2017

PREDICTIVE MODELLING OF FLIGHT LANDING DISTANCE USING SAS

FINAL PROJECT

NITISH GHOSAL

M12522228 | BANA6043-001

Summary

The objective of this project is to reduce the risk of landing overrun by identifying the different factors and how they impact the landing distance of commercial flights. Two raw data sets with over 800 and 150 observations respectively were studied for the purpose of this project. After merging the two datasets, removing duplicate and missing values, performing completeness and validity check and removing abnormal values we are left with a total of 831 observations (flight) that were used to fit the final linear regression model. We used exploratory data analysis and data visualization techniques to identify the relationships between our dependent variable and independent variables. After this we validated our findings by building a linear regression model to identify a linear equation that best fits our data. The study found that factors such as square of ground speed of an aircraft when passing over the threshold of the runway, height of an aircraft when it is passing over the threshold runway and make of aircraft have a direct impact on landing distance of flight. The relationship can be expressed by the equation: -

Distance = -945.44 + 0.27*(speed_ground1) + 14.05*(height) + 460.52*(aircraft_make)

*speed_ground1 = (speed_ground)*2

*aircraft_make: - 0 = airbus; 1 = boeing.

Chapter 1

Data Preparation

Data

The data is simulated landing data from 950 commercial flights provided in the form of two excel files FAA1.xls & FAA2.xls.

FAA1.xls has 800 rows and FAA2.xls has 150 rows.

Dataset description

- **Aircraft**: The make of an aircraft (Boeing or Airbus).
- **Duration** (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.
- **No_pasg**: The number of passengers in a flight.
- **Speed_ground** (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
- **Speed_air** (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
- **Height** (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway.
- **Pitch** (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway.
- **Distance** (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

Data Import

The first step of our data preparation process is to import the data in the two excel files into SAS.

CODE LOG RESULTS OUTPUT DATA

Line #

```
/*Import Dataset FAA1 */
PROC IMPORT OUT= WORK.FAA1 DATAFILE= "/home/ghosalnh0/statcomputing/FAA1.xls"
            DBMS=xls REPLACE;
    SHEET="FAA1";
    GETNAMES=YES;
RUN;
/*Import Dataset FAA2*/
PROC IMPORT OUT= WORK.FAA2 DATAFILE= "/home/ghosalnh0/statcomputing/FAA2.xls"
            DBMS=xls REPLACE;
    SHEET="FAA2";
    GETNAMES=YES;
RUN;
```

Merge Data

After importing the data into the SAS database, we merge the two tables FAA1 & FAA2 into a single table FAA for the purpose of analysis. On merging the two tables, we observe that there are some blank rows in the resultant table FAA. The resultant table FAA has 1000 observations and 8 variables.

```
/*Merge the two datasets*/
DATA WORK.FAA;
SET WORK.FAA1 WORK.FAA2;
RUN;
PROC PRINT DATA = WORK.FAA;
RUN;

/*Summary statistics for the merged uncleaned dataset*/
PROC MEANS DATA = WORK.faa MIN MAX MEAN MEDIAN STD VAR RANGE;

PROC FREQ DATA = WORK.FAA;
TABLES AIRCRAFT/MISSING;
```

The MEANS Procedure

Variable	Label	Minimum	Maximum	Mean	Median	Std Dev	Variance	Range
duration	duration	14.7642071	305.6217107	154.0065385	153.9480975	49.2592338	2426.47	290.8575036
no_pasg	no_pasg	29.0000000	87.0000000	60.1652632	60.0000000	7.4900041	56.1001619	58.0000000
speed_ground	speed_ground	27.7357153	141.2186354	79.2849940	79.4129094	19.3364178	373.8970524	113.4829200
speed_air	speed_air	90.0028586	141.7249357	103.7304174	100.8916770	10.6051134	112.4684305	51.7220771
height	height	-3.5462524	59.9459639	30.1392714	29.9044945	10.3593491	107.3161131	63.4922163
pitch	pitch	2.2844801	5.9267842	4.0192472	4.0153874	0.5260322	0.2767099	3.6423041
distance	distance	34.0807833	6533.05	1548.82	1267.44	948.6812561	899996.13	6498.97

The FREQ Procedure

aircraft				
aircraft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	50	5.00	50	5.00
airbus	450	45.00	500	50.00
boeing	500	50.00	1000	100.00

Data Cleaning

In this step, we clean the data by removing any missing values, duplicate values and abnormal values. We do this in three steps.

Step 1: Removing duplicates from the dataset

After this step , we are left with 851 observations in our dataset.

```
/*Removing duplicates from the combined dataset*/
PROC SORT DATA = WORK.FAA NODUPKEY OUT = WORK.FAA_UNIQUE;
BY AIRCRAFT NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DISTANCE;
RUN;
PROC PRINT DATA = WORK.FAA_UNIQUE;
RUN;
```

Step 2: Completeness Check

We perform the completeness check of each variable – examine if missing values are present. After that we delete rows with missing values.

```
/*Performing completeness check for the combined dataset*/
PROC MEANS DATA = WORK.faa_unique N NMISS MIN MAX MEAN MEDIAN STD VAR RANGE;

PROC FREQ DATA = WORK.FAA_UNIQUE;
TABLES AIRCRAFT/MISSING;
```

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev	Variance	Range
duration	duration	800	51	14.7642071	305.6217107	154.0065385	153.9480975	49.2592338	2426.47	290.8575036
no_pasg	no_pasg	850	1	29.0000000	87.0000000	60.1035294	60.0000000	7.4931370	56.1471018	58.0000000
speed_ground	speed_ground	850	1	27.7357153	141.2186354	79.4523229	79.6428041	19.0594903	363.2641710	113.4829200
speed_air	speed_air	208	643	90.0028586	141.7249357	103.7977237	101.1473493	10.2590370	105.2478395	51.7220771
height	height	850	1	-3.5462524	59.9459639	30.1442223	30.0931324	10.2877268	105.8373237	63.4922163
pitch	pitch	850	1	2.2844801	5.9267842	4.0093577	4.0082875	0.5288298	0.2796610	3.6423041
distance	distance	850	1	34.0807833	6533.05	1526.02	1258.09	928.5600816	862223.83	6498.97

The FREQ Procedure

aircraft				
aircraft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	1	0.12	1	0.12
airbus	450	52.88	451	53.00
boeing	400	47.00	851	100.00

We observe that variable speed_air has 643 missing values and duration has 51 missing values. At this stage, we do not remove the rows which have missing values for these variables since we do not want to lose data for other variables which may be relevant to our model.

