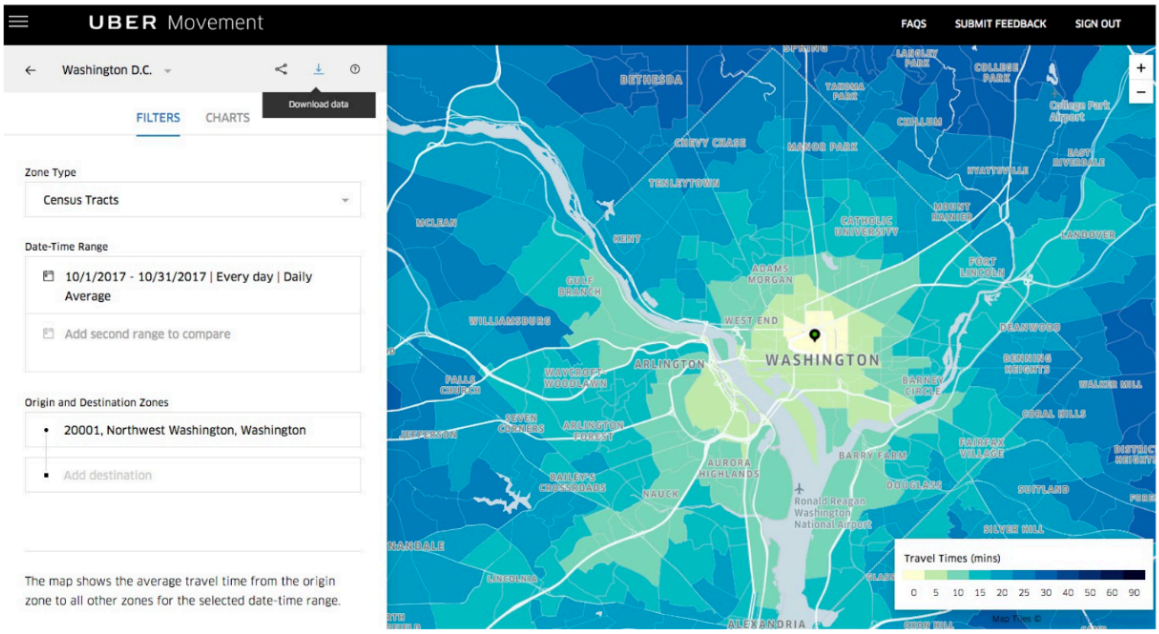
UML
Laurence (w4rner) Warner

Problem Set 1
Exploration & Computation

1. Obtain a dataset (preferably of substantive interest/domain expertise).

   Uber Movement is a service providing anonymized data to the public on over 2 billion trips to help urban planners and the public better understand traffic patterns.

   The following shows an example of the data that exists. It is semi-aggregated: it shows the average travel time between different city zones.

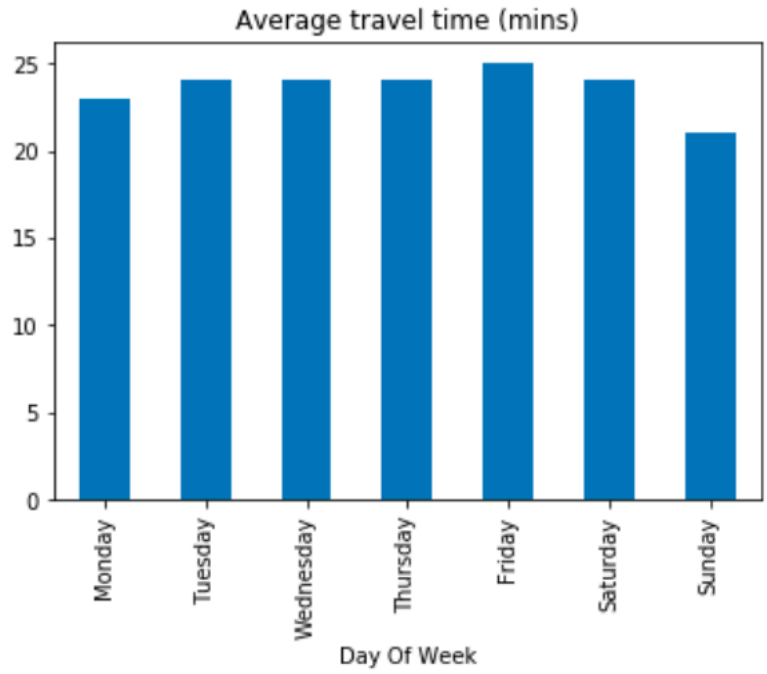


Let's look at the example of the links between two Census Tracts in this dataset and explore that: in San Francisco: 28th & Mission to 1400 Powell St.

