COMMERCIAL FLIGHT LANDING OVERRUN MODEL

BANA 6043 Project

SUMMARY

We are given data pertaining to 950 commercial flight landings. The objective is to understand the factors which affect the landing distance of a flight. This information will be used to predict and avoid runway overruns. It is found that the flight ground speed, aircraft manufacturer, and landing height are the important factors which have an impact on landing distance. Higher the ground speed or the landing height, the longer the landing distance; if the aircraft make is ‘Boeing’, it has a longer landing distance compared to an aircraft of ‘Airbus’ make, given all other parameters are the same.

Jagarlapudi Krishna Teja

UCID: M10896584

ABSTRACT

Background: Landing overrun also known as runway overruns are defined as situations when aircraft on takeoff or landing roll extends beyond the end of the runway (aviationknowledge.com). We are given a dataset which has data about a few variables including landing distance, for commercial flights.

Goal: To study the factors influencing the landing distance of a commercial flight.

Data: Given two excel files with 800 and 150 rows respectively. This is data pertaining to commercial flights and has landing distance and 7 other parameters.

Data cleaning: Rows containing abnormal values based on business knowledge are removed. Rows with same values for all variables are considered duplicates and are removed instead of considering them as coincidentally similar rows. The 950 rows of data lead to clean data from 836 commercial flights after cleaning.

Data exploration: 3 of the 8 variables are not normally distributed, with the landing distance being one among them. The air speed variable, another non-normally distributed variable, is dropped from the dataset since it has high correlation with another variable and has 75.5% missing values.

Model building: Of the remaining 6 variables used to predict landing distance 3 are significant factors – ground speed, aircraft manufacturer, and landing height. The final model has an adjusted R-squared of 84.96%

Final parameters: Below is the relation of the landing distance with each of the variables-

- Ground speed – positively correlated, i.e., higher => greater risk of landing overrun
- Landing height – positively correlated, i.e., higher => greater risk of landing overrun
- Aircraft manufacturer – Higher risk if 'Boeing' is the manufacturer as compared to 'Airbus'.

The model is able to explain the landing distance value given the above parameters 85% of the time

CONTENTS

Abstract 1
Background 1
Goal 1
Data 1
Data cleaning 1
Data exploration 1
Model building 1
Final parameters 1
Questions: 3
How many observations (flights) do you use to fit your final model? If not all 950 flights, why? 3
What factors and how they impact the landing distance of a flight? 3
Is there any difference between the two makes Boeing and Airbus? 3
Chapter 1: Data cleaning and preparation 4
Goal: 4
Observations and conclusions: 4
SAS Output: 4
SAS Code: 7
Chapter 2: Data exploration 10
Goal: 10
Observations and conclusions: 10
SAS Output: 10
SAS codes: 14
Chapter 3: Model Building 16
Goal: 16
Observations and conclusions: 16
SAS Output: 16
SAS Code: 17
Chapter 4: Model verification 18
Goal: 18
Observations and conclusions: 18
SAS Output: 18
SAS Code: 19

QUESTIONS:

HOW MANY OBSERVATIONS (FLIGHTS) DO YOU USE TO FIT YOUR FINAL MODEL? IF NOT ALL 950 FLIGHTS, WHY?

In the final model, 836 observations are used. The number reduced from 950 due to the following reasons:

- In the second data set of 150 rows, 'FAA.xls', data for 'duration' variable is not present. Comparing all the other columns of data it is found that all 'Boeing' data is a duplicate of the data in the 'FAA2.xls' file. This reduces the dataset to 850 rows.
- A total of 14 rows of abnormal data is present in the data set. This is distributed as follows –
    - Duration: 5 rows
    - Ground Speed: 3 rows
    - Air Speed: 1 row
    - Landing height: 10 rows
    - Landing distance: 2 rows

  This reduces the number of rows of data to a final 836 rows

WHAT FACTORS AND HOW THEY IMPACT THE LANDING DISTANCE OF A FLIGHT?

Below is the relation of the landing distance with each of the variables-

- Ground speed – positively correlated, i.e., higher => longer landing distance
- Landing height – positively correlated, i.e., higher => longer landing distance
- Aircraft manufacturer – longer landing distance if 'Boeing' is the manufacturer as compared to 'Airbus'.

The model is able to explain the landing distance value given the above parameters 85% of the time

IS THERE ANY DIFFERENCE BETWEEN THE TWO MAKES BOEING AND AIRBUS?

As suggested in the modelling process, the flights of Boeing make have a longer landing distance that a flight of Airbus make given that all other landing parameters are the same.

CHAPTER 1: DATA CLEANING AND PREPARATION

GOAL:

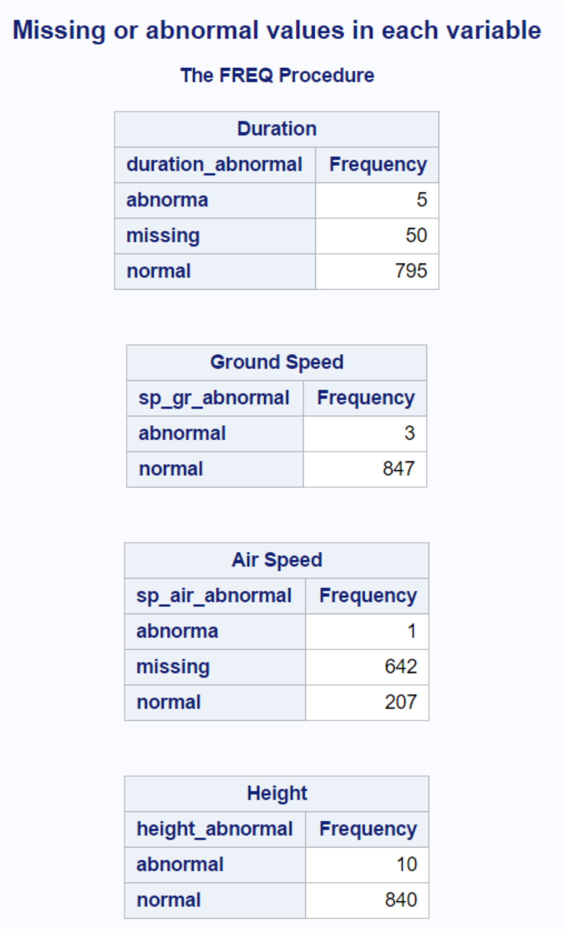
The goal of this step is to take the given data and make it ready for use for regression. This also involves the understanding of the various variables in the data set and how they are distributed.

OBSERVATIONS AND CONCLUSIONS:

- The data set ‘FAA2.xls’ has 200 rows of data with 50 blank rows at the end.
- The data for ‘duration’ variable is missing in ‘FAA2.xls’ file. Also, all ‘Boeing’ data is a duplicate of data from ‘FAA1.xls’ file.
- The flight air speed variable has missing values in both data sets – a total of 642 rows, 75% of rows.
- Variable for landing height has the highest number of abnormal values- 10 out of 850.
- Of the 8 variables, 1 is categorical and the remaining 7 are numerical.
- The variable distributions are as follows-
  - 3 normally distributed variables – Flight duration, Height, and Pitch.
  - 3 non-normal skewed variables – Air speed, Ground speed and Landing Distance.
  - 1 non-normal symmetric variable – Number of passengers.

