IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

CHAU PHAM

17-08-2024

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection using API and Web Scrapping
  - Data Wrangling
  - EDA with SQL
  - Interactive Visual Analytics with Folium and Interactive Dashboard with Plotly Dash
  - Predictive Analysis with Machine Learning Classifications
- Summary of all results
  - Overview of the cleaned data
  - Interactive analytics
  - Predictive analytics

Introduction

- Project background and context
  - 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.
- Problems you want to find answers
  - I want to determine the cost of a launch based on the successful probability of the first stage will land.
  - 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:
  - Using SpaceX Rest API
  - Using Web Scrapping form Wikipedia
- Perform data wrangling
  - Using one-hot encoding to clean null values and remove irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Linear Regression, KNN, SVM, DT models were built and evaluated for the best method

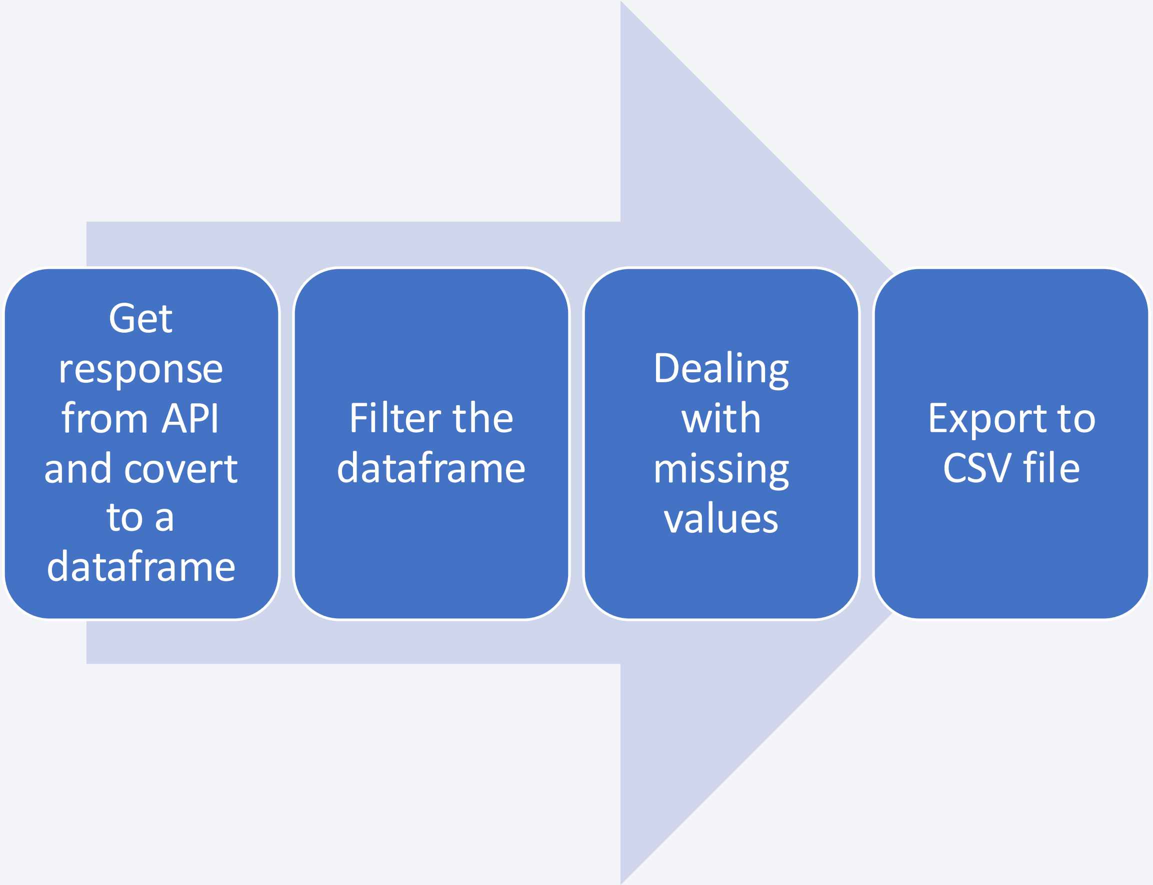
Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

