Iowa Liquor Sales Analysis Using HiveQL in Oracle Cloud: BDCE

Taimoor Khan, Eddie Castanon, Shuvadeep Kundu, Ankita Paul

Department of Information Systems, California State University,

Los Angeles

e-mail: [tkhan9@calstatela.edu](mailto:tkhan9@calstatela.edu) , [ecastan@calstatela.edu](mailto:ecastan@calstatela.edu) , [skundu@calstatela.edu](mailto:skundu@calstatela.edu) , apaul11@calstatela.edu

**Abstract:** Large scale data and its corresponding analytical tools have made it increasingly possible to make relational discoveries. These discoveries can impact various sectors from business, healthcare, government, education and much more. These behemoth data in existence are now able to be dissected and insights made visible. Using various technological tools such as Power BI, Hadoop File System, and Tableau; big data can be used to assert results, forecast possibilities, and provide solutions to an endless array of questions. In the USA and all over the World, companies have started utilizing Big Data Analytics to gain valuable business insight. We will implement the strategies and principles of Big Data Analytics to focus on sales of Liquor through the State of Iowa. The questions to focus on will revolve around Sales Pattern, Trend and any invisible insight that can’t be detected in normal report analysis, through analysis we will try to tackle these questions and gain valuable insight.

**1. Introduction**

Iowa is a state in the Midwestern United States. In this project, we have focused on the sales of Liquor in Iowa State and how it impacted the economy of the state of Iowa. The Alcoholic beverages consist of various types of drinks including likes of Tequila, Gin, Rum, Whiskey, Vodka and others. These analysis help in determining the trends that occur and a lot of valuable insights which help companies to change their strategies and equip themselves accordingly, these analysis help in making Business decisions based on data analytics.

The data set that we have used indicates the sales of Iowa Liquor mostly from year 2012 to 2019, the data set is being updated with real time data but we have not used data for the year of 2020 as it will not have a big enough sample size. The size of the data set we used is 4.13 GB. By performing analysis on this data set, we were able to determine Sales trends over the year, top 10 Counties by sale, top 10 categories that were sold in terms of volume and other important insights as well.

**2. Related Work**

In our efforts to complete our analytics project, we wanted to look for various sources of inspiration and guidance. We also wanted to ensure there was no plagiarism going on. There is currently plenty of works available to the public that document similar throughout projects and analysis. Our project & analysis focuses primarily on a type of sales trend regarding liquor sales in the state of Iowa.

[1] In research conducted by Mulford, Ledolter, and Fitzgerald, an analysis of Iowa’s liquor sales indicated that increased availability in alcohol yielded no change in consumption for Iowans. This research utilized various statistical analysis of sales trends to conclude their insights. Statistical algorithms were implemented to model Iowa’s liquor sales data. [2] In another study published in “Towards Data Science”, a predictive analysis of Iowa’s liquor data was conducted using Python scripting language and utilizing the Pandas library package. The research used dataframes and functions to create a predictive dataset. The prediction was a sales revenue of $292 million (which our data analysis proved this to be incorrect). [3] Our third referenced work is a study on Iowa’s Liquor sales using the data analysis software Rakam. The analysis aimed at looking into which liquor stores sold the most amount of liquor in the state, and which liquor type was the most sold the most. The research indicated that Hy-Vee #3 is the most popular liquor store in Iowa with the highest amount of liquor bottles sold. The most popular liquor bottles sold are Black Velvet and Fireball Cannon, constantly competing for first place. This research utilized the SQL scripting language such as ‘SELECT’, ‘JOIN’ and other queries, which could then be instantly visualized in Rakam’s dashboard to display the insights.

Although there are similar analysis goals all overall relating to sales volume, our research is still different. Our data was analyzed using Hive/Pig on cloud computing. We used Hive to create external tables from our data. We utilized the Hive queries to sort through the millions of lines of data to pull out data for our analysis. Our analysis also differs in the sense that it includes tempo-spatial representation using Excel 3D maps to demonstrate Iowa’s liquor sales throughout counties as the time progresses. The tools we utilized to gain insight greatly differed from the other studies, specifically in the implementation of the type of technology.

**3. Hardware Specifications**

In this project we have used Hadoop Cluster on Oracle Big Data Cloud Platform to extract and transform the Data, to create valuable insights. The specification for the system that we used, are as follows:

|  |  |
| --- | --- |
|  | **Oracle BDCE** |
| * **Memory** – 241.821 GB * **Cluster version** : Hadoop 2.7.1.2.4.2.0-258 * **Storage** – 1003.6 GB * **Nodes** – 3 (1 Master node , 2 Slave nodes) * **OCPU** – 32 * **CPU Speed** – 2.20 GHz * **Hive version**: Apache Hive version 1.2.1000.2.4.2.0-258 | |

**Table 1. Hardware Specifications**

**4. Background/ Existing Work**

Our work is based on already implemented forms of analysis that are widely used and implemented across various industries. Our data includes various types of formats for analysis. We utilized sales trend analysis, spatial analysis, temporal analysis and data visualization methods to assist us in answering our questions.

