BANA 6043-STATISTICAL COMPUTING

Project: Statistical Analysis to Reduce landing Overrun

Niharika Gupta- M13437287
Carl H Lindner School of Business: University of Cincinnati

Abstract: This report studies the factors that are involved in the landing distance of a commercial aircrafts with the inspiration to decrease the danger of flight landing overrun. The data analyzed is (isolated in two Excel records 'FAA-1.xls' and 'FAA-2.xls') containing landing information from 950 simulated commercial flights. Underneath it is an outline of the factors of each flight and the detailed report analyzes each factor in detail. This report is divided into five chapters covering various aspects to statistical analysis methods performed on our data sets such as: data exploration and data cleaning, descriptive study of variables, statistical modeling and model diagnostics, model validations, and remodeling. A portion of the procedures applied to the investigation are information cleaning systems, for example, consolidating records from various sources, performing legitimacy and fulfillment checks of the factors and end of copies and missing qualities. Other further developed methods that have been applied are the utilization of plots and correlation analysis of the variables, and linear regression modeling. The outcome of this report is to predict the factors involved and their degree of involvement in flight landing distance to come up with a mathematical equation that can be used by commercial in-flight software's which will warn pilots before landing if their flight has a risk of landing overrun and giving them ample time to make necessary adjustments.

Statistical Computing (BANA 6043 Project)

Gupta2na Niharika Gupta M13437287

Summary

We were provided 2 datasets with 800 and 150 observation each and the combined dataset had a total of 1000 rows and 8 columns because the combined dataset had 50 rows that were completely null. However, I used only 831 observation to complete the study because I removed were abnormal values (outliers), null rows and duplicate values. There were 100 duplicate rows and 19 abnormal values. After cleaning the dataset, I decided to deal with the abnormal values. There were 638 missing values for speed air and 50 for durations. I checked the co-relation between different variables in order to make decision about imputation. Because of strong co-relation between speed ground and speed distance I decided to calculate missing speed air values using the available speed ground values. I did not proceed with the imputation for the missing duration values as I didn't see any association between duration and distance. I then performed the univariate analysis to study the distribution of my variables and then studied landing distance with different variables using the Proc Chart and Proc Plot method which showed an association between distance and speed air, speed ground both, consistent with our prior findings with co-relation coefficients. I also noticed that landing distance for 'Boeing' aircraft is generally higher than 'Airbus' and I confirmed these findings using Proc T-test.

I again did the co-relation analysis after imputation and observed an even higher co-relation between speed air and speed ground obviously because speed air values were calculated from speed ground values. Next, in order to come up with a perfect model for my study, I performed regression analysis of our dependent variable distance with the independent variables. Based on the P values and parameter estimates, I kept removing variables till my model seemed fit for the study. My final model consisted of speed ground, height and type of aircraft. Next, I proceeded with the validation of my model and kept validating and remodeling until I came up with a model with R square value as high as 0.9669. My final model had only three variable (speed ground, height and type of aircraft) and these variables have positive co-relation with landing distance. I concluded my study after deriving a formula for calculating Landing distance using the values of these 3 variables.

Variables

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.

Contents

Statistical Computing (BANA 6043 Project)

Gupta2na **Niharika Gupta** **M13437287**

CHAPTER 1: DATA EXPLORATION AND DATA CLEANING

***Goal**: Importing the given datasets, combining them and exploring to check for outliers, missing values and duplicates and finally acting upon them accordingly.*

STEP 1: IMPORTING DATA FILES

Data sets were imported into SAS studio using the below code.

Observation:

Statistical Computing (BANA 6043 Project)

Gupta2na Niharika Gupta M13437287

- First data set has 800 rows and 8 columns and the second one has 200 rows (150 observations, 50 null rows) and 7 columns.

```
/*IMPORTING FIRST DATA SET*/

FILENAME REFFILE '/folders/myfolders/sasuser.v94/FAA1.xls';

PROC IMPORT DATAFILE=REFFILE
    DBMS=XLS
    OUT=WORK.FAA1
    REPLACE;
    GETNAMES=YES;
RUN;
PROC CONTENTS DATA=WORK.FAA1;
RUN;

/*IMPORTING SECOND DATASET*/

FILENAME REFFILE '/folders/myfolders/sasuser.v94/FAA2.xls';

PROC IMPORT DATAFILE=REFFILE
    DBMS=XLS
    OUT=WORK.FAA2
    REPLACE;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.FAA2; RUN;
```

Table: WORK.FAA1 | View: Column names | Filter: (none)

Total rows: 800 Total columns: 8 Rows 1-100

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152
6	boeing	137.59581722	55	75.014343744	.	41.21496259	4.203853398
7	boeing	73.023794916	54	54.4298029	.	24.03532163	3.8376457299
8	boeing	52.903187872	57	57.101661737	.	19.388837508	4.6436717769
9	boeing	155.51861605	61	85.443624251	.	35.375389749	4.2287278648
10	boeing	176.86203205	56	61.796710514	.	36.748816124	4.1843990127
11	boeing	158.4618984	61	53.778126741	.	46.355832902	5.5563991716

The CONTENTS Procedure

Data Set Name	WORK.FAA1	Observations	800
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	09/14/2019 09:36:06	Observation Length	72
Last Modified	09/14/2019 09:36:06	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)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	908
Obs in First Data Page	800
Number of Data Set Repairs	0
Filename	/tmp/SAS_work4EAE00000967_localhost.localdomain/SAS_work616E00000967_localhost.localdomain/faa1.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	671642
Access Permission	rw-rw-r--
Owner Name	sasdemo
File Size	128KB
File Size (bytes)	131072

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

The CONTENTS Procedure

Data Set Name	WORK.FAA2	Observations	200
Member Type	DATA	Variables	7
Engine	V9	Indexes	0
Created	09/14/2019 09:38:22	Observation Length	64
Last Modified	09/14/2019 09:38:22	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)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1021
Obs in First Data Page	200
Number of Data Set Repairs	0
Filename	/tmp/SAS_work4EAE00000967_localhost.localdomain/SAS_work616E00000967_localhost.localdomain/faa2.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	671648
Access Permission	rw-rw-r--
Owner Name	sasdemo
File Size	128KB
File Size (bytes)	131072

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	12	$12.	$12.	aircraft
7	distance	Num	8	BEST12.		distance
5	height	Num	8	BEST12.		height
2	no_pasg	Num	8	BEST12.		no_pasg

Table: WORK.FAA2 View: Column names Filter: (none)

Total rows: 200 Total columns: 7 Rows 1-100

	aircraft	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
5	boeing	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
6	boeing	55	75.014343744	.	41.21496259	4.203853398	1627.0681991
7	boeing	54	54.4298029	.	24.03532163	3.8376457299	805.30399317
8	boeing	57	57.101661737	.	19.388837508	4.6436717769	573.62178606
9	boeing	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548
10	boeing	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579
11	boeing	61	53.778126741	.	46.355832902	5.5563991716	1075.3717411
12	boeing	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506
13	boeing	54	93.391762435	92.869561214	32.223489271	3.8182761471	2128.708285
14	boeing	58	94.036412942	96.196460585	33.661226156	4.6361847249	2304.857574
15	boeing	63	63.540613553	.	26.402991875	3.8566584986	1089.9729531

STEP 2: COMBINING DATA SETS

I have used both Concatenation and Interleaving but I will be working on dataset combined using Concatenating.

1. CONCATENATING DATA SETS

Observation: The resulting data set has 1000 rows and 8 columns.

```
/*CONCATENATING DATA SETS*/
DATA COMBINED;
    SET FAA1 FAA2;
RUN;
PROC PRINT DATA=COMBINED;
RUN;
```

CODE LOG RESULTS OUTPUT DATA

Table: WORK.COMBINED View: Column names Filter: (none)

Total rows: 1000 Total columns: 8 Rows 1-100

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152
6	boeing	137.59581722	55	75.014343744	.	41.21496259	4.203853398
7	boeing	73.023794916	54	54.4298029	.	24.03532163	3.8376457299
8	boeing	52.903187872	57	57.101661737	.	19.388837508	4.6436717769
9	boeing	155.51861605	61	85.443624251	.	35.375389749	4.2287278648
10	boeing	176.86203205	56	61.796710514	.	36.748816124	4.1843990127
11	boeing	158.4618984	61	53.778126741	.	46.355832902	5.5563991716
12	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511
13	boeing	72.289633216	54	93.391762435	92.869561214	32.223489271	3.8182761471
14	boeing	187.59954737	58	94.036412942	96.196460585	33.661226156	4.6361847249
15	boeing	154.36870049	63	63.540613553	.	26.402991875	3.8566584986
16	boeing	165.54194536	69	48.774673273	.	31.228664837	3.9020460339
17	boeing	153.54633587	61	83.556493271	.	29.897473262	3.519783726
18	boeing	107.11331938	78	86.807962025	.	25.477015381	4.4142187986

2. INTERLEAVING DATA SETS BASED ON LANDING DISTANCE

Data sets are first sorted by distance and then interleaving is performed.

Observation:

- First 50 rows are blank which must be the 50 missing rows that we observed in the second dataset while importing. We will remove these rows before handling other missing values and abnormal data. First we need to verify these missing rows.

