## Summary

We have been given the flight data which consists of different attributes which may or may not be important for the analysis. The data is divided into two excel sheets which needs to be imported in SAS and analyzed. Our main goal is to study what factors and how they would impact the landing distance of a commercial flight. Landing data consists of landing distance and other parameters from 950 commercial flights. The data is simulated and not real.

We have done in the analysis in different sequential steps which consists of the following:

1. Data exploration and data cleaning
2. Data visualization
3. Modeling
4. Model checking
5. Re-exploration of data
6. Re-modeling
7. Finalizing the model

We will parse through each of the above steps in the report which will eventually lead us to the culmination of the report.

**The DATA**

**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.

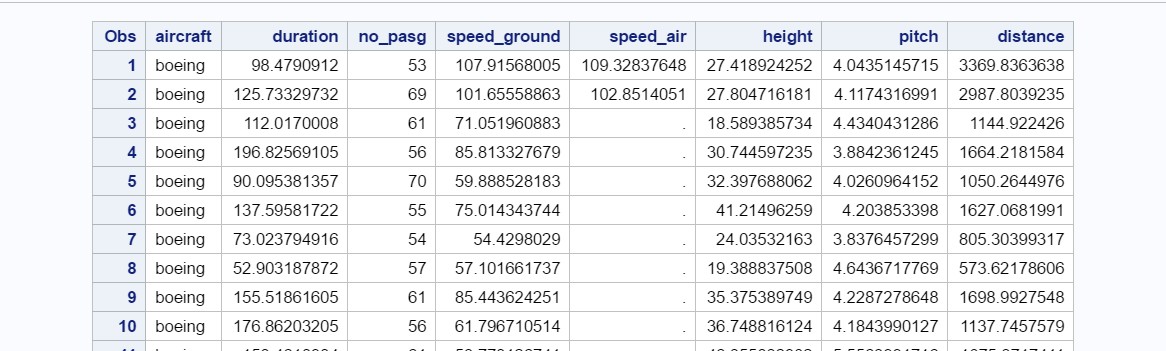
**Distance** (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

## DATA CLEANING

Data cleaning, deals with detecting and removing errors and inconsistencies from data to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in this case we are provided with data in two excel sheets FAA1 and FAA2 which share the same columns except for **duration** which is present in FAA2.xls only, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. To provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.

So, the first task was to import the data of both the sheets and combine them.

## Importing and Combining the data



**Code to Import the FAA1 excel file**

FILENAME REFFILE '/home/dassm0/FAA1.xls';

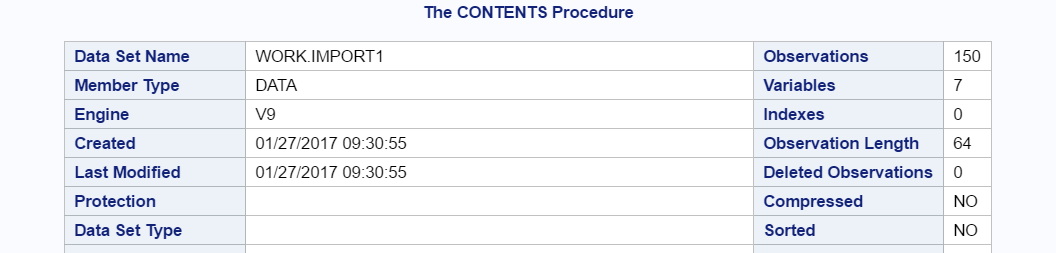
%web\_drop\_table(WORK.IMPORT); PROC IMPORT DATAFILE=REFFILE

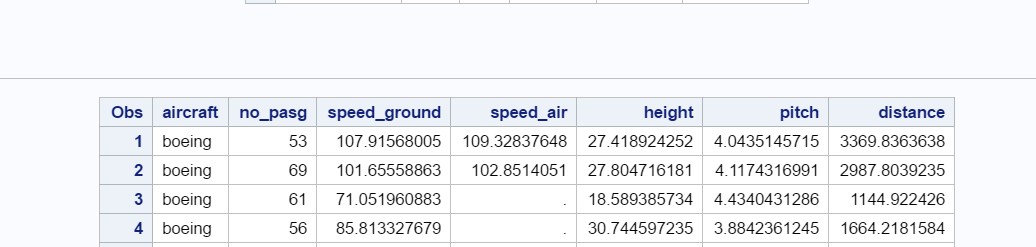
DBMS=XLS OUT=WORK.IMPORT; GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

