1. Introducción

- **Objetivo del Análisis**: Este informe presenta un análisis exploratorio de una base de datos de Airbnb en la ciudad de Nueva York. Los objetivos principales son limpiar los datos, realizar transformaciones y aplicar técnicas estadísticas y de aprendizaje automático para obtener insights útiles sobre los alojamientos en diferentes barrios.
- **Metodología**: La metodología incluye la limpieza y transformación de datos, análisis exploratorio de variables, y un modelo de regresión lineal para entender las relaciones entre las características de los alojamientos y el precio.

2. Limpieza de la Base de Datos

- **Descripción de los Datos**:
  - La base de datos contiene variables clave como `neighbourhood_group`, `room_type`, `price`, `minimum_nights`, `reviews_per_month`, entre otras.
- **Pasos de Limpieza**:
  - **Eliminación de Valores Duplicados**: Se identificaron y eliminaron registros duplicados para evitar redundancias.
  - **Eliminación de Columnas Irrelevantes**: Columnas como `id`, `host_name`, y `last_review` fueron eliminadas por no aportar al análisis.
  - **Manejo de Valores Nulos**:
    - Se eliminan las filas con nulos en `price`.
    - Los valores nulos en `reviews_per_month` se imputaron con 0, asumiendo que representan propiedades sin reseñas.
  - **Tratamiento de Outliers**:
    - En `price` y `minimum_nights`, se usó el rango intercuartílico (IQR) para eliminar valores extremos.
- **Justificación de Decisiones**: Se explican las decisiones en cada etapa para asegurar un conjunto de datos limpio y representativo.

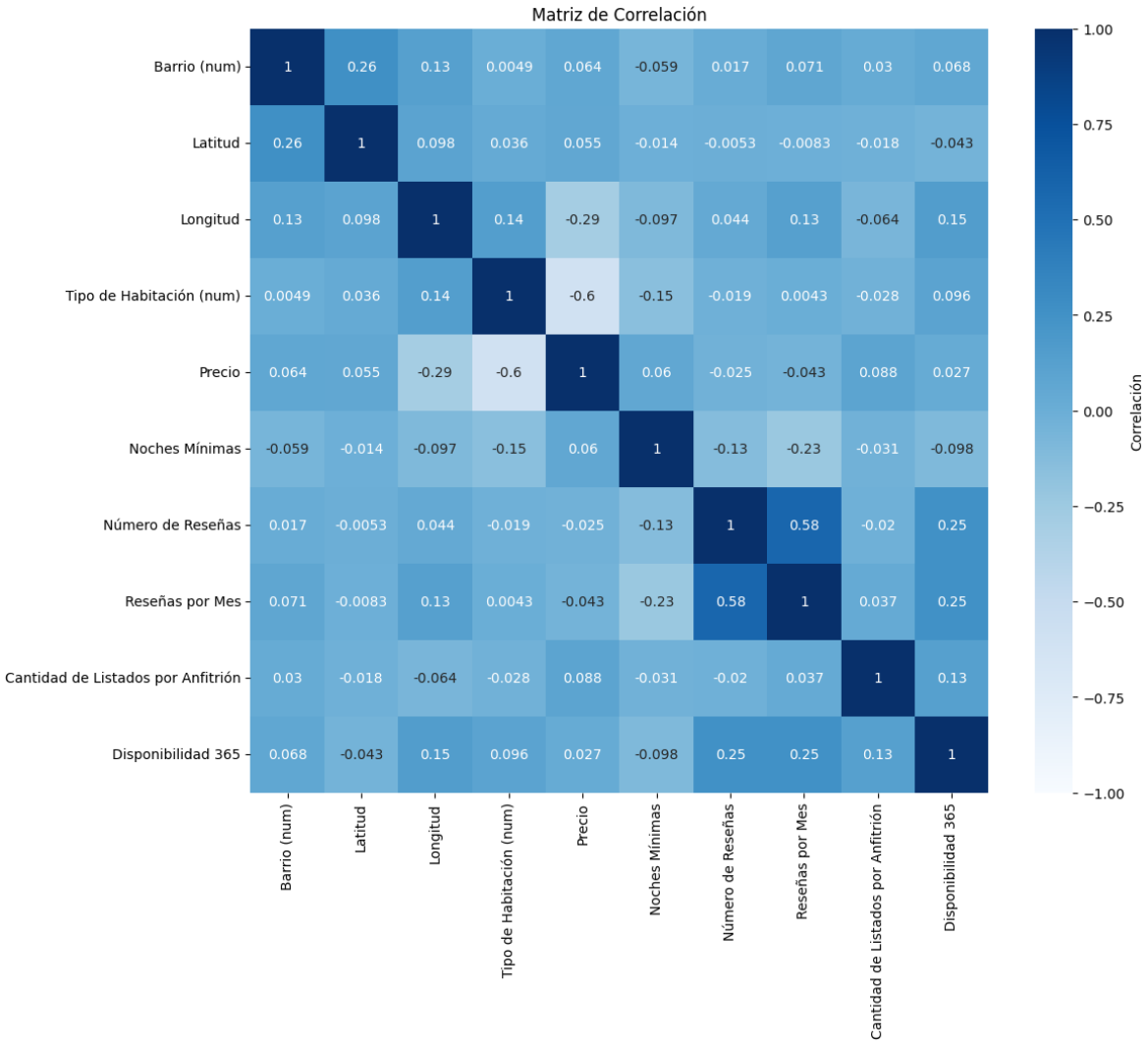
3. Transformación de Variables y Creación de `offer_group`

