Framing an Analytics Problem

Case 2 – Solar Electric Incentive Program – Project cost analysis case study

- The objective of this 2nd case is to understand the factors that can affect or contribute to the cost of a project.
- The dataset we use for this case study is has been sourced from New York State Energy Research and Development Authority(NYSERDA) and it includes the data points for Solar electric projects in the Incentive Program that started in December 2000.

Data Information

Variable	Description
City	Name of city for project location
County	Name of county for project location
Sector	Name of project sector. The sectors in this dataset are either Residential or Non-Residential
Program Type	Name of program type; either Residential/Small Commercial, Commercial/Industrial (Competitive), or Commercial/Industrial (MW Block)
Electric Utility	Name of electric utility for project location
Purchase Type	Solar photovoltaic project purchase agreement type. The purchase types are either Lease, Purchase or Power Purchase Agreement.
Date Application Received	Date project application was received by the program
Date Completed	Date NYSERDA recognized the project as interconnected and operational, and closed out the project application.
Project Status	Either Complete or Pipeline. Complete indicates projects that are interconnected and operational, and closed out the project application. Pipeline indicates projects with an active application that are not yet complete.
Total Inverter Quantity	Quantity of all inverters installed for project.
Total PV Module Quantity	Quantity of all photovoltaic (PV) modules installed for project.
Project Cost	Expected project installation cost in US dollars (USD), as reported by the solar project contractor.
$Incentive	Amount of project incentives paid by the program in USD.
Total Nameplate kW DC	The sum of kilowatt (kW) DC capacity ratings of the installed photovoltaic modules
Expected KWh Annual Production	Expected annual electricity production in kilowatt-hours (kWh)
Affordable Solar	Indicates if project is part of Affordable Solar program
Community Distributed Generation	Indicates if project Community Distributed Generation (Shared Solar)
Contractor	Name of entity responsible for installation of the project.

Our first task here would be to look at the type of data that has been made available to us for analyzing. Let's take a look at the datatypes of the variables.

Categorical Variables:

Binary:

Sector

Program Type

Project Status

Affordable Solar

Community Distributed Generation

Multi-level:

City

County

Electric Utility

Purchase Type

Date Time Type Variables:

Date Application Received

Date Completed

Continuous Variables:

Total Inverter Quantity

Total PV Module Quantity

Project Cost

Incentive

Total Nameplate kW DC

Expected KWh Annual Production

Task at Hand

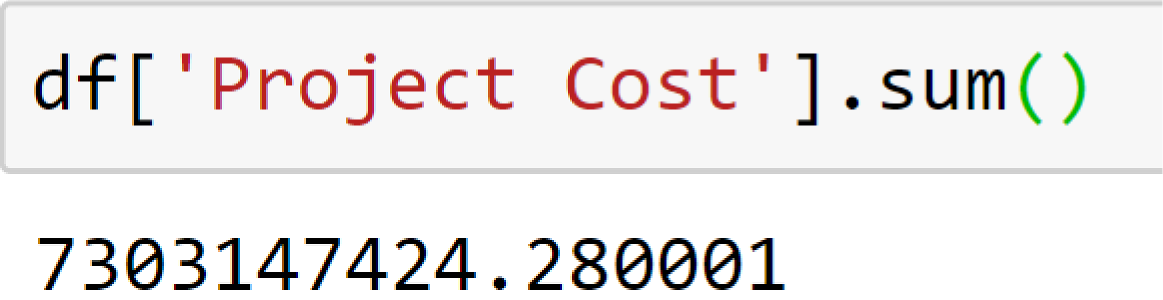
- The Data Scientists in the team have access to this data and they need to figure out how they can use this available data to find out different factors that can contribute to the overall project cost.
- The intent is to identify the avenues where there is a scope of reducing the cost of the project.

Some questions that can be raised initially that can act as a starting point to analyse the dataset

- What is the total project cost that has been estimated in Solar Electric projects since the beginning of this incentive program?
- How are the projects distributed across different cities?
- What is the current status of the projects? What proportion of projects are completed and up & running?
- More the number of photovoltaic modules installed, more would be the DC current generating capacity subsequently more would be the Expected annual electricity production in kilowatt-hours (kWh). Which plot can we use to verify this claim?
- Are higher costing projects estimated to produce more electricity in terms of annual production?
- Amongst the completed project, which project contractor seems to have exceeded the timeline of 5 years to complete the project since the date of application?
- Which type of program seems to have a higher cost of project? Does it adhere with the expected annual electricity production?
- Do incentives depend on the cost of the project?
- Every city can have multiple electric utilities. How can we visualize the contribution of different electric utilities to the cost of the project for the top 10 cities we previously identified?

Lets start answering these questions using the data at our disposal.

What is the total project cost that has been estimated in Solar Electric projects since the beginning of this incentive program?



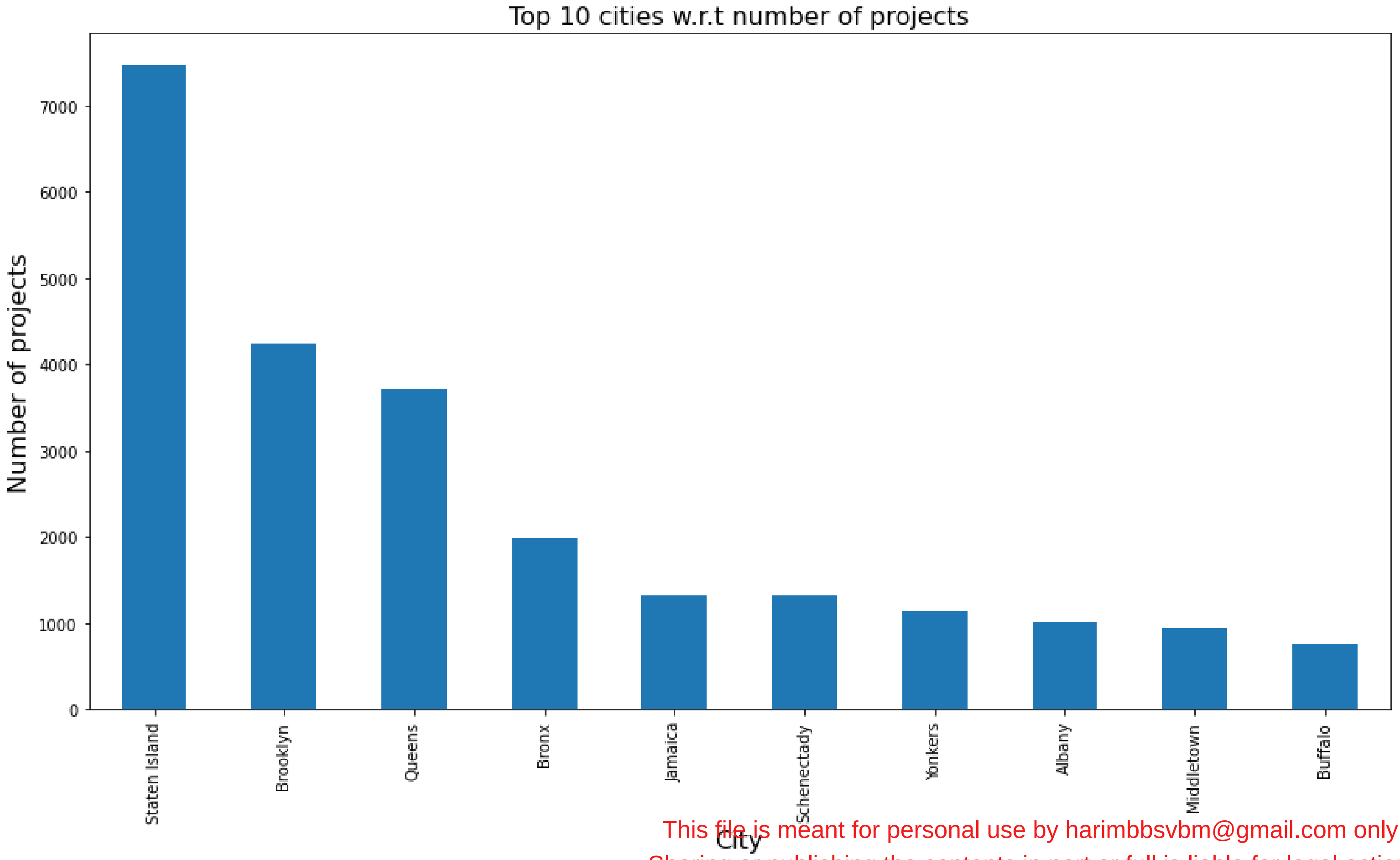
```
df['Project Cost'].sum()
```

```
7303147424.280001
```