SAS OUTPUT:

Frequency counts and total impact:



Missing or abnormal values in each variable

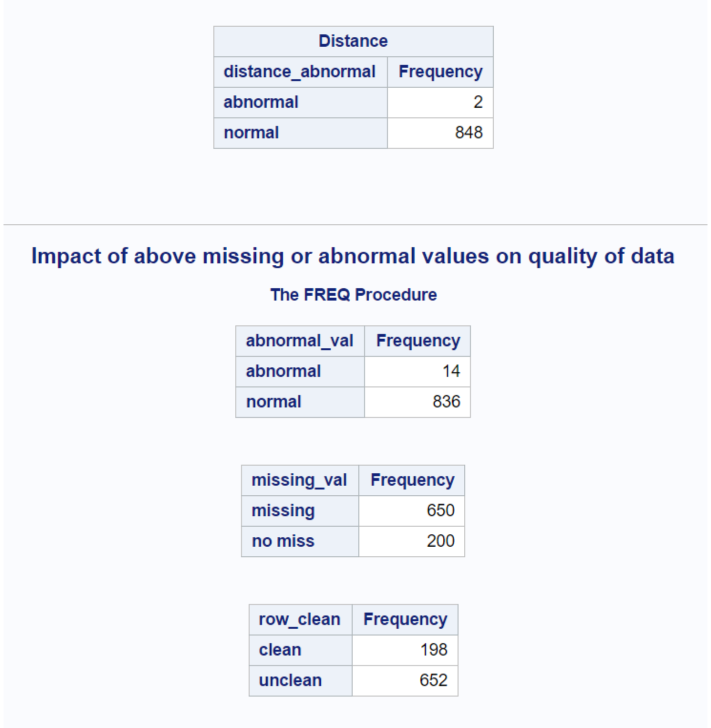
The FREQ Procedure

Duration	
duration_abnormal	Frequency
abnorma	5
missing	50
normal	795

Ground Speed	
sp_gr_abnormal	Frequency
abnormal	3
normal	847

Air Speed	
sp_air_abnormal	Frequency
abnorma	1
missing	642
normal	207

Height	
height_abnormal	Frequency
abnormal	10
normal	840



Distance	
distance_abnormal	Frequency
abnormal	2
normal	848

Impact of above missing or abnormal values on quality of data

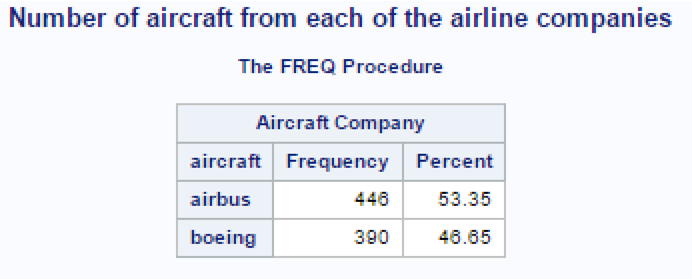
The FREQ Procedure

abnormal_val	Frequency
abnormal	14
normal	836

missing_val	Frequency
missing	650
no miss	200

row_clean	Frequency
clean	198
unclean	652

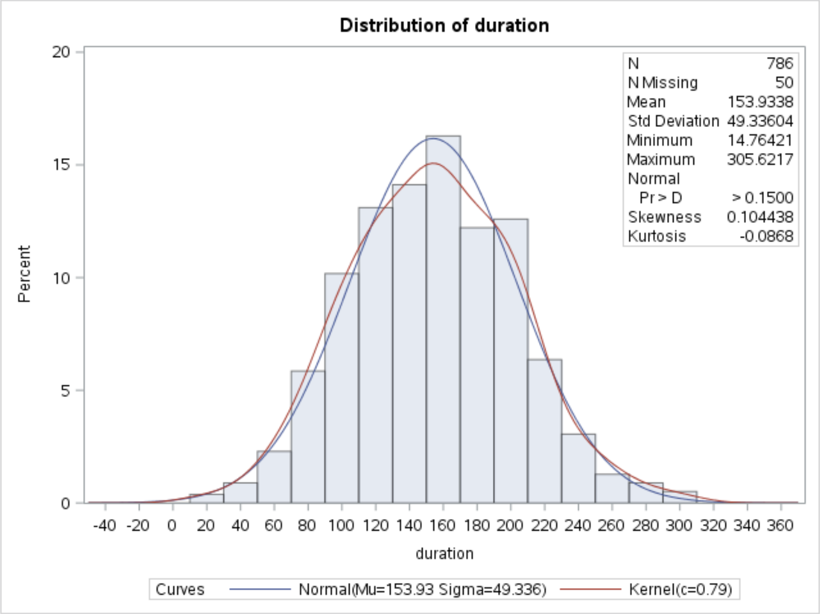
Distribution of each variable:

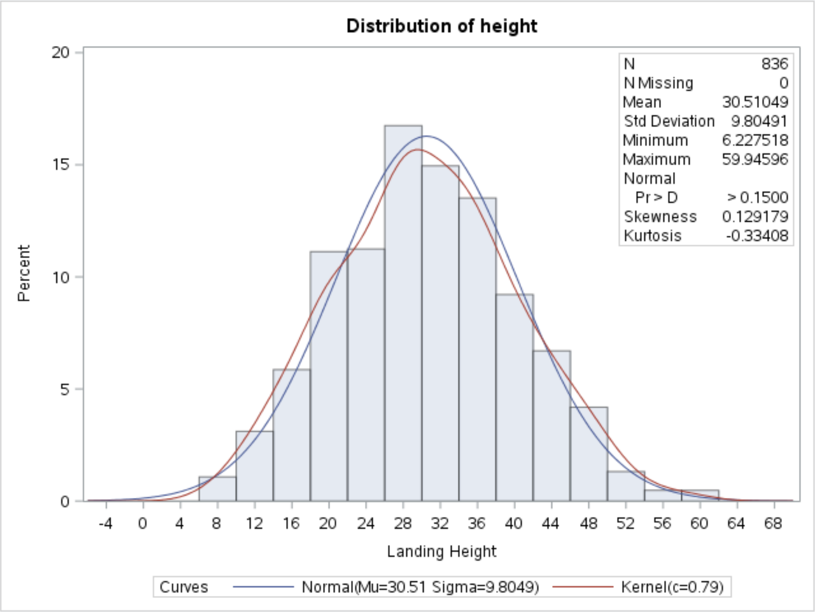


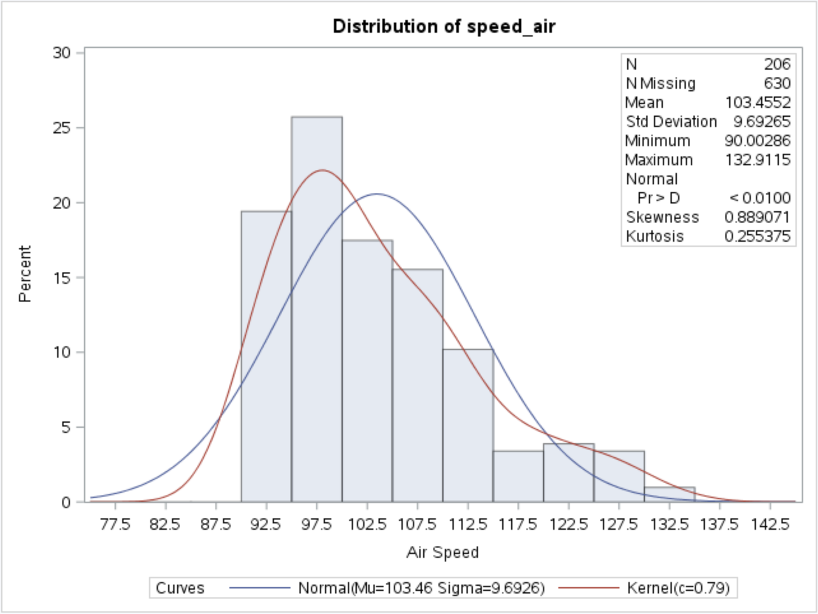
Number of aircraft from each of the airline companies

