Basic Statistical Analysis

As table 1.1 suggests, comparing the mean and median for all numerical attributes in our dataset, only subtle differences exist between these two measures. However, the standard deviation for each attribute is relatively large, which suggests high variations occur in the weather attributes.

The mode of each categorical attributes, as table 1.2 suggests, indicates the most frequent airlines, aircrafts, destination cities and other features. Since the sample is derived from the top 30 airports, the most frequent airlines and aircraft type correspond to the biggest airline companies and manufacturers. Therefore, normalization is not needed in most cases, since the data represent the population of airlines in general. However, in machine learning prediction analysis part, normalization is important for training and preventing the effects of bias.

In the first glance, high values of standard deviation, for example, attribute ‘cloud\_altitude’ whose deviation weight is almost 73 percent of its mean, imply high variation and instability of data features. Therefore, it is important to visualize these abnormal distributions of features and see if these abnormal patterns are due to high number of outliers.

|  |  |  |  |
| --- | --- | --- | --- |
| features | mean | median | std |
| cloud\_altitude | 15453.2 | 15000 | 11425.62 |
| temp | 22.82524 | 23 | 5.755422 |
| dewpoint | 22.82524 | 23 | 5.755422 |
| visibility | 9.640777 | 10 | 1.390022 |
| wind\_speed | 7.403883 | 7 | 2.94641 |
| gust\_speed | 0.848544 | 0 | 3.793962 |
| DurationMin | 150.0303 | 107.2333 | 133.5355 |

Table 1.1 Statistical Summary of Numerical Attributes: (Mean/Median/Standard Deviation)

|  |  |
| --- | --- |
| features | mode |
| ident | NKS559 |
| aircrafttype | A320 |
| originCity | New York, NY |
| destinationCity | Chicago, IL |
| airline | American Airlines |
| aircraft\_manuf | Boeing |
| Class (Delay) | 0 |

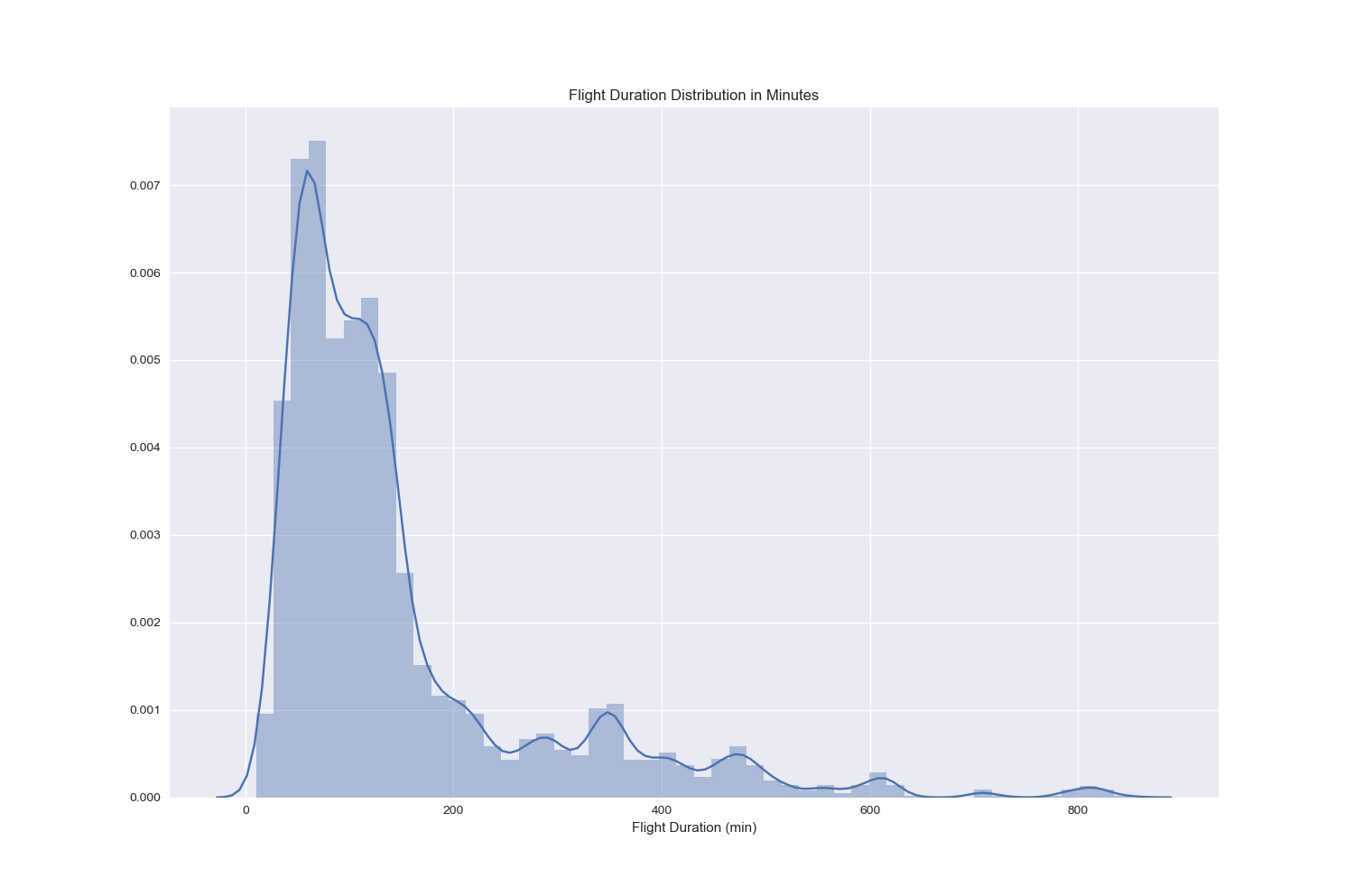
Table 1.2 Statistical Summary of Categorical Attributes: (Mode)