2. Choose a visual technique to illustrate your data (e.g., barplot, histogram, scatterplot).

Barplot

3. Now generate and present the visualization and describe what you see.



We see that daily average travel time is pretty consistent, but lower on Sunday.

4. Calculate the common measures of central tendency and variation, and then display your results.

```
mean      23.571429
std        1.272418
min       21.000000
25%       23.500000
50%       24.000000
75%       24.000000
max       25.000000
```

5. Describe the numeric output in substantive terms, e.g., a. What do these numeric descriptions of data reveal?

   Mean travel time is 23.5m, and the median is 24m.

   b. Why is this important?

   It is useful to know the central tendency of the data specifically.

c. What might you infer about the distribution or spread of the data?

Standard deviation is low, which reveals that there is little variation.

d. Etc.

We can infer that day of week is not a major determinant of Uber travel time.

UML

Laurence (w4rner) Warner

P Set 1

Critical Thinking

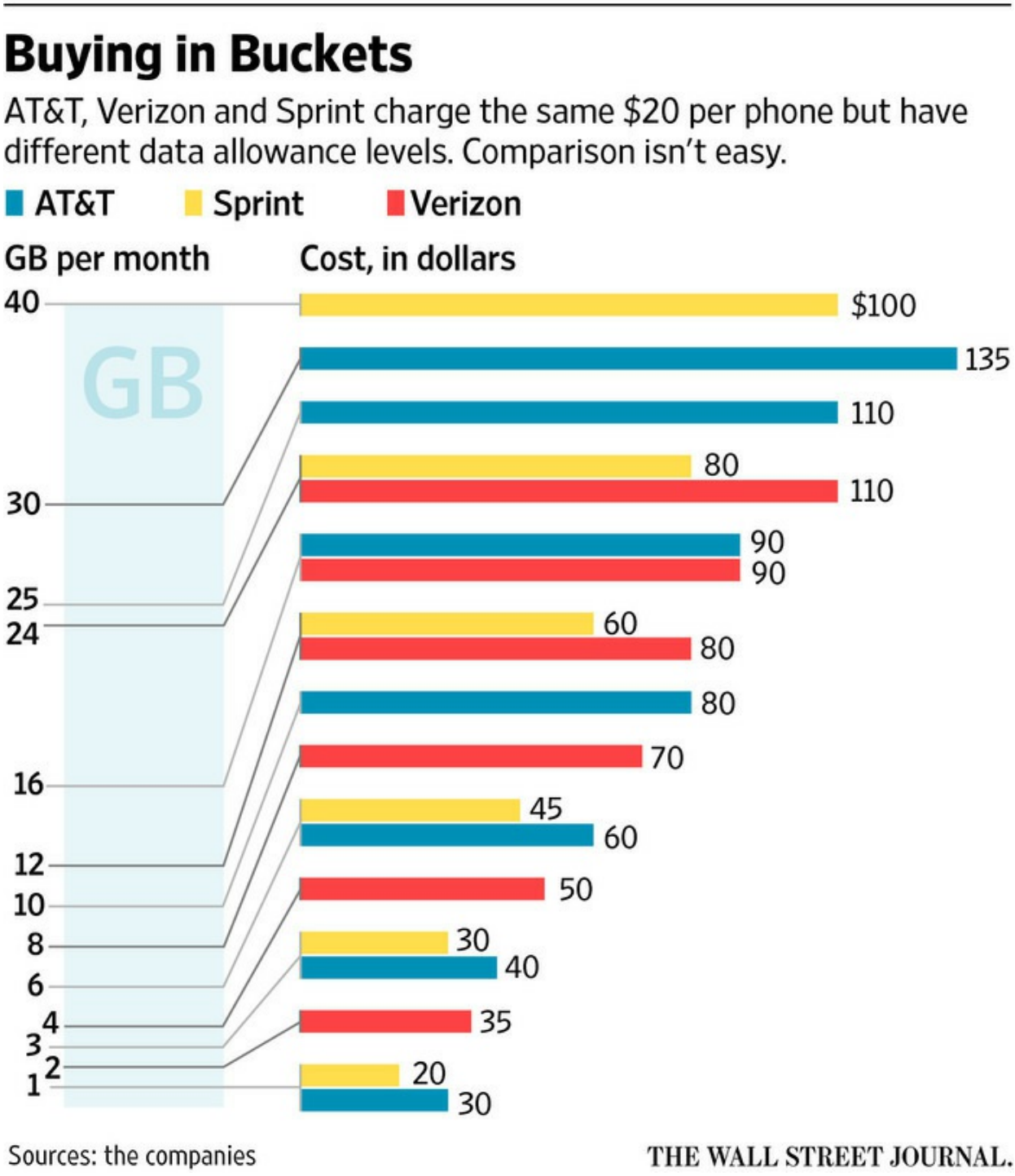
1. Describe the different information contained in/revealed by visual versus numeric exploratory data analysis. (Hint: Think of different examples of each and then what we might be looking for when leveraging a given technique).

