

## Informe Parcial 1

Presentado por David Polanía Mejía

[d.polaniam@uniandes.edu.co](mailto:d.polaniam@uniandes.edu.co)

202213328

1. análisis exploratorio y de calidad de datos a partir de los datos históricos suministrados.

### Eliminación de variables

Se eliminaron las siguientes variables que no aportaban valor al modelo, que generaban ruido o eran constantes.

- Id
- review rate number (Se elimino esta variable porque para el entrenamiento del modelo se requiere que no tenga relación con la variable objetivo.
- Country – Es una variable con alta cardinalidad y el valor de latitud y longitud pueden funcionar mejor.
- service fee
- neighbourhood

### Eliminación de valores nulos y/o repetidos

Se eliminaron los valores nulos o repetidos

### Limpieza de datos

- Price: Se eliminaron los caracteres especiales '\$' y se convirtió en una variable numérica.
- Price: Se filtraron los valores para que sean positivos y menores que 1000 ya que en este contexto no tiene sentido un número negativo y se eliminó el valor extremo superior a 1000.
- Construction year: Se determinó que se va a trabajar con los inmuebles con registros posteriores al año 2000
- Minimum nights: se entendió que este debe ser un número superior que 0, por lo cual se realizó este filtro.
- availability 365: Sólo se aceptaron números positivos

### Transformación de los datos

- One hot encoding: Se realizó una transformación one hot encoding para neighbourhood group y otras variables que fueran categóricas

## 2. Resumen de entrenamiento y modelo seleccionado.

- Min max: Se realizó una estandarización min max.

```
**Datos sin escalar**
```

```
Linear regression: 3443.7198533786454
```

```
**Datos escalados**
```

```
Linear regression: 0.034054426776815354
```

```
**Modelo Ridge:**
```

- **\*\*Alpha 0.1: 0.03404825426394409\*\***
- Alpha 0.2: 0.034057271504894523
- Alpha 0.9: 0.034184699132951926

```
**Modelo Lasso:**
```

- Alpha 0.1: 0.0371441038906897
- Alpha 0.2: 0.0371441038906897
- Alpha 0.9: 0.0371441038906897

- El modelo seleccionado es Modelo Ridge usando hiperparámetro alpha 0.1 ya que tuvo el mejor rendimiento.

## 3. definición del presupuesto de marketing que debe aprovisionar para promocionar los inmuebles menos populares.

Se determinó como inmueble no popular un valor de Predicted Number Of Reviews inferior al valor de la media que es 0.08 para este caso.

Finalmente, después de realizar los cálculos apropiados, se determina que el presupuesto que se debe establecer para el marketing para los inmuebles inpopulares es 1507.61 dolares.