Exploratory Data Analysis – Flight Landing Data

Submitted by: Akash Jain(jain2ar)

1. Analysis Summary:

The analysis is being done on the landing dataset to study what factors and how they impact the landing distance of the commercial flight. The analysis is a quintessential case of regression where distance will be the response variable.

In all there were 850 observations and 8 columns. Some of the columns like speed_air and duration were rejected from the analysis because their correlation with the distance was either too much or there was no correlation as all.

Upon the analysis of the remaining variables, it was found that all the variables are normally distributed except for the distance which happens to follow a lognormal distribution. There are also some outliers in the final dataset which needs to be addressed but we need additional information about those observations for their analysis. The final number of observations in the dataset after cleaning the data is 836.

2. Background about data

The analysis is being done on the datasets which contains the landing information of the flights. The whole data was divided into two files containing information related to the landing of the flights for Boeing and Airbus. The dataset contains the following variables:

a. Aircraft
b. Duration(in minutes)
c. No_pasg
d. Speed_ground(in miles per hour)
e. Speed_air(in miles per hour)
f. Height(in meters)
g. Pitch(in degrees)
h. Distance(in feet)

3. Data exploration and preparation

We start by importing the datasets into SAS by using the following code:

```
proc import datafile = "/home/jain2ar0/Midterm and Project/FAA1.xls" out=dataset1 dbms=xls replace;

proc import datafile="/home/jain2ar0/Midterm and Project/FAA2.xls" out=dataset2 dbms=xls;

run;
```

3.1 Structure of the datasets

After importing the datasets into sas and copying them into local datasets. We go on to check whether the structure of the datasets is correct or not.

SAS Code:

```
proc contents data=dataset1 varnum;
```

```
run;
proc contents data=dataset2 varnum;
run;
proc print data=dataset1(obs=20);
run;
proc print data= dataset2(obs=20);
run;
```

Output:

Dataset1:

The CONTENTS Procedure

Data Set Name	WORK.DATASET1	Observations	800
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	01/23/2017 21:53:31	Observation Length	72
Last Modified	01/23/2017 21:53:31	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Variables in Creation Order

#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	12	$12.	$12.	aircraft
2	duration	Num	8	BEST12.		duration
3	no_pasg	Num	8	BEST12.		no_pasg
4	speed_ground	Num	8	BEST12.		speed_ground
5	speed_air	Num	8	BEST12.		speed_air
6	height	Num	8	BEST12.		height
7	pitch	Num	8	BEST12.		pitch
8	distance	Num	8	BEST12.		distance

Dataset2:

The CONTENTS Procedure

Data Set Name	WORK.DATASET2	Observations	200
Member Type	DATA	Variables	7
Engine	V9	Indexes	0
Created	01/23/2017 21:53:31	Observation Length	64
Last Modified	01/23/2017 21:53:31	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Variables in Creation Order

#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	12	$12.	$12.	aircraft
2	no_pasg	Num	8	BEST12.		no_pasg
3	speed_ground	Num	8	BEST12.		speed_ground
4	speed_air	Num	8	BEST12.		speed_air
5	height	Num	8	BEST12.		height
6	pitch	Num	8	BEST12.		pitch
7	distance	Num	8	BEST12.		distance

Conclusion:

a. dataset1: The variables are as expected. There are 800 records and all have correct data types.
b. dataset2: There are only 7 variables instead of 8. Duration is missing. In all there are 200 observations
c. There is no unique variable or primary key in the datasets all the variables have relevant data except for missing values
d. Also in dataset2 there are 50 blank rows and hence we must delete them

3.2 Exploration

In section 3.1, we found out that there are 50 blank rows in dataset2, hence we deleted those missing rows using the following code:

SAS code:

```
options missing='';
data dataset2_v2;
set dataset2;
if missing(cats(of _all_)) then delete;
run;
```

To start exploring the data, we will start with merging the datasets together into a new dataset called dataset3. We will use match merge to remove duplicate observations.

