BANA 6043 Project

Summary:

- In the following project, two flight datasets were given to study what factors and how they would impact the landing distance of a commercial flight.
- The two datasets were firstly combined and trimmed in order to produce 950 observation and 8 variables.
- Looking at the dataset, it was observed there were few duplicate observations along with some abnormal values which were removed accordingly.
- Performed some basic summary analysis on each of the variables to detect whether there were any abnormal values and how were they distributed according to the aircraft type.
- In the next step, each predictor variable was plotted against landing_distance to detect any noticeable relation. Correlation Matrix was also computed to see the pearson correlation coefficients between our predictor and response.
- The variables which had relatively high values were then selected to be used in our linear regression model. However since speed_air and speed_ground were highly correlated among themselves, I decided to use Speed_air alone as it was giving me better R-Square and residual plots.
- The final model was built using Speed_air and height variable and the equation comes out to be:

  $\mathbf{Y = -5935.73 + 80.49*X_1 + 12.54*X_2}$

  Y = Landing_distance

  X_1 = Speed_air

  X_2 = Height

Ashutosh Singh

M13433470

Chapter 1: Data Understanding and Exploration

Dataset:

Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). Two Excel files: 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights).

Variable Names:

Aircraft (Make of an Aircraft)

duration (Duration of the flight in Minutes)

no_pasg (Number of passengers)

speed_ground (ground speed of an aircraft in mph when passing over the threshold of the runway)

speed_air (air speed of an aircraft in mph when passing over the threshold of the runway)

height (height of an aircraft in meters when it is passing over the threshold of the runway)

pitch (Pitch angle of an aircraft in degrees)

distance (landing distance of an aircraft in feet)

Importing the Datasets:

Importing FAA1.xls file:

```
PROC IMPORT out = WORK.FAA1 datafile= "/folders/myfolders/FAA1.xls"
    dbms=xls replace;
    sheet="FAA1";
    getnames=yes;
RUN;
PROC CONTENTS DATA=WORK.FAA1;
RUN;
```

After running the above command, we get Table **"WORK.FAA1"** with **800 rows and 8 columns**.

The CONTENTS Procedure

Data Set Name	WORK.FAA1	Observations	800
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	11/13/2019 12:09:22	Observation Length	72
Last Modified	11/13/2019 12:09:22	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	12	$12.	$12.	aircraft
8	distance	Num	8	BEST12.		distance
2	duration	Num	8	BEST12.		duration
6	height	Num	8	BEST12.		height
3	no_pasg	Num	8	BEST12.		no_pasg
7	pitch	Num	8	BEST12.		pitch
5	speed_air	Num	8	BEST12.		speed_air
4	speed_ground	Num	8	BEST12.		speed_ground

Importing FAA2.xls file:

```
PROC IMPORT out = WORK.FAA2 datafile= "/folders/myfolders/FAA2.xls"
    dbms=xls replace;
    sheet="FAA2";
    getnames=yes;
RUN;
PROC CONTENTS DATA=WORK.FAA2;
RUN;
```

After running the above command, we get Table **"WORK.FAA2"** with **200 rows and 7 columns.**

The CONTENTS Procedure

Data Set Name	WORK.FAA2	Observations	200
Member Type	DATA	Variables	7
Engine	V9	Indexes	0
Created	11/13/2019 12:16:16	Observation Length	64
Last Modified	11/13/2019 12:16:16	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	12	$12.	$12.	aircraft
7	distance	Num	8	BEST12.		distance
5	height	Num	8	BEST12.		height
2	no_pasg	Num	8	BEST12.		no_pasg
6	pitch	Num	8	BEST12.		pitch
4	speed_air	Num	8	BEST12.		speed_air
3	speed_ground	Num	8	BEST12.		speed_ground

Removing Empty rows in FAA2 Table:

After looking into the FAA2 table, it was found that the table contained many empty rows at the end. So, the following command was used to remove the empty rows:

```
OPTIONS MISSING='';
DATA FAA2_CLEANED;
       SET FAA2;
IF MISSING(CATS(of _all_)) THEN DELETE;
RUN;
```

After performing this step, the table **FAA2 contained 150 rows and 7 columns**.

Combining the Two Datasets:

```
DATA FAA_COMBINED;
       SET FAA1 FAA2_CLEANED;
RUN;
```

After performing this step, we get combined dataset of **950 rows and 8 columns**.

Looking at the table however we see that there are number of duplicate rows except the duration column. So, there might be the case that there are lot of records in FAA2 which are already in FAA1 but they don't have the duration column.

```
PROC SORT DATA=FAA_COMBINED OUT=FAA_COMBINED_UNIQUE NODUPKEY;
BY aircraft no_pasg Speed_Ground Speed_Air height pitch distance;
RUN;
PROC CONTENTS DATA=FAA_COMBINED_UNIQUE;
RUN;
```

The CONTENTS Procedure

Data Set Name	WORK.FAA_COMBINED_UNIQUE	Observations	850
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	11/13/2019 14:09:37	Observation Length	72
Last Modified	11/13/2019 14:09:37	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)		

After performing the above step, we have removed 100 duplicate records from our dataset.

Finding the Number of Missing Values for Each Variable:

```
PROC MEANS DATA=FAA_COMBINED_UNIQUE NMISS N;
TITLE NUMBER OF MISSING VALUES IN EACH VARIABLE;
RUN;
```

NUMBER OF MISSING VALUES IN EACH VARIABLE

The MEANS Procedure

Variable	Label	N Miss	N
duration	duration	50	800
no_pasg	no_pasg	0	850
speed_ground	speed_ground	0	850
speed_air	speed_air	642	208
height	height	0	850
pitch	pitch	0	850
distance	distance	0	850

Here N Miss denotes the number of Missing values for each variable. As we can see, ‘duration’ and ‘speed_air’ are the two variables with the most missing values.

Before taking any action on missing values of duration and Speed_air, we need to see the percent of missing values for each variable.

```
PROC SQL;
TITLE PERCENT OF MISSING VALUES FOR DURATION AND SPEED_AIR;
SELECT (NMISS(Duration) / COUNT(*))*100 AS miss_Duration_pct ,
 (NMISS(Speed_Air)/ COUNT(*))*100 AS miss_Speed_Air_pct
from FAA_COMBINED_UNIQUE
QUIT;
```

PERCENT OF MISSING VALUES FOR DURATION AND SPEED_AIR

miss_Duration_pct	miss_Speed_Air_pct
5.882353	75.52941

As we can see, the percent of missing values for Speed_Air is too high to remove them or to do any sort of Imputation as it will create unnecessary bias. In such cases, we have to report to the client about this situation and wait for their confirmation before dealing with the Speed_Air variable.

Detecting Abnormal values for Each Variable:

Now that we have found out missing values, let us see whether our dataset contains abnormal values according to the guidelines provided to us.

