Forecasting Future Cannabis Sales

A Modeling Project Based on Brand-Level Time Series Data

Ryan Holland, Chris Baker
Amateur Cannabis Consultants

Executive Summary

The cannabis industry has undergone immense growth in recent years, spurred on by ever-evolving shifts in public opinion and continued changes to legal regulation. As more states domestically and more nations abroad legalize recreational marijuana, cannabis retailers must continue exploring the possibility of new products and emerging markets. In this paper, as consultants for cannabis brand Cookies, we develop data science models the business can utilize to forecast commercial sales. We carefully merged disparate data sets detailing the last three years of cannabis sales for a variety of brands, extracted time-series features from the sales data, and aggregated product-level details into brand-level features to predict with. Then, we trained a variety of regression models, including Linear Regressors, K-Nearest Neighbor Regressors, Decision Tree Regressors, and Ensemble Bagged Regressors. We also attempted to optimize model inputs with dimensionality reduction and model parameters with grid search. In the end, our optimal Linear Regression model outperformed other models in terms of root mean square error and goodness-of-fit, accounting for 93% of the variation in monthly sales. Cookies will be able to leverage this model to continually anticipate sales success.

The most predictive feature in our model was data on the previous few months of sales, indicating that sales do not vary wildly from month to month. However, a few product-related features also had substantial and significant impacts on sales predictions. The greater variety of products and strains a brand offers, the more sales generated. Intuitively, users prefer having choices and value brands that cater to many different categories and product types. That said, options for flower, vapes, and edibles were more indicative of sales success than beverages, tinctures, and topicals. Additionally, brands that sold more hybrid products generated more sales, suggesting a potential consumer preference. Lastly, the longer a brand has been sold on the market, the greater the sales. Customer recognition and loyalty is essential, and luckily Cookies already succeeds in this regard.

1. Background

1.1. Industry Overview

Attributed to recent changes in legal regulation and continued shifts in public opinion, cannabis has become one of the fastest growing industries across many states. Nationwide sales totaled $21.3 billion in fiscal year 2020 and will hit $27.3 billion by the end of 2021. These figures are projected to continue growing, expected to reach $55.9 billion in 2026, following a compound annual growth rate of 14%. Additionally, these numbers only include legal brick-and-mortar dispensary sales, without factoring in other sales avenues like e-commerce, grocery stores, drug stores, and many other new channels that will emerge in the coming years. For example, CBD is expected to be approved as a food additive within the calendar year, and if so, annual retail sales of CBD products will exceed $17 billion by 2026. This immense growth in sales coincides with an evolved public opinion. 73% of US adults either consume cannabis or are open to it, up 10% over the last three years. Clearly, cannabis use is on the rise and as a result, retailers face a new set of challenges as they try to meet the needs of an ever-evolving and rapidly growing consumer base.

The ability to project sales would help retailers understand the brands, products, strains, flavors, and other factors that most influence consumer purchasing habits, empowering them to improve their bottom line and better serve the customers. However, the cannabis market is far more difficult to forecast than other industries for many reasons. First, cannabis has only been legalized recently, so there is not a massive wealth of past industry data on which to observe trends and base projections. Furthermore, each state is an isolated market with unique regulatory demands, legal roadblocks, and supply chain challenges. Therefore, sales can differ vastly from state to state.

Currently, 26 states only allow medical marijuana use, and 5 states have only legalized CBD sales. 13 states have fully legalized cannabis, with 8 having just approved recreational sales to begin shortly. These new markets are incredible growth areas for the industry. Illinois and Massachusetts, both states that just legalized, have already generated $825 million and $724 million in the first half of 2021 alone. New York is projected to bring in $2.9 billion in total legal sales by 2026. In addition to emerging domestic markets, international sales also present an interesting growth opportunity. Sales abroad totaled just $3.6 billion in 2020, but numbers are on the rise. Canadian markets saw a 54% increase and are expected to hit $6.7 billion in 2026. Many Canadian province officials strongly support cannabis and have campaigned to expand the number of dispensaries in their districts. Many markets are clearly welcoming massive growth.

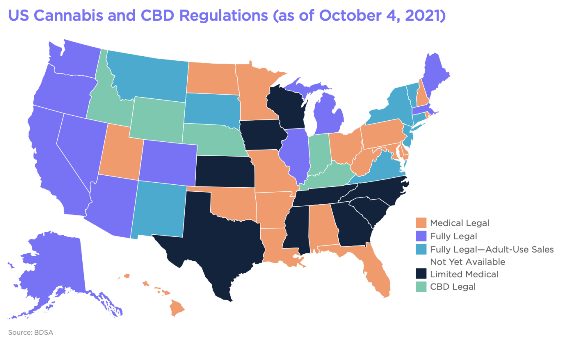

US Cannabis and CBD Regulations (as of October 4, 2021)
Medical Legal
Fully Legal
Fully Legal—Adult-Use Sales Not Yet Available
Limited Medical
CBD Legal
Source: BDSA

Fig 1. US states by legalization status.

Existing markets are also experiencing massive gains. California, Nevada, Oregon, and Colorado, the four most mature marijuana markets in the country, all saw double-digit growth in the last fiscal year. California's recreational market grew 24%. In Colorado, Washington, and Nevada, 50% of adults use cannabis, and this continues to rise.

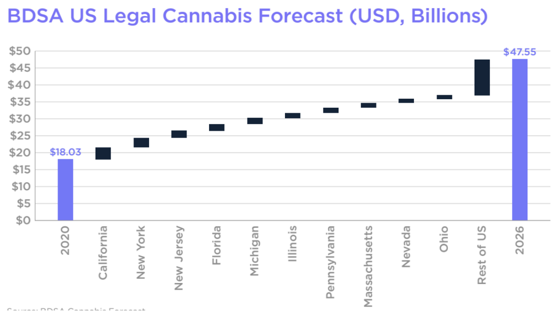


Fig 2. 2026 cannabis sales forecast by state.

1.2. Cookies

Cookies is one of the most established and exciting marijuana brands in the industry, selling cannabis products and a fashionable streetwear line. Their estimated annual revenue exceeds $86.1 million, driven by sales of flower, wax, vapes, and edibles, although they are best known for their signature girl scout cookies flower strain. Cookies currently operates 38 retail stores across seven states, and they are looking to expand into new markets. This includes opportunities abroad, with partnerships announced to launch in Israel, Spain, and Canada.

