Regression Analysis using SAS

Flight Landing – Risk of overrun

Statistical Computing
BANA 5143/6043 Project

Presented by:
Name: Rashmi Prathigadapa
UCID:M12853369

About the Project

Introduction:

Background: Flight landing.

Motivation: To reduce the risk of landing overrun.

Goal: To study what factors and how they would impact the landing distance of a commercial flight.

Data: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). See two Excel files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights).

Variable dictionary:

Aircraft: The make of an aircraft (Boeing or Airbus).

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

No_pasg: The number of passengers in a flight.

Speed_ground (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.

Speed_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.

Height (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway.

Pitch (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway.

Distance (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

Abstract

Landing overrun is problem for most flight landing operations. There are multiple factors which affect the landing distance. In this report we are trying to identify key factors affecting the landing distance of commercial flights namely Airbus and Boeing.

To determine the factors and quantify the impact of factors on landing distance, we created a linear regression model with the landing distance as dependent variable. Landing distance is largely dependent on ground speed of aircraft, pitch, aircraft type and height of the aircraft when it passes through the threshold of runway.

Using these covariates, we have built a linear regression equation, which can be used to predict the landing distance of an aircraft, given the values of factors influencing it. Below is the equation obtained:

Distance = -1049 + 454.45*(aircraft_name) + 0.27*(speed_ground1) + 14*(height) +21*(pitch)

(aircraft name=0 for airbus and 1 for boeing)

This equation specifies that:

1. If the aircraft_name value belongs to 0, it implies the model built for Airbus and 1 implies the model for Boeing.

2. For ‘Boeing’ aircraft type the predicted landing distance would be 454 units greater than the landing distance for ‘Airbus’ aircraft type.

3. For every one-unit increase in pitch there will be 21-unit increase in the predicted landing distance

4. For every one-unit increase in square of ground speed there will be 0.27-unit increase in the predicted landing distance

5. For every one-unit increase in height there will be 14-unit increase in the predicted landing distance

Table of contents

Chapter 1. Data Cleaning and Data Exploration

The datasets FAA1 and FAA2 were given. In order to use this data to build a linear model, we need to clean the data and prepare it for the modelling.

SAS Code:

```
/**************DATAPREPARATION****************/

/*IMPORTING THE DATASET FAA1 USING PROC IMPORT*/

PROC IMPORT
DATAFILE="/folders/myshortcuts/Stat_computing/Week3/FAA1.xls"
     DBMS=XLS
     OUT=FAA1;
     SHEET="FAA1";
RUN;
```

Output:

A dataset with 800 observations and 8 variables is created.

```
/*****IMPORTING THE DATASET FAA2 USING PROC IMPORT*****/

PROC IMPORT
DATAFILE="/folders/myshortcuts/Stat_computing/Week3/FAA2.xls"
     DBMS=XLS
     OUT=FAA2;
     SHEET="FAA2";
RUN;
```

Output:

A dataset with 150 observations and 7 variables is created.

```
/*****************************************************************/
 1. COMBINING DATASETS FROM DIFFERENT SOURCES USING SET STATEMENT
/*****************************************************************/

DATA FAA REST;
SET FAA1 FAA2;
IF AIRCRAFT="" THEN OUTPUT REST;
ELSE OUTPUT FAA;
RUN;
```

Output:

A dataset with 950 observations and 8 variables is created. Sample data is given below:

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	98.4790912	53	107.9156800465	109.32837648	27.418924252	4.0435145715	3369.8363838
2	boeing	125.73329732	69	101.6555886321	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	112.0170008	61	71.05196088308	.	18.589385734	4.4340431286	1144.922426
4	boeing	196.82569105	56	85.8133276789	.	30.744597235	3.8842361245	1664.2181584
5	boeing	90.095381357	70	59.88852818319	.	32.397688062	4.0260964152	1050.2644976
6	boeing	137.59581722	55	75.0143437441	.	41.21496259	4.203853398	1627.0681991
7	boeing	73.023794916	54	54.42980289964	.	24.03532163	3.8376457299	805.30399317
8	boeing	52.903187872	57	57.10166173716	.	19.388837508	4.6436717769	573.62178606
9	boeing	155.51861605	61	85.44362425143	.	35.375389749	4.2287278648	1698.9927548
10	boeing	176.86203205	56	61.79671051413	.	36.748816124	4.1843990127	1137.7457579
11	boeing	158.4618984	61	53.77812674124	.	46.355832902	5.5563991716	1075.3717411
12	boeing	180.61655753	54	141.2186353517	141.72493569	23.575935009	5.2168022511	6533.0476506
13	boeing	72.289633216	54	93.39176243452	92.869561214	32.223489271	3.8182761471	2128.708285
14	boeing	187.59954737	58	94.0364129417	96.196460585	33.661226156	4.6361847249	2304.857574
15	boeing	154.36870049	63	63.54061355285	.	26.402991875	3.8566584986	1089.9729531
16	boeing	165.54194536	69	48.7746732732	.	31.228664837	3.9020460339	943.06840443
17	boeing	153.54633587	61	83.55649327068	.	29.897473262	3.519783726	1793.5628232
18	boeing	107.11331938	78	86.80796202517	.	25.477015381	4.4142187986	1910.8768699
19	boeing	233.80249791	69	104.8084344839	103.86845794	43.882731896	3.2450978263	3213.985265
20	boeing	163.90650312	55	119.3804634966	120.44470797	38.558536007	3.7014493887	4524.2788621
21	boeing	97.477623266	63	73.53397633557	.	29.152465311	4.0140064257	1332.0387485
22	boeing	118.63054039	55	79.99481504199	.	29.366866101	4.4071812572	1515.9652753
23	boeing	126.54028789	70	94.78123028226	91.142068839	39.476298784	3.5949361476	2182.2207374
24	boeing	179.91591838	66	63.67116531366	.	19.574699606	4.2867337712	873.4408921
25	boeing	112.90009528	53	98.18041086249	99.135830727	28.152991316	3.9874712191	2586.6650864
26	boeing	56.64048966	66	72.95365823853	.	36.154157217	4.3878559157	1205.1280251
27	boeing	86.828911312	62	91.71453579219	92.874851912	28.773729478	3.3058880775	2313.3356963
28	boeing	157.35773231	57	72.32713077838	.	26.223285332	4.2231807894	1105.3658522
29	boeing	186.68141397	49	66.41723046402	.	44.692695788	4.1135438115	1176.0276765
30	boeing	140.23631155	65	118.7420047119	119.40214631	19.856192215	4.6462659602	4217.1294518
31	boeing	130.46356358	52	116.7134343365	117.65649967	36.195527446	3.8943524297	4240.0941825

Explanation:

This SET statement combines the data in two datasets vertically. The first dataset FAA1 contains 800 observations and the dataset FAA2 contains 150 observations. Also, the missing values are sent into dataset named REST (if any). Now the combined dataset FAA contains 950 observations.

