IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Elvis Mabika

10 August 2022

Outline

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

Executive Summary

Summary of methodologies

In this report we are trying to predict the cost of sending a rocket launch into orbit. This was done by finding out if the first stage in the launch lands successfully and then is reusable for other launches. If the first stage is reusable it can significantly cut the cost of launching a rocket. The data that was used was collected using the SpaceX REST API and web scrapping from related Wiki pages. The raw data was pre-processed in python and SQL. Analysis (Explanatory, dashboarding and predictive) was then done to gain more insights into the collected data

Executive Summary Cont…..

Summary of all results

- The best predictive classification model to predict if the first stage lands successfully was the Decision Tree Model with an accuracy score of 85%..
- Number of successfully launches increased as the number of attempts (flight numbers) increased. Overall success rate of SpaceX is 66.67% since it started.
- Launching sites are near costal areas and railway lines but far away from cities. From the 4 launching sites, KSC LC-39A and VAFB SLC 4E have a success rate of around 77%. Though CCAFS LC-40, has a success rate of 60%, but if the mass is above 10,000 kg the success rate is 100%.
- Orbits ES-L1, GEO, HEO and SSO had 100% success rates

Introduction

Commercial space travel is on a high rise with many private companies involved. The cost for space trave ranges from $62 to $165 million. Companies are making space travel affordable for everyone. However, most charge relatively higher charges with SpaceX having the lowest cost of around $62 million. It is reported that the company was able to change the economics of space flight by making use of its reusable rocket system. The purpose of this report is to answer the following questions:

1. Will stage one of a rocket launch be successful?
2. Which launch sites or payloads have more success flights?
3. How is launch success rate dependent on factors such as payload mass, orbit type, location and proximities of a launch site?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology

Data collection was done using the SpaceX rest API and also using web scrapping from Wikipedia.

Data wrangling and Processing

Used python and SQL. Data cleaning (missing values) and feature engineering was carried out before predictive analysis could be done

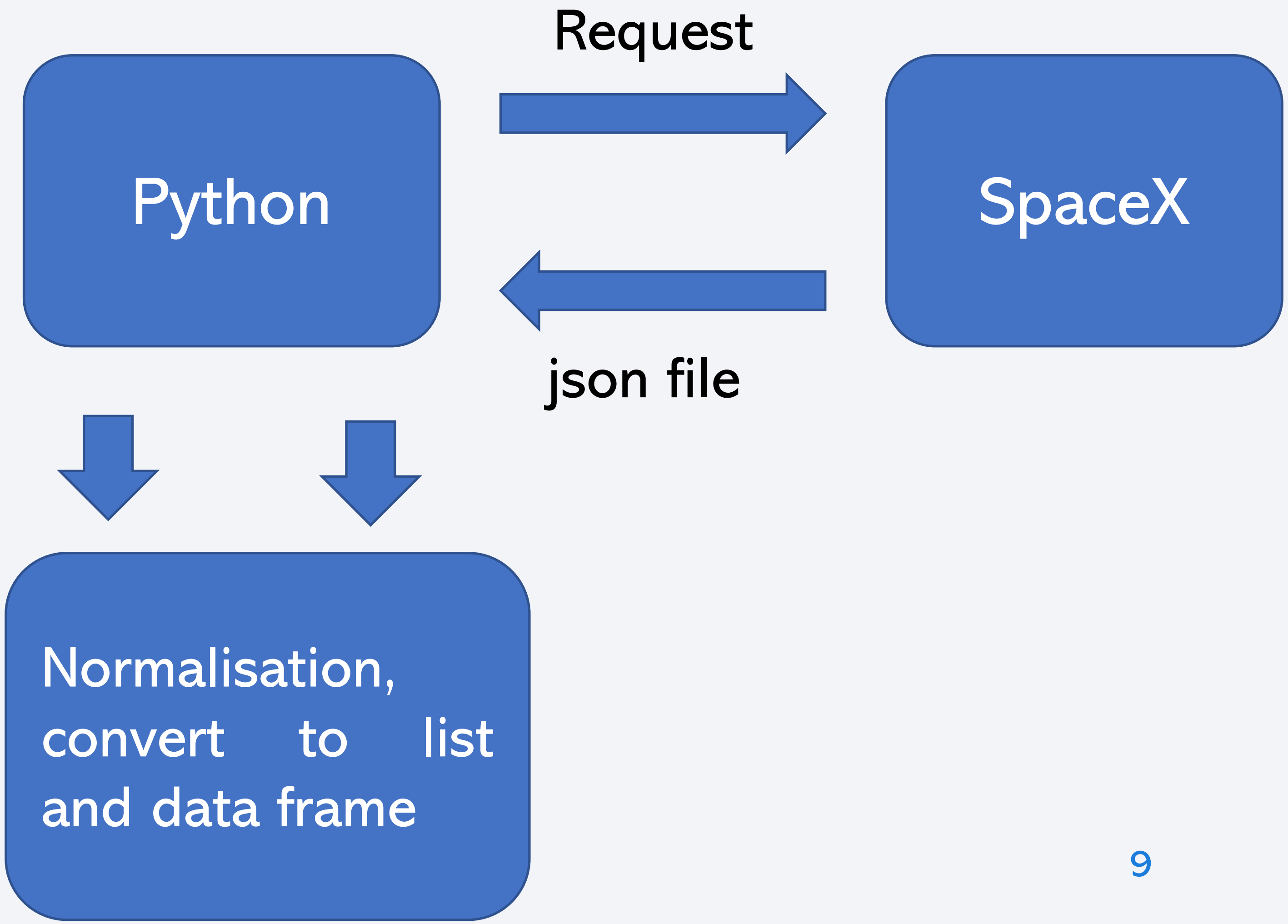
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
- Four classification were used (KNN, Logistic, Decision Trees and SVM were used)
- Fine tuning of hyperparameters for each model was done using GridSearchCV

Data Collection

- Data was collected using two methods namely web scrapping and a rest API. The SpaceX rest API was used to collect attributes of the data set such as booster versions, landing outcome, launch sites. The request method was used to get the json file that was then converted into a data frame.
- The request method and the BeautifulSoup package in python was used in web scrapping the data from Wikipedia

Data Collection – SpaceX API

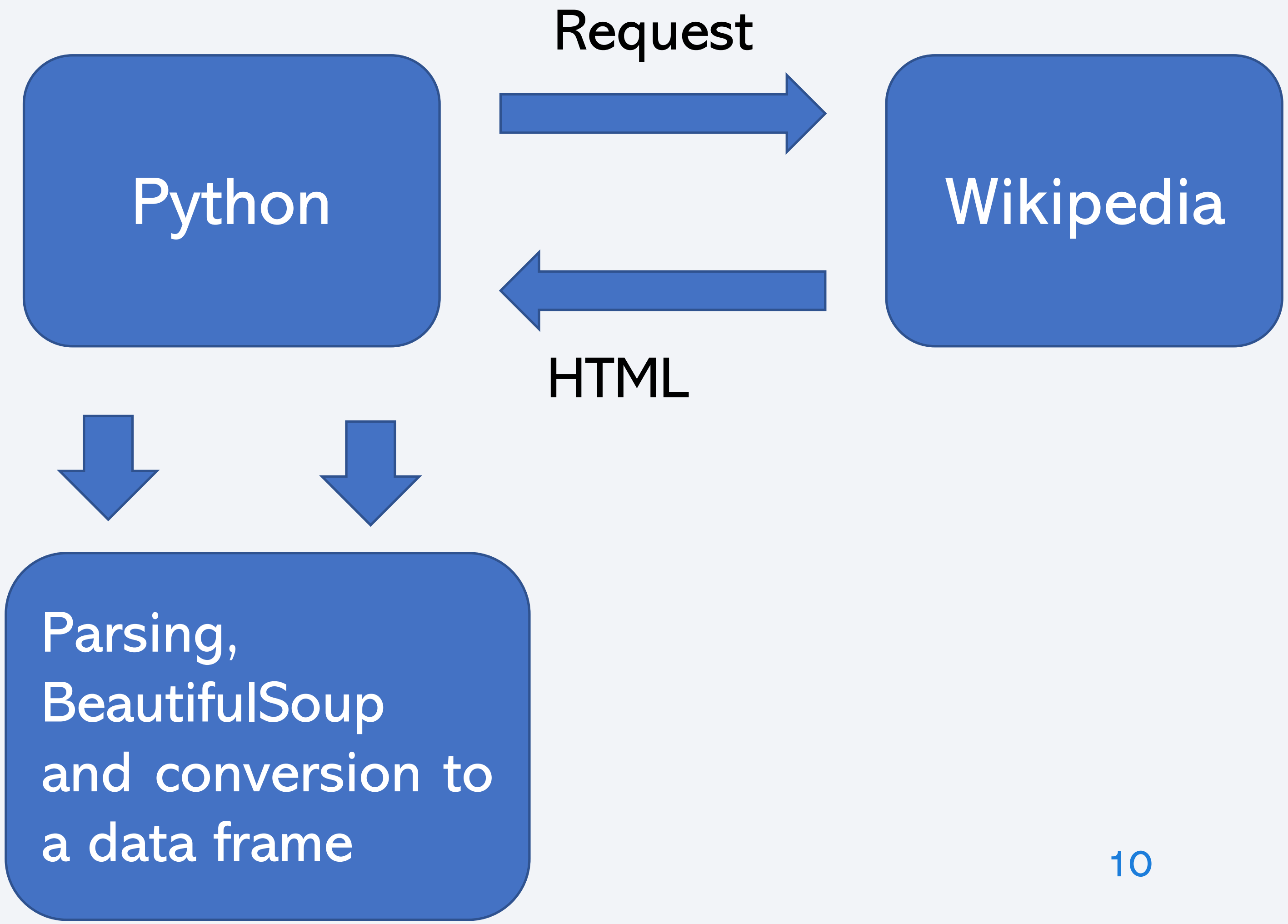
- We used get request method to the SpaceX API



https://github.com/Vinoro2002/IBM-Capestone-/blob/main/data_collection_api.ipynb

Data Collection - Scraping

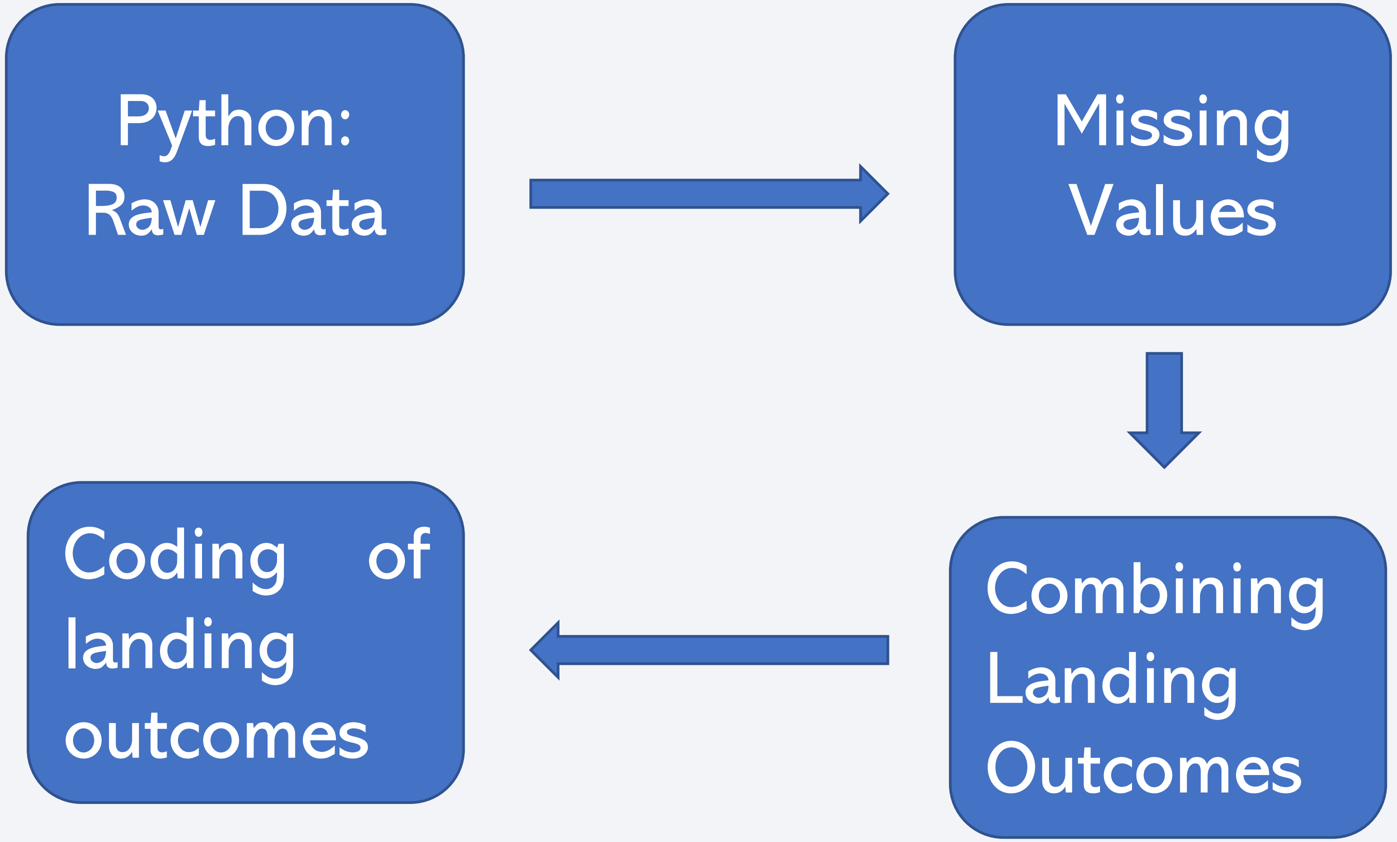
- We used get request method to HTML tables from Wikipedia. These were then parsed and converted into a data frame



https://github.com/Vinoro2002/IBM-Capestone-/blob/main/data_collection_webscrapping.ipynb

Data Wrangling

- Missing values were replaced with mean values. Landing outcomes were coded into two outcomes namely success/failure. The data was then normalized.



Python: Raw Data

Missing Values

Combining Landing Outcomes

Coding of landing outcomes

https://github.com/Vinoro2002/IBM-Capestone-/blob/main/data_wrangling.ipynb

EDA with Data Visualization

Plotted charts : we used scatter plots (Orbit, Payload launch site versus Flight number), bar plot (Orbit launch success rate) and a line plot (launch success yearly trend). These were used to get insights into correlations and trends that might exist among the variables.

https://github.com/Vinoro2002/IBM-Capestone/blob/main/EDA_data_viz.ipynb

EDA with SQL

SQL Queries done

- Total success/failure omission outcomes
- First landing outcome year
- Total and Average Payload carried by boosters
- Booster version that carried maximum payload

https://github.com/Vinoro2002/IBM-Capestone-/blob/main/EDA%20SQL.ipynb

Build an Interactive Map with Folium

Various map objects were used to plot the coordinates of the Launch sites in the map to analyze insights from the map.

The map objects include:

- Markers to denote the name of the particular Launch site using a label
- Circles to indicate the region of the coordinate in the defined circle radius
- MarkerCluster to indicate the cluster of the launch outcomes at a particular launch site (Whether the landing was successful or not)
- Lines measure the distance from the launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

https://github.com/Vinoro2002/IBM-Capestone-/blob/main/Data%20Viz%20Folium.ipynb

Build a Dashboard with Plotly Dash

The Dashboard has a scatter chart and a pie chart. The Dashboard was used as it gives more insights as compared static graphs. The interactions to the dashboard included dropdown list and a range slider to interact with a pie chart and a scatter point chart.

