

## SVKM’s NMIMS

Mukesh Patel School of Technology Management & Engineering AN Project Report ON

**Predictive Modelling of Flight Landing Distance**

Faculty Mentor: Submitted By:

MR SURAJ PATIL Ayush Patel (A232)

(Assistant Professor) 70011118040

MPSTME, NMIMS Batch: A

Shirpur Campus

Table

|  |  |  |
| --- | --- | --- |
| Sr No | Table Contents | Page No |
| 1 | Introduction | 3 |
| 2 | Data Description | 3 |
| 3 | Data Import and Merging | 4 |
| 4 | Data Cleaning | 5 |
| 5 | Data Summarization | 7 |
| 6 | Data Analysis and Visualization | 10 |
| 7 | Data Modelling | 17 |
| 8 | Conclusion | 21 |

**Introduction**

The objective of this project is to reduce the risk of landing overrun by identifying the different factors and how they impact the landing distance of commercial flights. Two raw data sets with over 800 and 150 observations respectively were studied for the purpose of this project. After merging the two datasets, removing duplicate and missing values, performing completeness and validity check and removing abnormal values I was left with a total of 831 observations (flight) that were used to fit the final linear regression model. I used exploratory data analysis and data visualization techniques to identify the relationships between our dependent variable and independent variables. After this I validated our findings by building a linear regression model to identify a linear equation that best fits our data.

**Data**

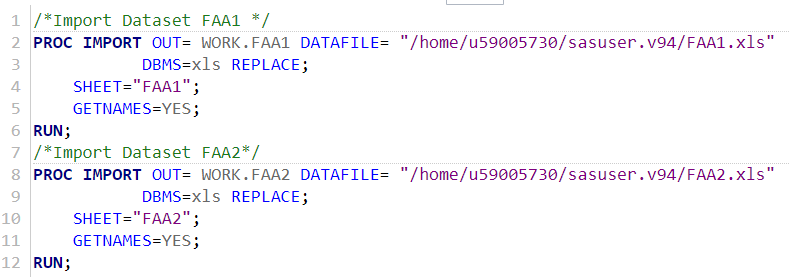
The data is simulated landing data from 950 commercial flights provided in the form of two excel files FAA1.xls & FAA2.xls. FAA1.xls has 800 rows and FAA2.xls has 150 rows.

**Data Description**

* Aircraft: The make of an aircraft (Boeing or Airbus).
* Duration (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.
* No\_pasg: The number of passengers in a flight.
* Speed\_ground (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
* Speed\_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
* Height (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway.
* Pitch (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway.
* Distance (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

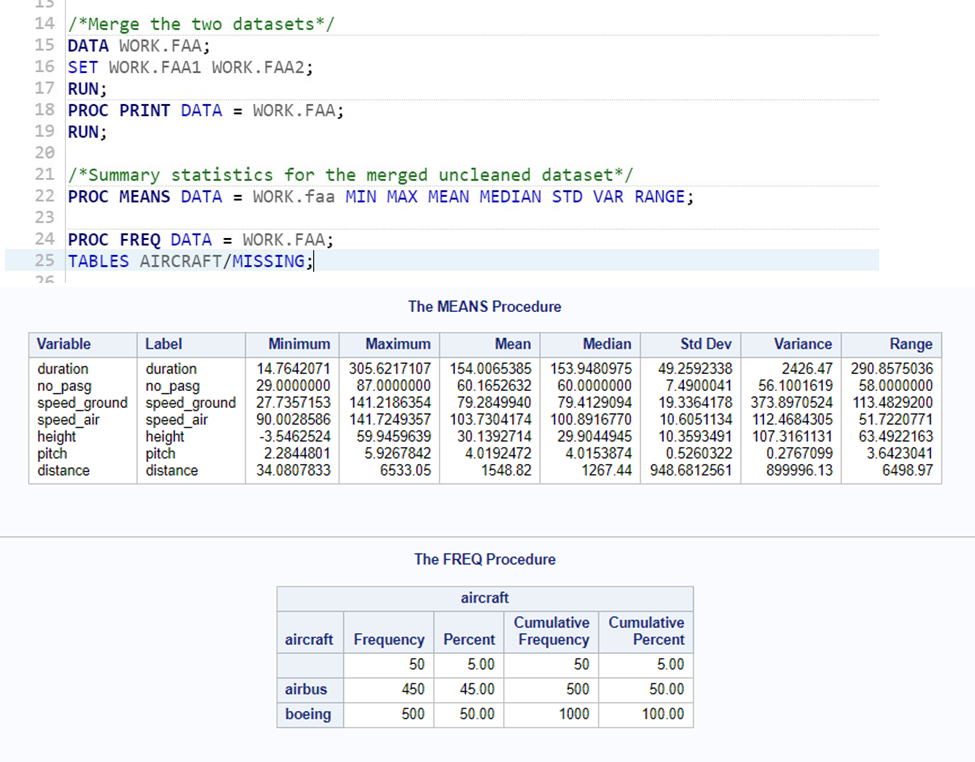
**Data Import**

The first step of our data preparation process is to import the data in the two excel files into SAS.



**Merge Data**

After importing the data into SAS database, we merge the two tables FAA1 and FAA2 into a single table FAA for the purpose of analysis. On merging the two tables, we observe that there are some blank rows in the table FAA. The FAA table has 1000 observation and 8 variables.

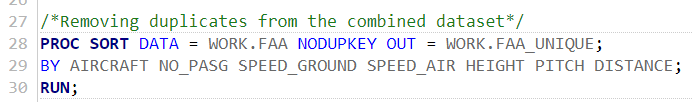


**Data Cleaning**

In this step, we clean the data by removing any missing values, duplicate values and abnormal values. We do this in three steps.

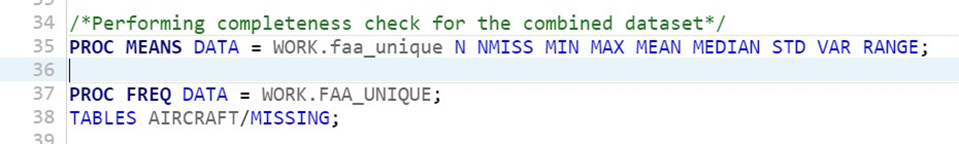
**Step 1:** Removing duplicates from the dataset.

