ISSN: 2771-7763

Research Article

Journal of Economic Research & Reviews

Cannabis Monthly Tax Revenue Analysis of Relevant Factors and Predictive Model

En-Chi Kuo[1], Dwayne Martin[2], Matthew Chin[3], Josh Cubero[4], Bahman Zohuri[5]*

[1]Golden Gate University, Ageno School of Business, San Francisco, California 94105, USA, Master of Science in Business Analytics Student Candidates

[2]Golden Gate University, Ageno School of Business, San Francisco, California 94105, USA, Adjunct Professor, Business Analytics School

***Corresponding author:**
Bahman Zohuri, Adjunct Professor, Business Analytics School, Golden Gate University, Ageno School of Business, San Francisco, California 94105, USA

Submitted: 15 Aug 2022; **Accepted**: 22 Aug 2022; **Published:** 02 Sep 2022

Citation: *En-Chi Kuo, Dwayne Martin, Matthew Chin, Josh Cubero, Bahman Zohuri, et al. (2022). Cannabis Monthly Tax Revenue Analysis of Relevant Factors and Predictive Model, J Eco Res & Rev, 2(3), 372-383*

Abstract

This analysis will focus on how legalized recreational and medicinal usage of cannabis would have an impact on a state's economy. We will look at the benefits and the lack thereof to find more details about the connection between cannabis and individuals' financial factors, such as their annual income, the gross domestic product if they are employed or not, and, if possible, how much they spend on healthcare or if they are receiving funding from the government to pay for any of their expenses. However, some states have begun to include cannabis as a legal drug for both medicinal and recreational reasons. More research has been done in the last half-century-plus to destigmatize cannabis' portrayal as a harmful drug, but as a new opportunity to study the unknowns of the plant and products made, as well as the economic growth that consumers have capitalized on, particularly in the last decade.

Keywords: Cannabis, Cannabis and Individual's Financial, Monthly Tax Revenue, Domestic Product, State of Economy, Artificial Intelligence, Machine Learning, Deep Learning.

Introduction

Efforts to legalize cannabis in the states have occurred as early as the late 1960s at the heart of the counterculture movement in the United States. Still, it would not be until the 2010s that people in their respective states began to have a say in cannabis legalization. There is still polarization in some parts of the country over its usage for medical or recreational purposes, but changes have come to find good out of the current growth. The state of Colorado is the second earliest but the most notable of the fifty states to support both medical and recreational legal purchasing and usage of marijuana and related cannabis products. The state's cannabis sales data are frequently cited to assess potential revenue and tax outcomes in an unpredictable, polarizing, growing, and potentially beneficial drug industry.

While the growth of the sale and production of cannabis has been a booming industry in the last five years, there are necessary taxes, fees, and regulations with which companies in the cannabis industry must comply. As the focus is Colorado, we will analyze the impact that the taxes from revenue had on the growth of the recreational side of cannabis. This will be achieved by using Colorado's data on cannabis and general data on economic states and governmental benefits that individuals were categorized under for machine learning models programmed across various Python language libraries. The information from these models will help provide an information gauge to predict future revenue values and opportunities in Colorado after the legalization of the sale, production, taxing, and usage and consumption of cannabis in the state.

History of Cannabis Legalization

The history of cannabis has been complex. The substance has transitioned from legal to illegal and then back to legal in certain states with specific stipulations. The complexity of cannabis is not only prevalent in the United States but the world over. However, for this study, the models and analysis will focus solely on cannabis in the United States.

Its history can be viewed globally, with consumption dating back as far as 600 BC in China. During these times, cannabis-based hemp was used for various products such as ropes and other durable goods, while others used cannabis for medicine and recreation. In the US, cannabis implementations can be traced back to the 1600s in Jamestown, where hemp plants were grown and required to export to England. While cannabis and hemp products saw significant use throughout history, the substance received less favorable treatment in the 20th century.

The United States Pure Food and Drug Act of 1906 was one of the earliest and most profound anti-cannabis legislation, significantly restricting medicinal and recreational cannabis use. Subsequently, states deemed cannabis a poisonous substance, regulating cannabis according to poison laws at the time.

The US continued its battle against cannabis by implementing the Marijuana Tax Act, adding a tax on companies conducting commercial operations with cannabis. The US Supreme Court subsequently overturned the Marijuana Tax Act in 1969. Then President Nixon responded by championing the Controlled Substances Act, which designated cannabis as a Schedule 1 drug. The Controlled Substances Act has had a significant impact on cannabis, as the act outlawed cannabis at the Federal level. Strict marijuana uses and propagation penalties followed the Controlled Substances Act, culminating in President George H. W. Bush's War on Drugs.

Cannabis use, however, has slowly made a return via medicinal use, with California emerging as the first state to decriminalize medicinal cannabis. Since then, many states have decriminalized medicinal marijuana, despite Federal regulations. According to the National Conference for State Legislatures, twenty-six states and the District of Columbia have cannabis decriminalization laws. The cannabis decriminalization movement has led to increased debate at all levels of government. The progress made since the counterculture movement to bring cannabis into the mainstream in American culture has emerged in the media and business. Studies have shown that marijuana has many health benefits, even though consumers can get "high" by either inhaling or consuming this drug. While still illegal at the federal level in the United States, efforts at the state level have been made since the 1970s for research to prove the benefits of this plant and how it can be used.

In November 2012, voters in the states of Washington and Colorado favored ballot measures to legalize the use of recreational marijuana effectively. In further detail, it allowed their states' governments to recognize producing, selling, purchasing, and using cannabis products to be allowed without any punishments, except for the federal government's regulations [1]. It was more than the health or medical benefits, but the drug could be a successful business under statewide regulations. Voters' approval of legal recreational cannabis usage and sale went into effect on December 6 and December 10 that year, respectively, for the states of Washington and Colorado. Due to disputes and arrangements made within their states' legislators, it was not Washington that was the first state to allow cannabis to be legally sold for recreational purposes formally, but it was Colorado.

It was not until January 1, 2014, that Colorado was the first state to formally allow cannabis entrepreneurs to legally pursue business in their state, which provided full rights for residents in the state to purchase and consume recreational cannabis products for their use with the Colorado Retail Marijuana Code, requiring sales tax for all items purchased containing this substance [3]. It has had a ripple effect on other states in passing legislation to allow cannabis to be sold for recreational or medicinal reasons in states such as Oregon, California, and New Jersey, among others. Allowing cannabis to be its industry for revenue with the support of government regulations and finances had a major change in the culture and perception of its portrayal over time. The hope is to predict future outcomes for the revenue and the cannabis retail sales taxes that were into their state's laws.

Current Legalization Talking Points

In March '21, 2022, NPR's Planet Money organization released a study citing cannabis legalization data, which indicated three positive impacts of legalizing cannabis. The study intended to debunk opponents of decriminalization talking points. First, around crime rates, the study found that legalized cannabis had no significant impact on the incidence of crime. States with decriminalization laws neither experienced an increase nor a decrease in violent crimes. Next, the NRP study found that decriminalized cannabis has not positively or negatively affected the occurrence of traffic fatalities.

