Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Universidad de Buenos Aires

Facultad de Ciencias Económicas

Escuela de Estudios de Posgrado

Carrera de Especialización en Métodos Cuantitativos para la Gestión y Análisis de Datos en Organizaciones

Trabajo Final de Especialización

Componentes principales de las contrataciones públicas de Buenos Aires.

Un análisis de los pliegos de bases y condiciones.

Autor: Félix Pedro Penna

Tutor: Roberto Abalde

[Diciembre 2019]

Resumen

El Gobierno de la Ciudad de Buenos Aires (GCBA) gestiona las compras y contrataciones de bienes y servicios de todo el sector público local a través del sistema Buenos Aires Compras (BAC). La información relativa a cada uno de los procesos de contratación se publica anticipadamente al los llamados a licitación, para permitir a los interesados evaluar condiciones y planificar ofertas. El análisis de esta información y el posterior trabajo de planificación conlleva una erogación significativa de recursos para los potenciales proveedores, que se desarrolla en un marco de competencia e incertidumbre sobre los resultados de la licitación.

Conocer los patrones que comparten aquellos procesos de convocatoria que resultan exitosos y diferenciarlos de aquellos que resultan cancelados es información relevante para la toma de decisiones de quienes deciden proveer al Gobierno, en tanto les permite reducir la incertidumbre respecto a los desenlaces de los procesos de contratación.

La cantidad de información disponible para analizar hace que, de querer aprovecharla por completo, escape a las posibilidades de los softwares clásicos de análisis de datos. Mediante el Análisis de Componentes Principales implementado en el lenguaje de programación Python, se extrae información relevante acerca de los patrones de comportamiento del conjunto de datos abiertos bajo análisis, obteniendo en qué medida los procesos de convocatoria que resultan completados exitosamente se distancian de aquellos que fracasan al cabo del período estipulado para su adjudicación.

Palabras clave:
Statistical Methods
Local Government Expenditures
Computer Programs
Open Data

Estructura

Introducción

El gasto del Estado en contrataciones públicas representan, en promedio, entre un 10% y un 15% del PBI en los países de América Latina, lo cual lo constituye como un factor crucial del crecimiento y el desarrollo económico de los países. En Argentina, el Gobierno Nacional y el Gobierno de la Ciudad de Buenos Aires (GCBA) gestionan de forma separada la adquisición de bienes y servicios, haciendo uso del presupuesto que disponen cada uno de los niveles de gobierno. (Volosin, 2015)

El GCBA dispone de un presupuesto millonario[1]. La cantidad de dinero que el Gobierno tiene a disposición para adquirir bienes de uso y servicios no personales (30 % del presupuesto ejecutado durante 2018 se destinó a la compra de bienes de uso y a la contratación de servicios no personales), hace que se constituya como el principal cliente de diversas compañías proveedoras de bienes y/o servicios y una oportunidad de negocio relevante para aquellas compañías que actualmente no lo son.

Desde el año 2011 el GCBA gestiona las compras y contrataciones de bienes y servicios de todo el sector público local a través del sistema Buenos Aires Compras (BAC). La administración central, los entes descentralizados,el poder legislativo, el poder judicial y todas las entidades creadas por la constitución de la Ciudad, se rigen bajo la ley 2095, la cual regula las normas básicas de los procesos de compras y contrataciones.

El sistema electrónico BAC registra cada uno de los actos administrativos que constituyen un proceso de compras realizado en el marco de la ley desde el año 2011, da soporte a cada una de las etapas fundamentales del mismo. La información correspondiente a la planificación, la licitación o convocatoria, la adjudicación y la contratación resultante queda almacenada en bases de datos que desde el año 2017

[1] https://www.buenosaires.gob.ar/presupuestoabierto

se encuentran disponibles en formato abierto en el portal de datos abiertos del Gobierno de la Ciudad[2].

Los datos abiertos de compras y contrataciones se constituyen como un insumo fundamental para los potenciales proveedores del gobierno: ponen a disposición información de relevancia para comprender el mercado de determinados tipos de bienes y servicios, contribuyen con la formulación acertada de presupuestos, brindan información al respecto de las capacidades del resto de los competidores, entre otras posibilidades vinculadas con la innovación de procesos y productos. (Prince, Jolias, Brys, 2013).

La cantidad de procesos de compras gestionados a través del BAC es creciente con el correr de los años. En 2014 la cantidad de procesos gestionados superó por primera vez las tres cifras (1488 procesos) y en 2018 se obtuvo la máxima cantidad registrada (6684 procesos). En consecuencia, la cantidad de datos disponibles para procesar y extraer información es lo suficientemente grande como para necesitar de un lenguaje de programación que logre manipular eficientemente la información. Asimismo, justifica el uso de algoritmos de aprendizaje automático para la identificación de patrones, variables relevantes y predicciones a través del entrenamiento de modelos.

La información incorporada en los pliegos de bases y condiciones de de cada uno de los procesos, se publican con la anticipación suficiente para garantizar que los potenciales proveedores de los ítems a contratar puedan estudiar las características de las ofertas, hacer consultas a las autoridades responsables de las compras públicas y decidir en un plazo preestablecido si participarán o no de las convocatorias

El análisis de las condiciones de los llamados a licitación y la preparación de ofertas conllevan un costo de recursos económicos y tiempo significativos, lo cual se desarrolla en un marco de competencia con otros potenciales adjudicatarios y en un escenario de incertidumbre sobre el resultado de la licitación.

[2] https://data.buenosaires.gob.ar

Conocer de forma anticipada el resultado más probable de las licitaciones de las compras de bienes y servicios es relevante para la previsibilidad y planificación de cualquier potencial proveedor. Permite ahorrar tiempos y costos de postulación de ofertas y asignar probabilidades al éxito en la adjudicación de los procesos.

El objetivo de la presente investigación es conocer la influencia de las características y condiciones de los pliegos de compras y contrataciones en el resultado de los procesos de compras llevados adelante por las distintas entidades del Gobierno de la Ciudad de Buenos Aires a través del sistema BAC. De esta manera, brindar información relevante para hacer más eficiente la gestión de los procesos de compras hacia dentro de la administración del GCBA, a la vez de proveer de una herramienta para que los actuales o potenciales proveedores puedan hacer un mejor uso de sus recursos.

Asimismo, en tanto objetivo específico, explorar patrones en los datos que permitan comprender, representar y diferenciar los procesos licitatorios exitosos (entendiendo como exitosos a aquellos procesos que resultan completos) de aquellos que resultan cancelados o desiertos. En ese sentido, será relevante identificar las características que resultan relativamente más importantes en la determinación del resultado de los procesos en cuestión.

Por último, presentar las implicancias y el potencial de implementación del método en problemas de negocio concretos, así como plantear las oportunidades que se presentan para el sector público al abordar desarrollos de este estilo.

Aprendizaje Automático y los datos de contratación pública

Las compras y contrataciones públicas constituyen un factor clave para el desarrollo de las funciones del sector público. Las distintas áreas del gobierno necesitan de ellas para cumplir con sus misiones y funciones (educación, salud, desarrollo económico, entre otras). Con el fin de promover la transparencia en lo que se considera el mercado más grande del mundo (según estimaciones de la organización Open Contracting Partnership los gobiernos de todo el mundo gastan alrededor de 9.5 billones de dólares en contratar bienes y servicios con compañías privadas), algunos países y ciudades publican en formato abierto sus datos de contratación pública.

Dadas las dimensiones del mercado, la cantidad de datos que publican los gobiernos nacionales y subnacionales hacen que sea un desafío interpretar la información que estos contienen. El uso de algoritmos y la capacidad de cómputo son herramientas centrales para la extracción de información de todos esos *Gigabytes* de datos, es por esta razón que los algoritmos, la estadística descriptiva, las técnicas de aprendizaje automático y los datos abiertos de contrataciones públicas se encuentran estrechamente relacionados.

1.1. Los datos abiertos del sector público: qué son y dónde encontrarlos

El Gobierno Abierto

Es necesario, antes de comenzar a hablar sobre datos abiertos, hacer mención al cambio de paradigma y cambio cultural hacia dentro de la gestión pública, donde la misma pasó de utilizar un modelo analógico, hermético y autorreferente, a uno digital, abierto, colaborativo y multidireccional. De esto último se trata el "Gobierno Abierto", una nueva forma de gobernar y de establecer vínculos entre el gobierno y los ciudadanos.

Una gestión gubernamental abierta supone una fluida comunicación e interacción recíproca entre gobierno y ciudadanía; la apertura de canales de diálogo para aprovechar la potencial contribución de los ciudadanos en el diseño de políticas públicas, la producción de bienes y servicios, el monitoreo y el control de gestión. Asimismo, precisa que los ciudadanos se involucren en estos nuevos canales participativos ocupando los diferentes roles en los que se le da lugar (Oslak, 2012).

El concepto de Gobierno Abierto se consolidó en el año 2009, tras la publicación del *Memorándum* de "Transparencia y Gobierno Abierto" publicado por el presidente de Estados Unidos de Norteamérica, Barack Obama. En el citado documento el presidente compromete a su administración a generar un nivel de apertura gubernamental sin precedentes y asimismo indica que el gobierno "*debe ser transparente(...) debe ser participativo(...) debe ser colaborativo".*

El proceso de digitalización del estado es un factor determinante en la producción y recopilación de datos de todo tipo y de gran volumen que surgen como resultado de las funciones básicas de la administración pública. El gobierno abierto cambia el paradigma de gestión de esos datos, los hace públicos y promueve el acceso y uso por parte de los ciudadanos. Daniel Lathrop y Laurel Rume, en su libro Open Government, mencionan que la información pública basada en datos abiertos es una forma de infraestructura, tan importante como otras infraestructuras y que la magia de los datos abiertos radica en que garantiza la transparencia y promueve la innovación, tanto por parte del Estado, como por parte de los ciudadanos.

Los datos abiertos en la Ciudad de Buenos Aires

Los datos abiertos, siguiendo lo que especifica la "Carta Internacional de Datos Abiertos", son aquellos registros digitales, puestos a disposición de forma libre, sin restricciones, de manera que puedan ser usados, reutilizados y distribuidos por cualquier persona, en cualquier momento y en cualquier lugar. Éstos se encuentran en el centro de una significativa transformación global que se da por la

tecnología y los medios digitales, con un enorme potencial para promover gobiernos, sociedad civil y organizaciones del sector privado más transparentes.

