PROJECT– FLIGHT LANDING ANALYSIS

PROJECT ID - BANA 5143/6043

BY – SYED IMAD HUSAIN (M12958531)

Purpose - To reduce the risk of landing overrun

Summary

This report was commissioned to examine the factors impacting the landing distance of a commercial flight to reduce the risk of landing overrun. The research draws attention towards the fact that different factors impact the landing distance differently based on the built of the plane. To avoid Landing Overrun(Distance > 6000 feet), below factors must be within these recommended limits –

Type of Aircraft	Speed Air(miles per hour)	Height(meters)
Airbus	182	NA
Boeing	190	29

Methods of analyses include data-preparation, Pearson correlation matrix, scatter plots, least square method of approximation amongst others. All analyses were performed in SAS University Edition on the FAA data sets provided during the start of the project. The summarization of the impact of contributing factors on Distance is as follows:

Built	Factor	Unit Increase	Variance in Unit Increase
Airbus	Speed Air	41	1
Boeing	Speed Air	43	1
	Height	15	2

Contents

1) Chapter 1 - Data Preparation

a) Importing data into SAS

i) Goal

Import data to analyze the quality of data sets provided. Two data sets have been provided namely FAA1 and FAA2 in excel format

ii) SAS code

```
FILENAME REFFILE '/folders/myfolders/FAA/FAA1.xls';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLS
	OUT=FAA.FAA1;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=FAA.FAA1; RUN;
```

```
FILENAME REFFILE '/folders/myfolders/FAA/FAA2.xls';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLS
	OUT=FAA.FAA2;
	GETNAMES=YES;
RUN;

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

iii) SAS output

Data Set Name	FAA.FAA1	FAA.FAA2
Observations	800	200
Variables	8	7
Indexes	0	0
Observation Length	72	64
Deleted Observations	0	0
Compressed	NO	NO
Sorted	NO	NO

Alphabetic List of Variables and Attributes in FAA1					
#	Variable	Type	Len	Format	Informat
1	aircraft	Char	12	12	12
8	distance	Num	8	BEST12.	
2	duration	Num	8	BEST12.	
6	height	Num	8	BEST12.	
3	no_pasg	Num	8	BEST12.	
7	pitch	Num	8	BEST12.	
5	speed_air	Num	8	BEST12.	
4	speed_ground	Num	8	BEST12.	

Alphabetic List of Variables and Attributes in FAA2					
#	Variable	Type	Len	Format	Informat
1	aircraft	Char	12	12	12
7	distance	Num	8	BEST12.	
5	height	Num	8	BEST12.	
2	no_pasg	Num	8	BEST12.	
6	pitch	Num	8	BEST12.	
4	speed_air	Num	8	BEST12.	
3	speed_ground	Num	8	BEST12.	

iv) Observations

- FAA2 has the same name of all variables as FAA1
- Variable Duration is missing from FAA2 whereas it is present in FAA1
- All variables except for Aircraft are numeric

v) Conclusion

- These datasets can be merged together for analysis

- Missing variable Duration will be left blank during merging process which can later be transformed based on requirement

b) Combining data sets from different sources

i) Goal

- Similar procedure needs to be applied on the datasets
- Combining datasets reduces redundancy in process
- Singular data feed for the Analytical Model
- Adding additional variable called 'FEED' which represents the source of information. Although this variable is immaterial to the analytical model, combining datasets without this variable means we may lose information. Furthermore, this variable can be used to indicate the defective dataset to which an abnormal observation belong

ii) SAS code

```
DATA FAA.FAA1_COPY;
SET FAA.FAA1;
FEED = "FAA1";
RUN;

DATA FAA.FAA2_COPY;
SET FAA.FAA2;
FEED = "FAA2";
RUN;

DATA FAA.FAA_COMBINED;
SET FAA.FAA1_COPY FAA.FAA2_COPY;
RUN;

PROC CONTENTS DATA=FAA.FAA_COMBINED; RUN;
```

iii) SAS output

Data Set Name	FAA.FAA_COMBINED
Observations	1000
Variables	9
Indexes	0
Observation Length	72
Deleted Observations	0
Compressed	NO
Sorted	NO

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

iv) Observations

- There are 50 rows from faa2 which are totally blank
- Speed_air has many missing values
- Duration values are missing for faa2 observations
- Some variables have anomalous values like Height is negative for few observations

v) Conclusion

- Before data can be fed to the model, data quality checks like the following needs to be performed
    - Data completeness
    - Anomalies, Outliers
    - Missing values

c) Performing the completeness check of each

i) Goal

- It is important to identify missing values for variables since a higher percentage of missing values can skew the outcome
- Before we perform this check, it is important to remove the 50 blank rows which are erroneously a part of the combined data set
- Identify duplicates

ii) SAS code

```
DATA FAA.FAA_COMBINED_CLEAN;
 SET FAA.FAA_COMBINED;
IF CMISS(AIRCRAFT,DISTANCE,DURATION,HEIGHT,NO_PASG,PITCH,SPEED_AIR,SPEED_GROUND)<8;
RUN;
PROC FORMAT;
	VALUE _NMISSPRINT LOW-HIGH="NON-MISSING";
	VALUE $_CMISSPRINT " "=" " OTHER="NON-MISSING";
PROC FREQ DATA=FAA.FAA_COMBINED_CLEAN;
	TITLE3 "MISSING DATA FREQUENCIES";
	TITLE4 H=2 "LEGEND: ., A, B, ETC = MISSING";
	FORMAT DURATION NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DISTANCE
		_NMISSPRINT.;
	FORMAT AIRCRAFT FEED $_CMISSPRINT.;
	TABLES (AIRCRAFT DURATION NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DISTANCE
		FEED) / MISSING NOCUM;
RUN;
PROC SQL;
SELECT COUNT(*) FROM (
SELECT    AIRCRAFT, MAX(DURATION) AS DURATION, NO_PASG, SPEED_GROUND, SPEED_AIR,
          HEIGHT, PITCH, DISTANCE, MIN(FEED) AS FEED
FROM  FAA.FAA_COMBINED_CLEANED
GROUP BY   AIRCRAFT, NO_PASG, SPEED_GROUND, SPEED_AIR, HEIGHT, PITCH, DISTANCE );
RUN;
```

iii) SAS output

aircraft		
aircraft	Frequency	Percent
NON-MISSING	950	100

speed_air		
speed_air	Frequency	Percent
.	711	74.84
NON-MISSING	239	25.16

distance		
distance	Frequency	Percent
NON-MISSING	950	100

duration		
duration	Frequency	Percent
.	150	15.79
NON-MISSING	800	84.21

height		
height	Frequency	Percent
NON-MISSING	950	100

speed_ground		
speed_ground	Frequency	Percent
NON-MISSING	950	100

no_pasg		
no_pasg	Frequency	Percent
NON-MISSING	950	100

pitch		
pitch	Frequency	Percent
NON-MISSING	950	100

count
834

iv) Observations

- Speed air has 75% missing values which may indicate that this variable cannot effectively contribute to the analysis or we may want to apply some rationale to estimate its values
- Duration has 16% missing values which is relatively small and a simpler process like substituting average values for missing values may work. However, it still depends on the analytical model to make this choice or not
- There are duplicate observations in the combined dataset

v) Conclusion

- 50 blank rows from faa2 dataset which were part of final dataset were removed
- Missing value computation may be needed to be applied for variables Speed air and Duration
- This gives us a partial idea of data quality. However, further analysis of non-missing values needs to be performed to determine the data quality for modeling purposes

d) Performing the validity check for each variable

i) Goal

- We need to study the variables to identify if they are in an acceptable range of values as per the data dictionary
- This will help us identify the number of outliers in the data set

ii) SAS code

