liquor

November 5, 2024

0.1 Analyzing Liquor Sales in Iowa for Inventory Optimization

Team 4: Brendan Wilcox, Mohammed Zaid Bin Haris, Nimisha Agarwal, Quan Nguyen

0.1.1 1. Problem Statement

Our project aims to equip a new liquor seller entering the Iowa market with the insights needed to quickly understand and adapt to local demand patterns. Entering a new market poses challenges, particularly in forecasting demand, managing inventory efficiently, and setting pricing that attracts customers while supporting profitability. By analyzing historical sales data, we will create a predictive framework tailored to Iowa's unique trends, helping this seller make data-informed decisions about stock, distribution, and pricing strategies.

Our study will focus on:

- **Forecasting Weekly/Monthly Sales**: We will build models that predict weekly and monthly sales, allowing the seller to maintain optimal inventory levels, reducing the risk of either stockouts or overstocking as they learn the rhythms of the Iowa market.
- **Understanding Customer Price Sensitivity and Predicting Optimal Prices**: We will assess customer sensitivity to price changes across product types, giving the seller a pricing guide that balances customer expectations with profitability. Additionally, we will explore price prediction models to recommend pricing levels that respond to market trends and demand cycles, supporting the seller in maintaining competitive yet profitable prices.

With these insights, the seller will be well-positioned to establish a strong foothold in the Iowa market, align operations with local demand patterns, and drive revenue through strategic pricing and inventory management from the outset.

0.1.2 2. Data Source

We will use the Iowa Liquor Sales dataset from Iowa's Alcoholic Beverages Division, accessible via data.iowa.gov. We are accessing the data set from Big Query Public Datasets. This dataset offers detailed records of liquor sales by Iowa Class "E" license holders, including grocery stores, liquor stores, and convenience stores, from January 1, 2012, to the present.

0.1.3 3. Loading the Dataset

[3]:
```python
from google.cloud import bigquery

client = bigquery.Client()
```

[ ]:
```python
sql = """
SELECT * FROM `bigquery-public-data.iowa_liquor_sales.sales`
"""
df = client.query(sql).to_dataframe()
df.head()
```

[ ]:
```
  invoice_and_item_number        date store_number  \
0          INV-15980300037  2018-11-29         5524
1          INV-20302500011  2019-06-28         4324
2          INV-73584900016  2024-08-26         2465
3          INV-21959900061  2019-09-17         2604
4             S26735800020  2015-07-15         2238

                          store_name                address          city  \
0          EAST SIDE LIQUOR & GROCERY        1116 E NEVADA ST  MARSHALLTOWN
1             DAYTON COMMUNITY GROCERY           22 NORTH MAIN        DAYTON
2                  SID'S BEVERAGE SHOP           2727 DODGE ST       DUBUQUE
3  HY-VEE WINE AND SPIRITS / LE MARS        1201 12TH AVE SW       LE MARS
4                   ADVENTURELAND INN  3200 ADVENTURELAND DR       ALTOONA

  zip_code                   store_location county_number    county  …  \
0  50158.0       POINT(-92.893113 42.044345)            64  MARSHALL  …
1    50530         POINT(-94.068439 42.26168)            94   WEBSTER  …
2  52003.0  POINT(-90.70505003 42.492316017)          None   DUBUQUE  …
3    51031         POINT(-96.18335 42.778257)            75  PLYMOUTH  …
4    50009         POINT(-93.49924 41.658513)            77      POLK  …

  item_number                     item_description pack bottle_volume_ml  \
0       26827           JACK DANIELS OLD #7 BLACK LBL   12             1000
1       43127                        BACARDI SUPERIOR   12             1000
2       87937                           JUAREZ SILVER   12             1000
3       64870                       FIREBALL CINNAMON   48              100
4       58838  JOSE CUERVO AUTHENTIC LIME MARGARITA    6             1750

  state_bottle_cost state_bottle_retail  bottles_sold  sale_dollars  \
0             18.89               28.34             4        113.36
1              9.50               14.25             4         57.00
2              9.00               13.50             4         54.00
3              0.90                1.35            96        129.60
4              8.20               12.30            36        442.80
```

```
   volume_sold_liters  volume_sold_gallons
0                 4.0                 1.05
1                 4.0                 1.05
2                 4.0                 1.05
3                 9.6                 2.53
4                63.0                16.64

[5 rows x 24 columns]
```

