1. Selección del dataset de trabajo

La ciudad seleccionada es Chicago, la capital de Illinois, Estados Unidos y su ciudad más poblada. Esta ciudad es de interés debido a que es una metrópolis que se encuentra cerca a un lago y este lago tiene una playa que visualmente se parece al mar.



El dataset se puede descargar en el siguiente enlace: <http://data.insideairbnb.com/united-states/il/chicago/2023-06-18/data/listings.csv.gz>

2. Entendimiento inicial de datos

NOTEBOOK: 1\_EDA\_inicial.ipynb

# Dimensiones de datos

Al igual que lo presente en el diccionario de datos, se tienen 75 columnas para 8.024 propiedades, donde se tiene una sola fila por cada propiedad.

# Tipos de datos

Al hacer una evaluación de los tipos de datos, se encuentran variables cuantitativas discretas y continuas, y variables categóricas de tipo nominal y ordinal. Después de hacer una eliminación de columnas innecesarias con los siguientes criterios (véase notebook):

* Contiene información del método de scraping
* Tienen información duplicada y menos limpia que en otras columnas.
* Información del host que agrupa información de varios listings, ya que el análisis de listings será asumiendo independencia entre estos.
* Información que necesitaría métodos adecuados de Natural Language Processing (NLP) para poder encontrar patrones.
* Primera y última fecha de review, porque la información de la tabla está agrupada y reviews\_per\_month puede resumir el comportamiento temporal de conteo de reviews.

Y de hacer correcciones en el tipo de dato tienen los siguientes tipos de datos:

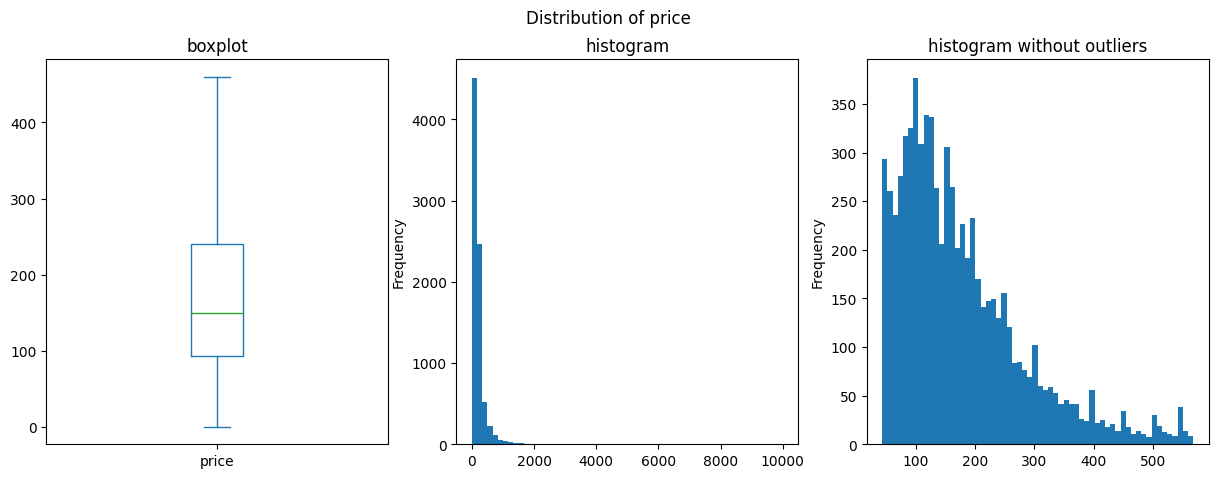
* **Categóricos nominales:** id, host\_response\_time, host\_is\_superhost, host\_identity\_verified, neughbourhood\_cleansed, property\_type, room\_type, amenities (requiere más preprocesamiento), has\_availability, instant\_bookable.
* **Categóricos ordinales:** bathrooms\_text.
* **Cuantitativa discreta:** accomodates, bedrooms, beds, minimum\_nights, maximum\_nights, minimum\_minimum\_nights, maximum\_minimum\_nights, maximum\_maximum\_nights, minimum\_maximum\_nights, availability\_30, availability\_60, availability\_90, availability\_365, number\_of\_reviews, number\_of\_reviews\_ltm, number\_of\_reveiws\_l30d
* **Cuantitativa continua:** latitude, longitude, price, minimum\_nights\_avg\_ntm, maximum\_nights\_avg\_ntm, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, review\_scores\_location, review\_scores\_value, reviews\_per\_month.