Cookies' success is attributed in part to masterful branding, driven by owner and co-founder Berner. Berner's Instagram has 482,000 followers and his YouTube has millions of views. Their name is key. As Berner explains, "Cookies is friendly, easy to digest. It's simple. Everybody loves cookies."

Cookies also builds strategic partnerships, such as Collins Ave, Minntz, and Grandiflora. Most notably, four Cookies retail stores are Lemonade branded locations, offering a popular sativa-centric menu. As a result of their quality product and excellent marketing, Cookies won this year's (2021) AdAge America's Hottest Brands award. As an established name with a massive customer base and imminent opportunities for growth and expansion, Cookies will benefit greatly from sales forecasting models.

2. Methodology

2.1. Provided Data

Cookies kindly provided five datasets on several aspects of the California cannabis industry from 2018-2021, including sales, units sold, and retail price for a variety of brands. The majority of data, including the sales figures to predict, were given on a brand-level. Although one dataset contained product-level sales, we ignored this because it only covered a small percentage of existing products and brands in the other datasets. Eliminating or imputing a large portion of the data would risk introducing more problems than the alternative of aggregating product-level data into brand features.

Therefore, our coding process began with data set construction, merging the provided tables into a single data frame that captured brand-level data with monthly sales as the target. We chose to predict total sales instead of total units sold since this is more useful from a business perspective. In the end, cannabis retailers want to maximize profits, and maximizing units sold may not have this effect, depending on prices and unit types.

In terms of merging the data, aggregating product features into brand features does have drawbacks. Each entry in the data frame represents a specific brand's sales for a specific month. Consequently, each brand has a different number of rows in the data frame depending on how long the brand has been on the market and how much sales data the brand provided. For each row, the time series data associated with that row will be unique, as these numbers differ from month to month. However, the brand features related to product offerings will have the same values for every row of that brand. In reality, some of the features (i.e. the number of products a brand sells), were likely not consistent for every month. Brands may have released new products or stopped offering certain products, and accurately capturing those changes in the data would certainly render our dataset more reliable. However, the data sets only provide product level data for 2021, so we assume the menu of products a brand offered this year has been the same menu of products offered each month since 2018.

2.2. Pre Data-Wrangling Research

First, our code loads the datasets, performs a few data type conversions and feature arithmetic, and then immediately handles outliers in the target feature. Observing a histogram of the Sales ($) feature, the data is heavily skewed with a number of outliers on the upper end. A few brands had months with uncharacteristically high profit. We addressed this by replacing outlier values with the largest Sales value that would not be considered an outlier. Therefore, these data points maintain their status as profitable months, without overly influencing the modeling. This helped ensure that any models developed will be accurate in fitting the more common occurrences in the data. While removing outliers may worsen performance when predicting outlier months, this step helps produce more general models that can be better applied to any average cannabis brand with ordinary sales.

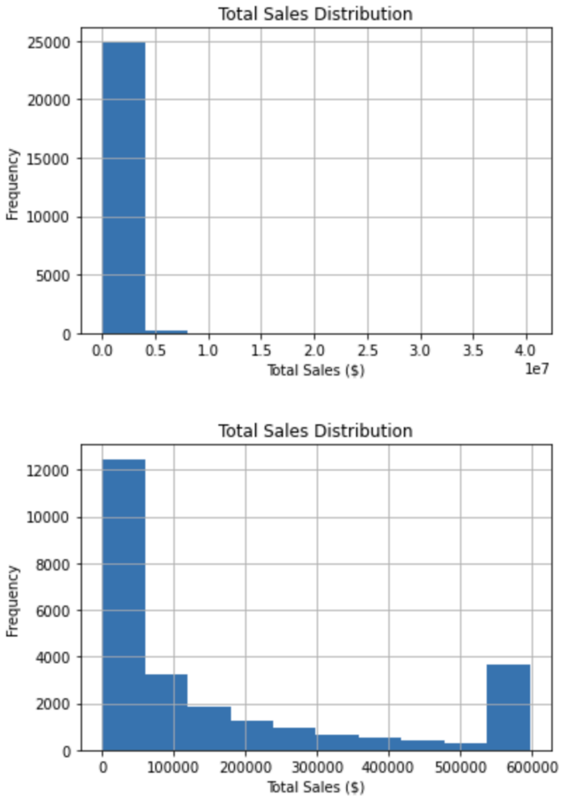


Fig 3. Distribution of monthly sales before (top) and after (bottom) imputing outlier values.

At this stage, the Sales target feature was linearly interpolated, replacing any missing values with the assumption that sales will progress linearly over time. This introduces some inaccuracy into our data set, but the assumption of steady growth is fairly reasonable and we preferred interpolation over unnecessarily experiencing rampid data loss. After replacement, the outlier instances are still on the very upper end of the spectrum of sales data.

Next, we iterated over each brand in the data set, extracting the rows/months listed for that brand, engineering new time series features, aggregating product-level features into brand-level features, and recombining each brand's data into a single unified data frame. Before collecting these brand attributes, we researched which aspects of brands would be most predictive of commercial sales success. For example, a BDSA 2021 national survey of cannabis users concluded that average consumers make their purchasing decisions based 31% on taste and flavor, 24% on THC content, and 22% on brand familiarity. While brand familiarity is difficult to quantify due to the qualitative nature of a sentiment that differs from person to person, a rough proxy is length of time on the market. The longer a brand has existed, the more likely it is for a user to know and trust the name. THC content and flavor options are easier features to engineer.

Beyond the BDSA data provided by Cookies, we also read a Deloitte consulting report on emerging trends of cannabis consumption in Canada. While this covers an entirely different country, Cookies is opening stores in Canada soon. Furthermore, the preferences of Canadian cannabis users might be similar to the preferences of Americans. Deloitte claims that 42% of the market prefers some form of smoking for marijuana consumption, including joints, flower, and vaping. Furthermore, they claim the primary motivation for cannabis consumption is stress relief, relaxation, and sleep. This could potentially indicate CBD content as an important factor, as CBD generally helps with these issues. We also analyzed a recent survey administered to 817 adult US cannabis users, conducted as a joint research venture between UCSD, UCLA, and other accredited universities. The survey determined the most important attributes affecting a person's willingness to buy cannabis products were quality, strain, price, THC content, and use of pesticides. For medical users, CBD content was a major factor. Aside from whether a brand uses pesticides, the other features are derivable from provided data.