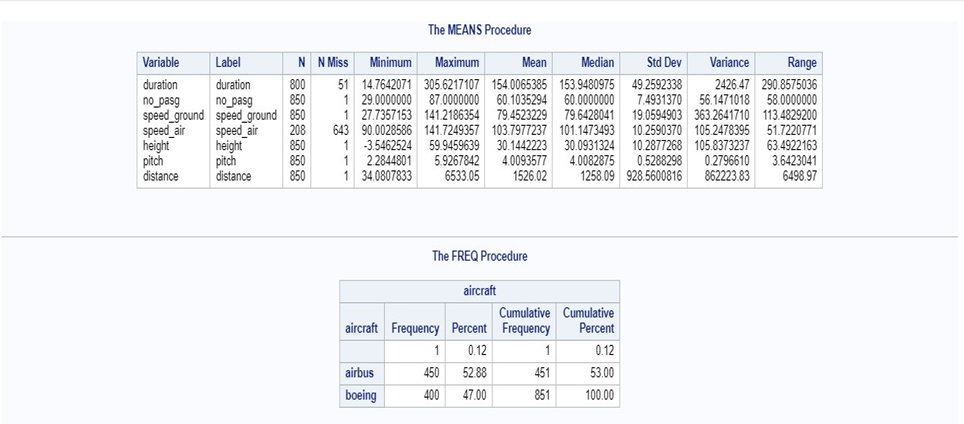
After this step, we are left with 851 observations in our dataset.



**Step 2:** Completeness check

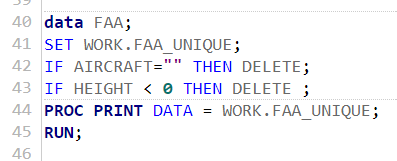
I performed the completeness check of each variable- examine if missing values are present. After that I deleted rows with missing values.





We observe that variable speed\_air has 643 missing values and duration has 51 missing values. At this stage, we do not remove the rows which have missing values for these variables since we do not want to lose data for other variables which may be relevant to our model.

We observe that height has negative value as minimum which cannot be realistically possible. Also, there is one blank row. We remove rows having negative height and blank value for all variables.

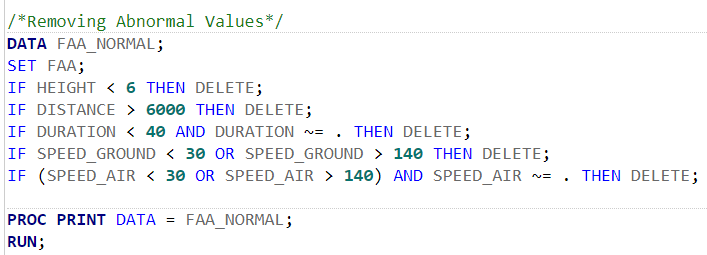


The dataset now has 845 observations.

**Step 3:** Validity check

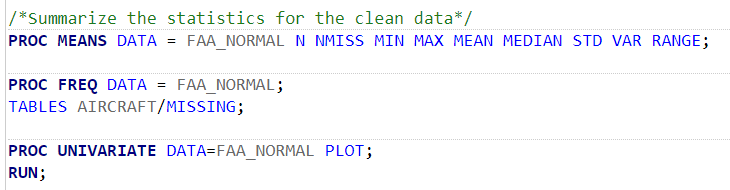
In this step, I performed the validity check of each variable. Examine if abnormal values are present and then remove rows with abnormal values. I defined abnormal values for each variable as:

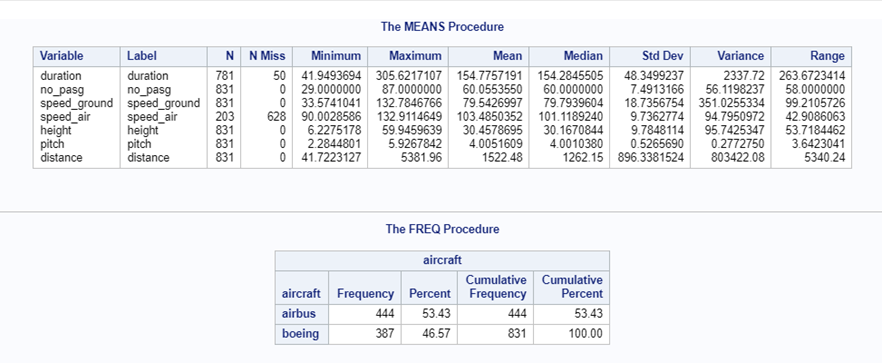
|  |  |
| --- | --- |
| **Variable** | **Abnormal Value** |
| Distance | > 6000 feet |
| Speed\_ground | < 30MPH or > 140MPH |
| Speed\_air | < 30MPH or > 140MPH |
| Duration | < 40 min |
| Height | > 6 meters |



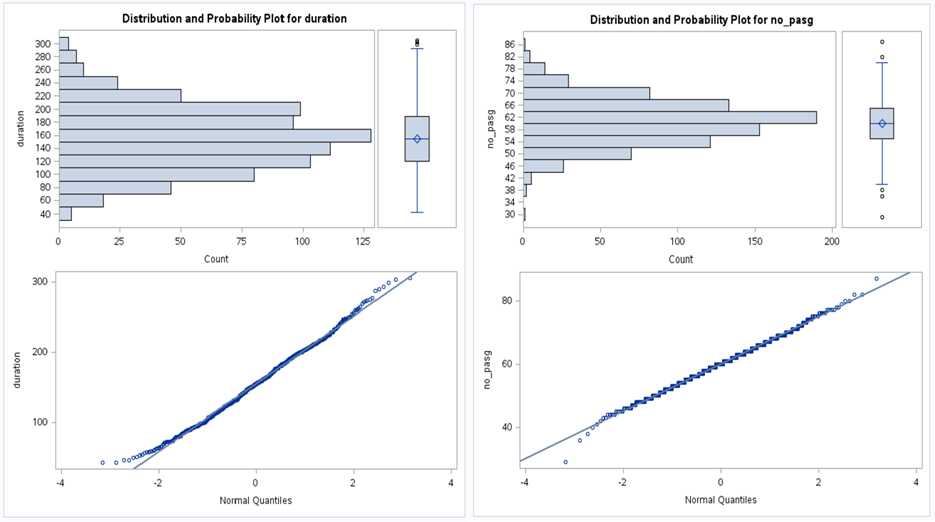
**Data Summarization**

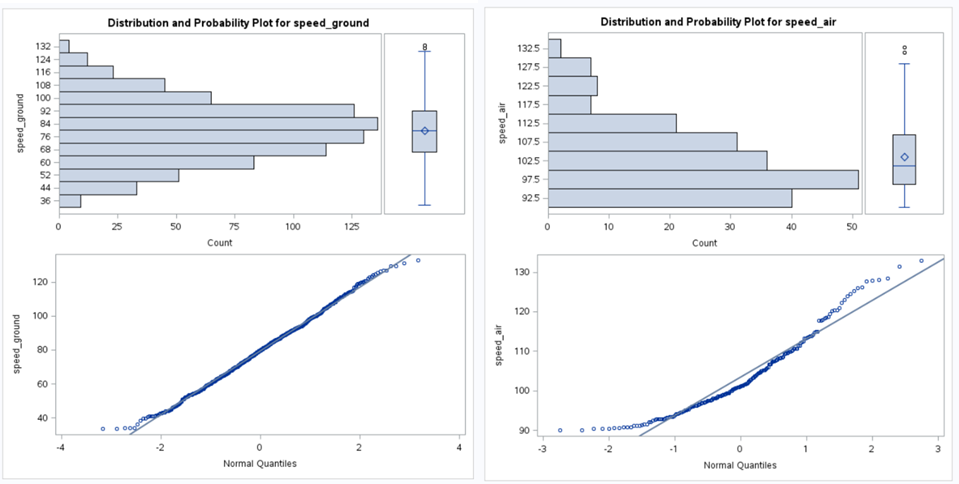
In this step, I summarized the statistics for the cleaned dataset and distribution of each variable.

