**Liquor Sales in Iowa from 2016 to 2019**

**Introduction**

This report gives an overview of the quantitative analysis of Liquor Sales in the state of Iowa in USA between 2012-2020. The analysis focuses on the most popular item based on the zip code and the percentage of sales per store from 2016 to 2019 . The report also illustrates the steps followed and the issues that occurred in order to extract the results of the research.

**Methodology**

The dataset used is the finance\_liquor\_sales.csv file. It consists of 196 rows and 24 columns. The survey was about a certain period of time so some data manipulation is needed.

To extract the specified amount of data a query was used in MySQL.The dataset had one table named finance\_liquor\_sales.The query to get all the columns of the table between the years 2016-2019 was:

*use liquorsales;*

*SELECT \**

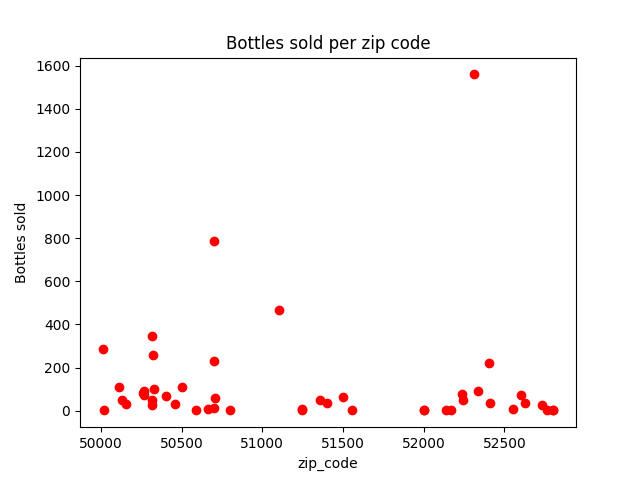
*from finance\_liquor\_sales*

*where YEAR(date) between 2016 and 2019;*

There were seventy-four rows returned. Then the output was exported to a csv file so that the data would be aggregated using Python and Pandas.

Aggregating the data in Python was the most difficult part of the analysis. Executing the right code to have the desired outcome was a long try and error process. Especially the quest to find the percentage of sales per store. Firstly, the data was tested for having null values and since there were a few ,dropna() method was used to remove empty cells. The code to get the sales per zip code and show that with a dot plot was the following:

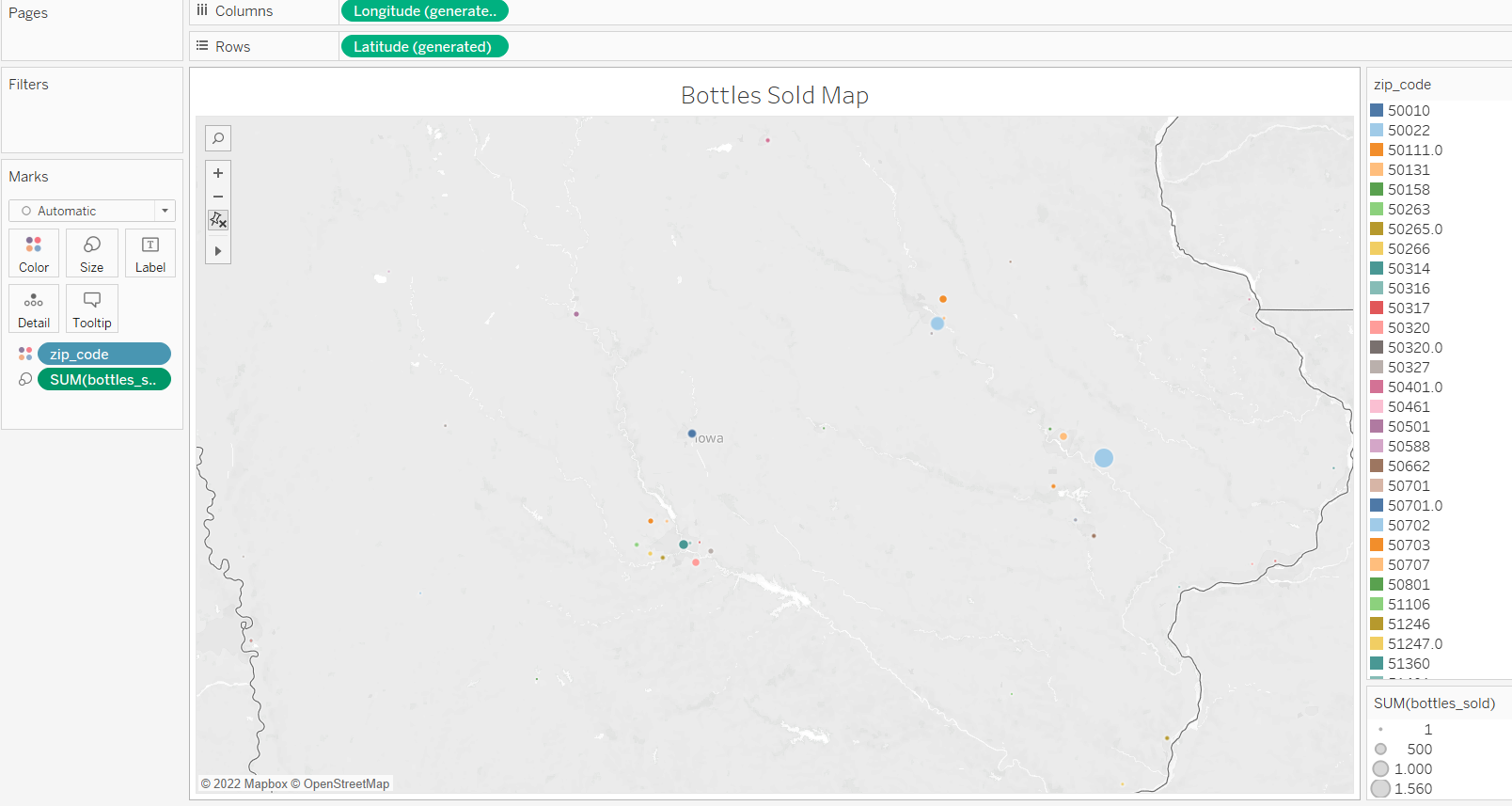
import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
  
  
data = pd.read\_csv('liquorsalesQ1.csv')  
  
remove\_nan=data.dropna(how='all')  
final=remove\_nan.copy()  
  
  
q=final.groupby('zip\_code').bottles\_sold.sum()  
  
plt.plot(q,'ro')  
plt.xlabel('zip\_code')  
plt.ylabel('Bottles sold')  
plt.title('Bottles sold per zip code')  
plt.show()