Additionally, NPR organization cited a report by the CATO Institute that decriminalization of cannabis has not caused the price of cannabis to crater but, on the contrary, has strengthened the price therein. Moreover, a study by Leafly and Whitney Economics determined that cannabis decriminalization had added 321K jobs in the US legal cannabis industry. Lastly, and most pertinent to this study, NPR found that cannabis decriminalization was a boon for states' budgets. Decriminalized cannabis tax revenue was found to have exceeded initial expectations, particularly in Colorado, where $387M in cannabis tax revenue. Furthermore, California collects $50M per month in cannabis tax revenue. Tax revenue is one of the most cited reasons for cannabis legalization.

Problem Statement

As previously stated in this study, twenty-six states, plus the District of Columbia, have implemented some form of cannabis decriminalization. With that, there are also twenty-six different methodologies for implementing taxation. For instance, Colorado, the first state to approve legal cannabis sales for adults, implements a wholesale excise and special retail tax, but only for recreational cannabis. Meanwhile, California implements a more robust cannabis taxation policy, with a per ounce tax on flowers, a tax on leaves and trim, an excise tax, and a standard sales tax. Lastly, Alaska has a more straightforward cannabis tax structure that implements a flat tax per ounce at wholesale. These are just three examples of cannabis tax models and implementation complexity. Next, this study will explore one state's method for estimating cannabis tax revenue.

Literature Review

In September '20, the Bureau of Business and Economic Research at the University of Montana published a report outlining their methodology for estimating cannabis tax revenue in Montana. The study was commissioned for the New Approach Montana initiative and was titled An Assessment of The Market and Tax Revenue Potential of Recreational Cannabis in Montana. The report submits that quantifies the estimated size of the recreational cannabis market in Montana and the possible tax revenue should the state decriminalize adult cannabis use. The research in this literature was conducted independently of the state of Montana and intended to be a political campaign to decriminalize cannabis consumption.

At a prominent level, the research uses sales from Montana residents and non-resident tourism to estimate potential tax revenue, with underlying estimations leading to forecast cannabis sales revenue. The study estimated that at a 20% cannabis tax, the state of Montana could generate a total of $43.4M to $52M in tax revenues from 2022-2026. This tax revenue estimation reflects a total recreational cannabis market between $217.2M and $259.8M.

The study used survey-based research on the use and incidence of services for resident and non-resident users. The survey data estimated that 14.3% of adults in Montana consumed marijuana in the previous 30 days, whereas by comparison, the US national average is 9.3%.

Additionally, the survey found that of that 14.3% of cannabis users, 22% say they use cannabis daily. Furthermore, daily cannabis users account for 67% of cannabis consumption. Lastly, the survey stated that 15% of tourists to states with recreational cannabis make visits to retail stores.

The first step Bureau of Business and Economic Research (BBER) took to create its model was to assess the cannabis market size in Montana. BBER first noted that estimating the cannabis market in Montana was complicated by the illicit cannabis market comprised of homegrown, organized production, and illegal imports from states or countries.

Consequently, BBER stated that there is no reliable sales volume data. BBER chose to follow Colorado's approach by implementing a model that considers consumption instead of sales volume. The study combines Montana consumption survey data and national usage survey data to compile a market size of all users in Montana.

Furthermore, the BBER study estimated that Montana's marijuana market size for consumption was 30.4 to 32.8 metric tons annually. Lastly, their research estimates that marijuana would retail for about $270 an ounce, for a total of $290M per year of retail sales. This estimate does not account for potential increases in legal cannabis consumption caused by price elasticity, changes in cultural norms, and the increased popularity of oils and edibles.

The BBER study did state that revenue estimates did come with several caveats. The first caveat is that these estimates assume suitably infrastructure such as regulations, supply, and licensing would be in place by 2022. Next, most cannabis tourism data are derived from mature markets, which benefit from being the first to market and might not reflect the actual outcomes in Montana.

Additionally, Montana's cannabis revenue could face competition from western states and Canada as more attractive regions to visit. Lastly, their study assumes that only leisure travelers to Montana will consume cannabis, while other visitors are not factored in.

The final and most poignant point in the BBER study is that there exist uncertainties in their model. Assumptions are made and identified throughout the research and the risks associated with adopting their model. Some of these risks include price decline created by production economies of scale, movement to consumption of edibles and oils, the potential inverse relationship between medicinal and recreational cannabis, the uncertainty that Montana tourists will consume cannabis products, and the cultural acceptance of cannabis post-legalization. Lastly, BBER recognizes that building a forecast model is difficult even when sufficient data is available but believes that the consumption survey-based data would be adequate to create a reasonable estimate of market size, retail sales, and tax revenue.

Dataset

In this study, we as authors have used capability and functionality of Artificial Intelligence (AI) along with its subsystems such as Machine Learning (ML) and Deep Learning (DL) to Analise our dataset required for this study.

Several key performance indicators (KPIs) will be used throughout this economic impact analysis of cannabis in the state of Colorado. In terms of the state's economy, the key financial indicators we will be looking at are the following: unemployment percentage rate in the state, Zillow Home Index, Supplemental Nutrition Assistance Program (SNAP), labor force participation, federal funding received, leisure hospitality, coincident economic activity index, business applications, all employees' nonfarm payroll, average hourly wage, and revenue. While all these distinct factors play key roles in determining the tax revenue value of the sale of cannabis in the state of Colorado, revenue is the most important because that key performance indicator best determines how said the state has dealt with any value in allowing business functions with cannabis. In particular, the tax revenue will help determine how much, if possible, the state's government benefitted from the cannabis business from products sold to consumers resulting in the state, city, and/or county sales taxes wherever applicable.

We will be using the overall revenue from Colorado to assess the economic potential that cannabis has on the state for business and any impacts resulting in more people moving to the state for job opportunities, lifestyle, or to find a different place to live. Further understanding of different economic metrics would help better gauge the direction the state is going towards potentially profiting off an up-and-coming drug industry. There are more than eight years of economic data provided by the St. Louis Federal Reserve to help provide more context into how affordable or not affordable marijuana products are. Determining the economic impact that cannabis has had on Colorado could be determined by the economic climate that gradually led to a massive increase in its population, especially in the later years of the 2010s decade from 2016 to 2019 and into 2020. As the first state to open a recreational business, such a reason for businesses and individuals who used cannabis to move to Colorado and take advantage of the laws they were protected to continue their actions around the drug legally and responsibly for the enjoyment of others.

