Applied Data Science Capstone

Space X

Renato Costa Machado

12/11/2022

SKILLS NETWORK

OUTLINE

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

EXECUTIVE SUMMARY

- Summary of methodologies
  - Data Collection via API, WEB Scraping
  - Exploratory Data Analysis (EDA) with Data Visualization
  - Eda with SQL
  - Interactive Map with Folium
  - Dashboards with Plotly Dash
  - Predictive Analysis
- Summary of results
  - Exploratory Data Analysis results
  - Interactive maps and dashboards
  - Predictive results

INTRODUCTION

The project

The aim of this project is to predict if the Falcon 9 first stage will successfully land. SpaceX says on its website that the Falcon 9 rocket launch cost 62 million dollars. Other providers cost upward of 165 million dollars each. The price difference is explained by the fact that SpacX can reuse the first stage. By determining for another company if it wants to compete with SpaceX for a rocket launch.

Questions to answer

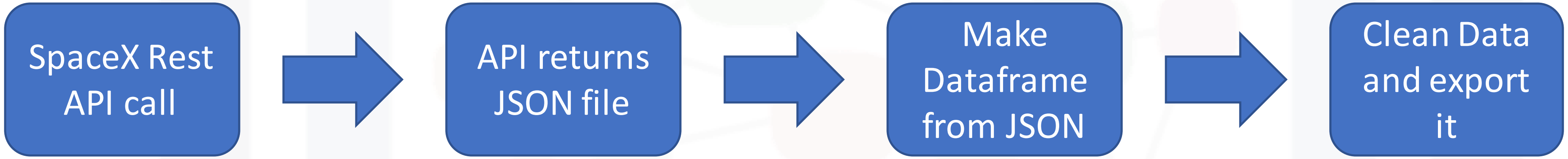
- What are the main characteristics of a successful or failed landing?
- What are the effects of each relationship of the rocket variables on the success or failure of a landing?
- What are the conditions which will allow SpaceX to achieve the best landing success rate?

METHODOLOGY

- Data collection methodology:
  - SpaceX rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

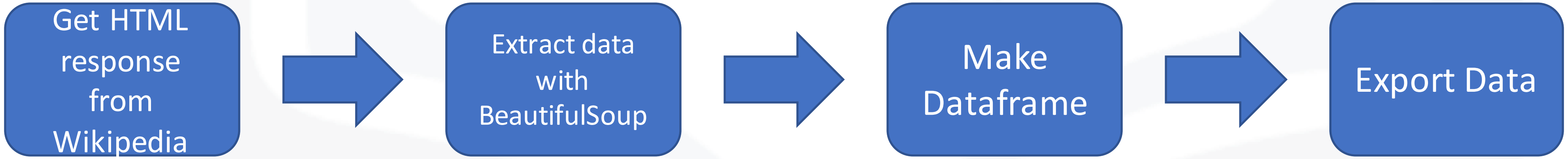
Data Collection

- Datasets are collected from Rest SpaceX API and webscrapping Wikipedia
  - The information obtained by the API are rocket, lunches, payload information.
    - The SpaceX Rest API URL is api.spacexdata.com/v4/



SpaceX Rest API call → API returns JSON file → Make Dataframe from JSON → Clean Data and export it

  - The information obtained by the webscrapping of Wikipedia are launches, landing, payload information.

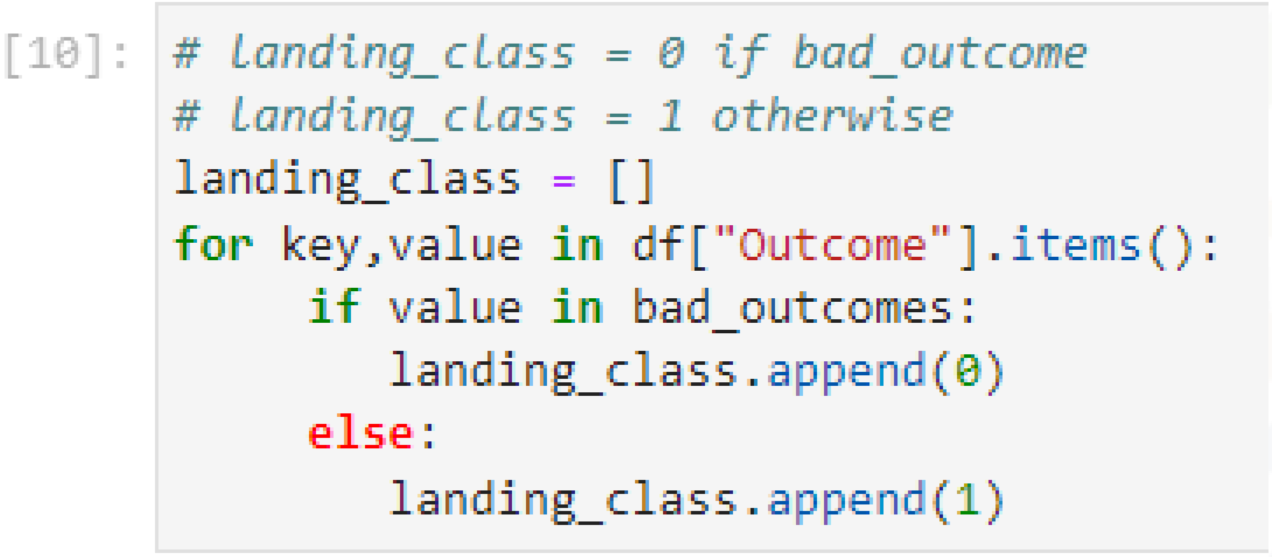


Get HTML response from Wikipedia → Extract data with BeautifulSoup → Make Dataframe → Export Data

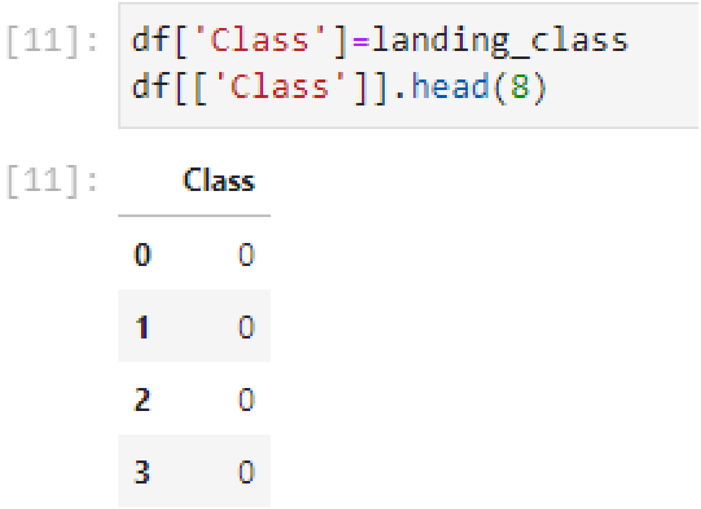
URL is https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Wranglin

- In the dataset, there are several cases where the booster did not land successfully
  - True Ocean, True RTLS, True ASDS means the mission has been successful.
  - False Ocean, False RTLS, False ASDS means the mission was a failure
- We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure



```
[10]: # landing_class = 0 if bad_outcome
      # landing_class = 1 otherwise
      landing_class = []
      for key,value in df["Outcome"].items():
          if value in bad_outcomes:
              landing_class.append(0)
          else:
              landing_class.append(1)
```



```
[11]: df['Class']=landing_class
      df[['Class']].head(8)
```

[11]:

	Class
0	0
1	0
2	0
3	0

EDA with Data Visualization

- Scatter Graphs
  - Flight Number vs Payload Mass
  - Flight Number vs Launch site
  - Payload vs Launch site
  - Orbit vs Flight Number
  - Payload vs Orbit Type
  - Orbit vs Payload Mass

- Bar Graph
  - Success rate vs Orbit

- Line Graph
  - Success rate vs Year

EDA with SQL

We performed SQL queries to gather and understand data from dataset:

- 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)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- 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 total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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

Built an Interactive MAP with Folium

Folium map object is a map centered on NASA Johnson Space Center at Houston, Texas

- Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker).
- Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map.Marker, folium.features.DivIcon).
- The grouping of points in a cluster to display multiple and different information for the same coordinates (folium.plugins.MarkerCluster).
- Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing (folium.map.Marker, folium.Icon).
- Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them (folium.map.Marker, folium.PolyLine, folium.features.DivIcon).

Build a Dashboard with Plotly Dash

Dashboard has dropdown, pie chart, rangeslider and scatter plot components

- Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown)
- Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (plotly.express.pie)
- Rangeslider allows a user to select a payload mass in a fixed range (dah_core_components.RangeSlider)
- Scatter chart shows the retationship between two variables, in particular Success vs Payload Mass (plotly.express.scatter)

Predictive Analysis (Classification)

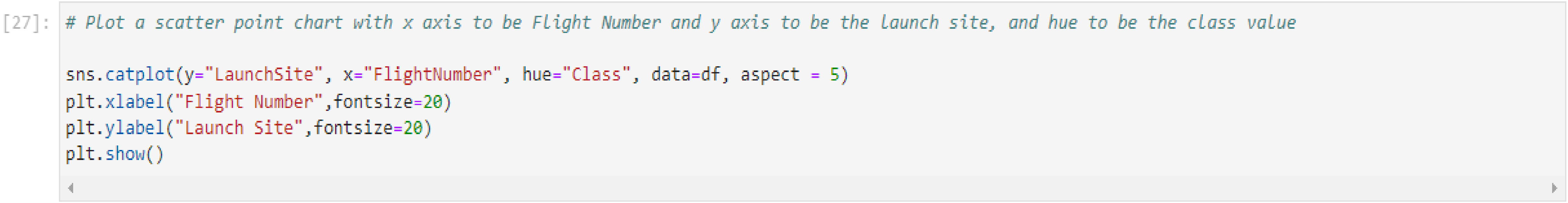
- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model evaluation
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen

SKILLS NETWORK

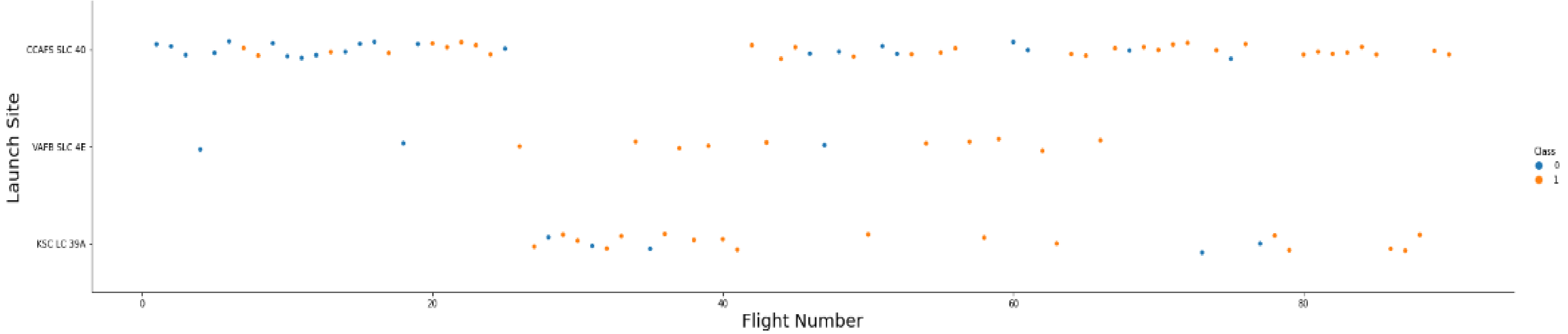
RESULTS

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

Flight Number vs. Launch Site

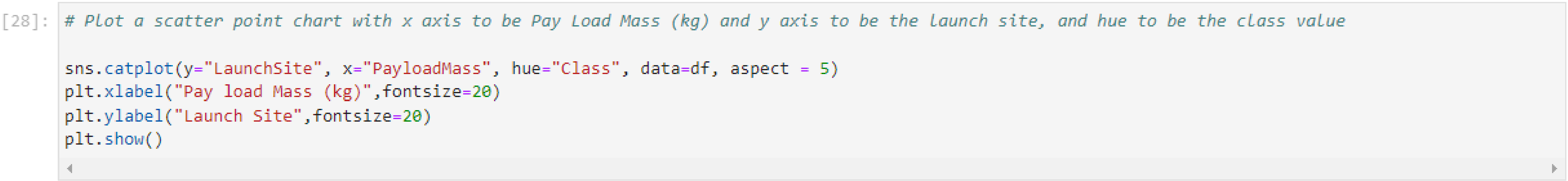


```
[27]: # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value

sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

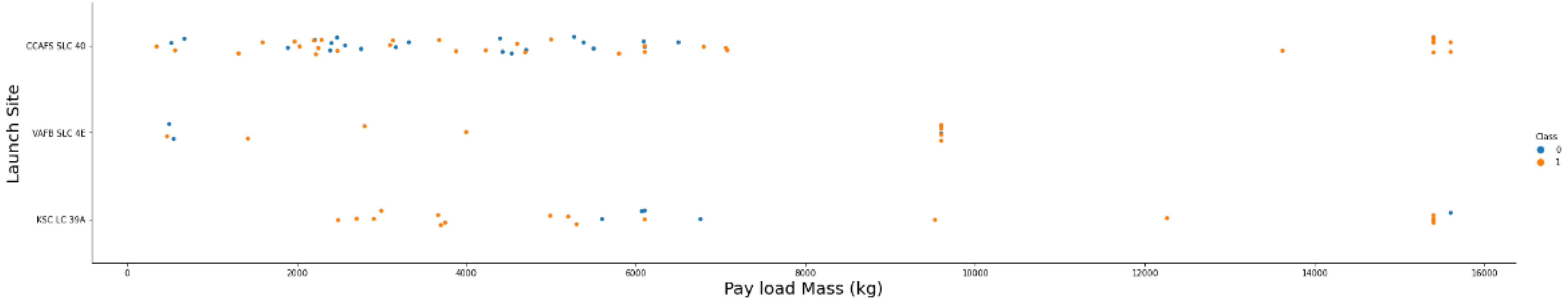


For each site the success rate is increasing

Payload vs. Launch Site

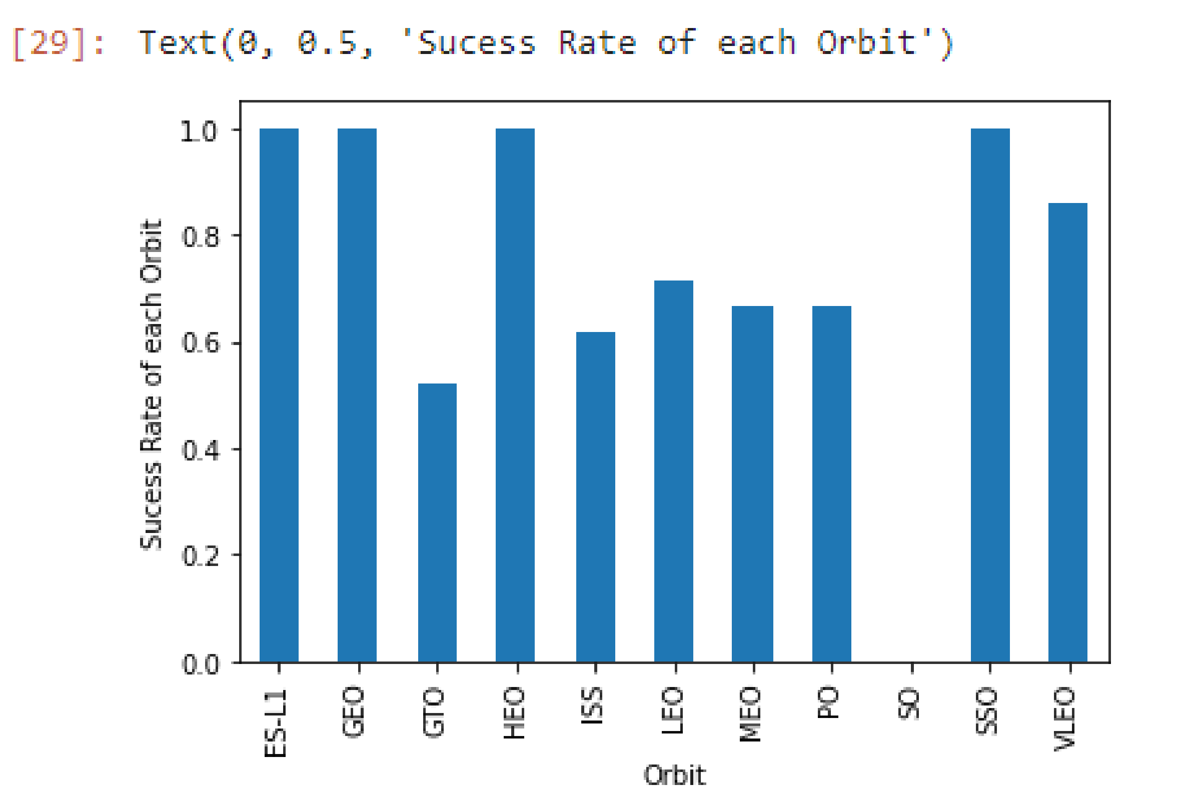


```
[28]: # Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value

      sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
      plt.xlabel("Pay load Mass (kg)",fontsize=20)
      plt.ylabel("Launch Site",fontsize=20)
      plt.show()
```



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

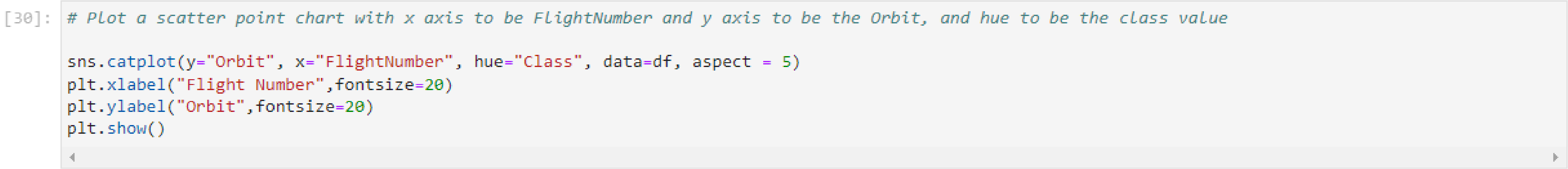
Success Rate vs. Orbit Type



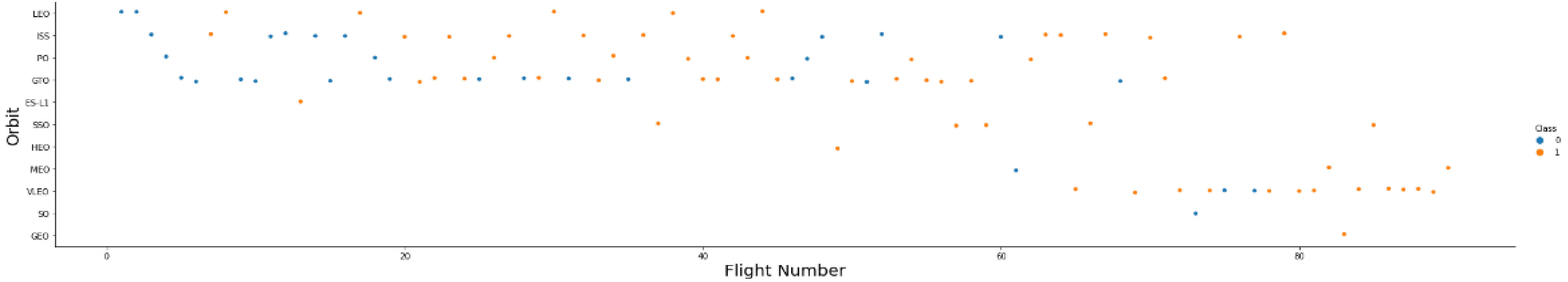
```
[29]: Text(0, 0.5, 'Sucess Rate of each Orbit')
```

ES-L1, GEO, HEO, and SSO have the best success rate

Flight Number vs. Orbit Type



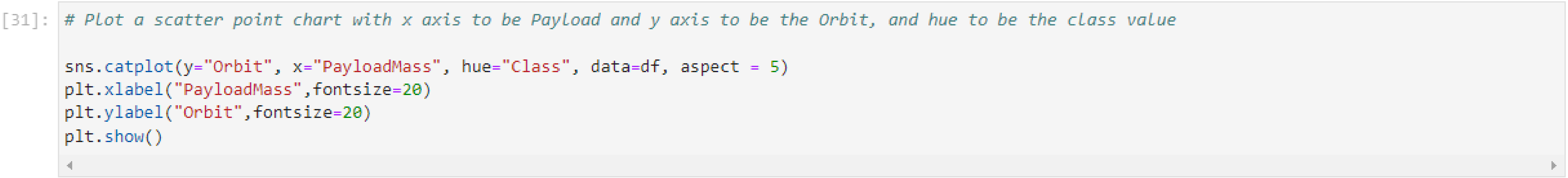
```
[30]: # Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value

      sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
      plt.xlabel("Flight Number",fontsize=20)
      plt.ylabel("Orbit",fontsize=20)
      plt.show()
```



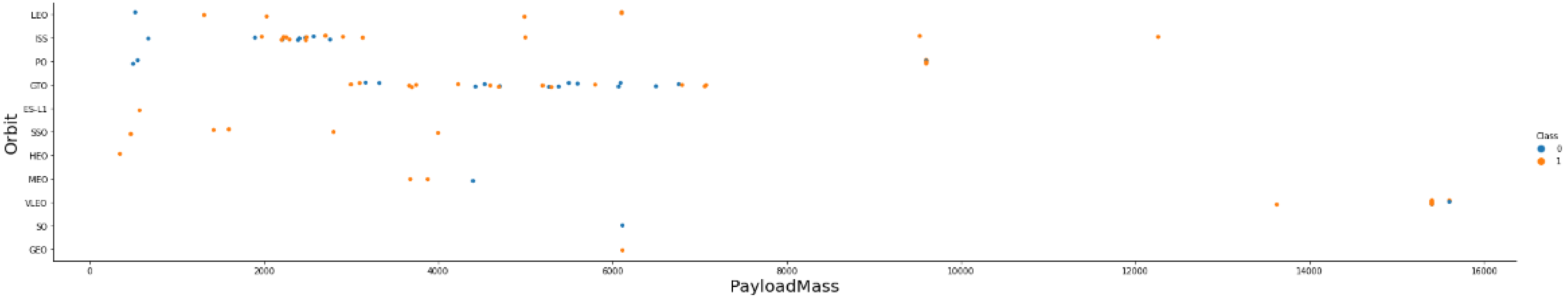
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

SKILLS NETWORK

Payload vs. Orbit Type

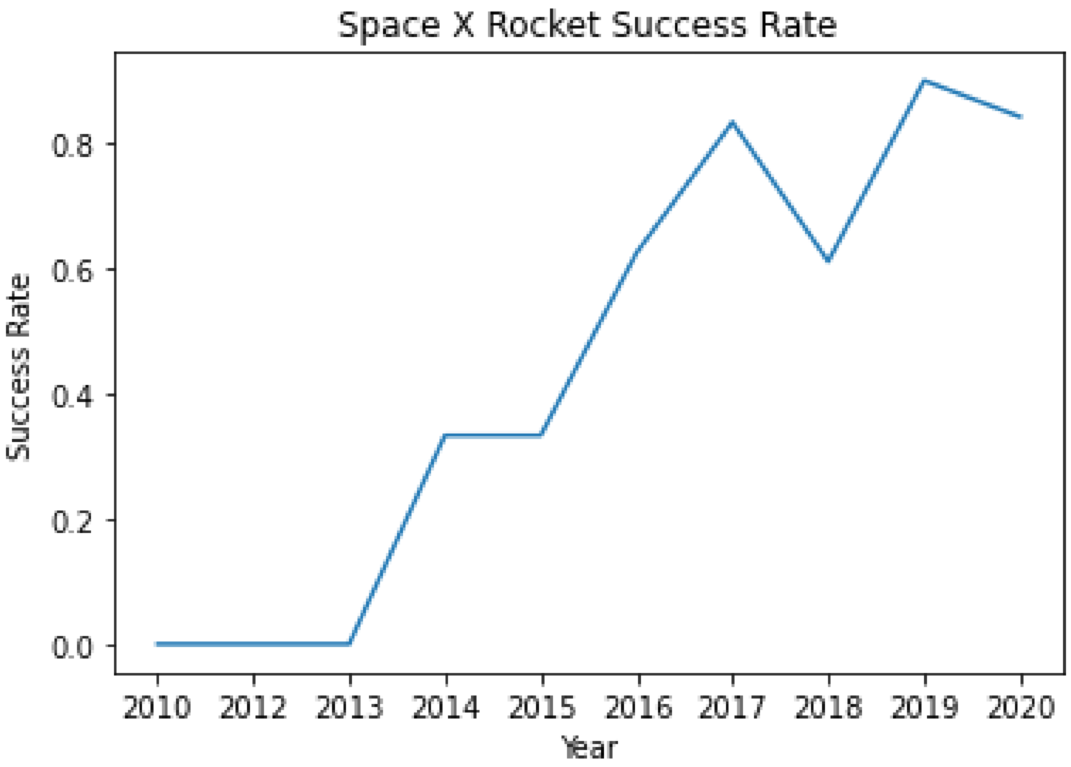


```
[31]: # Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value

      sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
      plt.xlabel("PayloadMass",fontsize=20)
      plt.ylabel("Orbit",fontsize=20)
      plt.show()
```



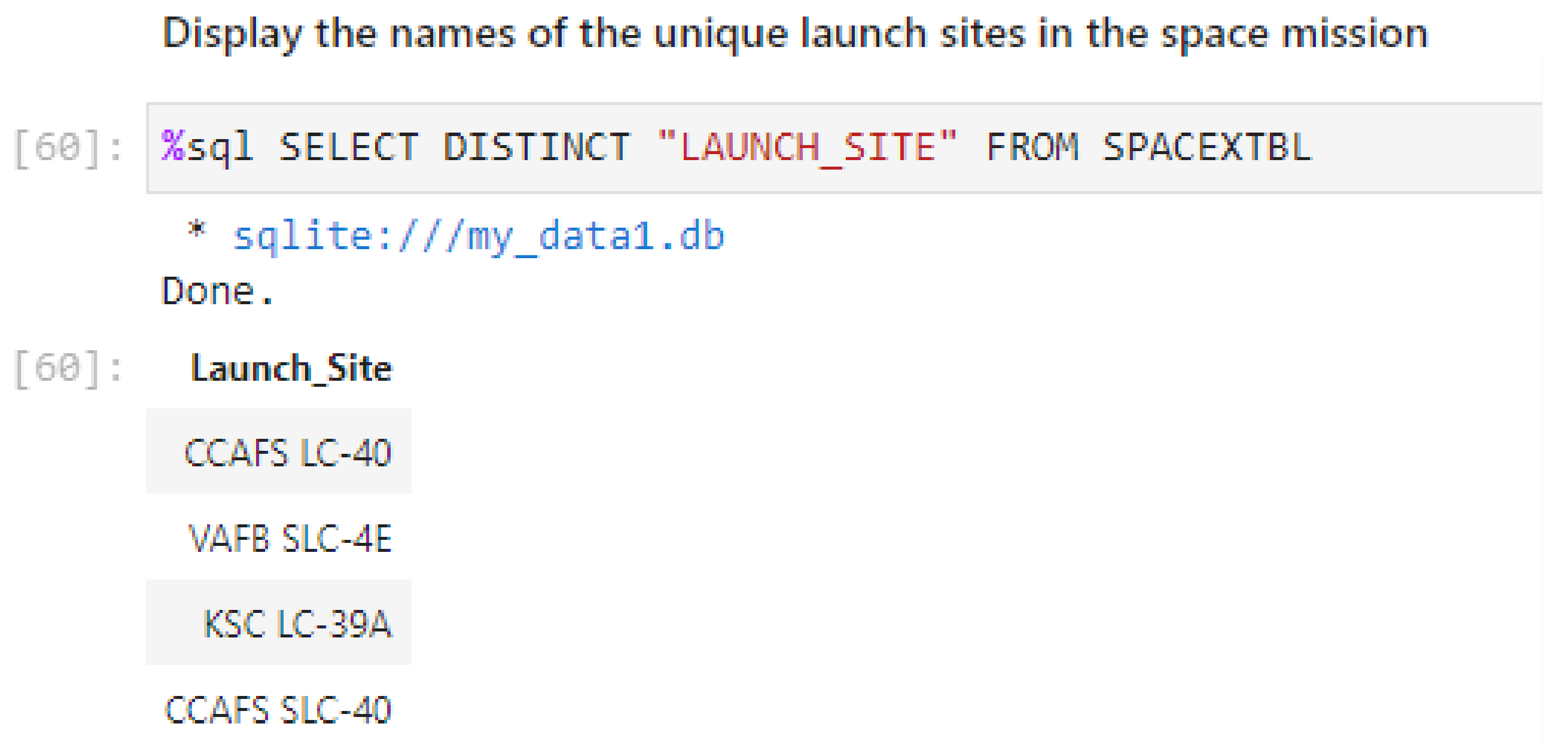
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



We can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names



Display the names of the unique launch sites in the space mission

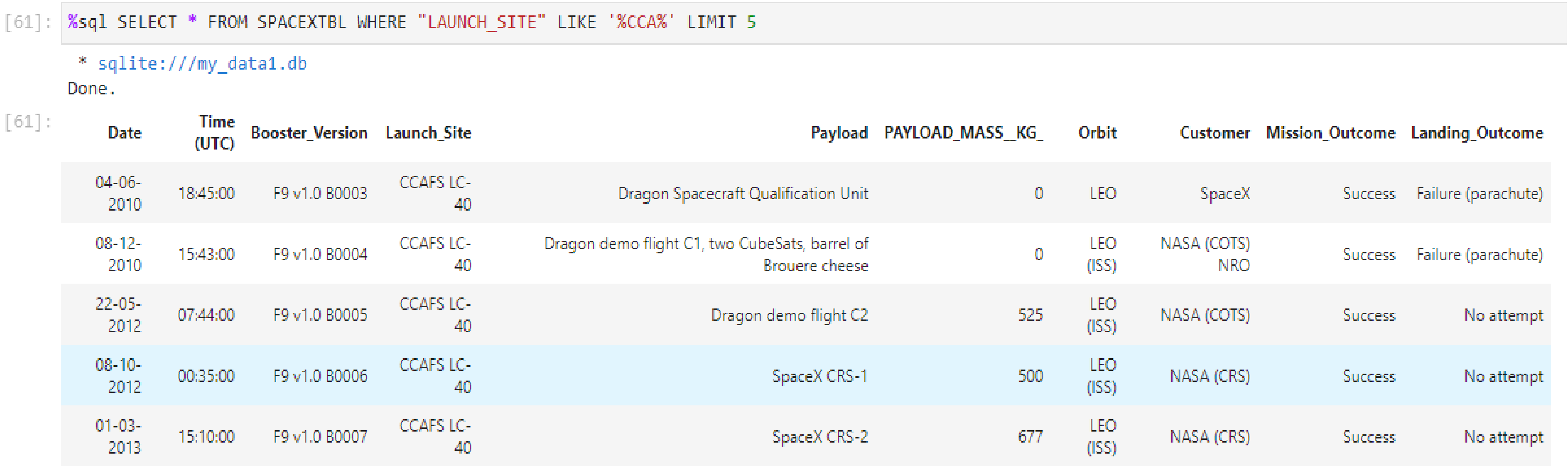
```
[60]: %sql SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

