IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jorge Sauma
June 2024

Outline

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

Executive Summary

SPACEX

Space X wants to predict the success of a rocket's first stage landing and how to improve such success rate.

This report details the analysis done using Space X launching data and describes which attributes are more likely to predict a successful landing.

The exploratory data analysis and SQL data findings showed that attributes such as Payload Mass and Launch site would be good predictors of success. Also, the experience gathered by the team is another important factor.

With this insights at hand, several models were used to predict the landing outcome and the Decision Tree model yielded the best results, however it did it by a slight margin. For all models, the accuracy score was good.

Introduction

One of the main competitive advantages of SpaceX in the race to create the best rockets is that the cost of launching new ones is much less than the competitors. This is because Space X is able to reuse the first stage of their rockets.

Therefore, the ability to predict if a first stage will successfully land so it can be reused is critical for the success of the company. It would allow the company to outbid other companies with higher launching costs.

In this report, we explain which are the best ways to predict if a first stage would be able to land safely so it can be reused in another launch.

The questions we want to answer are:

What data is needed to predict a successful landing?

What are the data requirements

How to collect and prepare such data?

Which model or models would be the best ones to use to predict a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - There are several sources that can be used to retrieve the needed data. Two of them that we are going to use are:
    - Extract data from Space X REST API (preferred method)
    - Use Web Scrapping to get data from sources like Wikipedia
- Perform data wrangling
  - The available data includes attributes like Flight Number, Date, Booster version, Payload mass Orbit, Launch Site, Outcome.
  - These attributes will be used to perform an exploratory analysis.

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL
    - Using SQL statements and visualization tools like scatter and line plots we will find out the right attributes to use.
- Perform interactive visual analytics using Folium and Plotly Dash
    - Folium and Plotly Dash are useful interactive tools to understand data.
- Perform predictive analysis using classification models
    - Once we decide the best attributes we would evaluate different prediction models, train and test them, and find out which has the best score (i.e. error rate, accuracy, etc.).

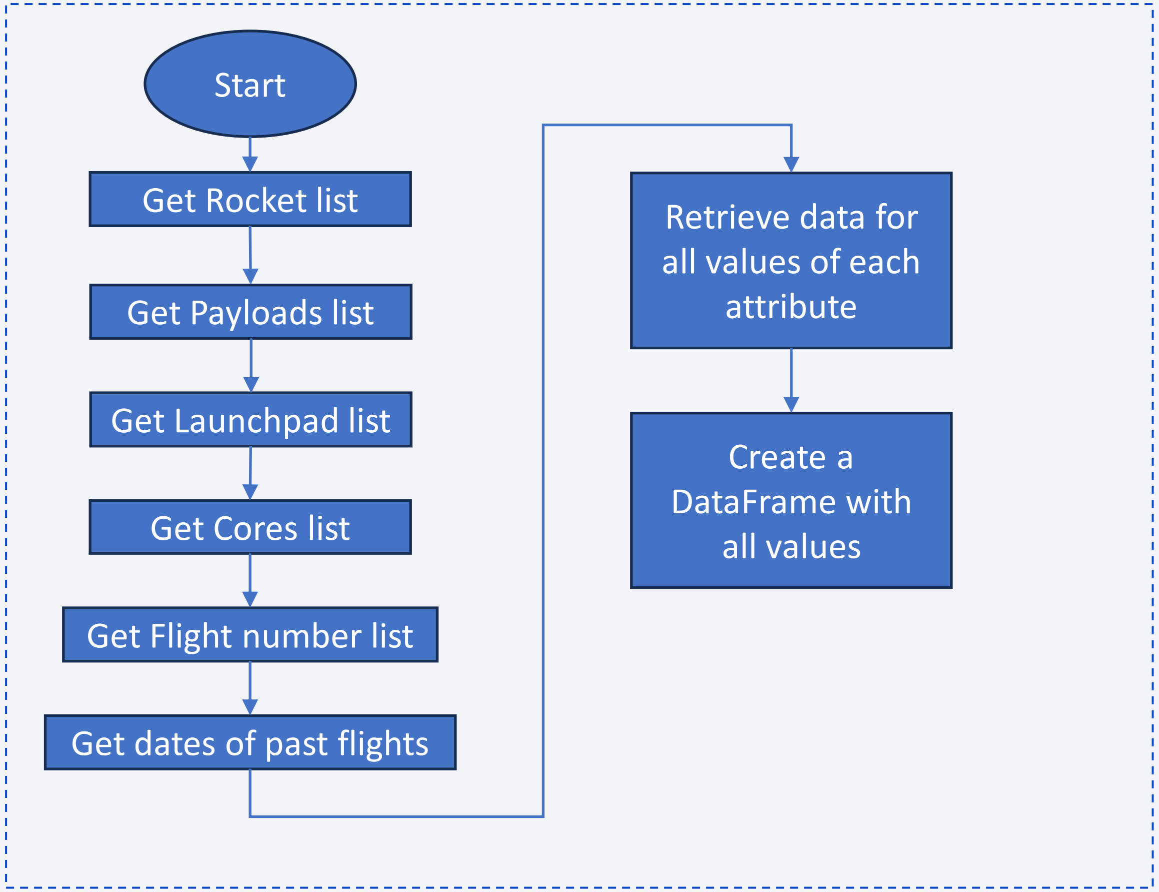
Data Collection

- As stated before, we did use two sources for getting data:
    - Extract data from Space X REST API (preferred method)
    - Use Web Scrapping to get data from sources like Wikipedia

Data Collection – SpaceX API

- REST API
- Source: https://api.spacexdata.com/v4/launches/past
- GitHub URL for the SpaceX API calls notebook:

https://github.com/JorgeSauma/IBM-DS-CapstoneProject/blob/main/Notebooks/jupyter-labs-spacex-data-collection-api.ipynb


Start

Get Rocket list

Get Payloads list

Get Launchpad list

Get Cores list

Get Flight number list

Get dates of past flights

Retrieve data for all values of each attribute

Create a DataFrame with all values

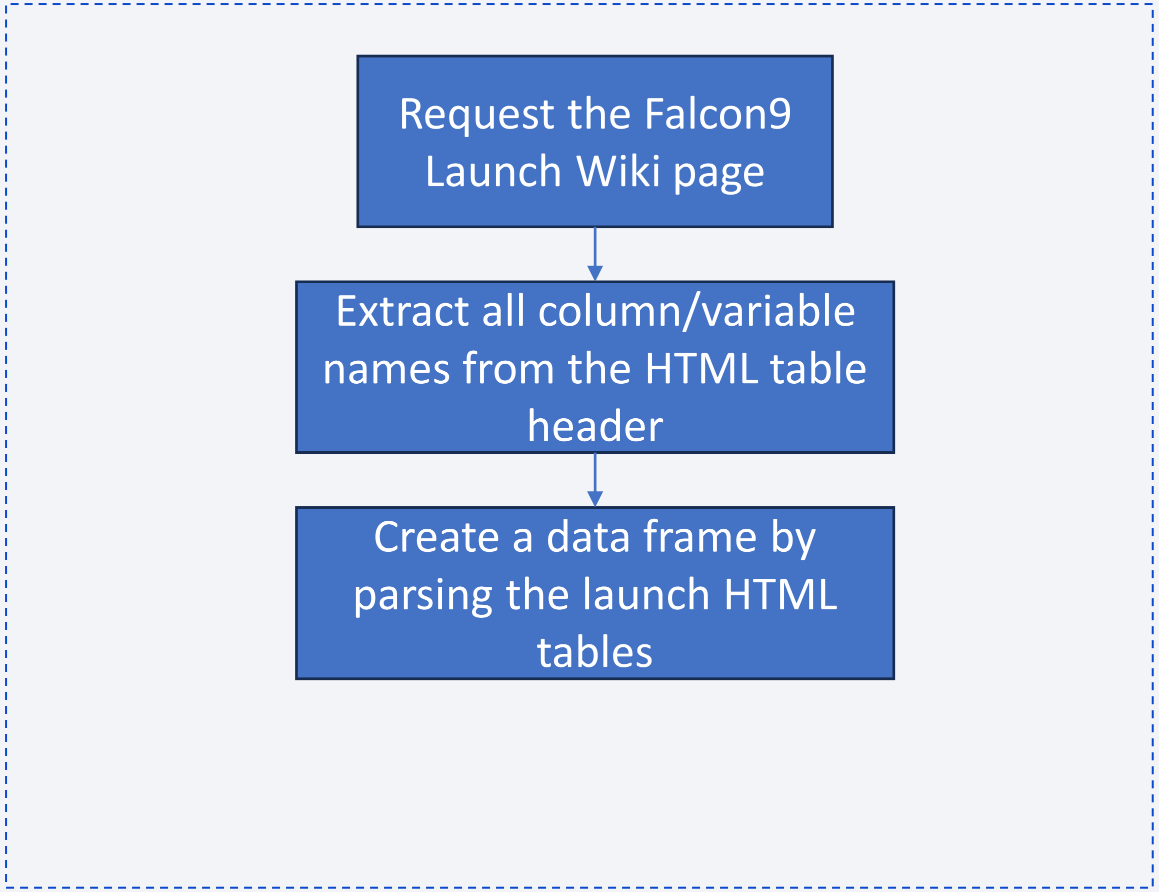
Data Collection - Scraping

- Web Scrapping using BeatifulSoup
- Source:

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

- GitHub URL for the SpaceX API calls notebook:

https://github.com/JorgeSauma/IBM-DS-CapstoneProject/blob/main/Notebooks/jupyter-labs-webscraping.ipynb



Request the Falcon9 Launch Wiki page

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

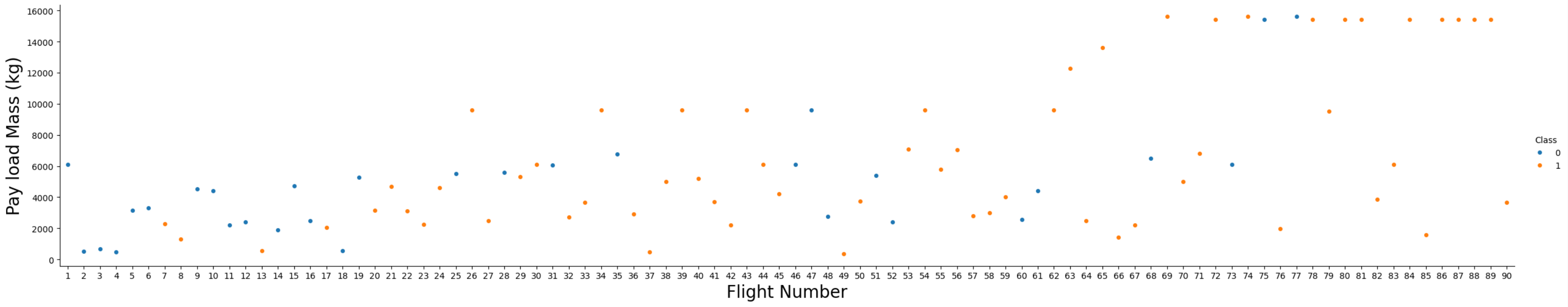
Data Wrangling

- Dealing with missing values:
  - Use the mean value for the PayloadMass missing values
  - Keep LandingPad missing values, representing when there wasn`t a landing pad
- Count values for:
  - Launches for each site
  - Orbits
  - Landing outcomes
    - For Landing Outcomes a label of 1 (Successful landing) and 0 (Failed landing) was defined.
- You need to present your data wrangling process using key phrases and flowcharts
- GitHub URL for Data Wrangling:
  - https://github.com/JorgeSauma/IBM-DS-CapstoneProject/blob/main/Notebooks/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

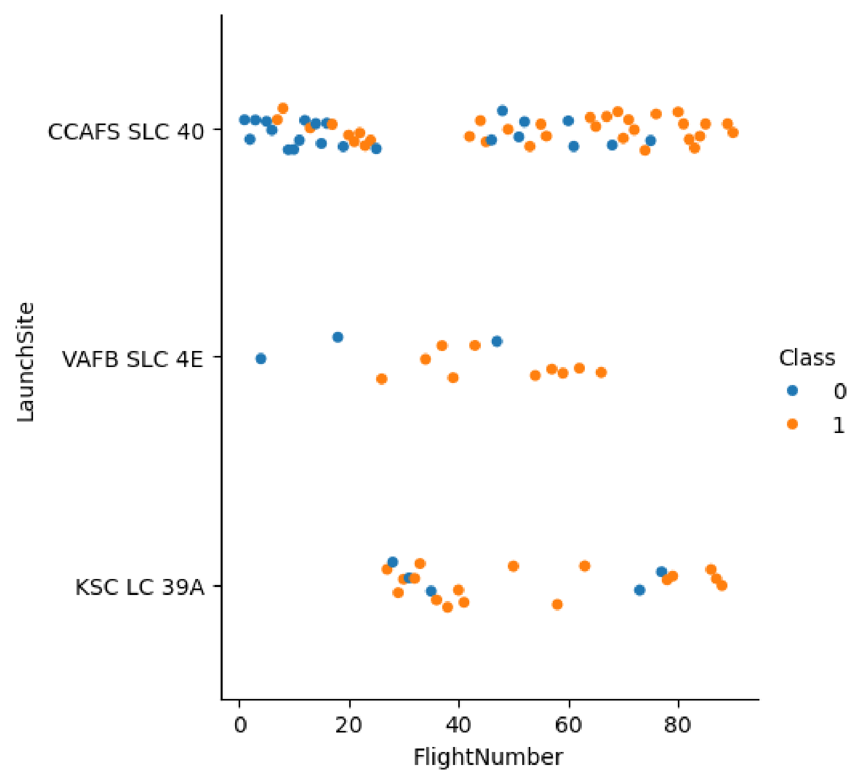
- We did several plots to understand the relationship between Outcome, Launch Site, Payload Mass, and Orbit.

