# Predicting Liquor Sales Profit and post-Pandemic Business Analytics

Author: Razvan Nelepcu | Mentor: Dhiraj Khanna | August 2021

# Introduction

## 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.

On the other hand, live does go on, and businesses must continue their business operations with the tools and possibilities they have available. During my journey in Business Management, I observed a lot of Executives making decisions on their intuition, even if data was available in some hard Drives right next to them. Data that could have been used to observe patterns, trends, or predictions. Data that can be used not just by the Executive level decision-making, but by Operations, Sales, Marketing, Financial or HR services as well. If there is data, we believe insights can be drawn from it.

That is the context in which we decide to use the spirits sales in Iowa since 2012 to present to identify possible Business Problems and to resolve them with Data Science.

## PROBLEMS IDENTIFICATION

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.

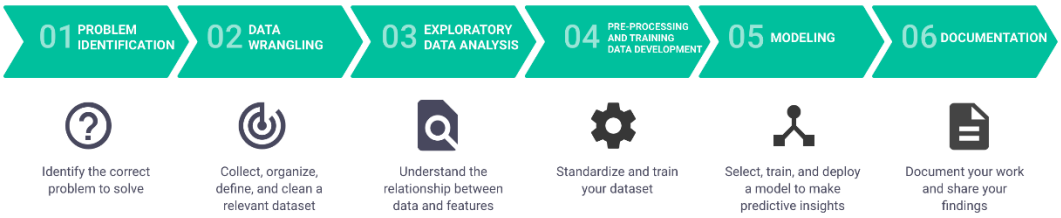
For future work we also identified 2 other Business Problems using the same data:

- 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.

Our approach was to tackle these problems in different part of the project for more clarity as well as to be able to use the output from some of them as input to those problems solved in the Modeling part.

### DELIVERABLE:

- All code - Jupiter Notebooks
  - Data wrangling
  - Exploratory Data Analysis
  - Pre-processing and training data development
  - Modeling
- Final Report of the Project
- Presentation Slide Deck




# Data Wrangling

### 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.

- **Invoice/Item Number** - Concatenated invoice and line number associated with the liquor order. This provides a unique identifier for the individual liquor products included in the store order
- **Date** - Date of order
- **Store Number** - Unique number assigned to the store who ordered the liquor
- **Store Name** - Name of store who ordered the liquor
- **Address** - Address of store who ordered the liquor
- **City** - City where the store who ordered the liquor is located
- **Zip Code** - Zip code where the store who ordered the liquor is located
- **Store Location** - Location of store who ordered the liquor. The Address, City, State and Zip Code are geocoded to provide geographic coordinates. Accuracy of geocoding is dependent on how well the address is interpreted and the completeness of the reference data used.
- **County Number** - Iowa county number for the county where store who ordered the liquor is located
- **County** - County where the store who ordered the liquor is located
- **Category** - Category code associated with the liquor ordered
- **Category Name** - Category of the liquor ordered.
- **Vendor Number** - The vendor number of the company for the brand of liquor ordered
- **Vendor Name** - The vendor name of the company for the brand of liquor ordered
- **Item Number** - Item number for the individual liquor product ordered
- **Item Description** - Description of the individual liquor product ordered
- **Pack** - The number of bottles in a case for the liquor ordered
- **Bottle Volume (ml)** - Volume of each liquor bottle ordered in milliliters
- **State Bottle Cost** - The amount that Alcoholic Beverages Division paid for each bottle of liquor ordered
- **State Bottle Retail** - The amount the store paid for each bottle of liquor ordered
- **Bottles Sold** - The number of bottles of liquor ordered by the store
- **Sale (Dollars)** - Total cost of liquor order (number of bottles multiplied by the state bottle retail)
- **Volume Sold (Liters)** - Total volume of liquor ordered in liters. (i.e. (Bottle Volume (ml) x Bottles Sold)/1,000)
- **Volume Sold (Gallons)** - Total volume of liquor ordered in gallons. (i.e. (Bottle Volume (ml) x Bottles Sold)/3785.411784)

In addition, we utilized other datasets regarding Demographics or per Capita Personal Income in the State of Iowa, available on the website mentioned above.

## SHOW ME THE DATA

One of the first issues we will encountered was dealing with the big number of records, and for this we used the Dask library to reduce the computational time.

As we can see, we had several features that were referting to the same unique entity, so before proceeding any futhers we had to take steps to ensure that we check our Data for quality.

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.

From the 2 describe methods used above we can conclude the following:

- there were 2772 unique Store Names, with Store Numbers between 2106 and 9946.
- there were 201 unique Counties, with County Numbers between 1 and 99. This is somethis that we explored further because the number of unique counties was double than the number of the codes associated with them.
- there were 527 unique Vendor Names, with Vendor Numbers between 10 and 987.
- the Pack feature ranges from 1 to 336. Again, the maximim value was quite high, so we explored this.
- the Bottle Volume ranged from 0 to 378,000 liters. We verified both the maximum and minimum of these values.

- the State Bottle Cost, State Bottle Retail, Bottles Sold, Sale (Dollars), Volume Sold (Liters) have had minimims at o(Zero) and unusual high values. We checked these outliers as well.

Each of these aspects was further analyzed in the Data Quality steps and in the Outliers Exploration.

## STORE NUMBER AND NAME

Using drop_duplicates and groupby method on our Dask DataFrame we were able to find that We had **2903** unique pairings of Store Name and Store Number, while we had only **2622** unique store numbers and **2722** unique store names.

```
mask = uniques_store.groupby('Store Number').count()
mask = mask.sort_values('Store Name', ascending = False)
duplicates = mask[mask['Store Name']>1].reset_index()
duplicates
```

| | Store Number | Store Name |
|---|---|---|
| 0 | 2663.0 | 4 |
| 1 | 4378.0 | 4 |
| 2 | 4152.0 | 3 |
| 3 | 5405.0 | 3 |
| 4 | 4824.0 | 3 |
| ... | ... | ... |
| 256 | 2501.0 | 2 |
| 257 | 2522.0 | 2 |
| 258 | 6171.0 | 2 |
| 259 | 2539.0 | 2 |
| 260 | 2591.0 | 2 |

261 rows × 2 columns

We had 261 different store that have the same Store number, but diferent Store Names.

Instead of going though all 261 Store Numbers to identify which one to keep, we started by looking at those combinations of Store Name and Store Number from our duplicates that have below 10 records in our sales data.

Then we had a look at the unique Name-Number duplicates and sorted them by Store Number.

It seemed that the error was in maintaining the same Store Name thought all the records, as we can see in the first 3 rows from the picture.

We used the Store Names that have the most amount of records in our data as the main one, and we replaced the others with that value. For this we used a groupby, merge and then a column replacement.

| | Store Number | Store Name | Invoice/Item Number |
|---|---|---|---|
| 3 | 2178.0 | Double D Liquor Store | 7560 |
| 1983 | 2178.0 | Double "D" Liquor Store | 12465 |
| 2451 | 2178.0 | "Double ""D"" Liquor Store" | 3265 |
| 17 | 2501.0 | Hy-Vee #2 / Ames | 49213 |
| 18 | 2501.0 | Hy-vee #2 / Ames | 54729 |
| ... | ... | ... | ... |
| 2897 | 6171.0 | Speedee Mart 1515 / Council Bluffs | 602 |
| 1623 | 9041.0 | S&B Farms Distillery | 265 |
| 1624 | 9041.0 | S&B Farmstead Distillery | 192 |
| 2499 | 9911.0 | Southern Glazers Wine & Spirits of Iowa | 980 |
| 2430 | 9911.0 | North American Spirits | 1005 |

## OTHER DATA CLEANING

We also verified the County and County Number features. While we had 201 disting County Names, we only had 99 distinct numbers. Doing a quick Wikipedia check we saw there are 99 counties in Iowa. The main issue with this was that some counties were writen in uppercase while others in lower or sentence case. After bringing all data to lower case, we still had 104 distinct counties and just 99 real ones.

We will upload a short csv file containing a list of all 99 counties from IOWA

```
iowa_counties = pd.read_csv('E:\Springboard/GitHub/Iowa_spirits_sales/data/raw/iowa_states.csv')
iowa_counties = iowa_counties['County'].str.lower()
iowa_counties
```

```
0          adair
1          adams
2      allamakee
3      appanoose
4        audubon
         ...
94     winnebago
95    winneshiek
96      woodbury
97         worth
98        wright
Name: County, Length: 99, dtype: object
```

```
wrong_c = set(counties) - set(iowa_counties)
wrong_c
```

```
{'buena vist', 'cerro gord', 'el paso', nan, "o'brien", 'obrien', 'pottawatta'}
```

We used the data from Wikipedia to create a dataset and with it we identified those values that were written wrong, or as it was the case of El Paso, counties that were not in Iowa.

## EXPLORING OUTLIERS

Using the describe methods applied on our numerical and categorical data we were able to identify the outlliers.

There were some extreme values(min or max) that we wanted to inspect:

- the Pack feature ranges from 1 to 336. We will check the max values.
- the Bottle Volume ranges from 0 to 378,000 liters. We will have to verify both the max and min values.
- the State Bottle Cost, State Bottle Retail, Bottles Sold, Sale (Dollars), Volume Sold (Liters) have all minimims at 0 and unusual high values. We will check the max and min value

It seems that there is type of beverage, Members Mark Silver Tequila that has 336 bottles in a pack. While it probably is not a pack, but more like a box or container, the data found in the records was valid and we kept it.

Checking the Bottle Volume (ml) max values, we see that 18 cases Smirnoff 1.75L/18 cases Captain 1.75L is recorded as having 378000ml, which is 378 liters. Dividing the 378 liters to the bottle size, 1.75, and then to the number of cases, 18, we see that each case has 12 bottles each. So a more accurate description of the item would be: 18 cases with 12 bottles each of Smirnoff 1.75L. But data is valid so was kept for future analysis.

## DEALING WITH MISSING DATA

Knowing that we have a total of 21,641,155 records in our data, let's explore how many missing values we have.

```
21641155 - df.count()
```

```
Invoice/Item Number            0
Date                           0
Store Number                   0
Address                    79992
City                       79991
Zip Code                   80036
Store Location           2075423
County Number             156794
County                    156794
Category                   16974
Category Name              25040
Vendor Number                  9
Vendor Name                    7
Item Number                    0
Item Description               0
Pack                           0
Bottle Volume (ml)             0
State Bottle Cost             10
State Bottle Retail           10
Bottles Sold                   0
Sale (Dollars)                10
Volume Sold (Liters)           0
Volume Sold (Gallons)          0
Store Name                     0
```

As we can see a lot of the missing data come from the Store information: Address, city, Zip Code, Store geolocation, County or County Number.

On the other hand, we have almost no missing data when it comes to the date, store number or name, items sold or transaction details.

