Post 01

Brian Hsu

October 24, 2017

Introduction

The purpose of this post is to look at a possible approach to analyzing multiple sets of data, each representing a different year, and to draw conclusions from the data. I will be using data from Transparent California, a site that aims to provide public access to salary data on employees for the State of California. More specifically, I will look at data involving the University of California system, and attempt to look at trends from 2011-2016 involving increases in salary.

Downloading and Reading the Data

To download the data, you can visit this site and click download records, or you can follow these instructions:

These commands need only be run once, within the R console. Be sure to replace DESIRED_DESTINATION with the desired file path.

```
# Download copy
download.file('https://transparentcalifornia.com/export/university-of-california-2016.csv', DESIRED_DESTINATION)
download.file('https://transparentcalifornia.com/export/university-of-california-2015.csv', DESIRED_DESTINATION)
download.file('https://transparentcalifornia.com/export/university-of-california-2014.csv', DESIRED_DESTINATION)
download.file('https://transparentcalifornia.com/export/university-of-california-2013.csv', DESIRED_DESTINATION)
download.file('https://transparentcalifornia.com/export/university-of-california-2012.csv', DESIRED_DESTINATION)
download.file('https://transparentcalifornia.com/export/university-of-california-2011.csv', DESIRED_DESTINATION)
```

Now, read the data, we need to load some libraries – readr is used to read the csv files, dplyr for data wrangling, and ggplot2 for data visualization.

```
library(readr)
library(dplyr)
library(ggplot2)
```

Now, we can use the read_csv function from the readr library.

```
dat2016 = read_csv("../data/university-of-california-2016.csv")
dat2015 = read_csv("../data/university-of-california-2015.csv")
dat2014 = read_csv("../data/university-of-california-2014.csv")
dat2013 = read_csv("../data/university-of-california-2013.csv")
dat2012 = read_csv("../data/university-of-california-2012.csv")
dat2011 = read_csv("../data/university-of-california-2011.csv")
```

