Unsupervised Machine Learning HW1

Yaoxi Shi

10 October 2019

Exploration & Computation

1.Dataset: Alcohol consumption and Happiness Scores by Countries

This dataset includes variables such as Country, Region, Hemisphere, Happiness Scores, Beer per Capital, Wine per Capital, GDP per Capital etc, collected from 122 countries in 2018. In this report, I am interested in visulizing how happiness is related to beer consumption among the countries across the world.

The below are some sample rows of the subdataset I am interested in:

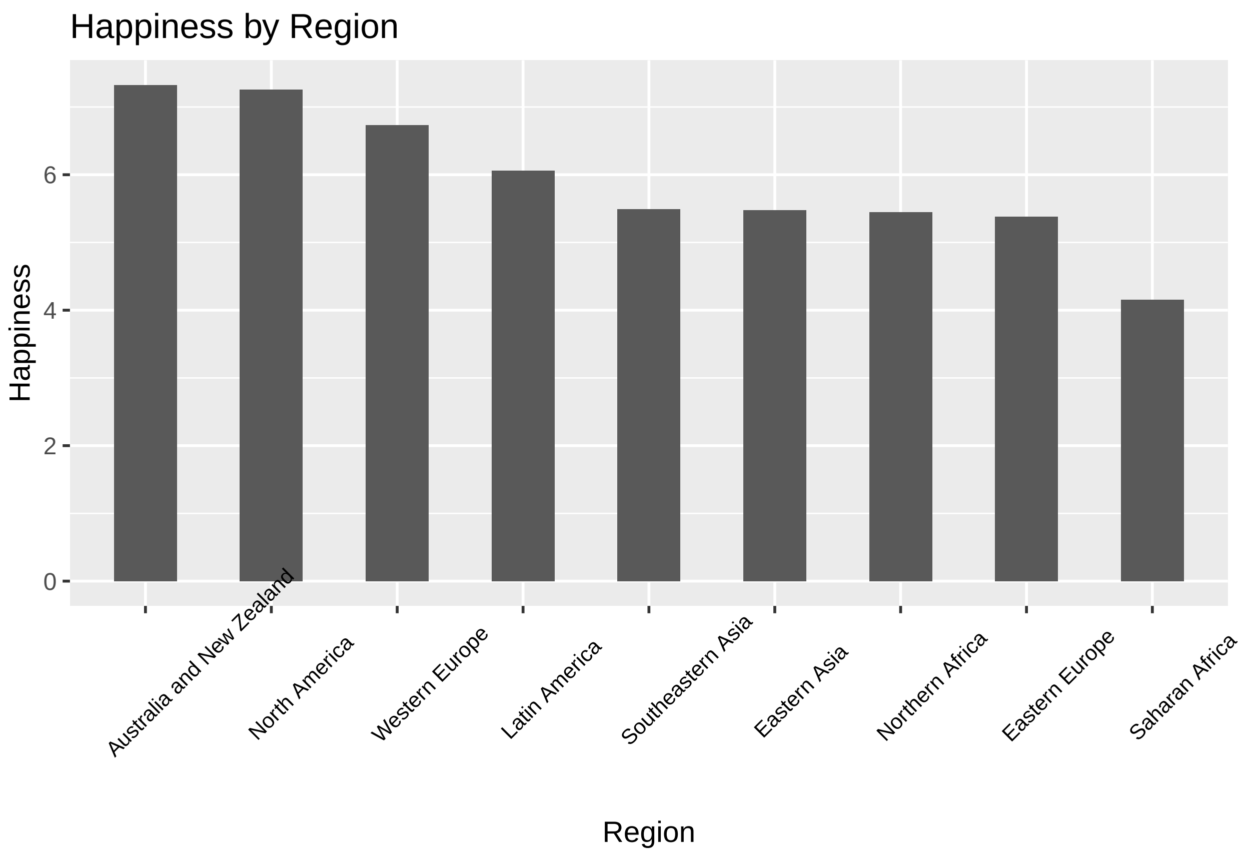
```
dataset <- data %>%
  dplyr::select(Country, Region, Hemisphere, HappinessScore, Beer_PerCapita)
sample_n(dataset, 5, replace = TRUE)
```

```
## # A tibble: 5 x 5
##   Country Region         Hemisphere HappinessScore Beer_PerCapita
##   <chr>   <chr>          <chr>               <dbl>          <dbl>
## 1 Serbia  Eastern Europe north                5.18            283
## 2 Chile   Latin America  south                6.70            130
## 3 Ukraine Eastern Europe north                4.32            206
## 4 Ireland Western Europe north                6.91            313
## 5 Benin   Saharan Africa north                3.48             34
```

