# DATA MANAGEMENT, WAREHOUSING, AND ANALYTICS

CSCI 5408
Fall 2018

# **ASSIGNMENT 4**

Gurjot Singh (B00811724)- gurjot@dal.ca

Devanshu Srivastava (B00810667)- dv960112@dal.ca

## Task Description

The task is to learn data analytics using data analytical dashboards. The focus is to skill in data visualization techniques involved in business intelligence, plot graphs, and big data sets, and learn how to identify patterns through simulation of data.

The data analytics visualization tool used is Power BI. It is a business analyst service provided by Microsoft and provide interactive visualizations with self-service business intelligence capabilities, where end users can create reports and dashboards by themselves, without having to depend on information technology staff or database administrators[1]. Data Visualization plays an important role in big data and advanced analytics projects. It is a technique for creating images and diagrams to communicate messages.

The provided dataset "Building Permits" was downloaded in a CSV format. The data was then loaded in the Power BI tool through the CSV file. The data was cleaned in the process to perform all the tasks. Geo-Spatial Visualization, Histogram, and Charts were created using visualization.

## Tool Selection

The visualization tool used is Power BI. This tool was chosen because we found this easy to use and reliable with doing intensive computations. Power BI is a suite of business analytics tools that deliver insights throughout your organization. It connects to hundreds of data sources, simplifies data prep, and drive ad hoc analysis. This tool produces beautiful reports and publishes them. We can create personalized dashboards with a unique, 360-degree view and scale it across the enterprise, with governance and security built-in[2].
Visualizing data is one of the core parts of Power BI i.e. it is a basic building block for creating visuals.
Power BI has a whole range of visualizations available by default - simple bar charts, pie charts, maps, and esoteric offerings like waterfalls, funnels, gauges, and more. Power BI Desktop also offers extensive page formatting tools, such as shapes and images that can enhance the dashboards.

## Data Loading

The dataset was downloaded from https://catalogue-hrm.opendata.arcgis.com/datasets/building-permits in CSV format. The dataset was loaded in Power BI using "get data" module. This imports the whole dataset which can be easily fetched from the "fields" view in the tool. All the columns were imported and are easily fetched from the fields.

# Data Cleaning

The data was cleaned before performing visualizations. The null values from all the fields were removed from the columns like DATE_OF_APPLICATION, DATE_OF_PERMIT_ISSUANCE which consists of large null value rows. The dates with null values were given a default date and for accessing the geo-coordinates we used STREET_NAME and appended the “Halifax Nova Scotia Canada” so that the location can be fetched easily for Halifax city.

# Plotting

Different plotting techniques were used in creating the graphs for the assignments.
**Task1- Geo-Spatial Visualization:** The main task is to plot the number of permits per location and the size of the point should represent the count of permits per location. A new column was created which consists of the street name merged with Halifax so that all the street names get matched in Halifax and provide accurate data. The column Location_Modified provided in the Location Field was developed as:

```
Location_Modified = Building_Permits[STREET_NAME]&",Halifax,Nova Scotia Canada"
```

The count of “Permit Numbers” was provided in the tool. The result obtained is displayed in Figure 1.