We observe that height has negative value as minimum which cannot be realistically possible. Also, there is one blank row. We remove rows having negative height and blank value for all variables.

```
data FAA;
SET WORK.FAA_UNIQUE;
IF AIRCRAFT="" THEN DELETE;
IF HEIGHT < 0 THEN DELETE ;
PROC PRINT DATA = WORK.FAA_UNIQUE;
RUN;
```

Our dataset now contains 845 observations.

Step 3: Validity Check

In this step, we perform the validity check of each variable – examine if abnormal values are present and then remove rows with abnormal values. We define abnormal values for each variable as per the following table -

Variable	Abnormal Value
Distance	> 6000 feet
Speed_ground	< 30MPH or > 140MPH
Speed_air	< 30MPH or > 140MPH
Duration	< 40 min
Height	> 6 meters

```
/*Removing Abnormal Values*/
DATA FAA_NORMAL;
SET FAA;
IF HEIGHT < 6 THEN DELETE;
IF DISTANCE > 6000 THEN DELETE ;
IF DURATION < 40 AND DURATION ~= . THEN DELETE;
IF SPEED_GROUND < 30 OR SPEED_GROUND > 140 THEN DELETE;
IF (SPEED_AIR < 30 OR SPEED_AIR > 140) AND SPEED_AIR ~= . THEN DELETE ;

PROC PRINT DATA = FAA_NORMAL;
RUN;
```

Data Summarizing

In this step, we summarize the statistics for our clean dataset and distribution of each variable.

```
/*Summarize the statistics for the clean data*/
PROC MEANS DATA = FAA_NORMAL N NMISS MIN MAX MEAN MEDIAN STD VAR RANGE;

PROC FREQ DATA = FAA_NORMAL;
TABLES AIRCRAFT/MISSING;

PROC UNIVARIATE DATA=FAA_NORMAL PLOT;
RUN;
```

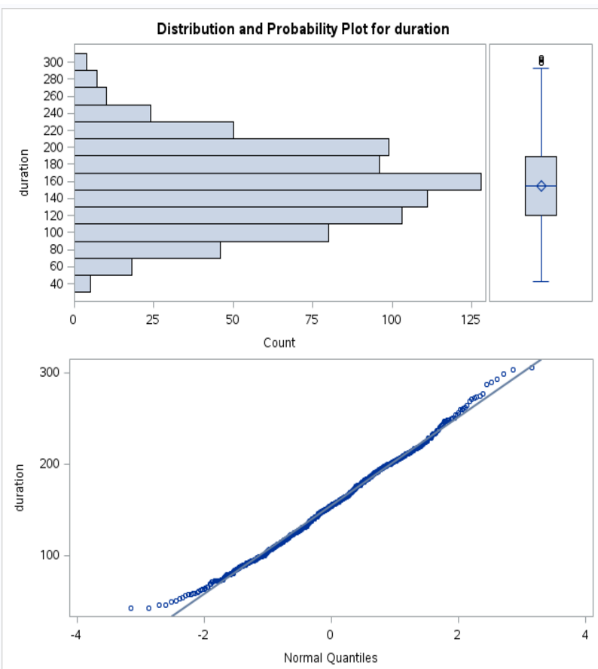
The MEANS Procedure

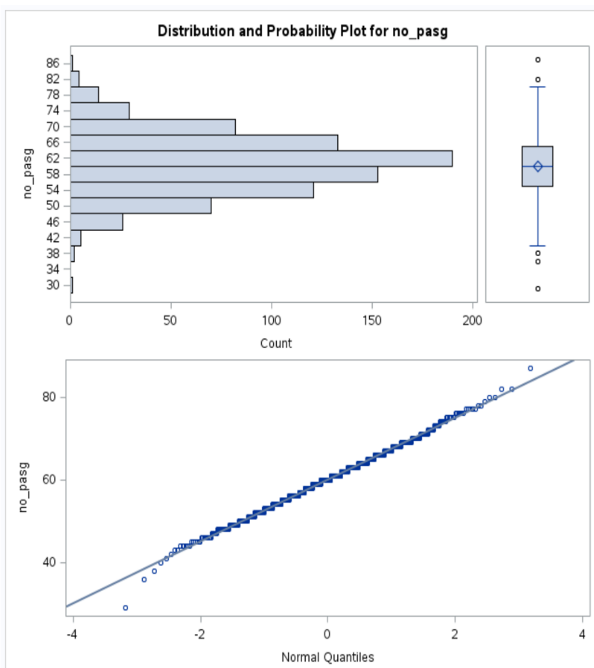
Variable	Label	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev	Variance	Range
duration	duration	781	50	41.9493694	305.6217107	154.7757191	154.2845505	48.3499237	2337.72	263.6723414
no_pasg	no_pasg	831	0	29.0000000	87.0000000	60.0553550	60.0000000	7.4913166	56.1198237	58.0000000
speed_ground	speed_ground	831	0	33.5741041	132.7846766	79.5426997	79.7939604	18.7356754	351.0255334	99.2105726
speed_air	speed_air	203	628	90.0028586	132.9114649	103.4850352	101.1189240	9.7362774	94.7950972	42.9086063
height	height	831	0	6.2275178	59.9459639	30.4578695	30.1670844	9.7848114	95.7425347	53.7184462
pitch	pitch	831	0	2.2844801	5.9267842	4.0051609	4.0010380	0.5265690	0.2772750	3.6423041
distance	distance	831	0	41.7223127	5381.96	1522.48	1262.15	896.3381524	803422.08	5340.24

