Flight Landing - Project

- Vaidiyanathan Lalgudi Venkatesan

Executive Summary

This report provides an analysis on the factors impacting the landing distance of a commercial flight and predict the same to reduce the risk of landing overrun. Methods of analysis include linear regression and hypothesis testing. The data used for analysis is landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). The We initially combine and clean the datasets, followed by exploratory data analysis. Results of data analyzed are used to answer the following questions:

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

   We use 832 observations to fit the model. This is because, the initial data consisted of exact duplicates and few abnormal variable values. We remove them during data preparation.

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

   Taking the FAA dataset, factors impacting landing distance are found as speed_ground, speed_air, height and pitch. Since speed_ground and speed_air are observed to have strong positive correlation, including both in our regression analysis will affect the variance inflation factor (VIF), which quantifies the severity of multicollinearity in an ordinary least squares regression analysis. Hence, we use only of one the variables (speed_ground in this case) for our analysis. Speed_ground is used instead of speed_air, since it covers more observations (speed_ground-832, speed_air-203) to fit the model. We develop a linear model giving the relationship between each variable and landing distance, mathematically written as follows:

   Distance ≈ -3037.954 + 42.056(Speed_ground) + 13.491(Height) + 200.859(Pitch)

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

   Yes! We find some differences between the two aircraft makes. It is observed that the height variable impacts the landing distance in airbus but not in Boeing aircraft. Also, it is seen that, on analyzing based on individual aircraft types, pitch does not impact the landing distance by much. The mathematical linear models for Boeing and Airbus aircraft types are summarized as below:

   Airbus:

   Distance ≈ -2522.891 + 42.554(Speed_ground) + 14.097(Height)

   Boeing:

   Distance ≈ -2522.455 + 41.739 (Speed_ground)

*All analysis is performed in 95% confidence interval

Data Preparation:

OBJECTIVE:

To prepare and clean the FAA data for further analysis and modelling. Following are the steps in this chapter, "Data Preparation":

1. To Combine data sets from different sources;
2. Perform the completeness check of each variable – examine if missing values are present;
3. Perform the validity check of each variable – examine if abnormal values are present;
4. Clean the data based on the results of Steps 2 and 3;
5. Summarize the distribution of each variable;

CODE, OUTPUT AND OBSERVATIONS:

Note: Dataset names are specified in italics in observation section.

1. **Combine Data sets:**

CODE:

```
/*1*/
/*combine datasets faa1 and faa2*/

FILENAME REFFILE '/folders/myfolders/SC/flight_data/FAA1.xls';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLS
	OUT=SC.flight_data_faa1;
	GETNAMES=YES;
RUN;

FILENAME REFFILE '/folders/myfolders/SC/flight_data/FAA2.xls';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLS
	OUT=SC.flight_data_faa2;
	GETNAMES=YES;
RUN;

data flight_data_combined;
set sc.flight_data_faa1 sc.flight_data_faa2;
run;

/*remove empty observations*/
data want;
set flight_data_combined;
if compress(cats(of _all_),'.')=' ' then delete;
run;
/*print first 15 observations of the total flight data*/
proc print data=flight_data_combined(obs=15);
run;
```

OUTPUT:

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
6	boeing	137.59581722	55	75.014343744	.	41.21496259	4.203853398	1627.0681991
7	boeing	73.023794916	54	54.4298029	.	24.03532163	3.8376457299	805.30399317
8	boeing	52.903187872	57	57.101661737	.	19.388837508	4.6436717769	573.62178606
9	boeing	155.51861605	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548
10	boeing	176.86203205	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579
11	boeing	158.4618984	61	53.778126741	.	46.355832902	5.5563991716	1075.3717411
12	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506
13	boeing	72.289633216	54	93.391762435	92.869561214	32.223489271	3.8182761471	2128.708285
14	boeing	187.59954737	58	94.036412942	96.196460585	33.661226156	4.6361847249	2304.857574
15	boeing	154.36870049	63	63.540613553	.	26.402991875	3.8566584986	1089.9729531

Figure 1.1: First 15 observations of the combined flight data

CODE BRIEF AND OBSERVATIONS:

The datasets FAA1 and FAA2 have been concatenated using **set** operator into a dataset called *flight_data_combined*. We then create another dataset *want* removing empty observations from *flight_data_combined.* Then we print the first 15 observations of *want*. (Figure 1.1)

---

2. Perform the completeness check of each variable – examine if missing values are present;

CODE:

```
/*2*/
/*Completeness check of each variable - examine if missing values are present*/
proc means data=want n nmiss;
run;
```

OUTPUT:

Variable	Label	N	N Miss
duration	duration	800	150
no_pasg	no_pasg	950	0
speed_ground	speed_ground	950	0
speed_air	speed_air	239	711
height	height	950	0
pitch	pitch	950	0
distance	distance	950	0

Figure 1.2: Number of missing values in each variable

CODE BRIEF AND OBSERVATIONS:

Here, we use **proc means** to find the number of observations and missing values in the combined dataset *want* (Figure1.2). It is observed that the variable contains 150 missing values and speed_air contains 711 missing values out of 950 observations. We will discuss ways to handle these missing values in upcoming chapters.

---

3. Perform the validity check of each variable – examine if abnormal values are present;

CODE:

```
/*3*//*observations with abnormal durations*/
data f_duration_abnormal;
set want;
where duration < 40;
run;
proc means data=f_duration_abnormal n ;
var duration;
run;

/*observations with abnormal SPEED_GROUND*/
data f_speed_ground_abnormal;
set want;
where speed_ground < 30 or speed_ground > 140;
run;
proc means data=f_speed_ground_abnormal n;
var speed_ground;
run;

/*observations with abnormal SPEED_air*/
data f_speed_air_abnormal;
set want;
where speed_air < 30 or speed_air > 140;
run;
proc means data=f_speed_air_abnormal n;
var speed_air;
run;

/*observations with abnormal height*/
data f_height_abnormal;
set want;
where height < 6;
run;
proc means data=f_height_abnormal n ;
var height;
run;

/*observations with abnormal distance*/
data f_distance_abnormal;
set want;
where distance >= 6000;
run;
proc means data=f_distance_abnormal n ;
var distance;
run;
```

OUTPUT:

Analysis Variable : duration duration
N
5

Analysis Variable : speed_ground speed_ground
N
5

Analysis Variable : speed_air speed_air
N
2

Analysis Variable : height height
N
12

Analysis Variable : distance distance
N
3

Figure 1.3: Abnormal values in each variable

CODE BRIEF AND OBSERVATIONS:

We perform the validity check of each variable in the dataset by looking for abnormal values. For example, we have a condition that the flight duration is usually more than 40 minutes. Hence, we filter for observations where the flight duration is less than 40 minutes. The same applies to other variables, such as ground and air speed must lie with 30-140mph, aircraft height should be greater than 6 meters, and the landing distance of the aircraft should be less than 6000 feet. We have found the number of abnormal values for each variable using **where** statement and **proc means**, which are shown in Figure 1.3.

---

4. CLEAN DATA SETS:

CODE:

```
/*4*/
/* LABELING ABNORMAL DURATIONS*/

data f_labeled;
set want;
if (duration ^=. and duration < 40) or (speed_ground ^=. and (speed_ground > 140 or
speed_ground < 30)) or (speed_air ^=. and (speed_air >140 or speed_air < 30)) or
(height < 6 and height ^= .) or (distance >=6000 and distance ^= .) then landing =0;
else landing =1;
run;
proc print data=f_labeled(obs=15);
run;
proc means data=f_labeled n nmiss;
```

```
run;

/*finding abnormal values*/
data f_c;
set f_labeled;
where landing = 0;
run;
proc means data=f_c n nmiss;
run;

/*removing obs with abnormal values, since it is relatively small to the data*/
data f_c;
set f_labeled;
where landing = 1;
run;
proc means data=f_c n nmiss;
run;

/*removing exact duplicates*/
proc sort data=f_c out=f_clean nodupkey;
by aircraft no_pasg speed_ground speed_air height pitch distance;
run;
proc means data=f_clean n nmiss;
run;
```

OUTPUT:

/* LABELING ABNORMAL DURATIONS*/

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	landing
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638	1
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235	1
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286	1144.922426	1
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584	1
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976	1
6	boeing	137.59581722	55	75.014343744	.	41.21496259	4.203853398	1627.0681991	1
7	boeing	73.023794916	54	54.4298029	.	24.03532163	3.8376457299	805.30399317	1
8	boeing	52.903187872	57	57.101661737	.	19.388837508	4.6436717769	573.62178606	1
9	boeing	155.51861605	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548	1
10	boeing	176.86203205	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579	1
11	boeing	158.4618984	61	53.778126741	.	46.355832902	5.5563991716	1075.3717411	1
12	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506	0
13	boeing	72.289633216	54	93.391762435	92.869561214	32.223489271	3.8182761471	2128.708285	1
14	boeing	187.59954737	58	94.036412942	96.196460585	33.661226156	4.6361847249	2304.857574	1
15	boeing	154.36870049	63	63.540613553	.	26.402991875	3.8566584986	1089.9729531	1

Figure 1.4(a): First 15 observations of *f_labeled* dataset. Landing=0 corresponds to abnormal value

/* proc means of f_labeled dataset*/

Variable	Label	N	N Miss
duration	duration	800	150
no_pasg	no_pasg	950	0
speed_ground	speed_ground	950	0
speed_air	speed_air	239	711
height	height	950	0
pitch	pitch	950	0
distance	distance	950	0
landing		950	0

Figure 1.4(b): no. of obs and missing values in *f_labeled* dataset

```
/*finding abnormal values*/
```

Variable	Label	N	N Miss
duration	duration	19	4
no_pasg	no_pasg	23	0
speed_ground	speed_ground	23	0
speed_air	speed_air	6	17
height	height	23	0
pitch	pitch	23	0
distance	distance	23	0
landing		23	0

Figure 1.4(c): no. of abnormal observations in *f_labeled* dataset

```
/*removing obs with abnormal values, since it is relatively small to the data*/
```

Variable	Label	N	N Miss
duration	duration	781	146
no_pasg	no_pasg	927	0
speed_ground	speed_ground	927	0
speed_air	speed_air	233	694
height	height	927	0
pitch	pitch	927	0
distance	distance	927	0
landing		927	0

Figure 1.4(d): *f_c* dataset with abnormal observations removed

```
/*removing exact duplicates*/
```

Variable	Label	N	N Miss
duration	duration	781	51
no_pasg	no_pasg	832	0
speed_ground	speed_ground	832	0
speed_air	speed_air	203	629
height	height	832	0
pitch	pitch	832	0
distance	distance	832	0
landing		832	0

Figure 1.4(e): *f_clean* dataset with exact duplicates removed

CODE BRIEF AND OBSERVATIONS:

The abnormal values are first labelled to analyse and handle them. The **if** statement specifies the condition for abnormality and this is labelled using a new variable called landing. The landing variable takes a value of **0** if the variable has an abnormal value, and **1** otherwise. This is stored in a new dataset *f_labeled.* (Figure 1.4(a)).

