Name: Varsha Agarwalla

UCID- M12910001

FAA Project

The goal of this project is to study the factors that would impact the landing distance of a commercial flight to reduce the risk of landing overrun.

We were provided with data for 950 flights initially and factors that include- aircraft type (airbus and Boeing), number of passengers, speed of air and that of ground, height and pitch of the flight, flight duration and the landing distance. When we come across these factors, we have a gut feeling that all these factors play an important role in determining the distance, but gut feeling isn't right when it comes to data. We study how distance is dependent on each of the variables(factors) by studying the correlation values. It gives us a better idea for building our model according. Since we work with sample data, there is a sampling error and it will never be a copy of population that it represents. We build a model which will perform well on the population rather than on sample. Also, the model should be able to perform on any kind of sampling data. This project helped us understand the factors individually and identify the impact each one has on each other and on the distance.

Every step performed on the data becomes important. Someone rightly said, "never trust the data, even if it is provided by your client!"

- Data cleaning- there were redundant observations and abnormality in the data which was treated
- Data visualization- plots help us get the complete picture of data, it helps us see how data is behaving and how they are correlated
- Correlation check- we have the plots, but to how the exact measure of data behavior, we need numbers. Hence, we perform correlation
- Modelling – this is the final but a crucial step. We must identify which factors impact and fit our model. We ensure that we don't overfit our model in any way just to improve our R^2. The model should be built in a way that it can be applied to any sample data or the entire population
- Model checking- checks performed on the residuals to see if the initial assumptions that come with model creation hold true

This project helped me understand the business case better with every step I performed. I am thankful to Prof. Liu for providing me with the opportunity to learn and apply the concepts.

Chapter 1: Data Preparation

Objective: Data Exploration and Data Cleaning

Analysis:

1. Step 1: load the given data-sets

Load the FAA1.txt file

```
FILENAME REFFILE '/folders/myfolders/GASUE34/FAA1.xls';

PROC IMPORT DATAFILE=REFFILE
        DBMS=XLS
        OUT=GASUE34.FAA1;
        GETNAMES=YES;
        SHEET="FAA1";
RUN;

PROC CONTENTS DATA=GASUE34.FAA1; RUN;
```

Output: 800 observations and 8 variables

LOAD THE FAA2.txt file

```
FILENAME REFFILE '/folders/myfolders/GASUE34/FAA2.xls';

PROC IMPORT DATAFILE=REFFILE
        DBMS=XLS
        OUT=GASUE34.FAA2;
        GETNAMES=YES;
        SHEET="FAA2";
RUN;

PROC CONTENTS DATA=GASUE34.FAA2; RUN;
```

Output: 150 observations and 7 variables

2. Step 2: Append the two data-sets

```
/*Created a copy of the data-sets before doing any analysis to keep the original data safe*/
    DATA FAA1_SORT;

    SET GASUE34.FAA1_EDITED;
    RUN;
    PROC PRINT DATA = FAA1_SORT;

    DATA FAA2_SORT;
    SET GASUE34.FAA2_EDITED; RUN;
```

```
PROC PRINT DATA = FAA2_SORT;

/*Sorted both the data-sets before appending- interleaving datasets*/
PROC SORT DATA = FAA1_SORT;
BY AIRCRAFT ;
PROC PRINT DATA = FAA1_SORT;
RUN;
PROC SORT DATA= FAA2_SORT;
BY AIRCRAFT;
PROC PRINT DATA = FAA2_SORT;
RUN;
/*Appended dataset creation*/
DATA FAA1_FAA2;
SET FAA1_SORT FAA2_SORT;
BY AIRCRAFT;
PROC PRINT DATA = FAA1_FAA2;
RUN;
```

Output: 950 observations and 8 variables

3. Step 3: Removing Duplicate records

I observed duplicates in the data for all the columns except for 'duration'. Since, the column is missing in the second data-set, it might be the case that the observations are repeated for few cases

Note: I have used SQL statements in SAS as most of our functions run SQL in the background. (Please let me know if you think otherwise)

Code to check the duplicate records-

```
SELECT *, COUNT(*) AS COUNT FROM FAA1_FAA2 GROUP BY aircraft, no_pasg, speed_ground, speed_air, height, pitch, distance having count > 1; RUN;
```

Since the case was similar for most of the records, please find a snippet of the duplicate records-

aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	COUNT
boeing	124.94457133	44	72.546668651	.	42.859879536	4.028501716	1321.1606709	2
boeing	.	44	72.546668651	.	42.859879536	4.028501716	1321.1606709	2
boeing	.	44	95.068731567	96.084198669	17.376660338	3.2011486815	2183.7983628	2
boeing	83.515006172	44	95.068731567	96.084198669	17.376660338	3.2011486815	2183.7983628	2
boeing	142.15534911	46	39.769294325	.	39.655921061	4.5992872267	1030.457488	2
boeing	.	46	39.769294325	.	39.655921061	4.5992872267	1030.457488	2
boeing	146.38562216	47	90.354274819	.	14.114268518	4.0402550027	1593.0611271	2
boeing	.	47	90.354274819	.	14.114268518	4.0402550027	1593.0611271	2
boeing	.	47	95.322576422	94.215160768	30.270100189	3.6451345759	2233.0489624	2
boeing	156.80568995	47	95.322576422	94.215160768	30.270100189	3.6451345759	2233.0489624	2
boeing	.	49	57.65125066	.	30.305425419	3.9341591214	981.8893648	2
boeing	162.45273186	49	57.65125066	.	30.305425419	3.9341591214	981.8893648	2
boeing	.	49	66.192530367	.	47.715701656	3.6191908432	1595.1338347	2
boeing	97.76437201	49	66.192530367	.	47.715701656	3.6191908432	1595.1338347	2
boeing	186.68141397	49	66.417230464	.	44.692695788	4.1135438115	1176.0276765	2
boeing	.	49	66.417230464	.	44.692695788	4.1135438115	1176.0276765	2
boeing	.	49	84.588609025	.	37.080439428	3.3443854922	1814.7887866	2
boeing	126.94651352	49	84.588609025	.	37.080439428	3.3443854922	1814.7887866	2
boeing	178.71333071	51	70.480194088	.	7.5824945838	4.8167893156	822.2286414	2
boeing	.	51	70.480194088	.	7.5824945838	4.8167893156	822.2286414	2
boeing	.	52	46.965489789	.	48.836222177	3.7268981671	1136.0148411	2
boeing	71.573834716	52	46.965489789	.	48.836222177	3.7268981671	1136.0148411	2
boeing	.	52	81.533090888	.	22.411979234	3.702074231	1587.3880099	2
boeing	196.46411848	52	81.533090888	.	22.411979234	3.702074231	1587.3880099	2
boeing	130.46356358	52	116.71343434	117.65649967	36.195527446	3.8943524297	4240.0941825	2
boeing	.	52	116.71343434	117.65649967	36.195527446	3.8943524297	4240.0941825	2
boeing	163.73992283	53	44.394275805	.	37.763521555	3.8154730289	996.87232711	2
boeing	.	53	44.394275805	.	37.763521555	3.8154730289	996.87232711	2
boeing	153.65742555	53	79.413854562	.	14.4479785	3.2463483541	1240.2804099	2
boeing	.	53	79.413854562	.	14.4479785	3.2463483541	1240.2804099	2
boeing	112.90009528	53	98.180410862	99.135830727	28.152991316	3.9874712191	2586.6650864	2
boeing	.	53	98.180410862	99.135830727	28.152991316	3.9874712191	2586.6650864	2

/*Checked the record count of duplicate records*/

```
PROC SQL;

SELECT COUNT, COUNT(*) FROM (SELECT *, COUNT(*) AS COUNT FROM FAA1_FAA2
 GROUP BY aircraft, no_pasg, speed_ground, speed_air, height, pitch, distance) GROUP BY
COUNT; RUN;
```

Output:

COUNT	
1	750
2	200

So, there are 100 duplicate records (count 2 means – two record with same data, so only 100 original records). I created a data-set removing duplicates.

