# Project : Data Analysis On Chicago Crime Data

**Author : Sohel Japanwala**

**Email : sohel.japanwala@gmail.com**

**LinkedIn : linkedin.com/in/soheljapanwala/**

# 1. Introduction

Crime is an international concern, but it is documented and handled in very different ways in different countries. In the United States, violent crimes and property crimes are recorded by the Federal Bureau of Investigation (FBI). Additionally, each city documents crime, and some cities release data regarding crime rates. The city of Chicago, Illinois releases crime data from 2001 onward online.
In this problem, we'll focus on one specific type of property crime, called "motor vehicle theft" (sometimes referred to as grand theft auto).


## 1.1 Problem Statement

- Through analysis, we deduct the crime trend over the years
- We try to analyze the areas more prone to crimes
- The nature of crimes need to be determined
- Details should be gathered regarding arrests

# 2. Getting started

## 2.1 Importing Packages

In [1]:
```
import numpy as np
import pandas as pd
import pandas_profiling

import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

from datetime import datetime, timedelta,time
```

## 2.2 Loading Data

In [2]:
```
crimedata_df=pd.read_csv("https://courses.edx.org/assets/courseware/v1/96f9b8f751467da3a4b8a5be33e32905/asset-v1:MITx+15.071x+2T2020+type@asset+block/mvtWeek1.csv")
crimedata_df.head()
```



```
C:\ProgramData\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3071: DtypeWarning: Columns (6,7) 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[2]:

| | ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Yea |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 8951354 | 12/31/12 23:15 | STREET | False | False | 623 | 6 | 69 | 201 |
| 1 | 8951141 | 12/31/12 22:00 | STREET | False | False | 1213 | 12 | 24 | 201 |
| 2 | 8952745 | 12/31/12 22:00 | RESIDENTIAL YARD (FRONT/BACK) | False | False | 1622 | 16 | 11 | 201 |
| 3 | 8952223 | 12/31/12 22:00 | STREET | False | False | 724 | 7 | 67 | 201 |
| 4 | 8951608 | 12/31/12 21:30 | STREET | False | False | 211 | 2 | 35 | 201 |

## 2.3 Description Of Dataset

- **ID**: a unique identifier for each observation
- **Date**: the date the crime occurred
- **LocationDescription**: the location where the crime occurred
- **Arrest**: whether or not an arrest was made for the crime (TRUE if an arrest was made, and FALSE if an arrest was not made)
- **Domestic**: whether or not the crime was a domestic crime, meaning that it was committed against a family member (TRUE if it was domestic, and FALSE if it was not domestic)
- **Beat**: the area, or "beat" in which the crime occurred. This is the smallest regional division defined by the Chicago police department.
- **District**: the police district in which the crime occured. Each district is composed of many beats, and are defined by the Chicago Police Department.
- **CommunityArea**: the community area in which the crime occurred. Since the 1920s, Chicago has been divided into what are called "community areas", of which there are now 77. The community areas were devised in an attempt to create socially homogeneous regions.
- **Year**: the year in which the crime occurred.
- **Latitude**: the latitude of the location at which the crime occurred.
- **Longitude**: the longitude of the location at which the crime occurred.

### 2.3.1 Shape Of Dataset