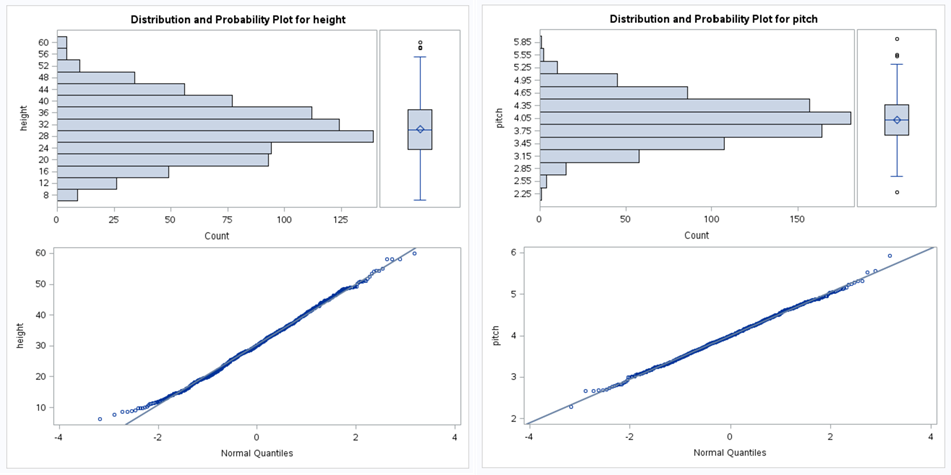


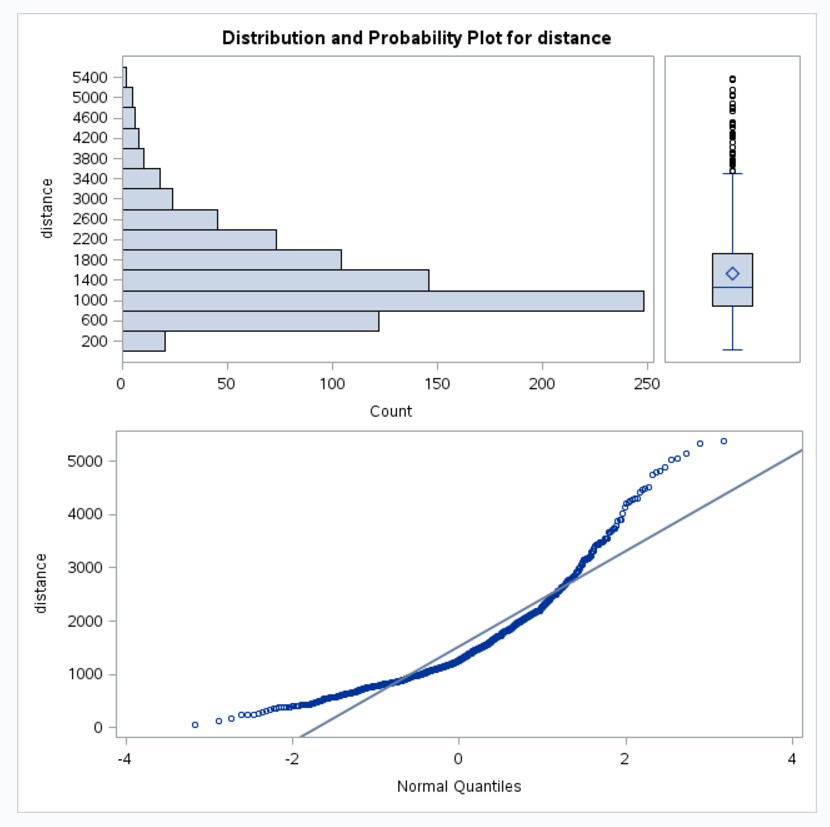


This cleaned dataset consists of 831 observations. Although, speed\_air has missing values for 628 observations and the duration has missing values for 50 observations, we don’t discard the rows having these variables at this point in the analysis.







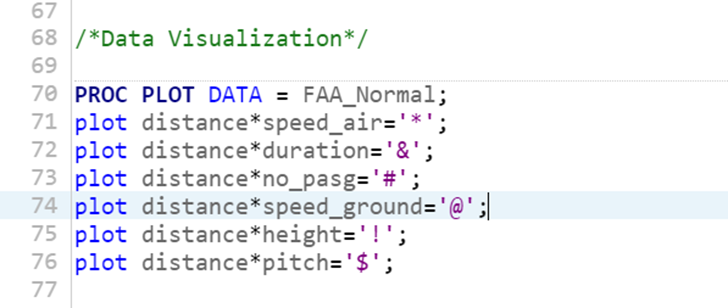


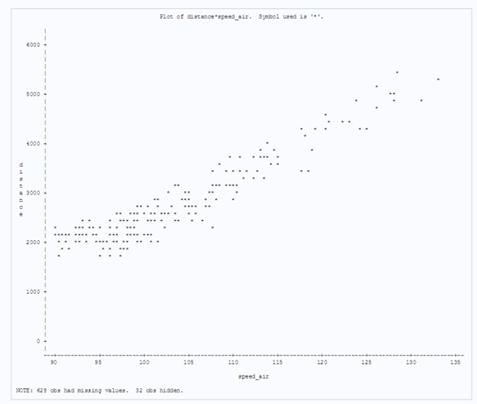
From the distributions of each variable, I observed that duration, no\_pasg, speed\_ground, height and pitch follow normal distribution but speed\_air and distance do not appear to be normally distributed.