SAS Code:

```
/* REMOVE DUPLICATES*/

PROC SORT DATA=FAA OUT = FAA_NO_DUP NODUPKEY DUPOUT=DUPS;
BY AIRCRAFT NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DISTANCE; RUN;
```

Output:

Out of 950 lines, there are 100 lines which were present in both FAA1 and FAA2 datasets, so they were considered as duplicate lines and were excluded into a dataset named “DUPS”. Rest of the 850 were considered for our analysis.

```
NOTE: There were 950 observations read from the data set WORK.FAA.
NOTE: 100 observations with duplicate key values were deleted.
NOTE: The data set WORK.FAA_NO_DUP has 850 observations and 8
variables.
NOTE: The data set WORK.DUPS has 100 observations and 8 variables.
NOTE: PROCEDURE SORT used (Total process time):
       real time           0.00 seconds
       cpu time            0.00 seconds
```

Explanation:

Using the NODUPKEY, duplicate records are excluded from our analysis and sent into a different dataset named DUPS.
In our data, the variable DURATION is present only in FAA1 but not FAA2. Therefore, that field must not be considered while removing duplicates. Out of 950 observations, 850 of them are unique while 100 are duplicate records, so they are excluded. Now we are left with 850 observations.

```
/*******************************************************/
 2. Performing the completeness check of each variable
     EXAMINE IF MISSING VALUES ARE PRESENT within fields
/*******************************************************/
```

SAS CODE:

```
PROC MEANS DATA= FAA_NO_DUP N NMISS MEAN MIN MAX MEDIAN STD STDDEV;
VAR _NUMERIC_;
OUTPUT OUT=FAA_STAT;
RUN;
```

Output:

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Minimum	Maximum	Median	Std Dev
duration	duration	800	50	154.0065385	14.7642071	305.6217107	153.9480975	49.2592338
no_pasg	no_pasg	850	0	60.1035294	29.0000000	87.0000000	60.0000000	7.4931370
speed_ground	speed_ground	850	0	79.4523229	27.7357153	141.2186354	79.6428041	19.0594903
speed_air	speed_air	208	642	103.7977237	90.0028586	141.7249357	101.1473493	10.2590370
height	height	850	0	30.1442223	-3.5462524	59.9459639	30.0931324	10.2877268
pitch	pitch	850	0	4.0093577	2.2844801	5.9267842	4.0082875	0.5288298
distance	distance	850	0	1526.02	34.0807833	6533.05	1258.09	928.5600816
of_all_		0	850	.	.	.	.	.

There are 50 missing values in the ‘duration’ column coming from the dataset FAA2. 642 missing values in ‘speed_air’. Also, to analyze the data the mean, median, standard deviation, maximum and minimum are calculated.

Alternatively, we can also use proc univariate.

```
/*STATISTICS USING UNIVARIATE TO IDENTIFY THE MISSING VALUES*/

PROC UNIVARIATE DATA=FAA_NO_DUP;
RUN;
```

Output:

Upon using the univariate, we get to know the number of observations, mean, standard deviation, skewness, variance etc.

Also, information about the missing values for each variable is given.

For variable 'duration':

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

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	50	5.88	100.00

For variable 'Speed Air':

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

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	642	75.53	100.00

Explanation:

Therefore, the variables Speed_air and Duration have missing values in them, which can either be imputed or ignored, based on the model requirement.

```
/*****************************************************************/
3. Performing the validity check of each variable –
     examine if abnormal values are present;
/*****************************************************************/

/*******ABNORMAL VALUES BASED ON THE VALIDITY RULES GIVEN******/


/*IF SPEED_GROUND VALUE IS LESS THAN 30MPH OR GREATER THAN 140MPH*
/*IF SPEED_AIR VALUE IS LESS THAN 30MPH OR GREATER THAN 140MPH*
/*IF DURATION OF FLIGHT IS LESS THAN 40 MINUTES*
/*IF HEIGHT OF FLIGHT ON RUNWAY IS LESS THAN 6 FEET*
/*IF RUNWAY DISTANCE IS LESS THAN 6000*/
```

SAS Code:

```
DATA FAA_CLEAN;
SET FAA_NO_DUP;
IF DURATION LT 40 AND DURATION NE . THEN ABNORMALITY="DURATION LESS
THAN 40 MINUTES";
```

```
ELSE IF (SPEED_GROUND LT 30 OR SPEED_GROUND GT 140) AND (SPEED_GROUND
NE .) THEN ABNORMALITY="SPEED WRT GROUND IS NOT BETWEEN 30-140 MPH";
ELSE IF (SPEED_AIR LT 30 OR SPEED_AIR GT 140) AND (SPEED_AIR NE .)
THEN ABNORMALITY="SPEED WRT GROUND IS NOT BETWEEN 30-140 MPH";
ELSE IF HEIGHT LT 6 THEN ABNORMALITY="HEIGHT IS LESS THAN 6 FT";
ELSE IF DISTANCE GT 6000 THEN ABNORMALITY="LANDING DISTANCE LESS THAN
6000";
ELSE ABNORMALITY="NONE";
RUN;
```

Output:

A new column with type of abnormality (if any) is created.

```
/* Exclude the abnormal values*/

DATA FAA_ABNORMAL FAA_NORMAL;
SET FAA_CLEAN;
IF ABNORMALITY="NONE" THEN OUTPUT FAA_NORMAL;
ELSE OUTPUT FAA_ABNORMAL;
RUN;

proc print data=faa_abnormal; run;
```

Output:

Dataset named “FAA_ABNORMAL”. These values are excluded for reasons in the ABNORMALITY column.

Obs	aircraft	duration	no_pasg	speed_ground	height	pitch	distance	ABNORMALITY
1	airbus	16.893454896	54	94.51105222271	37.476967053	4.1733221259	2162.92737	DURATION LESS THAN 40 MINUTES
2	airbus	150.94674427	58	66.42111946786	-2.915335901	3.1225583646	34.080783293	HEIGHT IS LESS THAN 6 FT
3	airbus	31.7016661	61	76.35417643285	30.991021813	2.8173796019	948.47376723	DURATION LESS THAN 40 MINUTES
4	airbus	163.52364053	62	72.02802425244	0.086105484	3.6220566648	537.91958189	HEIGHT IS LESS THAN 6 FT
5	airbus	157.91497689	68	56.49798666138	-0.067758596	4.6928768405	380.36298195	HEIGHT IS LESS THAN 6 FT
6	airbus	103.09084673	73	92.99494238128	-3.332387973	4.8305592948	1567.6657219	HEIGHT IS LESS THAN 6 FT
7	boeing	141.93411511	46	27.73571530329	24.400127629	4.3682093233	1323.7157777	SPEED WRT GROUND IS NOT BETWE
8	boeing	31.391008253	51	98.2198006656	52.473140903	4.1623371208	2808.3151244	DURATION LESS THAN 40 MINUTES
9	boeing	180.61655753	54	141.2186353517	23.575935009	5.2168022511	6533.0476506	SPEED WRT GROUND IS NOT BETWE
10	boeing	14.764207145	59	108.2916902859	46.930873666	4.8096217396	3645.6110025	DURATION LESS THAN 40 MINUTES
11	boeing	212.94303494	61	29.22765638171	23.349901124	4.3961881217	1076.855217	SPEED WRT GROUND IS NOT BETWE
12	boeing	283.76336844	62	58.88931238095	4.2644634439	4.7721930401	425.85856098	HEIGHT IS LESS THAN 6 FT
13	boeing	17.375513046	63	63.57042960985	28.406673108	3.9378640453	1032.4646189	DURATION LESS THAN 40 MINUTES
14	boeing	175.08462089	64	52.4931391022	-3.546252405	4.2132855404	581.38099947	HEIGHT IS LESS THAN 6 FT
15	boeing	119.92455279	64	136.6591583152	44.286109179	4.1694037368	6309.9459762	LANDING DISTANCE LESS THAN 60
16	boeing	119.64402906	68	70.17846387335	2.2051944554	3.7397746803	816.20664104	HEIGHT IS LESS THAN 6 FT
17	boeing	146.04337112	69	71.78730588315	-1.528129182	4.1994604645	738.65436932	HEIGHT IS LESS THAN 6 FT
18	boeing	124.37864547	72	60.36704372521	3.7889195211	3.7060888319	641.59956822	HEIGHT IS LESS THAN 6 FT
19	boeing	133.45985625	73	57.0452994941	1.2538552556	4.7153842391	371.27726086	HEIGHT IS LESS THAN 6 FT

Explanation:

In the problem, we are given certain constraints for the independent variables and if the values are failing to satisfy those constraints, they are excluded as abnormal values from our analysis. After applying those constraints, we are now left with 831 rows for our analysis.

```
/*5. SUMMARIZING THE DISTRIBUTION OF EACH VARIABLE (WHAT TABLES AND
FIGURES WILL YOU PRESENT?) */
```

- For each variable, box plot and a histogram were drawn to understand the distribution of each variable.
- From box plot, we see whether we have outliers in data.
- From histogram, we see whether each dependent variable follows a normal distribution.

