**Estudio de los Contactos Realizados al Sistema Único de Atención Ciudadana (SUACI)**Budziñski, Maia Ludmila1; Buttafuoco, Fernando2; Rodriguez, Joaquin3

1Facultad de Ciencias Naturales y Exactas, Universidad de Buenos Aires.

2

3

**Resumen**

Resumen del trabajo

**Palabras Clave**

Sistema Único de Atención Ciudadana, Cluster Analysis, Trámites, K-means.

# INTRODUCCIÓN

Sistema Único de Atención Ciudadana (SUACI) es el sistema oficial del Gobierno de la Ciudad a través del cual los ciudadanos tienen la posibilidad de efectuar solicitudes, reclamos, denuncias, quejas o reportes respecto a distintos servicios de la ciudad.  
El siguiente trabajo presenta fundamentalmente dos objetivos: la exploración de datos de la información proveniente de los contactos al SUACI desde el 2017 al 2019 y la inferencia de propiedades y estructuras de la distribución de los datos del año 2019, buscando encontrar grupos de muestras que posean alta relación entre sí mediante Cluster Analysis utilizando el algoritmo K-means.

# DESCRIPCIÓN DEL DATASET

En este trabajo, se utilizó el dataset ‘Información de los contactos realizados al Sistema Único de Atención Ciudadana (SUACI)’. Los datos incluidos en este estudio, provienen del reservorio de datos abiertos del GCBA (<https://data.buenosaires.gob.ar/dataset/sistema-unico-atencion-ciudadana>). Se utilizaron las bases con los registros correspondientes a los años completos 2017 y 2018, y parcial 2019 (Enero – Agosto). Estas se encuentran en formato *.csv*, correspondiendo un archivo para cada año. En ellas se encuentran todos los contactos generados en la Ciudad de Buenos Aires. En total, para los 3 años, se realizaron aproximadamente 2,2 millones de contactos. En cada uno de estos, se relevan los siguientes conceptos: contacto, fecha de ingreso, prestación, categoria, sub-categoria, tipo de contacto, comuna, barrio, calle, altura, esquina próxima, x e y

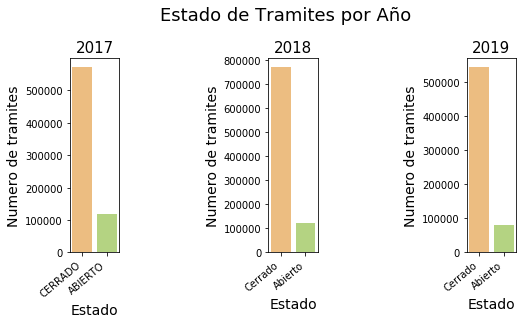
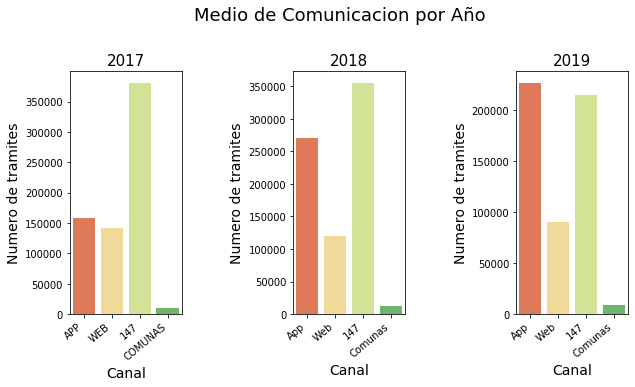
# ANÁLISIS EXPLORATORIO DE DATOS

Antes de comenzar el análisis, los datasets fueron limpiados de datos nulos utilizando el siguiente criterio:

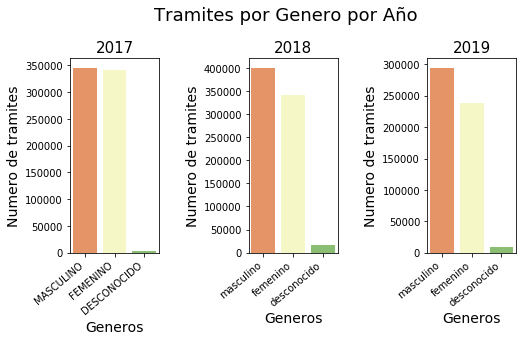
Eliminación de variable (feature): variable con más del 50% de registros nulos de datos nulos.