**4.1 Sales Trend Analysis**

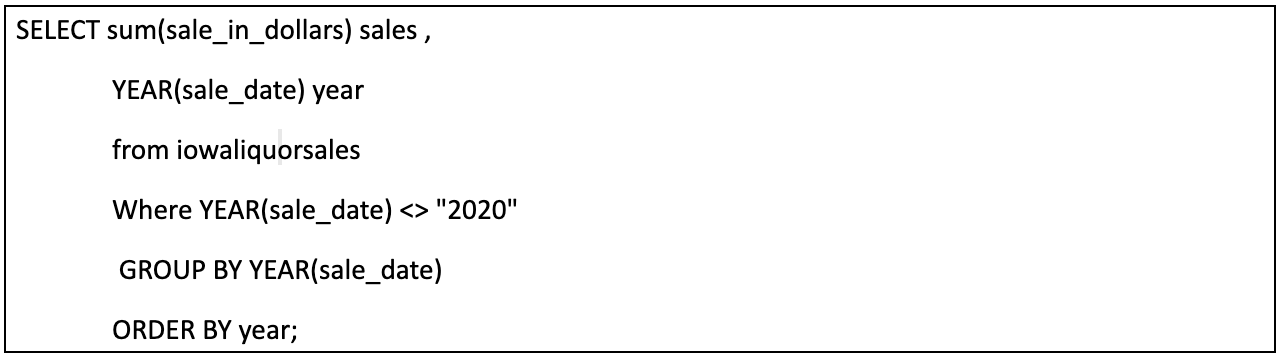
Sales Trend Analysis is the review of historical revenue results to better see patterns. This type of analysis uses financial analysis and budgeting which can indicate revenue growth or decline rates for a company or product. With the various columns and variables that our data set contained, we were able to accurately and efficiently gain insights using sales trend analysis. We were able to pinpoint such sales analysis such as: sales by type of liquor, sales by city, sales by county, sales over time, and much more. With said results, we were able to generate excel files that were then used for another type of analysis explained later in this section. This type of analysis with HiveQL was done in the lab of weblogs file analysis and sensor data analysis in Hive using Oracle Cloud.

**4.2 Temporal & Spatial Analysis**

A combination of Temporal & Spatial analysis were both utilized in our analysis. Historically, this type of analysis is used in data analysis when data is collected along time and space. This data is then used for research to describe phenomena occurring in a particular period at a particular period of time, or through a period of time. Our data did contain coordinate points and did include sales dates and locations across Iowa. Using Hive queries, we pulled out the records of the columns Invoice/Item number, Date of Sale, Longitude, Latitude and Sale Amount for sale record of a particular liquor category *Canadian Whiskies,* over the years. Once the table was created, we extracted the data in CSV file format which was imported over to .xlsx file and we were finally able to create visualizations on Power BI and Excel 3D maps. We did similar work of data extraction, Power BI and 3D Map visualizations in the labs where we analyzed weblog files, sensor data of HVAC systems, twitter sentiments and data of truck events.

**4.3 Hadoop FS/Hive QL/Oracle BDCE**

Oracle BDCE was remotely accessed utilizing the ssh command that was implemented all semester long. We utilized the wget command to access and download our data set from Amazon S3, which we based on the Hive Twitter Sentiment analysis conducted in a prior lab. Utilizing the Hadoop FS, we were able to load our data and create various directories to work from. Utilizing Hive QL, & with the guidance of prior labs, data analysis was performed to create hive queries and hive tables pertaining to our analysis in question. Statements derived from Hive’s SQL-like language allowed for us to script the query results. Hive QL statements were extremely essential in the execution of our insight which is the basis for our visualizations through Excel, Power BI and Tableau.

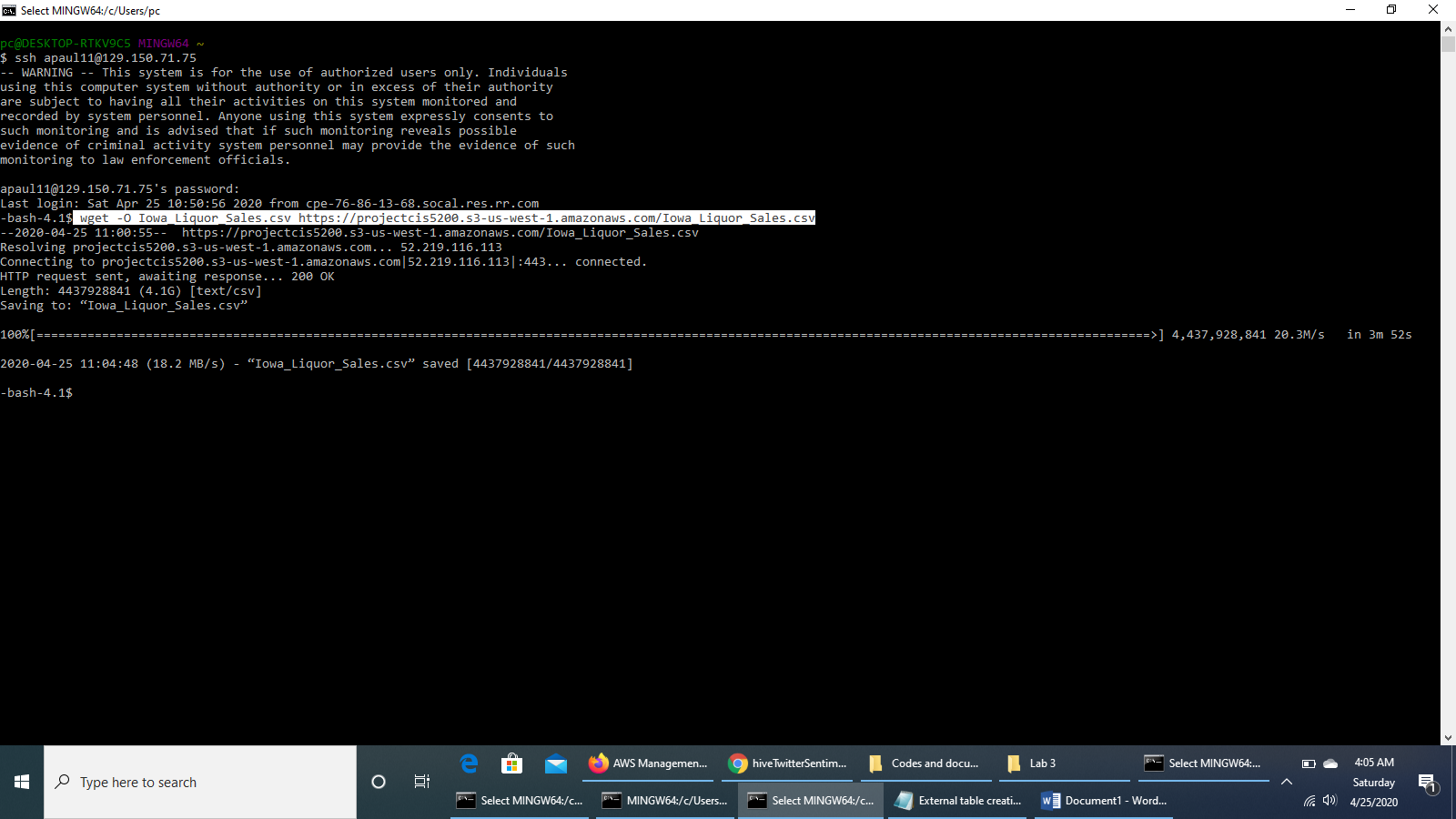


**Figure 1. Hive Query Example**

**5. Our Work**

We implemented models in Oracle Cloud BDCE using Hadoop Cluster. The dataset we used is related to Iowa Liquor Sales, the dataset we used is 4.13 GB in size. We used Hive Query Language, HDFS and Beeline Command to create Hive tables to generate our desired analysis and create tempo-spatial analysis for the used dataset. We used the full data set since Oracle BDCE enables us to do so.

