Aleafia Health

Customer Behavior Analysis and Value Prediction

Presented By: Mo Di Wang

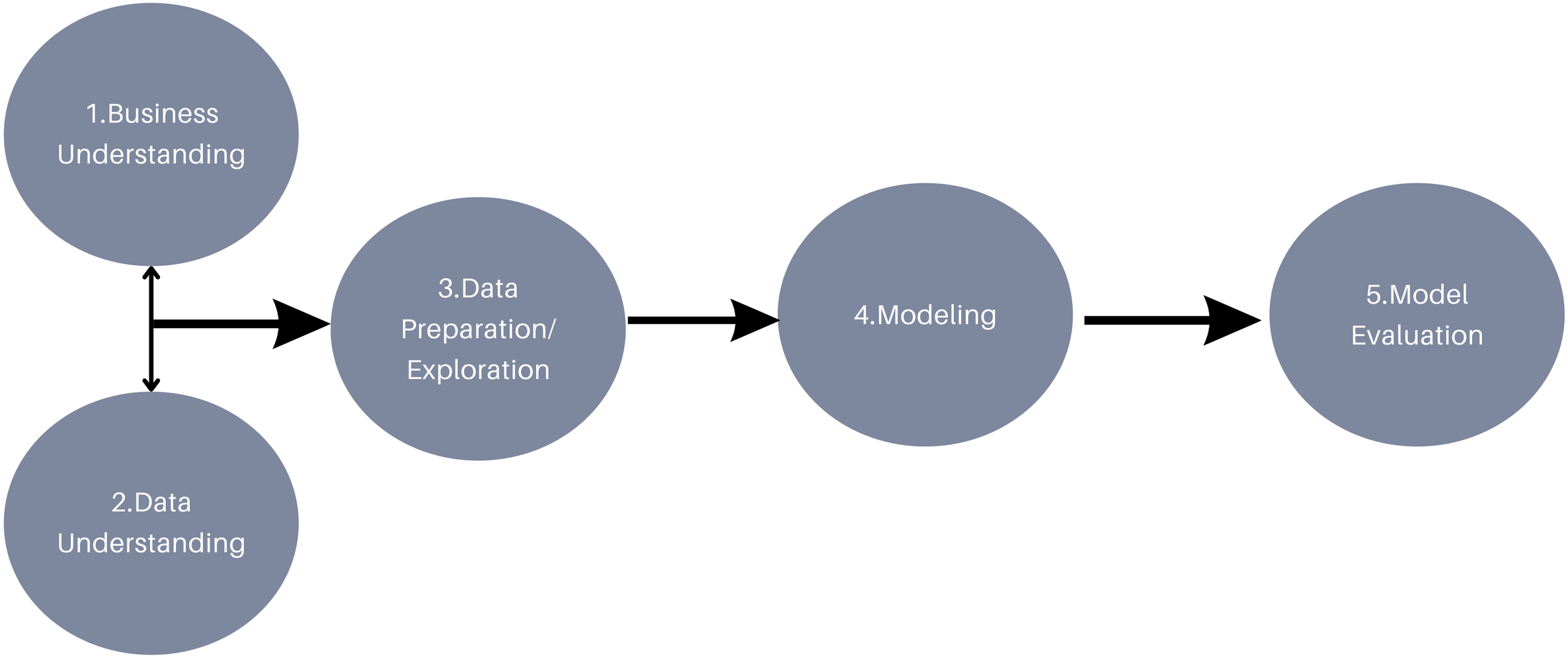
2022-08-15

Meeting Our Team

- 7+ years experience in Business Intelligence and Data Science.
- Experience in Retail/Manufacturing Industry
- Master's Degree From Queen's University (MFIT)
- Bachelor's Degree From University of Toronto
- Proficient in Python/SQL/BI Tools

Aleafia Health

SETTING THE CONTEXT: CRISP-DM APPROACH



Business Understanding: Discuss the business background and problem.

- Aleafia Health is a Licenced Producer of Medical and Recreational Cannabis

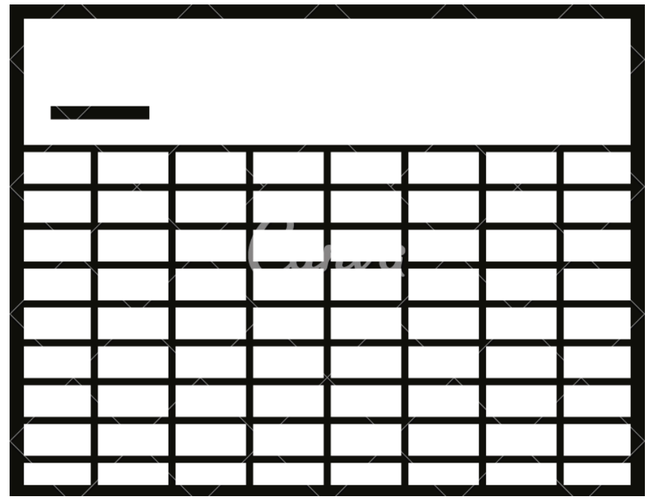
- The company builds relationships with clinics to sell Cannabis products directly to customers.