We identified some good methods that would allow us to better fill the missing data such as:

- Zip Code, or Address, or Location of a Store based on its number and replacing the NaNs with the corect values. But this gave a Memory Error, so we will have this Data Cleaning process in mind for future work.
- Checking the store name to extract city for those records that are missing the city value

- example store name: Sam's Club 6979 / Ankeny. In this case, Alkeny is the city, and we can do a split on the store name to extract and update the city.

But for the purpose of this project and considering the business problems in mind and the limitations of the local machine used in this project, we completed simple subtitutions for the remaining missing values.

We sorted our data's values by the missing features, making sure those coming from the same store were grouped up one after another, and making sure the NaNs are at the end. Then we used a forward fill to fill the mising data with data from the records that contained usefull data. Obviously,there are caveats to this method, but it is something that presumably gives good results.

```
df = df.sort_values(by = ['Store Number', 'Address', 'City', 'Zip Code', 'Store Location', 'County Number'], na_position='last')
```

```
df['Address'] = df['Address'].fillna(method = 'ffill')
```

```
df['City'] = df['City'].fillna(method = 'ffill')
```

```
df['Zip Code'] = df['Zip Code'].fillna(method = 'ffill')
```

```
df['Store Location'] = df['Store Location'].fillna(method = 'ffill')
```

```
df['County Number'] = df['County Number'].fillna(method = 'ffill')
```

```
df['County'] = df['County'].fillna(method = 'ffill')
```

```
21641155 - df.count()
```

```
Invoice/Item Number         0
Date                        0
Store Number                0
Address                     0
City                        0
Zip Code                    0
Store Location              0
County Number               0
County                      0
```

Finally, we removed some records that were missing information about the transaction and replaced the missing category names with 'Other'.

As a final step in our Data Wrangling, we conducted:

## 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.

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

For this step we will need the following fields:

1. Date - grouped by day, with values aggregated by Sum
2. Sale(Dollars)
3. Volume Sold(Gallons)
4. State Profit( calculated before aggregation as (State Bottle Retail - State Bottle Cost)* Bottles Sold )

5. Store Name
6. County
7. Vendor Name
8. City
9. Category Name
10. Item Description

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

In our pre-processing step we will conduct customer segmentation and we will use this as a feature in our Modeling. For this we will need the following features:

1. Date
2. Sale (Dollars)
3. Store Number
4. Store Name

Using aggregations will be create aditional features:

5. Recency
6. Frequency
7. Monetary Value

And as an output we will create score for each store and aggregated scores based on the 3 created features.

**Problem 3.** Predicting Profit for next month

The focus of our project will be to predict Profit for the State of Iowa department for the next month using time Series Analysis and Prediction.

The data used will be:

1. Date
2. state_profit( calculated before aggregation as (State Bottle Retail - State Bottle Cost)* Bottles Sold

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

**Problem 4.** Storage Capacity EDA

For the analysis on the State's Storage Capacity, we will be using windowing fucntions to determine the total volume of Liquor sold over a period of 7 days, 1 month, respectively 3 months. This will give us a clear idea of how much volume of Liquor the state has to store in its facilities, as well as how fast they have to resuply.

These inital Features will be used:

1. Date
2. Volume sold - per invoice

In our EDA we will be creating the following Features:

3. Weekly_vol
4. Monthly_vol
5. 3months_vol

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

For this step we will also import dome demographic data as well as income per capita data from Iowa.Gov.

As a first step we will have to choose one store/ store chain. Other features used in our process:We will merge the cust_segmentation created in the Pre-processing to the folowing existing features:

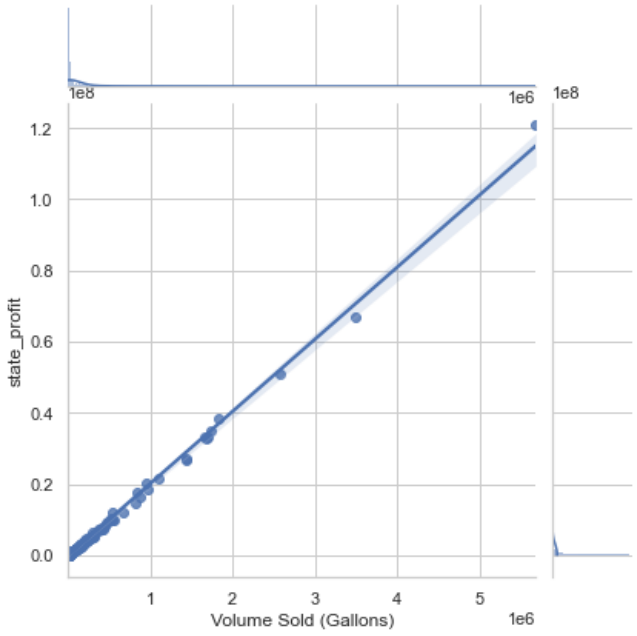
1. Store Location
2. Sales
3. Zip Code
4. Store Name
5. Store Number

We will need the following external feartures:

6. Population per County/ Zip Code
7. Income per County/ Zip Code

# Exploratory Data Analysis(EDA)

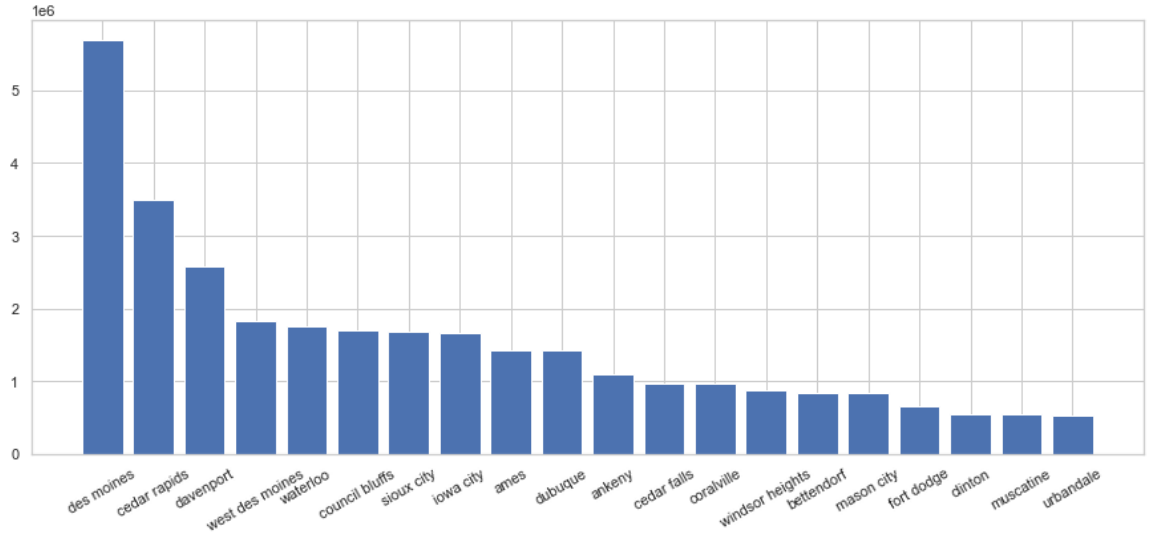
## UNIVARIATE, BIVARIATE AND MULTIVARIATE ANALYSIS




Here we can see our data aggregation on City. We can clearly see the correlation between Profit and the Volume Sold.

We can see that while most cities are averaging up to 1,000,000 Gallons, there are a few that are buying up to 5.5M Gallons.

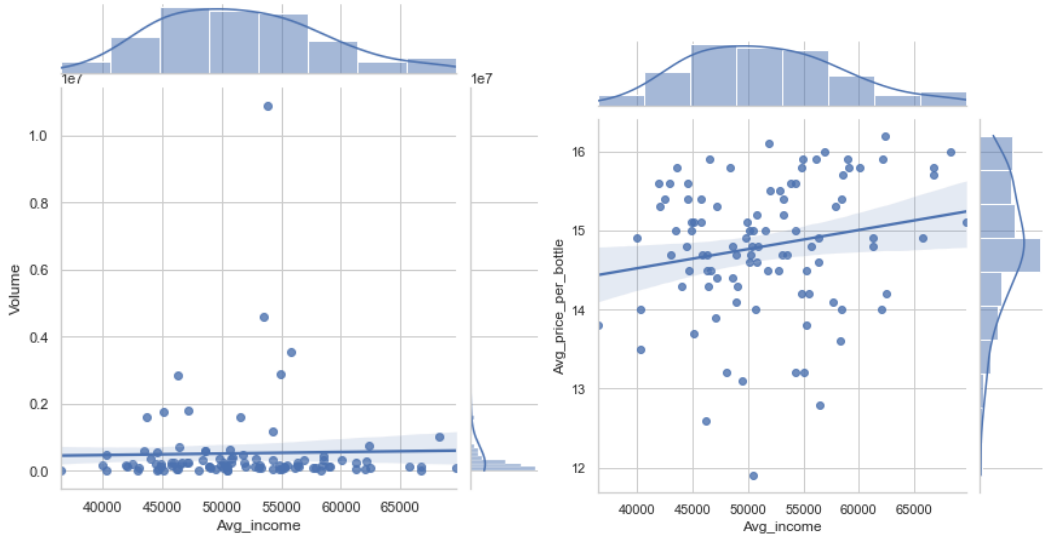
So next we explored our top 20 cities by volume bought.




## COUNTY EDA

To explore the County feature we imported some additional data from Iowa.gov: the **annual personal income per county.**

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

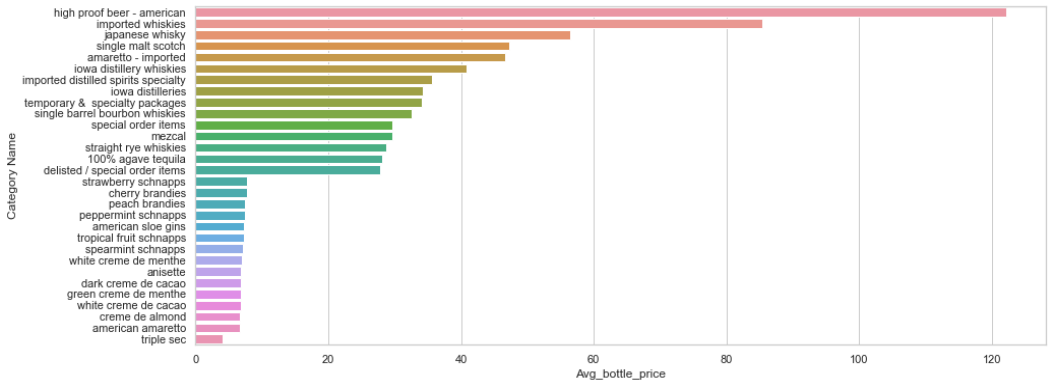



We can see that there are no correlation between how rich are the residents of a county and how much Volume of Spirits was sold in that county.