As mentioned before, the information generated by Basic Statistical Analysis, especially in terms of mean, median and standard deviation, is limited. Therefore, detection of outliers and visualize the distribution can help better understand the features.

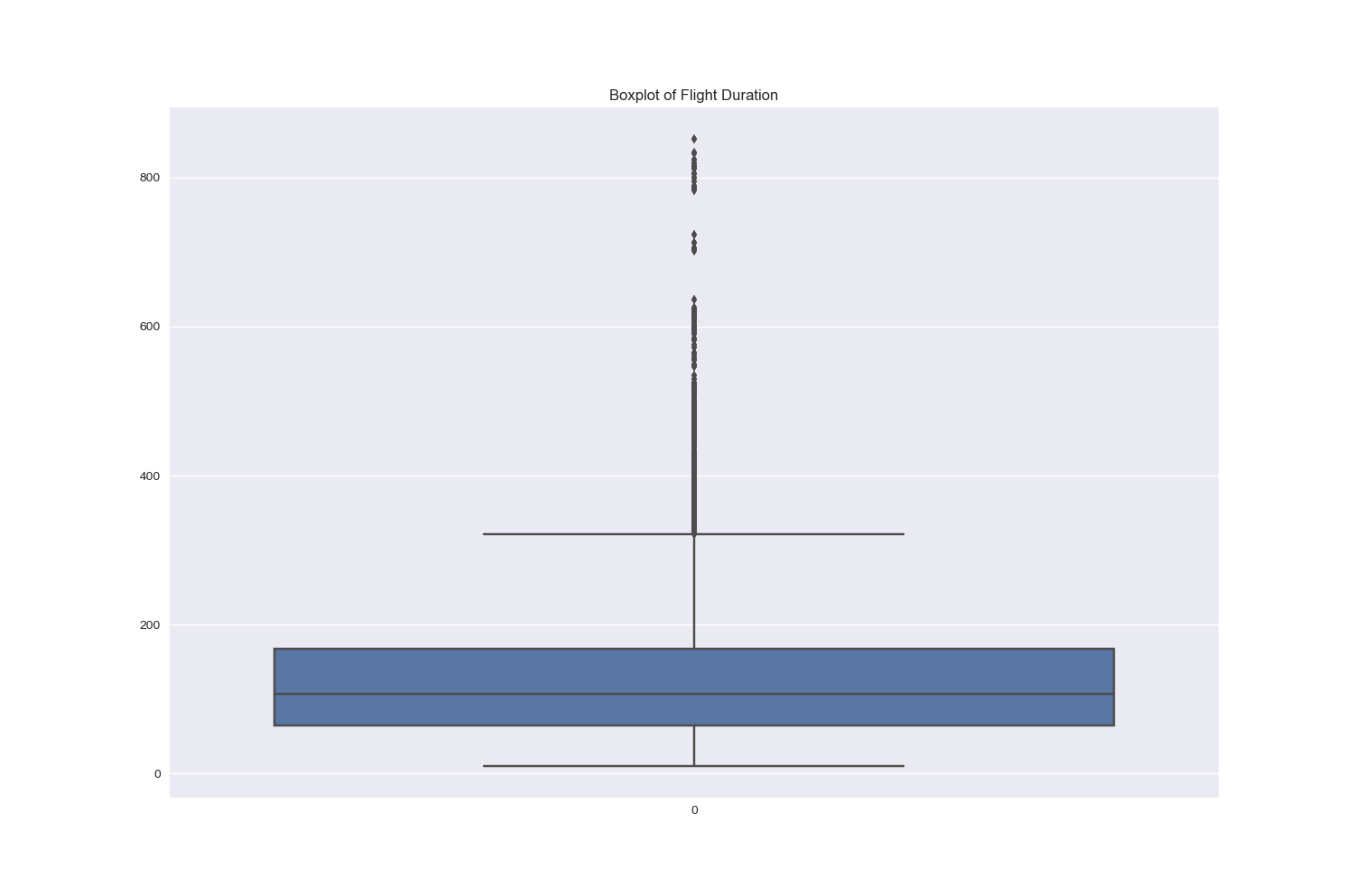
|  |  |
| --- | --- |
| Feature | Number of Outliers |
| cloud\_altitude | 9 |
| temp | 0 |
| dewpoint | 0 |
| visibility | 147 |
| wind\_speed | 3 |
| gust\_speed | 75 |
| DurationMin | 186 |

