Predicting profit and Driving Business Insights using Iowa spirits sales

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

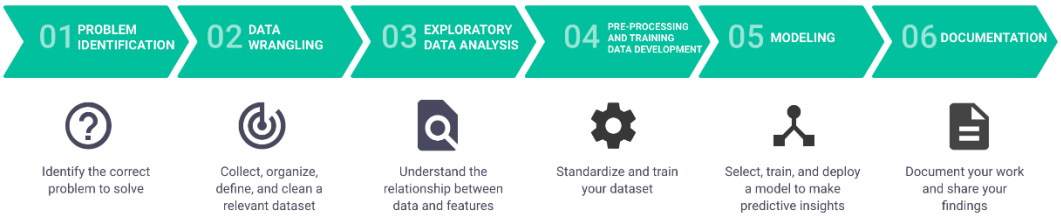
This project looks to resolve some issues that are of high importance for two distinct business entities: Iowa Department of Commerce, Alcoholic Beverages Division and Liquor Store Owners. These are the **5 Business Problems** that we identified and that were solved with this project:

- Exploration on what was the impact of Covid-19 on the Alcoholic Beverages Industry.
- Storage capacity management exploratory analysis for Iowa Department of Commerce, Alcoholic Beverages Division.
- 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.
- 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 0(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.