Homework 1

Important notes:

Submit via Canvas (Assignments tab).

Include your name on your write-up.

Turn in a PDF *or* a single link. The only acceptable submission format is PDF. As an alternative, you can post a single link to your write-up (most commonly a Google doc, but e.g. a published PDF or Markdown file on GitHub also works).

Submit early and often. If you are trying to upload your solution for the first time 1 second before the deadline and your wifi breaks, we will be sympathetic to your situation, but we will not relax course policies regarding deadlines or penalties for late homework submission. This is to prevent obvious forms of abuse and to ensure a level playing field for all students. You can also submit multiple times before the deadline! If you intend to be working on a problem right up until the deadline, we suggest submitting an early draft of your homework to "lock in credit" for the problems you've already solved, and then replacing that submission later on Canvas, once you've made your finishing touches. Canvas will save your last submission, over-writing any previous submissions.

Do not include your raw R code in your PDF write-up unless we explicitly ask for it.

Most people in the past have just created their write-ups in Word (or a similar program like Google Docs, Pages, etc.), pasted tables and plots in from R as required, and then saved the result as a PDF to upload to Canvas. This is the most popular option, with virtually no learning curve. It works great. However, a very cool, more advanced (but fun) option is to use a package called RMarkdown, which allows you to use R itself to produce publication-quality tables, plots, and shareable documents. (One of the main course resources was written in RMarkdown – you can see all the raw source code here.) It's *totally optional* for this course, but if you used your homework assignments as an excuse to learn RMarkdown, you'd be picking up a valuable, marketable skill, and your write-ups would look highly professional. For those familiar with the idea of a Jupyter Notebook from the Python world, RMarkdown is pretty similar.

Problem 1: playlists revisited

Return to our playlist data from a popular music streaming service contained in `plays_top50.csv`.

Part A

Consider two music streaming events: "plays Daft Punk" and "plays David Bowie". These variables are called `daft.punk` and `david.bowie` in this data set. Using the R functions `xtabs` and `prop.table`, make a 2x2 table of conditional probabilities, conditional on the levels of the `david.bowie` variable. The bottom right entry in your 2x2 table should display P(plays Daft Punk | plays David Bowie). Round the entries in your table to 3 decimal places.

You should not include any R syntax in your write-up, but you will need to report the table that you generate in the console. To transfer this table into your write-up, the simplest option is to copy and paste the table from R's console output. If you do this, **make sure you use a fixed-width font**, like Courier, to display the table—otherwise the columns won't line up properly. You can also format tables in another program if you'd prefer to do it that way (e.g., copying and pasting R table contents into Excel and adding borders for a nicer presentation).

Part B

Are the events "plays Johnny Cash" and "plays Pink Floyd" independent? Why or why not? Provide numerical evidence to support your answer.

Problem 2: Super Bowl ads

In February of 2021, the website fivethirtyeight.com ran a story that looked for interesting patterns in Super Bowl ads over the years. Here's how they described their approach:

> Like millions of viewers who tune into the big game year after year, we at FiveThirtyEight LOVE Super Bowl commercials. We love them so much, in fact, that we wanted to know everything about them ... by analyzing and categorizing them, of course. We dug into the defining characteristics of a Super Bowl ad, then grouped commercials based on which criteria they shared—and let me tell you, we found some really weird clusters of commercials.
>
> We watched 233 ads from the 10 brands that aired the most spots in all 21 Super Bowls this century, according to superbowl-ads.com. While we watched, we evaluated ads using seven specific criteria, marking every spot as a "yes" or "no" for each.

Go read the full story here. Make sure to download the corresponding data set, called superbowl.csv, from our Canvas site. (But watch the commercials at your own risk—you might not believe the levels of insensitivity in some of these ads that passed for socially acceptable even 10 or 15 years ago.)

Then come back to here to answer some questions about this data set, which has the following variables in it:

variable	type	description
year	number	Superbowl year
brand	categorical	Brand for commercial
superbowl_ads_dot_com_url	character	Superbowl ad URL
youtube_url	character	Youtube URL
funny	categorical	Contains humor
show_product_quickly	categorical	Shows product quickly
patriotic	categorical	Patriotic
celebrity	categorical	Contains celebrity
danger	categorical	Contains danger
animals	categorical	Contains animals
use_sex	categorical	Uses sexuality
view_count	number	Youtube view count
like_count	number	Youtube like count
dislike_count	number	Youtube dislike count

Please use this data to answer the following questions.

Part A

The authors drew attention to a cluster of commercials that they described as "DANGER + NOT TRYING TO BE FUNNY." As they put it:

> These ads probably aren't what you think of first when it comes to Super Bowl commercials. They feature danger, violence or injury, but not as the punchline of a joke. This cluster is home to a few real tear-jerkers and some attempts at inspirational unity.

