MapReduce Sales Analysis

Tiyasa Mukherjee

1. Problem Statement

The liquor industry plays a crucial role in the retail economy, especially in regions where sales are closely monitored and regulated. For businesses operating in this space, gaining a clear understanding of sales patterns is essential to staying competitive, aligning supply with demand, and running efficient operations.

In this context, we are analysing granular liquor sales data spanning from 2020 to 2025.

The goal is to uncover meaningful trends related to consumer behaviour, regional performance, and product preferences. These insights will support data-driven strategies aimed at optimizing inventory, maximizing profitability, and improving overall customer satisfaction.

2. Dataset Overview

The dataset used in this analysis is a comprehensive liquor sales record spanning 2020 to 2025, comprising over 4.4 GB of structured sales data. It includes detailed transactional entries from multiple liquor stores across different states, covering:

- Store metadata (e.g., *store_number, store_name, city, county*)
- Product & vendor details (e.g., *category_name, vendor_name, item_description*)
- Financials & volumes (e.g., *state_bottle_cost, bottles_sold, sale_dollars, volume_sold_liters*)
- Time-based fields (*date, year, month*, etc.)

3. Approach Overview

We adopted a systematic, end-to-end pipeline for ingestion, processing, and analysis.

1. **Data Cleaning and Preprocessing**: The dataset was loaded in chunks, cleaned for missing values, renamed for consistency, and enhanced with derived date fields for better temporal analysis.
2. **Data Ingestion**: The cleaned data was ingested into AWS RDS (MySQL) and subsequently transferred to HBase using Apache Sqoop with appropriate schema and structure.
3. **Batch Analysis Using MapReduce (MRJob):** The processed dataset was analyzed using MRJob classes in Python to compute revenue metrics, performance rankings, and sales trends.
4. **Recommendations:** Present the findings and data-driven strategies for optimizing store operations, vendor management, and product promotion.

3.1. Data Cleaning and Preprocessing

The original liquor sales dataset containing transaction-level information from 2020 to 2025. Due to its size (~4.4 GB), the dataset was loaded in chunks using Python's `pandas` library for efficient memory management. Note: I've considered the first 10,000 rows for the analysis

Several preprocessing steps were carried out:

- **Column Renaming:**
  To improve readability and consistency, columns were renamed.

  For example:

  - *state_bottle_retail → bottle_price*

- *state_bottle_cost → bottle_cost*
- *sale_dollars → sale_dollars_total*
- *volume_sold_liters → liters_sold*
- *volume_sold_gallons → gallons_sold*
- *item_description → product_name*

- **Missing and Null Value Handling:**
  Columns such as *vendor_name, category_name*, and *bottle_volume_ml* were checked for null values. Rows with critical missing data were dropped to preserve the integrity of the analysis.
- **Duplicate Removal:**
  Duplicate entries were identified using *invoice_item_number* and other transaction fields. These were removed to ensure accurate aggregations in later stages.
- **Inconsistency Fixes:**
  Inconsistent naming patterns in columns like *vendor_name* and *category_name* (e.g., mixed case, extra spaces) were standardized using string cleaning operations.
- **Date Parsing and Feature Engineering:**
  The *date* field was parsed into datetime format, and new fields were created:
  - *year*
  - *month*
  - *day*
  - *day_of_week*
- **Outlier Detection and Treatment:**
  Numeric columns like *bottles_sold, bottle_price*, and *sale_dollars_total* were evaluated for extreme outliers using statistical thresholds (e.g., IQR). Outliers were retained only if they reflected actual high-volume transactions.
- **Final Output Files:**
  - Cleaned CSV: `Liquor_Sales_df_new1.csv`
  - MRJob Input: `liquor_sales.txt` (tab-separated version for Hadoop/MapReduce compatibility)
- **Processed Dataset View:**