0.1.4 4. Dataset Description

[5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30305765 entries, 0 to 30305764
Data columns (total 24 columns):
 #   Column                   Dtype
---  ------                   -----
 0   invoice_and_item_number  object
 1   date                     dbdate
 2   store_number             object
 3   store_name               object
 4   address                  object
 5   city                     object
 6   zip_code                 object
 7   store_location           object
 8   county_number            object
 9   county                   object
 10  category                 object
 11  category_name            object
 12  vendor_number            object
 13  vendor_name              object
 14  item_number              object
 15  item_description         object
 16  pack                     Int64
 17  bottle_volume_ml         Int64
 18  state_bottle_cost        float64
 19  state_bottle_retail      float64
 20  bottles_sold             Int64
 21  sale_dollars             float64
 22  volume_sold_liters       float64
 23  volume_sold_gallons      float64
dtypes: Int64(3), dbdate(1), float64(5), object(15)
memory usage: 5.5+ GB
```

[6]:
```
df.describe()
```

[6]:

```
               pack  bottle_volume_ml  state_bottle_cost  state_bottle_retail  \
count  3.030576e+07      3.030576e+07       3.030576e+07         3.030576e+07
mean   1.211732e+01      8.739757e+02       1.081060e+01         1.622570e+01
std    7.798025e+00      6.220215e+02       1.345413e+01         2.017994e+01
min    1.000000e+00      0.000000e+00       0.000000e+00         0.000000e+00
25%    6.000000e+00      7.500000e+02       5.740000e+00         8.620000e+00
50%    1.200000e+01      7.500000e+02       8.500000e+00         1.275000e+01
75%    1.200000e+01      1.000000e+03       1.300000e+01         1.950000e+01
max    3.360000e+02      3.780000e+05       2.498902e+04         3.748353e+04

       bottles_sold  sale_dollars  volume_sold_liters  volume_sold_gallons