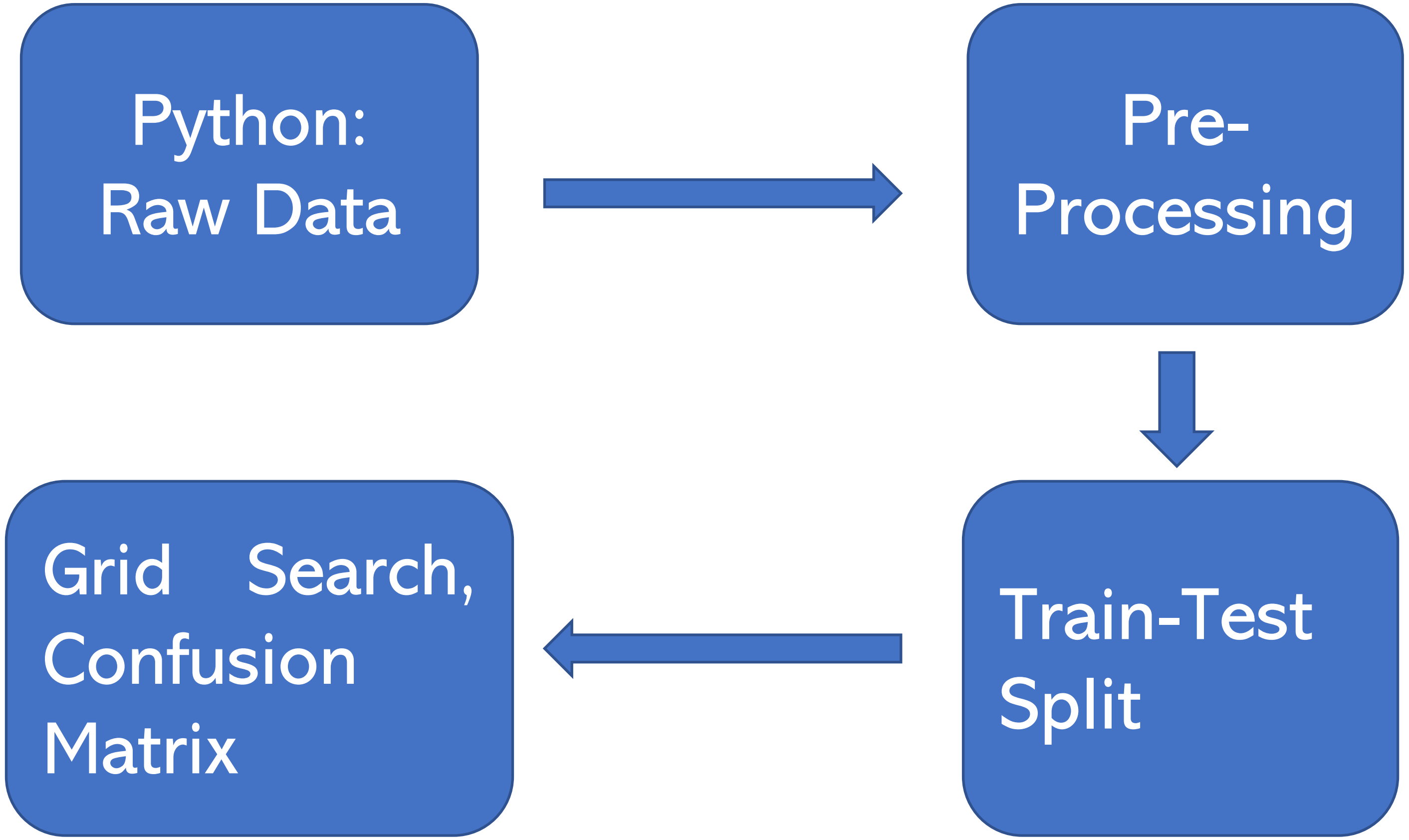
The graphs and interactions that were used were:

- Scatter chart to show the correlation
- Slider to select payload range
- Dropdown list to enable Launch Site selection

https://github.com/Vinoro2002/IBM-Capestone-/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

We first pre-processed our data allowing us to standardize it. We then split our data into training and testing data, We then trained the model and perform Grid Search, The Grid Search allowed us to find the hyperparameters that allow a given algorithm to perform best. We used various models namely, Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbours.



Python: Raw Data

Pre-Processing

Train-Test Split

Grid Search, Confusion Matrix

https://github.com/Vinoro2002/IBM-Capestone-/blob/main/predictive_analysis.ipynb

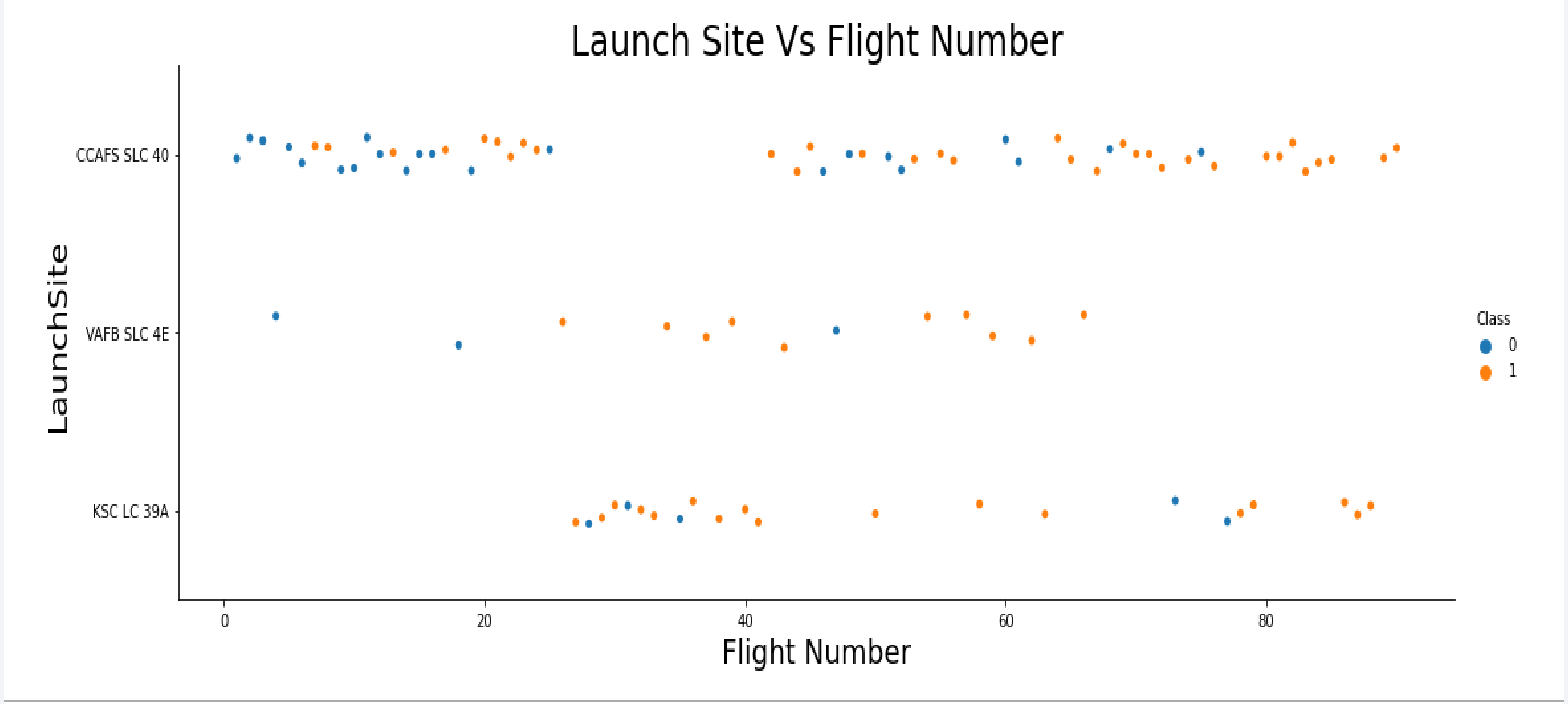
Results

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

Section 2

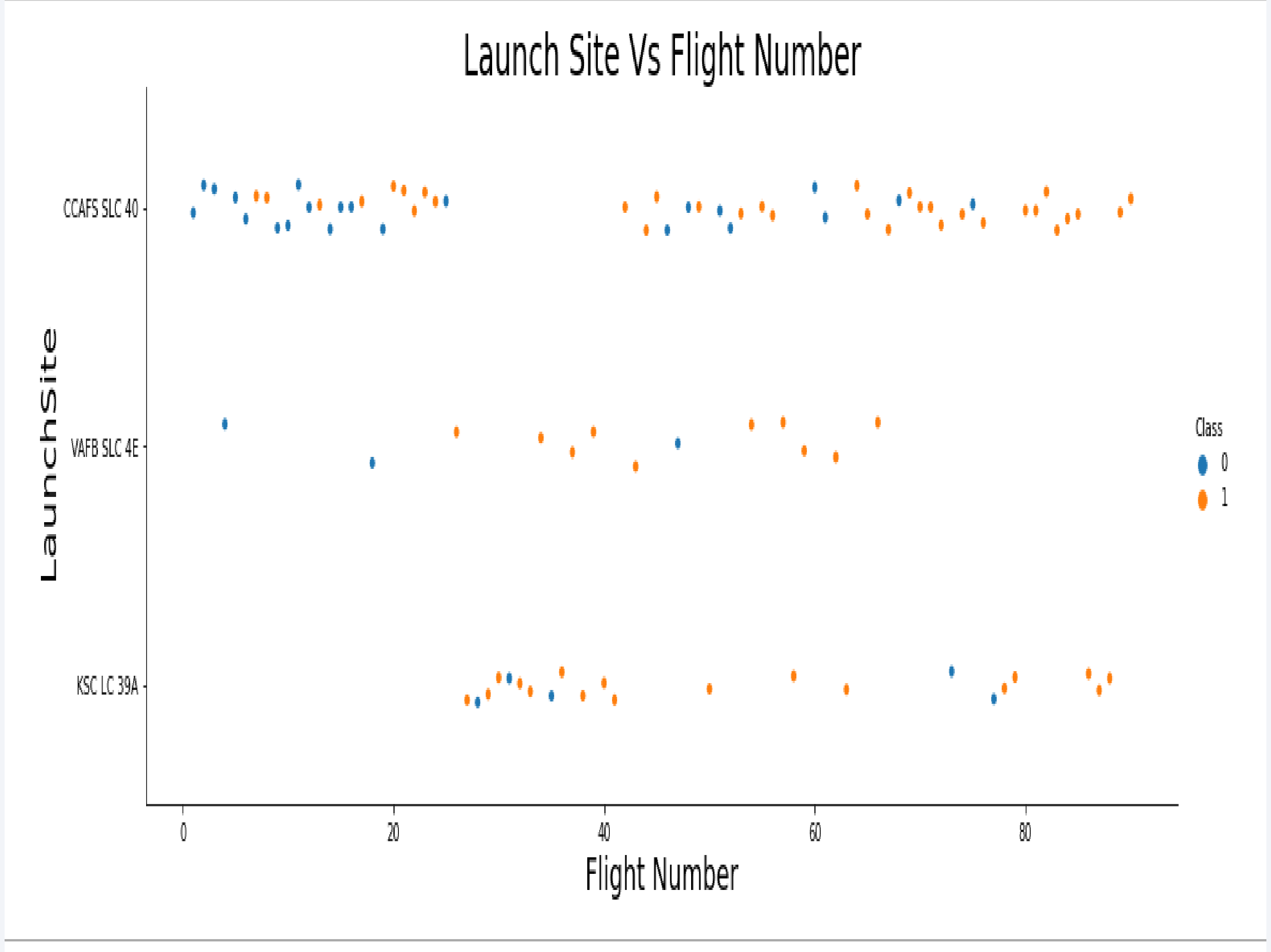
Insights drawn from EDA

Flight Number vs. Launch Site

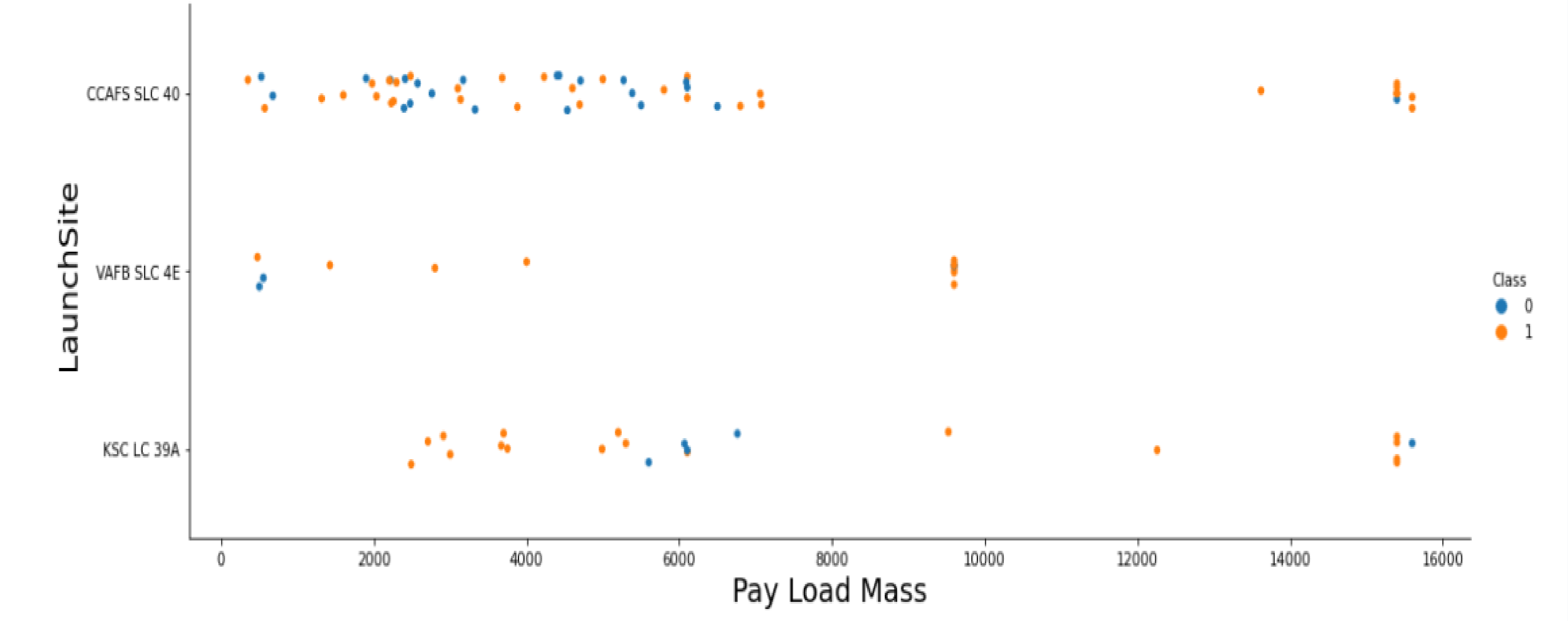


Flight Number vs. Launch Site

- We can see that as the company continued with launch attempts most were successful.
- CCAFS LC-40 had more attempted launches with the first attempts though being less successful
- The other two sites have a higher success rates , with KSC LC-39A having attempted launches at a latter stage.

