Proyecto Final - KeepCoding Mujeres en Tech II

Table of contents

Data Set

Arquitectura y validación de los datos

a. Muestreo y exploración inicial de los datos

```
[1] "El dataset tiene: 21278 filas y  17 columnas."
```

```
[1] "Las variables que tenemos son: "
```

```
 [1] "Room ID"                    "Name"
 [3] "Host ID"                    "Neighbourhood"
 [5] "Room type"                  "Room Price"
 [7] "Minimum nights"             "Number of reviews"
 [9] "Date last review"           "Number of reviews per month"
[11] "Rooms rent by the host"     "Availibility"
[13] "Updated Date"               "City"
[15] "Country"                    "Coordinates"
[17] "Location"
```

Para mayor facilidad en el manejo de los datos reemplazamos los espacios en los nombres de las columnas por guiones bajos "_" y se cambian los nombres por minúsculas.

```
colnames(data) = colnames(data) %>%  str_replace_all(' ', '_') %>% tolower()
```

Hacemos un análisis inicial de los datos, en este encontramos:

- La variable room_price tiene como valor máximo 9999, este valor se puede considerar como outlier pero usualmente se usa para codificar valores no válidos, por lo que se tendrá en cuenta para eliminar y no ser considerado como valor máximo observado.
- La variable number_of_reviews_per_month es numérica pero se ha cargado como character, tenemos que transformarla a numérica.
- Aunque las columnas room_id y host_id se cargan como variables cuantitativas, por sus características debemos transformarlas en variables de tipo texto.
- La variable minimum_nights tiene un valor máximo de 1125, tendremos que analizarlo para saber si se considera un outlier o si se puede corresponder con un valor anómalo.
- La variable coordinates es una variable que contiene la latitud y longitud en una misma cadena de texto, vamos a separar separar la variable usando la "," como separador y lo vamos a transformar a variable numérica.
- Las variables date_last_review y number_of_reviews_per_month tienen NA's, en el caso de number_of_reviews_per_month al ser cuantitativa se puede considerar imputar con un 0, ya que la interpretación de la variable no se vería afectada.

```
summary(data)
```

```
   room_id             name              host_id          neighbourhood
 Min.   :    6369   Length:21278       Min.   :     5154   Length:21278
 1st Qu.:18047743   Class :character   1st Qu.: 25506202   Class :character
 Median :28823834   Mode  :character   Median : 95416752   Mode  :character
 Mean   :26951219                      Mean   :120527709
 3rd Qu.:37708095                      3rd Qu.:208979694
 Max.   :44274350                      Max.   :356881304

 room_type           room_price     minimum_nights     number_of_reviews
 Length:21278       Min.   :   0.0   Min.   :   1.000   Min.   :  0.00
 Class :character   1st Qu.:  36.0   1st Qu.:   1.000   1st Qu.:  0.00
 Mode  :character   Median :  60.0   Median :   2.000   Median :  6.00
                    Mean   : 163.9   Mean   :   5.522   Mean   : 34.85
                    3rd Qu.: 100.0   3rd Qu.:   3.000   3rd Qu.: 38.00
```

```
                  Max.   :9999.0   Max.   :1125.000   Max.   :661.00

 date_last_review     number_of_reviews_per_month rooms_rent_by_the_host
 Min.   :2012-08-04   Length:21278                Min.   :  1.00
 1st Qu.:2019-10-06   Class :character            1st Qu.:  1.00
 Median :2020-02-23   Mode  :character            Median :  2.00
 Mean   :2019-10-20                               Mean   : 14.26
 3rd Qu.:2020-03-12                               3rd Qu.:  7.00
 Max.   :2020-07-28                               Max.   :244.00
 NA's   :5413
  availibility    updated_date           city              country
 Min.   :  0.0   Min.   :2020-06-11   Length:21278       Length:21278
 1st Qu.:  0.0   1st Qu.:2020-07-17   Class :character   Class :character
 Median :132.0   Median :2020-07-17   Mode  :character   Mode  :character
 Mean   :158.9   Mean   :2020-07-16
 3rd Qu.:335.0   3rd Qu.:2020-07-17
 Max.   :365.0   Max.   :2020-07-17

 coordinates          location
 Length:21278       Length:21278
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Aplicamos los cambios que hemos observado en el análisis inicial antes de realizar el análisis de las variables cuantitativas.

```
data[,c("room_id", "host_id")] %<>%  toString()
data$number_of_reviews_per_month %<>%  as.numeric()
data$lat <- data$coordinates %>% str_split_i(', ', 1) %>% as.numeric()
data$long <- data$coordinates %>% str_split_i(', ', 2) %>%  as.numeric()
```

Analizamos cuántos valores distintos tienen las variables cualitativas:

```
data %>% select_if(where(is.character)) %>% map_dfc(n_distinct)
```

```
# A tibble: 1 x 9
  room_id  name host_id neighbourhood room_type  city country coordinates
    <int> <int>   <int>         <int>     <int> <int>   <int>       <int>
```

```
1       1 20410       1           142         4    13       8       21278
# i 1 more variable: location <int>
```

Las variables City y Country deberían tener un único valor, ya que solo nos centramos en una ciudad y en un país.

