SUMMARY

Business problem:

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

The business problem required the study of various variables contributing in the variability of landing distance. The project started with data exploration of 950 records which after data cleaning got reduced to 781 records. Following this the correlation analysis was done to figure out the variables which might affect landing. 'speed_air', 'speed_ground','height', 'pitch' were the major contributing factors out of which 'pitch' affected the landing distance in accordance with aircraft make. Last but not the least, regression analysis was done which generated a linear model with highly significant R-square of 0.88.

To conclude, 'speed_air' impacted the landing distance the most as compared to any other variable (contribution of other variables was negligible as compared to 'speed_air'). 'speed_air' accounted for 88% of the variation in 'distance'

The fitted model is

distance = -5417.6 + 79.2 * speed_air

TABLE OF CONTENTS

CHAPTER 1

DATA UNDERSTANDING AND DATA EXPLORATION

Data Source and Tools

Datasets:

Two excel sheets named FAA1.xls and FAA2.xls containing flight information which were provided by the professor.

Tool Used:

SAS University Edition

Data Description

1. FAA1.xlsx
    - **a.** Contains 8 variables and 800 entries
    - **b.** Variables are:
        - **i.** *Aircraft:* The make of an aircraft (Boeing or Airbus).
        - **ii.** *Duration (in minutes):* Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.
        - **iii.** *No_pasg:* The number of passengers in a flight.
        - **iv.** *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.
        - **v.** *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.
        - **vi.** *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.
        - **vii.** *Pitch (in degrees):* Pitch angle of an aircraft when it is passing over the threshold of the runway.
        - **viii.** *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.
2. FAA2.xlsx
    - **a.** Contains 7 variables and 150 entries
    - **b.** All variables apart from “duration” in FAA1.xlsx

Data Exploration

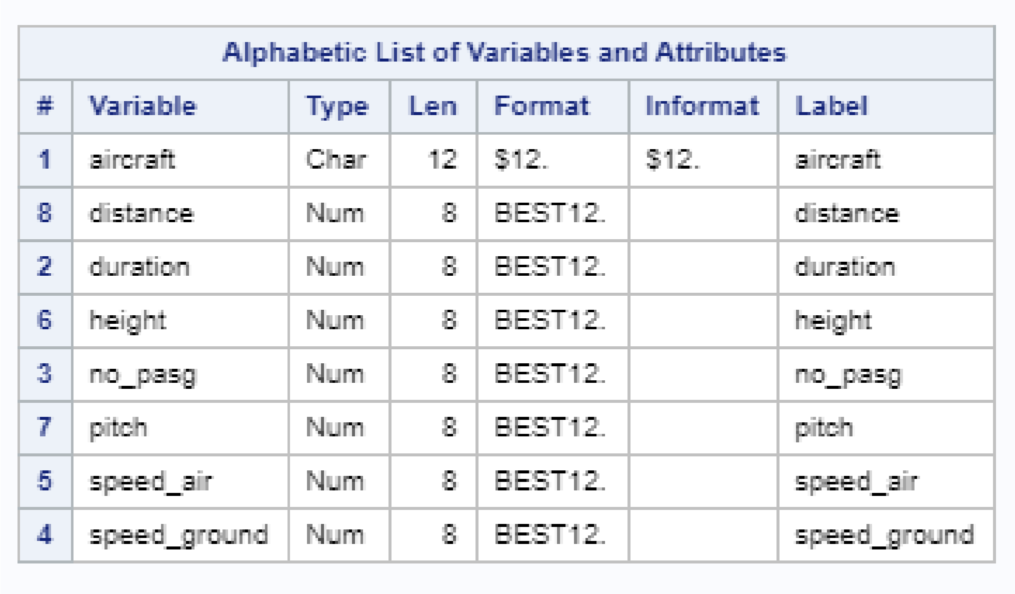
Variable Datatypes

Datatypes of the variables are examined to get an idea of the dataset.

SAS code:

```
proc contents data=FAA_Cleaned_Dataset;

run;
```

Output Snippet:



Alphabetic List of Variables and Attributes

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

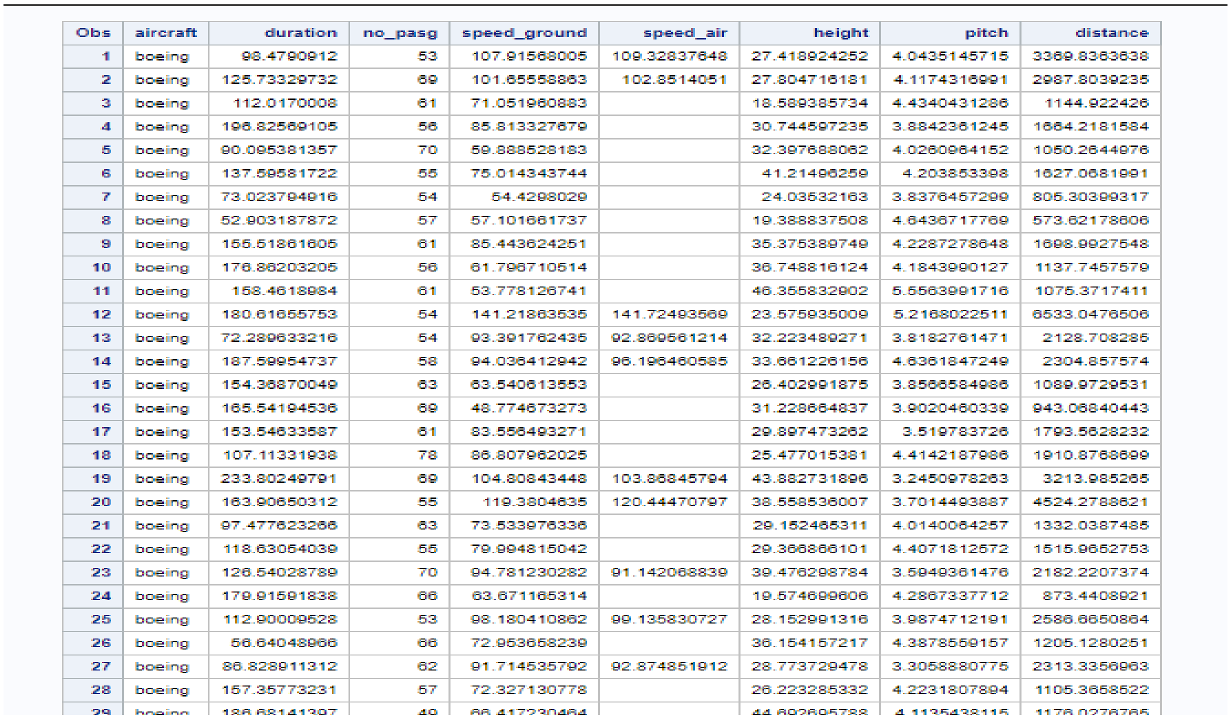
Aggregating Data

The two excel sheets are imported in SAS University Edition and the two datasets are joined to form a single dataset. Completely empty rows are removed from the dataset. The resultant dataset is of 950 observations and named ‘FAA_combined’.

SAS code:

```
proc import datafile="/folders/myfolders/bana 6043/Resources/FAA1.xls"

out=faa1

dbms=xls

replace;

getnames=yes;

run;


proc import datafile="/folders/myfolders/bana 6043/Resources/FAA2.xls"

out=faa2

dbms=xls

replace;

getnames=yes;

run;


options missing = ' ';

DATA FAA_combined;

SET faa1 faa2;

if missing(cats(of _all_)) then delete;

run;

proc print data= FAA_combined;

run;
```

Output Snippet:



Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	112.0170008	61	71.051960883		18.589385734	4.4340431286	1144.922426
4	boeing	196.82569105	56	85.813327679		30.744597235	3.8842361245	1664.2181584
5	boeing	90.095381357	70	59.888528183		32.397688062	4.0260964152	1050.2644976
6	boeing	137.59581722	55	75.014343744		41.21496259	4.203853398	1627.0681991
7	boeing	73.023794916	54	54.4298029		24.03532163	3.8376457299	805.30399317
8	boeing	52.903187872	57	57.101661737		19.388837508	4.6436717769	573.62178606
9	boeing	155.51861605	61	85.443624251		35.375389749	4.2287278648	1698.9927548
10	boeing	176.86203205	56	61.796710514		36.748816124	4.1843990127	1137.7457579
11	boeing	158.4618984	61	53.778126741		46.355832902	5.5563991716	1075.3717411
12	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506
13	boeing	72.289633216	54	93.391762435	92.869561214	32.223489271	3.8182761471	2128.708285
14	boeing	187.59954737	58	94.036412942	96.196460585	33.661226156	4.6361847249	2304.857574
15	boeing	154.36870049	63	63.540613553		26.402991875	3.8566584986	1089.9729531
16	boeing	165.54194536	69	48.774673273		31.228664837	3.9020460339	943.06840443
17	boeing	153.54633587	61	83.556493271		29.897473262	3.519783726	1793.5628232
18	boeing	107.11331938	78	86.807962025		25.477015381	4.4142187986	1910.8768699
19	boeing	233.80249791	69	104.80843448	103.86845794	43.882731896	3.2450978263	3213.985265
20	boeing	163.90650312	55	119.3804635	120.44470797	38.558536007	3.7014493887	4524.2788621
21	boeing	97.477623266	63	73.533976336		29.152465311	4.0140064257	1332.0387485
22	boeing	118.63054039	55	79.994815042		29.366866101	4.4071812572	1515.9652753
23	boeing	126.54028789	70	94.781230282	91.142068839	39.476298784	3.5949361476	2182.2207374
24	boeing	179.91591838	66	63.671165314		19.574699606	4.2867337712	873.4408921
25	boeing	112.90009528	53	98.180410862	99.135830727	28.152991316	3.9874712191	2586.6650864
26	boeing	56.64048966	66	72.953658239		36.154157217	4.3878559157	1205.1280251
27	boeing	86.828911312	62	91.714535792	92.874851912	28.773729478	3.3058880775	2313.3356963
28	boeing	157.35773231	57	72.327130778		26.223285332	4.2231807894	1105.3658522
29	boeing	186.68141397	49	66.417230464		44.692695788	4.1135438115	1176.0276765

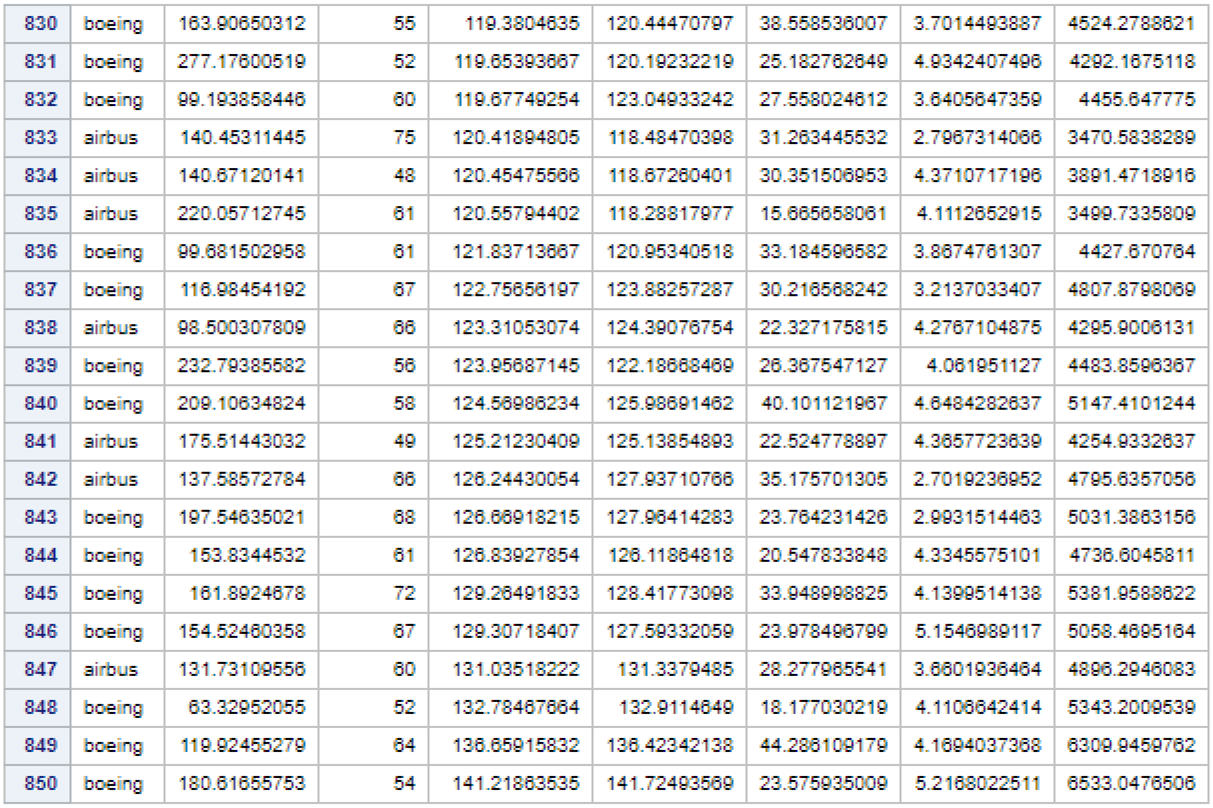
Removing Duplicate Records

Variables ‘speed_ground’ and ‘distance’ are used as key to uniquely identify each observation and remove duplicate records.
After removal of duplicate records, only 850 records are left. 100 records were removed.

SAS code:

```
PROC SORT data= FAA_combined NODUPKEY out= FAA_NoDup;

 BY speed_ground distance;

RUN;


proc print data= FAA_NoDup;

run;
```

Output Snippet:



830	boeing	163.90650312	55	119.3804635	120.44470797	38.558536007	3.7014493887	4524.2788621
831	boeing	277.17600519	52	119.65393667	120.19232219	25.182762649	4.9342407496	4292.1675118
832	boeing	99.193858446	60	119.67749254	123.04933242	27.558024612	3.6405647359	4455.647775
833	airbus	140.45311445	75	120.41894805	118.48470398	31.263445532	2.7967314066	3470.5838289
834	airbus	140.67120141	48	120.45475566	118.67260401	30.351506953	4.3710717196	3891.4718916
835	airbus	220.05712745	61	120.55794402	118.28817977	15.665658061	4.1112652915	3499.7335809
836	boeing	99.681502958	61	121.83713667	120.95340518	33.184596582	3.8674761307	4427.670764
837	boeing	116.98454192	67	122.75656197	123.88257287	30.216568242	3.2137033407	4807.8798069
838	airbus	98.500307809	66	123.31053074	124.39076754	22.327175815	4.2767104875	4295.9006131
839	boeing	232.79385582	56	123.95687145	122.18668469	26.367547127	4.061951127	4483.8596367
840	boeing	209.10634824	58	124.56986234	125.98691462	40.101121967	4.6484282637	5147.4101244
841	airbus	175.51443032	49	125.21230409	125.13854893	22.524778897	4.3657723639	4254.9332637
842	airbus	137.58572784	66	126.24430054	127.93710766	35.175701305	2.7019236952	4795.6357056
843	boeing	197.54635021	68	126.66918215	127.96414283	23.764231426	2.9931514463	5031.3863156
844	boeing	153.8344532	61	126.83927854	126.11864818	20.547833848	4.3345575101	4736.6045811
845	boeing	161.8924678	72	129.26491833	128.41773098	33.948998825	4.1399514138	5381.9588622
846	boeing	154.52460358	67	129.30718407	127.59332059	23.978496799	5.1546989117	5058.4695164
847	airbus	131.73109556	60	131.03518222	131.3379485	28.277965541	3.6601936464	4896.2946083
848	boeing	63.32952055	52	132.78467664	132.9114649	18.177030219	4.1106642414	5343.2009539
849	boeing	119.92455279	64	136.65915832	136.42342138	44.286109179	4.1694037368	6309.9459762
850	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506

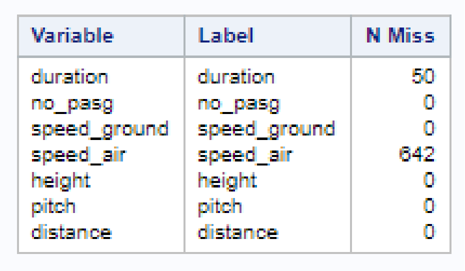
Checking Missing Values

‘duration’ has 50 missing values (6% approximately) and ‘speed_air’ has 642 missing values (75% approximately)

