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)

[1]:
```
# import the necessary libraries
import numpy as np
import pandas as pd

# import visualization
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()

```

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.

```
[2]:
# load the data
building_df = pd.read_csv('Files/Building_Permits.csv', low_memory=False)

# preview of the data
building_df.head()
```

Fig 2: Loading of the data

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

[2]:

	Permit Number	Permit Type	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Number Suffix	Street Name	Street Suffix	Unit	Unit Suffix	Description	Curre Stat
0	201505065519	4	sign - erect	05/06/2015	0326	023	140	NaN	Ellis	St	NaN	NaN	ground fl facade: to erect illuminated, electr...	expir
1	201604195146	4	sign - erect	04/19/2016	0306	007	440	NaN	Geary	St	0.0	NaN	remove (e) awning and associated signs.	issu
2	201605278609	3	additions alterations or repairs	05/27/2016	0595	203	1647	NaN	Pacific	Av	NaN	NaN	installation of separating wall	withdra
3	201611072166	8	otc alterations permit	11/07/2016	0156	011	1230	NaN	Pacific	Av	0.0	NaN	repair dryrot & stucco at front of bldg.	comple
4	201611283529	6	demolitions	11/28/2016	0342	001	950	NaN	Market	St	NaN	NaN	demolish retail/office/commercial 3-story buil...	issu

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

I. ‘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.

II. '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.

III. '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.

IV. '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.

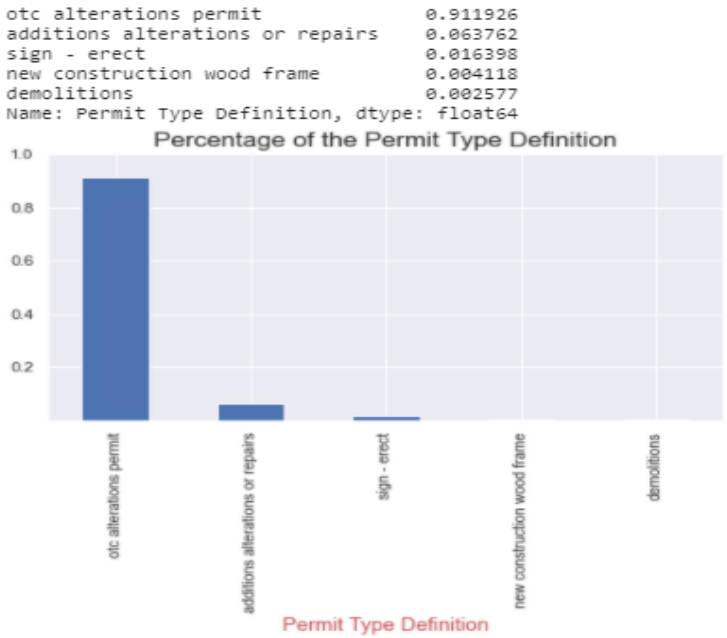

Fig 6: Bar plot of the percentage of the building permit type definition.

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.

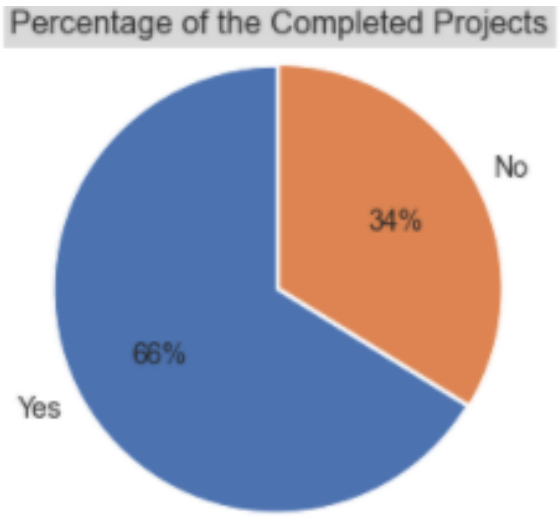

Fig 7: Pie chart of the Completed Projects.

- 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.

[37]:

	PermitAppliYear	Percentage (%)
0	2015	20.92
1	2016	19.75
2	2014	19.47
3	2013	18.88
4	2017	18.74
5	2018	2.23
6	2012	0.01

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.

[38]:

	Neighborhoods - Analysis Boundaries	Percentage (%)
0	Financial District/South Beach	12.87
1	Mission	6.77
2	Sunset/Parkside	5.79
3	West of Twin Peaks	4.95
4	Castro/Upper Market	4.05
5	Outer Richmond	3.92
6	South of Market	3.89
7	Noe Valley	3.78
8	Marina	3.75
9	Pacific Heights	3.62

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.

```
count    145934.000000
mean         32.284101
std         101.017673
min           0.000000
50%           0.000000
max        1740.000000
Name: PermitDuration(Days), dtype: float64
```

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.

[42]:

	Permit Creation Date	Neighborhoods - Analysis Boundaries	Issued Date	PermitDuration(Days)	Project_cost
171052	2013-02-07	Noe Valley	2017-11-13	1740	600000.0

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.

```
[43]: 1 family dwelling       0.301695
      apartments              0.253991
      office                  0.160492
      2 family dwelling       0.135690
      retail sales            0.041725
      food/beverage hndlng    0.029844
      tourist hotel/motel     0.010258
      vacant lot              0.009953
      residential hotel       0.006457
      warehouse,no frnitur    0.005347
      Name: Existing Use, dtype: float64
```

Fig 12: Percentage of 'Existing Use'