- The following plots were created:

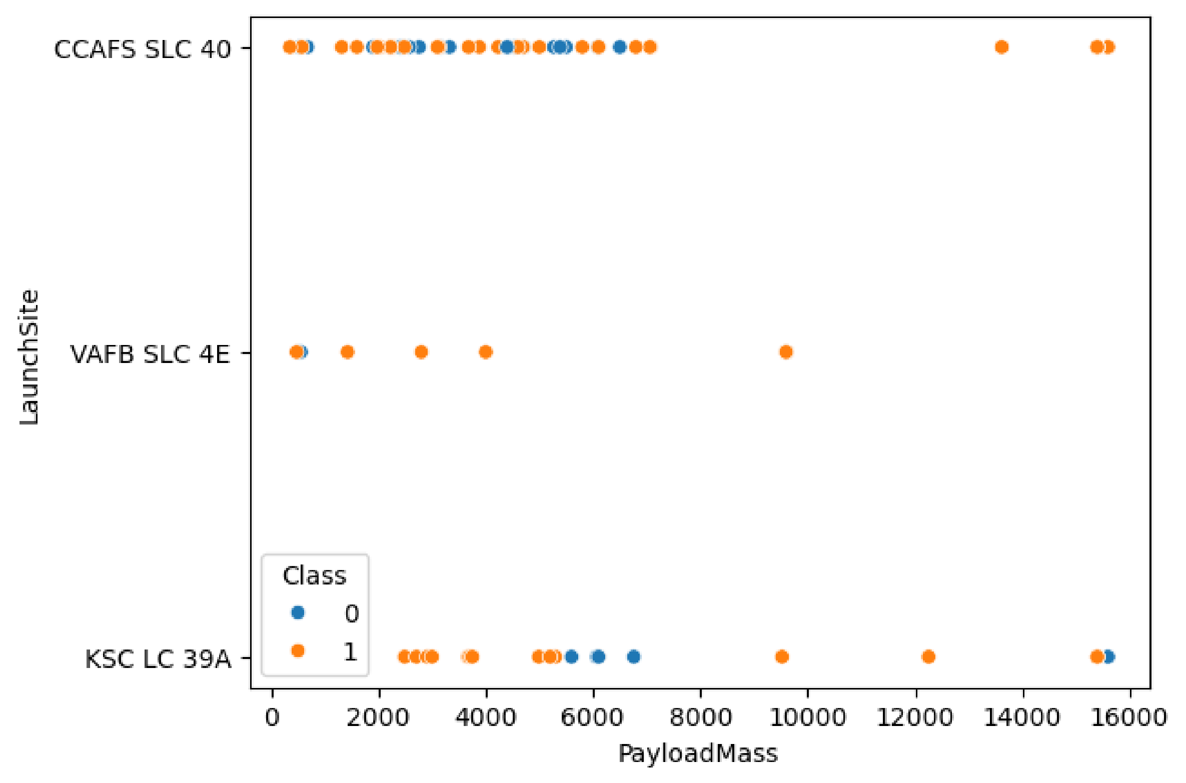


Outcome for each flight related to Payload Mass

EDA with Data Visualization

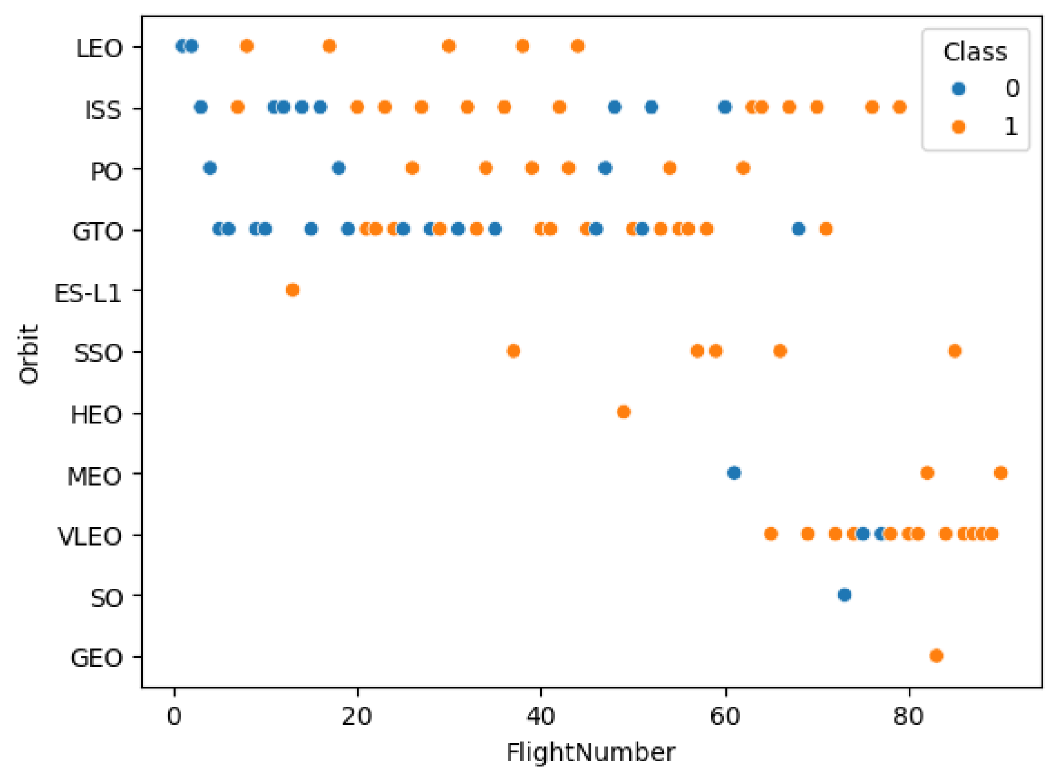


Outcome for each flight grouped by landing site

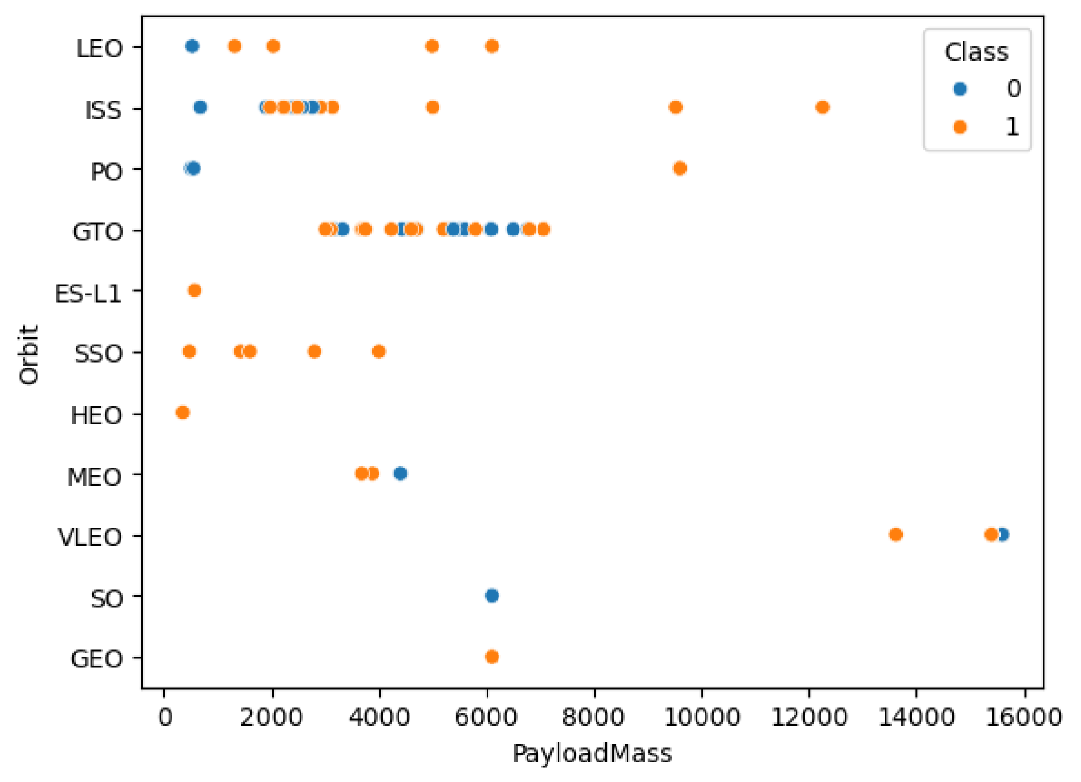


Payload Mass for each flight grouped by Launch Site

EDA with Data Visualization

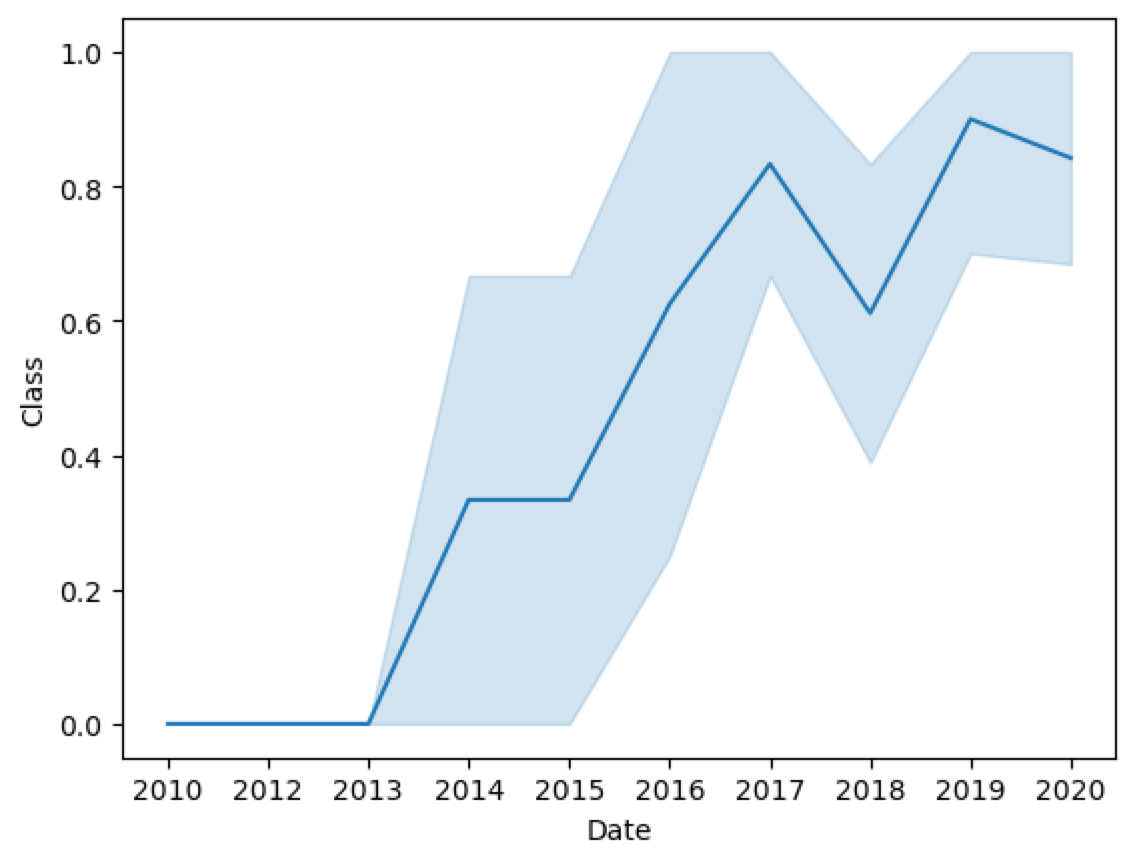


Outcome for each flight grouped by Orbit



Payload Mass for each flight grouped by Orbit

EDA with Data Visualization



Outcome progression

EDA with Data Visualization

GitHub URL for EDA with Data Visualization:

https://github.com/JorgeSauma/IBM-DS-CapstoneProject/blob/main/Notebooks/edadataviz.ipynb

EDA with SQL

- Besides the plots, we did another EDA using SQL statements.
- The following operations were done using SQL queries:
  - Display the names of the unique launch sites in the space mission:

```
select distinct(LAUNCH_SITE) from SPACEXTBL
```

  - Display 5 records where launch sites begin with the string 'CCA’:

```
select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

  - Display the total payload mass carried by boosters launched by NASA (CRS):

```
select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

```
select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1’
```

  - List the date when the first succesful landing outcome in ground pad was achieved:

```
select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

EDA with SQL

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

```
     select Booster_Version from SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND
(PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000)
```

- List the total number of successful and failure mission outcomes:

```
     select count(Mission_Outcome) from SPACEXTBL WHERE Mission_Outcome = 'Success' OR
Mission_Outcome = 'Failure (in flight)'
```

- List the names of the booster_versions which have carried the maximum payload mass:

```
     select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select
max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015:

```
     SELECT SUBSTR(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_site FROM
SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%' AND SUBSTR(Date,0,5) = '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:

```
     SELECT Landing_Outcome, COUNT(*) AS Numbers FROM SPACEXTBL WHERE Landing_Outcome LIKE
'Success%' AND Date BETWEEN '2010.06.04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY
Numbers DESC;
```

EDA with SQL

- GitHub URL for the SQL EDA:

https://github.com/JorgeSauma/IBM-DS-CapstoneProject/blob/main/Notebooks/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To make it easier to visualize the landing sites and outcomes, a Folium map was created and the following elements were added:
  - Marked launch sites
  - Clusters with outcomes for each launch for each site
  - Lines between launch site and coast, cities, railways, etc.
- All these elements help us to find geographical patterns about launch sites
- GitHub URL for Map with Folium:

https://github.com/JorgeSauma/IBM-DS-CapstoneProject/blob/main/Notebooks/lab_jupyter_launch_site_location_withFolium.ipynb

Build a Dashboard with Plotly Dash

- The dashboard includes a pie chart to show the landing outcome based on the selected launch site and a scatter plot that shows the landing outcome based on the payload.
- This dashboard was created to facilitate finding a pattern based on the landing outcome and the launch site or the Payload.
- GitHub URL for Plotly Dash:

https://github.com/JorgeSauma/IBM-DS-CapstoneProject/blob/main/Notebooks/Dash_script_IBMDS.python

Predictive Analysis (Classification)

- In order to define the most appropriate classification model to predict a successful outcome, we evaluated several methods:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-nearest neighbor
- Before applying these methods, the dataset was split in training and testing set, so we can evaluate the accuracy of the model and its ability to predict outcomes.
- GitHub URL for Predictive Analysis:

https://github.com/JorgeSauma/IBM-DS-CapstoneProject/blob/main/Notebooks/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

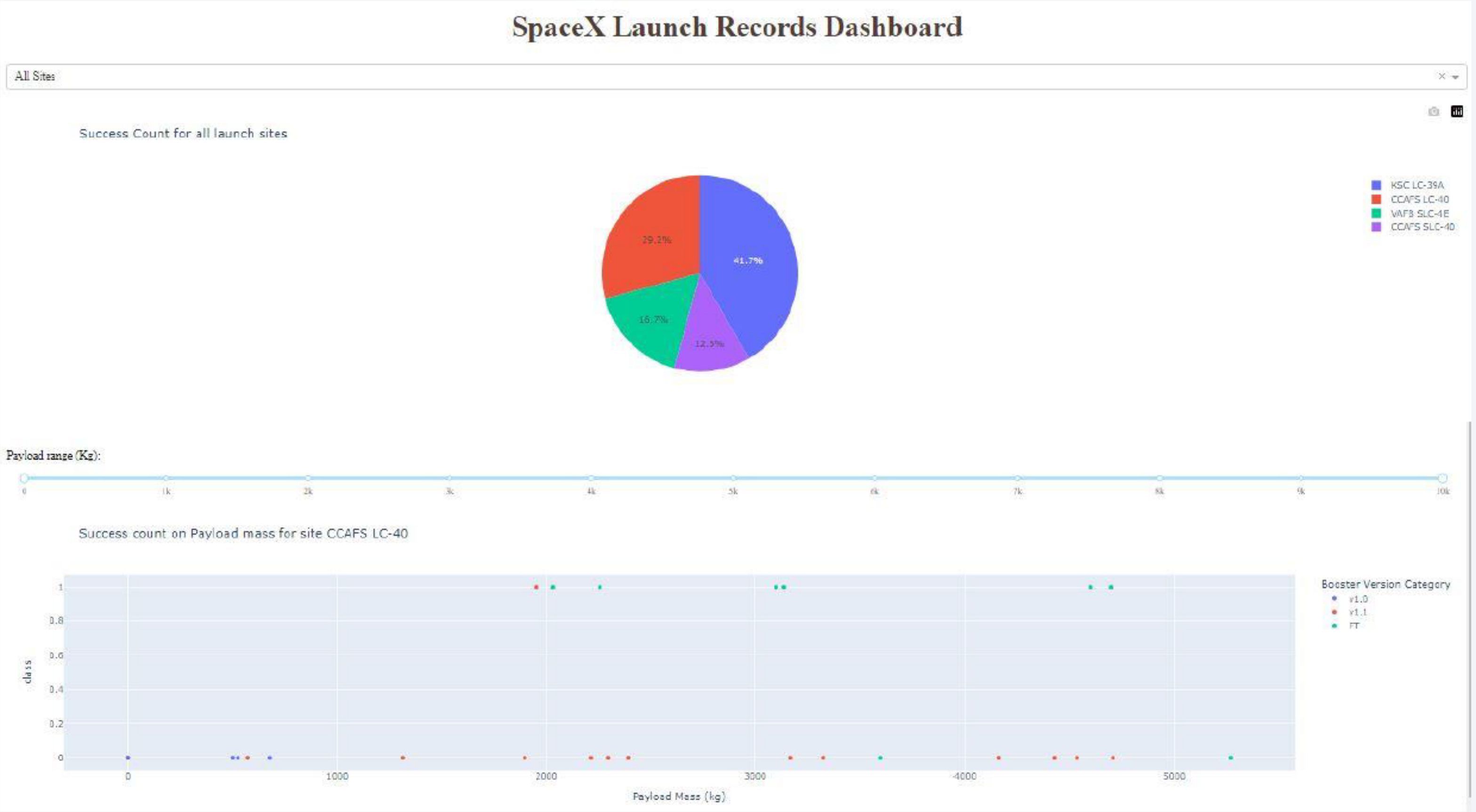
- Exploratory Data Analysis results:
  - Higher Payload Mass increases the chance of success
  - The orbit and the launching site may not be good predictors
  - The experience launching rockets has improved the chance of a successful landing
- Predictive analysis results:
  - The scores for each model are:

Logistic Regression:	0.8333333333333334
Support Vector Machine	0.8333333333333334
Decision Tree	0.8888888888888888
K-nearest neighbor	0.8333333333333334

