

Taller Final Pandas y Power Bi

Jesus David Pedraza – 01210372007

Docente: Deybeth Hernando Riaño Nuñez

Lenguajes de programación

Universidad de Santander (UDES)

Bucaramanga (Santander)

Ingenieria de software

18/11/2024

## Tabla de contenido

Introducción .....	3
Preguntas Propuestas Para responder con pandas .....	4
Preguntas propuestas para responder con Power BI .....	4
Propósito del análisis .....	5
Cómo lo voy a hacer .....	5
Resultados encontrados (Análisis de los datos) .....	5
Conclusión final .....	10

## Introducción

En este proyecto voy a analizar los datos del dataset **New York City Airbnb Open Data** y **FIFA World Cup**, usando Python y la biblioteca pandas (para el primer dataset) como herramientas principales y Power BI desktop (Para el segundo dataset). El objetivo es responder a cinco preguntas clave relacionadas con los alojamientos en Nueva York y la participación de estos equipos en los mundiales con varios factores más. Esto me permitirá explorar patrones y características importantes, como los precios, la disponibilidad y las reseñas. En cuanto al futbol, me permitirá analizar cuantas personas asistieron a ese mundial, el número de ganadores de mundiales, cuantas veces ha participado un equipo en los mundiales, las fechas y goles marcados en cada uno de ellos.

El conjunto de datos incluye información detallada, como las coordenadas geográficas, precios por noche, disponibilidad anual, cantidad de reseñas y otros datos interesantes que me ayudarán a sacar conclusiones valiosas.

También se contará con un mapa de referencia indicando donde se realizó el mundial con el número de asistentes y demás datos.

## Preguntas Propuestas Para responder con pandas

1. **¿Qué relación existe entre la ubicación (latitud, longitud) y el precio promedio por noche en cada grupo de barrios?**
  - Me interesa saber si el lugar donde está ubicado un alojamiento tiene un impacto claro en el precio promedio.
2. **¿Existen diferencias significativas en los precios y disponibilidad entre los alojamientos con más de 50 reseñas y aquellos con menos de 10?**
  - Quiero analizar si los alojamientos más comentados tienen características diferentes en cuanto a precio y disponibilidad.
3. **¿Qué barrios tienen las propiedades más disponibles al año (mayor disponibilidad 365 promedio) y cómo se relacionan con el precio promedio y el tipo de habitación?**
  - Aquí analizaré cuáles son las zonas con mayor disponibilidad y si esto influye en los precios y en el tipo de habitación que se ofrece.
4. **¿Cuáles son las características más comunes (precio, disponibilidad, reseñas, etc.) de los alojamientos en barrios con más de 500 listados activos?**
  - Mi objetivo es identificar patrones en las áreas con muchos listados y ver qué las hace diferentes del resto.
5. **¿Cuáles son los cinco barrios más populares en términos de número total de reseñas, y cómo varía el precio promedio en estos barrios?**
  - Por último, quiero saber cuáles son los barrios más destacados en cuanto a reseñas y cómo se comportan los precios en esas zonas.

## Preguntas propuestas para responder con Power BI

1. **¿Cuántos goles se anotaron, partidos se jugaron y equipos se clasificaron?**
  - Me interesa analizar si hay una tendencia en la evolución de estos datos a lo largo del tiempo, y si ciertos años destacan por un incremento significativo en alguno los factores.
2. **¿Qué países han ganado más Copas del Mundo?**
  - Quiero investigar si los países con más títulos tienen un desempeño consistentemente destacado en términos de goles anotados, partidos ganados, o clasificaciones consecutivas.
3. **¿Qué países han participado más en las Copas del Mundo?**

The screenshot shows a Jupyter Notebook environment. On the left is a file explorer with a sidebar containing icons for files, folders, and search. The main area of the file explorer shows a folder named 'sample\_data' containing a file named 'AirbnbNY.csv'. Below the file explorer is a status bar indicating 'Disco' and '75.24 GB de espacio disponible'.

