Homework 2

Problem 1 - Capital Metro UT Ridership

The file `capmetro_UT.csv` contains data from Austin's own Capital Metro bus network, including shuttles to, from, and around the UT campus. These data track ridership on buses in the UT area. Ridership is measured by an optical scanner that counts how many people embark and alight the bus at each stop. Each row in the data set corresponds to a 15-minute period between the hours of 6 AM and 10 PM, each and every day, from September through November 2018. The variables are:

- *timestamp*: the beginning of the 15-minute window for that row of data
- *boarding*: how many people got on board any Capital Metro bus on the UT campus in the specific 15 minute window
- *alighting*: how many people got off ("alit") any Capital Metro bus on the UT campus in the specific 15 minute window
- *day_of_week* and *weekend*: Monday, Tuesday, etc, as well as an indicator for whether it's a weekend.
- *temperature*: temperature at that time in degrees F
- *hour_of_day*: on 24-hour time, so 6 for 6 AM, 13 for 1 PM, 14 for 2 PM, etc.
- *month*: July through December

Your task in this problem is **to make two faceted plots** and to answer questions about them.

1. One faceted line graph that plots **average boardings** by hour of the day, day of week, and month. You should facet by day of week. Each facet should include three lines of average boardings (y) by hour of the day (x), one line for each month and distinguished by color. Give the figure an informative caption in which you explain what is shown in the figure and also address the following questions, citing evidence from the figure. Does the hour of peak boardings change from day to day, or is it broadly similar across days? Why do you think average boardings on Mondays in September look lower, compared to other days and months? Similarly, why do you think average boardings on Weds/Thurs/Fri in November look lower? (Hint: wrangle first, then plot.)

2. One faceted scatter plot showing boardings (y) vs. temperature (x), faceted by hour of the day, and with points colored in according to whether it is a weekday or weekend. Give the figure an informative caption in which you explain what is shown in the figure and also answer the following question, citing evidence from the figure. When we hold hour of day and weekend status constant, does temperature seem to have a noticeable effect on the number of UT students riding the bus?

These are exactly the kind of figures that Capital Metro planners might use to understand seasonal and intra-week variation in demand for UT bus service. These are also the kind of figures one would create in the process of building a model to predict ridership.

Notes. First, a feature of R is that it orders categorical variables alphabetically by default. This doesn't make sense for something like the day of the week or the month of the year. To reorder the days of the week and months in appropriate order, paste the following block of code into your R script at the top and execute it before you start further work on your plots for this problem:

```
# Recode the categorical variables in sensible, rather than alphabetical, order
capmetro_UT = mutate(capmetro_UT,
               day_of_week = factor(day_of_week,
                 levels=c("Mon", "Tue", "Wed","Thu", "Fri", "Sat", "Sun")),
               month = factor(month,
                 levels=c("Sep", "Oct","Nov")))
```

Second, please keep each figure + caption to a single page combined (i.e. two pages, one page for first figure + caption, a second page for second figure + caption).

Problem 2: Wrangling the Billboard Top 100

Consider the data in `billboard.csv` containing every song to appear on the weekly Billboard Top 100 chart since 1958, up through the middle of 2021. Each row of this data corresponds to a single song in a single week. For our purposes, the relevant columns here are:

- performer: who performed the song
- song: the title of the song
- year: year (1958 to 2021)
- week: chart week of that year (1, 2, etc)
- week_position: what position that song occupied that week on the Billboard top 100 chart.

Use your skills in data wrangling and plotting to answer the following three questions.

Part A: Make a table of the top 10 most popular songs since 1958, as measured by the *total number of weeks that a song spent on the Billboard Top 100.* Note that these data end in week 22 of 2021, so the most popular songs of 2021 will not have up-to-the-minute data; please send our apologies to The Weeknd.

Your table should have **10 rows** and **3 columns**: `performer`, `song`, and `count`, where `count` represents the number of weeks that song appeared in the Billboard Top 100. Make sure the entries are sorted in descending order of the `count` variable, so that the more popular songs appear at the top of the table. Give your table a short caption describing what is shown in the table.

(*Note*: you'll want to use both `performer` and `song` in any `group_by` operations, to account for the fact that multiple unique songs can share the same title.)

Part B: Is the "musical diversity" of the Billboard Top 100 changing over time? Let's find out. We'll measure the musical diversity of given year as *the number of unique songs that appeared in the Billboard Top 100 that year.* Make a line graph that plots this measure of musical diversity over the years. The x axis should show the year, while the y axis should show the number of unique songs appearing at any position on the Billboard Top 100 chart in any week that year. For this part, please filter the data set so that it excludes the years 1958 and 2021, since we do not have complete data on either of those years. Give the figure an informative caption in which you explain what is shown in the figure and comment on any interesting trends you see.

There are number of ways to accomplish the data wrangling here. We offer you two hints on two possibilities:

1) You could use two distinct sets of data-wrangling steps. The first set of steps would get you a table that counts the number of times that a given song appears on the Top 100 in a given year. The second set of steps operate on the result of the first set of steps; it would count the number of unique songs that appeared on the Top 100 in each year, *irrespective of how many times* it had appeared.
2) You could use a single set of data-wrangling steps that combines the `length` and `unique` commands.

Part C: Let's define a "ten-week hit" as a single song that appeared on the Billboard Top 100 for at least ten weeks. There are 19 artists in U.S. musical history since 1958 who have had *at least 30 songs* that were "ten-week hits." Make a bar plot for these 19 artists, showing how many ten-week hits each one had in their musical career. Give the plot an informative caption in which you explain what is shown.