SAS code:

```
proc sort data=dataset1;
```

```
by aircraft no_pasg speed_ground speed_air height pitch distance;
run;
proc sort data=dataset2_v2;
by aircraft no_pasg speed_ground speed_air height pitch distance;
run;
data dataset3;
merge dataset1 dataset2_v2;
by aircraft no_pasg speed_ground speed_air height pitch distance;
run;
proc contents data= dataset3 varnum;run;
```

SAS Output:

The CONTENTS Procedure

Data Set Name	WORK.DATASET3	**Observations**	850
Member Type	DATA	**Variables**	8
Engine	V9	**Indexes**	0
Created	01/31/2017 02:25:32	**Observation Length**	72
Last Modified	01/31/2017 02:25:32	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Conclusion:

a. There are 850 observations as expected and there are 8 columns in the dataset.

3.2.1 Check for missing values

To check for missing values in the dataset, we use the following SAS code:

SAS code:

```
proc means data=dataset3 nmiss;
run;
/*checking for missing values for aircraft*/
proc format;
value $missing_aircraft ' '='Missing Value' other='Correct Value';
run;
 proc freq data=dataset3;
format _CHAR_ $missing_aircraft.;
tables _CHAR_ / missing missprint nocum nopercent;
```

run;

SAS Output:

The MEANS Procedure

Variable	Label	N Miss
duration	duration	50
no_pasg	no_pasg	0
speed_ground	speed_ground	0
speed_air	speed_air	642
height	height	0
pitch	pitch	0
distance	distance	0

The FREQ Procedure

aircraft	
aircraft	Frequency
Correct Value	850

Conclusion:

a. There are missing values in speed_air and distance. Rest all the variables have complete data.
b. There are 5.8% of values are missing for speed_air and 75.53% values are missing for distance.

3.2.2 Check for correlation

To check for correlation in the dataset, we use the following SAS code:

SAS Code:

```
proc corr data=dataset3;run;
```

SAS Output:

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 800	-0.03391 0.3382 800	-0.06063 0.0866 800	0.04911 0.4898 200	-0.00678 0.8481 800	-0.03896 0.2710 800	-0.06209 0.0792 800
no_pasg no_pasg	-0.03391 0.3382 800	1.00000 850	-0.00517 0.8803 850	-0.00589 0.9327 208	0.01098 0.7492 850	-0.01490 0.6643 850	-0.03033 0.3771 850
speed_ground speed_ground	-0.06063 0.0866 800	-0.00517 0.8803 850	1.00000 850	0.98929 <.0001 208	-0.01607 0.6399 850	-0.03062 0.3727 850	0.86196 <.0001 850
speed_air speed_air	0.04911 0.4898 200	-0.00589 0.9327 208	0.98929 <.0001 208	1.00000 208	-0.06588 0.3444 208	0.00639 0.9270 208	0.94728 <.0001 208
height height	-0.00678 0.8481 800	0.01098 0.7492 850	-0.01607 0.6399 850	-0.06588 0.3444 208	1.00000 850	0.01284 0.7085 850	0.13624 <.0001 850
pitch pitch	-0.03896 0.2710 800	-0.01490 0.6643 850	-0.03062 0.3727 850	0.00639 0.9270 208	0.01284 0.7085 850	1.00000 850	0.10269 0.0027 850
distance distance	-0.06209 0.0792 800	-0.03033 0.3771 850	0.86196 <.0001 850	0.94728 <.0001 208	0.13624 <.0001 850	0.10269 0.0027 850	1.00000 850

Conclusion:

a. Speed_air and speed_ground is highly correlated. The correlation is positive and is also quite significant as suggested by the p-value.
b. Speed_air and speed_ground is highly correlated with distance. The correlation is positive and is also quite significant as suggested by the p-value.
c. The correlation between height and distance is significant but it is negligible.
d. Rest all the variable are not significantly correlated with each other.
e. Since speed_air and speed_ground is highly correlated, we can delete one of the fields from the dataset. Since 76% of the values from speed _air is missing, we will remove that column. This will also help in resolving multicollinearity.
f. Duration is not significantly correlated with either distance or any other variable Also, logically speaking, duration of the flight should not affect the landing distance. Hence, we will also remove it from the dataset.

3.2.3 Analysis of variables

We will analyse each variable present in the dataset to look for their distribution, identify outliers and unusual observations using basic statistics measure and plots.