That made us wonder: what's the relationship between danger and humor across *all* Super Bowl commercials in the sample?

To address this question, please use the data to estimate the following probabilities:

- P(danger = TRUE)
- P(danger = TRUE | funny = TRUE)

- P(danger = TRUE | funny = FALSE)

Please round your numbers to two decimal places. In light of these numbers, does it seem that ads using humor are *more* or *less* likely to feature danger than ads not using humor? Or, on the other hand, do `humor` and `danger` look nearly independent of each other?

Part B

The article also described a cluster of ads that bizarrely seemed to juxtapose "selling with sex" and "animals." As the author put it:

> There was a wide range of approaches in how advertisers combined these categories, though, with some more disturbing than others. On one end are ads that sell sex while an animal happens to be in one of the shots — the Bob Dole Pepsi ad shows him walking on the beach with his dog, and a Budweiser ad that centers on some crabs stealing a cooler of beer makes sure to sneak in frames of women in bikinis. These ads sell sex, and these ads have animals, but they're not really fundamentally intertwined.
>
> At the other unholy end of the spectrum, though, are Bud Light ads in which a talking chimp hits on a woman and a falcon brings back a woman's bra to its handler after attacking a city block on the hunt for beer. The only thing more unsettling than watching these bizarre commercials is realizing a whole boardroom approved these concepts for what was a likely multimillion-dollar ad spot. The commercials in this cluster really cover the full spectrum, so watch at your own risk.

Following on from this, please use the data to estimate the following probabilities. Round to two decimal places.

- P(animals=TRUE)
- P(animals=TRUE | use_sex=TRUE)
- P(animals=TRUE | use_sex=FALSE)

In light of these numbers, does it seem that ads using sexuality are *more* or *less* likely to feature animals than ads not using sexuality? Or, on the other hand, do `use_sex` and `animals` look nearly independent of each other?

Part C

The authors also highlighted a cluster of ads that combined "patriotic symbolism with celebrity endorsements."

Following on from this, please use the data to estimate the following probabilities. Round to two decimal places.

- P(celebrity=TRUE)
- P(celebrity=TRUE | patriotic=TRUE)
- P(celebrity=TRUE | patriotic=FALSE)

In light of these numbers, does it seem that ads using patriotic symbolism are *more* or *less* likely to feature celebrity endorsements than ads not using patriotic symbolism? Or, on the other hand, do `celebrity` and `patriotic` look nearly independent of each other?

Problem 3: Beauty, or not, in the classroom

The University of Texas at Austin, like every major university in the country, asks students to evaluate their courses and professors.The `profs.csv` file contains data on course-instructor evaluation surveys from a sample of 463 UT Austin courses. These data represent evaluations from 25,547 students and most major

academic departments. The data frame also includes information on characteristics of each course and facts about the professors such as seniority and demographics. Also included is a rating of each instructor's physical attractiveness, as judged by a panel of six students (3 males, 3 females) who were shown photos of the instructors. Key variables in the `prettyprofs.csv` data frame are:

- `eval`: the instructor's average teaching evaluation score, on a scale of 1 to 5, for courses in the sample
- `beauty`: the six panelists' average rating of the professor's physical attractiveness, shifted to have a mean of zero. For example, 2 is two points above average and -1 is one point below average.
- `minority`: is the professor from a non-white racial or ethnic minority?
- `age`: the professor's age in years
- `gender`: indicator of the professor's gender
- `credits`: indicator of whether the course is a single-credit elective ("single") or an academic course ("more")
- `division`: indicator of whether the course is a lower or upper division course
- `native`: indicator of whether the professor is a native English speaker
- `tenure`: indicator of whether the professor has tenure/is on the tenure track, or not
- `students`: the number of students who participated in the course evaluation survey
- `allstudents`: the number of students enrolled in the course
- `prof`: unique identifier variable for the professor

Use these data to address the following questions by creating plots and/or calculating summary statistics.

Format the plot professionally with clear labeling and consideration of best practices for effective plots.

Include in your write-up an image of each plot along with an informative caption below each plot. The caption may be typed in your write-up below the plot and does not have to be generated using ggplot's caption feature. The caption should consist of 1-2 sentences describing key features of the plot (if these are not already clear from the chart title and labels) and a short summary of key takeaways from your plot in its relevant context. Think of this caption as a walkthrough for your plot audience.

Part A. Create a histogram to display the overall data distribution of course evaluation scores.

Part B. Use side-by-side boxplots to show the distribution of course evaluation scores by whether or not the professor is a native English speaker.

Part C. Use a faceted histogram with two rows to compare the distribution of course evaluation scores for male and female instructors.

