USA BUSINESS SUBSIDY ANALYSIS

Gnana Teja Peddi

Table of Contents

Executive Summary:

The main purpose of our report is to understand the factors that affect – Subsidy Value received by an organization belonging to an industry across every state and city in the United States of America. Subsidy being one of the major economic drivers for every country especially for the United States of America the relevance of our analysis becomes even more critical to understand what makes the amount awarded to various organizations shoot up or go down. Our main agenda revolves around creating the best model from our set of independent variables that can predict the subsidy value to be awarded to the organization in any given region belonging to any specific industry in the USA. The biggest limitation of our work is the fact that Subsidy Value is dependent on a wide range of economic activities that promote the initiative which might be a result of political agenda or to drive economic growth in a specific industry. Therefore, I had to develop a strong model by working with the limited scope of data used in our analysis.

The approach taken by us for this project was to first analyse the independent variables individually & then understand the relationship between the independent variable with our dependant variable - Subsidy Value. Post this I started creating our best fit model that explains the variance in our dependant variable with respect to the independent variables.

Our key findings reveal that there is statistical difference between subsidy value(log) received by organisations across states.

Introduction to the topic:

What is subsidy? Most Subsidies are cash grants or loans that the government gives to businesses. It encourages activities the government wishes to promote. The subsidy depends on the amount of the goods or services provided. One level of government can also give subsidies to another. This includes federal grants given to state or local governments and state grants given to municipal governments.

The WTO(World Trade Organisation) has given a broader understanding of Subsidy through its definition- It says “subsidy” is any financial benefit provided by a government which gives an unfair advantage to a specific industry, business, or even individual.

According to the WTO there are five types of subsidies:

- Cash subsidies, such as the grants mentioned above
- Tax concessions, such as exemptions, credits, or deferrals
- Assumption of risk, such as loan guarantees
- Government procurement policies that pay more than the free-market price
- Stock purchases that keep a company's stock price higher than market levels

The importance of Subsidy as an economic driver:

Government funded subsidies help an industry by paying for part of the cost of the production of a good or service by offering Tax credits or reimbursements or by paying for part of the cost a consumer would pay to purchase a good or service.

- Increasing Production and Consumption - Governments seek to implement subsidies to encourage production and consumption in specific industries. On the supply side, government subsidies help an industry by allowing the producers to produce more goods and services. This shifts the supply curve to the right & thereby increases the overall supply of that good or service, increases the quantity demanded for that good or service and lowers the overall price of the good or service.
- Increasing Savings - Since the government helps suppliers through tax credits or reimbursements, the lower overall price of their goods and services is more than offset by the savings they receive.
- Tax Credits - On the consumer side, government subsidies can help potential consumers with the cost of a good or service, usually through tax credits. A great example of this is when consumers who refit their houses with solar panels receive a tax credit to offset the high price of purchasing the new solar panels. This helps specially the renewable industry by allowing more consumers to purchase the products associated with that industry.
- The Bottom Line Impact - Government subsidies can help an industry on both the supplier side and the consumer side. To implement subsidies, governments need to raise taxes or reallocate taxes from existing budgets. There is also an argument that incentives in the form of subsidies actually reduce the incentives of firms to cut costs.

Source of Data

SUBSIDY TRACKER is the first national search engine for economic development subsidies and other forms of government financial assistance to business. I extracted the subsidy information state wise for the United States of America from this website.

The dataset contains 67051 rows and 31 columns/variables.

Each row in the dataset represents the Subsidy value granted for a particular company by the Government.

The dataset has information about :

Subsidy award entries: 629,000 (400,000 state/local; 229,000 federal)

Subsidy programs: 1,039 (901 state/local; 138 federal)

Parent companies covered: 2,934

Data Description

Subsidy Value adjusted for Megadeal - The Dependent variable and it is the value(dollar amount) of subsidy granted.

Year: The year in which a specific subsidy (or portion of a multi-year subsidy) was awarded or disbursed.

Location: Each state/local recipient is from a specific state; I have added the state name for those entries. The raw data has entries of companies for 51 states in the USA. For federal entries this indicates the physical location of the recipient company rather than the source of the subsidy. Not all entries have this information because it is sometimes absent from the source documents, which are then listed under United States as the location information.

Subsidy Source: The source is the level of government (state, local or federal) of the agency which awarded the subsidy. Some entries have "multiple" in this field to indicate that the subsidy package included state as well as local components.

Ownership Structure: This provides basic information on the company such as ownership status.

Type of subsidy: various subsidy programs are divided into 19 broad categories, 14 for state/local program and 5 for federal programs

Research Questions

Primary question

What are the factors that strongly influence the Subsidy given by the Government to an organization/company?

Hypothesis:

1. There is statistically significant difference among the average Log Subsidy value awarded to companies in different regions.

2. Economically less progressive states are provided with more subsidies than the economically more progressive states.

3. Companies with multiple source (subsidies from state as well as local) receive the highest average (log)subsidy value.

4. All major sectors receive similar subsidy value and the major industry/sector does not play a role in determining the same.

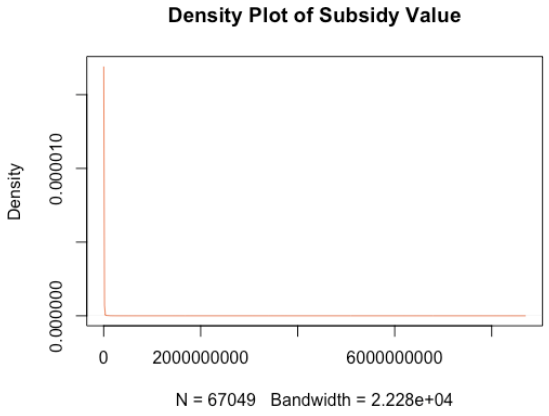
5. There is statistically significant correlation between the average Log Subsidy value awarded to companies and the no.of years.

6. The average Log Subsidy value awarded to companies by Republicans is more than the Democrats.

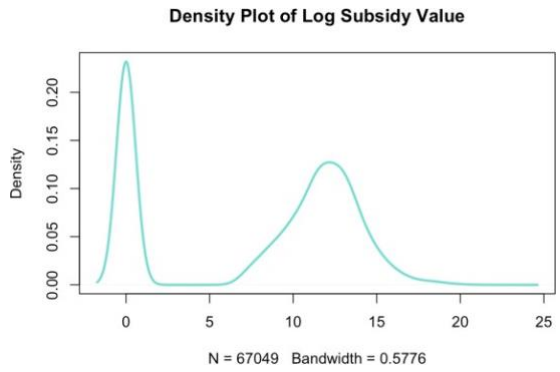
7. There is statistically significant correlation between the average Log Subsidy value awarded to companies and the unemployment rate.