Los datos abiertos habilitan a las empresas, la sociedad civil, la academia y la ciudadanía en general a generar valor a través de datos públicos y permiten el control social a las autoridades respecto de lo que estas estan haciendo con los fondos que administran en función de sus roles. Por otra parte, habilitan al gobierno a devolver a la sociedad un activo que resulta de actividades financiadas por ella misma.

Entre las consecuencias del uso de datos abiertos, se ha demostrado la existencia de impacto económico a través de la generación de valor agregado en nuevas plataformas y soluciones tecnológicas, la reducción de los niveles de corrupción y en consecuencia el aumento en la eficiencia del uso de los recursos del Estado y, por último, el empoderamiento de los ciudadanos para que puedan tomar mejores decisiones.

El acceso a los datos abiertos es un proceso que suele estar regulado por normativas y normalmente respeta estándares internacionales. Las recomendaciones sobre información pública de la OCDE y los ocho principios sobre los datos abiertos, identificados por Tim O´Reilly son el marco de referencia para la apertura de datos en el Gobierno. Los ocho principios consisten en: Completitud, en tanto toda la información es pública y por lo tanto no deben entregarse datos incompletos; Primarios, la fuente de la información debe ser primaria y debe publicarse tal como fué obtenida; Oportuna, la información publicada debe tener una temporalidad adecuada; Sin restricción de acceso, la información debe estar disponible para todos los tipos de usuarios; Procesable, los datos entregados deben poder ser procesados por una computadora; De acceso no discriminatorio, la información debe estar disponible para cualquier persona sin necesidad de registro previo; No propietaria, los datos no pueden estar en formatos asociados a una compañía o entidad, que requiera de un software propietario para

su uso; Licenciamiento libre, la información no debe estar sujeta a ningún tipo de derecho o patente. (Barros, 2012)

Para el caso de la Ciudad de Buenos Aires, el distrito donde hace foco el presente trabajo, dos decretos (156/2012 y 478/2013) establecen la creación del portal de datos abiertos del Gobierno de la Ciudad (data.buenosaires.gob.ar) y la apertura por defecto de todos los datos producidos por el gobierno (a menos que haya una normativa vigente que indique lo contrario). Asimismo, la Ley 104 de Acceso a la Información Pública y la Ley 1845 de Protección de Datos Personales, regulan la publicación y acceso a los datos e información.

Desde el año 2013, el Gobierno de la Ciudad de Buenos Aires publica los datos producidos por la administración pública local en su propio portal de datos abiertos. Los datos de Compras y Contrataciones públicas son uno de los conjuntos de datos destacados del portal, en tanto contiene el detalle de todos los procesos de contratación de bienes y servicios gestionados por las diferentes áreas del Gobierno a través del sistema electrónico de contratación "Buenos Aires Compras (BAC)"[3]. Mediante el sistema BAC se gestionan los procesos de compras desde su convocatoria a proveedores, pasando por su adjudicación, contratación y ejecución de los contratos. El detalle de la información publicada constituye a este conjunto de datos como uno de los activos más importantes para el desarrollo de investigaciones que intentan comprender la forma en la que el Gobierno de la Ciudad administra los fondos públicos.

1.2. Estándar de datos para la contratación abierta

El set de datos que contiene los registros de todos los procesos de compras gestionados mediante el sistema electrónico de contratación BAC, se encuentra constituído por un único recurso estructurado en formato Java Script[4] que sintetiza

[3] https://www.buenosairescompras.gob.ar/
[4] Javascript object notation, es un formato estructurado de texto sencillo, utilizado para el intercambio de datos.

y ordena la información. Esto se debe a que el Gobierno de la Ciudad de Buenos Aires adoptó el Estándar para las Contrataciones Abiertas (conocido como OCDS por sus siglas en inglés) en la publicación de sus datos de contratación en marzo del 2019.

¿Qué implicancias tiene la adopción de un estándar internacional para publicar datos de contratación?

Los datos que contienen los detalles de la contratación del Gobierno son una herramienta potente para realizar un seguimiento integral de los procesos de gestión de compras, obtener conocimiento sobre lo que ocurre con ellos, conocer los detalles de cada uno de los contratos firmados, corregir errores y eficientizar procesos. Estas bondades requieren de que los datos y los documentos de contratación se encuentren disponibles de forma estructurada y reutilizable.

Adoptar un estándar internacional garantiza la posibilidad de reutilización de los datos y el desarrollo de análisis e investigaciones a lo largo del tiempo. Así se habilita de forma permanente la posibilidad de análisis, se homogeniza el lenguaje del proceso de contratación, se simplifica su comprensión y se da lugar a la comparación de la información entre distritos que publican la información bajo las mismas reglas y estructuras.

Los datos utilizados para el presente trabajo de investigación se encuentran disponibles en formato abierto (estructurado, no propietario) csv o JSON en el portal de Datos Abiertos del Gobierno de la Ciudad de Buenos Aires y respetan el Estándar de Datos para la Contratación Abierta promovido por la organización Open Contracting Partnership[5].

Estándar de Datos para la Contratación Abierta

El OCDS es un estándar global de datos sin propietarios, utilizado para reflejar los ciclos de contratación de bienes, servicios, recursos y obras de forma completa, para las organizaciones del sector privado, el sector público y la sociedad

[5] https://www.open-contracting.org/data-standard/?lang=es

civil. El estándar permite a los usuarios de alrededor del mundo disponibilizar sus datos en formato reutilizable y legible por computadoras, con el objetivo de generar valor a partir de ellos, extraer información y crear herramientas que faciliten su comprensión.

El estándar fue diseñado y desarrollado para conectar los registros y documentos relativos a la contratación que recopilan las distintas organizaciones, con las necesidades de los usuarios: el mismo gobierno, la academia, los ciudadanos, la sociedad civil, el periodismo, entre otros.

Cada publicación de datos que bajo el estándar en cuestión, está compuesto por las siguientes etapas y secciones:

- Partes interesadas: el detalle de los participantes (internos y externos a la organización) que se ven involucrados en alguna parte del proceso de contratación.
- Planificación: contiene la información relativa a los antecedentes de un proceso de contratación. Puede contener información sobre el presupuesto de donde salen los fondos para la ejecución del contrato o los detalles de los proyectos que requieren de esa contratación.
- Convocatoria o llamado: contiene la información relativa a los anuncios de una organización que pretende adquirir determinados bienes, trabajos o servicios.
- Adjudicación: es la sección que anuncia las adjudicaciones relacionadas a cada uno de los procesos de convocatoria, siempre y cuando el proceso de selección haya resultado exitoso.
- Contratación: describe los detalles de los contratos perfeccionados luego de la adjudicación.
- Implementación: muestra el avance de la ejecución de cada uno de los contratos perfeccionados.

La publicación de datos de contratación del Gobierno de la Ciudad de Buenos Aires incluye tres de las etapas listadas: convocatoria o llamado;

adjudicación y contratación. Haciendo foco específicamente en la primera de las etapas, es desde donde se desarrolla la investigación incluida en este trabajo: la información contenida en la etapa de convocatoria está compuesta por variables cuantitativas y cualitativas que hacen un total de 37, con el siguiente detalle:

Tabla 1: Detalle de las variables del conjunto de datos

Variable	Descripción
ocid	Identificador global del proceso de compras
id	Identificador único del proceso de compras
date	Fecha del proceso
initiationType	Tipo de inicio del proceso
tag	Etiquetas del proceso
tender/id	Id de convocatoria
tender/title	Título de la convocatoria
tender/description	Descripción de la convocatoria
tender/status	Estado de la convocatoria
tender/procuringEntity/id	Id de entidad procuradora
tender/value/currency	Moneda de la convocatoria
tender/value/amount	Monto de la Convocatoria
tender/procuringEntity/name	Nombre de la entidad procuradora
tender/procurementMethod	Método de convocatoria
tender/procurementMethodDetails	Detalle del método de convocatoria
tender/additionalProcurementCategories	Categoría de la convocatoria
tender/tenderPeriod/startDate	Fecha de inicio de la convocatoria
tender/tenderPeriod/endDate	Fecha de fin de la convocatoria
tender/tenderPeriod/durationInDays	Duración de la convocatoria
tender/enquiryPeriod/startDate	Fecha de inicio del período de consultas
tender/enquiryPeriod/endDate	Fecha de fin del período de consultas
tender/enquiryPeriod/maxExtentDate	Fecha de cierre del período de consultas

language	Idioma
tender/items/0/id	Identificador del ítem a contratar
tender/items/0/description	Descripción del ítem a contratar
tender/items/0/quantity	Cantidad de ítems a contratar
tender/items/0/unit/name	Unidad de medida del ítem
tender/items/0/unit/scheme	Nombre del catálogo de unidades de medida del ítem
tender/items/0/classification/scheme	Nombre del catálogo de clasificación del ítem
tender/items/0/classification/id	Identificación del ítem
tender/items/0/unit/value/amount	Monto unitario del ítem
tender/items/0/unit/value/currency	Moneda del monto unitario
tender/documents/0/id	Identificador de los documentos del proceso
tender/documents/0/documentType	Tipo de documento
tender/documents/0/url	Link al documento
tender/documents/0/datePublished	Fecha de publicación de los documentos
tender/documents/0/language	Lenguaje de los documentos

Fuente: Elaboración propia en base a documentación de la publicación de los datos

1.3. Reutilización de los datos de contratación

Para que los datos abiertos que son publicados por organismos públicos sean reutilizados se requiere de garantizar una serie de condiciones que faciliten ese objetivo y que exceden a la selección de formatos y estándares de publicación. El escenario ideal es aquel donde el uso de los datos es completamente accesibles y no requiere permisos específicos, pero eso pocas veces ocurre en la práctica. La realidad indica otras condiciones, los datos suelen estar sujetos a derechos de propiedad intelectual y es por esta razón que la publicación de los mismos se acompaña por licencias de uso, que establecen un marco de actuación entorno a la

manipulación de los datos, propiciando una amplia difusión y capacidad de utilización por parte de los usuarios finales.