```
print('Ciudades únicas:')
```

```
[1] "Ciudades únicas:"
```

```
data$city %>% unique()
```

```
 [1] "Madrid"        "Sevilla"       "Girona"        "Barcelona"
 [5] "Lisbon"        "Florence"      "Istanbul"      "San-francisco"
 [9] "Brussels"      "London"        "Sydney"        "Mallorca"
[13] "Sicily"
```

```
print('Países únicas:')
```

```
[1] "Países únicas:"
```

```
data$country %>% unique()
```

```
[1] "Spain"          "Portugal"       "Italy"          "Turkey"
[5] "United states"  "Belgium"        "United kingdom" "Australia"
```

```
data = data[data$city=="Madrid",]
```

```
data %>% select_if(where(is.character)) %>% map_dfc(n_distinct)
```

```
# A tibble: 1 x 9
  room_id  name host_id neighbourhood room_type  city country coordinates
    <int> <int>   <int>         <int>     <int> <int>   <int>       <int>
1       1 20391       1           127         4     1       1       21255
# i 1 more variable: location <int>
```

b. Definir e implementar el Datawarehouse

```
data[is.na(data$number_of_reviews_per_month), 'number_of_reviews_per_month'] = 0
```

c. (Opcional) Ingesta de datos (ETL) y validación de que se ha cargado correctamente

Análisis Exploratorio

Hacer un estudio estadístico con R o Python, según preferencia personal, y averiguar cuales son las métricas adecuadas para el dataset. No olvidemos: a. Revisión de la calidad de los datos b. Detección outliers (rango de variables), imputación valores nulos. c. Boxplots, histogramas, etc. d. Normalización de los valores de las tablas (quitar tildes, "dobles espacios", etc.)

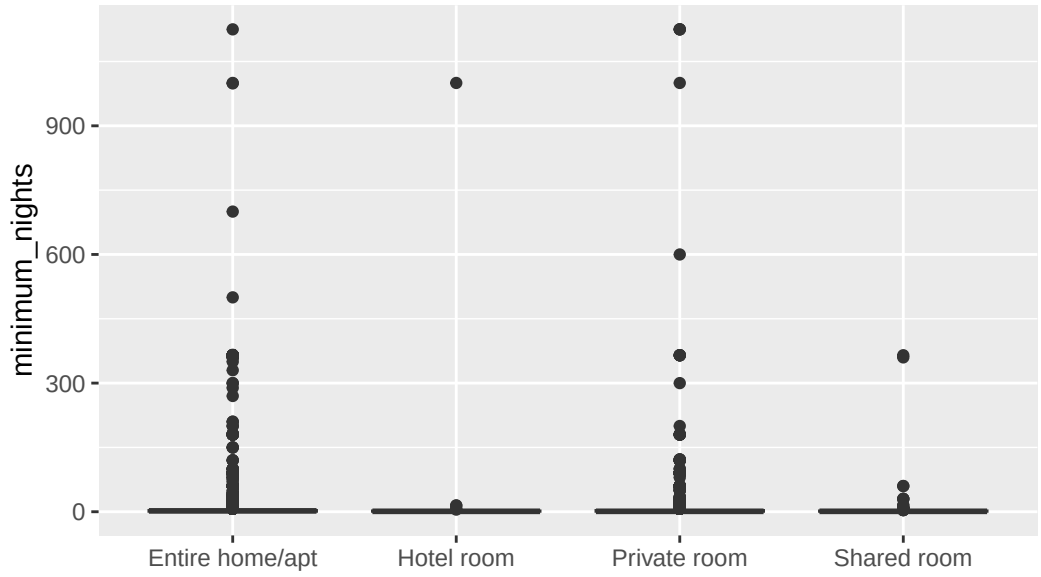
```
data %>% select_if(where(is.numeric)) %>% summary()
```