Table 1.3 Count of Outliers for Numerical Features in Flight Dataset (25-75 percentile)

According to table 1.3 above, feature visibility, gust speed and flight duration tend to have large numbers of outliers. Specifically, for flight duration, as histogram 2.1 and boxplot 2.2 suggest: there are many extremely long flights duration times which lie above 75 quantiles within the feature.

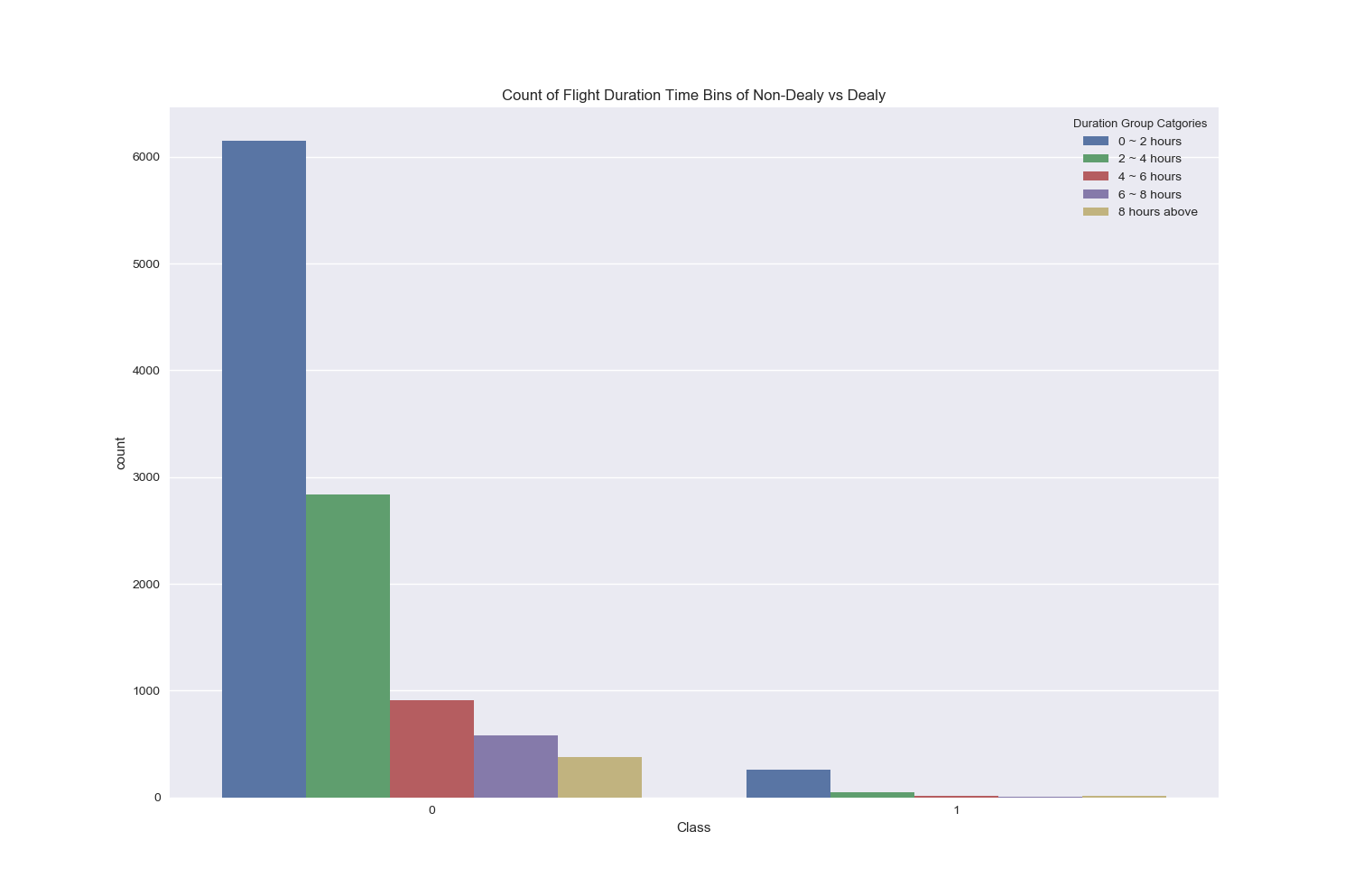


Graph 2.1 Histogram of Flights Duration in Minutes



Graph 2.2 Boxplot of Flights Duration in Minutes

However, these outliers, also known as long flights duration, tend to have association with the delay rate. As histogram 2.3 suggests, after binning the flights duration into 5 duration