Stat 21 Homework 4

Your name here   Collaborators: [list any collaborators here]

Due: Oct 13th, by 6pm ET

This assignment is due on to be submitted on Gradescope on **October 13** by **6:00pm ET**. Please use the `homework-q-and-a` channel on Slack to post any related questions or error messages.

General instructions for all assignments:

You must submit your completed assignment as a single **PDF** document uploaded to **Gradescope**. For instructions on how to do this, please watch this 2 minute video: https://youtu.be/KMPoby5g_nE. You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and the requested output. You can find a R markdown homework template here: http://www.swarthmore.edu/NatSci/sthornt1/Stat021/Stat21.html. Please make sure each problem is **clearly labeled** and that any handwritten components (such as pictures or equations) are easily readable as pictures within the R markdown document. You may want to use a service like CamScanner (https://www.camscanner.com/) to help you upload handwritten pages.

You are allowed to work with your classmates on this homework assignment but you must disclose the names of anyone you collaborate with at the top of your solutions. Each homework assignment is worth 20 points. One problem will be chosen at random to be graded for correctness and the other problems will be graded for completion. At the end of the semester, your lowest homework grade will be dropped. No homework solutions will be provided.

- Use this file as the template for your submission. You can delete unnecessary text (e.g. these instructions) but make sure to keep the nicely formatted "Problem 1", "Problem 2", "a.", "b.", etc
- Upload your knitted HTML or PDF file to the Homework 1 submission section on Gradescope. Name this file as: [SwatID]_stat21_hw03.pdf (e.g. and "sthornt1_stat21_hw04.pdf"). You only need to upload one file, but please make sure that your graphs, code, and answers to each question appear in the appropriate place. If we cannot see your code/graphs/answers, we cant give you credit for your work!
- Your file should contain the code to answer each question in its own code block. Your code should produce plots/output that will be automatically embedded in the output file.
- Each answer must be supported by written statements (unless otherwise specified).
- Include the name of anyone you collaborated with at the top of the assignment.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `gcookbook`, `knitr`

Use the information below to answer problems 1-4.

To see how much of a difference time of day makes on the speed at which she could download files, a college sophomore performed an experiment. She placed a file on a remote server and then proceeded to download it at three different time periods of the day (Early - 7:00am, Evening - 5:00pm, Late night 12:00am). She downloaded the file 48 times in all, 16 times at each time of day and recorded the time (in seconds) that the download took.

The following R code contains this data set, labeled `download_data`.

```
download_data <- tibble(time_of_day = c(rep("Early",16), rep("Evening",16), rep("Late Night", 16)),
                        time_sec = c(68,138,75,186,68,217,93,90,71,154,166,130,72,81,76,129,299,367,331
head(download_data)
```

```
## # A tibble: 6 x 2
##   time_of_day time_sec
##   <chr>          <dbl>
## 1 Early             68
## 2 Early            138
## 3 Early             75
## 4 Early            186
## 5 Early             68
## 6 Early            217
```

Problem 1

Identify the predictor and response variables. How many degrees of freedom will the treatment sum of squares term have? How many degrees of freedom will the sum squared error term have?

Solution Problem 1:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 2

Visualize the data with relevant box plots and normal probability plots. Do the assumptions necessary for an ANOVA test seem reasonable here? (Note, you'll need a normal probability plot for each of the times of day, the `filter()` function will come in handy here.)

Solution Problem 2:

[Write your solution here.]

Problem 3

State the null and alternative hypothesis for an ANOVA test of association between time of day and downloading speed. Perform the ANOVA test and report your conclusion in the context of the problem.

Solution Problem 3:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 4

Use the R function `TukeyHSD()` to preform a controlled, pairwise comparison of the difference in downloading time among all possible paired times of day. Interpret the results.

Solution Problem 4:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Use the information below to answer problems 5 and 6.

A college basketball player is taking a statistics course and has just learned about ANOVA models. She is curious about whether or not there is a difference in the average height (in in) NBA players can jump, depending on their team. She collects data on the maximum standing vertical leap distance for a random selection of $n = 42$ NBA players from her three favorite teams.

