Problem Set #1: Exploratory Data Analysis

Oliver Tang
Pritzker School of Molecular Engineering, University of Chicago
MACS 40800 Unsupervised Machine Learning (Autumn 2019)

Exploration & Computation

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

I will conduct research in computational biology in the following years towards my PhD degree, thus I selected a biochemical dataset from Kaggle named **Structral Protein Sequence**. Load it into R and make it a bit more tidy, leaving only statistics of interest.

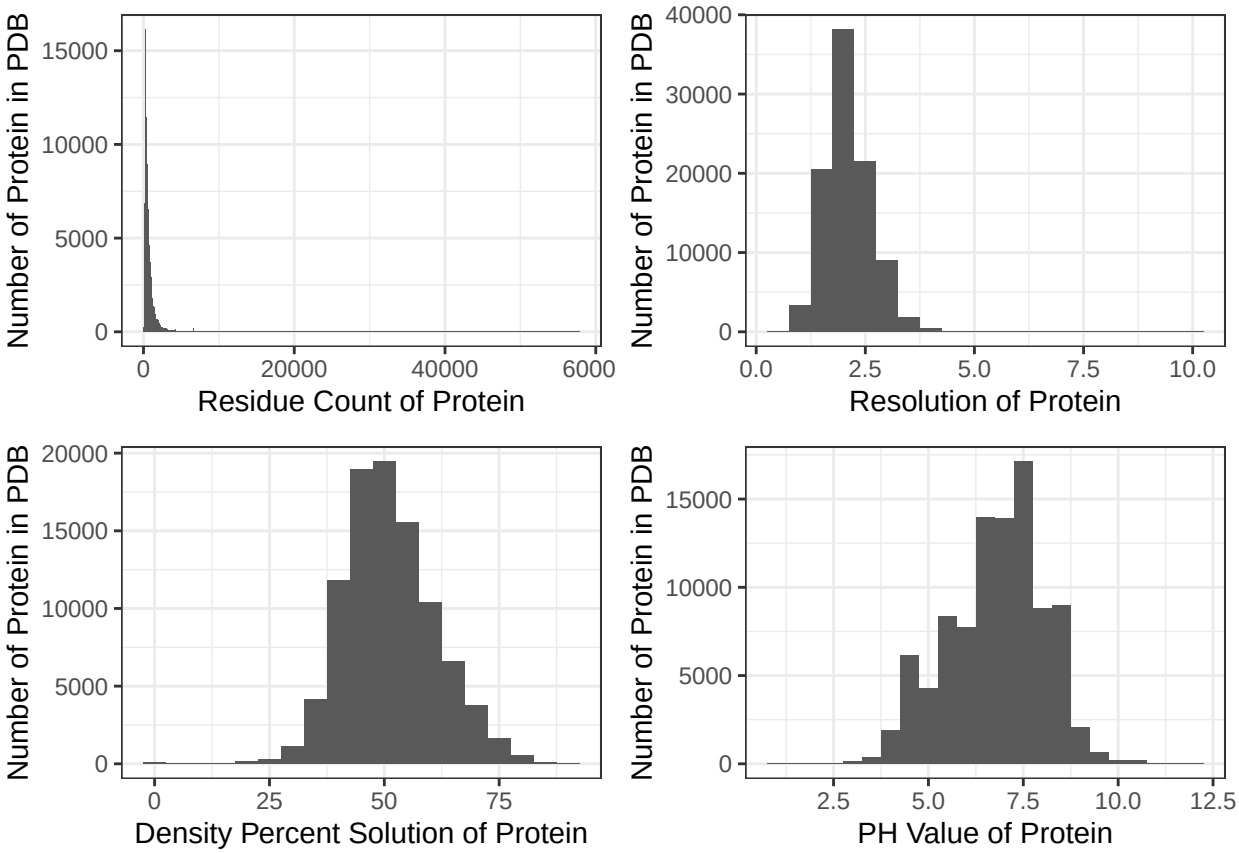
```
library(tidyverse)
library(gridExtra)
df0 = read.csv(file = "C:/Users/yifen/OneDrive/Desktop/MACS 408 UML/pdb_data_no_dups.csv")
df = df0 %>%
  dplyr::select(residueCount, resolution, densityPercentSol, phValue) %>%
  drop_na()
```

```
summary(df)
```

```
##   residueCount       resolution     densityPercentSol    phValue
##  Min.   :    5.0   Min.   : 0.480   Min.   : 0.00     Min.   : 1.000
##  1st Qu.:  277.0   1st Qu.: 1.750   1st Qu.:44.13     1st Qu.: 6.000
##  Median :  456.0   Median : 2.000   Median :50.07     Median : 7.000
##  Mean   :  722.1   Mean   : 2.113   Mean   :51.07     Mean   : 6.787
##  3rd Qu.:  838.0   3rd Qu.: 2.400   3rd Qu.:57.26     3rd Qu.: 7.500
##  Max.   :57792.0   Max.   :10.000   Max.   :92.00     Max.   :12.000
```