- In the past years, the marketing strategy in the company is to sell products at bundled and discounted prices.

- By analyzing historical sales performance, could we know the Total/Average customer value for the next few months (One or Three Months)?

Aleafia Health

Data Understanding: What resources are available?



- **Client Information Table**: Client ID, Age, Gender, Clinic, Prescription City, Address, Patient Status, Veteran Status, etc.

.

- **Sales Transaction Table**: Client ID, Purchase Date, Promotion Type, Product Purchased, Net Sales, Discounts, etc.

- **Location Table**: Business Regions, Prescription City,etc.

- **Education Service Fee Table**: Clinic Code, and Clinic Rate (%).

Data Understanding: What we observed from data?

Client Information Table	
# of Clients Registered	50000
Categorical Variable	22
Integer Variable	10
Boolean Variable	8
Missing Value	0

Sales Transaction Table	
# of Clients Purchased	20000
# of Orders	33000
Total Net Sales	99000000
Categorical Variable	8
Integer Variable	7
Boolean Variable	5
Missing Value	3000

- Not all patients in the client master list made a purchase in the last few months.

.

- Client-purchased locations vary very much (Spelling errors, cities abb (GTA, etc) .

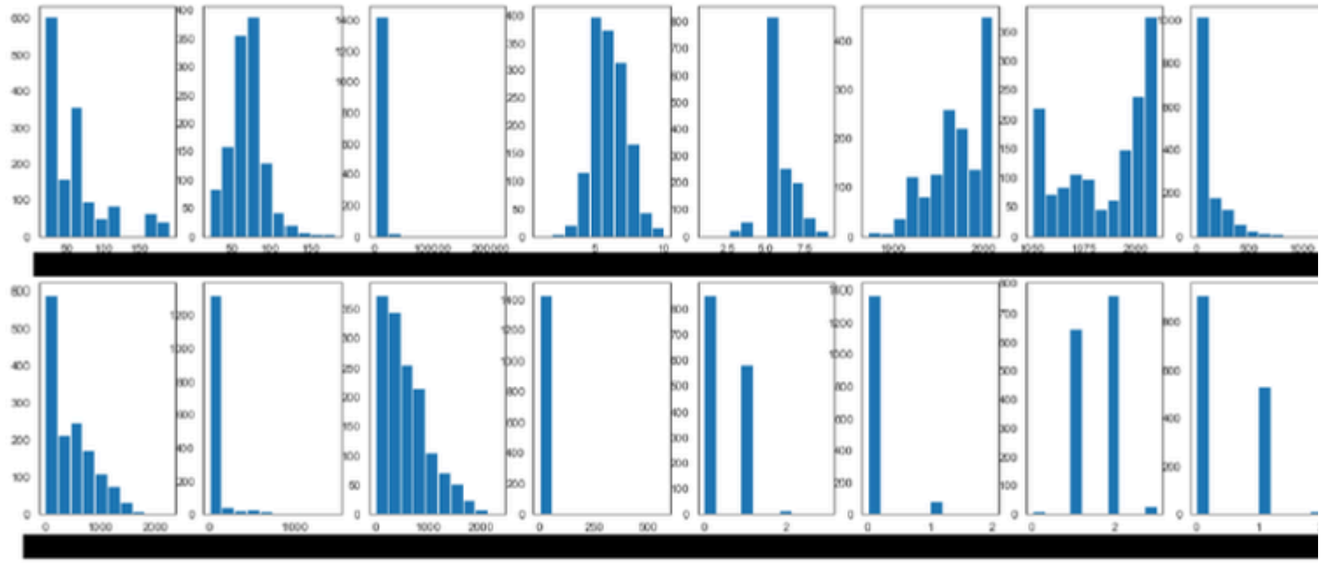
- There are data discrepancies in the sales transaction table, a column called wholesale price is either blank or price discrepancies.

