



Measuring the Cannabis Economy

Charles River Economics Labs

Shivant Krishnan (Project Lead)

Utkarsh Dandanayak, Kiran Duggirala, Nicolas Marin Gamboa, Cory McCormack
(Analysts)*

Fall 2023

*The authors are from the University of Chicago. They are grateful to to Dr. Katie Genadek, Dr. Rebecca Hutchinson, and Jacqueline Agnes Glasheen at the U.S. Census Bureau for their invaluable insights and guidance. We also thank Evan Rose, John List, and Victor Lima from the Department of Economics for their mentorship and valuable contributions to our project.

Contents

1	Team and Project Overview	3
1.1	Charles River Economics Labs	3
1.2	Partner Organization	3
1.3	Results in Brief	3
2	Introduction	3
2.1	Landscape of the Cannabis Industry	3
2.2	Motivation and Roadmap	4
3	Literature Review	6
4	Data and Methodology	7
4.1	Micro-Level	7
4.2	Aggregate	9
4.3	Classification via Deep Neural Networks	9
5	Economic Analysis	11
5.1	Product Pricing & Regional Heterogeneity	11
5.2	Sales and Taxation	15
6	Policy Implications	16
7	Conclusion	17
8	Appendix	19

1 Team and Project Overview

1.1 Charles River Economics Labs

Charles River Economics Labs is a group of undergraduate researchers at the University of Chicago, Harvard, and Stanford that completes pro-bono economics research for governments and other social enterprises. The organization is advised by graduate researchers and faculty members at Harvard and the University of Chicago. Its past and current research partners include the United Nations Development Programme, the Department of Transportation, the Environmental Protection Agency, the Brookings Institution, and the US Department of State, among others.

1.2 Partner Organization

The US Census Bureau’s Associate Director for Economic Programs (ADEP) team operates under the leadership of Nick Orsini, Associate Director. The directorate conducts more than 60 monthly, quarterly, and annual surveys covering various sectors of the economy, including 12 economic indicators and an economic census and census of governments in years ending in two and seven. The team is particularly interested in characterizing the US Cannabis industry, and requires support in collecting and linking business-level data with their internal databases.

1.3 Results in Brief

In our study of the Cannabis industry in 37 US states, we construct a novel comprehensive dataset to analyze trends in license issuance, dispensary ratings, product pricing, sales, and tax revenues. Notably, we find differences in the relationship between product price and THC content per unit between recent adopters of recreational cannabis (e.g. New York), and early adopters (e.g. California). This suggests that the business dynamics of the Cannabis industry vary significantly based on geographic and demographic features. Our research offers critical insights into the heterogeneous nature of the Cannabis market, emphasizing the implications for policy and business strategies.

2 Introduction

2.1 Landscape of the Cannabis Industry

The US Cannabis industry has rapidly evolved from an underground economy to a complex market with significant economic implications. For example, the worldwide Cannabis market is expected to reach a revenue of \$51.27 billion by the end of 2023, and a market volume of \$102.9 billion by 2028 assuming an annual growth rate of 14.95% (Popovici and French (2014)). The transformation has been characterized by a body of state-level legislation combined with an array of diverse tax regulations. Before characterizing the economic effects of this legislation, it is important to understand how the landscape of the US Cannabis industry has changed over the

past decade, and especially over the past few years. The state of Colorado made headlines when it became the first state to open recreational marijuana shops in January 2012. Since then, many more states have legalized marijuana even though it remains illegal at the federal level. As of December 2023, 37 US states and the District of Columbia have legalized medical marijuana, 20 of which (including DC) have also legalized recreational marijuana. Most recently, Rhode Island and Connecticut have legalized the recreational use of Cannabis for individuals 21 years and older.

Several states have also decriminalized small amounts of marijuana for personal consumption. In 2014, the District of Columbia passed legislation which made possession or transfer of marijuana (of one ounce or less) a civil violation, rather than a state crime. In the years following, states like New Jersey, Michigan, Arizona, Connecticut, Montana, Virginia, and New York voted to pass recreational Cannabis legislation, which was a notable shift in public policy and societal attitudes regarding Cannabis.

2.2 Motivation and Roadmap

It is important to define the "Cannabis Market" within the context of this study, especially due to the presence of substances such as Cannabidiol (CBD), Delta-8 THC, Delta-10 THC, and other hybrids. While chemically related to Delta-9 THC (the primary psychoactive component in Cannabis), these are extremely different from a legislative standpoint. Delta-9 THC is federally classified as a Schedule I drug under the DEA's Controlled Substances Act of 1970. In contrast to substances like Delta-8 THC and CBD, Delta-9 THC has been strictly prohibited by federal law. The 2018 Farm Bill was significant legislation that federally legalized hemp (Cannabis with less than 0.3% Delta-9 THC) and its variants, which include CBD. This has resulted in the proliferation of legal CBD and other non-Delta-9 products, which deliver a similar sensation to Delta-9 THC and are thus attractive to some consumers. In this study, we focus exclusively on the market for Delta-9 THC products, which are the substances legalized for medical and/or recreational use in several states, but have not been federally legalized as of December 2023. This focus allows us to analyze the market dynamics, regional differences, and regulatory impacts related to the *legal* Cannabis market, without the confounding influence of cannabis-adjacent substances. Thus, since products containing Delta-8, Delta-10, and other variants of THC have occupied a legal gray area, they are not the focus of this study.

Our analysis is situated within a landscape of Cannabis industry data that is somewhat incomplete. While there is an abundance of publicly-available macroscopic data, such as statewide revenue trends/figures, the industry lacks a holistic micro-level dataset that allows for analysis on the dispensary level. The available data also contains variations in the available information, with some states only offering current license issuance data, and other states only offering aggregated tax revenues for the current year. These gaps hinder our ability to conduct in-depth economic analysis and understand the finer contours of the market. As a result, we do not employ advanced econometric models or causal inference techniques to study pricing strategies or consumer preferences in the present study. However, by parsing together dispensary-level information from sources such

as WeedMaps and CannaPages, we are able to create a comprehensive dataset that encompasses detailed business information for 8,000+ Cannabis retailers across the United States. Similarly, we collect data on 3,500,000+ products in these stores across seven product categories. Based on the collected data on product types, pricing, demographics, sales patterns, and tax revenue histories, we are able to conduct descriptive analyses and draw preliminary conclusions regarding the interactions between regulatory frameworks (such as tax rate changes and legalization changes) and their corresponding market outcomes. By bridging these data gaps and employing novel web scraping and linkage procedures, we aim to provide a robust foundation for understanding the various economic impacts of Cannabis legalization. This research is particularly crucial in the current context in which the Cannabis industry is evolving faster than ever before, and policy decisions are often made in the absence of comprehensive data-driven insights.

To provide context, we present some facts on the Cannabis Industry in New Mexico, which presents a compelling case study for Cannabis legalization, usage, and industry dynamics. This will be a primary focus of our analysis of dispensary sales metrics. In April 2021, NM Governor Michelle Lujan Grisham signed the Cannabis Regulation Act into law, which allowed adults to legally purchase and possess up to two ounces of cannabis from licensed retailers. We choose new Mexico because the recent policy change demonstrates the short-term effects of transitioning from an illicit to a legal market. As a result, we can examine the early stages of the market without a significant lack of data. In particular, this state was one of the only ones in our dataset which contained detailed sales figures on a quarterly basis. Further, as a state with historically lower economic growth rates compared to national averages, the introduction of the cannabis industry may represent a significant development opportunity. The potential for job creation, business growth, and increased tax revenues is particularly salient for New Mexico’s economy. Finally, the state of New Mexico has already implemented unique regulations on Cannabis, such as microbusiness licenses, which allow for small-scale growing and retail of Cannabis substances). These licenses promote the involvement of small businesses in the Cannabis industry, so we can conduct a time series analysis using sales and business-level data to see how effective this policy has been over the past few years.

In what follows, we will first present a comprehensive literature review on current and previous academic research on the Cannabis industry. We then address our data and methodology, which includes in-depth discussions of the WeedMaps scraping algorithm, public data sources for tax revenue, and an approach to classifying store types via a Convolutional Neural Network (CNN). After this, we will showcase preliminary results from analysis of our novel dataset, specifically pertaining to product pricing, regional heterogeneity, and taxation dynamics. Finally, we discuss the policy implications of our analysis and suggest directions for future research. Threaded throughout the remainder of this paper are case study examples from states such as Washington, California, New York, and New Mexico, which strengthen our claims that there are significant differences between regional Cannabis economies. All of the visualizations and regression outputs associated with this project can be found in the [Appendix](#).

3 Literature Review

In this section, we summarize notable contributions and findings in the existing economic literature pertaining to the Cannabis industry.

A large proportion of the existing literature has sought to estimate demand curves for Cannabis, and examine whether it is a complement or substitute with respect to several other addictive goods industries, including opioids, alcohol, and cigarettes. For example, [Miller and Seo \(2021\)](#) examine potential substitution between Cannabis and alcohol (and between Cannabis and cigarettes) by estimating a differences-in-differences model that compares trends in alcohol and cigarette consumption across counties in Washington, and find that legalization results in a decrease in demand for both cigarettes and alcohol (5% and 15% respectively). Economists studying the Cannabis industry have repeatedly brought up lack of available (price and consumption) data as a significant concern for demand estimation studies (see [Davis et al. \(2016\)](#) and [Jacobi and Sovinsky \(2016\)](#) for example), and this is presumably a concern due to the varying legal status of cannabis consumption across contexts. We would like to emphasize, therefore, that our primary contribution to this body of literature is to develop a source of relevant product-level and dispensary-level information, and a methodology for repeatedly collecting this information over time, in order to facilitate economic analyses of the cannabis industry. Beyond cross-price elasticities, economists have also considered this complement / substitute relationship using alternative outcomes, such as opioid-related deaths ([Livingston et al., 2017](#)) and alcohol-related traffic fatalities ([Hansen et al., 2018](#)).

For further context, we examine recent findings in consumption behavior and tax policy analysis of the Cannabis industry. Studies like [Miller et al. \(2017\)](#) highlight how demographic factors such as age shape marijuana consumption patterns. Their key finding is that students at a University in Washington experienced significant increases in marijuana use after legalization, and the increase is of a larger magnitude for females, black students, and Hispanic students. This study uses a logistic regression methodology to estimate the probability of a student choosing to engage in a certain substance, since it is nearly impossible to track exact consumption patterns among college students in an academic study of this nature. Interestingly, there is no corresponding change in students' consumption of tobacco, alcohol, or other drugs, which suggests a lack of substitution between marijuana and other substances. Further, the increase in marijuana consumption was invariant to the age of the participants, with underage and legal-age students showing similar responses to legalization. This study allows us to hypothesize that marijuana legalization has differential demographic impacts. As such, we use this general framework in order to analyze Cannabis sales/consumption patterns across states with varying times of legalization. In contrast to Miller's study, the present research focuses on tax revenue and total sales figures (only available in some states) to proxy Cannabis usage rather than a likelihood measure for individual Cannabis consumption. Research by [Cooke et al. \(2018\)](#) extends this analysis to market segmentation, suggesting that consumer diversity drives product differentiation and marketing strategies within the cannabis industry. Such studies are pivotal in our understanding of the development of niche markets in the cannabis industry, and how producers can adjust their strategies to cater to a diverse

set of consumers. Similarly, a large body of the existing literature has focused on tax policy. Studies such as [Caulkins et al. \(2015\)](#) have explored the complexities of Cannabis taxation, noting that policymakers face a tradeoff between generating revenue and preventing a resurgence of the black market. The effectiveness of various tax structures, whether based on percentage of price, weight, or THC content, has been a focal point. This approach directly informs our analysis of average dispensary prices by product category, and has inspired us to model the relationship between price and THC content by category. This presents important tax implications due to the fact that some states (e.g. Illinois) have implemented potency-based tax policies. An understanding of these dynamics is critical for states navigating the transition to a legalized Cannabis market, and even more so for states with an existing market/taxation structure in place.

Lastly, our data collection procedure (described below) was informed by existing literature on sourcing online information pertaining to the cannabis economy. For example, [Pedersen et al. \(2018\)](#) document in their study of (medical) marijuana dispensaries that out of several online sources, Weedmaps tended to be the most accurate source of dispensary-level information. As a result, we decided to focus our primary dispensary-level data collection efforts on this source.

4 Data and Methodology

4.1 Micro-Level

The Cannabis industry presents a challenge in data collection due to the numerous public and private sources, often separated by states, which provide detailed business information for Cannabis dispensaries. As such, we focus on collecting micro-level data from three primary sources: licensure, dispensary, and product datasets. Some of these data are at the state-level and need to be merged together to create a holistic picture, while others are scraped on a national level without the need for additional manipulation.

Licensure data was compiled from 17 states that have active recreational dispensaries. We note that this does not represent all states where recreational marijuana has been legalized, because some states are still in the process of operationalizing dispensaries and have not been issued licenses yet. The collection process involved a manual extraction of data from official state government websites. Unfortunately, the collection of data across different states was not uniform. As of December 2023, 13 states have published license numbers, accounting for 88% of all licenses. Similarly, only 11 states published address information, reflecting 67% of all licenses. These data allows us to geographically locate licenses and conduct regional density analyses. However, other important variables, such as license expiration dates, had even less representation. Only five states, including California, Oregon, and Michigan – which collectively host the majority of dispensaries – provided this information, making up 48% of all licenses. This implies that as dispensaries proliferate, states are likely to increase the type of data that they provide which can be used in future research. Unfortunately, the license data currently provided is insufficient for in-depth data analysis. Mapping the issuance and expiration dates over time provides an incomplete picture due to the lack of data.

These data constraints indicate an early stage in the standardization of information disclosure by state authorities. As the Cannabis industry expands, we anticipate that states will enhance the availability and depth of license data. The current state of data, although preliminary, offers a starting point for examining the spatial distribution and regional dynamics of cannabis licensure, and provides a basis for future academic study.