- **Conversión de Variables Categóricas**: Las variables `neighbourhood_group` y `room_type` se transformaron a formato numérico para facilitar el análisis.
- **Creación de la Columna `offer_group`**:
  - Se creó una columna `offer_group` que contiene la cantidad de oferentes por `neighbourhood_group` usando `groupby`.
  - Esta columna permite un análisis agregado por grupo de vecindario.

4. Análisis de Correlación

- **Matriz de Correlación**:
  - Se generó una matriz de correlación con variables seleccionadas para identificar relaciones entre ellas.

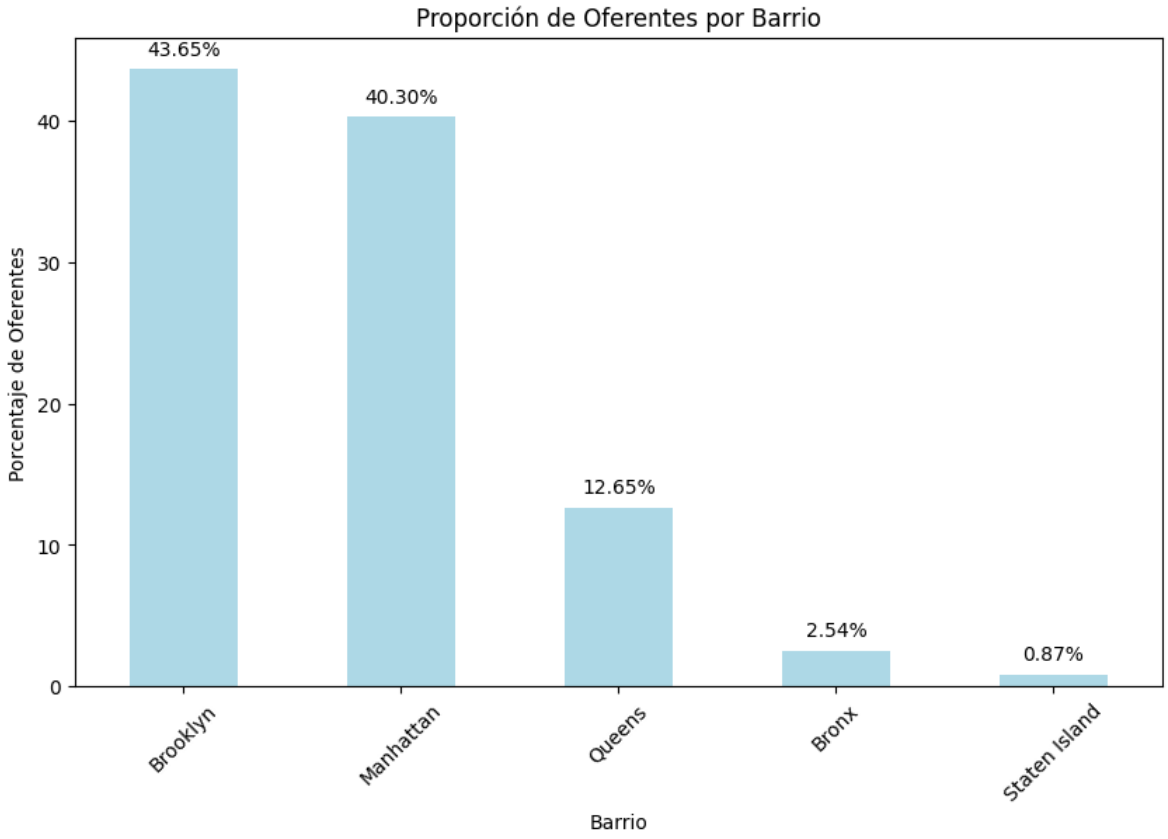
- **Interpretación de Resultados**: El análisis de la matriz de calor mostró relaciones relevantes entre variables, destacando posibles dependencias entre características como `price`, `minimum_nights`, y `room_type`.

