George Washington University

Influence of Income in the Elections 2016

Intro to Data Science/ Pofessor Brian Wright

Lilian Sao de Rivera
11-2-2018

Contents

List of Figures

List of Tables

Introduction

Thesis Statement

The downturn observed in the democrat party in the presidential elections of 2016 was influenced by factors related to income. This analysis does not present the underlying factors. Instead, an evaluation of the influence of income over voting is displayed to confirm o deny this first assessment.

Justification

The outcome of the election 2016 was a surprise for everybody. Since then, many have tried to explain the unexpected results. Reviewing the concepts around voting behavior is essential to understand better what led to the events of 2016. Two significant concerns shape the process of voting. The first one is to understand the person's motivations that drive him or her to choose one candidate over the other. The second one is the dynamics of the electoral process. These dynamics involve the direction in which the process evolves and the changes over time. The sources of individual voting behavior are public policy, government performance, and personal characteristics of the candidates. Party identification and general ideological orientations shape these three sources. These factors can change over time, for example, a public policy issue like health care can alter the way people vote in one election term. However, party affiliation and ideology can still be stable; one person does not change party affiliation from one election to another. The dynamics of the electoral process it has to do with the analysis of changes in party affiliation over time. The fluctuations between terms can be due to economic conditions or characteristics of a candidate for example; this are short-term changes. On the contrary, the long-term changes that last more than one particular election are due to changes in loyalties and party affiliations. Long-term changes do not occur frequently; the last one can be dated back to the New Deal realignment in 1930. However, there has been cumulative changes that are important(Institute of Social Research- University of Michigan, 2018).

Besides this understanding of how people vote, other aspects were important during the 2016 elections. The GOP president ran on a platform of job creation and security issues. The slogan used refers to the "forgotten ones," which meant that there were economic factors involved in the decision process. It is important to notice that the index of job creations was already in good shape and the economy was stable. Despite these statistics, there were some signals that people did not get the benefits of the economy. One indicator was the news reports, in the article by Long (2017) in CNN Money, the author interviewed several GOP voters who emphasize the necessity of job creation and the determination of the candidate to accomplish what he was set to do. These aspects denoted a desire for change in the economic area. Another indicator is the economic inequality. According to the Economic Policy Institute, the gap in incomes between 99% and 1% is growing. This increment in wealth for the 1% increases the average income per capita for the entire population, which introduces a bias in economic indexes. This bias may explain the lack of empathy of the democrat party towards this problem in the previous election.

This increasing gap could be explained in part in a study by Graham (2017) and published by Brookings that describes the productivity "paradox." This research looks into the disparities in income among different sectors. While technological workers get significant monetary rewards from their work, the blue-collar workers are affected by automation, transformation, and obsolescence of old technology. According to Graham, "While productivity and innovation produce aggregate economic gains, those left behind experience significant costs, which extend well beyond the monetary realm." The author continues explaining that the most vulnerable areas to the productivity "paradox" are the rural areas and small towns in the heartland which tend to have few social connections outside their communities. These interesting findings explain somehow the set of mind of rural areas in United States in 2016. Under these circumstances seems logical to think that income and generations of wages was a very important issue for Americans.

To sum up, when the concepts of voting behavior are mapped with the actual state of the rural area and the growing gap in income among the different sectors, there seems to be a justification for the downturn in democrat voting. Income, economic growth, and determination to execute appear to be relevant in the elections 2016. The analysis presented here intends to prove that income influenced this drop for the democrat party.

Methodology

There are three datasets involved in the analysis. The first dataset contains information about voting in years 2012, and 2016. The voting is summarized by county. This dataset also contains general information about the county. The second dataset contains the income per capita and income per household. The last dataset contains the abbreviation of each state. This information was needed to translate the median household per income and the abbreviation by the state. In these datasets, the states were in the extended version. Once the datasets were in the correct version of state, the information about the votes was joined to the median income dataset.

