Data Science Assignment - Ishaan Vasant

NYC DOB Permit Issuance - Exploratory Data Analysis

Importing Dependencies

In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import datetime
import warnings
warnings.filterwarnings('ignore')
from IPython.display import display
```

Changing Directory

In [3]:
```
cd
```

```
/Users/ishaan
```

In [4]:
```
cd Downloads
```

```
/Users/ishaan/Downloads
```

Reading and Displaying NYC DOB Permit Issuance Dataset

In [5]:
```
df = pd.read_csv('DOB_Permit_Issuance.csv')
```

In [284]:
```
# pd.options.display.max_columns = None
# display(df)
```

Data Insights

In [7]:
```
df.shape
```

Out[7]:
```
(3508249, 60)
```

In [285]:
```
# df.info()
```

• **Data has over 3.5 million entries and 60 features**

• **Data types vary from int and float to a range of objects**

Data Visualization

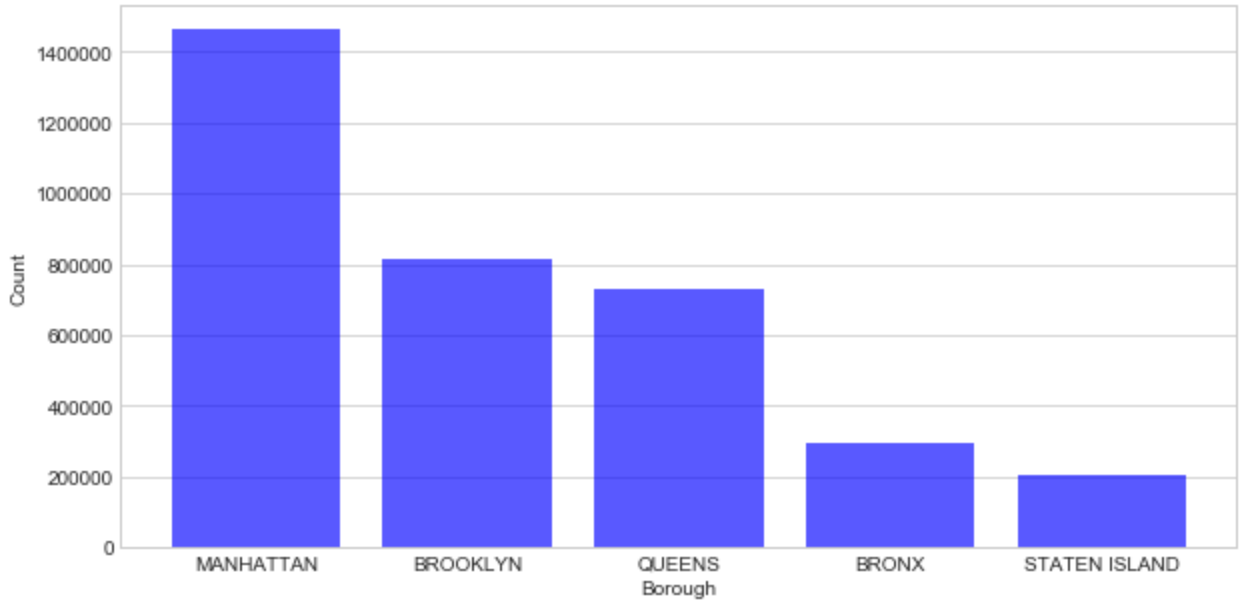
Boroughs

In [24]:
```
df['BOROUGH'].value_counts()
```

Out[24]:
```
MANHATTAN        1462213
BROOKLYN          813934
QUEENS            731242
BRONX             296495
STATEN ISLAND     204365
Name: BOROUGH, dtype: int64
```

In [87]:
```