```
In [3]: crimedata_df.shape
```

```
Out[3]: (191641, 11)
```



### 2.3.2 Summarising Data

```
In [4]: crimedata_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 191641 entries, 0 to 191640
Data columns (total 11 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   ID                   191641 non-null  int64  
 1   Date                 191641 non-null  object 
 2   LocationDescription  191641 non-null  object 
 3   Arrest               191641 non-null  bool   
 4   Domestic             191641 non-null  bool   
 5   Beat                 191641 non-null  int64  
 6   District             148589 non-null  object 
 7   CommunityArea        167038 non-null  object 
 8   Year                 191641 non-null  int64  
 9   Latitude             189365 non-null  float64
 10  Longitude            189365 non-null  float64
dtypes: bool(2), float64(2), int64(3), object(4)
memory usage: 13.5+ MB
```

In [5]:
```
crimedata_df.describe()
```

Out[5]:

| | ID | Beat | Year | Latitude | Longitude |
|---|---|---|---|---|---|
| **count** | 1.916410e+05 | 191641.000000 | 191641.000000 | 189365.000000 | 189365.000000 |
| **mean** | 4.968629e+06 | 1258.810771 | 2006.134940 | 41.844555 | -87.677782 |
| **std** | 2.342997e+06 | 702.530465 | 3.520285 | 0.085583 | 0.057333 |
| **min** | 1.310022e+06 | 111.000000 | 2001.000000 | 41.644710 | -87.925510 |
| **25%** | 2.832144e+06 | 722.000000 | 2003.000000 | 41.771618 | -87.718474 |
| **50%** | 4.762956e+06 | 1121.000000 | 2006.000000 | 41.850029 | -87.677597 |
| **75%** | 7.201878e+06 | 1733.000000 | 2009.000000 | 41.917116 | -87.635513 |
| **max** | 9.181151e+06 | 2535.000000 | 2012.000000 | 42.022878 | -87.524676 |

## 2.4 Data Profiling

In [6]:
```
crimeprofile=pandas_profiling.ProfileReport(crimedata_df)
crimeprofile.to_file(output_file="c:\\Project\CrimeDataprofile.html")
```



### 2.4.1 Dataset Profile Overview

Overview

Overview | Warnings 12 | Reproduction

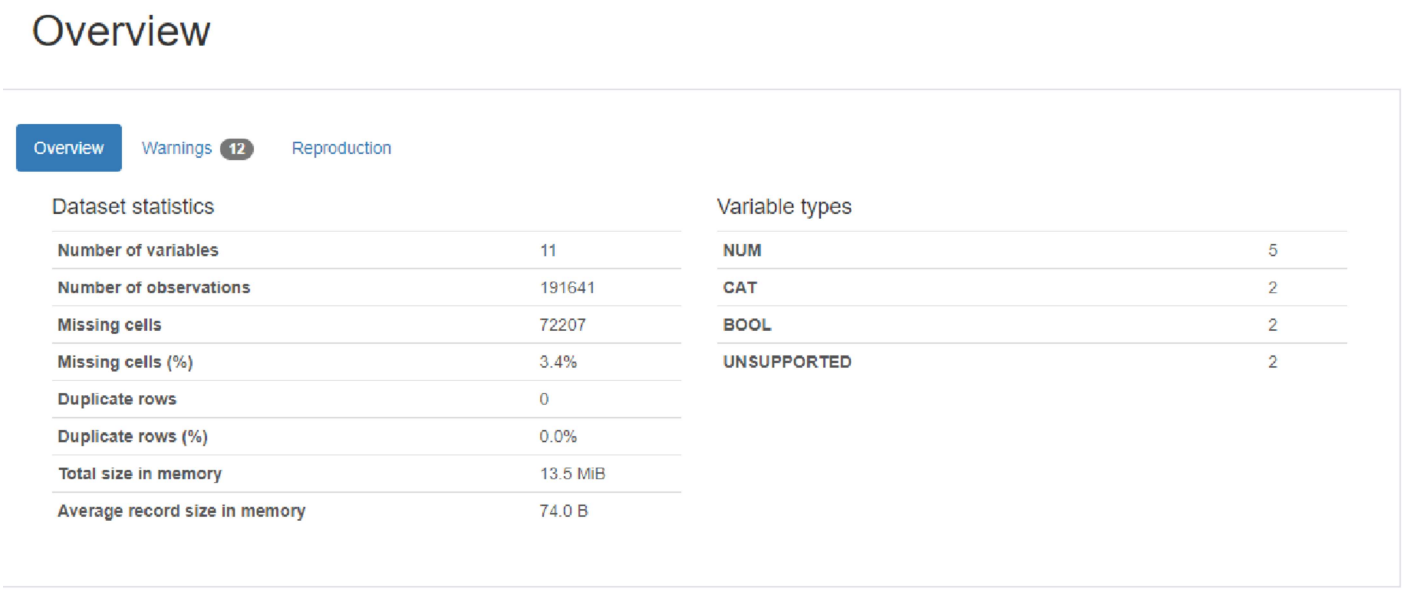
Dataset statistics

| | |
|---|---|
| Number of variables | 11 |
| Number of observations | 191641 |
| Missing cells | 72207 |
| Missing cells (%) | 3.4% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 13.5 MiB |
| Average record size in memory | 74.0 B |

Variable types

| | |
|---|---|
| NUM | 5 |
| CAT | 2 |
| BOOL | 2 |
| UNSUPPORTED | 2 |

### 2.4.2 Dataset Profile Warnings




### 2.4.3 Dataset Observations:



- **District** has 43052 (22.5%) missing values
- **CommunityArea** has 24603 (12.8%) missing values
- **Latitude** has 2276 (1.2%) missing values
- **Longitude** has 2276 (1.2%) missing values
- **District** is an unsupported type, check if it needs cleaning or further analysis
- **CommunityArea** is an unsupported type, check if it needs cleaning or further analysis

## 2.5 Data Preprocessing

### 2.5.1 Checking Missing Values

In [7]:
```
crimedata_df.isnull().sum()
```

Out[7]:
```
ID                        0
Date                      0
LocationDescription       0
Arrest                    0
Domestic                  0
Beat                      0
District              43052
CommunityArea         24603
Year                      0
Latitude               2276
Longitude              2276
dtype: int64
```

In [8]:
```
crimedata_df[crimedata_df.District.isnull()].head()
```

Out[8]:

| | ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Y |
|---|---|---|---|---|---|---|---|---|---|
| **36** | 9124818 | 12/31/12 0:01 | RESIDENCE-GARAGE | False | False | 932 | NaN | 61 | 2 |
| **49** | 8976358 | 12/30/12 18:30 | STREET | False | False | 1935 | NaN | 7 | 2 |
| **57** | 8951754 | 12/30/12 15:00 | SPORTS ARENA/STADIUM | False | False | 1233 | NaN | 31 | 2 |
| **86** | 9042047 | 12/29/12 23:30 | STREET | False | False | 1113 | NaN | 25 | 2 |
| **113** | 9046824 | 12/29/12 7:00 | ST E.. T | False | Fa.se | [illegible] | N..N | [illegible] | 2 |



In [9]:
```
crimedata_df[crimedata_df.CommunityArea.isnull()].head()
```

Out[9]:

| | ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea |
|---|---|---|---|---|---|---|---|---|
| **53161** | 6925027 | 5/16/09 18:00 | STREET | True | False | 1124 | 11 | NaN |
| **87794** | 5149102 | 11/20/06 8:00 | STREET | False | False | 823 | NaN | NaN |
| **128537** | 3360044 | 6/3/04 16:30 | SCHOOL, PUBLIC, GROUNDS | False | False | 1124 | 11 | NaN |
| **165055** | 2097457 | 4/24/02 19:30 | STREET | False | False | 1823 | NaN | NaN |
| **165066** | 2097283 | 4/24/02 14:00 | STREET | False | False | 814 | NaN | NaN |

In [10]:
```
crimedata_df[crimedata_df.Latitude.isnull()].head()
```

Out[10]:

| | ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Y |
|---|---|---|---|---|---|---|---|---|---|