```
 DATA FAA.FAA_COMBINED_OUTLIER;
SET FAA.FAA_COMBINED_CLEAN;
KEEP
AIRCRAFT_CHECK
DISTANCE_CHECK
DURATION_CHECK
HEIGHT_CHECK
NO_PASG_CHECK
PITCH_CHECK
SPEED_AIR_CHECK
SPEED_GROUND_CHECK
;
IF  CMISS(aircraft)=1 THEN AIRCRAFT_CHECK=' MISSING ';ELSE IF aircraft = 'boeing' OR aircraft = 'airbus' THEN
AIRCRAFT_CHECK = 'IN RANGE'; ELSE AIRCRAFT_CHECK = 'ANOMALY';
IF  MISSING(distance)=1 THEN DISTANCE_CHECK = 'MISSING ';ELSE IF distance < 6000 THEN DISTANCE_CHECK = 'IN
RANGE';ELSE DISTANCE_CHECK = 'ANOMALY';
IF  MISSING(duration)=1 THEN DURATION_CHECK = 'MISSING ';ELSE IF duration > 40 THEN DURATION_CHECK = 'IN
RANGE' ; ELSE DURATION_CHECK = 'ANOMALY';
```

```
IF  MISSING(HEIGHT)=1 THEN HEIGHT_CHECK = 'MISSING ';ELSE IF height = 6 OR HEIGHT > 6 THEN HEIGHT_CHECK =
'IN RANGE'; ELSE HEIGHT_CHECK = 'ANOMALY';
IF  MISSING(no_pasg)=1 THEN NO_PASG_CHECK = 'MISSING ';ELSE IF no_pasg > 0 THEN NO_PASG_CHECK = 'IN RANGE';
ELSE NO_PASG_CHECK = 'ANOMALY';
IF  MISSING(pitch)=1 THEN PITCH_CHECK = 'MISSING ';ELSE IF pitch > 0 THEN PITCH_CHECK = 'IN RANGE'; ELSE
PITCH_CHECK = 'ANOMALY';
IF  MISSING(speed_air)=1 THEN SPEED_AIR_CHECK = 'MISSING ';ELSE IF speed_air < 140 OR speed_air > 30 THEN
SPEED_AIR_CHECK = 'IN RANGE'; ELSE SPEED_AIR_CHECK = 'ANOMALY';
IF  MISSING(speed_ground)=1 THEN SPEED_GROUND_CHECK = 'MISSING ';ELSE IF speed_ground < 140 OR
speed_ground > 30 THEN SPEED_GROUND_CHECK = 'IN RANGE'; ELSE SPEED_GROUND_CHECK = 'ANOMALY';
RUN;

PROC SQL;
CREATE TABLE FAA.OUTLIER_SUMMARY AS
SELECT VARIABLE,SUM(MISSING) AS MISSING, SUM(IN_RANGE) AS IN_RANGE, SUM(ANOMALY) AS ANOMALY
FROM
(SELECT 'AIRCRAFT' AS VARIABLE,
CASE WHEN AIRCRAFT_CHECK = 'MISSING ' THEN 1 ELSE 0 END AS MISSING,
CASE WHEN AIRCRAFT_CHECK = 'IN RANGE' THEN 1 ELSE 0 END AS IN_RANGE,
CASE WHEN AIRCRAFT_CHECK = 'ANOMALY' THEN 1 ELSE 0 END AS ANOMALY
FROM FAA.FAA_COMBINED_OUTLIER
UNION ALL
SELECT 'DISTANCE' AS VARIABLE,
CASE WHEN DISTANCE_CHECK = 'MISSING ' THEN 1 ELSE 0 END AS MISSING,
CASE WHEN DISTANCE_CHECK = 'IN RANGE' THEN 1 ELSE 0 END AS IN_RANGE,
CASE WHEN DISTANCE_CHECK = 'ANOMALY' THEN 1 ELSE 0 END AS ANOMALY
FROM FAA.FAA_COMBINED_OUTLIER
UNION ALL
SELECT 'DURATION' AS VARIABLE,
CASE WHEN DURATION_CHECK = 'MISSING ' THEN 1 ELSE 0 END AS MISSING,
CASE WHEN DURATION_CHECK = 'IN RANGE' THEN 1 ELSE 0 END AS IN_RANGE,
CASE WHEN DURATION_CHECK = 'ANOMALY' THEN 1 ELSE 0 END AS ANOMALY
FROM FAA.FAA_COMBINED_OUTLIER
UNION ALL
SELECT 'HEIGHT' AS VARIABLE,
CASE WHEN HEIGHT_CHECK = 'MISSING ' THEN 1 ELSE 0 END AS MISSING,
CASE WHEN HEIGHT_CHECK = 'IN RANGE' THEN 1 ELSE 0 END AS IN_RANGE,
CASE WHEN HEIGHT_CHECK = 'ANOMALY' THEN 1 ELSE 0 END AS ANOMALY
FROM FAA.FAA_COMBINED_OUTLIER
UNION ALL
SELECT 'NO_PASG' AS VARIABLE,
CASE WHEN NO_PASG_CHECK = 'MISSING ' THEN 1 ELSE 0 END AS MISSING,
CASE WHEN NO_PASG_CHECK = 'IN RANGE' THEN 1 ELSE 0 END AS IN_RANGE,
CASE WHEN NO_PASG_CHECK = 'ANOMALY' THEN 1 ELSE 0 END AS ANOMALY
FROM FAA.FAA_COMBINED_OUTLIER
UNION ALL
SELECT 'PITCH' AS VARIABLE,
CASE WHEN PITCH_CHECK = 'MISSING ' THEN 1 ELSE 0 END AS MISSING,
CASE WHEN PITCH_CHECK = 'IN RANGE' THEN 1 ELSE 0 END AS IN_RANGE,
CASE WHEN PITCH_CHECK = 'ANOMALY' THEN 1 ELSE 0 END AS ANOMALY
FROM FAA.FAA_COMBINED_OUTLIER
UNION ALL
SELECT 'SPEED_GROUND' AS VARIABLE,
CASE WHEN SPEED_GROUND_CHECK = 'MISSING ' THEN 1 ELSE 0 END AS MISSING,
```

