Stat Computing – Final Project
BANA 6043
Section 003

Name	**M-Number**
Dhivya Rajprasad	M10857825

Abstract:

I have fitted the final model using 833 observations instead of the original 950 observations as 100 observations were duplicates and 17 observations were abnormal with respect to the defined parameters in the problem. The factors which impact the landing distance of a flight are the make of the aircraft, the ground speed of the aircraft and height of the aircraft when passing over the threshold of the runway. The relationship between the variables can be explained by the final regression equation

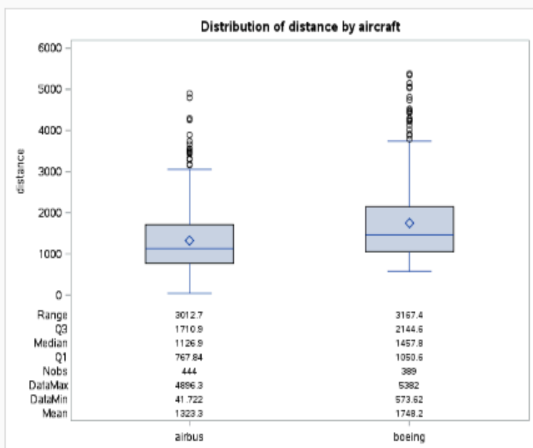
y(distance)= 2230.448-399.37x_1+0.69815$x_2{}^2$- 69.97x_2+13.44x_3

where x_1= make of an aircraft

x_2 = ground speed of the aircraft

x_3 = height of the aircraft when passing over the run way.

The landing distance is impacted by the make of the aircraft with higher landing distance for boeing when compared to airbus. The details are as below:



CHAPTER 1- DATA PREPARATION AND DATA CLEANING

GOAL:

Data cleaning is the first step in the process of data analysis and forms the most important step of the process. The quality of the output of the data analysis procedure depends on the quality of the input data. Data cleaning involves finding the errors and irregularities in the data and eliminating them from the data. The steps involved varies depending on the size of the data along with the number of variables being used.

The steps which have been used in this assignment are:

1. Importing multiple datasets
2. Checking for missing rows in datasets
3. Removing missing rows from the dataset
4. Concatenating the multiple datasets into a single dataset
5. Finding missing values and understanding the missing values
6. Duplicates check
7. Applying conditions to the data to subset the abnormal data and finding the count of the abnormal data
8. Removing abnormal data
9. Finding the basic parameters for all classes of data

The steps may be reiterated multiple times to get a cleaner dataset when more variables are given or larger datasets are considered.

STEP 1- Importing multiple datasets:

Proc Import statement is used to import the data sets with **Out** statement specifying the dataset name once it is imported. **Datafile** statement specifies the location of the file to be imported with **DBMS** statement to specify the database format to be imported along with **replace** function to replace any variables of the same name given before. **Getnames** statement specifies if the first record in the dataset should be taken as the variable names.

```
/*Import the dataset 1*/
proc import out=flightinfo1
datafile= "/folders/myfolders/sasuser.v94/FAA1.xls"
dbms= xls replace;
getnames= yes;

/*Import the dataset 2*/
proc import out=FAA2
datafile= "/folders/myfolders/sasuser.v94/FAA2.xls"
dbms= xls replace;
getnames= yes;
```

STEP 2 - Checking for missing rows in datasets

Proc Means statement is used to find the missing rows in the data (repeated missing variables across all variables-nmiss) along with other basic parameters of the data.

```
/*Checking for missing rows in datasets*/
title Basic Parameters for Dataset 1;
proc means data=flightinfo1 n nmiss mean median std max min;
title Basic Parameters for Dataset 2;
proc means data=FAA2 n nmiss mean median std max min;
```

Basic Parameters for Dataset1

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Maximum	Minimum
duration	duration	800	0	154.0065385	153.9480975	49.2592338	305.6217107	14.7642071
no_pasg	no_pasg	800	0	60.1325000	60.0000000	7.5271686	87.0000000	29.0000000
speed_ground	speed_ground	800	0	79.5414195	79.6428041	19.2348870	141.2186354	27.7357153
speed_air	speed_air	200	600	103.8294713	100.9933978	10.4118729	141.7249357	90.0028586
height	height	800	0	30.1217717	30.1467453	10.2761691	59.9459639	-3.5462524
pitch	pitch	800	0	4.0183751	4.0200665	0.5248160	5.9267842	2.2844801
distance	distance	800	0	1544.52	1267.44	938.2330999	6533.05	34.0807833

Basic Parameters for Dataset2

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Maximum	Minimum
no_pasg	no_pasg	150	50	60.3400000	60.5000000	7.3107717	78.0000000	44.0000000
speed_ground	speed_ground	150	50	77.9173910	76.5308198	19.8788997	141.2186354	29.2276564
speed_air	speed_air	39	161	103.2224489	100.2606698	11.6781942	141.7249357	90.1110133
height	height	150	50	30.2326030	29.2596657	10.8272955	58.0835448	-3.5462524
pitch	pitch	150	50	4.0238987	3.9877143	0.5342237	5.5563992	2.6689057
distance	distance	150	50	1571.77	1271.99	1005.55	6533.05	425.8585610

From the above output, we find that there are 50 missing rows in dataset 2 which is missing across all the variables and has to be removed.

STEP 3- Removing missing rows from the dataset:

We use **Options missing** statement to define the missing variables as a space. We then define a new dataset and **Set** the dataset 2 to it and find the missing values which is missing across all the variables and remove them.

```
/*Removing missing rows in dataset 2*/
options missing=' ';
data flightinfo2;
set FAA2;
if missing(cats(of _all_)) then delete;
run;
proc print data=flightinfo2;
run;
```

Basic Parameters for Dataset 2 after clearing missing rows

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Maximum	Minimum
no_pasg	no_pasg	150	0	60.3400000	60.5000000	7.3107717	78.0000000	44.0000000
speed_ground	speed_ground	150	0	77.9173910	76.5308198	19.8788997	141.2186354	29.2276564
speed_air	speed_air	39	111	103.2224489	100.2606698	11.6781942	141.7249357	90.1110133
height	height	150	0	30.2326030	29.2596657	10.8272955	58.0835448	-3.5462524
pitch	pitch	150	0	4.0238987	3.9877143	0.5342237	5.5563992	2.6689057
distance	distance	150	0	1571.77	1271.99	1005.55	6533.05	425.8585610

STEP 4- Concatenating the multiple datasets into a single dataset

We use **Set** statement to concatenate both the datasets into a single dataset.

```
/*Combining datasets*/
data dataset;
set flightinfo1 flightinfo2;
run;
```