For Duration

```
PROC SQL;
TITLE Number of Abnormal values for Duration;
SELECT count(*) as Abnormal_Values
FROM FAA_COMBINED_UNIQUE
WHERE duration < 40 AND MISSING(duration) = 0;
RUN;
```

Number of Abnormal values for Duration

Abnormal_Values
5

```
PROC SQL;
TITLE Number of Abnormal values for Speed_Ground;
SELECT count(*) as Abnormal_Values
FROM FAA_COMBINED_UNIQUE
WHERE speed_ground NOT BETWEEN 30 AND 140;
RUN;
```

Number of Abnormal values for Speed_Ground

Abnormal_Values
3

```
PROC SQL;
TITLE Number of Abnormal values for Speed_Air;
SELECT count(*) as Abnormal_Values
FROM FAA_COMBINED_UNIQUE
WHERE (speed_air NOT BETWEEN 30 AND 140) AND MISSING(speed_air) =
0;
RUN;
```

Number of Abnormal values for Speed_Air

Abnormal_Values
1

```
PROC SQL;
TITLE Number of Abnormal values for height;
SELECT count(*) as Abnormal_Values
FROM FAA_COMBINED_UNIQUE
WHERE height < 6;
RUN;
```

Number of Abnormal values for height

Abnormal_Values
10

```
PROC SQL;
TITLE Number of Abnormal values for distance;
SELECT count(*) as Abnormal_Values
FROM FAA_COMBINED_UNIQUE
WHERE distance > 6000;
RUN;
```

Number of Abnormal values for distance

Abnormal_Values
2

As we can see that each variable has some abnormal values, but they are not in large amount.

Removing the abnormal values:

Since there are few abnormal values, so we can delete the observations containing them.

```
DATA FAA_CLEANED;
       SET FAA_COMBINED_UNIQUE;
IF duration <= 40 AND MISSING(duration) = 0 THEN DELETE;
IF height < 6 AND MISSING(height) = 0 THEN DELETE;
IF distance > 6000 AND MISSING(distance) = 0 THEN DELETE;
IF (speed_air < 30 OR speed_air > 140) AND MISSING(speed_air) = 0
THEN DELETE;
IF (speed_ground < 30 OR speed_ground > 140) AND
MISSING(speed_ground) = 0 THEN DELETE;
RUN;
```

After performing this operation, we get new **dataset containing 831 rows and 8 columns.**

```
PROC CONTENTS DATA=FAA_CLEANED;
RUN;
```

The CONTENTS Procedure

Data Set Name	WORK.FAA_CLEANED	Observations	831
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	11/13/2019 21:21:52	Observation Length	72
Last Modified	11/13/2019 21:21:52	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Summarizing the Distribution of Each Variable:

Checking whether there are no more abnormal values and what is the range of each numeric variable:

```
PROC MEANS DATA = FAA_CLEANED N NMISS MEAN STD MIN MAX;
TITLE Basic Summary of each Variable;
RUN;
```

Basic Summary of each Variable

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Std Dev	Minimum	Maximum
duration	duration	781	50	154.7757191	48.3499237	41.9493694	305.6217107
no_pasg	no_pasg	831	0	60.0553550	7.4913166	29.0000000	87.0000000
speed_ground	speed_ground	831	0	79.5426997	18.7356754	33.5741041	132.7846766
speed_air	speed_air	203	628	103.4850352	9.7362774	90.0028586	132.9114649
height	height	831	0	30.4578695	9.7848114	6.2275178	59.9459639
pitch	pitch	831	0	4.0051609	0.5265690	2.2844801	5.9267842
distance	distance	831	0	1522.48	896.3381524	41.7223127	5381.96

As we can see from the output, there are no abnormal values and all values are within the normal range.

Summarizing variables for each type of aircraft:

```
PROC MEANS DATA=FAA_CLEANED NOPRINT;
BY aircraft;
OUTPUT OUT=FAA_summary1 MEAN(duration speed_ground ) =
MeanDuration MeanSpeed_G
STD(duration speed_ground) = StdDuration StdSpeed_G;
RUN;
PROC PRINT DATA=FAA_summary1;
Title Summary statistics of duration and speed for each airline;
RUN;
```

Summary statistics of duration and speed for each airline

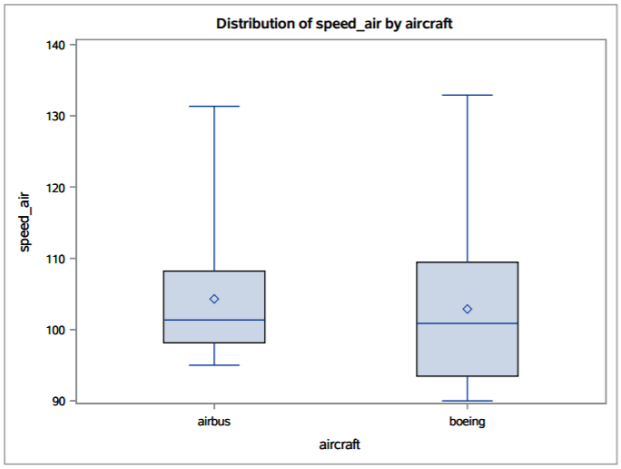
Obs	aircraft	_TYPE_	_FREQ_	MeanDuration	MeanSpeed_G	StdDuration	StdSpeed_G
1	airbus	0	444	156.90332993	80.249875727	49.18828988	16.954969277
2	boeing	0	387	152.60962427	78.731366049	47.446717931	20.582498993

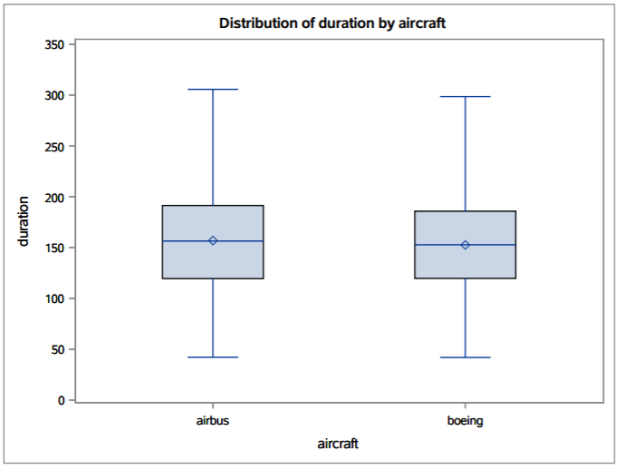
```
PROC MEANS DATA=FAA_CLEANED NOPRINT;
BY aircraft;
OUTPUT OUT=FAA_summary2 MEAN(height distance) =
MeanHeight MeanDistance
STD(height distance) = StdHeight StdDistance;
RUN;
PROC PRINT DATA=FAA_summary2;
Title Summary statistics of height and distance for each airline;
RUN;
```

