Mid Term Project
Aniket Mahurkar

Project Summary

Background: Flight landing.

Motivation: To reduce the risk of landing overrun.

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

Data: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models).

Variables:

- 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.

Chapter 1: Data Preparation

Goal:
The data preparation step aims at cleaning the raw dataset by treating the unwanted records and missing values. The step covers all activities required to construct the final data set that can be used for modelling.

1. Combining data sets from different sources

- Two data sets were read from excel sheets
- Excel sheets are converted into sas data sets faa1 and faa2

```
proc import out= faa1 datafile= "/home/mahurkav0/sasuser.v94/FAA1.xls"
        DBMS=xls replace;
    sheet="FAA1";
    getnames=yes;
run;

proc import out= faa2 datafile= "/home/mahurkav0/sasuser.v94/FAA2.xls"
        DBMS=xls replace;
    sheet="FAA2";
    getnames=yes;
run;
```

- The created sas datasets are cross checked for number of records, number of variables and format of variables

```
proc contents data=faa1;
run;

proc contents data=faa2;
run;
```

The CONTENTS Procedure

Data Set Name	WORK.FAA1	Observations	800
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	01/24/2017 04:43:37	Observation Length	72
Last Modified	01/24/2017 04:43:37	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)		

The CONTENTS Procedure

Data Set Name	WORK.FAA2	Observations	200
Member Type	DATA	Variables	7
Engine	V9	Indexes	0
Created	01/24/2017 04:43:37	Observation Length	64
Last Modified	01/24/2017 04:43:37	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)		

Observation: Data set faa2 has 50 missing records.
Conclusion: These 50 missing records needs to be removed for correct analysis of data

- Missing records from faa2 are cleared

```
data cleanfaa2;
set faa2;
if nmiss(of _numeric_) and cmiss(of _character_) > 0 then delete;
run;
```

- Data Sets are **concatenated** to create a combined data set combined_airlines for analysis

```
data combined_airlines;
set faa1
cleanfaa2;
run;
```

- The contents of dataset combined_airlines are checked for number of records and number of variables

```
proc contents data=combined_airlines;
run;
```

The CONTENTS Procedure

Data Set Name	WORK.COMBINED_AIRLINES	Observations	950
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	01/24/2017 04:49:26	Observation Length	72
Last Modified	01/24/2017 04:49:26	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)		

Observation: Concatenated data set has 950 records and 8 variables. one extra variable present only in faa1 is missing for 150 records of cleanfaa2.

Conclusion: There are few records which seem to be same in both the excel sheets, so we should keep only one record

- Checking common records

```
proc sort data=faa1 out=sort_faa1;
by aircraft       no_pasg       speed_ground speed_air height pitch    distance;
run;

proc sort data=cleanfaa2 out=sort_cleanfaa2;
by aircraft       no_pasg       speed_ground speed_air height pitch    distance;
run;


data common_airlines;
merge sort_faa1 (in=f1) sort_cleanfaa2 (in=f2) ;
by aircraft       no_pasg       speed_ground speed_air height pitch    distance;
if f1 and f2;
run;
```

The CONTENTS Procedure

Data Set Name	WORK.COMMON_AIRLINES	Observations	100
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	01/24/2017 15:06:02	Observation Length	72
Last Modified	01/24/2017 15:06:02	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)		

Observation: There are 100 records that are common
Conclusion: Remove these common records

- Removing common records from sorted data sets

```
data merged_airlines;
merge sort_faa1 (in=f1) sort_cleanfaa2 (in=f2) ;
by aircraft        no_pasg          speed_ground speed_air height pitch     distance;
run;
```

- Checking contents of merged data

```
proc contents data= merged_airlines;
run;
```

The CONTENTS Procedure

Data Set Name	WORK.MERGED_AIRLINES	Observations	850
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	01/24/2017 15:08:58	Observation Length	72
Last Modified	01/24/2017 15:08:58	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)		

Observations: there are 850 required records
Conclusion: The merged data set removed the common 100 records

2. Performing the completeness check of each variable – examining presence of missing values

- The dataset merged_airlines is checked for analyzing formats of variables

```
proc contents data=merged_airlines;
run;
```

Alphabetic List of Variables and Attributes

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

- Missing values are checked for each variable along with its mean, median and number of records

```
proc means data=merged_airlines nmiss;
run;
```

The MEANS Procedure

Variable	Label	N Miss	N	Mean	Median	Std Dev	Skewness
duration	duration	50	800	154.0065385	153.9480975	49.2592338	0.1214794
no_pasg	no_pasg	0	850	60.1035294	60.0000000	7.4931370	-0.0215023
speed_ground	speed_ground	0	850	79.4523229	79.6428041	19.0594903	0.1178254
speed_air	speed_air	642	208	103.7977237	101.1473493	10.2590370	1.0564046
height	height	0	850	30.1442223	30.0931324	10.2877268	-0.0956784
pitch	pitch	0	850	4.0093577	4.0082875	0.5288298	0.0061541
distance	distance	0	850	1526.02	1258.09	928.5600816	1.6349388

Observation: There are 50 missing values of variable duration, and there are 642 missing values of variable speed_air

Conclusion: The two variables contain a large amount of missing values which will impact the analysis.

3. Performing the validity check of each variable – examining presence of abnormal values

- Basis the definition of abnormal values in the project description, abnormal values are removed from the data
  - If the speed_ground value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
  - The landing aircraft is required to be at least 6 meters in height
  - The length of the airport runway is typically less than 6000 feet.
  - Duration should always be greater than 40 mins (we should not consider the missing values from 150 cleanfaa2 records)

```
data abnormal_merged_airlines;
set merged_airlines;
if (speed_ground <30 or speed_ground>140)
or (speed_air ne . and speed_air <30 or speed_air>140)
or height<6 or distance>6000 or (duration<40 and duration ne .);
run;

proc print data=abnormal_merged_airlines;
run;
```

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

Observation: 19 records are abnormal per the definition of abnormal data
Conclusion: **We should remove these 19 records** since these can have incorrect readings that can dilute our analysis

4. **Cleaning the data based on the results of Steps 2 and 3;**

- Based on results of 2 and 3 following cleaning needs to be done
  - Remove 19 abnormal records

```
data clean1_merged_airlines;
set merged_airlines;
if (speed_ground <30 or speed_ground>140)
or (speed_air ne . and speed_air <30 or speed_air>140)
or height<6 or distance>6000 or (duration<40 and duration ne .)then delete;
run;
```

- Checking the number of records in clean2_merged_airlines

The CONTENTS Procedure

Data Set Name	WORK.CLEAN1_MERGED_AIRLINES	Observations	831
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	02/12/2017 03:06:51	Observation Length	72
Last Modified	02/12/2017 03:06:51	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)		

Alphabetic List of Variables and Attributes

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

Observation: The clean1_combined_airlines dataset contains (850- 19 abnormal records) 831 records and 8 variables
Conclusion: This data is clean enough to analyze distribution of each variable

5. **Summarizing the distribution of each variable**

```
proc means data=clean1_merged_airlines n mean min max nmiss median stddev skewness;
run;
```

The MEANS Procedure

Variable	Label	N	Mean	Minimum	Maximum	N Miss	Median	Std Dev	Skewness
duration	duration	781	154.7757191	41.9493694	305.6217107	50	154.2845505	48.3499237	0.1898657
no_pasg	no_pasg	831	60.0553550	29.0000000	87.0000000	0	60.0000000	7.4913166	-0.0135746
speed_ground	speed_ground	831	79.5426997	33.5741041	132.7846766	0	79.7939604	18.7356754	0.0889029
speed_air	speed_air	203	103.4850352	90.0028586	132.9114649	628	101.1189240	9.7362774	0.8827269
height	height	831	30.4578695	6.2275178	59.9459639	0	30.1670844	9.7848114	0.1271445
pitch	pitch	831	4.0051609	2.2844801	5.9267842	0	4.0010380	0.5265690	0.0173051
distance	distance	831	1522.48	41.7223127	5381.96	0	1262.15	896.3381524	1.4763958

Observation:

- This gives a snapshot of all the numerical variables that we will be using for our analysis

