**Problem Statement –**

We have two datasets which are in the form of excel sheets named FAA1.xlsx and FAA2.xlsx. Both of these data sets contain the data about the flight landings. Both datasets have little variance.

Our problem statement is to study the factors which impact the landing conditions of a commercial flight with the 950 records of data.

**The given data and conditions are as follows –**

**FAA1** has 8 variables with total data count of 801 rows.

The variables are -

Aircraft, duration, no\_pasg, speed\_ground, speed\_air, height, pitch, distance

**FAA2** has 7 variables with total count of 151 rows.

**The variables are –**

Aircraft, no\_pasg, speed\_ground, speed\_air, height, pitch, distance

So, when we compare both the datasets, FAA2 has one variable missing.

Variables and its descriptions –

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 Preparation and Cleaning -**

**Step 1 –**

We will import both the datasets in SAS.

proc import out=faa1 datafile='/folders/myfolders/ecprg193/FAA1.xls'

dbms = xls replace;

getnames = yes;

run;

proc import out=faa2 datafile='/folders/myfolders/ecprg193/FAA2.xls'

dbms = xls replace;

getnames = yes;

run;

**Analysis on first glance is as follows –**

1) Duration column in dataset faa2 is not present.

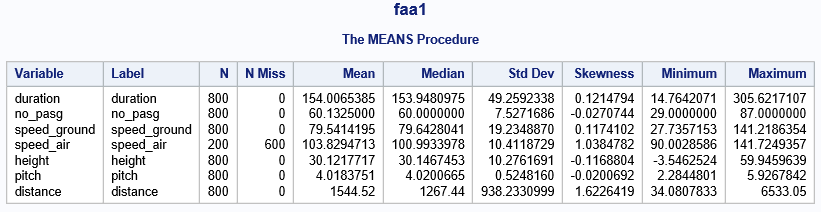
2) The first 101 datasets of faa2 are same as that of faa1 except for the duration column.

proc means data=faa1 n nmiss mean median std skew min max;

title 'faa1';

run;

The below table shows the characteristics of the FAA1 -

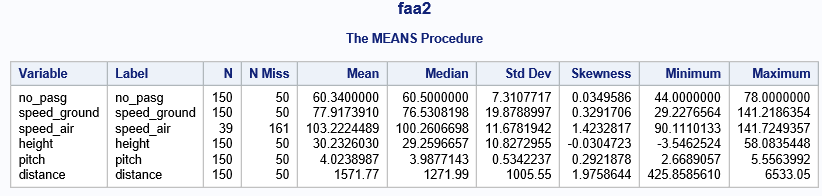


proc means data=faa2 n nmiss mean median std skew min max;

title 'faa2';

run;

The below table shows the characteristics of the FAA2 -



The above data can be used to show that –

1. All the variables except distance has a normal distribution as we can see that the mean and median are nearly equal for both data sets.
2. First data set has 600 missing values for speed\_air variable.
3. Similarly, there are missing values for each variable in faa2 dataset.
4. The minimum and maximum for the variables are comparable in both the data sets, so should not contribute to a big change after combining.

**Step 2 –**

We will combine records from both the datasets.

data combined\_faa;

set faa1 faa2;

data combined\_faa;

set combined\_faa;

if missing(coalesceC(of \_character\_)) then delete;

run;

After this we get 950 rows of data in a new data set named combined\_faa.

**Step 3 –**

After combining, we will remove all the duplicate fields from the combined dataset.

We had 100 rows of data in faa2 which had same data as in faa1. So, those should be removed as duplicates.

proc sort data=combined\_faa nodupkey;

by aircraft no\_pasg speed\_ground speed\_air height pitch distance;

run;

After this step, we are left with 850 rows of data.

**Step 4 –**

Now we have been provided with conditions where the data is abnormal. So, we should preferably remove those records before analyzing.

The conditions mentioned are as follows –

Duration (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.

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.

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.

**Abnormal conditions –**

**Duration<40min**

**140 MPH < Speed\_ground < 30 MPH**

**140 MPH < Speed\_air < 30 MPH**

**Height < 6 meters**

**Distance > 6000 feet**

We should remove all the records with the above abnormal conditions.

We will create three groups which will be –

0 – Abnormal Values

1 – Missing Values

2 – Normal Values

data combined\_faa\_v1;

set combined\_faa;

if speed\_ground=. or speed\_air=. or height=. then group=1;

else group=2;

if speed\_ground ne . and (speed\_ground lt 30.0 or speed\_ground gt 140.0) then group=0;

if speed\_air ne . and (speed\_air lt 30.0 or speed\_air gt 140.0) then group=0;

if height ne . and height<6 then group=0;

if distance ne . and distance > 6000 then group=0;

run;

Now we will create a new dataset combined\_faa\_v2 which will not have abnormal values.

Deleting the abnormal values by removing group 0 values.

data combined\_faa\_v2;

set combined\_faa\_v1;

if group=0 then delete;

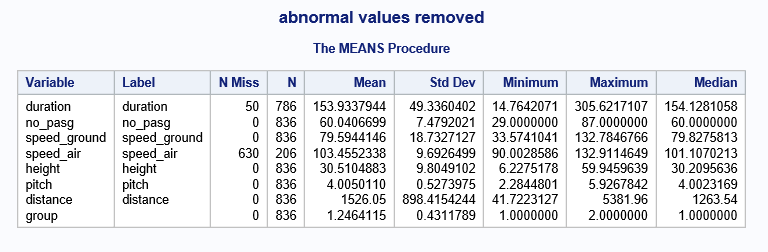
else output combined\_faa\_v2;

run;

Now we are left with 836 rows of data with 8 variables.

**Step 5 –**

Let us analyze the data available after we removed the abnormal values –



From the above analysis, we can see that –

1) Duration has 50 missing values.