Statistical Computing (BANA 6043 Project)

Gupta2na Niharika Gupta M13437287

```
/*INTERLEAVING DATA SETS*/

PROC SORT data=FAA1;
BY distance;

PROC SORT data=FAA2;
BY distance;

DATA COMBINED2;
SET FAA1 FAA2;
BY distance;

PROC PRINT DATA=COMBINED;
RUN;
```

39		.	.	.	.	.	.	.
40		.	.	.	.	.	.	.
41		.	.	.	.	.	.	.
42		.	.	.	.	.	.	.
43		.	.	.	.	.	.	.
44		.	.	.	.	.	.	.
45		.	.	.	.	.	.	.
46		.	.	.	.	.	.	.
47		.	.	.	.	.	.	.
48		.	.	.	.	.	.	.
49		.	.	.	.	.	.	.
50		.	.	.	.	.	.	.
51	airbus	132.46942492	80	100.01055305	100.891677	41.033010684	4.2975016214	2554.8330623
52	airbus	109.19713407	43	82.483044979	.	30.140024889	4.0896284195	1321.0000654
53	airbus	93.952926911	58	96.878686347	98.085883143	29.178095121	3.967524021	2008.2207232
54	airbus	45.635423091	60	93.793862117	.	42.830935448	4.271324799	2003.4386496
55	airbus	99.148062915	63	97.096913917	96.913737767	33.144245658	3.5162975656	2060.1694249
56	airbus	199.43713308	69	58.10907688	.	24.20102213	3.6341657268	418.01946274
57	airbus	141.96833358	63	85.849362338	.	46.468182053	3.4015527555	1492.6717204
58	airbus	203.13135186	55	90.261004686	.	42.318923044	3.2745717105	1446.8557482
59	airbus	199.79840009	55	61.712054098	.	34.976545351	4.0805575423	643.85634155
60	airbus	112.87149908	60	104.45540038	103.6715358	23.783587114	3.9026553246	2488.9984842
61	airbus	148.49500413	53	99.874522521	98.724063607	39.520425649	3.9041206536	2404.7430929
62	airbus	89.075548734	57	74.212201979	.	25.747432194	3.6751924109	852.77611439
63	airbus	109.79101574	50	85.88079228	.	33.466314922	3.462927709	1408.5685921
64	airbus	209.19366153	54	50.812930767	.	38.841316346	4.0338980996	566.92692802

Table: WORK.COMBINED2 View: Column names Filter: (none)

Total rows: 1000 Total columns: 8 Rows 1-100

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch
1		.	.	.	.	.	.
2		.	.	.	.	.	.
3		.	.	.	.	.	.
4		.	.	.	.	.	.
5		.	.	.	.	.	.
6		.	.	.	.	.	.
7		.	.	.	.	.	.
8		.	.	.	.	.	.
9		.	.	.	.	.	.
10		.	.	.	.	.	.
11		.	.	.	.	.	.
12		.	.	.	.	.	.

STEP 3: VERIFYING NULL ROWS

Null rows are verified below.

Observation:

- The results confirm that there are 50 null rows.
- There are other missing values too, but these will be handled later.

```
/* Determining missing values*/

Proc Sort Data = COMBINED;
    by aircraft;

PROC MEANS DATA=COMBINED NMISS;
    by aircraft;
RUN;
```

The MEANS Procedure

aircraft=' '

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

aircraft=airbus

Variable	Label	N Miss
duration	duration	50
no_pasg	no_pasg	0
speed_ground	speed_ground	0
speed_air	speed_air	364
height	height	0
pitch	pitch	0
distance	distance	0

aircraft=boeing

Variable	Label	N Miss
duration	duration	100
no_pasg	no_pasg	0
speed_ground	speed_ground	0
speed_air	speed_air	347
height	height	0
pitch	pitch	0
distance	distance	0

Statistical Computing (BANA 6043 Project)

Gupta2na Niharika Gupta M13437287

STEP 4: HANDLING NULL ROWS

Observation:

- A total of 950 rows remain after removing the null rows.

```
/*Removing null rows*/
DATA CLEAN;
    SET COMBINED;
    IF aircraft='' Then delete;
PROC PRINT DATA=CLEAN;
RUN;
```

Table: WORK.CLEAN View: Column names Filter: (none)

Total rows: 950 Total columns: 8 Rows 1-100

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitc
1	airbus	132.46942492	80	100.01055305	100.891677	41.033010684	4.29750162
2	airbus	109.19713407	43	82.483044979	.	30.140024889	4.08962841
3	airbus	93.952926911	58	96.878686347	98.085883143	29.178095121	3.9675240
4	airbus	45.635423091	60	93.793862117	.	42.830935448	4.2713247
5	airbus	99.148062915	63	97.096913917	96.913737767	33.144245658	3.51629756
6	airbus	199.43713308	69	58.10907688	.	24.20102213	3.63416572
7	airbus	141.96833358	63	85.849362338	.	46.468182053	3.40155275
8	airbus	203.13135186	55	90.261004686	.	42.318923044	3.27457171
9	airbus	199.79840009	55	61.712054098	.	34.976545351	4.08055754
10	airbus	112.87149908	60	104.45540038	103.6715358	23.783587114	3.90265532

STEP 5 : CHECKING FOR DUPLICATES

We need to check if there are any duplicate rows. In order to do that, dataset is first sorted and then duplicates are checked.

Observation:

- There are around 100 duplicate rows as the output shows a total of 850 rows.

```
/* DATA SET IS FIRST SORTED*/
proc sort data=CLEAN OUT=CLEANSORTED;
BY aircraft descending duration distance height no_pasg pitch speed_air speed_ground;
RUN;


/*CHECKING FOR DUPLICATES*/

PROC FREQ data=CLEANSORTED;
    TABLES aircraft*height*no_pasg*pitch*speed_air* speed_ground*distance/ noprint out=keylist;
RUN;
PROC PRINT;
    WHERE count ge 2;
RUN;
```

Statistical Computing (BANA 6043 Project)

Gupta2na **Niharika Gupta** **M13437287**

521	boeing	19.97791367	65	4.2333790632	.	70.256718159	1160.8697017	2	.
523	boeing	20.238480104	61	3.9437122363	.	76.529795789	1379.260881	2	.
524	boeing	20.343993391	57	3.8794887756	.	73.313432826	1217.3211669	2	.
528	boeing	20.783114081	59	3.7320281536	.	56.441314363	731.33170479	2	.
529	boeing	20.801467416	55	4.0098154916	.	63.193085808	829.07888581	2	.
540	boeing	21.772286622	78	4.5955283685	.	61.220375598	970.04651856	2	.
546	boeing	22.411979234	52	3.702074231	.	81.533090888	1587.3880099	2	.
554	boeing	23.349901124	61	4.3961881217	.	29.227656382	1076.855217	2	.
559	boeing	23.575935009	54	5.2168022511	141.72493569	141.21863535	6533.0476506	2	0.83682
562	boeing	23.839448756	62	3.3959225955	.	79.745315854	1322.2129905	2	.
564	boeing	23.936206647	60	3.227014688	.	59.092128964	987.55977466	2	.
566	boeing	24.03532163	54	3.8376457299	.	54.4298029	805.30399317	2	.
568	boeing	24.531203641	61	4.3804647597	96.033397859	98.206172066	2149.5567186	2	0.83682
580	boeing	25.477015381	78	4.4142187986	.	86.807962025	1910.8768699	2	.
582	boeing	25.684334148	54	4.4463077911	.	83.655266858	1470.7842588	2	.
589	boeing	26.223285332	57	4.2231807894	.	72.327130778	1105.3658522	2	.
594	boeing	26.402991875	63	3.8566584986	.	63.540613553	1089.9729531	2	.
595	boeing	26.470877122	60	3.8504358681	93.18623585	93.89933526	2157.4188427	2	0.83682
604	boeing	27.0634843	66	4.4386388552	.	64.946988369	920.28626467	2	.
611	boeing	27.418924252	53	4.0435145715	109.32837648	107.91568005	3369.8363638	2	0.83682
618	boeing	27.804716181	69	4.1174316991	102.8514051	101.65558863	2987.8039235	2	0.83682
620	boeing	27.862901056	63	3.6862217761	.	67.969699235	1237.473458	2	.
624	boeing	28.036760187	57	4.6915095248	.	47.883210688	745.65270838	2	.
626	boeing	28.152991316	53	3.9874712191	99.135830727	98.180410862	2586.6650864	2	0.83682
627	boeing	28.206547549	60	3.8902037695	.	49.168445223	662.1715576	2	.
634	boeing	28.406673108	63	3.9378640453	.	63.57042961	1032.4646189	2	.
636	boeing	28.543870421	71	4.6031173258	.	41.455692355	763.699563	2	.
640	boeing	28.773729478	62	3.3058880775	92.874851912	91.714535792	2313.3356963	2	0.83682
644	boeing	28.972879292	67	3.8124590751	.	78.954218512	1480.5722537	2	.
645	boeing	29.152465311	63	4.0140064257	.	73.533976336	1332.0387485	2	.
648	boeing	29.366866101	55	4.4071812572	.	79.994815042	1515.9652753	2	.
652	boeing	29.620479559	60	5.2258651627	.	63.238252725	832.51894862	2	.
655	boeing	29.897473262	61	3.519783726	.	83.556493271	1793.5628232	2	.
659	boeing	30.216568242	67	3.2137033407	123.88257287	122.75656197	4807.8798069	2	0.83682
660	boeing	30.270100189	47	3.6451345759	94.215160768	95.322576422	2233.0489624	2	0.83682
662	boeing	30.305425419	49	3.9341591214	.	57.65125066	981.8893648	2	.
666	boeing	30.598791007	67	4.2369678091	.	76.531843725	1236.3788062	2	.

Table: WORK.KEYLIST View: Column names Filter: (none)

Total rows: 850 Total columns: 9 Rows 1-100

	aircraft	height	no_pasg	pitch	speed_air	speed_ground	distanc
1	airbus	-3.332387973	73	4.8305592948	.	92.994942381	1567.66572
2	airbus	-2.915335901	58	3.1225583646	.	66.421119468	34.0807832
3	airbus	-0.067758596	68	4.6928768405	.	56.497986661	380.362981
4	airbus	0.086105484	62	3.6220566648	.	72.028024252	537.919581
5	airbus	6.2275177523	59	3.778378927	.	85.687131792	976.433685
6	airbus	8.559069177	66	3.9134477851	.	51.158228388	242.595886
7	airbus	9.1646259705	53	3.1601680959	.	85.943555546	1107.69707
8	airbus	9.688307724	52	3.3585464091	.	73.761115944	554.160987
9	airbus	9.6972160002	73	3.1724565811	107.78665842	104.70010613	2340.58279
10	airbus	10.099990802	51	4.4654606398	110.56964767	109.78499325	3054.41240
11	airbus	10.753780476	62	4.1132082314	.	86.239504411	1158.58917
12	airbus	11.198044368	66	4.2638480897	.	93.918984258	1504.74588
13	airbus	11.238021337	66	3.6038718062	.	63.817889502	370.470741
14	airbus	11.472293317	69	3.2632435552	.	76.614822894	636.741375
15	airbus	11.516139597	56	2.7990204762	.	74.337393872	485.194727
16	airbus	11.671763822	64	3.3398414652	.	84.34088335	1085.48413
17	airbus	11.716588101	62	3.8585210867	.	85.02644786	1287.7637
18	airbus	12.315795866	55	3.7213188282	.	81.69133242	829.186392
19	airbus	12.353622855	62	4.4082857766	.	77.729568204	821.019472
20	airbus	12.878217378	56	3.6021593263	.	82.093109962	899.588606

STEP 6: REMOVAL OF DUPLICATES

Duplicates are removed using the below code.

```
/*REMOVAL OF DUPLICATE VALUES*/

proc sort data=CLEANSORTED nodupkey out=NO_DUP;
by aircraft distance height no_pasg pitch speed_air speed_ground;
run;
```

CODE LOG RESULTS OUTPUT DATA

Table: WORK.NO_DUP View: Column names Filter: (none)

Total rows: 850 Total columns: 8 Rows 1-100

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch
1	airbus	150.94674427	58	66.421119468	.	-2.915335901	3.1225583646
2	airbus	190.7394255	77	47.882117055	.	14.835964361	2.7322842836
3	airbus	212.05403613	63	51.587044527	.	20.451285811	3.063686215
4	airbus	128.37336566	64	55.461625107	.	14.65127605	3.9792117538
5	airbus	237.40527671	48	53.774013118	.	28.260802216	3.1755295986
6	airbus	142.5876457	66	51.158228388	.	8.559069177	3.9134477851
7	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162
8	airbus	230.32398183	58	55.108631792	.	29.859498104	3.2599541617
9	airbus	175.53311361	61	65.037084787	.	13.807590435	3.4948549953
10	airbus	182.44776305	66	52.70784152	.	24.302641153	4.1859666088
11	airbus	214.78506113	59	56.285225619	.	19.097947487	3.9151278527
12	airbus	183.61849925	69	53.539242523	.	31.739422907	3.5237749131
13	airbus	261.38701422	68	57.08470944	.	15.761691561	3.8049960732
14	airbus	149.20859096	66	63.817889502	.	11.238021337	3.6038718062
15	airbus	208.87900668	56	53.888762395	.	23.496668191	4.0000159209
16	airbus	98.176296764	60	53.749134607	.	25.54578925	3.7142005991

Observation:

- A total of 850 rows and 8 columns remain after removal of duplicate values.

