FEDERAL AVIATION ADMINISTRATION (FAA) PROJECT

Data Analysis and Findings

Abstract

The Federal Aviation Administration wants to ensure better flight safety across the country, especially during flight landings. 'Landing Distance' plays a major role in it. Therefore it is important that FAA is be able to determine what factors affect the landing distance of a flight so that they can make better predictions to further enhance flight safety. This project analyzed two datasets from FAA using the SAS statistical tool package. The data was cleaned up and made analysis-ready. The clean-up process took into consideration missing values, abnormal observations, and duplicate entries. Following this data visualization and data exploration methods were used to gather basic characteristics about the data. Using a multiple linear regression procedure, the data was modeled to find how and to what degree the predictor variables affect the landing distance. Through a further statistical significance check the model was further streamlined to make the predictions more accurate. The study found that generally key variables like air speed and height affect the landing distance more than other variables. Moreover the study also found that the landing distance is affected in different ways for the two types of aircraft – Airbus and Boeing.

Akash Dash

dashah@mail.uc.edu

M12887489

Contents

Chapter 1 – Data Exploration, Cleaning and Preparation

1.1 Specific Goal for data exploration and cleaning on the FAA dataset:

To make the dataset ready for further modeling. From the first look on the datasets, here are some observations

- Each of the datasets seem to have values missing in certain variables.
- Moreover FAA1 contains a new variable 'duration' which is additional information that is missing in FAA2.
- There appear to be some redundant/duplicate observations among the two datasets

Activities we performed are -

- Imported the datasets into SAS
- Ran an initial validity and completeness check on each of the datasets
- Concatenated the datasets
- Ran before-clean-up summary statistics on the data
- Created flag for abnormal values
- Deleted abnormal values to further clean the data ; preserve observations and variables with missing values
- Used NODUP KEY to remove the duplicate observations.
- Generated summary statistics on the new cleaned-up dataset.

1.2 SAS Code

```
PROC IMPORT OUT=FAA1 DATAFILE= "/folders/myfolders/GASUE34_data/FAA1.xls"
            DBMS=XLS REPLACE;
     SHEET="FAA1";
     GETNAMES=YES;
RUN;

proc print data=FAA1;
run;

PROC IMPORT OUT=FAA2 DATAFILE= "/folders/myfolders/GASUE34_data/FAA2.xls"
            DBMS=XLS REPLACE;
     SHEET="FAA2";
     GETNAMES=YES;
RUN;

proc print data=FAA2;
run;

proc univariate data=FAA1;
 var Duration No_pasg Speed_ground Speed_air Height Pitch Distance;
 title name = 'FAA1_univ';
run;

proc univariate data=FAA2;
 var No_pasg Speed_ground Speed_air Height Pitch Distance;
```

```
 title name = 'FAA2_univ';
run;

/*Concatenate the datasets*/
data FAA;
       set FAA1 FAA2;
       title name = 'FAA';
       run;

proc print data=FAA;
run;

proc univariate data=FAA;
 var Duration No_pasg Speed_ground Speed_air Height Pitch Distance;
run;
data FAA_clean1;
       set FAA;
       title name = 'FAA_clean1';
       if duration < 40 and duration ^= . then abn = 1;
       else if speed_ground < 30 and speed_ground ^= . then abn = 1;
       else if speed_ground > 140 and speed_ground ^= . then abn = 1;
       else if speed_air < 30 and speed_air ^= . then abn = 1;
       else if speed_air > 140 and speed_air ^= . then abn = 1;
       else if height < 6 and height ^= . then abn = 1;
       else if distance > 6000 and distance ^= . then abn = 1;
       else abn = 0;
run;
proc print data=FAA_clean1;
run;

data FAA_clean2;
       set FAA_clean1;
       title name = 'FAA_clean2';
       if abn = 1 then delete;
       run;
proc print data=FAA_clean2;
run;
/*Let us do a validity check and completeness check on this dataset*/
proc univariate data=FAA_clean2;
var Duration No_pasg Speed_ground Speed_air Height Pitch Distance;
run;
```

1.3 SAS Output (Only relevant output shown here)

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

Moments (Ground Speed)			
N	832	**Sum Weights**	832
Mean	79.5235023	**Sum Observations**	66163.5539
Std Deviation	18.7325852	**Variance**	350.909747
Skewness	0.09117377	**Kurtosis**	-0.2329996
Uncorrected SS	5553163.53	**Corrected SS**	291606
Coeff Variation	23.5560364	**Std Error Mean**	0.64943554

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

Moments (Height)			
N	832	Sum Weights	832
Mean	30.4554041	Sum Observations	25338.8962
Std Deviation	9.77918085	Variance	95.632378
Skewness	0.12795758	Kurtosis	-0.33081
Uncorrected SS	851176.831	Corrected SS	79470.5061
Coeff Variation	32.1098377	Std Error Mean	0.3390321

Moments(Pitch)			
N	832	Sum Weights	832
Mean	4.00507998	Sum Observations	3332.22654
Std Deviation	0.52625729	Variance	0.27694674
Skewness	0.01777465	Kurtosis	-0.08739
Uncorrected SS	13575.9765	Corrected SS	230.142741
Coeff Variation	13.139745	Std Error Mean	0.01824469

Moments(Landing Distance)			
N	832	Sum Weights	832
Mean	1521.89391	Sum Observations	1266215.73
Std Deviation	895.95975	Variance	802743.873
Skewness	1.47826357	Kurtosis	2.5548043
Uncorrected SS	2594126168	Corrected SS	667080159
Coeff Variation	58.8713671	Std Error Mean	31.0618156

1.4 Observations and Conclusions

Exploration on the data helped us gather the following information –

- We observe that after clean-up, the number of observations have reduced from 950 to 833.
- Mean and median are very close for almost all variables, which means skew is not high
- There were missing values for many of the variables in each dataset
- From the min and max, we know there are abnormal values present in variables which were removed
- We also looked at summary (descriptive) statistics and normality of all the variable distributions.

