Marijuana
2019 REPORT

Qing (Tiffany) Huang
Sara Shafieian
Shanshan Wang
Sia Mbatia
Srijana Lawa
Yufei (Joy) Gu

MARIJUANA-2019 REPORT

Table of Contents

MARIJUANA-2019 REPORT

Executive Summary

Dataset Description

Our dataset came from an anonymous online survey extracted from priceofweed.com, a survey platform, for economic research purpose. The data set tracks marijuana sales history of 117,931 transactions across 50 U.S. states between January 2012 and October 2013.

What interested us at first glance was the binary variable showing recreational use legalization. On November 16, 2012 Colorado (CO) and Washington (WA) became the first two states to legalize the recreational use of cannabis following the passage of Amendment 64 and Initiative 502. Our dataset captures transactions made 11 months each before and after legalization in both states. We were eager to find out whether the changes in marijuana policies would correlate with price changes in these states, but also in related states.

Business Questions

Our purpose is to understand the legalized marijuana market. The legalization in the marijuana market was expected to boost the market. Our analysis is aimed at answering the following questions.

- What factors will impact marijuana price and how?
- Does recreational legalization have impact on the market in affected and related states?

Preparation for Analysis

Adding population data: We started our exploration with adding the population of each state to the existing dataset. In that way we could have a more objective judgment of certain variables with respect to population size. Both 2012 and 2013 population data were added.

Cleaning: Our main dataset was provided by an economics researcher using STATA software. To ensure we used variables relevant to our analysis, we reviewed and got rid of most of the one-hot encoded dummies and some redundant variables, and ended up with 33 variables.

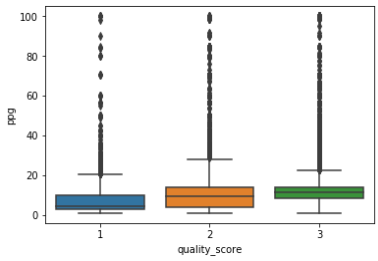
Methodology

Exploratory Data Analysis (EDA)

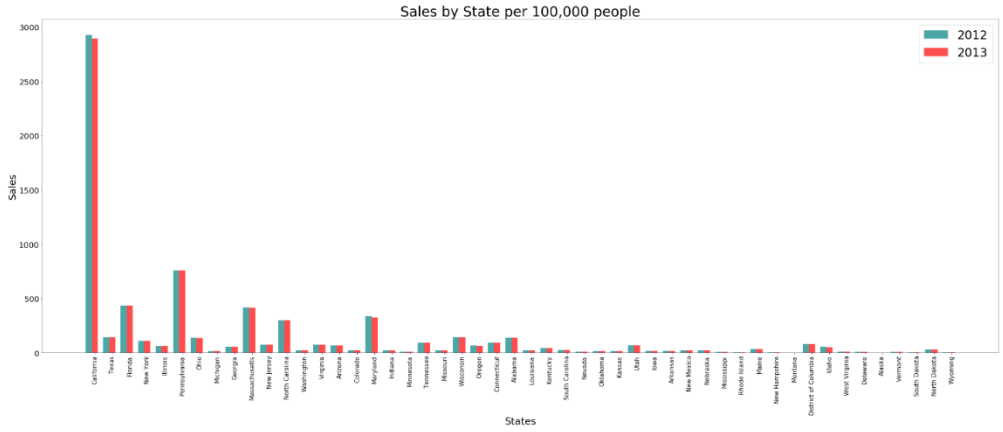
Simple analytical techniques were utilized in our EDA process such as visualization and correlation tables. Interesting findings include:

- *Price per gram* and *quantity* were significantly negatively correlated with each other.
- *Income* was highly correlated with *beer tax rate* and *distance*.
- *Price per gram* was correlated with product *quality*

	ppg	quantity	distance	beer
ppg	1.000000	-0.341761	0.102148	0.013852
quantity	-0.341761	1.000000	-0.052496	-0.002668
distance	0.102148	-0.052496	1.000000	0.052109
beer	0.013852	-0.002668	0.052109	1.000000
dayspostlegal	-0.008217	0.011010	-0.002670	0.065335
income	0.035579	0.013832	0.366855	-0.323608
quality_score_high	0.159564	0.055973	0.010424	0.034436
quality_score_medium	-0.126908	-0.056816	-0.005279	-0.034761



Then we assumed that the approximately 117,000 instances could reflect the true distribution of marijuana purchases across the US in terms of aggregated sales by state. After we adjusted the total sales by state by population, we observed that a few states had significantly higher per capita sales thus we divided all states into two groups. Since we wanted to see a more representative picture of the country, we decided to exclude the higher sales states such as California and Pennsylvania because they skewed the rest of the dataset.