SAS Code:

```
PROC UNIVARIATE DATA=FAA_NORMAL;
HISTOGRAM/NORMAL;
QQPLOT;
RUN;
```

Output: 1. For the variable 'Duration':

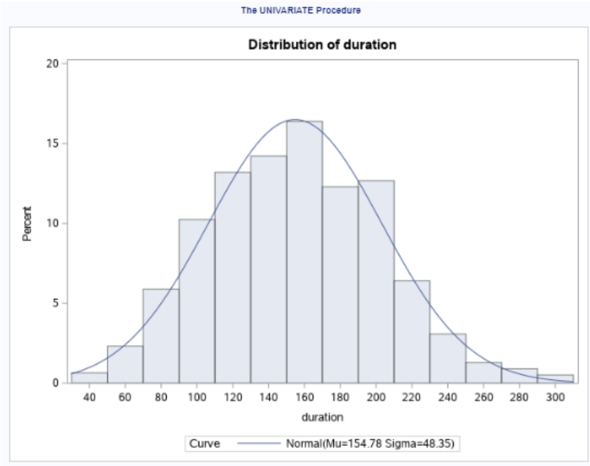
The UNIVARIATE Procedure
Variable: duration (duration)

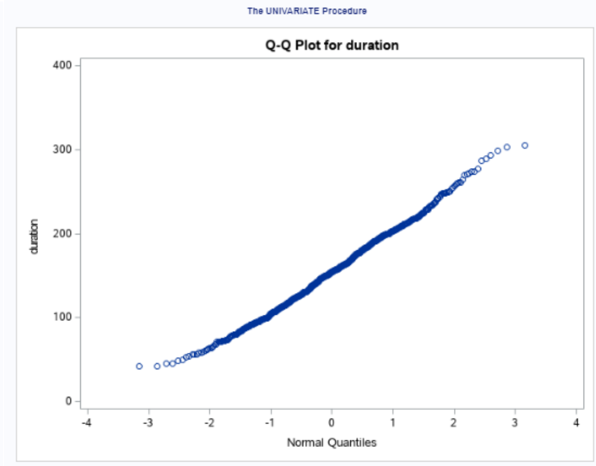
Moments			
N	781	Sum Weights	781
Mean	154.775719	Sum Observations	120879.837
Std Deviation	48.3499237	Variance	2337.71512
Skewness	0.18986566	Kurtosis	-0.1958773
Uncorrected SS	20532681.4	Corrected SS	1823417.79
Coeff Variation	31.2387007	Std Error Mean	1.73009629

Basic Statistical Measures			
Location		Variability	
Mean	154.7757	Std Deviation	48.34992
Median	154.2846	Variance	2338
Mode	.	Range	263.67234
		Interquartile Range	70.03148

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	89.46075	Pr > |t|	<.0001
Sign	M	390.5	Pr >= |M|	<.0001
Signed Rank	S	152685.5	Pr >= |S|	<.0001

This variable follows a normal distribution with little skew.





2. For variable No_pasg:

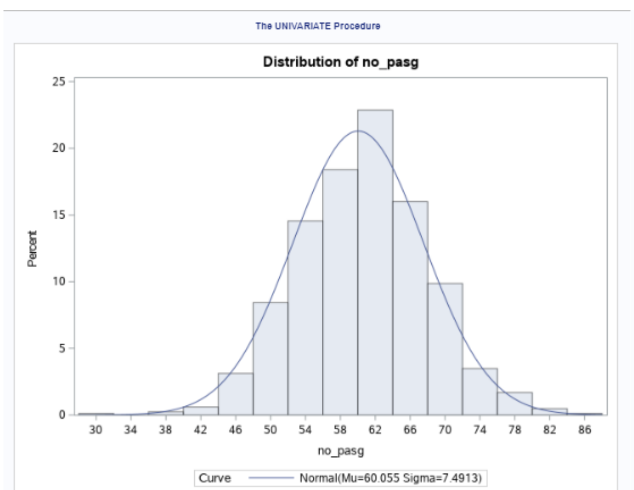
The UNIVARIATE Procedure
Variable: no_pasg (no_pasg)

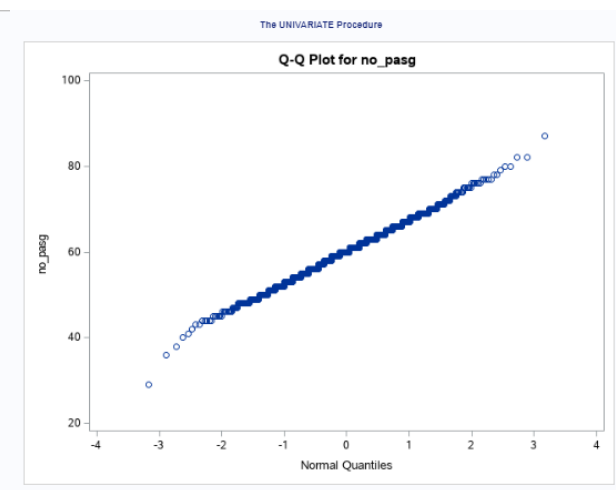
Moments			
N	831	Sum Weights	831
Mean	60.055355	Sum Observations	49906
Std Deviation	7.49131655	Variance	56.1198237
Skewness	-0.0135746	Kurtosis	0.30027454
Uncorrected SS	3043702	Corrected SS	46579.4537
Coeff Variation	12.4740193	Std Error Mean	0.25987089

Basic Statistical Measures			
Location		Variability	
Mean	60.05535	Std Deviation	7.49132
Median	60.00000	Variance	56.11982
Mode	61.00000	Range	58.00000
		Interquartile Range	10.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	231.0969	Pr > |t|	<.0001
Sign	M	415.5	Pr >= |M|	<.0001
Signed Rank	S	172848	Pr >= |S|	<.0001

No of passengers (No_pasg) follows normal distribution.





3. For variable Speed_Ground:

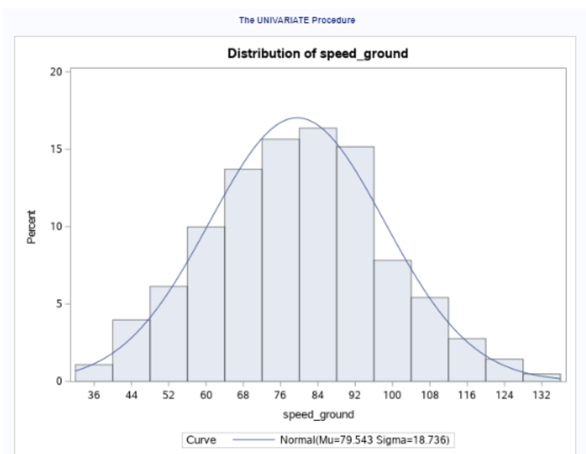
The UNIVARIATE Procedure
Variable: speed_ground (speed_ground)

