**BANA 6043 STATISTICAL COMPUTING PROJECT**

Submitted by Pravallika Kalidindi (M12823119)

**Abstract:**

Landing overrun is a problem for most flight landing operations. In this report we are trying to identify key factors affecting the landing distance of commercial flights.In order to determine and quantify the impact of factors on landing distance we created a linear regression model keeping landing distance as the dependent variable.

Landing distance of a flight is largely dependent on ground speed of aircraft, type of aircraft and height of aircraft when it passes through the threshold of the runway.

**√Distance = -15.8990 + (394.82 \* aircraft\_boeing) + (13.63 \* Height) + (0.0001743 \* speed\_ground^4)**

* For Boeing aircraft type, the predicted landing distance is 394 units greater than Airbus aircraft type
* For every unit increase in Height above threshold, there will be 13.63 units increase in landing distance.
* For every unit increase in ground speed to the power 4, the predicted landing distance will increase in 0.0001743 units.

**CHAPTER-1 : Data Cleaning**

Goal : The major goal of this project is to combine the data, clean it and make it ready for further analysis. Steps to achieve this goal are mentioned below in detail

1. Step 1 – Importing Excel files as data sets in SAS

Code –

**PROC** **IMPORT** OUT= WORK.FAA2 DATAFILE= "/FOLDERS/MYFOLDERS/FAA2.XLS"

DBMS=XLS REPLACE;

SHEET="FAA2";

GETNAMES=YES;

**RUN**;

**PROC** **IMPORT** OUT= WORK.FAA1 DATAFILE= "/FOLDERS/MYFOLDERS/FAA1 (1).XLS"

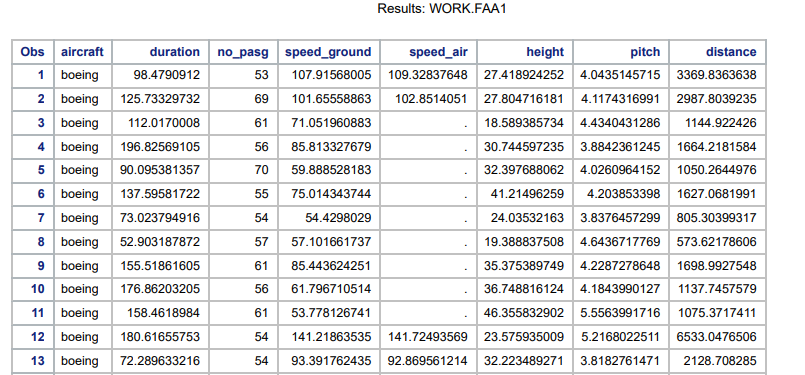
DBMS=XLS REPLACE;

SHEET="FAA1";

GETNAMES=YES;

**RUN**;

Result – Snapshot of Raw data



1. Step2 – Cleaning the Raw data before combining.

A quick look at the two data sets after importing into SAS has shown that there are empty rows. To clean this and remove the rows where all columns are empty, the following code is used.

Code used to detect empty rows –

OPTIONS MISSING=' ';

**DATA** FAA2\_SPEC;

SET FAA2;

IF MISSING(CATS(OF \_ALL\_)) ;

**RUN**;

OPTIONS MISSING=' ';

**DATA** FAA1\_SPEC;

SET FAA1;

IF MISSING(CATS(OF \_ALL\_)) ;

**RUN**;

Code used to overwrite existing data sets and removing empty rows

OPTIONS MISSING='';

**DATA** FAA2;

SET FAA2;

IF MISSING(CATS(OF \_ALL\_)) THEN DELETE;

**RUN**;

OPTIONS MISSING='';

**DATA** FAA1;

SET FAA1;

IF MISSING(CATS(OF \_ALL\_)) THEN DELETE;

**RUN**;

1. Step 3 – Combining the two data sets.

The technique used to combine the data sets is Concatenating.

Code to combine data sets-

**DATA** FLIGHT\_DETAILS;

SET FAA1 FAA2;

**RUN**;

The second data set has an extra column called duration. After combining the data sets, the new data set has all the common columns from two data sets and the new column “Duration”. Concatenating the records has forced all the first data set entries to have empty values for Duration column. Also it is observed that there are duplicate rows.

1. Step 4 – Removing Duplicate rows

Code –

**PROC** **SORT** DATA=FLIGHT\_DETAILS OUT=FLIGHT\_DETAILS NODUPKEY ;

BY AIRCRAFT NO\_PASG SPEED\_GROUND SPEED\_AIR HEIGHT PITCH DISTANCE;

**RUN**;

1. Step 5 – Completeness check for each variable

To see the counts of missing entries in each variable in data set, the following code is used.

