IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akhil Sharma
18-Jan-2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Data Collection using API and Web Scraping
- Exploratory Data Analysis using SQL and visualization
- Interactive visual analytics using Folium and Plotly Dash
- Predictive Analysis using Classification models (Logistic regression , SVM , Decision trees and K Nearest Neighbour)

Introduction

In this Project we will predict if the Falcon 9 first stage will land successfully.

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - SpaceX REST API
    - Web Scraping related wiki pages
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Classification models like Logistic Regression , SVM , Decision Tree , KNN etc are applied in the Spacex Dataset. It has been trained with Train data set and then tested on test data set. Confusion Matrix has been plotted to compare the various classification models.

Data Collection

Data Collection using API :- A get request to the SpaceX API.

Request to the SpaceX API

Clean the data

Data Collection using Web Scraping :-Web scrap Falcon 9 launch records with BeautifulSoup:

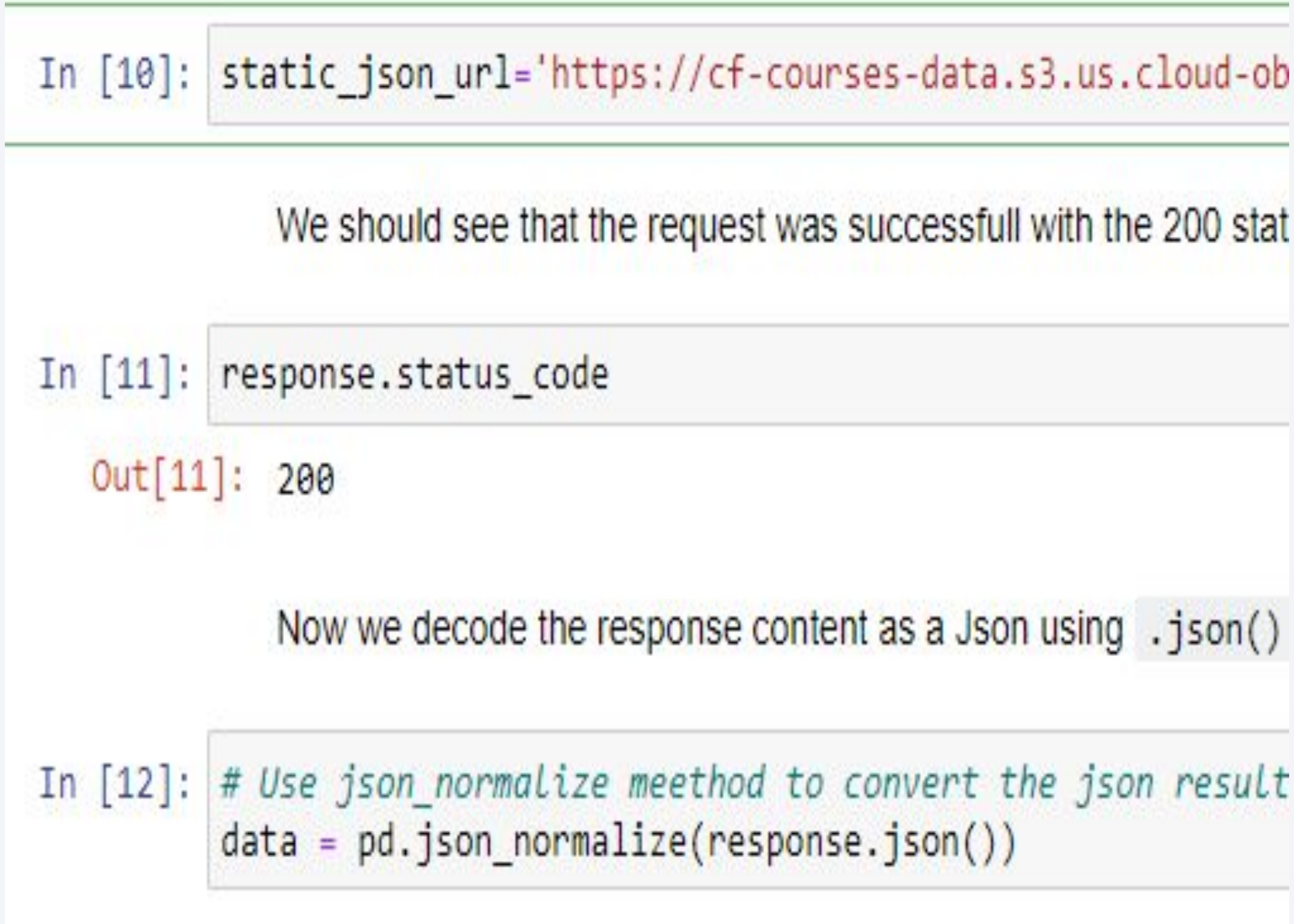
Extracted a Falcon 9 launch records HTML table from Wikipedia

Parsed the table and converted it into a Pandas data frame

Data Collection – SpaceX API

- GET request to read data from API.

- **GIT HUB LINK**



```
In [10]: static_json_url='https://cf-courses-data.s3.us.cloud-ob
```

We should see that the request was successfull with the 200 stat

```
In [11]: response.status_code

Out[11]: 200
```

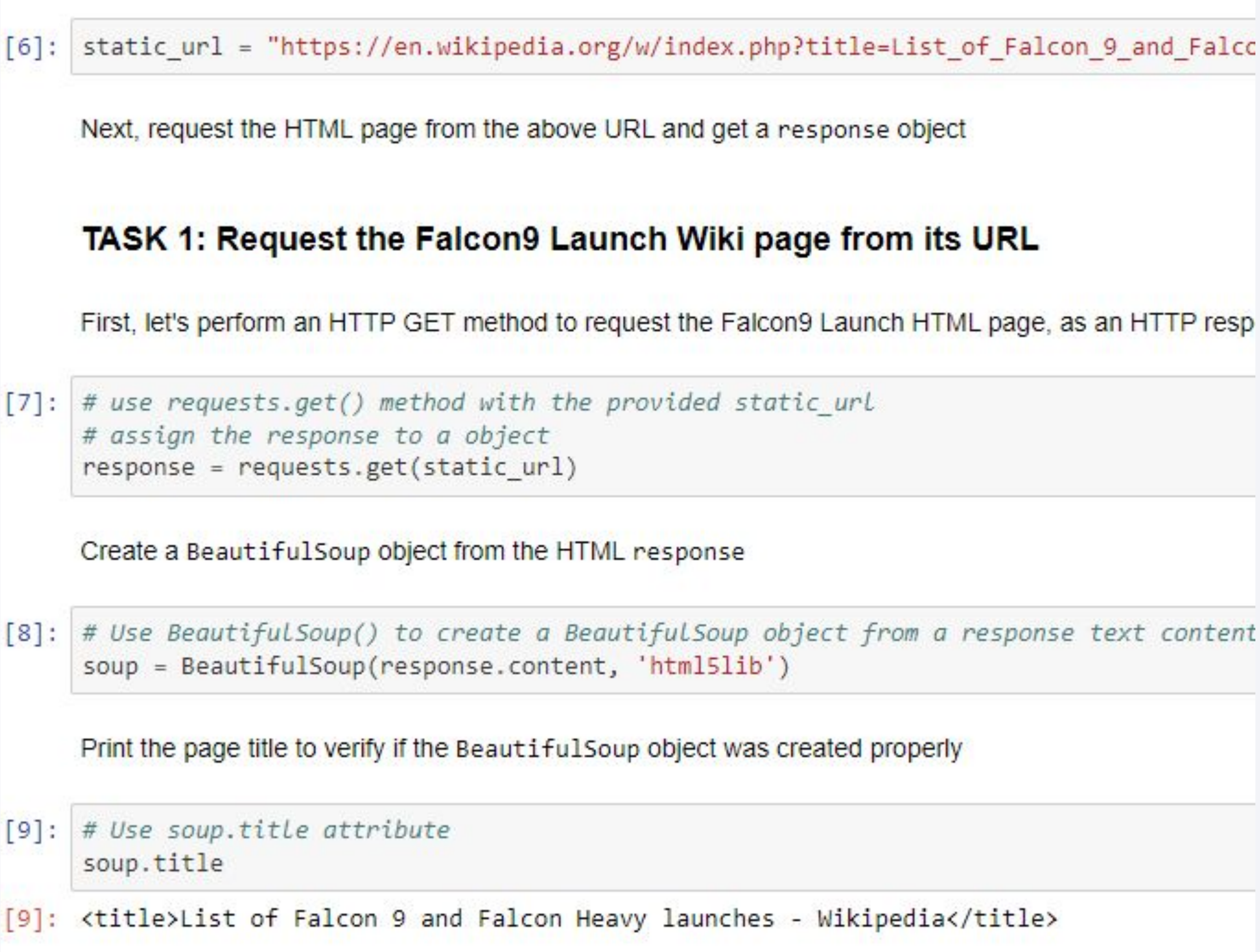
Now we decode the response content as a Json using `.json()`

```
In [12]: # Use json_normalize meethod to convert the json result
         data = pd.json_normalize(response.json())
```

Data Collection - Scraping

- Extracted a Falcon 9 launch records HTML table from Wikipedia using BeautifulSoup object.

- [GIT HUB LINK]



```
[6]: static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falco
```

Next, request the HTML page from the above URL and get a response object

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP resp

```
[7]: # use requests.get() method with the provided static_url
     # assign the response to a object
     response = requests.get(static_url)
```

Create a BeautifulSoup object from the HTML response

```
[8]: # Use BeautifulSoup() to create a BeautifulSoup object from a response text content
     soup = BeautifulSoup(response.content, 'html5lib')
```