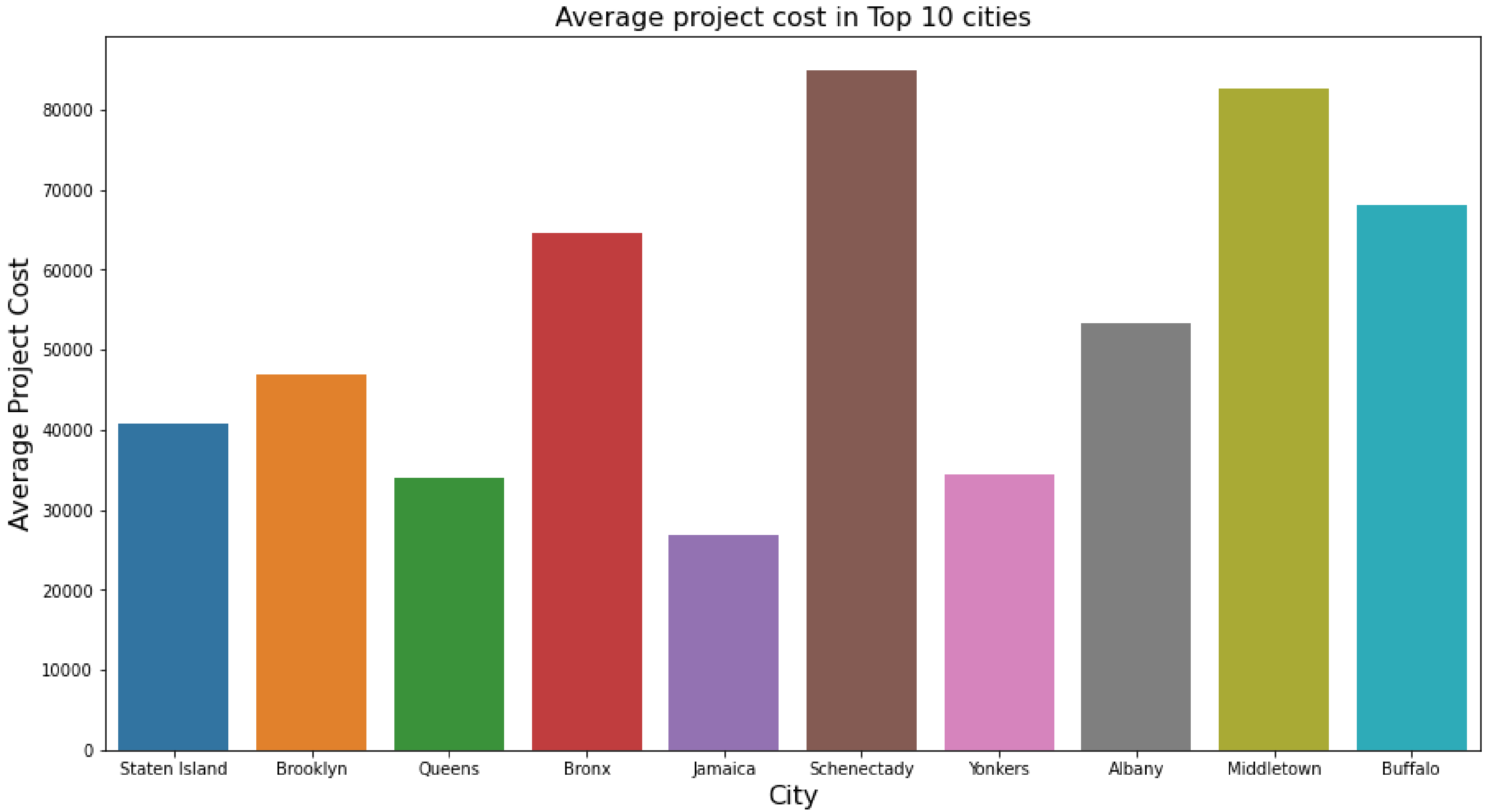
Approximately, $7.3 billion have been invested in the Solar Electric project since the beginning of the incentive program.

How are the projects distributed across different cities?

This question can further be divided to identify the top 10 cities that have most number of projects implemented/in progress as there might be many entries for city and visualizing all in a single graph becomes difficult.

