BANA 6043 – 003 Stat Computing – Project

Name: Jeevan Sai Reddy Beedareddy
UCID: M12368208
Mail ID: beedarjy@mail.uc.edu

Background: Flight landing.
Motivation: To reduce the risk of landing overrun.
Goal: To study what factors and how they would impact the landing distance of a commercial flight.

Executive Summary:

1. Landing Distance is majorly impacted by Speed_Ground, Speed_Air and Height. Speed_Air has around 75% missing values
2. Landing Distance also varies based on type of aircraft. For Airbus, mean Landing Distance is around 1300 meters where as for Boeing mean Landing Distance is around 1750. This difference of around 450 between two is statistically significant
3. Since, Landing Distance is dependent upon type of aircraft as well, we decided to build on two models one for Airbus and the other for Boeing. Number of observations for each aircraft is also almost same and hence building two models might not create any biases
4. Below is the final model equation for Airbus
   Distance = -2522.89061 + 42.55420*speed_ground + 14.09773*height
5. Below is the final model equation for Boeing
   Distance = -2008.46764 + 42.28538*speed_ground + 14.19682*height

Questions mentioned:

1. How many observations (flights) do you use to fit your **final** model? If not all 950 flights, why?
A. There were 831 observations in the final model. This was further split in to Airbus and Boeing.
   Removing exact duplicates reduced number of observations from 950 to 850
   Removing observations with abnormal values reduced number of observations from 850 to 831. Given there were very few abnormal values, treating them as an error and removing will not affect model
2. What factors and how they impact the landing distance of a flight?
A. In the final model, we chose Speed_Ground and height to be impacting Landing Distance. Increment in Speed_Ground by around 42 MPH would increase landing distance by 1 meter and increment in height by around 14 meters would increase landing distance by 1 meter
3. Is there any difference between the two makes Boeing and Airbus?
A. Yes. There is difference between two makes. For Airbus, mean Landing Distance is around 1300 meters where as for Boeing mean Landing Distance is around 1750. This difference of around 450 between two is statistically significant

Chapter 1: Data Preparation and Data Cleaning:

After defining business outcomes, first step is to prepare data suited for analysis. In order to do the same we need to follow below steps:

a. Combine data sets: Aggregate data from multiple sources and create one master data set. Call it as analytical data
b. Completeness check: Understand if there are any missing values in the data
c. Validity check: Understand if there are any abnormalities in the data
d. Clean data: Handle missing values and abnormalities
e. Summarize distribution

Part 1: Combine data sets:

Goal: We have two data sets provided FAA1 and FAA2. Need to combine the same and create one master dataset.

Process:

1. Load FAA1 dataset. There are 8 variables in the same. Check for any exact duplicates and remove exact duplicates i.e. if there are any two rows having exactly same values across all variables, keep only one row. This need to be done because having duplicate values might impact analysis results. Call it as FAA1_DATA.
2. Load FAA2 dataset. There are 7 variables in the same. Check for any exact duplicates and remove exact duplicates i.e. if there are any two rows having exactly same values across all variables, keep only one row. This need to be done because having duplicate values might impact analysis results. Call it as FAA2_DATA. Please note that there were around 50 observations where there was no data for any variable
3. Combine FAA1 and FAA2. In order to concatenate FAA1 and FAA2, generally it is preferred to have same number of variables in both data sets. Since, we don't have duration in FAA2 data set, create one temporary dataset from FFA1 by removing duration and concatenate this temporary dataset with FAA2. Call this concatenated dataset to be FAA1_FAA2_COMBINED. Also remove any exact duplicates from FAA1_FAA2_COMBINED. This step would give us around 851 observations and 7 variables
4. Attach duration to FAA1_FAA2_COMBINED: Create a primary key in FAA1_DATA (this has duration variable) by concatenating all variables other than duration i.e. AIRCRAFT, NO_PASG, SPEED_GROUND, SPEED_AIR, HEIGHT, PITCH, DISTANCE. Create another temporary dataset from FAA1_DATA with only primary key and duration. Call it as FFA1_DATA_PRIM_KEY. Similarly create primary key for FAA1_FAA2_COMBINED by concatenating variables AIRCRAFT, NO_PASG, SPEED_GROUND, SPEED_AIR, HEIGHT, PITCH, DISTANCE. Merge FAA1_FAA2_COMBINED and FAA1_DATA_PRIM_KEY by primary key created. Call it as FAA1_FAA2_COMBINED_V3
5. Remove rows where all observations have missing values. Call it as FAA1_FAA2_N_MISS. This would give us 850 observations and 9 variables including primary key created.

SAS Code:

```
/*------------ Chapter 1: Importing data in to SAS and combining datasets ------------------*/
/*------------ Import data set in to SAS ---------------------------------------------*/

proc import out=FAA1_DATA
        datafile="/folders/myfolders/STAT Computing/FAA1.xls"
        dbms=xls replace;
  sheet='FAA1';
run;

PROC PRINT DATA=FAA1_DATA;
RUN;

/*--------------- There might be cases where exact duplicates are present ---------------*/
/*--------------- Remove exact duplicates from imported data set ------------------------*/

proc sort data=FAA1_DATA
  out=FAA1_DATA
  NODUPRECS;
  by aircraft duration no_pasg speed_ground speed_air height pitch distance;
run;

PROC PRINT DATA=FAA1_DATA;
RUN;

/*---------------- Import FAA2 in to SAS ---------------------------------------------*/
proc import out=FAA2_DATA
        datafile="/folders/myfolders/STAT Computing/FAA2.xls"
        dbms=xls replace;
  sheet='FAA2';
run;

PROC PRINT DATA=faa2_data;
RUN;

/*--------------- Remove exact duplicates from imported data set ------------------------*/

proc sort data=FAA2_DATA
  out=FAA2_DATA
  NODUPRECS;
  by aircraft no_pasg speed_ground speed_air height pitch distance;
run;

PROC PRINT DATA=FAA2_DATA;
RUN;

/*-------------- Upon observing, FAA2 doesn't have duration. Hence first let's combine
```

```
data sets with out duration column ----------------------------------------------------*/
/*-------------- Create another data set without taking duration ------------------------*/
DATA FAA1_DATA_V2;
SET FAA1_DATA;
DROP DURATION;
RUN;

PROC PRINT DATA=faa1_data_v2;
RUN;

/*------- Combining FAA1 and FAA2 data sets. Here we are doing simple concatenation ------*/
DATA FAA1_FAA2_COMBINED;
	SET FAA1_DATA_V2 FAA2_DATA;
RUN;

PROC PRINT DATA=FAA1_FAA2_COMBINED;
RUN;

/*------- Removing exact duplicates from combined data set ------------------------------*/
proc sort data=FAA1_FAA2_COMBINED
  out=FAA1_FAA2_COMBINED
  NODUPRECS;
  by aircraft no_pasg speed_ground speed_air height pitch distance;
run;

PROC PRINT DATA=FAA1_FAA2_COMBINED;
RUN;

/*--------- Creating Primary Key for combined data set --------------------------------*/

DATA FAA1_FAA2_COMBINED_V2;
SET FAA1_FAA2_COMBINED;
length PRIM_KEY $ 5000;
PRIM_KEY = catx('__', AIRCRAFT, NO_PASG, SPEED_GROUND, SPEED_AIR, HEIGHT, PITCH, DISTANCE);
put _all_;
RUN;

PROC PRINT DATA=faa1_faa2_combined_v2;
RUN;



/*------- There is duration column in FAA1_DATA. Need to extract the same
and join back to above created data set ---------------------------------------------*/

DATA FAA1_DATA_PRIM_KEY;
SET FAA1_DATA;
length PRIM_KEY $ 5000;
```

```
PRIM_KEY = catx('__', AIRCRAFT, NO_PASG, SPEED_GROUND, SPEED_AIR, HEIGHT, PITCH, DISTANCE);
put _all_;
KEEP PRIM_KEY DURATION;
RUN;

PROC PRINT DATA = FAA1_DATA_PRIM_KEY;
RUN;

/*------- Sort two data sets before doing merge ---------------------------------------*/

PROC SORT DATA=faa1_faa2_combined_v2;
BY PRIM_KEY;
RUN;

PROC SORT DATA=faa1_data_prim_key;
BY PRIM_KEY;
RUN;

/*------- Merge two data sets by PRIM_KEY created ---------------------------------------*/
DATA FAA1_FAA2_COMBINED_V3;
MERGE FAA1_FAA2_COMBINED_V2 FAA1_DATA_PRIM_KEY;
BY PRIM_KEY;
RUN;

PROC PRINT DATA=faa1_faa2_combined_v3;
RUN;

/*---------- Removing row which has all missing values -----------------------------*/
DATA FAA1_FAA2_N_MISS;
SET FAA1_FAA2_COMBINED_V3;
WHERE PRIM_KEY ^= '.__.__.__.__.__.';
RUN;

PROC PRINT DATA=faa1_faa2_n_miss;
RUN;
```

