```
/* Create a library and read excel file and save it to a SAS dataset*/
Libname Predict "/home/u61478728/BAN130/Final Project";

PROC IMPORT datafile="/home/u61478728/BAN130/Final Project/FlightDelays.csv"
        DBMS=CSV out=Predict.flight replace;
    guessingrows=max;
RUN;

/*1. Perform the necessary “Handling Missing Data” operations to the missing values.
--Checking for missing values in Numeric variables using PROC MEANS*/
data Predict.Flight_Clean;
    set Predict.Flight(rename=('Flight Status'n=FLIGHT_STATUS Weather=WEATHER));

    if missing(FL_DATE) then
        delete;
run;

proc means data=predict.Flight_clean nmiss;
run;

title 'Final dataset after handling the missing data cleaning';

proc print data=predict.Flight_clean (obs=5) noobs;
run;

/*2 Create a new SAS dataset “FlightDelays” containing only one Origin plus a new
variable called DelayedFlight with values of 1 for delayed flight and 0 for none.*/
data Predict.FlightDelays;
    set Predict.Flight_clean;

    if ORIGIN='DCA' and Flight_Status='ontime' then
        DelayedFlight=0;
    else if ORIGIN='DCA' and Flight_Status='delayed' then
        DelayedFlight=1;
    else
        delete;
run;

title 'Top 5 observation of the new "FlightDelays" dataset';

proc print data=predict.flightdelays(obs=5);
run;

/*3. Generate a table for the average delay per day for each airport and plot the vertical
bar chart for the 7 days. */
proc format;
    value WEEKDAYS 1='Monday' 2='Tuesday' 3='Wednesday' 4='Thursday' 5='Friday'
        6='Saturday' 7='Sunday';
run;

data predict.FormattedData;
    set predict.Flight_clean;
    format DAY_WEEK WEEKDAYS.
       UPDATED_CRS_DEP_TIME time5.
       UPDATED_DEP_TIME time5.;
    UPDATED_CRS_DEP_TIME=input(put(CRS_DEP_TIME, z4.), hhmmss.);
    UPDATED_DEP_TIME=input(put(DEP_TIME, z4.), hhmmss.);
    DELAY_IN_MINS=intck('minutes', UPDATED_CRS_DEP_TIME, UPDATED_DEP_TIME);
    drop CRS_DEP_TIME DEP_TIME;

    If DELAY_IN_MINS lt -19 then
        delete;
run;

/* Procedure for table for JFK destination*/
proc sql;
    create table predict.DailyAverageDelay_JFK as select * from
        predict.FormattedData where DEST='JFK';
quit;

title 'TABLE for The Avg Delay Per Day for Destination Airport JFK';

proc print data=predict.DailyAverageDelay_JFK (Obs=5) noobs;
run;

title 'Bar Graph The Avg Delay Per Day for Destination Airport JFK';

proc sgplot data=predict.DailyAverageDelay_JFK;
```

```
    vbar DAY_WEEK / response=DELAY_IN_MINS stat=mean;
run;

/* Procedure for table for EWR destination*/
proc sql;
    create table predict.DailyAverageDelay_EWR as select * from
        predict.FormattedData where DEST='EWR';
quit;

title 'TABLE for The Avg Delay Per Day for Destination Airport EWR';

proc print data=predict.DailyAverageDelay_EWR (Obs=5) noobs;
run;

title 'Bar Graph The Avg Delay Per Day for Destination Airport EWR';

proc sgplot data=predict.DailyAverageDelay_EWR;
    vbar DAY_WEEK / response=DELAY_IN_MINS stat=mean;
run;

/*Procedure for table for LGA destination*/
proc sql;
    create table predict.DailyAverageDelay_LGA as select * from
        predict.FormattedData where DEST='LGA';
quit;

title 'Bar Graph The Avg Delay Per Day for Destination Airport LGA';

proc print data=predict.DailyAverageDelay_LGA (Obs=5) noobs;
run;

title 'Bar Graph The Avg Delay Per Day for Destination Airport LGA';

proc sgplot data=predict.DailyAverageDelay_LGA;
    vbar DAY_WEEK / response=DELAY_IN_MINS stat=mean;
run;

/*Procedure for table for BWI Origin*/
proc sql;
    create table predict.DailyAverageDelay_BWI as select * from
        predict.FormattedData where ORIGIN='BWI';
quit;

title 'TABLE for The Avg Delay Per Day for Origin Airport BWI';

proc print data=predict.DailyAverageDelay_BWI (Obs=5) noobs;
run;

title 'Bar Graph The Avg Delay Per Day for Origin Airport BWI';

proc sgplot data=predict.DailyAverageDelay_BWI;
    vbar DAY_WEEK / response=DELAY_IN_MINS stat=mean;
run;

/*Procedure for table for IAD Origin*/
proc sql;