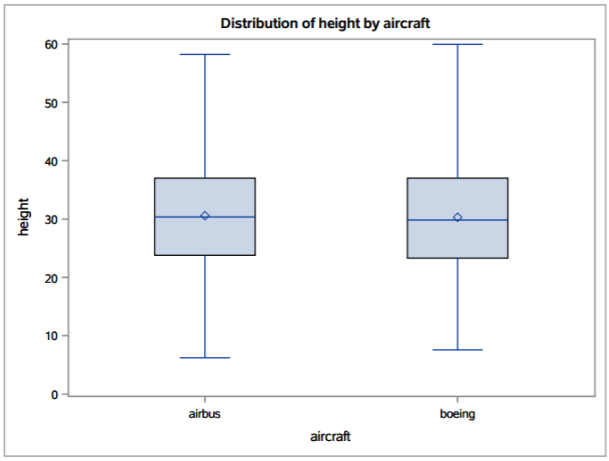
Summary statistics of height and distance for each airline

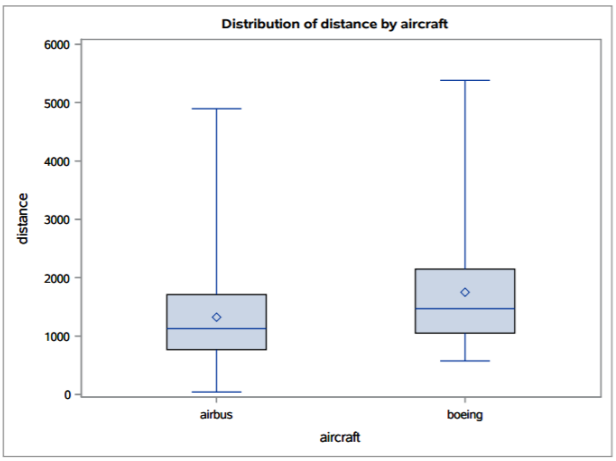
Obs	aircraft	_TYPE_	_FREQ_	MeanHeight	MeanDistance	StdHeight	StdDistance
1	airbus	0	444	30.589221821	1323.3169623	9.8543912304	791.92824815
2	boeing	0	387	30.307170716	1750.9832977	9.7149203775	953.85003001

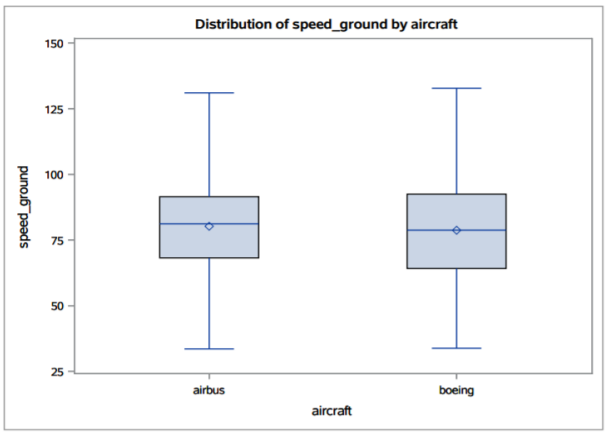
Distribution Charts:

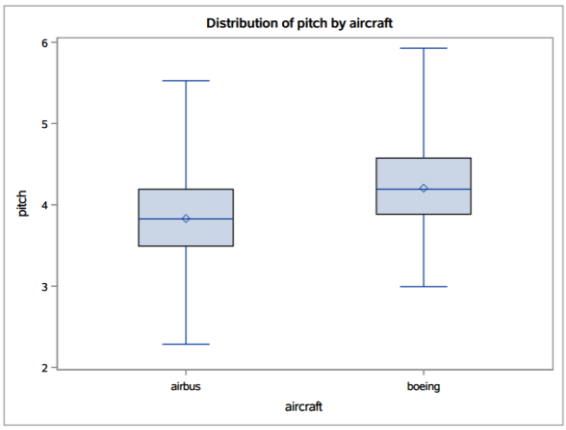


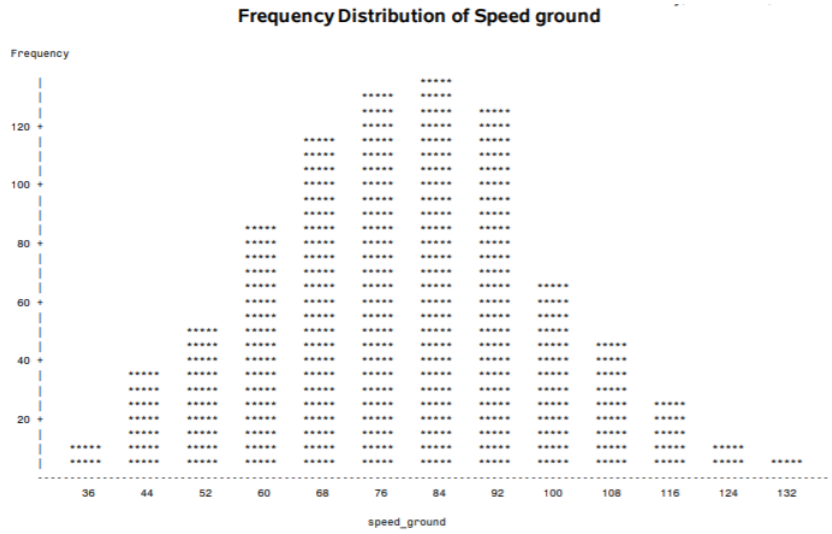



Distribution of height by aircraft
height
60
50
40
30
20
10
0
airbus
boeing
aircraft

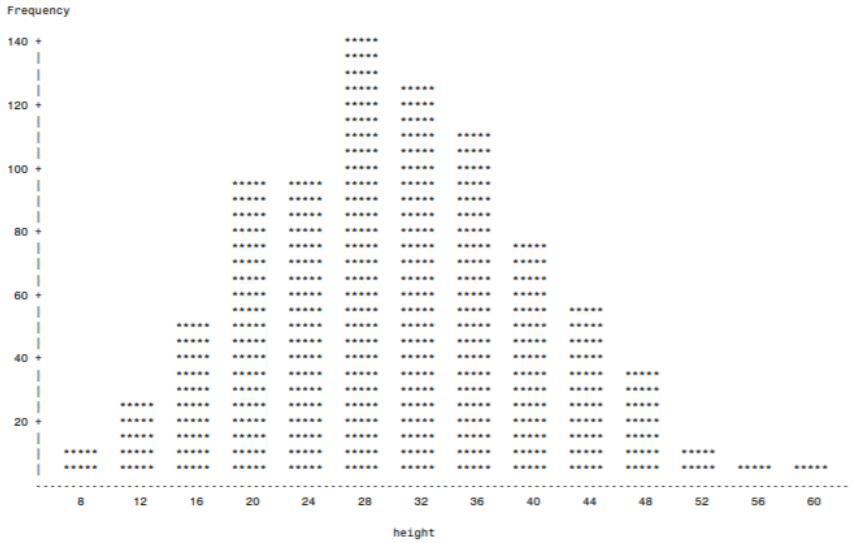

Distribution of distance by aircraft
distance
6000
5000
4000
3000
2000
1000
0
airbus
boeing
aircraft


