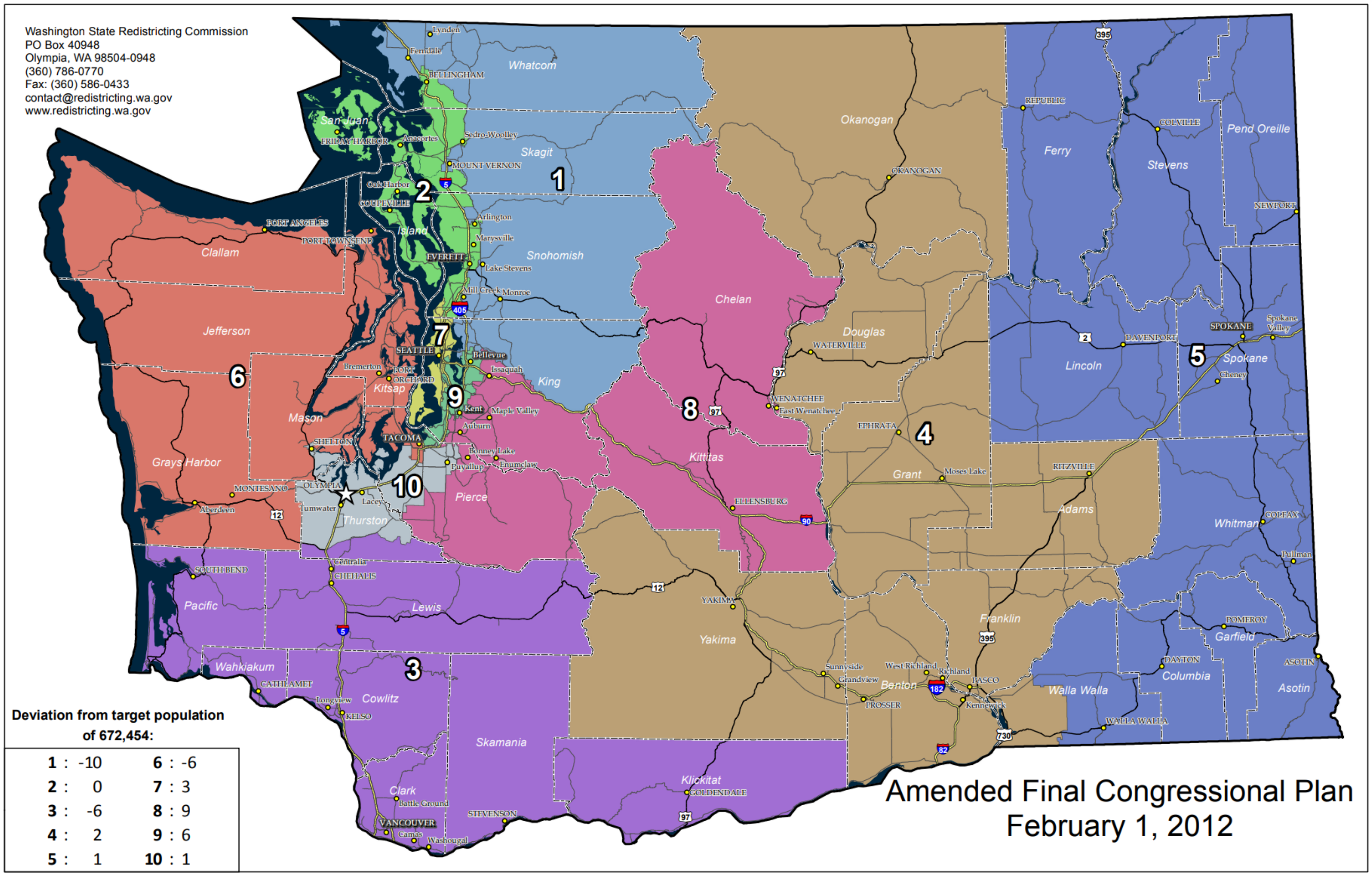
Team 1

TBANLT 540: Applied Regression Models – Final Project

Abdullah Mirreh
Rahul Kumar
Viviana Ramirez
Bibek Mahal
Randall Plyler

Autumn Quarter 2021


Washington State Redistricting Commission
PO Box 40948
Olympia, WA 98504-0948
(360) 786-0770
Fax: (360) 586-0433
contact@redistricting.wa.gov
www.redistricting.wa.gov