    create table predict.DailyAverageDelay_IAD as select * from
        predict.FormattedData where ORIGIN='IAD';
quit;

title 'TABLE for The Avg Delay Per Day for Origin Airport IAD';

proc print data=predict.DailyAverageDelay_IAD (Obs=5) noobs;
run;

title 'Bar Graph The Avg Delay Per Day for Origin Airport IAD';

proc sgplot data=predict.DailyAverageDelay_IAD;
    vbar DAY_WEEK / response=DELAY_IN_MINS stat=mean;
run;

/*Procedure for table for DCA Origin*/
proc sql;
    create table Predict.DailyAverageDelay_DCA as select * from
        Predict.FormattedData where ORIGIN='DCA';
quit;

title 'TABLE for The Avg Delay Per Day for Origin Airport DCA';
```

```
proc print data=Predict.DailyAverageDelay_DCA (Obs=5) noobs;
run;

title 'Bar Graph The Avg Delay Per Day for Origin Airport DCA';

proc sgplot data=Predict.DailyAverageDelay_DCA;
    vbar DAY_WEEK / response=DELAY_IN_MINS stat=mean;

    /*4. Produce a report showing the mean number of flights per day for each Carrier. Give
    a sample of a scatter plot for one of the Carrier. */
proc sql;
    create table predict.Carrier_AvgFlight as select CARRIER, DAY_WEEK,
        count(FL_NUM) as COUNT_FL from predict.FormattedData group by CARRIER,
        DAY_WEEK order by DAY_WEEK;
quit;

title 'Printing the first five observation of table Carrier_AvgFlight';

proc print data=predict.Carrier_AvgFlight (obs=5) noobs;
run;

title 'Average Flights daily by Each Carrier';

proc report data=predict.Carrier_AvgFlight;
    column CARRIER DAY_WEEK COUNT_FL;
    define CARRIER/ group;
    define DAY_WEEK/ ORDER=INTERNAL group;
    define COUNT_FL/ analysis MEAN;
run;

title 'SCATTERPLOT FOR Flights Per day for Carrier USAirways';

proc sgplot data=predict.Carrier_AvgFlight (where=(CARRIER='US'));
    scatter x=DAY_WEEK y=COUNT_FL/group=CARRIER;
    xaxis grid;
    yaxis grid;
run;

/*6. Plot a histogram for each of the quantitative variables. Based on the histograms and
summary statistics, answer the following question: Which variables have the largest
variabilities? */
title "Histogram for Quantitative Variable DAY_OF_MONTH";

proc sgplot data=predict.FormattedData;
    histogram DAY_OF_MONTH;
    density DAY_OF_MONTH;
run;

proc univariate data=predict.FormattedData normal plot;
    var DAY_OF_MONTH;
run;

title "Histogram for Quantitative Variable DAY_WEEK";

proc sgplot data=predict.FormattedData;
    histogram DAY_WEEK;
    density DAY_WEEK;
run;

proc univariate data=predict.FormattedData normal plot;
    var DAY_WEEK;
run;

title "Histogram for Quantitative Variable DISTANCE";

proc sgplot data=predict.FormattedData;
    histogram DISTANCE;
    density DISTANCE;
run;

proc univariate data=predict.FormattedData normal plot;
    var DISTANCE;
run;

title "Histogram for Quantitative Variable WEATHER";

proc sgplot data=predict.FormattedData;
    histogram WEATHER;
    density WEATHER;
```

```
run;

proc univariate data=predict.FormattedData normal plot;
    var WEATHER;
run;

title "Histogram for Quantitative Variable FL_NUM";

proc sgplot data=predict.FormattedData;
    histogram FL_NUM;
    density FL_NUM;
run;

proc univariate data=predict.FormattedData normal plot;
    var FL_NUM;
run;

title "Histogram for Quantitative Variable UPDATED_CRS_DEP_TIME";

proc sgplot data=predict.FormattedData;
    histogram UPDATED_CRS_DEP_TIME;
    density UPDATED_CRS_DEP_TIME;
run;

proc univariate data=predict.FormattedData normal plot;
    var UPDATED_CRS_DEP_TIME;
run;

PROC CONTENTS data=predict.FormattedData;
    title "Histogram for Quantitative Variable UPDATED_DEP_TIME";

proc sgplot data=predict.FormattedData;
    histogram UPDATED_DEP_TIME;
    density UPDATED_DEP_TIME;
run;

proc univariate data=predict.FormattedData normal plot;
    var UPDATED_DEP_TIME;
run;

/*6 Provide data summarization using four different Pivot tables to highlight different
facts about the dataset*/
proc freq data=predict.FormattedData;
    tables ORIGIN*DEST;
run;

proc freq data=predict.FormattedData;