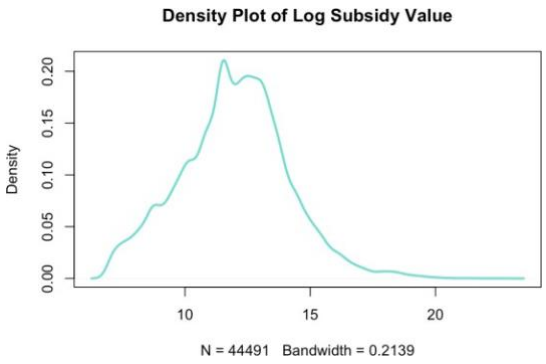
Data Analysis

Dependent Variable **-** The dependent variable, **subsidy value** is a continuous variable with the following density plot.



The density plot for the dependent variable was extremely right skewed. To handle this I performed log transformation. And to handle the zero inflation, I removed the zeros to get a fairly normal distribution.





The descriptive statistics of Log of subsidy value is as follows:

Min	1st Qu.	Median	Mean	3rd Qu	Max
6.909	10.597	11.999	11.972	13.305	22.887

Standard Deviation = 2.161491

Independent Variables Analysis:

1.Location has 51 state values which are then converted into a meaningful variable as regions in the USA and used as a factor variable for better analysis. Univariate analysis - Table, Levels: determines that there are six levels in the “Region” category – five regions as per the US official demarcations and one level added for the rows where the physical location of the company is unknown and subsidies are provided by the federal government. Also, doing bivariate analysis of

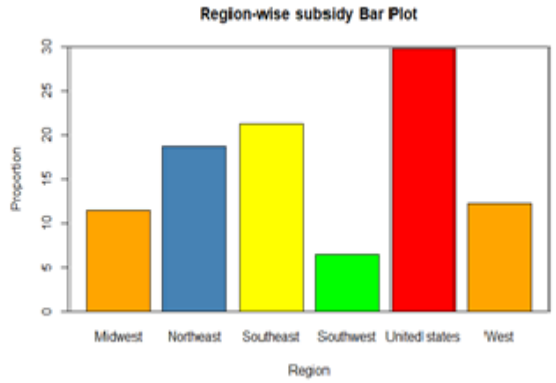
our factor variable – Location (Region) with our dependent variable – Subsidy Value (Log converted value)

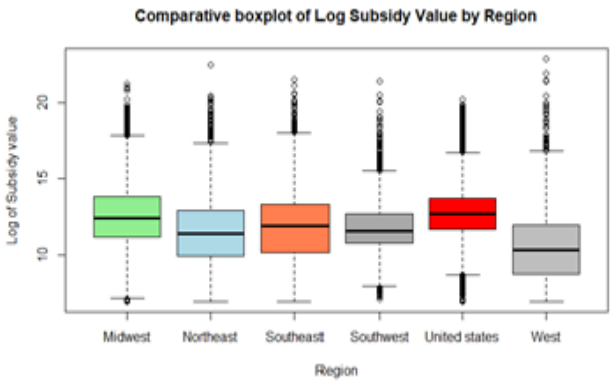
Table 1: Region wise number of subsidies awarded

Midwest	Northeast	Southeast	Southwest	United states	West
5058	8272	9367	2854	13134	5408

Table 2: Log of Subsidy value sorted from highest to lowest region wise.

Region	Log of Subsidy Value
United states	12.80993
Midwest	12.52348
Southwest	11.79545
Southeast	11.75189
Northeast	11.48007
West	10.49514





Anova test was performed to test the hypothesis:

Null hypothesis: There is no difference among the average Log Subsidy value awarded to companies in different regions.

Alternative hypothesis: There is statistically significant difference among the average Log Subsidy value awarded to companies in different regions.

Call: aov(formula = subval ~ Region, data = Subsidy_Data_merge)		
Terms:		
	Region	Residuals
Sum of Squares	25084.51	178443.60
Deg. of Freedom	5	44087
Residual standard error : 2.011848		

Summary:

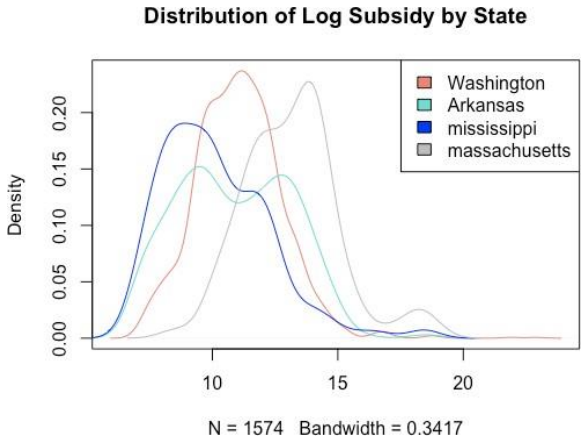
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Region	5	25085	5017	1239	<0.0000000000000002
Residuals	44087	178444	4		
Region	***				

The **Anova test** was performed in R to determine if there is a difference between the average Log Subsidy value awarded to companies in different regions of the US. The p-value is less than 0.05, and therefore we reject the null hypothesis. There is a statistically significant difference between the average Log Subsidy value awarded to companies in different regions.

To find if there is difference between the subsidies provided to economically more progressive and economically less progressive states :

I wanted to compare the Average value of Subsidy(log subsidy) provided to the economically less progressive states and the economically more progressive states.

I selected a few economically less progressive states(Arkansas, Mississippi) and compared the Average log of subsidy provided by the government to these states as compared to the average log of subsidy provided to the economically more progressive states(Washington, Massachusetts).



I also performed a **t-test** to test the hypothesis:

.

Alternative Hypothesis: Economically less progressive states are provided more Subsidies(log of subsidy) by the government as compared to economically more progressive states.

One Sample t-test		
95% confidence interval	10.45232	Inf
t	-6.7426	
p-value	1	

Surprisingly, the p- value signifies that the null cannot be rejected and hence the alternative hypothesis cannot be accepted. This implies that Economically more progressive states are provided more Subsidies(log of subsidy) by the government as compared to economically less progressive states.

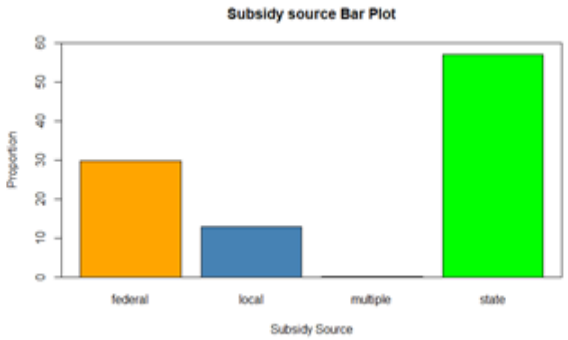
2. Source of subsidy is analyzed on a univariate level to find the levels in which the subsidy dataset is divided. Performing the univariate analysis, I could find that source of the subsidies is from four levels – state, federal, local and multiple (multiple being state as well as local). Highest number of subsidies are provided by state followed by federal subsidies. Relatively lesser number of companies are awarded multiple subsidies.