3.2.3.1 Analysis of Pitch

We use the following code to perform analysis on pitch:

SAS Code:

```
proc univariate data=dataset3 plot;
```

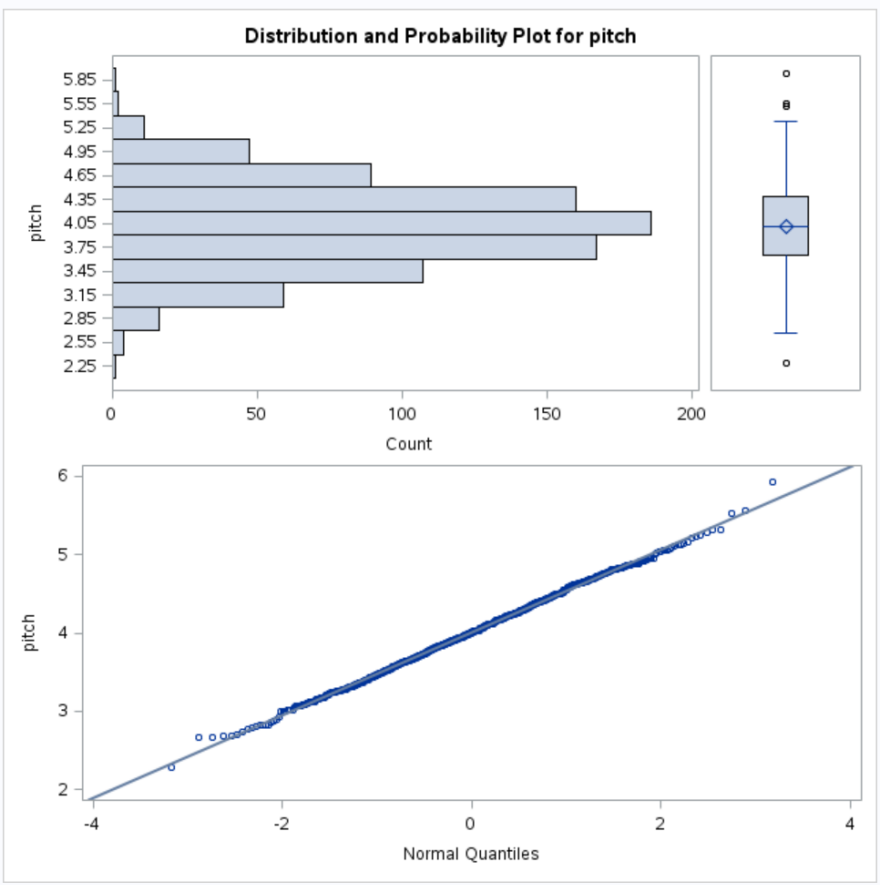
var pitch;
run;

SAS Output:

The UNIVARIATE Procedure
Variable: pitch (pitch)

Moments			
N	850	Sum Weights	850
Mean	4.0093577	Sum Observations	3407.95404
Std Deviation	0.52882984	Variance	0.279661
Skewness	0.00615409	Kurtosis	-0.1062347
Uncorrected SS	13901.139	Corrected SS	237.432186
Coeff Variation	13.1898892	Std Error Mean	0.01813871

Basic Statistical Measures			
Location		Variability	
Mean	4.009358	Std Deviation	0.52883
Median	4.008288	Variance	0.27966
Mode		Range	3.64230
		Interquartile Range	0.73712



Conclusions:

a. Pitch seems to be almost normally distributed as the mean and median are almost equal. This also suggest that there are some potential outliers

b. The distribution is slightly positively skewed but the measure is negligible
c. From the box plot, it is clear that we have some outliers in pitch and the distribution is normal as seen from the histogram and normal quantile plot.