The data is combined with two types of aircrafts: Boieng and Airbus
We check the distribution of our variables for these two types of aircrafts

```
proc means data=clean1_merged_airlines n mean min max nmiss median stddev skewness
maxdec=2;
by aircraft;
run;
```

The MEANS Procedure

aircraft=airbus

Variable	Label	N	Mean	Minimum	Maximum	N Miss	Median	Std Dev	Skewness
duration	duration	394	156.90	42.15	305.62	50	156.45	49.19	0.13
no_pasg	no_pasg	444	60.21	36.00	87.00	0	60.00	7.43	0.03
speed_ground	speed_ground	444	80.25	33.57	131.04	0	81.17	16.95	0.06
speed_air	speed_air	85	104.31	95.01	131.34	359	101.35	8.09	1.30
height	height	444	30.59	6.23	58.23	0	30.35	9.85	0.08
pitch	pitch	444	3.83	2.28	5.53	0	3.83	0.50	0.01
distance	distance	444	1323.32	41.72	4896.29	0	1126.89	791.93	1.39

aircraft=boeing

Variable	Label	N	Mean	Minimum	Maximum	N Miss	Median	Std Dev	Skewness
duration	duration	387	152.61	41.95	298.52	0	152.73	47.45	0.25
no_pasg	no_pasg	387	59.87	29.00	82.00	0	60.00	7.57	-0.05
speed_ground	speed_ground	387	78.73	33.82	132.78	0	78.77	20.58	0.14
speed_air	speed_air	118	102.89	90.00	132.91	269	100.88	10.76	0.82
height	height	387	30.31	7.58	59.95	0	29.84	9.71	0.18
pitch	pitch	387	4.20	2.99	5.93	0	4.19	0.49	0.06
distance	distance	387	1750.98	573.62	5381.96	0	1470.78	953.85	1.50

Observation: The 831 records are split amongst boeig and airbus as 387 and 444 respectively.

Chapter 2: Descriptive Study

Goal:
Descriptive study is a part of data understanding. It helps us to visualize the variation of data. The data visualization step aims at studying the distribution graph of variables in our data. It is also useful to study the variation of the response variable, distance against all predictor variables.

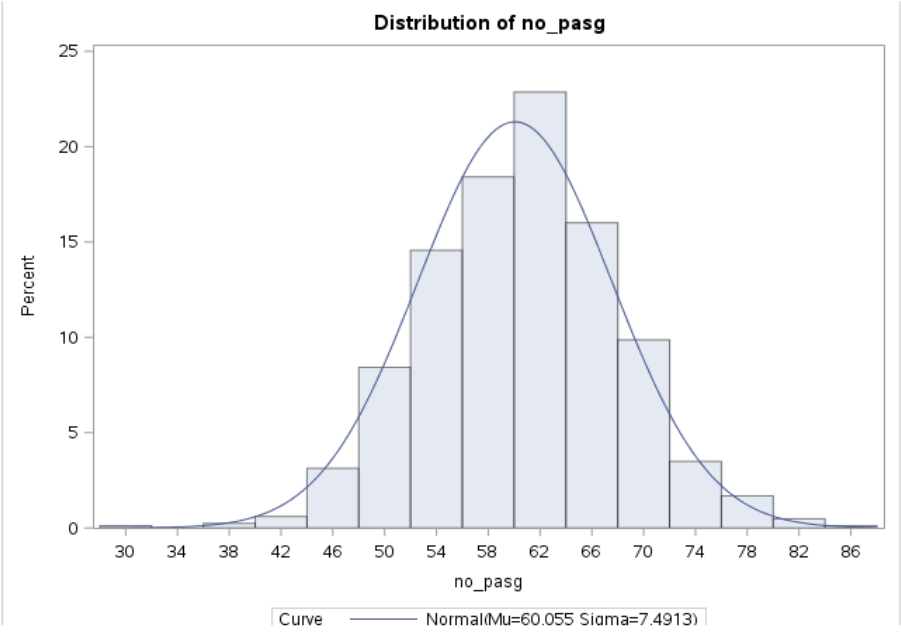
1. **Distribution study of each variable:**

- **Number of Passengers:**

```
proc univariate data=clean1_merged_airlines;
var no_pasg;
histogram no_pasg/normal;
run;
```

The UNIVARIATE Procedure
Variable: no_pasg (no_pasg)

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



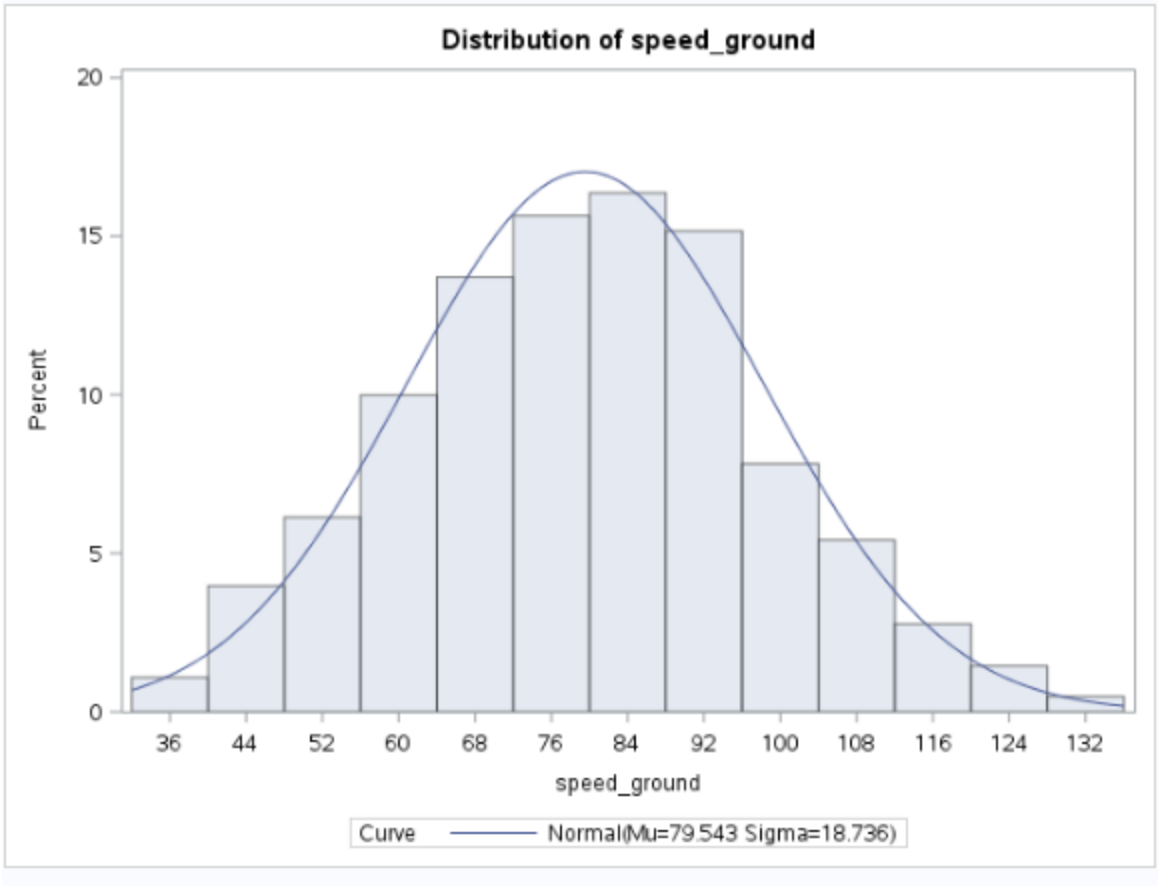
Observations and Conclusions: No_pasg variable is normally distributed

- **Speed_ground**

```
proc univariate data=clean1_merged_airlines;
var speed_ground;
histogram speed_ground /normal;
run;
```

The UNIVARIATE Procedure
Variable: speed_ground (speed_ground)

