IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yujie Wang
27/09/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data has been through Data Collection, Data Wrangling, EDA with Data Visualization, EDA with SQL and a series of process.
  - Use Plotly Dash to build the interactive visualization based on the Spacex data
  - Implement and evaluate four machine learning models for predictive analysis.
- Summary of all results
  - KSC LC-39A is the site which has the highest successful rate for launches.
  - The four models have the same accuracy which is 83.3%

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX. SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive.
- 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

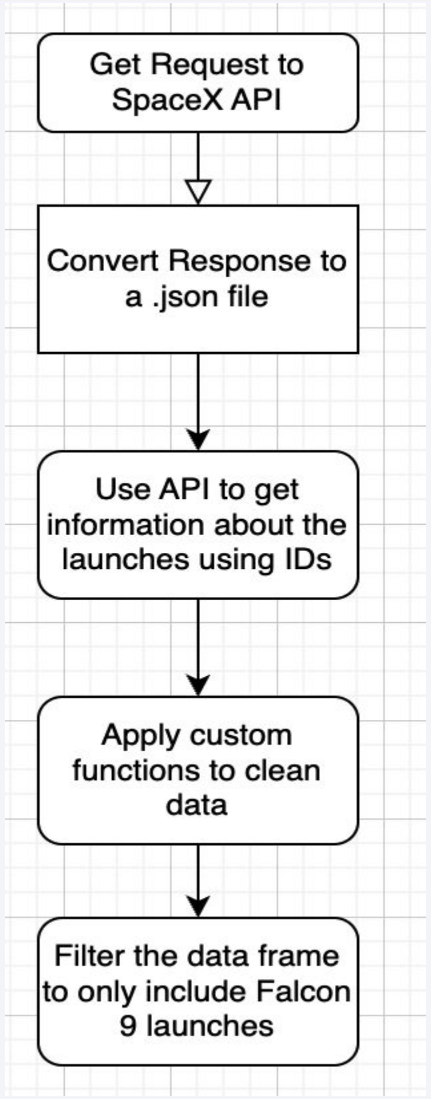
Methodology

Executive Summary

- Data collection methodology:
  - In this project, data is collected using two methods: Get data from SpaceX API and Use web scraping to collect Falcon 9 historical launch records.
- Perform data wrangling
  - Dealing with Missing Values and Create a landing outcome label using Python
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models: SVM, Decision Trees, Logistic Regression, KNN
  - Hyperparameter Selection: GridSearchCV
  - Evaluation Metrics: Accuracy

Data Collection – SpaceX API

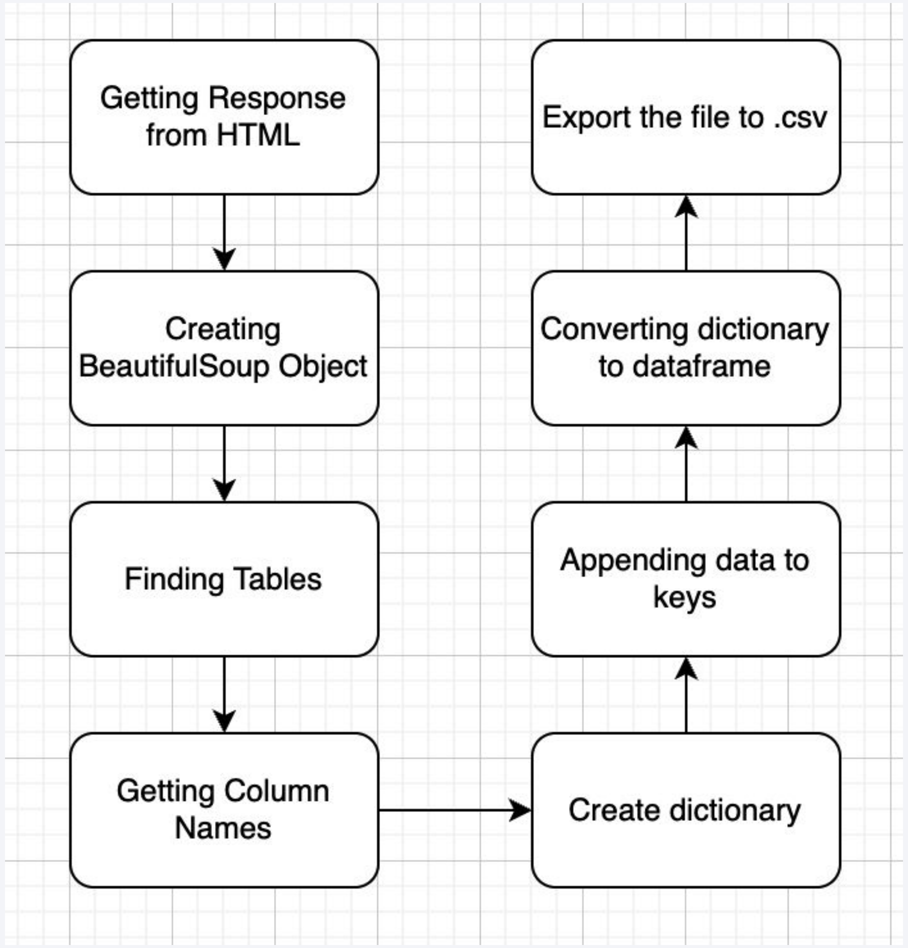
- The Steps:
  - Perform get request to SpaceX API
  - Convert the Json to Dataframe
  - Use API to get information about the launches using IDs
  - Apply custom functions to clean data
  - Filter the Dataframe to only include Falcon 9 launches
- The Github URL for this stage: [Data Collection- SpaceX API]()


Get Request to SpaceX API
Convert Response to a .json file
Use API to get information about the launches using IDs
Apply custom functions to clean data
Filter the data frame to only include Falcon 9 launches

Data Collection - Scraping

- The Steps:
  - Request the Falcon 9 Luanch Wiki page
  - Extract data from the HTML table header
  - Parse the launch HTML tables

- The Github URL for this stage: [Data Collection -Scraping]



Data Wrangling

- **Processing the data**:
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome of the orbits
  - Create a landing outcome label from Outcome column
- **The Github URL for this stage**: Data Wrangling

EDA with Data Visualization

- Use Scatter plot to explore the relationship:
    - between Flight Number and Launch Site
    - between Payload and Launch Site
    - between Flight Number and Orbit type
    - between Payload and Orbit type
- Use Bar plot to check is there are any relationship between success rate and orbit type
- Use Line plot to visualize the launch success yearly trend
- The Github URL for this stage: EDA with Data Visualization

EDA with SQL

- SELECT query and DISTINCT - the unique launch sites in the space mission
- WHERE clause, LIKE, and LIMIT -5 records where launch sites begin with 'CCA'
- SUM() and LIKE - total payload mass carried by boosters launched by NASA(CRS)
- AVG(() and LIKE - average payload mass carried by booster version F9 v1.1
- Min() to list the date when the first successful landing outcome was achieved
- WHERE clause and LIKE - list the names of the boosters which have success in drone ship
- COUNT() and GROUP BY clause - total number of successful and failure mission outcomes
- MAX() and Subquery - List the names of the booster_versions
- Substr() and LIKE- List the records for the months in year 2015.
- Group by and Order by - Rank the out of landing outcomes
- **The Github URL for this stage**: EDA with SQL

Build an Interactive Map with Folium

- To mark all launch sites on a map, use folium.Circle to add a highlighted circle area with a text label on a specific coo rdinate (folium.Marker)
- To mark the success or failed launches for each site on the map, create markers for launch records, Markcluster is a good way to simplify a map containing many markers having the same coordinate.
- To calculate the distance between a launch sit to its proximities
  - Add MousePosition to get coordinate
  - Use folium.Marker to show the distance from the coastline, railway, highway and city to the launch site.
  - Use PolyLine to drae the straight line from launch site to those signs listed above.
- The Github URL for this stage: Interactive Map with Folium

Build a Dashboard with Plotly Dash

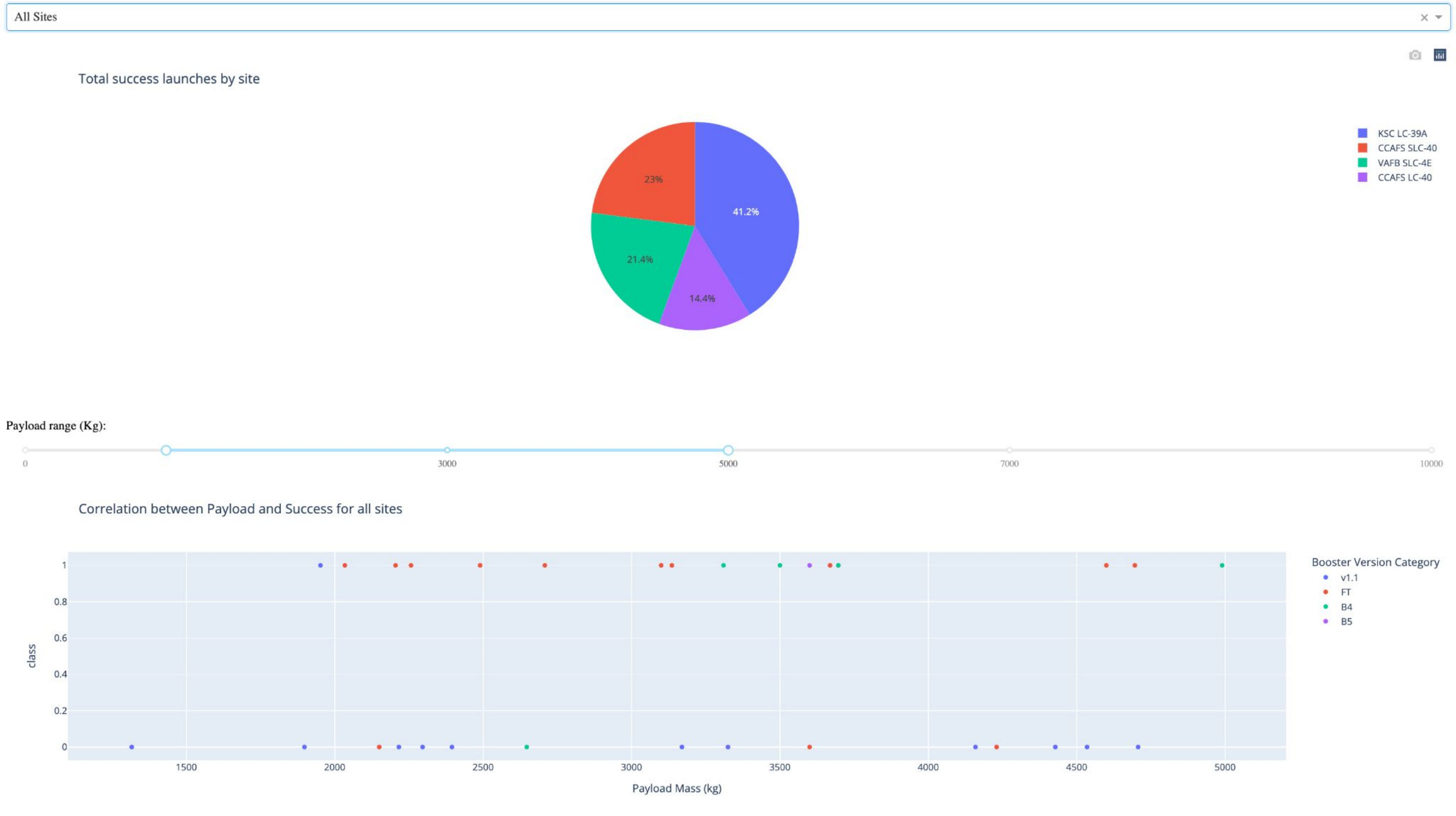
- Dash board and Plotly library were used to create the interactive visualization online. HTML was used to generate the web page.
- Pandas was used to load the data and generate the dataframe
- Drop-down was added as the input component to select the launch site.
- Pie chart was used to indicate the ratio of the successful launches.
- Rangeslider was added to narrow down the scale of payload mass.
- Scatter plot was added to show the relationship between the payload mass the success launches, with the colorful markers standing for different booster version.
- The Github URL for this stage: Build a Dashboard with Plotly Dash

Predictive Analysis (Classification)

- Build the Model
  - Load the Data and create the label
  - Split the data to the train set and test set
  - Normalize the train dataset
  - Create the Logistic Regression, SVM, Decision Trees, KNN model
  - Train the model (fit)
  - Test the data
- Evaluation Metrics: Accuracy and confusion matrix
- Hyperparameter Selection: GridSearch cross validation
- The Github URL for this stage: Predictive Analysis