Eliminación de instancia (sample): siempre y cuando no implique eliminar más del 20% del total de instancias del dataset.

De un análisis preliminar de las variables: Canal, Género, Estado, Tipo De Prestación, Evolución Temporal Comunas, Categoría, Sub-Categoría, Concepto y Tiempo De Resolución., se derivaron las siguientes observaciones generales:

En cuanto al **CANAL** de contacto, se observa que los medios de comunicación utilizados por los usuarios para comunicarse con el SUACI son tres (vía Web, Aplicación móvil y telefónicamente al 147). Además, se evidencia un aumento año a año del uso de la aplicación móvil como medio de comunicación predominante 

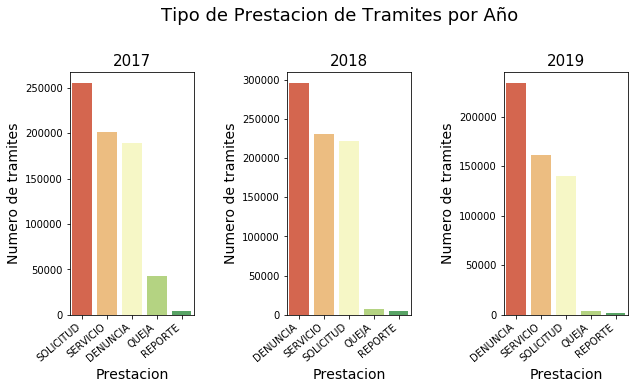
**Figura 1.** Countplot de los Medios de Comunicación utilizado por los usuarios para contactarse con el SUACI separados por Año (2017-2019).

En cuanto al **GÉNERO** de los contactos, se observa una clasificación en masculino, femenino o desconocido (la cual puede ser atribuida a que el usuario no se siente representado con ninguna de las opciones ofrecida o bien se trata de un dato faltante). Al respecto observamos que la cantidad de contactos de hombres, en los últimos años, ha superado la de las mujeres. 

**Figura 2.** Countplot de Trámites registrados por el SUACI, separados por Género y por Año (2017-2019).

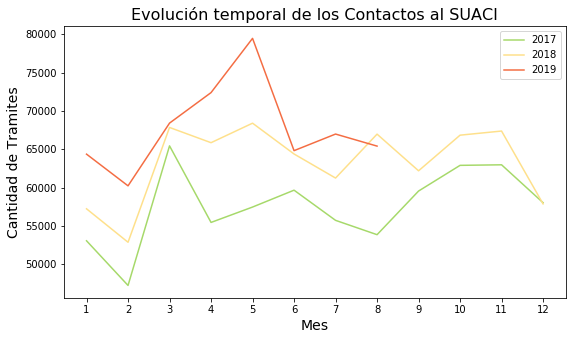
En cuanto al **ESTADO** de los trámites, El dataset no otorga gran información sobre la resolución de los contactos; las únicas medidas relacionadas a ello, son el estado (abierto o cerrado) y las fechas de inicio y fin del contacto. El que aún hayan contactos abiertos para 2017 y 2018 sugiere poca confiabilidad en este dato. De todas formas no hemos profundizado análisis en este tema (quedará para futuros análisis)

**Figura 3**. Countplot de los Trámites registrados por el SUACI, separados por el Estado de los mismos y por Año (2017-2019).

