Ashwita Saxena
M06119969
Statistical Computing FAA Project

Statistical Computing Project

Ashwita Saxena

M06119969

Summary: While studying what factors and how they impact the landing distance of a commercial flight, I used raw data with information about 950 flights. After cleaning the data, I was left with 811 flights. By doing exploratory data analysis, I found out that landing distance of a flight is highly correlated with the speed of the aircraft when in air and speed of the aircraft when on the ground. Different linear regression models gave different results as to which variable impacts the landing distance. Overall, the models explained that height of the aircraft when passing over the threshold and ground speed as well as air speed are factors that highly impact the landing distance. Boeing and Airbus type of aircrafts also impacted the landing distance of the aircrafts differently. The overall data processing and modeling procedure gave us good insight into finding out how independent variables impacted the dependent variable distance. However, the small sample size of 811 is not sufficient to imply the results on the population. We should collect more data and analyze it in order to be able to accurately predict the factors affecting the landing distance.

Ashwita Saxena
M06119969
Statistical Computing FAA Project

Goal: The goal of this report is to study what factors and how they would impact the landing distance of a commercial flight .

CHAPTER 1: DATA PREPERATION

Data Set

To study what factors and how they would impact the landing distance of a commercial flight, I imported 2 data sets into SAS – FAA1 and FAA2.

1. FAA1: This dataset contains 800 observations and 8 variables.
2. FAA2: This dataset contains 200 observations and 7 variables.

The variables are:

Aircraft: The make of an aircraft (Boeing or Airbus).
Duration (in minutes): Flight duration between taking off and landing.
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.
Speed_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway.
Height (in meters): The height of an aircraft when it is passing over the threshold of the runway.
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.

```
libname project '/folders/myfolders/Statistical Computing';

FILENAME data1 '/folders/myfolders/Statistical Computing/FAA1.xls';

PROC IMPORT DATAFILE=data1
    DBMS=XLS
    OUT=project.FAA1;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=project.FAA1; RUN;

FILENAME data2 '/folders/myfolders/Statistical Computing/FAA2.xls';

PROC IMPORT DATAFILE=data2
    DBMS=XLS
    OUT=project.FAA2;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=project.FAA2; RUN;
```

Data Set Name	PROJECT.FAA1	Observations	800
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	09/09/2018 16:24:46	Observation Length	72
Last Modified	09/09/2018 16:24:46	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)		

Data Set Name	PROJECT.FAA2	Observations	200
Member Type	DATA	Variables	7
Engine	V9	Indexes	0
Created	09/09/2018 16:37:59	Observation Length	64
Last Modified	09/09/2018 16:37:59	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)		

Merging Datasets:

I merged both the datasets to create one dataset for further analysis. The new dataset has 1000 observations and 8 variables.

```
/*merging data*/

data project.merge;
set project.faa1 project.faa2;
run;

proc print data=project.merge;
run;
```

```
73          data project.merge;
74          set project.faa1 project.faa2;
75          run;

NOTE: There were 800 observations read from the data set PROJECT.FAA1.
NOTE: There were 200 observations read from the data set PROJECT.FAA2.
NOTE: The data set PROJECT.MERGE has 1000 observations and 8 variables.
NOTE: DATA statement used (Total process time):
      real time           0.07 seconds
      cpu time            0.02 seconds
```

I explored the variables of the new dataset by using proc means.

```
/*exploring the new dataset*/

proc means data=project.merge;
run;

```

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
duration	duration	800	154.0065385	49.2592338	14.7642071	305.6217107
no_pasg	no_pasg	950	60.1652632	7.4900041	29.0000000	87.0000000
speed_ground	speed_ground	950	79.2849940	19.3364178	27.7357153	141.2186354
speed_air	speed_air	239	103.7304174	10.6051134	90.0028586	141.7249357
height	height	950	30.1392714	10.3593491	-3.5462524	59.9459639
pitch	pitch	950	4.0192472	0.5260322	2.2844801	5.9267842
distance	distance	950	1548.82	948.6812561	34.0807833	6533.05

Upon looking at the data, I figured that there were duplicate values in the data. I removed the duplicates by excluding the variable ‘duration’ since during my analysis I found out that there were a significant number of missing duration values. The resulting dataset had 850 values as 150 duplicate values were removed. Most of the missing duration values were removed from the dataset as a result.

```
/*removing duplicates from the data*/

proc sort data=project.merge out=project.remdup nodupkey;
by distance speed_ground height pitch no_pasg aircraft speed_air;
run;

proc print data=project.remdup;
run;

Data project.remdup_new;
Set project.remdup;
IF no_pasg= "." then delete;
RUN;

PROC PRINT Data = project.remdup_new;
PROC MEANS Data = project.remdup_new;
RUN;
```

I explored the new dataset to check out if it has any missing values.

Ashwita Saxena
M06119969


```
/*data exploration*/
proc contents data=project.remdup_new;
run;
PROC MEANS Data = project.remdup_new;
RUN;
proc freq data=project.remdup_new;
run;
```

The CONTENTS Procedure

Data Set Name	PROJECT.REMDUP_NEW	Observations	850
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	09/17/2018 11:56:26	Observation Length	72
Last Modified	09/17/2018 11:56:26	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)		

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
duration	duration	800	154.0065385	49.2592338	14.7642071	305.6217107
no_pasg	no_pasg	850	60.1035294	7.4931370	29.0000000	87.0000000
speed_ground	speed_ground	850	79.4523229	19.0594903	27.7357153	141.2186354
speed_air	speed_air	208	103.7977237	10.2590370	90.0028586	141.7249357
height	height	850	30.1442223	10.2877268	-3.5462524	59.9459639
pitch	pitch	850	4.0093577	0.5288298	2.2844801	5.9267842
distance	distance	850	1526.02	928.5600816	34.0807833	6533.05