**Data Analysis and Visualization**

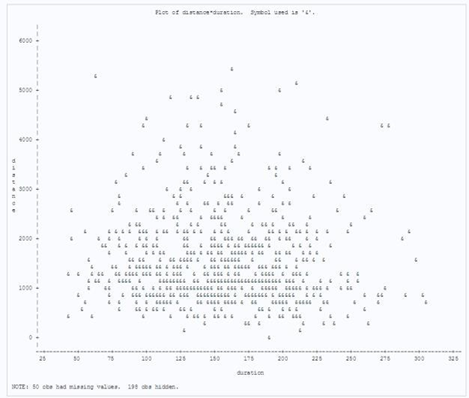
Using scatter plot to understand the relationships between all variables. A box plot was plotted to see the differences in landing distance because of its make.

**Step 1:** I plotted the graphs for our dataset. From these graphs, we can identify which of the independent variables has a linear relationship with the dependent variable distance.

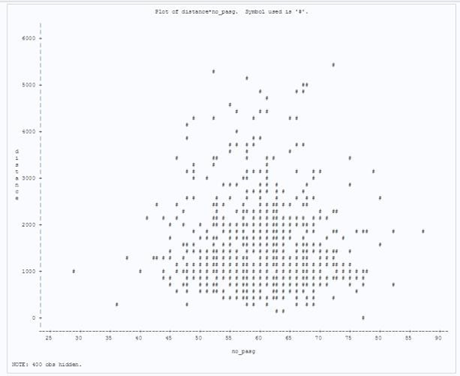




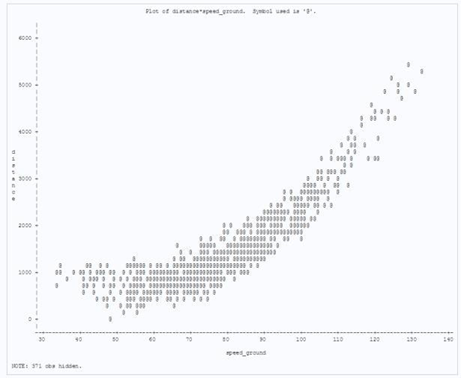
Distance vs Speed\_air



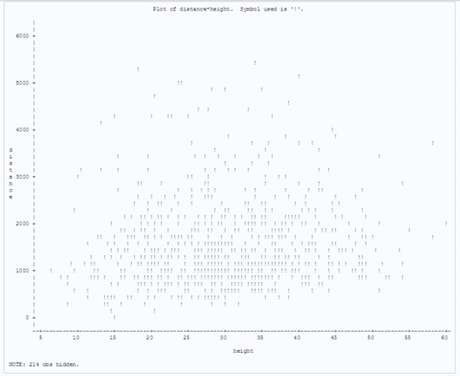
Distance vs Duration



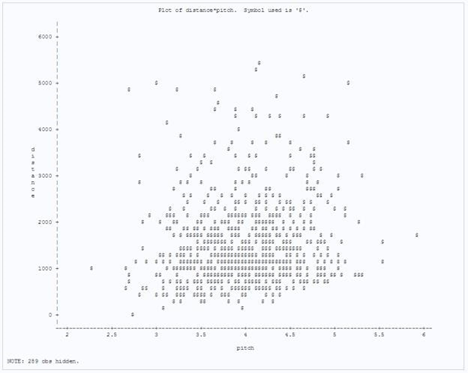
Distance vs No\_pasg



Distance vs Speed\_ground

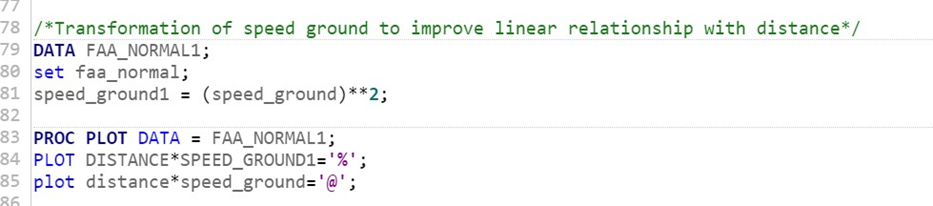


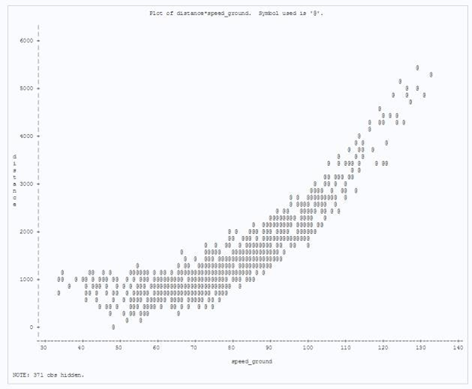
Distance vs Height



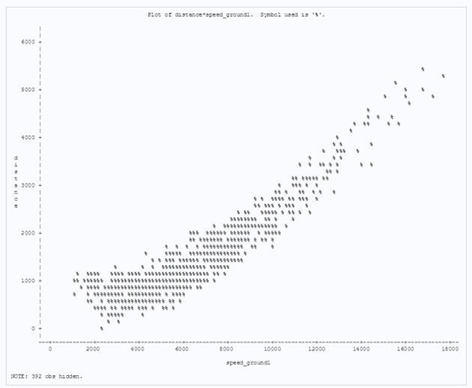
Distance vs Pitch

**Step 2:** I performed transformation of variable speed\_ground to increase linearity with distance. Also, I transformed the categorical variable aircraft into a binary variable so that it could be included in the model.