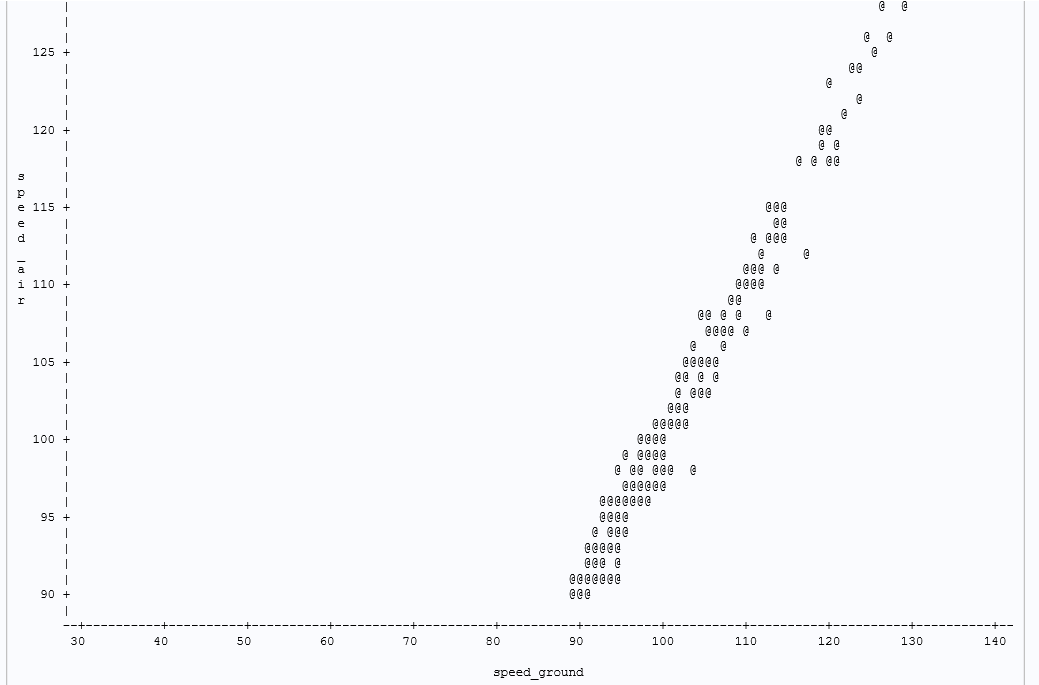
2) Speed\_air has 630 missing values.

Let us analyze both of these variables.

proc plot data=combined\_faa\_v2;

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

run;

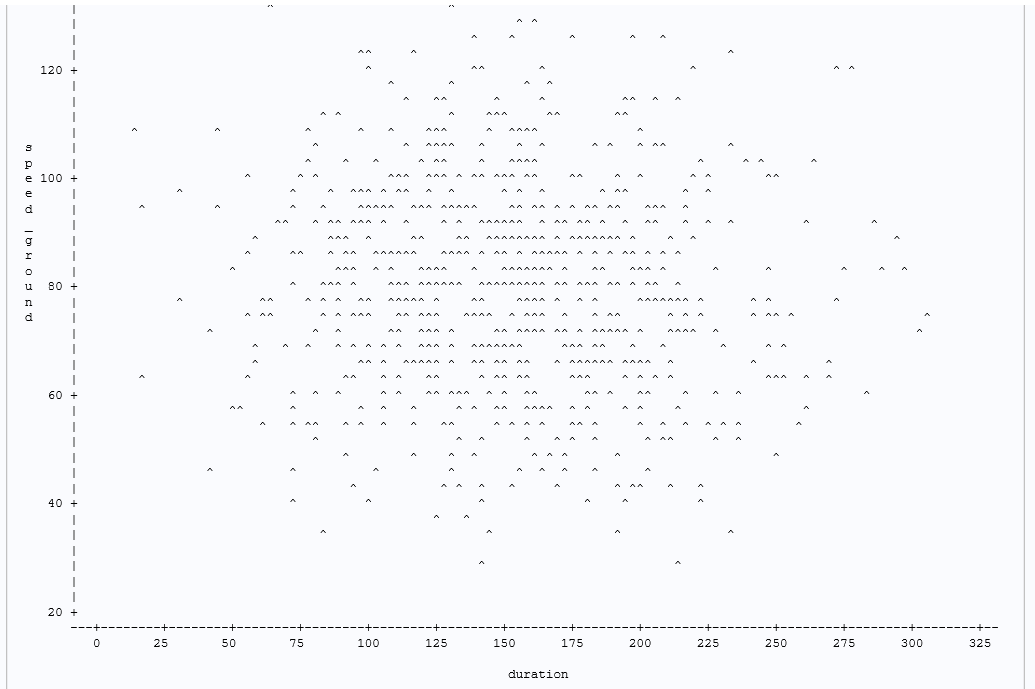


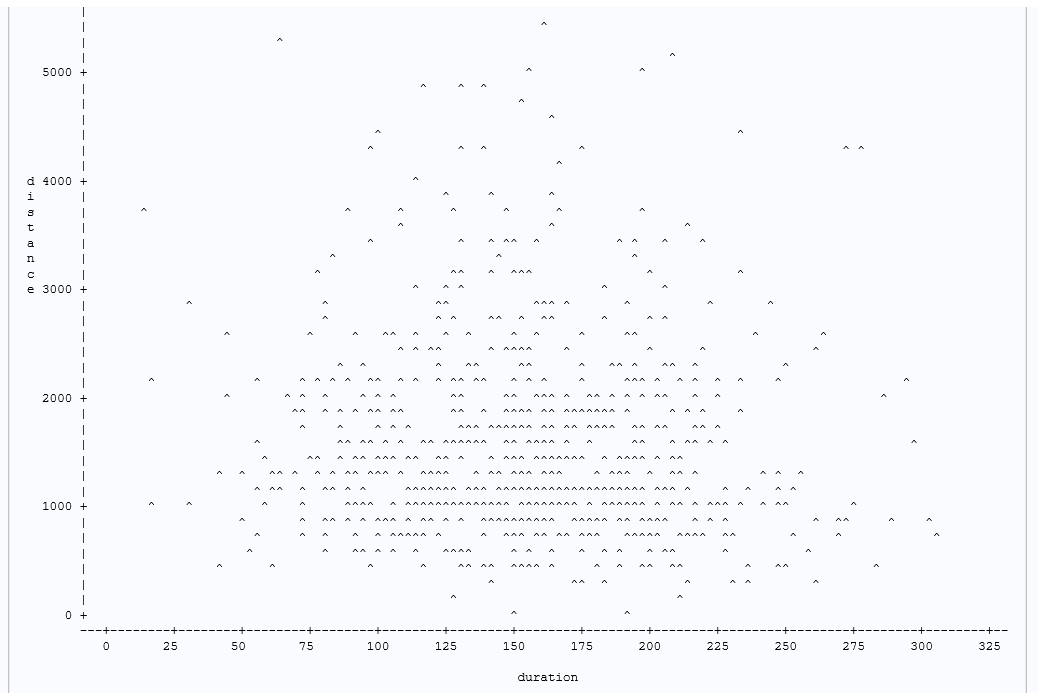
From the above plot, we can see that the speed\_ground and speed\_air are correlated. So, even if we remove this variable we can do the analysis. Moreover, we have 630 values missing, so it is better to drop the speed\_air variable.