| **36** | 9124818 | 12/31/12 0:01 | RESIDENCE-GARAGE | False | False | 932 | NaN | 61 | 2 |
| **49** | 8976358 | 12/30/12 18:30 | STREET | False | False | 1935 | NaN | 7 | 2 |
| **57** | 8951754 | 12/30/12 15:00 | SPORTS ARENA/STADIUM | False | False | 1233 | NaN | 31 | 2 |
| **86** | 9042047 | 12/29/12 23:30 | STREET | False | False | 1113 | NaN | 25 | 2 |
| **113** | 9046824 | 12/29/12 7:00 | STREET | False | False | 813 | NaN | 64 | 2 |

In [11]:
```
crimedata_df[crimedata_df.Longitude.isnull()].head()
```

Out[11]:

| | ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Y |
|---|---|---|---|---|---|---|---|---|---|
| **36** | 9124818 | 12/31/12 0:01 | RESIDENCE-GARAGE | False | False | 932 | NaN | 61 | 2 |
| **49** | 8976358 | 12/30/12 18:30 | STREET | False | False | 1935 | NaN | 7 | 2 |
| **57** | 8951754 | 12/30/12 15:00 | SPORTS ARENA/STADIUM | False | False | 1233 | NaN | 31 | 2 |
| **86** | 9042047 | 12/29/12 23:30 | STREET | False | False | 1113 | NaN | 25 | 2 |
| **113** | 9046824 | 12/29/12 7:00 | STREET | False | False | 813 | NaN | 64 | 2 |



## 2.5.2 Deleting MIssing Value Rows Of Latitude and Longitude (since it accounts for minor percentage of data)

In [12]:
```
crimedata_df.dropna(subset=['Latitude'],inplace=True)
crimedata_df.dropna(subset=['Longitude'],inplace=True)
```

## 2.5.3 Verification if data is deleted

In [13]:
```
crimedata_df.Latitude.isnull().sum()
```

Out[13]: 0

In [14]:
```
crimedata_df.Longitude.isnull().sum()
```

Out[14]: 0

### 2.5.3 Replacing Missing Values

In [15]:
```
crimedata_df.loc[crimedata_df.District=='100']
```

Out[15]:

| ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Year | Latitude |
|---|---|---|---|---|---|---|---|---|---|

In [16]:
```
crimedata_df.loc[crimedata_df.CommunityArea=='100']
```

Out[16]:

| ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Year | Latitude |
|---|---|---|---|---|---|---|---|---|---|

In [17]:
```
crimedata_df['District'].fillna("100",inplace=True)
crimedata_df['CommunityArea'].fillna("100",inplace=True)
```

### 2.5.4 Verifying if data was replaced



In [18]:
```
crimedata_df.District.isnull().sum()
```

Out[18]: 0

In [19]:
```
crimedata_df.CommunityArea.isnull().sum()
```

Out[19]: 0

### 2.5.5 Changing Data type

In [20]:
```
crimedata_df['Date']=pd.to_datetime(crimedata_df['Date'])
```

In [21]:
```
crimedata_df.LocationDescription=crimedata_df.LocationDescription.astype(str)
```

In [22]:
```
crimedata_df.District=crimedata_df.District.astype(str)
crimedata_df['District']=crimedata_df['District'].replace([' '],'0')
crimedata_df.District=crimedata_df.District.astype('float64')
#crimedata_df.District=crimedata_df.District.astype('int64')
```

In [23]:
```
crimedata_df.CommunityArea=crimedata_df.CommunityArea.astype(str)
crimedata_df['CommunityArea']=crimedata_df['CommunityArea'].replace([' '],'0')
crimedata_df.CommunityArea=crimedata_df.CommunityArea.astype('float64')
#crimedata_df.CommunityArea=crimedata_df.CommunityArea.astype('int64')
```

In [24]:
```
crimedata_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 189365 entries, 0 to 191640
Data columns (total 11 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   ID                   189365 non-null  int64
 1   Date                 189365 non-null  datetime64[ns]
 2   LocationDescription  189365 non-null  object
 3   Arrest               189365 non-null  bool
 4   Domestic             189365 non-null  bool
 5   Beat                 189365 non-null  int64
 6   District             189365 non-null  float64
 7   CommunityArea        189365 non-null  float64
 8   Year                 189365 non-null  int64
 9   Latitude             189365 non-null  float64
 10  Longitude            189365 non-null  float64
dtypes: bool(2), datetime64[ns](1), float64(4), int64(3), object(1)
memory usage: 14.8+ MB
```

## 2.6 Profiling Post Pre-Processing

In [25]:
```
crimeDataPostProfile=pandas_profiling.ProfileReport(crimedata_df)
crimeDataPostProfile.to_file(output_file="C:\\Project\CrimeDataPostProfile.html")
```



### 2.6.1 Dataset Post Pre-Processing Overview

Overview

Overview | Warnings 8 | Reproduction

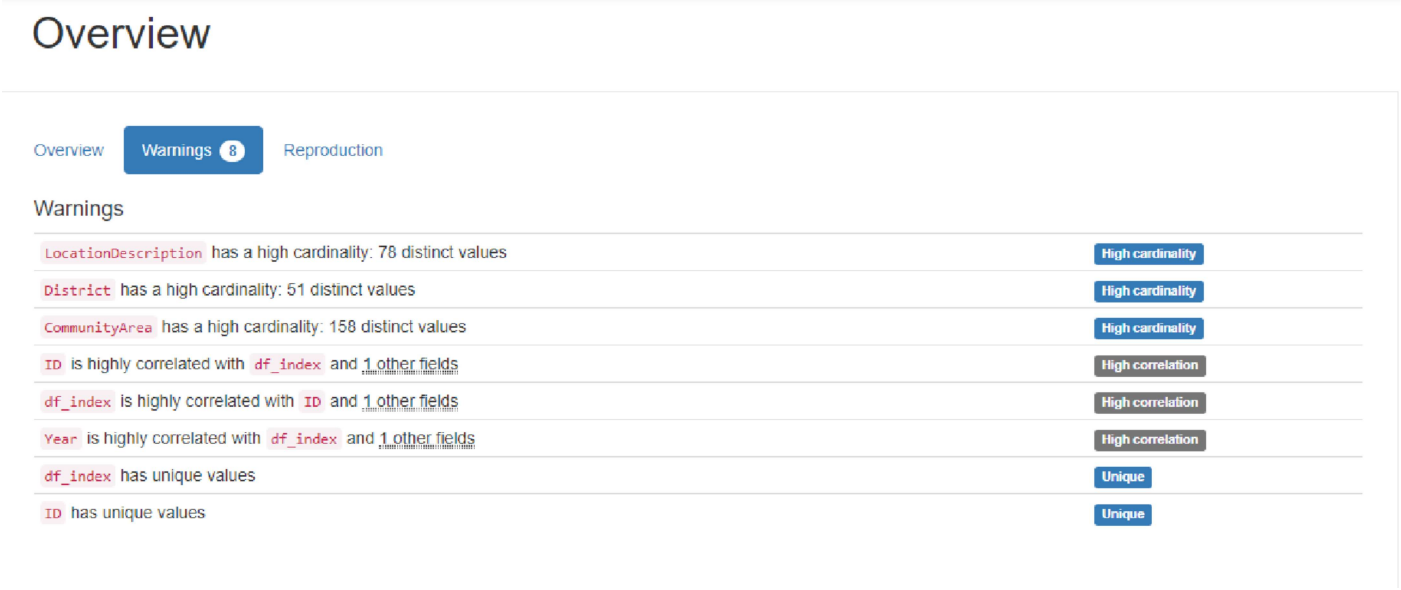
Dataset statistics

| | |
|---|---|
| Number of variables | 12 |
| Number of observations | 189365 |
| Missing cells | 0 |
| Missing cells (%) | 0.0% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 14.8 MiB |
| Average record size in memory | 82.0 B |

Variable types

| | |
|---|---|
| NUM | 6 |
| CAT | 3 |
| BOOL | 2 |
| DATE | 1 |

### 2.6.2 Dataset Post Pre-Processing Warning




# 3. EDA

### 3.1 Analyzing Columns For Unique Data



In [26]:
```
print(f"There are {crimedata_df.ID.nunique()} Unique Crime ID in the dataset")
```

