Analysis on Building Permit Data Set

Bi-Weekly By ELOMENA ENAHORO

Data analysis is the use of available data, which is visualized in different forms and used for various permutations and combinations to become familiar with the data and derive meaningful observations.
Descriptive statistics is about describing and summarizing data.
Some analysis involve:

i. Univariate analysis
ii. Bi-variate analysis
iii. Multivariate analysis

Importing the necessary Libraries and Data on Jupyter Notebook

In [1]:
```
#importing the necessary libraries and displaying the data head
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt

#importing the data set and parsing the date columns so python can recognize it as a date column
data = pd.read_csv("Building_Permits-1.csv", parse_dates=["Permit Creation Date", "Current Status Date", "Filed Date",
data.head()
```

```
//anaconda3/envs/py38/lib/python3.8/site-packages/IPython/core/interactiveshell.py:3165: DtypeWarning: Columns (22,32
) have mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

Out[1]:

	Permit Number	Permit Type	Permit Type Definition	Permit Creation Date	Block	Lot	Street Number	Street Number Suffix	Street Name	Street Suffix	...	Existing Construction Type	Existing Construction Type Description	Proposed Construction Type	Proposed Construction Type Description	Site Permit
0	201505065519	4	sign - erect	2015-05-06	0326	023	140	NaN	Ellis	St	...	3.0	constr type 3	NaN	NaN	NaN
1	201604195146	4	sign - erect	2016-04-19	0306	007	440	NaN	Geary	St	...	3.0	constr type 3	NaN	NaN	NaN
2	201605278609	3	additions alterations or repairs	2016-05-27	0595	203	1647	NaN	Pacific	Av	...	1.0	constr type 1	1.0	constr type 1	NaN
3	201611072166	8	otc alterations permit	2016-11-07	0156	011	1230	NaN	Pacific	Av	...	5.0	wood frame (5)	5.0	wood frame (5)	NaN
4	201611283529	6	demolitions	2016-11-28	0342	001	950	NaN	Market	St	...	3.0	constr type 3	NaN	NaN	NaN

5 rows × 43 columns

Libraries:

i. Pandas
ii. Numpy
iii. Seaborn
iv. Matplotlib

Data Cleaning

Preparing the data set for use usually following some tedious work of removing null values and transforming the data for analysis.
In this data set, the columns containing more than half null values were dropped.

```
'''Dropping Columns that have too much null values to perform analysis on'''

df = data1.drop(columns = ["Site Permit", "Voluntary Soft-Story Retrofit", "TIDF Compliance",
                           "Fire Only Permit", "Completed Date", "Unit Suffix", "Street Number Suffix",
                            "Structural Notification", "Unit"])
```

After the columns were dropped

```
Data columns (total 34 columns):
 #   Column                                  Non-Null Count   Dtype
---  ------                                  --------------   -----
 0   Permit Number                           198900 non-null  object
 1   Permit Type                             198900 non-null  int64
 2   Permit Type Definition                  198900 non-null  object
 3   Permit Creation Date                    198900 non-null  datetime64[ns]
 4   Block                                   198900 non-null  object
 5   Lot                                     198900 non-null  object
 6   Street Number                           198900 non-null  int64
 7   Street Name                             198900 non-null  object
 8   Street Suffix                           196132 non-null  object
 9   Description                             198610 non-null  object
 10  Current Status                          198900 non-null  object
 11  Current Status Date                     198900 non-null  datetime64[ns]
 12  Filed Date                              198900 non-null  datetime64[ns]
 13  Issued Date                             183960 non-null  datetime64[ns]
 14  First Construction Document Date        183954 non-null  datetime64[ns]
 15  Number of Existing Stories              156116 non-null  float64
 16  Number of Proposed Stories              156032 non-null  float64
 17  Permit Expiration Date                  147020 non-null  datetime64[ns]
 18  Estimated Cost                          160834 non-null  float64
 19  Revised Cost                            192834 non-null  float64
 20  Existing Use                            157786 non-null  object
 21  Existing Units                          147362 non-null  float64
 22  Proposed Use                            156461 non-null  object
 23  Proposed Units                          147989 non-null  float64
 24  Plansets                                161591 non-null  float64
 25  Existing Construction Type              155534 non-null  float64
 26  Existing Construction Type Description  155534 non-null  object
 27  Proposed Construction Type              155738 non-null  float64
 28  Proposed Construction Type Description  155738 non-null  object
 29  Supervisor District                     197183 non-null  float64
 30  Neighborhoods - Analysis Boundaries     197175 non-null  object
 31  Zipcode                                 197184 non-null  float64
 32  Location                                197200 non-null  object
 33  Record ID                               198900 non-null  int64
```

Uni-variate Analysis

The code below was used to crate different tables showing the frequency of the columns