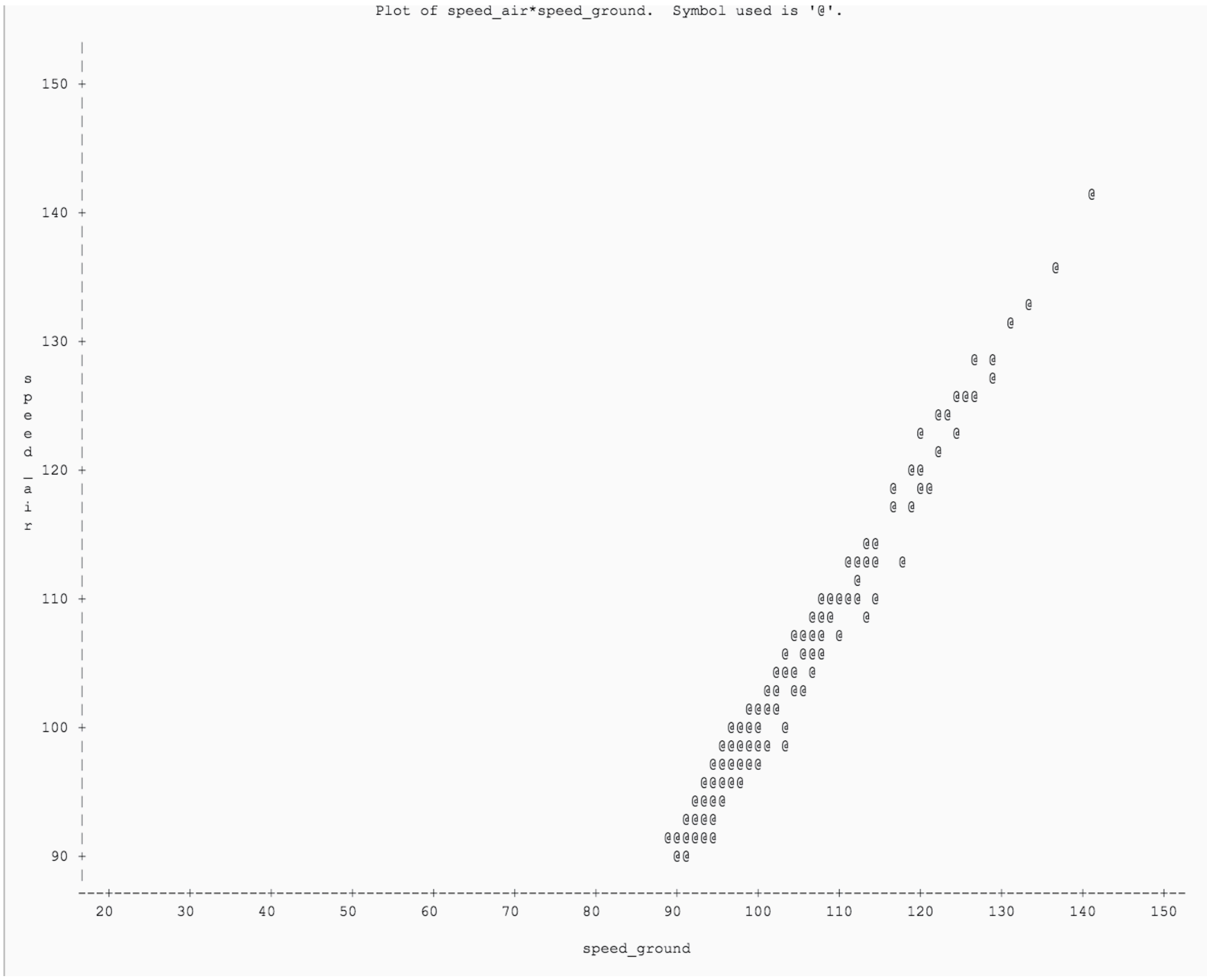
STEP 5- Finding missing values and understanding them

We use **Proc Means** statement to find the missing values after combining the dataset. We get a total of 950 observations with 150 missing values in duration (which is completely absent in dataset 2) and 711 missing values in speed_ air. We understand the correlation between speed_air and speed_ground which would have caused the missing values. There is no value of speed_air when speed_ground is less than 90 from the plot.

```
/*Finding missing values*/
proc means data=dataset n nmiss mean median std min max;
run;
/*Understanding the missing values*/
proc plot data=dataset;
plot speed_air*speed_ground='@';
run;
```

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
duration	duration	800	150	154.0065385	153.9480975	49.2592338	14.7642071	305.6217107
no_pasg	no_pasg	950	0	60.1652632	60.0000000	7.4900041	29.0000000	87.0000000
speed_ground	speed_ground	950	0	79.2849940	79.4129094	19.3364178	27.7357153	141.2186354
speed_air	speed_air	239	711	103.7304174	100.8916770	10.6051134	90.0028586	141.7249357
height	height	950	0	30.1392714	29.9044945	10.3593491	-3.5462524	59.9459639
pitch	pitch	950	0	4.0192472	4.0153874	0.5260322	2.2844801	5.9267842
distance	distance	950	0	1548.82	1267.44	948.6812561	34.0807833	6533.05



STEP 6- Duplicates check

We use **Proc Sort** statement to sort the data by all parameters (except duration which is absent in dataset2) along with **nodupkey** to remove the duplicates and export to a new dataset.
We use **Proc means** statement to find the basic parameters again after removing duplicates. The data has by 100 observations, which were exact duplicates to 850 observations.

```
/*Duplicates check*/
proc sort data=dataset nodupkey
out=dataset_nodups;
by aircraft no_pasg speed_ground speed_air height pitch
distance;
run;
proc means data=dataset_nodups n nmiss mean median min max;
run;
```

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Minimum	Maximum
duration	duration	800	50	154.0065385	153.9480975	14.7642071	305.6217107
no_pasg	no_pasg	850	0	60.1035294	60.0000000	29.0000000	87.0000000
speed_ground	speed_ground	850	0	79.4523229	79.6428041	27.7357153	141.2186354
speed_air	speed_air	208	642	103.7977237	101.1473493	90.0028586	141.7249357
height	height	850	0	30.1442223	30.0931324	-3.5462524	59.9459639
pitch	pitch	850	0	4.0093577	4.0082875	2.2844801	5.9267842
distance	distance	850	0	1526.02	1258.09	34.0807833	6533.05

STEP 7- Applying conditions to the data to subset the abnormal data and finding the count of the abnormal data

We use the conditions provided in the project to subset the data and find the abnormal data using **IF and Then** statements. We also define similar classes for the normal and missing data.
We then use **Proc Freq** statement to find the count of all the different classes and find 17 abnormal values and 638 missing values in the dataset. We don't remove the missing values as they are not consistent over all rows.

```
/*Finding abnormal values*/
data dataset2;
set dataset_nodups;
if duration=" " then Test='MISSING';
else if speed_air=" " then Test='MISSING';
else TEST='NORMAL';
if duration NE " " and duration<40 then Test='ABNORMAL' ;
if speed_ground NE " " and 140<speed_ground<30  then
Test='ABNORMAL' ;
if speed_air NE " " and 140<speed_air<30  then Test='ABNORMAL' ;
if height NE " " and height<6 then Test='ABNORMAL' ;
if distance NE " " and distance>6000 then Test='ABNORMAL' ;
run;
/*Count of abnormal data*/
proc freq data=dataset2;
table test;
run;
proc means data=dataset2 n nmiss mean median min max;
```