Distribution of speed_ground by aircraft
speed_ground
150
125
100
75
50
25
airbus
boeing
aircraft


Distribution of pitch by aircraft
pitch
6
5
4
3
2
airbus
boeing
aircraft

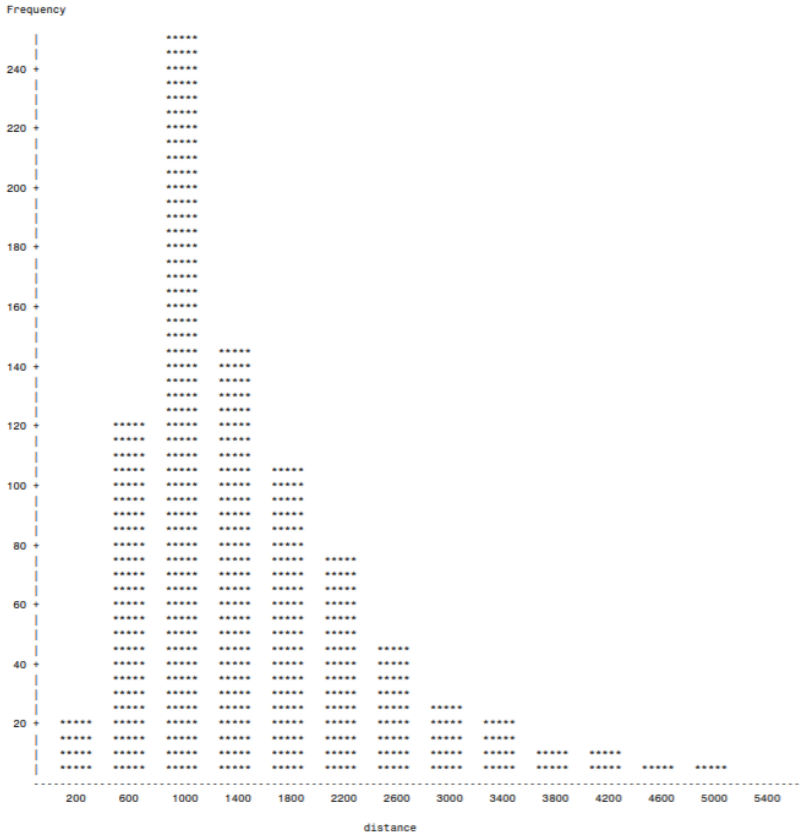

Frequency Distribution of Speed ground
Frequency
120
100
80
60
40
20
36
44
52
60
68
76
84
92
100
108
116
124
132
speed_ground

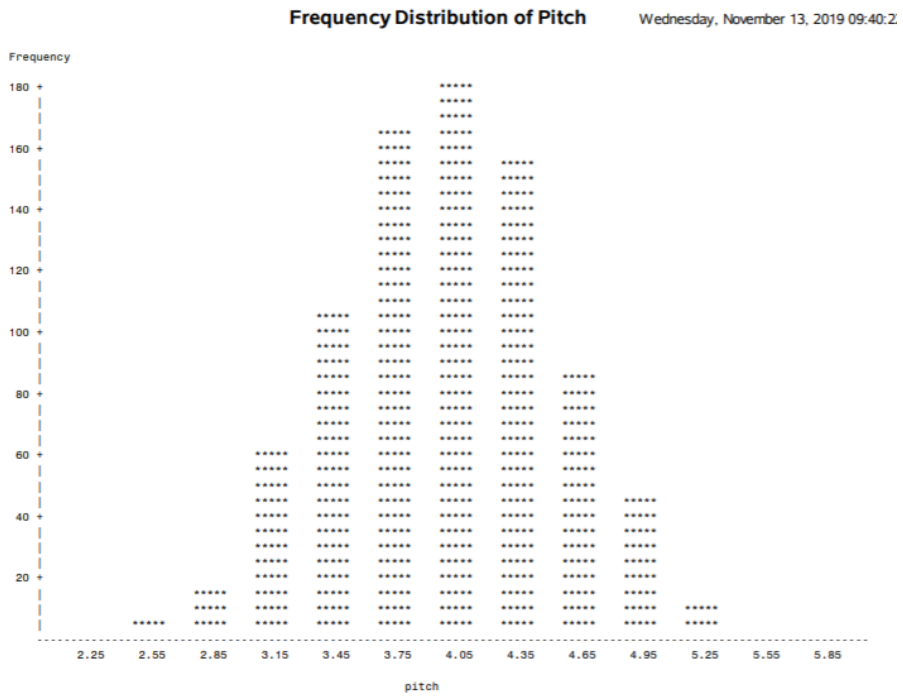
Frequency Distribution of Height



Wednesday, November 13, 2019 09:

Frequency Distribution of Distance





Summary of Chapter 1:

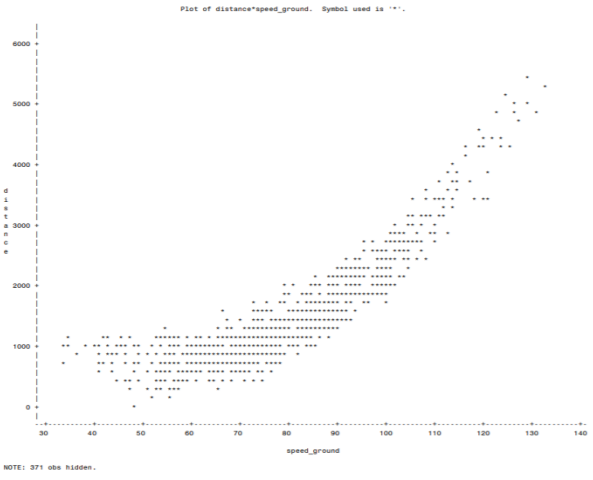
- The two datasets were combined to form a single consolidated dataset. The empty rows at the end were removed.
- The duplicates in both the tables were removed before calculating percent of missing values and number of abnormal values.
- Due to high percent of missing values in Speed_Air, the observations were not removed to avoid losing other useful information.
- After detecting abnormal values and removing them, our final dataset contained 831 rows and 8 columns.
- Finally, various summary tables and charts were used to summarize the distribution of each variable.

