Ciencia de Datos

Trabajo Práctico Integral

Grupo: nº10

1 Benjamin, Ian Nicolas, 80738

2 Pedernera, Nicolas, 83201

3 Ricse, Javier, 80122

4 Charra Marquez, Giuliano, 63162

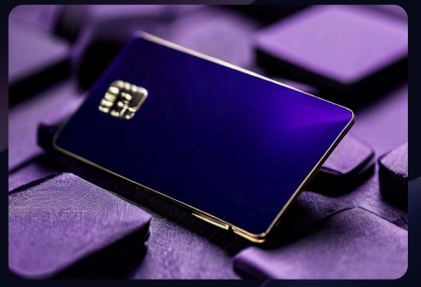
Trabajo Práctico Integral: Análisis de Datos de Airbnb en Nueva York

En este trabajo práctico integral, nuestro grupo analizará a fondo un dataset de Airbnb en Nueva York, explorando técnicas de Ciencia de Datos para obtener valiosos insights y realizar predicciones. Será un recorrido completo, desde la selección del dataset hasta la presentación final de nuestros hallazgos.

Selección del Dataset

Dataset de Cosecha de Vinos

Un conjunto de datos clásico sobre las características de diferentes tipos de vino. Interesante para modelar la calidad del vino, pero no tenia las suficientes variables para analizar.



Dataset de Fraudes de Tarjetas de Crédito

Un dataset sobre transacciones fraudulentas con tarjetas de crédito. Muy interesante para aplicar técnicas de detección de anomalías, pero no queríamos predecir un booleano.

Dataset de Airbnb en Nueva York (Seleccionado)

Este dataset contiene información detallada sobre anuncios de Airbnb en la ciudad de Nueva York. Es ideal para analizar tendencias de la industria de alojamiento.

Dataset de Airbnb a Nivel Mundial

Este dataset abarca información de anuncios de Airbnb a nivel mundial. Permitirá analizar tendencias globales en la industria de alojamiento, pero hay muchos lugares que carecen de datos.

Dataset de Airbnb en Nueva York

102599 Filas y 26 Columnas

Numero de Revisiones

ID del anuncio

Latitud y Longitud

Barrio

Disponibilidad 365 días

Año de construcción

Noches mínimas

Tipo de habitación

Reglas de la casa

Licencia

Host verificado

Precio

Proceso ETL

1

Extracción

Recopilamos el dataset de Airbnb en Nueva York a partir de Kaggle, asegurándonos de contar con una versión completa y actualizada.

2

Transformación

Realizamos una limpieza exhaustiva de los datos, manejando valores faltantes, corrigiendo formatos y eliminando inconsistencias.

3

Carga

Estructuramos el dataset de forma organizada y lo cargamos en un nuevo dataset para facilitar el análisis posterior.

Casos Problemáticos

Valores Faltantes

Insertamos el valor (moda y mediana) para completar estos datos.

Valores muy altos en 0

23500 valores de disponibilidad 365 son 0.