Distance vs Speed\_ground

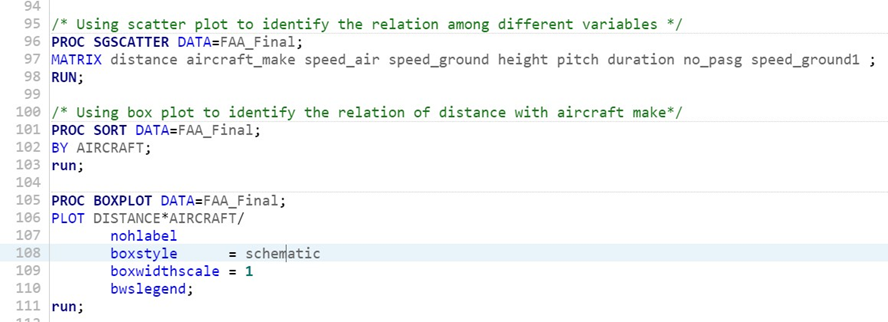


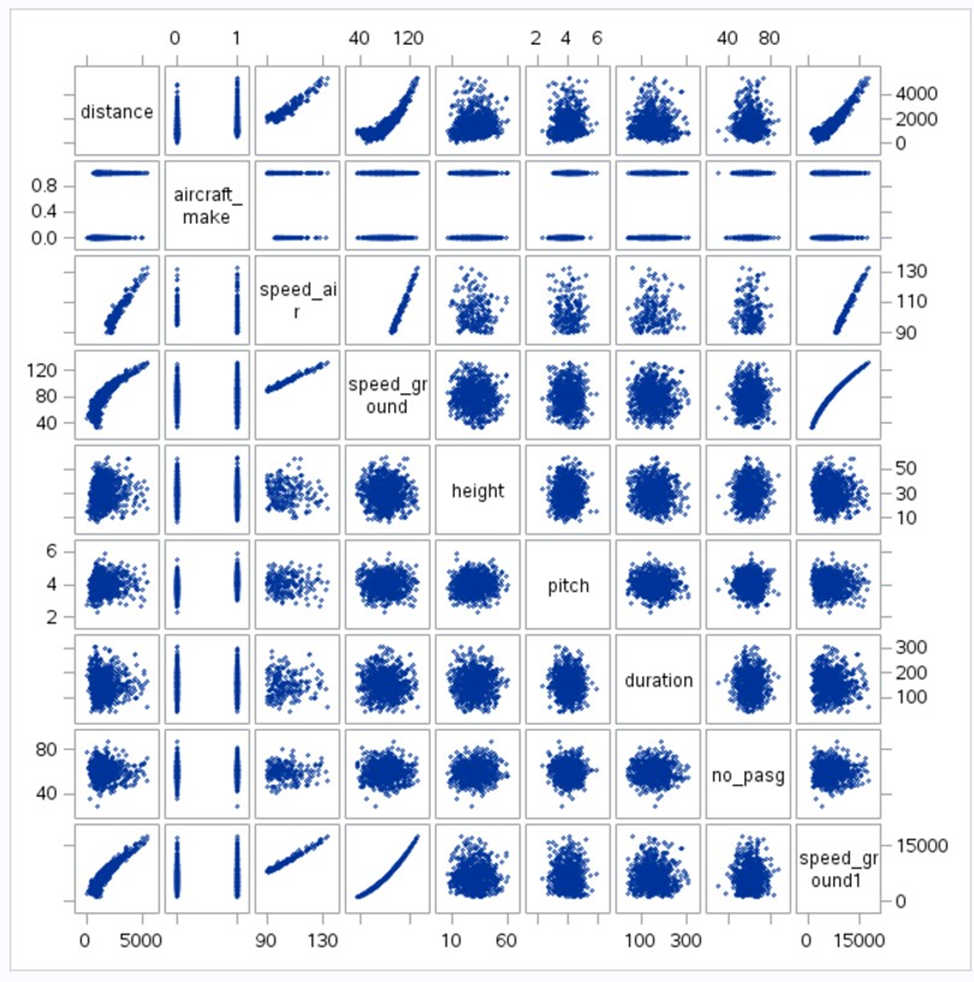
Distance vs (Speed\_ground)^2

One can observe that the first diagram has a slightly better linear relationship with distance.

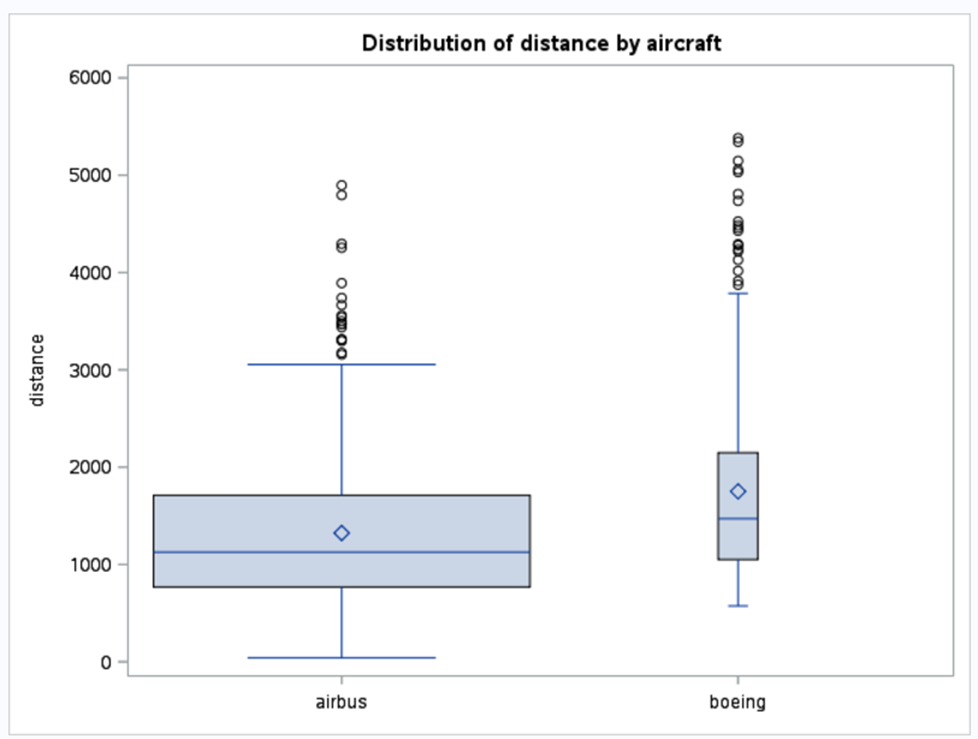


**Step 3:** Using scatter plot to understand the relationships between all variables. A box plot was plotted to see the differences in landing distance because of its make.



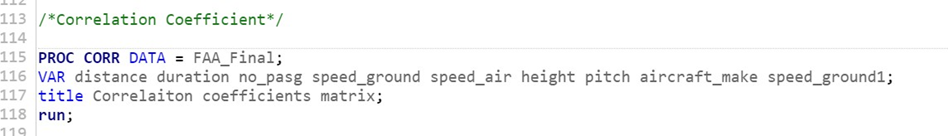


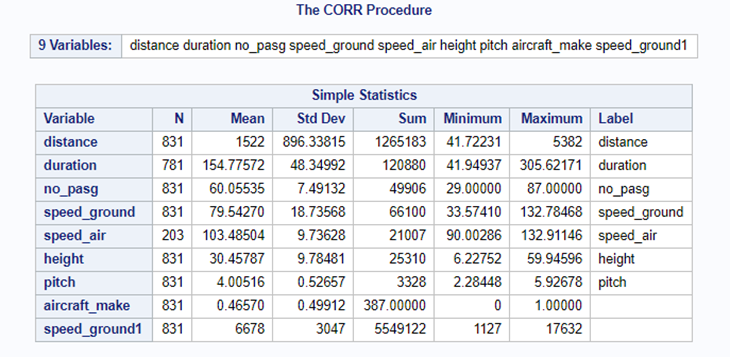
We see that speed\_air, speed\_ground and speed\_ground1 have some linear relationship with landing distance.

