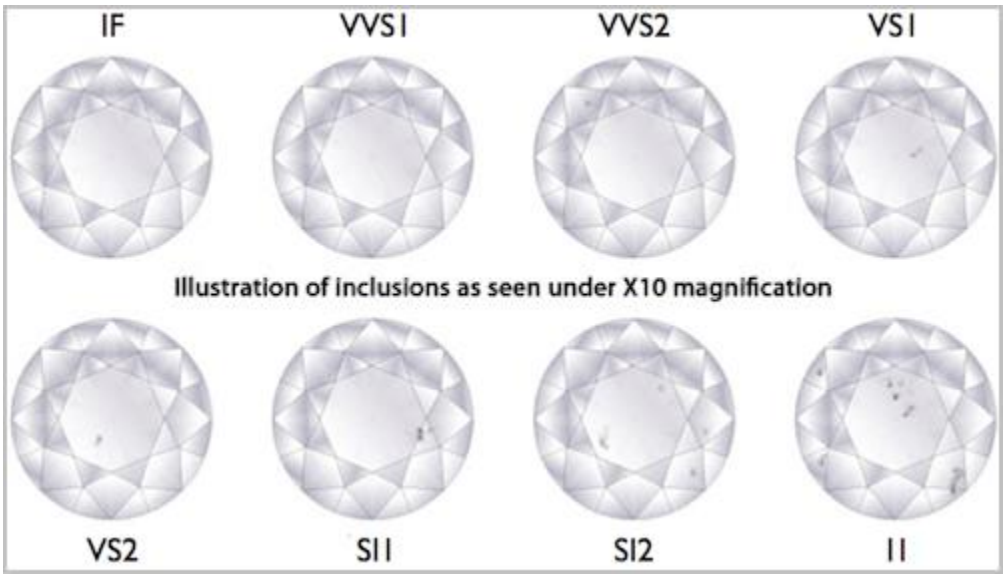
MACS 40800 – UML **Sharon Shen (12445609)**