# Variables de interés

En esta sección se selecciona un top 5 de variables de interés para el cálculo de rentabilidad, ordenando de mayor a menor prioridad. Las primeras 4 no presentan ningún valor nulo, lo cual es bastante beneficioso para el análisis.

## 1. Precio (price)

El precio es fundamental para este análisis, debido a que ninguna de las otras variables tiene en cuenta el valor de rentar la propiedad. Para este atributo se puede inferir un comportamiento altamente sesgado hacia valores entre $100 y $200 USD, donde la mediana se encuentra cerca de los $150USD. Cómo se puede observar en el siguiente gráfico existen propiedades que tienen valores de renta diarios de hasta $10 mil USD, lo cual es bastante lejano de lo normal por rango intercuartílico que es debajo de $400 USD por día.

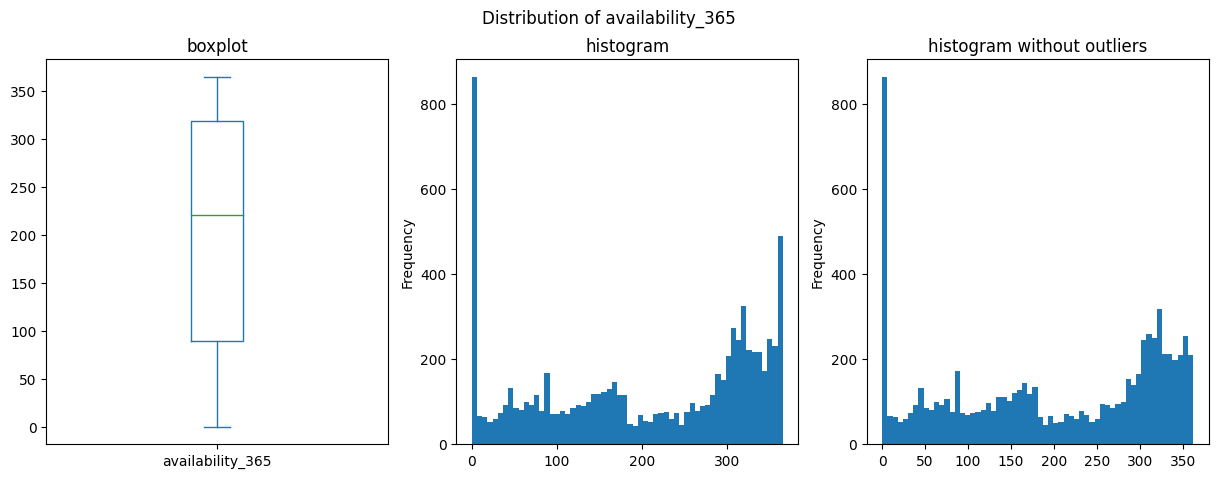


Afortunadamente esta variable no presenta ningún valor nulo, por lo que se puede usar para estimaciones de ganancia para cualquier propiedad.

## 2. Disponibilidad 365 (availability\_365)

El solo precio no es suficiente para estimar el margen de ganancia de una propiedad, ya que puede suceder el caso de que una propiedad tenga un alto precio de renta y no sea rentada en ningún día del año. Para suplir esta necesidad están las columnas de disponibilidad a 30, 60, 90 y 365 días en el futuro. Como esta ciudad es una metrópolis, el interés de visitarla ya sea por descanso o trabajo puede distribuirse todo el año, lo que sumado al hecho de que hay estaciones genera un posible comportamiento estacional no evaluable en este dataset. Suponiendo que tal comportamiento existe, es necesario tener información del uso de la propiedad durante todo el año, en caso de que no sea rentada en los siguientes días inmediatos.