We can see as well that there is a slight correlation between how wealthy are the residents of a county and the average bottle cost sold in that county. So wealthier people are buying more expensive drinks.

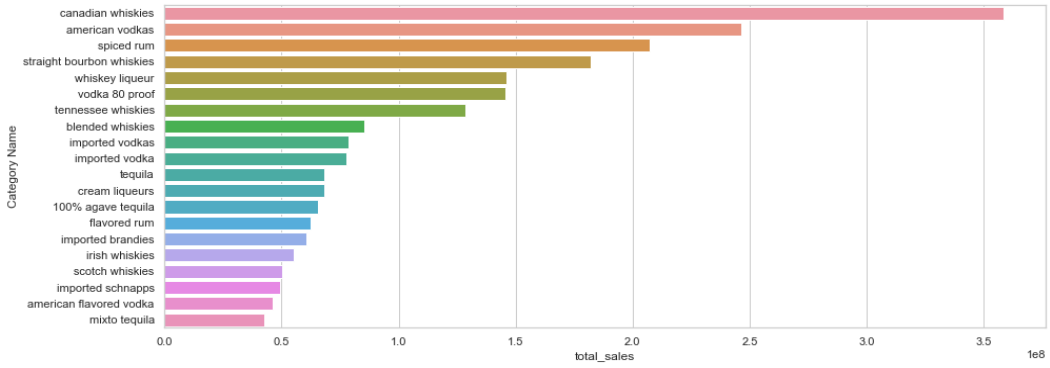
## SPIRITS CATEGORY EDA

We started by exploring the average bottle price per category.

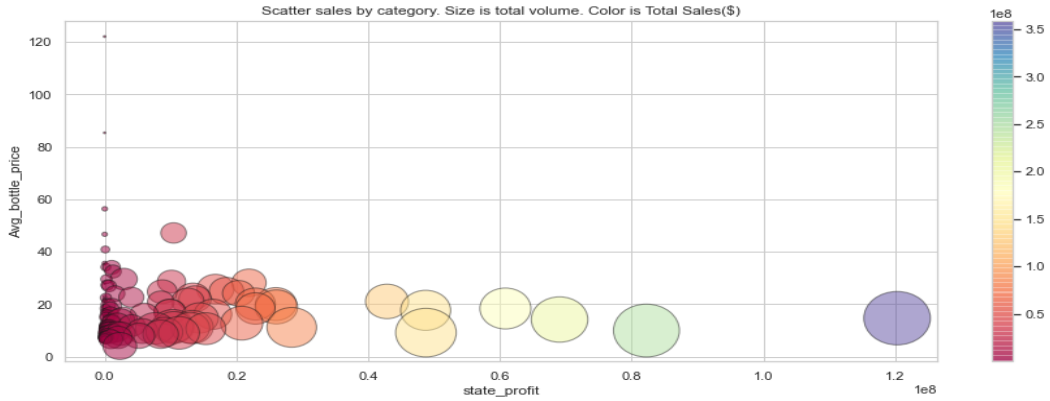



We can see that the High Proof Beer - American is the most expensive one with an average of over $120 per bottle, folowed by imported Wiskey and Japanese Wiskey. The least expensive one is the Triple Sec with $4.1 per bottle.

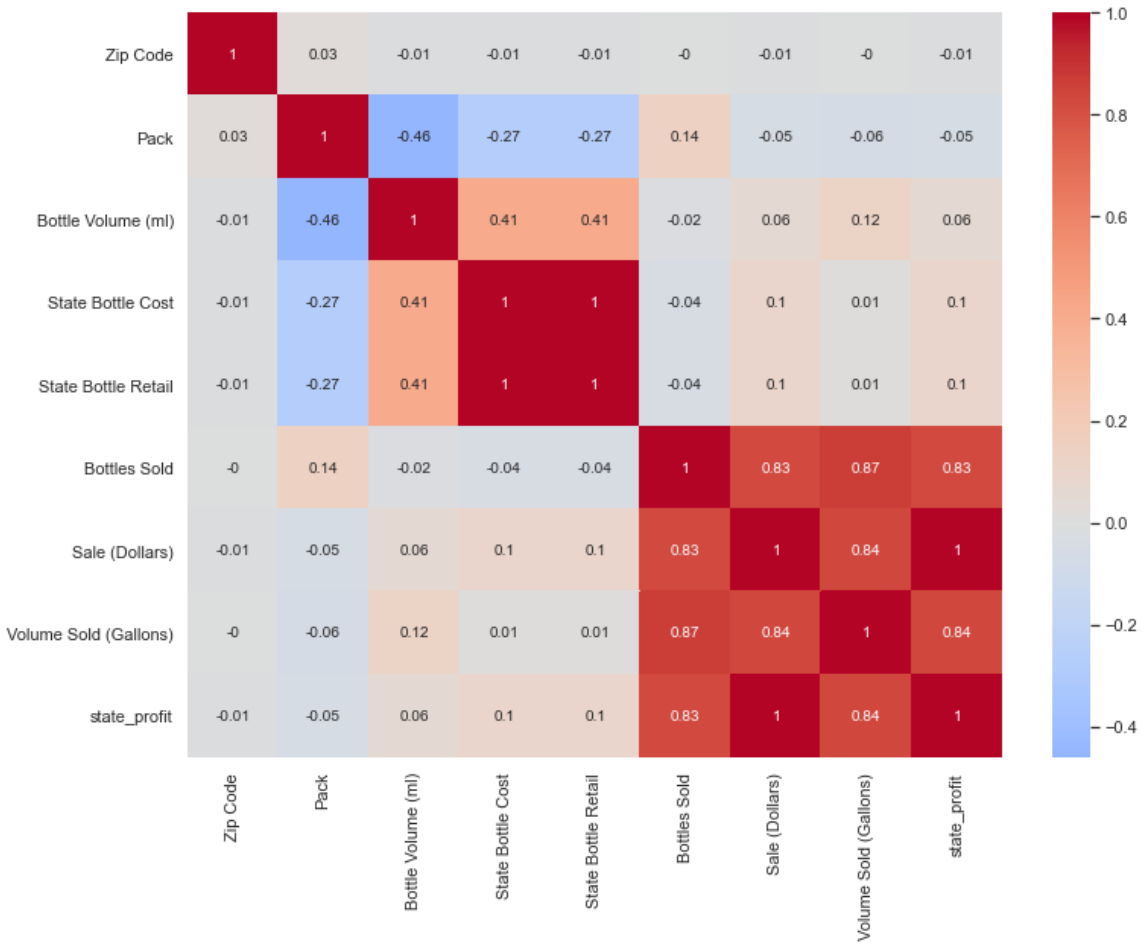
We also explored the most popular categories.




And finally, we explored a scatter plot of the categories with data regarding state profit,average bottle price, total volume and total sales.



When exploring the heatmap of our data we could see some of the expected correlations:




As espected we see that State Bottle cost and Retail price are strongly correlated, similarly to Sale($) and state_profit. Other high correlations are:

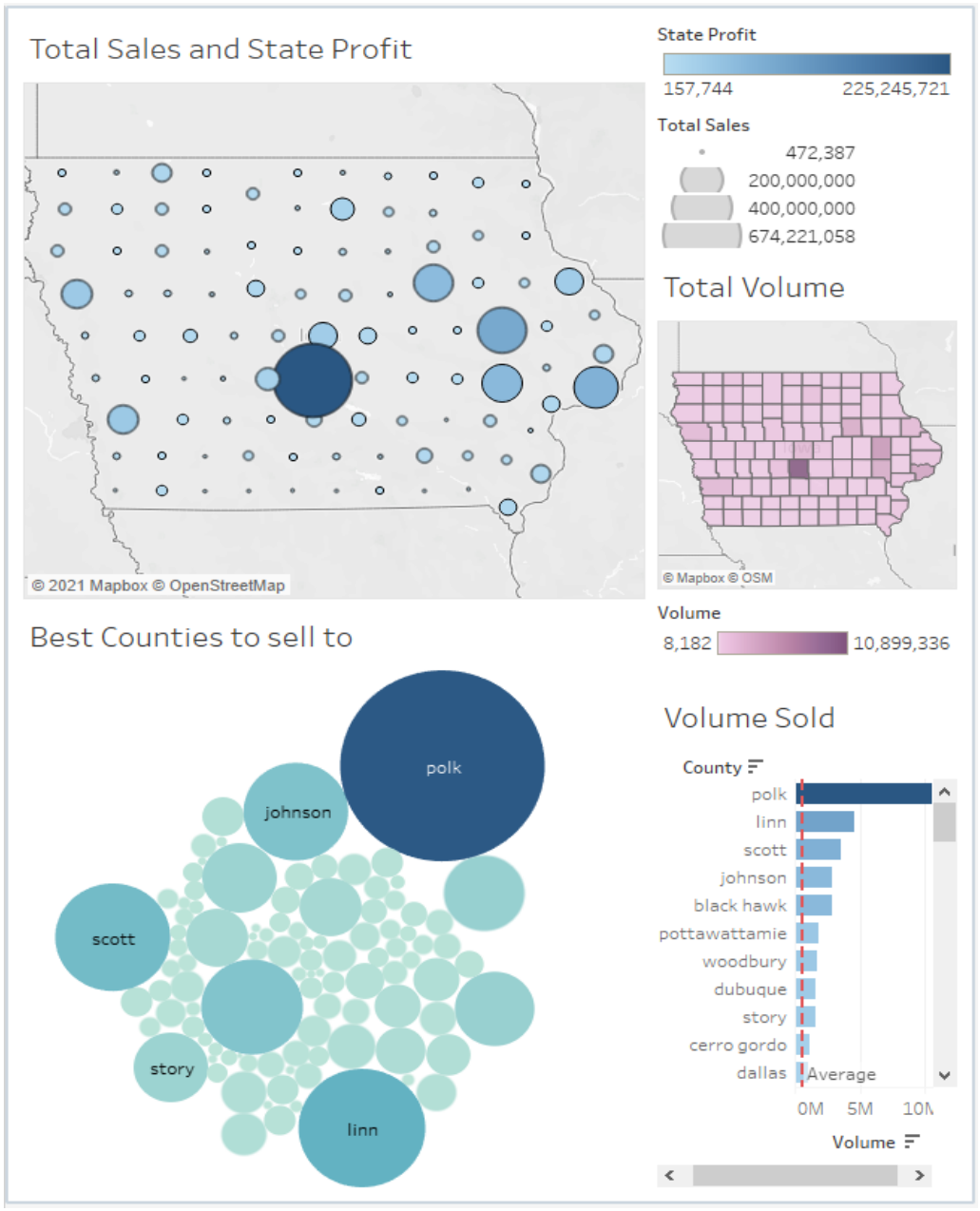
- Negative correlation of -0.46 between pack size and bottle Volume: the bigger the volume, the less bottles in a pack.
- Negative correlation of -0.27 between pack size and bottle Cost: the bigger the pack, the less bottles cost.
- Positive correlation of 0.41 between bottle Volume and bottle Cost: the bigger the Bottle, the the more it will cost.
- Finnaly, strong positive correlations (over 0.83) between Bottles Sold, Sale, Volume Sold and State Profit.

