Workearly: Python Data Bootcamp Course - Final Assignment

by Giorgos Fragkiadakis

Project Description

This project is designed to simulate a full workflow of a Data Analyst from getting data off the Database to manipulate it with the use of Python and Pandas module to present it through matplotlib module or Tableau.

The concept is that we are given a dataset that contains Liquor Sales in the state of Iowa in USA between 2012-2020 and we are asked to find the most popular item per zip code and the percentage of sales per store in the period between 2016-2019.

We are also asked to visualize the Data and present them in either a matplotlib format or in Tableau Public.

Every calculation and transformation of Data has to happen through a Python Script.

Project Implementation

- **Step 1.**

I loaded Dataset "**finance_liquor_sales.sql**" to MySQL Workbench and run it.

- **Step 2.**

In order to get all the columns of the table between the years 2016-2019 I run the following query:

```
SELECT *
FROM liquorsales.finance_liquor_sales
WHERE year(date) between 2016 and 2019
ORDER BY date;
```

- **Step 3.**

I exported the data to the "**finance_liquor_sales(2016_2019).csv**" file.

- **Step 4.**

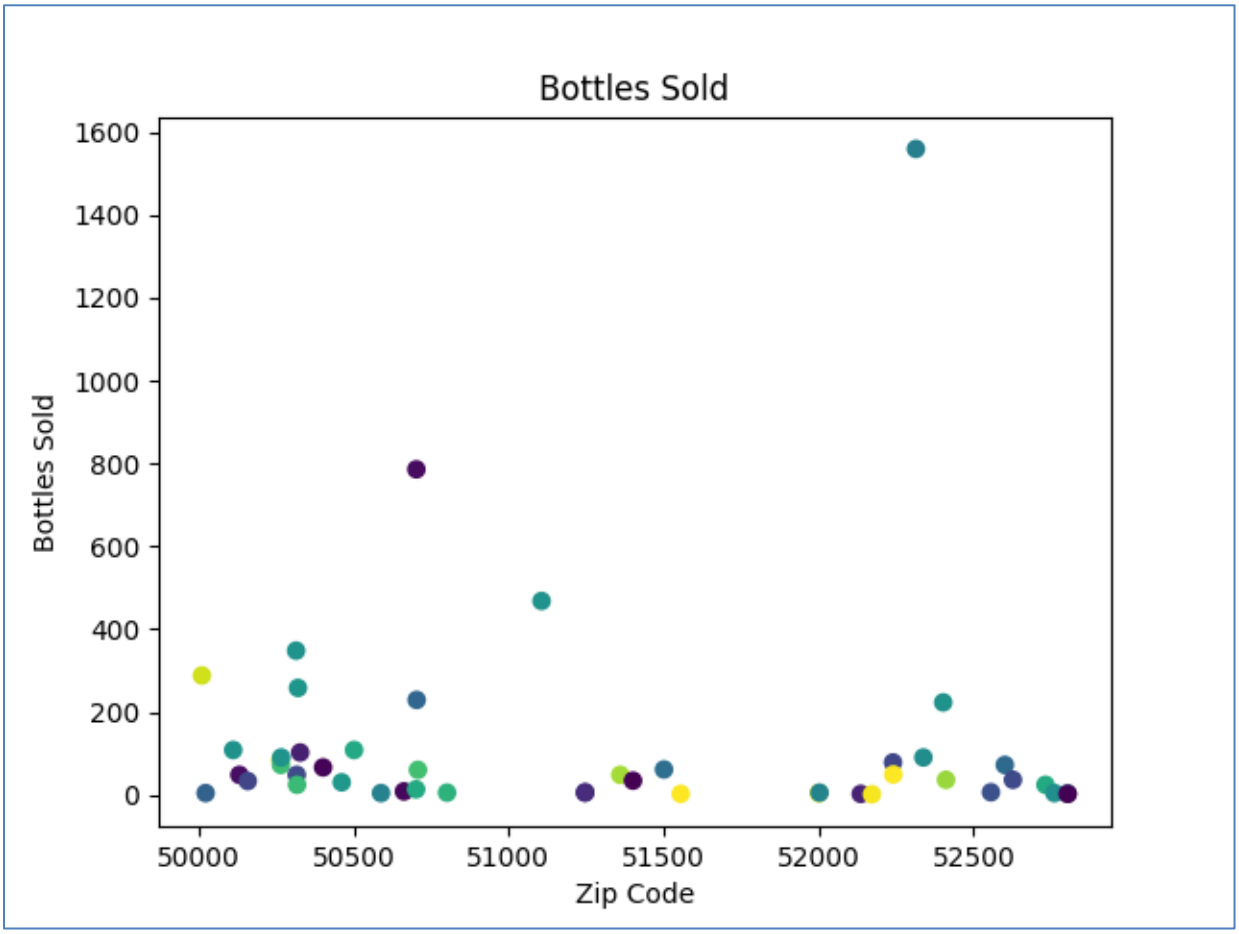
I used Python and Pandas in Pycharm ("**main.py**") to Aggregate the CSV data so I can group the data by zip code and calculate the total bottles sold per zip code using the following commands:

```
import pandas as pd

# Read the CSV file into a DataFrame
df = pd.read_csv("finance_liquor_sales(2016_2019).csv")

# Group the data by zip code and calculate the total bottles sold
grouped = df.groupby('zip_code')['bottles_sold'].sum()
```