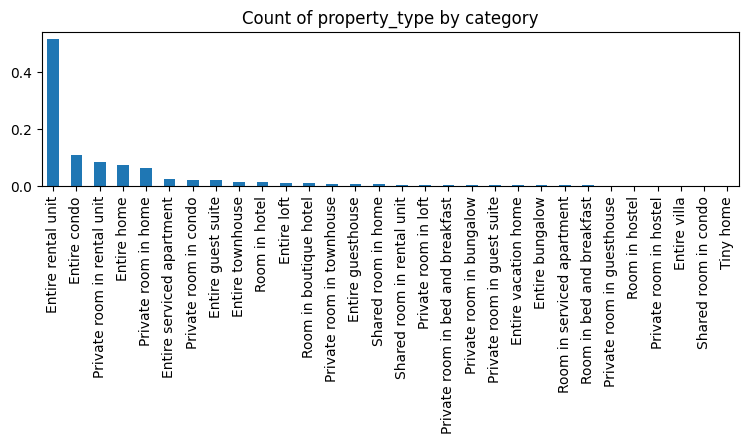
Siguiendo un estudio similar al desarrollado con precio, se tienen las siguientes gráficas:



En donde se evidencia que hay una serie de propiedades que no tienen ni un día disponible en el siguiente año. Este alto porcentaje de ocupación podría indicar un alto nivel de rentabilidad. Para los demás valores antes de 365 se tiene un patrón desordenado, relativamente uniforme, con media de disponibilidad entre 200 y 250 días, lo cual está cerca de ser el 60% del año. Por último se tiene una serie de propiedades que tienen los 365 días disponibles, representando una segunda moda de distribución que puede representar propiedades que no generan ganancia por estar cerradas o porque nadie las quiere rentar.

## 3. Tipo de propiedad (property\_type)

Como se vio anteriormente con el precio, hay propiedades que pueden llegar a ser rentadas por $10.000 USD, lo cual implica que existen múltiples necesidades que los huéspedes tienen en cuenta a la hora de rentar una propiedad. Uno de los factores que más puede influir en las necesidades es el tipo de propiedad, que puede ir desde un apartamento hasta una mansión completa, presentando el siguiente comportamiento:

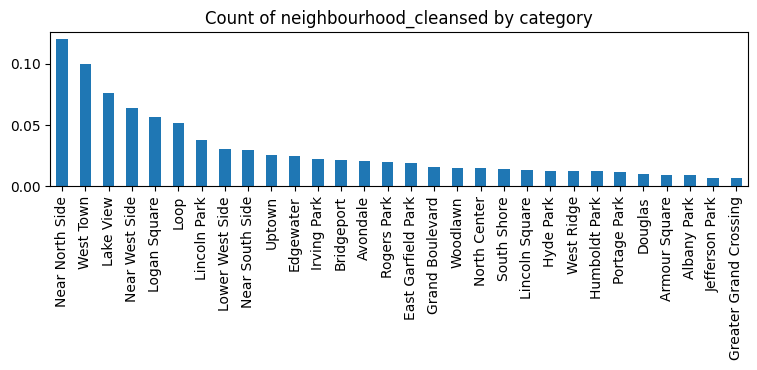


Hay 49 tipos de propiedades, donde más del 50% de las propiedades son rentadas como unidades completas, es decir, todo el edificio, o son rentadas como todo el apartamento. Estas unidades completas corresponden o a todo el recinto, o un apartamento, y al relacionarlo con el número de habitaciones se tiene una explicación de por qué la mediana de número de habitaciones son 4 por propiedad.

## 4. Barrio (neighbourhood\_cleansed)

Al igual que en Bogotá, puede ocurrir el hecho de que existan zonas especializadas en alojamiento para extranjeros o personas que desean vivir temporalmente en la ciudad. Estas zonas generalmente están cerca de lugares de interés, como sitios turísticos y centros de comercio y negocios. Siguiendo esta hipótesis, se tiene que algunos barrios pueden ser de mayor interés para los huéspedes y combinados con las variables anteriores pueden generar mayor rentabilidad.