[60]:

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

Launch Site Names Begin with 'CCA'



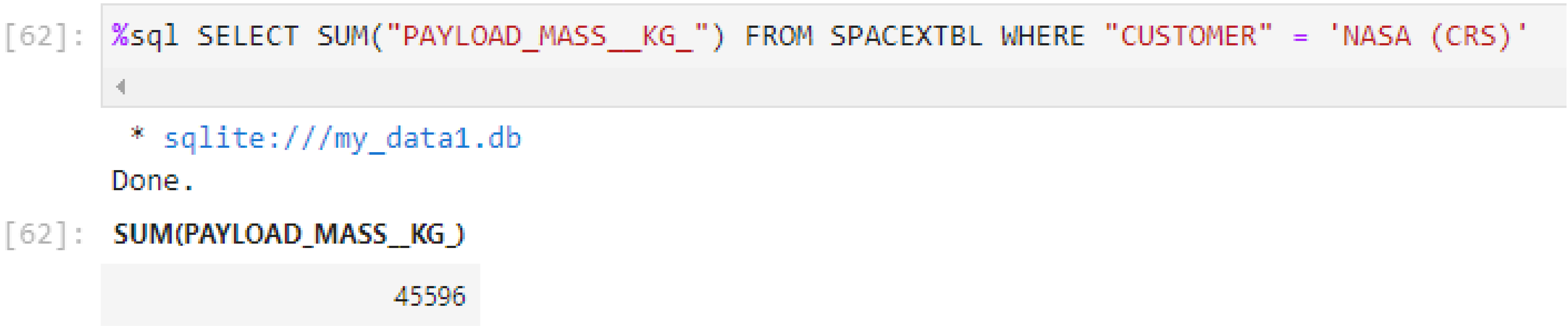
```
[61]: %sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

[61]:

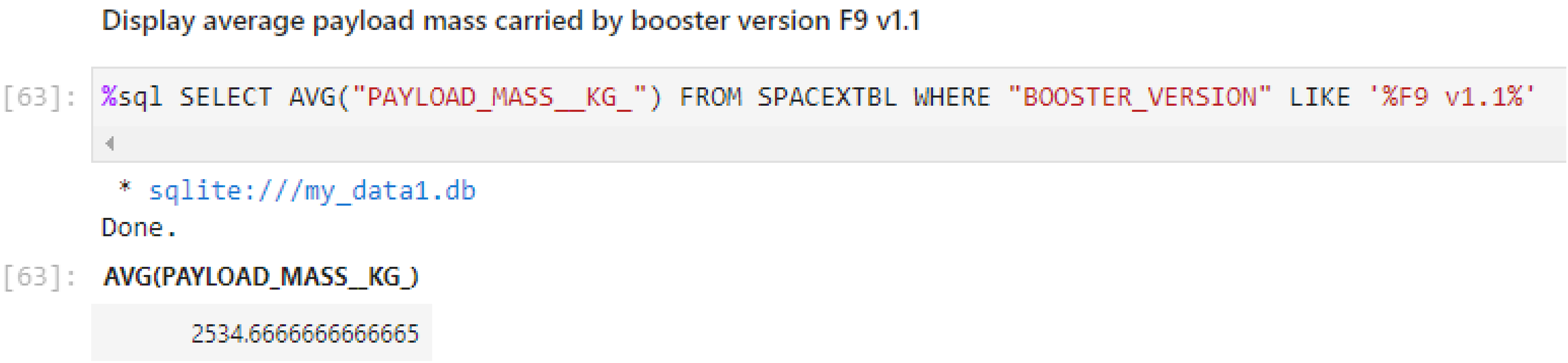
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass



```
[62]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.

[62]: SUM(PAYLOAD_MASS__KG_)
                      45596
```

Average Payload Mass by F9 v1.1



Display average payload mass carried by booster version F9 v1.1

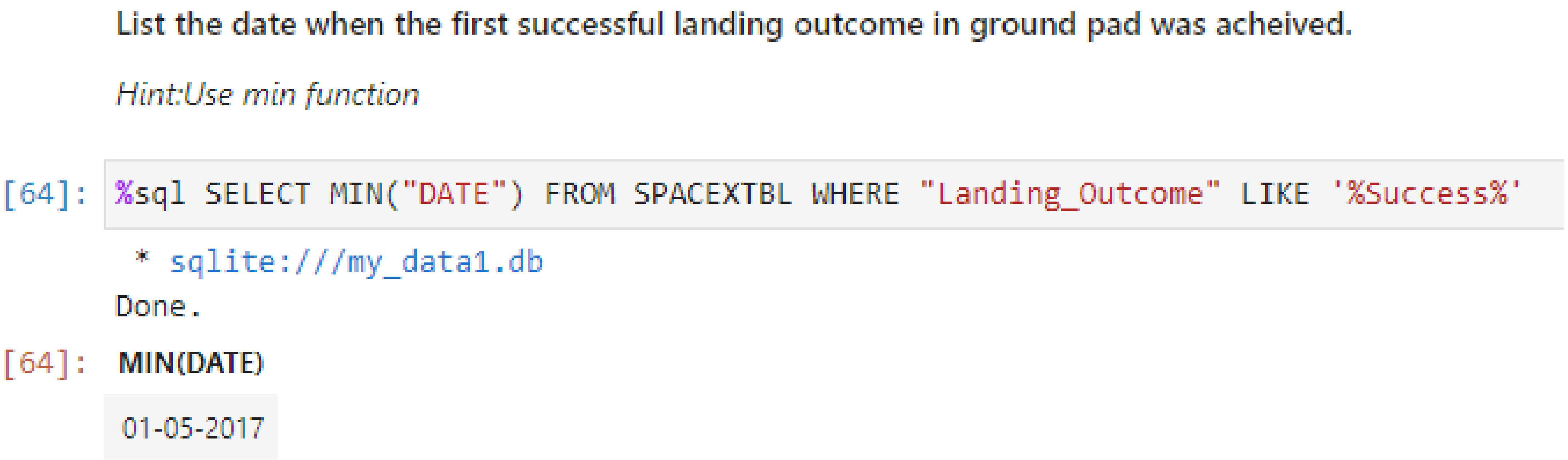
```
[63]: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

[63]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date



List the date when the first successful landing outcome in ground pad was acheived.