Code:

```
PROC SORT DATA = FAA1_FAA2 OUT =nodup_FAA1_FAA2 NODUPKEY;
BY aircraft no_pasg speed_ground speed_air height pitch distance;
RUN;
```

4. Step 4: Checked individual variable to identify outliers.

- **Speed_ground**

```
/*SPEED_GROUND*/

/*VALUE LESS THAN 30 AND > 140 IS ABNORAML*/
        PROC SORT DATA = nodup_FAA1_FAA2;
        BY AIRCRAFT ;
        RUN;
        PROC PRINT DATA = nodup_FAA1_FAA2;
        RUN;

        PROC UNIVARIATE DATA = nodup_FAA1_FAA2;
        BY AIRCRAFT;
        VAR SPEED_GROUND;
        RUN;
```

Output:

The UNIVARIATE Procedure
Variable: speed_ground (speed_ground)

aircraft=airbus

Moments			
N	450	Sum Weights	450
Mean	80.1994492	Sum Observations	36089.7521

aircraft=airbus

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
33.5741	1	120.558	446
36.4214	2	123.311	447
40.8018	3	125.212	448
41.1010	4	126.244	449
43.8528	5	131.035	450

Observation: no abnormality in the ground speed of airbus.

The UNIVARIATE Procedure
Variable: speed_ground (speed_ground)

aircraft=boeing

Moments			
N	300	Sum Weights	300
Mean	78.8615082	Sum Observations	23658.4525

The UNIVARIATE Procedure
Variable: speed_ground (speed_ground)

aircraft=boeing

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
27.7357	451	126.669	746
33.8230	452	126.839	747
34.1178	453	129.307	748
34.2221	454	132.785	749
34.3036	455	136.659	750

Observation: So, we observe that for boeing, we have only 1 out of 300 cases, with abnormal ground speed and it is very close to the threshold value.

Action: we will safely change the value to the minimum threshold value i.e. 30. I am taking this step because the value is very close to the threshold and it will not cause much of an impact.

Code:

```
DATA nodup_FAA12_SPEED;
SET nodup_FAA1_FAA2;
IF speed_ground <30 THEN SPEED_GROUND = 30 ;
RUN;
PROC PRINT DATA = nodup_FAA12_SPEED;
RUN;

PROC SORT DATA = nodup_FAA12_SPEED;
BY SPEED_GROUND;
RUN;
```

- **Distance**

```
/*DISTANCE*/
/* <= 6000 IS REQUIRED*/
PROC SORT DATA = nodup_FAA12_SPEED;
BY AIRCRAFT;
RUN;

PROC UNIVARIATE DATA = nodup_FAA12_SPEED;
 BY AIRCRAFT;
 VAR DISTANCE;
 RUN;
```

Output:

The UNIVARIATE Procedure
Variable: distance (distance)

aircraft=airbus

Moments			
N	450	Sum Weights	450
Mean	1318.18703	Sum Observations	593184.161

The UNIVARIATE Procedure
Variable: distance (distance)

aircraft=airbus

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
34.0808	99	3891.47	445
41.7223	12	4254.93	448
133.0869	18	4295.90	447
180.5652	35	4795.64	449
241.1610	26	4896.29	450

Observation: for aircraft = airbus, the distance has not crossed the allowable limit but the minimum value (34.0808) is less than 0.5% of the maximum value. We will have to adjust those values as they are bringing the mean down.

Action: I will set the lower values to at least 5% of the higher value that is possible i.e. 300. There are 6 observations with values < 300.

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	count
10	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162	250.68976141	1
12	airbus	190.7394255	77	47.882117055	.	14.835964361	2.7322842836	41.722312733	1
17	airbus	142.5876457	66	51.158228388	.	8.559069177	3.9134477851	242.59588646	1
18	airbus	212.05403613	63	51.587044527	.	20.451285811	3.063686215	133.08690985	1
26	airbus	237.40527671	48	53.774013118	.	28.260802216	3.1755295986	241.16096423	1
33	airbus	230.32398183	58	55.108631792	.	29.859498104	3.2599541617	270.83676243	1
35	airbus	128.37336566	64	55.461625107	.	14.65127605	3.9792117538	180.56522534	1
88	airbus	175.53311361	61	65.037084787	.	13.807590435	3.4948549953	280.80440304	1
99	airbus	150.94674427	58	66.421119468	.	-2.915335901	3.1225583646	34.080783293	1

The UNIVARIATE Procedure
Variable: distance (distance)

aircraft=boeing

Moments			
N	300	Sum Weights	300
Mean	1765.57667	Sum Observations	529673.001

The UNIVARIATE Procedure
Variable: distance (distance)

aircraft=boeing

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
371.277	499	5031.39	746
641.600	507	5058.47	748
653.616	505	5147.41	745
671.303	494	5343.20	749
690.001	539	6309.95	750

Observation: Here one record has value >6000 which is an outlier,
Action: we will bring it to our threshold limit and set it to 6000. I am taking this step because the set value will not cause much of difference in the mean of the entire variable. Also, here the minimum is > 300 so I will leave it as it is.

Code:

```
DATA nodup_FAA12_DIST;
SET nodup_FAA12_SPEED;
 IF DISTANCE < 300 THEN DISTANCE = 300 ;
IF DISTANCE > 6000 THEN DISTANCE = 6000 ;
RUN;
PROC PRINT DATA = nodup_FAA12_DIST;
RUN;
```

- **Height**

```
/*HEIGHT*/
/* >= 6 IS REQUIRED*/
PROC SORT DATA = nodup_FAA12_SPEED;
BY AIRCRAFT;
RUN;

PROC UNIVARIATE DATA = nodup_FAA12_SPEED;
 BY AIRCRAFT;
 VAR HEIGHT;
 RUN;
```

Output:

The UNIVARIATE Procedure
Variable: height (height)

aircraft=airbus

Moments			
N	450	Sum Weights	450
Mean	30.3196736	Sum Observations	13643.8531

The UNIVARIATE Procedure
Variable: height (height)

aircraft=airbus

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-3.3323880	352	52.3786	50
-2.9153359	99	53.4386	251
-0.0677586	39	54.1985	414
0.0861055	141	54.2760	69
6.2275178	275	58.2278	442

The UNIVARIATE Procedure
Variable: height (height)

aircraft=boeing

Moments			
N	300	Sum Weights	300
Mean	29.8967235	Sum Observations	8969.01704

The UNIVARIATE Procedure
Variable: height (height)

aircraft=boeing

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-1.52813	567	52.4731	696
1.25386	499	54.2760	491
2.20519	559	55.0935	602
3.78892	507	58.0818	729
8.72687	505	59.9460	613

Observation: There are observations with height <6.
Action: I will see the number of records having this as the case