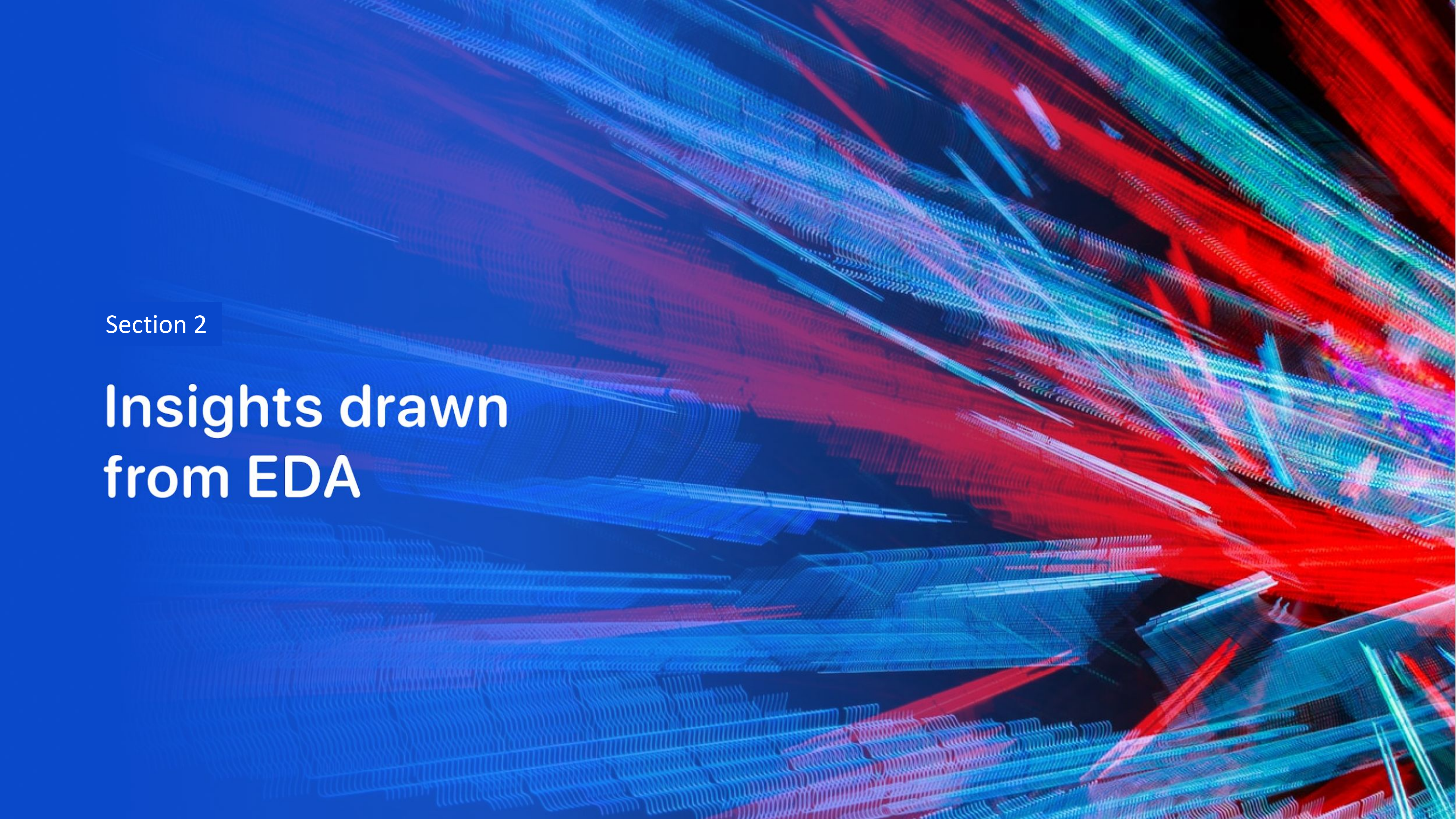
Results

- The number of launches from CCAFS SLC-40 are significantly higher than the launches from other sites.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).
- Logistic Regression, SVM, Decision Tree and KNN all has the same accuracy on the same dataset, which is 83.3%



All Sites

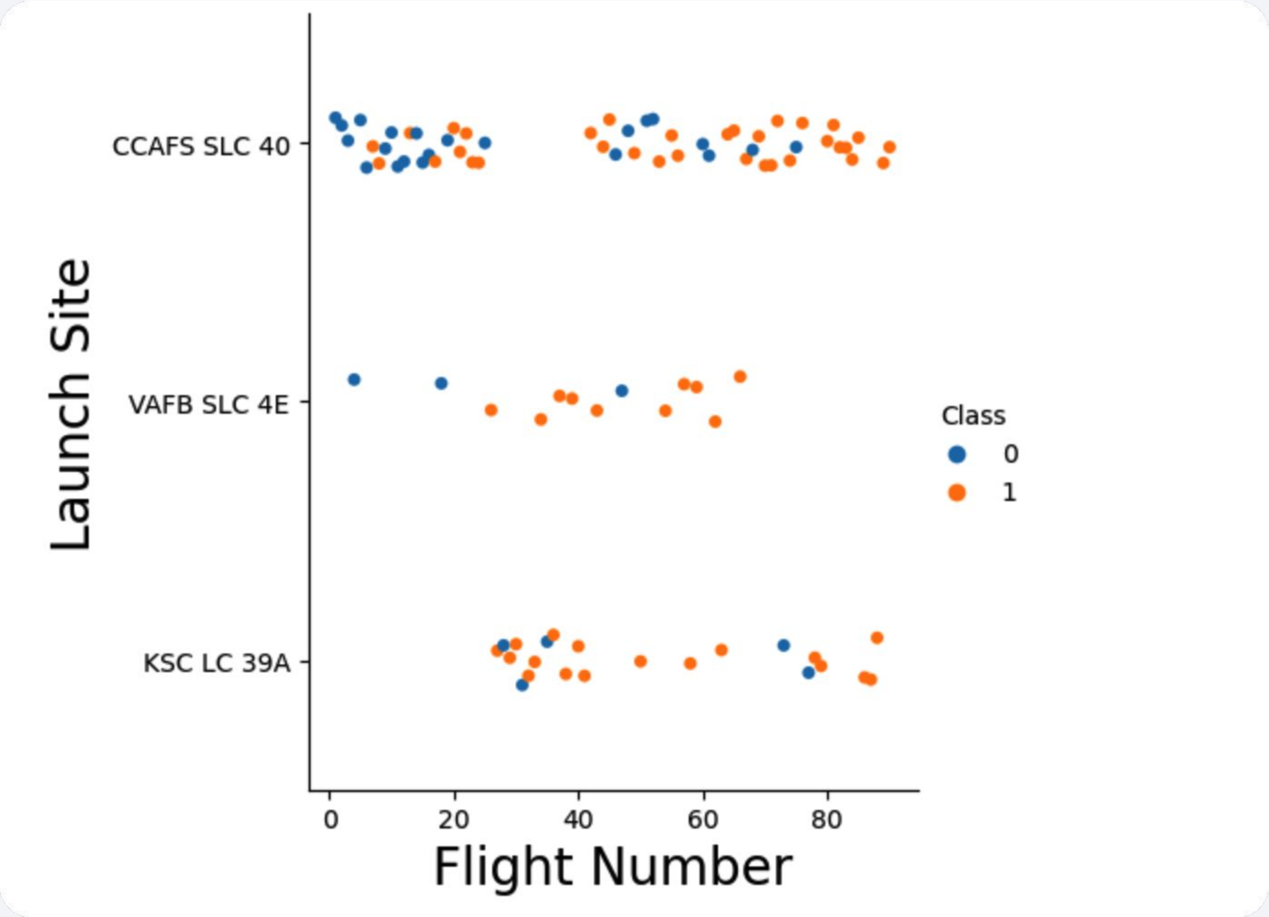
Payload range (Kg):



Section 2

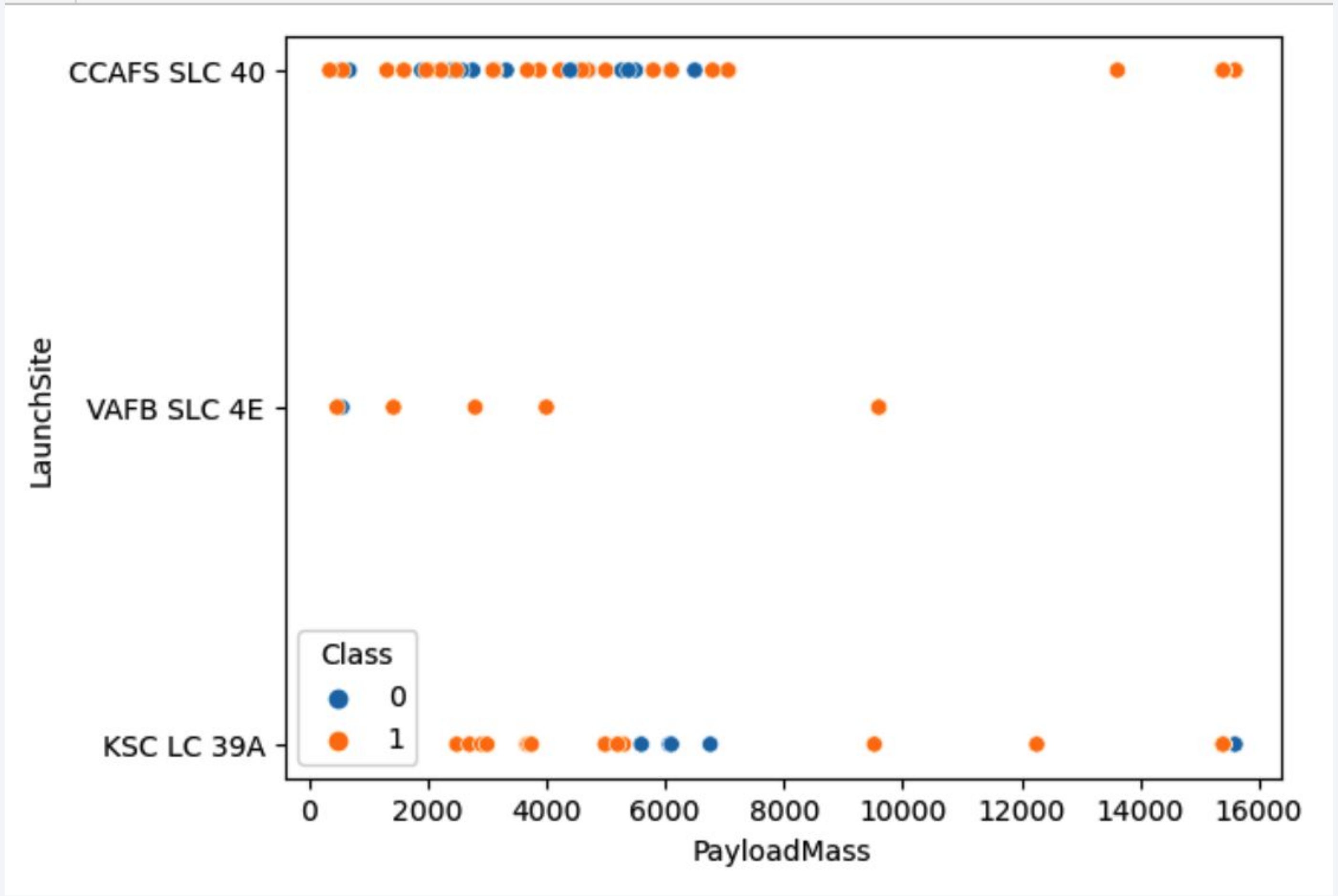
Insights drawn from EDA

Flight Number vs. Launch Site



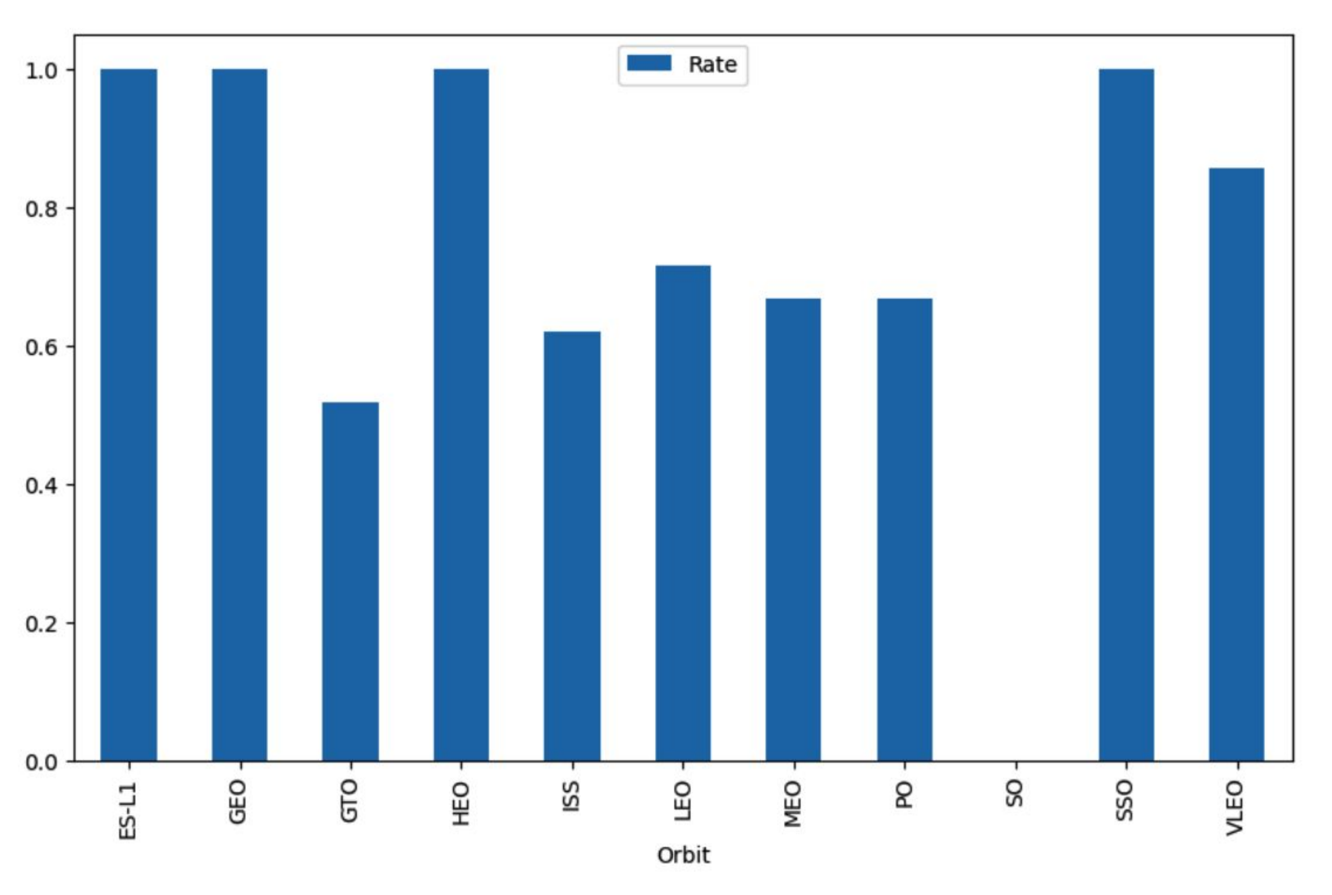
The number of launches from CCAFS SLC-40 are significantly higher than the launches from other sites.