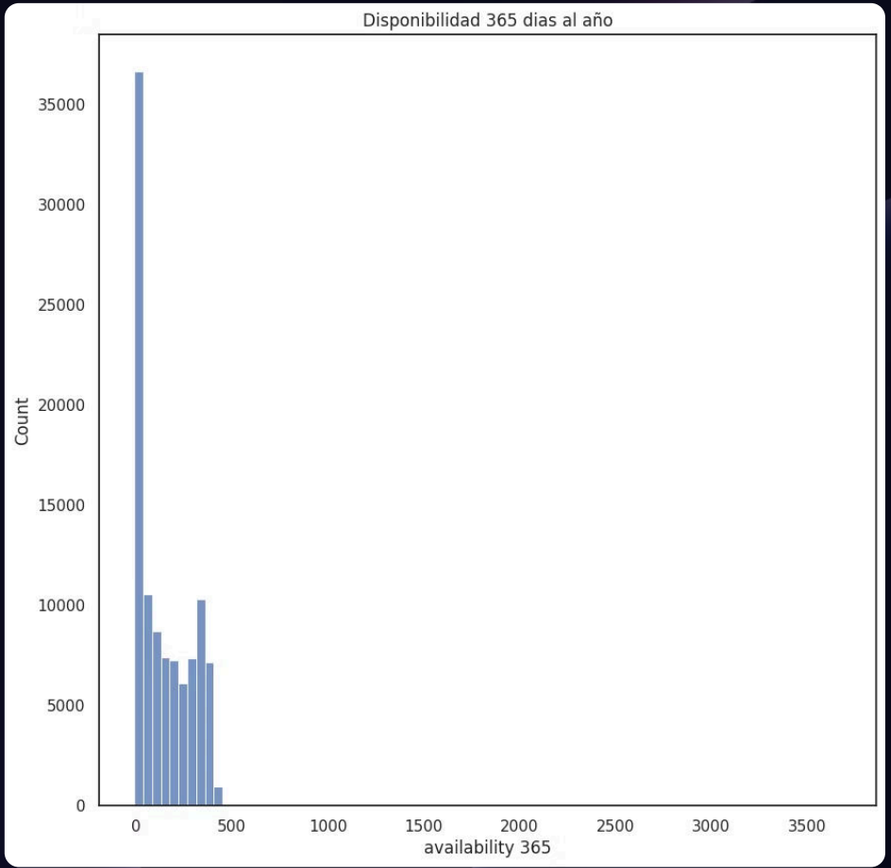
Valores Inconsistentes

Valores fuera del rango aceptado Noches Mínimas y Disponibilidad 365.

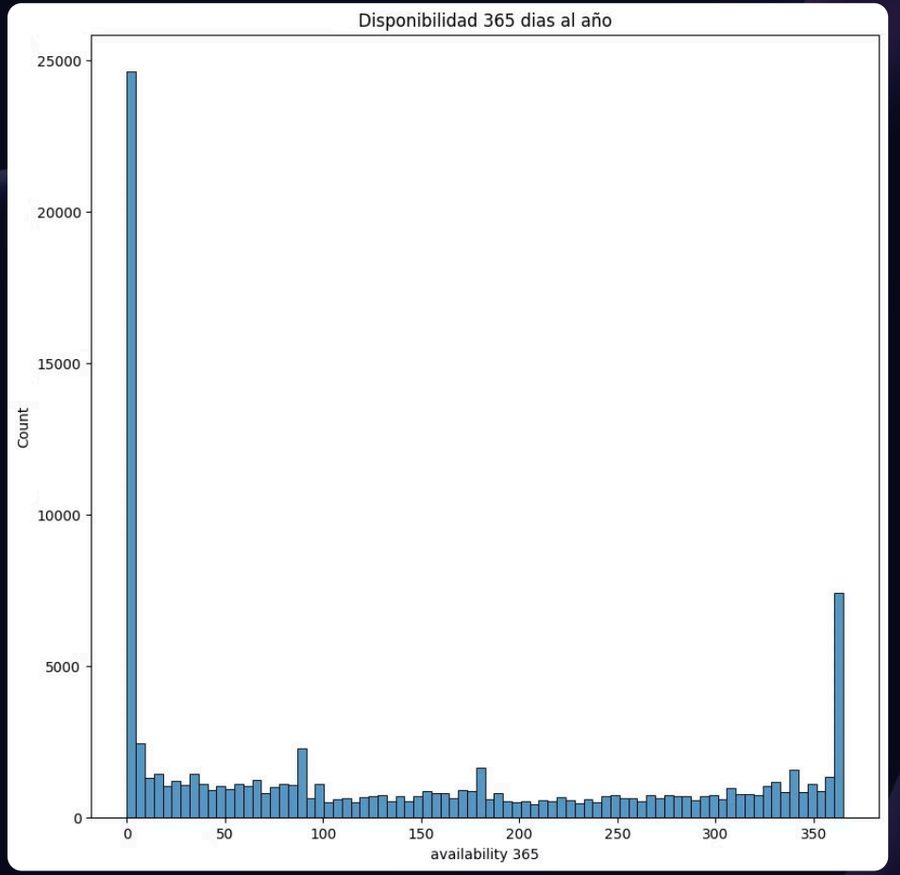
Categorías repetidas

Datos escritos incorrectamente.