The total count of observations with one or more abnormal variable is then found and stored in dataset *f_c* (Figure 1.4(c)). It is observed that: (Observations having abnormal values << Total observations). Hence, we delete the abnormal observations and store again in *f_c.* (Figure 1.4(d))

Finally, we remove the exact duplicates in the dataset using **proc sort, nodupkey** and store in a new dataset *f_clean.* As shown in Figure 1.4(e), from an initial 950 observations, after removing abnormal and exact duplicate observations, our final dataset consists of 832 observations.

5. **Summarize the distribution of each variable;**

CODE:

```
/*5*/
/*summarizing distribution of each variable*/

ods graphics / imagemap=on;

proc univariate data=WORK.f_clean;
	ods select Histogram;
	var duration no_pasg speed_ground speed_air height pitch distance ;
	histogram duration no_pasg speed_ground speed_air height pitch distance
		/ normal;
	inset n mean median std max min q3 q1/ position=ne;
run;

/*summary statistics for clean data*/
proc means data=f_clean n nmiss mean median std min max q3 q1;
run;
```

OUTPUT:

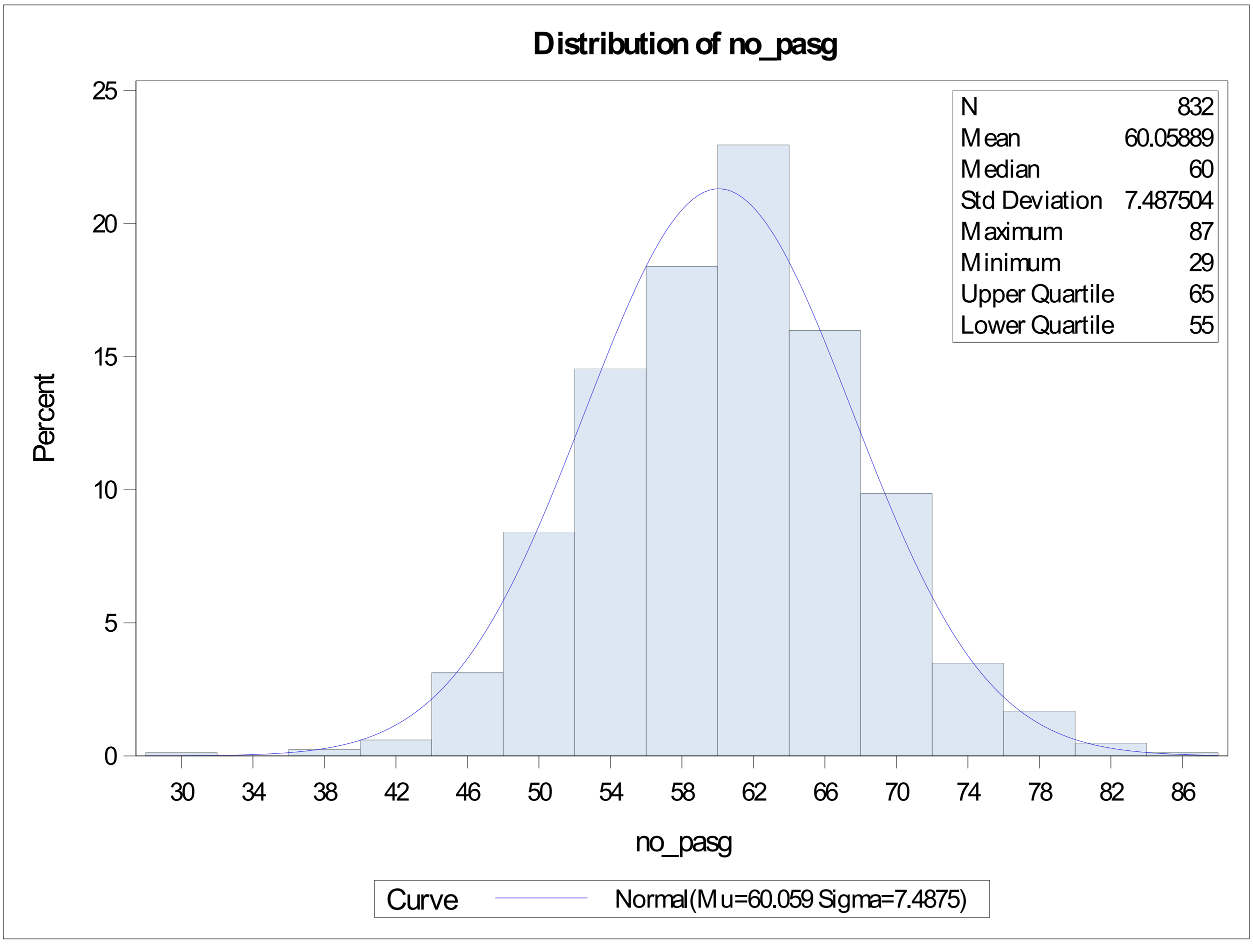


Figure 1.5(a): Distribution of the number of passengers in a flight

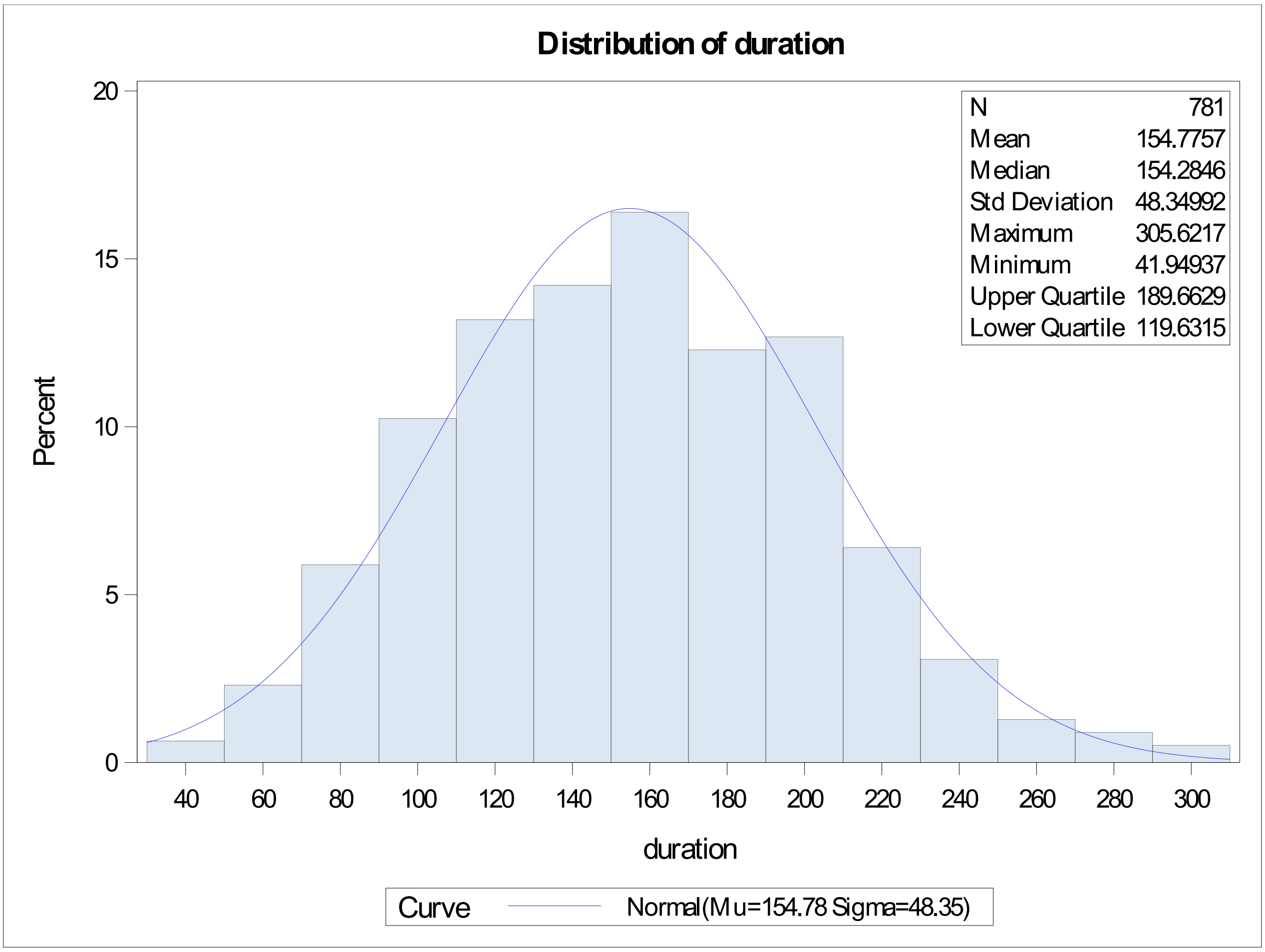


Figure 1.5(b): Distribution of the duration of a flight

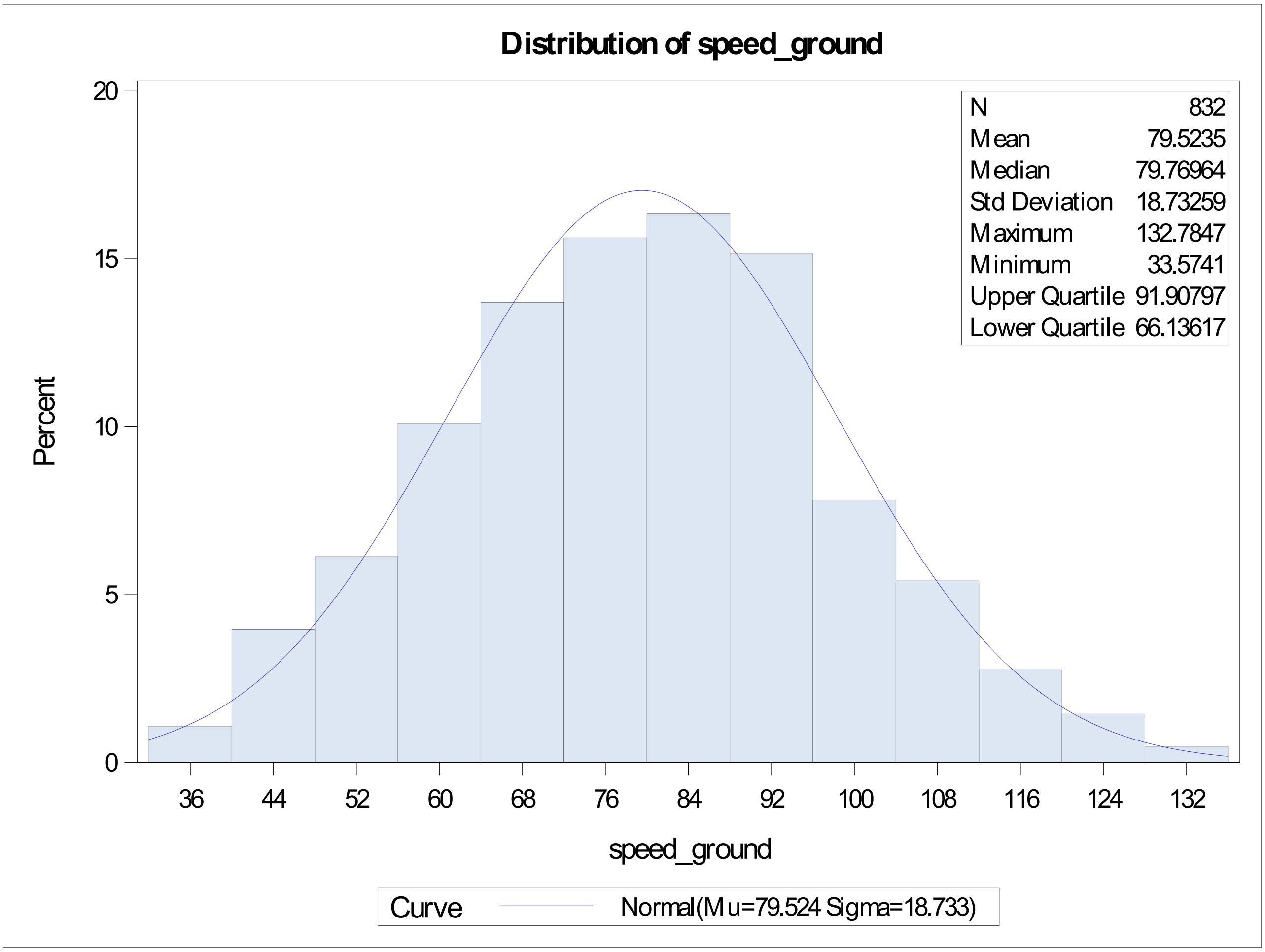


Figure 1.5(c): Distribution of the ground speed

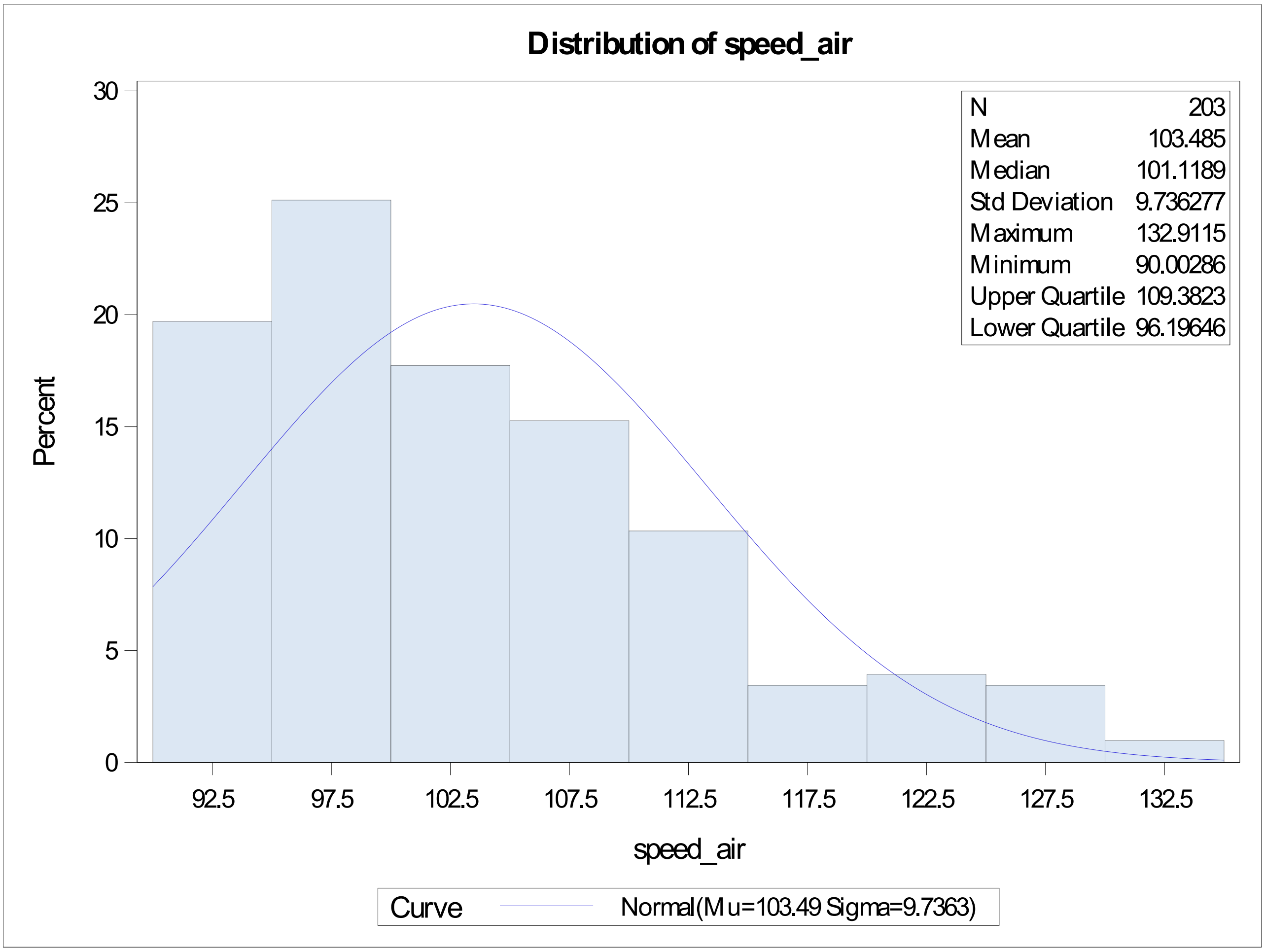


Figure 1.5(d): Distribution of the air speed

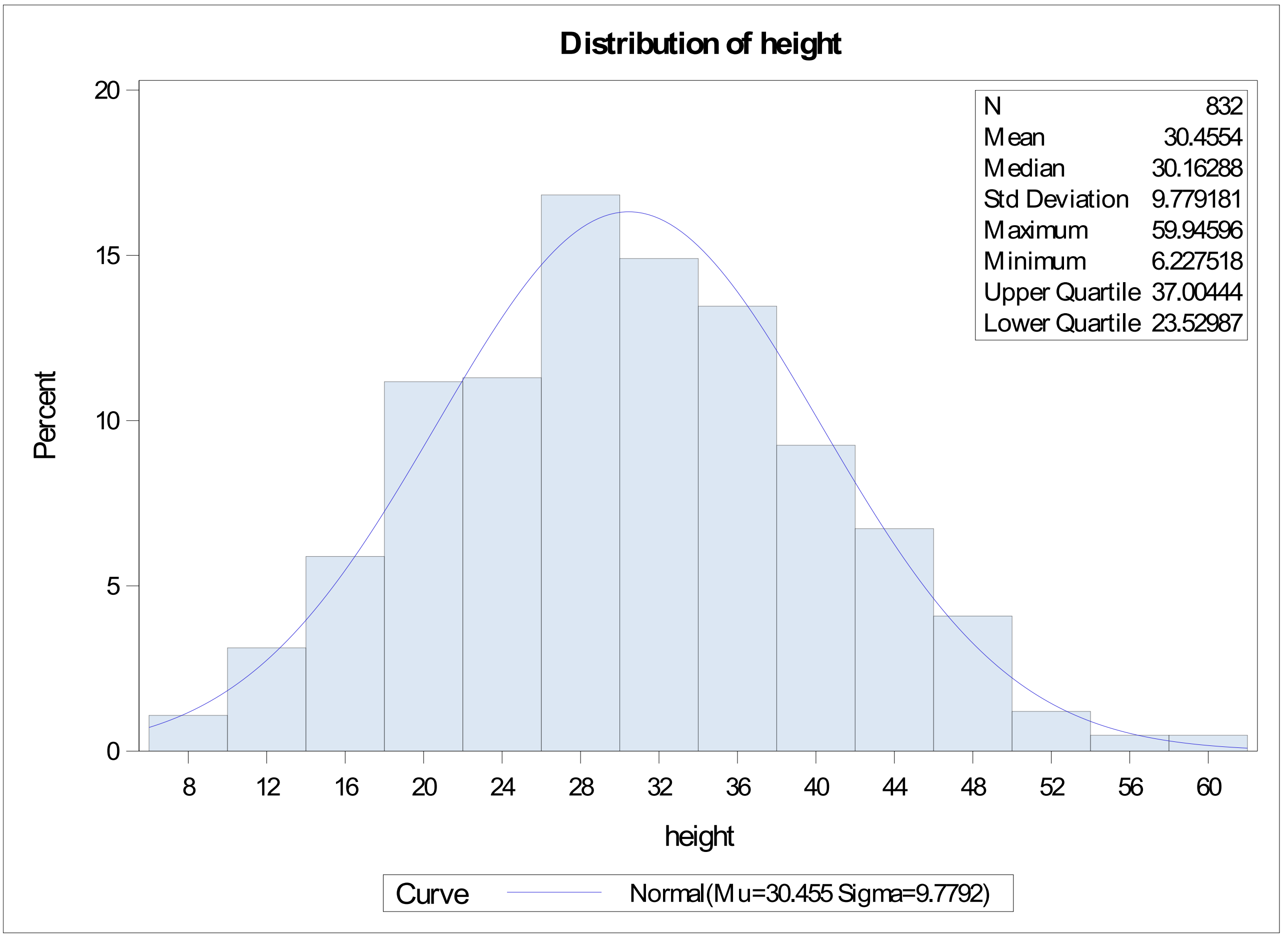


Figure 1.5(e): Distribution of height of the flight

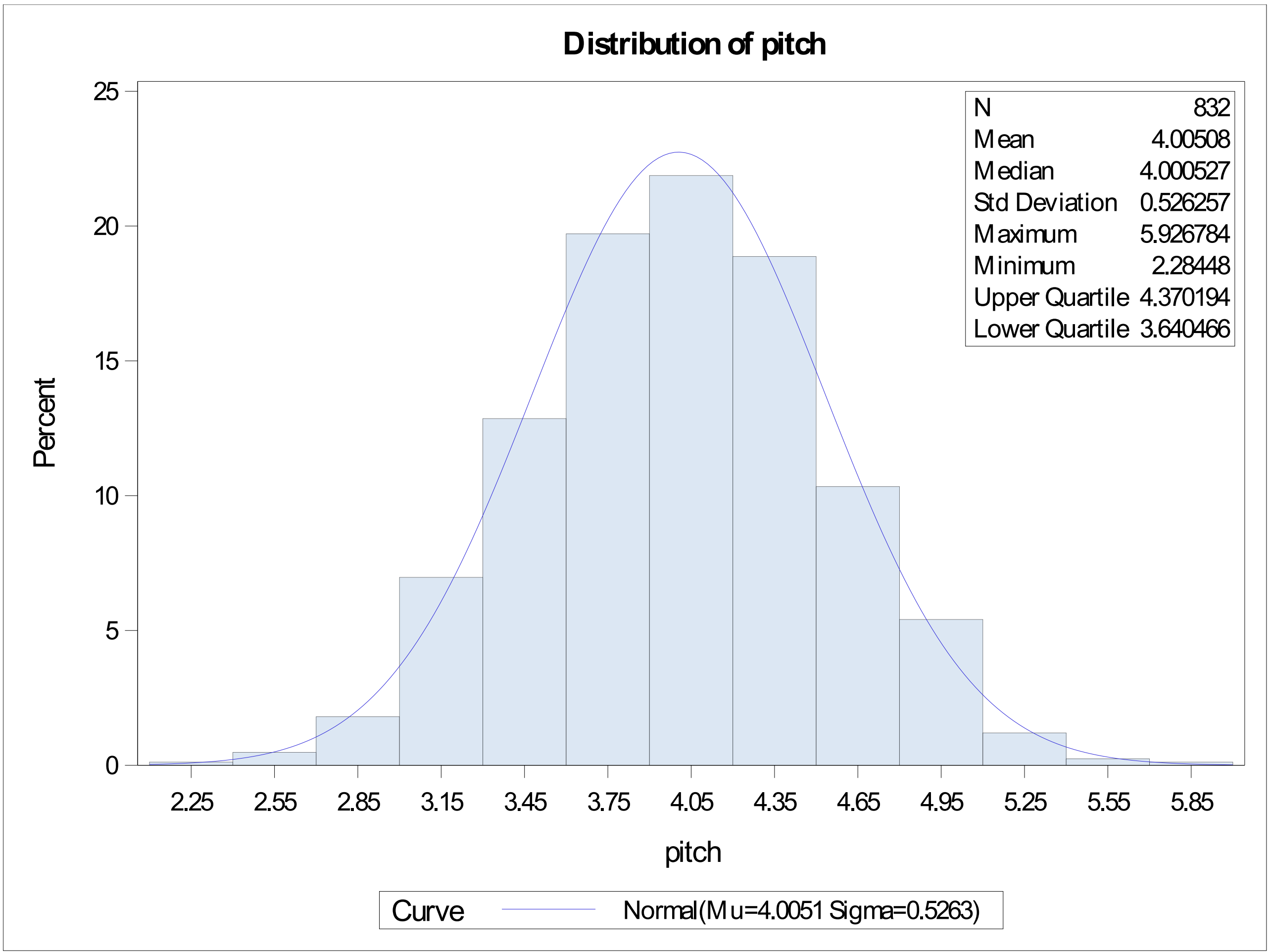


Figure 1.5(f): Distribution of the pitch

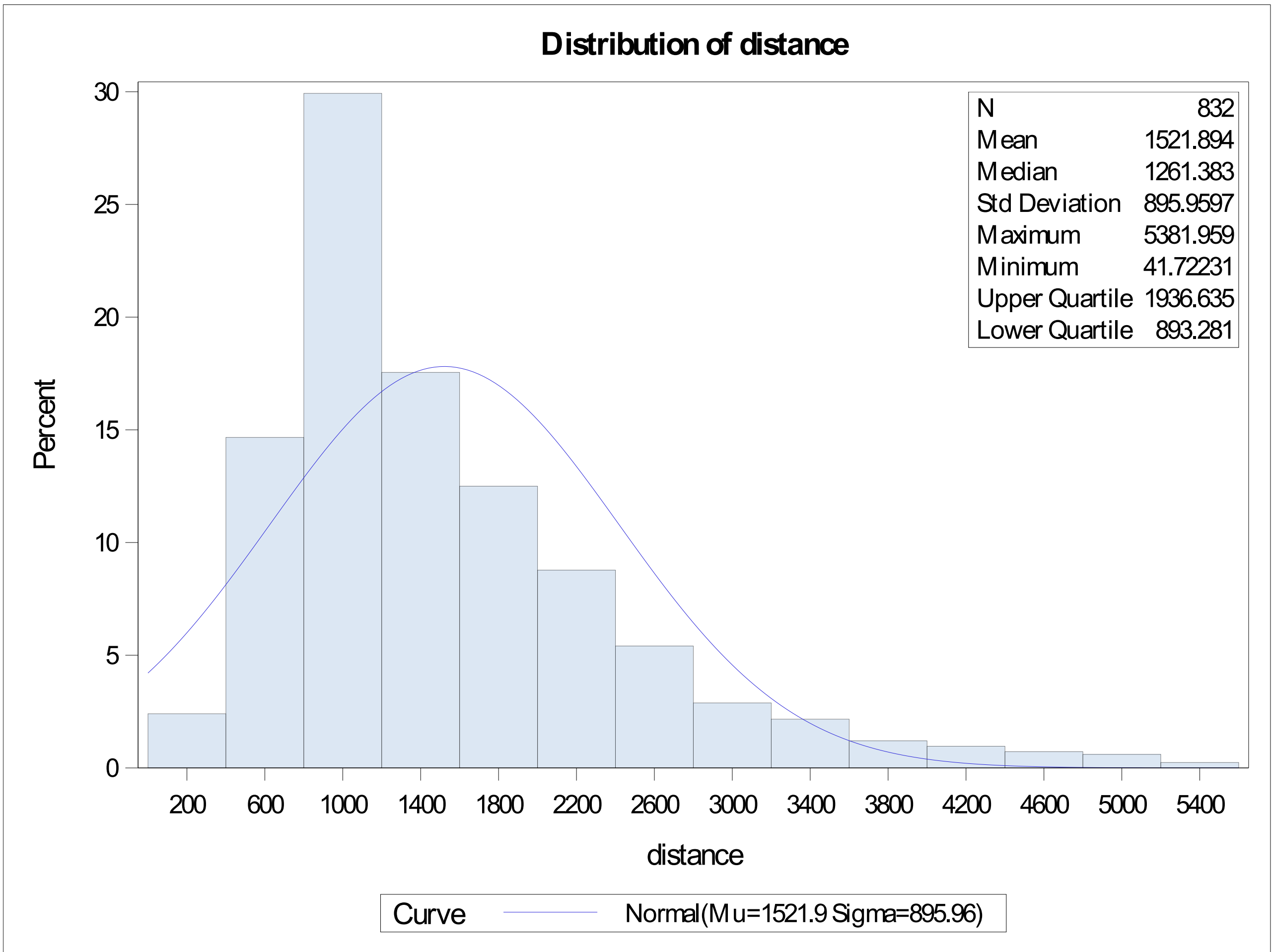


Figure 1.5(g): Distribution of the landing distance

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Minimum	Maximum	Upper Quartile	Lower Quartile
duration	duration	781	51	154.7757191	154.2845505	48.3499237	41.9493694	305.6217107	189.6629425	119.6314577
no_pasg	no_pasg	832	0	60.0588942	60.0000000	7.4875038	29.0000000	87.0000000	65.0000000	55.0000000
speed_ground	speed_ground	832	0	79.5235023	79.7696381	18.7325852	33.5741041	132.7846766	91.9079704	66.1361658
speed_air	speed_air	203	629	103.4850352	101.1189240	9.7362774	90.0028586	132.9114649	109.3823005	96.1964606
height	height	832	0	30.4554041	30.1628822	9.7791808	6.2275178	59.9459639	37.0044409	23.5298692
pitch	pitch	832	0	4.0050800	4.0005270	0.5262573	2.2844801	5.9267842	4.3701941	3.6404662
distance	distance	832	0	1521.89	1261.38	895.9597497	41.7223127	5381.96	1936.63	893.2809642
landing	landing	832	0	1.0000000	1.0000000	0	1.0000000	1.0000000	1.0000000	1.0000000

Figure 1.5(h): Summary statistics after data preparation

CODE BRIEF AND OBSERVATIONS:

We perform a univariate analysis and plot the distribution of each variable in the flight data using Histogram. The various inset statistics of each variable is shown in figures 1.5(a)-1.5(g). Also, we check the summary statistics of the clean data (Figure 1.5h), which shows that we have a final 832 observations from an initial 950, after the data preparation step. These 832 will be used for performing statistical analysis, which is explained in the next chapter.