Proc print;

**Observation:** As you can see above the FAA1.xls contains 800 records.



**Code to Import the FAA2 excel file**

FILENAME REFFILE1 '/home/dassm0/FAA2.xls';

PROC IMPORT DATAFILE=REFFILE1 DBMS=XLS OUT=WORK.IMPORT1; GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT1; RUN;

Proc print;

**Observation:** The FAA2.xls contains 150 records.

After we combine them using the SET function we have 950 records in total. Below is the code to combine both the excel sheets.

**Code**

Data combined;

set WORK.IMPORT WORK.IMPORT1; RUN;

proc print

## Data Cleaning:

Now, we have all the data we require to do the statistical Analysis but the first and the foremost step is cleansing the data.

We have been given a few **parameters** in the basis of which we can clean the data. I have tried cleaning the data with the given parameters and below are by observations:

* 1. **Duration** (in minutes): Flight duration should always be greater than 40min.

**Observation:** I tried deleting the rows with duration less than 40 mins, I was left with just 795 records. This means that there will be a loss of 155 records if we use duration as a parameter to cleanse the data. I have decided not to proceed with same.

* 1. **Speed\_ground:** Speed in Ground should be greater than 30 mph and greater than 140 mph.

**Observation:** After applying the required filter there were five records which didn’t fit the above parameters and were deleted.

* 1. **Speed\_Air:** Speed in Ground should be greater than 30 mph and greater than 140 mph. If this filter is applied I will just be left with 237 of 950 records in my dataset. Therefore, I am not using this parameter to filter the records.
  2. **Height**: The landing aircraft is required to be at least 6 meters high at the threshold of the runway. If this filter is applied there’s a loss of 12 records. Even when I consider the height to be less than 5 meters then also there is a loss of 12 records only. It was only when I kept the minimum height of 4meters there was a loss of 10 records which is a minute change from the previous result.
  3. **Regarding the Duplicates:** I can also see some duplicate records which can lead to inaccuracy in my data analysis. I have proceeded by deleting the duplicate records using **NODUPKEY.**

**Code for the above parameters:**

Data combined;

set WORK.IMPORT WORK.IMPORT1;

if height<4 then delete;

if ((speed\_ground < 27) or (speed\_ground > 145)) then delete; RUN;

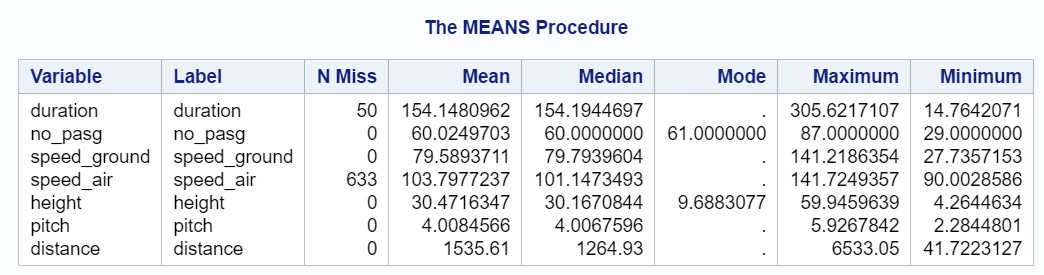
PROC SORT data=combined NODUPKEY

BY aircraft no\_pasg speed\_ground height pitch distance; RUN;

**Observation:** If I applied all the above filters or parameters for data cleaning than I am only left with 841 records out of 950. **Now, I am performing the statistical analysis on the combined data set after removing the undesired records.**

## Statistical Analysis

For the statistical Analysis I am using the **Means** procedure which will perform the basic descriptive analysis for the data sets. Below is the code and output for the same. Here, the dataset is comprised of 841 records in total. This is just the basic summarized information for each variable. Also, I am using the **Univariate** procedure to generate the Histogram for the attributes like speed\_ground, height etc.