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Minimum	Maximum
duration	duration	800	50	154.0065385	153.9480975	14.7642071	305.6217107
no_pasg	no_pasg	850	0	60.1035294	60.0000000	29.0000000	87.0000000
speed_ground	speed_ground	850	0	79.4523229	79.6428041	27.7357153	141.2186354
speed_air	speed_air	208	642	103.7977237	101.1473493	90.0028586	141.7249357
height	height	850	0	30.1442223	30.0931324	-3.5462524	59.9459639
pitch	pitch	850	0	4.0093577	4.0082875	2.2844801	5.9267842
distance	distance	850	0	1526.02	1258.09	34.0807833	6533.05

The FREQ Procedure

Test	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ABNORMA	17	2.00	17	2.00
MISSING	638	75.06	655	77.06
NORMAL	195	22.94	850	100.00

STEP 8- Removing abnormal data

We remove the abnormal data using **IF condition** to subset only the normal and missing data. We then use **Proc Freq** statement to find the count of all the different classes and find that the 17 abnormal values are removed.

```
/*Removing abnormal data*/
data dataset3;
set dataset2;
if upcase(Test)='MISSING' or upcase(Test)= 'NORMAL';
;
proc freq data=dataset3;
table test;
run;
proc means data=dataset3 n nmiss mean median min max;
run;
```

The FREQ Procedure

Test	Frequency	Percent	Cumulative Frequency	Cumulative Percent
MISSING	638	76.59	638	76.59
NORMAL	195	23.41	833	100.00

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Minimum	Maximum
duration	duration	783	50	154.8336063	154.2845505	41.9493694	305.6217107
no_pasg	no_pasg	833	0	60.0396158	60.0000000	29.0000000	87.0000000
speed_ground	speed_ground	833	0	79.4201043	79.7453159	27.7357153	132.7846766
speed_air	speed_air	203	630	103.4850352	101.1189240	90.0028586	132.9114649
height	height	833	0	30.4420643	30.1586800	6.2275178	59.9459639
pitch	pitch	833	0	4.0060661	4.0035958	2.2844801	5.9267842
distance	distance	833	0	1521.71	1262.15	41.7223127	5381.96

STEP 9- Finding the basic parameters for all classes of data

We find the basic parameters for all the classes of data using **Proc Means** statement again after sorting the data using **Proc Sort** by aircraft and by the test variables of missing and normal data.

```
/*Finding the basic parameters for all classes of data*/
proc sort data=dataset3;
by aircraft Test;
run;
proc means data=dataset3 n nmiss mean median min max;
by aircraft Test;
var duration no_pasg speed_ground speed_air  height pitch
distance;
run;
```

The MEANS Procedure

aircraft=airbus Test=MISSING

Variable	Label	N	N Miss	Mean	Median	Minimum	Maximum
duration	duration	317	50	157.6727079	157.8742146	42.1462262	305.6217107
no_pasg	no_pasg	367	0	60.1934605	60.0000000	36.0000000	87.0000000
speed_ground	speed_ground	367	0	75.1792484	76.4571738	33.5741041	113.4273968
speed_air	speed_air	8	359	103.0040324	101.9646303	97.6046173	114.8444490
height	height	367	0	30.5086484	30.2805716	6.2275178	54.2760427
pitch	pitch	367	0	3.8398256	3.8228940	2.2844801	5.5267842
distance	distance	367	0	1053.85	990.1458144	41.7223127	3738.65

aircraft=airbus Test=NORMAL

Variable	Label	N	N Miss	Mean	Median	Minimum	Maximum
duration	duration	77	0	153.7358907	150.6005316	45.5027789	264.5933748
no_pasg	no_pasg	77	0	60.3116883	62.0000000	41.0000000	80.0000000
speed_ground	speed_ground	77	0	104.4176711	101.1302048	92.9075918	131.0351822
speed_air	speed_air	77	0	104.4454218	100.8916770	95.0113646	131.3379485
height	height	77	0	30.9732535	31.2634455	9.6972160	58.2277997
pitch	pitch	77	0	3.7897389	3.8378238	2.7019237	4.9429731
distance	distance	77	0	2607.67	2440.38	1740.90	4896.29

aircraft=boeing Test=MISSING

Variable	Label	N	N Miss	Mean	Median	Minimum	Maximum
duration	duration	271	0	154.3552717	155.5186161	41.9493694	298.5223339
no_pasg	no_pasg	271	0	59.9852399	60.0000000	29.0000000	82.0000000
speed_ground	speed_ground	271	0	67.8871124	69.8802482	27.7357153	90.4995236
speed_air	speed_air	0	271				
height	height	271	0	30.4117844	30.1382747	7.5824946	59.9459639
pitch	pitch	271	0	4.2045533	4.1932644	3.0689057	5.9267842
distance	distance	271	0	1246.66	1176.03	573.6217861	2123.80

aircraft=boeing Test=NORMAL

Variable	Label	N	N Miss	Mean	Median	Minimum	Maximum
duration	duration	118	0	149.0213824	144.9462185	63.3295206	287.0025157
no_pasg	no_pasg	118	0	59.5084746	60.0000000	43.0000000	79.0000000
speed_ground	speed_ground	118	0	102.7846998	100.4315876	88.6875803	132.7846766
speed_air	speed_air	118	0	102.8909526	100.8783383	90.0028586	132.9114649
height	height	118	0	29.9578943	29.2481455	10.2552254	58.0817896
pitch	pitch	118	0	4.2084174	4.1941122	2.9931514	5.3106775
distance	distance	118	0	2899.88	2633.03	1780.11	5381.96

CONCLUSION:

We have now cleaned the data using different steps and iterations and have made the data ready for the next step of data analysis in accordance with CRISP DM methodology.

CHAPTER 2- DATA EXPLORATION

GOAL:

Data exploration is the second step in the process of data analysis and is an informative process to gain insights into the data that has been cleaned in the previous step. The process basically consists of variable identification as predictor and target variables and forming meaningful relationships between the classes of variables to understand the impact of variables. This stage involves creating the question which needs to be answered and identifying the information necessary to answer that particular question.

The steps which have been used in this assignment are:

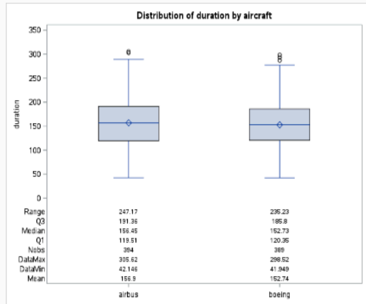
1. Plotting all the variables separated by the aircraft to understand the outliers and variation.
2. Plotting all the variables with distance to understand the effect of the variables on the landing distance.
3. Finding the Correlation between the variables to understand the relationship

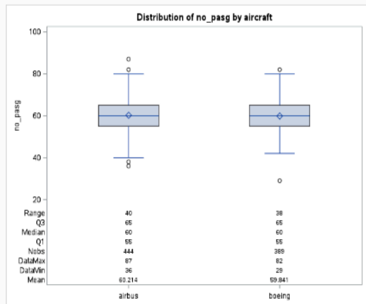
The steps may be reiterated multiple times to get a cleaner dataset when more variables are given or larger datasets are considered.