```
  room_price   minimum_nights     number_of_reviews
Min.   :   0   Min.   :   1.000   Min.   :  0.00
1st Qu.:  36   1st Qu.:   1.000   1st Qu.:  0.00
Median :  60   Median :   2.000   Median :  6.00
Mean   : 164   Mean   :   5.522   Mean   : 34.88
3rd Qu.: 100   3rd Qu.:   3.000   3rd Qu.: 38.00
Max.   :9999   Max.   :1125.000   Max.   :661.00
number_of_reviews_per_month rooms_rent_by_the_host  availibility
Min.   : 0.000              Min.   :  1.00         Min.   :  0.0
1st Qu.: 0.000              1st Qu.:  1.00         1st Qu.:  0.0
Median : 0.390              Median :  2.00         Median :133.0
Mean   : 1.065              Mean   : 14.25         Mean   :158.9
3rd Qu.: 1.540              3rd Qu.:  7.00         3rd Qu.:335.0
Max.   :27.250              Max.   :244.00         Max.   :365.0
     lat             long
Min.   :40.33   Min.   :-3.864
1st Qu.:40.41   1st Qu.:-3.708
Median :40.42   Median :-3.701
Mean   :40.42   Mean   :-3.694
3rd Qu.:40.43   3rd Qu.:-3.687
Max.   :40.56   Max.   :-3.524
```

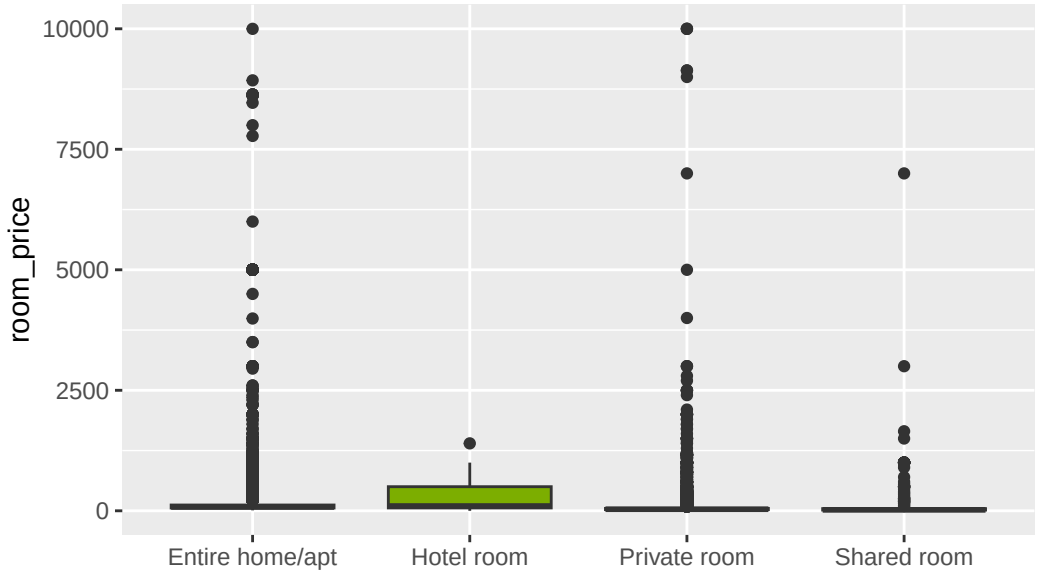
En las variables room_price y minimum_nights tiene un máximo muy lejano al tercer cuartil, vamos a analizar si estos valores se pueden considerar válidos o no.

```
data %>% select ('room_type','room_price', 'minimum_nights') %>%
  ggplot( aes(x=room_type, y=minimum_nights, fill=room_type)) +
```

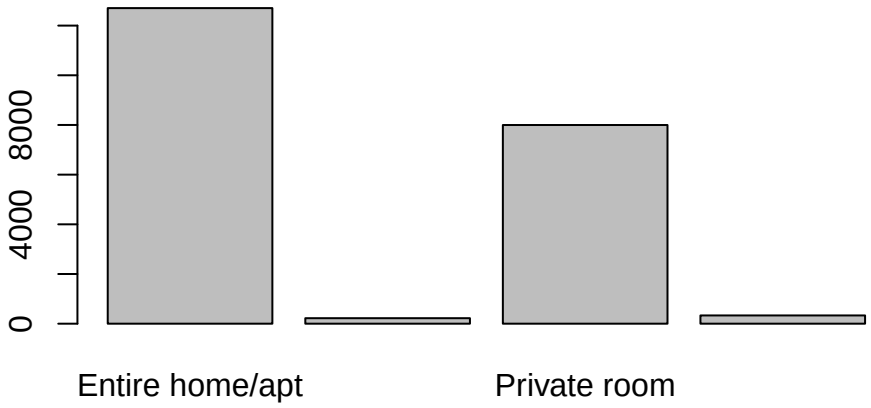
```
    geom_boxplot() +
 #   scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
  #  theme_ipsum() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("") +
    xlab("")
```




```
data %>% select ('room_type','room_price', 'minimum_nights') %>%
  ggplot( aes(x=room_type, y=room_price, fill=room_type)) +
    geom_boxplot() +
 #   scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
  #  theme_ipsum() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("") +
    xlab("")
```




```
data$room_type %>% table() %>% barplot()
```



Visualización de las métricas

A partir de los datos de Airbnb, obtén los KPIs que puedan ser de relevancia y contesta a través de un dashboard a una pregunta relevante que hagas sobre los datos. a. a. Se valorará el diseño final del dashboard. b. b. El uso de buenas prácticas. c. c. El cálculo de KPIs adecuados y el uso de campos calculados avanzados. d. d. El uso de vistas interactivas. Nota: En este ejercicio no habrá un dashboard modelo sino que se basará en valorar vuestras capacidades con el uso de la herramienta de Tableau y que podáis demostrar todo lo aprendido durante este bloque

Pre-procesamiento y Modelado

La tarea asignada es hacer un algoritmo de regresión lineal que prediga el precio de un inmueble en función de las características que elijáis

Informe

En esta etapa se debe de simular la presentación de resultados en un entorno real de empresa. Suposiciones iniciales Cuales han demostrado ser válidas y cuáles no. ¿Por qué? Métricas seleccionadas: ¿han sido las correctas o no? ¿por qué?