STEP 7: CHECKING FOR ABNORMAL VALUES (OUTLIERS)

The criteria for normal and abnormal values is listed in the variable description of the data. We need to determine if there are any abnormal values as these will be the outliers for our study.

```
/*CHECKING FOR ABNORMAL VALUES*/

DATA ABNORMAL;
    SET NO_DUP;
IF duration<= 40 and duration ^= . then DUR='ABNORMAL';else DUR='NORMAL';
IF speed_ground<30 OR speed_ground>140 then SPEED_G='ABNORMAL'; else SPEED_G='NORMAL';
IF (speed_air<30 OR speed_air>140) and speed_air ^= .  then SPEED_A='ABNORMAL'; else SPEED_A='NORMAL';
IF height<6 then HGHT='ABNORMAL'; else HGHT='NORMAL';
IF distance>6000 then LD='ABNORMAL';ELSE LD='NORMAL';
RUN;

proc print data=abnormal;
run;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	DUR	SPEED_G	SPEED_A	HGHT	LD
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286	1144.922426	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
6	boeing	137.59581722	55	75.014343744	.	41.21496259	4.203853398	1627.0681991	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
7	boeing	73.023794916	54	54.4298029	.	24.03532163	3.8376457299	805.30399317	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
8	boeing	52.903187872	57	57.101661737	.	19.388837508	4.6436717769	573.62178606	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
9	boeing	155.51861605	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
10	boeing	176.86203205	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
11	boeing	158.4618984	61	53.778126741	.	46.355832902	5.5563991716	1075.3717411	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
12	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506	NORMAL	ABNORMAL	ABNORMAL	NORMAL	ABNORMAL
13	boeing	72.289633216	54	93.391762435	92.869561214	32.223489271	3.8182761471	2128.708285	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
14	boeing	187.59954737	58	94.036412942	96.196460585	33.661226156	4.6361847249	2304.857574	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
15	boeing	154.36870049	63	63.540613553	.	26.402991875	3.8566584986	1089.9729531	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
16	boeing	165.54194536	69	48.774673273	.	31.228664837	3.9020460339	943.06840443	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
17	boeing	153.54633587	61	83.556493271	.	29.897473262	3.519783726	1793.5628232	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
18	boeing	107.11331938	78	86.807962025	.	25.477015381	4.4142187986	1910.8768699	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
19	boeing	233.80249791	69	104.80843448	103.86845794	43.882731896	3.2450978263	3213.985265	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
20	boeing	163.90650312	55	119.3804635	120.44470797	38.558536007	3.7014493887	4524.2788621	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
21	boeing	97.477623266	63	73.533976336	.	29.152465311	4.0140064257	1332.0387485	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
22	boeing	118.63054039	55	79.994815042	.	29.366866101	4.4071812572	1515.9652753	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
23	boeing	126.54028789	70	94.781230282	91.142068839	39.476298784	3.5949361476	2182.2207374	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
24	boeing	179.91591838	66	63.671165314	.	19.574699606	4.2867337712	873.4408921	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
25	boeing	112.90009528	53	98.180410862	99.135830727	28.152991316	3.9874712191	2586.6650864	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
26	boeing	56.64048966	66	72.953658239	.	36.154157217	4.3878559157	1205.1280251	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
27	boeing	86.828911312	62	91.714535792	92.874851912	28.773729478	3.3058880775	2313.3356963	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
28	boeing	157.35773231	57	72.327130778	.	26.223285332	4.2231807894	1105.3658522	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
29	boeing	186.68141397	49	66.417230464	.	44.692695788	4.1135438115	1176.0276765	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
30	boeing	140.23631155	65	118.74200471	119.40214631	19.856192215	4.6462659602	4217.1294518	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
31	boeing	130.46356358	52	116.71343434	117.65649967	36.195527446	3.8943524297	4240.0941825	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
32	boeing	142.15534911	46	39.769294325	.	39.655921061	4.5992872267	1030.457488	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
33	boeing	155.84557082	62	79.745315854	.	23.839448756	3.3959225955	1322.2129905	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
34	boeing	124.94457133	44	72.546668651	.	42.859879536	4.028501716	1321.1606709	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL

Observation:

- There are some abnormal values, but we need to determine their number in order to decide if we can remove them or not.
- The resulting data set has 13 columns now, with 5 new columns added to indicate normal and abnormal values.

STEP 8: CHECKING FOR THE NUMBER OF OUTLIERS AND STORING THEM IN A SEPARATE DATASET

Observation:

- The number of outliers is just 19 which is not too high thus we can remove them from the data set that we want to work on.
- We will however save them in a different dataset before deleting them from our main data set.

Statistical Computing (BANA 6043 Project)

Gupta2na Niharika Gupta M13437287

```
/*CHECKING FOR THE COUNT OF ABNORMAL VALUES AND STORING THEM IN A NEW DATASET*/

DATA ABNORMAL_VALUES;
 SET ABNORMAL;
WHERE DUR="ABNORMAL"OR SPEED_G='ABNORMAL'OR SPEED_G='ABNORMAL' OR SPEED_A='ABNORMAL' OR HGHT='ABNORMAL'OR LD='ABNORMAL';
RUN;

PROC PRINT DATA=ABNORMAL_VALUES;
RUN;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	DUR	SPEED_G	SPEED_A	HGHT	LD
1	airbus	150.94674427	58	66.421119468	.	-2.915335901	3.1225583646	34.080783293	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
2	airbus	157.91497689	68	56.497986661	.	-0.067758596	4.6928768405	380.36298195	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
3	airbus	163.52364053	62	72.028024252	.	0.086105484	3.6220566648	537.91958189	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
4	airbus	31.7016661	61	76.354176433	.	30.991021813	2.8173796019	948.47376723	ABNORMAL	NORMAL	NORMAL	NORMAL	NORMAL
5	airbus	103.09084673	73	92.994942381	.	-3.332387973	4.8305592948	1567.6657219	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
6	airbus	16.893454896	54	94.511052223	95.930926862	37.476967053	4.1733221259	2162.92737	ABNORMAL	NORMAL	NORMAL	NORMAL	NORMAL
7	boeing	133.45985625	73	57.045299494	.	1.2538552556	4.7153842391	371.27726086	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
8	boeing	283.76336844	62	58.889312381	.	4.2644634439	4.7721930401	425.85856098	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
9	boeing	175.08462089	64	52.493139102	.	-3.546252405	4.2132855404	581.38099947	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
10	boeing	124.37864547	72	60.367043725	.	3.7889195211	3.7060888319	641.59956822	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
11	boeing	146.04337112	69	71.787305883	.	-1.528129182	4.1994604645	738.65436932	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
12	boeing	119.64402906	68	70.178463873	.	2.2051944554	3.7397746803	816.20664104	NORMAL	NORMAL	NORMAL	ABNORMAL	NORMAL
13	boeing	17.375513046	63	63.57042961	.	28.406673108	3.9378640453	1032.4646189	ABNORMAL	NORMAL	NORMAL	NORMAL	NORMAL
14	boeing	212.94303494	61	29.227656382	.	23.349901124	4.3961881217	1076.855217	NORMAL	ABNORMAL	NORMAL	NORMAL	NORMAL
15	boeing	141.93411511	46	27.735715303	.	24.400127629	4.3682093233	1323.7157777	NORMAL	ABNORMAL	NORMAL	NORMAL	NORMAL
16	boeing	31.391008253	51	98.219800666	99.057514589	52.473140903	4.1623371208	2808.3151244	ABNORMAL	NORMAL	NORMAL	NORMAL	NORMAL
17	boeing	14.764207145	59	108.29169029	109.32758442	46.930873666	4.8096217396	3645.6110025	ABNORMAL	NORMAL	NORMAL	NORMAL	NORMAL
18	boeing	119.92455279	64	136.65915832	136.42342138	44.286109179	4.1694037368	6309.9459762	NORMAL	NORMAL	NORMAL	NORMAL	ABNORMAL
19	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506	NORMAL	ABNORMAL	ABNORMAL	NORMAL	ABNORMAL

STEP 9: HANDLING OUTLIERS

I removed the abnormal values using the below code.

```
/*REMOVING OUTLIERS*/
DATA NO_OUTLIERS;
    SET ABNORMAL;
IF DUR='ABNORMAL' THEN DELETE;
IF SPEED_G='ABNORMAL' then DELETE;
IF SPEED_A='ABNORMAL' THEN DELETE;
IF HGHT='ABNORMAL' then DELETE;
IF LD='ABNORMAL' then DELETE;
RUN;

PROC PRINT DATA=NO_OUTLIERS;
RUN;
```

Table: WORK.NO_OUTLIERS View: Column names Filter: (none)

Total rows: 831 Total columns: 13 Rows 1-100

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch
1	airbus	190.7394255	77	47.882117055	.	14.835964361	2.7322842836
2	airbus	212.05403613	63	51.587044527	.	20.451285811	3.063686215
3	airbus	128.37336566	64	55.461625107	.	14.65127605	3.9792117538
4	airbus	237.40527671	48	53.774013118	.	28.260802216	3.1755295986
5	airbus	142.5876457	66	51.158228388	.	8.559069177	3.9134477851
6	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162
7	airbus	230.32398183	58	55.108631792	.	29.859498104	3.2599541617
8	airbus	175.53311361	61	65.037084787	.	13.807590435	3.4948549953
9	airbus	182.44776305	66	52.70784152	.	24.302641153	4.1859666088
10	airbus	214.78506113	59	56.285225619	.	19.097947487	3.9151278527
11	airbus	183.61849925	69	53.539242523	.	31.739422907	3.5237749131
12	airbus	261.38701422	68	57.08470944	.	15.761691561	3.8049960732
13	airbus	149.20859096	66	63.817889502	.	11.238021337	3.6038718062
14	airbus	208.87900668	56	53.888762395	.	23.496668191	4.0000159209
15	airbus	98.176296764	60	53.749134607	.	25.54578925	3.7142005991
16	airbus	42.146226159	63	46.264718501	.	20.490711515	3.4819121545
17	airbus	179.99562817	66	57.243835228	.	35.325529113	3.0665057372
18	airbus	198.42235144	58	64.305146799	.	17.456593278	3.8221946371

Observation:

- A total of 831 observations remain after removing outliers.

