Name: Shrey Kumar Parth
Student ID: M13383610

STAT COMPUTING

DATA ANALYSIS FINAL REPORT
FLIGHT LANDING

Observation

Landing data with parameters (flight duration, air speed, ground speed, height of the flight, pitch(angle) of the flight) from 950 commercial flights (not real data set but simulated from statistical models) has been provided.

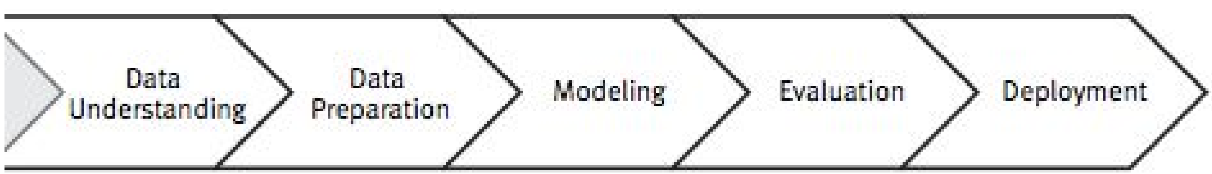
I created the final model using **832 observations** as 100 observations were duplicates and 18 were abnormal with respect to the defined parameters in the problem. The dataset was cleansed modified and linear regression was performed upon. The factors which impact the landing distance of a flight are the **model of the aircraft (airbus or boeing), the ground speed of the aircraft and height of the aircraft when passing over the threshold of the runway**. **And, the make of the aircraft (airbus or boeing) does play a significant role.**

The relationship between the variables can be explained by the final equation

$Y(\text{distance}) = -2052.896 + 42.78x_1 + 14.52x_2 - 501.57x_3$.

where x_1= speed_ground, x_2=height and x_3=aircraft_type

Method



Analysis Steps

CHAPTER 1(a): DATA UNDERSTANDING

Collect Initial Data

The data was provided by Prof. Liu in class. The data consists of two Excel files, ‘FAA1.xls’ (800 flights) and ‘FAA2.xls’ (150 flights).

Describe Data

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.

Explore Data

The data consists of 8 variables in total. The **'aircraft'** is categorical and the others are numerical.

Variable segmentation: Identification of the variables. A general hypothesis of what I think about the importance of variables in the analysis. Not all variables will have the same importance in the model so I have categorized the variables in 3 categories namely- **low, neutral and high.**

Low **-** *Duration, No_pasg*

Neutral- *Pitch*

High- *speed_ground, speed_air, height, aircraft*

Breakdown

Data cleaning is the first step in the process of data analysis and forms the most important step of the process.

Importing the data:

```
%web_drop_table(WORK.IMPORT2);
FILENAME REFFILE '/folders/myfolders/Assgnment2/FAA1
(Autosaved).xls';
PROC IMPORT DATAFILE=REFFILE
DBMS=XLS
OUT=WORK.IMPORT2;
GETNAMES=YES;
RUN;
PROC CONTENTS DATA=WORK.IMPORT2; RUN;
%web_open_table(WORK.IMPORT2);
```

Importing the data into SAS. Both the files have been concatenated and the duplicate values have been removed. The total number of individual observations is 800 and 200.

Data Set Name	WORK.IMPORT2	Observations	800
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	11/09/2019 22:23:13	Observation Length	64
Last Modified	11/09/2019 22:23:13	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)		

Data Set Name	WORK.IMPORT1	Observations	200
Member Type	DATA	Variables	7
Engine	V9	Indexes	0
Created	11/09/2019 22:29:26	Observation Length	64
Last Modified	11/09/2019 22:29:26	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)		

Verify Data Quality

Aircraft: Data Type: String, Categorical values - ‘Airbus’ and ‘Boeing’.

Duration: Data Type: Decimal, Numerical values.

Speed_ground: Data Type: Decimal, Numerical values.

Speed_air: Data Type: Decimal, Numerical values.

Height: Data Type: Decimal, Numerical values.

Pitch: Data Type: Decimal, Numerical values.

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	8	$8.	$8.	aircraft
8	distance	Num	8	BEST8.		distance
2	duration	Num	8	BEST8.		duration
6	height	Num	8	BEST8.		height
3	no_pasg	Num	8	BEST8.		no_pasg
7	pitch	Num	8	BEST8.		pitch
5	speed_air	Num	8	BEST8.		speed_air
4	speed_ground	Num	8	BEST8.		speed_ground

CHAPTER 1(b): DATA PREPARATION

In this step we'll have to clean our data for exploratory analysis and delve into it deeper. The basic steps of cleaning our dataset is to look for missing values and abnormal data which will skew our data.

Cleaning and Constructing Data

Concatenating Data:

We notice that there are only 150 rows in the 2nd file but it shows 200, that's because of an extra space in the aircraft column. We need to remove that when concatenating.

```
DATA IMPORT3;
  SET IMPORT2 IMPORT1;
  options missing = ' ';
  if missing(cats(of _all_)) then delete;
RUN;
```

Now we have 950 rows.

Data Set Name	WORK.IMPORT3	Observations	950
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	11/09/2019 22:44:21	Observation Length	64
Last Modified	11/09/2019 22:44:21	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)		

Getting rid of duplicate data:

We see that there are duplicate values in our data which we have to remove. Duplicate data not only increases our data size but it also leads to anomalies in analysis.

```
DATA IMPORT3;
  SET IMPORT2 IMPORT1;
  options missing = ' ';
  if missing(cats(of _all_)) then delete;
  proc sort data=IMPORT3 out=IMPORT3 nodupkey;
  by speed_ground no_pasg pitch height distance;
RUN;
```

Data Set Name	WORK.IMPORT3	Observations	850
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	11/09/2019 22:59:49	Observation Length	64
Last Modified	11/09/2019 22:59:49	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	YES
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Integrating and Formatting data

We now study the data carefully and get the basic numerical properties. Sorting the data according to aircraft type, we explore the data.

```
PROC CONTENTS DATA=WORK.IMPORT3;
RUN;
PROC SORT DATA=WORK.IMPORT3;
BY aircraft;
RUN;
PROC MEANS DATA=WORK.IMPORT3 MEAN MIN MAX STD NMISS ;
BY aircraft;
VAR duration no_pasg speed_air speed_ground height pitch ;
RUN;
```

