Unsupervised - HW1

Allison Collins

10/10/2019

Computation & Exploration

Load the data + investigate contents

For this homework, I will be using a dataset from the CDC which shows new HIV transmission cases (and associated HIV rates) with race and year breakdown (found via the master list posted on Piazza)

```
# Load the HIV data
HIVdata <- read_csv("/Users/allisoncollins/Documents/GitHub/Problem-Set-1/hiv-data-use.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   `Race/Ethnicity` = col_character(),
##   Cases = col_number(),
##   `Rate per 100000` = col_double()
## )
```

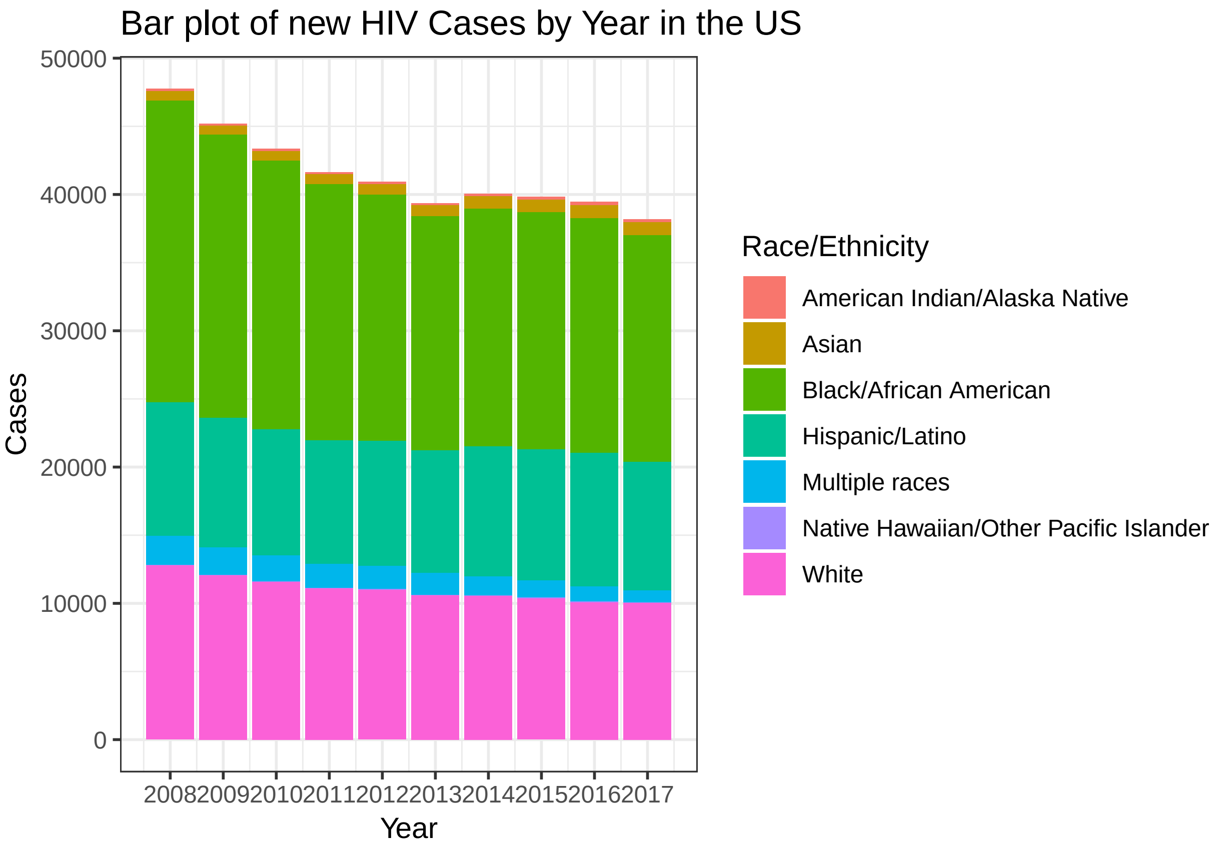
```
#Check quick column stats
summary(HIVdata)
```

```
##       Year      Race/Ethnicity         Cases          Rate per 100000
##  Min.   :2008   Length:70          Min.   :   41.0   Min.   : 5.500
##  1st Qu.:2010   Class :character   1st Qu.:  193.2   1st Qu.: 6.875
##  Median :2012   Mode  :character   Median : 1660.5   Median :12.250
##  Mean   :2012                      Mean   : 5939.0   Mean   :22.844
##  3rd Qu.:2015                      3rd Qu.:10313.0   3rd Qu.:29.200
##  Max.   :2017                      Max.   :22150.0   Max.   :74.300
```