Code-

**PROC** **FORMAT**;

VALUE $MISSFMT ' '='MISSING' OTHER='NOT MISSING';

VALUE MISSFMT . ='MISSING' OTHER='NOT MISSING';

**RUN**;

**PROC** **FREQ** DATA=FLIGHT\_DETAILS;

FORMAT \_CHAR\_ $MISSFMT.; /\* APPLY FORMAT FOR THE DURATION OF THIS PROC \*/

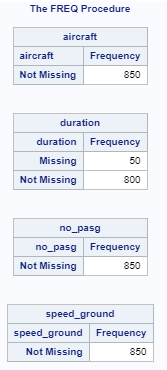
TABLES \_CHAR\_ / MISSING NOCUM NOPERCENT;

FORMAT \_NUMERIC\_ MISSFMT.;

TABLES \_NUMERIC\_ / MISSING NOCUM NOPERCENT;

**RUN**;

Output –

Out of 6 variables present , 4 of them (aircraft, speed\_ground, no\_pasg, height) have no missing values.

Duration variable has missing values – This happened because of combining the two data sets. Duration variable is present only in one of the two data sets.

Speed\_air also has almost 75% missing values. But this variable will be retained for now as it is an important variable. Similarly Duration variables will be left as it is for now even though there are missing values.

1. Step 7 - Performing the validity check of each variable

* No of passengers –

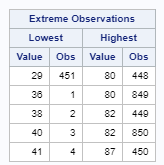
Code to check extreme values –

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR NO\_PASG;

**RUN**;

Result –



Since nothing is mentioned about the number of passengers and there are no abnormal values, this column is fine .

* Speed on ground –

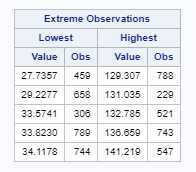
Code-

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR SPEED\_GROUND;

**RUN**;

Result –



There are abnormal values for Speed on ground . Speed on ground less than 30MPH and greater than 140 MPH are considered to be abnormal. These values are deleted.

Code-

**DATA** FLIGHT\_DETAILS;

SET FLIGHT\_DETAILS;

IF SPEED\_GROUND<**30** OR SPEED\_GROUND>**140** THEN DELETE;

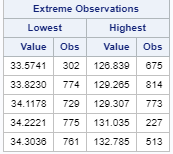
**RUN**;

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR SPEED\_GROUND;

**RUN**;

Result



* Speed in air-

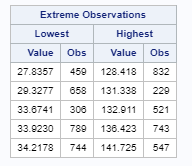
Code-

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR SPEED\_AIR;

**RUN**;

Result



Abnormal values are observed for Speed in Air. Values less than 30MPH and greater than 140MPH are considered to be abnormal. These values are deleted.

Code-

**DATA** FLIGHT\_DETAILS;

SET FLIGHT\_DETAILS;

IF SPEED\_AIR<**30** OR SPEED\_AIR>**140** THEN DELETE;

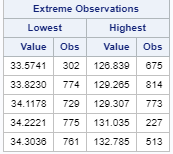
**RUN**;

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR SPEED\_AIR

**RUN**;

Result



* Height

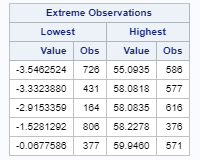
Code-

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR HEIGHT;

**RUN**;

Output –



It is observed that there are negative values in Height column which are not reasonable. The abnormal rows are deleted from the data set.

Code to delete rows with abnormal Height values

**DATA** FLIGHT\_DETAILS;

SET FLIGHT\_DETAILS;

IF HEIGHT<**0** THEN DELETE;

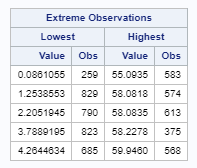
**RUN**;

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR HEIGHT;

**RUN**;

Output



* Pitch-

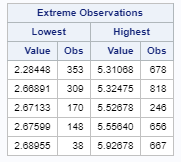
Code –

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR PITCH;

**RUN**;

Output –



No abnormal values are observed.

* Distance

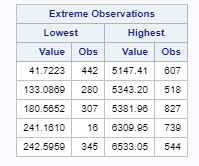
Code

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR DISTANCE;

**RUN**;

Output



Since the airport runway length is typically 6000 feet, distance values greater than 6000 are abnormal and we will delete them.

Code to delete abnormal values.

**DATA** FLIGHT\_DETAILS;

SET FLIGHT\_DETAILS;

IF DISTANCE>**6000** THEN DELETE;

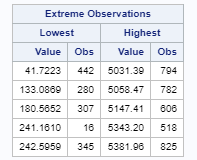
**RUN**;

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR DISTANCE;

**RUN**;

Output –



* Duration

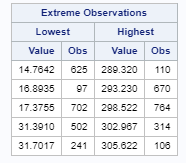
Code-

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR DUARTION;

**RUN**;

Output



Flight duration should be greater than 40 minutes so values less than 40 minutes are deleted.