CONCLUSION:

The flight datasets FAA1 and FAA2 have been combined, checked for missing values and abnormal values. Data cleaning, such as removing abnormal values and exact duplicates was performed, and the distribution of each variable plotted.

QUESTIONS IN DATA PREPARATION:

- Knowing how recent the data is very important in analysing the same. Very old data cannot be used to predict outcomes, as the flights might have changed, pilots might have changes and many other reasons
- Need to know if the given data is from the same airport, as data from different airports will make it difficult to predict accurately the risk of landing overrun, with factors such as weather, geographical location etc. being unknown factors
- Having flight timings will also help in developing actionable insights
- We also have some unknown factors such as the experience of pilots, and some irreducible errors such as the effect of weather on any given day (even if it is the same airport).

Exploratory Data Analysis:

OBJECTIVE:

To perform statistical analysis on the clean data and to study what factors and how they would impact the landing distance of a commercial flight. We perform the following steps:

1. Do the plots, to study the relationship between the dependent and independent variables;
2. Calculate the correlation between variables;
3. Do the regression analysis;
4. Model checking;

CODE, OUTPUT AND OBSERVATIONS:

1. **Plot Y (Dependent variable) vs X (Independent variables):**

CODE:

```
/*exporting clean data to an excel*/
proc export data=f_clean
dbms = xls
outfile='/folders/myfolders/SC/flight_data/FAA_clean.xls'
replace;
run;
/*import clean data from excel*/
FILENAME REFFILE '/folders/myfolders/SC/flight_data/FAA_clean.xls';
PROC IMPORT DATAFILE=REFFILE DBMS=XLS OUT=work.flight_clean;
GETNAMES=YES;
RUN;
/*create a macro to plot var1 vs var2, and use that to plot landing distance vs all
other variables*/
%macro plot1(dataset, var1, var2);
proc plot data=&dataset;
plot &var1*&var2;
title "&var1 vs &var2 in &dataset";
run;
%mend plot1;

%plot1(flight_clean, distance, duration);
%plot1(flight_clean, distance, no_pasg);
%plot1(flight_clean, distance, speed_ground);
%plot1(flight_clean, distance, speed_air);
%plot1(flight_clean, distance, height);
%plot1(flight_clean, distance, pitch);
```

OUTPUT:

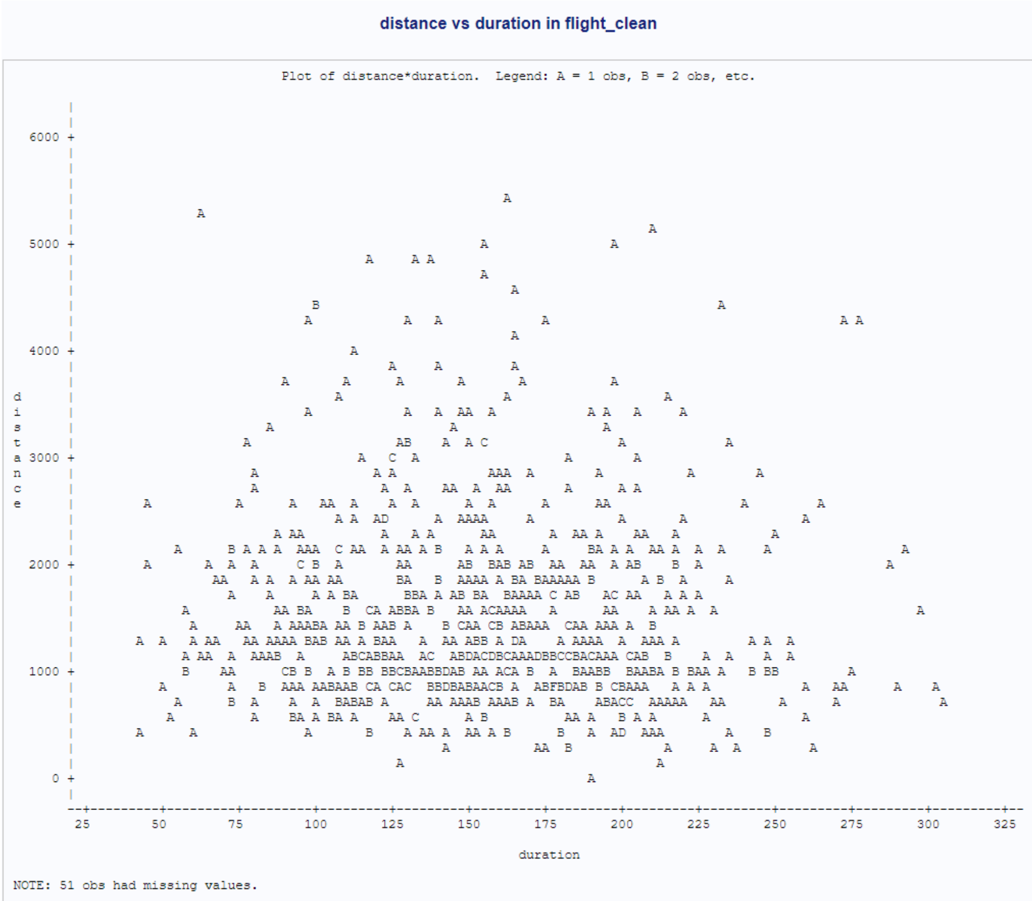


Figure 2.1(a): Plot of landing distance vs duration to visualize correlation

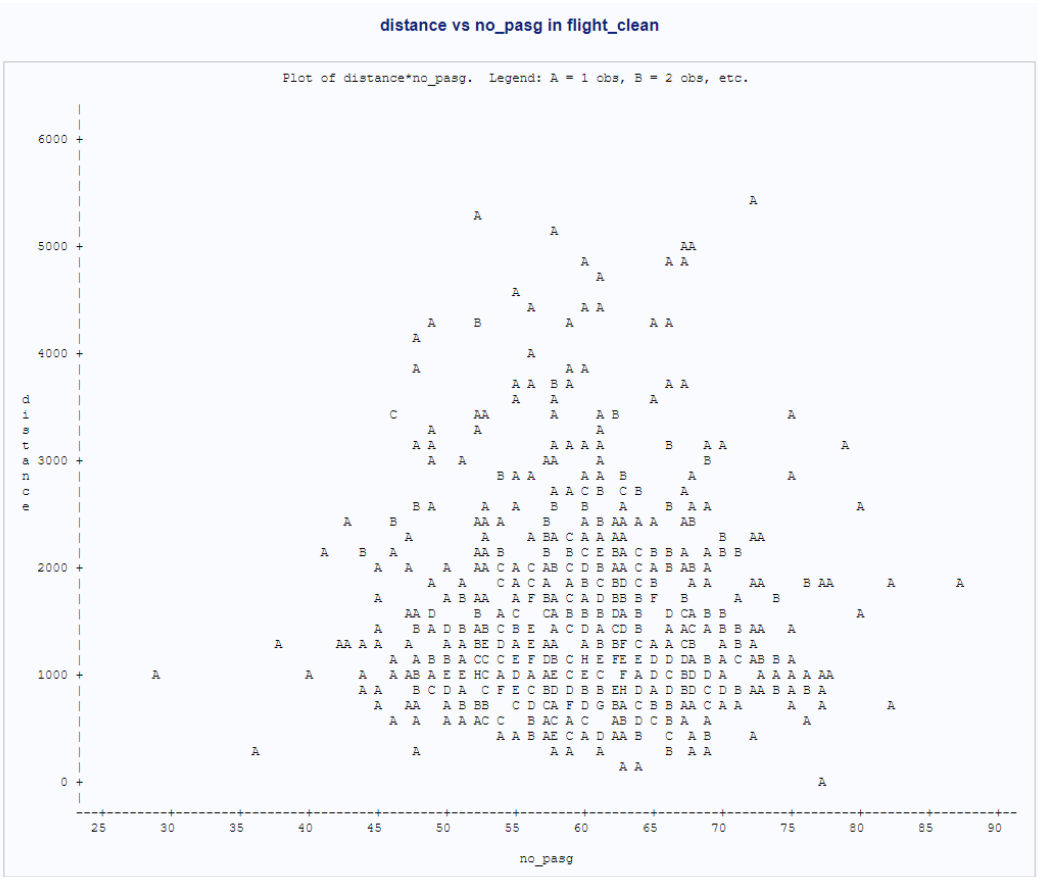


Figure 2.1(b): Plot of landing distance vs no_pasg to visualize correlation

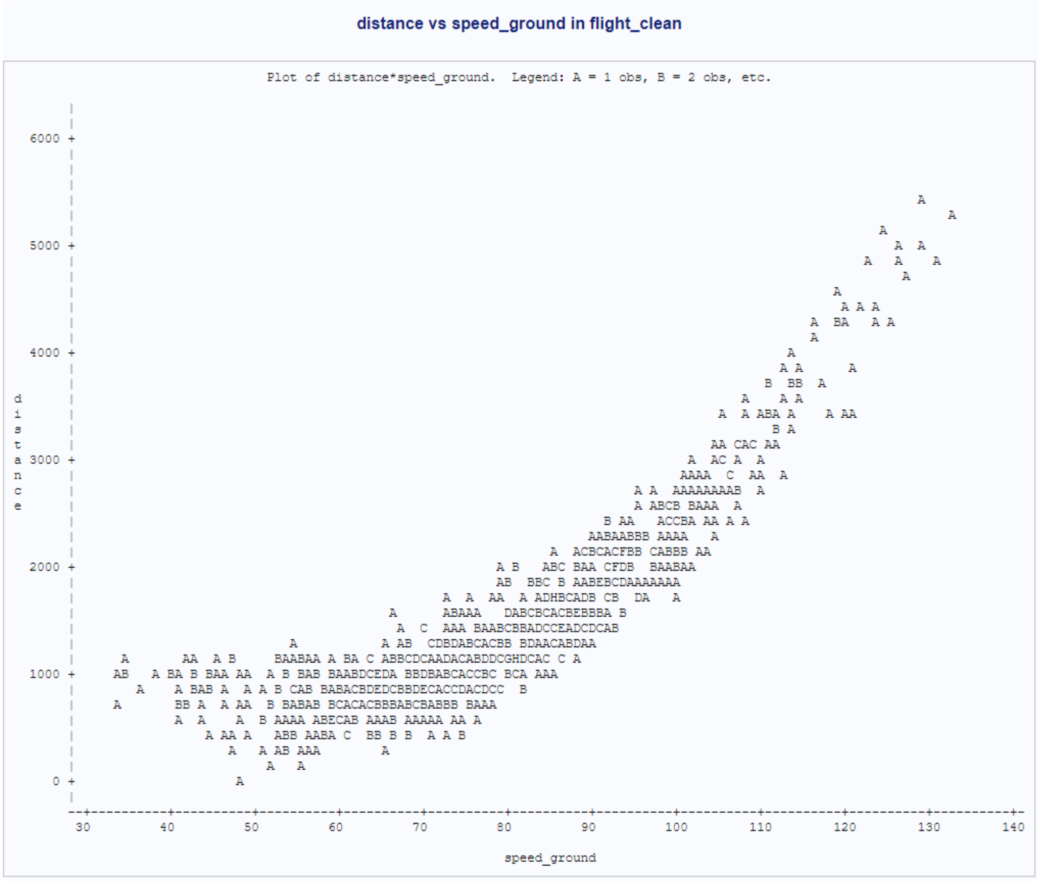


Figure 2.1(c): Plot of landing distance vs speed_ground to visualize correlation

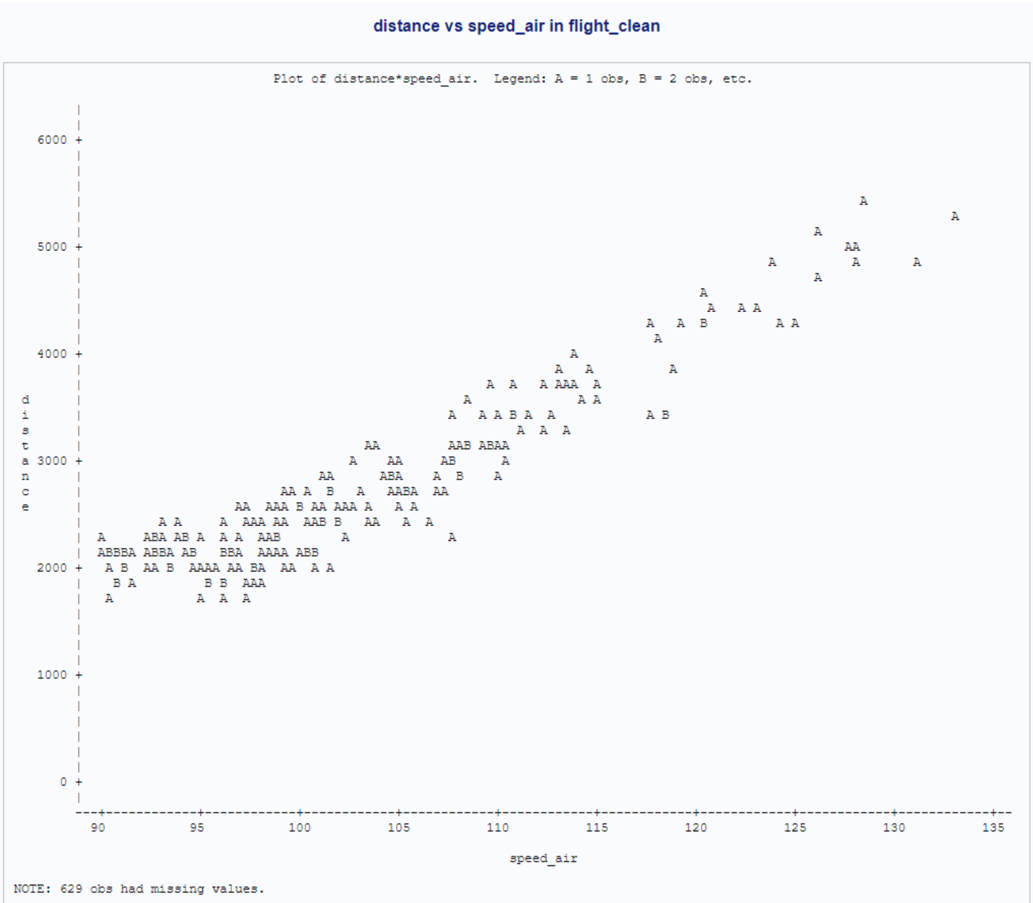


Figure 2.1(d): Plot of landing distance vs speed_air to visualize correlation

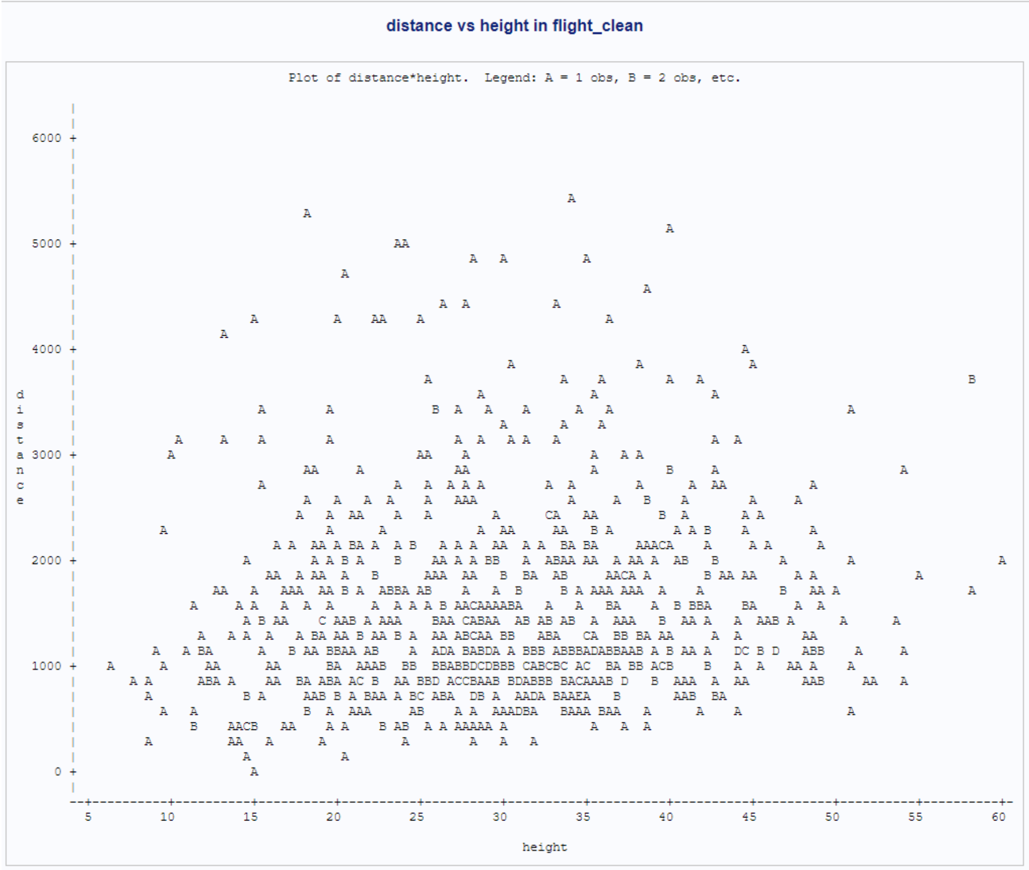


Figure 2.1(e): Plot of landing distance vs height to visualize correlation

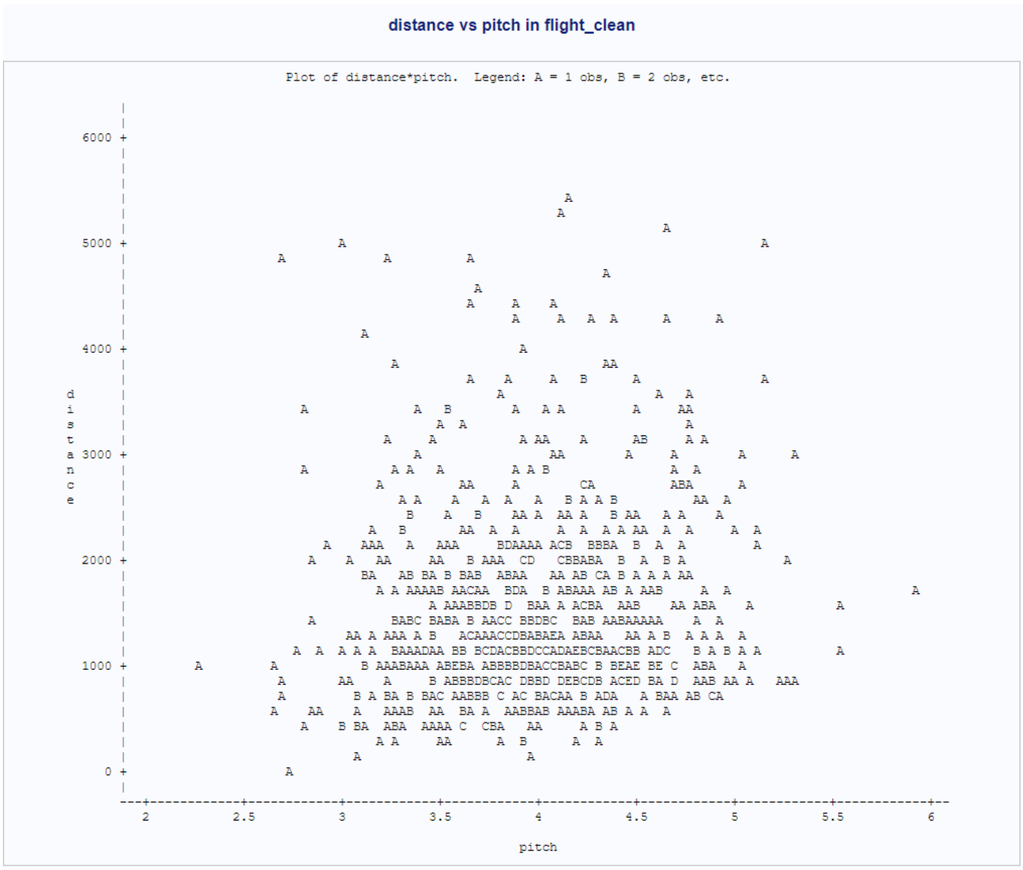


Figure 2.1(f): Plot of landing distance vs pitch to visualize correlation

CODE BRIEF AND OBSERVATIONS:

Initially we export the clean data into an excel and then import the same into a dataset called *flight_clean*. Then we use a macro to plot the landing distance with each independent variable to visualize correlation between them. This is shown in Figure 2.1(a) –(f).

We observe that there is a visible strong positive correlation between landing distance and variables speed_air and speed_ground.

---

2. <u>Calculate the correlation between variables</u>:

CODE:

```
/*to check correlation of variables with respect to the landing distance*/
proc corr data=flight_clean;
var duration no_pasg speed_ground speed_air height pitch;
with distance;
title Correlation coefficients with Landing Distance;
run;
/*speed air, speed ground, pitch and height have p value less than 0.05*/

/*to check the correlation between variables which affect landing distance*/
proc corr data=flight_clean;
var speed_ground speed_air pitch height;
title Correlation coefficients btw air and grnd speed;
run;
/*Here, we see that the speed air and speed_ground are correlated. Hence it is a
good practice to use only one of these. here we take speed ground*/
```

OUTPUT:

Correlation coefficients with Landing Distance

The CORR Procedure

1 With Variables:	distance
6 Variables:	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						
	duration	no_pasg	speed_ground	speed_air	height	pitch
distance distance	-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

Figure 2.2(a): Correlation of variables with respect to landing distance

Correlation coefficients between variables

The CORR Procedure

4 Variables:	speed_ground speed_air pitch height

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
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
pitch	832	4.00508	0.52626	3332	2.28448	5.92678	pitch
height	832	30.45540	9.77918	25339	6.22752	59.94596	height

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations				
	speed_ground	speed_air	pitch	height
speed_ground speed_ground	1.00000 832	0.98794 <.0001 203	-0.03898 0.2615 832	-0.05737 0.0982 832
speed_air speed_air	0.98794 <.0001 203	1.00000 203	-0.03927 0.5780 203	-0.07933 0.2606 203
pitch pitch	-0.03898 0.2615 832	-0.03927 0.5780 203	1.00000 832	0.02301 0.5074 832
height height	-0.05737 0.0982 832	-0.07933 0.2606 203	0.02301 0.5074 832	1.00000 832

Figure 2.2(b): Correlation between variables which affect landing distance

CODE BRIEF AND OBSERVATIONS:

We perform the procedure corr on each variable to identify its correlation with landing distance, on the *flight_clean* dataset. We observe that the variables speed_air and speed_ground have a strong positive correlation with the landing distance. Also, the height and pitch have a weak positive correlation to the landing distance. (we have taken a 95% confidence interval, and hence including height and pitch). This is shown in Figure 2.2(a).