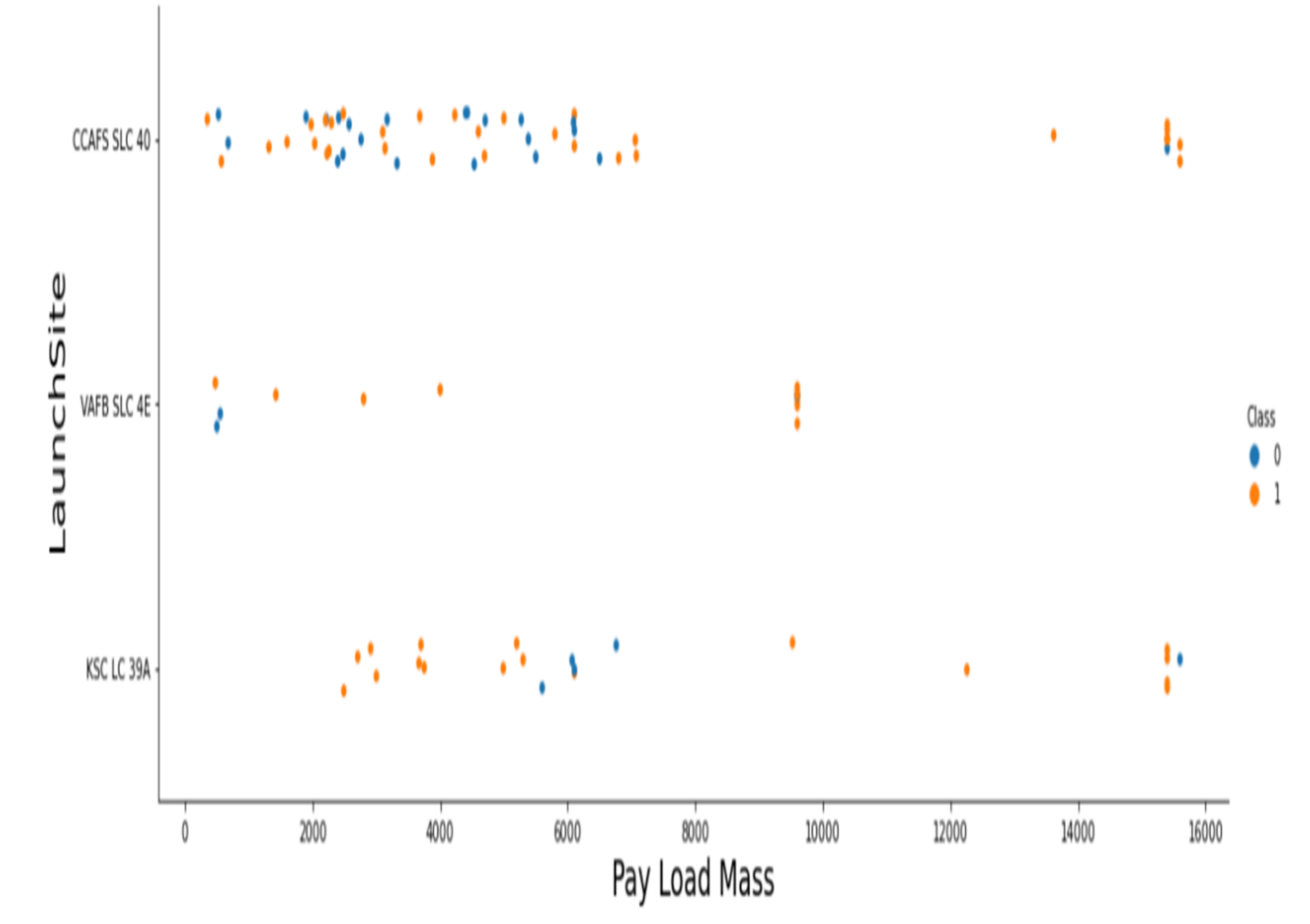


Payload vs. Launch Site



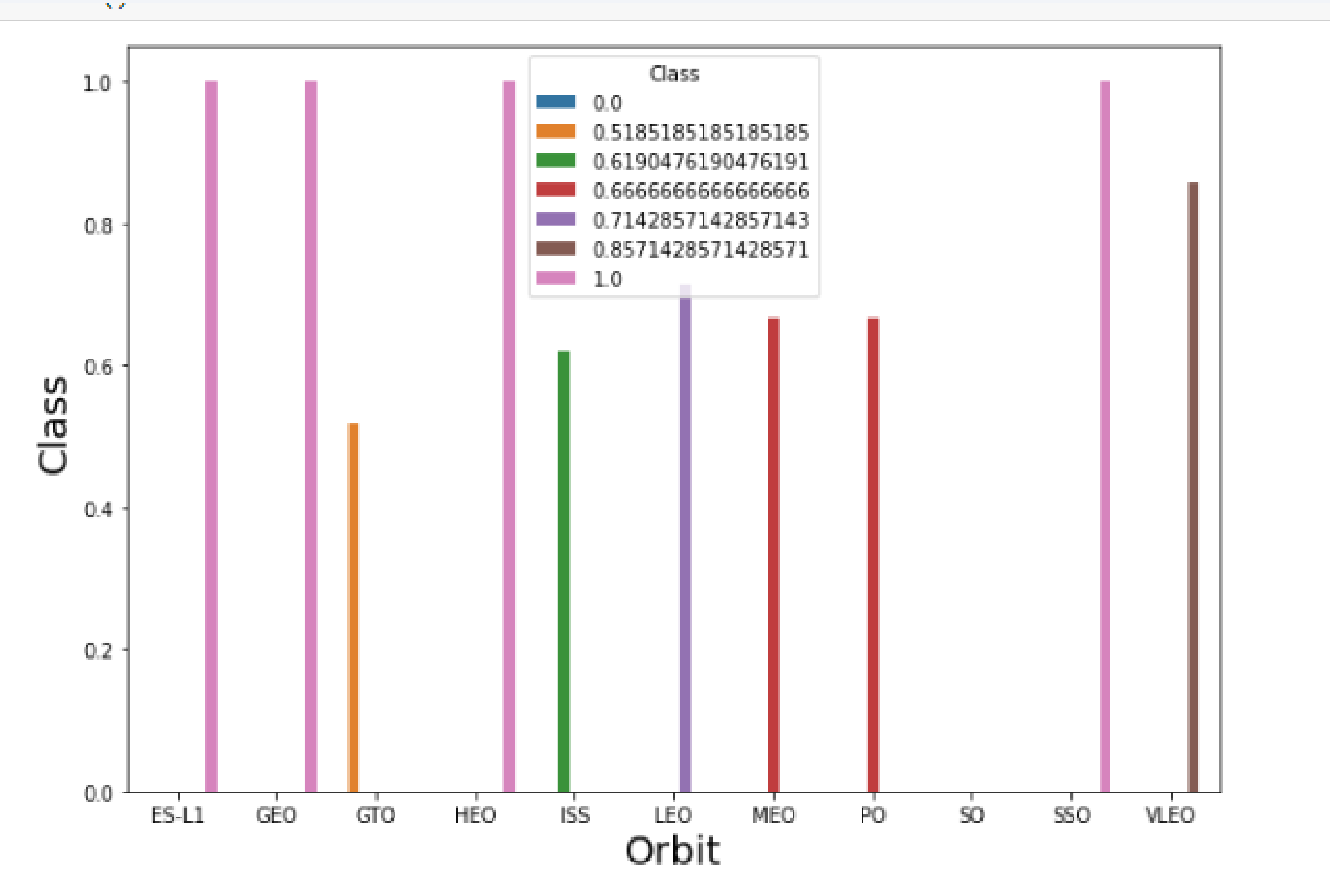
Payload vs. Launch Site

- Pay loads with mass greater than 10000kgs are more successful
- VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)
- KSC LC-39A has a 100% suscces rate for smaller payloads (under 5500kgs)

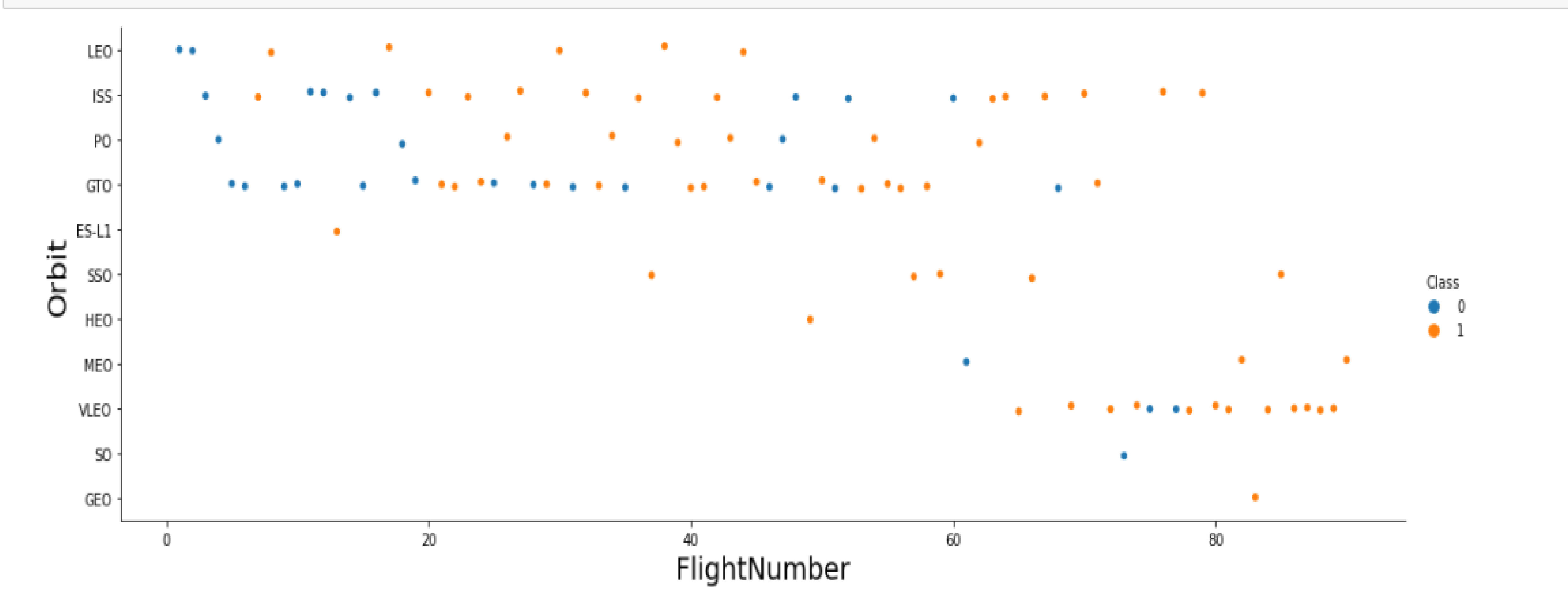


Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO had 100% success rates
- The lowest are with GTO (just above 50%) and SO with 0%



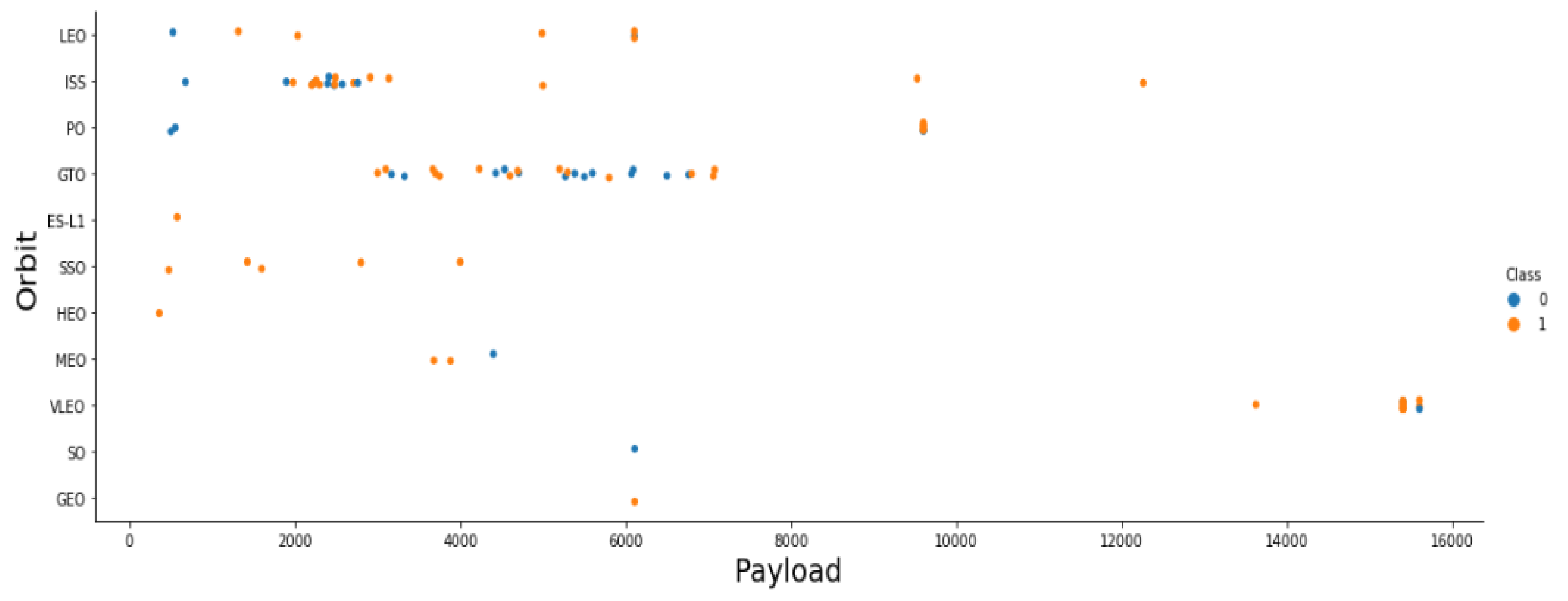
Flight Number vs. Orbit Type



Flight Number vs. Orbit Type

- LEO orbit has a success that appears to related to the number of flights
- The rest do not seem to have any relationship especially orbit GTO