Print the page title to verify if the BeautifulSoup object was created properly

```
[9]: # Use soup.title attribute
     soup.title

[9]: <title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

Data Wrangling

- Data Wrangling has been performed to perform Exploratory Data Analysis and determining the training Labels.
  - Calculated the number of launches on each site
  - Calculated the number and occurrence of each orbit
  - Calculated the number and occurence of mission outcome per orbit type
  - Created a landing outcome label from Outcome column

GIT Hub URL

EDA with Data Visualization

Below Charts are plotted for Exploratory Data Analysis :-

> Scatter Plot to see the relationship between Flight Num & Launch Site , , Launch Site & Payload , Orbit & Flight Num , Payload & Orbit Type.

> Bar Chart to observe the success rate by orbit type.

> Linehart to observe the success rate by Year.

GIT Hub URL

EDA with SQL

- *Display the names of the unique launch sites in the space mission*
- *Display 5 records where launch sites begin with the string 'CCA*
- *Display the total payload mass carried by boosters launched by NASA (CRS)*
- *List the date when the first successful landing outcome in ground pad was acheived*
- *List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*
- *List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015*
- *Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order*

- ***https://github.com/akhilsharma43/Firstrepo/blob/master/EDA-SQL.ipynb***

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- For each launch site, added a Circle object based on its coordinate (Lat, Long) values
- Markers are added for success/failure of each launch site.
- Calculated the distance of Launch sites with its proximities.

- **https://github.com/akhilsharma43/Firstrepo/blob/master/interactive_visual_analytics_lab.ipynb**

Build a Dashboard with Plotly Dash

- Pie chart has been added to show the success rate of various launch sites

- Slider has been added to select the payload range.

Predictive Analysis (Classification)

- Created a column for class.
- Standardized the space X data
- Data has been splitted into Test and train data.
- Best Hyperparameter for Logistic regression , SVM , KNN and Decision tree have been selected using Grid Search
- Classifier methods has been compared based on their result on test data.

- https://github.com/akhilsharma43/Firstrepo/blob/master/ML_prediction.ipynb

Results

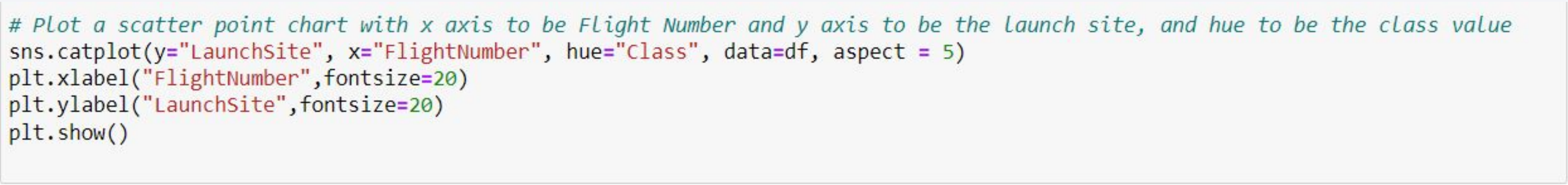
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

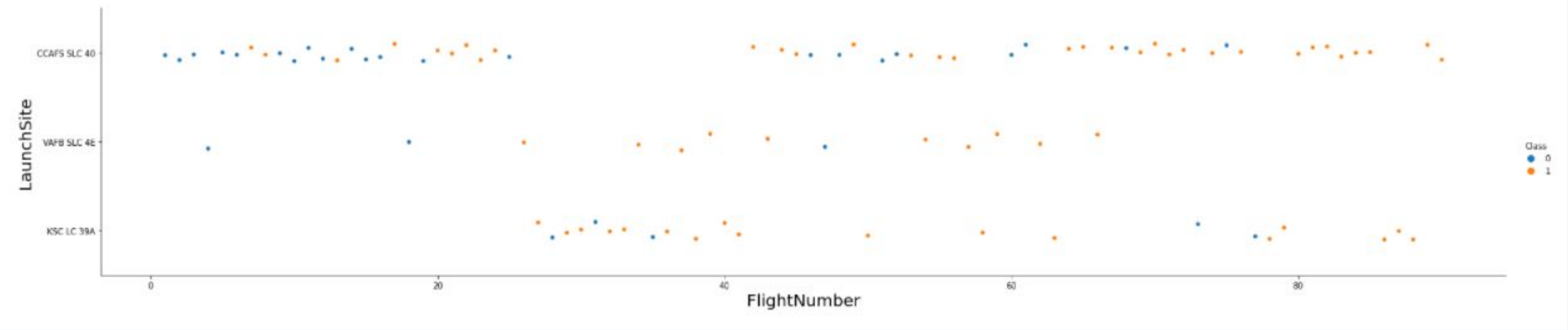
Insights drawn from EDA

Flight Number vs. Launch Site

scatter plot of Flight Number vs. Launch Site



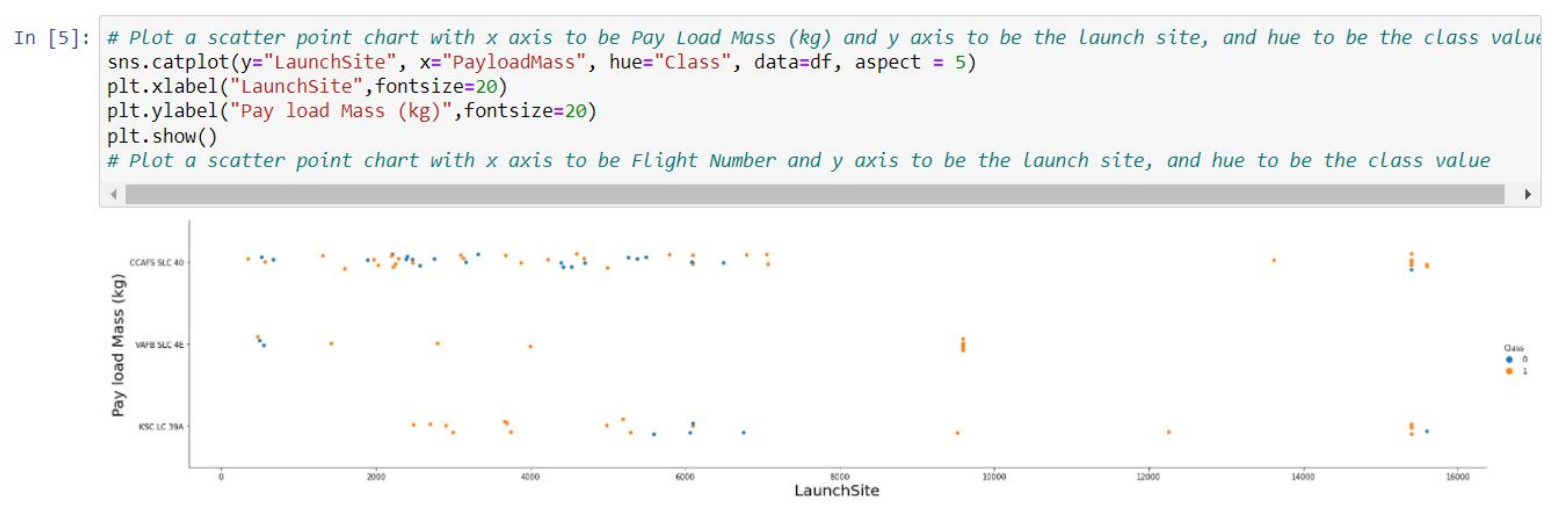
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



successful landing is more with less flight number except for CCAFS LS -40.

Payload vs. Launch Site

Scatter plot of Payload vs. Launch Site

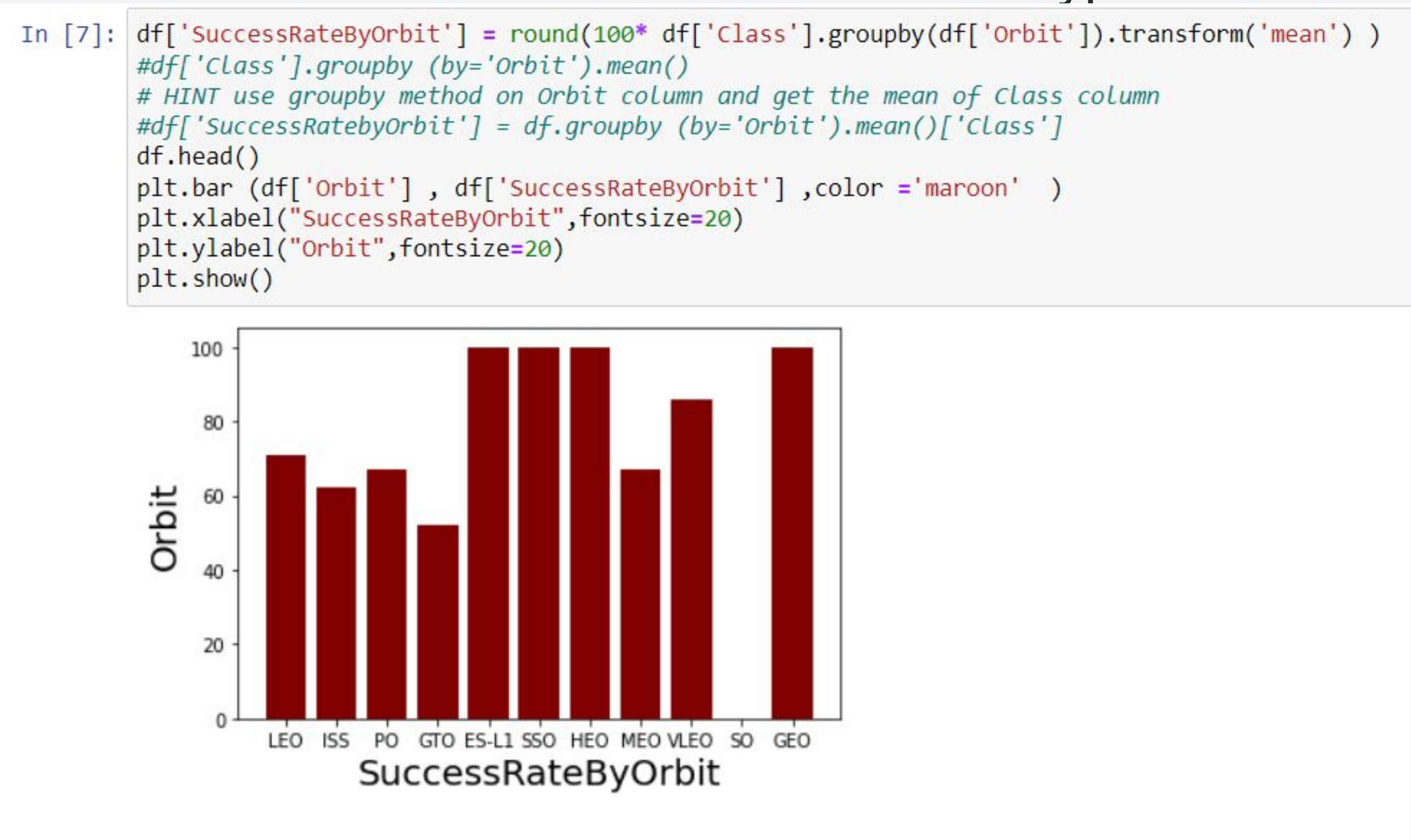


```
In [5]: # Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
        sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
        plt.xlabel("LaunchSite",fontsize=20)
        plt.ylabel("Pay load Mass (kg)",fontsize=20)
        plt.show()
        # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
```

For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type

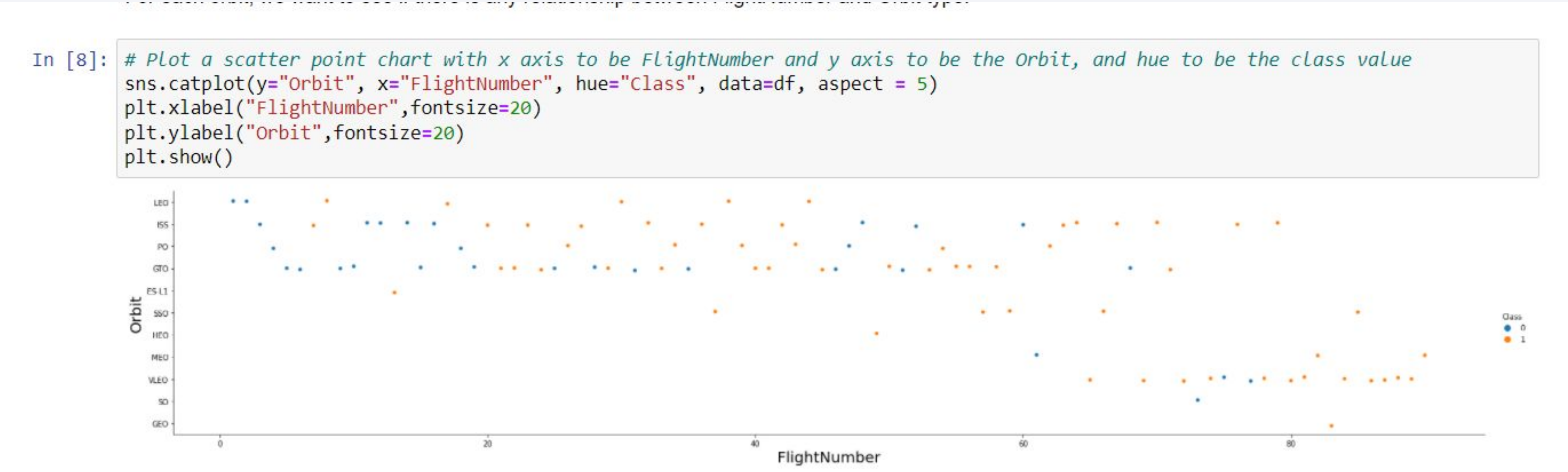


