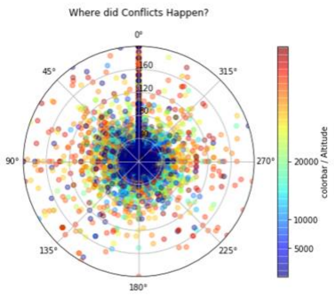
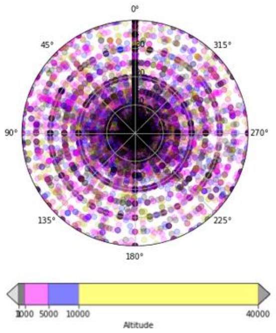
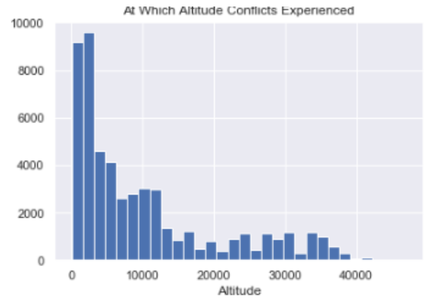
Exploratory Data Analysis Mini Project

For my Capstone Project #1, After I cleaned the data, I used some data visualization and EDA techniques to understand the interactions between the variables. The details are explained steps below:

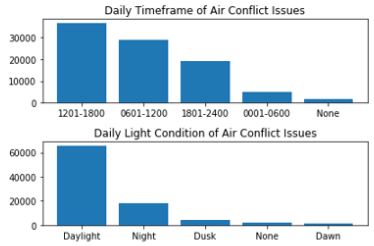
1. After classifying the safety incidents in four categories, air conflict issues were being chosen and the columns which has less than 10% information dropped. As a result, the new data set shape is 91670 x 58.
2. Then I started to research on altitude, location (radial, distance columns), time, light, phase, airspace class, month, year, state, primary cause and narrative (synopsis) variables.
3. Between all those variables there were inconsistencies because of having empty cells in each questionnaire. The tendency of the crew was using narrative and synopsis sections but arbitrary inputs for the remaining cells. For that reason, most of the variables has a big chunk of “None”, “0” or “NaN” inputs.
4. Radial, distance and altitude show us where the exact location of conflict issue. At least one of this three variables might have a “0” value since it was empty in original input file. I tried to visualize those three on a polar chart. “0” radial, “0” distance or “0” altitude has more data then the others. It creates a kind of unreliability on those data.





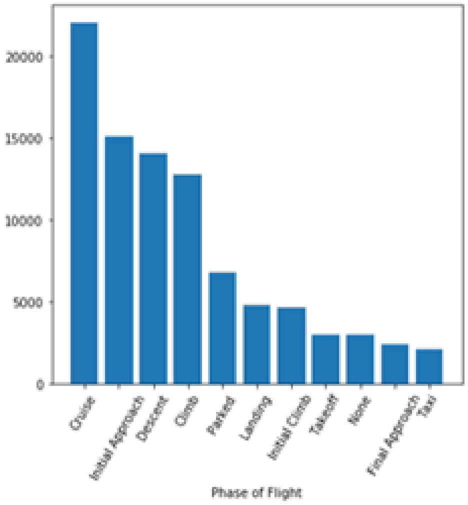


5. One other interaction analysis between the variables was time and light variables. Those two has similar characteristics by nature. The effects on the conflict issues should be same and this makes it redundant.

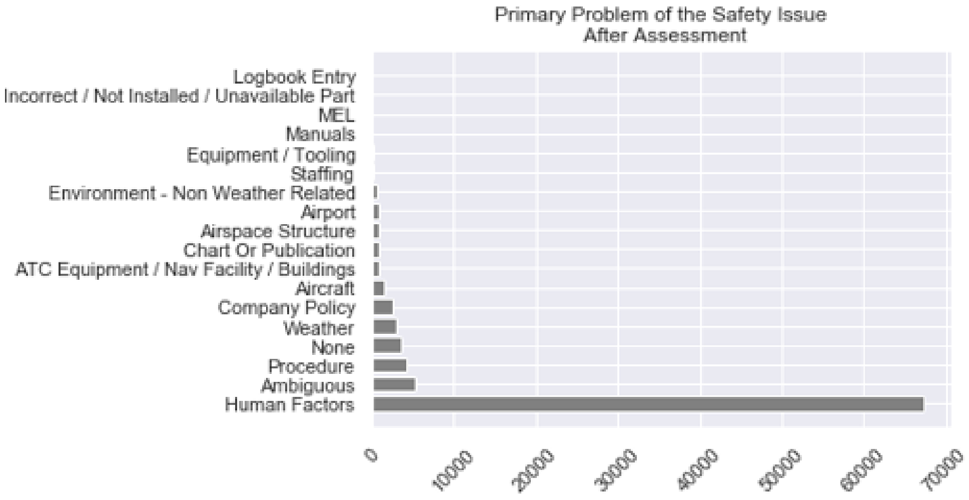


6. One big problem for this dataset is all information belongs to conflict issues. And we do not have the information of other flights which has no records of safety

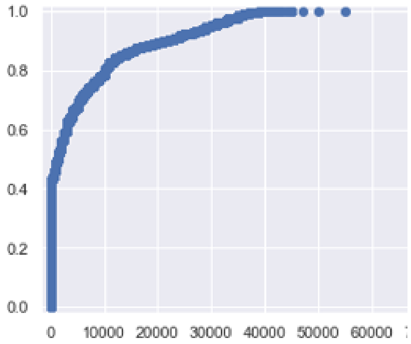
issues. So, for dependant column all records should be “1” and obviously this is not an appropriate target variable. To predict the safety issues won’t be possible with the current database. Then, I focused on “Phase”, “Anomaly” and “Pri_Problem” columns as the dependant variables.