<u>Summary stats before clean-up Observations = 950)</u>

Variables	Mean	Median	SD	Min	Max	Percent of abnormal	Percent of missing
Duration	154	153.9	49.25	14.76	305.62		20
No_pasg	60.16	60	7.49	29	87		5
Speed_ground	79.28	79.41	19.33	27.73	141.219		5
Speed_air	103.73	100.89	10.60	90.00	141.72		76
Height	30.13	29.90	10.35	-3.54	59.94		5

Pitch	4.019	4.015	0.526	2.28	5.92		5
Distance	1548.8	1267.4	948.68	34.08	6533.04		5
Total						2.4%	

Summary stats after clean-up (Observations = 833)

Variables	Mean	Median	SD	Min	Max	Pct of abnormal	Pct of missing
Duration	154.77	154.28	48.34	41.94	305.62		6.24
No_pasg	60.12	60	7.5	29	87		0.12
Speed_ground	79.4	79.6	18.8	33.5	132.7		0.12
Speed_air	103.3	100.8	9.88	90	132.9		75.63
Height	30.5	30.16	9.83	6.2	59.9		0.12
Pitch	4.01	4.00	0.52	2.28	5.9		0.12
Distance	1543.5	1281.2	906.9	41.7	5381.9		0.12
Total						0	

Chapter 2 – Data Visualization

2.1 Specific Goal for data visualization on the FAA dataset:

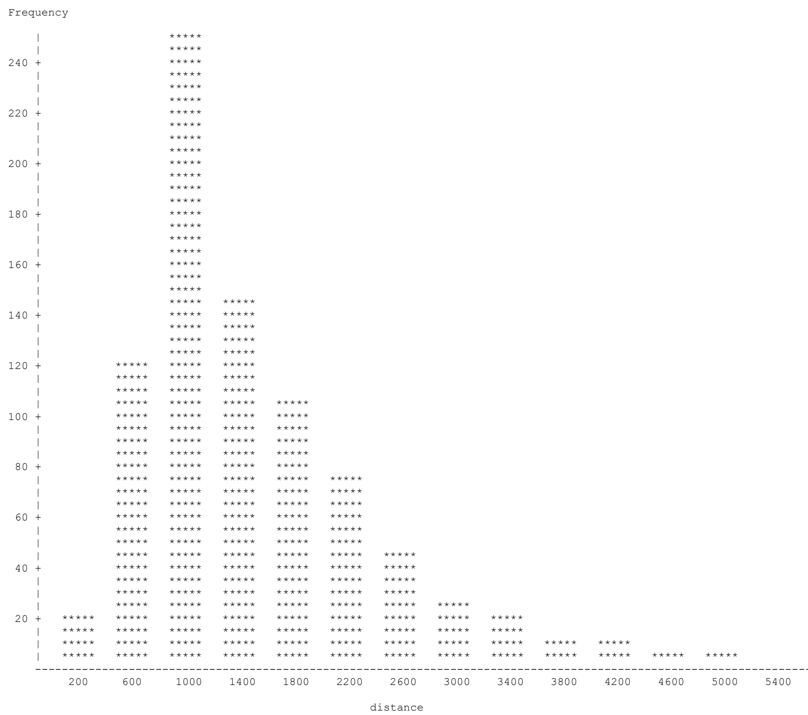
- To see through FREQ plots / histograms the distributions of all variables, whether they are normally distributed and whether they show any other trends.
- To see through X-Y plots how the response variable is distributed against each of the predictors. Also we want to see how each of the predictor variables look against the other in X-Y plots.
- To get an indicative idea of any correlations / associations existing between variables.

