Report about Final Assignment

ELEFTHERIADOU SOFIA

Περιεχόμενα

1. CONSEPT

We are given a dataset about Liquor Sales in the state of Iowa in United States of America between 2012-2020. What we want is to find the most popular item per zip code and the percentage of sales per store between 2016-2019. We are going to visualize the result either in matplotlib or Tableau Public.

2. GIVEN

We are given the file "finance_liquor_sales.sql" which contains the matrix about the Liquor Sales. The matrix has information about the location that a store is, the store number, the store's sales and invoice information. We are going to use this to extract the result that we want.

3. ADD DATASET TO WORKBENCH

First we have to open Workbench and connect to MySQL Server. Second, we click on **File->Open SQL Script** and we choose "finance_liquor_sales.sql". In order to get all the columns between 2016-2019, we use a Query that gets the year of the column date (YEAR(date)) of the matrix and sort all the columns ascending by date. So, we get a smaller matrix, which contains sales between years 2016 and 2019. The Query is the following:

```
SELECT * FROM finance_liquor_sales

WHERE year(date) BETWEEN 2016 AND 2019

ORDER BY date ASC;
```

4. EXPORT DATA FROM WORKBENCH

We want to extract the smaller matrix in order to use it on PyCharm to extract the result we are asked. We click on **Export button** above the result of the query and we save the new matrix on ".csv" file ("finance_liquor_sales 2016-2019.csv").

5. EXPORT RESULTS WITH PyCharm

We create a file named **"FinanceLiquor"** which will contain the python scripts about the two questions. So, in the **"main.py"**, we import pandas so we will have the ability to make aggregations (**import pandas as pn**). Then, we read the csv file that we have the data we need and then we convert the matrix into dataframe so we can make aggregation using Pandas:

```
data = pn.read_csv('finance_liquor_sales 2016-2019.csv')
data2 = pn.DataFrame(data)
```

We make sure that the column ‘zip_code’ has the right datatype and we create a smaller dataframe with the information we will use to find the most popular item per zip code.

```
data2['zip_code'] = data2['zip_code'].astype(int)
data3 = data2[["zip_code", "item_description", "bottles_sold"]]
```

We sort the data3 by ‘zip_code’ and ‘bottles_sold’, ascending and descending respectively and we group the result by ‘zip_code’. We print the result without index and we save it in a csv file named **“solution1.csv”**.

```
data4 = data3.sort_values(['zip_code','bottles_sold'],
ascending=[True,False]).groupby(['zip_code']).head(100)
print(data4.to_string(index=False))
data4.to_csv('solution_1.csv',index=False)
```

To find the percentage of sales per store, we choose the following columns from the first dataframe ‘store_number’ and ‘sale_dollars’, we sort ascending and group by ‘store_number’. We find the total sales by summing the column ‘sale_dollars’. We add a column named ‘percentage’ by divided ‘sale_dollars’ by ‘total_sales’ and multiplying the result by 100. We print the result and save the dataframe in a csv file named ‘solution2.csv’.