Details:

aircraft=airbus

Variable	Label	Mean	Minimum	Maximum	Std Dev	N Miss
duration	duration	156.1099583	16.8934549	305.6217107	49.7830202	50
no_pasg	no_pasg	60.2466667	36.0000000	87.0000000	7.4174927	0
speed_air	speed_air	104.2123333	95.0113646	131.3379485	8.0924561	364
speed_ground	speed_ground	80.1994492	33.5741041	131.0351822	16.9206507	0
height	height	30.3196736	-3.3323880	58.2277997	10.2505068	0
pitch	pitch	3.8317436	2.2844801	5.5267842	0.5004493	0

aircraft=boeing

Variable	Label	Mean	Minimum	Maximum	Std Dev	N Miss
duration	duration	151.9031187	14.7642071	298.5223339	48.7011866	0
no_pasg	no_pasg	59.9425000	29.0000000	82.0000000	7.5834005	0
speed_air	speed_air	103.5054579	90.0028586	141.7249357	11.5689208	278
speed_ground	speed_ground	78.6118058	27.7357153	141.2186354	21.1999092	0
height	height	29.9468397	-3.5462524	59.9459639	10.3387152	0
pitch	pitch	4.2091735	2.9931514	5.9267842	0.4874715	0

Missing Values:

- We see that the variable duration has 50 values missing out of total 850 observations. So that comes up to **5.88%** which is actually a very small number.
- The variable speed_air has a total of 642 values missing out of total 850 observations. That is actually **75.5%** and is a significant number.

```
proc sql;
select  (Nmiss(duration)*100)/count(*) as missing_dur,
      (Nmiss(no_pasg)*100)/count(*) as missing_no_pasg,
      (Nmiss(speed_air)*100)/count(*) as missing_speed_air,
      (Nmiss(speed_ground)*100)/count(*) as
missing_speed_ground,
      (Nmiss(height)*100)/count(*) as missing_height,
      (Nmiss(pitch)*100)/count(*) as missing_pitch
from import3;
quit;
```

Abnormal Values:

We have some abnormal data present in the Dataset which will interfere with our analysis. We have to get rid of these data points. Some of the conditions that we have been already provided to us.

- The duration of a normal flight should always be greater than 40min.
- The speed_ground should be less than 30MPH or greater than 140MPH for the landing to be considered normal.
- The landing aircraft is required to be at least 6 meters high at the threshold of the runway.

We now clean the data applying these conditions in the code.

```
PROC CONTENTS DATA=WORK.IMPORT3;
RUN;
PROC SORT DATA=WORK.IMPORT3;
BY aircraft;
RUN;
DATA IMPORT3;
SET WORK.IMPORT3;
DROP speed_air;
```

```
IF duration<40 and duration is not null then DELETE;
IF speed_ground<30 or speed_ground>140 then DELETE;
IF height<6 then DELETE;
RUN;
PROC PRINT DATA=IMPORT3;
RUN;
%web_open_table(WORK.IMPORT3);
```

We drop the rows containing these abnormal values to get a better and cleaner version of our data. Now we have a total of 832 rows.

Data Set Name	WORK.IMPORT3	Observations	832
Member Type	DATA	Variables	9
Engine	V9	Indexes	0
Created	18/09/2019 19:56:43	Observation Length	72
Last Modified	18/09/2019 19:56:43	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)		

In this chapter, we started with merging both the files so that we have all our data in one place. But since some of the data was redundant, we had to get rid of that.

In the next step we checked for missing and abnormal values in the data. The abnormal data points have been removed so that it doesn't interfere with our analysis.

Converting aircraft (categorical) to numerical type:

```
DATA IMPORT4;
SET IMPORT3;
if aircraft='airbus' then aircraft_type=1;
else aircraft_type=0;
run;
proc print;
Run;
```

This is how an instance of the table looks now.

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	null	aircraft_type
1	airbus	192.2829	64	33.5741		36.97069	4.358464	782.7174		1
2	airbus	126.0784	54	36.42139		33.7997	4.866111	869.0337		1
3	airbus		46	40.80179		24.40013	3.968209	620.0905		1
4	airbus	100.2531	61	41.10099		34.56394	3.490939	668.9332		1
5	airbus	128.6879	56	43.85281		34.11456	3.312578	554.0662		1
6	airbus	169.7494	65	43.92356		43.05077	3.365144	735.8234		1
7	airbus	134.6856	57	44.12614		27.18238	3.012829	417.5431		1
8	airbus	191.3109	66	44.25848		49.28774	3.69714	901.4067		1
9	airbus	42.14623	63	46.26472		20.49071	3.481912	383.5585		1
10	airbus	172.0493	36	47.48677		13.98481	4.29902	250.6898		1
11	airbus	117.7406	59	47.6798		28.60649	3.752047	406.0893		1
12	airbus	190.7394	77	47.88212		14.83596	2.732284	41.72231		1
13	airbus	92.17284	58	48.1728		34.75702	3.948344	558.3651		1
14	airbus	248.7291	58	49.46213		16.79618	3.761689	452.1313		1
15	airbus	209.1937	54	50.81293		38.84132	4.033898	566.9269		1
16	airbus		54	50.90311		35.72948	4.54404	597.9855		1
17	airbus	142.5876	66	51.15823		8.559069	3.913448	242.5959		1
18	airbus	212.054	63	51.58704		20.45129	3.063686	133.0869		1
19	airbus		59	52.09981		34.55297	4.88663	647.4882		1
20	airbus	133.7396	50	52.33337		18.0381	3.935193	513.4958		1
21	airbus	182.4478	66	52.70784		24.30264	4.185967	317.8127		1
22	airbus	60.53364	64	52.98413		23.865	4.322738	487.4687		1
23	airbus	106.3736	66	53.37241		51.00335	2.827557	538.9741		1
24	airbus	183.6185	69	53.53924		31.73942	3.523775	349.1585		1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]		[illegible]

CHAPTER 2: DESCRIPTIVE ANALYSIS

Mapping correlation

We noticed that the speed_air column is missing most of the values i.e. around 75%. Now we don't actually know how important this variable is or will be in our analysis so we can't just remove it. Removing this column will interfere with our analysis.

So, What can we do?