The proc means result shows that there are missing values in duration and speed_air and number of observations does not sum up to 850. Upon running proc freq, I confirmed the number of missing values. I am pasting only a section of the result for relevance purpose. The results show that there are 50 missing values in duration and 642 missing values in speed air. For my analysis, I will use speed_ground instead of speed_air as they are both very similar when compared. As such, I did not get rid of the missing values from speed_air and let it be as is.

duration				
duration	Frequency	Percent	Cumulative Frequency	Cumulative Percent
14.764207145	1	0.13	1	0.13
16.893454896	1	0.13	2	0.25
17.375513046	1	0.13	3	0.38
Frequency Missing = 50				

speed_air				
speed_air	Frequency	Percent	Cumulative Frequency	Cumulative Percent
90.002858582	1	0.48	1	0.48
90.111013336	1	0.48	2	0.96
90.367403727	1	0.48	3	1.44
Frequency Missing = 642				

Data Cleaning: As listed in the data description, I filtered out the abnormal values from the dataset and then looked at the variables by using proc means. The minimum and maximum values look good as per the filters applied.

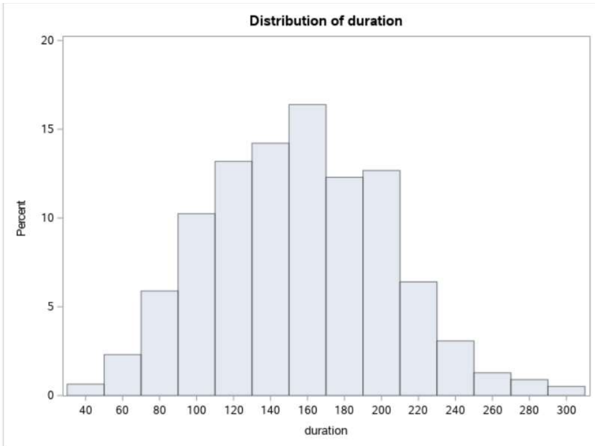
```
/*data cleaning*/
data project.clean;
set project.remdup_new;
if aircraft ='' then delete;
if duration ne '' then if duration<40 then delete;
if speed_ground < 30 then delete;
if speed_ground > 140 then delete;
if height < 6 then delete;
if distance > 6000 then delete;
run;

proc print data=project.clean;
run;
proc means data=project.clean;
run;
```

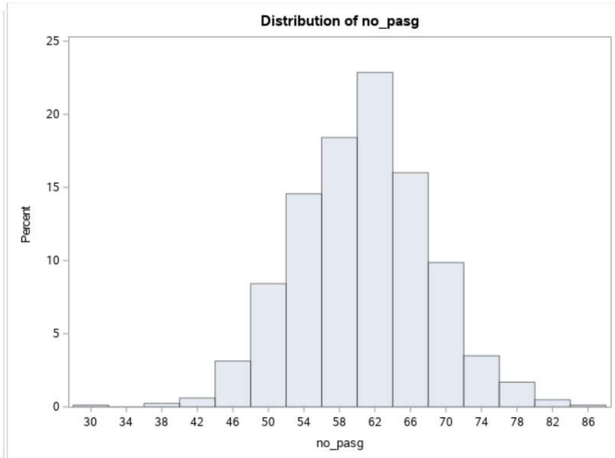
The MEANS Procedure

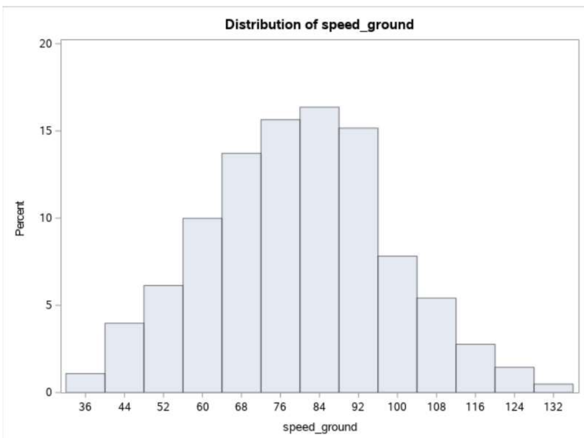
Variable	Label	N	Mean	Std Dev	Minimum	Maximum
duration	duration	781	154.7757191	48.3499237	41.9493694	305.6217107
no_pasg	no_pasg	831	60.0553550	7.4913166	29.0000000	87.0000000
speed_ground	speed_ground	831	79.5426997	18.7356754	33.5741041	132.7846766
speed_air	speed_air	203	103.4850352	9.7362774	90.0028586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

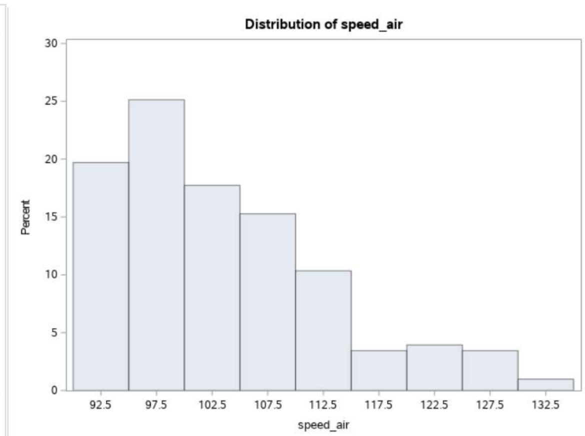
I looked that the distribution of each variable and determined that we don’t need to remove any extreme values as they all fall under our constraints requirements. The distributions look like follows:

