Pat Foley, Ty Helfrich, and Tom Hollerbach
Dr. Levkoff
ECON-494
April 18th, 2021

Descriptive Analytics Data Mining (Part 1)

Executive Summary:

Our group decided to create our own collection of data for the project. One of our group members works for a national chiropractic franchise and has access to several data points on every office in the company. We decided to tailor our project around key data points from every office in an effort to translate our in-class learning into a real-world scenario.

The company uses data collecting software that shows the real-time numbers of the offices. From this software, we were able to transfer data on the offices' monetary stats, (Total Income) and patient-related stats (Chiropractic visits, New Patient visits, Patient referrals) onto a data table. We then used outside sources to find the population of the city that the office is located in. For our categorical data we decided to use the state for which the office is located in, the year that the office opened, and the specific ownership model.

We wanted to use a time frame that was long enough to have an accurate look at what the clinics can do, while not being too long that the data went into the shutdowns during the COVID-19 pandemic. We ultimately settled in on using data from Quarter 1 of 2021 (January 1-March 31).

We created the table with a tidy format in mind. We kept three things in mind while creating our data table.

1. Every Column is a Variable
2. Every Row is an Observation
3. Every Cell has a Single Value

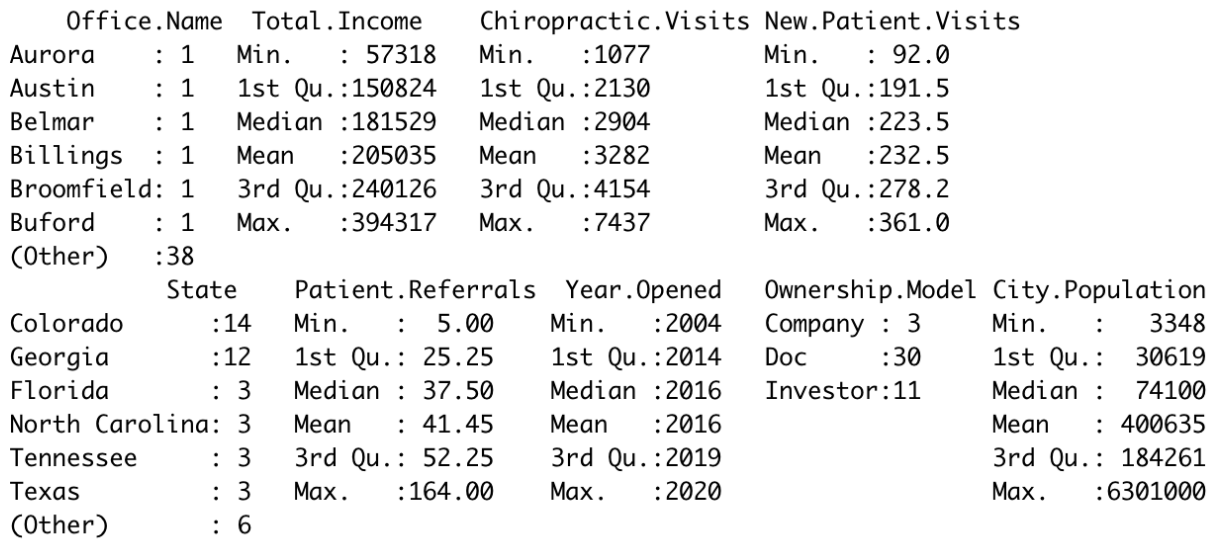
We dive deeper into the cleaning/creating of the data in the second section of this report.

Our data points of interest are the total income of the offices, and what factors contribute to the deviation in income between offices. What contributes to the difference between an office making almost $400,000 a quarter and an office making less than $100,000. Is it the location? Is it the ownership model? Is it based on the population of the city that the office is located? Questions that we will hypothesize and evaluate in portion three of this project.

Cleaning the Data:

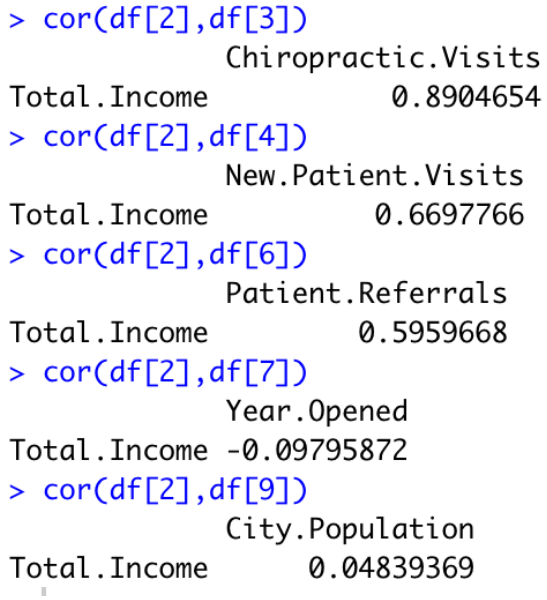
As stated in the executive summary, our data was sourced from one of our group member's national chiropractic franchise internship. Over the course of our project, the necessity of cleaning our data was not required to a full extent. With our three categorical variables of the state for which the office is located in, the year that the office opened, and the specific ownership model, there was not much cleaning of data that had to be done. One issue we struggled with was with the total income variable not being able to be converted in R studio. We were able to fix this by changing the variable's cell formatting from numeric to automatic. We also converted the ownership model to a binary variable by making the offices owned by investors or the company a 0 and offices owned by the doctor a 1. The following sections break down how we went about organizing the data and the correlations we found.

Exploratory Analysis:

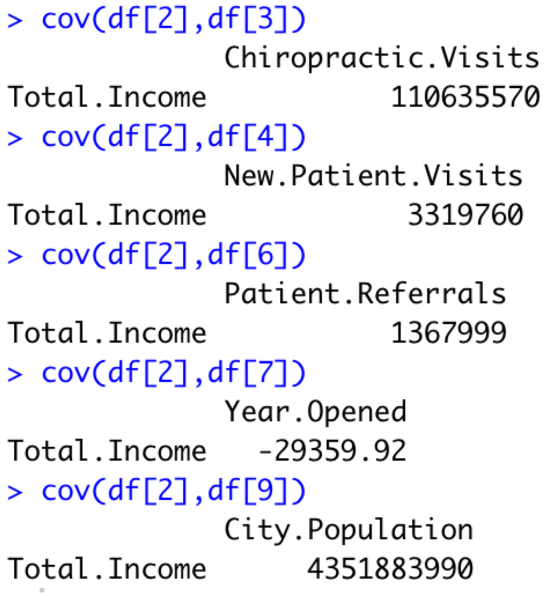


```
    Office.Name  Total.Income    Chiropractic.Visits New.Patient.Visits
 Aurora    : 1   Min.   : 57318   Min.   :1077        Min.   : 92.0     
 Austin    : 1   1st Qu.:150824   1st Qu.:2130        1st Qu.:191.5     
 Belmar    : 1   Median :181529   Median :2904        Median :223.5     
 Billings  : 1   Mean   :205035   Mean   :3282        Mean   :232.5     
 Broomfield: 1   3rd Qu.:240126   3rd Qu.:4154        3rd Qu.:278.2     
 Buford    : 1   Max.   :394317   Max.   :7437        Max.   :361.0     
 (Other)   :38                                                          
            State    Patient.Referrals  Year.Opened   Ownership.Model City.Population
 Colorado      :14   Min.   :  5.00    Min.   :2004   Company : 3     Min.   :   3348
 Georgia       :12   1st Qu.: 25.25    1st Qu.:2014   Doc     :30     1st Qu.:  30619
 Florida       : 3   Median : 37.50    Median :2016   Investor:11     Median :  74100
 North Carolina: 3   Mean   : 41.45    Mean   :2016                   Mean   : 400635
 Tennessee     : 3   3rd Qu.: 52.25    3rd Qu.:2019                   3rd Qu.: 184261
 Texas         : 3   Max.   :164.00    Max.   :2020                   Max.   :6301000
 (Other)       : 6                                                                  
```

In the summary of our data provided above, we can begin to better visualize our data with simple statistics across all variables. These simple statistics include the minimum, maximum, mean, and median of all numeric variables. Our main variable of interest is total income, so the first step we took was to analyze this variable's distribution.

