UML - Problem Set 1

Jesús Pacheco

10/11/2019

Exploration & Computation

1. Obtain a dataset

The dataset I chose is the labor force participation by gender in Mexico in last ~20 years. I'm interested in knowing the gender wage gap, so I started by looking at how the female labor force is changing. The data was taken from the UN Gender Statistics portal: https://genderstats.un.org/

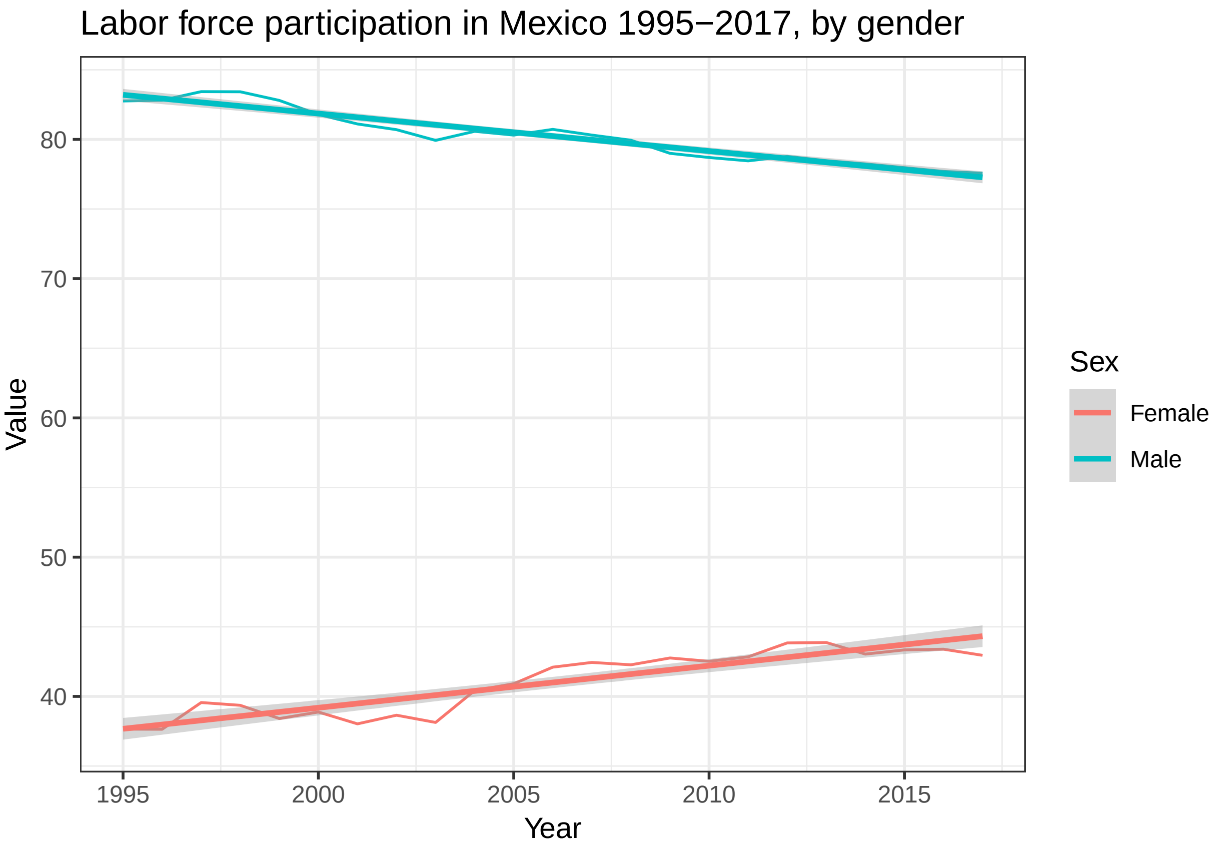
```
lfp <- read_csv("/Users/JesusPacheco/Downloads/Labour_force_participation_Mex.csv", skip=2, col_names=T
skim(lfp)
```

2. Choose a visual technique to illustrate your data

I chose a line showing the time trend.

3. Generate and present the visualization and describe what you see.

```
lfp %>%  filter(Sex != "Both sexes") %>% ggplot(aes(x=Year, y=Value, colour=Sex)) + geom_line() +
  geom_smooth(method = lm) +
  labs(title="Labor force participation in Mexico 1995-2017, by gender") + theme_bw()
```



We can see a substantial difference between labor force participation: around 80 percent of the males in ages 15 and above, while only half of that for women. We can see that the gap is slowly closing, since women appear to be working more outside the home after 2004.