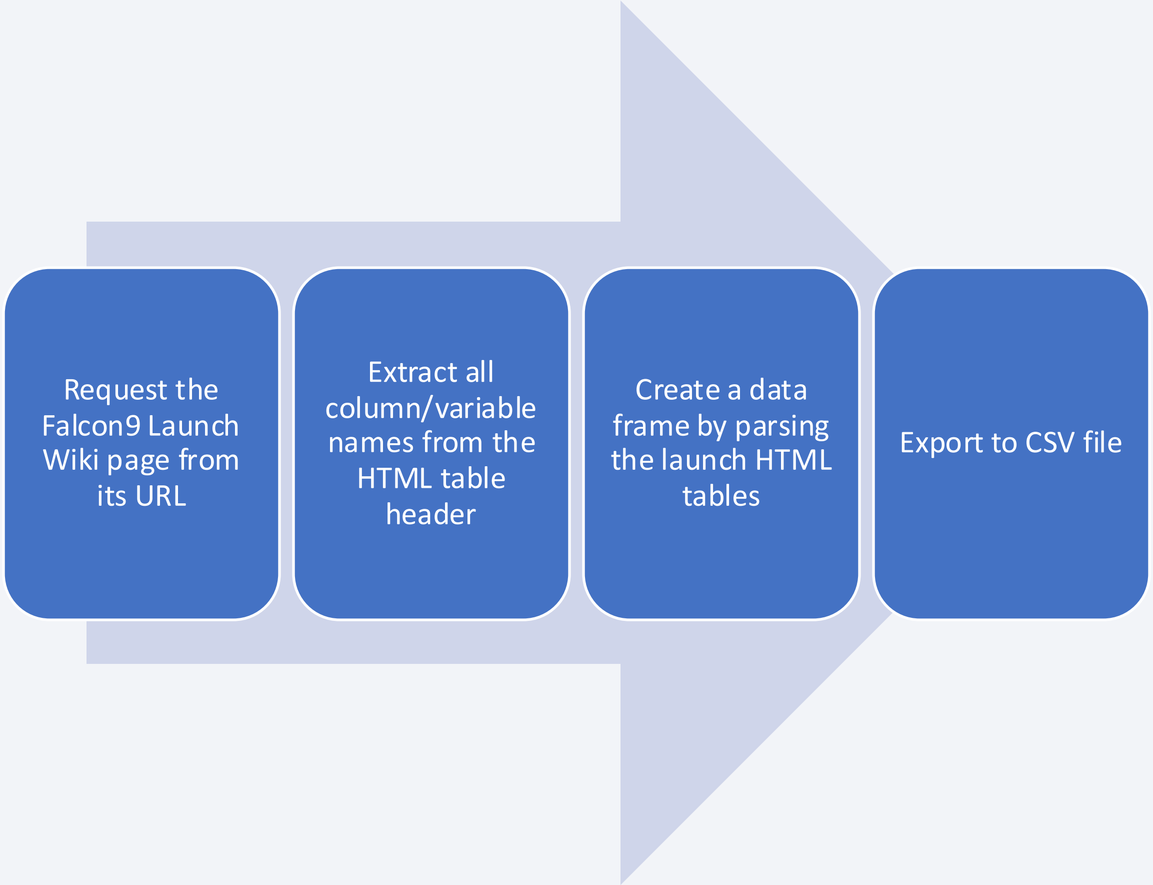
- GitHub: Link


Get response from API and covert to a dataframe
Filter the dataframe
Dealing with missing values
Export to CSV file

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- GitHub: [Link]



Request the Falcon9 Launch Wiki page from its URL

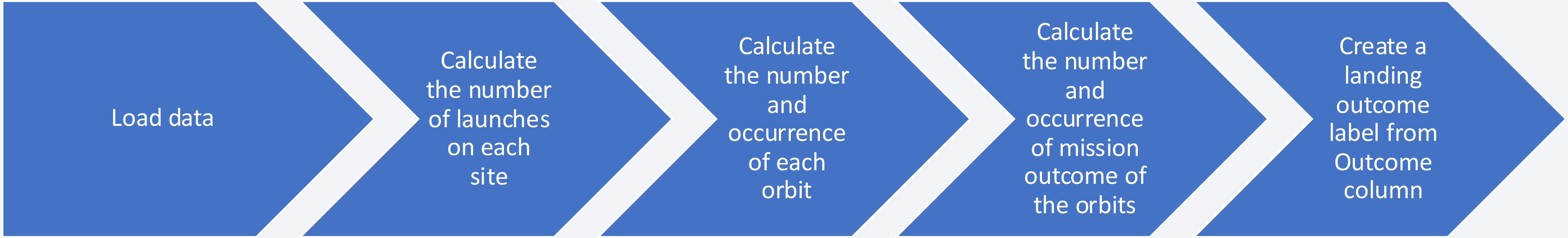
Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Export to CSV file

Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts



Load 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

- Github link: Link

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Catplot to visualize the relationship between flight number and payload.
  - Catplot to visualize the relationship between flight number and launch site.
  - Catplot to visualize the relational sip between payload and launch side.
  - Bar chart to visualize the relationship between success rate of each orbit type
  - Catplot to visualize the relationship between flight number and orbit type.
  - Catplot to visualize the relationship between payload and orbit type.
  - Line chart to visualize the launch success yearly trend.
- GitHub: Link

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - SQL queries performed include:
  - Displaying the name of the unique launch site in the space mission.
  - Displaying five record where launch sites begin with the string “KSC”.
  - Displaying the total payload mass carried by boosters launched by NASA (CRS).
  - Displaying average payload mass carried by booster version. F9 v1 .1.
  - Listing the data where the successful landing outcome in drone ship was achieved.
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000.
  - Listing the total number of successful and failure mission outcomes.
  - Listing the names of the booster version which have carried the maximum payload mass.
  - Listing the records which will display the month names, successful landing outcomes in ground pad, boosters Version, launch site for the months in year 2017.
  - Ranking the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order .

- GitHub URL: Link

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Geographical patterns about launch sites: the success/failed launches for each site on the map, calculate the distances between a launch site to its proximities.
  - A pie chart to show the total successful launches count for all sites and each site.
  - A scatter chart to show the correlation between payload and launch success.
- I added those objects to finding an optimal location for building a launch site
- GitHub URL: Link

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - A callback function to render the success-payload-scatter-chart scatter plot.
  - To visually observe how payload may be correlated with mission outcome for selected site.
- Explain why you added those plots and interactions:

  To answer these questions:

  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
- GitHub URL: Link

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model

  The SVM, KNN and Logistic Regression model achieved model achieved the highest accuracy at 83.3%, while the SVM performs the best in terms of Area Under the Curve at 0.958

- You need present your model development process using key phrases and flowchart

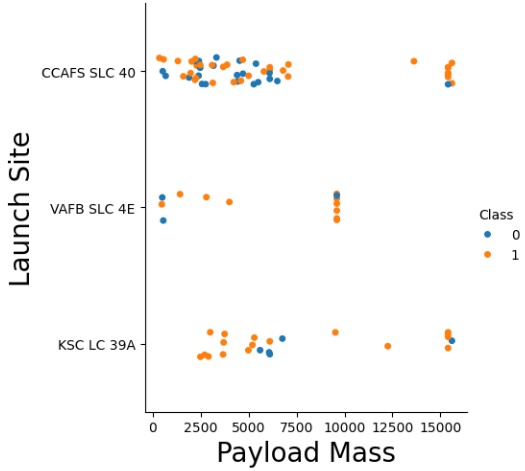
Load the DataFrame → Standardize data → Train - test split

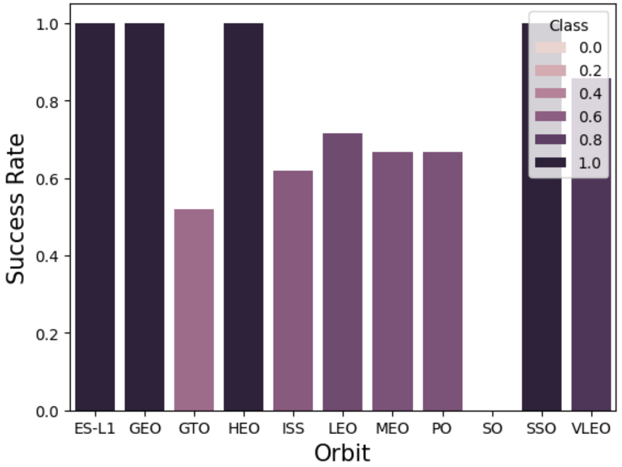
→ Use Logistic Regression, SVM, Decision Tree and KNN for Classification → Compare results from these models