```
In [8]: # displaying the occurrance of each variable in a column
        for i in df:
            a = df[i].value_counts()
            print("--------"*25)
            print(i)
            print()
            print(a)
```

Columns in the data set and their most frequent variable
For variables/columns which have unique and numerical values are not included below

Permit Type Definition:
 i. otc alteration permit(178,844)
ii. additions alterations or repairs(14,663).
Street Name:
 i. Market(5,443)
ii. California(4,587).
Description:
 i. Street Space(2,2163)
ii. reroofing(9,465).
Current Status:
 i. Complete(97,077)
ii. Issued(83,559).
Number of Existing Stories:
 i. 2.0(52,767)
ii. 3.0(45,470).
Number of Proposed Stories:
 i. 2.0(50,991)
ii. 3.0(46,802).
Existing Use:
 i. 1 Family dwelling(46,766)
ii. Apartments(40,798).
Existing Units:
 i. 1.0(47,346)
ii. 2.0(29,134).
Proposed Use:
 i. 1 Family dwelling(46,346)
ii. Apartments(43,032).
Proposed Units:
 i. 1.0(46,913)
ii. 2.0(26,884).
Existing Construction Type Description:
 i. Wood Frame 5(113,350)
ii. Construction type 1(28,072).
Proposed Construction Type Description:
 i. Wood Frame 5(114,382)
ii. Construction type 1(27,841).
Supervisor District:
 i. 3.0(28,649)
ii. 8.0(26,760).
Neighborhoods - Analysis Boundaries:
 i. Financial District/South Beach(21,816)
ii. Mission(14,681).

```
Zipcode:
 i.   94110.0(17,837)
ii.   94114.0(13,404).
```

Time Analysis:

Several date variables were parsed using date-time to be able to get the difference in date and wait time to get a permit.

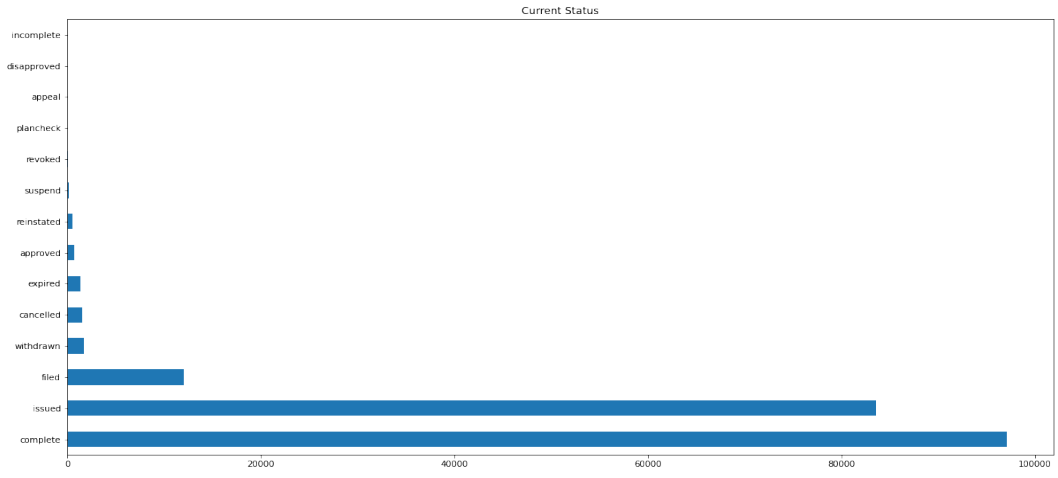
"Permit Creation Date" was subtracted from the "Issued Date" to get the exact wait time.