Code-

**DATA** FLIGHT\_DETAILS;

SET FLIGHT\_DETAILS;

IF DURATION<40 AND DURATION <> . THEN DELETE;

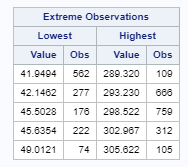
**RUN**;

**PROC** **UNIVARIATE** DATA=FLIGHT\_DETAILS;

VAR DURATION;

**RUN**;

Output-



1. Step 8 – Distribution of Variables

Code

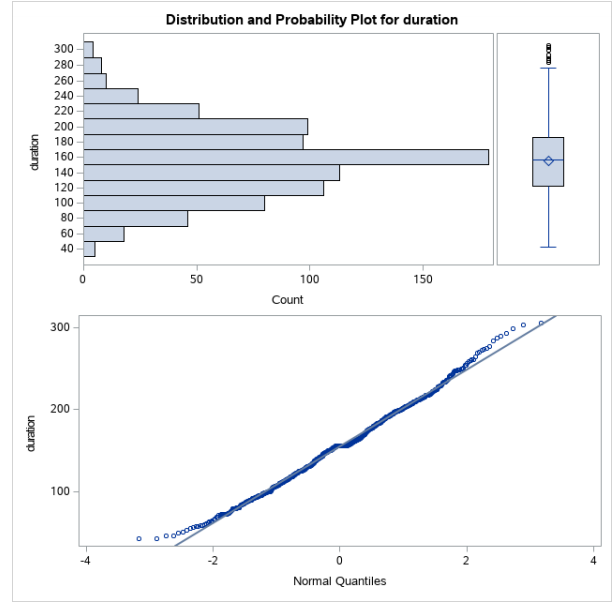
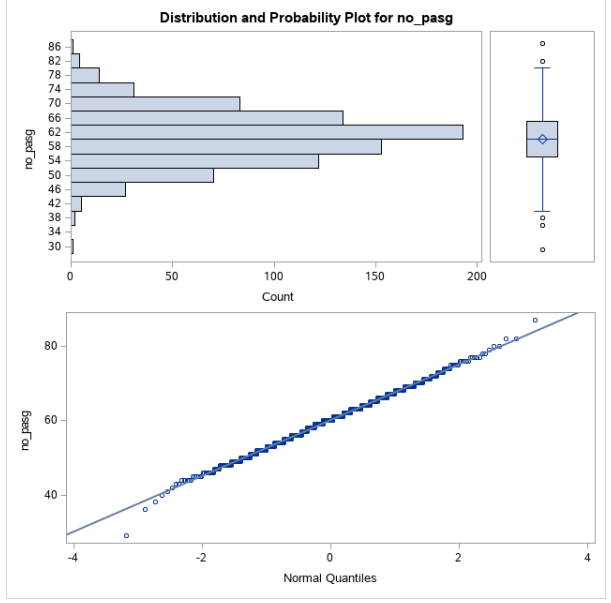
ODS SELECT PLOTS;

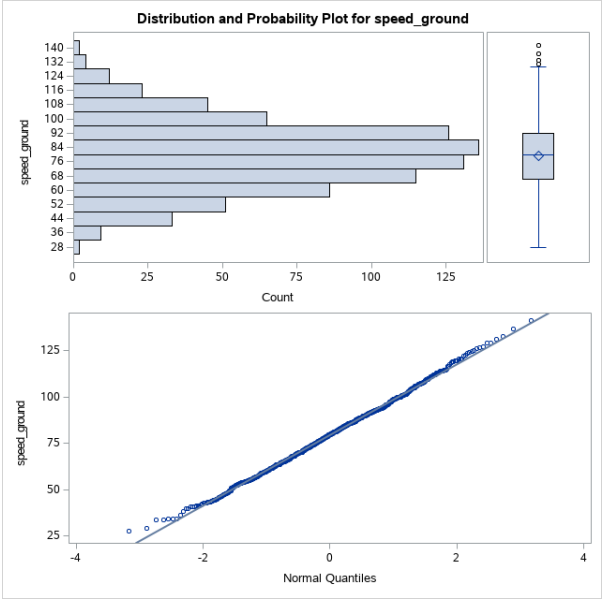
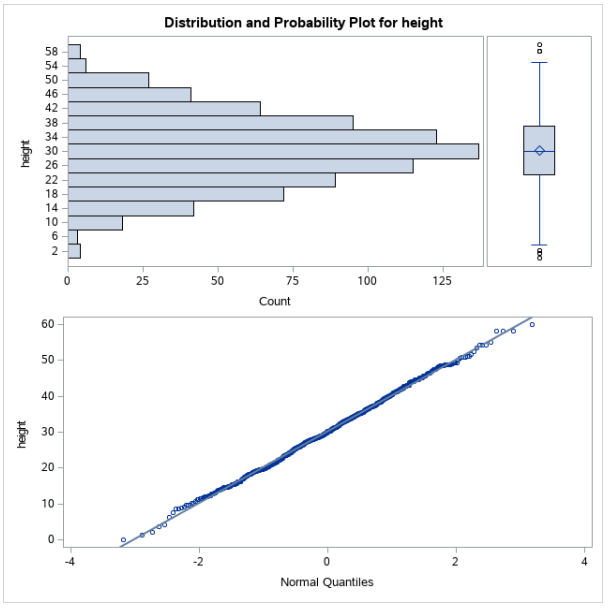
**PROC** **UNIVARIATE** DATA=Flight\_Details PLOTS;

