FAA PROJECT

Abstract

On analysis of 83.6% of the data provided, also categorized as the Normal clean data, the following has been observed and proved:

- 86.67% positive linear dependence of distance on speed_ground
- 24% dependence of distance on aircraft (mean distance of Boeing is 24.74% greater than Airbus)
- 10% positive linear dependence of distance on height
- 35% dependence of pitch on aircraft (mean pitch of Boeing is 9.5% greater than Airbus)
- Conclusion: The Landing threshold can be overshot with increase in speed on ground, increase in height, and more for Boeing aircrafts in comparison to Airbus.

Keerthi Gopalakrishnan
M12931398

FAA PROJECT: CHAPTER 1

DATA PREPARATION

Goal Statement:

The first chapter of the FAA Project is dedicated to preparing the data for further analysis. This involves combining the data, observing the variables, observing the missing values trend, and cleansing the data keeping sample size in mind. Thereby making it ready to use for advance analysis.

Implementation & Methodology:

1. **Data Import and Observation:**

   The data was imported onto data sets. It was combined and then observed.

   Code:

```
/* Combining the faa1 and faa2 data to perform analysis */
DATA Combined_Dataset;
    SET faa1 faa2_altered;

    /* below gives a basic understanding of the distribution of variables */
PROC MEANS DATA=combined_dataset n nmiss mean median stddev min max var;
RUN;
```

   Result:

FAA Combined Raw Data

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum	Variance
duration	duration	800	200	154.0065385	153.9480975	49.2592338	14.7642071	305.6217107	2426.47
no_pasg	no_pasg	950	50	60.1652632	60.0000000	7.4900041	29.0000000	87.0000000	56.1001619
speed_ground	speed_ground	950	50	79.2849940	79.4129094	19.3364178	27.7357153	141.2186354	373.8970524
speed_air	speed_air	239	761	103.7304174	100.8916770	10.6051134	90.0028586	141.7249357	112.4684305
height	height	950	50	30.1392714	29.9044945	10.3593491	-3.5462524	59.9459639	107.3161131
pitch	pitch	950	50	4.0192472	4.0153874	0.5260322	2.2844801	5.9267842	0.2767099
distance	distance	950	50	1548.82	1267.44	948.6812561	34.0807833	6533.05	899996.13

- On observing the above MEAN PROC data of combined raw data set, many samples fall under missing values mainly under ‘speed_air’ and ‘duration’.

Variable	Label	N	NPERC %	N Miss	N MissPERC %
duration	duration	800	80	200	20
no_pasg	no_pasg	950	95	50	5
speed_ground	speed_ground	950	95	50	5
speed_air	speed_air	239	23.9	761	76.1
height	height	950	95	50	5
pitch	pitch	950	95	50	5
distance	distance	950	95	50	5

- The percentage of missing values for a single variable ‘speed_air’ is 76.1% which is very high in terms of samples.

2. Data Handling

The following concepts have been applied to handle these missing values:

- All the missing values in ‘duration’ is from Faa2 sheet as it does not contain the column duration. If all missing values are deleted, that leads to completely omitting Faa2 sheet. This will not lead to good quality analytical results.

- Variable ‘speed_air’ contain several missing values which can affect the final sample size. On observing the values of speed_air and speed_ground, the values are in close approximation with each other. Hence instead of deleting these missing values, we continue our analysis on basis of the samples we have in ‘speed_air’ and ‘speed_ground’.

3. Data Cleaning:

- Now that all missing values will not be deleted, only those will be deleted that are common to the most relevant fields. From above table it can be seen that 50 missing values are common to all fields. That is, 5% of the total values will be deleted as it will not add to the final result and it will not affect the sample size largely.
- This step will also remove the presence of any duplicate values. Duplicate values have been checked on the basis of two variables ‘speed_ground’ and ‘distance’, owing to their unique decimal point structure.

Code:

```
/* Data Cleaning and Duplicate removal */

Data combined_dataset_clean;
    SET combined_dataset;
    IF distance=. then
        delete;
RUN;

proc sort data=combined_dataset_clean out=clean_ndups nodupkey;
by speed_ground distance;
run;

PROC MEANS DATA=clean_ndups n nmiss mean median stddev min max var;
title FAA Combined Clean Ndups Data;
RUN;
```

Result:

FAA Combined Clean Ndups Data

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum	Variance
duration	duration	800	50	154.0065385	153.9480975	49.2592338	14.7642071	305.6217107	2426.47
no_pasg	no_pasg	850	0	60.1035294	60.0000000	7.4931370	29.0000000	87.0000000	56.1471018
speed_ground	speed_ground	850	0	79.4523229	79.6428041	19.0594903	27.7357153	141.2186354	363.2641710
speed_air	speed_air	208	642	103.7977237	101.1473493	10.2590370	90.0028586	141.7249357	105.2478395
height	height	850	0	30.1442223	30.0931324	10.2877268	-3.5462524	59.9459639	105.8373237
pitch	pitch	850	0	4.0093577	4.0082875	0.5288298	2.2844801	5.9267842	0.2796610
distance	distance	850	0	1526.02	1258.09	928.5600816	34.0807833	6533.05	862223.83

- 150 samples were missing or duplicate values, that is, 15% of the samples were cleaned out to obtain a resulting data set of 85% of the samples, which is 850 samples.

4. Data Categorization:

- This is an extremely crucial step to the project. In the previous result, the minimum and maximum clearly show that there are many outliers, also known as ‘abnormal samples’. Categorizing is done to ensure that there are two datasets that can be separately observed and compared. First is the normal data (based on the project information), second is the abnormal data.
- Though not explicitly mentioned, a threshold of 6000 has been applied for ‘distance’. Any samples above 6000 feet ‘distance’ has been categorized as abnormal.