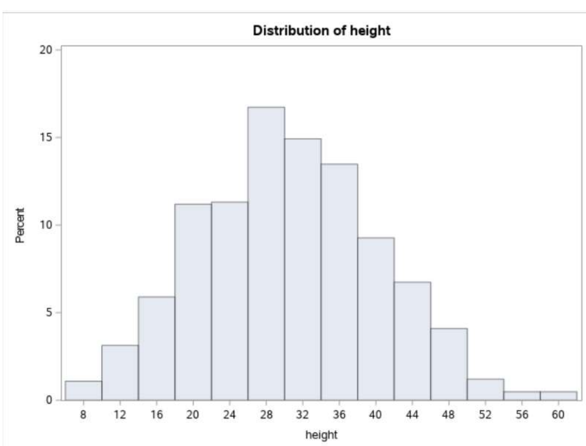
```
proc univariate data=project.clean;
histogram;
run;
```

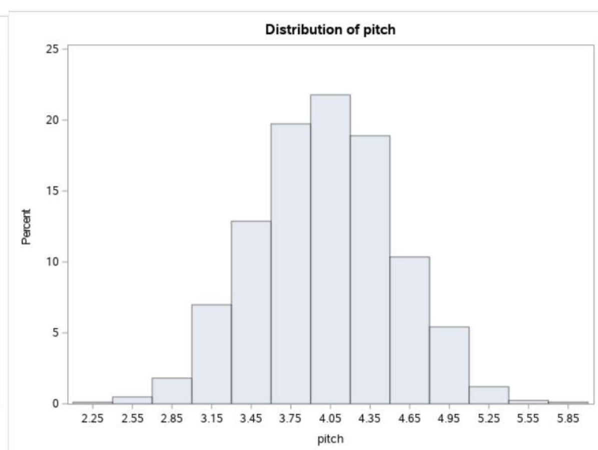










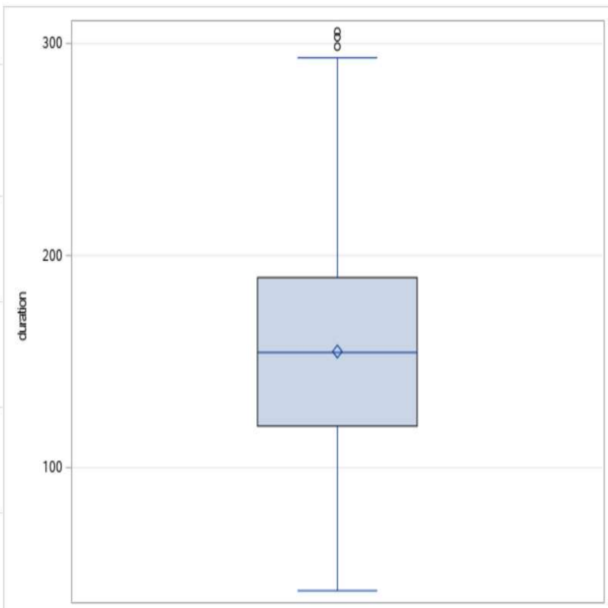


Upon looking at the minimum value of distance, I determined that I should remove the smallest two values from the data as the distance is too small for an aircraft to cover at landing. If the aircraft tries to stop after covering this distance, it would probably be similar to crashing.

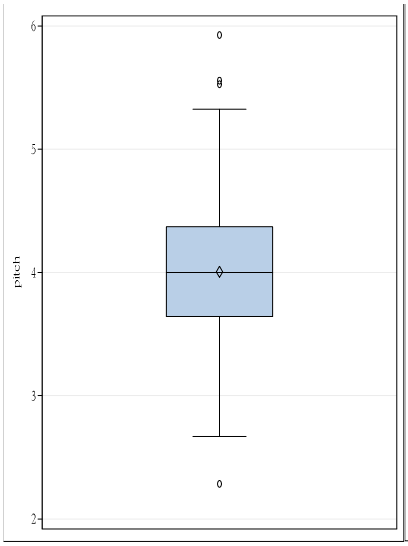
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
41.7223	1	5031.39	827
133.0869	2	5058.47	828
180.5652	3	5147.41	829
241.1610	4	5343.20	830
242.5959	5	5381.96	831

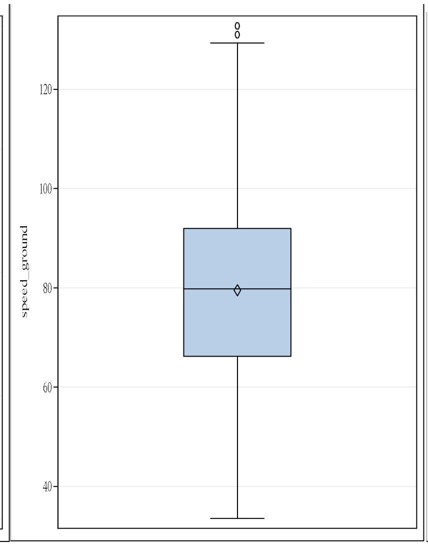
I also looked at box plots of each variable and removed all the outliers above and below min and max whiskers. For variable duration however, I only removed the values above the 99th percentile. Since distance has many outliers, I donot want to get rid of all of them as it will reduce our sample size by many.