values = df['BOROUGH'].value_counts().keys().tolist()
counts = df['BOROUGH'].value_counts().tolist()
plt.figure(figsize=(10,5))
plt.bar(values,counts,color='b',alpha=0.65)
plt.grid(axis='x', alpha=0)
plt.xlabel('Borough')
plt.ylabel('Count')
```

Out[87]:
```
Text(0,0.5,'Count')
```



• **Most of the construction permits apply to the borough of Manhattan with Staten Island gathering the least**

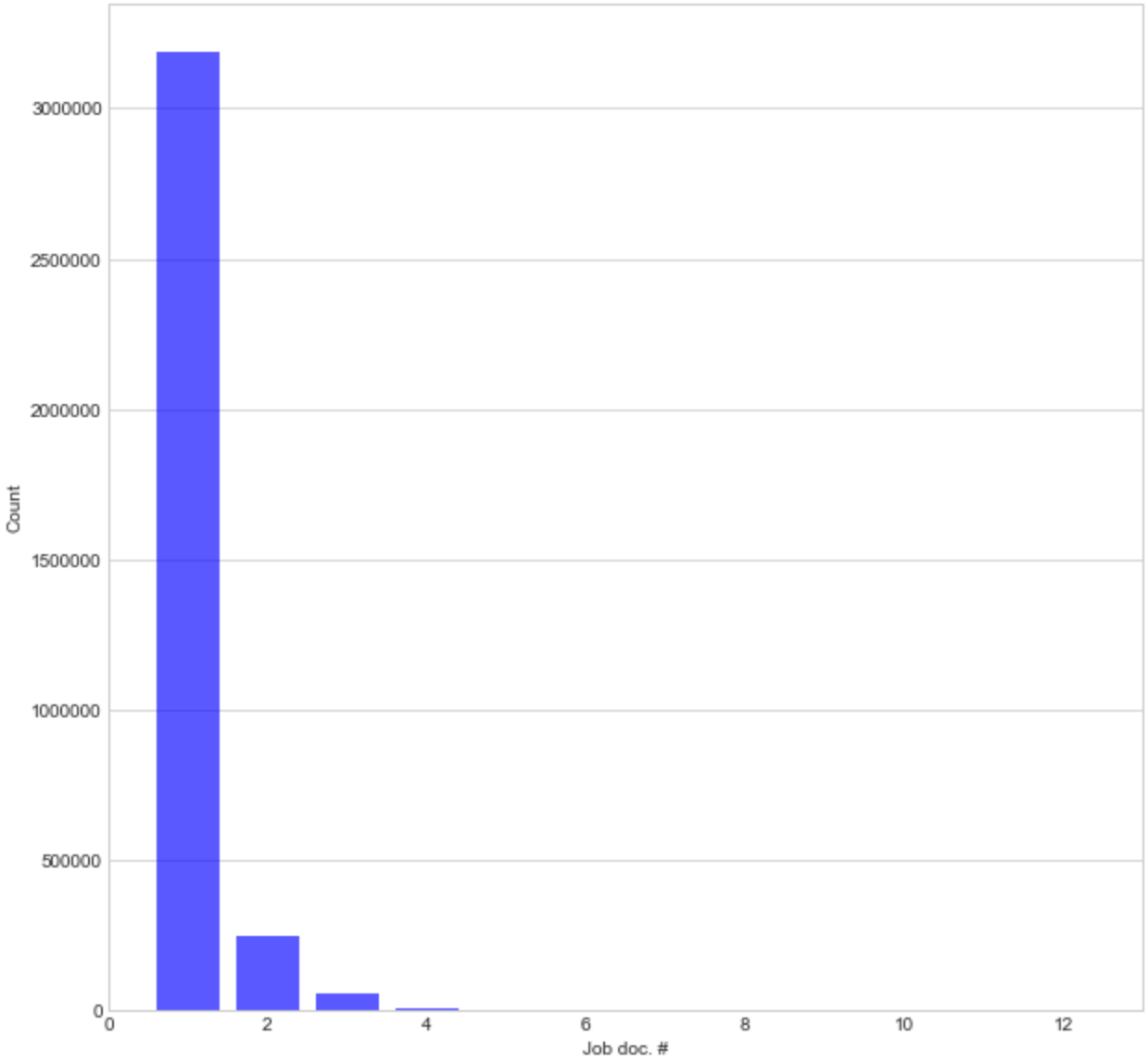
Job Document Number

In [15]:
```
df['Job doc. #'].value_counts()
```

Out[15]:
```
1     3189810
2      249293
3       55988
4        9471
5        2499
6         736
7         298
8         106
9          39
10          5
12          2
11          2
Name: Job doc. #, dtype: int64
```

In [83]:
```
values = df['Job doc. #'].value_counts().keys().tolist()
counts = df['Job doc. #'].value_counts().tolist()
plt.figure(figsize=(10,10))
plt.bar(values,counts,color='b',alpha=0.65)
plt.grid(axis='x', alpha=0)
plt.xlabel('Job doc. #')
plt.ylabel('Count')
```

Out[83]: Text(0,0.5,'Count')



- **Only Job Doc. # 1, 2 and 3 are significant with 1 gathering over 3 million permits**

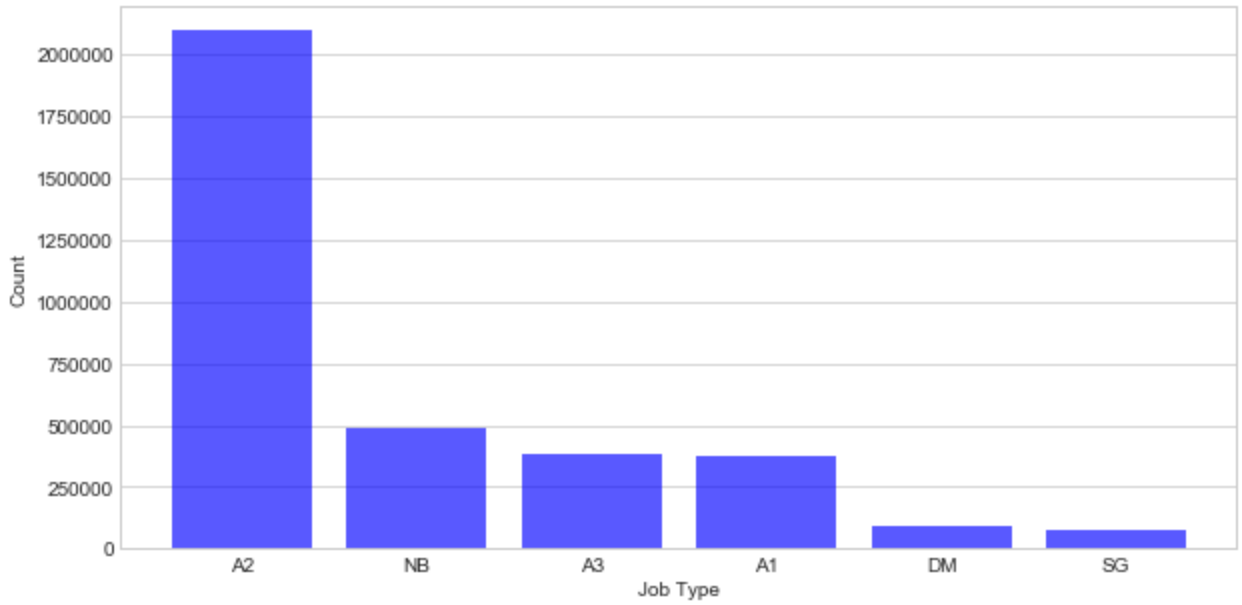
Job Type

