DS102 - Project Part 1

Data science is a collaborative activity. While you may talk with others about the project, we ask that you write your solutions individually. If you do discuss the project with others please include their name in your submission.
You must submit a typed up pdf on Gradescope for the project. We will not accept handwritten submissions. You can optionally use the provided LaTeX template found at https://github.com/ds-102/fa19/tree/master/project/project01.tex to type up your project.
Due by: 1:59pm, Tuesday 3rd December, 2019

This is part 1 (of 2) of the project. The setting is the following:

Setup You've been hired by a stealth-mode start up company to analyze data related to the potential market-share and competing companies. The company can't tell you yet exactly what they do, only that it has to do with scooters (the type in Figure 1).

Your job as a freelance data analyst is to look at this problem from several angles, and write up a report detailing your methods and your findings. This document details the questions the company would like answers to, as well as some questions you might investigate on your own to enhance the report.

Instructions We've provided an outline, including a set of prompts and subsections for you to fill out. Please **do not alter the section numbers**. If you wish to copy the outline into a different text editor (google docs, word), that's fine, but please keep the section and subsection structure consistent.

Some of the prompts are mandatory, and some are optional open-ended questions (marked with a star, *). We expect you to answer all questions without stars, select a single starred questions depending on what interests you most.

Figure 1: A scooter. Also an example template for inputting figures in latex.

Remember that this is an independent project. While you can talk to others, your analysis (especially for open ended questions) should be from your own ideas. Please submit your filled-in report as a pdf on gradescope. You must also email any code you use in a zip file to karlk@berkeley.edu with files clearly named to match each section, and subject line "DS102 - project01".

Grading The major components of your grade are:

- **Content**: in each part of the assignment there are structured questions for you to fill out, with point assignments given. There are also open ended prompts (denoted with *'s) that you can explore depending on which parts of the project interest you most. A full and correct set of responses for the non-starred problems will receive 80% credit; the remaining 20% will be assigned for open ended responses.
- **Completeness**: For all parts of the report, carefully and completely describe what you did. As a rule of thumb, a reader who has taken DS100 but not DS102 should be able to reproduce your analysis without referring to any of your code.
- **Creativity**: For the open ended problems, we expect to see a careful or creative integration of topics from class. This creativity takes effort, and you have the choice of which starred problems to put forth this effort on. Document why you're approaching questions with certain techniques, and if things don't work out as expected, that's ok!
- **Professionalism and Readability**: We expect full sentences with correct grammar and spelling. All axes in all figures should be labeled, with units when applicable.

Setup

You've been tasked with analyzing the public datasets released by three bike sharing companies in New York, Washington DC, and Chicago. Another data scientist has helpfully preprocessed the dataset for you in four files that can be found here: https://github.com/ds-102/fa19/tree/master/project/bikeshare.zip.

The ny.csv, chicago.csv, and dc.csv each contain the bike rentals that occurred in 2016 in their respective cities with each row representing a single rental.

The day.csv dataset contains daily bike rental information between the years 2011 and 2012 in Washington DC. The original bike rental dataset has also been merged with weather information and holiday information using other data sources.

If you're curious you can find the original datasets at the following links

- New York: https://www.citibikenyc.com/system-data
- Washington DC: https://www.capitalbikeshare.com/system-data
- Chicago: https://www.divvybikes.com/system-data
- Daily Summary: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

1 Preliminary Data Analysis

You decide to start by performing some exploratory analysis to see if you can find any interesting trends.

1.1 Demographic Information

In this section you should:

1. Plot the distribution of male to female riders for chicago.csv.
2. Plot the distribution of the gender column for ny.csv.
3. Given the results in Chicago, make an educated guess as to the mapping from numerical value to Male/Female/Unspecified within the ny.csv dataset.
4. Plot the distribution of the birth years of bike renters in Chicago and NY.
5. Discuss the results you observed in the age plots and whether this fits with your intuition. Would you remove any data? If so, why?

1.2 Rental Times

In this section you should:

1. Plot the three distribution of trip duration in minutes across all three cities.
2. Are the plots you generated useful? If not, plot them again so that the visualization is more useful.
3. Plot the start time of trips split by hour for all three cities.
4. Discuss the results you observed in the start time plots. Do they fit your intuition?