```
CASE WHEN SPEED_GROUND_CHECK = 'IN RANGE' THEN 1 ELSE 0 END AS IN_RANGE,
CASE WHEN SPEED_GROUND_CHECK = 'ANOMALY' THEN 1 ELSE 0 END AS ANOMALY
FROM FAA.FAA_COMBINED_OUTLIER
UNION ALL
SELECT 'SPEED_AIR' AS VARIABLE,
CASE WHEN SPEED_AIR_CHECK = 'MISSING ' THEN 1 ELSE 0 END AS MISSING,
CASE WHEN SPEED_AIR_CHECK = 'IN RANGE' THEN 1 ELSE 0 END AS IN_RANGE,
CASE WHEN SPEED_AIR_CHECK = 'ANOMALY' THEN 1 ELSE 0 END AS ANOMALY
FROM FAA.FAA_COMBINED_OUTLIER
)
GROUP BY VARIABLE
;
RUN;

PROC PRINT DATA = FAA.OUTLIER_SUMMARY ;
```

iii) SAS output

Obs	VARIABLE	MISSING	IN_RANGE	ANOMALY
1	AIRCRAFT	0	950	0
2	DISTANCE	0	947	3
3	DURATION	150	795	5
4	HEIGHT	0	938	12
5	NO_PASG	0	950	0
6	PITCH	0	950	0
7	SPEED_AIR	711	239	0
8	SPEED_GROUND	0	950	0

iv) Observations

- Distance, Duration and Height have anomalous values. The occurrences are relatively small
- Speed Air has 75% missing values
- Duration has 16% missing values

v) Conclusion

- Observations with anomalous values for Distance, Duration and Height can be removed from the dataset
- Speed Air has too many missing values and factors to estimate its values are insufficiently available , hence we may drop the column from analysis but let it reside in the dataset
- Values for Duration can be computed with enough accuracy hence it will remain as is for now

e) Cleaning the data based on earlier steps

i) Goal

- Based on Validity check for each variable, Distance, Duration and Height are to be removed from the dataset to remove anomalous data
- We will not drop variable Speed Air just because of its perceived statistical insignificance

ii) SAS code

```
DATA FAA.FAA_COMBINED_CLEANED;
SET FAA.FAA_COMBINED_CLEAN;
IF DURATION > 40 OR MISSING(DURATION)=1;
IF HEIGHT = 6 OR HEIGHT > 6;
IF DISTANCE < 6000;
RUN;
```

```
PROC SQL;
CREATE TABLE FAA.FAA_COMBINED_CLEANED AS (
SELECT     AIRCRAFT, MAX(DURATION) AS DURATION, NO_PASG, SPEED_GROUND, SPEED_AIR,
           HEIGHT, PITCH, DISTANCE, MIN(FEED) AS FEED
FROM  FAA.FAA_COMBINED_CLEANED
GROUP BY   AIRCRAFT, NO_PASG, SPEED_GROUND, SPEED_AIR, HEIGHT, PITCH, DISTANCE );
RUN;