The purpose of the analysis is to observe the relationship between income and the shifts of votes from the Democrat party to the Republican party. Two critical variables were defined to fulfill this goal.

Independent variable (Explanatory variable): the median income per household. The value of this variable is continuous. An extra value was created to discretize the information and make an analysis per group.

Dependent variable (Response variable): the downturn in votes from the democrat party, that is: votes from 2016 minus votes from 2012.

Descriptive variables (Display and exploratory purpose): Income per capita, state and county.

Statistical Inference Model

The analysis comprised of a categorical variable, household income, that it is needed to explain the downturn in votes, so the most appropriate model to check significance is One-Way ANOVA. The initial number of groups used to make the analysis was thirteen but due to the extreme similarity between groups a new categorization was designed base on the similarities and proximity in the scale of income. The final number of groups used were five. A markdown file accompanies this report. This report only shows the results of the last group (five elements), but the file contains both analyses.

Definition of Hypothesis

Ho: The difference between votes from 2016 and 2012 was not influenced by household income.

H1: The difference between votes from 2016 and 2012 was influenced by household income.

Tidy Data

Datasets

The datasets used in the analysis are listed in Table 1. The detailed description of the fields can be found on the reference sites listed in the reference list. Here we are going to describe only the fields relevant to the ANOVA test. The use of the three datasets is public. However, these datasets will be provided as a resource along with the markdown file.

Table 1.

List of datasets used in the analysis.

Dataset	Site	#Observations	#Variables
Votes.csv	Kaggle	3120	82
States.csv	50states	51	3
Incomepercounty.csv	Census-USA	3203	8

Cleaning

The information in the datasets needed some minor adjustments. The following are the steps that were taken in order to tidy the data.

a. States in income per county were abbreviated using the dataset states.

b. The names of counties in votes were cleaned. All of the names contain the word "county," this word was eliminated.

c. The word "city" in votes was lower case. This word was also capitalized in order to make a clean join with income per county.

d. Votes and income per county were joined to get one dataset with the data from votes and the data from median income.

e. These datasets contained 68 records that could not be joined. These observations contained the word “parish.” This information was not in the census database, so there were eliminated.

f. The income per county contained information of non-voting states; these states were not included in the final dataset.

g. An intermediate dataset was generated that contained information from votes and income per county where the counties identified as parish were not included.

h. A final dataset called votes_subset_final was generated. The final product contained the explanatory variables, the factor variable income_factor_II used in the ANOVA analysis, and the difference of percentages in votes from 2016 and 2012 for the Democratic party. The data for the Republican party was also calculated, but It was not used in the analysis.

i. Several fields were included in the final datasets but here are only described the variables that were relevant to the analysis. Table 2 has this information.

Table 2.

Final data set votes_subset_final

Variable	Type	Data
Income_factor_II	Factor	030k, 050k, 070k, 085k, >085k
MedHouseHoldIncome	Num	Median Household Income. This includes the income of the householder and all other individuals 15 years old and over in the household, whether they are related to the householder or not (*).
Dif	Num	%votes 2016 - %votes 2012
Income	Int	Per capita Income. It is the mean income computed for every man, woman, and child in a particular group including those living in group quarters (**).
State_abbr	Chr	State Abbreviation
County_name	Chr	County name

(*) Reference https://www.census.gov/quickfacts/fact/note/US/INC110216
(**) Reference https://www.census.gov/quickfacts/fact/note/US/INC910216

Exploratory Data Analysis

Descriptive Statistics

In table 3 can be observed that the means in the variable "dif" range from 0.05 to 0.07 between the groups 030k and 070k. This is an important value since the wining of the candidate in several states was close to this number. It is relevant to mention that the highest mean is the group of 50k, which could mean that this group was a major contributor to the GOP candidate. The other interesting statistic in this table is the average income in the group of 70k. The mean value for this group is around 56k which low compared to range. The income of this group is skewed to left.