Deviation from target population of 672,454:

1 :	-10	6 :	-6
2 :	0	7 :	3
3 :	-6	8 :	9
4 :	2	9 :	6
5 :	1	10 :	1

Amended Final Congressional Plan
February 1, 2012

Map provided by https://www.redistricting.wa.gov/

Contents

Abstract

Our paper uses the data related to State of Washington Employee salaries across all the colleges and universities. Our exploration is set out to understand the impact of salaries based on the location (i.e., congressional district) thereby allowing the state of Washington employees and their prospects to make an educated decision based of the existing data. By regressing the data, the findings indicate that the Washington State Congressional Districts 7, 8 and 6 in the order are the highest paid.

Introduction

In a college and/or university system, there are several different kinds of jobs and employment opportunities offered with various levels of experience. Outside of experience and job title, there factors that impact college and university employee salaries. One of the factors that impact salaries of employees is the location of the college and university. In the state of Washington, there are thirty-six different public community colleges and universities. Public community colleges and universities are scattered across different cities in the state, as well as ten congressional districts.

The article, "Visualizing the Higher Education Industry," from NewAmerica.org highlights the parallels drawn from funding to congressional school districts. "High employment and huge expenditures are complemented by a wide geographic reach: There are over 16 colleges in the average Congressional district, and there are no Congressional districts with fewer than four schools. By sector, these averages break down to about four public colleges and universities, eight for-profits, and four private nonprofits for every Congressional district. On average, schools in each district employ over 6,000 full time faculty, staff, and administrators and spend over a billion dollars on total operating costs," (Dancy & Laitinen).

Our analysis will dive into understanding the differences in salaries of teachers and educators within separate congressional districts. Often in states where there are multiple congressional districts, various factors impact the funding, salaries, and operation of schools. It's with our analysis we will draw whether these factors impact salaries and are correlated.

Data Set

The data set utilized for our analysis is from data.wa.gov, "Washington State – state worker salaries". Navigation to the dataset by following; from the main page, click on "Data Catalog," expand categories, and select "Labor." Under view types select "Datasets," and under tags select "salary." Result after these filters is Annual Salary 2010 thru 2013. Once clicking on the dataset, an option to extract as a csv is presented. Download the csv.

The selected dataset contains 339,763 observations (rows) of 8 variables (columns) with following headers and characteristics:

- **Agency:** numeric data type containing agency code
- **AgencyTitle:** character data type
- **EmployeeName:** character data type

- **JobTitle:** character data type containing job names for every employee thru the years
- **Salary2010:** numeric data type containing the salary for every employee on year 2010
- **Salary2011:** numeric data type containing the salary for every employee on year 2011
- **Salary2012:** numeric data type containing the salary for every employee on year 2012
- **Salary2013:** numeric data type containing the salary for every employee on year 2013

To analyze this data, we performed cleaning steps utilizing R studio. The first step was to unpivot all four columns containing SalaryYear. For this, the pivot_longer function was used. Second step was to remove "salary" string from newly unpivot data using mutate and substring functions. Those two steps allowed us to get a single salary per year per employee. After these steps our dataset contained 1,359,052 observations. Throughout the rest of the assignment, we chose to only handle five variables seen below:

- **AgencyTitle:** chr data type
- **EmployeeName:** chr data type
- **JobTitle:** chr data type
- **Year:** chr data type
- **Salary:** num data type

Because we wanted to analyze University and College employees' salaries on different Congressional Districts, we went and separated the data into Washington State Congressional Districts. For this, we first utilized the internet to search information on congressional districts. Initially we needed to know how many public universities and colleges were in the State of Washington. While this information could be found in the dataset provided by the fiscal > Washington State, we felt it was much more simplistic if we just searched up a list of all Washington State public higher education institutions.