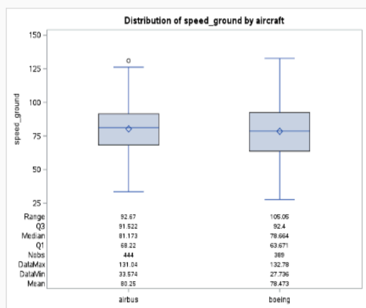
STEP 1- Plotting all the variables separated by the aircraft to understand the outliers and variation.

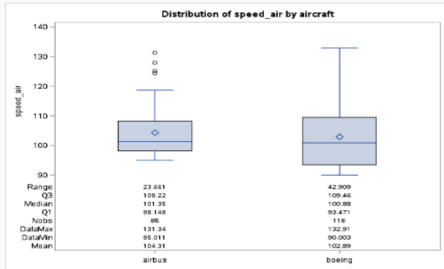
We use **Proc boxplots** on the dataset across all the variables with the aircraft type of Boeing and Airbus and the **Insetgroup** statement to have the data displayed in the box plots. We find there is a huge difference in the minimum value for distance between the aircrafts with Boeing having a higher minimum landing distance which is causing a marked change in the distribution of data.

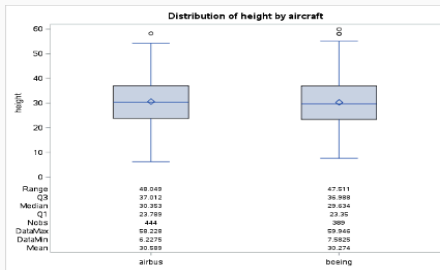
```
/*Plots for aircrafts*/
proc boxplot data=dataset3;
plot (duration no_pasg speed_ground speed_air height pitch
distance)*aircraft/
boxstyle = schematic
nohlabel;
insetgroup mean min max n nhigh nlow nout q1 q2 q3 range;
run;
```

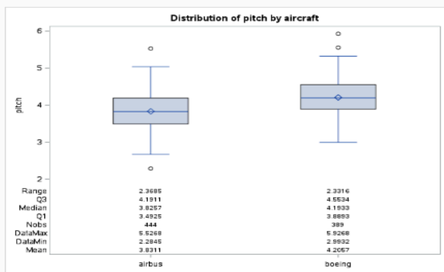

Distribution of duration by aircraft
duration
350
300
250
200
150
100
50
0
Range 247.17 235.23
Q3 191.36 185.8
Median 156.45 152.73
Q1 119.51 120.35
Nobs 394 389
DataMax 305.62 298.52
DataMin 42.146 41.949
Mean 156.9 152.74
airbus
boeing

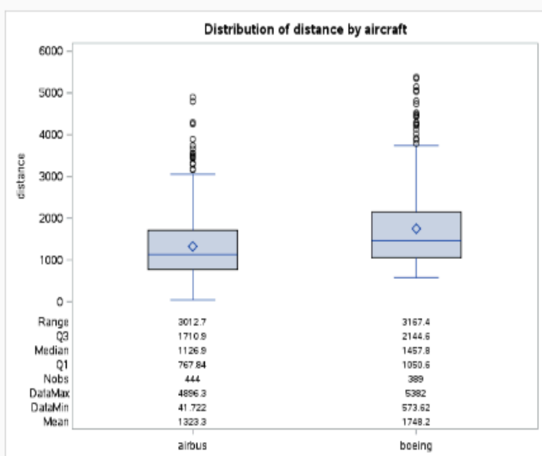

Distribution of no_pasg by aircraft
no_pasg
100
80
60
40
20
Range 40 38
Q3 65 65
Median 60 60
Q1 55 55
Nobs 444 389
DataMax 87 82
DataMin 36 29
Mean 60.214 59.841
airbus
boeing


Distribution of speed_ground by aircraft
speed_ground
150
125
100
75
50
25
Range 92.67 105.05
Q3 91.522 92.4
Median 81.173 78.664
Q1 68.22 63.671
Nobs 444 389
DataMax 131.04 132.78
DataMin 33.574 27.736
Mean 80.25 78.473
airbus
boeing


Distribution of speed_air by aircraft
speed_air
Range 23.661 42.909
Q3 108.22 109.46
Median 101.35 100.88
Q1 98.148 93.471
Nobs 85 118
DataMax 131.34 132.91
DataMin 95.011 90.003
Mean 104.31 102.89
airbus
boeing


Distribution of height by aircraft
height
Range 48.049 47.511
Q3 37.012 36.988
Median 30.353 29.634
Q1 23.789 23.35
Nobs 444 389
DataMax 58.228 59.946
DataMin 6.2275 7.5825
Mean 30.589 30.274
airbus
boeing