2.Visual Technique: Bar plot and scatter plot

3.1 Bar plot showing happiness scores of each region of the world.

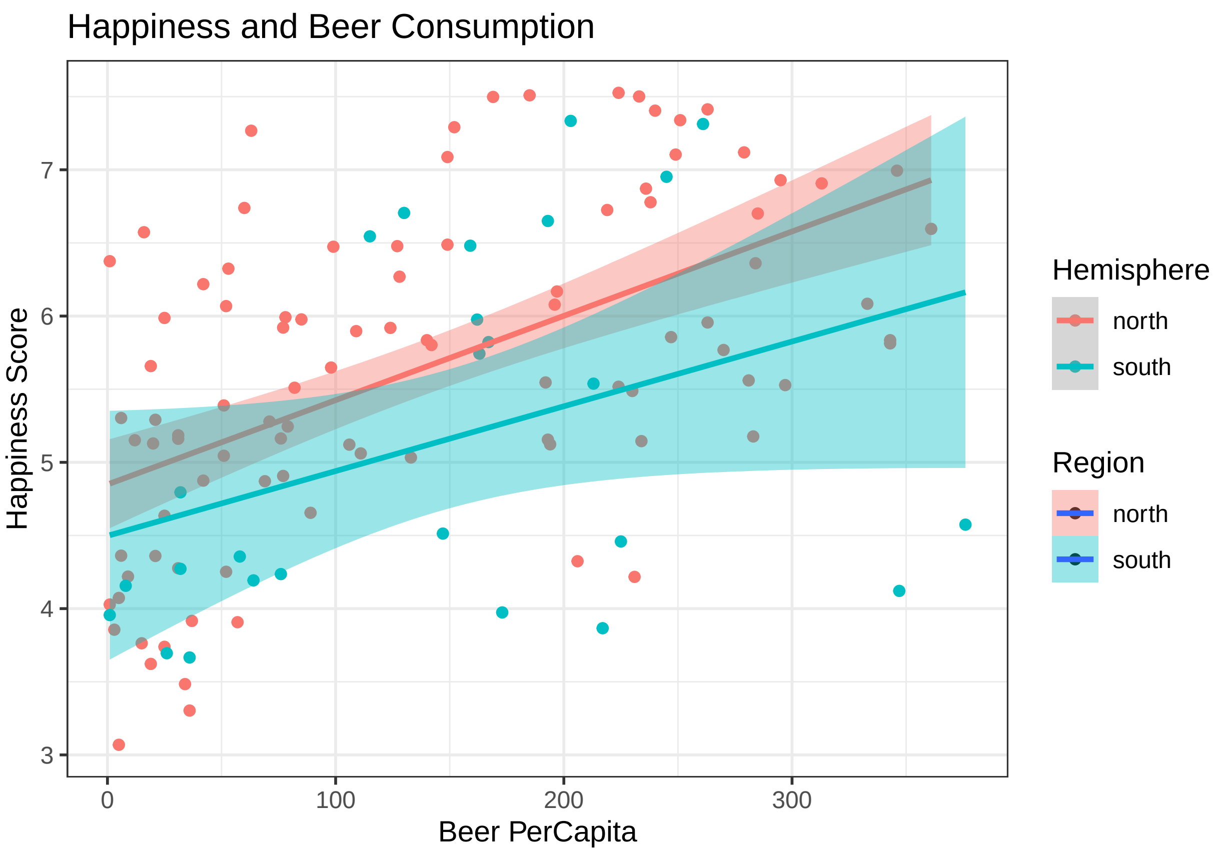
```
# Bar plot: Happiness by Region
dataset %>%
  group_by(Region) %>%
  summarize(mean_happiness = mean(HappinessScore, na.rm = TRUE)) %>%
  ggplot() +
  geom_bar(aes(x = reorder(Region, -mean_happiness), y =mean_happiness),
           stat = "identity", width = 0.5) +
  labs(x = "Region",
       y = "Happiness",
       title = "Happiness by Region") +
  theme(axis.text.x = element_text(size = 8,  color = "black",
                                   vjust = 0.5, hjust = 0.2, angle = 45))
```