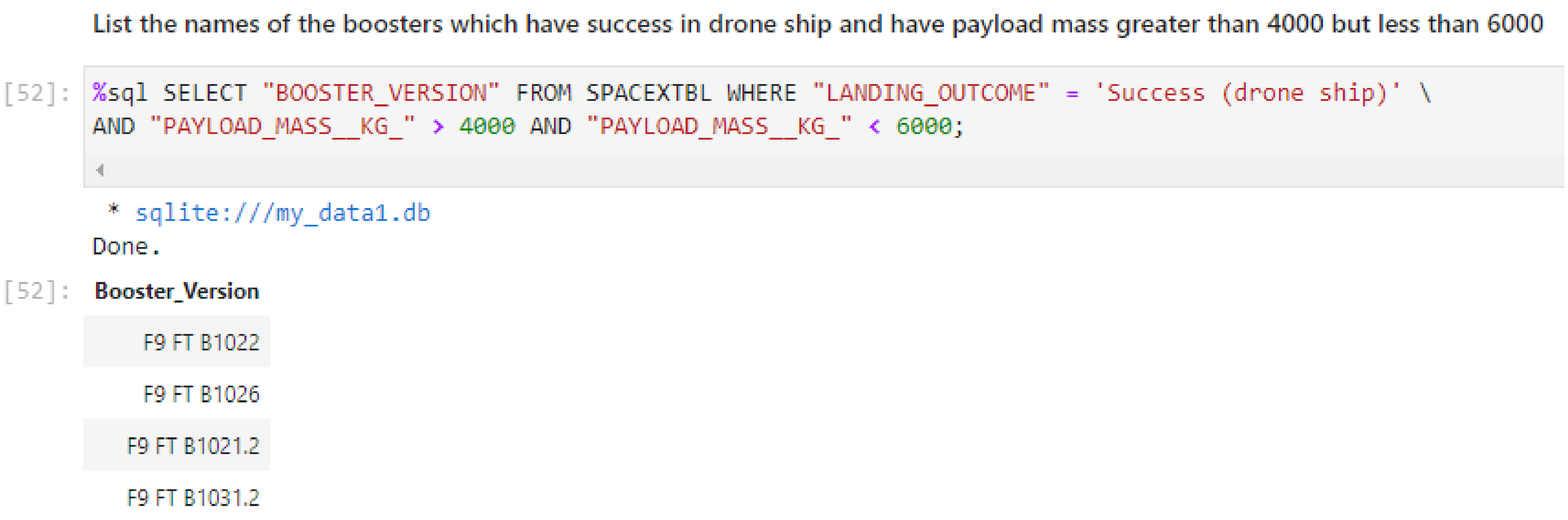
Hint:Use min function

```
[64]: %sql SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Success%'

 * sqlite:///my_data1.db
Done.
```

[64]:

MIN(DATE)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000



List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

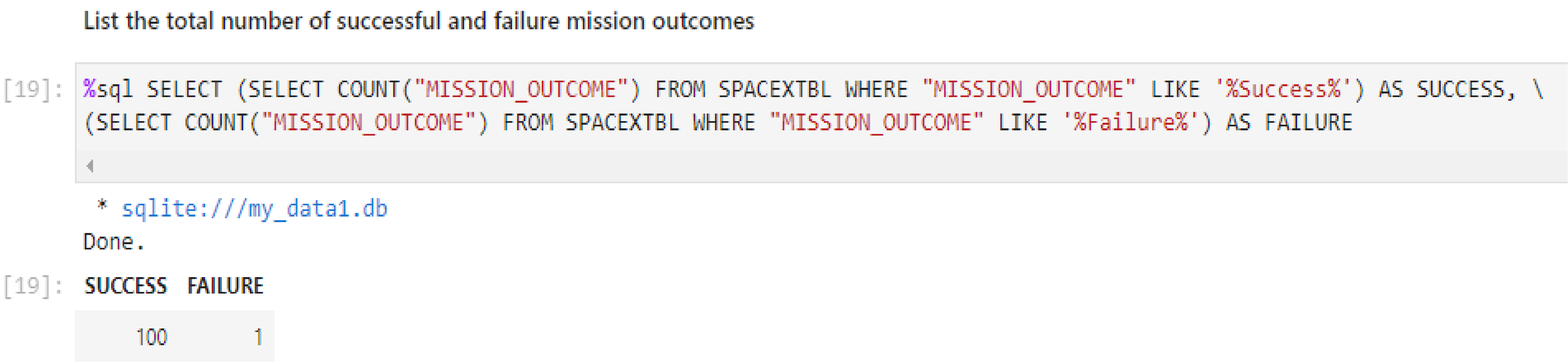
[52]:
```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING_OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

[52]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

[19]:

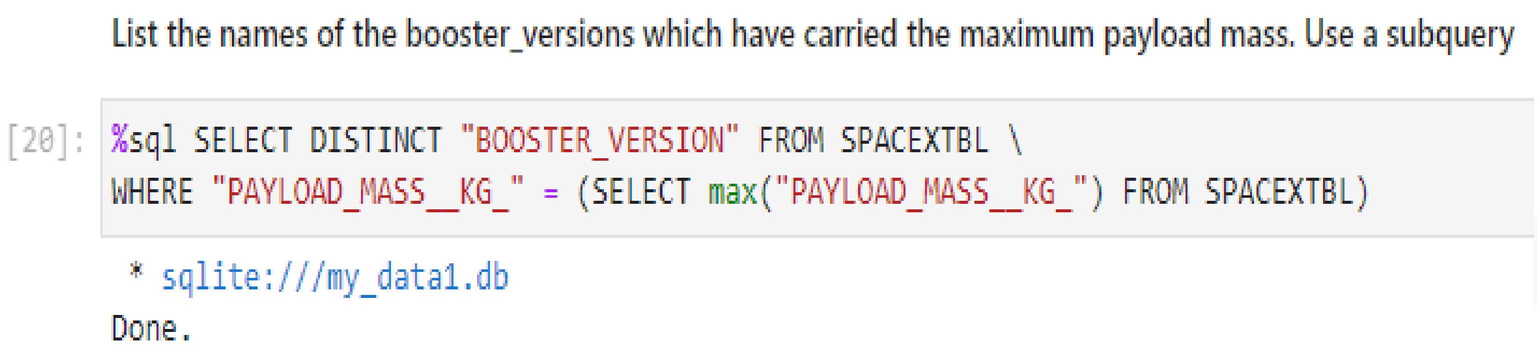
```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

```
 * sqlite:///my_data1.db
Done.
```

[19]:

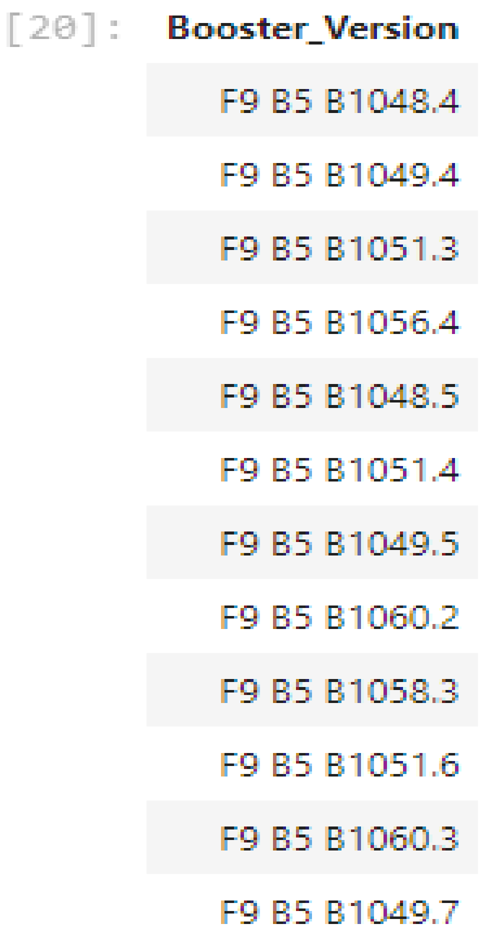
SUCCESS	FAILURE
100	1

Booster Carried Maximum Payload



List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

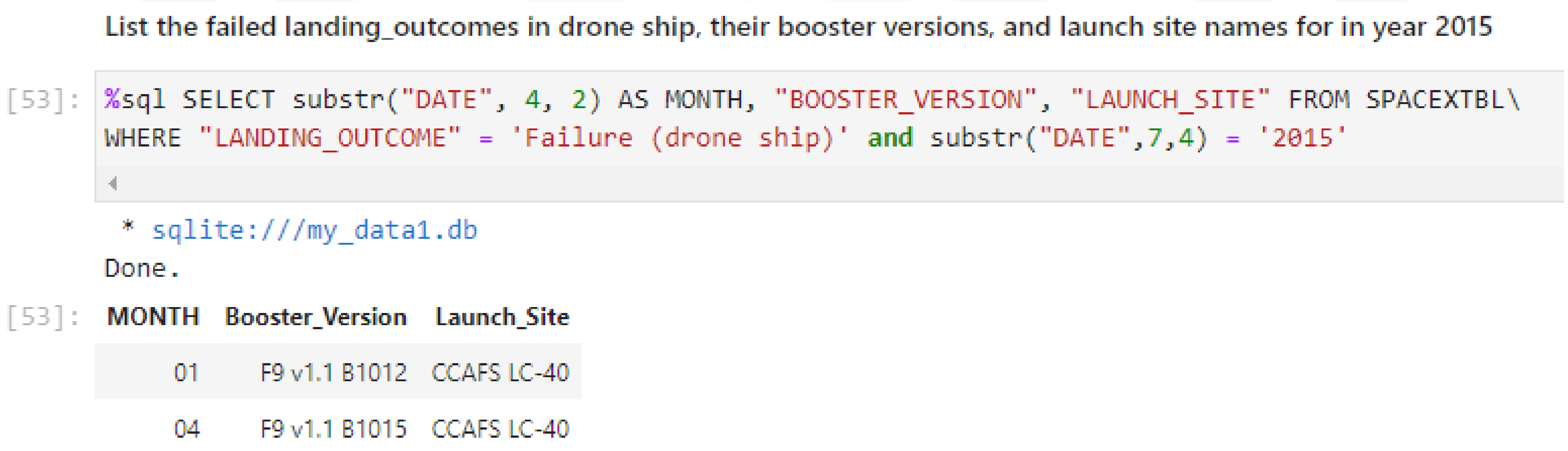
```
[20]: %sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
      WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)

 * sqlite:///my_data1.db
Done.
```



[20]:

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records



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

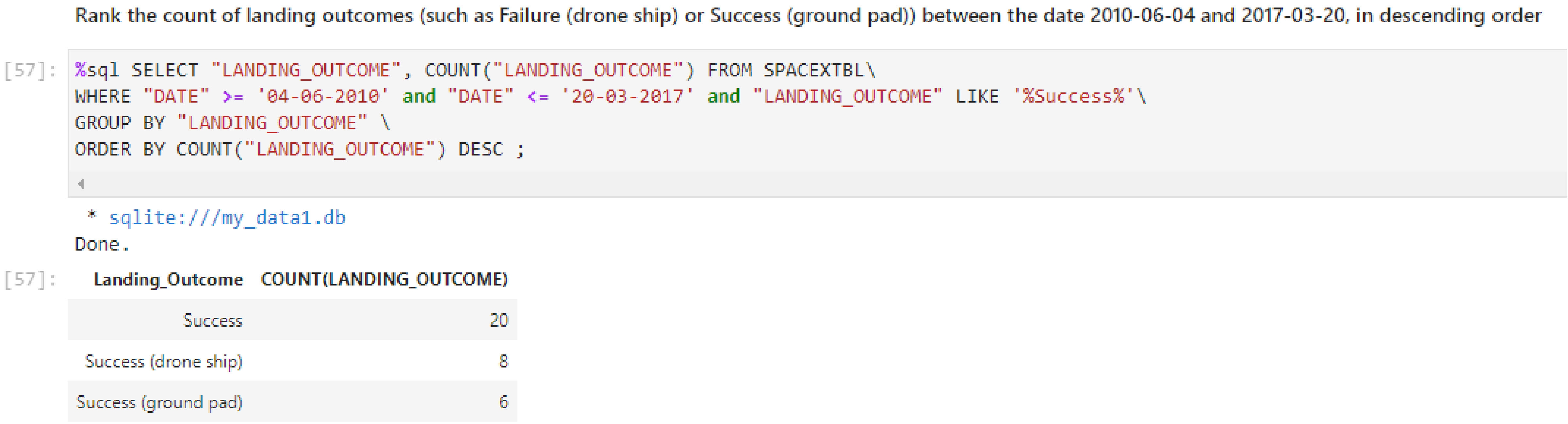
```
[53]: %sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
      WHERE "LANDING_OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

[53]:

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

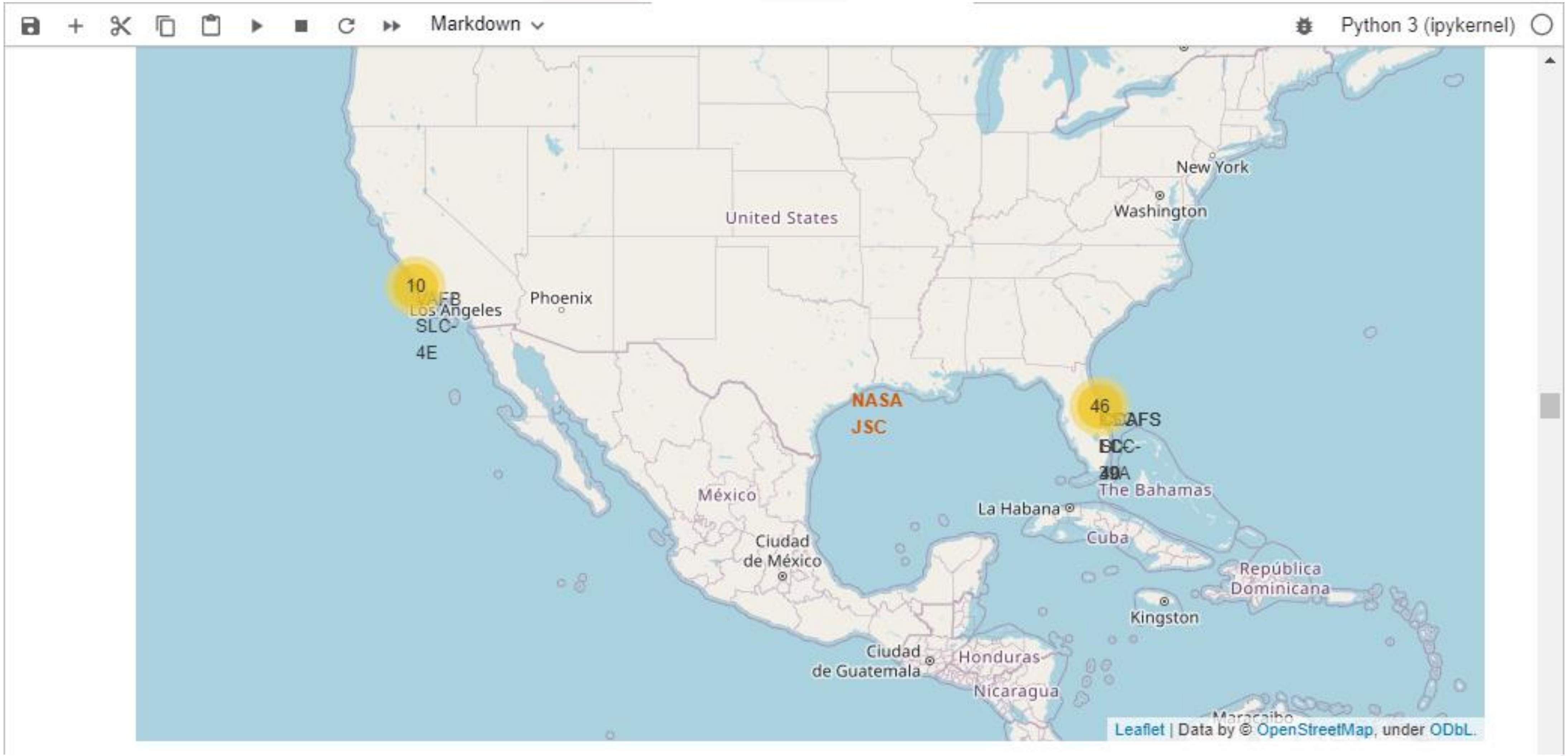
```
[57]: %sql SELECT "LANDING_OUTCOME", COUNT("LANDING_OUTCOME") FROM SPACEXTBL\
      WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING_OUTCOME" LIKE '%Success%'\
      GROUP BY "LANDING_OUTCOME" \
      ORDER BY COUNT("LANDING_OUTCOME") DESC ;
```