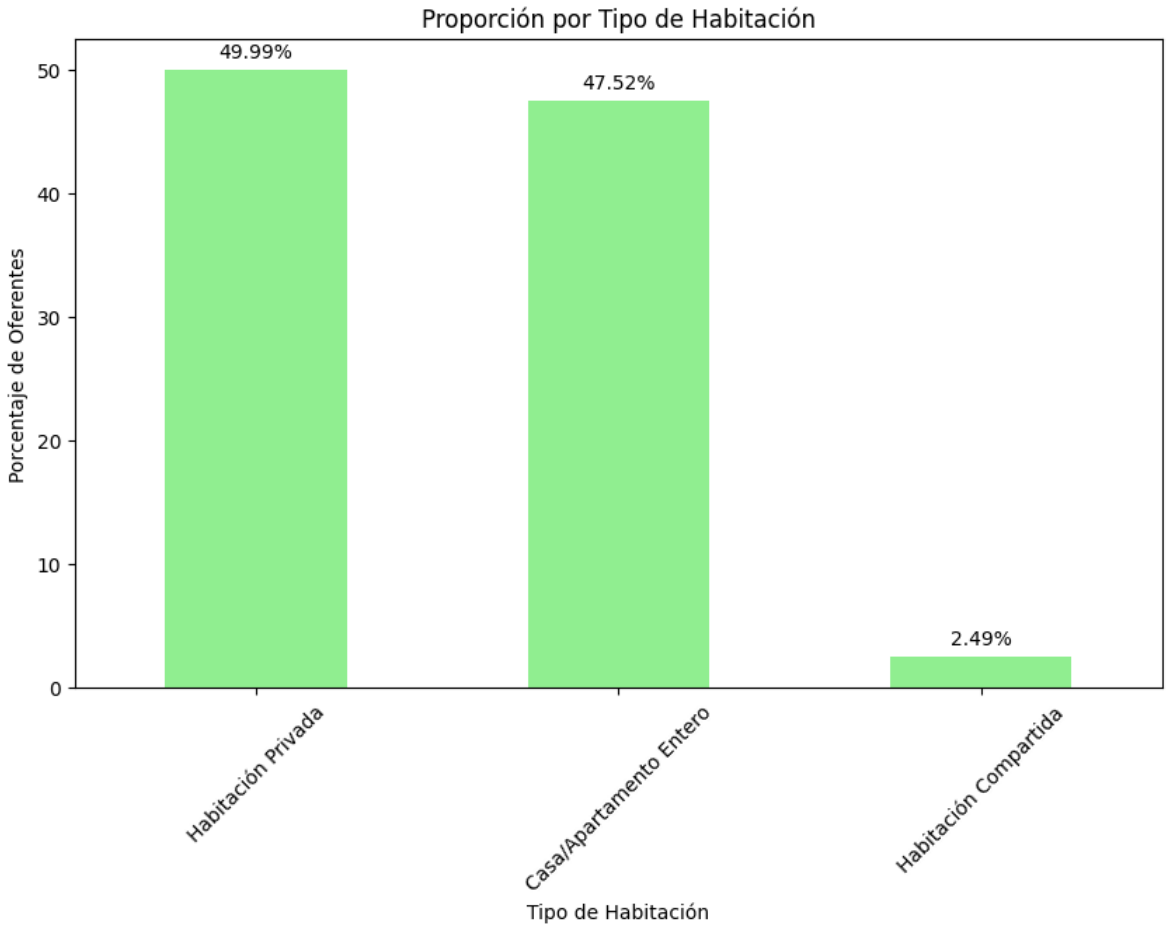


5. Análisis de Composición de Oferentes

- **Distribución por `neighbourhood_group`**:
  - Se calculó la proporción de oferentes por cada `neighbourhood_group` y se representó en un gráfico de barras.
  - **Resultados**: Brooklyn y Manhattan concentran la mayoría de los oferentes, con Queens en un distante tercer lugar.
- **Distribución por Tipo de Habitación**:
  - La proporción de tipos de habitación también se representó en un gráfico de barras.