    tables FLIGHT_STATUS*DAY_WEEK;
run;

proc freq data=predict.FormattedData;
    tables CARRIER*DAY_WEEK;
run;

proc freq data=predict.FormattedData;
    tables CARRIER*FLIGHT_STATUS;
run;

/*8 Data Reduction: Reduce the number of variables (columns) using the necessary
operation (e.g., domain knowledge). Store the result of this step in a new file
“FlightDelaysTrainingData.csv” */
data predict.FlightDelays_Reduced;
    set predict.FormattedData;
    drop TAIL_NUM Weather FL_NUM FL_DATE DAY_OF_MONTH DISTANCE;
run;

title "Printing the first 20 values FlightDelays_Reduced Dataset";

proc print data=predict.FlightDelays_Reduced (obs=20);
run;

proc export data=predict.FlightDelays_Reduced
        outfile="/home/u61478728/BAN130/Final Project/FlightDelaysTrainingData.csv"
        dbms=csv;
run;

/*9 Data Conversion: Some of the algorithms don’t comply with numerical data. The nonnumerical data in the
dataset is required to be converted. You need to provide a reference table for the transformed data. */
data predict.Flight_converted;
```

```
    set predict.FlightDelays(keep=Flight_Status Day_week);

    if Flight_Status='ontime' then
        Flight_Status_new=0;
    else if Flight_Status='delayed' then
        Flight_Status_new=1;
    else
        delete;
    Day_week_new=Day_Week;
    format day_week_new weekdays.;
run;

/* We have changed the variable Flight_Status and Day_week and given them new values
in variable Day_week_new and Flight_Status_new*/
title 'Printing the first 10 observations of converted data';

proc print data=predict.Flight_converted (obs=10) noobs;
    var Day_week Day_weeK_new Flight_Status Flight_Status_new;
run;

/* 10. Predict */
/* Showing average delays per day per carrier */
proc sql;
    create table predict.all_Delay as select CARRIER, DAY_WEEK, Delay_In_Mins from
        Predict.FormattedData group by CARRIER, DAY_WEEK, Delay_in_Mins order by
        DAY_WEEK;
quit;

title 'Report to display the data of delay in different carriers on all days';

proc report data=Predict.all_Delay;
    column CARRIER DAY_WEEK Delay_in_mins;
    define CARRIER/ group;
    define DAY_WEEK/ group;
    define delay_in_mins/ analysis mean;
    title 'Average Delays by Each Carrier per weekday';
run;

/* Showing delays per carrier */
title 'SG Plot to display the data of delay in different carriers';

proc sgplot data=Predict.all_Delay;
    scatter x=CARRIER y=Delay_in_mins /group=CARRIER;
    xaxis grid;
    yaxis grid;
run;

/* Showing delays per day */
title 'SG Plot to display the data of delay in different carriers on all days';

proc sgplot data=Predict.all_Delay;
    scatter x=DAY_WEEK y=Delay_in_mins /group=DAY_WEEK;
    xaxis grid;
    yaxis grid;
run;

title 'Report to display average delays of all carriers per day';

proc report data=Predict.all_Delay;
    column DAY_WEEK Delay_in_mins;
    define DAY_WEEK/ group;
    define delay_in_mins/ analysis mean;
run;

/*Sunday has the highest delays- best to avoid travelling on sunday*/
/*prediction to identify the best flight to take if traveling on Sunday*/
proc sql;
    create table predict.Sunday_Delay as select CARRIER, DAY_WEEK, Delay_In_Mins,
        count(FL_NUM) as No_flights from Predict.FormattedData WHERE DAY_WEEK=7 group
        by CARRIER, DAY_WEEK order by DAY_WEEK;
quit;

title 'Report to display the data of delay in different carriers on Sunday';
proc report data=Predict.Sunday_Delay;
    column CARRIER Delay_in_mins No_flights;
    define CARRIER/ group;
    define Delay_in_mins/ ORDER=internal mean analysis;
    define No_flights / ORDER=internal mean analysis;
run;
```

```
title 'SG Plot to display the data of delay in different carriers on Sunday';
proc sgplot data=Predict.Sunday_Delay(where=(DAY_WEEK=7));
    scatter x=CARRIER y=Delay_in_mins /group=CARRIER;
    xaxis grid;
    yaxis grid;
run;

/*Best to avoid RU and CO, Prefer US ,UA*. US would be the best option because there are enough flights-50.
UA only has 4*/
```