```
df.days_to_issue.describe()
```

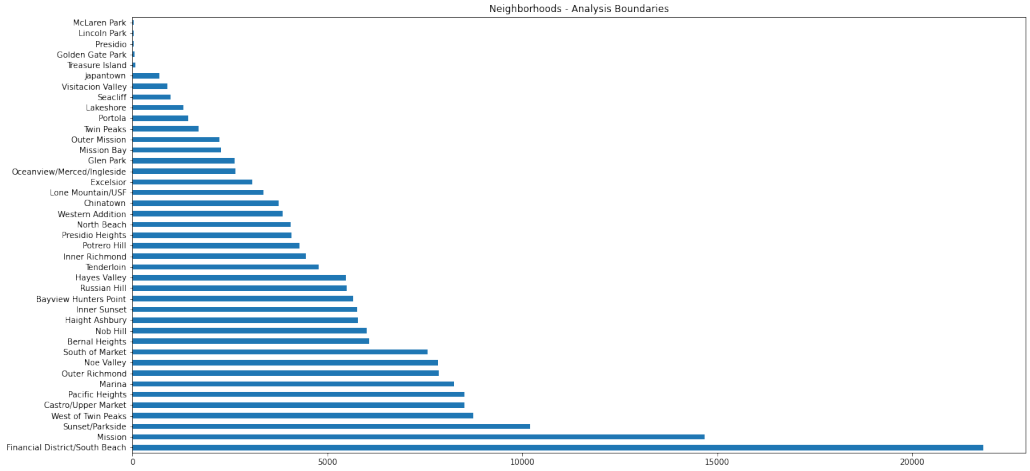
```
count    183960.000000
mean         26.054697
std          91.061716
min           0.000000
25%           0.000000
50%           0.000000
75%           6.000000
max        1740.000000
Name: days_to_issue, dtype: float64
```

The following are some of the data visualized:

i. The data visualization shown below is of the "Current Status" Column, it displays that a lot of building permits haved been issued out and even more have been completed while a very small portion of the building permits have been either disapproved, withdrawn, suspended or revoked.

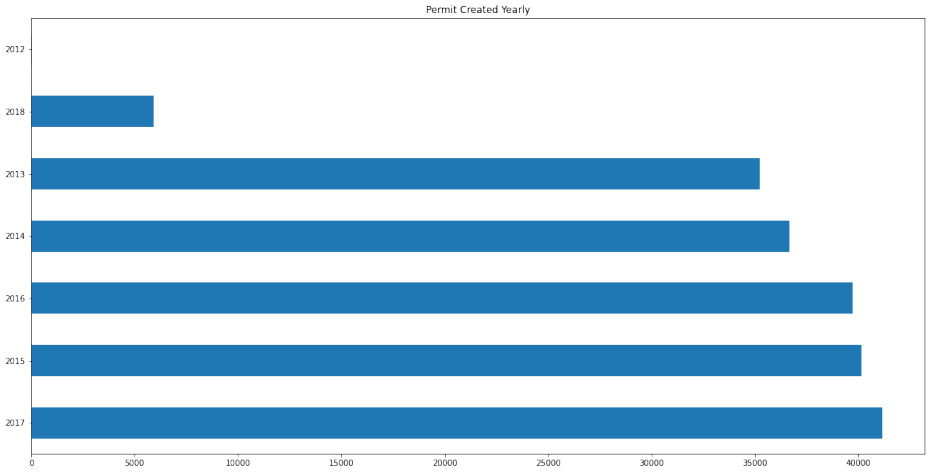


ii. The data visualization shown below is of the “Neighborhoods –Analysis Boundaries” Column, it displays that more building permit requests were made from the Financial District/South Beach neighboorhood.



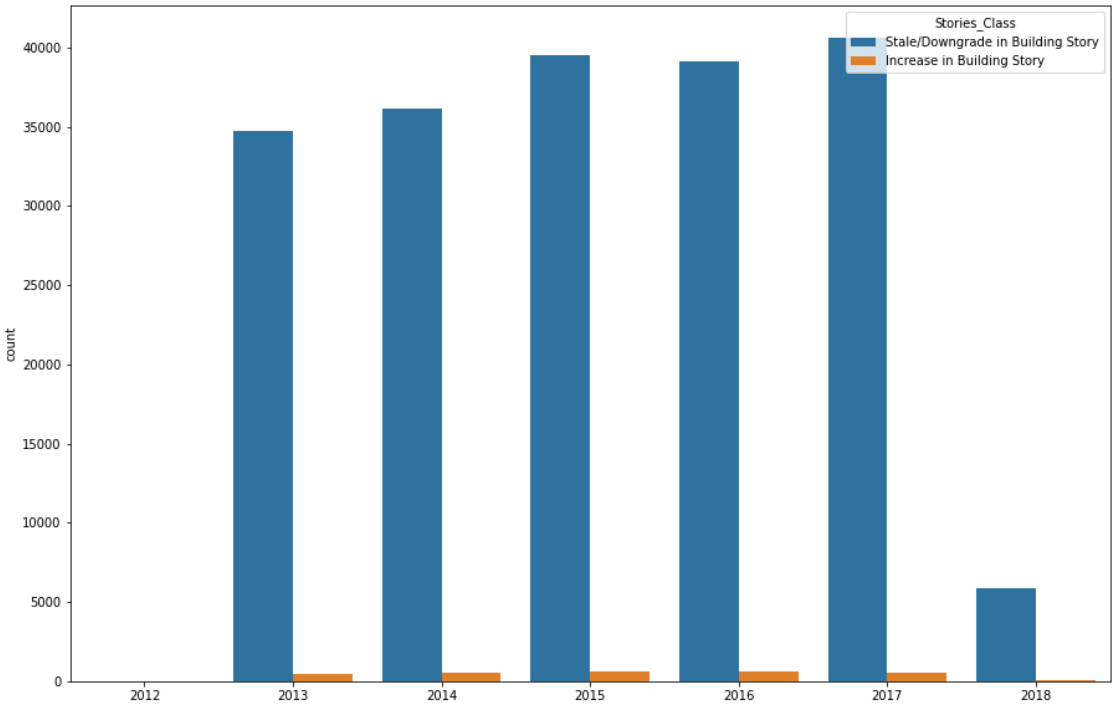