**Code:**

data simulation; set combined;

proc univariate data=simulation noprint; histogram;

Proc means data=combined n nmiss mean median mode max min; run;

**Output:**

**Observation-** From the above output we can see that speed air contains the maximum amount of missing values which is 633. This can highly impact our analysis as missing data can be misleading at times.The minimum and the maximum values can vary significantly due to the missing values.

Percent

**Distribution of duration**

20

15

10

Percent

5

0

20 40 60 80 100 120 140 160 180 200 220 240 260 280 300

duration

**Distribution of no\_pasg**

25

20

15

10

5

0

30 34 38 42 46 50 54

58

no\_pasg

62 66 70 74 78 82 86

Percent

**Distribution of pitch**

25

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |
|  |  |  |
|  |  |

20

15

Percent

10

5

0

2.25 2.55 2.85 3.15 3.45 3.75 4.05 4.35 4.65 4.95 5.25 5.55 5.85

pitch

**Distribution of speed\_ground**

20

15

10

5

0

28 36 44 52 60 68 76 84 92 100 108 116 124 132 140

speed\_ground

**Distribution of speed\_air**

40

30

20

10

0

93

99

105

111

117

speed\_air

123

129

135

141

**Distribution of height**

20

15

10

5

0

6 10 14 18 22 26 30 34 38 42 46 50 54 58

height

Percent

Percent

**Observation:** From the above histograms we can infer that all the parameters in the Flight Data except **speed\_air** follow a Normal distribution.

## Correlation Analysis

Correlation analysis is the process of studying the strength of that relationship with available statistical data. It determines whether a relationship between two variables is present, and how strong it might be. We have used the CORR procedure to find the correlation between the different attributes. Below is the code for the same

## Code

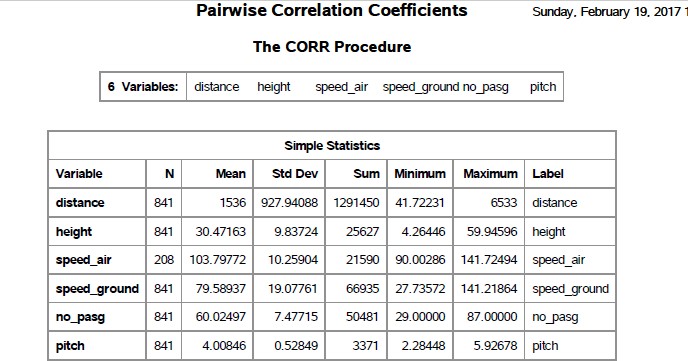
data simulation; set combined;

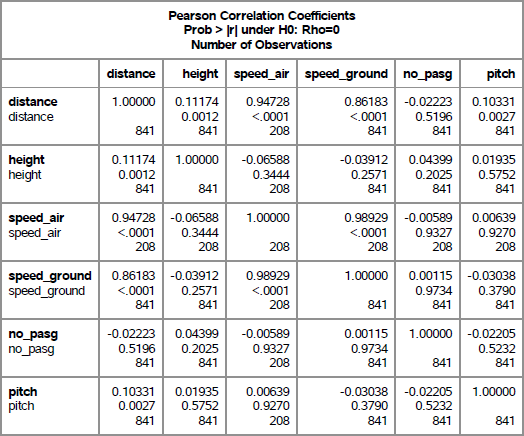
proc corr data=simulation;

var distance height speed\_air speed\_ground no\_pasg pitch; title Pairwise Correlation Coefficients;

run;

## Output





**Observation:** We can see that the landing distance is strongly correlated with speed\_ground and it is weakly correlated with the no\_pasg. This shows that the number of passengers in the flight effects the landing distance minimially. Also we can see a strong correlation between speed\_air and speed\_ground.