```
In [33]: df['Job Type'].value_counts()
```

```
Out[33]: A2    2094465
         NB     491768
         A3     384782
         A1     371457
         DM      90049
         SG      75728
         Name: Job Type, dtype: int64
```

```
In [94]: values = df['Job Type'].value_counts().keys().tolist()
         counts = df['Job Type'].value_counts().tolist()
         plt.figure(figsize=(10,5))
         plt.bar(values,counts,color='b',alpha=0.65)
         plt.grid(axis='x', alpha=0)
         plt.xlabel('Job Type')
         plt.ylabel('Count')
```

```
Out[94]: Text(0,0.5,'Count')
```



Data from the DOB Job Application Filings dataset was collected and it revealed what the above acronyms stand for: -

• A1 = Alteration Type I, A major alteration that will change the use, egress, or occupancy of the building.

• A2 = Alteration Type II, An application with multiple types of work that do not affect the use, egress, or occupancy of the building.

• A3 = Alteration Type III, One type of minor work that doesn't affect the use, egress, or occupancy of the building.

• NB = New Building, An application to build a new structure. “NB” cannot be selected if any existing building elements are to remain—for example a part of an old foundation, a portion of a façade that will be incorporated into the construction, etc.

• DM = Demolition, An application to fully or partially demolish an existing building.

• **It is seen that most permits were given to the A2 Job Type**

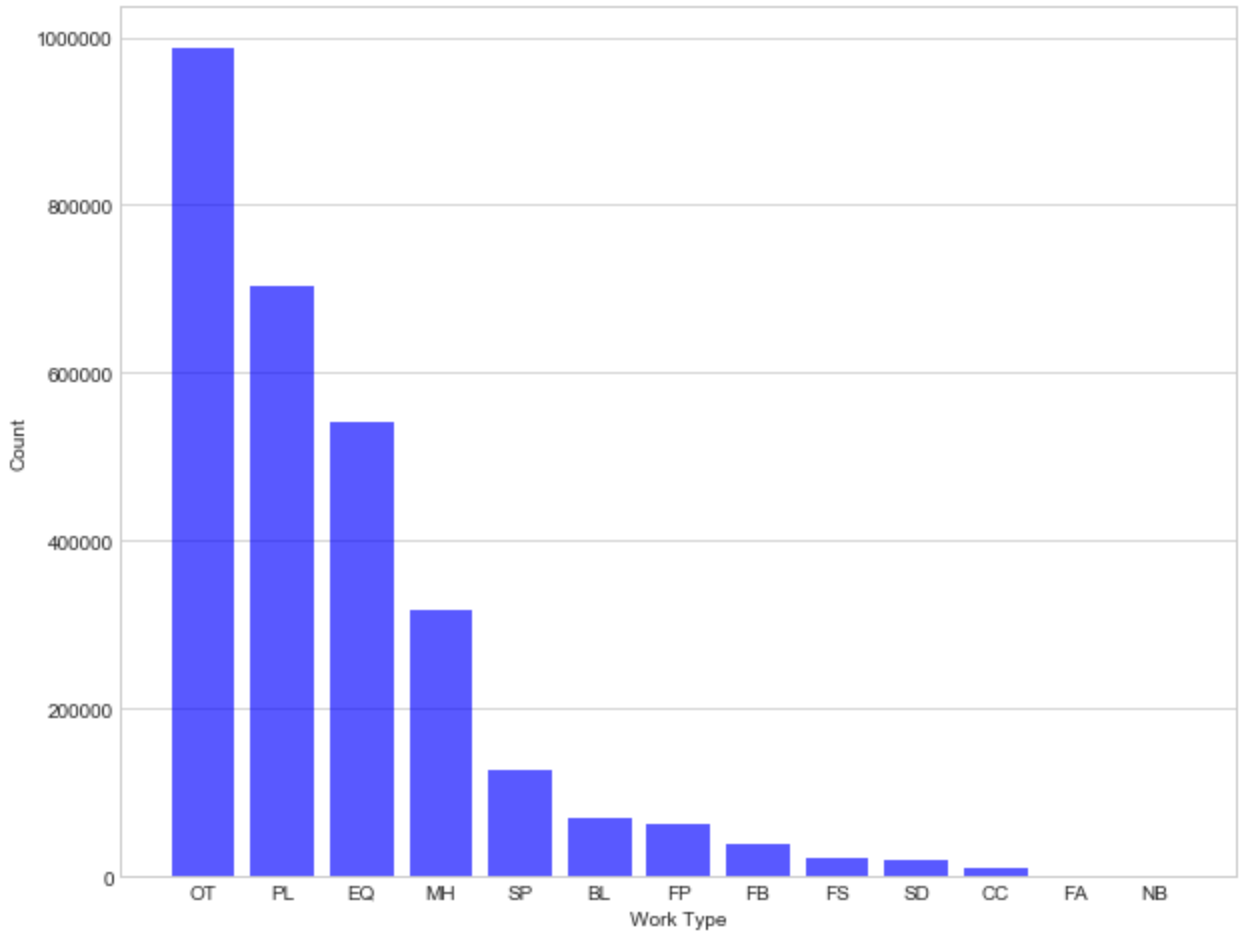
Work Type

In [34]:
```
df['Work Type'].value_counts()
```

Out[34]:
```
OT    986608
PL    704541
EQ    541170
MH    317608
SP    125786
BL     68916
FP     62084
FB     38900
FS     22558
SD     18775
CC     11341
FA       244
NB         1
Name: Work Type, dtype: int64
```

In [95]:
```
values = df['Work Type'].value_counts().keys().tolist()
counts = df['Work Type'].value_counts().tolist()
plt.figure(figsize=(10,8))
plt.bar(values,counts,color='b',alpha=0.65)
plt.grid(axis='x', alpha=0)
plt.xlabel('Work Type')
plt.ylabel('Count')
```

Out[95]:
```
Text(0,0.5,'Count')
```



Data from the NYC buildings website was collected and it revealed what the above acronyms stand for: -

BL: Boiler

CC: Curb Cut

CH: Chute

DM: Demolition and Removal

EQ: Construction Equipment

EW: Equipment Work

FA: Fire Alarm

FB: Fuel Burning

FN: Fence

FP: Fire Suppression

FS: Fuel Storage

MH: Mechanical/HVAC

OT: Other

PL: Plumbing

SD: Standpipe

SF: Scaffold

SG: Sign

SP: Sprinkler

• It is seen that most permits were given to the 'Other' category of work types and the second most were given to Plumbing work

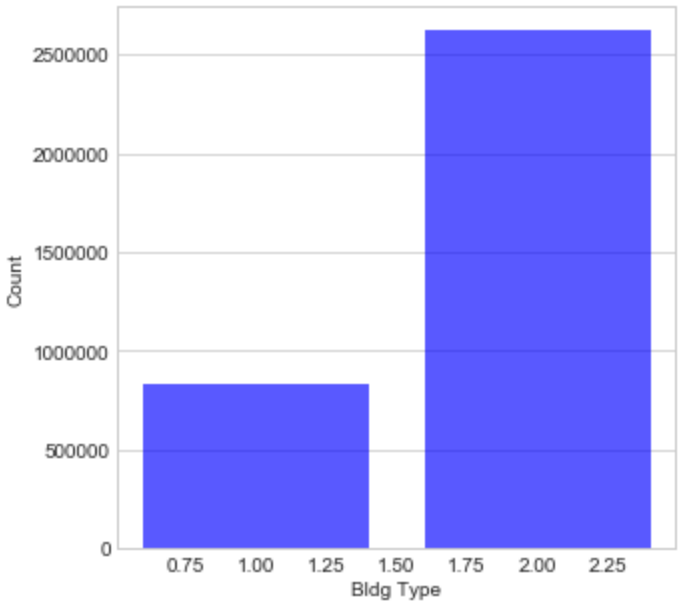
Building Type