En cuanto al **TIPO DE PRESTACIÓN** de los trámites de los tres años analizados, vemos que predominan las solicitudes (pedido o reclamo al GCBA sobre una materia de su competencia y por    la cual debe realizar una acción específica), seguidos por servicios (solicitud o consulta en base a un servicio prestado en forma directa por el GCBA ) y denuncias (pedido de inspección  o corroboración de una infracción a la normativa vigente, realizada por un tercero) y, en última instancia, se encuentran las quejas (agradecimiento, sugerencia, felicitación o queja relacionado con algún servicio o proceso prestado por el GCBA).

**Figura 4**. Countplot de los Trámites registrados por el SUACI, separados por Tipo de Prestación de los mismos y por Año (2017-2019).

En cuanto a la **EVOLUCIÓN TEMPORAL** de los trámites, en primer lugar se evidencia un aumento interanual en todos los meses. Esto podría deberse al notable crecimiento del uso de la aplicación móvil. Sumado a esto, en febrero se observa una caída de los contactos y un posterior repunte en los meses siguientes.



.

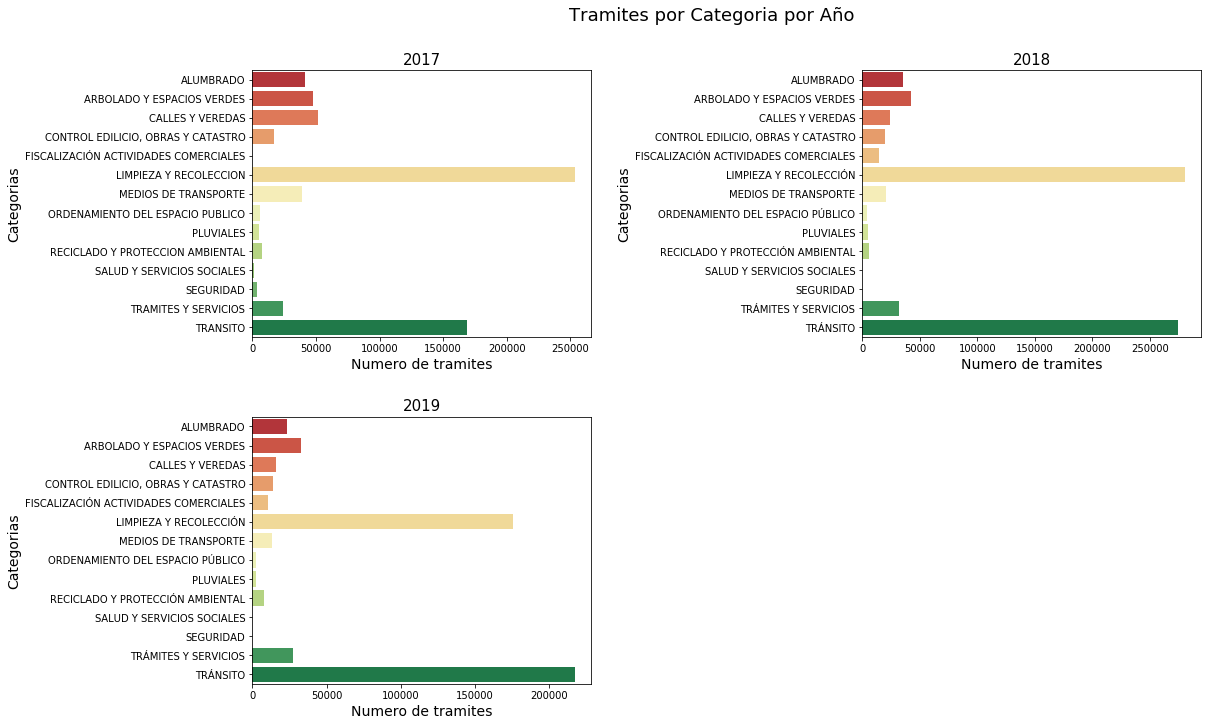
**Figura 5**. Cantidad de Trámites registrados por el SUACI en cada Mes, separados por Año (2017-2019).