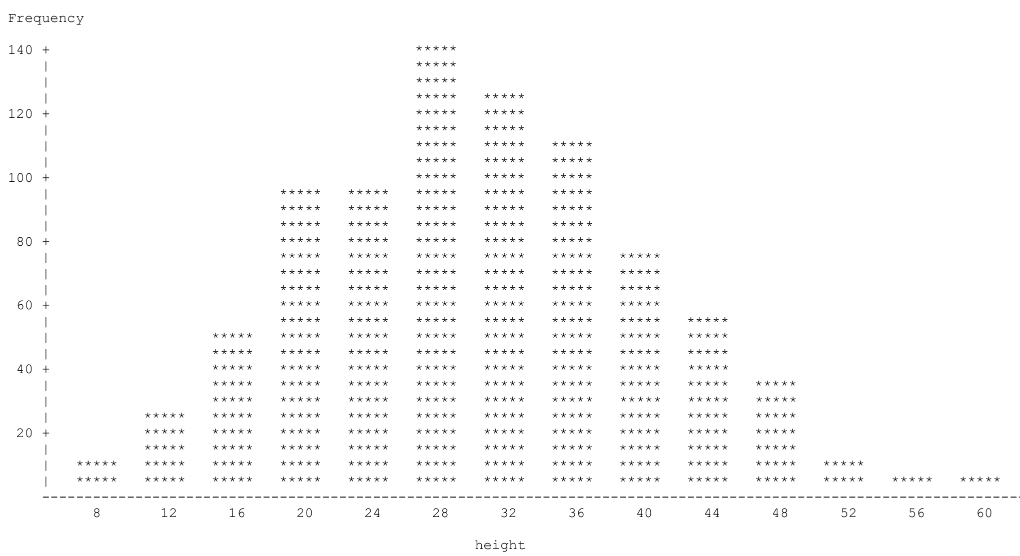
2.2 SAS Code

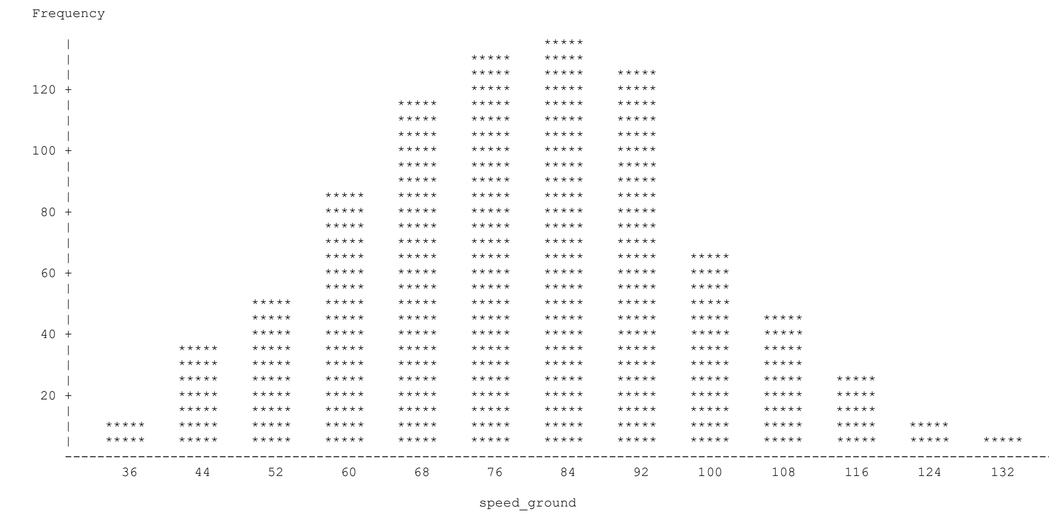
```
/*CHAPTER - 2 - DATA VISUALIZATIONS on FAA Dataset*/
proc chart data=FAA_clean3;
vbar Duration No_pasg Speed_ground Speed_air Height Pitch Distance;
run;

/*X-Y plots between the response variable and every predictor variable*/
proc plot data=FAA_clean3;
	plot Distance * Duration;
	plot Distance * No_pasg;
	plot Distance * Speed_ground;
	plot Distance * Speed_air;
	plot Distance * Height;
	plot Distance * Pitch;
	plot Distance * Aircraft;
run;

/*X-Y plots among predictor variables...selected few */
```

```
proc plot data=FAA_clean3;
       plot No_pasg * Duration;
       plot Height * Speed_ground;
       plot Speed_ground * Speed_air;
       plot Pitch * Height;
       plot Speed_ground * Pitch;
run;
```

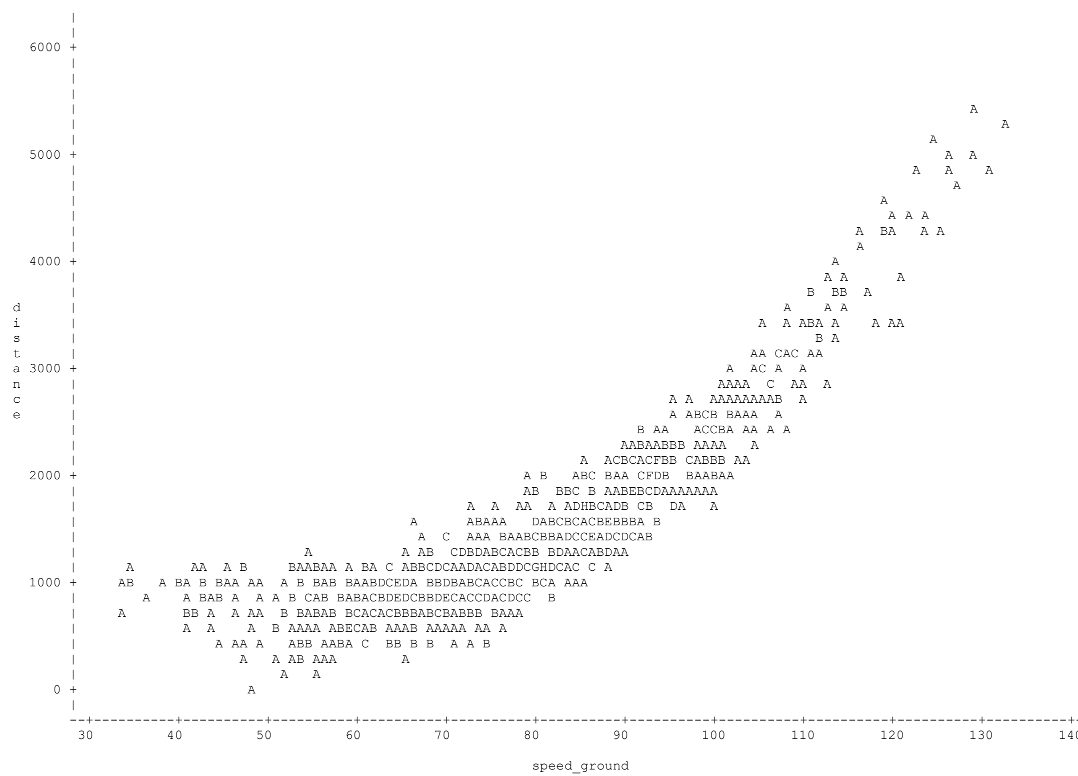
2.3 SAS Output (only few selected ones shown here)





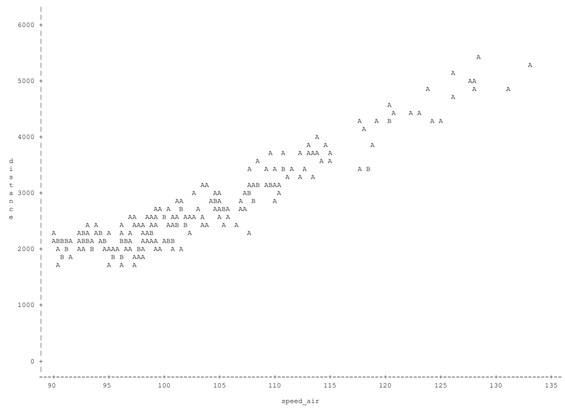


Plot of distance*speed_ground. Legend: A = 1 obs, B = 2 obs, etc.