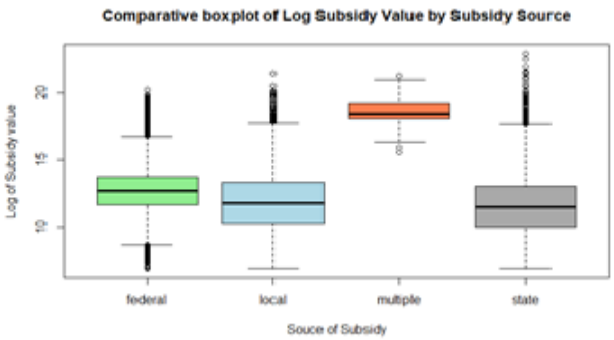
Table 1: Number of subsidies awarded based on source of subsidy

federal	local	multiple	state
13134	5698	84	25177

Table 2: Log of Subsidy value sorted from highest to lowest by subsidy source.

Subsidy Source	Log of Subsidy Value
multiple	18.57146
federal	12.80993
local	11.84063
state	11.50974





Anova test was performed to test the hypothesis:

Null hypothesis: There is no difference among the average Log Subsidy value awarded to companies by different source of subsidies.

Alternative hypothesis: There is statistically significant difference among the average Log Subsidy value awarded to companies by different source of subsidies. Companies with multiple source (subsidies from state as well as local) receive the highest average (log)subsidy value.

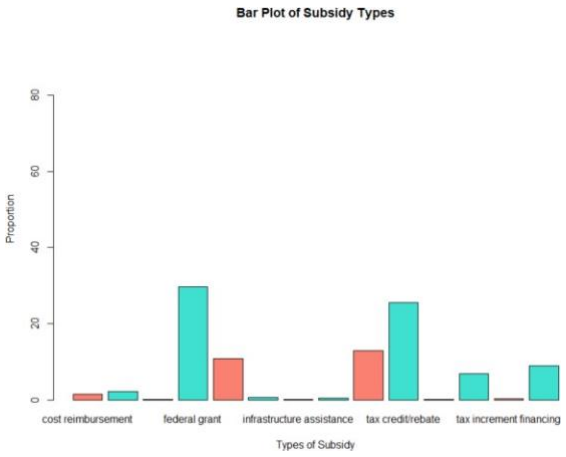
Call:aov(formula = subval ~ Subsidy.Source, data = Subsidy_Data_merge)		
Terms:		
	Subsidy Source	Residuals
Sum of Squares	18342.74	185185.36
Deg. of Freedom	3	44089
Residual standard error : 2.049454		

Summary:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SubsidySource	3	18343	6114	1456	<0.0000000000000002
Residuals	44089	185185	4		
Subsidy Source ***					

The **Anova test** was performed in R to determine if there is a difference among the average Log Subsidy values awarded to companies by different sources. The p-value is less than 0.05, and therefore we reject the null hypothesis. There is a statistically significant difference among the average Log Subsidy value awarded to companies by different sources. Companies with multiple source (subsidies from state as well as local) receive the highest average (log)subsidy value.

3. **Type of Subsidy** had 14 levels in the beginning when analyzed through univariate analysis.



The following table shows the average value of Log of subsidy for each Type of Subsidy

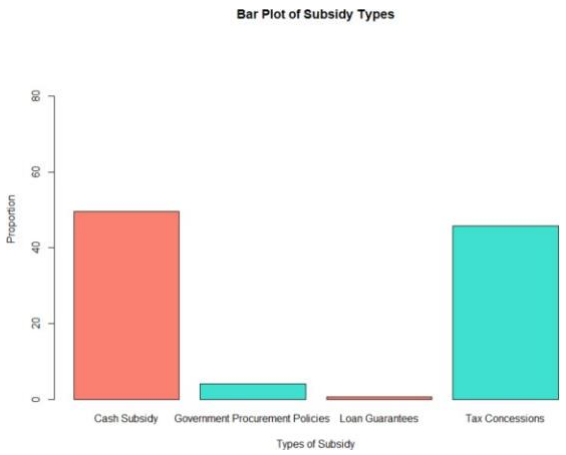
Type of Subsidy	log(Subsidy Value)
cost reimbursement	9.227905
enterprise zone	11.743916
federal allocated tax credit	15.854627
federal grant	12.79408
grant	11.958289
grant/loan hybrid program	13.3054
infrastructure assistance	13.848718
megadeal	18.670860
property tax abatement	11.753964
tax credit/rebate	11.496158
tax credit/rebate and grant	13.405853
tax credit/rebate; property tax abatement	11.736732
tax increment financing	13.848219
training reimbursement	10.962319

As per the **WTO**, subsidies are classified into four main classes. They are

1. Cash subsidies, such as the grants mentioned above.
2. Tax concessions, such as exemptions, credits, or deferrals.
3. Assumption of risk, such as loan guarantees.
4. Government procurement policies that pay more than the free-market price.
5. Stock purchases that keep a company's stock price higher than market levels.

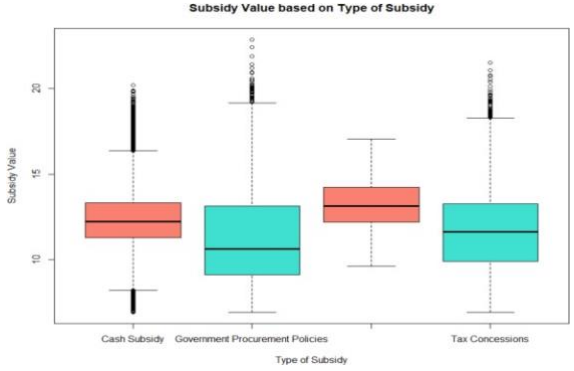
So I grouped these 14 types of subsidies into the above given five categories. None of them belonged to the fifth category. Hence I got a new column named Subsidy. Category with four levels.

The data distribution with the new column (Subsidy.Category) looks like this



After doing bivariate analysis, the following table shows the average Log of subsidy for different Subsidy categories.

Subsidy Category	**log(subsidy value)**
Cash Subsidy	12.28285
Government Procurement Policy	11.51728
Loan Guarantees	13.29876
Tax Concessions	11.61431



Anova test was performed to test the hypothesis:

Null hypothesis: There is no difference among the average Log Subsidy value for different types of subsidies.

Alternative hypothesis: There is statistically significant difference among the average Log Subsidy value for different type of subsidies.

Call: aov(formula = subval ~ Subsidy.Category, data = Subsidy_Data_merge)		
Terms:		
	Subsidy. Category	Residuals
Sum of Squares	5580.41	197947.69
Deg. of Freedom	3	44089
Residual standard error : 2.118898		

Summary:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Region	3	5580	1860.1	414.3	<0.0000000000000002

Residuals	44089	197948	4.5		
Subsidy Category	***				

From the above table, the p-value is less than 0.05, and therefore we reject the null hypothesis at 95% confidence level and accept the alternate hypothesis. There is a statistically significant difference among the average Log Subsidy value for different types of subsidies.