```
In [58]: df['Bldg Type'].value_counts()

Out[58]: 2.0    2619873
         1.0     833899
         Name: Bldg Type, dtype: int64
```

In [96]:
```
values = df['Bldg Type'].value_counts().keys().tolist()
counts = df['Bldg Type'].value_counts().tolist()
plt.figure(figsize=(5,5))
plt.bar(values,counts,color='b',alpha=0.65)
plt.grid(axis='x', alpha=0)
plt.xlabel('Bldg Type')
plt.ylabel('Count')
```

Out[96]:
```
Text(0,0.5,'Count')
```



• 2.6 million permits were given for 2-person occupancy type buildings and less than a million permits were given to single-person occupancy type buildings

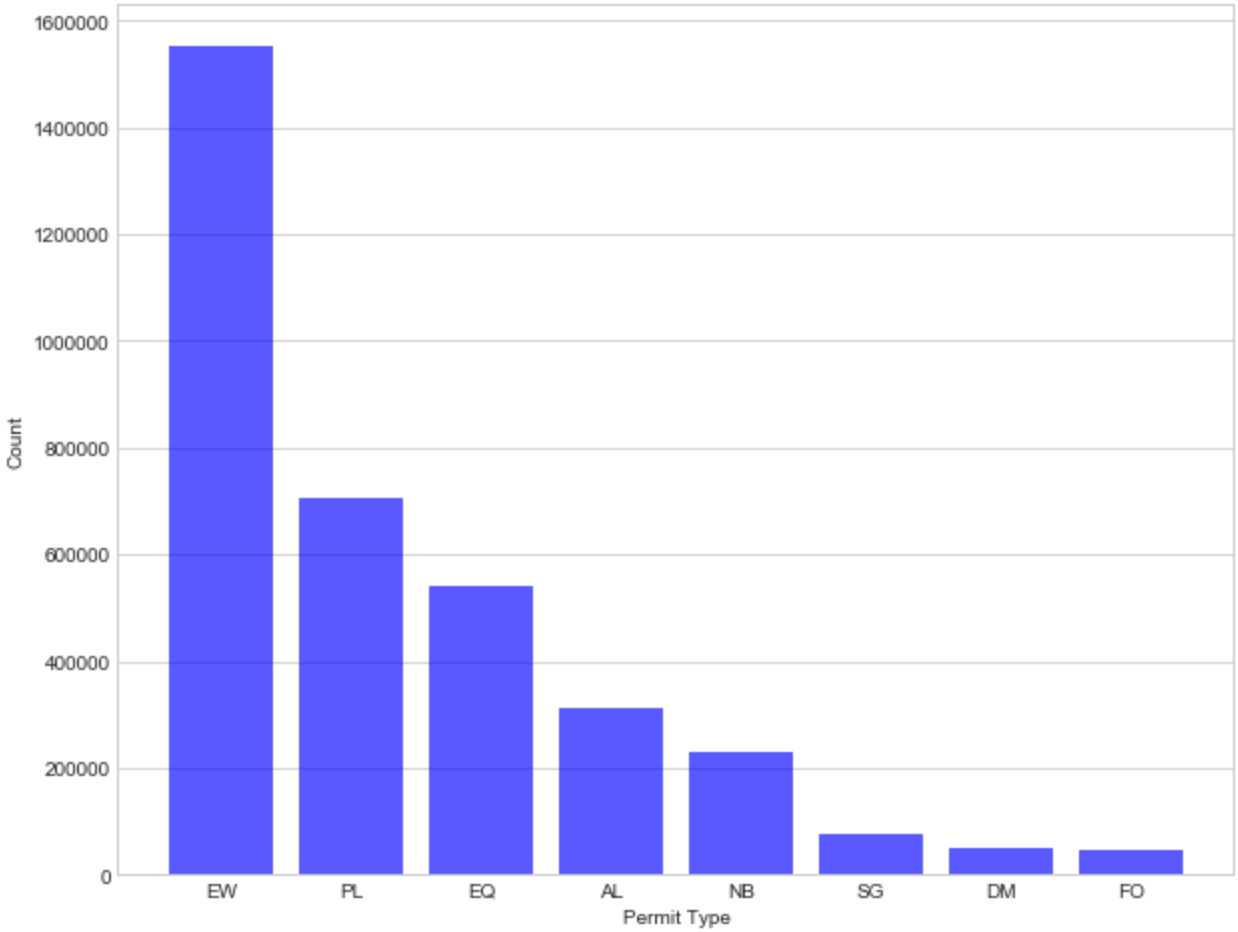
Permit Type

In [63]:
```
df['Permit Type'].value_counts()
```

Out[63]:
```
EW    1551850
PL     704540
EQ     541170
AL     312159
NB     227844
SG      75708
DM      50386
FO      44591
Name: Permit Type, dtype: int64
```

In [97]:
```
values = df['Permit Type'].value_counts().keys().tolist()
counts = df['Permit Type'].value_counts().tolist()
plt.figure(figsize=(10,8))
plt.bar(values,counts,color='b',alpha=0.65)
plt.grid(axis='x', alpha=0)
plt.xlabel('Permit Type')
plt.ylabel('Count')
```

Out[97]:
```
Text(0,0.5,'Count')
```



Data from the NYC buildings website was collected and it revealed what the above acronyms stand for: -

AL: Architectural

EW: Equipment Work

EQ: Construction Equipment

FO: Foundation

NB: New Building

PL: Plumbing

SG: Sign

DM: Demolition and Removal

• It is seen that most permits types were of the Equipment work category and the second most were of the Plumbing category

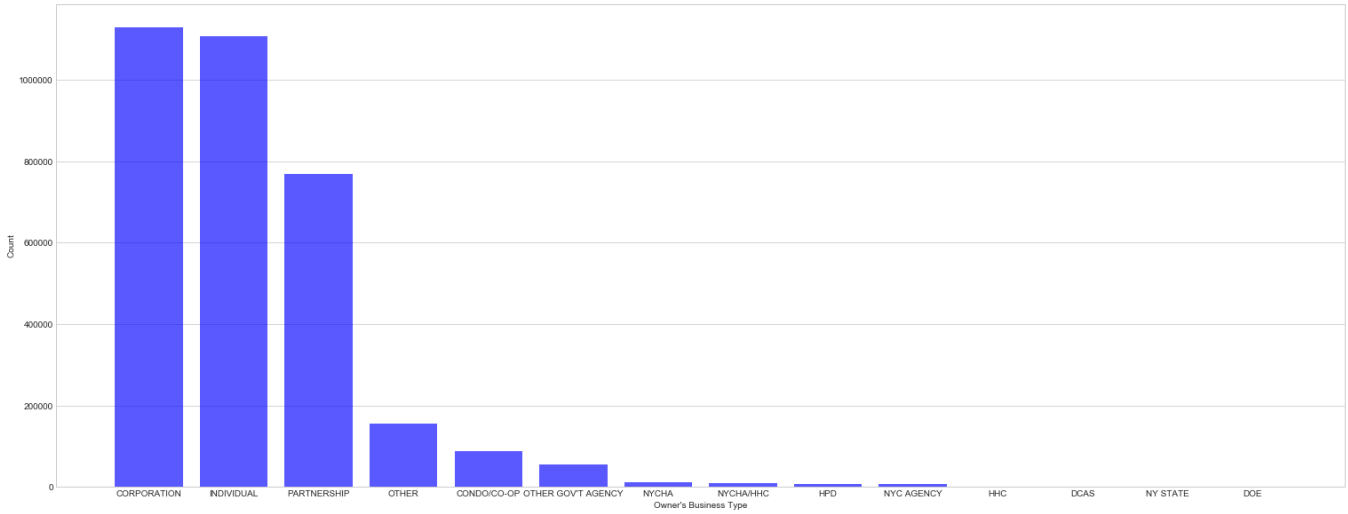