Análisis de Variables

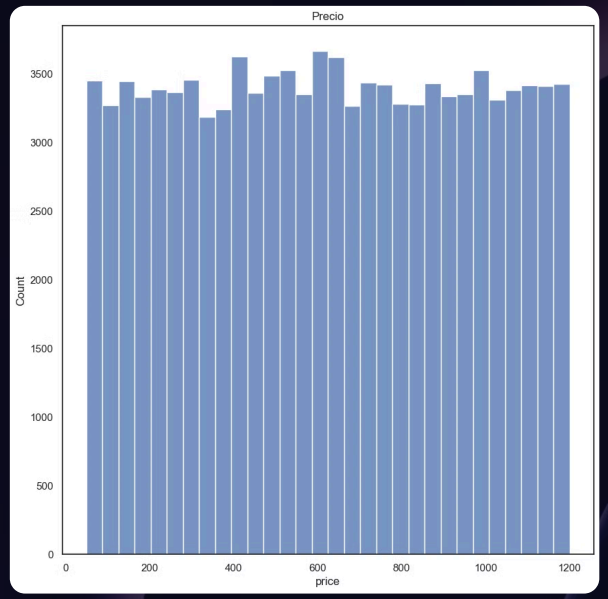


Se puede apreciar que la variable disponibilidad 365 días supera los valores limites y también contiene la mayoría de los valores en cero



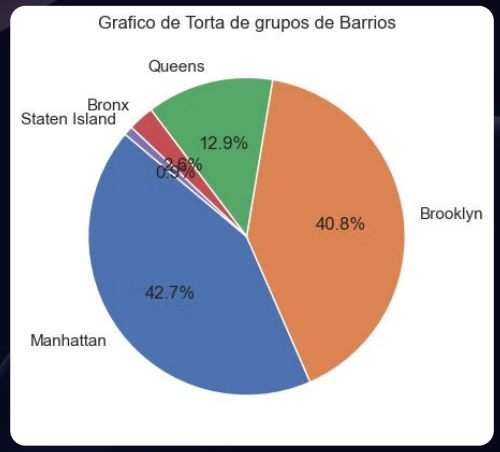
Luego de la limpieza podemos apreciar mejor el grafico

Análisis de Variables



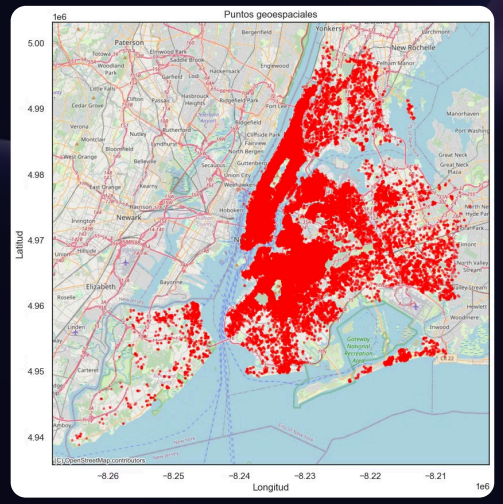
Variables Numéricas

Analizamos la distribución y tendencias de variables como precio, calificación, número de comentarios y número de huéspedes.



Variables Categóricas

Estudiamos la composición y frecuencia de variables como tipo de propiedad, estilo de listado y ubicación.



Variables Espaciales

Analizamos la evolución de los precios y ocupación a lo largo del tiempo, identificando patrones estacionales.

Selección de Variables

Variables Clave

Identificamos las variables más predictivas para el precio , como barrio y grupo de barrio, Tipo de Alquiler, tipo de cuarto, numero de reviews, etc.

