CS24200: Project 3

Due date: Saturday Apr 25, 11:59pm AoE (Anywhere on Earth)

Submit a Jupyter notebook with both the code that you used for analysis and your answers to the questions below to Blackboard.

Use the supplied Hospital Charge dataset downloaded from data.gov: `https://data.cms.gov/Medicare/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3` This dataset includes hospital-specific charges for the top 100 most frequently billed discharges, for more than 3000 hospitals in the US that receive Medicare Inpatient Prospective Payment System (IPPS) payments.

In this assignment, you will parse, transform, explore and analyze this dataset. Based on your analysis you will formulate and test hypotheses about the data. There are 12 columns and 163K rows in the data (see description on the website). 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 Distributions and Outliers (10 pts)

Read in the data and plot histograms/densities or scatter plots for each of the following. Make sure to label the axis of the plots appropriately.

(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 Transforming Data (10 pts)

There are 100 values of `DRG Definition`. Construct 100 `DRG Charges` features, one for each unique value of `DRG Definition`. The feature should record the `Average Covered Charges` for the specified `DRG` category. Then construct a transformed version of the data 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. Make sure to include missing values for any provider that doesn't have a charge for a specific DRG.

Prov.Id	Prov.State	DRG Charges 039	DRG Charges 057	DRG Charges 064	...
10001	AL	32963.07	20312.78	38820.39	...
10011	AL	13998.28	22074.08	24204.84	...

3 Correlation and Scatterplots (10 pts)

On the new transformed version of the data, explore the relationships among the 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.** 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.** Calculate the correlation among the selected features and report. Discuss whether the correlations support your observations from the scatterplot above.

4 Boxplots and T-tests (20 pts)

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

(a) **Boxplots.**

- Select six states that you think may exhibit differences in their hospital charges (consider e.g., geographic, size, 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(a). Make sure to use the same six selected states. (This will result in 3 box plots total.)

(b) **Formulate and test claim (1).**

- 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 (2).**

- 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 to assess your hypotheses. To do this 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. 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.