Once we gained such a list, we then searched up a map of Washington State that is color coded by Congressional Districts. The map was essential for our exercise to get an understanding of where known universities and colleges are located. As for the remaining universities and colleges, we gathered their zip codes and utilized "Find my US Representative" via Congress's website. Using an Excel sheet, we searched up each university and college's zip code on "Find my Representative" website and that allowed us to pinpoint the US Congressional District of such university or college in Washington State. Once we populated the Excel sheet, we were ready to conduct our analysis via R.

Analysis Plan

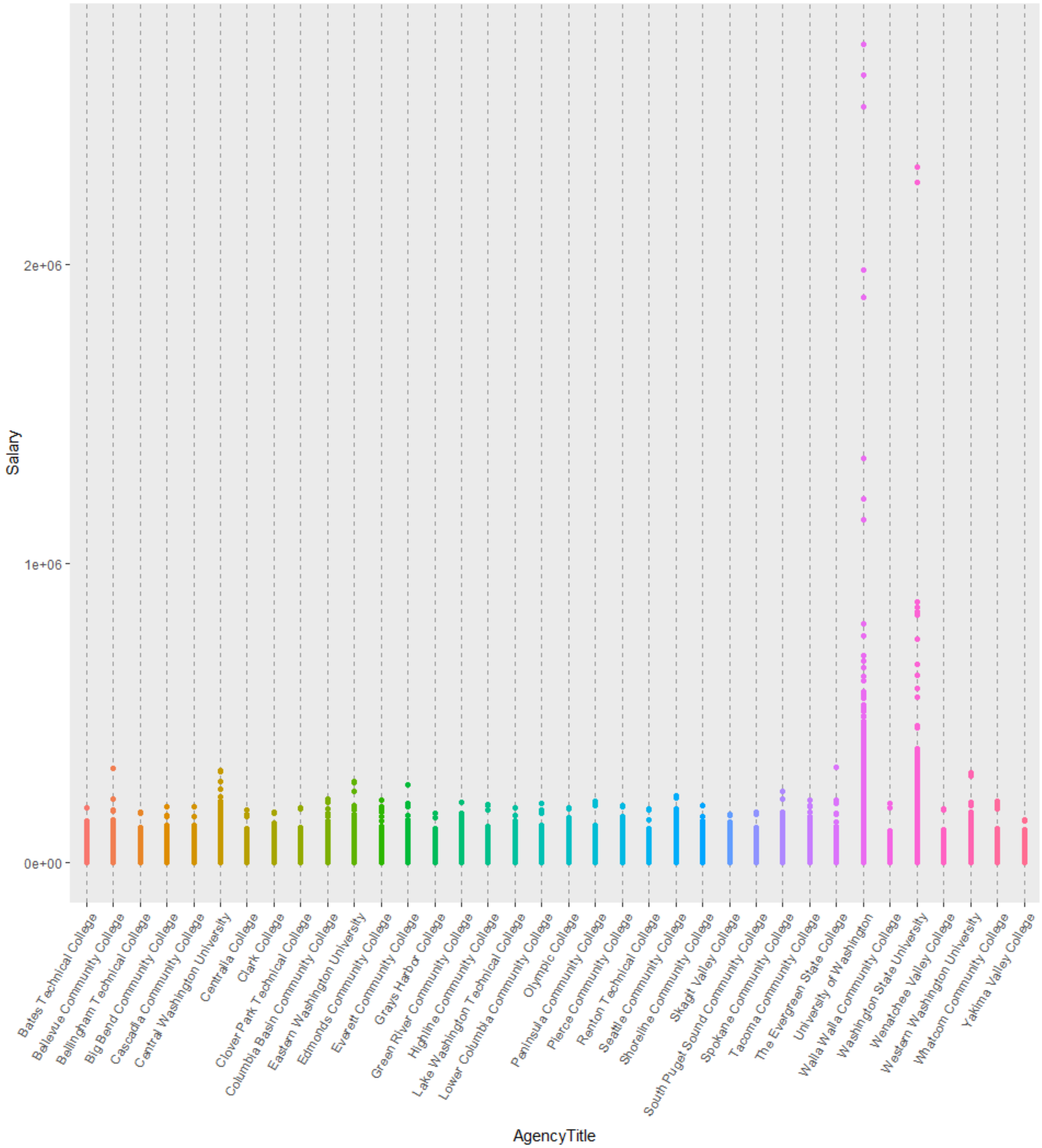
For the duration of this exercise, the group choose to utilize categorical Linear regression. The choice for the exploration question we decided was, "Is there an impact on salary for all the colleges based on congressional district in the state of Washington?" A plan for obtaining the solution is as follows:

Clean the data to transform to present for data analysis.

- Filter the data from the dataset specific to colleges and universities.
- Identify colleges and map them into respective congressional districts for grouping thereby establishing categorical classification of 10 districts.
- Plot the data and identified outliers i.e., University of Washington (UW), Washington State University (WSU). These will be included in the model to ensure completeness.
- Identify the reference factor i.e., the Congressional District 7 i.e., Seattle area.

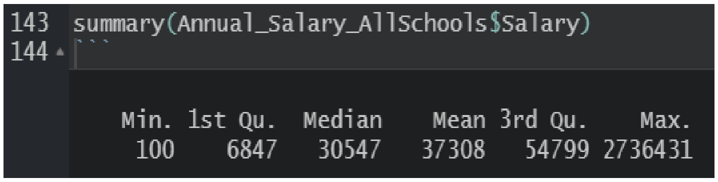
Results

Once the data cleaning phases have been complete, plots were generated utilizing ggplot to provide a visual overview of the data. Below you can see a scatterplot of salaries with their respective agencies.



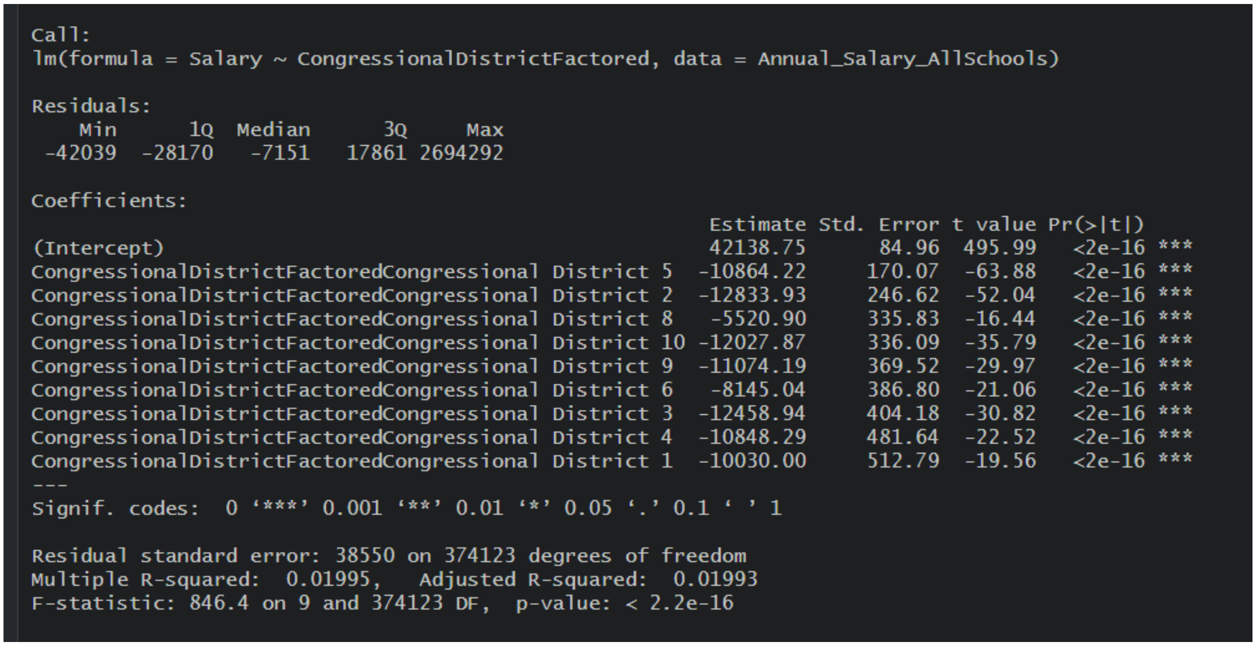
The visual above highlights the challenges that are already present with the data. Datasets with categorical variables and a large skewness often show a low r^2. For example, when you plot a linear, exponential, or polynomial fit line of the chart, it often would lie flat or very “unfit”. This is because the correlation between the variables is very low, and we will explore that more in a bit.