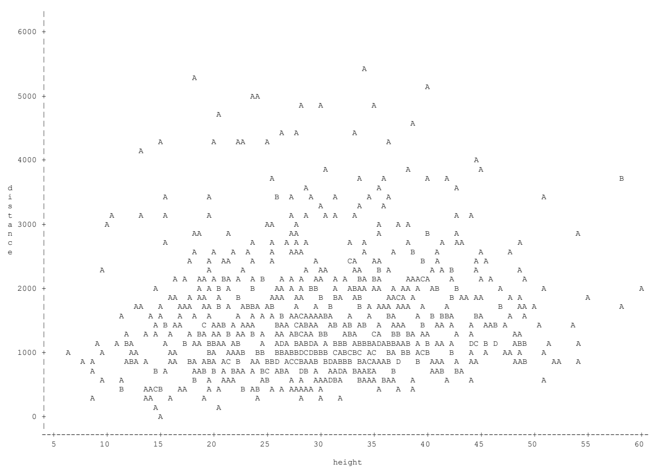
NOTE: 1 obs had missing values.

Plot of distance*speed_air. Legend: A = 1 obs, B = 2 obs, etc.



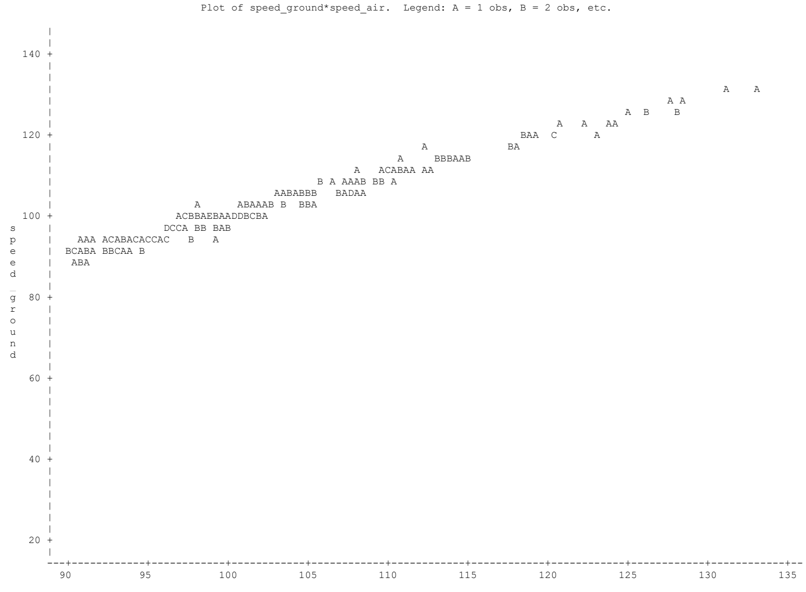
NOTE: 630 obs had missing values.

Plot of distance*height. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 1 obs had missing values.

Plot of speed_ground*speed_air. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 630 obs had missing values.

2.4 Observations and Conclusions

The FREQ plots/histograms and the X-Y plots gave us the following insights about the dataset and variables:-

(a) Through FREQ plots of each individual variable we wanted to see the distributions
(b) Almost all variables still have a normal distribution with some having little amount of skew. Which could be because there are many missing values and also because the data was collected from two different sources/datasets.
(c) Through X-Y plots we were able to get an indicative idea on correlations / associations.
(d) There is strong correlation/association between Landing Distance and Ground and Air speeds.
(e) There are no specific positive or negative patterns in the other X-Y plots but there are definitely contributions to response variable from other predictor variables as well.
(f) Some correlations exist between predictors but the strongest positive linear relationship is between Speed_ground and Speed_air

Chapter 3 – Modeling

3.1 Specific Goal:

Our goals for modeling were –

(a) To develop a statistically significant model (using multiple linear regression) which can help us accurately predict how Landing Distance is affected with changes in Predictor variables.
(b) To gain an idea into the correlations between Response and Predictors as well as any significant correlations between predictors which can hamper the accuracy of the model.

3.2 SAS Code:

```
/*Modeling*/
/*Actual Correlation Calculation between variables*/
proc corr data=FAA_clean3;
 var Distance Duration No_pasg Speed_ground Speed_air Height Pitch;
 title Correlation Coefficients;
run;


 /*Regression Model*/

proc reg data=FAA_clean3;

 model Distance=Duration No_pasg Speed_ground Speed_air Height Pitch / r;
 title Regression Model;
 run;

proc glm data=FAA_clean3;
class Aircraft;
model Distance= Aircraft Duration No_pasg Speed_ground Speed_air Height Pitch / solution estimates;
run;
```

3.3 SAS Output:

Correlation Procedure

7 Variables:	distance	duration	no_pasg	speed_ground	speed_air	height	pitch

Simple Statistics							
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
distance	832	1522	895.95975	1266216	41.72231	5382	distance
duration	781	154.77572	48.34992	120880	41.94937	305.62171	duration
no_pasg	832	60.05889	7.48750	49969	29.00000	87.00000	no_pasg
speed_ground	832	79.52350	18.73259	66164	33.57410	132.78468	speed_ground
speed_air	203	103.48504	9.73628	21007	90.00286	132.91146	speed_air
height	832	30.45540	9.77918	25339	6.22752	59.94596	height
pitch	832	4.00508	0.52626	3332	2.28448	5.92678	pitch

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 832	-0.05138 0.1514 781	-0.01801 0.6039 832	0.86627 <.0001 832	0.94210 <.0001 203	0.09953 0.0041 832	0.08710 0.0120 832
duration duration	-0.05138 0.1514 781	1.00000 781	-0.03639 0.3098 781	-0.04897 0.1716 781	0.04454 0.5364 195	0.01112 0.7564 781	-0.04675 0.1918 781
no_pasg no_pasg	-0.01801 0.6039 832	-0.03639 0.3098 781	1.00000 832	-0.00054 0.9877 832	-0.00616 0.9305 203	0.04688 0.1767 832	-0.01799 0.6043 832
speed_ground speed_ground	0.86627 <.0001 832	-0.04897 0.1716 781	-0.00054 0.9877 832	1.00000 832	0.98794 <.0001 203	-0.05737 0.0982 832	-0.03898 0.2615 832
speed_air speed_air	0.94210 <.0001 203	0.04454 0.5364 195	-0.00616 0.9305 203	0.98794 <.0001 203	1.00000 203	-0.07933 0.2606 203	-0.03927 0.5780 203
height height	0.09953 0.0041 832	0.01112 0.7564 781	0.04688 0.1767 832	-0.05737 0.0982 832	-0.07933 0.2606 203	1.00000 832	0.02301 0.5074 832
pitch pitch	0.08710 0.0120 832	-0.04675 0.1918 781	-0.01799 0.6043 832	-0.03898 0.2615 832	-0.03927 0.5780 203	0.02301 0.5074 832	1.00000 832