Chapter 2: Data Visualization and Descriptive Analysis

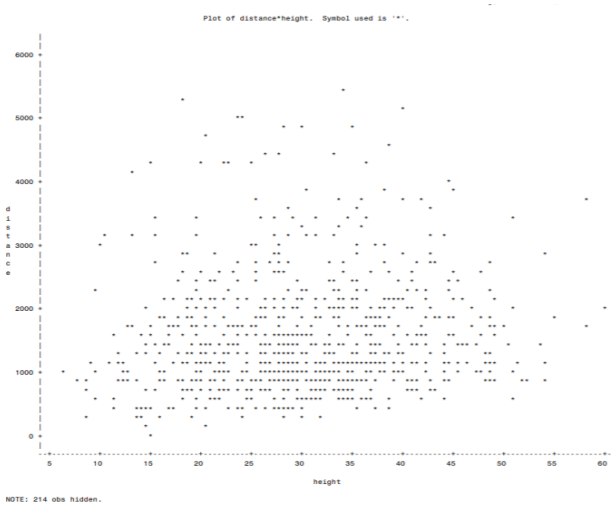
Now after cleaning the data and performing some basic summary analysis on each variable, it's time to perform some Exploratory Data Analysis to see how these predictors relate with our response variable (distance).

The following SAS code will generate various X-Y plots:

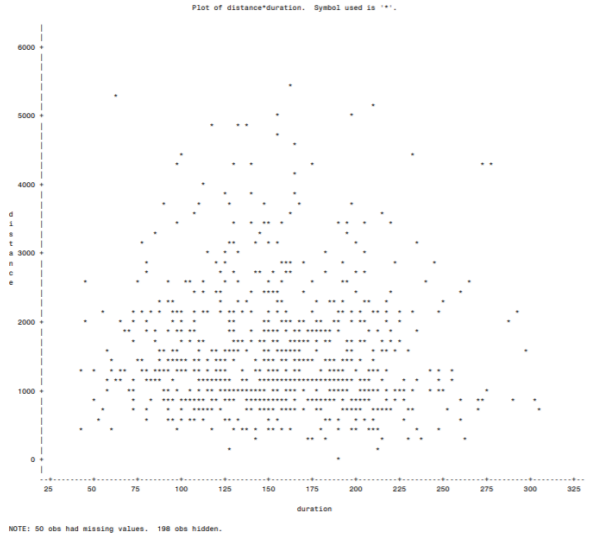
```
PROC PLOT DATA=FAA_CLEANED;
PLOT distance*Speed_ground='*';
PLOT distance*height='*';
PLOT distance*duration='*';
PLOT distance*Speed_air='*';
PLOT distance*no_pasg='*';
PLOT distance*pitch='*';
RUN;
```



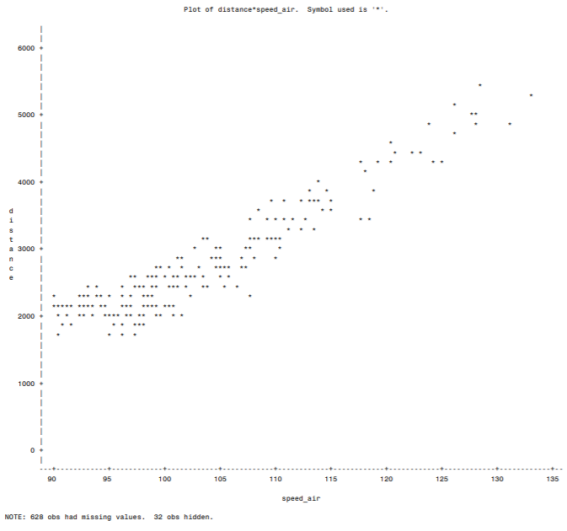
Speed Ground



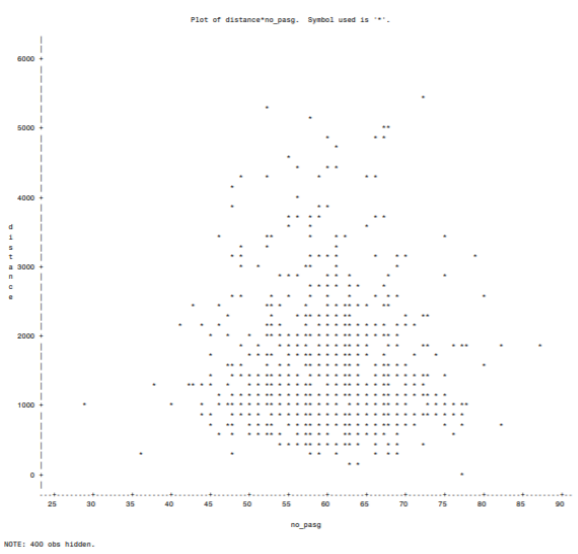
Height



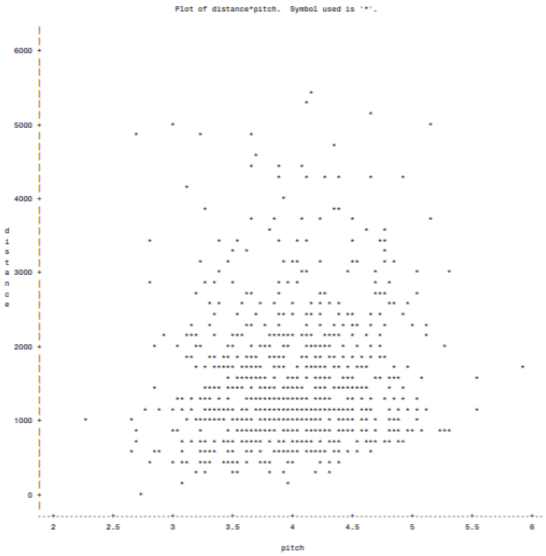
Duration



Speed Air



No of Passengers



Pitch

Looking at the above graphs, the relationship between predictor variable and some response variables like speed_air and speed_ground seem to be clear. Similar cannot be said about other predictor variables. In order to have more detailed understanding of these linear relationships, we need to compute correlation matrix.

Computing the Correlation Matrix:

In order to know which predictors are highly correlated with our response variable – distance, we need to compute the Correlation Matrix. We will build separate Correlation Matrix for each type of aircraft to see whether there is any difference between boeing and airbus.

```
PROC CORR DATA=FAA_CLEANED;
VAR Speed_Ground Speed_Air height pitch no_pasg duration distance;
TITLE Pairwise correlation coefficients;
BY aircraft;
RUN;
```

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

	speed_ground	speed_air	height	pitch	no_pasg	duration	distance
speed_ground	1.00000	0.98169	-0.03346	-0.00493	0.00906	-0.06055	0.90520
speed_ground		<.0001	0.4819	0.9176	0.8491	0.2305	<.0001
	444	85	444	444	444	394	444
speed_air	0.98169	1.00000	-0.00546	0.00007	-0.06372	0.01575	0.96411
speed_air	<.0001		0.9604	0.9995	0.5623	0.8918	<.0001
	85	85	85	85	85	77	85
height	-0.03346	-0.00546	1.00000	0.05128	0.02367	-0.01319	0.14494
height	0.4819	0.9604		0.2809	0.6189	0.7941	0.0022
	444	85	444	444	444	394	444
pitch	-0.00493	0.00007	0.05128	1.00000	-0.11802	-0.04402	0.07330
pitch	0.9176	0.9995	0.2809		0.0128	0.3835	0.1230
	444	85	444	444	444	394	444
no_pasg	0.00906	-0.06372	0.02367	-0.11802	1.00000	-0.02499	-0.00732
no_pasg	0.8491	0.5623	0.6189	0.0128		0.6209	0.8777
	444	85	444	444	444	394	444
duration	-0.06055	0.01575	-0.01319	-0.04402	-0.02499	1.00000	-0.07851
duration	0.2305	0.8918	0.7941	0.3835	0.6209		0.1198
	394	77	394	394	394	394	394
distance	0.90520	0.96411	0.14494	0.07330	-0.00732	-0.07851	1.00000
distance	<.0001	<.0001	0.0022	0.1230	0.8777	0.1198	
	444	85	444	444	444	394	444