```
There are 189365 Unique Crime ID in the dataset
```

In [27]:
```
print(f"There are {crimedata_df.Date.nunique()} unique Datestamps in the datas
et")
```

```
There are 130596 unique Datestamps in the dataset
```

In [28]:
```
print(f"There are {crimedata_df.LocationDescription.nunique()} Unique Location
Areas in the dataset")
```

```
There are 78 Unique Location Areas in the dataset
```

In [29]:
```
print(f"There are {crimedata_df.Beat.nunique()} Unique Beats in the dataset")
```

```
There are 300 Unique Beats in the dataset
```

In [30]:
```
print(f"There are {crimedata_df.CommunityArea.nunique()} Unique Community Area
s in the dataset")
```

```
There are 79 Unique Community Areas in the dataset
```

In [31]:
```
print(f"There are {crimedata_df.Year.nunique()} Unique Years in the dataset")
```

```
There are 12 Unique Years in the dataset
```

In [32]:
```
print(f"There are {crimedata_df.District.nunique()} Unique Districts in the da
taset")
```

```
There are 28 Unique Districts in the dataset
```

## 3.2 Year Wise Crime Data

In [33]:
```
groupbyYear_df=pd.DataFrame(crimedata_df.groupby('Year')['ID'].nunique())
groupbyYear_df.sort_values(['ID'],inplace=True)
groupbyYear_df.head(12)
```

Out[33]:

| Year | ID |
|---|---|
| **2009** | 12042 |
| **2012** | 13998 |
| **2007** | 14271 |
| **2008** | 14281 |
| **2010** | 15485 |
| **2011** | 15622 |
| **2006** | 16069 |
| **2005** | 16403 |
| **2003** | 16615 |
| **2004** | 16826 |
| **2002** | 17274 |
| **2001** | 20479 |

In [34]:
```
groupbyYear_df.describe()
```

Out[34]:

| | ID |
|---|---|
| **count** | 12.000000 |
| **mean** | 15780.416667 |
| **std** | 2096.747805 |
| **min** | 12042.000000 |
| **25%** | 14278.500000 |