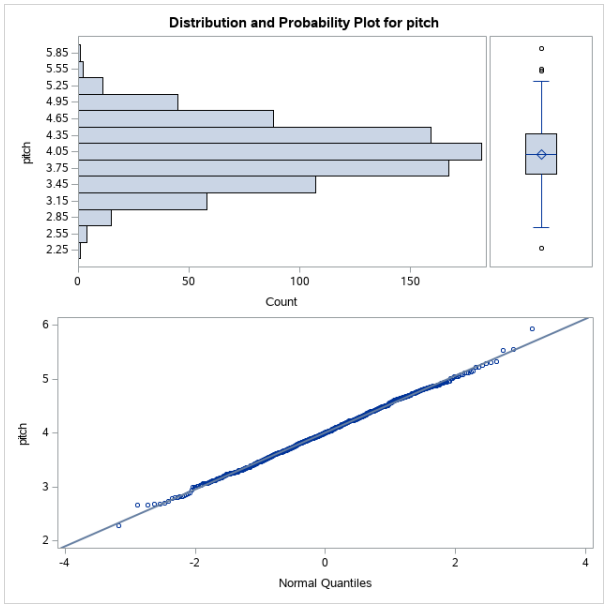
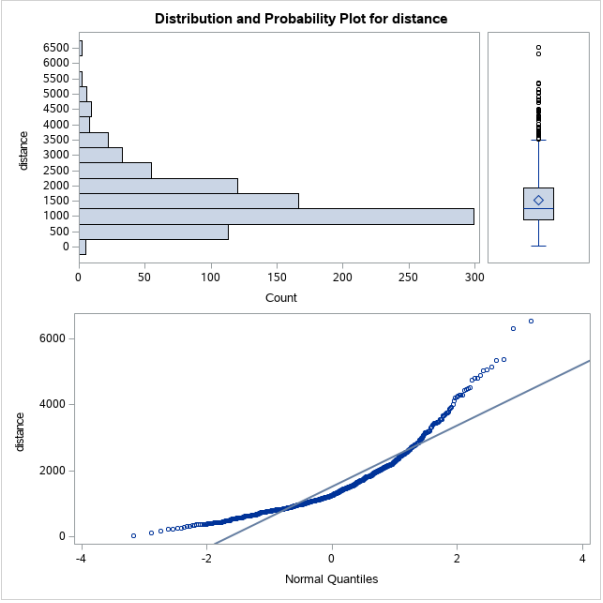
VAR DURATION NO\_PASG SPEED\_GROUND HEIGHT PITCH DISTANCE;

**RUN**;

Output

Observations -

As seen from the above graphs all variables follow close to normal distribution except for distance. The distance variable is heavily left skewed as seen from the histogram and Q-Q plot. We might have to transform this variable when we build our model.

**Chapter 2 - Descriptive Study**

Goal- Understand the correlation between the dependant variable and independent variables and analyse the plots.

1. Step 1 - Check the correlation between distance and other independent variables

Code-

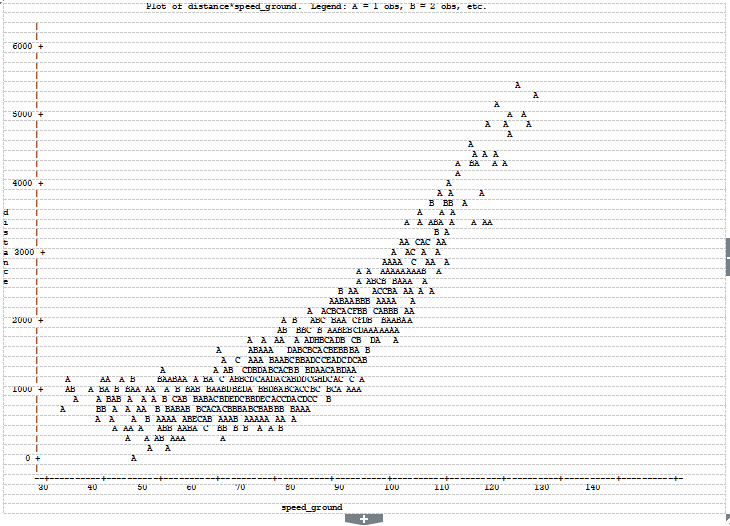
**PROC** **PLOT** DATA=FLIGHT\_DETAILS;