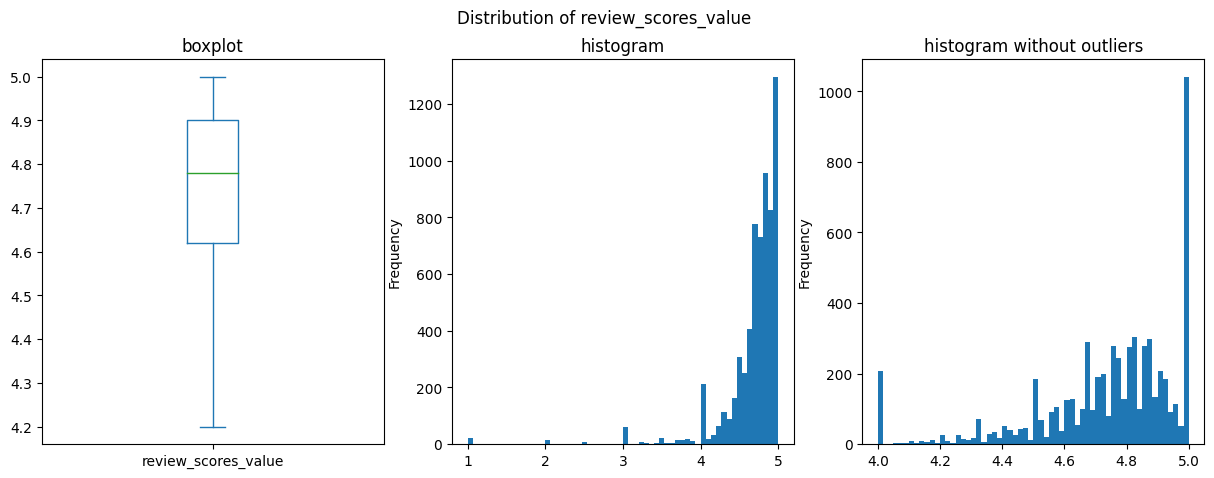
Al ser una variable categórica de alta cardinalidad, con 77 categorías, se tiene un comportamiento en el que solo 6 barrios se llevan más del 50% de todas las propiedades:



Donde estos barrios son: Near North Side, West Town, Lake View, Near West Side, Logan Square, Loop y Lincoln Park.

## 5. Review del valor (review\_scores\_value)

El valor podría llegar a resumir la información presente en los demás scores, dado que para calificar si un valor es justo se tiene en cuenta, checkin, ubicación, amenities y limpieza entre otras. Desafortunadamente, solo 80% de las propiedades tienen este valor y, como se puede observar en la siguiente imagen, los usuarios que califican suelen hacerlo con valores encima de 4.0:



No obstante, también hay una muy pequeña cola en valores de 1 y 2, por lo que esta variable se podría entender como un indicador de experiencias “extremas” de estancia. Esto quiere decir, el hecho de tener un review de valor implica que el usuario tuvo una experiencia excelente o una pésima.

3. Estrategia de análisis

Para el análisis se creará un estimador de ganancia definido como:

Con este estimador se desea calcular aproximadamente cuánto se ganaría en un año al invertir en la propiedad, aprovechando que el precio tiene formato diario y availability\_365 brinda una idea de el porcentaje de ocupación anual de la propiedad en días.

Luego se procederan a hacer comparaciones bivariadas de la ganancia contra variables categóricas y númericas como property\_type, neighbourhood\_cleansed y review\_scores\_value usando estadísticos de tendencia central como la media por grupo, así como cuartiles, desviación estándar y correlaciones. Para el análisis gráfico se tendrá para variables categóricas boxplots comparando contra la ganancia y scatters para las variables cuantitativas. Nótese que para algunas variables se requerirá un preprocesamiento que convierta de numérica a categórica como review\_scores\_value, pero el proceso es estándar.

Para comparar si los grupos top de las variables categóricas son realmente diferentes de los demás se aplica un test de Kolmogorov-Smirnov (KS) del cual se espera un pvalue debajo de 0.05 y un estadístico encíma de 0.20 para mostrar una fuerte separación. Finalmente se harán comparaciones trivariadas contra la ganancia usando bubbleplot, boxplot con hue y heatmaps. La primera para dos variables cuantitativas, la segunda para una variable cuantitativa y una categórica, y la tercera para dos variables categóricas.

