

**APLICACIÓN DE MODELOS DE CLUSTERING AL**

**ANALISIS DE PRECIOS DE LA OFERTA OFRECIDA POR AIRBNB**

**EN LOS DISTINTOS BARRIOS DE ESTADOS UNIDOS**

**UNIVERSIDAD TECNOLOGICA NACIONAL,** Buenos Aires

**Ingeniería Industrial**

**CATEDRA DE CIENCIA DE DATOS – 2023 – CURSO I5521**

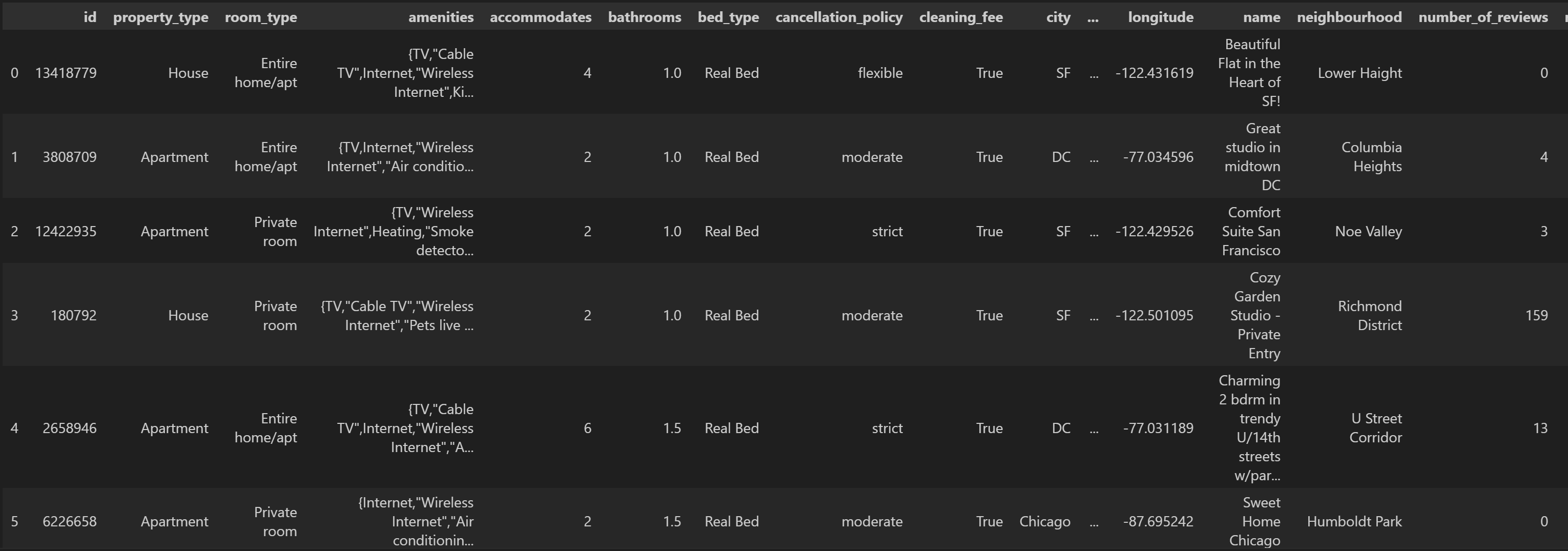
**ALUMNOS**

**PICCO, Lucas SALAS, Christian**

**ABSTRACT:** Nuestra intención es realizar un análisis exhaustivo de los datos obtenidos a través de la oferta ofrecida en Airbnb. Airbnb ha emergido como una plataforma popular para alquileres de propiedades en todo el mundo. El análisis de datos de los precios de Airbnb puede proporcionar valiosa información sobre los factores que influyen en las tarifas de alquiler. En este trabajo, se lleva a cabo un análisis detallado de los precios de Airbnb, centrándose en varias variables como el vecindario, las comodidades, la ubicación geográfica, entre otros.