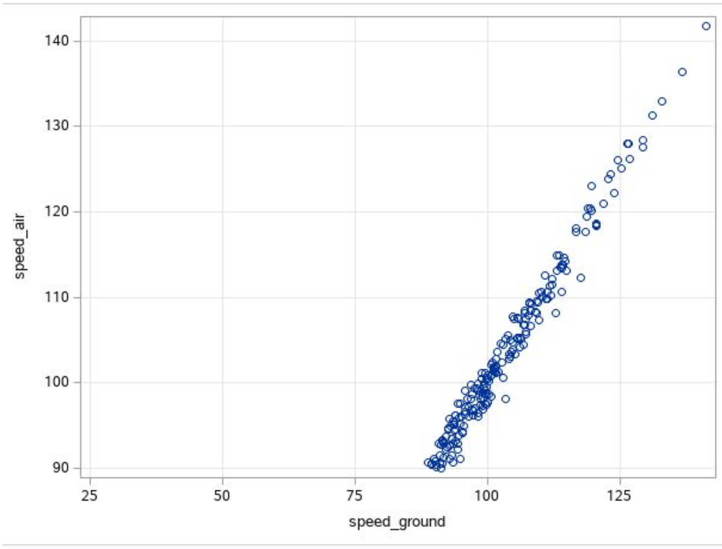
We notice that the values in columns, speed_ground and speed_air are very similar to each other. We'll take out the correlation of speed_air with the speed_ground column to see how much are they actually similar.

```
ods noproctitle;
ods graphics / imagemap=on;
proc corr data=WORK.IMPORT3 pearson nosimple noprob plots=none;
var speed_air;
with speed_ground;
Run;
```

1 With Variables:	speed_ground
1 Variables:	speed_air

Pearson Correlation Coefficients Number of Observations	
	speed_air
speed_ground speed_ground	0.98858 204

The correlation comes out to **0.988**, which is very high and almost nears 1. Now, let's see and plot these two variables together and see what do we get.



It forms a straight line , which essentially proves our hypothesis that these two variables are strongly correlated.

We have solved the problem of missing values in **speed_air** to one extent. Since there's a high correlation between the variables, we'll use the variable **speed_ground** wherever possible in our analysis.

Marginal Analysis

Now, we perform a marginal mapping of all the variables we have with the variable **'distance'** to check the significance of these variables with respect to distance.

```
ods noproctitle;
ods graphics / imagemap=on;
proc corr data=WORK.IMPORT3 pearson nosimple noprob plots=none;
var speed_ground height duration no_pasg speed_air pitch
aircraft;
with distance;
run;
```

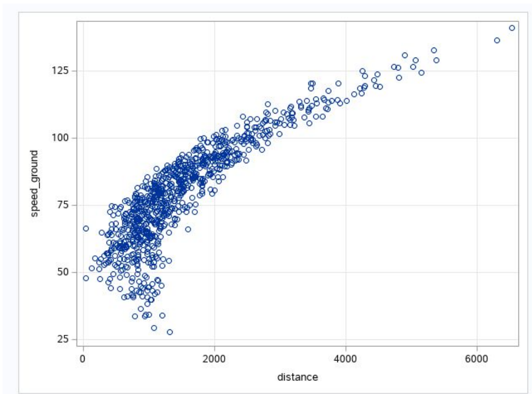
8 Variables:	duration no_pasg speed_ground speed_air height pitch aircraft_type distance

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
duration	782	154.73115	48.33503	121000	41.94937	305.62171	duration
no_pasg	832	60.06010	7.48806	49970	29.00000	87.00000	no_pasg
speed_ground	832	79.61135	18.82881	66237	33.57410	136.65916	speed_ground
speed_air	204	103.64650	9.98231	21144	90.00286	136.42342	speed_air
height	832	30.47449	9.79067	25355	6.22752	59.94596	height
pitch	832	4.00536	0.52628	3332	2.28448	5.92678	pitch
aircraft_type	832	0.53365	0.49917	444.00000	0	1.00000	
distance	832	1528	911.04506	1271493	41.72231	6310	distance

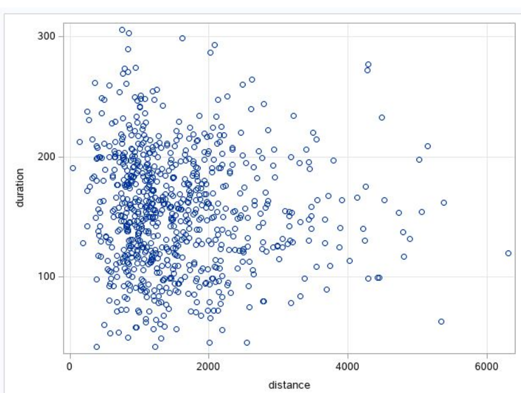
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	duration	no_pasg	speed_ground	speed_air	height	pitch	aircraft_type	distance
duration duration	1.00000 782	-0.03685 0.3034 782	-0.05144 0.1507 782	0.03261 0.6500 196	0.00979 0.7845 782	-0.04701 0.1891 782	0.04532 0.2056 782	-0.05525 0.1226 782
no_pasg no_pasg	-0.03685 0.3034 782	1.00000 832	0.00179 0.9589 832	0.00361 0.9591 204	0.04782 0.1682 832	-0.01773 0.6095 832	0.02199 0.5264 832	-0.01413 0.6840 832
speed_ground speed_ground	-0.05144 0.1507 782	0.00179 0.9589 832	1.00000 832	0.98858 <.0001 204	-0.05207 0.1334 832	-0.03777 0.2765 832	0.03630 0.2957 832	0.86618 <.0001 832
speed_air speed_air	0.03261 0.6500 196	0.00361 0.9591 204	0.98858 <.0001 204	1.00000 204	-0.05286 0.4528 204	-0.03470 0.6223 204	0.05629 0.4239 204	0.94426 <.0001 204
height height	0.00979 0.7845 782	0.04782 0.1682 832	-0.05207 0.1334 832	-0.05286 0.4528 204	1.00000 832	0.02348 0.4988 832	0.01254 0.7179 832	0.10655 0.0021 832
pitch pitch	-0.04701 0.1891 782	-0.01773 0.6095 832	-0.03777 0.2765 832	-0.03470 0.6223 204	0.02348 0.4988 832	1.00000 832	-0.35433 <.0001 832	0.08754 0.0115 832
aircraft_type	0.04532 0.2056 782	0.02199 0.5264 832	0.03630 0.2957 832	0.05629 0.4239 204	0.01254 0.7179 832	-0.35433 <.0001 832	1.00000 832	-0.24076 <.0001 832
distance distance	-0.05525 0.1226 782	-0.01413 0.6840 832	0.86618 <.0001 832	0.94426 <.0001 204	0.10655 0.0021 832	0.08754 0.0115 832	-0.24076 <.0001 832	1.00000 832