SAS code:

```
proc means data= FAA_NoDup NMISS;
run;
```

Output Snippet:



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

Verifying Data Quality

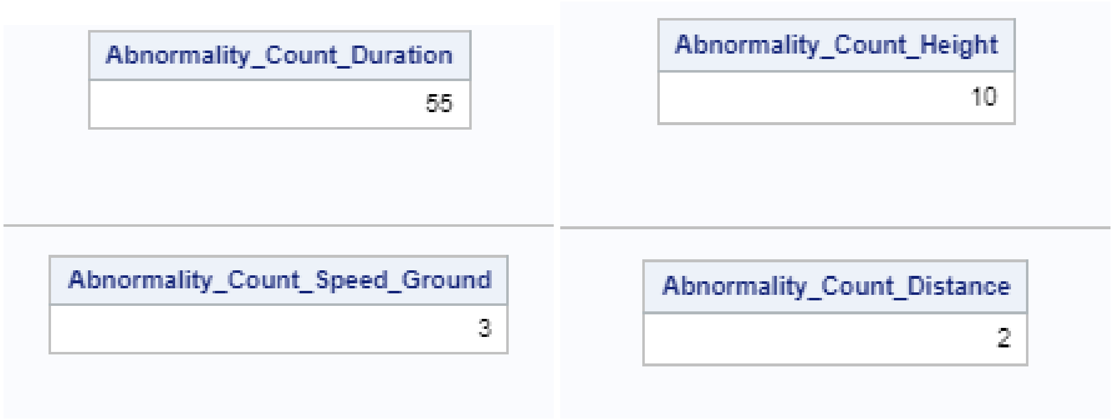
Abnormality within the dataset is detected and represented in tabular format.

Variable	Restrictions	Count of Violating observations	Percentage
duration	>40	55	6.5
speed_ground	30<s<140	3	0.3
height	>6	10	1.1
distance	<6000	2	0.2

SAS code:

```
proc sql;
 select count(*) as Abnormality_Count_Duration from FAA_NoDup
 where duration < 40
 ;
RUN;
proc sql;
 select count(*) as Abnormality_Count_Speed_Ground from FAA_NoDup
 where speed_ground < 30 or speed_ground > 140
 ;
RUN;
proc sql;
 select count(*) as Abnormality_Count_Height from FAA_NoDup
 where height < 6
 ;
RUN;
proc sql;
 select count(*) as Abnormality_Count_Distance from FAA_NoDup
 where distance > 6000
 ;
RUN;
```

Output Snippet:



Abnormality_Count_Duration
55

Abnormality_Count_Height
10

Abnormality_Count_Speed_Ground
3

Abnormality_Count_Distance
2

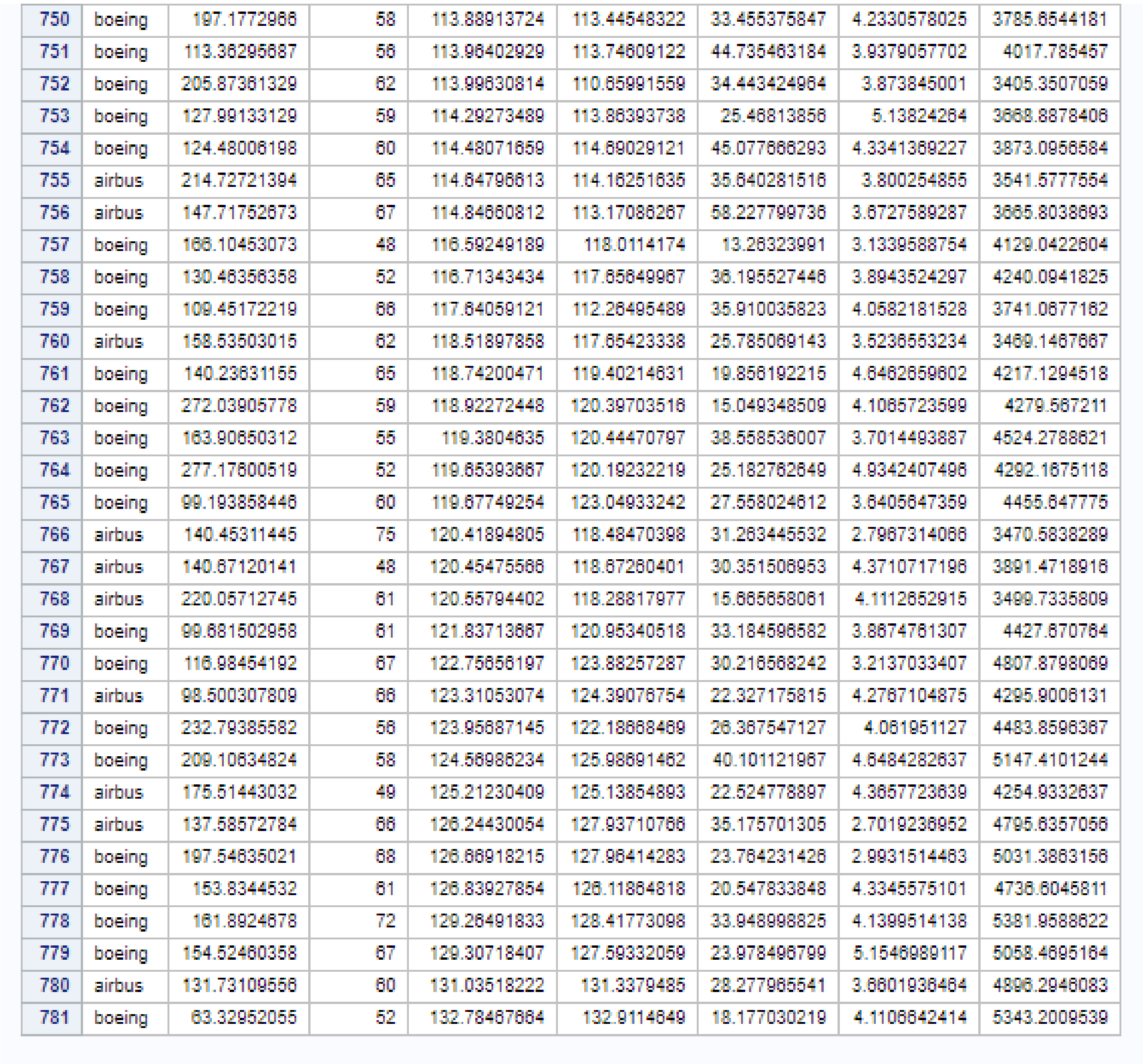
Removing abnormal values

Removing abnormalities present in the dataset. Abnormalities are defined in the variable description of the data.

SAS code:

```
data FAA_cleaned;

set FAA_NoDup;

if height<6 then delete;

if distance > 6000 then delete;

if speed_ground < 30 or speed_ground > 140 then delete;

if duration < 40 then delete;

run;


proc print data=FAA_cleaned;

run;
```

Output Snippet:



750	boeing	197.1772966	58	113.88913724	113.44548322	33.455375847	4.2330578025	3785.6544181
751	boeing	113.36295687	56	113.96402929	113.74609122	44.735463184	3.9379057702	4017.785457
752	boeing	205.87361329	62	113.99630814	110.65991559	34.443424964	3.873845001	3405.3507059
753	boeing	127.99133129	59	114.29273489	113.86393738	25.46813856	5.13824264	3668.8878406
754	boeing	124.48006198	60	114.48071659	114.69029121	45.077666293	4.3341369227	3873.0956584
755	airbus	214.72721394	65	114.64796613	114.16251635	35.640281516	3.800254855	3541.5777554
756	airbus	147.71752673	67	114.84660812	113.17086267	58.227799736	3.6727589287	3665.8038693
757	boeing	166.10453073	48	116.59249189	118.0114174	13.26323991	3.1339588754	4129.0422604
758	boeing	130.46356358	52	116.71343434	117.65649967	36.195527446	3.8943524297	4240.0941825
759	boeing	109.45172219	66	117.64059121	112.26495489	35.910035823	4.0582181528	3741.0677162
760	airbus	158.53503015	62	118.51897858	117.65423338	25.785069143	3.5236553234	3469.1467667
761	boeing	140.23631155	65	118.74200471	119.40214631	19.856192215	4.6462659602	4217.1294518
762	boeing	272.03905778	59	118.92272448	120.39703516	15.049348509	4.1065723599	4279.567211
763	boeing	163.90650312	55	119.3804635	120.44470797	38.558536007	3.7014493887	4524.2788621
764	boeing	277.17600519	52	119.65393667	120.19232219	25.182762649	4.9342407496	4292.1675118
765	boeing	99.193858446	60	119.67749254	123.04933242	27.558024612	3.6405647359	4455.647775
766	airbus	140.45311445	75	120.41894805	118.48470398	31.263445532	2.7967314066	3470.5838289
767	airbus	140.67120141	48	120.45475566	118.67260401	30.351506953	4.3710717196	3891.4718916
768	airbus	220.05712745	61	120.55794402	118.28817977	15.665658061	4.1112652915	3499.7335809
769	boeing	99.681502958	61	121.83713667	120.95340518	33.184596582	3.8674761307	4427.670764
770	boeing	116.98454192	67	122.75656197	123.88257287	30.216568242	3.2137033407	4807.8798069
771	airbus	98.500307809	66	123.31053074	124.39076754	22.327175815	4.2767104875	4295.9006131
772	boeing	232.79385582	56	123.95687145	122.18668469	26.367547127	4.061951127	4483.8596367
773	boeing	209.10634824	58	124.56986234	125.98691462	40.101121967	4.6484282637	5147.4101244
774	airbus	175.51443032	49	125.21230409	125.13854893	22.524778897	4.3657723639	4254.9332637
775	airbus	137.58572784	66	126.24430054	127.93710766	35.175701305	2.7019236952	4795.6357056
776	boeing	197.54635021	68	126.66918215	127.96414283	23.764231426	2.9931514463	5031.3863156
777	boeing	153.8344532	61	126.83927854	126.11864818	20.547833848	4.3345575101	4736.6045811
778	boeing	161.8924678	72	129.26491833	128.41773098	33.948998825	4.1399514138	5381.9588622
779	boeing	154.52460358	67	129.30718407	127.59332059	23.978496799	5.1546989117	5058.4695164
780	airbus	131.73109556	60	131.03518222	131.3379485	28.277965541	3.6601936464	4896.2946083
781	boeing	63.32952055	52	132.78467664	132.9114649	18.177030219	4.1106642414	5343.2009539

The new count of the number of the observations is 781 which means 69 observations were removed out of 850 as they had abnormal data.

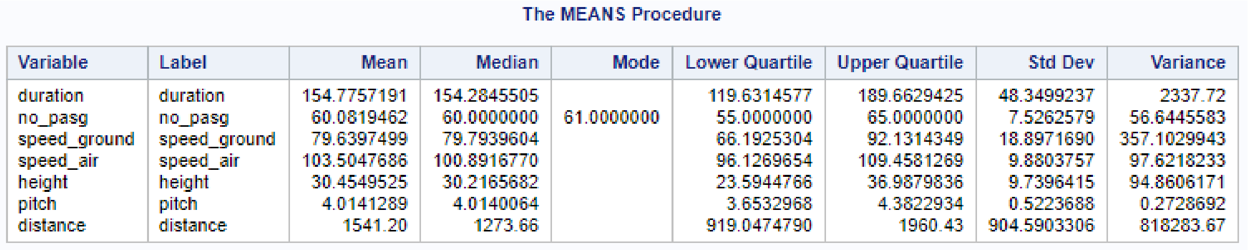
Getting Summary Statistics

Summary Statistics of the different variables are fetched. These include mean, median, mode, lower quartile, upper quartile, standard deviation, variance.

SAS code:

```
proc means data= FAA_cleaned MEAN MEDIAN mode q1 q3 std var;
run;
```

Output Snippet:



The MEANS Procedure

Variable	Label	Mean	Median	Mode	Lower Quartile	Upper Quartile	Std Dev	Variance
duration	duration	154.7757191	154.2845505		119.6314577	189.6629425	48.3499237	2337.72
no_pasg	no_pasg	60.0819462	60.0000000	61.0000000	55.0000000	65.0000000	7.5262579	56.6445583
speed_ground	speed_ground	79.6397499	79.7939604		66.1925304	92.1314349	18.8971690	357.1029943
speed_air	speed_air	103.5047686	100.8916770		96.1269654	109.4581269	9.8803757	97.6218233
height	height	30.4549525	30.2165682		23.5944766	36.9879836	9.7396415	94.8606171
pitch	pitch	4.0141289	4.0140064		3.6532968	4.3822934	0.5223688	0.2728692
distance	distance	1541.20	1273.66		919.0474790	1960.43	904.5903306	818283.67

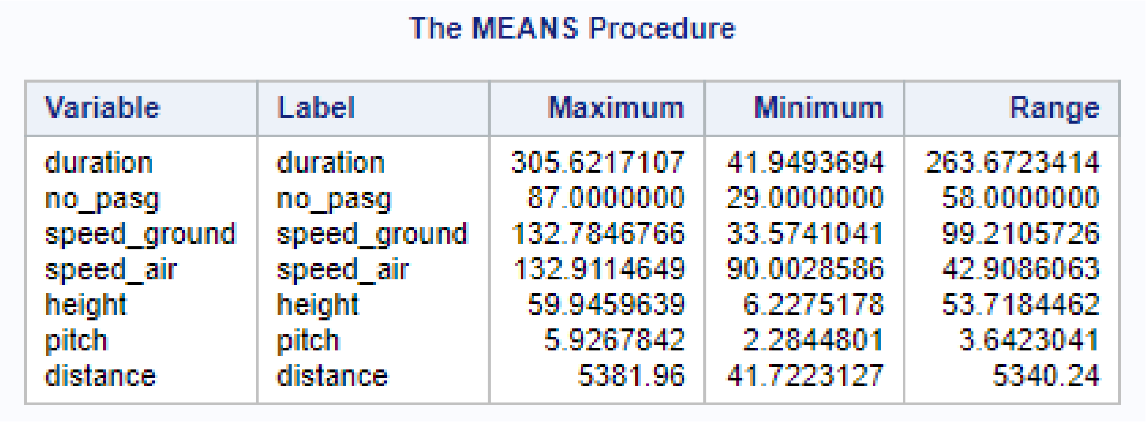
Checking range, minimum and maximum

Large interval ranges are observed in the table with duration leading the list.

SAS code:

```
proc means data= FAA_cleaned max min range;
run;
```

Output Snippet:



The MEANS Procedure

Variable	Label	Maximum	Minimum	Range
duration	duration	305.6217107	41.9493694	263.6723414
no_pasg	no_pasg	87.0000000	29.0000000	58.0000000
speed_ground	speed_ground	132.7846766	33.5741041	99.2105726
speed_air	speed_air	132.9114649	90.0028586	42.9086063
height	height	59.9459639	6.2275178	53.7184462
pitch	pitch	5.9267842	2.2844801	3.6423041
distance	distance	5381.96	41.7223127	5340.24

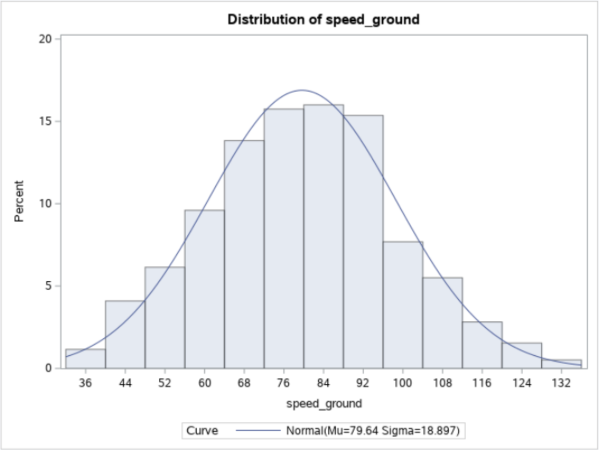
Variable Distribution

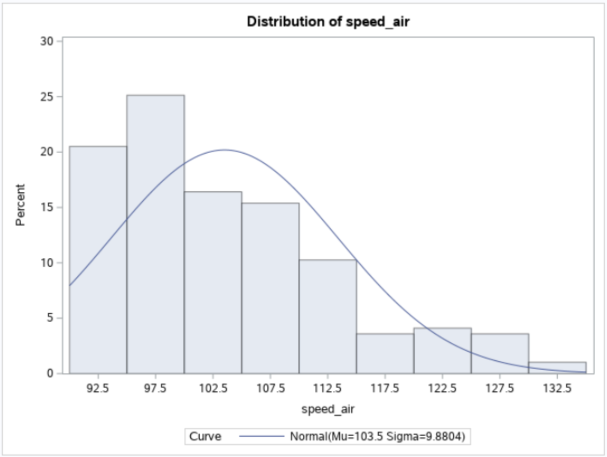
All the variables are taken into account and their distribution is plotted. This will help in finding out whether the data is symmetric or not.

