Liquor Sales in Iowa, USA 2016-2019

Initial dataset and questions to answer

The initial dataset contained the Liquor sales in the state of Iowa (USA) between the years 2012-2020 however we have been asked to find:

- the most popular item per zipcode and
- the percentage of sales per store

in the period between 2016-2019.

Description of steps taken

A SQL query was used to extract all the columns between 2016 – 2019 in a CSV format from Workbench. Then, utilising Python and the Pandas library, the CSV file was imported and a dataframe was created to be further analysed.

I printed the information about the dataframe which contains 24 columns and 74 entries, 0 to 73. The data types are object, integers and float and there also some missing values on the variables 'store_location' (9) and 'category_name'(6). However, to answer the questions I kept only the following 7 columns: 'invoice_and_item_number', 'store_number', 'store_name', 'zip_code', 'item_description', 'bottles_sold', 'sale_dollars'.

For the first question, I grouped the data based on zip code and item description, I found the sum of the 'bottles_sold' for each zipcode, sorted the values in descending order and printed the first row.
For the second question, at first, I found the total of the 'sale_dollars', I created a new column called '%_of_sales' and calculated the percentage with the following script:

```
df["%_of_sales"] =(df['sale_dollars'] / sum_of_sales) * 100
```

Then, I grouped the data by 'store_number', found the sum of '%_of_sales' for each row and sorted the values in descending order.

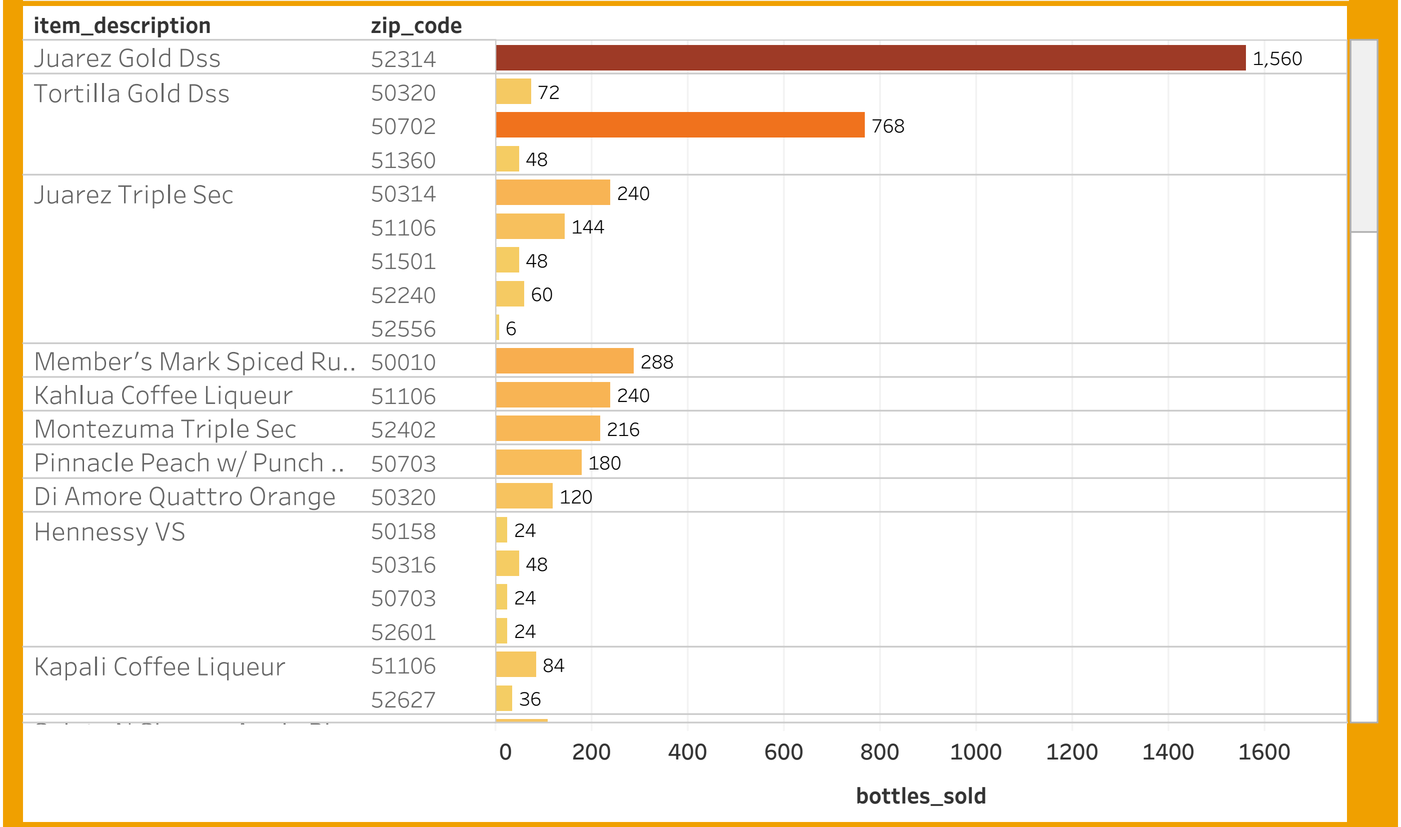
For the visualization of the data, I opened the CSV file in Tableau, created the graphs for the questions and then I created a dashboard with the two graphs. A link has been attached:
https://public.tableau.com/views/Finalassignment_16744901562800/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link
The first graph is a horizontal bar chart showing the sum of bottles sold in each zip code and the item description.
The second graph is also a horizontal bar chart showing the percentage of sales for each store. To calculate the percentage, I converted the 'sale_dollars' variable to measure, I chose the measure sum and from the quick table calculation I chose the percent of total.