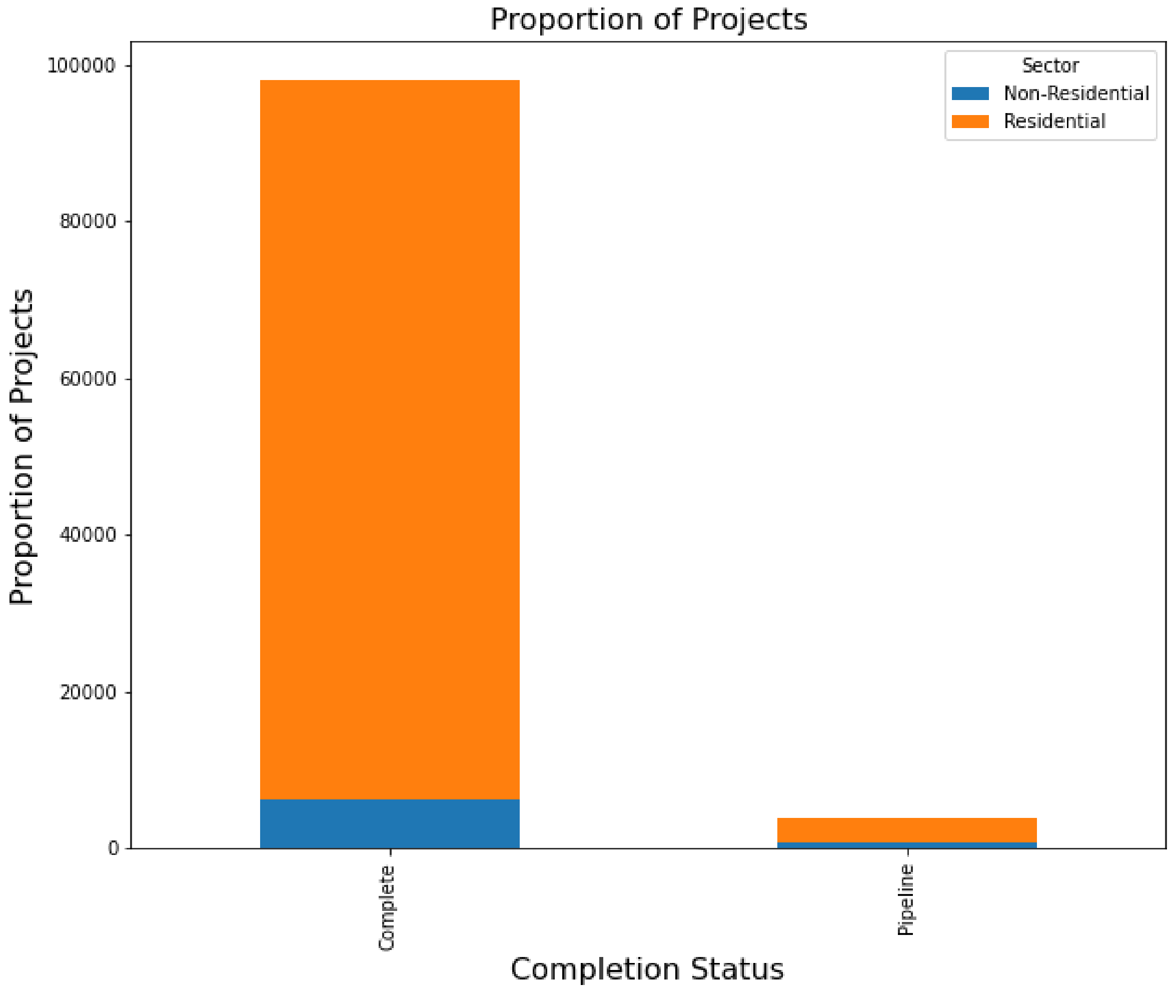


Among the top 10 cities with most number of projects, which cities has the most project cost on an average?



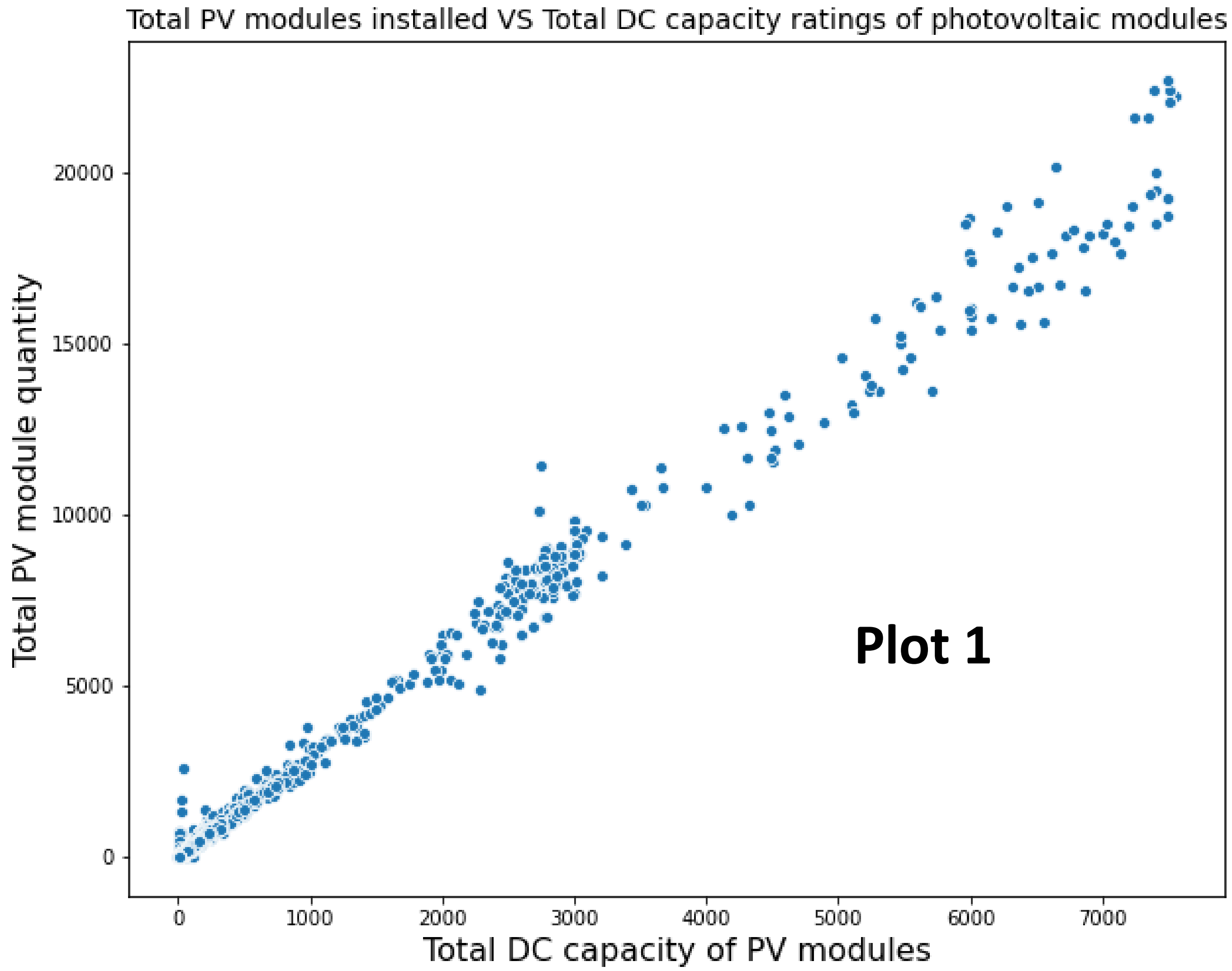
Even though Staten Island has the most number of projects, the mean project cost is highest in the city of Schenectady.

What is the current status of the projects? How many projects are completed and up & running from each sector?



Approximately 95% projects have been completed. Majority of the completed projects are from Residential sector.