Dispensary data was compiled from the listing of dispensaries available on weedmaps.com. Weedmaps has a dispensary map feature that allows users to search for cannabis dispensaries centered around their current location. The map displays at most 31 dispensaries on one page, and allows users to click through pages to display dispensaries at further distances away from the centered location. The site also contains a set of maps for major cities in each state for the map to center on. So for each state, the first city in alphabetical order was chosen as the states' representative map and all pages of data on the map were searched through (we validated that each city's map contains data on every dispensary, making the choice of map arbitrary). Then for each dispensary on each page, Python Selenium was used to click through to each store page, and BeautifulSoup was used to collect the store's name, phone number, business descriptors (i.e. medical vs. recreational stores, presence of online ordering, etc.), the store's average user rating, and how many reviews were left on the site. Then, the dispensary's "Store Details" button was clicked on, and the dispensary's address and license numbers were scraped. Finally, the dispensary's url was recorded. This process yielded 6701 dispensaries, across the US and Canada (including Puerto Rico). Finally, we searched through the addresses to extract the US state the dispensary was located in, if applicable. This yielded 5217 dispensaries located in one of the 50 states, with dispensaries present in 37 states (see Figure 3 for a map of dispensary counts by state).

Finally, product level information was collected from each dispensary's listed product menu using the Scrapy framework, which uses "spiders" to collect data. For a given dispensary's url, one spider collected data on each listed product's name, price, THC and CBD content, weight, and product category (notable categories including Flower, Edibles, Concentrates, Preroll, and Drinks) as available. After removing duplicates, 2,089,805 products were recorded. Next, information regarding weight and THC content were scraped from the product name for those products lacking explicit weights and THC content. Then, the weights were standardized to units of milligrams (using the same conversion factor as Weedmaps of 28 grams per ounce), and THC content was standardized to milligrams of THC. Finally, the products data was joined with dispensary level data using the url column to allow matching of products to states. Filtering down to products from dispensaries within the 50 US states yielded 1,798,245 products. It is worth noting that not all dispensaries display their product information on Weedmaps: among states with over 50 dispensaries in the dispensary data, 24.6% of dispensaries displayed no product data. The distribution of non-reporting dispensaries is also not even between states. For example, while California had product data for 92.5% of dispensaries, Washington only had data for 33.3% of dispensaries. This likely leads to sampling bias in the analysis of products data.

4.2 Aggregate

Aggregate tax data is available for states that have legalized medical and recreational cannabis. Understanding the ways that states can collect revenue from cannabis taxes is critical to policy suggestions that maximize tax revenue. States tax cannabis in three ways: price based, weight based, and potency based. Price based taxes take the form of excise or sales taxes as a percentage of sales prices. This is the most popular form of taxes, with 21 states implementing some form of price based taxation. State and local governments can both take advantage of such taxes. Some states have low tax rates, such as Missouri at 6%, whereas other states such as Washington tax cannabis at 37%. Weight based taxes are used in Alaska, Colorado, Maine, Nevada, and New Jersey. These states tax the cultivator by weight for different parts of the plant. For example, Alaska taxes cultivators \$50 per ounce for mature flowers and \$25 per ounce for immature flowers. The important distinction here is that cultivators are taxed instead of the final consumer. Lastly, potency based taxes are sometimes levied in addition to other taxes. A key example of this is Illinois, where an additional percentage based tax is proportionally levied based on a product's THC content. The structure of our tax revenue and tax rate data, along with the variations in this data across states, are essential to our analysis.

4.3 Classification via Deep Neural Networks

To further be able to further analyze data from the Cannabis Industry it can be useful to classify the various facets of the industry's ecosystem. Online services such as Cannapages offer such classifications for the segments of the industry they list divying up storefronts and related services as follows: Testing Labs, Commercial Suppliers, Smoke Shops, CBD/Hemp Stores, Temp Agencies, Kratom/Herbal Shops, Dispensaries, Attorneys, Security Companies, Contractors, Grow Stores, Consultants, Doctors, Lounges, CBD Massage, Cannabis Investors, Vape Shops. To be able to port over such classification schemes based on only the store name we introduced deep neural networks commonly used in language classification problems within machine learning. In particular, we attempted temporal convolutional network (TCN) and transformer architectures.

For this task we took the scraped Cannapages data containing store names and the classification label and prepared it for use by the neural network. The industry standard way to do this for language models is to separate sentences into small pieces known as tokens (each accompanied by a token index) and then transforming sentences into length standardized arrays of the corresponding indices. For this project we used SentencePiece [?? \(sen\)](#) Both neural network architectures then go through a proceeding known as embedding which takes the index arrays and turns them into high dimensional vectors that the neural network is able to process to give a probability distribution on the labels. As with most classification tasks we then use cross-entropy loss to train the model. Mathematically, this procedure is as follows, for an array of indices x the model θ defines a probability distribution $p_\theta(y|x)$ on the labels (given by some finite set Y with $y \in Y$) and then we define loss as:

$$\mathcal{L}(p_\theta(y|x)) = \mathbb{E}[-\log(p_\theta(y|x))]$$

where the expectation is taken with respect to the dataset's induced distribution on $X \times Y$. The network is then optimized using the rule

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(p_\theta(y|x))$$

where the gradient is computed through backpropagation and η is the learning rate.

To understand TCNs it is necessary to understand the general family of models they emerge from: Convolutional Neural Networks (CNNs). CNN first became prominent in the context of computer vision and were (until recently) the state-of-the-art for image classification and generation. A traditional (linear/fully-connected) neural network is a series of layers L^i that on input v give

$$L^i(v) = \sigma(W^i v + B^i)$$

Where W^i is some matrix, B^i a vector in the codomain of W^i , and σ some activation function (such as ReLU or a sigmoid) and produce the distribution above by applying in sequence and then (softmax) normalizing the final output which has dimensionality the number of layers.

A convolutional neural network is somewhat more sophisticated. In the setting of images (2D-convolution) the i th layer takes a three dimensional tensor $[I]_{x,y,n}$ of shape X, Y, N (in the image context N is the number of features/channels, for example, in images this starts as 3 corresponding to red, green, and blue) and outputs $[J]_{x,y,i_2}$ of shape X, Y, N' as follows:

$$J_{x,y,n'} = \sigma\left(\sum_{1 \leq i,j \leq s, 1 \leq n \leq N} K_{n',i,j,n}^i I_{x+i,y+j,n} - B_{n'}^i\right)$$

where K^i is a tensor of shape N', X, Y, N and B is a vector of shape N' .

Versions of temporal convolutional networks have been shown to be successful at tasks such as language classification when modified sufficiently. A TCN, introduced in [Bai et al. \(2018\)](#), is a 1 dimensional convolutional network which has been modified to have two extra features: causation and dilation. As 1 dimensional arrays can be interpreted as sequences, TCNs are causal in the sense that the kernel above is modified so that the i th element in the array defined by the layer upon some input depends on convolution with only the first i elements of the array in the input. This creates some mathematical problems which are solved with a process called dilation in which the $x+i, y+j$ used by the kernel to produce the i th entry are replaced by some $x+i*g, y+j*g$ (in the 2D case) where g is some natural number. Modern temporal convolutional networks introduce some tools which have enjoyed great success in other subfields of machine learning such as residual connections [He et al. \(2015\)](#).

The other style of architecture attempted is one native to the world of language models: the transformer. Introduced in 2017 [Vaswani et al. \(2023\)](#), this style of model has been widely used for language generation in Large Language Models. The details of how each layer functions in such a

model would bring us far astray from our goal and thus are not presented here.

Both of these styles of architectures are commonly able to achieve high accuracy rates as seen in Raihan (2021) and Allibhai (2022). However, the use of very similar architectures at a variety of embedding dimensions and distinct learning rate schedules gave both architectures a test error rate of 61.30% upon training. This prompts one of two conclusions: either the data on Cannapages is mislabelled (which is possibly but unlikely to be significant) or that the classification above has classes that overlap in their roles in the cannabis economy as described through their branding. As an example, the models proved inept at separating Smoke Shops from Vape Shops which indicates that the two types of stores share branding and, it would be reasonable to speculate, have significant overlap in their role in the cannabis economy.

5 Economic Analysis

5.1 Product Pricing & Regional Heterogeneity

First, we analyzed the relationship between price and THC content in each state. A correlation matrix on the full product data (see Figure 10) shows very weak correlation between all variables. However, filtering down to looking at only flower products shows strong positive correlations between price, THC content, and weight (see Figure 11). Looking to the correlation between price and THC content in other product categories, we see a correlation of 0.025 for Concentrates products and 0.37 for Edibles and Gummies. The differences in the relationship between THC and price by category indicates the peculiarities of each product category. For concentrates, the extremely weak relationship between THC and price indicates that other factors such as quality of the distillation or flavor are the determiners of price rather than THC content, and this is supported by POM (2023). For edibles, the moderate correlation between THC and price indicates that THC content is one determining factor when consumers pick an edibles product, but other factors like the underlying food being sold or the quality of food are equally important. Finally, strong correlation in flowers product is likely because there is not much other product but the THC in the flower, meaning there is less room to discriminate based on other factors. This means that there is no universal relationship between THC and price, which highly suggests that different categories of cannabis products represent truly differentiated products, and that cannabis is not a single consumer good. Additionally, in both the full product data and flower product data, there is near 0 correlation between the dispensary’s Average Rating and price, and an only weak correlation of 0.077 between the number of reviews and price in the flower products data. This indicates that prices do not significantly vary due to the perceived quality or popularity of a given dispensary, indicating that quality differences occur on the product level rather than the dispensary level.

Since the relationship between THC and price was strongest in flower products, and since flower products were the most represented product in our dataset (see Figure 31), we decided to analyze the relationship between THC content and price between states for flower products specifically. We ran regressions by cutting off products with more than 15 mg of THC or a larger price than \$500,

which never represented more than 0.1% of the dataset and thus represented very clear outliers. The full regression results are in [Figure 33](#), and the regressions for the states of Colorado, California, Illinois, New Jersey, New York, and Washington are pictured in [Figure 15](#), [Figure 13](#), [Figure 17](#), [Figure 18](#), [Figure 20](#), and [Figure 23](#) respectively. We chose these states since they represent three states which were early adopters of cannabis legalization and three states which legalized later. Looking at the graphs, we first note a strong presence of outlier with very lower THC content but very high prices. The particularly large presence of low-THC premium products indicates a strong market for low-THC products, with the differentiator either being the inherent appeal of low-THC products with customers or a different ratio of other cannaboids that the Weedmaps data does not measure. We also see that the linear trend is very strong in later adopter states New York and New Jersey, while the trend becomes weaker in early adopters like Washington and Colorado. We can interpret the varying R^2 between these states as indicators for market diversification in these states: if the R^2 is lower, the market is more diversified as there is more variation in price for a given THC content. So, the initial comparison between early and late adapters seems to indicate that the market becomes more diversified as the market has time to develop. We verify this by regressing the time since legalized sales began against the R^2 of the price-THC regression (see [Figure 27](#)). The positive relationship confirms the idea that the market for flower products becomes more diversified with more time since legalization.

Next, analyzing the data we noticed that there were often repeated product names in the dispensary data, corresponding to the same product being sold by a chain of dispensaries in different locations. These duplicated products often displayed small variations in price and THC content due to the regional effects of . To average out these regional effects, we repeated the regression procedure on the data grouped by product name. The full regression results are in [Figure 34](#), the regressions for the same list of states are pictured in [Figure 14](#), [Figure 12](#), [Figure 16](#), [Figure 19](#), [Figure 21](#), and [Figure 22](#) respectively, and the R^2 on the grouped regression vs time since legalization relationship graph is visible in [Figure 26](#). We compare the R^2 between the original method (called the Raw R^2) and the grouping method (called the Grouped R^2) in [Figure 24](#). This shows that for most states the R^2 is nearly the same as the original R^2 , and that for the rest the R^2 substantially increases. This is likely because the micro regional effects are averaged out of the data, making the relationship between R^2 and THC stronger in those states where the micro regional variations are strongest (eg. Michigan, where we have many many dispensaries recorded). Comparing [Figure 27](#) and [Figure 26](#), we note that the positive relationship between R^2 and time since legalization also improves after grouping by product.

Finally, we note the effect of regional hubs on market diversification. First, we separate Western states' cannabis markets from Eastern states' markets, where Western states are defined as all states containing the Rocky Mountains and westward. This definition makes sense as a distinction between Eastern and Western cannabis markets, since they are completely separated by a string of Great Plains states, none of which have legalized recreational cannabis sales (see [Figure 5](#)). Additionally, by this definition of western states, six large western states had legalized cannabis before the first

eastern states, which suggests that they might have different cannabis markets. So, we display the R^2 vs. Legalization Time plot with an additional color distinction between Eastern and Western states in [Figure 28](#). In the figure, we observe that among states that Western states as a group do seem to have a more diversified market, with late adopters who are also western having lower R^2 than other eastern late adopters. The main outlier to this trend – having an R^2 above 0.8 despite being a western early adopter – is Alaska, which is less likely to reap regional benefits to diversification on account of its geographic isolation and much smaller market. Therefore, we can conclude regional effects play a large part in determining market diversification.

We also present descriptive results on price variations by state for two product categories for which we have the most data (flower and concentrate). [??](#) presents a heatmap of price per gram by state for concentrate products. First, we can note that there does not seem to be a clear relationship between legalization date and prices. There are outliers at both ends of the distribution - for example, Ohio (which legalized recreational use in 2023) seems to have the lowest price per gram (at approximately \$25), while Washington (legalized in 2012) follows closely at approximately \$29 per gram in our data. Perhaps the most striking anomaly here is Montana, with a price per gram of over \$600. Speculating on the reasons for this anomalous result, we suggest that this may be due to a small sample of product data, and that the average is being driven upwards by a small number of expensive outliers. This is possible given that Montana can be considered a late legalizer (2020) and therefore Weedmaps data for this state may be limited (we report a store count of approximately 140 from Weedmaps for Montana, however many stores do not report a “menu” of products and their respective prices - this is described in more detail in the Data and Methodology section). Other geographic determinants of price that might explain the distribution in [??](#) include cost of living, incomes, supply and availability of other (illicit) drugs, and regulatory environments. We display analogous price results for flower in [Figure 1](#). It is interesting to note that flower prices correspond slightly more closely to legalization dates by state - many of the early legalizers (such as Colorado, Washington, and Massachusetts) appear at the lower end of the price distribution (at approximately \$8.5, \$10, and \$10.5 per gram respectively), with Oregon being the most prominent exception (which has the highest price per gram at \$53). Furthermore, relative to concentrate products, regions at or close to the west coast seem to have comparatively higher average prices than their east coast counterparts.