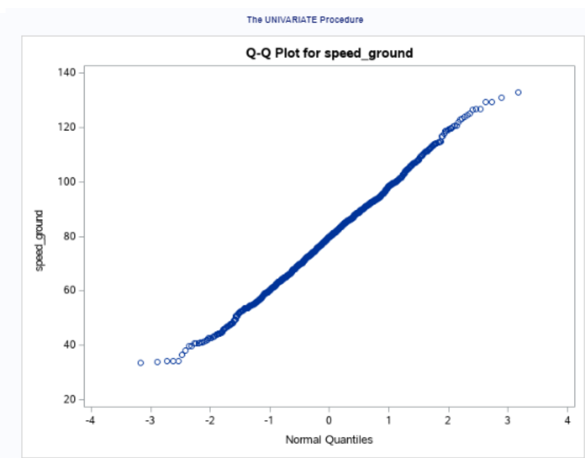
Moments			
N	831	Sum Weights	831
Mean	79.5426997	Sum Observations	66099.9835
Std Deviation	18.7356754	Variance	351.025533
Skewness	0.08890294	Kurtosis	-0.2324866
Uncorrected SS	5549122.33	Corrected SS	291351.193
Coeff Variation	23.5542363	Std Error Mean	0.64993338

Basic Statistical Measures			
Location		Variability	
Mean	79.54270	Std Deviation	18.73568
Median	79.79396	Variance	351.02553
Mode	.	Range	99.21057
		Interquartile Range	25.75708

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	122.3859	Pr > |t|	<.0001
Sign	M	415.5	Pr >= |M|	<.0001
Signed Rank	S	172848	Pr >= |S|	<.0001

The variable Speed_ground follows normal distribution.





4. For variable Speed_air:

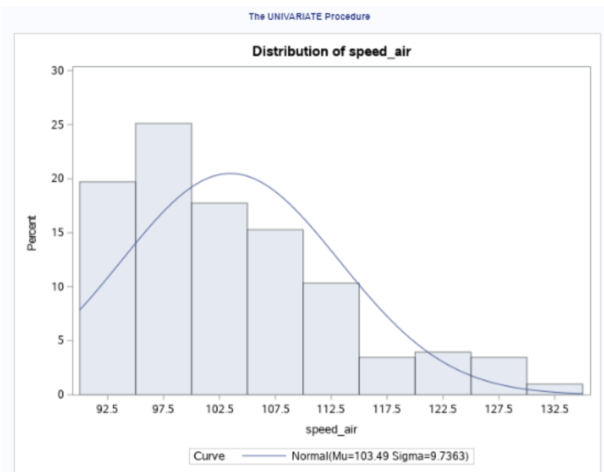
The UNIVARIATE Procedure
Variable: speed_air (speed_air)

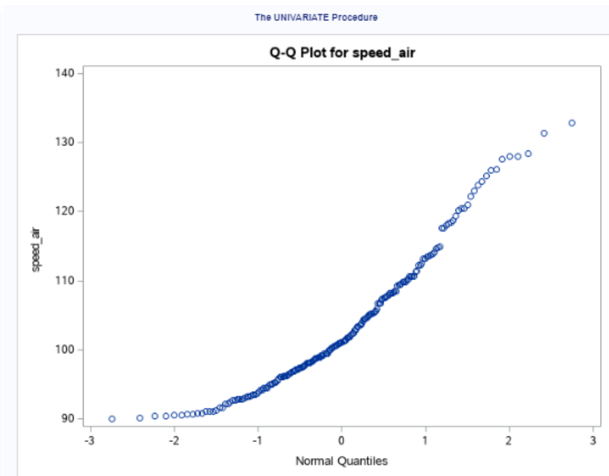
Moments			
N	203	Sum Weights	203
Mean	103.485035	Sum Observations	21007.4621
Std Deviation	9.73627738	Variance	94.7950972
Skewness	0.88272686	Kurtosis	0.23173679
Uncorrected SS	2193106.57	Corrected SS	19148.6096
Coeff Variation	9.40839162	Std Error Mean	0.68335271

Basic Statistical Measures			
Location		Variability	
Mean	103.4850	Std Deviation	9.73628
Median	101.1189	Variance	94.79510
Mode	.	Range	42.90861
		Interquartile Range	13.18584

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	151.4372	Pr > |t|	<.0001
Sign	M	101.5	Pr >= |M|	<.0001
Signed Rank	S	10353	Pr >= |S|	<.0001

The air speed is not all normal.





5. For variable height:

The UNIVARIATE Procedure
Variable: height (height)

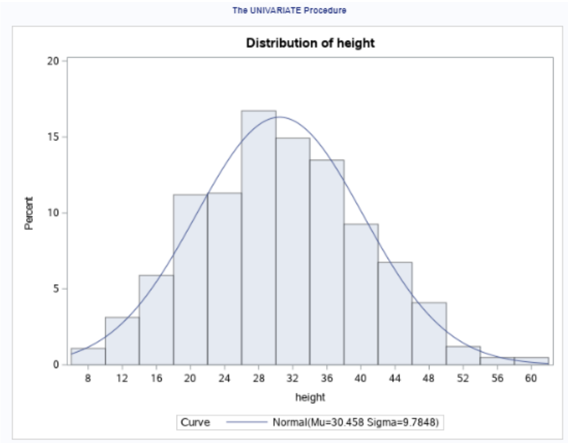
Moments			
N	831	Sum Weights	831
Mean	30.4578695	Sum Observations	25310.4896
Std Deviation	9.78481143	Variance	95.7425347
Skewness	0.12714447	Kurtosis	-0.3338733
Uncorrected SS	850369.892	Corrected SS	79466.3038
Coeff Variation	32.1257251	Std Error Mean	0.33943135

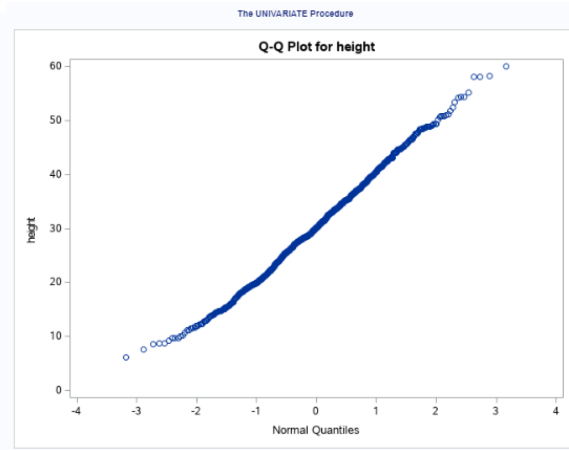
Basic Statistical Measures			
Location		Variability	
Mean	30.45787	Std Deviation	9.78481
Median	30.16708	Variance	95.74253
Mode	9.68831	Range	53.71845
		Interquartile Range	13.48443

Note: The mode displayed is the smallest of 49 modes with a count of 2.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	89.73205	Pr > |t|	<.0001
Sign	M	415.5	Pr >= |M|	<.0001
Signed Rank	S	172848	Pr >= |S|	<.0001

This variable height follows normal distribution with slightly right skewed.





