**Ezeofor Chinedu Emmanuel**

# Bi-weekly Challenge 1

## Exploratory Data Analysis (EDA) of Building Permits in the US.

The processes taken during this analysis has been split into four (4) parts which include:

1. Importation of the necessary libraries.
2. Data ingestion
3. Data cleaning
4. Data Exploration

**Tool used:**

|  |
| --- |
| * Python (Jupyter notebook) |

**Importation of the necessary libraries.**

The necessary libraries (packages) were imported as shown in figure 1. They include:

|  |  |
| --- | --- |
| * Pandas * Numpy | * Matplotlib and Seaborn (Visualization) |

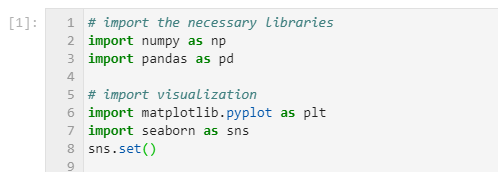


Fig 1: Importation of important libraries

**Data Ingestion**

The ‘csv file’ containing the data was loaded using pandas. It is shown in figure 2 below.

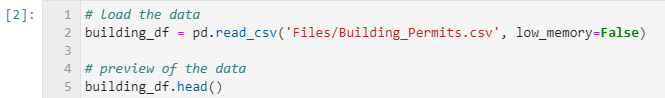


Fig 2: Loading of the data

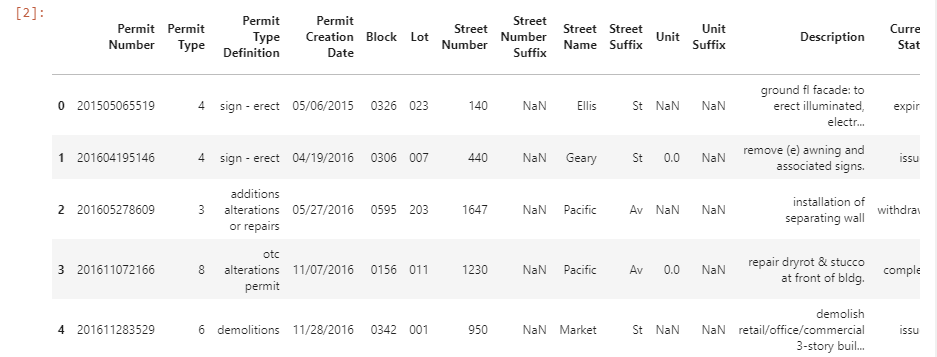
The first five records of the data are shown in figure 3 below.

Fig 3: Preview of the data

The data set has 198900 records (rows) and 43 variables (columns).

**Data Cleaning**

The data set has lots of missing values. It also has some variables that are not of interest to this analysis. The missing values were correctly dropped and using feature engineering, new features (variables) were created.

**Data Exploration**

The data set was explored in order to find patterns and insights.

* Univariate analysis

The histograms of the numerical variables are shown in figure 4

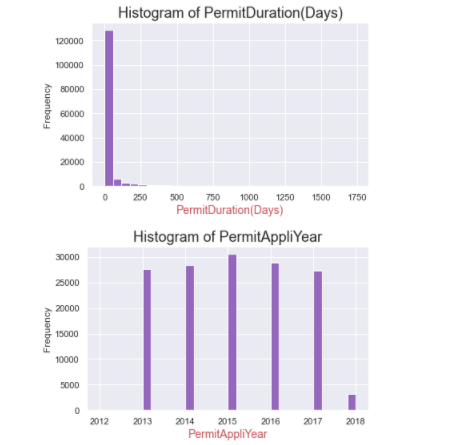


Fig 4: Histograms of numerical variables

1. ‘PermitDuration(Days)': this variable follows an exponential distribution with outliers on the right

hand side of the distribution. It has a mean of about 32 days, a standard deviation of 101 days which

shows that the data points are spread apart from the mean and a median of 0 days.

1. 'PermitAppliYear': The distribution of this variable follows a fairly normal distribution. It has a mean of about 2015, a standard deviation of 1.4 which shows that the data points are very close to the mean and a median of 2015.
2. 'Project\_cost': The distribution of this variable follows an exponential distribution. The box plot

shows that it contains too many outliers. It has a mean of about $160,000, a standard deviation of

about $4,000,000 which shows that the data points are spread apart from the mean and a median of

$12,500.

1. 'Num\_of\_Stories': The distribution of this variable follows an exponential distribution. It has a mean of about 6 stories, a standard deviation of 9 stories which shows that the data points are far from the

mean and a median of 3 stories.

The boxplot of the ‘project cost’ of each observation in the data set is shown in figure 5. It verifies the fact that the project cost variable contains a lot of outliers in the upper region i.e too many high values which are not representative of the whole data.



Fig 5: Box plot of the Project Cost

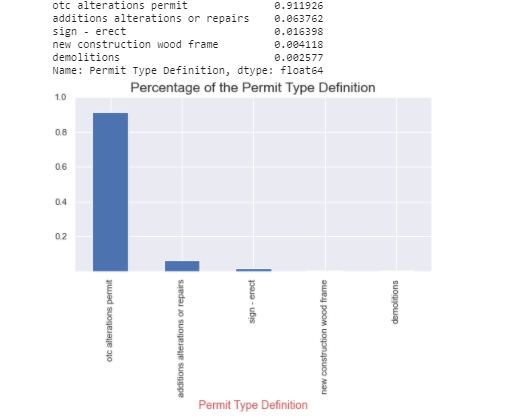
The frequency distribution (in percentage) and bar chart of ‘Permit Type Definition’ of the top five types of building permits is shown below in figure 6.

Fig 6: Bar plot of the percentage of the building permit type definition.

It is clearly seen that ‘otc alterations permit’ is by far the most common kind of building permit filled. ‘additions alterations or repairs’ is a distant second.

After the building permit has been issued, the ratio of the completed building projects is shown in figure 7.

A value of 66% shows that most projects have been completed.

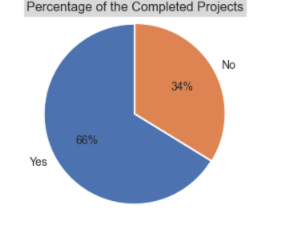


Fig 7: Pie chart of the Completed Projects.

* Multivariate Analysis

From the analysis in figure 8, it can be seen that most of the building permit applications were made in the year 2015. 2012 had the least number of permit applications.

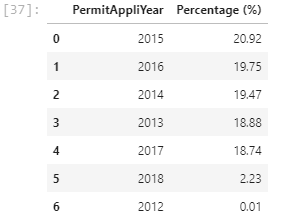


Fig 8: Percentage of Year the Building Permit Application was created or filed

Figure 9 below shows the neighbourhoods with the most building permit applications.

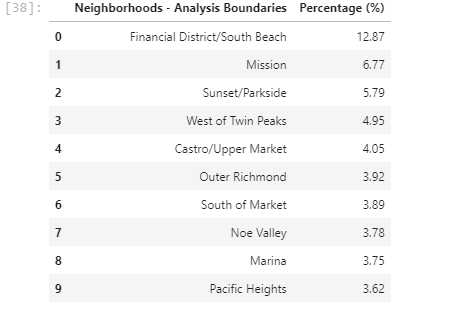


Fig 9: Percentage of the Neighbourhoods that got Building Permit Application Approved

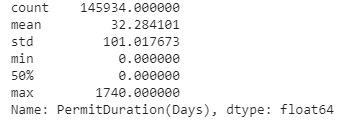
The average time it takes for a building permit to be approved is between o to 26 days (figure 10). The high value of the standard deviation shows that the data points are spread out over a large range of values from the mean meaning that it could in fact take more than 26 days.

Fig 10: Descriptive statistics of the PermitDuration (in days) variable

The building permit application that took the longest time to be approved is shown in figure 11 below.

Fig 11: Building Permit Application that too the longest to be approved.

The most common type of 'Existing Use' for the building permit application is shown in figure 12.

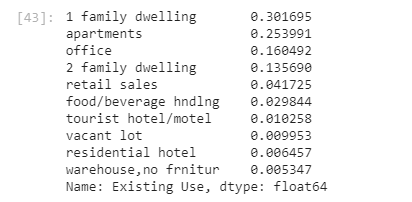


Fig 12: Percentage of 'Existing Use'