4. Ownership Structure

Ownership structure has 38 levels. Since it has 38 levels, I found out top 10 ownership structures with the highest subsidy values.

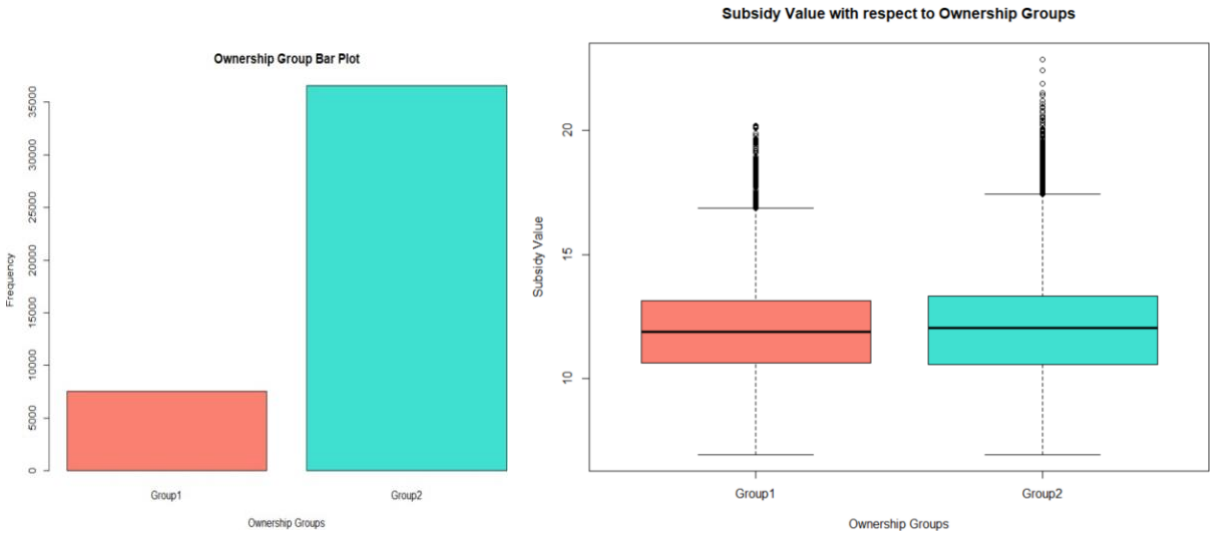
Ownership Structure	log(Subsidy Value)
joint venture (owned 55-50 by phillips 66 and enbridge)	16.38474
joint venture (owned 50-50 by celanese and kureha)	15.52053
non-profit	15.19882
government-owned	14.93605
joint venture (owned 50-50 by cenovus energy and phillips 66)	14.52093
joint venture (owned by group of utilities)	13.72331
joint venture (owned 50-50 by valero energy and darling ingredients)	13.51426
joint venture (owned 50-50 by huntsman and kronos worldwide, which is owned by valhi)	13.28813
joint venture (owned by group of utilities)	13.72331
joint venture (owned 50-50 by philips 66 and chevron)	13.17056

And then I noticed that all the ownership structures that start with “joint venture” are ideally the same with different kind of Joint Ventures with different companies.

So I decided to merge them together into a single category and it was reduced to 11 levels which include:

"alaska native-owned" , "cooperative" ,"employee-owned" ,"government-owned" , "government sponsored & publicly traded", "Joint Venture" , "mutual","non-profit" ,"out of business" "privately held", "publicly traded"

Since Publicly Traded has got most of the data points, I again grouped these 11 levels into two groups with "publicly traded" as one group and all the rest into another group.



However, log of subsidy value was almost the same for both groups when examined by average.

Ownership.Group	**log(subsidy value)**
Group1	11.92365
Group2	11.95935

Anova test was performed to find out if there is any statistical difference between different ownership groups:

Null hypothesis: There is no difference among the average Log of Subsidy value given to different ownership groups

Alternative hypothesis: There is statistically significant difference among the average Log Subsidy value given to different ownership groups

Call: aov(formula = subval ~ Ownership.Group data = Subsidy_Data_merge)		
Terms:		
	Ownership.Group	Residuals
Sum of Squares	7.97	203520.14
Deg. of Freedom	1	44091
Residual standard error : 2.148467		

Summary:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Region	1	8	7.967	1.726	0.189
Residuals	44091	203520	4.616		
Ownership.Group					

The p- value indicates that we fail to reject the null and there is statistically significant correlation between the average Log Subsidy value awarded to different ownership groups.

.

5. Year:

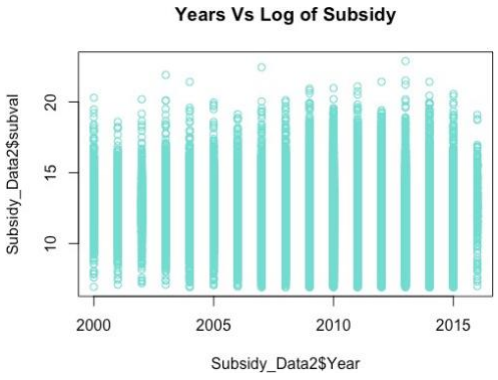
I first started by substituting the median value of year for the missing values in this Year variable. The univariate analysis of the variable year has the following descriptive statistics:

Min	1st Qu.	Median	Mean	3rd Qu	Max
1983	2007	2010	2009	2012	2017

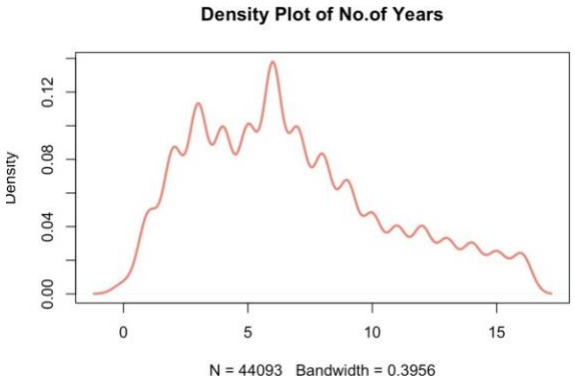
Standard Deviation = 4.046997

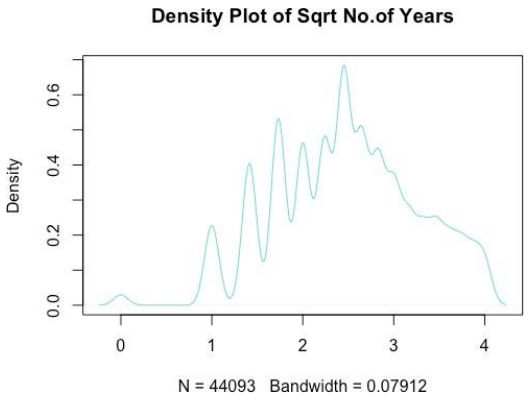
The univariate analysis also showed that Year less than 2000 and greater than 2016 dont have as many data points, so I took a subset of the dataset without these years for our analysis.