While there is much variability in the kind of economic data that is collected in states such as Washington, Oregon, or California, Colorado has systematically included cannabis business data to address different concerns and needs among different people. We chose Colorado over other states because it formally allowed recreational selling and consumption of cannabis the earliest versus other states, having started in January 2014. Washington, on the other hand, despite voter approval on a measure favoring the use of recreational cannabis, did not officially recognize it as a legal business and industry practice until July 2014. Colorado had public data available to find current and potential connections to the advantages and disadvantages that cannabis has had on its economy, the growth of its population and entrepreneurship, and the higher amounts of variation possible, which have led to a boom of overall development in the state.

Several definitions will be referenced throughout this analysis. All the economic data comes from the St. Louis Branch of the Federal Reserve.

The first key performance indicator of use to compare Colorado's sales of cannabis is the unemployment rate in the state. The unemployment rate is the percentage of people who are not actively in a job at a company or organization and are receiving money and other forms of compensation for the time they are working. The lower the unemployment rate in the state is, the more likely there are individuals or groups of people who can purchase cannabis products. While that is often the case it does not consider the fact that even some people whose status is unemployed can also afford said products for their usage.

Next, the Zillow Home Value Index (ZHVI) uses Zillow's estimates to assess the housing market for the state by determining a general value of prices in various housing types. It includes single-family and multi-family homes of assorted sizes and is a widely used measurement to provide a "typical" price anywhere between the top 35 percent and the top 65 percent of the prices of most homes (Zillow, n.d.). The estimates they give on the value of a home are provided in an average range, with the percentile not being in the top 25 percent, nor is it in the bottom 25 percent. It includes houses and condominiums but does not include the prices for apartments or homes with rent fees but provides an estimate on what one house could be worth in whole if it were bought.

Both state and federal governments provide benefits to low-income families and individuals who may experience financial challenges paying for food and groceries. The Supplemental Nutrition Assistance Program, known as SNAP, provides such welfare opportunities, expanding the then-known federal food stamps program (U.S. Census Bureau, n.d.). This measurement determines how many Coloradoans received SNAP benefits whenever they experienced poverty, near-poverty, or other hardship situations. Monthly numbers from the St. Louis Federal Reserve's Economic Database can provide information on how challenging it is for people on these benefits to get the nourishment they need to survive without having to worry about not being able to afford groceries or putting food on the table.

Fourth, the labor force participation rate is the percentage of individuals who are either currently employed or had been working at the time the data was collected. This is not representative of those who are eligible to work because this is a measurement of those who were working at the time they were surveyed or if they are currently employed. Participation in employment is considered as one choosing to work and receive voluntary compensation in the form of money and other incentives, if applicable. This could be tied to our analysis to determine how many individuals who were actively working had spent money on cannabis products for their consumption and enjoyment.

The fifth key economic indicator in the dataset identified the federal funds effective rate, known as "FEDFUNDS." This percentage determines the interest rate on money borrowed from banks or other organizations which users would have to pay back on such as credit cards, mortgages, and loans, among other payment purposes ("Federal funds," n.d.). Federal funds are determined by how the Federal Reserve Banks across the US when the interest rates are adjusted on money between the Fed and the banks. They research and evaluate overall financial markets in the US and around the world with consumers' salaries, spendings, and investments.

Leisure Hospitality is the next key economic indicator for use in this analysis. This monthly measurement consists of employment data from the U.S. Bureau of Labor Statistics primarily focused on the number of employees in the hospitality and leisure industries in the state of Colorado. Often this is centered around individuals who participate in or take advantage of hospitality businesses which cover the needs for travelers, tourists, business trips, retail, and food, among other things. It was expected according to the St. Louis FRED data that this industry faced major losses during the pandemic because less people worked in jobs and fewer people traveled to Colorado or other states.

The Coincident Economic Activity Index consists of the four-following metrics: unemployment rate, nonfarm payroll enrollment, mean manufacturing hours worked by employees, and monetary compensation. All of this is focused on individuals in Colorado who work in fields where most jobs may be in demand or are offered but with not as many applicants considering those positions in the industries. This could help find connections between how many employed and unemployed individuals were able to purchase cannabis products without any financial problems or pressures with or without payment or with a stable income.

Business Applications is another key performance metric that shows how many businesses filed papers to be formally recognized as businesses in the state for sales, taxes, and financial incentives and benefits from the government if possible. The data has shown an upward rise in the number of applications filed in the state since Colorado opened to cannabis business and the set regulations in January 2014, which has remained consistently upward over a span of eight-plus calendar years of legalized business.

The metric" All Employees, Non-Farm Payroll" defines the number of Coloradoans who are working stable, paid jobs that contribute to the state's economy as well as the country's economy with respect to the Gross Domestic Product (GDP). This number determines the number of people who are employees of a nationally or state-recognized company, agency, or other form of an organization where these employers compensate its employees and maintain the constant flow of money across various kinds of transactions.

The Average Hourly Wage is the dollar-per-hour rate for any individual who is employed for any stable job. Individuals who are considered in the mean hourly wage include hourly employees, contractors, employees on non-hourly stipends, and salaried employees who have yearly pay. The analysis of this metric can gauge amount individuals can spend on cannabis and what factors their income has on if consumers can purchase these products, and if so, at what quantity or quantities, to see if they buy more month after month.

The final key performance indicator in the economic data is revenue. This measurement is important because it determines how much was made in sales for products. It represents how much investments are made in cannabis in the state per month. As the earliest state to allow recreational business to cannabis, in more than eight years of data collection, the amount of money generated from businesses selling cannabis has gradually increased until the COVID-19 pandemic in March 2020, which saw an exponential spike in revenue. Thus, taxes on the sales increased leading up to the pandemic and experienced this massive growth when Coloradoans and Americans had to stay home. There has been an increase in the access points outside of brick-and-mortar stores selling products with this substance which also factors in the amount of expected sales tax that the state was to receive from all transactions.

Exploratory Data Analysis

The aim of this project is to find economic affection after Cannabis was legalized by the Colorado government in 2014. We will be applying Python as the technique tool for further exploratory data analysis along with the time series analysis. First, we will be using a heatmap to explain the correlation between the variables in the datasets. Based on the graph below (Figure 1, generated AI, ML, and DL capability and functionality), we can find that the diagonal squares from top left to bottom right are all in the lightest color, whereas the number closer to one means the six independent variables have a strong correlation with the independent variable.

The graphs below are the count of each variable (figure 2). The unemployment rate is mostly between 2% to 3%, as the average rate is lower, which means that there will be more people who are affordable to cannabis. Zillow Home Index has the maximum count in the price of $40,000, and most of the month is under the maximum price. The requested SNAP funding is mostly either between $400,000 to $500,000 or none. The labor force participation is at an average of 67%, as the higher percentage of the available workers in the workforce, the higher willingness will the customers to purchase cannabis-related products. The rate of FEDFUNDS has a maximum count of rate under 0.3. The count of leisure hospitality is focused on 295 to 350. The coincident economic activity index is around 140 with a count above 12.5. Business applications are between 6,000 to 7,500. All employees' non-farm payroll is around 2,700. Foremost the average hourly wage is around $27 per hour. And lastly, the most count of revenue is about $2,000,000. Overall, the unemployment rate remains low along with over 66% of labor force participation, and with the stability of the Zillow home index, people have more spare money to treat themselves, which buying cannabis products may be one of their choices.