Airbus

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

	speed_ground	speed_air	height	pitch	no_pasg	duration	distance
speed_ground	1.00000	0.99048	-0.08263	-0.04755	-0.01043	-0.04361	0.90050
speed_ground		<.0001	0.1046	0.3509	0.8379	0.3922	<.0001
	387	118	387	387	387	387	387
speed_air	0.99048	1.00000	-0.12922	-0.02499	0.02104	0.05264	0.97760
speed_air	<.0001		0.1631	0.7882	0.8211	0.5713	<.0001
	118	118	118	118	118	118	118
height	-0.08263	-0.12922	1.00000	0.00492	0.07297	0.03558	0.06920
height	0.1046	0.1631		0.9232	0.1519	0.4852	0.1743
	387	118	387	387	387	387	387
pitch	-0.04755	-0.02499	0.00492	1.00000	0.11215	-0.02132	-0.06504
pitch	0.3509	0.7882	0.9232		0.0274	0.6759	0.2017
	387	118	387	387	387	387	387
no_pasg	-0.01043	0.02104	0.07297	0.11215	1.00000	-0.05091	-0.01785
no_pasg	0.8379	0.8211	0.1519	0.0274		0.3178	0.7262
	387	118	387	387	387	387	387
duration	-0.04361	0.05264	0.03558	-0.02132	-0.05091	1.00000	-0.01064
duration	0.3922	0.5713	0.4852	0.6759	0.3178		0.8347
	387	118	387	387	387	387	387
distance	0.90050	0.97760	0.06920	-0.06504	-0.01785	-0.01064	1.00000
distance	<.0001	<.0001	0.1743	0.2017	0.7262	0.8347	
	387	118	387	387	387	387	387

Boeing

Summary of Chapter 2:

In both aircraft types, the correlation of distance is high with speed_air and speed_ground and also with height although not as high as the other two. The correlation of height is however greater in 'Airbus' as compared to 'Boeing'.

A very interesting observation from the above table is that two predictor variables speed_air and speed_ground are highly correlated among themselves in each of the case which may explain why both of them are also correlated equally with distance.

Chapter 3: Statistical Modeling

After performing the above steps, we will try to fit the model that best explains our data.

Generalized form of linear regression is:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Where

β_0 = Intercept

β_1 = Coefficient of Speed_Air, X_1 = Speed_Air

β_2 = Coefficient of Height, X_2= Height

We will build the model separately for each aircraft type to see if there is any difference between the two coefficient values.

```
PROC REG DATA=FAA_CLEANED;
MODEL distance=Speed_Air Height;
BY aircraft;
TITLE Regression analysis of the Flights data set;
RUN;
```

aircraft=airbus

Number of Observations Read	444
Number of Observations Used	85
Number of Observations with Missing Values	359