proc plot data=combined\_faa;

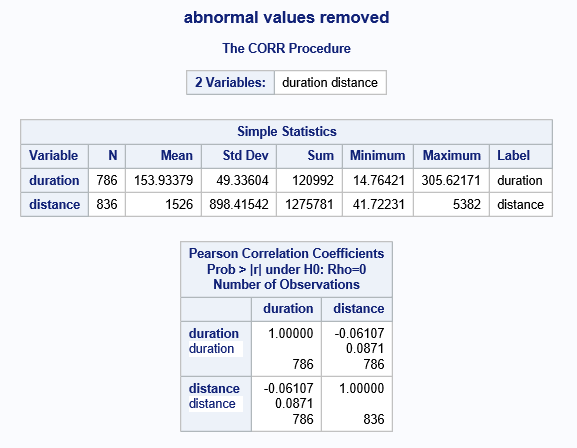
plot speed\_ground\*duration='^';

run;





Here we see no relation of duration with speed\_ground and distance. We plan to check the correlation.



There seems to be no correlation between duration and distance. But as we have only 50 records missing, we plan to move ahead with this variable.

So, after the data preparation and cleaning we have the following dataset-

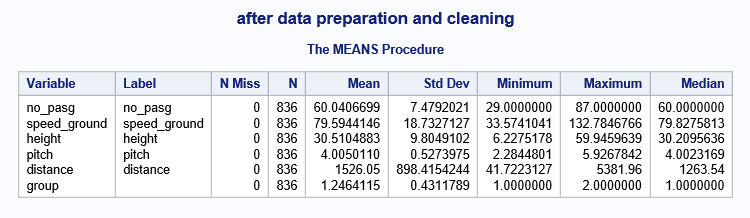
data combined\_faa\_v3;

set combined\_faa\_v2;

drop speed\_air;

run;

**This dataset has 836 records.**



**Data Exploration –**

In this section, we will try to gain insights from the clean data available to us. We will analyze individual variables and check its relations with other variables. This will help us in understanding the variation in data and how we can implement a modelel.

**1)Univariate Analysis –**

no\_pasg -



| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 836 | **Sum Weights** | 836 |
| **Mean** | 60.0406699 | **Sum Observations** | 50194 |
| **Std Deviation** | 7.47920208 | **Variance** | 55.9384637 |
| **Skewness** | -0.0103497 | **Kurtosis** | 0.30478069 |
| **Uncorrected SS** | 3060390 | **Corrected SS** | 46708.6172 |
| **Coeff Variation** | 12.4568931 | **Std Error Mean** | 0.25867361 |

The above table shows us that the variables no\_pasg is nearly a normal distribution. The mean, median and mode are nearly equal. We have some outliers but we cannot remove those because it can have an impact on analysis.

speed\_ground –



| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 836 | **Sum Weights** | 836 |
| **Mean** | 79.5944146 | **Sum Observations** | 66540.9306 |
| **Std Deviation** | 18.7327127 | **Variance** | 350.914524 |
| **Skewness** | 0.08550087 | **Kurtosis** | -0.2394114 |
| **Uncorrected SS** | 5589300.05 | **Corrected SS** | 293013.627 |
| **Coeff Variation** | 23.53521 | **Std Error Mean** | 0.64788441 |

This variable also looks nearly normal as the mean, median and mode are nearly equal. Though some outliers are present, we may not remove the outliers.

height -



| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 836 | **Sum Weights** | 836 |
| **Mean** | 30.5104883 | **Sum Observations** | 25506.7682 |
| **Std Deviation** | 9.8049102 | **Variance** | 96.1362641 |
| **Skewness** | 0.12917906 | **Kurtosis** | -0.3340755 |
| **Uncorrected SS** | 858497.735 | **Corrected SS** | 80273.7805 |
| **Coeff Variation** | 32.1361956 | **Std Error Mean** | 0.33910991 |

This variable also looks nearly normal as the mean, median and mode are nearly equal. Though some outliers are present, we may not remove the outliers.

pitch -



| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 836 | **Sum Weights** | 836 |
| **Mean** | 4.00501101 | **Sum Observations** | 3348.1892 |
| **Std Deviation** | 0.52739751 | **Variance** | 0.27814813 |
| **Skewness** | 0.00858989 | **Kurtosis** | -0.0889377 |
| **Uncorrected SS** | 13641.7883 | **Corrected SS** | 232.253692 |
| **Coeff Variation** | 13.168441 | **Std Error Mean** | 0.01824042 |

This variable also looks nearly normal as the mean, median and mode are nearly equal. Though some outliers are present, we may not remove the outliers because it may have some hidden information.

distance -



| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 836 | **Sum Weights** | 836 |
| **Mean** | 1526.0539 | **Sum Observations** | 1275781.06 |
| **Std Deviation** | 898.415424 | **Variance** | 807150.275 |
| **Skewness** | 1.46471488 | **Kurtosis** | 2.48355071 |
| **Uncorrected SS** | 2620881139 | **Corrected SS** | 673970479 |
| **Coeff Variation** | 58.8718017 | **Std Error Mean** | 31.0723472 |

This is a non-normal distribution which does not have mean, median and mode nearly equal. This can be a lognormal distribution.