Code:

```
/* calculate %*/

DATA ABNORMAL NORMAL;
    SET clean_ndups;

    IF speed_ground < 30 OR speed_ground > 140 OR height < 6 OR distance > 6000 THEN
            OUTPUT ABNORMAL;
    ELSE
        OUTPUT NORMAL;
RUN;

/* Understanding of the variable distribution of abnormal and normal datasets */
PROC MEANS DATA=abnormal n nmiss mean median stddev min max var;
TITLE FAA Abnormal Cleaned Data;
RUN;

PROC MEANS DATA=normal n nmiss mean median stddev min max var;
TITLE FAA Normal Cleaned Data;
RUN;
```

Result:

FAA Abnormal Cleaned Data

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum	Variance
duration	duration	14	0	158.0905971	148.4950577	46.2337949	103.0908467	283.7633684	2137.56
no_pasg	no_pasg	14	0	63.8571429	64.0000000	7.6445772	46.0000000	73.0000000	58.4395604
speed_ground	speed_ground	14	0	70.9674145	63.3940816	33.2506376	27.7357153	141.2186354	1105.60
speed_air	speed_air	2	12	139.0741785	139.0741785	3.7487367	136.4234214	141.7249357	14.0530270
height	height	14	0	8.2729105	1.7295249	14.6050499	-3.5462524	44.2861092	213.3074840
pitch	pitch	14	0	4.2689172	4.2907474	0.5697167	3.1225584	5.2168023	0.3245771
distance	distance	14	0	1524.18	690.1269688	2113.28	34.0807833	6533.05	4465961.53

FAA Normal Cleaned Data

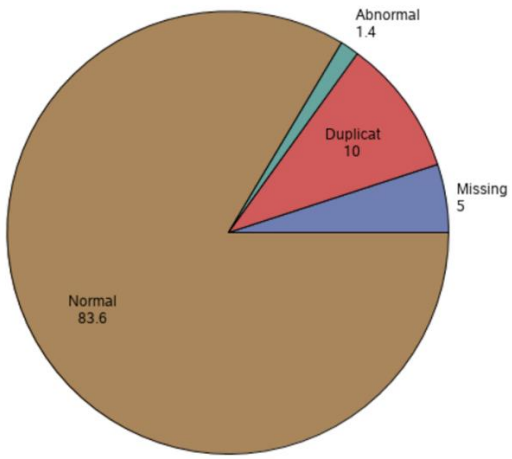
The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum	Variance
duration	duration	786	50	153.9337944	154.1281058	49.3360402	14.7642071	305.6217107	2434.04
no_pasg	no_pasg	836	0	60.0406699	60.0000000	7.4792021	29.0000000	87.0000000	55.9384637
speed_ground	speed_ground	836	0	79.5944146	79.8275813	18.7327127	33.5741041	132.7846766	350.9145236
speed_air	speed_air	206	630	103.4552338	101.1070213	9.6926499	90.0028586	132.9114649	93.9474613
height	height	836	0	30.5104883	30.2095636	9.8049102	6.2275178	59.9459639	96.1362641
pitch	pitch	836	0	4.0050110	4.0023169	0.5273975	2.2844801	5.9267842	0.2781481
distance	distance	836	0	1526.05	1263.54	898.4154244	41.7223127	5381.96	807150.27

- The normal data set has 836 samples, that is 83.6% of the total data provided can be used for advanced analysis, that can be trusted to provide high quality results.
- 1.4% of the total data is abnormal and is best to not be included in the analysis of generalization of variables' dependence.

Conclusion

- 76.1% samples of the total data had missing values. Hence Data was not deleted on basis of only missing values.
- 5% of the total data was missing values common to all variables. These common missing values' rows were deleted.
- 10% of the total data were duplicate values. These were removed too.
- 85% of the samples were found to be clean and can be used for analysis after cleaning of data.
- 83.6% of the total data has normal values and 1.4% has abnormal values, based on threshold conditions. The normal data set is the final one that will be used for advanced analysis and to draw conclusions.



-

Above shows the pie chart of sample distribution represented in %.

Questions and Further Details

1) Is there a reason behind having two datasheets (FAA1 and FAA2)? What is the relevance of each or how are they different?
2) What is the reason behind not have duration in sheet FAA2?
3) It has been mentioned that the runway length is 6000 feet. However, if we see maximum of 'distance' some flights have surpassed 6000 feet. Can a threshold be suggested for 'distance'? Like 5200 feet.
4) How old are the samples in the provided data sets, from which year have they been taken?
5) What range do these samples cover (range of data representation, are there samples belonging to more than one airport?)?
6) Is it possible to receive information on the runway threshold, and touch down zone area?
7) What were the measures taken earlier to avoid landing overrun?
8) Details on weather conditions in and around the airport will be required, along with the geographical nature/landscape around the runway.
9) Do these flights in the sheet belong to day or night or mix of both, is it possible to categorize them?
10) A map of the infrastructure of the airport, including allocated parking areas would be helpful.

FAA PROJECT: CHAPTER 2

EXPLORATORY DATA ANALYSIS

Goal Statement:

This chapter is aimed at analyzing the data in all permutations and combinations of variables. Variables will be distributed in terms of one another and the results will be interpreted in order to draw valuable conclusions.

Implementation and Methodology:

1. **Observing Data types**:
   - This step is performed to understand the data type each variable belongs to. This will be essential in order to decide the best analytical model that can be used for the variable.

   Code:

```
62 /* content */
63 proc contents data=normal;
64 run;
```

   Results:

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	6	$6.	$6.	aircraft
8	distance	Num	8	BEST12.		distance
2	duration	Num	8	BEST12.		duration
6	height	Num	8	BEST13.		height
3	no_pasg	Num	8	BEST8.		no_pasg
7	pitch	Num	8	BEST12.		pitch
5	speed_air	Num	8	BEST12.		speed_air
4	speed_ground	Num	8	BEST12.		speed_ground

   - It can be seen that aircraft is a character variable, and will require a different analytical model in comparison to the other distribution variables.

2. **TTEST Analysis of variables across 'aircraft' type**:

   - Since aircraft is a categorical variable, using a two sample TTEST on its dependence on variables will be the first step to analysing inter-dependence.
   - The analysis will be done based on the hypothesis that if the mean of the distributed variables are equal for 'Boeing' and

- Hypothesis:

    - H0: Mean of distributed variables (distance, height, pitch, speed_air, speed_ground, no_pasg, duration) are equal across both types of aircrafts.
      H0 : **μ1= μ2**
    - Ha: Mean of distributed variables (distance, height, pitch, speed_air, speed_ground, no_pasg, duration) are not equal across both types of aircrafts.
      Ha : **μ1!= μ2**

Code:

```
66 /* TTEST for aircraft */
67
68 PROC TTEST DATA=NORMAL;
69 CLASS AIRCRAFT;
70 VAR distance speed_air speed_ground height duration pitch no_pasg;
71 RUN;
72
```

Result:

The TTEST Procedure

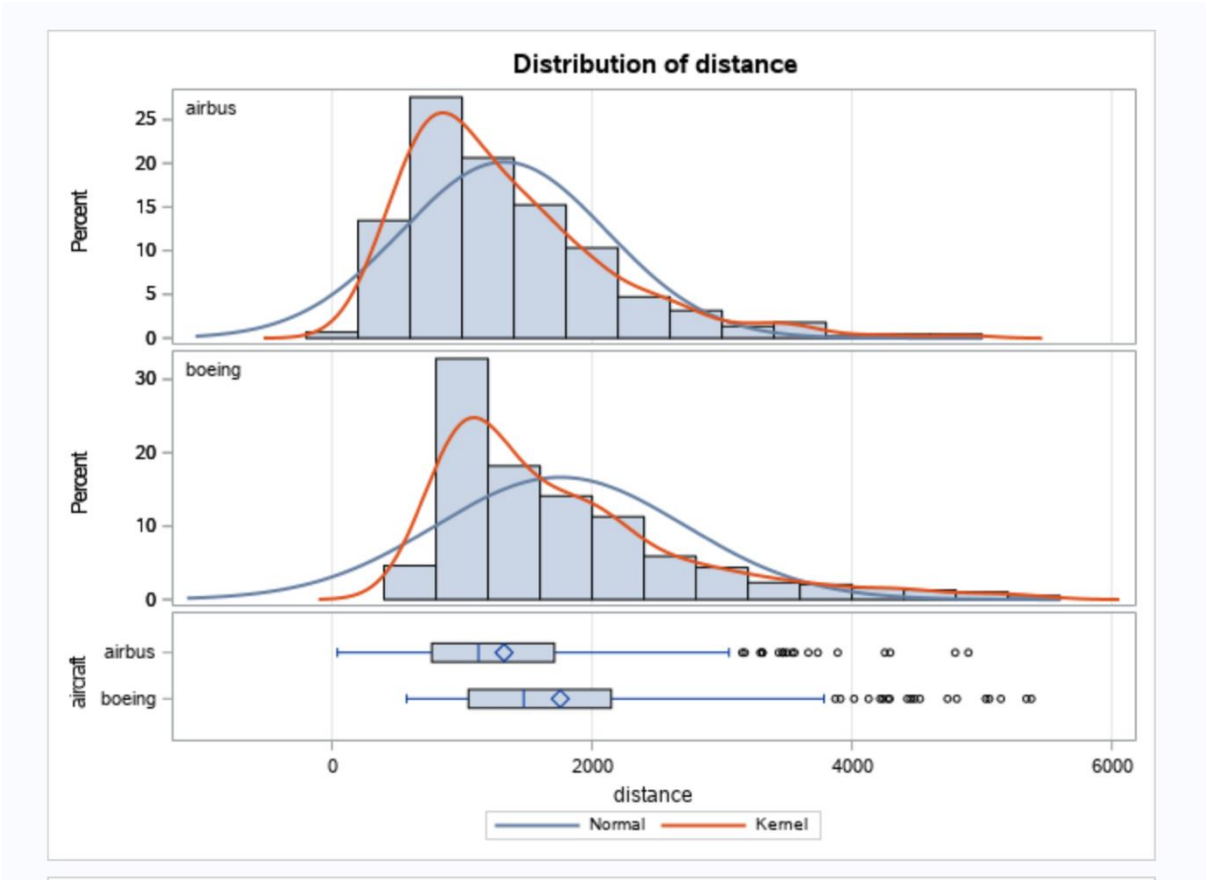
Variable: distance (distance)

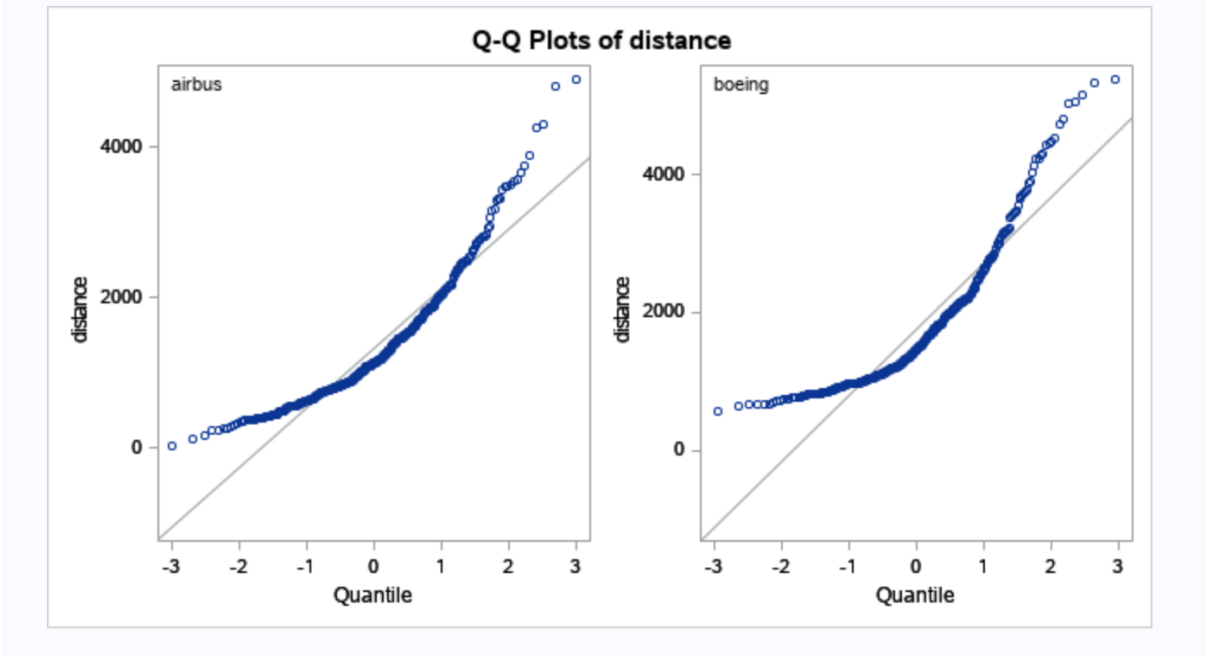
aircraft	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus		446	1324.4	791.3	37.4714	41.7223	4896.3
boeing		390	1756.7	957.2	48.4691	573.6	5382.0
Diff (1-2)	Pooled		-432.4	872.6	60.4970		
Diff (1-2)	Satterthwaite		-432.4		61.2647		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1324.4	1250.7	1398.0	791.3	742.6	847.0
boeing		1756.7	1661.4	1852.0	957.2	894.4	1029.5
Diff (1-2)	Pooled	-432.4	-551.1	-313.6	872.6	832.7	916.6
Diff (1-2)	Satterthwaite	-432.4	-552.6	-312.1			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	834	-7.15	<.0001
Satterthwaite	Unequal	756.67	-7.06	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	389	445	1.46	0.0001