**OBJETIVOS:** Los objetivos de este estudio son realizar un análisis exploratorio de datos para comprender la distribución de precios en Airbnb, identificar patrones y tendencias a través de diversas variables, y aplicar técnicas de clustering para segmentar las propiedades en grupos similares en función de sus características. De esta manera podremos lograr mediante diferentes técnicas de predicción y entrenamiento de modelos llegar a un estimativo de precio con respecto a las propiedades y sus rentabilidades. Entender los datos y poder segmentarlos es el elemento clave para poder realizar un análisis valido. Una vez la información es segmentada de manera inteligente podemos determinar las diferencias entre el porque de esos datos. Al saber el porqué de esos datos podemos cambiar intervenir en ellos de la manera que nos sea mas provechosa para nuestro negocio, emprendimiento o mismo necesidades sociales.

**Descripción del Dataset :** El conjunto de datos utilizado en este análisis contiene información sobre propiedades de Airbnb, recopilada en forma de dataset. Consta de 19309 entradas y 29 variables, incluyendo vecindario, precio, comodidades, latitud, longitud, etc.



Ej. Fig. 1.

Luego se realizó una búsqueda de los valores dentro del dataset, esto nos permite detectar que variables son potables para su análisis, así como cuales descartar para tener un mejor aprovechamiento de la información y la analítica, en la siguiente figura se observan los resultados obtenidos.



Al realizar el análisis de la cantidad de valores nules se puede

Observar que los mismos se acumulan en al menos 15 variables.

A la hora de decidir que variables eliminar se tomo en cuenta

La correlación entre las mismas y su utilidad con respecto al precio

Se entiende que al tratarse de alquileres de propiedades las variables

Mas influyentes siempre serán, el barrio, las acomodaciones, las

Valoraciones de los anteriores huéspedes y la cantidad de

Habitaciones. Es por esto por lo que se decidieron eliminar las siguientes

Variables: *columnas\_eliminar = ['amenities','review\_scores\_rating','description','first\_review','zipcode',*

*'thumbnail\_url','cleaning\_fee','thumbnail\_url','cleaning\_fee','*

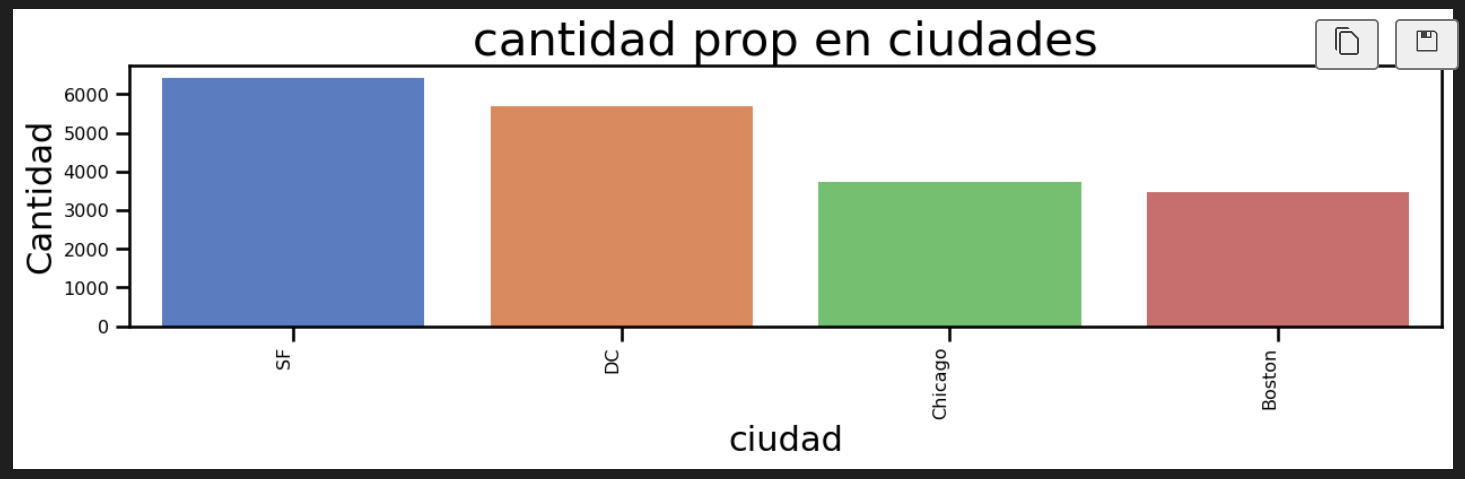
*host\_has\_profile\_pic','host\_identity\_verified','host\_response\_rate','*

*last\_review','host\_since']*

Ej. Fig. 2.