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



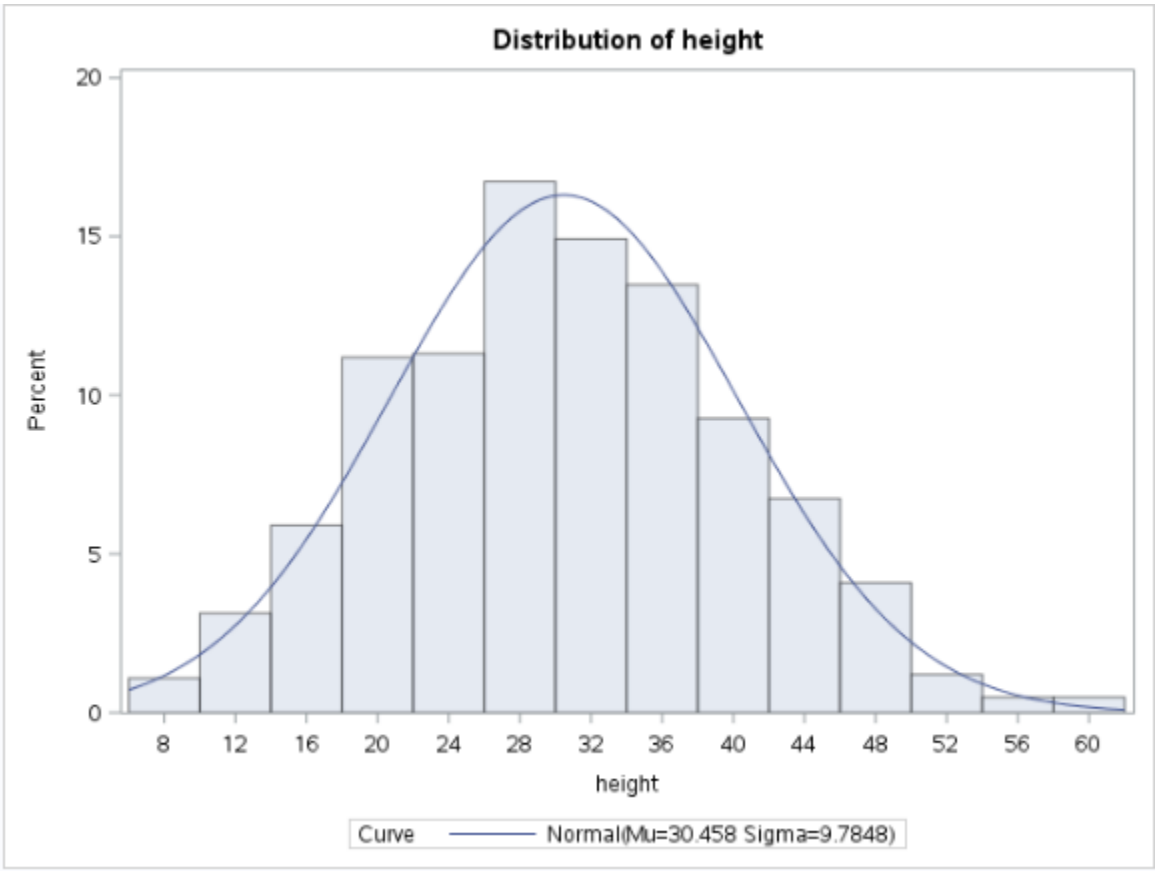
Observations and Conclusions: Speed_ground variable is normally distributed

- **Height:**

```
proc univariate data=clean1_merged_airlines;
var height;
histogram height /normal;
run;
```

The UNIVARIATE Procedure
Variable: height (height)

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



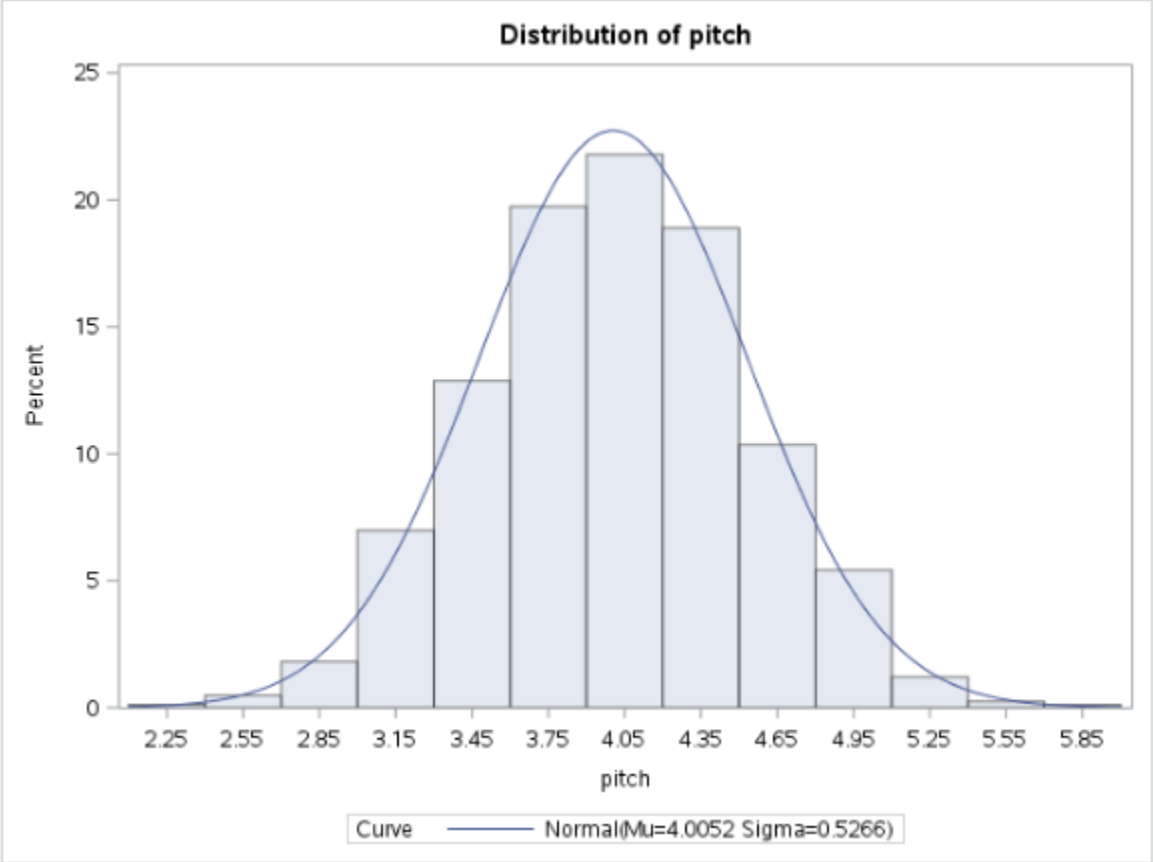
Observations and Conclusions: The graph is linearly distributed and has a positive skewness

- **Pitch**

```
proc univariate data=clean1_merged_airlines;
var pitch;
histogram pitch /normal;
run;
```

The UNIVARIATE Procedure
Variable: pitch (pitch)