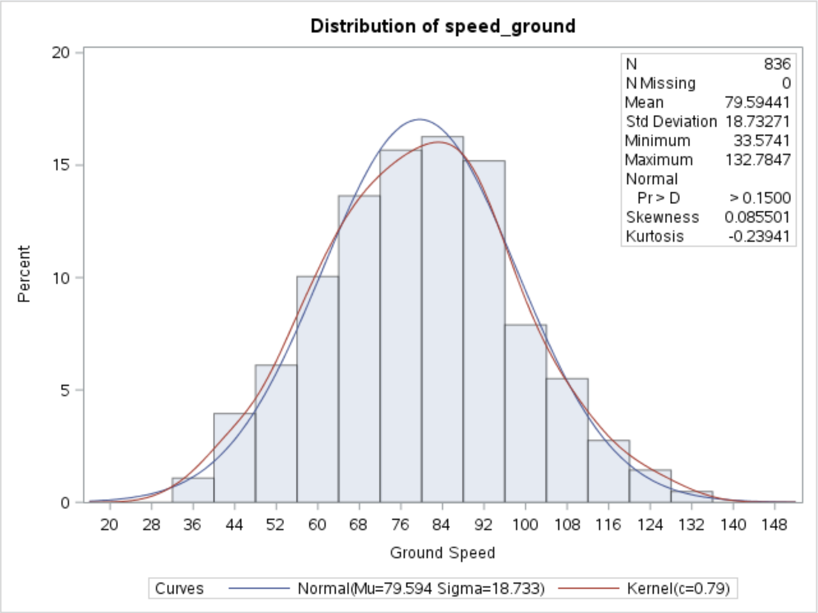
The FREQ Procedure

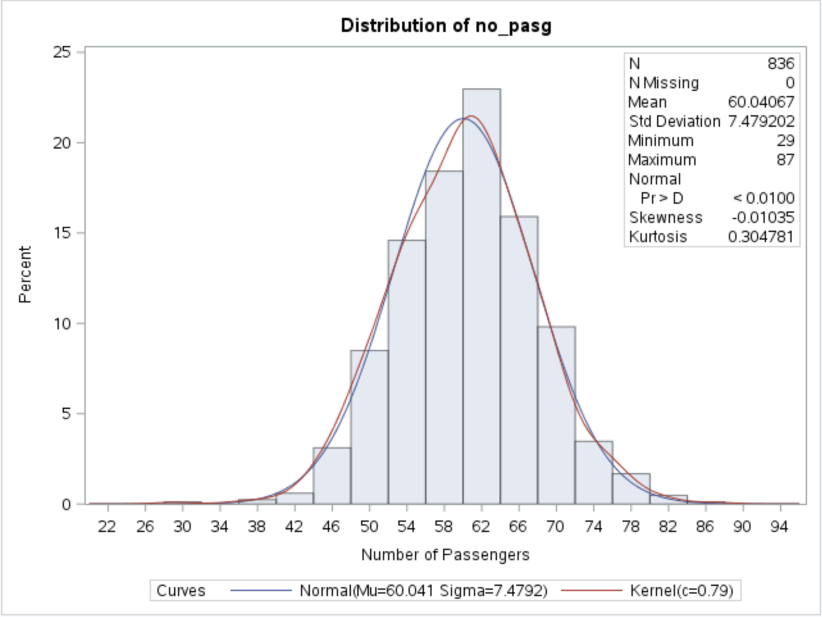
Aircraft Company		
aircraft	Frequency	Percent
airbus	446	53.35
boeing	390	46.65

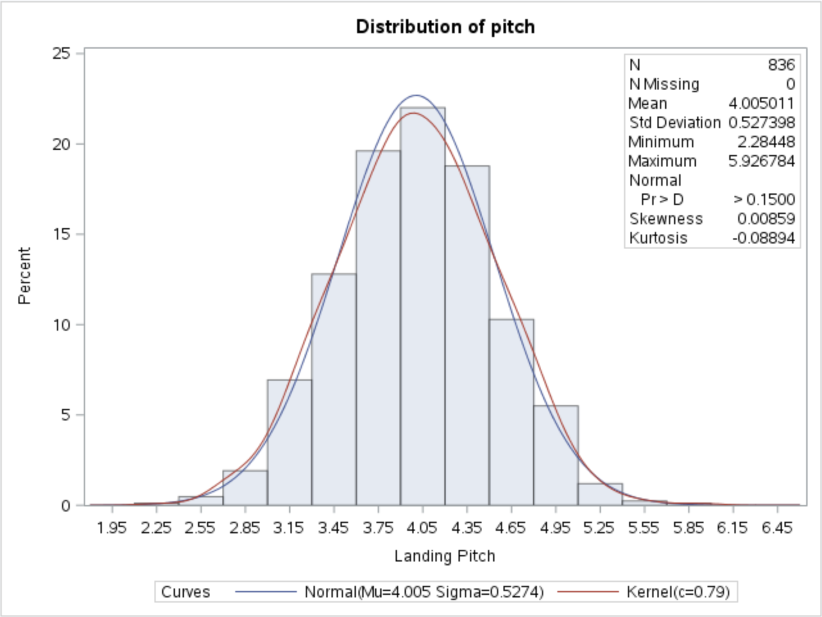


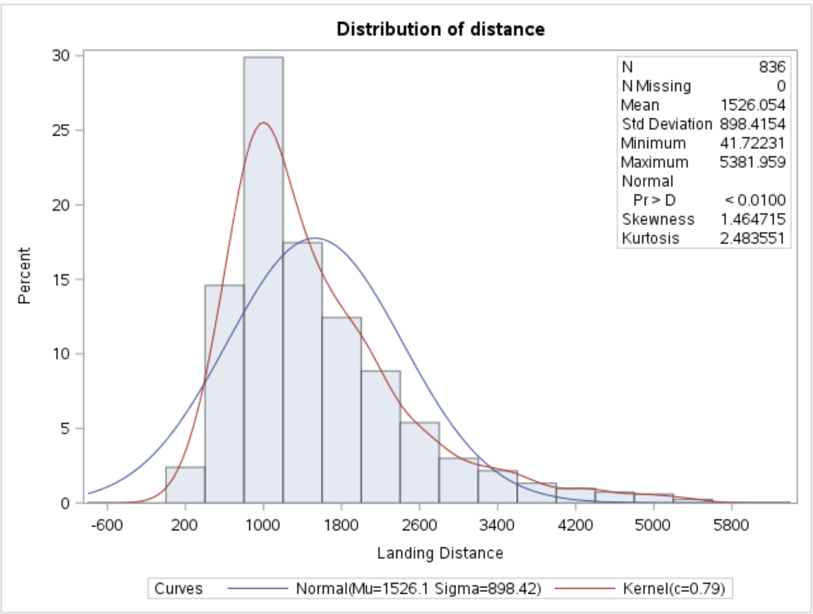



Distribution of speed_air
N 206
N Missing 630
Mean 103.4552
Std Deviation 9.69265
Minimum 90.00286
Maximum 132.9115
Normal
Pr > D < 0.0100
Skewness 0.889071
Kurtosis 0.255375
Percent
0
5
10
15
20
25
30
77.5
82.5
87.5
92.5
97.5
102.5
107.5
112.5
117.5
122.5
127.5
132.5
137.5
142.5
Air Speed
Curves
Normal(Mu=103.46 Sigma=9.6926)
Kernel(c=0.79)


Distribution of speed_ground
N 836
N Missing 0
Mean 79.59441
Std Deviation 18.73271
Minimum 33.5741
Maximum 132.7847
Normal
Pr > D > 0.1500
Skewness 0.085501
Kurtosis -0.23941
Percent
0
5
10
15
20
20
28
36
44
52
60
68
76
84
92
100
108
116
124
132
140
148
Ground Speed
Curves
Normal(Mu=79.594 Sigma=18.733)
Kernel(c=0.79)


Distribution of no_pasg
N 836
N Missing 0
Mean 60.04067
Std Deviation 7.479202
Minimum 29
Maximum 87
Normal
Pr > D < 0.0100
Skewness -0.01035
Kurtosis 0.304781
Percent
0
5
10
15
20
25
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
Number of Passengers
Curves
Normal(Mu=60.041 Sigma=7.4792)
Kernel(c=0.79)





(Note that all histograms had few basic metrics in the insets- number of observations, number of missing observations, mean, standard deviation, minimum, maximum, skewness, kurtosis and the p-value for K-S test for normality)