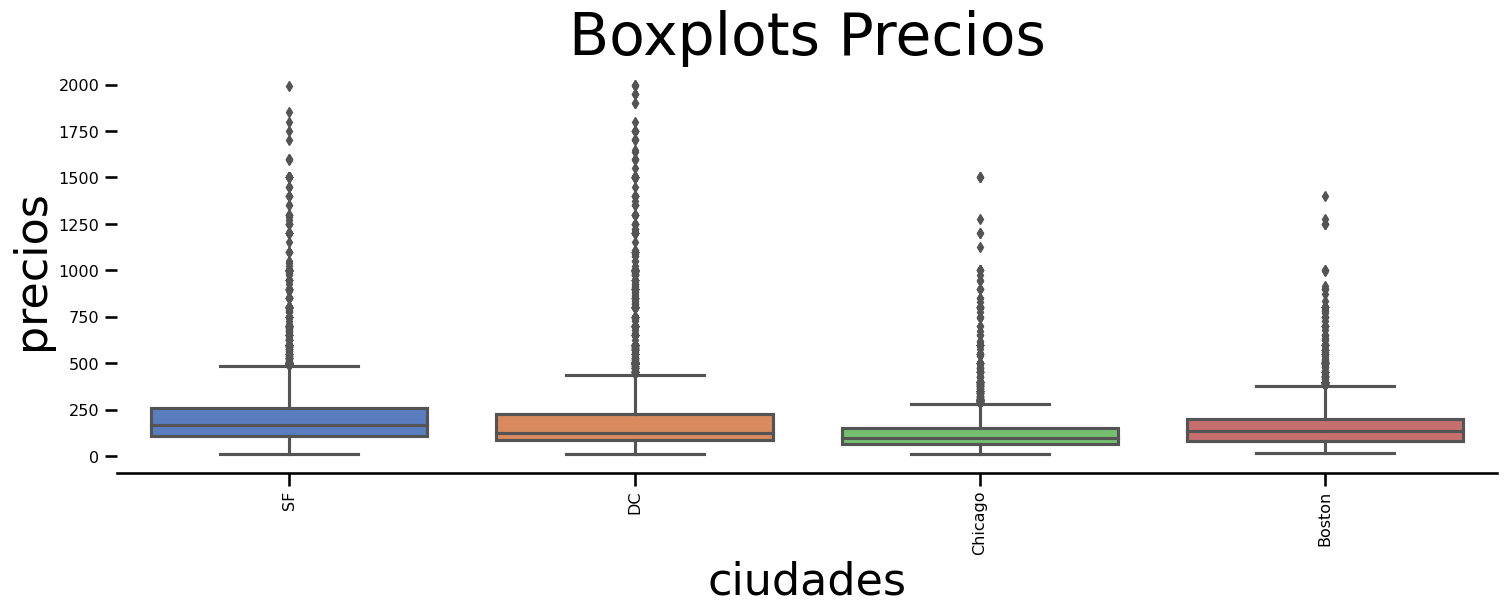
**ANALISIS EXPLORATORIO DE DATOS:** Se llevó a cabo un análisis exploratorio que reveló la distribución de los precios, relaciones entre variables, y patrones geoespaciales. Se observaron variaciones significativas en los precios entre distintos vecindarios y se identificaron correlaciones entre ciertas comodidades y tarifas de alquiler.

Se comenzó por un análisis del agrupamiento de propiedades con respecto a las distintas ciudades observándose que las aglomeraciones mas fuertes se dan en: SF, DC, Chicago y Boston.