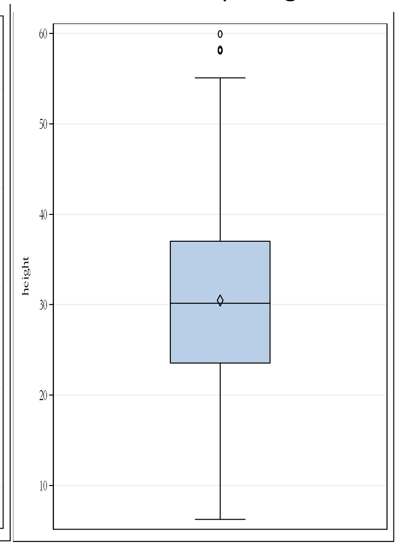
```
proc sgplot data=PROJECT.CLEAN;
        vbox duration;
        yaxis grid;
    run;
proc sgplot data=project.clean;
        vbox pitch;
        yaxis grid;
    run;
proc sgplot data=project.clean;
        vbox speed_ground;
        yaxis grid;
    run;

proc sgplot data=project.clean;
        vbox height;
        yaxis grid;
    run;
proc sgplot data=project.clean;
        vbox distance;
        yaxis grid;
    run;
proc sgplot data=project.clean;
        vbox speed_air;
        yaxis grid;
    run;
proc sgplot data=project.clean;
        vbox no_pasg;
        yaxis grid;
    run;
```



duration



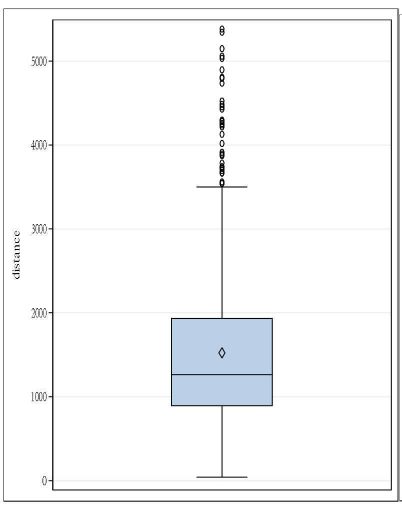


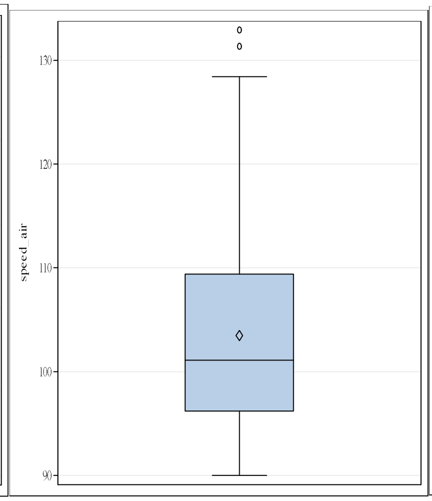


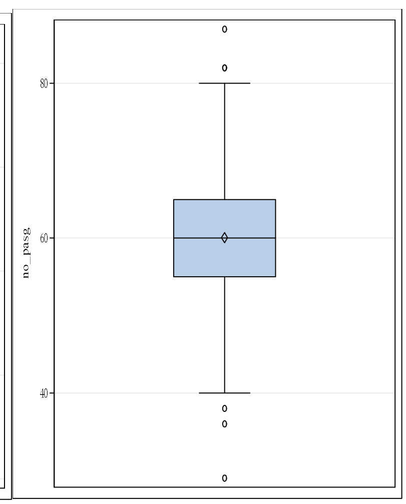
Pitch

speed_ground

height







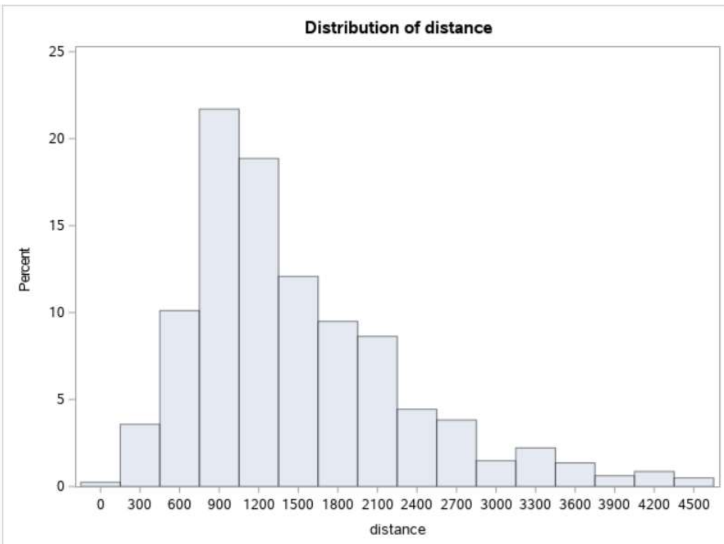
Distance

speed_air

no_pasg

```
data project.final;
set project.clean;
 if duration>294 then delete;
 if pitch>5.4 then delete;
 if pitch<2.3 then delete;
 if speed_ground>130 then delete;
 if height>56 then delete;
 if distance<180 then delete;
 if distance>4737 then delete;
 if speed_air>125.5 then delete;
run;
```

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
duration	duration	762	154.2826570	47.8364705	41.9493694	293.2299603
no_pasg	no_pasg	811	60.0036991	7.5337093	29.0000000	87.0000000
speed_ground	speed_ground	811	78.9646313	18.1554068	33.5741041	125.2123041
speed_air	speed_air	192	102.2520293	8.3281113	90.0028586	125.1385489
height	height	811	30.3689940	9.6150446	6.2275178	55.0935091
pitch	pitch	811	4.0045932	0.5116301	2.6689057	5.3247470
distance	distance	811	1480.20	819.1654654	41.7223127	4524.28



Distribution of landing distance after outlier treatment.