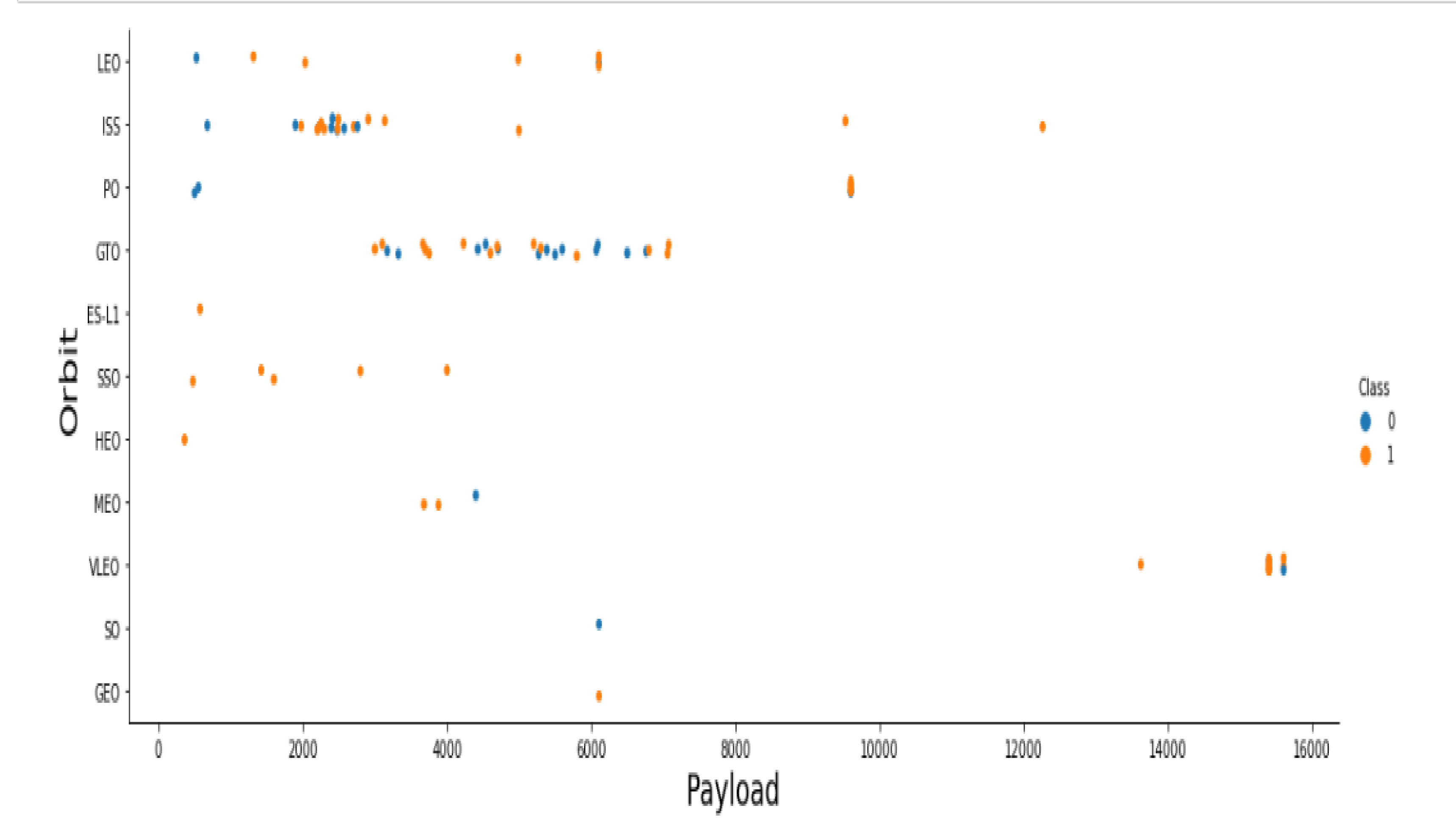
Payload vs. Orbit Type



Payload vs. Orbit Type

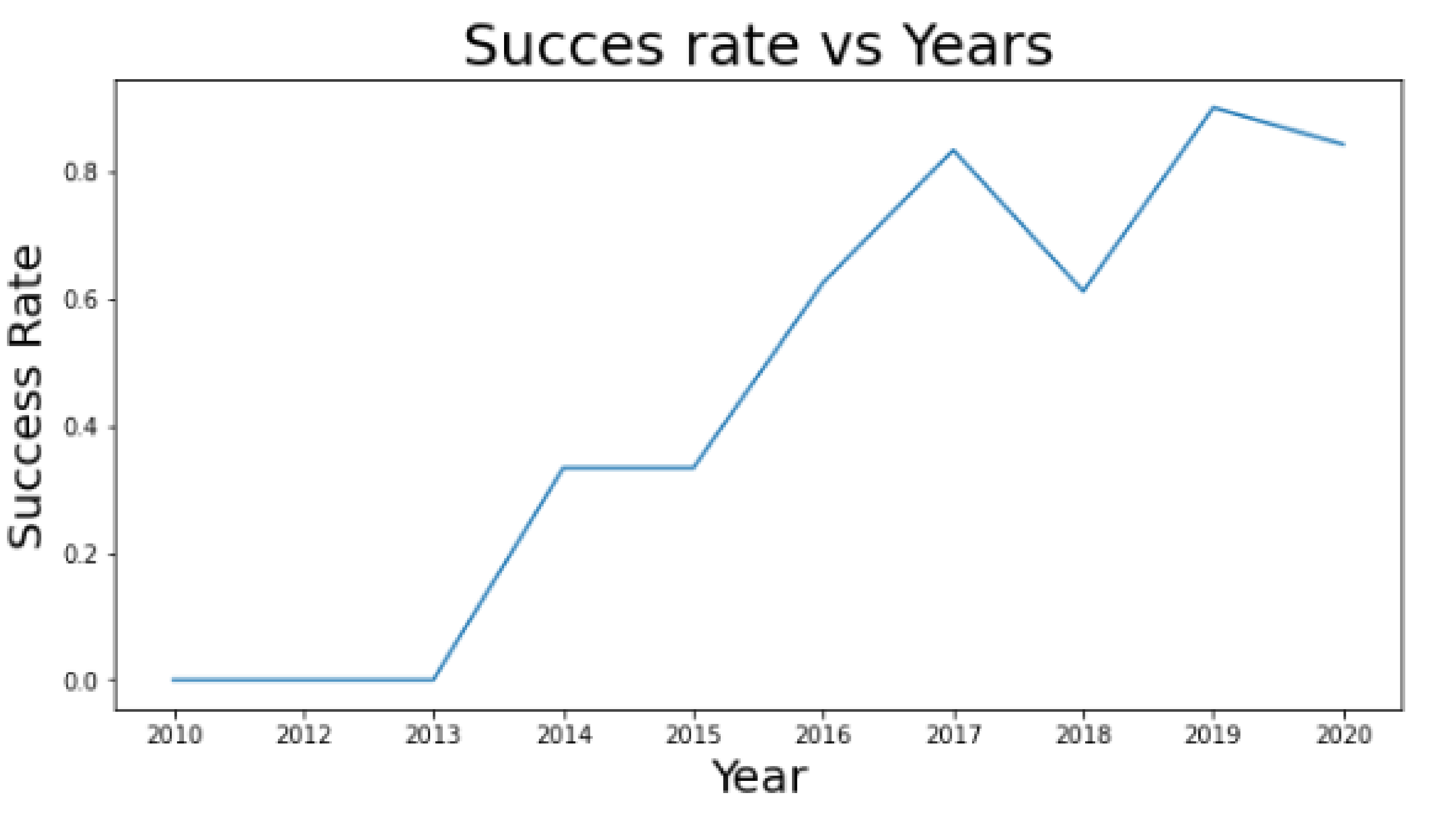
- PO, LEO and ISS have positive success rates as the heavy loads increase. The same can not be said of the rest and in particular GTO

- We also note that VLEO has more launches with pay loads above 10000kgs with a success rate of 80%



Launch Success Yearly Trend

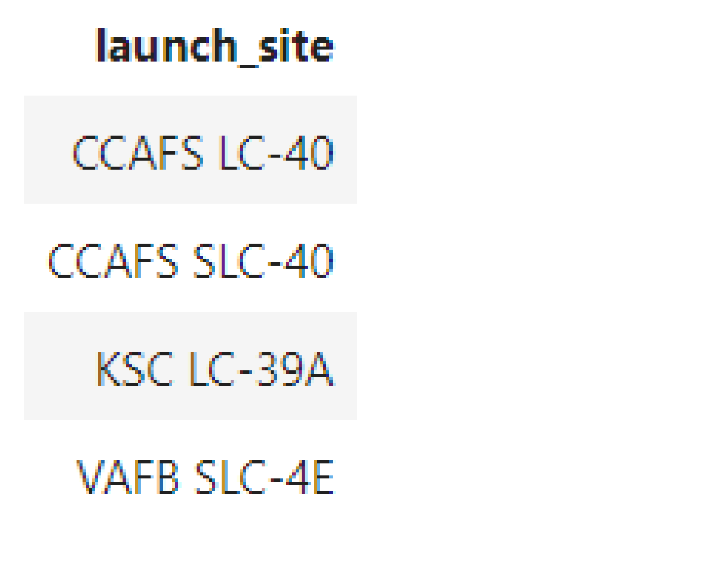
- We see that the success rate has been on the rise since 2013
- There was however a huge dip between 2017-2018



All Launch Site Names

- The unique launch sites used by Space X are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E. This was achieved by using the Distinct function of SQL:

```
%%sql
SELECT DISTINCT(launch_site)
from SpaceX
```



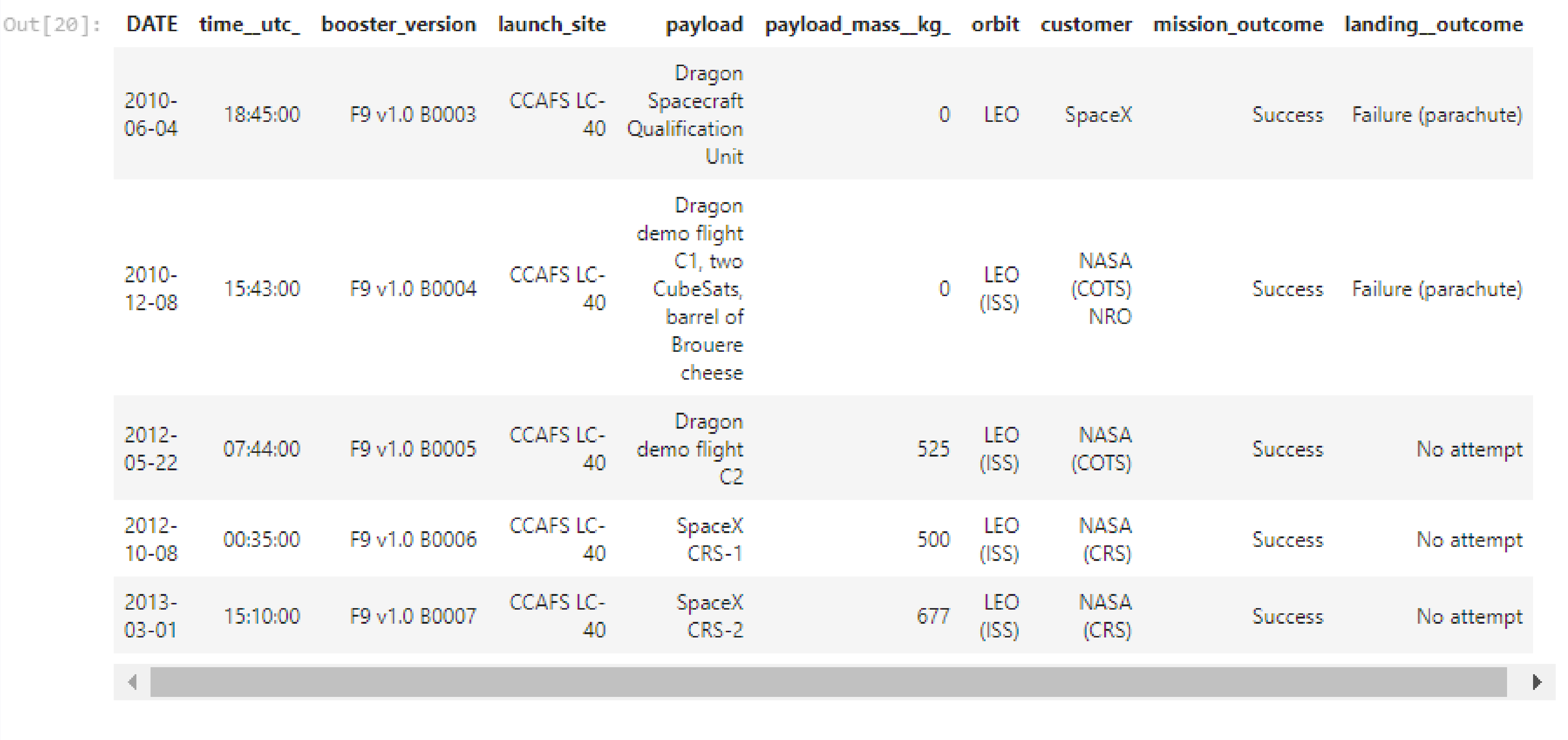
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- The query makes use of the WHERE clause and the wildcard % function to find launch sites records that begin CCA. To produce only 5 records we used the LIMIT clause:

```
%%sql
SELECT*
from SpaceX
WHERE launch_site LIKE 'CCA%'
LIMIT 5;
```

Launch Site Names Begin with 'CCA'

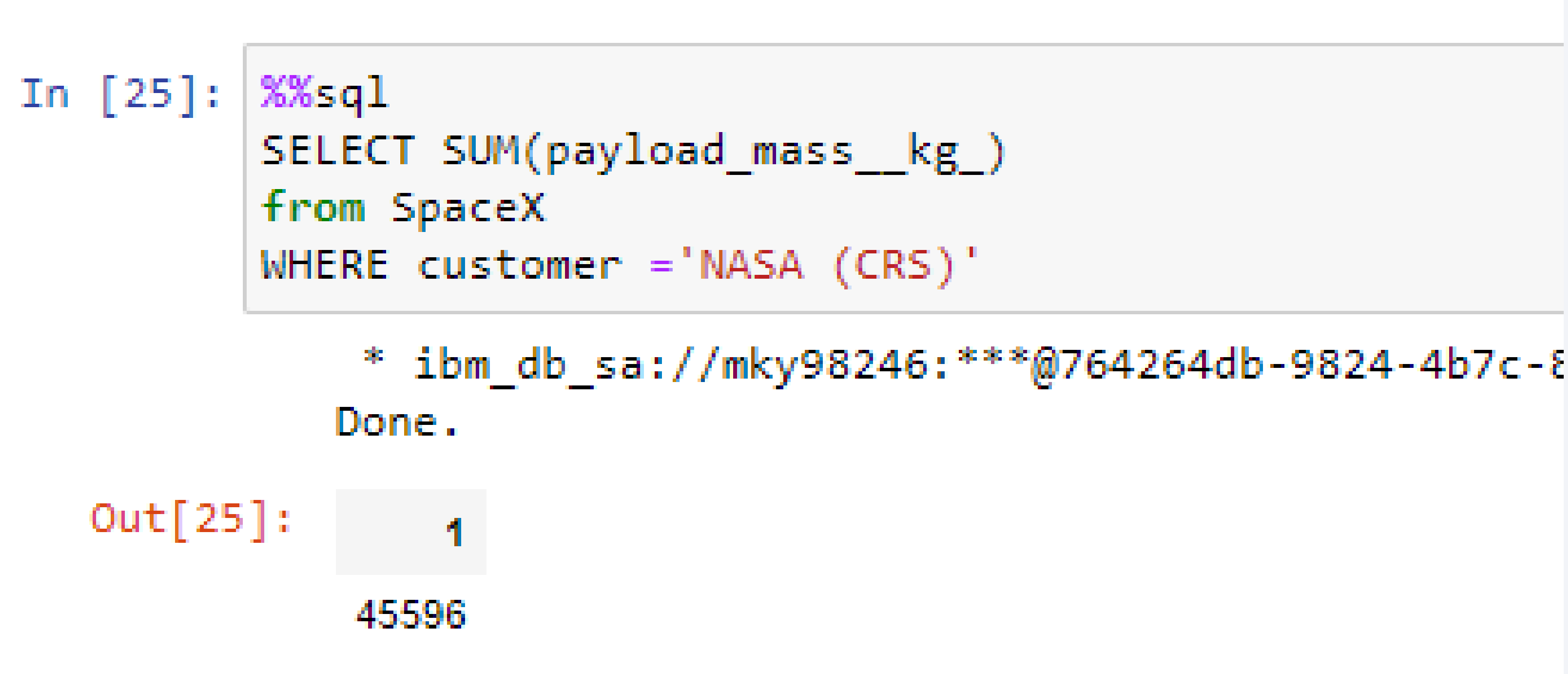


Out[20]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Query made use of the SUM function



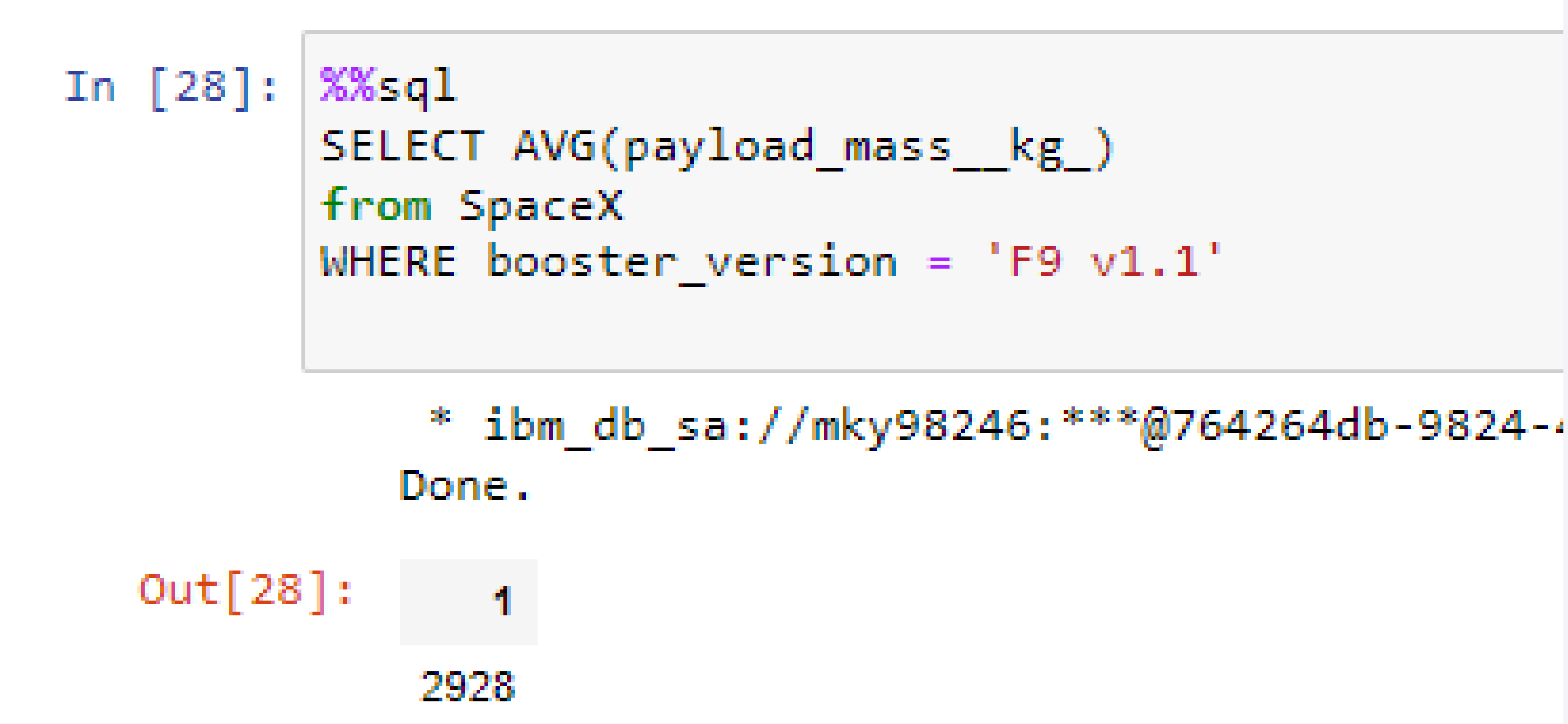
```
In [25]: %%sql
         SELECT SUM(payload_mass__kg_)
         from SpaceX
         WHERE customer ='NASA (CRS)'
```

```
 * ibm_db_sa://mky98246:***@764264db-9824-4b7c-
Done.
```

```
Out[25]:     1
         45596
```