Beyond prices, we study regional discrepancies in the characteristics of the cannabis economy by looking at store counts and average ratings by state. [Figure 2](#) depicts the number of dispensaries listed on Weedmaps by state. It is important to note that although Weedmaps does not list every licensed dispensary in the US, this plot still gives us a reasonable approximation of the relative state-level distribution of stores, assuming that the site does not significantly overreport or underreport dispensaries for certain states (and we do not have reason to believe that this would not be the case). We highlight two interesting descriptive findings from [Figure 2](#).

First, note that store counts for many states remain quite low despite legalizing recreational use a few years ago (e.g. consider Illinois, New York, Vermont, New Jersey). We speculate that

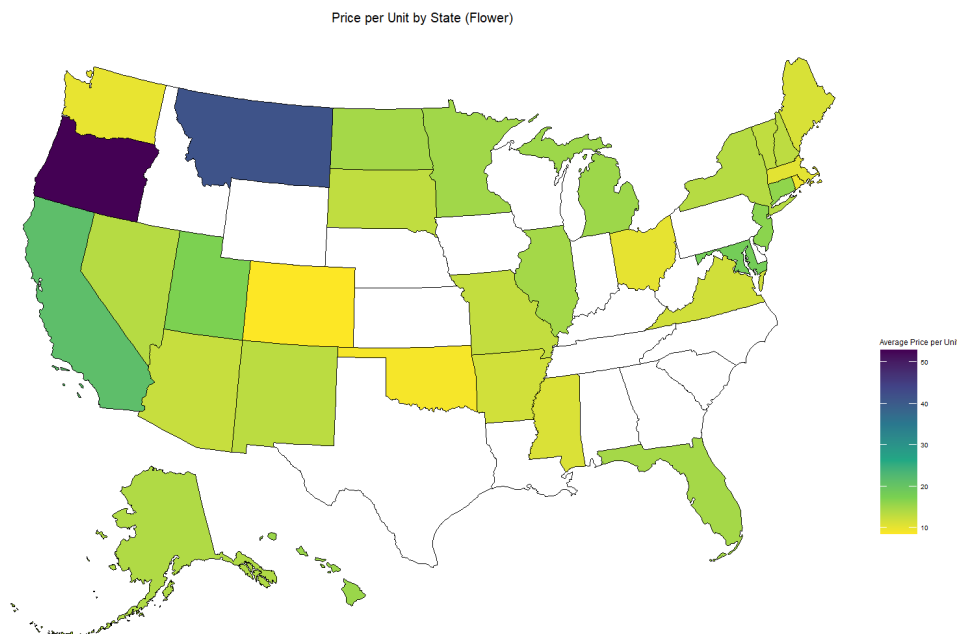


Figure 1: Average Price per Unit (Flower).

this phenomena might be driven both by slower rates of business creation and business growth during the COVID pandemic, as well as the fact that there is often a nontrivial time lag between the date of legalization and the first sale of legal recreational cannabis. The latter is likely a consequence of the time commitment associated with developing a formal licensing system for cannabis dispensaries (see Data and Methodology section for further discussion). Second, in this map we also observe the case of a notable outlier, namely Oklahoma. Despite the fact that Oklahoma has yet to legalize cannabis for recreational use (and legalized medical use cannabis in 2018, much later than many other states), we find more than 650 Weedmaps-listed dispensaries for the state (2nd only to the 833 dispensaries for California). Again it is difficult to prescribe causal determinants of the high dispensary growth in Oklahoma specifically, but one potential driver of this result might be the comparatively more lenient regulatory framework employed in the state (e.g. there are no specific qualifying conditions that are required in order to obtain a medical cannabis card).

Lastly, from our Weedmaps data collection we also obtained average ratings for each dispensary, allowing us to compare store ratings across states. Figure 6 gives average Weedmaps store rating by state. Although there is generally little variation in average ratings, the outliers at the low end of the distribution tend to be states which have legalized more recently (e.g Virginia, New York, Minnesota). This is perhaps unsurprising, and might be attributed to the initial “growing pains” that are associated with business development in novel industries.

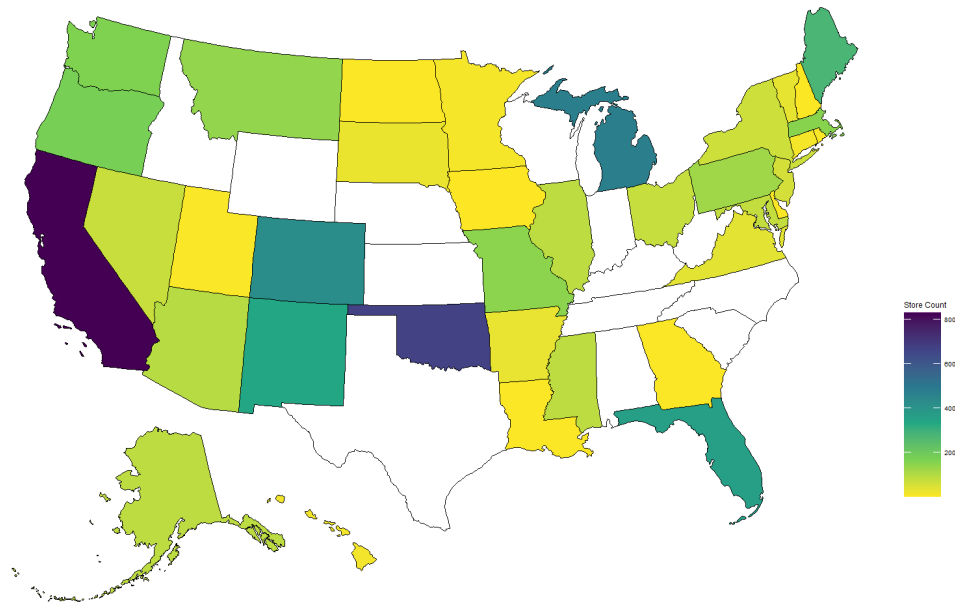


Figure 2: Store Count per State.

5.2 Sales and Taxation

First, in Figure 4 we see that there is a positive relationship between the number of dispensaries and total tax revenue by the state, although there are large error bounds in the prediction. Since many states with lower dispensary counts are recent legalizers, this positive relationship suggests that as the cannabis market grows and diversifies and dispensaries proliferate in response, states can expect an increase in tax revenue. Additionally, when looking at the relationship between the tax rate and the slope between price and THC content (across all states), we see little to no correlation, which is a surprising result. The data points are relatively spread out across the range of slopes, while the tax rate remains narrow. Since tax rates do not have a strong association with how the price of Cannabis products respond to THC content, this suggests that tax rates do not necessarily dictate the market’s valuation of THC content. This may be a result of the structure of the tax regime in most states, which are uniform excise taxes on wholesalers and retailers and are thus agnostic of THC. However, running a similar regression between tax rate and the intercept term in the THC vs price regression – which indicates the effect of tax rate on the baseline price level based off of which THC prices are determined – yield a similarly null result, meaning that the uniform tax regime also does not raise prices as a whole. Therefore, we cannot definitively conclude the effect of tax rates on the prices, which might be due to peculiarities of the cannabis economy that we are unable to measure from our available dataset. This result might also be a result of the confounding variables of uneven legalization regimes and timelines across states which obscure the effect of tax specifically on cannabis prices.

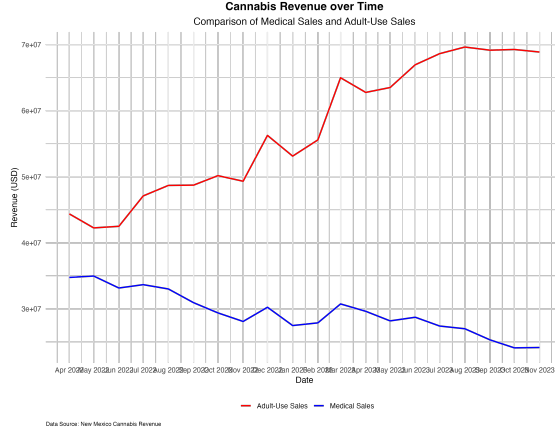


Figure 3: New Mexico Sales Volume by Sales Type.

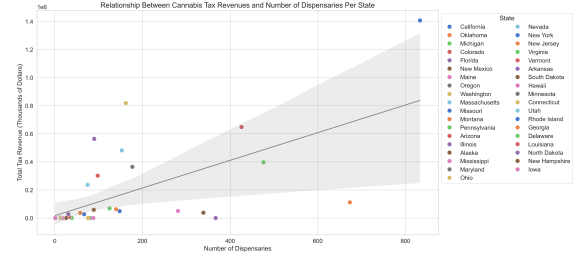


Figure 4: Tax Revenue vs. Number of Dispensaries.

6 Policy Implications

Based on the descriptive analysis above, we can suggest several policy implications within the domains of taxation strategies/regulation, production decisions, and market dynamics.

From our results on sales in Figure 3, we confirm our intuition that the volume of recreational Cannabis sales increases after legalization in New Mexico, while the volume of medicinal sales decreases. This separation of market trends suggests that regulatory frameworks could be adjusted so that medical patients retain access to medical Cannabis, possibly through price adjustments for those products or tax exemptions. The need for such a policy would depend on the patient demographics of New Mexico, but the observed shift in consumer preferences could be leveraged by policymakers to better serve the medical Cannabis population. Additionally, the observed relationship between tax revenue and number of dispensaries indicates that market expansion is beneficial for fiscal purposes. Policymakers could consider issuing more micro-business licenses to "kick-start" dispensaries and growing sites, which would be particularly influential in areas with sizeable illicit markets.

Now, returning to our analysis of the relationship between price and THC content per unit, we can suggest implications on the production side. The relationship between THC and price varies both across states and across product categories. To reiterate, we find that the strongest relationships exist between price and THC content for products in the "flower" category, and the trend is stronger in late-adopter states (e.g. NY, NJ) while it is weaker in early-adopter states (CA, WA, CO). The variance across states and product categories suggests a differentiated market, where consumer valuations of THC content (and willingness-to-pay) are heterogeneous. In states where the relationship between price and THC is stronger, producers can increase the proportion of high-THC products sold. Such price discrimination based on product potency could allow producers to increase profit by meeting the higher willingness-to-pay for THC with a higher price. On the other hand, results from early-adopter states like California suggests that other product qualities (apart from THC) may have an influence on product price, especially for flower. This allows us to

conclude that producers offer highly differentiated products, and that consumers place a premium on this diversity. Perhaps state-specific strategies that cater to heterogeneous preferences can prove beneficial for both consumers and producers of Cannabis. Developing a balance of policy measures and production strategies is critical to ensuring that the Cannabis industry is economically vibrant while also aligning with consumer welfare.

7 Conclusion

The present study provides a foundational and descriptive understanding of the Cannabis market, but it also reveals areas for future study. For example, studies could estimate the demand curves for "indoor-grown" vs "outdoor-grown" Cannabis products, a common distinction between types of marijuana flower. If consumers strongly preferred diversity in products, then we'd expect to see different demand curve elasticities for each of these groups. Such insights could be beneficial for both production and regulation policies. Additionally, further linguistic analysis can be run on low and high price products at the same THC level to uncover which factors lead to price differentiation other than THC content, giving deeper insight into consumers' emerging preferences as the cannabis economy diversifies. Finally, based on our analysis of the different pricing properties of different product categories, future studies could attempt to cluster different product categories together into groups of products that share a similar market and similar demand properties, enabling researchers to construct a representative basket of cannabis goods that encompasses all major types of cannabis products.

However, there are several limitations of our study. We lacked time series data on sales and product prices, which prevented an in-depth analysis of price shocks and revenue dynamics. The reliance on cross-sectional data limits our ability to employ causal inference techniques and investigate the relationships between legalization and several outcome variables. Similarly, the lack of demographic information in our consumer data prevents a rigorous analysis of consumer profiles, which can have important implications for Cannabis firms' marketing strategies. Another key limitation in our product data analysis was the inability to differentiate between specific products in terms of popularity and specific sales numbers, making strong conclusions about how consumer preferences reacted to different products and pricing regimes difficult. Finally, we were limited by the uneven data availability between states, from the difference in product data availability on Weedmaps to the lack of uniformity in licensure data between states. This made strong conclusions about the effect of states' different policy decisions on the cannabis economy hard to ascertain since we could not always make apples-to-apples comparisons between different states' data.