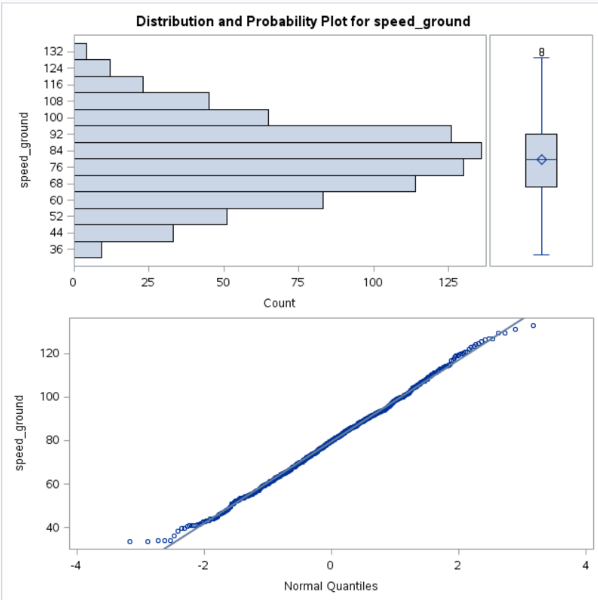
The FREQ Procedure

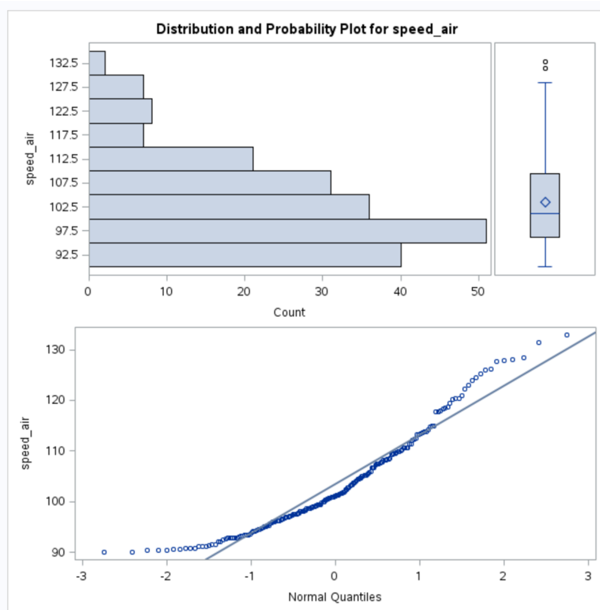
aircraft				
aircraft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
airbus	444	53.43	444	53.43
boeing	387	46.57	831	100.00

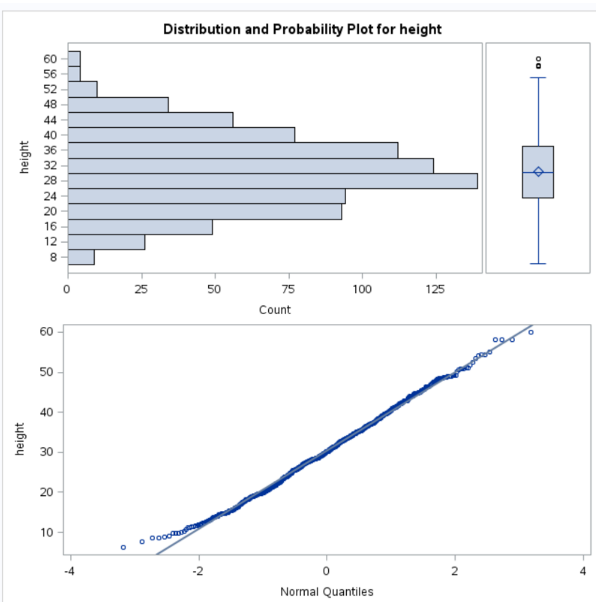
Our clean dataset consists of 831 observations. Although speed_air has missing values for 628 observations and duration has missing values for 50 observations, we don't discard the rows having these variables at this point in our analysis.

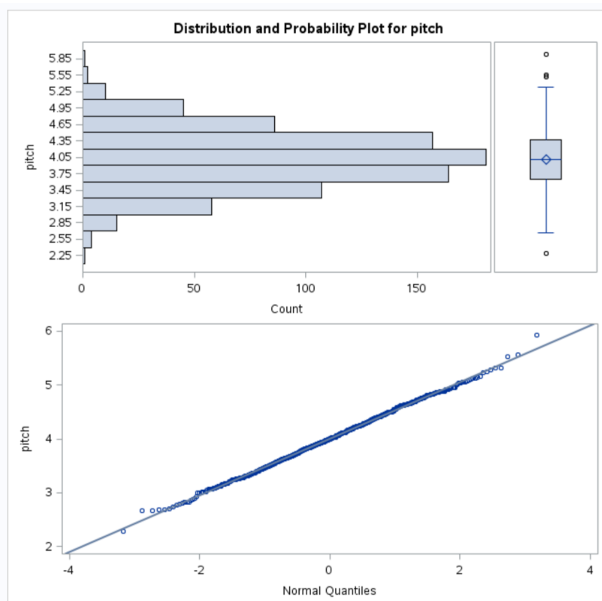


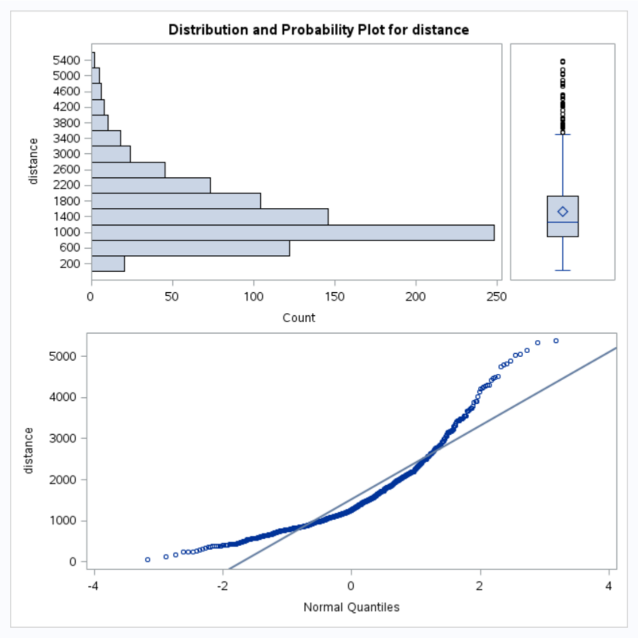



Distribution and Probability Plot for speed_ground
speed_ground
132
124
116
108
100
92
84
76
68
60
52
44
36
0 25 50 75 100 125
Count
120
100
80
60
40
-4 -2 0 2 4
Normal Quantiles


Distribution and Probability Plot for speed_air
speed_air
132.5
127.5
122.5
117.5
112.5
107.5
102.5
97.5
92.5
0 10 20 30 40 50
Count
130
120
110
100
90
-3 -2 -1 0 1 2 3
Normal Quantiles


Distribution and Probability Plot for height
height
60
56
52
48
44
40
36
32
28
24
20
16
12
8
0 25 50 75 100 125
Count
60
50
40
30
20
10
-4 -2 0 2 4
Normal Quantiles


Distribution and Probability Plot for pitch
pitch
5.85
5.55
5.25
4.95
4.65
4.35
4.05
3.75
3.45
3.15
2.85
2.55
2.25
0 50 100 150
Count
6
5
4
3
2
-4 -2 0 2 4
Normal Quantiles



From the distributions of each variable we observe that duration, no_pasg, speed_ground, height and pitch follow normal distribution but speed_air and distance do not appear to be normally distributed.

Questions

While doing the data preparation for our dataset, I came up with following questions: -

- What should be taken as unique identifier for our dataset?
- Speed_air is not given for 628 rows and duration for 50 rows, will this impact our analysis result? Should we remove rows that contain missing values or should we perform imputation to complete the dataset?
- Which variables should be taken into account for data modelling?

