Data preprocessing

Reading the Airport codes data

I splitted the coordinates column into Longitude and coordinates

Reading the Flights data

The flights dataset contains 1915886 rows and 16 columns. Most of the columns have the wrong data type (`object`) as shown above. Some of the columns such as columns 3, 13 and 14 have a mixed type as indicated by the warning when loading the CSV into a dataframe. We are going to use the `sanitize_df_columns` function from the `utils` module to clean these columns.

We observe that there are missing values in some column of the flights dataset. Let get a look at the missing values and deal with them.

This dataset has a lot of missing values. The most important columns are `OP_CARRIER_FL_NUM` which have 40 missing values, `DEP_DELAY` which have 50351 missing values, `ARR_DELAY` which have 55991 missing values and `OCCUPANCY_RATE` which have 310 missing values and `DISTANCE` which have 1,913,806 missing values. These are the most important columns because they are used in the calculation of business metrics(busiest routes, most profitable routes, most popular routes, etc.).

There are 2 options when dealing with these missing values, we either drop the missing values or replace them with an inputed value which could be the mean, median, mode of the other values or even with zero.

- The `OP_CARRIER_FL_NUM` column has 40 missing values and this column does not have any specific meaning. We will drop these values.
- For the `ARR_DELAY` column and `DEP_DELAY` column, we can safely assume that the missing values are zero. because these values represent (`55,991/1,915,886 = 2.93%`) and (`50,351/1,915,886 = 2.63%`) of the total dataset, so the absence must mean that there was no delay.
- For the `DISTANCE` and `OCCUPANCY_RATE` columns, we will inpute the missing values with the median of the other values because any completed flights will have some distance between the origin and the destination and a non-zero occupancy rate. The median was chosen instead of the mean because the median is more robust to outliers.

The flights dataset does not have any missing values now except for `AIR_TIME` which is not relevant for the purpose of our analysis. We can move on with loading and cleaning other datasets

Reading the tickets data

The `INTIN_FARE` column has the wrong data type (`Object`) because the column includes some characters that are not numbers. We are going to use the `sanitize_df_columns` function from the `utils` module to clean this `tickets` data as well.

After the clean up all columns have the relevant data types.