Now, we can run some basic functions to look at the structure of the newly-imported data frames.

```
summary(dat2016)
```

```
##  Employee Name       Job Title            Base Pay
##  Length:291141      Length:291141      Min.   :  -3494
##  Class :character   Class :character   1st Qu.:   3734
##  Mode  :character   Mode  :character   Median :  25533
##                                        Mean   :  41952
##                                        3rd Qu.:  61808
##                                        Max.   :1012846
##
##   Overtime Pay        Other Pay          Benefits        Total Pay
##  Min.   : -8585.0   Min.   : -53433   Min.   :     0   Min.   :      2
##  1st Qu.:     0.0   1st Qu.:      0   1st Qu.:     0   1st Qu.:   4347
##  Median :     0.0   Median :      0   Median :  2130   Median :  27645
##  Mean   :   705.1   Mean   :   6109   Mean   : 10712   Mean   :  48765
##  3rd Qu.:     0.0   3rd Qu.:   1652   3rd Qu.: 20946   3rd Qu.:  65906
##  Max.   :154153.0   Max.   :3277299   Max.   :128406   Max.   :3577299
##  NA's   :1          NA's   :5
##  Total Pay & Benefits      Year         Notes              Agency
##  Min.   :      2      Min.   :2016   Length:291141      Length:291141
##  1st Qu.:   4456      1st Qu.:2016   Class :character   Class :character
##  Median :  30978      Median :2016   Mode  :character   Mode  :character
##  Mean   :  59477      Mean   :2016
##  3rd Qu.:  86007      3rd Qu.:2016
##  Max.   :3633894      Max.   :2016
##
##     Status
##  Length:291141
##  Class :character
##  Mode  :character
##
##
##
##
```

```
summary(dat2015)
```

```
##  Employee Name       Job Title            Base Pay       Overtime Pay
##  Length:281514      Length:281514      Min.   : -2668   Min.   : -2817.0
##  Class :character   Class :character   1st Qu.:  3520   1st Qu.:     0.0
##  Mode  :character   Mode  :character   Median : 24620   Median :     0.0
##                                        Mean   : 40531   Mean   :   687.1
##                                        3rd Qu.: 59742   3rd Qu.:     0.0
##                                        Max.   :971205   Max.   :135091.0
##
##    Other Pay           Benefits        Total Pay       Total Pay & Benefits
##  Min.   : -54002   Min.   :     0   Min.   :      1   Min.   :      1
##  1st Qu.:      0   1st Qu.:     0   1st Qu.:   4158   1st Qu.:   4266
##  Median :      0   Median :  2046   Median :  26667   Median :  29878
##  Mean   :   5730   Mean   : 10453   Mean   :  46947   Mean   :  57401
##  3rd Qu.:   1600   3rd Qu.: 20429   3rd Qu.:  63696   3rd Qu.:  83351
##  Max.   :3214771   Max.   :122650   Max.   :3514771   Max.   :3570343
##  NA's   :4
##       Year         Notes              Agency             Status
##  Min.   :2015   Length:281514      Length:281514      Length:281514
##  1st Qu.:2015   Class :character   Class :character   Class :character
##  Median :2015   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2015
##  3rd Qu.:2015
##  Max.   :2015
##
```

```
summary(dat2014)
```

```
##  Employee Name       Job Title            Base Pay       Overtime Pay
##  Length:275257      Length:275257      Min.   :-24433   Min.   : -5041.0
##  Class :character   Class :character   1st Qu.:  3364   1st Qu.:     0.0
##  Mode  :character   Mode  :character   Median : 23852   Median :     0.0
##                                        Mean   : 39529   Mean   :   624.8
##                                        3rd Qu.: 58937   3rd Qu.:     0.0
##                                        Max.   :981215   Max.   :131322.0
##
##    Other Pay           Benefits        Total Pay       Total Pay & Benefits
##  Min.   : -70340   Min.   :     0   Min.   :      1   Min.   :      1
##  1st Qu.:      0   1st Qu.:     0   1st Qu.:   3970   1st Qu.:   4066
##  Median :      7   Median :  1525   Median :  25763   Median :  28558
##  Mean   :   5478   Mean   :  9728   Mean   :  45631   Mean   :  55360
##  3rd Qu.:   1854   3rd Qu.: 19223   3rd Qu.:  62764   3rd Qu.:  81563
##  Max.   :3176127   Max.   :118442   Max.   :3476127   Max.   :3526895
##  NA's   :9
##       Year         Notes              Agency             Status
##  Min.   :2014   Length:275257      Length:275257      Length:275257
##  1st Qu.:2014   Class :character   Class :character   Class :character
##  Median :2014   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2014
##  3rd Qu.:2014
##  Max.   :2014
##
```