1.3 Further Exploration

In this section you should:

1. Visualize three more attributes.
2. Discuss any insight you obtained from these three new visualizations.
3. Pick one of the three attribute and plot its distribution if you haven't already. Explore the data further to get a plausible explanation for the shape of the distribution. For example you could explore whether the attribute is correlated with another attribute.
4. Given the insights you obtained from the data, write down a hypothesis that you think is important and how you would go about testing it.

1.4 Optional*: Creating a new dataset

You want to investigate the change in rental behavior from 2011 to 2016 and also the change in rental behavior between cities. To do this you want to transform ny.csv, chicago.csv, and dc.csv into three new datasets that resemble day.csv.

Read the attribute information section found at https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset. Now try to reproduce a few of the columns described there while using ny.csv, chicago.csv, and dc.csv as your data source. Submit the newly created files as ny_daily.csv, chicago_daily.csv, and dc_daily.csv.

If you're feeling particularly ambitious you can try to reproduce the columns describing weather but you'll have to find an external dataset that has daily weather information for each city. This might be time consuming and isn't compulsory.

Describe the dataset creation procedure you followed. Write about any interesting conclusions you made by comparing the three new datasets with day.csv, support your conclusion with graphs and statistics.

2 Hypothesis Testing

Suppose that, for privacy reasons, the company has to change its policy and make the indicator of whether each bike ride corresponds to a casual (customer) or non-casual (subscriber) user confidential. In the future the company will no longer have access to this feature. However, this feature is particularly useful, because depending on the type of user the company might want to give them different promotions and advertisement. Therefore, in this part we will design a test for whether the user is casual or not. We will also try to prevent too many false discoveries, i.e. falsely declaring a casual user as a non-casual one.

You will be given general guidelines for the problem below. However, there is no single correct solution. Feel free to use any additional Python packages, as well as any hyperparameters you might want to add. You can also reuse any code from previous homework/lab solutions.

Pick your favorite city among the three we have data from (New York, Washington DC or Chicago). In the rest of the problem, you will only use the data set corresponding to that city. Let n denote the total number of samples. For each $i \in \{1, \dots, n\}$, denote by Y_i the indicator whether the user is casual ($Y_i = 0$) or non-casual ($Y_i = 1$). For $i \in \{1, \dots, n\}$, let X_i denote the feature vector consisting of the following 3 features: trip duration, ride start time and ride stop time. As the start and stop time, simply take the *hour* of the corresponding time. For example, if the ride starts at 13:05, take 13 as the start time. Similar for stop time.

Split the data set into three uniformly-without-replacement randomly sampled subsets: S_1 (60% of the whole data set), S_2 (20% of the whole data set) and S_3 (20% of the whole data set). Recall that we need the model of the data under the null in order to compute p-values. We will use S_1 and S_2 to learn this model, and then compute p-values on S_3.

1. Consider the logistic regression model

$$\mathbb{P}(Y_i = 1 | X_i) = \frac{1}{1 + e^{-\theta^\top X_i}},$$

for some $\theta = (\theta_1, \theta_2, \theta_3)$. Use S_1 to train a logistic regression model for predicting Y_i from X_i according to the above model. Denote by θ_* the learned value of θ. Consider using methods available in `sklearn`.

2. For each sample (X_i, Y_i) in S_2, compute $s_i^{(2)} = \frac{1}{1+e^{-\theta_*^\top X_i}}$. For each (X_j, Y_j) in S_3, compute $s_j^{(3)}$ analogously. Denote by $S_{2,0}$ the subset of S_2 that consists of casual users ($Y_i = 0$). For each sample j in S_3, compute a p-value as

$$P_j = \frac{1}{|S_{2,0}|} |\{i \in S_{2,0} : s_i^{(2)} > s_j^{(3)}\}|,$$

where as usual we use $|\cdot|$ to denote the cardinality of a set. Plot two histograms: one of null p-values ($Y_i = 0$, casual riders) and one of non-null p-values ($Y_i = 1$, non-casual riders). What do you observe?