SAS CODE:

```
/*#####################################################################################################
Import data and combine multiple files*/
PROC IMPORT DATAFILE = "/home/jagarlka0/sasuser.v94/Mid-Term Project/FAA1.xls" OUT = faa1 replace dbms=xls;
SHEET = "FAA1"; GETNAMES = yes;
RUN;
PROC IMPORT DATAFILE = "/home/jagarlka0/sasuser.v94/Mid-Term Project/FAA2.xls" OUT = faa2 replace dbms=xls;
SHEET = "FAA2"; GETNAMES = yes;
RUN;
```

```
/*########################################################################################
Clean the table - Remove blank rows
manual observation of data sets showed that faa2.xls has blank rows at the end*/
options missing = ' ';
data faa2;
  set faa2;
  if missing(cats(of _all_)) then delete;
run;


/*########################################################################################
Merge two tables by the column names. Manual observation of data shows two things-
1) The second data set has only 7 variables compared to 8 in the first
2) Some rows in the second dataset have same values as the first, except of course for the missing variable*/
proc sort data=faa1;
by aircraft no_pasg speed_ground speed_air height pitch distance;
run;
proc sort data=faa2;
by aircraft no_pasg speed_ground speed_air height pitch distance;
run;
data faa;
merge faa1 faa2;
by aircraft no_pasg speed_ground speed_air height pitch distance;
label aircraft='Aircraft Company' no_pasg='Number of Passengers' speed_ground='Ground Speed' speed_air='Air Speed'
height='Landing Height' pitch='Landing Pitch' distance='Landing Distance';
run;


/*########################################################################################
Find empty values for each variables*/
proc means data=faa n nmiss mean median;
run;
/*Above proc showed distance and air speed variables have missing values.*/


/*########################################################################################
Find abnormal values in each variables*/

/*Duration - a normal flight duration is grater than 40 min
Ground Speed - must be between 30 mph and 140 mph
Air Speed - must be between 30 mph and 140 mph
Height - must be greater than 6 m
Distance - Normally less than 6000 ft. */
data faa_abnormal;
set faa;
if duration=. then duration_abnormal='missing'; /*We know this variable has missing values*/
else if duration<40 then duration_abnormal='abnormal';
else duration_abnormal='normal';
label duration_abnormal = 'Duration';
if speed_ground<30 or speed_ground>140 then sp_gr_abnormal='abnormal';
else sp_gr_abnormal='normal';
label sp_gr_abnormal = 'Ground Speed';
if speed_air=. then sp_air_abnormal='missing'; /*We know this variable has missing values*/
else if speed_air<30 or speed_air>140 then sp_air_abnormal='abnormal';
else sp_air_abnormal='normal';
label sp_air_abnormal = 'Air Speed';
if height<6 then height_abnormal='abnormal';
```

```
else height_abnormal='normal';
label height_abnormal = 'Height';
if distance>6000 then distance_abnormal='abnormal';
else distance_abnormal='normal';
label distance_abnormal = 'Distance';
/*Consolidate missing and abnormal value columns*/
if duration_abnormal='missing' or sp_air_abnormal='missing' then missing_val='missing';
else missing_val='no missing values';
if duration_abnormal='abnormal' or sp_gr_abnormal='abnormal' or sp_air_abnormal='abnormal' or
height_abnormal='abnormal' or distance_abnormal='abnormal' then abnormal_val='abnormal';
else abnormal_val='normal';
if missing_val='missing' or abnormal_val='abnormal' then row_clean='unclean';
else row_clean='clean';
run;


/*##############################################################################################
Observe missing values and abnormal values in each variable*/
proc freq data=faa_abnormal;
title 'Missing or abnormal values in each variable';
tables duration_abnormal sp_gr_abnormal sp_air_abnormal height_abnormal distance_abnormal /nocum nopercent nocol;
run;
/*Find impact of missing data or abnormal values*/
proc freq data=faa_abnormal;
title 'Impact of above missing or abnormal values on quality of data';
tables abnormal_val missing_val row_clean /nocum nopercent nocol;
run;


/*##############################################################################################
Removing the unclean rows of the data. Also remove temporary variables created in previous steps*/
data faa_clean;
set faa_abnormal;
if abnormal_val='abnormal' then delete;
drop duration_abnormal sp_gr_abnormal sp_air_abnormal height_abnormal distance_abnormal abnormal_val missing_val
row_clean;
run;


/*##############################################################################################
Observe each variable*/
/*Airline*/
proc freq data=faa_clean;
title 'Number of aircraft from each of the airline companies';
tables aircraft /nocum;
run;

title ;
proc univariate data=faa_clean noprint;
var duration height speed_air speed_ground no_pasg pitch distance;
histogram /kernel normal(noprint);
inset n nmiss mean std min max normal(ksdpval)skewness kurtosis/position=ne;
run;
```

Chapter 2: Data exploration

Goal:

The goal of this step is to explore the data further and find patterns and relations between the various variables and the landing distance (dependent variable). We also look to see if there are any interactions among the independent variables themselves.

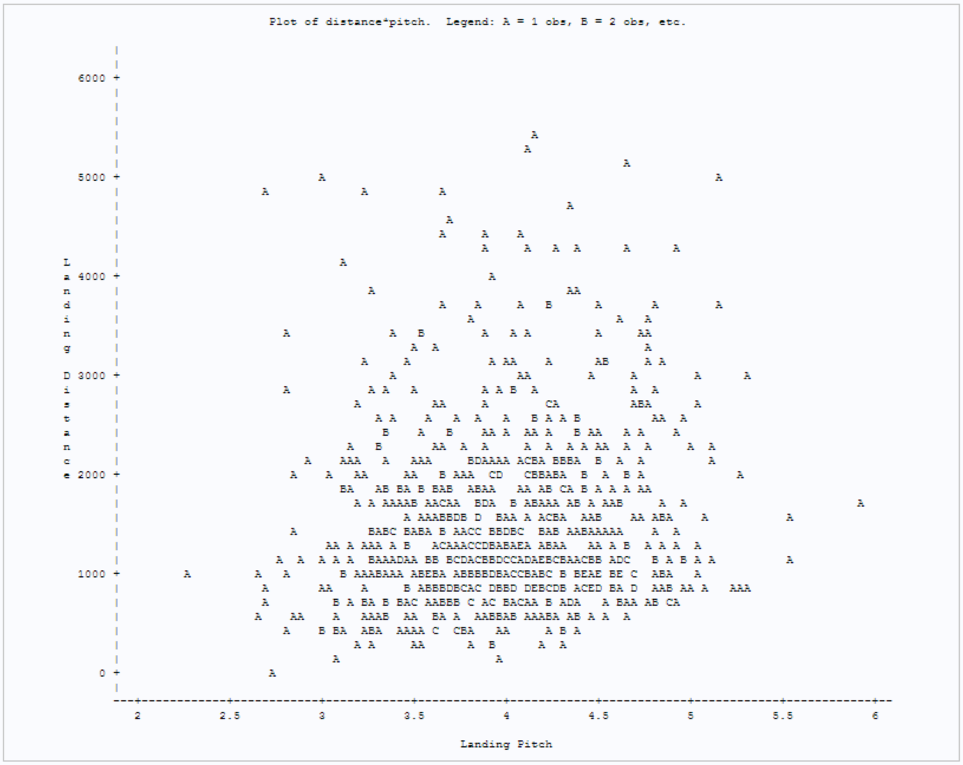
Observations and conclusions:

- Scatter plots of landing distance vs each of the variable are plotted to understand the relation between the variables and the landing distance.
    - It is visibly seen that landing distance has an almost linear relationship with ground and air speed.
    - It is also seen that 'Boeing' flights have a higher average landing distance compared to the 'Airbus' ones.
- Correlation amongst all the variables is checked to see if any variables have an interaction amongst themselves.
    - It is seen that air speed is highly correlated with ground speed with a correlation factor of 0.988.
    - Since 75% of air speed values are missing, and it has such a high correlation with ground speed, this variable can be dropped.

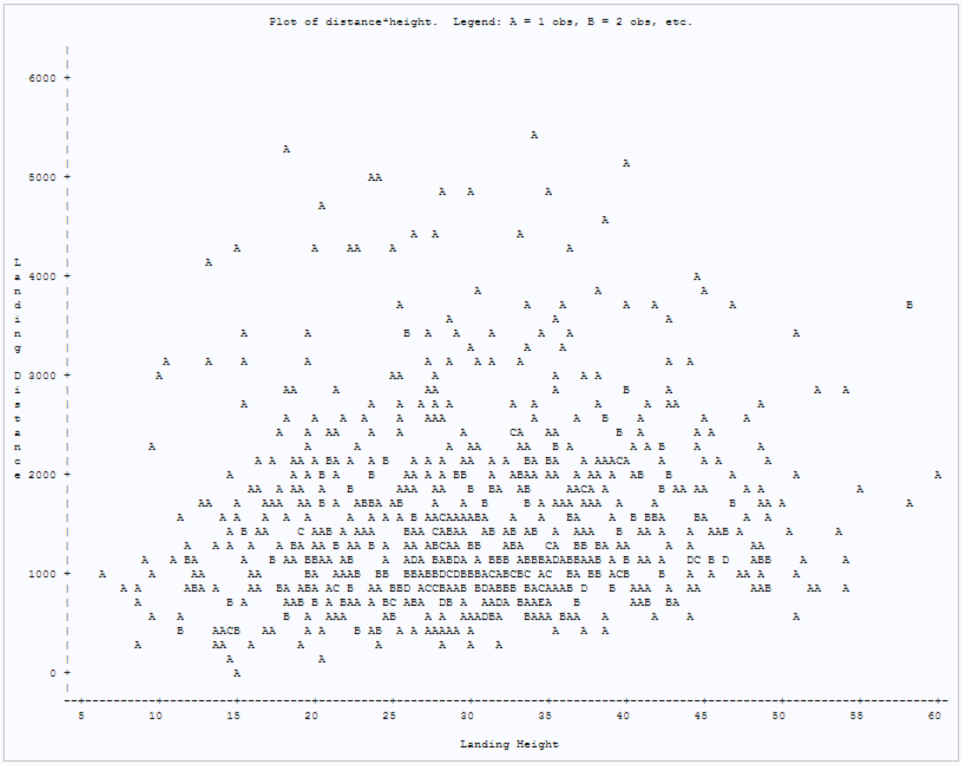
SAS Output:

Scatter plots:

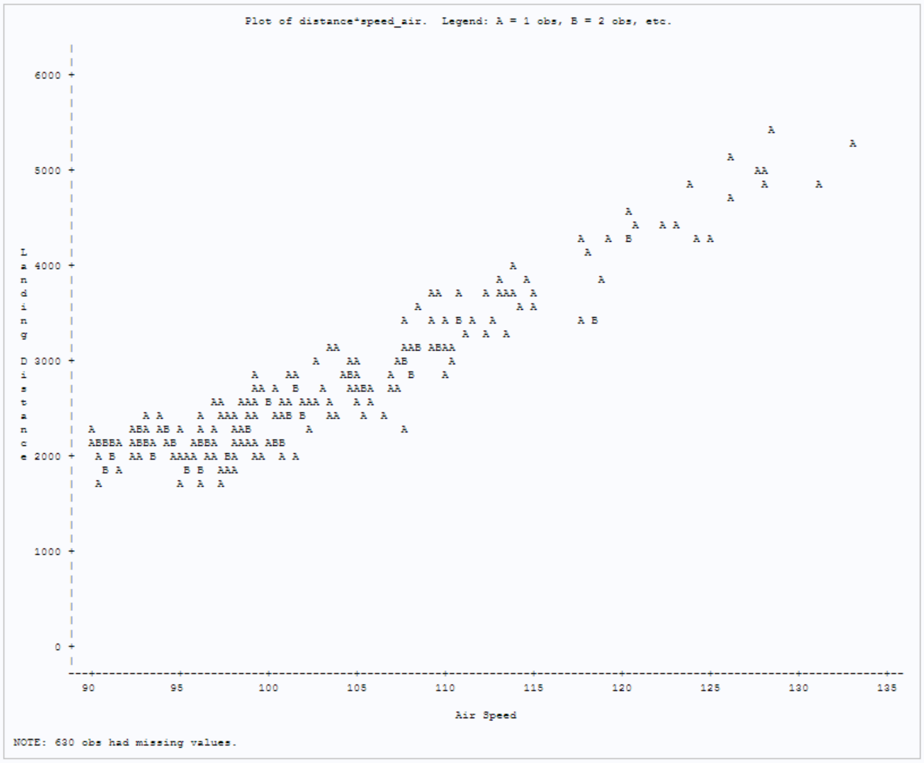
Plot of Landing distance vs Pitch



Plot of Landing distance vs Height

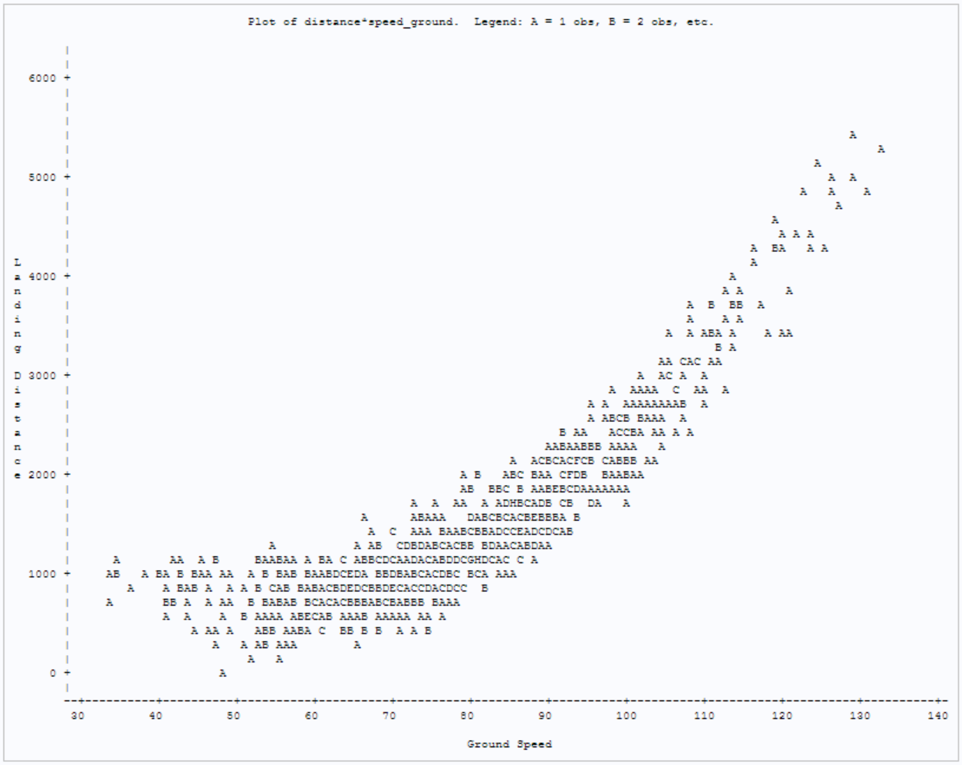


Plot of Landing distance vs Air speed

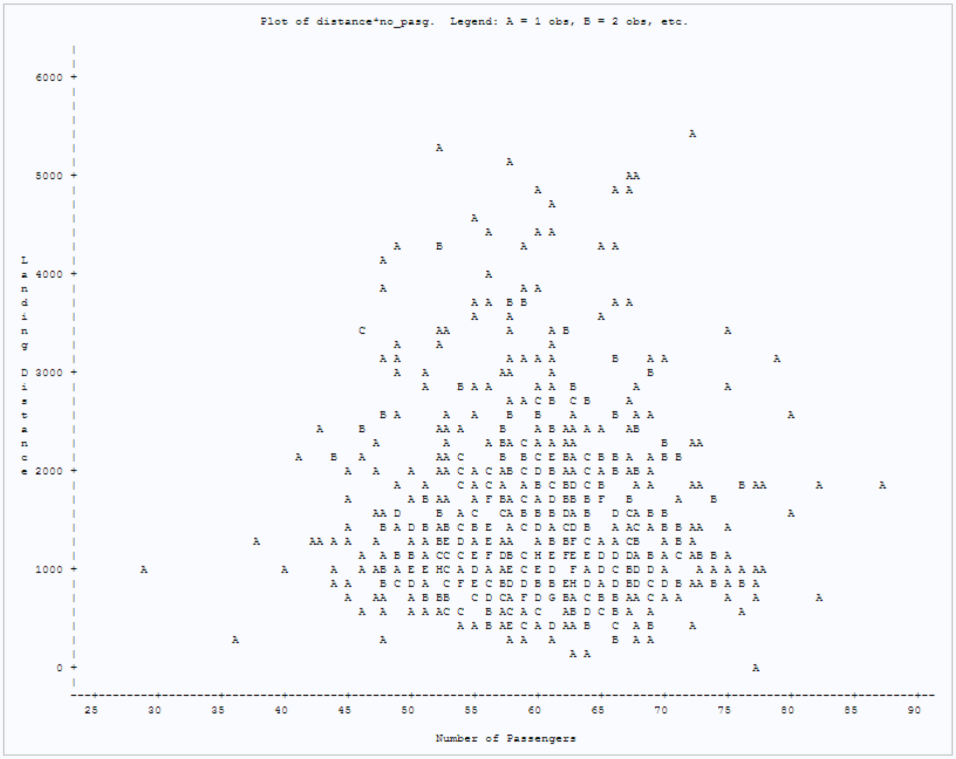


NOTE: 630 obs had missing values.