Then, we calculate the correlation between variables affecting landing distance. We observe that speed_ground and speed_air have strong positive correlation, as shown in Figure 2.2(b). Including both in our regression analysis will affect the **variance inflation factor (VIF)**, which quantifies the severity of **multicollinearity** in an ordinary least squares regression analysis. Hence, we move forward and use only of one the variables (speed_ground in this case) in our analysis. Speed_ground is used instead of speed_air, since it covers more observations (speed_ground-832, speed_air-203) to fit the model.

---

3. <u>Do the Regression analysis</u>:

CODE:

```
proc reg data=flight_clean;
model distance = speed_ground height pitch/ r;
output out=diagnostics residual=residuals;
title Regression analysis of FAA dataset;
run;
```

OUTPUT:

Regression analysis of FAA dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	832
Number of Observations Used	832

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	524763080	174921027	1017.69	<.0001
Error	828	142317078	171881		
Corrected Total	831	667080159			

Root MSE	414.58477	R-Square	0.7867
Dependent Mean	1521.89391	Adj R-Sq	0.7859
Coeff Var	27.24137		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-3037.95491	136.30923	-22.29	<.0001
speed_ground	speed_ground	1	42.05675	0.76956	54.65	<.0001
height	height	1	13.49149	1.47340	9.16	<.0001
pitch	pitch	1	200.85997	27.35515	7.34	<.0001

Figure 2.3: Regression analysis on the FAA data

CODE BRIEF AND OBSERVATIONS:

We do the regression analysis on the FAA data. Here, we are regression landing distance on speed_ground, height and pitch. From the Figure 2.3, we see that the R squared value is .78, which explains how much we could cover the variability of the response data around the mean (R squared value is from 0-1 and 0 means none of the variability is covered). Generally, r squared value of 0.8 might overfit the data, leading to higher test MSE. Mathematically, we can write this linear relationship as

$$Y \approx \beta_0 + \beta_1 X$$
(2.1)

Where, β_0 is the intercept term and β_1 is the slope, that is, the average increase in Y with one unit increase in X. From the above table parameter estimates, this can be written as

$$Distance \approx -3037.954 + 42.056(Speed\_ground) + 13.491(Height) + 200.859(Pitch) \qquad (2.2)$$

You might read "≈" as *"is approximately modeled as"*.

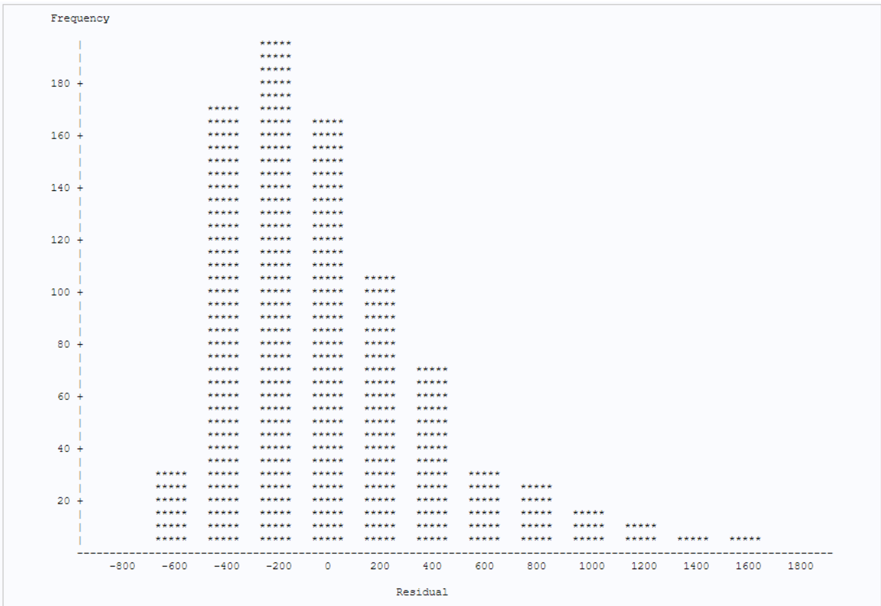
Also, we save the residuals to a new dataset called *diagnostics,* for model checking.

4. **Model Diagnostics:**

CODE:

```
proc chart data=diagnostics;
vbar residuals;
run;
proc ttest data = diagnostics;
var residuals;
run;
```

OUTPUT:



The TTEST Procedure

Variable: residuals (Residual)

DF	t Value	Pr > |t|
831	0.00	1.0000

Figure 2.4: proc chart and ttest for model diagnostics

CODE BRIEF AND OBSERVATIONS:

We first plot the residuals to check its distribution. We find that it is normal but skewed slightly. Then we perform ttest to check if the mean=0. Since the p value is high, we fail to reject the null hypothesis, which states that the mean=0. Hence, the residuals confirm our model to predict the lading distance. (Figure 2.4)

Exploratory Data Analysis for each aircraft type:

We perform the same for each aircraft type to study what factors and how they would impact the landing distance in each case.