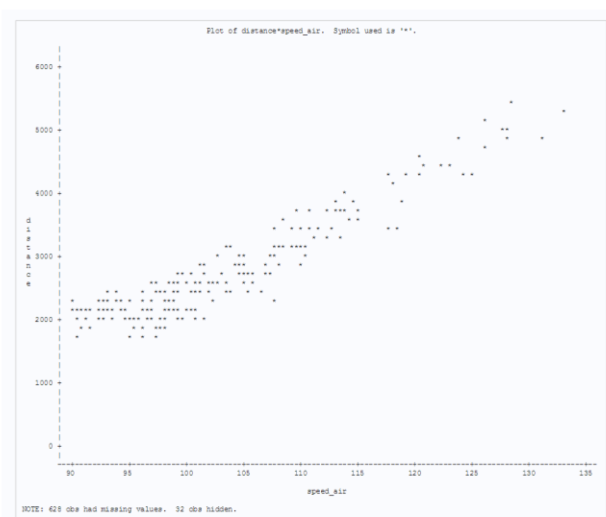
Chapter 2

Exploratory Data Analysis & Data Visualization

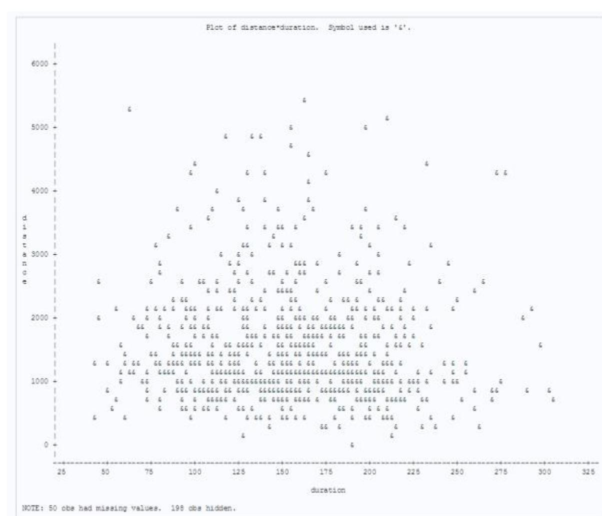
The objective is to study the co-relation between independent and dependent variables by exploring and visualizing data. Here we will study how distance (dependent variable) is co-related with other variables. We will then eliminate those variables from the regression model which do not have a strong correlation with our dependent variable.

Step 1: We plot graphs for our dataset. From these graphs, we can identify which of the independent variables has a linear relationship with our dependent variable distance.

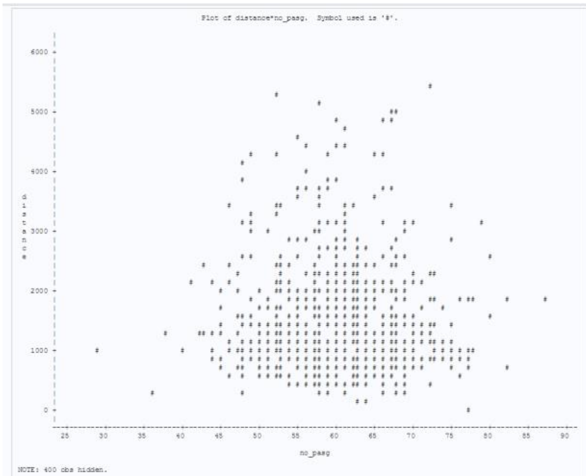
```

/*Data Visualization*/

PROC PLOT DATA = FAA_Normal;
plot distance*speed_air='*';
plot distance*duration='&';
plot distance*no_pasg='#';
plot distance*speed_ground='@';
plot distance*height='!';
plot distance*pitch='$';