From the bar plot, we could find that people living in Australia, New Zealand, North America, Western Europe have the highest happiness scores, while peopel living in Northern Africa, Eastern Europe, and Saharan Africa feel least happy. Also, the countries in the region that have relatively higher happiness scores are developed contries, and on the other side, countries with lower happiness scores are mostly the developing countries.

3.2 Scatter plot showing the relationship between beer consumption and happiness scores, by hemisphere.

```
# Scatter plot: Beer Consumption and Happiness Score
ggplot(dataset, aes(x = Beer_PerCapita, y = HappinessScore,
                    fill = Hemisphere, color = Hemisphere)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(x = "Beer PerCapita",
       y = "Happiness Score",
       title = "Happiness and Beer Consumption", fill = "Region") +
  theme_bw()
```



From the figure, we can see that happiness are generally positively related with beer consumption, more beer consumption per capital, the happiness score of the country is higher. Also, the happiness scores of the countries in north hemisphere is higher than the countries in the south, but the trend of beer consumtion on happiness is similar.

4.Central tendency and variation of the data

```
## Skim summary statistics
##  n obs: 122
##  n variables: 5
##  group variables: Hemisphere
##
## -- Variable type:character ------------------------------------------------
##  Hemisphere variable missing complete  n min max empty n_unique
##       north  Country       0       96 96   4  22     0       96
##       north   Region       0       96 96  12  17     0        8
##       south  Country       0       26 26   4  15     0       26
##       south   Region       0       26 26  13  25     0        3
##
## -- Variable type:numeric --------------------------------------------------
##  Hemisphere       variable missing complete  n   mean     sd   p0   p25
##       north Beer_PerCapita       0       96 96 134.94 106.31 1    36.75
##       north HappinessScore       0       96 96   5.63   1.11 3.07  5.04
##       south Beer_PerCapita       0       26 26 147.27 100.51 1    59.5
##       south HappinessScore       0       26 26   5.15   1.23 3.67  4.17
##     p50    p75    p100     hist
##  107.5  231.5  361    <U+2587><U+2585><U+2582><U+2582><U+2583><U+2583><U+2582><U+2582>
##    5.65   6.47   7.53 <U+2582><U+2583><U+2583><U+2587><U+2586><U+2587><U+2585><U+2586>
##  160.5  210.5  376    <U+2587><U+2583><U+2582><U+2587><U+2587><U+2582><U+2581><U+2582>
##    4.54   6.35   7.33 <U+2586><U+2587><U+2581><U+2581><U+2583><U+2581><U+2583><U+2583>
```

5.Describe the numeric output in substantive terms

The data reveals that the average happiness score of the north hemisphere (5.63) is a bit higher than south (5.15), and the countries in the north have smaller variations. However, the beer consumption per capital in the north (134.94) is lower than in the south (147.27), and north hemisphere has larger variations.

This summary data gives us overall picture of the happiness scores and the beer consumption in north and south hemisphere, it's easy to make the comparison among the groups. Also, according to the distributions of the data, I found that the happiness scores of the countries in the north hemisphere is close the normal distribution, but for the countries in the south hemisphere, the happiness scores are more divergent, some countries have really low scores while other countries have very high happiness scores (countries like Australia). This distribution reveals that the gap between the levels of happiness of the countries in the south hemisphere is much larger than north.