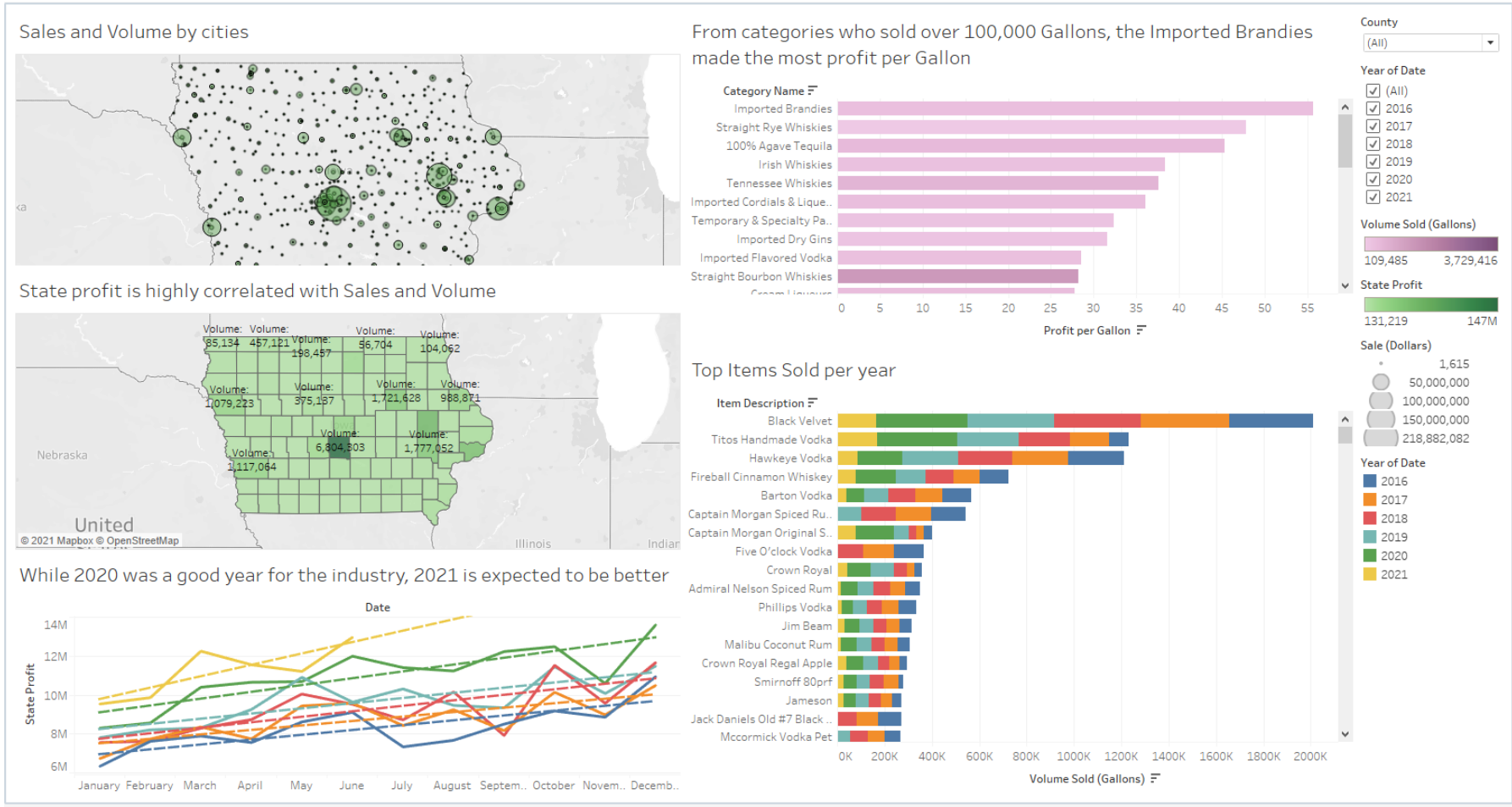
# The impact of Covid-19 on the Alcoholic Beverages Industry

Using the cleaned subset from our first notobook we then proceeded with exploration on how the pandemic impacted the Spirits industry.

To explore the covid impact on the Alcoholic industry in Iowa we have the folowing directions that we want to analyze:

1. a scatter plot to show the evolution of sales, volume sold and state profit per County. We will also use two Tableau Dashboard to better visualize the data.
2. a time series that shows the same 3 variables trends

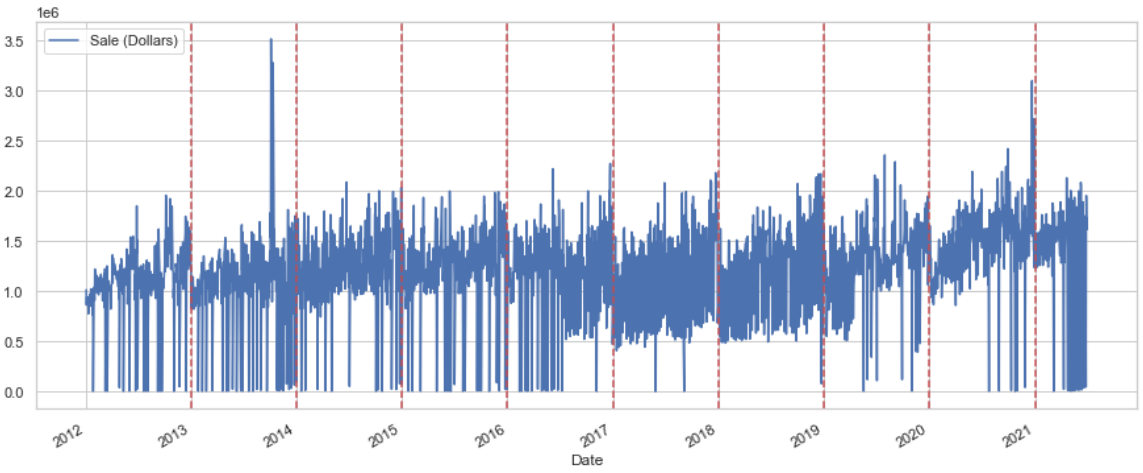


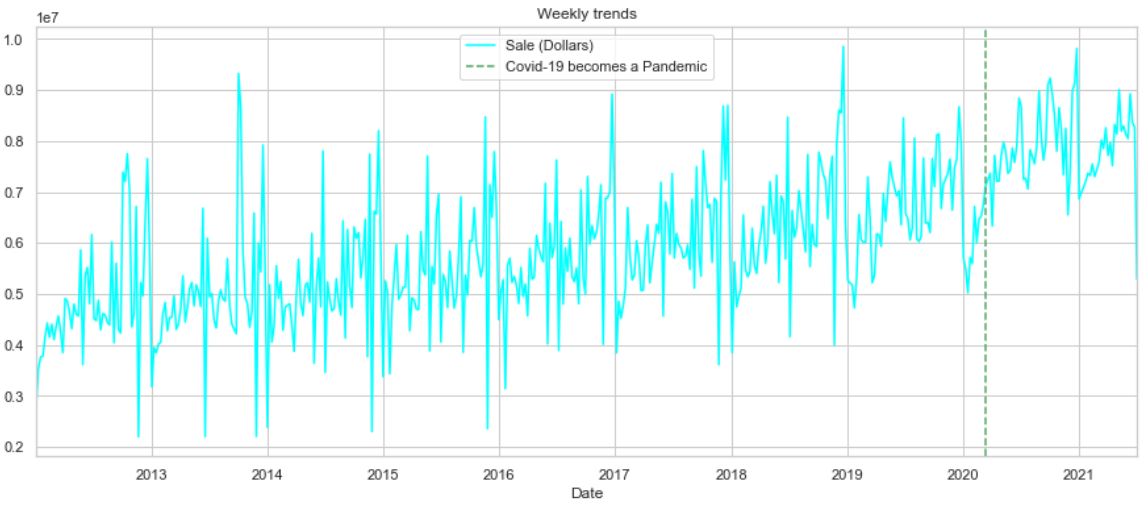


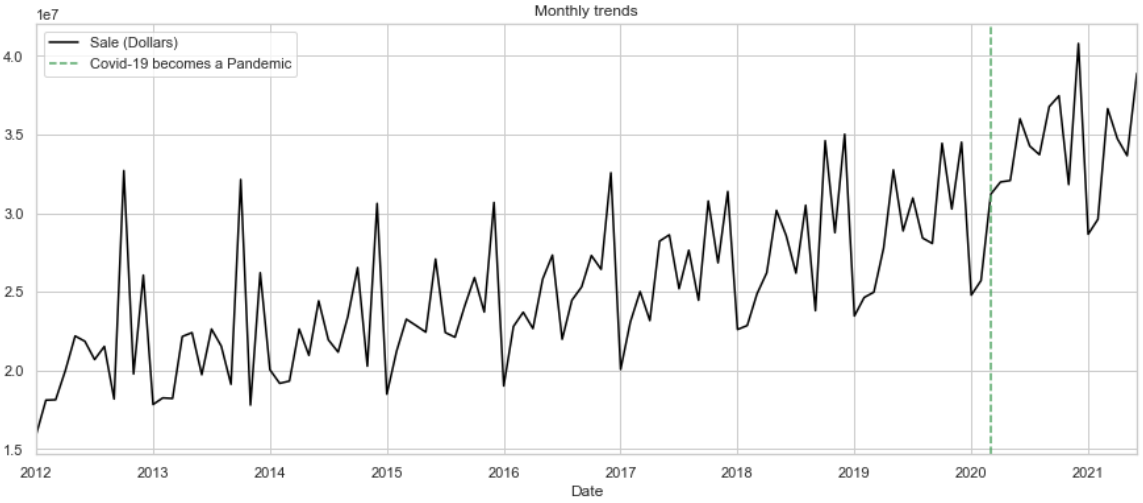

This is the second Dashboard created with Tableau, with focus on yearly trends, cities, categories and items sold. This is the link to the Tableau Public work: https://public.tableau.com/views/QuickEDA/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link

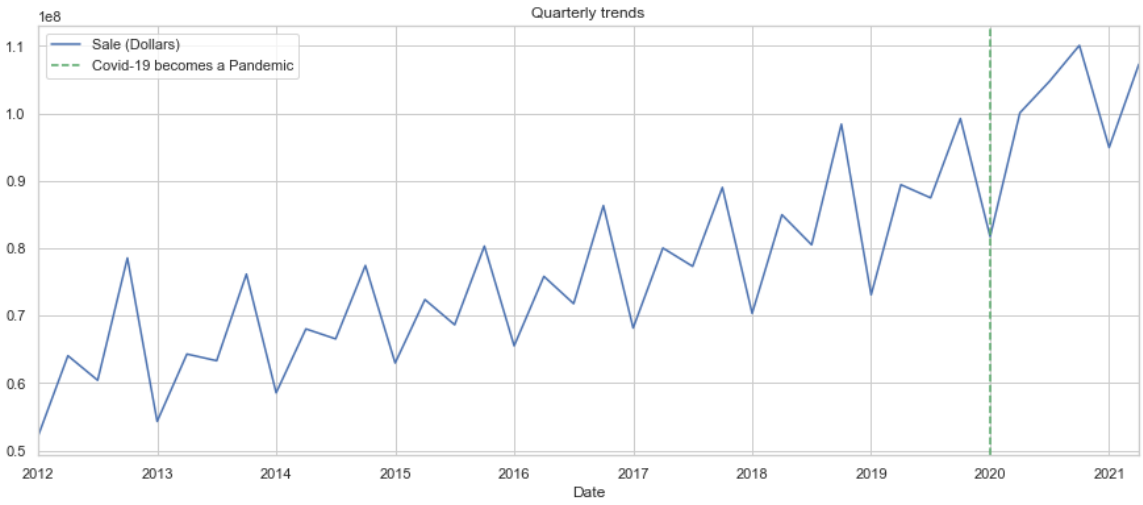
We used Tableau to create a representative Dashboard. We also used Ipywidgets to create interactive graphs to represent evolution of sales, volume of Spritis thought the years.