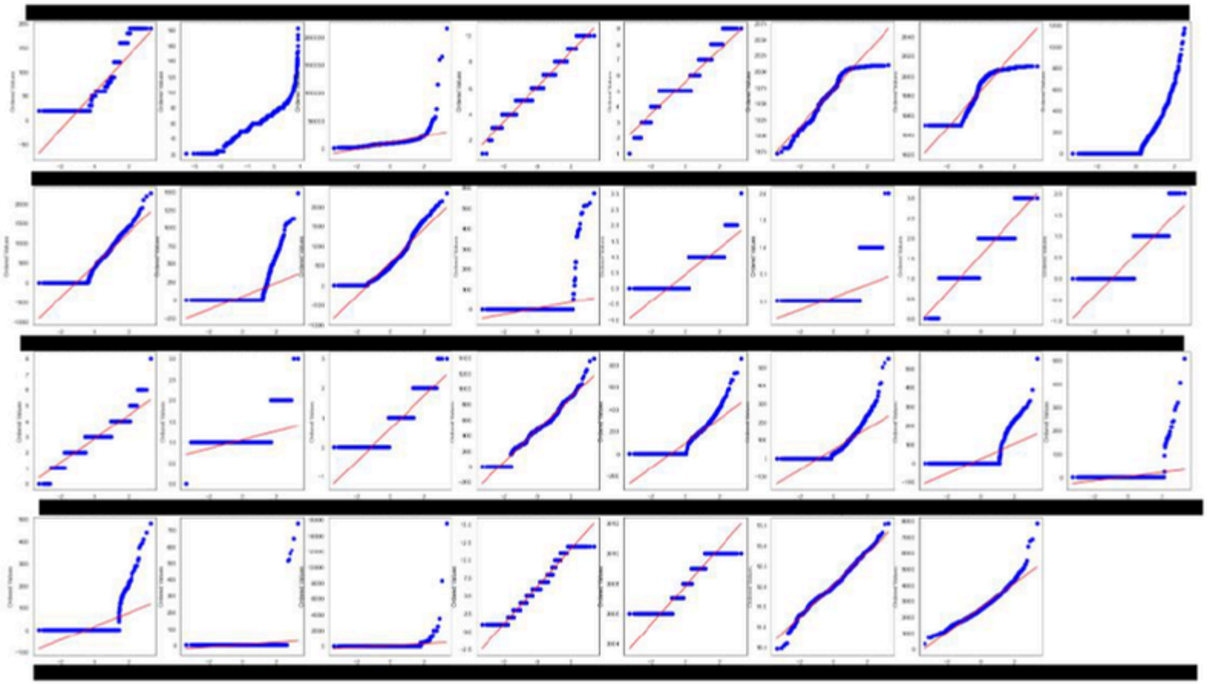
- People of different ages had a purchase.

Aleafia Health

Data Understanding: What we observed from data?