SAS Output:
Output dataset: FAA1_FAA2_N_MISS
Number of Variables: 9
Number of Observations: 850

Snapshot of output:

Obs	aircraft	no_pasg	speed_ground	speed_air	height	pitch	distance	PRIM_KEY
1	airbus	46	104.07757658	103.40921036	19.7157721	4.1043931104	2494.8046454	airbus__46__104.07757658__103.40921036__19.7157721__4.10
2	airbus	46	40.801786477	.	24.400127629	3.9682093233	620.09051196	airbus__46__40.801786477__.__24.400127629__3.9682093233_
3	airbus	48	61.570704648	.	21.785707448	4.3511947442	560.53392302	airbus__48__61.570704648__.__21.785707448__4.3511947442_
4	airbus	50	84.219908138	.	32.542946798	3.318828622	1485.4400456	airbus__50__84.219908138__.__32.542946798__3.318828622__
5	airbus	51	62.484050366	.	26.53804471	3.8228939729	749.48028928	airbus__51__62.484050366__.__26.53804471__3.8228939729__
6	airbus	51	83.630692914	.	23.302265488	4.5566399591	1460.4181796	airbus__51__83.630692914__.__23.302265488__4.5566399591_
7	airbus	52	72.036625004	.	24.740341243	3.6279838777	648.02156805	airbus__52__72.036625004__.__24.740341243__3.6279838777_
8	airbus	52	73.761115944	.	9.688307724	3.3585464091	554.16098701	airbus__52__73.761115944__.__9.688307724__3.3585464091__
9	airbus	52	89.577029476	.	35.463228123	3.834651479	1390.8995718	airbus__52__89.577029476__.__35.463228123__3.834651479__
10	airbus	54	50.903105868	.	35.729484049	4.5440403076	597.98554514	airbus__54__50.903105868__.__35.729484049__4.5440403076_
11	airbus	54	67.456935552	.	41.334169856	3.8581993926	877.06227359	airbus__54__67.456935552__.__41.334169856__3.8581993926_
12	airbus	54	80.24779883	.	48.426731903	3.289757889	1303.6900358	airbus__54__80.24779883__.__48.426731903__3.289757889__1
13	airbus	54	83.071912777	.	37.317578277	3.4734612582	1338.6101651	airbus__54__83.071912777__.__37.317578277__3.4734612582_
14	airbus	54	86.425045711	.	14.748572684	3.5418381552	1476.177543	airbus__54__86.425045711__.__14.748572684__3.5418381552_
15	airbus	55	68.751529748	.	48.277120042	4.2626359629	1079.1170993	airbus__55__68.751529748__.__48.277120042__4.2626359629_
16	airbus	56	73.974086384	.	32.455027763	3.0805850946	769.49665785	airbus__56__73.974086384__.__32.455027763__3.0805850946_
17	airbus	56	86.528840828	.	40.94901507	3.7270256473	1437.6338566	airbus__56__86.528840828__.__40.94901507__3.7270256473__
18	airbus	57	88.418098446		45.02439155	3.7036944046	1616.3360538	airbus 57 88.418098446 45.02439155 3.7036944046

Part 2: Completeness Check:

Goal: In this step we would be checking for missing values amongst all variables. PROC UNIVARIATE can be used to check for missing values for all numerical variables. Use PROC FREQ to check for missing values for categorical variables. There is only one categorical variable i.e. AIRCRAFT.

SAS Code:

```
/*Step 2: Checking for Missing values, outliers and any abnormalities in data ---------*/
/*---------- Doing univariate analysis. This helps us understand missing values -------*/

/*---------- Univariate would give us missing percentage as well as summary stats
for all numerical variables ----------------------------------------------------------*/

PROC UNIVARIATE DATA=FAA1_FAA2_N_MISS PLOT;
RUN;

/*----- Do frequency distribution to understand missing values for categorical variables*/
/*----- Only aircraft is categorical variable ----------------------------------------*/

PROC FREQ DATA=FAA1_FAA2_N_MISS;
RUN;

/*---------- Below are missing value percentages --------------------------------------*/
/*---------- speed_air - 643 ~ 75% --------------------------------------------------*/
/*---------- Duration - 50 ~ 5% ----------------------------------------------------*/
```

SAS Output:

Only Speed_Air and Duration have missing values. All other variables don't have any. Below is the part of output where missing values are presented

Speed_Air:

Missing Values			
Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	642	75.53	100.00

Duration:

Missing Values			
Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	50	5.88	100.00

Frequency distribution for aircraft:

The FREQ Procedure

aircraft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
airbus	450	52.94	450	52.94
boeing	400	47.06	850	100.00

Part 3: Validity Check

Goal: Check for any abnormalities in the data. Below are rules for checking abnormalities

1. Duration needs to greater than 40 mins. Else abnormal
2. Ground speed needs to be in the range of 30 – 140 MPH. Else abnormal
3. Air speed needs to be in the range of 30 – 140 MPH. Else abnormal
4. Height needs to grater than 6 meters. Else abnormal
5. Distance needs to be less than 6000 feet. Else abnormal