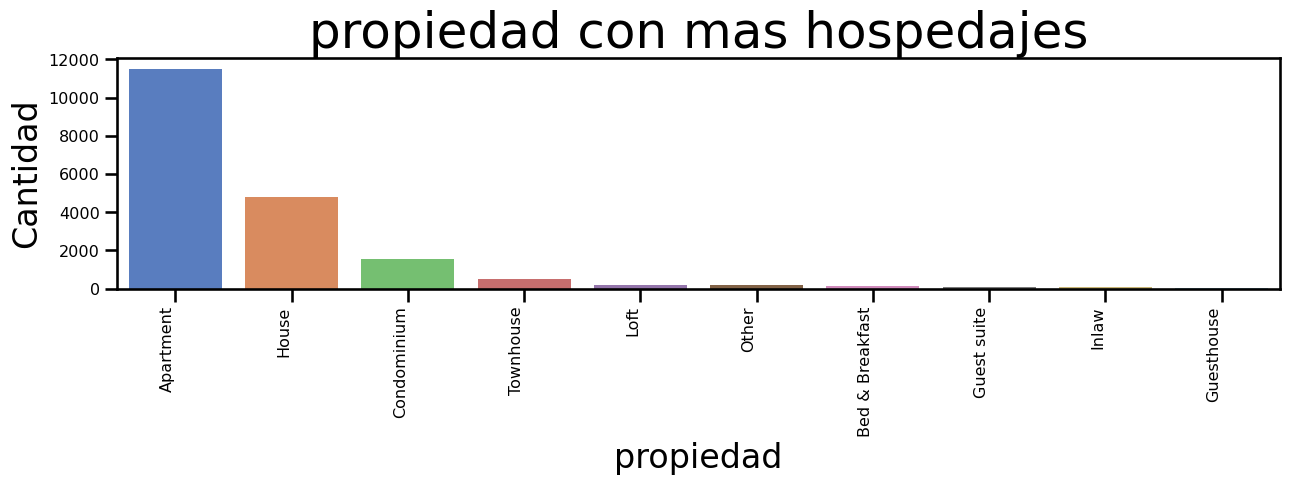
Ej. Fig. 3.

Continuamos con un análisis de los precios por medio de un Boxplot para lograr una segmentación de los precios y lograr percatarnos de los famosos datos “outlayers”, en nuestro caso lo precios se concentran entre 50 USD Y 250 USD. Los precios apreciados fuera del Boxplot pertenecen a otro tipo de propiedades. En este caso los precios mas altos se encuentran en SF y DC.