Table 3.

Descriptive statistics for the dataset votes_subset_final.

Income Grp.	N	Dif Mean	Dif Mode	Dif Median	HHold Income Mean	HHold Income Mode	HHold Income Median
>085k	36	0.0124	0.0399	0.0215	95145.53	87782.81	92431.0
030k	126	**-0.0617**	-0.0367	-0.0487	**27025.62**	29007.77	27405.5
050k	2062	**-0.0772**	-0.0487	**-0.0726**	41087.65	42330.88	41483.5
070k	727	**-0.0564**	-0.0349	-0.0530	**56299.40**	52129.59	54836.0
085k	93	-0.0145	-0.0107	-0.0108	76421.52	72461.96	75604.0

Graphical presentation

The first graphical explorations presented in figure 1 below shows the boxplots and violin graph for the swing states in the elections 2016. These states show a low income that seems homogenously distributed between 40k and 50k. The violin graph shows that the income is spread down the 50k which means low wages. The only states in this graph that present a homogenous spread in the y-axis is Nevada which voted Democrat in the final results.

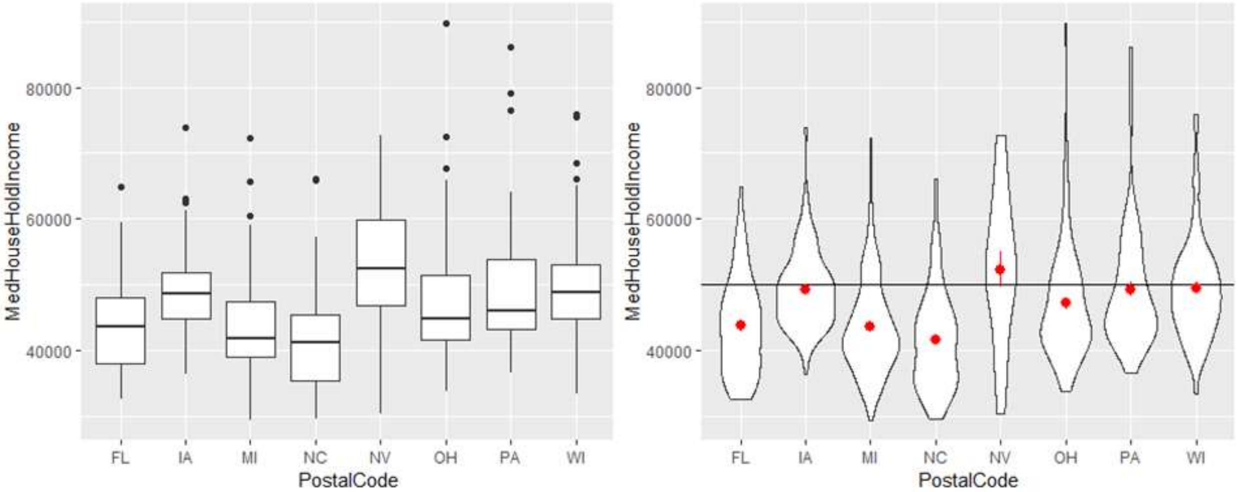

Figure 1. Boxplot on the left and Violin plot on the right of swing states elections 2016.

Since the information presented about swing states seemed too homogenous for the ANOVA test, it was decided to include all the states in the datasets. However, when the graphs were compared among each other the patter seemed alike. It can be seen in the figure 2 below that the graph on all states (right) presents a similar distribution than the graph with just the swing states (left). Each dot represent a county.