3.2.3.2 Analysis of Height

We use the following code to perform analysis on height:

SAS Code:

```
proc univariate data=dataset3 plot;
var height;
run;
```

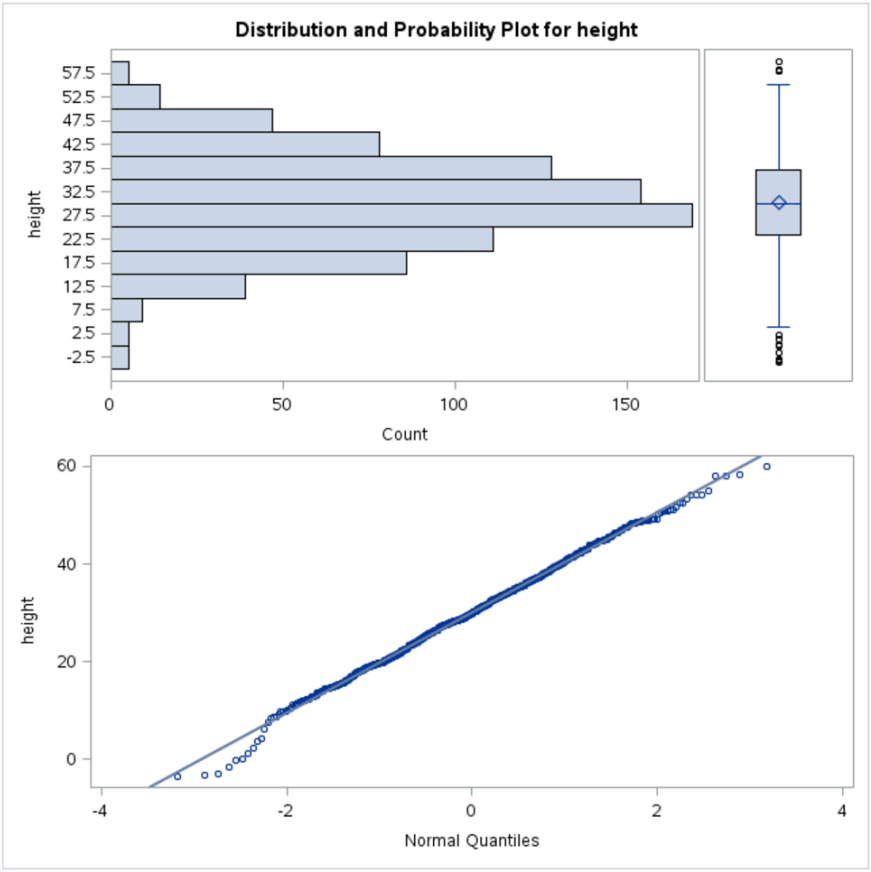
SAS Output:

The UNIVARIATE Procedure
Variable: height (height)

Moments			
N	850	Sum Weights	850
Mean	30.1442223	Sum Observations	25622.589
Std Deviation	10.2877268	Variance	105.837324
Skewness	-0.0956784	Kurtosis	0.10262214
Uncorrected SS	862228.907	Corrected SS	89855.8878
Coeff Variation	34.1283538	Std Error Mean	0.35286612

Basic Statistical Measures			
Location		Variability	
Mean	30.14422	Std Deviation	10.28773
Median	30.09313	Variance	105.83732
Mode	9.68831	Range	63.49222
		Interquartile Range	13.69231

Quantiles (Definition 5)	
Level	Quantile
100% Max	59.94596
99%	53.43862
95%	47.38932
90%	43.91020
75% Q3	36.99458
50% Median	30.09313
25% Q1	23.30227
10%	17.21471
5%	13.80759
1%	3.78892
0% Min	-3.54625



Conclusion:

a. Height seems to be almost normally distributed as the mean and median are almost equal
b. The distribution is slightly negatively skewed and slightly peaked but the measure is negligible
c. There are also 1% potential outliers in height which can be seen from the quantiles as the minimum height for safe landing should be 6 meters.

3.2.3.3 Analysis of Distance

We use the following code to perform analysis on distance:

SAS Code:

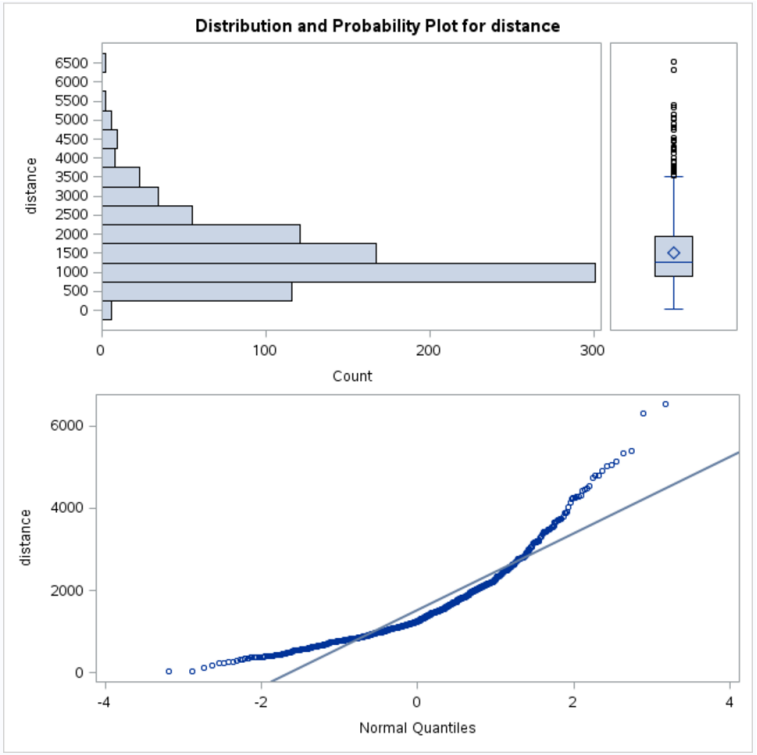
```
proc univariate data=dataset3 plot;
var distance;
run;
```

SAS Output:

The UNIVARIATE Procedure
Variable: distance (distance)

Moments			
N	850	Sum Weights	850
Mean	1526.02309	Sum Observations	1297119.63
Std Deviation	928.560082	Variance	862223.825
Skewness	1.63493883	Kurtosis	3.5837272
Uncorrected SS	2711462540	Corrected SS	732028027
Coeff Variation	60.8483636	Std Error Mean	31.849348

Basic Statistical Measures			
Location		Variability	
Mean	1526.023	Std Deviation	928.56008
Median	1258.092	Variance	862224
Mode		Range	6499
		Interquartile Range	1054



Conclusion:

a. There is quite a difference between mean and median and this suggests that there are outliers that too towards the upper limit of the data
b. The distribution is also positively skewed and highly peaked. The observation for outliers conforms with the box plot
c. Distance is not normally distributed and to apply regression to this dataset, log transformation must be applied on distance as it follows a lognormal distribution.

3.2.3.4 Analysis of Speed_ground

We use the following code to perform analysis on Speed_ground:

SAS Code:

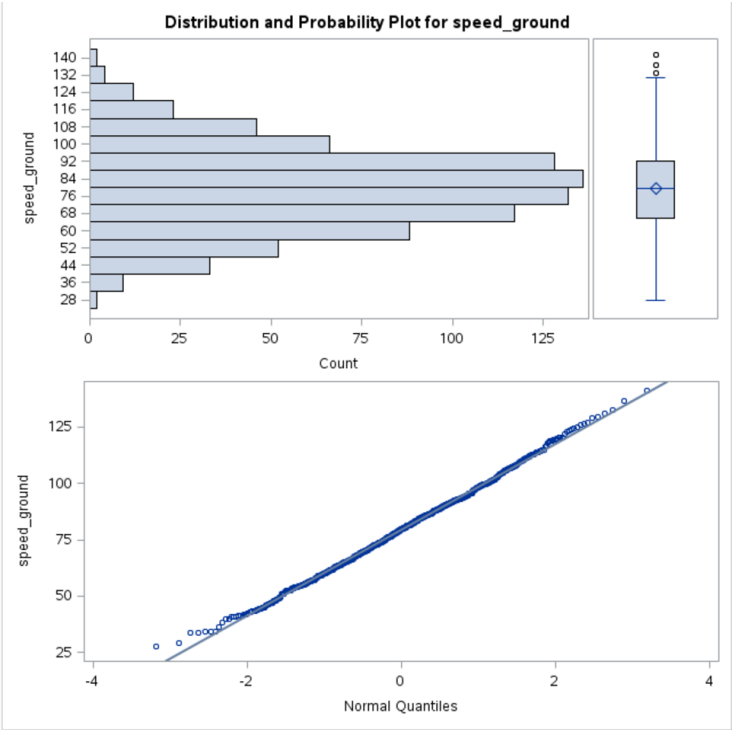
```
proc univariate data=dataset3 plot;
var speed_ground;
run;
```