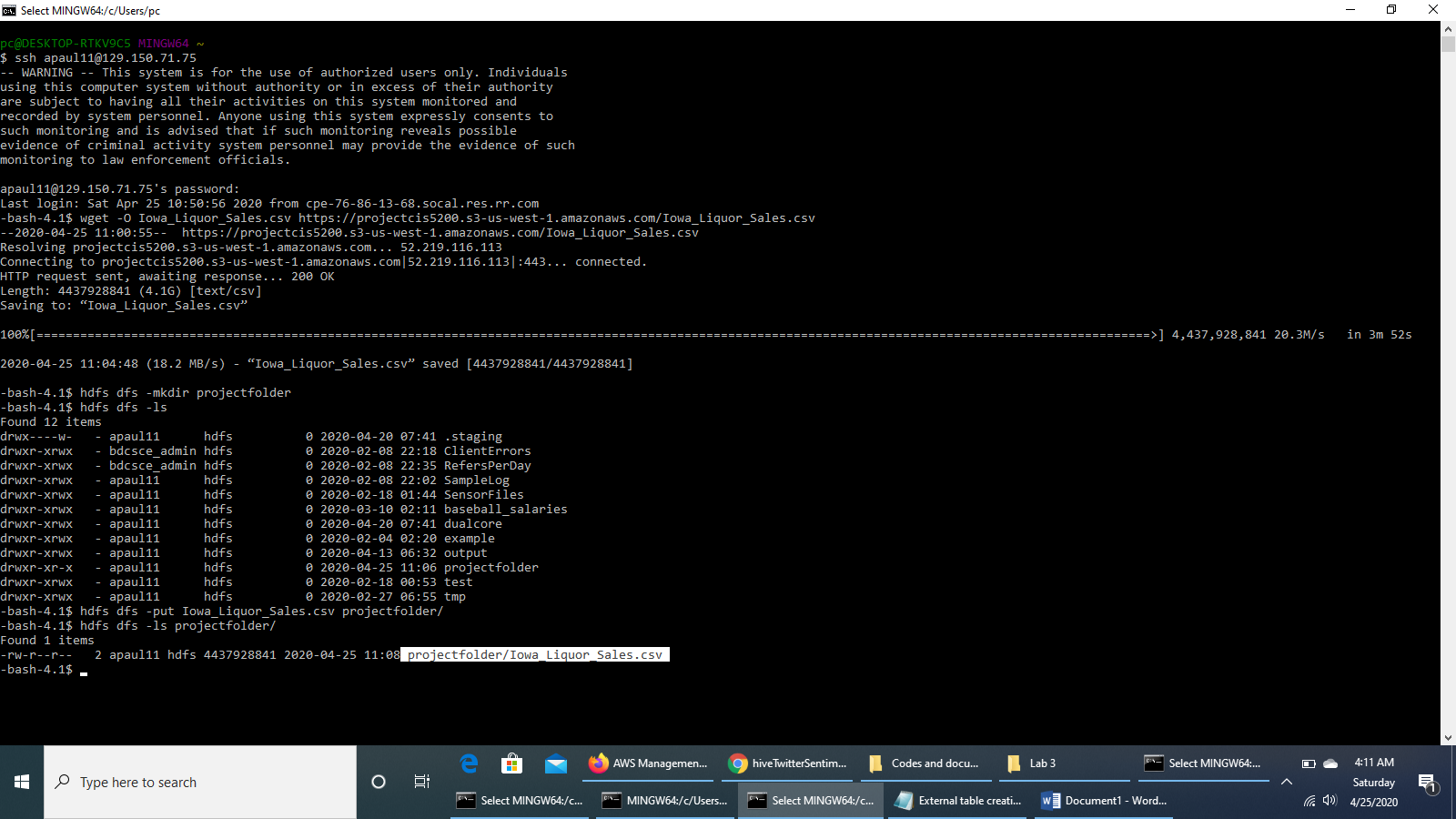
**5.1 Dataset Utilization (Oracle Cloud BDCE)**



**Figure 2. Downloading dataset**

First we linked the Oracle account to SSH to our server at BDCE by using the IP address provided by our Professor. Then, we downloaded the data files ‘iowa\_Liquor\_Sales’ and ‘Cleaned\_location\_Data’, in CSV format from our Amazon S3 repository.

After downloading the data we proceeded to upload it into the Hadoop File System Directory. Then we put the files into two different directories, *projectfolder* and *tempo\_spatial\_data*.