groups, ranging from ‘1-2 hours’ to ‘8 hours and above’, we can see there are around 600 observations of long-time flights (more than 8 hours) in our dataset. The major reasons of binning flights duration are that this feature is continuous and consists of a range of data from 0 to more than 1000, and binning the time can help distinguish the airlines as short-duration to long-duration flights contextually, which also serve as an interesting factors that correlated with delay rate as the paper will illustrate in later part.

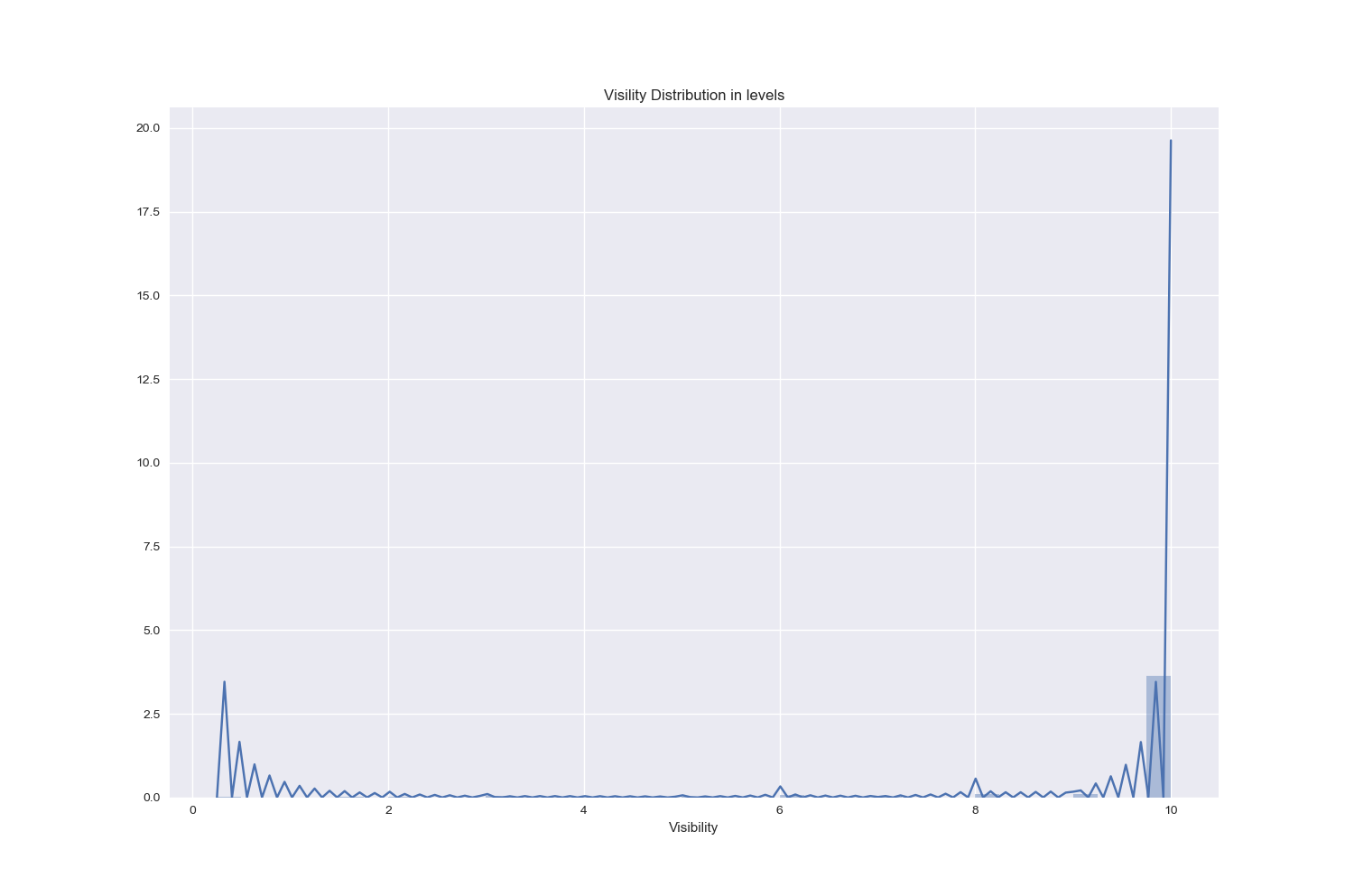
Since these observations are derived from real flights dataset, which are correct and justifiable since long-time flights do exist in major airports.