```



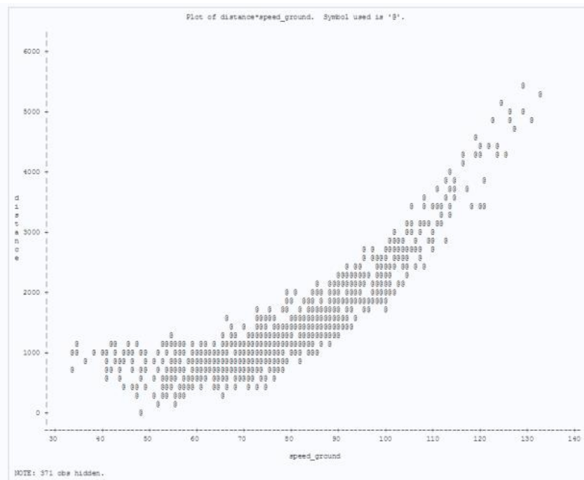
Distance vs Speed_air



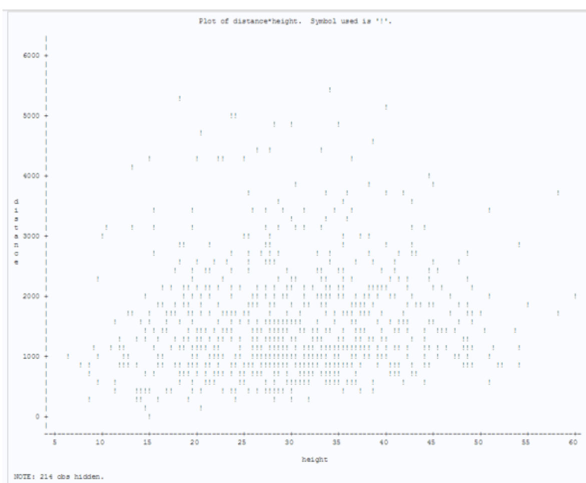
Distance vs duration



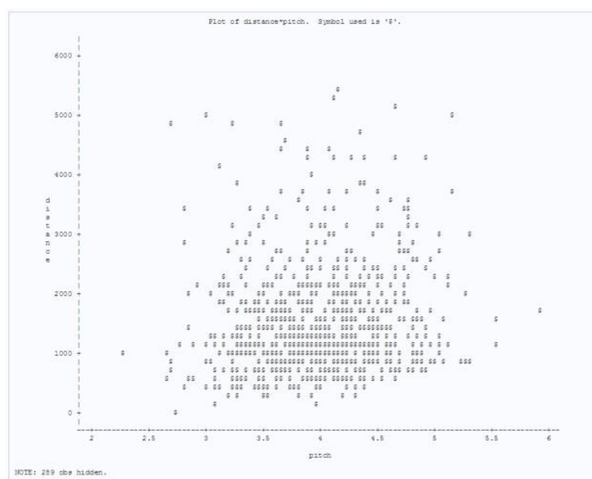
Distance vs no_pasg



Distance vs Speed_ground



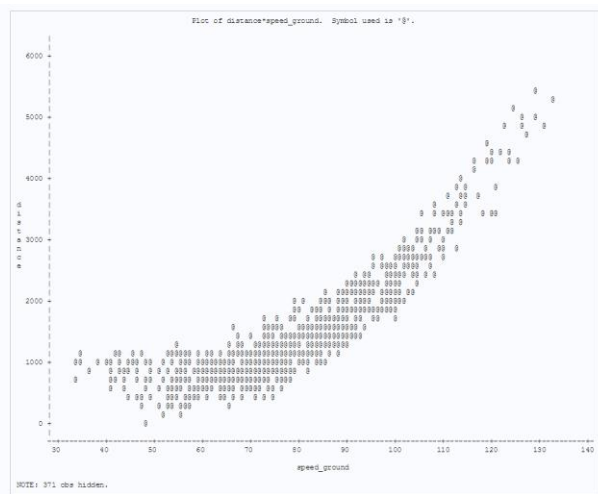
Distance vs height



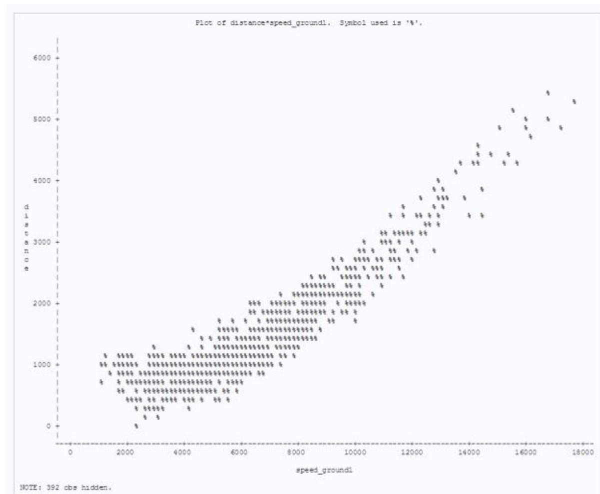
Distance vs pitch

Step 2: We will perform transformation of variable speed_ground to increase linearity with distance. Also, we will transform the categorical variable aircraft into a binary variable so that it could be included in our model.

```
/*Transformation of speed ground to improve linear relationship with distance*/
DATA FAA_NORMAL1;
set faa_normal;
speed_ground1 = (speed_ground)**2;

PROC PLOT DATA = FAA_NORMAL1;
PLOT DISTANCE*SPEED_GROUND1='%';
plot distance*speed_ground='@';
```



Distance vs speed_ground



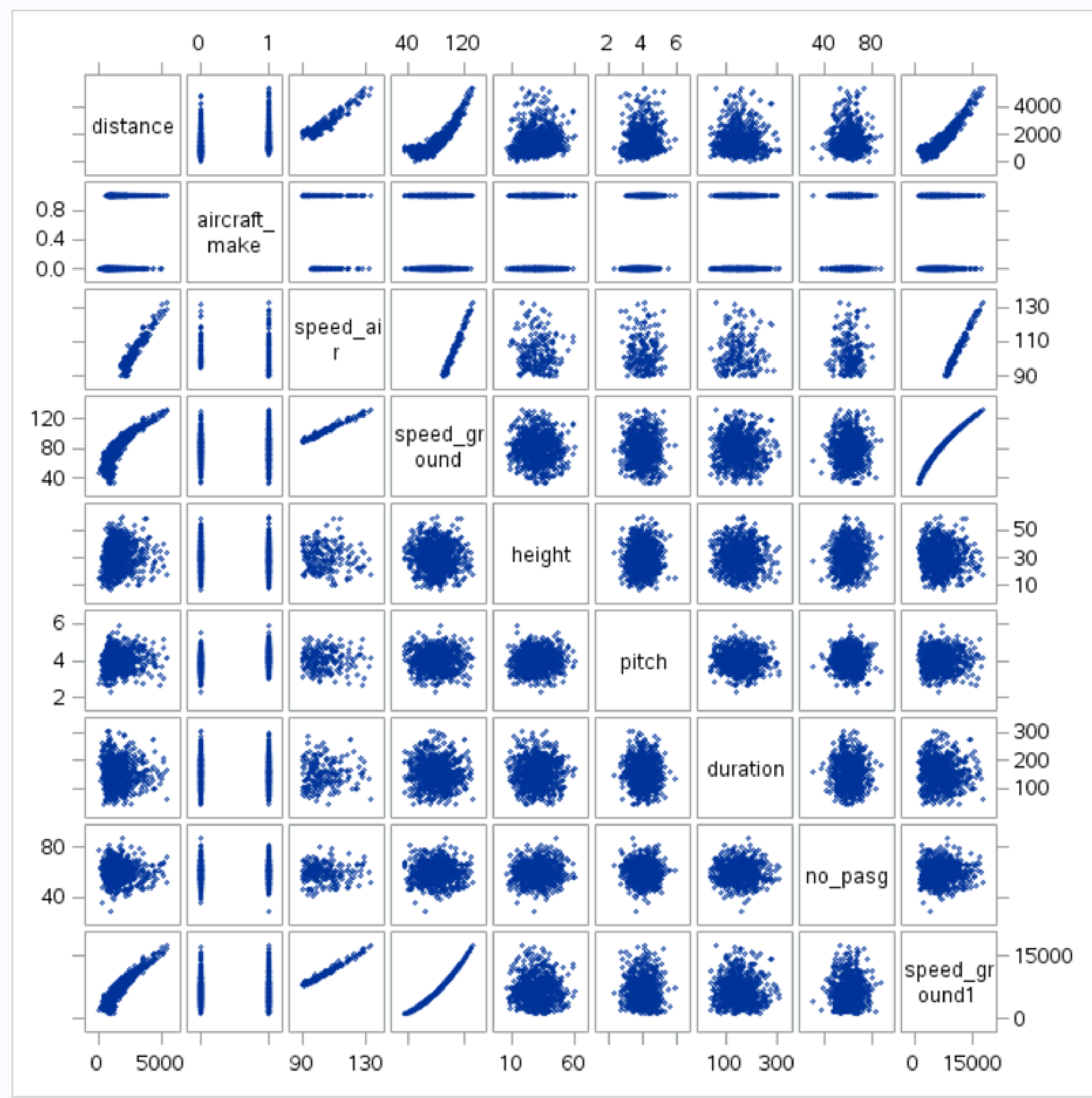
Distance vs (speed_ground)^2

We observe that speed_ground1 has a slightly better linear relationship with distance.

