In [ ]:
```
##Project 2
##Continue project 1 with updating dataset to code in Python and develop new f
##Will Wang
##02/23/24
```

In [34]:
```
## choose dataset from https://opendataphilly.org/datasets/ the Affordable_Hou
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("Affordable_Housing.csv")
df
df.head()
```

Out[34]:

	OBJECTID	FISCAL_YEAR_COMPLETE	PROJECT_NAME	DEVELOPER_NAME	ADDRESS	PRO
0	1	2019.0	Gloria Casarez Residence (1315 N 8th)	Project HOME	1315 N 8TH ST	Sp
1	2	2019.0	Roberto Clemente Homes	Nueva Esperanza	3921-61 N 5TH ST	
2	3	2019.0	Henry Avenue Senior Campus I	NewCourtland Elder Services	3232 HENRY AVE	
3	4	2019.0	Villas Del Caribe	HACE	161-71 W ALLEGHENY AVE	
4	5	2019.0	Cantrell Place	Presby's Inspired Life	447 CANTRELL ST	

Part 1: complete the remaining project 1 in Python

In [26]:
```
## Fisrt to show the range of the years when the data of the affordable housing
arrange_year = df.sort_values(by='FISCAL_YEAR_COMPLETE') ##use sort_values to s
range_year = (arrange_year['FISCAL_YEAR_COMPLETE'].dropna().min(), arrange_yea
print("Range of years:", range_year)
```

```
Range of years: (1995.0, 2019.0)
```

1.1 : Which project has the most units? the most rental units top 5? the most special needs units top 5?

In [29]:
```
##first clean the data column of TOTAL_UNITS
project02_dataset_units = df.dropna(subset=['TOTAL_UNITS'])


most_units = project02_dataset_units['TOTAL_UNITS'].max()
print("Most units:", most_units)
most_unit_proj_name = project02_dataset_units.loc[project02_dataset_units['TOTA
print("Project(s) with the most units:", most_unit_proj_name)
```

```
top_5_rental = project02_dataset_units[project02_dataset_units['PROJECT_TYPE']
top_5_rental_name = top_5_rental['PROJECT_NAME'].tolist()
print("Top 5 rental projects:", top_5_rental_name)


top_5_sn = project02_dataset_units[project02_dataset_units['PROJECT_TYPE'] == '
top_5_sn_name = top_5_sn['PROJECT_NAME'].tolist()
print("Top 5 special needs projects:", top_5_sn_name)
```

```
Most units: 470.0
Project(s) with the most units: 293    Southwark Plaza
Name: PROJECT_NAME, dtype: object
Top 5 rental projects: ['Southwark Plaza', 'Casa Farnese 202', 'Casa Farnese 2
02', 'Walnut Park Plaza', 'Carl Mackley']
Top 5 special needs projects: ['Germantown YMCA', 'Preservation Projects', 'St
ation House', 'Ruth Williams House', 'St. John the Evangelist House']
```

1.2 :Which project has the most accessible units? the most accessible units top 5?

In [30]:

```
##first clean the data column of accessible_UNITS
project02_dataset_accunits = df.dropna(subset=['ACCESSIBLE_UNITS'])


most_acc_units = project02_dataset_accunits['ACCESSIBLE_UNITS'].max()
print("Most accessible units:", most_acc_units)
most_acc_proj_name = project02_dataset_accunits.loc[project02_dataset_accunits
print("Project(s) with the most accessible units:", most_acc_proj_name)


top_5_acc = project02_dataset_accunits.nlargest(5, 'ACCESSIBLE_UNITS')
top_5_acc_name = top_5_acc['PROJECT_NAME'].tolist()
print("Top 5 projects with the most accessible units:", top_5_acc_name)
```

```
Most accessible units: 75.0
Project(s) with the most accessible units: 5    Ann Thomas Presbyterian
Name: PROJECT_NAME, dtype: object
Top 5 projects with the most accessible units: ['Ann Thomas Presbyterian', 'Vi
llas Del Caribe', 'Ruth Williams House', 'Centennial Village', 'Cantrell Plac
e']
```

1.3 :The most accessible proportion units top 5? the least accessible proportion top5