Graph 2.3 Distribution of Flights Duration Bins and Number of Delays

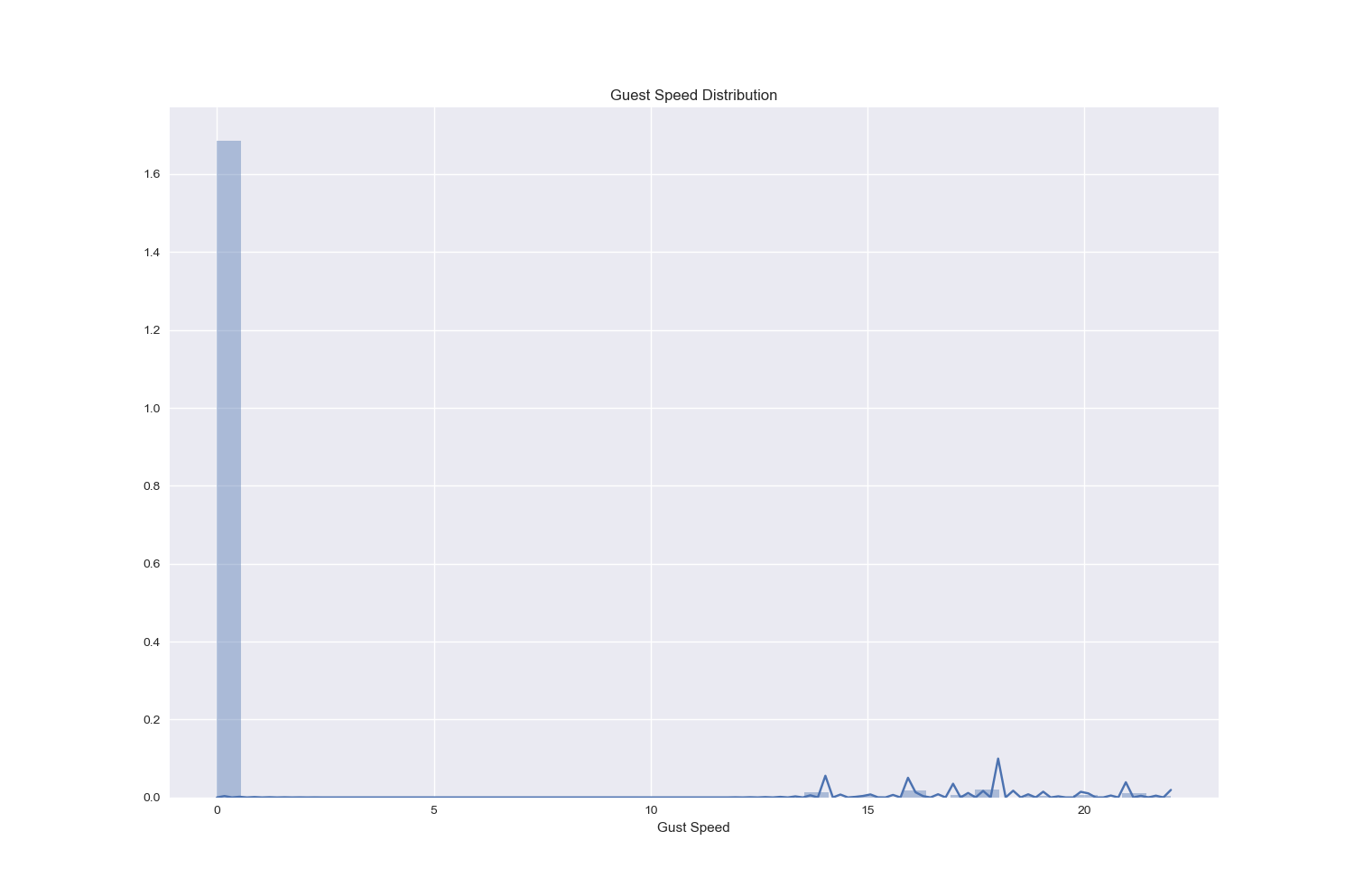
Surprisingly, the distribution of delayed (Class 1) and not-delayed (Class 0) in term of flights duration time are similar, which suggests that there is a potential correlation between flights duration and likelihood of delay. Therefore, the outliers in flight duration time feature can not be removed and are valuable in future machine learning predication analysis.

Graph 2.4 and 2.5 show the distribution of another two features visibility and gust speed whose outliers’ numbers are high. As the distribution of ‘Visuality’, the distribution of discrete levels suggests that there are few levels between 2 to 8, and rather have more 10 and 0, 2 in the dataset. The same logic goes with the distribution of ‘Gust Speed’ in graph 2.5.