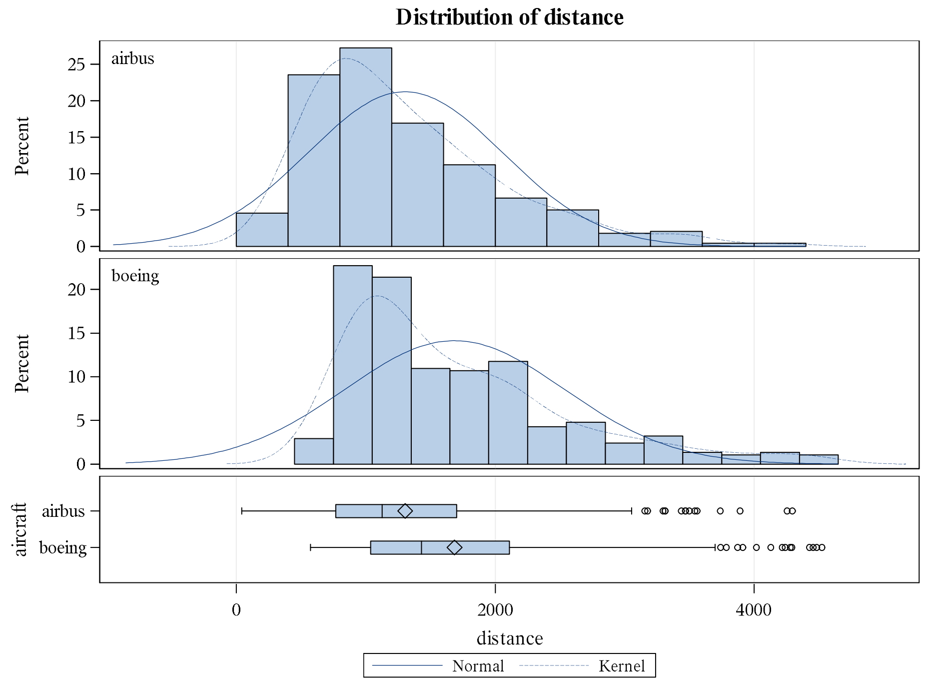
My final dataset has 811 observations and 8 variables. The distribution of the variables pitch, height, speed_ground and duration looks somewhat close to a normal distribution. However, distance is highly skewed to the right. Speed_air is also skewed to the right and number of passengers is slightly skewed to the left.

CHAPTER 2: EXPLORATORY DATA ANALYSIS

I did a ttest to find out how different categories of aircraft (Boeing/airbus) impacted the distance.

```
/*ttest for aircraft*/
proc ttest data=project.final;
class aircraft;
var distance;
run;
```

Ashwita Saxena
M06119969
Statistical Computing FAA Project



The TTEST Procedure

Variable: distance (distance)

aircraft	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus		437	1304.4	752.1	35.9800	41.7223	4295.9
boeing		374	1685.6	847.0	43.7982	573.6	4524.3
Diff (1-2)	Pooled		-381.1	797.3	56.1631		
Diff (1-2)	Satterthwaite		-381.1		56.6820		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1304.4	1233.7	1375.1	752.1	705.4	805.6
boeing		1685.6	1599.5	1771.7	847.0	790.4	912.5
Diff (1-2)	Pooled	-381.1	-491.4	-270.9	797.3	760.3	838.1
Diff (1-2)	Satterthwaite	-381.1	-492.4	-269.9			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	809	-6.79	<.0001
Satterthwaite	Unequal	752.95	-6.72	<.0001

Equality of Variances

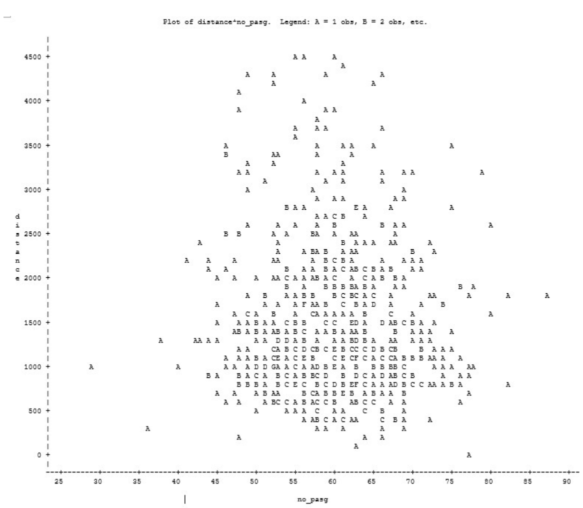
Method	Num DF	Den DF	F Value	Pr > F
Folded F	373	436	1.27	0.0170

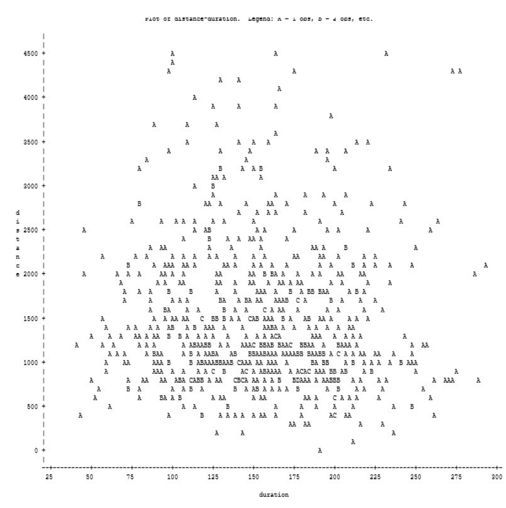
This test shows that mean landing distance for airbus is 1304 feet whereas for boeing, it is 1685.6 feet. The standard deviation of Boeing is also greater than that of airbus showing that distance of boeing is more spread across its mean as compared to airbus. Since P value of Folded F is less than .05, we look at the Satterthwaite test. The p value os Satterthwaite test is less than .0001 which means that we can reject the null that the landing distance for the two different types of planes is significantly different.