Distribution of distance
airbus
boeing
Percent
aircraft
distance
Normal
Kernel


Q-Q Plots of distance
airbus
boeing
distance
Quantile

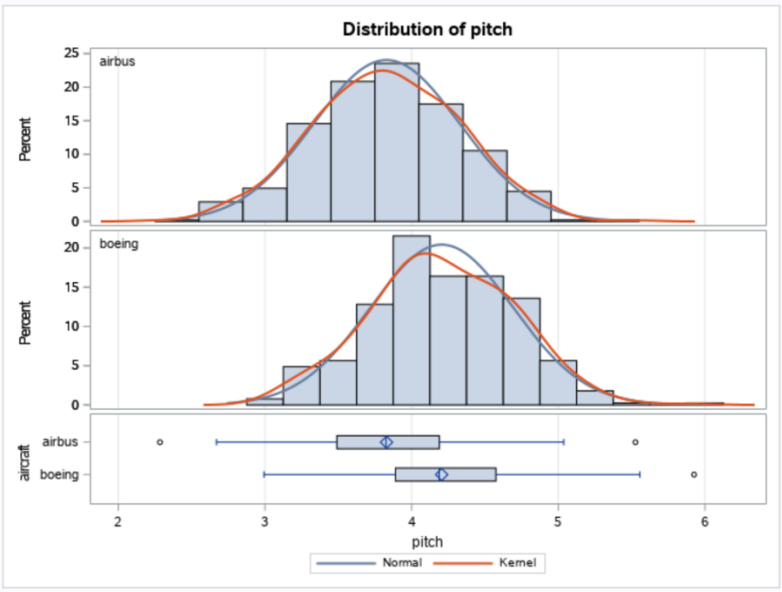
Variable: pitch (pitch)

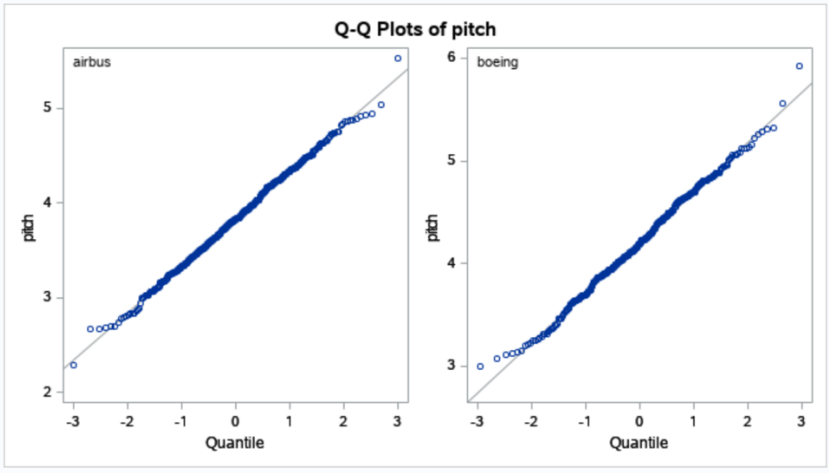
aircraft	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus		446	3.8296	0.4976	0.0236	2.2845	5.5268
boeing		390	4.2056	0.4881	0.0247	2.9932	5.9268
Diff (1-2)	Pooled		-0.3759	0.4932	0.0342		
Diff (1-2)	Satterthwaite		-0.3759		0.0341		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		3.8296	3.7833	3.8759	0.4976	0.4669	0.5325
boeing		4.2056	4.1570	4.2542	0.4881	0.4561	0.5250
Diff (1-2)	Pooled	-0.3759	-0.4430	-0.3088	0.4932	0.4706	0.5180
Diff (1-2)	Satterthwaite	-0.3759	-0.4430	-0.3089			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	834	-11.00	<.0001
Satterthwaite	Unequal	823.05	-11.01	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	445	389	1.04	0.6986





Procedure

- From above charts and tables the following test information is inferred:

  a. The Pr > F is less than 0.05 for **Distance,** hence considering the Satterthwaite model's p-value, 0.0001 < 0.05. Therefore, there is strong evidence to reject the **H0.** That is, the mean of distance of 'Boeing' is different from the mean of 'Airbus'. This can be interpreted to inform us that the type of 'aircraft' and 'distance' are closely interrelated at a 95% confidence level.
  b. The Pr > F is greater than 0.05 for **Pitch,** hence considering the Pooled model's p-value, 0.0001 < 0.05. Therefore, there is strong evidence to reject the **H0.** That is, the mean of pitch of 'Boeing' is different from the mean of 'Airbus'. This can be interpreted to inform us that the type of 'aircraft' and 'pitch' are closely interrelated at a 95% confidence level.
  c. Using similar interpretation, it was observed that height, speed_air, speed_ground, no_pasg, duration have p-values of 0.764, 0.32, 0.26, 0.5,0.188. Hence these variables are not inter-dependent on 'aircraft'. That is, **H0** was not rejected.
  d. It can also be suggested that distance has a larger influence of 'aircraft' than pitch. On observing the difference in mean, the distance mean for Boeing is 32.62% greater than the distance mean for Airbus. Whereas for pitch the mean of Boeing is greater by 10.52%.
  e. It can be concluded that landing distance is greater for Boeing than Airbus on an average. Pitch is greater for Boeing than Airbus on an average.

3. X-Y Plot Analysis:

- This step of analysis is extremely crucial in finding out the influence of distributed variables on each other.
- Since aircraft is a character variable, it is converted to a numerical form of 1s and 0s. Where 1 represents Airbus and 0 represents Boeing.