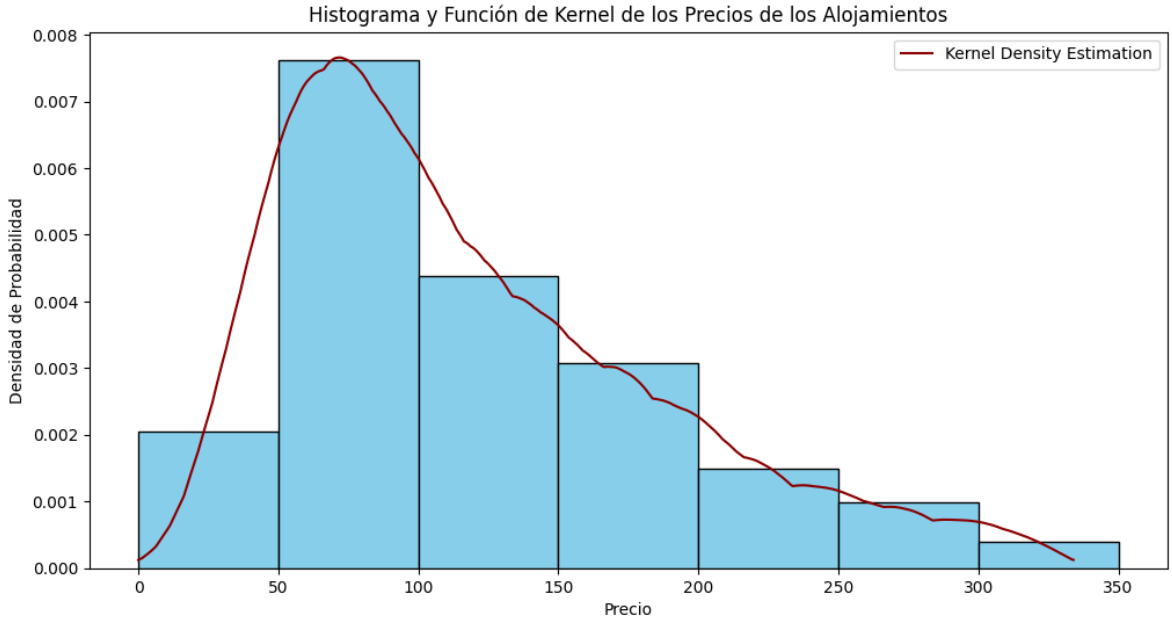
- **Resultados**: La mayoría de los alojamientos son habitaciones privadas o apartamentos enteros, con pocas habitaciones compartidas.