SAS Code:

```
/*---------- Using limits provided, checking for abnormalities in data -----------------*/
/*---------- Since speed_air has so many missing values, make sure that only non-missing
values are taken in to consideration -------------------------------------------------*/

DATA FAA1_FAA2_N_MISS;
SET FAA1_FAA2_N_MISS;
IF DURATION > 40 OR (DURATION = .) THEN DURATION_MEASURE = 'NORMAL';
ELSE DURATION_MEASURE = 'ABNORMAL';
IF SPEED_GROUND >= 30 AND SPEED_GROUND <= 140 THEN SPEED_GROUND_MEASURE = 'NORMAL';
ELSE SPEED_GROUND_MEASURE = 'ABNORMAL';
IF (SPEED_AIR >= 30 AND SPEED_AIR <= 140) OR (SPEED_AIR = .) THEN SPEED_AIR_MEASURE =
'NORMAL';
ELSE SPEED_AIR_MEASURE = 'ABNORMAL';
IF HEIGHT >= 6 THEN HEIGHT_MEASURE = 'NORMAL';
ELSE HEIGHT_MEASURE = 'ABNORMAL';
IF DISTANCE <= 6000 THEN DISTANCE_MEASURE = 'NORMAL';
ELSE DISTANCE_MEASURE = 'ABNORMAL';
RUN;

PROC PRINT DATA=FAA1_FAA2_N_MISS;
RUN;

PROC FREQ DATA=FAA1_FAA2_N_MISS;
TABLES DURATION_MEASURE SPEED_GROUND_MEASURE SPEED_AIR_MEASURE HEIGHT_MEASURE
DISTANCE_MEASURE;
RUN;
```

SAS Output:

Every variable has few abnormalities. Percentage of abnormalities is very minimal. Please note that and Duration Air Speed also have missing values. They need not be considered abnormal as per above definition.

The FREQ Procedure

DURATION_MEASURE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ABNORM	5	0.59	5	0.59
NORMAL	845	99.41	850	100.00

SPEED_GROUND_MEASURE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ABNORM	3	0.35	3	0.35
NORMAL	847	99.65	850	100.00

SPEED_AIR_MEASURE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ABNORM	1	0.12	1	0.12
NORMAL	849	99.88	850	100.00

HEIGHT_MEASURE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ABNORM	10	1.18	10	1.18
NORMAL	840	98.82	850	100.00

DISTANCE_MEASURE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ABNORM	2	0.24	2	0.24
NORMAL	848	99.76	850	100.00

Below are sample statistics before handling abnormalities and missing values:

Variable	Count	Mean	Standard Deviation	Minimum	Maximum	Count Missing	%Missing
distance	850	1526	928.56008	34.08078	6533	0	0.00%
no_pasg	850	60.10353	7.49314	29.00000	87.00000	0	0.00%
speed_ground	850	79.45232	19.05949	27.73572	141.21864	0	0.00%
speed_air	208	103.79772	10.25904	90.00286	141.72494	642	75.53%
height	850	30.14422	10.28773	-3.54625	59.94596	0	0.00%
pitch	850	4.00936	0.52883	2.28448	5.92678	0	0.00%
duration	800	154.00654	49.25923	14.76421	305.62171	50	5.88%

Part 4: Handle abnormalities and missing values

Goal: Since we found there are few abnormalities in our variables, need to handle the same. Also instances of abnormalities are very less such and hence can remove the same:

Only duration and Air Speed has missing values. Duration has around 5% and Air speed has around 75% missing values. For Duration, we can handle it by replacing missing values with median of non-missing values. For Air Speed, since there are 75% missing values, we need not handle the same and have it in current form

SAS Code:

```
/*---------- Using limits provided, checking for abnormalities in data -----------------*/
/*---------- Since speed_air has so many missing values, make sure that only non missing
values are taken in to consideration -----------------------------------------------*/

DATA FAA1_FAA2_N_MISS;
SET FAA1_FAA2_N_MISS;
IF DURATION > 40 OR (DURATION = .) THEN DURATION_MEASURE = 'NORMAL';
ELSE DURATION_MEASURE = 'ABNORMAL';
IF SPEED_GROUND >= 30 AND SPEED_GROUND <= 140 THEN SPEED_GROUND_MEASURE = 'NORMAL';
ELSE SPEED_GROUND_MEASURE = 'ABNORMAL';
IF (SPEED_AIR >= 30 AND SPEED_AIR <= 140) OR (SPEED_AIR = .) THEN SPEED_AIR_MEASURE =
'NORMAL';
ELSE SPEED_AIR_MEASURE = 'ABNORMAL';
IF HEIGHT >= 6 THEN HEIGHT_MEASURE = 'NORMAL';
ELSE HEIGHT_MEASURE = 'ABNORMAL';
IF DISTANCE <= 6000 THEN DISTANCE_MEASURE = 'NORMAL';
ELSE DISTANCE_MEASURE = 'ABNORMAL';
RUN;

PROC PRINT DATA=FAA1_FAA2_N_MISS;
RUN;
```

```
PROC FREQ DATA=FAA1_FAA2_N_MISS;
TABLES DURATION_MEASURE SPEED_GROUND_MEASURE SPEED_AIR_MEASURE HEIGHT_MEASURE
DISTANCE_MEASURE;
RUN;

/*----------- Now let's delete abnormal values ----------------------------------*/

DATA FAA1_FAA2_N_MISS_V2;
SET FAA1_FAA2_N_MISS;
IF DURATION_MEASURE = 'ABNORM' OR SPEED_GROUND_MEASURE = 'ABNORM' OR
SPEED_AIR_MEASURE
= 'ABNORM' OR HEIGHT_MEASURE = 'ABNORM' OR DISTANCE_MEASURE = 'ABNORM' THEN DELETE;
RUN;

/*---------- Rechecking for missing values ----------------------------------------*/

PROC FREQ DATA=FAA1_FAA2_N_MISS_V2;
TABLES SPEED_GROUND_MEASURE SPEED_AIR_MEASURE HEIGHT_MEASURE DISTANCE_MEASURE;
RUN;

/*---------- Handling Missing Values for duration ----------------------------------*/
proc stdize data = FAA1_FAA2_N_MISS_V2
reponly method = MEDIAN out = FAA1_FAA2_N_MISS_V3;
var DURATION;
run;
```

Output:

We can do PROC UNIVARIATE to check if there are still missing values. Now, there shouldn’t be any missing values in DURATION. Also total number of observations would be around 831
Below are sample statistics after handling abnormalities and missing values:

Variable	Count	Mean	Standard Deviation	Minimum	Maximum	Count Missing	%Missing
distance	831	1522	896.33815	41.72231	5382	0	0.00%
no_pasg	831	60.05535	7.49132	29.00000	87.00000	0	0.00%
speed_ground	831	79.54270	18.73568	33.57410	132.78468	0	0.00%
speed_air	831	103.48504	9.73628	90.00286	132.91146	642	77.25%
height	831	30.45787	9.78481	6.22752	59.94596	0	0.00%
pitch	831	4.00516	0.52657	2.28448	5.92678	0	0.00%
duration	831	154.74617	46.87113	41.94937	305.62171	0	5.88%

