Samrat Kunwar(s5108723)

  Due: 6 Oct 2019

food violations report (2010ICT-Software Technology)

**Executive summary**

As industries are becoming more and more data driven, analysis of data has become very important to keep track of the activities being performed. Python helps in the data analysis process with its built-in libraries NumPy, Pandas, MatPlot and so on.

This document talks about the analysis of data for a government organization with the help of two excel files thus provided. It queries the data’s from the files to a database and performs indicated tasks. Through the implementation of different tasks we find the violations count and graph the average violations over time of different postcode and companies.

Through the analysis it is found that the data collected holds a violation count total of 906014. Similarly, with the graphs show that, the highest violation rate belongs to the postcode 90012 ranging above 200 over the 8th and 9th months period and lowest holders are the post code from 90001 to 1222. California however fluctuates with highest average violation reaching nearly 17.

Based on the graph of Burger king and McDonalds, Burger King has the highest violation rate reaching almost 1.5 average violation rate on the 8 the month. The average violation rate for McDonalds, remains constant with few time gaps touching an average violation rate of 1 .

**Table of Content**

**Introduction ………………………………………………………….3**

**Procedure ……………………………………………………………..4**

**Database Structure ……………………………………………….5**

**Violation Counts ……………………………………………………7**

**Violation count over time ……………………………………..8**

**Conclusion ……………………………………………………………9**

**Introduction**

In the world where everything is advancing towards digitalization, data is starting to be considered as gold. Data is being transferred through a system over a large volume and as industries are becoming more and more data driven, those data thus obtained acts as a critical milestone for improvement. Government organizations are no exceptions as they store data of all the things carried out under their arms. Those data help them to make decisions on what needs to be done based on the result obtained. The data keep track of the activities performed to keep things regulated. So, it is important for the data to be well organized and managed for analysis at any time.

This report goes through how python libraries were implemented to analyse those data. Similarly, it determines the database structure and schemas while displaying all the list of violations in a new excel file called “ViolationsTypes.xlsx”. It then concludes with the graph plotting of violations based on different conditions.

**Procedure**

We need to analyse the data provided to us in two excel files, “inspections.xlsx” and “violations.xlsx”. We use python and its modules to analyse the data. We use python libraries like NumPy for multi-dimensional array, Pandas for Data-frame and Matplot for plotting graphs of data thus obtained.

For the purpose of analysis, the task was divided into four parts. In the first part, we scripted Creatdb\_food.py file. It creates a database with the specific name and adds a Data-Frame for the inspections and violations table.

The second file i.e. sql\_food.py queries all the data from the workbooks and enters them to corresponding table in the database. It lists the distinct businesses that have at least 1 violation and orders it in alphabetical order.

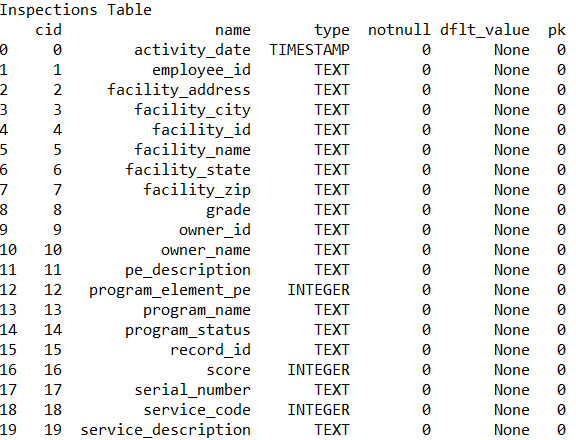
The third file i.e. excel\_food.py, creates a new workbook called “VIolationsTypes.xlsx” and stores the code of violation, description of violation and counts the number of violations made. It then counts the total violation based on the data thus obtained.

The fourth file Numpy\_food.py, plots the data into graph based on the conditions provided.