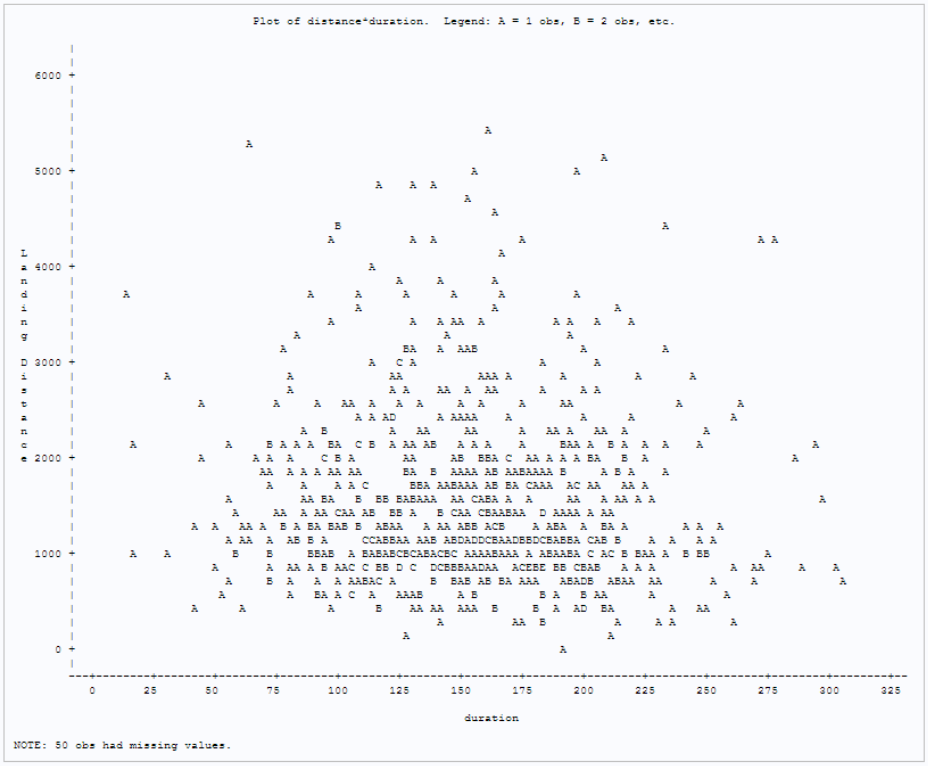
Plot of Landing distance vs Ground speed



Plot of Landing distance vs Number of passengers



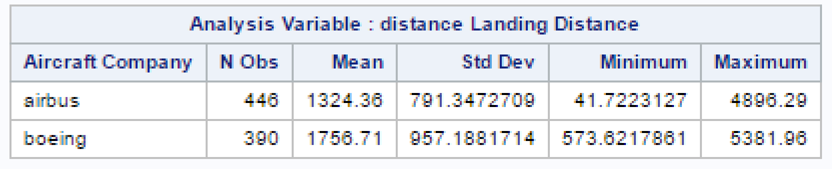
Plot of Landing distance vs Flight duration



```
Plot of distance*duration.  Legend: A = 1 obs, B = 2 obs, etc.
```

```
NOTE: 50 obs had missing values.
```

Landing distance metrics by Aircraft manufacturer

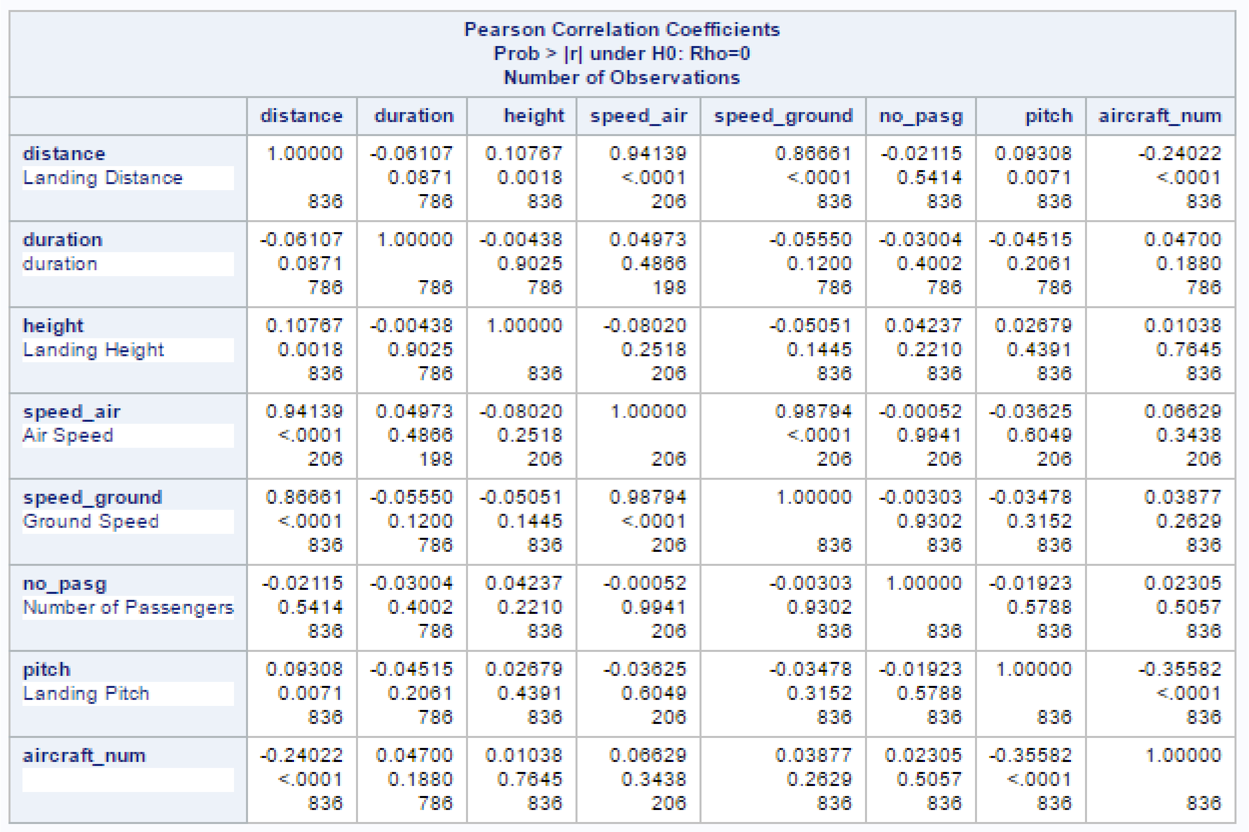


Analysis Variable : distance Landing Distance					
Aircraft Company	N Obs	Mean	Std Dev	Minimum	Maximum
airbus	446	1324.36	791.3472709	41.7223127	4896.29
boeing	390	1756.71	957.1881714	573.6217861	5381.96

Correlation table

(The variable for aircraft manufacturer being a categorical text variable needs to be converted into a numeric variable before it can used for correlation or other calculations. Hence a new dummy variable called ‘aircraft_num’ is created, with ‘Boeing’ = 0, and ‘Airbus’ = 1)

(Correlation table in next page)



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	distance	duration	height	speed_air	speed_ground	no_pasg	pitch	aircraft_num
distance Landing Distance	1.00000 836	-0.06107 0.0871 786	0.10767 0.0018 836	0.94139 <.0001 206	0.86661 <.0001 836	-0.02115 0.5414 836	0.09308 0.0071 836	-0.24022 <.0001 836
duration duration	-0.06107 0.0871 786	1.00000 786	-0.00438 0.9025 786	0.04973 0.4866 198	-0.05550 0.1200 786	-0.03004 0.4002 786	-0.04515 0.2061 786	0.04700 0.1880 786
height Landing Height	0.10767 0.0018 836	-0.00438 0.9025 786	1.00000 836	-0.08020 0.2518 206	-0.05051 0.1445 836	0.04237 0.2210 836	0.02679 0.4391 836	0.01038 0.7645 836
speed_air Air Speed	0.94139 <.0001 206	0.04973 0.4866 198	-0.08020 0.2518 206	1.00000 206	0.98794 <.0001 206	-0.00052 0.9941 206	-0.03625 0.6049 206	0.06629 0.3438 206
speed_ground Ground Speed	0.86661 <.0001 836	-0.05550 0.1200 786	-0.05051 0.1445 836	0.98794 <.0001 206	1.00000 836	-0.00303 0.9302 836	-0.03478 0.3152 836	0.03877 0.2629 836
no_pasg Number of Passengers	-0.02115 0.5414 836	-0.03004 0.4002 786	0.04237 0.2210 836	-0.00052 0.9941 206	-0.00303 0.9302 836	1.00000 836	-0.01923 0.5788 836	0.02305 0.5057 836
pitch Landing Pitch	0.09308 0.0071 836	-0.04515 0.2061 786	0.02679 0.4391 836	-0.03625 0.6049 206	-0.03478 0.3152 836	-0.01923 0.5788 836	1.00000 836	-0.35582 <.0001 836
aircraft_num	-0.24022 <.0001 836	0.04700 0.1880 786	0.01038 0.7645 836	0.06629 0.3438 206	0.03877 0.2629 836	0.02305 0.5057 836	-0.35582 <.0001 836	1.00000 836