References

- Sentence piece. <https://github.com/google/sentencepiece>.
(2023, December). The Difference Between Low-Cost and High-Cost Grams of Concentrate.
- Allibhai, J. (2022, Jul). Introduction to text classification using transformers.
- Bai, S., J. Z. Kolter, and V. Koltun (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling.
- Caulkins, J. P., B. Kilmer, M. A. R. Kleiman, R. J. MacCoun, G. Midgette, P. Oglesby, R. L. Pacula, and P. Reuter (2015, January). Considering Marijuana Legalization: Insights for Vermont and Other Jurisdictions. Technical report, RAND Corporation.
- Cooke, A., B. Freisthler, and E. Mulholland (2018, July). Examination of Market Segmentation among Medical Marijuana Dispensaries. *Substance use & misuse* 53(9), 1463–1467.
- Davis, A. J., K. R. Geisler, and M. W. Nichols (2016, June). The price elasticity of marijuana demand: evidence from crowd-sourced transaction data. *Empirical Economics* 50(4), 1171–1192.
- Hansen, B., K. S. Miller, and C. Weber (2018, March). Early Evidence on Recreational Marijuana Legalization and Traffic Fatalities.
- He, K., X. Zhang, S. Ren, and J. Sun (2015, Dec). Deep residual learning for image recognition.
- Jacobi, L. and M. Sovinsky (2016, August). Marijuana on Main Street? Estimating Demand in Markets with Limited Access. *American Economic Review* 106(8), 2009–2045.
- Livingston, M. D., T. E. Barnett, C. Delcher, and A. C. Wagenaar (2017, November). Recreational Cannabis Legalization and Opioid-Related Deaths in Colorado, 2000–2015. *American Journal of Public Health* 107(11), 1827–1829.
- Miller, A. M., R. Rosenman, and B. W. Cowan (2017, August). Recreational marijuana legalization and college student use: Early evidence. *SSM - Population Health* 3, 649–657.
- Miller, K. and B. Seo (2021, March). The Effect of Cannabis Legalization on Substance Demand and Tax Revenues. *National Tax Journal* 74(1), 107–145.
- Pedersen, E. R., M. Zander-Cotugno, R. A. Shih, J. S. Tucker, M. S. Dunbar, and E. J. D’Amico (2018). Online Methods for Locating Medical Marijuana Dispensaries: Practical Considerations for Future Research. *Cannabis (Research Society on Marijuana)* 1(2), 22–35.
- Popovici, I. and M. T. French (2014, April). Cannabis Use, Employment, and Income: Fixed-effects Analysis of Panel Data. *The journal of behavioral health services & research* 41(2), 185–202.
- Raihan, D. (2021, Apr). Deep learning techniques for text classification.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2023). Attention is all you need.

8 Appendix

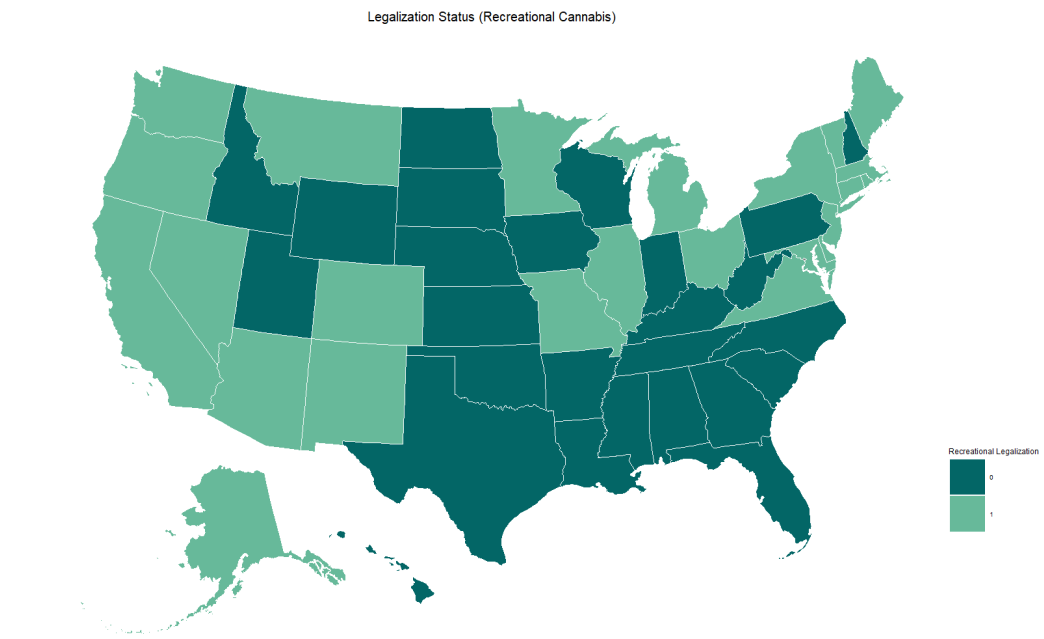


Figure 5: Legalization status by state (1 indicates state has legalized for recreational use)

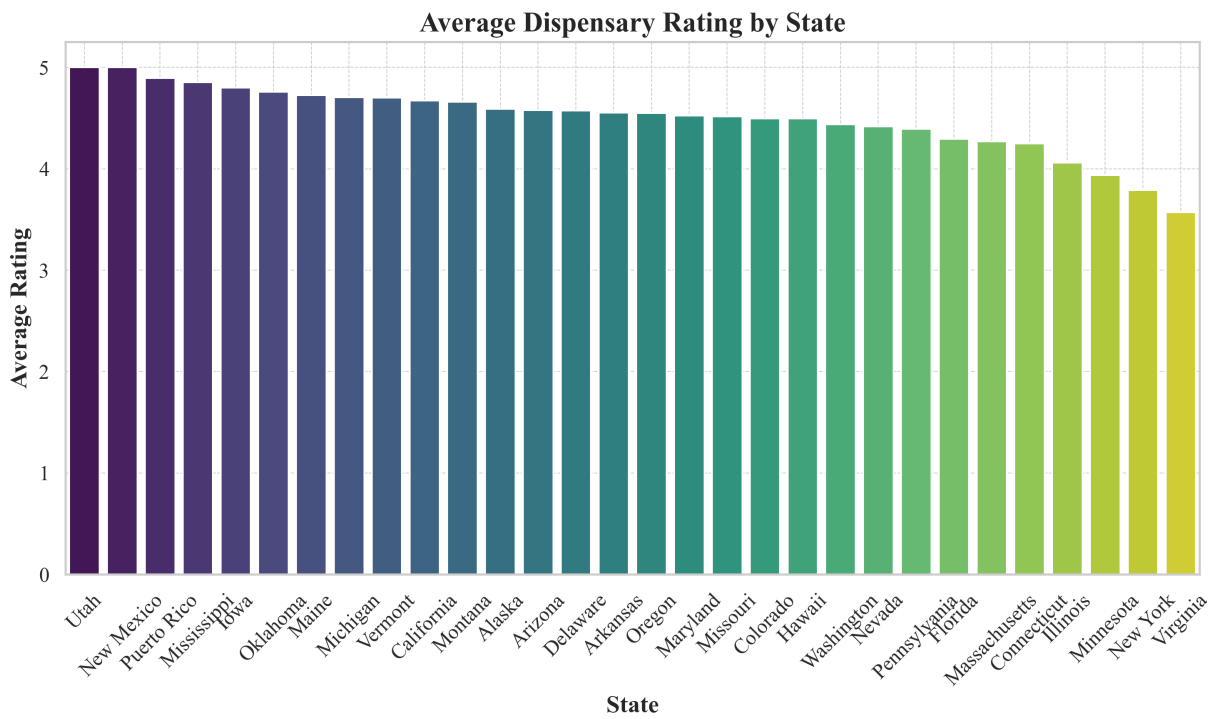


Figure 6: Average Ratings by State.

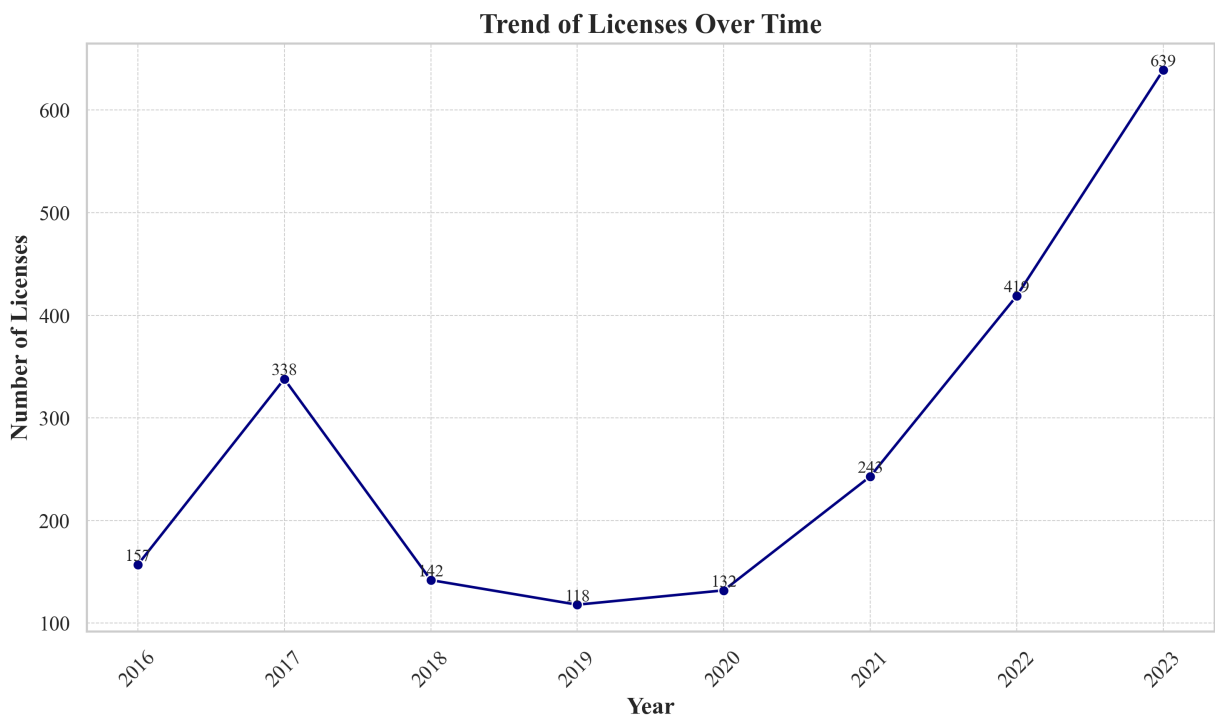


Figure 7: License Trend Over Time.

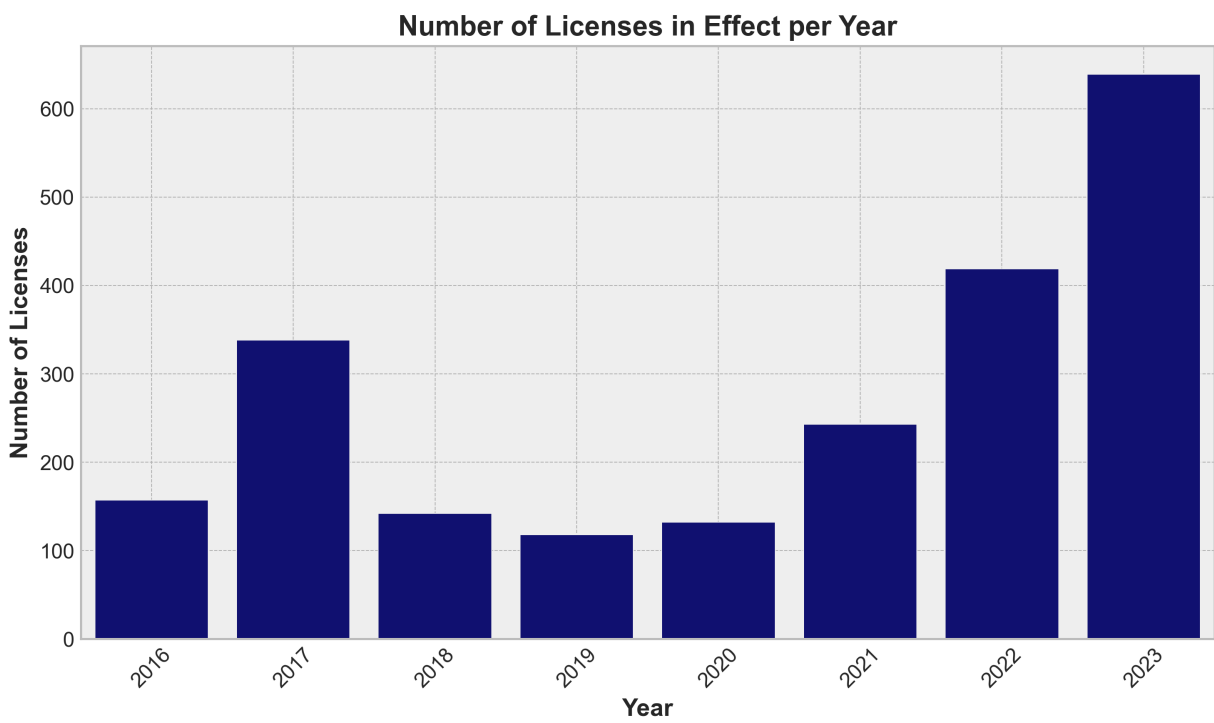


Figure 8: Licenses by Year.

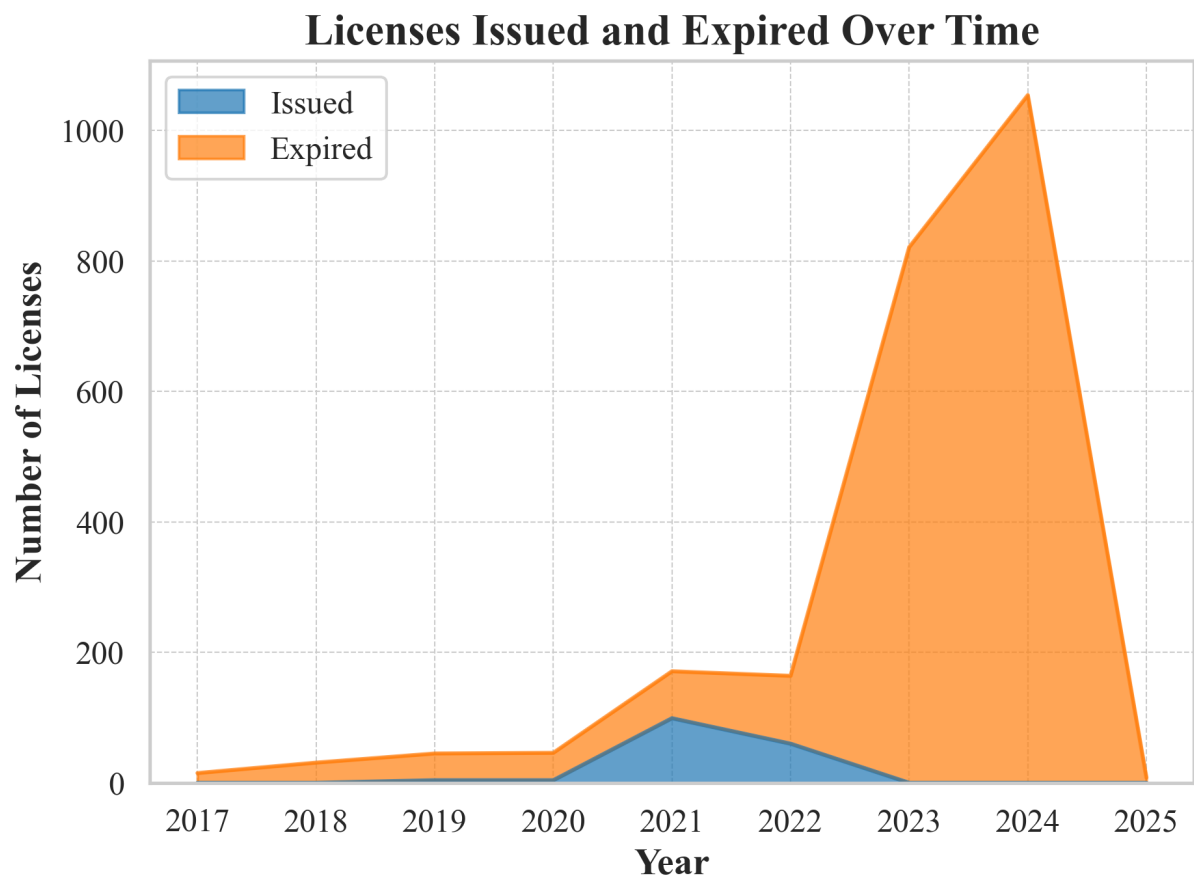


Figure 9: Licenses Issued and Expired Over Time.