invoice_item_number	date	store_number	store_name	address	city	zip_code	county_number	county
S24127700024	2015-02-19	3678	Smoke Shop, The	1918 SE 14TH ST	DES MOINES	50320.0	77.0	Polk
S19323500030	2014-06-03	2607	Hy-Vee Wine and Spirits / Shenandoah	520 SO FREMONT	SHENANDOAH	51601.0	73.0	Page
S23334500013	2015-01-06	4810	Kum & Go #518 / Ankeny	3603 NE OTTERVIEW CIRCLE	ANKENY	50021.0	77.0	Polk
S15034600007	2013-10-09	4583	Kum & Go #5100 / Manson	208 MAIN ST	MANSON	50563.0	13.0	Calhoun
S25185100053	2015-04-21	5080	C's Liquor Store	719 2ND AVE W	SPENCER	51301.0	21.0	Clay
S26178600169	2015-06-15	2506	Hy-Vee #1044 / Burlington	3140 AGENCY	BURLINGTON	52601.0	29.0	Des Moines
S11599200028	2013-04-11	2630	Hy-Vee Drugstore #2 / WDM	1010 60TH ST	WEST DES MOINES	50266.0	77.0	Polk
S14039300026	2013-08-21	3916	Smokin' Joe's #5 Tobacco and Liquor	1115 ALBIA RD	OTTUMWA	52501.0	90.0	Wapello
S14777200004	2013-09-25	4073	Uptown Liquor, Llc	306 HWY 69 SOUTH	FOREST CITY	50436.0	95.0	Winnebago
S28698700004	2015-10-27	2578	Hy-Vee / Charles City	901 KELLY ST	CHARLES CITY	50616.0	34.0	Floyd
S15825800006	2013-11-20	4465	HOME TOWN FOOD ON 4	714 S EAST ST	POMEROY	50575.0	13.0	Calhoun
S11716100052	2013-04-17	3990	Cork and Bottle / Oskaloosa	309 A AVE WEST	OSKALOOSA	52577.0	62.0	Mahaska
S28977300019	2015-11-10	4743	No Frills Supermarkets #791 / Counci	1801 VALLEY VIEW DR	COUNCIL BLUFFS	51503.0	78.0	Pottawattamie

category	category_name	vendor_number	vendor_name	item_number	item_description	pack	bottle_volume_ml	state_bottle_cos
1031200.0	Vodka Flavored	380	Phillips Beverage Company	41783	Uv Blue Raspberry Vodka Mini	6	500	4.8
1062200.0	Puerto Rico & Virgin Islands Rum	434	Luxco-St Louis	45277	Paramount White Rum	12	1000	4.3
1062200.0	Puerto Rico & Virgin Islands Rum	35	Bacardi U.S.A., Inc.	43121	Bacardi Superior Rum Mini	12	500	5.5
1081200.0	Cream Liqueurs	305	Mhw Ltd	73050	Rumchata	6	750	12.
1081390.0	Imported Schnapps	421	Sazerac Co., Inc.	69713	Dr. McGillicuddy's Peach Mini	12	500	4.9
1081600.0	Whiskey Liqueur	260	Diageo Americas	66206	Piehole Cherry Pie Mini	12	500	4.
1071100.0	American Cocktails	395	Proximo	58838	Jose Cuervo Authentic Lime Margarita	6	1750	8.
1012100.0	Canadian Whiskies	260	Diageo Americas	11294	Crown Royal Canadian Whisky	24	375	7.6
1012100.0	Canadian Whiskies	55	Sazerac North America	12407	Canadian Ltd Whisky	12	1000	5.
1031080.0	Vodka 80 Proof	260	Diageo Americas	37426	Popov Vodka 80 Prf Traveler	12	750	4.
1011100.0	Blended Whiskies	434	Luxco-St Louis	24156	Hawkeye Blend Whiskey	12	750	3.3
1032200.0	Imported Vodka - Misc	35	Bacardi U.S.A., Inc.	34436	Grey Goose Vodka L'orange	6	750	17.9
1022100.0	Tequila	395	Proximo	89198	Jose Cuervo Especial Reposado Tequila	6	1750	20.2