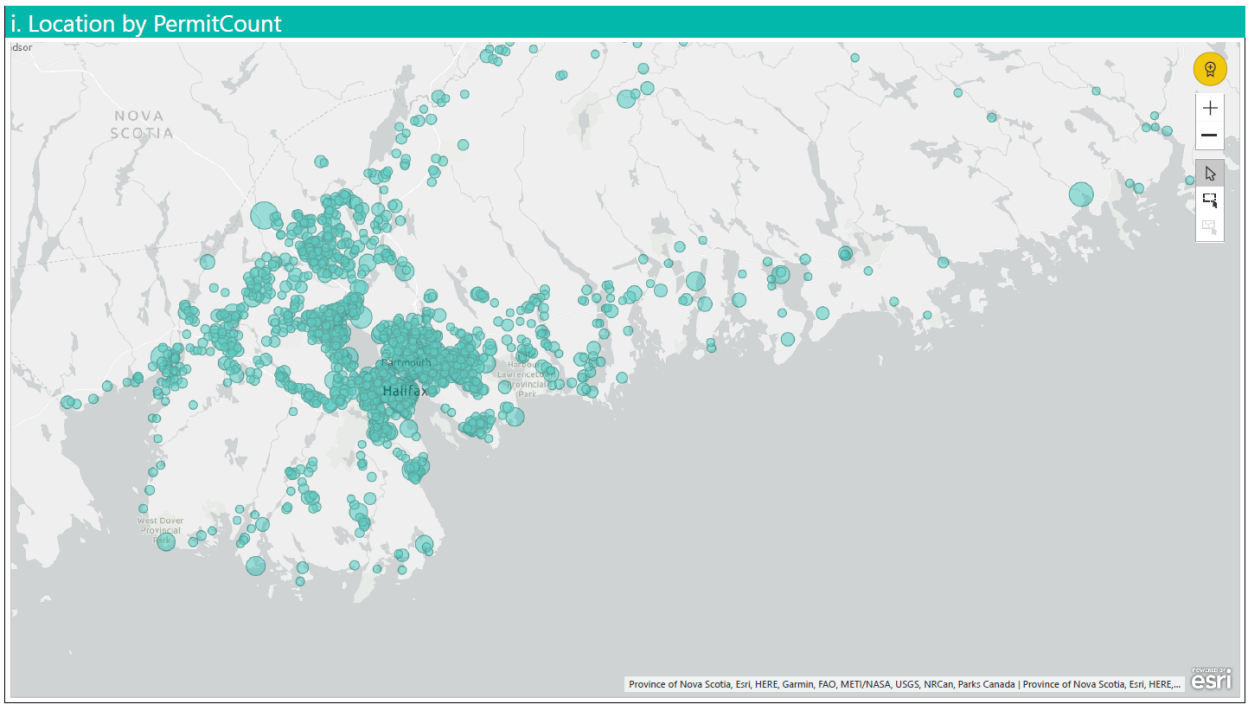



**Figure1: Geo- Spatial Visualization for the number of permits per location**

**Task2- Histogram:** The main task is to create a histogram which represents the count of permits for each permit type. “Permit Type” and the Count of “Permit Number” were provided for this task as the plots for the histogram. The result displayed us the count of permits for all the permit types. The result obtained is displayed in Figure 2.