PROC CONTENTS DATA = FAA.FAA_COMBINED_CLEANED; RUN;
```

iii) SAS output

Data Set Name	FAA.FAA_COMBINED_CLEANED
Observations	834
Variables	9
Indexes	0
Observation Length	72
Deleted Observations	0
Compressed	NO
Sorted	NO

iv) Observations

- 20 rows were removed as part of Data cleaning
- Speed Air and Duration are variables which have missing values and they persist
- 96 rows were deleted since they were duplicates

v) Conclusion

- Need to either remove Speed Air or identify technique to calculate its values effectively
- Duration's missing values may be replaced by average values

f) Summarizing the distribution of each

i) Goal

Before we start with the Analytical model, a statistical overview of all of the variables will provide us with insights

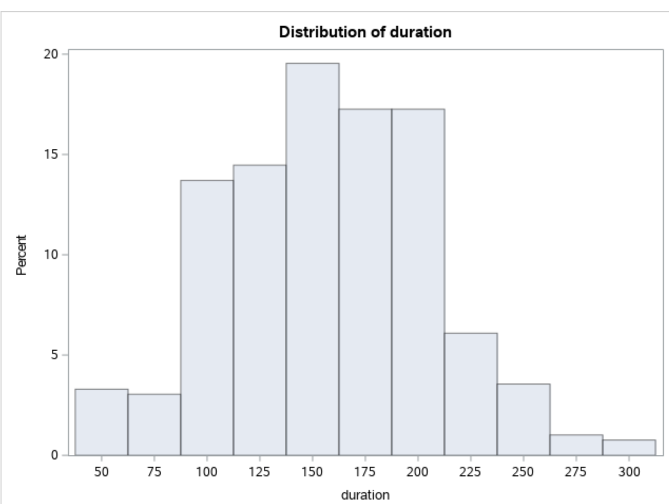
ii) SAS code

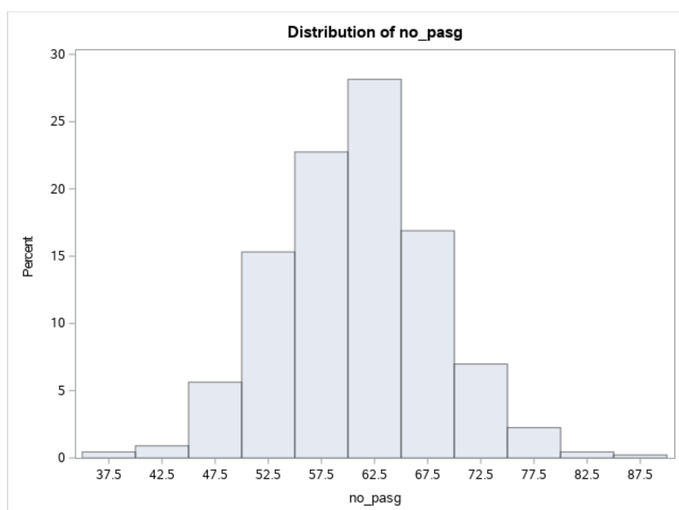
```
PROC MEANS DATA=FAA.FAA_COMBINED_CLEANED CHARTYPE MEAN MEDIAN STD MIN MAX N VARDEF=DF;
	VAR DURATION NO_PASG SPEED_GROUND HEIGHT PITCH DISTANCE;
	CLASS AIRCRAFT;
	TITLE 'STATISTICAL SUMMARY FOR COMBINED DATASET BY AIRCRAFT'; RUN;
PROC SORT DATA=FAA.FAA_COMBINED_CLEANED OUT=WORK.SORTTEMPTABLESORTED; BY AIRCRAFT; RUN;
PROC UNIVARIATE DATA=WORK.SORTTEMPTABLESORTED;
	ODS SELECT HISTOGRAM;
	VAR DURATION NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DISTANCE;
	HISTOGRAM DURATION NO_PASG SPEED_GROUND HEIGHT PITCH DISTANCE;
	BY AIRCRAFT;  TITLE 'FREQUENCY DISTRIBUTION FOR VARIABLES BY AIRCRAFT'; RUN;
PROC DELETE DATA=WORK.SORTTEMPTABLESORTED;
RUN;
TITLE;
```