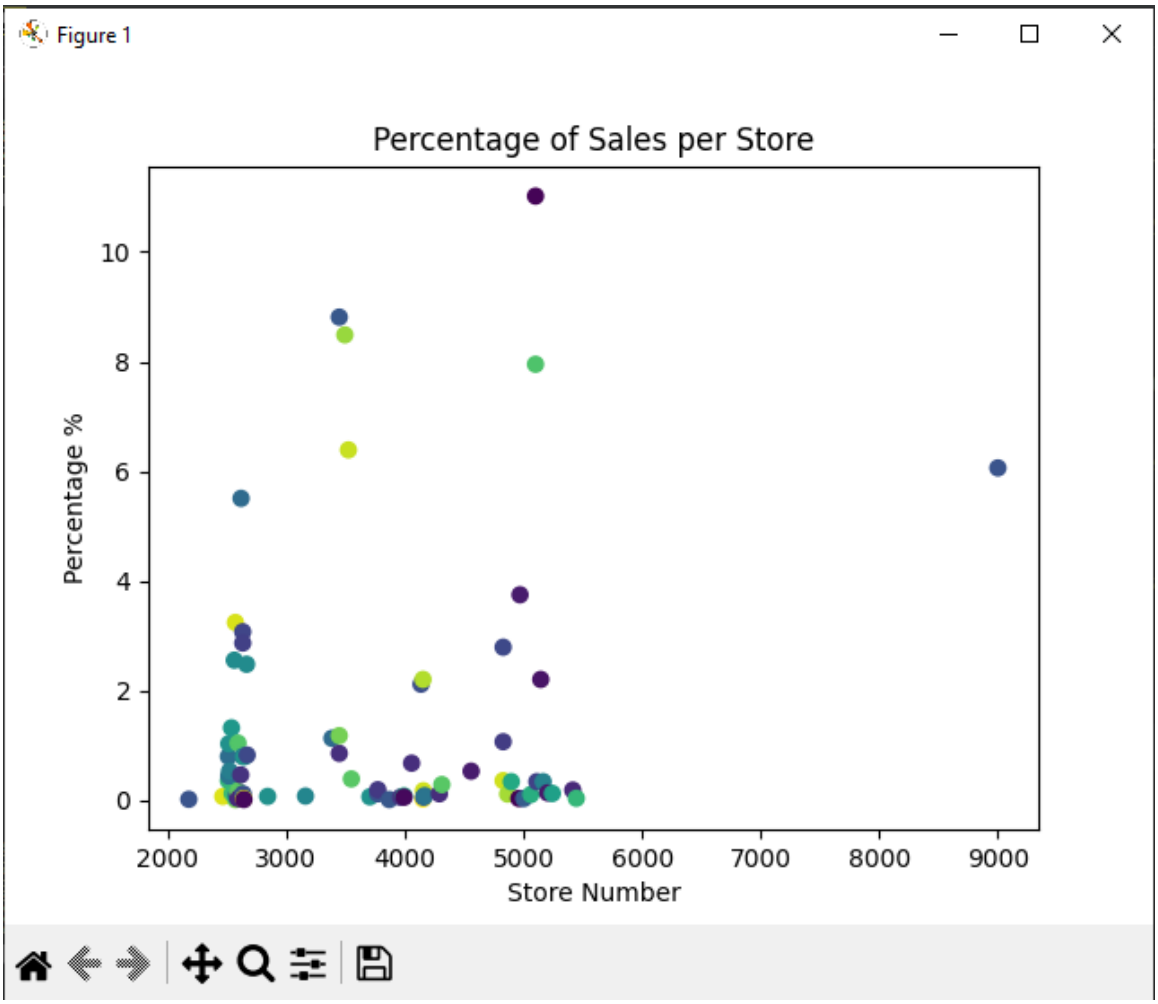
```
data5 = data2[['store_number', 'sale_dollars']]
data6 = data5.sort_values(['store_number'],
ascending=True).groupby(['store_number']).head(100)
total_sales = data6['sale_dollars'].sum()
data6['percentage'] = (data6.sale_dollars/total_sales)*100
print(data6.to_string(index=False))

#save the new csv file for solution 2 a
data6.to_csv('solution_2.csv', index=False)
```

6. PLOTTING BY matplotlib

In file **‘matplot.py’**, we have the visualization of the most popular item per zip code. We use a scatter plot with ‘zip_code’ on x-axis and ‘bottles_sold’ on y-axis. Also we make random colors every time. The plot is the following:

In **"matplot2.py"** we have the visualization for the percentage of sales per store. We have 'store_number' in x-axis and 'sale_dollars' in y-axis. The result is below:

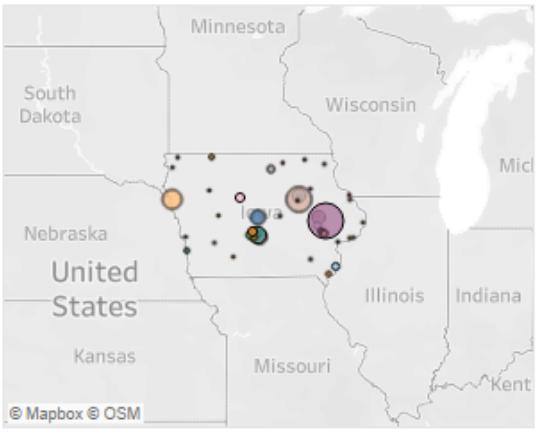


7. PLOTTING IN TABLEAU PUBLIC

We have also the result plotted in Tableau Public. The link is below and so is the visualization:

https://public.tableau.com/views/FinanceLiquor/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

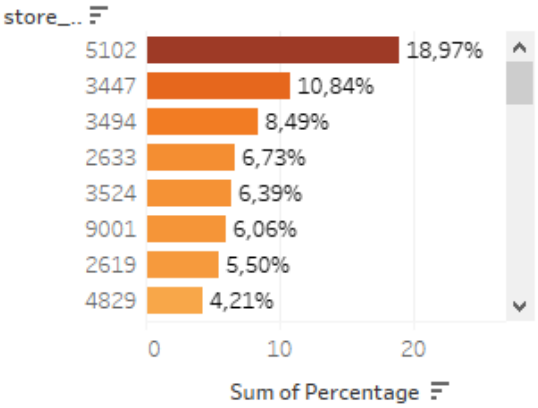
Bottles Sold Map



ottles Sold Matrix

p_code	item_description	
0010	Member's Mark Spiced Ru..	288
50022	Paramount Triple Sec	4
50111	Saints N Sinners Apple Pie	108
50131	Platinum 7x Vodka	48
50158	Hennessy VS	24
	Kahlua Coffee	3
	Kamora Coffee Liqueur	6
50263	Jagermeister w/2 Shot Gl..	84
50265	Kahlua Coffee	72
50266	Avion Silver w/Powell & M..	90
50314	Hennessy Vs Cognac	60
	Juarez Triple Sec	240

Sales Percentage per Store



Sales of Percentage per Store - Bubbles