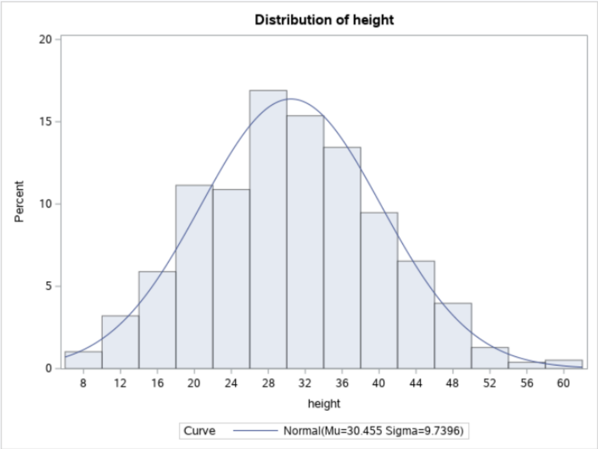
SAS code:

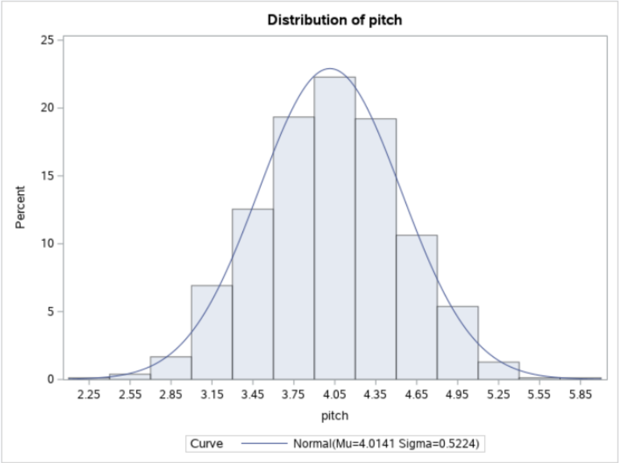
```
PROC UNIVARIATE DATA=FAA_cleaned;
HISTOGRAM speed_ground / NORMAL;
Run;

PROC UNIVARIATE DATA=FAA_cleaned;
HISTOGRAM speed_air / NORMAL;
Run;

PROC UNIVARIATE DATA=FAA_cleaned;
HISTOGRAM height / NORMAL;
Run;

PROC UNIVARIATE DATA=FAA_cleaned;
HISTOGRAM pitch / NORMAL;
Run;

PROC UNIVARIATE DATA=FAA_cleaned;
HISTOGRAM duration / NORMAL;
Run;

PROC UNIVARIATE DATA=FAA_cleaned;
HISTOGRAM distance / NORMAL;
Run;
```

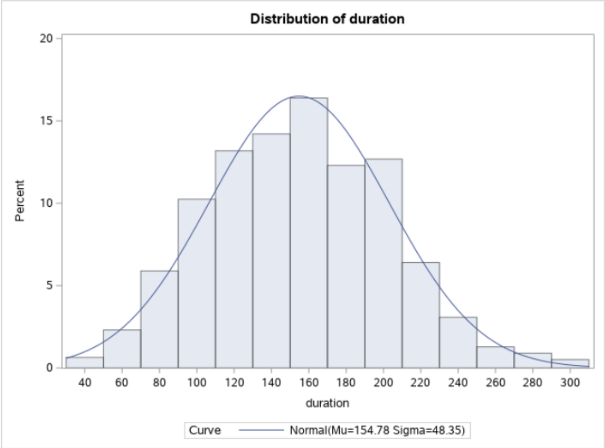
Output Snippet:

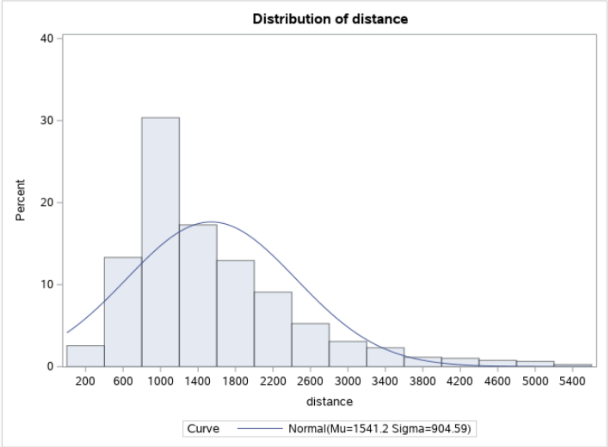












The graphs show that almost all the variables except distance, speed_air follows a normal distribution i.e. values are symmetrically distributed.

CHAPTER 2

DATA VISUALIZATION AND DESCRIPTIVE ANALYSIS

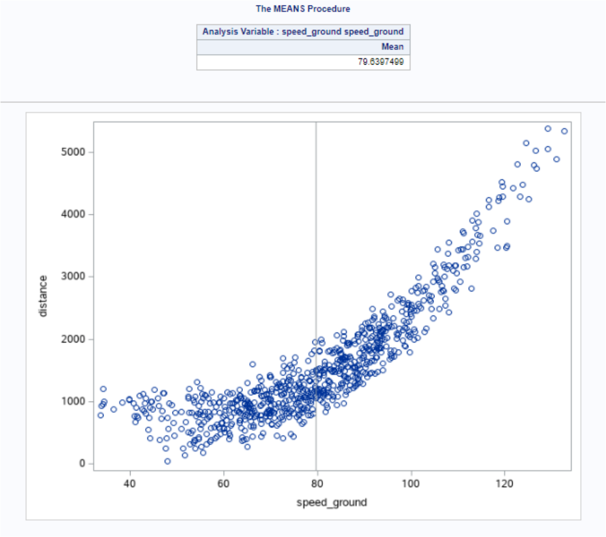
Plot XY graphs

The overall business problem is identifying the factors affecting distance, thus we plot distribution of ‘distance’ against other variables.

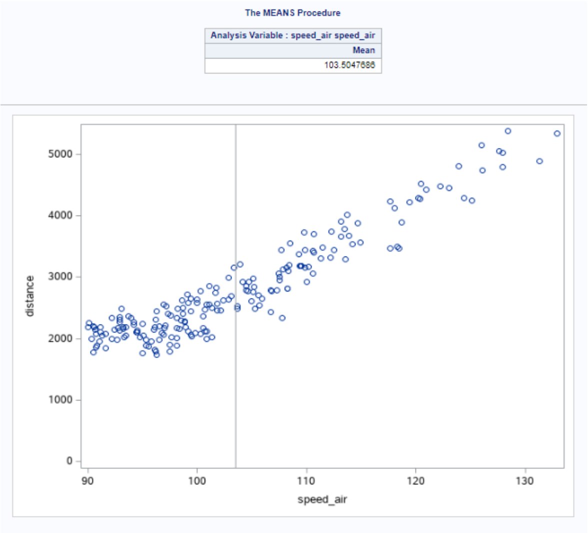
SAS code:

```
proc means data=FAA_cleaned noprint; output out= FAA_mean_summary mean(speed_ground speed_air duration height pitch
no_pasg) =
  speed_ground_average speed_air_average duration_average height_average pitch_average no_pasg_average;
run;
proc print data=FAA_mean_summary;
run;
data FAA_summary;
set FAA_mean_summary FAA_cleaned;
run;

proc sgplot data=FAA_summary;
  scatter x=speed_ground y=distance;
  refline speed_ground_average /Axis=x;
run;
proc sgplot data=FAA_summary;
  scatter x=speed_air y=distance;
  refline speed_air_average /Axis=x;
run;
proc sgplot data=FAA_summary;
  scatter x=duration y=distance;
  refline duration_average /Axis=x;
run;
proc sgplot data=FAA_summary;
  scatter x=height y=distance;
  refline height_average/Axis=x;
run;
proc sgplot data=FAA_summary;
  scatter x=pitch y=distance;
  refline pitch_average/Axis=x;
run;
proc sgplot data=FAA_summary;
  scatter x=no_pasg y=distance;
  refline no_pasg_average/Axis=x;
run;
```