4. Desarrollo de la estrategia

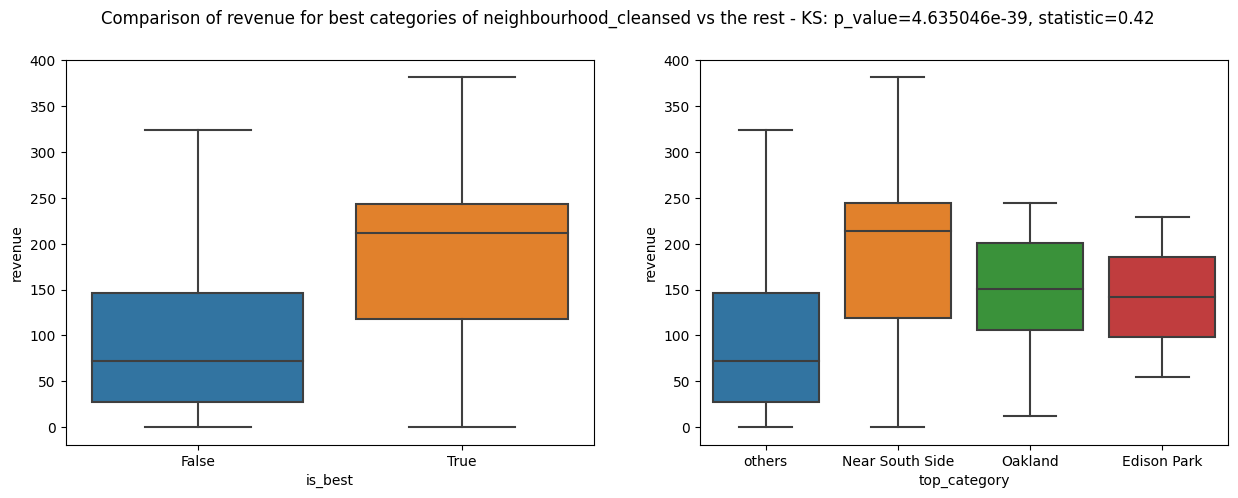
Se genera un notebook de preprocesamiento y búsqueda de insights con la estrategia descrita en la sección 3.

5. Generación de resultados

Acorde al proceso descrito en la sección 4, se recomiendan propiedades con las siguientes características:

# Barrio (neighbourhood\_cleansed)

A partir del revenue diario se puede establecer que las propiedades más rentables se encuentran en los barrios: Near South Side, Oakland y Edison Park. En la siguiente gráfica se puede visualizar el hecho de que en estos tres barrios, el 75% de las propiedades genera ganancias que están por encima de $140 USD diarios, mientras que para el resto de barrios se tiene totalmente lo contrario. En términos de distribuciones, la distribución de estos barrios es diferente del resto, con un estadístico de Kolmogorov-Smirnov (KS) por encima de 0.20.

