**Flights delays and weather conditions at “Newark”, “JFK” and “La Guardia” airports**

**Context**

In this project I focused on the analysis of the relationship between the length of flights delays and weather conditions in 2017 for flights departing from the New York airports of “Newark Liberty International Airport”, “La Guardia Airport” and “JFK International Airport”.

The business questions I decided to answer to are the following:

* Does the weather effectively influence the flights departure delays and which weather condition impact mostly?
* Do the different airlines or the specific departure airports have a role to play together with the weather conditions with the length of departure delays?

**Business intelligence and data-driven decision making**

*What insights can the business/organisation gain from your analysis and how will your analysis help the business/organisation make better decisions?*

The first answer that could be useful for the Airports business is if or not actually weather has an impact on the length of the delays. In fact, if the answer is yes, then a second step would be wondering why the weather has such an influence on those delays. Obviously, it is not realistic that Airports ban the flights when certain weather conditions happens (with very extreme exceptions of course), so it would be useful to analyse if this influence is mostly registered for specific airlines of for specific airports for example in order to make the appropriate activities in order to reduce the passengers’ discomfort due to long delays.

**Domain knowledge and the business context**

*Briefly describe the business/organisation and where your analysis fits within its aims/activity.*

For this project, I assumed to have been hired form Newark Airport with the aim of investigating the length of flights delays and weather conditions for flights departing from three New York Airports: “Newark Liberty International Airport”, “La Guardia Airport” and “JFK International Airport”.

Having an idea of whether or not the weather influence departures delays in the three airports can be useful for the business to find out which interventions it would be wise to adopt in order to reduce the delays and all the negative implications that these delays could imply for these three New York airports in terms of credit, efficiency and cost savings.

In particular, knowing if the weather has an higher impact in specific airports or for specific airlines, can be extremely useful in order to apply more focused solutions.

**Data**

**Internal and external data sources**

*Briefly describe your data sources and which were internal (provided by the organisation) and external (not sourced from the organisation’s data)*

I used three main data sources:

* `flights` - Internal – This dataset contains information about flights departing from one of the following three New York airports: “Newark Liberty International Airport”, “La Guardia Airport” and “JFK International Airport” in 2017. Specifically, in addition to identifiers like their number, and talinum, also data about their starting and ending points departure and arrival times and delays, duration and length of the flights. It is a dataset provided by the internal data of the New York airports.
* `airlines` - Public - This is a public dataset, which contains the airlines’ names and their shortcuts.
* `weather` - External – This is an external dataset containing information about specific weather features recordings for the three airports of interest in 2017.

**Types of data**

*What kind of data did you work with? E.g. categorical and numerical data and their sub-types.*

**Data formats**

*What format did your data come in? E.g. all downloaded flat files (CSV) or any data from APIs, scraping etc.*

**Data quality and bias**

*Briefly describe the quality of the data and whether you have any reasons to suggest the data is biased e.g. only data from a specific demographic even though a broader demographic would be of interest to the organisation.*

**Ethics**

**Ethical issues in data sourcing and extraction**

*Do you have any ethical concerns regarding the sourcing and extraction of your data?*

**Ethical implications of business requirements**

*Are there any ethical implications of the business requirements?*

**Analysis**

**Stages in the data analysis process**

*What were the main stages in your data analysis process?*

* First of all I could have decided to measure the departure delays in terms of number of delays or in terms of their length. I decided to concentrate on their length, since it seems to me much more interesting if I am in aim of comparing the delays with the weather conditions.
* Comparing the `weather` and `flights` datasets came to light that the weather recording information in the dataset I used, are only available for the departure airports. Since the `departure delays` and `arrival delays` variables are strongly correlated and the aim of the project is to focus mainly on three New York airports, I decided to concentrate most on the impact that weather has on the departure delays. Even if I am convinced that could be a good extension that of looking for a further dataset containing weather information also on the destination airports and analyse in which way the arrival place weather influence the arrival delays too or to analyse how much the weather conditions impact on the arrival delays at the airports of interests which I assume must be strongly correlated to the departure delays too. However, this could be an interesting point to analyse for a possible extension.
* Finally, since I am interested in the impact that the weather has on the departure delays and I it doesn’t seem to me reasonable that early departures can be caused by weather ( a part from specific and critical conditions that we are not considering in this analysis), I also decided to consider only positive values of departure delays, which correspond to the effective delays and not to the early departures. Furthermore, the first statistical analysis I performed seems to validate this assumption, since the correlation between departure delays and weather conditions seems to be a bit higher if we take into account only effective delays.
* The main stages of my analysis can be then summarised as follows:
* First of all, I decided to delete from the `weather` dataset the column containing values for temperature, dewpoint, pressure, humidity and precipitation. This was because the not Nas values in this columns where more or less the 10% of the dataset, which induced me not to consider them statistically significant for my analysis.
* Then, I started having a look of the correlation coefficients between departure delays and the remaining weather features. From this analysis I realised that there is a very strong positive correlation between the wind speed and the wind gust speed. For this reason I also decided to delete the `wind gust` column from my dataset since any correlations that flights delays can have with the wind speed automatically imply the same correlation with the wind gust speed too. At the end of this process my left weather features were: “wind\_speed”, “wind\_dir” and “visibility”.
* Using visualisations and linear models I realised that dividing the values of the three weather features in groups was helpful for having a better idea of how each of them influence most the delays (I also used linear models for defining a better way of grouping them).
* Using visualisations I also analysed if for different airlines and different airports, the three weather features have a different impact on the departure delays.
* Finally, I made a prediction on future departure delays greater or not than 10 minutes using a random forest algorithm, which I compared to a logistic regression algorithm too.

**Tools for data analysis**

*What were the main tools you used for your analysis?*

**Descriptive, diagnostic, predictive and prescriptive analysis**

*Please report under which of the below categories your analysis falls* ***and why*** *(can be more than one)*

**Descriptive Analytics** tells you what happened in the past.

**Diagnostic Analytics** helps you understand why something happened in the past.

**Predictive Analytics** predicts what is most likely to happen in the future.

**Prescriptive Analytics** recommends actions you can take to affect those outcomes.