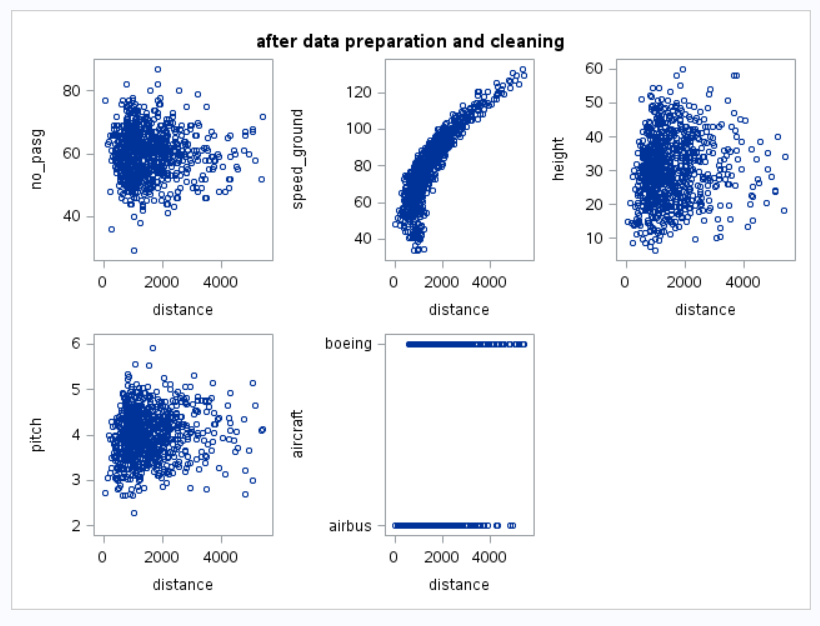
**2) Bivariate Analysis –**

We find the bivariate analysis, which is the relation of one variable with another variable.

proc sgscatter data=combined\_faa\_v3;

plot(no\_pasg speed\_ground height pitch aircraft)\*distance;

run;



From this we can see that there is a linear and positive relationship between distance and speed\_ground. But if we want to increase the linearity we can try to transform any of the variable. The only variable which is nearly linear is speed\_ground, so we should try transformation on that variable.

Before doing that let us check the correlations of present variables.

**3) Correlation –**

Now we will find the correlation between various variables to check the type of relation between them.

For that, first of all we will convert our aircraft to numeric values so that we can find the correlation with distance.

data combined\_faa\_v3;

set combined\_faa\_v3;

if aircraft="boeing" then aircraft\_type=0;

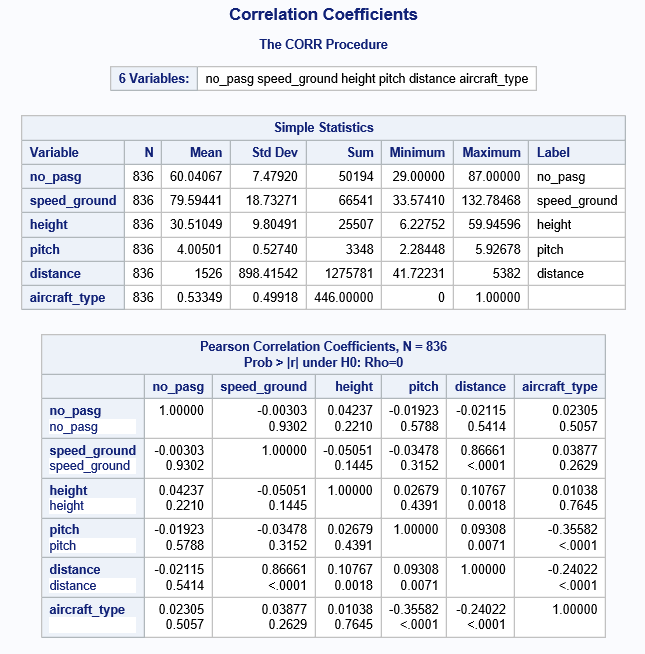
else aircraft\_type=1;

proc corr data=combined\_faa\_v3;

var no\_pasg speed\_ground height pitch distance aircraft\_type;

title Correlation Coefficients;

run;



From the above correlation table, we find that the distance is correlated with speed\_ground.

Let us now check after transforming our speed\_ground variable. We can try to check its square.

data combined\_faa\_v4;

set combined\_faa\_v3;

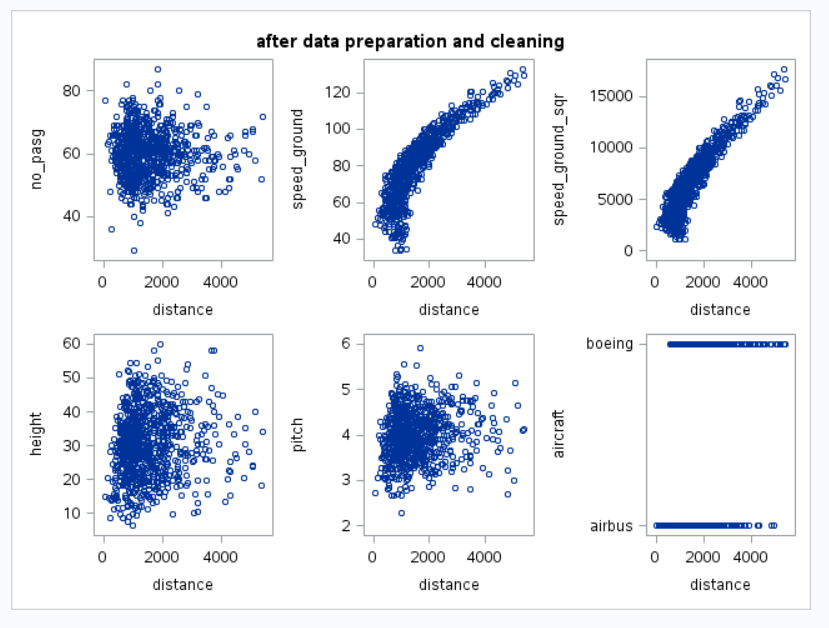
speed\_ground\_sqr = speed\_ground\*\*2;