Step 10 DROPPING THE COLUMNS CREATED FOR ABNORMAL AND NORMAL VALUES

Next, we will drop the new variables ('normal', 'abnormal) that we created since we don't need them.

```
/*DROPPING COLUMNS CREATED FOR ABNORMAL VALUES SINCE WE DON'T NEED THEM ANYMORE*/
DATA FLIGHTS_CLEAN;
SET NO_OUTLIERS;
DROP DUR SPEED_G SPEED_A HGHT LD;
RUN;

PROC PRINT DATA=FLIGHTS_CLEAN;
RUN;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	airbus	190.7394255	77	47.882117055	.	14.835964361	2.7322842836	41.722312733
2	airbus	212.05403613	63	51.587044527	.	20.451285811	3.063686215	133.08690985
3	airbus	128.37336566	64	55.461625107	.	14.65127605	3.9792117538	180.56522534
4	airbus	237.40527671	48	53.774013118	.	28.260802216	3.1755295986	241.16096423
5	airbus	142.5876457	66	51.158228388	.	8.559069177	3.9134477851	242.59588646
6	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162	250.68976141
7	airbus	230.32398183	58	55.108631792	.	29.859498104	3.2599541617	270.83676243
8	airbus	175.53311361	61	65.037084787	.	13.807590435	3.4948549953	280.80440304
9	airbus	182.44776305	66	52.70784152	.	24.302641153	4.1859666088	317.81268659
10	airbus	214.78506113	59	56.285225619	.	19.097947487	3.9151278527	321.51632716
11	airbus	183.61849925	69	53.539242523	.	31.739422907	3.5237749131	349.15851648
12	airbus	261.38701422	68	57.08470944	.	15.761691561	3.8049960732	350.60240534
13	airbus	149.20859096	66	63.817889502	.	11.238021337	3.6038718062	370.47074159
14	airbus	208.87900668	56	53.888762395	.	23.496668191	4.0000159209	375.32596789
15	airbus	98.176296764	60	53.749134607	.	25.54578925	3.7142005991	378.82578267
16	airbus	42.146226159	63	46.264718501	.	20.490711515	3.4819121545	383.55849778
17	airbus	179.99562817	66	57.243835228	.	35.325529113	3.0665057372	383.57772124
18	airbus	198.42235144	58	64.305146799	.	17.456593278	3.8221946371	383.90578116
19	airbus	207.69159848	61	54.542338048	.	19.610845089	3.9730540461	397.01200564

STEP 11 CHECKING FOR OTHER MISSING VALUES AFTER REMOVING ROWS WITH NO AIRCRAFT NAME

OBSERVATIONS:

1. There are 628 missing values for speed_air which accounts for around 75.57 percent of the total observations.
2. There are 50 missing values for duration.

```
/*CHECKING FOR MISSING VALUES*/
proc format;
 value $missfmt ' '='Missing' other='Not Missing';
 value  missfmt  . ='Missing' other='Not Missing';
RUN;

proc freq data=FLIGHTS_CLEAN;
format _CHAR_ $missfmt.;
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;
```

The FREQ Procedure

aircraft	Frequency
Not Missing	831

DUR	Frequency
Not Missing	831

SPEED_G	Frequency
Not Missing	831

SPEED_A	Frequency
Not Missing	831

HGHT	Frequency
Not Missing	831

LD	Frequency
Not Missing	831

duration	Frequency
Missing	50
Not Missing	781

no_pasg	Frequency
Not Missing	831

speed_ground	Frequency
Not Missing	831

speed_air	Frequency
Missing	628
Not Missing	203

height	Frequency
Not Missing	831

pitch	Frequency
Not Missing	831

distance	Frequency
Not Missing	831

STEP 12 DEALING WITH MISSING VALUES

First, I need to check the co-relation between different variables to see how important a variable is for this study and how are the variable related to each other to determine the method of imputation.

I have used the CORR method to determine pairwise co-relation coefficient. Highlighted columns below indicate a strong co-relation between the 2 variables and the following can be concluded:

```
/*finding co-relation between different variables*/

proc corr data=FLIGHTS_CLEAN;
var distance duration height no_pasg pitch speed_air speed_ground;
title Pairwise correlation coefficients;
run;
```

Pairwise correlation coefficients

The CORR Procedure

7 Variables:	distance duration height no_pasg pitch speed_air speed_ground

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	831	1522	896.33815	1265183	41.72231	5382	distance
duration	781	154.77572	48.34992	120880	41.94937	305.62171	duration
height	831	30.45787	9.78481	25310	6.22752	59.94596	height
no_pasg	831	60.05535	7.49132	49906	29.00000	87.00000	no_pasg
pitch	831	4.00516	0.52657	3328	2.28448	5.92678	pitch
speed_air	203	103.48504	9.73628	21007	90.00286	132.91146	speed_air
speed_ground	831	79.54270	18.73568	66100	33.57410	132.78468	speed_ground

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations							
	distance	duration	height	no_pasg	pitch	speed_air	speed_ground
distance distance	1.00000 831	-0.05138 0.1514 781	0.09941 0.0041 831	-0.01776 0.6093 831	0.08703 0.0121 831	0.94210 <.0001 203	0.86624 <.0001 831
duration duration	-0.05138 0.1514 781	1.00000 781	0.01112 0.7564 781	-0.03639 0.3098 781	-0.04675 0.1918 781	0.04454 0.5364 195	-0.04897 0.1716 781
height height	0.09941 0.0041 831	0.01112 0.7564 781	1.00000 831	0.04699 0.1760 831	0.02298 0.5082 831	-0.07933 0.2606 203	-0.05761 0.0970 831
no_pasg no_pasg	-0.01776 0.6093 831	-0.03639 0.3098 781	0.04699 0.1760 831	1.00000 831	-0.01793 0.6057 831	-0.00616 0.9305 203	-0.00013 0.9969 831
pitch pitch	0.08703 0.0121 831	-0.04675 0.1918 781	0.02298 0.5082 831	-0.01793 0.6057 831	1.00000 831	-0.03927 0.5780 203	-0.03912 0.2599 831
speed_air speed_air	0.94210 <.0001 203	0.04454 0.5364 195	-0.07933 0.2606 203	-0.00616 0.9305 203	-0.03927 0.5780 203	1.00000 203	0.98794 <.0001 203
speed_ground speed_ground	0.86624 <.0001 831	-0.04897 0.1716 781	-0.05761 0.0970 831	-0.00013 0.9969 831	-0.03912 0.2599 831	0.98794 <.0001 203	1.00000 831

Observations:

- There is a strong co-relation between ground speed and landing distance (0.86624).
- There is a strong co-relation between air speed and landing distance, thus speed air is an important variable that we need to keep (0.94210).
- There is a strong-co relation between speed air and speed ground(0.98974).
- There is no direct co-relation between distance and duration as circled below. Thus the missing duration values can be left as such.

STEP 12 SUBSTITUTING MISSING SPEED AIR VALUES USING IMPUTATION

Since there are many missing speed_air values, this can cause a great amount of bias, thus I need to replace missing values using imputation method. There are two approaches. I can either substitute the missing values with the mean of the present speed air values. From the dataset, it can be observed that all the speed air values present are greater than 90. This means all the values below 90 are missing and it won't be accurate to use this method to calculate the missing values. Second method is to use regression method for imputation. Since there is a strong co-relation between air speed and ground speed, I can use the ground speed to calculate the missing speed air values.

1. I first created a new variable with a value equal to the difference between speed air and speed ground.

```
/*Difference Analysis of Speed air and ground to imputate missing values*/

data SPEEDAG_DIFF;
set FLIGHTS_CLEAN;
difference= speed_ground-speed_air;
run;
proc print data=SPEEDAG_DIFF;
run;

proc means data=SPEEDAG_DIFF;
var difference;
run;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	airbus	190.7394255	77	47.882117055	.	14.835964361	2.7322842836	41.722312733
2	airbus	212.05403613	63	51.587044527	.	20.451285811	3.063686215	133.08690985
3	airbus	128.37336566	64	55.461625107	.	14.65127605	3.9792117538	180.56522534
4	airbus	237.40527671	48	53.774013118	.	28.260802216	3.1755295986	241.16096423
5	airbus	142.5876457	66	51.158228388	.	8.559069177	3.9134477851	242.59588646
6	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162	250.68976141
7	airbus	230.32398183	58	55.108631792	.	29.859498104	3.2599541617	270.83676243
8	airbus	175.53311361	61	65.037084787	.	13.807590435	3.4948549953	280.80440304
9	airbus	182.44776305	66	52.70784152	.	24.302641153	4.1859666088	317.81268659
10	airbus	214.78506113	59	56.285225619	.	19.097947487	3.9151278527	321.51632716
11	airbus	183.61849925	69	53.539242523	.	31.739422907	3.5237749131	349.15851648
12	airbus	261.38701422	68	57.08470944	.	15.761691561	3.8049960732	350.60240534
13	airbus	149.20859096	66	63.817889502	.	11.238021337	3.6038718062	370.47074159
14	airbus	208.87900668	56	53.888762395	.	23.496668191	4.0000159209	375.32596789
15	airbus	98.176296764	60	53.749134607	.	25.54578925	3.7142005991	378.82578267
16	airbus	42.146226159	63	46.264718501	.	20.490711515	3.4819121545	383.55849778
17	airbus	179.99562817	66	57.243835228	.	35.325529113	3.0665057372	383.57772124
18	airbus	198.42235144	58	64.305146799	.	17.456593278	3.8221946371	383.90578116
19	airbus	207.69159848	61	54.542338048	.	19.610845089	3.9730540461	397.01200564