SAS Output:

The UNIVARIATE Procedure
Variable: speed_ground (speed_ground)

Moments			
N	850	Sum Weights	850
Mean	79.4523229	Sum Observations	67534.4744
Std Deviation	19.0594903	Variance	363.264171
Skewness	0.11782542	Kurtosis	-0.1030934
Uncorrected SS	5674182.15	Corrected SS	308411.281
Coeff Variation	23.9885879	Std Error Mean	0.65373512

Basic Statistical Measures			
Location		Variability	
Mean	79.45232	Std Deviation	19.05949
Median	79.64280	Variance	363.26417
Mode		Range	113.48292
		Interquartile Range	26.21296



Conclusion:

a. There is slight difference between mean and median and this suggests that there are outliers that too towards the upper limit of the data
b. The distribution is also slightly positively skewed and peaked and almost follows the normal distribution
c. The observation for outliers conforms with the box plot. And the normal quantile plot confirms the normal distribution.

3.2.3.5 Analysis of No_pasg

We use the following code to perform analysis on No_pasg:

SAS Code:

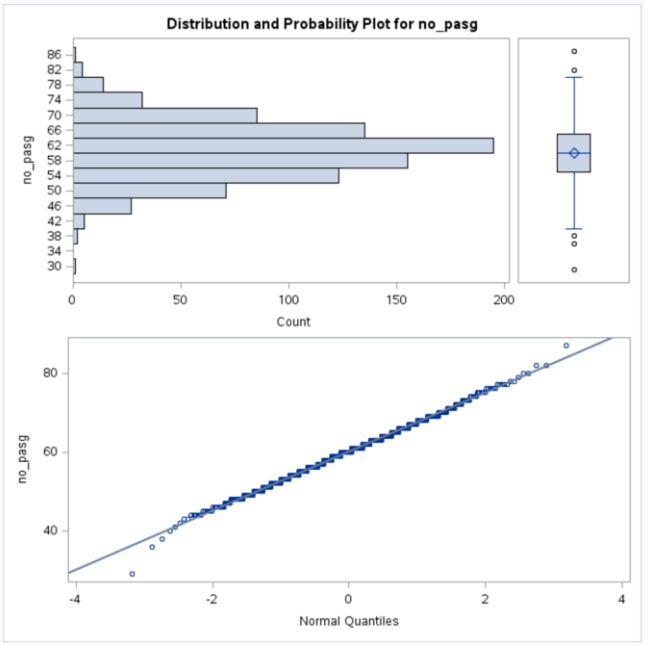
```
proc univariate data=dataset3 plot;
var no_pasg;
run;
```

SAS Output:

The UNIVARIATE Procedure
Variable: no_pasg (no_pasg)

Moments			
N	850	Sum Weights	850
Mean	60.1035294	Sum Observations	51088
Std Deviation	7.49313698	Variance	56.1471018
Skewness	-0.0215023	Kurtosis	0.27570748
Uncorrected SS	3118238	Corrected SS	47668.8894
Coeff Variation	12.4670499	Std Error Mean	0.25701248

Basic Statistical Measures			
Location		Variability	
Mean	60.10353	Std Deviation	7.49314
Median	60.00000	Variance	56.14710
Mode	61.00000	Range	58.00000
		Interquartile Range	10.00000



Conclusion:

a. There is slight difference between mean and median and this suggests that there are outliers in the data
b. The distribution is also slightly negatively skewed and peaked and almost follows the normal distribution
c. The observation for outliers conform with the box plot. The normal quantile plot confirms the normal distribution with somewhat step like distribution.