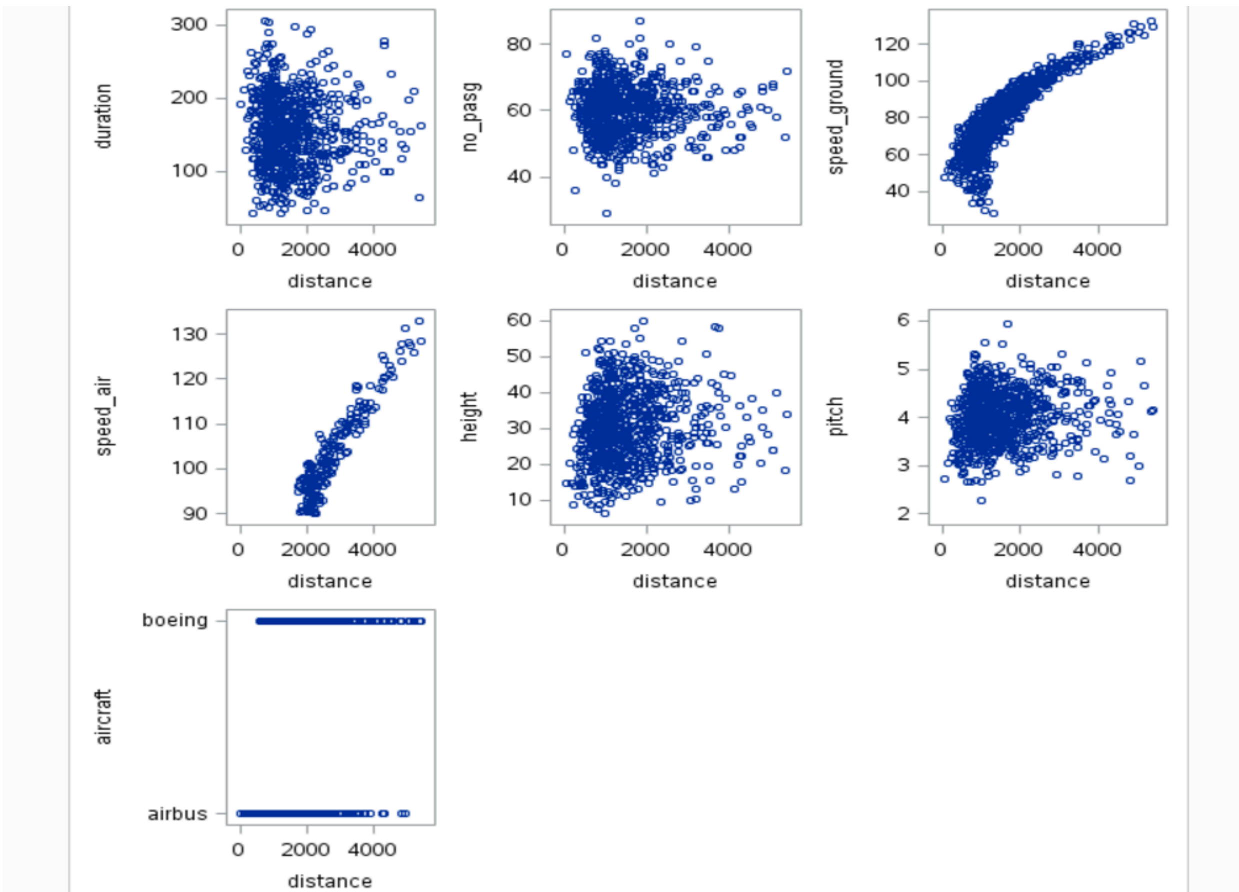

Distribution of pitch by aircraft
pitch
Range 2.3685 2.3316
Q3 4.1911 4.5534
Median 3.8257 4.1933
Q1 3.4925 3.8893
Nobs 444 389
DataMax 5.5268 5.9268
DataMin 2.2845 2.9932
Mean 3.8311 4.2057
airbus
boeing


Distribution of distance by aircraft
distance
Range 3012.7 3167.4
Q3 1710.9 2144.6
Median 1126.9 1457.8
Q1 767.84 1050.6
Nobs 444 389
DataMax 4896.3 5382
DataMin 41.722 573.62
Mean 1323.3 1748.2
airbus
boeing

<u>STEP 2- Plotting all the variables with distance to understand the effect of the variables on the landing distance.</u>

We use **Proc Sgscatter** and the **Plot function** to plot multiple plots of all the variables with distance in a single plot. We see a strong linear and positive relationship only for Speed_ground and Speed_air with the distance.

```
/*Plots for distance*/
proc sgscatter data=dataset3;
plot(duration no_pasg speed_ground speed_air height pitch
aircraft)*distance;
run;
```



<u>STEP 3- Finding the Correlation between the variables to understand the relationship.</u>

We use **Proc Corr** to correlate the data by the variables to understand the correlation between the different variables pairwise. We find a strong correlation between distance and speed_air and speed_ground as expected from the plots. We also find that there is a correlation between speed_ground and speed_air which was shown before when understanding the missing data.

```
/*Correlation*/
proc corr data=dataset3;
var duration no_pasg speed_ground speed_air height pitch
distance;
Title Pairwise Correlation Coefficients;
run;
```

Pairwise Correlation Coefficients

The CORR Procedure

7 Variables:	duration no_pasg speed_ground speed_air height pitch distance

Simple Statistics							
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
duration	783	154.83361	48.33498	121235	41.94937	305.62171	duration
no_pasg	833	60.03962	7.49821	50013	29.00000	87.00000	no_pasg
speed_ground	833	79.42010	18.87950	66157	27.73572	132.78468	speed_ground
speed_air	203	103.48504	9.73628	21007	90.00286	132.91146	speed_air
height	833	30.44206	9.77839	25358	6.22752	59.94596	height
pitch	833	4.00607	0.52626	3337	2.28448	5.92678	pitch
distance	833	1522	895.41965	1267584	41.72231	5382	distance

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations							
	duration	**no_pasg**	**speed_ground**	**speed_air**	**height**	**pitch**	**distance**
duration duration	1.00000 783	-0.03545 0.3219 783	-0.05161 0.1491 783	0.04454 0.5364 195	0.01019 0.7758 783	-0.04578 0.2006 783	-0.05203 0.1458 783
no_pasg no_pasg	-0.03545 0.3219 783	1.00000 833	0.00563 0.8710 833	-0.00616 0.9305 203	0.04814 0.1651 833	-0.01932 0.5776 833	-0.01729 0.6182 833
speed_ground speed_ground	-0.05161 0.1491 783	0.00563 0.8710 833	1.00000 833	0.98794 <.0001 203	-0.05271 0.1285 833	-0.04340 0.2108 833	0.86078 <.0001 833
speed_air speed_air	0.04454 0.5364 195	-0.00616 0.9305 203	0.98794 <.0001 203	1.00000 203	-0.07933 0.2606 203	-0.03927 0.5780 203	0.94210 <.0001 203
height height	0.01019 0.7758 783	0.04814 0.1651 833	-0.05271 0.1285 833	-0.07933 0.2606 203	1.00000 833	0.02180 0.5299 833	0.09994 0.0039 833
pitch pitch	-0.04578 0.2006 783	-0.01932 0.5776 833	-0.04340 0.2108 833	-0.03927 0.5780 203	0.02180 0.5299 833	1.00000 833	0.08633 0.0127 833
distance distance	-0.05203 0.1458 783	-0.01729 0.6182 833	0.86078 <.0001 833	0.94210 <.0001 203	0.09994 0.0039 833	0.08633 0.0127 833	1.00000 833

CONCLUSION:

We have now explored the data in accordance with CRISP DM methodology using different steps and iterations and have understood the variables which have to be modelled to understand the relationship with landing distance which are- speed_air, speed_ground, height and pitch as their p values are less than 0.05 making statistically significant relationship.

CHAPTER 3 AND CHAPTER 4- MODELING AND MODEL CHECKING

GOAL:

Modeling is the third step in the process of data analysis and is used to model the predictor variables which are found to have an impact on the target variables found in the data exploration stage. In this step, typically various modeling techniques are selected and then applied and all the parameters calculated are then calibrated to obtain the optimal values. There are many modeling techniques which can be applied depending on the number of variables and the extent of data that is being provided.

The steps which have been used in this assignment are:

1. Regression Analysis
2. Regression Analysis using significant variables
3. Regression Analysis using transformed variable
4. Check the distribution of residuals by plotting to check for independance
5. Check if the distribution of residuals is normally distributed
6. Perform hypothesis testing on the residuals

STEP 1- Using regression analysis on the distance variable with respect to all variables.