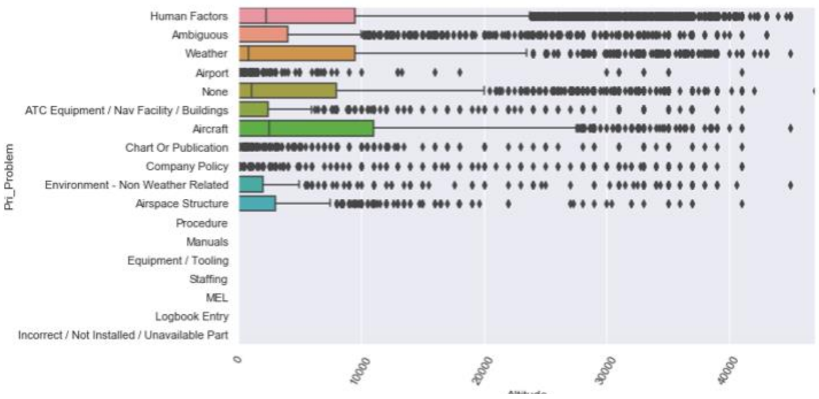
a. Phases: This variable represents the phase of flight such as “take-off”, “taxi”, “cruise” or “landing” and is highly inter-related with the other location variables and has no relationship with time, light, meteorology… etc. variables. Additionally, to predict this variable by location variables won’t provide any kind of value-added to the customer. Because location variables are showing the parameters relative to airports. It is a different way to define phase.
b. Anomaly: During the data wrangling part, the anomaly variable was used to categorize the data and between four categories “air conflict” issues were selected. And now the target column become more homogenous. Also, other groups of variables are not strongly related with “Anomaly” column.
c. Pri_Problem: This variable represents the root cause of the conflict issue. The biggest chunk of data is Human Factors. This variable is the best candidate for target column. Because, firstly there is a huge value-added in this research to understand the safety issue’s reason. Second, the data in this column is pretty consistent and pure analyzed through human experiences. The next step should be to understand the relations with other variables.

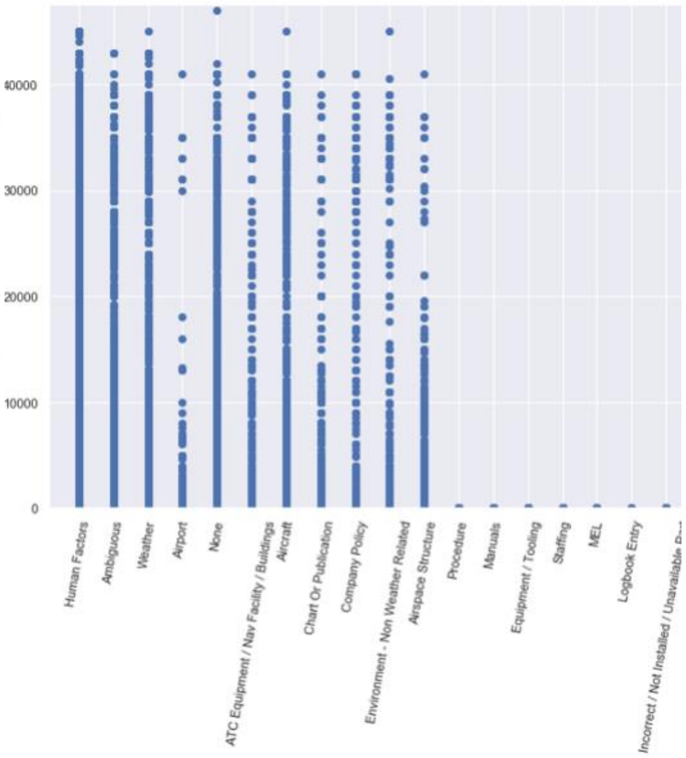


7. If we analyse the relationship between independent variables and Pri_Problem column, the findings were not so strong. As an example I will show “Altitude” variable distribution on target column.
8. Emprical Cumulative Distribution Function of “Altitude” column is depicted below:



9. The boxplot and scatterplot of Altitude and Pri_Problem column is as follows:





10. As a result; the most suitable model for this project would provide the root cause analysis for aviation safety issues. Pri_Problem column will be the target (dependant) variable. But still there is not a very powerful connection with independent variable and dependant variable. Further research needed for a better machine learning model. Other options can be “Narrative” and “Synopsis” columns as the independent variables. But those columns are not numerical and the data is formed by texts. My decision will be going on those variables to find a decoding as numerical and formulate a relationship with the target column.