Part 5: Summarize distribution

Goal: Plot distributions to get some sense of data

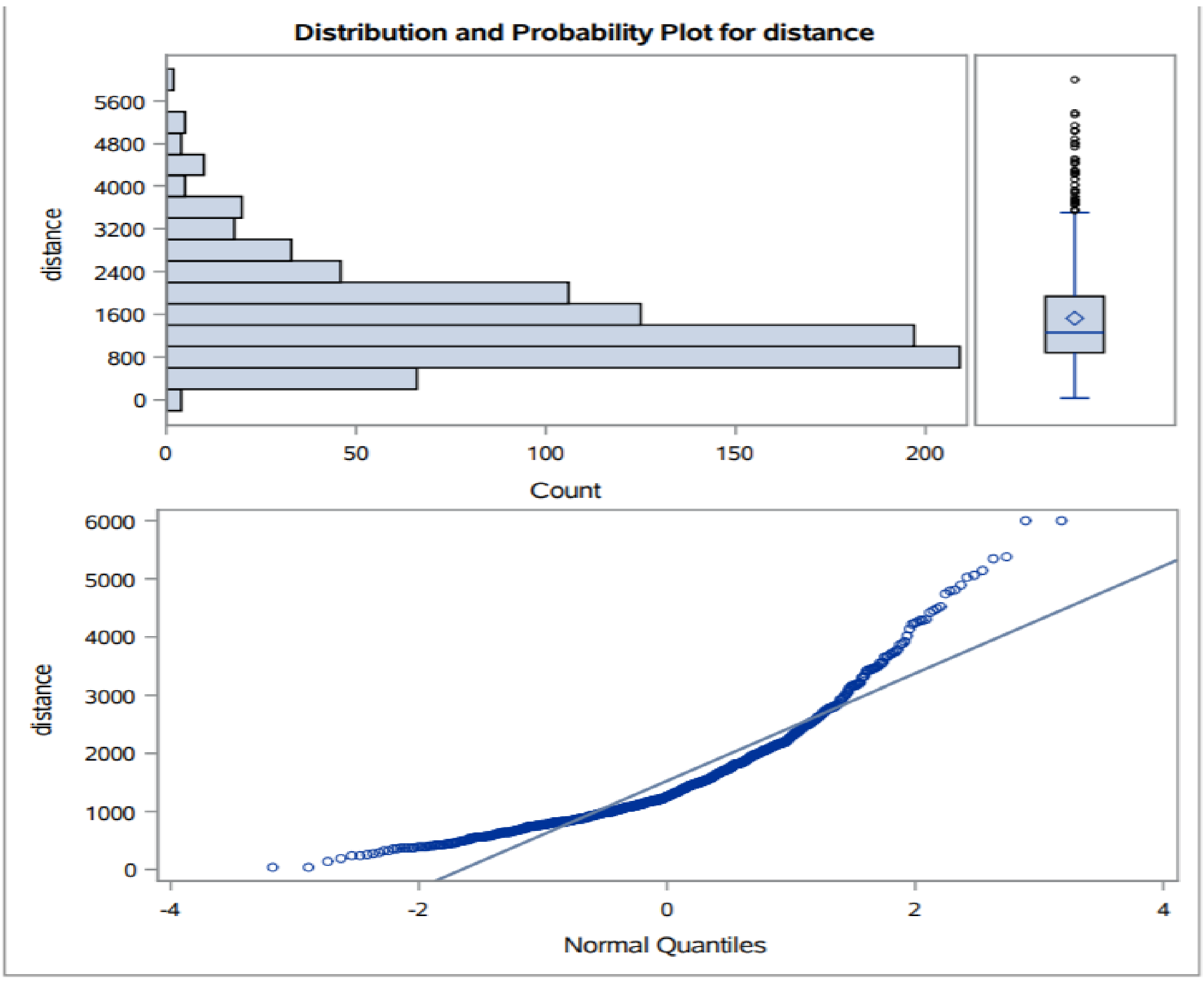
This can be done with PROC Univariate. Please check for normality of data and also box plots

SAS Code:

```
PROC UNIVARIATE data= FAA1_FAA2_N_MISS_V3 PLOT;
RUN;
```

Output:

For Distance is skewed towards right because of presence of very high distances. Also remember that we did handle for abnormalities in distance. We need not treat for outliers given that there is a possibility of high distances.



All other variables are having almost normal distribution

Chapter 2: Exploratory Data Analysis

After cleaning data, we need to perform Exploratory Data Analysis to get insights regarding data.

Part 1: Bi-Variate Analysis

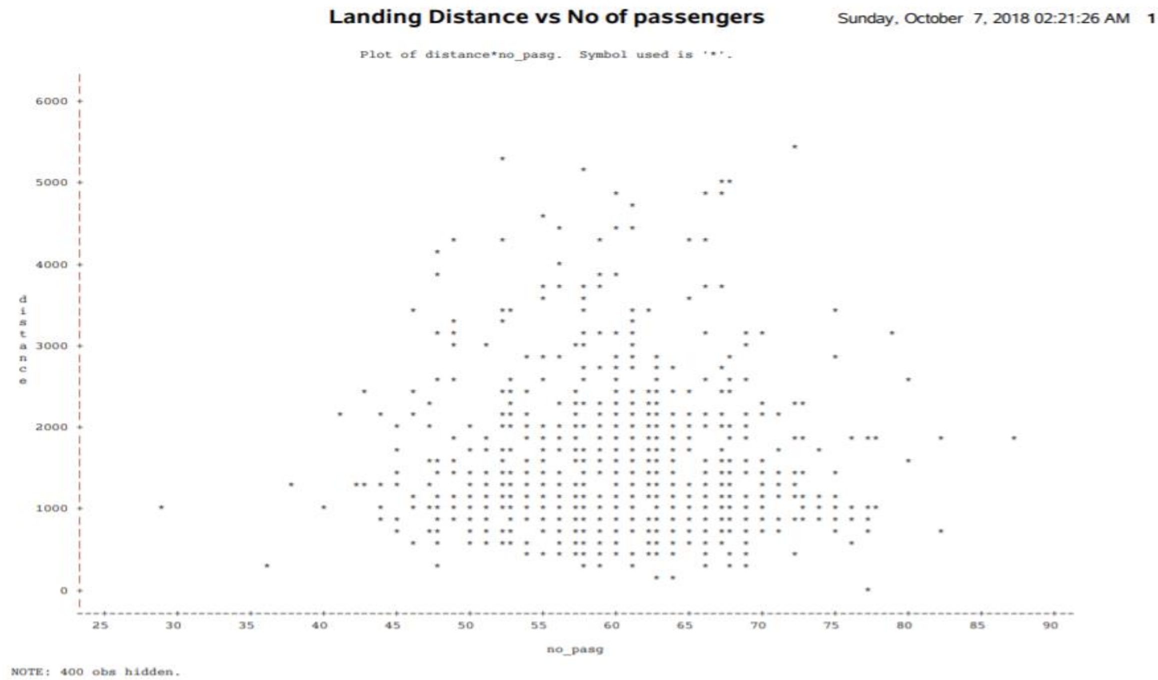
Goal: To understand relationship between landing distance and all independent variables

SAS Code:

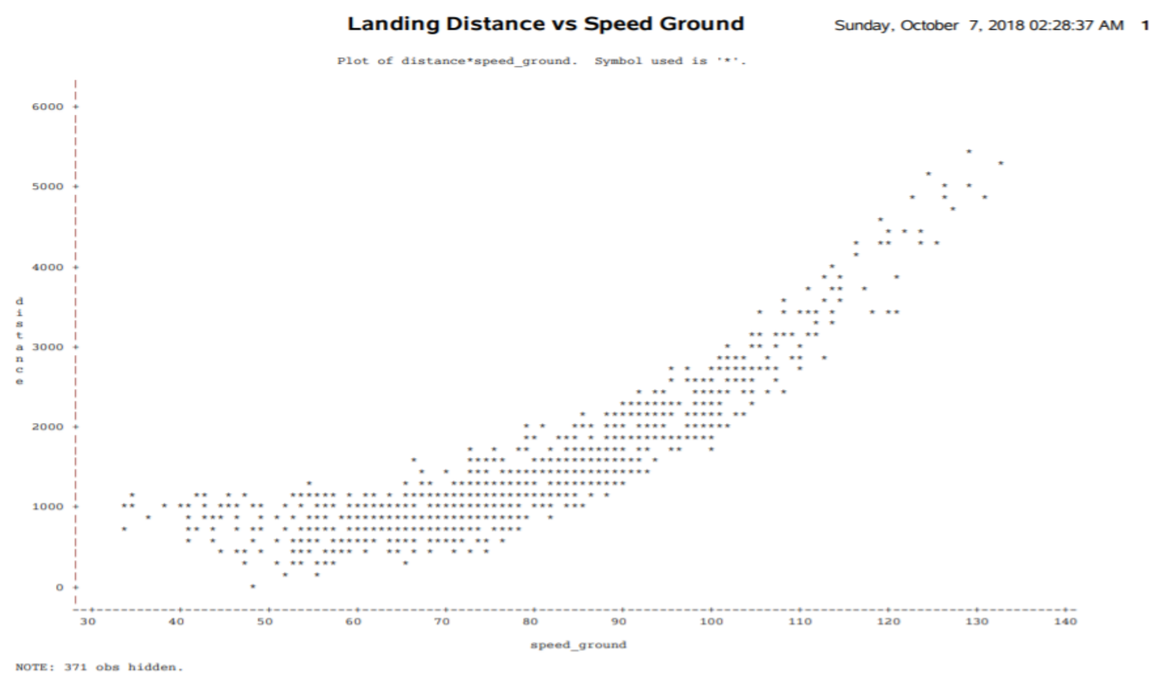
```
/*------ Plotting landing distance against all independent variables ---------------------------------------------*/

title "Landing Distance vs No of passengers";
proc plot data=FAA1_FAA2_N_MISS_V3;
plot distance*no_pasg = '*';
run;

title "Landing Distance vs Speed Ground";
proc plot data=FAA1_FAA2_N_MISS_V3;
plot distance*speed_ground = '*';
run;

title "Landing Distance vs Speed Air";
proc plot data=FAA1_FAA2_N_MISS_V3;
plot distance*speed_air = '*';
run;

title "Landing Distance vs Height";
proc plot data=FAA1_FAA2_N_MISS_V3;
plot distance*height = '*';
run;

title "Landing Distance vs Pitch";
proc plot data=FAA1_FAA2_N_MISS_V3;
plot distance*pitch = '*';
run;

title "Landing Distance vs Duration";
proc plot data=FAA1_FAA2_N_MISS_V3;
plot distance*duration = '*';
run;
```

Output:

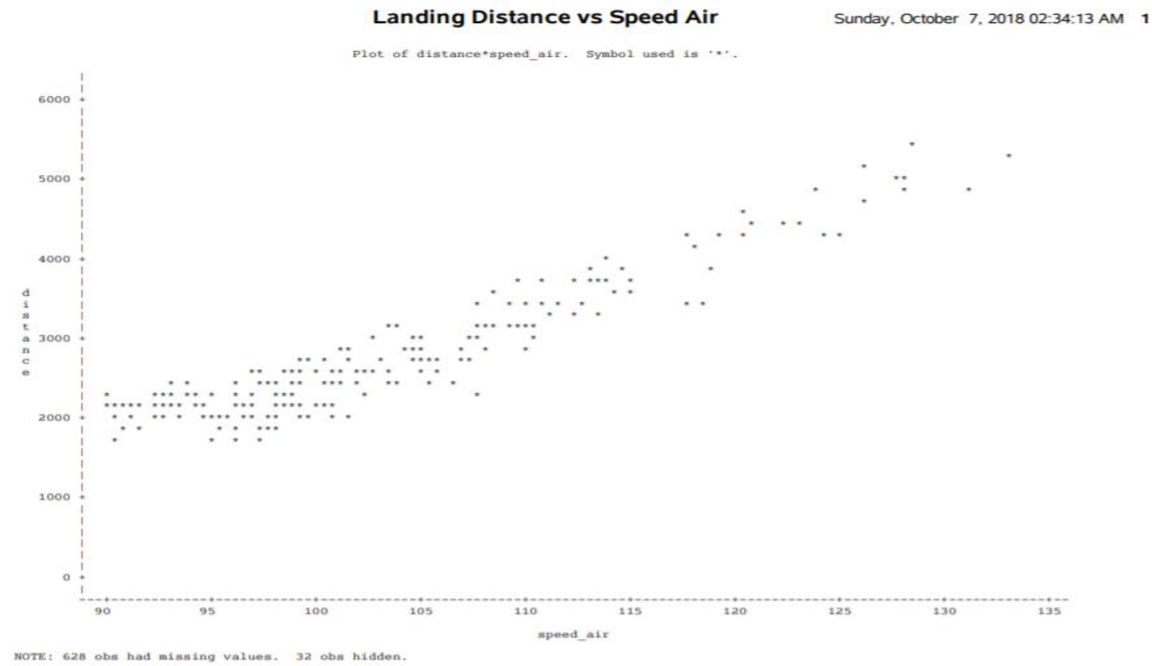
Landing Distance vs No of Passengers: There is no evident linear relationship



