SUMMARY

This project has been conducted with the objective of predicting landing overruns for aircrafts using the given data.

Landing overruns are a major problem in the aviation industry as they may can damage to aircrafts as well as the crew and passengers on-board. In past, many aircrafts have fallen victim to this and hence, this project aims at reducing risk of overrun by predicting if an aircraft may have a landing overrun or not based on certain parameters.

The given data is in the form of two tables containing 950 observations in total and having seven variables namely Aircraft, Distance, Duration, Ground Speed, Air Speed, Height and Pitch. After cleaning the data and performing sanity checks according to a given set of conditions, only 780 observations were found to be fit for modeling.

A linear regression model was fit to the data and factors Aircraft Type, Ground Speed, square of Ground Speed and Height were found to be important. Air Speed was removed from the modeling process because of having 75% missing values. Comparison of regression models by aircraft type has also been done in the process.

The linear equation was found to be:

Distance = 2187.5 + (-401.95) * Aircraft Code + (-69) * Ground Speed + 0.69 * Square of Ground Speed + 13.66 * Height

Here aircraft code = 1 for “airbus” and 0 for “boeing”. The <u>r-squared</u> value for the model was **0.98** and <u>Root mean squared distance</u> on the entire dataset was **135.69** metres.

The model was checked for diagnostics and the residuals were found to follow the assumptions of Independence, 0 mean, normal distribution and constant variance.

Due to high R-squared of the model, the model has been checked for overfitting by performing validation on testing dataset after splitting the given dataset into testing and training. The model was found to be consistent across different sets of test dataset. The details of validation have been left out from this report due to its limited scope.

As a result, using this model, given the required set of parameters, we can accurately predict the landing distance of an aircraft and hence its possibility of overrun if we have the prior knowledge of airstrip length. This information can be used to intimate pilots in advance of any risk of overrun before they land.

Chapter – 1

DATA EXPLORATION AND DATA CLEANING

OBJECTIVE:

The objective of this exercise is to perform sanity checks on the data, report inconsistencies and undertake measures to clean the data.

PROCEDURE:

Data Exploration is a step by step procedure to understand the quality and parameters of data. Here, I have combined both the data sources and run then explored the aggregated table. The steps are as follows:

1. Uploading the data on the SAS On-Demand server.

```
LIBNAME INPUT '/home/saxenapi0/GASUE34_data/';

PROC IMPORT DATAFILE='/home/saxenapi0/GASUE34_data/FAA1.xls'
    DBMS=XLS
    OUT=INPUT.FAA1;
    GETNAMES=YES;
RUN;

PROC PRINT DATA = INPUT.FAA1(OBS=10);
RUN;

PROC IMPORT DATAFILE='/home/saxenapi0/GASUE34_data/FAA2.xls'
    DBMS=XLS
    OUT=INPUT.FAA2;
    GETNAMES=YES;
RUN;

PROC PRINT DATA = INPUT.FAA2(OBS=10);
RUN;

```

FAA1

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
6	boeing	137.59581722	55	75.014343744	.	41.21496259	4.203853398	1627.0681991
7	boeing	73.023794916	54	54.4298029	.	24.03532163	3.8376457299	805.30399317
8	boeing	52.903187872	57	57.101661737	.	19.388837508	4.6436717769	573.62178606
9	boeing	155.51861605	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548
10	boeing	176.86203205	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579

FAA2

Obs	aircraft	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
5	boeing	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
6	boeing	55	75.014343744	.	41.21496259	4.203853398	1627.0681991
7	boeing	54	54.4298029	.	24.03532163	3.8376457299	805.30399317
8	boeing	57	57.101661737	.	19.388837508	4.6436717769	573.62178606
9	boeing	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548
10	boeing	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579

2. Combining the datasets

```
DATA COMBINED;
SET INPUT.FAA1 INPUT.FAA2;
RUN;

PROC PRINT DATA=COMBINED(OBS=10);
RUN;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
6	boeing	137.59581722	55	75.014343744	.	41.21496259	4.203853398	1627.0681991
7	boeing	73.023794916	54	54.4298029	.	24.03532163	3.8376457299	805.30399317
8	boeing	52.903187872	57	57.101661737	.	19.388837508	4.6436717769	573.62178606
9	boeing	155.51861605	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548
10	boeing	176.86203205	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579

3. Performing a univariate analysis on each variable (Frequency on categorical variables)

   a. **Categorical Variable**: Aircraft

```
PROC FREQ DATA=INPUT.COMBINED(KEEP=AIRCRAFT);
RUN;
```

The FREQ Procedure

aircraft				
aircraft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
airbus	450	47.37	450	47.37
boeing	500	52.63	950	100.00

b. Numerical Variables:

```
TITLE 'SUMMARY OF VARIABLES PRIOR TO CLEANING';
PROC MEANS DATA = INPUT.COMBINED N MEAN MEDIAN STDDEV MIN MAX NMISS;
RUN;
```

The detailed output for PROC MEANS.
Below is the summary of important parameters of these 7 variables:

SUMMARY OF VARIABLES PRIOR TO CLEANING

The MEANS Procedure

Variable	Label	N	Mean	Median	Std Dev	Minimum	Maximum	N Miss
duration	duration	800	154.0065385	153.9480975	49.2592338	14.7642071	305.6217107	50
no_pasg	no_pasg	850	60.1035294	60.0000000	7.4931370	29.0000000	87.0000000	0
speed_ground	speed_ground	850	79.4523229	79.6428041	19.0594903	27.7357153	141.2186354	0
speed_air	speed_air	208	103.7977237	101.1473493	10.2590370	90.0028586	141.7249357	642
height	height	850	30.1442223	30.0931324	10.2877268	-3.5462524	59.9459639	0
pitch	pitch	850	4.0093577	4.0082875	0.5288298	2.2844801	5.9267842	0
distance	distance	850	1526.02	1258.09	928.5600816	34.0807833	6533.05	0

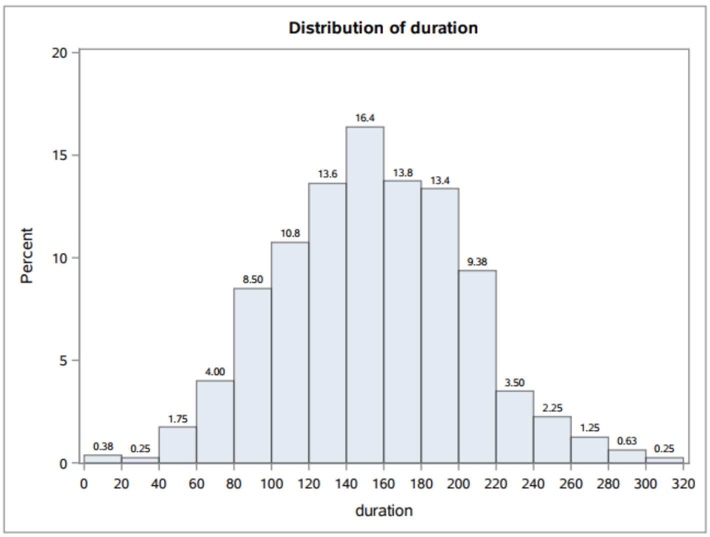
<u>Note:</u> Since its very important to observe the percentage of outliers of each variable before removing them so that we can keep a track of how many records are removed because of which variable, I am looking at the distributions of the variables and their outliers before cleaning the data.

4. <u>Notable Observations</u>:
   a. General Observations:
      i. No unique key such as 'Flight ID' is present in the datasets.
      ii. Dataset FAA2 has first 100 observations having same values as those of FAA1. These can be removed after consent of the client.
   b. Variable-specific observations:
      i. Categorical Variable: <u>Aircraft</u>
         1. 'Aircraft' is a categorical variable and the records are well distributed between the two aircraft types – boeing and airbus, thus ensuring no bias in terms of proportion.
         2. However, there are be a difference in the parameters of these two types of aircrafts which can be explored further in Bivariate analysis.

ii. Numerical Variables:

1. <u>Duration:</u>
   a. There are **15.79** % missing values of duration.
   b. It is given that duration of a flight should always be > 40 mins but from data we can see that the minimum duration is 14.76 mins. Hence, records with duration <= 40 min are outliers. This percentage is very small(0.63%) as we can see from the histogram.

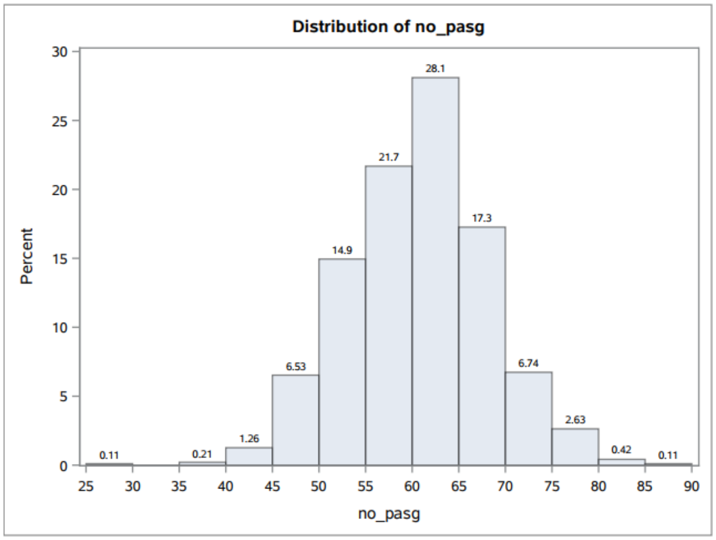
```
PROC UNIVARIATE DATA = INPUT.COMBINED NOPRINT;
    HISTOGRAM DURATION/MIDPOINTS=20 ENDPOINTS=20 TO 320 BY 20 BARLABEL=PERCENT;
RUN;
```



   c. Variable is normally distributed.

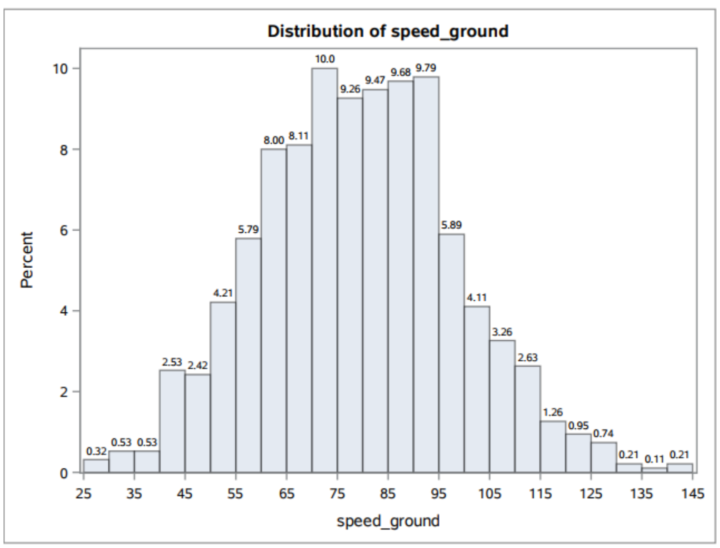
2. <u>No. of Passengers:</u>
   a. No missing values
   b. Variable is normally distributed

