IDS706 Project: Cloud SQL

Project Overview:

-In this project, I will explore some insights using the Iowa Liquor Retail Sales Data.

-The Data would be explored inside the Big Query Platform in GCP.

-In addition to the result of SQL queries, I would also make plots using the internal UI. I find it efficient and enjoyable getting insights through the different plotting and visualizing UIs from Big Query Platform.

Data:

-This data contains every wholesale purchase of liquor in the State of Iowa by retailers for sale to individuals since January 1, 2012. The Sate of Iowa controls the wholesale distribution of liquor intended for retail sale, which means this dataset offers a complete view of retail liquor sales in the entire state. The dataset contains every wholesale order of liquor by all grocery stores, liquor stores, convenience stores, etc., with details about the store and location, the exact liquor brand and size, and the number of bottles ordered. In this project we will just draw some of the insights as examples from the data.

Insights in the form of EDA:

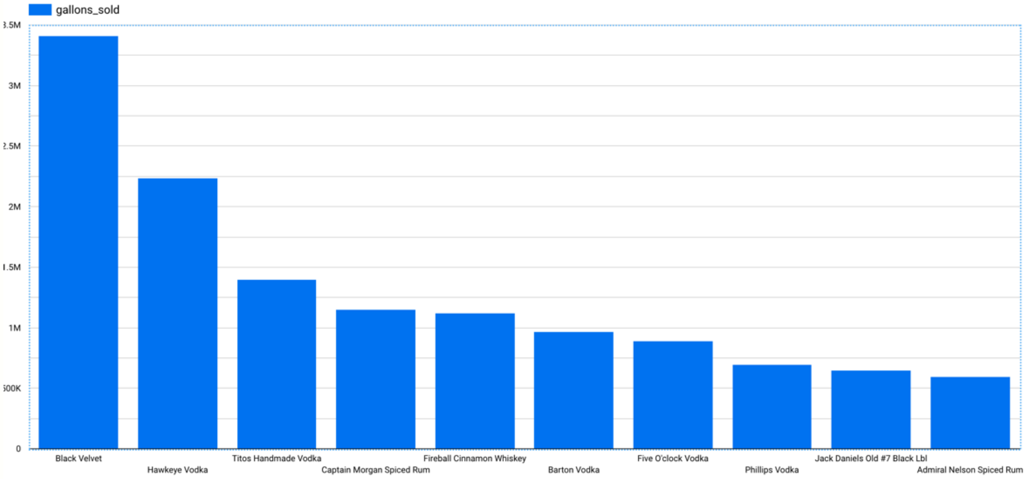
(1) What is the most popular consumed liquor in Iowa?

-Insight: The most popular consumed liquor is Black Velvet, around 3.4 million gallons sold. It is followed by Hawkeye Vodka and Titos Handmade Vodka.

-SQL:

```
SELECT item_description,ROUND(SUM(volume_sold_gallons),2) AS gallons_sold
FROM `bigquery-public-data.iowa_liquor_sales.sales`
GROUP BY 1
ORDER BY 2 DESC;
```

-Data Visualization:



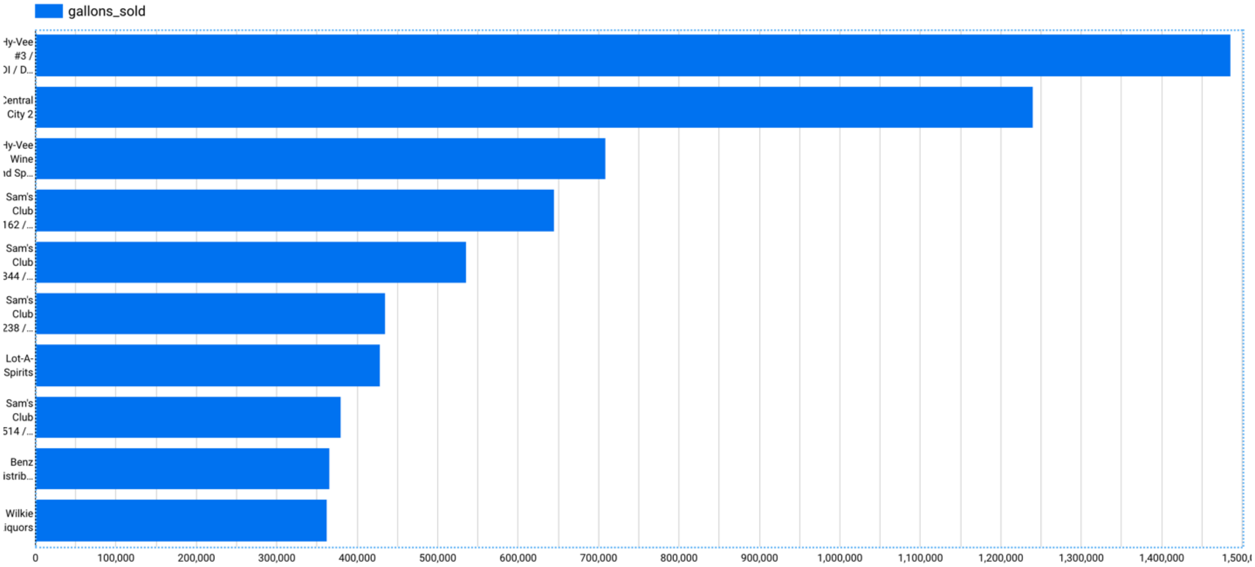
(2) What store have sold the most gallons of liquor?