Correlation

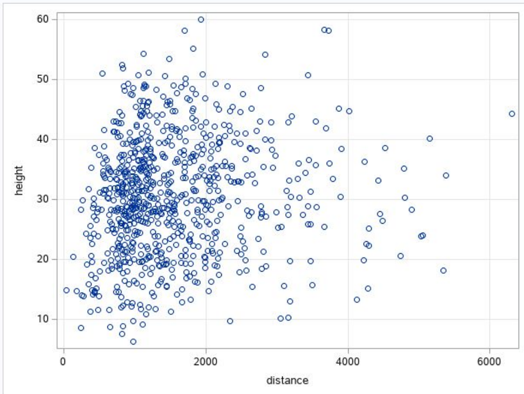
Here we try to see the correlation between the variables and distance by plotting scatter plots with them.

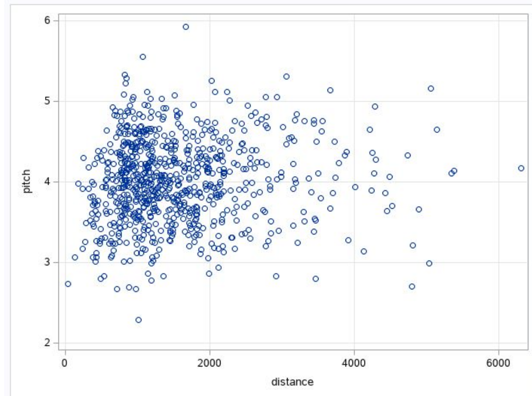


Correlation of **speed_ground** with distance looks promising as the graph forms a straight line.

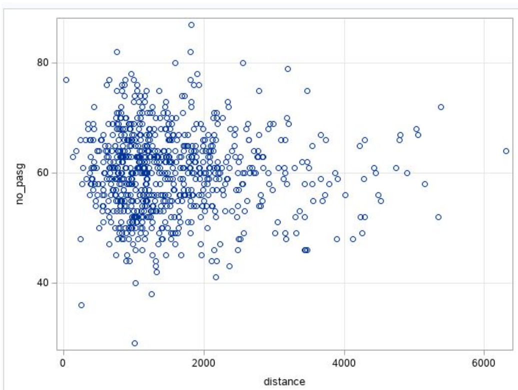


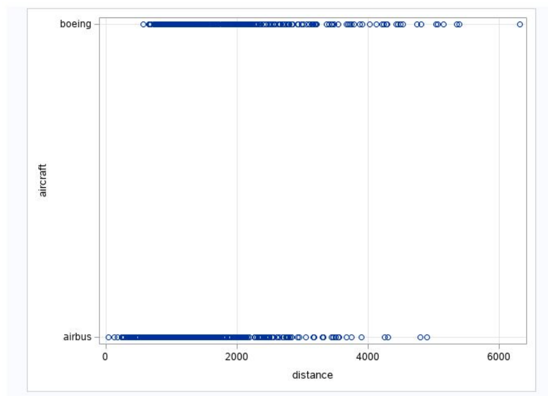
Duration doesn't relate much to the distance and forms negative correlation which is evident here.





The same goes for **height**. It doesn't form a strong correlation**. Pitch** doesn't forms a strong correlation too.





No_pasg shows negative correlation in respect to distance.

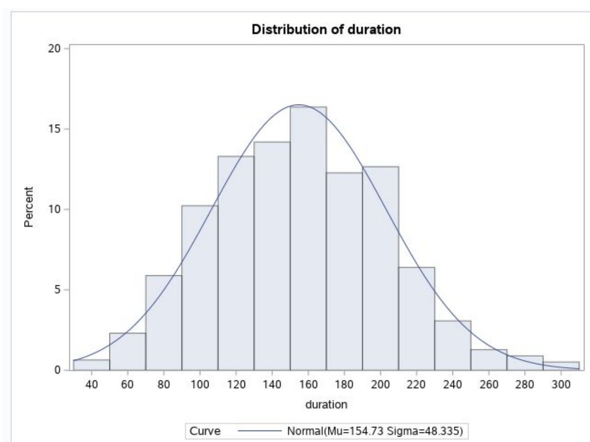
Visualizing Data

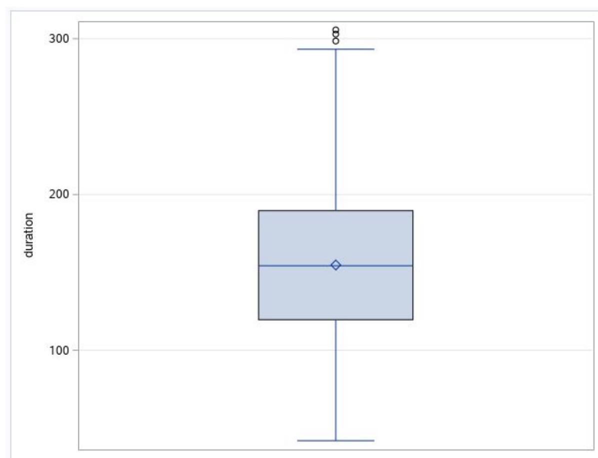
Here we try to explore the data provided by querying and visualizing it for inconsistencies, missing values , wrong values. This step helps us in getting to know the data better.

We plot the **histogram distribution and box plot charts** of the variables to get a visualization of the distribution.