4. Common measures of central tendency and variation

```
lfp %>%  filter(Sex != "Both sexes") %>% group_by(Sex)  %>%  summarize(Average = mean(Value), Median =me
```

```
## # A tibble: 2 x 4
##   Sex    Average Median Variance
##   <chr>    <dbl>  <dbl>    <dbl>
## 1 Female    41     42.1     5.02
## 2 Male      80.2   80.3     3.58
```

```
lm(Value ~Year, data =lfp, subset=(Sex=="Female"))
```

```
##
## Call:
## lm(formula = Value ~ Year, data = lfp, subset = (Sex == "Female"))
##
## Coefficients:
## (Intercept)         Year
##   -565.3393       0.3023
```

```
lm(Value ~Year, data =lfp, subset=(Sex=="Male"))
```

```
##
## Call:
```

```
## lm(formula = Value ~ Year, data = lfp, subset = (Sex == "Male"))
##
## Coefficients:
## (Intercept)         Year
##    620.3130      -0.2692
```

5. Describe the numeric output in substantive term

As we could imagine from looking at the graph, we see that over the last 22 years, in Mexico men are twice as likely as women to have a job outside the home. The variation over the time period, as expected, is not very big since we would expect these social phenomena to move at a slow pace. But we do see an increase in the female participation. Furthermore, from a simple linear model we can get an estimate of the trend. From those trends we can calculate

$$42.95 + .3023x = 77.57 - .2692x$$

$$x = 53.6$$

Hence, keeping this trend it would take 53 years for woman to have equal participation in the job market. This type of statistics, although not very rigorous, are usually shown to highlight a notion of how fast something is increasing/decreasing. The data seem very relevant for policy implications. Actually in the Mexican current policy context, public daycare programs have suffer important cuts since last year.

Critical Thinking

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

When we are using the visualization, we are usually looking at how the data is distributed, to discover any trend, abnormality or even to "pre-test" some theory that we might have *a priori*. When we use numeric EDA, we are looking at single estimates for a given measurement. Although this technique is more precise, sometimes it can give not very helpful estimates or even bias our notion of the data as highlighted by Anscombe's quartet.

2. Two examples of "bad" visualizations

First one:

This graph is incredibly misleading, it's trying to make your eye recognize a huge increase in the tax rate: the second bar is 5 or more times larger than the first when the rate is only going from 35 to 39.6 percent. I believe this effect happens, to different extents, everytime we see axis not starting at zero.

Second one:

This chart is trying to show the number of features each new version of Microsoft Word added to the previous one, but it is doing a terrible job. First, there's a chronological order but the pie chart is not good for conveying that order. But it isn't good at conveying dimensions either since we need to measure angles in our minds: it's hard to see if the 6.0 had more or less features than the 2.0 just by looking at the chart. Third, the 3D feature is almost never helpful in this type of graphs, it only adds more difficulty to measure the angles. Finally, I would argue that this many colors are also not helpful (not only because I'm color blind, but because it's hard to keep track of so many colors).