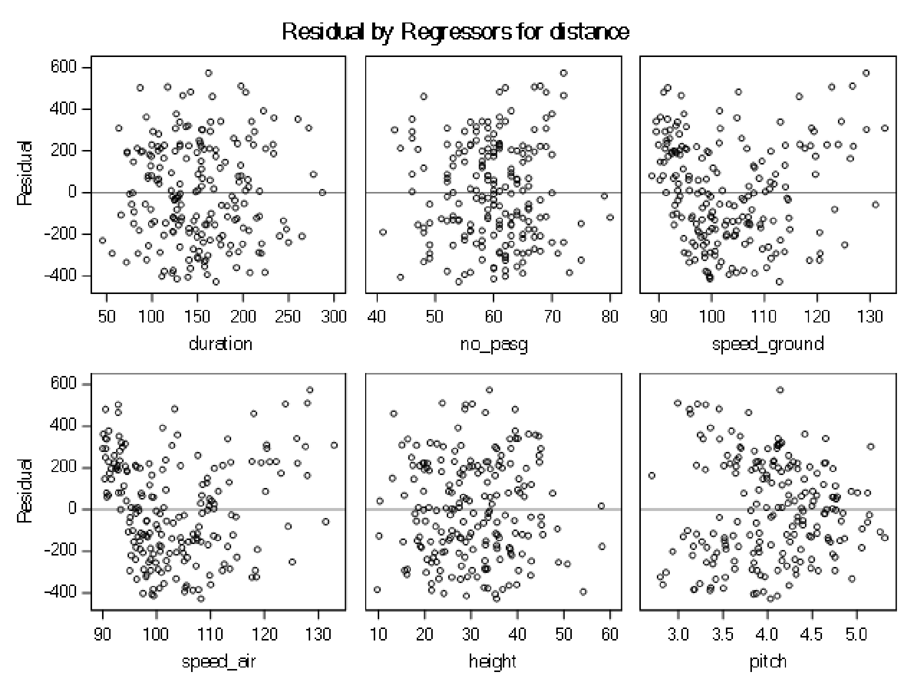
Regression Procedure

Number of Observations Read	833
Number of Observations Used	195
Number of Observations with Missing Values	638

Root MSE	242.43886	**R-Square**	0.9173
Dependent Mean	2784.49158	**Adj R-Sq**	0.9147
Coeff Var	8.70676		

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	122625808	20437635	347.72	<.0001
Error	188	11050001	58777		
Corrected Total	194	133675809			

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-6249.84344	291.58960	-21.43	<.0001
duration	duration	1	0.02246	0.36763	0.06	0.9514
no_pasg	no_pasg	1	-3.34026	2.48183	-1.35	0.1800
speed_ground	speed_ground	1	-2.27284	11.56846	-0.20	0.8445
speed_air	speed_air	1	82.90693	11.75796	7.05	<.0001
height	height	1	12.65927	1.87055	6.77	<.0001
pitch	pitch	1	123.73575	31.32815	3.95	0.0001



GLM Procedure

Class Level Information		
Class	Levels	Values
aircraft	2	airbus boeing

Source	DF	Type I SS	Mean Square	F Value	Pr > F
aircraft	1	3978580.0	3978580.0	220.08	<.0001
duration	1	493579.7	493579.7	27.30	<.0001
no_pasg	1	46433.3	46433.3	2.57	0.1107
speed_ground	1	119246400.9	119246400.9	6596.32	<.0001
speed_air	1	3377980.8	3377980.8	186.86	<.0001
height	1	3142800.4	3142800.4	173.85	<.0001
pitch	1	9500.9	9500.9	0.53	0.4694

Source	DF	Type III SS	Mean Square	F Value	Pr > F
aircraft	1	7669468.447	7669468.447	424.25	<.0001
duration	1	7080.086	7080.086	0.39	0.5322
no_pasg	1	37368.555	37368.555	2.07	0.1522
speed_ground	1	5523.066	5523.066	0.31	0.5811
speed_air	1	3110148.129	3110148.129	172.04	<.0001
height	1	3134551.016	3134551.016	173.39	<.0001
pitch	1	9500.871	9500.871	0.53	0.4694

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-5791.657272	B	163.2345690	-35.48	<.0001
aircraft airbus	-437.942766	B	21.2621163	-20.60	<.0001
aircraft boeing	0.000000	B	.	.	.
duration	0.127632		0.2039443	0.63	0.5322
no_pasg	-1.981166		1.3779690	-1.44	0.1522
speed_ground	-3.546367		6.4160142	-0.55	0.5811
speed_air	85.546892		6.5220701	13.12	<.0001
height	13.675598		1.0385572	13.17	<.0001
pitch	-13.489744		18.6077460	-0.72	0.4694

3.4 Observations and Conclusions:

Our observations from the modeling exercise are as follows –

(a) From the Pearson correlation matrix we observe the following –

- The statistically significant(based on p-value of 0.0001) and strong +ve correlations are the following:
  - Landing distance * Ground speed - 86%
  - Landing distance * Air speed - 94%
  - Ground speed * Air speed - 98%

(b) From the Regression output we observe the following -

- SAS used only 195 observations out of 833 to build the model.
- Majority of observations were left out because of one or more missing values. So our guess is that statistically the model would have been more robust had SAS used more observations in model building
- The model is Distance = -6249.84 + 0.022*Duration - 3.34*No_pasg - 2.27*Speed_ground + 82.9*Speed_air + 12.65*Height + 123.73*Pitch

Chapter 4 – Model Checking

4.1 Specific Goal:

(a) Check whether residuals:

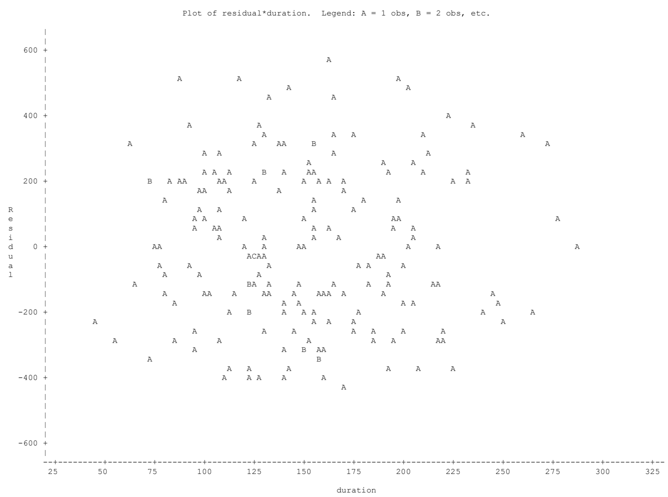
- Are normally distributed (test hypothesis)
- Have mean around zero (test hypothesis)
- Have constant variance (check plots)

(b) After (a) is satisfied, go on to interpret our model

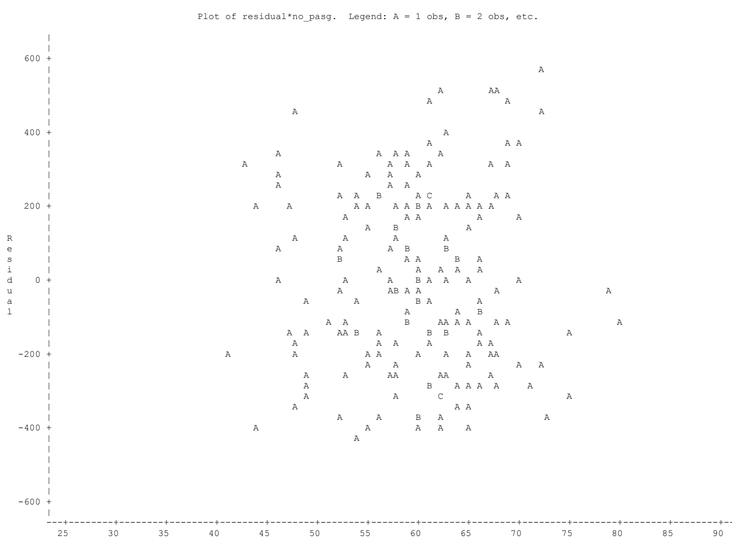
4.2 SAS Code:

```
/*Model Checking*/

proc reg data=FAA_clean3;
 model Distance=Duration No_pasg Speed_ground Speed_air Height Pitch / r;
 output out=diagnostics r=residual;
 run;

  /*Checking Residual Plots*/
proc plot data=diagnostics;
 plot Residual*Duration;
 plot Residual*No_pasg;
 plot Residual*Speed_ground;
 plot Residual*Speed_air;
 plot Residual*Height;
 plot Residual*Pitch;
run;

proc means data=diagnostics t prt;
var Residual;
run;

  /*Checking Normal distribution for Residuals*/
proc chart data=diagnostics;
 vbar Residual;
run;
proc univariate data=diagnostics;
var Residual;
histogram;
run;
```

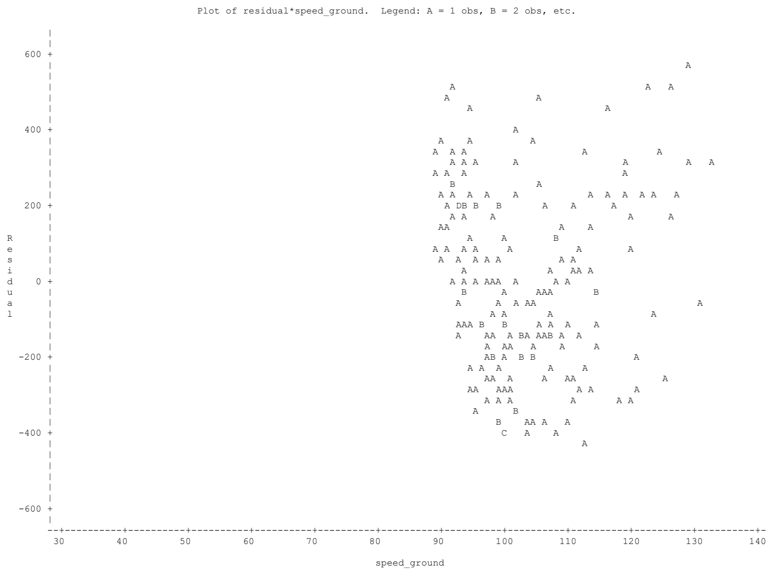
4.3 SAS Output: (Only selected few have been shown so as not to make the chapter bulky)



NOTE: 638 obs had missing values.



NOTE: 638 obs had missing values.



NOTE: 638 obs had missing values.