366	airbus	177.19166302	69	92.808986814	.	29.646414652	4.1643874773	1960.4768127	.
367	airbus	216.87640251	45	91.618595738	.	38.324199382	4.7436314527	1967.6109937	.
368	airbus	162.71456234	54	98.1873349	96.515044653	29.625193292	3.053515269	1975.1105677	1.67229
369	airbus	161.23109432	62	100.63019337	100.87746159	20.288654665	2.8616402098	1990.8532282	-0.24727
370	airbus	45.635423091	60	93.793862117	.	42.830935448	4.271324799	2003.4386496	.
371	airbus	93.952926911	58	96.878686347	98.085883143	29.178095121	3.967524021	2008.2207232	-1.20720
372	airbus	217.12308376	66	94.81425838	97.631341718	33.058365517	3.8235547791	2017.6011486	-2.81708
373	airbus	.	68	99.590180051	97.60461733	42.895159857	3.6363220884	2027.0504154	1.98556
374	airbus	157.57858379	64	101.37496522	101.35383408	21.405235851	3.2399654772	2028.9675819	0.02113
375	airbus	127.30609635	60	99.689595816	99.459199839	28.970863899	4.1071025677	2032.7742548	0.23040
376	airbus	190.79170859	58	93.672281401	95.098821771	33.630978811	3.4592153332	2048.4941733	-1.42654
377	airbus	223.95233137	56	99.325035899	99.397364536	36.782682861	3.1800709689	2059.5377377	-0.07233
378	airbus	99.148062915	63	97.096913917	96.913737767	33.144245658	3.5162975656	2060.1694249	0.18318
379	airbus	108.72206895	64	99.817383459	100.30989703	16.215280593	4.314091631	2080.219501	-0.49251
380	airbus	130.85088326	63	97.379504716	96.729593787	22.499125622	3.8106767874	2092.8564123	0.64991
381	airbus	95.847324908	62	96.922695596	99.836951088	21.418855429	3.8448922111	2093.9965956	-2.91426
382	airbus	139.31381028	44	99.596841547	99.160266345	35.187030092	3.8402667146	2116.080919	0.43658
383	airbus	154.23622614	65	102.87894145	100.66930601	24.347516208	2.9357567095	2119.3135955	2.20964
384	airbus	247.49599004	66	100.75477196	100.83679396	19.028711072	3.1679780428	2123.1470877	-0.08202
385	airbus	161.09215447	64	94.526255342	.	49.239984741	4.5063742177	2134.1932905	.
386	airbus	190.8336252	71	93.79901336	.	45.784826363	3.9702077249	2143.274116	.

2. I then applied the means procedure on the "difference" variable. The results show that the mean of the difference variable is -0.0738829.

```
/*PROC MEANS ON THE NEW VARIABLE "DIFFERENCE"*/
proc means data=SPEEDAG_DIFF;
var difference;
run;
```

The MEANS Procedure

Analysis Variable : difference

N	Mean	Std Dev	Minimum	Maximum
203	-0.0738829	1.5321314	-3.4350609	5.3756363

3. Next I added this mean value to the speed ground to impute missing speed air value

```
/* Imputing missing values of Speed_air*/

data FLIGHTS_IMPUTE;
set FLIGHTS_CLEAN;
if speed_air='.' then speed_air=speed_ground-0.0738829;
run;
proc print data=FLIGHTS_IMPUTE;
run;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	airbus	190.7394255	77	47.882117055	47.808234155	14.835964361	2.7322842836	41.722312733
2	airbus	212.05403613	63	51.587044527	51.513161627	20.451285811	3.063686215	133.08690985
3	airbus	128.37336566	64	55.461625107	55.387742207	14.65127605	3.9792117538	180.56522534
4	airbus	237.40527671	48	53.774013118	53.700130218	28.260802216	3.1755295986	241.16096423
5	airbus	142.5876457	66	51.158228388	51.084345488	8.559069177	3.9134477851	242.59588646
6	airbus	172.04931209	36	47.486765029	47.412882129	13.984809941	4.2990197162	250.68976141
7	airbus	230.32398183	58	55.108631792	55.034748892	29.859498104	3.2599541617	270.83676243
8	airbus	175.53311361	61	65.037084787	64.963201887	13.807590435	3.4948549953	280.80440304
9	airbus	182.44776305	66	52.70784152	52.63395862	24.302641153	4.1859666088	317.81268659
10	airbus	214.78506113	59	56.285225619	56.211342719	19.097947487	3.9151278527	321.51632716
11	airbus	183.61849925	69	53.539242523	53.465359623	31.739422907	3.5237749131	349.15851648
12	airbus	261.38701422	68	57.08470944	57.01082654	15.761691561	3.8049960732	350.60240534
13	airbus	149.20859096	66	63.817889502	63.744006602	11.238021337	3.6038718062	370.47074159
14	airbus	208.87900668	56	53.888762395	53.814879495	23.496668191	4.0000159209	375.32596789
15	airbus	98.176296764	60	53.749134607	53.675251707	25.54578925	3.7142005991	378.82578267
16	airbus	42.146226159	63	46.264718501	46.190835601	20.490711515	3.4819121545	383.55849778
17	airbus	179.99562817	66	57.243835228	57.169952328	35.325529113	3.0665057372	383.57772124
18	airbus	198.42235144	58	64.305146799	64.231263899	17.456593278	3.8221946371	383.90578116
19	airbus	207.69159848	61	54.542338048	54.468455148	19.610845089	3.9730540461	397.01200564
20	airbus	198.8694626	58	61.127025526	61.053142626	22.51924496	3.4452935393	397.54283343
21	airbus	138.23492748	59	64.813671941	64.739789041	14.780336473	4.2209516551	402.93130074
22	airbus	117.7405601	59	47.679801523	47.605918623	28.606492697	3.7520465798	406.08928283
23	airbus	156.4574255	59	72.591244009	72.517361109	14.082627958	3.7352438518	408.97922734
24	airbus	134.685641	57	44.126135466	44.052252566	27.182379063	3.0128294313	417.54307265
25	airbus	199.43713308	69	58.10907688	58.03519398	24.20102213	3.6341657268	418.01946274
26	airbus	153.43737249	61	54.593475453	54.519592553	29.148613678	3.2715631978	424.16175773
27	airbus	142.46940356	66	63.918959665	63.845076765	30.153465712	3.0650903012	428.99182821
28	airbus	199.50282008	72	58.512146812	58.438263912	38.590930739	3.5470466288	432.73222589
29	airbus	162.52084082	62	74.91630489	74.84242199	14.399868335	3.1129303364	433.66000569
30	airbus	117.25925859	61	66.079801141	66.005918241	14.565008188	3.7122960454	436.14783447
31	airbus	129.67260359	68	67.955731399	67.881848499	14.717433562	4.3023776071	436.67877934
32	airbus	248.72910776	58	49.46213365	49.38825075	16.796177769	3.7616893064	452.13126383

Step 13 UNIVARIATE ANALYSIS TO SEE THE DISTRIBUTION OF DIFFERENT VARIABLES

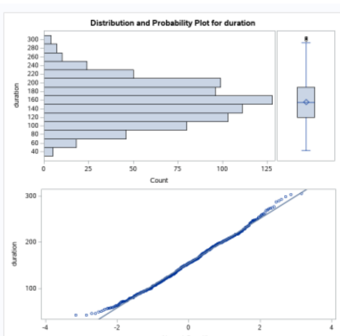
After dealing with both missing and abnormal values, I can start exploring the dataset to learn more about it. I will first apply the univariate analysis to study the distribution of different variables.

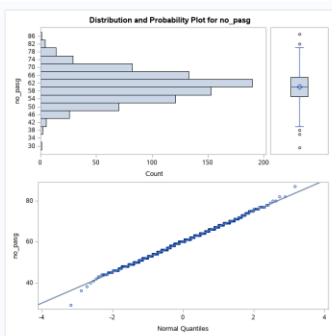
Observations:

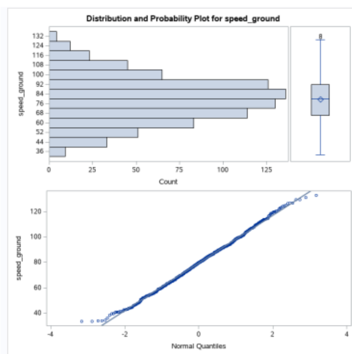
- All the variables are normally distributed except distance which is exponentially distributed.

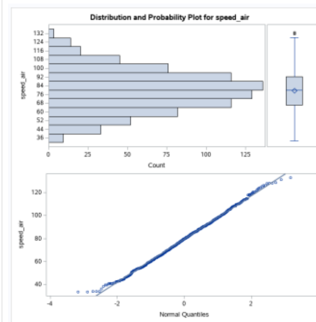
Univariate Analysis Result with graphs.pd

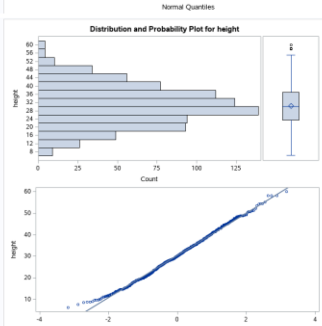
```
PROC UNIVARIATE DATA=FLIGHTS_IMPUTE PLOT;
RUN;
```

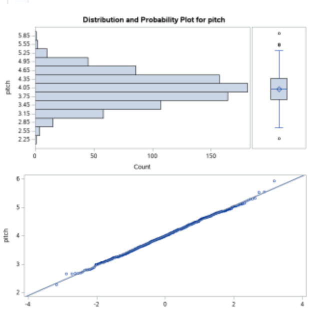


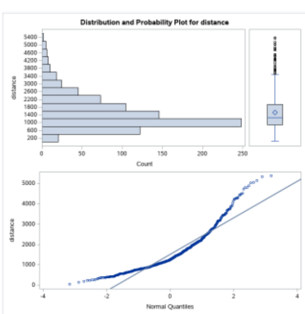












Conclusion

1. Initial dataset has 950 observations (plus 50 null rows).
2. There were 100 duplicate values, 628 missing speed air values and 50 duration values and around 19 outliers.
3. After removing the duplicates, and outliers and dealing with the missing values, the prepared dataset has 830 observations and 8 columns. Our dataset is now ready for descriptive analysis.

CHAPTER 2: DESCRIPTIVE STUDY OF VARIABLES

Goals: To study association of landing distance with different variables and try to find variables of significance.

STEP 1: STUDYING LANDING DISTANCE WITH OTHER VARIABLES

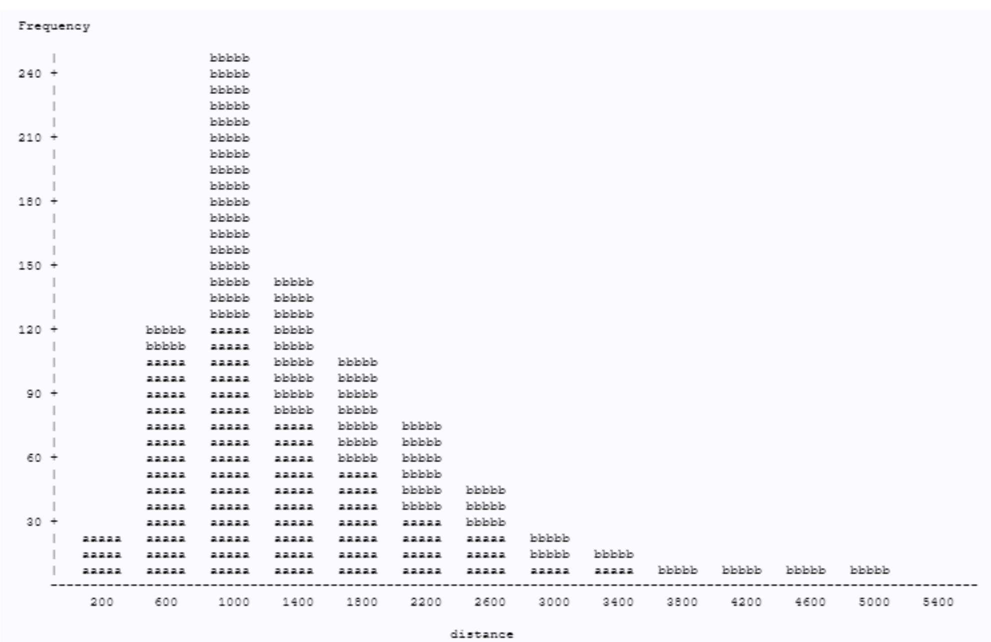
Since we are studying landing distance, I will study the effects of different variables on the landing distance.

LANDING DISTANCE AND AIRCRAFT TYPE

Observations:

Landing distance is exponentially distributed, and it appears to be higher for Boeing aircrafts in general. I will perform another test to confirm this observation.

