Homework 4

Problem 1

For this exercise, you'll need to download the `GasPrices.csv` data set from the class website. This data set came from a student project in the spring of 2016, in a different class. It was a pretty awesome project! We'll let the students who did the project describe things in their own words:

> Have you ever been driving through town looking to make a quick stop to fill up your car with gas and noticed that different gas stations are advertising different gas prices? Have you ever stopped to wonder why this might be the case? Could there be some underlying factors responsible for this noticeable difference in price, specifically for the same, regular unleaded mix of gas on the same day at the same time?
>
> To observe prices and other traits of gas stations firsthand, we visited 101 gas stations in the Austin area. We split the city into east and west sections with Lamar Blvd.~serving as the dividing line. At each gas station, we observed all necessary characteristics while staying in the car. We used the Maps app to determine the address and zip codes of the gas stations and the transportation feature within Maps on the iPhone to locate the gas stations themselves. We input the data directly into an Excel spreadsheet. Once we had visited all 101 gas stations, we used the US Census Bureau's American Fact Finder to input the median income for each zip code.

Needless to say, these students knocked it out of the park for effort. Let's look at their data set and use it to answer some questions. There are lots of variables in this data set, but for our purposes here, the important ones are as follows:

- ID: Order in which gas stations were visited
- Name: Name of gas station
- Price: Price of regular unleaded gasoline, gathered on Sunday, April 3rd, 2016
- Highway: Is the gas station accessible from either a highway or a highway access road?
- Stoplight: Is there a stoplight in front of the gas station?
- Competitors: Are there any other gas stations in sight?
- Zipcode: Zip code in which gas station is located
- Income: Median Household Income of the ZIP code where the gas station is located based on 2014 data from the U.S. Census Bureau
- Brand: ExxonMobil, ChevronTexaco, Shell, or Other.

The theories

People have a lot of pet theories about what explains the variation in prices between gas stations. Here are several such theories:

A) Gas stations charge more if they lack direct competition in sight.

B) The richer the area, the higher the gas prices.

C) Gas stations at stoplights charge more.

D) Gas stations with direct highway access charge more.

E) Shell charges more than all other non-Shell brands.

Which of these theories seem true, and which are unsupported by data? Take each theory one by one and assess the evidence for the theory in this data set.

Your discussion of each theory should include three mini-sections:
- Claim: a statement of the theory itself.
- Evidence: the evidence for or against the theory, in the form of any relevant numerical and/or visual summaries If the theory looks correct, provide an estimate of the effect size: that is, how large is the difference (e.g. for highway vs. non-highway gas stations) and/or association (e.g. between income and price)? If the theory is unsupported by the data, explain why.
- Conclusion: your conclusion about whether the theory is supported or unsupported by the data.

No single theory should require more than a single typed page to assess, including any figures, tables, or numbers you use to make your case. Less than 1 page per theory is perfectly acceptable—indeed, preferable, as long as you can cover all the bases.

In assessing the evidence for each theory, make sure you appropriately deal with a major issue: **uncertainty.** Remember, this is just a sample of gas stations, and there is always uncertainty when generalizing from a specific sample to a wider population. Therefore, you need to quote confidence intervals rather than just single-number estimates for these differences. So, for example, it's not enough to say something like, "the difference in price between gas stations on and off the highway is X cents," and to draw your conclusion from this statement. Instead, you have to say something like, "the difference in price between gas stations on an off the highway is somewhere between L and U, with 95% confidence," and to draw your conclusion from this interval.

Problem 2

The file `sclass.csv` contains data on nearly 30,000 used Mercedes S-Class vehicles sold on cars.com. These are big, very expensive, luxurious cars used frequently by chauffeurs. The variables of interest here are `price`, `mileage`, `trim` (i.e. submodel), `color`, and `year`.

Use your knowledge of statistical inference to answer the following questions. In each case, make a *sensible choice* regarding how much to round your answers (i.e,. don't ask the instructor team).

Part A: Filter the data set down to include only those cars where `year == 2011` and `trim == "63 AMG"`. Based on these 116 cars, compute a 95% bootstrap confidence interval for the average mileage of 2011 S-Class 63 AMGs that were hitting the used-car market when this data was collected.

