David Wang

MGMT-6962

Analysis of industry tax rate, compensation and dividend

With the country nearing midterm election, we frequently hear in the news cycle about giant corporations paying little to no taxes. It is important to note that the tax rate referred throughout this project is the total taxes paid, not just the commonly noted corporate tax rate. This peaked my interest as to how well employees at these corporations are paid, and whether the same cost saving approach is taken in compensating their workers. The data-set also comes with a few extra factors we can use, with this we hope to see if we can predict the dividends paid by each industry. The initial conjecture coming into this project is that there is a positive relationship between industry tax rate and industry wages, and as for the second model we believe there is a high correlation between our dependent variable; dividends paid and our independent variables; tax rate per industry, undistributed profits per industry, and employee compensation per industry.

The data set used comes from the bureau of economic analysis(BEA) which is a branch of the department of Commerce. The BEA is tasked to provide official macroeconomic and industry statistics. The data set we are concerned with is the National Income and Product Accounts. The data set comes in eight excel documents(BEA refers to these excel documents as sections) each containing multiple tables, each excel document pertains to a specific topic such as, GDP, personal income/spending, and industry statistics. For this research project we will be using the data provided in section-1 and section-6 excel sheets. Section-1 contains CPI data which we will need to adjust our fiscal data for inflation. Section 6 contains industry statistics

such as persons engaged in production by industry, wages and salaries by industry, net interest by industry, etc. The data provided by the BEA consists entirely of quantitative data, spanning across more than 60 distinct industries from 1929 to the present.

The data we are using spans across 90 years and during that time many new industries have sprung up and disappeared during that time, before we can use our data-set we need to standardize the industries we are using across the entire time period of the data-set. To achieve this requires quite a bit of work, the dataset provided by the BEA does not contain the footnotes describing the changes and additions made to each industry, thus we need to cross reference with the annual report submitted for that year. The following changes thus need to be made.

For the mining sector, the most recent report shows this sector being condensed down to three industries; “oil and gas extraction”, “ mining except oil and gas”, and “support activities for mining”, this means that for the years 1929 to 1948 we need to combine “anthracite mining”, “bituminous soft coal mining”, “nonmetallic mining and quarrying” down into a single industry and rename to “mining except oil and gas”, the leftover “crude petroleum and natural gas” then needs to renamed to “oil and gas extraction”. The years spanning from 1948 to 2000 “metal mining”, “coal mining”, and “nonmetallic minerals except fuels” needs to be combined into “mining except oil and gas”, and “oil and gas extraction” can be left as is.

The “Transportation and public utilities” sector from 1929 to 1948 contained the “transportation industry”, “communications and broadcasting industry”, and “electric, gas and sanitary services industry”. From the year 2000 and forward “electric, gas and sanitary services” was extracted and made its own sector, the “communications and broadcasting industry” was also extracted but placed in a new sector named “Information”. The “transportation industry”

was then combined with the new “Storage industry” making up a new sector called “Transportation and storage”. To standardized this, for the time period 1929 to 1948 we need to rename “communications and broadcasting industry” to “Utilities” and remove the “Storage industry” from 2000 and onward since it didn't exist from 1929 to 1999.

The wood products industry before the year 2000 was split into two seperate industries, “lumber and wood products” and “furniture and fixtures”, thus we need to combine the two industries and rename it to “wood products”.

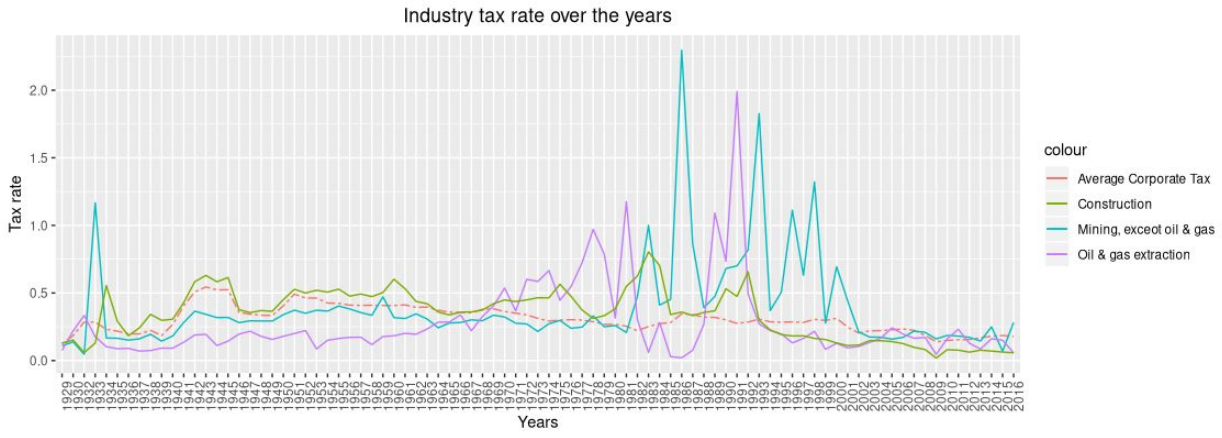
The manufacturing sector contains a wide range of industries categorized as either durable goods or nondurable goods. First taking a look at the durable goods category, we see that the only changes that need to be made are those relating to that of the wood industry. From 1929 to 2000 the now consolidated wood industry was split into two industries, “Lumbar and wood products”, “furniture and fixtures”. For the nondurable goods category, we need to combine “Food and kindred products” and “Tobacco products” into “Food and beverage and tobacco products”, “Apparel and other textile products” and “Leather and leather products” into “Apparel and leather and allied products” for the years 1929 to 2000.