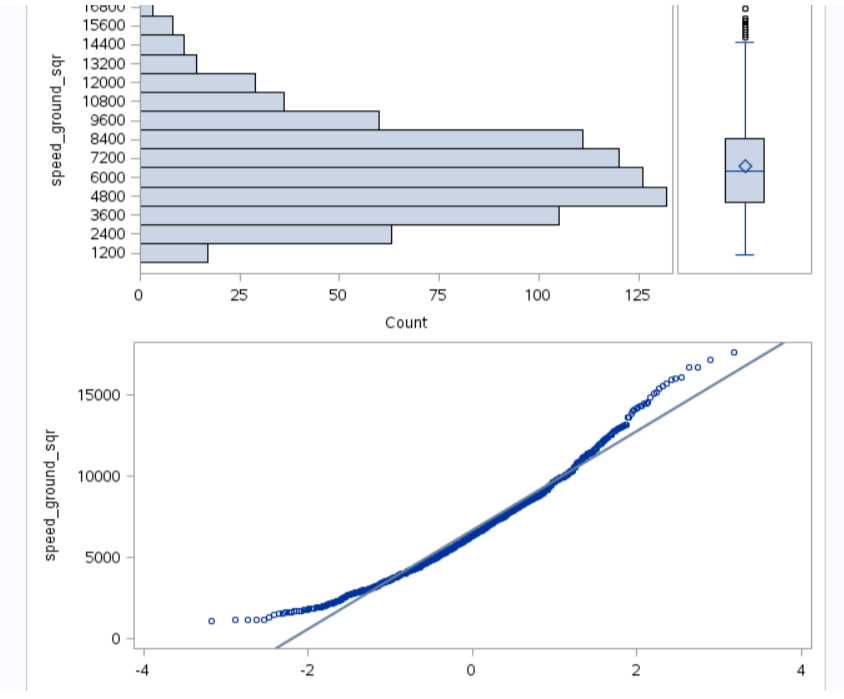
run;

proc sgscatter data=combined\_faa\_v4;

plot(no\_pasg speed\_ground speed\_ground\_sqr height pitch aircraft)\*distance;

run;





The plot above shows that the square of the speed\_ground is more linear that the speed\_ground. This should be better when we use for modeling.

Let us also check the correlation coefficients.

data combined\_faa\_v4;

set combined\_faa\_v4;

if aircraft="boeing" then aircraft\_type=0;

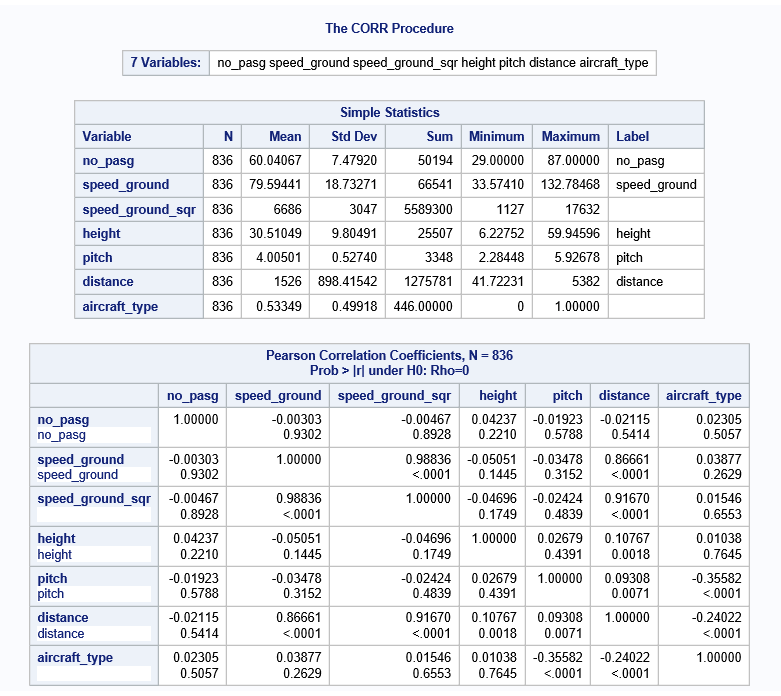
else aircraft\_type=1;

proc corr data=combined\_faa\_v4;

var no\_pasg speed\_ground speed\_ground\_sqr height pitch distance aircraft\_type;

title Correlation Coefficients;

run;



Now when we compare the values of correlation of speed\_ground and speed\_ground\_sqr, the value is greater for speed\_ground\_sqr.

**Modeling –**

We will build a model which we can be used to show the dependence of the response variable on the independent variables. We are trying to find a model which can show the variation of distance based on other variables.

Linear regression is a type of analysis which tries to find if one or more variable can represent a dependent variable.

The prerequisites of linear regression are –

* Linear relationship
* Multivariate normality
* No or little multicollinearity
* No auto-correlation
* Homoscedasticity

We can either transform the predictor or response variable. We choose to transform the predictor.

We can achieve a high value of R square if we consider both speed\_ground and speed\_ground\_sqr.

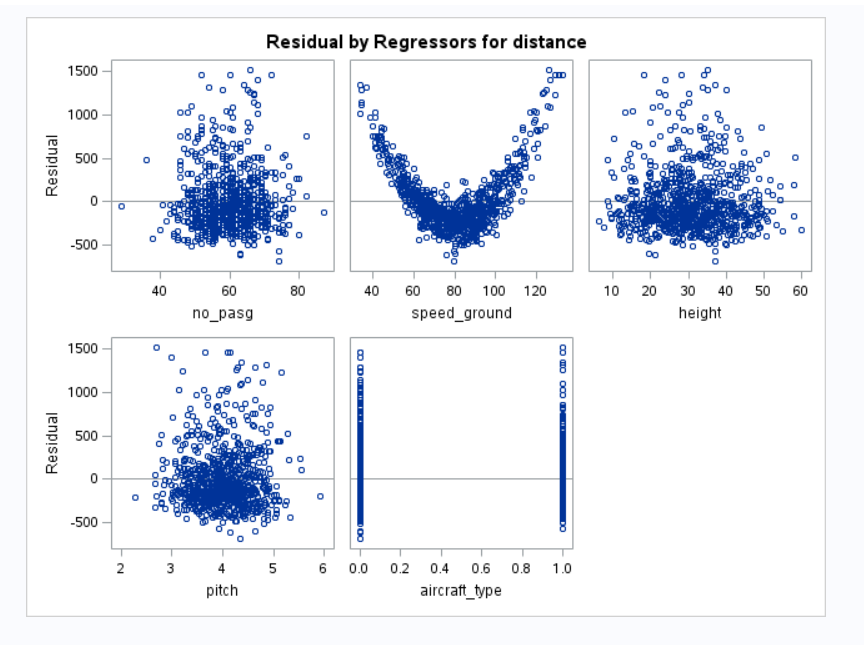
Let us try creating a model using the speed\_ground as the distance and speed\_ground are highly correlated.