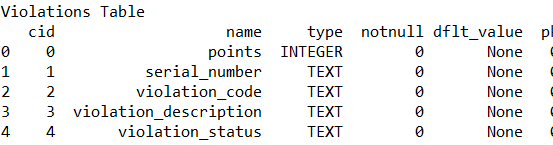
**Database structure**

The database thus created contains two main able i.e. inspections and violations table. The following image provides the data structure and types of data that it stores.

For the inspections table, the following image provides the table structure and the attributes present in the inspections table.

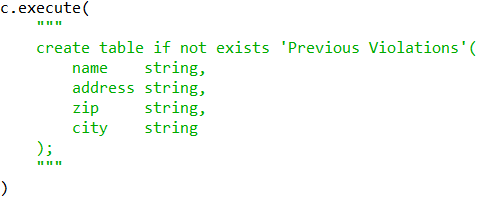


Similarly, the following image shows the structure of the violations table.



As we see from the images that almost all the data is stored in the form of text(string).

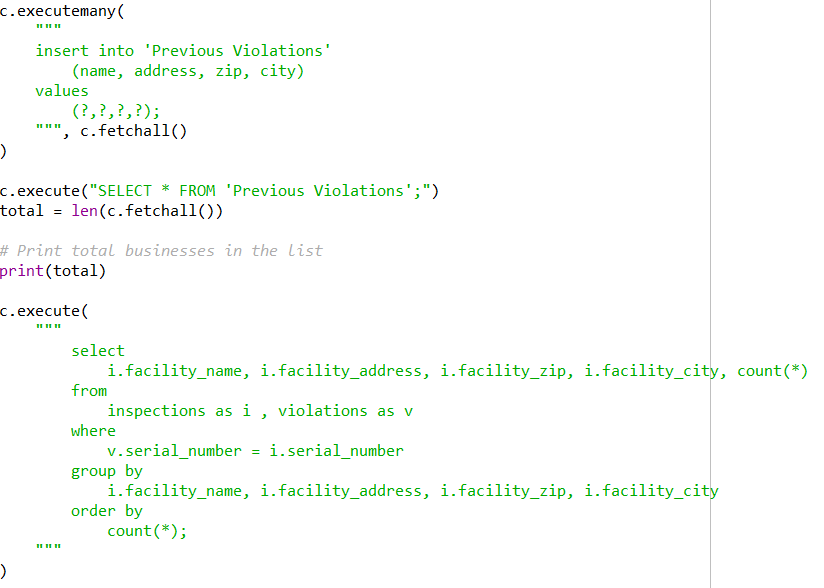
The third table Previous Violations is created through querying of the two tables. The schema and table execution is a follows.



To make sure that we don’t get error, we add the following command before executing the create command.



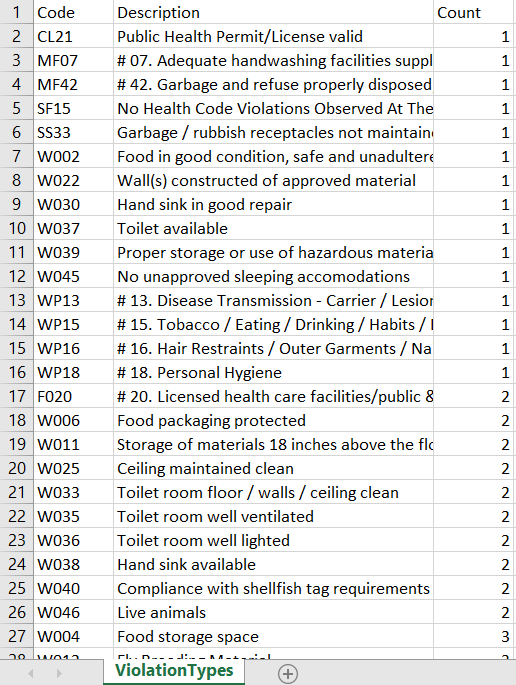
The following image shows the insertion of data into the table.