Create a visualization

I will use a bar plot, so that we can visualize the change over time in HIV cases.

```
HIVdata %>%
ggplot(aes(x = Year, y=Cases,fill=`Race/Ethnicity`)) +
  geom_bar(stat="identity",position="stack") +
  labs(x = "Year",
       y = "Cases",
       title = "Bar plot of new HIV Cases by Year in the US") +
  scale_x_continuous(breaks=c(2008,2009,2010,2011,2012,2013,2014,2015,2016,2017))+
  theme_bw()
```



From the above bar chart, we can see that the number of HIV cases has been decreasing over time – and glancing at the chart, it would appear this decrease over time holds across racial / ethnic groups.

Calculate central measures of tendency + variation

For this, we will look at the rate per 100000 variable (as the raw number of cases varies given variation in the size of different ethnic groups in the US)

```
group_by(HIVdata, Year) %>%
 summarise(mean=mean(`Rate per 100000`), sd=sd(`Rate per 100000`),
           median = median(`Rate per 100000`))
```

```
## # A tibble: 10 x 4
##     Year  mean    sd median
##    <int> <dbl> <dbl>  <dbl>
##  1  2008  30.5  30.7   14.9
##  2  2009  28.3  28.3   13.7
##  3  2010  25.0  24.1   12.3
##  4  2011  23.4  21.9   13.4
##  5  2012  22.7  20.5   12.2
##  6  2013  21.0  19.1   10.9
##  7  2014  20.4  18.0    9.8
##  8  2015  20.5  16.8   16.4
##  9  2016  19.0  16.5   11.9
## 10  2017  17.7  15.3   11.9
```

In creating the above summary chart, what it depicts is for each year, the mean, median, and standard deviation of the HIV rate per 100000 people (this would thus be with regard to variation across the different racial groups within a given year). The mean HIV rate decreases over time (with a plateau from 2013-2015 and very small increase one year).The median decreases on the whole, but spikes in 2015 and then again declines. This is important as it implies HIV rate is decreasing over time - which we would hope to see given new treatment and prevention options and information campaigns. Moreover, the standard deviation is steadily decreasing over time which is important since it allows us to infer there is less spread between racial groups over time.

```
group_by(HIVdata, `Race/Ethnicity`) %>%
 summarise(mean=mean(`Rate per 100000`), sd=sd(`Rate per 100000`),
           median = median(`Rate per 100000`))
```

```
## # A tibble: 7 x 4
##   `Race/Ethnicity`                        mean     sd median
##   <chr>                                  <dbl>  <dbl>  <dbl>
## 1 American Indian/Alaska Native           9.1   1.43    9
## 2 Asian                                   5.98  0.316   5.95
## 3 Black/African American                 58.7   7.92   55.6
## 4 Hispanic/Latino                        23.5   2.31   22.8
## 5 Multiple races                         43.7  17.9    42.6
## 6 Native Hawaiian/Other Pacific Islander 12.4   2.27   12.2
## 7 White                                   6.47  0.527   6.35
```

In creating the above summary chart, what it depicts is for each race, the mean, median, and standard deviation of the HIV rate per 100000 people (this would thus be with regard to variation across the years for a given racial/ethnic group). This is an important view of the data, as it allows us to understand disparities between racial groups. We can see that Black/African American group has the highest mean and median HIV rate, followed by multiple race individuals, and then Hispanic/Latino. However, the multiple race racial/ethnic group has significantly higher standard deviation than other groups. Asians have the lowest mean and median rate, followed closely by whites. Notably, the standard deviation is also low for these two groups, which allows us to infer that over this 10 year period the rate did not vary widely (because of this measure telling us the magnitude of the distance to the mean; we know that the different rates over the different years were very close to the mean)