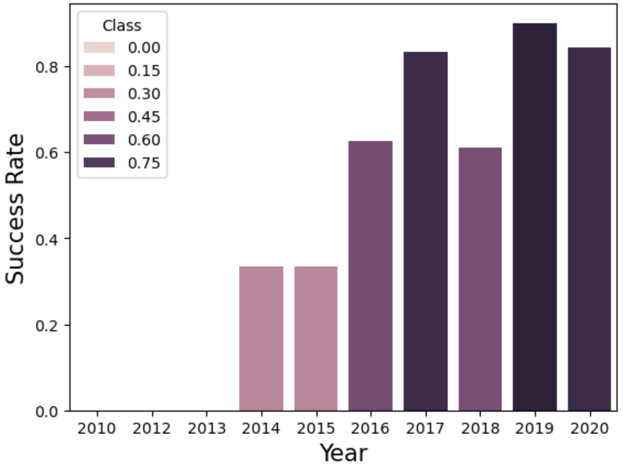
- GitHub URL: Link

Results

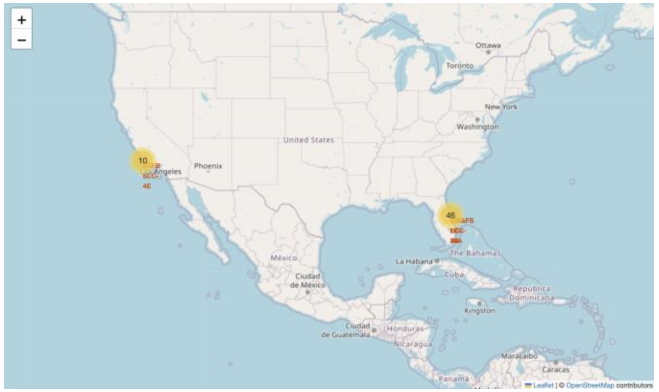
- Exploratory data analysis results

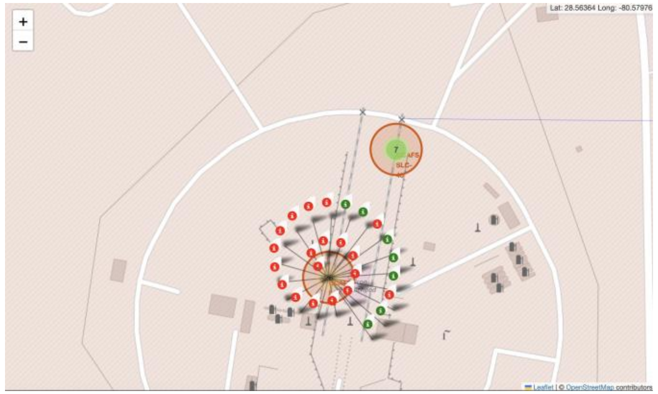






- Interactive analytics demo in screenshots




+
−
Lat: 28.56364 Long: -80.57976
7
Leaflet | © OpenStreetMap contributors

+
−
Lat: 28.58082 Long: -80.61622
False Cape
33
0.97 KM
0.58 KM
0.86 KM
Leaflet | © OpenStreetMap contributors