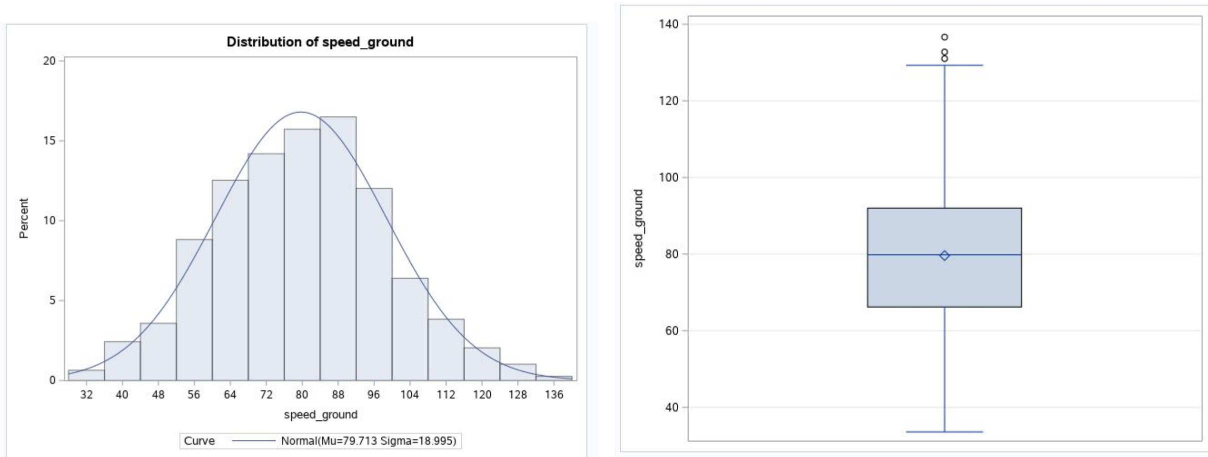
```
ods noproctitle;
ods graphics / imagemap=on;
/* Exploring Data */
proc univariate data=WORK.IMPORT3;
ods select Histogram;
var duration speed_ground height pitch;
histogram duration speed_ground  height pitch /normal;
run;
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=WORK.IMPORT3;
vbox distance /;
yaxis grid;
run;
ods graphics / reset;
```