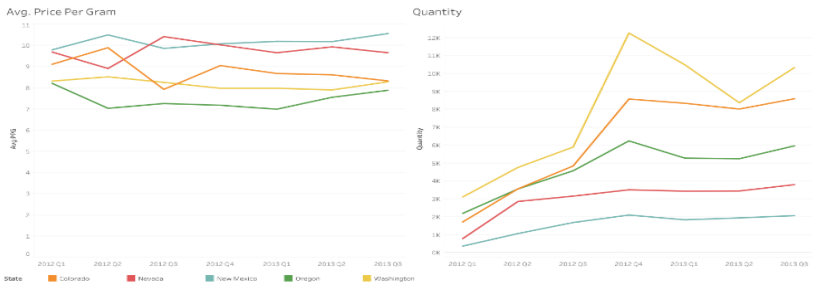


With the aim to narrow down the scope to a few representative states, we assumed that similar geographic location and political climate would relate to similar marijuana consumer attitudes. Thus, we chose three states that are both democrat liberal west coast states close to CO and WA: New Mexico (NM), Nevada (NV), and Oregon (OR).

Subsequently, we streamlined the dataset to the following 11 variables:

- Numerical variables: *days post legal*, *distance*, *income*, *ppg*, *quantity*
- Categorical variables: *state*, *legal*, *quality_score*, with *year* and *state* as dummies

Most purchases were made after the legalization of the medical use of marijuana. Recreational use was made legal on November 6, 2012 (Q4 2012). There were sharp decreases in prices and sales volumes right after that date, except for the price in Washington. The volume continued to increase after Q3, 2013, while the price showed a relatively stable trend. The quantities of products sold in the two fully legalized states showed a stronger fluctuation, signaling the impact of the legalization. The sales in these two states were much higher than those of the other three states. Q2 could be the slow season as there were decreases in sales volume for both years indicating it would be a good time for a vendor to apply seasonal promotions to boost sales.

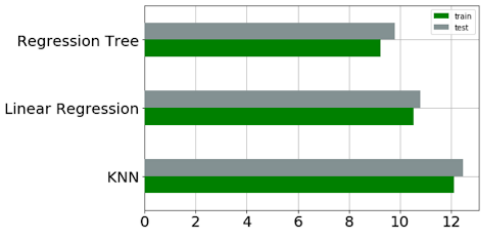


Outliers and Independent Feature Transformations

When we ran a number of regressions including the outliers of ppg and dummies of month, year and state, population and income, the adjusted was R^2 less than 0.07, MSE was high, residuals were not normally distributed, and had heteroscedasticity error. After failing to attain useful effect of our independent features, we removed *ppg* outliers and decided not to use income, population and month for the regression.

Regression Analysis

The prediction variables we choose to do our analysis include days after legalization, legal status, quantity, distance, year dummy and quality score, to analyze their impact on our target price per gram (*ppg*). The results showed the adjusted R^2 was only 0.182.



	train	test
Linear Regression	10.52	10.8
Regression Tree	9.23	9.79
KNN	12.1	12.47

We compared the results of KNN, decision tree, and linear regression models but all measures showed that our regression results were invalid.

Clustering with Multilinear Regression (MLR)

We conducted k- mean clustering analysis with quantity, distance, dayspostlegal and quality_score only. We decided not to include the dummy variable during the clustering because it had huge impact in number of clusters obtained. After comparing the Silhouette scores, we decided to separate dataset into two clusters - Cluster 0 and Cluster 1

We ran MLR with all the numerical variables to predict our target, *ppg*, for both clusters separately. However, the results were disappointing as clusters 0 and 1 had R^2 of 0.0988 and 0.0952, and mean squared errors of 40.75 and 24.39 respectively. Still there was heteroscedasticity error in both the regression model.

Artificial Neural Networks

Since we were not able to get the best model from regression and clustering, we thought of using neural network. We created a neural network with four hidden layers and 32 neuron in each layer. We used Rectified Linear unit for the activation of each hidden layer. In order to avoid overfitting we used the dropout method with drop rate 10%. The least MSE value we could obtain for train and test was 9.693 and 10.268 respectively, which shows that there may still be overfitting. So we thought of moving into discretization method.

