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 had 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.

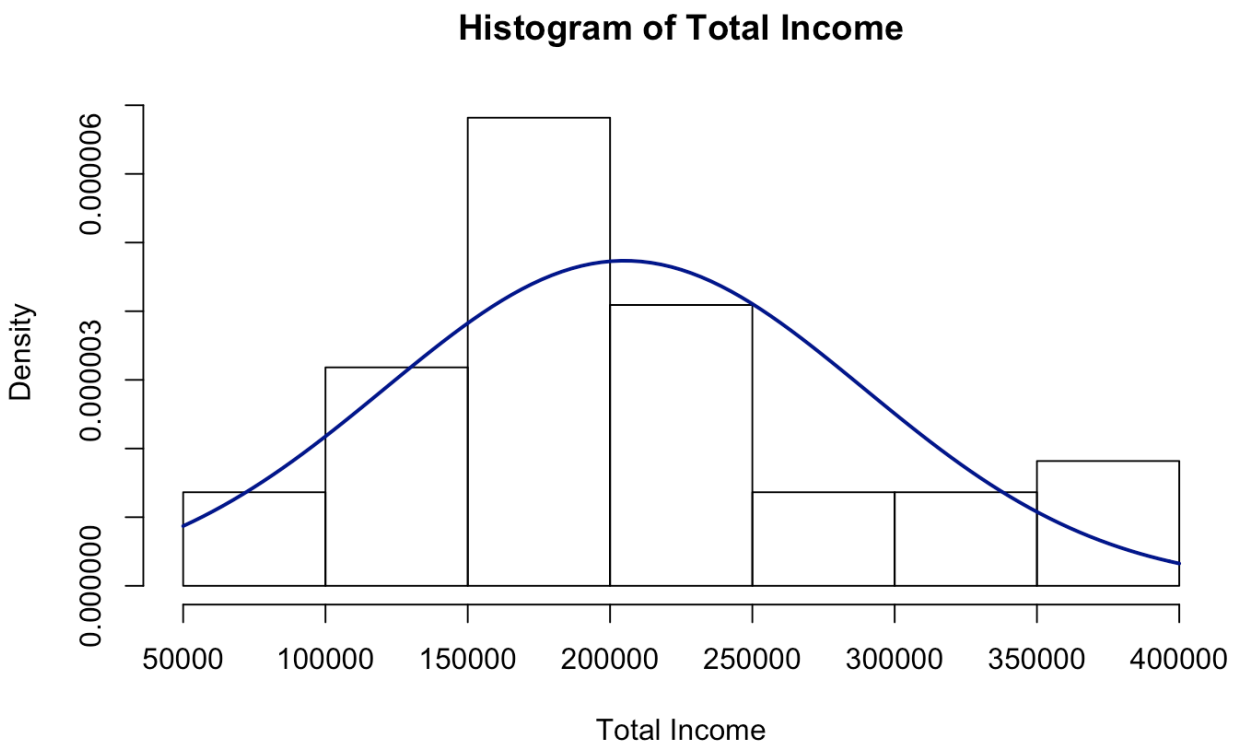
Cleaning the Data:

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.