Distribution of Variables

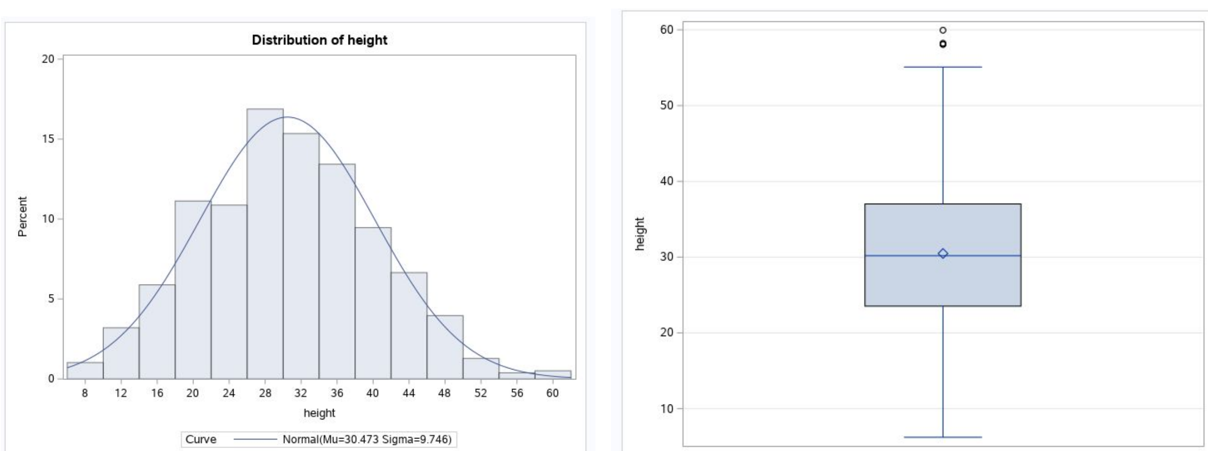




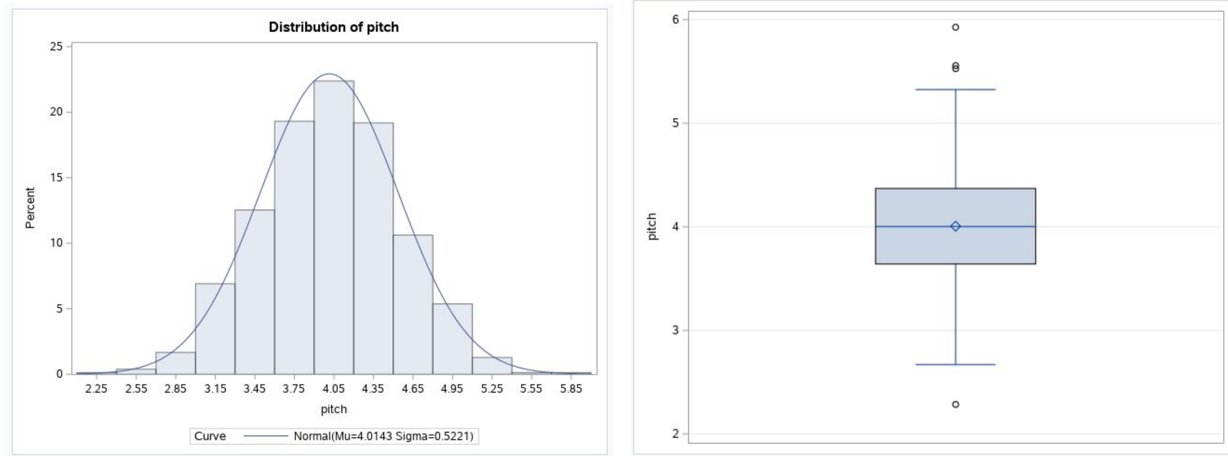
Duration



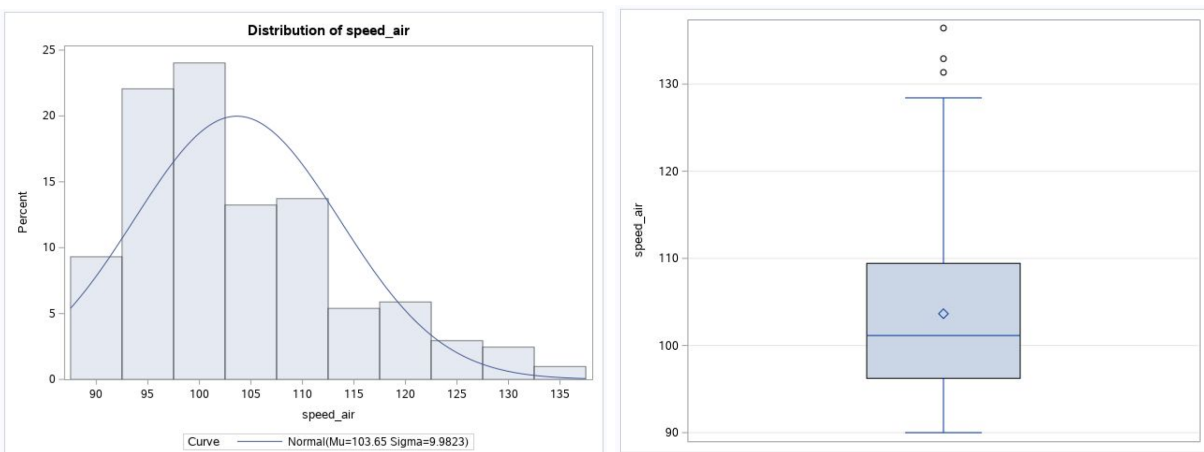
Speed_ground



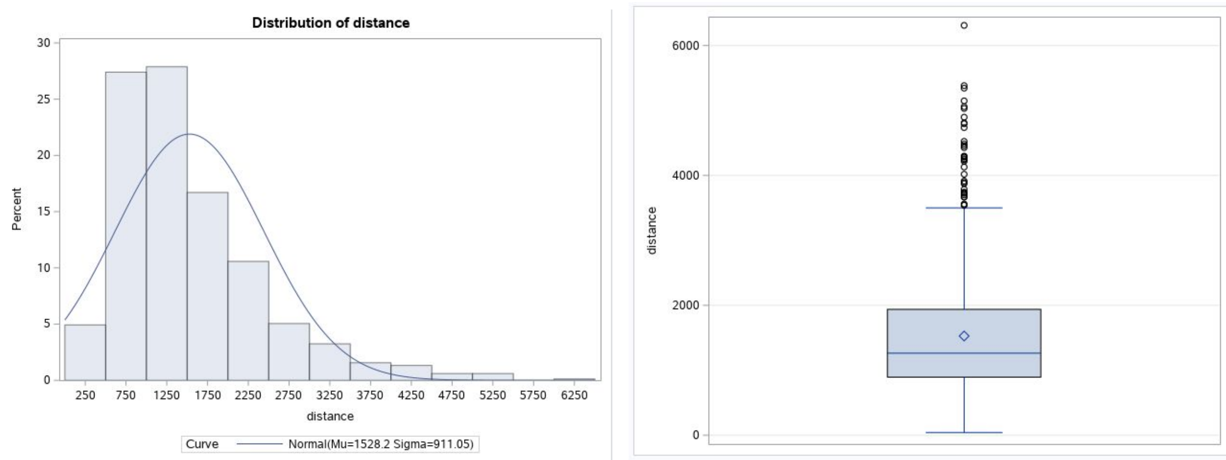
Height



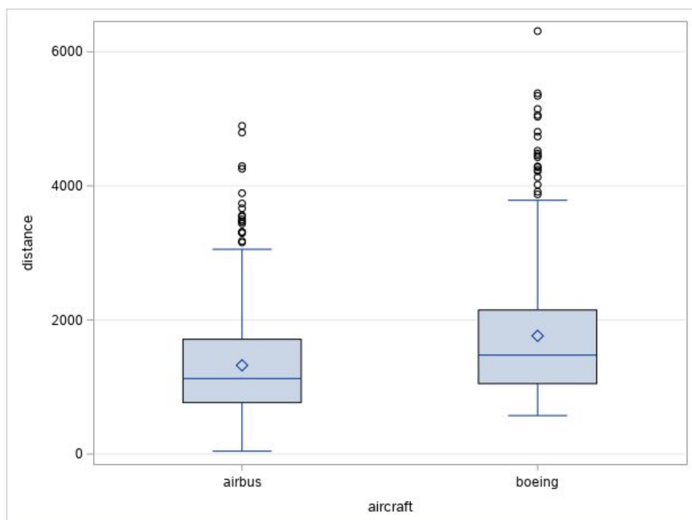
Pitch



Speed_air



Distance



Aircraft

Most of the variables follow a normal distribution here, except **speed_air and distance which are skewed to the right.** **Speed_air**'s distribution shows that the data set is trimmed. There is a very high variability and most of the data points are outliers.

Association with Distance

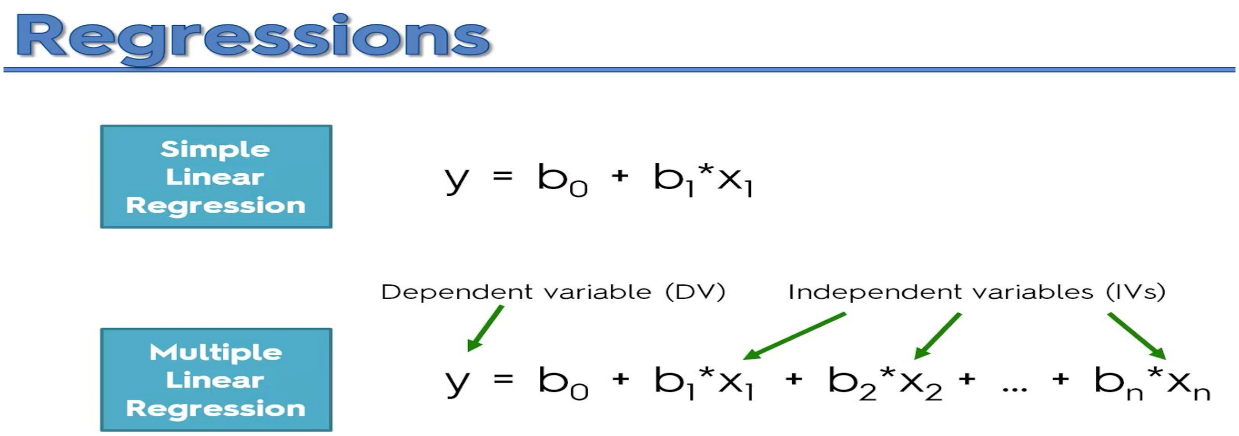
VARIABLE	YES/NO	POSITIVE/NEGATIVE	STRENGTH	SHAPE
speed_ground	Yes	Positive	0.86118	Linear
duration	No	Negative	- 0.05525	Linear
height	Yes	Positive	0.10655	Linear
pitch	Yes	Positive	0.08754	Linear
No_pasg	No	Negative	- 0.01413	Non Linear
aircraft	Yes	Negative	- 0.24076	Linear

CHAPTER 3: MODELLING

Modeling is used to model the predictor variables which are found to have an impact on the target variables.