Payload vs. Launch Site



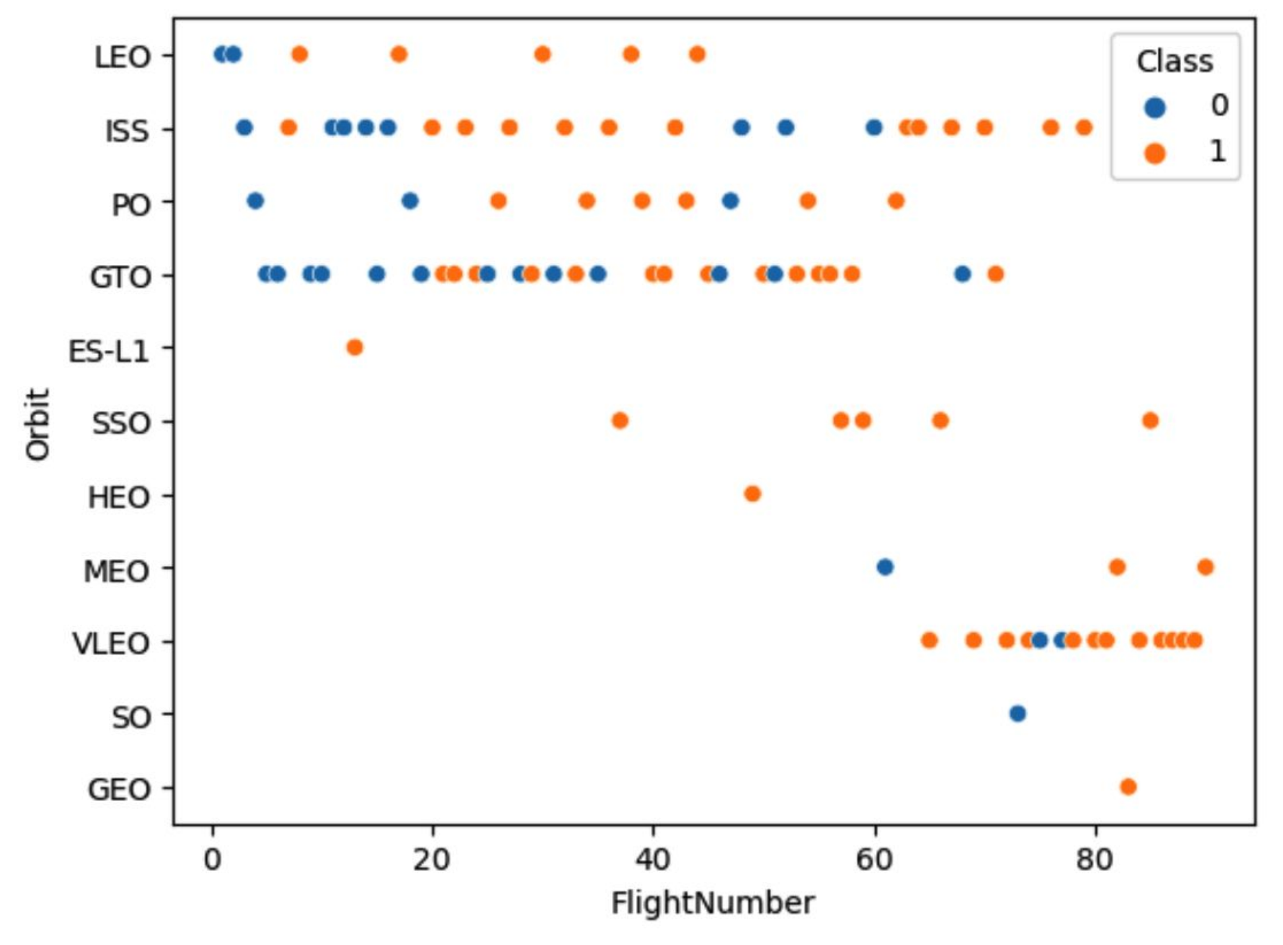
For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type



The orbit types of ES-L1, GEO, HEO and SSO achieved the highest success rate while SO didn’t success even once.

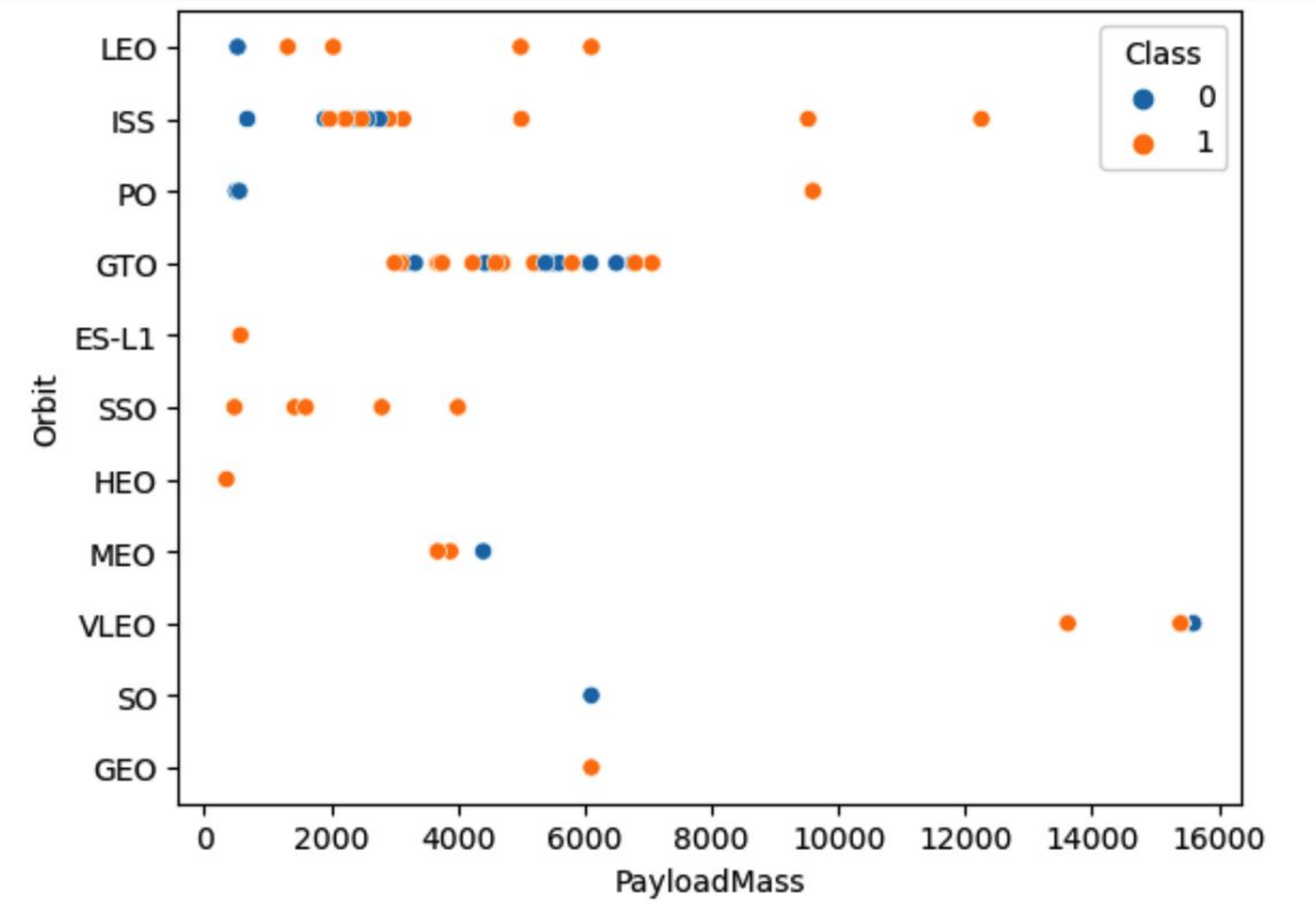
Flight Number vs. Orbit Type



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.

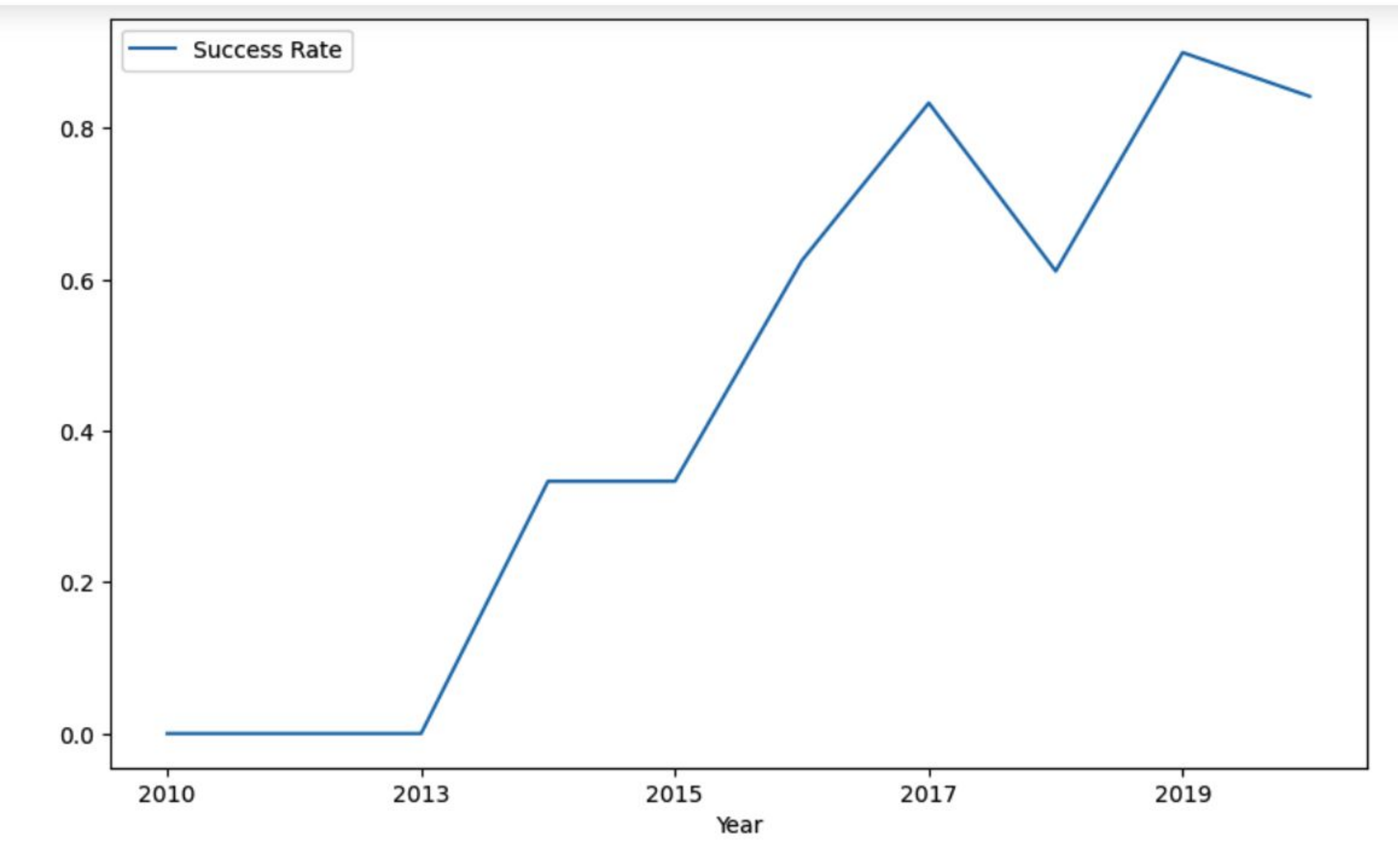
Payload vs. Orbit Type

- 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.