state_bottle_cost	state_bottle_retail	bottles_sold	sale_dollars	volume_sold_liters	volume_sold_gallons	year	month	day	day_of_week	longitude	latitude
4.89	7.34	2	14.68	1.0	0.26	2015	2	19	Thursday	-93.597011	41.570844
4.34	6.51	12	78.12	12.0	3.17	2014	6	3	Tuesday	-95.385111	40.761736
5.54	8.31	1	8.31	0.5	0.13	2015	1	6	Tuesday	-93.572458	41.760989
12.5	18.75	6	112.5	4.5	1.19	2013	10	9	Wednesday	-94.534532	42.517855
4.96	7.44	1	7.44	0.5	0.13	2015	4	21	Tuesday	-95.147741	43.14521
4.9	7.35	1	7.35	0.5	0.13	2015	6	15	Monday	-91.136655	40.814666
8.2	12.3	6	73.8	10.5	2.77	2013	4	11	Thursday	-93.790534	41.584979
7.65	11.48	1	11.48	0.38	0.1	2013	8	21	Wednesday	-92.437224	41.009342
5.5	8.25	12	99.0	12.0	3.17	2013	9	25	Wednesday	-93.633306	43.261538
4.5	6.75	12	81.0	9.0	2.38	2015	10	27	Tuesday	-92.67556000000000	43.066993
3.36	5.04	24	120.96	18.0	4.76	2013	11	20	Wednesday	-94.678054	42.542993
17.97	26.96	1	26.96	0.75	0.2	2013	4	17	Wednesday	-92.648153	41.296228
20.25	30.38	6	182.28	10.5	2.77	2015	11	10	Tuesday	-95.81877500000000	41.24394

3.2. Data Ingestion

3.2.1. Setting up RDS instance

To facilitate structured querying and downstream analytics, the cleaned liquor sales dataset was ingested into an AWS RDS MySQL instance.

Below is the complete setup and ingestion workflow:

Infrastructure and Network Setup:

1. **EC2 Instance Setup**:
   - **Instance Type**: `t2.medium` (2 vCPUs, 4 GiB RAM)
   - Used primarily for initial access, key storage, and verification tasks.
   - A custom key pair (`my_key_val_tm.pem`) was generated and used for SSH access.
2. **EMR Cluster Setup**:
   - **EMR Version**: `7.4.0`
     - **Hadoop** – Core distributed storage and processing framework
     - **Hive** – SQL-like interface for querying structured data on HDFS
     - **HBase** – NoSQL database for scalable random-access reads/writes
     - **Sqoop** – For transferring data between RDS (MySQL) and HBase
     - **ZooKeeper** – Coordination service used by HBase for managing distributed nodes
   - **Master Node**: `m4.xlarge`
   - **Network Configuration**:
     - VPC: `vpc-0fe9c1440e2d1e6ad` (shared with RDS instance for communication)
     - Security groups allowed inbound traffic on port `3306` from the EMR's group to the RDS instance.
3. **RDS Instance Setup**:
   - **Engine**: MySQL 8.0
   - **Instance Class**: `db.t3.medium`
   - **Storage**: 20 GiB (General Purpose SSD)
   - **Connectivity**:
     - Publicly accessible
     - Bound to the same VPC as EMR
     - Security group allowed MySQL/Aurora (port `3306`) from EMR's security group

3.2.2. Upload data into RDS via MySQL Workbench

1. Workbench Connection:
   - Connected using the RDS endpoint:

     liquordbinstance.cxzwuat5vlim.us-east-1.rds.amazonaws.com

   - Used standard MySQL credentials (admin, password).

2. Schema Creation**:**
    - The liquor_sales table was created using the given schema.
3. Workbench Settings for Local Import:
    - Enabled local file upload:

      Preferences → SQL Editor → MySQL Session → Enable "Allow LOAD DATA LOCAL INFILE"

4. Ingestion Command:
    - Loaded data using:

      ```
      LOAD DATA LOCAL INFILE
      '/Users/tiyasamukherjee/Downloads/Liquor_Sales_cleaned_df.csv'
      INTO TABLE liquor_sales
      FIELDS TERMINATED BY ','
      ENCLOSED BY '"'
      LINES TERMINATED BY '\n'
      IGNORE 1 ROWS;
      ```

5. Validation Queries:
    - Ensured data was correctly inserted using:

      ```
      SELECT COUNT(*) FROM liquor_sales;
      SELECT * FROM liquor_sales LIMIT 10;
      ```