Una vez establecidas las condiciones de uso de los datos abiertos[6], la reutilización de los mismos sucede dentro de un "ecosistema de datos abiertos". Allí, el organismo es la parte que provee de los datos subiendolos a portales centralizados; luego, desarrolladores y académicos procesan los datos y crean servicios; a continuación, los ciudadanos consumen la información, redundando en un mayor control, eficiencia en la administración pública y mayor apertura de datos.

Los datos de compras y contrataciones, sobre los que hace foco este trabajo, suelen utilizarse para evaluar cartelización de empresas, manipulación de ofertas, comportamiento colusorio, falta de competitividad, fraude e irregularidades en la contratación. En ocasiones se combinan técnicas de aprendizaje automático y de análisis estadístico inferencial para predecir la colusión a través de carteles de manipulación de ofertas, donde se destaca la implementación del modelo de regresión *Lasso* y ensambles de árboles de decisión (Huber, Imhof, 2018). Por otra parte, métodos de aprendizaje no supervisado también son utilizados para detectar patrones de comportamiento (Heijnen, Haan, Soetevent, 2015). También existen casos de utilización de redes neuronales y redes neuronales profundas en la predicción de comportamiento irregular en los procesos licitatorios utilizando datos de compras y contrataciones públicas (Sun, Sales, 2018). En cuanto a la selección de variables relevantes para explicar los resultados de los procesos, la literatura presenta una serie de técnicas que se engloban en la práctica denominada "*Feature Engineering*" (Géron, A., 2017) y por otro lado métodos de estadística clásica como los análisis de correlación y de regresión múltiple (Miñano Pérez, Castejón Costa, 2018).

[6] Para el caso de la Ciudad de Buenos Aires, la licencia bajo la cual se publican los datos abiertos es la Creative Commons 2.5 by AR. Permite copiar, distribuir, adaptar y transformar la información siempre y cuando se lo indique haciendo referencia a la fuente de los datos.

A pesar de que en la literatura no abundan los ejercicios prácticos donde se evalúen los resultados de los procesos licitatorios en base a los detalles y características del mismo, las técnicas utilizadas para abordar dicha problemática en este desarrollo son las que se encuentran frecuentemente en los ejercicios que utilizan conjuntos de datos de contratación.

Aquí se estudiarán las particularidades de la contratación pública del Gobierno de la Ciudad de Buenos Aires desde un enfoque cuantitativo, exploratorio y analítico. En primer lugar, el análisis exploratorio de los datos servirá para comprender el comportamiento de las variables que hacen a la problemática tratada; luego, se presentará un ejercicio de implementación de Análisis de Componentes Principales a fin de visualizar e interpretar el comportamiento de los procesos de compras públicas, extraer patrones y comprender sus características principales.

Se pondrá foco en la etapa de convocatoria a proveedores. En ella, se especifican una serie de características que hacen al proceso de contratación (plazos para realizar consultas, duración de la etapa licitatoria, monto de la contratación, bienes o servicios a contratar, entre otros). Luego, una vez que el período de convocatoria ha terminado, se especifica el resultado del mismo, si resultó adjudicado, desierto, cancelado o si ha finalizado satisfactoriamente. Estudiando las características del proceso de convocatoria, se podrá identificar aspectos en común de los procesos exitosos y contrastarlos con aquellos que no lograron alcanzar los resultados esperados, a fin de enunciar las causas posibles de tales resultados.

Aspectos metodológicos del algoritmo PCA

Un problema central en el análisis multivariante es el de reducción de la dimensionalidad: si es posible representar a un conjunto de variables P en un menor número de variables r, se ha reducido la dimensión a costa de una pérdida de información. El análisis de componentes principales (PCA) es el algoritmo seleccionado para representar la etapa de convocatoria de los procesos de compras del Gobierno de la Ciudad de Buenos Aires.

2.1 El análisis de datos multivariantes

Para describir cualquier situación real, las características de cualquier proceso o situación, hace falta tener en cuenta simultáneamente a una serie de variables. El análisis de datos multivariante se compone de métodos estadísticos que permiten estudiar los elementos de una población que se describen a través de una multiplicidad de variables de distinto tipo: cualitativas y cuantitativas, medidas en distintas escalas y siguiendo distintas distribuciones.

El análisis multivariado persigue un listado de objetivos específicos orientados a resumir las variables observadas, encontrar grupos o asociaciones entre las observaciones, clasificar nuevas observaciones en grupos existentes y relacionar dos conjuntos de variables. Obtener indicadores que resuman la información de las variables originales, con una pequeña pérdida de información tiene sus ventajas: permiten graficar la información para facilitar la interpretación y el análisis, comparar distintos grupos de datos o momentos del tiempo y comprender la realidad en un formato más accesible. Encontrar grupos permite hacer una división o diferenciación entre observaciones puntuales y encontrar aspectos que estas tienen en común y cómo se diferencian entre si. Si los grupos se encuentran definidos a priori, la clasificación permite vincular a nuevas observaciones en estos grupos preestablecidos de forma de asociar a estas últimas con las características de

los grupos originales. Como condición necesaria de los tres objetivos anteriores, se encuentra el de relacionar conjuntos de variables entre sí a fin de comprender su estructura y relación de dependencia.

El análisis multivariado tiene aplicación en distintos campos de la ciencia, desde la biología hasta la psicología y las ciencias sociales, donde las aplicaciones de los métodos para problemas específicos han contribuido con el desarrollo del análisis estadístico en conjunto.

Parte relevante del análisis multivariado es el análisis gráfico y de valores atípicos, que contribuyen con una descripción homogénea y de fácil interpretación de los datos. En el presente trabajo, el análisis de valores atípicos es necesario para la implementación del algoritmo seleccionado. De esta manera, se evita distorsionar la matriz de correlación con valores heterogéneos respecto a la distribución del resto de los datos y se permite una mejor representación gráfica.

Los datos atípicos son aquellas observaciones que parecen haberse generado de forma distinta al resto de los datos. Pueden ocurrir por errores de medición o por la heterogeneidad intrínseca de la variable estudiada.

Los efectos de los datos atípicos pueden ser negativos y graves. Una sola observación puede distorsionar la media y el desvío estándar de la distribución, a la vez de destruir las relaciones existentes en ellas, mientras que también tiene influencia en la matriz de correlación entre las distintas variables que constituyen el análisis.

Hay dos formas fundamentales para tratar con la heterogeneidad de los datos, la primera es utilizar estimadores robustos, diseñados para ser poco afectados por la contaminación de los atípicos, la segunda es la detección de atípicos y eliminación de los mismos para trabajar con datos homogéneos. En cuanto a la segunda forma, utilizada en este trabajo, el primer paso para implementarla es identificar observaciones sospechosas utilizando diagramas de caja o la fórmula:

$$(1)\quad \frac{|x_i - med(x)|}{Meda(x)} > 4$$

donde $med(x)$ es la mediana de las observaciones y $Meda(x)$ es la mediana de las observaciones absolutas. Luego, una vez detectado, eliminar las observaciones del conjunto de datos es el segundo paso. A continuación, puede repetirse el procedimiento para detectar valores atípicos de la nueva distribución, hasta que no queden valores que cumplan el criterio de la función (1).

2.2 La utilidad del PCA en el análisis de datos de contrataciones

El Análisis de Componentes Principales es un método algebraico y estadístico útil para representar patrones en los datos y resaltar las similaridades y diferencias entre ellos, destacándose entre las técnicas de análisis exploratorio de los datos. Intenta describir la información que contienen los datos utilizando un número menor de variables que las que constituyen el conjunto de datos original, partiendo de la idea de que si las variables originales se encuentran con algún grado de correlación será posible sustituirlas por un nuevo conjunto de variables menor, sin grandes pérdidas de información.

Hottelling en 1933 fue quien desarrolló la técnica para permitir representar óptimamente observaciones en una dimensión reducida respecto a la original, el primer paso para identificar posibles variables latentes o no observadas, que están generando variabilidad en los datos. Siendo a su vez, una técnica que facilita la representación e interpretación de los datos.

El procedimiento consiste en buscar ejes o dimensiones que sean combinación lineal de las variables originales de forma que se minimice la pérdida de información inicial siempre y cuando estos sean linealmente independientes. Partiendo de estas condiciones, el objetivo básico del método consiste en reducir el número de variables originales, tomando como nuevas a los ejes o componentes hallados, considerando seleccionar un número tal que la pérdida de varianza total sea reducida y aún así se cumplan los objetivos de simplificar y reestructurar la información inicial. La reducción de la dimensión va a permitir simplificar posteriores análisis, que se harán a partir de un menor número de variables;

representar gráficamente las observaciones en una dimensión reducida; interpretar las relaciones entre las variables observadas y eliminar la información redundante.

Implementación del método

Partiendo de una matriz que contenga el conjunto de datos a analizar, habiendo determinado que al menos un par de las variables representadas en los datos tienen correlación distinta de cero, hace falta observar las unidades de medidas de las mismas (todas ellas deben ser cuantitativas). Si estas están representadas en escalas muy distintas entre sí, será necesario implementar una transformación lineal a los datos a través de la estandarización. De esta manera, restando la media y dividiendo por el desvío estándar a cada una de las observaciones de cada variable, se obtiene una distribución con media igual a cero y varianza igual a uno, lo cual habilita a realizar el análisis de componentes principales utilizando la matriz de correlaciones de las variables. El método de componentes principales se ve afectado por la escala de las variables que utiliza, partiendo de que busca maximizar las varianzas de las componentes que resultan del mismo. Si un componente varía relativamente más que otro por la escala de la variable en que están medidos, el método va a determinar que la variable que maximiza la varianza es aquel que por su escala evidencia una mayor dispersión. (Peña, 2012)

Luego, el método requiere de calcular los autovalores y autovectores a partir de la matriz de correlación. En primer lugar se extraen los autovalores y se los ordena de mayor a menor; a continuación, se extraen los autovectores de cada autovalor, sobre los cuales se construyen las componentes principales. Existen tantas componentes principales como variables originales tenga el conjunto de datos y se las ordena en orden de importancia según su autovalor asociado. La varianza de cada componente principal es equivalente al valor de su autovalor asociado y constituye una medida de la información que ese componente puede explicar.