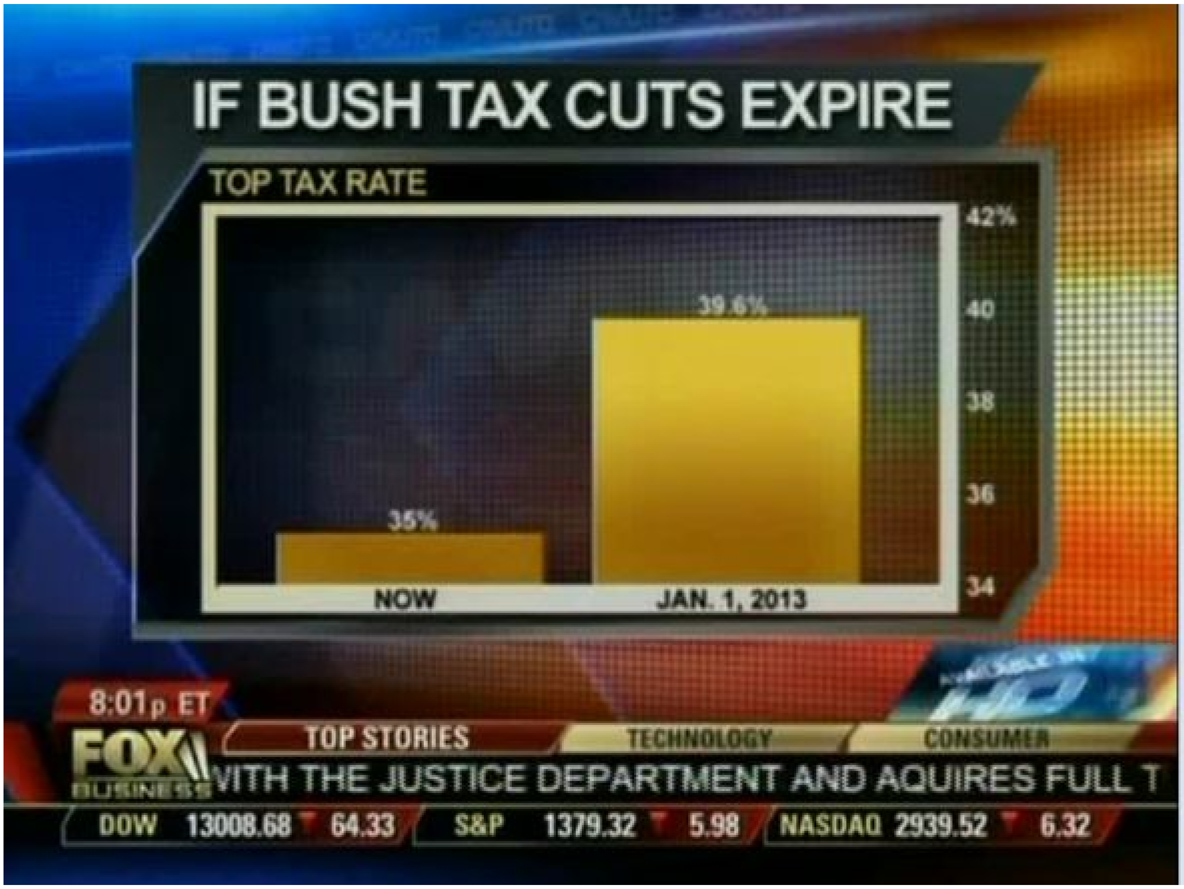


Figure 1: https://www.mediamatters.org/static/images/item/fbn-cavuto-20120731-bushexpire.jpg