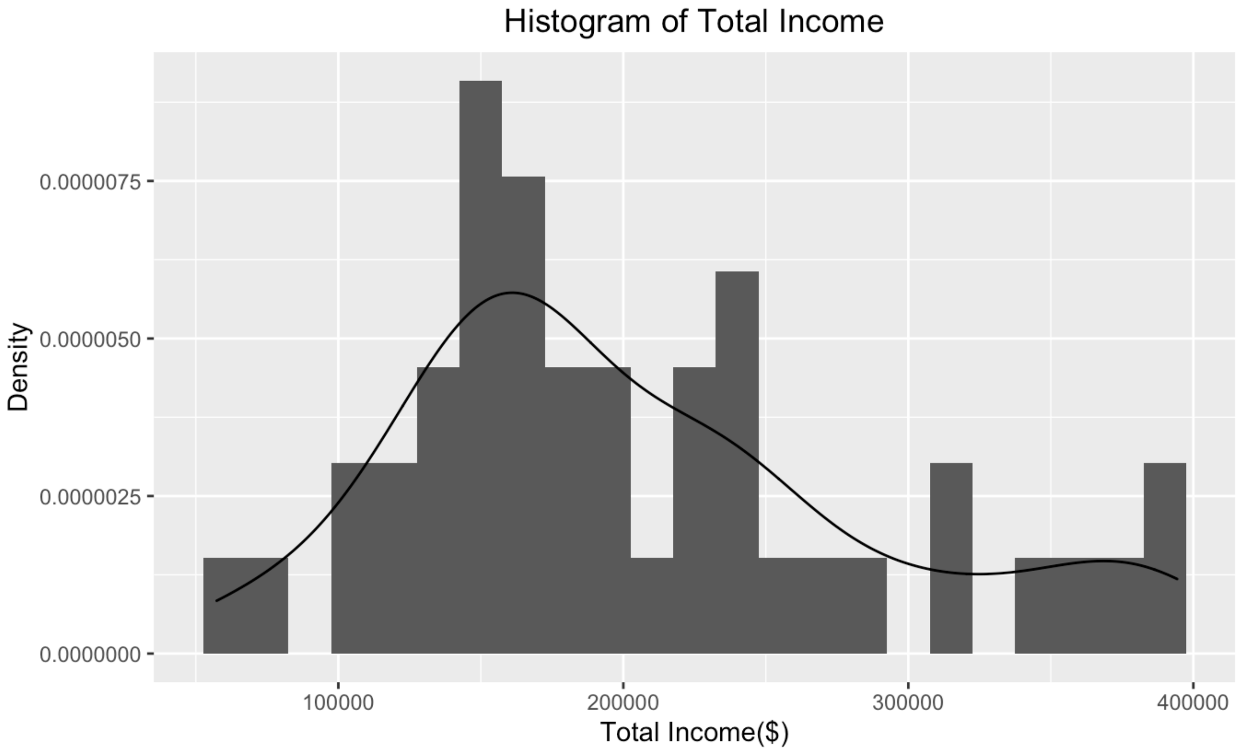


```
> cor(df[2],df[3])
             Chiropractic.Visits
Total.Income           0.8904654
> cor(df[2],df[4])
             New.Patient.Visits
Total.Income          0.6697766
> cor(df[2],df[6])
             Patient.Referrals
Total.Income         0.5959668
> cor(df[2],df[7])
             Year.Opened
Total.Income -0.09795872
> cor(df[2],df[9])
             City.Population
Total.Income      0.04839369
```

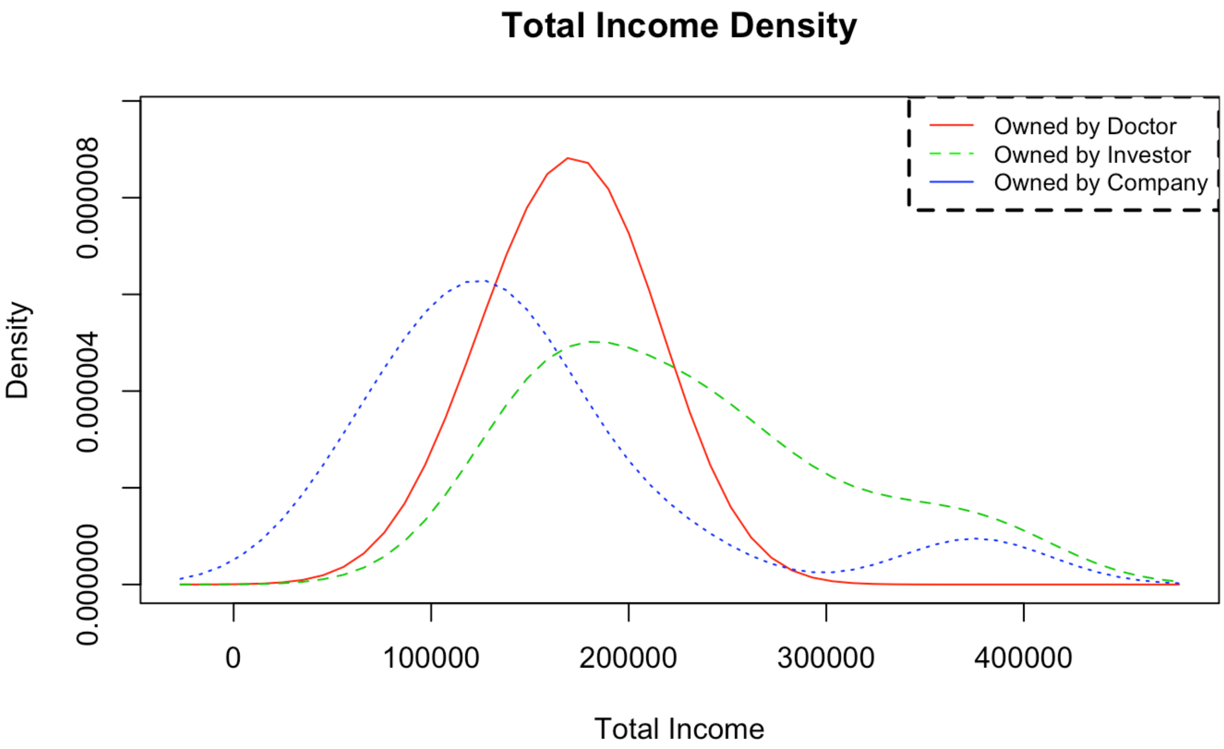


```
> cov(df[2],df[3])
             Chiropractic.Visits
Total.Income           110635570
> cov(df[2],df[4])
             New.Patient.Visits
Total.Income            3319760
> cov(df[2],df[6])
             Patient.Referrals
Total.Income           1367999
> cov(df[2],df[7])
             Year.Opened
Total.Income   -29359.92
> cov(df[2],df[9])
             City.Population
Total.Income      4351883990
```

