Lab 2 Report:

Name: Anh Nguyen UCID: nguye2aq

Here are what I've done, my code and rationale:

1. Combining data sets from different sources;

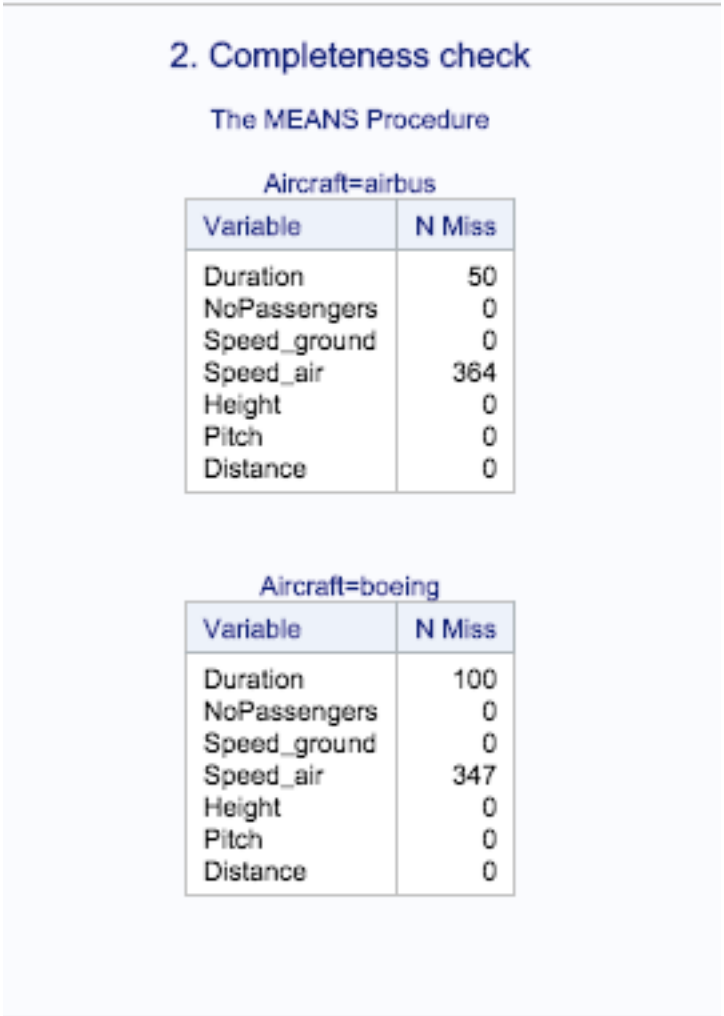
```
*1 Combining data sets from different sources;
data FAA1;
Filename FAA1xls '/home/u42905097/Class 2/FAA1.txt';
infile FAA1xls delimiter='09'x;
input Aircraft $ Duration NoPassengers Speed_ground Speed_air Height Pitch Distance;
run;

data FAA2;
Filename FAA2xls '/home/u42905097/Class 2/FAA2.txt';
infile FAA2xls delimiter='09'x;
input Aircraft $ NoPassengers Speed_ground Speed_air Height Pitch Distance;
run;

proc sort data=faa1;
by Aircraft;
run;

proc sort data = faa2;
by Aircraft;
run;

data FAACombined;
set FAA1 FAA2;
by Aircraft;
run;

title '1. Combination';
proc print data=FAACombined;
run;
```

* I couldn't figure out how to infile a .xls file (there were something wrong with the way strings were input), so I tried to do it using a .txt file. I found and replaced all blank cells with '.'. Then used the delimiter '09'x for tabs, since the text file automatically turned cells into tabs.
* Since the FAA2 doesn't have the Duration variable, to avoid losing variables and observations, I decided that interweaving concatenation is the best method to combine 2 files.

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

```
*2. Performing the completeness check of each variable – examine if missing values are present;
title '2. Completeness check';
proc means data = FAACombined nmiss;
by Aircraft;
var duration NoPassengers Speed_ground Speed_air Height Pitch Distance;
run;
```

There are a lot of missing value presented, in Duration and Speed_air.



2. Completeness check

The MEANS Procedure

Aircraft=airbus

Variable	N Miss
Duration	50
NoPassengers	0
Speed_ground	0
Speed_air	364
Height	0
Pitch	0
Distance	0

Aircraft=boeing

Variable	N Miss
Duration	100
NoPassengers	0
Speed_ground	0
Speed_air	347
Height	0
Pitch	0
Distance	0

There are in total 150 missing Duration datapoints and 711 missing Speed_air datapoints.

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

```
*3. Performing the validity check of each variable – examine if abnormal values are present;
title '3. Validity check';
data FAAValidity;
set faacombined;
if Duration>40 or Duration=. then ValidDuration = 1;
else ValidDuration = 0;
if Speed_ground <30 or Speed_ground>140 then ValidSpeed_ground = 0;
else ValidSpeed_ground = 1;
if (Speed_air >=30 and Speed_air<=140) or Speed_air=. then ValidSpeed_air = 1;
*if Speed_air=. then ValidSpeed_air = 1;
else ValidSpeed_air = 0;
if Height>6 then ValidHeight = 1;
else ValidHeight = 0;
if Distance<6000 then ValidDistance = 1;
else ValidDistance = 0;
run;

proc print data=FAAValidity;
```

```
run;
```

I encode the abnormal datapoint as 0 and normal datapoint as 1, creating a new variable for each validity check.

4. Cleaning the data based on the results of Steps 2 and 3;

```
*4. Cleaning the data based on the results of Steps 2 and 3;
title '4. Data Cleaning';
data FAACleaned;
set FAAValidity;
if ValidDuration=1 and ValidSpeed_ground=1 and ValidSpeed_air=1 and ValidHeight=1
and ValidDistance = 1;
drop ValidDuration ValidSpeed_ground ValidSpeed_air ValidHeight ValidDistance;
run;

proc print data = FAACleaned;
run;
```

I deleted all the data where the Validity index is 0. As you can see in the difference of observations, there were 23 abnormal observations and were deleted.

5. Summarizing the distribution of each variable (what tables and figures will you present?);

```
 *5. Summarizing the distribution of each variable (what tables and figures will you present?);
title '5. Distributions of variable';
proc means data = faacleaned n mean median nmiss std range q1 q3;
by Aircraft;
Var Duration NoPassengers Speed_ground Speed_air Height Pitch Distance;
run;

proc univariate data = faacleaned plot;
by Aircraft;
var Duration Height Distance NoPassengers;
run;
```

I utilized the Proc Means and Proc Univariate to summarize the distribution of each variable.

I specified the numbers of observations, mean, median, missing obs, range, lower and upper quantiles of each variable. (Sorted by aircraft)

I displayed the Distribution and Probability plot, Count distribution Scatter diagram and trendline for each variable. (Sorted by aircraft)

I also displayed box plots by groups by variables.

In reality, which figures and graphics I choose to present will depend on the purpose of the analysis.

6. **List all the questions you have in the course of data preparation (for your own reference)**

I have a few questions:

- Should we eliminate the observations with missing data? Without data, the missing data could have been abnormal, and the reason could be mis-measurement, therefore the data points could skewed our analysis. However, we can't dismiss an entire source of data because of a missing variable (FAA2 doesn't have duration of flight)

- The height requirements are dubbed "non-typical", not "abnormal". Should we drop these datapoints during cleaning all the same?