2. Choose a visual technique to illustrate your data: *Histogram*.

```
p1 = ggplot(data = df) +
  geom_histogram(aes(x=residueCount), binwidth=100) +
  labs(x="Residue Count of Protein",
       y="Number of Protein in PDB") + theme_bw()
p2 = ggplot(data = df) +
  geom_histogram(aes(x=resolution), binwidth=0.5) +
  labs(x="Resolution of Protein",
       y="Number of Protein in PDB") + theme_bw()
p3 = ggplot(data = df) +
  geom_histogram(aes(x=densityPercentSol), binwidth=5) +
  labs(x="Density Percent Solution of Protein",
       y="Number of Protein in PDB") + theme_bw()
p4 = ggplot(data = df) +
  geom_histogram(aes(x=phValue), binwidth=0.5) +
  labs(x="PH Value of Protein",
       y="Number of Protein in PDB") + theme_bw()
grid.arrange(p1, p2, p3, p4, ncol = 2)
```



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

1. For most of the proteins, they have residue count around several hundred, only a few with residue count for thousands or even up to 50,000.
2. The statistics of the other three properties align better with normal distribution, with some slight deformation.
3. The summary above has already shown substantial facts about the overall distribution, which is confirmed by the complete histogram. Moreover, more detailed distribution pattern could be investigated by inspecting the histogram.

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

```
se = function(x){sqrt(var(x)/(length(x)-1))}
com_measures = matrix(c(
  mean(df$residueCount),mean(df$resolution),mean(df$densityPercentSol),mean(df$phValue),
  median(df$residueCount),median(df$resolution),median(df$densityPercentSol),median(df$phValue),
  var(df$residueCount),var(df$resolution),var(df$densityPercentSol),var(df$phValue),
  se(df$residueCount),se(df$resolution),se(df$densityPercentSol),se(df$phValue)),
  ncol=4,byrow=TRUE)
colnames(com_measures) = c("Residue Count", "Resolution", "Density%", "PH Value")
rownames(com_measures) = c("mean", "median", "variance", "standard error")
com_measures
```

```
##                Residue Count  Resolution    Density%    PH Value
## mean              722.07591 2.112629954 51.07080575 6.787291632
## median            456.00000 2.000000000 50.07000000 7.000000000
## variance        920673.85731 0.301693383 99.76949538 1.611695858
## standard error       3.11507 0.001783191  0.03242753 0.004121511
```

Regarding "common measures", four kinds of measurements are chosen, namely arithmetic mean, median, variance and standard error, which is shown by row in the table above.

5. Describe the numeric output in substantive terms.

1. For resolution, density% and pH value, the difference between arithmetic mean and median is small, which means the distribution of these data is close to normal with only few outliers, which is in consistency with the histogram above. For residue count, however, the arithmetic mean is nearly two times of the median, meaning in this case, the arithmetic mean is significantly enlarged by the minor proteins who have enormous number of residues. The level of median clearly shows that most of the proteins have residue count around several hundred, aligning with the histogram, too.
2. Although the variances seem a bit too large compared to the data, given the big volume of this data set (nearly ten thousand entries), the calculated standard deviation is small, meaning that the data should be concentrated to some extent. It can also be found with the histogram.
3. As I said above regarding the summary, these common measures can grasp the "main idea" of the statistics as a whole, even without looking into the detailed distribution graph. They can give data scientists a quick hint of what to do next at most of the time.