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



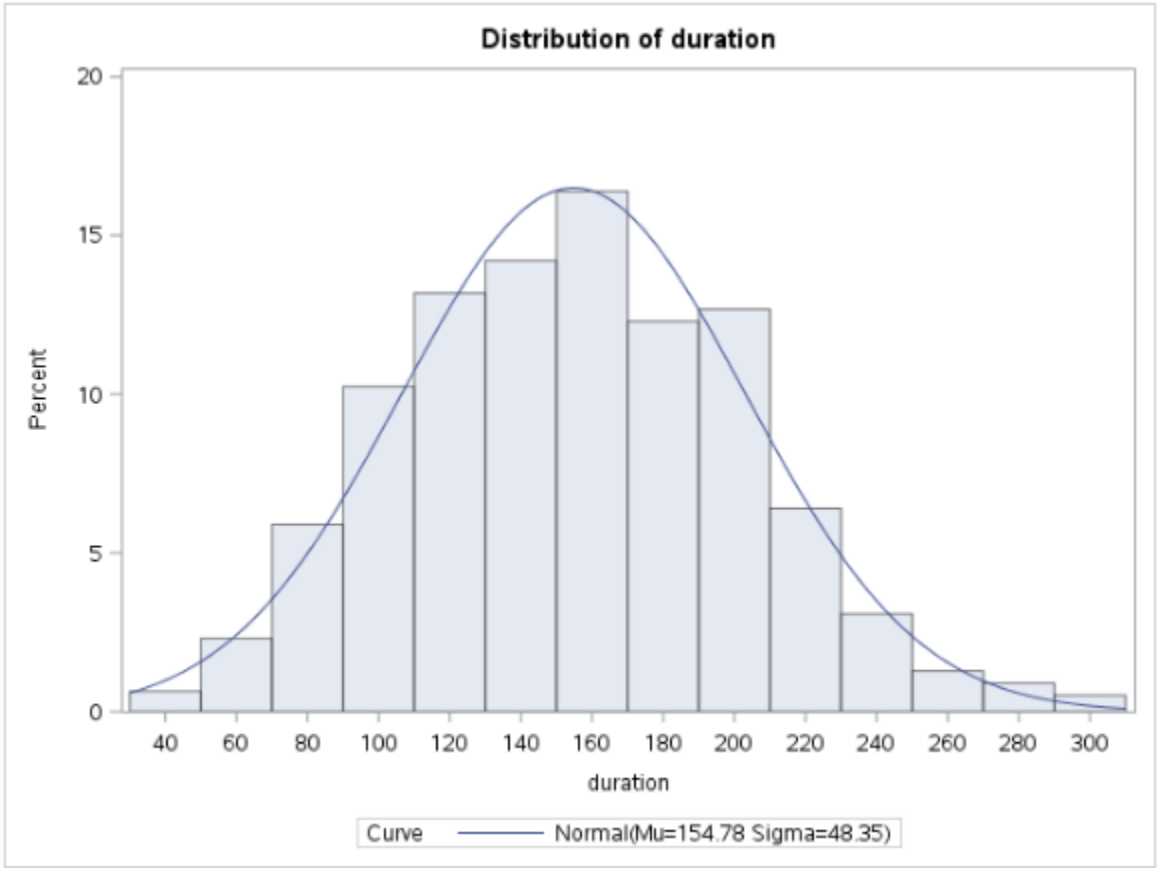
Observations and Conclusions: The graph is normally distributed

- **Duration:**

```
proc univariate data=clean1_merged_airlines;
var pitch;
histogram pitch /normal;
run;
```

The UNIVARIATE Procedure
Variable: duration (duration)

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

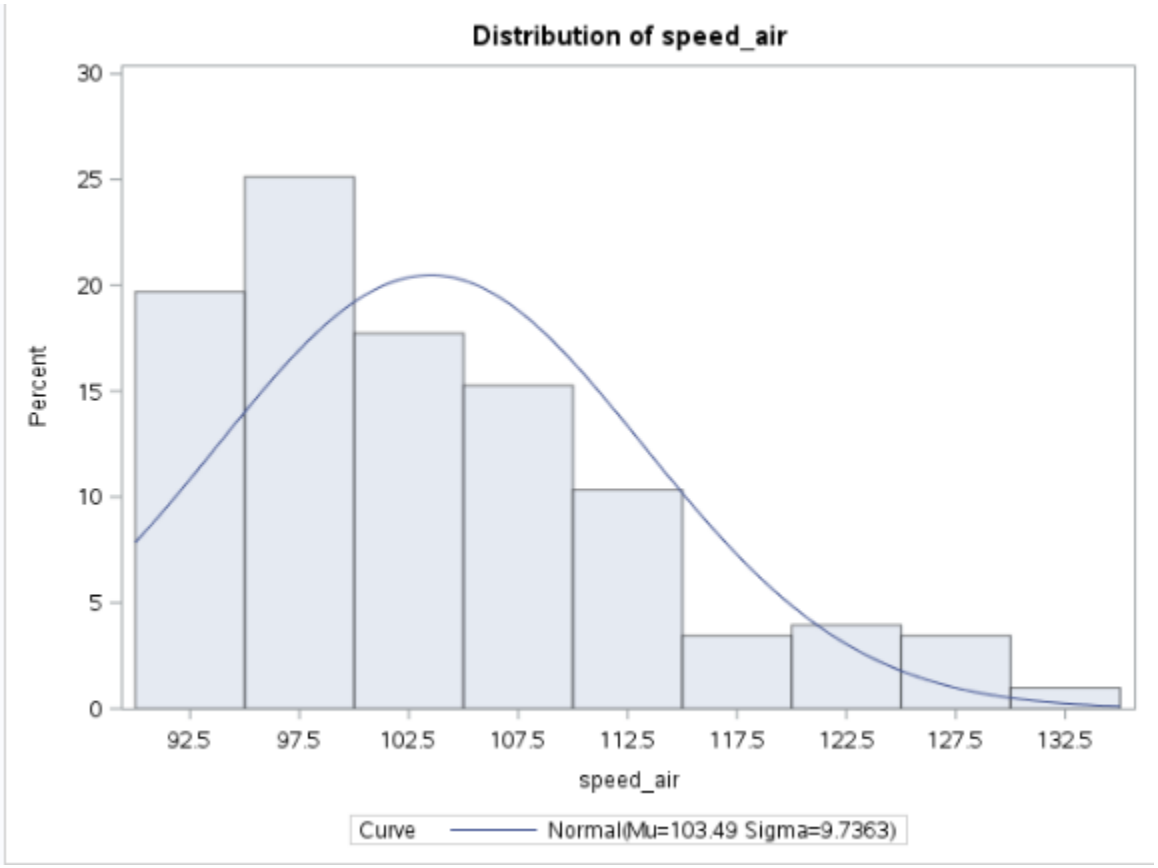


Observations and Conclusions: The graph is normally distributed

- **Speed_air**

The UNIVARIATE Procedure
Variable: speed_air (speed_air)

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



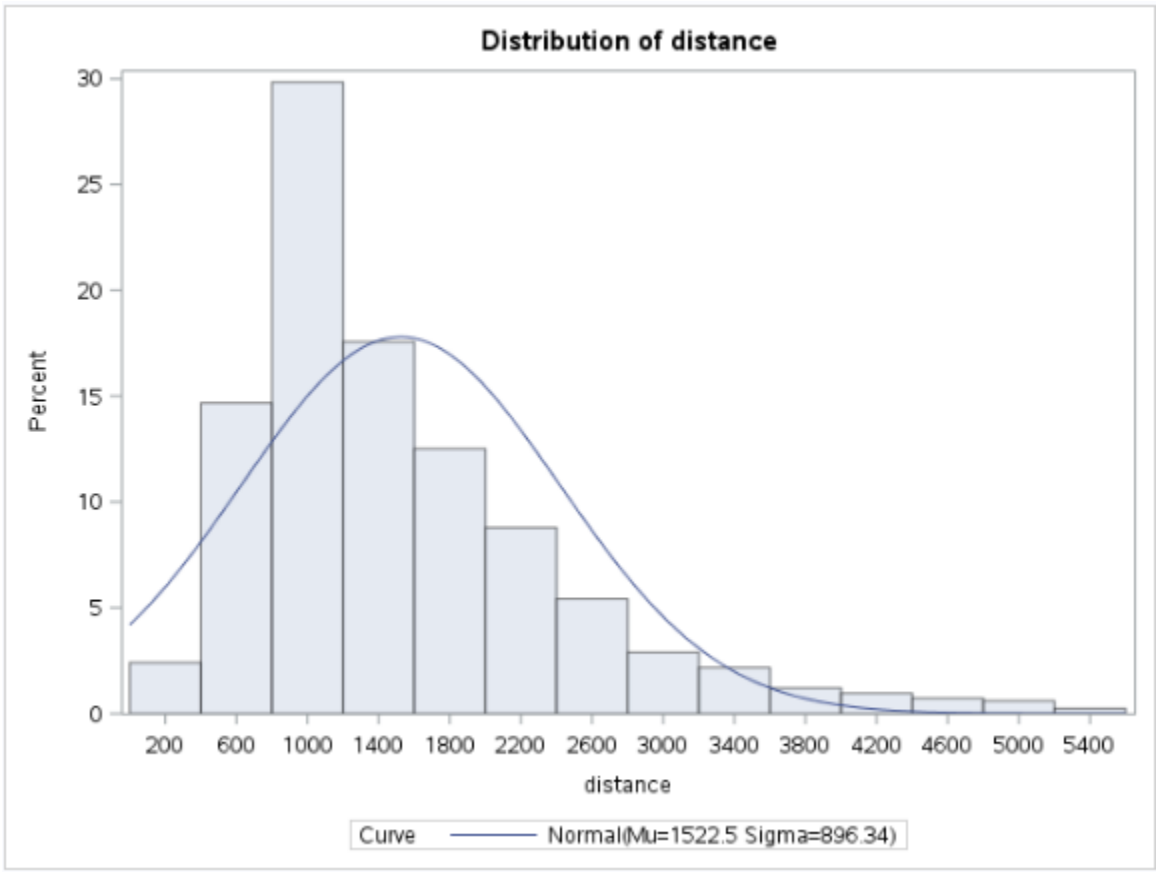
Observations and Conclusions: Due to a lot of missing values the graph appears to be clipped from the left side, otherwise the variable is normally distributed.