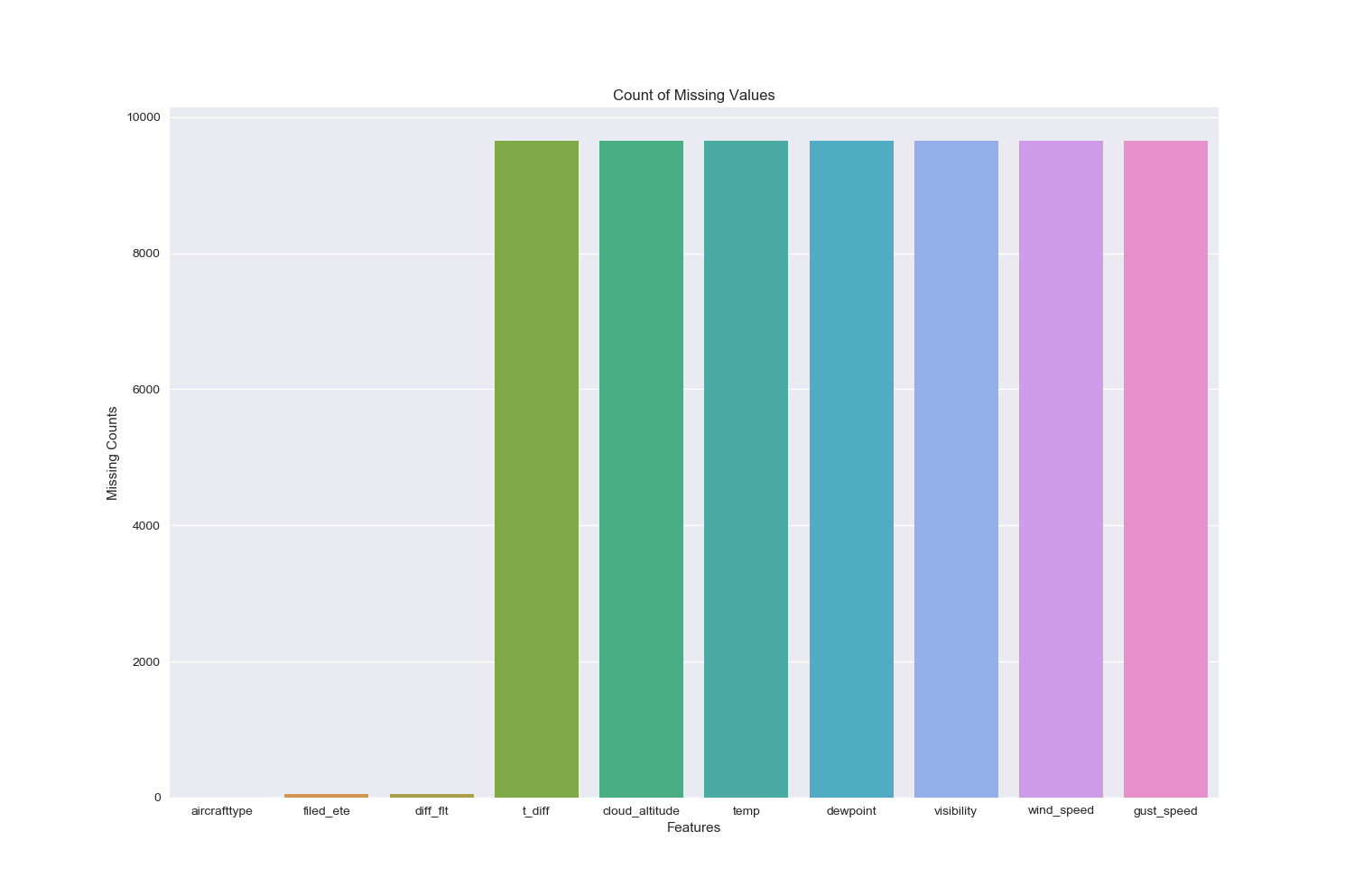
Therefore, the mean and standard deviations can not be used as the only indicators to explain the outliers. In this case, the outliers can not be removed due to the distribution below.