Next we plotted our data aggregated by day, week, month and quarter.

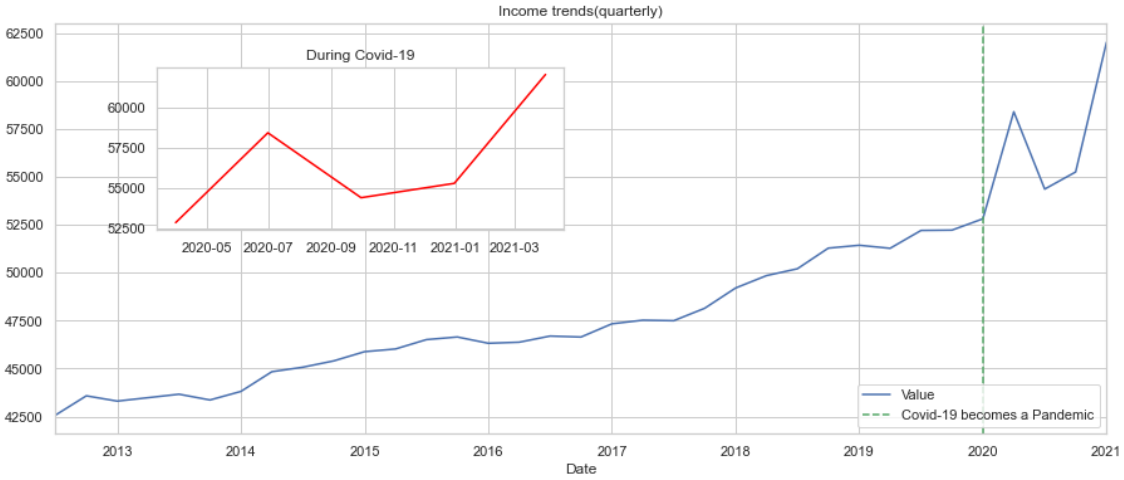


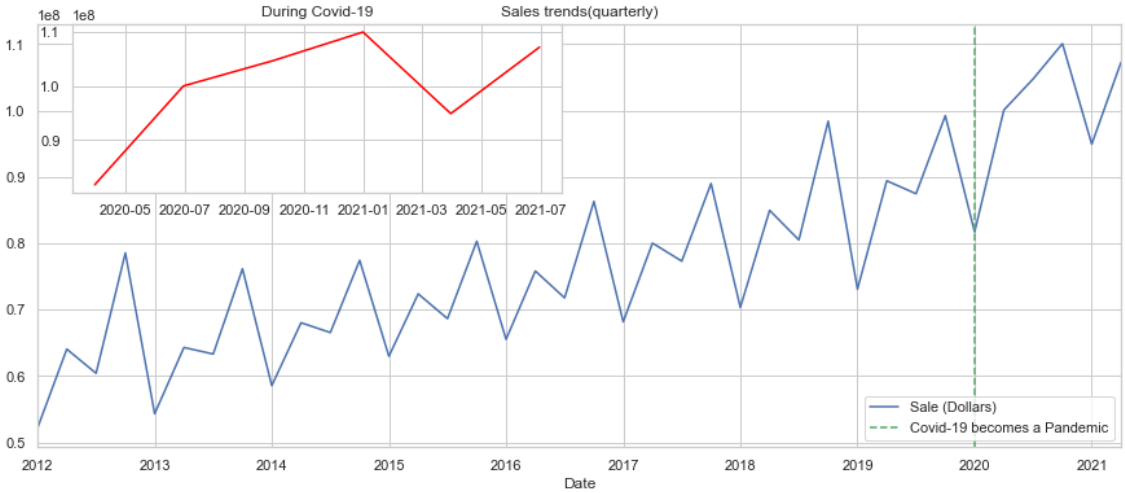






To see if the trends in Sales might have be dependent on any other than the Pandemic influence, we got some additional data, the quarterly income and the population growth. Let's see how the Income and Sales trends are plotted.



The total population seems irrelevant to the Business problem we are looking after: the impact of Covid-19 on the industry.

But the other 2 quarterly graphs are quite telling.

Let's explore the Income first.

- For the average income, we can see a big jump by more than 5.000 USD from Q1 to Q2 of 2020, right after Covid-19 was declared a pandemic and Pandemic Stimulus was send to businesses and to all US residents. Before this, it took a whole 3 years for the average income to increase by 5000 USD.
- Next, we can see the income dropping suddenly, back to where the income trend from the previous quarters would have predict the income to be.
- The evolution from Q3 to Q4 of 2020 was also the expected one.
- And lastly, we observe another big jump of more than 5000 USD from Q1 of 2021 to Q2 of 2021. That was probably associated as well with a stimulus check awarded by the US Government.

Regarding the Sales amount, we can observe a general pre-pandemic trend:

- Q1 is a bad quarter for sales
- then we see a big jump in Q2
- Q3 is slightly worse than Q2
- another big jump in Q4
- followed by a big regression from Q4 to Q1 of next year.

We can explain these by the fact that in Q4 there are a lof of holidays when people are going to parties and gatherings.

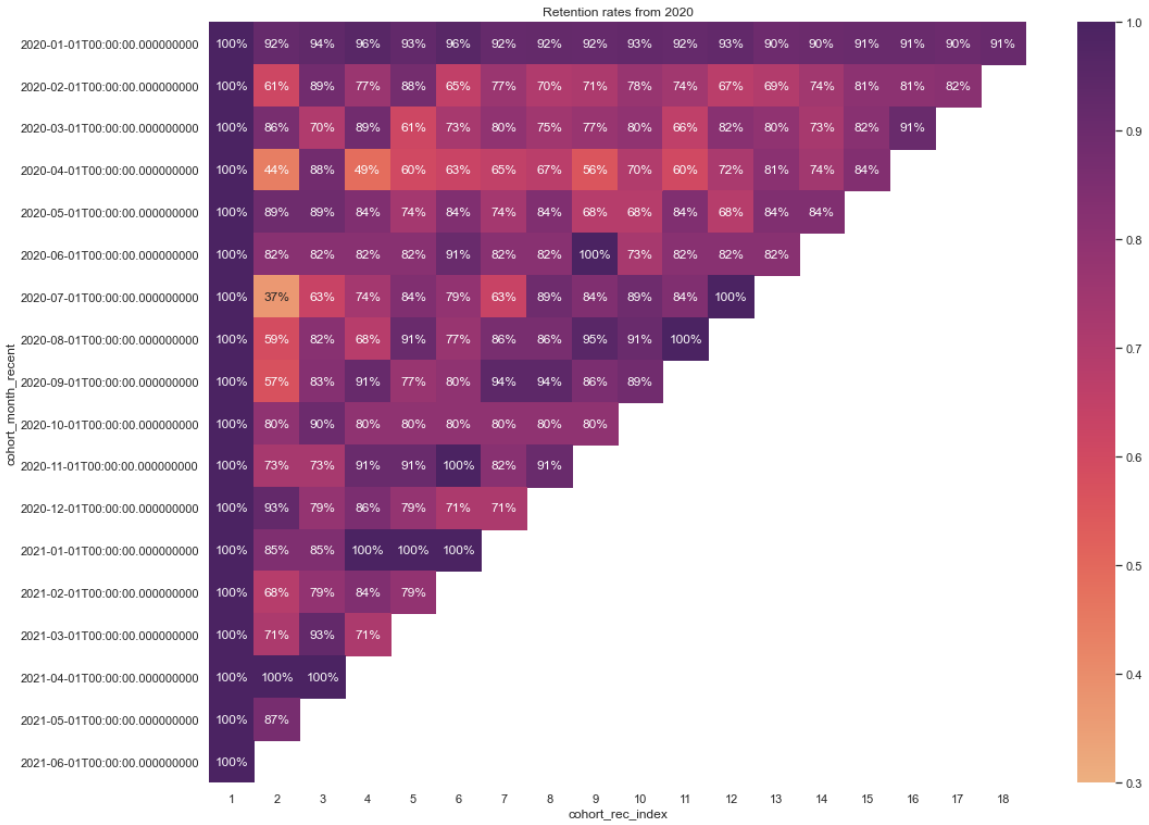
But for 2020 we see something different:

- the jump from Q1 to Q2 in 2020 was so high that the Sales in Q2 were higher than those of Q4 the previous year. This has not happened in any of the previous 8 years.
- the valley that was ussualy between Q2 and Q3 was now gone, and the sales continued an ascendent trend from Q2 to Q3 and then to Q4. And this happened with a lot of restrictions to bars and restaurant because of Covid-19.

To conclude our short analysis, there are certain correlations between the Pandemic and the growth in the Beverages industry. This growth might as well be explained by the fact that the average income increased as well during the same period.

# Pre-proccessing – Customer Segmentation with RFM

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




We can see that, unlike the retail industry where the retantion 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.

## RFM SEGMENTATION

This is a behavioral customer segmentation based on three metrics:

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

The RFM values can be grouped in multiple ways:

- Percentiles(for e.g. quantiles)

- Pareto 80/20 cut
- Custom - based on business knowledge

We are going to use percentile-based grouping.

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

| | 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

## BUILDING RFM SEGMENTS

Using the data created, we will calculate quartile value for each column and name them R, F and M.