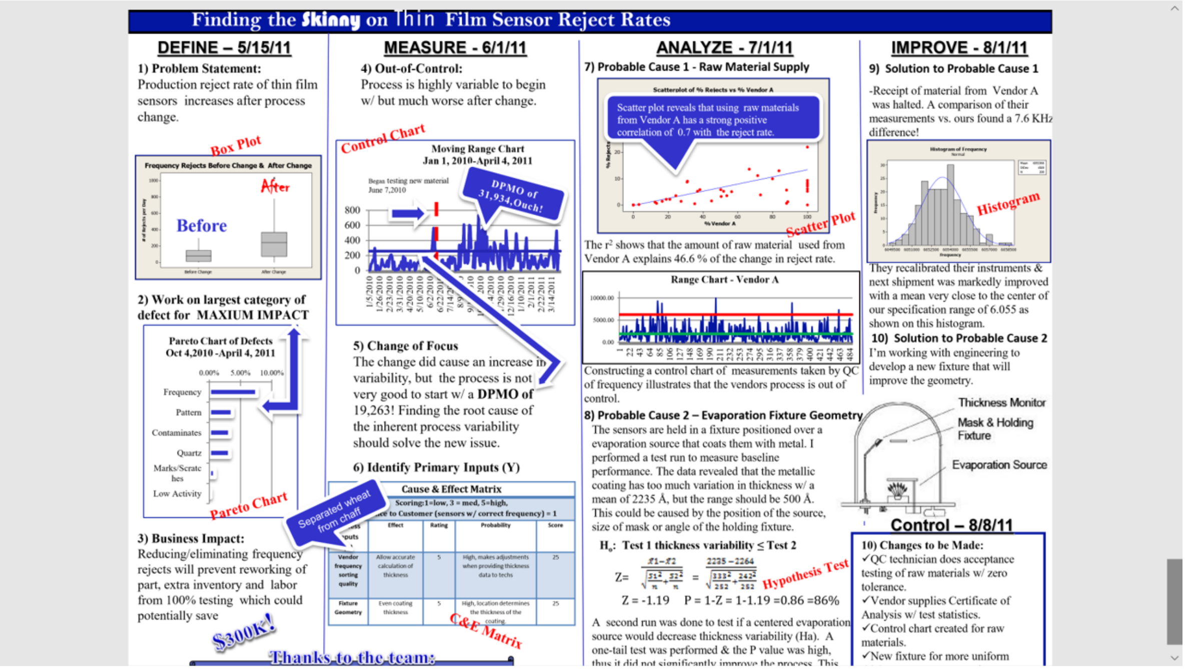
Critical Thinking

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

Visual exploratory data analysis provides us with more direct sense of the data, easier for us to detect patterns, structure and trends. Also human brain are usually more sensitive to the visual patterns, it make the complicated dataset easierand faster for the human brain to process and understand. Numeric exploratory data analysis is more rational, it provides a more specific result, and also can reveal more accurate internal relationships with the variables, which maybe hard to observe from the visual patterns.

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

a)



Finding the Skinny on Thin Film Sensor Reject Rates

DEFINE – 5/15/11

1) Problem Statement:
Production reject rate of thin film sensors increases after process change.

2) Work on largest category of defect for MAXIUM IMPACT

3) Business Impact:
Reducing/eliminating frequency rejects will prevent reworking of part, extra inventory and labor from 100% testing which could potentially save

$300K!

Thanks to the team:

MEASURE - 6/1/11

4) Out-of-Control:
Process is highly variable to begin w/ but much worse after change.

5) Change of Focus
The change did cause an increase in variability, but the process is not very good to start w/ a **DPMO of** 19,263! Finding the root cause of the inherent process variability should solve the new issue.

6) Identify Primary Inputs (Y)

Cause & Effect Matrix

Scoring:1=low, 3 = med, 5=high,
...ce to Customer (sensors w/ correct frequency) = 1

...ess Inputs	Effect	Rating	Probability	Score
Vendor frequency sorting quality	Allow accurate calculation of thickness	5	High, makes adjustments when providing thickness data to techs	25
Fixture Geometry	Even coating thickness	5	High, location determines the thickness of the coating	25

ANALYZE - 7/1/11

7) Probable Cause 1 - Raw Material Supply

Scatter plot reveals that using raw materials from Vendor A has a strong positive correlation of 0.7 with the reject rate.

The r^2 shows that the amount of raw material used from Vendor A explains 46.6 % of the change in reject rate.

Constructing a control chart of measurements taken by QC of frequency illustrates that the vendors process is out of control.