- **Distance**:

```
proc univariate data=clean1_merged_airlines;
var distance;
histogram distance /normal;
run;
```

The UNIVARIATE Procedure
Variable: distance (distance)

Moments			
N	831	Sum Weights	831
Mean	1522.48287	Sum Observations	1265183.27
Std Deviation	896.338152	Variance	803422.083
Skewness	1.47639585	Kurtosis	2.54813164
Uncorrected SS	2593060185	Corrected SS	666840329
Coeff Variation	58.8734473	Std Error Mean	31.093626



The UNIVARIATE Procedure
Fitted Normal Distribution for distance (distance)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	1522.483
Std Dev	Sigma	896.3382

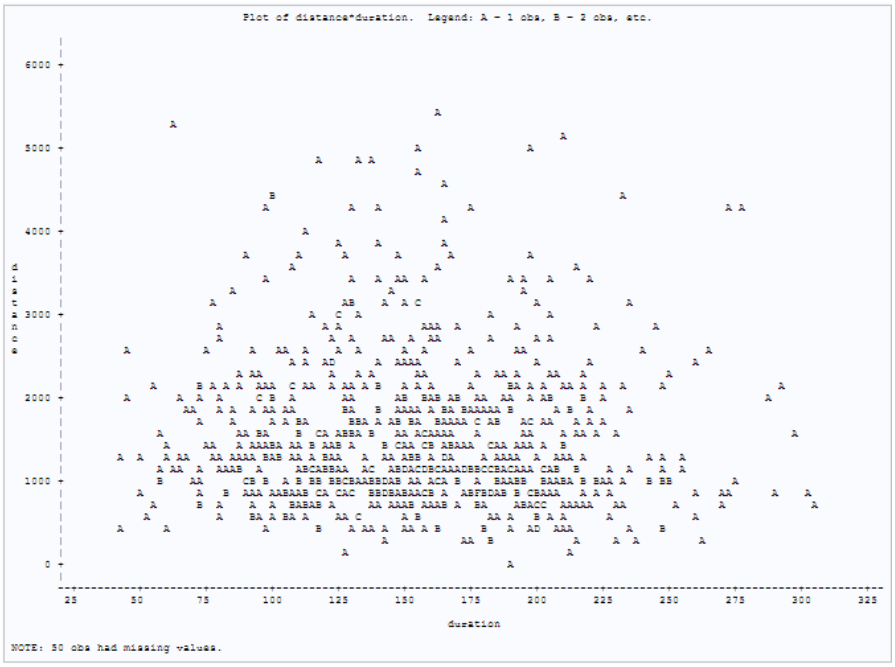
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.1170655	Pr > D	<0.010
Cramer-von Mises	W-Sq	4.3765851	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	26.0814686	Pr > A-Sq	<0.005

Observations and Conclusions:

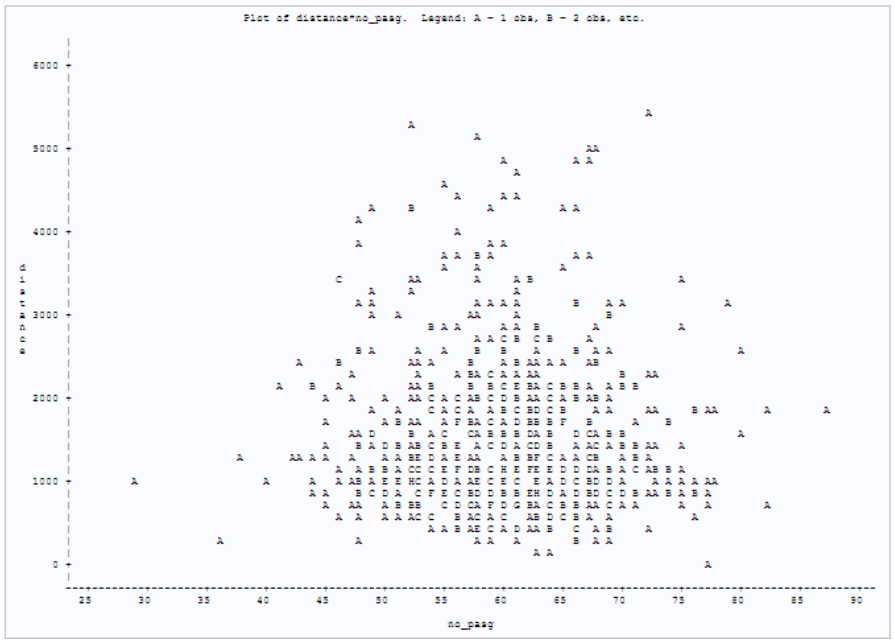
- The graph is <u>not</u> linearly distributed and appears to be lognormal.

2. <u>Study the variation of response variable, distance, against all predictor variables</u>

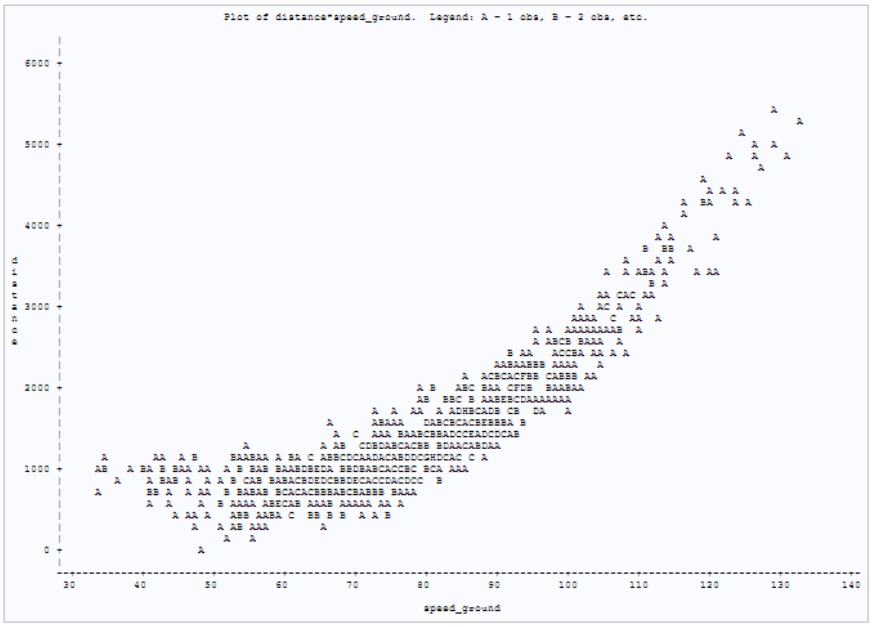
```
proc plot data=clean1_merged_airlines;
plot distance*duration;
plot distance * no-pasg;
plot distance *speed_ground;
plot distance * speed_air;
plot distance * height;
plot distance * pitch;
run;
```