Graph 2.4 Distribution of Visibility Levels in Flights Dataset

 Graph 2.5 Distribution of Guest Speed in Flights Dataset

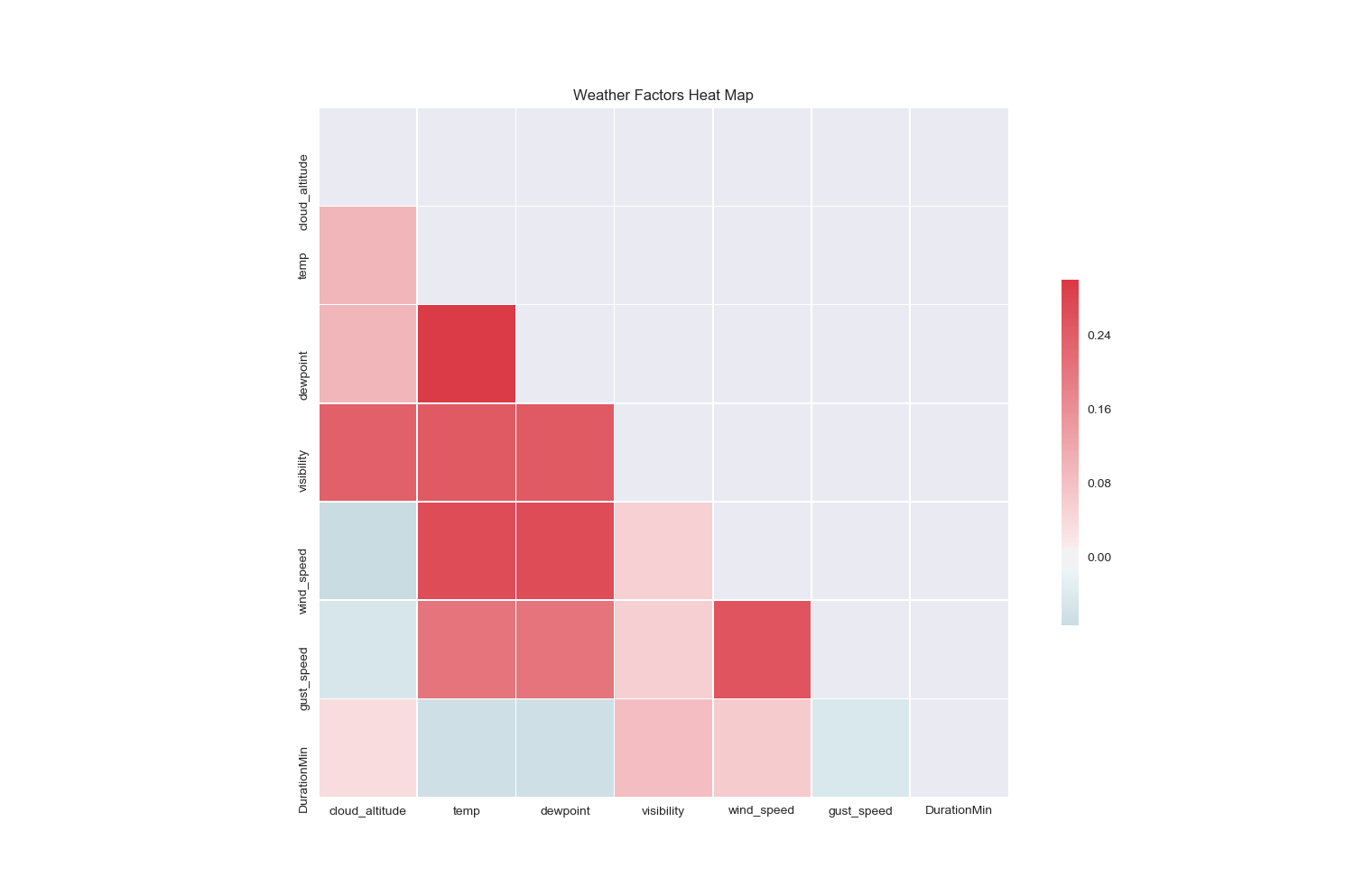
In order to further proceed to analyze flights dataset, the missing values can have negative effects for detecting correlations and building machine learning models. As shown in graph 2.6, the weather features tend to have many missing rows compared to the categorical features related to the information of airlines, destinations, aircraft types and so on. However, removing all the rows containing missing values of weather features will also affect the completeness of other features. Instead, keeping the missing cells under weather features and separating these two types of features to analyze separately is a better option.

Graph 2.6 Count of Missing Values in Flights Dataset

Correlation Analysis

In order to reduce the dimensionality of flights dataset, which contains 41 features in the raw cleaned data frame, finding the correlated features is important to generate better results in machine learning models, such as Naïve Bayes whose assumption is the independency of all features.

Graph 3.1 is the heat map of all numerical features, mostly from weather features that shown the correlation ratio ranging from negative 0.1 to positive 0.3. As the graph suggests, features such as temperature and dewpoint, wind speed and gust speed are highly correlated, with ratio more than 0.2. Such observations can be further analyzed by using Pearson’s R test.