One more interesting point when analyzing the graph is the outlier points on “University of Washington” and “Washington State University”. After a brief analyzing of the data, these outlier points are those of football coaches at the two largest universities in the state. The salaries of these positions are more than $2 million, far outweighing the average $37,308 salary of all state school workers. The summary statistics for all school salaries can be seen above.



```
summary(Annual_Salary_AllSchools$Salary)
```
```

    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     100    6847   30547   37308   54799 2736431 
```

After creation of categorical variables, the team created three Linear Regression models to analyze the data. The first one detailing all the congressional districts with colleges and universities, the second with just congressional districts, and the third with just colleges and universities. For the purposes of the assignment and analyzing the impact on congressional districts and salaries, the second model *lm(formula = Salary ~ CongressionalDistrictFactored, data = Annual_Salary_AllSchools)* will be used.

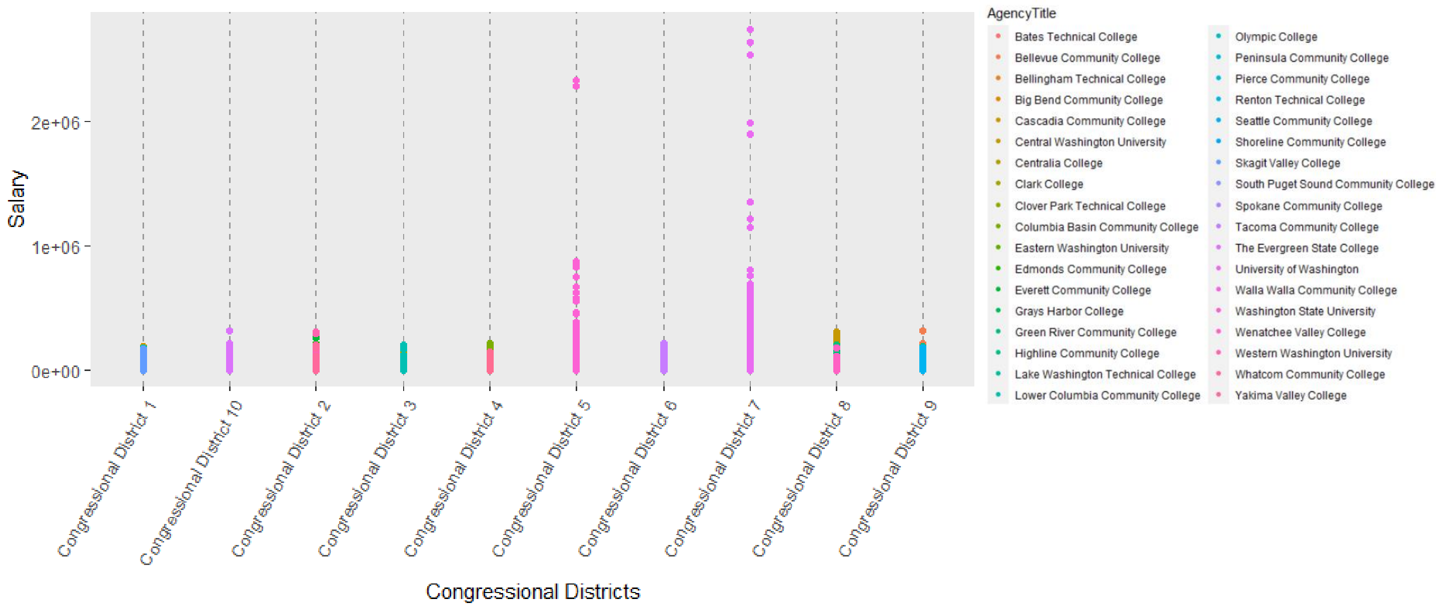


```
Call:
lm(formula = Salary ~ CongressionalDistrictFactored, data = Annual_Salary_AllSchools)

Residuals:
    Min      1Q  Median      3Q     Max 
 -42039  -28170   -7151   17861 2694292 

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                             42138.75      84.96  495.99   <2e-16 ***
CongressionalDistrictFactoredCongressional District 5  -10864.22     170.07  -63.88   <2e-16 ***
CongressionalDistrictFactoredCongressional District 2  -12833.93     246.62  -52.04   <2e-16 ***
CongressionalDistrictFactoredCongressional District 8   -5520.90     335.83  -16.44   <2e-16 ***
CongressionalDistrictFactoredCongressional District 10 -12027.87     336.09  -35.79   <2e-16 ***
CongressionalDistrictFactoredCongressional District 9  -11074.19     369.52  -29.97   <2e-16 ***
CongressionalDistrictFactoredCongressional District 6   -8145.04     386.80  -21.06   <2e-16 ***
CongressionalDistrictFactoredCongressional District 3  -12458.94     404.18  -30.82   <2e-16 ***
CongressionalDistrictFactoredCongressional District 4  -10848.29     481.64  -22.52   <2e-16 ***
CongressionalDistrictFactoredCongressional District 1  -10030.00     512.79  -19.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38550 on 374123 degrees of freedom
Multiple R-squared:  0.01995,	Adjusted R-squared:  0.01993 
F-statistic: 846.4 on 9 and 374123 DF,  p-value: < 2.2e-16
```