| **50%** | 15845.500000 |
| **75%** | 16667.750000 |
| **max** | 20479.000000 |

## Observations:

- The crime rate shows a decreasing trend over the years
- On an average, 15,867 crimes are committed each year

## 3.3 District Wise Crime Data



In [35]:
```
groupbyDistrict_df=pd.DataFrame(crimedata_df.groupby('District')['ID'].nunique
())
groupbyDistrict_df.sort_values(['ID'], inplace=True)
groupbyDistrict_df.head(10)
```

Out[35]:

| | ID |
|---|---|
| **District** | |
| **31.0** | 1 |
| **0.0** | 4 |
| **23.0** | 5 |
| **21.0** | 29 |
| **20.0** | 2494 |
| **1.0** | 2598 |
| **18.0** | 3425 |
| **24.0** | 3930 |
| **13.0** | 4013 |
| **12.0** | 4090 |

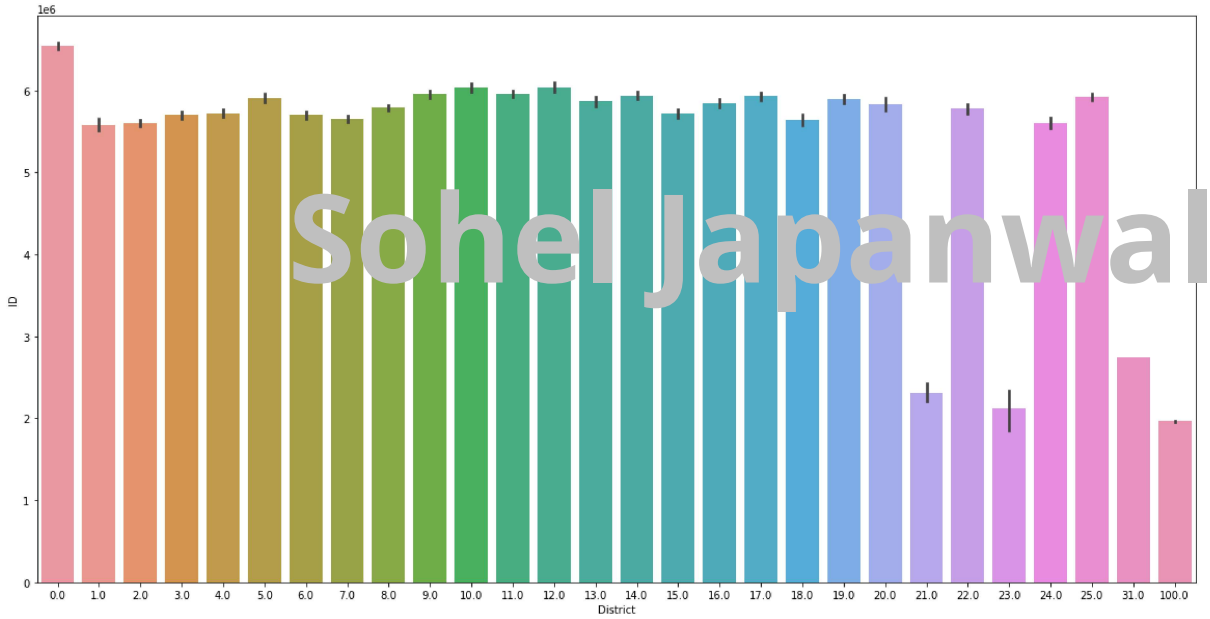
In [36]:
```
groupbyDistrict_df.describe()
```

Out[36]:

| | ID |
|---|---|
| **count** | 28.000000 |
| **mean** | 6763.035714 |
| **std** | 7505.972666 |
| **min** | 1.000000 |
| **25%** | 3803.750000 |
| **50%** | 5682.000000 |
| **75%** | 7828.250000 |
| **max** | 40782.000000 |

In [37]:
```
plt.figure(figsize=(20,10))
sns.barplot(y=crimedata_df["ID"], x=crimedata_df["District"])
```

Out[37]: <matplotlib.axes._subplots.AxesSubplot at 0x1d199f4a790>





## Observations:

- On an average, 5,568 crimes were committed per district

## 3.4 Location Wise Crime Data

In [38]:
```
groupbyLocationDescription_df=pd.DataFrame(crimedata_df.groupby('LocationDescr
iption')['ID'].nunique())
groupbyLocationDescription_df.sort_values(['ID'],inplace=True)
groupbyLocationDescription_df.head(78)
```

Out[38]:

| LocationDescription | ID |
|---|---|
| JAIL / LOCK-UP FACILITY | 1 |
| CTA TRAIN | 1 |
| AIRPORT BUILDING NON-TERMINAL - SECURE AREA | 1 |
| AIRPORT EXTERIOR - SECURE AREA | 1 |
| NEWSSTAND | 1 |
| ... | ... |