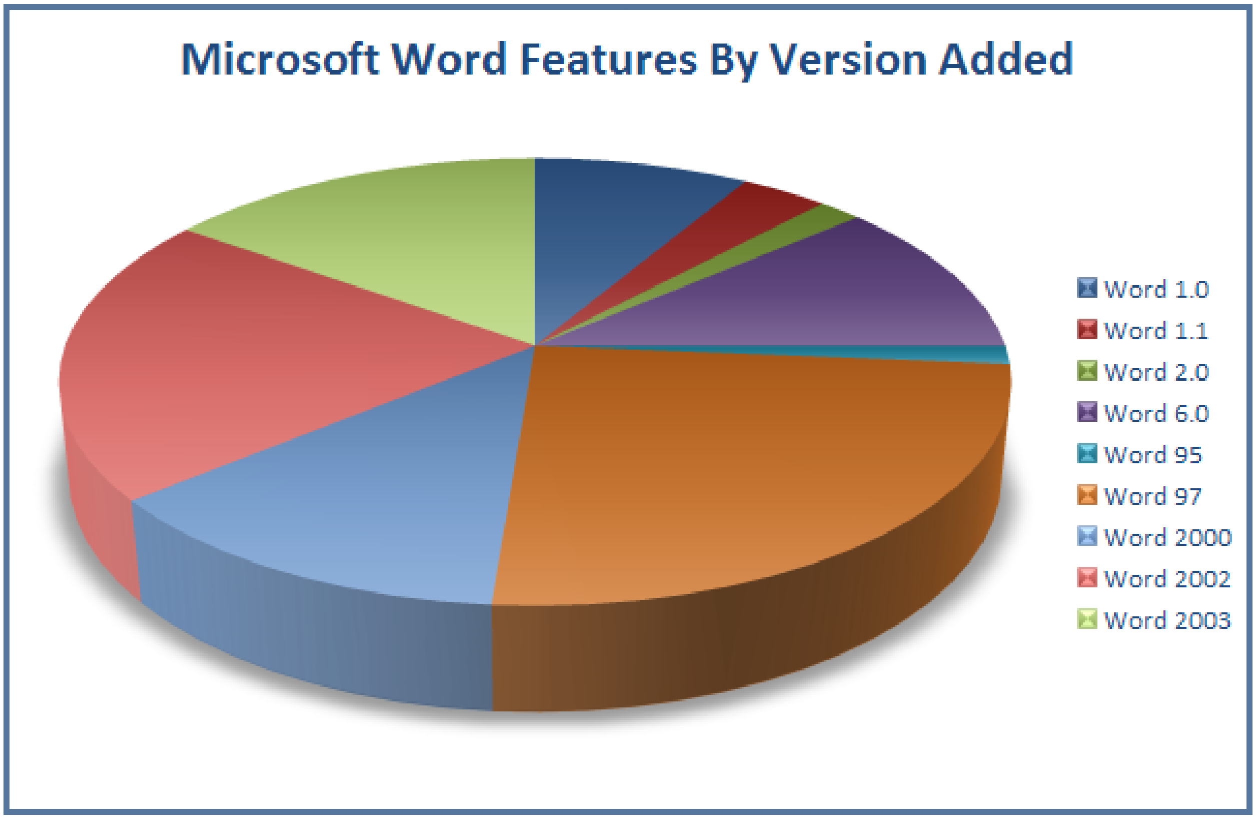
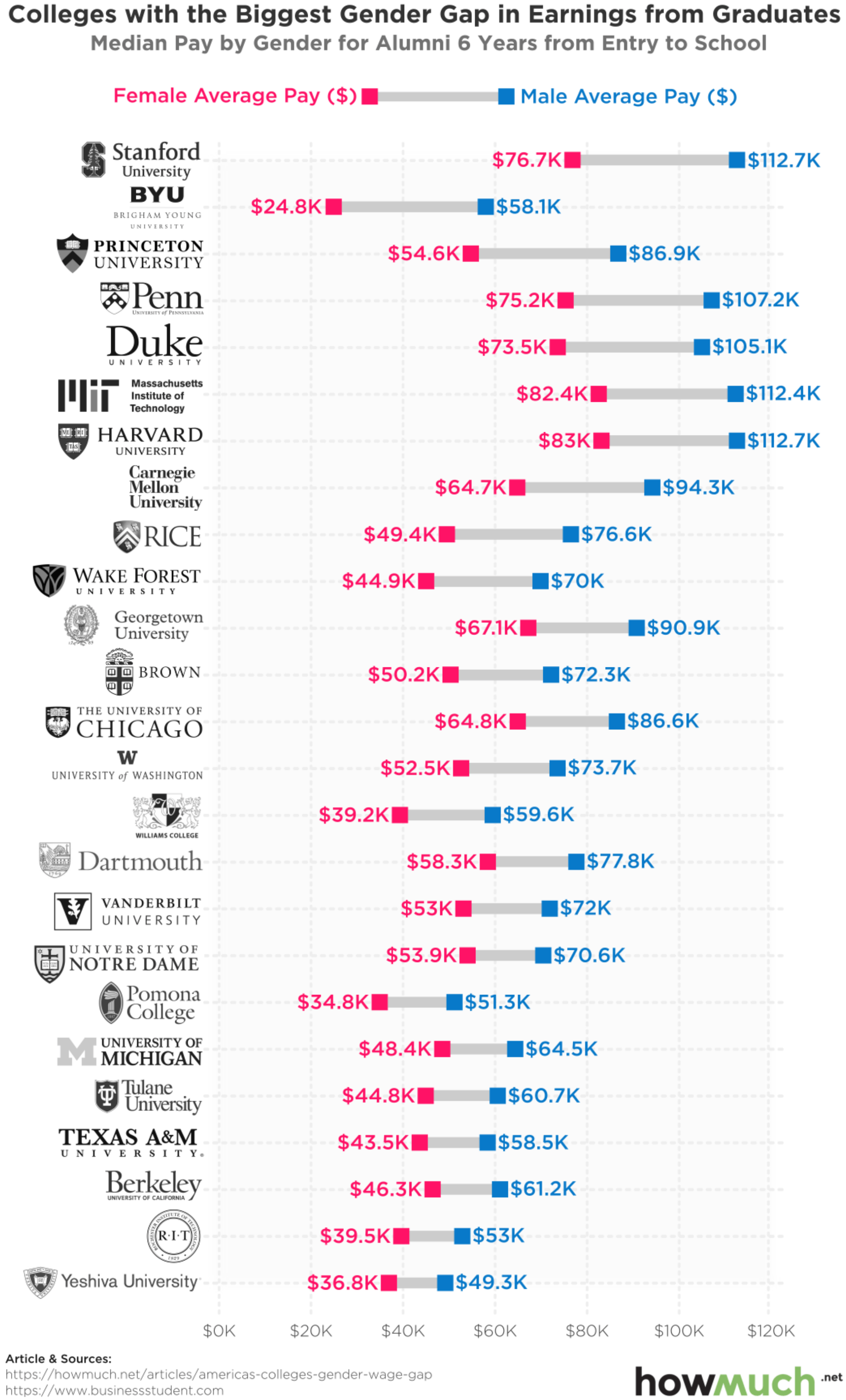