Average Payload Mass by F9 v1.1

- To obtain the result the AVG function was used with the WHERE clause being used to filter data only for F9 v1.1



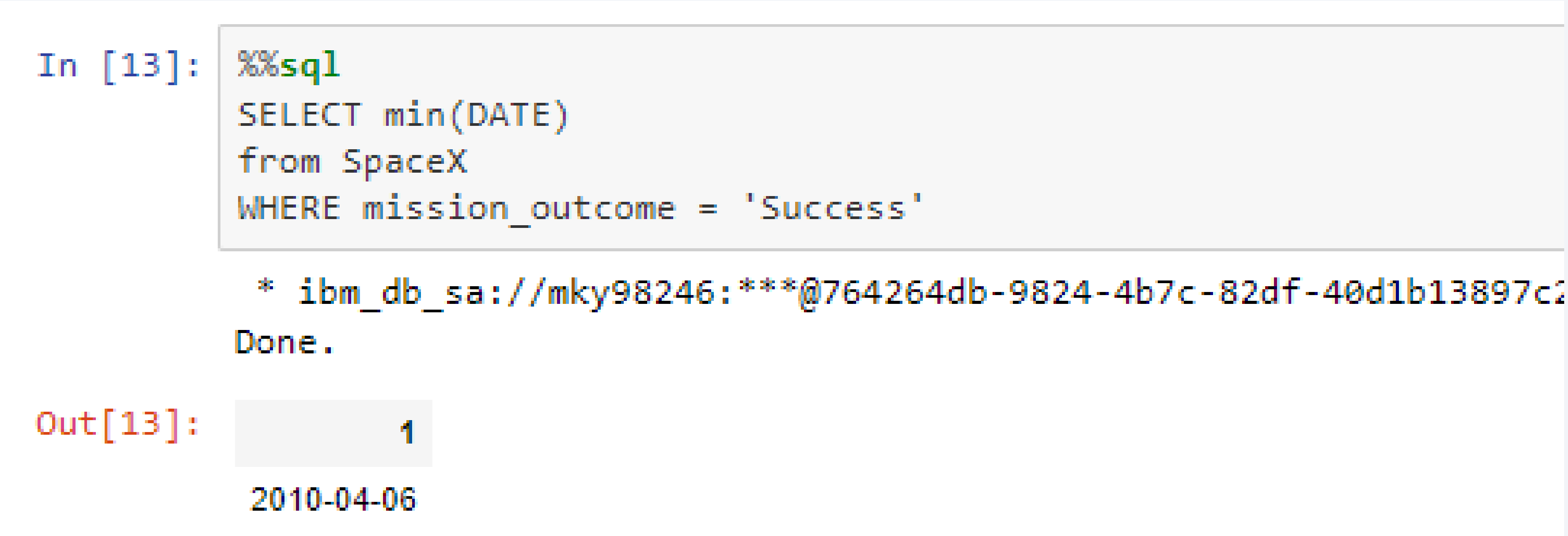
```
In [28]: %%sql
         SELECT AVG(payload_mass__kg_)
         from SpaceX
         WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://mky98246:***@764264db-9824-
Done.
```

```
Out[28]:    1
         2928
```

First Successful Ground Landing Date

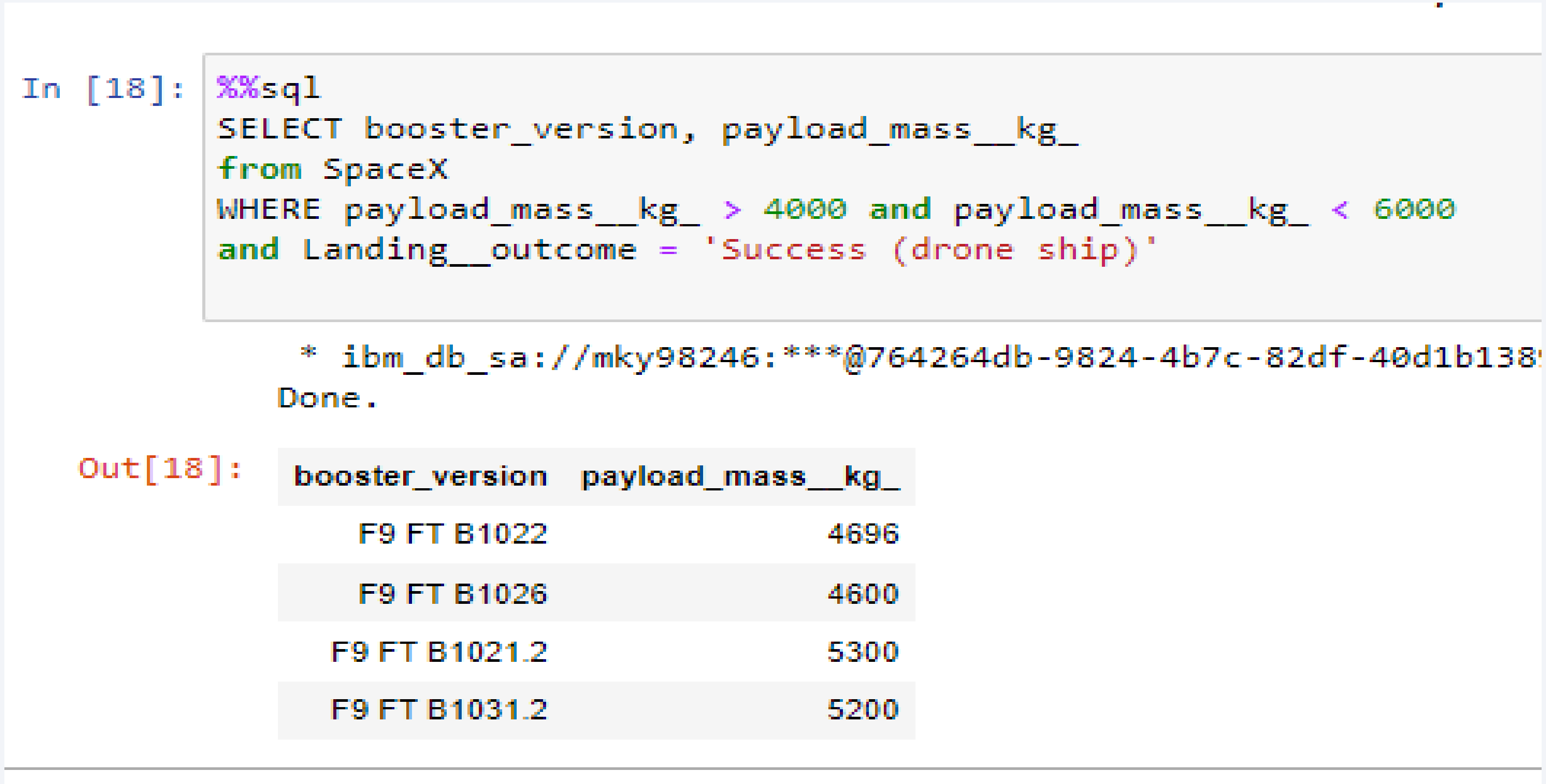
- We use the MIN function to obtain the earliest date and filtered the ones that where successful by the WHERE clause.



```
In [13]: %%sql
         SELECT min(DATE)
         from SpaceX
         WHERE mission_outcome = 'Success'

          * ibm_db_sa://mky98246:***@764264db-9824-4b7c-82df-40d1b13897c
         Done.

Out[13]:          1
          2010-04-06
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Two statements were used in WHERE clause together with the and statement to obtain the range and also to filter records for success drone ship



In [18]:

```
%%sql
SELECT booster_version, payload_mass__kg_
from SpaceX
WHERE payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
and Landing__outcome = 'Success (drone ship)'
```

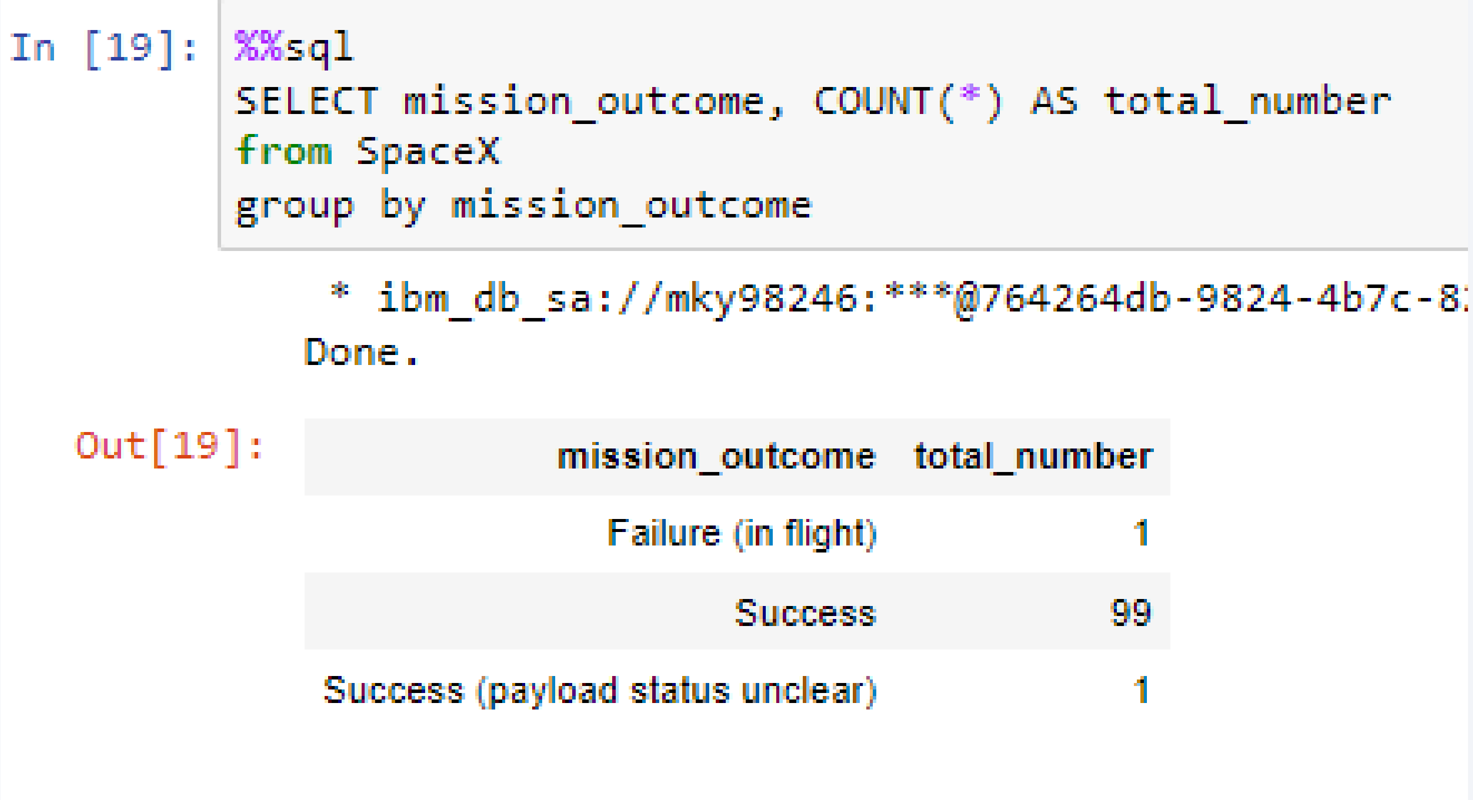
```
 * ibm_db_sa://mky98246:***@764264db-9824-4b7c-82df-40d1b138
Done.
```

Out[18]:

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

COUNT function was use to find the total with the GROUP function being used to categorize the status of the mission outcomes