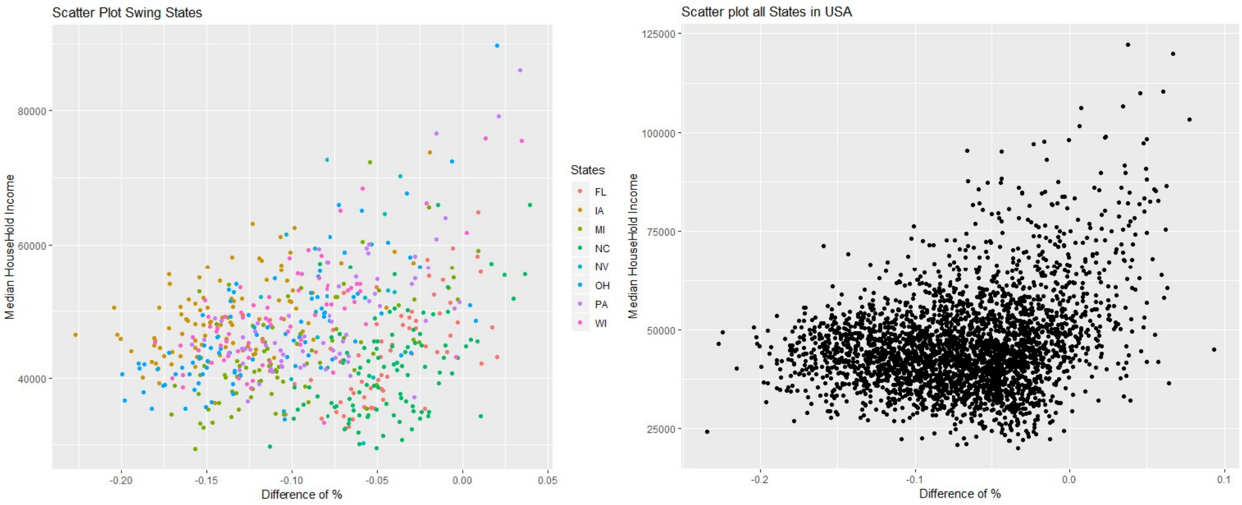

Figure 2. Scatter plot of swing States on the left and all states on the right.

The figure 3 below shows other interesting information. Even thou the dataset includes all the states the salaries are skew to the left, which means that most of the population is below of 50k which is low. In this graph can be seen also that the downturn was general, most counties presented a negative value.

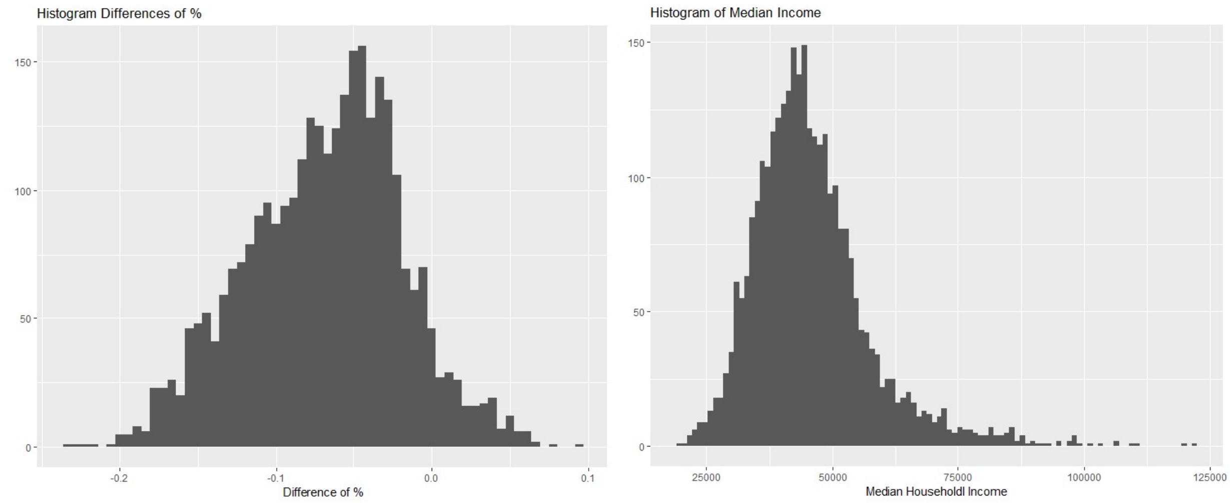

Figure 3. Histograms of differences in % and salaries.