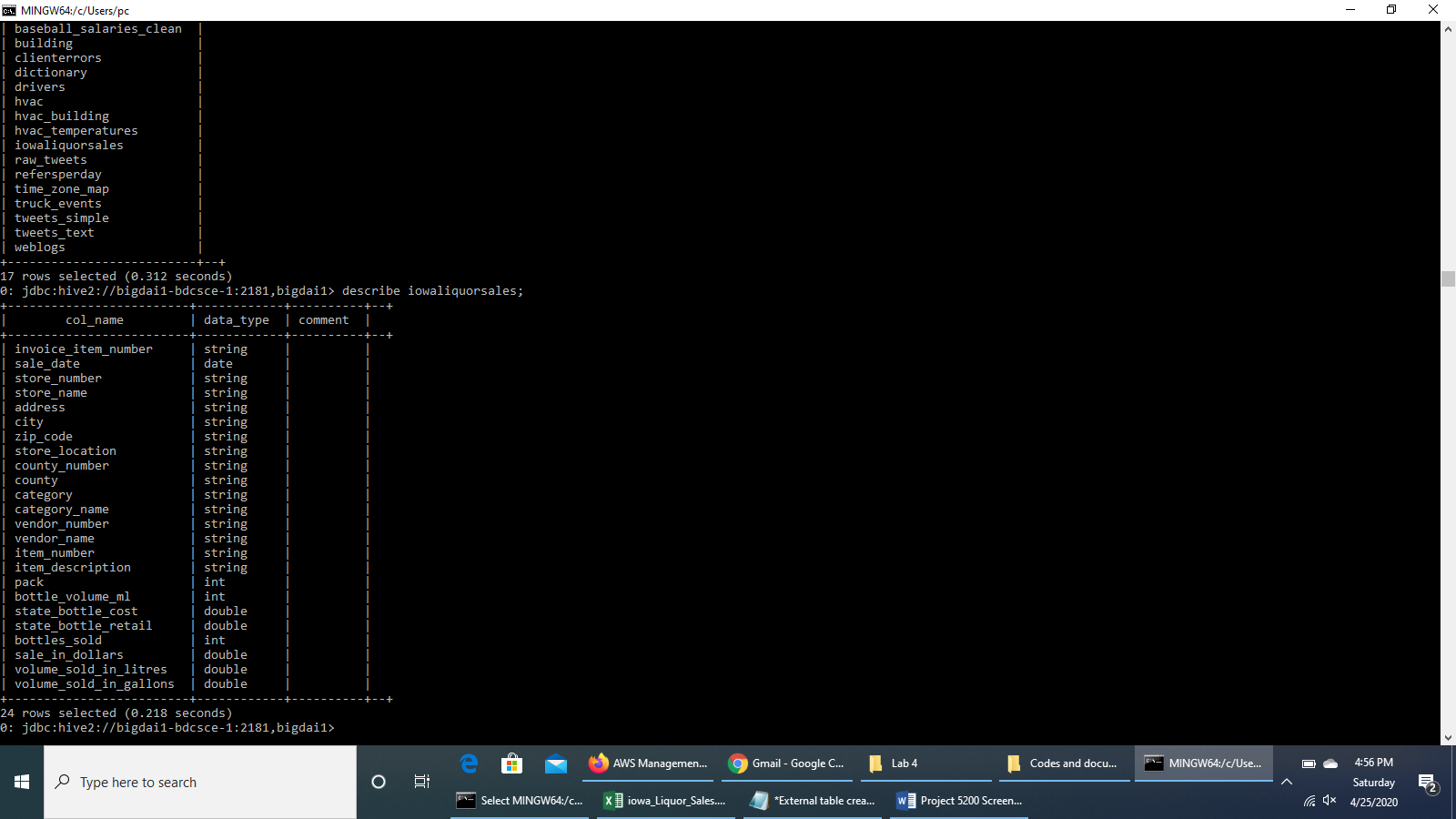


**Figure 3. Dataset Storage**

This option enables us to utilize datasets better in further steps.