```
PROC UNIVARIATE DATA = INPUT.COMBINED NOPRINT;
    HISTOGRAM NO_PASG/MIDPOINTS=5 ENDPOINTS=25 TO 90 BY 5 BARLABEL=PERCENT;
RUN;
```



3. <u>Ground Speed:</u>
    a. No missing values
    b. Given range of ground speed = 30 – 140 . Data outside this interval is outlier(**1.59**%).

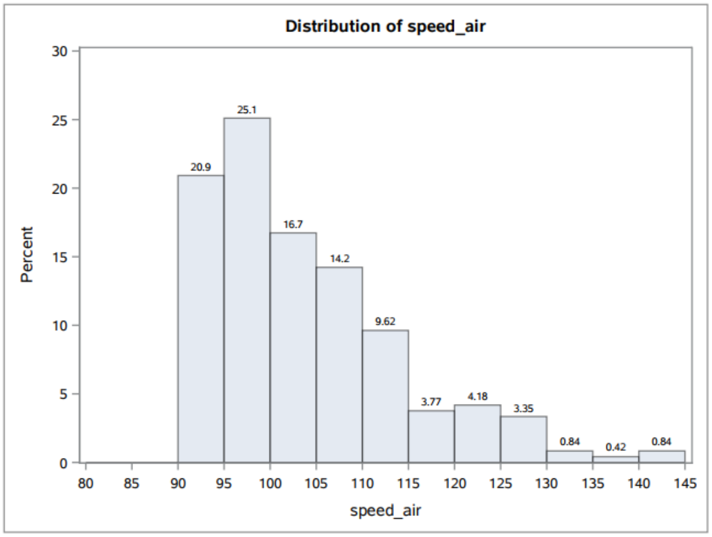
```
PROC UNIVARIATE DATA = INPUT.COMBINED NOPRINT;
    HISTOGRAM SPEED_GROUND/MIDPOINTS=5 ENDPOINTS=25 TO 145 BY 5 BARLABEL=PERCENT;
RUN;
```



    c. Data looks normally distributed.

4. <u>Air Speed :</u>
    a. **74.84** % missing values
    b. Valid range of ground speed = 30 – 140. Values outside this range are outliers (0.84%).

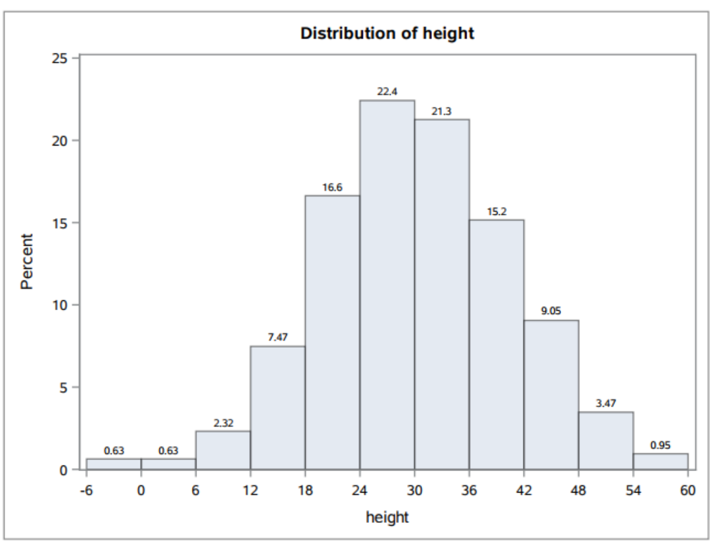
```
PROC UNIVARIATE DATA = INPUT.COMBINED NOPRINT;
    HISTOGRAM SPEED_AIR/MIDPOINTS=5 ENDPOINTS=80 TO 145 BY 5 BARLABEL=PERCENT;
RUN;
```



c. Variable is left skewed.

5. <u>Height:</u>
    a. No missing values.
    b. Minimum height required = 6 m. Values less than this are outliers(**1.26**%).

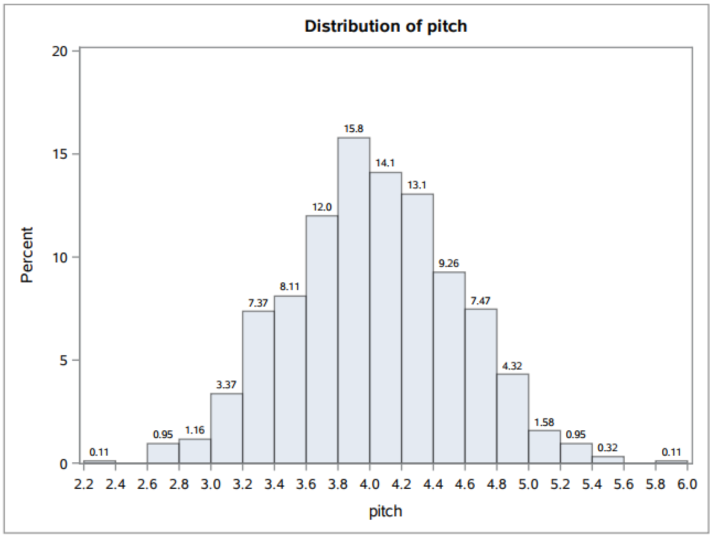
```
PROC UNIVARIATE DATA = INPUT.COMBINED NOPRINT;
    HISTOGRAM HEIGHT/MIDPOINTS=6 ENDPOINTS=-6 TO 6 BY 6 BARLABEL=PERCENT;
RUN;
```



c. Variable is normally distributed.

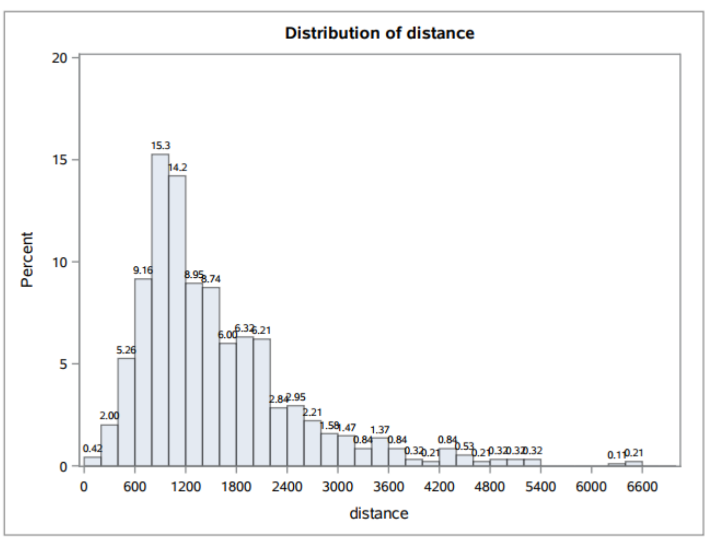
6. <u>Pitch:</u>
    a. No missing values.
    b. Data is normally distributed.

