AIT 580 – Final Project
SOI Individual Income Tax Statistics – 2017

Introduction

The project deals with the data set SOI Individual Income Tax Statistics , this ZIP Code data are based on population data that was filed and processed by the IRS during the 2017 calendar year which has 166537 rows and 153 column 179mb file. Out of which 19 columns like State, zip code, gross income number of single joint returns etc are chosen to conduct the analysis for this paper. The paper conducted analysis to find strength between different variables and to show patterns and trends of each state with spread in income, taxes.[1]

Who is collecting the data?

The publisher is IRS Research, Analysis, and Statistics (RAS) Statistics of Income (SOI) and is maintained by Kevin Pierce. The data is collected by Internal Revenue Service (IRS) and is available at irs.gov/statistics for public access. The ZIP code data for the Income Division Statistics were tabulated using the individual income tax returns submitted with the Internal Revenue Service (IRS) for the 12-month period from 1st January 2018 to 31st December 2018. Annual statistical reporting was mandated by the 1916 Tax Act as it affects individuals. The Data Book of the IRS is issued annually by the IRS, which includes fiscal year-based statistical tables. The study includes statistics on revenue recovery, refunds allocation, law enforcement, public aid, and the budget and workforce.[1]

Need for SOI Data

The data SOI gathers, analyzes, and publishes are utilized by several government departments, researchers and public. Through analyzing this dataset, In the United States we may examine the individual tax returns. We will observe the pattern in tax returns between the US states and zip codes.

Potential questions from the study

- We could analyze the average total income of each state.
- To analyze strength of association between variables.
- Relation between different variables like number of returns, dependents, exemptions etc.?
- Which state has high tax returns?

Privacy

The data collected is an open database that the public can use. There is no privacy. Negative tax returns with gross income are excluded. The limitation contained in the data is that we cannot use this dataset to do policy evaluations. ZIP codes with less than 100 returns is classified as 99999.

Does not include returns with adjusted gross deficit. Due to transparency safety procedures or exclusion from returns, State totals cannot be comparable with State totals reported elsewhere by SOI.[1]

Quality

Certain people may not have the requirement to submit a federal income tax report and this data may not represent the whole U.S. population. There are few missing data columns values in the dataset which are simply represented my zero and the results of these values in the dataset can cause inaccurate predictions and visualizations may vary.[1]

Requirements and Resources Needed

Hardware:

- Processor : Intel Core ™ i5-6200 CPU 2.40 Ghz
- Installed memory (RAM) : 8.00GB (7.90Gb usable)
- System type : 64-bit Operating System, x64-processor
- Windows Edition : Windows 10 Enterprise

Software:

- RStudio is used for Data exploration, transformation, and visualization.
- Tableau is used for visualization.
- Jupyter (Python) for analysis and visualizations.
- MS Excel for Data exploration, Cleaning.

Dataset Description

The dataset is a 196 Mb .csv format file. The data set has 166537 rows and 153 columns of the SOI Individual Income Tax Statistics – 2017.

```
> setwd ("D:/GMU DAE/AIT580/project")
> SOITax <- read.csv ("17zpallagi.csv", sep=",")
> dim (SOITax)
[1] 166537    153
```

The 153 attributes in the dataset are scaled down and I have chosen to work on 19 Variables.

```
> Taxstats <- SOITax [, c (2:20)]
> write.csv (Taxstats, file = "Taxstats.csv")
> dim (Taxstats)
[1] 166537     19

> names(Taxstats)
 [1] "STATE"      "zipcode"    "agi_stub"   "N1"         "mars1"      "MARS2"
 [7] "MARS4"      "ELF"        "CPREP"      "PREP"       "DIR_DEP"    "N2"
[13] "NUMDEP"     "TOTAL_VITA" "VITA"       "TCE"        "VITA_EIC"   "RAC"
[19] "ELDERLY"
```

Metadata

19 different attributes in the dataset are:

Variable	Description	Type	Value
State	The State associated with the ZIP code	Char	Two-digit State abbreviation code "AK","AL","AR"
Zipcode	5-digit Zip code	int	0, 35004, 35004
Agi_stub	Size of adjusted gross income	int	1 = $1 under $25,000 2 = $25,000 under $50,000 3 = $50,000 under $75,000 4 = $75,000 under $100,000 5 = $100,000 under $200,000 6 = $200,000 or more
N1	Number of returns	num	802640, 499070
Mars1	Number of single returns	num	474470, 218590
Mars2	Number of joint returns	num	99850, 137460
Mars4	Number of head of household returns	num	216600, 129760
ELF	Number of electronically filed returns	num	717050, 448190
Cprep	Number of computer prepared paper returns	num	44090, 26230
Prep	Number of returns with paid preparer's signature	num	269560, 156410
Dir_dep	Number of returns with direct deposit	num	580390, 365010
N2	Number of exemptions	num	1259760, 985860
Numdep	Number of dependents	num	475750, 352150
Total_vita	Total number of volunteer prepared returns	num	25570, 11550
Vita	Number of volunteer income tax assistance (VITA) prepared returns	num	17310, 7570
Tce	Number of tax counseling for the elderly (TCE) prepared returns	num	8250, 3980
Vita_etc	Number of volunteer prepared returns with Earned Income Credit	num	5670, 370, 0
Rac	Number of refund anticipation check returns	num	212170, 120760
Elderly	Number of elderly returns	num	150660, 112510