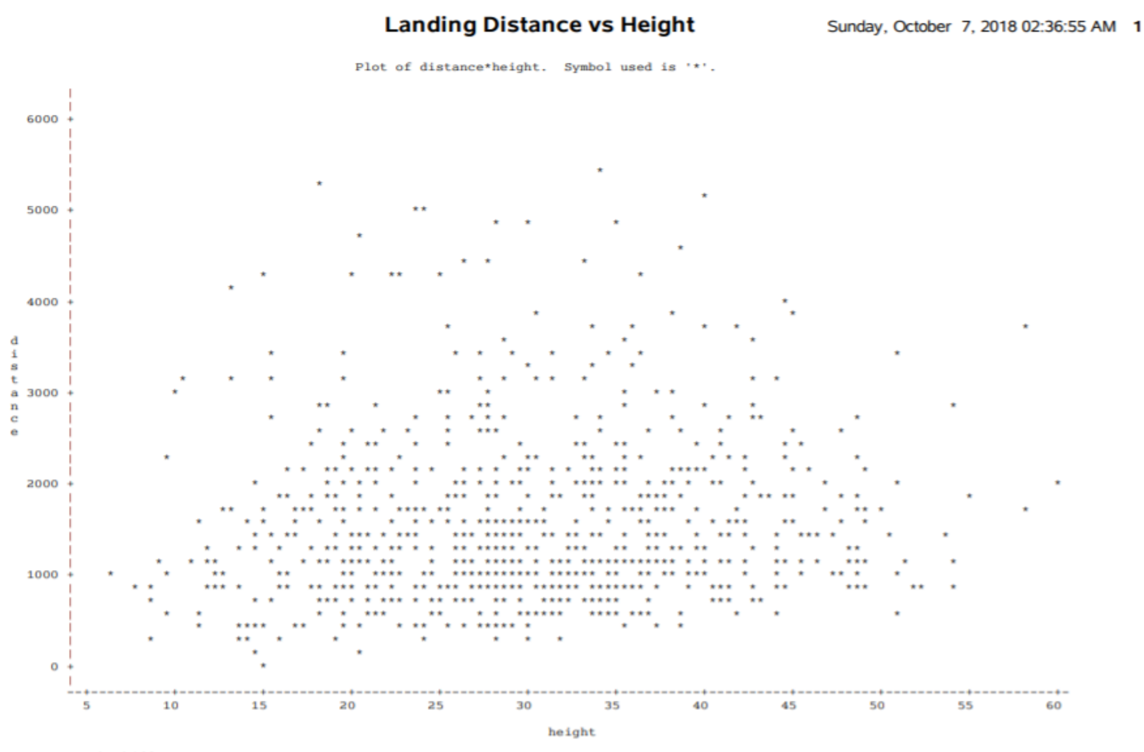
Landing Distance vs Speed Ground: There is evident linear relationship



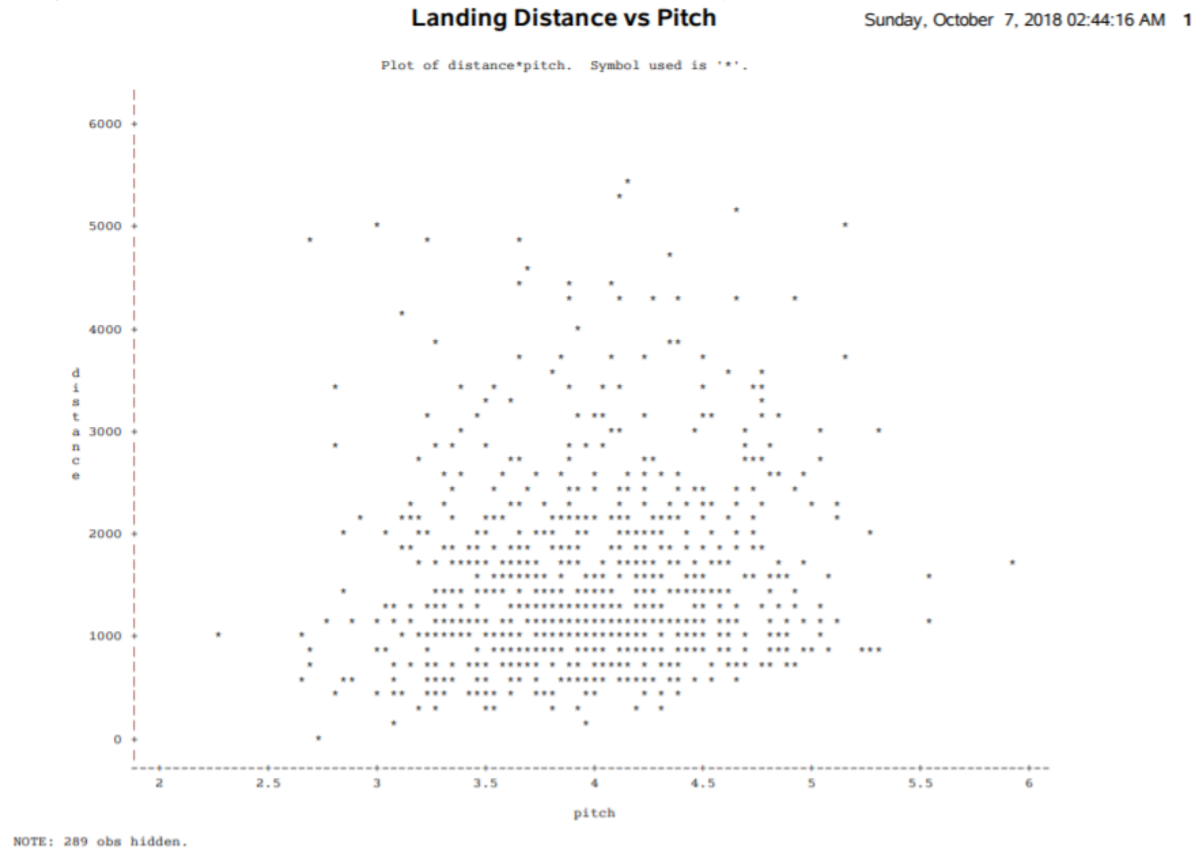
Landing Distance vs Speed Air: There is evident linear relationship for values that are not missing



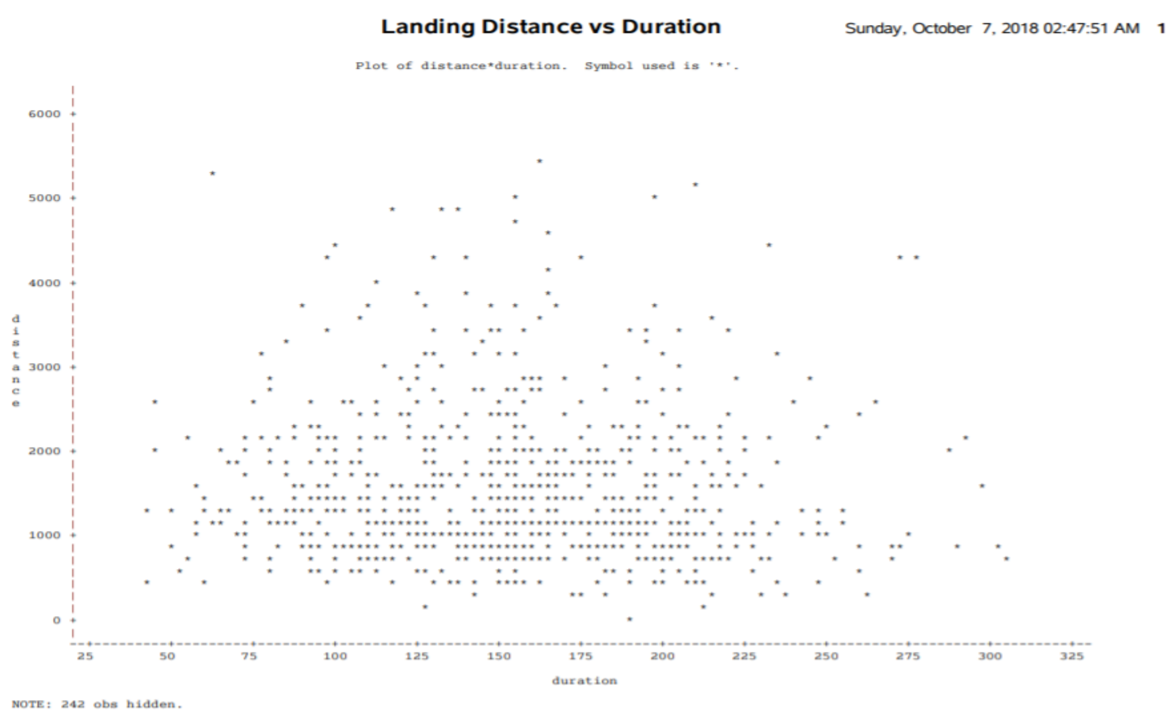
Landing Distance vs Height: There is no evident linear relationship