proc reg data=combined\_faa\_v4;

model distance = no\_pasg speed\_ground\_sqr height pitch aircraft\_type/r spec;

output out=faa\_regression r=residual;

run;



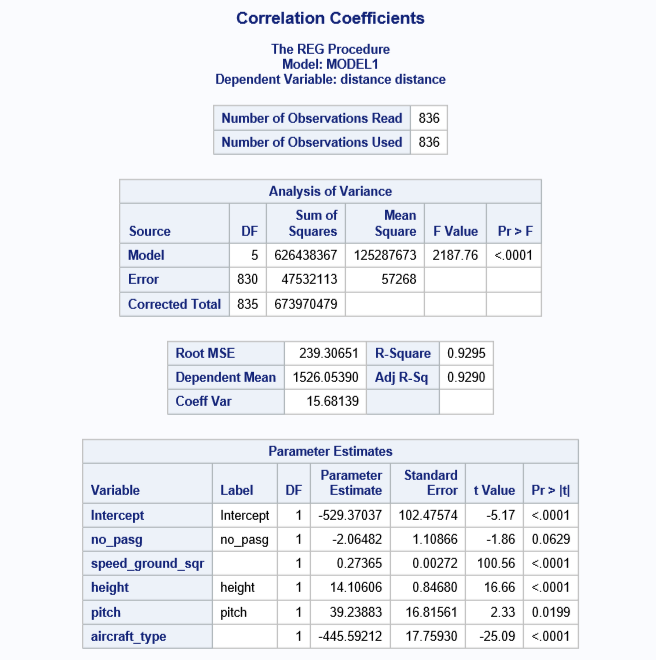
The relation between speed\_ground and residual is not linear and is a u shape. So, let us try a new model using speed\_ground\_sqr. Trying to remodel.

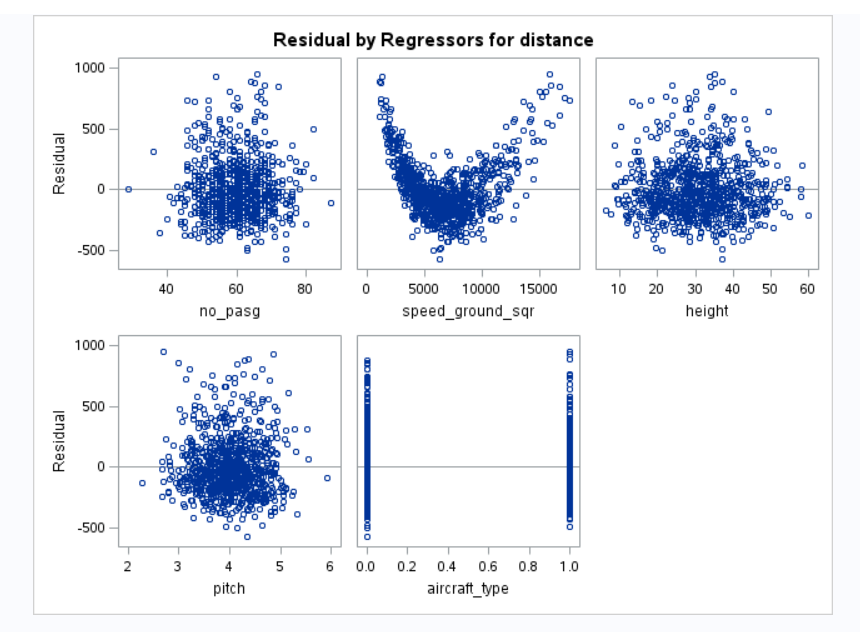
proc reg data=combined\_faa\_v4;

model distance = no\_pasg speed\_ground\_sqr height pitch aircraft\_type/r spec;

output out=faa\_regression r=residual;

run;





Still it has the same u shape. Again trying to remodel.

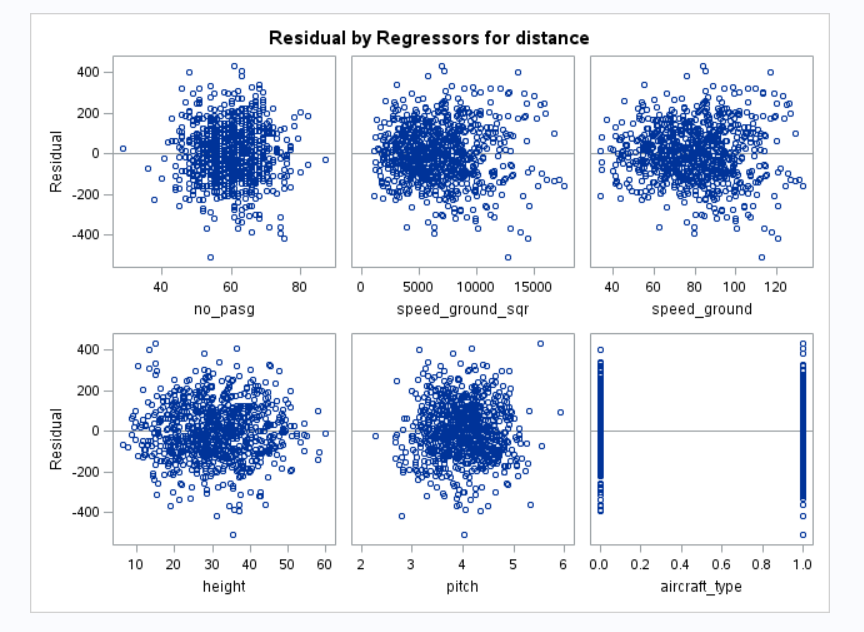
Let us try both speed\_ground and speed\_ground\_sqr.