```
PROC UNIVARIATE DATA = INPUT.COMBINED NOPRINT;
    HISTOGRAM PITCH/MIDPOINTS=0.2 ENDPOINTS=2.6 TO 6 BY 0.2 BARLABEL=PERCENT;
RUN;
```



7. <u>Distance:</u>
    a. No missing values.
    b. Length is typically less than 6000m but more runway is not a problem for overrun risk.
    c. Data is left skewed but the percentage of longer runways is very less.

```
PROC UNIVARIATE DATA = INPUT.COMBINED NOPRINT;
    HISTOGRAM DISTANCE/MIDPOINTS=200 ENDPOINTS=0 TO 7000 BY 200 BARLABEL=PERCENT;
RUN;
```



DATA CLEANING:

On the basis of the above observations, the data can be cleaned using the given rules (mentioned in the above observations as well).

<u>Deduping Data:</u>

SAS Code:

```
/* DE-DUPING DATA */

PROC SORT DATA=INPUT.COMBINED NODUPKEY;
BY AIRCRAFT NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DISTANCE;
RUN;

PROC CONTENTS DATA=INPUT.COMBINED;
RUN;
```

Output:

The CONTENTS Procedure

Data Set Name	INPUT.COMBINED	Observations	850
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	09/22/2017 00:14:54	Observation Length	72
Last Modified	09/22/2017 00:14:54	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)		

<u>Variable – wise filtering:</u>

In addition to filtering every variable as given in the problem statement, I have removed Distance < 100 as passenger airplanes cannot land in such small distances and Distance >= 6000 due to abnormality. As a result, 2 additional rows have been removed.

SAS Code:

```
DATA INPUT.COMBINED_CLEANED_EXT;
SET INPUT.COMBINED;
IF SPEED_GROUND >= 30 AND SPEED_GROUND <=140;
IF (SPEED_AIR >= 30 AND SPEED_AIR <=140) OR SPEED_AIR = '.';
IF DURATION>40 AND DURATION NE '.';
IF HEIGHT>=6;
IF DISTANCE>=100 AND DISTANCE <=6000; /* SINCE PASSENGER AIRCRAFTS
CANNOT LAND IN VERY LOW DISTANCES AND DISTANCE >= 6000 IS AN ABNORMALITY */
RUN;
```

Output:

780 observations are left after cleaning.

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162	250.68976141
2	airbus	188.01797726	38	85.180842251	.	37.028793691	4.1216901717	1257.0092519
3	airbus	93.540807771	40	80.627416679	.	28.60255713	3.6234201886	1021.0888117
4	airbus	123.30242152	41	97.568203986	96.978436701	38.409192953	3.5322719834	2167.7576915
5	airbus	109.19713407	43	82.483044979	.	30.140024889	4.0896284195	1321.0000654
6	airbus	139.31381028	44	99.596841547	99.160266345	35.187030092	3.8402667146	2116.080919
7	airbus	214.22048507	45	72.490616757	.	33.228125197	4.3693164876	748.7667918
8	airbus	182.7116757	45	77.805502137	.	20.189958388	4.178015403	905.49788375
9	airbus	197.43183449	45	81.375317855	.	46.285569727	4.2052754575	1459.5022976
10	airbus	115.86922387	45	86.875879978	.	34.838071106	3.7997683715	1262.1538907

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
770	boeing	79.705863144	75	106.7461226	106.73317595	18.346201583	4.8074017332	2785.855295
771	boeing	130.94961924	76	44.732763125	.	32.782994552	4.861881592	874.79864397
772	boeing	147.03191592	76	63.597942325	.	36.489042355	4.4917734289	1051.9369604
773	boeing	219.72115595	76	88.103462433	.	42.085495821	4.6540097977	1927.0536775
774	boeing	130.16891519	77	55.086685785	.	38.032817792	4.0971206341	998.09700633
775	boeing	172.56012205	77	82.29713755	.	44.758716354	4.2293090445	1809.27205
776	boeing	228.17710591	78	61.220375598	.	21.772286622	4.5955283685	970.04651856
777	boeing	107.11331938	78	86.807962025	.	25.477015381	4.4142187986	1910.8768699
778	boeing	128.93810992	79	106.93389135	108.42651323	30.457709156	4.8421492	3203.3188407
779	boeing	161.82569155	80	82.509055403	.	36.680194026	4.685310032	1590.3719225
780	boeing	194.4671661	82	40.815188666	.	22.618444074	4.8765952309	761.4850777

<u>Summarizing cleaned data:</u>

SAS Code :