8) Probable Cause 2 – Evaporation Fixture Geometry
The sensors are held in a fixture positioned over a evaporation source that coats them with metal. I performed a test run to measure baseline performance. The data revealed that the metallic coating has too much variation in thickness w/ a mean of 2235 Å, but the range should be 500 Å. This could be caused by the position of the source, size of mask or angle of the holding fixture.

H_o: Test 1 thickness variability ≤ Test 2

$$Z = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\frac{s1^2}{n} + \frac{s2^2}{n}}} = \frac{2235 - 2264}{\sqrt{\frac{333^2}{252} + \frac{242^2}{252}}}$$

Z = -1.19   P = 1-Z = 1-1.19 =0.86 =86%

A second run was done to test if a centered evaporation source would decrease thickness variability (Ha). A one-tail test was performed & the P value was high, thus it did not significantly improve the process. This

IMPROVE - 8/1/11

9) Solution to Probable Cause 1

-Receipt of material from Vendor A was halted. A comparison of their measurements vs. ours found a 7.6 KHz difference!

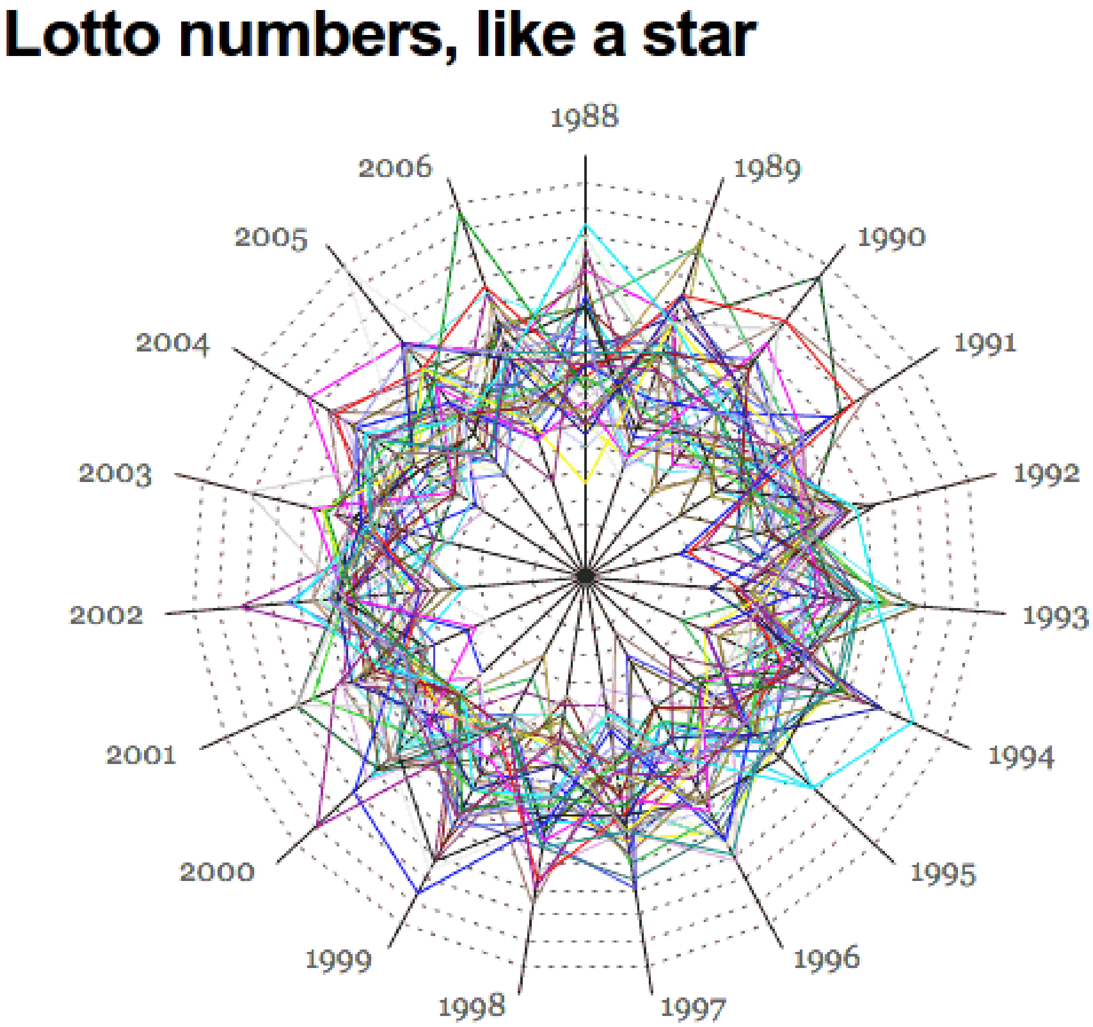
They recalibrated their instruments & next shipment was markedly improved with a mean very close to the center of our specification range of 6.055 as shown on this histogram.