We use **Proc Reg** to perform regression analysis for distance as response variable and all the remaining variables except aircraft as the predictor variables.
We find that only 195 values were considered due to lot of missing values.
We find p values below 0.05 level of significance for aircraft, speed_air and height showing statistically significant relationship with distance, which is also true from the correlation matrix.
But we find no relation for speed_ground as found from correlation matrix as speed_ground and speed_air is highly correlated.
We use speed_ground for further analysis as it does not have any missing values.
We find that adjusted R square value is 0.9738 i.e. 97.38% of the variability is explained by this model.

```
/*Regression Analysis*/
data dataset4;
set dataset3;
if aircraft='boeing' then aircraftcode = 0;
else aircraftcode = 1;
proc reg data=dataset4;
model distance= aircraftcode duration no_pasg speed_ground
speed_air height pitch ;
title Regression analysis of the data set;
run;
```

Regression analysis of the data set

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	833
Number of Observations Used	195
Number of Observations with Missing Values	638

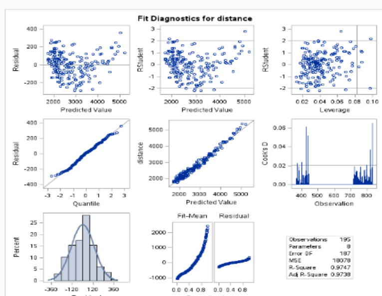
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	130295276	18613611	1029.64	<.0001
Error	187	3380533	18078		
Corrected Total	194	133675809			

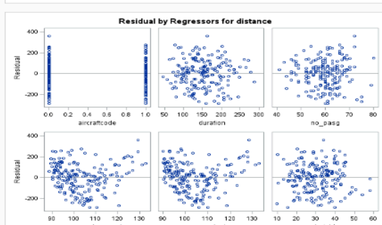
Root MSE	134.45339	R-Square	0.9747
Dependent Mean	2784.49158	Adj R-Sq	0.9738
Coeff Var	4.82865		

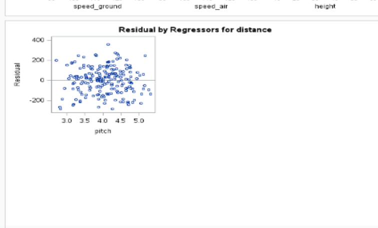
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5791.65727	163.23457	-35.48	<.0001
aircraftcode		1	-437.94277	21.26212	-20.60	<.0001
duration	duration	1	0.12763	0.20394	0.63	0.5322
no_pasg	no_pasg	1	-1.98117	1.37797	-1.44	0.1522
speed_ground	speed_ground	1	-3.54637	6.41601	-0.55	0.5811
speed_air	speed_air	1	85.54689	6.52207	13.12	<.0001
height	height	1	13.67560	1.03856	13.17	<.0001
pitch	pitch	1	-13.48974	18.60775	-0.72	0.4694

Regression analysis of the data set

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance







STEP 2- Using regression analysis using only significant variables.

We use **Proc Reg** using only the three significant variables in the model. We find that 833 observations are now taken into account. The adjusted R square value predicts about 84% of the variability. We find all the variables are statistically significant for distance.

```
/*Second iteration using significant variables*/
proc reg data=dataset4;
model distance= aircraftcode speed_ground height;
title Regression analysis of significant variables;
run;
```

Regression analysis of significant variables

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	833
Number of Observations Used	833

Analysis of Variance

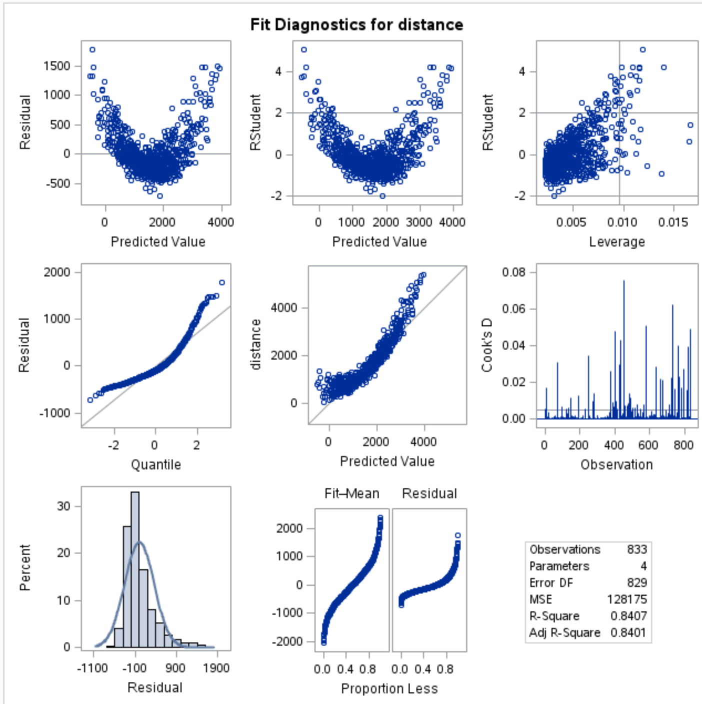
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	560821234	186940411	1458.48	<.0001
Error	829	106256689	128175		
Corrected Total	832	667077923			

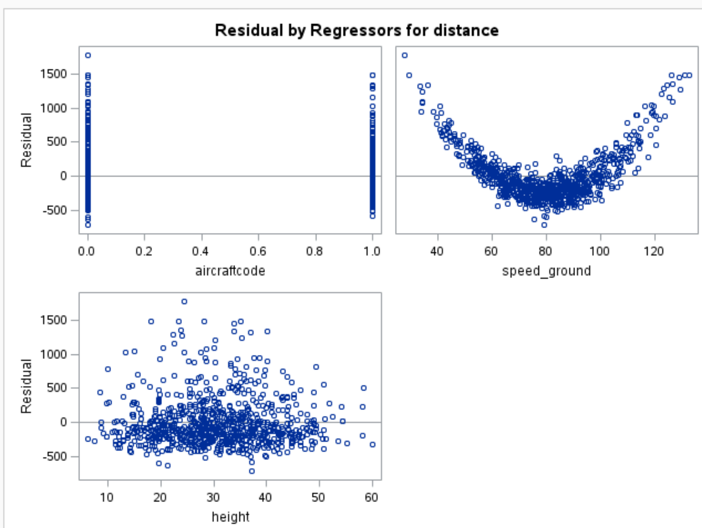
Root MSE	358.01471	R-Square	0.8407
Dependent Mean	1521.70929	Adj R-Sq	0.8401
Coeff Var	23.52714		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1952.65302	68.54959	-28.49	<.0001
aircraftcode		1	-503.51307	24.89500	-20.23	<.0001
speed_ground	speed_ground	1	41.82800	0.65910	63.46	<.0001
height	height	1	13.82158	1.27131	10.87	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance





STEP 3- Regression analysis using transformed variable

We use a new transformed variable for speed_ground which is a square of the variable in accordance with the transformation principles. We then use **Proc Reg** on the dataset with the four variables to achieve an optimum model. We get an adjusted R square of 97.60% to explain the variability for this model. We also find our assumptions of no relationship in residuals satisfied in the model. The final regression equation can be summed up as

y(distance)= 2230.448-399.37x_1+0.69815$x_2{}^2$- 69.97x_2+13.44x_3
where x_1= make of an aircraft
x_2 = ground speed of the aircraft
x_3 = height of the aircraft when passing over the run way.

```
/*Third Iteration to achieve randomness of residuals*/
data dataset5;
set dataset4;
speed_ground_squared = speed_ground**2;
proc reg data=dataset5;
model distance= aircraftcode speed_ground_squared speed_ground
height;
title Regression analysis adjusting for speed_ground;
run;
```

Regression analysis adjusting for speed_ground

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	833
Number of Observations Used	833