Challenges I faced

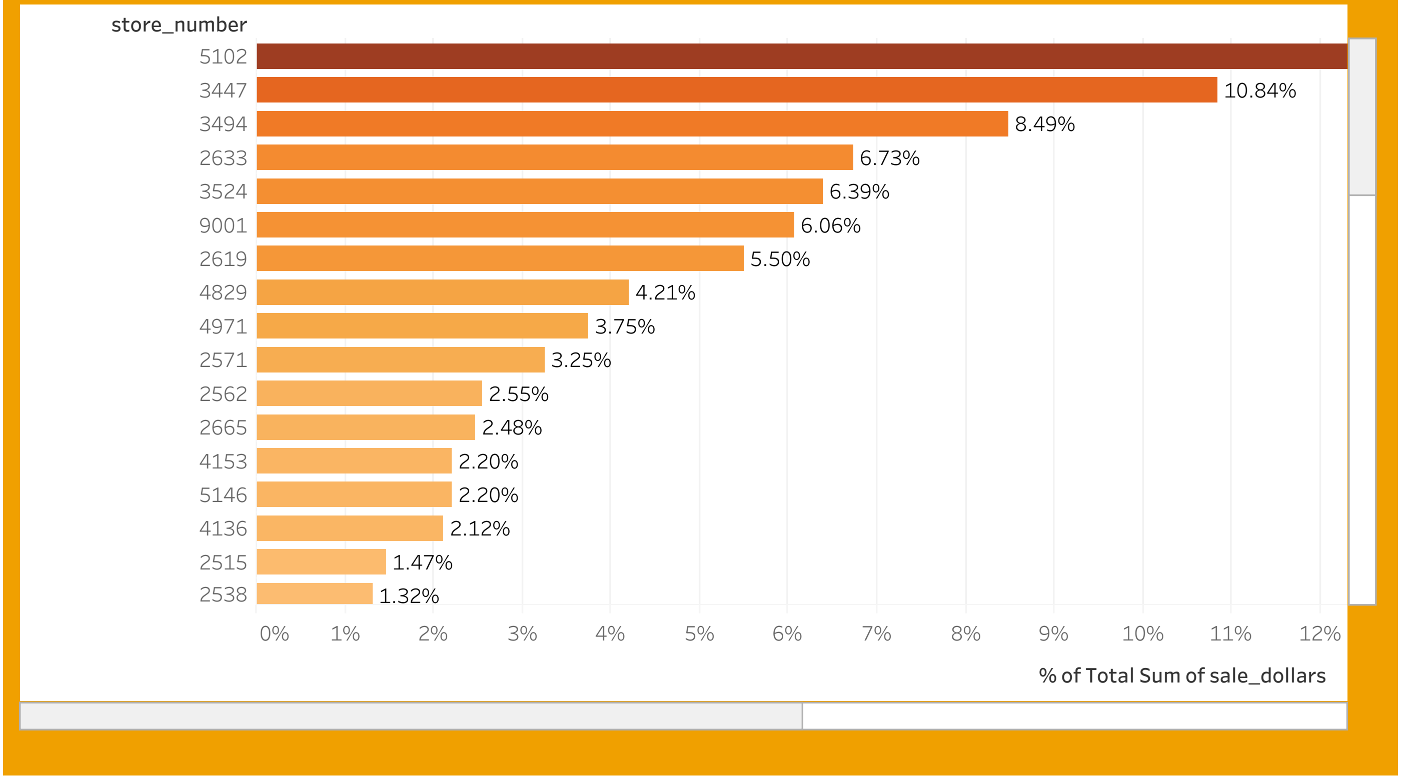
During the initial steps (importing the file and analysing the data) I did not face any significant difficulty as I was using google and the provided material to find everything I needed.
The main challenge I faced was plotting the data using the matplotlib. Even though I was searching on google I kept getting errors or wrong values. Specifically, I had difficulty to plot the aggregated variable 'bottles_sold' and for the second graph I was trying to create an horizontal bar chart however, the bars were displayed as lines and only 3 of them were shown on the graph.

Liquor Sales in Iowa, USA 2016-2019

Most popular item sold based on zipcode



Percentage of sales per store


store_number
% of Total Sum of sale_dollars