```
summary(dat2013)
```

```
##  Employee Name       Job Title            Base Pay       Overtime Pay
##  Length:268442      Length:268442      Min.   :-17633   Min.   : -1905.0
##  Class :character   Class :character   1st Qu.:  3326   1st Qu.:     0.0
##  Mode  :character   Mode  :character   Median : 23585   Median :     0.0
##                                        Mean   : 37745   Mean   :   592.8
##                                        3rd Qu.: 55980   3rd Qu.:     0.0
##                                        Max.   :931424   Max.   :138620.0
##
##    Other Pay           Benefits        Total Pay       Total Pay & Benefits
##  Min.   :-345255   Min.   :-16086   Min.   :      1   Min.   :  -1223
##  1st Qu.:      0   1st Qu.:     0   1st Qu.:   3923   1st Qu.:   3923
##  Median :      0   Median :  1509   Median :  25488   Median :  28235
##  Mean   :   5182   Mean   :  8703   Mean   :  43520   Mean   :  52223
##  3rd Qu.:   1521   3rd Qu.: 17089   3rd Qu.:  59525   3rd Qu.:  76226
##  Max.   :2442860   Max.   :108954   Max.   :2639609   Max.   :2675371
##  NA's   :1         NA's   :2
##       Year         Notes              Agency
##  Min.   :2013   Length:268442      Length:268442
##  1st Qu.:2013   Class :character   Class :character
##  Median :2013   Mode  :character   Mode  :character
##  Mean   :2013
##  3rd Qu.:2013
##  Max.   :2013
##
```

```
summary(dat2012)
```

```
##  employee_name       job_title           base_pay
##  Length:262416      Length:262416      Min.   :  -802.1
##  Class :character   Class :character   1st Qu.:  3461.8
##  Mode  :character   Mode  :character   Median : 24468.9
##                                        Mean   : 37223.7
##                                        3rd Qu.: 54892.5
##                                        Max.   :935006.4
##   overtime_pay        other_pay       total_benefits       total_pay
##  Min.   :  -458.8   Min.   :-108400   Length:262416      Min.   :      1
##  1st Qu.:     0.0   1st Qu.:      0   Class :character   1st Qu.:   4152
##  Median :     0.0   Median :      0   Mode  :character   Median :  26027
##  Mean   :   595.1   Mean   :   4919                      Mean   :  42738
##  3rd Qu.:     0.0   3rd Qu.:   1749                      3rd Qu.:  58329
##  Max.   :118640.3   Max.   :1968243                      Max.   :2234192
##  total_pay_benefits      year         notes           jurisdiction_name
##  Min.   :      1    Min.   :2012   Length:262416      Length:262416
##  1st Qu.:   4152    1st Qu.:2012   Class :character   Class :character
##  Median :  26027    Median :2012   Mode  :character   Mode  :character
##  Mean   :  42738    Mean   :2012
##  3rd Qu.:  58329    3rd Qu.:2012
##  Max.   :2234192    Max.   :2012
```

```
summary(dat2011)
```

```
##  employee_name       job_title           base_pay
##  Length:259043      Length:259043      Min.   :  -571.2
##  Class :character   Class :character   1st Qu.:  3336.4
##  Mode  :character   Mode  :character   Median : 23337.0
##                                        Mean   : 35694.9
##                                        3rd Qu.: 52914.7
##                                        Max.   :896563.2
##   overtime_pay         other_pay         total_benefits
##  Min.   :   -68.01   Min.   : -44591.7   Length:259043
##  1st Qu.:     0.00   1st Qu.:      0.0   Class :character
##  Median :     0.00   Median :      2.1   Mode  :character
##  Mean   :   555.83   Mean   :   4614.8
##  3rd Qu.:     0.00   3rd Qu.:   1389.6
##  Max.   :132950.39   Max.   :2659880.2
##    total_pay         total_pay_benefits       year         notes
##  Min.   :      0.5   Min.   :      0.5   Min.   :2011   Length:259043
##  1st Qu.:   4014.0   1st Qu.:   4014.0   1st Qu.:2011   Class :character
##  Median :  24989.1   Median :  24989.1   Median :2011   Mode  :character
##  Mean   :  40865.5   Mean   :  40865.5   Mean   :2011
##  3rd Qu.:  55881.0   3rd Qu.:  55881.0   3rd Qu.:2011
##  Max.   :2884880.2   Max.   :2884880.2   Max.   :2011
##  jurisdiction_name
##  Length:259043
##  Class :character
##  Mode  :character
##
##
##
```

Wrangling the Data

Looking at the imported data, it appears that the organization changed the column naming scheme in 2012 – this is a minor inconvenience; however, it is relatively easy to change the column names. Furthermore, it appears that another column was added, the status column, that is not present in the earlier data sets. Therefore, we will remove the status column, since it is largely empty, to ensure that all the data sets are roughly similar.

```
# Remove the status column
dat2016 = select(dat2016, -Status)
dat2015 = select(dat2015, -Status)
dat2014 = select(dat2014, -Status)
```

Now, we can make the column names uniform.

```
# Set column names for dat2011 equal to those of dat2016
colnames(dat2011) = colnames(dat2016)
colnames(dat2012) = colnames(dat2016)
```

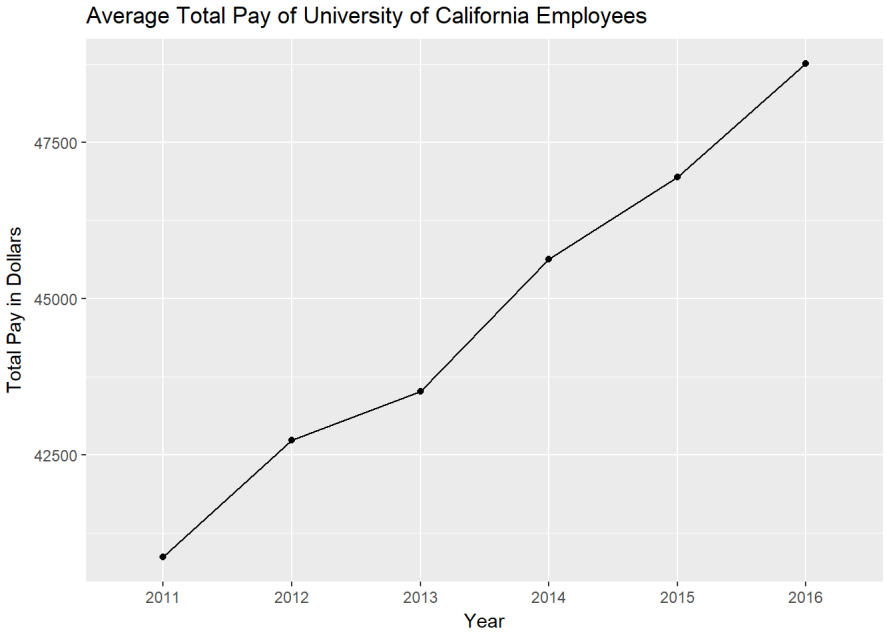
Manipulating and Plotting the Data

Now that we have imported the data and made it uniform, we can take a look at the data sets. To find the mean total pay for each year, we can use the mean command on the column ‘Total Pay’

```
combined_data = data.frame(
  mean = c(mean(dat2011$`Total Pay`),
  mean(dat2012$`Total Pay`),
  mean(dat2013$`Total Pay`),
  mean(dat2014$`Total Pay`),
  mean(dat2015$`Total Pay`),
  mean(dat2016$`Total Pay`)),
  year = c("2011", "2012", "2013", "2014", "2015", "2016")
)
```

Now that we have the means in a data frame, we can use ggplot to graph the data.

```
# Plot the average salary through the years
ggplot(combined_data, aes(x = year, y = mean)) + geom_point() + geom_line(aes(group = 1)) + xlab("Year") + ylab("T
otal Pay in Dollars") + ggtitle("Average Total Pay of University of California Employees")
```



The geom_point command adds the points to the graph, the xlab command adds the xlable “Year”, the ylab command adds the y label “Total Pay in Dollars”, and ggtitle adds the title of the graph. It appears that the average salary has increased in a relatively linear fashion – the points appear to follow a line, more or less. We can add a lowess line to show that.

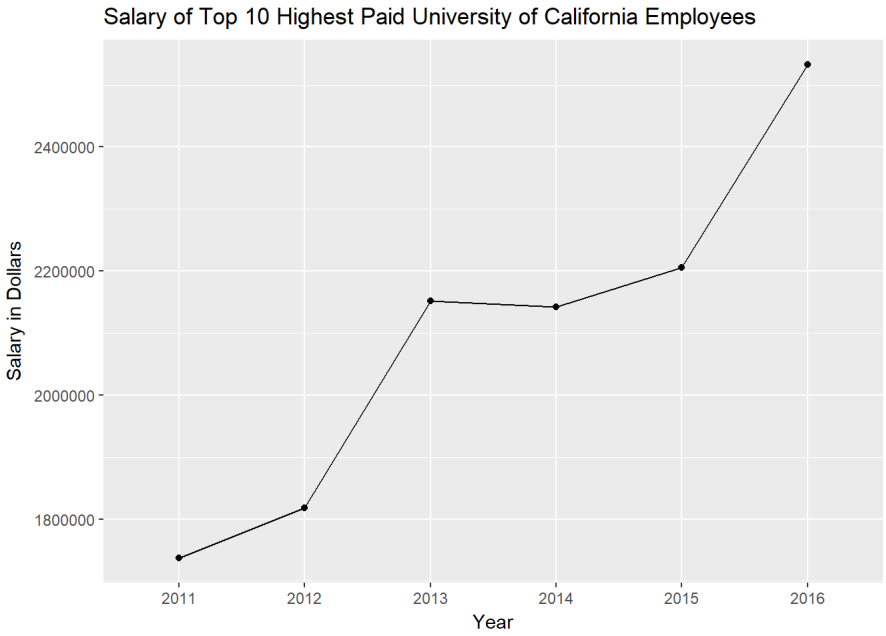
Now that we have a more generalized look at the data, we can take a deeper look into the data. For example, we can look at how the pay of the top-ten paid employees has changed.

```
# Slice the top 10, arranged by Total Pay and Benefits descending
top10_16 = dat2016 %>%
  arrange(desc(`Total Pay & Benefits`)) %>%
  slice(1:10)
top10_15 = dat2015 %>%
  arrange(desc(`Total Pay & Benefits`)) %>%
  slice(1:10)
top10_14 = dat2014 %>%
  arrange(desc(`Total Pay & Benefits`)) %>%
  slice(1:10)
top10_13 = dat2013 %>%
  arrange(desc(`Total Pay & Benefits`)) %>%
  slice(1:10)
top10_12 = dat2012 %>%
  arrange(desc(`Total Pay & Benefits`)) %>%
  slice(1:10)
top10_11 = dat2011 %>%
  arrange(desc(`Total Pay & Benefits`)) %>%
  slice(1:10)

# Add a new column to the combined data frame with the means of the top 10 salaries
combined_data = mutate(combined_data, top10 = c(
  mean(top10_11$`Total Pay & Benefits`),
  mean(top10_12$`Total Pay & Benefits`),
  mean(top10_13$`Total Pay & Benefits`),
  mean(top10_14$`Total Pay & Benefits`),
  mean(top10_15$`Total Pay & Benefits`),
  mean(top10_16$`Total Pay & Benefits`)))
```

Now, we can again use ggplot to create a graph.

```
ggplot(combined_data, aes(x = year, y = top10)) + geom_line(aes(group = 1)) + geom_point() + xlab("Year") + ylab("
Salary in Dollars") + ggtitle("Salary of Top 10 Highest Paid University of California Employees")
```



This graph seems much less linear – there was a huge jump in 2013, and another in 2016. This is probably because we only looked at the top 10 employees; a single massive increase in pay could have a much larger effect given the smaller sample size.

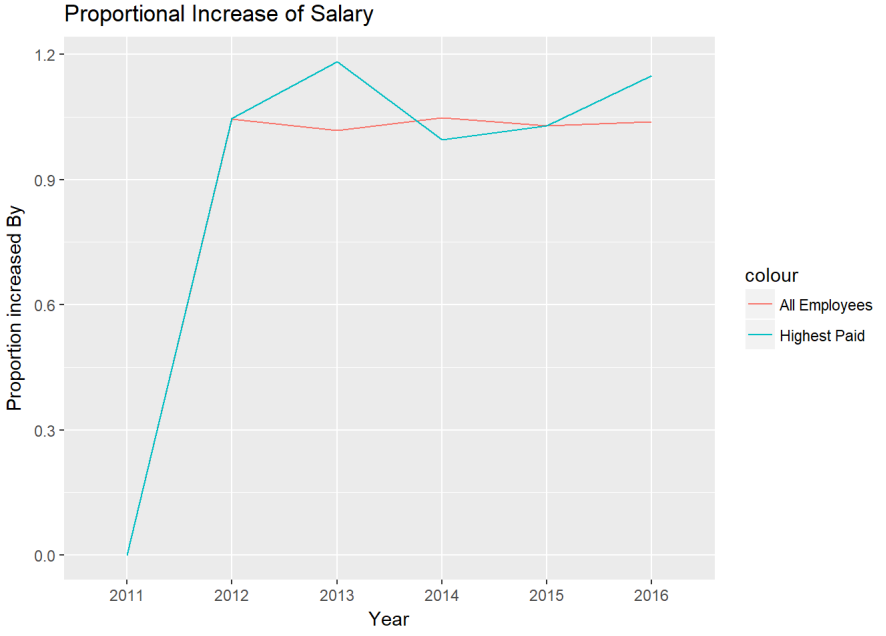
Now, we can look at the proportion that the salary increased by, for all employees and for the top 10 highest paid.

```
### Get the proportion of increase
combined_data = mutate(combined_data,
                       prop_increase = c(0,
                                         combined_data$mean[2] / combined_data$mean[1],
                                         combined_data$mean[3] / combined_data$mean[2],
                                         combined_data$mean[4] / combined_data$mean[3],
                                         combined_data$mean[5] / combined_data$mean[4],
                                         combined_data$mean[6] / combined_data$mean[5]))

combined_data = mutate(combined_data,
                       top10_prop_increase = c(0,
                                               combined_data$top10[2] / combined_data$top10[1],
                                               combined_data$top10[3] / combined_data$top10[2],
                                               combined_data$top10[4] / combined_data$top10[3],
                                               combined_data$top10[5] / combined_data$top10[4],
                                               combined_data$top10[6] / combined_data$top10[5]))
```

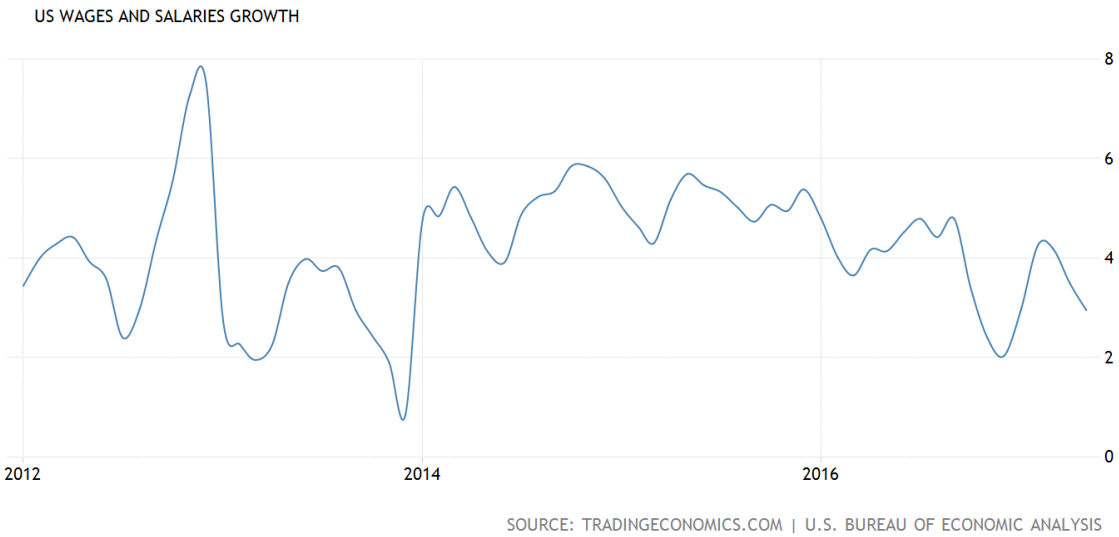
Now that we have the proportion of increase for both, we can use ggplot to create a plot.

```
# Create a graph using the new data
ggplot(combined_data, aes(x = year)) +  geom_line(aes(y = prop_increase, color = "All Employees", group = 1)) + ge
om_line(aes(y = top10_prop_increase, color = "Highest Paid", group = 1)) +  xlab("Year") +  ylab("Proportion incre
ased By") + ggtitle("Proportional Increase of Salary")
```



It appears that the salary of the highest paid employees did not increase at a rate much greater than that of all employees, indicating that the average salary is increasing at a somewhat equal rate for most employes.

Looking at wages and salary growth within the United States for the past 5 years yields an interesting comparison: using data from TradingEconomics.com, the rate of growth can be compared.



It appears that the rate of growth for employees in the University of California system is much lower than that of the average US employee; however, without a more substantial comparison perhaps involving more factors and variables, it is difficult to make such a conclusion.

In Summary

The purpose of this post was to find a way to analyze a number of different data sets, and draw conclusions from them. In particular, I wanted to look at whether or not the highest paid employees were getting paid more and more, relative to the average employee. However, of course, there is still more analysis to be done – I don't possess the math knowledge or background to use more complicated analyses, but I am certain that there exist better, more complex ways to look at the data. The main take-aways from this post as lessons in how to download freely available data sets from the internet, and how to manipulate them into a usable form. Much of the first portion of this post was dedicated to manipulating the data, and the second portion was dedicated to analyzing it.

Much of the data that we have used thus far in Stat 133 has been related to basketball, and it is often packaged in a relatively easy-to-use manner. Here, I explored a data set that was not as easy-to-use, with differing column names and many, many rows. Perhaps the most important thing to take away from this post is the ability of R to be used in a variety of situations – publicly available data sets are available for download, just a google search away, and knowledge of R and various libraries makes it possible to analyze that data and draw conclusions from it.

In conclusion, the purpose of my post was to experiment with manipulating a data set on my own; one that had not been prepared for easy use with R, and to use the data to draw conclusions about salaries within the University of California system. In this, I think I succeeded – while much of my code could likely be simplified, I feel that I was able to manipulate the data and create both visually appealing and informative graphs that helped illustrate my conclusion. Throughout this post, I was able to use various concepts from the class, including data manipulation, wrangling, and visualization. My post deepened my understanding of R and packages like ggplot and dplyr, and I hope that it was informative for any readers as well.

=====

References

"Transparent California." Transparent California, https://transparentcalifornia.com/.

"United States Wages and Salaries Growth 1960-2017 | Data | Chart." United States Wages and Salaries Growth | 1960-2017 | Data | Chart, https://tradingeconomics.com/united-states/wage-growth.

"Read CSV in R with Example of How to Read CSV in R." RProgramming.net, http://rprogramming.net/read-csv-in-r/

"Data Wrangling and Feature Engineering with dplyr" https://www.youtube.com/watch?v=Ds6arVTWwDc

"Beautiful plotting in R" http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/

"Download File from the Internet" https://stat.ethz.ch/R-manual/R-devel/library/utils/html/download.file.html

"R - Data Frames" https://www.tutorialspoint.com/r/r_data_frames.htm