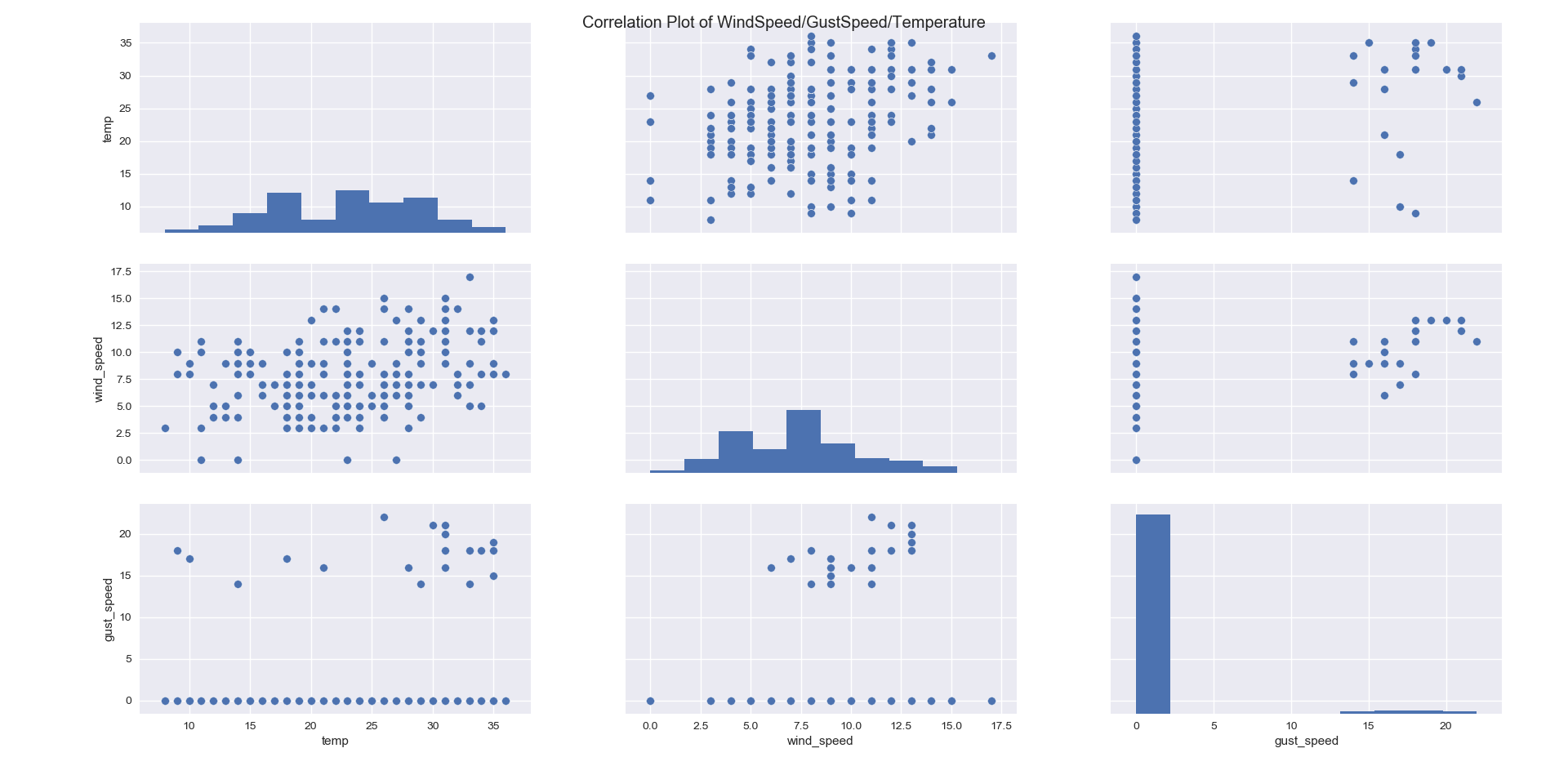
Graph 3.1 Weather Features Correlation Heat Map

Table 3.2 also suggests the Pearson’s R ratio between top three pairs of weather features (Temperature, Wind Speed and Gust Speed); surprisingly, many pairs of weather features are highly correlated. These correlations imply that the weather factors tend to influence each other in a regular pattern, and the dimensionality can be reduced by reducing the highly correlated features.

|  |  |  |  |
| --- | --- | --- | --- |
| Correlation | Temperature | Wind Speed | Gust Speed |
| Temperature | 1 | 0.2695 | 0.20389 |
| Wind Speed | 0.26952 | 1 | 0.25490 |
| Guest Speed | 0.20389 | 0.25490 | 1 |

Table 3.2 Pearson’s R rations between top three pairs of weather features

Graph 3.3 below also suggests the same results for correlations. The scatter plots between temperature and wind speed shows a positive linear correlation, even the correlation is not as strong as a linear approximation. Besides, the temperature and gust wind/guest wind and wind speed form a straight line of scatter points, as shown on the third subplots in first and second lines and first/second subplots in third lines. This observation is interesting because it seems to suggest the independency, as gust speed will always remain to be 0-2 no matter how temperature and wind speed changes. However, the Pearson’s R results show that these pairs are not independent and further analysis will be needed to visualize these discrepancies.

Graph 3.3 Correlation Matrix of Top Three Correlated Weather Features in Flights Dataset