```
/* SUMMARY OF DATA AFTER CLEANING */

TITLE 'SUMMARY OF VARIABLES AFTER CLEANING';
PROC MEANS DATA = INPUT.COMBINED_CLEANED_EXT N MEAN MEDIAN STDDEV MIN MAX NMISS;
RUN;
```

Output:

SUMMARY OF VARIABLES AFTER CLEANING

The MEANS Procedure

Variable	Label	N	Mean	Median	Std Dev	Minimum	Maximum	N Miss
duration	duration	780	154.7296117	154.2603883	48.3637632	41.9493694	305.6217107	0
no_pasg	no_pasg	780	60.0602564	60.0000000	7.5066225	29.0000000	87.0000000	0
speed_ground	speed_ground	780	79.6804648	79.8275813	18.8749855	33.5741041	132.7846766	0
speed_air	speed_air	195	103.5047686	100.8916770	9.8803757	90.0028586	132.9114649	585
height	height	780	30.4749769	30.2400354	9.7297907	6.2275178	59.9459639	0
pitch	pitch	780	4.0157723	4.0153874	0.5206798	2.2844801	5.9267842	0
distance	distance	780	1543.13	1277.47	903.5729476	133.0869099	5381.96	0

DECISION/CONCLUSION:

1. 100 rows of FAA2 have same values as those of FAA1. These have been removed.
2. Removing missing duration values:
   - Total rows = **850**
     Missing duration values = 50 (**5.9**%) which is within limits and hence should be removed.
3. Air Speed should not be used for analysis as ~75% values are missing. Due to this high number, even imputation is not a good option.
   - From the distribution, it looks like values < 90 are missing from the data due to some kind of data capture issue.
4. Distance variable is left skewed but can be approximated to a normal distribution.

QUESTIONS:

1. Is there any alternate data available to compensate for the variable air speed ?
2. Is it okay to use transformations to convert a skewed distribution to a normal distribution ?
3. Should we do a univariate analysis BY Aircraft type as well ?

Chapter – 2

DESCRIPTIVE STUDY

OBJECTIVE :

The objective of this exercise is to explore X-Y plots and identify the need for any transformations. It also includes doing multivariate analysis on variables to determine correlation.

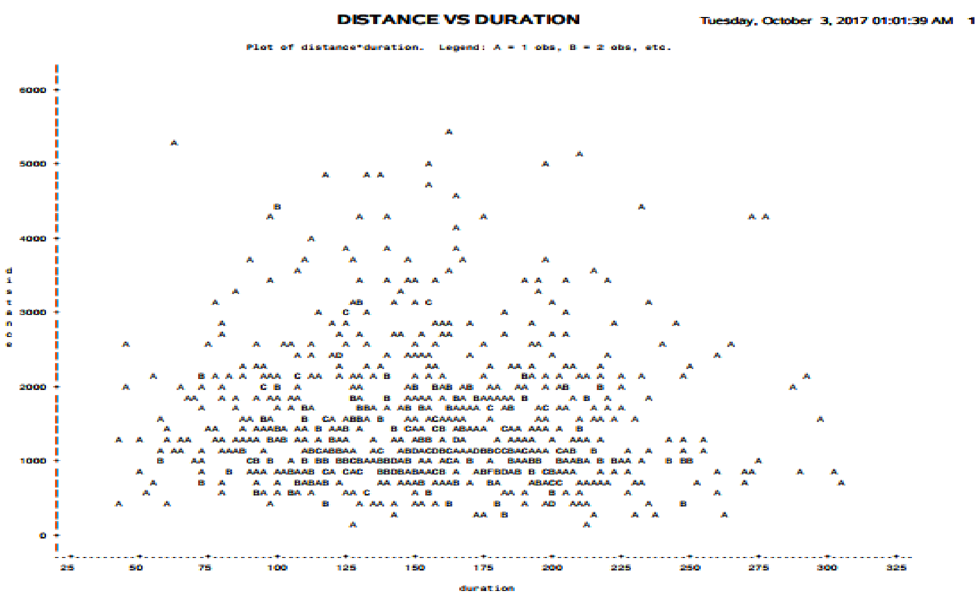
PROCEDURE :

Data visualizations involve two steps :

1) **Creating X- Y Plots:**

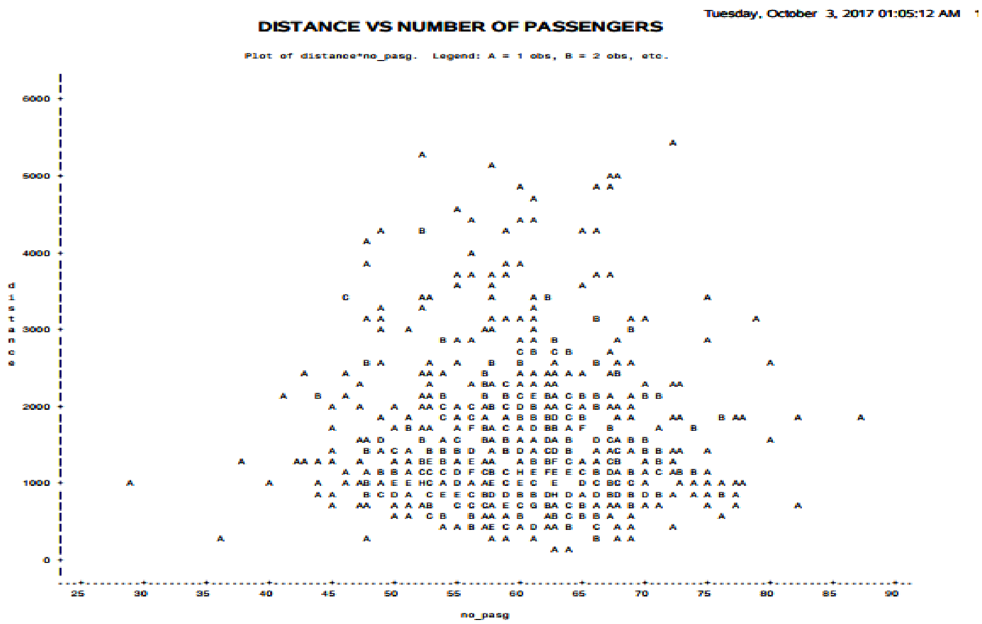
Distance vs Duration: Duration is *randomly distributed* vs Distance.

```
TITLE 'DISTANCE VS DURATION';
PROC PLOT DATA=INPUT.COMBINED_CLEANED_EXT;
    PLOT DISTANCE*DURATION;
    RUN;
```



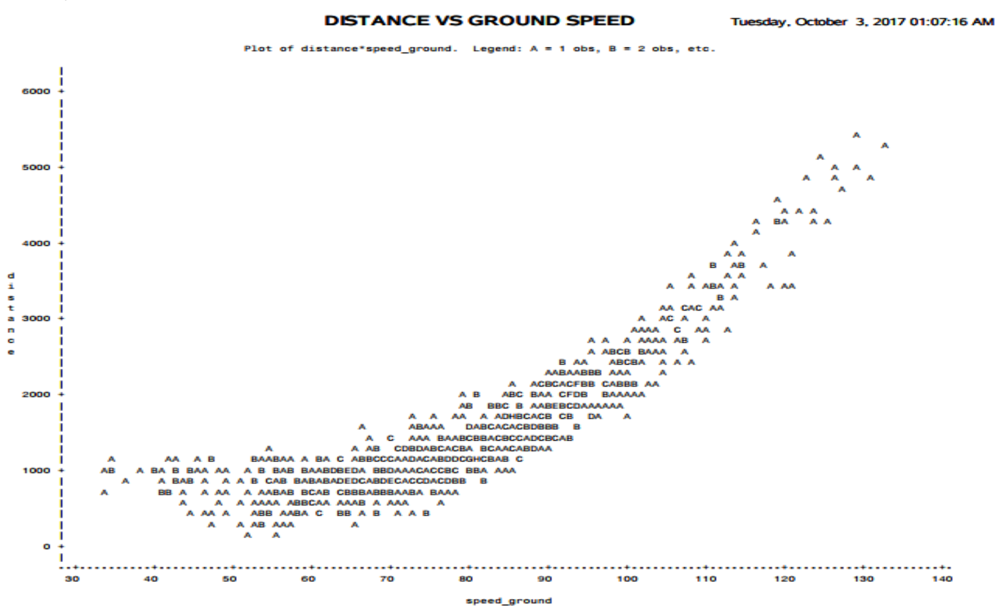
Distance vs Number of Passengers: Number of passengers is *randomly distributed* vs Distance

```
TITLE 'DISTANCE VS NUMBER OF PASSENGERS';
PROC PLOT DATA=INPUT.COMBINED_CLEANED_EXT;
    PLOT DISTANCE*NO_PASG;
    RUN;
```



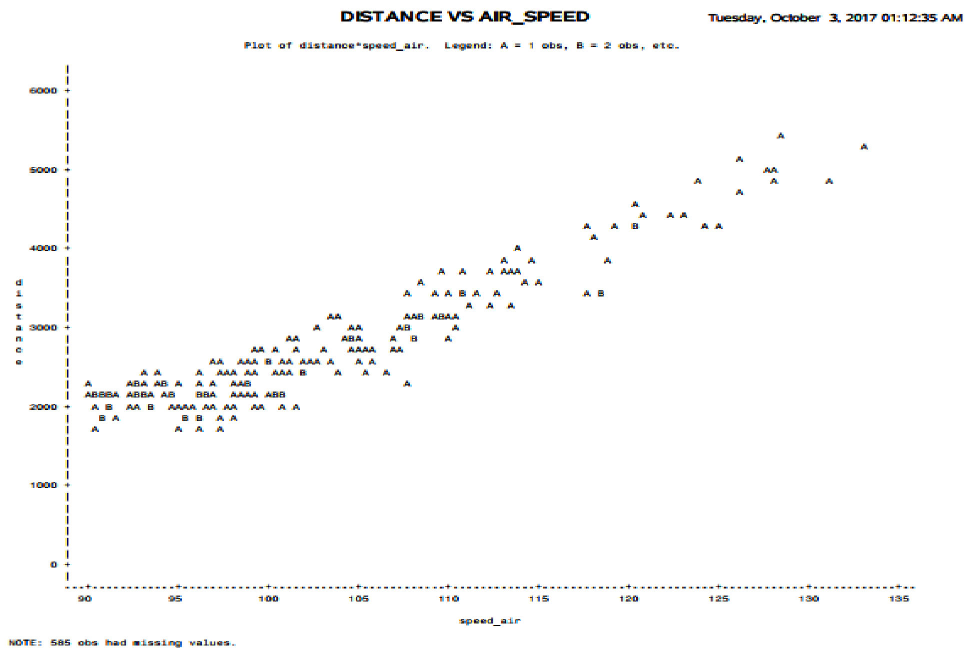
Distance vs Ground Speed: There seems to be *a trend* in Ground Speed vs Distance plot. It can be an exponential or squared trend and hence a transformation (preferably a squared one can be taken to make it linear).