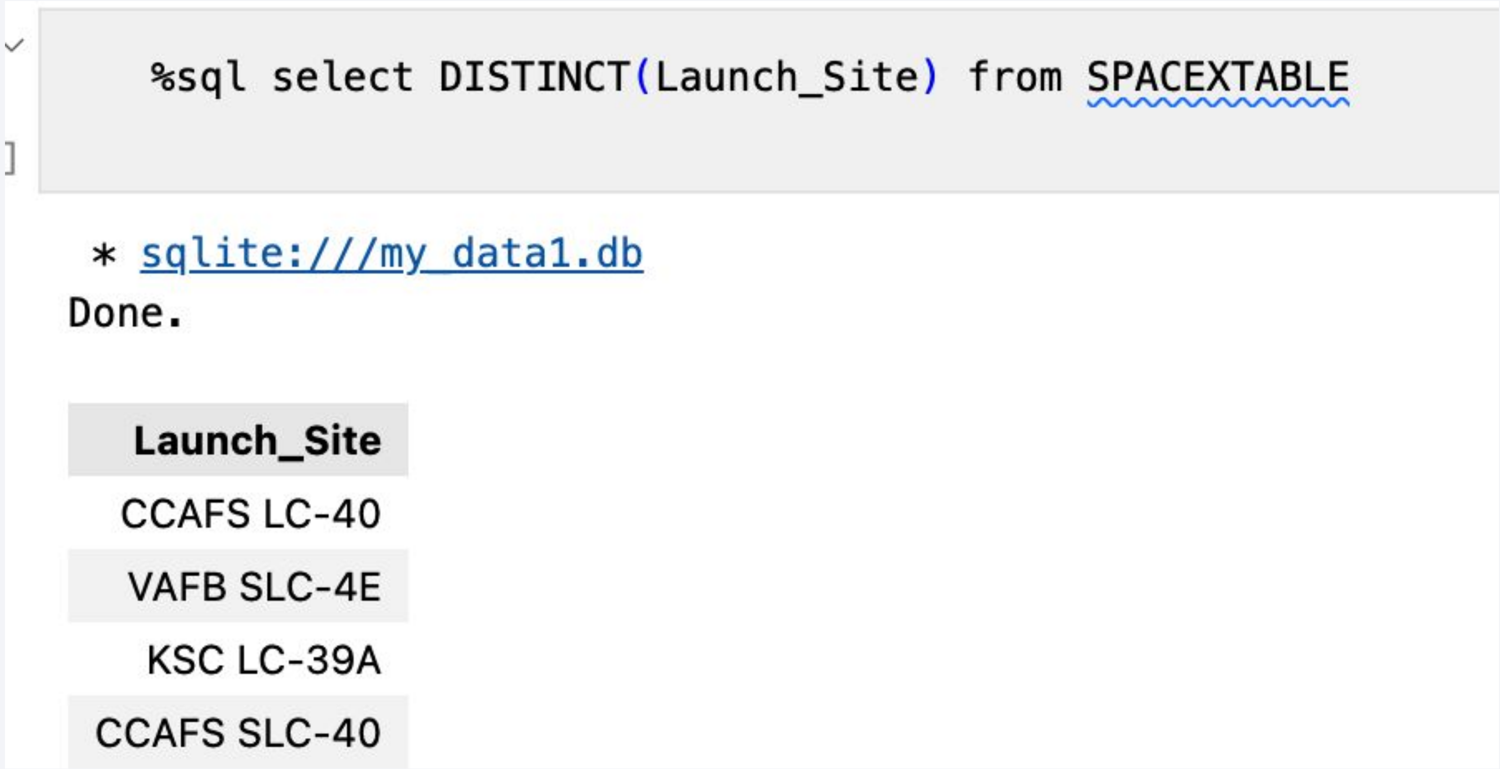
Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites



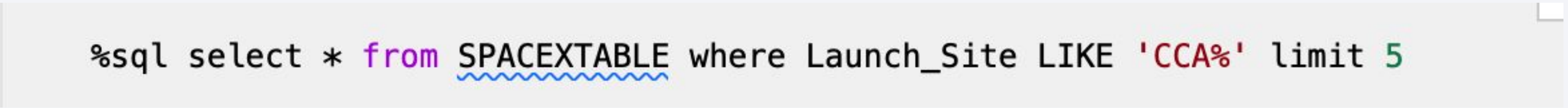
```
%sql select DISTINCT(Launch_Site) from SPACEXTABLE
```

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

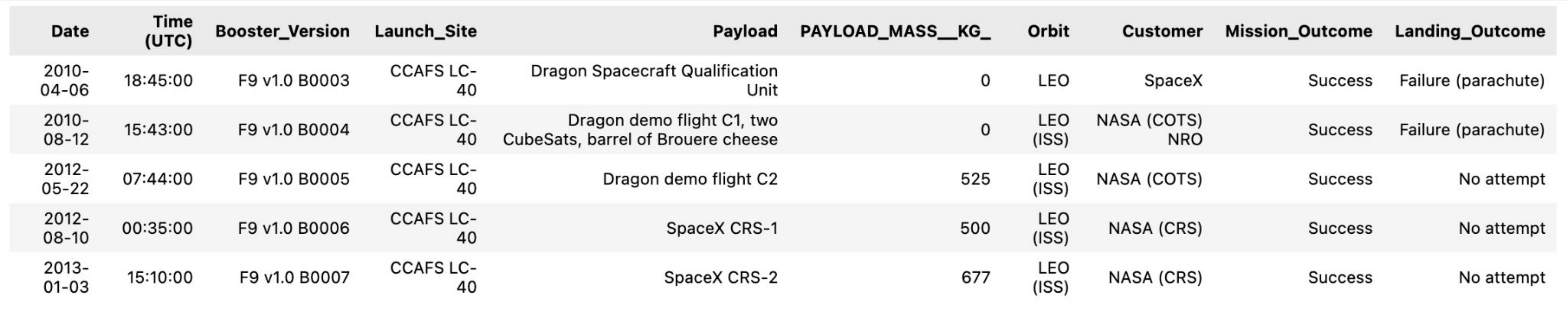
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

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



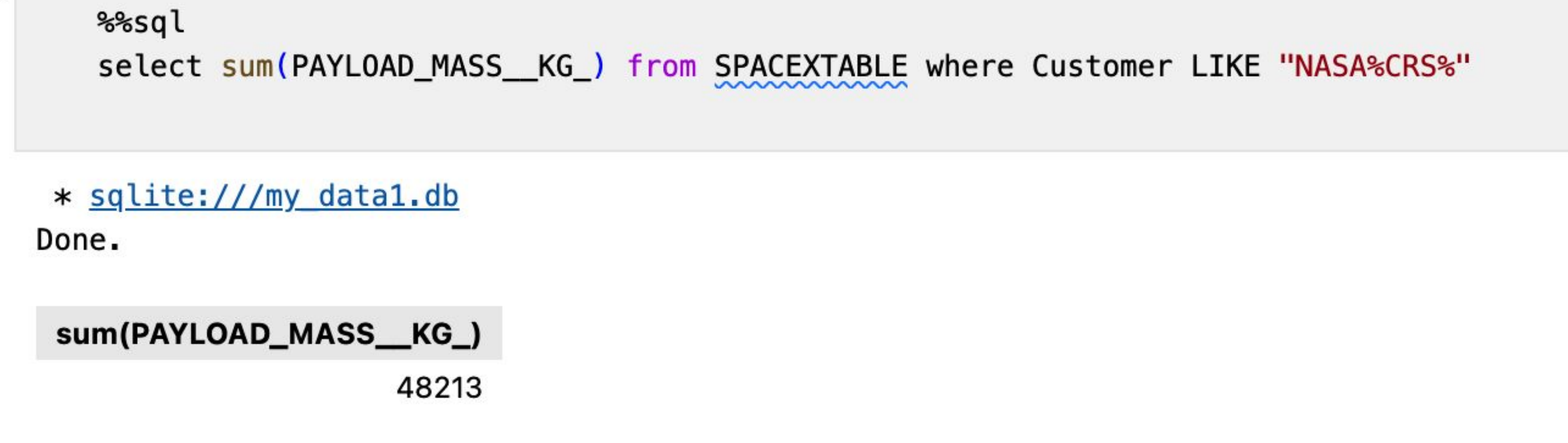
```
%sql select * from SPACEXTABLE where Launch_Site LIKE 'CCA%' limit 5
```



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA



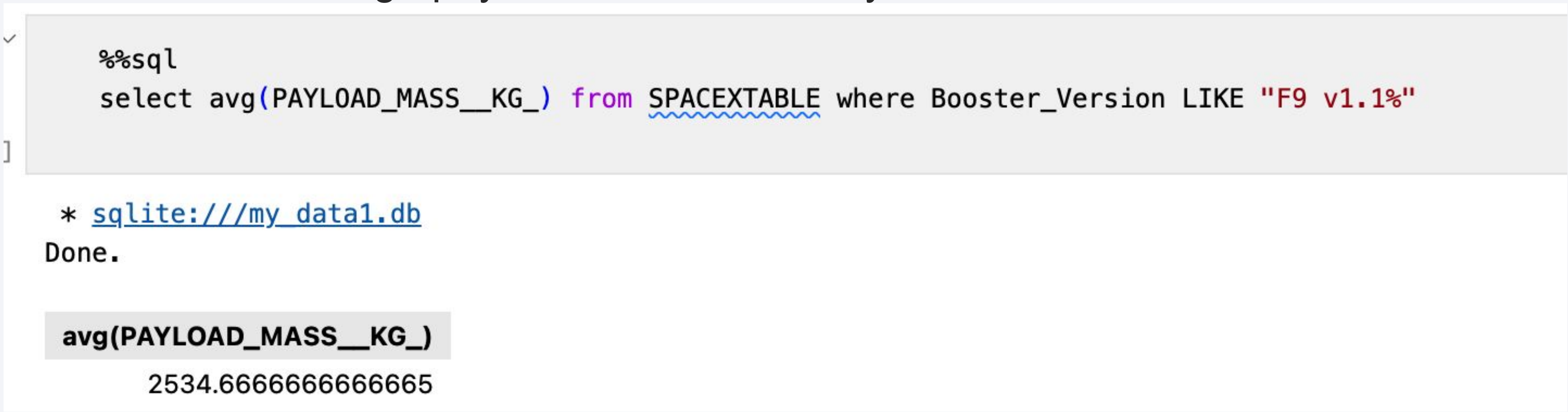
```
%%sql
select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer LIKE "NASA%CRS%"
```

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

sum(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

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



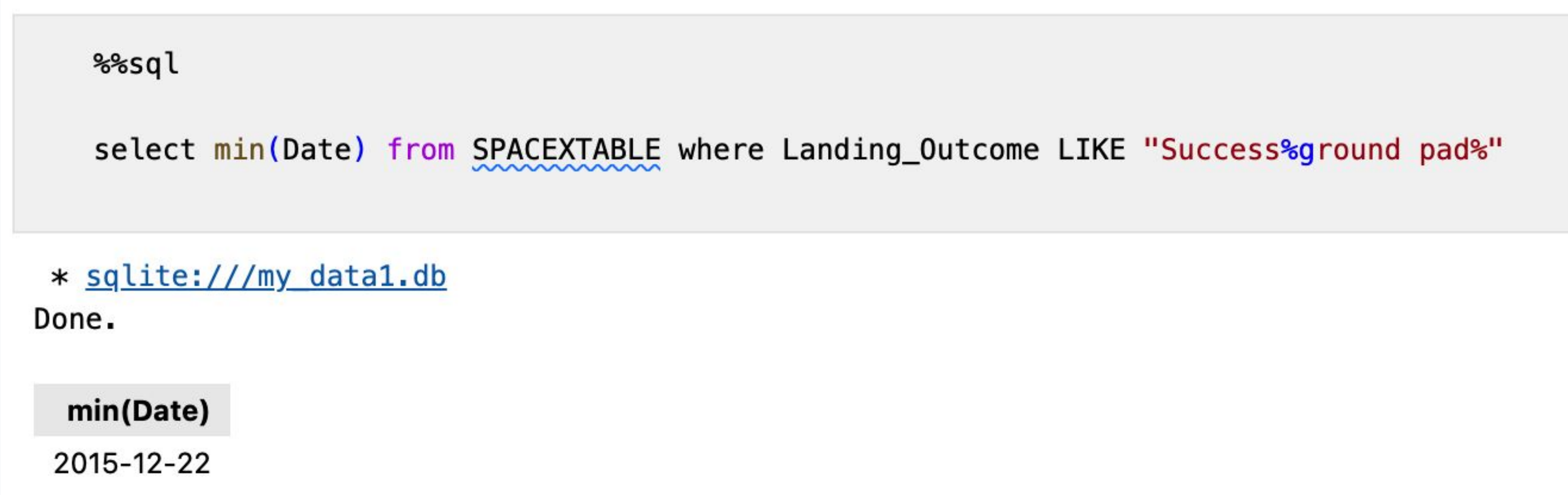
```
%%sql
select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version LIKE "F9 v1.1%"
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

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

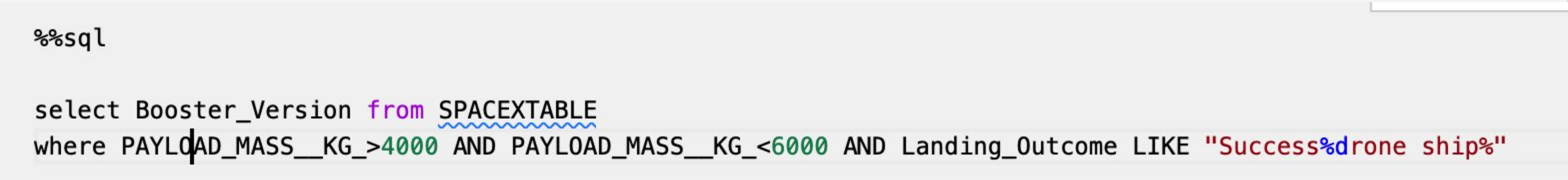