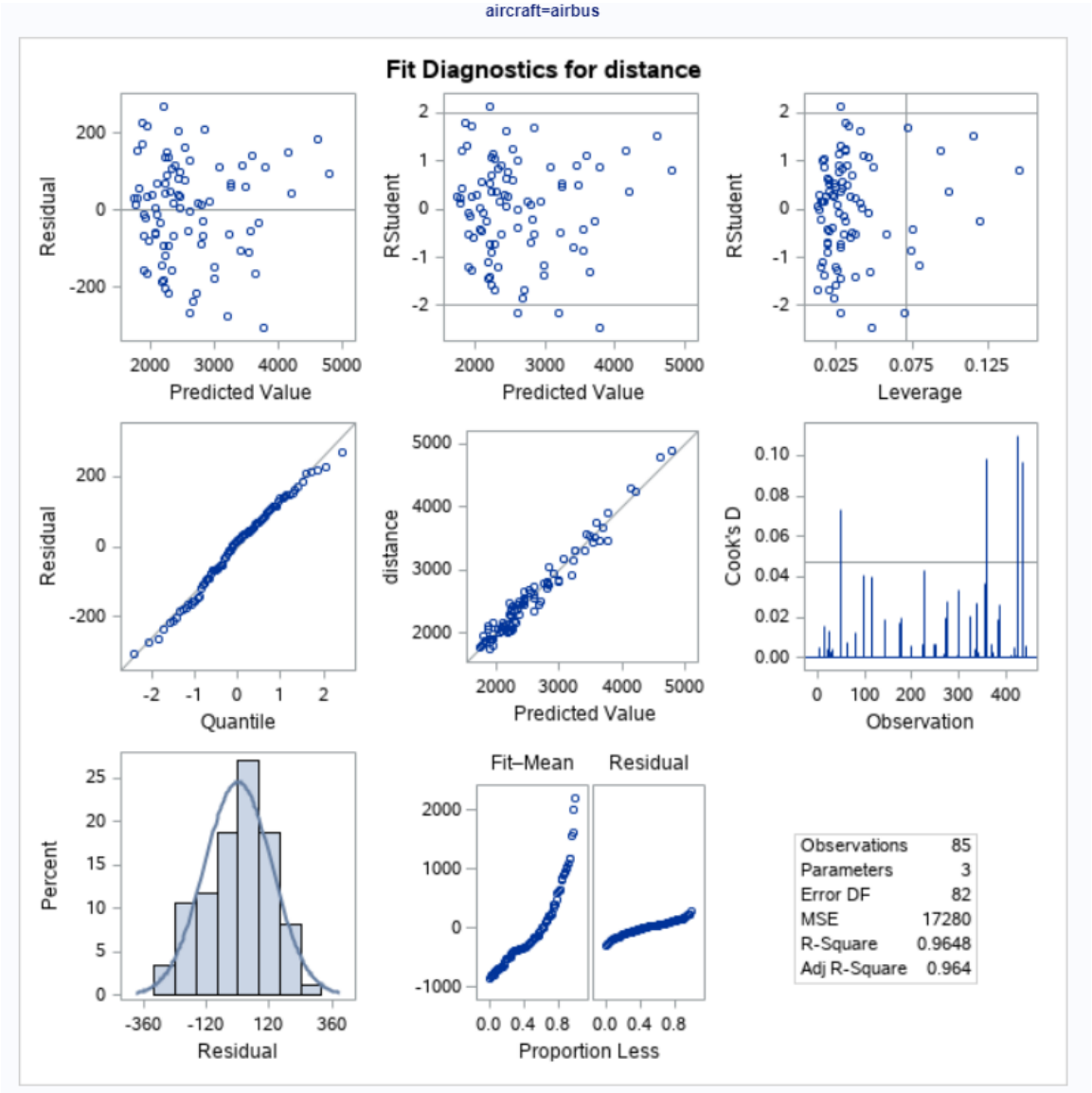
Analysis of Variance

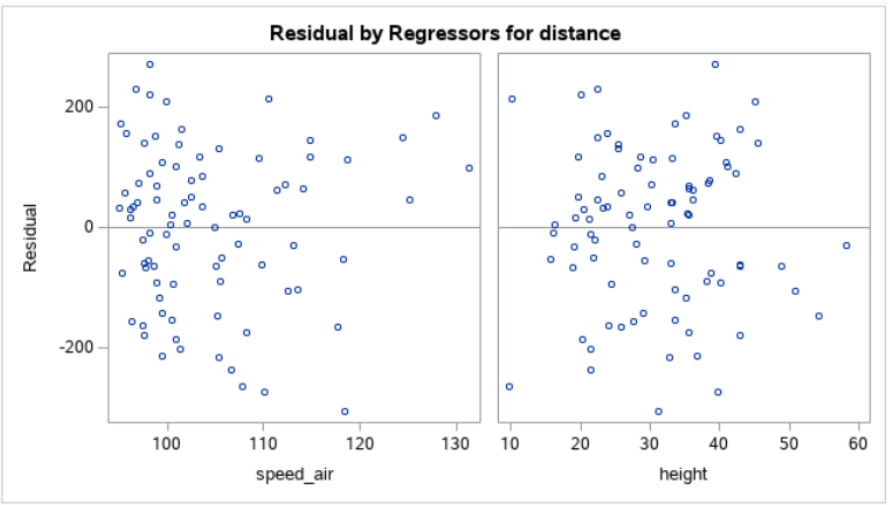
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	38848283	19424141	1124.06	<.0001
Error	82	1416988	17280		
Corrected Total	84	40265270			

Root MSE	131.45469	R-Square	0.9648
Dependent Mean	2600.86007	Adj R-Sq	0.9640
Coeff Var	5.05428		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-6425.58744	191.16956	-33.61	<.0001
speed_air	speed_air	1	82.60127	1.77303	46.59	<.0001
height	height	1	13.31185	1.46766	9.07	<.0001


aircraft=airbus
Fit Diagnostics for distance
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
distance
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 85
Parameters 3
Error DF 82
MSE 17280
R-Square 0.9648
Adj R-Square 0.964


Residual by Regressors for distance
Residual
speed_air
height

For Boeing:

aircraft=boeing

Number of Observations Read	387
Number of Observations Used	118
Number of Observations with Missing Values	269

Analysis of Variance

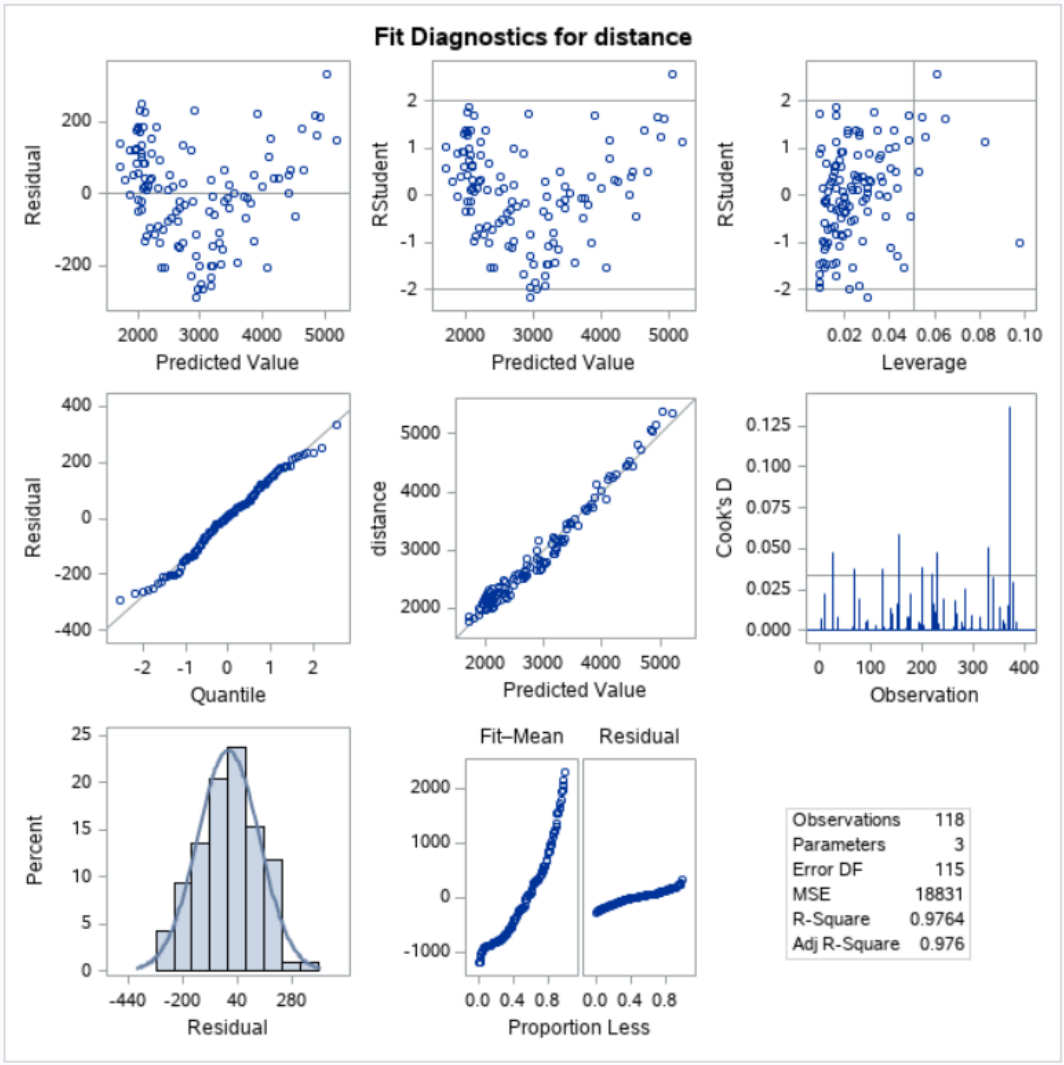
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	89619398	44809699	2379.54	<.0001
Error	115	2165596	18831		
Corrected Total	117	91784994			