count  3.030576e+07  3.030576e+07        3.030576e+07         3.030576e+07
mean   1.087976e+01  1.462172e+02        9.146541e+00         2.413392e+00
std    3.070450e+01  5.168930e+02        3.641226e+01         9.619213e+00
min   -7.680000e+02 -9.720000e+03       -1.344000e+03        -3.550400e+02
25%    3.000000e+00  3.600000e+01        1.500000e+00         4.000000e-01
50%    6.000000e+00  7.740000e+01        4.800000e+00         1.260000e+00
75%    1.200000e+01  1.500000e+02        1.050000e+01         2.770000e+00
max    1.500000e+04  2.795573e+05        1.500000e+04         3.962580e+03
```

The Iowa Liquor Sales dataset contains approximately 30 million records with 24 columns. Key features include:

- **Store Information**: Fields such as store ID, name, address, city, and county, which will allow us to analyze geographic differences in demand.
- **Product Information**: Details such as product ID, description, category, vendor, and bottle volume, enabling the categorization of products for a more targeted analysis.
- **Sales Data**: Key metrics like order date, number of bottles sold, retail price, total sales amount, and volume sold in liters/gallons, which are essential for forecasting sales trends.

The dataset contains a mixture of text (store and product descriptions), numeric (sales figures, bottle costs), and datetime (order dates) variables, making it well-suited for both time-series and categorical analyses.

0.1.5 5. Visualisations

[6]:

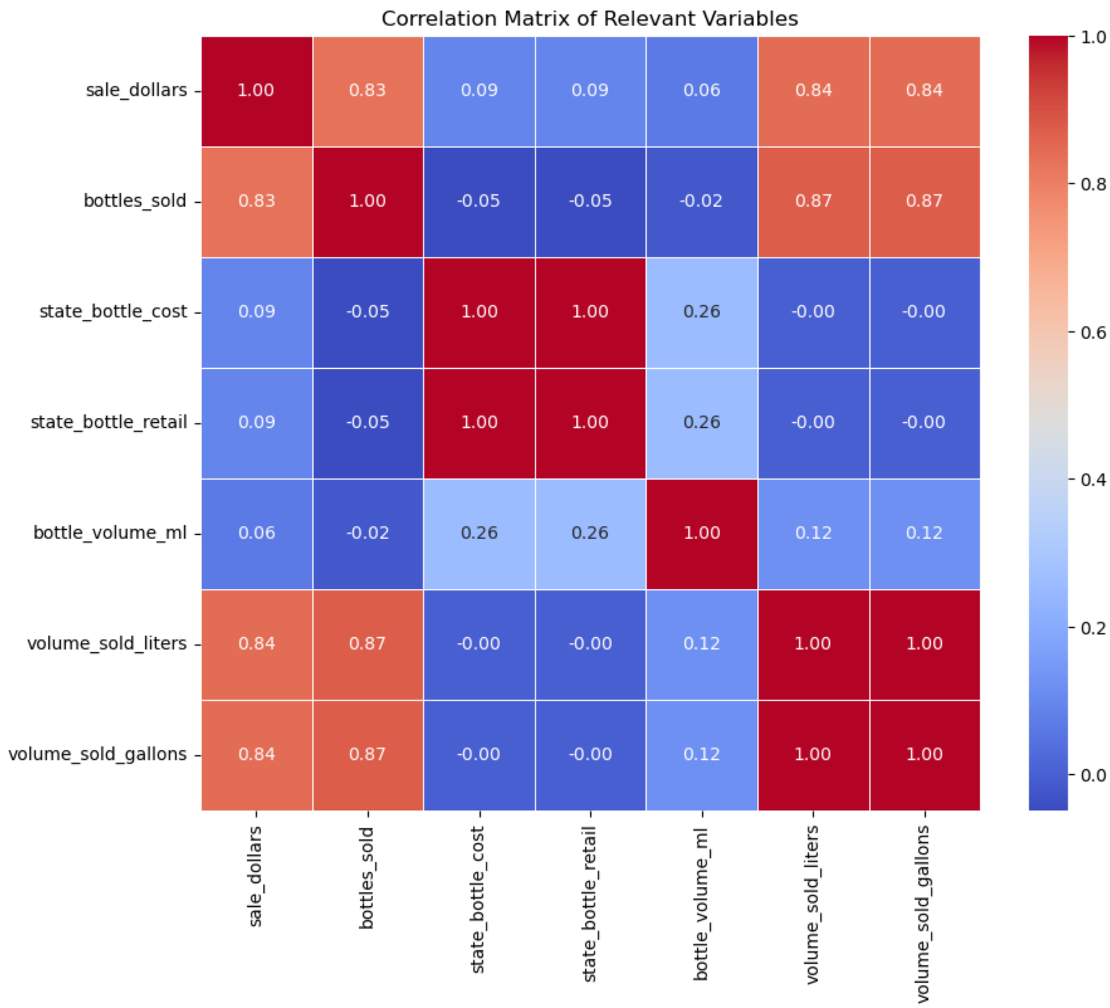
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

relevant_columns = [
    'sale_dollars',
    'bottles_sold',
    'state_bottle_cost',
    'state_bottle_retail',
    'bottle_volume_ml',
    'volume_sold_liters',
    'volume_sold_gallons'
```

```
]

correlation_matrix = df[relevant_columns].corr()

plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Correlation Matrix of Relevant Variables")
plt.show()
```