10) Solution to Probable Cause 2
I'm working with engineering to develop a new fixture that will improve the geometry.

Control – 8/8/11

10) Changes to be Made:
✓QC technician does acceptance testing of raw materials w/ zero tolerance.
✓Vendor supplies Certificate of Analysis w/ test statistics.
✓Control chart created for raw materials.
✓New fixture for more uniform

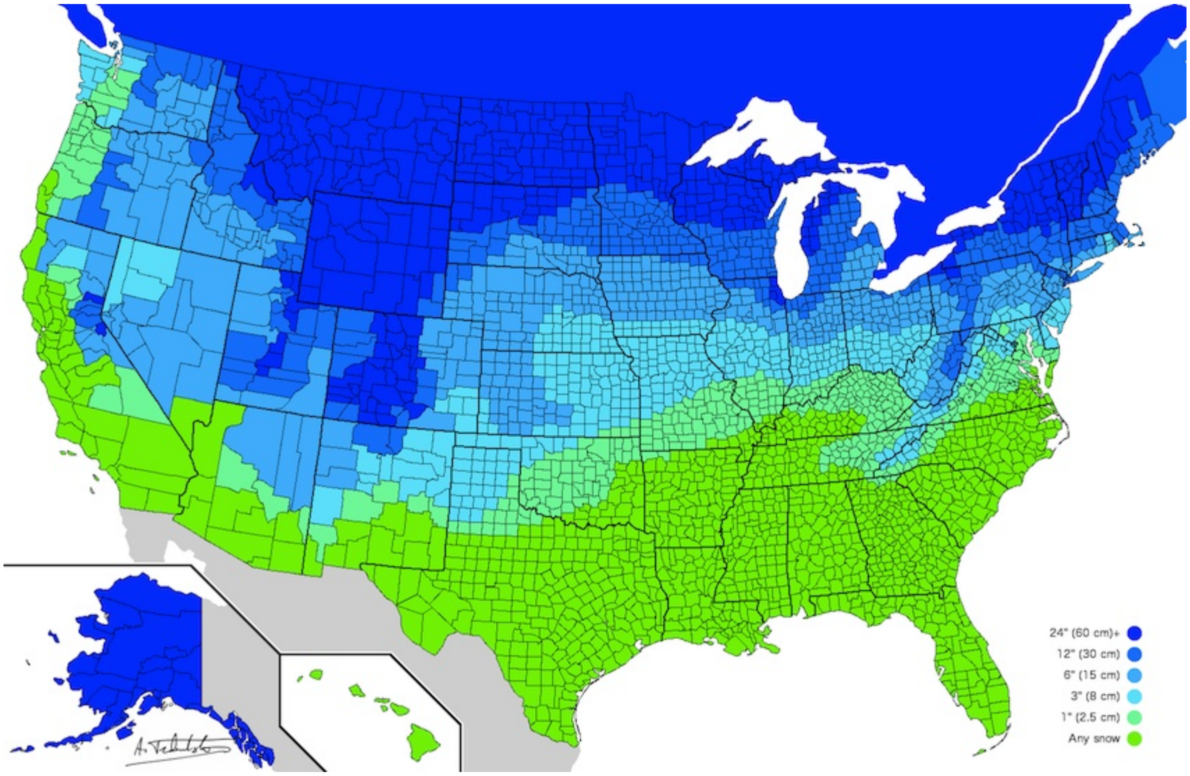
This figure contains too much information in one graph and it's too camplicated for human brain to process and get the information it wants to convey. Also, some of the labels are overlapping and cover the original information. Overall, it's too crowded, not organized, and inefficient to process.