Multi-Variate Analysis

i. Below displays the permit request made per year, showing that the least request was made in 2012 and it has increased over the years till 2017, with the year 2018 only in February.

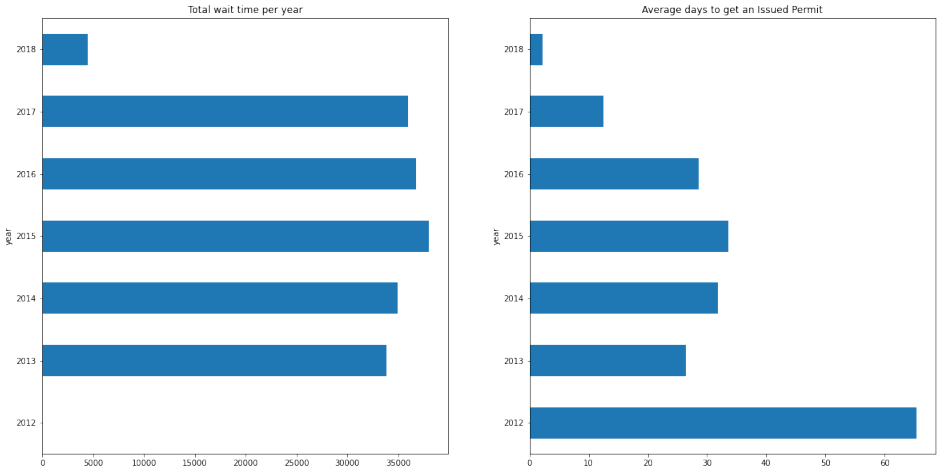


ii. Below displays year vs the proposed building class i.e. if an upgrade was being made to the building level.

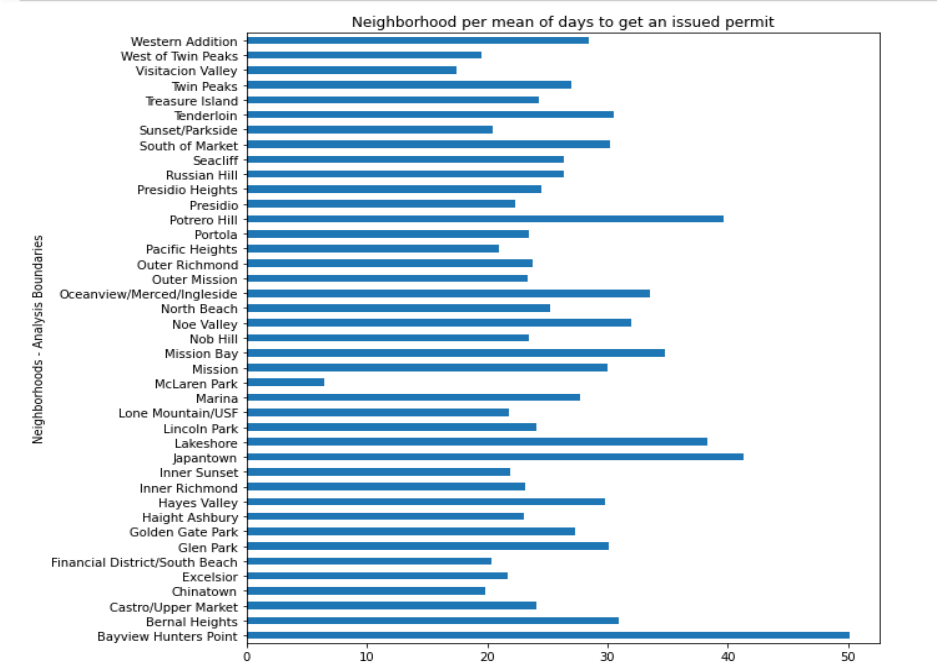
Stories_Class	Decrease in Building Story	Increase in Building Story	Stale in Building Story	Total
year				
2012	0	0	22	22
2013	15	459	34754	35228
2014	13	533	36140	36686
2015	14	642	39494	40150
2016	32	599	39090	39721
2017	14	544	40598	41156
2018	2	70	5865	5937
Total	90	2847	195963	198900



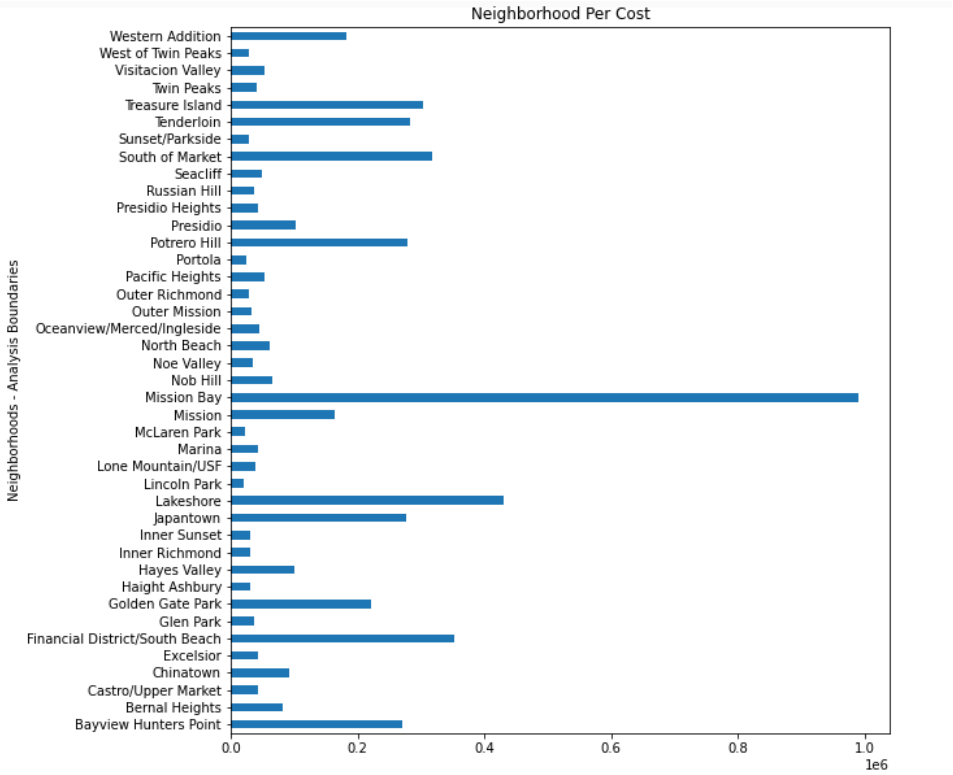
iii. Below displays the comparisons between the “year” and the “days to issue Building Permit”.
The first figure displaying the total wait time per year, with 2015 being the most and the second figure displays the average days per year, with 2018 being the least which signified that getting a permit in 2018 took lesser days.



iv. Below displays the comparisons between the “Neighborhoods” and the “mean days to issue Building Permit”, with the least mean being McLaren Park i.e. Building permits requests from McLaren Park too lesser days to get the Building Permits Issued.



v. Below displays the comparisons between the “Neighborhoods” and the “Cost”, with Mission Bay having the highest mean i.e. the Mission Bay neighborhood on an average spends more money on building than any other neighborhood.



vi. Below displays the comparisons between the “Year” and the “Permit Type Definition”, otc alteration permit being the most requested permit in 2017 .
It is obvious that the “Permit Type Definition” increases over the years, with 2018 only accounted until February.