The plot of Log Subsidy Vs Year makes us think Year could also be a categorical variable here.



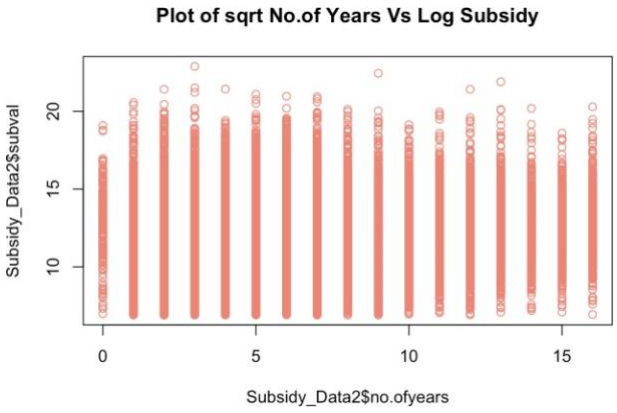
I create a column no.of years (no.of years since the subsidy was given) from Year as abs(max(Subsidy_Data2$Year)-Subsidy_Data2$Year).





To make the distribution better, we use sqrt transformation for the variable no.of years

Bivariate analysis : Sqrt No.of Years Vs Log Subsidy



YEAR and log subsidy value have a weak negative correlation with a correlation coefficient of -0.1.

No.of years and log subsidy value have a weak positive correlation with a correlation coefficient of 0.089

Cor.Test was performed to test the hypothesis

Null hypothesis: There is no correlation between the average Log Subsidy value awarded to companies for different no.of years.

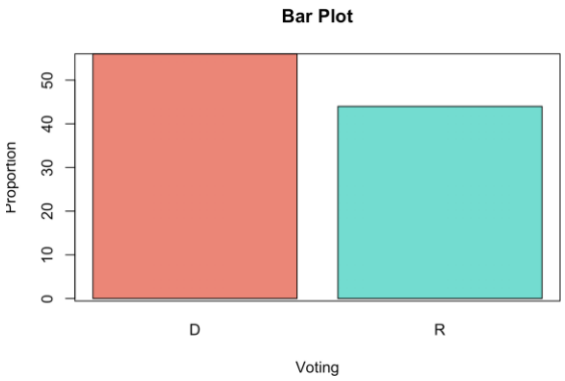
Alternative hypothesis: There is statistically significant correlation between the average Log Subsidy value awarded to companies and the no.of years.

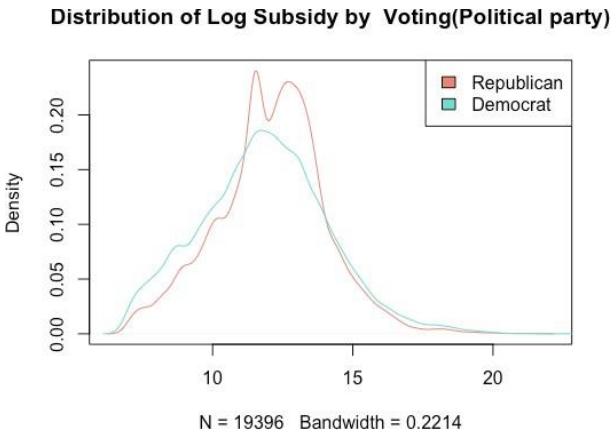
Pearson's product-moment correlation		
95% confidence interval	0.08029941	0.09881757
cor	0.08956623	
p-value	0.00000000000000022	

The p- value indicates that we can reject the null and accept the alternative hypothesis that there is statistically significant correlation between the average Log Subsidy value awarded to companies and the no.of years.

6. Voting:

To make the analysis more interesting, I created a column **Voting**- a categorical variable with two levels- R(Republican) and D(Democrat)-(Since our dataset had no value for level 'Other', I dropped it). The values were designated to the data points for this variable based on the Year and the Location(state).





The univariate shows that Democrats have given more number of subsidies as compared to the Republicans.

Democrats	Republicans
24697	19396

By doing the bivariate analyses of Voting(Political Party) Vs Log Subsidy, we notice that the mean values of Log Subsidy is a little higher for Republicans than the democrats.

Voting	Average	Median	Standard Deviation
Democrats	11.8	11.9	2.28
Republicans	12.1	12.2	1.96

Anova test was performed to test the hypothesis:

Null hypothesis: There is no difference among the average Log Subsidy value awarded to companies by Democrats and Republicans

Alternative hypothesis: There is statistically significant difference among the average Log Subsidy value awarded to companies by Democrats and Republicans.

Call: aov(formula = subval ~ Voting, data = Subsidy_Data_merge)		
Terms:		
	Voting	Residuals
Sum of Squares	609.13	202918.98
Deg. of Freedom	1	44091
Residual standard error : 2.145292		

Summary:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Voting	1	609	609.1	132.4	<0.0000000000000002
Residuals	44091	202919	4.6		
Voting ***					

The ANOVA test between the political party and the Log Subsidy, shows that there is a statistical difference between the average Log of Subsidy of the two levels in the Voting(political party).

Also, the t test shows that the average value of log subsidy of Republicans is higher than that of the Democrats.

7. Parent Sector :

Parent Sector has 50 values which are then converted into meaningful variable as sectors in the USA and used as a factor variable for better analysis. They have been divided into 4 main sectors like Primary, Secondary, Tertiary and Quaternary based on type of sector. There are four types of industry. These are primary, secondary, tertiary and quaternary.

Primary industry involves getting raw materials e.g. mining, farming and fishing.

Secondary industry involves manufacturing e.g. making cars and steel.

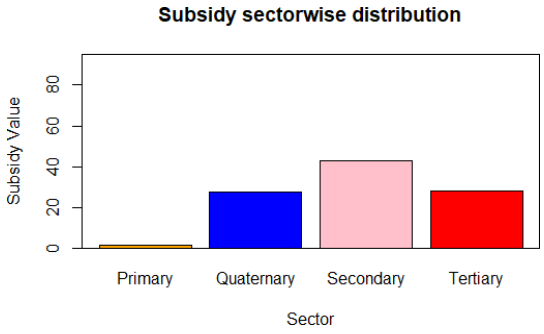
Tertiary industries provide a service e.g. teaching and nursing.