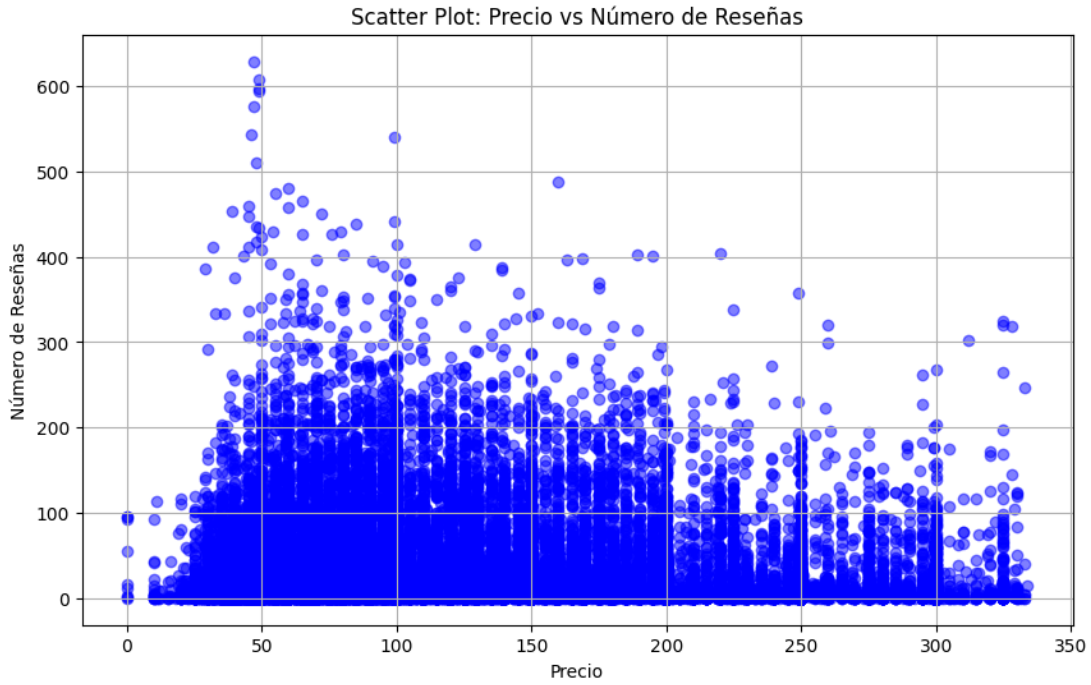
6. Análisis de Precios

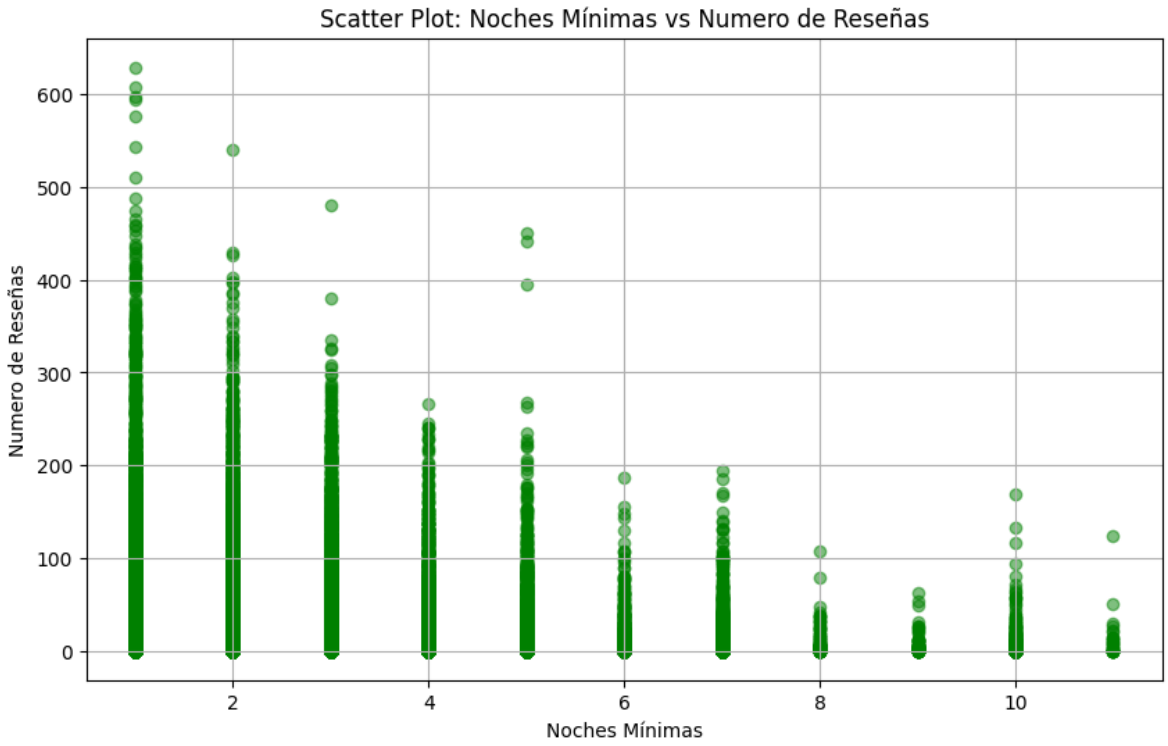
- **Histograma y Función de Densidad de Kernel**:
  - Se generó un histograma del precio de los alojamientos con una función de densidad de kernel para observar la distribución.
  - **Resultados**: La mayoría de los precios se concentran en el rango de 0 a 100, con un promedio de 119.81.
- **Resumen Estadístico**:
  - **Métricas**: Precio mínimo ($0), máximo ($334) y promedio ($119.81).
  - **Media de Precio por `neighbourhood_group` y Tipo de Habitación**:
    - Se observan diferencias en los precios medios entre barrios y tipos de habitación, con Manhattan y apartamentos enteros mostrando valores más altos.