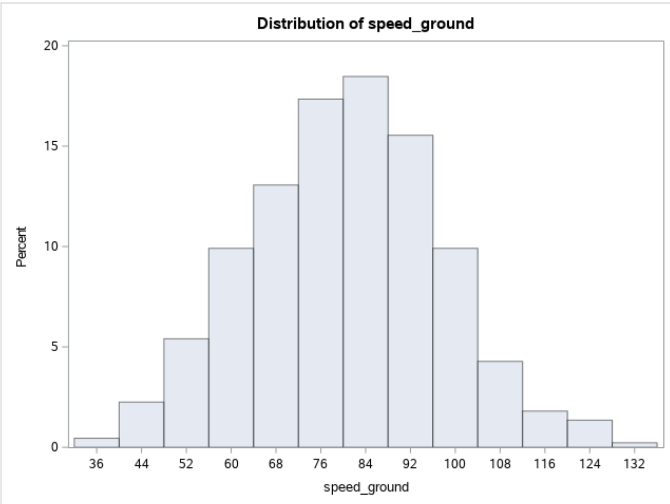
iii) SAS output

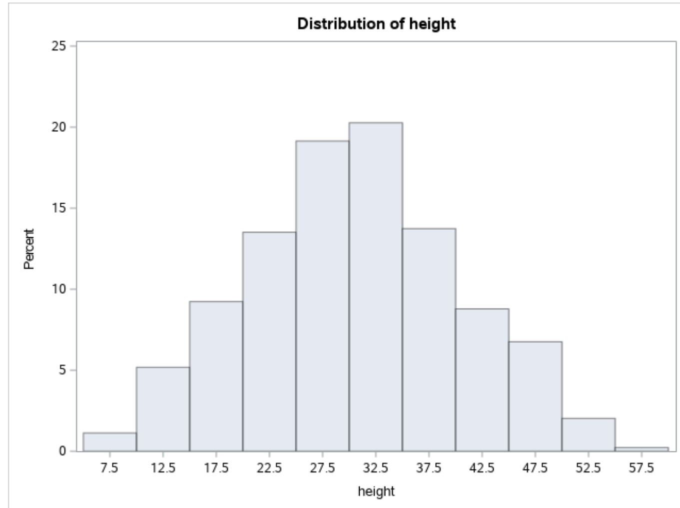
Aircraft	N	Var	Mean	Median	Std Dev	Min	Max	N
Airbus	444	duration	156.9033	156.4468	49.18829	42.14623	305.6217	394
		no_pasg	60.21396	60	7.426491	36	87	444
		speed_ground	80.24988	81.17257	16.95497	33.5741	131.0352	444
		speed_air	104.3098	101.3538	8.089587	95.01136	131.3379	85
		height	30.58922	30.3532	9.854391	6.227518	58.2278	444
		pitch	3.831139	3.825723	0.496079	2.28448	5.526784	444
		distance	1323.32	1126.89	791.9282	41.72231	4896.29	444
Boeing	390	duration	152.7373	152.7312	47.42622	41.94937	298.5223	389
		no_pasg	59.84872	60	7.57585	29	82	390
		speed_ground	78.4348	78.64482	20.8293	27.73572	132.7847	390
		speed_air	102.891	100.8783	10.76242	90.00286	132.9115	118
		height	30.26931	29.62704	9.688849	7.582495	59.94596	390
		pitch	4.205039	4.192321	0.487322	2.993151	5.926784	390
		distance	1746.32	1457.76	951.7122	573.6218	5381.96	390

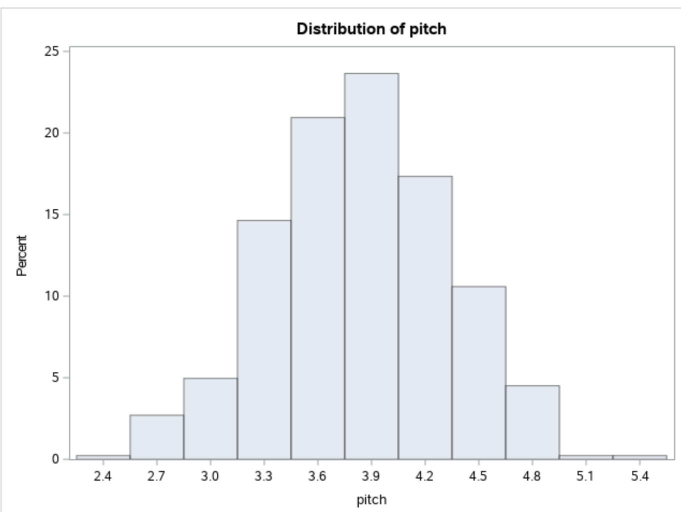
Frequency Distribution for Variables by Aircraft **aircraft=airbus**

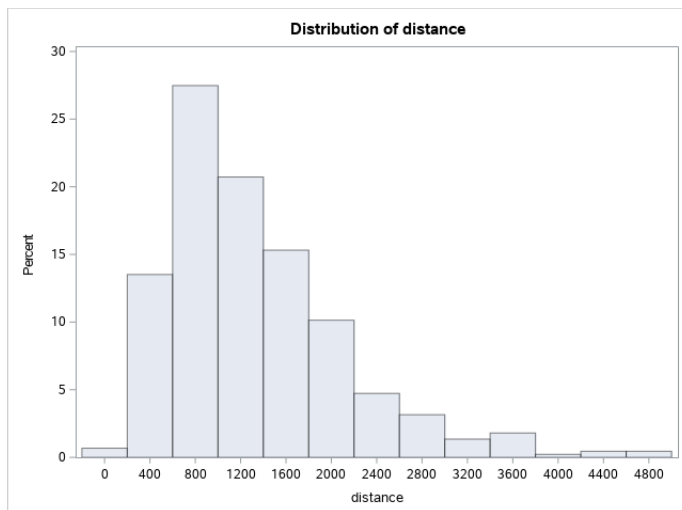


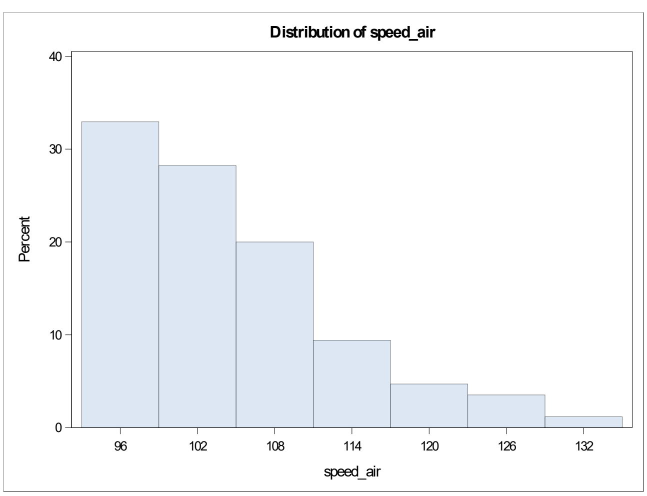



Distribution of speed_ground
Percent
speed_ground
0
5
10
15
20
36
44
52
60
68
76
84
92
100
108
116
124
132

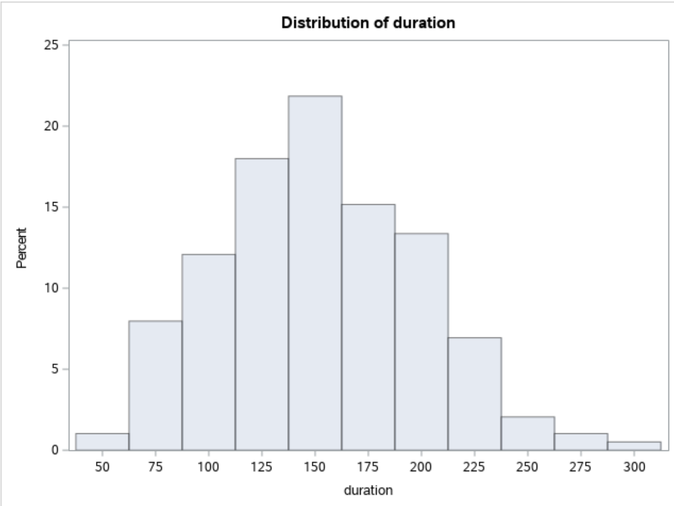

Distribution of height
Percent
height
0
5
10
15
20
25
7.5
12.5
17.5
22.5
27.5
32.5
37.5
42.5
47.5
52.5
57.5

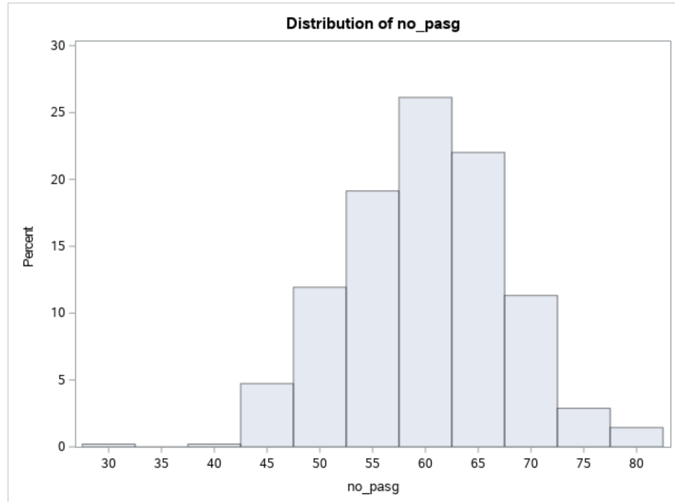

Distribution of pitch
Percent
pitch
0
5
10
15
20
25
2.4
2.7
3.0
3.3
3.6
3.9
4.2
4.5
4.8
5.1
5.4