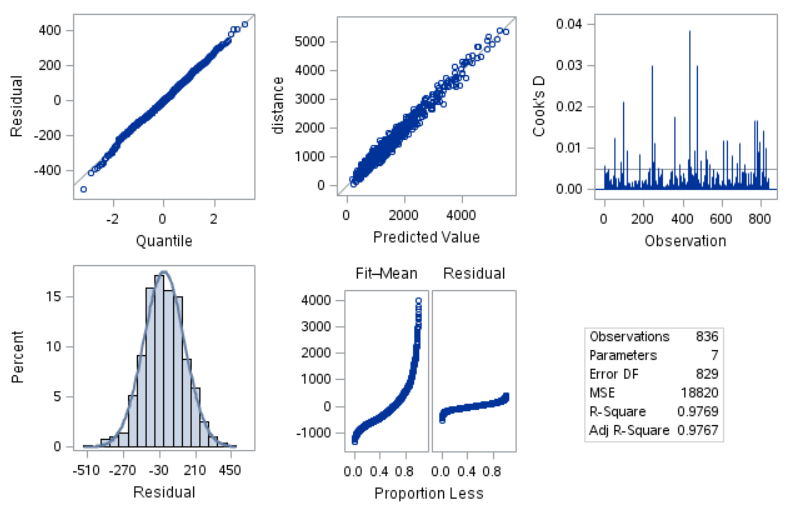
proc reg data=combined\_faa\_v4;

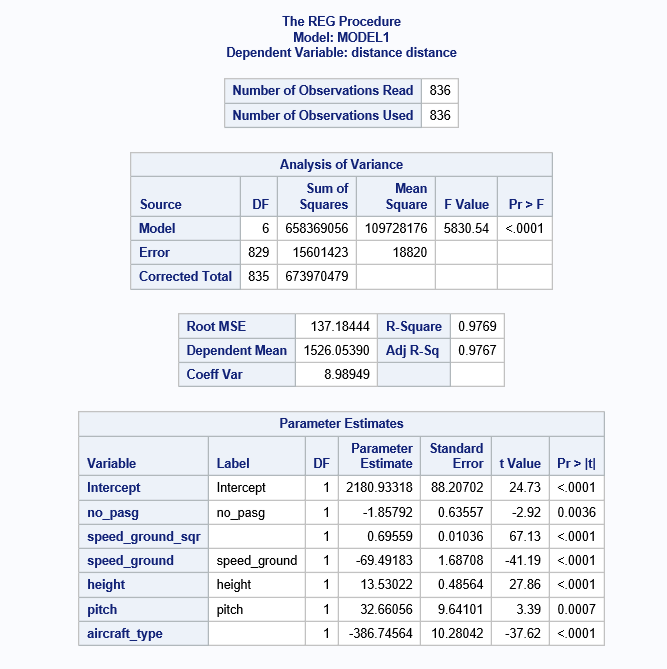
model distance = no\_pasg speed\_ground\_sqr speed\_ground height pitch aircraft\_type/r spec;

output out=faa\_regression r=residual;

run;







After taking speed\_ground and speed\_ground\_sqr the R-square value has also increased.

We see that the value of R square is 0.9769, looking at which we can say that the model fits the equation.

To create an equation, we check the p values of the variables. If the value of p is greater than .05 we drop the variable or else, we keep it.

**Creating the equation -**

Y = b0 + b1X1 + b2X2 + E

Y = Distance

B0 = 2180.93

B1 = -386.74

X1 = aircraft\_type

B2 = 0.695

X2 = speed\_ground\_sqr

X3=speed\_ground

B3= -69.49

B4 = 13.53

X4 = height

B5 = 32.66

X5 = pitch

**Distance = 2180.93 -386.74(aircraft\_type) + 0.695(speed\_ground\_sqr) -69.49(speed\_ground) +13.53(height) + 32.66(pitch)**

**Model Checking –**

In model checking we will have following criteria –

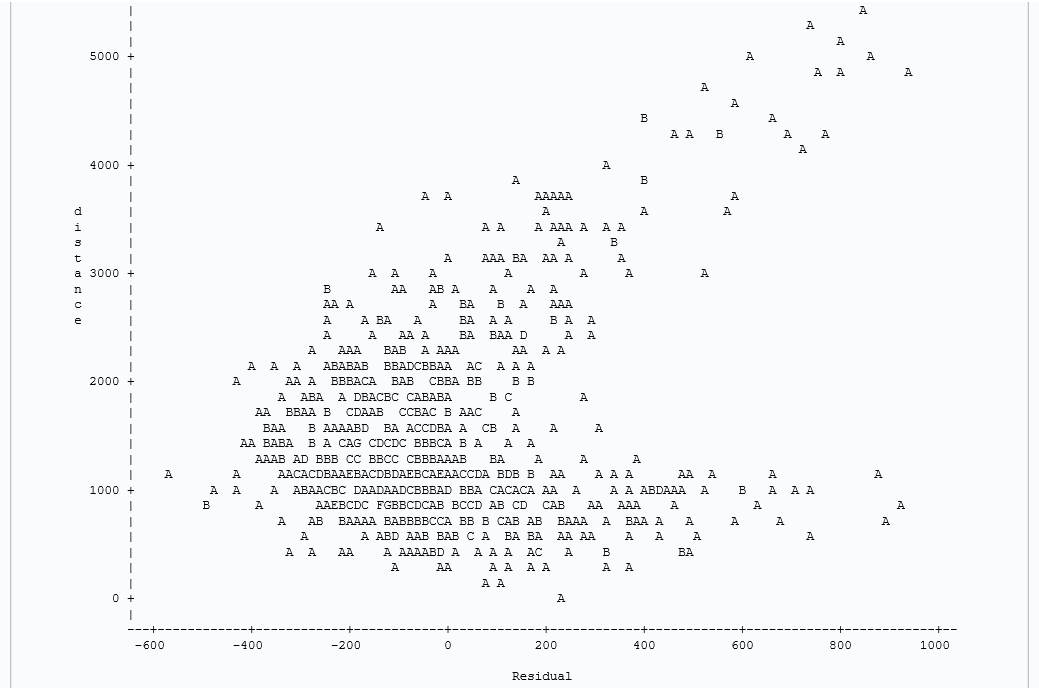
* Independent
* Normally Distributed
* Mean 0
* Constant Variance