6. For variable Pitch:

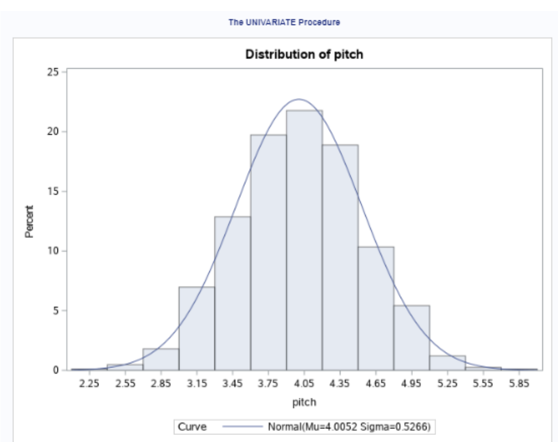
The UNIVARIATE Procedure
Variable: pitch (pitch)

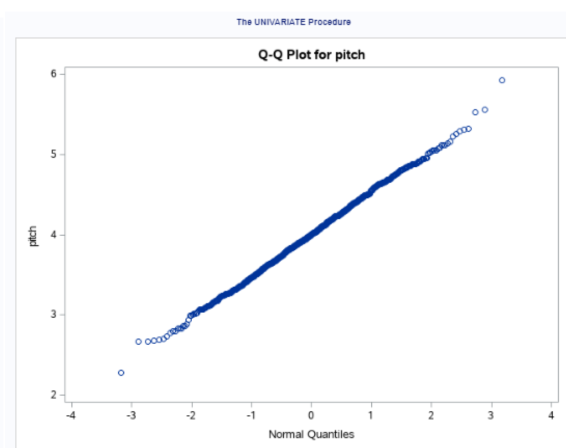
Moments			
N	831	Sum Weights	831
Mean	4.00516086	Sum Observations	3328.28868
Std Deviation	0.52656905	Variance	0.27727496
Skewness	0.01730511	Kurtosis	-0.0907921
Uncorrected SS	13560.4698	Corrected SS	230.138218
Coeff Variation	13.1472634	Std Error Mean	0.01826648

Basic Statistical Measures			
Location		Variability	
Mean	4.005161	Std Deviation	0.52657
Median	4.001038	Variance	0.27727
Mode	.	Range	3.64230
		Interquartile Range	0.73067

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	219.2629	Pr > |t|	<.0001
Sign	M	415.5	Pr >= |M|	<.0001
Signed Rank	S	172848	Pr >= |S|	<.0001

The variable pitch follows normal distribution.





7. For variable Distance:

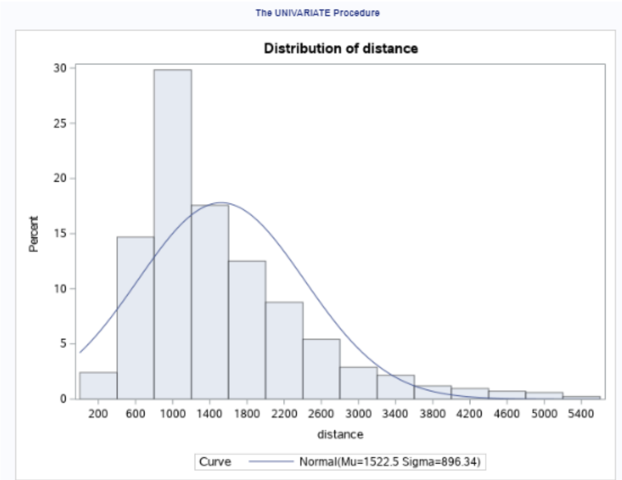
The UNIVARIATE Procedure
Variable: distance (distance)

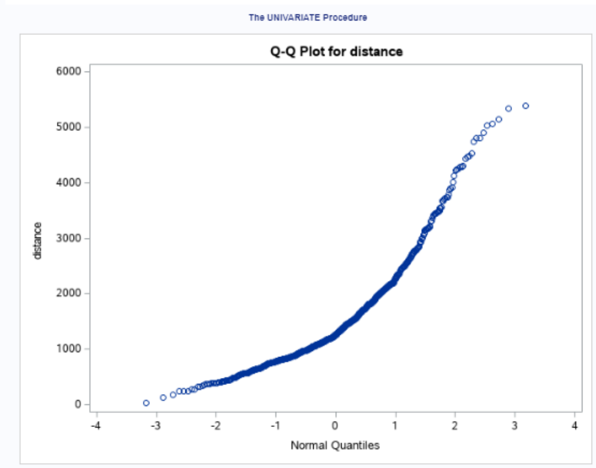
Moments			
N	831	Sum Weights	831
Mean	1522.48287	Sum Observations	1265183.27
Std Deviation	896.338152	Variance	803422.083
Skewness	1.47639585	Kurtosis	2.54813164
Uncorrected SS	2593080185	Corrected SS	666840329
Coeff Variation	58.8734473	Std Error Mean	31.093626

Basic Statistical Measures			
Location		Variability	
Mean	1522.483	Std Deviation	896.33815
Median	1262.154	Variance	803422
Mode	.	Range	5340
		Interquartile Range	1044

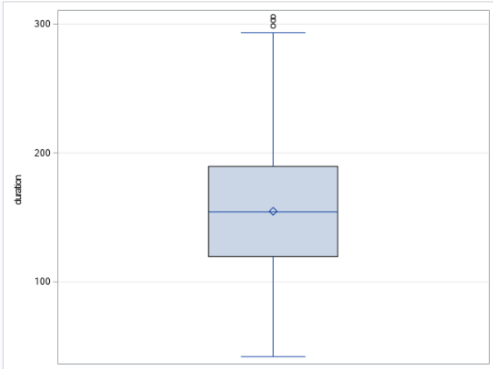
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	48.96447	Pr > |t|	<.0001
Sign	M	415.5	Pr >= |M|	<.0001
Signed Rank	S	172848	Pr >= |S|	<.0001

The variable distance does not follow normal distribution.

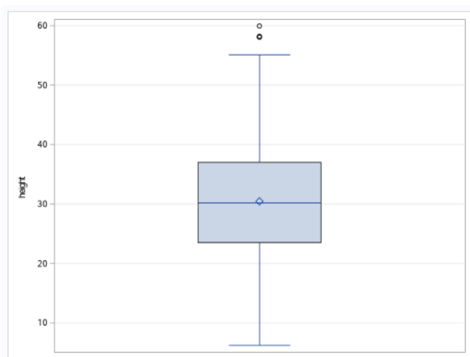




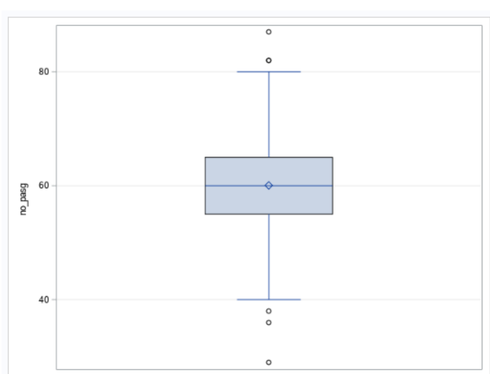

```
/*Duration trend*/

proc sgplot data=FAA_NORMAL;
      vbox duration/;
      yaxis grid;
run;
```



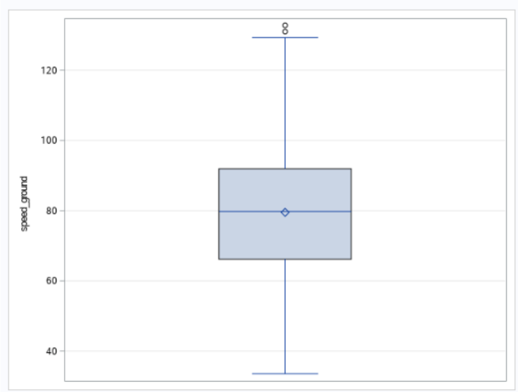

```
/*Height trend*/
proc sgplot data=FAA_NORMAL;
	vbox height/;
	yaxis grid;
run;
```