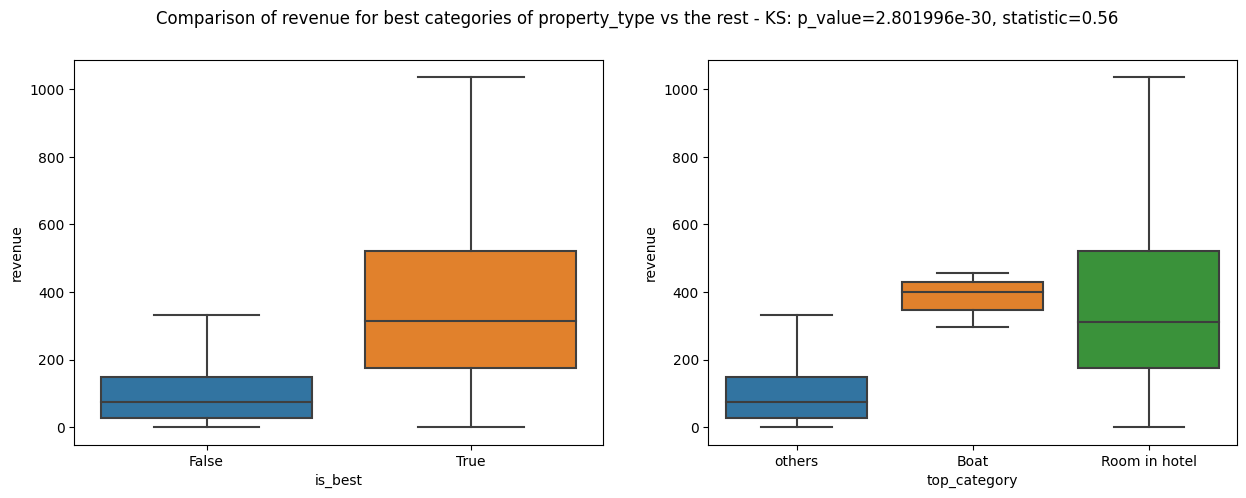


Yendo al detalle de lo observado para cada barrio individualmente, se tiene lo siguiente:

* Visualmente y ordenando por revenue medio se tiene que el barrio con propiedades más rentables es Near South Side, con una rentabilidad media de $213 USD por propiedad. Seguido de Oakland y Edison Park donde 75% de las propiedades genera más de $97 USD diarios.
* En Near South Side, se tiene un centro de comercio principal de Chicago, según Wikipedia, y esto también se refleja en el hecho de que se tengan 237 propiedades disponibles en este barrio.

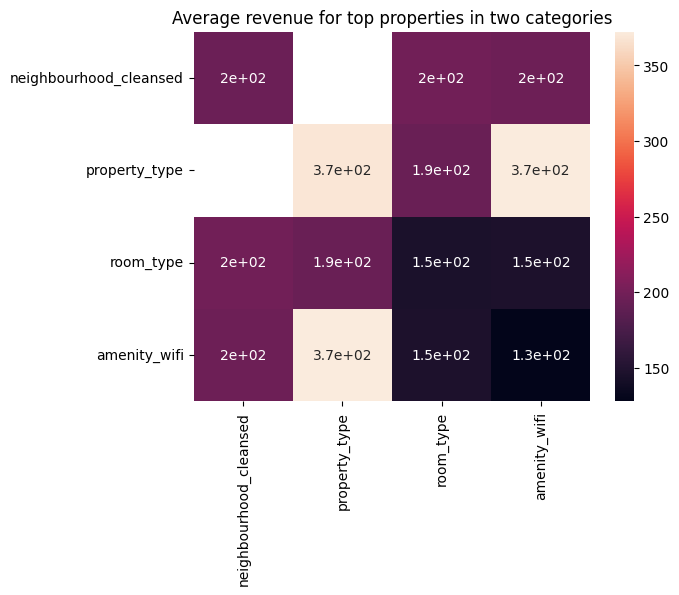
# Tipo de propiedad (property\_type)

De manera similar al barrio, el tipo de propiedad también influencia positivamente la cantidad de revenue recibido, donde con un estadístico de KS de 0.56 se tiene como tipo de propiedad más rentable los botes, seguido de las habitaciones de hotel, ya que el 75% de estos tipos de propiedades presentan un valor de revenue encima de $ 200 USD diarios, caso que no ocurre con los demás tipos de propiedad:



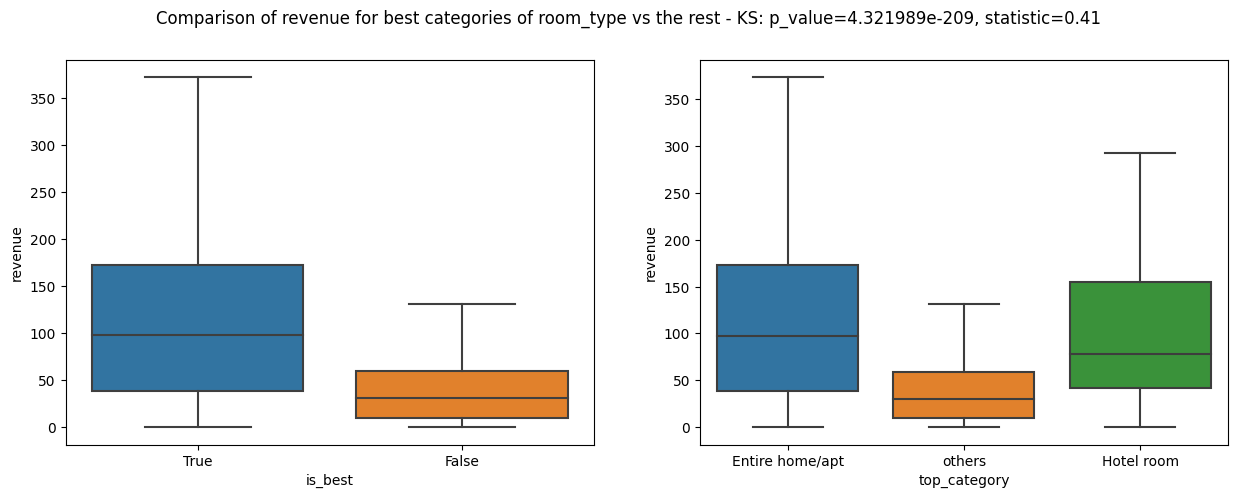
No obstante, solo hay 3 botes en todo el dataset, lo cual puede presentar una subrepresentación de la ganancia real, o por el lado positivo, una oportunidad de invertir en un negocio de baja competencia. Respecto de las habitaciones de hotel se tienen 98, pero al ser Chicago una metrópolis, se podrían encontrar habitaciones desaprovechadas por fuera del dataset para generar ganancia.

Al observar la ganancia promedio entre el cruce de mejores tipos de propiedad y barrios se encuentra que no existe ni un ejemplo disponible. Lo cual, de manera similar a los botes, presenta el potencial de incurrir en un negocio que tenga muy baja competencia.



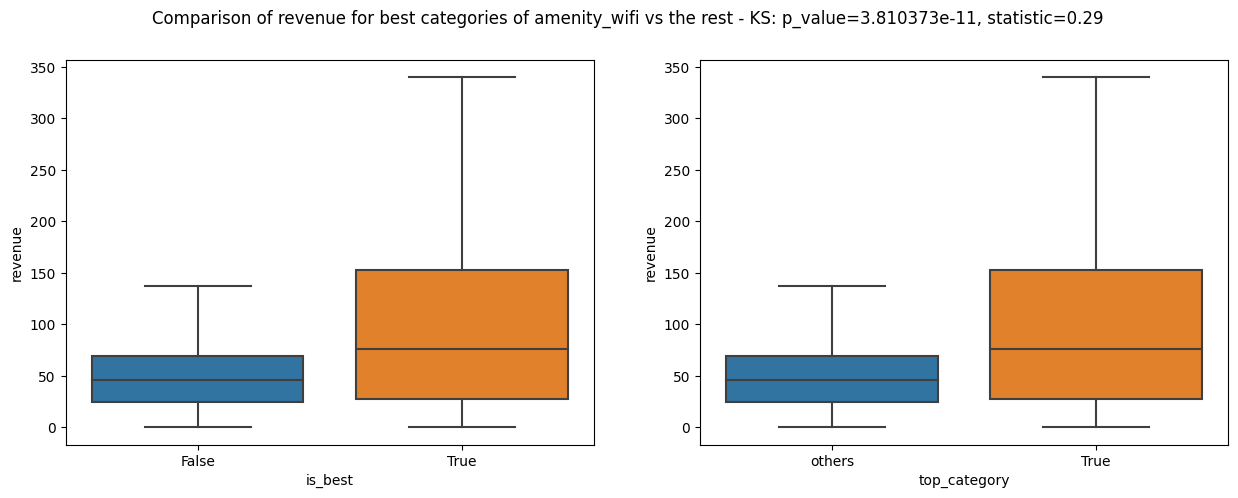
# Tipo de habitación (room\_type)

Coincidiendo con el tipo de propiedad, se tiene a las propiedades que se rentan en su totalidad y las habitaciones de hotel como las más rentables, en donde ambas categorías tienen un 75% de propiedades con una ganancia diaria por encima de $40 USD, mientras que para casi el 75% de los demás tipos ocurre lo contrario.