Next, we will further explore the monthly data with the five selected variables below (figure 3~7). The economic status before the year 2020 is stable, with a drop in the unemployment rate, and an overall increase in labor force participation, average hourly wage, and tax revenue risen upon the Zillow home index. A good economic status means people will have more excessive disposal income. According to Maslow's Hierarchy of Needs, after fulfilling the basic living needs, our next goal is accomplishing the psychological needs, which I consider purchasing cannabis products as enjoyment either with friends or to relax by ourselves. However, as Coronavirus disease has widely spread into the world beginning in December 2019 (Gharehgozli, the pandemic has made devastating harm to the global economy. [3] The economy took around one year to recover, which may affect the desire to spend on leisure products or events [4].

Lastly, we will present the seasonal pattern in the time series with the selected five variables mentioned in the previous paragraph (figure 8~12). The seasonal pattern of time series graphs has comparably significant changes from the monthly exploratory graphs. The main event that causes all the variables to be altered by the global wide pandemic. Aside from the period before 2020, the economy remains solid. The only variable that has not been affected by the pandemic is the tax revenue, which fluctuation has begun in early 2019. We conclude that people are more likely to have stable living conditions to spend more money on enjoying their living

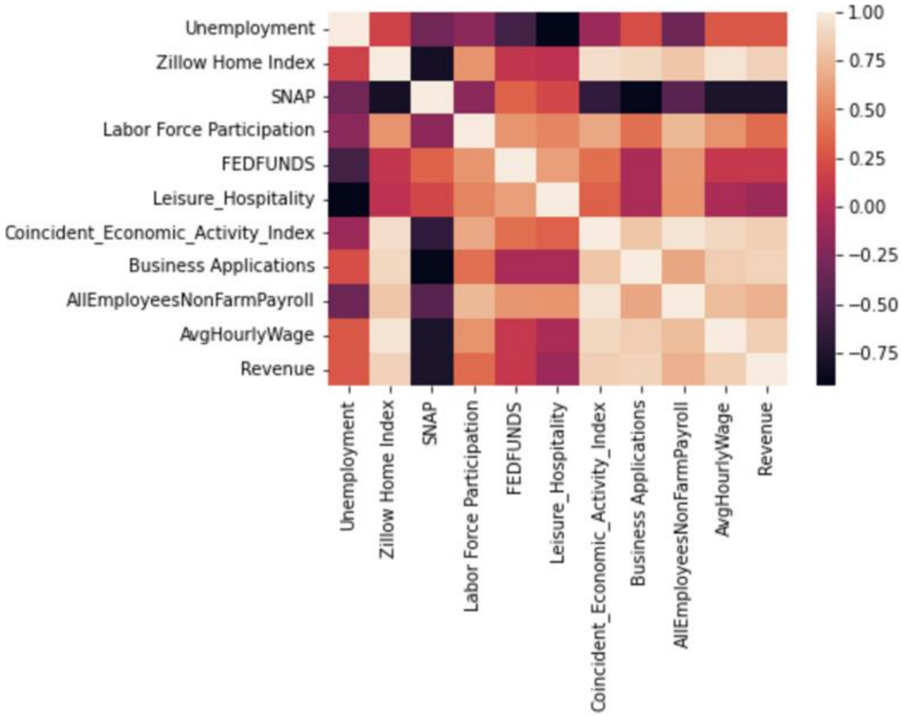


Figure 1: The Heatmap of Colorado Cannabis Economic
(Authors using algorithmic Payton using AI, ML and DL systems)

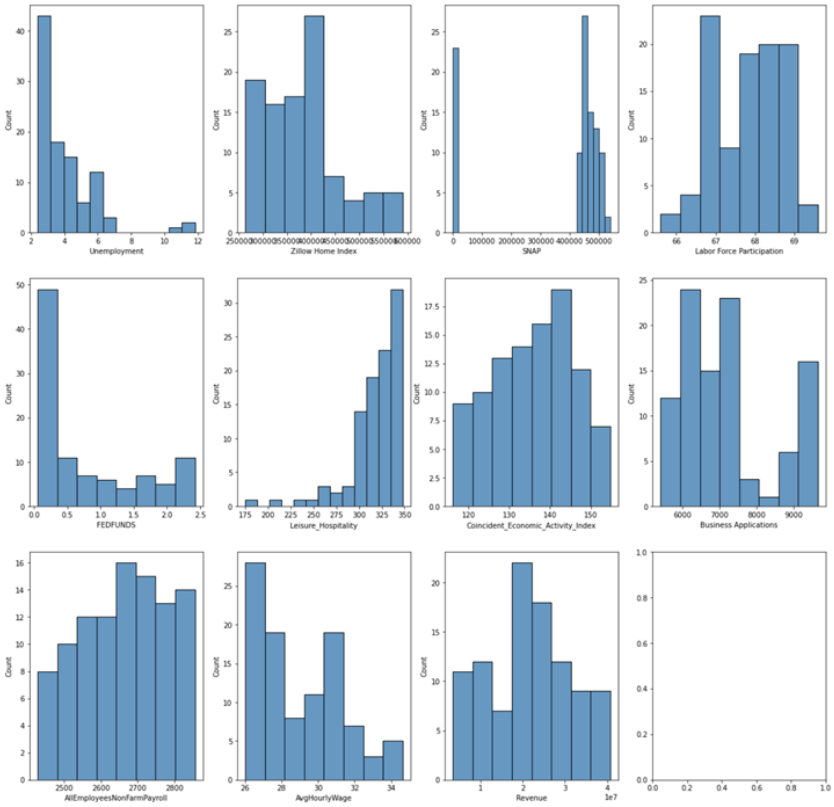


Figure 2: Histogram of Colorado Cannabis Economic Data
(Authors using algorithmic Payton using AI, ML and DL systems)