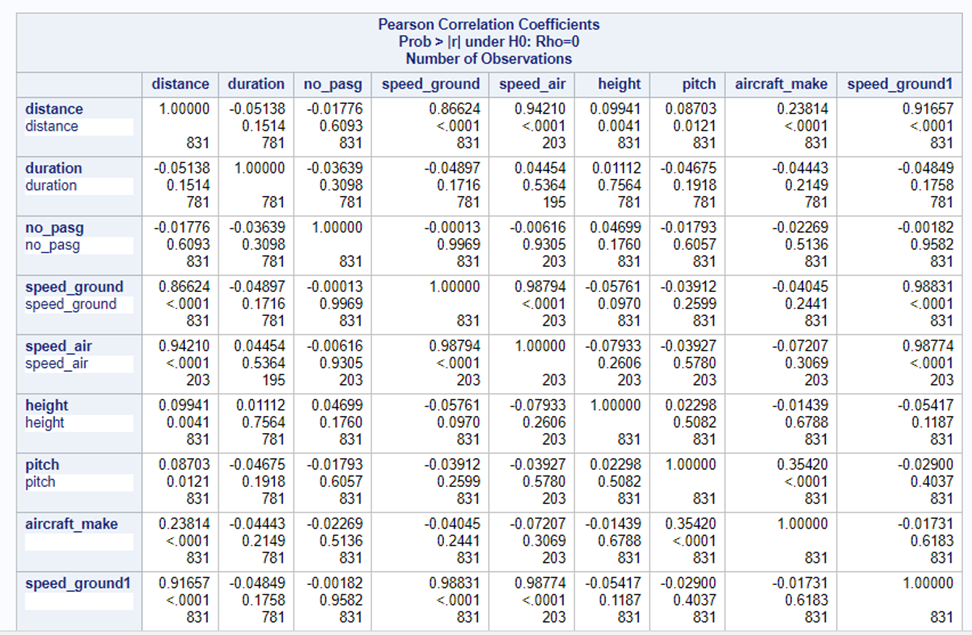


We observe that there is indeed a difference in distance because of aircraft make; for airbus make the landing distance mean, median is lower than that for Boeing.

**Step 4:** Next step is to validate the linear relationships found by the various plots; that is finding the strength of linear relationship between dependent and independent variables. We also need to check if there is any linear relationship between any independent variables. We will only keep one of the related independent variables in our model in order to avoid multicollinearity problems in our analysis.







From the correlation matrix, it is observed that variables aircraft\_make, speed\_air, speed\_ground, speed\_ground1, height have significant correlation with distance variable.

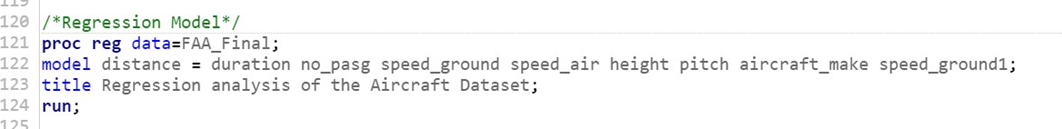
We also observe that speed\_air and speed\_ground are highly correlated and speed\_ground1 is highly correlated with both speed\_air and speed\_ground.

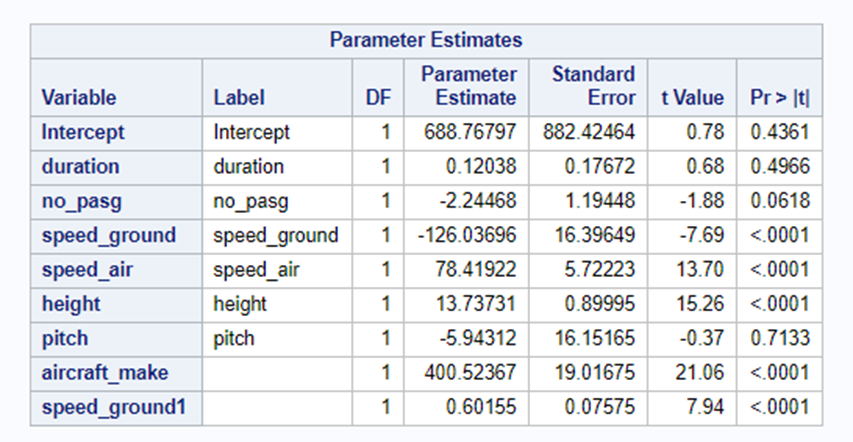
Since there is high correlation between speed\_air and speed\_ground we can drop one of these two variables. We choose to drop speed\_air from our model as it has missing values for almost 75 percent of the observations in our dataset and also because we wish to fit our model to as many observations as possible, so we choose to keep speed\_ground.

**Data Modelling**

The objective is to build a model that best describes the relationship between our dependent variable and independent variables. We will use linear regression to build our model since we observed that there is strong correlation between distance and other independent variables.