The central code editor displays the following Python code:

```
import pandas as pd

df = pd.read_csv('AirbnbNY.csv', header=0)

df
```

Below the code editor, the data is displayed as a table with 12 columns: id, name, host\_id, host\_name, neighbourhood\_group, neighbourhood, latitude, longitude, room\_type, price, minimum\_nights, and maximum\_nights. The table contains 5 rows of data, with the first row being the header and the subsequent rows representing individual listings. The table is truncated on the right side, indicated by an ellipsis.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	maximum_nights
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149		1
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225		1
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150		3
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89		1
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80		10
...	...	...	...	...	...	...	...	...	...	...	...	...

At the bottom of the notebook, a status bar shows 'Disco' and '75.24 GB de espacio disponible'.

- Relación entre ubicación y precio promedio por noche en cada grupo de barrios

```

0 s [26] def relacion_ubicacion_precio(df):
        grouped = df.groupby('neighbourhood_group').agg([
            'latitude': 'mean',
            'longitude': 'mean',
            'price': 'mean'
        ]).rename(columns={'price': 'avg_price_per_night'})
        return grouped

result_1 = relacion_ubicacion_precio(df)
print(result_1)

```

neighbourhood_group	latitude	longitude	avg_price_per_night
Bronx	40.848305	-73.884552	87.496792
Brooklyn	40.685036	-73.951190	124.383207
Manhattan	40.765062	-73.974607	196.875814
Queens	40.731531	-73.872775	99.517649
Staten Island	40.610225	-74.105805	114.812332

Manhattan tiene el precio promedio más alto (\$196.88), ya que es el corazón turístico y económico de la ciudad. El Bronx tiene el precio más bajo (\$87.50), lo que puede ser reflejo de una menor demanda turística o en opciones más económicas. En Brooklyn (\$124.38) y Queens (\$99.52), encontramos precios más accesibles, ideales para quienes buscan estar cerca de Manhattan sin gastar tanto. Staten Island, aunque no es tan turística, tiene precios intermedios (\$114.81), probablemente debido a su carácter más residencial y tranquilo.

**Conclusión:** Los precios reflejan claramente la popularidad y el atractivo de cada área. Manhattan es la opción Premium, mientras que el Bronx y Queens son perfectos para quienes buscan ahorrar.

- Diferencias en precios y disponibilidad según el número de reseñas

```

0 s [27] def diferencias_precios_disponibilidad(df):
        high_reviews = df[df['number_of_reviews'] > 50].agg({'price': 'mean', 'availability_365': 'mean'})
        low_reviews = df[df['number_of_reviews'] < 10].agg({'price': 'mean', 'availability_365': 'mean'})
        return pd.DataFrame({'high_reviews': high_reviews, 'low_reviews': low_reviews})

result_2 = diferencias_precios_disponibilidad(data)
print(result_2)

```

	high_reviews	low_reviews
price	128.707573	163.732724
availability_365	164.266274	95.430996

Aquí los alojamientos con más de 50 reseñas tienen un precio promedio menor (\$128.71) pero están disponibles más días al año (164 días). Esto nos dice que los alojamientos con muchas reseñas son más confiables y populares, por lo que tienden a estar reservados con más frecuencia. En cambio, los alojamientos con menos de 10 reseñas tienen un precio promedio más alto (\$163.73) y menos disponibilidad (95 días), probablemente porque son nuevos o menos valorados, lo que limita su demanda.

**Conclusión:** Los alojamientos con más reseñas son más asequibles y tienen más reservas, mientras que los menos revisados intentan compensar con precios más altos.

- **Barrios con mayor disponibilidad y su relación con precio y tipo de habitación**

```
def barrios_mayor_disponibilidad(df):
    grouped = df.groupby('neighbourhood').agg({
        'availability_365': 'mean',
        'price': 'mean',
        'room_type': lambda x: x.mode()[0]
    }).rename(columns={'availability_365': 'avg_availability', 'price': 'avg_price'})
    top_neighbourhoods = grouped.sort_values('avg_availability', ascending=False).head(5)
    return top_neighbourhoods

result_3 = barrios_mayor_disponibilidad(data)
print(result_3)
```

	avg_availability	avg_price	room_type
neighbourhood			
Fort Wadsworth	365.000000	800.000000	Entire home/apt
Co-op City	364.000000	77.500000	Private room
Willowbrook	351.000000	249.000000	Entire home/apt
Eastchester	333.461538	141.692308	Entire home/apt
Richmondtown	300.000000	78.000000	Entire home/apt

Algunos barrios destacan por tener alta disponibilidad, pero con características muy diferentes:

- **Fort Wadsworth** tiene una disponibilidad completa de 365 días, pero con un precio demasiado alto (\$800).
- **Co-op City** y **Richmondtown**, ofrecen precios muy buenos (\$77.50 y \$78) y también tienen una gran disponibilidad. Son ideales para estadías económicas.
- Otros barrios, como **Willowbrook** y **Eastchester**, tienen precios medios y alta disponibilidad, lo que puede atraer a quienes buscan alojamientos de largo plazo.