3.2.3.6 Analysis of Duration

We use the following code to perform analysis on Duration:

SAS Code:

```
proc univariate data=dataset3 plot;
var duration;
run;
```

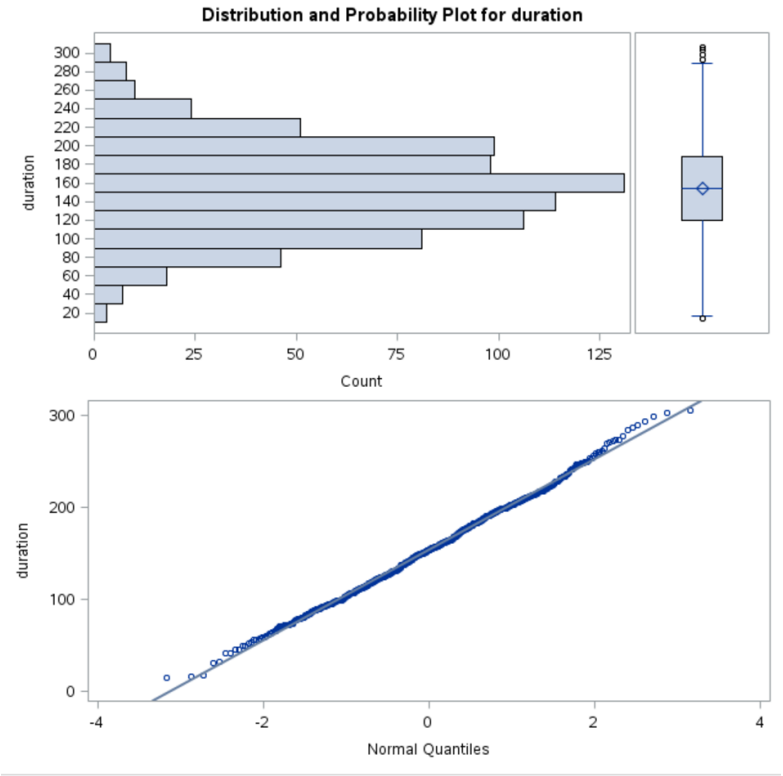
SAS Output:

The UNIVARIATE Procedure
Variable: duration (duration)

Moments			
N	800	Sum Weights	800
Mean	154.006538	Sum Observations	123205.231
Std Deviation	49.2592338	Variance	2426.47211
Skewness	0.12147943	Kurtosis	-0.0551851
Uncorrected SS	20913162.3	Corrected SS	1938751.22
Coeff Variation	31.9851574	Std Error Mean	1.74157691

Basic Statistical Measures			
Location		Variability	
Mean	154.0065	Std Deviation	49.25923
Median	153.9481	Variance	2426
Mode		Range	290.85750
		Interquartile Range	69.44330

Quantiles (Definition 5)	
Level	Quantile
100% Max	305.6217
99%	275.6969
95%	234.1229
90%	214.4738
75% Q3	188.9179
50% Median	153.9481
25% Q1	119.4746
10%	92.0313
5%	74.4080
1%	45.5691
0% Min	14.7642



Conclusion:

a) There is slight difference between mean and median and this suggests that there are outliers in the data
b) The distribution is also slightly positively skewed and peaked and almost follows the normal distribution
c) Per the condition given in the dataset, the duration should be atleast 40 mins and hence after looking at quantiles, we can say that atleast 1% of the observations are outliers
d) The normal quantile plot confirms the normal distribution.

3.2.3.7 Analysis of Speed_air

We use the following code to perform analysis on Speed_air:

SAS Code:

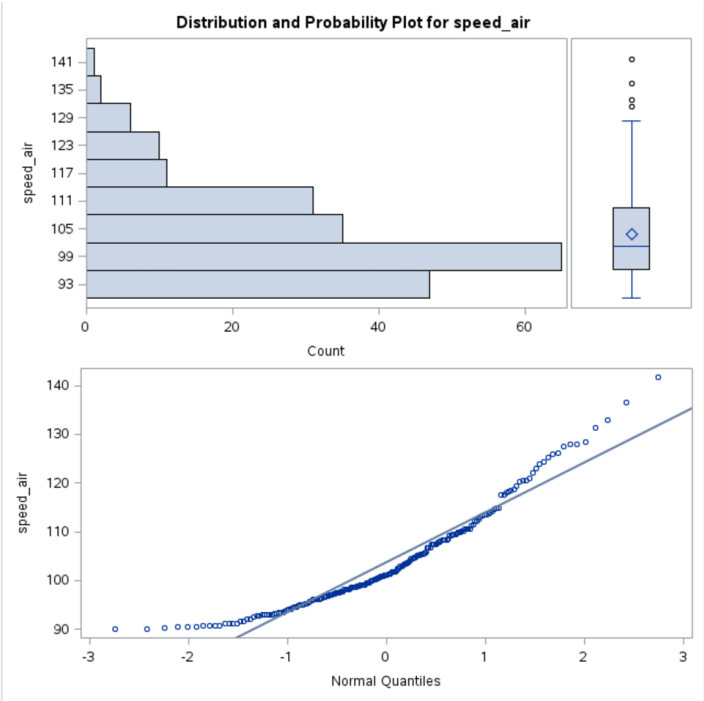
```
proc univariate data=dataset3 plot;
var speed_air;
run;
```