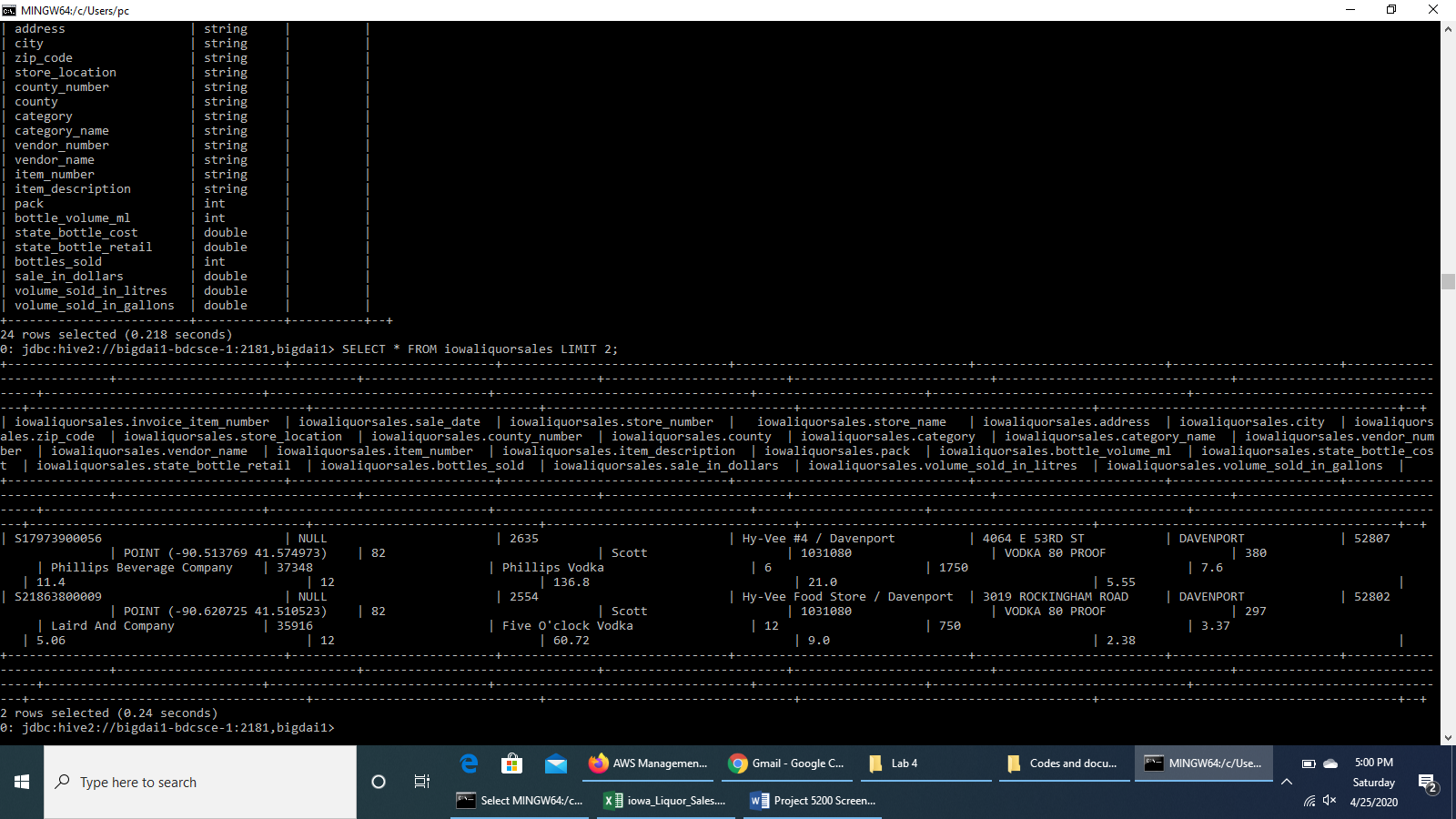
**5.2 External Table Creation**

We created an external table in Hive by name of ‘iowaliquorsales’ and then we checked if it was created properly by using Hive Shell.



**Figure 4. IowaLiqourSales External Table**

We modified this external table to read values in the tables.



**Figure 5. IowaLiqourSales External Table with values**

We altered 4 columns by changing their data types and changed the date format to utilize this dataset in better way.

**5.3 Hive Queries for Analyzing Data:**

We analyzed the dataset by using Hive Queries. First we have used queries that gave us output of sales of liquor in terms of Dollar from year 2012 to 2019.

|  |  |
| --- | --- |
| **Sales** | **Year** |
| 198524922 | 2012 |
| 201545077 | 2013 |
| 205225955 | 2014 |
| 211075281 | 2015 |
| 239263649 | 2016 |
| 286802051 | 2017 |
| 306403425 | 2018 |
| 323272184 | 2019 |

**Table 2. Output for Sales Trend over the years**

This table helped us to analyze the insights in the liquor sales over the years.

The Second Hive Query created was used to sort countries in descending order in terms of sales amount in Dollars and ranked top 10 counties in terms of sales amount.

|  |  |
| --- | --- |
| **Sales** | **County** |
| 467166579 | POLK |
| 173088633 | LINN |
| 150261318 | SCOTT |
| 128496823 | JOHNSON |
| 117662679 | BLACK HAWK |
| 65079282 | STORY |
| 64078347 | WOODBURY |
| 55271223 | DUBUQUE |
| 43276988 | DALLAS |
| 39991152 | POTTAWATTA |

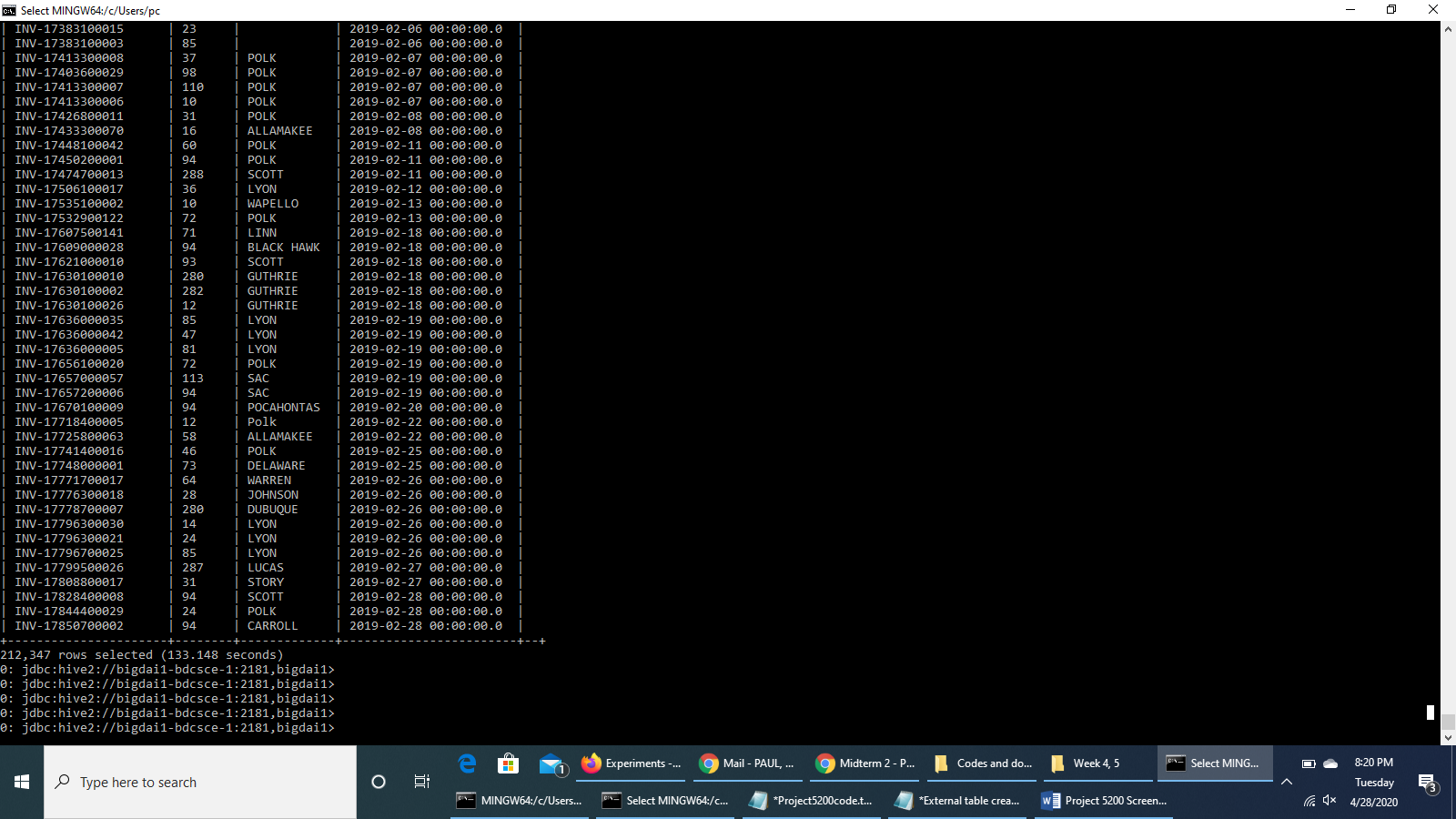
**Table 3: Output for Top 10 counties ranked in terms of Sales**