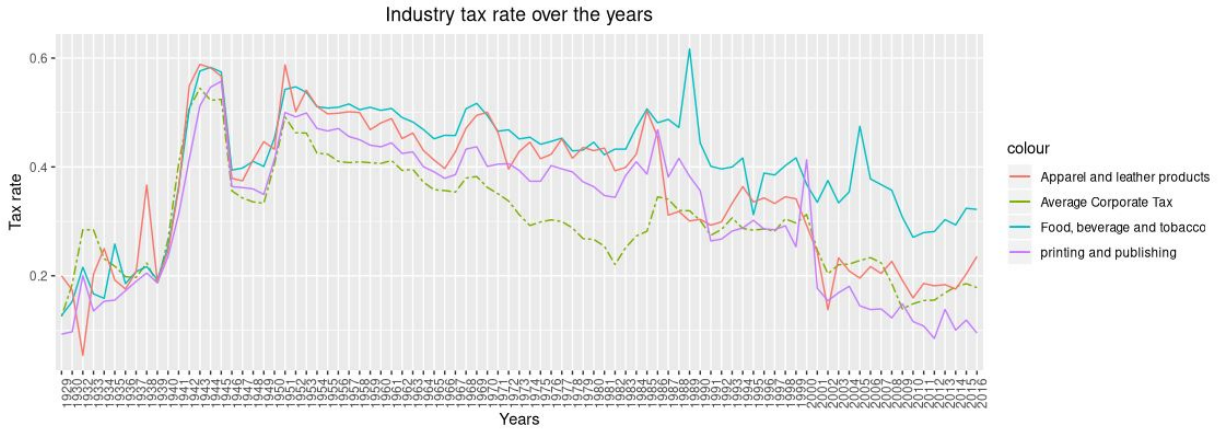
The Finance and insurance sector requires only one small change, for the time period 1929 to 2000, we need to combine ‘Insurance carriers” and “insurance agents, brokers, and service” into “Insurance carriers and related activities”.

After standardizing our industries, we still need to adjust the monetary data for inflation so that we can compare dollar amounts on the same scale. To adjust for inflation we can use the Consumer Price Index (CPI) provided in the data-set, CPI is a measurement that examines the weighted average of prices of a basket of consumer goods and services (e.g transportation, food,

medical care). We can use the following formula to calculate the inflation multiplier ((B – A)/A) + 1, where A is the starting CPI and B is the ending CPI. Using the above formula we can construct a matrix of inflation multiplier. Multiplying our data with the inflation matrix, we now have a inflation adjusted data-set.

Now that we are done with the data wrangling, we can begin our analysis of data. First we want to calculate and examine the tax rate of each industry, the data-set does not provide a direct tax rate that we can use, but it does provide data for profit before tax and profit after tax, with R's dataframe arithmetic abilities we can easily calculate what the tax rates are.