Una vez obtenidos todos los componentes principales, es necesario decidir con cuántos de ellos se va a representar el problema original. Para resolver esto, hace falta determinar el porcentaje de la variabilidad total que se quiere incorporar en la representación y el que se va a obviar. Si los autovalores asociados a las componentes que se descartan son bajos, la pérdida de información será relativamente poca. (Smith, L. I. 2002)

Para representar gráficamente a cada una de las observaciones en los nuevos ejes seleccionados (cada uno de los CP), hace falta obtener la "puntuación" de la observación: es la coordenada de cada una de ellas en las componentes elegidas. Esta operación nos devuelve los datos originales, expresados en las componentes que seleccionamos para representar el problema. (Peña, 2012)

2.3 Implementar PCA en Python

Python es un lenguaje de programación simple y potente, se reconoce como el lenguaje más popular para la computación científica. Utilizado tanto en la industria como en la academia, es especialmente elegido para el desarrollo de algoritmos y de análisis y explotación de datos.

Basado en el lenguaje de programación C, de código abierto y disponible para una variedad de sistemas operativos. Es un lenguaje útil para la interpretación del código gracias a la sintaxis que utiliza. Entre sus ventajas se encuentra la curva de aprendizaje rápida, que permite iniciarse en la programación en modo texto. Contiene una gran cantidad de librerías, tipos de datos y funciones incorporadas que facilitan la producción y desarrollo de código, sin la necesidad de escribirlo desde cero. (Rossum, G. 1995)

A las facilidades de interpretación y producción, se le incorpora el rendimiento. Python es un lenguaje con mayor velocidad de cómputo y menor uso de memoria que, por ejemplo, lenguajes como R y Matlab. La madurez y estabilidad de las librerías numéricas y de cálculo (como Numpy y Scipy) sumado a

la calidad de la documentación disponible hacen de python un lenguaje atractivo y conveniente para diferentes audiencias y usuarios. (McKinney, W. 2011)

El uso de la librería Scikit Learn

Scikit learn es un módulo de Python que integra una amplia variedad de algoritmos de aprendizaje automático, tanto aquellos de aprendizaje supervisado como los de aprendizaje no supervisado. Este módulo mantiene una implementación orientada a brindar facilidad para el usuario y se encuentra completamente integrada con el lenguaje Python.

Scikit learn surgió como respuesta a la creciente necesidad de utilizar algoritmos implementados en lenguajes de programación por parte de aquellos interesados en analizar grandes cantidades de datos, aunque sin experiencia en las ciencias de la computación.

Entre los objetivos del proyecto de desarrollo de Scikit Learn se encuentra el de priorizar su fácil implementación, dejando en segundo lugar la incorporación de características y variables. Basa su desarrollo en el uso de herramientas colaborativas y de acceso libre para toda la comunidad de usuarios, a quienes alientan a incorporar comentarios y mejoras. La documentación de Scikit Learn incorpora más de 300 páginas de guías y documentación, tutoriales e instrucciones de instalación, para garantizar el uso del módulo por parte de usuarios de todo nivel y experiencia.

Implementar el análisis de componentes principales con el lenguaje de programación Python es una tarea que se ve simplificada por la existencia de librerías o bibliotecas. Estas concentran la implementación del algoritmo codificado en un lenguaje de programación, de forma que no sea necesario ejecutar cada uno de los cálculos que requiere el mismo, sino que estos se ejecutan de manera integrada al hacer uso de la biblioteca en cuestión.

La biblioteca que contiene el modelo de análisis de componentes principales en el lenguaje Python es Scikit Learn. Es de código abierto y adaptable a diversos

contextos de análisis e investigación, permite la implementación de una serie de modelos de aprendizaje automático y herramientas para minería y análisis de datos, entre los que se destacan las técnicas de regresión, clasificación, clustering y reducción de la dimensionalidad, siendo esta última la categoría donde se ubica el PCA.

PCA en python es un algoritmo utilizado para descomponer un set de datos multivariado, en un conjunto de componentes ortogonales que explican la máxima porción de la varianza total. El modelo se utiliza como un objeto de transformación, que consume el conjunto de datos original para construir componentes y luego puede expresar un nuevo conjunto de datos (sea el original u otro) en término de esos componentes seleccionados.

Una característica relevante de la implementación de PCA en python es la centralización de los datos sin escalarlos ni estandarizarlos. Por esta razón, en la implementación del algoritmo con datos cuyas variables revistan escalas muy diversas entre sí, es necesario estandarizar previamente cada una de ellas.

La instalación del algoritmo de Análisis de Componentes Principales en Python requiere de ejecutar una sencilla línea de código:

from sklearn.decomposition import PCA

Entre los parámetros disponibles para seleccionar, se encuentra la cantidad de componentes principales a extraer del set de datos utilizado. Una vez fijado dicho parámetro y entrenado el algoritmo, se permite la extracción de una serie de atributos, entre los que se destacan el puntaje de cada una de las observaciones respecto de cada componente principal, la varianza explicada por cada componente, el ratio de la varianza explicada por cada componente, los valores singulares de cada componente, entre otros.

(Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011))

Los plazos y montos preestablecidos, determinantes del resultado de los procesos

La primera parte del análisis de los datos de contratación pública del GCBA se realizó utilizando el algoritmo de Análisis de Componentes Principales con el lenguaje de programación Python. Esto permite realizar los cálculos que requiere el análisis de forma eficiente y rápida, a la vez de seleccionar entre distintas opciones para visualizar gráficamente la información.

3.1. Selección de variables analizadas

Teniendo en cuenta que el algoritmo PCA tiene como requisito para su utilización el uso de un conjunto de datos con variables cuantitativas, fue necesario llevar adelante un proceso de selección de variables relevantes para el análisis y el procesamiento. Del total de 37 variables originales, un grupo de 4 variables se encuentran dentro de las cuantitativas: tender/value/amount (el monto total por el que se est dispuesto a hacer la convocatoria), tender/tenderPeriod/durationInDays (la duración total en días del proceso de convocatoria), tender/items/0/quantity (la cantidad de bienes o servicios que se quiere contratar el el proceso en curso), tender/items/0/unit/value/amount (el precio unitario de los bienes o servicios en cuestión). Teniendo en cuenta que el monto total por el que se está dispuesto a hacer la convocatoria es equivalente a la multiplicación de la cantidad de ítems por el precio unitario de los mismos, se descartó para el análisis a la variable que contiene el precio unitario de los bienes o servicios a contratar. De esta manera el análisis puede enfocarse en los procesos de compras en lugar de los ítems que se buscan contratar en cada uno de esos procesos (por cada proceso puede contratarse más de un ítem distinto).

A continuación, al conjunto de datos compuesto por las tres variables cuantitativas relevantes y agrupado a nivel de proceso de convocatoria, se le extrajo

aquellas observaciones que evidenciaban información faltante y se seleccionaron los procesos de convocatoria cuyo procedimiento de selección de proveedores es el de Licitación Pública, el proceso bajo el cual compiten por la provisión de un bien o servicio las empresas e individuos que desean contratar con el estado. Por último, se acompañaron las variables cuantitativas por una selección de variables cualitativas, a fin de graficar las distintas observaciones segmentando las mismas según variables cualitativas de interés.

Para facilitar la interpretación de los ejercicios propuestos, se renombraron las variables cuyos nombres se extraían del set de datos original: tender/value/amount se puede encontrar a continuación como "Monto total pliego", "tender/tenderPeriod/durationInDays" como "Duración de la Licitación" y "tender/items/0/quantity" como "Cantidad de ítems".

3.2. Información contenida en las variables seleccionadas

Habiendo obtenido el conjunto de datos necesario, se procedió a realizar un análisis exploratorio estadístico de la información que éste contiene. Se seleccionaron los datos de procedimientos de licitación pública para los períodos 2017 y 2018, obteniendo el siguiente detalle:

Tabla 2: Estadísticas descriptivas del conjunto de datos 2017

	Monto total pliego	Duración de la licitación	Cantidad de ítems
OBS	1024.00	1024.00	1024.00
MEDIA	4952064.71	8.93	91951.09
DESVÍO	18764525.81	5.82	1467236.32
MÍNIMO	1.00	1.00	1.00
25%	251228.88	5.00	8.00
50%	1279130.00	7.00	221.00
75%	3275090.00	12.00	7200.00
MÁXIMO	300904271.88	56.00	43200096.00

Fuente: Elaboración propia

Tabla 3: Estadísticas descriptivas del conjunto de datos 2018

	Monto total pliego	Duración de la licitación	Cantidad de ítems
OBS	1382.00	1382.00	1382.00
MEDIA	15043609.54	9.46	361206.88
DESVÍO	254762715.93	5.74	8336381.48
MÍNIMO	5000.00	1.00	1.00
25%	270000.00	6.00	12.00
50%	1169851.49	8.00	332.00
75%	3340300.00	12.00	10000.00
MÁXIMO	6714315602.44	61.00	227302562.00

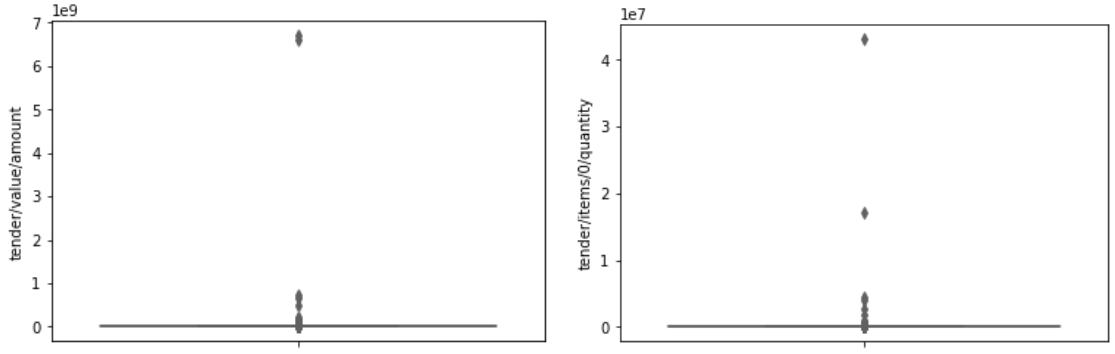
Fuente: Elaboración propia

El análisis estadístico descriptivo permite observar el comportamiento de los datos. En los períodos analizados se aprecia una dispersión significativa en la variable que indica el monto total del proceso de compras y en aquella que indica la cantidad de ítems que se espera contratar. Teniendo en cuenta el desvío estándar y el valor máximo y mínimo de ambas distribuciones, se puede hipotetizar con la existencia de valores atípicos.