Part B: Filter the data set down to include only those cars where `year == 2014` and `trim == "550"`. Based on this sample of 2889 cars, compute a 95% bootstrap confidence interval for the proportion of all 2014 S-Class 550s that were painted black. Hint: you might find this easiest if you use `mutate` to first define a new variable, `isBlack`, that is either TRUE or FALSE depending on whether the car is black.

Problem 3

Like any TV network, NBC conducts does market research on how viewers respond to TV shows (both its own shows and the shows of other competing networks). The data in nbc_pilotsurvey.csv contains the results of some of that research. Each row of this data frame shows the responses of a single viewer (the `Viewer` variable) to the "pilot" episode[1] of a single TV show (`Show` variable). The remaining variables encode the viewers reactions to the show. Viewers were asked to rate the strength of their agreement on a 1-5 scale (where 5 means "strongly agree") with various statements about the show, such as "This show made me feel happy" or "I found this show confusing."[2]

Use this data to answer the questions below. For each Part (A, B, C), your response should include four sections:

[1] I.e. the first episode of the show ever made.

[2] In fact, all the questions labeled Q1 were "This show made me feel…" questions, whereas all the questions labeled Q2 were "I found this show…" questions.

1) Question: What question are you trying to answer?
2) Approach: What approach/statistical tool did you use to answer the question?
3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.)
4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set.

These questions are fairly simple, so we'd expect each of these four sections for each part to be quite short—surely no more than 1-3 sentences each.

Your confidence intervals for all parts below should be constructed using a built-in R function that reports "large sample" confidence intervals (i.e. based on the Central Limit Theorem). **While you shouldn't include raw R code, make sure to state which R function you used in each "Approach" section.**

Part A. Consider the shows "Living with Ed" and "My Name is Earl." Who makes people happier: Ed or Earl? Construct a filtered data set containing only viewer responses where `Show == "Living with Ed"` or `Show == "My Name is Earl"`. Then construct a 95% confidence interval for the difference in mean viewer response to the `Q1_Happy` question for these two shows. Is there evidence that one show consistently produces a higher mean `Q1_Happy` response among viewers?

Part B. Consider the shows "The Biggest Loser" and "The Apprentice: Los Angeles." Which reality/contest show made people feel more annoyed? Construct a filtered data set containing only viewer responses where `Show == "The Biggest Loser"` or `Show == "The Apprentice: Los Angeles"`. Then construct a 95% confidence interval for the difference in mean viewer response to the `Q1_Annoyed` question for these two shows. Is there evidence that one show consistently produces a higher mean `Q1_Annoyed` response among viewers?

Part C. Consider the show "Dancing with the Stars." This show has a straightforward premise: it is a dancing competition between couples, with each couple consisting of a celebrity paired with a professional dancer. Per Wikipedia: "Each couple performs predetermined dances and competes against the others for judges' points and audience votes."

Despite the simplicity of this format, it seems that some Americans nonetheless find the show befuddling, as evidenced by our survey data on the `Q2_Confusing` question, which asked survey respondents to agree or disagree with the statement "I found this show confusing." Any response of 4 or 5 indicated that the survey participant either Agreed (4) or Strongly Agreed (5) that "Dancing with the Stars" was a confusing show.

Construct a filtered data set containing only viewer responses where `Show == "Dancing with the Stars"`. Based on this sample of respondents, what proportion of American TV watchers would we expect to give a response of 4 or greater to the "Q2_Confusing" question? Form a large-sample 95% confidence interval for this proportion and report your results.

Problem 4: EBay

In this problem, you'll analyze data from an experiment run by EBay in order to assess whether the company's paid advertising on Google's search platform was improving EBay's revenue. (It was certainly improving Google's revenue!) In fiscal year 2020, more than 80% of Google's reported $182 billion in revenue came from its advertising system. Google AdWords has advertisers bid on certain keywords (e.g., "iPhone" or "toddler shoes") in order for their clickable ads to appear at the top of the page in Google's search results. These links are marked as an "ad" by Google, and they're distinct from the so-called "organic" search results that appear lower down the page.