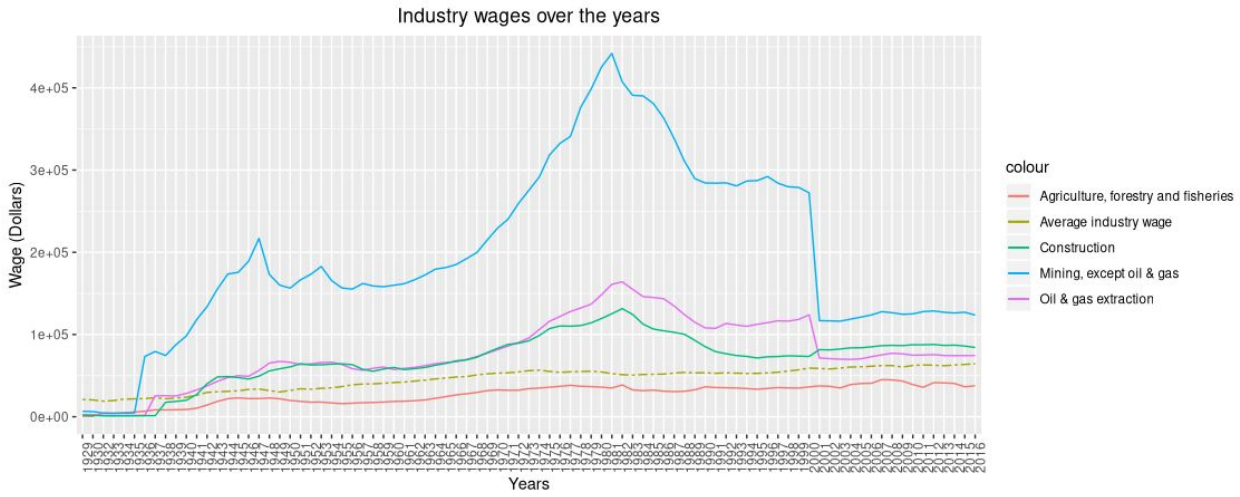


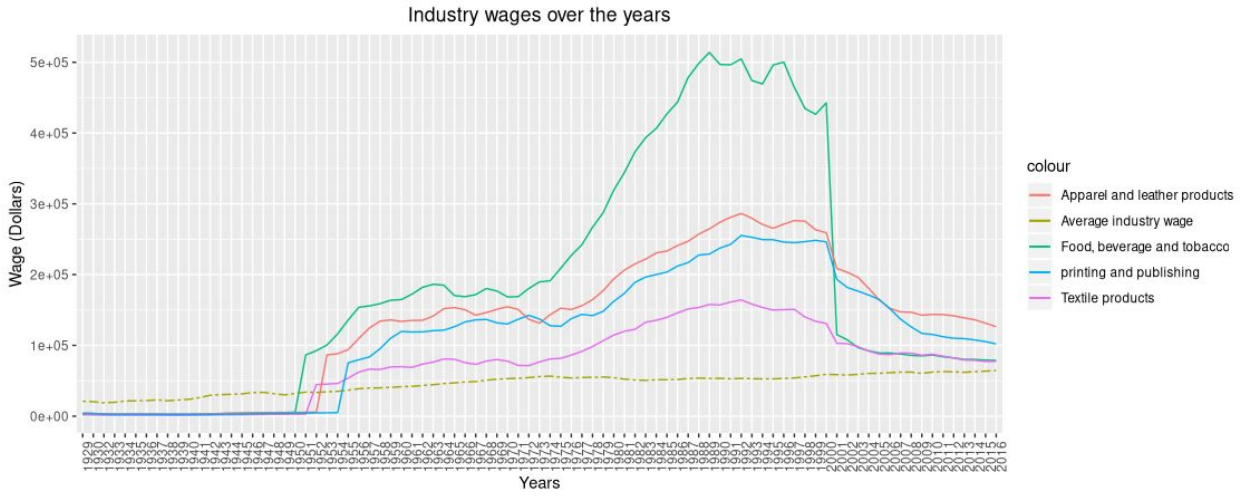


The above charts show the tax rate for a few selected industries, the dotted line in both graphs shows the average tax rate across all industries. We note the variance in taxes paid across the industries, this can mostly be explained by their business model which determines how much labor-related taxes and excise taxes are paid, these taxes are paid on top of the regular corporate tax. Overserving the trend of the dotted line we can see that it would appear the general trend for industry tax rates since 1985 has been on a steady decline. We also note that there are multiple major spikes throughout the graph, these spikes can be explained by either global economic events or tax reforms. For example in 1936 Revenue Act of 1936 was passed which imposed additional taxes on undistributed profits tax, another example is the energy crisis during the 1970's followed by the oil glut in the late 80's.

Some concerns were had regarding the tax rate while researching for this project, often times there is a lag period before you can really see effects on tax rates. Things such as tax shields, tax credits and tax deferrals can average out the effects of their finances across a period of time, thus tax paid for the current period might be deferred taxes that were owed from 2 to 3 years back or it could be a reduced tax via either a tax credit or a tax shield.

The other variable we want to look at is employee wages per industry.