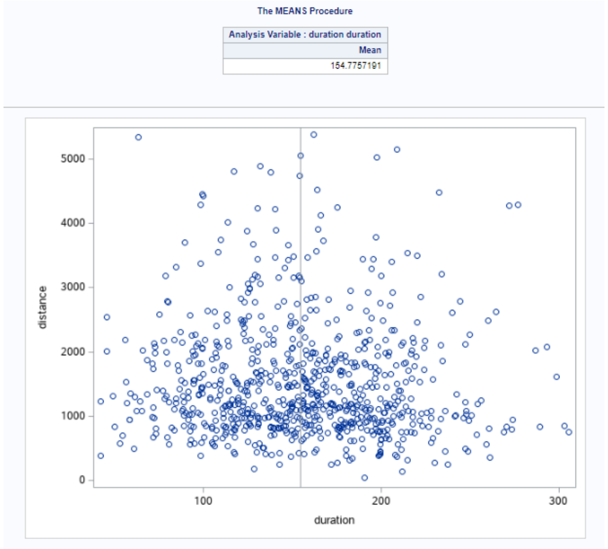
Output Snippet:



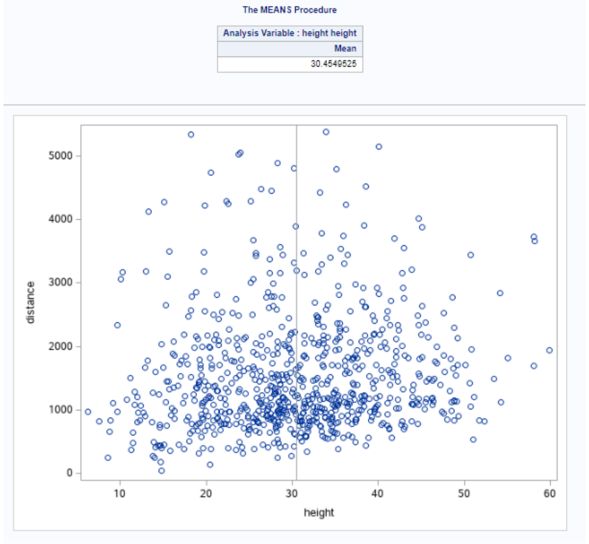
```
The MEANS Procedure

Analysis Variable : speed_ground speed_ground
                                         Mean
                                   79.6397499
```



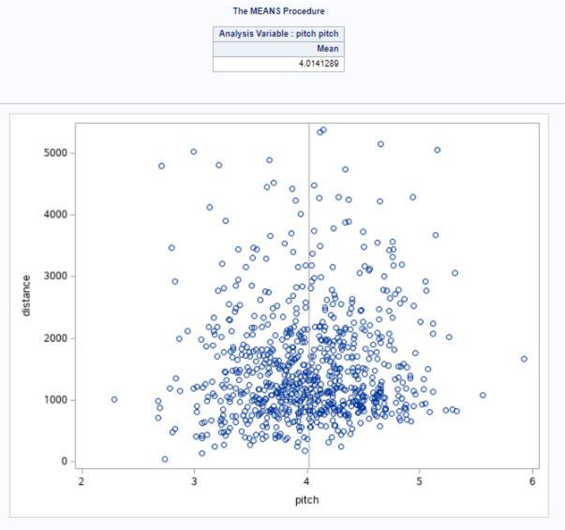
```
The MEANS Procedure

Analysis Variable : speed_air speed_air
                                   Mean
                            103.5047686
```



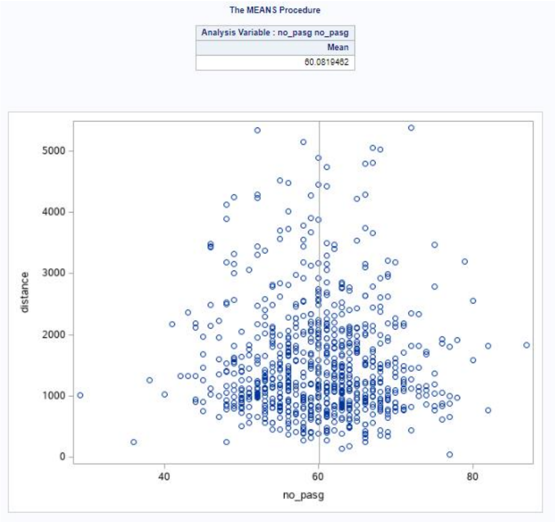
```
The MEANS Procedure

Analysis Variable : duration duration
                                 Mean
                          154.7757191
```



```
The MEANS Procedure

Analysis Variable : height height
                             Mean
                       30.4549525
```



```
The MEANS Procedure

Analysis Variable : pitch pitch
                           Mean
                      4.0141289
```



```
The MEANS Procedure

Analysis Variable : no_pasg no_pasg
                               Mean
                         60.0819462
```

The graph plots show that speed_air forms kind of a linear relation with distance. ‘speed_ground’ does not exactly form linear relation with distance. Other variables, that are pitch, duration and height does not form any relation with distance and are randomly distributed. Additionally, data provided for speed_air has the minimum as 90 miles/hr.

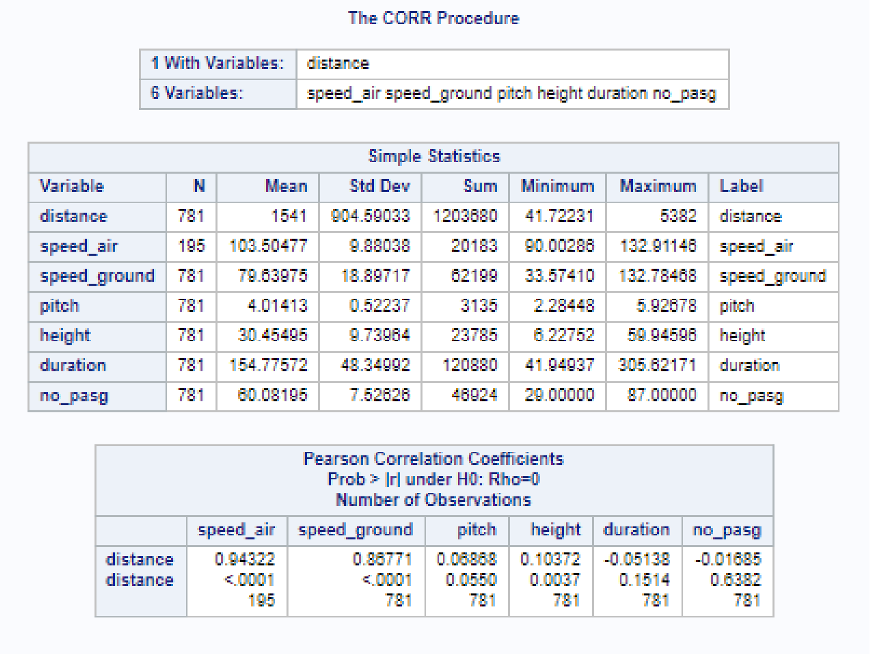
Correlation Analysis

Part A Analysis:

SAS code:

```
PROC CORR data = FAA_cleaned;

VAR speed_air speed_ground pitch height duration no_pasg;

WITH distance;
```

Output Snippet:



The CORR Procedure

1 With Variables:	distance
6 Variables:	speed_air speed_ground pitch height duration no_pasg

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	781	1541	904.59033	1203680	41.72231	5382	distance
speed_air	195	103.50477	9.88038	20183	90.00286	132.91146	speed_air
speed_ground	781	79.63975	18.89717	62199	33.57410	132.78468	speed_ground
pitch	781	4.01413	0.52237	3135	2.28448	5.92678	pitch
height	781	30.45495	9.73984	23785	6.22752	59.94598	height
duration	781	154.77572	48.34992	120880	41.94937	305.62171	duration
no_pasg	781	60.08195	7.52626	46924	29.00000	87.00000	no_pasg

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

	speed_air	speed_ground	pitch	height	duration	no_pasg
distance distance	0.94322 <.0001 195	0.86771 <.0001 781	0.06868 0.0550 781	0.10372 0.0037 781	-0.05138 0.1514 781	-0.01685 0.6382 781

