Predicting Liquor Sales Profit and post-Pandemic Business Analytics

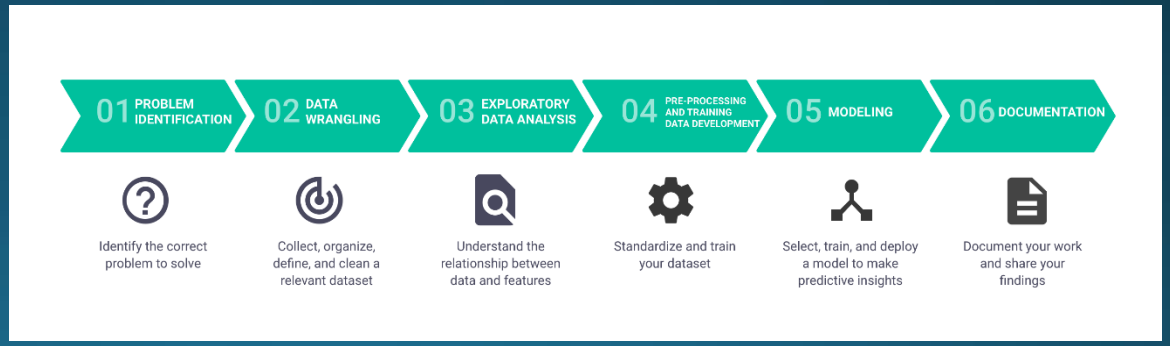
Author: Razvan Nelepcu

Mentor: Dhiraj Khanna

Project organization

The project follows all steps of Data Science workflow divided in the following major steps:

- Problem Identification
- Data wrangling
- Exploratory Data Analysis
- Pre-processing and training data development
- Modeling
- Documentation



Context

The Covid-19 pandemic is the major event of the 3rd Millennium that changed lives and businesses.

We know that some businesses went bankrupt or had to close their businesses overnight, such as bars, restaurant, hotels, or aviation. Some other flourished, such as online shopping, grocery stores or delivery services.

And nowadays more than ever we need data to understand how the pandemic impacted different parts of economy.

Our project will be directed on analyzing the spirits sales in Iowa since 2012 to present.

Problem statement

This project looks to resolve some issues that are of high importance for two distinct business entities: Iowa Department of Commerce, Alcoholic Beverages Division and Liquor Store Owners. These are the 3 Business Problems that we identified and that were solved with this project:

- Exploration on what was the impact of Covid-19 on the Alcoholic Beverages Industry.
- Cohort Analysis and Customer Segmentation using RFM(Recency, Frequency and Monetary value) and Unsupervised Learning
- Using time series analysis and predictions to predict profit for next month for Iowa Dept of Commerce from spirits.

For future work we also identified 2 other Business Problems using the same data:

- Storage capacity management exploratory analysis for Iowa Department of Commerce, Alcoholic Beverages Division.
- Lastly, we want to assist a hypothetical liquor store owner in Iowa in expanding to new locations throughout the state.

DATA COLLECTION AND ORGANIZATION

The main dataset used is the Iowa Liquor Sales database from Data.Iowa.gov.

It contains more than 24 million records of spirits purchase of Class “E” liquor licenses by product and date of purchase from January 1, 2012, to current, data provided and updated monthly by Iowa Department of Commerce, Alcoholic Beverages Division, each record with 24 descriptive columns as described in the figure below.

The data contains labels such as Invoice number, Store, Address, Zip Code, Geographical Location, beverage category, vendor name, Item Description, State Bottle Cost, State Bottle Retail, Bottles Sold and Sale.

The fact that the data is exhaustive for all the sales of this kind in the state of Iowa was a great statistical feature of our data because we were working with the whole population of sales of this category of alcoholic beverages and not with just a sample, which allowed us to create powerful business insights with great confidence levels.

FEATURES EXPLORATION

To clean data, we need to explore data.

First, we used the describe method to explore both numerical and categorical data. This allowed us to see the range of values, the count of values, as well as the count of distinct values for categorical features.

To ensure data quality we had to make sure the features were matching. We had 3 features for which we also had numbers:

- Store Name & Store Number
- County Name & County Number
- Category & Category number
- Vendor Name & Vendor number
- Item description & Item number

Form all these we had to make sure the pairings were unique.

Then we continued and explored the outliers of our data, with no significant findings

PREPARING OUR DATA INTO BUSINESS RELEVANT SUBSETS AND SAVING DATA

For each identified problem we selected the relevant features and for some we performed groupby methods. We identified 3 problems to be approached with our project and 2 Business problems for Future work. These are those tackled in this project:

Problem 1. Exploration on what was the impact of Covid-19 on the Alcoholic Beverages Industry

Problem 2. Cohort Analysis and Customer Segmentation using RFM(Recency, Frequency and Monetary value)

Problem 3. Using Time-Series to predict Profit for next month

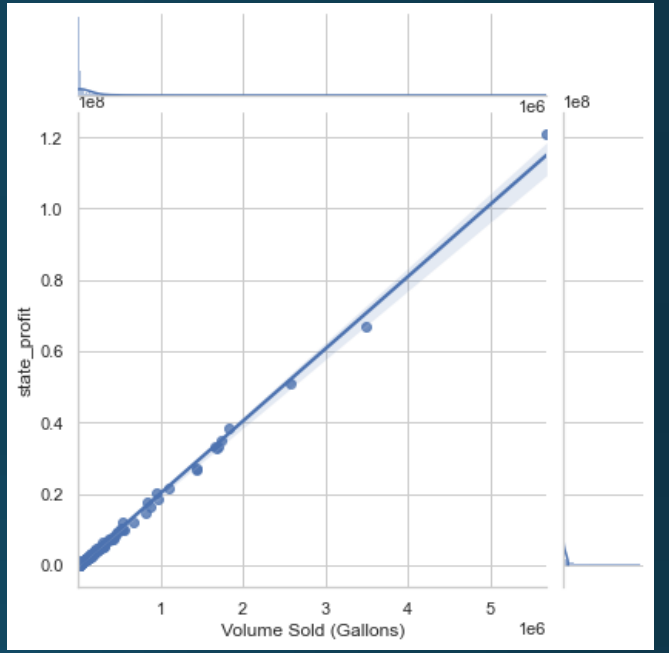
And these are for future work:

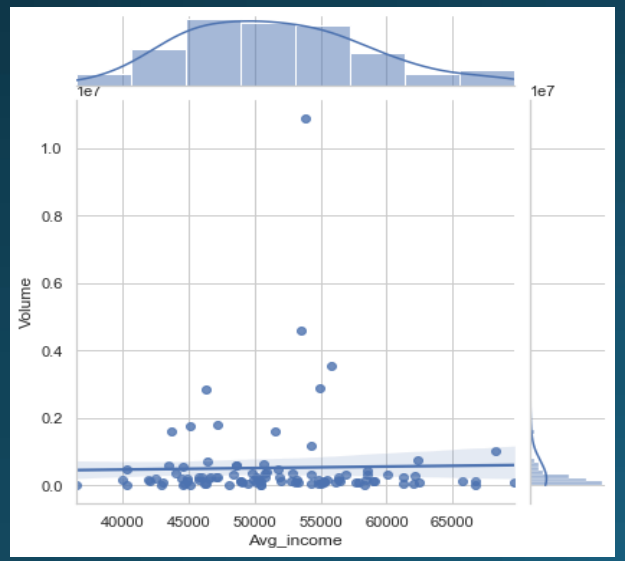
Problem 4. Storage Capacity EDA

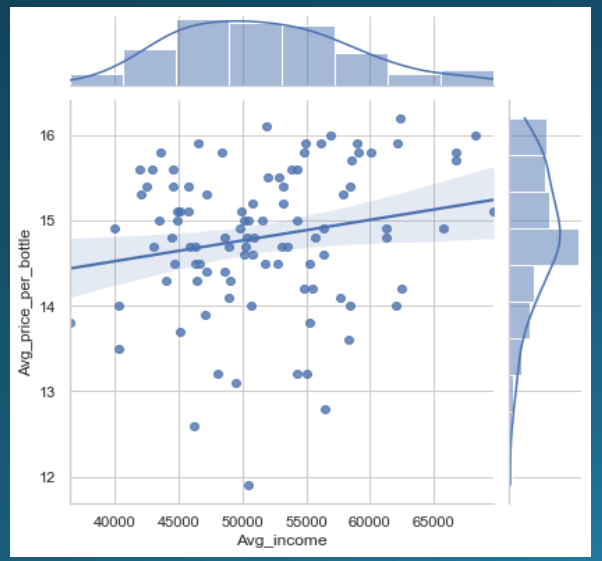
Problem 5. Assisting a hypothetical liquor store owner in Iowa in expanding to new locations throughout the state