- **All the scores are very similar, but the Decision Tree is the one with a slightly better prediction rate.**

Results

- Interactive analytics demo in screenshots :



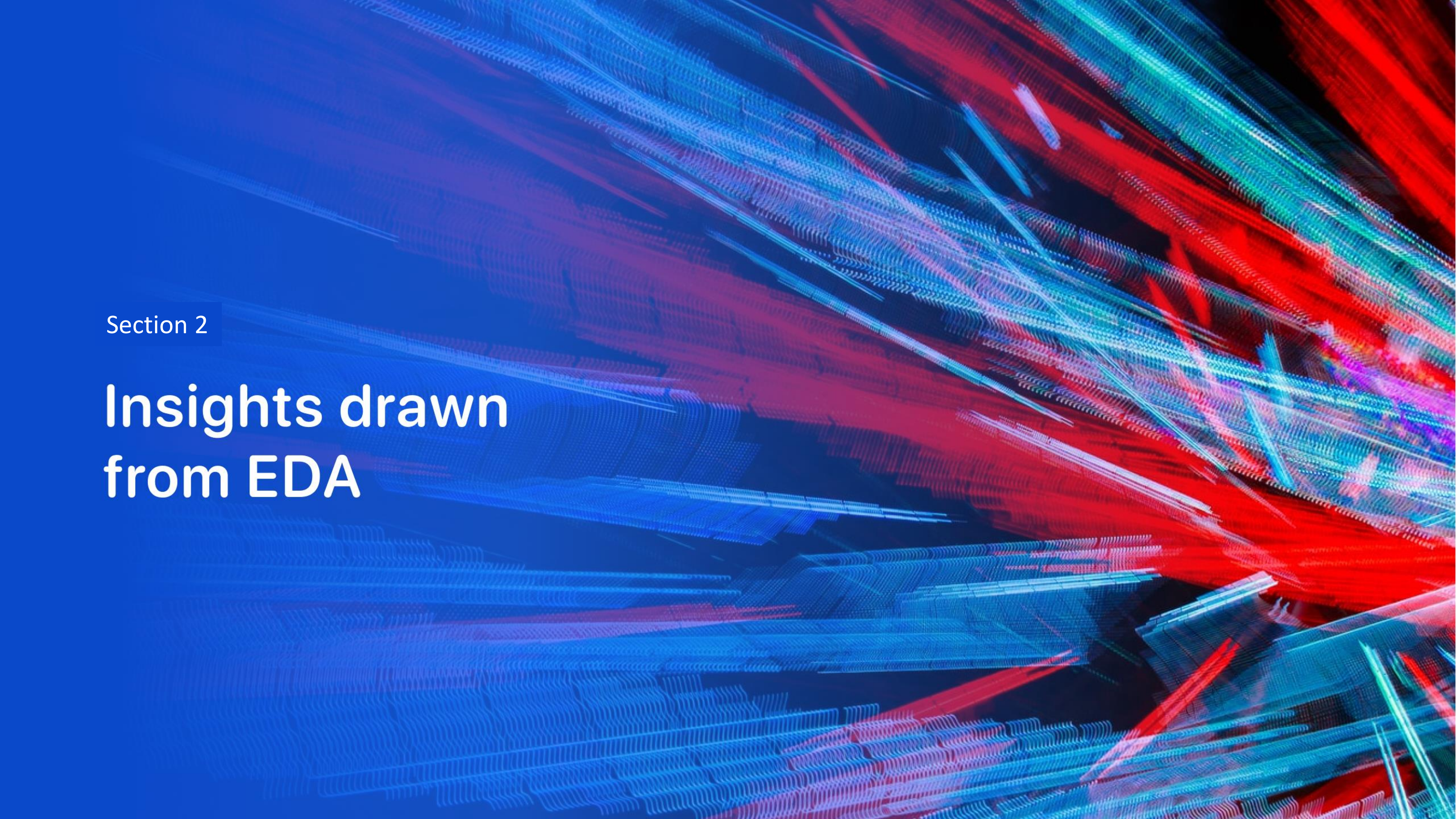
SpaceX Launch Records Dashboard

All Sites

Success Count for all launch sites

Payload range (Kg):

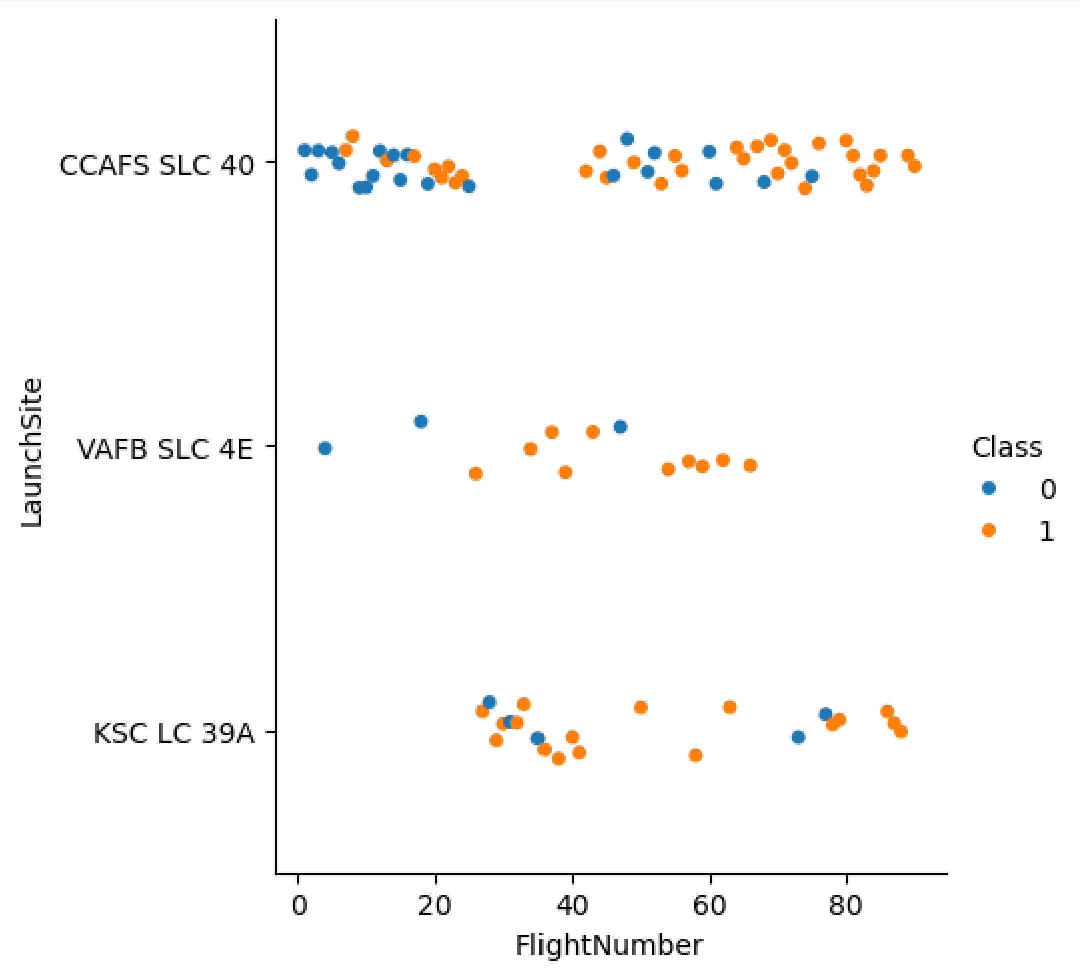
Success count on Payload mass for site CCAFS LC-40



Section 2

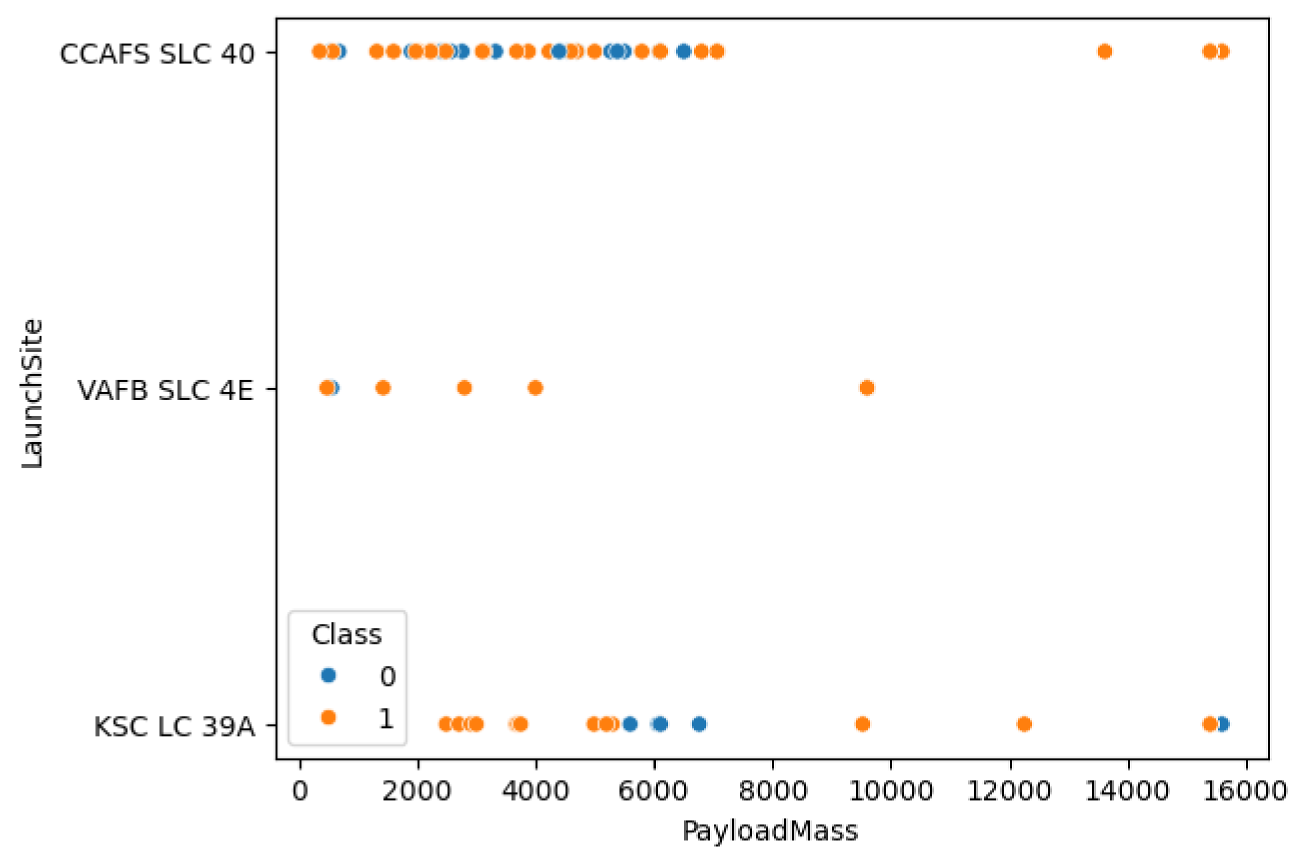
Insights drawn from EDA

Flight Number vs. Launch Site



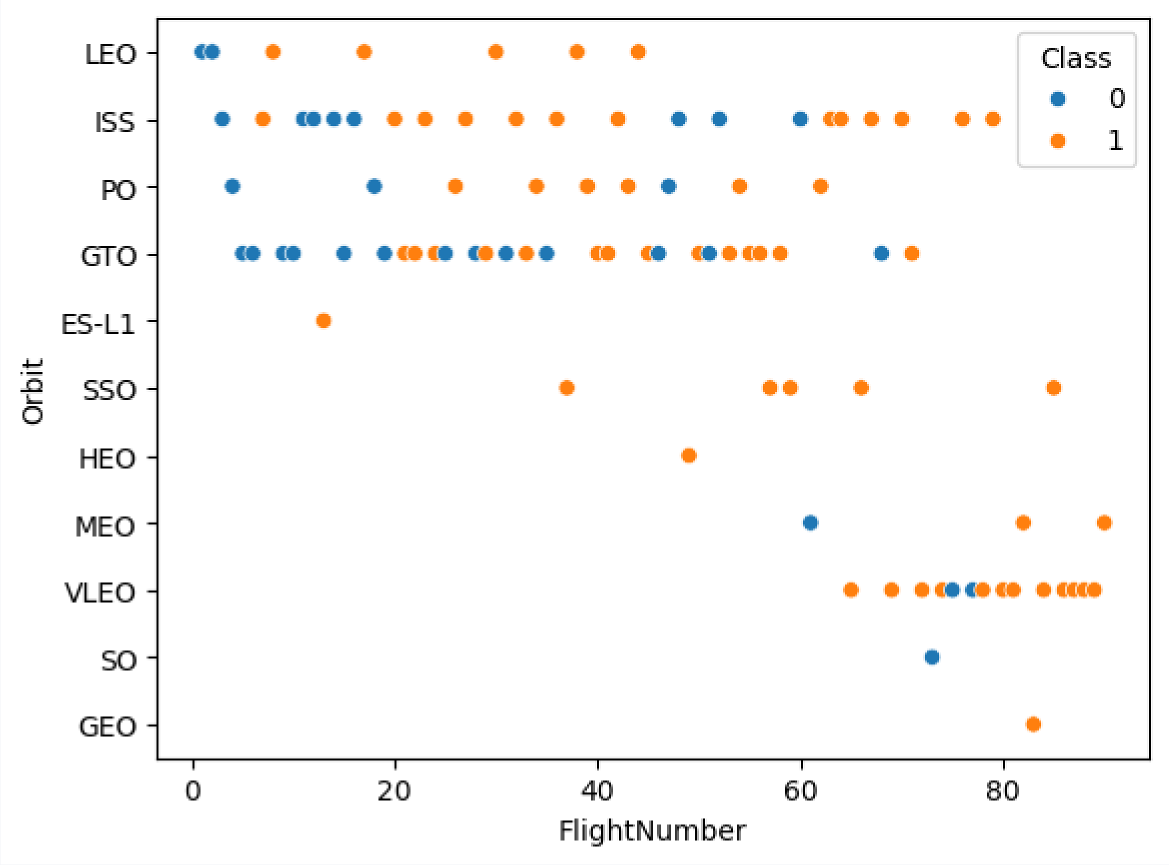
- From the plot we see that the outcome improves after more flights are performed.
- Each site has both success landings and failures and it is hard to select one site over other.

Payload vs. Launch Site



- From this plot is clear that a higher Payload Mass means a higher chance of a successful landing.

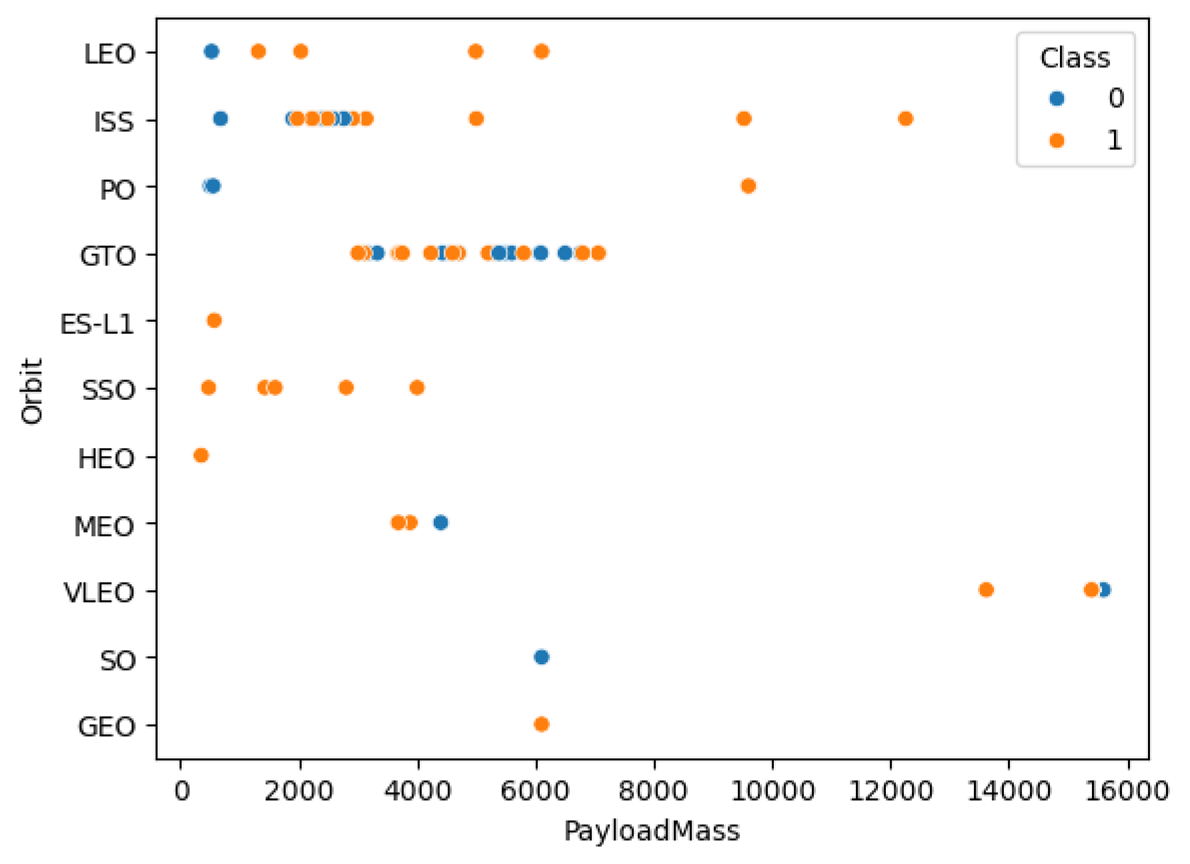
Flight Number vs. Orbit Type



- This plot shows that the orbit may not be a good predictor of outcome.

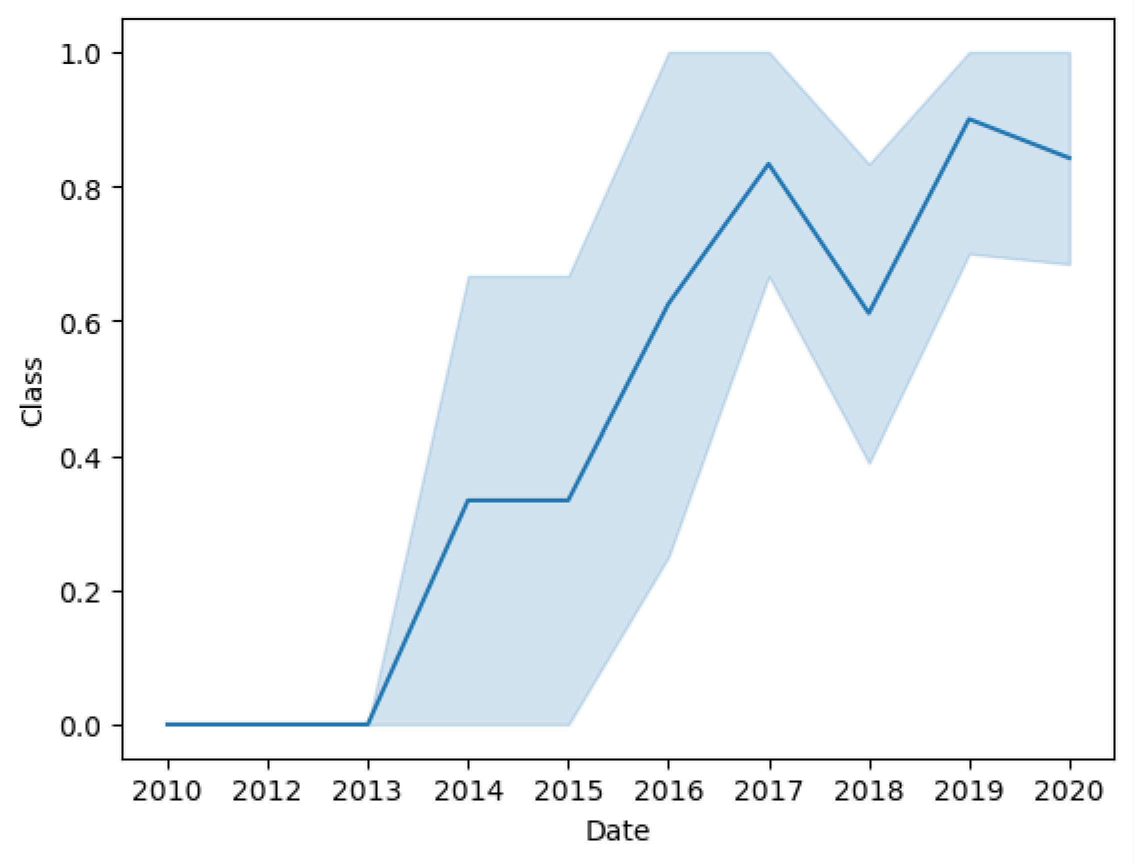
- Only rockets sent to SSO seems to be always successful, but we would need to add other factors to the analysis.

Payload vs. Orbit Type



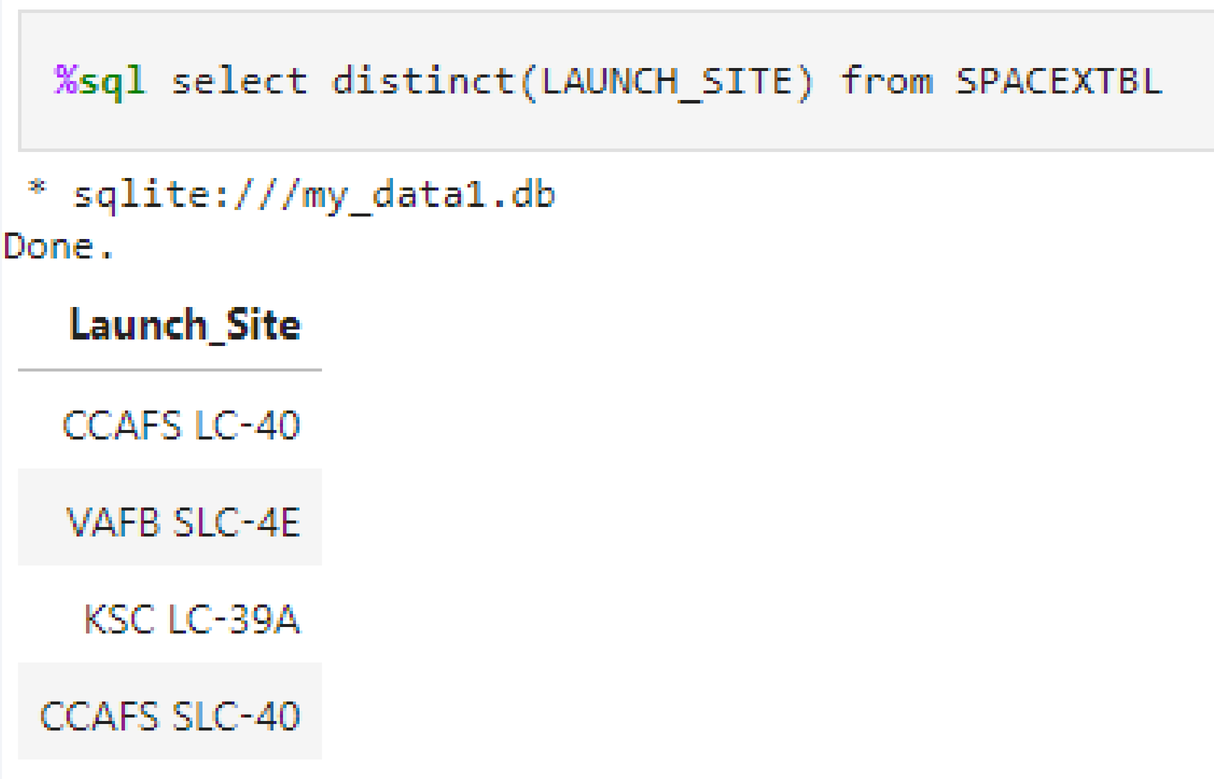
- Again, it seems like a higher Payload Mass means a better outcome, regardless of the orbit.

Launch Success Yearly Trend



- The graph clearly shows that the landing outcome is getting better.
- This means that the team is learning how to improve the landing process.

All Launch Site Names



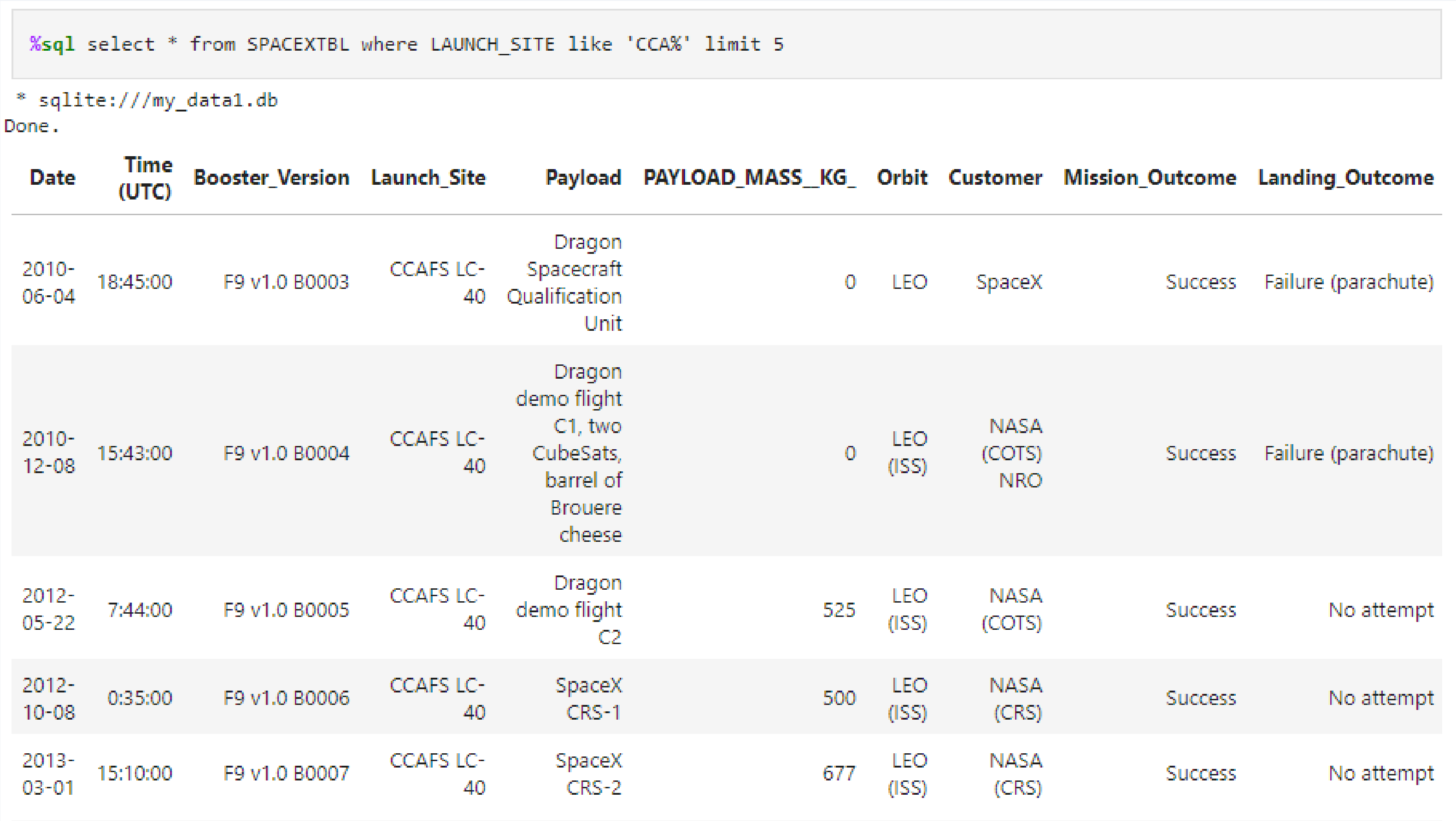
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

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

- There are currently 4 sites where rockets have been launched.

Launch Site Names Begin with 'CCA'



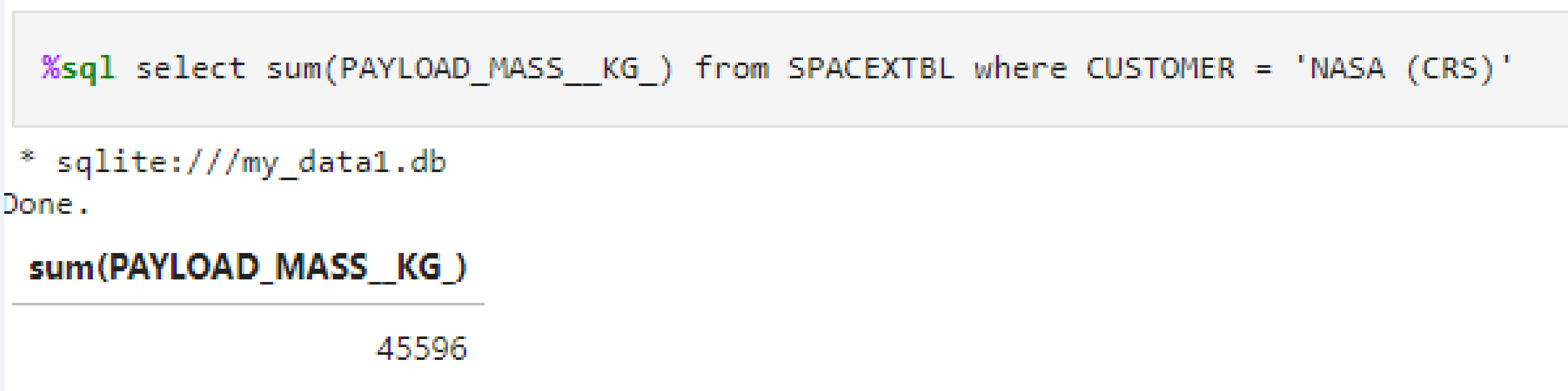
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- There are two sites which name starts with CCA

Total Payload Mass



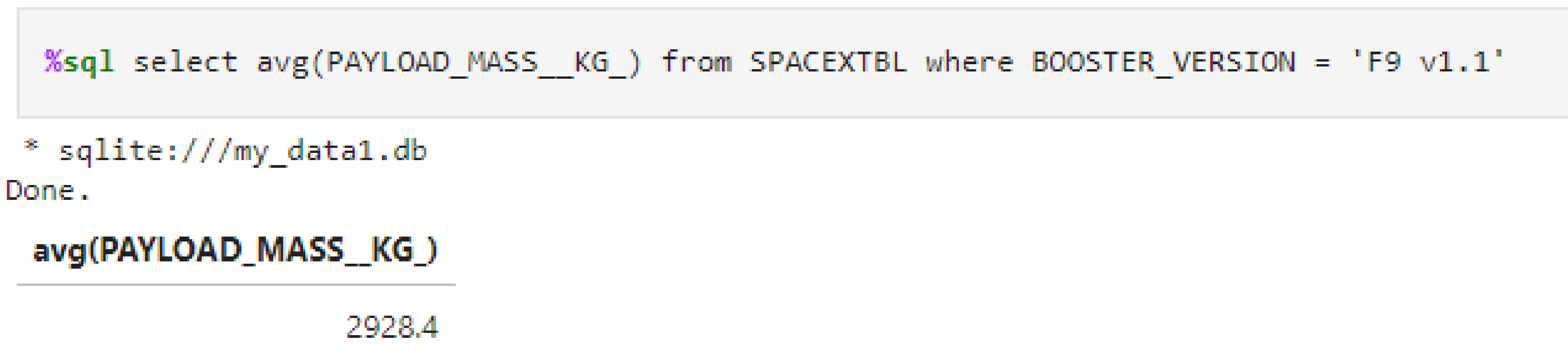
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

- The total of Payload used for launches is 45596 Kg.

Average Payload Mass by F9 v1.1



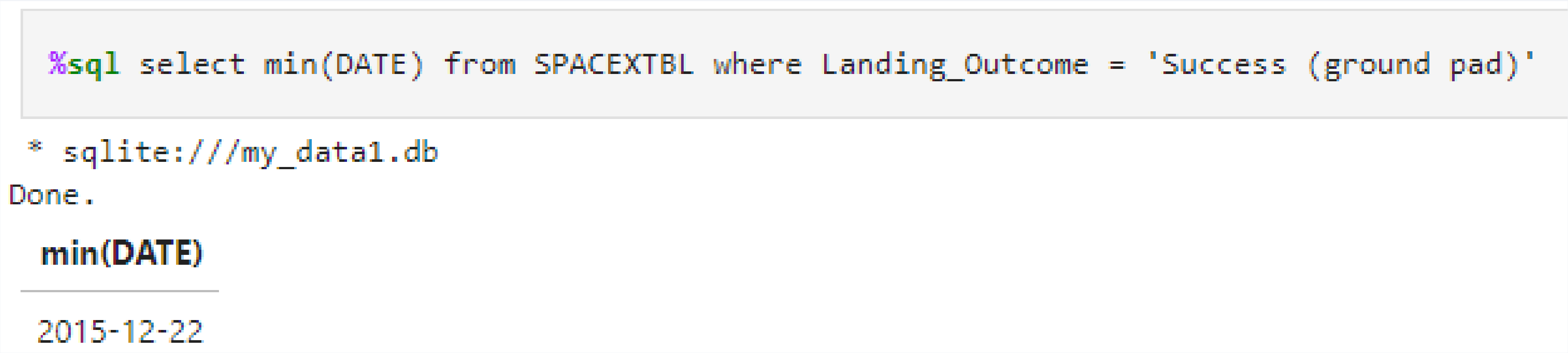
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

- The amount of Payload carried by Booster F9 1.1 is 2928.4

First Successful Ground Landing Date



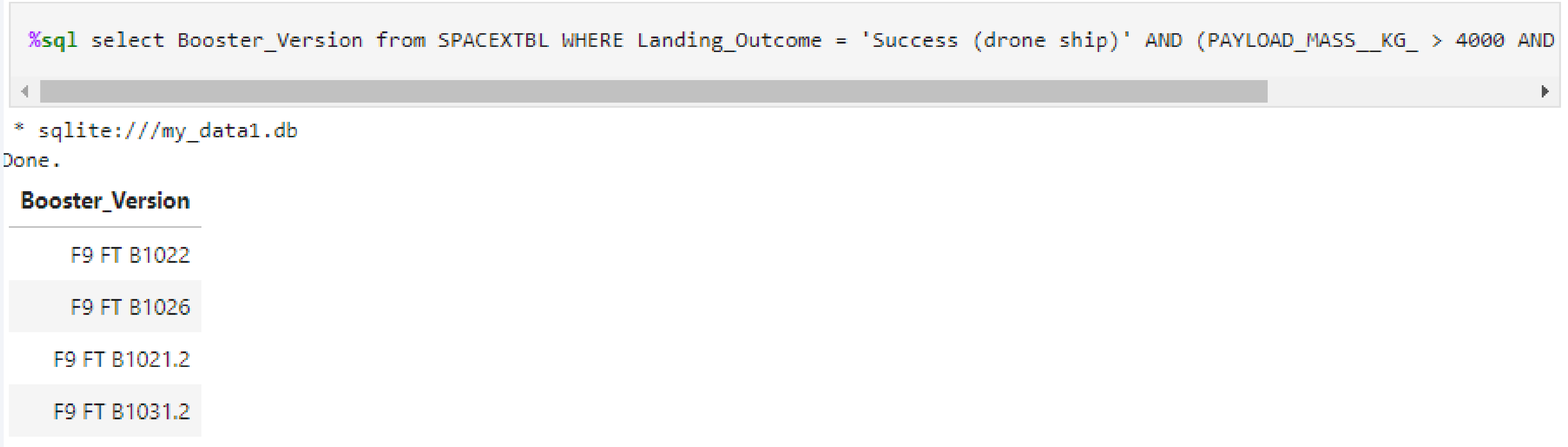
```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(DATE)
2015-12-22