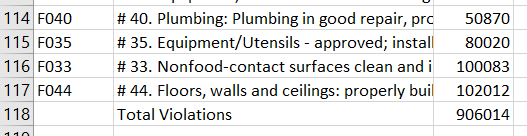


**Violation counts**

Through the analysis of the data we created a workbook called “ViolationTypes.xlsx”. Through the analysis, it was found that the lowest violations was caused by CL21 with the violation count 1 and the highest was F004 with the highest count of 102012.

The total violation count was 906014.





**Violations over time**

Over time the violation count has increased so much that it can’t be computed manually. As we see from the result of previous section, there are about 906014 violation counts. With the help of pythons library, MatPlot and NumPy we plot those data into a graph while taking certain conditions under consideration.

The figure below shows the graph thus plotted to determine the violation caused over time.

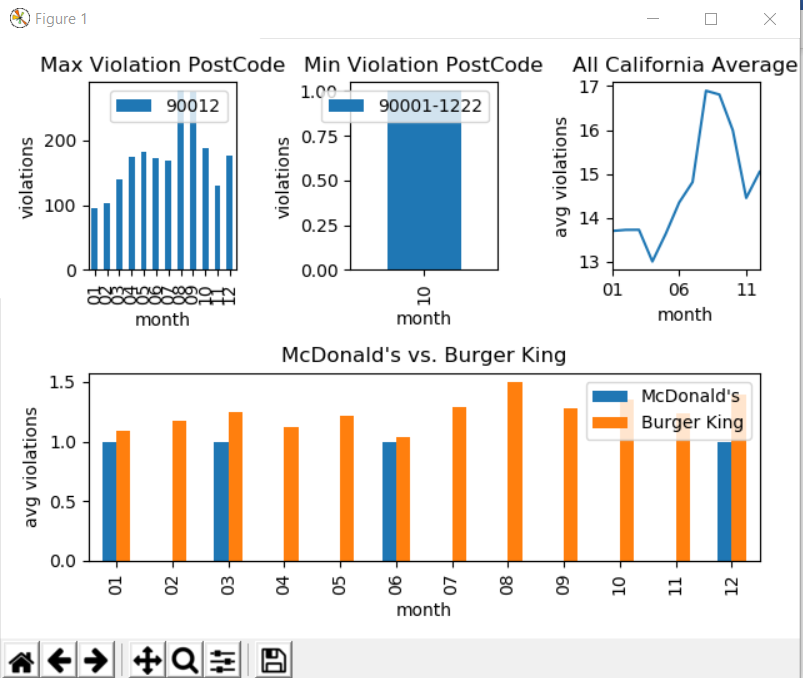


Figure 1 shows the highest number of violations per month for postcode with highest total violation. The highest violation rate over month was by postcode ‘90012’ which occurred in the last 8th and 9th months.

Figure 1 shows the lowest number of violations per month for postcode with lowest total violation. The lowest violation is 1 caused by postcode from 9000-1222 over a 10 month period.

Figure 3 shows the number of violations per month for California over 11 months period. It shows the highest violation almost touches 17 and the lowest violation rate is at the range of 13.

Figure 4 shows the average number of violations per month for all McDonalds and Burger King. We can see that Burger King has the highest violation rate on the 8the month whereas McDonald shows a series of consistency with few month gaps.

**Conclusion**

Thus, through the analysis we found that the total violation count reached about 906014 with the lowest count of one made by CL21 to WP18. Similarly, with the help of graphs, we can analyse that the highest violation rate belongs to the postcode 90012 ranging above 200 over the 8th and 9th months period. Similarly, the lowest violation rate holders are the post code from 90001 to 1222. California however fluctuates with highest average violation reaching nearly 17.

Based on the graph of Burger king and McDonalds, Burger King has the highest violation rate reaching almost 1.5 average violation rate on the 8 the month. The average violation rate for McDonalds, remains constant with few time gaps touching an average violation rate of 1 .