Owner's Business Type

In [36]:
```
df["Owner's Business Type"].value_counts()
```

Out[36]:
```
CORPORATION           1130020
INDIVIDUAL            1106781
PARTNERSHIP            769321
OTHER                  155929
CONDO/CO-OP             87028
OTHER GOV'T AGENCY      55623
NYCHA                   11523
NYCHA/HHC                9118
HPD                      6865
NYC AGENCY               6849
HHC                      1277
DCAS                     1199
NY STATE                 1178
DOE                       950
Name: Owner's Business Type, dtype: int64
```

In [103]:
```
values = df["Owner's Business Type"].value_counts().keys().tolist()
counts = df["Owner's Business Type"].value_counts().tolist()
plt.figure(figsize=(26,10))
plt.bar(values,counts,color='b',alpha=0.65)
plt.grid(axis='x', alpha=0)
plt.xlabel("Owner's Business Type")
plt.ylabel('Count')
```

Out[103]:
```
Text(0,0.5,'Count')
```



• Around 1.1 million permits were for buildings whose owner's business type was either 'Individual' or 'Corporation' with 'Partnership' gaining close to 800,000 permits

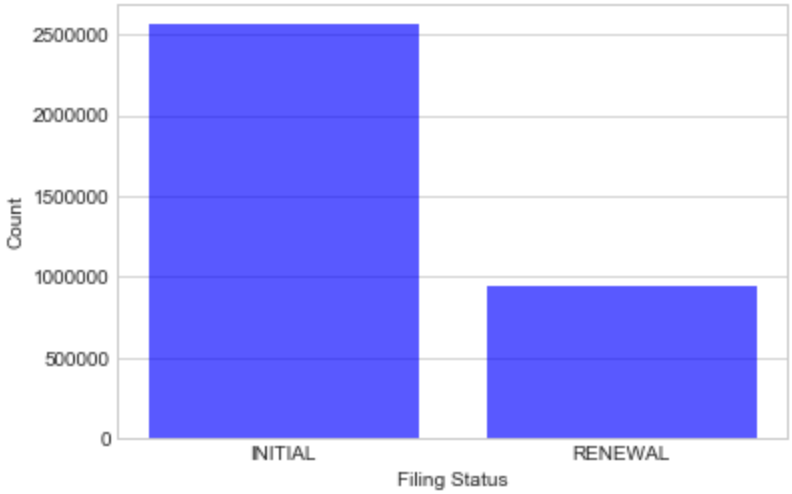
Filing Status

In [46]:
```
df['Filing Status'].value_counts()
```

Out[46]:
```
INITIAL    2559794
RENEWAL     948455
Name: Filing Status, dtype: int64
```

In [104]:
```
values = df['Filing Status'].value_counts().keys().tolist()
counts = df['Filing Status'].value_counts().tolist()
plt.bar(values,counts,color='b',alpha=0.65)
plt.grid(axis='x', alpha=0)
plt.xlabel('Filing Status')
plt.ylabel('Count')
```

Out[104]:
```
Text(0,0.5,'Count')
```