Density Plot For Numerical Variables

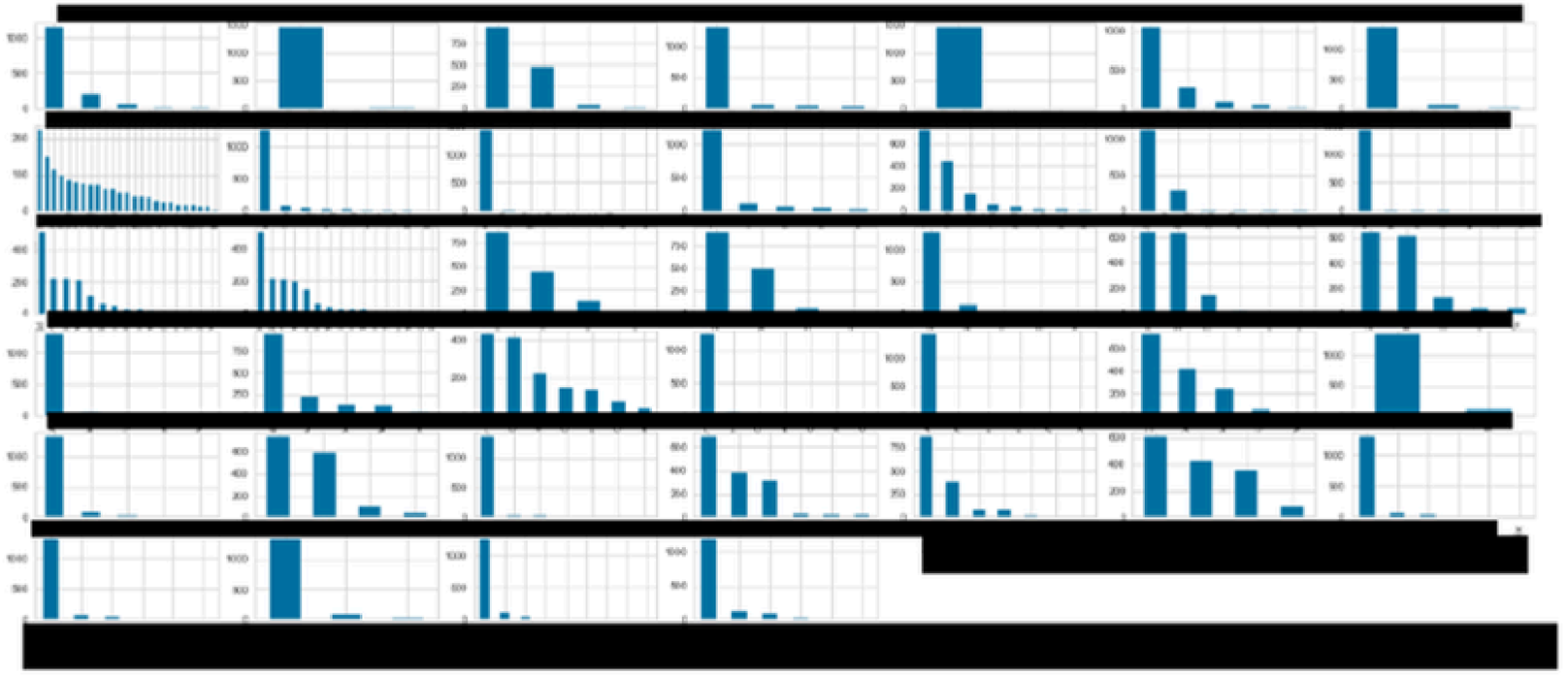




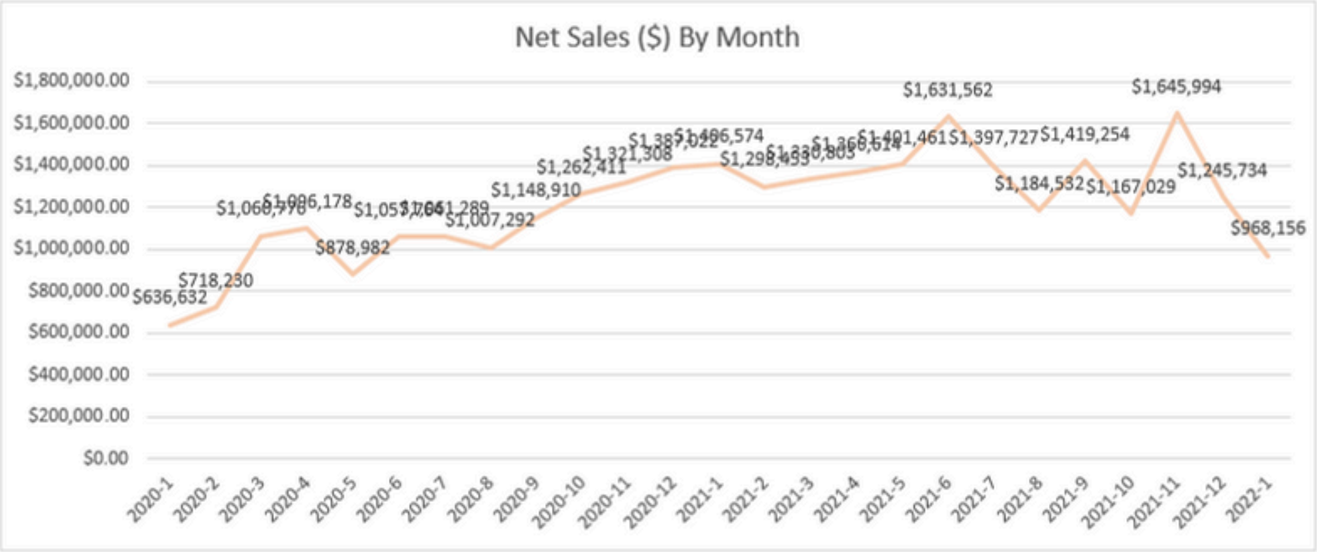
Aleafia Health

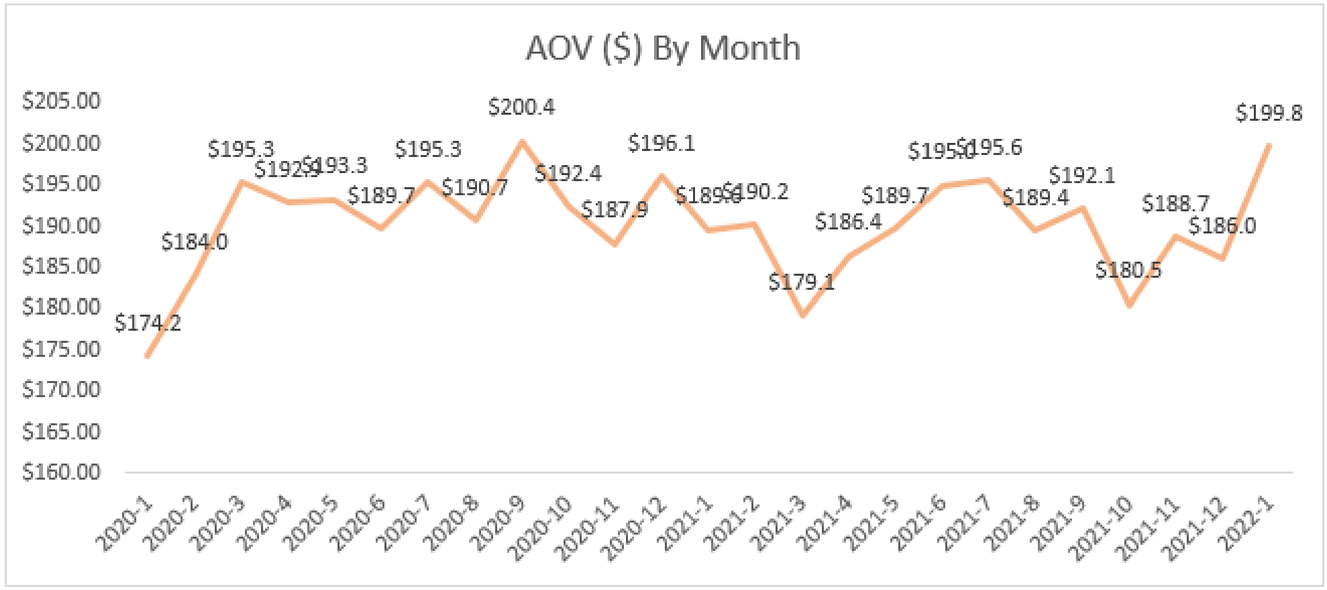
Data Understanding: What we observed from data?

Distribution For Categorical Variables.



Data Exploration: What can we learn from the data?





Average Spending By Age By Gender

Age Bucket	Female	Male	Undisclosed	Total
0-24	$631	**$900**	$430	$752
25-35	$620	$777	$636	$707
36-45	$645	**$1,000**	$356	$812
46-55	$658	**$958**	**$1,087**	$888
55+	$649	$800	$400	$732

- The sales and average order value have an inverse relationship.
- EDA process provides ideas about the relationship between data. It explains what has happened in business (BI).

Data Preparation: Feature Engineering

Aleafia Health

Transforming Transactional Sales into Consolidated Sales Table

Client ID	Purchase Date	Net Sales ($)
99999	2020-10-01	$58
99999	2020-11-01	$79
99999	2020-12-01	$21
12345	2021-10-13	$55
12345	2021-01-12	$32
12345	2021-02-13	$45

→

Client ID	3-Month_F0_Sales	3-Month_F1_Sales	3-Month_F2_Sales
99999	$158	$0	$160
12345	$55	$77	$209
45678	177	66	0

Prescription City
Promotion Type
Province
Clinic

→

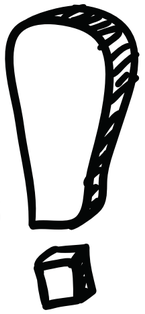
Gender
Veteran Status
Patient Status

→

One Hot Encoder (F1, F2 column features)

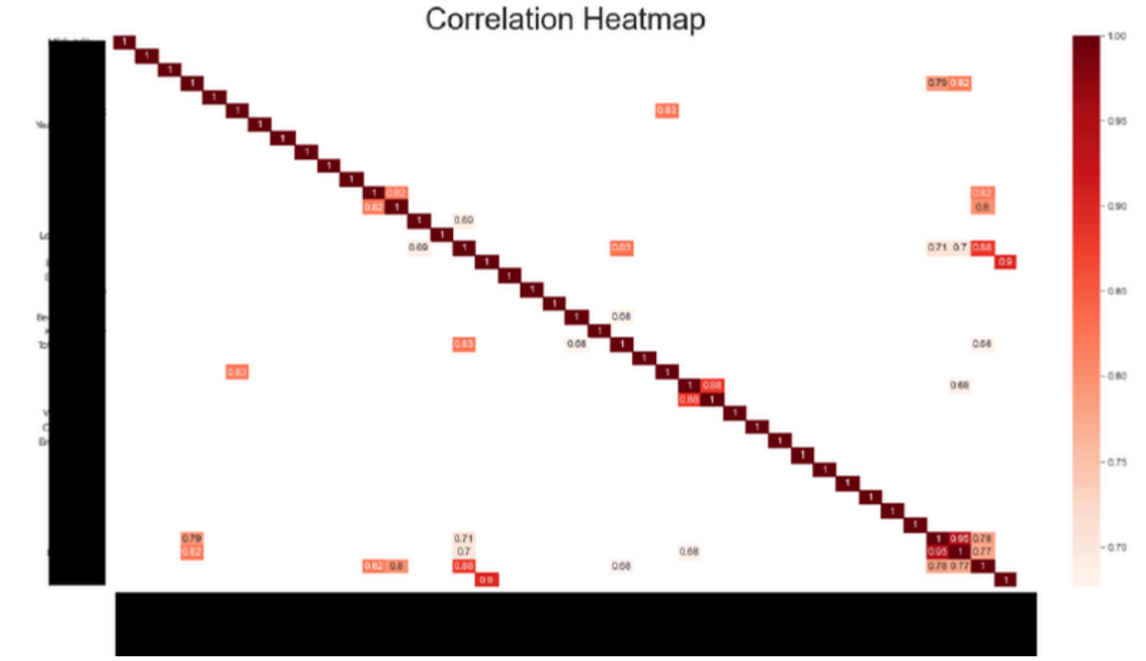
Target Encoding

Target Label:
3-Month_F4_Sales

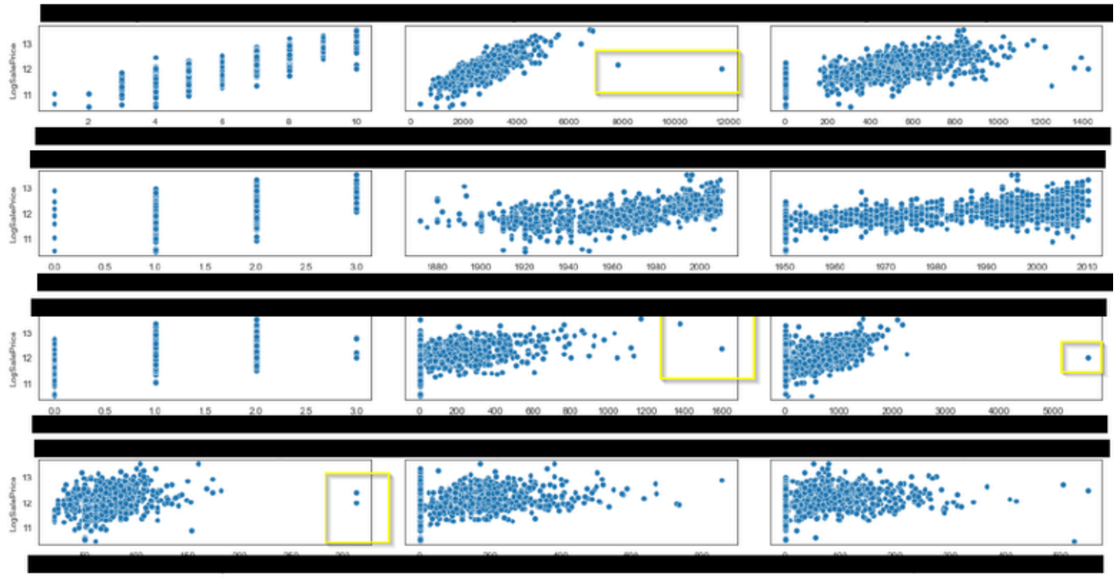


Data Preparation: Feature Engineering (Cont)

- Removing irrelevant features such as client ID, and columns that are before encoding.
- Correlation Analysis against the target label. Removing features that are less relevant to the target label.
- Removing Outliers. Identify the noise observations and remove them from the dataset.
- Log the features that are off from the best line of fit



Highest Correlation with Log Sale Price Before Removing Outliers



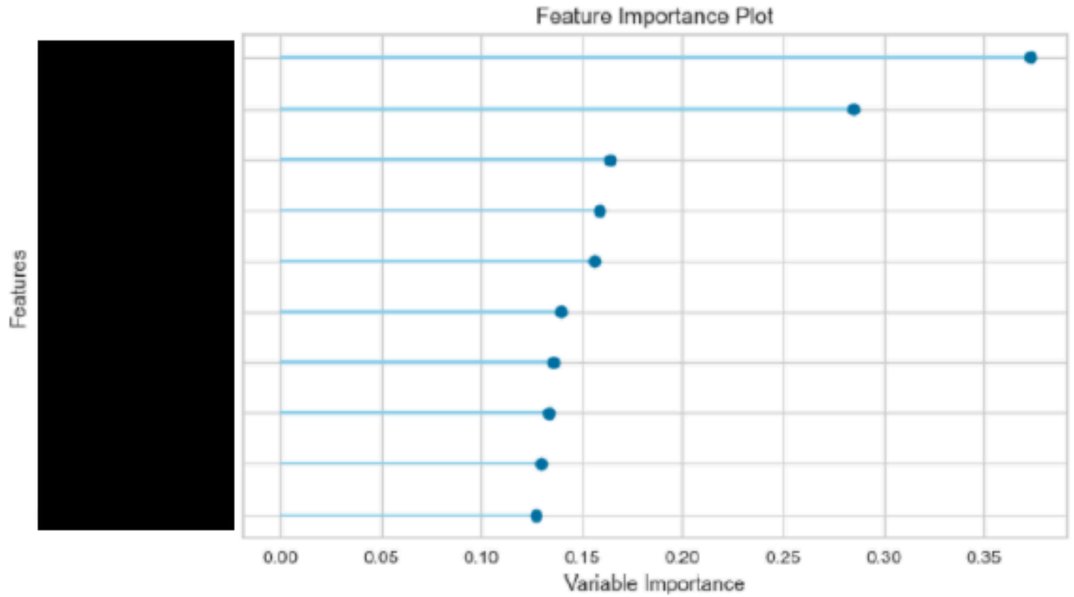
Modeling: Which model performs the best?

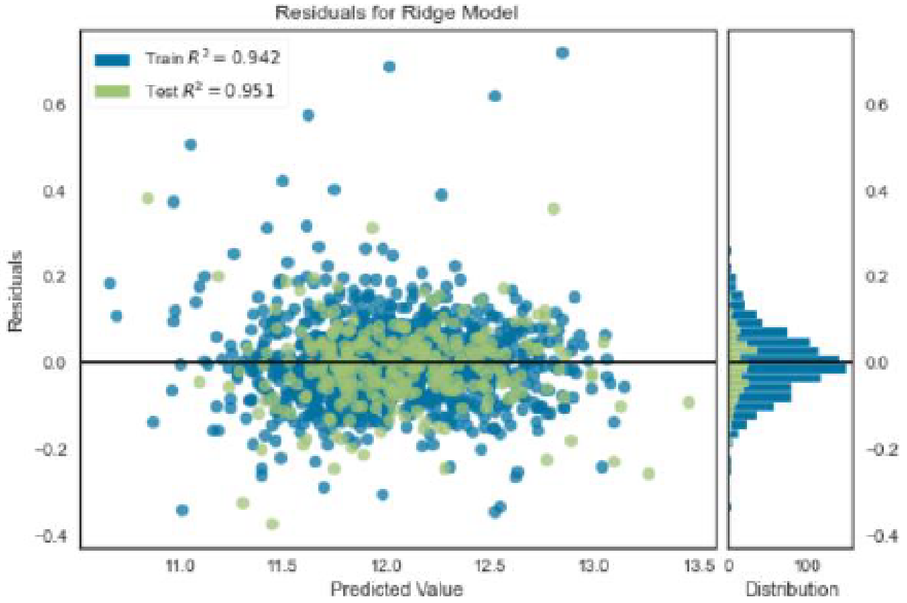
Aleafia Health

- Data Normalization
- Stratified Train Test Split
- Stratified K-Fold Cross-Validation

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
br	Bayesian Ridge	0.0825	0.0152	0.1210	0.9020	0.0094	0.0069	0.1610
catboost	CatBoost Regressor	0.0821	0.0151	0.1214	0.9025	0.0094	0.0069	3.8950
ridge	Ridge Regression	0.0848	0.0161	0.1247	0.8958	0.0097	0.0071	0.0690
gbr	Gradient Boosting Regressor	0.0879	0.0169	0.1287	0.8907	0.0100	0.0074	0.5240
lightgbm	Light Gradient Boosting Machine	0.0882	0.0171	0.1296	0.8890	0.0101	0.0074	0.6290
omp	Orthogonal Matching Pursuit	0.0888	0.0183	0.1328	0.8819	0.0103	0.0074	0.0490
xgboost	Extreme Gradient Boosting	0.0941	0.0186	0.1353	0.8786	0.0105	0.0079	0.9890
rf	Random Forest Regressor	0.0962	0.0202	0.1400	0.8710	0.0109	0.0081	0.9140
huber	Huber Regressor	0.0949	0.0225	0.1475	0.8551	0.0116	0.0080	0.5780
et	Extra Trees Regressor	0.1031	0.0233	0.1502	0.8499	0.0117	0.0086	1.0440
par	Passive Aggressive Regressor	0.1208	0.0294	0.1698	0.8068	0.0132	0.0101	0.0860
ada	AdaBoost Regressor	0.1351	0.0313	0.1760	0.7961	0.0136	0.0113	0.4640
knn	K Neighbors Regressor	0.1269	0.0317	0.1765	0.7947	0.0137	0.0106	0.0730
dt	Decision Tree Regressor	0.1514	0.0443	0.2099	0.7070	0.0163	0.0127	0.0390
lasso	Lasso Regression	0.3046	0.1540	0.3913	-0.0069	0.0301	0.0254	0.0860
en	Elastic Net	0.3046	0.1540	0.3913	-0.0069	0.0301	0.0254	0.0510
llar	Lasso Least Angle Regression	0.3046	0.1540	0.3913	-0.0069	0.0301	0.0254	0.9110
lr	Linear Regression	1.8651	551.6533	13.0053	-4082.8618	0.1996	0.1554	1.1270