Descriptive Statistics in R

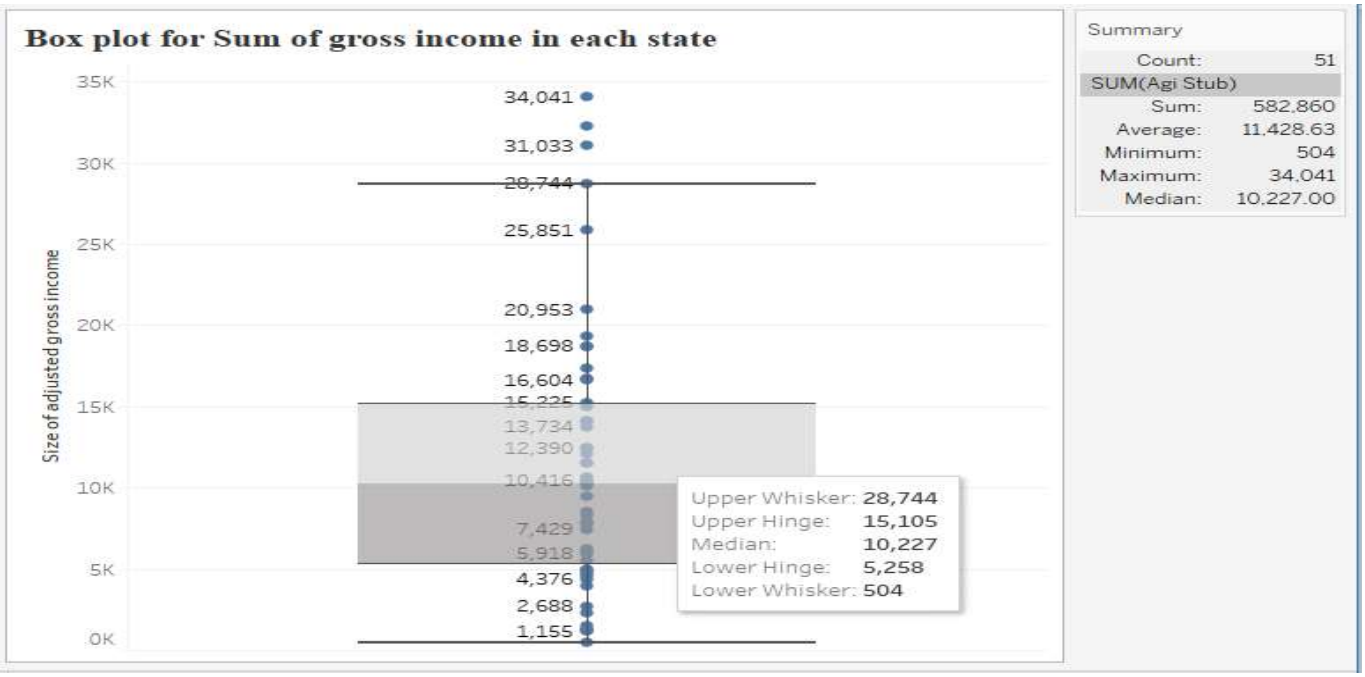
```
STATE           zipcode         agi_stub         N1                mars1
TX     :  9726   Min.   :    0   Min.   :1.0   Min.   :      0   Min.   :      0
NY     :  9222   1st Qu.:27030   1st Qu.:2.0   1st Qu.:     70   1st Qu.:      0
CA     :  8867   Median :48876   Median :3.0   Median :    250   Median :     70
PA     :  8213   Mean   :48870   Mean   :3.5   Mean   :   1798   Mean   :    857
IL     :  7386   3rd Qu.:70601   3rd Qu.:5.0   3rd Qu.:   1020   3rd Qu.:    340
OH     :  5987   Max.   :99999   Max.   :6.0   Max.   :5824360   Max.   :4130490
(Other):117136
```

```
MARS2              MARS4              ELF              CPREP
Min.   :      0.0   Min.   :     0.0   Min.   :      0   Min.   :     0.0
1st Qu.:     40.0   1st Qu.:     0.0   1st Qu.:     60   1st Qu.:     0.0
Median :    110.0   Median :    20.0   Median :    220   Median :     0.0
Mean   :    646.6   Mean   :   257.4   Mean   :   1595   Mean   :   109.2
3rd Qu.:    380.0   3rd Qu.:    90.0   3rd Qu.:    910   3rd Qu.:    60.0
Max.   :1757700.0   Max.   :982390.0   Max.   :4980210   Max.   :475240.0

PREP             DIR_DEP             N2              NUMDEP
Min.   :      0   Min.   :      0   Min.   :      0   Min.   :      0
1st Qu.:     40   1st Qu.:     40   1st Qu.:    150   1st Qu.:     40
Median :    150   Median :    140   Median :    520   Median :    150
Mean   :    957   Mean   :   1115   Mean   :   3450   Mean   :   1120
3rd Qu.:    560   3rd Qu.:    590   3rd Qu.:   2040   3rd Qu.:    590
Max.   :3387570   Max.   :3233490   Max.   :8578160   Max.   :3338670

TOTAL_VITA            VITA              TCE            VITA_EIC
Min.   :     0.00   Min.   :     0.00   Min.   :    0.00   Min.   :    0.00
1st Qu.:     0.00   1st Qu.:     0.00   1st Qu.:    0.00   1st Qu.:    0.00
Median :     0.00   Median :     0.00   Median :    0.00   Median :    0.00
Mean   :    37.66   Mean   :    18.86   Mean   :   18.83   Mean   :    4.71
3rd Qu.:     0.00   3rd Qu.:     0.00   3rd Qu.:    0.00   3rd Qu.:    0.00
Max.   :183220.00   Max.   :109160.00   Max.   :74070.00   Max.   :43090.00

RAC              ELDERLY
Min.   :     0.0   Min.   :     0.0
1st Qu.:     0.0   1st Qu.:    30.0
Median :    30.0   Median :    70.0
Mean   :   251.7   Mean   :   431.8
3rd Qu.:   100.0   3rd Qu.:   270.0
Max.   :993600.0   Max.   :1003170.0
```