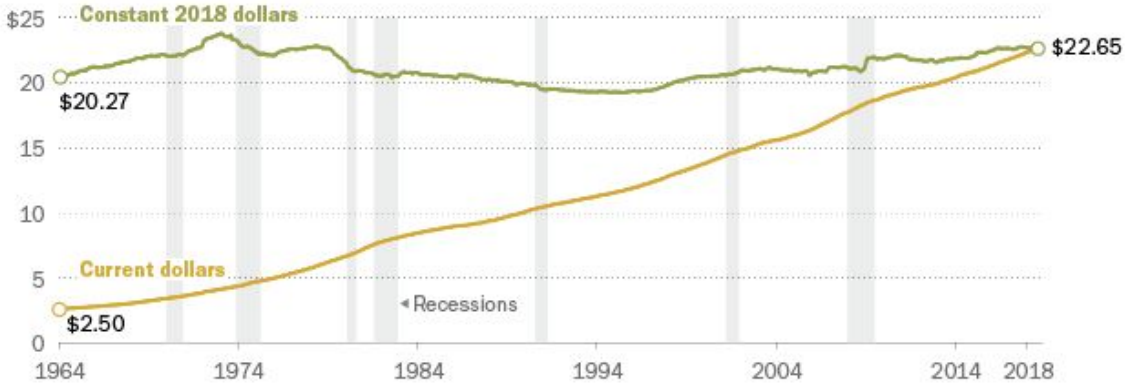


We see the greatest increase in wages and purchasing power happening between 1970's through the 1980's and once we reach the millennia we saw a huge decrease in purchasing power of employees, back to the same level as that of 1960's almost 40 years ago. At first i thought this

was a calculation error on my part, but after further research it seems that this is the case

Americans' paychecks are bigger than 40 years ago, but their purchasing power has hardly budged

Average hourly wages in the U.S., seasonally adjusted



Note: Data for wages of production and non-supervisory employees on private non-farm payrolls. "Constant 2018 dollars" describes wages adjusted for inflation. "Current dollars" describes wages reported in the value of the currency when received. "Purchasing power" refers to the amount of goods or services that can be bought per unit of currency.
Source: U.S. Bureau of Labor Statistics.

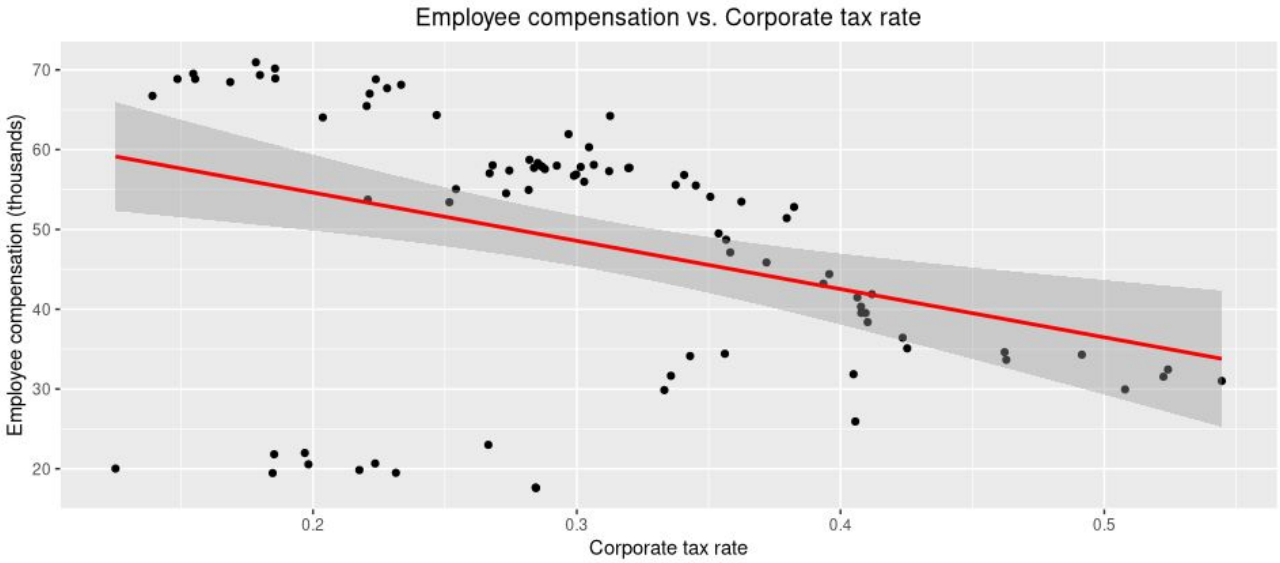
PEW RESEARCH CENTER

The above graphic is taken from a pew research report. It shows that the purchasing power in 2018 is the same as that of 1964.