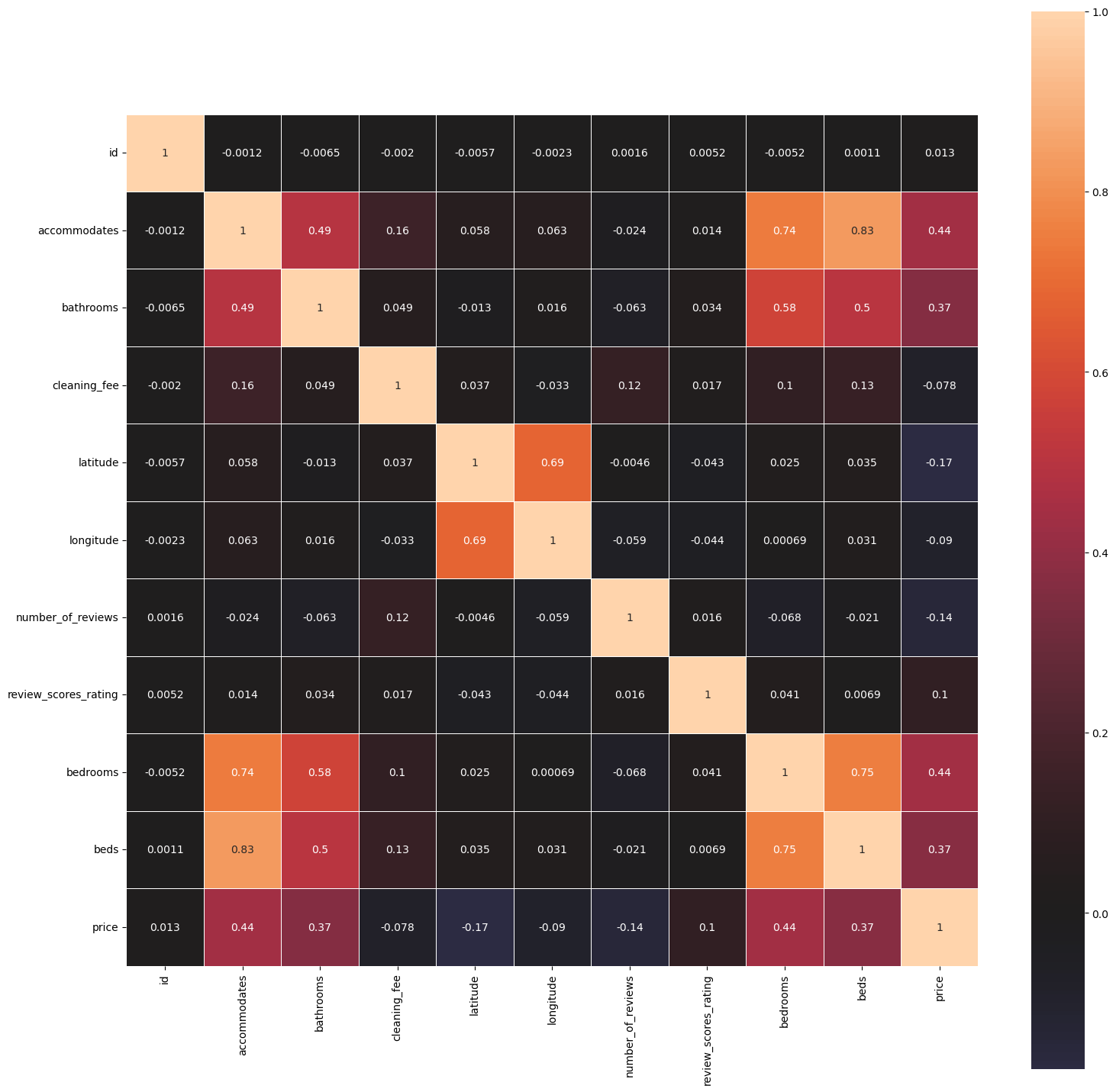


Ej. Fig. 4.

También se realizó un análisis de el tipo de propiedad con la mayor cantidad de hospedajes



**MATRIZ DE CORRELACIÓN**



**MACHINE LEARNING**

Una vez que se cargó data set, limpio la base datos eliminando registros null, borrando registros repetidos y visualizando resultado a través de diferentes gráficos de barra y boxplot.

Se procede a través modelo aprendizaje supervisado realizar varios modelos de regresión:

* Modelo de regresión lineal
* KNN regressor
* Super Vector Regressor
* Random Forest Regressor

Iniciamos primero con llamando a las librearías para trabajar con los diferentes modelos. Por ejemplo:

from sklearn import preprocessing

from sklearn import linear\_model

Primero eliminamos las variables que son string o numéricas y que no forman parte del modelo como: reserva, política cancelación, tipo cama, número de visitas.

Luego procedemos a definir la variable dependientes e independientes X (variables para modelo) Y (Precio a predecir).

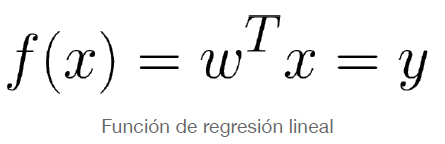
X esta formado por: Ciudad, barrio, tipo de propiedad, nro de baños, número de camas.

Define las variables de enteramiento y prueba: Xtrain, Xtest, Ytrain e Ytest tomando un 20% de los datos para el testeo. Se entrena las variables a través de una función.

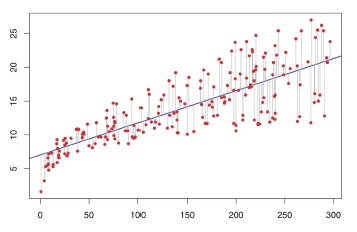
Escalamos las variables de entrenamiento y prueba para que no halla disparada a la hora de combinación lineal y posterior modelo de regresión.

**Modelo de regresión lineal**

El modelo consiste en establecer una relación entre las variables dependientes (ciudad, tipo propiedad, barrio, nro. de baños y camas) una variable independiente Y (precio) a predecir. El modelo consiste en formar una línea recta que muestra la relación entre variables.