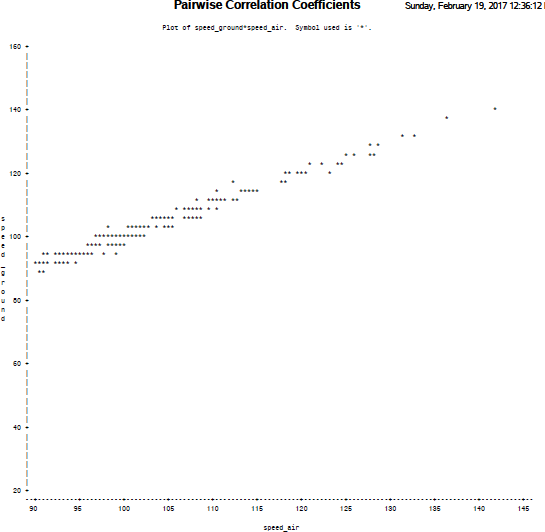
## Plot Procedure

We will now plot a graph between Speed\_air and Speed\_ground using the procedure plot.

## Below is the code:

proc plot data=simulation;

plot speed\_ground\* speed\_air='\*' ; run;



The above graph further testifies that the speed\_air and speed\_ground are closely related. Also, while using the proc means we found out that the speed\_air has a lot of missing values.

Therefore, we will not use speed\_air in further analysis.

## Regression Analysis

The next and one of the vital step is to perform the regression analysis. The main goal is to find model that describes the relationship between a response variable and several predictor (explanatory) variables. Here the response variable is Landing Distance and the remaining attributes are our predictors. Since the **no\_pasg** is weakly correlated with the landing distance, we will not use this as a predictor to build our model.

Also since we have been asked to find if there is any diffence between the Airbus or the boeing aircraft we will segregate the data for boeing and airbus respectively.

## Code to Segregate the Data

data simulation1 simulation2 ; set combined;

if(aircraft='boeing') then output simulation1; if(aircraft='airbus') then output simulation2; run;

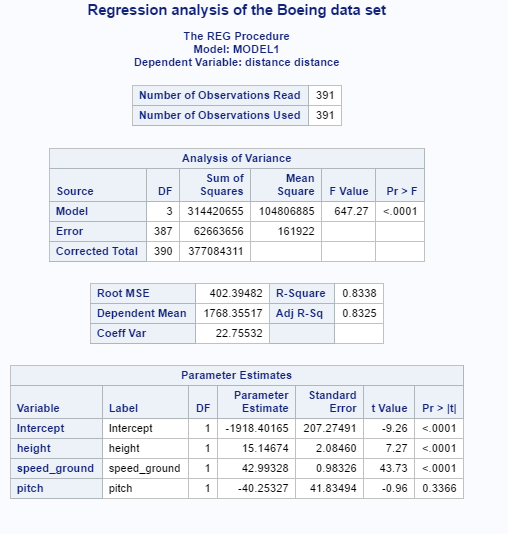
## Below is the code to build our model and perform the regression Analysis for the Segregated Data