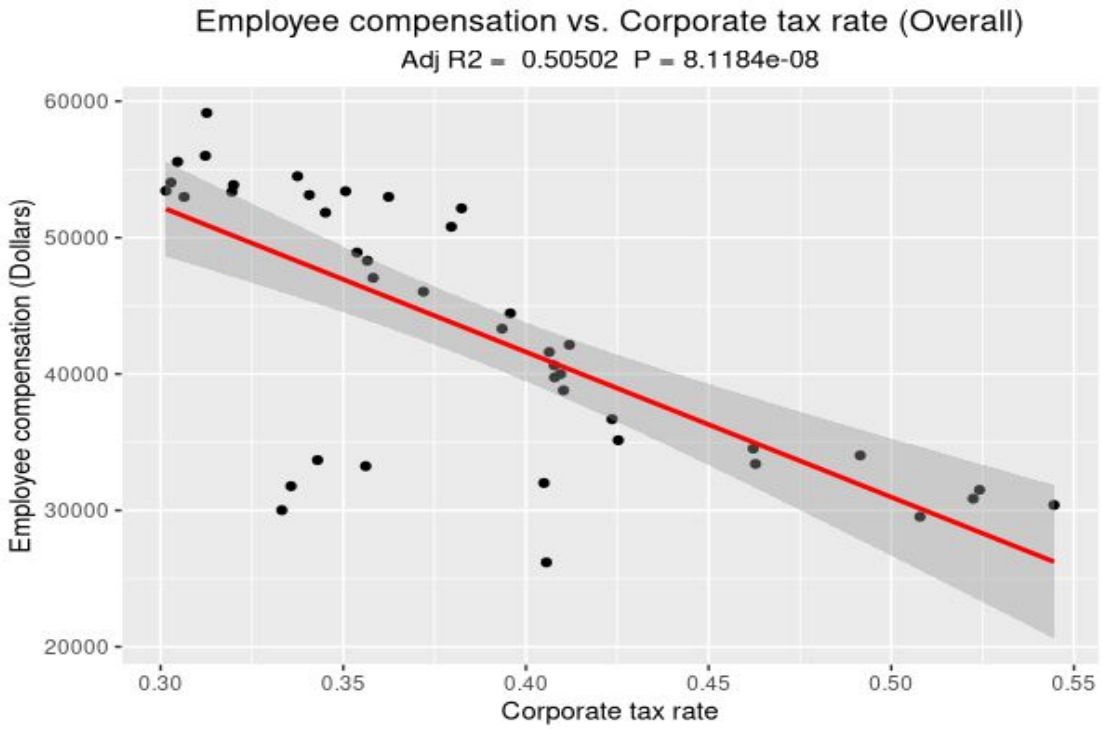
Moving on to modeling our data, the first relationship we wish to model is between the two variables industry tax rate and industry wages. Since our goal to predict one variable with the other a linear regression would make the most sense here. After organizing our data we can use a for loop to generate linear regression models for all 42 final industries.

The first graph generated shows the linear regression on the overall employee compensation and overall corporate tax rate, we see along the bottom there is a group of 10 or so outliers. The p-value for the below graph is 2.17e-8 which is low enough to reject the null hypothesis and a

adjusted R-squared value or 0.23, which means that the fitted line only explains about 26% of the data points.