Additionally, we were interested in the influence of edible products on sales. Industry trends suggest demand for these products will skyrocket. Sales grew 24% this fiscal year and should outpace the rest of the market through 2026. Notably, edibles offer the benefit of controlled dosage and familiar consumption methods, which will be appealing to first time users as new markets emerge. Beverage sales in particular grew 60% from 2020 to 2021, suggesting these products can drive new growth

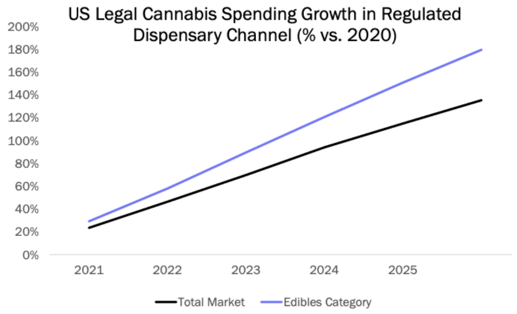


Fig 4. Edible sales forecasted over the total market.

Both edibles and beverages were included as two of the eight categories on which we split products, along with flower, prerolls, concentrates, vapes, tinctures, and topicals. The originally provided labels (i.e. inhalables and consumables) are too broad. Most consumers have specific preferences between, for example, prerolls and vapes (both are inhalables) or tinctures and gummies (both are consumables), so we differentiated products on a lower level. BDSA claims massive differences in market share between these product types. As of 2021, flower accounts for the largest portion of total sales ($760M), concentrates and vapes are next ($588M), followed by edibles/beverages ($286M), prerolls ($243M), tinctures ($38M), and topicals ($17M). These figures are helpful, as data about more popular categories may be predictive.

2.3. Feature Engineering

Below is every feature we derived or constructed, along with descriptions about exactly what each feature represents and why we theorized each one may be predictive of sales success. We also include thoughts on our data imputation strategy for each variable, if applicable.

Sales ($), Previous Month: Time series feature containing the total sales the month prior. Sales will likely not change substantially in adjacent months, so past sales data may be very predictive. We dropped rows that were missing this feature.

Sales ($), Past 3 Month Average: Time series feature with rolling average of sales in the prior three months. Similar to above, past sales data should be correlated to future sales. For months without three prior months of data, the average was taken over the last one or two months.

Change in Sales ($), Previous Month: Time series feature describing how much the sales increased or decreased between the prior two months. Past trends may be correlated to the future changes. For rows without two previous months of sales, we assume sales two months ago were 0, and the change is equivalent to the previous month's sales.

Units Sold, Previous Month ; Units Sold, Past 3 Month Average ; Units Sold, Previous Month: All similar to above, but for units sold instead of sales.

Time on Market: Time series feature estimating how long the brand has been available, calculated as the number of days from the first month listed for the brand to the most recent month. This is not an exact estimate, as some months may have been on the market for a long time before 2018.

Total Product Count: This counts the total number of unique products a brand is offering. This could easily be correlated with sales; perhaps the more products a brand offers, the more sales they have.

Average Product Price: The average retail price of products under this brand. The higher the price, the more money made on each purchase but also cheaper products may sell more often, so the potential correlation here is unclear.

Flower Product Count, Flower Product Share, and Average Flower Product Price: These features describe the number of flower products, the percentage of flower products compared to the total number of products, and the average price of flower products. Consumers may have preferences for different types of products. These features are also calculated for Prerolls, Edibles, Beverages, Vapes, Concentrates, Tinctures, and Topicals.

Sativa Product Count, Sativa Product Share: These features describe the number of sativa products the brand sells and what percentage of the strain products they sell are sativa. Users might be more likely to buy from brands that are sativa-focused. These features are repeated for indica and hybrid.

Total Strains: This counts the total number of unique strains (by name) that a brand sells. We also compute similar features specifically for the number of flower strains, preroll strains, vape strains, and concentrate strains.

Edibles, Avg THC/Item: The provided data lists total THC for edible products only. This feature cross calculates THC/item by dividing total THC by the number of items in a pack. THC potency might be an important factor for consumers. This could be positively or negatively correlated; it is unclear whether users prefer higher or lower THC content.

Vapes, Average Volume, Prerolls, Average Volume: For vapes and prerolls, the data lists volume rather than THC content. The volume of cannabis in these products could have an effect on sales.

Sells CBD Products: This feature is a boolean for whether the brand sells any CBD products. CBD is gaining traction as an in demand product, most notably for medical marijuana users.

Sells Devices: This feature is a boolean for whether the brand sells any devices in addition to actual cannabis. Brands that sell devices may gain sales by offering devices when others do not, or they may increase awareness by branding their devices.

Sells Flavored Products, Available Flavors: As mentioned, many marijuana users prefer flavored products. One feature is a boolean for whether the brand sells flavored products. The other feature counts how many unique flavors the brand offers.

Note that all of the non-time series features are guaranteed to be non-null. Either the value can be calculated exactly, or they could be 0, and our code ensures this to be the case. To handle missing time series data, we dropped rows that lacked previous month data to predict with, as this data will likely be the most predictive. We also dropped any rows where missing sales target values could not be linearly interpolated. Of note, all the time series construction, imputation, interpolation was done before the train/test split. After splits, there may not be enough data for some brands to properly calculate new or missing values, so all the feature engineering was done prior. For the test set, we assumed companies would have access to all prior sales for use in predictions. However, we dropped features that would be problematic to include in models, such as Units Sold, which describes sales for the current month and would not be known at the time of forecasting. Finally, after this feature engineering, we standardized the data frame with a StandardScaler. Every feature in our data set is numeric and subject to this normalization.

2.4. Modeling Methods

After data pre-processing, we ran methods to determine basic statistics on each feature and compute correlations between feature pairs. This gave insight into underlying data relationships.

For modeling, we began by executing a train/test split with 80% training data and 20% testing. We wrote auxiliary helper functions for evaluating a model, optimizing a model, cross-validating a model, and creating a residual plot, in order to reduce redundant code and render the entire codebase more legible. Then, we started with Linear Regression. After this initial attempt, we attempted optimizing performance through PCA and manual dimensionality reduction. For the manual option, we selected which features to remove based on observations in the correlation matrix, attempting to remove strong collinearity. After this, we tried using a K-Nearest Neighbors model, a Decision Tree model, and an ensemble Bagged Regression model. For each model type, including Linear Regression, we first tried with the default parameters and then tried optimizing the parameters with GridSearch. We also tried using a perceptron model, but it never converged.