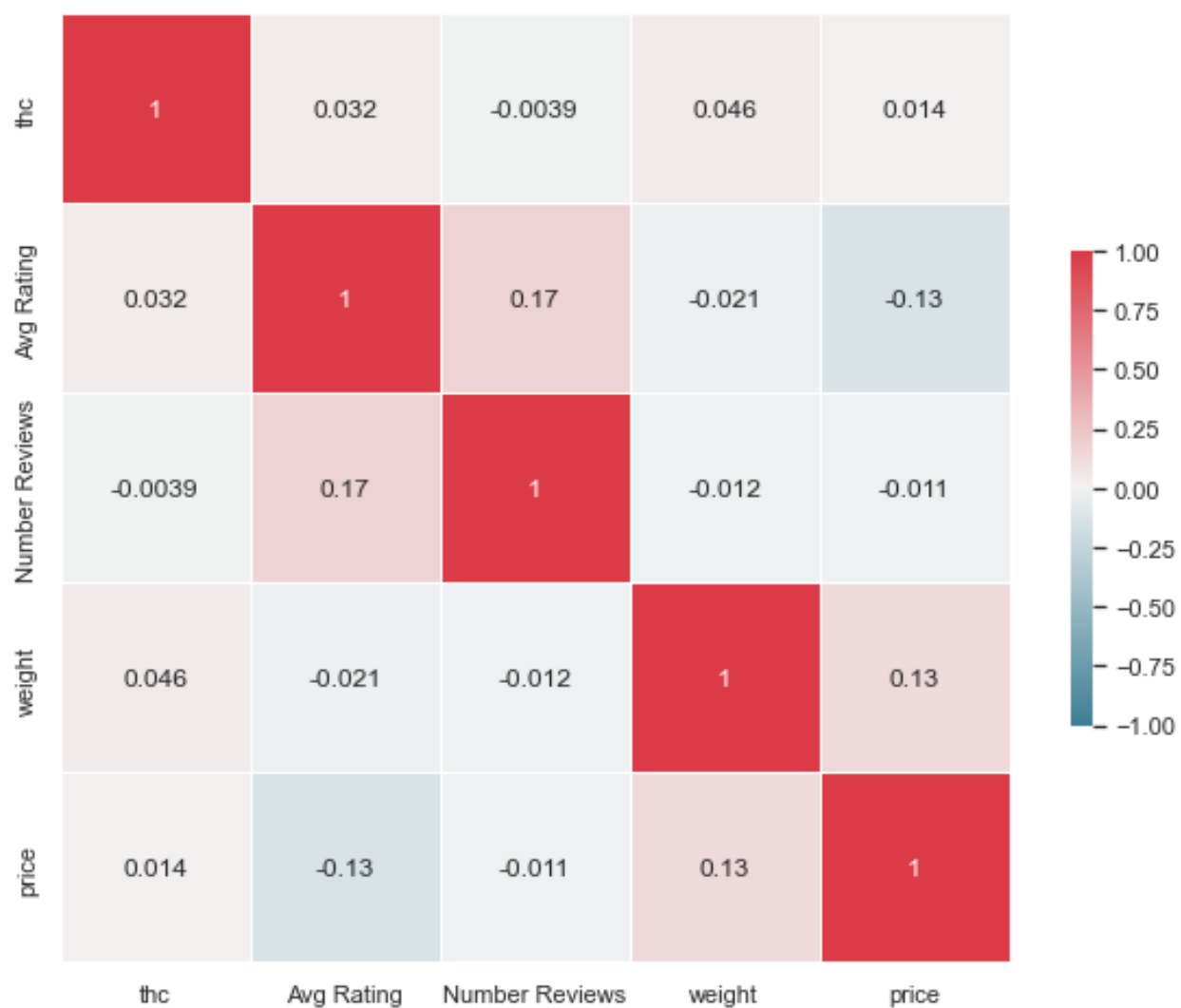


Figure 10: Full Products Data Correlation Matrix.



Figure 11: Flower Products Data Correlation Matrix.

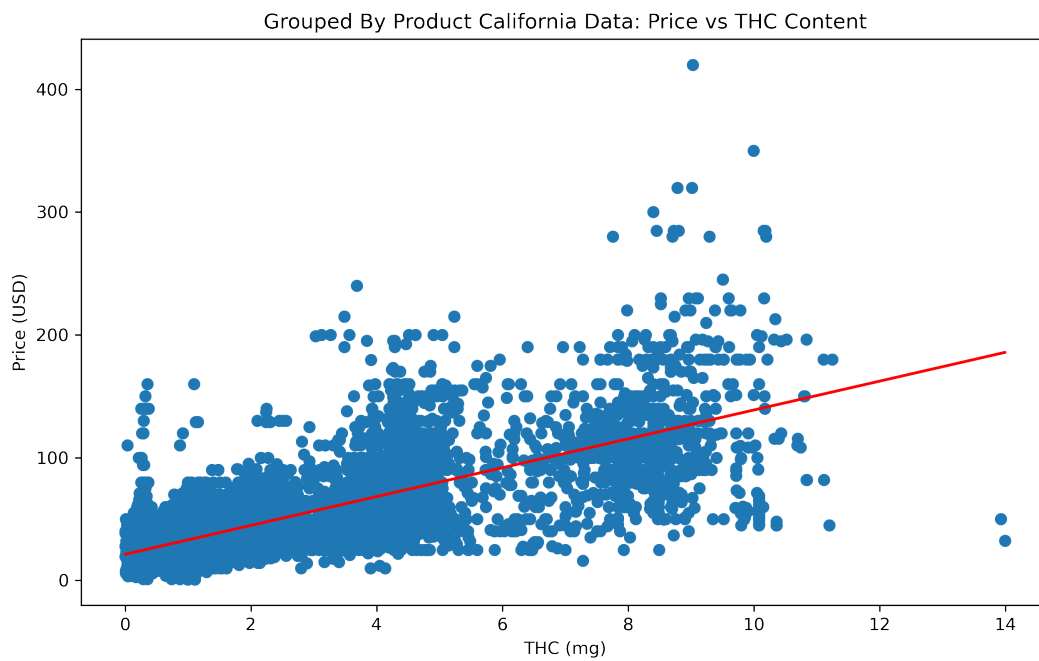


Figure 12: California Grouped Regression.

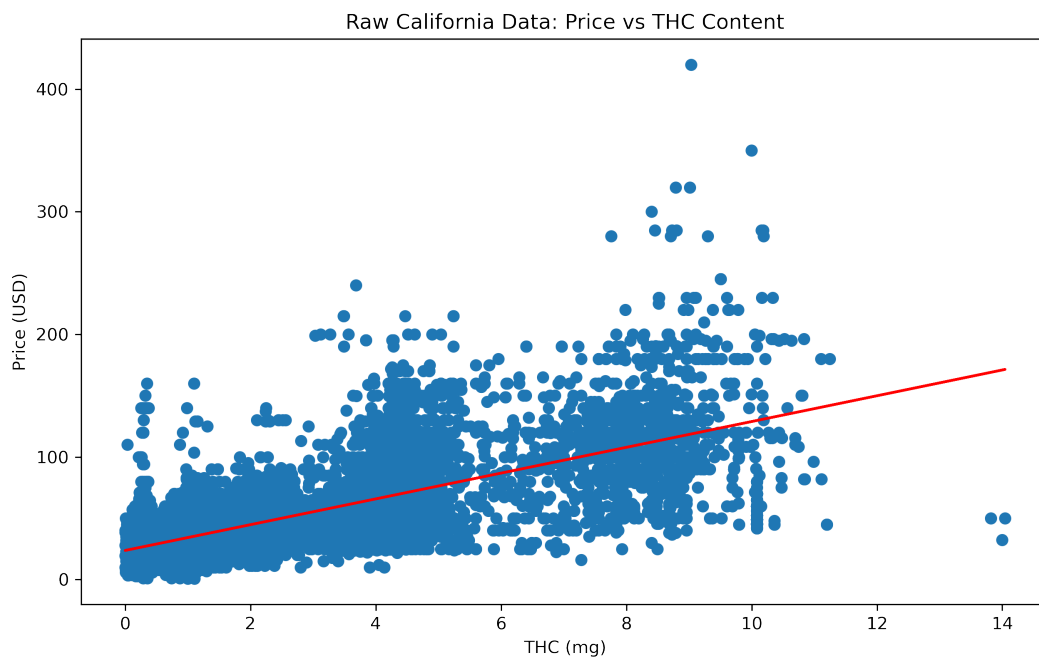


Figure 13: California Raw Regression.

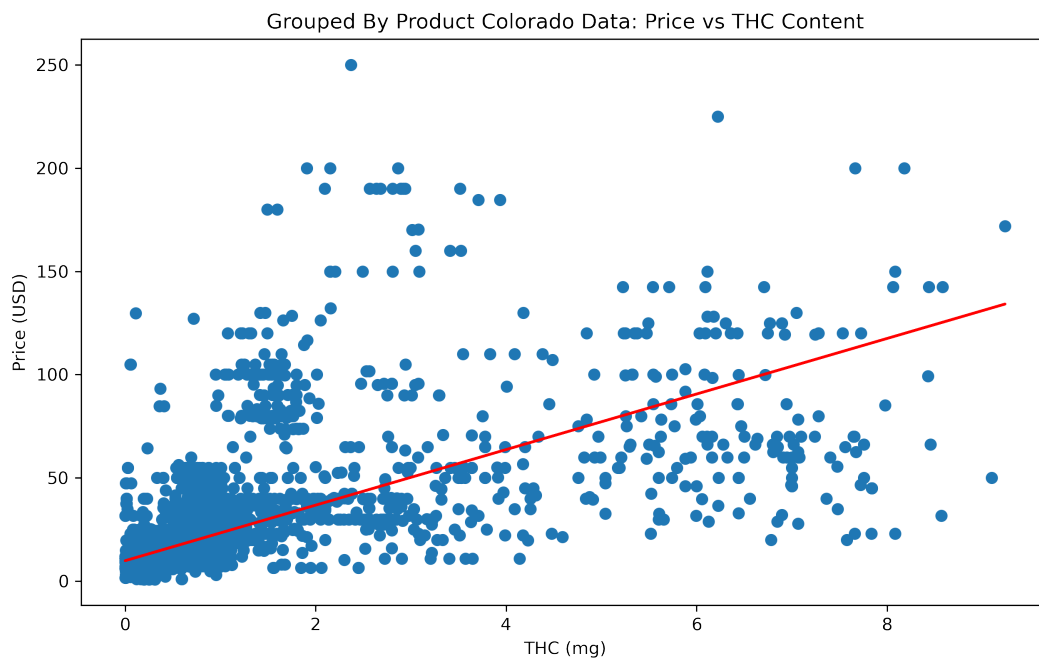


Figure 14: Colorado Grouped Regression.

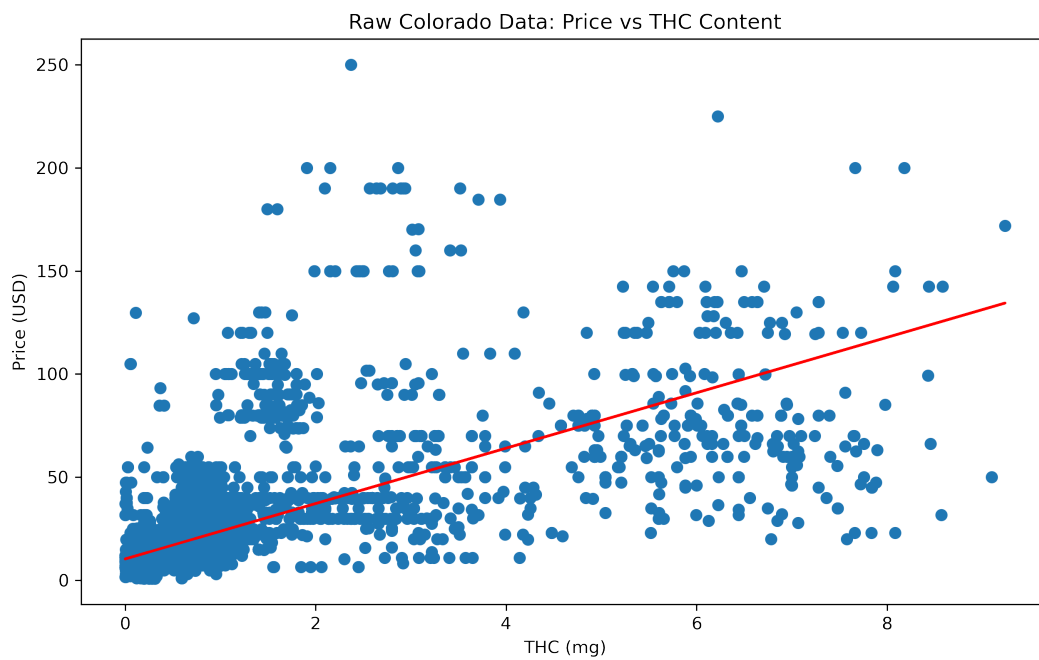


Figure 15: Colorado Raw Regression.

