Liquor Sales Analysis Report

Truong Cong Quoc Thang, Vishal Sharma, Sudeep Sidhu

Prepared by: Data Engineer – Hadoop & MapReduce Module

Dataset: Liquor_Sales.csv (about 20 million records)

1. Objective of Analysis

The purpose of this analysis is to evaluate liquor sales performance across different dimensions using distributed batch processing with Hadoop & MapReduce. Key goals include:

- Calculating total revenue by store
- Identifying top-selling liquor categories
- Analyzing regional sales by county
- Assessing store and vendor performance
- Detecting sales trends over time

2. Processing Methodology

The cleaned dataset was ingested from HBase and processed using Python's mrjob framework on a local MapReduce runner. Six jobs were implemented to extract insights:

1. Total Revenue by Store – MRTotalRevenueByStore: Calculates sales by store
2. Top-Selling Categories – MRTopSellingLiquorCategories: Aggregates bottles sold by item
3. County-Level Analysis – MRCountyLevelSalesAnalysis: Summarizes sales and volume by county
4. Store Performance – MRStorePerformanceAnalysis: Measures bottle count, revenue, transaction count, and average sale per transaction
5. Sales Trends Over Time – MRLiquorSalesTrends: Intended to track monthly revenue, but no valid date records were available
6. Vendor Performance – MRVendorPerformance: Evaluates supplier efficiency via volume, revenue, and transaction metrics

Intermediate results were written to .txt files and visualized in Jupyter Notebook for interpretation.

3. Analysis Results

3.1. Total Revenue by Store

Store performance varies significantly, with ALGONA leading at **$4.13M** and ALTA VISTA trailing at **$28.5K**.

3.2. Top-Selling Categories

The most sold liquor was White & Berry with **97,398 bottles**, followed by Peppermint Schnapps and White & Berry Mini. Flavored and seasonal items dominate the top ranks.

3.3. County-Level Sales

Anomalously high sales appear under counties labeled "Inc.", "LLC" — suggesting vendor names were mistakenly parsed into the county field. For example:

- "Inc.": **$217.28M** from **16.84 million bottles**

This indicates a need for data cleaning on geographic fields.

3.4. Store Performance

- ALGONA: Highest transaction count (**5,922**) and total revenue
- ALTA VISTA: Highest average sale per transaction (**$158.92**) despite low total revenue

Performance varies with some stores prioritizing volume, others average value per transaction.

3.5. Sales Trends Over Time

No monthly trend could be generated due to missing or malformed date fields in the dataset.

3.6. Vendor Performance

- DIAGEO AMERICAS: Highest total sales **($1.41M)** and bottle count
- Anchor Distilling: Highest average sale per transaction **($1,386.21)** from only 4 transactions

Some vendors show high individual sale values despite low volume.

4. Insights & Recommendations

- **Clean geographic fields** (County) and ensure valid Date formatting to improve temporal and regional analysis.
- **Focus inventory on top-selling items** like White & Berry, especially in lower-performing counties.
- **Retain strong vendor partnerships** (e.g. DIAGEO AMERICAS) and monitor high-value niche suppliers.

- **Implement store-level dashboards** to track transactional metrics and optimize marketing strategies.