Microsoft Word Features By Version Added
Word 1.0
Word 1.1
Word 2.0
Word 6.0
Word 95
Word 97
Word 2000
Word 2002
Word 2003

Figure 2:

3. Two examples of "good" visualizations

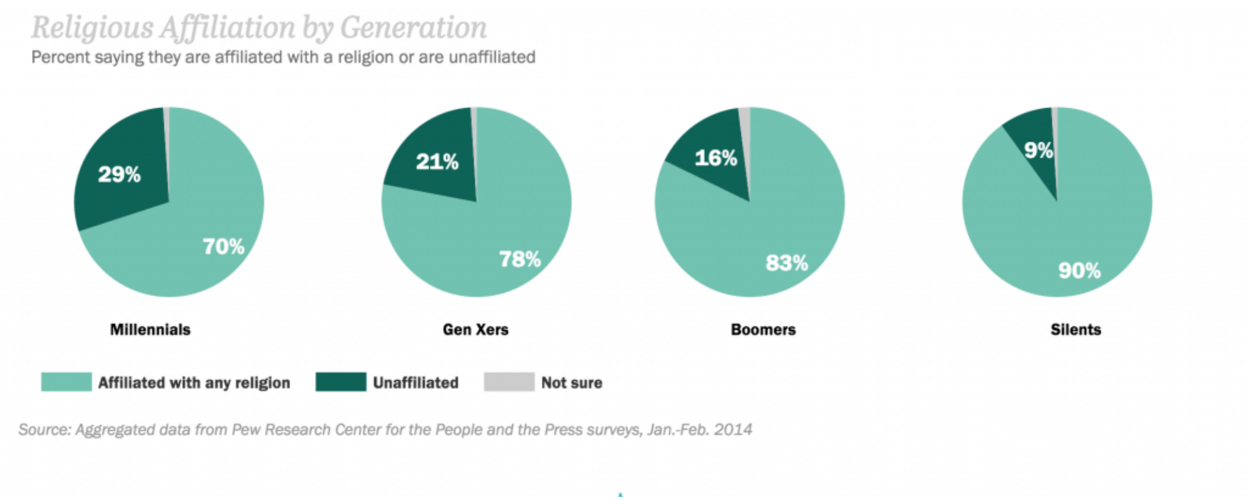


Colleges with the Biggest Gender Gap in Earnings from Graduates

Median Pay by Gender for Alumni 6 Years from Entry to School

Article & Sources:
https://howmuch.net/articles/americas-colleges-gender-wage-gap
https://www.businessstudent.com

First one: I like this visualization because it shows three different pieces of information at the same time without making the it too crowded with information. We can see the male median wage, female median wage and the gap between them. We can compare overall wages between universities and also how big the gap is and start to look for possible causal interpretations for that. Overall, the images, the logos and the colors, everything seems pretty neat and clear to me.



Source: Aggregated data from Pew Research Center for the People and the Press surveys, Jan.-Feb. 2014

Second one:

For the second one I choose a good example of the use of a pie chart. I think in this case, our brain is not too troubled with just comparing two angles. Having the graphs ordered as they are, we can see relatively easily that the amount of proportion of people affiliated with a certain religion is increasing with the generation: newer generations are less prone to religion. In the graph we can see the numbers as well and the “pie”" design makes it less boring (although, other types of grapghs could have done the trick as well).

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

EDA is important to get to know the data. It helps the research process with the overall understanding of a certain phenomenon or relation. It can help with small adjustments to our priors, to direct our minds towards patterns that might be useful in the next steps of the process. It can also keep us from following wrong directions.

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

Exploratory is the kind of analysis you do before you have a formal (“tidy”) research question. Ideas for research come, as stated by Tukey, from exploration. An example of exploratory analysis would be data visualization, to try to find relations between variables, let the data unveil potential causal relations.

Confirmatory is the part of the research process in which the objective is to test if the theory (priors) is supported by the facts (data). An example of confirmatory analysis, as I understand it, would be to run a regression and test a hypothesis on the coefficient to see if the expected relationship exists in the data.

The versus part is because sometimes, it is taught that confirmatory analysis is the only one necessary for inquiries, but Tukey argues that we actually need both. Important questions demand confirmatory analysis, but those right questions come sometimes from exploration, hence the role of exploratory analysis is not be undermined.