/\*1-Importing Input files FAA1 and FAA2\*/

FILENAME REFFILE '/home/satyasmc0/Stat computing class/FAA1.xls';

PROC IMPORT DATAFILE=REFFILE

DBMS=XLS

OUT=WORK.faa1;

GETNAMES=YES;

sheet=FAA1;

run;

FILENAME REFFILE '/home/satyasmc0/Stat computing class/FAA2.xls';

PROC IMPORT DATAFILE=REFFILE

DBMS=XLS

OUT=WORK.faa2;

GETNAMES=YES;

sheet=faa2;

proc means data=faa1 n mean std range min max nmiss;

proc means data=faa2 n mean std range min max nmiss;

run;

/\*2-from the means procedure, it is evident that the faa1 and faa2 belong to the same population

thus data can be combined\*/

data faa3;

set faa1 faa2;

proc means data=faa3 n mean std range min max nmiss;

run;

/\*3- Checking for duplicates\*/

proc sort data= faa3 nodupkey

out = faa4;

by pitch;

proc print data=faa4;

run;

/\*4- Removing missing values\*/

data faa4;

set faa4;

if missing(aircraft) then delete;

run;

proc sort data=faa4;

by aircraft;

proc means data=faa4 n mean std range min max nmiss;

title 'All data summary- unique';

run;

/\*No missing values found in summary- duration is missing for 50 values\*/

/\*5- Checking and removing abnormal values\*/

data faa\_normal;

set faa4;

if duration=. then miss='yes';

if speed\_ground=. then miss='yes';

if speed\_air=. then miss='yes';

if height=. then miss='yes';

if distance=. then miss='yes';

if pitch=. then miss='yes';

if duration<=40 and duration <> . then abnormal='yes';

if speed\_ground<30 or speed\_ground>140 and speed\_ground <> . then abnormal='yes';

if (speed\_air<30 or speed\_air>140) and speed\_air <> . then abnormal='yes';

if height<6 and height <> . then abnormal='yes';

if distance>6000 and distance <> . then abnormal='yes';

run;

proc sort data=faa\_normal;

by abnormal miss;

proc print data=faa\_normal;

proc means data = faa\_normal n nmiss min max ;

run;

/\*Since abnormal values are a very small percentage of the entire data, deleting them\*/

data faa\_normal;

set faa\_normal;

if abnormal='yes' then delete;

drop abnormal;

drop miss;

proc means data = faa\_normal n nmiss min max ;

run;

/\*We end up with 831 values with their summary\*/

/\* Comparing distributions indicates the speed\_air is a truncated dist so better to seggregate\*/

proc chart data= faa\_normal;

vbar speed\_air;

run;

proc chart data= faa\_normal;

vbar speed\_ground;

run;

data faa\_normal;

set faa\_normal;

if speed\_air= . then Group = 0; else Group = 1;

proc print data=faa\_normal;

proc means data = faa\_normal n nmiss min max;

run;

/\*Exploring Variable distributions\*/

proc univariate data=faa\_normal;

class group;

var speed\_ground;

histogram speed\_ground;

proc univariate data = faa\_normal;

class group;

var height;

histogram height;

proc univariate data = faa\_normal;

class group;

var pitch;

histogram pitch;

proc univariate data = faa\_normal;

class group;

var no\_pasg;

histogram no\_pasg;

/\*Height, pitch and passenger count variables are nearly normal\*/

/\*Exploring Variable Correlations\*/

proc corr data=faa\_normal;

var duration speed\_air speed\_ground no\_pasg pitch height distance;

run;

/\*The correlation matrix shows us that there is no impact of passenger count and duration on distance

Also as expected speed air and speed ground are heavily correlated\*/

/\*We can plot to see if there is any non-linear correlation\*/

proc gplot ; plot distance\*height;

proc gplot ; plot distance\*pitch;

proc gplot ; plot distance\*no\_pasg;

/\*Drop duration and passenger count\*/

data faa\_trim;

set faa\_normal;

drop duration no\_pasg;

run;

proc means data=faa\_trim n nmiss min max;

run;

/\*Impact of aircraft class\*/

proc ttest data=faa\_trim;

class aircraft;

var distance;

title 'Mean distance across Airbus and Boeing';

run;

/\*pvalue<alpha so refer Satterthwaite section implying unequal variances\*/

/\*p value of ttest implies that the mean equality can be rejected

Created a dummy variable that can be used in regression\*/

data faa\_final;

set faa\_trim;

if aircraft= 'boeing' then planetype = 0; else planetype = 1;

proc means data=faa\_final;

run;

/\*Regressing the landing distance against variables

- building 2 models in order to not delete out the air speed\*/

proc reg data = faa\_final;

model distance = planetype speed\_air height;

Title 'Regression when air speed is available';

proc reg data = faa\_final;

model distance = planetype speed\_ground height;

title 'Regression when air speed is unavailable;