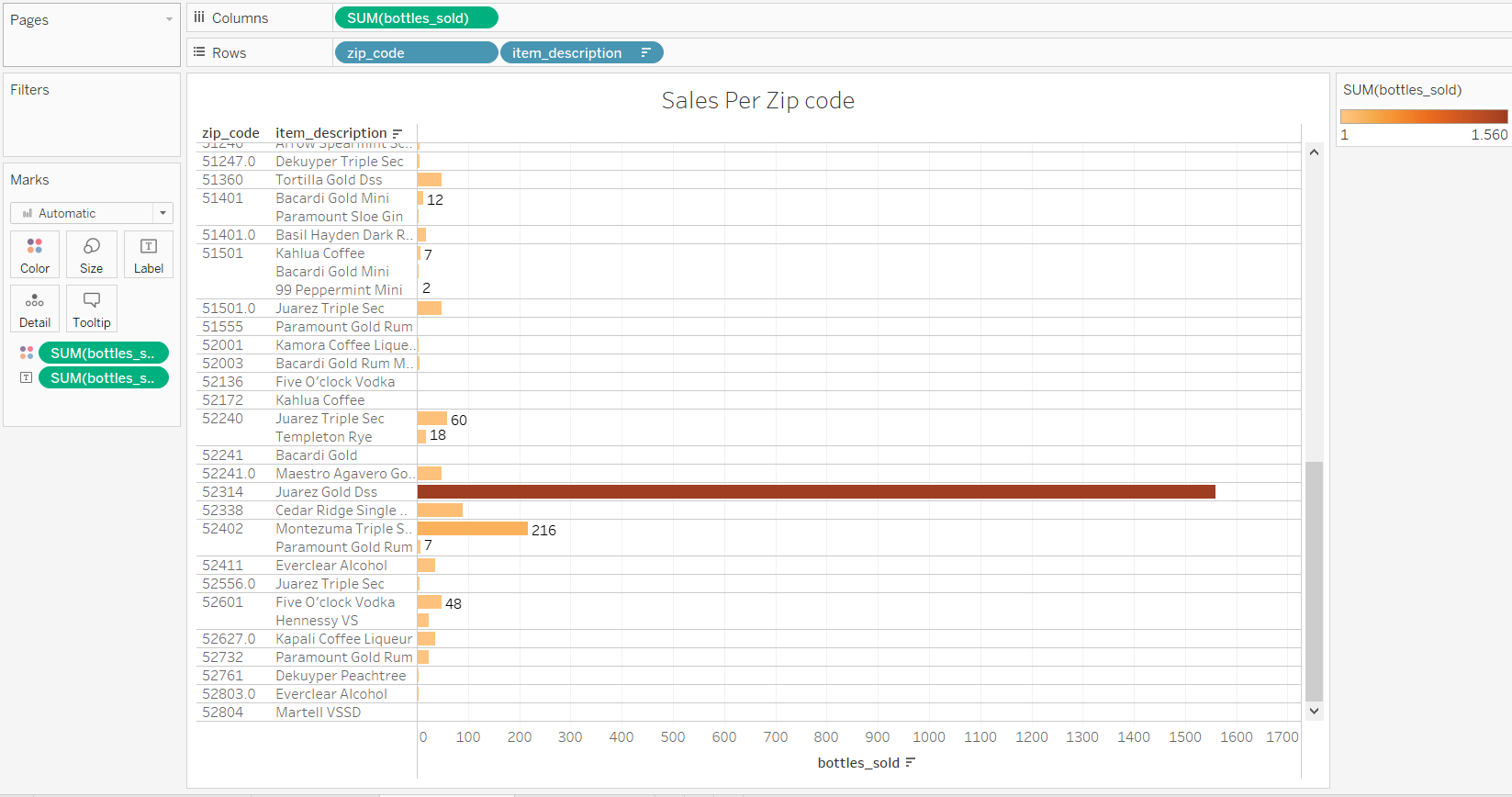


Moving on, to calculate the percentage of sales per store some additional lines of code was needed. The counter is used to find the number of columns and iterate afterwards to get the percentages of the sold bottles per store number.