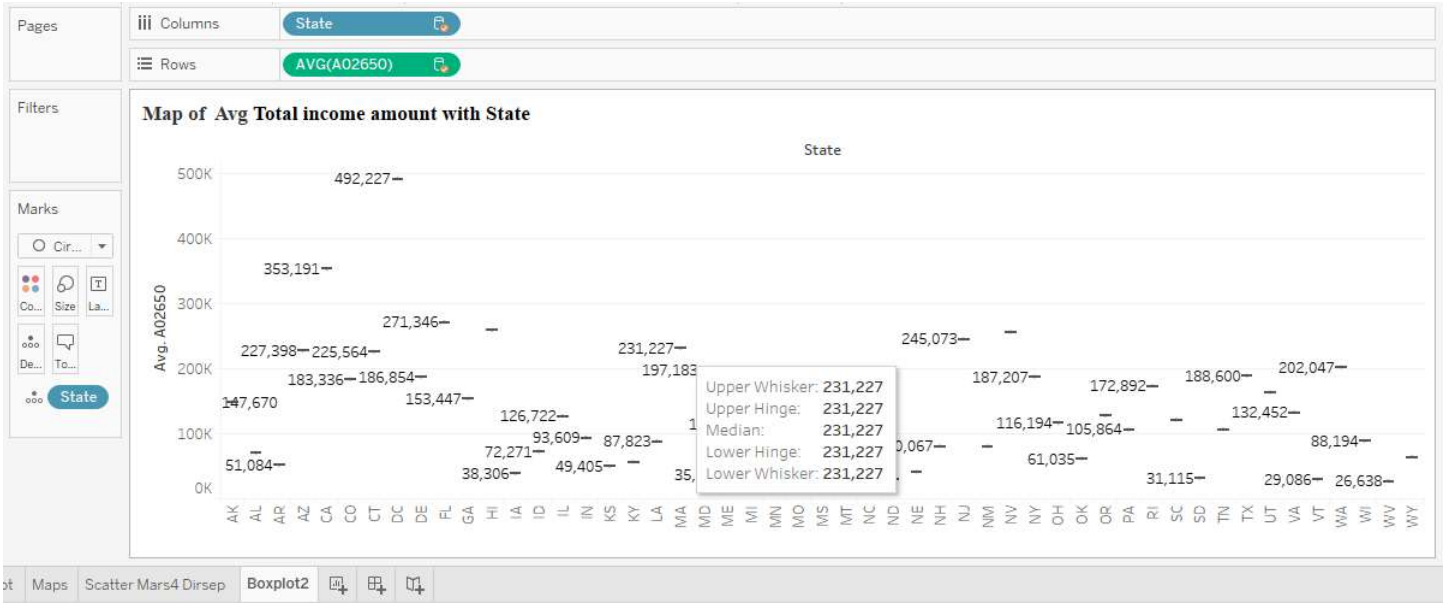
Findings

Box Plots

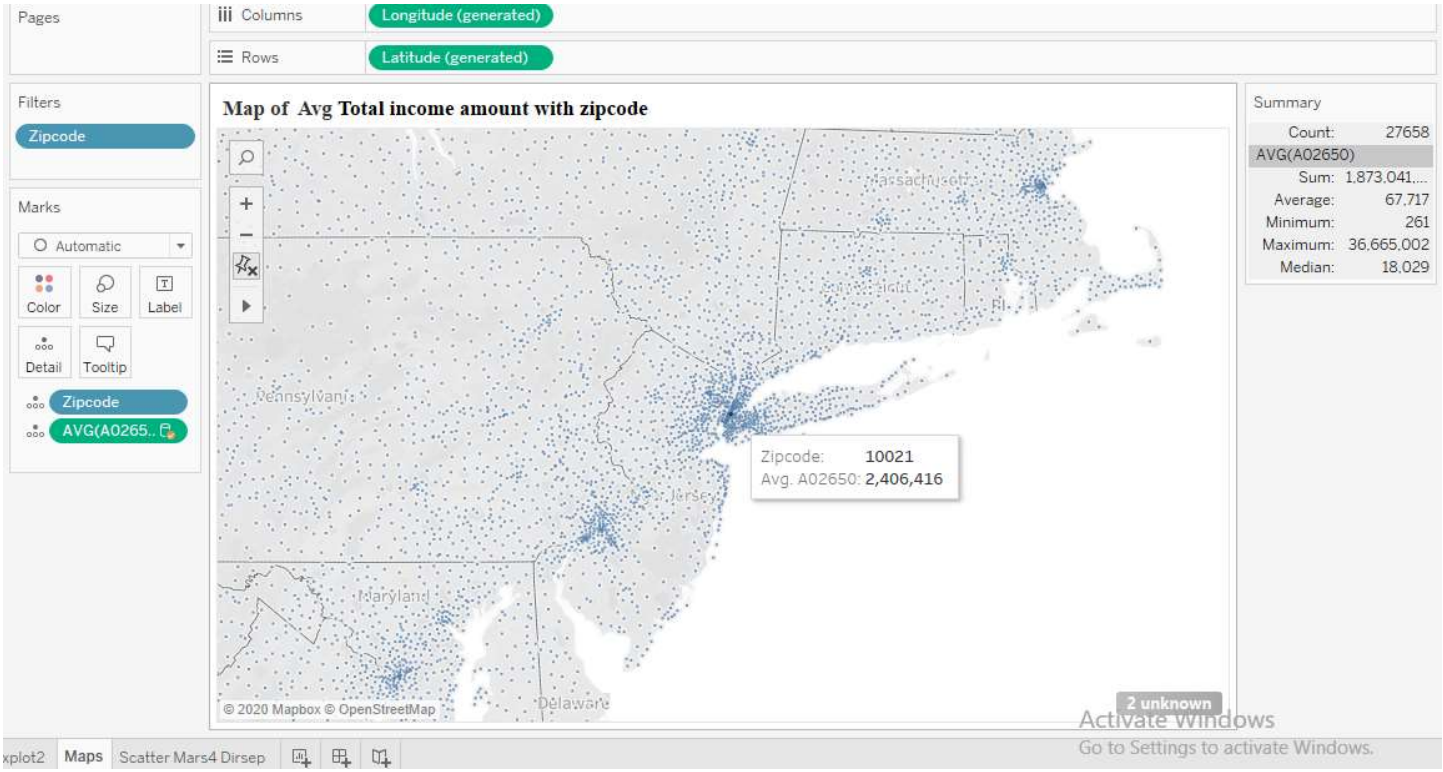
- Figure1: Boxplot shows Texas has the highest sum of Size of adjusted gross income (AGI_STUB) with 34,041.



- Figure 2: DC has the average total income amount (A02650) of 492,227 in the year 2017.



- Figure 3: Zip code with 10021 which is NYC Lower Manhattan has average total income amount (A02650) of 2,406,416.



Scatter Plots

- This scatter plot gives the relationship between the number of head of household returns (MARS4) and the number of returns with direct deposit (DIR_DEP). These variables have a linear relation.