Notes:

1) As with the previous problem, you might find this easier to accomplish in two distinct sets of data wrangling steps.
2) Make sure that the individuals names of the artists are readable in your plot, and that they're not all jumbled together. If you find that your plot isn't readable with vertical bars, you can add a `coord_flip()` layer to your plot to make the bars (and labels) run horizontally instead.
3) By default a bar plot will order the artists in alphabetical order. This is acceptable to turn in. But if you'd like to order them according to some other variable, you can use the `fct_reorder` function, described in this blog post. This is optional.

Problem 3: regression practice

Download the data in `creatinine.csv`. Each row is a patient in a doctor's office. The variables are:

- age: patient's age in years.
- creatclear: patient's creatine clearance rate in mL/minute, a measure of kidney health (higher is better).

Use this data, together with your knowledge of linear regression, to answer three questions:

A) What creatinine clearance rate should we expect for a 55-year-old?
B) How does creatinine clearance rate change with age? (This should be a number with units ml/minute per year.)
C) Whose creatinine clearance rate is healthier (higher) for their age: a 40-year-old with a rate of 135, or a 60-year-old with a rate of 112? Briefly explain your reasoning.

Problem 4: probability practice

Part A. A shady used car dealer has 30 cars, and 10 of them are "lemons" (that is, mechanically faulty used cars), but you don't know which cars they are. If you buy 3 cars, what is the probability that you will get at least one lemon?

Part B. We throw two dice (each with the usual 6 sides, numbered 1-6). What is the probability that the sum of the two numbers is odd? What is the probability that the sum of the two numbers is less than 7? What is the probability that the sum of the two numbers is less than 7, given that it is odd? Are these two events independent? Hint: just count cases.

Part C. Visitors to your website are asked to answer a single survey question before they get access to the content on the page. Among all of the users, there are two categories: Random Clicker (RC), and Truthful Clicker (TC). There are two possible answers to the survey: yes and no. Random clickers would click either one with equal probability. You are also giving the information that the expected fraction of random clickers is 0.3. After a trial period, you get the following survey results: 65% said Yes and 35% said No. What fraction of people who are truthful clickers answered yes? Hint: use the rule of total probability.

Part D. Imagine a medical test for a disease with the following two attributes:

- The sensitivity is about 0.993. That is, if someone has the disease, there is a probability of 0.993 that they will test positive.
- The specificity is about 0.9999. This means that if someone doesn't have the disease, there is probability of 0.9999 that they will test negative.
- In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it (or 0.000025 as a decimal probability).

Suppose someone tests positive. What is the probability that they have the disease?

Part E. If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. Suppose that on average across all days, an aircraft is present with probability 0.05. Let the events A and R be defined as follows: A = an aircraft is present, R = the radar registers an aircraft presence. What is $P(A \mid R)$, the conditional probability that an aircraft is present, given that the radar registers an aircraft presence?

Problem 5: modeling soccer games with the Poisson distribution

Go read the article "One match to go!", by Spiegelhalter and Ng. In this article, the authors describe how they formulated an approach for predicting the probability of different outcomes for soccer matches based on "attack strength" and "defense weakness," all within the context of a Poisson model. It is better than the simple approach we took in class, though probably not as good as what actual bookmakers (i.e. in Las Vegas) use.

Now go get the data from the 2018-19 English Premiere League soccer season. These are in the files "epl_2018-19_away.csv" and "epl_2018-19_home.csv", which give the home and away performance for all 20 teams. These two files allow you to replicate the analysis described in the "One Match to Go!" article. Specifically, the columns of interest in each file are "GF" and "GA", which are "Goals For" and "Goals Against," respectively. So, for example, in the "epl_2018-19_away.csv" you'll notice that Manchester City has 38 for GF and 11 for GA. That means in their away games that season, Manchester City scored 38 goals and allowed 11 goals by their opponents. Note: these data were downloaded from here.

Replicate Spiegelhalter and Ng's approach using the 2018-19 data to answer the following two questions:

1. What are your estimated probabilities of win/lose/draw results for a match between Liverpool (home) and Tottenham (away)?
2. What about Manchester City (home) versus Arsenal (away)?

Notes and requirements:

- Spiegelhalter and Ng did the calculations for the whole League, but you certainly don't need to; these two games will suffice. Note also that you'll use the full season's worth of data to estimate the model they describe.
- This is one problem where you might find it easier to use a spreadsheet program rather than R. This is perfectly acceptable. Just be sure to explain what you've done; no need to submit a spreadsheet if you go this route.
- Your write-up doesn't need the break down the results by all possible game scores (1-0, 1-1, etc); just summarize the probability of a draw or win for each team. The one exception is that, in your "Approach" section, you should include an explicit math formula, based on the PMF of a Poisson distribution, that shows how you could calculate the probability of a 2-1 victory for the home team in each of the two games. This formula will serve as an example of your approach for calculating all the necessary probabilities. You can insert an equation directly into the document, or you can simply include a handwritten equation that you took a photo of.
- Format your write-up in the following sections, some of which might be quite short:
  1) Question: What questions are you trying to answer?
  2) Approach: What approach/statistical tool did you use to answer the questions?
  3) Results: What evidence/results did your approach provide to answer the questions? (E.g. any numbers, tables, figures as appropriate.)
  4) Conclusion: What are your conclusions about your questions? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set.
- Summarize the Spiegelhalter and Ng approach in your own words in your Approach section of your write-up. It's fine to assume independence between the teams' scores but state this assumption in the Approach section.
- The course references show you two ways to solve this kind of problem in R: one using the PMF of the Poisson distribution and the other using Monte Carlo simulation. Make sure you are clear in your Approach section about which overall method you used.
- Don't get confused by the "Pts" (points) column in the data sets. This isn't goals. "Points" are how the league crowns a winner, with a team getting 3 points for a win and 1 point for a draw.
- GP means "games played."