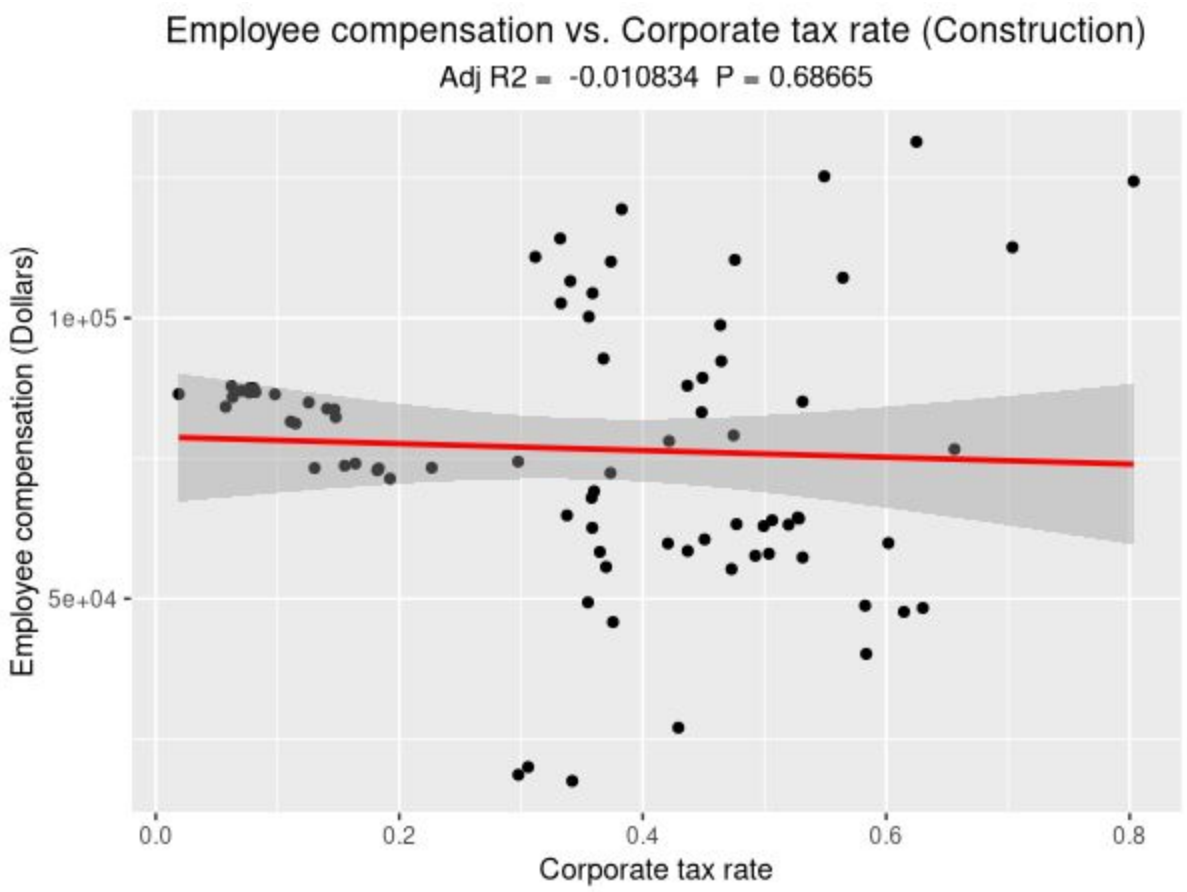


The next plot generated removes the bottom outliers, we should make note that these outliers occur in periods of economic instabilities, for example about half of the outliers occur during the great depression era, with one during 1978 and the rest during the 2008 recession. We see that with the removal of the outliers, this improved our p-value to 8.11e-8 and drastically improved our adjusted R-squared value to 0.505 which is about double that of our previous value.



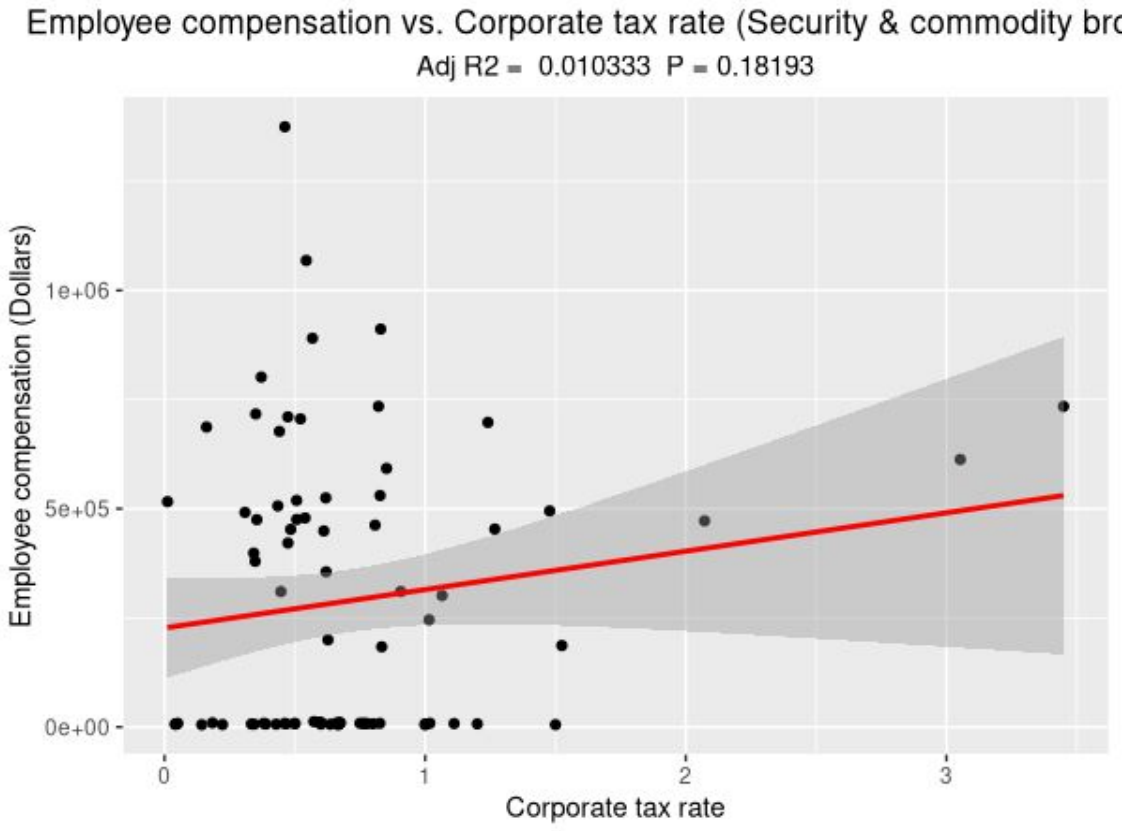
We should note that just because there exists outliers does not mean the removal of them are justified. The two graphs are shown to illustrate the difference between removal of the outlier and keeping the outlier. What we can take away from this is that during stable economic periods there is a clear linear relationship between industry tax rate and industry employee wages, but this relationship seems to break down during turbulent economic periods, thus we might possibly use this model to predict oncoming recessions.