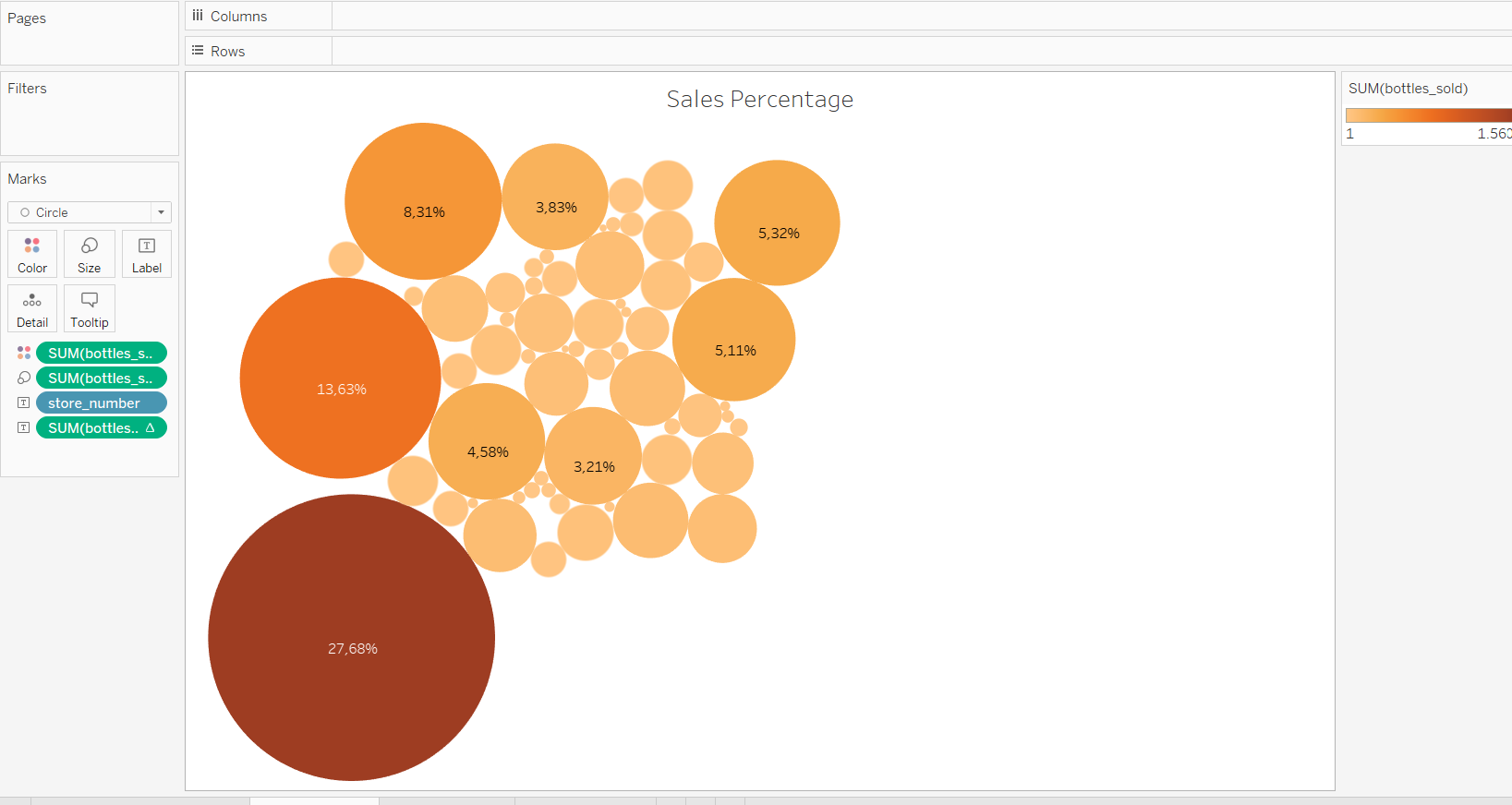
counter =0  
for col in data.columns:  
 counter = counter +1  
  
  
q=final.groupby('store\_number').bottles\_sold.sum()  
percentages=[]  
for i in range(counter -1):  
 percentages.append(q)  
 percentages.append(100-np.sum(percentages))  
  
print(percentages)