Ashwita Saxena
M06119969
Statistical Computing FAA Project

I used proc plot to plot each variable with distance. I observed that speed_air and speed_ground have a strong relationship with distance. I explored it further by creating scatterplots for the two variables

```
proc plot data=project.final;
plot distance*no_pasg;
run;
proc plot data=project.final;
plot distance*duration;
run;
proc plot data=project.final;
plot distance*speed_ground;
run;
proc plot data=project.final;
plot distance*speed_air;
run;
proc plot data=project.final;
plot distance*height;
run;
proc plot data=project.final;
plot distance*pitch;
run;
/* some correlation observed between distance and speed ground and distance and speed air*/

proc sgplot data=project.final;
   reg x=speed_air y=distance / nomarkers;
   title 'scatterplot';
   scatter x=speed_air y=distance;
run;
proc sgplot data=project.final;
   reg x=speed_ground y=distance / nomarkers;
   title 'scatterplot';
   scatter x=speed_ground y=distance;
run;
```

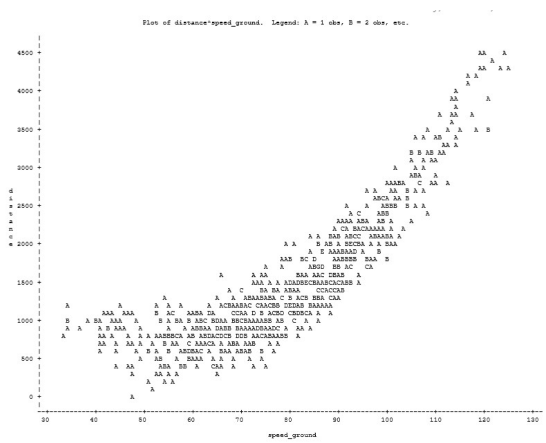


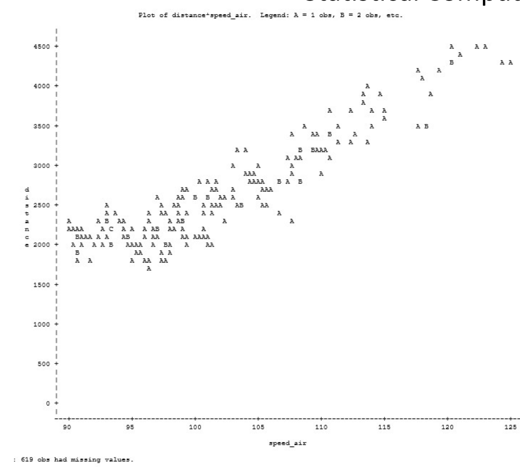


Distance*no_pasg

Distance*Duration

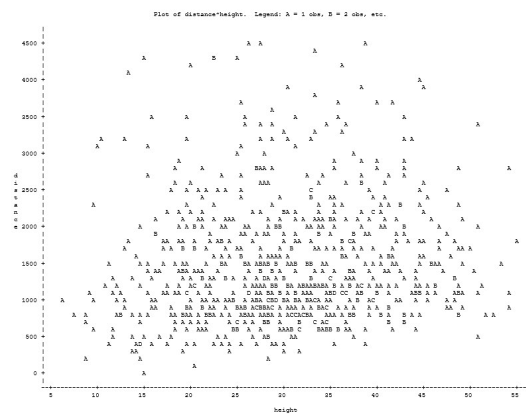
Ashwita Saxena
M06119969
Statistical Computing FAA Project

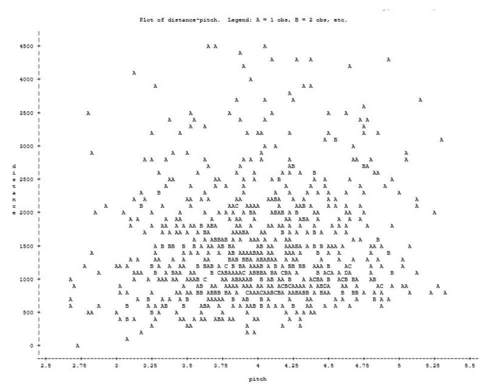




Distance*speed_ground

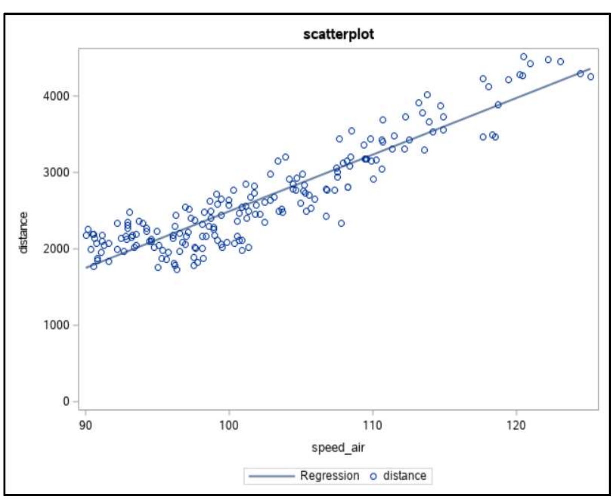
Distance*speed_air

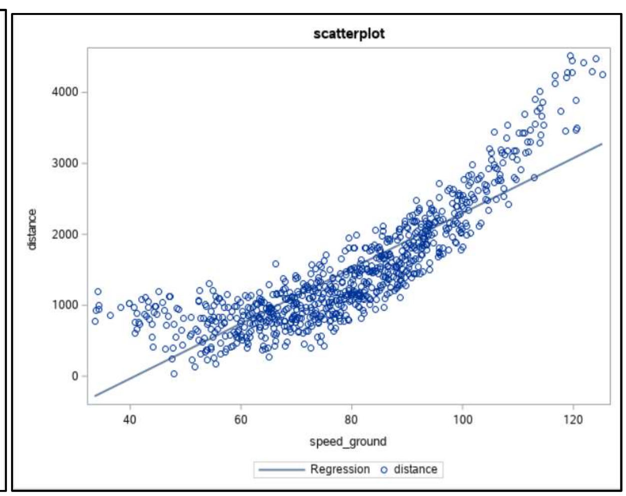




Distance*height

Distance*pitch





The scatterplots show that there is a linear relationship between distance and speed_air and distance and speed_ground.