Primero se entrena las variables para el modelo con: **model.fit(xtrain,ytrain)** luego se realiza la predicción: **predictions=model.predict(xtest).** La visualización del modelo queda representada de esta forma:



Los puntos rojos son valores reales etiqueta. La línea azul es la función aprendida que minimiza los residuos o distancia entre Ypred y Yreal.

El acuerry es una métrica para evaluar el modelo en nuestro caso. MSE error cuadrático medio nos da mse\_rl de 0.29. Nos indica que el modelo tiene un error de predecir de 0.29 dólares por cada dólar de hospedaje. Es un error aceptable para predecir el precio del hospedaje.

**Modelo KNN Regressor**

En este modelo de regresión pasa por diferentes etapas de: almacenamiento de variables entrenamiento, selección de valores K, cálculos de distancias, selección K más cercanos, promedio de los valores más cercanos.

Llamamos a la biblioteca: **scikit-learn.**

Primero a través de la función **KNeighborsRegressor ()** buscamos los vecinos k más cercanos tendrán un peso en la predicción. Se define los nro. de vecinos para la búsqueda.

Aplicamos la **funcion GridSearchCV()** evalúa el rendimiento del modelo para cada combinación de parámetros. Así obtenemos los mejores parámetros y R2 que explica la varianza de Y del modelo.

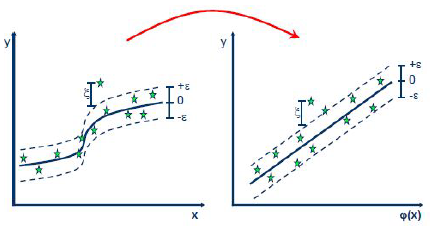
El acuerry para este modelo KNN regresor nos da un MSE:

Mse\_knn = 0.2912

**Support Vector Regressor**

Es un modelo de regresión basado en soporte vectorial para predecir valores numéricos. Encontrar la mejor función que se ajuste a los datos. Minimice la diferencia entre valor real del Precio de hospedajes y la de predicción con las variables de modelo.

* y : predicción precio
* X : vector de entrada
* W: vector de peso
  + B: termino de sesgo



* primero se define los parámetros para modelo.
* **GridSearchCV()** a través de combinaciones buscamos los mejores parámetros
* **regressor\_svr.predict()** predecimos con las variables de testeo.
* Calificamos el modelo con Acuerry de valor:

**Mse\_svg = 0.289**

**Randon Forest Regressor**

Modelo de regresión que consiste en algoritmo de random forest. Combina varios árboles de decisión para una predicción más precisa.

Se basa en 2 partes:

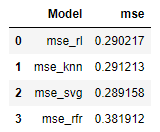
* construcción de subconjuntos de muestras aleatorias de entrenamiento.
* Votación predicción. Cada subconjunto realiza una predicción basado en el conjunto de entrada. La predicción del randon forest es promedio de la predicción de cada subconjunto.

Primero se definieron parámetros para modelo. Luego la función **GridSearchCV()** con los parámetros de muestras aleatorias, nro cross validation, y el modelo RFR. Se obtuvieron los mejores parámetros con un R2 0.5 indica la varianza de las etiquetas.   
Al final definimos el rendimiento del modelo a través de Acuerry de:

Mse\_rfr = 0.3819.

**RESULTADO**

Luego de aplicar diferentes modelos de regresión. Podemos evaluar qué modelo nos indicó mejor rendimiento de acuerry para predecir el precio de hospedaje en diferentes propiedades.