```
In [19]: %%sql
         SELECT mission_outcome, COUNT(*) AS total_number
         from SpaceX
         group by mission_outcome
```

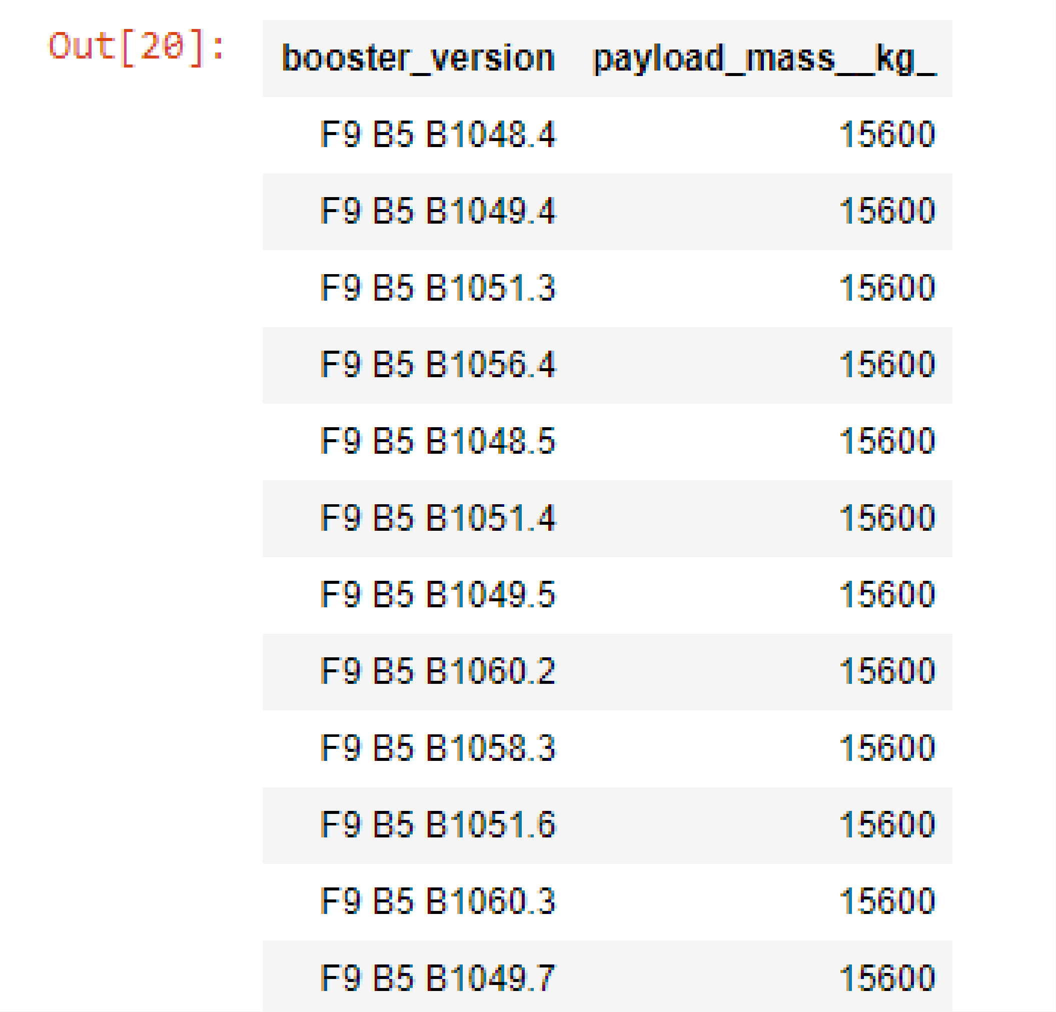
```
 * ibm_db_sa://mky98246:***@764264db-9824-4b7c-8
Done.
```

Out[19]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

A subquery was use to obtain the result together with the WHERE clause. The subquery contained the MAX function

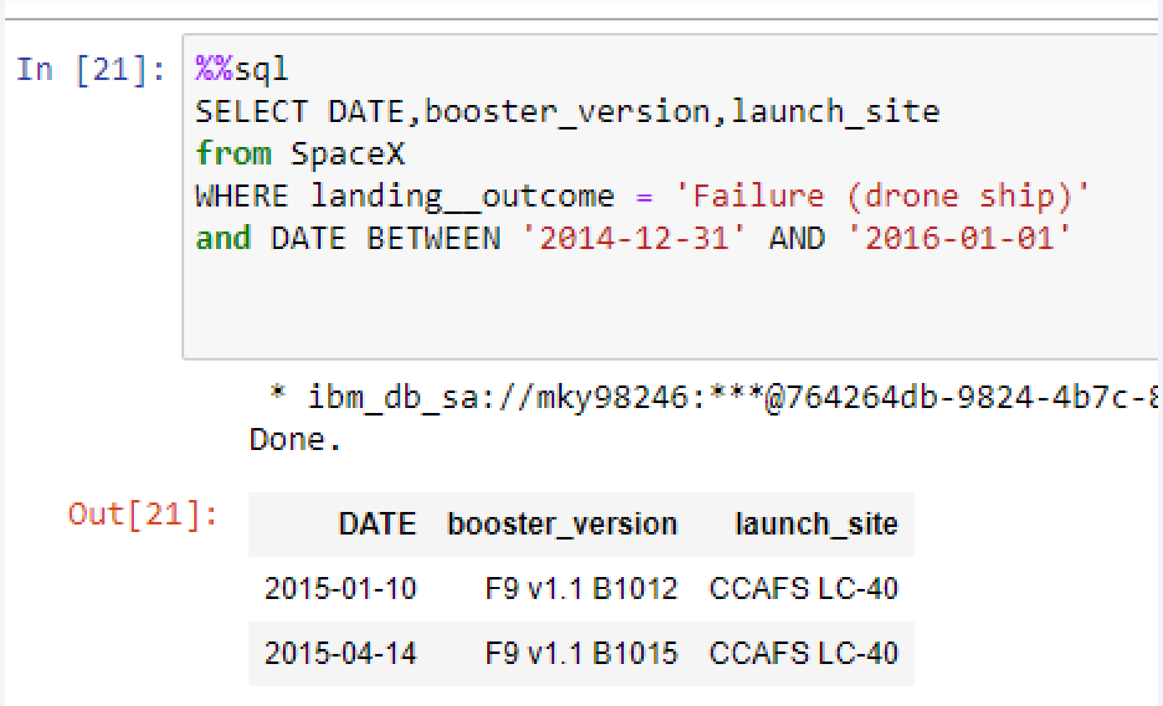


Out[20]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

We made use of the SELECT function to choose which columns to appear an the used two statements in the WHERE clause, with the first filtering only failure outcomes for drone ships with the second filtering the year 2015. We used the BETWEEN function to obtain the year range.



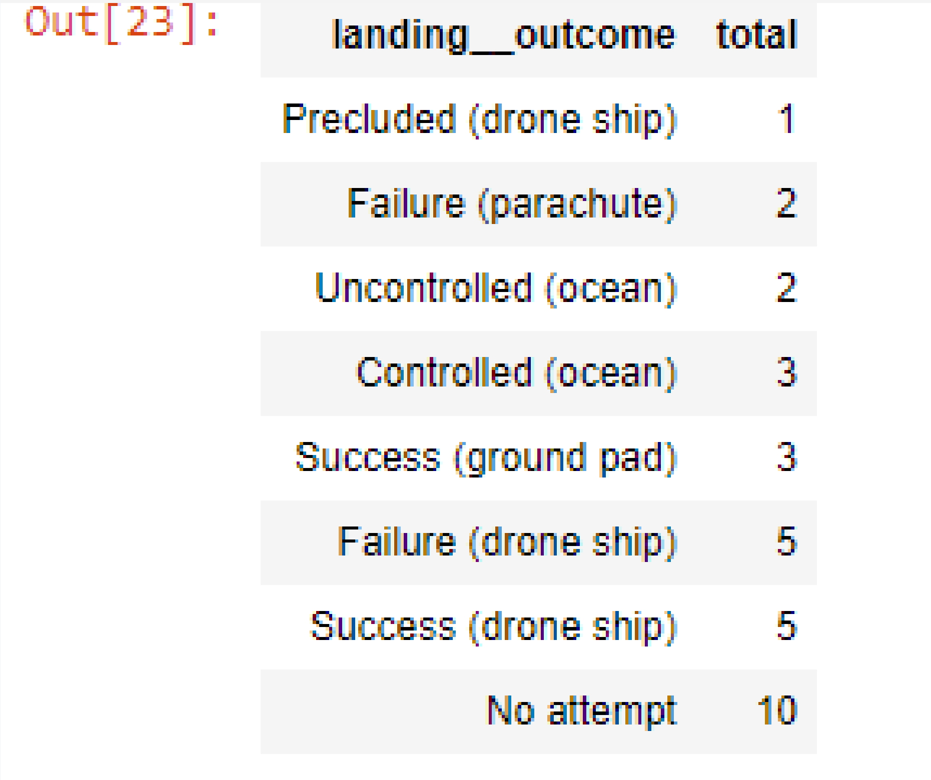
```
In [21]: %%sql
         SELECT DATE,booster_version,launch_site
         from SpaceX
         WHERE landing__outcome = 'Failure (drone ship)'
         and DATE BETWEEN '2014-12-31' AND '2016-01-01'

          * ibm_db_sa://mky98246:***@764264db-9824-4b7c-
         Done.
```

Out[21]:

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

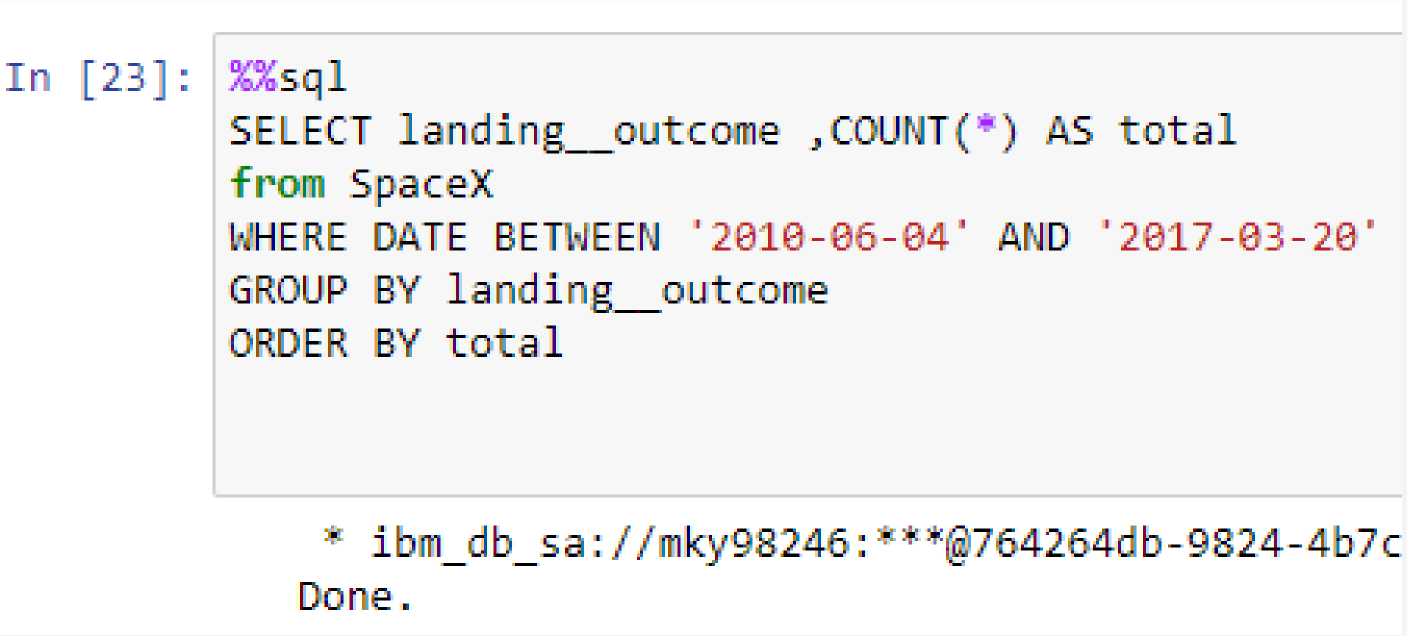
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

To obtain this result we used the WHERE clause, to filter the date, GROUP BY to summarize an then ORDER BY to sort them in ascending order



Out[23]:

landing__outcome	total
Precluded (drone ship)	1
Failure (parachute)	2
Uncontrolled (ocean)	2
Controlled (ocean)	3
Success (ground pad)	3
Failure (drone ship)	5
Success (drone ship)	5
No attempt	10

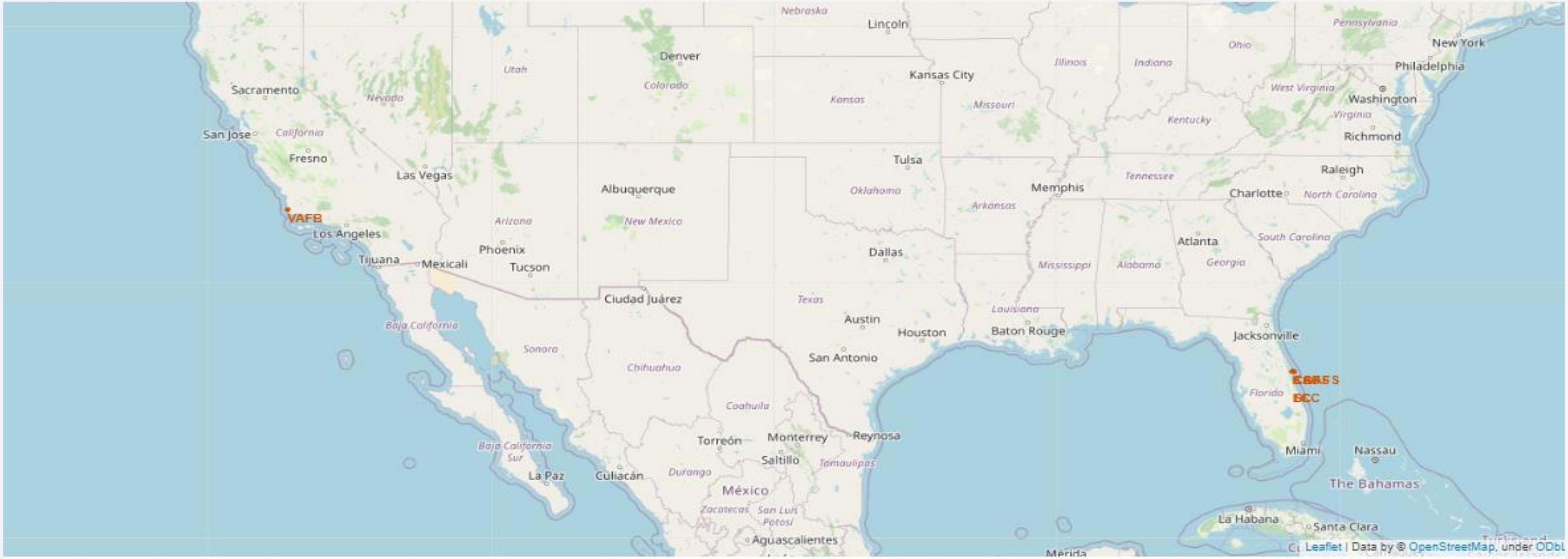


```
In [23]: %%sql
         SELECT landing__outcome ,COUNT(*) AS total
         from SpaceX
         WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
         GROUP BY landing__outcome
         ORDER BY total