Mediante gráficos de caja se confirma la hipótesis planteada, que da lugar al procedimiento de detección y eliminación de valores atípicos utilizando la fórmula 1 presentada previamente.

La detección de valores atípicos arrojó cantidades de observaciones superiores a 100 en ambos períodos analizados. Desechar más de 100 valores atípicos es probablemente una pérdida de información significativa del problema en estudio, por lo que se procedió a dividir la muestra en dos mitades iguales, utilizando a la mediana de cada uno de los períodos como valor límite entre ambas.

Gráfico: Gráfico de caja para detectar valores atípicos en el período 2017 y 2018



Fuente: Elaboración propia

De esta manera, el análisis se realiza sobre dos muestras del período 2017 y sobre dos muestras del período 2018. Para cada período, la primera muestra contiene las observaciones que corresponden a valores de la variable "Monto total pliego" que se encuentran por debajo de la mediana de la distribución original y la segunda muestra contiene las observaciones con valores de la misma variable que se encuentran por encima de la mediana de la distribución original.

En ambos períodos, la muestra 1, aquella que contiene a los valores por debajo de la mediana, presentó una distribución homogénea, por lo que no fué necesario eliminar valores atípicos. En la muestra 2, por el contrario, surgió la necesidad de realizar nuevamente la detección de valores atípicos y su posterior eliminación para homogeneizar el comportamiento de los datos.

La división de la distribución original de ambos períodos en valores por debajo y por encima de la mediana y posteriormente la eliminación de valores atípicos para la muestra 2, mejora el comportamiento de la distribución y conforma un conjunto de datos apropiado para realizar el análisis de componentes principales.

Sin valores atípicos la muestra de datos es más homogénea, no se encuentra condicionada por eventos con ocurrencia de baja frecuencia y mejora los resultados y la interpretabilidad del análisis algebraico y gráfico de los datos.

3.3. Utilizando PCA para comprender los procesos de compras

Como se indicó anteriormente, los datos de contratación pública están constituidos por un conjunto numeroso de variables y observaciones. Interpretar la información que proveen los datos no es tarea sencilla y es por eso que se recurre a herramientas y algoritmos estadísticos para extraer patrones y conclusiones de los mismos. El análisis de componentes principales logra aportar resultados relevantes en la comprensión de los datos bajo análisis, en tanto procede reduciendo la dimensión del conjunto original, a fin de simplificar la comprensión de los patrones que los datos evidencian.

Para efectuar el análisis, en primer lugar es necesario seleccionar un período temporal, para reducir el efecto distorsivo de la inflación sobre los precios de los bienes y en consecuencia de los montos de las convocatorias. En este trabajo se analizan por separado los datos correspondientes a los años 2017 y 2018, utilizando dos muestras para cada uno de los períodos.

Luego, se estandarizan los datos para cada uno de los períodos analizados, esto es condición necesaria para ejecutar el algoritmo, en tanto las variables de interés se encuentran expresadas en distintas escalas. A continuación, sobre los datos estandarizados se calculan las matrices de covarianza, obteniendo los siguientes resultados:

Tabla 1: Matriz de correlación 2017 muestra 1

	Monto total pliego	Duración de la licitación	Cantidad de ítems
Monto total pliego	1,0010	0.1057916	0.21337999
Duración de la licitación	0.1057916	1,0010	0.02225987
Cantidad de ítems	0.21337999	0.02225987	1,0010

Fuente: Elaboración propia

Tabla 2: Matriz de correlación 2017 muestra 2

	Monto total pliego	Duración de la licitación	Cantidad de ítems
Monto total pliego	1.00255102	-0.03362703	0.11085849
Duración de la licitación	-0.03362703	1.00255102	0.03780484
Cantidad de ítems	0.11085849	0.03780484	1.00255102

Fuente: Elaboración propia

Tabla 3: Matriz de correlación 2018 muestra 1

	Monto total pliego	Duración de la licitación	Cantidad de ítems
Monto total pliego	1.00144928	0.21781388	0.24891959
Duración de la licitación	0.21781388	1.00144928	0.07215991
Cantidad de ítems	0.24891959	0.07215991	1.0014492

Fuente: Elaboración propia

Tabla 4: Matriz de correlación 2018 muestra 2

	Monto total pliego	Duración de la licitación	Cantidad de ítems
Monto total pliego	1.00194175	0.09059408	0.11623805
Duración de la licitación	0.09059408	1.00194175	0.04474245
Cantidad de ítems	0.11623805	0.04474245	1.00194175

Fuente: Elaboración propia

La existencia de correlación entre las variables permite a posteriori reducir la dimensionalidad del conjunto de datos. En los registros correspondientes a la muestra 1 del año 2017 y a la muestra 1 del 2018, se observa un valor de correlación por encima del resto entre la variable que representa el monto total del pliego y la variable que contiene la cantidad de ítems que se espera contratar. Para procesos de compras con montos totales relativamente bajos, existe una relación directa entre éste y la cantidad de ítems que se espera contratar. En ambas muestras existe también correlación directa entre la variable que representa el monto total del pliego y la variable que indica la duración total del proceso de licitación.

En lo registros correspondientes a las muestras 2 de ambos períodos analizados, las correlaciones entre variables presentan valores más bajos. Existe una correlación directa entre la variable del monto total del proceso y entre la variable con la cantidad de ítems que se espera contratar, pero en este caso, la correlación entre el monto total del proceso y la duración total del proceso licitatorio es cercana a cero.

Partiendo de cada una de las matrices de correlación se realiza el cálculo de autovalores y autovectores. Los valores de los autovalores nos permiten conocer la varianza explicada por cada uno de los componentes principales asociados a ellos, es la información necesaria para determinar cuánta información estará dejando afuera la representación de los componentes principales con una dimensión de variables inferior a la original.

Autovalores 2018 Muestra 1

[1.37001561 0.70433861 0.92999361]

Autovectores 2018 Muestra 1

[0.6678198 0.74363843 0.03191545]
[0.50222115 -0.41854042 -0.75670194]
[0.54935474 -0.52136915 0.65298053]

Autovalores 2018 Muestra 2

[1.17291528 0.87406988 0.95884009]

Autovectores 2018 Muestra 2

[0.65240586 0.75261153 0.08912059]
[0.4957115 -0.33481817 -0.80135317]
[0.57326845 -0.5669856 0.59151552]

Autovalores 2017 Muestra 1

[1.24950126 0.77207284 0.98429675]

Autovectores 2017 Muestra 1

[0.69285977 0.71880657 0.0571178]
[0.35266032 -0.26870523 -0.89634156]
[0.62894836 -0.64118219 0.4396696]

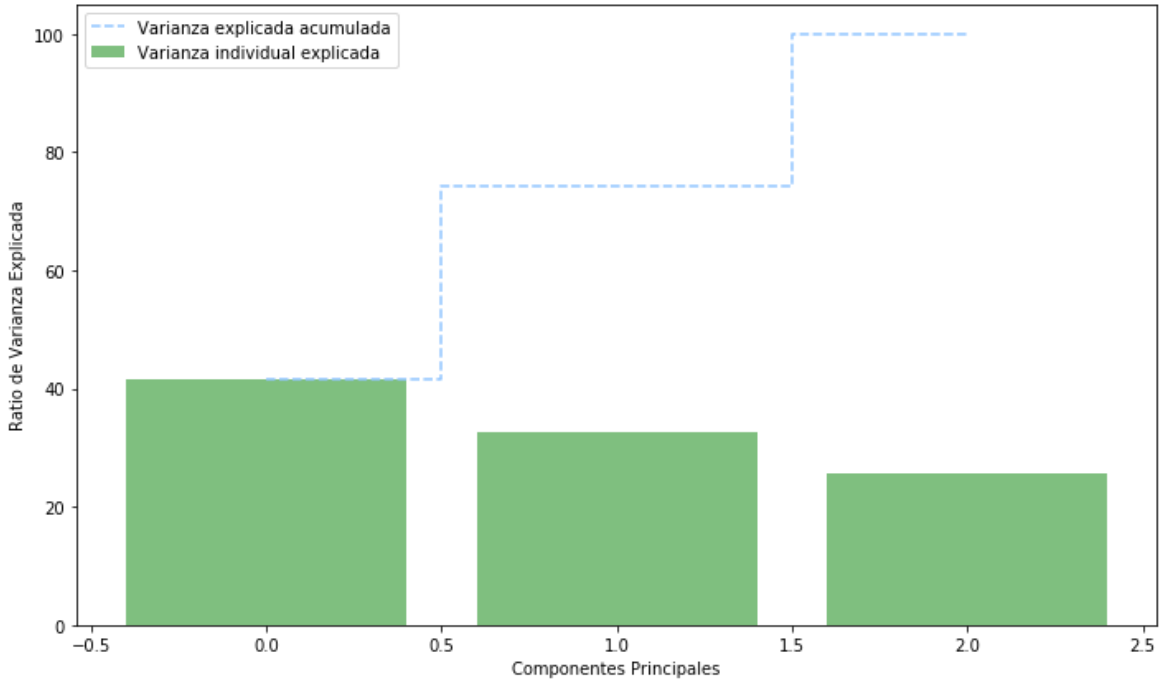
Autovalores 2017 Muestra 2

[[0.87212601 1.11349727 1.02202978]]

Autovectores 2017 Muestra 2

[0.65622876 0.70199753 -0.27670071]
[0.36121981 0.02969298 0.93200782]
[-0.66248326 0.71156011 0.23409003]

