Adnan Aygunduz  
Capstone #1 Milestone Report  
Springboard DS, March 2019

**Can Air Accidents and Incidents Be Predicted?**

**Problem:** Air accidents and incidents can cost lives and huge resources. Understanding the factors behind them based on the previous data we have can help authorities to take preventative measures in advance and save lives and valuable assets. Thus, this project will answer the question if we can create algorithms in order to predict before it happens or not based on the collection of data we have so far. Since air transportation safety and specifically flight safety is the paramount, government organizations such as National Transportation Safety Board (NTSB) and commercial air carriers will benefit from the outcomes of this study.

**Data:** The NTSB aviation accident database contains information from 1962 and later about civil aviation accidents and selected incidents within the United States, its territories and possessions, and in international waters. Generally, a preliminary report is available online within a few days of an accident. Factual information is added when available, and when the investigation is completed, the preliminary report is replaced with a final description of the accident and its probable cause. Full narrative descriptions may not be available for dates before 1993, cases under revision, or where NTSB did not have primary investigative responsibility.

**Data Set:** [ntsb-aviation-accident-dataset](https://public.opendatasoft.com/explore/dataset/ntsb-aviation-accident-dataset/information/?dataChart=eyJxdWVyaWVzIjpbeyJjaGFydHMiOlt7InR5cGUiOiJsaW5lIiwiZnVuYyI6IkNPVU5UIiwic2NpZW50aWZpY0Rpc3BsYXkiOnRydWUsImNvbG9yIjoiI0QwNTM1NiJ9XSwieEF4aXMiOiJldmVudF9kYX)

**Data Wrangling:**

1. **What kind of cleaning steps did you perform?**

###### After Prioritizing the Data Fields I followed the basic steps below as a checklist to clean my data set.I mainly used pandas. Before I got started I ran some commands such as "head()", "info()", "describe()", "pandas\_profiling.ProfileReport()" and "isna().any()" to better understand the dataset that I am dealing with.

1. Get Rid of Duplicates

With this command I got rid of all repeated rows in my dataset

df.drop\_duplicates(subset=None, keep='False', inplace=False)

1. Get Rid of Extra Spaces.

To get rid of seen or unseen extra spaces (white spaces) I've applied

df.column\_name.apply(lambda x: x.strip())

1. Select and Treat All Blank Cells.

column\_name.fillna("") or column\_name.dropna() methods to data frame.

1. Standardize existing columns and create new ones

*# make a list of the header row and strip up to the 4th letter. This is the location and year information*  
cols1 = list(df.columns)  
cols1 = [str(x)[:4] **for** x **in** cols1]

1. Change Text to Lower/Upper/Proper Case.

.apply(lambda x :x.replace('-','\_')

1. **How did you deal with missing values, if any?**

First I found out the total missing values and their percentage to the data frame.

mis\_val = df.isnull().sum()

# Percentage of missing values

mis\_val\_percent = 100 \* df.isnull().sum() / len(df)

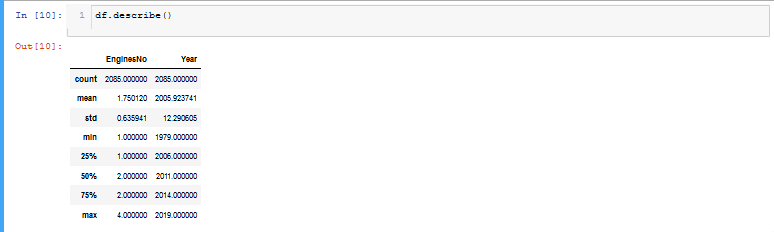
#### For missing categorical data simply labeled them as ’Missing’ or 'N/A' !

For missing numeric data, first flagged them as missing and then filled them according to data content, sometimes with '0's and sometimes with mean and sometimes with ffill() methods.