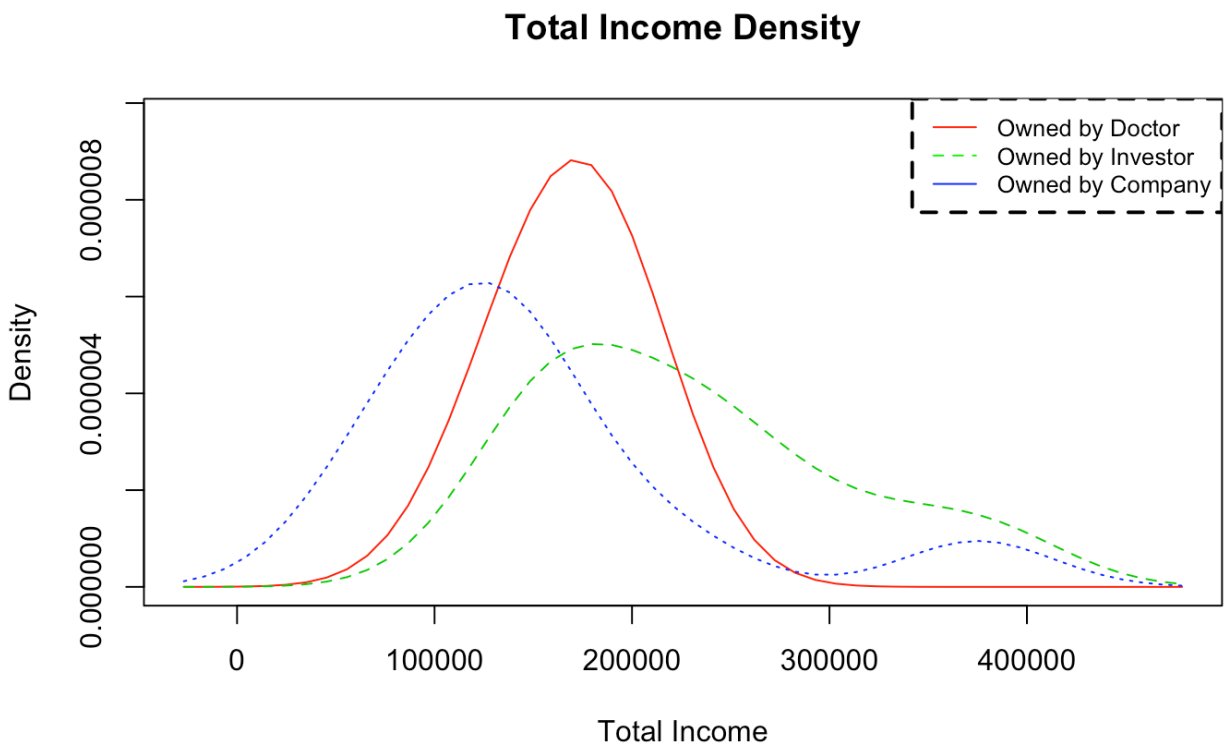
Exploratory Analysis:

```
    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
```

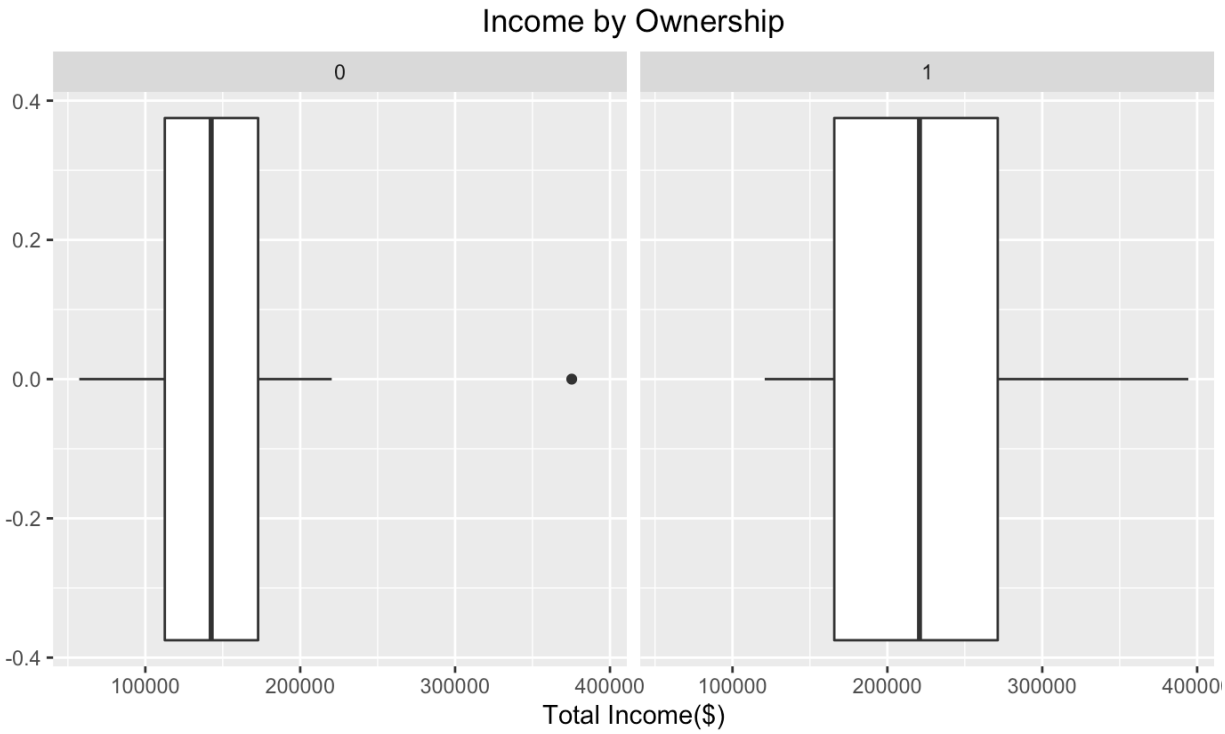
In the summary of our data provided above, we can begin to better visualize our data with simple statistics across all variables. These simple statistics include the minimum, maximum, mean, and median of all numeric variables. Our main variable of interest is total income, so the first step we took was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



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.