Gráfico 1: Varianza explicada por autovalores 2017 muestra 1

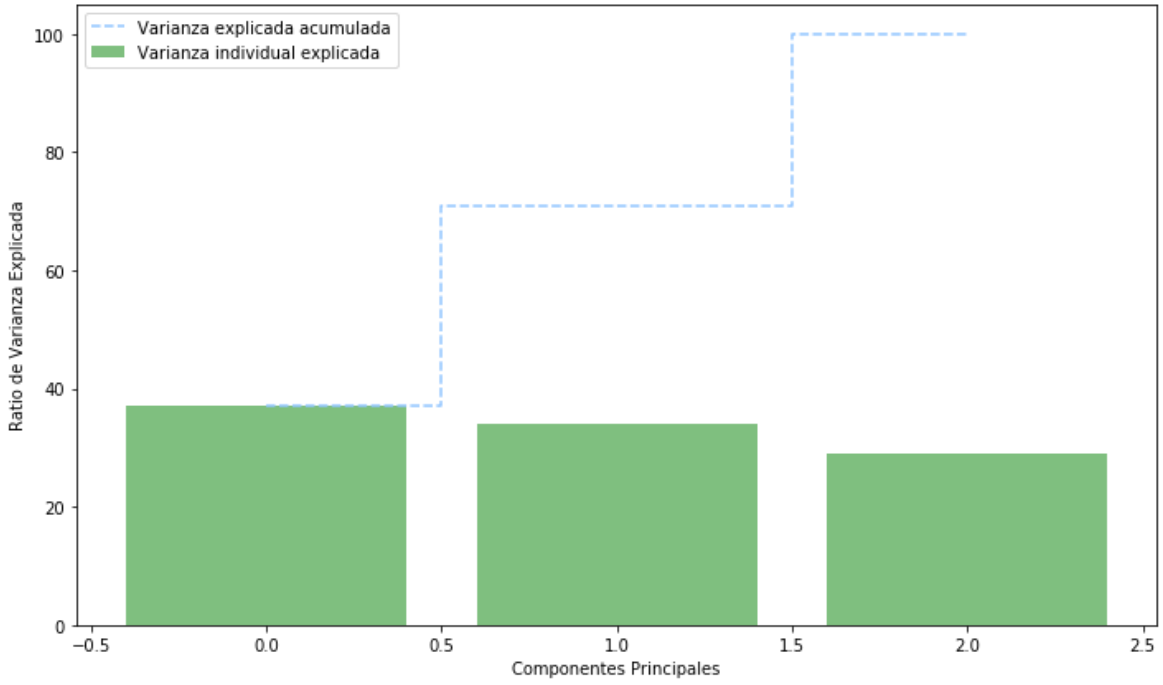


Fuente: Elaboración propia

Para el caso de la primera muestra del año 2017, las dos primeras componentes principales explican en total el 74.31% de la varianza total. Representar el conjunto de datos inicial con estas dos primeras componentes implica una pérdida de información superior al 25%, lo cual constituye una porción significativa de la varianza explicada total. Las dos primeras componentes principales explican respectivamente el 41.56% y 32.74% de la varianza.

Las dos primeras componentes principales de la segunda muestra del 2017 explican en conjunto el 71% de la varianza total. Representar el conjunto de datos inicial con estas dos componentes principales implica una pérdida de información del 29%, superior a la muestra 1.

Gráfico 2: Varianza explicada por autovalores 2017 muestra 2

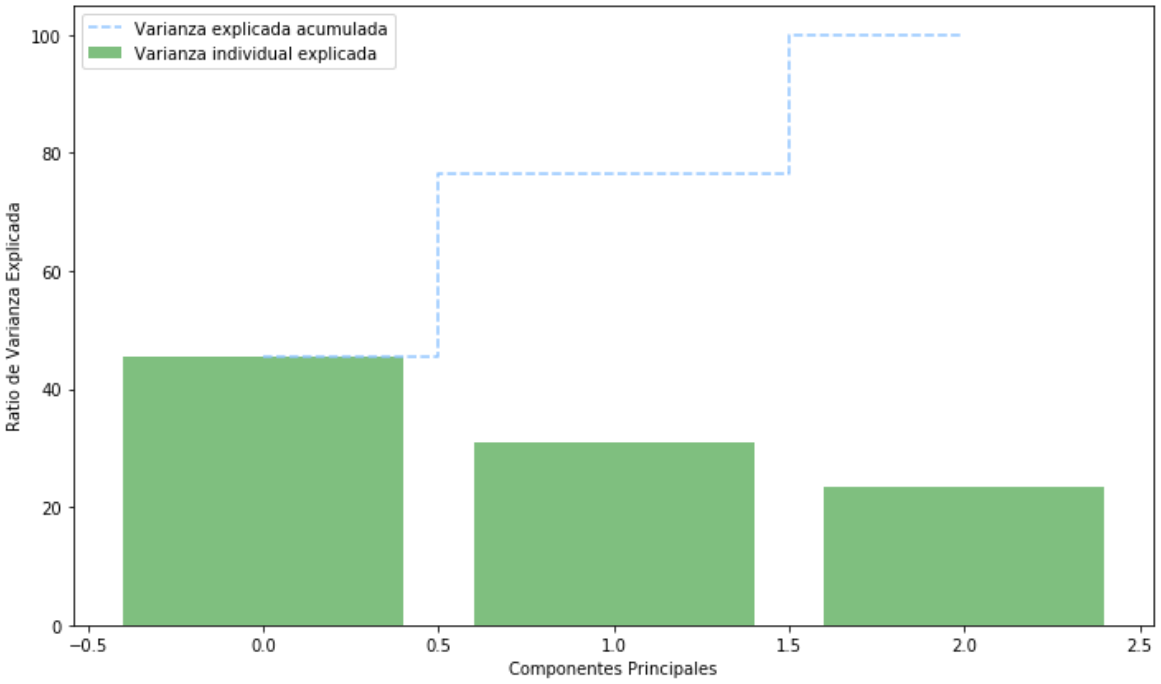


Fuente: Elaboración propia

Las componentes principales de los datos correspondientes a la muestra 1 de 2018 explican en conjunto el 76.55% de la variabilidad total. Al igual que para el caso de 2017, desestimar la tercera componente principal determina una pérdida de información significativa, en este caso cercana al 24%.

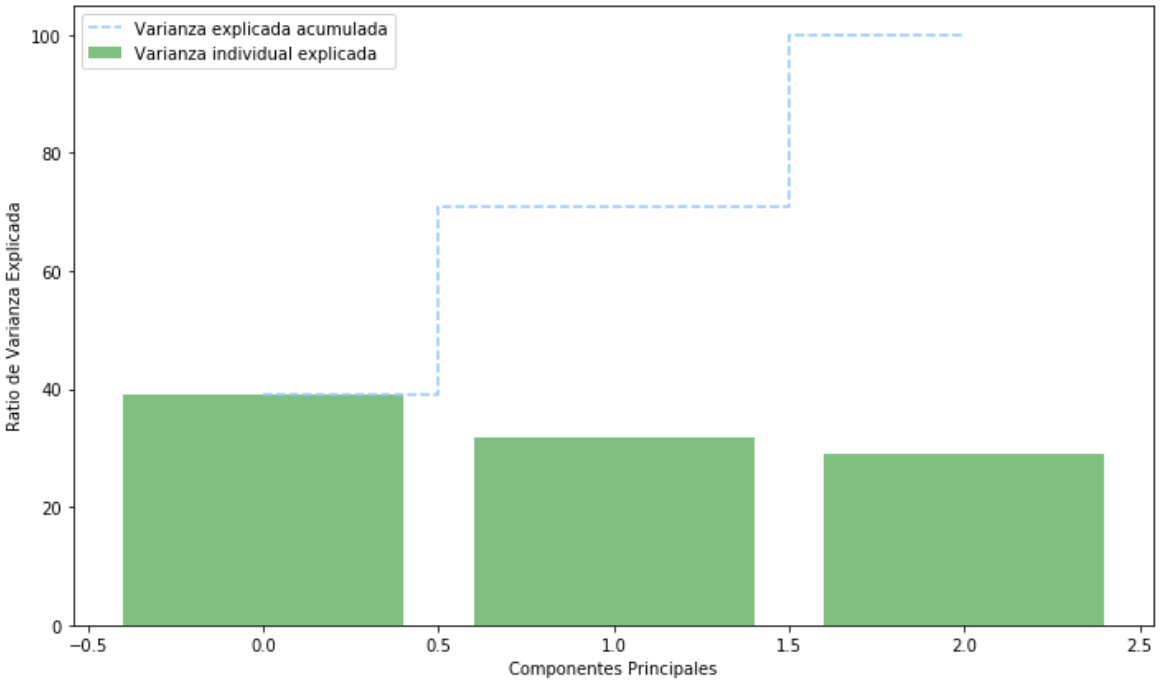
La muestra 2 del periodo 2018 se comporta de forma similar a la muestra 2 del período anterior. Las dos primeras componentes principales representan el 70.92% de la varianza total de los datos. Desestimar la tercera componente principal implica una pérdida de información del 29.07%, superior a lo que ocurre con la muestra 1 del período 2018.

Gráfico 3: Varianza explicada por autovalores 2018 muestra 1



Fuente: Elaboración propia

Gráfico 4: Varianza explicada por autovalores 2018 muestra 2



Fuente: Elaboración propia

Conociendo los valores de los autovalores y la varianza explicada que aportan cada una de las componentes principales asociadas es posible seleccionar la cantidad de componentes que va a contener el análisis. Para los cuatro casos explorados se seleccionan las dos primeras componentes principales, a fin de facilitar su representación gráfica posterior, aún sabiendo que el escenario donde se presenta la menor pérdida de la información es en las muestras que contienen los valores relativamente más bajos de la distribución de la variable que explica el monto total del proceso de compras (muestra 1).

A través de la carga factorial de cada una de las variables originales con respecto de las componentes principales seleccionadas, se puede conocer cuanto correlacionan entre sí y en consecuencia cuán representadas están cada una de las variables en las componentes seleccionadas.

Los resultados de la primera implementación del PCA con datos de las dos muestras del período 2017 arroja las siguientes cargas factoriales:

Tabla 5: Cargas factoriales análisis 2017 muestra 1

	PC1	PC2
tender/value/amount	0.774486	-0.056668
tender/tenderPeriod/durationInDays	0.394208	0.889276
tender/items/0/quantity	0.703045	-0.436204

Fuente: Elaboración propia

Tabla 5: Cargas factoriales análisis 2017 muestra 2

	PC1	PC2
tender/value/amount	0.740765	-0.279732
tender/tenderPeriod/durationInDays	0.031333	0.942218
tender/items/0/quantity	0.750855	0.236654

Fuente: Elaboración propia

En ambas muestras, la primera componente combina directamente la primera y tercera variable, el monto y la cantidad de ítems que se espera contratar. La segunda componente principal está altamente correlacionada con la duración del

período de convocatoria en ambas muestras de forma directa. Las etiquetas de las componentes principales se definen en consecuencia de su relación con las variables, "precio y cantidad" para la componente 1, "duración del proceso" para la componente 2.