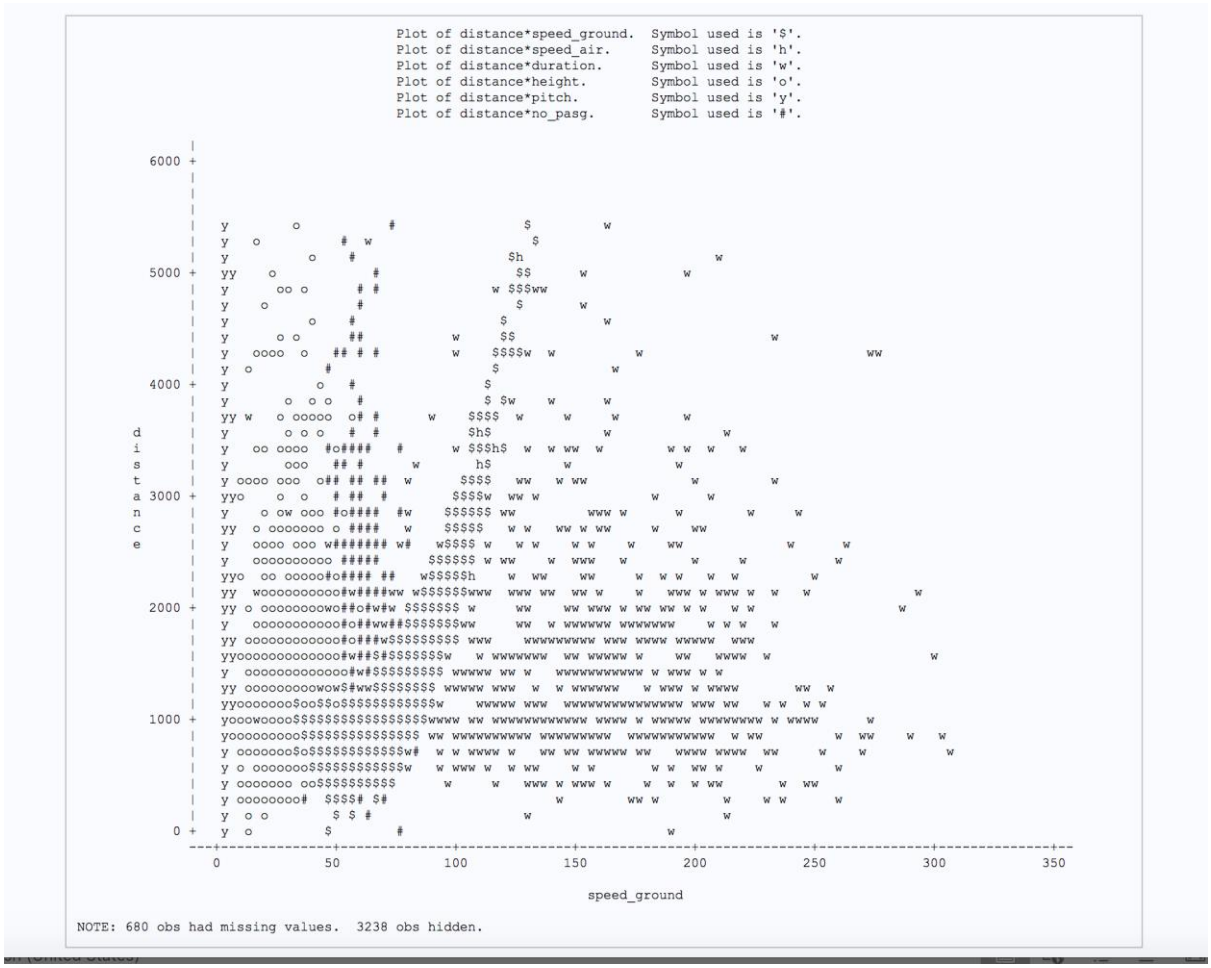
Code:

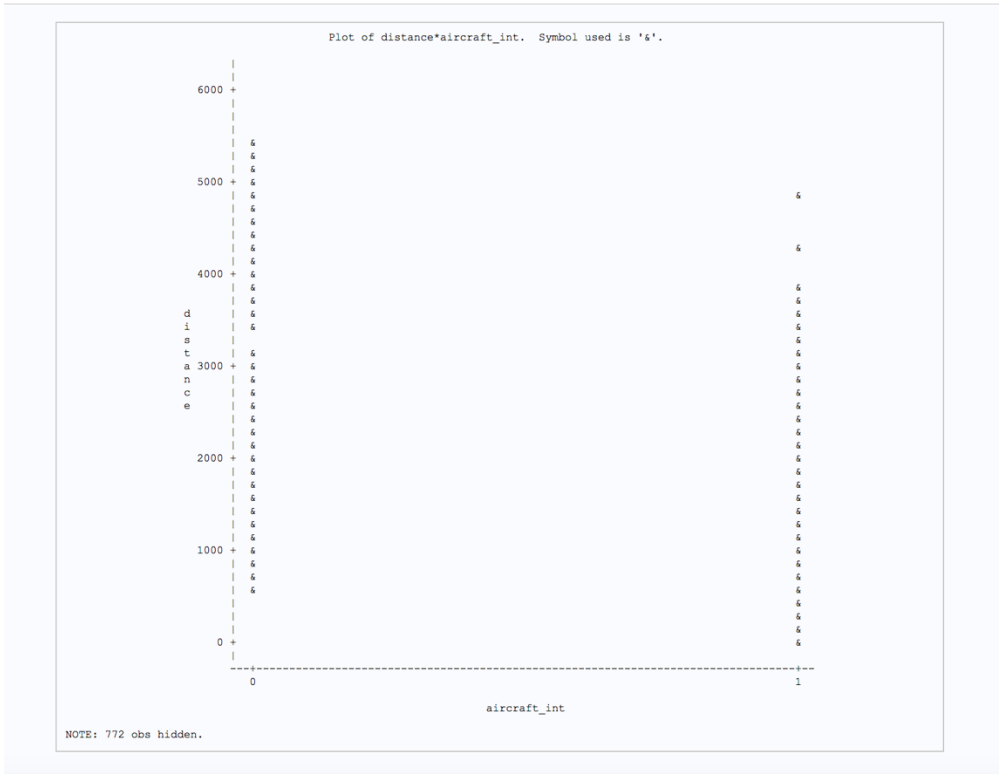
```
/* convert type of aircraft to integer counterparts */
Data normal_converted;
set normal;
if aircraft='airbus' then aircraft_int=1;
else aircraft_int=0;
run;

/* X-Y Plot of dependence */

PROC PLOT data=normal_converted;
plot distance*speed_ground = '$'
distance*speed_air = 'h'
distance*duration = 'w'
distance*height = 'o'
distance*pitch = 'y'
distance*no_pasg = '#'/ overlay  ;
plot distance*aircraft_int = '&';
run;
```

Result:


Plot of distance*speed_ground. Symbol used is '$'.
Plot of distance*speed_air. Symbol used is 'h'.
Plot of distance*duration. Symbol used is 'w'.
Plot of distance*height. Symbol used is 'o'.
Plot of distance*pitch. Symbol used is 'y'.
Plot of distance*no_pasg. Symbol used is '#'.
distance
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
250
300
350
speed_ground
NOTE: 680 obs had missing values. 3238 obs hidden.


Plot of distance*aircraft_int. Symbol used is '&'.
distance
6000
5000
4000
3000
2000
1000
0
0
1
aircraft_int
NOTE: 772 obs hidden.

- Giving the first and most valuable source of evidence, the first chart indicates a very strong dependence of Landing Distance on speed_air and speed_ground. The dependence is linear and positive, that is as speed_ground or air increases there is a good probable chance of Landing distance being overshot.
- Using a similar approach it was observed that other variables do not have any significant inter-dependence on each other.

Conclusion:

- Landing Distance exhibits strongest dependence on speed on ground and in air from the X-Y plots.
- Landing Distance also depends on aircraft. It has been observed from TTEST that the mean of Landing Distance of Boeing aircraft is 32.6% greater than the mean of Airbus aircraft.
- From the X-Y plot there is a small indication of linear increase of distance with height. However, there is no strong evidence for the same.
- Pitch has been observed to be different for different aircrafts. It is greater for Boeing by 10.52%.
- No substantial interrelation has been observed between any other combinations of variables.