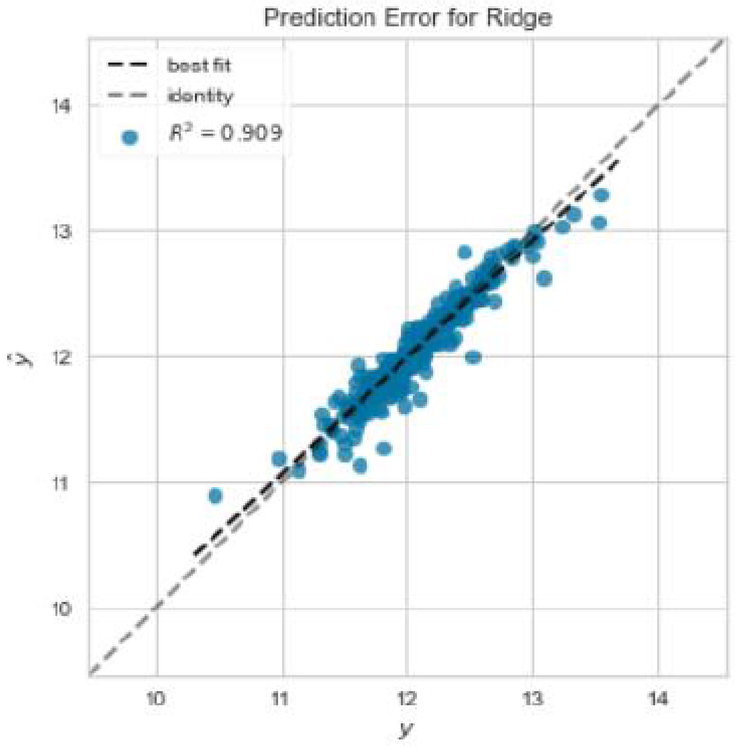
Evaluation: Which model performs the best?





	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0649	0.0073	0.0852	0.9467	0.0066	0.0054
1	0.0880	0.0131	0.1145	0.9178	0.0089	0.0074
2	0.0860	0.0265	0.1627	0.8519	0.0124	0.0072
3	0.0981	0.0215	0.1466	0.8944	0.0116	0.0083
4	0.0840	0.0175	0.1322	0.8825	0.0103	0.0070
5	0.0755	0.0090	0.0948	0.9255	0.0074	0.0063
6	0.0809	0.0129	0.1135	0.9159	0.0089	0.0068
7	0.0721	0.0094	0.0968	0.9395	0.0074	0.0060
8	0.0898	0.0189	0.1375	0.8697	0.0107	0.0076
9	0.0877	0.0164	0.1280	0.8727	0.0100	0.0074
Mean	0.0827	0.0152	0.1212	0.9016	0.0094	0.0069
SD	0.0092	0.0058	0.0234	0.0304	0.0018	0.0008

Evaluation: Which model performs the best?



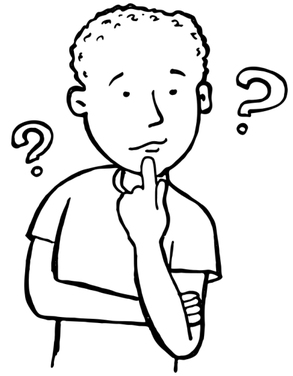
- The prediction errors are very close to the best fit line. The errors possess a normal distribution. There is no normality issue in the errors. The violation of the assumptions of the regression is not present.

Aleafia Health

Final Output:

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Ridge Regression	0.0664	0.0085	0.0921	0.9511	0.0071	0.0055

Client ID	3-month_F4_Sales
12345	$899
45678	$768
99999	$433



- Review Marketing Strategy.
- Further breaking down the spending into different product categories or products. Re-evaluating sustainable or unprofitable products.

What Did I learn from the project?

Aleafia Health

For Myself

- Gaining more business knowledge.
- Developing deeper machine learning techniques for performing the best results.
- How to effectively communicate the technical terms to non-tech stakeholders.

For Business

- The first machine learning technique in the business. It opens the opportunities to look at things differently from traditional approaches.
- In this case, if we keep doing what we do now. We will end up in the same situation and maybe even decline.

Aleafia Health

Q&A