 * ibm_db_sa://mky98246:***@764264db-9824-4b7c
Done.
```

Section 3

Launch Sites Proximities Analysis

Map of all launch sites



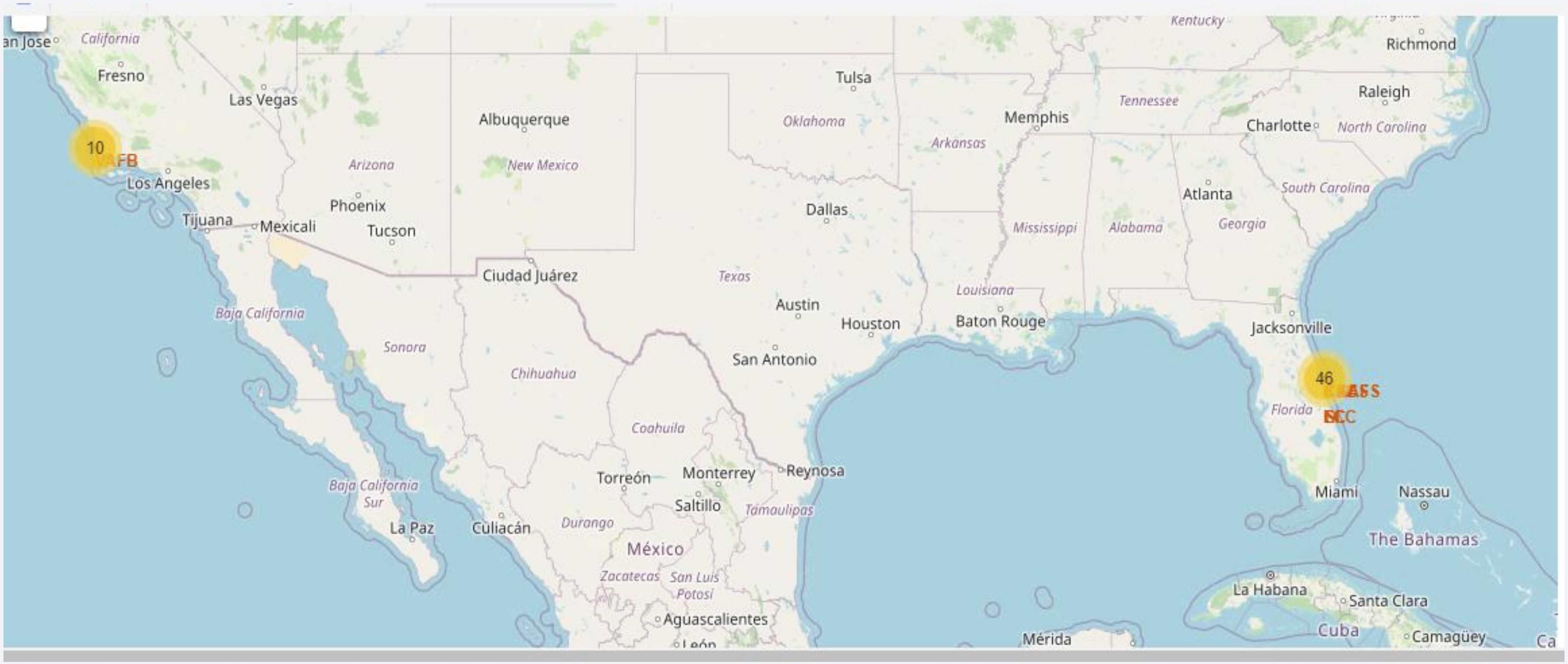
Map of all launch sites: Elements and findings

- All launching sites are closer to costal areas

- Three of them CCAFS LC-40, CCAFS SLC-40, KSC LC-39A are close to each other with the CCAs very close to each other.

- VAFB SLC-4E is the other side of the cost

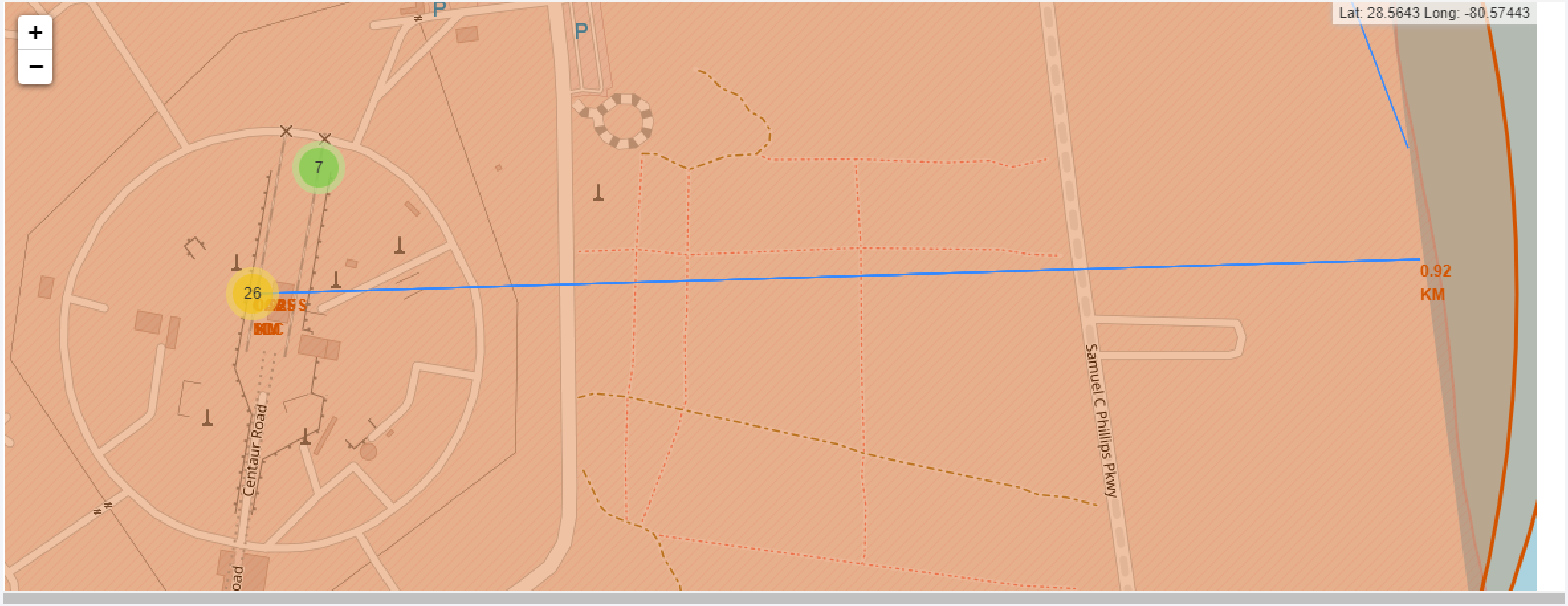
Successful/Failed launches per site



Important elements and findings on the screenshot

Site	Successful	Failed	Total	Success Rate
KSC LC-39A	10	3	13	76.9%
CCAFS SLC-40	7	19	26	26.92%
CCAFS LC-40	3	4	7	42.8%
VAFB SLC-4E	4	6	10	40.0%
Total	24	32	56	42.9

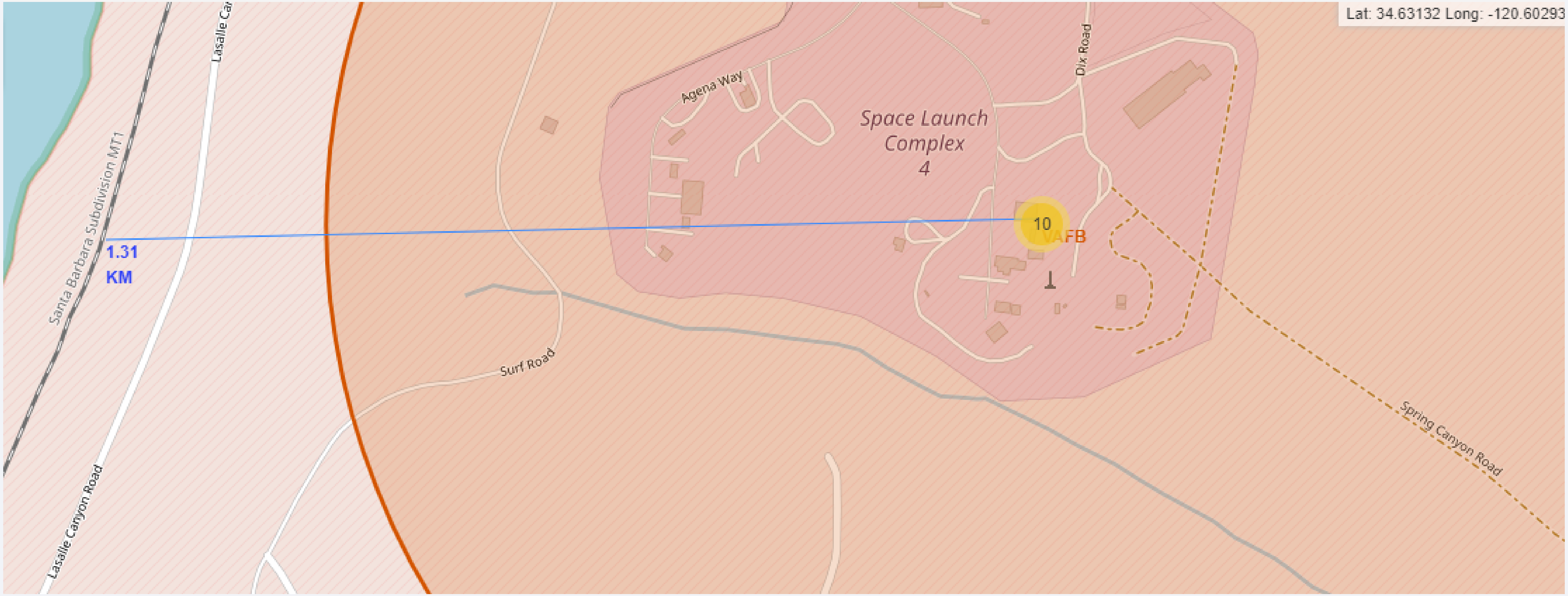
Distance of CCAFS SLC-40 to Coastline



Important elements and findings on the screenshot

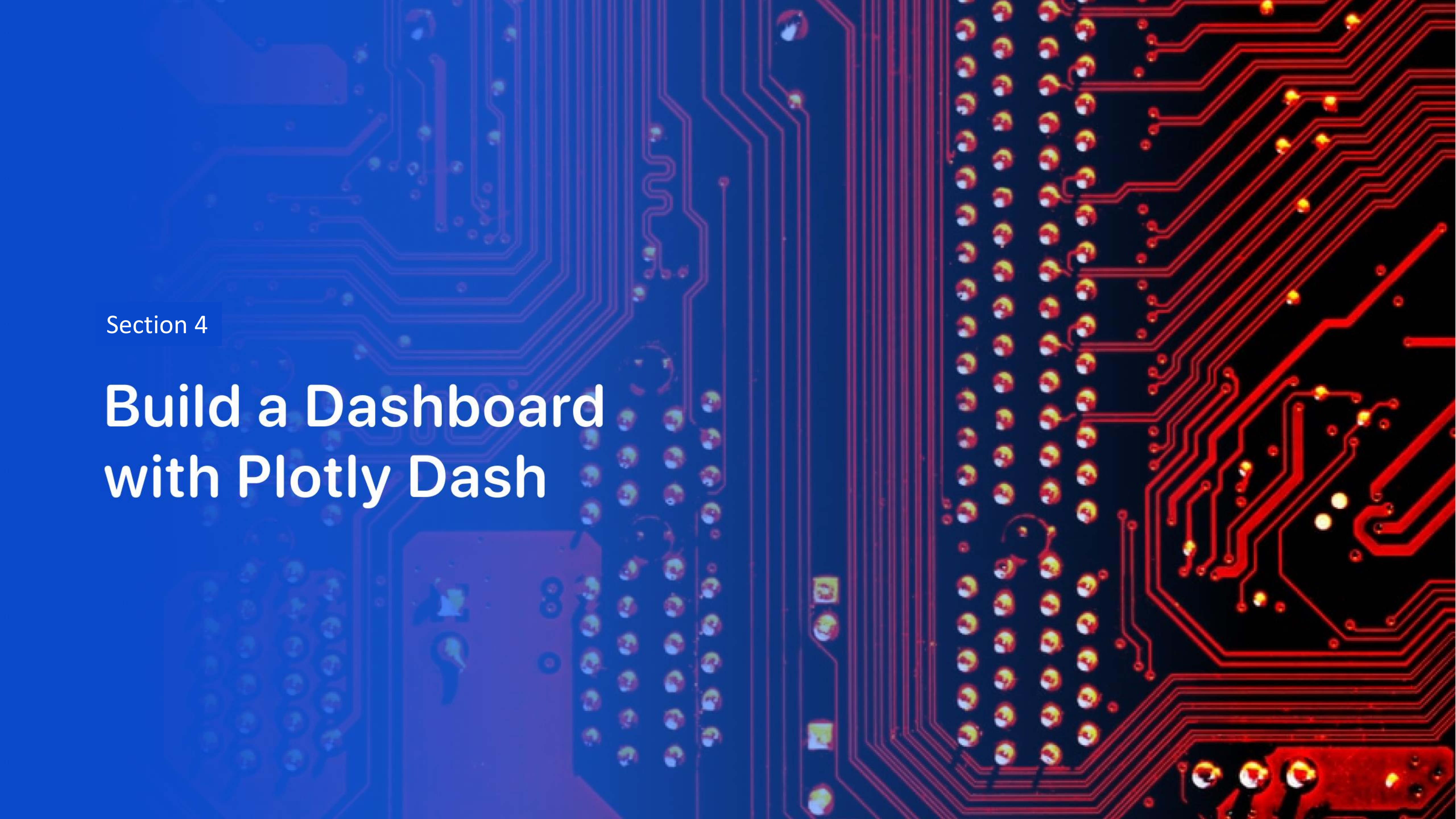
CCAFS SLC-40 is 0.92 kms from the coastline. According to Live Science, East Coast locations are desirable because any rockets leaving Earth's surface and traveling eastward get a boost from the Earth's west-to-east spin.

Distance from VAFB SLC-4E to Santa Rail



Important elements and findings on the screenshot

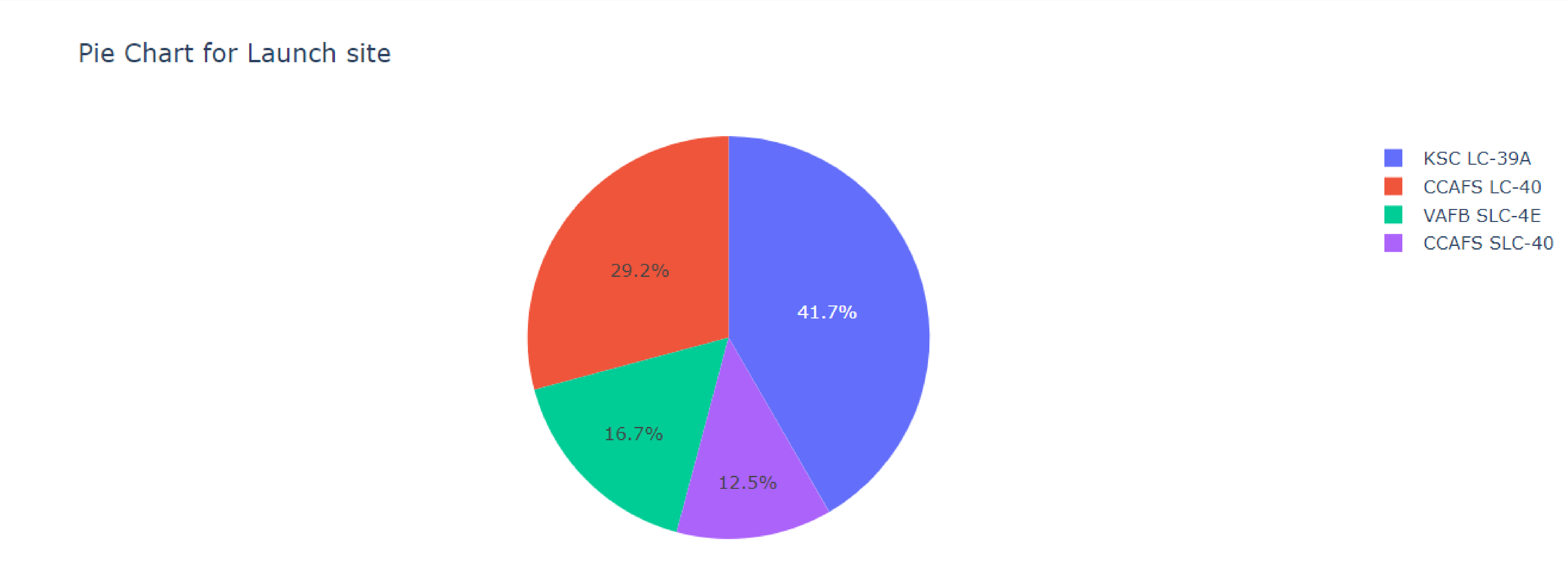
VAFB SLC-4E is 1.3 kms from railway line. These sites are near railway lines to make it easy and cheaper to transport them since they are heavy.



Section 4

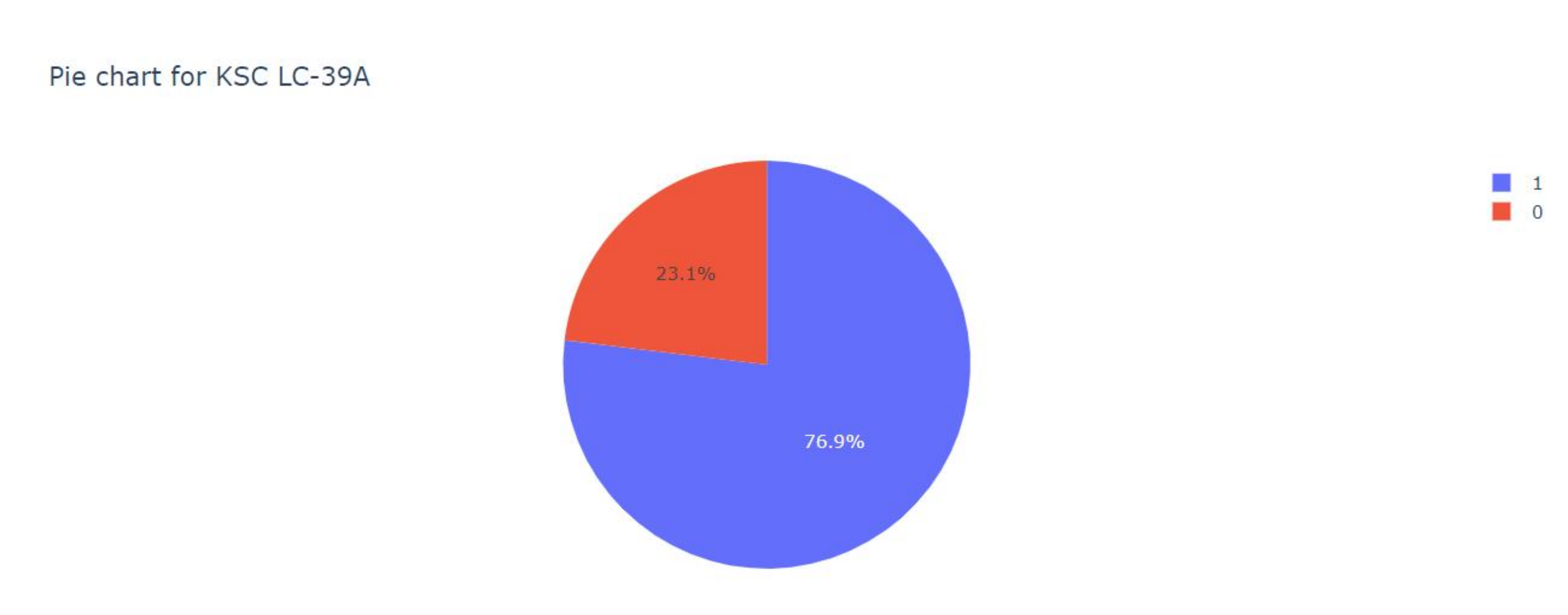
Build a Dashboard with Plotly Dash

Launch Success Count: All Sites