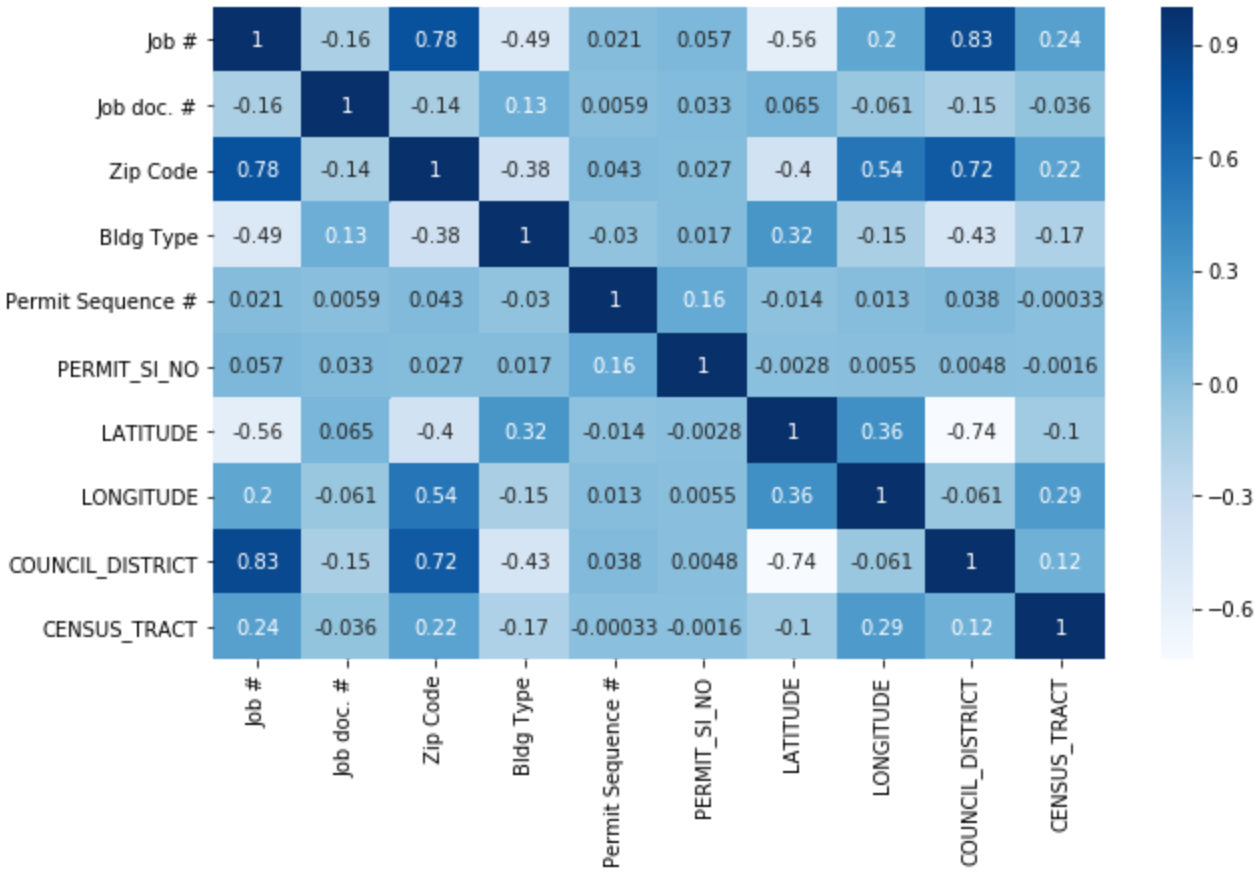


- **Over 2.5 million permits were Initial permits and just under 1 million permits were renewals**

Correlation

In [51]:
```
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(),cmap='Blues',annot=True)
```

Out[51]: <matplotlib.axes._subplots.AxesSubplot at 0x1a23072b70>



• The first digit of the Job Numbers are assigned based on which borough it represents: -

• 1 = Manhattan

• 2 = Bronx

• 3 = Brooklyn

• 4 = Queens

• 5 = Staten Island

Therefore, it makes perfect sense that the correlation heatmap shows a positive correlation between the Job # and geographical features such as Zip Code and Council_District (council district for the building's address, geocoded based on address)

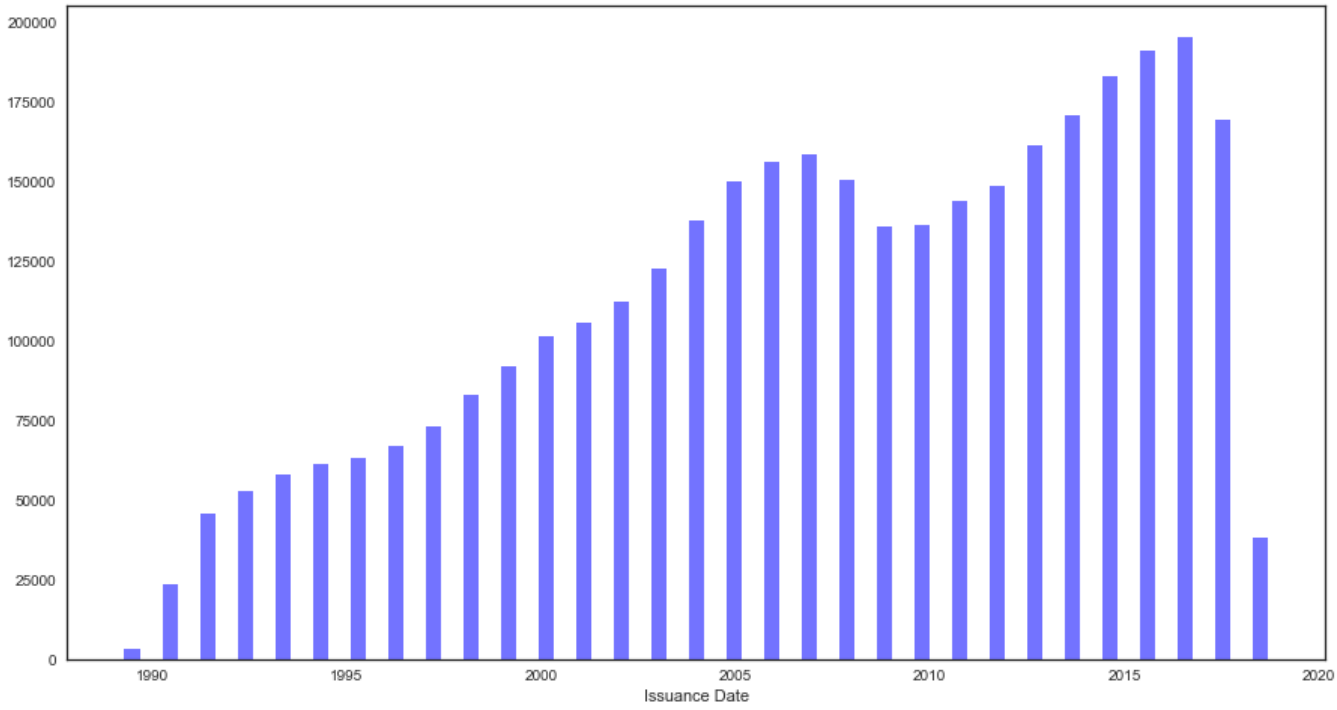
Issuance Date

The number of permit issuances for each year from 1989 to 2019 (until March) were computed and plotted below

In [49]:
```
df1=df['Issuance Date']
df2=df1[df1.notnull()]
df3=pd.DatetimeIndex(df2).year
```

In [179]:
```
plt.figure(figsize=(15,8))
sns.set(style="white")
sns.distplot(df3,bins=31,kde=False,hist_kws={"histtype": "bar", "rwidth": 0.4,"al
pha": 0.55, "color": "b"})
```

Out[179]: <matplotlib.axes._subplots.AxesSubplot at 0x1a47d25400>



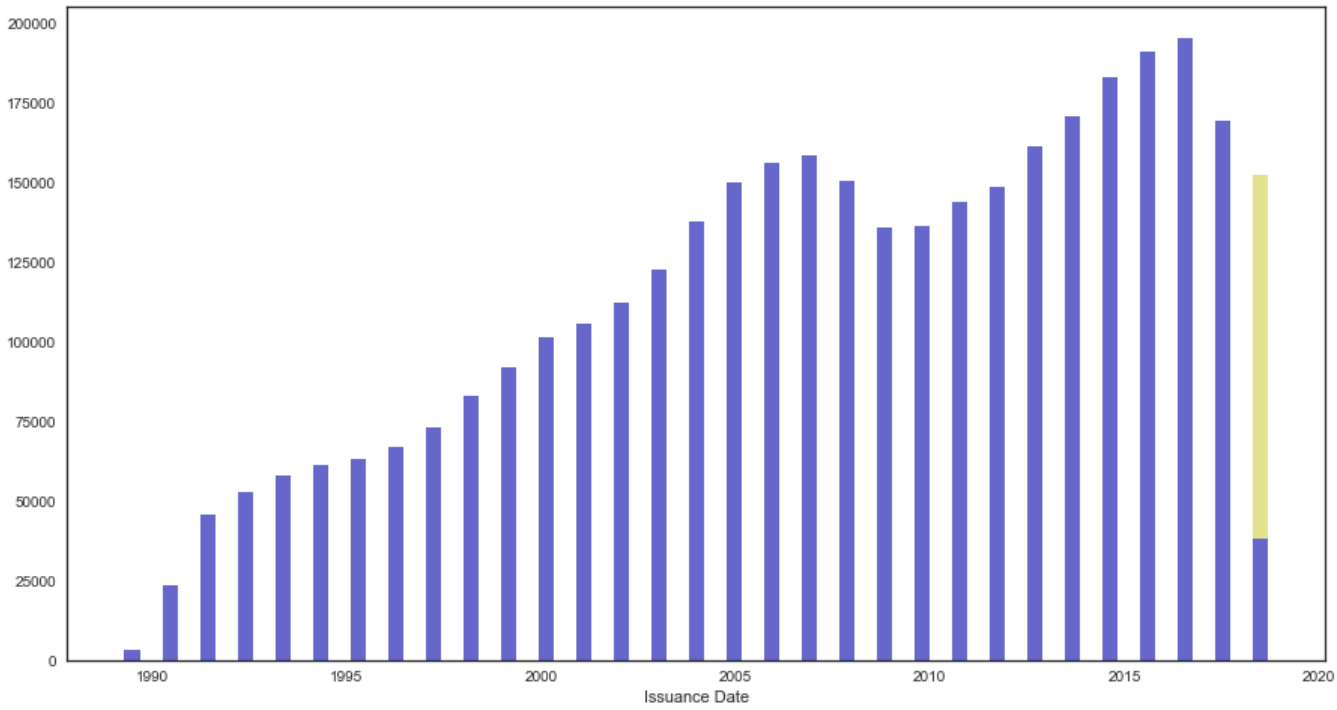
• It is seen that the number of construction permits issued have been increasing steadily over the past 30 years. 2008 and 2009 saw a decrease in the number of construction permits issued and that could be a side effect of The Great Recession - the financial crisis that hit the world economy in 2008.