```
In [7]: df['SuccessRateByOrbit'] = round(100* df['Class'].groupby(df['Orbit']).transform('mean') )
        #df['Class'].groupby (by='Orbit').mean()
        # HINT use groupby method on Orbit column and get the mean of Class column
        #df['SuccessRatebyOrbit'] = df.groupby (by='Orbit').mean()['Class']
        df.head()
        plt.bar (df['Orbit'] , df['SuccessRateByOrbit'] ,color ='maroon'  )
        plt.xlabel("SuccessRateByOrbit",fontsize=20)
        plt.ylabel("Orbit",fontsize=20)
        plt.show()
```

- Orbits GEO , SSO , ES-L1 and HEo have higher success rate.

Flight Number vs. Orbit Type

Scatter point of Flight number vs. Orbit type



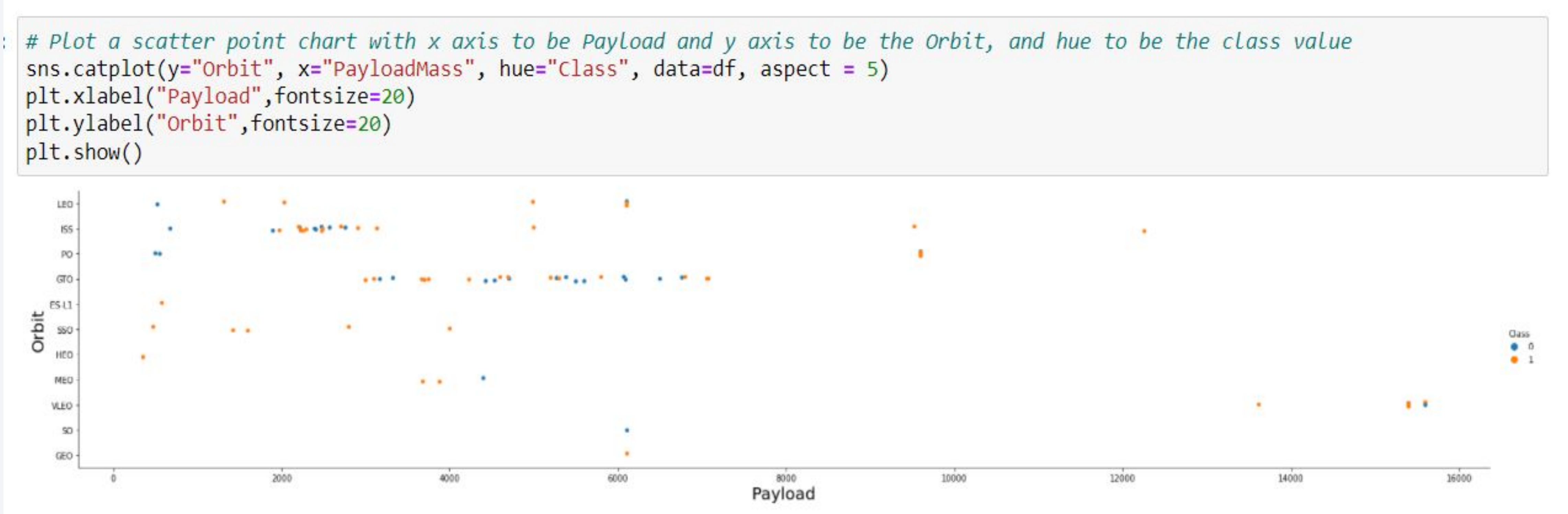
For each orbit, we want to see if there is any relationship between FlightNumber and Orbit type.

```
In [8]: # Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
        sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
        plt.xlabel("FlightNumber",fontsize=20)
        plt.ylabel("Orbit",fontsize=20)
        plt.show()
```

LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

- Scatter point of payload vs. orbit type

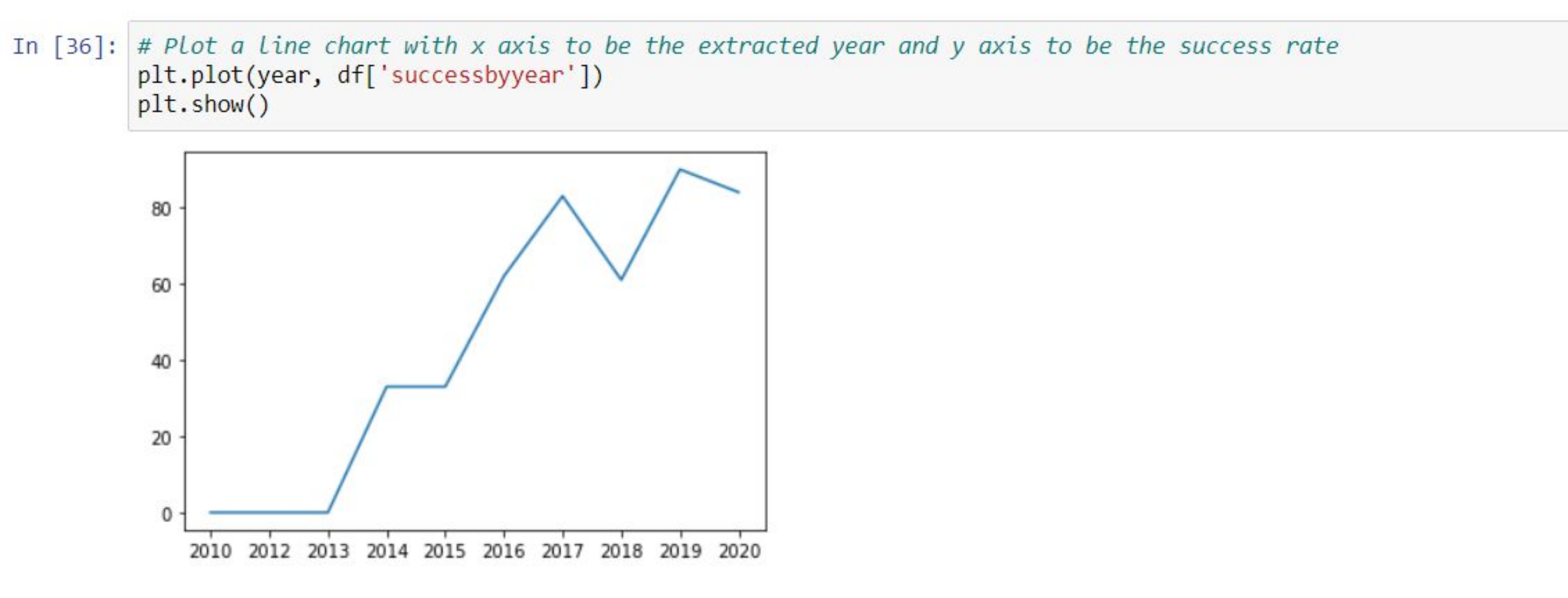


```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Line chart of yearly average success rate

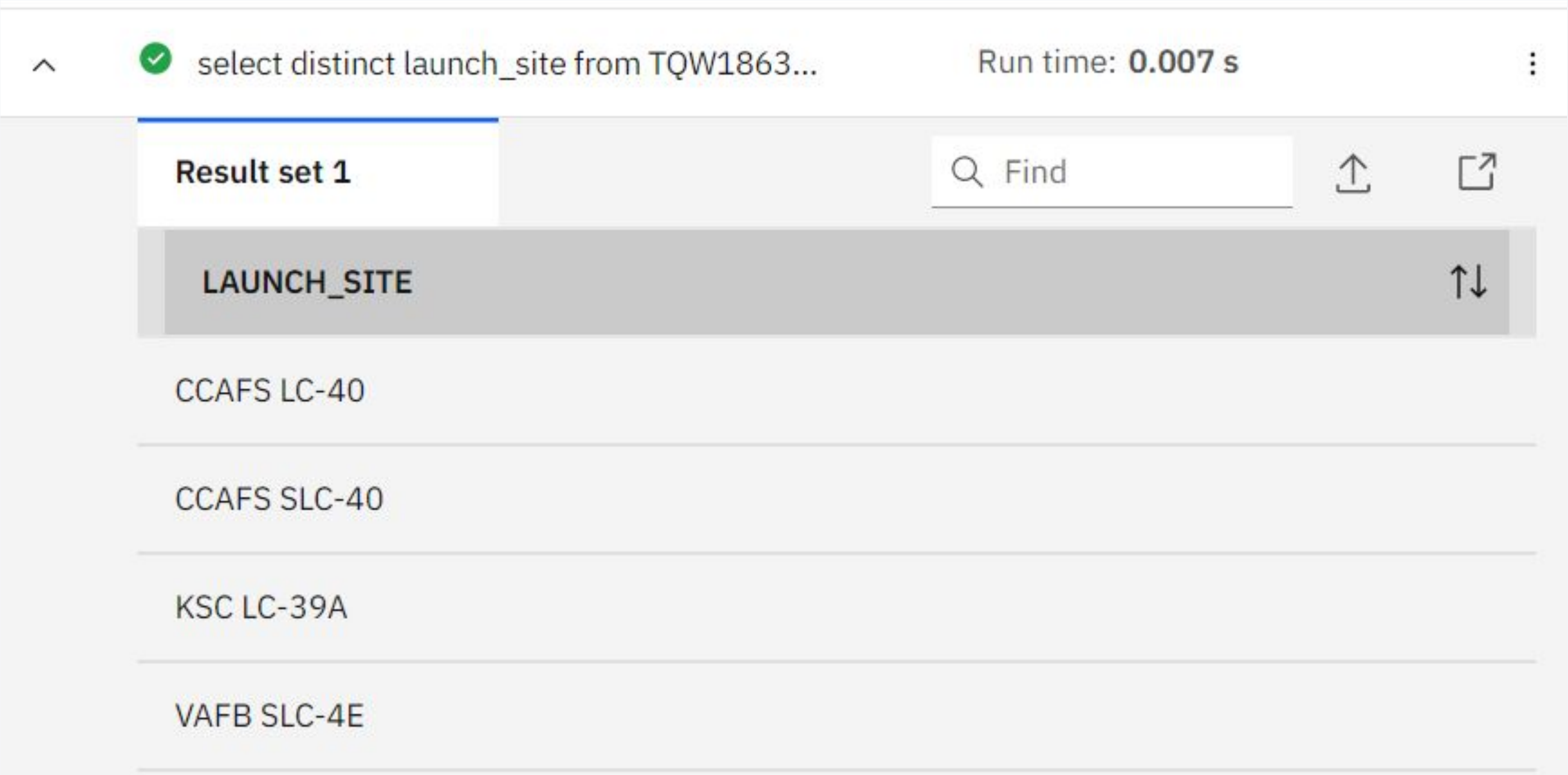


In [36]:
```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
plt.plot(year, df['successbyyear'])
plt.show()
```

Success rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites

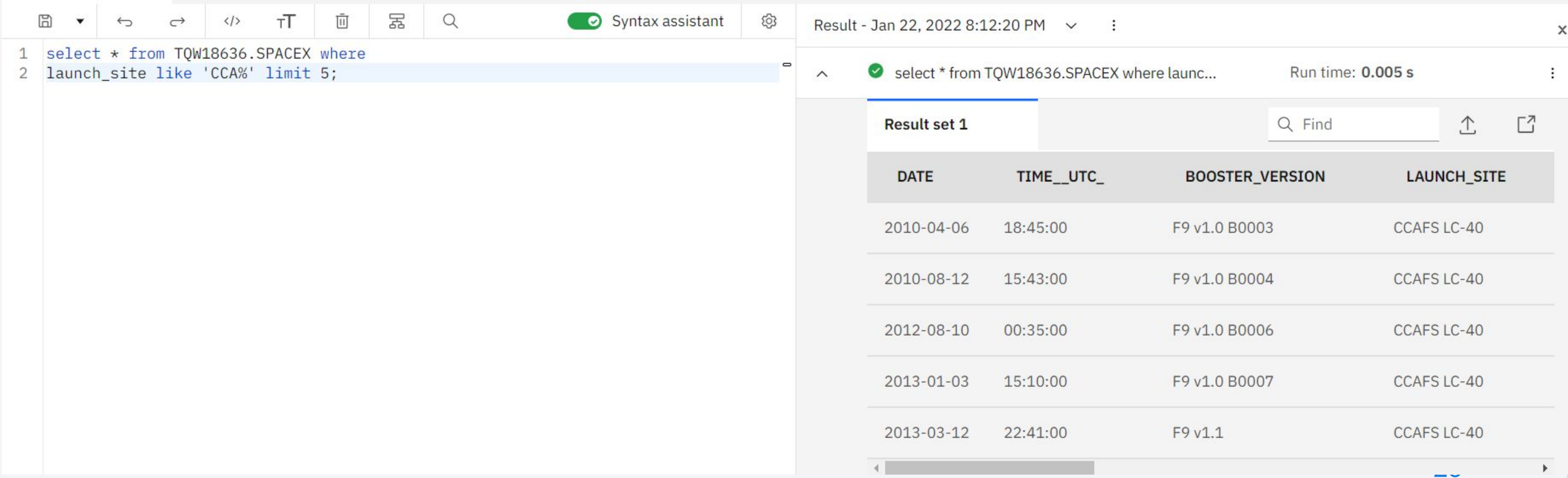


select distinct launch_site from TQW1863... Run time: **0.007 s**

LAUNCH_SITE
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`



```
select * from TQW18636.SPACEX where
launch_site like 'CCA%' limit 5;
```