- Correlation matrix shows the speed_air, speed_ground both are highly correlated to distance with value of 0.94322 and 0.86771, respectively.
- speed_air is restricted to 195 observations.
- pitch and height shows a fairly low correlation
- duration and no_pasg have a negligible negative correlation
- speed_air has the highest correlation with a value of 0.94322

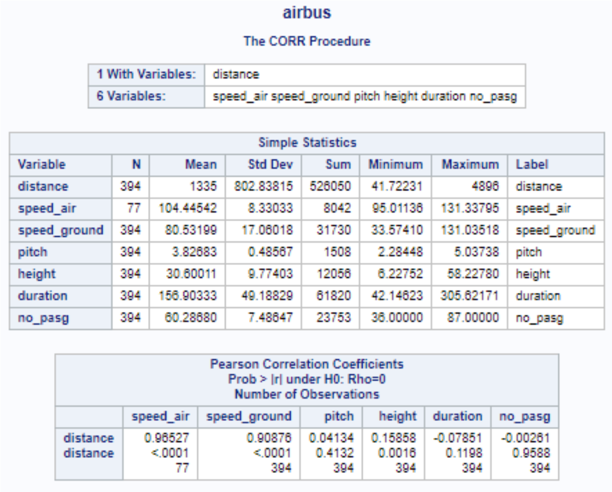
Let try individual analysis by aircraft to be more sure about the correlation.

Part B Analysis:

SAS code:

```
data airbus;
set FAA_cleaned (where=(aircraft='airbus')); run;
PROC CORR data = airbus;
TITLE "airbus";
VAR speed_air speed_ground pitch height duration no_pasg;
WITH distance;
data boeing (where=(aircraft='boeing'));
set FAA_cleaned; run;
PROC CORR data = boeing;
TITLE "boeing";
VAR speed_air speed_ground pitch height duration no_pasg;
WITH distance;
```

Output Snippet:



airbus

The CORR Procedure

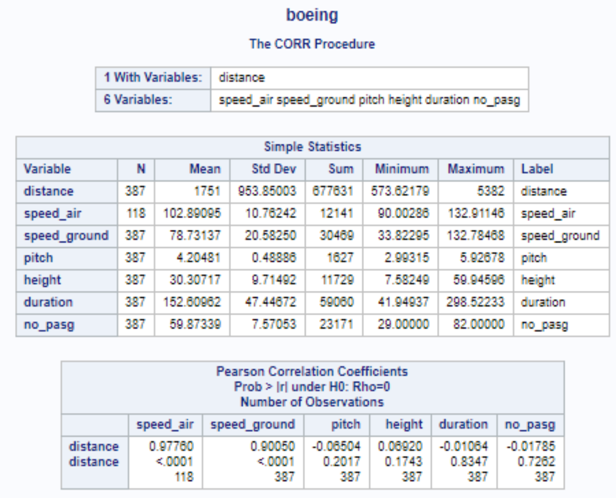
1 With Variables:	distance
6 Variables:	speed_air speed_ground pitch height duration no_pasg

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	394	1335	802.83815	526050	41.72231	4896	distance
speed_air	77	104.44542	8.33033	8042	95.01136	131.33795	speed_air
speed_ground	394	80.53199	17.06018	31730	33.57410	131.03518	speed_ground
pitch	394	3.82683	0.48567	1508	2.28448	5.03738	pitch
height	394	30.60011	9.77403	12056	6.22752	58.22780	height
duration	394	156.90333	49.18829	61820	42.14623	305.62171	duration
no_pasg	394	60.28680	7.48647	23753	36.00000	87.00000	no_pasg

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

	speed_air	speed_ground	pitch	height	duration	no_pasg
distance distance	0.96527 <.0001 77	0.90876 <.0001 394	0.04134 0.4132 394	0.15858 0.0016 394	-0.07851 0.1198 394	-0.00261 0.9588 394



boeing

The CORR Procedure

1 With Variables:	distance
6 Variables:	speed_air speed_ground pitch height duration no_pasg

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	387	1751	953.85003	677631	573.62179	5382	distance
speed_air	118	102.89095	10.76242	12141	90.00286	132.91146	speed_air
speed_ground	387	78.73137	20.58250	30469	33.82295	132.78468	speed_ground
pitch	387	4.20481	0.48886	1627	2.99315	5.92678	pitch
height	387	30.30717	9.71492	11729	7.58249	59.94596	height
duration	387	152.60962	47.44672	59060	41.94937	298.52233	duration
no_pasg	387	59.87339	7.57053	23171	29.00000	82.00000	no_pasg

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

	speed_air	speed_ground	pitch	height	duration	no_pasg
distance distance	0.97760 <.0001 118	0.90050 <.0001 387	-0.06504 0.2017 387	0.06920 0.1743 387	-0.01064 0.8347 387	-0.01785 0.7262 387

Following points can be inferred from the correlation matrices for the aircraft ‘airbus’ and ‘boeing’

- duration and no_pasg still shows negligible negative correlation thus should be not considered in the model
- ‘pitch’ shows a positive correlation for ‘airbus’ but shows a negative correlation for ‘boeing’, thus ‘pitch’ is directly affected by change of aircraft.
- ‘speed_air’ and ‘speed_ground’ still shows a high positive correlation.
- ‘height’ shows a positive correlation of 0.15858 and 0.06920 but the magnitude is not that significant as compared to ‘speed_air’

CHAPTER 3

STATISTICAL ANALYSIS

Correlation Matrix gave various insights about the variables which can be included in the model.

Both ‘speed_ground’ and ‘speed_air’ displays high correlation but ‘speed_air’ will be used in regression model as ‘speed_air’ has the highest correlation and its distribution follows a linear model to a great extent.

Assuming landing distance can be predicted by a linear function of a speed_air.

$\mathbf{Y = \beta_o + \beta_1 X + \varepsilon}$, where

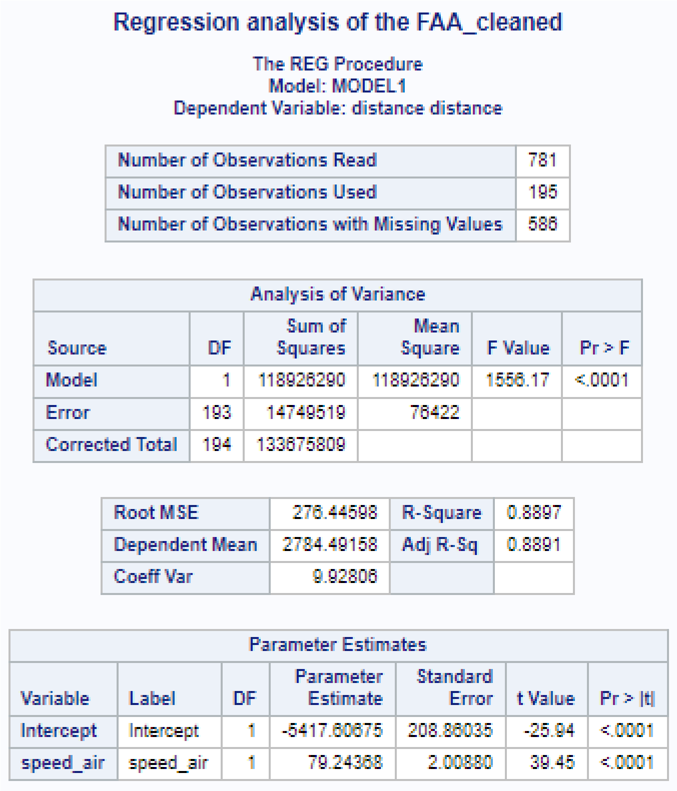
β_o, β_1 = unknown parameters, more specifically β_o = intercept , β_1 = slope, Y = landing distance, X = speed_air and ε is error term.