```
/* Landing distance sorted by aircrafts*/

proc chart data=FLIGHTS_IMPUTE;
vbar distance / subgroup = aircraft;
run;
```



LANDING DISTANCE AND AIRCRAFT TYPE BY TTEST (TO CONFIRM ASSOCIATION BETWEEN AIRCRAFT TYPE AND DISTANCE)

Observations:

- It is consistent with the prior observation that Boeing aircraft require a higher landing distance than airbus.

```
/* TTEST */
proc ttest data=FLIGHTS_IMPUTE;
class aircraft;
var distance;
run;
```

The TTEST Procedure

Variable: distance (distance)

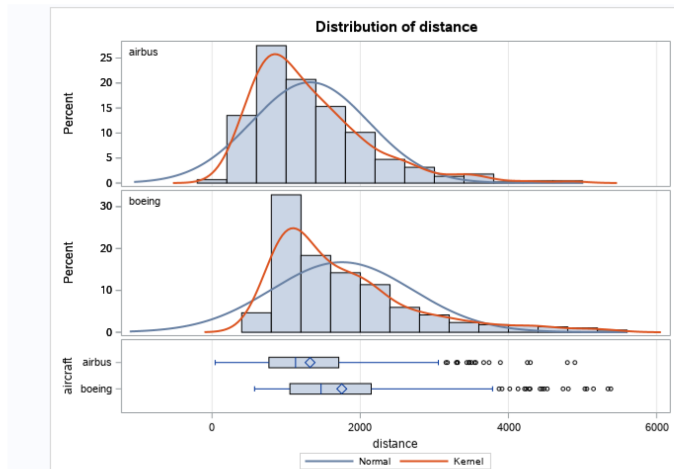
aircraft	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus		444	1323.3	791.9	37.5833	41.7223	4896.3
boeing		387	1751.0	953.9	48.4869	573.6	5382.0
Diff (1-2)	Pooled		-427.7	871.1	60.5772		
Diff (1-2)	Satterthwaite		-427.7		61.3472		

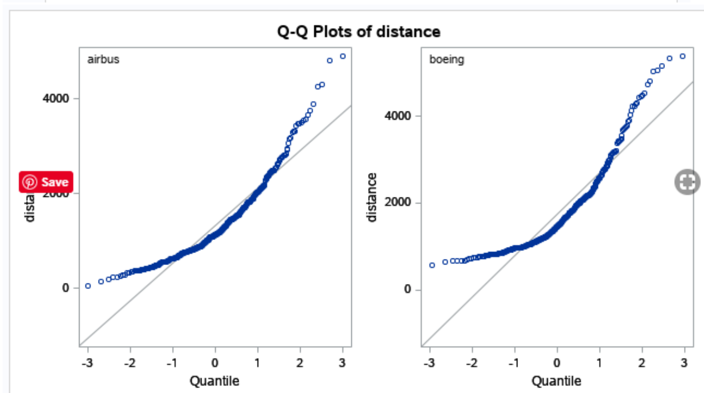
aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1323.3	1249.5	1397.2	791.9	743.0	847.8
boeing		1751.0	1655.7	1846.3	953.9	891.1	1026.2
Diff (1-2)	Pooled	-427.7	-546.6	-308.8	871.1	831.1	915.1
Diff (1-2)	Satterthwaite	-427.7	-548.1	-307.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	829	-7.06	<.0001
Satterthwaite	Unequal	752.49	-6.97	<.0001

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	386	443	1.45	0.0002



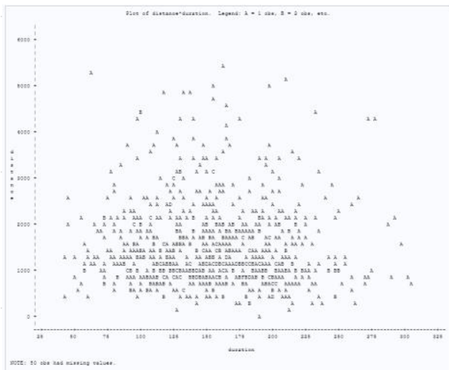


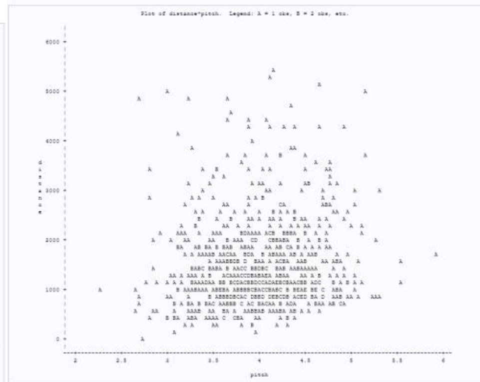
LANDING DISTANCE AND OTHER VARIABLES

Observations:

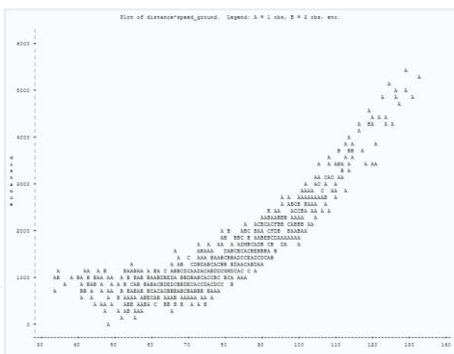
- The distribution of all variables are random except speed air and ground speed.
- Speed air and speed ground show strong co-relation when plotted against the variable distance. This is consistent with prior findings of high co-relation co-efficient between speed air and distance, and speed ground and distance.

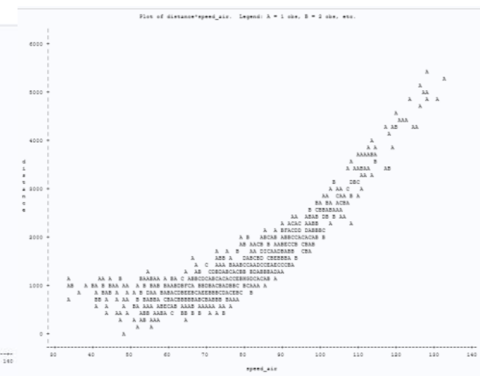
```
/* plotting distance and duration*/

PROC PLOT DATA=FLIGHTS_IMPUTE;
PLOT DISTANCE*DURATION;
RUN;


/* plotting distance and pitch*/


PROC PLOT DATA=FLIGHTS_IMPUTE;
PLOT distance*pitch;
RUN;

/* plotting distance and ground speed*/


PROC PLOT DATA=FLIGHTS_IMPUTE;
PLOT distance*speed_ground;
RUN;


/* plotting distance and air speed*/


PROC PLOT DATA=FLIGHTS_IMPUTE;
PLOT distance*speed_air;
RUN;

/* plotting distance and number of passengers*/


PROC PLOT DATA=FLIGHTS_IMPUTE;
PLOT distance*no_pasg;
RUN;
```

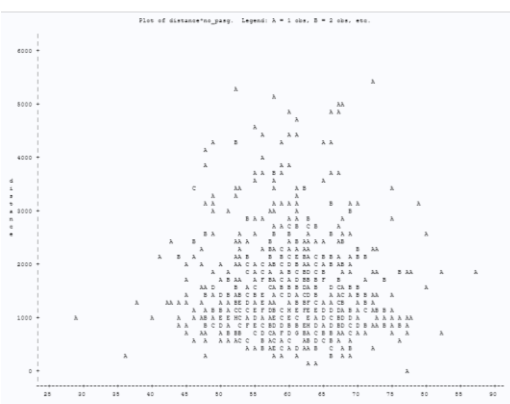


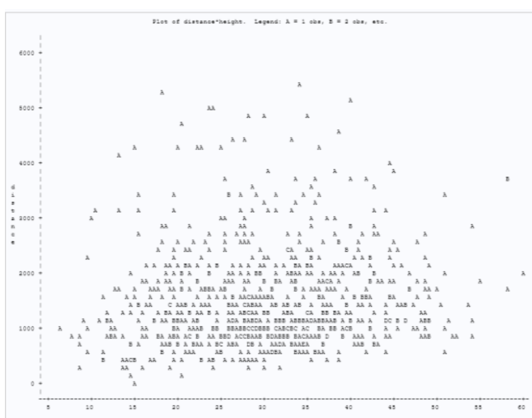


```
/* plotting distance and height*/


PROC PLOT DATA=FLIGHTS_IMPUTE;
PLOT distance*height;
RUN;
```









STEP 2: STUDYING CO-RELATION BETWEEN DIFFERENT VARIABLE AFTER IMPUTATION

I will again study the co-relation between different coefficients after substituting the missing speed air values.