Landing Distance vs Pitch: There is no evident linear relationship



Landing Distance vs Duration: There is no evident linear relationship



Part 2: To understand if make of aircraft has any impact on landing distance

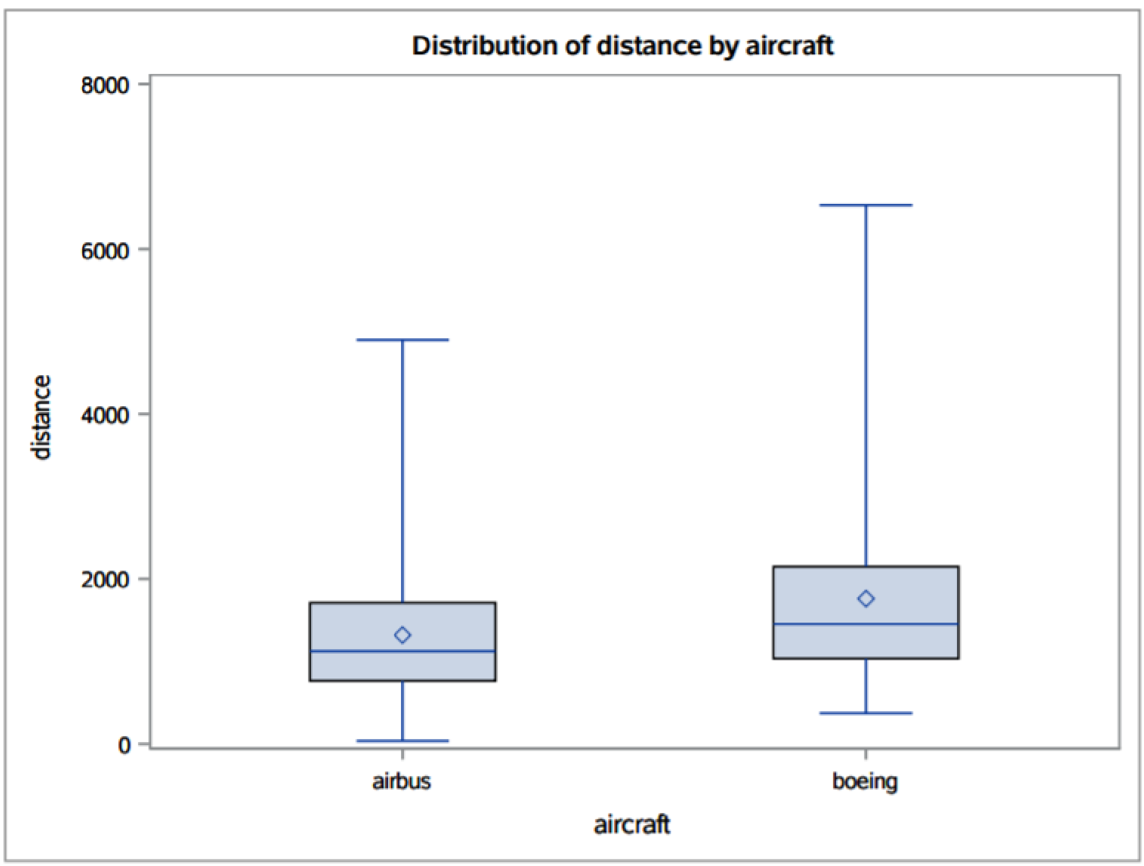
Goal: To study landing distance by aircraft and to conclude impact of the same. Here we will be checking for box-plot as well as performing two sample t-test

SAS Code:

```
/*--- To plot box-plot of distance across types of aircrafts -----------------*/
PROC BOXPLOT DATA=FAA1_FAA2_COMBINED_V3;
PLOT DISTANCE*AIRCRAFT;
TITLE TO UNDERSTAND DIFFERENCES BETWEEN AIRBUS AND BOEING;
RUN;

/*--- Performing Two Sample T-Test ----------------------------------------*/
PROC TTEST DATA=FAA1_FAA2_COMBINED_V3;
CLASS AIRCRAFT;
VAR DISTANCE;
RUN;
```

Output:

TO UNDERSTAND DIFFERENCES BETWEEN AIRBUS AND BOEING



The TTEST Procedure

Variable: distance (distance)

aircraft	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus		450	1318.2	792.3	37.3516	34.0808	4896.3
boeing		400	1759.8	1012.2	50.6123	371.3	6533.0
Diff (1-2)	Pooled		-441.7	902.5	62.0193		
Diff (1-2)	Satterthwaite		-441.7		62.9027		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1318.2	1244.8	1391.6	792.3	743.7	847.8
boeing		1759.8	1660.3	1859.3	1012.2	946.6	1087.7
Diff (1-2)	Pooled	-441.7	-563.4	-319.9	902.5	861.5	947.6
Diff (1-2)	Satterthwaite	-441.7	-565.1	-318.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	848	-7.12	<.0001
Satterthwaite	Unequal	753.38	-7.02	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	399	449	1.63	<.0001

From box-plot as well as two sample T-Test, it is evident there is a difference between landing distances of two types of air-crafts. Boeing has high landing distances and also it is statistically significant from Two sample T-Test

Since it is evident that two types of aircrafts have different landing distances, we would be required two separate models for each type of air-craft. Also, both types of aircrafts have around similar number of observations. Hence two models might not create any biases in the output

Chapter 3: Model Iterations and Model Selection

Goal: From chapter 2, we concluded that we would be building two models each for Airbus and Boeing. In order to do the same, we need to iterate on first few basic models and select the one that fits best in the end.

Part 1: Check for Correlation

Goal: Checking for correlation helps us understand impact of all dependent variables on landing distance. Also, there has to minimum / zero correlation amongst independent variables. Building the correlation matrix also helps us identify such cases.