• The number of issuances for the year 2019 was projected below based on the current rate :-

In [144]:
```
a=38070*3
b=np.arange(1,a)
c=(b*2019)/b
df4=df3.values
df5=np.append(df4,c)
```

In [182]:
```
sns.set(style="white")
fig, ax = plt.subplots(figsize=(15,8))
sns.distplot(df5,bins=31,ax=ax,kde=False,hist_kws={"histtype": "bar", "rwidth":
0.4,"alpha": 0.45, "color": "y"})
sns.distplot(df3,bins=31,ax=ax,kde=False,hist_kws={"histtype": "bar", "rwidth":
0.4,"alpha": 0.55, "color": "b"})
```

Out[182]: <matplotlib.axes._subplots.AxesSubplot at 0x1a4b693828>



• It is seen that the decline seen in 2008 and 2009 seems to be repeating itself, thus drawing the possible conclusion that we may be heading into another era of financial crisis.

• The NYChist dataset was observed which provides the number of people emplyed (annually) in various industries from the year 1990 to the year 2018. The data about the employees from the "Construction of Buildings" category were extracted and the trend was plotted.

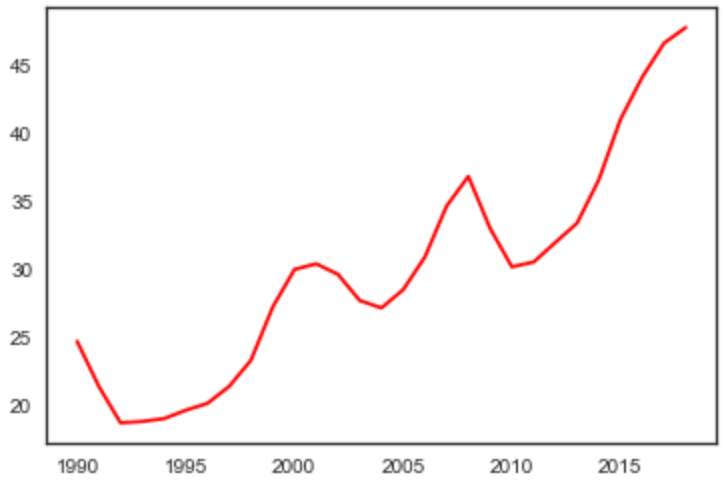
In [286]:
```
df6 = pd.read_excel('NYChist.xls')
# df6
```

In [287]:
```
# df6.info()
```

In [288]:
```
df7=df6[df6['INDUSTRY_TITLE']=='Construction of Buildings']
# df7
```

In [224]:
```
plt.plot(df7['YEAR'],df7['ANNUAL'],color='r')
```

Out[224]:
```
[<matplotlib.lines.Line2D at 0x1a53b4c2b0>]
```



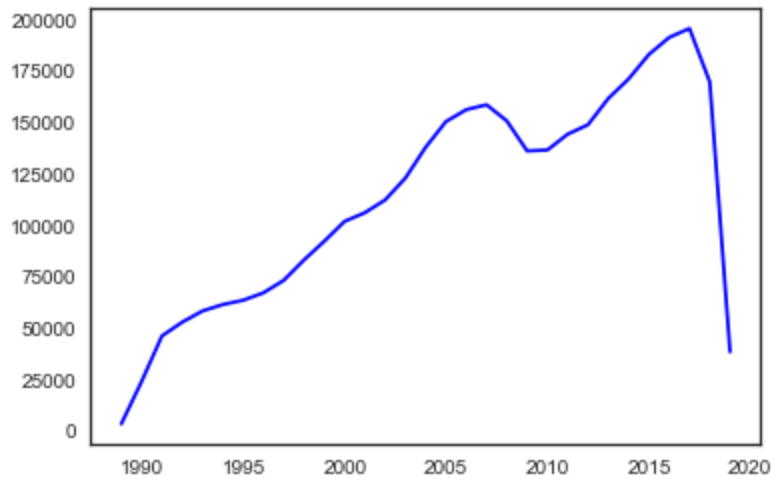
In [237]:
```
values = df3.value_counts().keys().tolist()
counts = df3.value_counts().tolist()
```

In [275]:
```
v=np.array(values)
c=np.array(counts)
df8=np.column_stack((v,c))