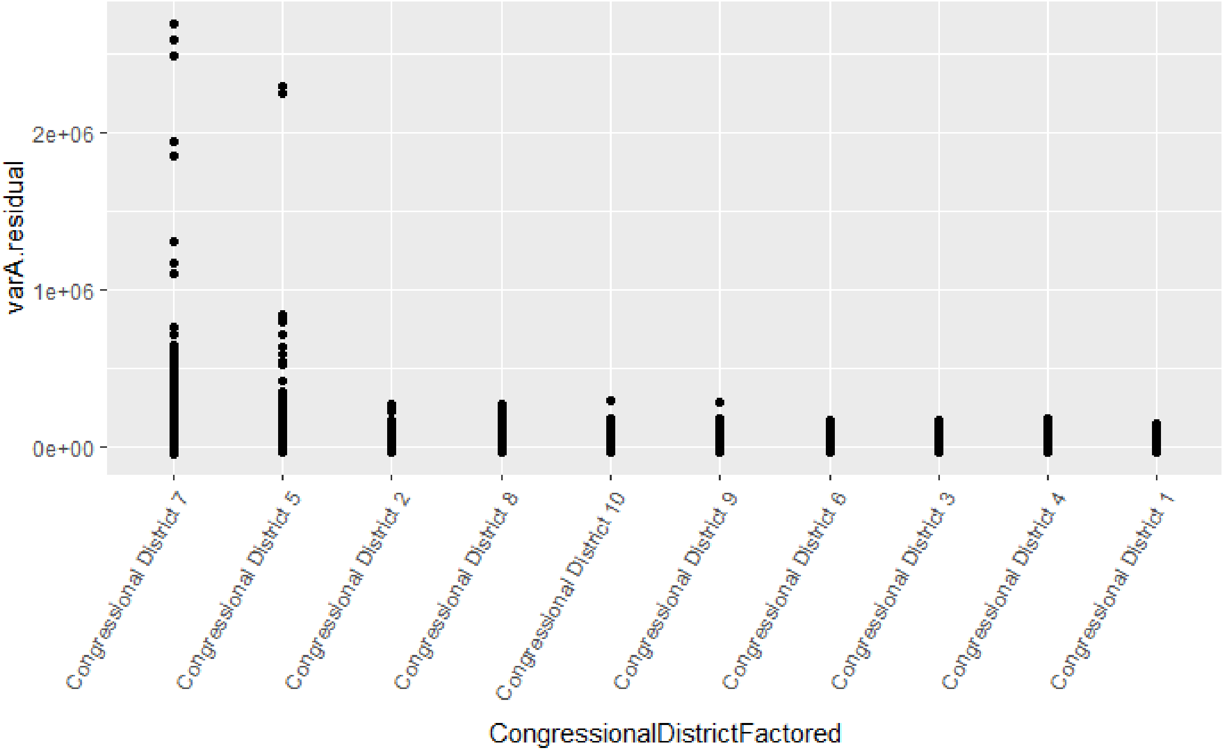
As seen above, the r^2 and adjusted r^2 are very low which indicate a low correlation. Each of the congressional districts show they are very significant which is a great sign in predicting their impact on the intercept estimate. As shown above, the Intercept which we set as “CongressionalDistrictFactoredCongressional District 7” which is the Greater Seattle Region, shows the highest salary for congressional district. This would make sense as the Seattle Region has the largest population density, largest tax structure, and largest economic impact in the region.