Respecto a los trámites por **COMUNAS**, el análisis evidencia en primera instancia un comportamiento muy estable a lo largo de los 3 años. Adicionalmente, se observa que las 5 comunas con mayor número de contactos son las ubicadas en la zona norte de CABA.

. 

**Figura 6.** Countplot de los Trámites registrados por el SUACI, separados por Comuna y por Año (2017-2019) y mapa de Capital Federal de Buenos Aires, seccionado por Comunas como referencia.

Analizando en detalle los tramites por Categorías, Sub-Categorías y Conceptos, observamos que, sumando todos los contactos por categoría, existe una significativa concentración de contactos en las **CATEGORÍAS** “LIMPIEZA Y RECOLECCIÓN” y “TRANSITO”.



**Figura 7**. Countplot de los Trámites registrados por el SUACI, separados por Categoría y por Año (2017-2019)

Los contactos realizados al SUACI se catalogan en primer lugar en 213 conceptos, que es el nivel más bajo de categorización. Esos conceptos se agrupan en 58 subcategorías, las cuales a su vez pertenecen a 17 categorías diferentes, siendo éste último el orden superior. Al tomar las Categorías predominantes, observamos que la mayor parte de los contactos se concentran en las **SUB-CATEGORÍAS** “RESIDUOS VOLUMINOSOS” Y “DENUNCIA VIAL” en los tres años analizados.



**Figura 8**. Countplot de los Trámites registrados por el SUACI, separados por Año (2017-2019) y (a) la Sub-Categoria Limpieza y Recolección o (b) la Sub-Categoria Transito.

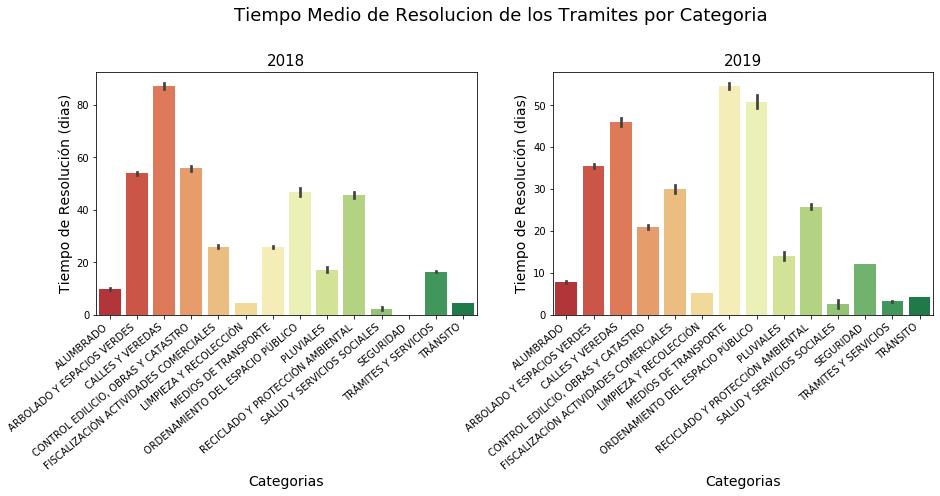
Finalmente, al tomar sólo las principales Sub-categorías predominantes, observamos que año a año los trámites se concentran en los **CONCEPTOS** “RETIRO DE ESCOMBROS” y “VEHICULO MAL ESTACIONADO”.

**Figura 9**. Countplot de los Trámites registrados por el SUACI, separados por Año (2017-2019) y (a) el Concepto Tramites de Denuncia Vial o (b) el Concepto Residuos Voluminosos.