```
%%sql

select min(Date) from SPACEXTABLE where Landing_Outcome LIKE "Success%ground pad%"
```

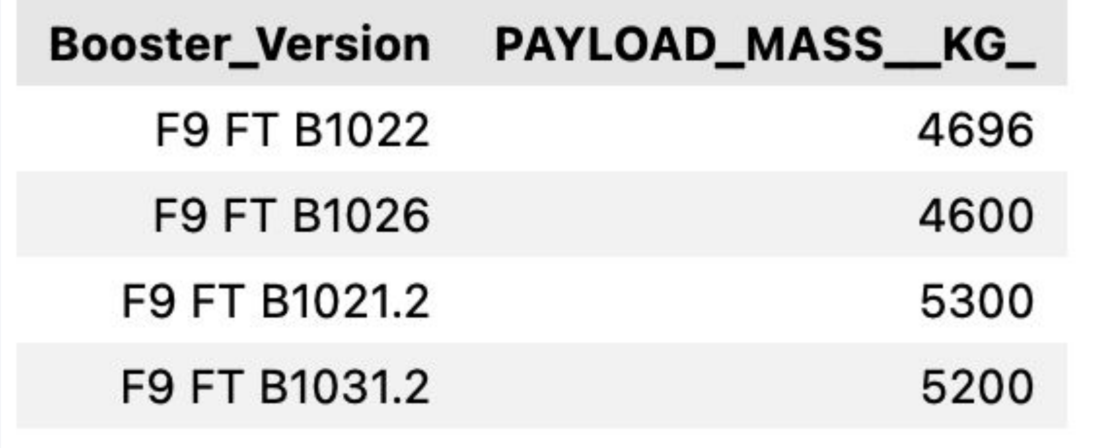
```
 * sqlite:///my_data1.db
Done.

 min(Date)
 2015-12-22
```

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



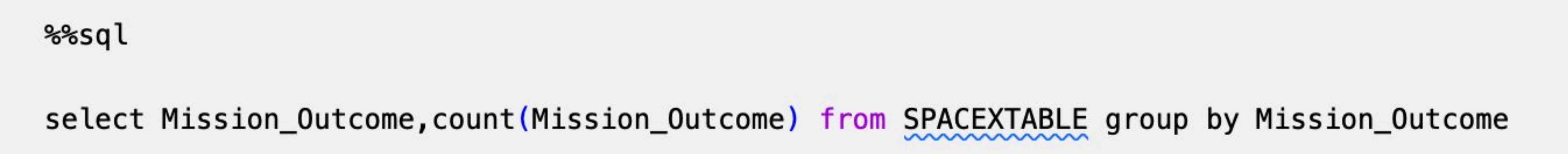
```
%%sql

select Booster_Version from SPACEXTABLE
where PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000 AND Landing_Outcome LIKE "Success%drone ship%"
```



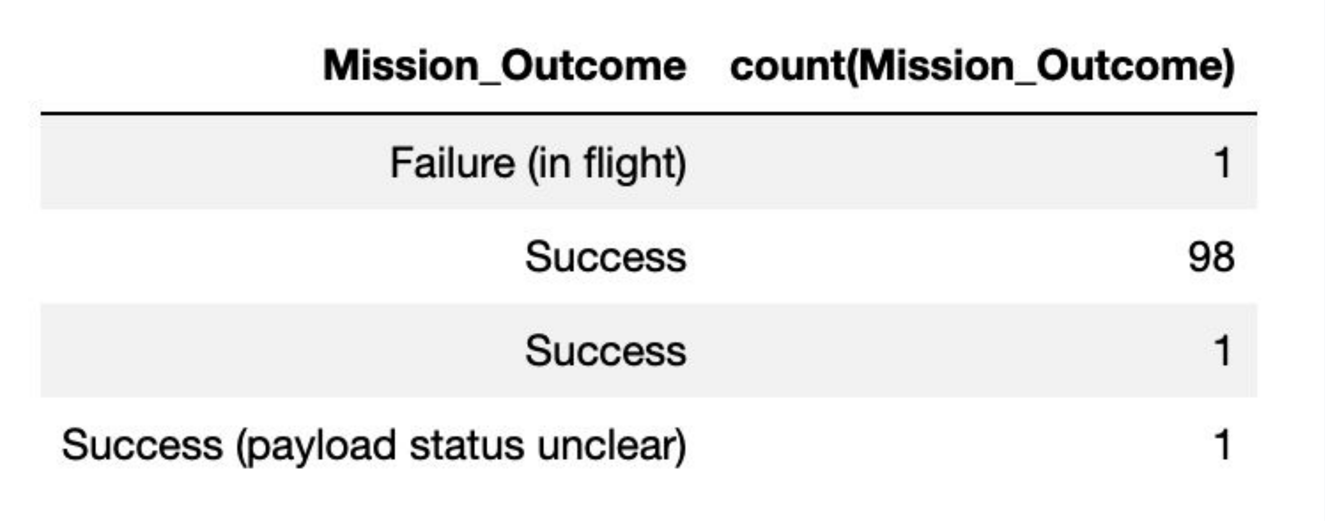
Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



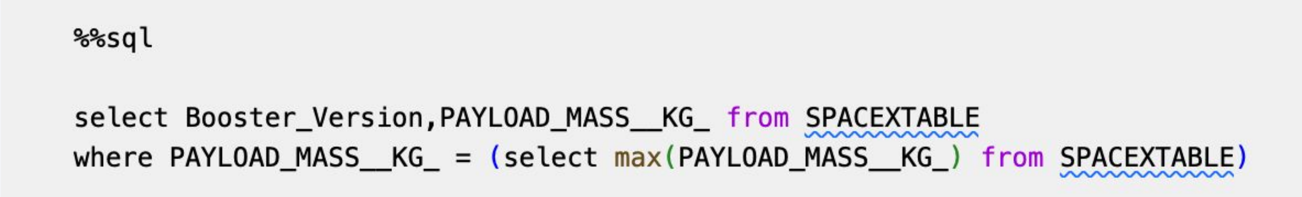
```
%%sql

select Mission_Outcome,count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```



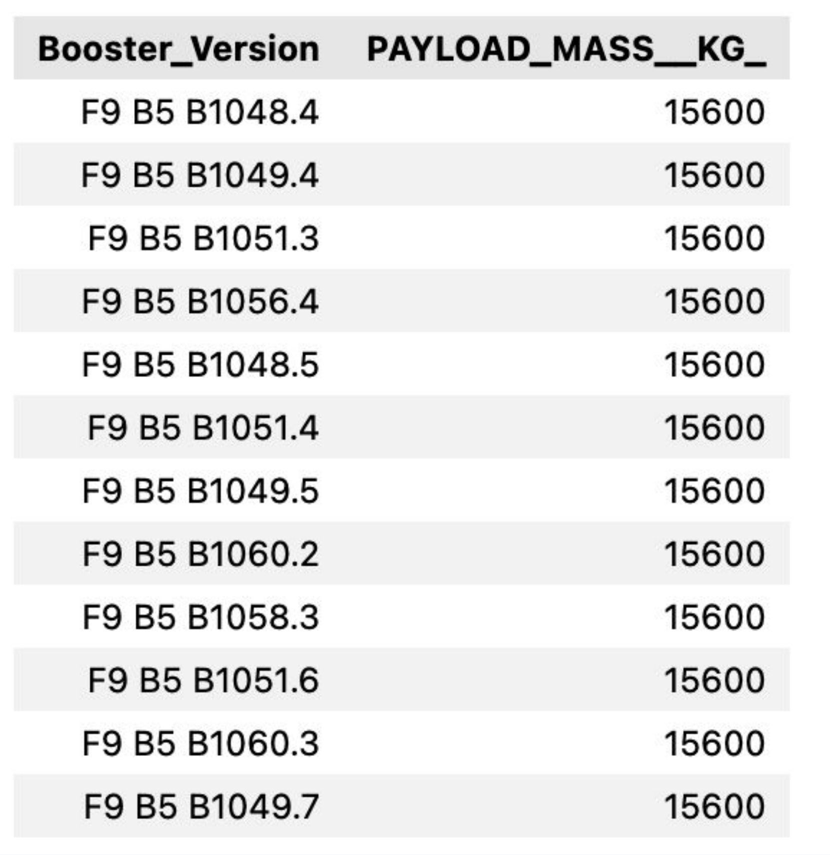
Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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



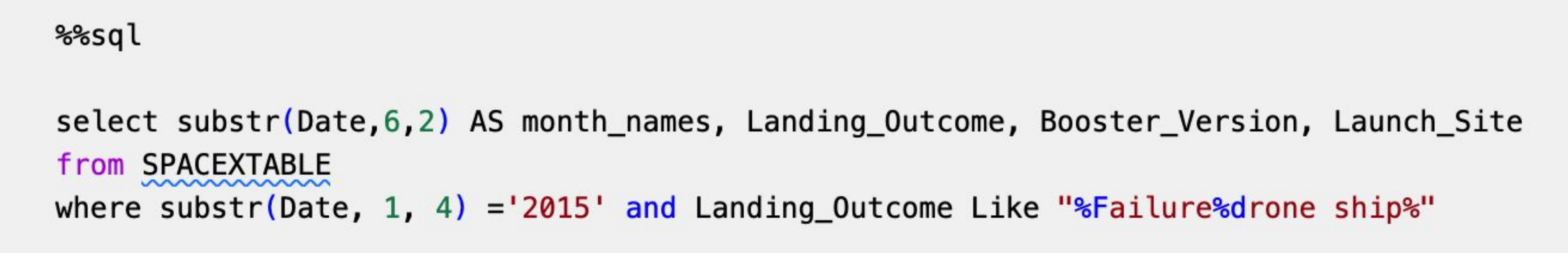
```
%%sql

select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```



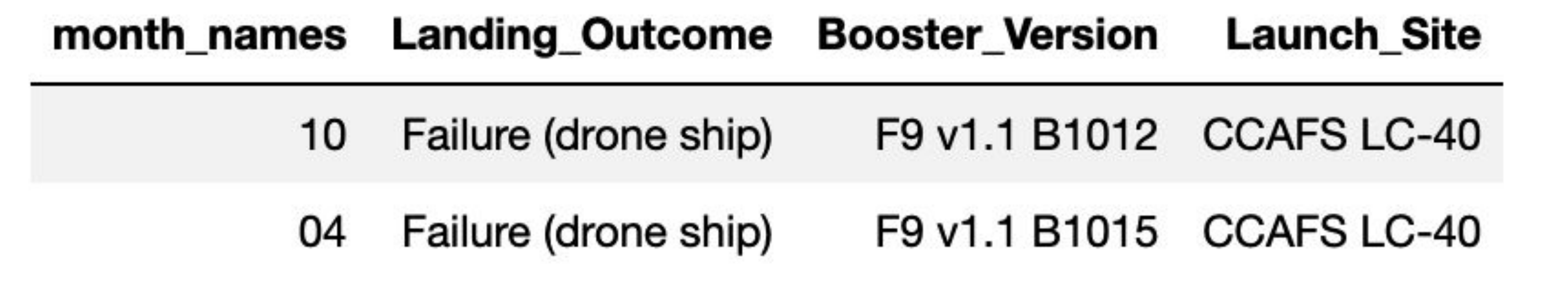
Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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



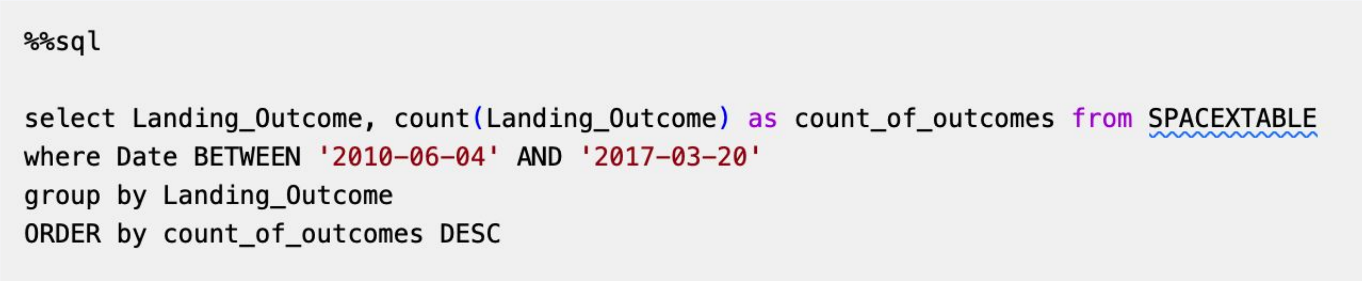
```
%%sql