| GAS STATION | 2084 |
| ALLEY | 2285 |
| OTHER | 4495 |
| PARKING LOT/GARAGE(NON.RESID.) | 14620 |
| STREET | 154761 |

78 rows × 1 columns



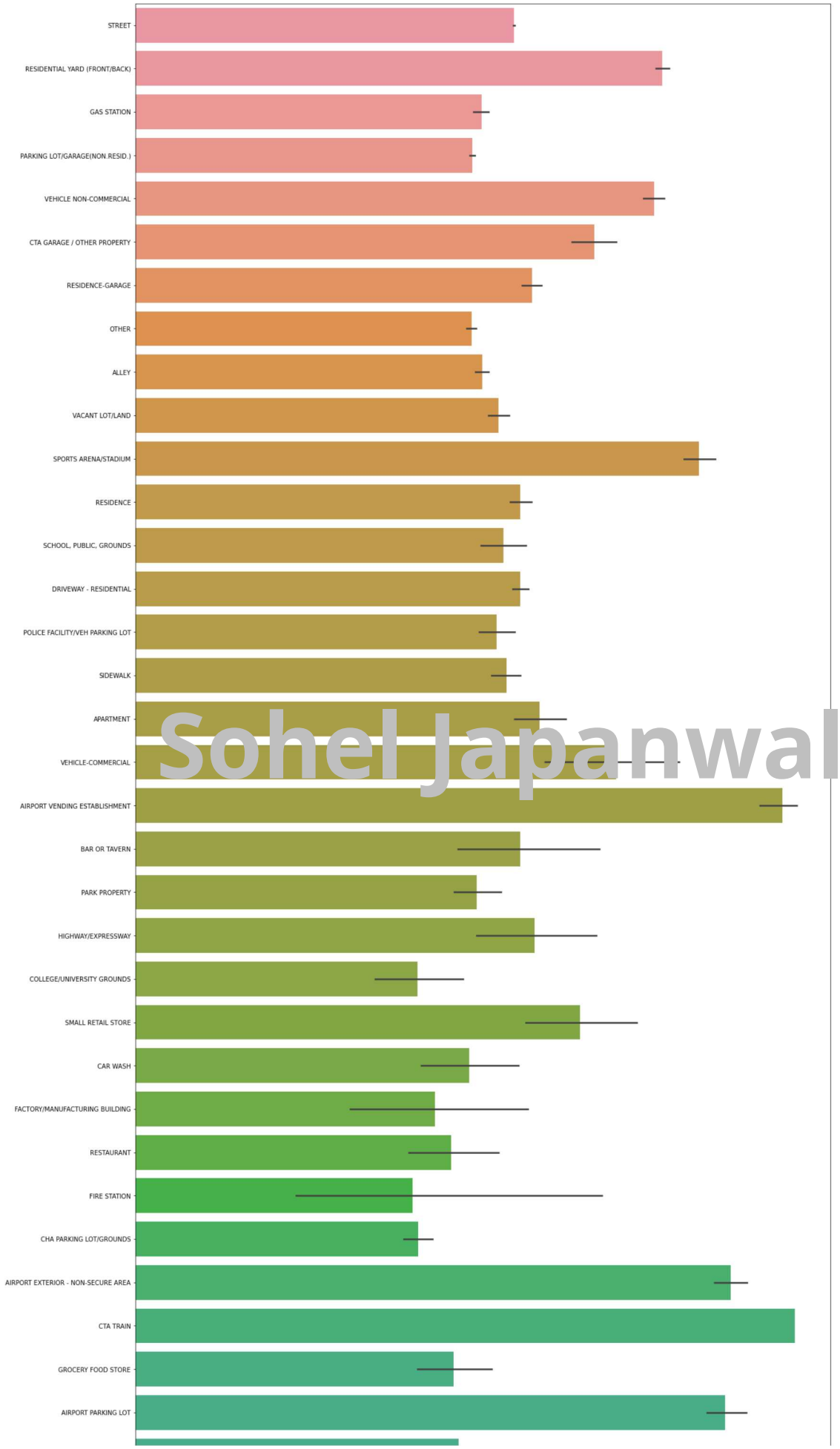
In [39]:
```
groupbyLocationDescription_df.describe()
```

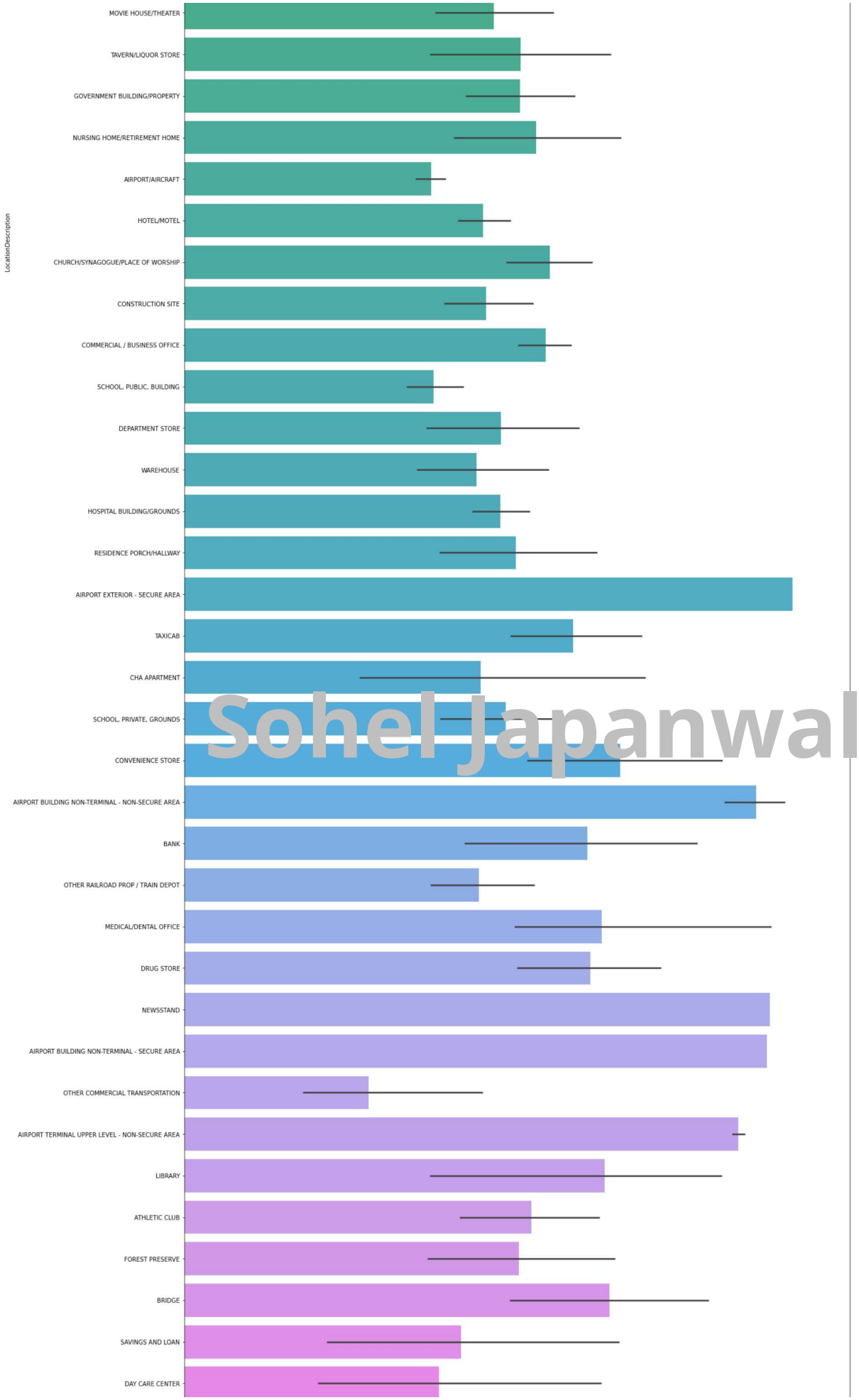
Out[39]:

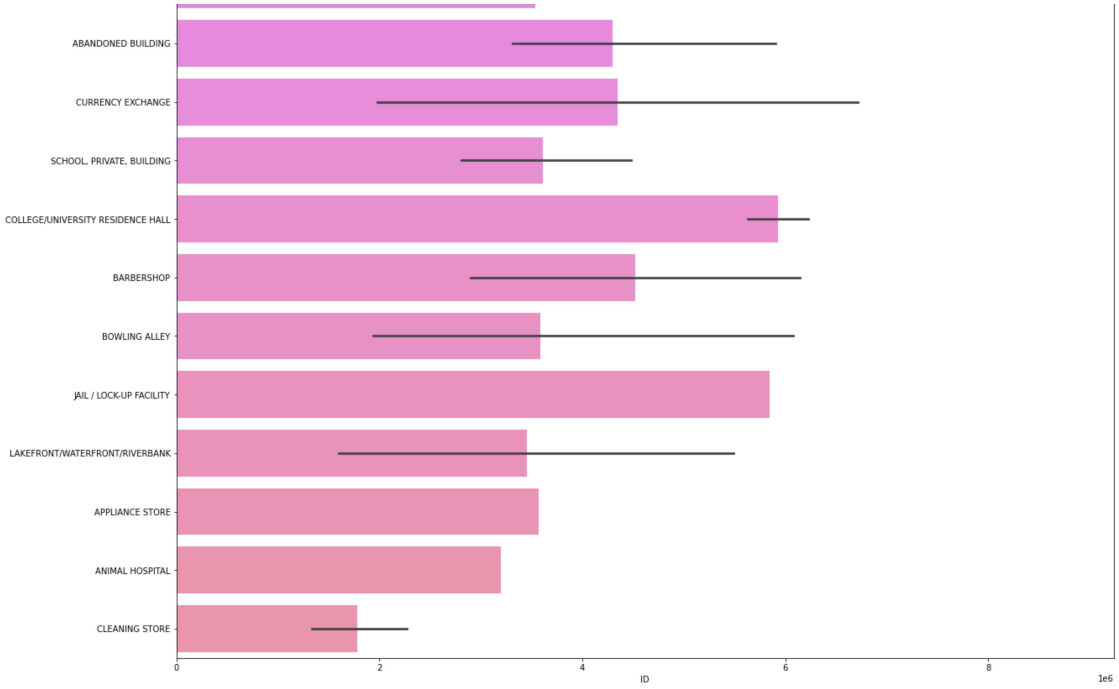
| | ID |
|---|---|
| count | 78.000000 |
| mean | 2427.756410 |
| std | 17560.767781 |
| min | 1.000000 |
| 25% | 4.250000 |
| 50% | 20.500000 |
| 75% | 140.500000 |
| max | 154761.000000 |

In [40]:
```
plt.figure(figsize=(20,100))
sns.barplot(x=crimedata_df["ID"], y=crimedata_df["LocationDescription"])
```

Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x1d199f4a310>








## Observations:

- The crime rate in open spaces is more than in closed and secured areas
- On an average, 2,456 crimes are committed per location



## 3.5 Community Area Wise Crime Data

In [41]:
```
groupbyCommunityArea_df=pd.DataFrame(crimedata_df.groupby('CommunityArea')['ID'].nunique())
groupbyCommunityArea_df.sort_values(['ID'],inplace=True)
groupbyCommunityArea_df.head(10)
```

Out[41]:

| CommunityArea | ID |
|---|---|
| **0.0** | 15 |
| **9.0** | 77 |
| **74.0** | 117 |
| **47.0** | 187 |
| **12.0** | 231 |
| **55.0** | 255 |
| **36.0** | 380 |
| **54.0** | 382 |
| **10.0** | 473 |
| **72.0** | 528 |

In [42]:
```
groupbyCommunityArea_df.describe()
```

Out[42]:

| | ID |
|---|---|
| **count** | 79.000000 |
| **mean** | 2397.025316 |
| **std** | 3015.695412 |
| **min** | 15.000000 |
| **25%** | 806.000000 |
| **50%** | 1695.000000 |
| **75%** | 2855.000000 |
| **max** | 24370.000000 |



In [43]:
```
plt.figure(figsize=(20,10))
```

Out[43]:
```
<Figure size 1440x720 with 0 Axes>
```

