Trabajo Práctico N°2

Profesora: Noelia Romero
Asistente: Victoria Oubiña

Reglas de formato y presentación

Fecha de entrega: domingo 21 de abril a las 23:59.
Contenidos: limpiar la base de datos provista y aplicar métodos de análisis descriptivo.

Modalidad de Entrega

- Al finalizar el trabajo práctico deben hacer un último commit en su repositorio de GitHub con el mensaje *Entrega final del tp*.
- Asegúrense de haber creado una carpeta llamada TP2. Deben entregar un reporte (pdf) y el código (jupyter notebook). Ambos deben estar dentro de esa carpeta.
- Deberán enviar el link a su repositorio -para que pueda ser clonado y corregido- al siguiente correo: v.oubina@gmail.com.
- La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
  - No envíen el correo hasta no haber terminado y estar seguros de que han hecho el commit y push a la versión final que quieren entregar. Debido a que se pueden tomar hasta 3 días de extensión a lo largo del curso, no corregiremos sus tareas hasta no tener el repositorio.
  - No hagan nuevos push después de haber entregado su versión final. Esto generaría confusión acerca de qué versión es la que quieren que se les corrija.

Reglas de formato y presentación

- El trabajo debe tener una extensión máxima de 8 páginas (no se permite Apéndice). Se espera una buena redacción en la resolución del práctico.
- El informe debe ser entregado en formato PDF, con los gráficos e imágenes en este mismo archivo.
- Entregar el código con los comandos utilizados, identificando claramente a qué inciso corresponde cada comando.

Parte I: Limpieza de la base

La base de datos provista contiene información sobre oferentes de Airbnb en la ciudad de Nueva York.

1. Realicen una limpieza de la base.

   (a) Tengan en cuenta si hay valores duplicados y eliminénlos.

   (b) Eliminen las columnas que no tienen información de interés.

   (c) Luego de leer el artículo Missing-data imputation[1], decidan qué hacer con los missing values e implementen su decisión. Justifiquen su elección. Pueden usar información del paper mencionado o de otras fuentes (cítenlas si las usan) para explicar los problemas que pueden surgir por su estrategia elegida.

   (d) Si hay observaciones con *outliers* o valores que no tienen sentido, tomen una decisión. Expliquen las decisiones tomadas.

   (e) Transformen las variables `'neighbourhood_group'` y `'room_type'` a variables numéricas.

   (f) Con la ayuda de los comandos `groupby` y `join` o `merge`, creen una columna que tenga la cantidad de oferentes por "Neighbourhood group". Llamen a esa nueva columna `offer_group`.

Parte II: Gráficos y visualizaciones

1. Una vez hecha la limpieza, realicen una matriz de correlación con las siguientes variables:
   `'neighbourhood_group'`, `'latitude'`, `'longitude'`, `'room_type'`, `'price'`, `'minimum_nights'`, `'number_of_reviews'`, `'reviews_per_month'`, `'calculated_host_listings_count'`, `'availability_365'`. Comenten los resultados. Utilicen alguno de los comandos disponibles en este link para graficar la matriz de correlación.
   Nota: consideren cómo es conveniente incluir las variables que originalmente eran categóricas para poder interpretar mejor la matriz de correlación.

2. Respondan las siguientes preguntas: ¿Cuál es la proporción de oferentes por "Neighbourhood group"? ¿Y por tipo de habitación? Además, realicen gráficos para mostrar estas composiciones y comenten los resultados.

3. Realicen un histograma de los precios de los alojamientos. Comenten el gráfico obtenido. Además, respondan las siguientes preguntas: ¿cuál es el precio mínimo, máximo y promedio? ¿Cuál es la media de precio por "Neighbourhood group" y por tipo de habitación?

[1] También pueden consultar el siguiente artículo para más información.

4. Realicen dos scatter plots con dos variables de interés en cada uno. Comenten.

5. Utilicen el análisis de componentes principales para graficar las variables en dos dimensiones. Comenten los resultados obtenidos (qué porcentaje de la varianza se logra explicar con dos componentes, cómo son los *loadings*, si ven algún patrón en el gráfico).

Parte III: Predicción

El objetivo de esta parte del trabajo es intentar predecir los precios de los alojamientos.

1. Eliminen de la base todas las variables relacionadas al precio.

2. Partan la base en una base de prueba (test) y una de entrenamiento (train) utilizando el comando `train_test_split`. La base de entrenamiento debe comprender el 70% de los datos, y la semilla a utilizar (*random state instance*) debe ser 201. Establezca a `price` como su variable dependiente en la base de entrenamiento (vector y). El resto de las variables serán las variables independientes (matriz X). Recuerden agregar la columna de unos (1).

3. Implementen una regresión lineal y comenten los resultados obtenidos.