```
 * sqlite:///my_data1.db
Done.
```

[57]:

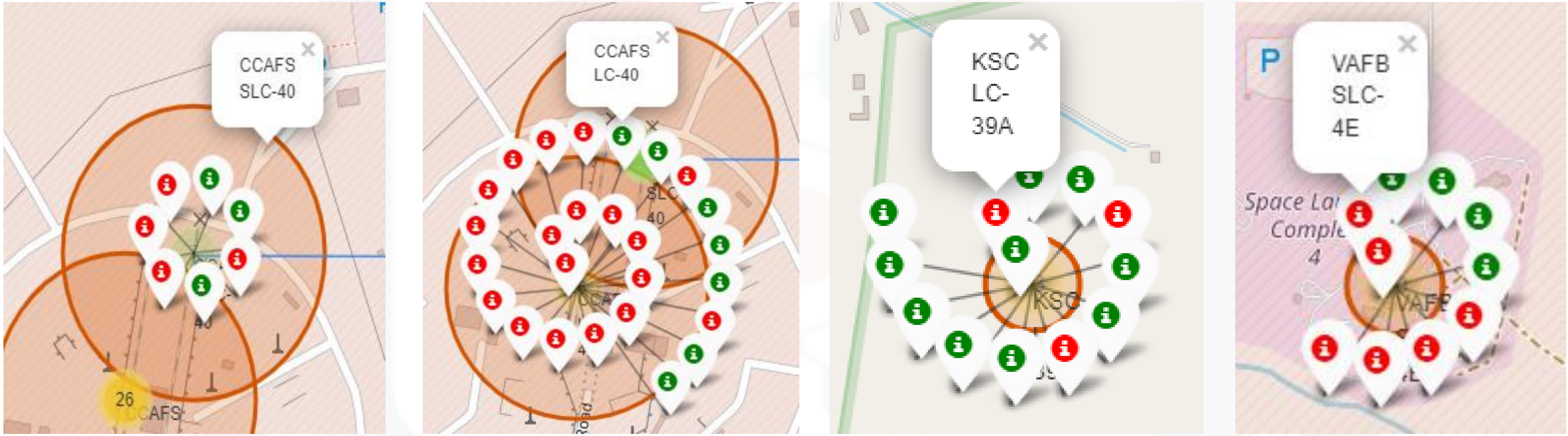
Landing_Outcome	COUNT(LANDING_OUTCOME)
Success	20
Success (drone ship)	8
Success (ground pad)	6

Folium Map – Ground Stations



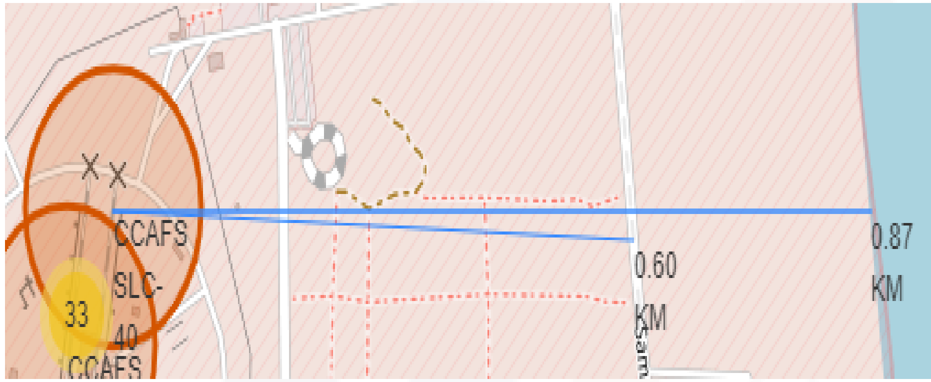
SKILLS NETWORK

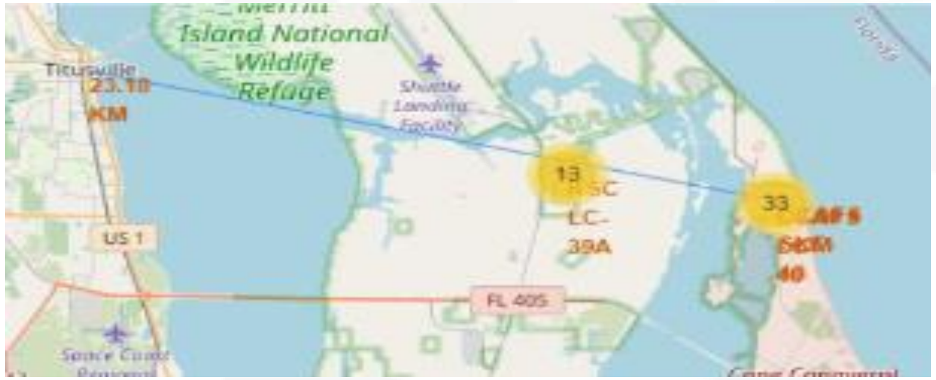
Folium Map – Color Labeled Markers

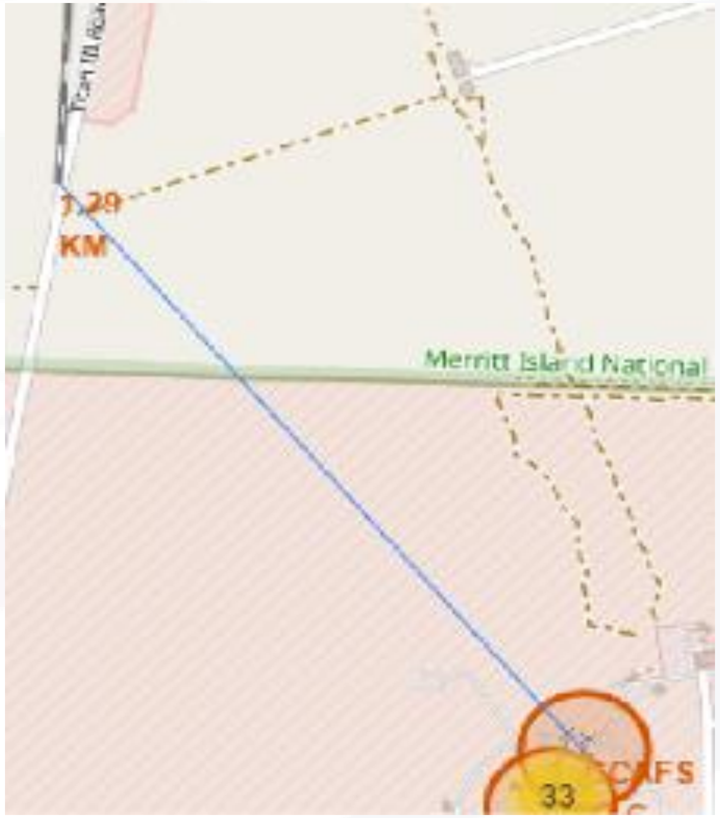


Green marker represents successful launches. Red marker represents unsuccessful launches.

Folium Map – Distances between CCAFS SLC-40 and its proximities

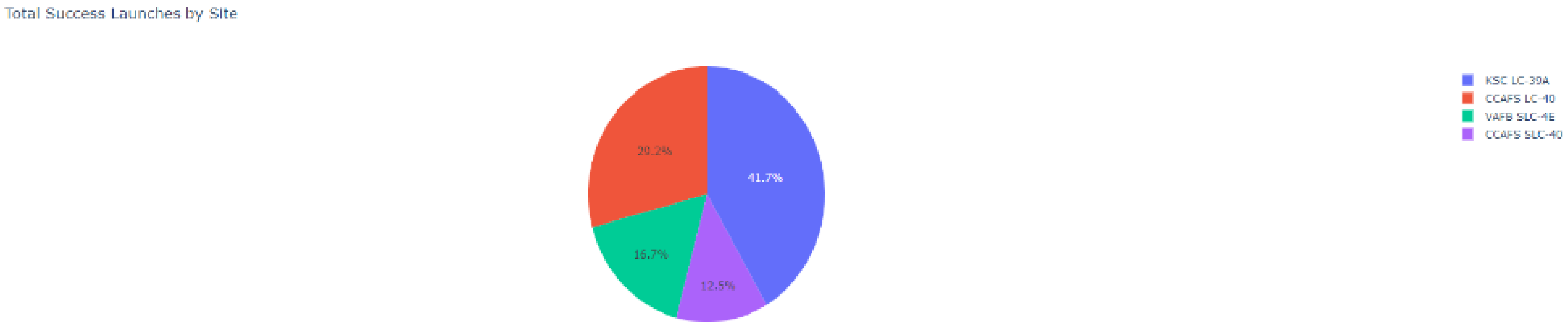






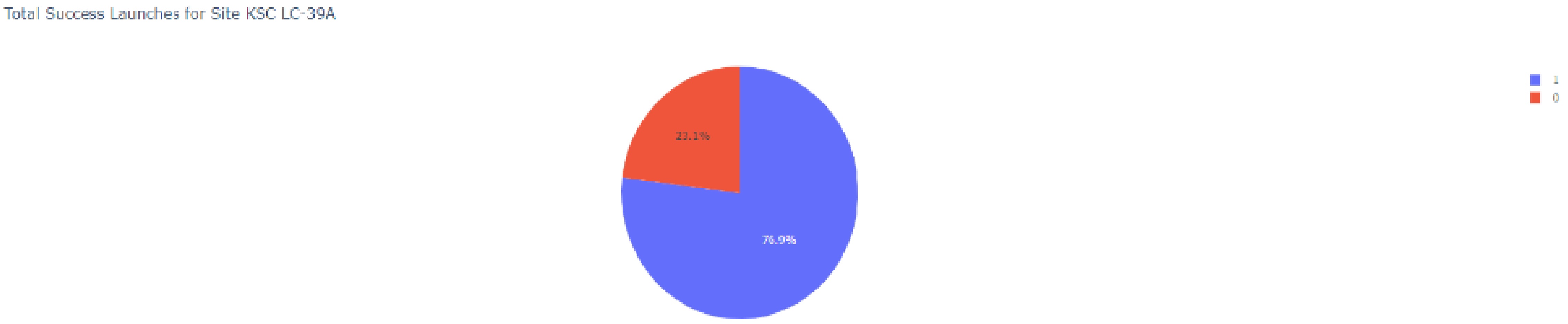
- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? No

DASHBOARD – Total success by site



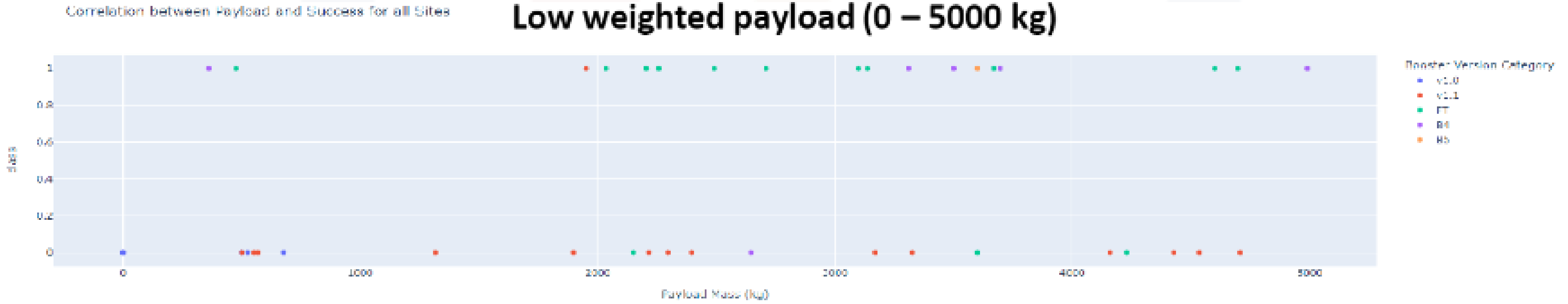
KSC LC-39A has the best success rate of launches

DASHBOARD – Total success launches for Site KSC LC-39A

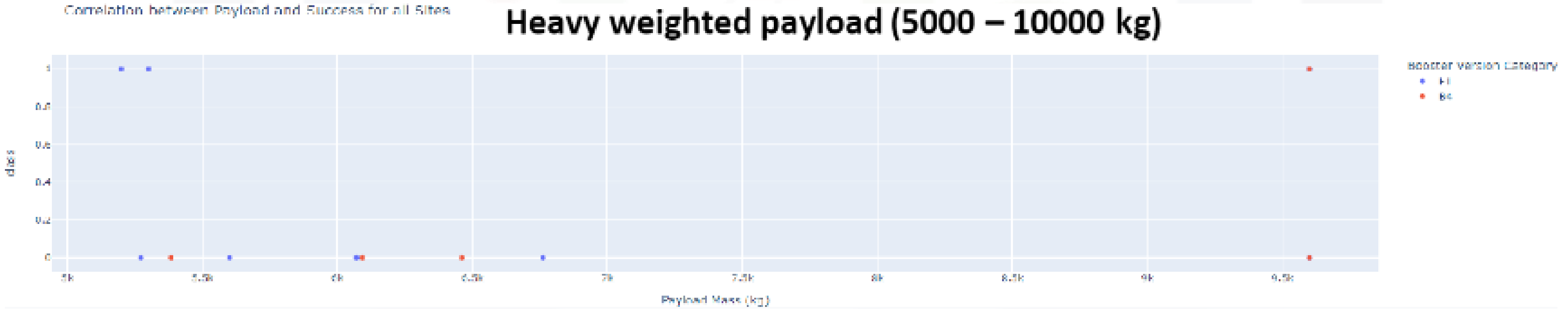


KSC LC-39A has achieved a 76.9% success rate

DASHBOARD – Payload mass vs Outcome for all sites with different payload mass selected



Low weighted payload (0 – 5000 kg)

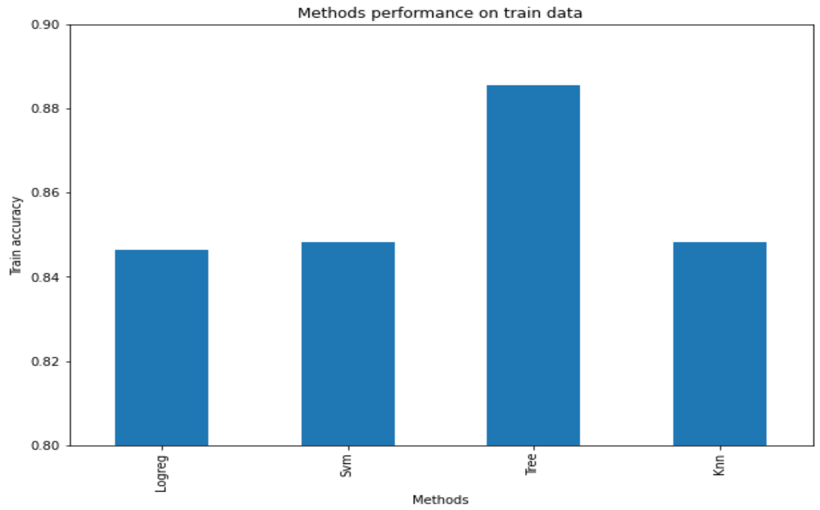


Heavy weighted payload (5000 – 10000 kg)

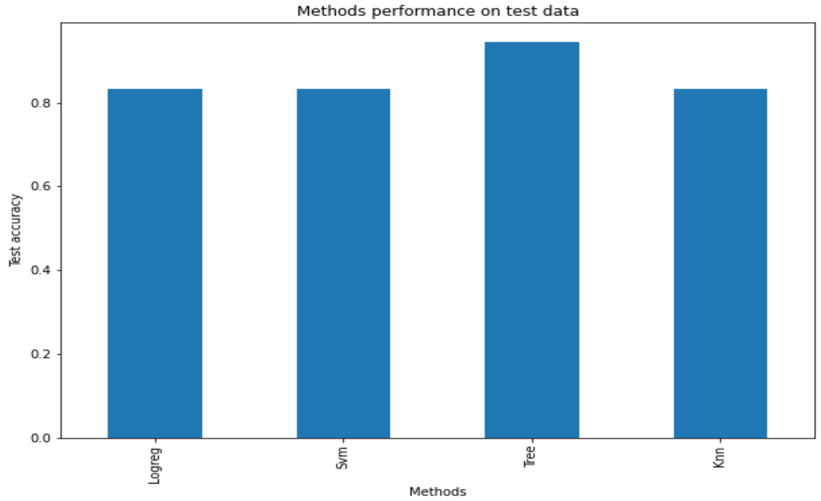
Low weighted payloads have a better success rate than the heavy

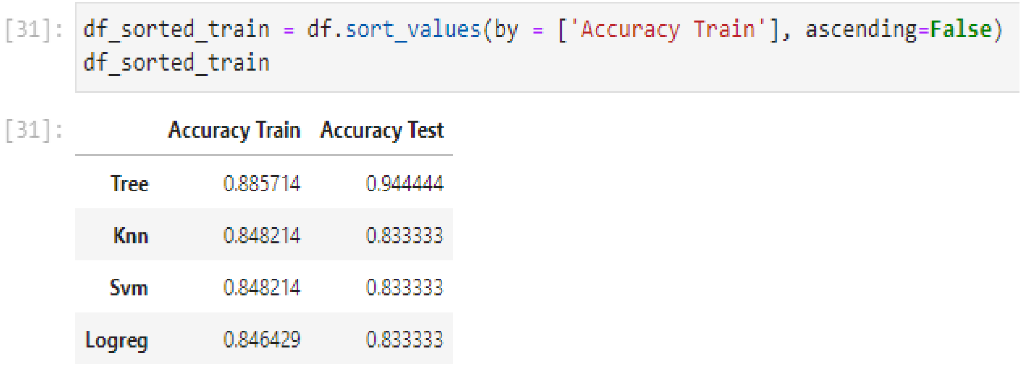
Classification Accuracy