```
/*finding co-relation between different variables after imputation*/

proc corr data=FLIGHTS_IMPUTE;
var distance duration height no_pasg pitch speed_air speed_ground;
title Pairwise correlation coefficients;
run;
```

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

	distance	duration	height	no_pasg	pitch	speed_air	speed_ground
distance distance	1.00000 831	-0.05138 0.1514 781	0.09941 0.0041 831	-0.01776 0.6093 831	0.08703 0.0121 831	0.86780 <.0001 831	0.86624 <.0001 831
duration duration	-0.05138 0.1514 781	1.00000 781	0.01112 0.7564 781	-0.03639 0.3098 781	-0.04675 0.1918 781	-0.04645 0.1948 781	-0.04897 0.1716 781
height height	0.09941 0.0041 831	0.01112 0.7564 781	1.00000 831	0.04699 0.1760 831	0.02298 0.5082 831	-0.05631 0.1048 831	-0.05761 0.0970 831
no_pasg no_pasg	-0.01776 0.6093 831	-0.03639 0.3098 781	0.04699 0.1760 831	1.00000 831	-0.01793 0.6057 831	-0.00056 0.9871 831	-0.00013 0.9969 831
pitch pitch	0.08703 0.0121 831	-0.04675 0.1918 781	0.02298 0.5082 831	-0.01793 0.6057 831	1.00000 831	-0.03616 0.2978 831	-0.03912 0.2599 831
speed_air speed_air	0.86780 <.0001 831	-0.04645 0.1948 781	-0.05631 0.1048 831	-0.00056 0.9871 831	-0.03616 0.2978 831	1.00000 831	0.99918 <.0001 831
speed_ground speed_ground	0.86624 <.0001 831	-0.04897 0.1716 781	-0.05761 0.0970 831	-0.00013 0.9969 831	-0.03912 0.2599 831	0.99918 <.0001 831	1.00000 831

Observations:

1. High co-relation between speed air and distance.
2. High correlation between speed ground and distance.
3. Even higher co-relation between speed air and speed ground since the missing speed air values were calculated from speed ground.

CONCLUSION

1. There seems to be high co-relation of distance with speed air and speed ground both.
2. Imputation increased the co-relation coefficient for speed air and speed ground.

CHAPTER 3: STATISTICAL MODELING

Goals: To use a linear regression model and study the relationship of the dependent variable (distance) with independent variables (aircraft, duration, no. of passengers, speed air, speed ground, pitch and height).

STEP 1: ASSIGNING AIRCRAFT TYPE A NUMERICAL VALUE IN ORDER TO PERFORM REGRESSION ANALYSIS

Since aircraft is not a numerical value, I will assign both the types of aircraft a numerical value in order to perform the regression analysis. The generated table will thus have 831 rows and 9 columns.

Statistical Computing (BANA 6043 Project)

Gupta2na Niharika Gupta M13437287

```
/*assigning numerical value to aircraft type*/

data flights_impute2;
set flights_impute;
if aircraft="airbus"
then Air_type=0; else Air_type=1;
run;
proc print data=flights_impute2;
run;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	Air_type
1	airbus	190.7394255	77	47.882117055	47.808234155	14.835964361	2.7322842836	41.722312733	0
2	airbus	212.05403613	63	51.587044527	51.513161627	20.451285811	3.063686215	133.08690985	0
3	airbus	128.37336566	64	55.461625107	55.387742207	14.65127605	3.9792117538	180.56522534	0
4	airbus	237.40527671	48	53.774013118	53.700130218	28.260802216	3.1755295986	241.16096423	0
5	airbus	142.5876457	66	51.158228388	51.084345488	8.559069177	3.9134477851	242.59588646	0
6	airbus	172.04931209	36	47.486765029	47.412882129	13.984809941	4.2990197162	250.68976141	0
7	airbus	230.32398183	58	55.108631792	55.034748892	29.859498104	3.2599541617	270.83676243	0
8	airbus	175.53311361	61	65.037084787	64.963201887	13.807590435	3.4948549953	280.80440304	0
9	airbus	182.44776305	66	52.70784152	52.63395862	24.302641153	4.1859666088	317.81268659	0
10	airbus	214.78506113	59	56.285225619	56.211342719	19.097947487	3.9151278527	321.51632716	0
11	airbus	183.61849925	69	53.539242523	53.465359623	31.739422907	3.5237749131	349.15851648	0
12	airbus	261.38701422	68	57.08470944	57.01082654	15.761691561	3.8049960732	350.60240534	0
13	airbus	149.20859096	66	63.817889502	63.744006602	11.238021337	3.6038718062	370.47074159	0
14	airbus	208.87900668	56	53.888762395	53.814879495	23.496668191	4.0000159209	375.32596789	0
15	airbus	98.176296764	60	53.749134607	53.675251707	25.54578925	3.7142005991	378.82578267	0
16	airbus	42.146226159	63	46.264718501	46.190835601	20.490711515	3.4819121545	383.55849778	0
17	airbus	179.99562817	66	57.243835228	57.169952328	35.325529113	3.0665057372	383.57772124	0
18	airbus	198.42235144	58	64.305146799	64.231263899	17.456593278	3.8221946371	383.90578116	0
19	airbus	207.69159848	61	54.542338048	54.468455148	19.610845089	3.9730540461	397.01200564	0

STEP 2: REGRESSION ANALYSIS OF EACH INDEPENDENT VARIABLE WITH DISTANCE

Observations:

- Considering a significance value of 0.05, 5 variables are significant i.e. Air_type, speed_ground, speed_air, pitch and height. I can drop duration and number of passengers from the model since they don't seem to have any impact on the distance variable.

```
/*Aircraft type and distance Regression analysis*/

proc reg data=flights_impute2;
model distance= Air_type;
title regression analysis of the flights dataset;
run;


/*duration and distance Regression analysis*/
proc reg data=flights_impute2;
model distance= duration;
title regression analysis of the flights dataset;
run;

/*no_pasg and distance Regression analysis*/
proc reg data=flights_impute2;
model distance= no_pasg;
title regression analysis of the flights dataset;
run;

/*speed_ground and distance Regression analysis*/
proc reg data=flights_impute2;
model distance= speed_ground;
title regression analysis of the flights dataset;
run;

/*speed_air and distance Regression analysis*/
proc reg data=flights_impute2;
model distance= speed_air;
title regression analysis of the flights dataset;
run;

/*pitch and distance Regression analysis*/
proc reg data=flights_impute2;
model distance= pitch;
title regression analysis of the flights dataset;
run;

/*height and distance Regression analysis*/
proc reg data=flights_impute2;
model distance= height;
title Regression analysis of the flights dataset;
run;
```

regression analysis of the flights dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	37818390	37818390	49.84	<.0001
Error	829	629021939	758772		
Corrected Total	830	666840329			

Root MSE	871.07516	R-Square	0.0567
Dependent Mean	1522.48287	Adj R-Sq	0.0556
Coeff Var	57.21412		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1323.31696	41.33940	32.01	<.0001
Air_type		1	427.66634	60.57720	7.06	<.0001

regression analysis of the flights dataset

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	1	1685114	1685114	2.06	0.1514
Error	779	636576145	817171		
Corrected Total	780	638261260			

Root MSE	903.97507	R-Square	0.0026
Dependent Mean	1541.20394	Adj R-Sq	0.0014
Coeff Var	58.65383		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1689.99417	108.54521	15.57	<.0001
duration	duration	1	-0.96133	0.66944	-1.44	0.1514

regression analysis of the flights dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	210253	210253	0.26	0.6093
Error	829	666630076	804138		
Corrected Total	830	666840329			

Root MSE	896.73720	R-Square	0.0003
Dependent Mean	1522.48287	Adj R-Sq	-0.0009
Coeff Var	58.89966		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1650.07564	251.45979	6.56	<.0001
no_pasg	no_pasg	1	-2.12459	4.15497	-0.51	0.6093

regression analysis of the flights dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	500382567	500382567	2492.03	<.0001
Error	829	166457762	200793		
Corrected Total	830	666840329			

Root MSE	448.09981	R-Square	0.7504
Dependent Mean	1522.48287	Adj R-Sq	0.7501
Coeff Var	29.43217		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1773.94071	67.83878	-26.15	<.0001
speed_ground	speed_ground	1	41.44219	0.83017	49.92	<.0001

regression analysis of the flights dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	502185777	502185777	2528.40	<.0001
Error	829	164654552	198618		
Corrected Total	830	666840329			

Root MSE	445.66611	R-Square	0.7531
Dependent Mean	1522.48287	Adj R-Sq	0.7528
Coeff Var	29.27232		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1773.67565	67.35033	-26.34	<.0001
speed_air	speed_air	1	41.45855	0.82450	50.28	<.0001

regression analysis of the flights dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5050617	5050617	6.33	0.0121
Error	829	661789712	798299		
Corrected Total	830	666840329			

Root MSE	893.47569	R-Square	0.0076
Dependent Mean	1522.48287	Adj R-Sq	0.0064
Coeff Var	58.68543		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	929.15082	237.91686	3.91	0.0001
pitch	pitch	1	148.14188	58.89635	2.52	0.0121

Regression analysis of the flights dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6590108	6590108	8.27	0.0041
Error	829	660250221	796442		
Corrected Total	830	666840329			

Root MSE	892.43586	R-Square	0.0099
Dependent Mean	1522.48287	Adj R-Sq	0.0087
Coeff Var	58.61714		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1245.11620	101.27189	12.29	<.0001
height	height	1	9.10657	3.16581	2.88	0.0041

STEP 3: REGRESSION ANALYSIS OF SIGNIFICANT VARIABLES TOGETHER WITH DISTANCE

```
/*regression Anlaysis of significant variables with distance*/
proc reg data=flights_impute2;
model distance= speed_ground speed_air height pitch Air_type;
title regression analysis of the flights dataset;
run;
```

regression analysis of the flights dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	567652644	113530529	944.30	<.0001
Error	825	99187686	120227		
Corrected Total	830	666840329			

Root MSE	346.73837	R-Square	0.8513
Dependent Mean	1522.48287	Adj R-Sq	0.8504
Coeff Var	22.77453		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2634.18236	116.17392	-22.67	<.0001
speed_ground	speed_ground	1	-9.08733	15.94304	-0.57	0.5688
speed_air	speed_air	1	51.47616	15.91780	3.23	0.0013
height	height	1	13.97350	1.23327	11.33	<.0001
pitch	pitch	1	34.23338	24.51583	1.40	0.1630
Air_type		1	481.36930	25.80388	18.65	<.0001

Observations:

- From the observations, I derieved the below formula:

  Distance= -2634.18 + (-9.08733* Speed_ground) +(51.47616*speed_air) +(13.97350*height) +(34.23338*pitch) +(481.36930*Air_type)

- Considering a significance level of 0.05 (i.e. 5%) or less I can say that only variables speed_air, height, and Air_type are significant.
- Although R_sq 0.8513 is a high value, that means that the model fits the data well. However, I will try to find a better model that fits the data in a more accurate way and has an even higher R_sq value.
- The significance value of the Pitch variable (1.078) suggests that it does not fit in my model and needs to be removed.
- Significance value for speed ground (0.5688) no longer fits in my previous already inferred correlation between speed_ground and speed_air. Since speed_air and speed_ground are dependent on each other thus in the regression model they might not co-exist and only one of these two variables might fit the data better. I will drop speed ground as most of the values for speed air are missing and what I have are only predicted values, so not as reliable.