```
<Figure size 1440x720 with 0 Axes>
```

## Observations:

- On an average, 2,141 crimes are committed per community area

## 3.6 Beat Wise Crime Data

In [44]:
```
groupbyBeat_df=pd.DataFrame(crimedata_df.groupby('Beat')['ID'].nunique())
groupbyBeat_df.sort_values(['ID'],inplace=True)
groupbyBeat_df.head(10)
```

Out[44]:

| Beat | ID |
|---|---|
| **130** | 1 |
| **114** | 4 |
| **1221** | 5 |
| **1214** | 7 |
| **1225** | 8 |
| **1215** | 9 |
| **1234** | 9 |
| **121** | 12 |
| **1235** | 20 |
| **215** | 21 |

In [45]:
```
groupbyBeat_df.describe()
```



Out[45]:

| | ID |
|---|---|
| **count** | 300.000000 |
| **mean** | 631.216667 |
| **std** | 344.272103 |
| **min** | 1.000000 |
| **25%** | 390.000000 |
| **50%** | 617.000000 |
| **75%** | 825.250000 |
| **max** | 1605.000000 |

## Observations:

- On an average, 623 crimes are committed per beat

## 3.7 Analysis Of Domestic Crime

In [46]:
```
groupbyDomestic_df=pd.DataFrame(crimedata_df.groupby('Domestic')['ID'].nunique
())
groupbyDomestic_df.sort_values(['ID'],inplace=True)
groupbyDomestic_df.head(10)
```

Out[46]:

| Domestic | ID |
|---|---|
| True | 411 |
| False | 188954 |

In [47]:
```
groupbyDomestic_df.describe()
```

Out[47]:

| | ID |
|---|---|
| count | 2.000000 |
| mean | 94682.500000 |
| std | 133320.033845 |
| min | 411.000000 |
| 25% | 47546.750000 |
| 50% | 94682.500000 |
| 75% | 141818.250000 |
| max | 188954.000000 |



## Observations:

- Significantly more crimes are committed against outsiders than against family members

## 3.8 Analysis Of Domestic Crime Over The Years

In [48]:
```
groupbyYearDom_df=pd.DataFrame(crimedata_df.groupby('Year')['Domestic'].sum())
groupbyYearDom_df.sort_values(['Domestic'],inplace=True)
groupbyYearDom_df.head(12)
```

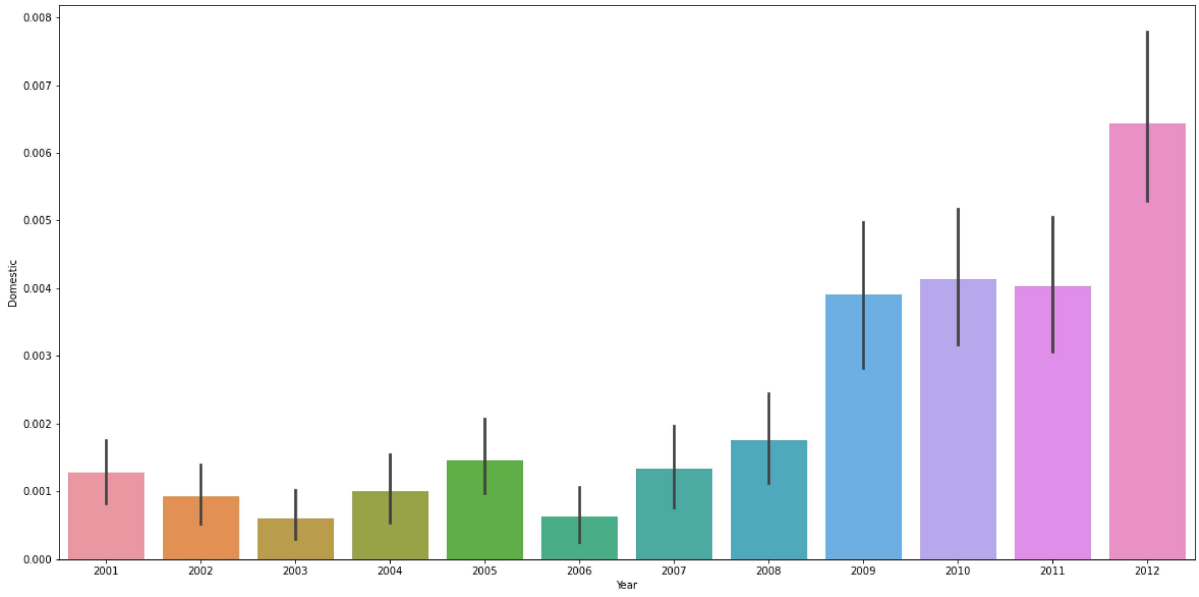
Out[48]:

| Year | Domestic |
|---|---|
| **2003** | 10.0 |
| **2006** | 10.0 |
| **2002** | 16.0 |
| **2004** | 17.0 |
| **2007** | 19.0 |
| **2005** | 24.0 |
| **2008** | 25.0 |
| **2001** | 26.0 |
| **2009** | 47.0 |
| **2011** | 63.0 |
| **2010** | 64.0 |
| **2012** | 90.0 |

In [49]:
```
plt.figure(figsize=(20,10))
sns.barplot(y=crimedata_df["Domestic"], x=crimedata_df["Year"])
```



Out[49]: <matplotlib.axes._subplots.AxesSubplot at 0x1d1a218e520>




## Observations:

- Domestic crimes have shown an increasing trend over the years

## 3.9 Analysis Of Arrest

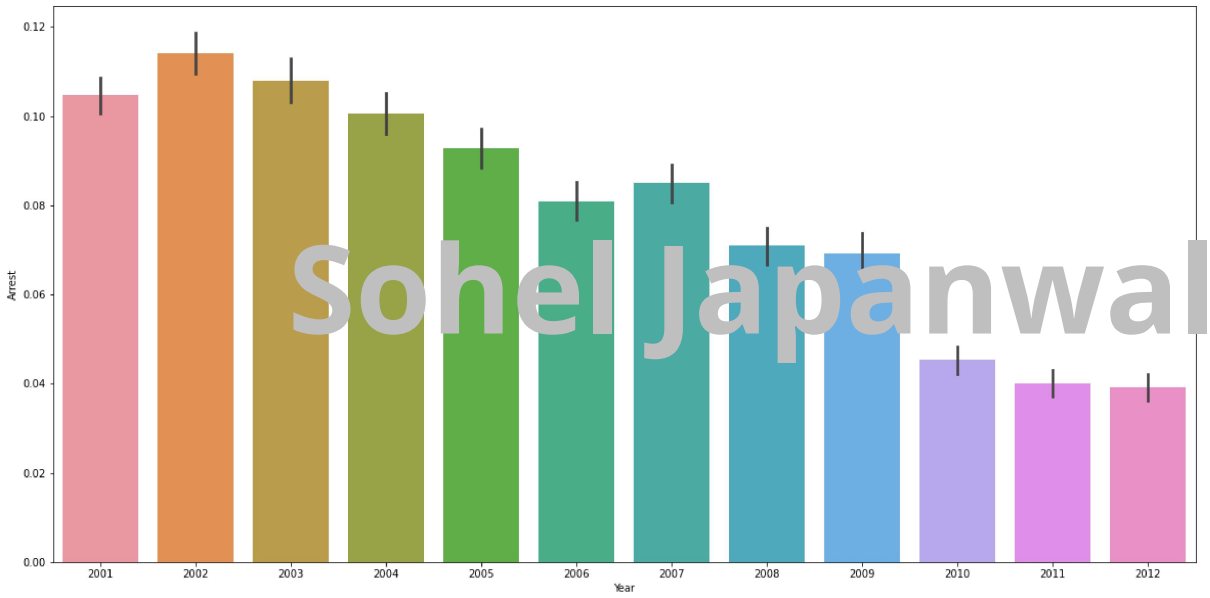
In [50]:
```
groupbyArrest_df=pd.DataFrame(crimedata_df.groupby('Arrest')['ID'].nunique())
groupbyArrest_df.sort_values(['ID'],inplace=True)
groupbyArrest_df.head(10)
```

Out[50]:

| Arrest | ID |
|---|---|
| True | 15354 |
| False | 174011 |

In [51]:
```
plt.figure(figsize=(20,10))
sns.barplot(y=crimedata_df["Arrest"],x=crimedata_df["Year"])
```

Out[51]: <matplotlib.axes._subplots.AxesSubplot at 0x1d199e9c730>






### Observations:

- Large numbers of crimes are those in which an arrest was not made

## 3.10 Day Wise Analysis

In [52]:
```
crimedata_df["Day"]=crimedata_df["Date"].dt.day_name()
crimedata_df.head()
```

Out[52]:

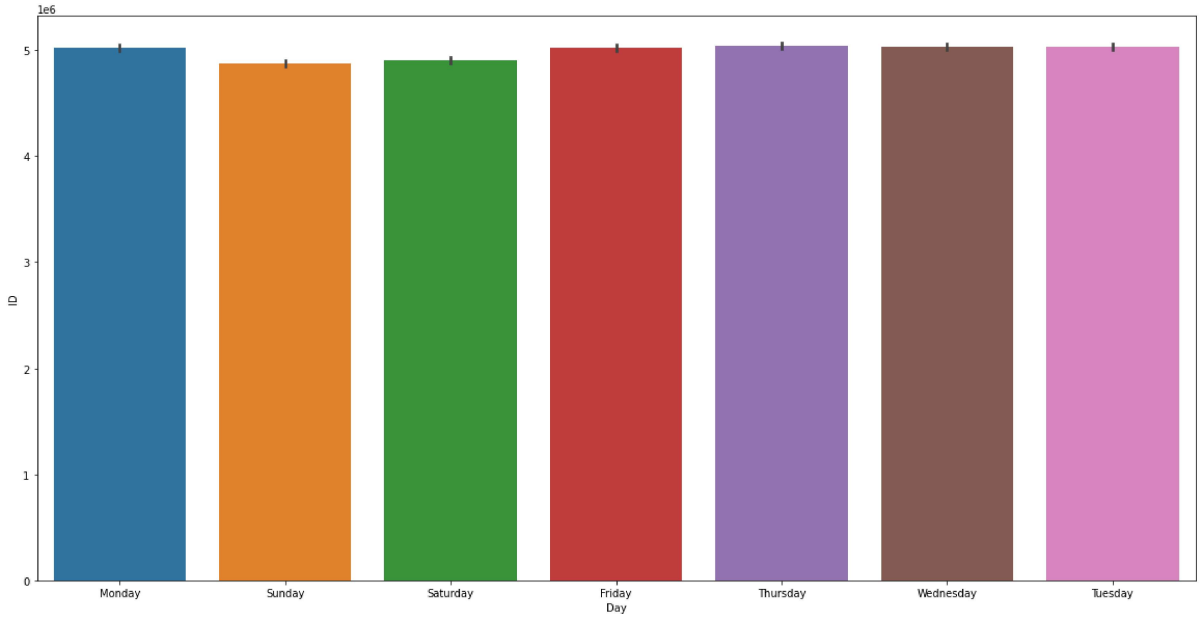
| | ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Yea |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 8951354 | 2012-12-31 23:15:00 | STREET | False | False | 623 | 6.0 | 69.0 | 201 |
| **1** | 8951141 | 2012-12-31 22:00:00 | STREET | False | False | 1213 | 12.0 | 24.0 | 201 |
| **2** | 8952745 | 2012-12-31 22:00:00 | RESIDENTIAL YARD (FRONT/BACK) | False | False | 1622 | 16.0 | 11.0 | 201 |
| **3** | 8952223 | 2012-12-31 22:00:00 | STREET | False | False | 724 | 7.0 | 67.0 | 201 |
| **4** | 8951608 | 2012-12-31 21:30:00 | STREET | False | False | 211 | 2.0 | 35.0 | 201 |

In [53]:
```
crimedata_df.groupby("Day")["ID"].nunique().sort_values()
```

Out[53]:
```
Day
Sunday       26025
Tuesday      26491
Saturday     26772
Thursday     26995
Monday       27065
Wednesday    27089
Friday       28928
Name: ID, dtype: int64
```

In [54]:
```
plt.figure(figsize=(20,10))
