STAT 380 Midterm

Spring 2019

YOUR NAME

Due: 03/01/2019 11:59PM

Integrity statment. The work submitted for this assessment is entirely my own. I have neither given or recieved unauthorized assistance during the assessment, and/or I will speak with Dr. Beckman privately if I am now aware or later become aware of any activity that may be in violation of academic integrity policies stated in the course syllabus or Penn State policy. I understand that the content of my completed exam may not be shared, used, or reproduced for any reason without expressed written permission of Dr. Matthew Beckman (mdb268@psu.edu).

Front matter

```
# always clean up R environment
rm(list = ls())

# set RNG seed (don't change this for the midterm)
set.seed(822)
```

0.1 Introduction

This assessment will include exploration of `flights` data, which can be accessed from the `nycflights13` R package, as well as weather data provided by the United States National Climactic Data Center (NCDC) provided to you in the GitHub Classroom repository provided to accompany the take-home portion of the midterm exam.

- "nycWeather2013.csv" contains 2013 daily weather data for three airports in the New York City area.
- "nycWeather2013_DataSummary.pdf" is the data description document provided by the NCDC to accompany our specific data set.

0.2 Logistics

Logistics

Canvas: Submit your completed HTML R Notebook (with embedded Rmd) to Canvas before the due date. If you are unable to render the R Notebook, you may submit your Rmd directly for partial credit.

GitHub: You are required to make at least 4 substantive GitHub commits during your work on the midterm. Also, your completed assignment must be entirely reproducible, meaning a grader with access only to your GitHub Repo should be able to execute your Rmd document and produce your final HTML R Notebook with absolutely no modification required.

Grading

- Overall
  - [5 pts] STAT 380 style guide use

  - [5 pts] GitHub commits
  - [5 pts] Reproducibility
  - [5 pts] Overall quality
- Part 1: Principal Components Analysis of JFK Weather Station
  - [10 pts] Data preparation
  - [15 pts] PCA of JFK weather station
- Part 2: Predictive Modeling of LaGuardia Flight Delays
  - [12 pts] Data Preparation
  - [4 pts] Partition Test and Training Data
  - [12 pts] Predictive modeling
  - [10 pts] Visualization

Part 1: Principal Components Analysis of JFK Weather Station

1.1 Data preparation

Task 1.1.1 The data are "complete", but have been coded in an unconventional way. Recode all use of `NA` as `0` in the NCDC data.

Task 1.1.2 A few of the variables in the provided data cannot be used for any meaningful analysis because every entry is identical. Identify which variables they are, report a bullet list indicating the name and description of these variables, and then remove them from your data prior to analysis.

1.2 Principal components analysis for JFK weather station

Task 1.2.1 Perform principal components analysis on data from the JFK airport weather station (only). Standardize all variables prior to analysis.

Task 1.2.2 Plot the proportion of variance explained by each principal component on a scree plot.

Task 1.2.3 How many principal components would be required to explain *at least* 76% of the variability in the JFK weather station data. Be sure to show your work.

Part 2: Predictive Modeling LaGuardia Flight Delays

2.1 Data preparation

Create `LaGuardiaFlights` data according to the following instructions:

1. Subset the `flights` data to include only the carriers that had at least 1000 flights depart from LaGuardia
   - call the resulting data frame `LaGuardiaFlights`
2. `LaGuardiaFlights` should include the following variables
   - `tenMinDelay`: **response** {TRUE / FALSE}; indicates departure delay greater than 10 minutes; remove cases with missing response
   - `hour`: hour of departure using 24 hr clock
   - `weekend`: {TRUE / FALSE} indicates flight departed on a Saturday or Sunday
   - `AWND`: average wind speed reported from the weather station at LaGuardia Airport
   - `PRCP`: precipitation reported from the weather station at LaGuardia Airport
   - `SNOW`: snowfall reported from the weather station at LaGuardia Airport

- `fog`: {TRUE / FALSE} "Yes" if the weather station at LaGuardia Airport reported any of the following:
  - `WT22`: "Ice fog or freezing fog"
  - `WT01`: "Fog, ice fog, or freezing fog (may include heavy fog)", or
  - `WT02`: "Heavy fog or heaving freezing fog (not always distinguished from fog"

Task 2.1.1 How many total cases are in the `Training` set?

Task 2.1.2 What percentage of these flights departed at least 10 minutes late?

Task 2.1.3 What percentage of these flights departed on the "Weekend"?

Task 2.1.4 What percentage of these flights departed on a day with "Fog"?

2.2 Partition test and training data

Task 2.2.1 partition 20% of the `LaGuardiaFlights` data for `Test` data and use the remaining 80% for `Training` data (Note: make sure `set.seed(822)` is properly specified in the Front Matter of your Rmd document)

2.3 Predictive modeling

Task 2.3.1 Fit a null model to the `Training` data, and report the accuracy of the model.

Task 2.3.2 Fit a logistic regression model to the `Training` data. Show the resulting confusion matrix, and report the accuracy of the model.

Task 2.3.3 Fit a *k*-nearest neighbors model to the `Training` data. Show the resulting confusion matrix, and report the accuracy of the model.

Task 2.3.4 Fit a naive Bayes model to the `Training` data. Show the resulting confusion matrix, and report the accuracy of the model.

2.4 Visualization

** Task 2.4.1 Create a visualization that shows the predicted probability of 10 minute delay based on hour of departure and precipitation using a logistic regression classifier. Specifically,**

- The predicted probability of 10 minute delay should be clearly displayed for all plausible combinations of hour and precipitation.
- overlay `Test` data with clear indication of the observed response
- demonstrate good plotting practices