7. Scatter Plots y Relaciones entre Variables

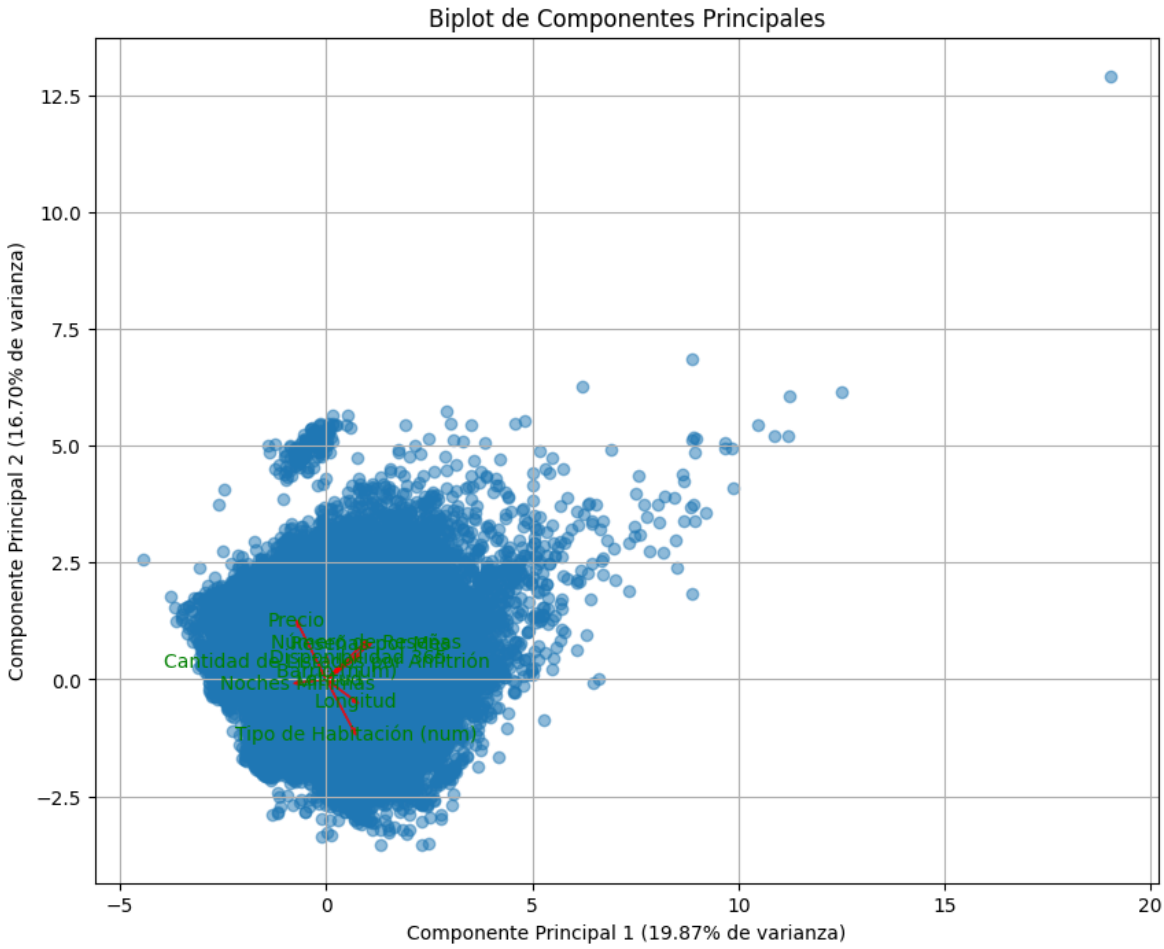
- **Scatter Plot: Precio vs. Número de Reseñas**:
  - Muestra la relación entre el precio y el número de reseñas, sugiriendo que los alojamientos con precios más bajos tienden a tener más reseñas.
- **Scatter Plot: Noches Mínimas vs. Número de Reseñas**:
  - Este gráfico sugiere que las propiedades con menos noches mínimas tienen mayor cantidad de reseñas, lo que podría indicar mayor accesibilidad.