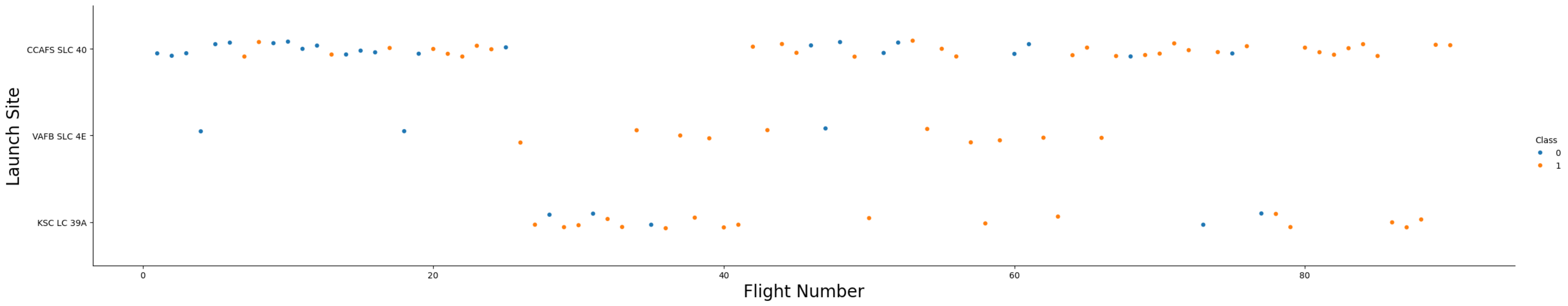
Results

- Predictive analysis results
  - ✓ The SVM, KNN and Logistic regression models are the best in term of prediction accuracy of the dataset.
  - ✓ Low weighted payloads perform better than the heavier payloads.
  - ✓ The success rate for SpaceX launch is directly proportional time in years they will eventually perfect the launches.
  - ✓ KSC LC-39A had the most successful launches from all the sites.
  - ✓ Orbit GEO, HEO, SSO, ES L1 has the best success rate.