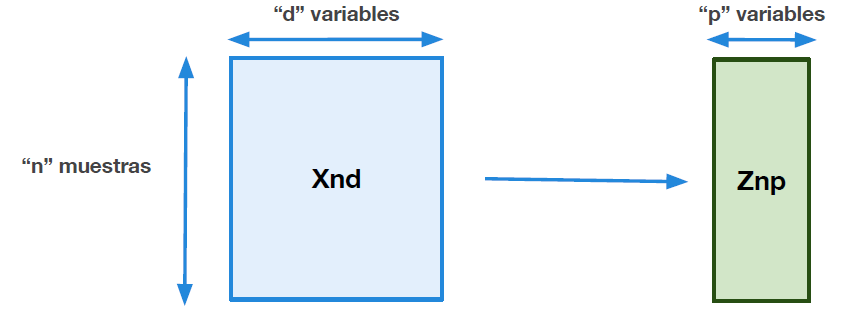


El modelo con menor Error Cuadrático medio (MSE) es el de regresión lineal con un rendimiento de mse\_rl de 0.2902 lo cual nos indica que 0.29 dólares por cada dólar que se predice en los precios de diferentes hospedajes según diferentes variables dependientes.

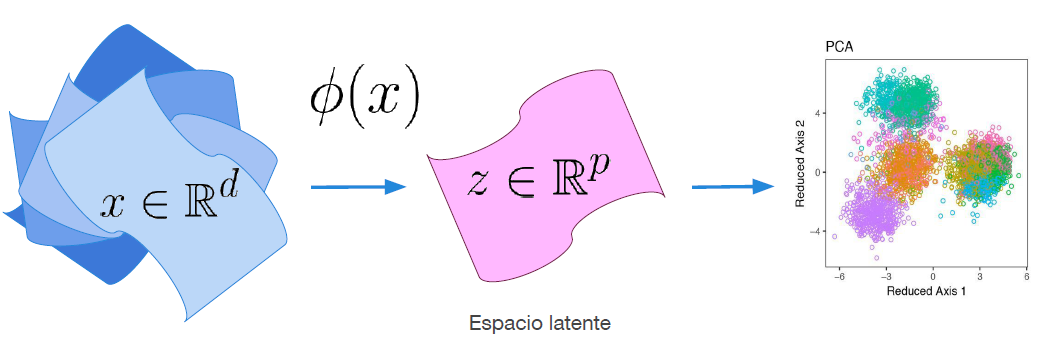
**METODO DE REDUCCION DIMENSIONES PCA**

El modelo consiste en reducción el número de variables (dimensiones) en un conjunto de datos que contenga las variables más relevantes de base datos para definir el precio del hospedaje. Para esto aplicamos el método más conocido Análisis Componentes Principales PCA.

El método consiste en: identificar los componentes principales que represente los datos. Procesar calculando los vectores y valores propios asociados. Al final proyectar los datos originales en un nuevo espacio de menor dimensión definidos por los vectores propios.



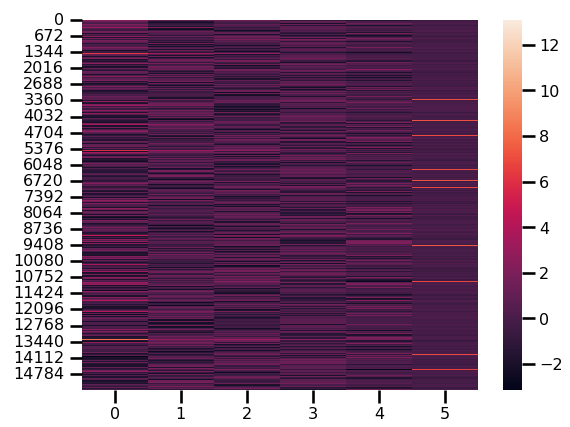
Trasformar la matriz Xnd de conjunto datos origen a una Matriz Znp de menor dimensión.



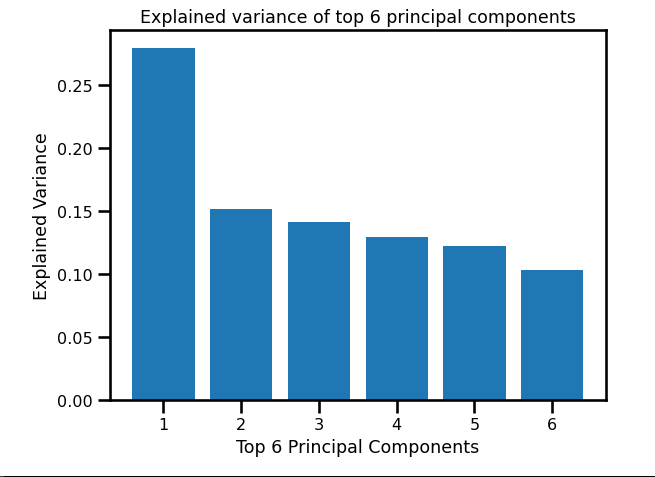
Aprender una función que proyecte los datos del conjunto X un uno Z de menor dimensión. Ejemplo de la figura una PCA de 2 dimensiones.

