Data Science School by Workearly – Final Assignment

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.

Project Implementation

- **Steps 1-3**

I added the given dataset to MySQL Workbench and created the following query in order to get the Data for the asked time period. Then I exported the Data in an excel file, rather than a csv one, because I faced some issues relating with the compatibility of the MS Office that I have.

```
use liquorsales;

select * from finance_liquor_sales

where year(date)>=2016 and year(date)<=2019;
```

- **Step 4**

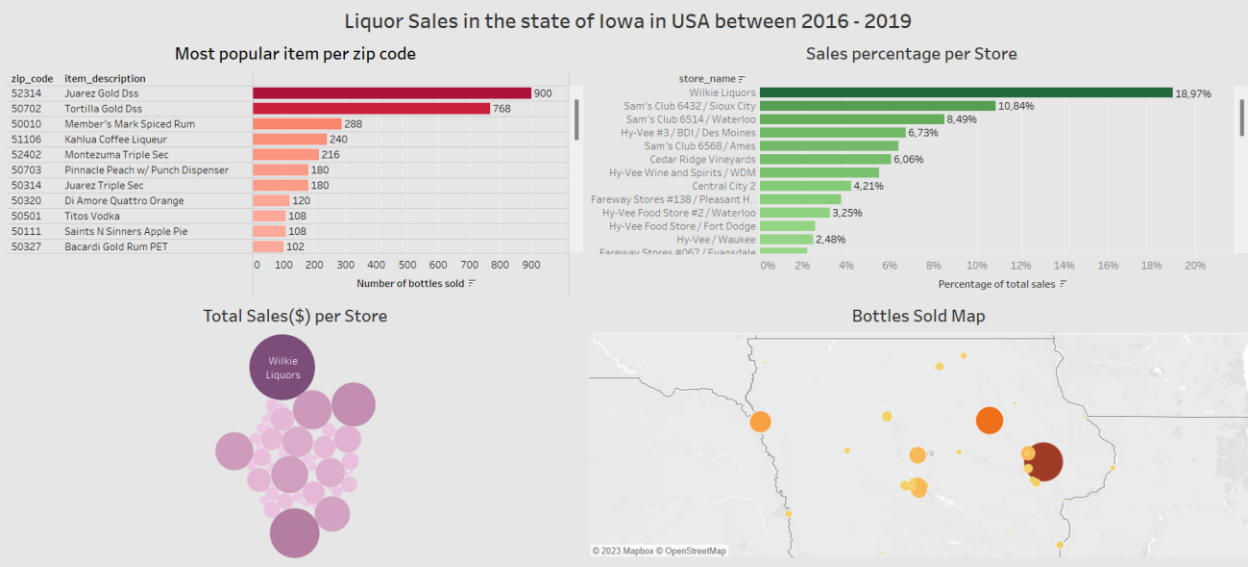
I used Python and Pandas in order to aggregate the xlsx data. I choose to create two new Pandas DataFrames with the most popular item per zip code and the percentage of sales per store respectively, and I exported them in two xlsx files.

Final_Assignment.py

```python
import pandas as pd

# create a DataFrame based on the given dataset for the asked time period
data = pd.read_excel("liquorsales2016_2019.xlsx")
df = pd.DataFrame(data)

# group the DataFrame by zip codes and find the most popular item for each
grouped_zips = df.groupby(["zip_code"])
max_bottles_index = grouped_zips["bottles_sold"].idxmax()
most_popular = df.loc[max_bottles_index, ['zip_code', 'item_description', 'bottles_sold']]
most_popular = most_popular.sort_values(by="bottles_sold", ascending=False)
most_popular = most_popular.reset_index(drop=True)
most_popular.to_excel("Most popular item per zip code.xlsx", index=False)

# group the DataFrame by stores and find the contribution of each store in the total sales
total_sales = df["sale_dollars"].sum()
grouped_stores = df.groupby(["store_name"])["sale_dollars"].sum()/total_sales*100
grouped_stores.to_excel("Percentage of sales per store.xlsx", float_format='%.4f')

```

- **Step 5**

I presented my Data from the newly made xlsx files with the help of Tableau. I also published the following Dashboard in Tableau Public.



Project Difficulties

The main difficulty that I faced was that I was not very used to manipulating data with pandas, so I did some search online in order to find the right modules for each task. Once I understood the proper use of some modules, the whole procedure seemed pretty easy.