-Insight: The store that had the most gallons of liquor sold turns out to be Hy-Vee #3 / BDI / Des Moine. A little less than 1,500,000 gallons are sold.

-SQL:

```
SELECT store_name, store_location, ROUND(SUM(volume_sold_gallons),2) AS gallons_sold
FROM `bigquery-public-data.iowa_liquor_sales.sales`
GROUP BY store_name, store_location
ORDER BY gallons_sold DESC
```

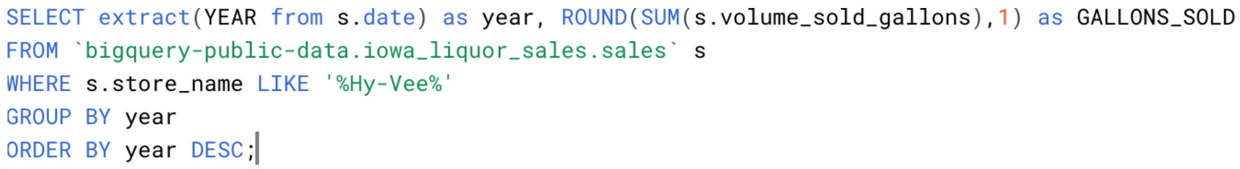
-Data Visualization:



(3) Suppose we are a group of analytics hired by the Hy-Vee corporation to see the amount of liquor sold in the affiliated stores across Iowa each year, and we want to see how they change across the past years.

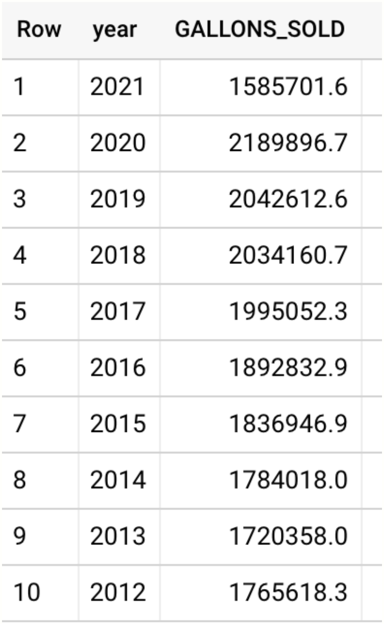
-Insight: As we can observe from the result, each year the liquor sold in the Hy-Vee stores are gradually increasing (Since 2021 have not yet ended, and the data only counted until August, 2021, we would exclude 2021 for now).

-SQL:



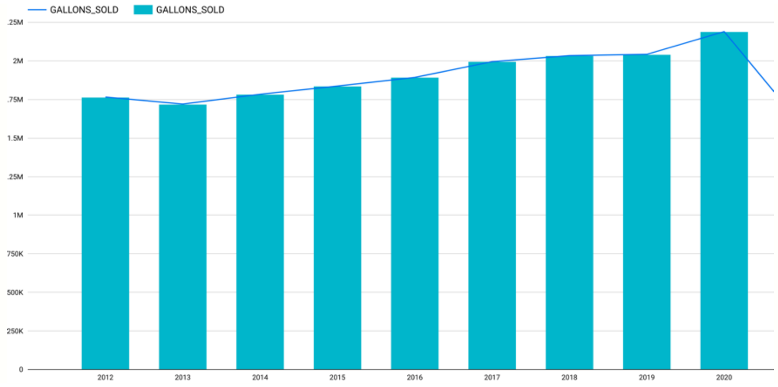
```
SELECT extract(YEAR from s.date) as year, ROUND(SUM(s.volume_sold_gallons),1) as GALLONS_SOLD
FROM `bigquery-public-data.iowa_liquor_sales.sales` s
WHERE s.store_name LIKE '%Hy-Vee%'
GROUP BY year
ORDER BY year DESC;
```

-Outcome:



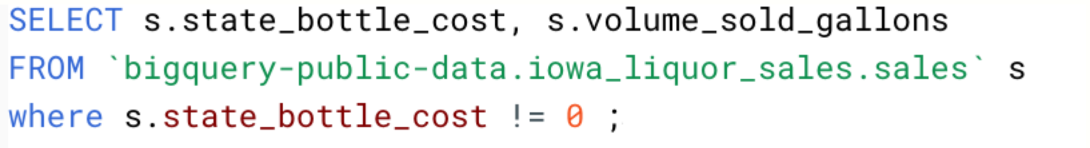
Row	year	GALLONS_SOLD
1	2021	1585701.6
2	2020	2189896.7
3	2019	2042612.6
4	2018	2034160.7
5	2017	1995052.3
6	2016	1892832.9
7	2015	1836946.9
8	2014	1784018.0
9	2013	1720358.0
10	2012	1765618.3

-Data Visualization:

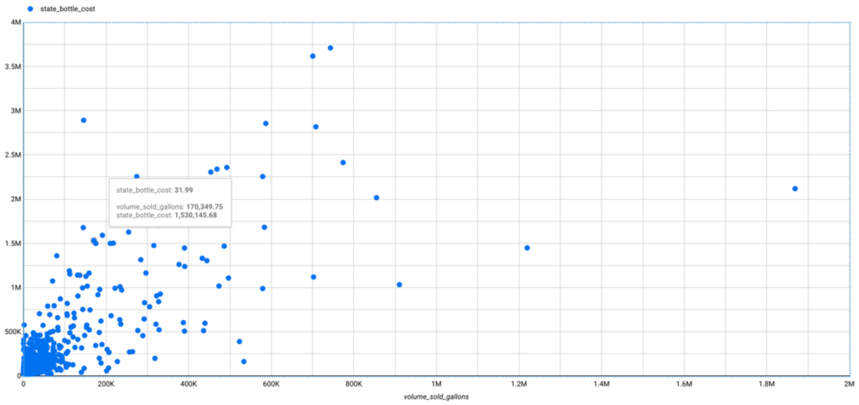


(4) Suppose I am interested in statistical analysis and want to know if any particular predictor can possibly be significant or if any interaction may possibly exist between the predictors, the Big Query can help a lot too.

-Let the response variable be gallons of liquor sold, and the predictor be the state_bottle_cost:

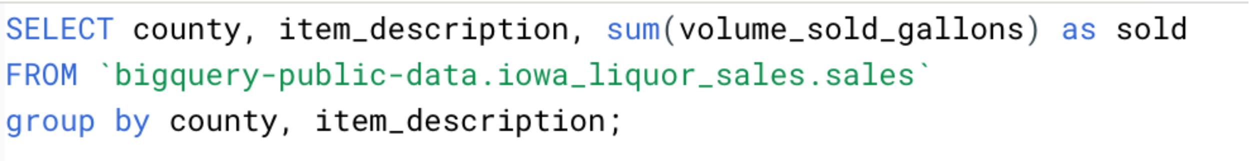


```
SELECT s.state_bottle_cost, s.volume_sold_gallons
FROM `bigquery-public-data.iowa_liquor_sales.sales` s
where s.state_bottle_cost != 0 ;
```

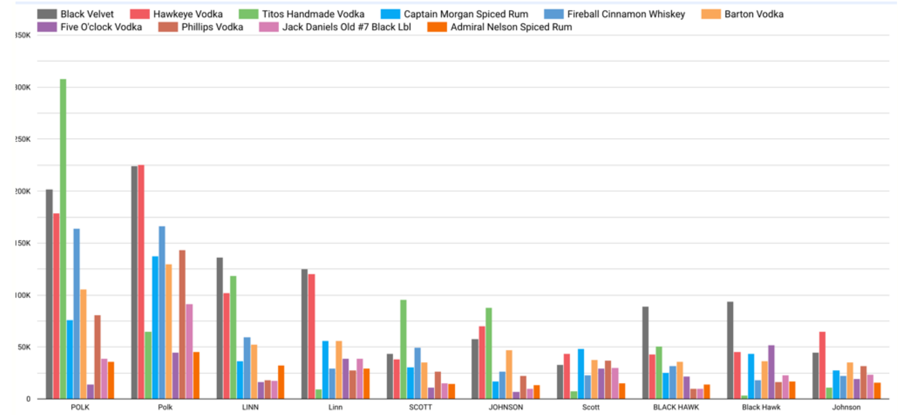


-Insight: There is no trend for the relationship between the amount of liquor sold and the price of the alcohol division paid for each bottle. Therefore state_bottle_cost may not be a significant predictor of my model.

-For the interaction, suppose we want to see if the liquor sold vs. county differs by the product of liquor, implying that the interaction exists:



```
SELECT county, item_description, sum(volume_sold_gallons) as sold
FROM `bigquery-public-data.iowa_liquor_sales.sales`
group by county, item_description;
```



-Insight: The trend of sold vs. county by item_description differs, so interaction may exist between county and the product of liquor. However, there is not enough data in some splitting categories, so this interaction term might not be significant, therefore may not be included in the final model.