En el trabajo procedemos a:

* Llamar a la librería **sklearn.decomposition import PCA**
* **Definimos los compoentes extraer n = 6 y PCA.**
* **Funcion de transfomracion de Xtrain\_scal pca.fit\_transform(xtrain\_scal)**
* **Obtenemos los autovalores.**
* **Las varianzas de cada compoente.**
* **Observamos la matriz correlación del conjunto reducido PCA**



* Obsrevamos los autovalores obtenidos del PCA. Donde los autovectores indican como se proyectan los datos X sobre componetes principales Y. Autovectores indican nivel importancias cada componente PCA.



El grafico de barra observamos que los primeros 2 componentes de PCA visualizamos el 48% de la variabilidad de los datos de las 10 dimensiones que era originalmente el conjunto de datos X.

**CONCLUSION**

Luego de realizar carga de la data set de Precios de hospedajes en EEUU. Procedimos a una limpieza de los datos eliminando registros vacíos y duplicados. Ya Limpiados los datos realizamos Visualizaciones de Boxplots de como varían los precios en cada ciudad, grafico de barra de cantidad propiedades en cada ciudad otro grafico de barra indicando que tipo propiedad se alojan más los clientes. Las visualizaciones pudimos encontrar diferentes resultados e conclusiones de la base datos.

Luego buscamos predecir el precio de hospedaje utilizando diferentes modelos de regresión como: Lineal, randon forest, KNN, support vector regressor. Para elegir un modelo que mejor explique el conjunto de datos procedimos a evaluar el rendimiento de cada uno a través del Acuerry es Error Cuadrático Medio. El mse que menor valor dio fue el de Regresión Lineal con mse\_rl de 0.29 que nos indica un error o distancia a predecir de 0.29 dólares de cada dólar de precio del alojamiento.

Finalizar procedimos a realizar un Modelo de reducción de dimensiones PCA. Buscamos reducir el número de variables o dimensiones en un conjunto menor de componentes que mejor explique la variabilidad del conjunto de datos originales. Identificamos componentes principales, calculamos los auto vectores y auto valores que van definir el nuevo espacio de menor dimensión. En nuestro caso reducimos de conjunto de 10 variables a 6 dimensiones donde los 1er 2 componentes representan 48 % de la variabilidad del conjunto original de 10 dimensiones.

En lo personal el trabajo nos permitió a traves de Python procesar una base datos Real de Hospedajes en diferentes ciudades de EEEUU usando herramientas de Python. También aplicar conceptos de Machine Learning para predecir el precio de un alojamiento con respecto a diferentes variables como ciudad, barrio, tipo propiedad, nro. baños, ambientes y etc.

**REFERENCIAS**

● Apuntes Cluster IA. 2023 - Cátedra de Ciencia de Datos UTN - FRBA. Consultado el 16/11/2020

● Página Oficial Numpy.. Disponible en: https://pandas.pydata.org/ ● Página Oficial Pandas.. Disponible en: https://pandas.pydata.org/ ● Página Oficial Seaborn.. Disponible en: https://seaborn.pydata.org/ ● Página Oficial Sklearn Preprocessing MinMaxScaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html ● Hastie, Trevor.; Friedman, J. H. (Jerome H.) (2001). The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations

● Logistic Regression. Disponible en: http://www.fao.org/tempref/AG/Reserved/PPLPF/ftpOUT/Gianluca/stats/Logistic%20Regressi on,%20A%20SelfLearning%20Text,%202Ed%20(Statistics%20For%20Biology%20And%20Health)%20(David%20 G%20Kleinbaum,%20Mitchell%20Klein)%200387953973.pdf