3. Now you should have $|S_3|$ p-values. Run the Benjamini-Hochberg algorithm under level 0.2 on these p-values, and compute the false discovery proportion (FDP) and sensitivity. Repeat the whole procedure from above for 200 different randomly sampled S_1, S_2 and S_3 (you don't have to repeat the visualizations). You should have 200 sets of $|S_3|$ p-values, and get 200 different false discovery proportions and sensitivities. Report the average FDP and sensitivity over these 200 trials. Is the average FDP above or below 0.2? Can you explain why it is or isn't?

2.1 Optional*: Improving the Average Sensitivity

Try to improve the average sensitivity of the BH procedure by making the distribution of non-null p-values "smaller". In other words, you want to make the histogram of non-null p-values more skewed "to the left". Report the new average sensitivity, compare it to the old average sensitivity, and describe in detail how you achieved this improvement. One suggestion for this would be to pick different features, or transform them in some way (i.e. you might want to add some features on top of the three from above, or, for example, instead of trip duration you can look at f(trip duration), for some function f). You can also use a completely different way to generate p-values, but you need to make sure that they are valid, meaning that they are approximately uniform under the null and that BH controls FDR under 0.2. If you choose to use a different way to generate p-values, you have to include a plot of null p-values and non-null p-values.

3 Gaussian Mixture Models of Trip Durations

In the previous part you performed Hypothesis Testing to distinguish between Casual and Registered users of the bike sharing service. Moving forwards, you would like to understand the distribution of trip durations so that you can better choose the scooter your startup will use. You assume that the main factor influencing the length of the trip is whether or not a customer is a subscriber.

1. Describe why it is reasonable to believe that should be a difference in the distributions of trip durations of subscribers and non-subscribed customers. Use figures from the data to back up your argument.

2. Use the Expectation-Maximization (E-M) Algorithm to learn a mixture of two Gaussians that describes the distribution trip-durations of length **less than one hour** in the `chicago.csv` dataset. Run this multiple times from different initializations. Do your results change drastically depending on the initialization? What are the means and variances of the fitted Gaussians?

3. Given the output of the E-M Algorithm, which of the distributions captures the behavior of the subscribed customers? For each customer, in the dataset, calculate the

posterior probability that the customer is from this distribution.

4. If you design a classifier which classifies a customer as a Subscriber if their posterior probability is greater that 0.5, what is the error of this classifier given the true User Types in the `chicago.csv` dataset?

5. Use the classifier derived from the `chicago.csv` dataset on the `ny.csv` and `dc.csv` datasets. How does this classifier perform? How does the performance compare to that in the previous part? (Make sure that the data for each city is in comparable units.)

3.1 Optional*: Learning Mixtures of Bivariate Gaussians

In the previous section you learned a mixture of univariate that described trip durations in the `chicago.csv` dataset. In this optional part you will learn a mixture of bivariate Gaussians using the start-times of the data.

1. Describe why it is reasonable to believe that the start-times of customers will help you distinguish between subscribers and non-subscribed customers. Use figures from the data to back up your argument.

2. Use the Expectation-Maximization (E-M) Algorithm to learn a mixture of two bivariate Gaussians that describes the distribution trip-durations of length **less than one hour** and start-times in the `chicago.csv` dataset. Make sure that the start-times are in the form of minutes-past-midnight.

   Run this multiple times from different initializations. Do your results change drastically depending on the initialization? What are the means and covariances of the fitted Gaussians? What do these numbers imply?

3. As before, design a classifier that classifies a customer as a Subscriber if their posterior probability of being from one of the Gaussians is greater that 0.5. How does this classifier perform compared to the one which used only the trip-durations? Why is the performance better or worse?

4. Experiment designing classifiers based off of mixtures of different numbers of bivariate Gaussians. How well do these classifiers generalize to the other cities in the data? Why do you think the performance is consistent or inconsistent depending on the city?

4 Causality and Experiment Design.

4.1 2SLS to estimate the effect of precipitation on #bike rentals.

You would like to know more about what causes a higher or lower number of rentals on a given day. More specifically, you might want to determine the effect that weather conditions have on the number of bike rentals.