b)

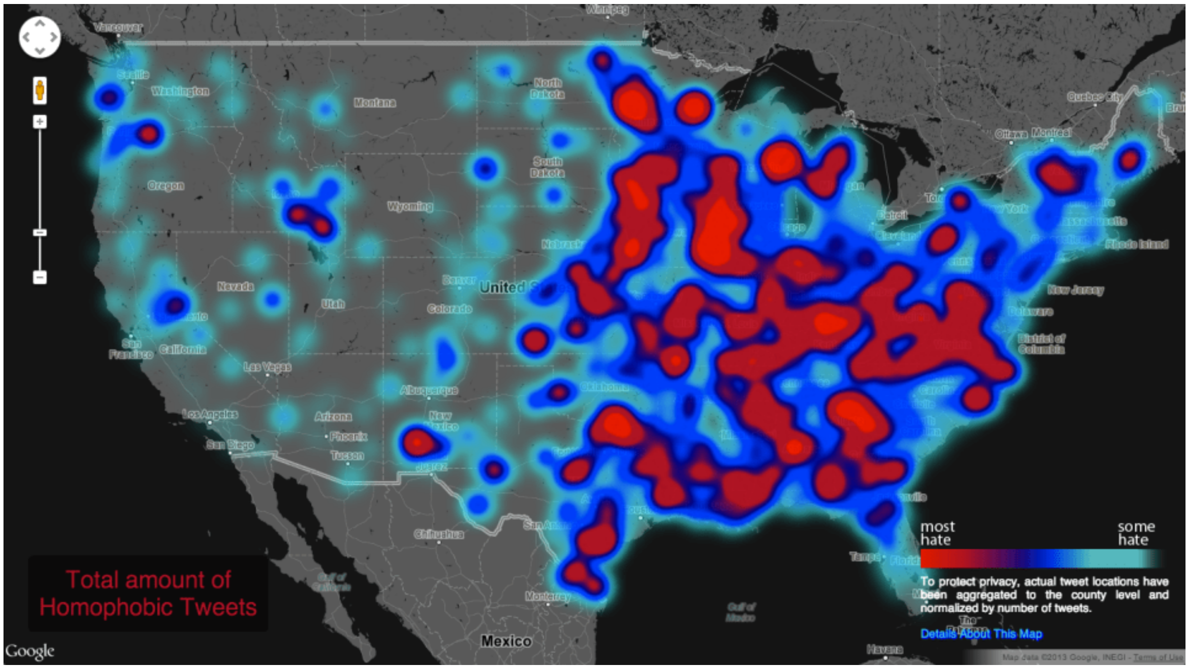
The title of this figure doesn't say anything about what this figure means. There are too many colorful lines overlapping together that readers are hard to tell where each line goes. Also, there is no label showing the meaning of each color. The graph is overall very confusing.

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



a)

This figure is trying to show the average amount of snowfall it typically takes to cancel school in different regions in US. The color and the label of this figure are very helpful for readers to understand the information. Also, visualizing the informarion on the map is very convenient for readers to match to the specific location.



b)

This picture is showing the geography of tweets that are against gays, races and the disabled. Using the heatmap to present the degree of hate in each region of the country is very clear, readers could the information at the first glance.

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

EDA is very helpful at the very beginning of the analysis of the unfamiliar dateset, because it can offer researcher a general sence of the pattern and the structure of the data, what kind of research question the given dataset could answer and what is the best methods to use for later analysis. EDA is also useful in other stages of the research, which could help researchers to find "out of box" insights.

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

According to John Tukey, confirmatory analysis is using statistical methods to test the hypothesis. An example is that given the hypothesis that beer consumption would incearse happiness, then use the statistical tool to analyse the collected data to verify whether the hypothesis is acceptable.

Exploratory analysis is that given a dataset, explore the answerable reseaech question of the dataset or generate hypothesis from the data. For example, using the visualization tools such as bar plot, scatter plot to explore the relationships of the datapoints.