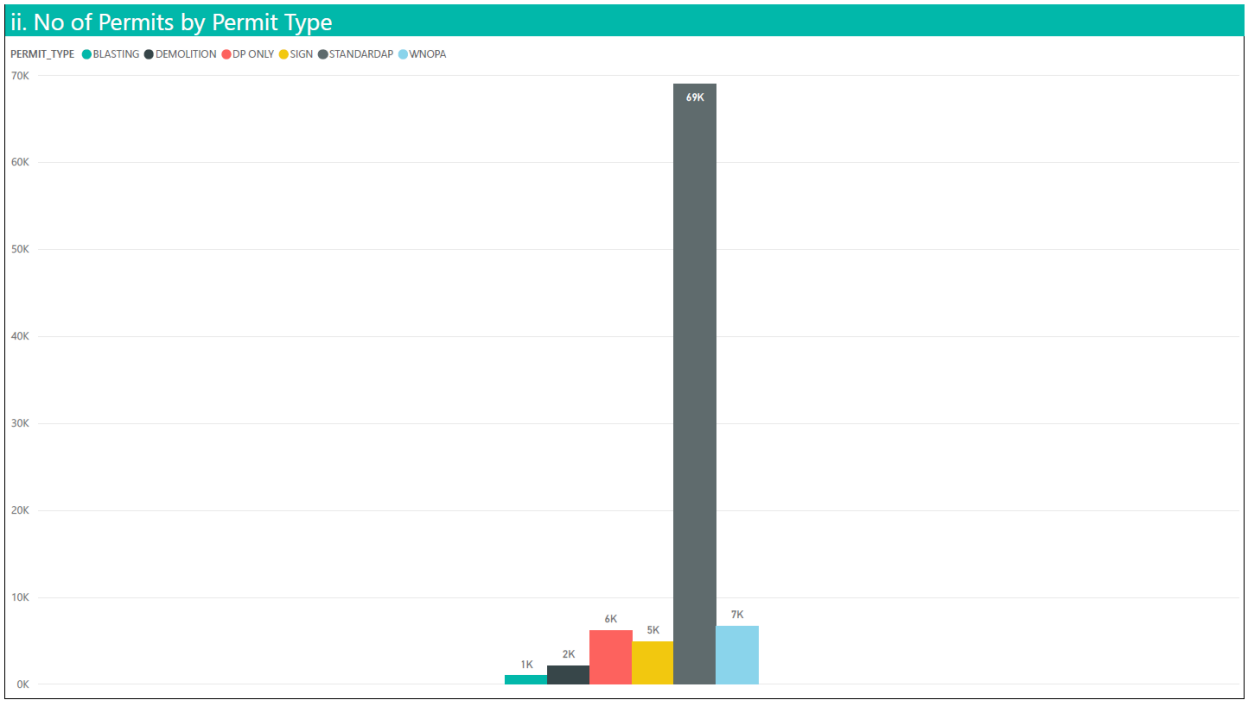



**Figure 2: Histogram representing the count of Permits per Permit Type**

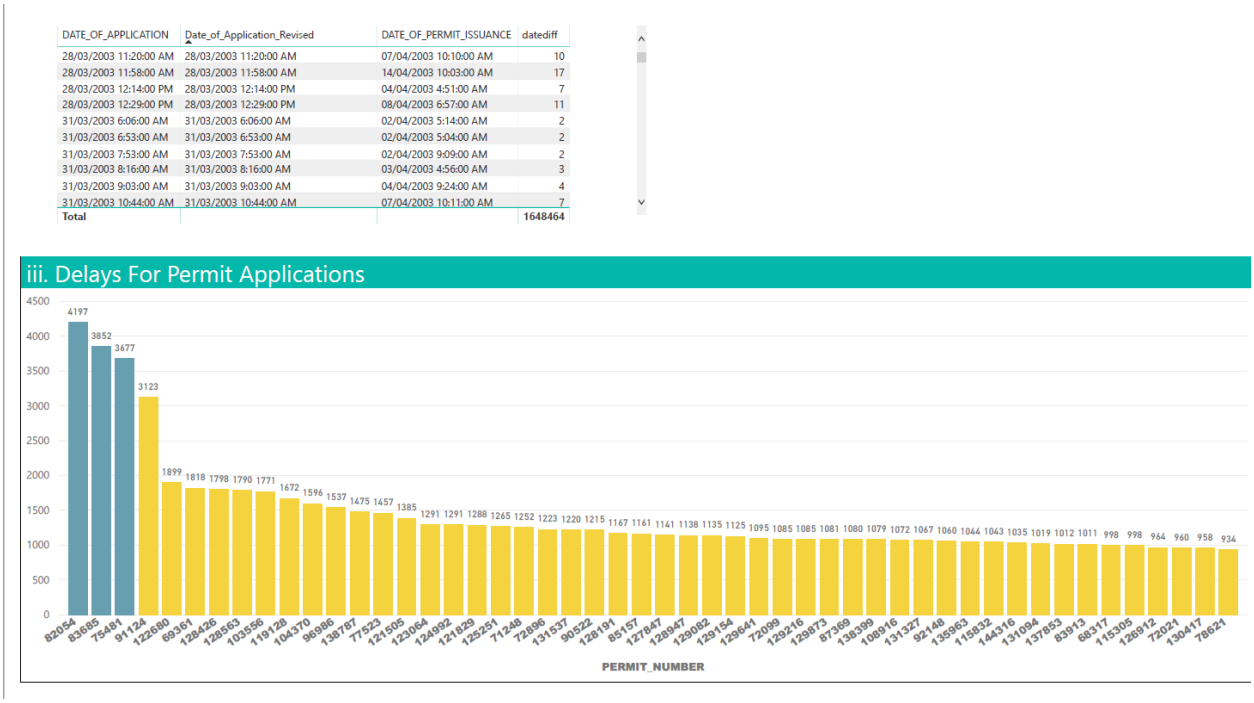
**Task3- Time Delays for Permit Applications:** The task is to calculate the intervals between the permit application date and the permit issuance date and then plot the time delays for all permit applications. For this first, the “Date_of_Application” was cleaned and the null values were replaced with a default date. Program for Cleaning and modifying the “Date_of_Application” is as follows:

```
Date_of_Application_Revised =
if(isblank(Building_Permits[DATE_OF_APPLICATION].[Date]),DATE(1950,01,01),Building_Permits[DATE_OF_APPLICATION])
```