Critical Thinking

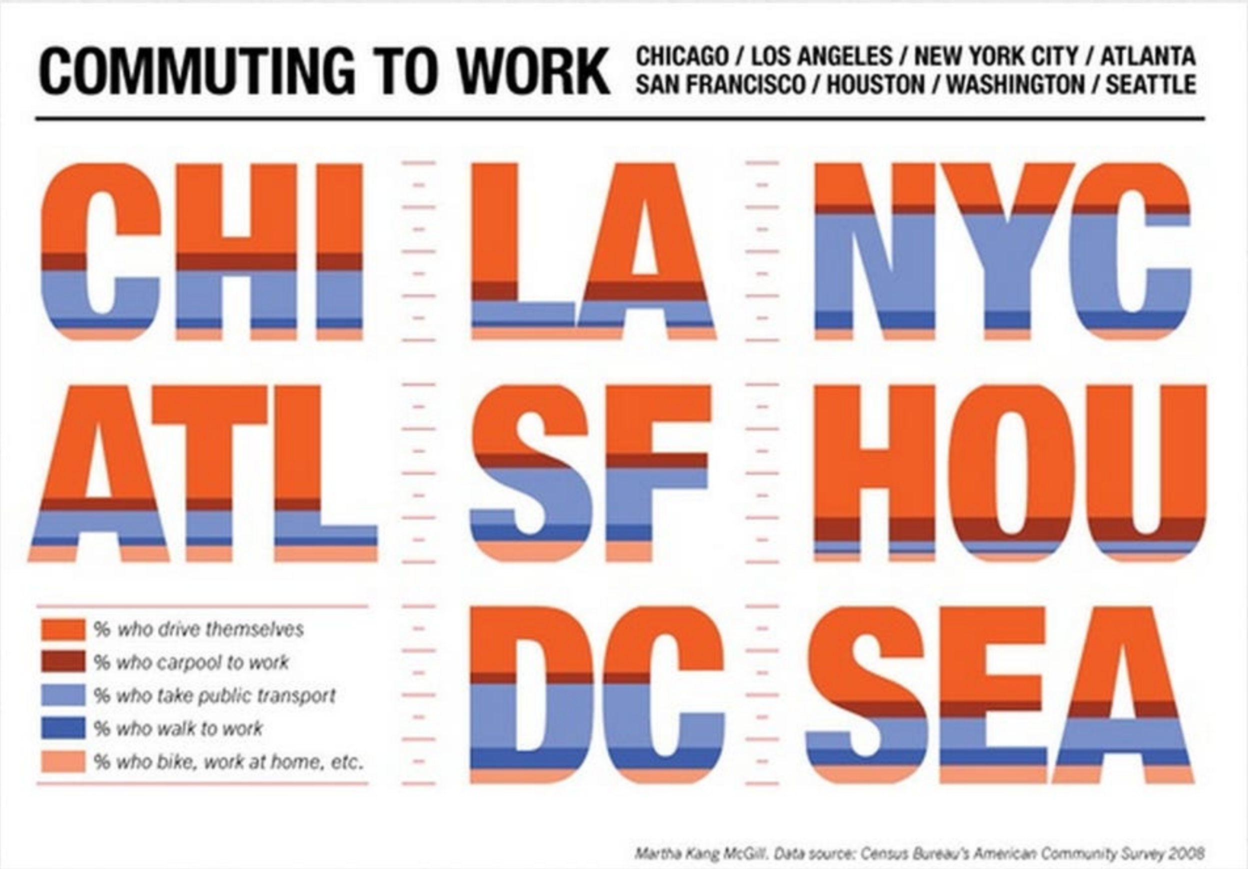
1. Describe the different information contained in/revealed by visual versus numeric exploratory data analysis. (Hint: Think of different examples of each and then what we might be looking for when leveraging a given technique).

Both visual and numeric techniques can be useful in exploratory data analysis. Numeric analysis can be useful in giving us quick “fast facts” – if we want to know what the average value is, how variable the data is, etc. This can be particularly useful if we are thinking across groups; for example, if we want to look how different racial, age, or geographic groups compare. If we were talking about income and wanted to get a sense of where the disparities lie, quick numeric summaries could allow us to make those comparisons.