Using the daily UCI data, run a two stage least squares regression to calculate the effect of adverse weather (`weathersit` > 1) vs nice weather (`weathersit` $= 1$) on the total number of rentals.

Since we are given an observed data set, we will use instrumental variables analysis and 2 stage least squares regression. In particular, temperature is a potential confounding variable for weather and the number of bike rentals. To overcome this, use humidity as an instrumental variable.

4.1.1 The causal model.

In this section you should:

1. Draw the graph of the causal model, with arrows between variables to denote causality. Your graph should include the following variables: temperature, weathersit, humidity, and number of rentals.

2. Clearly describe the assumptions necessary for 2SLS analysis. For any assumptions you can check with the data set, do so. For the remainder of the assumptions, give your best guess as to whether they hold given the problem setting, and discuss how you would test them (this could mean collecting more data).

4.1.2 2 Stage Least Squares

In this section you should:

1. Describe the 2SLS procedure, as it applies to the variables above. Describe the procedure at a level such that someone who has taken DS100, but not DS102, would be able to reproduce the analysis (without looking at your code).

2. Report the resulting treatment affect of `weathersit` on the total number of bike rentals. Interpret this treatment effect estimate in terms of the variables of the problem.

3. Report the resulting treatment affect of weathersit on the number of casual bike rentals, and on the number of registered bike rentals.

4.1.3 Discussion

In this section you should discuss the following:

1. Give a three sentence summary of the question, how and why we tested it, and the results of the analysis.

2. Interpret the treatment effect estimates that resulted for the number of casual rentals and for the number of registered rentals. Is the magnitude of the treatment effect higher for one group than another? Give at least one possible reason for the difference/similarity you find.

3. Discuss the applicability of the chosen model (causal graph from above) to this problem. Are there any variables that might missing from the causal graph? Are there any arrows that are missing?

4. If you had the opportunity to design your own study (and collect new data) to test the effect of adverse weather on the number of rentals, what would that study look like? Explain at a high level the design decisions you would make and why (keep your response to no more than 5 sentences).

4.2 Optional*: Simulating Experiment Design Approaches via Sub-sampling.

Define a hypothesis that you'd like to test given the data at hand. Then, imagine this data set were not available to you, and you needed to collect new data to answer your hypothesis. In particular, you need to design and carry out an experiment. Use what we've learned about experiment design to formulate your sampling strategy. Then, using any of the data sets given, sample instances from these data sets to simulate running your experiment. (Note: this means you will have to pick a hypothesis and a sampling design that can be simulated by taking draws from the data given).

As an example, you could imagine testing whether there are more bike rentals on weekends or weekdays, on average. Your experiment design might randomly or deterministically pick certain days to observe the number of riders on that day. For every day of observation you must pay someone to count the number of rentals, thus each day of observations costs you data. You might define two strategies: spend half of your money sampling weekdays, and spend half of your money sampling weekends. Or perhaps you think weekend rental counts are more variable, so you spend more samples observing weekend days. To simulate the results of these methods, you would take subsets of the days from the *observed* data set which correspond to the strategy of your design. This sub-sampling of the data would need to simulate the sampling strategy you chose.

Such a simulation study could compare different strategies for picking these days, and the resulting inferences made. A good simulation study will also give a notion of variability due to any randomness in the design. In the discussion of your experimental simulation methodology, you may want to address sources of randomness, nuisance variables, or unobserved/uncontrollable variables that you sampling methodology does or does not account for.

This is an open-ended problem, so the choice of hypothesis and resulting experiment design are up to you. As a starting point, the following are good things to address if you choose to do this part of the project:

- Define the hypothesis you would like to test.
- Define the experimental design you would use to test this hypothesis.
- Describe a data sub-sampling experiment to simulate the effect of your experiment design. Potentially also describe how you evaluate success of the experiment.
- Carry out the sub-sampling experiment and report your findings. Give a brief discussion of the implications of your findings.
- Discuss the pros and cons of a sub-sampling simulation experiment to test your experimental design.

You should not feel limited to addressing only these things, nor should you feel obligated to do all of them if you focus on some parts more than others. The intent of this section is to use sub-sampling to evaluate your design, so it would be wise to pick a hypothesis and experiment that are amenable to this.