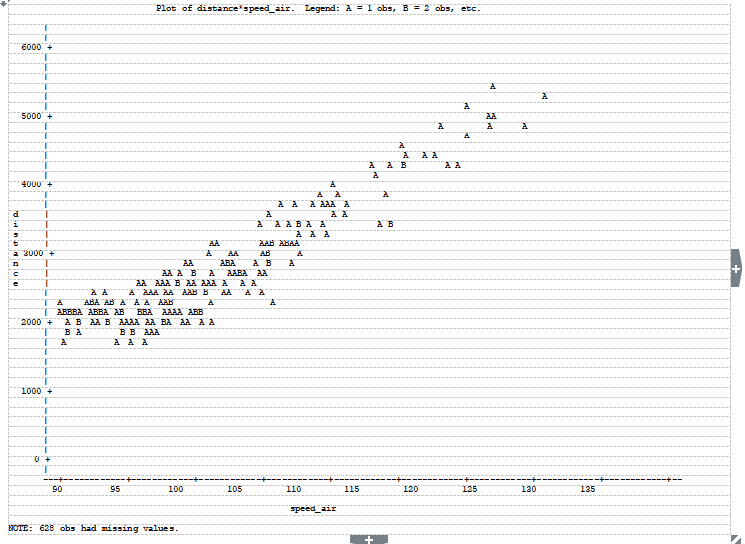
PLOT DISTANCE\*(NO\_PASG SPEED\_GROUND SPEED\_AIR HEIGHT DURATION PITCH);

**RUN**;

Output-

* We cannot see any strong linear trend between distance and other variables - no\_pasg, height, duration, pitch.
* Speed\_ground seems to have a non linear second order relationship with distance. We will further examine this relationship using residual plots and transform this variable if needed.
* Speed\_air has a linear relationship with distance. But because speed\_air has 75% missing values, this relationship cannot be certain.





2. STEP 2 - Checking the strength of relationship between dependent and independent variables.

Before building the model, it is important to check the correlation strength between independent and dependent variables. We also need to check if independent variables are highly correlated. We will only take one among those highly correlated independent variables.

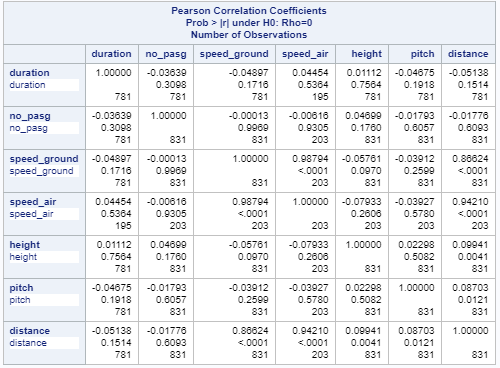
Code -

**PROC** **CORR** DATA=FLIGHT\_DETAILS;

**RUN**;

Observations-

* There is no strong relation between distance and other independent variables - duration and no\_pasg (p value for hypothesis testing that there is no correlation between the two variables is greater than 0.05 significance level).
* There is a strong relationship between speed\_ground and distance , speed\_air and distance. Hence speed\_ground and speed\_air can be strong predictors of distance
* Speed\_ground and speed\_air also have a very strong correlation. We can use one of these two variables to predict distance. Since we know speed\_air has 75% missing values, we will not use speed\_air but we will use speed\_ground to predict the distance.
* Pitch and height also have significant correlation with distance.



3.Step 3 - Creating binary variables for Aircraft type

Code -

**DATA** FLIGHT\_DETAILS;

SET FLIGHT\_DETAILS;

IF AIRCRAFT='BOEING' THEN DO; AIRCRAFT\_BOEING=**1**; AIRCRAFT\_AIRBUS=**0**;

END;

**RUN**;

**DATA** FLIGHT\_DETAILS;

SET FLIGHT\_DETAILS;

IF AIRCRAFT='AIRBUS' THEN DO; AIRCRAFT\_BOEING=**0**; AIRCRAFT\_AIRBUS=**1**;

END;

DROP AIRCRAFT;

**RUN**;

# **CHAPTER 3: Data Modelling**

1.Step 1 - Iteration 1

Run proc reg with all variables expect for speed\_air. We are not including speed\_air because we have seen previously that there is a high correlation between speed\_air and speed\_ground. Hence we will only use one of the two variables and we decided to drop speed\_air because it has 75% missing values.

Code-

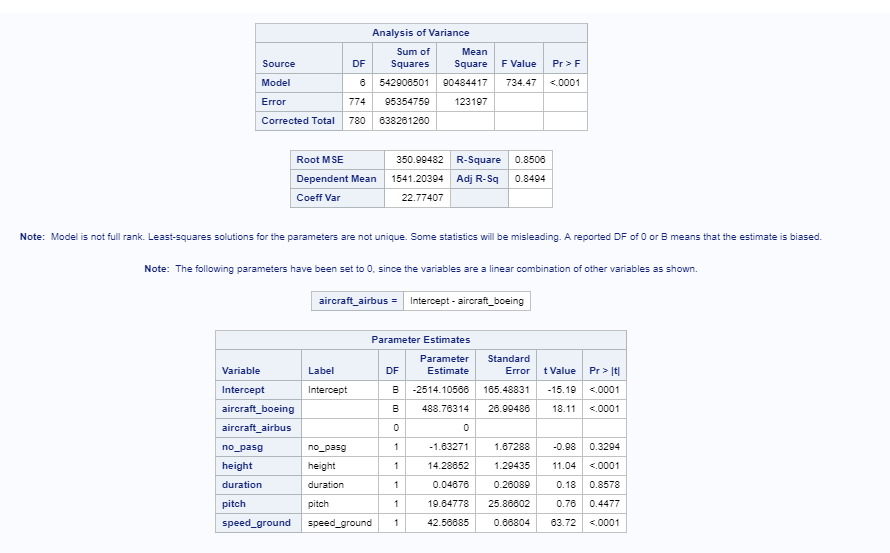
**PROC** **REG** DATA=FLIGHT\_DETAILS;