After the new column is generated, the intervals between the permit application date and permit issuance date are calculated using the “DateDiff” function in Power BI and stored in a new column as “datediff”. Program for Calculating the Date Difference between the issuance and application date is as follows:

```
datediff =
DATEDIFF(Building_Permits[Date_of_Application_Revised],Building_Permits[DATE_OF_PERMIT_ISSUANCE],DAY)
```

A Clustered column chart is then generated with the Permit Number and datediff as the values in the fields. The blank values have been filtered out using the tool. The top three permits that took the longest time to be processed has been highlighted in the graph. The results obtained are displayed in Figure 3.



| DATE_OF_APPLICATION | Date_of_Application_Revised | DATE_OF_PERMIT_ISSUANCE | datediff |
|---|---|---|---|
| 28/03/2003 11:20:00 AM | 28/03/2003 11:20:00 AM | 07/04/2003 10:10:00 AM | 10 |
| 28/03/2003 11:58:00 AM | 28/03/2003 11:58:00 AM | 14/04/2003 10:03:00 AM | 17 |
| 28/03/2003 12:14:00 PM | 28/03/2003 12:14:00 PM | 04/04/2003 4:51:00 AM | 7 |
| 28/03/2003 12:29:00 PM | 28/03/2003 12:29:00 PM | 08/04/2003 6:57:00 AM | 11 |
| 31/03/2003 6:06:00 AM | 31/03/2003 6:06:00 AM | 02/04/2003 5:14:00 AM | 2 |
| 31/03/2003 6:53:00 AM | 31/03/2003 6:53:00 AM | 02/04/2003 5:04:00 AM | 2 |
| 31/03/2003 7:53:00 AM | 31/03/2003 7:53:00 AM | 02/04/2003 9:09:00 AM | 2 |
| 31/03/2003 8:16:00 AM | 31/03/2003 8:16:00 AM | 03/04/2003 4:56:00 AM | 3 |
| 31/03/2003 9:03:00 AM | 31/03/2003 9:03:00 AM | 04/04/2003 9:24:00 AM | 4 |
| 31/03/2003 10:44:00 AM | 31/03/2003 10:44:00 AM | 07/04/2003 10:11:00 AM | 7 |
| **Total** | | | **1648464** |



**Figure 3: Time Delays for Permit Applications**

**<u>Task4</u>- <u>Plot to identify the Building Types</u>:** The task is to create a plot of building types where colour coding is used to identify the different types. The "Alternate_Building_Types" column was filtered to remove the null rows through Power BI. Then the column was provided in the field and a Donut Chart was developed which then provided the complete colour coding of different types of building types. The result has been displayed in Figure 4.

