**Reporte Inversiones inmobiliarias para alquiler vacacional**

**Entendimiento inicial**

El data set inicial llamado **milan\_listings** consta de 24.346 registros de propiedades de alquiler dentro de la aplicación de Airbnb y 75 columnas con diferentes características de la propiedad y su anfitrión.

Dentro de la información de la propiedad se tiene dimensiones cómo vecindario, tipo de propiedad, descripción, capacidad de huéspedes, número de habitaciones, baños y camas, comodidades, minimo y máximo de noches, precio por noche, etc.

Por el lado del anfitrión se tiene información de su nombre, ubicación, comportamiento de respuesta, foto, cantidad de propiedades que maneja, si es verificado entre otras.

En el análisis inicial vemos que hay varias dimensiones que deben ser descartadas por diferentes motivos:

* **Completitud**: neighborhood\_overview (50%), host\_about (50%), host\_neighbourhood (63%), neighbourhood (46%), neighbourhood\_group\_cleansed (100%), calendar\_updated (100%)
* **Relevancia (valores únicos o no descriptivos**): scrape\_id, last\_scraped, source, has\_availability, calendar\_last\_scraped.

**Limpieza y organizacion de la data**

Se llevan acabo los siguientes ajustes de algunos campos para limpiar la data original y sacar mayor provecho a la base.

* Tratar el **id** y **host\_id** cómo valores categóricos.
* Separar el **host\_location** en ciudad y pais.
* Nuevo campo **host\_since\_days** de días de experiencia del host.
* Revisando los anuncios (url) de diferentes casos de “0 baños” dentro de **bathrooms\_text**, se supone que las propiedades de tipo “Entire home/apt” que tengan 0 baños son errores y tienen al menos 1 baño. Por otro lado, los que sean “Private room” tienen al menos 1 baño compartido. Y finalmente los que contentan “Half-bath” se consideran igualmente baños compartidos. Con esto se limpia el campo original y se crea el categórico **bathrooms\_group**.
* Se eliminan las propiedades que tienen **minimum\_nights** de estadía mayor a 15, ya que estos generalmente son para estudiantes o no corresponden a actividades vacacionales. Se conserva aproximadamente el 96% de la base igualmente.
* Se transforma el valor de campo **price** que viene como string con diferentes signos a un valor de numérico entero. Se eliminan las propiedades que no tienen **price** (nan). Adicional a esto, al revisar los datos atípicos de **price** nos damos cuenta que los valores que están por encima de 9999 dolares son errores de digito por lo que se descartan.

**Top 5**

A continuación, está el top 5 de atributos relevantes para el analisis de inversiones, sin embargo, hay relaciones entre estos atributos y otras dimensiones que pueden generar insights valiosos.

**Price**

Variable cuantitativa de tipo entero, promedio de 167.95 dolares y desviación estándar de 342.61

**A white sheet of paper with black lines

Description automatically generated**

El histograma muestra que aun cuando se eliminaron ciertos outliers sigue existiendo propiedades muy alejadas de la media de precio, con los percentiles podemos ver que el 70% de las propiedades están entre 55 y 165 dolares por noche.

A black and white screen with numbers and percentages

Description automatically generated

De igual forma como es de esperarse por el histograma, se tiene un valor positivo de Skewness de 17.19 y uno muy alto de Kurtosis de 378.72 ya que hay un gran número de propiedades con precios mucho más altos que la media.

A black and white screen

Description automatically generated

**Review\_scores\_rating**

Variable cuantitativa continua, promedio 4.7 y desviación estándar de 0.42.

A white sheet with blue squares

Description automatically generated

El 95% de las propiedades tienen un score sobre 4. Y el 50% sobre 4.8. Cómo se puede ver con los percentiles y con la grafica el comportamiento de los reviews es muy positivo por lo que tiene una asimetría negativa -3.68, e igualmente existen algunos valores atípicos hacia valores bajos por lo que la curtosis es de 21.7

A screen shot of a computer screen

Description automatically generated

**Neighbourhood\_cleansed**

Variable cualitativa tipo string, tiene 88 valores distintos y el top 5 de vecindarios son Buenos Aires – Venezia, Duomo, Sarpi, Brera y Loreto.

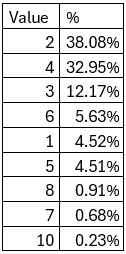
**A graph of a bar chart

Description automatically generated with medium confidence**

**Accommodates**

Variable cuantitativa de tipo entera, promedio 3.23 y desviación estándar de 1.49:

**A black and white screen with numbers and percentages

Description automatically generated **

**A graph of a person with a number of red squares

Description automatically generated with medium confidence**

Las propiedades con capacidad de 2 personas representan el 38% y las de 4 el 33%.

**Bathrooms\_text**

Para este caso se decidió crear un nuevo campo para reducir el número de valores y tomarlo como una variable categórica. Todos los baños compartidos asi sea uno o varios se asignaron a un mismo grupo, y los baños privados se separaron en cantidades “1”, “2 a 3” y “4+”.

De acuerdo con esto el 87.5% (19.370) de propiedades tienen un único baño, y el 10.7% (2.384) tienen 4 o más baños.

A white background with red squares

Description automatically generated

**Estrategia de Análisis**

Para establecer la mejor opción de inversión inmobiliaria teniendo en cuenta la data que tenemos a disposición, realizaremos diferentes analisis multivariado con el uso de gráficos, por ejemplo, graficas de caja para comparar distribuciones y medias, y de pruebas estadísticas, como la t-test o la z-test, para encontrar los valores de los atributos que hacen que una propiedad pueda ser publicada por un mayor precio y que afecta el rating dado a la propiedad.

Por otro lado, para incentivar la reserva de la propiedad a adquirir, realizar el mismo analisis entre las caracteristicas de los anfitriones y las comodidades mencionadas y el score y el numeró de reviews de las propiedades.