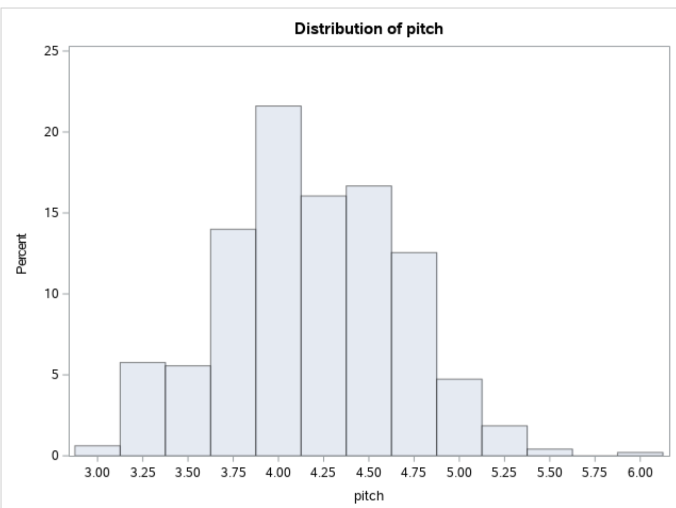

Distribution of distance
Percent
distance
0
5
10
15
20
25
30
0
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800

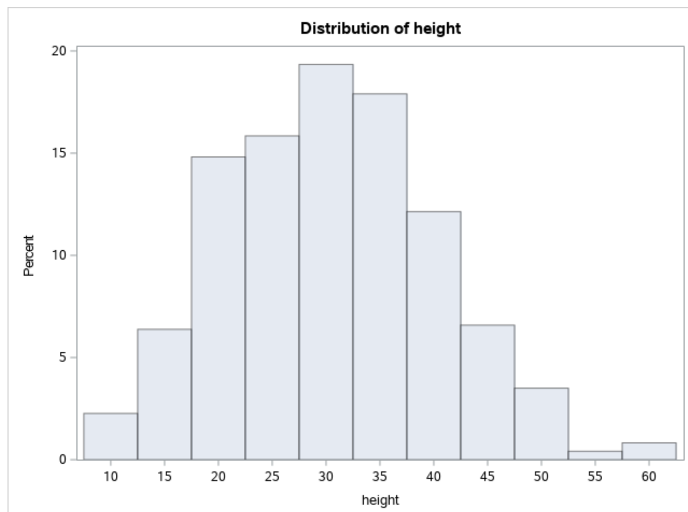

Distribution of speed_air
Percent
speed_air
0
10
20
30
40
96
102
108
114
120
126
132

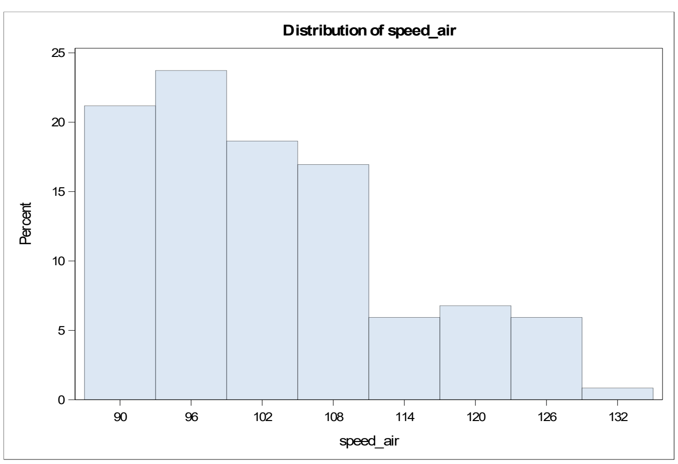
Frequency Distribution for Variables by Aircraft aircraft=boeing

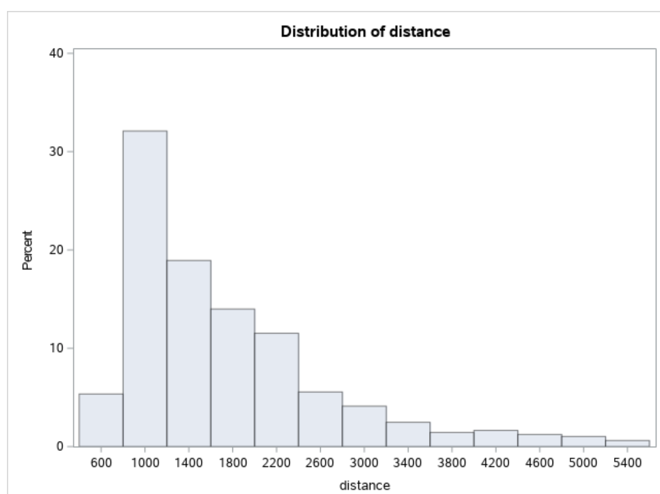


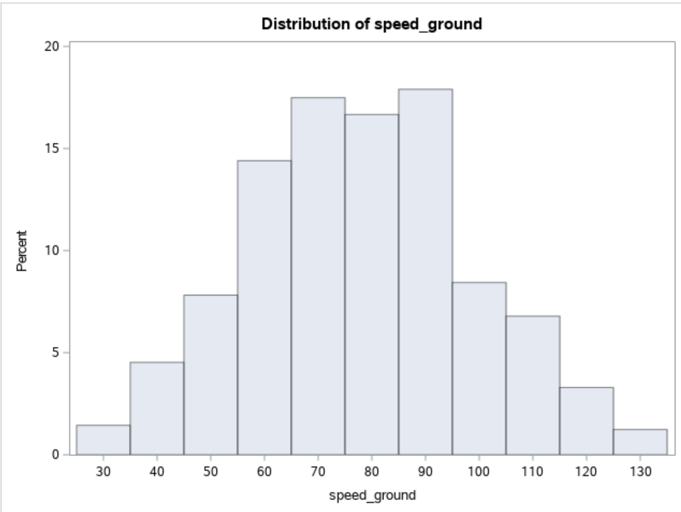












iv) Observations

- The Mean and Median are very close to each other
- This indicates that variables are distributed approximately normally
- Observing Freq distribution graphs, it becomes evident that all variables are arranged in an approximate normal distribution fashion with some extent of symmetry

v) Conclusion

- Data preparation is now complete
- Dataset is ready to be modeled
- 16.6 % rows were lost in the cleaning process
- The final dataset has 834 rows and 9 variables

2) Chapter 2 – Exploratory Data Analysis

a) Identify linear correlation between variables

i) Goal

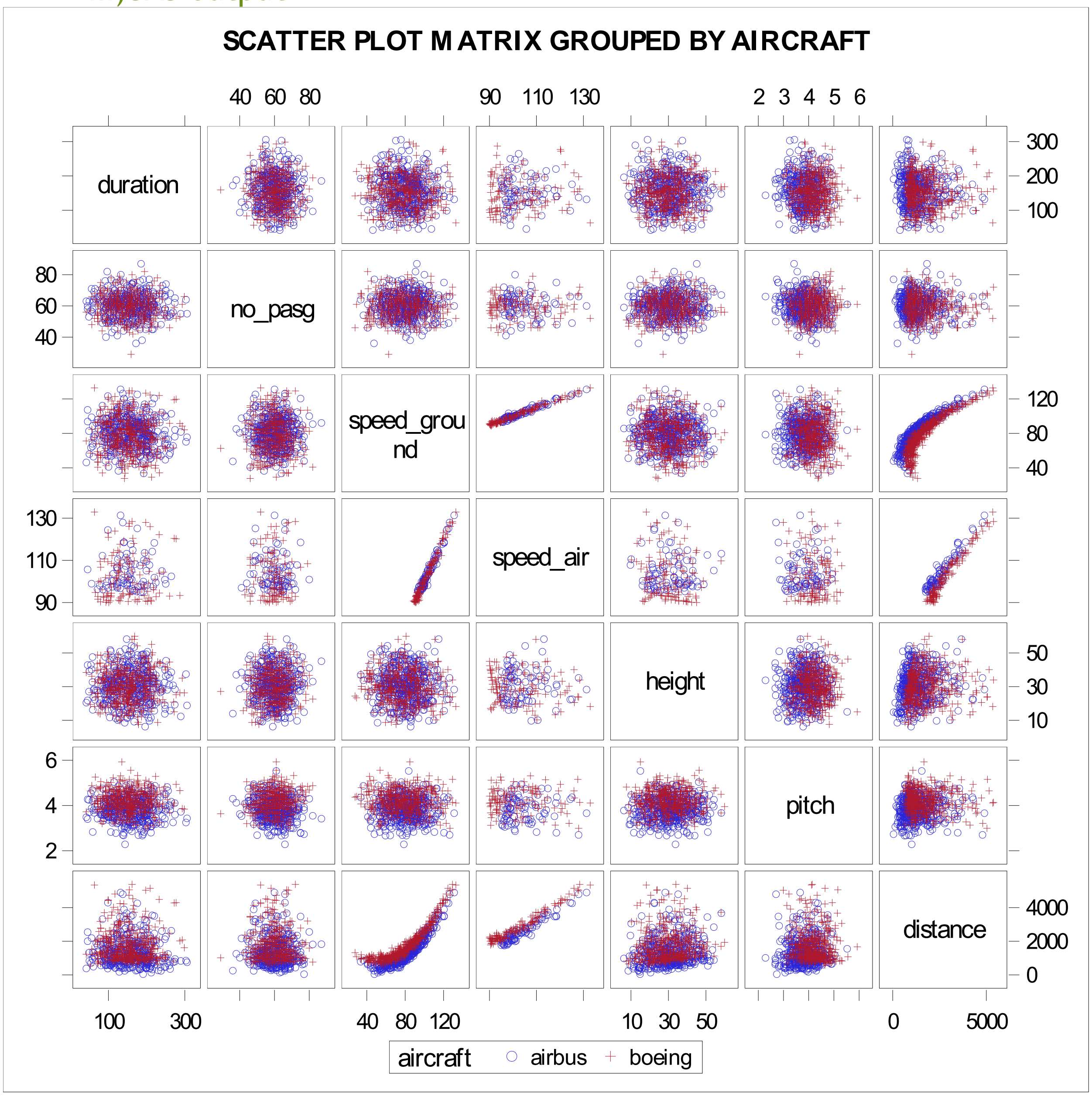
The intent of this chapter is to identify linear correlation between the variables and understanding their impact on the final decision variable that is Distance (landing Distance). We will identify the relationships between variables with the help of a mixed scatter plot. Also, since we have a categorical variable in the mix 'Aircraft', analyses and their interpretation will be grouped by Aircraft type.

ii) SAS code