```
[35]: (0.8, 0.9)
```



```
[36]: Text(0.5, 1.0, 'Methods performance on test data')
```



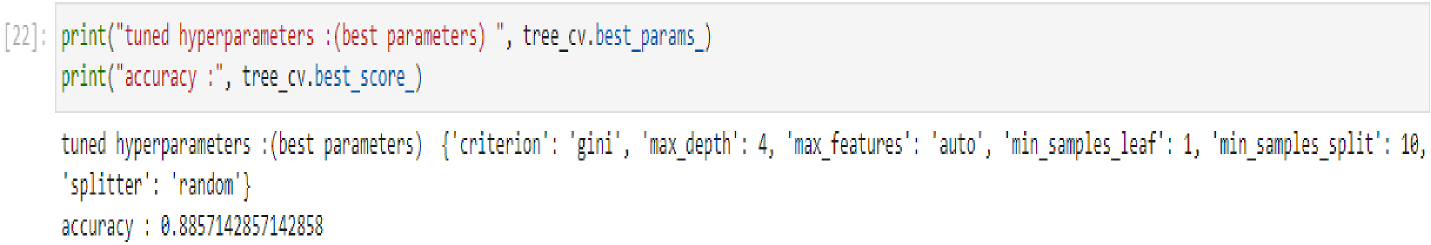


```
[31]: df_sorted_train = df.sort_values(by = ['Accuracy Train'], ascending=False)
      df_sorted_train
```

[31]:

	Accuracy Train	Accuracy Test
Tree	0.885714	0.944444
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333

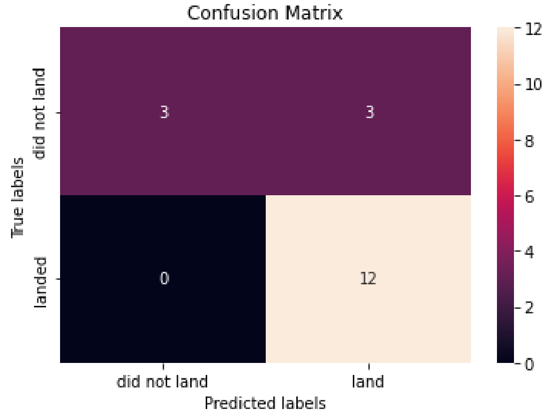
Decision tree has the best performance in the accuracy test



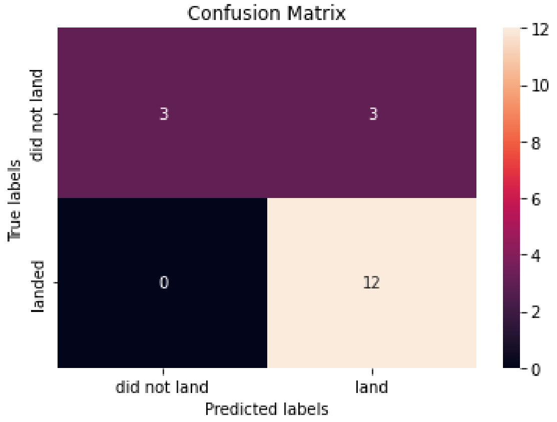
```
[22]: print("tuned hyperparameters :(best parameters) ", tree_cv.best_params_)
      print("accuracy :", tree_cv.best_score_)
```

```
tuned hyperparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 10,
'splitter': 'random'}
accuracy : 0.8857142857142858
```

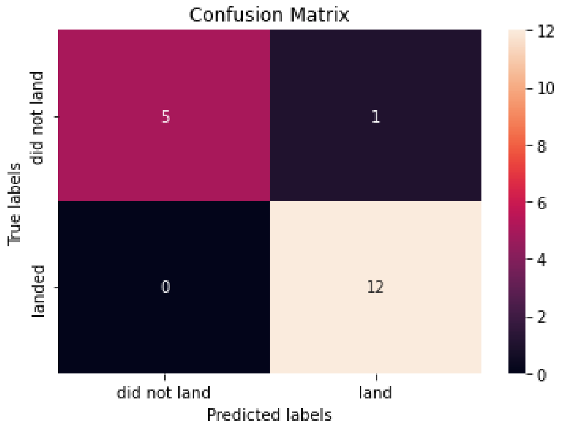
Confusion Matrix



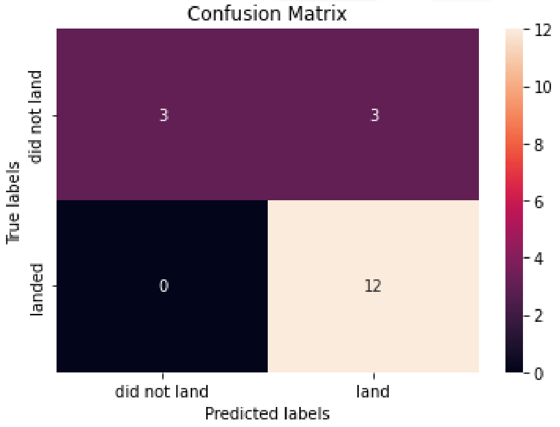
Logistic Regression



SVM

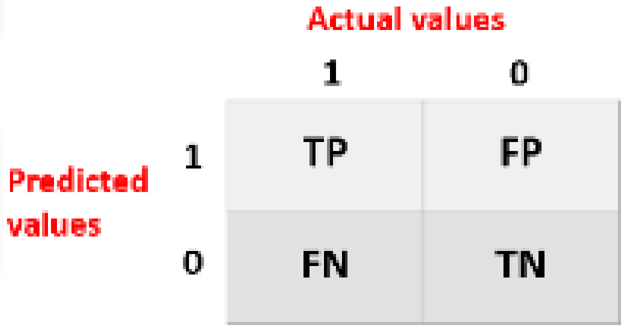


Decision Tree



KNN

The confusion matrix of decision tree confirms that it is the best model



		Actual values	
		1	0
Predicted values	1	TP	FP
	0	FN	TN

CONCLUSION

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches;
- The orbits with the best success rates are GEO, HEO, SSO e ES-L1;
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass;
- Low weighted payloads perform better than the heavy weighted payloads;
- KSC LC-39 A is the best launch site. With the current data, we cannot explain why some launch sites are better than others;
- We choose the Decision Tree Algorithm as the best model. We choose Decision Tree Algorithm because it has a better test accuracy.