Los resultados de la implementación del PCA con datos de las dos muestras del período 2018 arroja las siguientes cargas factoriales:

Tabla 7: Cargas factoriales análisis 2018 muestra 1

	PC1	PC2
tender/value/amount	0.781667	0.030778
tender/tenderPeriod/durationInDays	0.587838	-0.729734
tender/items/0/quantity	0.643007	0.629709

Fuente: Elaboración propia

Tabla 7: Cargas factoriales análisis 2018 muestra 2

	PC1	PC2
tender/value/amount	0.706563	-0.087267
tender/tenderPeriod/durationInDays	0.536862	0.784688
tender/items/0/quantity	0.620857	-0.579214

Fuente: Elaboración propia

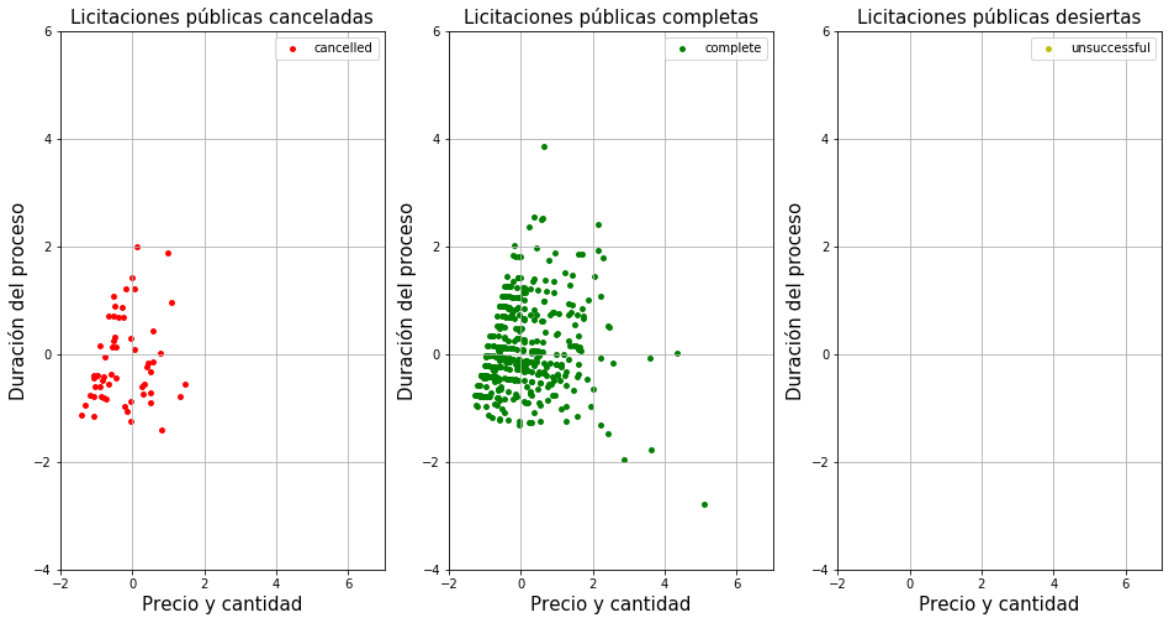
Al igual que en 2017, la primera componente principal de ambas muestras combina directamente a la primera y tercera variable, el monto y la cantidad de ítems que se espera contratar. Por su parte, la segunda componente principal está altamente correlacionada con la duración del período de convocatoria en ambas muestras, aunque de manera indirecta o negativa en la primera de ellas y de manera directa en la segunda. Las etiquetas de las componentes principales se definen en consecuencia de su relación con las variables, "precio y cantidad" para la componente 1, "duración del proceso" para la componente 2.

Para todas las muestras de los dos períodos analizados las etiquetas de las componentes principales son las mismas. La interpretación de la componente principal 1 es de igual manera en las dos muestras de ambos períodos y la

interpretación de la componente principal 2 varía únicamente en la muestra 1 del período 2018.

Para interpretar el resultado del algoritmo y representar el conjunto de datos en función de las componentes principales seleccionadas se utiliza el recurso gráfico. A continuación se visualizan los puntajes de cada una de las observaciones en función de las nuevas coordenadas, identificando con color verde a aquellos procesos de compra que logran completarse, con color rojo a aquellos procesos de compras que resultan cancelados y con color amarillo aquellos procesos que resultan desiertos.

Gráfico 5: Contrataciones 2017 muestra 1 en función de PC1 y PC2

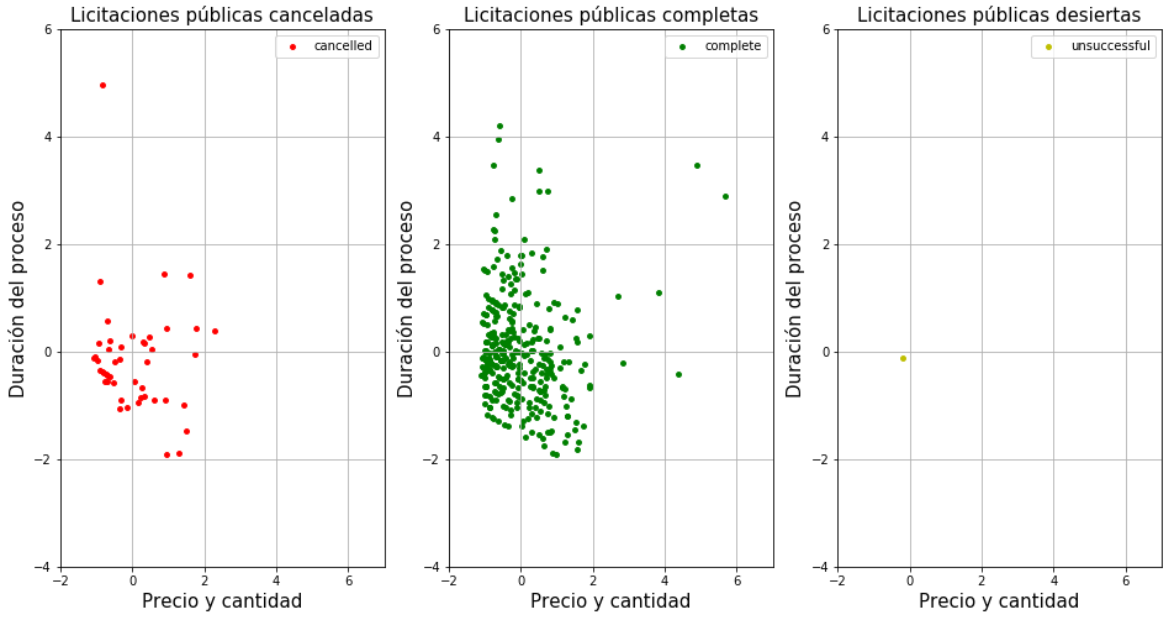


Fuente: Elaboración propia

El análisis gráfico de la primera muestra del 2017 muestra una concentración de puntos en el cuadrante que va del -2 al 2 en ambos ejes, para cada una de las categorías graficadas. Este comportamiento complejiza el análisis y la interpretación de los resultados. Sin embargo, se puede destacar una leve predominancia de procesos completos a la derecha del valor 2 del eje "precio y cantidad", evidenciando que aquellos procesos con altos montos, que buscan

adquirir cantidades relativamente grandes de ítems resultan completos exitosamente. No hay procesos de compras que resultan desiertos en la muestra 1 del período 2017.

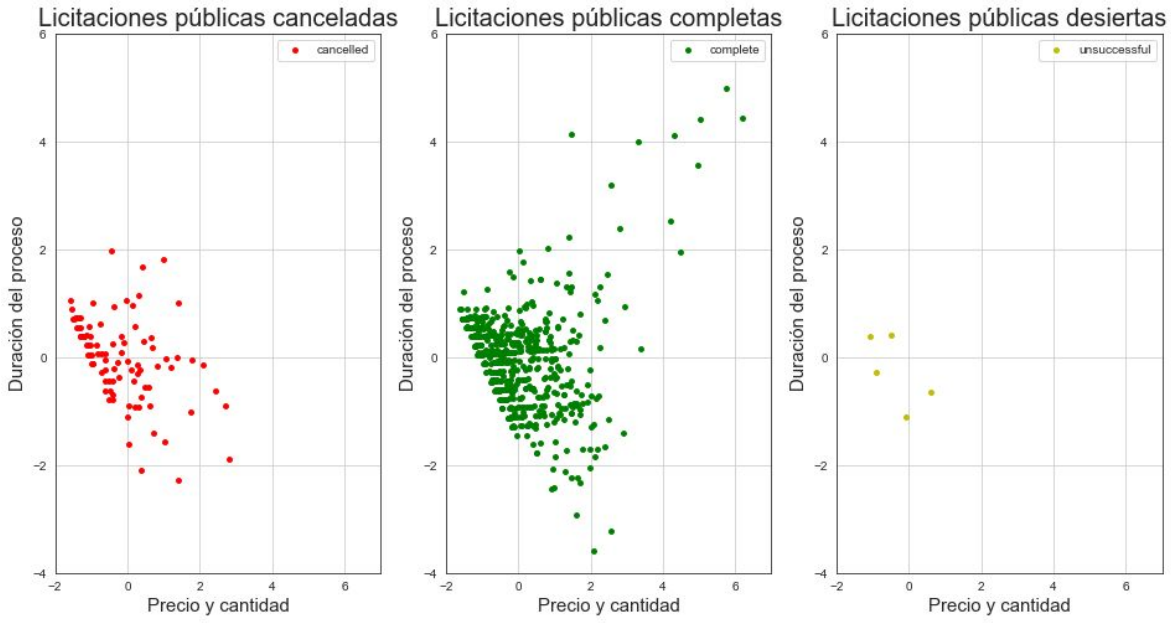
Gráfico 4: Contrataciones 2017 muestra 2 en función de PC1 y PC2



Fuente: Elaboración propia

El análisis gráfico de la segunda muestra de 2017 muestra una predominancia de los procesos completos en los valores positivos del eje "Duración del proceso", sobre todo aquellos que superan el valor 1. La mayoría de los procesos cancelados no alcanzan a superar el valor 2 del mismo eje. Lo mismo ocurre con el eje "precio y cantidad", los procesos cancelados no superan el valor 2 del mismo. Del análisis gráfico de la muestra 2 del 2017 se puede concluir que los procesos que se destacan en cuanto a la duración del proceso de convocatoria (tienen tiempos de licitación más amplios) pueden identificarse fácilmente como completos. Por otra parte, no hay un patrón claro que cumplan los procesos cancelados y no hay una cantidad suficiente de procesos desiertos como para extraer una conclusión al respecto.

