MSIS 510
Individual Assignment 3

Individual Assignment 3
Due at the beginning of the class on Nov. 21
Total: 8 points

The file "FlightDelay.csv" contains information on all commercial flights departing the Washing, DC area and arriving at New York during January 2004. For each flight, there is information on the departure and arrival airports, the distance of the route, the scheduled time and date of the flight, and so on. The variable that we are trying to predict is whether or not a flight is delayed. A second different goal is profiling flights: finding out which factors are associated with a delay. The description of the predictors is as follows:

DAY_WEEK:	day of the week
CRS_DEP_TIME:	scheduled local departure time. Broken down into 16 hours between 6:00AM and 22:00PM
DEP_TIME:	actual local departure time. Broken down into hours
ORIGIN:	DCA (Reagan National)
	IAD (Dulles)
	BWI (Baltimore-Washington Int'l)
DEST:	JFK (Kennedy)
	LGA (LaGuardia)
	EWR (Newark)
CARRIER:	eight airline codes
DISTANCE:	distance between airports in miles
Weather:	whether there was a weather-related issue

Answer all the questions in Word, and upload it to Canvas. You do not need to paste your R codes.

MSIS 510
Individual Assignment 3

1. Partition the data using the pre-generated training indexes provided in "TrainIndex.rda". You need to place "TrainIndex.rda" in the same folder as "FlightDelay.csv". You can replace the following codes for training and validation partition

```
set.seed(888)
train.index <- sample(1:nrow(delays.df), nrow(delays.df) * 0.7     )
train.df <- delays.df[train.index, ]
valid.df <- delays.df[-train.index, ]
```

   with

```
load("TrainIndex.rda")  # which stores the variable train.index
train.df <- delays.df[train.index, ]
valid.df <- delays.df[-train.index, ]
```

2. Fit a classification tree to the flight delay variable using all the relevant predictors. Do not include DEP_TIME (actual departure time) in the model since it is unknown at the time of prediction (we are generating our predictions of delays before the plane takes off). Use the default tree generated by rpart(). Describe the resulting tree in terms of rules.

3. Apply this default tree on validation set and write down the confusion matrix. **How** can you compute the accuracy based on the confusion matrix?

4. If you need to fly between DCA and EWR on a Monday, would you be able to use this tree to predict for the flight delay? What other information would you need? Is it available in practice?

MSIS 510
Individual Assignment 3

5. Use this classification tree to predict whether the following flight will be delayed or not.

CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	ORIGIN	Weather	DAY_WEEK
17_18	CO	14_15	EWR	199	DCA	No	Tue

6. Use the **same** partitioned training data and validation data as in 1). Fit another default tree, besides DEP_TIME, this time also excluding the Weather predictor.
    a. Visualize the resulting tree using prp() and paste it here.

    b. How is this tree used for classification? (What is the rule for classifying?)

    c. What can you conclude from this tree compared to the previous one? Which variable is the core for causing flight delays?