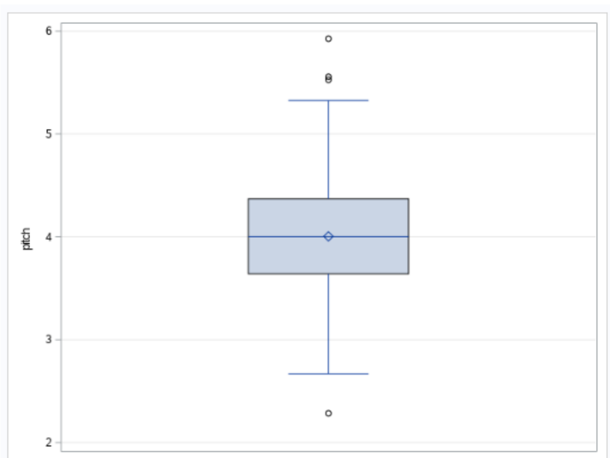

```
/*No_pasg trend*/
proc sgplot data=FAA_FINAL;
	vbox no_pasg/;
	yaxis grid;
run;
```



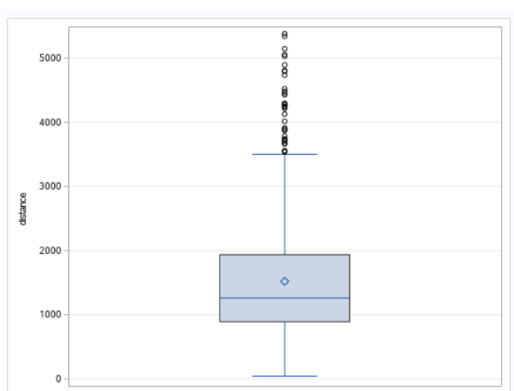

```
/*Speed_ground trend*/
proc sgplot data=FAA_FINAL;
	vbox speed_ground/;
	yaxis grid;
```




```
/*Pitch trend*/

proc sgplot data=FAA_FINAL;
      vbox pitch/;
      yaxis grid;
run;
```




```
/*Distance trend*/

proc sgplot data=FAA_FINAL;
      vbox distance/;
      yaxis grid;
run;
```



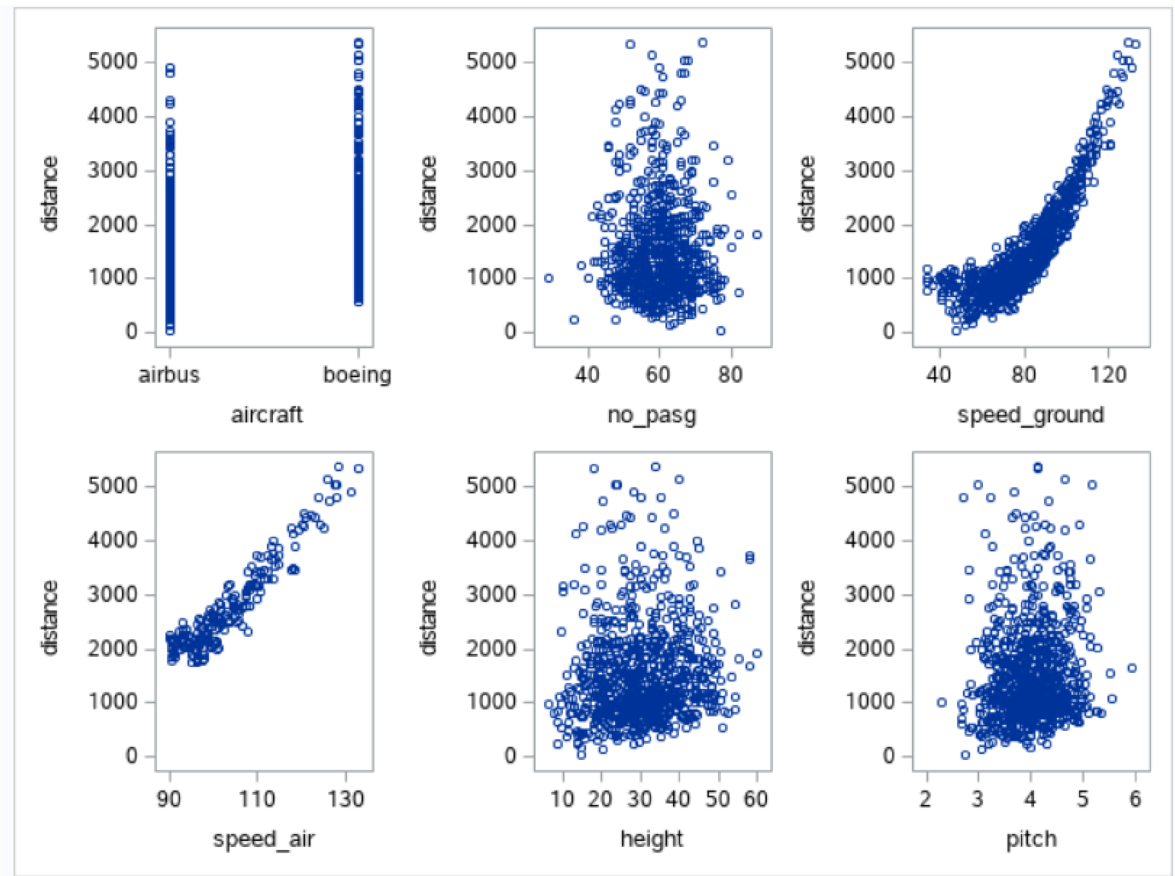
Chapter 2. Data visualization

In order to prepare the data for linear modeling, let us check for linearity between the variables and understand the correlation. If there is a linear relationship between the dependent and independent variable, the plot will be a straight line else it will be scattered.

SAS Code:

```
PROC SGSCATTER DATA=FAA_NORMAL;
PLOT DISTANCE*(AIRCRAFT NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH);
RUN;
```

Output:



Explanation:

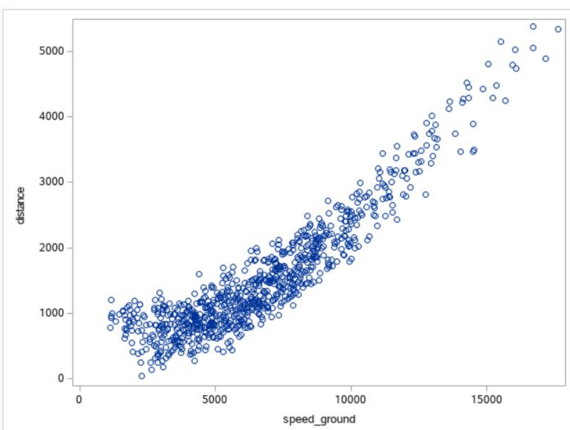
From the above graphs, we get to see the relationship between our dependent and independent variables. The columns speed_air and speed_ground seem to have collinearity with distance.

To increase the linearity with Y, transform the variable speed_ground (x) in our modeling to x^2 or exp(x)

SAS Code:

```
DATA FAA_NORMAL1;
SET FAA_NORMAL;
SPEED_GROUND1=SPEED_GROUND**2;
RUN;


PROC SGSCATTER DATA=FAA_NORMAL1;
PLOT DISTANCE*(SPEED_GROUND1);
RUN;
```

Output:



After performing the transformation, the variable seems to have a much better linear relationship.