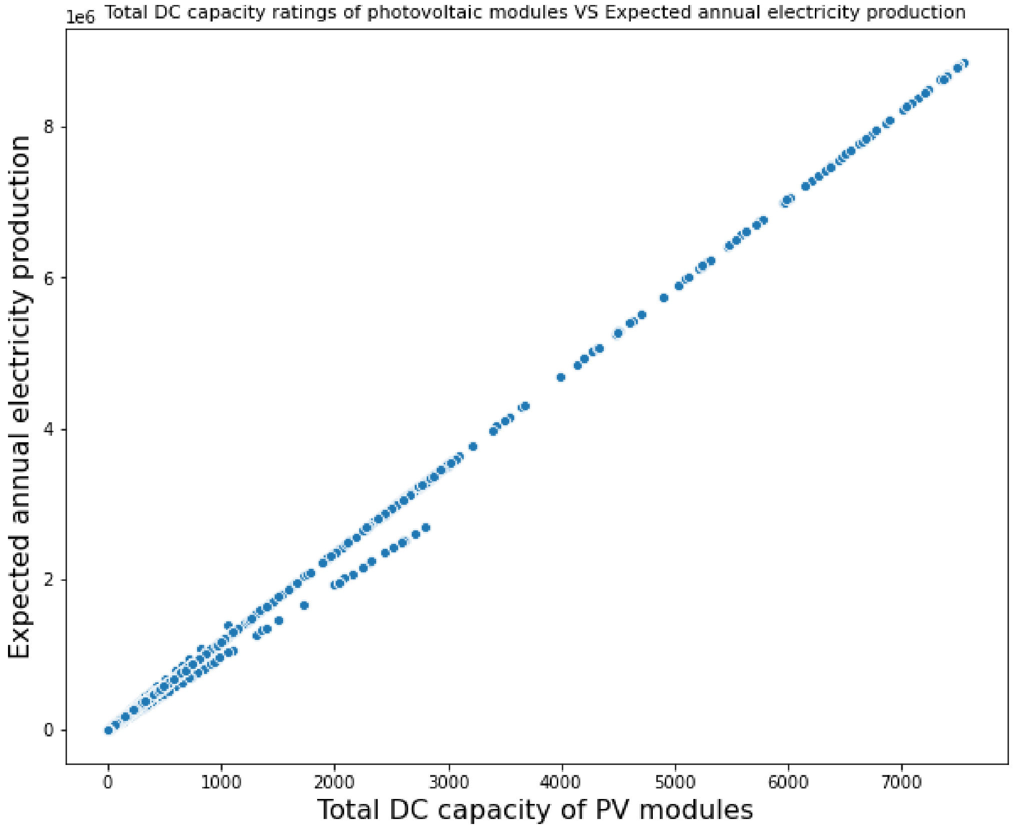
More the number of photovoltaic modules installed, more would be the DC current generating capacity subsequently more would be the Expected annual electricity production in kilowatt-hours (kWh). Use an appropriate plot to verify this claim.

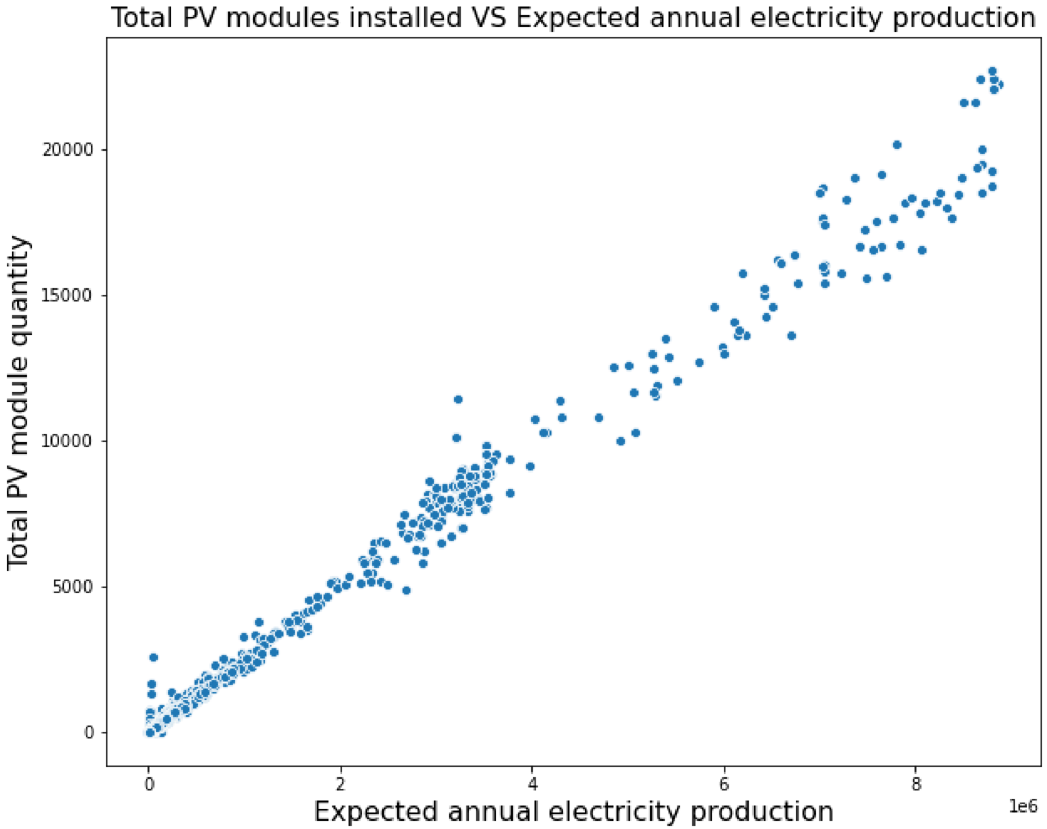


Plot 1

Continued on next slide….