Multivariate linear regression

Multiple linear regression attempts to model the relationship between two or more explanatory **variables** and a response **variable** by fitting a **linear** equation to observed data.



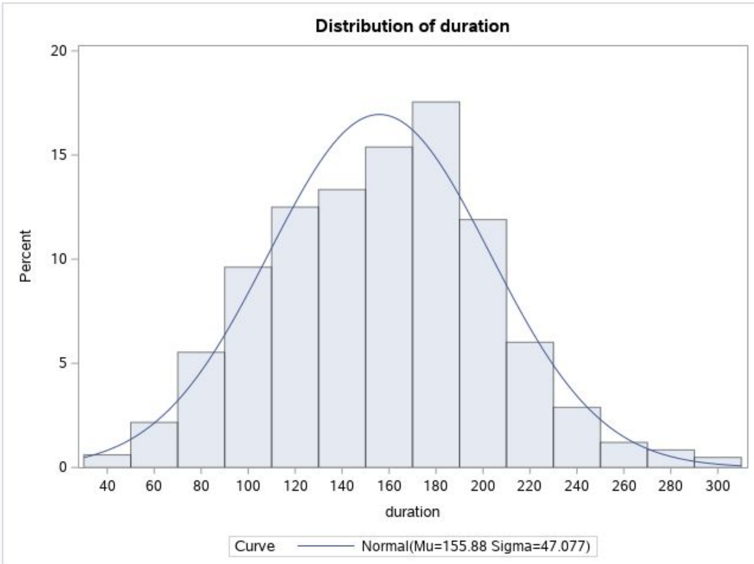
Imputing missing values

Before proceeding to the modelling, I want to impute the missing values in the column **duration.** We know the missing percentage in this column is 5.8%. We can impute values in this column by mean, median or random numbers between the minimum and maximum values.

I have imputed the missing values by **midrange.** Midrange replaces missing target variable values with the maximum value for the variable plus the minimum value for the variable divided by two. The midrange is a rough measure of central tendency that is easy to calculate.

```
proc stdize data=work.import3 out=work.import4
oprefix=old_duration          /* prefix for original variables */
reponly                  /* only replace; do not standardize */
method=MIDRANGE;            /* or MEDIAN, MINIMUM, MIDRANGE, etc. */
var duration;      /* you can list multiple variables to impute */
run;
```

The missing values of duration is now replaced. Looking at the distribution once again.



Modelling

Now we proceed with the modelling. Now it's stated that we will not be using **duration and no_pasg as our predictor variables because they show a negative correlation** with distance and speed_air is being replaced by just speed_ground.

Using all variables

```
proc reg data=work.import4;
model distance=speed_ground duration height pitch no_pasg
aircraft_type;
title Regression analysis of the simulated data set;
Run;
```

Regression analysis of the simulated data set

Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	832
Number of Observations Used	782
Number of Observations with Missing Values	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	562065225	93677537	734.02	<.0001
Error	775	98907855	127623		
Corrected Total	781	660973080			

Root MSE	357.24367	R-Square	0.8504
Dependent Mean	1547.30208	Adj R-Sq	0.8492
Coeff Var	23.08817		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2067.95710	173.27532	-11.93	<.0001
speed_ground	speed_ground	1	42.96336	0.67576	63.58	<.0001
duration	duration	1	0.02115	0.26549	0.08	0.9365
height	height	1	14.67718	1.31531	11.16	<.0001
pitch	pitch	1	19.44910	26.32649	0.74	0.4603
no_pasg	no_pasg	1	-1.48027	1.70242	-0.87	0.3848
aircraft_type		1	-494.44575	27.45434	-18.01	<.0001