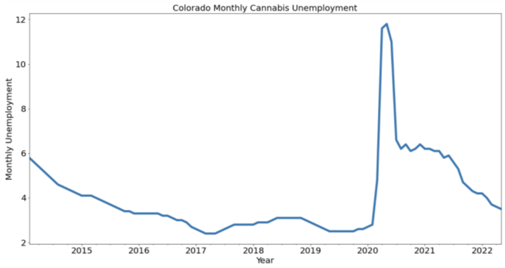


Figure 3: Colorado Monthly Cannabis in Unemployment

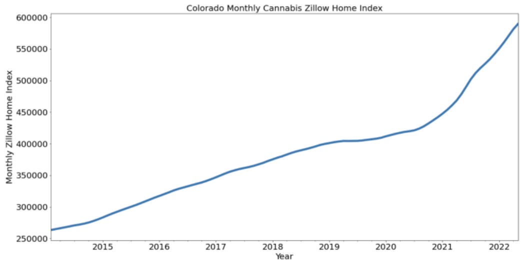


Figure 4: Colorado Monthly Cannabis in Zillow Home Index

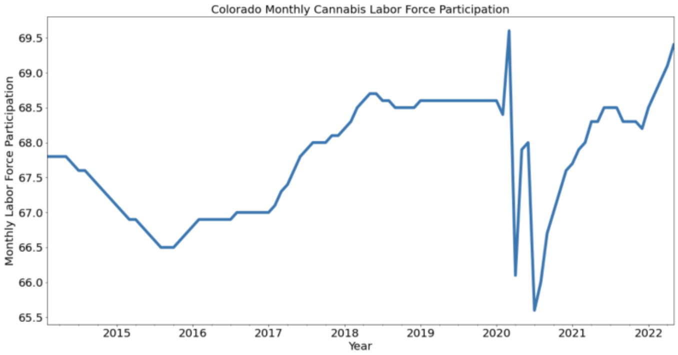


Figure 5: Colorado Monthly Cannabis in Labor Force Participation

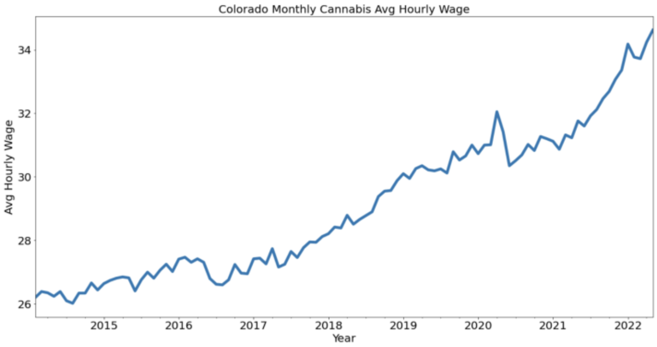


Figure 6: Colorado Monthly Cannabis in Avg hourly Wage

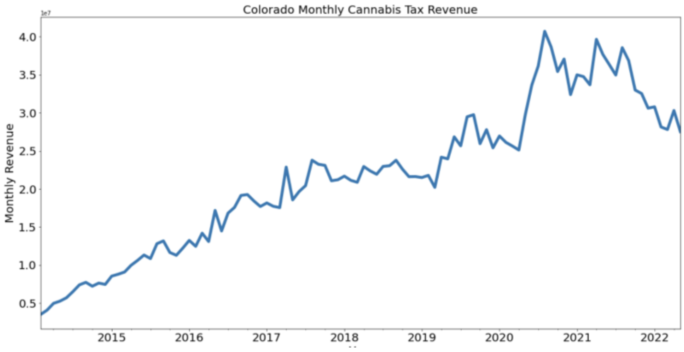


Figure 7: Colorado Monthly Cannabis in Tax Revenue

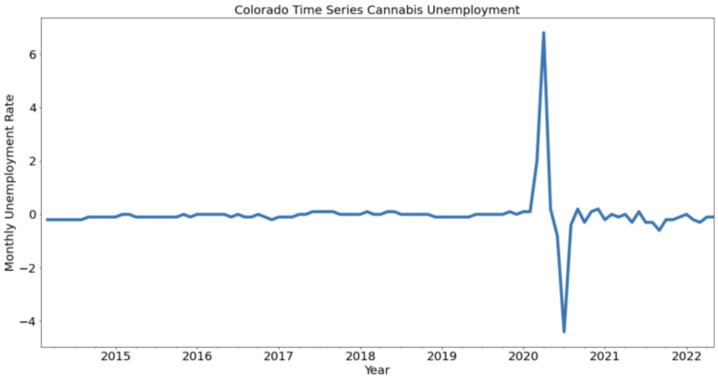


Figure 8: Colorado Monthly Cannabis Time Series in Unemployment

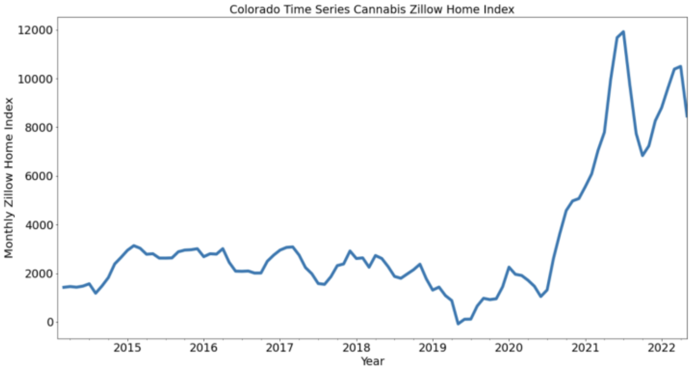


Figure 9: Colorado Monthly Cannabis Time Series in Zillow Home Index

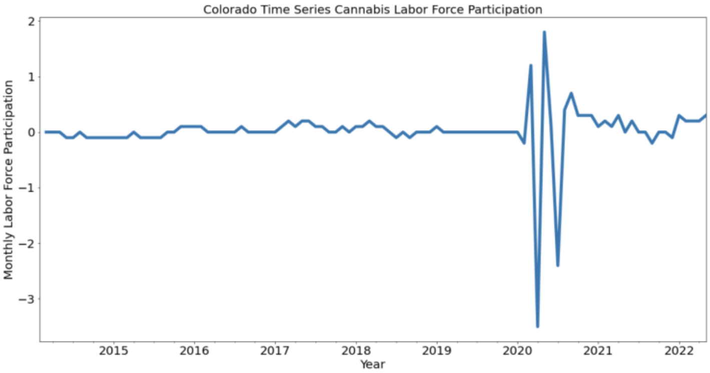


Figure 10: Colorado Monthly Cannabis Time Series in Labor Force Participation

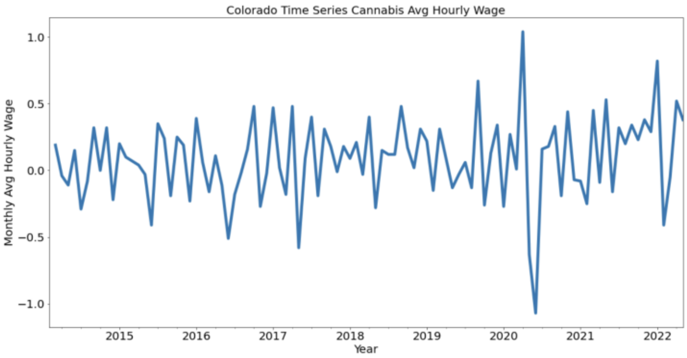


Figure 11: Colorado Monthly Cannabis Time Series in Avg hourly Wage

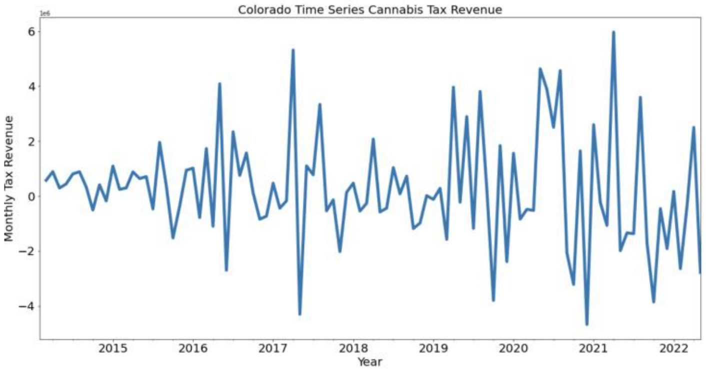


Figure 12: Colorado Monthly Cannabis Time Series in Tax Revenue

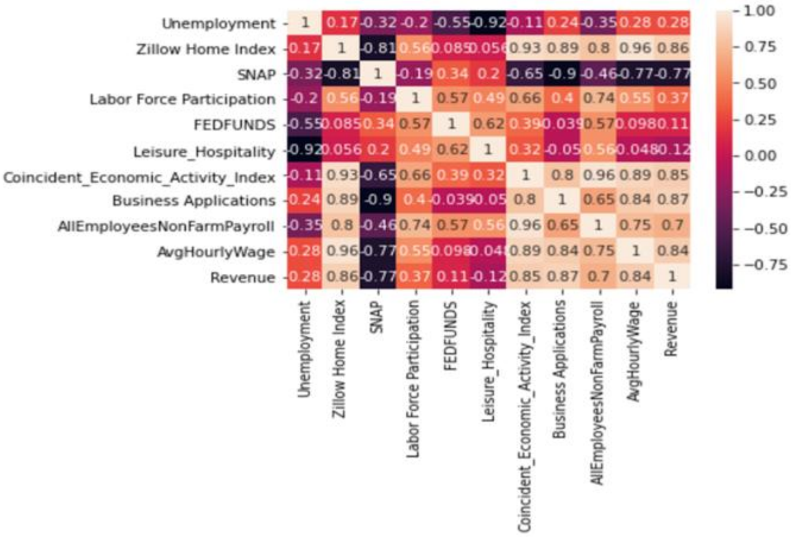


Figure 13: Correlation with Heatmap Data Points

In conclusion, the continuous growth of the economy has provided people with a secure income source that allows them to make excess purchases other than basic living goods. However, as the pandemic started, the shutdown economic has affected our living for a year to recover. The next session will be discussing machine learning models and the analysis as it presented in Figure 13.

Machine Learning Models and Analysis

Zillow Home Index Positive correlation

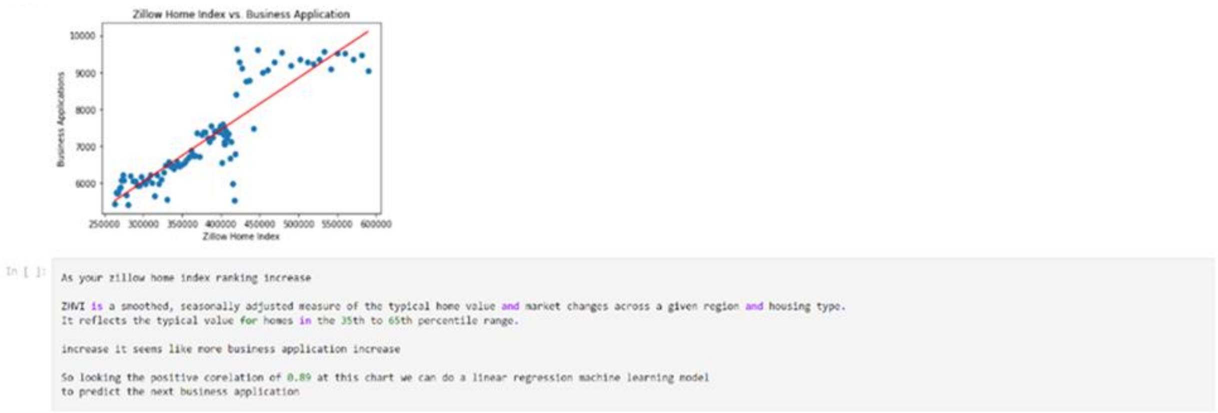
- Coincident Economic Activity Index
- Business Application
- All Employee Non-Farm Payroll
- Revenue
- Labor Force Participation

Zillow Home Index Negative or zero correlation

- SNAP
- Unemployment
- FED FUNDS
- Leisure Hospitality

Based on this correlation matrix chart we can validate inputs or aka features for our machine model to pick which of them will make out model more accurate and choose those inputs so the machine can understand positive or negative relationships of each feature. For example, Zillow Home index will give a zero or negative correlation to SNAP which is government food stamps.

Using supervised learning models, we are going to use one of the most general prediction models which is linear regression to predict business application vs home index. The second part we used a random forest model which another supervised learning to predict if our machine can learn from the one hundred rows of data and five features which have high correlation to home index, wages, and pay to predict any patterns on revenue. So, a machine learns from home, pay, and business participation can predict revenue. Warning due to the lack of data we have our predictions could be extremely low and can overfit. Meaning our machine can memorize our model and become inaccurate or irrelevant. One hundred records are small to make a machine model due to the lack of information the machine to learn and can memorize aka overfitting. See Figure 14



```
As your zillow home index ranking increase

ZHVI is a smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type.
It reflects the typical value for homes in the 35th to 65th percentile range.

increase it seems like more business application increase

So looking the positive corelation of 0.89 at this chart we can do a linear regression machine learning model
to predict the next business application
```