Visual better for:
Anomalies, clustering

Numeric better for:
Specificity.

2. Find (and include) two examples of “bad” visualizations and tell me precisely why they’re bad.
3. Find (and include) two examples of “good” visualizations and tell me precisely why they’re good.

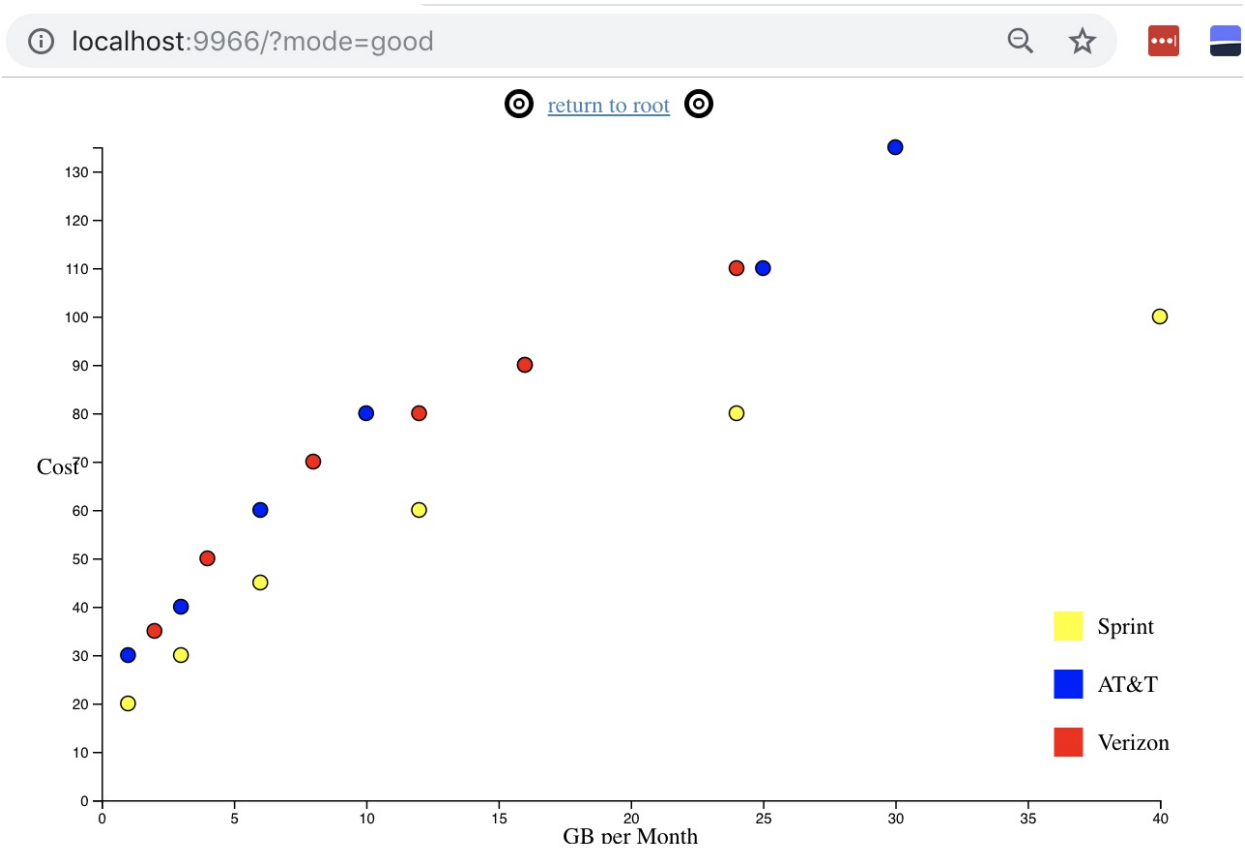
First Bad:
A WSJ viz, syndicated by viz.wtf



Crucially, it doesn't allow for easy calculation of the ratio of Cost to GB per month.