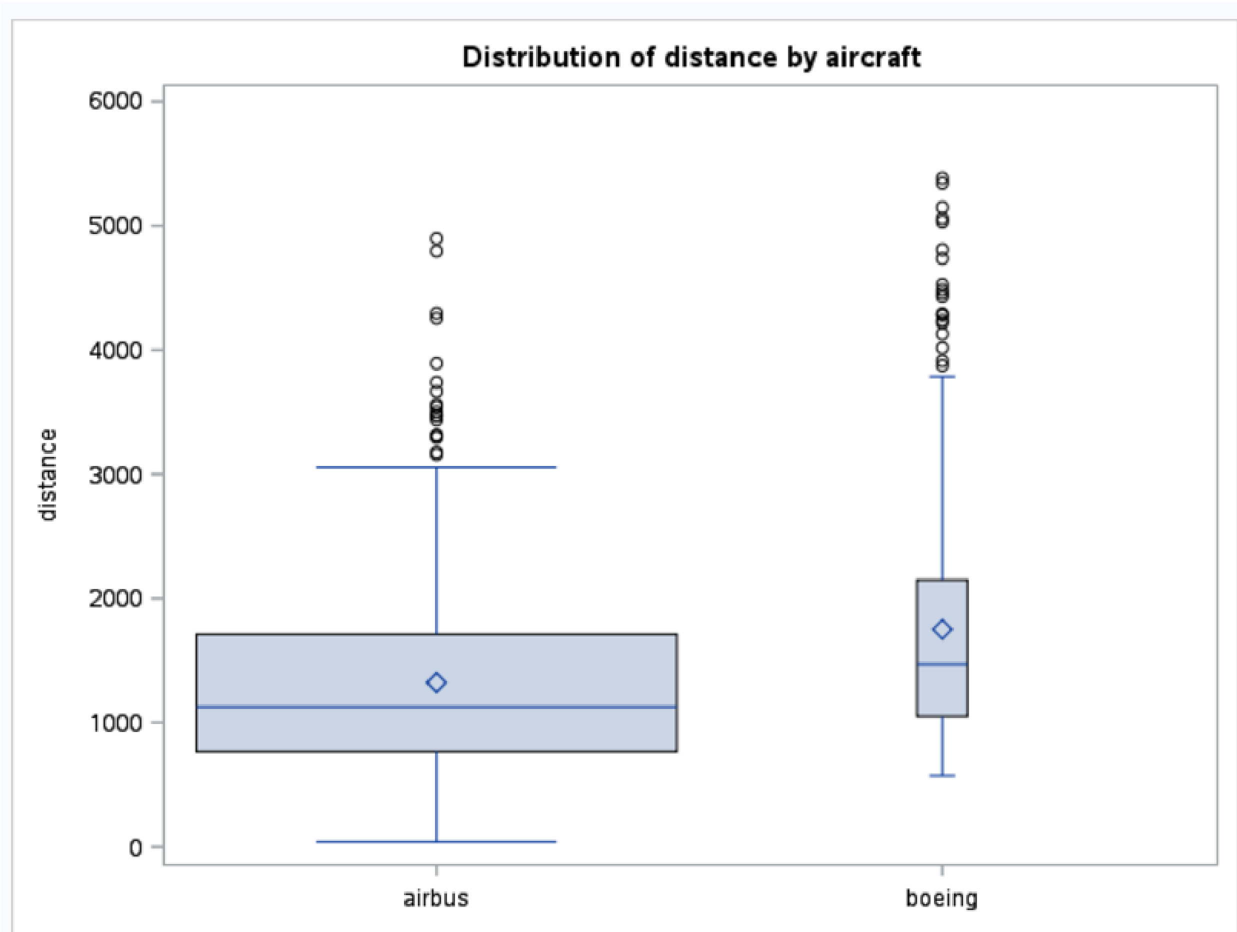
```
/*Transfroming categorical column Aircraft from non-numeric to binary*/
DATA FAA_Final;
SET FAA_Normal1;
IF aircraft = 'boeing' then aircraft_make = 1;
ELSE aircraft_make = 0;
PROC PRINT DATA = FAA_Final;
RUN;
```

Step 3 : Using scatter plot to understand the relationships between all variables. A box plot was plotted to see the differences in landing distance because of its make.

```
/* Using scatter plot to identify the relation among different variables */
PROC SGSCATTER DATA=FAA_Final;
MATRIX distance aircraft_make speed_air speed_ground height pitch duration no_pasg speed_ground1 ;
RUN;

/* Using box plot to identify the relation of distance with aircraft make*/
PROC SORT DATA=FAA_Final;
BY AIRCRAFT;
run;

PROC BOXPLOT DATA=FAA_Final;
PLOT DISTANCE*AIRCRAFT/
        nohlabel
        boxstyle      = schematic
        boxwidthscale = 1
        bwslegend;
run;
```



We see that speed_air, speed_ground and speed_ground1 have some linear relationship with landing distance.



We observe that there is indeed a difference in distance because of aircraft make; for airbus make the landing distance mean, median is lower than that for boeing.

Step 4: Next step is to validate the linear relationships found by the various plots; that is finding the strength of linear relationship between dependent and independent variables. We also need to check if there is any linear relationship between any independent variables. We will only keep one of the related independent variables in our model in order to avoid multicollinearity problems in our analysis.

```
/*Correlation Coefficient*/

PROC CORR DATA = FAA_Final;
VAR distance duration no_pasg speed_ground speed_air height pitch aircraft_make speed_ground1;
title Correlaiton coefficients matrix;
run;
```

The CORR Procedure

9 Variables:	distance duration no_pasg speed_ground speed_air height pitch aircraft_make speed_ground1

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	831	1522	896.33815	1265183	41.72231	5382	distance
duration	781	154.77572	48.34992	120880	41.94937	305.62171	duration
no_pasg	831	60.05535	7.49132	49906	29.00000	87.00000	no_pasg
speed_ground	831	79.54270	18.73568	66100	33.57410	132.78468	speed_ground
speed_air	203	103.48504	9.73628	21007	90.00286	132.91146	speed_air
height	831	30.45787	9.78481	25310	6.22752	59.94596	height
pitch	831	4.00516	0.52657	3328	2.28448	5.92678	pitch
aircraft_make	831	0.46570	0.49912	387.00000	0	1.00000	
speed_ground1	831	6678	3047	5549122	1127	17632	