Result - Jan 22, 2022 8:12:20 PM

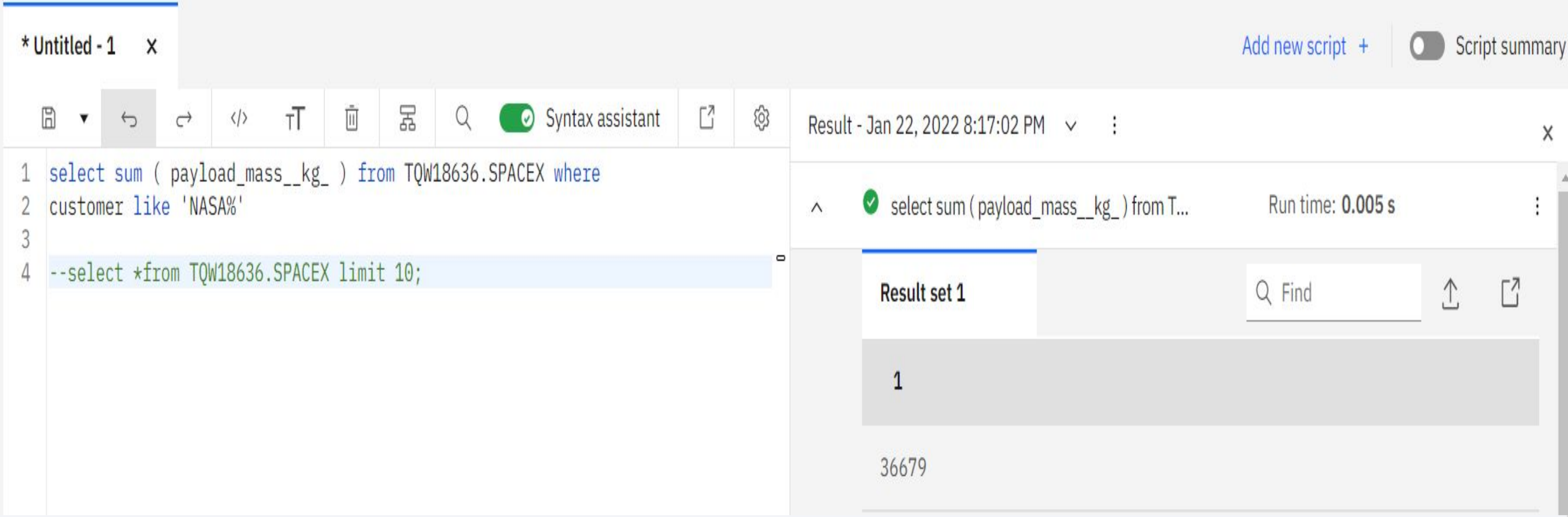
select * from TQW18636.SPACEX where launc... Run time: 0.005 s

Result set 1

DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA



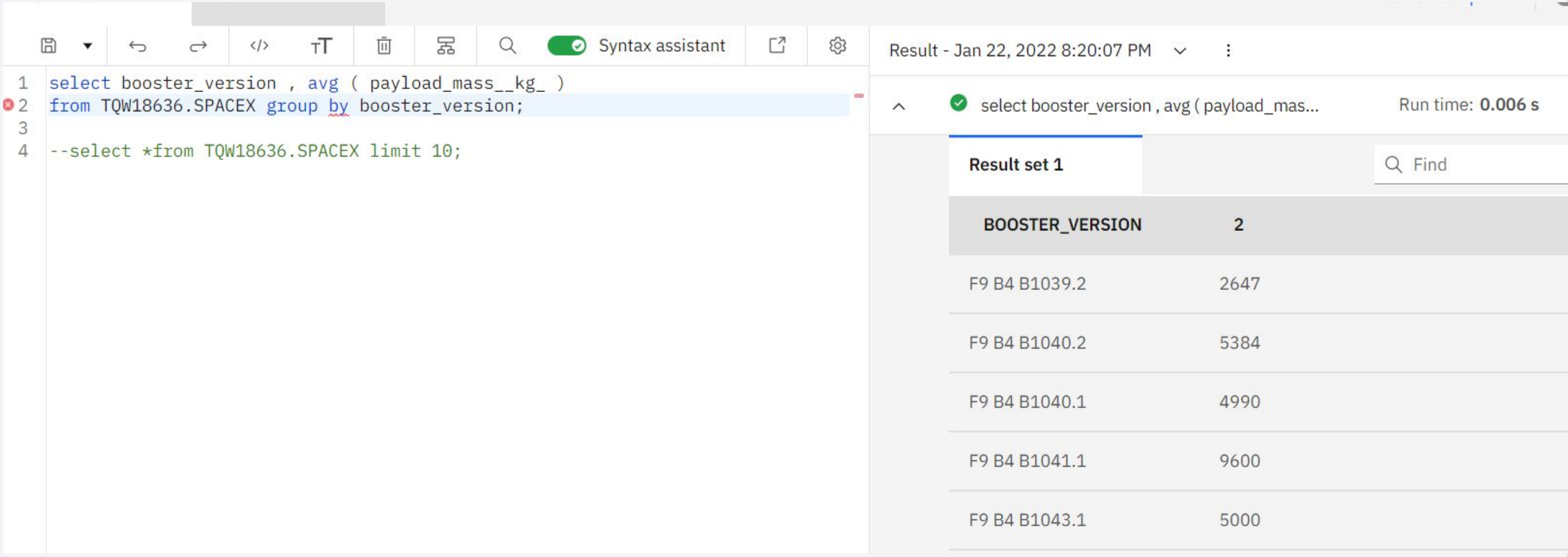
```
select sum ( payload_mass__kg_ ) from TQW18636.SPACEX where
customer like 'NASA%'

--select *from TQW18636.SPACEX limit 10;
```

select sum (payload_mass__kg_) from T...

1
36679

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1



```
select booster_version , avg ( payload_mass__kg_ )
from TQW18636.SPACEX group by booster_version;

--select *from TQW18636.SPACEX limit 10;
```

Result - Jan 22, 2022 8:20:07 PM

select booster_version , avg (payload_mas...

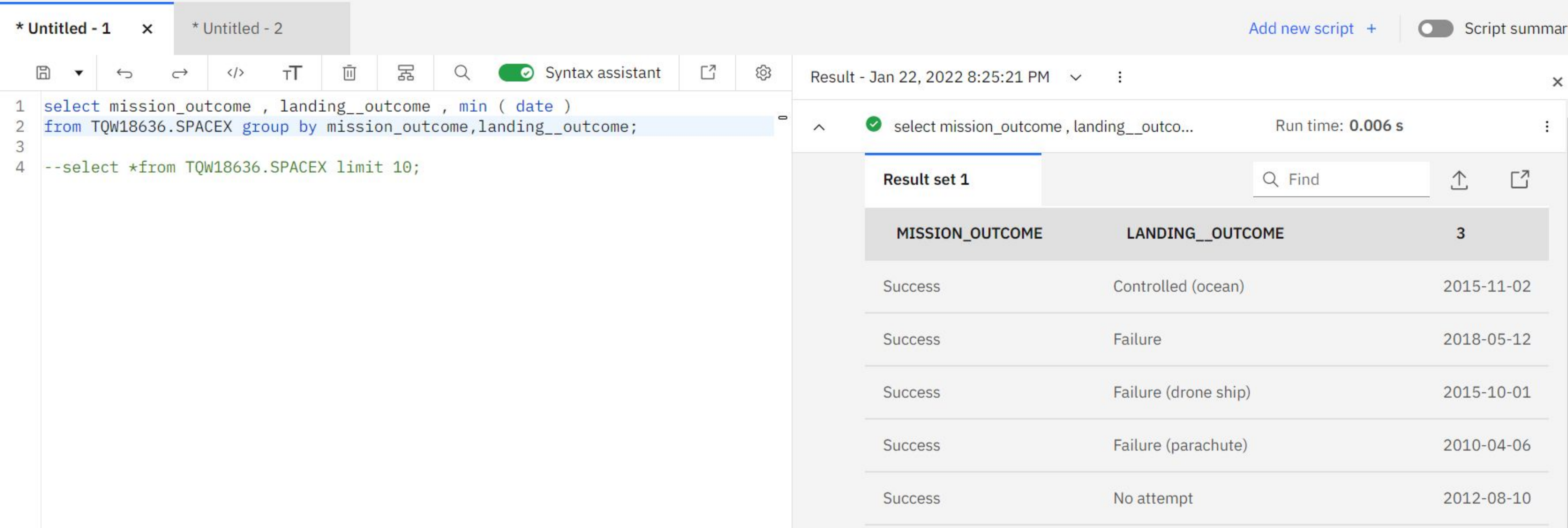
Run time: **0.006 s**

Result set 1

BOOSTER_VERSION	2
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1043.1	5000

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad



```
select mission_outcome , landing__outcome , min ( date )
from TQW18636.SPACEX group by mission_outcome,landing__outcome;

--select *from TQW18636.SPACEX limit 10;
```

Result - Jan 22, 2022 8:25:21 PM

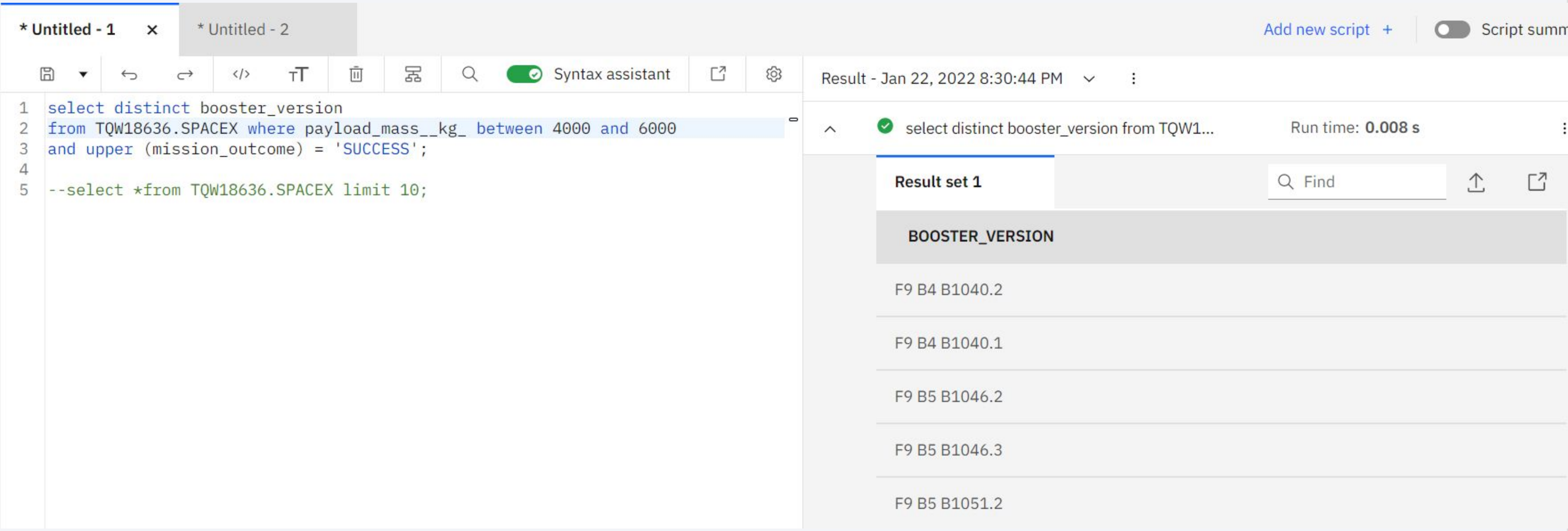
select mission_outcome , landing__outco... Run time: **0.006 s**