```
OPTIONS VALIDVARNAME=ANY;
ODS NOPROCTITLE;
ODS GRAPHICS / IMAGEMAP=ON;
```

```
/* SCATTER PLOT MATRIX MACRO */
%MACRO SCATTERPLOTMATRIX(XVARS=, TITLE=, GROUPVAR=);
     PROC SGSCATTER DATA=FAA.FAA;
          MATRIX &XVARS / %IF(&GROUPVAR NE %STR()) %THEN
               %DO; GROUP=&GROUPVAR LEGEND=(SORTORDER=ASCENDING)
%END;      ; TITLE &TITLE; RUN; TITLE;
%MEND SCATTERPLOTMATRIX;
%SCATTERPLOTMATRIX(XVARS=DURATION NO_PASG SPEED_GROUND SPEED_AIR
HEIGHT PITCH DISTANCE,  TITLE="SCATTER PLOT MATRIX GROUPED BY
AIRCRAFT", GROUPVAR=AIRCRAFT);
```

iii) SAS output



Variables	DURATION	NO_PASG	SPEED_GROUND	SPEED_AIR	HEIGHT	PITCH	DISTANCE
DURATION	NA	NO	NO	NO	NO	NO	NO
NO_PASG	NO	NA	NO	NO	NO	NO	NO
SPEED_GROUND	NO	NO	NA	Strong Positive	NO	NO	Strong Positive
SPEED_AIR	NO	NO	Strong Positive	NA	NO	NO	Strong Positive
HEIGHT	NO	NO	NO	NO	NA	NO	NO
PITCH	NO	NO	NO	NO	NO	NA	NO
DISTANCE	NO	NO	Strong Positive	Strong Positive	NO	NO	NA

iv) Observations

As observed from the scatter plot, the only correlations that exist are strong positive correlation and between

- Speed Air and Speed ground
- Speed Air and Distance
- Speed Ground and Distance

v) Conclusion

We need to verify our understanding of these correlations with the help of Pearson Correlation Matrix to quantify the observation. There also exists collinearity amongst two of the predictors in our scope i.e. Speed Air and Speed Ground.

b) Quantify linear correlation between Variables

i) Goal

Verify if the correlations identified from Scatter plot are of any statistical relevance and quantify the outcome. This analysis is grouped by Aircraft type to discover differences if any

ii) SAS code

```
ODS NOPROCTITLE;
ODS GRAPHICS / IMAGEMAP=ON;

PROC SORT DATA=FAA.FAA OUT=WORK.SORTTEMPTABLESORTED;
      BY AIRCRAFT;
RUN;
```

```
PROC CORR DATA=WORK.SORTTEMPTABLESORTED PEARSON NOSIMPLE;
      VAR DURATION NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH;
      WITH DISTANCE;
      BY AIRCRAFT;
RUN;

PROC DELETE DATA=WORK.SORTTEMPTABLESORTED;
RUN;


ODS NOPROCTITLE;
ODS GRAPHICS / IMAGEMAP=ON;

PROC SORT DATA=FAA.FAA OUT=WORK.SORTTEMPTABLESORTED;
      BY AIRCRAFT;
RUN;

PROC CORR DATA=WORK.SORTTEMPTABLESORTED PEARSON NOSIMPLE  PLOTS=NONE;
      VAR SPEED_GROUND;
      WITH SPEED_AIR;
      BY AIRCRAFT;
RUN;

PROC DELETE DATA=WORK.SORTTEMPTABLESORTED;
RUN;
```

iii) SAS output

aircraft=airbus

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

aircraft=boeing

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

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations						
	duration	**no_pasg**	**speed_ground**	**speed_air**	**height**	**pitch**
distance distance	-0.01266 0.8034 389	-0.01672 0.7421 390	0.89345 <.0001 390	0.97760 <.0001 118	0.07138 0.1595 390	-0.06493 0.2007 390

aircraft=airbus

1 With Variables:	speed_air
1 Variables:	speed_ground

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	speed_ground
speed_air speed_air	0.98169 <.0001 85

aircraft=boeing

1 With Variables:	speed_air
1 Variables:	speed_ground

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	speed_ground
speed_air speed_air	0.99048 <.0001 118

iv) Observations

Pearson Correlation Coefficients							
Prob > |r| under H0: Rho=0							
Aircraft	**Value**	**duration**	**no_pasg**	**speed_ground**	**speed_air**	**height**	**pitch**
airbus	coefficient	-0.07851	-0.00732	0.9052	0.96411	0.14494	0.0733
	p-value	0.1198	0.8777	<.0001	<.0001	0.0022	0.123
	N	394	444	444	85	444	444
boeing	coefficient	-0.01266	-0.01672	0.89345	0.9776	0.07138	-0.06493
	p-value	0.8034	0.7421	<.0001	<.0001	0.1595	0.2007
	N	389	390	390	118	390	390

The p-value indicates the probability of null hypothesis being true. The null hypothesis in Pearson Correlation test is that there exists no linear correlation within the variables. If we set the confidence interval for the hypotheses test as 99%, we can see that when aircraft is an Airbus, Speed ground, Speed air and Height impact the (landing) distance. Also, when aircraft is an airbus, only Speed ground and Speed air impact the (landing) distance.

Another point of observation of this result is that the number of values used for each test represented by N. As we can see, the values used for this test are very less in case of Speed air as compared to other variables

As also observed, there is very high correlation amongst the predictors Speed Ground and Speed Air. Since there are many missing values for Speed Air, we will leave it out of the model.

v) Conclusion

With a 99% confidence interval we can state that the following correlations exist with the variable Distance -

- For Aircraft Airbus
  - Speed Ground – Strong Positive
  - Height – Weak Positive
- For Aircraft Boeing
  - Speed Ground – Strong Positive

The above-mentioned correlations can now be used to create a Linear Regression Model which can help us study (landing) Distance.

3) Chapter 3 – Linear Regression Model

a) Linear Model for Aircraft Boeing

i) Goal

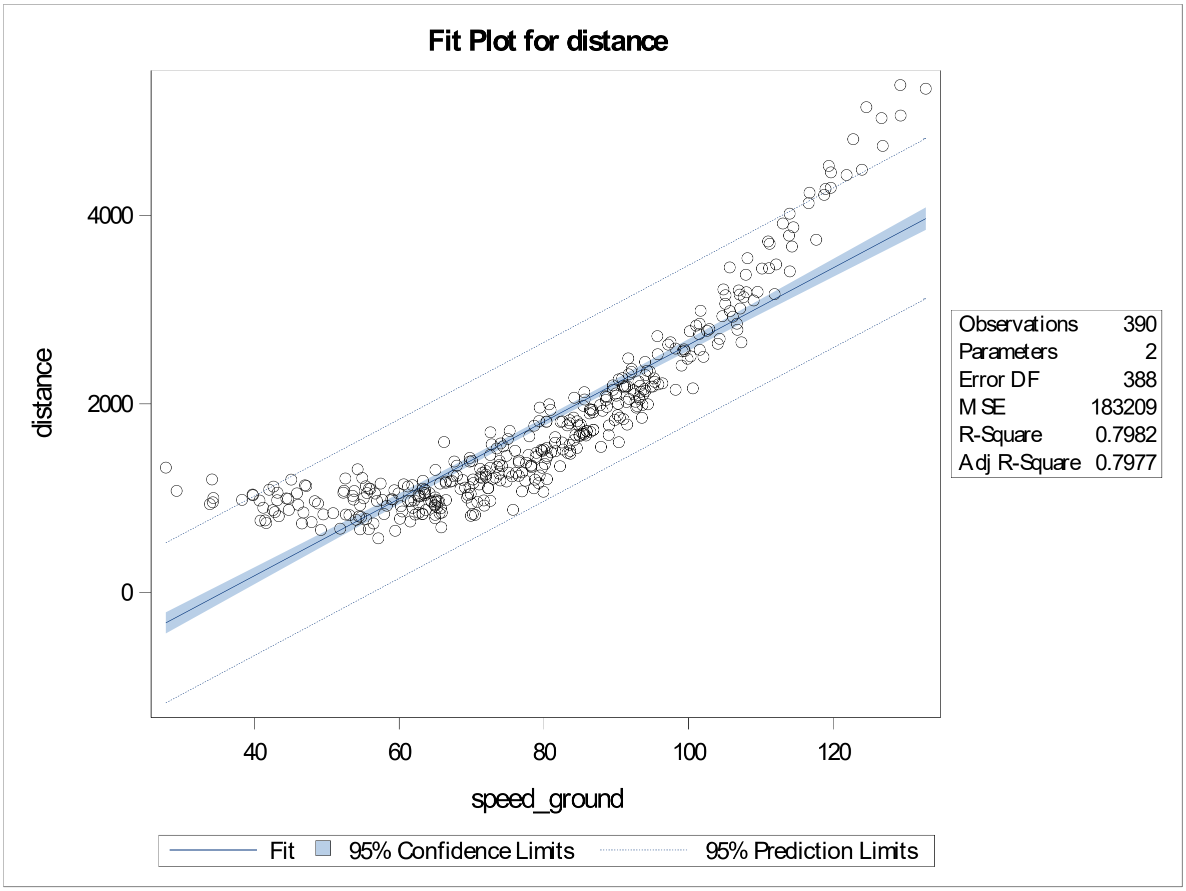
Create a linear model of the form $Y = \beta_0 + \beta_1 X_1$ to estimate the value of Distance (Y) when Speed Ground (X_1) is known where β_0 and β_1 is the coefficient for Speed Air.

ii) SAS code