Finally, we compared the RMSE, MAE, and R^2 for each model. Our preferred comparison metric was RMSE. We cross-validated the best linear model, the best ensemble model, and the best of the other models we tried developing. For these three cross validated models, we also created residual plots of the predicted sales vs the actual sales to visualize performance. The metrics and graphs are below.

3. Results

3.1. Pre-Modeling Correlation

Refer to Appendix 7.1. for a table summarizing the results of the correlation matrix for target value Sales ($). Regarding some of the key correlation observations, previous month's sales and past three months average sales are both extremely correlated with current sales, with correlation coefficients exceeding 0.94. These will likely be the most predictive features in any model. Past data on units sold was also correlated, with coefficients a bit lower around 0.87. Also positively correlated to sales with substantial coefficients were total product count, strain count, and hybrid product count, with lower correlations for variables like time in the market and whether the brand sells CBD. In terms of product type, it appears that statistics on flower and vape offerings are most indicative of sales success, perhaps suggesting that traditional flower and vapes are more substantial drivers of cannabis sales than other methods, or simply that these products sell for more or sell more often. For example, flower product count and vape product count were more correlated than the corresponding features for other product types. Some features, such as flavors and THC content, were surprisingly fairly uncorrelated with sales.

3.2. Linear Regression

To begin modeling, we fit a default Linear Regression model with no special parameters. This provided a baseline metric for comparing future performance against this simple model.

Root Mean Squared Error:	58836.82
Mean Absolute Error:	32714.91
R^2:	0.93

Next, we attempted to improve performance via dimensionality reduction. First, we ran principal component analysis, cutting the number of input features from 51 to 10. The accuracy scores were much worse, suggesting that the default linear model was not overfit due to high dimensionality.

Root Mean Squared Error:	77646.49
Mean Absolute Error:	56198.51
R^2:	0.87

As an alternative to PCA, we also tried executing manual dimensionality reduction, computing the most correlated feature pairs and dropping the feature less correlated with sales. For example, we dropped product share variables because product counts were correlated and more predictive of total sales. We also dropped product type strain counts, as total strain count encapsulated all this data in a more predictive variable. We tried many permutations of features to remove, but none of our attempts led to improvement, corroborating the theory that high dimensionality did not trigger any overfitting here. Therefore, moving forward, for all future models, the entire data set was used.

Root Mean Squared Error:	58915.05
Mean Absolute Error:	32806.79
R^2:	0.93

Finally, we used GridSearchCV to optimize Linear Regression. The scoring metric to minimize was RMSE. The parameters we tried were alternating values for fit_intercept and normalization. The optimal model was actually just the default model.

Root Mean Squared Error:	58836.82
Mean Absolute Error:	32714.91
R^2:	0.93

3.4. KNN Regression

As an alternative to linear regression, we also tried using K-Nearest Neighbors regression, once again starting with the default model and parameters.

Root Mean Squared Error:	66116.31
Mean Absolute Error:	35358.03
R^2:	0.91

We then attempted to optimize the KNN regressor with n_neighbors of 3, 5, 10, and 20 and both manhattan and euclidean distance. The optimal model used manhattan distance and 5 neighbors. This optimal model slightly reduced the error.

Root Mean Squared Error:	65255.66
Mean Absolute Error:	34219.96
R^2:	0.91

3.5. Decision Tree Regression

Next, we trained a default Decision Tree Regressor, using MSE as the splitting criterion parameter.

Root Mean Squared Error:	81005.42
Mean Absolute Error:	43066.85
R^2:	0.86

We then attempted to optimize the regressor, still using MSE as the criterion but iterating through splitting on the best split or a random split, max depths of 2, 3, 5, 10, and 15, and feature selection of auto, sqrt, or log2. The optimal model split on the best feature, considered all features for splits, and had a max depth of 5. The optimized model was a huge improvement over the default model.

Root Mean Squared Error:	59900.76
Mean Absolute Error:	32946.34
R^2:	0.92

3.3. Ensemble Regression

Lastly, we tried training an ensemble bagged regressor, first using the default parameters.

Root Mean Squared Error:	61267.6
Mean Absolute Error:	33154.99
R^2:	0.92

Finally, we optimized the model by trying both Decision Tree and KNN regressors as the base regressors, and 5, 10 and 20 estimators for each. The optimal model bagged 20 Decision Tree Regressors, slightly improving performance.

Root Mean Squared Error:	60718.28
Mean Absolute Error:	32379.11
R^2:	0.92

3.6. Comparison of Models

Below is a table summarizing the accuracy metrics of each model, sorted by best to worst RMSE. The optimal Linear, Decision Tree, and Bagged models performed best in terms of RMSE, MAE, and R^2.

Model	RMSE	MAE	R^2
Optimized Linear	58836.82	32714.91	0.93
Default Linear	58836.82	32714.91	0.93
Linear w/ Dim. Reduction	58915.05	32806.79	0.93
Optimized Dec. Tree	59900.76	32946.34	0.92
Optimized Bagged	60718.28	32379.11	0.92
Default Bagged	61267.60	33154.99	0.92
Optimized KNN	65255.66	34219.96	0.91
Default KNN	66116.31	35358.03	0.91
Linear w/ PCA	77646.49	56198.51	0.87
Default Dec. Tree	81005.42	43066.85	0.86

3.7. Cross-Validation

This process narrowed down the models based on performance working with an arbitrary train/test split. Next, we performed 10-fold cross validation on the three models: optimized Linear Regression, Decision Tree Regression, and Bagged Regression, to acquire summary metrics. We also built residual plots rendering predicted sales vs actual sales. The line above the data, predicted = actual, represents an ideal model that makes perfect predictions and has no error. The closer the data points are to this line on average, the better the model performs.

3.7.1. Linear Cross-Validation

Root Mean Squared Error:	59090.54
Mean Absolute Error:	33354.58
R^2:	0.93

The optimized Linear Regression model produces an RMSE score of 59090.54, which is about 27% of the standard deviation in sales. MAE is 15% of the standard deviation in sales. The model also has an R^2 of 0.93, suggesting that 93% of the variance in monthly sales can be collectively explained by the input features given this Linear Regression model.