Figure 14: Zillow Home Index vs Business Application

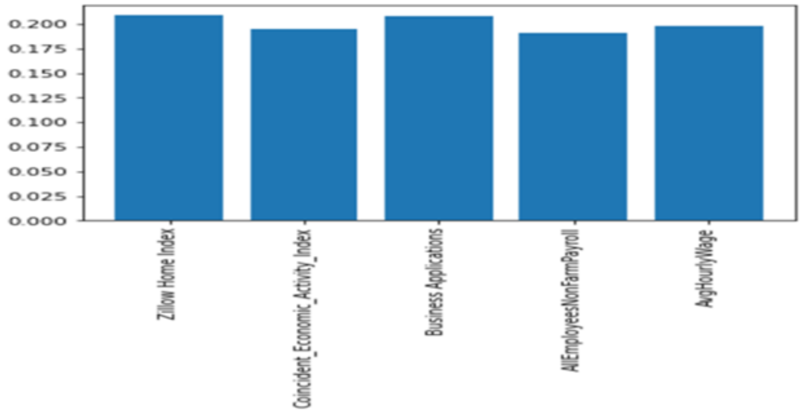
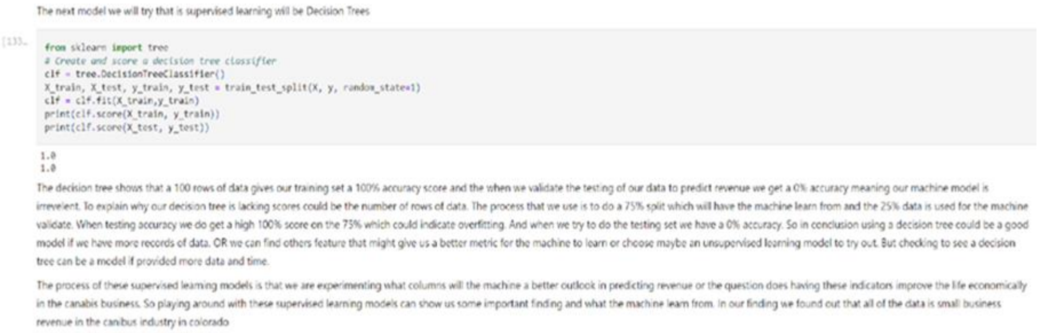


Figure 15: Correlation with five important data points



The next model we will try that is supervised learning will be Decision Trees

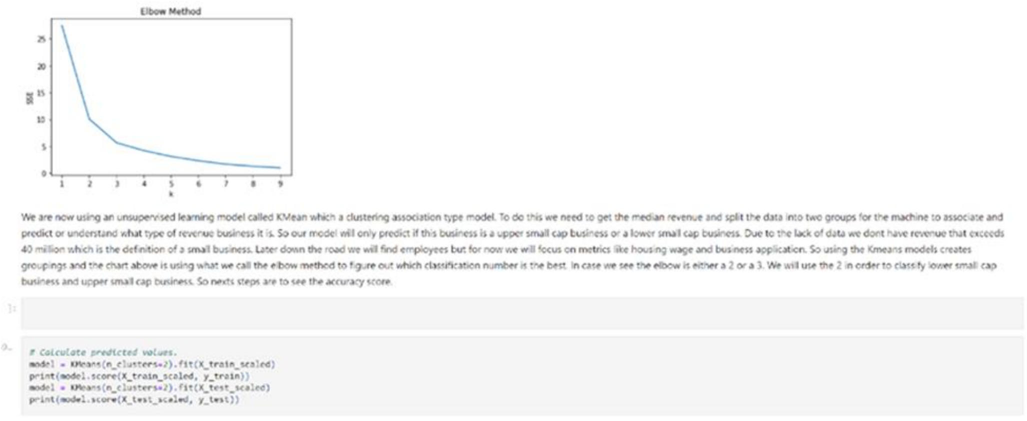
```
from sklearn import tree
# Create and score a decision tree classifier
clf = tree.DecisionTreeClassifier()
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
clf = clf.fit(X_train,y_train)
print(clf.score(X_train, y_train))
print(clf.score(X_test, y_test))
```

```
1.0
1.0
```

The decision tree shows that a 100 rows of data gives our training set a 100% accuracy score and the when we validate the testing of our data to predict revenue we get a 0% accuracy meaning our machine model is irrevelent. To explain why our decision tree is lacking scores could be the number of rows of data. The process that we use is to do a 75% split which will have the machine learn from and the 25% data is used for the machine validate. When testing accuracy we do get a high 100% score on the 75% which could indicate overfitting. And when we try to do the testing set we have a 0% accuracy. So in conclusion using a decision tree could be a good model if we have more records of data. OR we can find others feature that might give us a better metric for the machine to learn or choose maybe an unsupervised learning model to try out. But checking to see a decision tree can be a model if provided more data and time.

The process of these supervised learning models is that we are experimenting what columns will the machine a better outlook in predicting revenue or the question does having these indicators improve the life economically in the canabis business. So playing around with these supervised learning models can show us some important finding and what the machine learn from. In our finding we found out that all of the data is small business revenue in the canibus industry in colorado

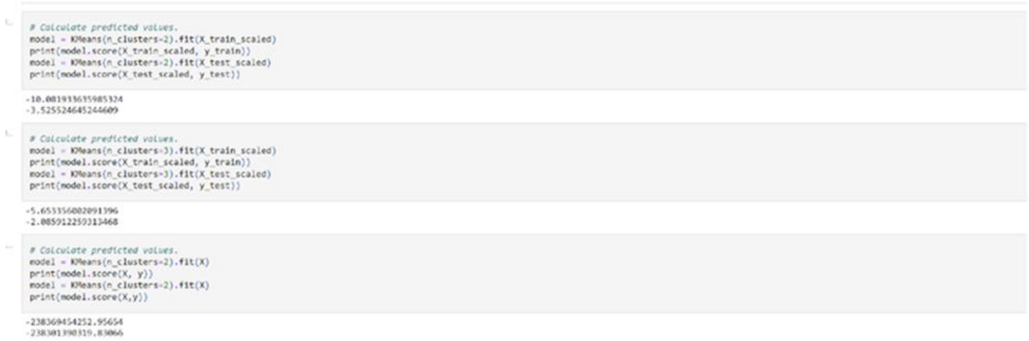
Figure 16: Supervised learning using Decision tree



We are now using an unsupervised learning model called KMean which a clustering association type model. To do this we need to get the median revenue and split the data into two groups for the machine to associate and predict or understand what type of revenue business it is. So our model will only predict if this business is a upper small cap business or a lower small cap business. Due to the lack of data we dont have revenue that exceeds 40 million which is the definition of a small business. Later down the road we will find employees but for now we will focus on metrics like housing wage and business application. So using the Kmeans models creates groupings and the chart above is using what we call the elbow method to figure out which classification number is the best. In case we see the elbow is either a 2 or a 3. We will use the 2 in order to classify lower small cap business and upper small cap business. So nexts steps are to see the accuracy score.

```
# Calculate predicted values.
model = KMeans(n_clusters=2).fit(X_train_scaled)
print(model.score(X_train_scaled, y_train))
model = KMeans(n_clusters=2).fit(X_test_scaled)
print(model.score(X_test_scaled, y_test))
```

Figure 17: Unsupervised learning model using the Elbow Method



```
# Calculate predicted values.
model = KMeans(n_clusters=2).fit(X_train_scaled)
print(model.score(X_train_scaled, y_train))
model = KMeans(n_clusters=2).fit(X_test_scaled)
print(model.score(X_test_scaled, y_test))
```

```
-10.081933635985324
-3.525524645244609
```

```
# Calculate predicted values.
model = KMeans(n_clusters=3).fit(X_train_scaled)
print(model.score(X_train_scaled, y_train))
model = KMeans(n_clusters=3).fit(X_test_scaled)
print(model.score(X_test_scaled, y_test))
```

```
-5.653356002091396
-2.085912259313468
```

```
# Calculate predicted values.
model = KMeans(n_clusters=2).fit(X)
print(model.score(X, y))
model = KMeans(n_clusters=2).fit(X)
print(model.score(X,y))
```

```
-238369454252.95654
-238301390319.83066
```