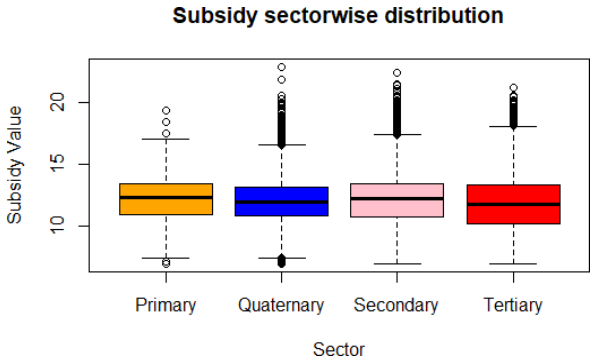
Quaternary industry involves research and development industries e.g. IT.

Also, doing bivariate analysis of our factor variable – Parent Sector with our dependent variable – Subsidy Value (Log converted value) I noticed that the count variation is very widely distributed and there are 18k records in Tertiary but just 621 in Primary and hence I have regrouped and combined Primary and Quaternary for a better analysis.

Table 1: Sector wise number of subsidies awarded

Primary	Secondary	Tertiary	Quaternary
621	12209	18990	12273





Anova test was performed to test the hypothesis:

Null hypothesis: There is no difference among the average Log Subsidy value awarded to different Sectors.

Alternative hypothesis: There is statistically significant difference among the average Log Subsidy value awarded to companies to different major parent sectors.

Call:aov(formula = subval ~ Subsidy.Source, data = Subsidy_Data_merge)		
Terms:		
	ParentSector	Residuals
Sum of Squares	1165.59	202362.52
Deg. of Freedom	2	44090
Residual standard error :2.142373		

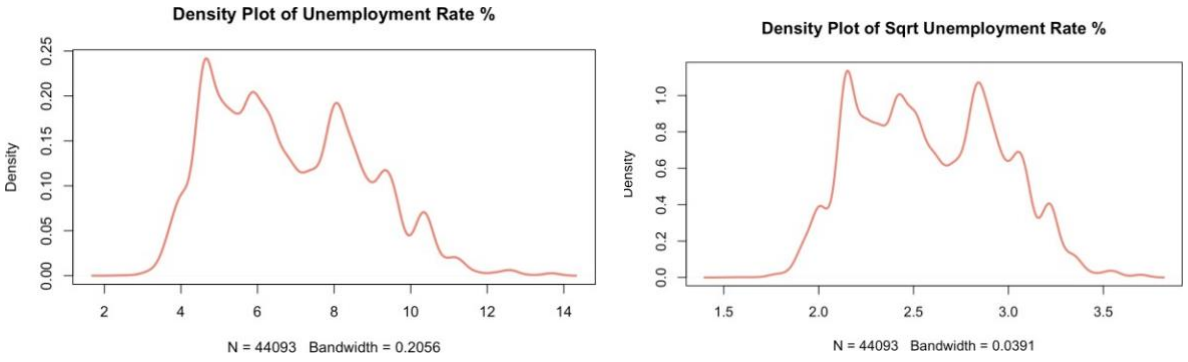
Summary:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ParentSector	2	1166	6114	582.8	<0.0000000000000002
Residuals	44090	202363	4.6		

The **Anova test** was performed in R to determine if there is a difference between the average Log Subsidy value awarded to different Sectors of the US. The p-value is less than 0.05, and therefore we reject the null hypothesis. There is a statistically significant difference between the average Log Subsidy value awarded to different sectors.

8. Unemployment Rate:

Another interesting variable that I wanted to analyze was 'Unemployment Rate'. So I created a column 'uemprate_percent' in which unemployment rate percent values were designated to the data points for this variable based on the Year and the Location(state).



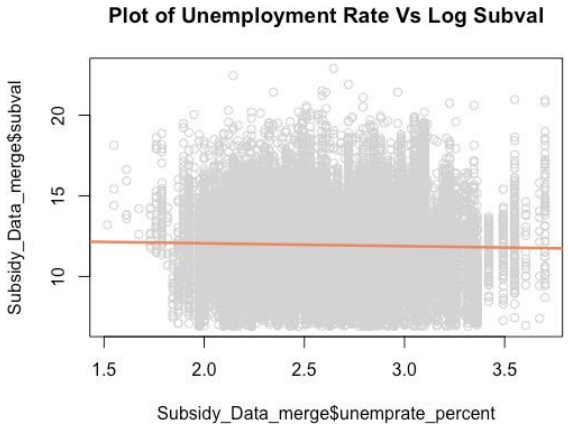
I did the square root transformation on the Unemployment rate to conform with normal distribution.

Summary of the sqrt Unemployment Rate:

Min	1st Qu.	Median	Mean	3rd Qu	Max
1.517	2.280	2.550	2.586	2.881	3.701

Standard Deviation = 0.3687947

The correlation between the Sqrt Unemployment rate and Log Subsidy is very low (coefficient = -0.03).



Cor.Test was performed to test the hypothesis:

Null hypothesis: There is no correlation between the average Log Subsidy value awarded to companies and the unemployment rate.

Alternative hypothesis: There is statistically significant correlation between the average Log Subsidy value awarded to companies and the unemployment rate.

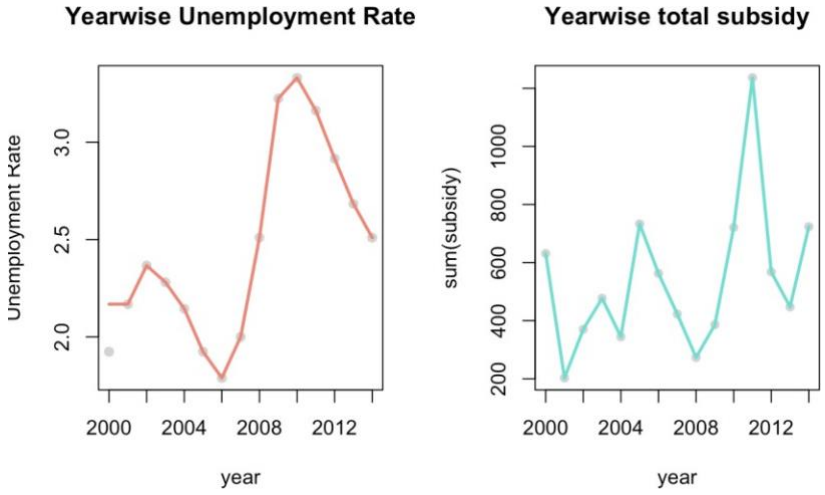
Pearson's product-moment correlation		
95% confidence interval	-0.03978111	-0.02113053

cor	-0.03045847	
p-value	0.0000000001584	

The p-value indicates that there is statistically significant correlation between the unemployment rate and average Log of subsidy at 95% confidence.

We shall now consider ‘Florida’(swing state)

The Plot of year wise sqrt Unemployment Rate and year wise sum of log subsidy value is shown below.



Observe that in the year 2010, the unemployment rate is at its peak. As a reactive measure the sum of subsidy in 2011 is at its peak. Consequently, the unemployment rate decreases in the following years.