- **Step 5.**

I used Matplotlib and Numpy to present my Data in a scatter plot using the following commands:

```
import matplotlib.pyplot as plt
import numpy as np

# Get unique zip codes and generate random colors
unique_zip_codes = grouped.index.unique()
num_colors = len(unique_zip_codes)
colors = np.random.rand(num_colors)

# Scatter plot for bottles sold by zip code
plt.scatter(grouped.index, grouped.values, c=colors)
plt.xlabel('Zip Code')
plt.ylabel('Bottles Sold')
plt.title('Bottles Sold')

# Display the plot
plt.show()
```



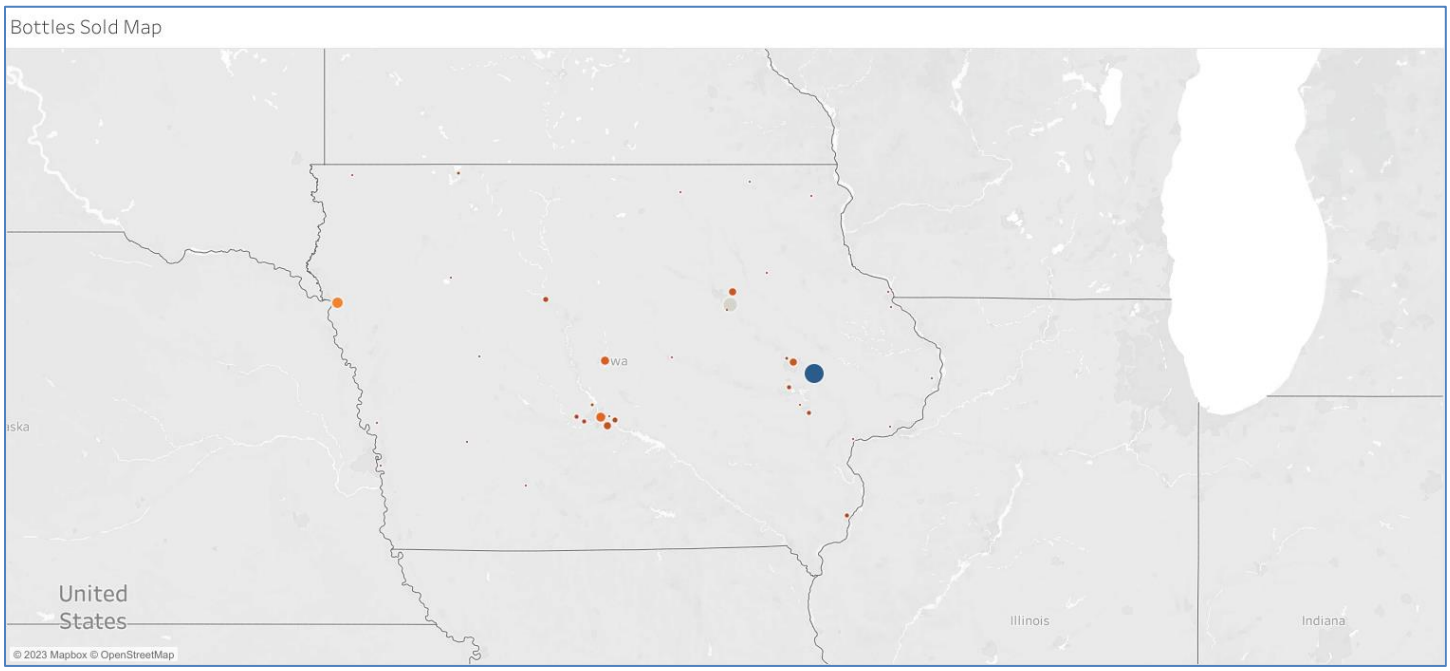
- **Step 6.**

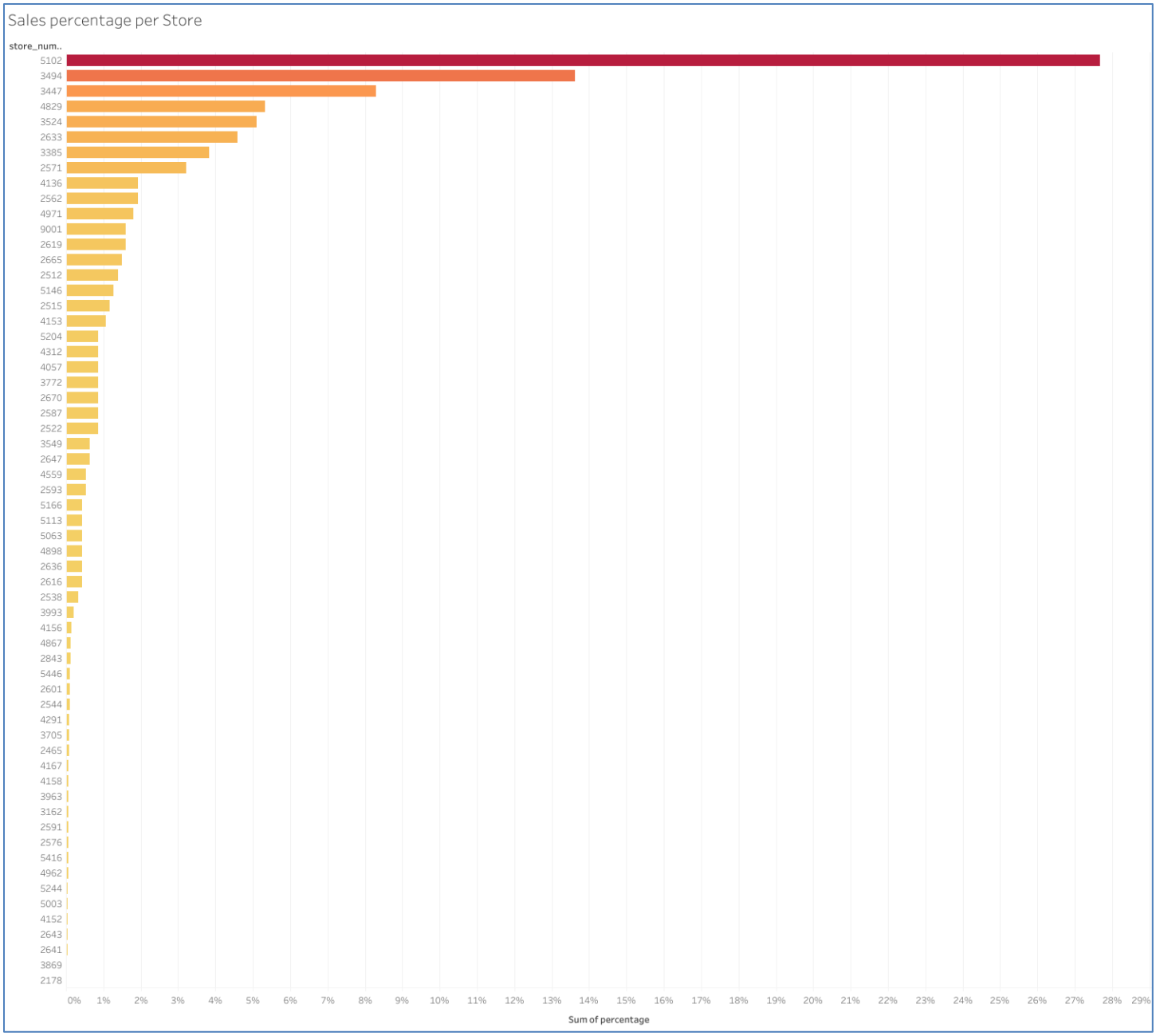
In order to get the Most Popular Item Sold based on Zip Code I used the following command and exported to "**most_popular_per_zip_code.csv**":

```
# Read the CSV file into a DataFrame
df = pd.read_csv("finance_liquor_sales(2016_2019).csv")

# Get the most popular item sold based on zip code
most_popular_per_zip_code = pd.DataFrame(df.groupby(by=['zip_code',
'item_description']).sum()['bottles_sold']).reset_index()

# Convert zip_code column to integer
most_popular_per_zip_code['zip_code'] = most_popular_per_zip_code['zip_code'].astype(int)

# Sort the values by Zip Code
most_popular_per_zip_code.sort_values(by=['zip_code', 'bottles_sold'], ascending=[True, False])

# Export it in csv file
most_popular_per_zip_code.to_csv('most_popular_per_zip_code.csv')
```