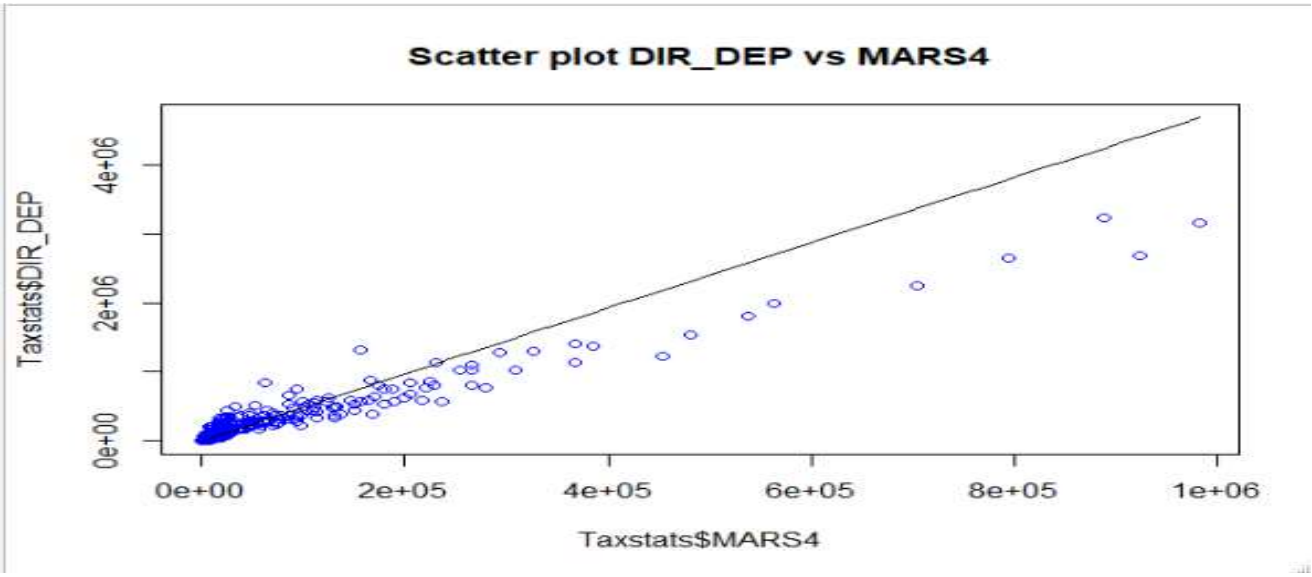


Figure 4: MARS4 and Dir_dep Scatter Plot in R

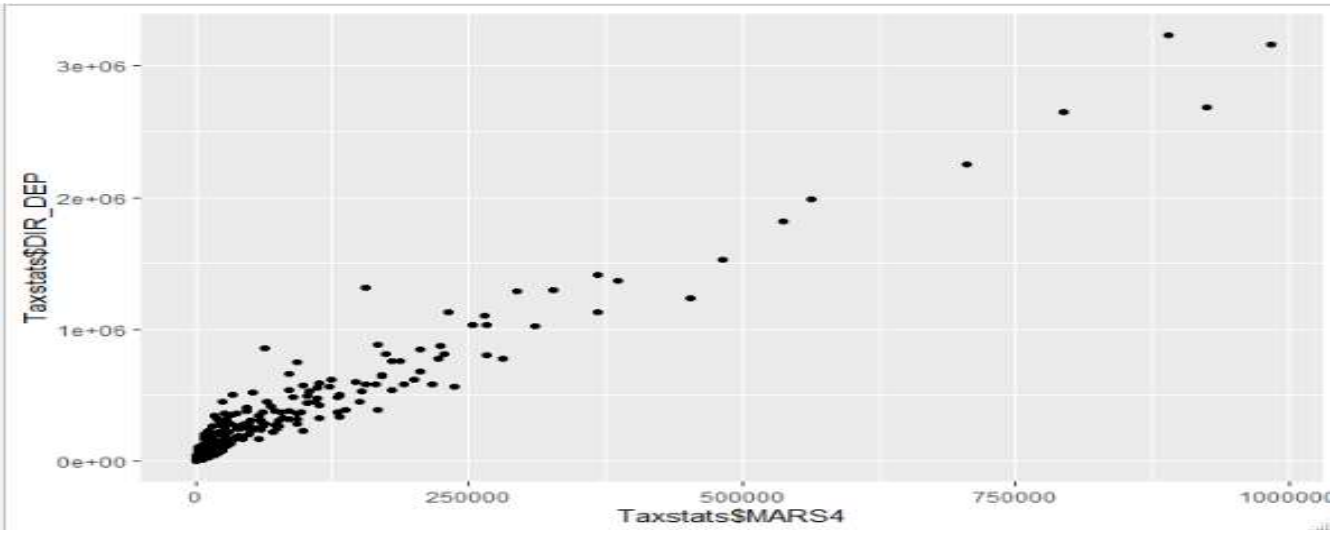


Figure 4: qplot for MARS4 and Dir_dep in R

Correlation analysis

- Correlation test is performed between MARS4 and DIR_DEP. It tests for the strength of the association between two continuous variables.

