Since the legalization of recreational marijuana in November 2016, recreational marijuana has been a booming industry in the state of California. From the fiscal year of 2020-2021 alone, nearly a billion dollars of tax revenue was generated in sales, despite the vast majority of the industry remaining underground. As in any field of business, brands may want to be able to accurately predict future sales based on time, and the products they put out. In this study, we used machine learning and data analytical tools to attempt to extract and accurately predict future sales of a company based on features of the company. As a result, we were able to predict the sales of any company during a certain month with ~90% accuracy%.

Background: As aforementioned, major and minor cannabis retailers alike inevitably want to be able to accurately predict future sales. Although the market is largely underground, it is expansive enough that the licensed brand sales is more than sufficient enough to create meaningful analysis. But their existence and dominant presence accounts for a strong influence in the sales of legal cannabis for companies like Cookies, yet it cannot be directly reflected in any way through data, which only exists for licensed brands. Its novelty has led to an explosive increase in sales over the years, unprecedented in history because history is so short – we only have so very few years of recorded sales, and many companies starting up have no such history to work with at all. Details about location, demographics, as well as migration from black market cannabis sales to legal ones are all outside the domain of most available data. The market is still rather unstable and cultural changes and social stigma towards the use of cannabis is volatile and difficult if not impossible to predict through data. However we will operate within these conditions nonetheless.

In the context of the specific datasets, it was proven difficult to combine a dataset which measures a brand's sales per month with a certain products sale in an unknown timeframe. In the

process of trying to adding a notion of time to BrandDetails, the researcher lost all sense of products. In fact, the original intended goal of the research was to make a predictive model in which given the year (from 2018 up to 2021), month, brand name, and product name, to generate the predicted sales of this particular model during that month. However, complications with making functional code lost the notion of products completely, and instead the goal was rerouted to be to predict a brand's sales based on month.

Methodology: The research gathers data from four separate dataframes, three of which detail every brand's sales by $, units sold, and average retail prices by month, and a fourth of which has no such immediately obvious timeframe and instead groups by product into separate rows based on price increment. The researcher tried to introduce a sense of time towards the fourth dataset, but due to incompetence was not able to properly code it so. Instead, he elected to introduce a sense of products through categorization and CBD/THC, as well as the number of types of products into an amalgamation of the first three. However, the notion of individual products was lost in the process, and the resulting predictive model was left to predict an entire brand's sales rather than a brand's product. The researcher than employed a series of predictive models including linear regression, principal component analysis, and gradient boosting to try to accurately predict validation data through training and testing a separate dataset, which would later be enhanced via cross-verification. The researcher than selected the best model which was the gradient boosting model with cross validation, which determined that the highest values of initial variables i.e. for learning rate gave the most accurate model. Then, it used the model to predict any brand and month's sales based on this.

Results:

Linear Regression

```
10184      8949.5
14044    227501.0
214      630737.0
25747     11611.1
14940    172428.0
           ...
4443      56186.4
23516    128196.0
2777     169458.0
413       14502.5
11257    200517.0
Name: Total Sales ($), Length: 5056, dtype: float64
[ 13614.81823535 256161.52089473 437345.39548284 ... 185771.28848582
   6131.388766   198203.19404728]
explained_variance:  0.7667
r2:  0.7666
MAE:  48333.9412
MSE:  8643912260.9985
RMSE:  92972.6425
```

```
Results from Grid Search

The best estimator across ALL searched params:
Lasso(alpha=1)

The best score across ALL searched params:
0.8429486899832297

The best parameters across ALL searched params:
{'alpha': 1}
13796      5848.92
12397     31712.00
7799      97719.30
23641    216149.00
16337     77426.50
           ...
16798     50182.60
553       69510.40
20942      1163.68
24066    325654.00
22365     39825.00
Name: Total Sales ($), Length: 5056, dtype: float64
[ -1062.66971513  41587.895227   120492.44156441 ...   2885
 494200.7976688  224766.32610326]
explained_variance:  0.8401
r2:  0.8401
MAE:  38595.5978
MSE:  5698107791.7346
RMSE:  75485.8119
```

Linear regression with Lasso regularization

```
13796      5848.92
12397     31712.00
7799      97719.30
23641    216149.00
16337     77426.50
           ...
16798     50182.60
553       69510.40
20942      1163.68
24066    325654.00
22365     39825.00
Name: Total Sales ($), Length: 5056, dtype: float64
[  8379.51235308  26540.738697   114593.78148329 ...  14571.2276721
 428767.14953926 156749.64277901]
explained_variance:  0.8167
r2:  0.8166
MAE:  42201.9331
MSE:  6533034862.128
RMSE:  80827.1914
```

GBR

```
Results from Grid Search

The best estimator across ALL searched params:
GradientBoostingRegressor(learning_rate=0.2, random_state=0)

The best score across ALL searched params:
0.8862534812902025

The best parameters across ALL searched params:
{'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 100}
13796      5848.92
12397     31712.00
7799      97719.30
23641    216149.00
16337     77426.50
           ...
16798     50182.60
553       69510.40
20942      1163.68
24066    325654.00
22365     39825.00
Name: Total Sales ($), Length: 5056, dtype: float64
[  4539.55702373  31462.87872194 130201.71232353 ...   1229.37474065
 481628.08653585  76113.29808946]
explained_variance:  0.8955
r2:  0.8955
MAE:  27369.0355
MSE:  3723519682.4119
RMSE:  61020.6496
```

GBR with Cross validation on parameters (grid search)

Discussion: This research, analysis, and model as well as its implications is severely limited by the ineptitude of the data researcher responsible in properly coding what he had in mind, as well as confusion about the intended goal of prediction. As a result, the model is admittedly very limited in its ability to make predictions that would be useful to a company in realistic executive scenarios, even if it can accurately predict a brand's sales in a month given some information and a follow-up function which generated average stats if the given month is not within the available time frame. If the researcher was adept at Python, he would have instead made the model predict a certain product's sales of a brand base done in a month. Cookies would likely find more use in

with such a predictive model and see whether it is worth introducing a new product it has not yet tried, and the resultant predictive model would be able to use product category information in a time-sensitive manner to estimate the revenue generated from such a product, and thus gauge whether or not such a product would be worthwhile to produce for sale.

Conclusion: The research however does indicate that total units sold as well as average monthly revenue were huge indicators in determining price, but perhaps more importantly the time or month in which the sales were conducted was just as if not more important – perhaps a testament to the explosive growth of the cannabis industry in recent years. Further research can be done in properly implementing the model the researcher had in mind, but was unable to create.