Distance vs Duration: The distribution is not strongly correlated



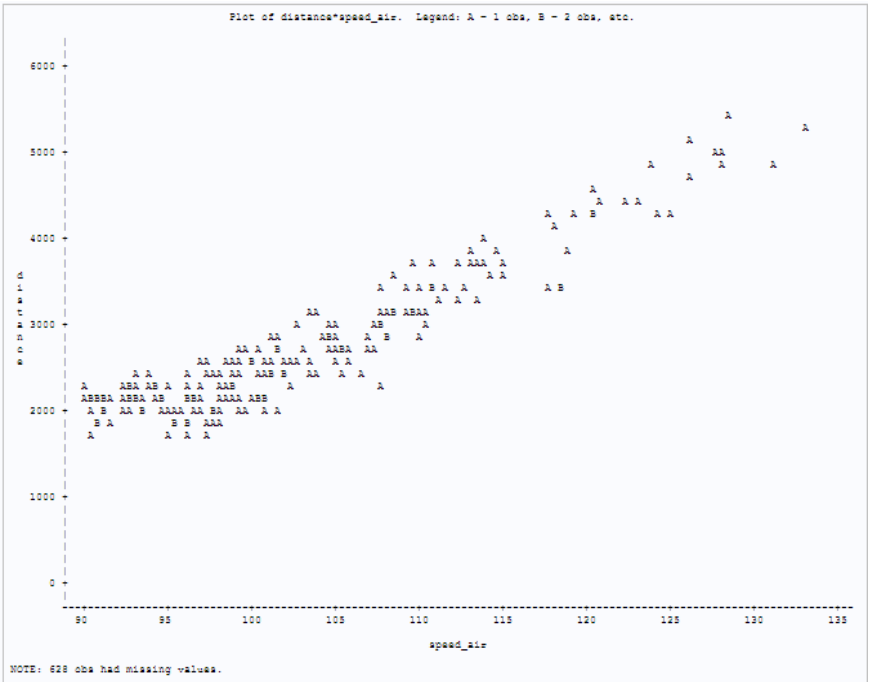
Distance vs No_pasg: Distance is not strongly correlated to number of passenger.



Distance vs Speed_ground:

Distance is highly correlated.

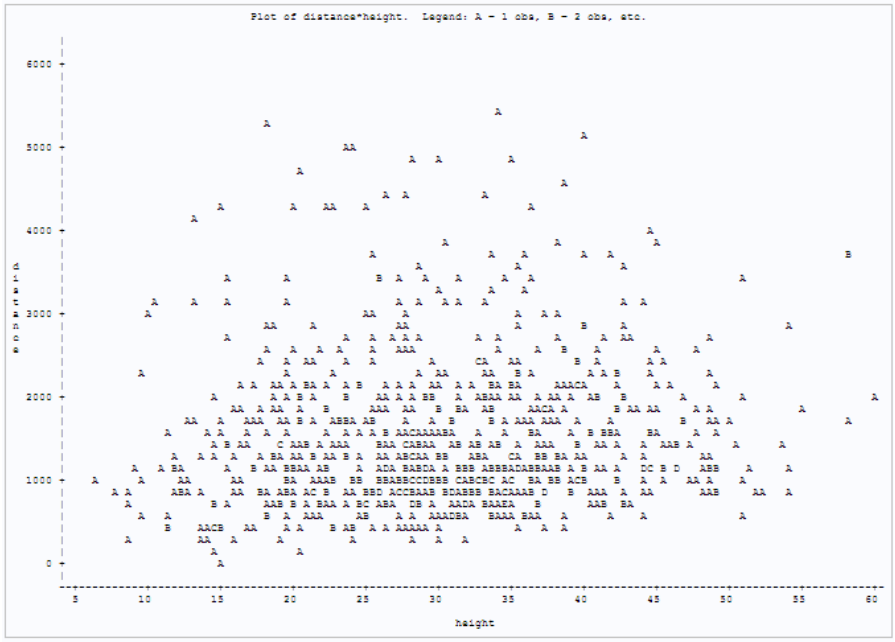
Correlation is not linear



Distance vs Speed_air

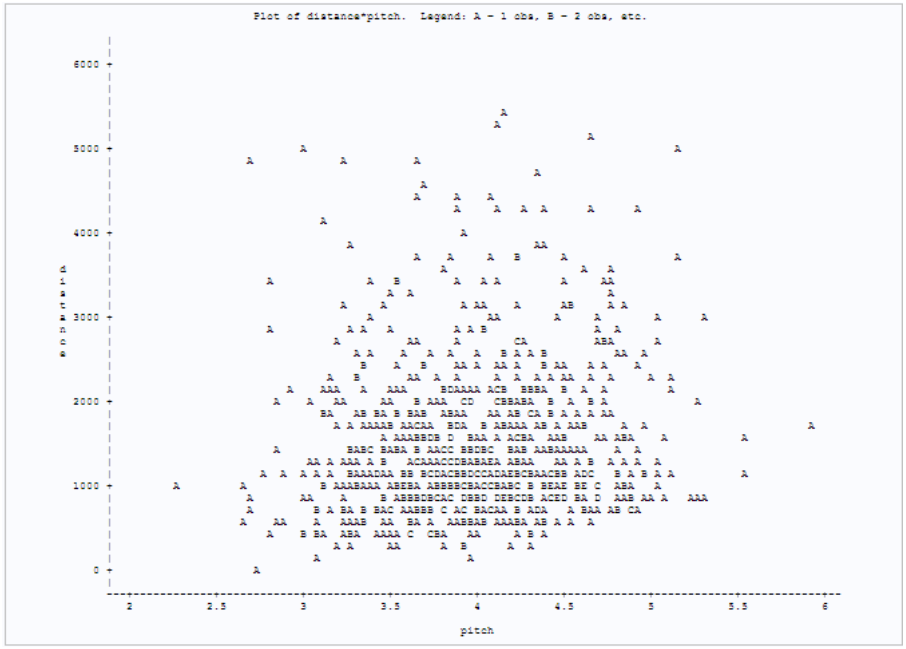
Distance is highly correlated to speed_air.

Correlation is not linear.



Distance vs pitch

Distance is not highy correlated to height



Distance vs pitch

Distance is not highly correlated to pitch

3. Study the correlation using Pearson correlation coefficients

- We further check correlation between numeric variables

```
proc corr data = clean1_merged_airlines;
run;
```

The CORR Procedure

7 Variables:	duration no_pasg speed_ground speed_air height pitch distance

Simple Statistics

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

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

	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
duration duration	1.00000 781	-0.03639 0.3098 781	-0.04897 0.1716 781	0.04454 0.5364 195	0.01112 0.7564 781	-0.04675 0.1918 781	-0.05138 0.1514 781
no_pasg no_pasg	-0.03639 0.3098 781	1.00000 831	-0.00013 0.9969 831	-0.00616 0.9305 203	0.04699 0.1760 831	-0.01793 0.6057 831	-0.01776 0.6093 831
speed_ground speed_ground	-0.04897 0.1716 781	-0.00013 0.9969 831	1.00000 831	0.98794 <.0001 203	-0.05761 0.0970 831	-0.03912 0.2599 831	0.86624 <.0001 831
speed_air speed_air	0.04454 0.5364 195	-0.00616 0.9305 203	0.98794 <.0001 203	1.00000 203	-0.07933 0.2606 203	-0.03927 0.5780 203	0.94210 <.0001 203
height height	0.01112 0.7564 781	0.04699 0.1760 831	-0.05761 0.0970 831	-0.07933 0.2606 203	1.00000 831	0.02298 0.5082 831	0.09941 0.0041 831
pitch pitch	-0.04675 0.1918 781	-0.01793 0.6057 831	-0.03912 0.2599 831	-0.03927 0.5780 203	0.02298 0.5082 831	1.00000 831	0.08703 0.0121 831
distance distance	-0.05138 0.1514 781	-0.01776 0.6093 831	0.86624 <.0001 831	0.94210 <.0001 203	0.09941 0.0041 831	0.08703 0.0121 831	1.00000 831

Observation and Conclusion:

- the correlation between speed_ground and speed_air speed_air is 0.99, with p value less than 0.05, suggesting speed_ground is definitely highly correlated to speed_air
- The correlation between duration and distance is -0.06, suggesting that duration is not correlated to distance

Conclusion:

- Because of the strong positive relation, close to 1, between speed_ground and speed_air we can conclude that: as speed_ground increases or decreases, speed_air also increases or decreases respectively.
- **So, speed_air is not adding any additional value in the analysis of distance and we can study effect on distance with speed_ground instead of speed_air**
- The duration of flight is not impacting the landing distance
- Also, duration has many missing values which will dilute our analysis. **So, we can remove the column of duration from analysis**

Chap 3: Statistical Modelling

Goal:
In the statistical modelling step, we check which parameters are associated with the variable to be modelled, distance in this case.
Further, the direction and magnitude of the association are checked. Model is created for linear regression of predictors against response variable.
The model is checked and remodeled to fit the data perfectly.