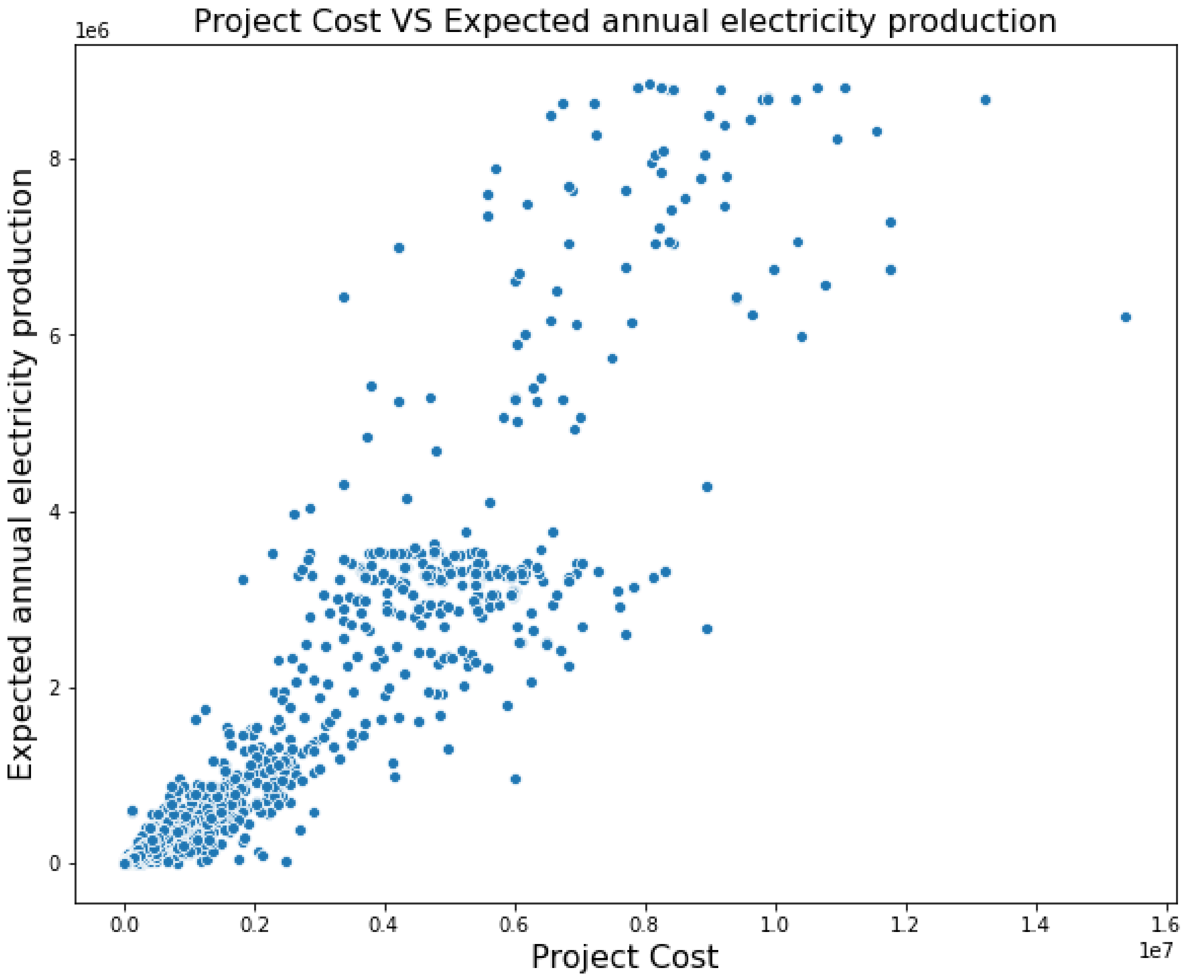
Continued from previous slide…


Total DC capacity ratings of photovoltaic modules VS Expected annual electricity production


Total PV modules installed VS Expected annual electricity production

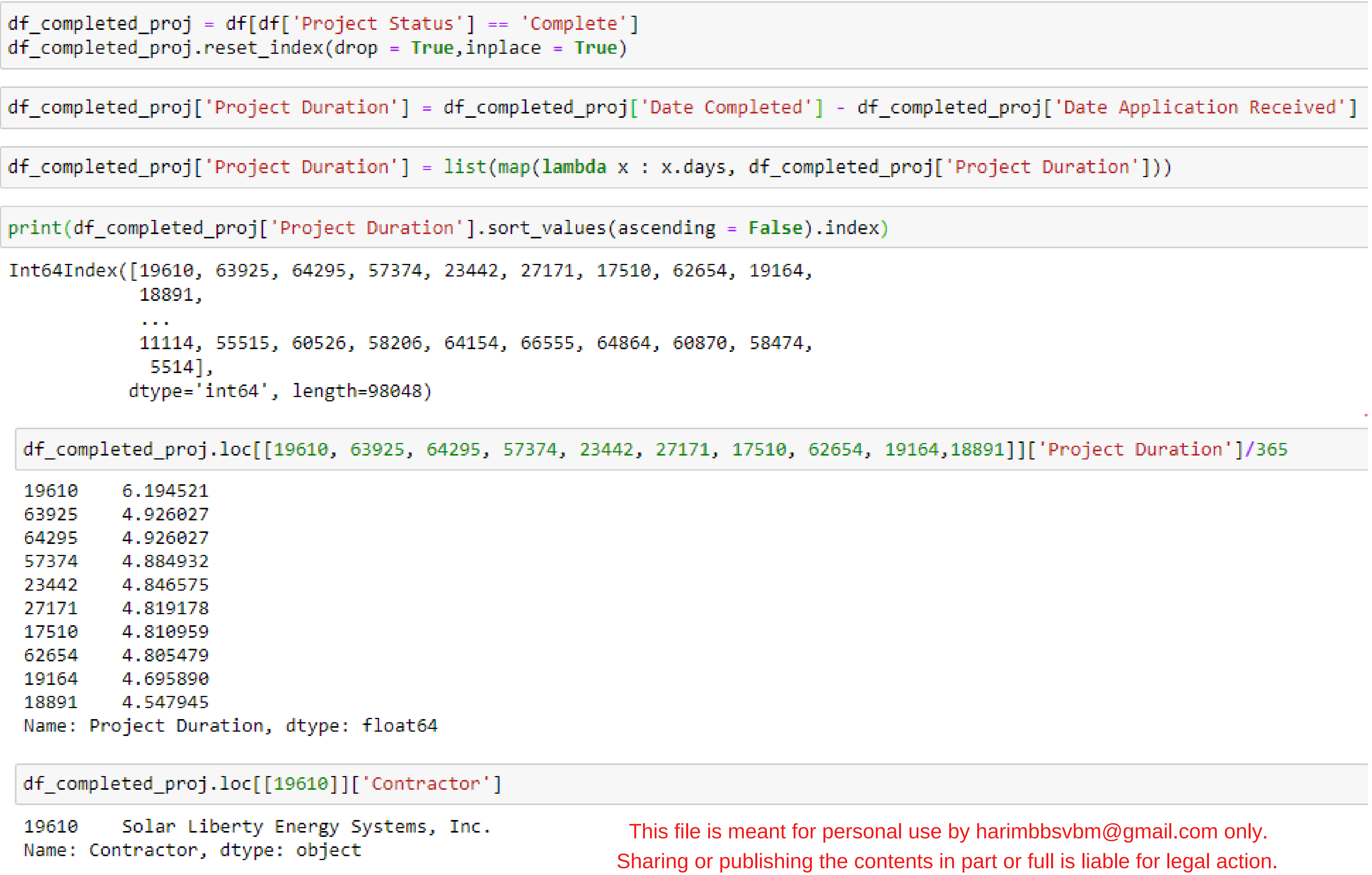
Using a statistical test (ANOVA), this can further be verified by formulating a hypothesis as follows:
H0 (Null Hypothesis): correlation for plot 1 = correlation for plot 2 = correlation for plot 3
Ha (Alternative Hypothesis): correlation for plot 1 != correlation for plot 2 != correlation for plot 3

Are higher costing projects estimated to produce more electricity in terms of annual production?



Looking at the plot, we can see there seems to be a positive correlation between project cost and Expected annual electricity production. So, quite clearly it looks like higher costing project might produce more electricity. But, its also important to keep in mind that "Correlation does not imply Causation".