Gráfico 4: Contrataciones 2018 muestra 1 en función de PC1 y PC2

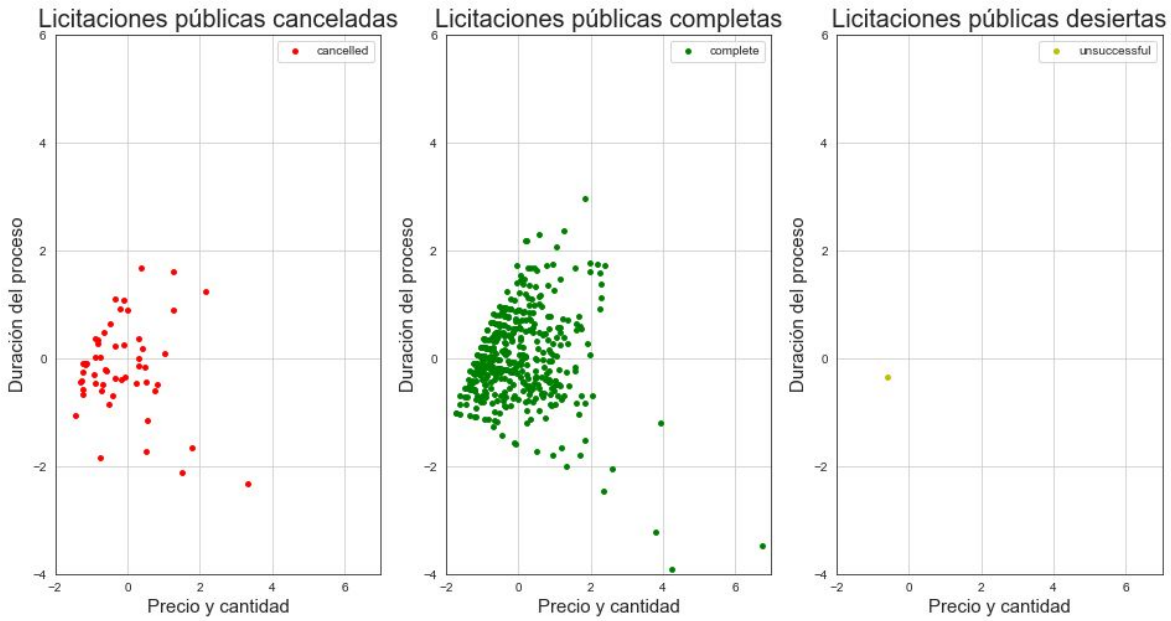


Fuente: Elaboración propia

El la muestra 1 del período 2018 se aprecia una clara predominancia de los procesos completos en valores altos de ambos componentes principales. Aquellos procesos con montos y cantidades de ítems relativamente altas, pero que a su vez revisten de tiempos de licitación relativamente cortos, son en todos los casos completos. Esto es una diferencia clara respecto de los datos del 2017, en 2018 parecieran haberse acortado los períodos de licitación y aún asi la mayoría de ellos resultaron completos.

Una hipótesis sobre la justificación de este comportamiento es la realización de los Juegos Olímpicos de la Juventud en Buenos Aires durante 2018, donde la urgencia por ajustarse a los cronogramas de la organización probablemente haya afectado a la duración de los procesos de compras, los cuales mostraron montos y cantidades elevadas por la magnitud que significa el evento. Probar esa hipótesis no es parte de este trabajo, queda planteada para futuras investigaciones entorno a esta problemática.

Gráfico 5: Contrataciones 2018 muestra 2 en función de PC1 y PC2



Fuente: Elaboración propia

En la segunda muestra del período 2018 no está tan claro el comportamiento de los procesos. Se destaca que la amplia mayoría de los procesos que superan el valor 2 del eje "Precio y cantidad", aquellos procesos con montos o cantidades altas, terminan por completarse exitosamente. En esta muestra, la cual contiene a las observaciones con valores de la variable "monto del proceso" que están por encima de la mediana de la distribución original, se muestran algunos registros por debajo del valor -2 del eje "Duración del proceso", contradiciendo al comportamiento de la muestra 1, de lo cual se puede extraer que el patrón identificado para la primera de las muestras se cumple solo en el caso de procesos de compras con montos totales relativamente bajos.

Conclusión

El análisis de componentes principales es una herramienta útil para extraer información de conjuntos de datos multivariados, representar gráficamente las observaciones e interpretar los patrones y el comportamiento de los registros analizados. Su implementación en el lenguaje de programación Python, a través de la biblioteca Scikit Learn, hace eficiente y performante su implementación utilizando grandes cantidades de datos.

La variabilidad presente en los datos de compras y contrataciones de Buenos Aires se debe a una serie de factores incorporados en el análisis. El sistema electrónico de contratación utilizado para gestionar las compras de bienes y servicios del Gobierno de la Ciudad, el mismo que genera los datos que nutrieron esta investigación, se ha estado utilizando desde el año 2011 para la contratación de todo tipo de bienes y servicios, que varían en montos y cantidades en amplio rango, evidenciado en el análisis exploratorio de los datos. El hecho de que el mismo sistema se utilice desde el año 2011, hace que los datos de precios y valores se encuentren distorsionados por el efecto de la inflación y por otra parte, la variedad de bienes y servicios que se contratan contribuye con la amplitud de los rangos de precios que se registran en los datos.

Para menguar los efectos negativos de estas cuestiones detectadas en el análisis exploratorio de los datos, se realizó el análisis en dos muestras para cada uno de los períodos analizados. Los resultados indican que existe cierta dificultad para identificar procesos de compras cancelados respecto de los completos con la técnica utilizada. Sin embargo, fue posible identificar características puntuales de aquellos procesos que se completaron exitosamente. Se pudo apreciar que aquellos procesos con precios y cantidades relativamente altas son en su mayoría procesos que se completan exitosamente, lo cual se ve reflejado especialmente en la muestra 1 de 2017 y la muestra 2 de 2018.

Una particularidad extraída del análisis es el comportamiento de los datos de la muestra 1 del 2018. Allí los procesos que alcanzan a completarse y se diferencian del resto de los procesos completados y cancelados son aquellos que evidencian una duración del proceso licitatorio relativamente más corta. Como se indicó, en el período 2018 aquellos procesos con montos relativamente bajos (muestra 1) evidencian una duración del proceso licitatorio relativamente corta y aún así lograron completarse exitosamente. La realización de los Juegos Olímpicos de la Juventud es la justificación de este comportamiento que se plantea como hipótesis para testear en futuros trabajos relacionados con la temática.

El comportamiento de los procesos de compras y el condicionamiento de las características de los pliegos de bases y condiciones varían con el tiempo y según el "tamaño" del proceso de compras en cuestión. El período inflacionario, el presupuesto anual, las prioridades del gobierno, las políticas públicas impulsadas y los eventos extraordinarios son quienes configuran la política de compras y determinan su comportamiento en consecuencia.

El hecho de que la fuente de los datos sea un set publicado en formato abierto en el portal de datos abiertos del Gobierno de la Ciudad es garantía de replicabilidad y actualización de este análisis. Ésta característica de los datos utilizados es condición necesaria para que el análisis pueda ser implementado en organizaciones del sector público y el sector privado. En el primero de los casos, la relevancia radica en el aporte de certidumbre para la planificación de ofertas, donde se destinan recursos y tiempos valiosos; de esta manera contribuye con el aumento de la competitividad de las empresas y la diversificación del mercado, la reducción de precios y aumento de calidad de los bienes disponibles. En el segundo, la relevancia radica en el aporte de información relevante para la gestión eficiente de los recursos públicos y para la optimización del proceso de compras, lo cual redunda en una mejor distribución de los bienes y servicios que provee el Estado.

Referencias bibliográficas

- Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".
- Heijnen, P., Haan, M. A., & Soetevent, A. R. (2014). Screening for collusion: a spatial statistics approach. Journal of Economic Geography, 15(2), 417-448.
- Huber, M., & Imhof, D. (2018). Machine learning with screens for detecting bid-rigging cartels. Université de Fribourg.
- Jake VanderPlas (2017) Python Data Science Handbook
- Lozares Colina, C., & López-Roldán, P. (1991). El análisis de componentes principales: aplicación al análisis de datos secundarios. Papers: revista de sociologia, (37), 031-63.
- M. Pal (2005) Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26:1, 217-222
- McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing, 14.
- Medina Merino (2017) Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python
- Nieto Mora, L. A. (2016). Aplicación de Data Mining en la Gestión del Plan Anual de Contratación en las Universidades públicas del Ecuador. Caso de Estudio Universidad Técnica de Ambato (Master's thesis, Universidad Técnica de Ambato. Facultad de Ingeniería en Sistemas, Electrónica e Industrial. Dirección de Posgrado. Maestría en Gestión de Bases de Datos).
- Oszlak, O. (2012). Gobierno abierto: promesas, supuestos, desafíos.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.
- Peña, D. (2013). Análisis de datos multivariantes. McGraw-Hill España.
- Pérez, P. M., & Costa, J. L. C. (2008). Capacidad predictiva de las variables cognitivo-motivacionales sobre el rendimiento académico. REME, 11(28), 1-13.
- Pinto Miranda, C. A. (2016). Recomendaciones para la mejora de gestión del proceso de compras públicas a partir del análisis de reclamos recibidos por Chilecompra.
- Rossum, G. (1995). Python reference manual.
- Smith, L. I. (2002). A tutorial on principal components analysis.
- Sun, T., & Sales, L. J. (2018). Predicting public procurement irregularity: An application of neural networks. Journal of Emerging Technologies in Accounting, 15(1), 141-154.
- VanderPlas, J. (2016). Python data science handbook: essential tools for working with data. " O'Reilly Media, Inc.".
- Volosín, N. (2015). Datos abiertos, corrupción y compras públicas.

Páginas web:

- http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_comp_reducido.pdf
- https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government
- https://opendatacharter.net/principles-es/
- https://scikit-learn.org/stable/modules/decomposition.html#decompositions