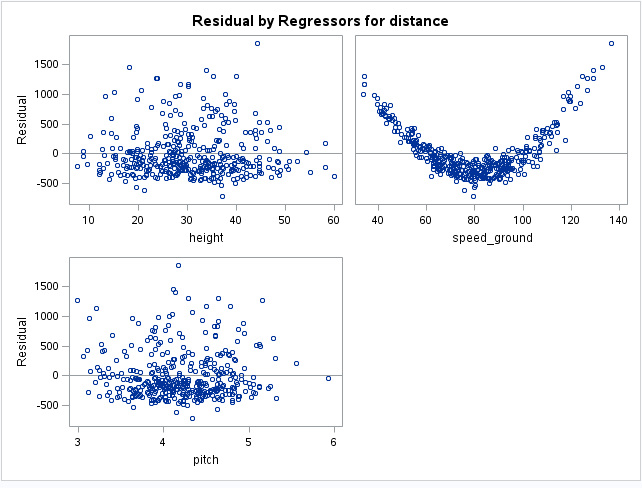
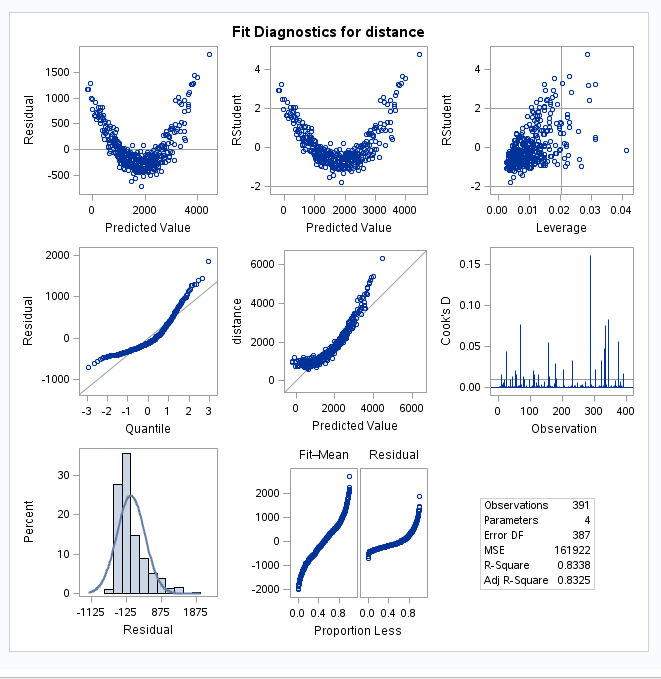
**Regression Analysis on Airbus Data**

data simulation1 simulation2 ; set combined;

if(aircraft='boeing') then output simulation1; if(aircraft='airbus') then output simulation2; run;