**Conclusión:** Alta disponibilidad puede significar exclusividad (como en Fort Wadsworth) o accesibilidad económica (como en Co-op City). Todo depende del barrio y del tipo de viajero que se hospede.

- **Características comunes en barrios con más de 500 listados activos**

```

0 s ▶ def características_comunes(df):
    high_listings = df.groupby('neighbourhood').filter(lambda x: len(x) > 500)
    summary = high_listings.agg({
        'price': 'mean',
        'availability_365': 'mean',
        'number_of_reviews': 'mean',
        'calculated_host_listings_count': 'mean'
    }).rename({
        'price': 'avg_price',
        'availability_365': 'avg_availability',
        'number_of_reviews': 'avg_reviews',
        'calculated_host_listings_count': 'avg_listings_count'
    })
    return summary

result_4 = características_comunes(data)
print(result_4)

```

avg_price	156.826220
avg_availability	104.305559
avg_reviews	22.881542
avg_listings_count	7.266576
dtype:	float64

En estos barrios encontramos un precio promedio de \$156.83, lo que no es tan elevado considerando la cantidad de opciones disponibles. También tienen una disponibilidad promedio de 104 días, lo que muestra que la demanda es alta, pero no a tal punto que se llegue a considerar saturada. también hay un promedio de 22 reseñas por alojamiento, lo que indica que son áreas confiables para los viajeros. Otro punto interesante es que muchos anfitriones tienen varias propiedades activas.

**Conclusión:** Estos barrios son zonas muy activas con opciones para todos los gustos y se ajusta a la economía de los clientes.

- **Cinco barrios más populares por número total de reseñas y variación en precio promedio**

```

0 s ▶ def barrios_mas_populares(df):
    popular_neighbourhoods = df.groupby('neighbourhood').agg({
        'number_of_reviews': 'sum',
        'price': 'mean'
    }).rename(columns={'number_of_reviews': 'total_reviews', 'price': 'avg_price'})
    top_neighbourhoods = popular_neighbourhoods.sort_values('total_reviews', ascending=False).head(5)
    return top_neighbourhoods

result_5 = barrios_mas_populares(data)
print(result_5)

```

	total_reviews	avg_price
neighbourhood		
Bedford-Stuyvesant	110352	107.678244
Williamsburg	85427	143.802806
Harlem	75962	118.974041
Bushwick	52514	84.800406
Hell's Kitchen	50227	204.794178



Estos barrios son los favoritos de los viajeros, cada uno con su propio carácter:

1. **Bedford-Stuyvesant** lidera en reseñas (110,352) con un precio accesible (\$107.68), perfecto para los viajeros que buscan ahorrar.
2. **Williamsburg** es moderno y tiene precios más altos (\$143.80), ideal para quienes buscan un ambiente moderno.
3. **Harlem** combina tradición y modernidad, con un precio razonable (\$118.97) y una alta cantidad de reseñas (75,962).
4. **Bushwick** destaca por ser económico (\$84.80) y popular entre jóvenes y viajeros mochileros.
5. **Hell's Kitchen**, con precios premium (\$204.79), es un destino más exclusivo para los viajeros que buscan exclusividad.

**Conclusión:** Estos barrios ofrecen una mezcla de accesibilidad, estilo y exclusividad, dependiendo del presupuesto y las preferencias de los viajeros.

## Conclusión final

Este proyecto me permitió profundizar en el análisis de datos y en el uso de pandas. Además, creo que los resultados pueden dar una idea interesante de cómo funciona el mercado de Airbnb en Nueva York, considerando tanto las características de los alojamientos como el comportamiento de los usuarios.