- Since speed_ground is highly correlated to distance, we try modelling distance based on speed_ground

```
proc reg data=clean1_merged_airlines;
model distance =speed_ground;
plot distance
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

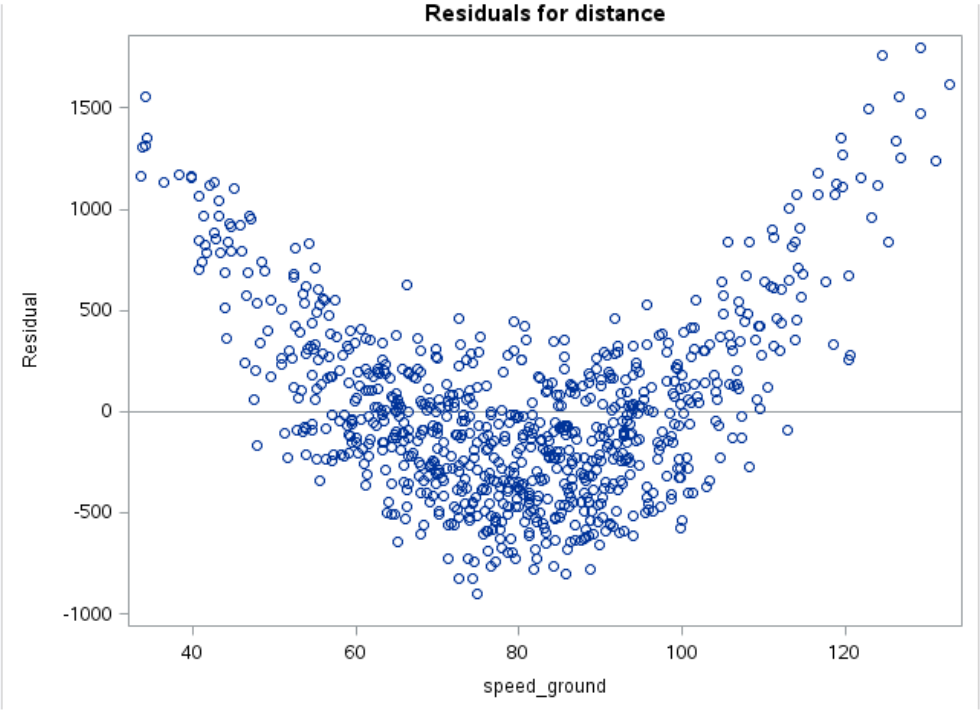
Analysis of Variance

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



Observation: The residual vs speed_ground relation is not linear but it is a U-shaped curve

Conclusion: The U-shaped graph above suggests that a quadratic term of speed_ground needs to be used to model distance

- **Remodel 1**: We add the square of speed_ground variable to our predictors
  As the quadratic equation suggests $y=ax^2+bx+c$, we need to model Y as a function of both square_speed_ground and speed_ground

```
data clean2_merged_airlines;

set clean1_merged_airlines;

square_speed_ground=speed_ground**2;

run;
```

```
proc reg data=clean1_merged_airlines;
model distance =speed_ground square_speed_ground;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

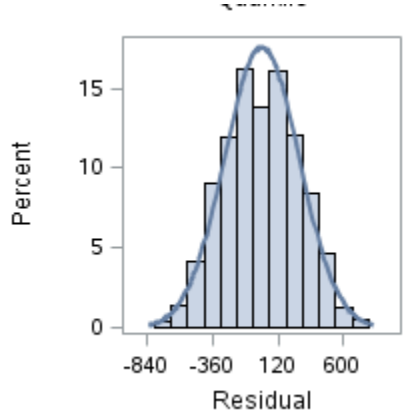
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2899.35562	131.15792	22.11	<.0001
speed_ground	speed_ground	1	-81.53482	3.31462	-24.60	<.0001
square_speed_ground		1	0.76503	0.02038	37.54	<.0001

Basis the estimated parameters, following is a suggested model:

Distance = 0.76 (speed_ground)^2 – 81.5 (speed_ground) + 2899.3

The output of regression modelling provides, the spread and normality of residual. These graphs provide a preliminary check to tell us if the model is fit for the data.



Observation and Conclusion:

- The scattered distribution of residual against speed_ground suggests that residual is independent of speed_ground
- The random distribution of residual suggests that the variance is constant and has a mean of zero. Suggesting that the model is a good fit for data