Although the majority of the industries exhibit the same linear relationship similar to that of the overall trend shown above, there are a couple outlier industries where the linear relationship completely breaks down or becomes weak enough to be negligible . The first example is the construction industry, the generated graph shows what seems to be an almost horizontal line indicating that as the tax rate varies up and down the employee does not seem to be affected that much.



The linear regression for the construction industry shows a p-value of 0.686 which is extremely high, meaning we must accept the null hypothesis wherein there exist no relationship between construction industry tax rate and construction industry wages, our adjusted R-squared value returns -0.0108, in practice the lowest R-squared should return is a zero, since at that point we can do no better than just using the mean value. Which means that if the regression line performs worse than using the mean value, the r-squared value we calculate will be negative. Possible explanation for this phenomena could be due to the heavy unionization of this particular industry, such that any wage fluctuation due to tax rates would be protected by union bargained deals.

The second industry that exhibits the breakdown of the linear relationship, is the security and commodity brokerage industry. The linear regression plot generated below display what seems like a vertical relationship, the range of wages are spread equally across all tax rates, this would suggest that wages are not impacted by the effects of industry tax rates. These findings are backed up by the linear regressions's statistical values, the p-value for the regression is 0.182 which means that the null hypothesis stands true. The adjusted r-squared value is 0.0103 meaning only 1% of variables are explained by the model, but considering the 3 right most outlier points, if they were removed the adjusted r-squared value would most likely lie in the negatives.