Code:

```
PROC PRINT DATA = nodup_FAA12_SPEED;WHERE HEIGHT < 6; RUN;
```

Output:

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	count
39	airbus	157.91497689	68	56.497986661	.	-0.067758596	4.6928768405	380.36298195	1
99	airbus	150.94674427	58	66.421119468	.	-2.915335901	3.1225583646	300	1
141	airbus	163.52364053	62	72.028024252	.	0.086105484	3.6220566648	537.91958189	1
352	airbus	103.09084673	73	92.994942381	.	-3.332387973	4.8305592948	1567.6657219	1
499	boeing	133.45985625	73	57.045299494	.	1.2538552556	4.7153842391	371.27726086	1
507	boeing	124.37864547	72	60.367043725	.	3.7889195211	3.7060888319	641.59956822	1
559	boeing	119.64402906	68	70.178463873	.	2.2051944554	3.7397746803	816.20664104	1
567	boeing	146.04337112	69	71.787305883	.	-1.528129182	4.1994604645	738.65436932	1

Action: there are 8 obs. Which is not meeting the threshold requirement. I will drop these records, assuming data as incorrect *(will consult with client for the step taken)*

Code:

```
proc sql;
create table nodup_FAA12_height as
select * from nodup_FAA12_DIST where height > 6;
run;
```

output: 840 observations and 8 variables

```
PROC UNIVARIATE DATA = nodup_FAA12_height;
BY AIRCRAFT;
VAR HEIGHT;
RUN;
```

Output:

The UNIVARIATE Procedure
Variable: height (height)

aircraft=airbus

Moments			
N	446	Sum Weights	446
Mean	30.6055661	Sum Observations	13650.0825

The UNIVARIATE Procedure
Variable: height (height)

aircraft=airbus

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
6.22752	196	52.3786	101
8.55907	344	53.4386	417
9.16463	75	54.1985	116
9.68831	55	54.2760	393
9.69722	428	58.2278	374

The UNIVARIATE Procedure
Variable: height (height)

aircraft=boeing

Moments			
N	296	Sum Weights	296
Mean	30.2814095	Sum Observations	8963.2972

The UNIVARIATE Procedure
Variable: height (height)

aircraft=boeing

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
8.72687	460	52.4731	491
8.82517	470	54.2760	712
9.68831	497	55.0935	549
11.97407	498	58.0818	544
11.99037	528	59.9460	539

- **Duration**

```
/*DURATION*/

/* The duration of a normal flight should always be greater than 40min.  */

    PROC UNIVARIATE DATA = nodup_FAA12_height;
    BY AIRCRAFT;
    VAR DURATION;
    RUN;
```

Output:

The UNIVARIATE Procedure
Variable: duration (duration)

aircraft=airbus

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
16.8935	97	273.591	260
31.7017	241	274.218	334
42.1462	278	289.320	110
45.5028	177	302.967	313
45.6354	223	305.622	106

The UNIVARIATE Procedure
Variable: duration (duration)

aircraft=airbus

Moments			
N	396	Sum Weights	396
Mean	156.233604	Sum Observations	61868.5071

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	50	11.21	100.00

Observation: for airbus, I observed 2 observations have recorded distance less than 40 minutes.

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	count
7	airbus	214.22048507	45	72.490616757	.	33.228125197	4.3693164876	748.7667918	1
11	airbus	216.87640251	45	91.618595738	.	38.324199382	4.7436314527	1967.6109937	1
16	airbus	237.40527671	48	53.774013118	.	28.260802216	3.1755295986	241.16096423	1
19	airbus	221.59245844	48	76.36854228	.	19.395151702	4.8736423429	932.57187751	1
33	airbus	203.32103587	50	64.658141026	.	35.122301566	3.5884171024	789.54314456	1
52	airbus	202.27604539	52	59.558781395	.	31.830496872	4.9226883242	650.3277785	1
57	airbus	201.66902084	52	75.801424413	.	38.404360138	3.1195493137	1051.5697597	1
67	airbus	260.50189235	53	63.489000555	.	35.371357104	3.9722083719	812.16317052	1
69	airbus	253.72768469	53	67.726377103	.	28.100904032	4.8166027921	714.5146546	1
72	airbus	202.10909397	53	76.334805748	.	50.745931362	3.5419962833	1052.6915505	1
73	airbus	203.25433498	53	81.27102299	.	26.163307183	4.1746185951	1141.4540114	1
84	airbus	209.19366153	54	50.812930767	.	38.841316346	4.0338980996	566.92692802	1
86	airbus	259.09791674	54	54.953323257	.	30.954303406	4.026187428	561.44690581	1
87	airbus	236.13989521	54	61.261575207	.	26.568850134	4.3822934229	455.51323477	1
97	airbus	16.893454896	54	94.511052223	95.930926862	37.476967053	4.1733221259	2162.92737	1
99	airbus	201.18798178	54	99.017401284	98.221920891	39.319647422	4.3935253271	2481.2581248	1

Looks like the data is not consistent. Consider the last two rows, here the height, speed, pitch of both airbuses is almost similar. So, it will take some time to land, which will be > 16 minutes.

237	airbus	218.87542131	61	70.701596429	.	34.608578408	3.3317046104	737.56722749	1
241	airbus	31.7016661	61	76.354176433	.	30.991021813	2.8173796019	948.47376723	1
242	airbus	215.14222119	61	76.358050779	.	34.755564291	4.6254327939	1085.8222616	1

Similarly, here the duration recorded as 31 is faulty. Similar is the case observed with the obs. With no duration values written.

Action: I will delete the rows with rows with duration less than 40 or greater than 200.

The UNIVARIATE Procedure
Variable: duration (duration)

aircraft=boeing

Moments			
N	296	Sum Weights	296
Mean	153.670793	Sum Observations	45486.5546

The UNIVARIATE Procedure
Variable: duration (duration)

aircraft=boeing

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
14.7642	584	272.039	587
31.3910	491	277.176	503
41.9494	537	287.003	508
56.5506	466	293.230	617
63.3295	504	298.522	688

Code: (to check the case for Boeing aircraft)

```
PROC PRINT DATA = nodup_FAA12_height; WHERE (duration > 200 or duration < 40) and
aircraft = 'boeing';RUN;
```

Output:

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	count
450	boeing	206.06572604	44	61.847975974	.	26.939627352	3.9372737398	896.58091126	1
454	boeing	212.29018	46	89.533713205	90.626181428	35.494742904	4.0010380484	2148.1079287	1
455	boeing	260.03401677	46	91.656696218	93.054834352	44.601461257	3.6923081945	2482.6563138	1
460	boeing	228.96326107	47	59.439028553	.	8.7268654925	4.0336203157	653.61609083	1
462	boeing	211.88872874	48	43.174799431	.	34.918422109	4.025280804	1057.1531205	1
465	boeing	209.67206752	48	76.515741847	.	41.40728351	4.1932644003	1398.2397027	1
472	boeing	213.98450886	49	80.394057703	.	16.962413199	4.0980200281	1531.2870582	1
484	boeing	217.50312565	51	53.691922053	.	19.911567238	3.1126360731	984.42773287	1
486	boeing	211.17454032	51	65.104876757	.	27.033533644	4.4171773513	980.45281693	1
491	boeing	31.391008253	51	98.219800666	99.057514589	52.473140903	4.1623371208	2808.3151244	1
493	boeing	222.70208536	52	39.725711308	.	33.265348033	4.4522817052	1037.914549	1