- Starting in 12/22/2015, the chance of landing success has been increasing.

Successful Drone Ship Landing with Payload between 4000 and 6000



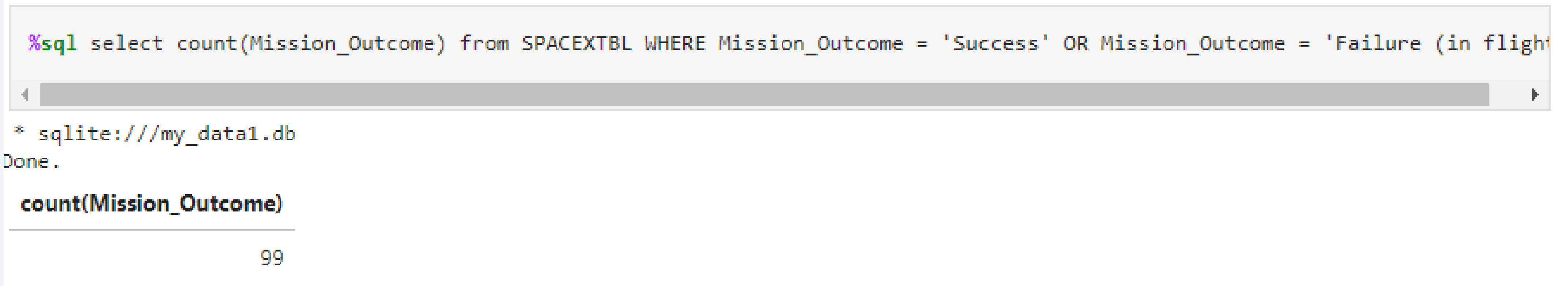
```
%sql select Booster_Version from SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ > 4000 AND
```

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

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

- These boosters have a positive record of success landing with Payloads between 4000 and 6000 Kg.