- **Remodel 2**: But since other variables are also correlated to distance, we now remodel distance taking into consideration other predictors from the dataset.

```
proc reg data=clean2_merged_airlines;

model distance =speed_ground square_speed_ground height pitch no_pasg;

run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Root MSE	225.79067	R-Square	0.9369
Dependent Mean	1522.48287	Adj R-Sq	0.9365
Coeff Var	14.83042		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1853.02482	144.52538	12.82	<.0001
speed_ground	speed_ground	1	-78.27909	2.75129	-28.45	<.0001
square_speed_ground		1	0.74829	0.01691	44.26	<.0001
height	height	1	12.97942	0.80350	16.15	<.0001
pitch	pitch	1	158.67206	14.93055	10.63	<.0001
no_pasg	no_pasg	1	-2.19279	1.04760	-2.09	0.0366

Observation and Conclusion: Considering the p value of hypothesis:

H0: Beta=0 for all the variables

The p value is in less than 0.05 so H0 is rejected and we take into consideration impact of all variables

Also, the value of R-square is higher than the previous model.

Thus, data better fits the model.

The suggested model is:

Distance = 0.75 (speed_ground)^2 – 78.3 (speed_ground) + 158.7 (pitch) + 12.97 (height) – 2.2 (no_pasg)

Explanation:

1. When height increases by 1 meter, the distance increases by 12.97 feet
2. When pitch increases by 1 degree, the distance increases by 158.7 feet
3. When number of passengers increase by 1, the distance decrease by 2.2 feet
4. When speed_ground increases till 104 miles/hour, the distance decreases. Thereafter the distance increases with increase in speed_ground.

Chapter 4: Testing Model

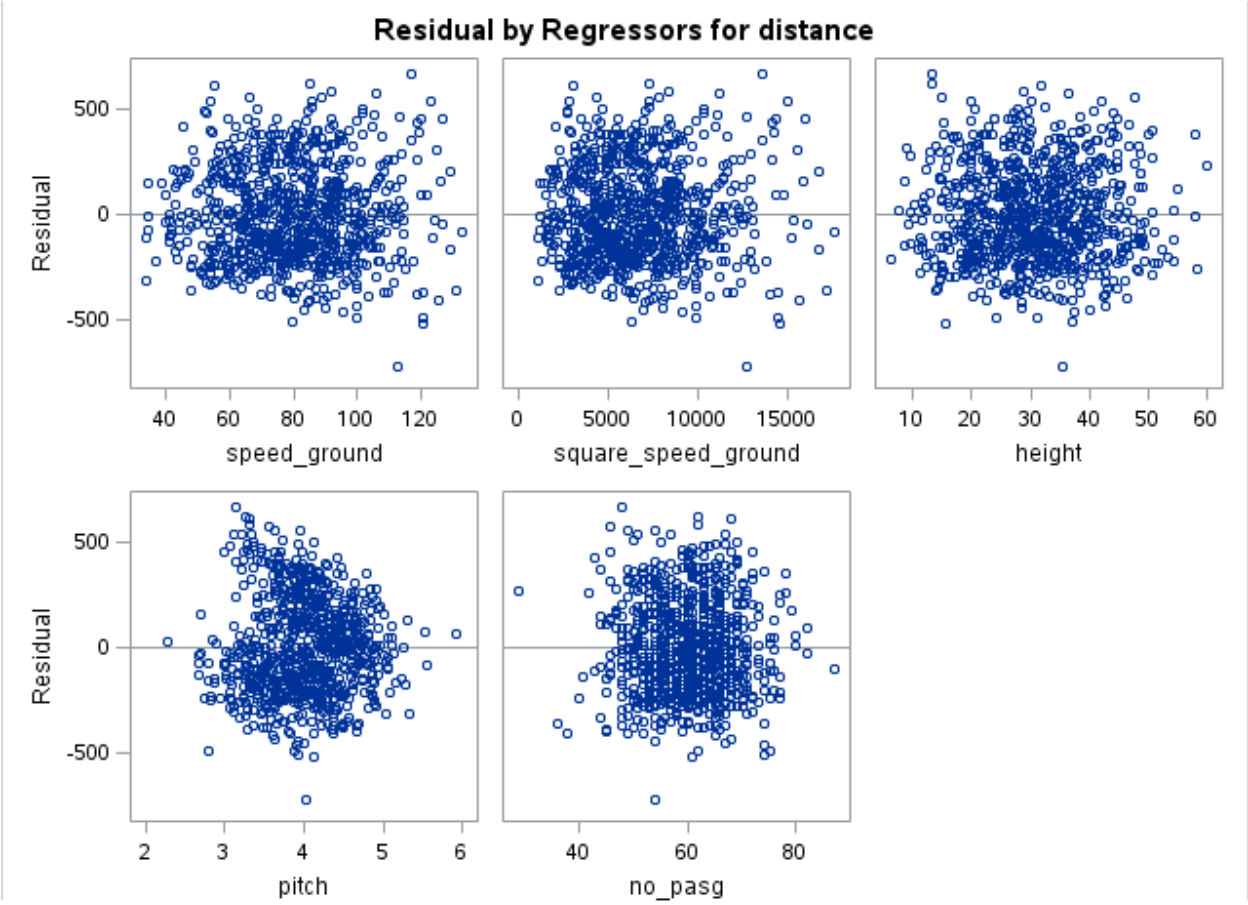
Goal:

It is important to thoroughly evaluate the created model and review the steps executed to create it, to be certain the model properly achieves the business objectives.

The model is tested to check if the residual is independent of the predictors and if residual is normally distributed.

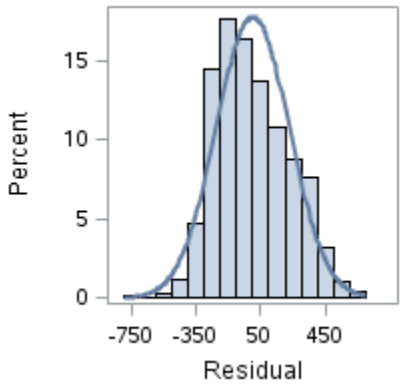
Following steps are used to test the model:

1. **Checking if residual is independent and checking constant variance**



Observation and Conclusion: Residual is scattered when plotted against predictors. Thus, residual is independent of predictors.
Also, residual has a constant variance.

2. **Checking if residual is normally distributed**



Observation and Conclusion: The histogram suggests that residual is normally distributed.

3. **Checking means of residual**

The MEANS Procedure

Analysis Variable : residual Residual	
t Value	Pr > |t|
-0.00	1.0000

Observation and Conclusion: p value test for the hypothesis: mean is equal to zero has a perfect value of 1. Suggesting the hypothesis is accepted and residual has to mean of 0.

Thus, the final model is tested and is found to fit the dataset.

Question & Answers:

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

   I found 831 observations fit for my model.
   Out of 950 flights, there were 19 records that had abnormal values. I believe, these values were recorded as a reason for the error. It was necessary to remove these abnormal values for creating a perfect model;

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

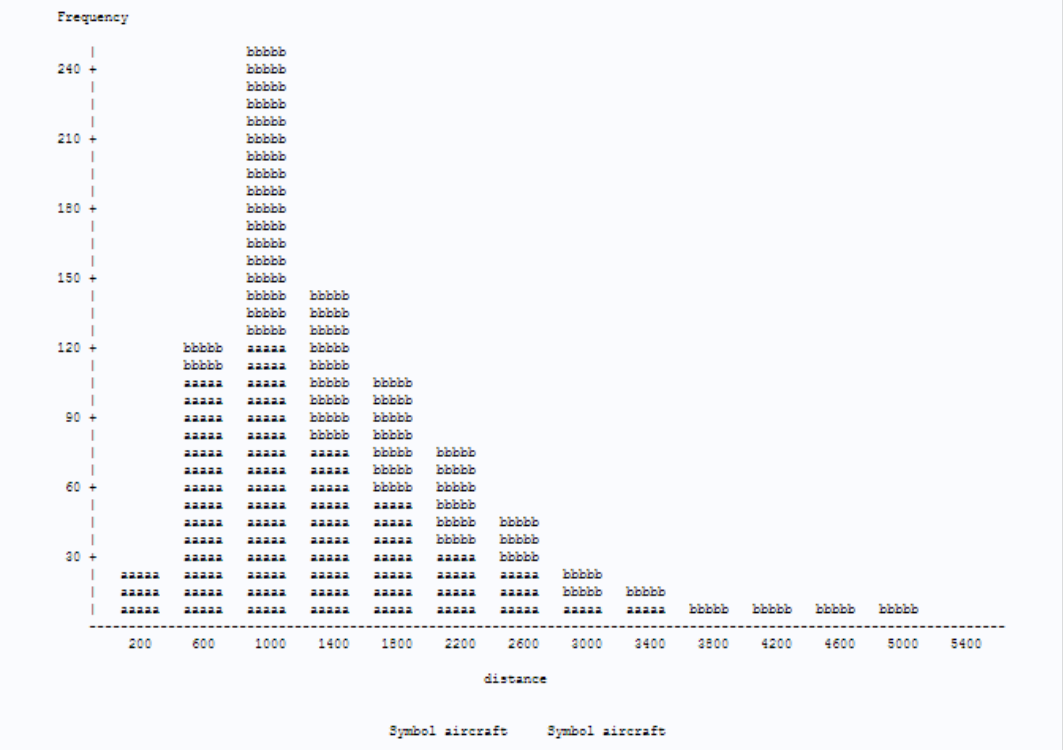
   - Speed_air is correlated to speed_ground which is being used in creating the model. So, we need not include speed_air in our mode
   - Duration variable has no correlation with distance. Thus, we exclude duration from our model.

Variable	Associated to Distance?	Direction of association	Shape of association	Size of association
Speed_ground	Yes	+ve	Parabolic (Quadratic equation)	0.75 (speed_ground)2 – 78.3 (speed_ground)
Pitch	Yes	+ve	Linear	158.7
Height	Yes	+ve	Linear	12.97
No_pasg	Yes	-ve	Linear	2.2
Speed_air	No	n/a	n/a	n/a
Duration	No	n/a	n/a	n/a

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

   Yes, the distribution of records for Boeing and Airbus are different. Please find below detailed discussion:

```
proc chart data=clean2_merged_airlines;
vbar distance/subgroup=aircraft;
run;
```



Observation and Conclusion: The spread of distance for airbus and boeing is different as seen from the chart above.

- We check the summary statistics of this distribution:

```
data airbus;
set clean2_merged_airlines;
if aircraft="airbus";
run;
data boeing;
set clean2_merged_airlines;
if aircraft="boeing";
run;
```

```
proc means data=airbus n  nmiss mean std min max range;
var distance;
TITLE summary statistics for airbus;
run;
```

```
proc means data=boeing n nmiss mean std min max range;
var distance;
TITLE summary statistics for boeing;
run;
```

summary statistics for airbus

The MEANS Procedure

Analysis Variable : distance distance						
N	N Miss	Mean	Std Dev	Minimum	Maximum	Range
444	0	1323.32	791.9282481	41.7223127	4896.29	4854.57

summary statistics for boeing

The MEANS Procedure

Analysis Variable : distance distance						
N	N Miss	Mean	Std Dev	Minimum	Maximum	Range
387	0	1750.98	953.8500300	573.6217861	5381.96	4808.34

- Proc TTest Procedure to test variance of means:

The TTEST Procedure
Variable: distance (distance)

aircraft	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)		-427.7	871.1	60.5772		

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

Observation and Conclusion:

Hypothesis of equality of variances is less than 0.05

Thus, we check the p value of Unequal variances

The probability of t is <0.0001 suggesting means of distance for Boeing and Airbus is significantly different.