FAA PROJECT: CHAPTER 3

DATA REGRESSION & EVALUATION

Goal Statement:

This chapter aims at evaluating the propositions made in the previous chapters, with reliable correlation outputs, regression models, and model cross evaluation for certainty of quality of results.

Implementation and Methodology:

1. Correlation Analysis:

- From the above results, it was noticed that besides speed-distance dependence, no other strong ties were observed.
- Hence to get a more detailed and accurate output, correlation is used.
- This step will clearly indicate the percentage of dependence and direction as well.

Code:

```
/* correlation analysis */
 PROC CORR DATA=normal_converted;
 VAR speed_ground speed_air duration height pitch no_pasg aircraft_int;
 with distance;
 run;

 /* correlation analysis */
 PROC CORR DATA=normal_converted;
 VAR aircraft_int;
 with pitch;
 run;
```

Result:

The CORR Procedure

1 With Variables:	distance
7 Variables:	speed_ground speed_air duration height pitch no_pasg aircraft_int

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	836	1526	898.41542	1275781	41.72231	5382	distance
speed_ground	836	79.59441	18.73271	66541	33.57410	132.78468	speed_ground
speed_air	206	103.45523	9.69265	21312	90.00286	132.91146	speed_air
duration	786	153.93379	49.33604	120992	14.76421	305.62171	duration
height	836	30.51049	9.80491	25507	6.22752	59.94596	height
pitch	836	4.00501	0.52740	3348	2.28448	5.92678	pitch
no_pasg	836	60.04067	7.47920	50194	29.00000	87.00000	no_pasg
aircraft_int	836	0.53349	0.49918	446.00000	0	1.00000	

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

	speed_ground	speed_air	duration	height	pitch	no_pasg	aircraft_int
distance distance	0.86661 <.0001 836	0.94139 <.0001 206	-0.06107 0.0871 786	0.10767 0.0018 836	0.09308 0.0071 836	-0.02115 0.5414 836	-0.24022 <.0001 836

The CORR Procedure

1 With Variables:	pitch
1 Variables:	aircraft_int

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
pitch	836	4.00501	0.52740	3348	2.28448	5.92678	pitch
aircraft_int	836	0.53349	0.49918	446.00000	0	1.00000	

Pearson Correlation Coefficients, N = 836
Prob > |r| under H0: Rho=0

	aircraft_int
pitch pitch	-0.35582 <.0001

- Above results provide the following insights:

    a. The Landing Distance is strongly dependent on speed_air speed_ground aircraft, as can be observed from the p-value of these variables, at a 95% confidence level. The distance and speed is correlated by 90% average. That is, speed is one of the main factors to consider for landing distance.
    b. The distance and aircraft is correlated by 24% average. That is, Boeing can contribute towards landing overrun in comparison to Airbus.
    c. A new finding, the distance and height of the aircraft is correlated by 10.7% positively.
    d. A new finding, the distance and pitch of the aircraft is correlated by 9.3% positively.
    e. The aircraft and pitch is correlated by 35% average. That is, pitch is greater for Boeing in comparison to Airbus.
    f. Using similar approach other variables were validated, however no new correlation was found.

2. Regression Analysis:

- While the previous analysis gives us an understanding of association of variables with each other, This step forms the ultimate layer of analysis in terms of prediction.
- From earlier results we can understand that ‘Distance’ is the object of interest. That Is, Landing Distance should be within the permissible range.
- Using Regression Analysis, we can predict ‘Distance’ from an estimate of the remaining variables.
- It is expected that greater landing distance would be associated with greater speed in air and ground, it would be lower in ‘Airbus’ than in ‘Boeing’, greater with height, and greater with pitch.

Code:

```
/* regression analysis */
PROC REG DATA=normal_converted;
MODEL distance = speed_ground duration height pitch no_pasg aircraft_int;
title Regression analysis of Distance;
run;
```

Result:

Regression analysis of Distance

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	836
Number of Observations Used	786
Number of Observations with Missing Values	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	549715032	91619172	746.82	<.0001
Error	779	95566373	122678		
Corrected Total	785	645281405			

Root MSE	350.25458	R-Square	0.8519
Dependent Mean	1544.88304	Adj R-Sq	0.8508
Coeff Var	22.67192		

- For above Model an r-squre of 0.85 is got. To make the model more efficient, based on correlation results, duration and no_pasg can be removed, since that does not add value.

Code:

```
/* regression analysis */
PROC REG DATA=normal_converted;
MODEL distance = speed_ground height pitch aircraft_int;
title Regression analysis of Distance;
run;
```

Result:

Root MSE	348.00927	R-Square	0.8507
Dependent Mean	1526.05390	Adj R-Sq	0.8500
Coeff Var	22.80452		

- It can be seen that the r-squre value has largely not changed. Hence those two variables can be removed from the model.
- Using a similar approach, the model was further altered for efficiency. The final regression model, with greatest r-square value is as shown below:

Code:

```
/* regression analysis */
 PROC REG DATA=normal_converted;
 MODEL distance = speed_ground aircraft_int height;/* pitch aircraft_int;*/
 title Regression analysis of Distance;
 run;
```

Result:

Regression analysis of Distance

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	836
Number of Observations Used	836

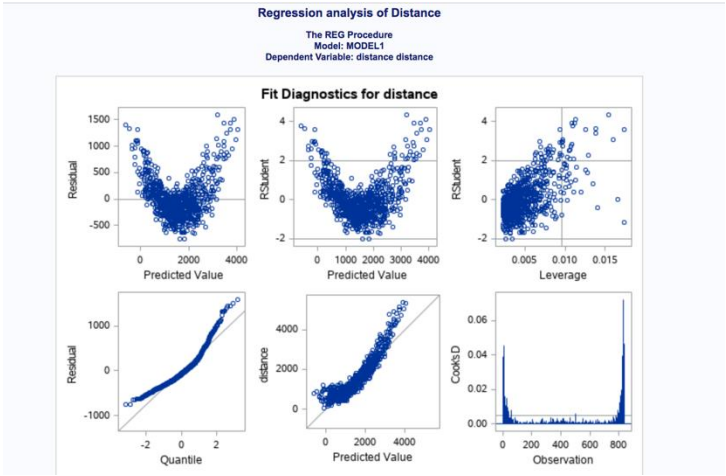
Analysis of Variance

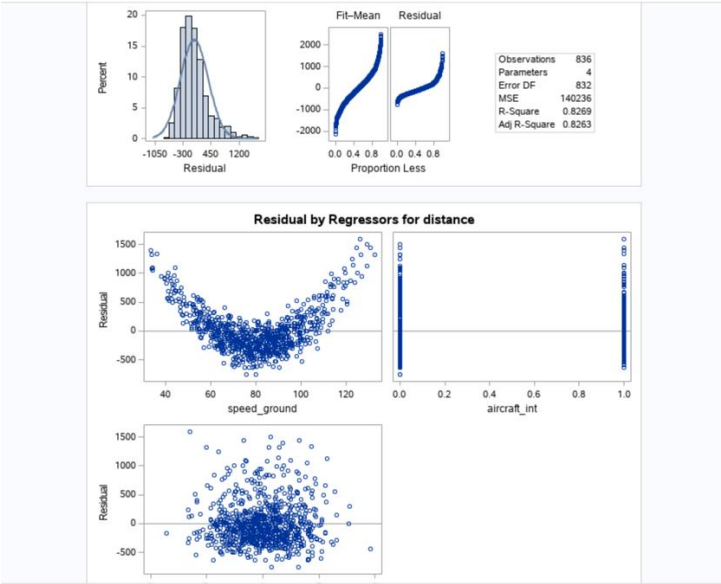
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	572972240	190990747	1573.34	<.0001
Error	832	100998240	121392		
Corrected Total	835	673970479			