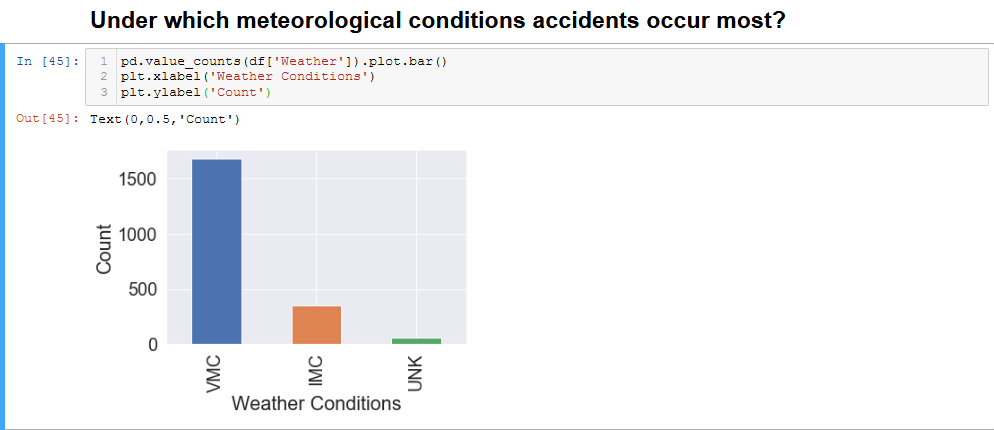
1. **Were there outliers, and how did you handle them?**

There were minor outliers in my data set since it's consisting of mostly categorical data. With those minor outliers I applied 'numpy.std ' standard deviation method to highlight any points above (Mean + 2\*SD) and any points below (Mean - 2\*SD) .

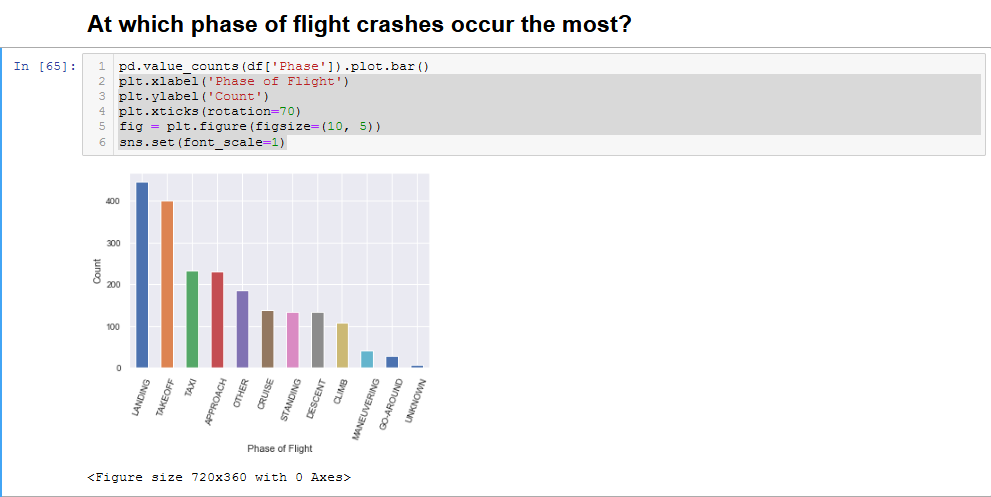
**Initial Findings**

****

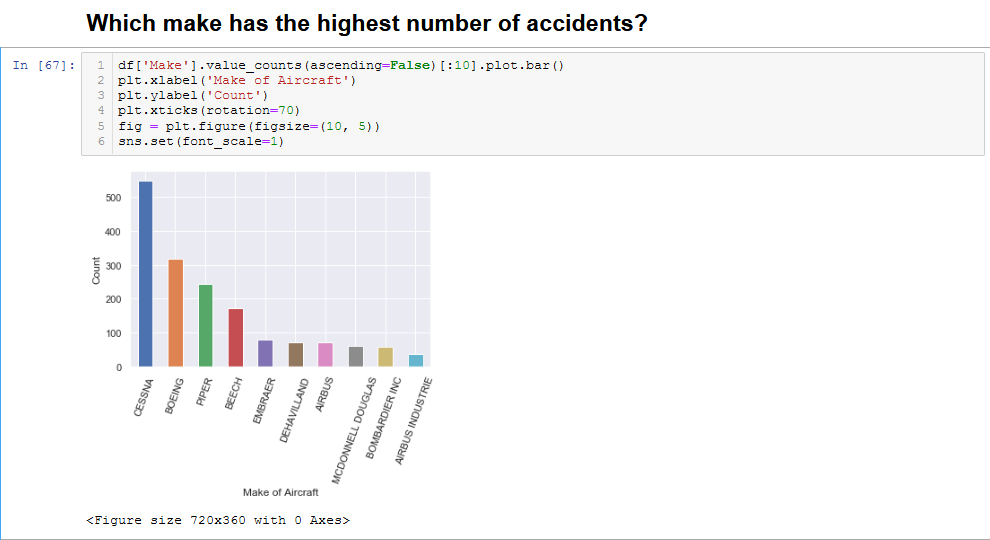
**This data set is consisting of all categorical data except the 'EngineNo' number of engines column. Therefore analyses are based on bar charts created from value counts.**