df9=sorted(df8,key=lambda x: x[0])
val, count = zip(*df9)
```

In [276]:
```
plt.plot(val,count,color='b')
```

Out[276]:
```
[<matplotlib.lines.Line2D at 0x1a5539f3c8>]
```

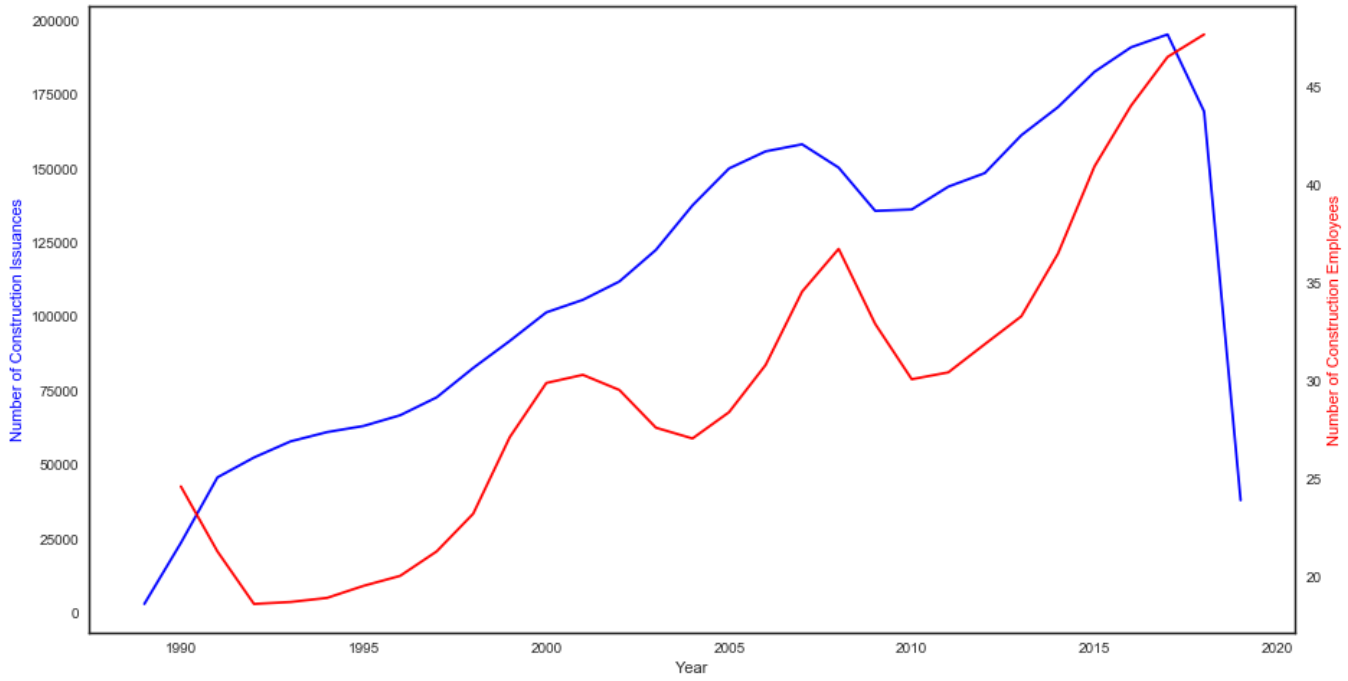


In [283]:
```
fig, ax1 = plt.subplots(figsize=(15,8))

ax2 = ax1.twinx()
ax2.plot(df7['YEAR'],df7['ANNUAL'], 'r-')
ax1.plot(val, count, 'b-')

ax1.set_xlabel('Year')
ax2.set_ylabel('Number of Construction Employees', color='r')
ax1.set_ylabel('Number of Construction Issuances', color='b')

plt.show()
```



• The trend of the number of construction employees is very similar to the trend of the number of permit issuances and again, that makes perfect sense. A key area to notice again is the 2008-2009 timeline where both graphs go through a steep depression. This confirms the direct correlation and impact of the financial crisis on the number of permits being issued and therfore on the number of construction workers being employed.