The Means Procedure

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

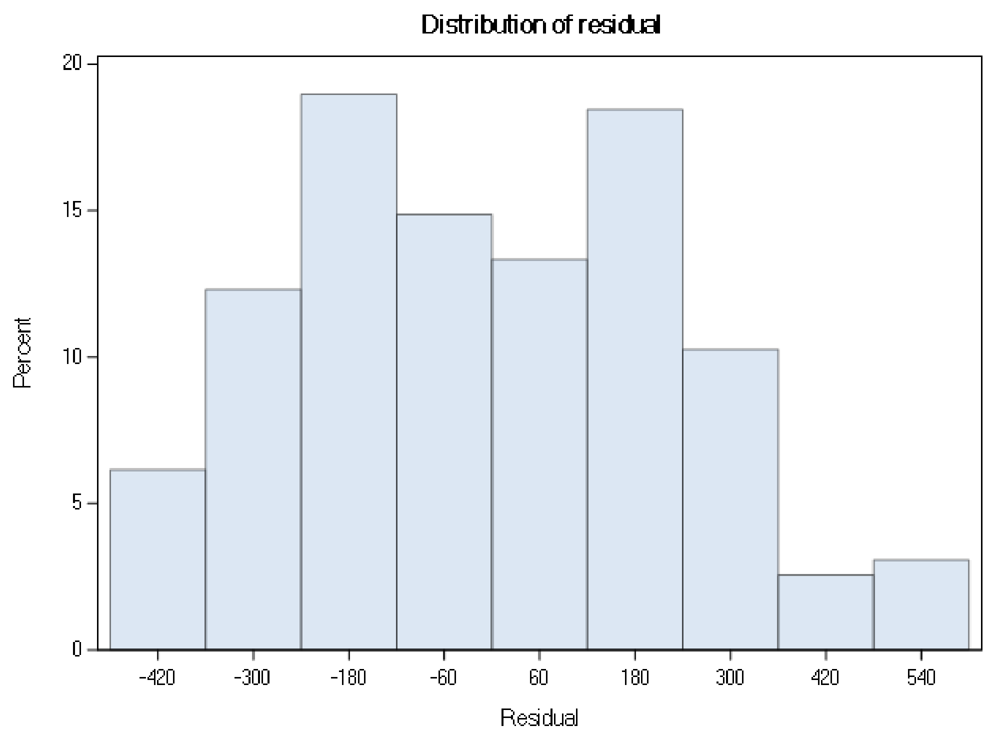
The Univariate Procedure (Residuals)

Moments			
N	195	**Sum Weights**	195
Mean	0	**Sum Observations**	0
Std Deviation	238.660362	**Variance**	56958.7686
Skewness	0.19374508	**Kurtosis**	-0.8658332
Uncorrected SS	11050001.1	**Corrected SS**	11050001.1
Coeff Variation	.	**Std Error Mean**	17.0908235

Basic Statistical Measures			
Location		**Variability**	
Mean	0.0000	**Std Deviation**	238.66036
Median	-16.8380	**Variance**	56959
Mode	.	**Range**	1002
		Interquartile Range	390.88347

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	0	**Pr > |t|**	1.0000
Sign	**M**	-4.5	**Pr >= |M|**	0.5668
Signed Rank	**S**	-131	**Pr >= |S|**	0.8686

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	573.714
99%	511.518
95%	377.591
90%	310.277
75% Q3	200.938
50% Median	-16.838
25% Q1	-189.946
10%	-311.520
5%	-373.791
1%	-414.050
0% Min	-427.885



4.3 Observations and Conclusions:

(a) The X-Y plot of residuals against each predictor variable indicate that –

- Residuals are uniformly distributed on both sides of the mean(0).
- Variance is constant.

(b) The one sample t-test shows that Ho is true, which means that the mean of residuals distribution is zero.

(c) Checking the histogram of the residuals, we see it appears bi-modal but normal. This could be because initial data collection was from two different sources.

(d) From SAS output we see that

- p-value for t-test is very high, which means we can't reject the null hypothesis here (Ho: Residuals are normally distributed).
- Therefore, now we can go ahead and interpret the model.

Model interpretation

- The p-value for the F-test is significant. So we conclude the model as a whole is statistically significant
- Checking the p-values for each predictor variable we see that the ones statistically significant are - Air speed and Height
- R-squared is 91.73, which means the model explains 91.73% of variability in Landing Distance.
- So, we now go ahead and remodel. This time keeping out the predictors which are not statistically significant.

Chapter 5 – Remodeling and Model Checking

5.1 Specific Goal:

Our goals for remodeling were –

(a) To verify how our model looks like after keeping out the variables not statistically significant, so that we can more accurately predict how Landing Distance is affected with changes in Predictor variables.
(b) To gain an idea into the correlations between Response and Predictors as well as any significant correlations between predictors which can hamper the accuracy of the model.
(c) To also check whether there is any difference in how predictors affect response for the two different classes of aircrafts – Airbus and Boeing.

5.2 SAS Code:

```
/*Remodeling*/
proc reg data=FAA_clean3;
 model Distance= Speed_air Height / r;
 output out=diagnostics_new r=residual;
 run;

/*Model Rechecking*/
  /*Checking Residual Plots*/
proc plot data=diagnostics_new;
  plot Residual*Speed_air;
 plot Residual*Height;
run;

proc means data=diagnostics_new t prt;
var Residual;
run;

  /*Checking Normal distribution for Residuals*/
proc chart data=diagnostics_new;
 vbar Residual;
run;

proc univariate data=diagnostics_new;
var Residual;
histogram;
run;
```

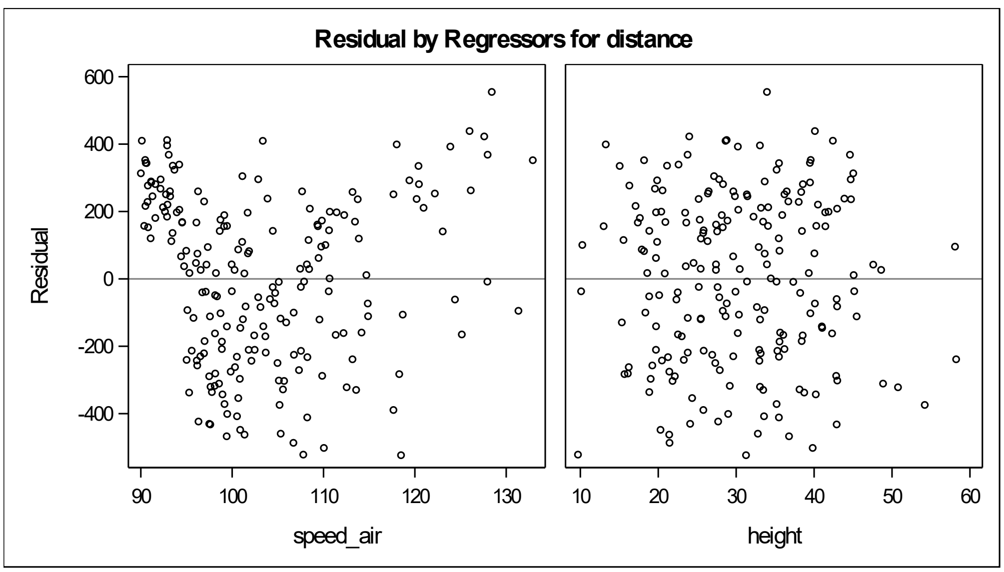
5.3 SAS Output:

Number of Observations Read	833
Number of Observations Used	203
Number of Observations with Missing Values	630

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	123925115	61962557	988.01	<.0001
Error	200	12542854	62714		
Corrected Total	202	136467968			

Root MSE	250.42817	**R-Square**	0.9081
Dependent Mean	2774.67289	**Adj R-Sq**	0.9072
Coeff Var	9.02550		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-5935.72777	201.33934	-29.48	<.0001
speed_air	speed_air	1	80.49493	1.81545	44.34	<.0001
height	height	1	12.54522	1.87638	6.69	<.0001



Means of the residuals

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

The Univariate Procedure and Hypothesis Testing for mean = 0

Moments			
N	203	Sum Weights	203
Mean	0	Sum Observations	0
Std Deviation	249.185343	Variance	62093.3351
Skewness	-0.1752555	Kurtosis	-0.9602506
Uncorrected SS	12542853.7	Corrected SS	12542853.7
Coeff Variation	.	Std Error Mean	17.4893824

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	249.18534
Median	17.39980	Variance	62093
Mode	.	Range	1078
		Interquartile Range	418.24460

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	4.5	Pr >= |M|	0.5746
Signed Rank	S	122	Pr >= |S|	0.8847

5.4 Interpretations and Conclusions:

Model Checking

(a) The X-Y plots of residuals against each predictor variable show that -
- Residuals are almost equally distributed on both sides of the mean(0).
- Variance appears to be constant.

(b) The one sample t-test shows that Ho is true that is the mean of residuals distribution is zero

(c) Checking the histogram of the residuals, we see it appears slightly bi-modal. This could be because initial data collection was from two different sources

(d) From SAS output we see that

- p-value for t-test is very high, which means we can't reject the null hypothesis here (Ho: Residuals are normally distributed).
- Now we can go ahead and interpret the model.

Final Model and Interpretation

(a) The p-value for the F-test is significant. So we conclude the model as a whole is statistically significant
(b) Checking the p-values for predictor variables Height and Air Speed we see that they are statistically significant.
(c) R-squared is 90.81, which means the model explains 90.81% of variability in Landing Distance. R-squared has come down a little bit as we removed insignificant variables, which means we lost information. But this model is a lot simpler too.
(d) Though the intercept is also statistically significant it doesn't have much relevance because the values of Air speed and Height as zeroes are out of dataset range.
(e) So our final model is

Landing Distance = -5935.72 + 80.5 * Air speed + 12.5 * Height

- 1 unit of increase in Air speed will result in 80.5 units of increase in Landing Distance, keeping all other variables constant.
- 1 unit of increase in Height will result in 12.5 units of increase in Landing Distance, keeping all other variables constant.

Comparison between aircrafts

- ✓ For Airbus planes, statistically significant variables to predict Landing Distance tend to be - Air speed, Ground Speed, and Height. (From PROC GLM and Class procedure)
- ✓ For Boeing planes, Landing distance predictors are Air speed and Height ((From PROC GLM and Class procedure)
- ✓ Sample estimate of the predictor variable 'height' is more in case of Boeing than Airbus, which means a unit change in height affects the landing distance more for Boeing than Airbus planes.
- ✓ Sample estimate for the predictor 'air speed' is more in case of Airbus than Boeing, which means a unit change in Air speed affects landing distance more for Airbus than Boeing.

Project Questions and Answers

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

We used 833 observations in total for modeling the data. Though the two datasets had 950 flights observations in total, upon exploration we found that there were observations with both abnormal values and duplicate entries. A basic data clean-up was done to make the combined dataset ready for analysis. In this process the final observations came down to 833. A point to be noted here is that observations with missing values have been preserved in the process to avoid loss of valuable information.

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

Upon running our regression model, we found that the two factors – Air speed and Height – of an aircraft when it enters the runway, are major contributors to landing distance. The model and interpretations of the effect are as follows.

Landing Distance = -5935.72 + 80.5 * Air speed + 12.5 * Height

- 1 unit of increase in Air speed will result in 80.5 units of increase in Landing Distance, keeping all other variables constant.
- 1 unit of increase in Height will result in 12.5 units of increase in Landing Distance, keeping all other variables constant.

Though other factors such as Pitch, Duration and Number of passengers also might affect the Landing distance, our model removed them based on a statistical significance test at a significance level of 0.0001.

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

Yes, there is a difference in factors and how they affect landing distance for the two different types of aircrafts. The difference is described below.

- ✓ For Airbus planes, statistically significant variables to predict Landing Distance tend to be - Air speed, Ground Speed, and Height. (From PROC GLM and Class procedure)
- ✓ For Boeing planes, Landing distance predictors are Air speed and Height ((From PROC GLM and Class procedure)
- ✓ Sample estimate of the predictor variable ‘height’ is more in case of Boeing than Airbus, which means a unit change in height affects the landing distance more for Boeing than Airbus planes.
- ✓ Sample estimate for the predictor ‘air speed’ is more in case of Airbus than Boeing, which means a unit change in Air speed affects landing distance more for Airbus than Boeing.