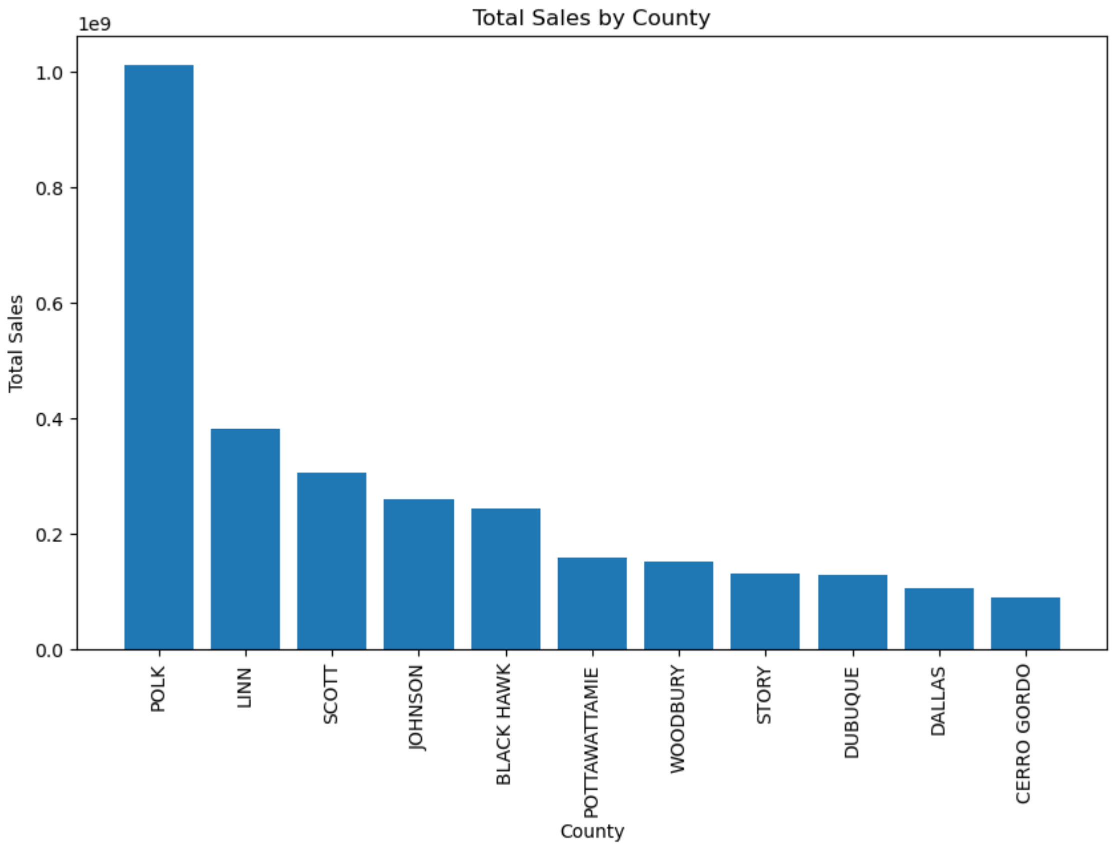


Summary The correlation matrix shows strong positive correlations between sale_dollars, bottles_sold, volume_sold_liters, and volume_sold_gallons, suggesting that higher sales are associated with larger quantities sold. Meanwhile, state_bottle_cost and state_bottle_retail have minimal correlation with other variables, indicating that price per bottle does not strongly influence total sales volume.

It is intutitive that there would be strong positive correlations between sale_dollars, bottles_sold, volume_sold_liters, and volume_sold_gallons, as they are the main inputs in the Price * Quantity = Revenue equation. Because of this, creating a correlation matrix of these variables was not about finding unexpected insights, but about performing a sanity check to make sure that our data made sense before proceeding withing further analysis.

[14]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
groupby_county = df.groupby('county')
sum_sales = groupby_county['sale_dollars'].sum().sort_values(ascending=False)
top10_sum_sales = sum_sales[0:11]

plt.figure(figsize=(10, 6))
plt.bar(top10_sum_sales.index, top10_sum_sales.values)
plt.xlabel('County')
plt.ylabel('Total Sales')
plt.title('Total Sales by County')
plt.xticks(rotation=90);
plt.show()
```



Summary This chart displays the top 10 counties by total liquor sales, with Polk County significantly outpacing the others. Linn and Scott counties follow, but there is a steep drop in sales after Polk, highlighting its dominance in liquor sales among the top counties.

This graph is interesting to us because it shows that georgraphy will play a factor in our sales forecast. With that, we will need to consider "county" as a variable in our model and pay attention to other categorical variables in general as we continue our analysis.

0.1.6 6. Anaysis Plan and Metrics

The anticipated results of our project include predictive models that offer clear insights into demand forecasting, inventory optimization, and pricing strategies to support a new liquor seller entering the Iowa market.

- **Sales Forecasting**: Using time series analysis (e.g., ARIMA, Prophet, LSTM), we'll predict weekly and monthly sales to help the seller maintain optimal inventory. Models will be evaluated using metrics like MAE, RMSE, and MAPE for accuracy.
- **Price Estimation**: We'll assess price sensitivity and predict optimal prices by analyzing how price changes impact sales across product categories. Evaluation will include MAE, RMSE, and cross-validation for stability, with methods to improve accuracy through feature selection and tuning.
- **Expected Results**: Accurate sales forecasts to manage inventory, elasticity scores identifying price points with the most demand impact, and predictive pricing recommendations for competitive positioning.

We'll compare models based on accuracy, feature selection, and parameter tuning to find the best-performing approach. These insights will enable data-driven decisions on inventory, pricing, and demand planning, providing a strong market entry strategy for the seller.

0.1.7 7. Potential Implications

The results of this project will provide actionable insights for the new liquor seller to make data-driven decisions in inventory management, demand planning, and pricing.

In practice, our predictive models will enable the seller to:

- **Optimize Inventory**: Accurately forecast demand, reducing costs from stockouts or overstocking.
- **Anticipate Demand Spikes**: Prepare for high-demand periods, like holidays, ensuring product availability.
- **Set Profitable Prices**: Use price sensitivity insights to attract customers while maximizing revenue.

This project minimizes risks in entering the Iowa market, supporting efficient operations, customer satisfaction, and profitability. The predictive framework can also scale as the business grows, making it a valuable long-term tool.

0.1.8 8. Proposal Discussion

We met with Professor Nachiketa Sahoo on 30th October and 4th November to discuss our project proposal and received feedback on our approach.

[ ]:

```

```