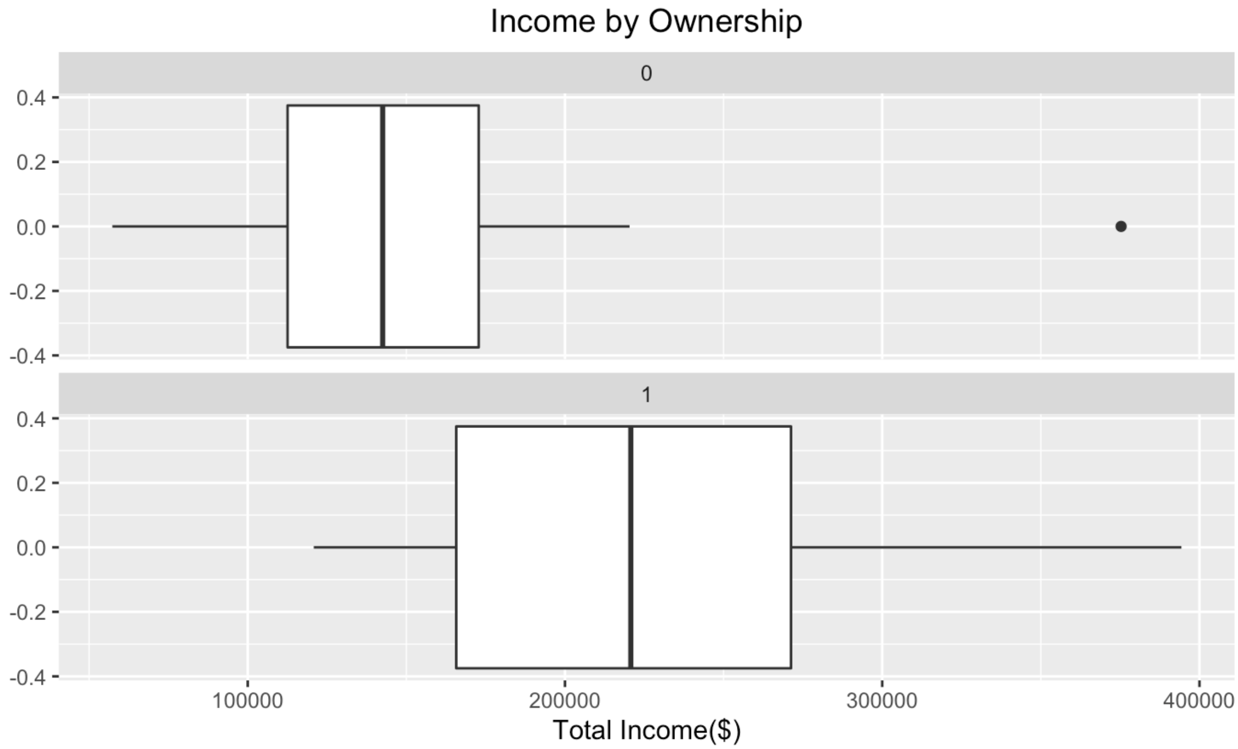
Our next step was to look into the relationships of our numeric variables with the total income. As expected, chiropractic visits, new patient visits, and patient referrals had the highest correlation with total income. With a correlation coefficient of 0.89, we decided to further analyze the relationship between total income and chiropractic visits.



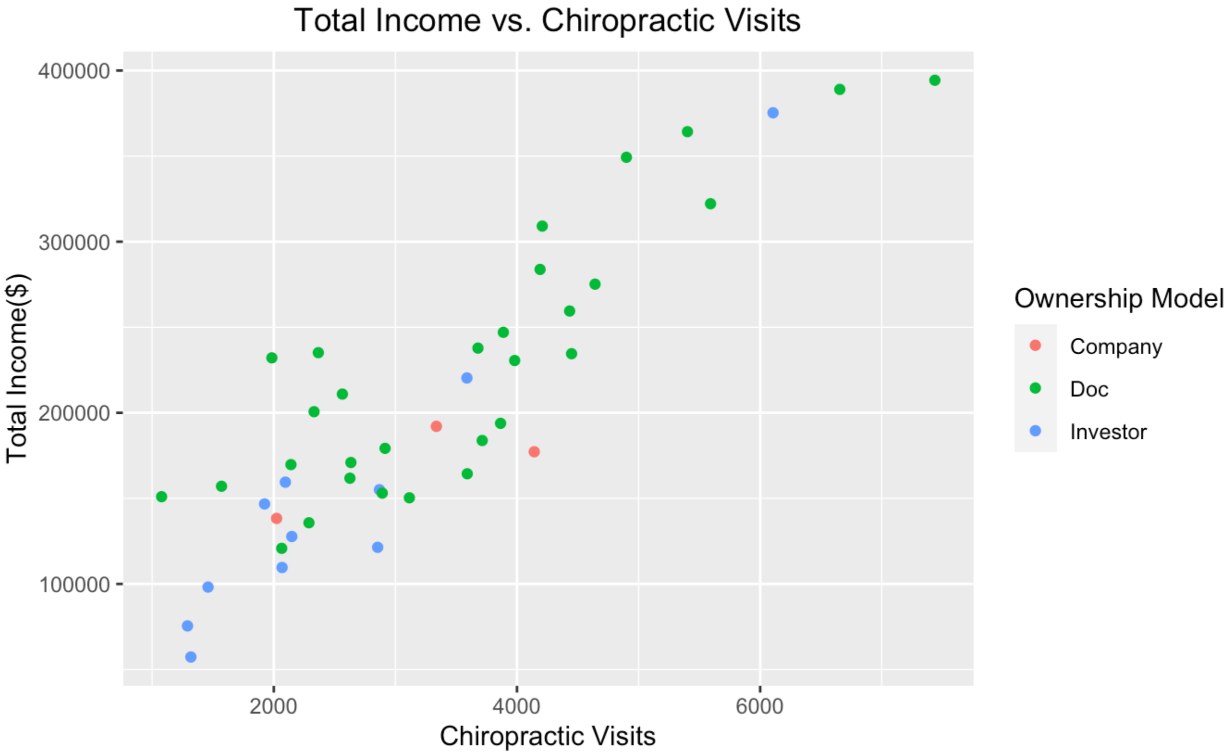
In this histogram, we can see that the data is relatively normally distributed. In the real world, we never expect our data to be a perfectly normal distribution. With only 38 observations, this distribution is much closer to a normal distribution than we originally had expected. As more offices are opened and analyzed across the nation, we would expect to see the total incomes more accurately represent the normal distribution. With more data in the right tail than the left, we can say that the data is right-skewed.



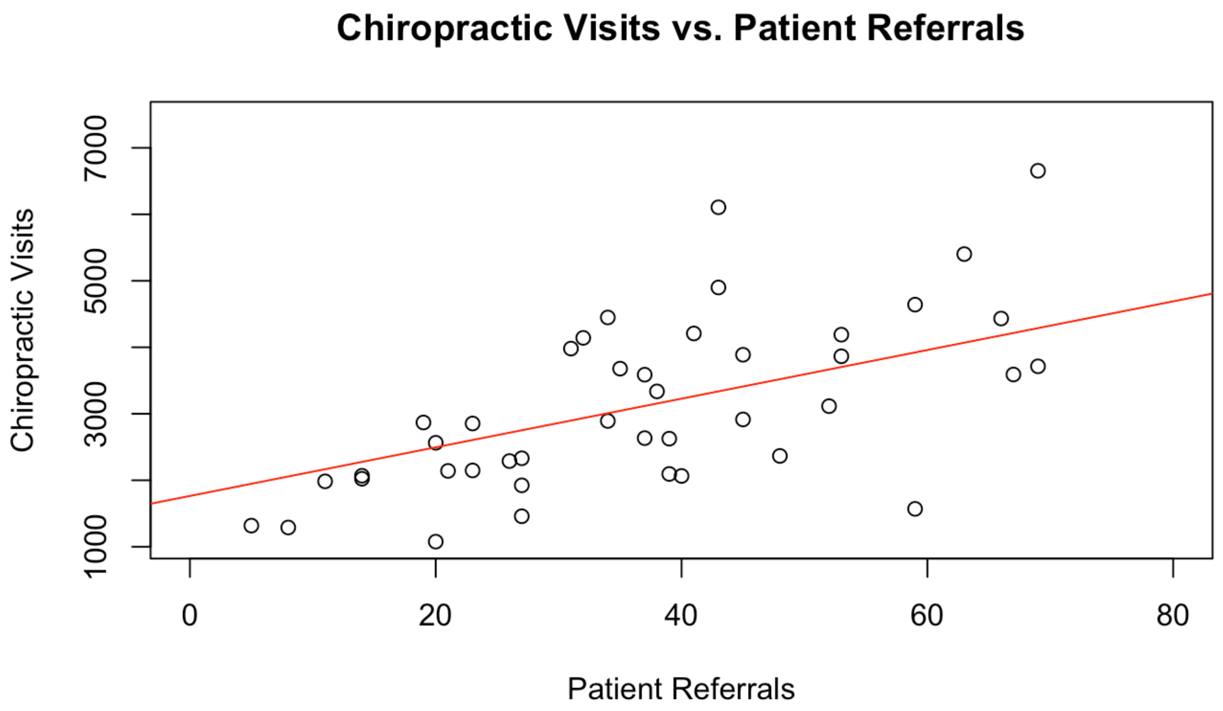
Now that we better understood the distribution of total income on its own, we wanted to investigate how the ownership model may affect the total income. To do so we overlaid the income distributions of each ownership model. By doing so we can see that each ownership model has a relatively similar mean income. Offices owned by the company or investors seem to have comparatively more offices in the upper-income class, but we must not look into this too closely because of the small sample size of each subgroup.



Next, we facet wrapped the total incomes by ownership models to obtain two box and whisker plots for comparison. Here, a 0 represents an office owned by an investor or the company, while a one represents an office owned by the doctor. This comparison very clearly portrays the greater total income of the offices owned by the doctor which may hint at the doctors being better equipped to run their business than an outside investor. On the left, we see one outlier with a significantly larger total income, while on the right side the longer right whisker informs us of more offices in the high-income bracket resulting in a larger mean total income.

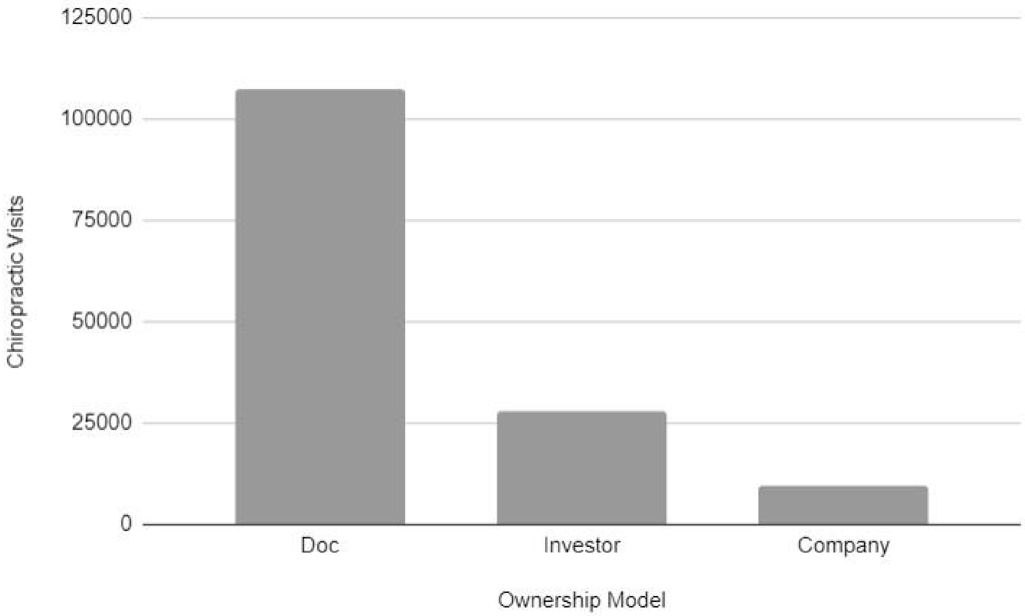


In the scatter plot above, a clear positive relationship can be seen between the number of chiropractic visits and total income. The total income looks as if it begins to plateau around 5,000 visits indicating a non-linear relationship, which we will investigate in part two. Each dot represents a unique office and their color indicates the ownership model which can be seen in the top left corner. This relationship comes as no surprise, but what is interesting to see is the minimal overlap of the data.



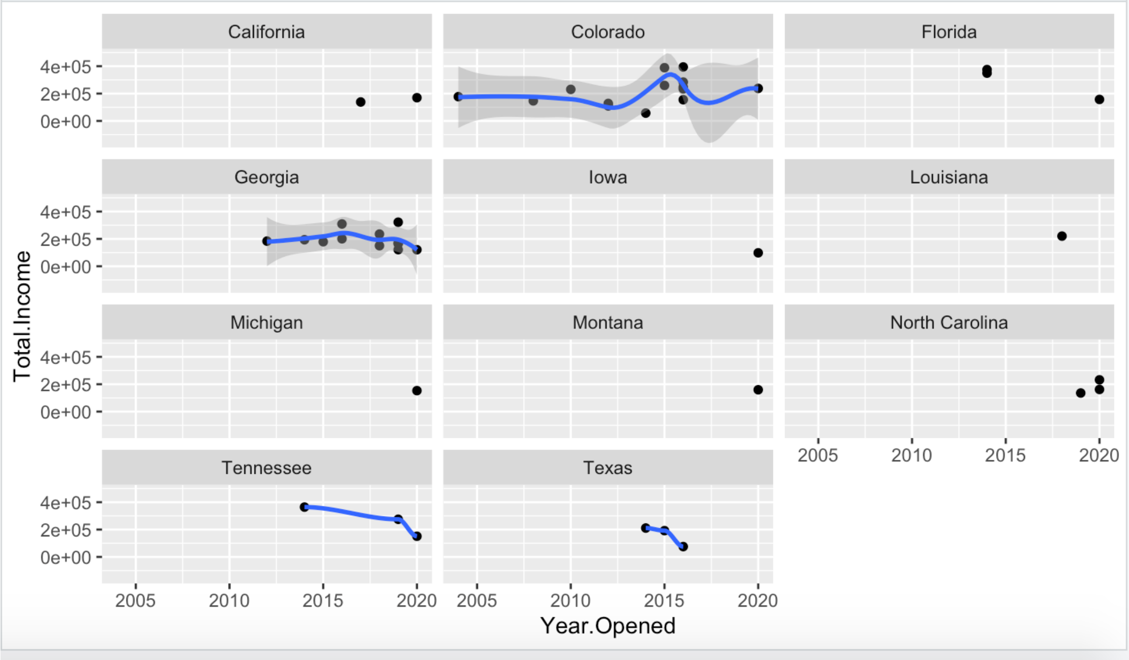
Now that we knew total income and chiropractic visits have a positive linear relationship, we wanted to investigate which variables had the greatest impact on the number of visits. The first logical relationship to examine was between patient referrals and chiropractic visits. We excluded outliers and added a best fit line to see, as expected, another positive linear relationship.

Ownership Model vs. Chiropractic Visits



The final variables we decided to investigate was the ownership model vs Chiropractic visits. As expected, an office received more visits if it was owned under the ownership model of a Doctor compared to an investor or company. The data is aggregated to display the summed amount of all doctor chiro visits, investor visits, and company visits.

Total Income By Year Opened and State



We wanted to see how much of an effect time opened would be for each office based on the state that they are located in. We found that consistently, offices opened between 2014 and 2018 have the highest total incomes, regardless of the state that they are located in. We can assume that the reason offices opened after 2018 did not perform as well last quarter is because they are still gaining a recurring customer base, and this is confirmed by seeing that offices opened after 2018 have a higher new patient number than offices opened before 2018. The interesting information here is that offices opened before 2018 have a lower total income than the other two categories of office. We believe this could have something to do with fatigue, both from a customers point of view and an owners point of view. Another explanation for this could be that the older offices have older equipment and decor which could lead to them losing customers to newer offices.

Part 1 Conclusion:

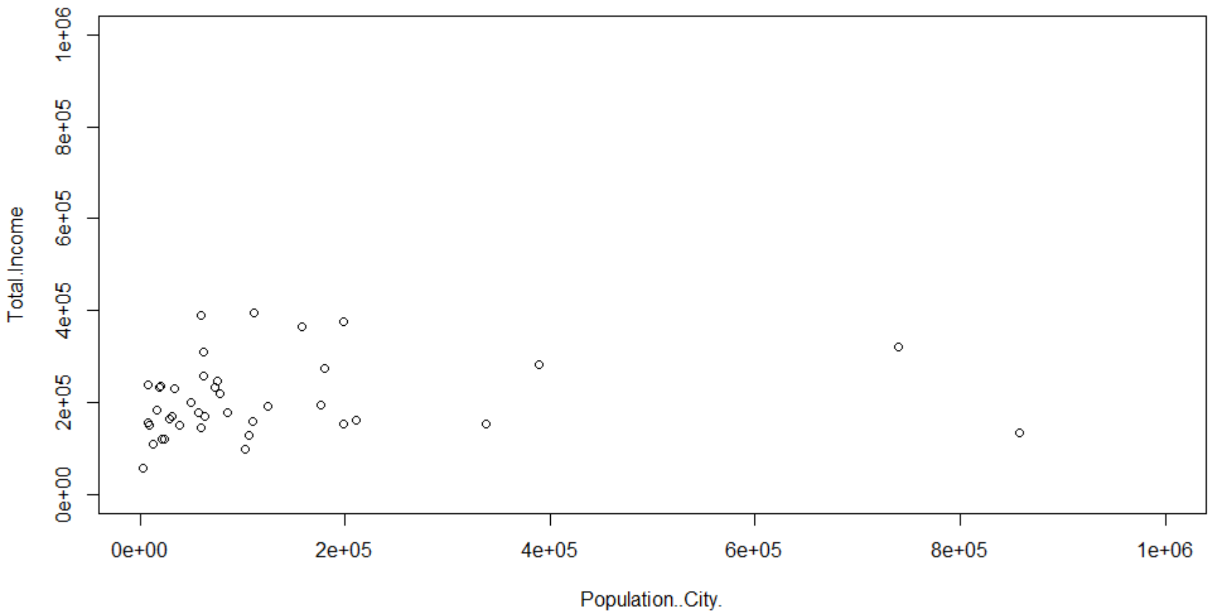
This project has really opened our eyes to the possibilities of statistical analysis in a real world setting. We were able to take raw data, both numerical and categorical, and create a tidy data table, which would turn into a deep statistical analysis that was molded into a detailed descriptive report for a local business. Some of the results were predictable such as chiropractic visits being correlated with total income, and some came as a surprise such as which ownership models were the most successful. The power of descriptive statistics lies in organization, and accuracy. Often, business owners will have a preconceived notion of what is working and what isn't. A statistical analysis like the one that we were able to create this week factors out all bias and opinion, and displays the facts. The first part of our project will be instrumental in the development of our second part of our project, which we assume will be based around predictive and prescriptive statistics. Somewhere in the data, is a perfect equation for an office to follow, we can't wait to dig deeper and solve it. The answer to every question a business has is often sitting in front of that business' face, staring right back like an unsolved puzzle. Only when the business organizes the data and runs analysis is the puzzle solved. There is truth in numbers, we just have to discover it.

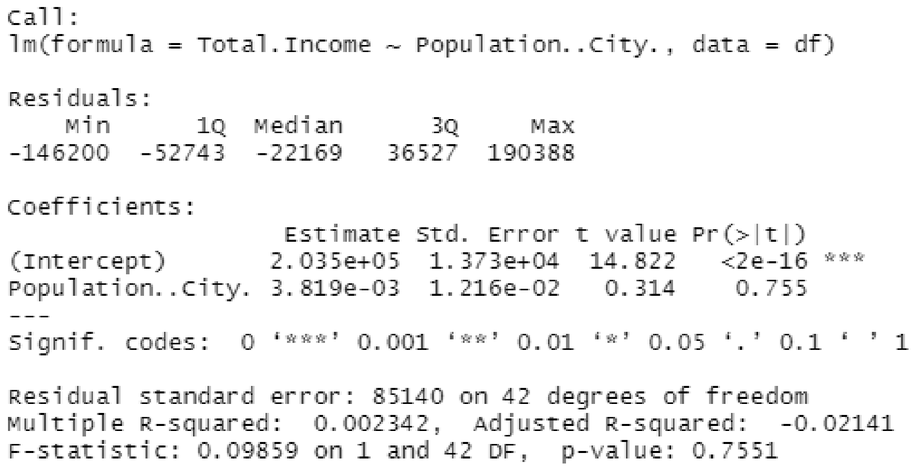
Part 2: Executive Summary

Our data collection process continued from part 1 of the project into part 2. We decided to settle on Total Income being our group's same y-variable across the linear regression models. Out of all independent variables amongst our data set, the total income variable was the best one fitted to portray the relationship and interpretation of our data. The 3 different linear regression models with a single variable we chose to identify were 1) Total Income vs Population of city, 2) Total income vs Chiropractic visits, 3) Total Income vs Patient referrals.