Amongst the completed projects, which project contractors seems to have exceeded the deadline of 5 years to complete the project since the date of application?



```
df_completed_proj = df[df['Project Status'] == 'Complete']
df_completed_proj.reset_index(drop = True,inplace = True)
```

```
df_completed_proj['Project Duration'] = df_completed_proj['Date Completed'] - df_completed_proj['Date Application Received']
```

```
df_completed_proj['Project Duration'] = list(map(lambda x : x.days, df_completed_proj['Project Duration']))
```

```
print(df_completed_proj['Project Duration'].sort_values(ascending = False).index)
```

```
Int64Index([19610, 63925, 64295, 57374, 23442, 27171, 17510, 62654, 19164,
            18891,
            ...
            11114, 55515, 60526, 58206, 64154, 66555, 64864, 60870, 58474,
             5514],
           dtype='int64', length=98048)
```

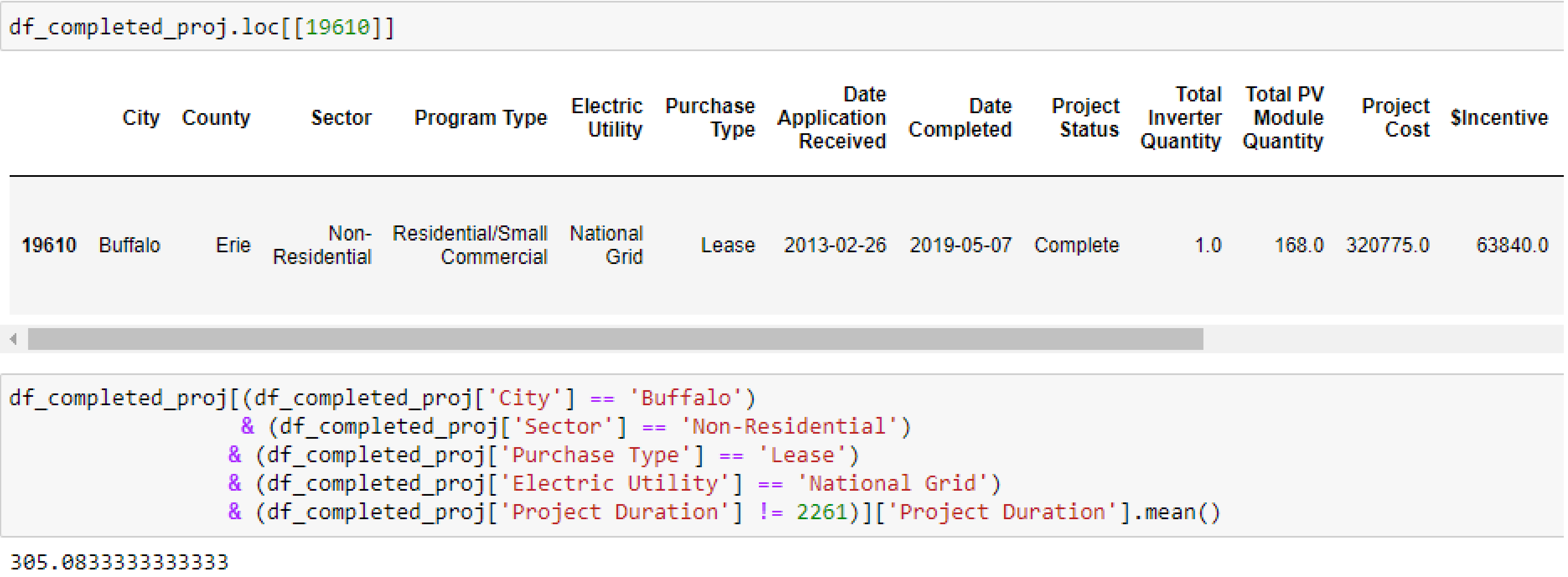
```
df_completed_proj.loc[[19610, 63925, 64295, 57374, 23442, 27171, 17510, 62654, 19164,18891]]['Project Duration']/365
```

```
19610    6.194521
63925    4.926027
64295    4.926027
57374    4.884932
23442    4.846575
27171    4.819178
17510    4.810959
62654    4.805479
19164    4.695890
18891    4.547945
Name: Project Duration, dtype: float64
```

```
df_completed_proj.loc[[19610]]['Contractor']
```

```
19610    Solar Liberty Energy Systems, Inc.
Name: Contractor, dtype: object
```

How much time did similar projects take for completion on an average?



```
df_completed_proj.loc[[19610]]
```

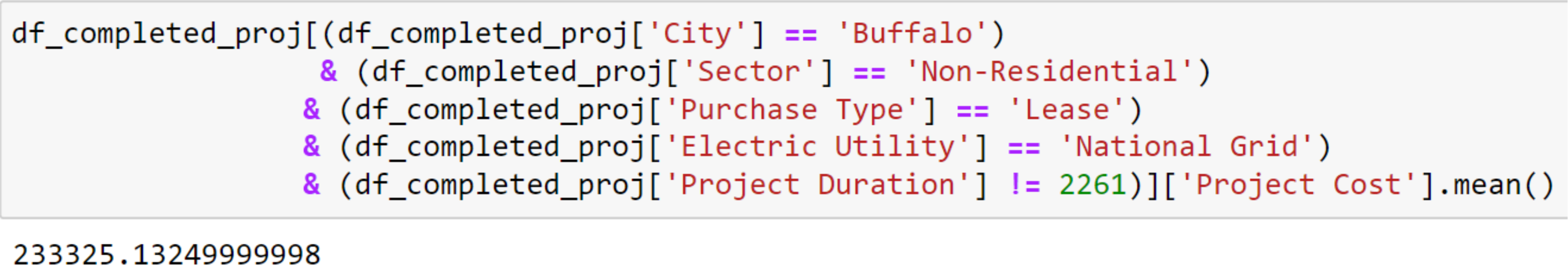
	City	County	Sector	Program Type	Electric Utility	Purchase Type	Date Application Received	Date Completed	Project Status	Total Inverter Quantity	Total PV Module Quantity	Project Cost	$Incentive
19610	Buffalo	Erie	Non-Residential	Residential/Small Commercial	National Grid	Lease	2013-02-26	2019-05-07	Complete	1.0	168.0	320775.0	63840.0

```
df_completed_proj[(df_completed_proj['City'] == 'Buffalo')
                   & (df_completed_proj['Sector'] == 'Non-Residential')
                  & (df_completed_proj['Purchase Type'] == 'Lease')
                  & (df_completed_proj['Electric Utility'] == 'National Grid')
                  & (df_completed_proj['Project Duration'] != 2261)]['Project Duration'].mean()
```

```
305.0833333333333
```

Contractor named "Solar Liberty Energy Systems, Inc." has taken the highest time(more than 6 years) to install the project and get it up & running. It takes just about 1 year to implement similar projects. Whereas this project took about 6 years to complete.

How much money was invested in similar projects on an average?