Next, we used a query to highlight the categories of liquor with highest volume sold.

|  |  |  |
| --- | --- | --- |
| **Volume** | **Sales** | **County** |
| 74956842 | 219044703 | CANADIAN WHISKIES |
| 37087643 | 74699762 | WHISKEY LIQUEUR |
| 30385401 | 45535241 | IMPORTED VODKA |
| 25229906 | 15606657 | PUERTO RICO & VIRGIN ISLANDS RUM |
| 20750759 | 130601226 | VODKA 80 PROOF |

**Table 4. Output for Top 5 Categories of liquor with highest volume sold**

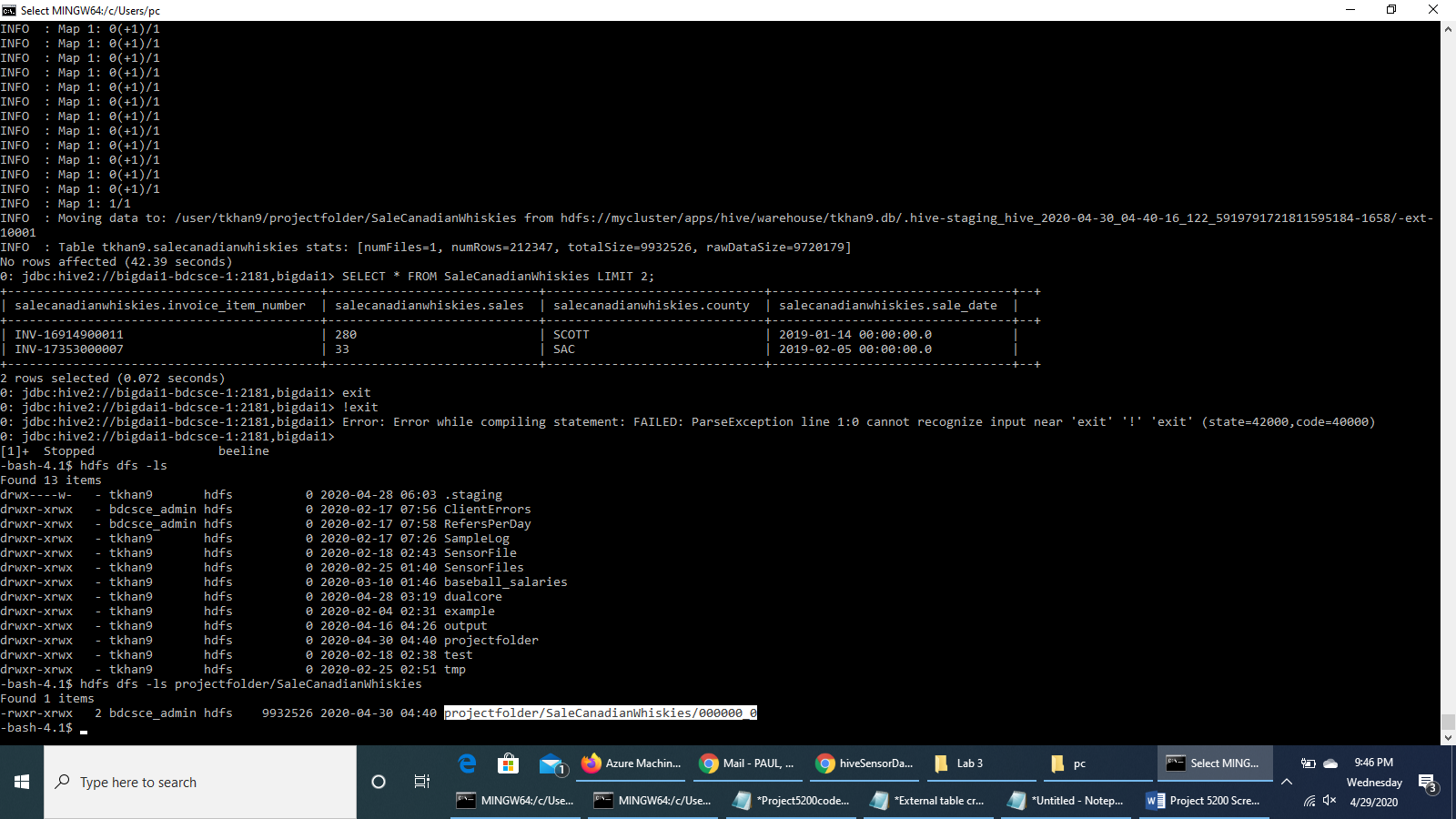
After this we have used Hive Query again to extract the data for ‘Canadian Whiskies’ Liquor sold along with item/invoice number in different counties.