Pearson's product-moment correlation

```
data:  Taxstats$MARS4 and Taxstats$DIR_DEP
t = 1683.6, df = 166535, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9715887 0.9721218
sample estimates:
      cor
0.9718565
```

- Variables have a significant correlation coefficient of 0.9718565 and p-value < 2.2e-16.

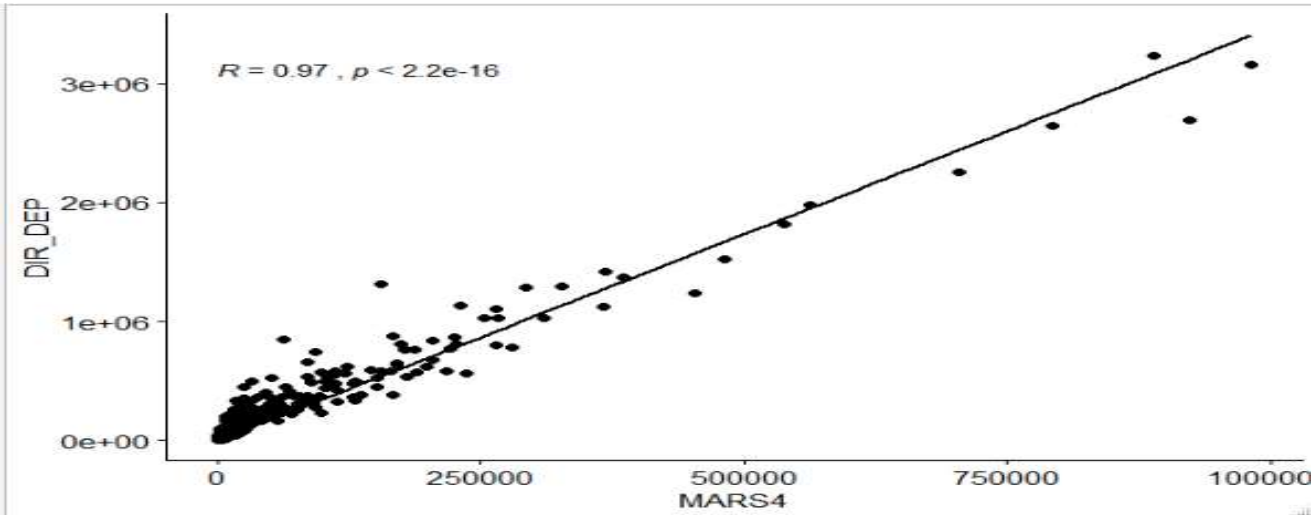


Figure 5: MARS4 and DIR_DEP Correlation plot

Regression Analysis

Hypothesis test

- The process to determine whether to reject a null hypothesis, based on sample data is calle d hypothesis test. Hypothesis Testing is used to determine if the probability of the given var iables is true. We reject null hypothesis and accepts an alternative hypothesis as P-VALUE is very less we can observe that the value of p is < 2.2e-16.

```
Residuals:
    Min      1Q    Median    3Q     Max
-203738       2       33       46   190856
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.642e+01  3.802e+00  -12.21   <2e-16 ***
DIR_DEP      2.724e-01  1.618e-04 1683.56   <2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1550 on 166535 degrees of freedom
Multiple R-squared:  0.9445,     Adjusted R-squared:  0.9445
F-statistic: 2.834e+06 on 1 and 166535 DF, p-value: < 2.2e-16
```

Welch Two Sample t-test

- A two-sample location check that is used to evaluate the hypothesis that two populations h ave equal means.