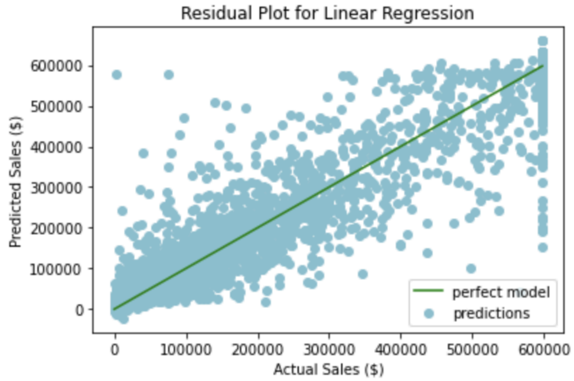


Fig 5. Linear Regression predicted vs. actual sales.

3.7.2. Decision Tree Cross-Validation

Root Mean Squared Error:	60162.61
Mean Absolute Error:	33416.07
R^2:	0.92

The Decision Tree model produces 2% more error than the Linear Regression model, and an R^2 that explains 1% less of the variance. The residual plot is also different than the Linear Regression graph, predicting the sales data in tiers rather than truly continuous, which may account for some of the error. This model is not substantially worse than Linear Regression, but the linear model is slightly more accurate, more simple, and more intuitive.

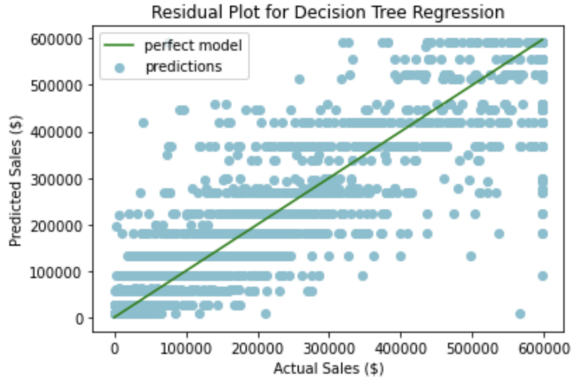


Fig 6. Decision Tree predicted vs. actual sales.

3.7.3. Ensemble Regression Cross-Validation

Root Mean Squared Error:	60790.6
Mean Absolute Error:	32969.07
R^2:	0.92

The optimal ensemble regressor produced RMSE higher than the Linear or Decision Tree models, but with a slightly lower MAE. The R^2 metric is identical to that of Decision Tree, but still 1% less predictive than our Linear Regression model.

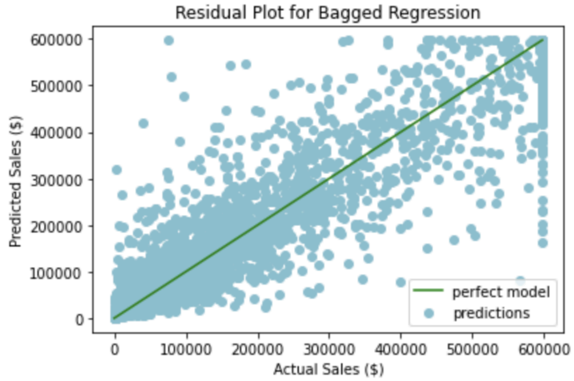
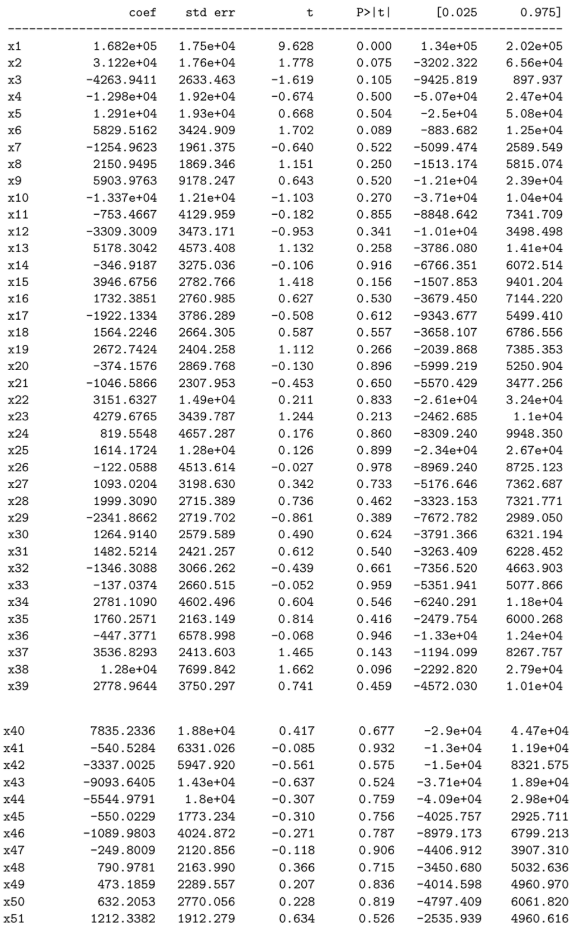


Fig 7. Bagged Regression predicted vs. actual sales.

3.8. Best Model Selection

Observing the cross-validation results, the optimal Linear Regression model clearly performed best, with the lowest RMSE and highest R^2 score. We would suggest this model as the forecasting tool for Cookies to project future commercial sales.

Looking deeper into the model itself, the weights of each feature, along with p-values and t-values, are listed below, along with discussion as to what each metric represents and which features are the most indicative and most confident predictors in the model.



```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          1.682e+05   1.75e+04      9.628      0.000    1.34e+05    2.02e+05
x2          3.122e+04   1.76e+04      1.778      0.075   -3202.322    6.56e+04
x3         -4263.9411   2633.463     -1.619      0.105   -9425.819     897.937
x4         -1.298e+04   1.92e+04     -0.674      0.500   -5.07e+04    2.47e+04
x5          1.291e+04   1.93e+04      0.668      0.504    -2.5e+04    5.08e+04
x6          5829.5162   3424.909      1.702      0.089    -883.682    1.25e+04
x7         -1254.9623   1961.375     -0.640      0.522   -5099.474    2589.549
x8          2150.9495   1869.346      1.151      0.250   -1513.174    5815.074
x9          5903.9763   9178.247      0.643      0.520   -1.21e+04    2.39e+04
x10        -1.337e+04   1.21e+04     -1.103      0.270   -3.71e+04    1.04e+04
x11         -753.4667   4129.959     -0.182      0.855   -8848.642    7341.709
x12        -3309.3009   3473.171     -0.953      0.341   -1.01e+04    3498.498
x13         5178.3042   4573.408      1.132      0.258   -3786.080    1.41e+04
x14         -346.9187   3275.036     -0.106      0.916   -6766.351    6072.514
x15         3946.6756   2782.766      1.418      0.156   -1507.853    9401.204
x16         1732.3851   2760.985      0.627      0.530   -3679.450    7144.220
x17        -1922.1334   3786.289     -0.508      0.612   -9343.677    5499.410
x18         1564.2246   2664.305      0.587      0.557   -3658.107    6786.556
x19         2672.7424   2404.258      1.112      0.266   -2039.868    7385.353
x20         -374.1576   2869.768     -0.130      0.896   -5999.219    5250.904
x21        -1046.5866   2307.953     -0.453      0.650   -5570.429    3477.256
x22         3151.6327   1.49e+04      0.211      0.833   -2.61e+04    3.24e+04
x23         4279.6765   3439.787      1.244      0.213   -2462.685     1.1e+04
x24          819.5548   4657.287      0.176      0.860   -8309.240    9948.350
x25         1614.1724   1.28e+04      0.126      0.899   -2.34e+04    2.67e+04
x26         -122.0588   4513.614     -0.027      0.978   -8969.240    8725.123
x27         1093.0204   3198.630      0.342      0.733   -5176.646    7362.687
x28         1999.3090   2715.389      0.736      0.462   -3323.153    7321.771
x29        -2341.8662   2719.702     -0.861      0.389   -7672.782    2989.050
x30         1264.9140   2579.589      0.490      0.624   -3791.366    6321.194
x31         1482.5214   2421.257      0.612      0.540   -3263.409    6228.452
x32        -1346.3088   3066.262     -0.439      0.661   -7356.520    4663.903
x33         -137.0374   2660.515     -0.052      0.959   -5351.941    5077.866
x34         2781.1090   4602.496      0.604      0.546   -6240.291    1.18e+04
x35         1760.2571   2163.149      0.814      0.416   -2479.754    6000.268
x36         -447.3771   6578.998     -0.068      0.946   -1.33e+04    1.24e+04
x37         3536.8293   2413.603      1.465      0.143   -1194.099    8267.757
x38          1.28e+04   7699.842      1.662      0.096   -2292.820    2.79e+04
x39         2778.9644   3750.297      0.741      0.459   -4572.030    1.01e+04

x40         7835.2336   1.88e+04      0.417      0.677    -2.9e+04    4.47e+04
x41         -540.5284   6331.026     -0.085      0.932    -1.3e+04    1.19e+04
x42        -3337.0025   5947.920     -0.561      0.575    -1.5e+04    8321.575
x43        -9093.6405   1.43e+04     -0.637      0.524   -3.71e+04    1.89e+04
x44        -5544.9791    1.8e+04     -0.307      0.759   -4.09e+04    2.98e+04
x45         -550.0229   1773.234     -0.310      0.756   -4025.757    2925.711
x46        -1089.9803   4024.872     -0.271      0.787   -8979.173    6799.213
x47         -249.8009   2120.856     -0.118      0.906   -4406.912    3907.310
x48          790.9781   2163.990      0.366      0.715   -3450.680    5032.636
x49          473.1859   2289.557      0.207      0.836   -4014.598    4960.970
x50          632.2053   2770.056      0.228      0.819   -4797.409    6061.820
x51         1212.3382   1912.279      0.634      0.526   -2535.939    4960.616
```

Fig 8. Ordinary Least Squares statistical output.

The table above summarizes our Ordinary Least Squares stats model. Take note of several columns. The coefficient is the variable weight, measuring the effect changes in the input variable has on our independent variable, in this case sales. Generally, this is the slope in slope intercept form. Std error estimates the standard deviation of the coefficient throughout the data. T values are a measurement of precision for determining the coefficient. P > |t| uses the T value to produce a P value, measuring how likely it is for the coefficient to be explained by chance. The last columns are measurements of coefficients within 95% of our data (two standard deviations) and generally do not include outliers. Notably, nearly half of our features have a negative coefficient, which implies that as the feature value increases and everything else is held constant, the predicted sales decrease as well. A slight majority of features have positive coefficients, so as those features increase, sales predictions also increase.

4. Discussion

4.1. Linear Model Evaluation

In summary, our linear regression model was able to fit and predict the data with a fairly low RMSE, meaning nearly all observed values were close to the mean line of the expected value, and a high R^2, meaning most of the variation in sales could be jointly explained by the model inputs. We tried to train other models too, but increasing our model complexity typically triggered less optimal results. Even when we ran cross validation, the Linear Regression model still emerged as the victor. This model could potentially be employed by Cookies as a forecasting tool to anticipate upcoming sales.

Accounting for T and P values, features that have a higher magnitude of T and a lower P demonstrate good precision, and high confidence that the input feature associated is important. This can be noted, for example, with our first sales feature x1, which corresponds to the previous month's sales. The P value is incredibly low, meaning that the prior month's sales data is unlikely to have the strong predictive effect it has due to chance. For every additional dollar in sales last month, the current month's sales are predicted to be $160,000 higher. The last 3 month's average sales have a similarly large effect and low P value. The only other feature with a significantly low P value was x38, which is the number of hybrid products a brand sells. This feature was seen to be highly correlated to sales in pre-modeling. However, it is interesting our model is more confident that hybrid products have a real effect on sales compared to total products/strains, which were just as correlated pre-modeling. This suggests that brands looking to grow sales should add more hybrid products to their inventory. A few features including time on market, indica product share, sativa product share, or vape product share, all had relatively low P values but more varied T values and much smaller effects/coefficients.