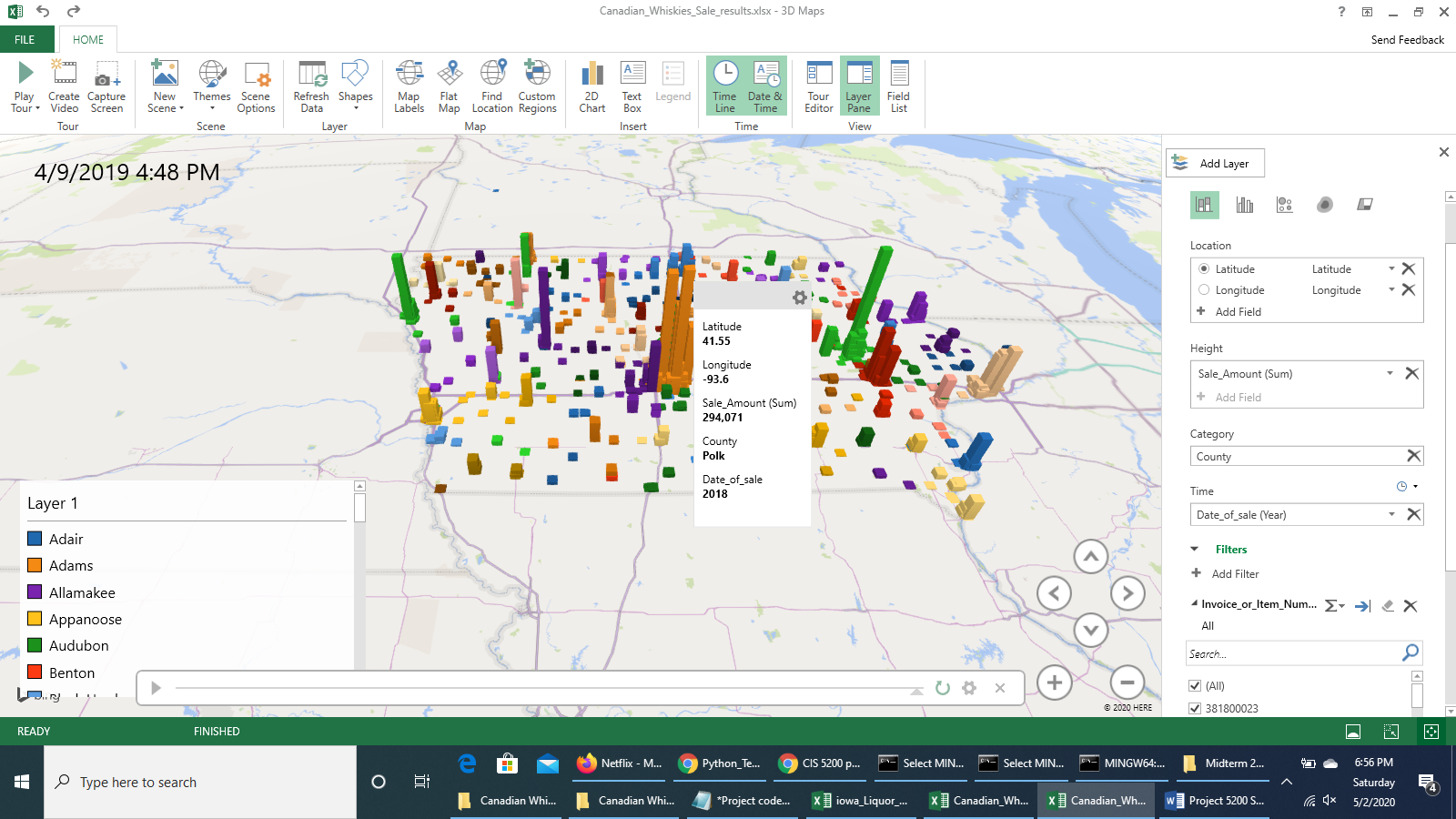
**Figure 6. Output for Sale of Canadian Whiskies in 2019**

**5.4 Output Data Extraction**

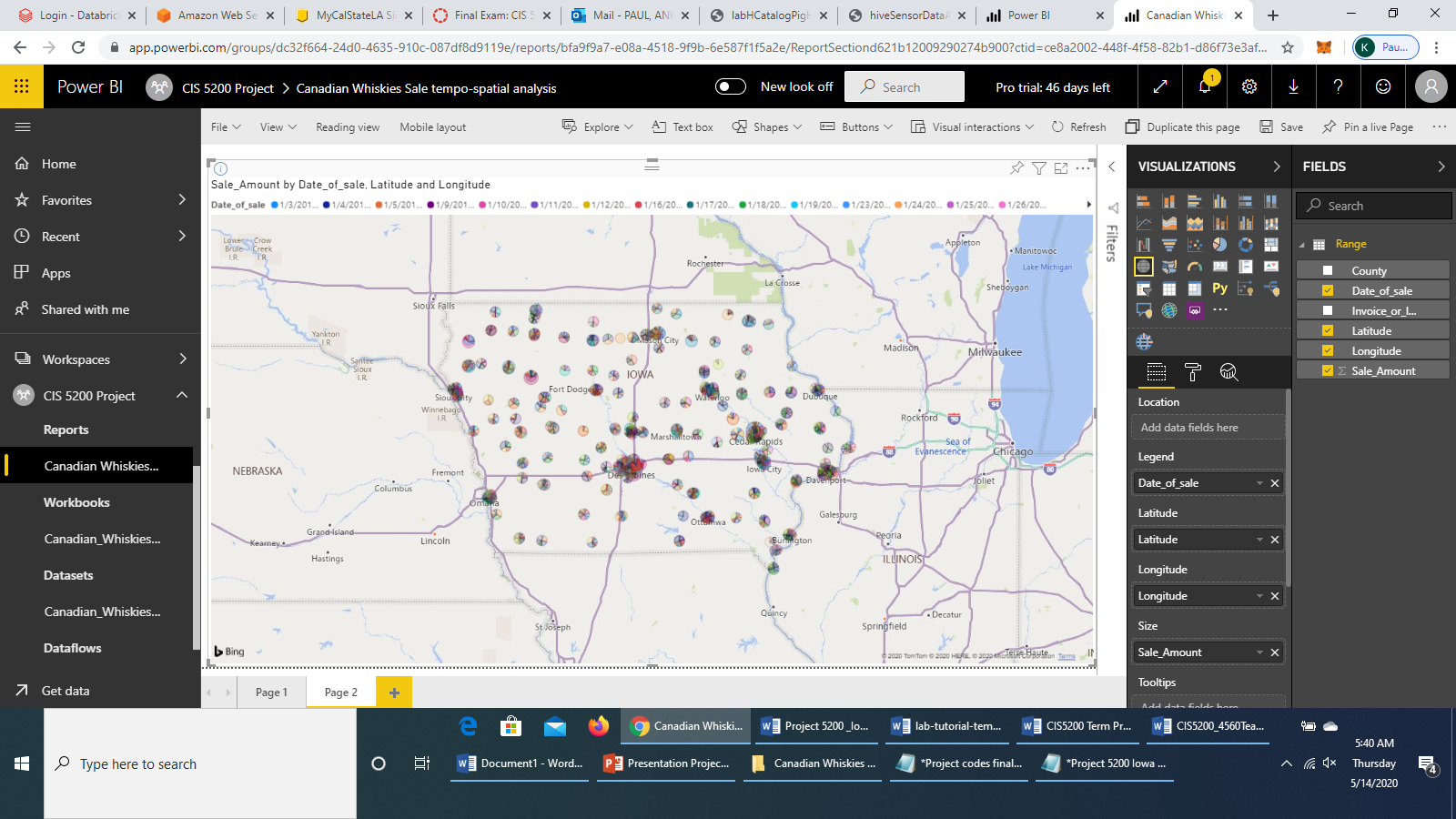
We have created a Hive Table using Hive Queries to extract the output data of the analysis of *Sale of Canadian Whiskies in 2019* to the local computer system.