Nobody pays for the organic search results; pages get featured here if Google's algorithms determine that they're among the most relevant pages for a given search query. But if a customer clicks on one of the sponsored "Ad" search results, Google makes money. Suppose, for example, that EBay bids $0.10 on the term "vintage dining table" and wins the bid for that term. If a Google user searches for "vintage dining table" and ends up clicking on the sponsored EBay link from the page of search results, EBay pays Google

$0.10 (the amount of their bid).[3]

For a small company, there's often little choice but to bid on relevant Google search terms; otherwise their search results would be buried. But a big site like EBay doesn't necessarily have to pay in order for their search results to show up prominently on Google. They always have the option of "going organic," i.e. **not** bidding on any search terms and hoping that their links nonetheless are shown high enough up in the organic search results to garner a lot of clicks from Google users. So the question for a business like EBay is, roughly, the following: does the extra traffic brought to our site from paid search results—above and beyond what we'd see if we "went organic"—justify the cost of the ads themselves?

To try to answer this question, EBay ran an experiment in May of 2013. For one month, they turned off paid search in a random subset of 70 of the 210 designated market areas (DMAs) in the United States. A designated market area, according to Wikipedia, is "a region where the population can receive the same or similar television and radio station offerings, and may also include other types of media including newspapers and Internet content." Google allows advertisers to bid on search terms at the DMA level, and it infers the DMA of a visitor on the basis of that visitor's browser cookies and IP address. Examples of DMAs include "New York," "Miami-Ft. Lauderdale," and "Beaumont-Port Arthur." In the experiment, EBay randomly assigned each of the 210 DMAs to one of two groups:

- the *treatment* group, where *advertising on Google AdWords for the whole DMA was paused for a month*, starting on May 22.
- the *control* group, where *advertising on Google AdWords continued* as before.

In `ebay.csv` you have the results of the experiment. The columns in this data set are:

- DMA: the name of the designated market area, e.g. New York
- rank: the rank of that DMA by population
- tv_homes: the number of homes in that DMA with a television, as measured by the market research firm Nielsen (who defined the DMAs in the first place)
- adwords_pause: a 0/1 indicator, where 1 means that DMA was in the treatment group, and 0 means that DMA was in the control group.
- rev_before: EBay's revenue in dollars from that DMA in the 30 days before May 22, before the experiment started.
- rev_after: EBay's revenue in dollars from that DMA in the 30 days beginning on May 22, after the experiment started.

The outcome variable of interest is the **revenue ratio** at the DMA level, i.e. the ratio of revenue after to revenue before for each DMA. (You'll need to use `mutate` to define this outcome yourself.) If EBay's paid search advertising on Google was driving extra revenue, we would expect this revenue ratio to be systematically lower in the treatment-group DMAs versus the control-group DMAs. On the other hand, if paid search advertising were a waste of money, then we'd expect the revenue ratio to be basically equal in the control and treatment groups. Two explanatory notes here:

- We use the ratio rather than the absolute difference because the DMAs differ enormously in population and therefore revenue.
- We wouldn't necessarily expect the before-and-after revenue ratio to be 1 (i.e. similar revenue before and after the experiment), even in the control-group DMAs. That's because, like any retailer, EBay's sales exhibit a lot of seasonal patterns and might be lower in some months across the board, regardless of paid search. That's why the important question isn't whether the revenue is the same before and after in the treatment-group DMAs, but whether the before-and-after **ratio** is the same for the treatment group as for the control group.

Your task is to compute the difference in revenue ratio between the treatment and control DMAs and provide a 95% confidence interval for the difference. Use these results to assess the evidence for whether the revenue ratio is the same in the treatment and control groups, or whether instead the data favors the idea that paid

[3]There's huge variability in the market price of different search terms. The market price per click for a search term like "insurance" or "attorney" or "MBA programs" might be $50 or more. Google makes a *fortune* on these popular search terms. For stuff you might buy on EBay, the market price is usually a lot less.

search advertising on Google creates extra revenue for EBay. Make sure you use at least 10,000 Monte Carlo simulations in any bootstrap simulations.

Your write-up for this problem should include four sections:

1) Question: What question are you trying to answer?
2) Approach: What approach/statistical tool did you use to answer the question?
3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.)
4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set.

It is certainly possibly in this case for each of these four sections to be only 1-3 sentences long, although you can take longer if you feel you need it. Make a sensible judgment about how much to round.