SAS code:

```
proc reg data=FAA_cleaned;

model distance=speed_air;

title Regression analysis of the FAA_cleaned;

run;
```

Output Snippet:

Part A Analysis:



Regression analysis of the FAA_cleaned

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	781
Number of Observations Used	195
Number of Observations with Missing Values	586

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	118926290	118926290	1556.17	<.0001
Error	193	14749519	76422		
Corrected Total	194	133675809			

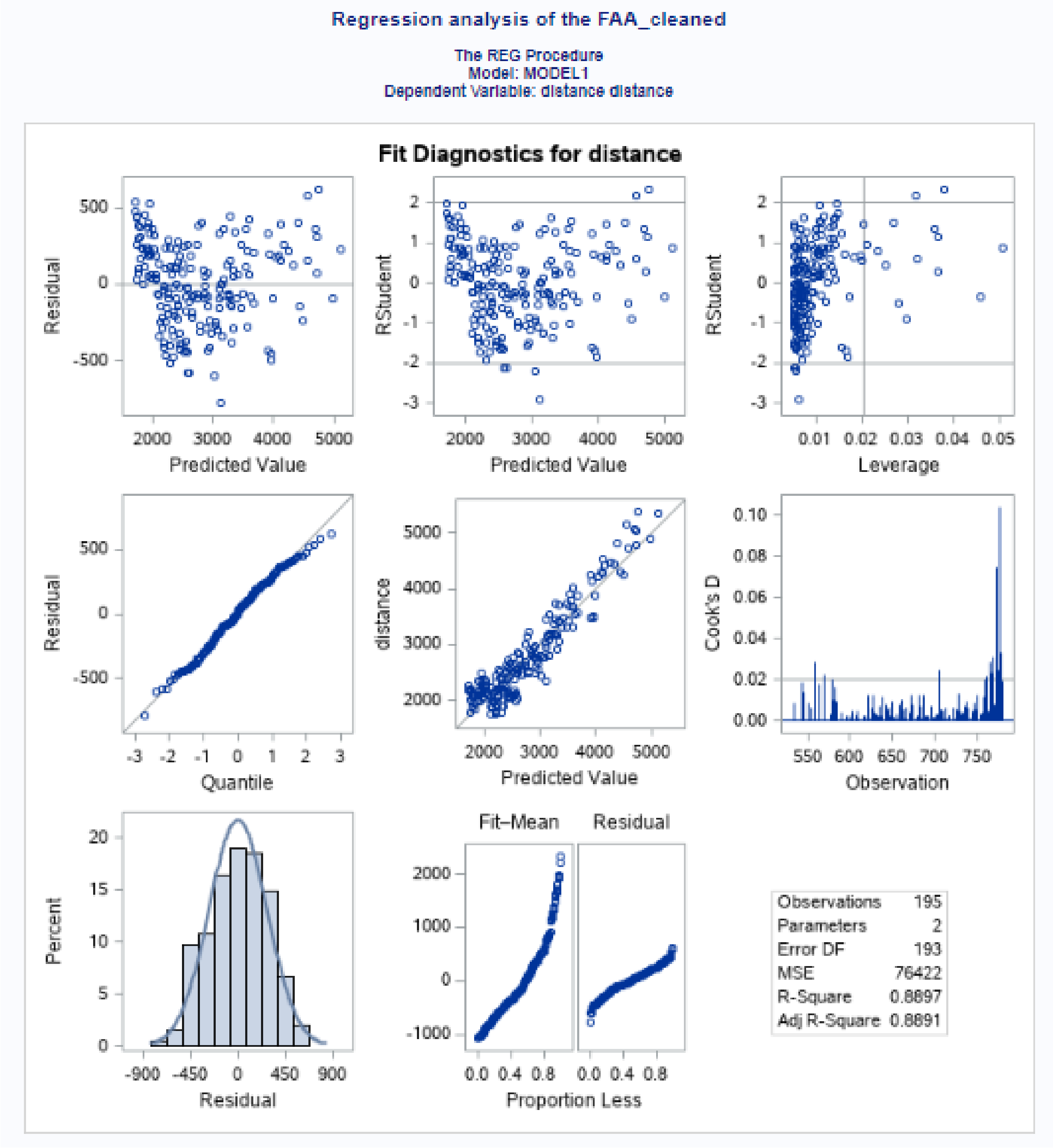
Root MSE	276.44598	R-Square	0.8897
Dependent Mean	2784.49158	Adj R-Sq	0.8891
Coeff Var	9.92806		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5417.60675	208.86035	-25.94	<.0001
speed_air	speed_air	1	79.24368	2.00880	39.45	<.0001

- The F statistic for the overall model is highly significant (=1556.17, <0.0001), indicating that the model explains a significant portion of the variation in the data. This model estimates two parameters, β_o and β_1; thus, the degrees of freedom are 1. The corrected total is 194.
- The "Parameter Estimates" table contains the estimates of β_o and β_1. The table also contains the *t* statistics and the corresponding *p*-values for testing whether each parameter is significantly different from zero. The *p*-values (*t*=-25.94, *p*<0.0001 and *t*=39.45, *p*<0.0001) indicate that the intercept and speed_air parameter estimates, respectively, are highly significant.

Part B Analysis:

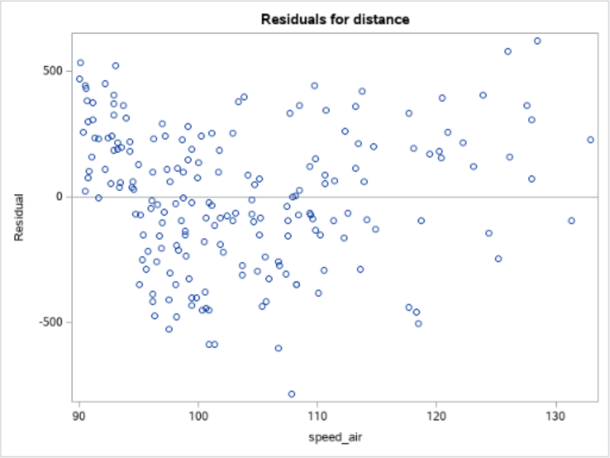


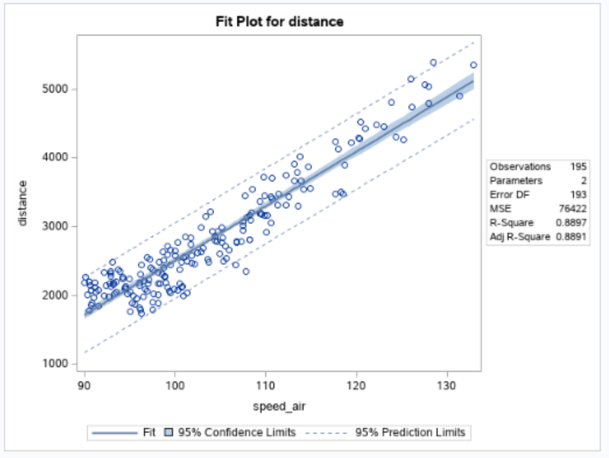
Regression analysis of the FAA_cleaned

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Fit Diagnostics for distance

Observations	195
Parameters	2
Error DF	193
MSE	76422
R-Square	0.8897
Adj R-Square	0.8891





Observations	195
Parameters	2
Error DF	193
MSE	76422
R-Square	0.8897
Adj R-Square	0.8891

- A trend in the residuals generally indicates non-constant variance in data. Since these residuals have no apparent trend, the analysis can be considered acceptable.
- The R-square and Adj R-square are used in assessing the fit of the model; values close to 1 indicate a better fit. The R-square of 0.8897 indicates that ‘speed_air’ accounts for 88% of the variation in ‘distance’.

The fit plot of distance shows that the relation is linear. From the parameter estimates, the fitted model:

distance = -5417.6 + 79.2 * speed_air

QUESTIONS AND ANSWERS

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

The final model includes 195 observations. Data cleaning removed abnormal values and duplicate records and filtered down to 781 records. All the observations were not used as it would result in incorrect results. Also, the correlation matrix showed that speed_air has the highest correlation with a value of 0.94322. Thus, speed_air with 195 records was included to generate the best fit model.

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

The landing distance is affected by different factors in different ways:

- speed_air: Affects the landing distance directly. Increase in speed increases the landing distance
- pitch: Affects the landing distance both directly and inversely but with very less magnitude. It varies depending on the aircraft make.
- height: Affects the landing distance directly but with very less magnitude. Increase in height slightly increases the landing distance.
- duration,no_pasg: They do not affect the landing distance significantly. Hence they can be ignored.
- speed_ground: Affects in the same way as speed_air but with less magnitude.

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

During data analysis, certain variables were changing their direction w.r.t aircraft make. 'Pitch' was one such factor, it showed a positive correlation of 0.04134 for 'airbus' but shows negative correlation of -0.06504 for 'boeing' w.r.t landing distance. Thus difference between the two makes Airbus and Boeing definitely affects certain variables in the way they relate to landing distance.