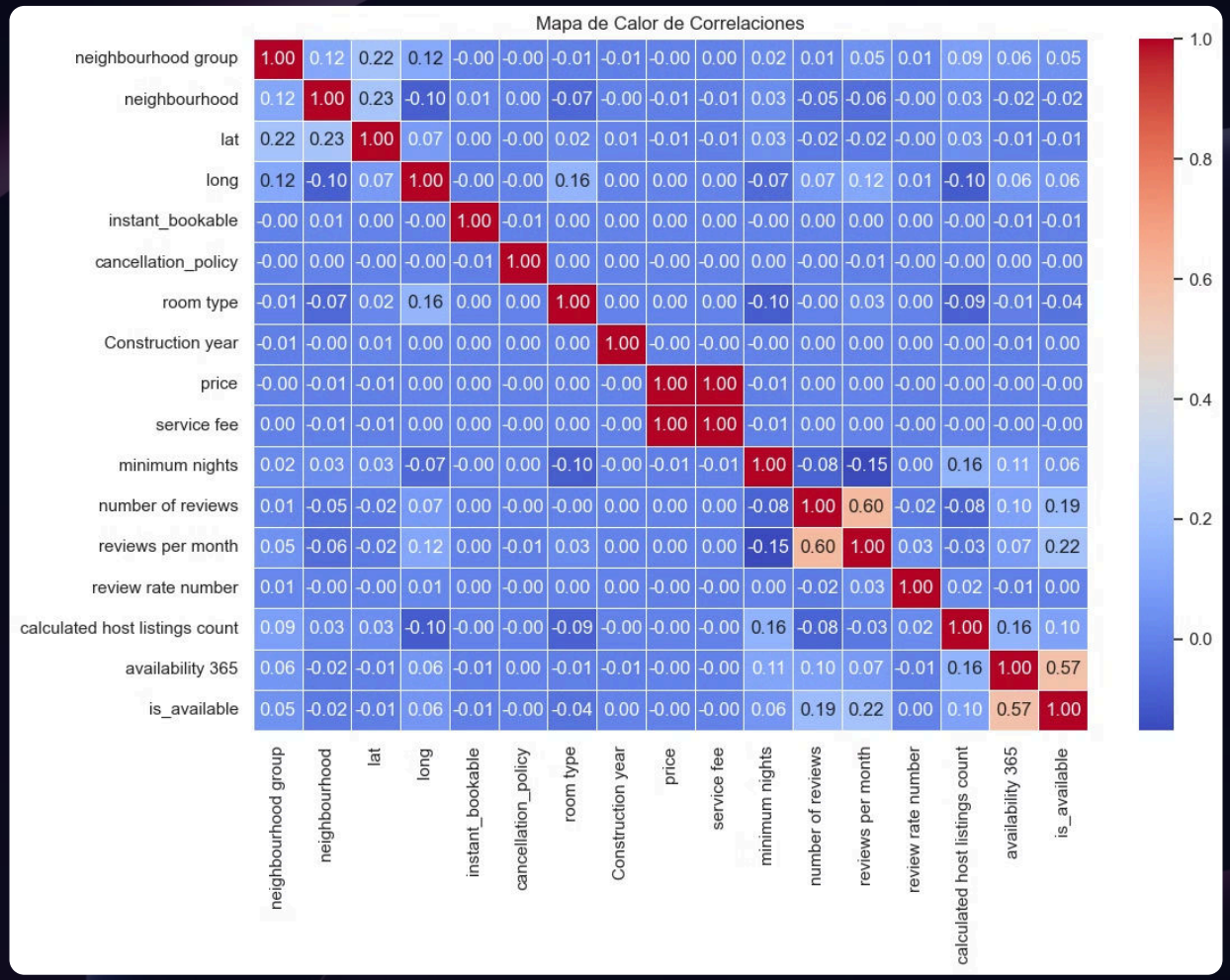
Variables Complementarias

Seleccionamos variables adicionales que podrían aportar información valiosa, como reseñas, disponibilidad al año y noches minimas.

Variables a Descartar

Variables con poca relevancia como el año de construcción, reserva instantánea

Mapa de Calor de Correlaciones



Análisis de Modelos

1

Métricas de Desempeño

Evaluamos los modelos utilizando métricas como R-cuadrado y error cuadrático medio.

2

Importancia de Variables

Analizamos qué variables tienen mayor impacto en las predicciones de cada modelo.

3

Ajuste de Hiperparámetros

Optimizamos los hiperparámetros de los modelos para mejorar su rendimiento.