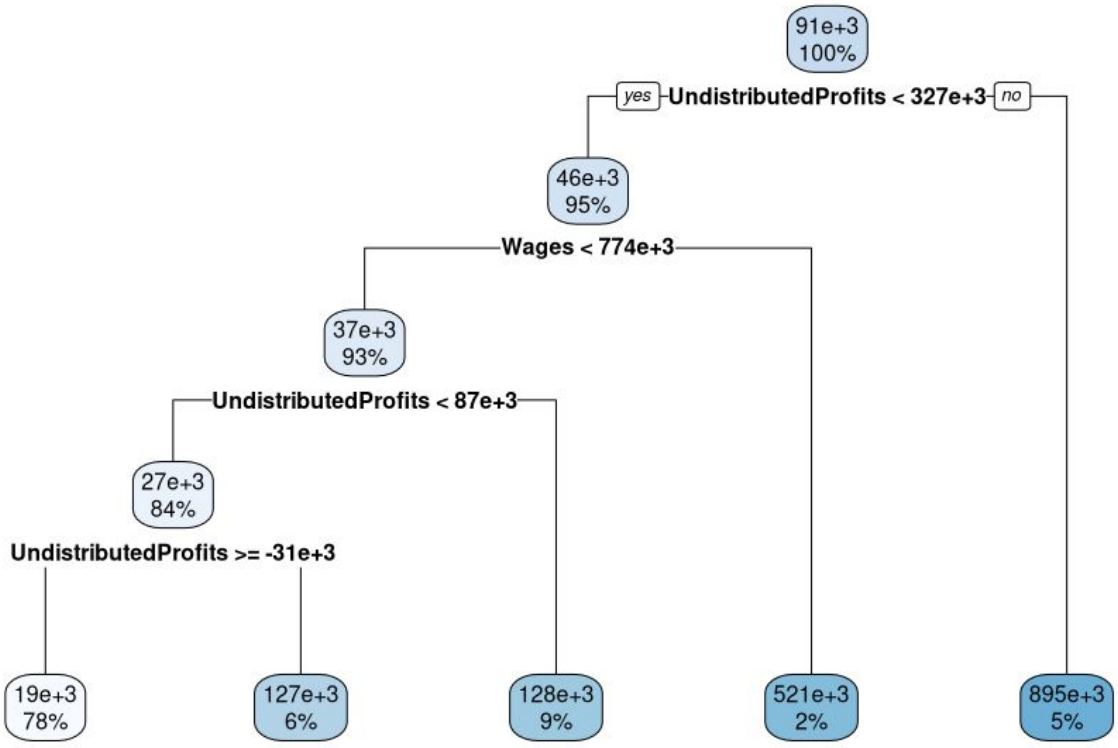


Using the linear regression model we created, we can try to predict the overall tax rate given the the average US worker wage. The average wage we will use is the quarterly news

release by the Bureau of Labor Statistics, which reported a median wage of $47,060 per year for the first quarter of 2019. Using the predict function results in a tax rate of 36.17%, this figure would be within ±1.0% if it were not for the Tax Cuts and Jobs Act which slashed corporate taxes down to 21%. The 21% tax rate would place us in the outlier category, if we were to assume that the outlier hypothesis we came up with earlier is correct, this would suggest that we are near an economic recession. This hypothesis is somewhat backed up by the current news cycle, with the yield curve inversion, The ISM Manufacturing Index close to dipping below 50 (above 50 indicate growing sector ; below 50 means shrinking), and the trade war dragging on, a recession definitely seems like a possibility.

For the next model we want to see if we can predict the dividends paid out by each industry, given their tax rate, employee compensation and undistributed profits. The model chosen for this was a decision tree. There are in general two types of decision trees, categorical variable decision trees, and continuous variable decision trees, we are more concerned with the latter variety. Tree based learning algorithms are perhaps one the most popular supervised learning methods, since they are easy to understand and provide an excellent visualization on model behavior, and second unlike linear models, tree based methods are able to map nonlinear relationships quite well, meaning that any non-linearity between factors do not affect the performance of the tree. Of course there also exists some downsides to using a decision tree. Decision trees are very prone to overfitting, this happens when the algorithm generates an overly complex tree that becomes too defined, a technique called pruning can mitigate this to some extent. Another downside of a decision tree is that when working with continuous numerical variables, there is information loss as the tree categorizes the variables into different classes.

Below is the decision tree generated, we see that given the three factors only two were used in the generated tree. The most significant factor appears to be undistributed profits, this finding makes sense since the dividends paid out comes from undistributed profits.



To estimate how our model performs, we can use k-fold cross validation, where k = 10. This involves fitting the same model multiple times using a different selection of our data-set. The accuracy result of our model was quite low coming in at only 30.82%. This would suggest that there does not exist a strong correlation between the three factors used to predict dividend paid per industry.

To sum up the results of this project, with outlier points removed we found a strong relationship between tax rate per industry and employee wages per industry, we also found that

these outlier points appear during times of economic turbulence. Using this model we tried to predict the overall industry tax rate given 2019's first quarter average income, the result suggested a 36% tax rate, we found this we be within 1% accuracy if it were not for the Tax Cuts and Jobs Act which slashed corporate taxes down to 21%. If our initial hypothesis linking turbulent economy with outliers were correct this would suggest the bearing of an economic recession. The second model we attempted to predict dividends paid per industry, using tax rates, undistributed profits, and employee wages as variables, the regression tree showed poor results, indicating that the relationship is weak.

During the project, we often found that the data points we have access to is not enough, macroeconomic trends spans multiple years, and any economic events can take years to show up in the fiscal data. At the start of the project 89 data points across 89 years seemed like an adequate amount of data, but in actuality when adjusted for outliers it's quite a low number, improvements on this can be made by manually collecting quarterly data from each year, this quadrupling the amount of data point we would have access to.

Github Link: https://github.com/DWShuo/GDP-data-analysis
R-packages: readxl. Stringr. Ggplot2. blscrapeR. Rpart. Rpart.plot, MASS. caret
References:

- https://www.bls.gov/news.release/pdf/wkyeng.pdf
- https://www.nytimes.com/2019/07/28/business/economy/economy-recession.html
- https://www.cbsnews.com/news/2018-taxes-some-of-americas-biggest-companies-paid-little-to-no-federal-income-tax-last-year/
- https://en.wikipedia.org/wiki/Tax_Cuts_and_Jobs_Act_of_2017
- https://en.wikipedia.org/wiki/Revenue_Act_of_1936