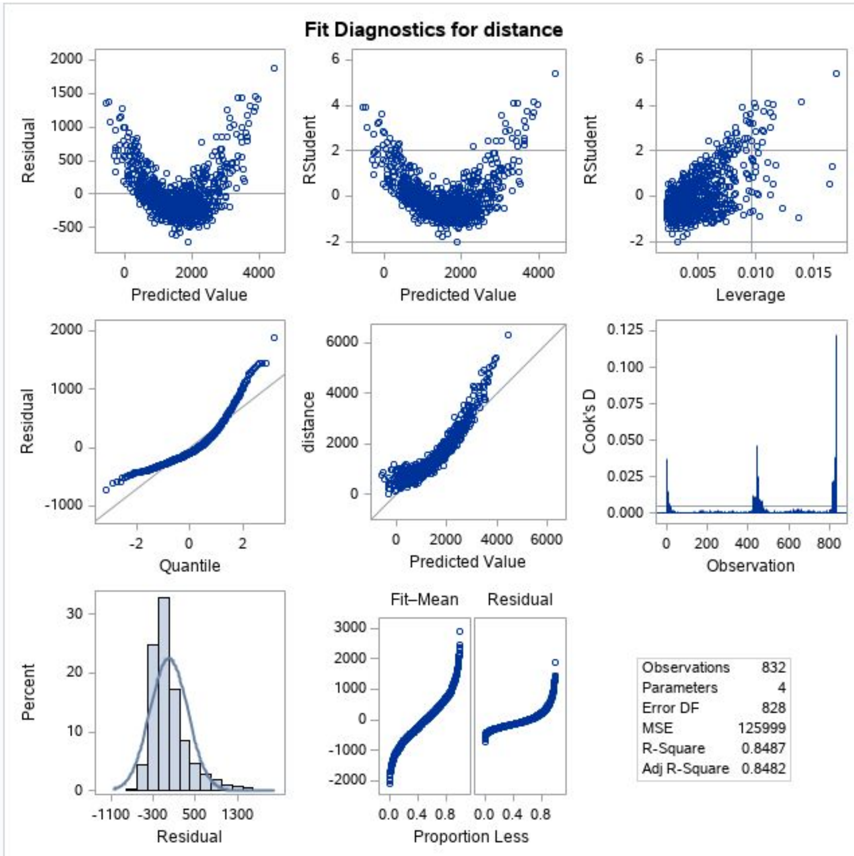
Using significant variables

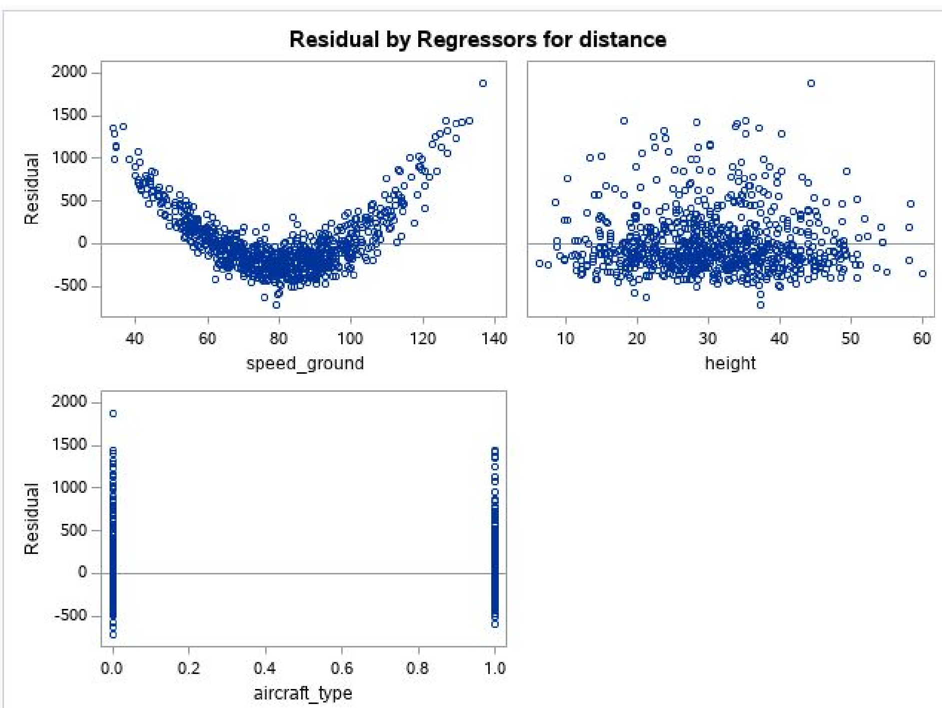
It can be seen that not all variables are significance in the model we just ran. We'll remove **duration, pitch** and **no_pasg** and create a new model.

```
proc reg data=work.import4;
model distance=speed_ground height aircraft_type;
title Regression analysis of the simulated data set;
Run;
```

Root MSE	354.96382	R-Square	0.8487
Dependent Mean	1528.23704	Adj R-Sq	0.8482
Coeff Var	23.22701		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2052.89639	68.30110	-30.06	<.0001
speed_ground	speed_ground	1	42.78669	0.65531	65.29	<.0001
height	height	1	14.52014	1.25952	11.53	<.0001
aircraft_type		1	-501.57254	24.68710	-20.32	<.0001


Fit Diagnostics for distance
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
distance
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 832
Parameters 4
Error DF 828
MSE 125999
R-Square 0.8487
Adj R-Square 0.8482


Residual by Regressors for distance
Residual
speed_ground
height
Residual
aircraft_type

After the model has run, We have got our parameter estimates of the prediction. We see that the p-values for the 3 variables **speed_ground, height, aircraft_type** are all less than 0.0001. The R-squared is 0.8487, meaning that approximately 84% of the variability of distance is accounted for by the variables in the model.

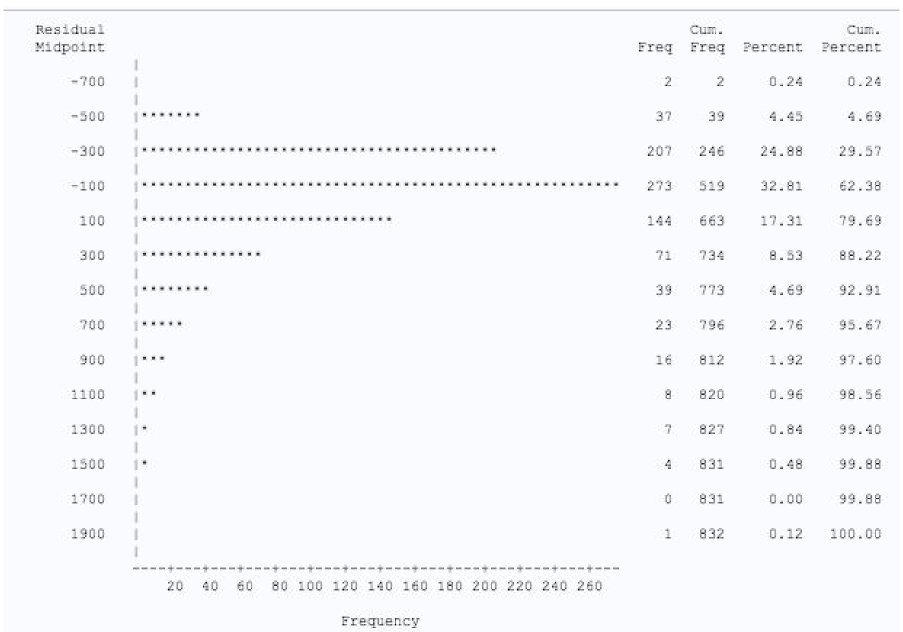
So we reject the null hypothesis which states that "The landing distance of the flight is not affected by these variables."

The final equation comes out $\mathbf{Y(distance) = -2052.896 + 42.78}\boldsymbol{x_1}\mathbf{ + 14.52}\boldsymbol{x_2}\mathbf{ - 501.57}\boldsymbol{x_3}$.
where x_1= speed_ground, x_2=height and x_3=aircraft_type

Residual Analysis

We now check the distribution of residuals.

```
proc reg data=work.import4;
model distance=speed_ground height aircraft_type/r;
title Regression analysis of the simulated data set;
output out=diagnostics r=residual; Run;
proc chart data= diagnostics; hbar residual; run;
```



The normal distribution of residual is slightly skewed to the right.

Conclusion

We are done with chapters of Data Mining and Modeling. We imported the data from multiple sources and joined them and then proceeded to understand the various numerical metrics of the dataset.
After getting a thorough understanding of the parameters we proceeded to clean the data by removing the missing and the abnormal values which would have possibly interfered with our analysis.
We have now modeled and checked the model in accordance with CRISP DM methodology using different steps and iterations.

Name: Shrey Kumar Parth
Student ID: M13383610