Total Number of Successful and Failure Mission Outcomes



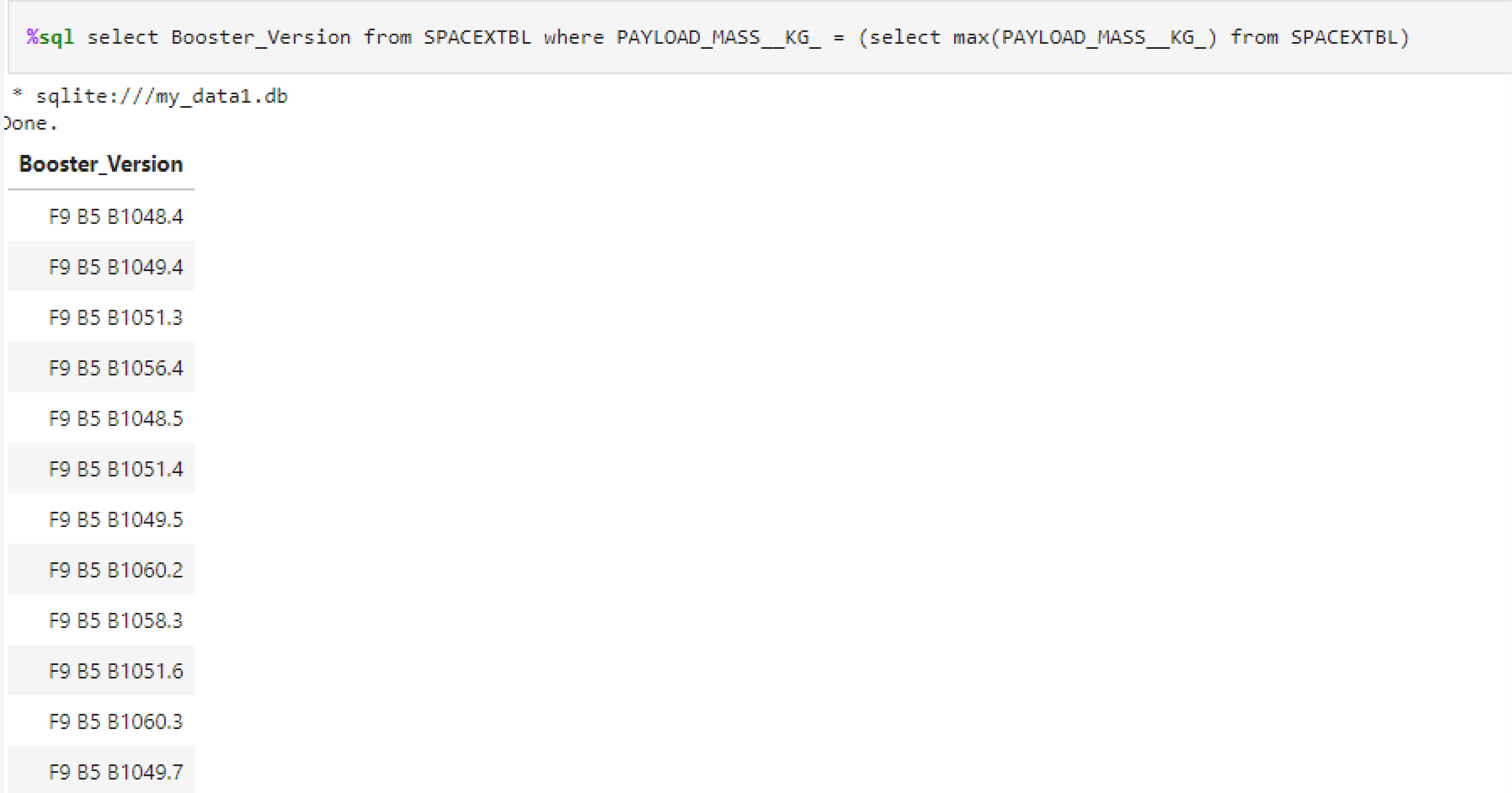
```
%sql select count(Mission_Outcome) from SPACEXTBL WHERE Mission_Outcome = 'Success' OR Mission_Outcome = 'Failure (in fligh
```

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

count(Mission_Outcome)
99

- There are 99 missions where the outcome was registered.

Boosters Carried Maximum Payload



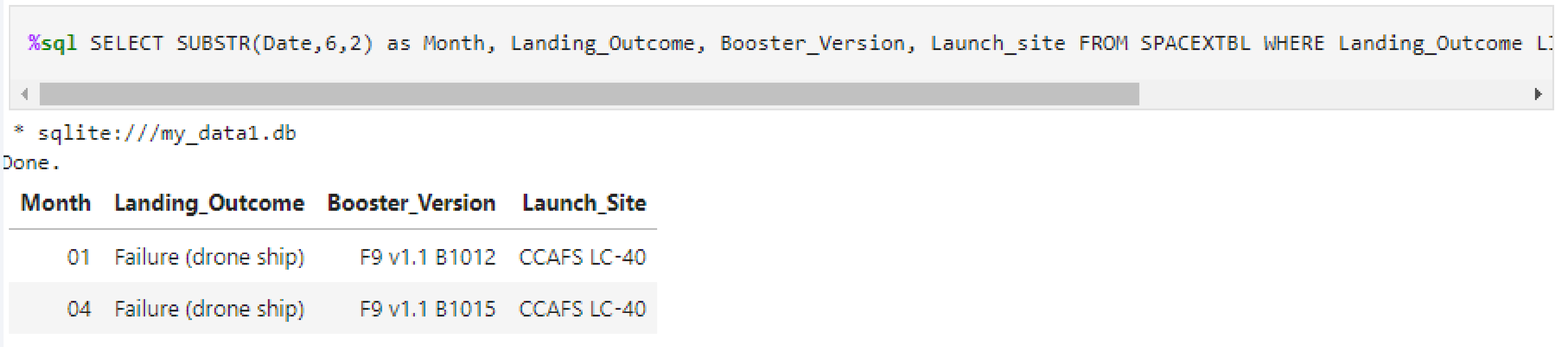
```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

- These Boosters had carried the highest Payload

2015 Launch Records



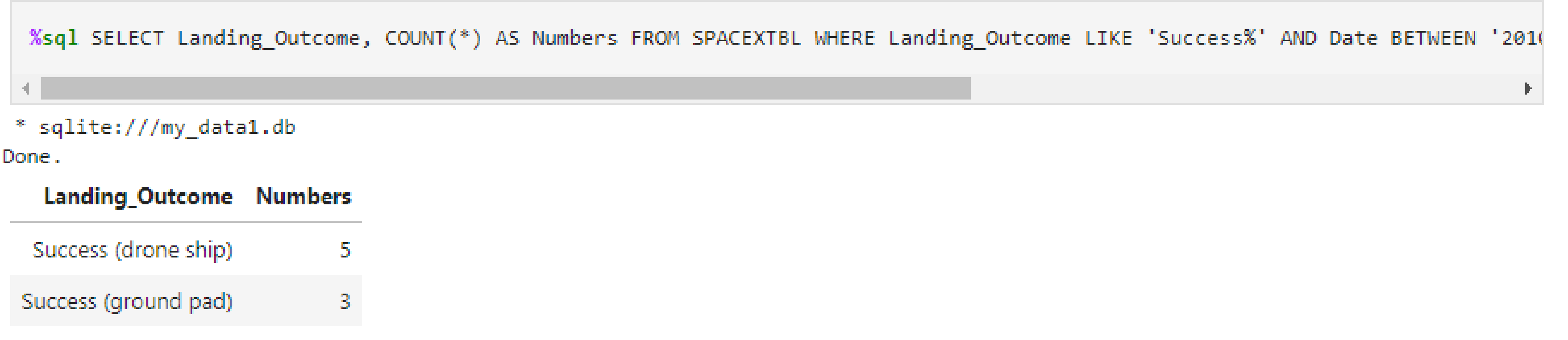
```
%sql SELECT SUBSTR(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_site FROM SPACEXTBL WHERE Landing_Outcome LI
```

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

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

- Failure results for year 2015

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