```
TITLE 'DISTANCE VS GROUND SPEED';
PROC PLOT DATA=INPUT.COMBINED_CLEANED_EXT;
    PLOT DISTANCE*SPEED_GROUND;
    RUN;
```



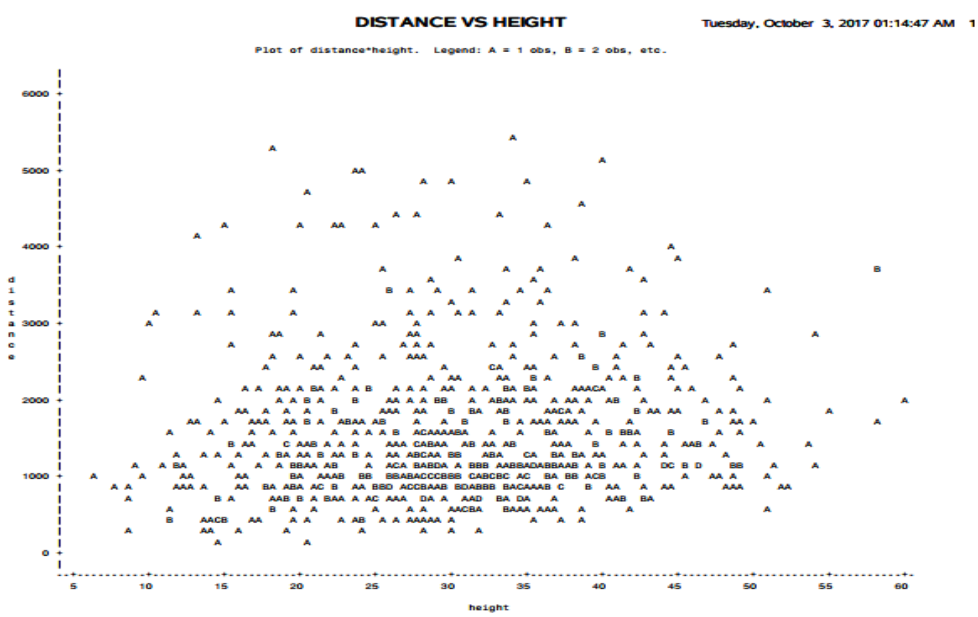
Distance vs Air Speed: The plot of Distance vs Air Speed is *highly linear*.

```
TITLE 'DISTANCE VS AIR_SPEED';
PROC PLOT DATA=INPUT.COMBINED_CLEANED_EXT;
    PLOT DISTANCE*SPEED_AIR;
    RUN;
```



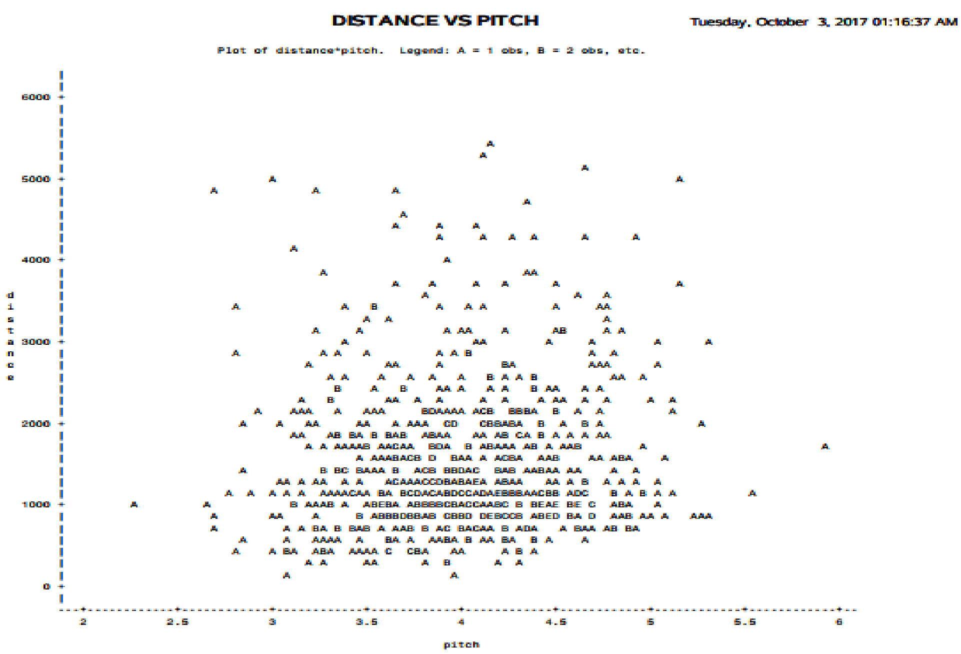
<u>Distance vs Height</u>: The plot of Distance vs Height looks *randomly distributed*.

```
TITLE 'DISTANCE VS HEIGHT';
PROC PLOT DATA=INPUT.COMBINED_CLEANED_EXT;
    PLOT DISTANCE*HEIGHT;
    RUN;
```



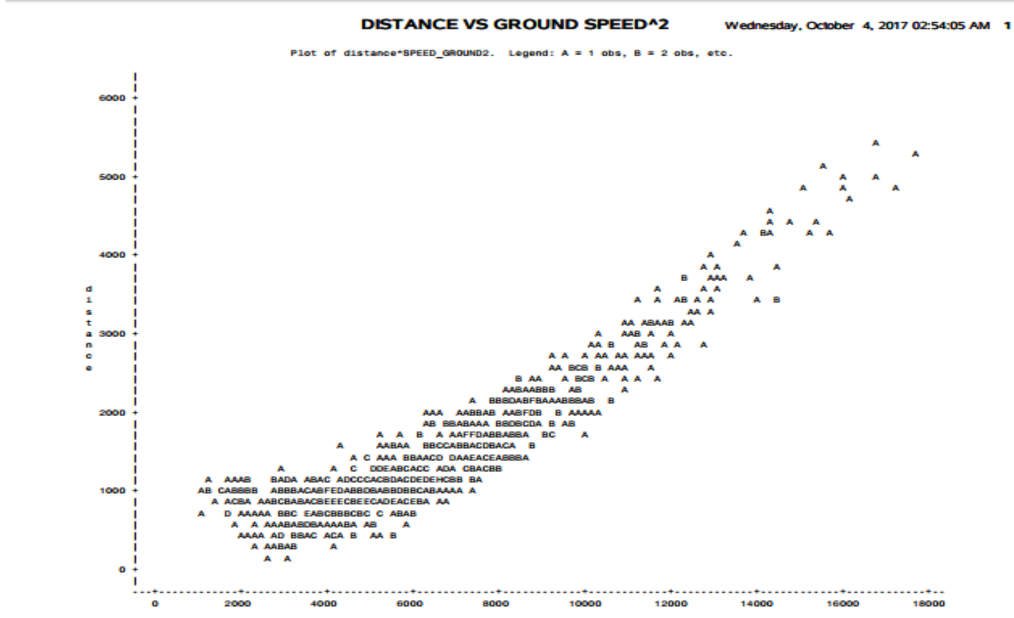
<u>Distance vs Pitch</u>: The plot of Distance vs Pitch is *randomly distributed.*