Root MSE	348.41371	R-Square	0.8501
Dependent Mean	1526.05390	Adj R-Sq	0.8496
Coeff Var	22.83102		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2021.76515	67.01475	-30.17	<.0001
speed_ground	speed_ground	1	42.45203	0.64497	65.82	<.0001
aircraft_int		1	-497.01201	24.17455	-20.56	<.0001
height	height	1	14.22549	1.23139	11.55	<.0001





- Above results provide the following insights:

  a. Pr > F is 0.0001, hence there is a linear relationship between dependent variable 'distance' and independent variables.
  b. The average expected error in 'distance' prediction is 374. That is, there is a 24% error associated with this model.
  c. The R-squre is 0.85 which indicates good fit of the model.
  d. Speed_ground, height, and aircraft type all have parameters with value less than 0.05, which indicates strong influence on dependent variable.
  e. Hence it can be concluded that, greater speed is related to greater landing distance, greater height is related to greater landing distance by a small parameter, and Boeing is related to greater landing distance than Airbus.

3. Model Diagnostics:

- The fitted regression model from above will now be evaluated.
- The assumption during regression is that noise is noise is independent, normally distributed, mean 0, with constant variance.

Code: Univariate

```
113 /* model checking */
114 PROC REG DATA=normal_converted;
115 MODEL distance = speed_ground aircraft_int height / r;
116 output out=diagnostics r=residual;
117 run;
118
119 proc univariate data=diagnostics;
120 run;
121
```

Result:

The UNIVARIATE Procedure
Variable: residual (Residual)

Moments			
N	836	Sum Weights	836
Mean	0	Sum Observations	0
Std Deviation	347.787256	Variance	120955.976
Skewness	1.57149451	Kurtosis	2.97899271
Uncorrected SS	100998240	Corrected SS	100998240
Coeff Variation	.	Std Error Mean	12.0284738

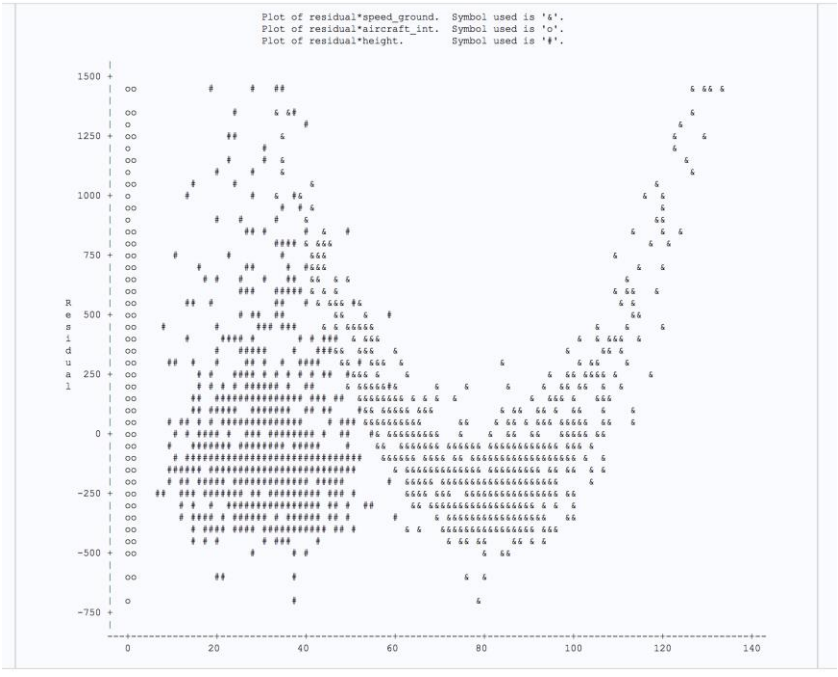
Basic Statistical Measures			
Location		Variability	
Mean	0.0000	Std Deviation	347.78726
Median	-89.1533	Variance	120956
Mode	.	Range	2183
		Interquartile Range	357.52102

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-106	Pr >= |M|	<.0001
Signed Rank	S	-31807	Pr >= |S|	<.0001

Code: Plotting of Residual

```
proc plot data=diagnostics;
plot residual*speed_ground='&'
 residual*aircraft_int="o"
residual*height = '#'/overlay;
title Plot of residual with model indepedent variables;
run;
```