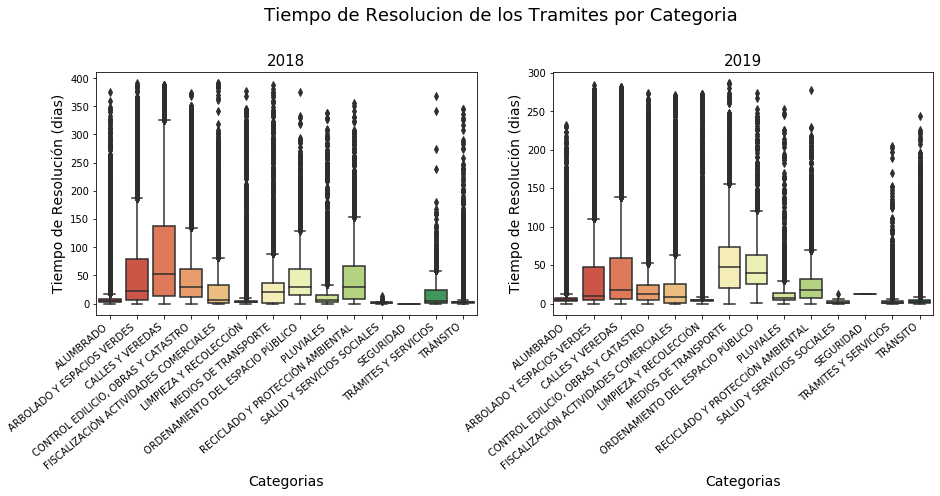
Sugerimos para futuros análisis, segregar estos Conceptos del resto y analizar ambos grupos de trámites (con y sin los datos de los trámites de Conceptos de mayor predominancia) por separado, para poder analizar otros

Conceptos a que quedan rezagados como consecuencia del gran peso que poseen los Conceptos principales, “RETIRO DE ESCOMBROS” y “VEHICULO MAL ESTACIONADO”.

Para identificar los **TIEMPO DE RESOLUCIÓN** de los contactos, creamos una nueva feature a partir de la diferencia entre entre la fecha de fin de contacto y la fecha de inicio. Esto solo pudo realizarse para los datos de 2018 y 2018 ya que el dataset del 2017 no contaba con datos sobre la fecha de fin de contacto. Posteriormente, Para profundizar un poco en el análisis, realizamos un boxplot con los datos.



**Figura10.** Análisis de la media del Tiempo de Resolución de los trámites registrados por el SUACI, separados por Año (2018-2019).



**Figura 11**. Boxplot del Tiempo de Resolución de los trámites registrados por el SUACI, separados por Año (2018-2019).

Dada la considerable presencia de outliers, no arribamos a ninguna conclusión. Queda para futuros análisis extraer dichos outliers y analizar por separado ambos grupos; sin embargo, queda claro que la media no resulta el parámetro estadístico idóneo para este caso.

# MATERIALES Y MÉTODOS

Nuestro objetivo en el presente informe es, a partir de los contactos realizados al SUACI, identificar zonas en la Ciudad Autónoma de Buenos Aires con similares patrones de comportamiento poblacional. Para ello realizamos un análisis de *clustering* de los datos.

## Pre-procesamiento de los datos

**Selección de variables**

Considerando el volúmen de datos que se tienen, tomamos únicamente el dataset con los contactos realizados en el año calendario 2019. Tomamos esta decisión a partir de la necesidad de reducir la cantidad de samples por el elevado costo computacional de los algoritmos involucrados. Esto redujo el dataset a aproximadamente 620mil muestras.

Utilizamos la variable "categoria" que aporta una categorización descriptiva del tipo de contacto, con un nivel de apertura adecuado. A efectos del problema propuesto no nos interesa el canal a través del cual se realizó el contacto ni el género del denunciante, así como la fecha en que se registró el mismo, en tanto tomamos el horizonte temporal de todos los contactos realizados durante el año calendario 2019.