The figure 4 is the last graph presented in the exploratory analysis. This figure presents the distribution by group. Again it can be seen a concentration in the group of 50k. It is important to note that this group consist of households income greater than 30k and below 50k. This range is just above the poverty line, however 50k in income is not a good number also, since it is not income per capita but per household.

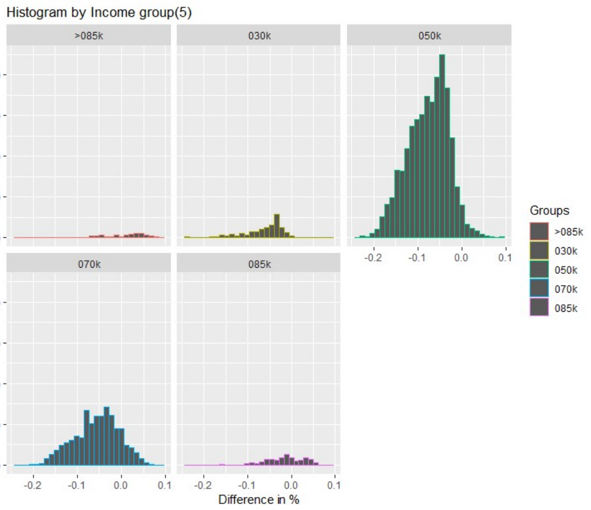

Figure 4. Histogram of salaries by income group.

Measures of Variance and Initial Correlation

The final analysis is the variance a correlation between household income and the differences in % in voting for the Democrat party. It can be observed in the table 4 that the standard deviation is alike with the exception of the70k group. It is important to note that this group had a mean of 56k, which mean that was skewed to the left. The range between the groups present big differences among the 30k,50k, and 70k, meanwhile the standard deviation in income is similar between the groups of 050k and 70k.

Table 4.

Variances of Differences in % of voting and household income.

		*Dif	Dif		Dif	Dif	**Hhi		Hhi	Mhi	Mhi
Income	n	SD	IQR	Dif Min	Max	Range	SD	Hhi IQR	Min	Max	Range
>085k	36	0.0420	0.0662	-0.0660	0.0777	0.1437	9754.502	11414.50	85297	122238	36941
030k	126	0.0417	0.0435	-0.2344	0.0017	0.2361	2443.469	3417.75	19986	30000	10014
050k	2062	0.0467	0.0671	-0.2268	0.0935	0.3203	5173.100	8159.75	30023	50000	19977
070k	727	0.0500	0.0689	-0.2043	0.0634	0.2677	5188.025	7226.50	50027	69965	19938
085k	93	0.0433	0.0658	-0.1595	0.0621	0.2216	4439.947	7848.00	70238	84951	14713

The initial correlation between household income and difference in % of voting is **0.2760845**. Th is number means that is a low correlation, however this analysis tries to present an influence of in come in the turndown of voting in the Democrat party; income is not the only factor. A 0.27 out 1.00 says that ¼ of the turndown could be caused by income. The ANOVA analysis in the next s ection will provide more information.