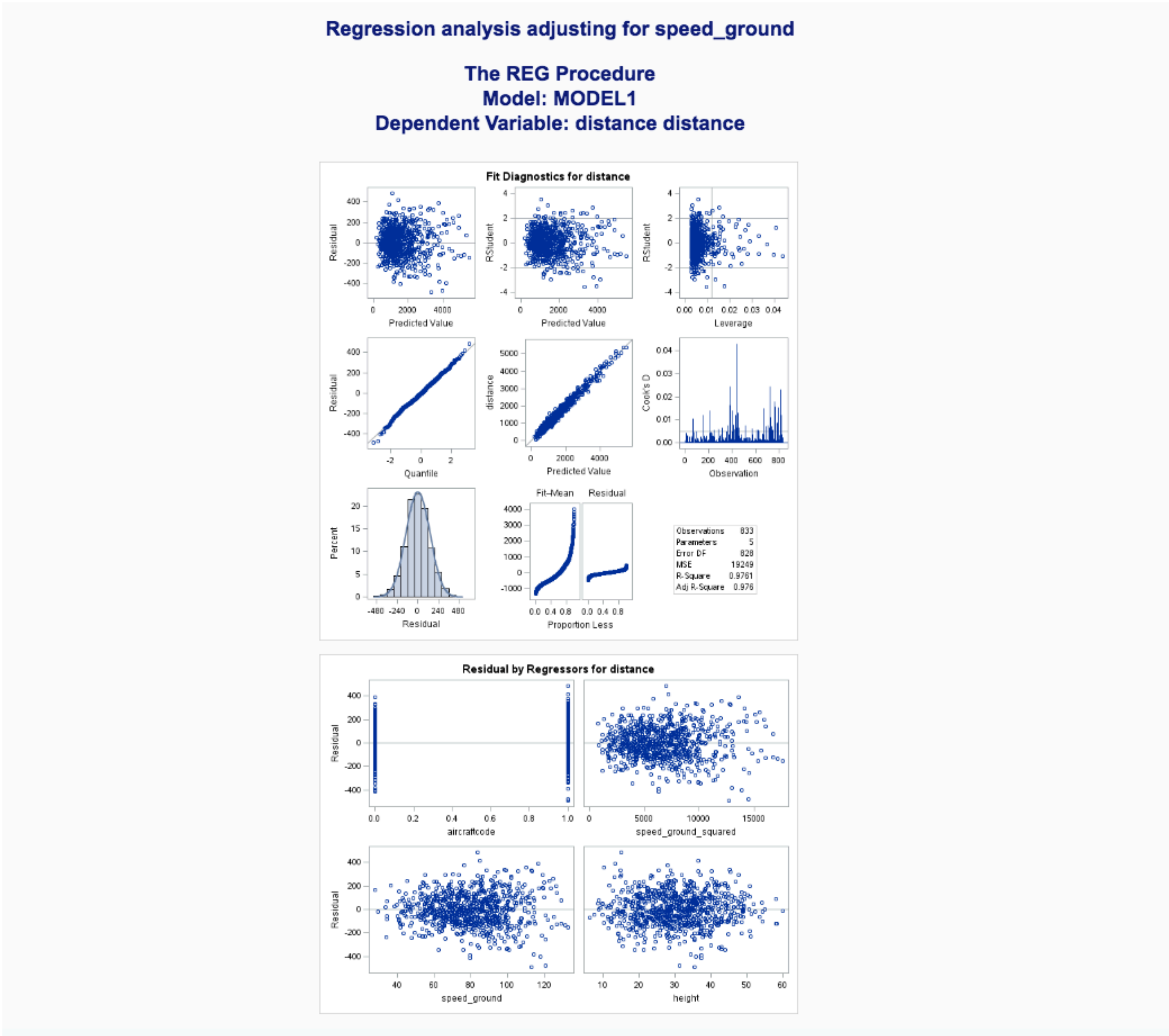
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	651139538	162784884	8456.68	<.0001
Error	828	15938385	19249		
Corrected Total	832	667077923			

Root MSE	138.74169	R-Square	0.9761
Dependent Mean	1521.70929	Adj R-Sq	0.9760
Coeff Var	9.11749		

Parameter Estimates

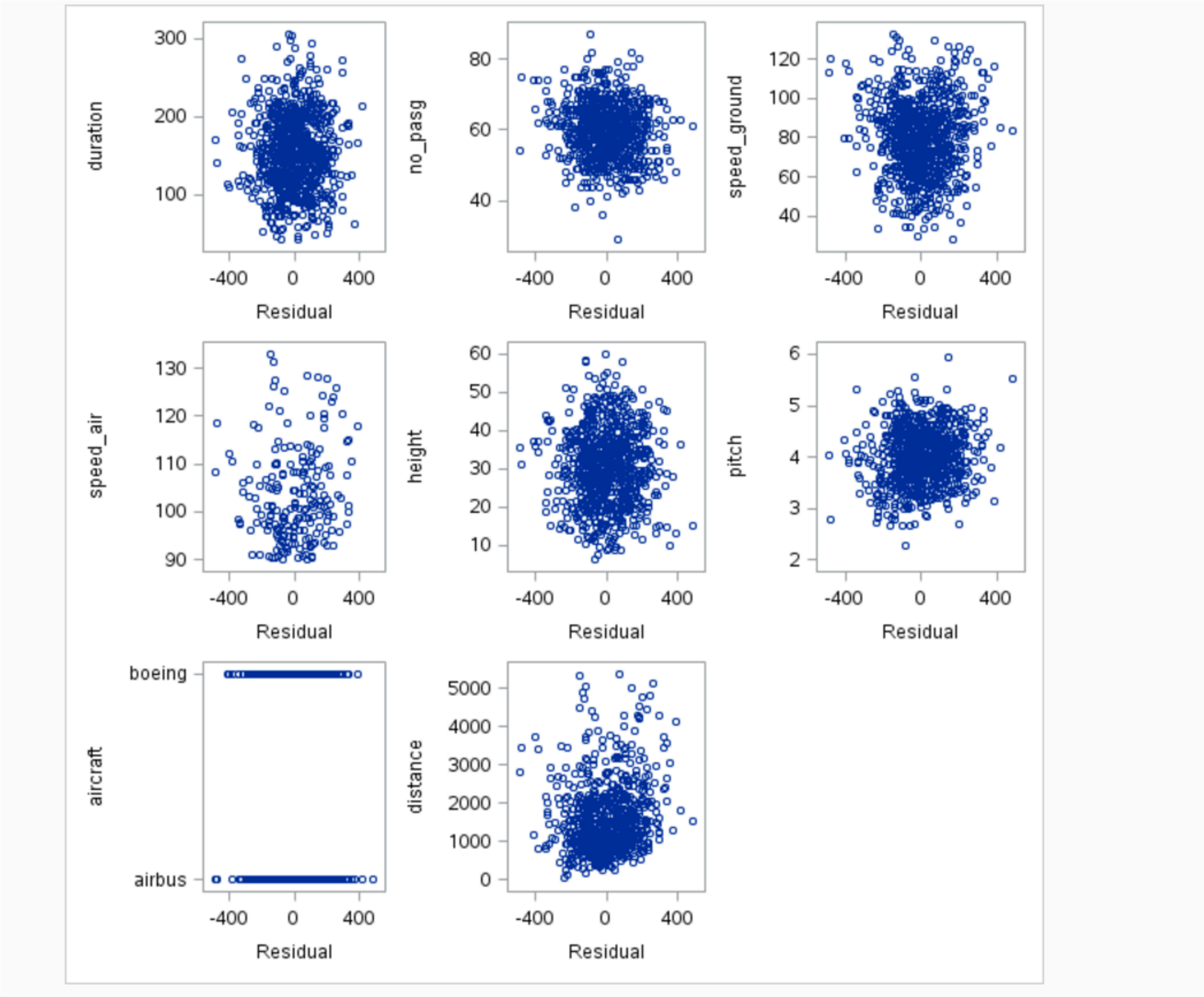
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2230.44821	66.59631	33.49	<.0001
aircraftcode		1	-399.37924	9.76662	-40.89	<.0001
speed_ground_squared		1	0.69815	0.01019	68.50	<.0001
speed_ground	speed_ground	1	-69.97520	1.65206	-42.36	<.0001
height	height	1	13.44866	0.49270	27.30	<.0001