- **Step 7.**

I used "finance_liquor_sales(2016_2019).csv" in Tableau in order to show the Bottles Sold Map and Sales Percentage per Store in a Dashboard:
https://public.tableau.com/app/profile/georgios.fragkiadakis/viz/Final_16890795733310/FinalDashboard?publish=yes

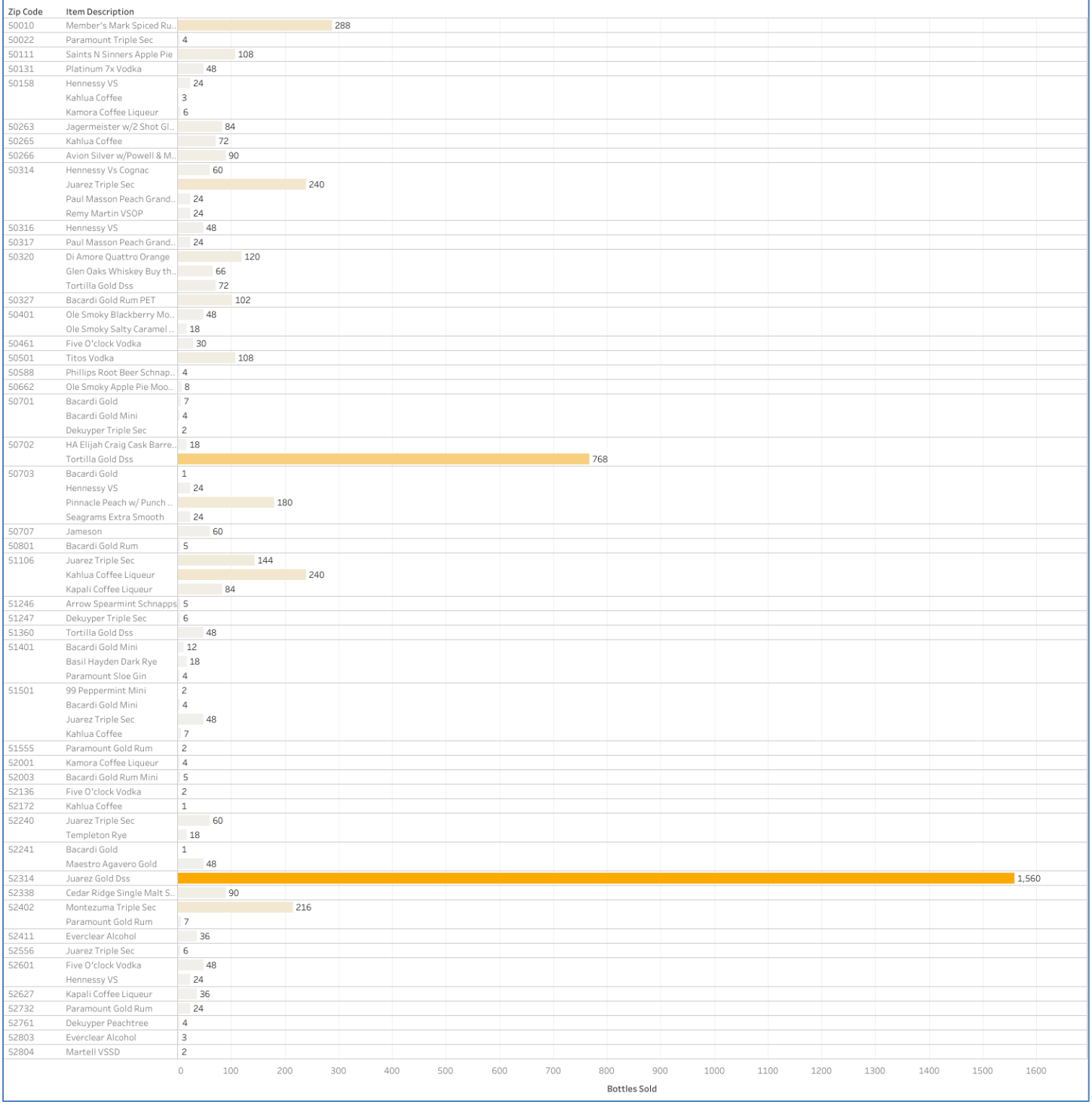


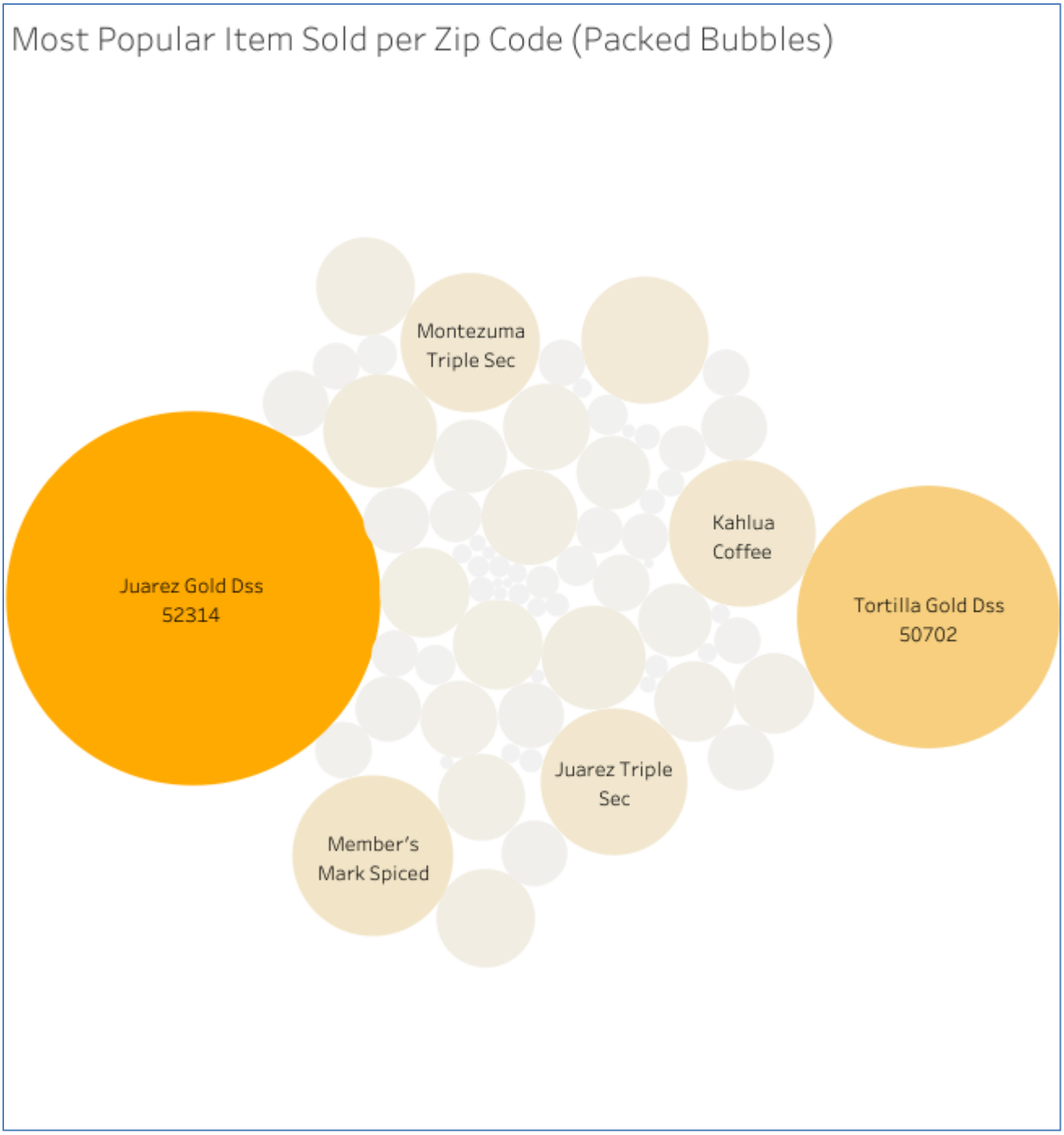


- **Step 8.**

I used " most_popular_per_zip_code.csv" in Tableau in order to show the Most Popular Item per Zip Code in Bar Chart and Packed Bubble Chart:
https://public.tableau.com/app/profile/georgios.fragkiadakis/viz/MostPopularItemSoldperZipCode/Dashboard?publish=yes

Most Popular Item Sold per Zip Code (Bar Chart)





Project Difficulties

The only difficulty that I had was to understand and get the data that we were asked to.