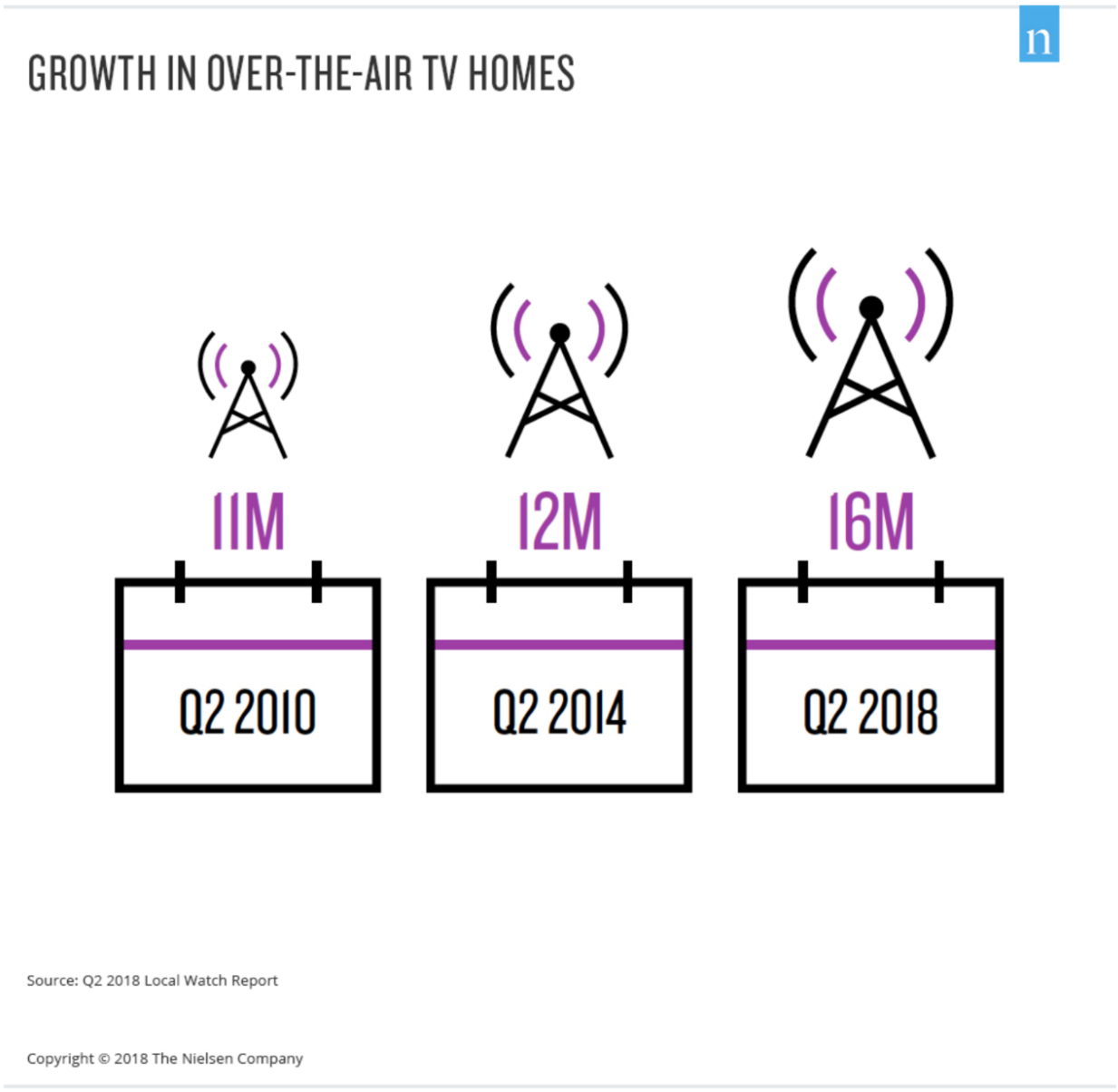
Critical Thinking

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

1. Numeric EDA, like the summary function and the commmon measures in last section, can give us a general idea of what this dataset might look like as a whole, and guide the following procedures in real-world practice. Since they do not require much computational resource to compute, it does no harm to get some numeric data first before doing visualization.
2. Visual EDA, say, the histogram in last section, and other visual techniques like barplot and scatterplot, can give us more detailed information regarding the dataset, and only on this stage, with certain statistical math, can solid conclusion be drawn. Compared to numeric EDA, visual ones take more time to compute and render, which may be more obvious on larger dataset. So, it has a trade-off, which the researchers should take into account.