Discretization of Independent Features

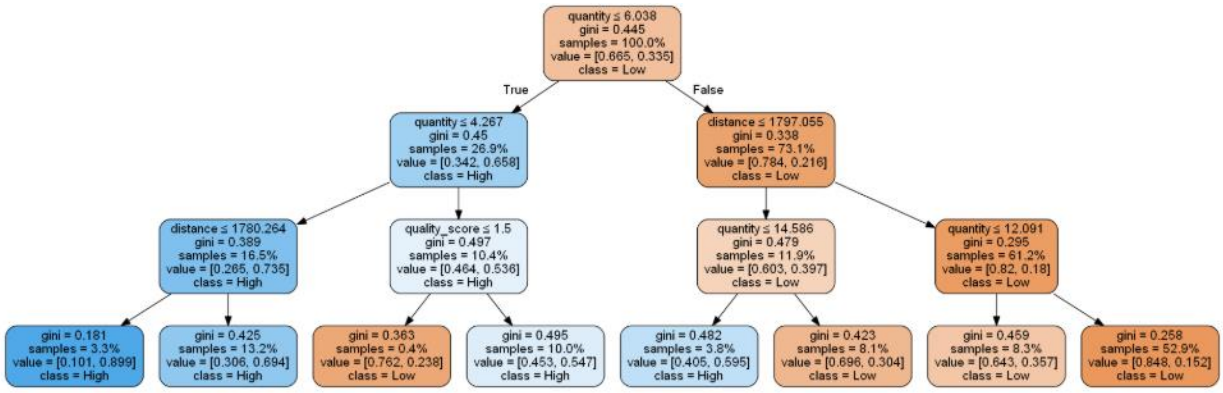
With yet another insignificant result, we tried to discretize the continuous variables *distance*, *quantity*, and *dayspostlegal*. Along with the *quality*, and *state* and *legal* dummies, a regression model yielded heteroskedastic residuals once again, and as before, our adjusted R^2 increased to 0.163 with a MSE of 10.81 – a better result but still not valid enough to explain *ppg* variations.

Segmenting Quantities

As *quantity* is repeatedly shown to have a strong influence on *ppg*, a suggested alternative was to review how *ppg* changes are affected by the other variables for each quantity segment as there are 10 distinct quantities reported for purchase. Of the 10 segments, the two below had the most promising variation outcomes, though still too low to make any inferences on.

Classification: Binning Price per Gram

We had not gotten any useful information from trying to predict *ppg*, so we switched our approach by transforming it into a category and predicting propensity for different outcomes. Our *ppg* values were classified into high and low categories, split the bins at the $ 14.61 mark. With a classification model, we found the most meaningful model for interpretation from our model measures, including an accuracy score of 0.7598.



Quantity was again the most significant variable for explaining *ppg* while the legal status and states didn't have any importance (0.00). We decided to use this model to make final inferences about how ppg changes are affected by other features.

Feature	Importance
quantity	0.881981
distance	0.111159
quality_score	0.006860

Measure	***High ppg***	***Low ppg***
Precision	0.65583	0.80507
Recall	0.59422	0.84312
F-Measure	0.62351	0.82366

Conclusion

From our objectives, we were able to answer our question in the following ways:

Factors Impacting Marijuana Price

The ***quantity*** proved to be the most critical factor in predicting price. In all our models, even the inconclusive ones, we found that quantity had high predictive power. We also come to this conclusion looking at the trend lines in the quarters in 2012 and 2013 leading up to and right after recreational legalization and in our categorical analysis in the classification model.

We believe this outcome to be plausible because people's demand of a good can influence the suppliers' actions in how they price a good. We did notice that consumption increased even though prices didn't fluctuate much post-legalization. This could be attributed to the increased social acceptance of recreational marijuana consumption which leads the consumers to increase their purchase quantities. *Distance* from the Sinaloa Cartel as well as quality of the product seemed to have an impact on predicting *ppg*, but none came close to the impact of *quantity*.

Impact of Recreational Legalization in Affected and Related States

We theorized that legalization would stabilize prices and may indirectly influence the demand and therefore quantity purchased by consumers. We find from our model of choice that legalization does not seem to have any direct impact on the price per gram in the affected states and states of same geographical and political characteristics. This could be because the type of consumer has stayed the same from before and after legalization, and they therefore already had access points with stable prices before the laws went into effect.

Final Thoughts

Since 2013, recreational legalization has been instated in 9 more states, signaling growth in the marijuana market which may present growth opportunities for entrepreneurs looking to enter or maintain their standing in the market. A next interesting analysis would be focused on ways to advise dealers on how to maximize profits. For example, in our analyses, Q2 seemed like a good opportunity to have deals because of slow sales. More business centered review of consumption would benefit dealers with the increasing acceptance of recreational marijuana use in the U.S.