Result set 1

MISSION_OUTCOME	LANDING__OUTCOME	3
Success	Controlled (ocean)	2015-11-02
Success	Failure	2018-05-12
Success	Failure (drone ship)	2015-10-01
Success	Failure (parachute)	2010-04-06
Success	No attempt	2012-08-10

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



```
select distinct booster_version
from TQW18636.SPACEX where payload_mass__kg_ between 4000 and 6000
and upper (mission_outcome) = 'SUCCESS';

--select *from TQW18636.SPACEX limit 10;
```

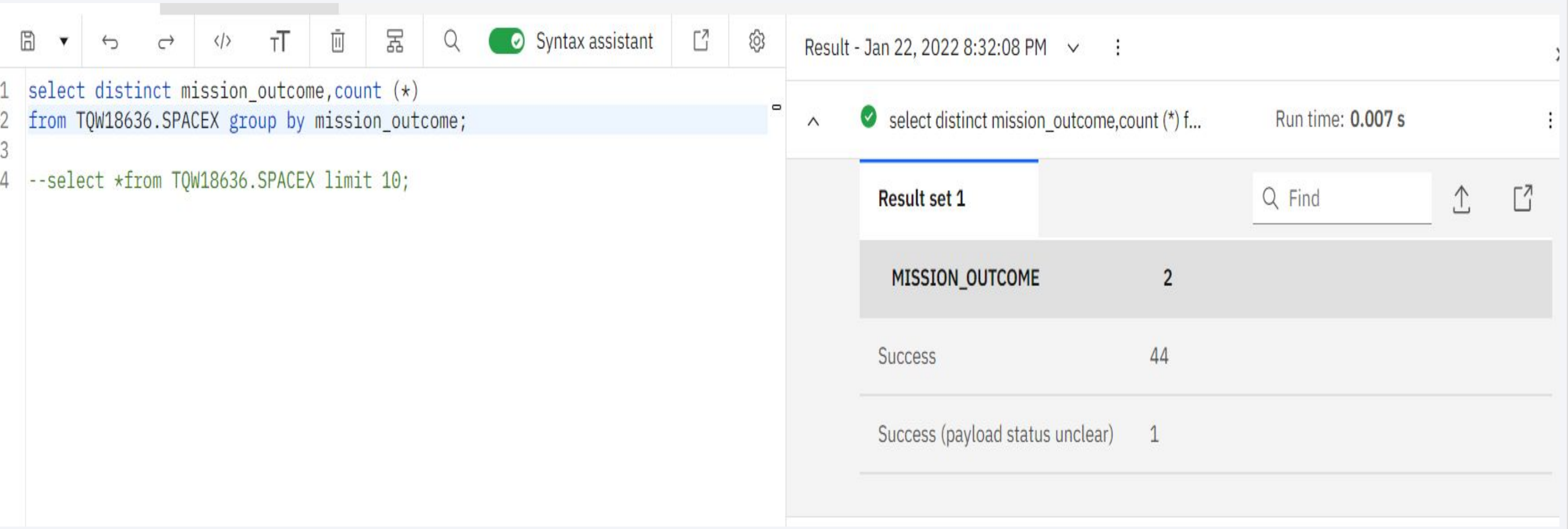
Result - Jan 22, 2022 8:30:44 PM

select distinct booster_version from TQW1... Run time: **0.008 s**

BOOSTER_VERSION
F9 B4 B1040.2
F9 B4 B1040.1
F9 B5 B1046.2
F9 B5 B1046.3
F9 B5 B1051.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



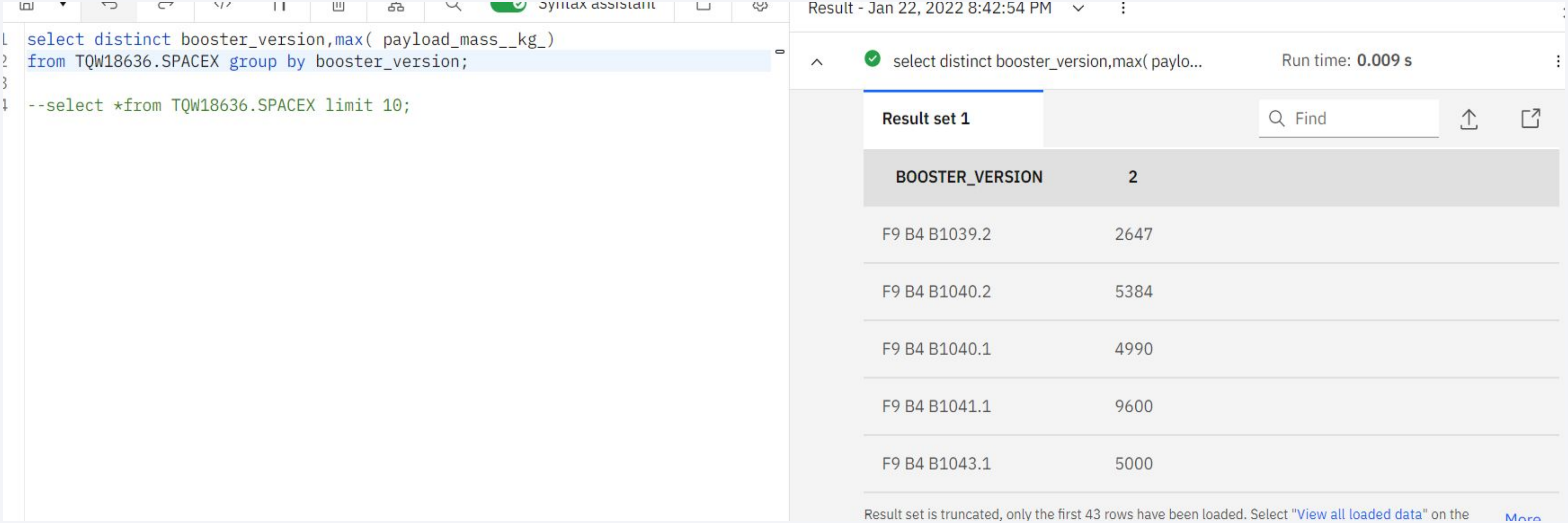
Syntax assistant
Result - Jan 22, 2022 8:32:08 PM
select distinct mission_outcome,count (*) f...
Run time: 0.007 s
Result set 1
Find

```
select distinct mission_outcome,count (*)
from TQW18636.SPACEX group by mission_outcome;

--select *from TQW18636.SPACEX limit 10;
```

MISSION_OUTCOME	2
Success	44
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass



```
select distinct booster_version,max( payload_mass__kg_)
from TQW18636.SPACEX group by booster_version;

--select *from TQW18636.SPACEX limit 10;
```

Result - Jan 22, 2022 8:42:54 PM

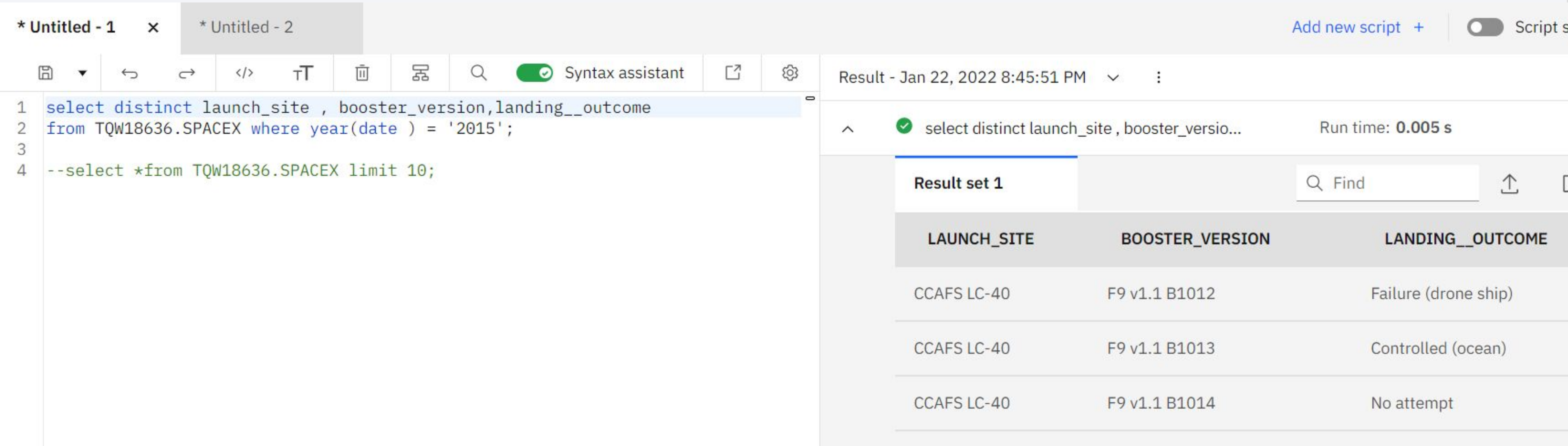
select distinct booster_version,max(paylo...   Run time: **0.009 s**

BOOSTER_VERSION	2
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1043.1	5000

Result set is truncated, only the first 43 rows have been loaded. Select "View all loaded data" on the

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



```
select distinct launch_site , booster_version,landing__outcome
from TQW18636.SPACEX where year(date ) = '2015';

--select *from TQW18636.SPACEX limit 10;
```

Result - Jan 22, 2022 8:45:51 PM

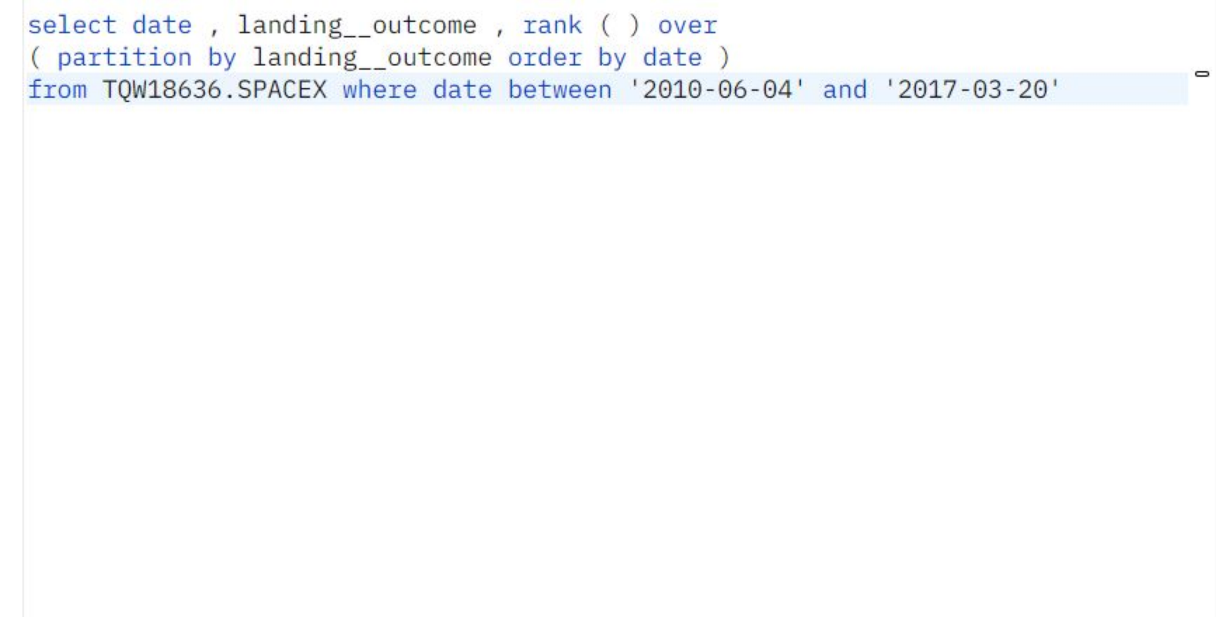
select distinct launch_site , booster_versio... Run time: **0.005 s**