select substr(Date,6,2) AS month_names, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTABLE
where substr(Date, 1, 4) ='2015' and Landing_Outcome Like "%Failure%drone ship%"
```



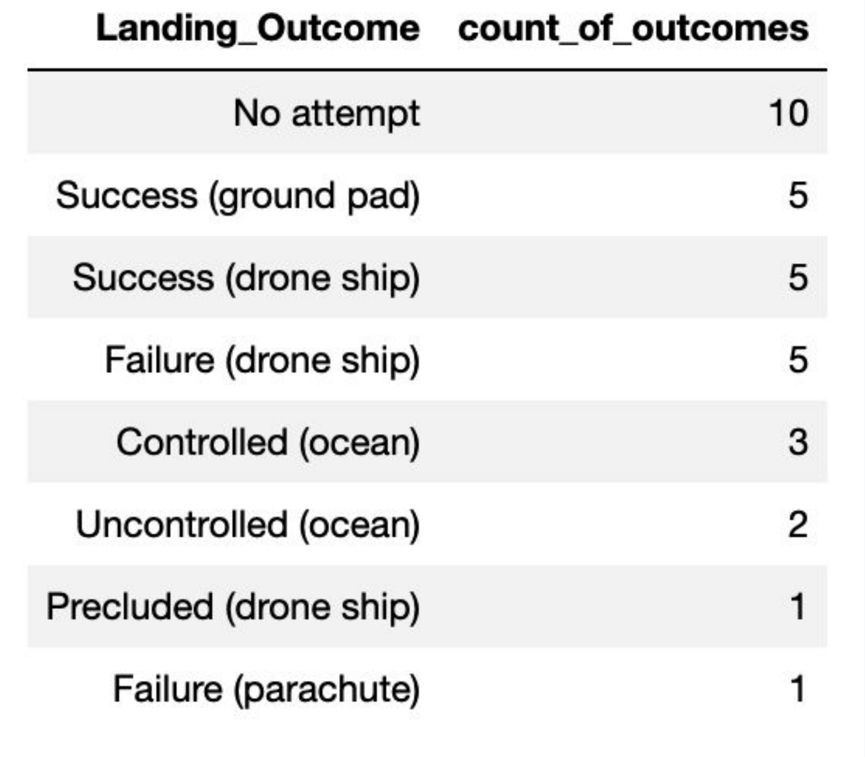
month_names	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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



```
%%sql

select Landing_Outcome, count(Landing_Outcome) as count_of_outcomes from SPACEXTABLE
where Date BETWEEN '2010-06-04' AND '2017-03-20'
group by Landing_Outcome
ORDER by count_of_outcomes DESC
```

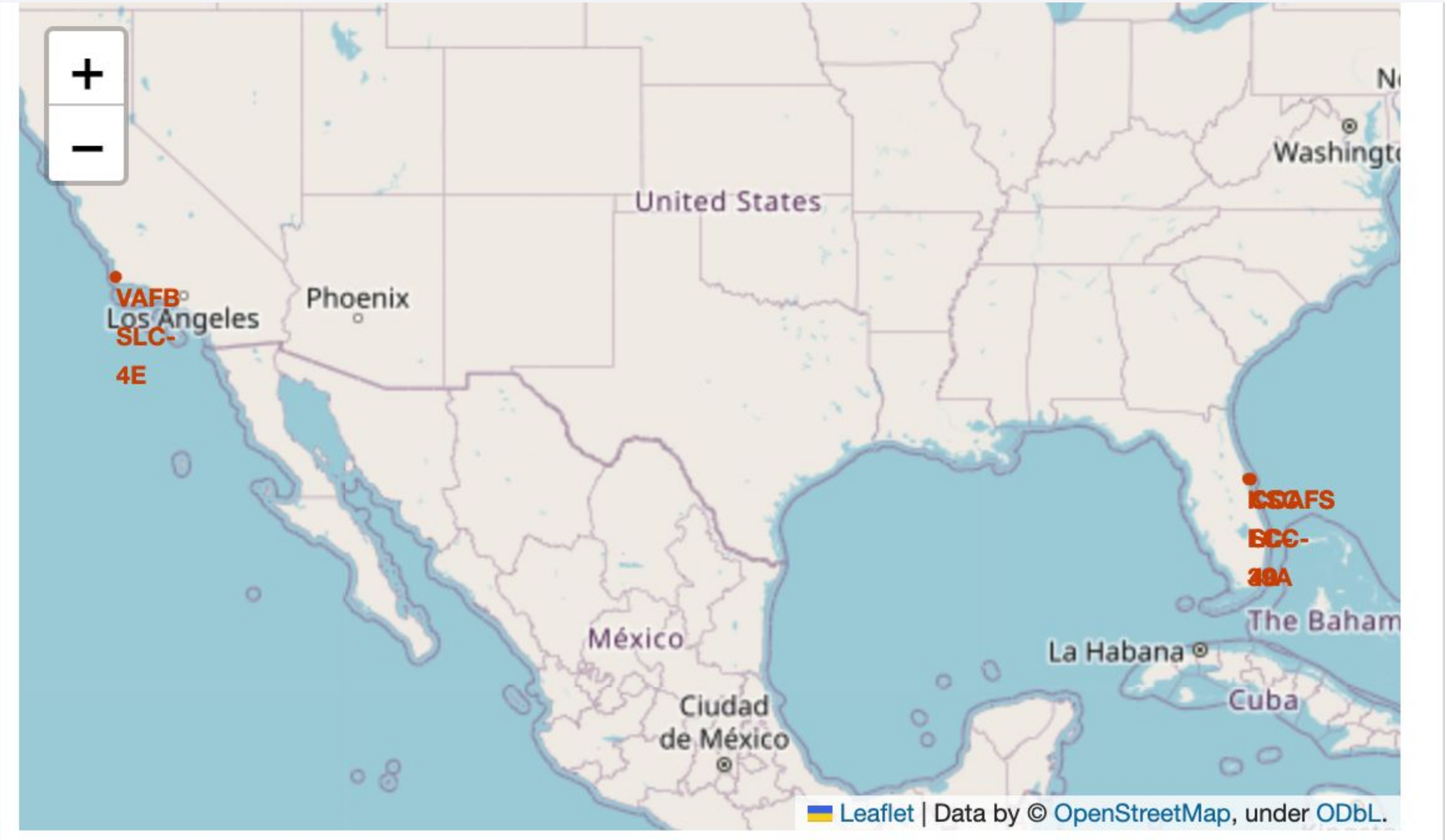


Landing_Outcome	count_of_outcomes
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

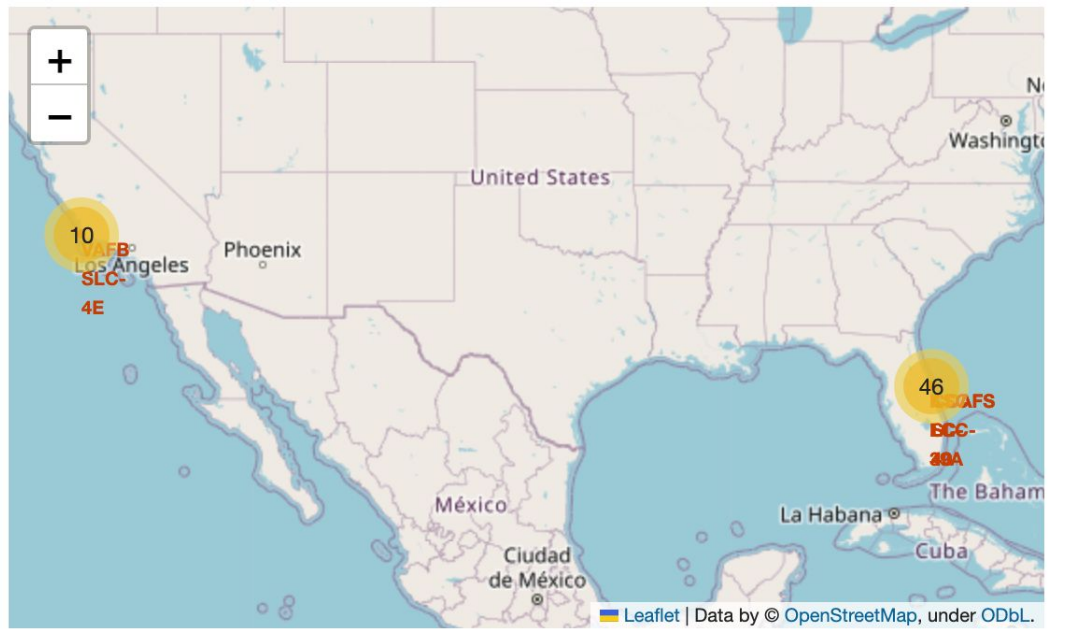
All Launch Sites on a map

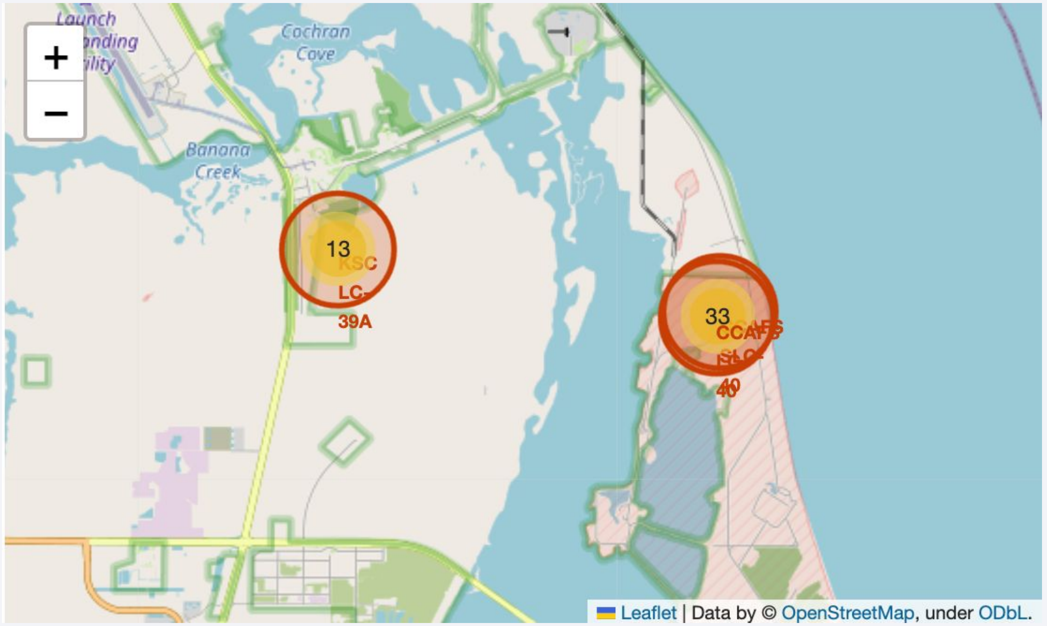


One site - nearby west coastline

Others - nearby east coastline

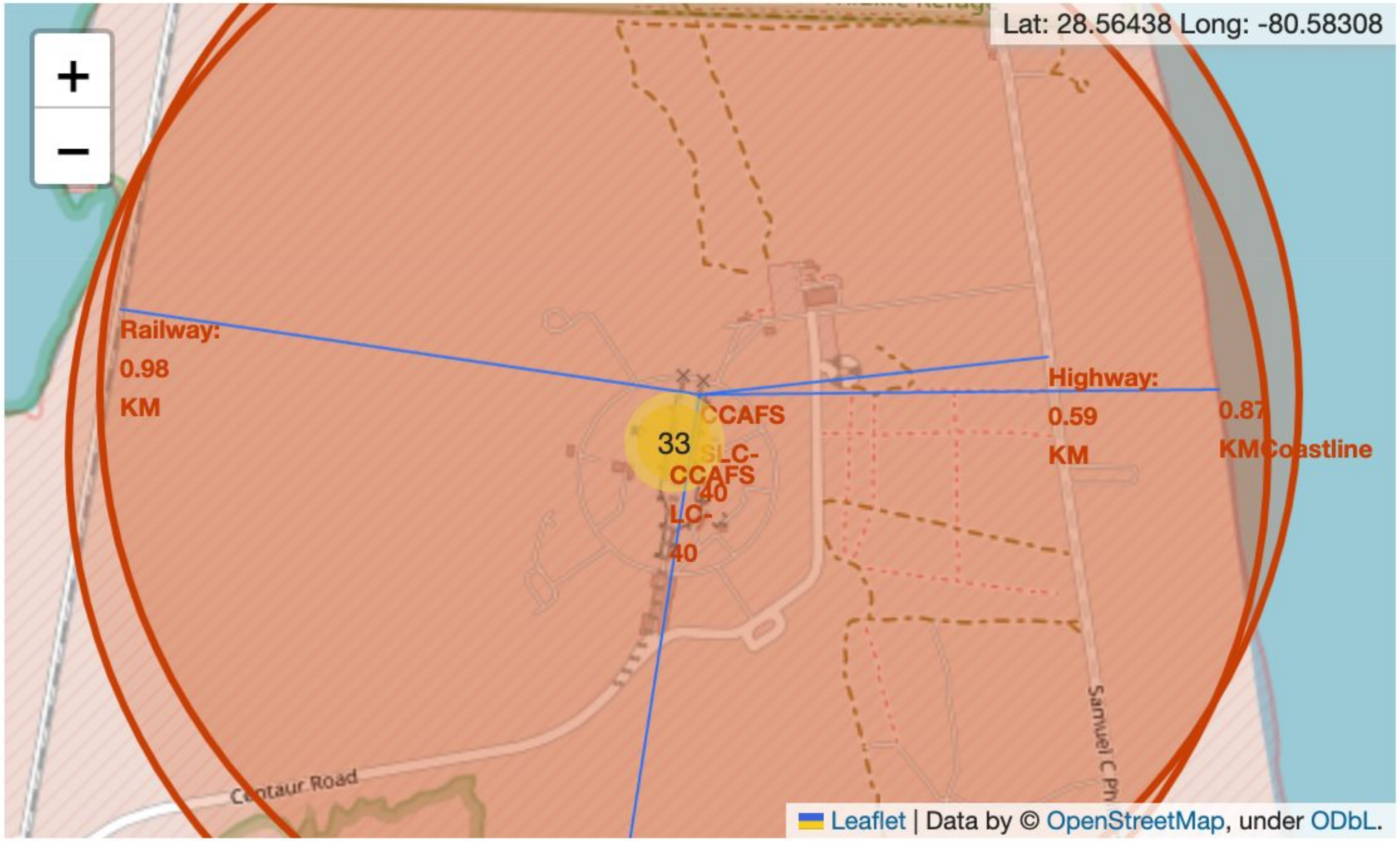
The success/failed launches for each site





The launches have the high success rate are mostly from the eastern coastline.

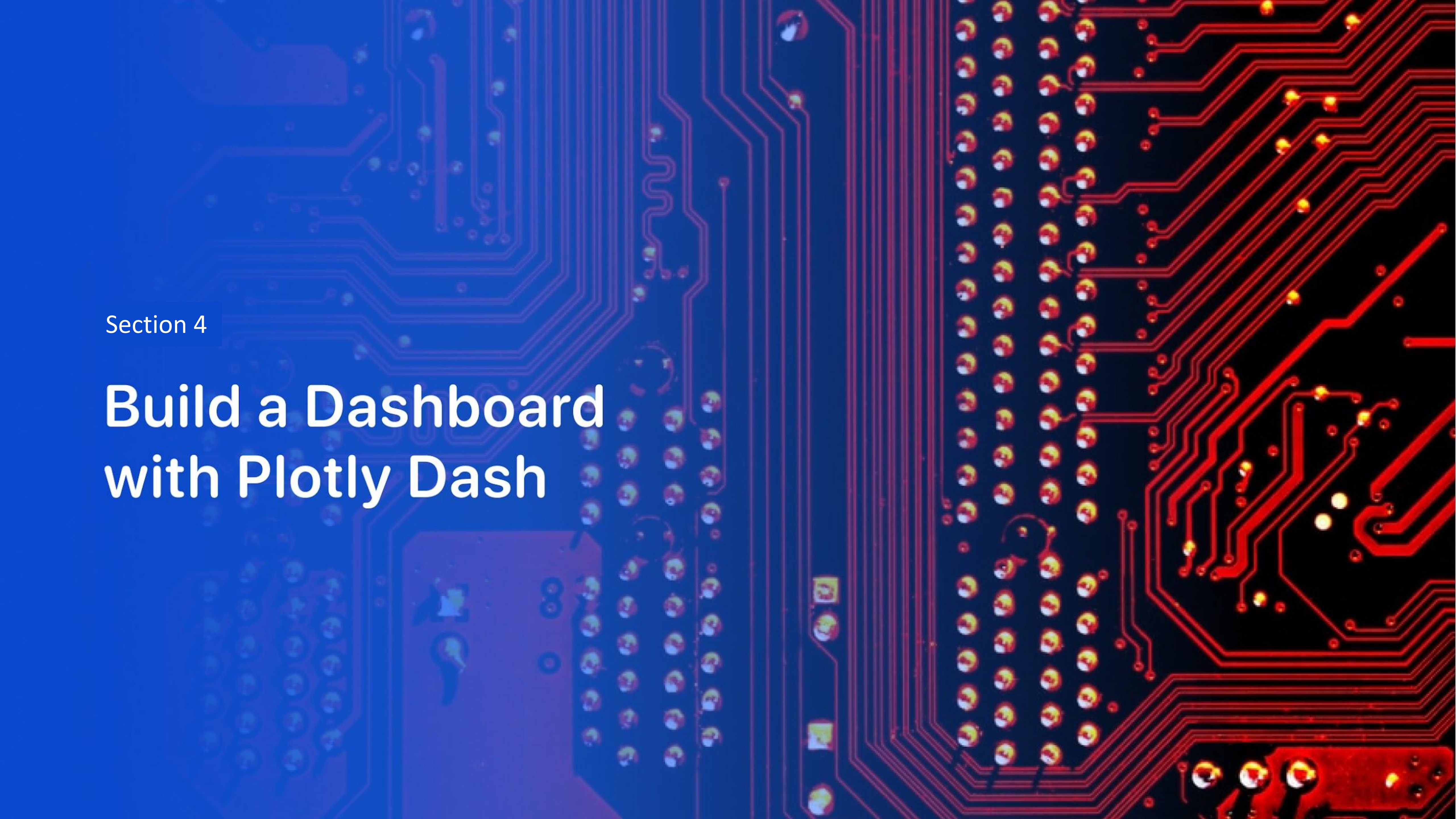
Launch Site and its proximities



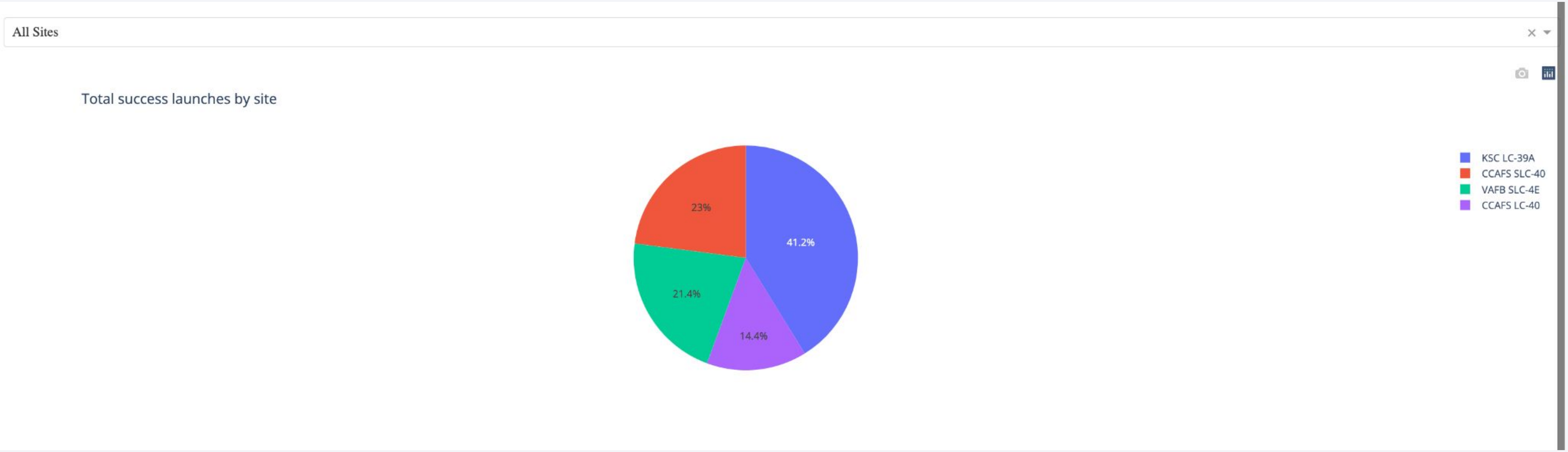
It can be seen that the blue straight lines mark the distance from the launch site to its proximities. Among the proximities, highway is closest to the launch site, whose distance is 0.59km.

Section 4

Build a Dashboard with Plotly Dash



Total success launches by all sites

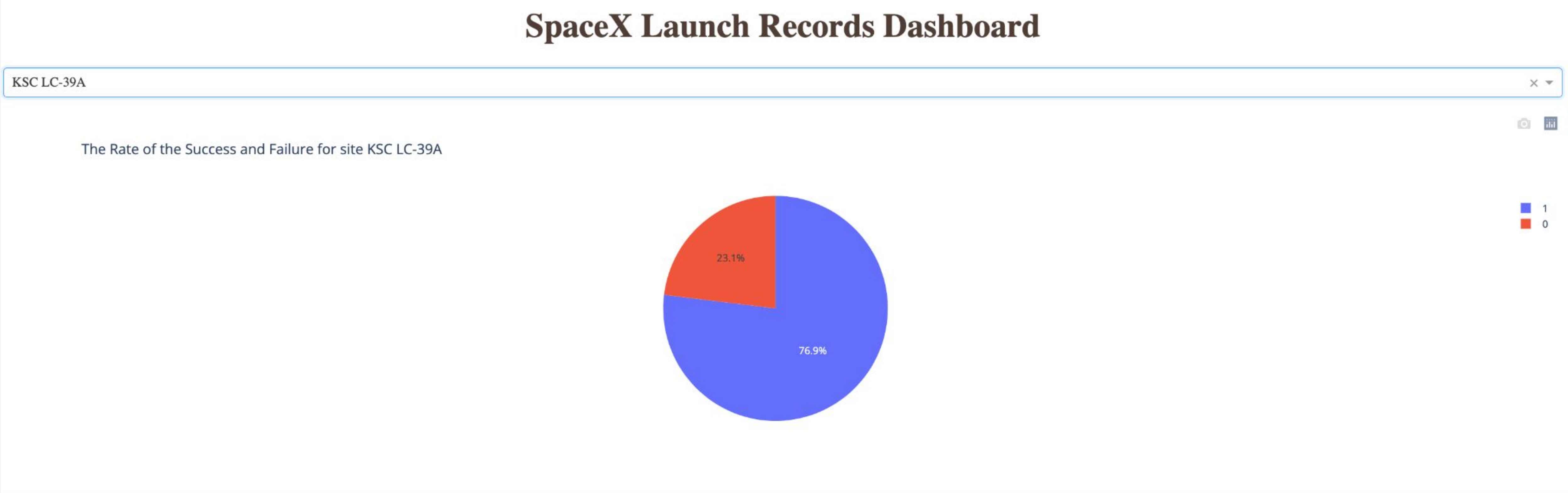


It indicates that:

KSC LC-39A has the highest success rate among all sites with 41.2%

CCAFS LC-40 is opposite, which has the lowest success rate among all sites with 14.4%

The launch site with the highest rate



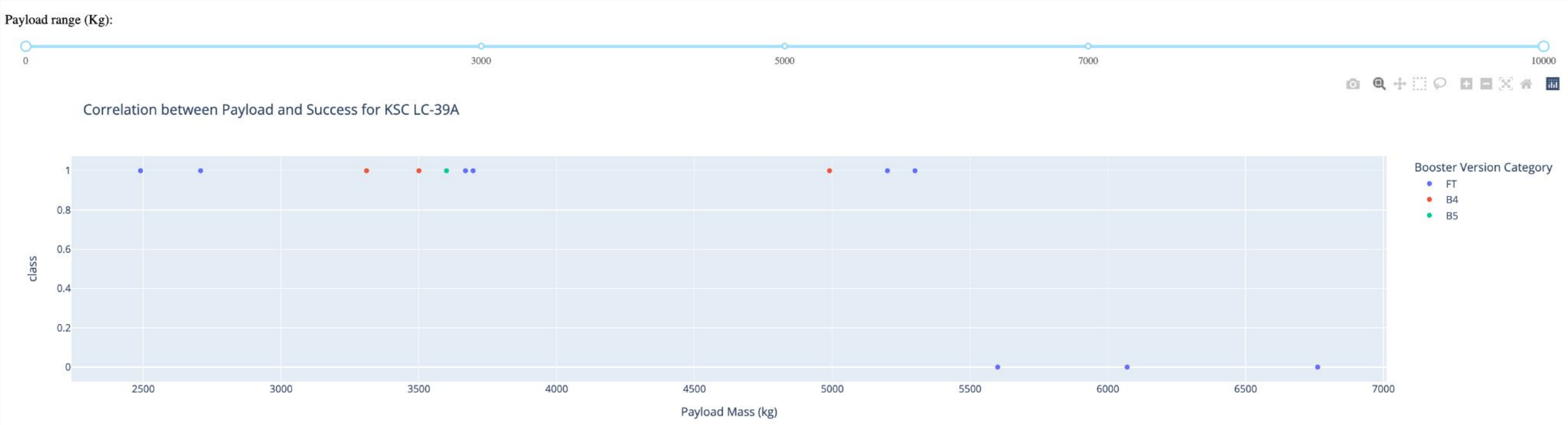