## Output



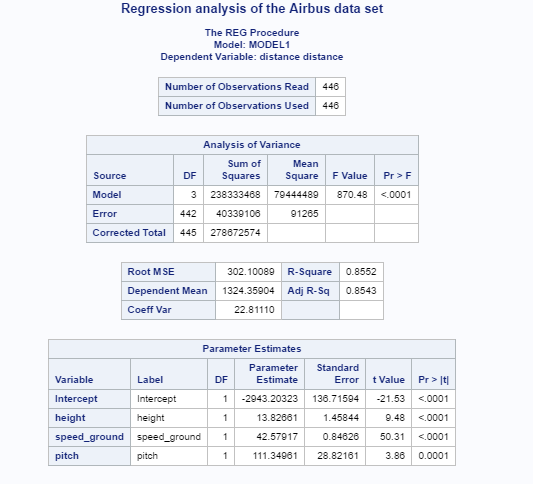


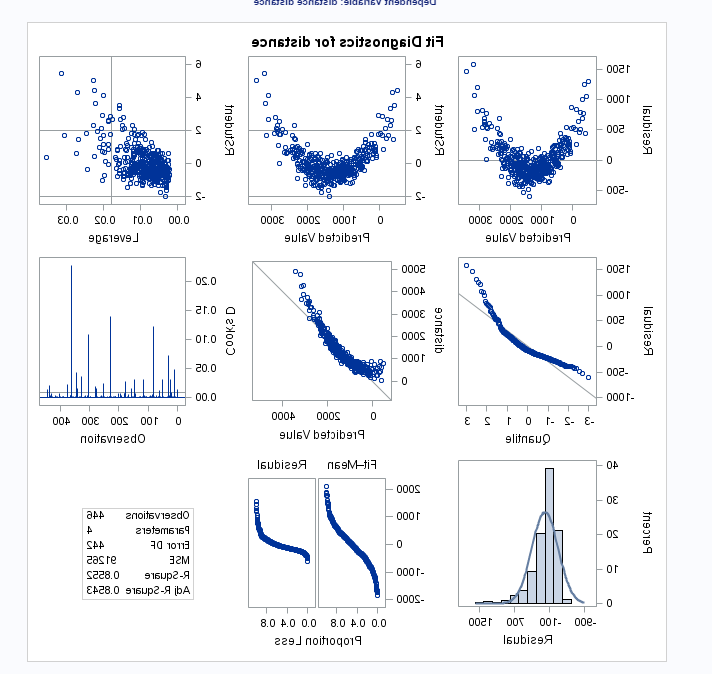
**Regression Analysis on Airbus Data**

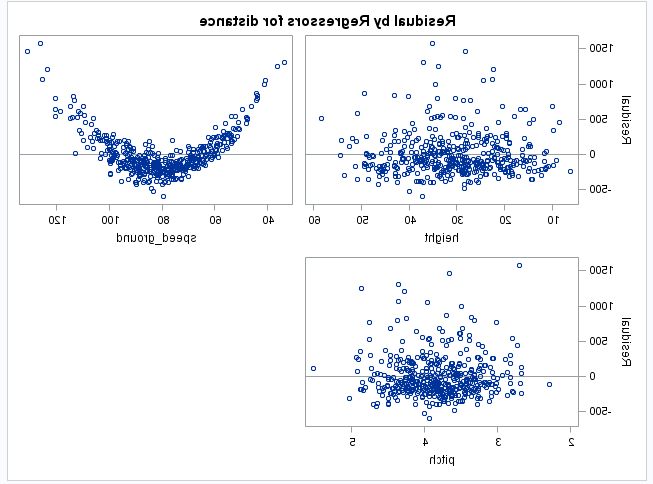
data simulation1 simulation2 ; set combined;

if(aircraft='boeing') then output simulation1; if(aircraft='airbus') then output simulation2; run;

## Output







**Observation**

**Analysis of variance**

**Airbus Data Set and Boeing Data Set**

If we look that the p-value of the **model**, we find that it is smaller than 0.05. Therefore, we will reject our null hypothesis that all the model coefficients are 0. Also, we can concluded that the model is statistically significant.

**The R Square and the adjusted R square**

* 1. quared is the proportion of variance in the dependent variable (distance) which can be explained by the independent variables (speed\_ground). This is an overall measure of the strength of association and does not reflect the extent to which any independent variable is associated with the dependent variable. The r-squared value for Boeing and Airbus data set is 83 percent and 85 percent respectively, which needs to be further improved. For better results we use the adjusted R for understanding model accuracy since it penalizes the addition of extraneous predictors to the model.

**Pr > |t|**- This column shows the 2-tailed p-values used in testing the null hypothesis that the coefficient (parameter) is 0. Using an alpha of 0.05:

**Airbus Data Set**

**All the predictors used in building the model have p**-value less than 0.05. Therefore, they are all statically significant.

**Boeing Data Set**

**All the predictors except pitch** used in building the model have p-value less than 0.05. Therefore, except pitch statically significant. While rebuilding the model, we will drop pitch as a predictor.

**Fit diagnostics**-

**Here we are concerned about the fit of the model**. For this reason, we will analyze the various residual plots. A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

The residual vs. Predicted value plot shows a slightly suboptimal fit but will still give you a good general sense of the relationship, even if it’s not perfect. Also, the residual vs. Quantile plot is slightly non-linear which again denotes suboptimal fit.

If we compare the residual plot with different repressors or distance, we find that all the plots are randomly distributed except the plot with speed. This leads us to conclude the reason of the pattern observed in the graph between the residual and the predicted value. Sometimes patterns like this indicate that some variable needs to be transformed. There’s room for improvement in our model.

**Transformation**

As decided we want a better fit for our model. Therefore, we will transform the speed\_ground by cubing it and check if the new generated model has a better fit.

Also, we will check if the adjusted R value has improved or not after the transformation of the attribute.