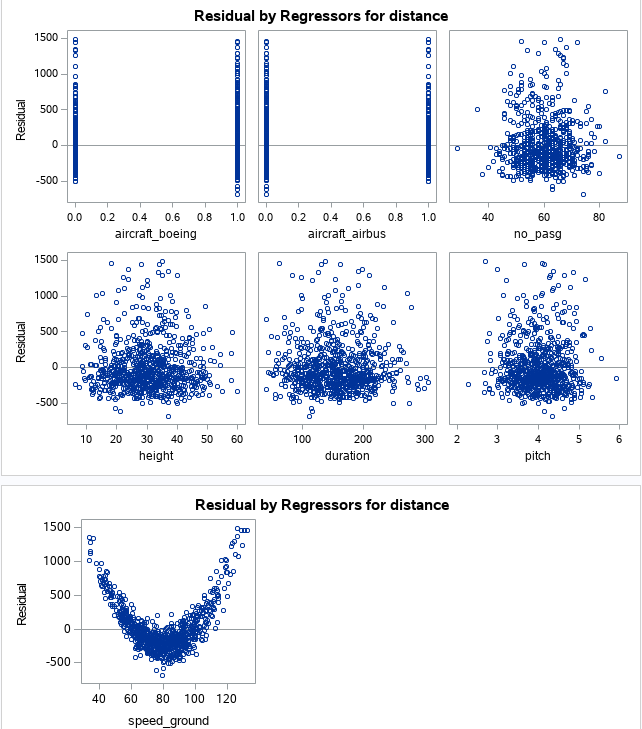
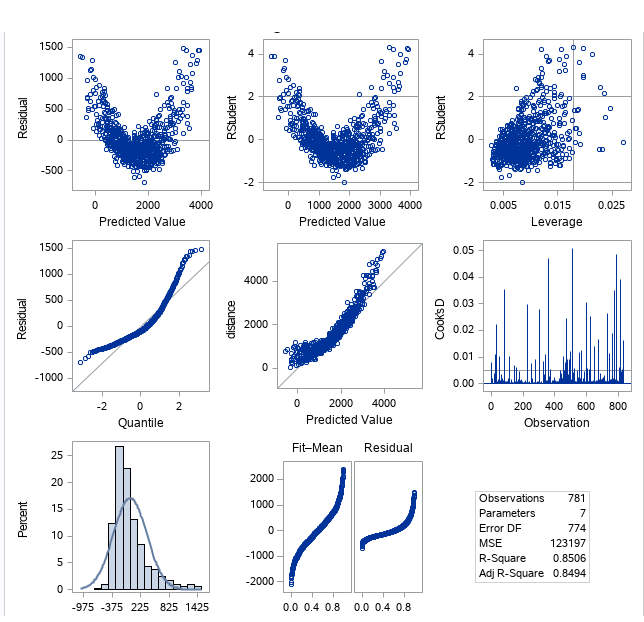
MODEL DISTANCE= AIRCRAFT\_BOEING AIRCRAFT\_AIRBUS NO\_PASG HEIGHT DURATION PITCH SPEED\_GROUND;

OUTPUT OUT=OUTP\_REG R=RES RESIDUAL=OUTPUT\_RESIDUAL;

**RUN**;

Output -



Observations -

* There are 831 rows in the data set out of which duration column has 50 missing values. These observations are not used in the regression model.
* Analysis of variance table shows a p-value of less than 0.05 level of significance. So we can reject the null hypothesis that all coefficients for independent variables are zero. We can hence conclude that there is some dependence between dependent and independent variables.
* The model predicted an adjusted R-square of 80% which suggests that a significant proportion of variability in dependent variable is predicted by independent variables.
* From the parameter estimate table, we can see that variables such as no\_pasg , duration and pitch are not significant. Their p-value is greater than 0.05 level of significance and hence we accept the null hypothesis that their coefficient is zero. We will drop these variables in our second iteration of model building.
* Variables such as speed\_ground, height, aircraft\_boeing, aircraft\_airbus are significant variables and we keep them in our second iteration of model building.
* Plot between residual and independent variables show that there exists a second order relationship between speed\_ground and distance. We will address this issue in our second iteration.

Assumptions of Linear Regression-

* The plot between residuals and independent variables should be identically distributed. This is not the case here as seen in the graph.
* The Q-Q plot shows that the residuals are not following normal distribution.
* The U-shaped curve between residuals and predicted values suggests that there is some non linear term in the model. Speed-ground might be causing non linearity.

2.Step 2 - Second iteration of model building.