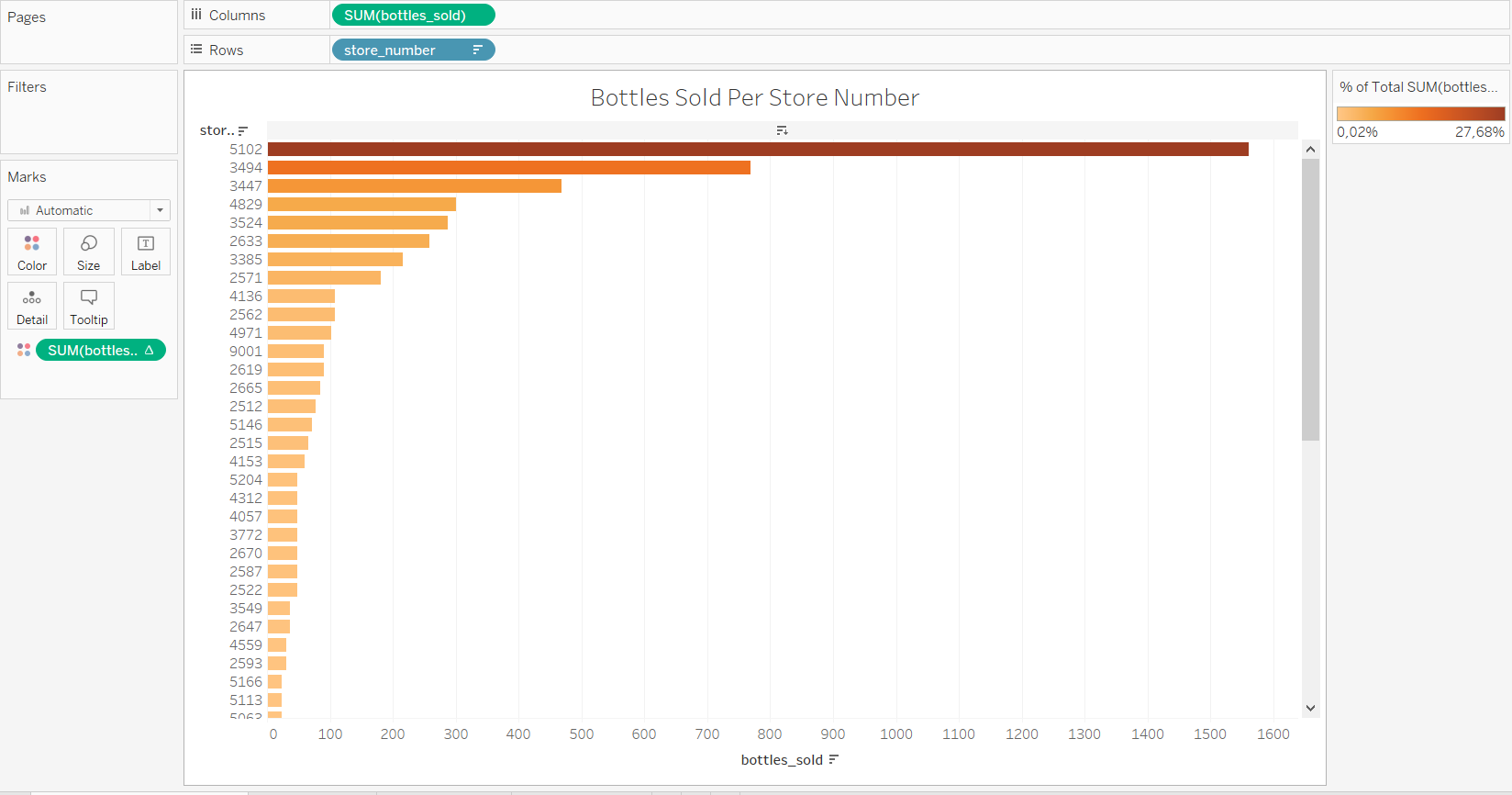
The csv file had the proper format so there was no need to use data interpreter. As a result the data was inserted in Tableau. Regarding the bottles sold based on the zip code a symbol map was selected and the zip code was added in the marks to the size option. Additionally, the sales per zip coded were illustrated with a text table and with a more detailed representation since the item description column was also added in the graph. The bottles sold were sorted in descending order by item description and with color and label marks.





For the second part of the results, a packed bubbles graph was used to demonstrate the sales percentage by store number. The tricky part was to label the percentage of the sales for each store and not the sum. That was accomplished by moving the bottles sold measure to the filters section and having a quick table calculation to compute the current value as a percentage of total. That was added to the label and color marks as well as the store number. Finally, the sales per store number was illustrated with a horizontal bar chart too where the sum of bottles sold was sorted in descending order.





**Results**

All things considered, after extracting, editing and illustrating data the aggregation in Python was the most complicated part of the assignment. The way to have the right output that represented the most popular item was demanding and needed a lot of research. The results show that the most popular item is the “Juarez Gold Dss” and located around the area with a zip code of 52314. Furthermore, the store with the greatest sales number is the 5102 with a total of 1560 sales. In conclusion, the process of the specific data analysis was a learning experience that provided me with the necessary insight to understand my weaknesses and see where to focus and keep working more on.