Observation: here we see the similar inconsistent data in the last two rows for duration. So I will delete the records with duration < 40

Code:

```
DATA nodup_FAA12_DUR;

SET nodup_FAA12_HEIGHT;

IF DURATION < 40  AND DURATION <> . THEN DELETE;

      PROC UNIVARIATE DATA = nodup_FAA12_DUR;
```

```
BY AIRCRAFT;
VAR duration;
RUN;
```

- **Speed air**

```
/*SPEED_AIR*/
/*VALUE LESS THAN 30 AND > 140 IS ABNORAML*/

PROC UNIVARIATE DATA = nodup_FAA12_DUR;
 BY AIRCRAFT;
 VAR speed_air;
 RUN;
```

The UNIVARIATE Procedure
Variable: speed_air (speed_air)

aircraft=airbus

Moments			
N	86	Sum Weights	86
Mean	104.212333	Sum Observations	8962.26066

Missing Values			
Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	360	80.72	100.00

The UNIVARIATE Procedure
Variable: speed_air (speed_air)

aircraft=boeing

Moments			
N	91	Sum Weights	91
Mean	103.582666	Sum Observations	9426.02261

Missing Values			
Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	205	69.26	100.00

Observation: more than 80% for airbus and 70% for Boeing have blank speed_Air. I will not be able to provide any insights, so I plan to keep the column as it is.

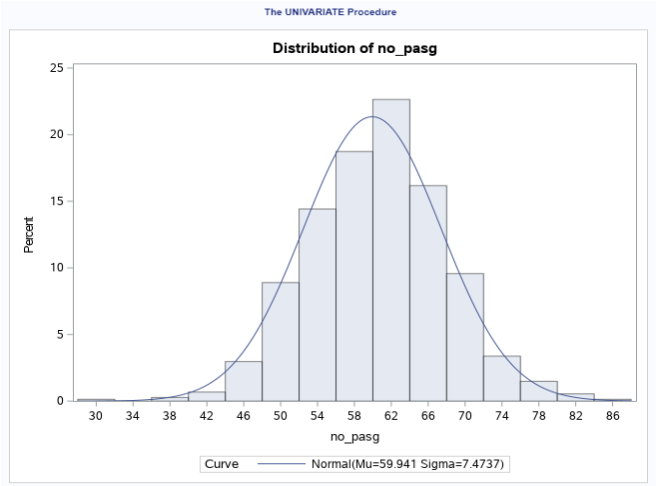
- **No_pasg**

Since, we aren't sure about the total capacity of an airbus/Boeing, we will look at the summary statistic of no_pasg. The values are consistent and hence we will not make any changes to the column values.

Code:

```
PROC UNIVARIATE DATA=nodup_FAA12_air;
VAR NO_PASG;
HISTOGRAM NO_PASG / NORMAL;
Run;
```

Output:



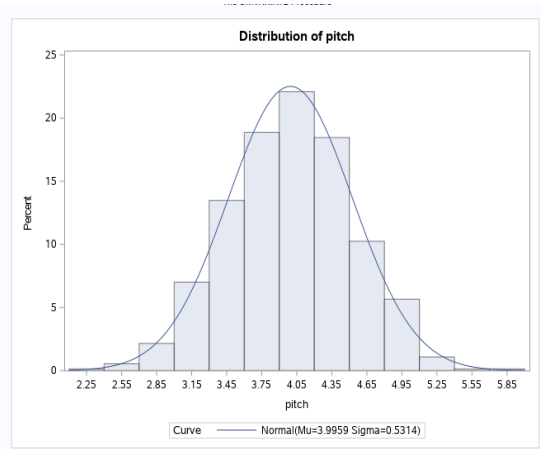
- **Pitch_Angle**

Since no threshold value is provided, I looked at the overall distribution and it seems it is close to normal. So, I will not make any changes to the column value.

Code:

```
PROC UNIVARIATE DATA=nodup_FAA12_air;
VAR PITCH;
HISTOGRAM PITCH / NORMAL;
Run;
```

Output:



So, my final data-set is ready for further processing with 835 observations and 8 variables, with 444 observations for airbus and 391 observations for Boeing

5. Step 5: Asking relevant questions based on initial data understanding and cleaning performed so as to confirm the steps taken.

Following was the observation made on each of the variable:

Variable Name	Total Obs (After Data Cleaning)	% Missing Values	Mean	Median	SD	Minimum	Maximum
Speed_ground	850	0	79.4	79.64	19.05	30	141.219
Distance	850	0	1525.03	1258.09	923.52	34	6000
Height	840	0	30.5	30.18	9.8	6.22	59.94
Duration	835	5.99	154.82	154.28	48.298	41.94	298.32
Speed_air	835	75.45	103.8322	101	10.3	90	141.725
no_pasg	835	0	60.03	60	7.49	29	87
Pitch	835	0	4	4	0.52	2.28	5.92

While cleaning data, we made few observations. I would list them down as questions. Please help me understand-

I. How are the two aircrafts different? Is there any other aircraft type that we aren't capturing here?
II. How is the frequency of these flights?
III. Is the data collected from a single airport?
IV. What is the time-period of the data?
V. How experienced were the pilots?
VI. Is the no_pasg column the actual number of passengers for the travel? What is the allowable limit for an aircraft? How many seats remain vacant on an average?
VII. The height column had negative values, can you help me understand the airport type?
VIII. Can you help me understand how pitch is measured and its threshold values?
IX. I observed most of the values in speed_air is blank. Can you help me understand how it affects the flight, so that I can gauge if we will need that information in our data?
X. How do we track the duration of the flights? I want to understand if they are tracked from any automated system? Most of the values are blank and some have captured low values, which seems inconsistent.

Once we have the questions answered and discussed on the data cleaning steps that are performed, we build our model.