For Airbus:

SAS Code:

```
/*--- Generating correlation matrix for Airbus ------------------------------*/
PROC CORR DATA=FAA1_FAA2_N_MISS_V3;
VAR DISTANCE NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DURATION;
WHERE AIRCRAFT = 'airbus';
RUN;
```

SAS Output:

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

	distance	no_pasg	speed_ground	speed_air	height	pitch	duration
distance distance	1.00000 444	-0.00732 0.8777 444	0.90520 <.0001 444	0.96411 <.0001 85	0.14494 0.0022 444	0.07330 0.1230 444	-0.07420 0.1185 444
no_pasg no_pasg	-0.00732 0.8777 444	1.00000 444	0.00906 0.8491 444	-0.06372 0.5623 85	0.02367 0.6189 444	-0.11802 0.0128 444	-0.02323 0.6254 444
speed_ground speed_ground	0.90520 <.0001 444	0.00906 0.8491 444	1.00000 444	0.98169 <.0001 85	-0.03346 0.4819 444	-0.00493 0.9176 444	-0.05654 0.2345 444
speed_air speed_air	0.96411 <.0001 85	-0.06372 0.5623 85	0.98169 <.0001 85	1.00000 85	-0.00546 0.9604 85	0.00007 0.9995 85	0.01523 0.8900 85
height height	0.14494 0.0022 444	0.02367 0.6189 444	-0.03346 0.4819 444	-0.00546 0.9604 85	1.00000 444	0.05128 0.2809 444	-0.01227 0.7966 444
pitch pitch	0.07330 0.1230 444	-0.11802 0.0128 444	-0.00493 0.9176 444	0.00007 0.9995 85	0.05128 0.2809 444	1.00000 444	-0.04102 0.3885 444
duration duration	-0.07420 0.1185 444	-0.02323 0.6254 444	-0.05654 0.2345 444	0.01523 0.8900 85	-0.01227 0.7966 444	-0.04102 0.3885 444	1.00000 444

It is clear that speed_ground and speed_air are highly correlated with distance. But speed_air has only 85 observations and is also highly correlated with speed_ground. Hence, we can choose only speed_ground instead of speed_air.

For Boeing:

SAS Code:

```
/*--- Generating correlation matrix for Boeing ------------------------------*/
PROC CORR DATA=FAA1_FAA2_N_MISS_V3;
VAR DISTANCE NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DURATION;
WHERE AIRCRAFT = 'boeing';
RUN;
```

SAS Output:

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

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

It is clear that speed_ground and speed_air are highly correlated with distance. But speed_air has only 118 observations and is also highly correlated with speed_ground. Hence, we can choose only speed_ground instead of speed_air

Part 2: Model Iterations and Model Selection

In order to finalize on model and final equation, we need to perform few iterations. In each iteration, we will be removing one variable and will be checking for change in R-Square. Adding too many variables which don't contribute to R-Square leads to overfitting. Hence need to keep only variables that would impact R-Square

For airbus:
SAS Code:

```
/*--- Model Iterations and Selection of Model -----------------------------*/
/*--- Iter 1 -------------------------------------------------------------*/
```

```
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = NO_PASG SPEED_GROUND HEIGHT PITCH DURATION;
WHERE AIRCRAFT = 'airbus';
RUN;

/*--- Iter 2 ----------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = SPEED_GROUND HEIGHT PITCH DURATION;
WHERE AIRCRAFT = 'airbus';
RUN;

/*--- Iter 3 ---------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = SPEED_GROUND HEIGHT DURATION;
WHERE AIRCRAFT = 'airbus';
RUN;

/*--- Iter 4 ---------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = SPEED_GROUND HEIGHT;
WHERE AIRCRAFT = 'airbus';
RUN;

/*--- Iter 5 ---------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = SPEED_GROUND;
WHERE AIRCRAFT = 'airbus';
RUN;
```

SAS Output:

Below is the table explaining variables in each iteration, R-Square

Iteration Number	Variables Selected	R-Square	Comments
Iter 1	NO_PASG, SPEED_GROUND, HEIGHT, PITCH, DURATION	0.8553	In this iteration, we didn’t consider SPPED_AIR owing to high correlation with SPEED_GROUND and high missing values
Iter 2	SPEED_GROUND HEIGHT PITCH DURATION	0.8552	Even with removing NO_PASG, R-Square didn’t change significantly. Hence can filter out NO_PASG
Iter 3	SPEED_GROUND HEIGHT DURATION	0.8506	Even with removing PITCH, R-Square didn’t change significantly. Hence can filter out PITCH
Iter 4	SPEED_GROUND HEIGHT	0.8501	Even with removing DURATION, R-Square didn’t change significantly. Hence can filter out DURATION

Iter 5	SPEED_GROUND	0.8194	By removing HEIGHT, R-Square changed significantly. Hence can't filter out the same

Hence, we will be choosing ITER 4 and below are few model characteristics

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

For every 42 MPH increment in speed_ground, distance increases by 1 meter and for every 14 meters increment in height, distance increases by 1 meter.

Hence, final equation for AIRBUS would be

Distance = -2522.89061 + 42.55420*speed_ground + 14.09773*height

For Boeing:
SAS Code:

```
/*--- Iter 1 ----------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = NO_PASG SPEED_GROUND HEIGHT PITCH DURATION;
WHERE AIRCRAFT = 'boeing';
RUN;

/*--- Iter 2 ------------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = SPEED_GROUND HEIGHT PITCH DURATION;
WHERE AIRCRAFT = 'boeing';
RUN;

/*--- Iter 3 -----------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
```

```
MODEL DISTANCE = SPEED_GROUND HEIGHT DURATION;
WHERE AIRCRAFT = 'boeing';
RUN;

/*--- Iter 4 --------------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = SPEED_GROUND HEIGHT;
WHERE AIRCRAFT = 'boeing';
RUN;

/*--- Iter 5 --------------------------------------------------------------------*/
PROC REG DATA=FAA1_FAA2_N_MISS_V3;
MODEL DISTANCE = SPEED_GROUND;
WHERE AIRCRAFT = 'boeing';
RUN;
```

SAS Output:

Below is the table explaining variables in each iteration, R-Square

Iteration Number	Variables Selected	R-Square	Comments
Iter 1	NO_PASG, SPEED_GROUND, HEIGHT, PITCH, DURATION	0.8330	In this iteration, we didn't consider SPPED_AIR owing to high correlation with SPEED_GROUND and high missing values
Iter 2	SPEED_GROUND HEIGHT PITCH DURATION	0.8327	Even with removing NO_PASG, R-Square didn't change significantly. Hence can filter out NO_PASG
Iter 3	SPEED_GROUND HEIGHT DURATION	0.8322	Even with removing PITCH, R-Square didn't change significantly. Hence can filter out PITCH
Iter 4	SPEED_GROUND HEIGHT	0.8317	Even with removing DURATION, R-Square didn't change significantly. Hence can filter out DURATION
Iter 5	SPEED_GROUND	0.8109	By removing HEIGHT, R-Square changed significantly. Hence can't filter out the same

Hence, we will be choosing ITER 4 and below are few model characteristics

Root MSE	392.36824	**R-Square**	0.8317
Dependent Mean	1750.98330	**Adj R-Sq**	0.8308
Coeff Var	22.40845		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2008.46764	104.75662	-19.17	<.0001
speed_ground	speed_ground	1	42.28538	0.97362	43.43	<.0001
height	height	1	14.19682	2.06276	6.88	<.0001

Hence final equation for BOEING would be

Distance = -2008.46764 + 42.28538*speed_ground + 14.19682*height