* In this interation, we will only include significant variables which are speed\_ground , height, aircraft\_boeing and aircraft\_airbus .
* To address the non linearity, we will try using speed\_groud^2 and speed\_ground^4.

Code -

**DATA** FLIGHT\_DETAILS;

SET FLIGHT\_DETAILS;

SPEED\_2 = SPEED\_GROUND\*SPEED\_GROUND;

SPEED\_4 = SPEED\_2 \* SPEED\_2;

**RUN**;

Model building -

Code -

**PROC** **REG** DATA=FLIGHT\_DETAILS;

MODEL DISTANCE= AIRCRAFT\_BOEING AIRCRAFT\_AIRBUS HEIGHT SPEED\_2;

OUTPUT OUT=OUTP\_REG R=RES RESIDUAL=OUTPUT\_RESIDUAL;

**RUN**;

**PROC** **REG** DATA=FLIGHT\_DETAILS;

MODEL DISTANCE= AIRCRAFT\_BOEING AIRCRAFT\_AIRBUS HEIGHT SPEED\_4;

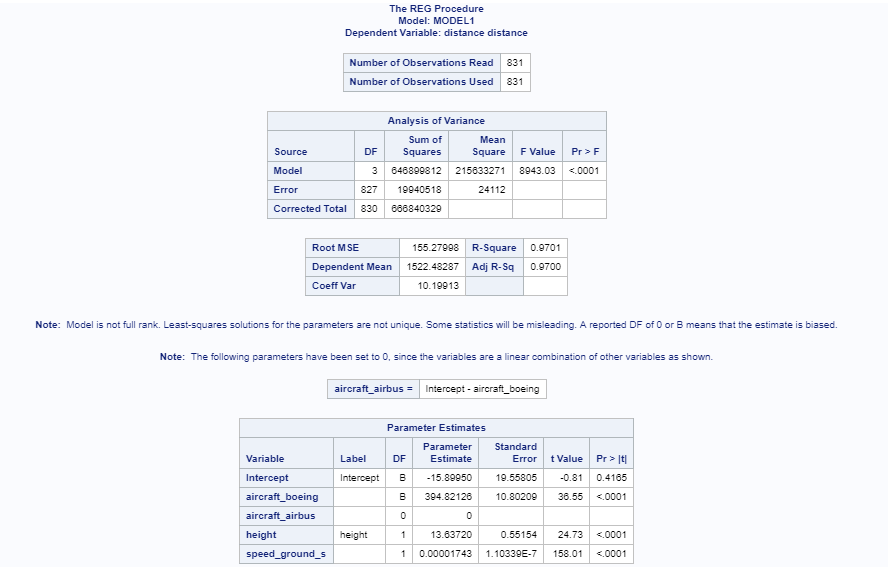
OUTPUT OUT=OUTP\_REG R=RES RESIDUAL=OUTPUT\_RESIDUAL;

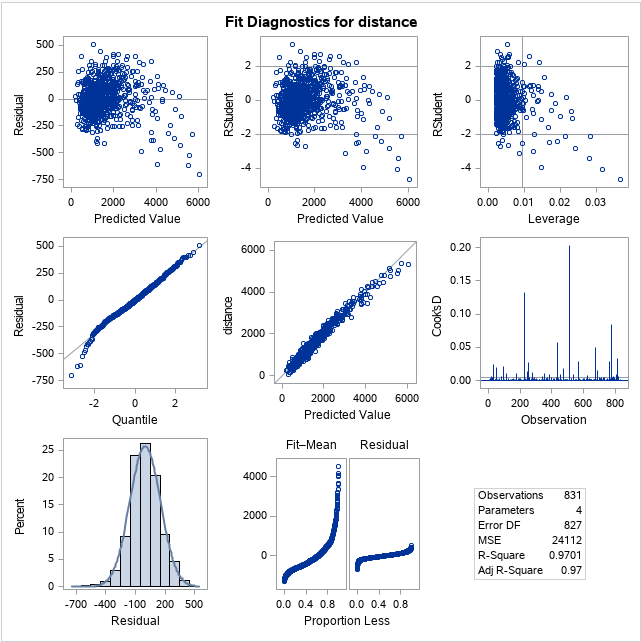
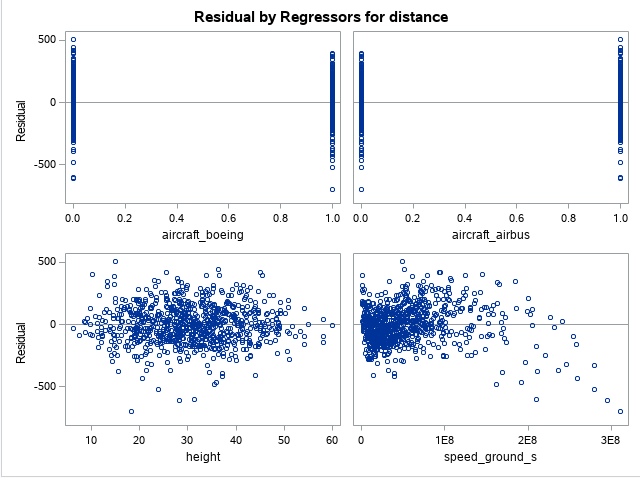
**RUN**;

Observations-

It has been seen that using Speed\_ground to the power 4 has better fit statistics.