One-Way ANOVA Analysis

The following is the results of the ANOVA analysis. The results show a high significance level which means that the results are outside of the probability threshold. Income is a factor in turndown in voting in the Democrat party.

```
                   Df Sum Sq Mean Sq F value Pr(>F)
Income_factor_II    4  0.774 0.19341   86.98 <2e-16 ***
Residuals        3039  6.757 0.00222


Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Tukey HSD test was executed in order to determine the contribution of the groups in the results of the ANOVA test. It can be seen in the table 5 below that almost all the groups show a high significance level with the exception of the 030k-070k combination.

Table 5.

Tukey HSD to check correlation of groups

Group	Dif	lwr	Upr	p adj
030k-050k	0.0154436	0.0036327	0.0272545	0.0033463
070k-050k	0.0207574	0.0152060	0.0263088	0.0000000
085k-050k	0.0626748	0.0490312	0.0763184	0.0000000
>085k-050k	0.0895426	0.0679055	0.1111796	0.0000000
070k-030k	0.0053138	-0.0071060	0.0177335	0.7698225
085k-030k	0.0472312	0.0296364	0.0648261	0.0000000
>085k-030k	0.0740990	0.0497763	0.0984216	0.0000000
085k-070k	0.0419174	0.0277436	0.0560913	0.0000000
>085k-070k	0.0687852	0.0468099	0.0907604	0.0000000
>085k-085k	0.0268677	0.0016043	0.0521312	0.0305464

Analysis of Residuals

The following step is to check the normality of the results. The result of the Shapiro-Wilk test below show significance, however the qqplot graph and the histogram graph of residuals in the figure 5 show that the residuals are normal. There can be conclude that the model still is good for the results that is showing, income continues to be an important factor.