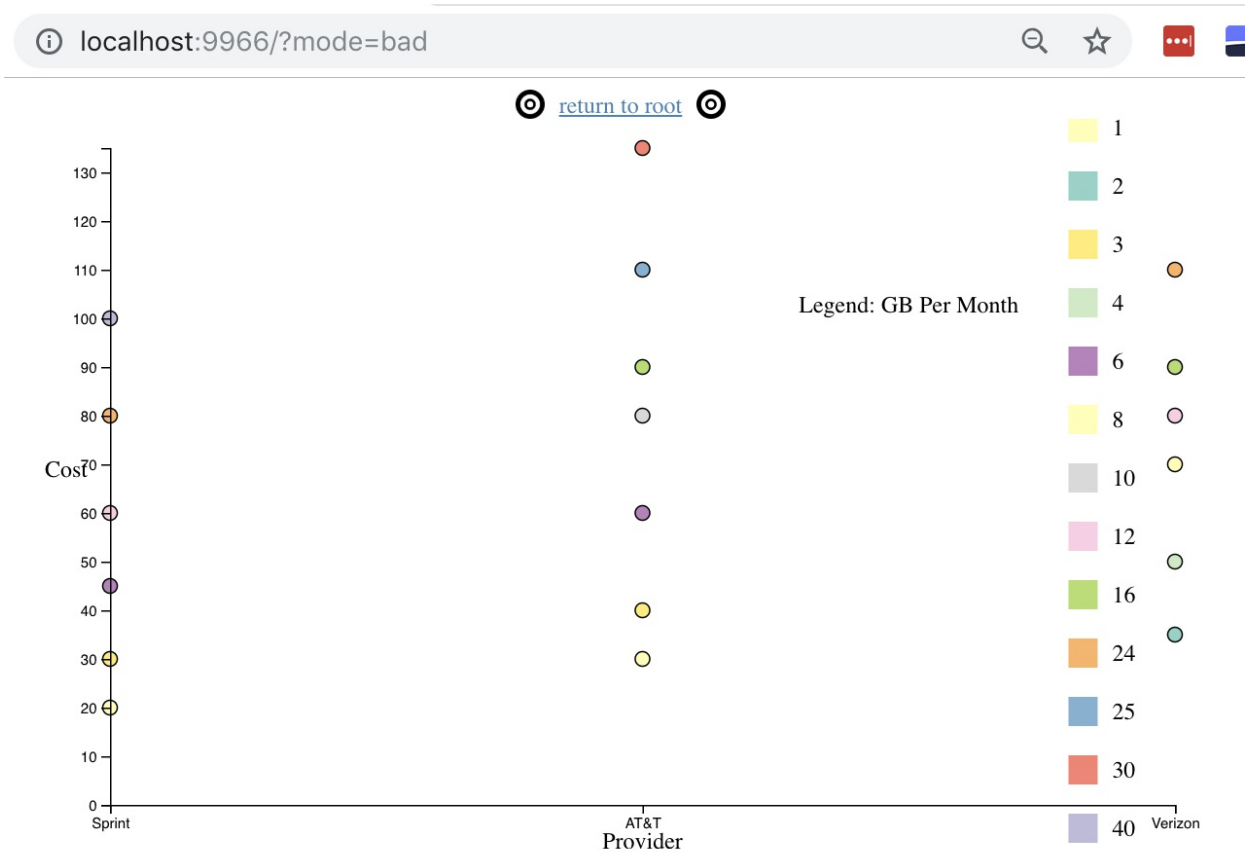
Inspired by this, I made two original visualizations.
I used the same single dataset about Mobile Phone Contracts with the following three variables: Mobile Provider, GB of data, Cost per month.
I challenged myself to create a good & a bad viz.

Good:



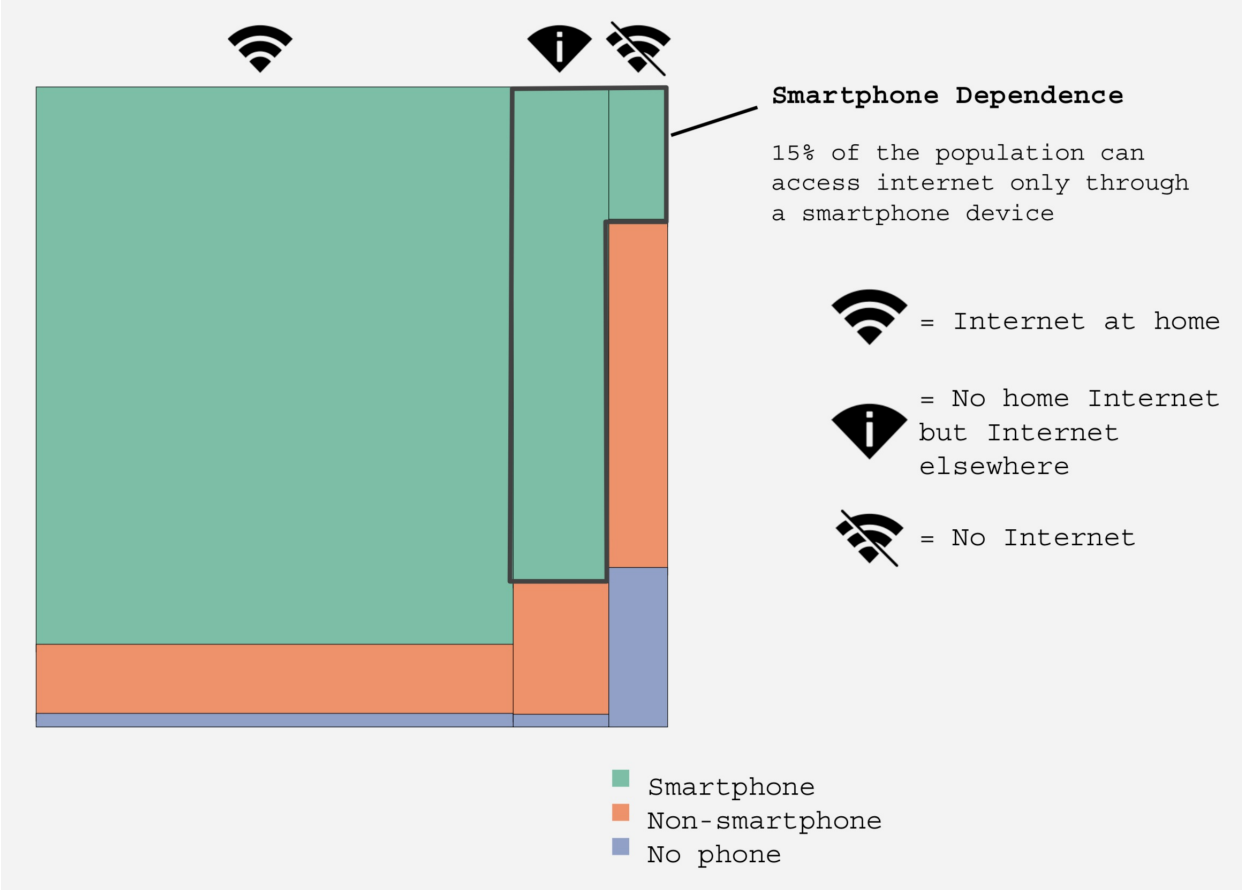
The common scale of the axes iood for judgements and comparisons of ratio quantities. The Color encoding for provider is good because there are only 3 categories.

Bad:



Hard to learn the size because the I use an array of 12 categorical colors to encode GB per Month, meaning we don't make use of the fact it is in an ordered variable. Really hard to compare that many different colors categorically.
Also, makes it difficult to make ratio judgements because distance between two colors not clear.

Another good one I've made for my thesis using D3.js:



It is a type of treemap called a mosaic plot, used to show a population divided into groups by two variables. It then identifies a specific population using an annotation.
Improvements:
Move keys onto the axis in which they are used for easier understanding.

4. When might we use EDA and why/how does it help the research process?

When we don't need specific concrete answers, but instead want to know what questions to ask.

5. What did John Tukey mean by “confirmatory” versus “exploratory”? Give me an example for each.

"Exploratory is an attitude, a flexibliity and a reliance on display". It is about finding the right question rather than right answer.
Example: automate scatter plots between all pairs of variables and look at them.

Confirmatory is when you are trying to confirm or disprove a stated hypothesis.
Example: the Logistic Regression analysis in my thesis.