4.2. Key Sales Predictors

As discussed earlier, the cannabis market is widely projected to grow, and Cookies will likely benefit from launching quality new products, or curating their current inventory to maximize sales. Growth in the industry is forecast 10% annually. Further, the COVID-19 pandemic has created lockdowns, remote working opportunities, and potentially more interest in recreational activities such as cannabis use. From our data, we believe some key indicators of success in the industry include prior sales and units sold, existing time in the market, total product count and strain count, specifically hybrid product count, and finally whether the brand sells CBD. Correlations also indicate vape and flower products may be prioritized as major drivers of sales, in comparison to other categories. These key indicators are supported by our high correlation numbers discussed in section 3.1 as well as articles on cannabis growth. These features also had decent magnitudes of positive correlation in our linear regression model for projecting sales. Overall, their usefulness as predictors is generally substantiated with confidence.

These same concepts are supported by research and intuition. There is value in offering a variety of cannabis products and a variety of strains, because consumers typically like to be given more options. Time on the market is important and essential in establishing a well known brand that returning customers can trust. The observed importance of hybrid products was the most revelatory, perhaps indicating that users prefer buying hybrid options.

As an aside, the amount of beverages, topicals, and tinctures a brand sells generally does not correlate much with sales. The sales correlations were weak and their effect in the model is questionable. Thus, cannabis companies probably should not focus on stocking many topical products or worrying about their pricing for beverages. Selling flower, vapes, concentrates, or edibles, as well as the sativa and indica ratios of these categories, are far more correlated and more lucrative in driving profit.

4.3. Recommendations for Cookies

From our research we understand that forecasted sales, growth, and profits of the cannabis industry are increasing annually at a rapid rate. Cookies is smart to pursue data science as a tool for better understanding their consumers. We recommend Cookies leverage our model to optimize sales..

As noted in our key predictors, we have seen that flower is the most popular selling category in the industry, followed closely by vapes. However, the popularity of edibles is projected to grow nearly threefold in the next 5 years. Although this trend was not necessarily captured in our modeling, we trust the consensus industry sentiment that edible use will rise. Furthermore, we have also seen that offering a large variety and inventory in popular cannabis categories is strongly correlated to sales, unlike topical, tincture, and beverage products.

Therefore we recommend that Cookies perform inventory restructuring to offer more products in the cannabis categories that generate sales such as flower and vapes. They could also focus less or produce less of the products that do not correlate well with sales such as tinctures and topicals. Cookies should also consider launching more products in the edible category since this section of the industry is expected to triple in sales soon.

Our last recommendation is for Cookies to offer more hybrid products, perhaps forming strategic partnerships with hybrid-focused brands. Cookies already has Lemonade as their sativa-centric sister brand, but our modeling indicates hybrid strains drive more sales. This is another trend to consider.

4.4. Next Steps for Analytic Work

In a 2021 Thrillist article titled “Best Dispensaries in LA”, Cookies is listed as a highly recommended dispensary for selling edibles, wax, flowers, and baked goods. Not only is Cookies part of a growing industry, the brand is already reputable and well established, offering a wide variety of products. The business is set up well to adapt, possessing the infrastructure and inventory to make changes based on our data investigation and modeling.

However, even though our models were able to predict features that highly correlated with sales, there are still some other features that could play an important role in determining the success of a cannabis company in the near future. For example, features we would like to include in the future are: average consumer distance to dispensary, average customer purchase volume at a dispensary, user’s preferred methods of access (i.e. pickup, delivery), average frequency of purchases, and percentage of sales completed online vs in-person retail.

We believe these features, if collected, would add to our current set of key predictors and allow us to make additional practical recommendations, such as whether Cookies should invest more time and effort into their flagship website upgrades or focus on in-store customers. Furthermore, Cookies is still a developing business along with Lemonnade and as they expand into other countries. There are still opportunities in the future for them to grow in many locations both locally and internationally, especially as regulations loosen or become lifted. Thus, finding their next hotspots to start another store or even finding out which stores may need a relocation could be important in driving more profits in the long run. Geographic data about store foot-traffic and average consumption by locale could be added to enable this analysis.

Lastly, we read articles provided in the assignment specifications, but we did not personally work on the data that BDSA used to generate their statistics on industry-wide trends, such as edible growth and beverage sales. We did see a slight correlation in edibles being a good predictor of sales as well as beverages not being well correlated to sales. However, since we are recommending that Cookies explore creating more edible products, there is still value in using data science to guide their new product development. Through user test groups we can collect data on favorite kinds of baked goods, flavors, candies, snacks, chocolates, and price point information for such products. We can also use data about consumer edible purchases and average cost to help reinforce the statistics proposed by BDSA. If these models are consistent with our current model predictions and the BDSA articles, Cookies can benefit from a wonderful new product launch that consumers helped develop in the planning process.

Lastly, although we are confident in our current predictions, it should be noted that data science can and should be used continuously to update strategies in reaction to new trends and help Cookies stay ahead of their competition. Cookies should continue to contract data scientists to run updated models in the future. In this way, Cookies can repeatedly revisit the state of the cannabis industry to account for new changes and update their inventory and sales strategies accordingly.

5. Conclusion

In summary, we were tasked with helping Cookies develop initial predictive models to help them keep a competitive edge as one of the largest and fastest growing cannabis brands in the world. Our aim was to develop a predictive model that would accurately forecast sales and give us insights based on various consumer and market features to help us guide Cookies toward lucrative, bountiful, and advantageous business decisions.

We began researching the current state of the market by reading information provided by BDSA and other articles pertaining to cannabis sales in the United States. From these articles, it was clear that sales for this industry are projected to grow by about 10% annually over the next 5 years. We also obtained multiple csv files covering data on features relevant to the consumer market for the cannabis industry. Our first task was to merge the datasets so we could link relevant information into one data frame. By doing this, we could visualize the data frame as well as train, test, and deploy a predictive model that would benefit Cookies in their business strategy and sales forecasting.

We then sanitized features by rounding values, converting types, and performing feature crosses. We adjusted for outliers to prevent skewing model results based on anomalies in the data set. This yielded a more safe and balanced distribution on frequencies when we plotted the target variable, Total Sales. Next, we developed basic time series features to augment the data set by interpolating units and sales, imputing missing price points (using averages), and adding time features.

Collecting statistics on brand-level product data was our next focus. We accumulated details, count, product ratio, and average price for the different types of cannabis products and strains. Some of these figures were adjusted such as boolean to numeric value conversions for further modeling. Each of these accumulated features was iterated over to be aggregated into our full data frame.