```
 Shapiro-Wilk normality test

data:  residuals_model_II
W = 0.99053, p-value = 2.455e-13
```

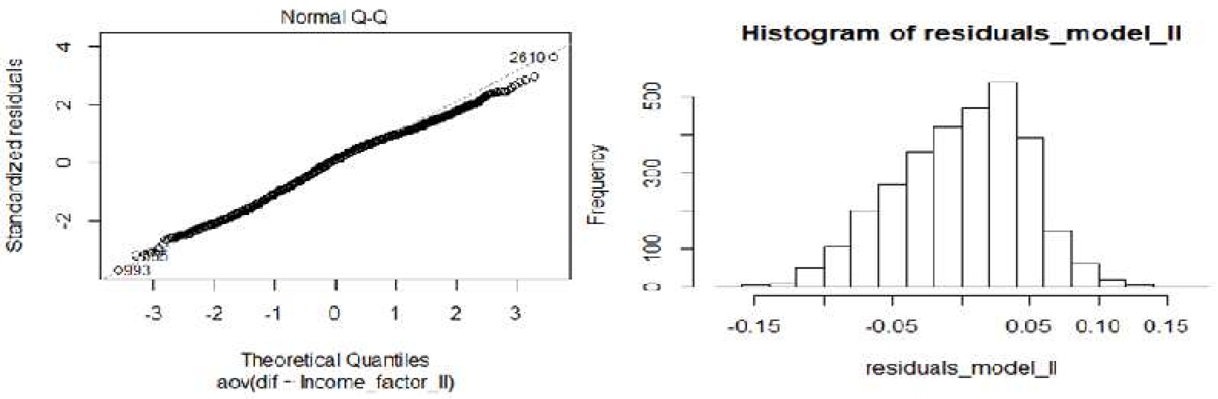

Figure 5. Graphics of qqplot and histogram of residuals.

Homogeneity of variances

In order to check for the homogeneity of variances of the groups involved in the model the Levene test was execute over the ANOVA model. We can observe from the results in the next paragraph that the test gave a significant result which means that the variances are not homogenous. However, the variances can be seen in the figure 6 below that shows a similarity between the variances in width an height. The group that seems out of place is the group of 030k.

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value   Pr(>F)
 group    4  4.5585 0.001135 **
      3039
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

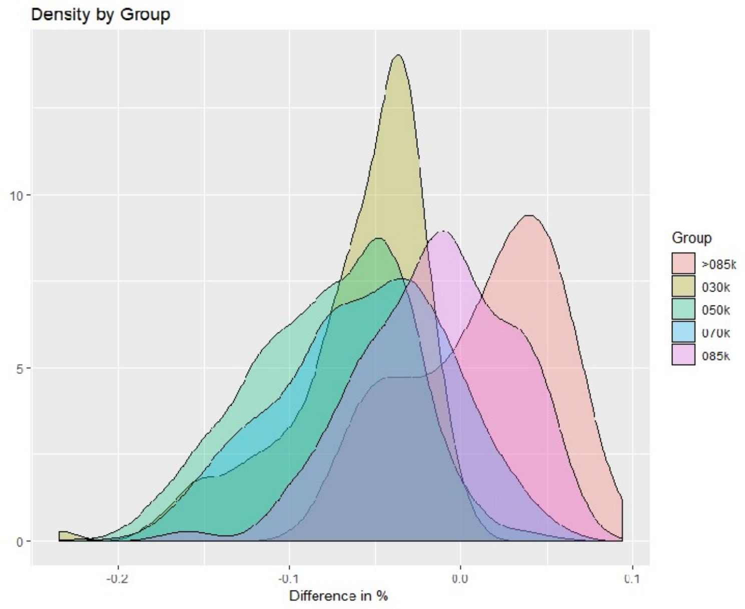

Figure 6. Density of difference in %.

Results

- The residuals show a normal distribution in the histogram graph and the qqplot graph.
- The Living test was statistic significance but the density graph shows a similar density with the exception of the 30k group.
- The ANOVA shows also significance and Tukey HSD between group shows the same, with the exception of 030k- >70k.

Hypothesis

- Ho: is rejected
- H1: is accepted

Conclusion

The analysis shows that income influenced the drop of voting from 2016 and 2012 in the Democrat party.

Additional Observations

- The groups between 50k and 70k seem to be similar.
- It would be interesting to analyze the dataset without 030k, and >085k and dividing the groups 50k and 70k by additional factors, for example, middle class, lower middle class, rural area, urban area.
- This analysis shows that income influence over the results of the voting, but there are other factors associated that would be interesting to look.

- According to Graham, the inequality among different groups is already a problem. Maybe a thorough review of the inequality would give some additional information about the results of the 2016 elections.

- This analysis shows that the groups above the 30k and below 70k show some tendency in the choices made in voting in 2016. It also shows that the group above the 30k and 50k contains the majority of counties which was an unexpected outcome. This information could be a start point to look at inequality and the status of the middle class in the USA in the rural area.

- Graham also states that the increase in salaries and job continues in the more advanced technological industries, so even thou the voting in 2016 show that changes are needed, these changes may not be reaching the places that asked for them.

Bibliography

Digital Properties, LLC. (2018). *STATE ABBREVIATIONS*. Retrieved from 50states.com: https://www.50states.com/abbreviations.htm

Graham, C. (2017, December 1). *The human costs of the productivity paradox in the USA: Insights from metrix of well-being.* Retrieved from Brookings: https://www.brookings.edu/research/the-human-costs-of-the-productivity-paradox-in-the-usa-insights-from-metrics-of-well-being/

Institute of Social Research- University of MIchigan. (2018). *Voting Behavior*. Retrieved from Teaching & Learning with ICPSR: https://www.icpsr.umich.edu/icpsrweb/instructors/setups/voting.jsp

Kaggle. (2018). *Dataset Voting 2016-2012*. Retrieved from Kaggle: https://www.kaggle.com/joelwilson/2012-2016-presidential-elections

Long, H. (2017, May 16). *Trump voters want jobs. Not noise about Russia*. Retrieved from CNN Money: https://money.cnn.com/2017/05/16/news/economy/donald-trump-voters-jobs-russia/index.html

Price, M. S. (2018, July 19). *The new gilded age- Income inequality in the U.S. by state, metropolitan area, and county*. Retrieved from Economic Policy Institute: https://www.epi.org/publication/the-new-gilded-age-income-inequality-in-the-u-s-by-state-metropolitan-area-and-county/

U.S.A. Department of Commerce. (2016). *American Fact Finder*. Retrieved from United States Census Bureau: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_13_5YR_DP03&prodType=table