**Below is code for transformation followed by regression For Airbus Dataset**

**data sim1;**

**set simulation2; speed=speed\_ground\*speed\_ground\*speed\_ground; drop speed\_ground;;**

**proc reg data=sim1;**

**model distance= height speed pitch ;**

**title Regression analysis of the Airbus data set; run;**

**For Boeing Dataset data sim1;**

**set simulation1;**

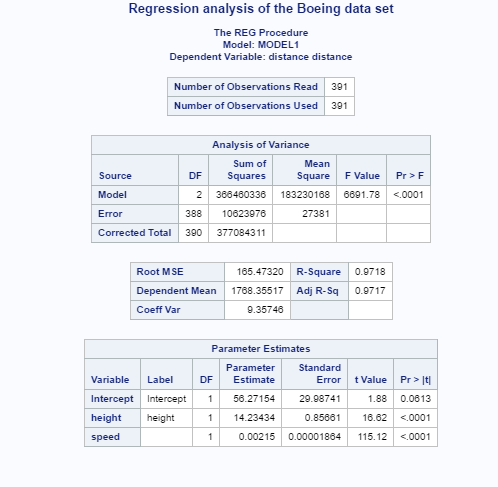
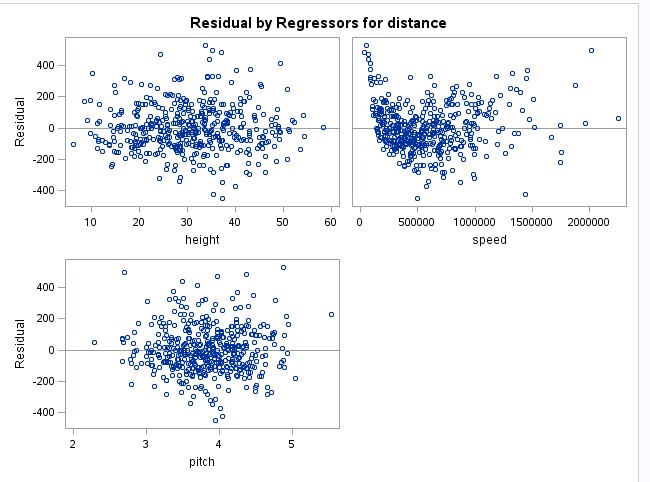
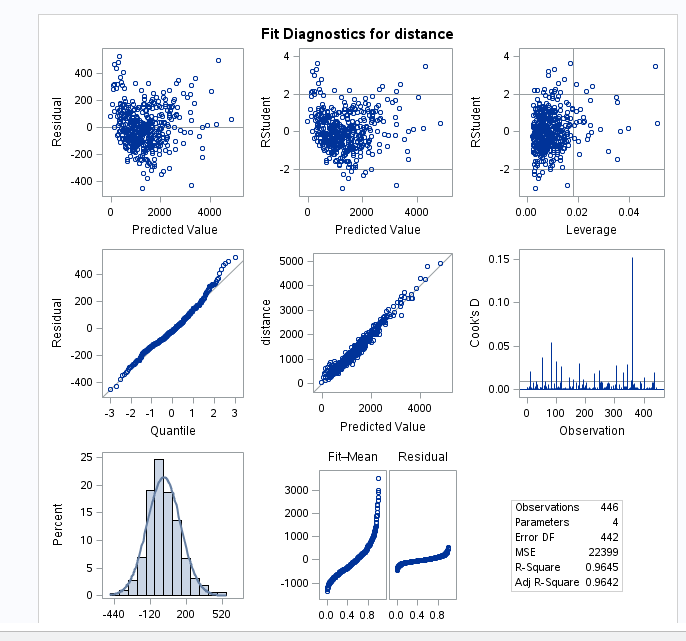
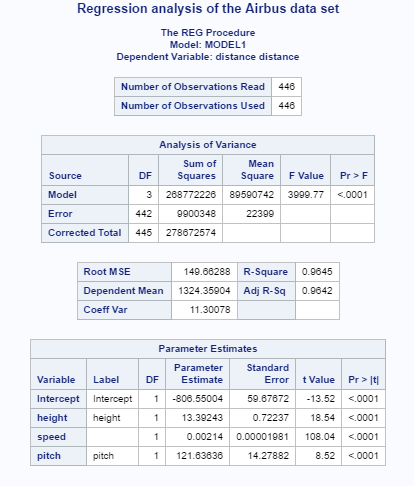
**speed=speed\_ground\*speed\_ground\*speed\_ground; drop speed\_ground;;**