In [31]:

```
##first clean the data column of accessible_UNITS
project02_dataset_proportion = df.dropna(subset=['TOTAL_UNITS', 'ACCESSIBLE_UNI

project02_dataset_proportion['ACCESSIBLE_PROPORTION'] = project02_dataset_prop

top_5_prop = project02_dataset_proportion.sort_values(by='ACCESSIBLE_PROPORTIO
top_5_prop_name = top_5_prop['PROJECT_NAME'].tolist()
print("Top 5 projects with the highest accessible proportion:", top_5_prop_nam

tail_5_prop = project02_dataset_proportion.sort_values(by='ACCESSIBLE_PROPORTI
tail_5_prop_name = tail_5_prop['PROJECT_NAME'].tolist()
print("Bottom 5 projects with the lowest accessible proportion:", tail_5_prop_
```

```
Top 5 projects with the highest accessible proportion: ['Ann Thomas Presbyteri
an', 'Bigham Place', 'Centennial Village', 'North Star Point Breeze', 'Roberto
Clemente Homes']
Bottom 5 projects with the lowest accessible proportion: ['Fairthorne Senior',
'APM Preservation', 'NewCourtland at Allegheny', 'Nativity BVM Senior', "St. R
aymond's House"]
```

```
/var/folders/b0/1wbtc4s57xb600v54tm8t_kh0000gn/T/ipykernel_23330/3359270527.p
y:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/st
able/user_guide/indexing.html#returning-a-view-versus-a-copy
  project02_dataset_proportion['ACCESSIBLE_PROPORTION'] = project02_dataset_pr
oportion['ACCESSIBLE_UNITS'] / project02_dataset_proportion['TOTAL_UNITS']
```

1.4 :The visitable units proportion of total units top 5? of all accessible units top 5? the visitable units proportion in this dataset?

In [32]:
```
##first clean the data column of VISITABLE_UNITS
project02_dataset_visitable = df.dropna(subset=['ACCESSIBLE_UNITS', 'VISITABLE_


most_visit_units = project02_dataset_visitable['VISITABLE_UNITS'].max()
print("Most visitable units:", most_visit_units)
most_visit_name = project02_dataset_visitable.loc[project02_dataset_visitable[
print("Project(s) with the most visitable units:", most_visit_name)


top_5_visit = project02_dataset_visitable.nlargest(5, 'VISITABLE_UNITS')
top_5_visit_name = top_5_visit['PROJECT_NAME'].tolist()
print("Top 5 projects with the most visitable units:", top_5_visit_name)
```

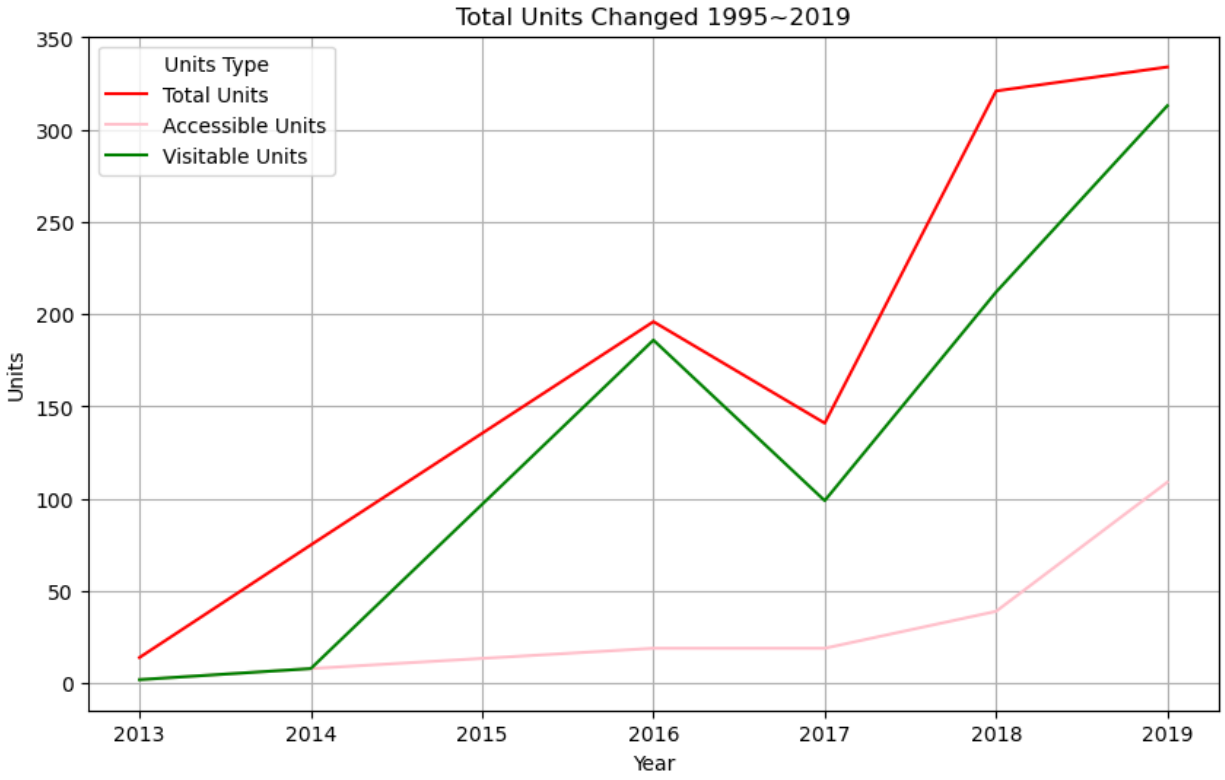
```
Most visitable units: 88.0
Project(s) with the most visitable units: 10    Ruth Williams House
Name: PROJECT_NAME, dtype: object
Top 5 projects with the most visitable units: ['Ruth Williams House', 'Ann Tho
mas Presbyterian', 'Nativity BVM Senior', 'Cantrell Place', 'Villas Del Carib
e']
```