Problem Set #1: Exploratory Data Analysis

Exploration & Computation

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

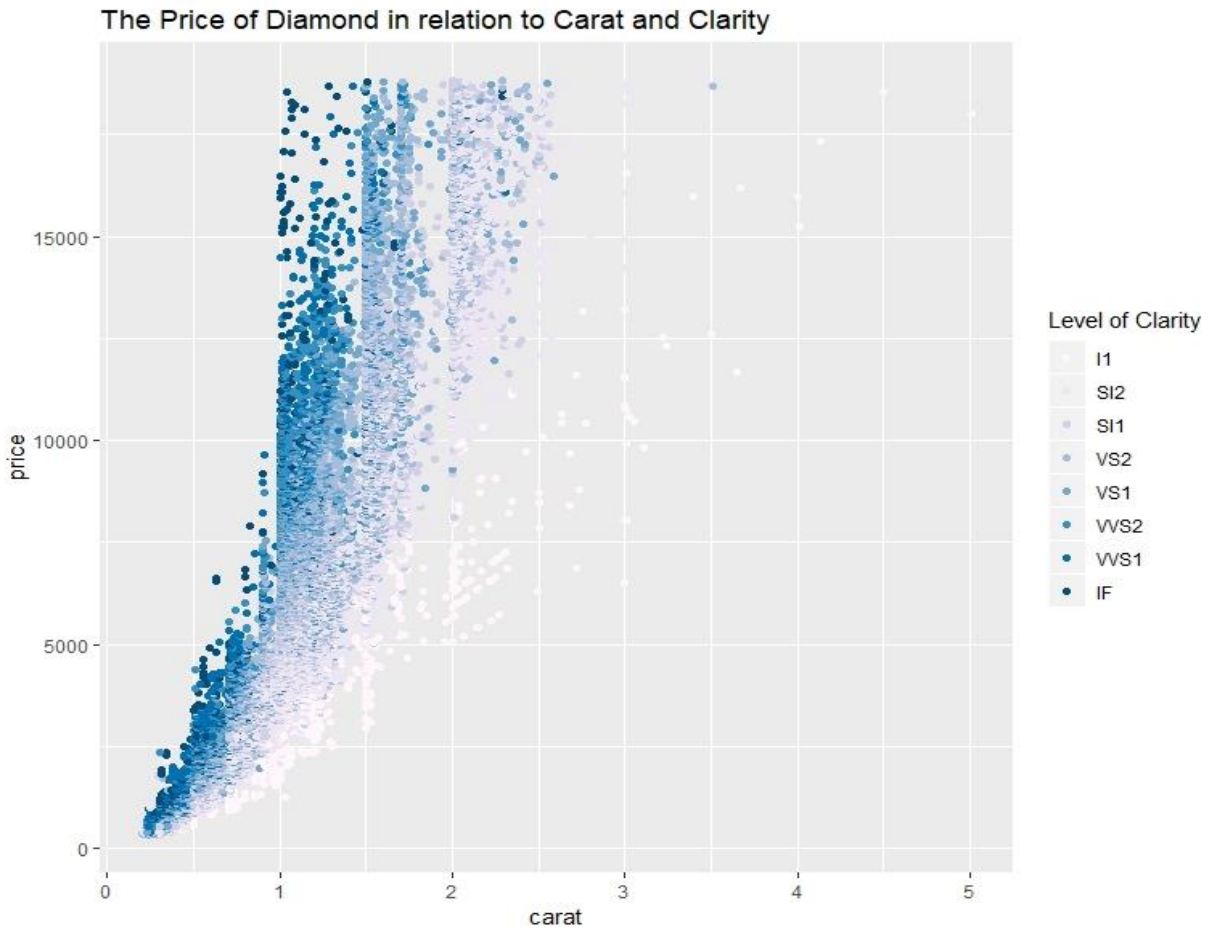


For this exercise, I utilized the dataset containing information on the price of diamonds, based on different attributed associated with this particular consumer good. Variables include the 4 Cs, i.e. carat, cut, color and clarity, as well as physical measurements such as the table width and depth of diamonds. The above diagram provides an illustration of one of the attributes, clarity, where IF represents the highest standard, i.e. the purest, and II the lowest.

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

I will be using a scatterplot to explore the price distribution of the observations with respect to some of the key attributes.

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



This scatterplot shows the price distribution of diamonds in relation to their carat and clarity level. On the x-axis, we have carat level from 0 to 5; each dot represents one observation, and they are color-coded such that observations with higher clarity would be closer to blue in this blue-to-white gradience. From the scatterplot, we can see that as the diamond gets larger in size, i.e. carat level is higher, the price tends to be higher; similarly, diamonds in pure forms receive higher price tags. Interestingly, data is mostly concentrated within the range between 1 carat and 2 carats, and it is harder to have higher clarity diamonds once it reaches a threshold carat level. In this case, a rough estimate from the graph would be around 3 carats.

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

For the variable carat, the mean is 0.7979, and the median is 0.7, with a standard deviation of 0.47. The maximum carat observed is 5, while the lowest is 0.2.

5. Describe the numeric output in substantive terms, e.g.,

a. What do these numeric descriptions of data reveal?

Based on the descriptive statistics, we can understand the average attribute level of diamonds, and how they vary from each other.

b. Why is this important?

It can help us compare different types of diamonds and be informative on one particular observation in terms of where it stands among the group.

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

For example, by comparing the mean and median of the carat variable, we can conclude that the distribution of the diamonds is slightly skewed to the right since the mean is larger than median, which means there is a long tail of high carat diamonds.

******R-code Attached Below******

```
# Load packages for visualization
library(tidyverse)
library(skimr)

# Use a preloaded dataset called "Diamonds" in R and find summary statistics
summary(diamonds)

# Find descriptive statistics through pastecs packages
install.packages("pastecs")
library(pastecs)
stat.desc(diamonds$carat)

# Use boxplot to visualize the price of diamond based on clarity
ggplot(data = diamonds) +
  geom_boxplot(aes(x = clarity, y = price)) +
  scale_color_discrete(guide=F) +
  labs(x = "Level of Clarity",
       y = "Price",
       title = "Price of Diamond by Clarity") +
```

```
  theme_classic()

# Install a package for more colorful visualization
install.packages("RColorBrewer")
library(RColorBrewer)

# Use scatterplot to show price in relation to carat and clarity level
ggplot(diamonds, aes(x=carat, y=price)) +
  geom_point(aes(color=clarity)) +
  labs(title = "The Price of Diamond in relation to Carat and Clarity") +
  scale_color_brewer(name="Level of Clarity", palette="PuBu")
```

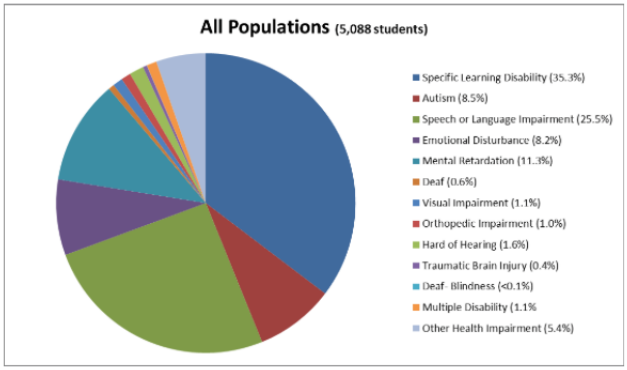
******End of R-Code******

Critical Thinking

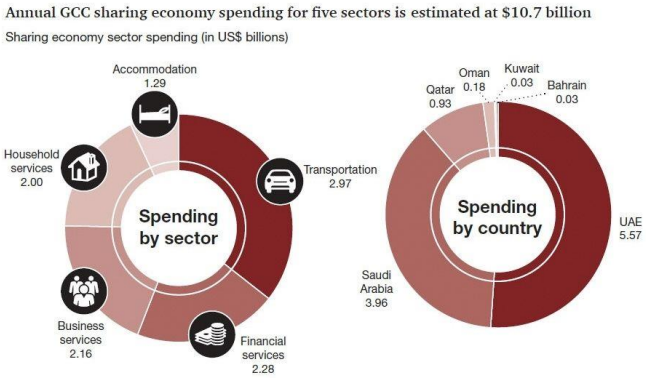
1. Describe the different information contained in/revealed by visual versus numeric exploratory data analysis.

Through visual tools, we are able to discover hidden patterns and underlying structure in the dataset. It displays data in a more presentable manner by adding graphics and color spectrums to specify the difference or relationship around the data points. For numeric exploratory data analysis, we look for central tendency, for example, mean of the dataset, and the amount of variability among the observations of a dataset. We are also presented with evidence on the data's distribution with the notions of quantiles.

2. Find (and include) two examples of "bad" visualizations and tell me precisely why they're bad.

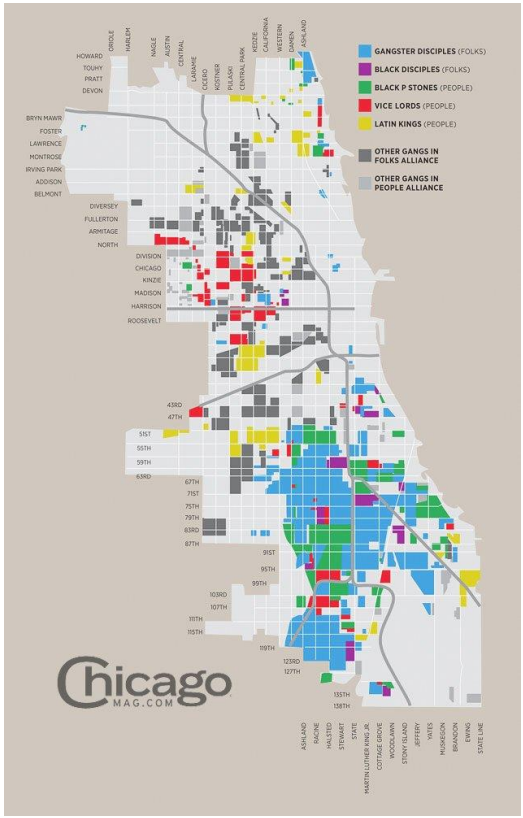


This is a bad graph because it includes too many categories, the sizes of which are hard to compare and identify for human eyes, given the choice of pie chart.



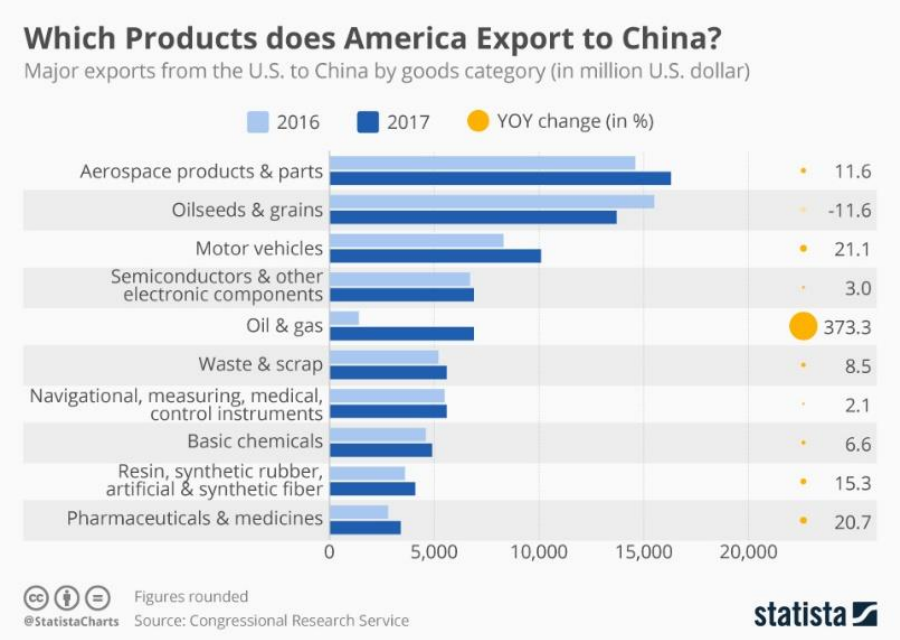
This chart is confusing because while it tries to present two set of findings side-by-side, the sizes and colors of the ring are the same and therefore audience are misled to think the total spending of countries combined is equal to the total spending of sectors.

3. Find (and include) two examples of "good" visualizations and tell me precisely why they're good.



This is a neatly constructed graph because it assigns each gang with its distinct color, and by color-coding the map, audience are able to identify instantly whereabout in Chicago are there gang scenes, and if so, which gang is presented at that neighborhood.

The diagram below is also an illustration of what makes a good graph, because it ranks products based on export level from high to low, and shows the yearly change by presenting 2016 and 2017 bars right next to each other.



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

We can use EDA when we are trying to identity patterns out of a given dataset, or under conditions that we are not sure where to start an analysis of the relationship between variables. EDA can be very helpful in research when there is lack of identifiable independent and dependent variables; and can be used to check if our assumptions are appropriate by providing inference to the distribution of data.

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

Confirmatory data analysis is when we start our research with a hypothesis and use data to try and test this specific hypothesis, a process which we have a set of independent variables and dependent variables, i.e. "confirming" an association. For example, researchers who look at the effective of sex education program would hypothesize that abstinence-only program can yield greater reduction in teenage pregnancy rate as compared to comprehensive sex education, and by showing statistically significant difference in change in rate between the two categories, one can confirm or reject null hypothesis. On the other hand, exploratory data analysis does not require a pre-set hypothesis and is not used to produce estimates or predict outcome. In exploratory process, we are trying to identify what are some key variables worth looking at. In other words, we don't have specific expectation for certain outcome, and we can find some unknown patterns through the analysis. For example, conducting web scraping on a corpus of wall street journal article titles, hedge funds are able to identify previously unspotted but potentially profitable companies in some hot industries worth investing in.