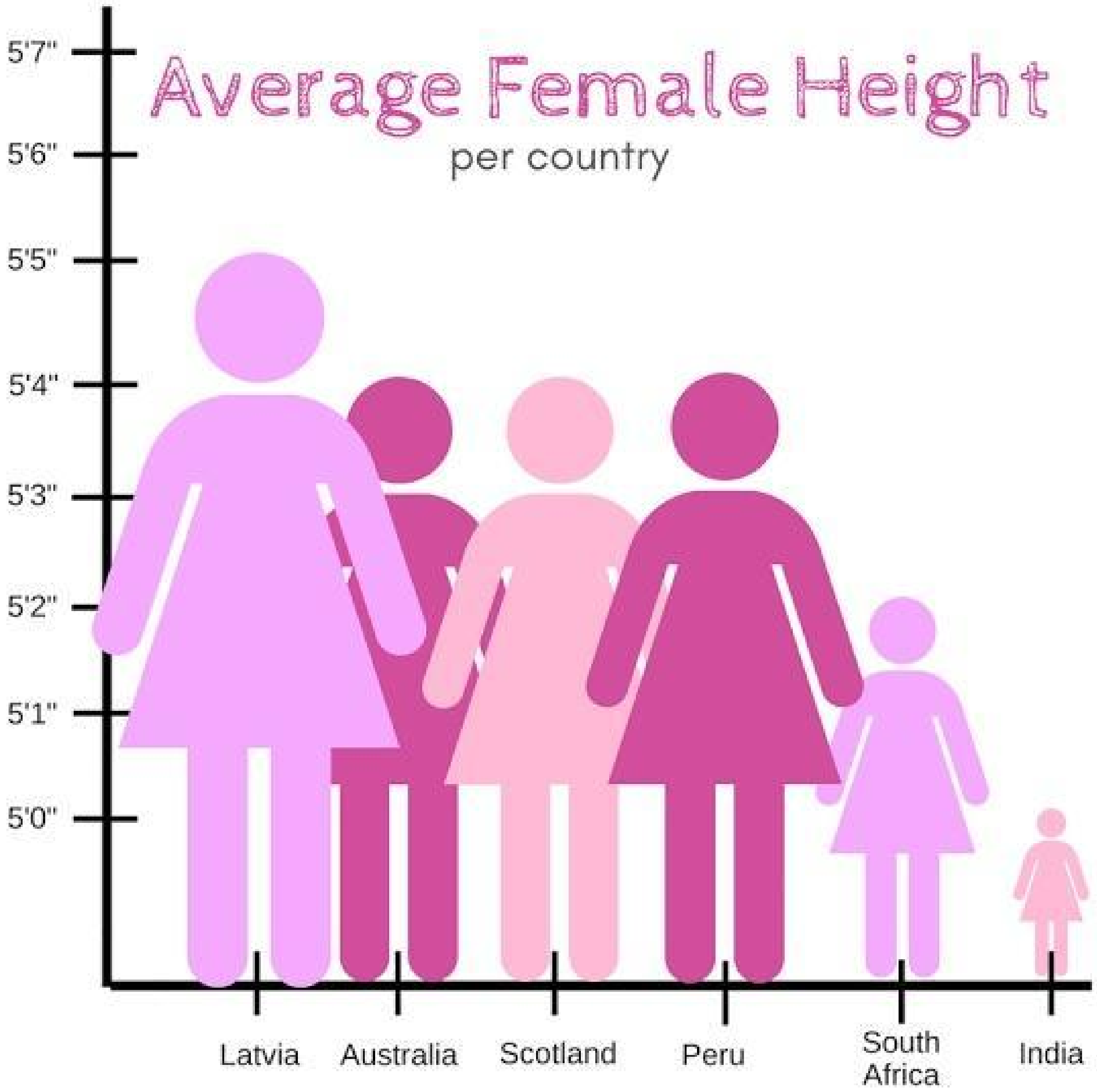
Visual techniques can be very useful for looking at trends over time; while this can also be captured in numeric summaries, visually it can be easier to see a gradual increase/decrease, break in the trend, etc. Further, if we are concerned about data cleaning or understanding where outliers could potentially be, this would be shown in a chart (say a box and whisker plot), but the mean would smooth out the presence of an outlier.