**Independent –**

proc plot data=faa\_regression;

plot distance\*residual;

run;



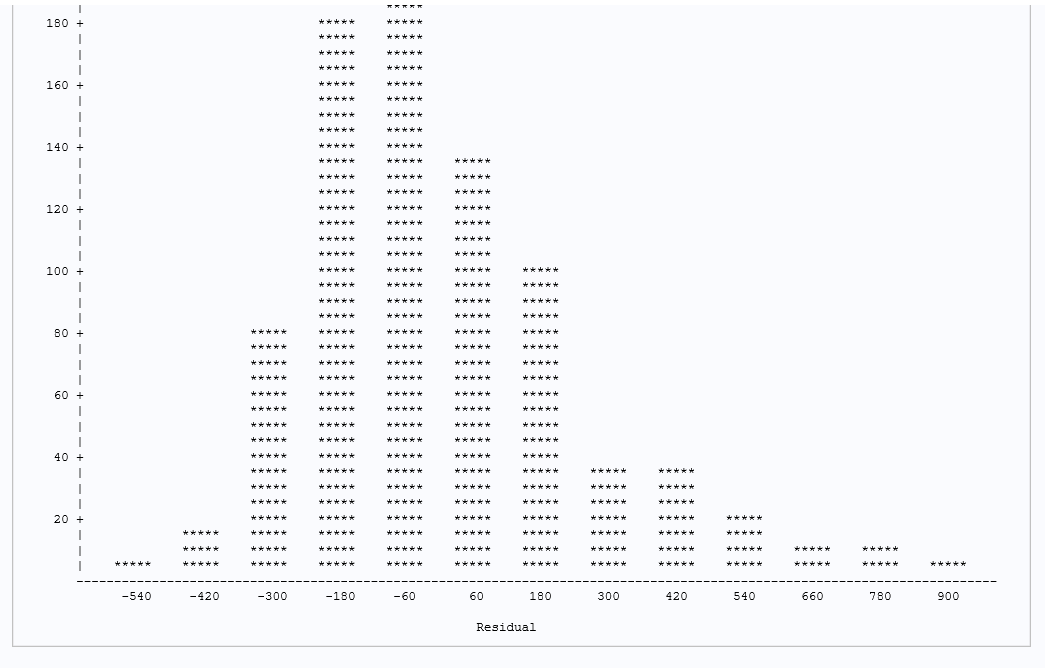
**Normally distributed –**

proc univariate data=faa\_regression;

histogram/normal;

var residual;

run;

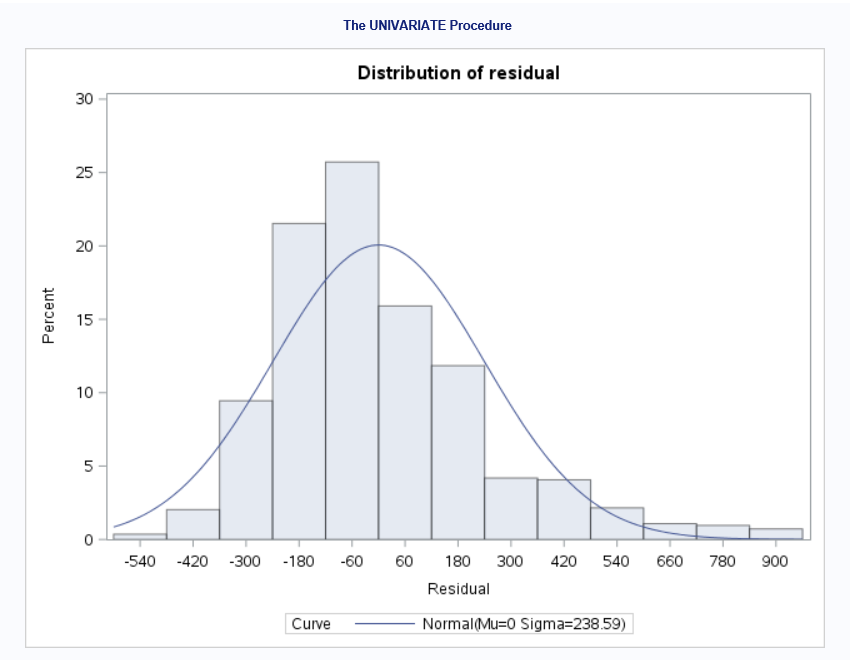


**Mean 0 –**

proc chart data=faa\_regression;

vbar residual;

run;



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

I have fit 836 observations in the final model. In my views, all the 950 observations were not eligible for a model because of the following reasons –

1. 100 observations from FAA2 were exactly duplicate of the observations from FAA1, but FAA2 had duration column missing. But, as I have dropped the duration column, those 100 observations have become exact replicas. So, it is preferred to remove those duplicates. After this we are left with 850 observations.
2. We also remove the abnormal values which are mentioned in problem statement.

After removing those, we are left with 836 observations.

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

After modelling, we are left with this equation

**Distance = 2180.93 -386.74(aircraft\_type) + 0.695(speed\_ground\_sqr) -69.49(speed\_ground) +13.53(height) + 32.66(pitch)**

Above equation shows that the distance is impacted by five variables. We do not consider the speed\_air, no\_pasg and duration variables.

The other variables like no\_pasg, duration, speed\_air did not have a strong correlation like the other four variables. So, we removed these and kept the others.

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

Yes, there is a difference in both the makes of aircraft. If we analyze the response variable against the aircrafts, we can see the difference.

proc ttest data=combined\_faa\_v4;

var distance;

class aircraft\_type;

run;

The mean and median of distance for Boeing is greater than that of Airbus.

The minimum and maximum are also very different.