```
PROC REG DATA=FAA.FAA ALPHA=0.05 ;
      WHERE AIRCRAFT="boeing";
      MODEL DISTANCE=SPEED_GROUND /;
ODS SELECT PARAMETERESTIMATES FITPLOT   ;
RUN;
```

iii) SAS output

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1455.59057	84.54609	-17.22	<.0001
speed_ground	speed_ground	1	40.82254	1.04189	39.18	<.0001



iv) Observations

For the values of the coefficients listed in the output, we have a p-value very small, meaning that the values have statistical significance. This model predicts the value of Landing Distance solely based on Speed Ground.

v) Conclusion

We can state with 95% confidence interval that for Aircraft=Boeing

Distance = -1455.6 + 40.8*Speed_Ground, where

β_0 = Intercept = -1455.6 +- 84.6

β_1 = Coefficient of Speed Ground = 40.8 +- 1.04

b) Linear Model for Aircraft Airbus

i) Goal

Create a linear model of the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ to estimate the value of Distance (Y) when Speed Ground (X_1) is known where β_0 is the intercept, β_1 is the coefficient for Speed Ground and β_2 is the coefficient for Height.

ii) SAS code

```
PROC REG DATA=FAA.FAA ALPHA=0.05;
      WHERE AIRCRAFT="airbus";
      MODEL DISTANCE=SPEED_GROUND HEIGHT/;

ODS SELECT PARAMETERESTIMATES  ;
RUN;
```

iii) SAS output

Parameter Estimates						
Variable	Label	DF	Parameter	Standard	t Value	Pr > |t|
	Intercept	1	-2522.89061	85.19508	-29.61	<.0001
Intercept	speed_ground	1	42.5542	0.86152	49.39	<.0001
height	height	1	14.09773	1.48228	9.51	<.0001

iv) Observations

For the values of the coefficients listed in the output, we have a p-value very small, meaning that the values have statistical significance. This model predicts the value of Landing Distance based on Speed Ground and Height.

v) Conclusion

We can state with 95% confidence interval that for Aircraft=Boeing

Distance = -2522.9 + 42.6*Speed_Ground + 14.1*Height, where

β_0 = Intercept = -2522.9 +- 85.2

β_1 = Coefficient of Speed Ground = 42.6 +- 0.9

β_2 = Coefficient of Height = 14.1 +- 1.5

4) Chapter 4 – Summary Questions and Limitation

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

Both data sets together contained 1000 rows of data. However, only 834 rows were used to fit the model. Following is a brief account of the removed rows:

- 50 blank rows
- 24 anomalous rows
- 92 duplicate rows

b) What factors and how they impact the landing distance of a flight?

Different factors impact the landing distance differently based on type of aircraft:

- Airbus
  - Speed Air (miles per hour) – Unit increases in Speed air increases the Landing distance(feet) by 41 units
- Boeing
  - Speed Air (miles per hour) - Unit increase in Speed air increases the Landing distance(feet) by 43 units
  - Height (meters) - Unit increase in Speed air increases the Landing distance(feet) by 15 units

c) Is there any difference between the two makes, Boeing and Airbus?

Technically, it is factual that there exists a difference between the makes of aircraft because they are designed differently and have different parameters. These differences were also illustrated during the data analyses phase where similar parameters had statistically significant differences in terms of averages and deviations based on the make. Furthermore, it was observed that some factors did impact one type of make alone. For example, while regressing Distance for Airbus, it was observed that Height did play a role whereas so was not the case with Boeing.

d) Limitations of data analyses

- We have analyzed the data set with the intent of drawing only linear correlations
- Boeing and Airbus are two different classes of airplane and hence they shouldn't be grouped together for analysis. Hence, a singular criterion being applied for anomaly detection to both groups of data may not yield pragmatic results
- Since Air Speed = Ground Speed – Wind Speed, we can calculate Speed Air, a variable is having 76% missing values. However, Airport conditions or locations are unknown and so is any approximation criterion for Wind Speed. Is it possible to obtain Wind Speed from client or any other relevant measure to determine Speed Air? Doing this can also help us know if Speed Air is really a factor impacting Landing distance
- While Data preparation step, I considered all the observations since the dataset was very small compared to real-life datasets. When dealing with real life situations, I think it would make more sense to operate on a sample of raw data to determine anomalies. However, since we are sampling, we may not be able to determine the best way of anomaly detection. So, I was wondering if Anomaly detection criteria is only the result of our understanding of the values of a variable can take based on the data dictionary?
- I think that this step should involve a lot of feedback and rework steps. Based on our study, we can reach back to the client to suggest changes and obtain other variables if possible. Also, the collective summary of anomaly detection and removal must be sent to the client for his approval before we move towards modeling
- What is the most comprehensive way to code anomaly identification in SAS? I used IF statements in the DATA step and I think it is a time-consuming process
- Also, in the process, there are multiple questions related to the data itself:
    - Where are airports located?
    - Where will the outcome of this project be applied?
    - Elaboration of parameters like time, day, date, weather, etc. during data collection?
    - More diversification of the variable Aircraft?