 **Figure 7. Output Data Extraction**

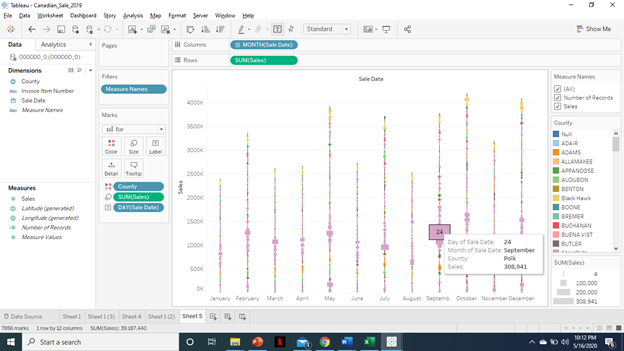
We also extracted the output data of Sale of *Canadian Whiskies* over the years across all the counties, for tempo-spatial analysis. The data was extracted in CSV format and was saved in *.xlsx*, format after including column headers in order to create visuals for deriving valuable business insights. We did Tempo-Spatial Analysis using Excel 3D Map, Visual Analysis using Tableau and Power BI Tool.



**Figure 8. Tempo-Spatial Analysis on Excel 3D Map**



**Figure 9. Tempo-Spatial Analysis on Microsoft Power BI**



**Figure 10. Temporal Analysis of Sale of Canadian Whiskies in 2019 on Tableau**

**6. Conclusion**

Hadoop, Hive, Amazon AWS, Tableau, and Excel are all powerful tools which helped solidify our results and come to our conclusions regarding our dataset. When it came time for us to answer all our previous questions, we were able successfully answer everything we had sought out to find out at the initiation of our research. Through the use of data visuals, we were also able to display the results and found out some interesting insights for Iowan’s and their liquor purchasing habits.

**6.1 Questions and Answers:**

1. What do the people of Iowa like drinking the most? Which spirit/liquor is the most popular?

(a) Canadian Whiskies sold the most volume of liquor in Iowa, with a combined generated sales revenue of $219.04 billion (for all years) which was the highest of all liquor sales in Iowa up to 2019. This could be correlated to Iowa’s geographical proximity to Canada.

2. What are the cities or counties in Iowa who purchase the most liquor? What implications does this have?

(a) The city of Des Moines came out as the city with the highest purchasing amount. This could be correlated with Des Moines being Iowa’s most populous city.

(b) Polk County came out on top for sales in all of Iowa’s counties. This might be due to the fact that the county of Polk is also the most populous county in Iowa.

3. What are the top 5 alcoholic categories sold in Iowa?

In order from top to bottom: Canadian Whiskies, Vodka 80 Proof, Whiskies Liqueur, Imported Vodka, and Puerto Rico & Virgin Island Rums.

**6.2 Further Insights**

* Iowans have been progressively purchasing more years ever since the commencement of the 2015 fiscal year. Our data shows that from 2012 through 2014, liquor sales remained stagnant. However, starting 2015, sales started to rise with 2019 being the highest year on record with a total sales revenue of $323 billion.
* Our data did show the top 5 liquor choices for Iowans. From those top choices, we found out that Vodka 80 proof came in 2nd place. When comparing the volume (in liters) of Vodka 80 proof to the remaining 3 choices, we found out that Whisky Liqueur, Imported Vodkas and Puerto Rico & Virgin Island Rums sold more volume of actual product than Vodka 80 proof. How is that possible? The conclusion we came up with was that it might be possible that Alcohol 80 Proof might be selling less product for a higher price, resulting in more profit.

### References

[1] Mulford, H & Ledolter, Johannes & Fitzgerald, J. (1992), “Alcohol availability and consumption: Iowa Sales Data Revisited”, *Journal of studies on alcohol*. 53. 487-94. 10.15288/jsa.1992.53.487.

[2] Michael Salmon, “Predictive Modeling with Iowa State Liquor Sales Data”, [Online] Available on:

https://towardsdatascience.com/predictive-modeling-with-iowa-state-liquor-sales-data-e45342081b83

[3] “Analyzing Liquor Sales in Iowa” · Rakam.io Blog

Berke Ozenc-Burak Kabakci-Can Ozuysal - https://rakam.io/blog/analyzing-liquor-sales-iniowa/

[4] lab2HiveLogAnalysis\_Oracle.pdf

[5] hiveSensorDataAnalysisLab\_oracle.pdf

[6] hiveTwitterSentimentAnalysisLab\_oracle.pdf

[7] Hive, Chapter 11, *Cloudera*

[8] Original dataset url: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy

[9] Amazon S3 link to iowa\_Liquor\_Sales.csv: https://projectcis5200.s3-us-west-1.amazonaws.com/iowa\_Liquor\_Sales.csv

[10] Amazon S3 link to Cleaned\_location\_data.csv: https://projectcis5200.s3-us-west-1.amazonaws.com/Cleaned\_location\_data.csv

[11] Youtube Link to 3D Map Video capture: https://youtu.be/nYV3qjIAUEI

[12] Link of visualization on Power BI: https://app.powerbi.com/groups/dc32f664-24d0-4635-910c-087df8d9119e/reports/bfa9f9a7-e08a-4518-9f9b-6e587f1f5a2e?ctid=ce8a2002-448f-4f58-82b1-d86f73e3afdd