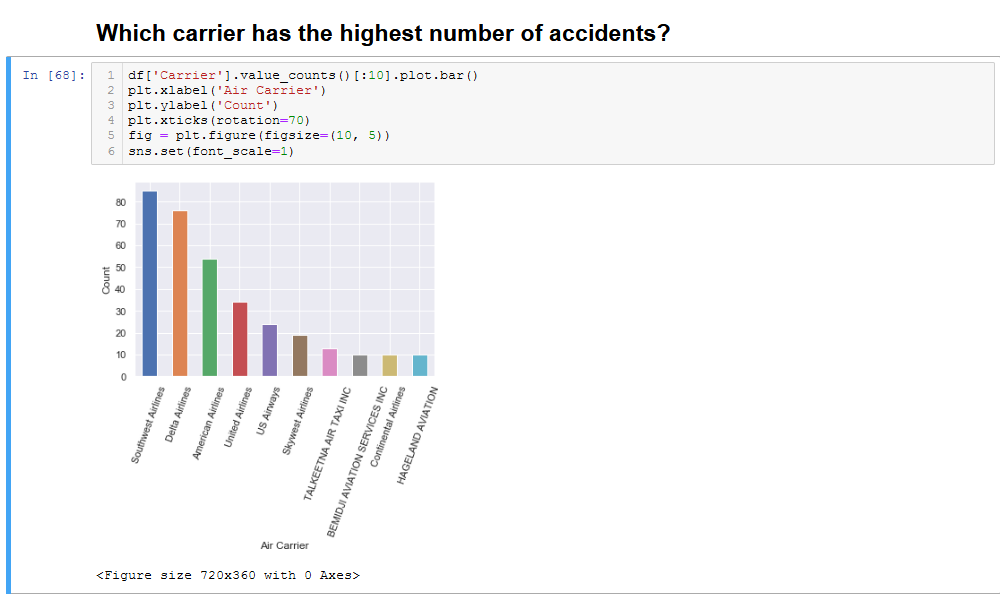
Accidents occured in Visual Meteorological Conditions more than the Instrument Meteorological Conditions. IFR flights relies on aircraft instruments. In VMC, visual references are used as primary data. Technological developments in aviation instruments can effect the accuracy of the flight under IFR conditions.



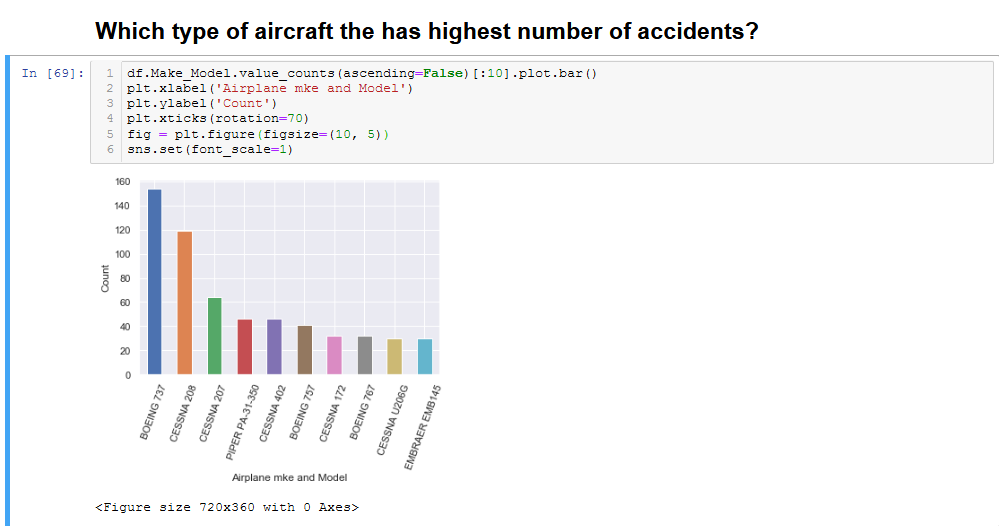
Based on the chart above landing and take-off phases of a flight are the most critical time frames.