Since Aircraft column is a categorical variable and it has 2 levels – Airbus and Boeing, create the numerical values 0 and 1 to represent them.

```
DATA FAA_FINAL;
SET FAA_NORMAL1;
IF COMPRESS(UPCASE(AIRCRAFT)) ="BOEING" THEN AIRCRAFT_TYPE=1;
ELSE AIRCRAFT_TYPE=0;
RUN;
```

Now we calculate the Pearson's correlation coefficient to understand the relationship between the variables.

```
PROC CORR DATA=FAA_FINAL;
VAR AIRCRAFT_TYPE DURATION NO_PASG SPEED_GROUND SPEED_GROUND1
SPEED_AIR HEIGHT PITCH DISTANCE;
RUN;
```

The CORR Procedure

9 Variables:	AIRCRAFT_TYPE duration no_pasg speed_ground SPEED_GROUND1 speed_air height pitch distance

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
AIRCRAFT_TYPE	831	0.46570	0.49912	387.00000	0	1.00000	
duration	781	154.77572	48.34992	120880	41.94937	305.62171	duration
no_pasg	831	60.05535	7.49132	49906	29.00000	87.00000	no_pasg
speed_ground	831	79.54270	18.73568	66100	33.57410	132.78468	speed_ground
SPEED_GROUND1	831	6678	3047	5549122	1127	17632	
speed_air	203	103.48504	9.73628	21007	90.00286	132.91146	speed_air
height	831	30.45787	9.78481	25310	6.22752	59.94596	height
pitch	831	4.00516	0.52657	3328	2.28448	5.92678	pitch
distance	831	1522	896.33815	1265183	41.72231	5382	distance

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

	AIRCRAFT_TYPE	duration	no_pasg	speed_ground	SPEED_GROUND1	speed_air	height	pitch	distance
AIRCRAFT_TYPE	1.00000 831	-0.04443 0.2149 781	-0.02269 0.5136 831	-0.04045 0.2441 831	-0.01731 0.6183 831	-0.07207 0.3069 203	-0.01439 0.6788 831	0.35420 <.0001 831	0.23814 <.0001 831
duration duration	-0.04443 0.2149 781	1.00000 781	-0.03639 0.3098 781	-0.04897 0.1716 781	-0.04849 0.1758 781	0.04454 0.5364 195	0.01112 0.7564 781	-0.04675 0.1918 781	-0.05138 0.1514 781
no_pasg no_pasg	-0.02269 0.5136 831	-0.03639 0.3098 781	1.00000 831	-0.00013 0.9969 831	-0.00182 0.9582 831	-0.00616 0.9305 203	0.04699 0.1760 831	-0.01793 0.6057 831	-0.01776 0.6093 831
speed_ground speed_ground	-0.04045 0.2441 831	-0.04897 0.1716 781	-0.00013 0.9969 831	1.00000 831	0.98831 <.0001 831	0.98794 <.0001 203	-0.05761 0.0970 831	-0.03912 0.2599 831	0.86624 <.0001 831
SPEED_GROUND1	-0.01731 0.6183 831	-0.04849 0.1758 781	-0.00182 0.9582 831	0.98831 <.0001 831	1.00000 831	0.98774 <.0001 203	-0.05417 0.1187 831	-0.02900 0.4037 831	0.91657 <.0001 831
speed_air speed_air	-0.07207 0.3069 203	0.04454 0.5364 195	-0.00616 0.9305 203	0.98794 <.0001 203	0.98774 <.0001 203	1.00000 203	-0.07933 0.2606 203	-0.03927 0.5780 203	0.94210 <.0001 203
height height	-0.01439 0.6788 831	0.01112 0.7564 781	0.04699 0.1760 831	-0.05761 0.0970 831	-0.05417 0.1187 831	-0.07933 0.2606 203	1.00000 831	0.02298 0.5082 831	0.09941 0.0041 831
pitch pitch	0.35420 <.0001 831	-0.04675 0.1918 781	-0.01793 0.6057 831	-0.03912 0.2599 831	-0.02900 0.4037 831	-0.03927 0.5780 203	0.02298 0.5082 831	1.00000 831	0.08703 0.0121 831
distance distance	0.23814 <.0001 831	-0.05138 0.1514 781	-0.01776 0.6093 831	0.86624 <.0001 831	0.91657 <.0001 831	0.94210 <.0001 203	0.09941 0.0041 831	0.08703 0.0121 831	1.00000 831

Output:

Since the p-value for distance vs no_pasg is less than 0.05 it does not have any relation with the distance. Thus, we infer that the no_pasg and aircraft_type don't play a significant role in explaining our response variables.

Also, there is high correlation between distance vs speed_ground and distance vs speed_air, one of the variables can be dropped. Since speed_air has missing values in it, we can drop the speed_air from our modeling.

```
DATA FAA_FIN2(DROP=SPEED_AIR);
SET FAA_FINAL;
RUN;
```

Chapter 3. Modeling

The objective of modelling is to build an equation for the response variable to understand its dependence on the independent variables chosen. We concerned with finding a model that describes the relationship between distance and several predictor (explanatory) variables by regression. A linear model has the form Y = b0 + b1X + ε. The constant b0 is called the intercept and the coefficient b1 is the parameter estimate for the variable X. The ε is the error term. ε is the residual that cannot be explained by the variables in the model.

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.

```
PROC REG DATA=FAA_FIN2;
MODEL DISTANCE= DURATION AIRCRAFT_TYPE SPEED_GROUND1 HEIGHT PITCH/r
spec;
output out = FAA_FIN3 R=RESIDUAL;
RUN;
```

Output:

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	781
Number of Observations with Missing Values	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	593731773	118746355	2066.69	<.0001
Error	775	44529486	57457		
Corrected Total	780	638261260			

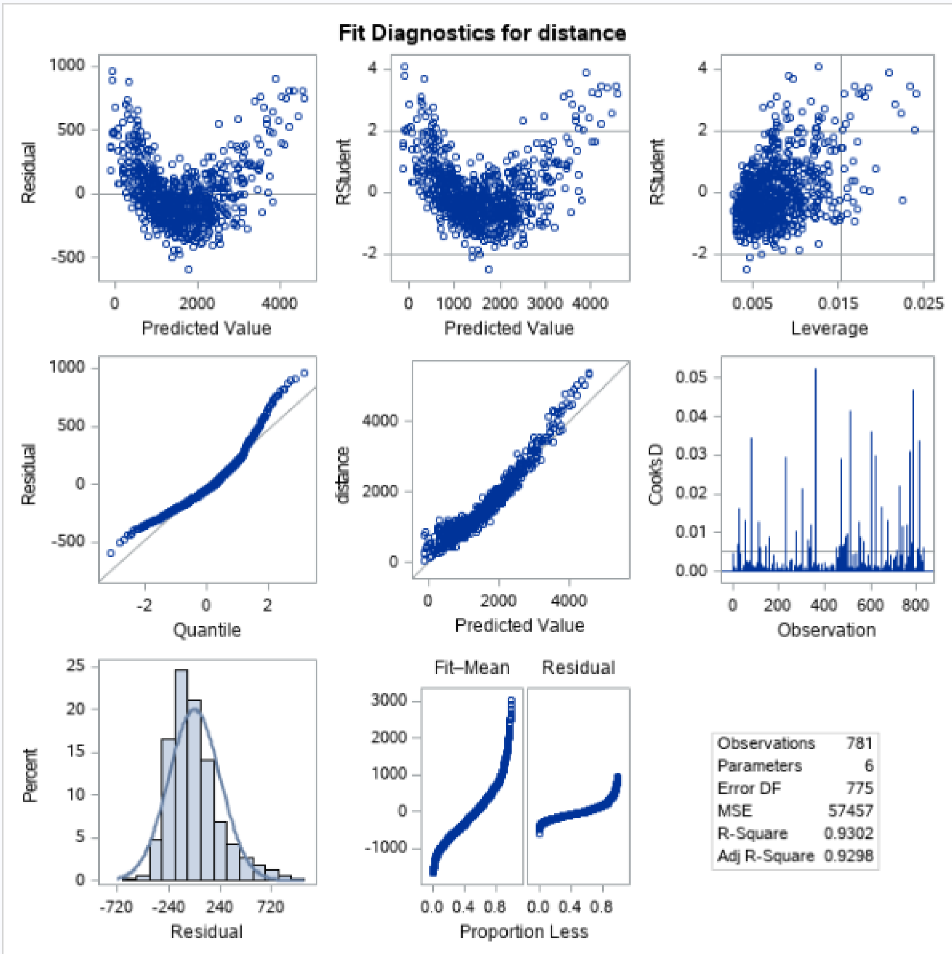
Root MSE	239.70274	R-Square	0.9302
Dependent Mean	1541.20394	Adj R-Sq	0.9298
Coeff Var	15.55295		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1049.86800	82.12698	-12.78	<.0001
duration	duration	1	0.07278	0.17802	0.41	0.6828
AIRCRAFT_TYPE		1	454.45984	18.42096	24.67	<.0001
SPEED_GROUND1		1	0.27407	0.00280	97.94	<.0001
height	height	1	14.17052	0.88318	16.04	<.0001
pitch	pitch	1	21.67065	17.66355	1.23	0.2202