For recency, the smaller the value, the higher the score. For the other two metrics, the higher the value, the higher the score.

| | 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 |

## 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.

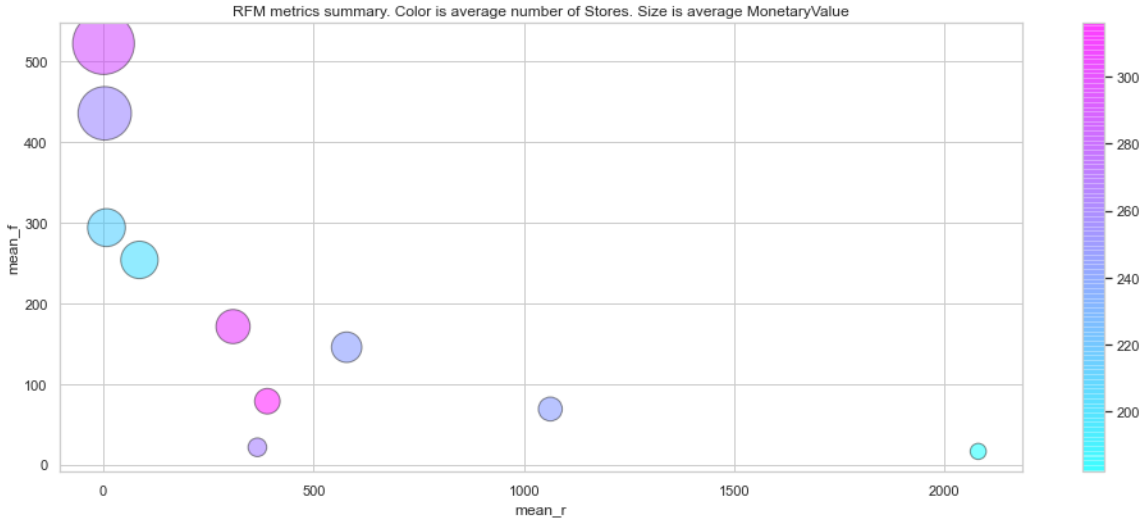
| | Frequency | MonetaryValue | Recency | R | F | M | RFM_Segment | RFM_Score |
|---|---|---|---|---|---|---|---|---|
| Store Name | | | | | | | | |
| 'Da Booze Barn / West Bend | 82 | 204930.02 | 0 | 4 | 2 | 2 | 422 | 8 |
| 10th Hole Inn & Suite / Gift Shop | 13 | 16221.98 | 2 | 4 | 1 | 1 | 411 | 6 |
| 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 |

| | Recency | Frequency | MonetaryValue | |
|---|---|---|---|---|
| | mean | mean | mean | count |
| RFM_Score | | | | |
| 3 | 2080.0 | 16.2 | 24726.9 | 182 |
| 4 | 366.6 | 21.3 | 44589.0 | 261 |
| 5 | 1062.9 | 68.7 | 120773.1 | 243 |
| 6 | 390.2 | 78.4 | 156734.3 | 316 |
| 7 | 578.7 | 145.3 | 316769.9 | 243 |
| 8 | 308.7 | 170.9 | 500678.2 | 303 |
| 9 | 86.3 | 253.5 | 720828.4 | 202 |
| 10 | 7.9 | 293.5 | 762875.3 | 211 |
| 11 | 3.7 | 435.1 | 3012210.0 | 256 |
| 12 | 0.9 | 521.6 | 5408756.7 | 289 |

## METRICS SUMMARY PER RFM SCORES

Finally, let's view our Summary metrics per RFM score.

Another way to classify Customers based on RFM besides RFM segment or score is to use Unsupervised clustering.




## PREPROCESSING DATA BEFORE K-MEANS UNSUPERVISED CLUSTERING

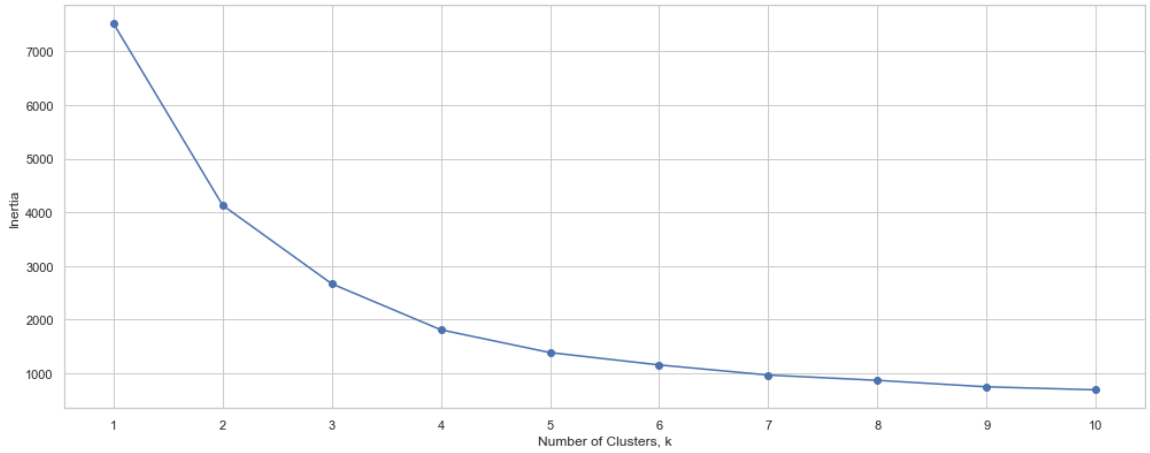
The prerequisites for K-Means are the following:

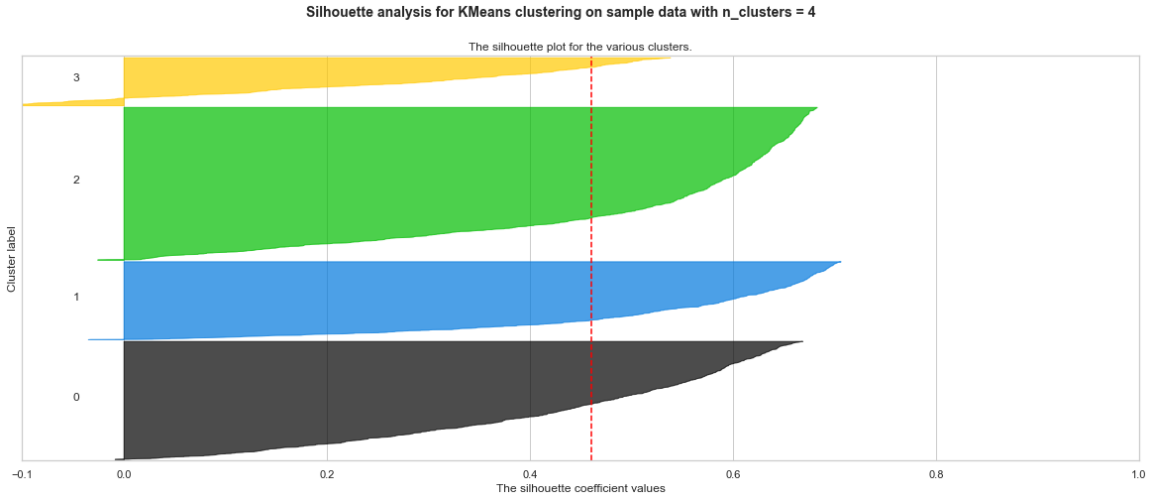
- Symetric distribution of variables(not skewed)
- Variable with same mean
- Varaible with same variance

## KMEANS CLUSTERING

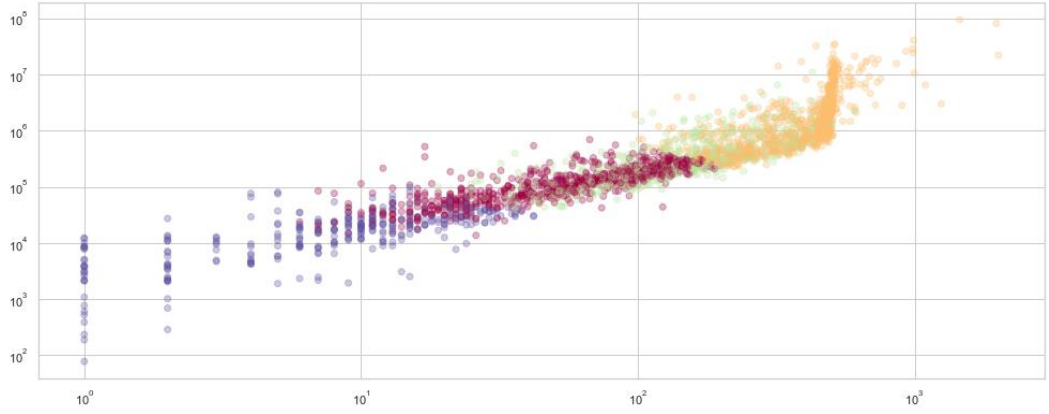
As methods of choosing the optimal K clusters we will use:

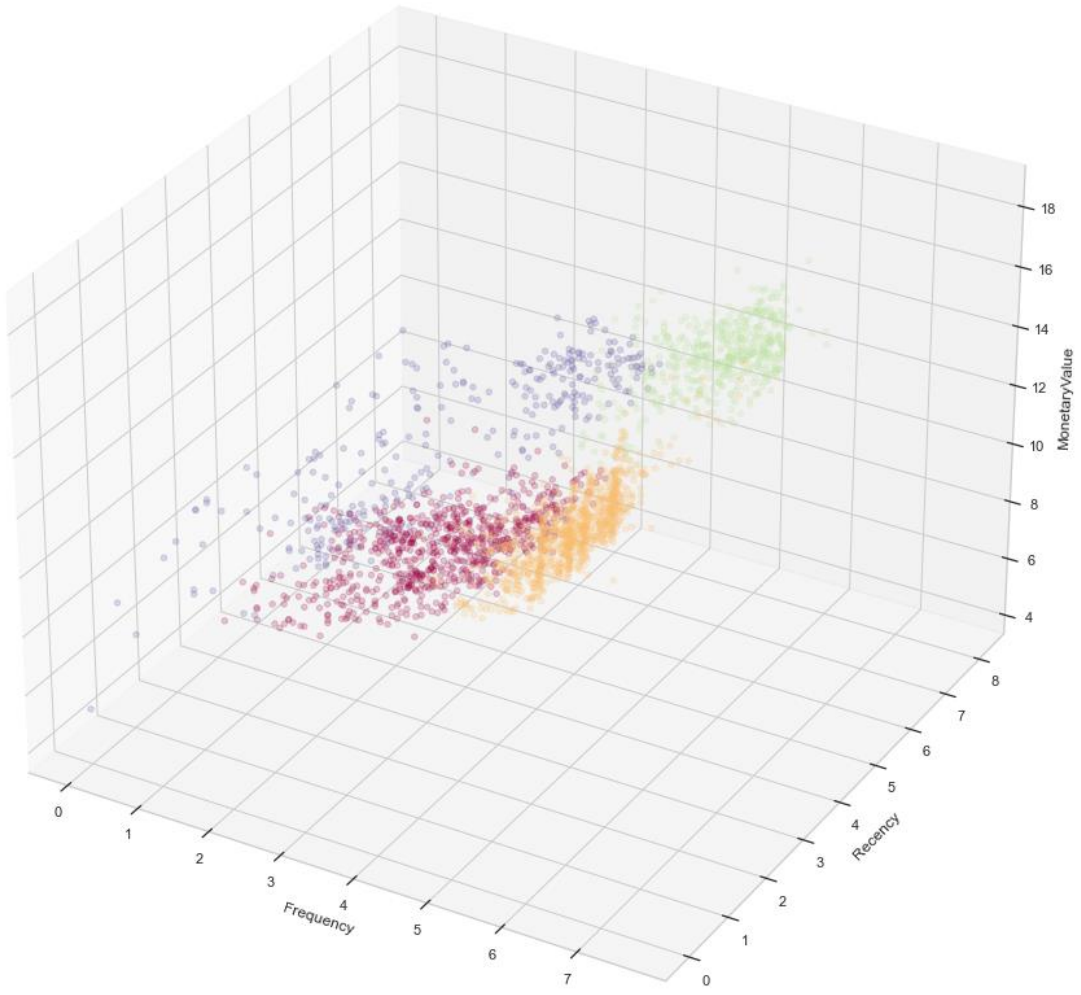
- Elbow method
- Silhouette scores and plots




Based on the 2 metrics we decided to choose 4 clusters for our stores.






| | 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 |

It seems that Cluster o has the folowing characteristics:

- low recency
- medium - low frequency

- medium - low monetaryValue
- These stores have bought recently, but only bought 67 times in the past 8 years, with an average of 137000 USD. These might be the smaller liquor stores that are buying consistent just every other month because it takes a lot to sell their products.

Cluster 1 has the folowing characteristics:

- low recency
- very high frequency
- very high monetaryValue
- These stores have bough recently, with a high frequency of 390 times in the past 8 years( an average of once per week) and a MonetaryValue of 2,700,000. This is also the biggest group, with 963 stores. This are the big Liquor stores, with a lot of traffic and that are looking for constant resupply.

Cluster 2 has the folowing characteristics:

- high recency
- very high frequency
- very high monetaryValue
- O average these stores have not bough from the State for 1500 days. That is almost 5 year. So even their frequency is not very low and the Monetary Value ios quite high, these are the Stores who stopped buying from the State for different reasons. These is one of the main group to be targeted with Marketing Campaigns.

Cluster 3 has the folowing characteristics:

- high recency
- very low frequency
- very low monetaryValue
- These are the small stores that stopped buying from the state more than 3 years ago. Their average Monetary Value is only 20000USD and they only bought 11 in the last 8 years on average. this is a group that doesn't present any interest.

## Modeling

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 of methods and 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

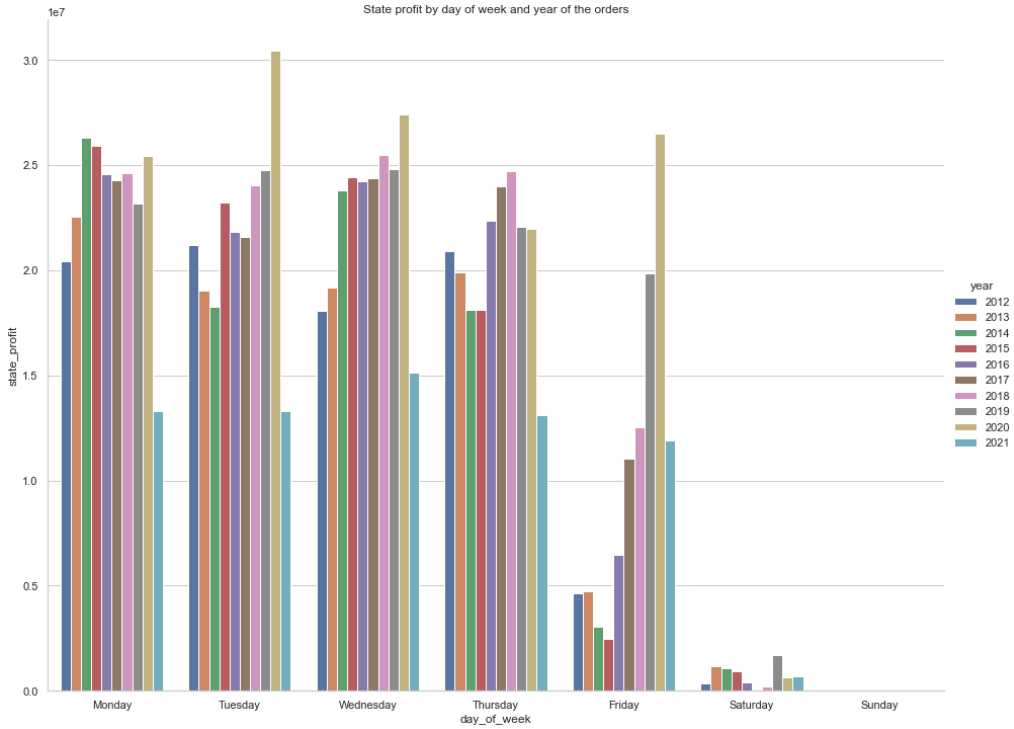
Choosing the metrics

- RMSE
- MAE

Models Tested

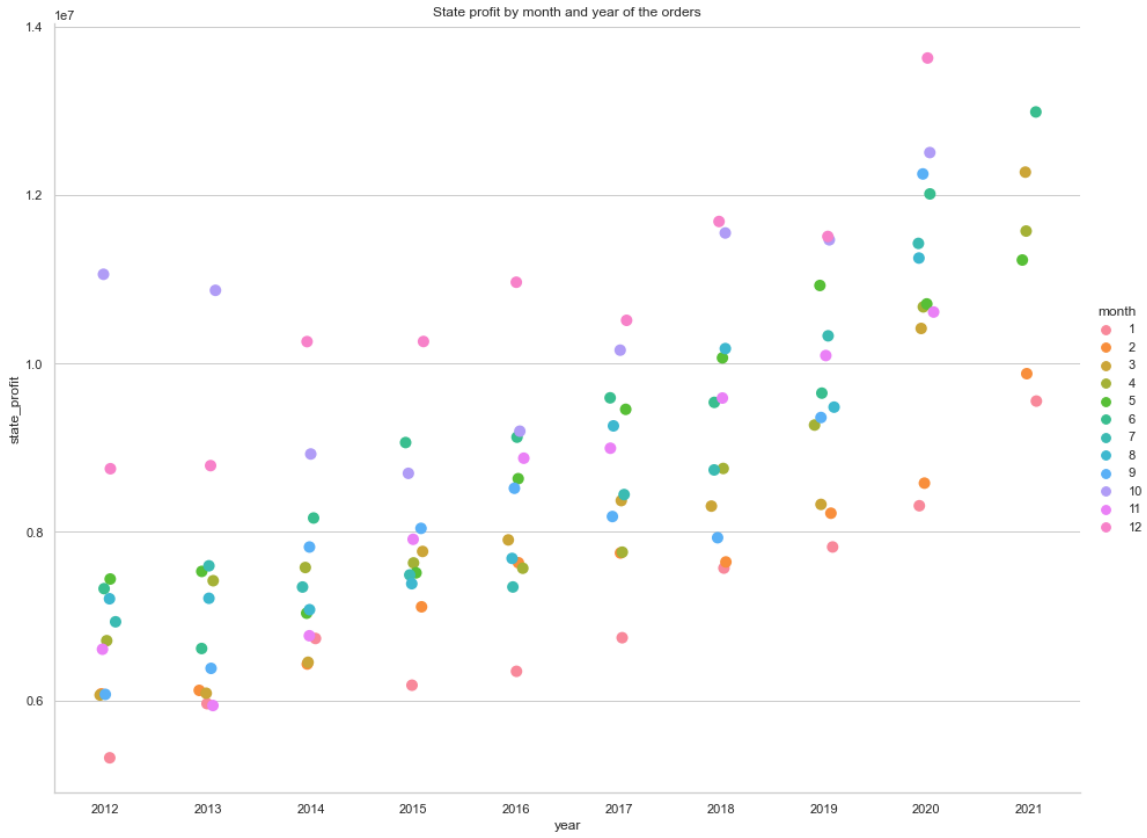
- AutoRegressive model (AR 1)
- BIC(Bayesian Information Criterion) adjusted AR model
- Moving Average Model(MA)
- Autoregressive Moving Average (ARMA)
- Seasonal autoregressive integrated moving average (SARIMAX)

We began by exploring the profit by day of the week and month.




We can observe that:

- There are no sales happening on Sunday.
- There are very few sales on Saturday.
- If before 2019 there were way fewer Sales(and profit) On Friday orders, since then we can see taht the amount of profit from Friday orders is within the same values as those from any other workday.
- Sales on Tuesday, Wednesday and Friday were at the highest values in 2020.




## VALIDATION SET

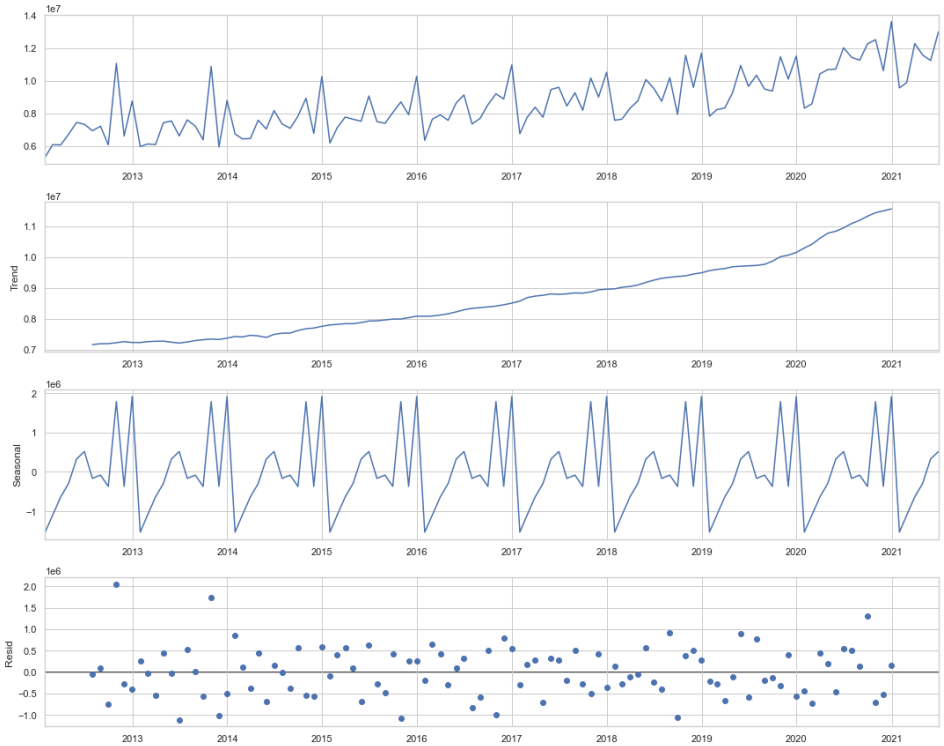
So far from the start of the exercise we worked with data from 2012 to June 30th 2021. We recently checked and on Iowa.gov website they recently uploaded another batch of data, containing Sales from July 2021.