3.2.3. Upload data into HBase from RDS

1. SSH into EMR Master Node:

   ```
   ssh -i your-key.pem hadoop@<EMR-Master-Public-DNS>
   ```

2. Download MySQL Connector JAR:

   ```
   wget https://downloads.mysql.com/archives/get/p/3/file/mysql-connector-java-5.1.49.tar.gz
   tar -xvzf mysql-connector-java-5.1.49.tar.gz
   cp mysql-connector-java-5.1.49/mysql-connector-java-5.1.49-bin.jar /usr/lib/sqoop/lib/
   ```

3. Configure Networking:
    - EMR and RDS were placed in the **same VPC (**vpc-0fe9c1440e2d1e6ad**)**.
    - RDS security group was updated to **allow inbound MySQL (port 3306)** access from EMR's security group.
    - **Public endpoint** of RDS was used (liquordbinstance.cxzwuat5vlim.us-east-1.rds.amazonaws.com).
4. Run Sqoop Import to HBase:

   ```
   sqoop import \
     --connect jdbc:mysql://liquordbinstance.cxzwuat5vlim.us-east-
   1.rds.amazonaws.com:3306/liquor_sales \
     --username admin \
     --password <your_password> \
     --table liquor_sales \
     --hbase-table liquor_sales_hbase \
     --column-family salesinfo \
     --hbase-row-key invoice_item_number \
     --split-by store_number \
     --driver com.mysql.jdbc.Driver
   ```

3.3. Batch Processing Using MapReduce (in Colab using MRJob)

We performed parallel processing of the cleaned liquor sales data using the **MRJob framework in Google Colab**, simulating a MapReduce environment. The analysis was based on the `/content/liquor_sales.txt` file (tab-separated) derived from the cleaned CSV.

The following insights were generated:

Total Revenue by Store:

The top revenue-generating stores include:

- Store `4209`: $1,733.64
- Store `4214`: $1,386.33
- Store `4233`: $987.21
  These stores indicate consistent high-volume sales and should be prioritized for stock replenishment and promotional activities.

Top-Selling Liquor Categories

Most sold categories (by volume) were:

- Vodka 80 Proof: 10,791 bottles
- Vodka Flavored: 3,090 bottles
- American Grape Brandies: 2,192 bottles
- Whiskey Liqueur: 2,224 bottles
- Tequila: 2,157 bottles
  These categories are ideal candidates for targeted promotions and bulk stocking.

County-Level Sales Analysis

Counties showing high sales performance include:

- **Scott**: $47,299.09
- **Story**: $23,986.87
- **Webster**: $10,018.31
  Note: County name inconsistencies (e.g., "SCOTT" vs. "Scott") were handled during preprocessing.

Store Performance Analysis

We analyzed store performance on revenue and volume:

- Store `4209`: 168 bottles, $1,733.64
- Store `4214`: 109 bottles, $1,386.33
- Store `4233`: 63 bottles, $987.21
  These stores show strong throughput per transaction and can benefit from larger inventories.

Trends in Liquor Sales Over Time

Time-based sales analysis revealed peak periods in:

- **May 2012**: $19,938.20
- **May 2015**: $21,360.29
- **July 2012**: $17,178.14
  The months of **May–August** typically see a surge in sales, indicating seasonal demand spikes.

Vendor Performance

Top revenue-generating vendors include:

- **Bacardi U.S.A., Inc.**: $42,983.14
- **Brown-Forman Corporation**: $42,048.40
- **Sazerac North America**: $26,396.28
  Vendors with consistent high sales should be engaged for strategic partnerships and volume discounts.

4. Recommendations (Based on 10,000 Rows)

- **Optimize Store Inventory**: Focus on top-performing stores like 4209 and 4214 for inventory expansion.
- **Boost Promotions**: Highlight best-selling categories like Vodka and Tequila in promotional campaigns.
- **Seasonal Stocking**: Increase stock around May–August based on observed historical peaks.
- **Vendor Engagement**: Strengthen partnerships with vendors like Bacardi and Brown-Forman for better pricing and supply chain benefits.
- **Data Consistency**: Resolve inconsistent naming (e.g., counties and vendors) for accurate reporting in full-dataset analysis.