The following R code contains this data set, labeled `jump_data`. (Note this is made-up data.)

```
jump_data <- tibble(team = c("Boston Celtics","Chicago Bulls","Miami Heat","Boston Celtics","Miami Heat
                    height = c(36.1, 35.0 ,34.0 ,33.7 ,33.0, 33.0, 32.5, 32.4, 32.5, 32.0, 32.0, 31.6, 3
30.5, 29.9, 30.0, 29.4, 29.6, 29.5, 29.4, 18.9, 29.1, 29.0, 28.5, 28.4, 28.1, 27.5, 27.6, 27.1, 26.9, 2
head(jump_data)
```

```
## # A tibble: 6 x 2
##   team           height
##   <chr>           <dbl>
## 1 Boston Celtics   36.1
## 2 Chicago Bulls    35
## 3 Miami Heat       34
## 4 Boston Celtics   33.7
## 5 Miami Heat       33
## 6 Miami Heat       33
```

Problem 5

Create three box plots to visualize the jump height for the players from these three teams. (Don't forget to make sure that team is understood to be a factor variable in R.) Then create three normal probability plots for the jump height scores for each time. Remember to standardize the scores first. (Also, you may want to use the `filter` function in R to extract and plot the data for each of the three teams separately.) Comment on the appropriateness of the ANOVA assumptions based on these plots.

Solution Problem 5:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 6

Let $\mu_{Celtics}$, μ_{Bulls}, μ_{Heat} represent the average jump height for each of the three basketball teams. State the null and alternative hypotheses for the ANOVA test. Perform the test in R and report the conclusions at an $\alpha = 0.10$ level of significance.

Solution Problem 6:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Use the information below to answer problems 7-10.

The increase in the passage of voter identification laws ("Voter ID") has been labelled as a push to reduce fraudulent or illegal voting in the United States. Most of these laws have come into effect by requiring voters to present a photo ID either at time of registration or at the time of voting. The National Conference of State Legislatures (NCSL) identified four different levels of voter identification requirements: Strict Photo ID Law, Non-Strict Photo ID, Strict Non-Photo ID Law, and NonStrict Non-Photo ID. Non-strict identification laws are defined as "at least some voters without acceptable identification have an option to cast a ballot that will be counted without further action on the part of the voter". Strict identification laws are defined as "voters without acceptable identification must vote on a provisional ballot and also take additional steps after Election Day for it to be counted."

In 2004, not a single state enforced a photo identification requirement at the polls, but since then the numbers have increased drastically. Although designed to reduce fraudulent voting, several studies have found that there are a very small number of fraudulent votes to begin with, so the laws have very little impact. Rather, these laws have acted to raise the costs of voting and reduce convenience, especially for certain groups of voters who are unlikely to have a photo ID.

The following (real) data set contains information on what level of voter identification law was in place for each state in the 2016 US presidential election (no low, non-strict non-photo ID law, strict non-photo ID law, non-strict photo ID law, strict photo ID law) and what was the percentage of voter turnout for each state that year. (No law states do not require a government issued ID in order to submit a ballot.)[1]

```
voter_turnout <- tibble(law_type = c(rep("No law",18),
                                     rep("Non-strict non-photo", 14),
                                     rep("Strict non-photo",2),
                                     rep("Non-strict photo",8),
                                     rep("Strict photo",8)),
                        turnout_rate = c(52.8,56.5,61.6,68.4,70.7,66.4,67.2,74.1,62.8,57.4,64.4,54.5,56.8,64
head(voter_turnout)
```

```
## # A tibble: 6 x 2
##   law_type turnout_rate
##   <chr>           <dbl>
## 1 No law           52.8
## 2 No law           56.5
## 3 No law           61.6
## 4 No law           68.4
## 5 No law           70.7
## 6 No law           66.4
```

Problem 7

Suppose we want to test whether or not there is an association between the voter ID laws a state implements and the voter turnout rate. State the null and alternative hypotheses for a relevant ANOVA test. The report the p-value of the test and interpret the conclusion in the context of the problem.

Solution Problem 7:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 8

Create five normal probability plots, one for each type of voter ID law. Does the normality condition seem reasonable here? Disregarding the plots, does the independence assumption seem reasonable?

Solution Problem 8:

```
## Uncomment this line and put any r-code you used for your solution here
```

[1] Source for voter turnout data: https://digitalcommons.colby.edu/cgi/viewcontent.cgi?article=1980&context=honorstheses

Problem 9

Display the ANOVA table for this model. (Make sure the categorical variable is a factor in R.) What is an estimate of the variance of voter turnout among each of the states with different voter ID laws?

Solution Problem 9:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 10

Plot the data with box plots for voter turnout. Does the equal variance assumption seem reasonable? What does this mean about the estimate in Problem 9?

Solution Problem 10: [Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```