I conducted a correlation analysis to find out the percentage of correlation between the variables. The results are as below.

```
/*correlation analysis*/
proc corr data=project.final;
var distance pitch height speed_ground speed_air no_pasg duration;
run;
```

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

	distance	pitch	height	speed_ground	speed_air	no_pasg	duration
distance distance	1.00000 811	0.10470 0.0028 811	0.10349 0.0032 811	0.86058 <.0001 811	0.91868 <.0001 192	-0.04177 0.2348 811	-0.04482 0.2166 762
pitch pitch	0.10470 0.0028 811	1.00000 811	0.03528 0.3156 811	-0.03548 0.3129 811	-0.00576 0.9368 192	-0.01118 0.7506 811	-0.03942 0.2771 762
height height	0.10349 0.0032 811	0.03528 0.3156 811	1.00000 811	-0.06262 0.0747 811	-0.09905 0.1717 192	0.04799 0.1721 811	0.00831 0.8188 762
speed_ground speed_ground	0.86058 <.0001 811	-0.03548 0.3129 811	-0.06262 0.0747 811	1.00000 811	0.98340 <.0001 192	-0.01409 0.6886 811	-0.04504 0.2143 762
speed_air speed_air	0.91868 <.0001 192	-0.00576 0.9368 192	-0.09905 0.1717 192	0.98340 <.0001 192	1.00000 192	-0.08082 0.2651 192	0.07049 0.3417 184
no_pasg no_pasg	-0.04177 0.2348 811	-0.01118 0.7506 811	0.04799 0.1721 811	-0.01409 0.6886 811	-0.08082 0.2651 192	1.00000 811	-0.04157 0.2517 762
duration duration	-0.04482 0.2166 762	-0.03942 0.2771 762	0.00831 0.8188 762	-0.04504 0.2143 762	0.07049 0.3417 184	-0.04157 0.2517 762	1.00000 762

Distance is highly correlated with speed_ground with a coefficient of correlation of 86%. Speed air is also highly correlated with distance with a correlation coefficient of 91.86%. Based on the exploratory data analysis, we expect to find ground_speed and ground_air to be significant factors that impact landing distance.

CHAPTER 3: MODELING

In order to interpret how aircraft type airbus interacts with landing distance, I created a dummy variable (aircraft_dummy). When the aircraft type is airbus, the dummy value is 1 and when the aircraft type is boeing, the dummy value is 0. I could also do the same for Boeing, but I stick to airbus for this model.

```
/* creating dummy variable for aircraft*/
data project.dummy;
set project.final;
if aircraft='airbus' then aircraft_dummy=1;
else aircraft_dummy=0;
run;
```

MODEL 1: I created my first model with landing distance as the dependent variable and all the other variables as the independent variables.

```
proc reg data=project.dummy;
model distance= no_pasg duration speed_ground height pitch speed_air aircraft_dummy/*/r*/;
title 'Regression Model with all the variables';
/*output out=diagnostics r=residual*/;
run;
```

Regression Model with all the variables

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	811
Number of Observations Used	184
Number of Observations with Missing Values	627

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	81074872	11582125	723.66	<.0001
Error	176	2816870	16005		
Corrected Total	183	83891743			

Root MSE	126.51065	R-Square	0.9664
Dependent Mean	2665.15304	Adj R-Sq	0.9651
Coeff Var	4.74684		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5413.59510	171.35412	-31.59	<.0001
no_pasg	no_pasg	1	-3.38287	1.33777	-2.53	0.0123
duration	duration	1	0.16406	0.19593	0.84	0.4035
speed_ground	speed_ground	1	-1.16788	6.21483	-0.19	0.8512
height	height	1	13.60271	1.04417	13.03	<.0001
pitch	pitch	1	-2.83371	18.53208	-0.15	0.8786
speed_air	speed_air	1	79.67244	6.37711	12.49	<.0001
aircraft_dummy		1	-417.38135	20.71198	-20.15	<.0001

The overall model is statistically significant with a P-value <0.0001. The model has an R-Squared value of 96.64% which means that this model explains 96% of the variability of the data around its mean. However, when we look at the significance of each of the independent variables, we can see that only 4 variables (no_pasg, height, speed_air and aircraft_dummy) are statistically significant at 95% confidence level. However, This model only used 184 out of the 811 observations. Majority of the data has not been used in this model because of missing values. Hence we need to change some parameters in our model.