1.5 :plot the changes of total numbers of TOTAL_UNITS, ACCESSIBLE_UNITS, VISITABLE_UNITS changed by year

In [35]:
```
##first clean the data columns
project02_dataset_year = df.dropna(subset=['TOTAL_UNITS', 'ACCESSIBLE_UNITS',


total_units_yearly = project02_dataset_year.groupby('FISCAL_YEAR_COMPLETE').agg
    'TOTAL_UNITS': 'sum',
    'ACCESSIBLE_UNITS': 'sum',
    'VISITABLE_UNITS': 'sum'
}).reset_index()
```

```
plt.figure(figsize=(10, 6))
plt.plot(total_units_yearly['FISCAL_YEAR_COMPLETE'], total_units_yearly['TOTAL_
plt.plot(total_units_yearly['FISCAL_YEAR_COMPLETE'], total_units_yearly['ACCESS
plt.plot(total_units_yearly['FISCAL_YEAR_COMPLETE'], total_units_yearly['VISITA
plt.title('Total Units Changed 1995~2019')
plt.xlabel('Year')
plt.ylabel('Units')
plt.legend(title='Units Type')
plt.grid(True)
plt.show()
```



Summary and Thoughts: In this project the main idea is to see the changes of affordable houses changed by year. There are three main factors: the total units, the accessible units, and the visitable units. and all of them can somehow show the affordable places for people in Philadelphia area. the total units means the properties listed under the Phialdelphia gov for people to search, but some data, mostly between early 1995 to 2010 are NA which means those properties listed not availeble during those times, it can also be seen in the plot. The accessible units are the affordable properties that could be on the list for people need special access to contact, but due to high demand and low (compared with the needs) supply, it grows slowly, which can show how the real affordable places in local Philly are still under high pressure, especailly consider the transportion that in local Philly the accessible units means it could help people with disability and lack of cars. The visitable units are increasing with the total number but still not enough, this part in my understading is how people can go to visit in person . This dataset contains the street number but not zip code, which I think it could be improved to make a better map visualable that how the units locate in Philly. The other part could be improved is there's no data of the prices for each/average units, which could show how each place's afforablity in local market.

Part 2: City_Facilities_pub dataset to show local city facilities in Philly and mapping

In [39]:
```
## create new dataset City_Facilities_pub

df_2 = pd.read_csv("City_Facilities_pub.csv")
df_2
df_2.head()
```

Out[39]:

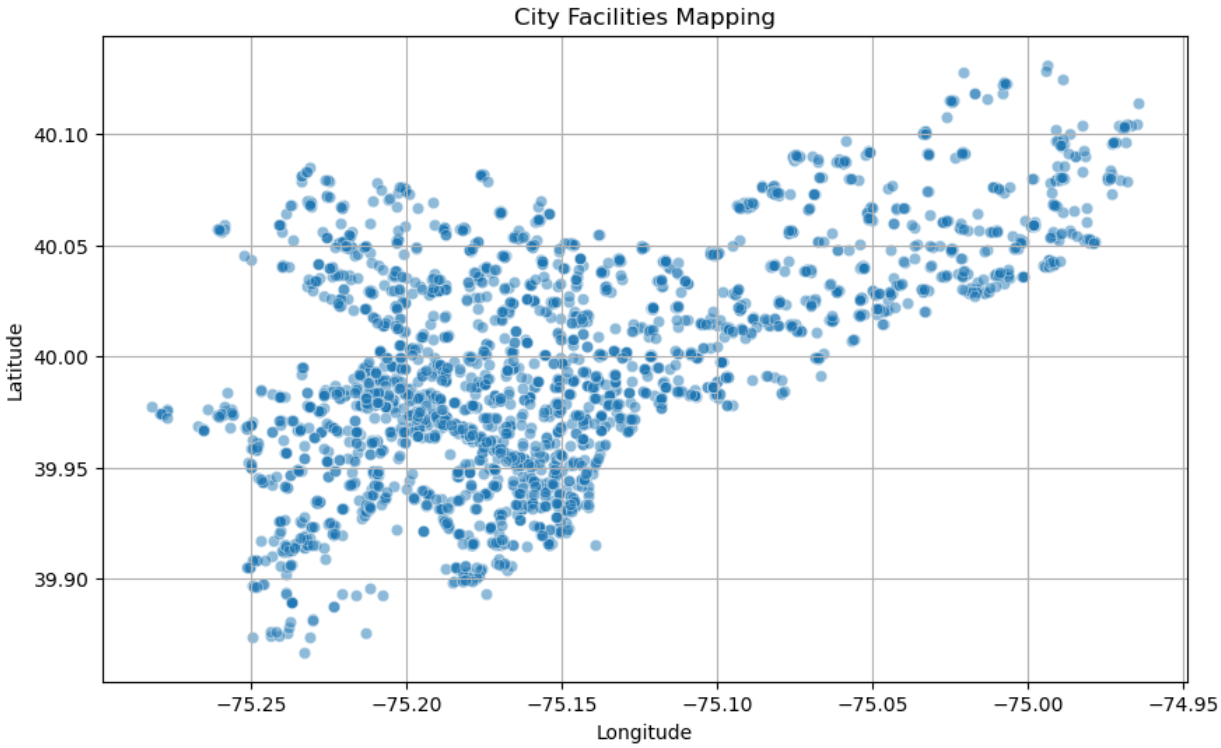
	X	Y	OBJECTID	ASSET_ADDR	ASSET_NAME	OPA_ADDR	SITE_NAME	SI
0	-75.152273	39.978575	1	1737 N 11TH ST	11th & Cecil B. Moore Playground Stands	NaN	11th & Cecil B. Moore Playground	
1	-75.223147	39.919949	2	2800 S 63RD ST	63rd & Lindbergh Storage Building 2	NaN	63rd & Lindbergh Park	
2	-75.223402	39.920237	3	2800 S 63RD ST	63rd & Lindbergh Storage Building 1	NaN	63rd & Lindbergh Park	
3	-75.242966	39.910034	4	2604 ISLAND AVE	75th & Chelwynde Park	NaN	75th & Chelwynde Park	
4	-75.240686	40.059310	5	600 PORT ROYAL AVE	Al Pearlman Sports Center Barn	NaN	Al Pearlman Sports Center	

5 rows × 34 columns

In [42]:
```
## create the mapping of the city facilities' locations by using x,y

import matplotlib.pyplot as plt
import seaborn as sns   ##show mapping use seaborn library

# Plotting
plt.figure(figsize=(10, 6))
sns.scatterplot(x='X', y='Y', data=df_2, alpha=0.5)
plt.title('City Facilities Mapping')
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.grid(True)
plt.show()
```



Summary and Thoughts: In this part the idea is to show how the city facilities in local Philadelphia, we can see most city facilities concentrate near the central area, and the south part is more dense than the north, this could help locals to decide when choosing to rent or buy an affordable housing compared with Part 1. It can be developped when adding the affordable housing data in the map.

Part 3: arrests_citywide dataset

In [46]:
```
## create new dataset arrests_citywide
df_3 = pd.read_csv("arrests_citywide.csv")
df_3
df_3.head()
```

Out[46]:

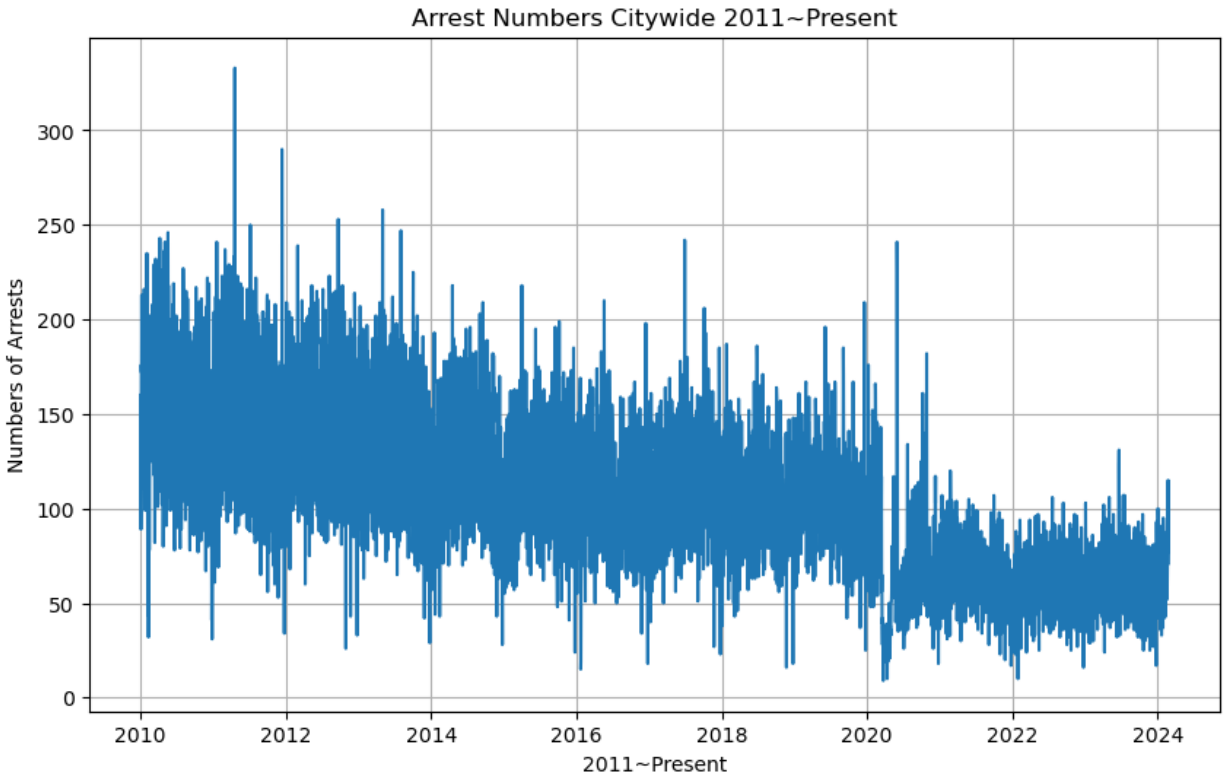
	offense_category	day	defendant_race	count	objectid
0	Aggravated Assault	2010-06-13 04:00:00+00	Latinx	1	38734696
1	Aggravated Assault	2010-06-13 04:00:00+00	White	3	38734697
2	Aggravated Assault	2010-06-14 04:00:00+00	Black	12	38734698
3	Aggravated Assault	2010-06-15 04:00:00+00	Black	14	38734699
4	Aggravated Assault	2010-06-15 04:00:00+00	Latinx	4	38734700

In [53]:
```
df_3['day'] = pd.to_datetime(df_3['day']).dt.date ##convert the format to day
df_3['day']

arrests_num = df_3.groupby('day')['count'].sum().reset_index() ##groupby the da

# Plotting
plt.figure(figsize=(10, 6))
```

```
plt.plot(arrests_num['day'], arrests_num['count'])
plt.title('Arrest Numbers Citywide 2011~Present')
plt.xlabel('2011~Present')
plt.ylabel('Numbers of Arrests')
plt.grid(True)
plt.show()
```



Summary and Thoughts: In this part the idea is to show how the daily arrests in local Philadelphia, we can see during the past 14 years (data is from 2011 to present) the overall trend is going down, and one specific timing was within 2020 which could be estimated as the pandemic effect. This dataset could be improved by using mapping factors to identified the police arrest locations that could be combined with the upper two dataset to mapping a whole new affordable housing locations where people could better consider.

Overall Summary: By analysis the three dataset we can have a better picture of local Philadelphia city when people choose affordable housing, not only by the factor of the housing type and its avalebility but also its convienence with other public facilities, and maybe the most importance, the safety. The project also shows how different local datas which seem like unralateble could be analyzed into a new picture if using a proper tech and tool. This combination clearly shows the potential of these data analysis into building not a better personal choice in housing but also how a better city could develop.

In [ ]: