Trabajo Final Data Wrangling

CRITERIOS DE EVALUACIÓN

Deberán presentar un jupyter notebook resolviendo las consignas propuestas para este trabajo práctico. En dicho documento deberán:

- Explicar los procedimientos y decisiones (usar las celdas de texto de Google Colab)
- Comentar el código
- Llegar a los resultados esperados

Se presentarán 10 consignas a resolver. Si se considera que existe algún análisis adicional a ser mencionado y realizado, podrá realizarse.

El trabajo a presentar debe tener carácter de informe, por lo que además de la correcta resolución de las tareas planteadas se tendrá en consideración la prolijidad y claridad explicativa a la hora de evaluar el trabajo.

IMPORTANTE: Cada consigna asume que se está trabajando partir del dataset obtenido en el punto anterior.

CONSIGNAS

El objetivo general es realizar un análisis exploratorio de los anuncios inmobiliarios de algunas provincias de la Patagonia Argentina reportados por el portal Airbnb.

Los datos están conformados por cuatro archivos csv:

- *Neuquen.csv*
- *Rio_Negro.csv*
- *Chubut.csv*
- *Tierra_del_Fuego.csv*

Poseen las siguientes columnas:

Nombre de columna	Descripción
id_alojamiento	Identificador de la publicación
name	Nombre y descripción del alojamiento
category	Tipo de alojamiento
rating	Puntuación del alojamiento

city	Ciudad (No confundir con las 4 provinciasde los archivos, esto es por ejemplo una ciudad dentro de Neuquen o Rio negro)
latitude	Latitud en grados
longitude	Longitud en grados
characteristics	Características del alojamiento
check_in	Día de check-in para el cual se consultó la disponibilidad del alojamiento
check_out	Día de check-out para el cual se consultó la disponibilidad del alojamiento
price_discounted	Precio con descuento (si lo hay)
price_original	Precio original (sin descuento)

1. Carga de datos y armado del dataset

Leer los cuatro archivos de propiedades en alquiler temporal de las ciudades de la Patagonia Argentina y unirlos en un único dataframe creando una nueva columna que se llame ***provincia*** y contenga el nombre de la provincia, extraído del nombre del archivo.
Recomendación: Luego de unir, usen *reset_index(drop=True)* para que los índices les queden prolijos.

Analizar la cantidad de columnas y observaciones de la tabla.

Aclaración para este punto y el resto del trabajo: Habrá dos columnas, ***provincia*** *que habrán generado y* ***city,*** *que ya viene como información dentro del archivo. Esas columnas referirán correspondientemente a las provincias y ciudades de las publicaciones de airbnb.*

2. Análisis de duplicados

Los datos de los anuncios pueden encontrarse repetidos. Analizar si existen registros duplicados (iguales para todas sus columnas) y de existir, eliminarlos. Además, reportar si las columnas tienen valores repetidos solo mirándolas de a una. Dar una recomendación sobre qué les parecería mejor hacer en cada caso.
¿Es lo mismo un duplicado en la columna ***price_original*** que en ***id_alojamiento***?

3. Análisis de datos faltantes y limpieza

Mostrar el porcentaje de datos faltantes de cada columna
¿Cuáles son las 3 columnas con más datos faltantes?

Sobre la columna ***ratings***:

1. Llevarla a un formato donde solo quede el número con punto en vez de coma (ej: 1.5), es decir, la columna tiene que quedar de tipo ***float64*** al ejecutar ***df.info() .***
2. Luego, imputar los valores faltantes utilizando la media de esos valores.
3. Crear una variable llamada habitaciones ***best_reviews*** que para cada fila tenga el valor *True* si el rating es 5.0 y el valor *False* si tiene un rating menor.
4. ¿Cuál es el porcentaje de publicaciones que cumplen el criterio de **best reviews**?

Sobre la columna ***price_original***:

1. Llevarla a un formato donde solo quede el número entero (ej: 107), es decir, la columna tiene que quedar de tipo ***int64*** al ejecutar ***df.info()***

Crear una columna ***discounted*** que tenga valor True si la publicación tenía descuento y False si no la tenía.

Sobre la columna ***price_discounted***:

1. Imputar los valores faltantes por el valor de ***price_original*** de esa fila.
2. Llevarla a un formato donde solo quede el número entero (ej: 107), es decir, la columna tiene que quedar de tipo ***int64*** al ejecutar ***df.info()***

4. Análisis general de precios

¿Cuántos alquileres con descuento hay?
¿Cuál es el alquiler más caro? ¿Y el más barato?
Mostrar 2 histogramas de los precios originales y los precios con descuento.
¿Observa alguna diferencia?

5. Discretización de precios

Crear una nueva columna ***price_category*** en base a la distribución de precios (***price_original***) pero discretizando en intervalos de igual ancho con la siguientes etiquetas para esos bins: ['bajo', 'medio', 'alto']. Hacer un gráfico de barras con esta nueva columna ¿Qué se observa?

6. Análisis por ciudad

Agrupar por la columna ***provincia***, de modo que queden 4 grupos. Sobre ese DataFrame agrupado responder las preguntas:
¿Cuántos alquileres con descuento hay por provincia?
De cada provincia: ¿Cuál es el alquiler más caro? ¿Y el más barato?.
Para responder, muestre el DataFrame agrupado incluyendo únicamente las columnas pertinentes (y con nombres de columnas adecuados y claros).

¿Qué ciudad tiene más publicaciones con descuento? ¿En qué ciudad se encuentra el precio más caro? ¿Y el más barato?

Por último, realice 2 gráficos de cajas (boxplots) mostrando en uno la distribución de precios originales por provincia, y en otro la distribución de ratings por provincia.

7. Características del alojamiento

Poniendo el foco en la columna ***characteristics***, ¿El dataset se encuentra en formato *tidy?* ¿Por qué?

Deberán construir las columnas **baños**, **dormitorios y camas** con la cantidad de baños, dormitorios y camas y obtenidos de la columna ***characteristics****.* Notarán que esta columna posee valores de tipo string que se parecen a las listas de Python. Aclaración: El orden de esos datos en esas listas no está asegurado. Además podrían faltar datos, en ese caso completar con *NaN*.

Este dataset es real y fue scrapeado de sitio de Airbnb, en particular este punto muestra el tipo de situaciones con las que hay que lidiar al limpiar datos.

Realice 3 gráficos de barras mostrando la cantidad de publicaciones en función del número de baños, habitaciones y camas. Por ejemplo, para el gráfico de baños, se espera que una barra indique cuantas publicaciones tienen 1 baño, cuantas tienen 2 baños, etc.

8. Detección de outliers

Crear una variable ***price_standarized*** en base la estandarización de la variable ***price_original.*** *Identifique outliers con el criterio de 3 desvíos estándar visto en clase, y reporte la cantidad de outliers para cada ciudad (****city****) ordenados de mayor a menor.*

Mirando los resultados en general (no por ciudad) y observando en particular la columna generada ***price_category.*** Indique cuantos outliers hay por cada categoría de precio (alto / medio / bajo) y grafíquelos en un gráfico de barras.
¿Les parece que hay algo raro? ¿Los outliers tendrían que estar siempre en bajos o en altos? ¿Podrían estar en medios?
Si se hubiera realizado una discretización de igual frecuencia ¿Cómo cree que hubiera cambiado esto?

9. Análisis estacional y de duración

Analizar si existe una variación estacional en los precios agrupando por mes de check_in. Mostrar esta variación en un gráfico de barras. Considerando la región de la Argentina de donde se tomaron los datos, ¿le parece que tienen sentido esos resultados?

Duración de las estancias. Calcular la duración de cada estancia utilizando las columnas **check_in** y **check_out**. Crear una nueva columna **stay_duration** que contenga la duración en días.

¿Existe alguna relación entre la duración de la estancia y el precio del alojamiento? Graficar esta ambas variables en un scatter plot y comentar lo observado.

10. Correlación entre precio y características

Habiendo creado las columnas **bathrooms**, **bedroom y beds** en el punto 7, agrupar el dataset por cada una de ellas, calcular el promedio de ***price_original*** *y* graficar cada una utilizando gráficos de barras ¿Qué observan? ¿Hay alguna tendencia? ¿Siempre se cumple?

Posteriormente, sobre el dataset sin agrupar, calcular la correlación de ***price_original*** Vs. cada una de esas columnas (**bathrooms**, **bedroom y beds**) ¿Cuál se asocia más fuertemente con el precio? (Recordar: Correlación no implica causalidad).