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

	distance	duration	no_pasg	speed_ground	speed_air	height	pitch	aircraft_make	speed_ground1
distance distance	1.00000 831	-0.05138 0.1514 781	-0.01776 0.6093 831	0.86624 <.0001 831	0.94210 <.0001 203	0.09941 0.0041 831	0.08703 0.0121 831	0.23814 <.0001 831	0.91657 <.0001 831
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	-0.04443 0.2149 781	-0.04849 0.1758 781
no_pasg no_pasg	-0.01776 0.6093 831	-0.03639 0.3098 781	1.00000 831	-0.00013 0.9969 831	-0.00616 0.9305 203	0.04699 0.1760 831	-0.01793 0.6057 831	-0.02269 0.5136 831	-0.00182 0.9582 831
speed_ground speed_ground	0.86624 <.0001 831	-0.04897 0.1716 781	-0.00013 0.9969 831	1.00000 831	0.98794 <.0001 203	-0.05761 0.0970 831	-0.03912 0.2599 831	-0.04045 0.2441 831	0.98831 <.0001 831
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	-0.07207 0.3069 203	0.98774 <.0001 203
height height	0.09941 0.0041 831	0.01112 0.7564 781	0.04699 0.1760 831	-0.05761 0.0970 831	-0.07933 0.2606 203	1.00000 831	0.02298 0.5082 831	-0.01439 0.6788 831	-0.05417 0.1187 831
pitch pitch	0.08703 0.0121 831	-0.04675 0.1918 781	-0.01793 0.6057 831	-0.03912 0.2599 831	-0.03927 0.5780 203	0.02298 0.5082 831	1.00000 831	0.35420 <.0001 831	-0.02900 0.4037 831
aircraft_make	0.23814 <.0001 831	-0.04443 0.2149 781	-0.02269 0.5136 831	-0.04045 0.2441 831	-0.07207 0.3069 203	-0.01439 0.6788 831	0.35420 <.0001 831	1.00000 831	-0.01731 0.6183 831
speed_ground1	0.91657 <.0001 831	-0.04849 0.1758 781	-0.00182 0.9582 831	0.98831 <.0001 831	0.98774 <.0001 203	-0.05417 0.1187 831	-0.02900 0.4037 831	-0.01731 0.6183 831	1.00000 831

From the correlation matrix, it is observed that variables aircraft_make, speed_air, speed_ground, speed_ground1, height have significant correlation with distance variable.

We also observe that speed_air and speed_ground are highly correlated and speed_ground1 is highly correlated with both speed_air and speed_ground.

Since there is high correlation between speed_air and speed_ground we can drop one of these two variables. We choose to drop speed_air from our model as it has missing values for almost 75 percent of the observations in our dataset and also because we wish to fit our model to as many observations as possible, so we choose to keep speed_ground.

Chapter 3

Data Modelling

The objective of this chapter is to build a model that best describes the relationship between our dependent variable and independent variables. We will use linear regression to build our model since we observed in the previous chapter that there is strong correlation between distance and other independent variables.

A linear equation has the form

$$Y = mX + C + \epsilon$$

Where C = intercept

Coefficient m is the parameter estimate for variable X

ϵ is the error

Step 1: We run the initial model by including all the independent variables.

```
/*Regression Model*/
proc reg data=FAA_Final;
model distance = duration no_pasg speed_ground speed_air height pitch aircraft_make speed_ground1;
title Regression analysis of the Aircraft Dataset;
run;
```

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	688.76797	882.42464	0.78	0.4361
duration	duration	1	0.12038	0.17672	0.68	0.4966
no_pasg	no_pasg	1	-2.24468	1.19448	-1.88	0.0618
speed_ground	speed_ground	1	-126.03696	16.39649	-7.69	<.0001
speed_air	speed_air	1	78.41922	5.72223	13.70	<.0001
height	height	1	13.73731	0.89995	15.26	<.0001
pitch	pitch	1	-5.94312	16.15165	-0.37	0.7133
aircraft_make		1	400.52367	19.01675	21.06	<.0001
speed_ground1		1	0.60155	0.07575	7.94	<.0001

From the parameter estimates, we observe that p-values for duration, no_pasg and pitch are greater than 0.05. Therefore, all of them are dropped from the model equation since they are statistically insignificant. We can also drop speed_air from our model since we observed in the previous chapter that it has high collinearity with speed_ground as well as speed_ground1.

Step 2: Run the regression analysis again after removing statistically insignificant variables identified in Step 1.

We also know that speed_ground and speed_ground1 have high collinearity. So, we will keep one of the two independent variables in our model. We will run two regressions, one with speed_ground and other with speed_ground1, along with other independent variables such as aircraft_make height and choose the best model between the two.

```
proc reg data=FAA_Final;
model distance = speed_ground height aircraft_make;
title Regression analysis of the Aircraft Dataset;
run;
```

Regression analysis of the Aircraft Dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	566080053	188693351	1548.72	<.0001
Error	827	100760276	121838		
Corrected Total	830	666840329			

Root MSE	349.05344	R-Square	0.8489
Dependent Mean	1522.48287	Adj R-Sq	0.8484
Coeff Var	22.92659		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2512.24333	68.19743	-36.84	<.0001
speed_ground	speed_ground	1	42.40242	0.64830	65.41	<.0001
height	height	1	14.14783	1.24046	11.41	<.0001
aircraft_make		1	496.04524	24.29753	20.42	<.0001

```
proc reg data=FAA_Final;
model distance = speed_ground1 height aircraft_make;
title Regression analysis of the Aircraft Dataset;
run;
```

Regression analysis of the Aircraft Dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	618905549	206301850	3559.25	<.0001
Error	827	47934780	57962		
Corrected Total	830	666840329			

Root MSE	240.75350	R-Square	0.9281
Dependent Mean	1522.48287	Adj R-Sq	0.9279
Coeff Var	15.81322		

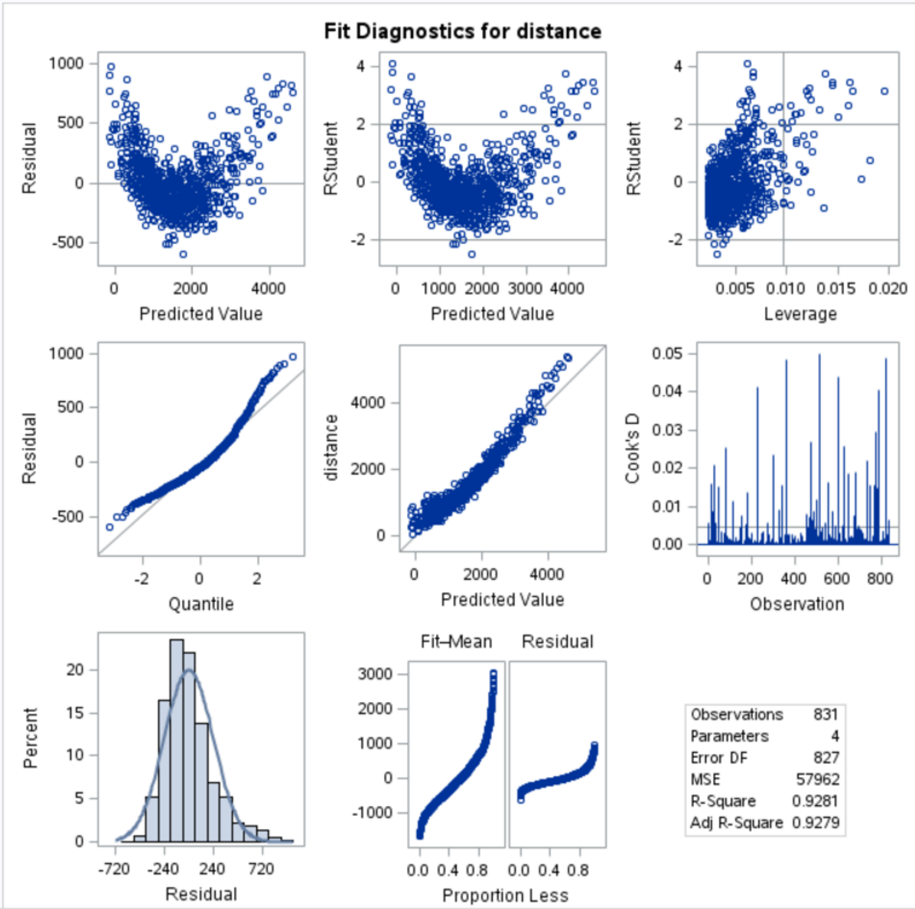
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-945.44865	34.77435	-27.19	<.0001
speed_ground1		1	0.27335	0.00275	99.52	<.0001
height	height	1	14.05637	0.85540	16.43	<.0001
aircraft_make		1	460.52232	16.74721	27.50	<.0001