Multiple Regression :

I first built the model with the important independent variables(based on the univariate and bivariate analysis) and the Log Subsidy value (dependent variable).

```
                                           Estimate Std. Error t value             Pr(>|t|)
(Intercept)                                12.72802    0.08975 141.810 < 0.0000000000000002 ***
Subsidy.Sourcelocal                        -0.55786    0.04825 -11.562 < 0.0000000000000002 ***
Subsidy.Sourcemultiple                      6.38049    0.22314  28.594 < 0.0000000000000002 ***
Subsidy.Sourcestate                        -0.63035    0.03710 -16.990 < 0.0000000000000002 ***
Ownership.GroupGroup2                       0.04413    0.02533   1.743             0.081409 .
RegionNortheast                            -1.06350    0.03709 -28.677 < 0.0000000000000002 ***
RegionSoutheast                            -0.75029    0.03555 -21.106 < 0.0000000000000002 ***
RegionSouthwest                            -0.45604    0.04784  -9.532 < 0.0000000000000002 ***
RegionWest                                 -1.99686    0.04022 -49.653 < 0.0000000000000002 ***
no.ofyears                                  0.07723    0.01289   5.991         0.0000000021 ***
Subsidy.CategoryGovernment Procurement Policies -0.19275    0.05494  -3.508             0.000451 ***
Subsidy.CategoryLoan Guarantees             1.74365    0.11573  15.067 < 0.0000000000000002 ***
Subsidy.CategoryTax Concessions             0.43066    0.02895  14.875 < 0.0000000000000002 ***
ParentSectorQuaternary                     -0.31991    0.08187  -3.907         0.0000934117 ***
ParentSectorSecondary                      -0.01454    0.08112  -0.179             0.857780
ParentSectorTertiary                       -0.20056    0.08196  -2.447             0.014404 *

Residual standard error: 1.975 on 44077 degrees of freedom
Multiple R-squared:  0.1552,    Adjusted R-squared:  0.1549
F-statistic: 539.7 on 15 and 44077 DF,  p-value: < 0.00000000000000022
```

The Adjusted R- squared for the model was 15.49%, that is the model could explain 15.49% variance in the Log of subsidy value. Some variables used were statistically insignificant in determining the Log of subsidy at 95% confidence.

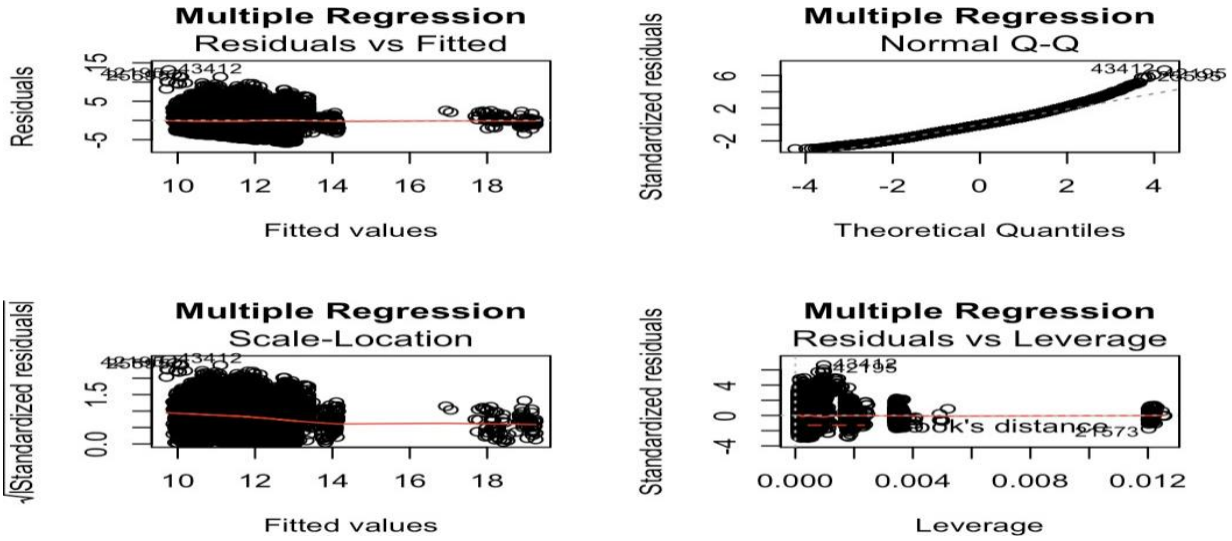
After multiple trials that involved regrouping the variables(parent sector), including some variables (unemployment rate and voting(political party)), changing the reference category/releveling, I built the following model.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	12.17912	0.09473	128.564	< 0.0000000000000002	***
Subsidy.Sourcelocal	-3.21866	0.09680	-33.250	< 0.0000000000000002	***
Subsidy.Sourcemultiple	-3.83518	0.30650	-12.513	< 0.0000000000000002	***
Subsidy.Sourcestate	-3.07229	0.08676	-35.413	< 0.0000000000000002	***
unemprate_percent	0.12853	0.02891	4.446	0.00000876822459	***
Ownership.GroupGroup2	0.05222	0.02453	2.129	0.0333	*
RegionNortheast	-1.15444	0.04118	-28.037	< 0.0000000000000002	***
RegionSoutheast	-0.40091	0.03883	-10.325	< 0.0000000000000002	***
RegionSouthwest	-0.51053	0.05107	-9.996	< 0.0000000000000002	***
RegionWest	-1.89639	0.04076	-46.529	< 0.0000000000000002	***
no.ofyears	0.09657	0.01360	7.102	0.00000000000125	***
Type.of.Subsidyenterprise zone	2.81251	0.10266	27.397	< 0.0000000000000002	***
Type.of.Subsidyfederal allocated tax credit	2.88760	0.23360	12.361	< 0.0000000000000002	***
Type.of.Subsidygrant	2.74260	0.08685	31.578	< 0.0000000000000002	***
Type.of.Subsidygrant/loan hybrid program	4.20967	0.13674	30.786	< 0.0000000000000002	***
Type.of.Subsidyinfrastructure assistance	3.88852	0.23596	16.479	< 0.0000000000000002	***
Type.of.Subsidymegadeal	9.92028	0.22221	44.644	< 0.0000000000000002	***
Type.of.Subsidyproperty tax abatement	2.67374	0.08424	31.740	< 0.0000000000000002	***
Type.of.Subsidytax credit/rebate	2.75892	0.08200	33.645	< 0.0000000000000002	***
Type.of.Subsidytax credit/rebate and grant	3.88483	0.27604	14.074	< 0.0000000000000002	***
Type.of.Subsidytax credit/rebate; property tax abatement	3.19423	0.10373	30.795	< 0.0000000000000002	***
Type.of.Subsidytax increment financing	4.33237	0.18846	22.988	< 0.0000000000000002	***
Type.of.Subsidytraining reimbursement	2.03187	0.08507	23.883	< 0.0000000000000002	***
ParentSector1Quaternary	-0.05582	0.02528	-2.208	0.0272	*
ParentSector1Secondary	0.19297	0.02255	8.556	< 0.0000000000000002	***
VotingR	-0.14881	0.02852	-5.218	0.00000018136021	***

Residual standard error: 1.915 on 44067 degrees of freedom
Multiple R-squared: 0.2056, Adjusted R-squared: 0.2052
F-statistic: 456.3 on 25 and 44067 DF, p-value: < 0.00000000000000022

The adjusted R- squared for the model was 20.52%, that is the model could explain 20.52% variance in the Log of subsidy value. All variables used were statistically significant in determining the Log of subsidy at 95% confidence. I then plotted the model to observe the regression diagnostics.