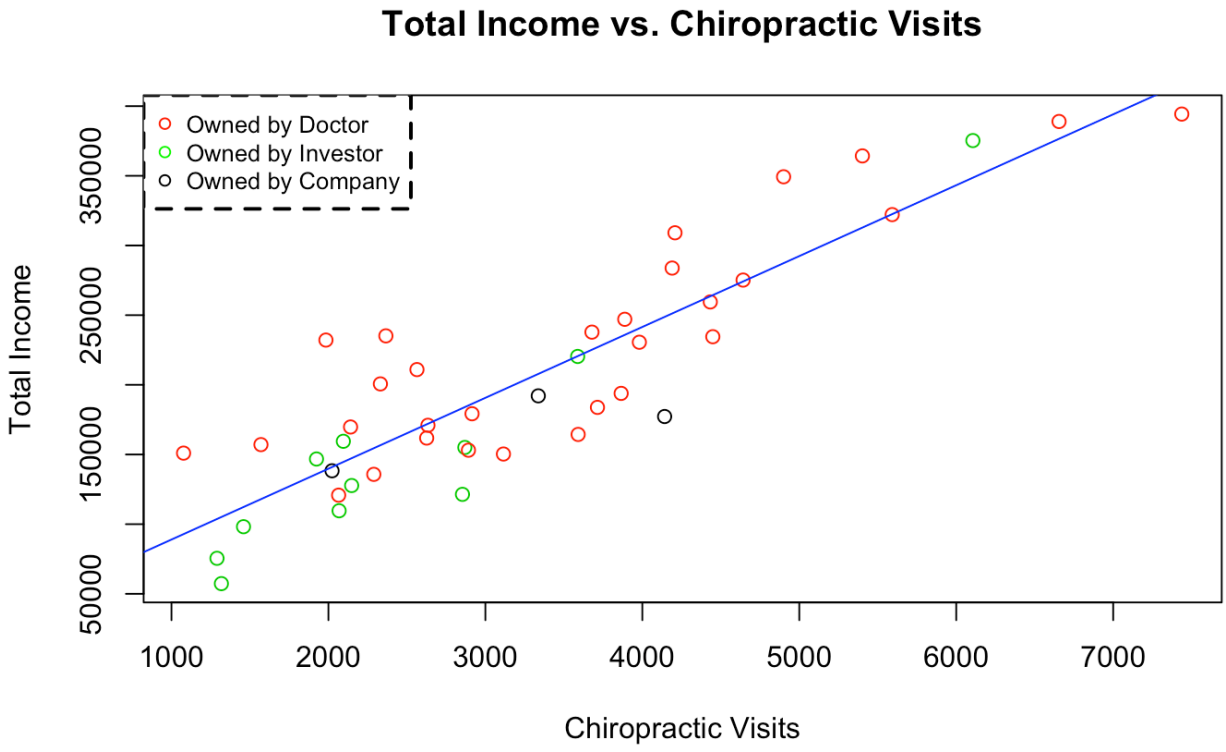


Next, we facet wrapped the total incomes by ownership models to obtain two box and whisker plots for comparison. Here, a 0 represents an office owned by an investor or the company, while a one represents an office owned by the doctor. This comparison very clearly portrays the greater total income of 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.

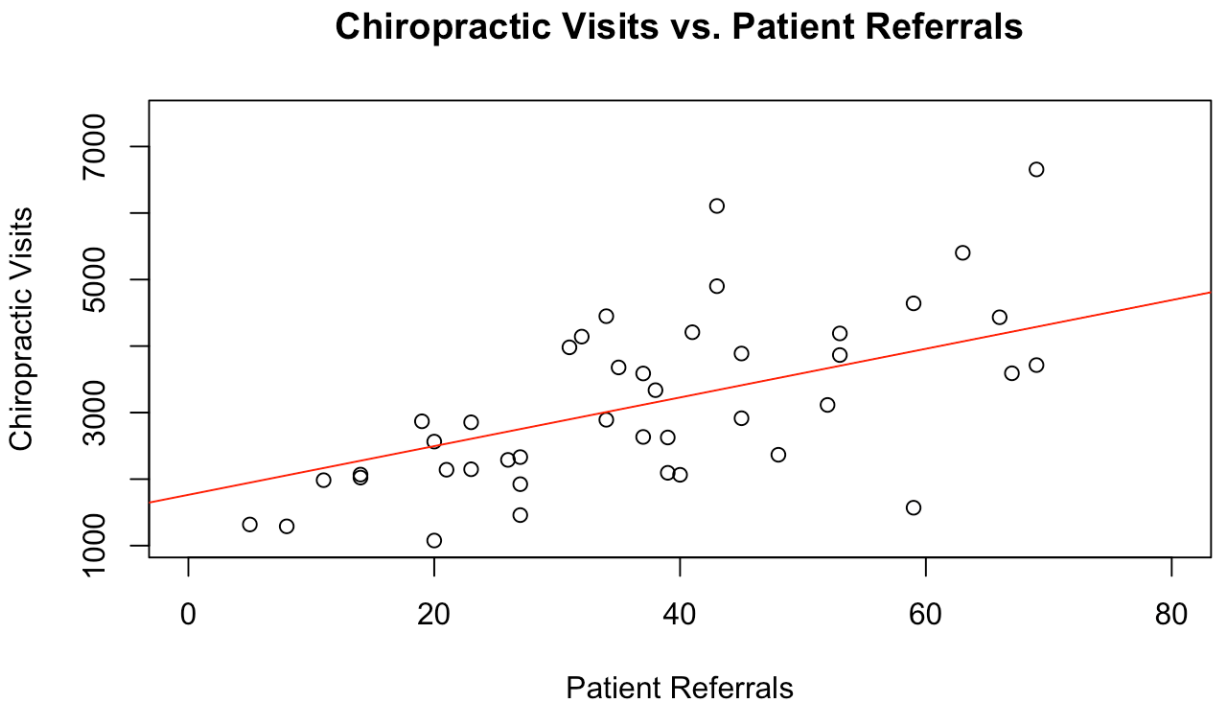
```
> cor(df[2],df[3])
             Chiropractic.Visits
Total.Income           0.8904654
> cor(df[2],df[4])
             New.Patient.Visits
Total.Income          0.6697766
> cor(df[2],df[6])
             Patient.Referrals
Total.Income         0.5959668
> cor(df[2],df[7])
             Year.Opened
Total.Income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

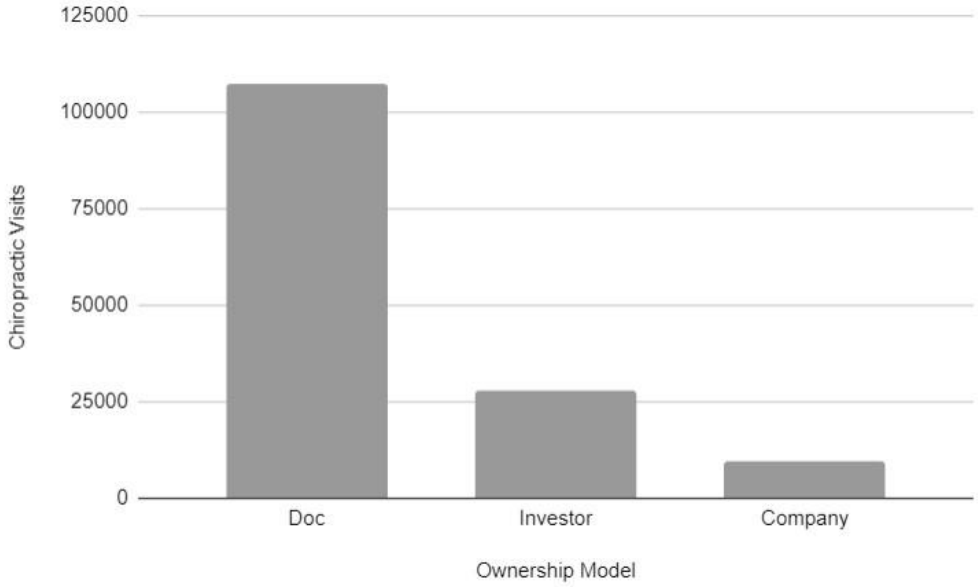


In the scatter plot above, a clear positive linear relationship is highlighted by the blue line of best fit. Each dot represents a unique office and their color indicates the ownership model which can be seen in the top left corner. This relationship comes as no surprise, but what is interesting to see 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.

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.