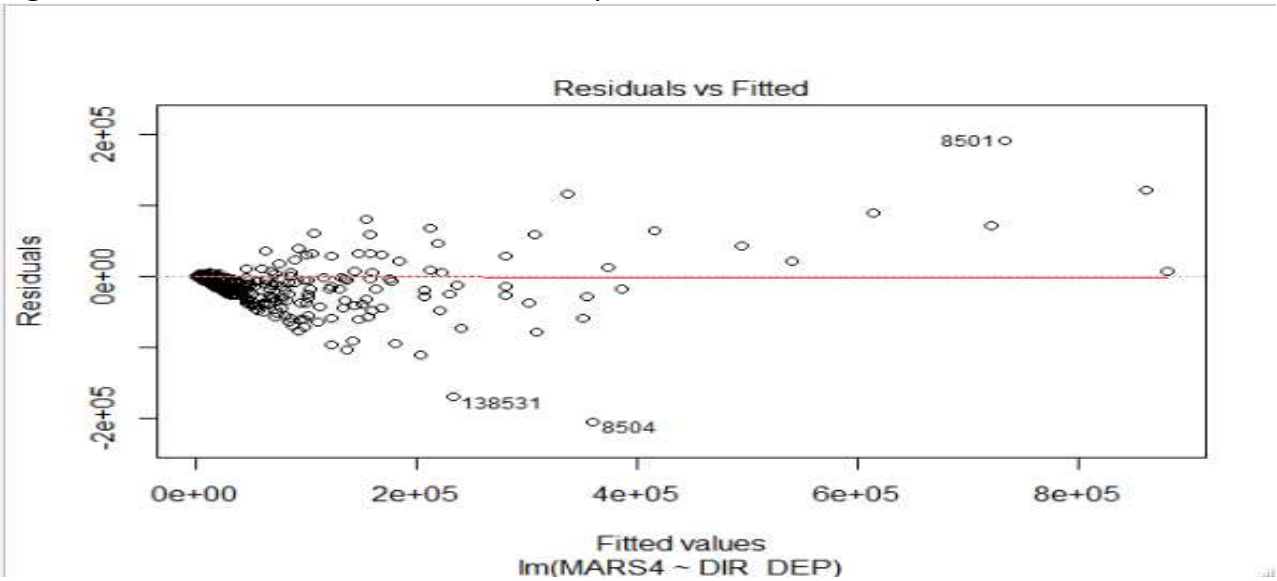
```
data:  Taxstats$MARS4 and Taxstats$DIR_DEP
t = -14.362, df = 192544, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -975.0010 -740.8337
sample estimates:

mean of x mean of y
 257.3954 1115.3127
```

Pearson's Chi-squared test

- Tests for the strength of the association between two categorical variables

```
data:  Taxstats$MARS4 and Taxstats$DIR_DEP
X-squared = 65167186, df = 949351, p-value < 2.2e-16
```

Figure 6: observe a linear line at 0. The points 8501, 138531 and 8504 are the outliers.



Linear Regression Plot:

Linear regression is performed between the number of head of household returns (MARS4) and the number of returns with direct deposit (DIR_DEP).

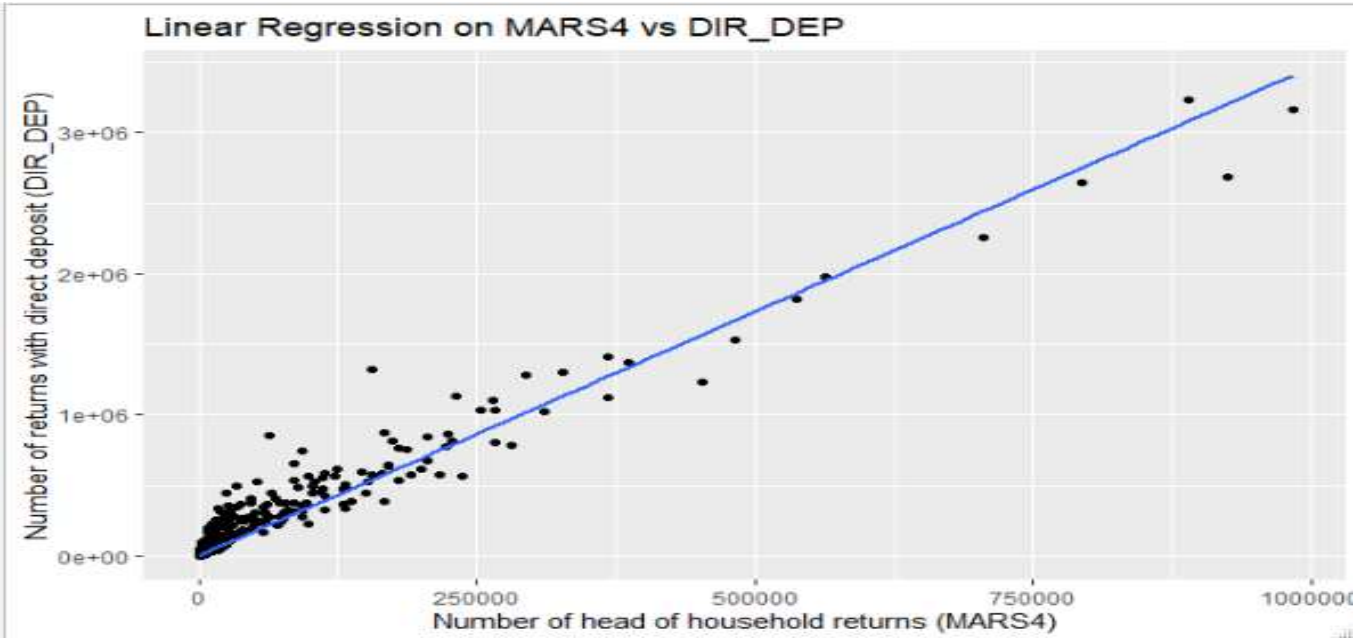


Figure 7: Linear regression plot

Clustering analysis

The different types of data available in the data set are classified into clusters using clustering. k-means clustering is a method of vector quantization, but k-means is a crude heuristic. We have used elbow criterion method which is a visual method. The analysis is done on N1, Mars1 and Mars2 variables.