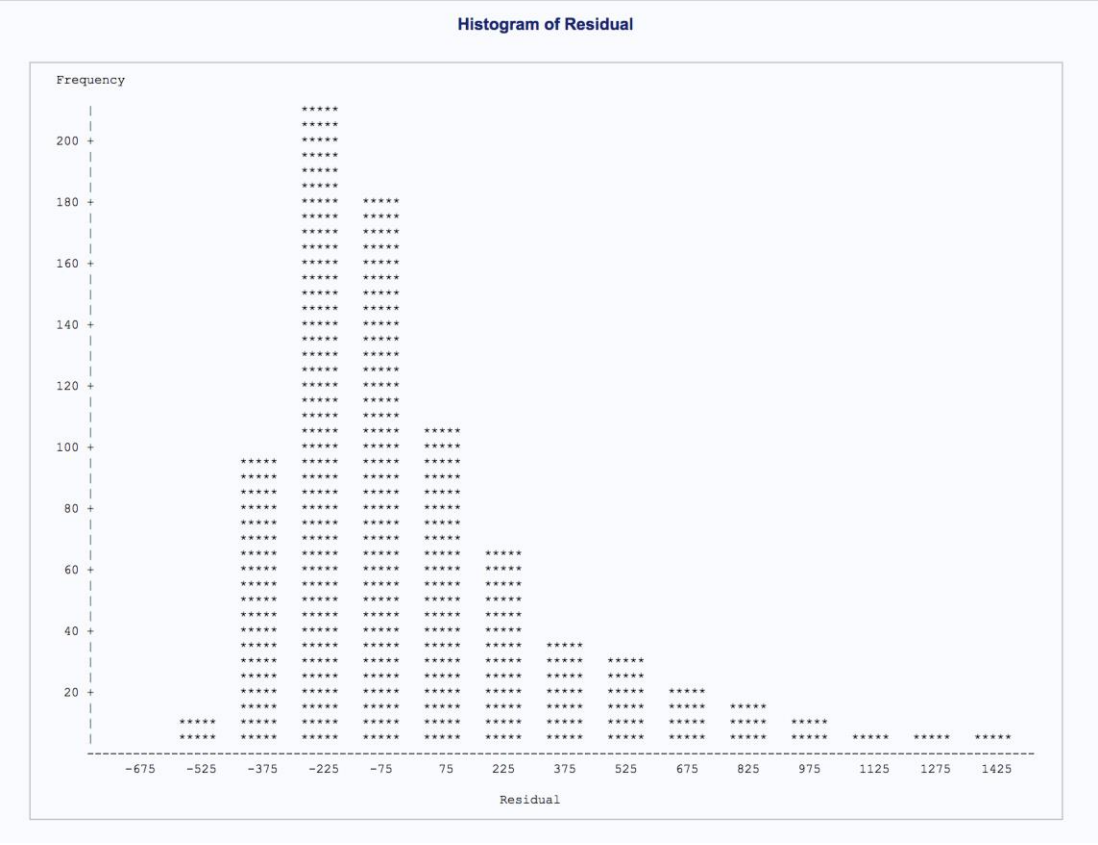
Result:



Code: Histogram of residual

```
126 proc chart data=diagnostics;
127 vbar residual;
128 run;
```

Result:



Code: TTEST for mean of residual

```
proc ttest data=diagnostics;
var residual;
run;
```

Result:

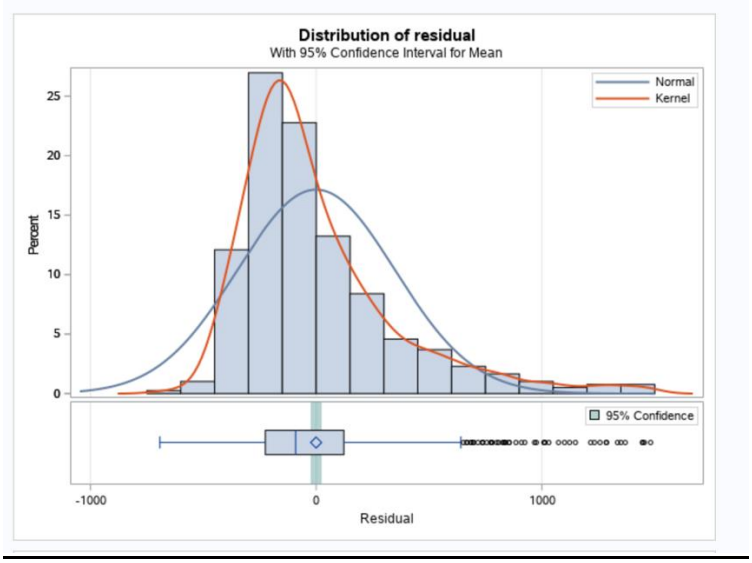
The TTEST Procedure

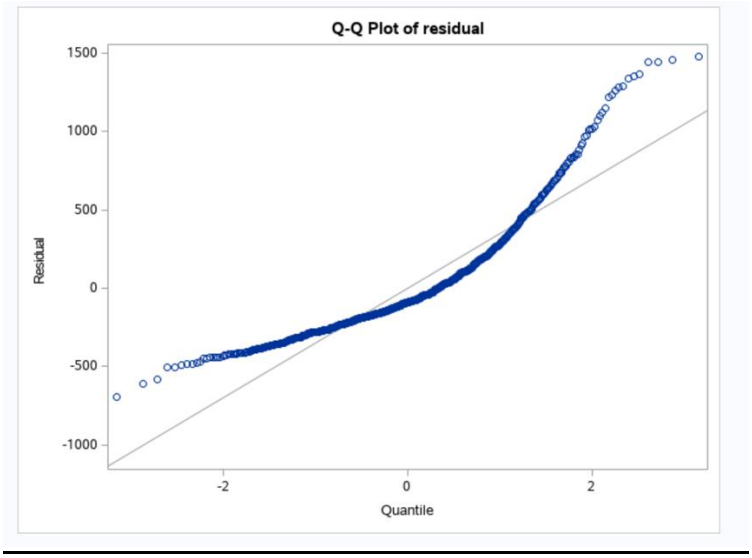
Variable: residual (Residual)

N	Mean	Std Dev	Std Err	Minimum	Maximum
786	-182E-14	348.9	12.4453	-692.5	1482.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-182E-14	-24.4301	24.4301	348.9	332.5	367.1

DF	t Value	Pr > |t|
785	-0.00	1.0000





Conclusion:

- A strong correlation matrix indicated linear dependence of distance on speed_ground, height, pitch. It further showed that distance is greater for Boeing than Airbus.
- Using understanding of correlation matrix, a regression model was fitted. The model provided the following conclusions at a 95% confidence level:
  a. There can be a greater chance of overshooting Landing Distance threshold with increase in speed on ground.
  b. There can be a greater chance of overshooting Landing Distance threshold with Boeing aircrafts than Airbus.
  c. There can be a greater chance of overshooting Landing Distance threshold with greater height.
- The fitted model's residuals were observed, and were found to be giving 0 mean, normal. Distribution, constant variance.

Write your short answers to these questions:

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

836 Observations have been used to fit the model because, 100 were duplicate, 14 were abnormal according to project.pdf.
Hence to receive accurate results and maintain high sample size, all missing values were not deleted.

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

- 86.67% positive linear dependence of distance on speed_ground
- 24% dependence of distance on aircraft (mean distance of Boeing is 24.74% greater than Airbus)
- 10% positive dependence of distance on height
- 9% positive dependence of distance on pitch

However, after regression model, it was noticed that pitch does not contribute significantly to distance.

Hence the final factors are:

a. Speed_ground – largely (as speed increases, distance increases)
b. Aircraft type – moderately (Distance is greater for Boeing than Airbus)
c. Height – slightly (as height increases, distance increases)

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

There is considerable difference in the two makes Boeing and Airbus:

a. The mean of distance for Boeing is 32% greater than mean of distance for Airbus. This indicates that there is greater landing threshold violation threat for Boeing in comparison to Airbus.
b. The mean of pitch for Boeing is 10.5% greater than mean of pitch for Airbus. This indicates that pitch could be an indirect contributor to Landing distance threshold via aircraft type.