Result set 1

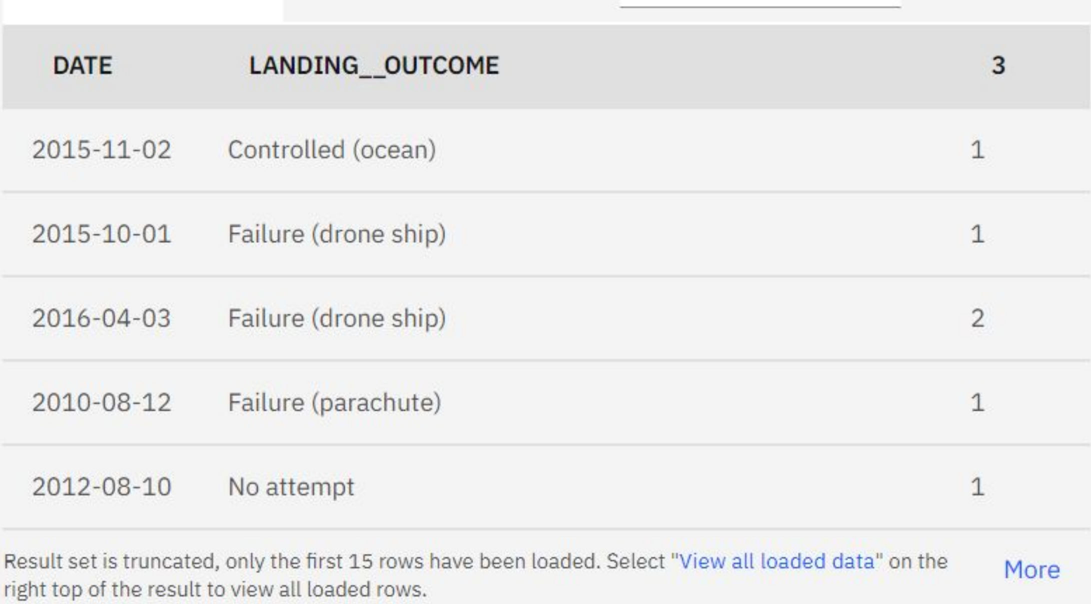
LAUNCH_SITE	BOOSTER_VERSION	LANDING__OUTCOME
CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
CCAFS LC-40	F9 v1.1 B1013	Controlled (ocean)
CCAFS LC-40	F9 v1.1 B1014	No attempt

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order



```
select date , landing__outcome , rank ( ) over
( partition by landing__outcome order by date )
from TQW18636.SPACEX where date between '2010-06-04' and '2017-03-20'
```



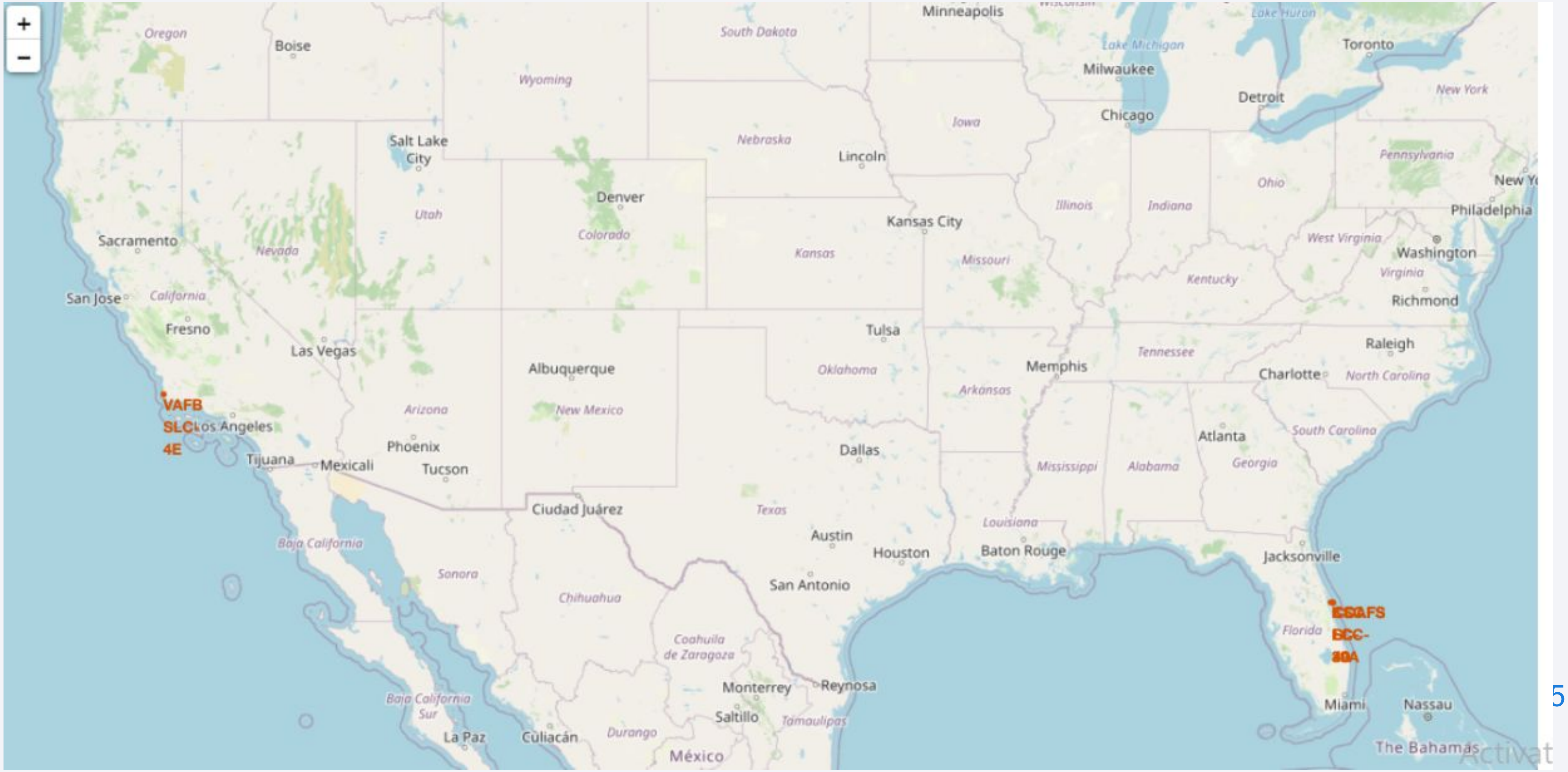
DATE	LANDING__OUTCOME	3
2015-11-02	Controlled (ocean)	1
2015-10-01	Failure (drone ship)	1
2016-04-03	Failure (drone ship)	2
2010-08-12	Failure (parachute)	1
2012-08-10	No attempt	1

Result set is truncated, only the first 15 rows have been loaded. Select "View all loaded data" on the right top of the result to view all loaded rows.