In pre-modeling, we calculated summary statistics and correlations on our data set. The data showed strong correlations between new sales and past sales, as well as with popular strains and products. We then trained various models to find the best fit for our data. We compared models between Linear Regression, Decision Tree Regression, K-Nearest Neighbors Regression, and Ensemble Regression, of which Linear Regression performed with the smallest RMSE and largest R^2 value. After running grid search optimization, 10-fold cross validation, and comparing residual plots and scores between high-performing models, our Linear Regression model still had the best results. Our last step in observing the data was calculating feature weights and significance using ordinary least squares.

From our research and predictive modeling, we can confidently conclude that Linear Regression fits the data best. In our model, previous sales and units sold were correlated highly with total sales. Given that the reports we have studied expect the industry to grow, this is a great predictor of sales forecasting. Other indicators based on correlation and modeling stats are hybrid products, flower products, vape products, and time on the market. We substantiated the predictive capabilities of our model by validating with ordinary least squares.

Cookies is well established, sells a wide variety of products, and has been marketing extremely well, but there are still avenues of improvement we can explore for improving sales. Cookies may consider restructuring inventory to prioritize flower, vape, and hybrid products, which were shown to be very predictive of increased sales. Also, while edibles in our model were slightly less predictive, there is an expectation of category growth. In contrast, our data shows that offering topicals, beverages, and tinctures is not as correlated to sales, so limiting the stock or costs spent to produce these items could be beneficial while focusing on, once again, flowers, vapes, and edibles. Those categories are where the money is made.

Looking ahead, we could run models on Cookies web traffic, online sales, customer location, and many other important features. What we know for sure is that cannabis sales are increasing and people enjoy the Cookies and Lemonnade brand. Hopefully, with the help of data science guiding upcoming business decisions, Cookies can take their brand and their profits to a new level.

6. References

BDSA. (2021, September 27). *California Cannabis Market*. BDSA. Retrieved December 3, 2021, from https://bdsa.com/wp-content/uploads/2021/10/California-Cannabis-Market.pdf.

BDSA. (2021, October 11). *Essential Cannabis Insights*. Vol 4, Issue 7. Retrieved December 3, 2021, from https://bdsa.com/wp-content/uploads/2021/10/Essential-Cannabis-Insights-Fall-2021-Market-Forecasts-Update-1.pdf

Deckard, E., Goodson, G., Fergus, J. (2021, April 14). *The 19 Best Dispensaries in Los Angeles. Thrillist*. Retrieved December 3, 2021, from https://www.thrillist.com/lifestyle/los-angeles/best-dispensaries-in-los-angeles-la.

Lukas, J. (2021, July 22). *2021 State of Cannabis*. BDSA. Retrieved December 3, 2021, from https://bdsa.com/wp-content/uploads/2021/07/2021-State-of-Cannabis.pdf.

Roberts, C. (2016, February 22). *Million Dollar Cookie: How Berner Built a Business Empire on Marijuana*. SF Weekly. Retrieved December 3, 2021, from https://www.sfweekly.com/news/million-dollar-cookie-how-berner-built-a-business-empire-on-marijuana/.

Schaneman, B. (2021, October 8). *How to Build a Global Flower Brand: Q&A with Cookies Co-founder Berner*. MJBizDaily. Retrieved December 3, 2021, from https://mjbizdaily.com/how-rapper-berner-built-a-major-marijuana-brand-cookies/.

Whitmore, M., Belz, M., Cacioppo, L., Iorio, P., Peters, T., Thakore, C. (2016). *Recreational Marijuana - Insights and Opportunities*. Deloitte. Retrieved December 3, 2021, from https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/Analytics/ca-en-analytics-DELOITTE%20Recreational%20Marijuana%20POV%20-%20ENGLISH%20FINAL_AODA.pdf.

Zhu B, Guo H, Cao Y, An R, Shi Y. *Perceived Importance of Factors in Cannabis Purchase Decisions: A Best-worst Scaling Experiment*. Int J Drug Policy. 2021 May; 91:102793. doi: 10.1016/j.drugpo.2020.102793. Epub 2020 May 29. PMID: 32482489; PMCID: PMC7704653. Retrieved December 3, 2021, from https://pubmed.ncbi.nlm.nih.gov/32482489/

7. Appendix

7.1. Correlation Table

Feature	Correlation Coefficient
Sales ($), Previous Month	*0.960694*
Sales ($), Past 3 Month Avg	*0.949696*
Units Sold, Previous Month	*0.880918*
Units Sold, Past 3 Month Avg	*0.875456*
Total Product Count	*0.473135*
Total Strains	*0.456741*
Flower Strains	*0.410579*
Hybrid Product Count	*0.390610*
Flower Product Count	*0.388338*
Preroll Strains	*0.381650*
Preroll Product Count	*0.369874*
Average Vape Product Price	*0.368463*
Sativa Product Count	*0.359821*
Time on Market	*0.357373*
Indica Product Count	*0.352755*
Vapes, Average Volume (mg)	*0.344276*
Concentrate Strains	*0.322558*
Concentrate Product Count	*0.315638*
Vape Strains	*0.297202*
Vape Product Count	*0.293073*
Average Concentrate Product Price	*0.279774*
Average Preroll Product Price	*0.272425*
Vape Product Share	*0.260202*
Sells CBD Products	*0.254731*
Edible Product Count	*0.210648*
Average Flower Product Price	*0.209106*
Hybrid Product Share	*0.192813*
Indica Product Share	*0.183814*
Sells Devices	*0.172127*
Prerolls, Average Volume (mg)	*0.169533*
Sativa Product Share	*0.150904*
Change in Sales ($), Previous Month	*0.143776*
Concentrate Product Share	*0.141298*
Available Flavors	*0.140653*
Average Edible Product Price	*0.133151*
Tincture Product Count	*0.126048*
Sells Flavored Products	*0.119614*
Preroll Product Share	*0.114252*
Average Tincture Product Price	*0.095585*
Change in Units Sold, Previous Month	*0.081244*
Flower Product Share	*0.052394*
Average Beverage Product Price	*0.042708*
Beverage Product Count	*0.034428*
Topical Product Count	*0.021800*
Average Topical Product Price	*-0.009530*
Edibles, Average THC/Item (mg)	*-0.012231*
Tincture Product Share	*-0.041620*
Edible Product Share	*-0.046314*
Avg Product Price	*-0.054648*
Beverage Product Share	*-0.076534*
Topical Product Share	*-0.126615*