**proc reg data=sim1;**

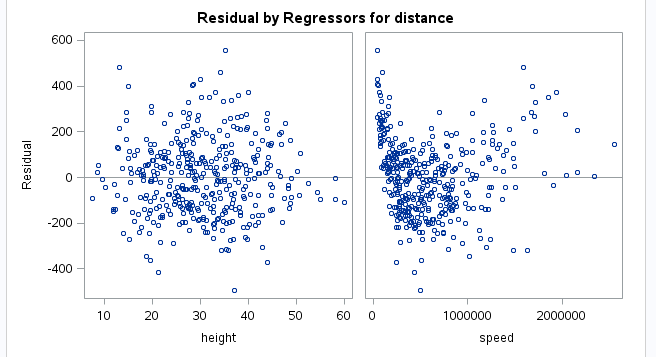
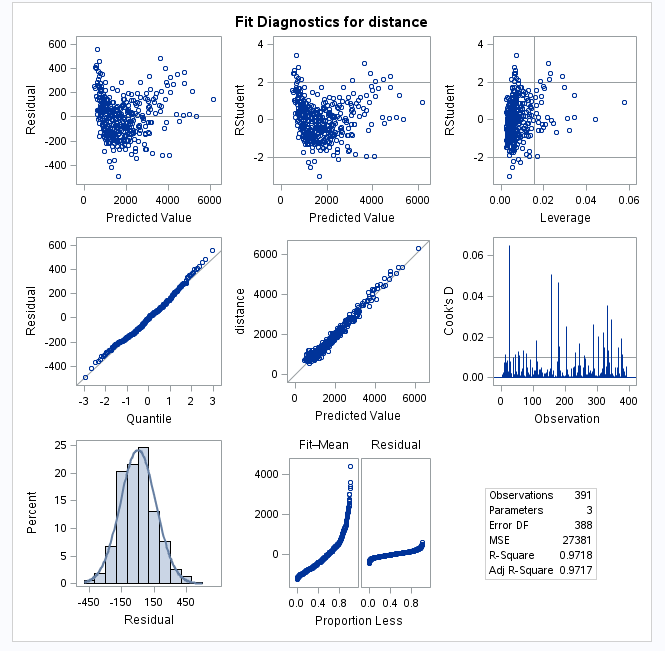
**model distance= height speed ;**

**title Regression analysis of the Airbus data set; run;**

**Output for Airbus DataSet**



**Output for Boeing Dataset**



## Observations

A number of improvements can be seen in both the models after the transformation of one of the predictor. They are

* + - The p-value is less than 0.05 for all the predictors. Therefore all the predictors are statistically significant for determing the model.
    - The adjusted-R vaue has significantly improved and both the models have an improved accuracy.
    - The plot of residual withits p all redictors is randomly scattered now.
    - If we look at the fit diagnostics of both the model The residual vs. Predicted value plot is scattered now and shows a nearly optimal fit.The normal distribution of residual is much better than before.

## Questions

1. **How many observations (flights) do you use to fit your final model? If not all 950 flights, why?**

There were a total of 950 records earlier out of which we were only left with 837 records after the required Data cleaning. In the data cleaning process we removed the duplicates after the combing the two datasheets. We also removed a few of the records based on the requirements of the user.

## What factors and how they impact the landing distance of a flight?

As shown above, I have created two models based on the different aircraft types. The factors that impact the landing distance are mainly height, speed\_ground and pitch(*not for the boeing aircrafts*). Amongst all the factors the speed\_ground hass a greataer impact on landing distance than the remaining factors.Therefore, if we have a desired speed\_ground it will positively reflect in the landing distance of the aircraft.

## Is there any difference between the two makes Boeing and Airbus?

Yes, there is a significant difference between the the two makes. I created two regression models to see if there is any difference between two models. I found out that pitch is statiscally insignificant in determing the landing distance of Boeing Aircraft while does impact the landing distance of the Airbus aircrafts.