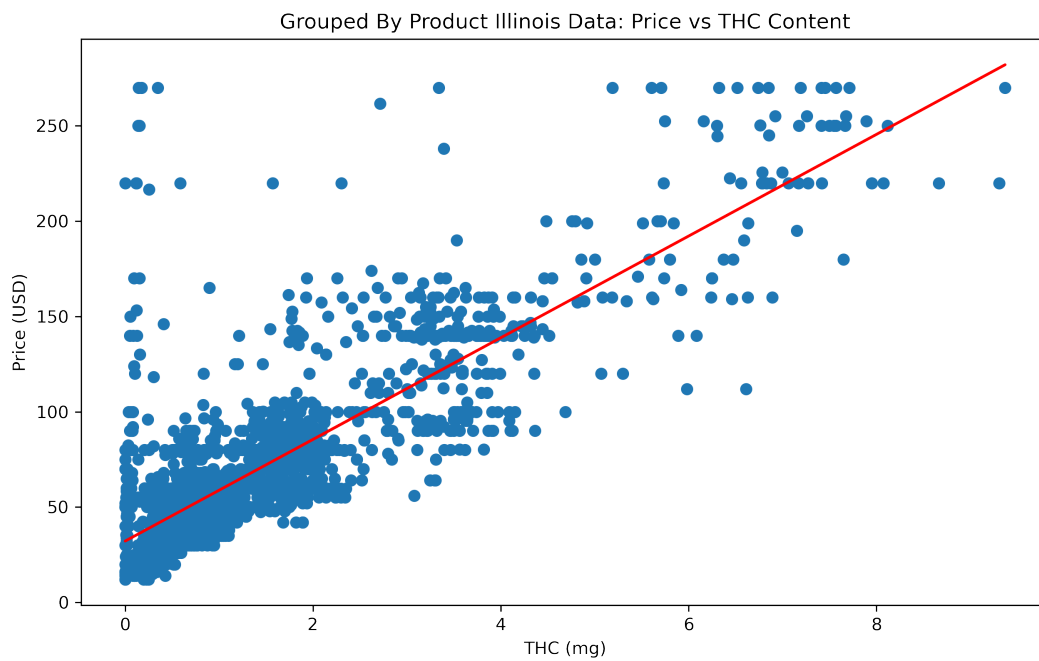


Figure 16: Illinois Grouped Regression.

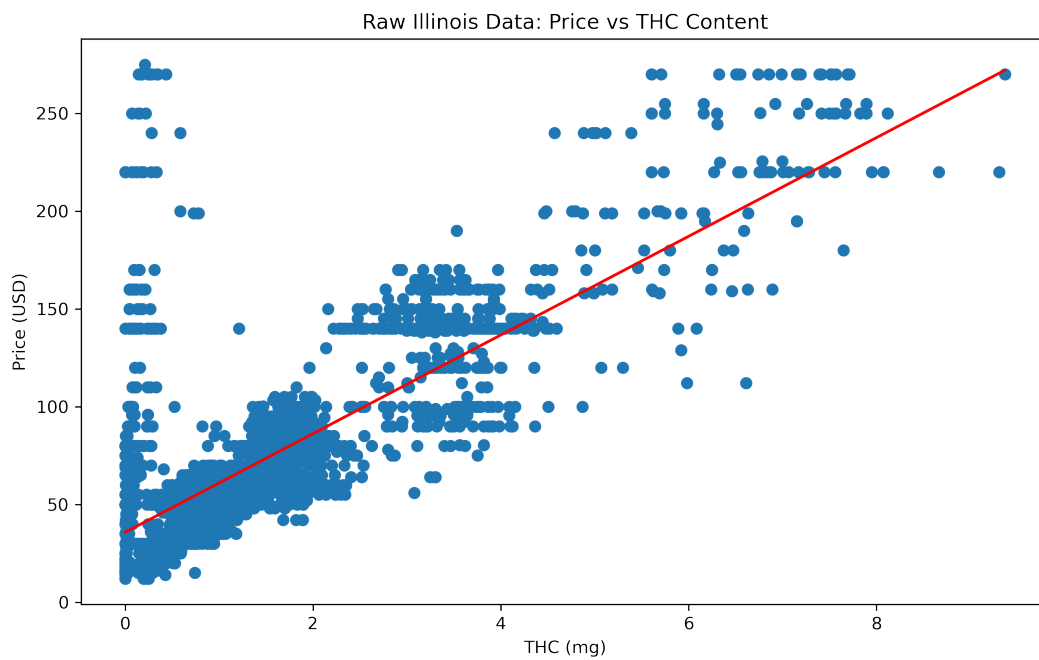


Figure 17: Illinois Raw Regression.

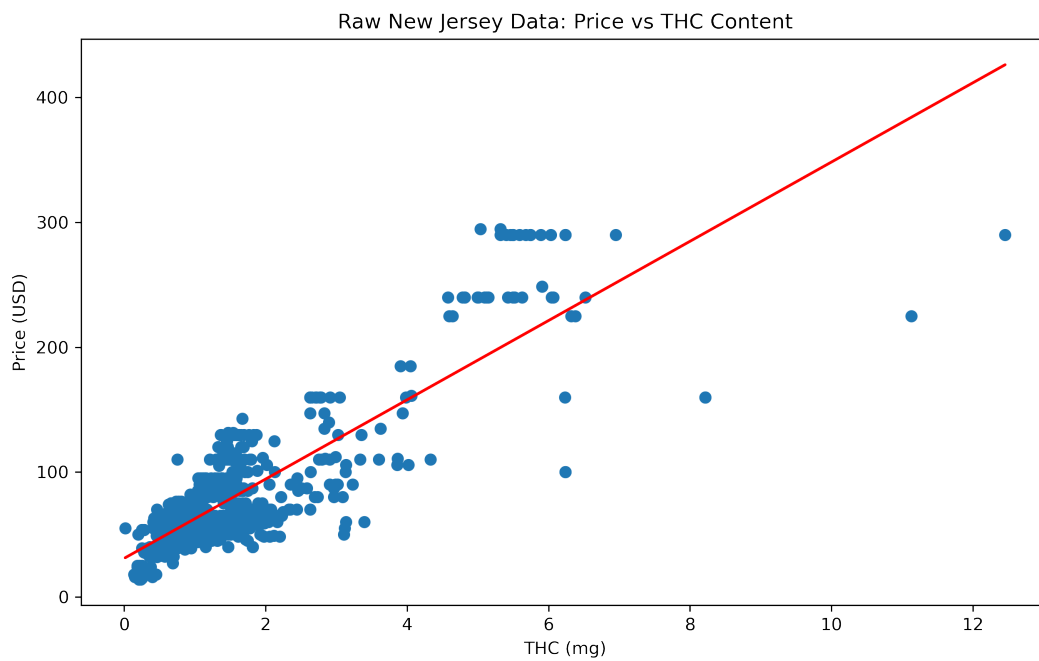


Figure 18: New Jersey Raw Regression.

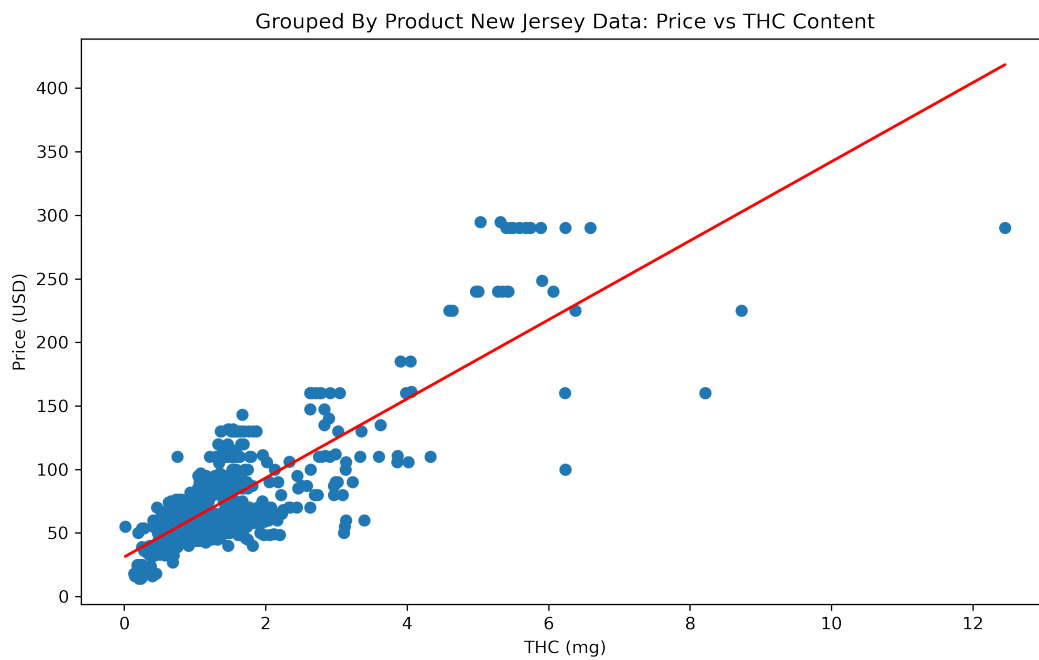


Figure 19: New Jersey Grouped Regression.

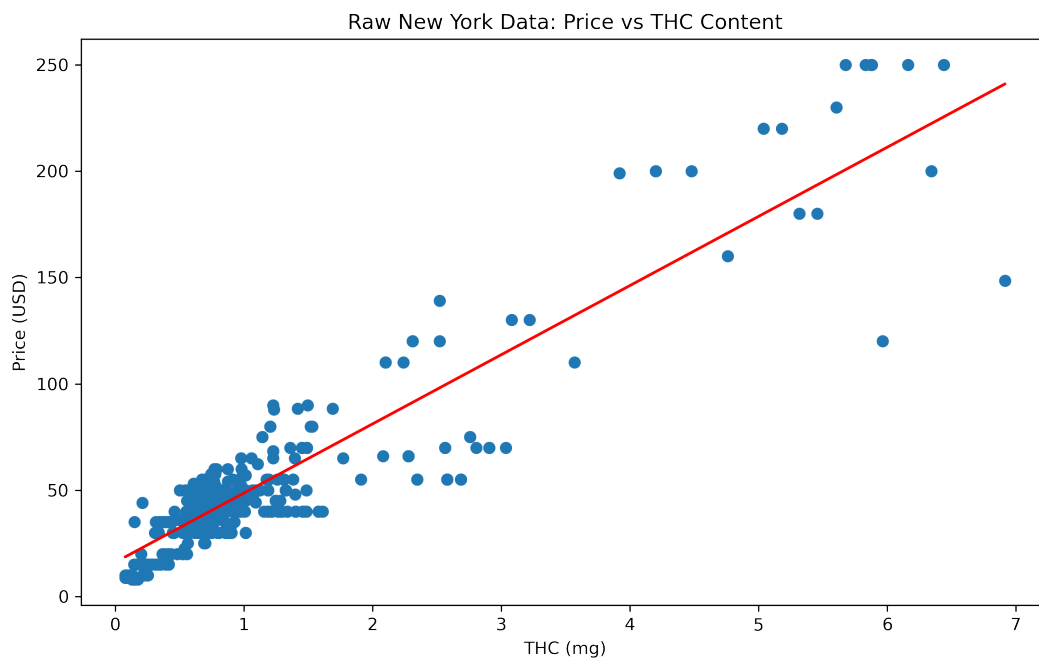


Figure 20: New York Raw Regression.

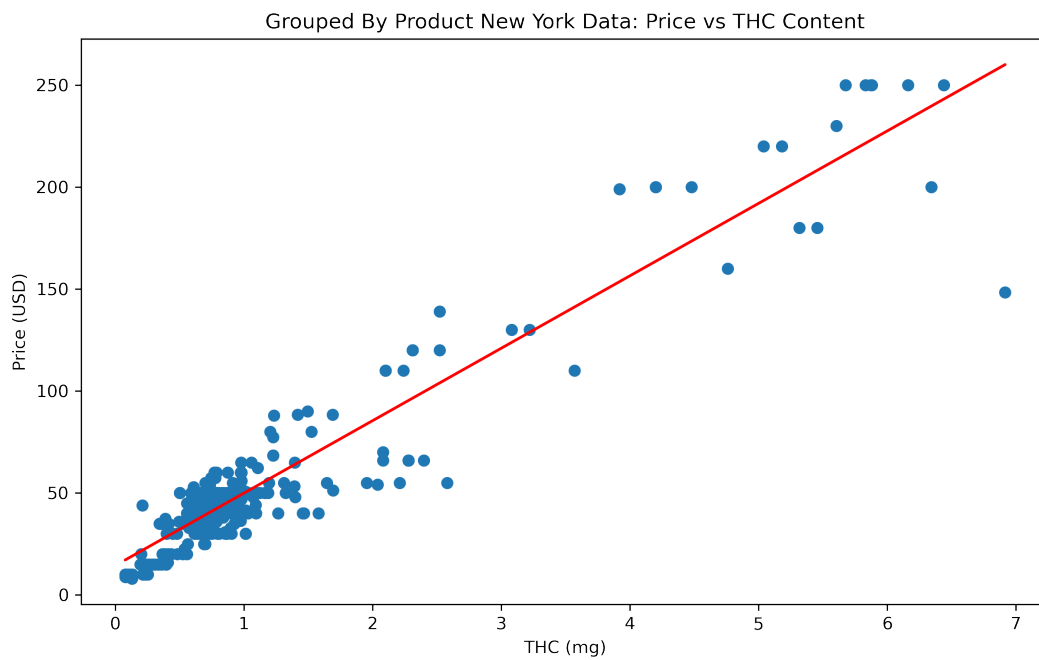


Figure 21: New York Grouped Regression.