sns.barplot(y=crimedata_df["ID"],x=crimedata_df["Day"])
```

Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x1d19a70aa60>




### Observations:

- The day of the week does not affect the number of crimes committed



## 3.11 Time Wise Analysis

In [55]:
```
crimedata_df["Time"]= crimedata_df["Date"].dt.time
crimedata_df.head()
```

Out[55]:

| | ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Yea |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 8951354 | 2012-12-31 23:15:00 | STREET | False | False | 623 | 6.0 | 69.0 | 201 |
| 1 | 8951141 | 2012-12-31 22:00:00 | STREET | False | False | 1213 | 12.0 | 24.0 | 201 |
| 2 | 8952745 | 2012-12-31 22:00:00 | RESIDENTIAL YARD (FRONT/BACK) | False | False | 1622 | 16.0 | 11.0 | 201 |
| 3 | 8952223 | 2012-12-31 22:00:00 | STREET | False | False | 724 | 7.0 | 67.0 | 201 |
| 4 | 8951608 | 2012-12-31 21:30:00 | STREET | False | False | 211 | 2.0 | 35.0 | 201 |

In [64]:
```
crimedata_df["Time_category"]= crimedata_df["Time"].apply(lambda x:"DayTime" if(x>time(6) and x<time(18)) else "NightTime")
```

Out[64]:
```
'\nfor value in crimedata_df["Time"]:\n    \n    if (value>=time(6) and value
<=time(12)):\n        print(f"greater than 6 less than 12",value)\n        c
rimedata_df["Time_category"]=crimedata_df["Time_category"].apply(lambda x:"6
to 12")\n        \n    elif(value>time(12) and value<time(18)):\n        prin
t(f"greater than 12 less than 18",value)\n                                \n    elif
(value>time(18) and value<time(0)):\n        print(f"greater than 18 less tha
n 00",value)\n                            \n    elif(value>time(0) and value<time
(6)):\n        print(f"greater than 00 less than 06",value)\n'
```



In [66]:
```
crimedata_df.groupby("Time_category")["ID"].nunique().sort_values()
```

Out[66]:
```
Time_category
DayTime       80358
NightTime    109007
Name: ID, dtype: int64
```

In [65]:
```
crimedata_df.head()
```

Out[65]:

| | ID | Date | LocationDescription | Arrest | Domestic | Beat | District | CommunityArea | Yea |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 8951354 | 2012-12-31 23:15:00 | STREET | False | False | 623 | 6.0 | 69.0 | 201 |
| **1** | 8951141 | 2012-12-31 22:00:00 | STREET | False | False | 1213 | 12.0 | 24.0 | 201 |
| **2** | 8952745 | 2012-12-31 22:00:00 | RESIDENTIAL YARD (FRONT/BACK) | False | False | 1622 | 16.0 | 11.0 | 201 |
| **3** | 8952223 | 2012-12-31 22:00:00 | STREET | False | False | 724 | 7.0 | 67.0 | 201 |
| **4** | 8951608 | 2012-12-31 21:30:00 | STREET | False | False | 211 | 2.0 | 35.0 | 201 |

### Observations:

- More crimes occur during evening and night time



## 4. Conclusions:

- The crime rate shows a decreasing trend over the years
- On an average, 15,867 crimes are committed each year
- On an average, 5,568 crimes were committed per district
- The crime rate in open spaces is more than in closed and secured areas
- On an average, 2,456 crimes are committed per location
- On an average, 2,141 crimes are committed per community area
- On an average, 623 crimes are committed per beat
- Significantly more crimes are committed against outsiders than against family members
- Arrest of crimes have decreased over the years
- Domestic crimes show an increasing trend
- More crimes tend to happen during evening/ nightime

## 5. Actionable Insights:

- The city security authorities are doing a good job of keeping the society safe. They would be required to do so in the future
- The secuity forces should focus their efforts more on open and unsecured spaces
- The security forces should focus more on the Beats, Community Area and Districts with the highest crimes
- More social awareness campaigns should be undertaken that promote brotherhood and good moral behaviour even amongst strangers
- The authorities have to look deeper as to why crime related arrests have been decreasing

# Project : Data Analysis On Chicago Crime Data

**Author : Sohel Japanwala**

**Email : sohel.japanwala@gmail.com**

**LinkedIn : linkedin.com/in/soheljapanwala/**