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

```r
#Visualization 1
knitr::include_graphics("bad_viz_1.jpeg")
```



```r
#Visualization 2
knitr::include_graphics("bad_viz_2.jpg")
```



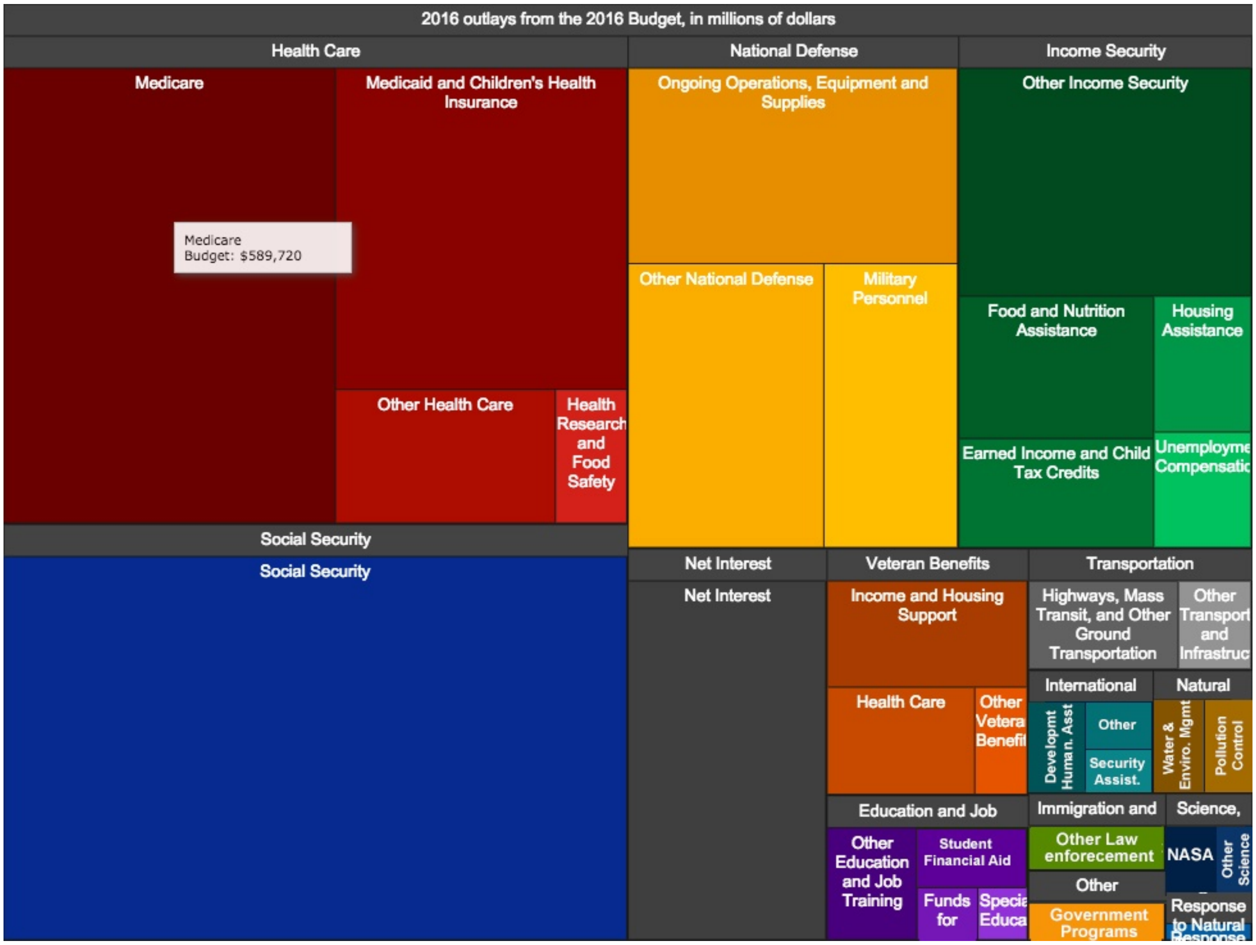
The above 2 images are examples of bad data visualizations, which I chose because they seemed like "cool" ways to show the data, but I think it came at the expense of people understanding the data.

In the first one, it is hard to compare across cities having to look at all the acronyms which go both up and down and left to right, vs on a chart where we could look in one direction and more easily see the difference in modes of transportation taken to work (as the size of the different bar components grows or shrinks).Also, because of the various shapes it is hard to grasp what the % is – it's harder to see what 40% of the letter A is vs of a bar. Lastly, is there a meaning of letter width? It's not clear and again makes it harder to understand the percent of people taking that form of transport (image source - business insider)