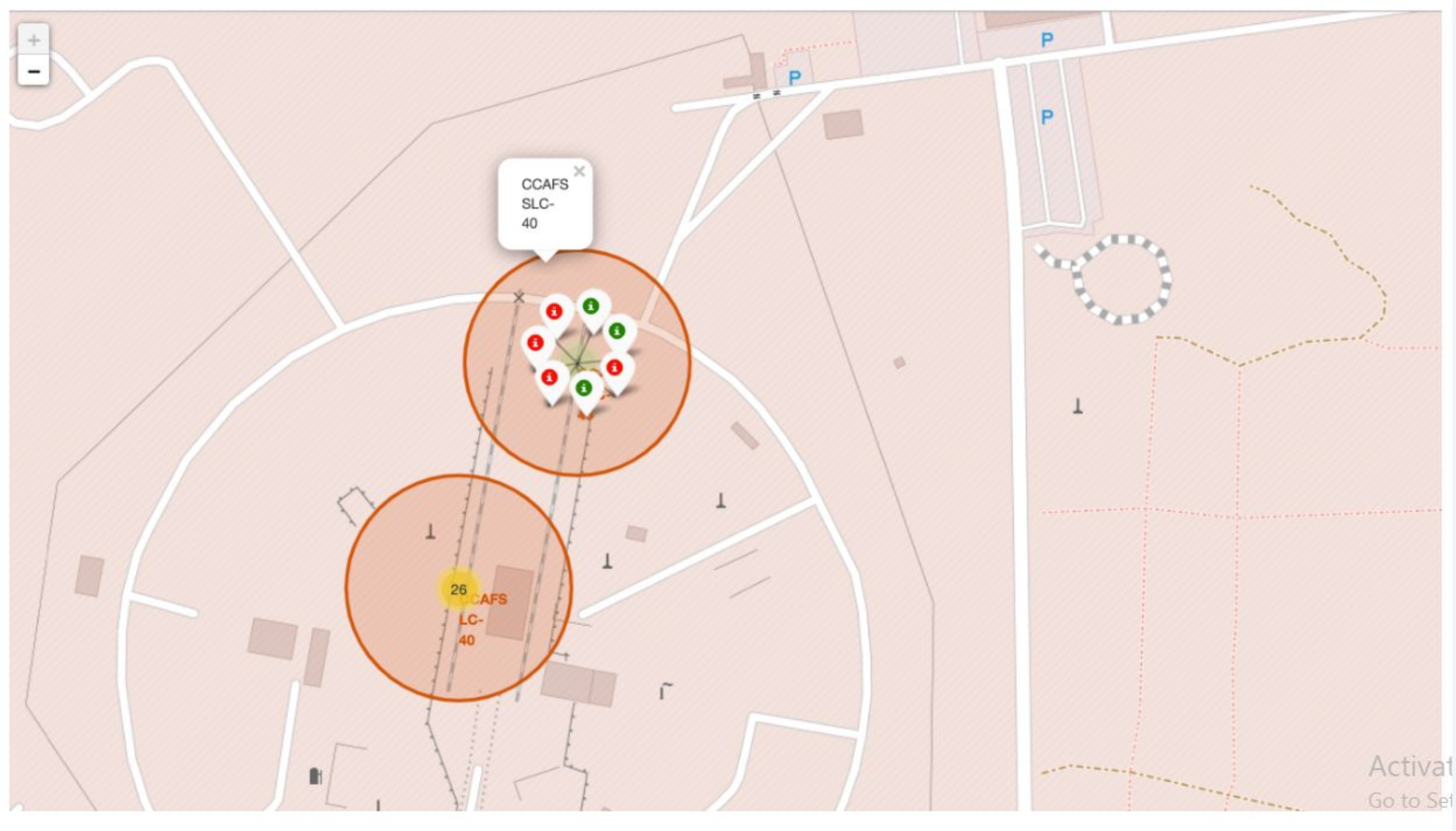
Section 4

Launch Sites Proximities Analysis

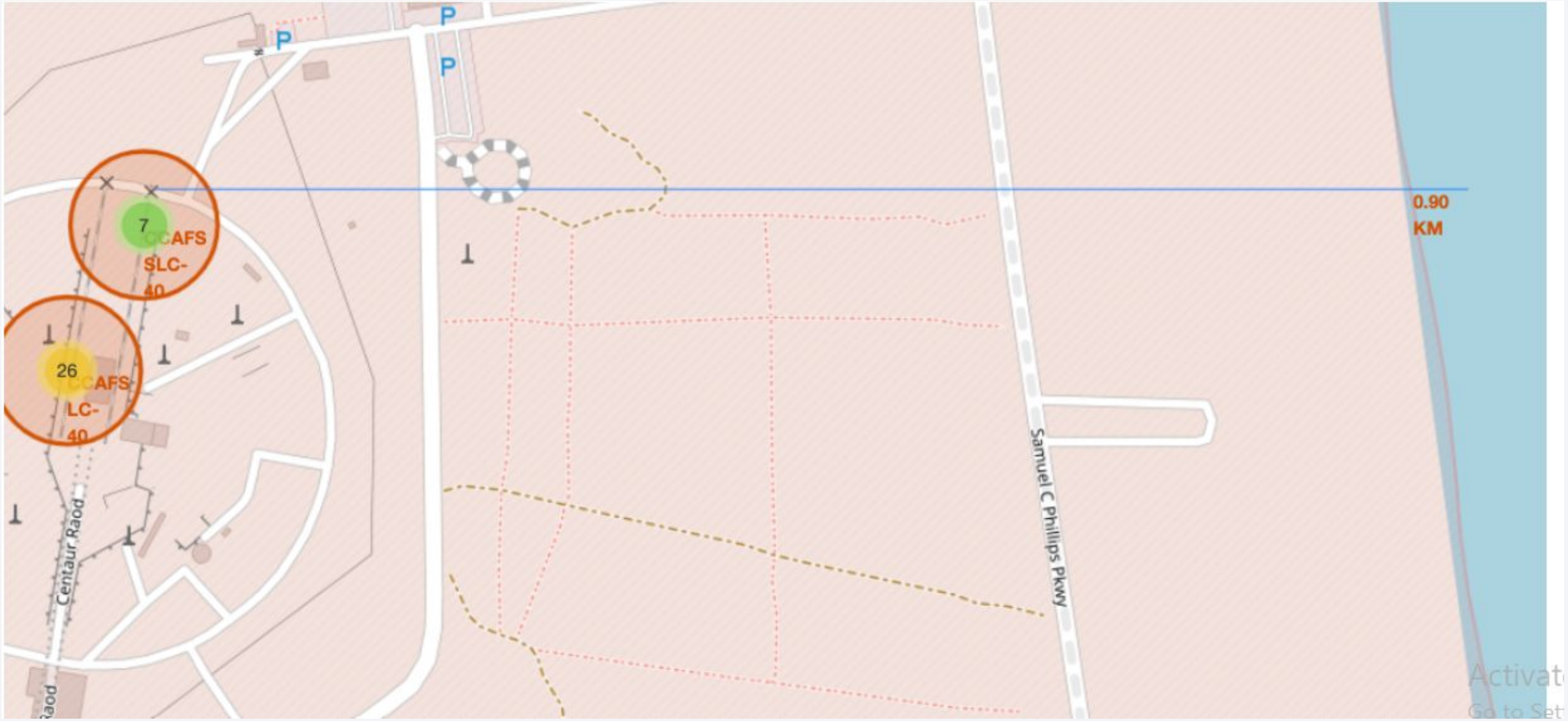
Launch Site Locations



Success/Failure at Launch Sites

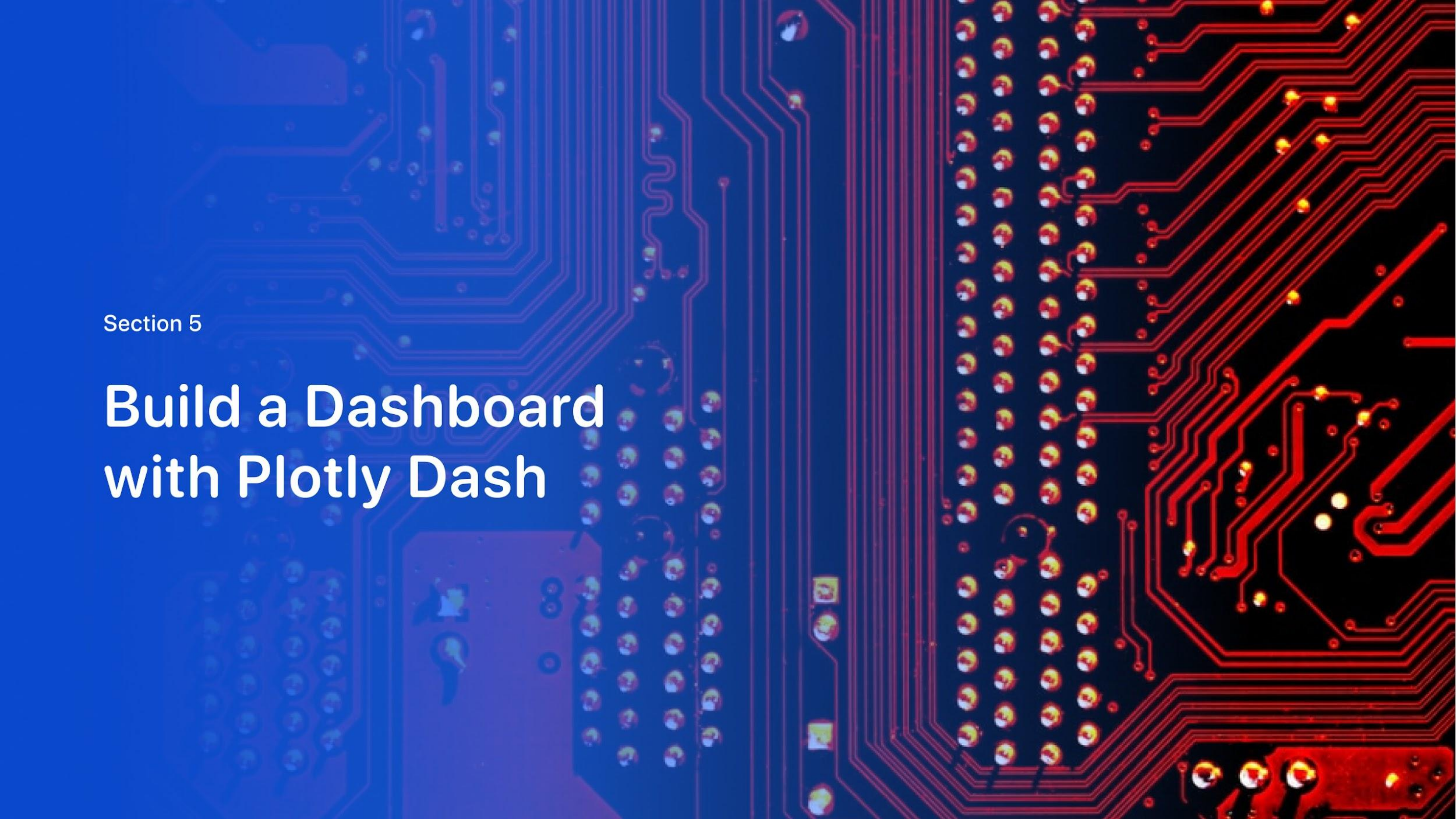


Distance of Proximities from Launch Site

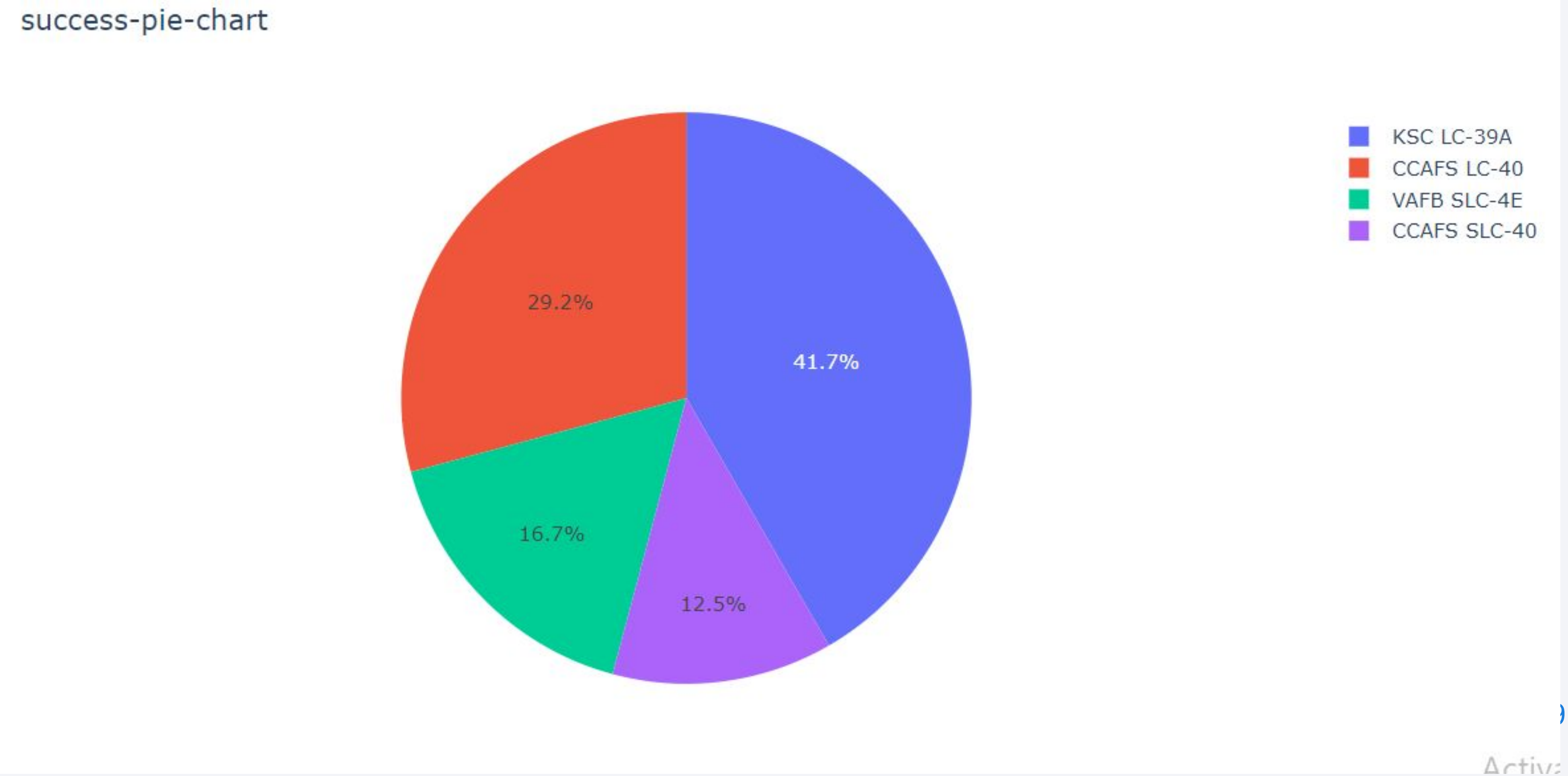


Section 5

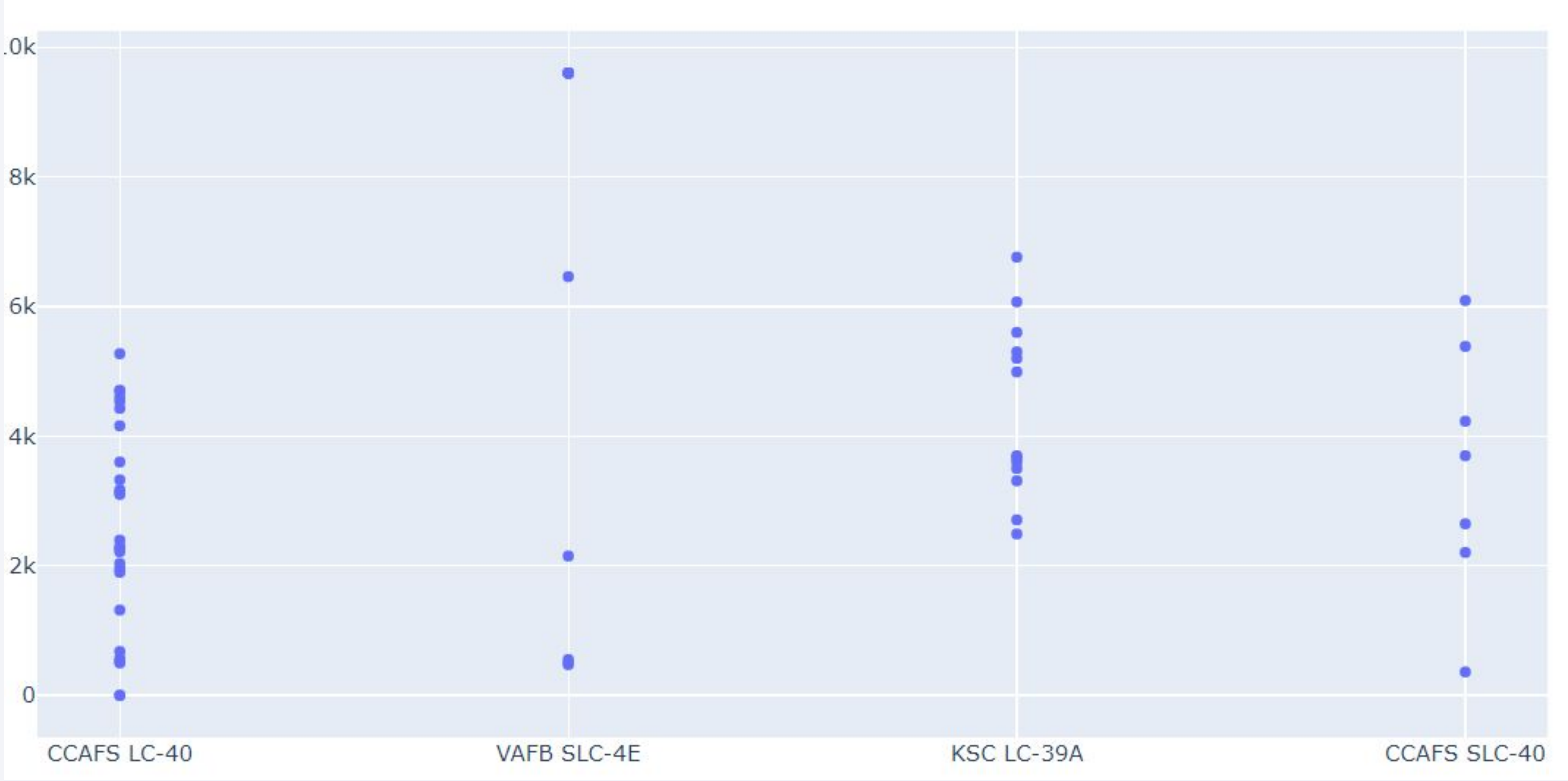
Build a Dashboard with Plotly Dash



Success Pie Chart



<Dashboard Screenshot 2>



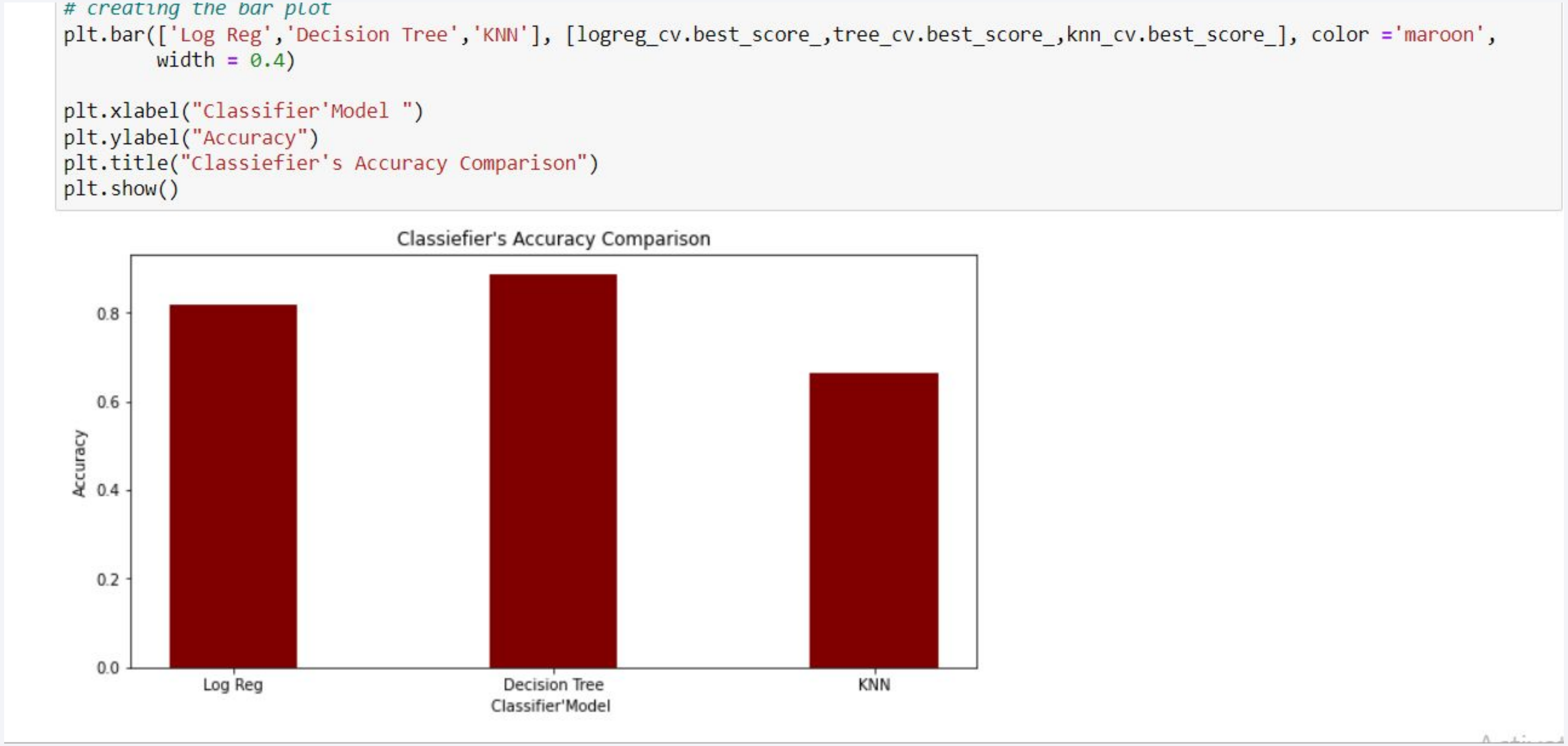
.0k
8k
6k
4k
2k
0
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Section 6

Predictive Analysis (Classification)

Classification Accuracy

Visualize the built model accuracy for all built classification models, in a bar chart

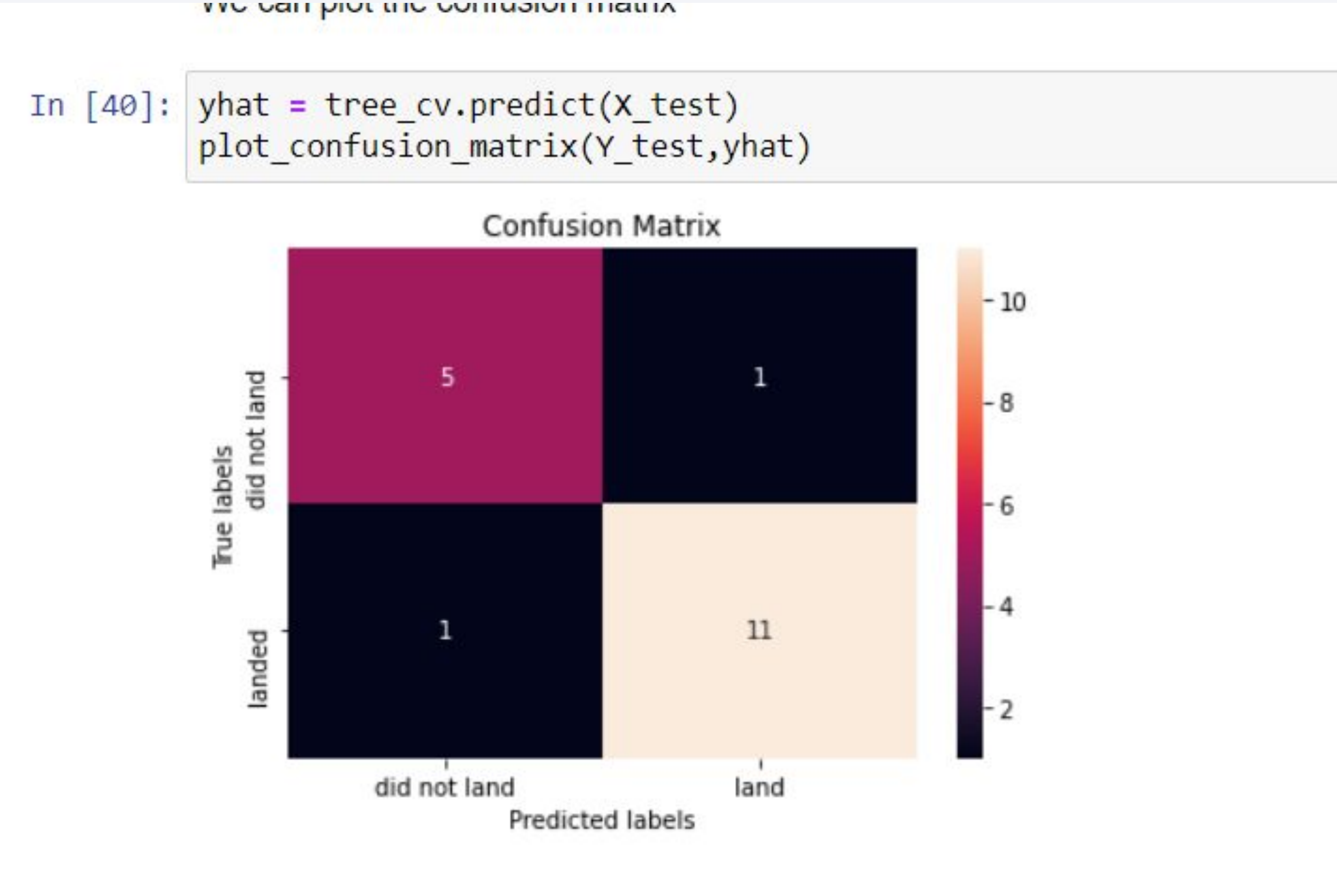


```
# creating the bar plot
plt.bar(['Log Reg','Decision Tree','KNN'], [logreg_cv.best_score_,tree_cv.best_score_,knn_cv.best_score_], color ='maroon',
        width = 0.4)

plt.xlabel("Classifier'Model ")
plt.ylabel("Accuracy")
plt.title("Classiefier's Accuracy Comparison")
plt.show()
```

Decision Tree has the highest Accuracy among all classifier’s models/

Confusion Matrix

- Confusion Matrix for Decision Tree. 1 outcome predicted by model as Landed was actually not landed. Similarly 1 outcome predicted as not landed is actually landed.



We can plot the confusion matrix

```
In [40]: yhat = tree_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```

Conclusions

- Decision Tree performed better than other classifier models in Spacex Data set.
- Its Accuracy is around 89%.
- 1 outcome is False positive and 1 outcome is False Negative out of total 18 outcomes.

Thank you!