Elbow method is to run k-means clustering on SOI Individual Tax Statistics-2017 for a range of values of k and for each value of k we calculate the sum of squared errors (SSE).

K-Means clustering - Elbow criterion

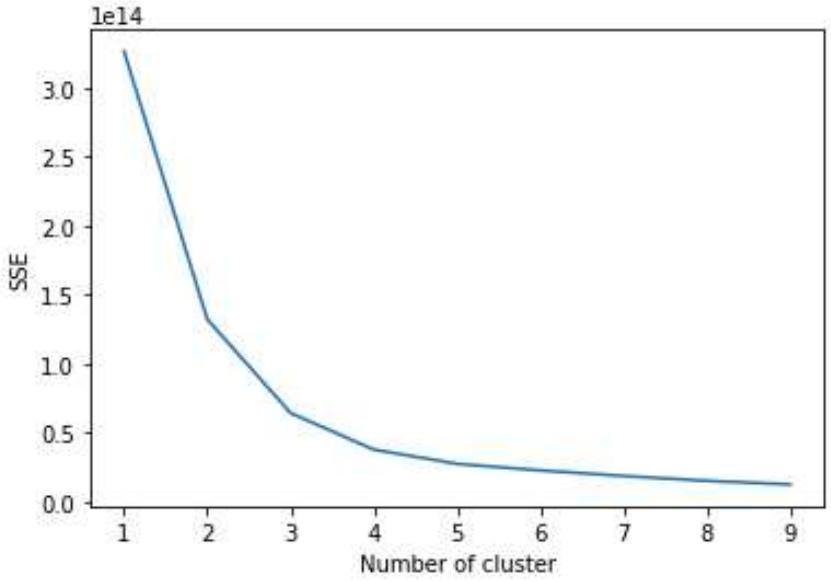


Figure 8: Above line graph, the "elbow" on the arm is the value of optimal k clusters. We must select the small value of k that still has a low SSE.

A higher Silhouette Coefficient score relates to a model with better-defined clusters

For n_clusters=2, The Silhouette Coefficient is 0.9985601768386141
For n_clusters=3, The Silhouette Coefficient is 0.9977758593012598
For n_clusters=4, The Silhouette Coefficient is 0.9968426238930497
For n_clusters=5, The Silhouette Coefficient is 0.9956175345035024
For n_clusters=6, The Silhouette Coefficient is 0.995616742698524
For n_clusters=7, The Silhouette Coefficient is 0.9940027975238743
For n_clusters=8, The Silhouette Coefficient is 0.9939783520229314
For n_clusters=9, The Silhouette Coefficient is 0.9927665802771889
For n_clusters=10, The Silhouette Coefficient is 0.9941384253978325

Number of returns N1, single returns Mars1, and joint returns Mars2 variables where Optimal number of cluster is 3. So, choosing *n_clusters=3* is the optimal number of cluster since we have 3 variables in the dataset.

Results

- Boxplot shows Texas has the highest sum of Size of adjusted gross income (AGI_STUB) with 34,041.
- DC has the average total income amount (A02650) of 492,227 in the fiscal year 2017.
- Zip code with 10021 which is NYC Lower Manhattan has average total income amount (A02650) of 2,406,416.
- The relationship between the number of head of household returns (MARS4) and the number of returns with direct deposit (DIR_DEP) is the significant correlation coefficient of 0.9718565. These variables have a linear relation.
- The analysis done on Number of returns N1, single returns Mars1, and joint returns Mars2 variables where Optimal cluster is 3 and Silhouette Coefficient is 0.997775

Explain/define terms

a. **linear regression:** This is an approach to model the relationship between a dependent variable and one or more independent variables.
b. **Clustering:** It a technique to group similar observations into several clusters based on the observed values of several variables for individual.
c. **Silhouette:** a method of interpretation and validation of consistency within clusters of data. The silhouette ranges from −1 to +1.

Reference

[1] 17zpallagi, SOI. Statistics of Income-2017. Accessed https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2017-zip-code-data-soi Oct 31, 2019