```
%sql SELECT Landing_Outcome, COUNT(*) AS Numbers FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success%' AND Date BETWEEN '201
```

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

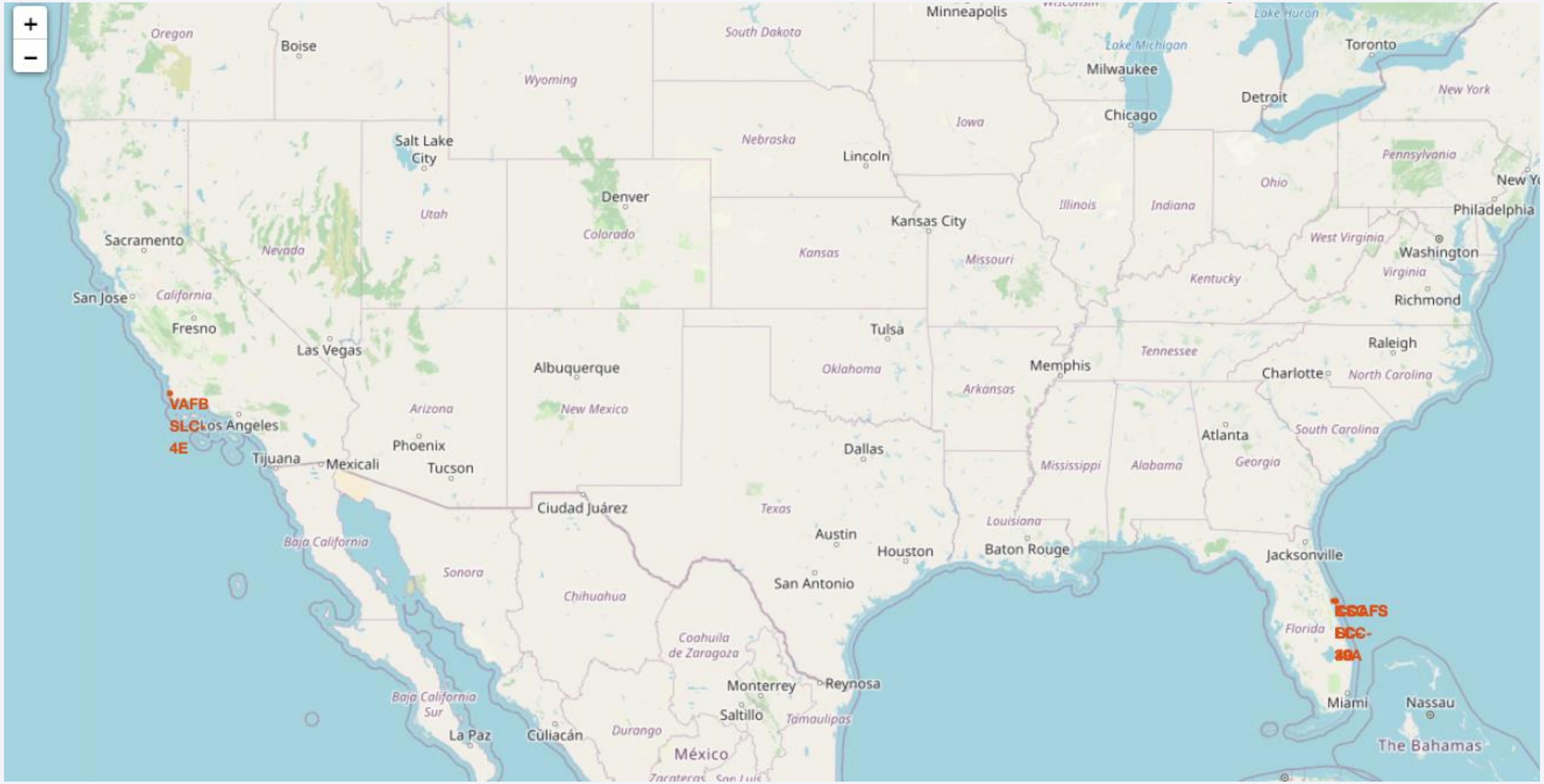
Landing_Outcome	Numbers
Success (drone ship)	5
Success (ground pad)	3

- Landing outcomes that happened between 2010-06=04 and 2017-03-20

Section 3

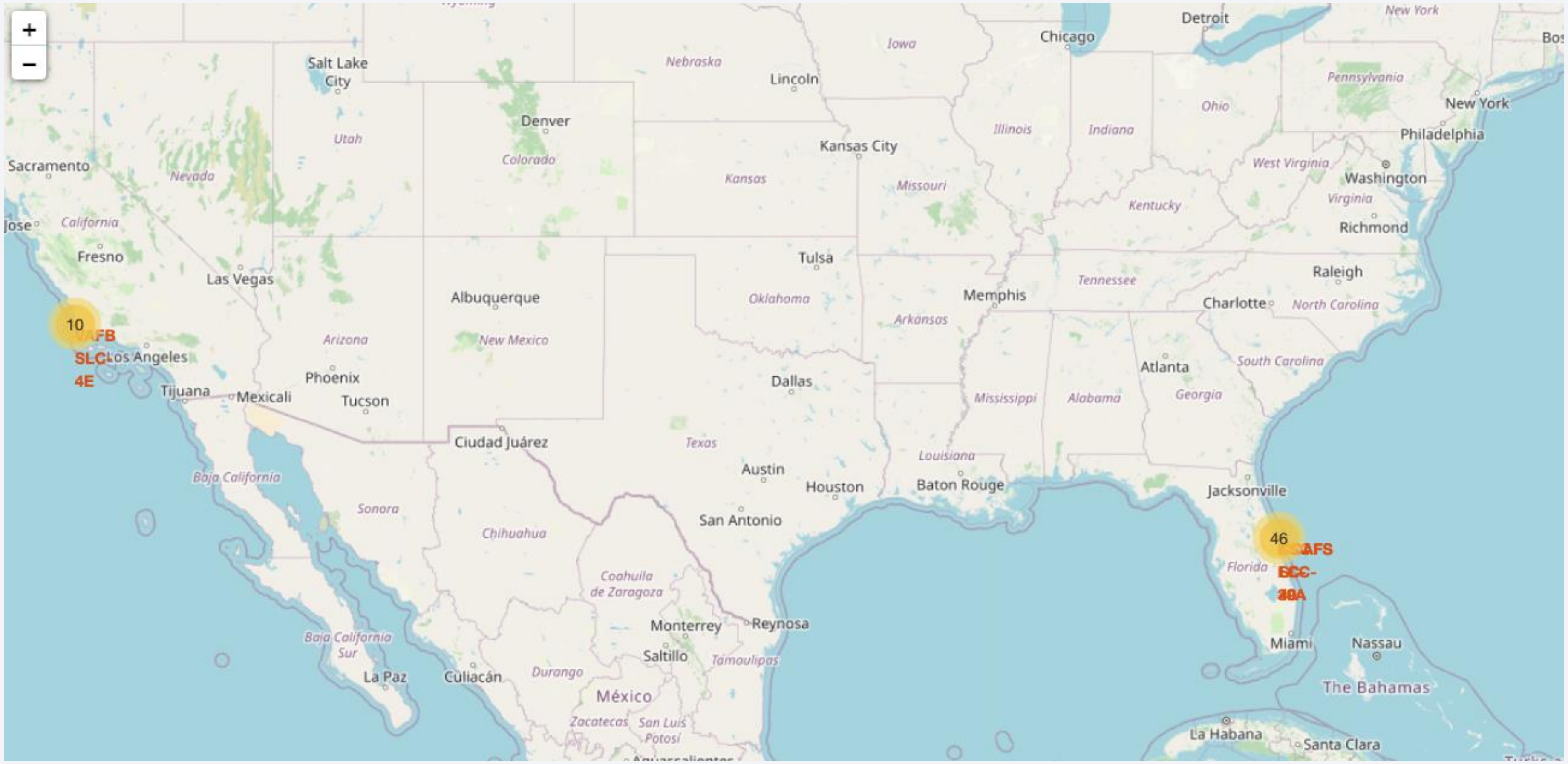
Launch Sites Proximities Analysis

Launch sites map



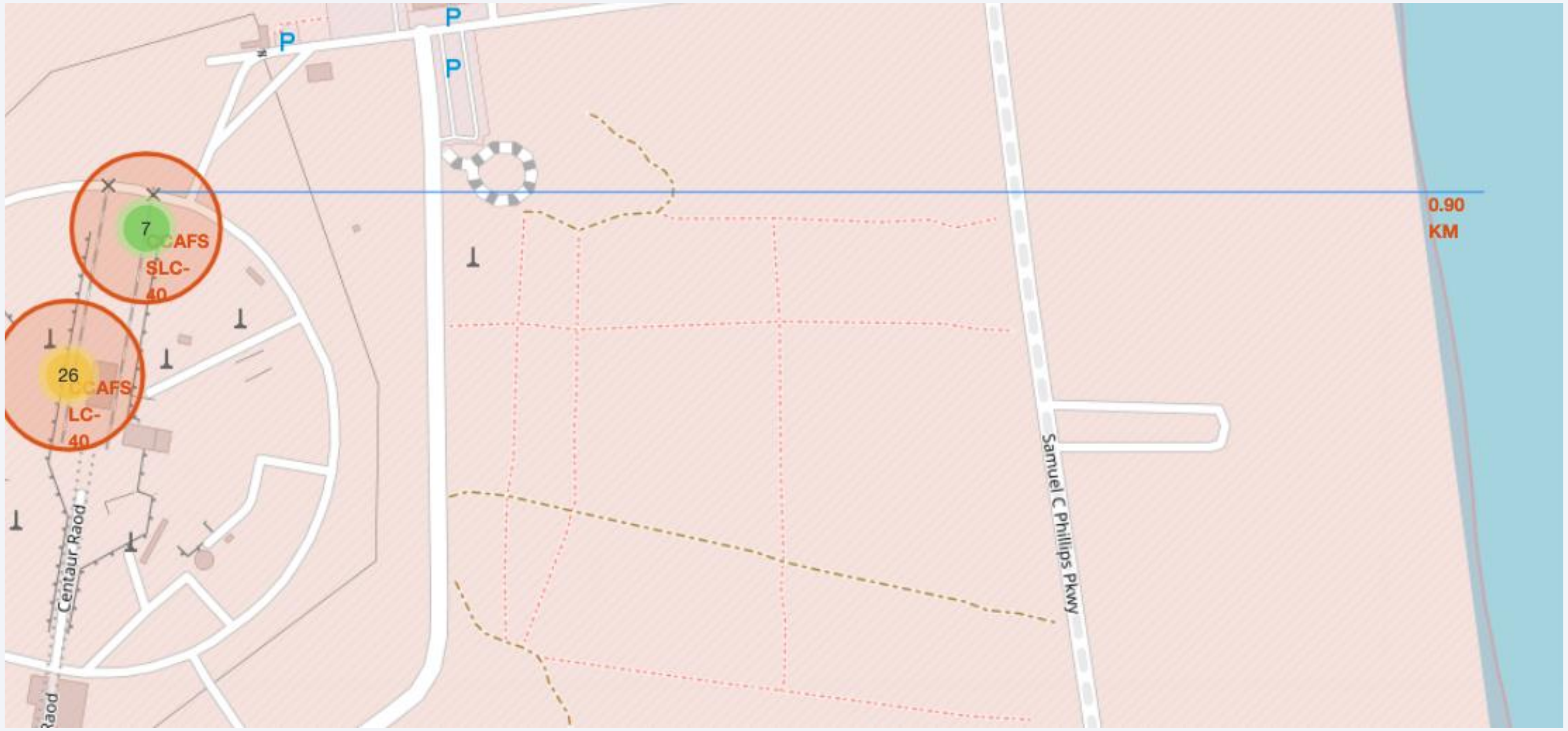
- This map shows where the launching sites are located

Number of launches for each site

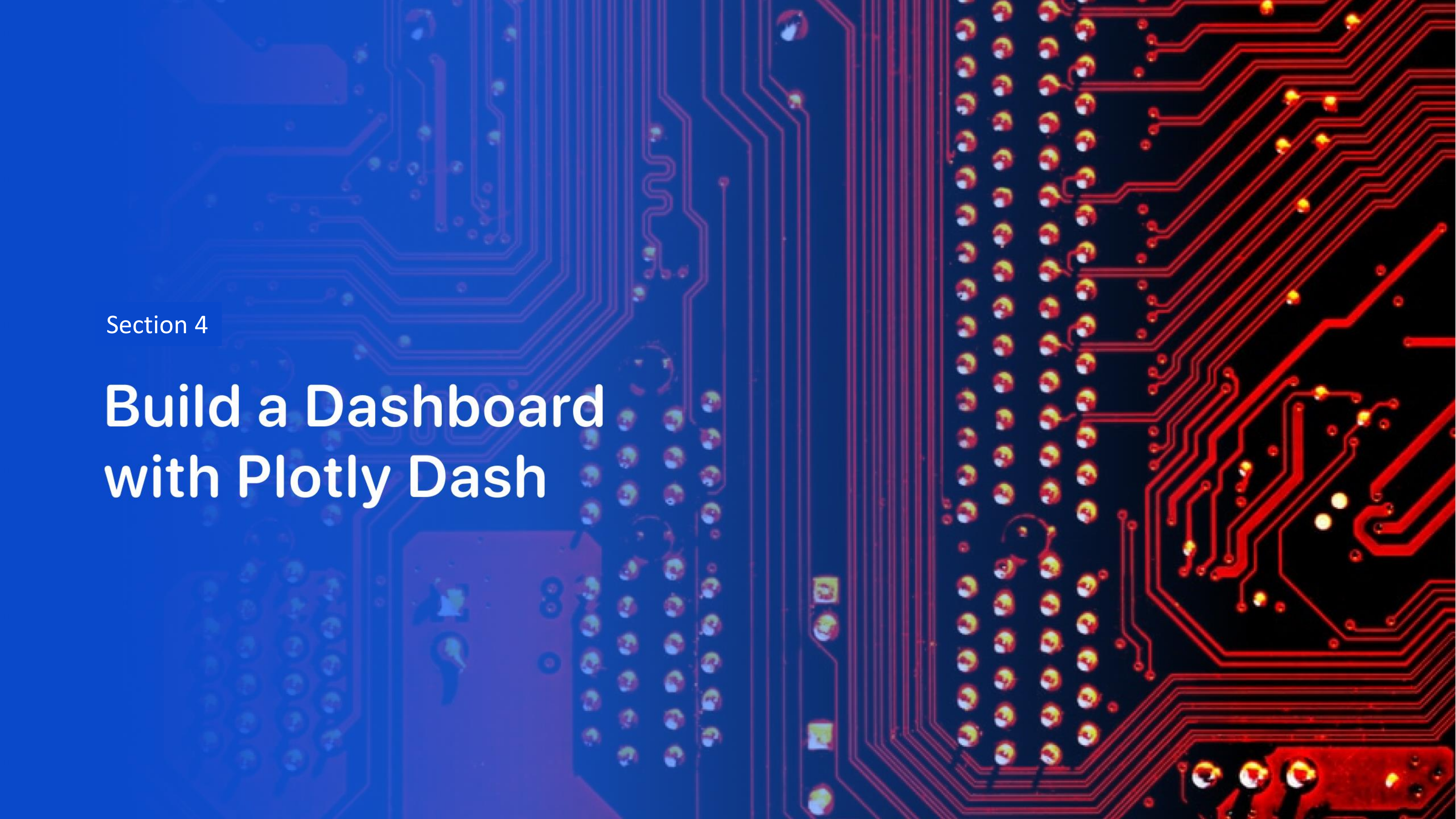


- This map shows the number of launches for each site.

Distance to coast



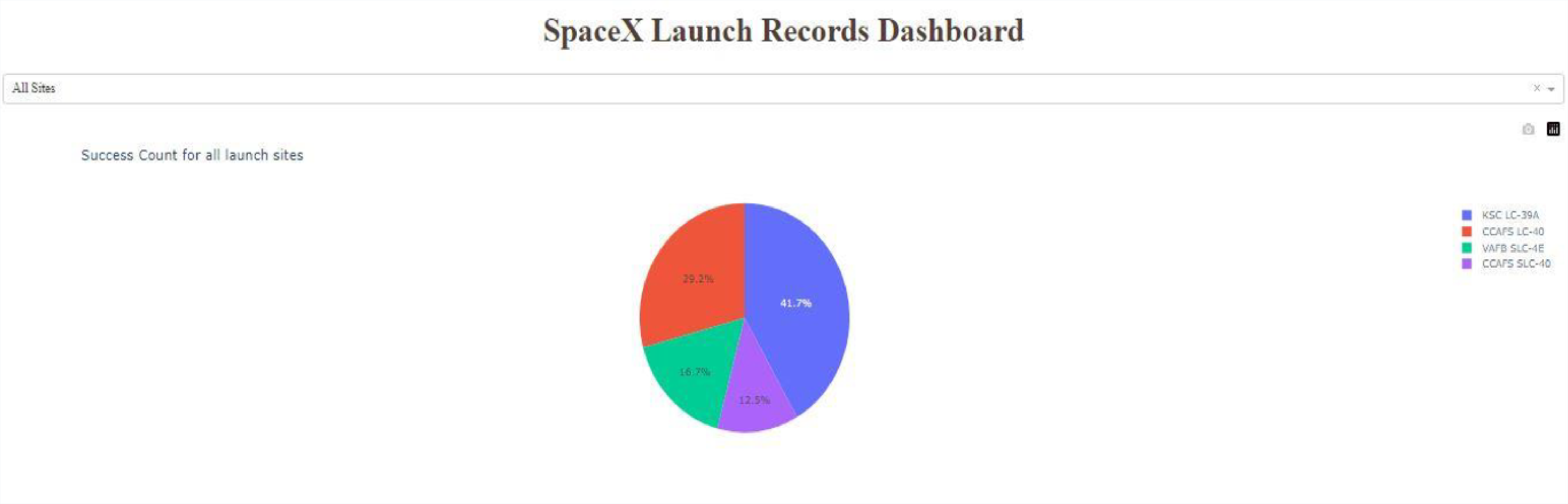
- This map shows the distance to the coast and the success rate for each launch.



Section 4

Build a Dashboard with Plotly Dash

Launch success for all sites

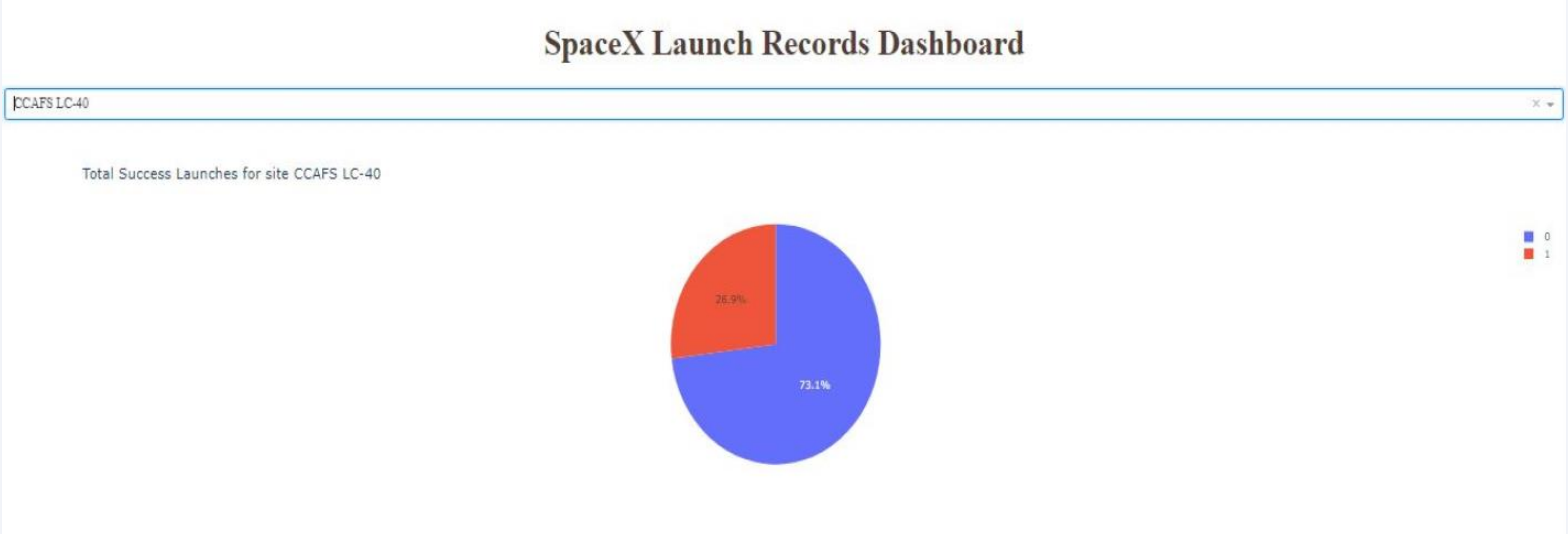


SpaceX Launch Records Dashboard

All Sites

- This interactive graph shows that site KSC LC-39A has the best success count.

Data for specific site

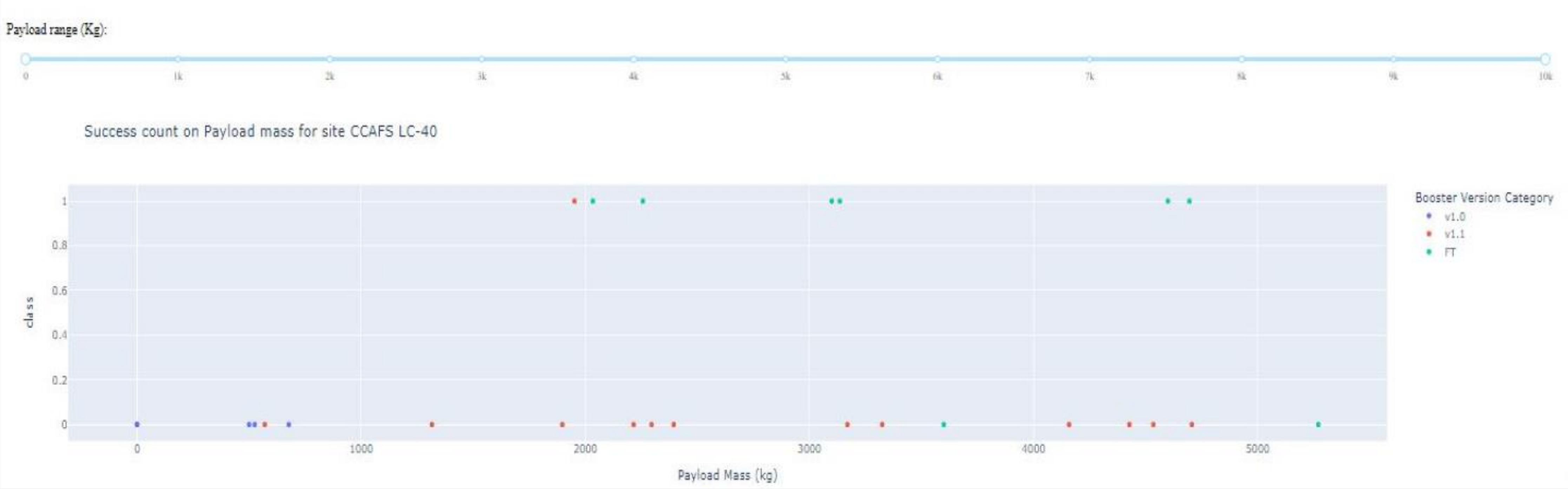


SpaceX Launch Records Dashboard

CCAFS LC-40

- Data for a specific site can be shown, like the success/failures count for site CCAFS LC/40

Landing result and Payload Mass

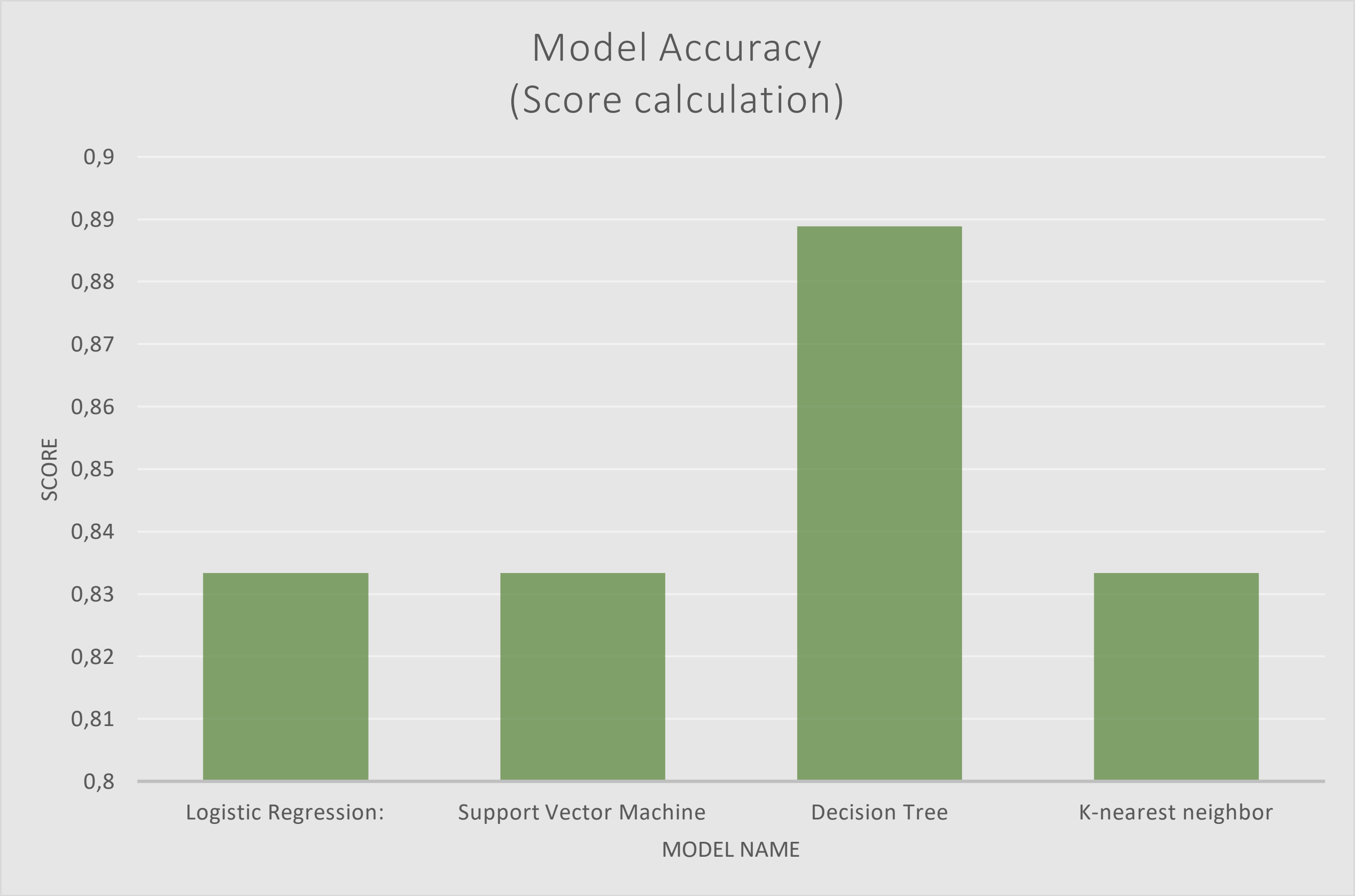


- Payload analysis. The “Range” widget can be moved to modify the Payload Mass and show the outcome.

Section 5

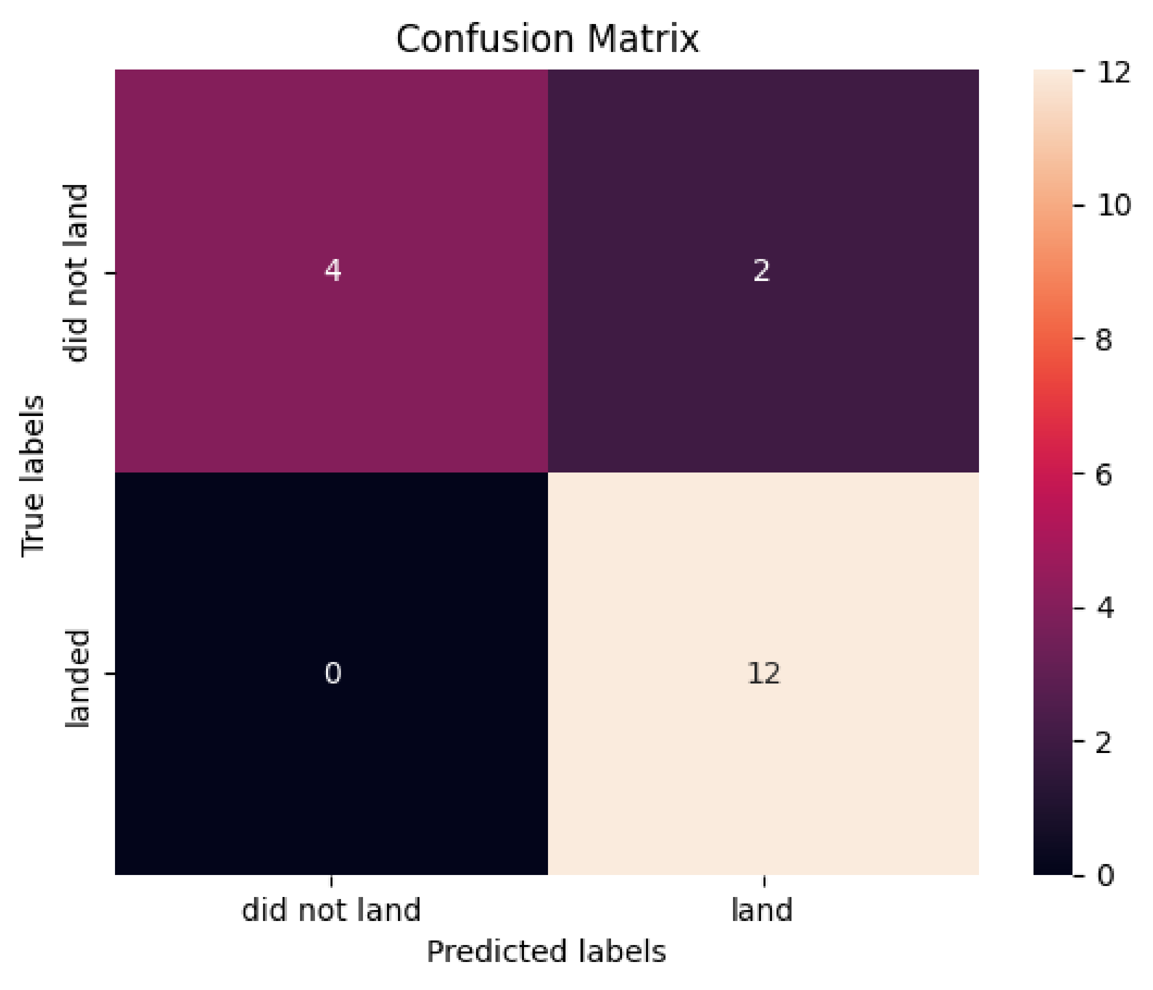
Predictive Analysis (Classification)

Classification Accuracy



- The accuracy of all models is very similar. The one that is slightly higher is for for the Decision Tree model.

Confusion Matrix for Decision Tree Model



- The matrix shows that the number of False Positive and False Negative are low. In this case, it would mean a prediction of successful landing when the rocket crashed, or the prediction of a failure, when the landing was successful.

Conclusions

- There seems to be better outcome (i.e. a successful landing) when the Payload Mass is higher.
- Also, each launching site has different outcome rates (success vs failure) and site KSC LC 39A seems to be the best place to launch.
- Orbit is not really a good predictor of success or failure. Also, it may be difficult to restrict missions to reach only a specific orbit.
- Another clear trend is that the success rate has been improving with time. This means that the experience gathered by the team is reflected in the landing outcome.

Conclusions

- Regarding the prediction model, all models we tried yielded an acceptable value, however the results for the Decision Tree was slightly better.
- The Confusion Matrix for the Decision Tree model shows a low number of False Positives and False Negatives, therefore we recommend to use this model to predict future outcomes.
- As a final note, if, for some reason, the Decision Tree model is hard to use, any of the other models could be used. Even better, a couple of models can be used to predict a result and check if all of them agree in the result.

Appendix

- None at this time.

Thank you!