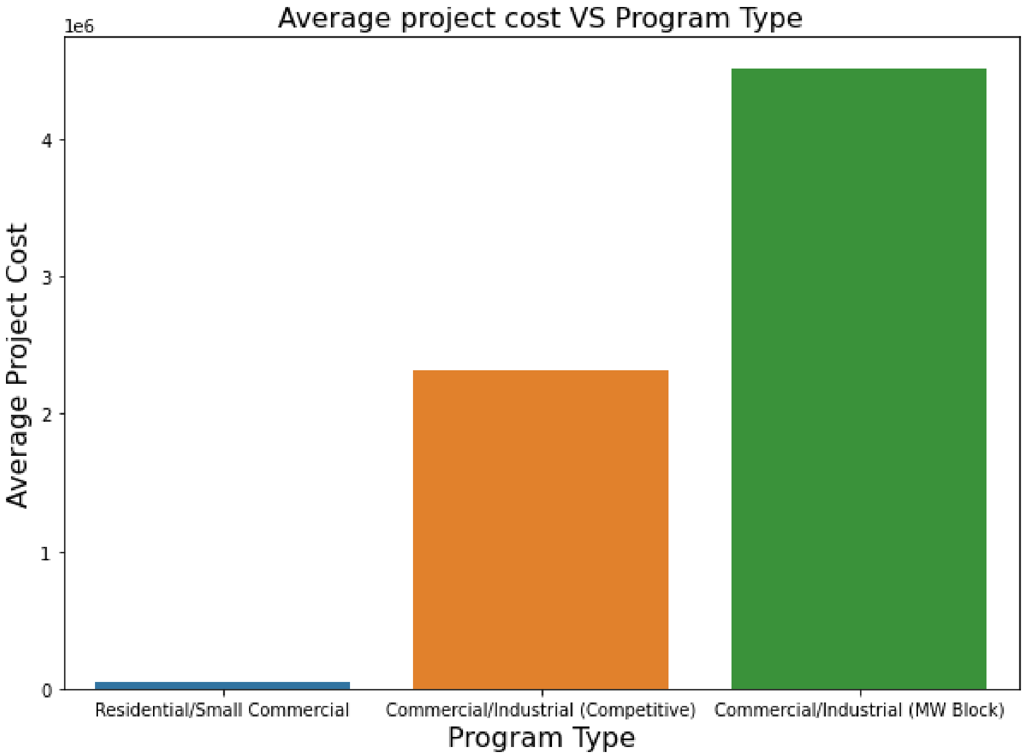
```
df_completed_proj[(df_completed_proj['City'] == 'Buffalo')
                  & (df_completed_proj['Sector'] == 'Non-Residential')
                 & (df_completed_proj['Purchase Type'] == 'Lease')
                 & (df_completed_proj['Electric Utility'] == 'National Grid')
                 & (df_completed_proj['Project Duration'] != 2261)]['Project Cost'].mean()
```

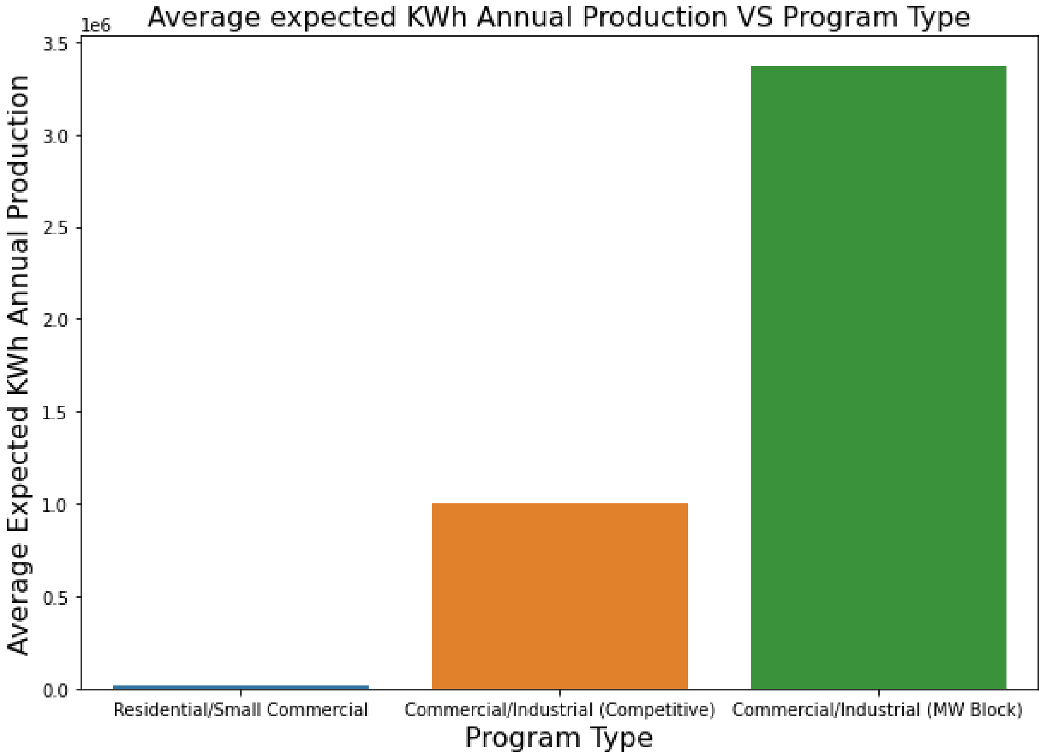
```
233325.13249999998
```

On an average, approximately 233 thousand dollars were invested in similar projects, but about 320 thousand dollars were invested in this specific project which is more than the average.

More details regarding the reasons for the delay in completion of the project as well as the higher cost of the project should be discussed with the contractor.