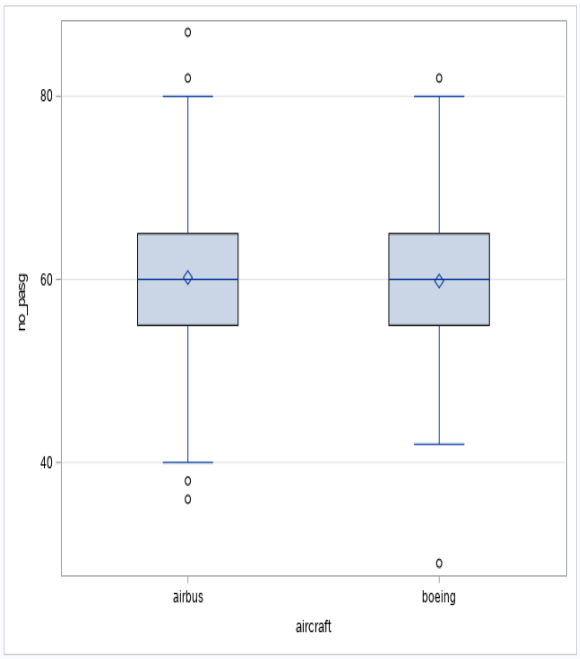
Chapter 2: Exploratory Data Analysis

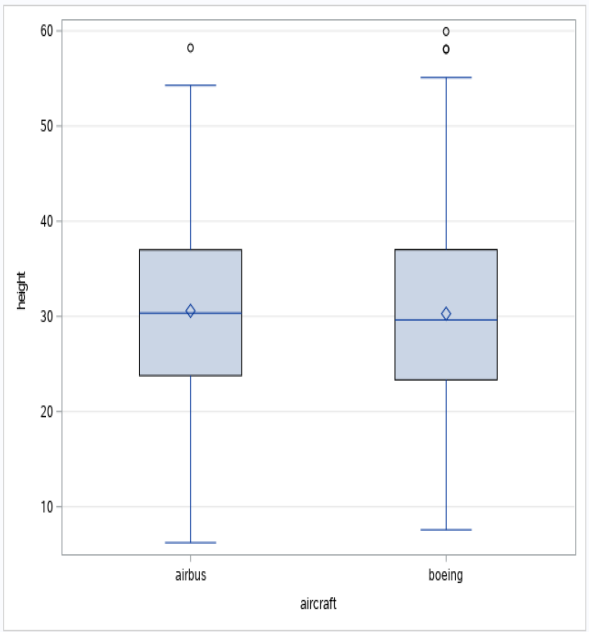
Objective: Data Visualization to see the relationship between dependent and independent variables

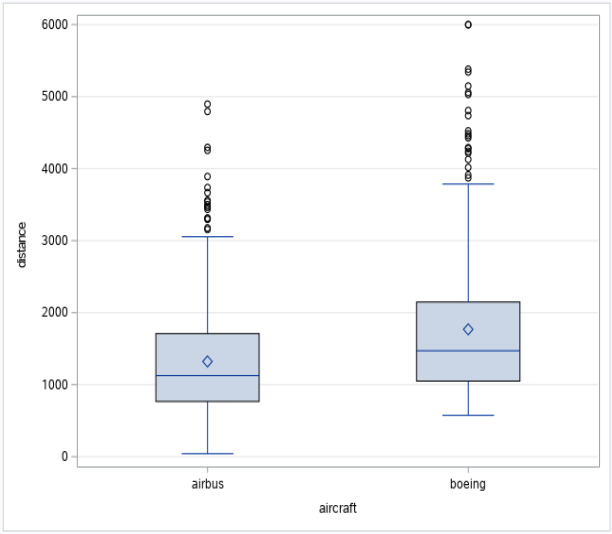
Analysis:

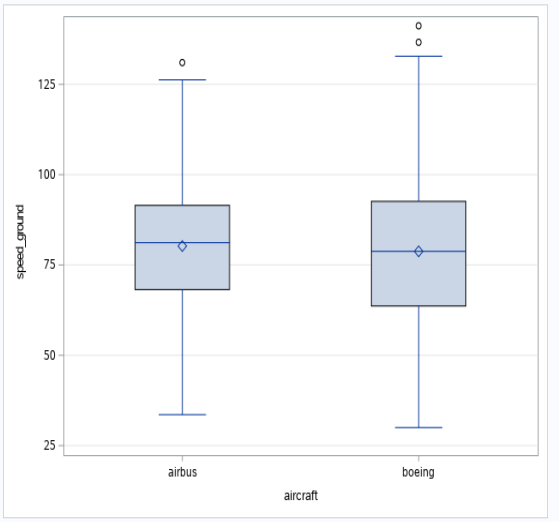
1. Step1: Plotted the box plot for all the above variables to check for the outliers

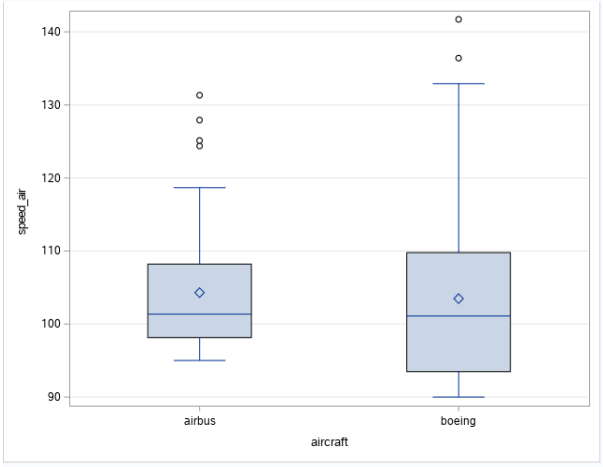
```
proc sgplot data=WORK.NODUP_FAA12_AIR;
      vbox no_pasg /category=aircraft;
      yaxis grid;
run;

proc sgplot data=WORK.NODUP_FAA12_AIR;
      vbox height /  category=aircraft;
      yaxis grid;
run;

proc sgplot data=WORK.NODUP_FAA12_AIR;
      vbox distance / category=aircraft;
      yaxis grid;
run;

proc sgplot data=WORK.NODUP_FAA12_AIR;
      vbox speed_ground / category=aircraft;
      yaxis grid;
run;

proc sgplot data=WORK.NODUP_FAA12_AIR;
      vbox speed_Air / category=aircraft;
      yaxis grid;
run;

proc sgplot data=WORK.NODUP_FAA12_AIR;
      vbox pitch / category=aircraft;
      yaxis grid;
run;

 proc sgplot data=WORK.NODUP_FAA12_AIR;
      vbox duration / category=aircraft;
      yaxis grid;
run;
```

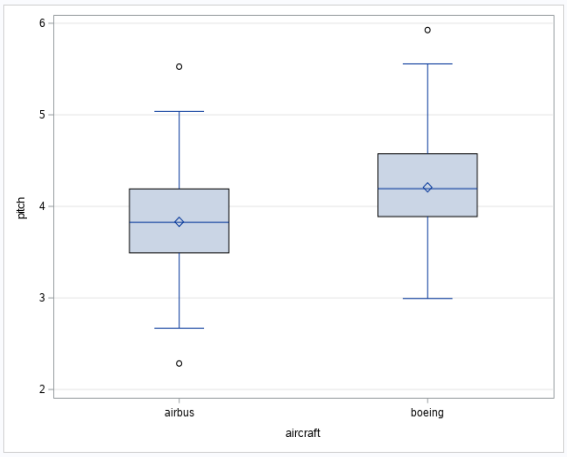

no_pasg
80
60
40
airbus
boeing
aircraft

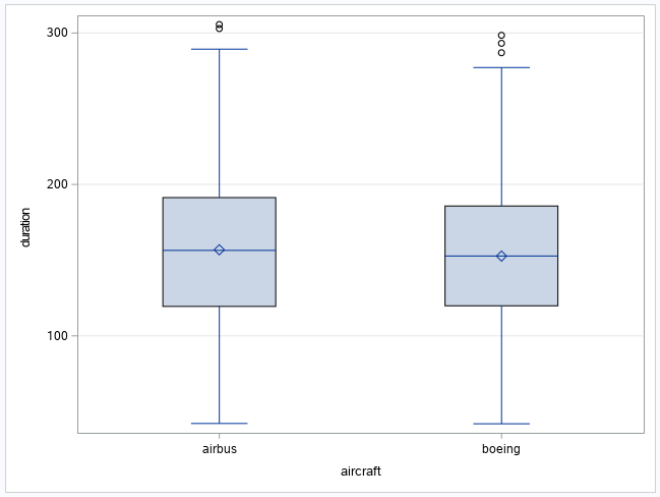

height
60
50
40
30
20
10
airbus
boeing
aircraft


distance
6000
5000
4000
3000
2000
1000
0
airbus
boeing
aircraft


speed_ground
125
100
75
50
25
airbus
boeing
aircraft


speed_air
140
130
120
110
100
90
airbus
boeing
aircraft


pitch
6
5
4
3
2
airbus
boeing
aircraft



Code to remove the outliers after observation:

```
DATA PROJECT1;
SET NODUP_FAA12_AIR;
IF NO_PASG  < 40  or NO_PASG  > 80 THEN DELETE;
IF HEIGHT > 56 THEN DELETE;
IF DISTANCE > 3060 and aircraft = 'airbus' THEN DELETE;
IF DISTANCE > 3790 AND AIRCRAFT = 'boeing' THEN DELETE;
IF SPEED_GROUND > 130 AND AIRCRAFT = 'airbus' THEN DELETE;
IF SPEED_GROUND > 136 AND AIRCRAFT = 'boeing' THEN DELETE;
IF SPEED_AIR > 120 AND AIRCRAFT = 'airbus'  THEN DELETE;
IF SPEED_AIR > 135 AND AIRCRAFT = 'boeing'  THEN DELETE;
IF (PITCH > 5.4  OR PITCH < 2.5) AND AIRCRAFT = 'airbus' THEN DELETE;
IF PITCH > 5.6 AND AIRCRAFT = 'boeing' THEN DELETE;
IF DURATION > 300 AND AIRCRAFT = 'airbus'  THEN DELETE;
IF DURATION > 280 AND AIRCRAFT = 'boeing'  THEN DELETE;

PROC PRINT DATA = PROJECT1;
RUN;
```

2. Step2: Plot to check the relation between landing distance and other variables:

```
PROC PLOT DATA = PROJECT1;
PLOT DISTANCE*NO_PASG;
PLOT DISTANCE*SPEED_GROUND;
PLOT DISTANCE*PITCH;
PLOT DISTANCE*DURATION;
PLOT DISTANCE*HEIGHT;
PLOT DISTANCE*AIRCRAFT;
PLOT DISTANCE*SPEED_AIR;
RUN;
```

Observation: The data was quite linear for speed_air and speed_ground and appeared scattered for plots of height and pitch with respect to the distance. Hence, I cannot make a judgement about how

correlated the variables are with respect to the landing distance. Hence, I will check for the exact correlation value, and its strength.

3. Step3: Correlation check of the variables with landing distance

Since there are two categories of aircraft, I will study them separately to understand which variables help us to study the variability in landing distance and if that depends on the aircraft type.

```
PROC CORR DATA = PROJECT1;
VAR DISTANCE NO_PASG SPEED_GROUND PITCH DURATION HEIGHT SPEED_AIR;
by aircraft;
TITLE CORRELATION WITH LANDING DISTANCE;
```

Here the results are based on the null hypothesis i.e. – there is no linear relationship between the variable x and the dependent variable y(landing distance).

For airbus:

CORRELATION WITH LANDING DISTANCE

The CORR Procedure

aircraft=airbus

7 Variables:	distance no_pasg speed_ground pitch duration height speed_air

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	418	1224	627.50980	511476	41.72231	3054	distance
no_pasg	418	60.27990	7.08703	25197	40.00000	80.00000	no_pasg
speed_ground	418	78.70034	15.45048	32897	33.57410	112.82932	speed_ground
pitch	418	3.83164	0.48251	1602	2.66891	5.03738	pitch
duration	370	156.19113	48.81362	57791	42.14623	289.32049	duration
height	418	30.49549	9.81285	12747	6.22752	54.27604	height
speed_air	67	100.87066	4.03916	6758	95.01136	110.56965	speed_air

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

	distance	no_pasg	speed_ground	pitch	duration	height	speed_air
distance distance	1.00000 418	-0.00085 0.9861 418	0.89699 <.0001 418	0.09311 0.0572 418	-0.07827 0.1329 370	0.14596 0.0028 418	0.84935 <.0001 67
no_pasg no_pasg	-0.00085 0.9861 418	1.00000 418	0.00743 0.8797 418	-0.07037 0.1510 418	-0.01367 0.7933 370	-0.00126 0.9795 418	-0.00235 0.9849 67
speed_ground speed_ground	0.89699 <.0001 418	0.00743 0.8797 418	1.00000 418	-0.00644 0.8956 418	-0.05654 0.2781 370	-0.07199 0.1417 418	0.92879 <.0001 67
pitch pitch	0.09311 0.0572 418	-0.07037 0.1510 418	-0.00644 0.8956 418	1.00000 418	-0.03511 0.5008 370	0.09267 0.0584 418	0.00815 0.9478 67
duration duration	-0.07827 0.1329 370	-0.01367 0.7933 370	-0.05654 0.2781 370	-0.03511 0.5008 370	1.00000 370	-0.00502 0.9233 370	0.07958 0.5456 60
height height	0.14596 0.0028 418	-0.00126 0.9795 418	-0.07199 0.1417 418	0.09267 0.0584 418	-0.00502 0.9233 370	1.00000 418	-0.10399 0.4024 67
speed_air speed_air	0.84935 <.0001 67	-0.00235 0.9849 67	0.92879 <.0001 67	0.00815 0.9478 67	0.07958 0.5456 60	-0.10399 0.4024 67	1.00000 67

Observation:

a. No_pasg: the p_value is quite high, hence we can't reject the null hypothesis. That implies there is no linear dependency between no. of passengers and landing distance.
b. Speed_ground : landing distance and speed of the ground are highly correlated.
c. Pitch: P value is close to 0.05, hence will check the impact by fitting in the model

d. Duration: the p_value is quite high, hence we can't reject the null hypothesis. That implies there is no linear dependency between duration and landing distance
e. Height: linear relationship exists between landing distance and height
f. Speed_air: linear relationship exists between landing distance and speed of the air

Action: As per this obersation, I believe the landing distance is dependent on the all the variables except for no_pasg and duration of flight. Also, the speed_ground and speed_Air are highly correlated, and thus any one of the variable will be to explain the variability. I would consider only one of the two while modelling. The data for speed_air is very thin, hence I would consider speed_groud for my analysis.

For boeing:

CORRELATION WITH LANDING DISTANCE

The CORR Procedure

aircraft=boeing

7 Variables:	distance no_pasg speed_ground pitch duration height speed_air

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	361	1599	717.81386	577150	573.62179	3786	distance
no_pasg	361	59.90859	7.45840	21627	42.00000	80.00000	no_pasg
speed_ground	361	76.18530	18.81584	27503	30.00000	117.64059	speed_ground
pitch	361	4.21324	0.47269	1521	3.06891	5.55640	pitch
duration	361	150.99268	45.64796	54508	41.94937	270.59595	duration
height	361	30.17710	9.42973	10894	7.58249	55.09351	height
speed_air	97	99.18823	6.98039	9621	90.00286	113.86394	speed_air

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

	distance	no_pasg	speed_ground	pitch	duration	height	speed_air
distance distance	1.00000 361	-0.01502 0.7762 361	0.88469 <.0001 361	-0.01593 0.7629 361	-0.06105 0.2473 361	0.10461 0.0470 361	0.94429 <.0001 97
no_pasg no_pasg	-0.01502 0.7762 361	1.00000 361	0.01379 0.7940 361	0.09753 0.0642 361	-0.05656 0.2838 361	0.09443 0.0731 361	0.00440 0.9659 97
speed_ground speed_ground	0.88469 <.0001 361	0.01379 0.7940 361	1.00000 361	-0.01735 0.7425 361	-0.09234 0.0797 361	-0.08722 0.0980 361	0.97777 <.0001 97
pitch pitch	-0.01593 0.7629 361	0.09753 0.0642 361	-0.01735 0.7425 361	1.00000 361	-0.03008 0.5689 361	0.03110 0.5559 361	0.17738 0.0822 97
duration duration	-0.06105 0.2473 361	-0.05656 0.2838 361	-0.09234 0.0797 361	-0.03008 0.5689 361	1.00000 361	0.04988 0.3447 361	-0.01445 0.8883 97
height height	0.10461 0.0470 361	0.09443 0.0731 361	-0.08722 0.0980 361	0.03110 0.5559 361	0.04988 0.3447 361	1.00000 361	-0.18223 0.0740 97
speed_air speed_air	0.94429 <.0001 97	0.00440 0.9659 97	0.97777 <.0001 97	0.17738 0.0822 97	-0.01445 0.8883 97	-0.18223 0.0740 97	1.00000 97

Observation:

a. No_pasg: the p_value is quite high, hence we can't reject the null hypothesis. That implies there is no linear dependency between no. of passengers and landing distance.
b. Speed_ground : landing distance and speed of the ground are highly correlated.
c. Pitch: P value is very high, hence we can't reject the null hypothesis. That implies there is no linear dependency between pitch and landing distance
d. Duration: the p_value is quite high, hence we can't reject the null hypothesis. That implies there is no linear dependency between duration and landing distance
e. Height: linear relationship exists between landing distance and height
f. Speed_air: linear relationship exists between landing distance and speed of the air

The observations are almost similar to that of airbus. However, in case of boeing, pitch is being insignificant.

Chapter 3: Building a model

Objective: building a model to describe the relation between the dependent variable and several independent variables.

Analysis:

1. Step1: Building a model

I will perform linear regression here to understand the variability in Y (landing distance) which is caused by the other independent variables. I will study the model separately for the two aircrafts.

There are certain assumptions that are made for the error terms, i.e. they are independednt of each other, noramlly distributed along the regression line, mean is 0 and the variance is constant. We will study the same once we have our model to check it it holds true.

For Airbus:

```
PROC REG DATA = WORK.PROJECT1 (where=(aircraft='airbus'));
 MODEL DISTANCE =  SPEED_GROUND
 HEIGHT PITCH / VIF;
 TITLE REGRESSION ANALYSIS FOR AIRBUS
 RUN;
```

REGRESSION ANALYSIS FOR AIRBUS RUN

Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	418
Number of Observations Used	418

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	140475673	46825224	817.07	<.0001
Error	414	23725814	57309		
Corrected Total	417	164201487			

Root MSE	239.39242	R-Square	0.8555
Dependent Mean	1223.62675	Adj R-Sq	0.8545
Coeff Var	19.56417		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-2489.28603	116.13264	-21.43	<.0001	0
speed_ground	speed_ground	1	37.04876	0.76073	48.70	<.0001	1.00521
height	height	1	13.05879	1.20293	10.86	<.0001	1.01387
pitch	pitch	1	104.11426	24.40091	4.27	<.0001	1.00866

Here, the null hypothesis is - none of the independent variables have a significant relationship with the dependent variable, and alternate hypothesis is – atleast one of the variable has a significant relationship with distance.
ANOVA only tells us there are dependent or not, hence we look at the t-value provided by Parameter estimates to understand how dependent or independent it is. Also, if the p-value is <0.001, it means it has a significant relationship with the dependent variable.

Observation:

The value of R^2 is 85% , which is closer to the adjusted R^2 value. The adjusted R^2 also plays an important role in case of overfitting. Overfitting happens when we force too many variables into the model, which may be insignificant and will inflate R^2.

Thus, we can write the regression line equation as –
Landing distance = -2489.29 + 37.04 (speed_ground) + 13.06 (height) + 104.11 (pitch)

For Boeing:

```
PROC REG DATA = WORK.PROJECT1 (where=(aircraft='boeing'));
 MODEL DISTANCE =  SPEED_GROUND
 HEIGHT  / VIF;
 TITLE REGRESSION ANALYSIS FOR BOEING
 RUN;
```

REGRESSION ANALYSIS FOR BOEING RUN

Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	361
Number of Observations Used	361

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	151357641	75678820	793.71	<.0001
Error	358	34134783	95349		
Corrected Total	360	185492423			

Root MSE	308.78561	R-Square	0.8160
Dependent Mean	1598.75417	Adj R-Sq	0.8149
Coeff Var	19.31414		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-1439.74586	89.30925	-16.12	<.0001	0
speed_ground	speed_ground	1	34.36005	0.86824	39.57	<.0001	1.00767
height	height	1	13.94333	1.73247	8.05	<.0001	1.00767

Observation:

The speed of the ground and height of aircraft play a significant impact on the landing distance. Also, the pitch doesn't play a role here like in case of airbus.
The adjusted R^2 and R^2 are close to each other, thus there is no overfitting of the model. We also study the variance inflation which is nothing but the inverse of R^2 and it ranges from 1 to infinity. It gives us an idea of overfitting. If the value is greater than 5, we are overfitting and we need to check for the variables stuffed in the model.

Thus, we can write the regression line equation as –
Landing distance = -1439.75 + 34.36 (speed_ground) + 14 (height)

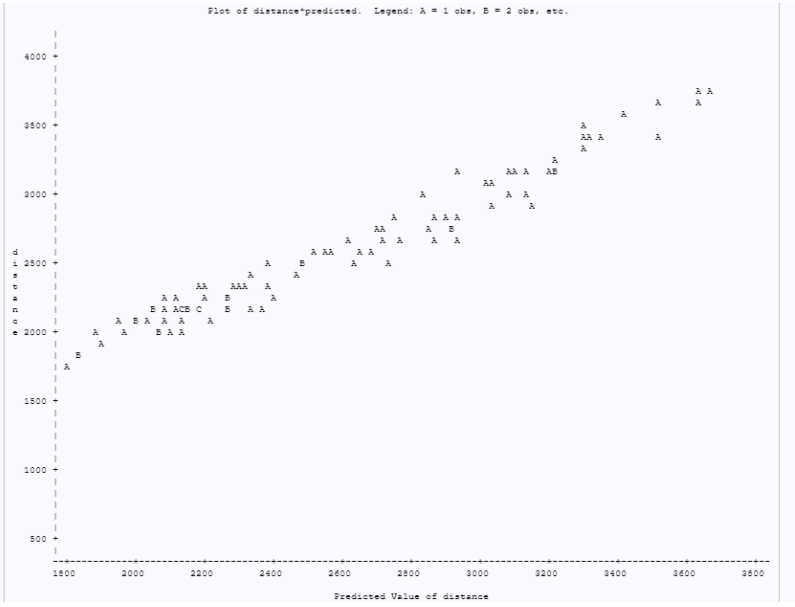
Aircraft Type	Total Obs	Significant independent variables	Reggresion Value
Airbus	418	Speed_ground, height, pitch	0.8555
Boeing	361	Speed_ground, height	0.816

2. Step2: Model checking

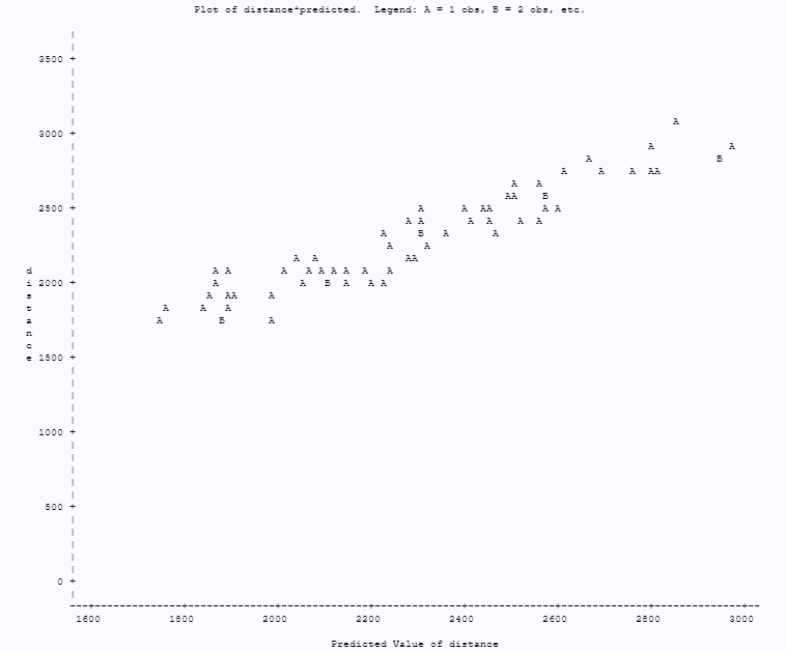
Once we have the model, we have to check if the initial assumptions made were holding true or not.

a. landing distance and predicted value obtained should be highly correlated

For Boeing,



For Airbus,



b. means of the residual should be 0

For Boeing

The MEANS Procedure

Analysis Variable : RESIDUAL Residual				
N	Mean	Std Dev	Minimum	Maximum
97	5.438216E-13	104.3367390	-247.4681185	218.5670429

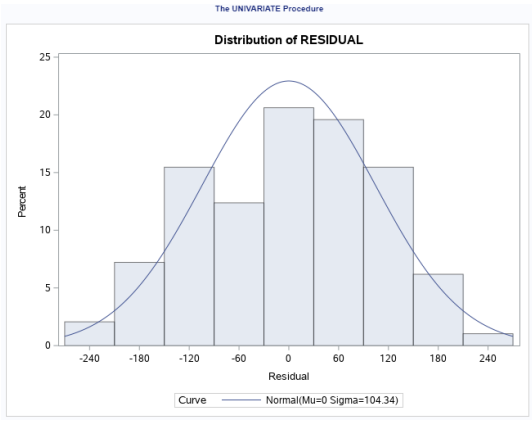
For Airbus

The MEANS Procedure

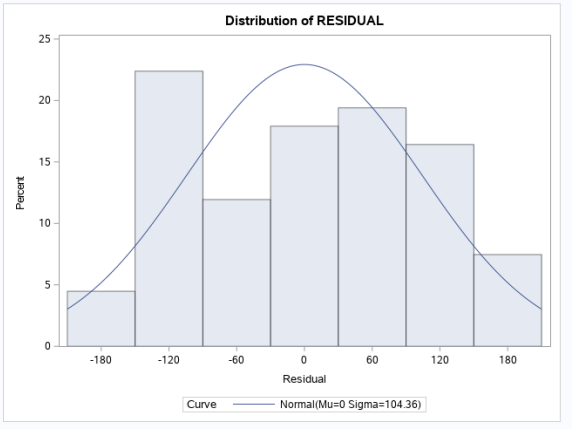
Analysis Variable : RESIDUAL Residual				
N	Mean	Std Dev	Minimum	Maximum
67	-1.20474E-12	104.3639918	-198.3028573	200.6466524

c. The residual should be noramlly distributed

For Boeing



For Airbus



Questions:

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

I have used 734 observations in my final model- 361 (Boeing) and 418 (Airbus) comapred to the original given 950 records.

a. One of the files had missing duration of flight information, and when observed closely, we see that there were 100 duplicate records
b. There were threshold values provided for each of the variables- hence, removed 8 observations where the height threshold wasn't met. Also, the duration of the flight was less than 40 minutes for 7 observations, which was removed
c. Once I had completed the initial cleaning, I further removed the outliers by plotting the box plot for each of the variables. This step was taken so that as to carry out the regression and find the best fit line. In case of outliers, the fit line tends to move in the direction of outliers and hence, increases the error value. Then I was left with 779 observations, and 8 variables
d. Finally, after observing the values of correlation, I observed that not all the variables were significant to the landing distance and hence ended up using 734 observations to fit the model

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

Since, we were provided with two different flight types with no information regarding how different they are from each other, I have observed them separately.
After studying the correlation values, we see that in case of

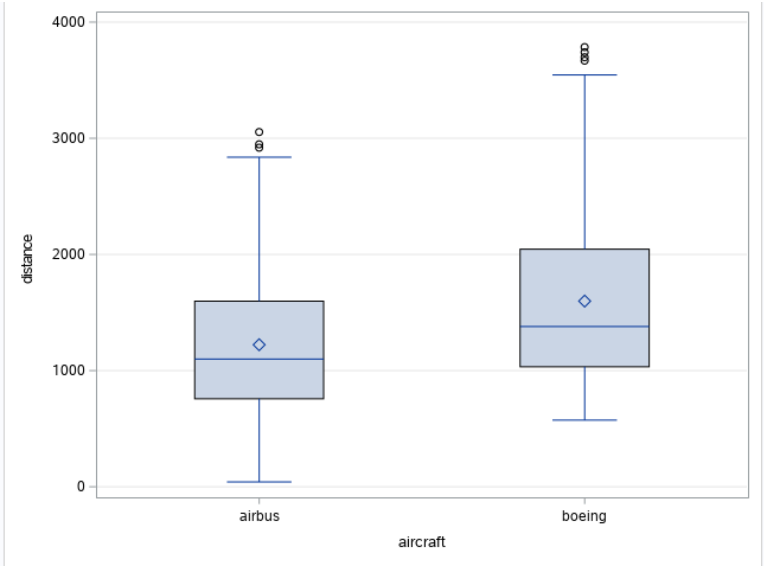
a. Airbus: speed of the ground and air, pitch and height play a significant role in determining the landing distance. We also see that the speed of ground and that of air are highly correlated and dependent on each other, which means any one of them will be able to explain the variability in landing distance. Since the data for speed of air was sparse, we can consider speed of the ground instead. Also, the landing distance is independent of the duration of the flight and the number of passengers onboarded.

   Once we have the dependent variables and how they fit the model, we see that the pitch, height and speed of the ground have positive impact on the landing distance

b. Boeing: we observe results similar to that of airbus, however in case of boeing aircraft, landing distance is independent of the pitch of the aircraft besides flight duration and number of passengers. However, the height and speed of the ground has a positive impact on the landing distance of the flight.

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

Yes the make of the aircraft are different.
Firstly, when we check the plots for both aircrafts with distance, we observe that the means of both the aircrafts are different, hence we consider them different while applying regression modelling.



Secondly, When we fit the model, we see that pitch of Boeing doesn't impact the landing distance whereas in case of airbus, pitch plays a significant role in determining the landing distance.