Since this data was not used in any of our analysis, it is the perfect candidate for the Validation Set.

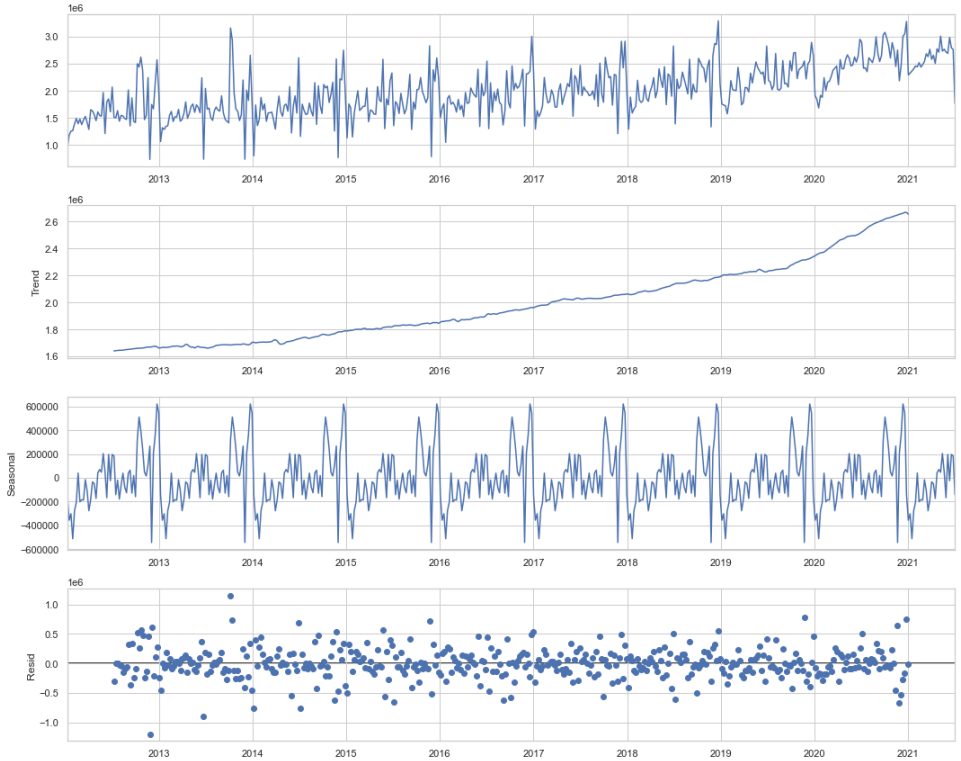
## DECOMPOSITION OF AGGREGATED DATA

We will create decompositions of our aggregated data.

Monthly data:




Weekly data:



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

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

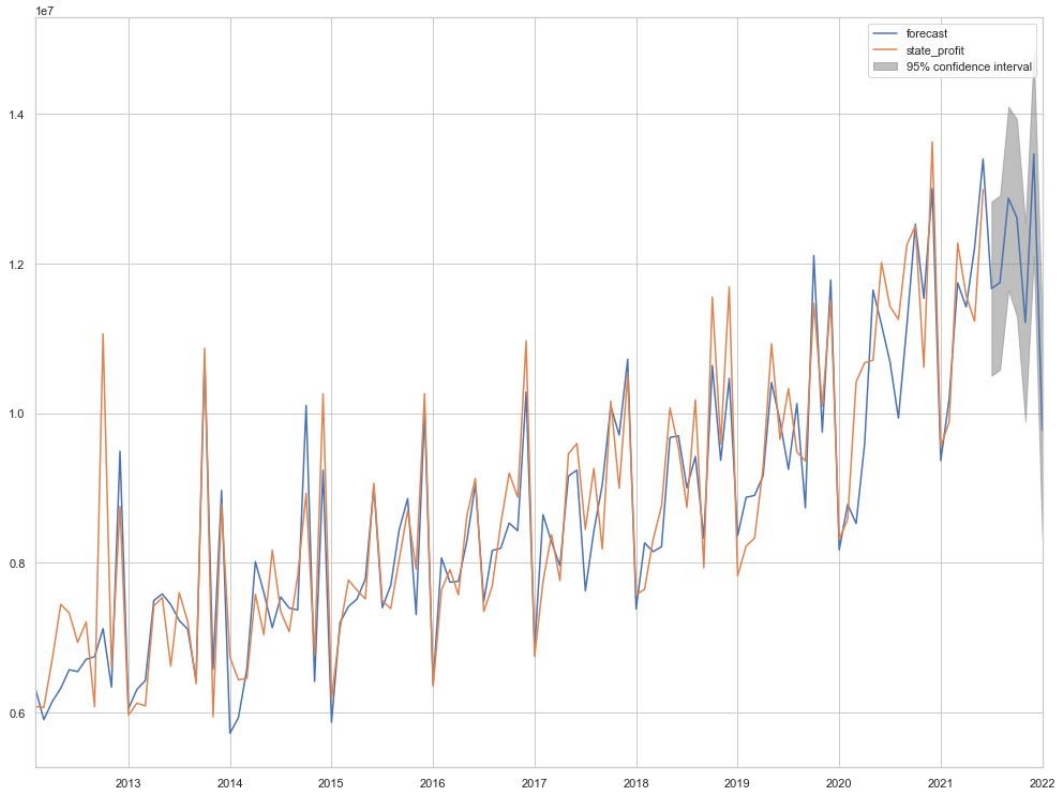
| | data_agg | model | params | rmse | mae |
|---|---|---|---|---|---|
| 139 | month_agg | SARIMAX | p=2, d=1, q=10, seasonal_order=(1,1,0,12) | 3330.86 | 3330.86 |
| 136 | month_agg | SARIMAX | p=2, d=0, q=10, seasonal_order=(1,1,0,12) | 42648.82 | 42648.82 |
| 88 | month_agg | ARIMA | p=0, d=2, q=2 | 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 |

| | data_agg | model | params | rmse | mae | Accuracy on Validation |
|---|---|---|---|---|---|---|
| 139 | month_agg | SARIMAX | p=2, d=1, q=10, seasonal_order=(1,1,0,12) | 3330.86 | 3330.86 | 99.971591 |
| 136 | month_agg | SARIMAX | p=2, d=0, q=10, seasonal_order=(1,1,0,12) | 42648.82 | 42648.82 | 99.636252 |
| 88 | month_agg | ARIMA | p=0, d=2, q=2 | 52372.53 | 52372.53 | 99.553320 |
| 112 | month_agg | ARIMA | p=0, d=2, q=2 | 52372.53 | 52372.53 | 99.553320 |
| 77 | month_agg | AR | p =16 | 61794.20 | 61794.20 | 99.472963 |
| 128 | month_agg | SARIMAX | p=1, d=0, q=8, seasonal_order=(1,1,0,12) | 61840.06 | 61840.06 | 99.472572 |

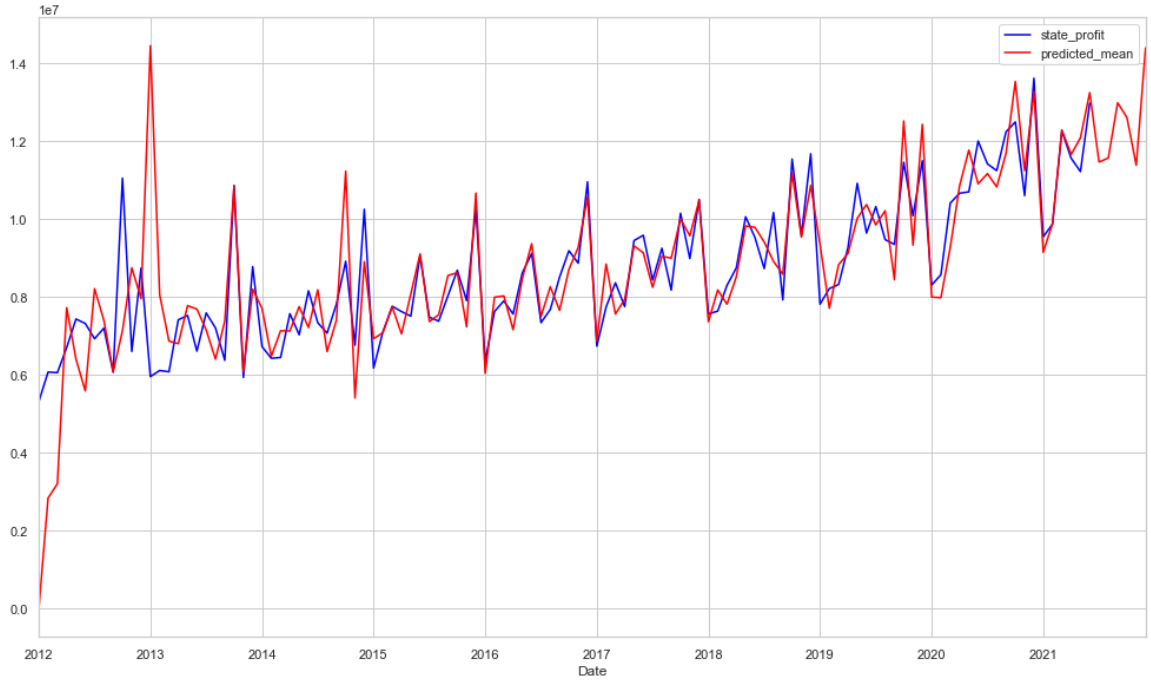
As we can see, the Accuracy of predictions on the Validation set is 99% with all the top models.

These are 2 graphs of the models at the top.

AR16 – on monthly aggregated data.




Best model - SARIMAX with p=2, d=1, q=10, seasonal_order=(1,1,0,12) on monthly aggregations



# Future Work

For future work we still want to explore the last identified Business Problem – helping a store owner to pick a new location for a store so that he will maximize sales.

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 best model.

## 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 foreseable 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 atract new customers.
- Expand to new locations within the same state on to other states.

The Covid-19 pandamic 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.