Model Diagnostics:



The **final model** after removing the outliers has the following msummary:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	12.18006	0.09465	128.691	< 0.0000000000000002	***
Subsidy.Sourcelocal	-3.21313	0.09671	-33.223	< 0.0000000000000002	***
Subsidy.Sourcemultiple	-3.83230	0.30621	-12.515	< 0.0000000000000002	***
Subsidy.Sourcestate	-3.07077	0.08668	-35.428	< 0.0000000000000002	***
unemprate_percent	0.12801	0.02888	4.432	0.000009355261	***
Ownership.GroupGroup2	0.05189	0.02451	2.117	0.0342	*
RegionNortheast	-1.15402	0.04114	-28.052	< 0.0000000000000002	***
RegionSoutheast	-0.40206	0.03879	-10.364	< 0.0000000000000002	***
RegionSouthwest	-0.51099	0.05103	-10.014	< 0.0000000000000002	***
RegionWest	-1.89852	0.04072	-46.623	< 0.0000000000000002	***
no.ofyears	0.09689	0.01358	7.132	0.000000000001	***
Type.of.Subsidyenterprise zone	2.81082	0.10256	27.406	< 0.0000000000000002	***
Type.of.Subsidyfederal allocated tax credit	2.88809	0.23338	12.375	< 0.0000000000000002	***
Type.of.Subsidygrant	2.74141	0.08677	31.594	< 0.0000000000000002	***
Type.of.Subsidygrant/loan hybrid program	4.20777	0.13661	30.800	< 0.0000000000000002	***
Type.of.Subsidyinfrastructure assistance	3.88713	0.23574	16.489	< 0.0000000000000002	***
Type.of.Subsidymegadeal	9.91788	0.22200	44.675	< 0.0000000000000002	***
Type.of.Subsidyproperty tax abatement	2.66850	0.08416	31.706	< 0.0000000000000002	***
Type.of.Subsidytax credit/rebate	2.75781	0.08192	33.663	< 0.0000000000000002	***
Type.of.Subsidytax credit/rebate and grant	3.87948	0.27578	14.067	< 0.0000000000000002	***
Type.of.Subsidytax credit/rebate; property tax abatement	3.18827	0.10363	30.765	< 0.0000000000000002	***
Type.of.Subsidytax increment financing	4.37862	0.18881	23.191	< 0.0000000000000002	***
Type.of.Subsidytraining reimbursement	2.03151	0.08499	23.902	< 0.0000000000000002	***
ParentSector1Quaternary	-0.05587	0.02525	-2.212	0.0269	*
ParentSector1Secondary	0.19241	0.02253	8.539	< 0.0000000000000002	***
VotingR	-0.14890	0.02849	-5.226	0.000000173719	***

Residual standard error: 1.914 on 44063 degrees of freedom
Multiple R-squared: 0.2061, Adjusted R-squared: 0.2057
F-statistic: 457.6 on 25 and 44063 DF, p-value: < 0.00000000000000022

The model can explain 20.57% variance in the Log of Subsidy value. All the independent variables, numerical and all levels of categorical are statistically significant in determining the Log of Subsidy value.

The interpretation of the model is as follows:

- The average log of subsidy value provided by local subsidy source is $3.213 lower as compared to average log of subsidy provided by federal source(reference category), holding other variables at their constant. Similarly, we can interpret the average log of subsidy provided by other sources as compared to the reference category(Federal subsidy source).
- A unit increase in unemployment rate increases the log of subsidy by 0.12801$, on average, holding other variables at their constant.
- The average log of subsidy value provided by Ownership group 2 is $0.05189 higher as compared to average log of subsidy provided by Ownership group 1(reference category), holding other variables at their constant.
- The average log of subsidy value provided to the Northeast Region is $1.15402 lower as compared to average log of subsidy provided to the Midwest Region(reference category, holding other variables at their constant. Similarly, we can interpret the average log of subsidy provided to the other Regions as compared to the reference category(Midwest Region).
- A unit increase in the number of years increases the log of subsidy by $0.09689, on average, holding other variables at their constant.
- The average log of subsidy value provided as enterprise zone subsidy is $2.81082 higher as compared to average log of subsidy provided as cost reimbursement(reference category),

holding other variables at their constant. Similarly, we can interpret the average log of subsidy provided as the other types of subsidy as compared to the reference category(cost reimbursement).

- The average log of subsidy value provided to Quaternary sector is $0.05587 lower as compared to average log of subsidy provided to Primary-Tertiary sector(reference category), holding other variables at their constant. Similarly, we can interpret the average log of subsidy provided to the other sectors as compared to the reference category(Primary-Tertiary sector).
- The average log of subsidy value provided by the Republican Government is $0.14890 lower as compared to average log of subsidy provided by the Democrats(reference category), holding other variables at their constant.

Conclusion:

The key findings from our analysis are as follows:

- The average value of the log of subsidy value was different for across regions.
- Companies with multiple sources of subsidy receive highest amount of subsidies.
- The average value of the log of subsidy is the highest for Loan Guarantee
- The average value of log subsidy given by Republicans is higher than that of the Democrats.
- The average value of log of subsidy is greatest for Secondary Sector, on an average.

Experimental Design Limitations

- The dataset does not contain the information about all the companies that received subsidies by the government in the United states of America for the time frame used in our analysis (2007 to 2017) and the dataset is not a random sample.
- The dependent variable(Subsidy Value) was highly right skewed and zero inflated. In order to do the analysis, I had to drop the observations with Subsidy value=0 and do the log transformation on the same. This led to the reduction in the number of observations from 67051 to 44491.
- I was unable to use all the variables in the dataset because of the following reasons:
    - Missing values (more than 20%)
    - Too many levels
- The subsidy value could be dependent on a wide range of factors, not limited to the variables present in our dataset.

References

https://www.thebalance.com/government-subsidies-definition-farm-oil-export-etc-3305788

https://www.thoughtco.com/us-farm-subsidies-3325162

http://www.geography.learnontheinternet.co.uk/topics/industrytheory.html

http://www.businessdictionary.com/definition/economic-sector.html

https://www.statista.com/statistics/189414/unemployment-rate-in-florida-since-1992/

http://www.dlt.ri.gov/lmi/laus/us/annavg.htm