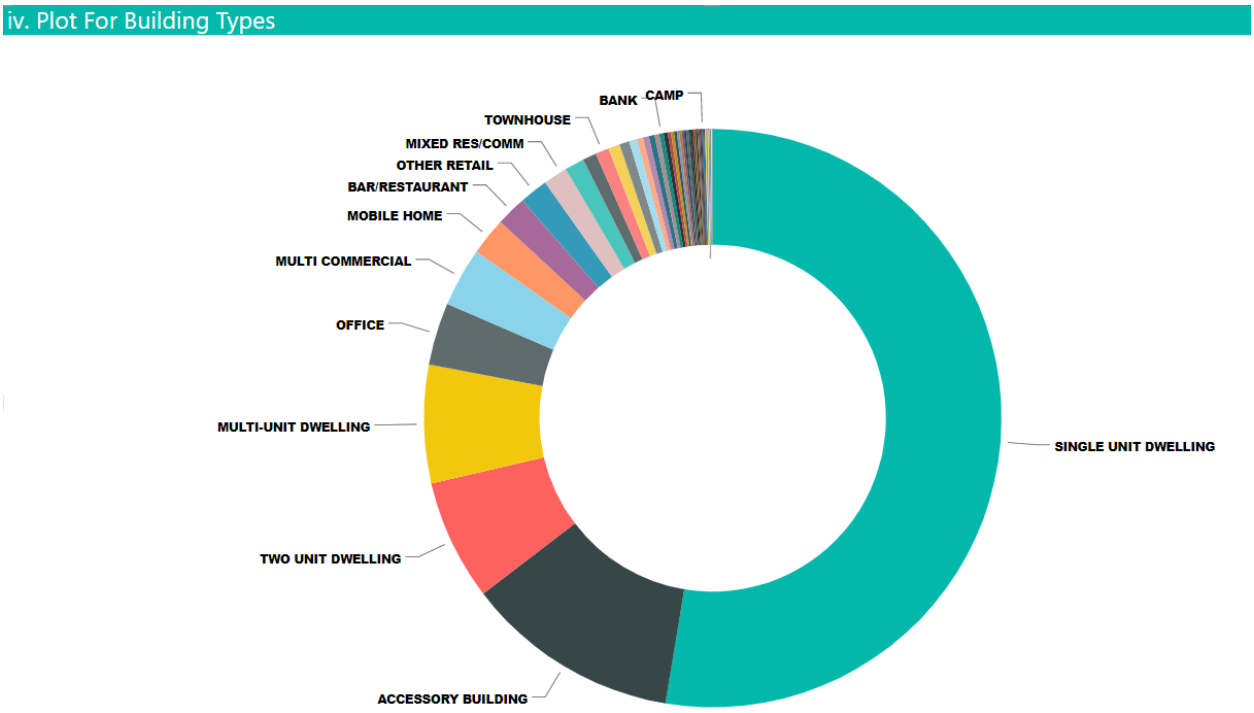



**Figure 4: Plot to identify the Different Building Types**

# Analysis

Graphs are a common method to visually illustrate relationships in the data. The purpose of a graph is to present data that are too numerous or complicated to be described adequately in the text and in less space[3].

In the first task, the Geo-Spatial Visualization developed displayed the number of permits for each location. This provided us the comparative result of the count of permit numbers for all the locations and using the graph we were able to compare different locations and their count. The graph also dynamically displayed the location based on the size of the permit numbers, for example, “Sackville” has a big count of permit numbers i.e. 1589 has a bigger plot than others. Also on highlighting a specific location provided complete details of that location that we are providing in the tool. So to conclude, using Geo-Spatial Visualization, comparison among different locations is very useful in the analysis.

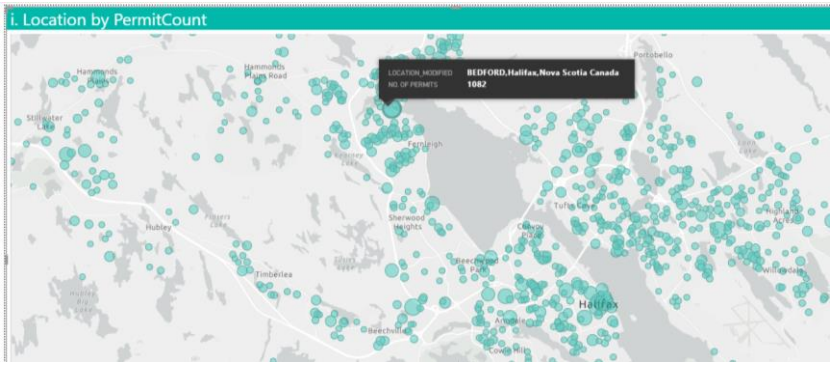



Figure 5: Analysis of Geo-Spatial Visualization

In the second task, histogram explains the count of permit type for respective permit numbers. Using the histogram, we can easily figure out the permit types which has high permit counts with complete details and also the types which have less permit counts. For example, “Standardap” has the highest permit count whereas, “blasting” has the minimum count. This type of visualization is very useful for comparing the values of different types.

In the third task, the clustered column chart displays the complete details of the delays for all the permit applications. From the graph, we can easily figure out that the permit numbers 82054, 83685, and 75481 took the longest time to be processed. This analysis is very useful when we are comparing data containing a lot of columns, so this type of graph will provide the complete list and display the results.

In the fourth task, we have created the donut pie chart which provides the different building types using an enhanced color code. From the graph, we can analyse that the “Single unit Dwelling” is the maximum types of building types. On highlighting a specific building type, the graph provides complete details of the count of the selected building type. This type of visualization is very useful when we need to find the size of the different type of objects.

# References

[1]"Power BI | Interactive Data Visualization BI Tools", *Powerbi.microsoft.com*, 2018. [Online]. Available: https://powerbi.microsoft.com/en-us/. [Accessed: 01- Nov- 2018].

[2]https://docs.microsoft.com/en-us/power-bi/guided-learning/visualizations?tutorial-step=1

[3]"What is graph? definition and meaning", *BusinessDictionary.com*, 2018. [Online]. Available: http://www.businessdictionary.com/definition/graph.html. [Accessed: 01- Nov- 2018].