Definimos entonces un nuevo dataset con las variables estrictamente importantes para nuestro ejercicio, comprendidas por:

* Las variables dummies obtenidas a partir de la variable categórica "categoría"
* El par latitud-longitud, para que el algoritmo le asigne también un peso a la proximidad geográfica de los contactos a la hora de clusterizar

La dimensión final del dataset a *clusterizar* es (626101, 19)

**4.1.2 Estandarización**

Las coordenadas latitud y longitud presentan problemas de continuidad (por ejemplo, en el caso del antemeridiano de Greenwich, donde la longitud toma valores de +180 y -180 a ambos lados, representando un salto gigante en la distancia para dos puntos próximos entre sí). Restringiendo el dominio a la Ciudad de Buenos Aires, no tenemos este problema, y por otro lado, al expresarse los valores en decimales, la diferencia máxima de los valores en este dominio es inferior a la unidad. Por tanto, no hay diferencias en el orden de magnitud de la diferencia entre las variables y, en principio, no sería necesario escalar nuestras variables

De todos modos, decidimos realizar un escalado de nuestras variables y guardarlo en un dataset separado del original a fines comparativos.

**4.1.3 Principal Component Analysis (PCA)**

Realizamos un PCA a partir del dataset original para reducir la dimensionalidad de la matriz y poder trabajar con un número reducido de componentes principales que representaran la mayor varabilidad posibles

## Modelo de Aprendizaje

Utilizamos K-means para realizar un Cluster Analysis de los datos originales, estandarizados y las dos componentes principales que obtuvimos de realizar el PCA sobre los datos originales. K-means es un algoritmo de clasificación no supervisada que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.

El algoritmo consta de tres pasos:

Inicialización: una vez escogido el número de grupos, *k*, se establecen *k* centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.

Asignación objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano.

Actualización centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso. El algoritmo *k-means* resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster. Los objetos se representan con vectores reales de *d* dimensiones (*x1,x2,…,xn*) y el algoritmo *k-means* construye *k*grupos donde se minimiza la suma de distancias de los objetos, dentro de cada grupo *S*={*S1,S2,…,Sk*} a su centroide. El problema se puede formular de la siguiente forma:

Se obtienen *k*grupos o clusters con su correspondiente centroide *μi*.

En cada actualización de los centroides, desde el punto de vista matemático, imponemos la condición necesaria de extremo a la función E(*μi*) que, para la función cuadrática (1) es:

y se toma el promedio de los elementos de cada grupo como nuevo centroide.

Para medir la calidad de los clusters se computó el Silhouette Index (*S*) utilizando la distancia media intra-cluster (*a*) y la distancia media inter-cluster cercano (*b*) para cada muestra.

Dónde,

*a(i)* es la diferencia promedio de un objeto con respecto a todos los demás objetos en el mismo grupo

*b(i)* es la diferencia promedio del objeto con todos los objetos en el grupo más cercano.

# RESULTADOS

# DISCUSIÓN Y CONCLUSIONES

Las principales ventajas del método *k-means* son que es un método sencillo y rápido. Pero es necesario decidir el valor de ***k*** y el resultado final depende de la inicialización de los centroides. En principio no converge al mínimo global sino a un mínimo local.

# REFERENCIAS

The Elements of. Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. February 2009. Trevor Hastie · Robert Tibshirani.Jain,

Jain, A.K. (2010) Data Clustering: 50 Years beyond K-Means. Pattern Recognition Letters, 31, 651-666. <http://dx.doi.org/10.1016/j.patrec.2009.09.011>

Chaturvedi, Nikhil & Rajavat, Anand. (2013). An Improvement in K-mean Clustering Algorithm Using Better Time and Accuracy. International Journal of Programming Languages and Applications. 3. 13-19. 10.5121/ijpla.2013.3402.

Shinde, Sachin V. and Bharat A. Tidke. “Improved K-means Algorithm for Searching Research Papers.” (2014).