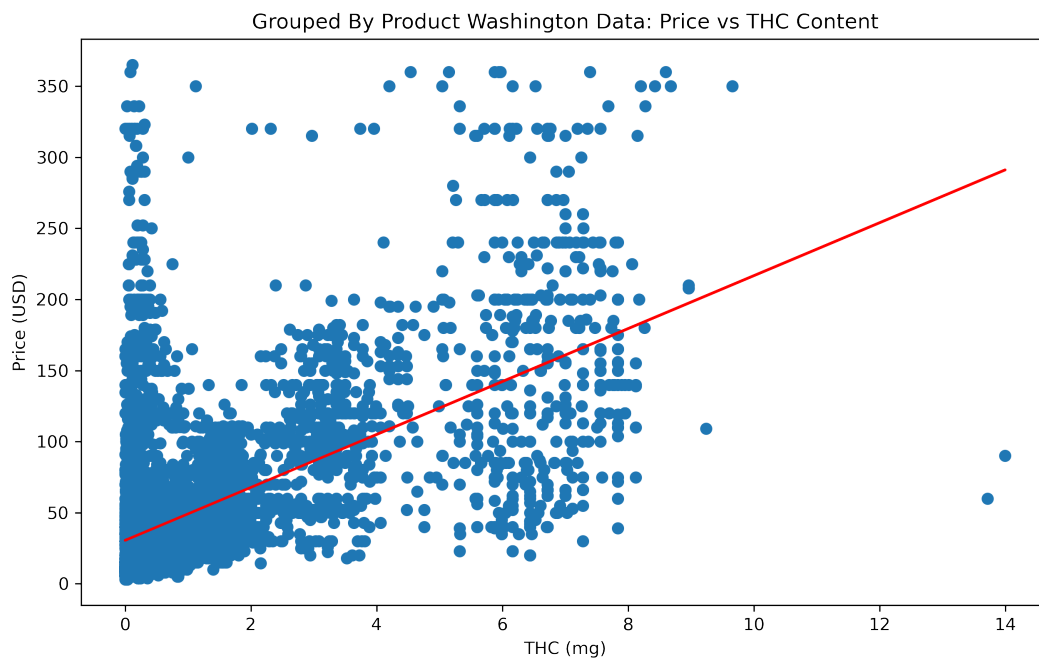


Figure 22: Washington Grouped Regression.

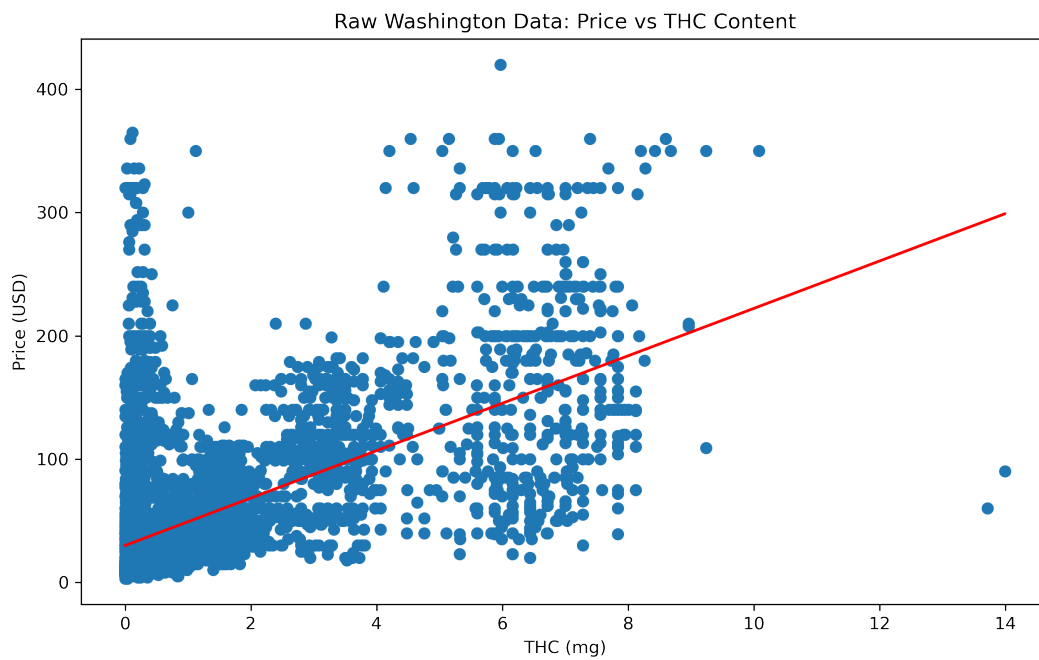


Figure 23: Washington Raw Regression.

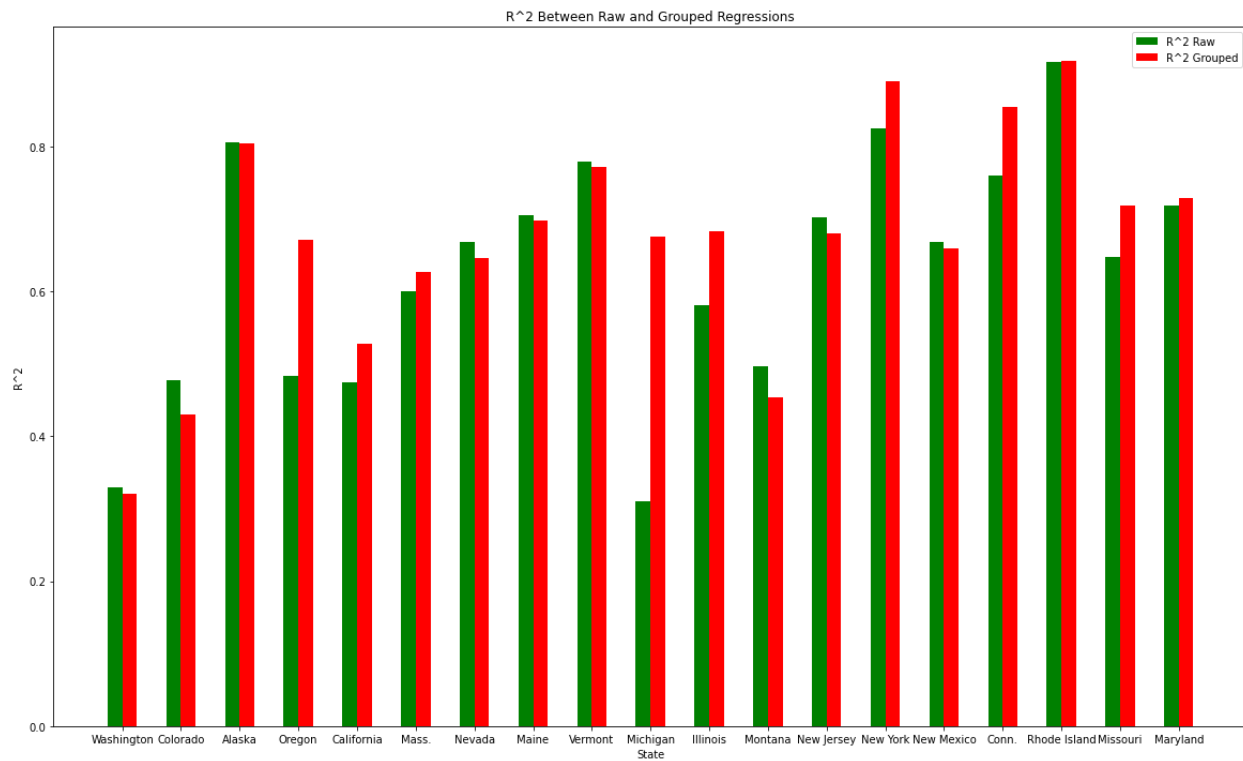


Figure 24: R^2 in Raw vs. Grouped Regressions

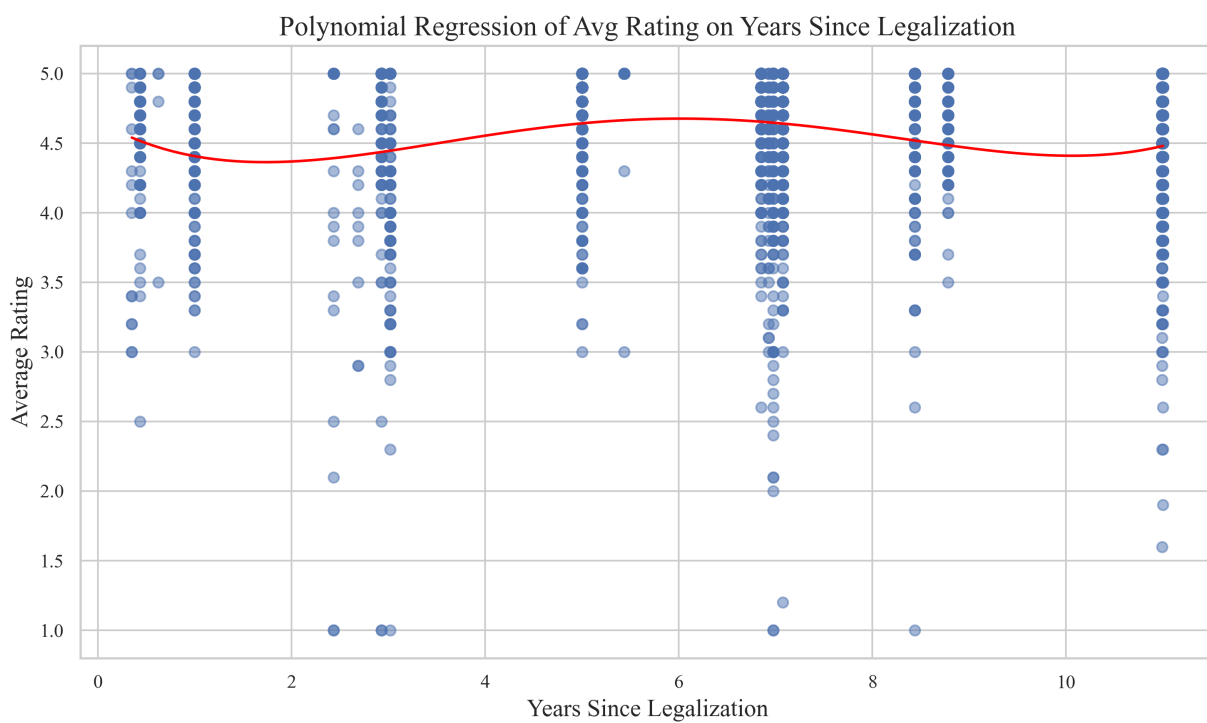


Figure 25: Polynomial Regression Plot with degree 3.

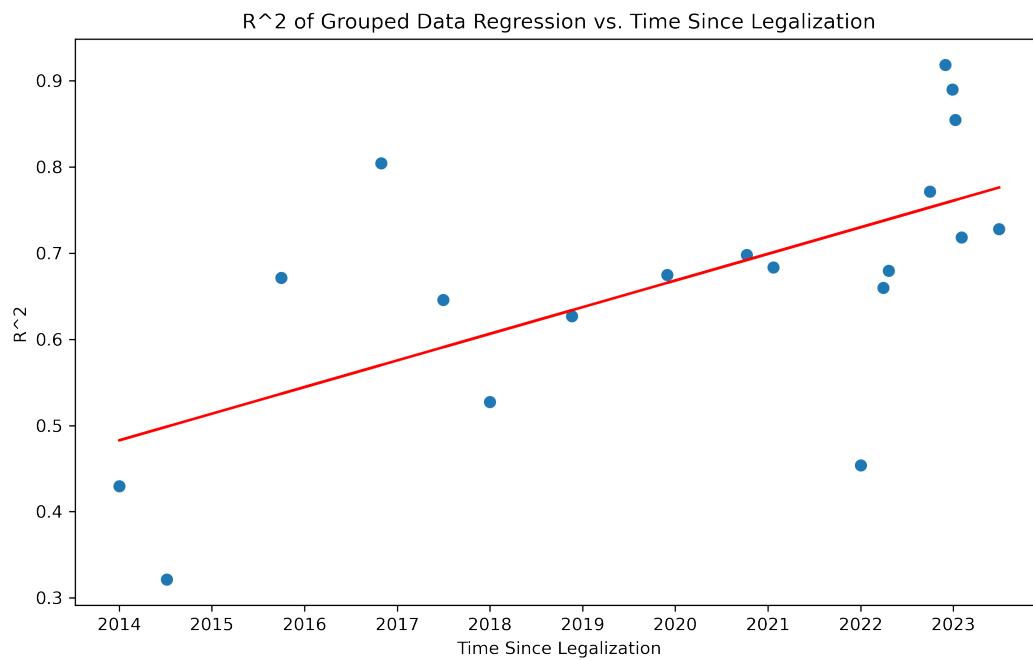


Figure 26: R^2 vs. Time (Grouped).