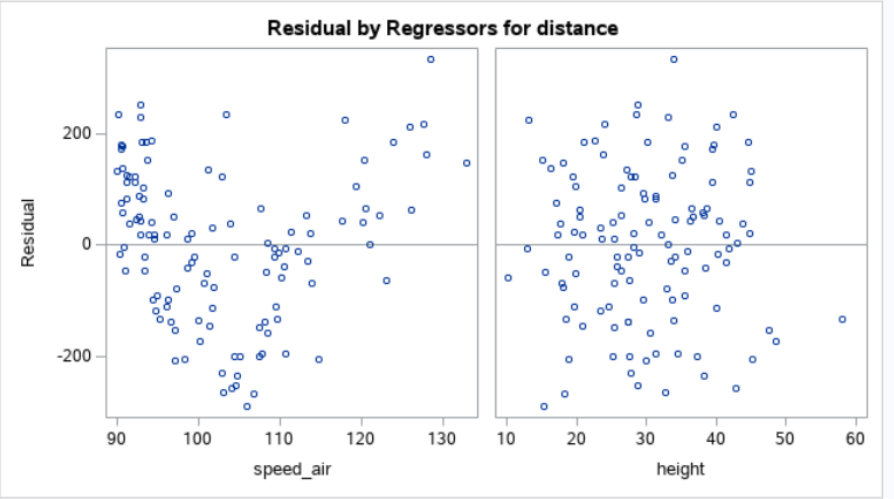
Root MSE	137.22708	R-Square	0.9764
Dependent Mean	2899.87705	Adj R-Sq	0.9760
Coeff Var	4.73217		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5956.13890	134.83861	-44.17	<.0001
speed_air	speed_air	1	81.99642	1.18876	68.98	<.0001
height	height	1	13.99719	1.39327	10.05	<.0001

aircraft=boeing





In both of the aircraft types, the coefficients of speed_air and height are coming nearly same along with very small p-values which suggest that $\beta_1 = \beta_2 \neq 0$. Also, the residuals seem to follow constant variance with some outliers in the predictor and response variable. The R-Square and Adj-R-Sq are coming close to 1 which is an indication that our model is good.

Since there is not much difference between our two models, we can combine them to make model for our combined dataset.

```
PROC REG DATA=FAA_CLEANED;
MODEL distance=Speed_Air Height;
TITLE Regression analysis of the Flights data set;
RUN;
```

Number of Observations Read	831
Number of Observations Used	203
Number of Observations with Missing Values	628

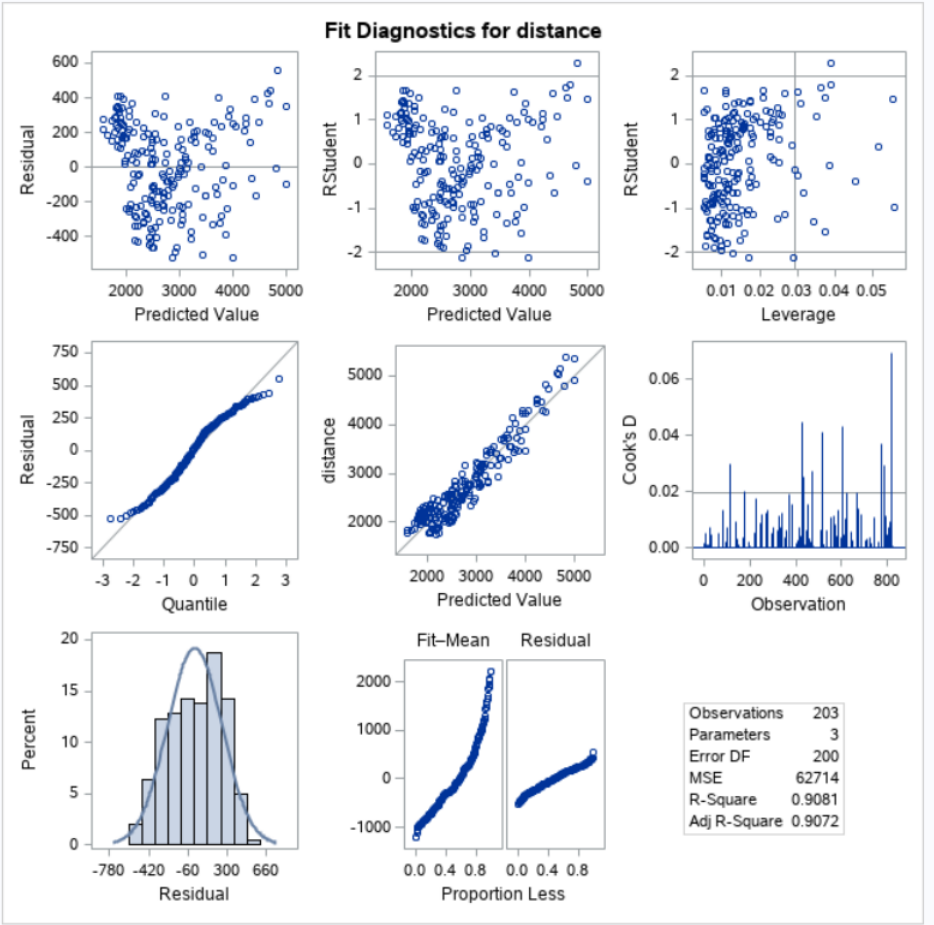
Root MSE	250.42817	R-Square	0.9081
Dependent Mean	2774.67289	Adj R-Sq	0.9072
Coeff Var	9.02550		

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			

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



Summary of Chapter 3:

The final model equation (**Model 1**) that was built using above variables was:

$Y = -5935.73 + 80.49*X_1 + 12.54*X_2$, Where X_1=Speed_Air and X_2=Height.

Speed_Air was used instead of speed_ground because it was performing relatively better in terms of R-Square and Residual plots. However, it comes with its own problem that there are lot of missing values in speed_air.

Another model (**Model 2**) was built which used Speed_ground as X_1 instead of Speed_air. The equation for that model comes out to be:

$Y = -2224.76 + 41.85*X_1 + 13.72*X_2$, Where X_1=Speed_Ground and X_2=Height.

The R-Square for the above equation comes out to be 0.7727 which is comparatively small as compared to the previous one.

Short Answers to these Questions

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

   831 observations were used to fit my final model. All 950 observations were not used because there were some duplicate values and abnormal values which needed to be removed because they would affect our final model.

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

   The major factors that look like impacting the landing distance of the flight were coming out to be – speed_air, speed_ground and height from the correlation matrix.

   However, since speed_air and speed_ground was highly correlated among themselves I decided to go with speed_air alone. Also, there was not any great amount of difference between two models built for each aircraft type so it was not a considerable factor in my model.

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

   There were some differences between the distributions of variables like distance and pitch for the two makes as seen in the Box plots and summary tables. In Correlation Matrix, the same variables were showing large values in both cases although there was some difference in the height variable. However when the variables were used in my model, there is no significant difference between the estimated coefficients.