Interpreting the data further, we can see that all the remaining districts have predicted less salaries, while some have more than others. The worst district salary relationship shows that congressional district 2 comes in last, with an estimated salary of $29,255 ($42,183-$12,833). If we look at the map on the front page, we can see that congressional district 2 encompasses the Oak harbor, Camano Island region of Washington.

Drawing an inference from the data, if a state school worker wanted to make the largest salary centered on our limited examination (based on all salaries), the person would want to work in the 2nd or 7th congressional districts. The 8th congressional district encompassing the central part of Washington including Renton, Kent, and Federal Way. This would be the third largest locations of high salaries.



Above is another visual utilizing ggplot that details all the schools and universities placed in their respective congressional district. Note that congressional district 7 and 5 both have outliers that give them a strong weight in the regression analysis. The ggplot visual also details certain congressional districts have overall higher salaries. What the data does not tell is the impact of what job has on salary. An example is that an English professor may make $30,000 and a football coach makes $2 million annually. This would have a large impact on the overall skewness of the regression. For this, the R^2 is not a great predictor of determining salary.



Residual plot surfaces the margin of error variance compared across all the congressional districts. Congressional District 7 shows higher residual which is possible given the amount of variation in the salaries of which UW is part of.

Limitations of Study and Conclusion

Throughout the process there were a few limitations worth mentioning. A low R^2 is justified as there is not enough variables explain the nature of the job that drive the salary, although the relative variation of salaries across all the congressional districts is determined. Performing timeseries analysis would have been possible if there were enough years, however the dataset only had four years which is not sufficient historical data to predict the salary increase year over year. For future analysis, our recommendation would be to analyze the different career professions to better understand the impact on salary.

To summarize our paper, there is a relationship between salary and schools in congressional districts. Though the difference between the districts is subtle, there is enough evidence to support a pay difference between the congressional school districts. The downside to our equation is that many factors contribute to salaries that cannot be quantified, however, we can draw a conclusion based on evidence that various congressional districts compensate state school workers differently.

Works Cited

Dancy, Kim and Amy Laitinen. *Visualizing the Higher Education Industry*. 14 October 2015. Article. 10 December 2021. <https://www.newamerica.org/education-policy/edcentral/the-higher-education-industry/ >.

Representatives, United States House of. *Find Your Representative*. 10 December 2021. Website. 10 December 2021. <https://www.house.gov/representatives/find-your-representative >.

State of Washington. *Annual Salary 2010 through 2013*. 22 January 2015. Dataset. 10 November 2021. <https://data.wa.gov/browse?category=Labor&limitTo=datasets&tags=salary>.

Washington State. *Draw Your WA*. 10 December 2021. 10 December 2021. <https://www.redistricting.wa.gov/>.