8. Análisis de Componentes Principales (PCA)

- **Biplot de Componentes Principales**:
  - Se realizó un PCA para reducir la dimensionalidad y visualizar las variables en dos componentes.
  - **Resultados**: Los dos primeros componentes explican el 36.57% de la varianza total. Se observa que algunas variables están alineadas con el componente principal, mostrando patrones en los datos.



9. Eliminación de Variables Relacionadas con el Precio

- **Proceso**:
  - Se eliminan las variables relacionadas con el precio para preparar el modelo sin depender de estas características.
- **Verificación**: Se confirma la lista de columnas antes y después de la eliminación para asegurar que las variables de precio han sido excluidas correctamente.

10. División de Datos en Entrenamiento y Prueba

- **División de Datos**:
  - Usando `train_test_split`, se dividió la base en 70% para entrenamiento y 30% para prueba, estableciendo `price` como variable dependiente y el resto como independientes.
- **Resultado**:
  - Se muestran las dimensiones de los conjuntos de entrenamiento y prueba para verificar la correcta división de datos.

```
Dimensiones de X_train: (27797, 11)
Dimensiones de X_test: (11914, 11)
Dimensiones de y_train: (27797,)
Dimensiones de y_test: (11914,)
```

11. Regresión Lineal

- **Implementación**:
  - Se implementó una regresión lineal para predecir `price` utilizando las variables independientes en el conjunto de entrenamiento.
- **Resultados de Coeficientes**:
  - Se presenta una tabla con los coeficientes obtenidos, explicando la influencia de cada variable en la predicción del precio.

Tabla de estimaciones de los coeficientes:

	Variable	Coeficiente
0	Intercepto	-28941.641740
1	Barrio (num)	5.581201
2	Latitud	102.938761
3	Longitud	-336.712324
4	Tipo de Habitación (num)	-72.614350
5	Noches Mínimas	-1.775857
6	Número de Reseñas	-0.066078
7	Reseñas por Mes	-1.265950
8	Cantidad de Listados por Anfitrión	0.138941
9	Disponibilidad 365	0.072055

12. Evaluación del Modelo Fuera de Muestra

- **Métricas de Error**:
  - Se calcularon MSE, RMSE y MAE en los conjuntos de entrenamiento y prueba.
- **Resumen de Resultados**:
  - Los errores son similares entre ambos conjuntos, indicando que el modelo generaliza bien y no está sobreajustado.
  - El RMSE es de aproximadamente 129 unidades, lo cual es razonable en relación con el valor promedio de los precios.
- **Conclusión**: El modelo es adecuado para este conjunto de datos, mostrando predicciones consistentes dentro y fuera de la muestra.

```
Tabla de errores (entrenamiento y prueba):
  Métrica  Entrenamiento       Prueba
0     MSE    2536.371271  2565.787165
1    RMSE      50.362399    50.653600
2     MAE      37.675060    37.739036
Resumen de Resultados:
1. Errores similares en entrenamiento y prueba:
   Las métricas MSE, RMSE y MAE son muy parecidas en ambos conjuntos, lo cual es una buena señal.
   Esto indica que el modelo no está sobreajustado y funciona bien en datos nuevos.

2. Promedio de error en las predicciones:
   El RMSE, que está en la misma escala que el precio, muestra que el modelo se desvía en promedio
   alrededor de 129 unidades del valor real. Este error es razonable y estable tanto en entrenamiento
   como en prueba.

3. Consistencia del modelo:
   Los errores pequeños y consistentes entre entrenamiento y prueba indican que el modelo generaliza bien,
   es decir, que puede hacer predicciones precisas en datos que no ha visto antes.

En resumen: El modelo de regresión lineal es adecuado para este conjunto de datos.
Sus predicciones en datos nuevos son consistentes con las de entrenamiento, lo que indica que es confiable
para hacer predicciones fuera de la muestra.
```

13. Conclusión

- **Resumen de Hallazgos**: Los análisis realizados proporcionan una visión integral de la estructura de precios y tipos de alojamientos en Nueva York.
- **Reflexión sobre el Modelo**: La regresión lineal demuestra ser útil para el análisis de precios, y los valores de error indican que las predicciones son confiables.
- **Posibles Aplicaciones**: Este tipo de análisis puede aplicarse en estudios de mercado para optimizar precios y estrategias de alquiler en plataformas de alojamiento.