SpaceX Launch Records Dashboard

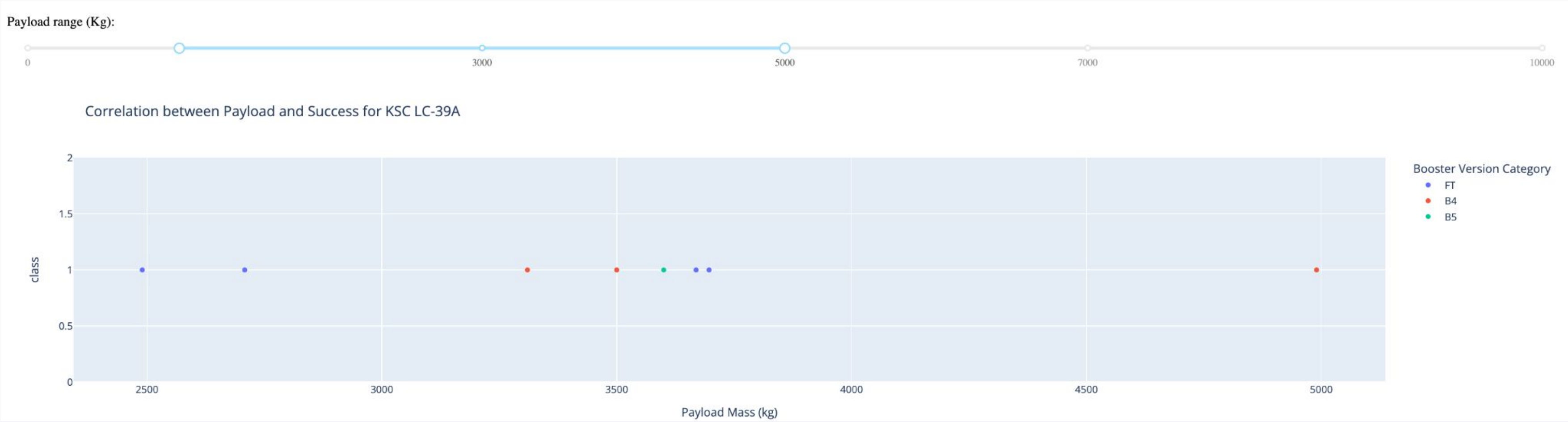
KSC LC-39A

The Rate of the Success and Failure for site KSC LC-39A

It shows that the KSC LC-39A has the highest success rate with 76.9% success ratio, with payload range 2000 - 7000 (kg)

Payload vs Launch outcome





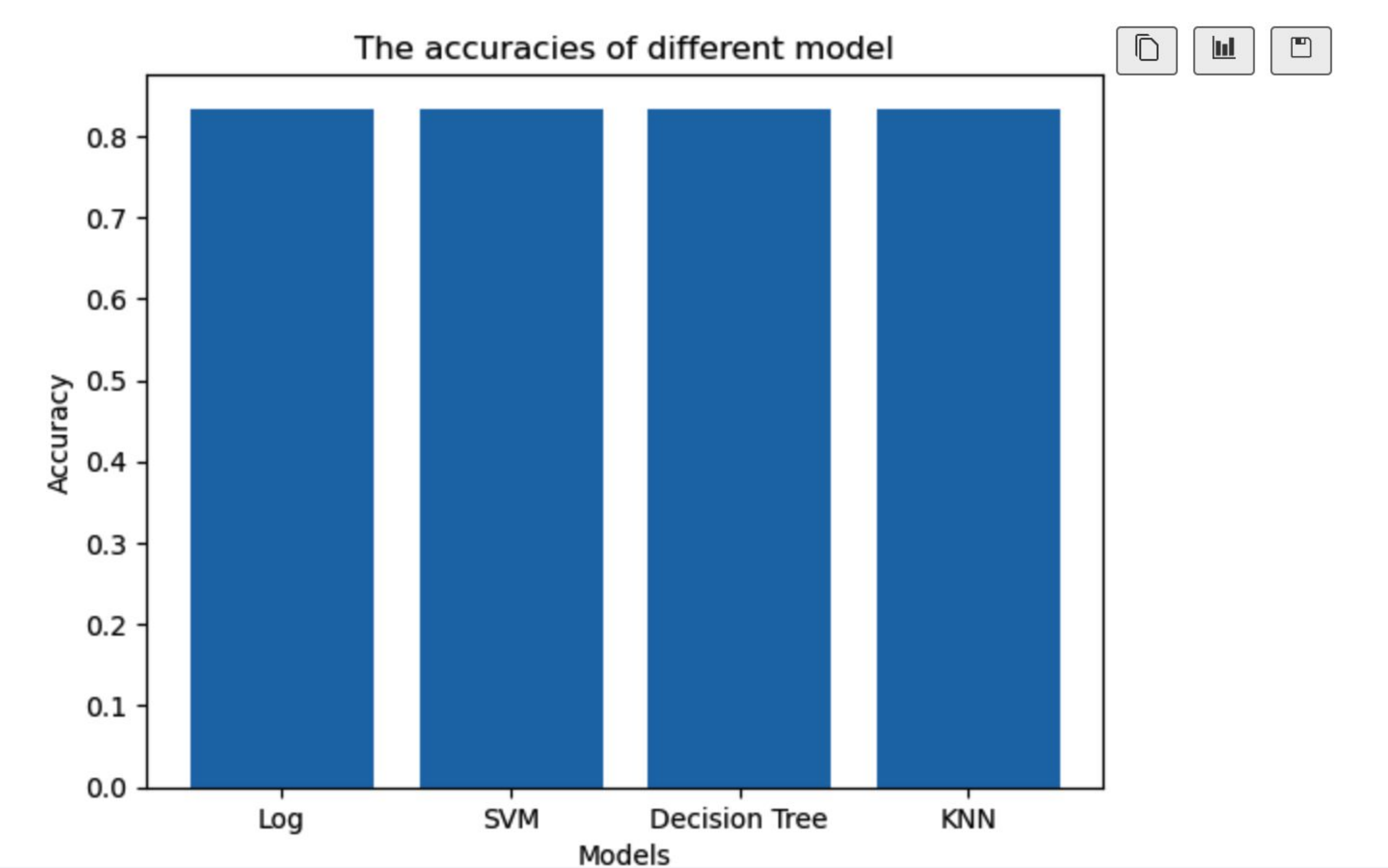
It’s clear that within the payload range of 2000 - 5500 (kg), the launches from the KSC LC-39A site all success.

Other than that, within all successful launches, the booster version of FT account for the most of them.

Section 5

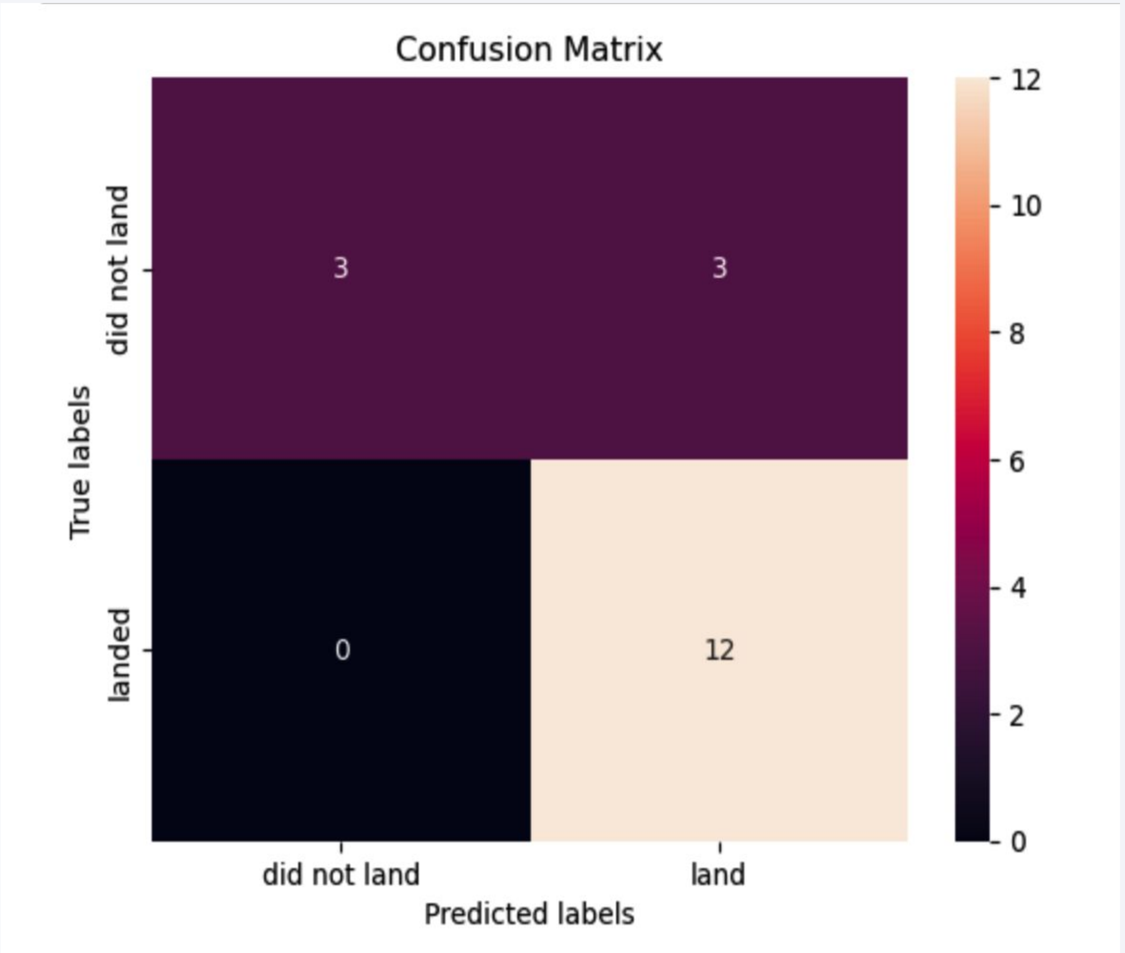
Predictive Analysis (Classification)

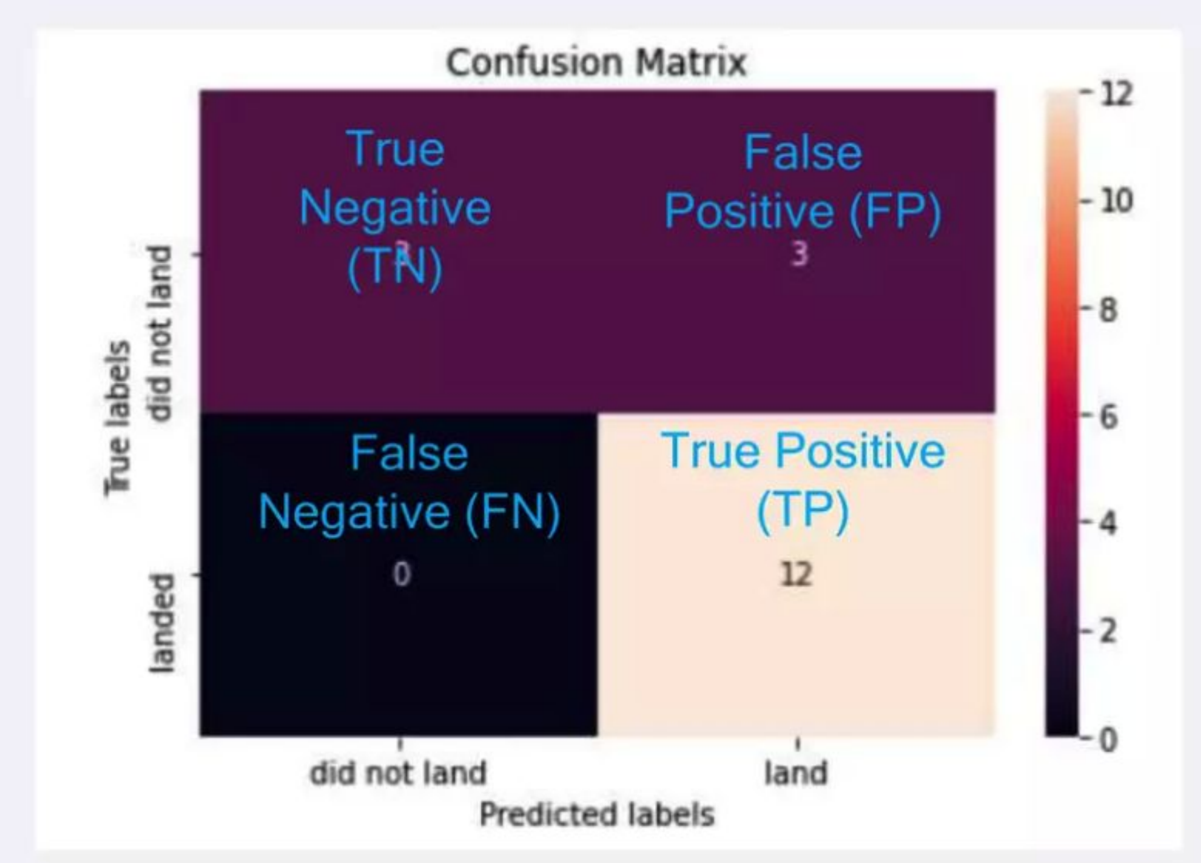
Classification Accuracy



All four models achieved the same accuracy, which is around 83.3%.

Confusion Matrix





Conclusions

- It can be found that the site with the highest successful score was KSC LC-39A
- On KSC LC-39A, within the range from 2000 kg to 5500kg, there was a higher successful rate of launches.
- The four models evaluated in this project all achieved the same performance, with the accuracy of 83.3%
- The distance to its proximities can be seen in the interactive visualization which is built by Plotly and Dash.

Appendix

- The code can be found on the Github:

https://github.com/RooNat/IBM_Data_Science/tree/drafts/Applied%20Data%20Science

Thank you!