Final Assignment Python Data: Liquor Sales Analysis

Iraklis Marios Papigkiotis

9/1/2024

Summary

The purpose of this project involved the analysis of Liquor sales across the US in order to find the products with the most sales and the stores that achieved the higher sales. The most successful product found to be DISTILLED SPIRITS SPECIALTY and the most sales were achieved in Iowa in a city with zipcode 52314. Regarding the store with the most sales achieved, this was no other than Wilkie Liquors with approximately 18% of the total sales. A secondary analysis was performed in order to calculate the percentage of sales of the most successful item in each zipcode but no insights came up due to the fact that a great number of stores in the dataset appear to sell only one product and it was either the most famous or a random product making the percentage of sales of this product in those stores skyrocket to 100% while on others to 0%. In order to have a better insight, more data need to be gathered and those data need to be more spreaded in time (e.g from 2010- 2024 and not 2016-2019)

1. Introduction

- **Objectives:**
  - **Data Extraction and Cleaning:** Extract pertinent data and conduct preliminary cleaning to handle missing or inconsistent data.
  - **Data Analysis: Employ Python with Pandas** or SQL to analyze the dataset, identifying the predominant item per zipcode and calculating the proportion of sales for each store between 2016 and 2019
  - **Data Visualization:** Represent the analyzed data aesthetically and informatively using tools like matplotlib, seaborn, **plotly**, or integrate with visualization platforms such as Tableau Public, Power BI, Looker Studio, or HEX for a more interactive experience.

2. Data Description

- **Data Sources:** https://storage.googleapis.com/courses_data/Assignment%20CSV/finance_liquor_sales.csv
- **Data Description:** Liquor Sales data across the US between 2012-2022
- **Data Cleaning:**
  1. Small dataset → Need to preserve the data and not drop anything
  2. Search for inconsistent data
  3. Replace null values
  4. Ensure no data loss and healthy data

3. Methodology

- **Use of Pandas** for versatile manipulation of the data frame
- **Use of plotly** for the creation of interactive and informative graphics

4. Analysis

After cleaning the inconsistent data and replacing them in order to prevent data loss I moved straight to the objectives as the dataset was complicated and could not observe any patterns without grouping data and visualize them.

- **Objective 1:** Grouped all the data based on zipcode and product in order to sum the bottles sold and find the most successful item in each zipcode
- **Objective 2:** Grouped all the data based on store, summed them and calculated the percentage of sales of each store
- **Visualizations of Objectives 1,2 emerge after compiling the final_assignment_papigkiotis.py**
- **(Optional) Objective 3:** Grouped and merged different data frames in order to calculate the percentage of sales of the most successful item in each zipcode of each store

5. Results

- **Key Findings:**
  - **Most successful product:** DISTILLED SPIRITS SPECIALTY
  - **Store with the most percentage sales:** Wilkie Liquors achieved **18.2%** of the total sales. **Location**: Iowa, **Zipcode**: 52314
- **Visualizations:** compile '**final_assignment_papigkiotis.py'**
- **Interpretation:** The outcome of my analysis showed that Wilkie Liquors is the dominant store in the US and the big percentages of sales that emerge are due to the sale of DISTILLED SPIRITS SPECIALTY. This product is a traditional type of whiskey produced by them which also makes this store the main supplier in the market. Apart from that, the distribution of products in the market varies only slightly from zipcode to zipcode and the sales produced by each store follows the same distribution. Now, regarding the distribution of the percentage of sales of the most successful product in each zipcode and store, no insight could be extracted due to the small dataset. In order to perform such visualization the dataset should be spreaded more in time and to have more entries as well as a great number of stores sold only one product.