SAS CODES:

```
/*#############################################################################################
Create dummy variable for character column - aircraft*/
data faa_clean;
set faa_clean;
if aircraft='boeing' then aircraft_num=0;
else aircraft_num=1;

/*#############################################################################################
Observe plot of each variable with landing distance to understand distribution*/
proc plot data = faa_clean;
plot distance*pitch;
plot distance*height;
plot distance*speed_air;
plot distance*speed_ground;
plot distance*no_pasg;
plot distance*duration;
run;

proc means mean std min max;
class aircraft;
var distance;
run;

/*#############################################################################################
Observe correlation between all the variables*/
proc corr data = faa_clean;
var distance duration height speed_air speed_ground no_pasg pitch aircraft_num;
run;
```

```
/*Above proc shows air speed variable is highly correlated with ground speed.
Since it has too many missing values, it is being dropped. It's impact/significance can be gathered from ground speed*/
data faa_clean;
set faa_clean;
drop speed_air;
run;
```

CHAPTER 3: MODEL BUILDING

GOAL:

The goal of this step is to perform the regression activity which will identify the factors which have an impact on the dependent variable. Once identified, it also provides the values of the coefficients against the factors in a linear equation as below:

$Y = a_1x_1 + a_2x_2 + a_3x_3 + ...$

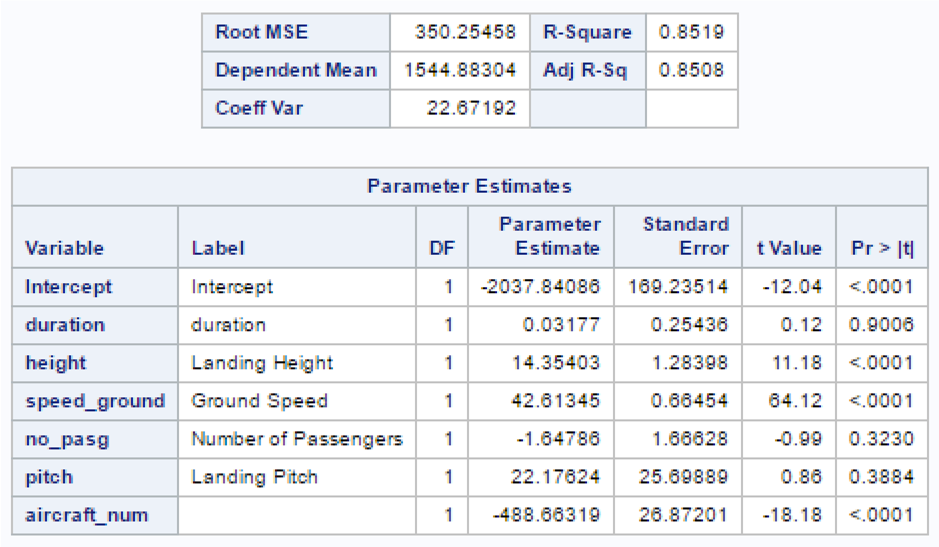
where, Y is the dependent variable, x_i are the independent important factors and a_i are the coefficients, which indicate their impact on Y.

OBSERVATIONS AND CONCLUSIONS:

- The first iteration of the regression procedure shows that the flight duration, number of passengers and landing pitch do not have an impact on the landing distance.
  - The p-value of the t-statistic is higher than 5% and hence these variables do not influence the dependent variable.
  - These variables are independent from each other as can be seen in the correlation table above. So, as opposed to modelling best practices, where one variable is dropped at a time and another iteration of regression is performed, we can drop all three variables at once and re-run the regression
- The second iteration of the regression shows that the remaining variables DO have an impact on the landing distance and are thus to be considered the important factors.
  - The coefficients are –
    - Landing height – 14.22
    - Ground speed – 42.45
    - Aircraft make - -497.01 (0=Boeing, 1=Airbus)
  - The adjusted R-squared value for the equation is 84.96%

SAS OUTPUT:

Regression output – first iteration:

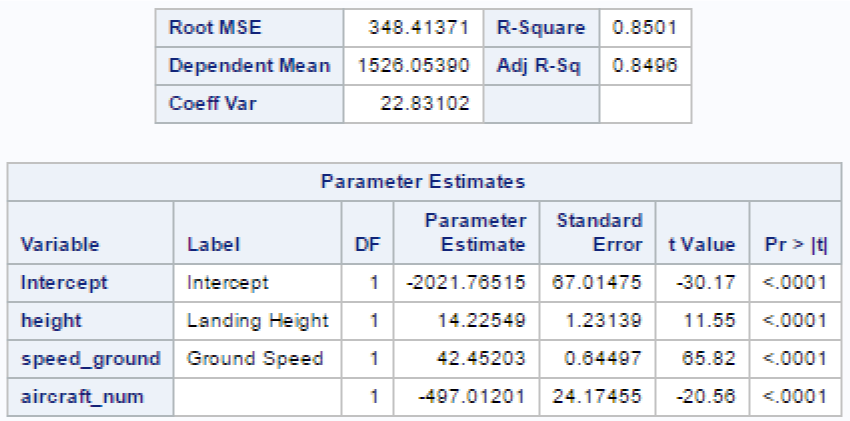


Root MSE	350.25458	R-Square	0.8519
Dependent Mean	1544.88304	Adj R-Sq	0.8508
Coeff Var	22.67192		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2037.84086	169.23514	-12.04	<.0001
duration	duration	1	0.03177	0.25436	0.12	0.9006
height	Landing Height	1	14.35403	1.28398	11.18	<.0001
speed_ground	Ground Speed	1	42.61345	0.66454	64.12	<.0001
no_pasg	Number of Passengers	1	-1.64786	1.66628	-0.99	0.3230
pitch	Landing Pitch	1	22.17624	25.69889	0.86	0.3884
aircraft_num		1	-488.66319	26.87201	-18.18	<.0001

Regression output – second iteration:



Root MSE	348.41371	R-Square	0.8501
Dependent Mean	1526.05390	Adj R-Sq	0.8496
Coeff Var	22.83102		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2021.76515	67.01475	-30.17	<.0001
height	Landing Height	1	14.22549	1.23139	11.55	<.0001
speed_ground	Ground Speed	1	42.45203	0.64497	65.82	<.0001
aircraft_num		1	-497.01201	24.17455	-20.56	<.0001

SAS CODE:

```
/*##############################################################################################################
Begin the model building using proc reg*/
proc reg data=faa_clean;
model distance= duration height speed_ground no_pasg pitch aircraft_num;
run;

/*Above proc showed that the three variables duration, no_pasg and pitch are not significant.
These variables all have very low corelation among themselves as seen in proc corr output
Hence removing all of them at once from the regression procedure*/
proc reg data=faa_clean;
model distance= height speed_ground aircraft_num;
output out=faa_res r=residuals;
run;
```

CHAPTER 4: MODEL VERIFICATION

GOAL:

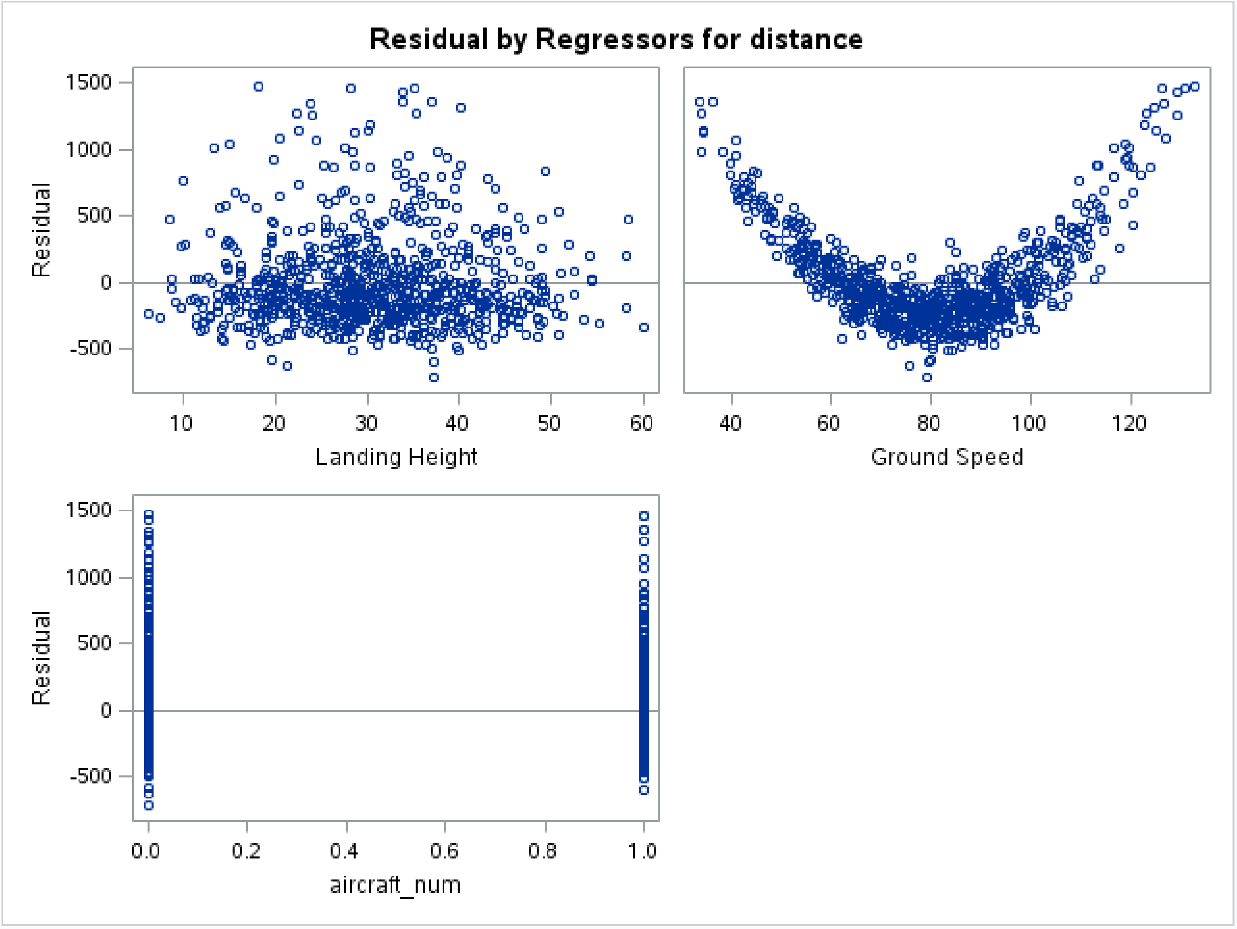
This step involves checking residuals for normality, and 0 mean; thus indirectly verifying the assumptions made to use linear regression.

OBSERVATIONS AND CONCLUSIONS:

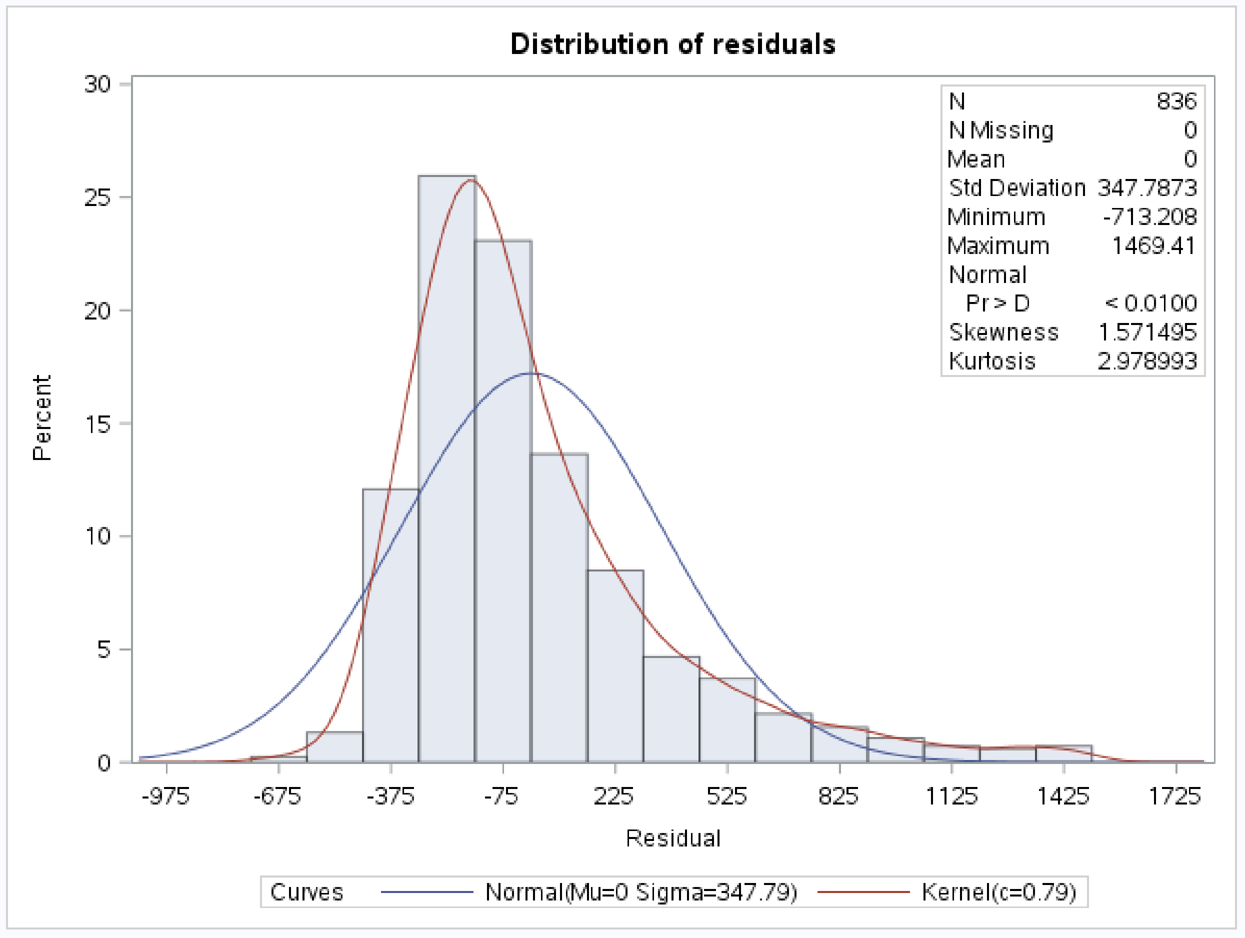
- Checking the plots of the residuals against the three variables-
  - There is no particular linear relationship between the residuals and the variables
- Normality test for residuals as seen on the histogram with p-value for K-S test shows that the residuals are normally distributed
- The t-test with null hypothesis that mean is zero is tested and we cannot reject the null hypothesis.

SAS OUTPUT:

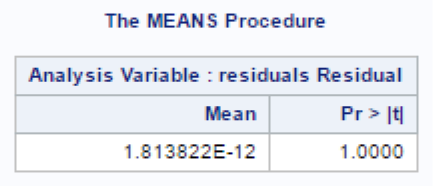
Plots of residuals vs the variables:



Normal tests



Test for mean=0



The MEANS Procedure

Analysis Variable : residuals Residual	
Mean	Pr > |t|
1.813822E-12	1.0000

SAS CODE:

```
/*###################################################################################################
Checking assumptions on residuals*/
/*Check for normality of residuals*/
proc univariate data=faa_res noprint;
var residuals;
histogram /kernel normal(noprint);
inset n nmiss mean std min max normal(ksdpval)skewness kurtosis/position=ne;
run;
```

```
/*Check for zero mean of residuals*/
proc means data=faa_res mean prt;
var residuals;
run;
```