CODE:

```
/*AIRCRAFT TYPE - AIRBUS*/
data airbus;
```

```
set flight_clean;
if aircraft = 'airbus';
run;

%plot1(airbus, distance, duration);
%plot1(airbus, distance, no_pasg);
%plot1(airbus, distance, speed_ground);
%plot1(airbus, distance, speed_air);
%plot1(airbus, distance, height);
%plot1(airbus, distance, pitch);

/*to check correlation of variables with respect to the landing distance in Airbus*/
proc corr data=airbus;
var duration no_pasg speed_ground speed_air height pitch;
with distance;
title Correlation coefficients with Landing Distance in Airbus;
run;
/*speed air, speed ground and height have p value less than 0.05*/

/*to check the correlation between variables which affect landing distance*/
proc corr data=airbus;
var speed_ground speed_air height;
title Correlation coefficients between variables;
run;

proc reg data=airbus;
model distance = speed_ground height;
title Regression analysis of FAA dataset airbus;
run;

/*AIRCRAFT TYPE - BOEING*/
data boeing;
set flight_clean;
if aircraft = 'boeing';
run;

%plot1(boeing, distance, duration);
%plot1(boeing, distance, no_pasg);
%plot1(boeing, distance, speed_ground);
%plot1(boeing, distance, speed_air);
%plot1(boeing, distance, height);
%plot1(boeing, distance, pitch);

/*to check correlation of variables with respect to the landing distance in Boeing*/
proc corr data=boeing;
var duration no_pasg speed_ground speed_air height pitch;
with distance;
title Correlation coefficients with Landing Distance in boeing;
run;

/*to check the correlation between varaibles which affect landing distance*/
proc corr data=boeing;
var speed_ground speed_air;
title Correlation coefficients between variables;
run;
```

```
proc reg data=boeing;
model distance = speed_ground;
title Regression analysis of FAA dataset boeing;
run;
```

OUTPUT:

AIRBUS:

Correlation coefficients with Landing Distance in Airbus

The CORR Procedure

1 With Variables:	distance
6 Variables:	duration no_pasg speed_ground speed_air height pitch

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	444	1323	791.92825	587553	41.72231	4896	distance
duration	394	156.90333	49.18829	61820	42.14623	305.62171	duration
no_pasg	444	60.21396	7.42649	26735	36.00000	87.00000	no_pasg
speed_ground	444	80.24988	16.95497	35631	33.57410	131.03518	speed_ground
speed_air	85	104.30976	8.08959	8866	95.01136	131.33795	speed_air
height	444	30.58922	9.85439	13582	6.22752	58.22780	height
pitch	444	3.83114	0.49608	1701	2.28448	5.52678	pitch

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

	duration	no_pasg	speed_ground	speed_air	height	pitch
distance distance	-0.07851 0.1198 394	-0.00732 0.8777 444	0.90520 <.0001 444	0.96411 <.0001 85	0.14494 0.0022 444	0.07330 0.1230 444

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

	speed_ground	speed_air	height
speed_ground speed_ground	1.00000 444	0.98169 <.0001 85	-0.03346 0.4819 444
speed_air speed_air	0.98169 <.0001 85	1.00000 85	-0.00546 0.9604 85
height height	-0.03346 0.4819 444	-0.00546 0.9604 85	1.00000 444

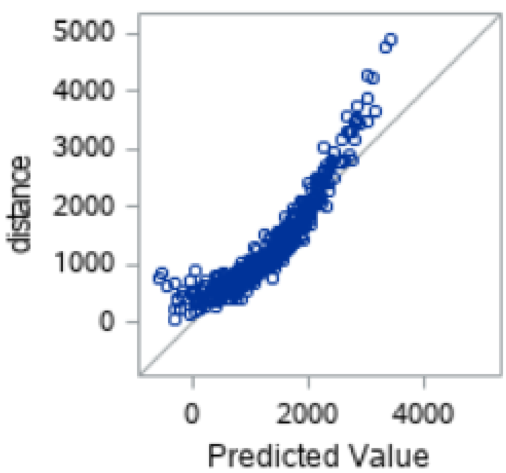
Regression analysis of FAA dataset airbus

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	444
Number of Observations Used	444

Root MSE	307.26984	R-Square	0.8501
Dependent Mean	1323.31696	Adj R-Sq	0.8495
Coeff Var	23.21967		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2522.89061	85.19508	-29.61	<.0001
speed_ground	speed_ground	1	42.55420	0.86152	49.39	<.0001
height	height	1	14.09773	1.48228	9.51	<.0001



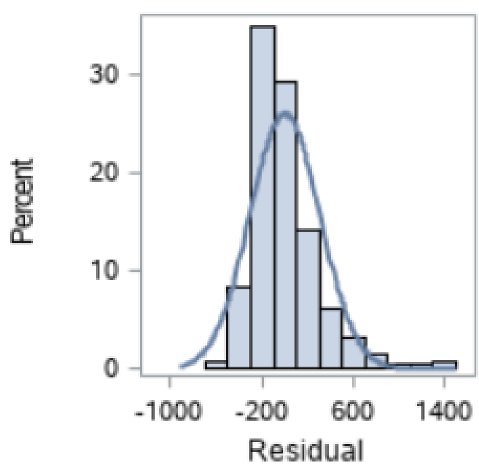


Figure 2.5: Analysis of landing distance for aircraft type – AIRBUS

BOEING:

Correlation coefficients with Landing Distance in boeing

The CORR Procedure

1 With Variables:	distance
6 Variables:	duration no_pasg speed_ground speed_air height pitch

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	388	1749	953.31500	678663	573.62179	5382	distance
duration	387	152.60962	47.44672	59060	41.94937	298.52233	duration
no_pasg	388	59.88144	7.56241	23234	29.00000	82.00000	no_pasg
speed_ground	388	78.69229	20.57029	30533	33.82295	132.78468	speed_ground
speed_air	118	102.89095	10.76242	12141	90.00286	132.91146	speed_air
height	388	30.30227	9.70284	11757	7.58249	59.94596	height
pitch	388	4.20413	0.48841	1631	2.99315	5.92678	pitch

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

	duration	no_pasg	speed_ground	speed_air	height	pitch
distance distance	-0.01064 0.8347 387	-0.01864 0.7143 388	0.90064 <.0001 388	0.97760 <.0001 118	0.06953 0.1717 388	-0.06391 0.2091 388

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

	speed_ground	speed_air
speed_ground speed_ground	1.00000 388	0.99048 <.0001 118
speed_air speed_air	0.99048 <.0001 118	1.00000 118

Regression analysis of FAA dataset boeing

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	388
Number of Observations Used	388

Root MSE	414.80899	R-Square	0.8112
Dependent Mean	1749.13145	Adj R-Sq	0.8107
Coeff Var	23.71514		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1535.45506	83.36840	-18.42	<.0001
speed_ground	speed_ground	1	41.73962	1.02507	40.72	<.0001

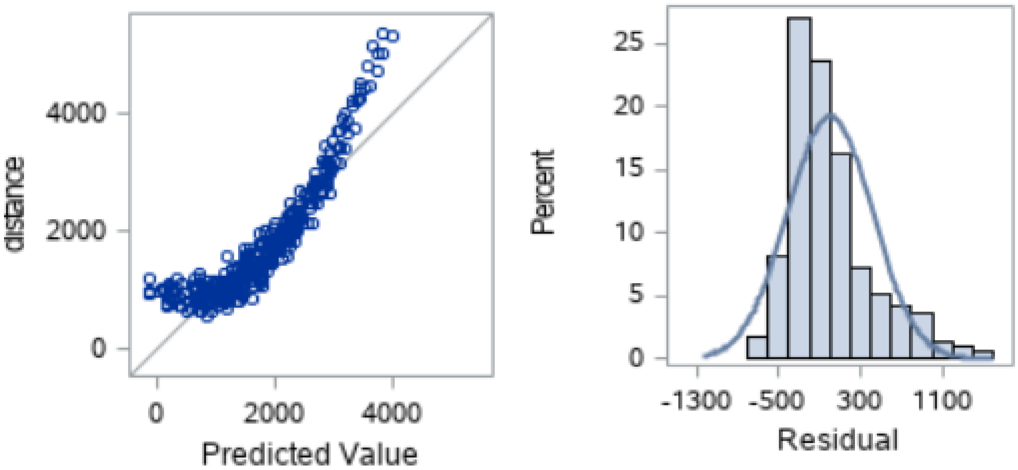


Figure 2.6: Analysis of landing distance for aircraft type – BOEING

CODE BRIEF AND OBSERVATIONS:

Finally, we perform our analysis on different aircraft type to study the difference in parameters affecting landing distance for different aircraft types.

Airbus:
As shown in Figure 2.5, for Airbus, we observe that there is a total of 444 observations. Using these observations, we find the correlation of different variables with respect to the landing distance. Compared to the complete FAA dataset, where both ground and air speed, height and pitch affect landing distance, the airbus aircraft type pitch does not have a correlation with respect to the landing distance at 95% confidence level. Also, since speed_ground and speed_air are correlated, including both in our regression analysis will affect the variance inflation factor (VIF). Hence, we use speed_ground and height alone to predict the landing distance. Performing regression of distance on these variables, we get a mathematical linear relationship as below:

$$Distance \approx -2522.891 + 42.554(Speed\_ground) + 14.097(Height) \qquad (2.3)$$

Boeing:
As shown in Figure 2.6, for Boeing, we observe that there is a total of 388 observations. Using these observations, we find the correlation of different variables with respect to the landing distance. Compared to the complete FAA dataset, where both ground and air speed, height and pitch affect landing distance, the airbus aircraft type pitch and height does not have a correlation with respect to the landing distance at 95% confidence level. Also, since speed_ground and speed_air are correlated, including both in our regression analysis will affect the variance inflation factor (VIF). Hence, we use speed_ground alone to predict the landing distance. Performing regression of distance on speed_ground, we get a mathematical linear relationship as below:

$$Distance \approx -2522.455 + 41.739\ (Speed\_ground) \qquad (2.3)$$

CONCLUSION:

In this chapter, we analyzed what factors and how they would impact the landing distance of a commercial flight. Considering the entire FAA data, the variables speed_ground, speed_air, height and pitch impact landing distance. We also found that there is a difference in the impact of variables between aircraft types. The variable height affects Airbus but not Boeing. Finally, we regress landing distance on each correlated variable, check for VIF. All the above prediction is done at 95% confidence (alpha = 0.05).