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

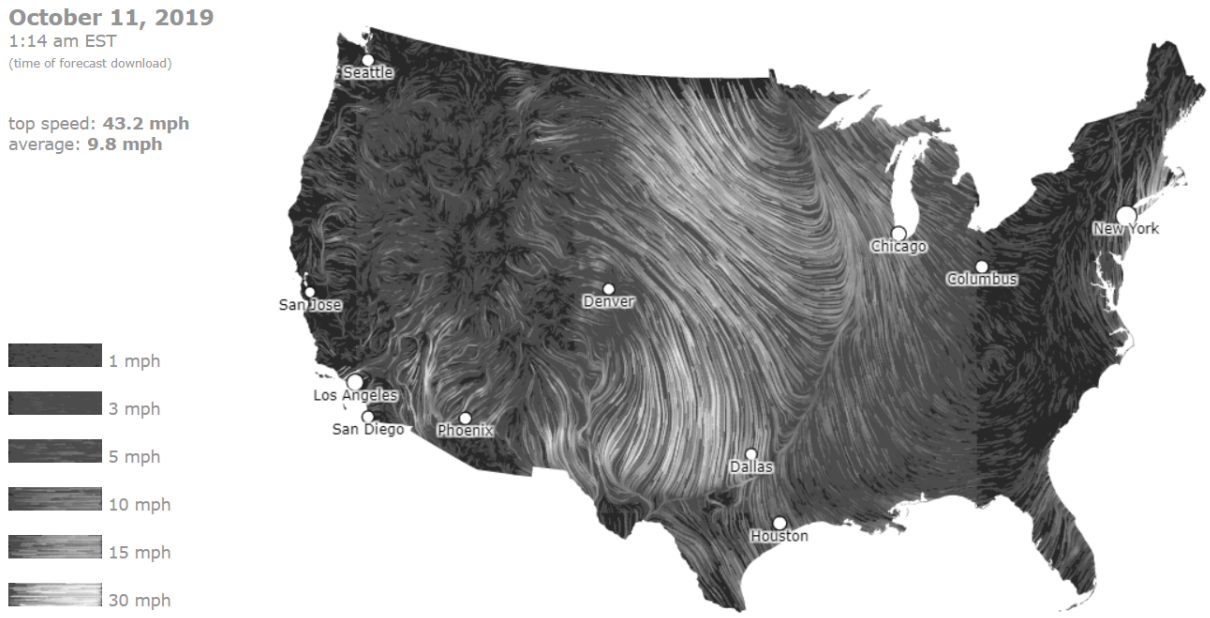
1. The histogram for residue count in last section. Because most of the data are concentrated at left side, in a very tiny part of the whole graph, leaving blank at most of the place. It is surely not good for clear visualization of the data. I think a good way to work it out may be changing the two axes to logarithm ones to fit the large numbers.
2. Pic from WTF Visualization.



Human eyes are poor when measuring angles, it is the same with sizes, and even poorer when comparing the difference between sizes, needless to say that in this graph, the size is not propotional to the data. Thus, it does not make sense to use the size of icons to visualize the data. Line charts, in this case, would do a way better job.

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

1. The histogram for three properties other than the residue count in last section. The distribution of the data is clearly shown, and the space on the charts is efficiently used. If time permits and is really needed, they can be redone with smaller band width, to show more details of the distribution.
2. Pic from Wind Map



This graph is really impressive, with clearly labeled map and legend, and moreover, fancy yet suitable animation. One can easily understand the overall wind speed and direction inside United States with just a glimpse of this site.

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

1. EDA is suitable when we are given a dataset yet do not have a clear question in mind to answer, or we want to explore more what the dataset can tell us about other than a hypothesis-testing task.
2. EDA is an idea for researchers to utilize the most of the information out of the dataset. It is like extract all the value from the given dataset, thus to maximize the possibility to make some new findings.

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

1. Confirmatory data analysis is more of a complete and definite work flow, with a clear hypothesis in mind and trying to modify the dataset with certain formal modelling, ultimately proving or falsifying the hypothesis. For instance, when we want to test if a professor gives significantly higher grades than another, we grab the data and calculate the mean, variance, do the hypothesis testing, and so on so forth.
2. Exploratory data analysis is "an attitude", according to Dr. Tukey, suggesting that there is much more information that is embedded in a dataset than we though, and EDA is a way to extract these information, which would be beyond the touch of a confirmatory data analysis. A good example would be the lab we did in class, trying to find some general patterns of people's view to politics.