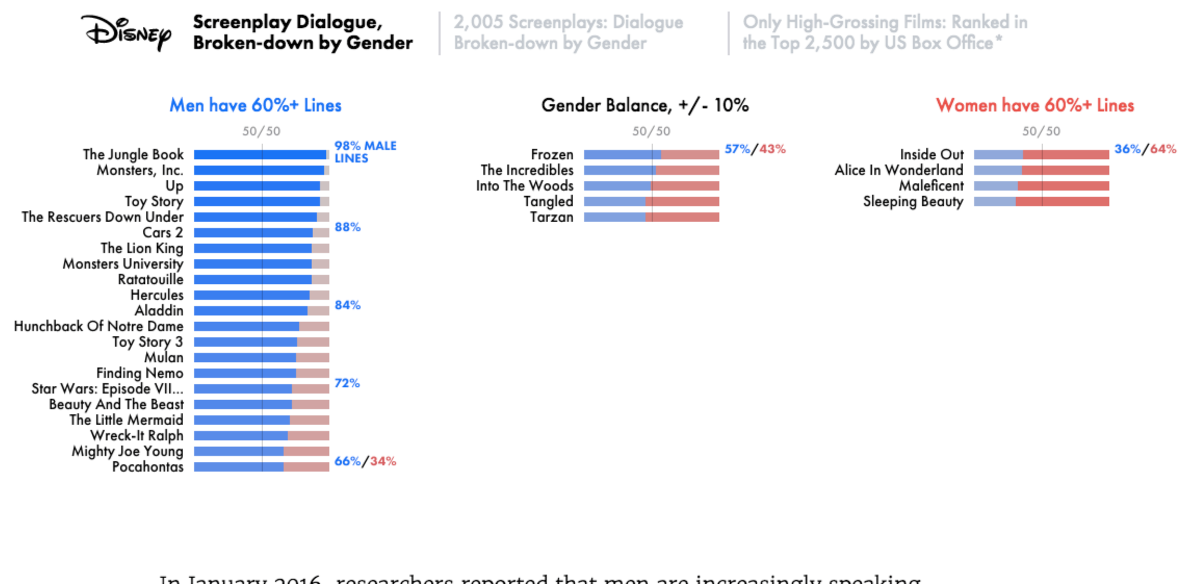
In the second visual, because the y-axis does not start at 0, the scaling is misleading. It appears that the height of the average woman in India is ~1/5 that of Latvia, but in actuality it is a 5 inch difference. Further, in trying to get creative and use shapes for the bars, they are all different widths, which again could imply different sizes of women (image source - Tumblr)

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

```r
#Visualization 1
knitr::include_graphics("good_viz_1.jpg")
```



```r
#Visualization 2
knitr::include_graphics("good_viz_2.png")
```



The above 2 images are examples of good data visualizations.

I chose the first one not because it is an example of a super pretty visualization but rather because it makes often nebulous information easy to comprehend. This visual depicts the US budget (source - tableau article; office of management and budget), which people often have misperceptions about. It is actually interactive where you can hover for more information, but I included a static version which I think is still a good example. By grouping the different categories in similar color schemes, it allows us to see big picture how that category is size-wise, and then within that, breaks out individual components. Overall, it clearly conveys how the money is being spent; what the largest categories are, and dispels myths about the budget/spend.

In the second visual, I think this shows how clean images can be highly effective. The grouping by percentage of words spoken makes it easy to understand what group the films split into – and also, what the overall distribution of films where men vs. women get more speaking roles (since the list with males speaking a higher percentage is much longer) Then within each group, the use of the red/blue bars with different levels of vibrancy (and ordered by percentage) make it clear to understand what the distribution is (source: Tableau article)

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

EDA is very useful at the offset, to get an overall sense of what the data is suggesting before formulating a research question – it can help generate initial hypotheses (such as on disparities among different groups, an overall time trend, etc.)

Moreover, it can be very useful in the data cleaning phase of the research process, in helping identify outliers for example, that we might want to exclude from analysis. Thus it can help formulate the question up front - and then it serves an additional purpose again in data gathering/model generation.

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

Confirmatory analysis refers to seeking out an answer to a question - that drives the design of the research, data collection, study and analysis. He explains typically there is this singular focus given that it is hard to have alternative analyses in the context of confirming in/not in favor of a null hypothesis. An example could be a study to determine if white students perform better than black students on the SATs, where confirming if this relationship in performance between the two groups exists is the only outcome.

Explanatory data analysis is often thought of as descriptive statistics, but Tukey says that it is more about the attitude/flexibility. For example, if we think about race and SAT score, we might seek to understand if data suggests a relationship between race and SAT performance. We could make charts that look at average scores by race, or the distribution, or time trends.