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
20	50.61	0.0002

Sum of Residuals	1.517274E-9
Sum of Squared Residuals	44529486
Predicted Residual SS (PRESS)	45457434



The F value is as high as 2066 and R square is .9302 which shows that the independent variables very clearly explain our response variable distance and thus we are in a position to obtain our equation. The p value for the duration is greater than 0.05 and the parameter estimate is 0.07 which is nearly 0, this variable is not considered in the equation.

So, it should not be a part of the equation. Rest all variables have their pvalue greater than 0.05 thus they make our equation.

BUILDING the EQAUTION:

$Y = b0 + b1X1 + b2X2 + \varepsilon$
Y = distance
B0 = -1049

B1= 454.45
X1=aircraft_type
B2=0.27
X2=speed_ground1
B3=14
X3=height
B4=21
X4=pitch
Distance = -1049 + 454.45(aircraft_name) + 0.27(speed_ground1) + 14(height) + 21(pitch)

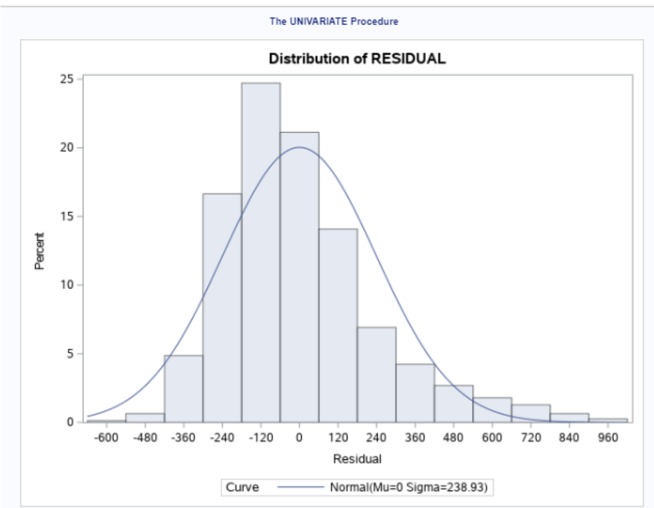
4. Model Checking

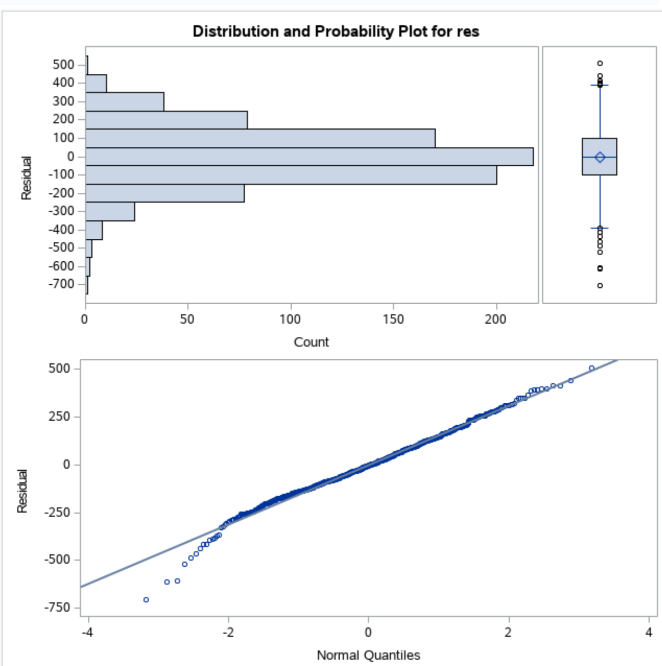
The objective of model checking is to check the assumptions for the noise terms. They are assumed to be: 1. Independent 2. Normally distributed. 3. Mean 0 4. Constant Variance

We will validate that the residuals are independent as it is an assumption of linear regression by examining the residuals of our final model. Specifically, we will use diagnostic statistics from REG as well as create an output dataset of residual values for PROC UNIVARIATE to test.

```
/*STATISTICS*/
PROC UNIVARIATE DATA=FAA_FIN3 NORMAL;
HISTOGRAM/NORMAL;
VAR RESIDUAL;
RUN;
```

The p of Chi-square value is less than 0.05. The distribution of the residuals.





The mean value is 0.

The UNIVARIATE Procedure
Variable: RESIDUAL (Residual)

Moments			
N	781	Sum Weights	781
Mean	0	Sum Observations	0
Std Deviation	238.933223	Variance	57089.085
Skewness	1.11473719	Kurtosis	1.63562663
Uncorrected SS	44529486.3	Corrected SS	44529486.3
Coeff Variation	.	Std Error Mean	8.54970291

Basic Statistical Measures			
Location		Variability	
Mean	0.0000	Std Deviation	238.93322
Median	-44.2062	Variance	57089
Mode	.	Range	1558
		Interquartile Range	280.40470

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-74.5	Pr >= |M|	<.0001
Signed Rank	S	-19402.5	Pr >= |S|	0.0020

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.932406	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.099607	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.296826	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	13.98464	Pr > A-Sq	<0.0050

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

The final model built consists of below linear equation:

Distance = -1049 + 454.45*(aircraft_name) + 0.27*(speed_ground1) + 14*(height) +21*(pitch)

(aircraft name=0 for airbus and 1 for boeing)

This equation specifies that:

1. If the aircraft_name value belongs to 0, it implies the model built for Airbus and 1 implies the model for Boeing.

2. For ‘Boeing’ aircraft type the predicted landing distance would be 454 units greater than the landing distance for ‘Airbus’ aircraft type.

3. For every one-unit increase in pitch there will be 21-unit increase in the predicted landing distance

4. For every one-unit increase in square of ground speed there will be 0.27-unit increase in the predicted landing distance

5. For every one-unit increase in height there will be 14-unit increase in the predicted landing distance

Questions Answered:

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

We are originally given 950 records of data. After data cleaning, we are left with 831 observations that can be used for model building. 100 records were dropped because they are duplicate and post that we dropped a few records based on the definitions in the Variable Dictionary.

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

Distance = -1049 + 454.45(aircraft_name) + 0.27(speed_ground1) + 14(height) + 21(pitch)

(aircraft name=0 for airbus and 1 for boeing)

- Landing Distance is a dependent variable which depends on height, speed_ground and type of aircraft
- A unit increase in height increases landing distance by 14.
- Speed_ground ^ 2 shows a change in Landing Distance in terms of 0.27***(speed_ground)^2**
- For 'Boeing, the formulated distance will be 454.45 units lesser than 'Airbus' keeping all other factors the same.

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

We can see that the aircraft variable is very significant since the probability is < 0.0001 in the Parameter Estimates Table. We can see that the distance. For 'Airbus', the formulated distance will be 454.45 units greater than for 'Boeing' keeping all other factors the same.

~End of the Project~