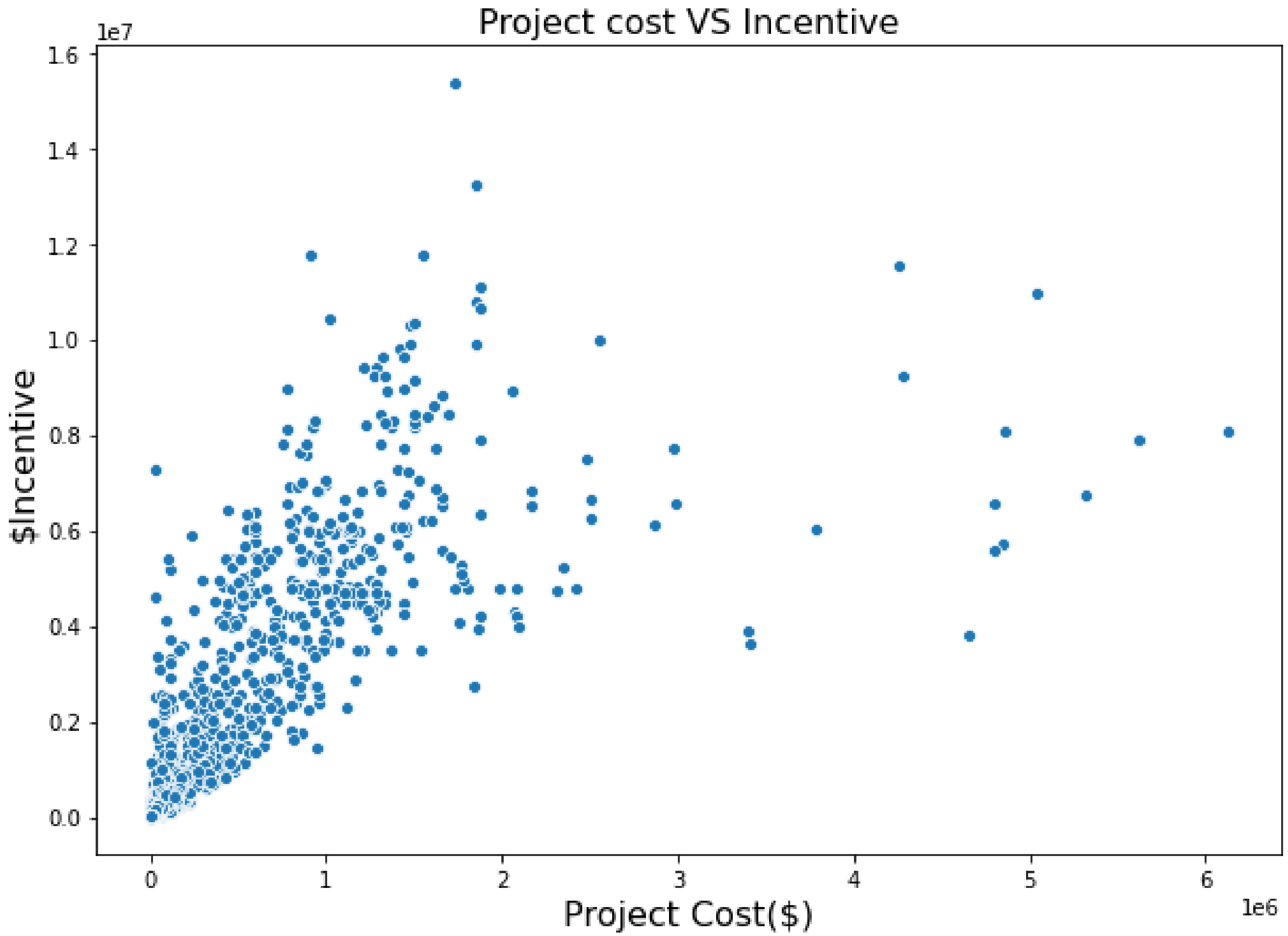
Which type of program seems to have a higher cost of project? Does it adhere with the expected annual electricity production?





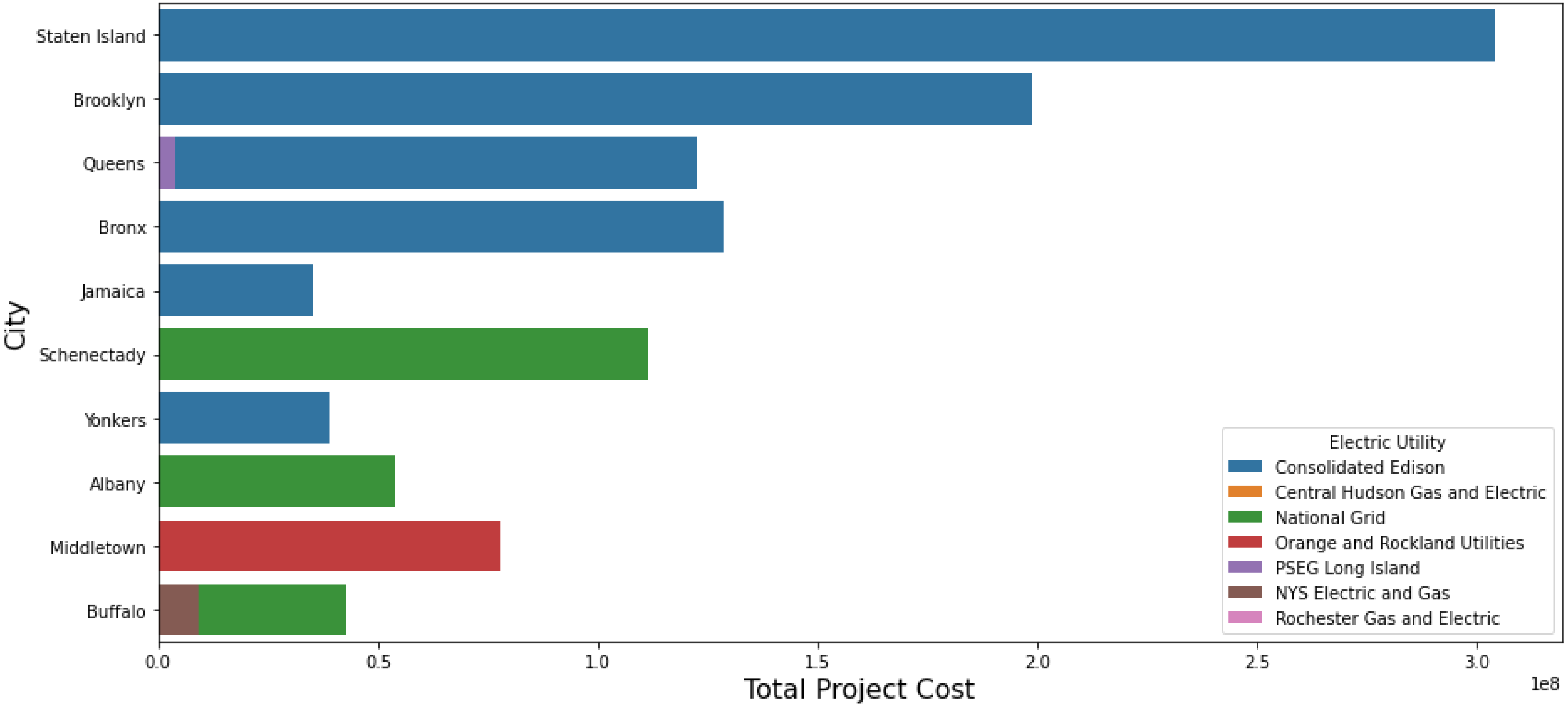
As we can see from the plots, the project cost is higher for "Commercial/ Industrial(MegaWatt Block)" program type which is quite obvious as it is expected to generate the highest amount of annual electricity production.

Do incentives depend on the cost of the project?



There seems to be a positive correlation between the project cost and the incentives. But this might not actually be the only criteria for deciding the amount of incentive. Remember the phrase, "Correlation does not imply causation". Its highly likely that there can be other factors involved to decide the incentive amount.

Every city can have multiple electric utilities. How can we visualize the contribution of different electric utilities to the cost of the project for the top 10 cities we previously identified?



From the above stacked bar plot, we can clearly see that most used electric utility is “Consolidated Edison”. “Orange and Rockland Utilities” is the only electric utility available in Middletown. “National Grid” is the only utility available in Schenectady and Albany and a majorly used utility in Buffalo.

- This data can further be refined and used for prediction of the project cost as well as the incentives.
- This dataset has a vast scope of drilling down to a specific case/situation and analyzing the data revolving around a certain situation, as we did in case of identifying the project contractor who did not meet the 5 year completion deadline.
- This is just an introduction to what can be the possibilities of exploring a certain dataset to find meaningful insights w.r.t cost analysis. If you have set an objective in mind, you can narrow down this analysis to just follow the path to reach the objective which will also save a lot of time.

Thank You!

END!