Figure 18: Machine learning using a predictive model

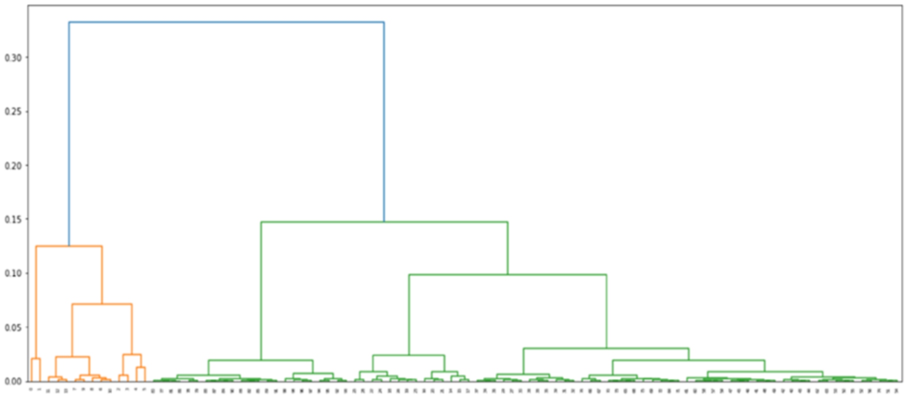


Figure 19: Family tree of two subgroups in the cannabis business using wage, housing, and employment

We did three supervised learning linear regression random forest; decisions trees and two unsupervised learning K means and hierarchical clustering. The above charts show the machine able to classify a family tree of two subgroups in the cannabis business using wage, housing, and employment. We can tell based on a subset which shows in the green one association and in the orange another. Machine learning is based on the one hundred rows of data two classification and can predict if you give the data will choose groups or green. A hierarchical clustering is a machine model that breaks into a family tree format and uses that to classify our groups when given the data. Looking closer you can break down these as grouping of the revenue of the cannabis revenue. Showing that we did have a lot of our data in the lower small gap business model which could be the green while our orange could be the high-end revenue business model.

Finding on Machine Learning

During this analysis we did many techniques to explore if the data we have available can predict or recommend cannabis revenue. Our finding was to make a correlation matrix which shows positive and negative relationships of the data. This is useful because we can find specific columns that can be used in more machine learning models in the future. We started finding Revenue has a high positive correlation of 0.84 with Average Hourly Wage, 0.87 Business Application, 0.85 Coincident Economic Activity Index and 0.86 Zillow Home Index. We did also find a negative correlation in revenue like snap -0.77 and no correlation in FED FUNDS, and unemployment. So, looking through a correlation matrix helps us find a mutual relationship or connection between the columns of our data to apply some machine learning model to predict or recommend.

We started off with a linear regression model, a supervised learning model where we can predict by using trends and fitting a line to predict future outcomes. We started charting Zillow Home Index vs. Business Applications. In doing so we can validate and see the line chart can fit the best fitted line to predict business Application which is a positive indicator to cannabis revenue. Our findings showed a positive correlation, but a high p value in the relationship. We then used scikit to learn to make the linear regression model and predict future business applications to the Zillow home index.

Later, we used other models such as Random Forest and Decision trees. To use this technique and remove any bias we used a train test split method. For the machine to not be biased. We split our data into 75% and 25% validation. And in doing so we got high p values scores of 100% accuracy for our training. When we validate our testing set the results show 100% accuracy. This will mean due to our one hundred columns dataset the machine could have overfitted the data.

Unsupervised learning models we considered and tried out are the K means. Which applies clustering and we wanted to do revenues and created two categories. We used a technique to preprocess the data and our findings on this machine learning model gave negative scores. The p value was extremely low, showing difficulty in accuracy. We tried many diverse types of clustering like 2 classifications to 3 classification and still got negative scores.

The goal of the unsupervised learning model was to find out what type of revenue business it was. To divide the revenue into two groups: low cap revenue businesses or high cap revenue business. We found out the splitting those types made our model biased due to the small amount of data we have and a not so even data set between the groups. This could be the reason for the poor negative p value.

We learned when doing a machine learning unsupervised model, we need to have an even spread of data if not the computer can cluster data and sway into biased decision-making or become completely irrelevant. By having more samples of data and equal parts of the data from both classes not based on the median split could give us better results. The last and final unsupervised learning model is hierarchical clustering where the data is broken into two groups.

For illustration and depiction of all these analyses based on Machine Learning Technique, please refer to Figure 14 through Figure 19.

Conclusion

In conclusion we applied a total of four machine learning models each telling a different story and finding we can use it to explore more to the idea of cannabis revenue. We applied many preprocessing techniques to optimize our results. The first model was linear regression on two features of Zillow Home index and business applications showing a strong correlation between. With that information we moved to decision trees and random forest tree models. With this it gave the most promising scores to evaluate out and our model might have learned something, but it was still optimistic because of our small sample size data showing 100% accuracy which might indicate overfitting.

Later we moved into the two unsupervised learning models and did K mean clustering technique based on revenue. This model had deficient performance when we tried different alternate techniques to improve the p-value scores if the model can even predict something. We found a data problem and our scores were 0% accurate. We then used another clustering technique that would be hierarchical clustering and found the breakdown of the two groups is extremely different showing a gap in revenue. We can apply a better machine model knowing what the breakdown can be instead of the median cut. The best outcome is the decision tree model which helped us get a recommendation system to calculate revenue using the four highly correlated features like home wage and avg pay.

References

1. Hartman, M. (2022, May 31). Cannabis Overview.
2. Seaton, M. (2019, April 19). Collecting the history of marijuana legalization in Colorado. History Colorado.
3. McLeod, S. (2007). Maslow's hierarchy of needs. Simply psychology, 1(1-18).
4. Gharehgozli, O., Nayebvali, P., Gharehgozli, A. et al. Impact of COVID-19 on the Economic Output of the US Outbreak's Epicenter. EconDisCliCha 4, 561–573 (2020).
5. Barkey, P. M., & Sonora, R. (2020). An Assessment of The Market and Tax Revenue Potential of Recreational Cannabis in Montana.
6. Boesen, Ulrik. "A Road Map to Recreational Marijuana Taxation." Tax Foundation (2020).
7. Project, M. P. (n.d.). Cannabis Tax Revenue in States that Regulate Cannabis for Adult Use. MPP.
8. Recreational Marijuana ProCon.org. (2018, November 13). Procon.org; Britannica.
9. U.S. Census Bureau, SNAP Benefits Recipients in Colorado [BRCO08M647NCEN], retrieved from FRED.
10. Zillow, Zillow Home Value Index (ZHVI) for All Homes Including Single-Family Residences, Condos.