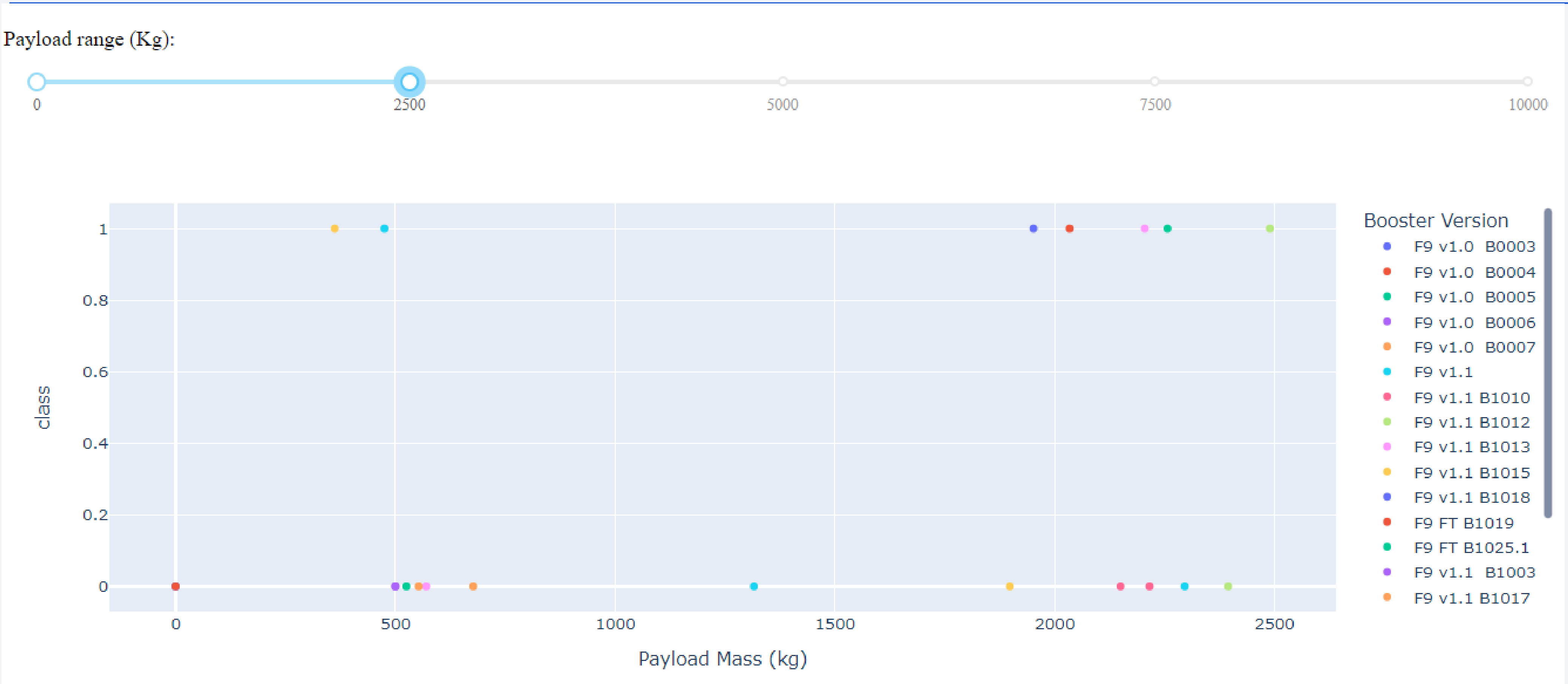
The Launch Site with the highest success rate is KSC LC – 39A. It contributes 41.7% of the total success missions for Space X

Launch site with highest launch success



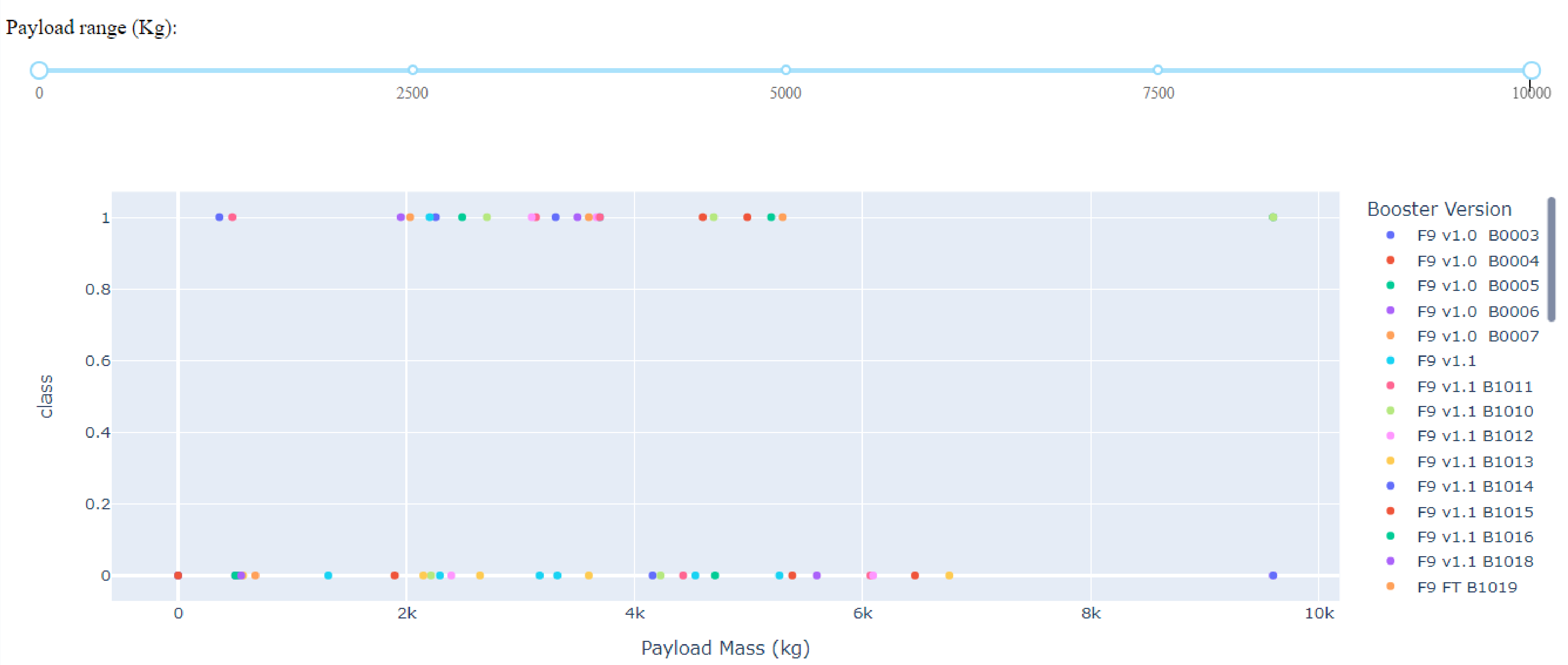
KSC LC -39A has a launch success rate of 76.9%

Scatter Plot: Payload Vs Launch Outcome: 2500 Kgs



- Payloads that are under 2500kgs have a success rate of 36.8%.

Scatter Plot: Payload Vs Launch Outcome: All



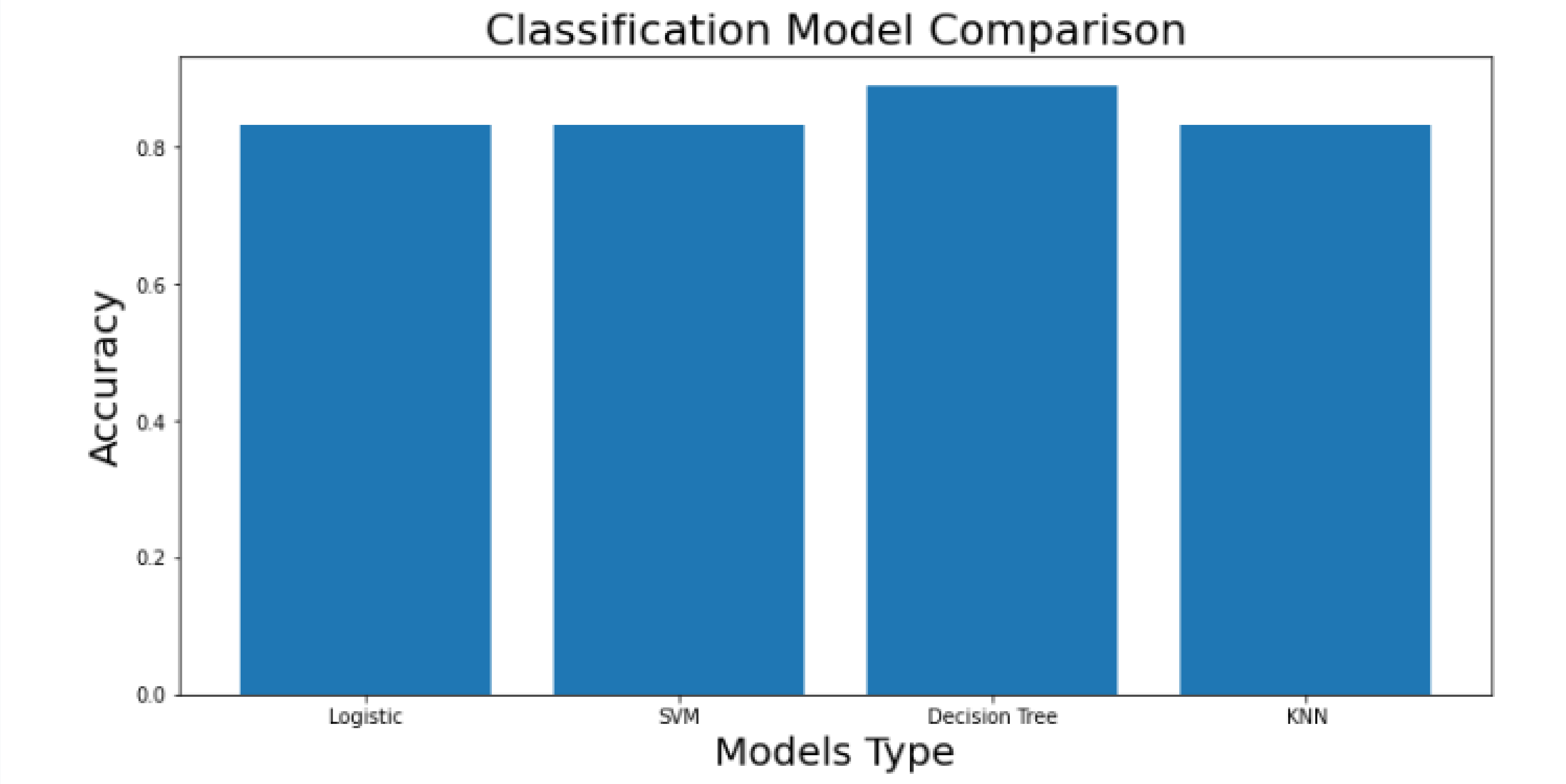
Most successful launch outcomes were between 2000-6000 kgs. 85% of all successful outcomes where within that range.

Section 5

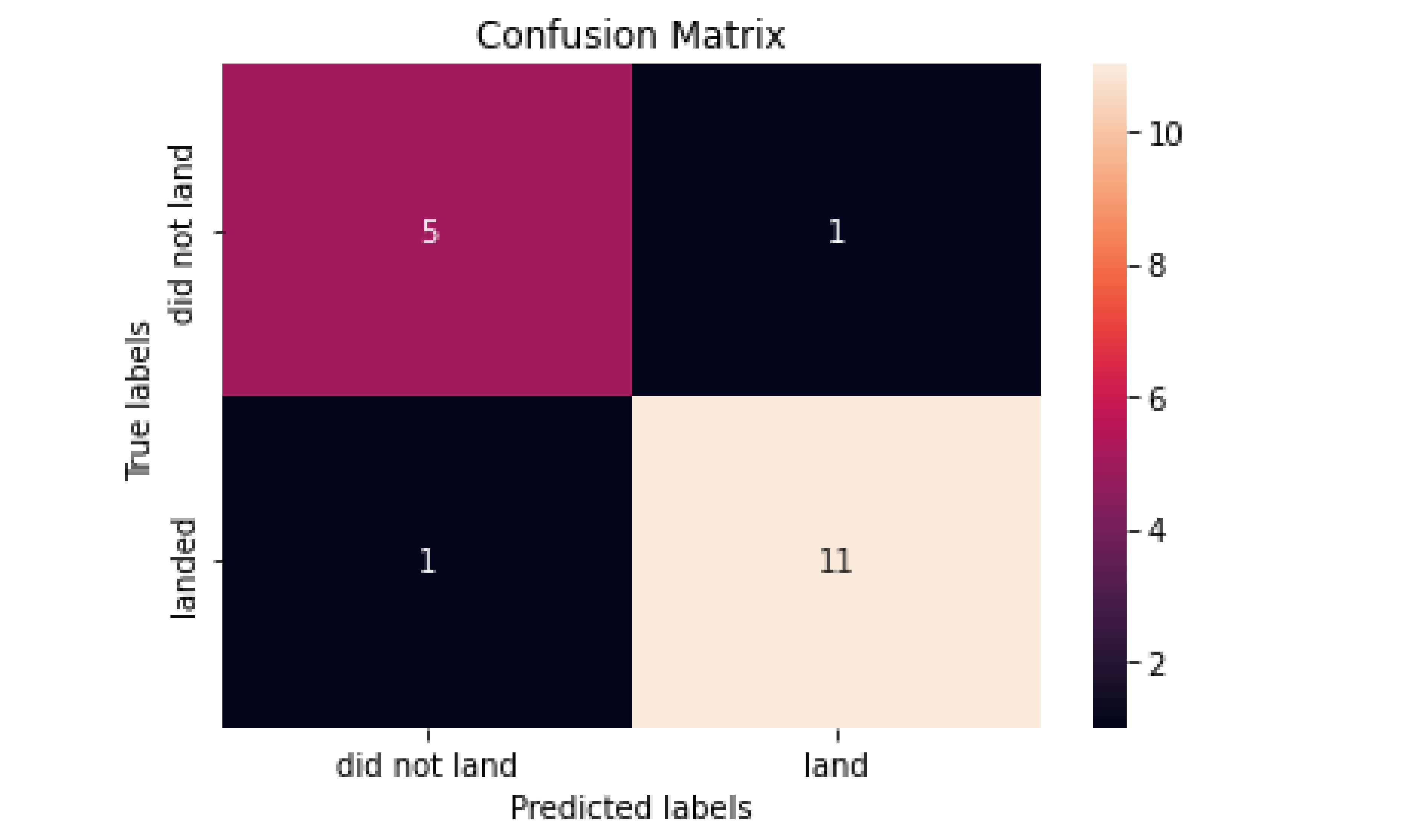
Predictive Analysis (Classification)

Classification Accuracy

The Decision Tree model highest classification accuracy of 85%



Confusion Matrix



Conclusions

- Wef found out tha the Decision Tree model giave us the highest accuracy in predicting launch outcome of a new launch mission. And whether the first stage will be reused.
- This will be beneficial in approximating the cost of he new mission
- Launching from site KSC LC-39A and VAFB SLC 4E will give us a beter chance of success and lowering of the costs. Also launches into Orbits ES-L1, GEO, HEO and SSO are mostly likely to be successful

References

1. Why do rockets launch from Florida?, Karen Rowan, Live Science Staff , May 30, 2020, https://www.livescience.com/32721-why-are-rockets-launched-from-florida.html

Appendix

1. https://github.com/Vinoro2002/IBM-Capestone-

Thank you!