```
TITLE 'DISTANCE VS PITCH';
PROC PLOT DATA=INPUT.COMBINED_CLEANED_EXT;
    PLOT DISTANCE*PITCH;
    RUN;
```



<u>Distance vs Ground Speed^2</u>: The plot of Distance vs Ground Speed^2 is *highly linear.*

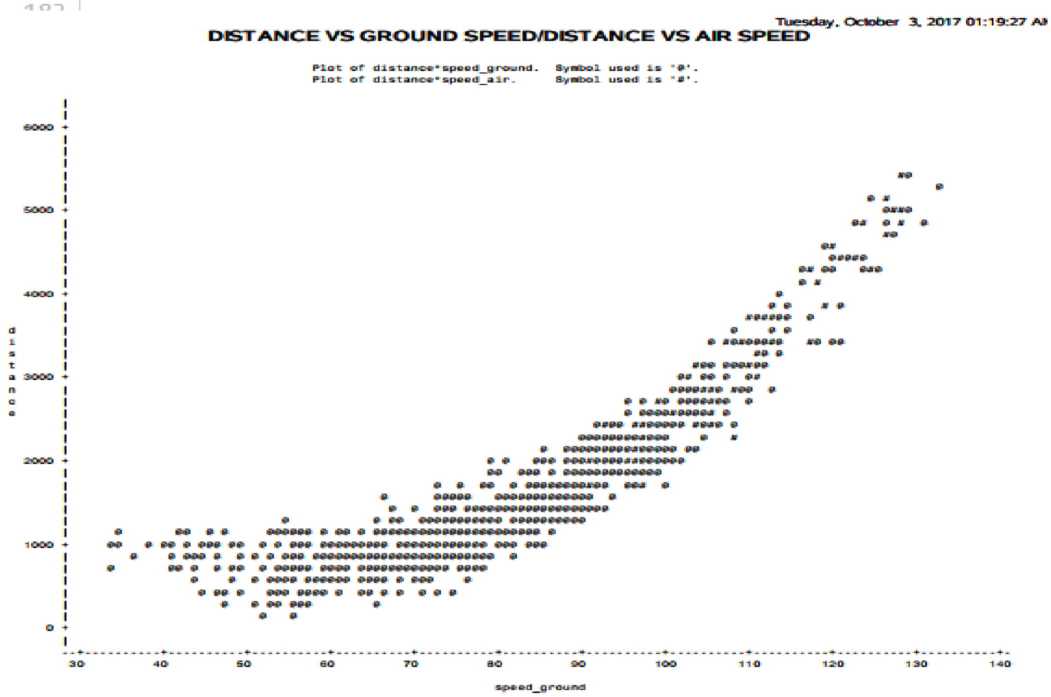
```
TITLE 'DISTANCE VS GROUND SPEED^2';
PROC PLOT DATA=INPUT.COMBINED_CLEANED_EXT_3;
    PLOT DISTANCE*(SPEED_GROUND2);
    RUN;
```



<u>Overlay Plot :</u>

<u>Distance*Speed_Ground / Distance*Speed_Air Overlay</u>: Air Speed appears to be *falling within the distribution* of ground speed. Hence, **<u>we can exclude air speed</u>** from our regression model. However, we can double check this using correlation of variables.

```
TITLE 'DISTANCE VS GROUND SPEED/DISTANCE VS AIR SPEED';
PROC PLOT DATA=INPUT.COMBINED_CLEANED_EXT;
    PLOT DISTANCE*SPEED_GROUND='@' DISTANCE*SPEED_AIR='#'/OVERLAY;
    RUN;
```



<u>2) Correlation Matrix:</u>

```
PROC CORR DATA=INPUT.COMBINED_CLEANED_EXT;
VAR DISTANCE DURATION NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH;
TITLE 'CORRELATION COEFFICIENTS WITH DIST_LOG'
RUN;
```

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

	distance	duration	no_pasg	speed_ground	speed_air	height	pitch
distance distance	1.00000 780	-0.04991 0.1638 780	-0.01213 0.7352 780	0.86724 <.0001 780	0.94322 <.0001 195	0.10065 0.0049 780	0.06382 0.0749 780
duration duration	-0.04991 0.1638 780	1.00000 780	-0.03868 0.2807 780	-0.04747 0.1854 780	0.04454 0.5364 195	0.01268 0.7237 780	-0.04460 0.2134 780
no_pasg no_pasg	-0.01213 0.7352 780	-0.03868 0.2807 780	1.00000 780	0.00338 0.9250 780	0.00002 0.9998 195	0.04214 0.2397 780	-0.00742 0.8361 780
speed_ground speed_ground	0.86724 <.0001 780	-0.04747 0.1854 780	0.00338 0.9250 780	1.00000 780	0.98835 <.0001 195	-0.05532 0.1226 780	-0.05729 0.1099 780
speed_air speed_air	0.94322 <.0001 195	0.04454 0.5364 195	0.00002 0.9998 195	0.98835 <.0001 195	1.00000 195	-0.08673 0.2280 195	-0.04827 0.5028 195
height height	0.10065 0.0049 780	0.01268 0.7237 780	0.04214 0.2397 780	-0.05532 0.1226 780	-0.08673 0.2280 195	1.00000 780	0.02985 0.4051 780
pitch pitch	0.06382 0.0749 780	-0.04460 0.2134 780	-0.00742 0.8361 780	-0.05729 0.1099 780	-0.04827 0.5028 195	0.02985 0.4051 780	1.00000 780

Observations from Correlation Matrix:

1) Ground Speed and Air Speed are ***highly correlated to the target variable Distance.***

2) Ground Speed and Air Speed are also ***highly correlated to each other***. Hence, we should exclude Air Speed from our model as it has only 195 valid observations (only 25%) and its effect is taken care of by variable Ground Speed.

3) All other variables in the correlation matrix are ***independent*** to each other.

Chapter – 3

STATISTICAL MODELING

OBJECTIVE :

The objective of this exercise is to perform modeling on the given data and get model parameters. These parameters will then help in creating an equation for linear regression.

PROCEDURE :

Data modeling including preparing data for modeling as well as building the model. This is followed by checking model diagnostics.

1) Data Preparation for modeling:

Dummitizing variable 'AIRCRAFT' – Since, SAS does not automatically dummitize categorical variables, we will need to do this manually by creating a column for AIRCRAFT CODE having value 1 for "Airbus" and 0 for "Boeing" (Variable Transformation – 1).

```
DATA INPUT.COMBINED_CLEANED_EXT_2;
SET INPUT.COMBINED_CLEANED_EXT;
FORMAT AIRCRAFT_CODE 1.;
IF AIRCRAFT = "airbus" THEN AIRCRAFT_CODE = 1;
ELSE AIRCRAFT_CODE = 0;
RUN;

PROC PRINT DATA=INPUT.COMBINED_CLEANED_EXT_2(OBS=10);
RUN;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	AIRCRAFT_CODE
1	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162	250.68976141	1
2	airbus	188.01797726	38	85.180842251	.	37.028793691	4.1216901717	1257.0092519	1
3	airbus	93.540807771	40	80.627416679	.	28.60255713	3.6234201886	1021.0888117	1
4	airbus	123.30242152	41	97.568203986	96.978436701	38.409192953	3.5322719834	2167.7576915	1
5	airbus	109.19713407	43	82.483044979	.	30.140024889	4.0896284195	1321.0000654	1
6	airbus	139.31381028	44	99.596841547	99.160266345	35.187030092	3.8402667146	2116.080919	1
7	airbus	214.22048507	45	72.490616757	.	33.228125197	4.3693164876	748.7667918	1
8	airbus	182.7116757	45	77.805502137	.	20.189958388	4.178015403	905.49788375	1
9	airbus	197.43183449	45	81.375317855	.	46.285569727	4.2052754575	1459.5022976	1
10	airbus	115.86922387	45	86.875879978	.	34.838071106	3.7997683715	1262.1538907	1

2) Modeling:

Baseline Model : Using all variables

```
PROC REG DATA=INPUT.COMBINED_CLEANED_EXT_2;
    MODEL DISTANCE = AIRCRAFT_CODE DURATION NO_PASG SPEED_GROUND HEIGHT PITCH;
    TITLE 'LINEAR REGRESSION FOR DISTANCE VS OTHER FACTORS';
RUN;
```

LINEAR REGRESSION FOR DISTANCE VS OTHER FACTORS

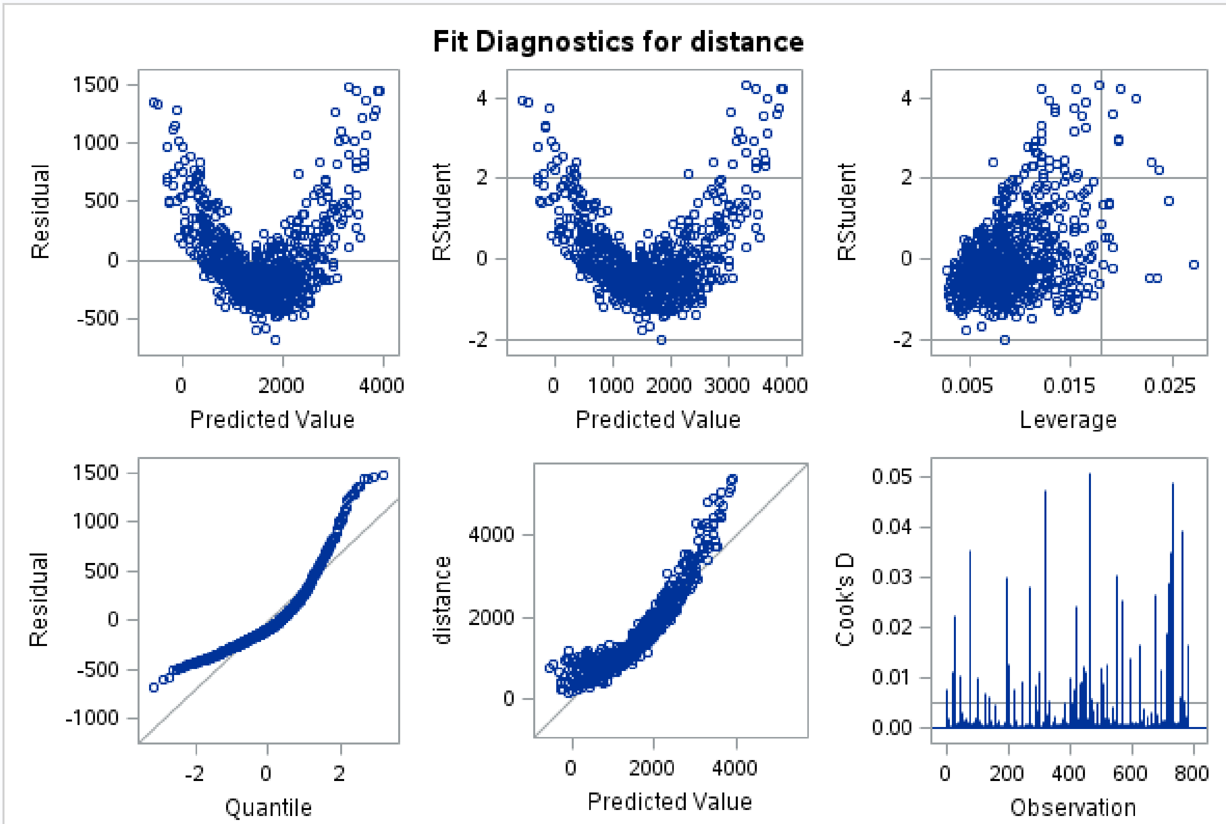
The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	780
Number of Observations Used	780

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	540794487	90132414	731.73	<.0001
Error	773	95215445	123177		
Corrected Total	779	636009932			

Root MSE	350.96512	R-Square	0.8503
Dependent Mean	1543.12635	Adj R-Sq	0.8491
Coeff Var	22.74377		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2031.11974	170.50138	-11.91	<.0001
AIRCRAFT_CODE		1	-488.91335	26.99294	-18.11	<.0001
duration	duration	1	0.04038	0.26094	0.15	0.8771
no_pasg	no_pasg	1	-1.78030	1.67848	-1.06	0.2892
speed_ground	speed_ground	1	42.61463	0.66949	63.65	<.0001
height	height	1	14.37168	1.29671	11.08	<.0001
pitch	pitch	1	21.84655	25.94634	0.84	0.4001



Inferences:

1) Duration, number of passengers and pitch **<u>are not significant</u>**. Only Aircraft Code, Ground Speed and Height are important (Air Speed had already been pulled out of consideration for modeling).

2) The overall residuals for Distance follow a curve and are not homoscedastic which is a pre-requisite condition for linear regression. Hence, we can use a transformation to make it homoscedastic.

<u>Redefining Model:</u>

1) **Variable Transformation for re-modeling:**

```
/* VARIABLE TRANSFORMATION - 2 */

DATA INPUT.COMBINED_CLEANED_EXT_3;
SET INPUT.COMBINED_CLEANED_EXT_2;
FORMAT SPEED_GROUND2 5.;
SPEED_GROUND2 = SPEED_GROUND**2;
RUN;

PROC PRINT DATA=INPUT.COMBINED_CLEANED_EXT_3(OBS=10);
RUN;
```

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	AIRCRAFT_CODE	SPEED_GROUND2
1	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162	250.68976141	1	2255
2	airbus	188.01797726	38	85.180842251	.	37.028793691	4.1216901717	1257.0092519	1	7256
3	airbus	93.540807771	40	80.627416679	.	28.60255713	3.6234201886	1021.0888117	1	6501
4	airbus	123.30242152	41	97.568203986	96.978436701	38.409192953	3.5322719834	2167.7576915	1	9520
5	airbus	109.19713407	43	82.483044979	.	30.140024889	4.0896284195	1321.0000654	1	6803
6	airbus	139.31381028	44	99.596841547	99.160266345	35.187030092	3.8402667146	2116.080919	1	9920
7	airbus	214.22048507	45	72.490616757	.	33.228125197	4.3693164876	748.7667918	1	5255
8	airbus	182.7116757	45	77.805502137	.	20.189958388	4.178015403	905.49788375	1	6054
9	airbus	197.43183449	45	81.375317855	.	46.285569727	4.2052754575	1459.5022976	1	6622
10	airbus	115.86922387	45	86.875879978	.	34.838071106	3.7997683715	1262.1538907	1	7547

2) Re-modeling:

```
PROC REG DATA=INPUT.COMBINED_CLEANED_EXT_3;
    MODEL DISTANCE = AIRCRAFT_CODE SPEED_GROUND SPEED_GROUND2 HEIGHT;
    TITLE 'LINEAR REGRESSION FOR DISTANCE VS OTHER FACTORS';
RUN;
```

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2199.32400	91.36737	24.07	<.0001
AIRCRAFT_CODE		1	-394.95050	10.49384	-37.64	<.0001
duration	duration	1	0.00954	0.10052	0.09	0.9244
no_pasg	no_pasg	1	-1.51640	0.64662	-2.35	0.0193
speed_ground	speed_ground	1	-68.94907	1.69465	-40.69	<.0001
SPEED_GROUND2		1	0.69192	0.01039	66.61	<.0001
height	height	1	13.67775	0.49965	27.37	<.0001
pitch	pitch	1	17.50047	9.99562	1.75	0.0804

3) Refining the model based on p-values:

```
/* REFINED MODEL BASED ON P-VALUES */

PROC REG DATA=INPUT.COMBINED_CLEANED_EXT_3;
    MODEL DISTANCE = AIRCRAFT_CODE SPEED_GROUND SPEED_GROUND2 HEIGHT/R;
    TITLE 'LINEAR REGRESSION FOR DISTANCE VS OTHER FACTORS'/R;
    OUTPUT OUT=INPUT.DIAGNOSTICS R=RESIDUALS;
RUN;
```

LINEAR REGRESSION FOR DISTANCE VS OTHER FACTORS/R

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

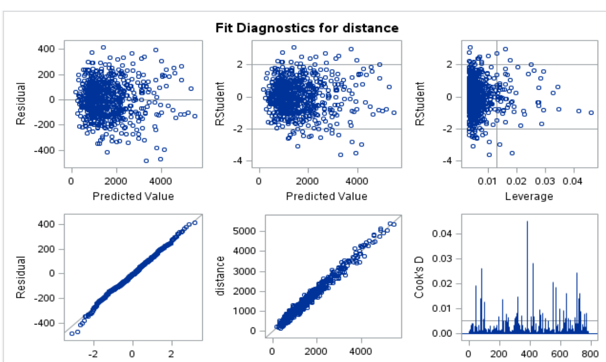
Number of Observations Read	780
Number of Observations Used	780

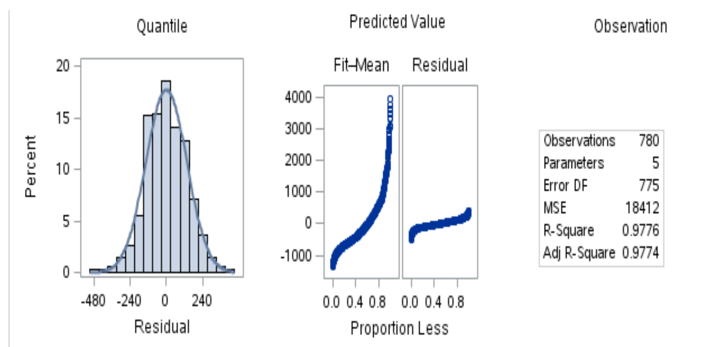
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	621740717	155435179	8442.11	<.0001
Error	775	14269214	18412		
Corrected Total	779	636009932			

Root MSE	135.69042	R-Square	0.9776
Dependent Mean	1543.12635	Adj R-Sq	0.9774
Coeff Var	8.79322		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2187.49597	69.08938	31.66	<.0001
AIRCRAFT_CODE		1	-401.95385	9.83725	-40.86	<.0001
speed_ground	speed_ground	1	-69.01536	1.70046	-40.59	<.0001
SPEED_GROUND2		1	0.69220	0.01042	66.40	<.0001
height	height	1	13.66057	0.50064	27.29	<.0001

Output Statistics							
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D
1	251	260.1629	16.4830	-9.4732	134.7	-0.070	0.000
2	1257	1435	8.1879	-177.9982	135.4	-1.314	0.001
3	1021	1112	7.5363	-90.4630	135.5	-0.668	0.000
4	2168	2166	9.1143	1.8335	135.4	0.014	0.000
5	1321	1214	7.4778	107.0039	135.5	0.790	0.000
6	2116	2259	8.8215	-142.6815	135.4	-1.054	0.001
7	749	873.9035	7.6240	-125.1367	135.5	-0.924	0.001
8	905	881.9166	9.1150	23.5812	135.4	0.174	0.000
9	1460	1385	10.8808	74.1393	135.3	0.548	0.000
10	1262	1490	7.8253	-227.8172	135.5	-1.682	0.002





Model Interpretation:

We can summarize the coefficient estimates of different variables in the model as under:

Factor	Coefficient
Aircraft Code	-401.95385
Ground Speed	-69.01536
Square of Ground Speed	0.69220

Height	13.66

We can now build the equation for linear regression using these estimates. Given the intercept of 2187.5, the linear regression equation is:

Distance = 2187.5 + (-401.95) * Aircraft Code + (-69) * Ground Speed + 0.69 * Square of Ground Speed + 13.66 * Height

The equation can be interpreted in terms of different variables as:

Aircraft Code: For change in aircraft from boring to airbus, the distance estimate reduces by 401.95 meters.

Ground Speed: For 1 unit change in ground speed, distance changes by -69 metres due to linear value but for every unit change in square value of ground speed, distance changes by 0.69 meters.

Height: For 1 unit change in heightm the distance changes by 13.66 metres.

3) Model Diagnostics:

Checking for following assumptions on the residuals:

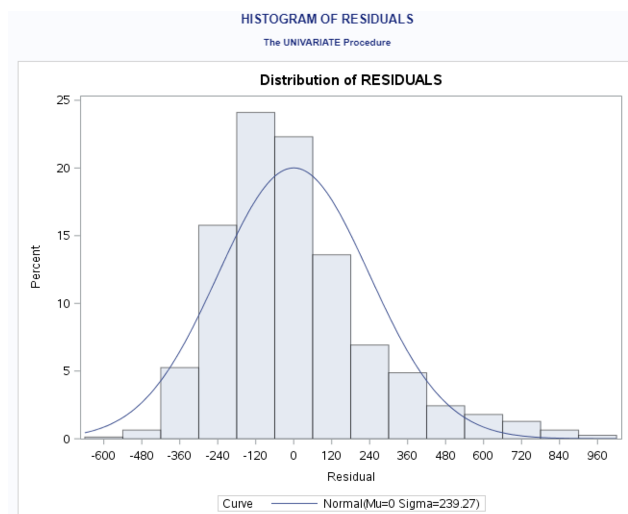
1) Independent 2) Normally Distributed 3) Mean = 0 4) Constant Variance

HISTOGRAM OF RESIDUALS

The UNIVARIATE Procedure
Fitted Normal Distribution for RESIDUALS (Residual)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	239.2711

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.0950522	Pr > D	<0.010
Cramer-von Mises	W-Sq	2.2884986	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	13.8744452	Pr > A-Sq	<0.005



Inference:

From above we can see that the residuals follow all the assumptions and their normality can be proved through p<0.01 in Kolmogorov-Smirnov, Cramer-con Mises and Anderson Darling tests above. Their constant variance can be seen from Residuals plot in the model.

Write Short answers to questions:

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

Ans: I used a total of 780 observations in my final model. Remaining 150 got removed as a part of the cleaning process below:

S.No.	Action	Rows Removed	Remaining
1	De-duplicating the combination of two datasets	100	850
2	30 <= Ground Speed <= 140	3	847
3	30 <= Air Speed <= 140	1(common with above)	847
4	Duration >40 and duration is not missing	55	792
5	Height >= 6	10	782
6	100 <= Distance <= 6000	2	780

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

Ans: The effect of all factors on impact of Distance is given by the following equation:

Distance = 2187.5 + (-401.95) * Aircraft Code + (-69) * Ground Speed + 0.69 * Square of Ground Speed + 13.66 * Height

Number of passengers, **duration and pitch** do not significantly affect the model as can be seem from their X-Y plots as well as linear regression p-values.

Aircraft code, **Ground Speed**, **Square of Ground Speed** and **Height** significantly affect the model and 98% of the variance is explained by these factors.

Aircraft code has a step effect whereas height as a positive liner relationship with distance. Ground speed affects distance in a combination of linear form and squared form.

The relationship of significant variables can also be seen from their X-Y plots and p-values in the model.

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

Ans: Doing Regression by Variable 'Aircraft'

```
PROC REG DATA=INPUT.COMBINED_CLEANED_EXT_3;
    MODEL DISTANCE = SPEED_GROUND HEIGHT PITCH;
    BY AIRCRAFT;
    TITLE 'LINEAR REGRESSION FOR DISTANCE VS OTHER FACTORS';
RUN;
```

For aircraft = airbus:

For aircraft = boeing :

Root MSE	126.61103	R-Square	0.9753
Dependent Mean	1338.44280	Adj R-Sq	0.9750
Coeff Var	9.45958		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1387.37847	118.32718	11.72	<.0001
speed_ground	speed_ground	1	-69.35575	2.65386	-26.13	<.0001
SPEED_GROUND2		1	0.69276	0.01618	42.80	<.0001
height	height	1	13.49183	0.65769	20.51	<.0001
pitch	pitch	1	111.48717	13.29018	8.39	<.0001

Root MSE	129.03533	R-Square	0.9819
Dependent Mean	1750.98330	Adj R-Sq	0.9817
Coeff Var	7.36931		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2540.52743	102.67600	24.74	<.0001
speed_ground	speed_ground	1	-69.92102	2.02139	-34.59	<.0001
SPEED_GROUND2		1	0.69882	0.01244	56.20	<.0001
height	height	1	13.45304	0.67850	19.83	<.0001
pitch	pitch	1	-75.92910	13.46232	-5.64	<.0001

Observation : Though the RMSE is same for both, there is a large difference in coefficient of pitch as well as Intercept.

Further exploring by *T-Test* on Distance and Pitch:

Distance:

```
PROC TTEST DATA=INPUT.COMBINED_CLEANED_EXT_3;
VAR DISTANCE;
CLASS AIRCRAFT;
RUN;
```

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	778	-6.54	<.0001
Satterthwaite	Unequal	751.49	-6.54	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	386	392	1.42	0.0006

Since, the variances are equal, through Pooled method, null hypothesis is rejected, so there is a significant difference between landing distances of two types of aircrafts.

Pitch:

```
PROC TTEST DATA=INPUT.COMBINED_CLEANED_EXT_3;
VAR PITCH;
CLASS AIRCRAFT;
RUN;
```

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	778	-10.78	<.0001
Satterthwaite	Unequal	777.43	-10.78	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	386	392	1.02	0.8161

Since, the variances are unequal, through Satterthwaite method, null hypothesis is rejected, so there is a significant difference between pitch of two types of aircrafts.