<u>MODEL 2</u>:

```
/*model without speed_air as it has many missing values*/
proc reg data=project.dummy;
model distance= no_pasg duration speed_ground height pitch aircraft_dummy;
title 'Model without speed_air';
run;
```

Model without speed_air

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	811
Number of Observations Used	762
Number of Observations with Missing Values	49

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	439510774	73251796	721.39	<.0001
Error	755	76664235	101542		
Corrected Total	761	516175010			

Root MSE	318.65661	R-Square	0.8515
Dependent Mean	1496.70736	Adj R-Sq	0.8503
Coeff Var	21.29051		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1849.44150	157.68429	-11.73	<.0001
no_pasg	no_pasg	1	-2.72366	1.52983	-1.78	0.0754
duration	duration	1	0.10077	0.24231	0.42	0.6776
speed_ground	speed_ground	1	40.14111	0.63545	63.17	<.0001
height	height	1	14.18358	1.21226	11.70	<.0001
pitch	pitch	1	31.62391	24.36237	1.30	0.1947
aircraft_dummy		1	-462.83505	24.81210	-18.65	<.0001

The overall model is statistically significant with a P-value <0.0001. The model has an R-Squared value of 0.8515 which means that this model explains 85.15% of the variability of the data around its mean. Even though this model has a lower r squared value, it looks better than the previous model since it uses 762 of the 811 observations. However, at 95% confidence level, only speed_ground, height and aircraft dummy are statistically significant. We could say that for every unit increase in the speed_ground, the landing distance increases by 40.14 feet. For every meter increase in height of the threshold of the runway, landing distance increases by 14.18 feet.

<u>Model 3</u>:

```
/*model without speed_air and duration*/
proc reg data=project.dummy;
model distance= no_pasg speed_ground height pitch aircraft_dummy;
title 'Model with no missing values(all variables but speed_air and duration)';
run;
```

Model with no missing values(all variables but speed_air and duration)

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	811
Number of Observations Used	811

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	462322047	92464409	916.52	<.0001
Error	805	81213922	100887		
Corrected Total	810	543535968			

Root MSE	317.62692	R-Square	0.8506
Dependent Mean	1480.20136	Adj R-Sq	0.8497
Coeff Var	21.45836		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1872.67777	145.72265	-12.85	<.0001
no_pasg	no_pasg	1	-3.26794	1.48363	-2.20	0.0279
speed_ground	speed_ground	1	40.08398	0.61720	64.95	<.0001
height	height	1	14.09436	1.16583	12.09	<.0001
pitch	pitch	1	50.25077	23.36543	2.15	0.0318
aircraft_dummy		1	-455.63420	23.99483	-18.99	<.0001

The overall model is statistically significant with a P-value <0.0001. The model has an R-Squared value of 0.8506 which means that this model explains 85.06% of the variability of the data around its mean. This model uses all the 811 observations and all the variables are statistically significant at 95% confidence level. Hence we can say, for one more passenger, the landing distance decreases by 3.2 feet. For every unit increase in the speed_ground, the landing distance increases by 40.08 feet. For every meter increase in height of the threshold of the runway, landing distance increases by 14.09 feet. For every unit increase in the pitch, the landing distance increases by 50.25 feet. If the aircraft is airbus, the landing distance would be 455 feet less than Boeing. Or if the aircraft was Boeing, the landing distance would be 455 feet more than airbus.

<u>Model 4:</u> In Model 4, I excluded the dummy variable.

```
/*model with no dummy*/
proc reg data=project.dummy;
model distance= speed_ground height pitch no_pasg;
title 'Model with speed_ground, height, pitch and No_pasg';
run;
```

Model with speed_ground, height, pitch and No_pasg

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	811
Number of Observations Used	811

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	425944620	106486155	729.88	<.0001
Error	806	117591348	145895		
Corrected Total	810	543535968			

Root MSE	381.96200	R-Square	0.7837
Dependent Mean	1480.20136	Adj R-Sq	0.7826
Coeff Var	25.80473		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2638.45195	168.39459	-15.67	<.0001
speed_ground	speed_ground	1	39.45321	0.74113	53.23	<.0001
height	height	1	13.23680	1.40091	9.45	<.0001
pitch	pitch	1	207.89423	26.26467	7.92	<.0001
no_pasg	no_pasg	1	-3.85435	1.78375	-2.16	0.0310

The overall model is still statistically significant but the r squared value is .7873. This is a weaker model as compared to the ones above.

<u>Conclusion:</u>

I chose model 3 as the best model for its model fit (r squared value), significance of the variables intercepts and overall statistical significance.

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

   My final model has 811 observations. The raw data of 950 flights contained around a 100 duplicates which had to be removed. With the remaining 850 observations, I applied filters to variables as given in the dataset description. Any values above or below the defined thresholds were deleted (as explained in chapter 1). I also looked at boxplots of each variable to identify outliers and removed those from my dataset as well. My final dataset ready for modeling has 811 observation. Also, the model 3 uses all 811 observations.

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

   All models show different factors that impact landing distance. However, Model 3, at 95% confidence level, shows that for one more passenger, the landing distance decreases by 3.2 feet. For every unit increase in the speed_ground, the landing distance increases by 40.08 feet. For every meter increase in height of the threshold of the runway, landing distance increases by 14.09 feet. For every unit increase in the pitch, the landing distance increases by 50.25 feet. If the aircraft is airbus, the landing distance would be 455 feet less than Boeing. Or if the aircraft was Boeing, the landing distance would be 455 feet more than airbus. However at 99% confidence interval, we can say that only speed_ground, height, and type of aircraft impacts the landing distance of the flight. Since pitch and no_pasg become statistically insignificant.

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

According to model 3, if the aircraft is airbus, the landing distance would be 455 feet less than Boeing. Or if the aircraft was Boeing, the landing distance would be 455 feet more than airbus. Also, looking at the ttest to check how different aircraft types are related to landing distance, we find out that (as also explained in the ttest section above) - the mean landing distance for airbus is 1304 feet whereas for boeing, it is 1685.6 feet. The standard deviation of Boeing is also greater than that of airbus showing that distance of boeing is more spread across its mean as compared to airbus. Since P value of Folded F is less than .05, we look at the Satterthwaite test. The p value os Satterthwaite test is less than .0001 which means that we can reject the null that the landing distance for the two different types of planes is significantly different.