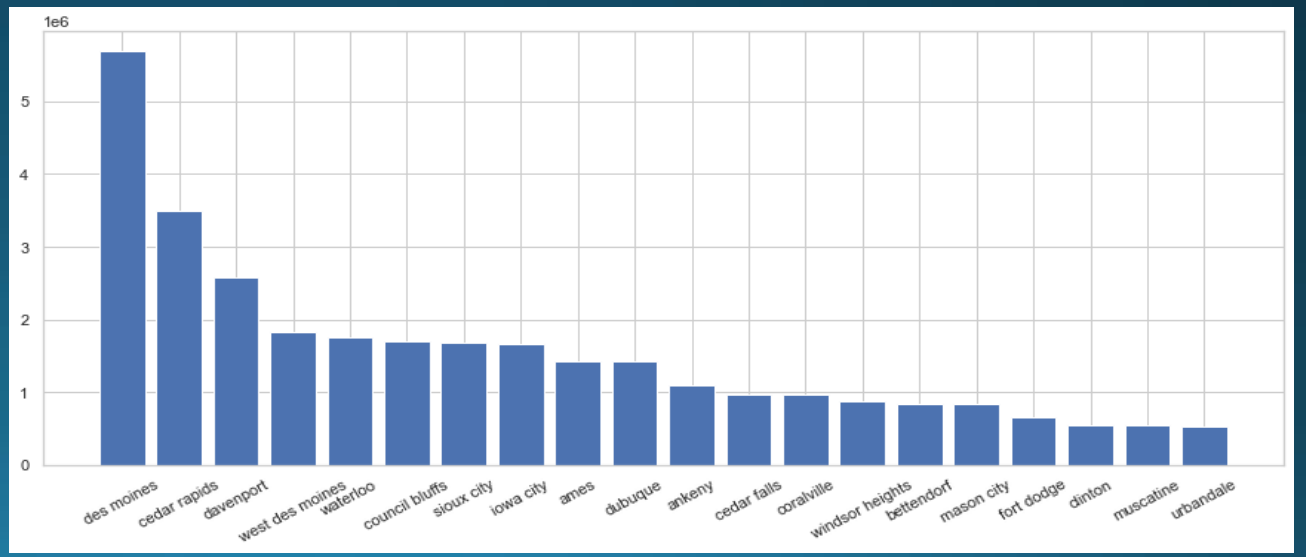
UNIVARIATE, BIVARIATE AND MULTIVARIATE ANALYSIS

We also wanted to explore if the wealthier counties are buying more or if they are buying the more expensive drinks.





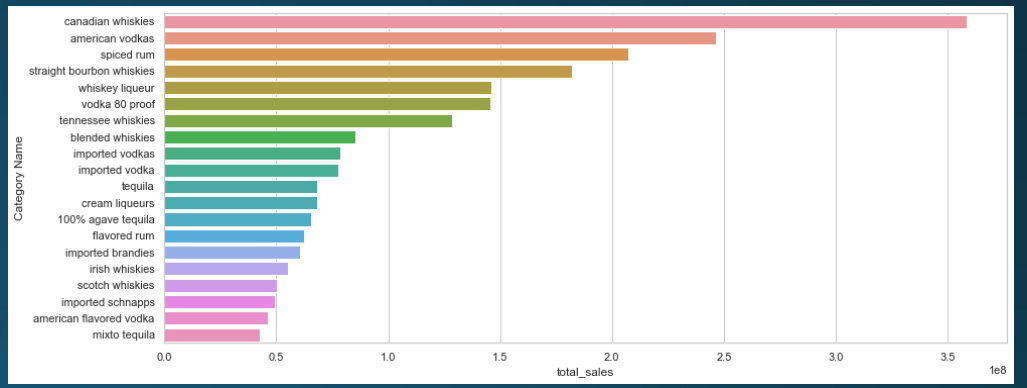


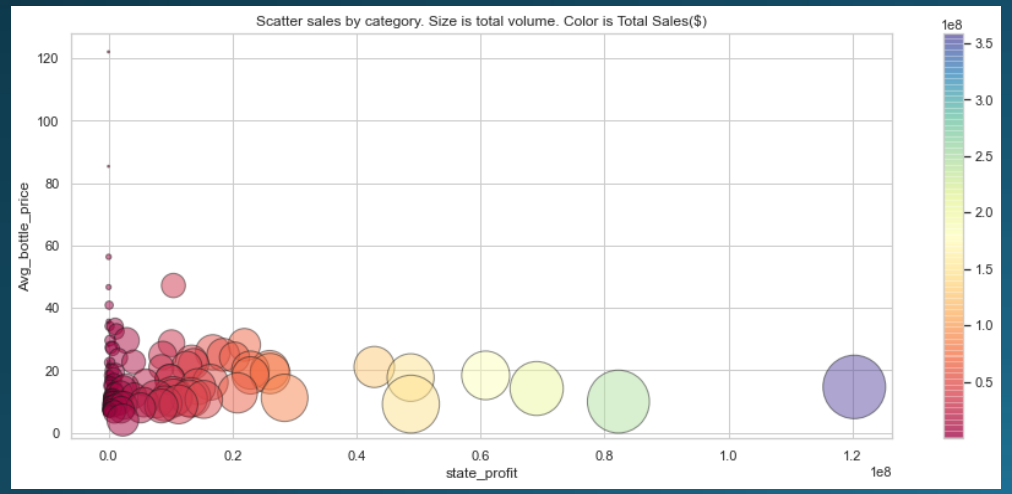


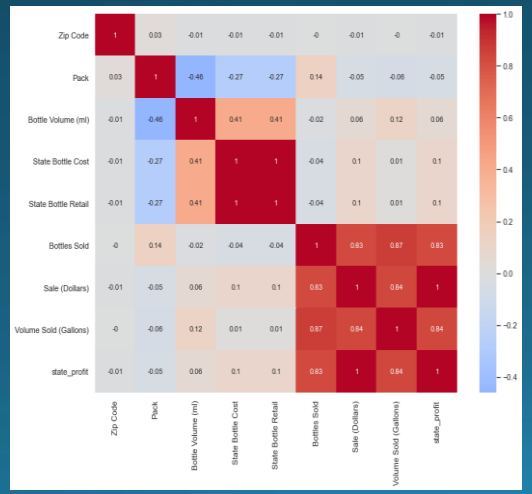
UNIVARIATE, BIVARIATE AND MULTIVARIATE ANALYSIS

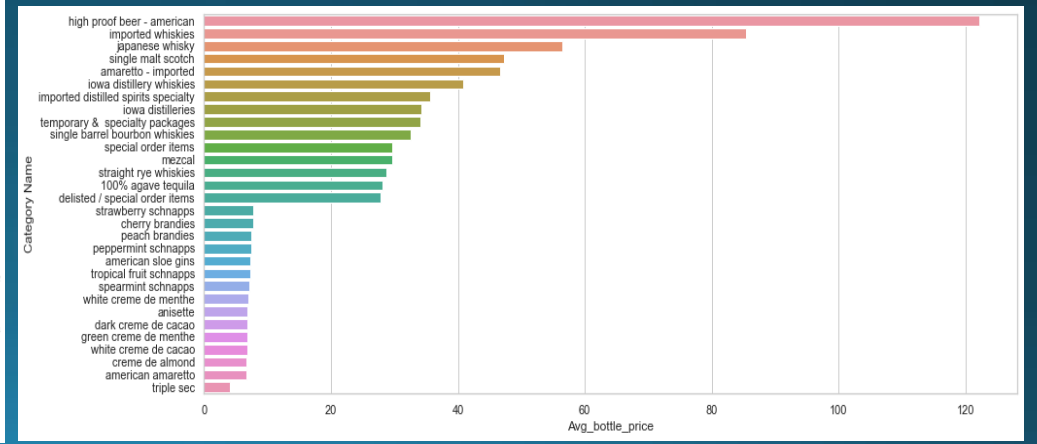
We also explored:

- the average bottle price per category, ranking these by sales and cost and Volume sold.
- a scatter plot of the categories with data regarding state profit, average bottle price, total volume and total sales.
- the heatmap of our numerical features correlations.





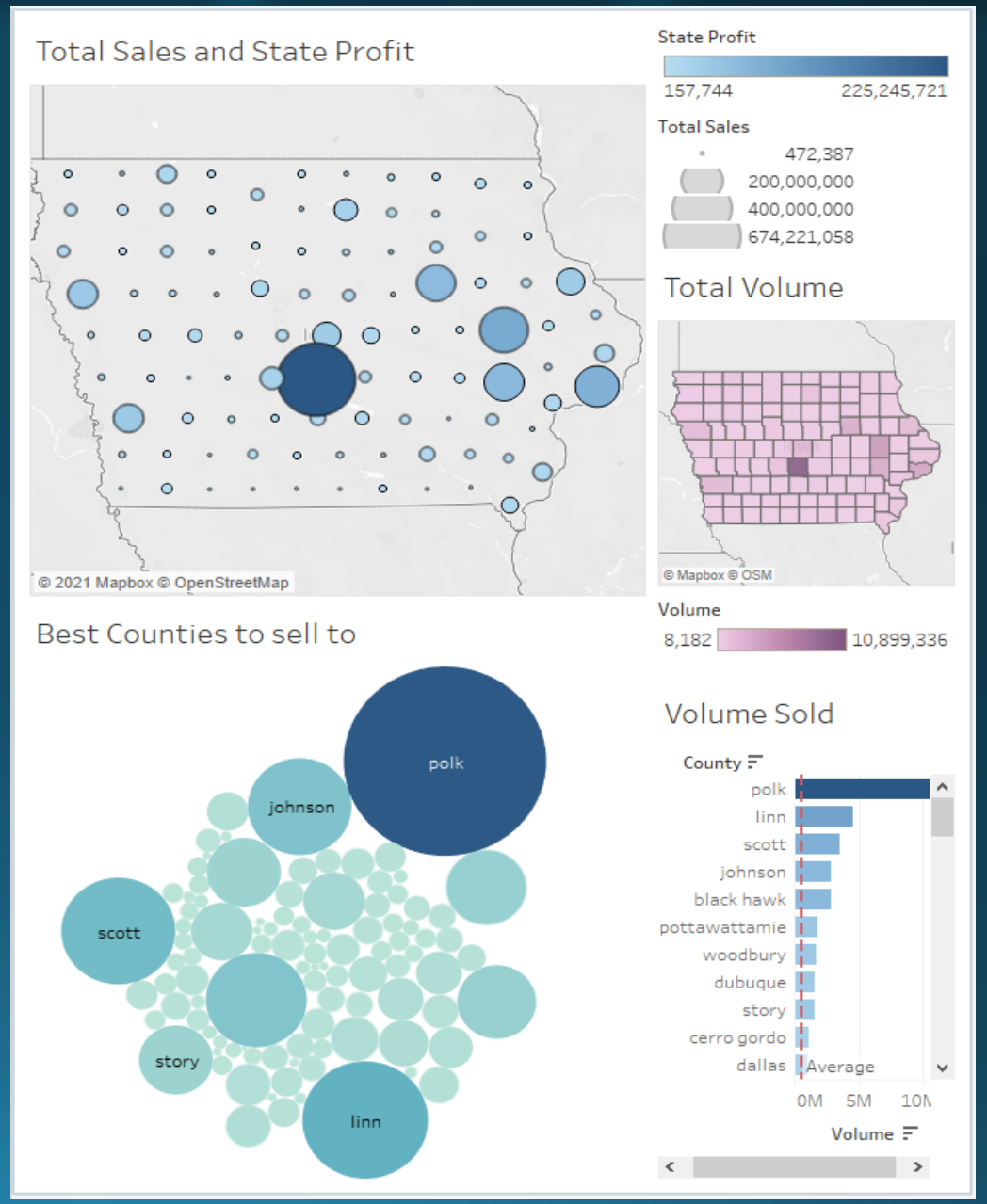




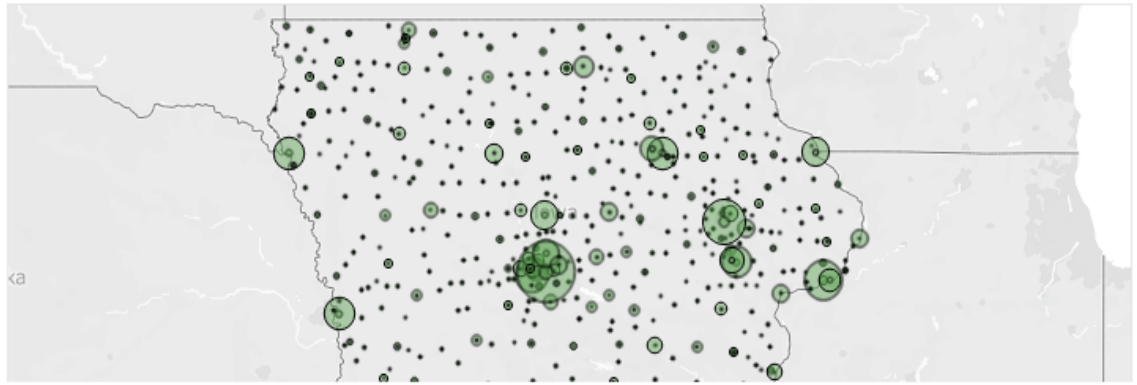
THE IMPACT OF COVID-19 ON THE ALCOHOLIC BEVERAGES INDUSTRY

We used Tableau to create two dashboard to help us visualize the data.

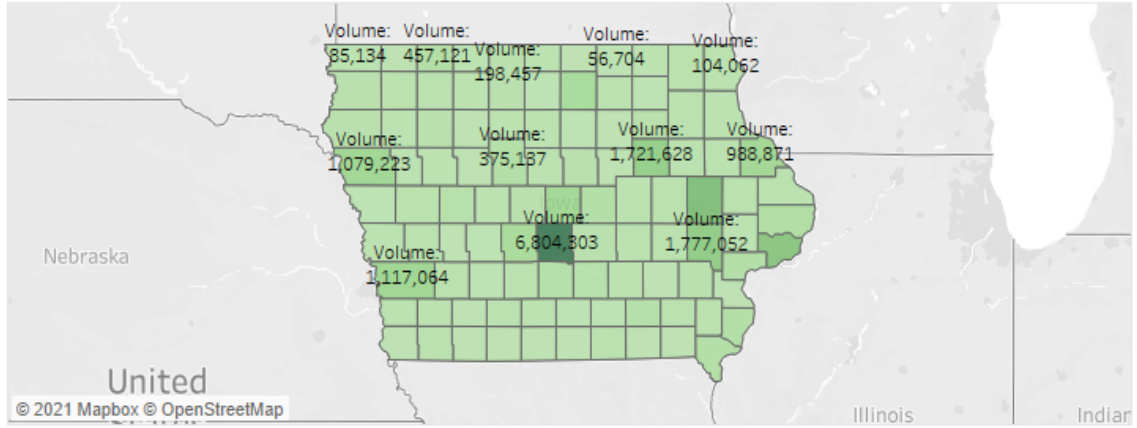
This is the link to the Tableau Public work: https://public.tableau.com/views/QuickEDA/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link



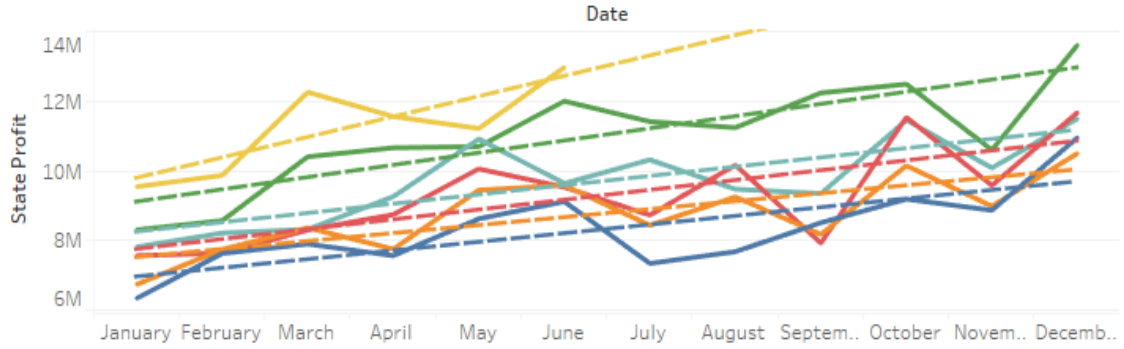
Sales and Volume by cities



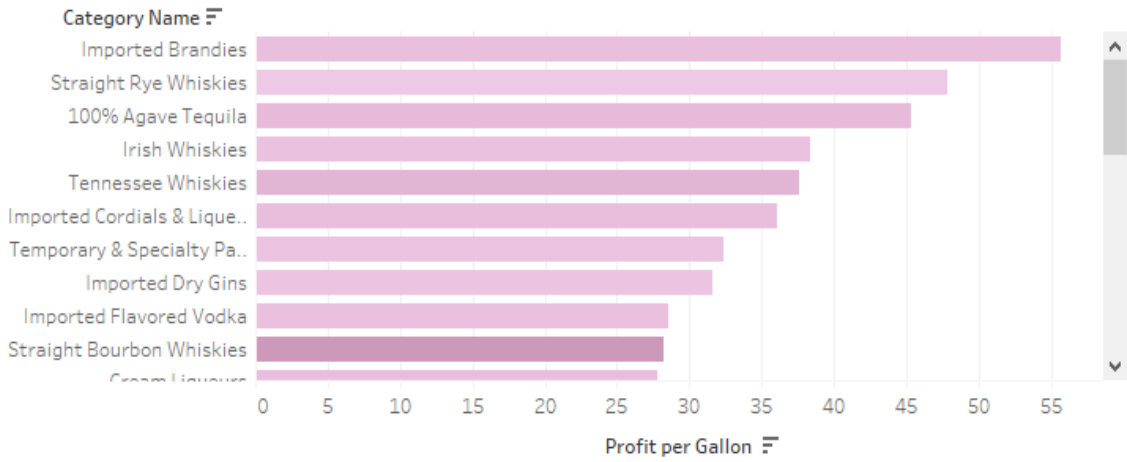
State profit is highly correlated with Sales and Volume



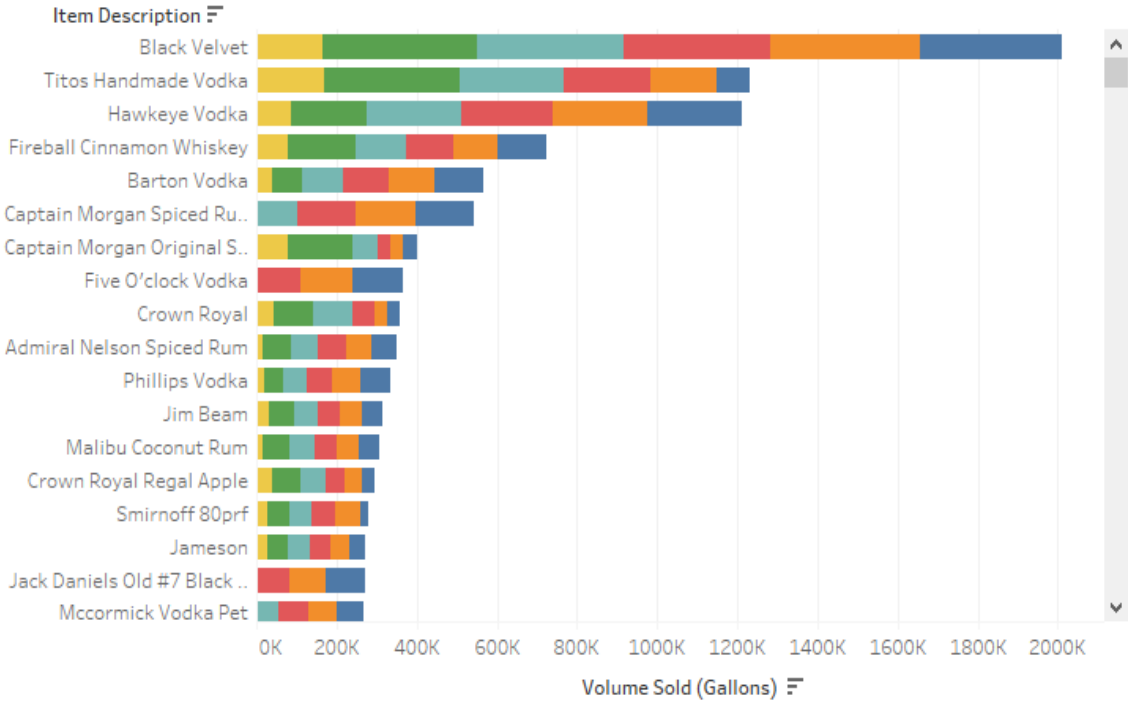
While 2020 was a good year for the industry, 2021 is expected to be better

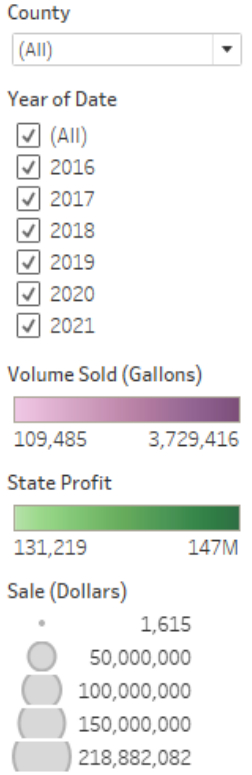


From categories who sold over 100,000 Gallons, the Imported Brandies made the most profit per Gallon

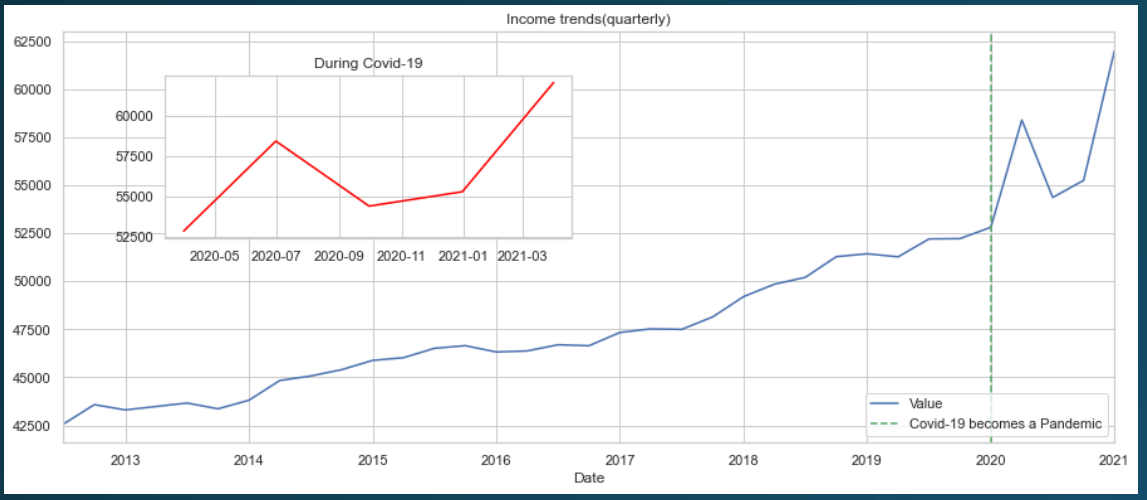


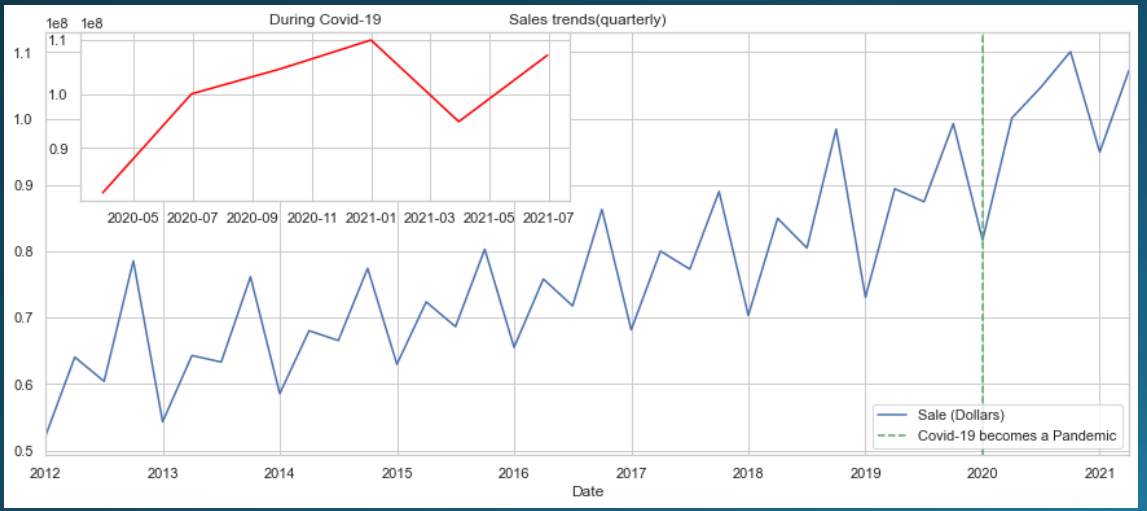
Top Items Sold per year





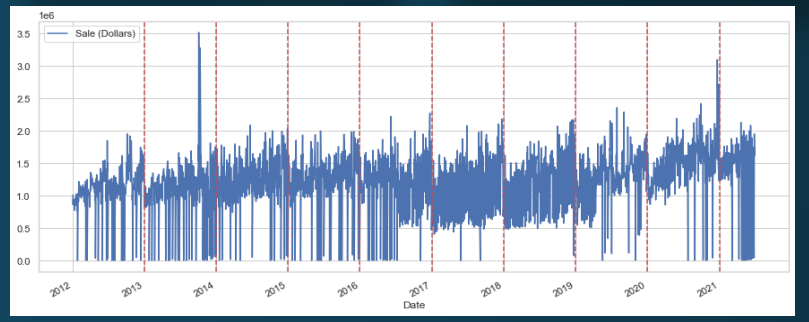
THE IMPACT OF COVID-19 ON THE ALCOHOLIC BEVERAGES INDUSTRY

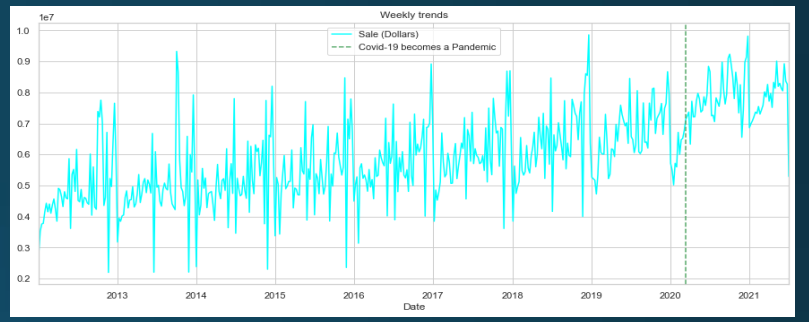


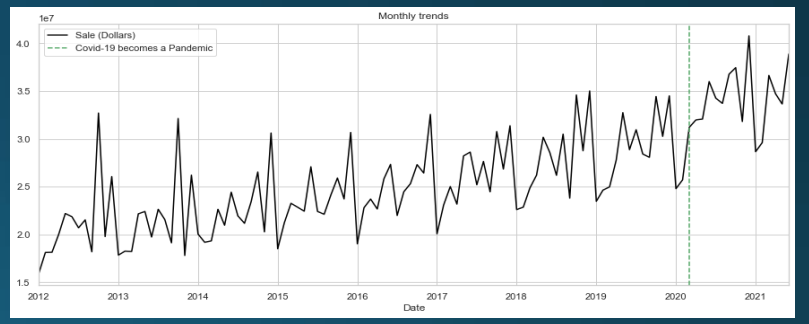


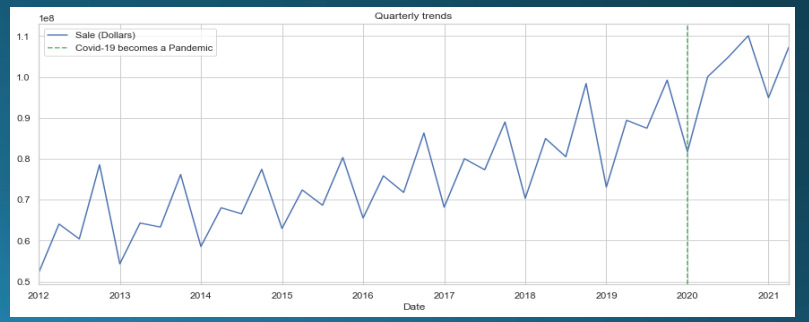
We also aggregated our total Sales feature on days, weeks, month and quarters to visualize trends and seasonality.

To see if the trends in Sales might be dependent on any other than the Pandemic influence, we got some additional data, the quarterly income and the population growth.







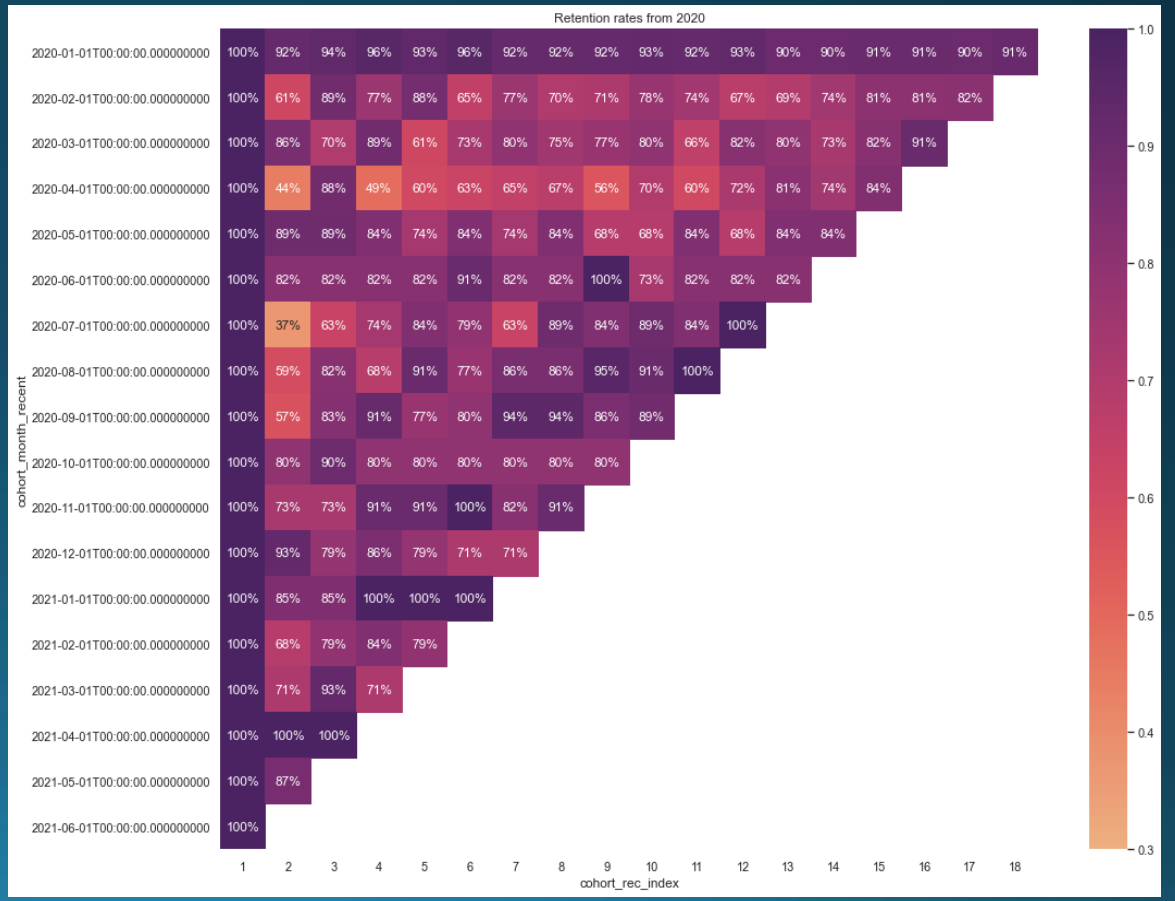


PRE-PROCCESSING – COHORT ANALYSIS

We began our analysis with a customer cohort Analysis. More precisely with a Store retention table, for data since 2020.

We can see that, unlike the retail industry where the retention rates are substantially lower, in this case the stores that are buying from the state department are more prone to order again every month.

We can see that for the months of July and August 2020, all the stores that made their first purchase in these 2-month ended up ordering again in June of 2021.



PRE-PROCCESSING – CUSTOMER SEGMENTATION WITH RFM

This is a behavioral customer segmentation based on three metrics:

- Recency(R)
- Frequency(F)
- Monetary Value(M)

The process of percentile grouping involves:

- Sorting the customers based on each of the 3 metrics
- Breaking customers into a pre-defined number of groups of equal size
- Assign a label to each group

Building RFM segment and RFM score

- For the RFM segment we will just concatenate the 3 metrics.
- For the RFM score we will add the 3 metrics.

	Frequency	MonetaryValue	Recency
Store Name			
'Da Booze Barn / West Bend	82	204930.02	0
10th Hole Inn & Suite / Gift Shop	13	16221.98	2
16th Ave BP / Cedar Rapids	11	26917.82	37
1st Ave BP / Cedar Rapids	12	20904.70	51
1st Stop Beverage Shop	235	1851350.37	5
...	...	...	...
Z's Quickbreak	177	251324.69	1780
Zapf's Pronto Market	207	316601.53	467
goPuff / Ames	79	148069.68	6
goPuff / Iowa City	47	185477.64	1
k food mart / Monticello	1	3758.37	592

2506 rows × 3 columns

	Frequency	MonetaryValue	Recency	R	F	M
Store Name						
'Da Booze Barn / West Bend	82	204930.02	0	4	2	2
10th Hole Inn & Suite / Gift Shop	13	16221.98	2	4	1	1
16th Ave BP / Cedar Rapids	11	26917.82	37	2	1	1
1st Ave BP / Cedar Rapids	12	20904.70	51	2	1	1
1st Stop Beverage Shop	235	1851350.37	5	3	3	4
...	...	...	...	...	...	...
Z's Quickbreak	177	251324.69	1780	1	3	2
Zapf's Pronto Market	207	316601.53	467	1	3	3
goPuff / Ames	79	148069.68	6	3	2	2
goPuff / Iowa City	47	185477.64	1	4	2	2
k food mart / Monticello	1	3758.37	592	1	1	1

	Frequency	MonetaryValue	Recency	R	F	M	RFM_Segment	RFM_Score
Store Name								
'Da Booze Barn / West Bend	82	204930.02	0	4	2	2	422	8
10th Hole Inn & Suite / Gift Shop	13	16221.98	2	4	1	1	411	6
16th Ave BP / Cedar Rapids	11	26917.82	37	2	1	1	211	4
1st Ave BP / Cedar Rapids	12	20904.70	51	2	1	1	211	4
1st Stop Beverage Shop	235	1851350.37	5	3	3	4	334	10
...	...	...	...	...	...	...	...	...
Z's Quickbreak	177	251324.69	1780	1	3	2	132	6
Zapf's Pronto Market	207	316601.53	467	1	3	3	133	7
goPuff / Ames	79	148069.68	6	3	2	2	322	7
goPuff / Iowa City	47	185477.64	1	4	2	2	422	8
k food mart / Monticello	1	3758.37	592	1	1	1	111	3

CUSTOMER SEGMENTATION WITH RFM AND UNSUPERVISED LEARNING CLUSTERING

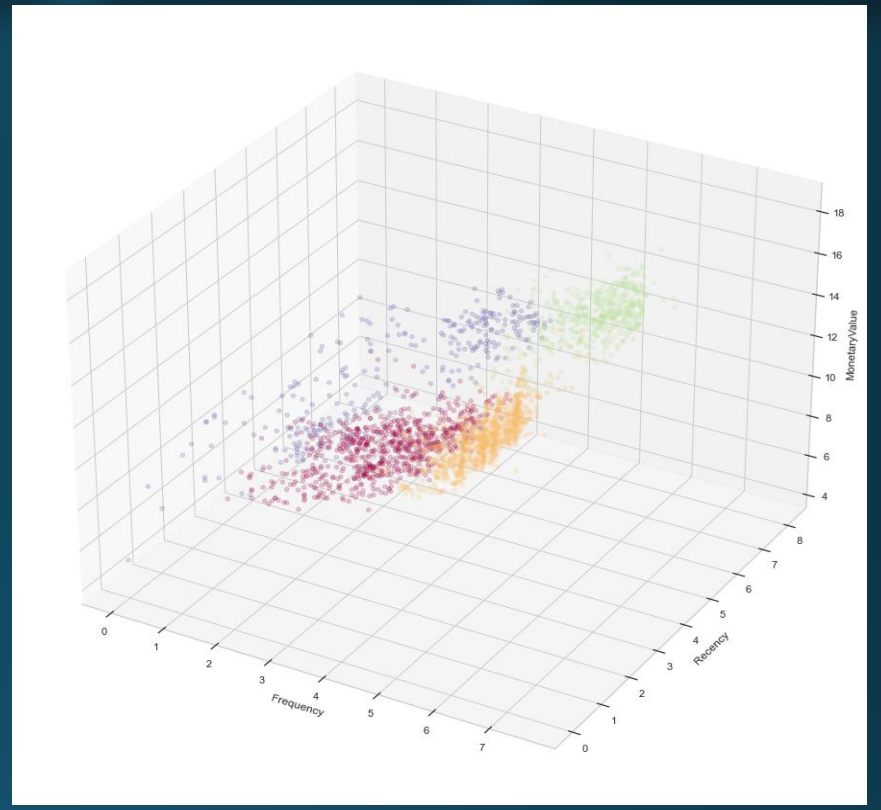
Using the RFM values we can furthermore create a clustering of similar Stores using Unsupervised Learning

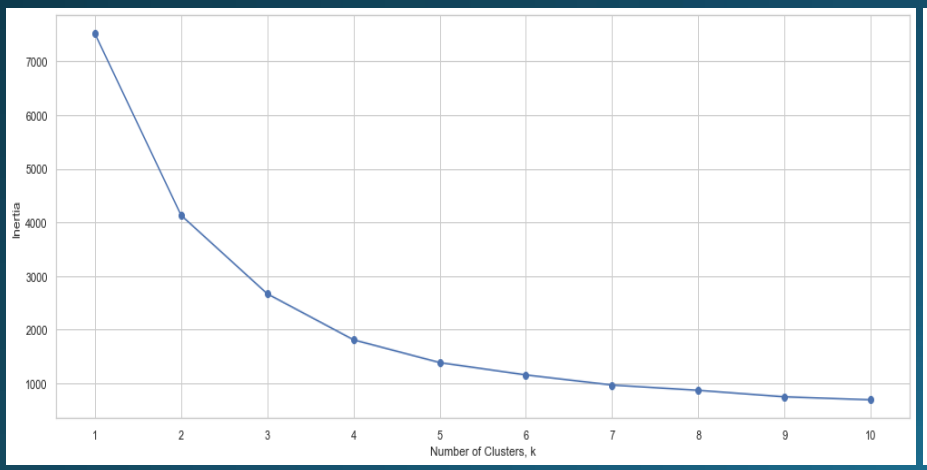
As methods of choosing the optimal K clusters we will use:

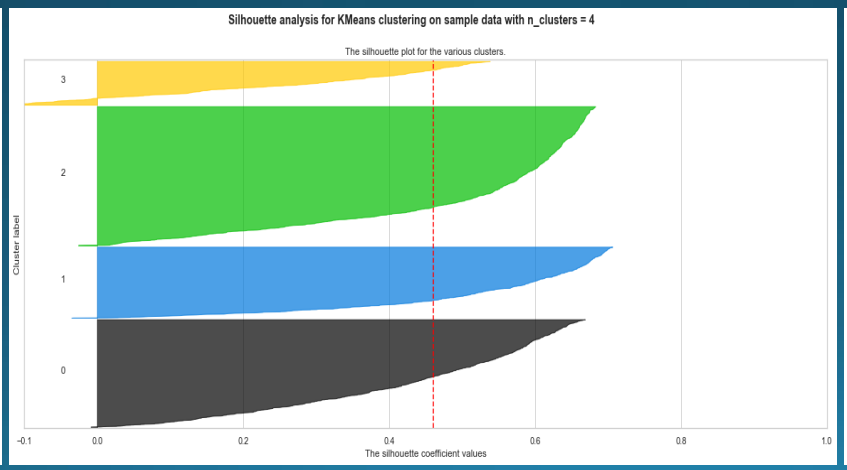
- Elbow method
- Silhouette scores and plots

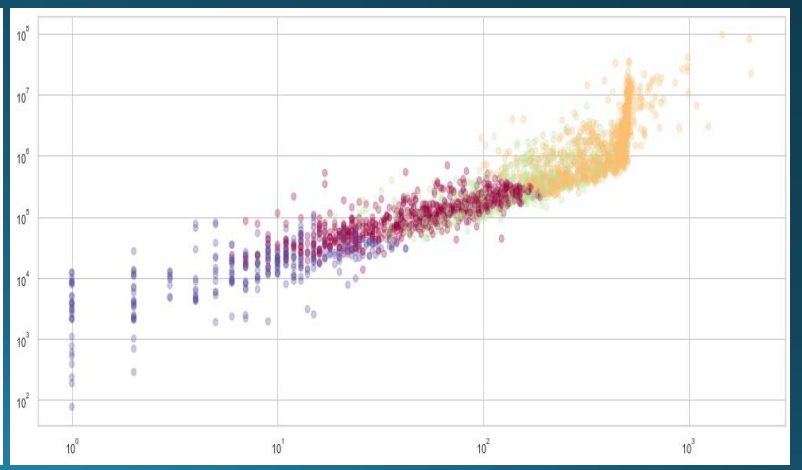
We decided on 4 clusters.

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
cust_segmentation				
0	7.7	67.1	137658.4	745
1	5.3	390.9	2705848.9	963
2	1507.5	166.5	504989.9	495
3	1166.0	11.5	20683.2	303









TIME SERIES ANALYSIS

The main target and goal of our project is using time series analysis and predictions to forecast Iowa's Dept. of Commerce profits from spirits Sales. We plan to follow a series tools to conduct the analysis, transformation and modeling of our data:

Time series decomposition

- Trend
- Seasonality
- Noise(Residuals)

Stationarity

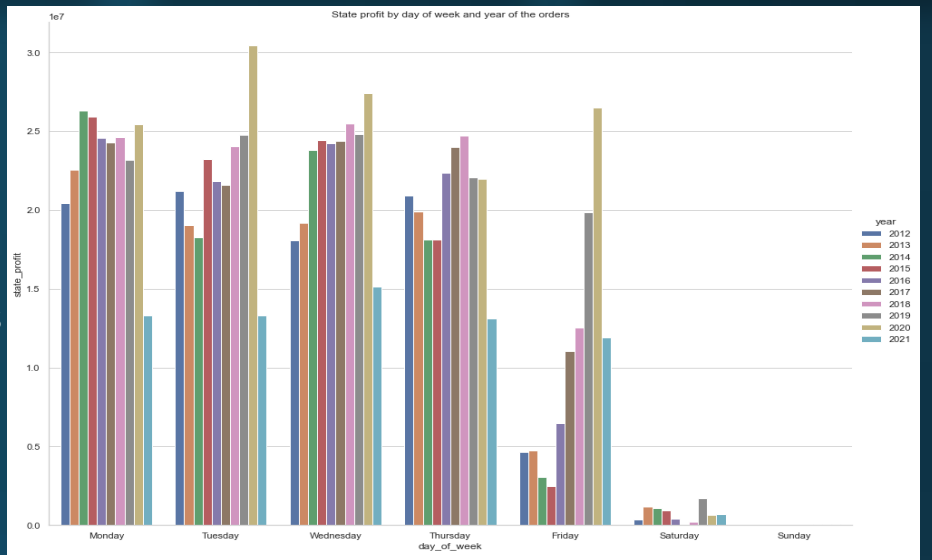
- AC and PAC plots
- Dickey-Fuller test

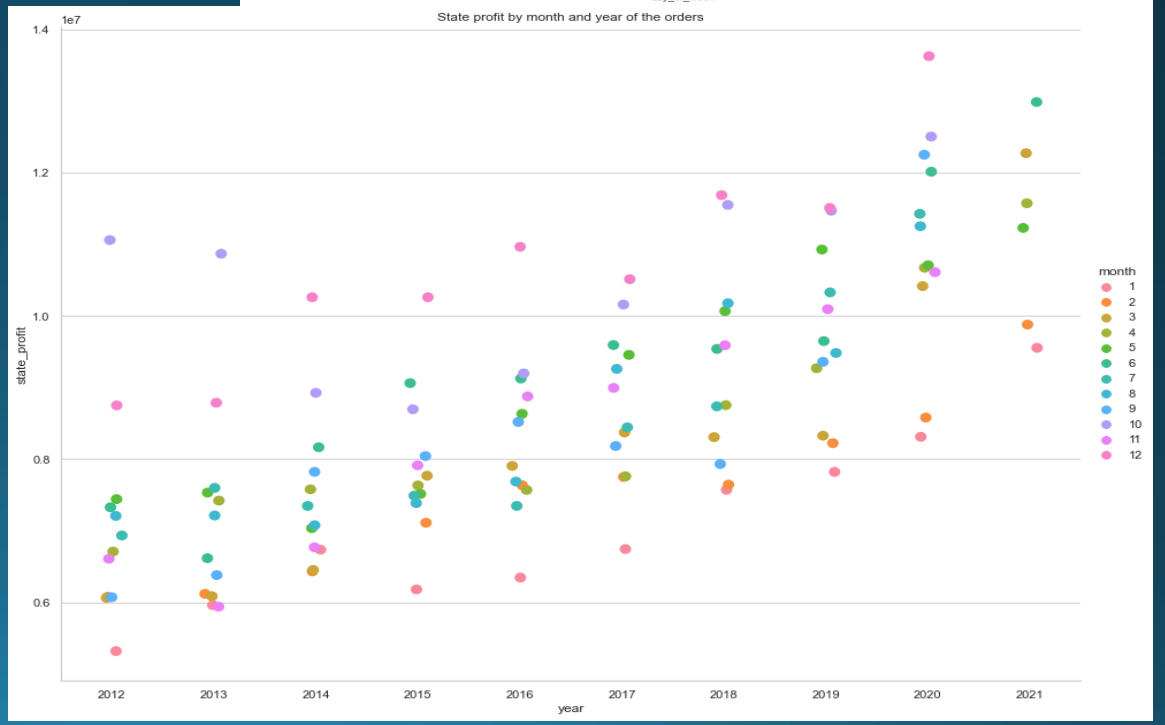
Choosing the metrics

- RMSE
- MAE

Models Tested

- AutoRegressive model (AR 1)
- BIC(Bayesian Information Criterion) adjusted AR model
- Moving Average Model(MA)
- Autoregressive Moving Average (ARMA)
- Seasonal autoregressive integrated moving average (SARIMAX)



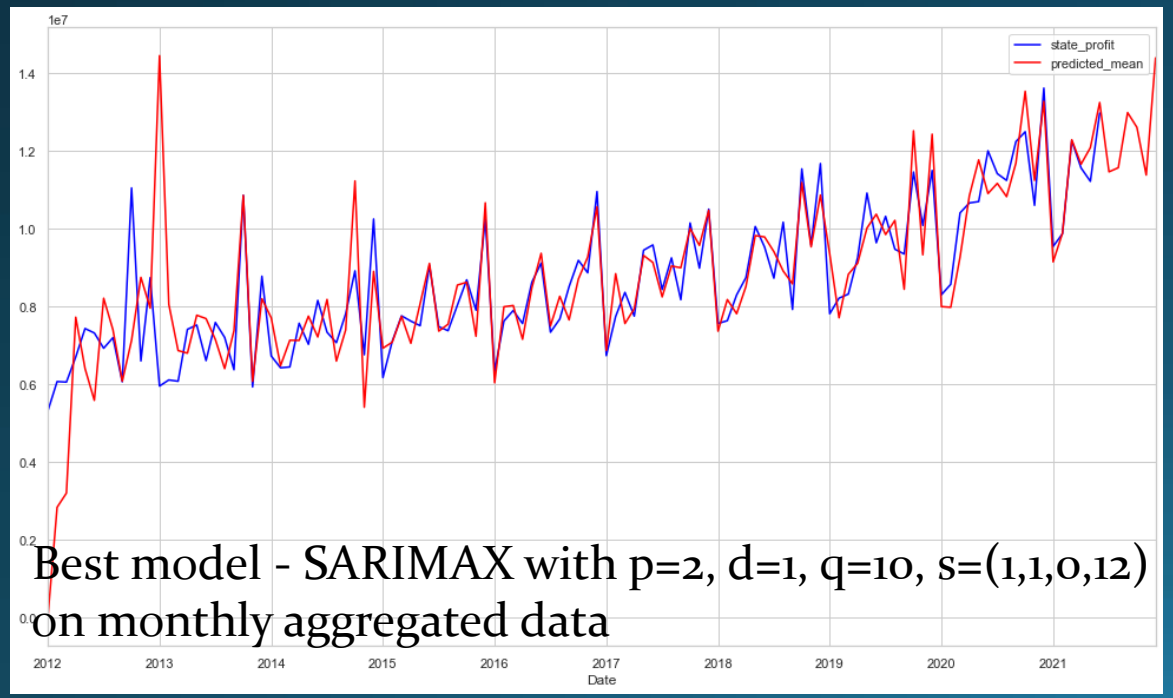


MODELING

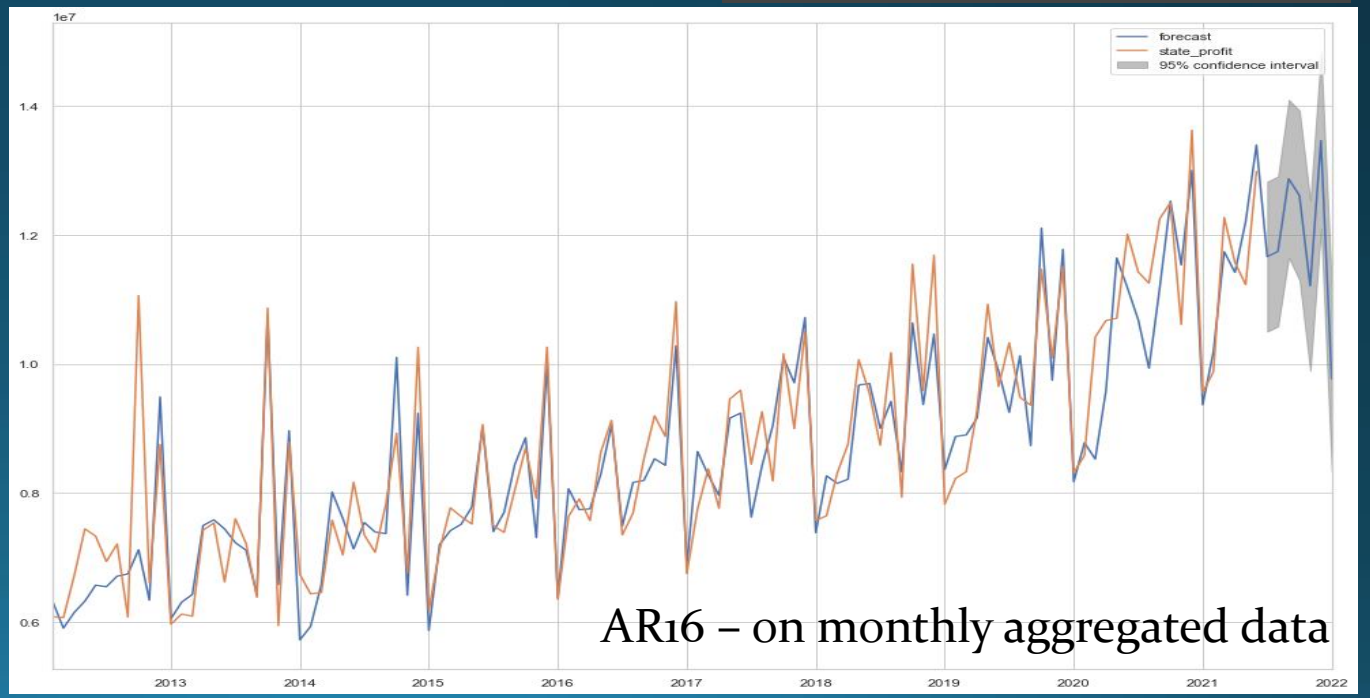
We used the RMSE as the metric to classify the models. We also used the multitud
of models with 2 different data: daily aggregations and Monthly aggregations.

This are the top 20 models with the highest scores on the Validation set.

	data_agg	model	params	rmse	mae
139	month_agg	SARIMAX	p=2, d=1, q=10, seasonal_order=(1,1,0,12)	3330.86	3330.86
136	month_agg	SARIMAX	p=2, d=0, q=10, seasonal_order=(1,1,0,12)	42648.82	42648.82
88	month_agg	ARIMA	p=0, d=2, q=2	52372.53	52372.53
112	month_agg	ARIMA	p=0, d=2, q=2	52372.53	52372.53
77	month_agg	AR	p =16	61794.20	61794.20
128	month_agg	SARIMAX	p=1, d=0, q=8, seasonal_order=(1,1,0,12)	61840.06	61840.06
59	daily_agg	SARIMAX	p=1, d=0, q=10, seasonal_order=(1,1,0,7)	69173.20	56291.66
64	daily_agg	SARIMAX	p=2, d=0, q=9, seasonal_order=(1,1,0,7)	69176.27	55936.69
65	daily_agg	SARIMAX	p=2, d=0, q=10, seasonal_order=(1,1,0,7)	69359.03	56583.06
57	daily_agg	SARIMAX	p=1, d=0, q=8, seasonal_order=(1,1,0,7)	69399.36	55900.98
58	daily_agg	SARIMAX	p=1, d=0, q=9, seasonal_order=(1,1,0,7)	69756.80	55610.50
70	daily_agg	SARIMAX	p=6, d=0, q=9, seasonal_order=(1,1,0,7)	70033.00	58067.88
69	daily_agg	SARIMAX	p=6, d=0, q=8, seasonal_order=(1,1,0,7)	70210.09	58422.77
63	daily_agg	SARIMAX	p=2, d=0, q=8, seasonal_order=(1,1,0,7)	70225.39	55741.00
71	daily_agg	SARIMAX	p=6, d=0, q=10, seasonal_order=(1,1,0,7)	70282.11	58368.37
13	daily_agg	ARIMA	p=8, d=0, q=9	70381.38	57362.76
53	daily_agg	ARIMA	p=8, d=0, q=9	70381.38	57362.76
11	daily_agg	ARMA	p=8, q=9	70381.38	57362.76
12	daily_agg	ARMA	p=8, q=10	70393.36	58973.98
10	daily_agg	ARMA	p=8, q=8	70928.67	58771.67



Best model - SARIMAX with p=2, d=1, q=10, s=(1,1,0,12) on monthly aggregated data



AR16 – on monthly aggregated data

FUTURE WORK

For future work we still want to explore the 2 identified Business Problems – helping a store owner to pick a new location for a store so that he will maximize sales, and EDA on Storage Capacity.

Another aspect is that the data will become available for future testing and prediction.
Every month new data will be uploaded so we can use that to test our models predictions accuracies.

RECOMMENDATIONS

Ptofit per month is a very important metric for any company in every industry. Predicting the profit evolution allows companies to plan and make Business Operations decisions for the foreseeable future.

Future Profit per month is also an indicator of the value a company is having, and the potential growth. This will determine the need to invest in possible expansions:

- Increase the number or quantities of available product to sell, along with investing in Marketing to attract new customers.
- Expand to new locations within the same state on to other states.

The Covid-19 pandemic EDA and the RFM Customer Segmentations that we conducted are great tools for the Executive level to understand trends and customer base so they can make informed strategical business decisions.

Thank you!

Questions?