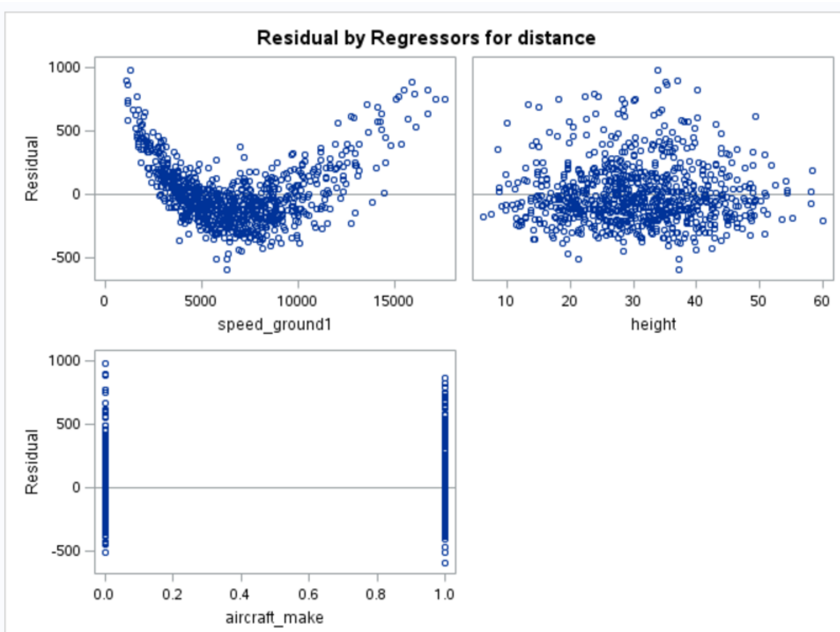
Comparing the R-Square and Adj R-sq values , we choose the the model containing independent variable speed_ground1 since we can predict the value of our dependent variable with greater variability (~93 percent compared to ~85 percent) and less noise (~7 percent compared to ~15 percent).

Regression analysis of the Aircraft Dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance



Observations	831
Parameters	4
Error DF	827
MSE	57962
R-Square	0.9281
Adj R-Square	0.9279



Model equation -

Distance = -945.44 + 0.27*(speed_ground1) + 14.05*(height) + 460.52*(aircraft_make)

Questions to be answered

1. **How many observations (flights) do you use to fit your final model? If not all 950 flights, why?**
   We used 831 observations to fit our final model. The rest 119 observations were removed from our data set in the various data prep steps taken in Chapter 1 such as removing duplications and missing values, performing completeness check and removing abnormal values.

2. **What factors and how they impact the landing distance of a flight?**
   As seen from the model equation, landing distance depends on three factors:
   i. Speed_ground1, i.e. square of speed_ground. There will be an increase of 0.27 units in landing distance for every unit increase in square of speed_ground, keeping all other parameters constant.
   ii. Height. There will be an increase of 14.05 units in landing distance for every unit increase in height, keeping all other parameters constant.
   iii. Aircraft_make. For Boeing make aircraft the landing distance would be 460.52 units greater than that for Airbus make aircraft, keeping all other parameters constant.
3. **Is there any difference between the two makes Boeing and Airbus?**
   Yes there is a significant difference between the two makes of aircraft. Landing distance has a direct correlation with aircraft make.

   We perform GLM and T test to study the difference in the populations of the two makes and their impact on landing distance.

```
proc glm data = FAA_Final;
class aircraft;
model distance = speed_ground height aircraft_make;


proc ttest data = FAA_Final;
class aircraft;
VAR distance;
```

The GLM Procedure

Class Level Information		
Class	Levels	Values
aircraft	2	airbus boeing

Number of Observations Read	831
Number of Observations Used	831

The GLM Procedure
Dependent Variable: distance distance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	566080052.9	188693351.0	1548.72	<.0001
Error	827	100760276.3	121838.3		
Corrected Total	830	666840329.2			

R-Square	Coeff Var	Root MSE	distance Mean
0.848899	22.92659	349.0534	1522.483

Source	DF	Type I SS	Mean Square	F Value	Pr > F
speed_ground	1	500382566.8	500382566.8	4106.94	<.0001
height	1	14916377.4	14916377.4	122.43	<.0001
aircraft_make	1	50781108.7	50781108.7	416.79	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
speed_ground	1	521207999.7	521207999.7	4277.87	<.0001
height	1	15848830.0	15848830.0	130.08	<.0001
aircraft_make	1	50781108.7	50781108.7	416.79	<.0001

The TTEST Procedure
Variable: distance (distance)

aircraft	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus	444	1323.3	791.9	37.5833	41.7223	4896.3
boeing	387	1751.0	953.9	48.4869	573.6	5382.0
Diff (1-2)		-427.7	871.1	60.5772		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1323.3	1249.5	1397.2	791.9	743.0	847.8
boeing		1751.0	1655.7	1846.3	953.9	891.1	1026.2
Diff (1-2)	Pooled	-427.7	-546.6	-308.8	871.1	831.1	915.1
Diff (1-2)	Satterthwaite	-427.7	-548.1	-307.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	829	-7.06	<.0001
Satterthwaite	Unequal	752.49	-6.97	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	386	443	1.45	0.0002

Since F-value is greater than 1, we can say that there is a significant difference between the two aircraft makes. We observe that there is a difference in the mean, standard deviation , variation and impact on landing distance of the two makes.