Single Variable Linear Regressions:

Total Income vs. Population of City

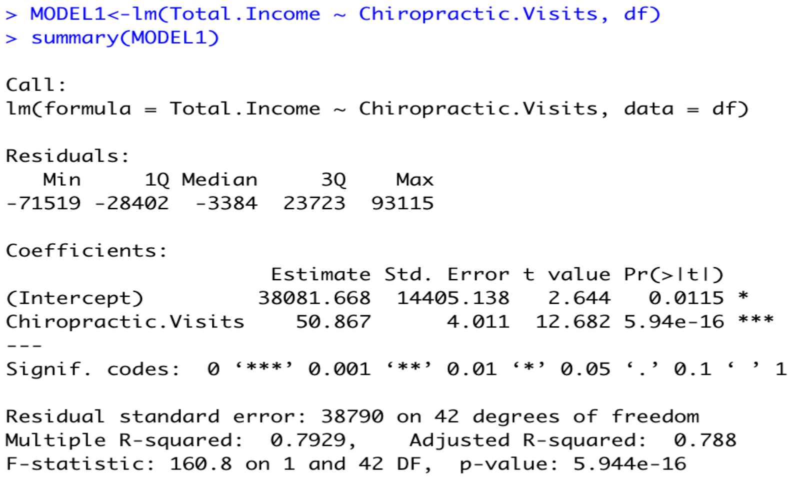




```
Call:
lm(formula = Total.Income ~ Population..City., data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-146200  -52743  -22169   36527  190388 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.035e+05  1.373e+04  14.822   <2e-16 ***
Population..City. 3.819e-03  1.216e-02   0.314    0.755    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85140 on 42 degrees of freedom
Multiple R-squared:  0.002342,  Adjusted R-squared:  -0.02141 
F-statistic: 0.09859 on 1 and 42 DF,  p-value: 0.7551
```

As seen by the graph above, there is a linear relationship between Total Income of the office and the population of that city where the office is located. As population increases, there is also a slight increase in the total income. To reach this graph I created a model named “Model1” and included the code: MODEL1<-lm(Total.Income ~ Population..City., df)... I then plotted the two variables, creating an xlim of 1,000,000 and ylim of 1,000,000 in order to see the relationship between the two variables. Without the x lim and y lim, the data was portrayed incorrectly and it was difficult to even see a relationship between the two because of how extensive and large the population by city data was. The p value is .7551 meaning we cannot reject the null hypothesis that the data is statistically significant.

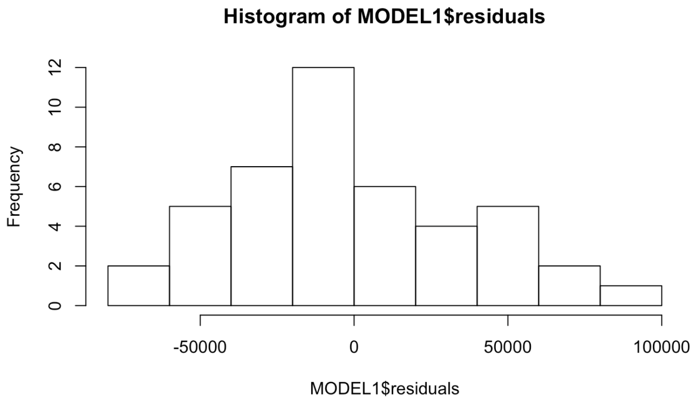
Total.Income = 38081 + 50.86(Chiropractic.Visits)



```
> MODEL1<-lm(Total.Income ~ Chiropractic.Visits, df)
> summary(MODEL1)

Call:
lm(formula = Total.Income ~ Chiropractic.Visits, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-71519 -28402  -3384  23723  93115 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         38081.668  14405.138   2.644   0.0115 *  
Chiropractic.Visits    50.867      4.011  12.682 5.94e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38790 on 42 degrees of freedom
Multiple R-squared:  0.7929,	Adjusted R-squared:  0.788 
F-statistic: 160.8 on 1 and 42 DF,  p-value: 5.944e-16
```

In the single variable linear regression above, our independent variable is the number of chiropractic visits and the dependent variable is total income. This model has an R-squared of

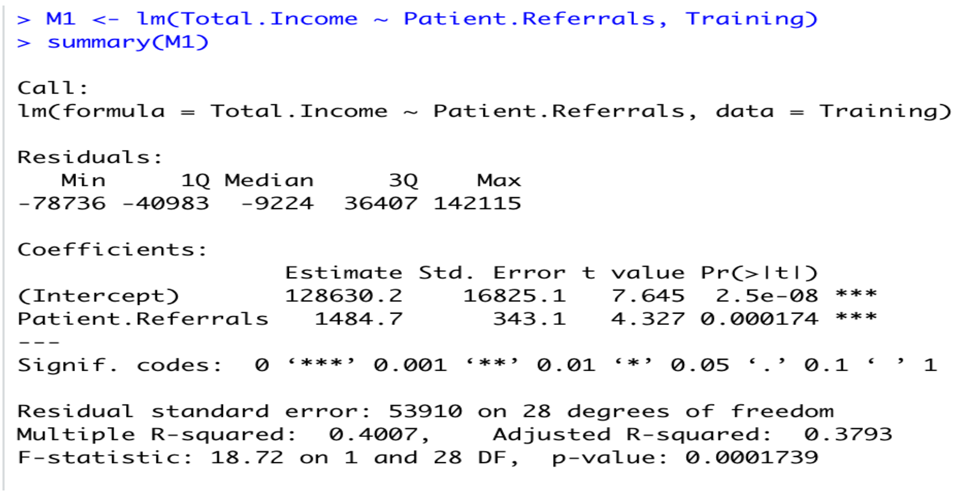
0.7929 indicating that approximately 79% of the variation in total income can be attributed to variations in the number of chiropractic visits. Chiropractic.Visits is significant above the 99% level with a beta parameter of 50.867. From this, we would expect to see a $50 increase in total income for each additional visit. Overall, the model is statistically significant with a p-value very close to zero.



Running the jarque bera test, we obtained a p-value of 0.4698. This means we cannot reject the null hypothesis that the residuals are normally distributed.

By breaking the data up into 70% training and 30% testing, we were able to obtain an in-sample error of $38,617 and an out-of-sample error of $38,800. Comparing this out-of-sample error to the average total income we get a relative error of 18.92% which we hope to improve upon with our multivariate regression models.

Total Income vs. Patient Referrals



```
> M1 <- lm(Total.Income ~ Patient.Referrals, Training)
> summary(M1)

Call:
lm(formula = Total.Income ~ Patient.Referrals, data = Training)

Residuals:
   Min     1Q Median     3Q    Max 
-78736 -40983  -9224  36407 142115 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       128630.2    16825.1   7.645  2.5e-08 ***
Patient.Referrals   1484.7      343.1   4.327 0.000174 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53910 on 28 degrees of freedom
Multiple R-squared:  0.4007,	Adjusted R-squared:  0.3793 
F-statistic: 18.72 on 1 and 28 DF,  p-value: 0.0001739
```

For our final single variable regression we wanted to run a model that would answer the question “Does having happy patients return as total income on the business, through referrals?” with an adjusted r square of .3793 and a p-value of .0002, we had found a regression equation that would meet the confidence interval of .99. With nearly 40% of total income being accounted for by patient referrals alone, we feel confident in saying that our regression equation would help this business estimate the total income of an office based on their quarterly patient referrals. The final equation is as follows:

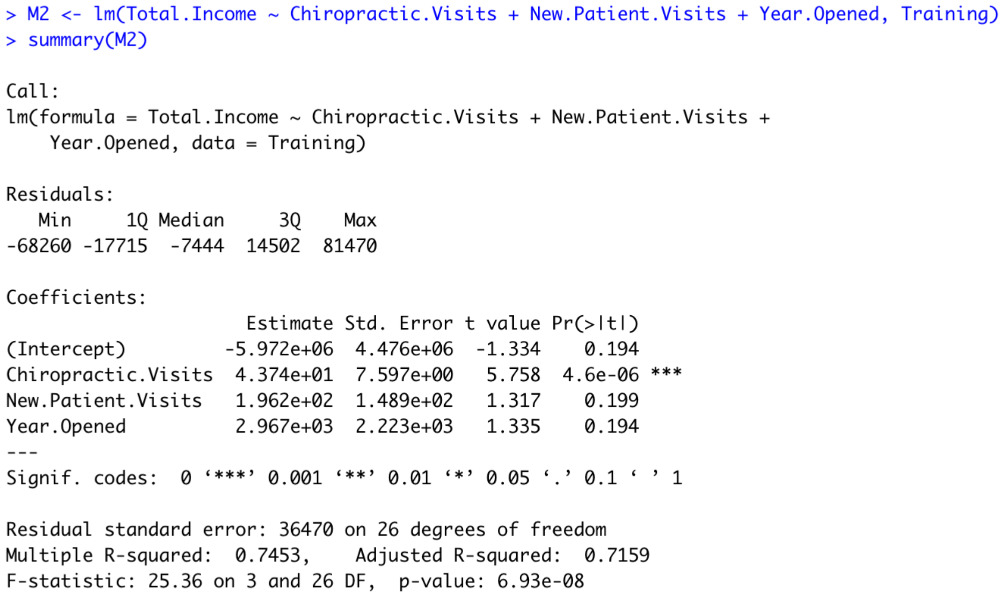
Total Income=128630.2+1484.7(Patient Referrals)

We typically find it valuable to test the regression equation by plugging an existing offices' true total income vs. their predicted total income. The total income of the Downtown Colorado Spring location (chosen at random) for Quarter 1 of 2021 was $177,252. The predicted Total Income for Downtown CS would be Total Income=128630.2+1487.7(32) or $176,236.6. An extremely close estimation, granted a lucky plucked sample for the sake of the argument.

By breaking the data into 70% training and 30% testing we were able to obtain an in-sample error of $52,085 and an out-of-sample error of $96,067.

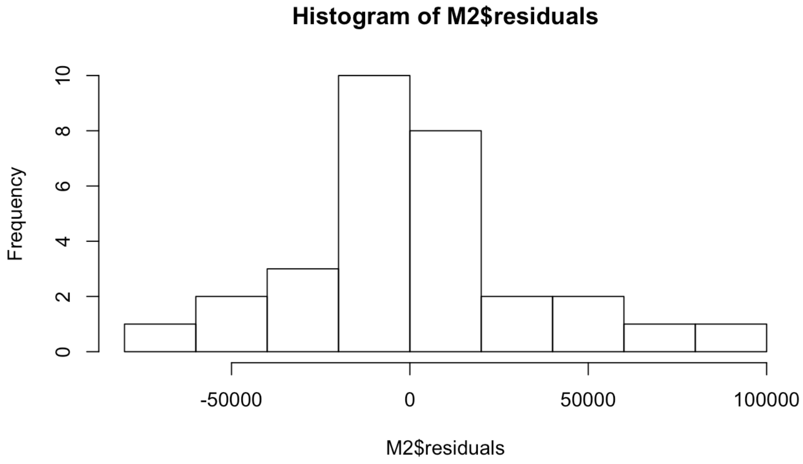
Multiple Variable Linear Regressions:

Total.Income = -5.96*10^6 + 0.43(Chiropractic.Visits) + 0.019(New.Patient.Visits) + 0.0029(Year.Opened)



```
> M2 <- lm(Total.Income ~ Chiropractic.Visits + New.Patient.Visits + Year.Opened, Training)
> summary(M2)

Call:
lm(formula = Total.Income ~ Chiropractic.Visits + New.Patient.Visits + 
    Year.Opened, data = Training)

Residuals:
   Min     1Q Median     3Q    Max 
-68260 -17715  -7444  14502  81470 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -5.972e+06  4.476e+06  -1.334    0.194    
Chiropractic.Visits  4.374e+01  7.597e+00   5.758  4.6e-06 ***
New.Patient.Visits   1.962e+02  1.489e+02   1.317    0.199    
Year.Opened          2.967e+03  2.223e+03   1.335    0.194    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36470 on 26 degrees of freedom
Multiple R-squared:  0.7453,	Adjusted R-squared:  0.7159 
F-statistic: 25.36 on 3 and 26 DF,  p-value: 6.93e-08
```

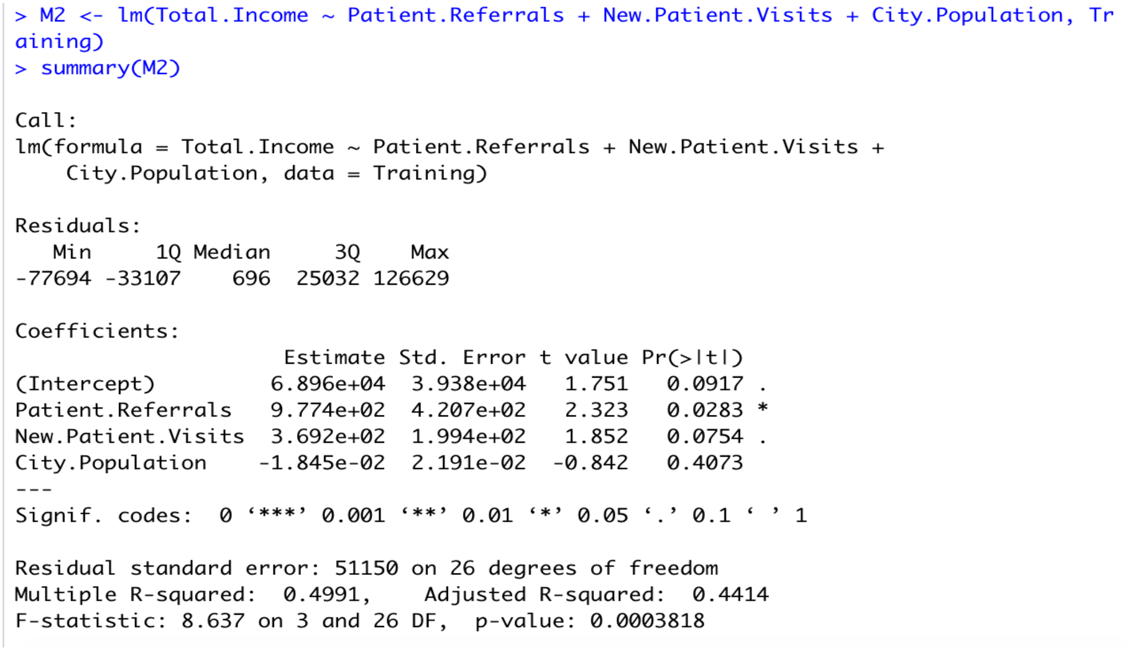
In this multivariate regression, we analyzed the relationship of number of chiropractic visits, new patient visits and year opened with total income. This model has an R-squared of 0.7453 indicating that approximately 74% of the variation in total income can be attributed to variations in these three independent variables. The numbers seen above were calculated using the training dataset which contains 70% of the data. Number of chiropractic visits stayed significant above the 99% level as we saw in the single variable regression model. To our surprise New.Patient.Visits was not a statistically significant variable. We based this assumption off of our prediction of a strong correlation between New.Patient.Visits and Chiropractic.Visits. All three estimated parameters are near zero giving minimal impact to total income. Overall, the model is statistically significant with a p-value very close to zero. With a standard error higher than all of the parameters and two of the three variables being statistically insignificant while the overall model is significant, it is likely our model suffers from multicollinearity.