**Step 1:** We run the initial model by including all the independent variable.

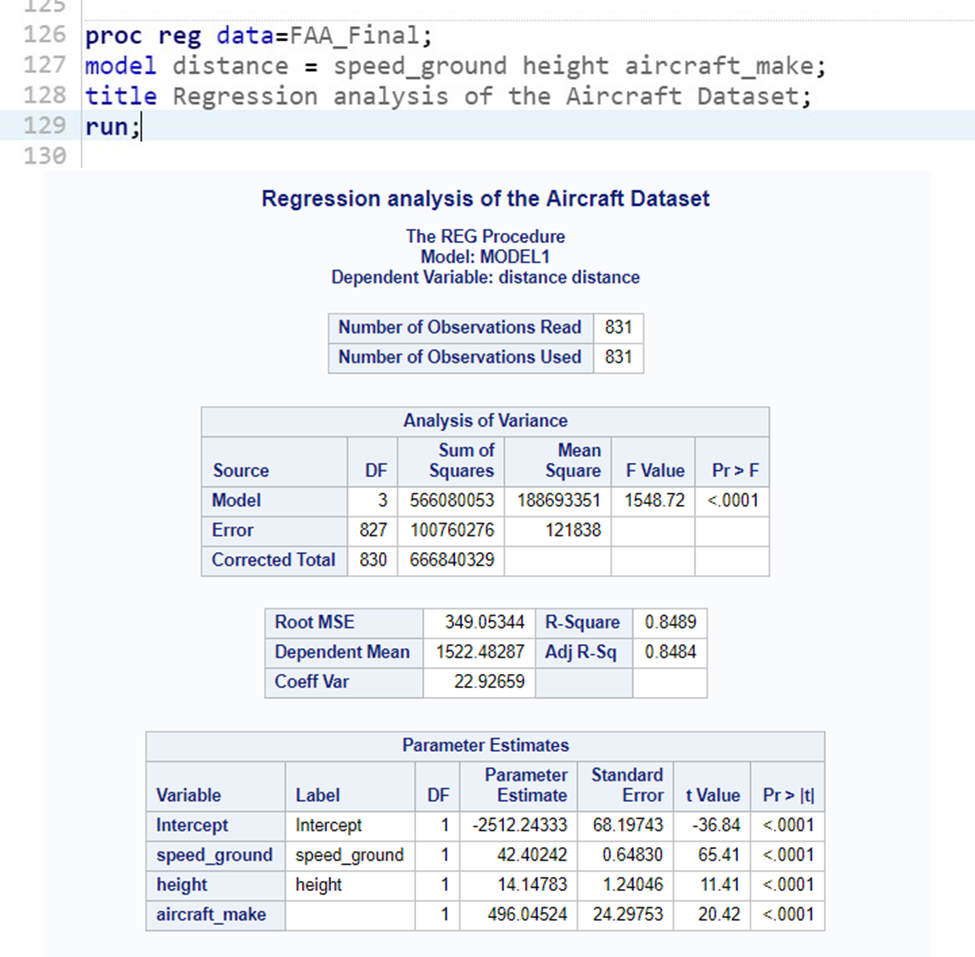


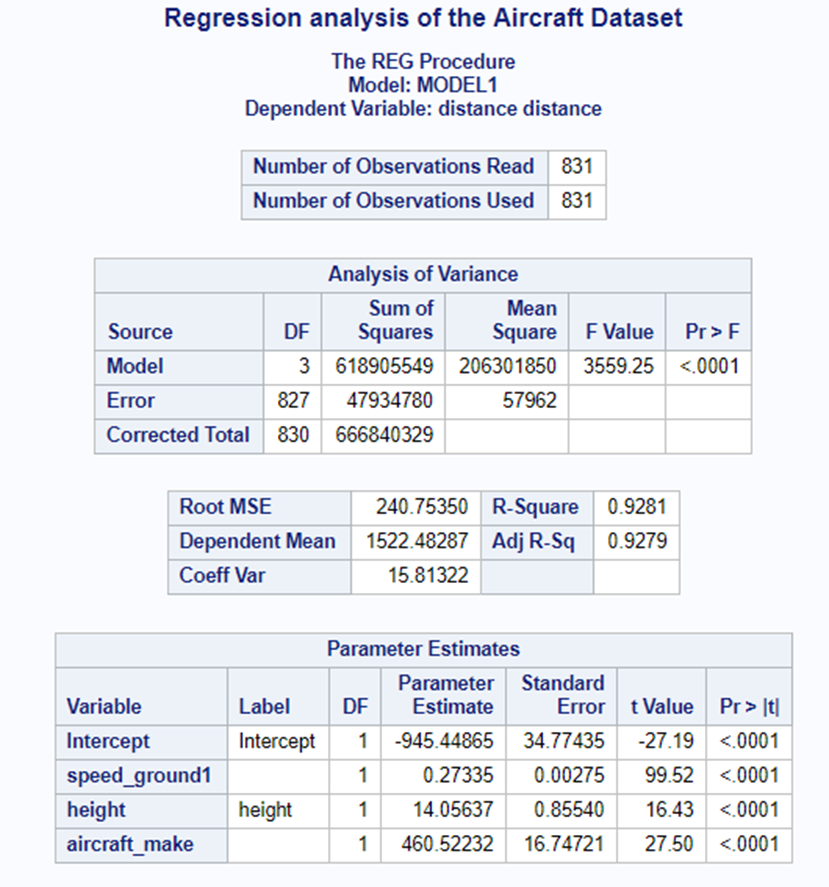


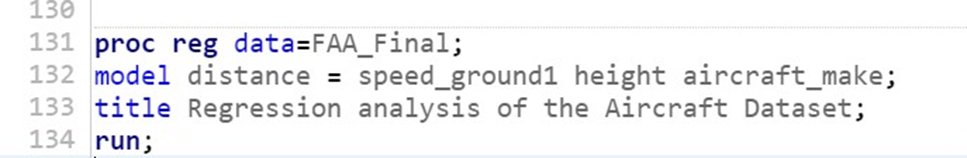
From the parameter estimates, we observe that p-values for duration, no\_pasg and pitch are greater than 0.05. Therefore, all of them are dropped from the model equation since they are statistically insignificant. We can also drop speed\_air from our model since we observed in the previous chapter that it has high collinearity with speed\_ground as well as speed\_ground1.

**Step 2:** Run the regression analysis again after removing statistically insignificant variables identified in Step 1.

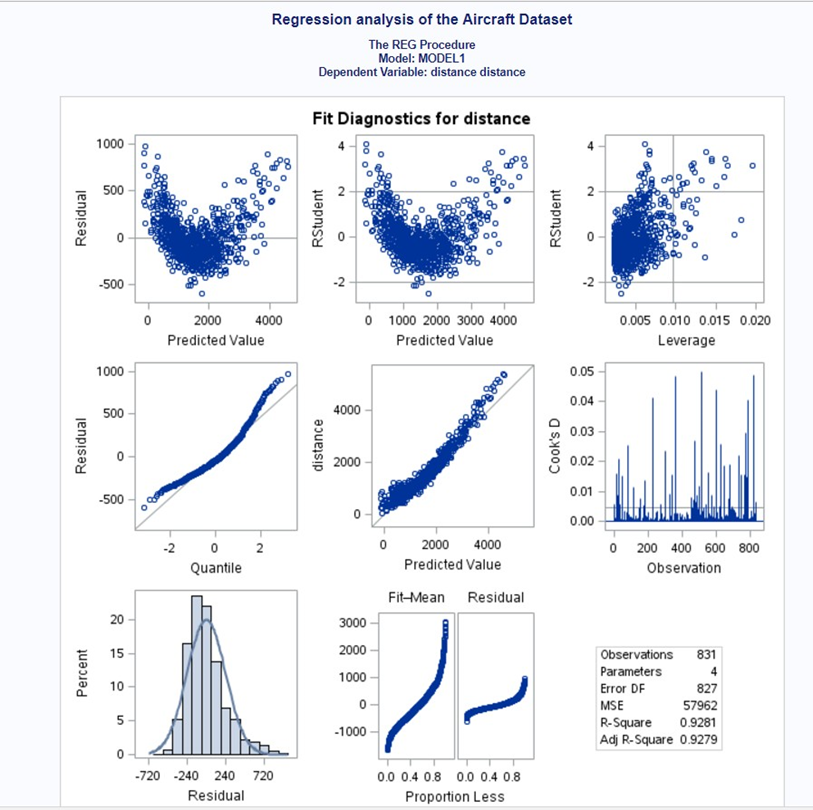
We also know that speed\_ground and speed\_ground1 have high collinearity. So, we will keep one of the two independent variables in our model. We will run two regressions, one with speed\_ground and other with speed\_ground1, along with other independent variables such as aircraft\_make height and choose the best model between the two.