# Amenities (amenity\_wifi)

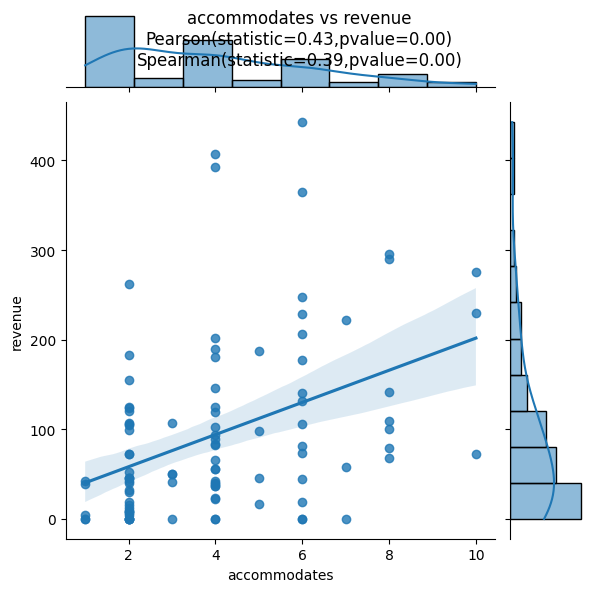
Estando en la era de Big Data, se tiene que el acceso a internet es un recurso básico y necesario para poder vivir bien en la actualidad. Esto se refleja en el hecho de que el 75% de las propiedades sin wifi llega a ganar lo que gana el 50% de las que si lo tienen, es decir, un valor por encima de los $75 USD.

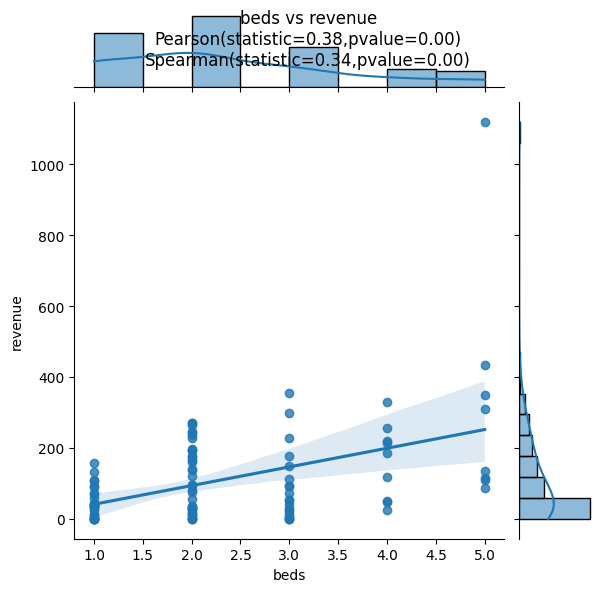


Al volver a la matriz de ganancia media entre categorías top, se tiene que se puede hacer una alta ganancia de $350 USD diarios promedio con propiedades que sean habitaciones de hotel y tengan wifi.

# Número de huéspedes y camas (accommodates y beds)

Estas fueron las únicas variables numéricas que presentaron una correlación moderada con el revenue, con valores de cerca del 0.40. Estas correlaciones pueden llegar a ser engañosas en el sentido de que la mayoría de datos se concentra entre valores de 1 a 4 y la tendencia de la correlación puede estar siendo bastante afectada por propiedades altamente lujosas, cuyo precio de renta diario es alto y tienen un alto número de habitaciones o camas.





Con estos resultados no se recomienda guiarse por el número de habitaciones o camas, ya que incluso los resultados de las variables categóricas demuestran que las habitaciones de hotel individuales pueden generar mayor rentabilidad.

# Conclusión

Los mejores tipos de propiedades son aquellas en los barrios Near South Side, Oakland y Edison Park. El primer barrio es un barrio a la costa del lago Michigan, por lo que puede ser una oportunidad de alta ganancia para rentar botes sin sufrir por altos niveles de competencia. Lo mismo sucede con Oakland, pero al igual que Edison Park, al ser zonas de gentrificación, lo mejor sería optar por comprar habitaciones de hotel. Sin importar el tipo de propiedad que se compre y su ubicación, hay que asegurarse de que siempre tenga wifi para asegurar la mayor ganancia posible.