Running the jarque bera test, we obtained a p-value of 0.6118. This means we cannot reject the null hypothesis that the residuals are normally distributed.

Running the model on the training and testing datasets, we were able to obtain an in-sample error of $33,954 and an out-of-sample error of $40,848. Comparing this out-of-sample error to the average total income we get a relative error of 19.92%.

Total Income= 68,960+977.4(Patient Referrals)+369.2(New Patient Visits)-.0018(City Population)

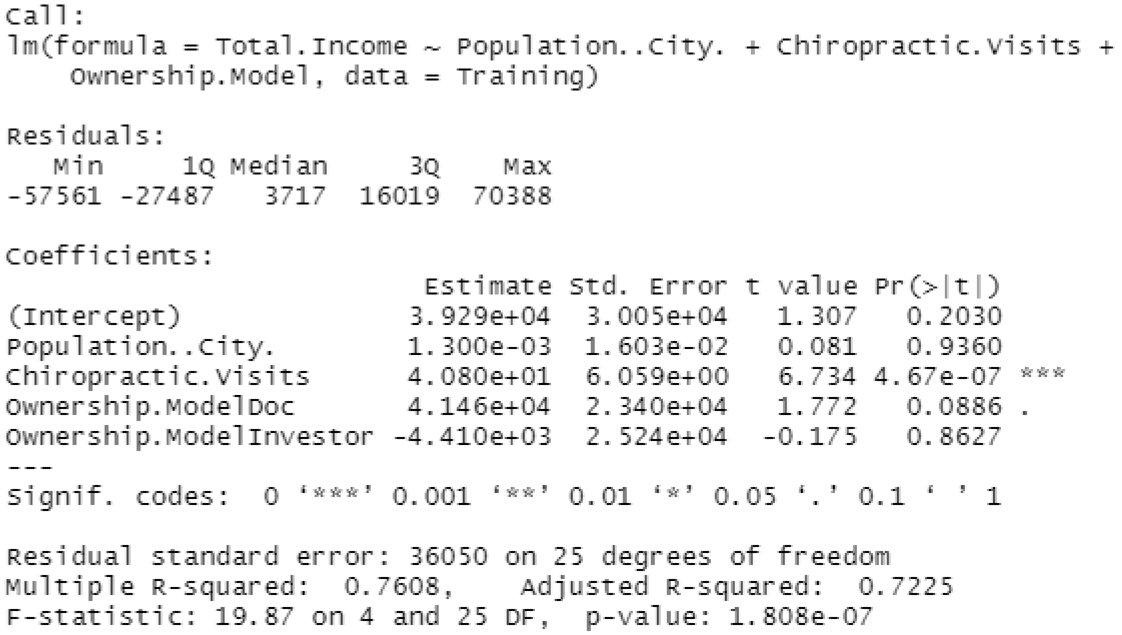


```
> M2 <- lm(Total.Income ~ Patient.Referrals + New.Patient.Visits + City.Population, Tr
aining)
> summary(M2)

Call:
lm(formula = Total.Income ~ Patient.Referrals + New.Patient.Visits + 
    City.Population, data = Training)

Residuals:
   Min     1Q Median     3Q    Max 
-77694 -33107    696  25032 126629 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)  
(Intercept)         6.896e+04  3.938e+04   1.751   0.0917 .
Patient.Referrals   9.774e+02  4.207e+02   2.323   0.0283 *
New.Patient.Visits  3.692e+02  1.994e+02   1.852   0.0754 .
City.Population    -1.845e-02  2.191e-02  -0.842   0.4073  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51150 on 26 degrees of freedom
Multiple R-squared:  0.4991,	Adjusted R-squared:  0.4414 
F-statistic: 8.637 on 3 and 26 DF,  p-value: 0.0003818
```

In an effort to raise the predictive value of the patient referrals regression equation, we decided to add on New Patient Visits and City Population. We found that while the adjusted R squared increased, it was a marginally small increase from .379 in the singular regression to .44 in the multiple regression. The p-value is still within our confidence range of .99, so we continue to reject that patient referrals is not a relevant vactor when it comes to total income.

Running the model on the training and testing datasets, we were able to obtain an in-sample error of $47,616 and an out-of-sample error of $100,477.

Total Income vs. Population/city, Chiropractic Visits, and Ownership Model



```
Call:
lm(formula = Total.Income ~ Population..City. + Chiropractic.Visits +
    Ownership.Model, data = Training)

Residuals:
   Min     1Q Median     3Q    Max
-57561 -27487   3717  16019  70388

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.929e+04  3.005e+04   1.307   0.2030
Population..City.        1.300e-03  1.603e-02   0.081   0.9360
Chiropractic.Visits      4.080e+01  6.059e+00   6.734 4.67e-07 ***
Ownership.ModelDoc       4.146e+04  2.340e+04   1.772   0.0886 .
Ownership.ModelInvestor -4.410e+03  2.524e+04  -0.175   0.8627
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36050 on 25 degrees of freedom
Multiple R-squared:  0.7608,    Adjusted R-squared:  0.7225
F-statistic: 19.87 on 4 and 25 DF,  p-value: 1.808e-07
```

Our second multivariate regression, we analyzed the data of population per city, chiropractic visits, and the ownership model to the total income variable. This model had an R squared of .7608 which in regards to the above analysis, means that 76% of the variations in total income can be attributed to the variations amongst the other three variables of population, chiropractic visits, and ownership model. As seen from the model, the Chiropractic Visits is very minal in the significance of the model. The p value is within our 99% confidence interval so we can say that the model is statistically significant. The in sample error related to this model was calculated at $65,907.61 and the out of error sample was calculated at $135,483.8.

Conclusion:

Predictive analytics give us the ability to quantify outcomes that were for a long time considered to be unquantifiable. Just as we are an accumulation of all of our decisions, a business's effectiveness is a result of its numbers. An adjusted R squared reminds us that there is a perfect regression equation for any statistic, the difficulty is in uncovering the variables. In all three six of our regression models, we discovered a different R squared, a different p-value and of course a different regression equation; yet, they all offered different insight as to what each individual office should focus on to increase their total income. The model that we felt was the most accurate was *Total.Income = -5.96*10^6 + 0.43(Chiropractic.Visits) + 0.019(New.Patient.Visits) + 0.0029(Year.Opened).* We chose this model because it has both the highest adjusted R square and the lowest out of sample error while maintaining statistical significance on a 99% confidence interval. Business analytics are a key to a more efficient business, if these chiropractic offices can focus their resources on gaining more chiropractic visits, referrals, and open up in a smaller town, odds are they will have a higher total income. In an uncontrollable world, businesses should take every opportunity to put the odds in their favor. A luxury that comes with understanding their numbers.