CS24200: Project 1

Due date: Sunday October 27, 11:59pm

Note: You can use *Python* and/or *R* to complete this assignment. Submit a PDF with both the code that you used for analysis and your answers to the questions below to *Gradescope*. Your homework must be typed. *Jupyter Notebook* is highly recommended to use for this project. Moreover, you can directly export a PDF file from the Notebook (with your code and answers) as the final submission. If you choose to use other ways to generate the file, please make sure the corresponding code attached right behind your answers for each question. [1]

Use the supplied Hospital Charge dataset downloaded from Data.CMS.gov:
`https://data.cms.gov/Medicare-Inpatient/Inpatient-Prospective-Payment-System-IPPS-Provider/tcsp-6e99`
This dataset includes hundreds of hospital-specific charges from more than 3000 hospitals in the US that receive Medicare Inpatient Prospective Payment System (IPPS) payments in Fiscal Year 2017. All the payments are under Medicare based on a rate per discharge using the Medicare Severity Diagnosis Related Group (MS-DRG). This dataset has been updated on Augest 28, 2019. For more details, please refer to CMS.gov.

In this assignment, you will parse, explore, transform and analyze this dataset. Based on your analysis, you will formulate and test hypotheses about the data. There are 12 columns and 196K rows in the data. Each row is a hospital record. You will focus your analysis on the following columns:

- 1: `DRG Definition`
- 2: `Provider Id`
- 6: `Provider State`
- 9: `Total Discharges`
- 10: `Average Covered Charges`
- 11: `Average Total Payments`
- 12: `Average Medicare Payments`

[1] Last revised: Oct. 14, 2019; Posted: Oct. 15, 2019

1 Data Exploration (16 pts)

1.1 Overview (8 pts, 2pts each)

Read in the data and try to answer the following questions:

(a) How many `Provider Id` in the dataset are different? And which state has the most number of unique providers out all the other states?

(b) What is the value of mean, median, and standard deviation for `Hospital Discharges` in FY 2017?

(c) How many `DRG Definitions` from this dataset are unique? And which one was the coded the most in FY 2017?

(d) Which category of `DRG` has the least number of hospital discharges? And how many `Total Discharge` are there for this `DRG` category?

1.2 Distributions and Outliers (8 pts, 2pts each)

Plot histograms/densities or scatter plots for each of the following. Make sure to label the axis of the plots appropriately. And explain why you choose certain types of graph to visualize those data.

(a) `Total Discharges`

(b) `Average Covered Charges`

(c) `Average Total Payments` vs. `Average Medicare Payments`

(d) `Average Covered Charges` vs. `Average Medicare Payments`

In each plot, identify at least one outlier and discuss whether the outlier(s) are surprising or expected given the location of the `Provider`.

2 Data Processing (10 pts)

2.1 Feature Creation (4 pts)

Recall **1.1(c)**, for each unique value of `DRG Definition`, we can create a corresponding feature '`DRG Charge`', which records the `Average Covered Charges` for each specified `DRG` category regrading the `Provider`. In the following, design a method [2] you may use to create 100 `DRG Charges` features through the top 100 most frequently billed discharges of `DRG` in the dataset.

2.2 Transforming Data (4 pts)

Construct a transformed version of the data in **2.1** that only includes the `Provider Id`, `Provider State`, and the 100 new `DRG Charges` features. For example, the data should look like the format in the table below.

Prov.Id	Prov.State	DRG Charges 039	DRG Charges 057	DRG Charges 064	...
10001	AL	41130.56	25434.17	46240.00	...
10005	AL	14450.08	NaN	26866.23	...

2.3 Quality Control (2 pts)

Identify potential issues from the data that you generated in **2.2**:

- Make sure your dataset include missing values for any provider that doesn't have a charge for a specific `DRG` category. List common ways to handle those missing items.
- Check whether there are any duplicated rows and/or columns in your dataset.

[2] Note: You may provide either pseudo code or brief description or both of them in your answer.

3 Data Analysis & Interpretation (34 pts)

3.1 Correlation and Scatterplots (12 pts)

On the new transformed version of the data in **Section 2**, explore the relationships among those 100 `DRG Charges` features. Identify two pairs of `DRG Charges` features with high positive associations and two pairs with low positive associations. For each of the four pairs of features:

(a) **Plot scatterplots (8 pts)**
Plot a scatterplot to show their relationship. Make sure to label both axis of the plot with the feature names. Discuss whether the observed relations are interesting or expected, given the `DRG` category names. (This will result in 4 scatter plots total.)

(b) **Compute correlations (4 pts)**
Calculate the correlation among the selected features and report them. Discuss whether the correlations support your observations from the scatterplot above.

3.2 Boxplots and T-tests (22 pts)

On the new transformed version of the data, explore how the `DRG Charge` features vary with `Provider State`.

(a) **Boxplots (6 pts)**

- Select six states that you think may exhibit differences in their hospital charges (consider e.g., geographic, size, population, political differences). Find a `DRG Charge` feature that shows some variation across the six selected states. Plot a box plot to show the variation (i.e., the six `Provider States` vs. the selected `DRG Charge`). Make sure to label both axes of the plot with the appropriate attribute names/values.
- Select two other `DRG Charge` features to repeat **3.2(a)**. Make sure to use the same six selected states. (This will result in 3 box plots total.)

(b) **Formulate and test claim: Part I (6 pts)**

- Based on the three box plots, identify the pair of states that you think have the most significant differences in their charges for a *single* `DRG` category. Explicitly state your hypothesis in terms of H_0 and H_1.
- Perform a two-sample Student's t-test to assess your hypotheses. State whether you are performing a one-sided or two-sided test. Report the resulting t statistic and p-value. Discuss whether the results support your claim(s).

(c) **Formulate and test claim: Part II (10 pts)**

- Based on the three box plots, identify a different pair of states that you think have a significant difference in their charges *across all three* selected `DRG` categories. Explicitly state your hypothesis in terms of H_0 and H_1.

- Perform a two-sample *paired* Student's t-test [3] to assess your hypotheses. Report the resulting t statistic and p-value. Discuss whether the results support your claim(s).
- Repeat the test as above, but use an *unpaired* t-test this time. Report the differences and discuss what (if any) impact there is on your assessment of significance.

[3] Hint: To do a paired t-test, you will need to concatenate the values from the three selected `DRG` categories into a single vector, one for each state. If the samples are different sizes (e.g., each state has a different number of providers), just randomly downsample from the state with more providers to reduce it to the same size as the state with fewer providers.