Output -



Observations -

* There are 831 rows in the data set .All of these observations are not used in the regression model.
* Analysis of variance table shows a p-value of less than 0.05 level of significance. So we can reject the null hypothesis that all coefficients for independent variables are zero. We can hence conclude that there is some dependence between dependent and independent variables.
* The model predicted an adjusted R-square of 97% which has significantly increased since iteration 1. This suggests that a significant proportion of variability in dependent variable is predicted by independent variables.
* From the parameter estimate table, we can see that all dependent variables are significant and signs are intuitive. Their p-value is less than 0.05 level of significance and hence we reject the null hypothesis that their coefficient is zero.

Assumptions of Linear Regression-

* The plot between residuals and independent variables should be identically distributed. This is satisfied here as seen in the graph. The randomness in variance has significantly reduced from the first iteration.
* The Q-Q plot shows that the residuals are following an approximate normal distribution. We will further examine this using a normalcy test.

Code-

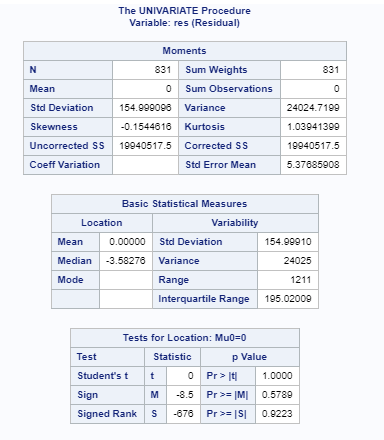
**PROC** **UNIVARIATE** DATA=OUTP\_REG

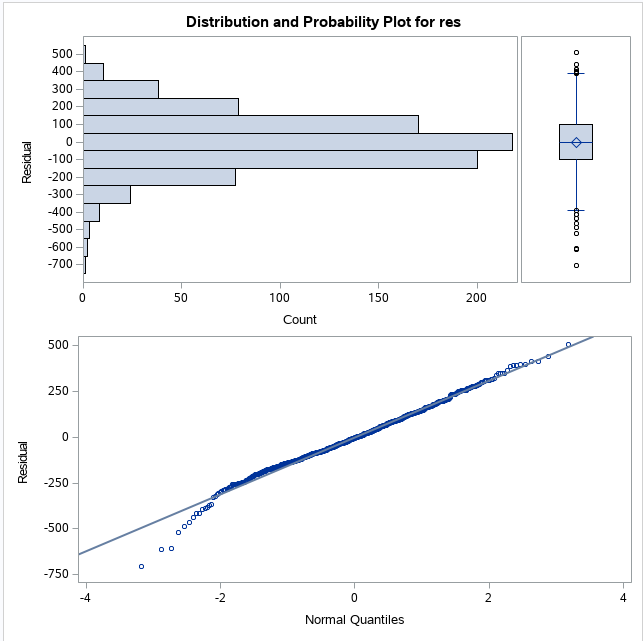
NORMAL PLOT;

VAR RES;

**RUN**;

Output -



The test for normality shows that the residuals are not following a normal distribution (Shapiro-Wilk p-value is less than 0.05. So we reject the hypothesis that residuals are following a normal distribution).

But from the Q-Q plot and the histogram we can see that the residuals are quite close to a normal distribution.

Summary :

Landing distance of a flight is largely dependent on ground speed of aircraft, type of aircraft and height of aircraft when it passes through the threshold of the runway.

**√Distance = -15.8990 + (394.82 \* aircraft\_boeing) + (13.63 \* Height) + (0.0001743 \* speed\_ground^4)**

* For Boeing aircraft type, the predicted landing distance is 394 units greater than Airbus aircraft type
* For every unit increase in Height above threshold, there will be 13.63 units increase in landing distance.
* For every unit increase in ground speed to the power 4, the predicted landing distance will increase in 0.0001743 units.

**Answers to Questions**

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

**Answer**: 831 observations are used to fit the model. 119 observations have been deleted to remove abnormal values in ground speed of aircraft, air speed of aircraft, height above threshold of runway, duration of flight etc.

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

**Answer**: Landing distance of a flight is largely dependent on ground speed of aircraft, type of aircraft and height of aircraft when it passes through the threshold of the runway.

* For Boeing aircraft type, the predicted landing distance is 394 units greater than Airbus aircraft type
* For every unit increase in Height above threshold, there will be 13.63 units increase in landing distance.
* For every unit increase in ground speed to the power 4, the predicted landing distance will increase in 0.0001743 units.

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

**Answer** : Yes, there is a significant difference between the make of two aircrafts, Boeing and Airbus. From our final model, we can say that :

For Boeing aircraft type, the predicted landing distance is 394 units greater than Airbus type.