**\* [ ] Describe the core message or hypothesis for your project.**

GENERALLY - We wanted to try and understand the forces that caused the variance in medical procedure costs for two specific procedures. Diabetes and Joint Replacement. These 2 procedures were chosen because 1.) they are relatively common and 2.) they are not immediately life threatening and therefore a patient may be able to research in advance which hospitals would be more

affordable. Around 2013 hospitals standardized all producers by DRG codes specifically for Medicare purposes and were required to make their costs more transparent to the public ( this is roughly 2-3 years after Obamacare ACA was passed.

From a 2013 Washington Post article , "

In downtown New York City, two hospitals 63 blocks apart varied by 321 percent in the prices they charged to treat complicated cases of asthma or bronchitis. One charged an average of $34,310; the other billed, on average, $8,159.

The highest average charge for a lower joint replacement was $36,000 by University of Maryland Medical Center in Baltimore, much lower than the highest rates in other states.  
Elsewhere, Las Colinas Medical Center just outside Dallas billed Medicare, on average, $160,832 for lower joint replacements.

we were able to utilize very recent data from Medicare.gov and CMS.gov for nation wide average hospital charge by 100 top procedures as well as survey results per hospital combined with census info by zip code (income, population ).

From the get go the Medicare data was rather complete and clean. So we performed rough heatmap to determine using standard deviation which states were charging more than the national average. Based on those result we evaluated NY State, NY City and California State, as well as the Nation as a whole.

**OUR HYPOTHESIS**

**Attempted Hypos :**

1. **Hypothesis**: There is a significant difference between ( customer satisfaction ) and (procedure charge)

**Null**: There is no significant difference between overall customer satisfaction and procedure charge

**B.) Hypothesis:** There is a significant difference between overall [ hospital rating ] and [ procedure charge ]

**Null:** There is no significant difference between [overall hospital rating] and [ procedure charge]

**C.) Hypothesis:** There is a significant different between [ population density ] and [ procedure charge].

**Null:** There is no significant difference between population density and medical procedure cost.

**D.) Hypothesis:** There is a significant different between [ HOUSEHOLD INCOME PER HOSP AREA ] and [ procedure charge].

**Null:** There is no significant difference between[ HOUSEHOLD INCOME PER HOSP AREA ] and [procedure charge]

**Unattempted Hypos :**

**E.)-Hypothesis:** There is a significant [procedure cost] vs [private or government medical centers].

**Null:** There is not a significant cost difference knee and diabetes procedures across private government

medical centers.

**F.)-Hypothesis:** There is a significant difference between ( income) and (customer satisfaction)

**\* [ ] Describe the questions you and your group found interesting, and what motivated you to answer them**

we wanted to know if their was any difference between east coast and west coast charges and mid west. Also want to evaluate national if hospital ratings affect amounts charged. Also we want to compare rural to suburban, to urban areas and medium house income high, mid, low income. Also wanted to understand why there was such as difference in cost in Manhattan. In Manhattan it appeared there was correlation between a hosp overall rating and the avg procedure charge. Meaning the higher the rating the higher the cost and conversely lower rating = lower cost.

**\* [ ] Summarize where and how you found the data you used to answer these questions**

we obtained most of our data from data.gov , Medicare.gov included in a large zip file. They also provided the ability to create visualization maps on their website (we did not use) We obtained census data from a pip install library for python / jupyter. We also pulled in some additional survey data from Socrata linked api

**\* [ ] Describe the data exploration and cleanup process (accompanied by your Jupyter Notebook)**

Medicare Data National, State, City:

- Many Merges occurred with CSV's saved at each point

1.) from the Medicare national file we extracted out Joint Replacement and Diabetes created separate CVSs

2.) then using those as the base file, we did a left merge with survey data always using the ' Hospital Provider Id ' as the merge on key value

3.) later Ron converted survey results into Numerics so they could be statistically evaluated

4.) Tyler used python install library to extract nation census data per zip code, some challenges were experienced taking census list and combining to Dataframe/csv

5.) then Brandon took Tyler's census data and added categories/ buckets for High,Mid,Low income, and Urban, Suburban, Rural, categories numeric and text this was useful if we wanted to drill down further or color our scatters by income or city /region type

6.) all of that was left merged into our base national, and columns reorganized for ease of view

7.) next 3 states CA, NY, MS extract from each for 6 csv's, 3 for joint, 3 for diab.

8.) clean performed to drop rows without census or containing text

9.) from their ttest , regression, and chart analysis performed

**\* [ ] Describe the analysis process (accompanied by your Jupyter Notebook)**

1.) first we produced a variety of charts to understand our data, sorting by hospital charge amount, high to low, then looking at top 15 charges. We evaluated those top 15 charting, hospital rating, patients serviced, hospital charge and Medicare reimbursement amounts, also showing averages per city, state, and national looking for relationships.

2.) Then we were taught about ttest, p-value, chi squares, and regressions , so we switched gears

3.) we generated 4-5 hypothesis to test looking for significant p-values

4.) Ron produced a nice ttest and scatter chart combo,that was sort of automated, which brandon took further and eventually setup a Batch process to run all 7 ttests and chart analysis with polynomial regression lines 1 and 4 dimensional, for each Medical procedure file and to save out unique file names for chart names, and final csv for ttest regression analysis

5.) batching proved very useful especially when Ron suggested recharting all 4 dimensional Polyfit charts with 1 dimensional charts for a clearer trend line

6.) ALSO figuring out how to format and display the true p-value with 30 decimals on our charts helped immensely

**\* [ ] Summarize your conclusions. This should include a numerical summary (i.e., what data did your analysis yield), as well as visualizations of that summary (plots of the final analysis data)**

HYPOTHESIS RESULTS

**\* [ ] Discuss the implications of your findings. This is where you get to have an open-ended discussion about what your findings "mean".**

HYPOTHESIS CONCLUSIONS

**\* [ ] Tell a good story! Storytelling through data analysis is no different than in literature. Find your narrative and use your analysis and visualization skills to highlight conflict and resolution in your data.**

# Technical Requirements