Part D. Create a scatterplot to visualize the extent to which there may be an association between the professor's physical attractiveness (x) and their course evaluations (y).

Problem 4: SAT scores for UT students

The data in `utsat.csv` contains the SAT scores and graduating college GPAs for every UT student who entered UT in a specific, recent year (it's definitely this century, but I'm not saying which just to maximize data anonymity here), and went on to graduate from UT within 6 years. The variables in this data set are:

- SAT.V: score on the verbal section of the SAT (200-800)
- SAT.Q: score on the quantitative section of the SAT (200-800)
- SAT.C: combined SAT score
- School: college or school at which the student first matriculated (not necessarily where they ended up)
- GPA: college GPA upon graduation, on a 4-point scale
- Status: this should be G, for graduated, for everyone in this data set

Your task in this problem is to make a single table of summary statistics of this data. The table should show the following summary statistics for SAT verbal scores, SAT math scores, and graduating GPA across the whole sample: mean, standard deviation, inter-quartile range (IQR), 5th percentile, 25th percentile, median (50th percentile), 75th percentile, and 95th percentile.

Your table should have three rows (SAT Verbal, SAT Quantitative, GPA) and 8 columns of numbers (one column for each summary statistic). Use R to calculate these summary statistics, and then format them into a nice, professional-looking table using a program like Word, Excel, Google Sheets, etc. Paste this nicely-formatted table into your write-up.[1]

Include a caption for your table. The caption may be typed in your write-up below the table and does not have to be generated using an R command. The caption should consist of 2-3 sentences describing what the table shows. Think of this as an orientation to the table for a reader who's encountering it for the first time.

Problem 5: bike sharing

Bike-sharing systems are a new generation of traditional bike rentals where the whole process from rental to return is automatic. There are thousands of municipal bike-sharing systems around the world (e.g. Citi bikes in NYC or "Boris bikes" in London), and they have attracted a great deal of interest because of their important role in traffic, environmental, and health issues—especially in the wake of the Covid-19 pandemic, when ridership levels on public-transit systems have plummeted.

These bike-sharing systems also generate a tremendous amount of data, with time of travel, departure, and arrival position recorded for every trip. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility patterns across a city.

Bike-sharing rental demand is highly correlated to environmental and seasonal variables like weather conditions, day of week, time of year, hour of the day, and so on. In this problem, you'll look at some of these demand-driving factors using the `bikeshare.csv` data from the course Canvas page. This data set contains a two-year historical log (2011 and 2012) from the Capital Bikeshare system in Washington D.C. The raw data is publicly available at http://capitalbikeshare.com/system-data. These data have been aggregated on an hourly and daily basis and then merged with weather and seasonal data.

The variables in this data set are as follows:

- instant: unique record identifier for each row
- dteday: date
- season: season (1:spring, 2:summer, 3:fall, 4:winter)
- yr: year (0: 2011, 1:2012)
- mnth: month (1 to 12)
- hr: hour (0 to 23)
- holiday: whether the day is holiday or not
- weekday: day of the week (1 = Sunday)
- workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit: a weather situation code with the following values
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp: Normalized temperature in Celsius. The actual values are divided by 41 (max)
- total: count of total bike rentals that hour, including both casual and registered users

Your task in this problem is to prepare three figures. To make these figures, you will need to combine the ideas from our lesson on Plots with our lesson on Data wrangling. In other words, you won't be able to make these plots by calling `ggplot` on the raw data we've provided. First, you'll need to use some of our six key

[1] If you're using RMarkdown, feel free to read up on the `kable` function in the `knitr` library. You can see an example of how I used this function in our course packet here. And if you ever need help with RMarkdown, just ask!

data verbs from the Data Wrangling lesson to get the data into an appropriate form. Only then will you actually be able to create these plots.

- Plot A: a line graph showing average hourly bike rentals (`total`) across all hours of the day (`hr`).
- Plot B: a faceted line graph showing average bike rentals by hour of the day, faceted according to whether it is a working day (`workingday`).
- Plot C: a faceted bar plot showing average ridership (y) **during the 9 AM hour** by weather situation code (`weathersit`, x), faceted according to whether it is a working day or not. (Remember that you can focus on a specific subset of rows of a data set using `filter`.)

Your write-up should include each plot, together with an informative caption (i.e., written paragraph) below each plot. Think of this caption paragraph as a walkthrough for your plot audience – perhaps what you would say in a live presentation to incorporate the plot into your narrative. Your caption should clearly explain the plot itself (e.g., what the axes represent and what the panels show). Don't forget to specify variable units. The caption should also contain a one-sentence *take-home lesson* of what we have learned about ridership patterns from the plot.