Modelos Utilizados

Regresión Lineal

- **Error cuadrático medio**: 109526.6
- **R-cuadrado:** -0.000201

Random Forest

- **Error cuadrático medio:** 70733.8
- **R-cuadrado**: 0.35

SVR

- **Error cuadrático medio:** 110325.717758978
- **R-cuadrado:** -0.0075

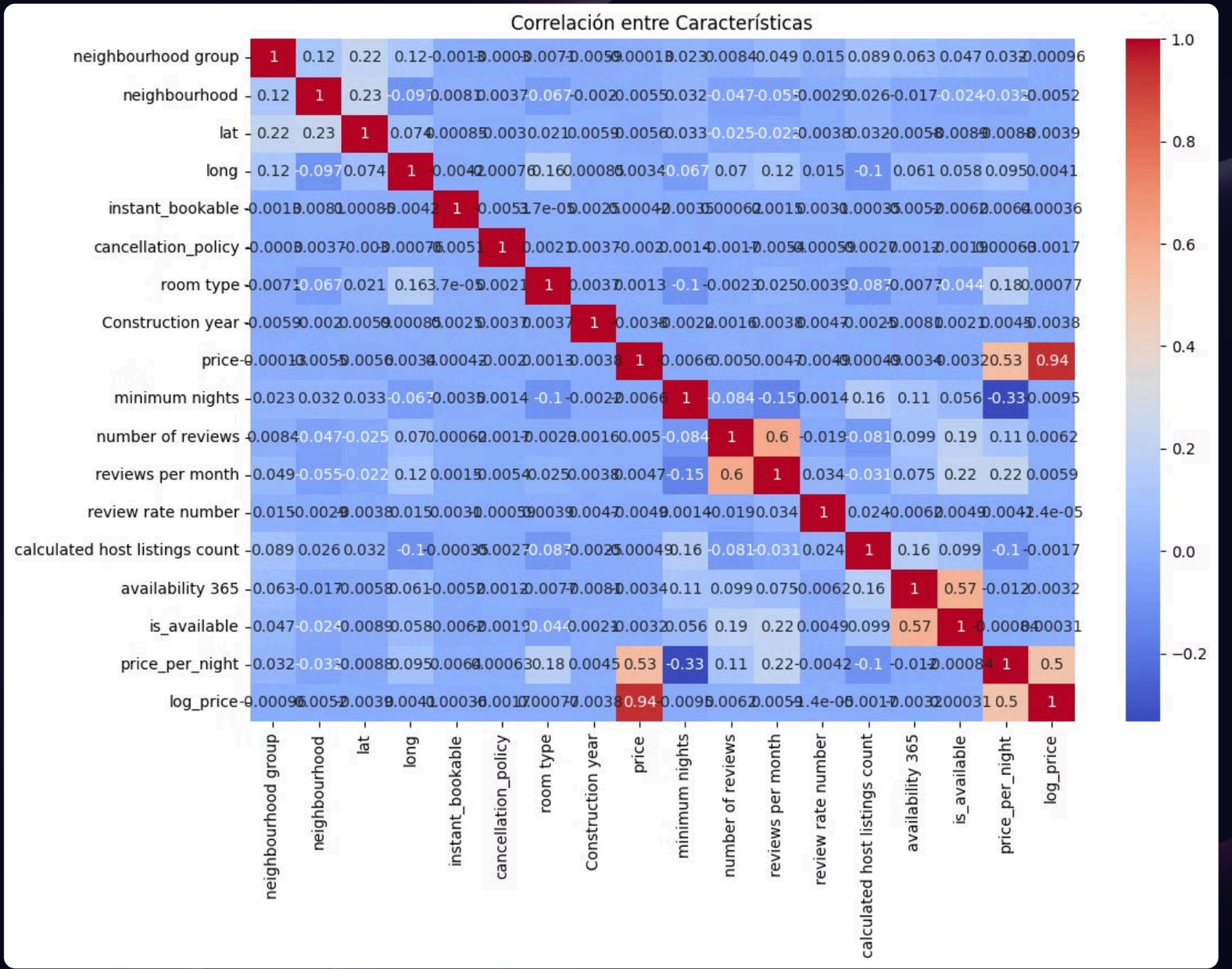
Redes Neuronales

- **Error cuadrático medio**: 109527.6
- **R-cuadrado:** -0.0002

Acciones a mejorar el modelo

- Eliminar columnas no relevantes(Barrios, grupos de barrios, política de cancelación, reservación en el momento y año de construcción).
- Eliminar sobreajuste.
- Realizar validación cruzada y ajuste de hiperparámetros.
- Realizar transformación logarítmica de columna precio.
- Crear nueva variable para análisis Precio Por Noche.
- Cantidad final de columnas: 13.

Mapa de Correlación Final



Resultados Obtenidos

Columna Precio Logarítmico

- **Error cuadrático medio**: 133.3
- **R-cuadrado**: 0.998

Columna Precio por Noche

- **Error cuadrático medio:** 14750.7
- **R-cuadrado**: 0.72

Análisis de Resultados y Conclusiones

1. Elección, Limpieza y Preparación del Dataset
2. Recomendaciones sobre propiedades
3. Conclusiones

Muchas Gracias !!