STEP 4: REGRESSION ANALYSIS AFTER REMOVING SPEED AIR AND PITCH

```
/*regression Anlaysis of significant variables with distance after removing speed_air and pitch*/
proc reg data=flights_impute2;
model distance= speed_ground height Air_type;
title Regression analysis of the flights dataset;
run;
```

Regression analysis of the flights dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	566080053	188693351	1548.72	<.0001
Error	827	100760276	121838		
Corrected Total	830	666840329			

Root MSE	349.05344	R-Square	0.8489
Dependent Mean	1522.48287	Adj R-Sq	0.8484
Coeff Var	22.92659		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2512.24333	68.19743	-36.84	<.0001
speed_ground	speed_ground	1	42.40242	0.64830	65.41	<.0001
height	height	1	14.14783	1.24046	11.41	<.0001
Air_type		1	496.04524	24.29753	20.42	<.0001

Observation:

- The model obtained using the above variables is the following:

  Distance= -2512.24333+ (42.40242*speed_ground) + (14.14783*height) + (496.04524*Air_type)

- High value of R_square (0.8489) means that the model fits the data well even after removing the speed_air and pitch. However, I will stilltry to find a better model after model validation to see if there exists a model with even higher R_square value.

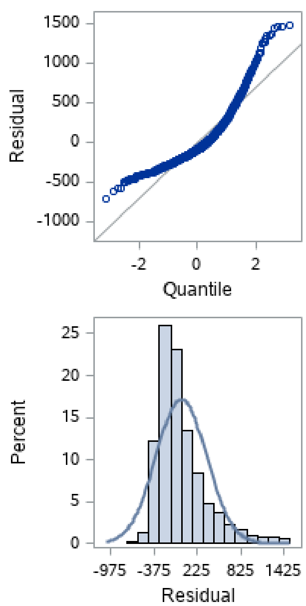
CONCLUSION:

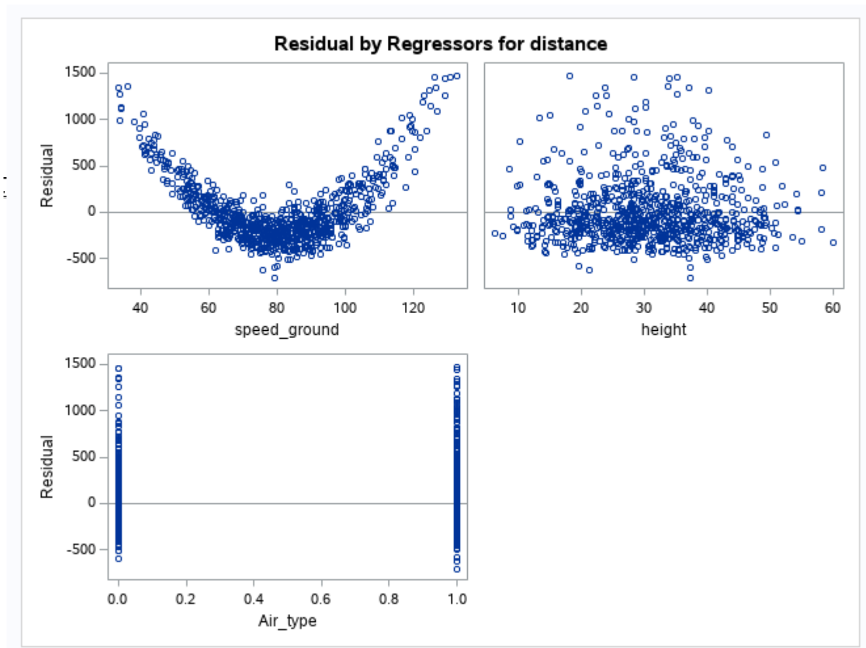
- All the selected variables have a positive impact on the dependent variable distance. Thus, higher the values of the independent variables, the greater the landing distance and eventually greater the risk of landing overrun.
- Regression analysis of each independent variable with distance shows Air_type, speed_ground, speed_air, pitch and height are significant, and duration and pitch can be dropped.
- Regression analysis of remaining variables together with distance shows that pitch is no longer significant and speed air and speed ground cannot co-exist in the model as they are dependent on each other. Pitch and speed air are then dropped.
- Regression analysis done with remaining variables and a formula for calculating landing distance is derived.
- Current model has a high R square values (0.8489) indicating it fits our data well but I will try and find a better model.

CHAPTER 4: MODEL VALIDATION

Goals: To apply model validation and validate normal distribution and residual regression.

```
/*regression Anlaysis of significant variables with distance model validation*/
proc reg data=flights_impute2;
model distance= speed_ground height Air_type;
title Regression analysis of the flights dataset;
output out=diagnostics r=residual;
run;
```





Observations: Also, two more inferences can be derived from the fit diagnostics model plots:

• The Quantile Residual plot demonstrates that the normal distribution of the residuals is not fairly met.

• The residual percent chart doesn't follow a normal distribution and is a little skewed to the right.

As seen in the residual regression chart above, the residuals for both height and Air_type don't look to follow any pattern and are scattered. However, speed_ground seems to follow a pattern.

CONCLUSION:

1. Residuals failed the normality test and I should study the model diagnostics of the individual variables in order to better meet the model diagnostics requirements.
2. The linear model generated does not pass our model diagnostic requirements and transformations are required on this data as residuals have a curved pattern.

CHAPTER 5: REMODELING AND MODEL VALIDATION

Goals: In the previous chapter I learnt that the residuals for both height and Air_type didn't follow any pattern and are scattered. But speed_ground seems to follow a pattern. So, in order to solve this problem, I can introduce a quadratic term in this variable which might probably fix this pattern and get a scattered chart.

```
/*Modifying the speed_ground variable to adjust to quadratic equation*/
Data flights_impute3;
set flights_impute2;
Format speed_ground2;
speed_ground2= speed_ground**3;
run;

/*regression Anlaysis of significant variables with distance after adding the quadratic equation*/
proc reg data=flights_impute3;
model distance= speed_ground2 height Air_type;
title Regression analysis of the flights dataset;
output out=diagnostics r=residual;
run;
```

Regression analysis of the flights dataset

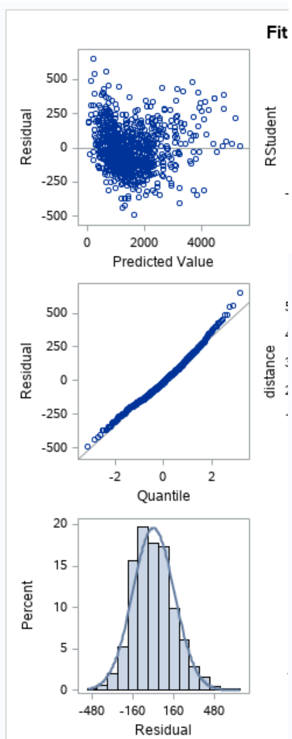
The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

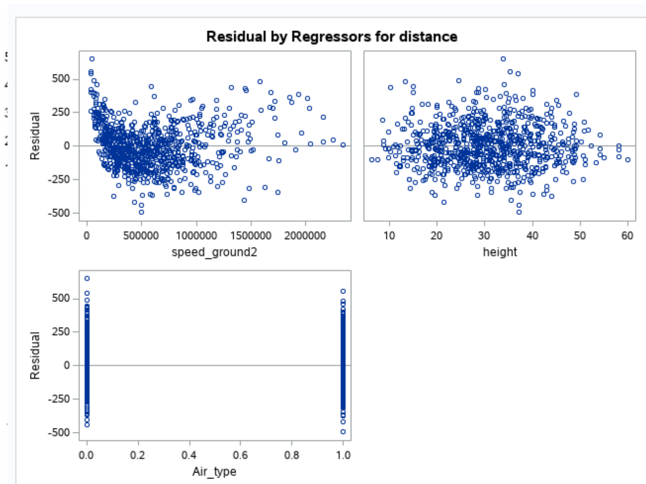
Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	644794996	214931665	8062.86	<.0001
Error	827	22045333	26657		
Corrected Total	830	666840329			

Root MSE	163.26969	R-Square	0.9669
Dependent Mean	1522.48287	Adj R-Sq	0.9668
Coeff Var	10.72391		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-354.75055	21.44332	-16.54	<.0001
speed_ground2		1	0.00214	0.00001426	150.02	<.0001
height	height	1	13.86544	0.58000	23.91	<.0001
Air_type		1	425.35176	11.35549	37.46	<.0001





Observations:

- After introducing the quadratic term to the equation, the new model looks like the following: **Distance= -354.75055+ (0.00214*speed_ground^3) + (13.86544* height) + (425.35176*Air_type)**
- As seen in the below image, this model fits in all our requirements as validated below:
    1. The model results in an extremely high R-Square value of (0.9669)
    2. All variables have a p value lower than 0.05 significance
    3. Parameter estimates for all three variables speed_ground, height and Air_type is positive
    4. Normality assumption is fulfilled where residuals are normally distributed around 0
    5. The quantile residual chart is not curved
    6. No patterns can be seen in the residual regression chart and are scattered

CONCLUSION:

- I have found a better model as compared to all my previous models and it fulfills all my requirements of model diagnostic validation.
- This system can promptly alert pilots when during the landing procedure the ‘distance’ value is anticipated to be greater than 6000 according to the formula, and pilot can take appropriate actions to reduce the chances of landing overrun.