STEP 4: Check the distribution of residuals by plotting

We can check the distribution of residuals by plotting all the residual against all the different variables using scatter plots. We don't find any relationship being explained by the residuals which satisfies the first condition of error terms being independent.

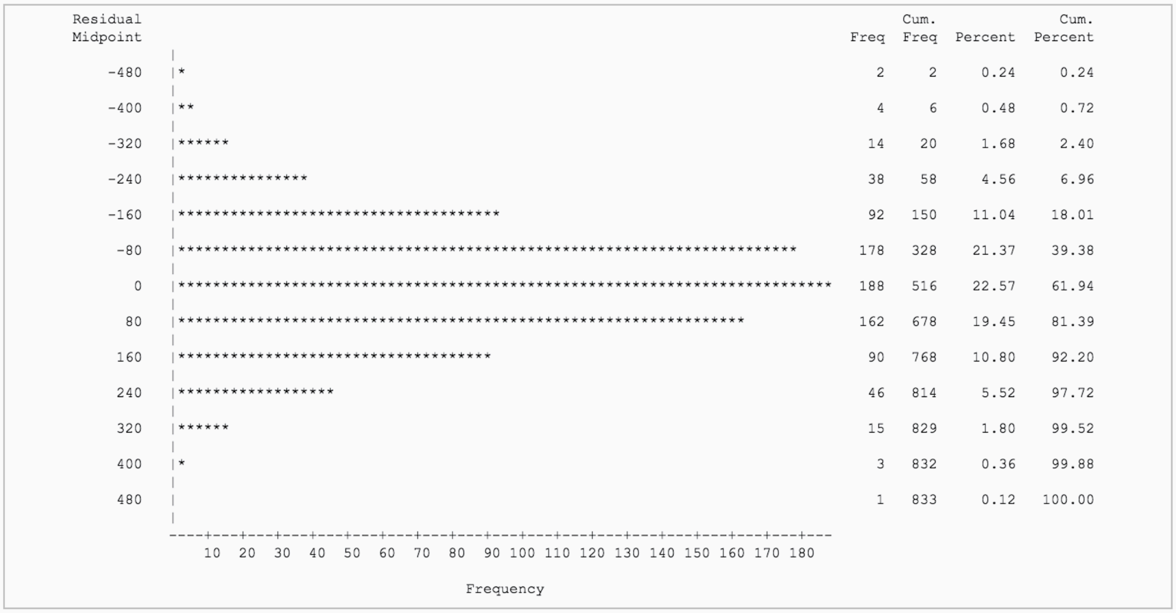
```
/*Plotting the residuals*/
proc sgscatter data=diagnostics;
plot(duration no_pasg speed_ground speed_air height pitch
aircraft distance)*residual;
run;
```



STEP 5: Check if the distribution of residuals is normally distributed

We use **Proc Chart** function on the diagnostic data and plot the histogram of the residuals which will test for normality of the graphs and is found to be normal.

```
/*Plot the histogram of residuals*/
proc chart data= diagnostics;
hbar residual;
run;
```



STEP 6: Perform hypothesis testing on the residuals

We use **Proc TTest** to test the residuals for the normality and for mean=0. We find that the data is normally distributed and that the p value is around 1 which helps us to accept the null hypothesis of mean=0.

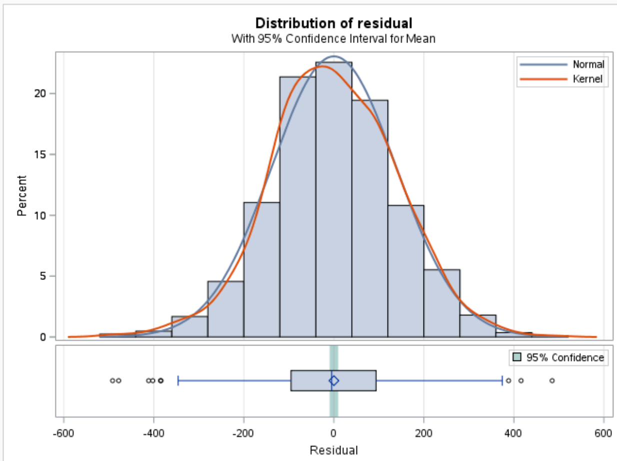
```
/* T Test*/
proc ttest data=diagnostics;
var residual;
run;
```

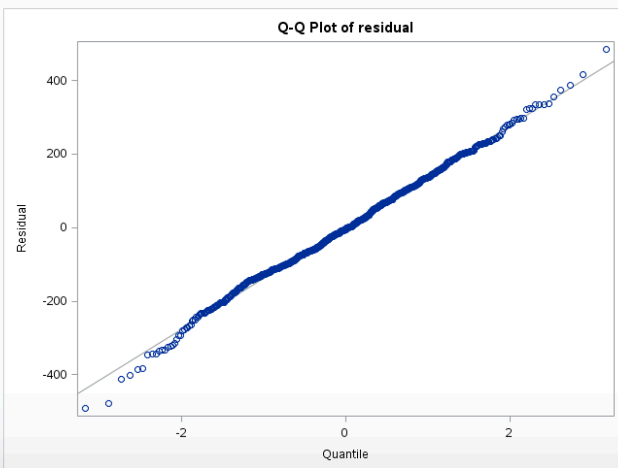
The TTEST Procedure
Variable: residual (Residual)

N	Mean	Std Dev	Std Err	Minimum	Maximum
833	-357E-14	138.4	4.7955	-491.8	485.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-357E-14	-9.4128	9.4128	138.4	132.1	145.4

DF	t Value	Pr > |t|
832	-0.00	1.0000





CONCLUSION:

We have now modeled and checked the model in accordance with CRISP DM methodology using different steps and iterations. We have also reframed our model with transformation to account for more variability and to also use predictor variables to define the relationship instead of the residuals. We find all the assumptions of the regression model satisfied which is

1. The error terms are independent.
2. The error terms are normally distributed.
3. Their mean is zero
4. They have constant variance.