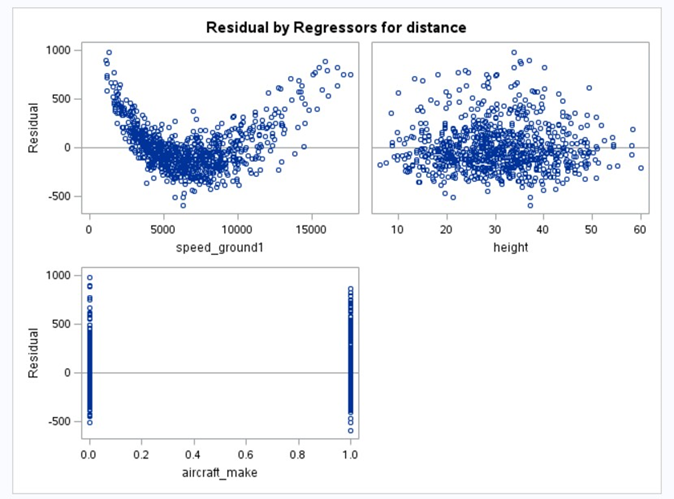






Comparing the R-Square and Adj R-sq values , we choose the model containing independent variable speed\_ground1 since we can predict the value of our dependent variable with greater variability (93 percent compared to 85 percent) and less noise ( 7 percent compared to 15 percent).





**Conclusion**

1. **How many observations (flights) did you use to fit your final model?**

**Ans:** We used 831 observations to fit our final model. The rest 119 observations were removed from our data set in the various data prep steps taken above such as removing duplications and missing values, performing completeness check and removing abnormal values.

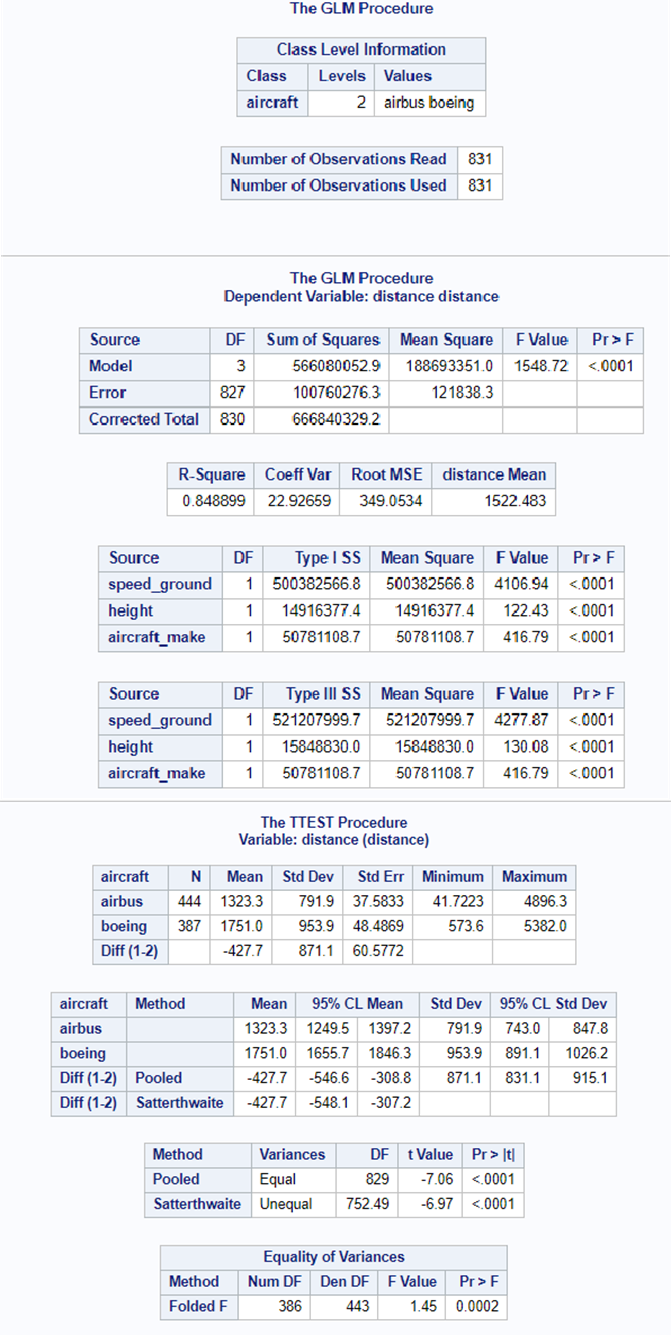
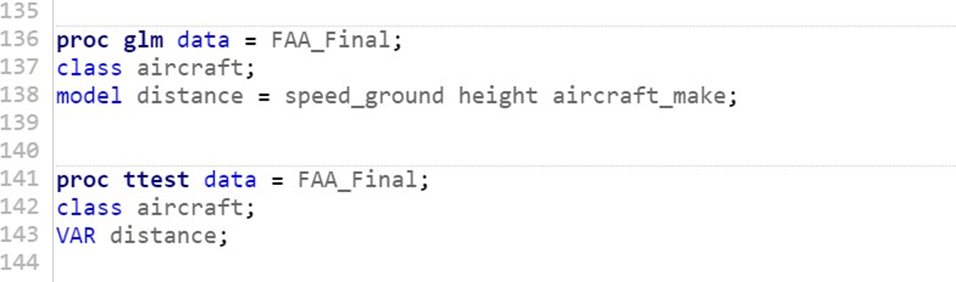
1. **What factors and how they impact the landing distance of a flight?**

**Ans:** As seen from the model equation, landing distance depends on three factors:

* Speed\_ground1, i.e. square of speed\_ground: There will be an increase of 0.27 units in landing distance for every unit increase in square of speed\_ground, keeping all other parameters constant.
* Height: There will be an increase of 14.05 units in landing distance for every unit increase in height, keeping all other parameters constant.
* Aircraft\_make: For Boeing make aircraft the landing distance would be 460.52 units greater than that for Airbus make aircraft, keeping all other parameters constant.

1. **Is there any difference between the two Boeing and Airbus?**

**Ans:** Yes, there is a significant difference between the two makes of aircraft. Landing distance has a direct correlation with aircraft make. We performed the GLM and T test to study the difference in the populations of the two makes and their impact on landing distance.



Since F-value is greater than 1, we can say that there is a significant difference between the two aircraft makes. We observe that there is a difference in the mean, standard deviation, variation and impact on landing distance of the two makes.