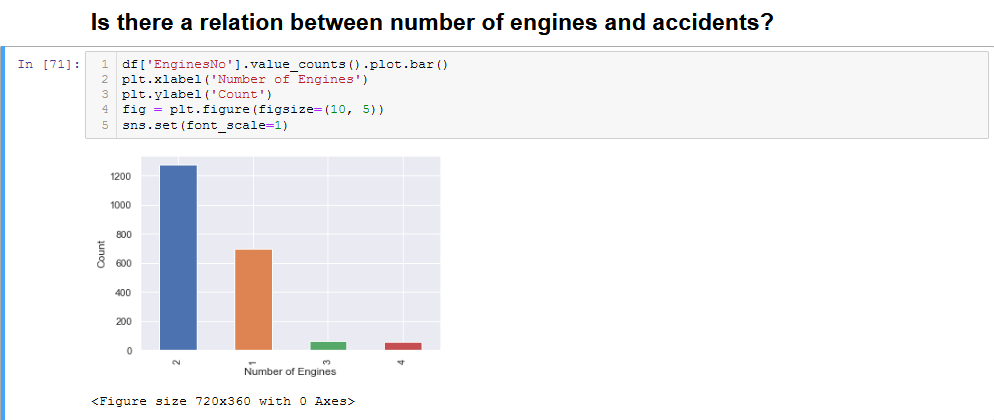
To be able to make a sound judgment total flights hours of each make should be taken into consideration.



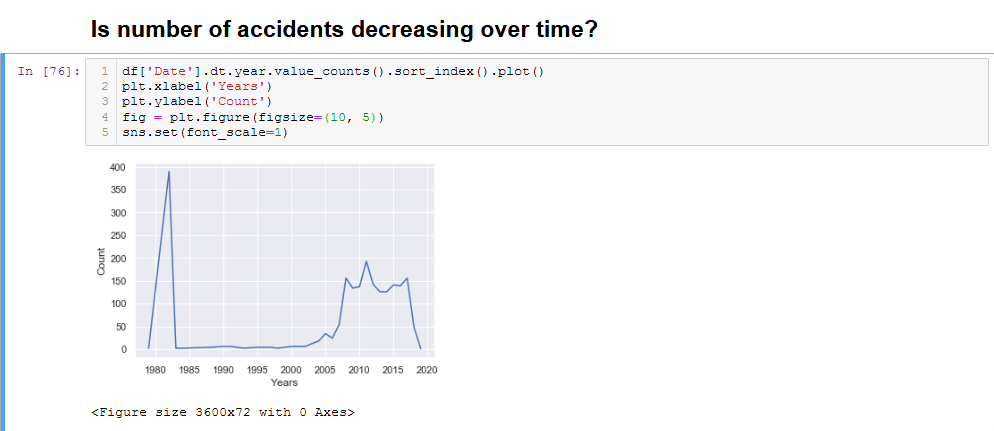
To be able to make a sound judgment total flights hours of each carrier should be taken into consideration.



Time distribution should be observed between each accident of a same make and model.



Majority of the aircrafts involved in an accident were twin engine. Total flight hours should also be considered.

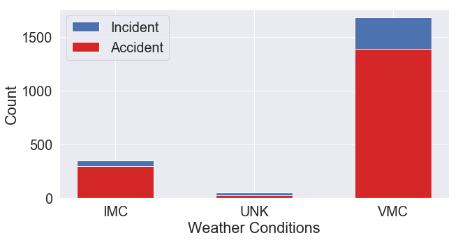


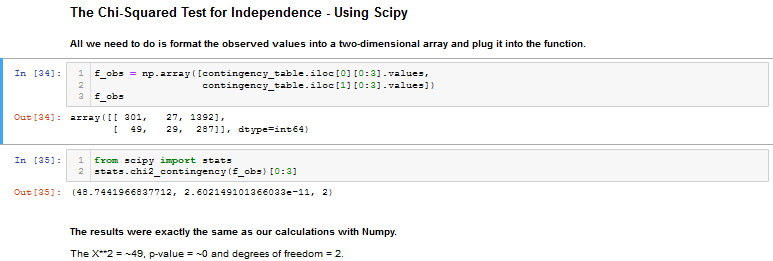
Decreased number of accident between 1982 and 2004 might be the result of applied new practices after huge number of accidents.

**The Null and Alternate Hypotheses**

H0:There is no statistically significant relationship between weather conditions and investigation type.

Ha:There is a statistically significant relationship between weather conditions and investigation type.





Conclusions

With a p-value < 0.05 , we can reject the null hypothesis. There is definitely some sort of relationship between Accident and the Incident column. We don't know what this relationship is, but we do know that these two variables are not independent of each other.