SAS Output:

The UNIVARIATE Procedure
Variable: speed_air (speed_air)

Moments			
N	208	Sum Weights	208
Mean	103.797724	Sum Observations	21589.9265
Std Deviation	10.259037	Variance	105.24784
Skewness	1.0564046	Kurtosis	0.90174387
Uncorrected SS	2262771.53	Corrected SS	21786.3028
Coeff Variation	9.88368204	Std Error Mean	0.71133623

Basic Statistical Measures			
Location		Variability	
Mean	103.7977	Std Deviation	10.25904
Median	101.1473	Variance	105.24784
Mode		Range	51.72208
		Interquartile Range	13.19078



Conclusion:

a) There is substantial difference between mean and median and this suggests that there are extreme outliers in the data
b) The distribution is highly skewed which can be confirmed by the value of skewness and the histogram.

3.3 Data Cleaning

As per the conclusion from section 3.2.2, we remove the columns, speed_air and duration from the dataset. To remove these columns, we use the following code:

SAS Code:

```
data dataset3_v2;
set dataset3;
drop speed_air duration;
run;
```

SAS Output: dataset3_v2 created.

The remaining columns in the dataset are, Aircraft, No_pasg, Speed_ground, Height, Pitch and distance. According to the dataset description, there are certain constraint for these variables like:

a. Speed_ground less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal
b. The landing aircraft is required to be at least 6 meters high at the threshold of the runway
c. The length of the airport runway is typically less than 6000 feet.

We must make sure that out dataset conforms to these conditions and hence we will use the following code to clean the garbage values in the data.

SAS Code:

```
data dataset3_v3;
set dataset3_v2;
if speed_ground<30 or speed_ground>140 then delete;
if height < 6 then delete;
if distance > 6000 then delete;
run;
```

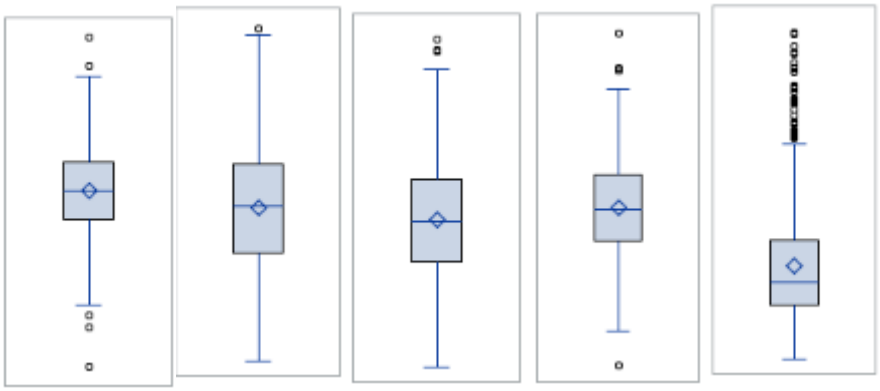
SAS Output: dataset3_v3 created.

Checking the variables again using the following SAS code:

```
proc univariate data=dataset3_v3 plot;run;
```

SAS Output:

Box Plots to analyse outliers:



No_Pasg Speed_Ground Height Pitch Distance

Conclusion:

a. No of passengers has some outliers.
b. Speed ground has 1 outliers
c. Height has 3 outliers
d. Pitch has some outliers
e. Distance has some outliers.

Questions:

1) There are certain observations in Distance where the values are 41 feet or less than 200 feet. Are these some special cases like accident? Additional information is needed for these cases
2) There are certain observations in no_of passengers where the values are 30-38 and some values between 82-86. These observations can't be considered as outliers as there is nothing to be bothered about. Are these any special cases?