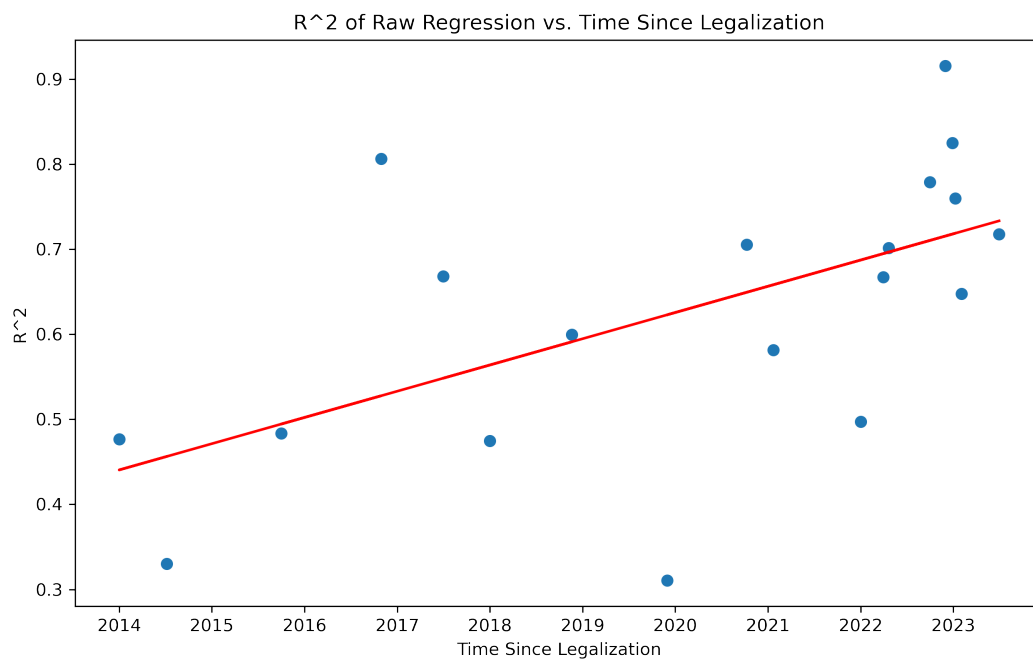


Figure 27: R^2 vs. Time (Raw).

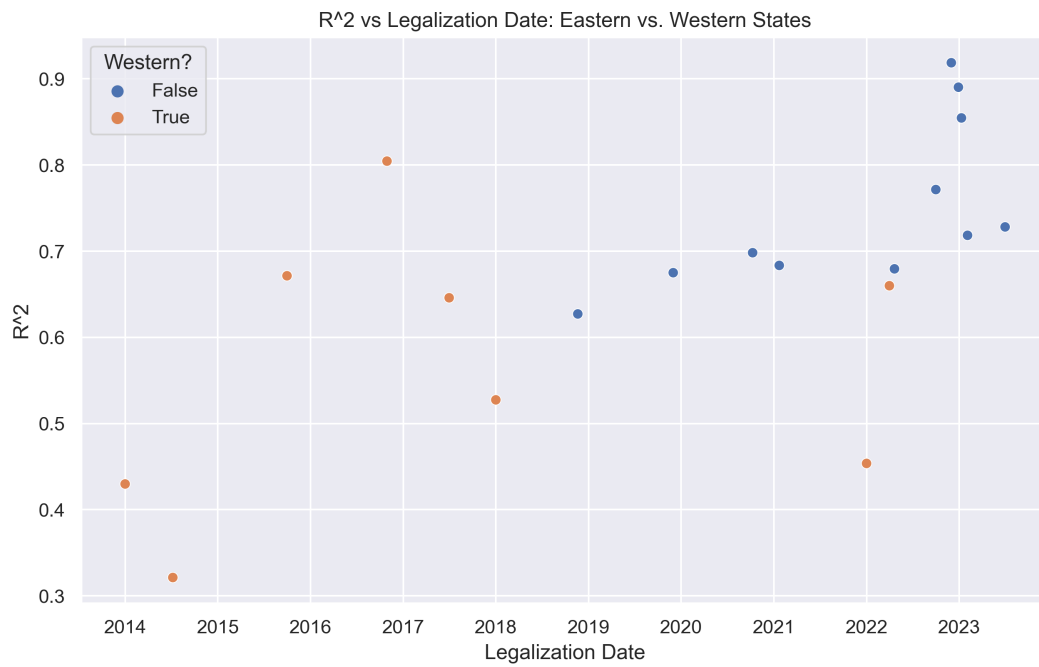


Figure 28: R^2 vs. Time: Eastern vs Western States

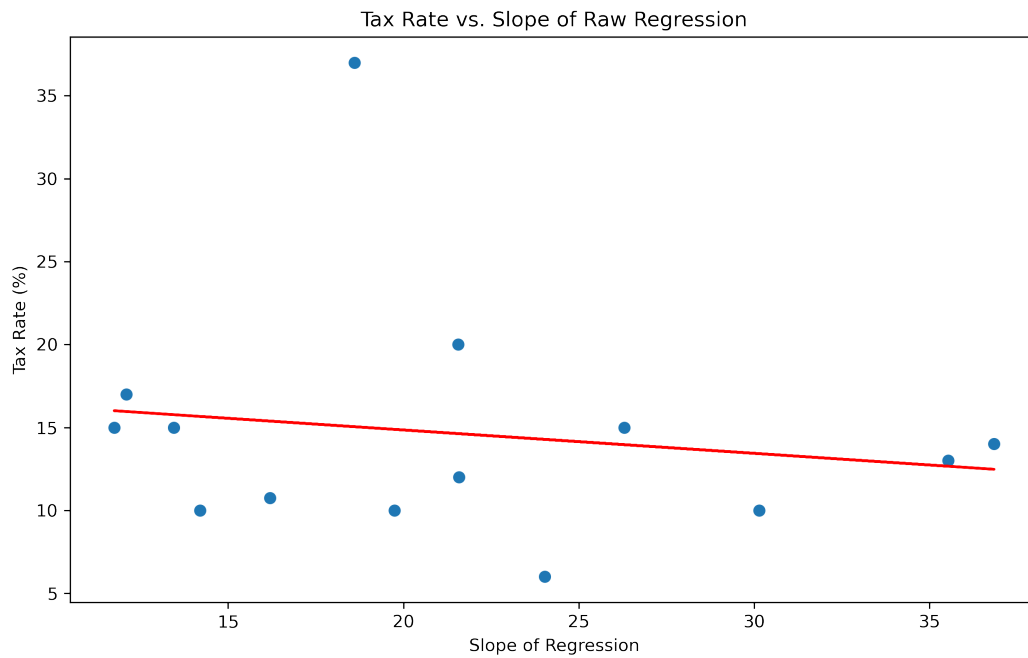


Figure 29: (Raw) Slope of Line of Best-Fit vs. Tax Rate.



Figure 30: (Grouped)Slope of Line of Best-Fit vs. Tax Rate.

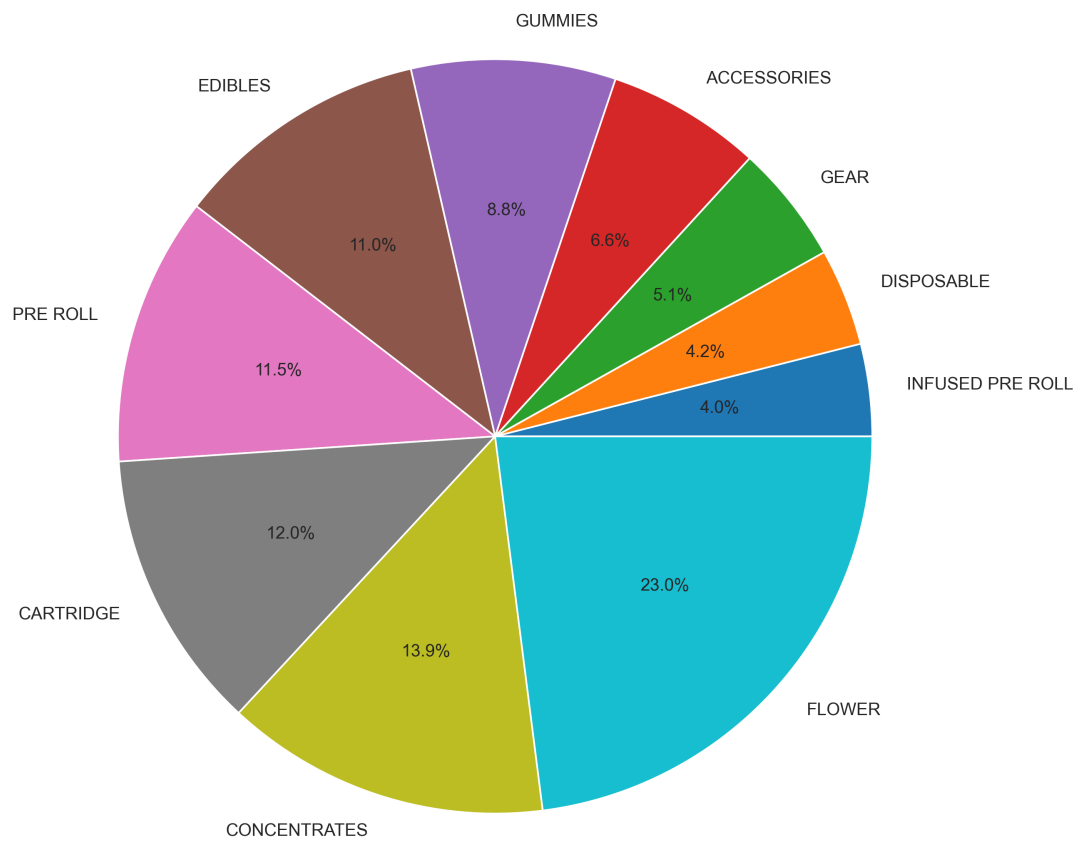


Figure 31: Distribution of Products by Category Among Top 10 Products

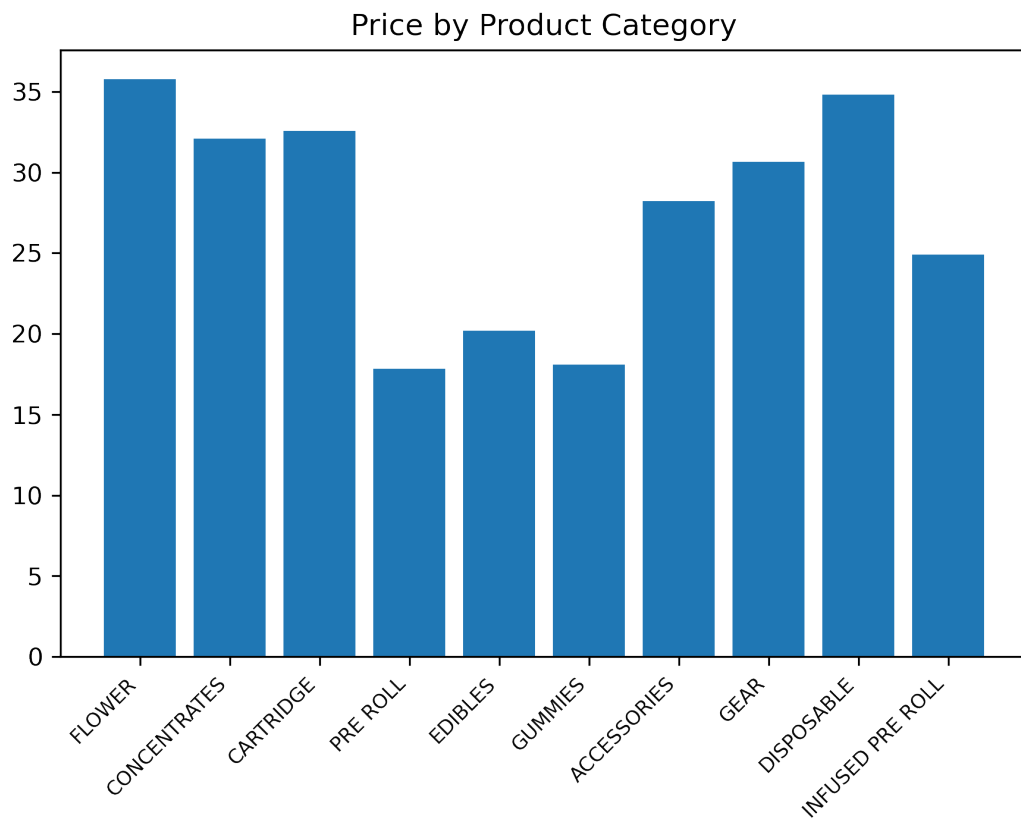


Figure 32: Avg. Price by Product Category.

state	N	intercept	slope	R^2
Washington	12525	30.229447	18.852539	0.330180
Colorado	8217	10.370818	13.425114	0.476832
Alaska	2604	7.680589	35.062591	0.806287
Oregon	6170	11.187983	6.984155	0.483558
California	32452	23.763011	10.490104	0.474754
Massachusetts	3862	21.573594	15.481906	0.599720
Nevada	5454	14.818843	26.709365	0.668021
Maine	3067	13.700186	19.526492	0.705613
Vermont	151	17.380852	37.678846	0.779172
Michigan	34615	25.543069	6.470901	0.310443
Illinois	7147	35.705490	25.235498	0.581371
Montana	1613	11.153873	22.401604	0.497228
New Jersey	1415	30.897328	31.737513	0.701674
New York	521	16.095833	32.520990	0.824994
New Mexico	5605	9.093980	22.455060	0.667412
Connecticut	330	18.164021	18.312281	0.759732
Rhode Island	122	9.108577	30.287489	0.915831
Missouri	6053	24.989377	23.173453	0.647845
Maryland	6218	24.119003	26.381168	0.717891

Figure 33: Raw Regression Results

state	N	intercept	slope	R^2
Washington	10878	30.549466	18.609198	0.321266
Colorado	6252	9.916466	13.447511	0.429820
Alaska	2469	7.863878	34.997479	0.804540
Oregon	4304	6.807935	12.094636	0.671449
California	19668	21.269142	11.749441	0.527451
Massachusetts	2713	21.423185	16.190651	0.627173
Nevada	3831	15.912089	26.294729	0.645913
Maine	2214	13.420669	19.747667	0.698230
Vermont	150	17.863045	36.844302	0.771526
Michigan	23408	14.391559	14.194577	0.675046
Illinois	3747	32.142611	26.664484	0.683487
Montana	926	12.176516	21.556611	0.453676
New Jersey	1255	30.992317	31.117811	0.679505
New York	329	14.401976	35.530202	0.890316
New Mexico	2842	9.926496	21.584497	0.659918
Connecticut	159	15.135064	19.557045	0.854657
Rhode Island	102	9.305608	30.146636	0.918720
Missouri	2868	23.511060	24.031274	0.718429
Maryland	4932	23.783017	26.927135	0.728162

Figure 34: Grouped Regression Results