Section 2

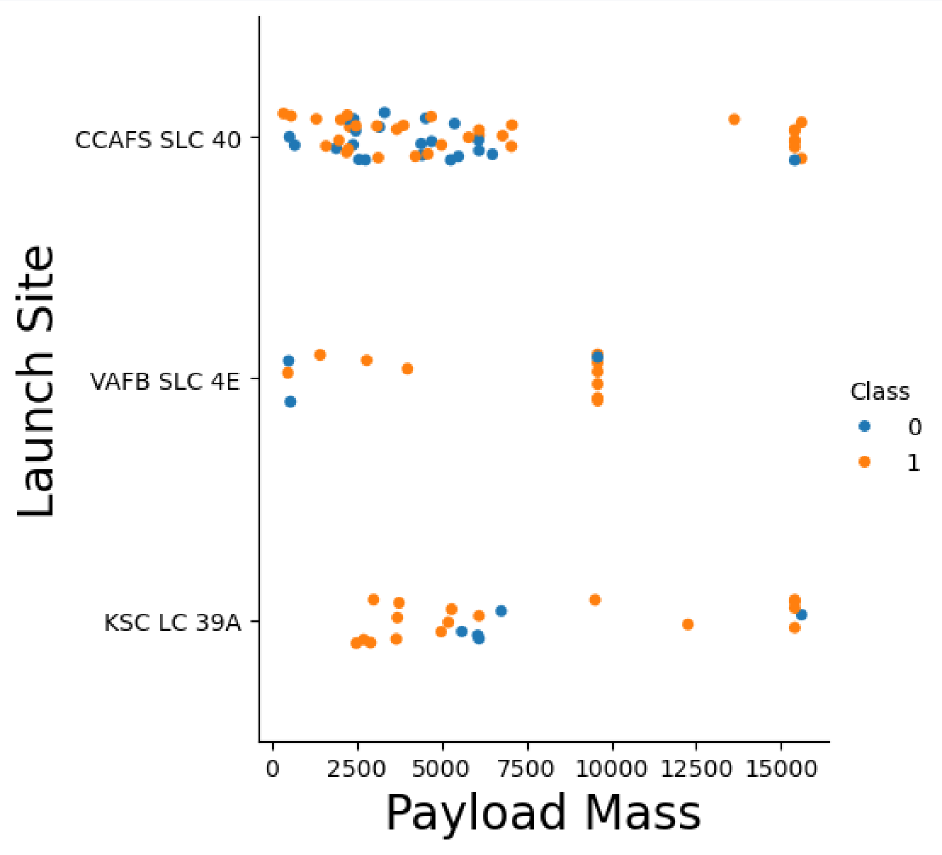
Insights drawn from EDA

Flight Number vs. Launch Site



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

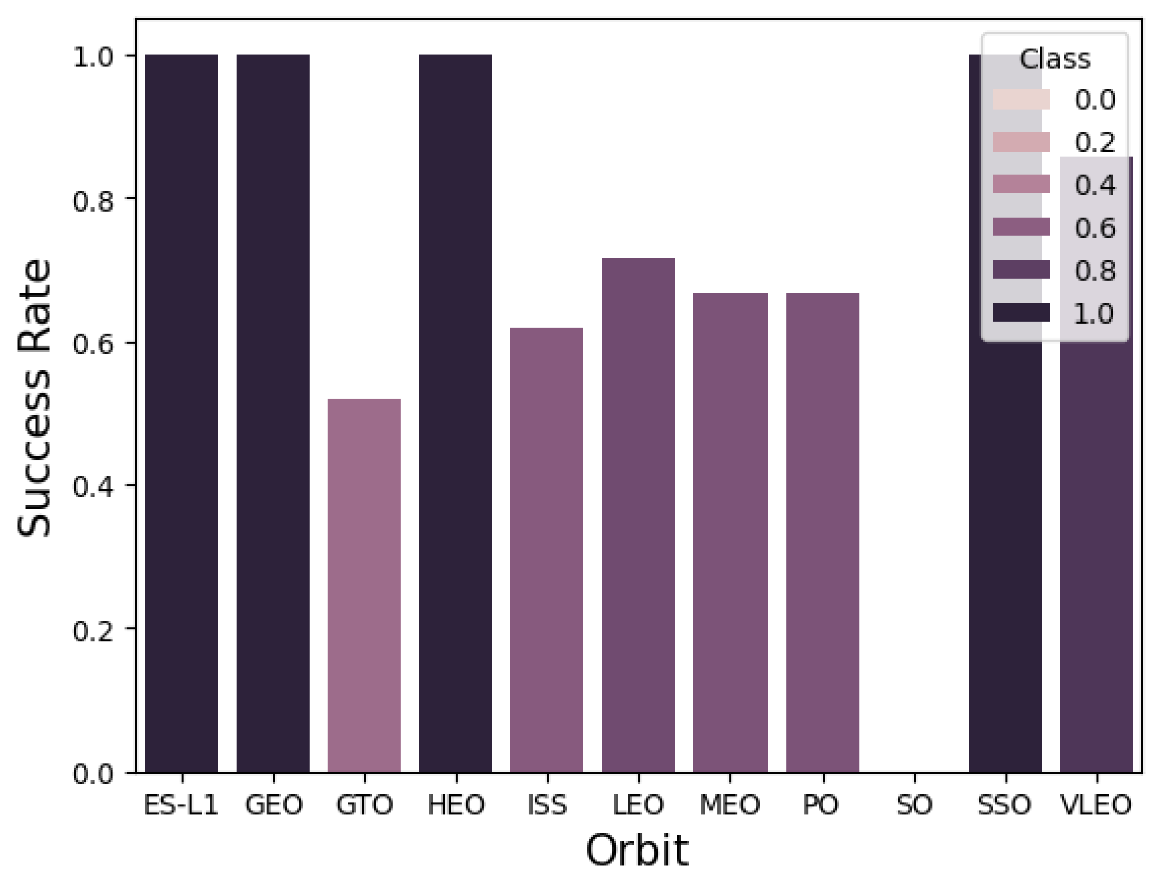
Payload vs. Launch Site



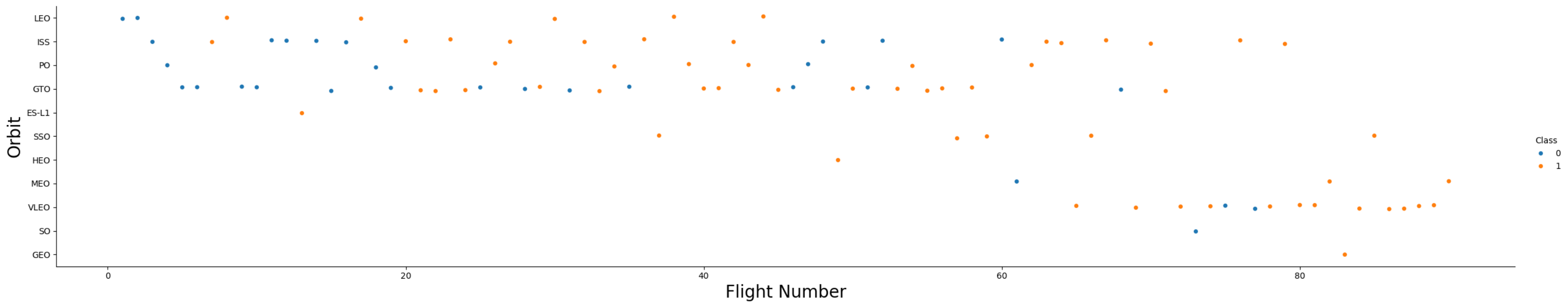
- For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000)
- The majority are rockets launched for light payload (less than 10000)

Success Rate vs. Orbit Type

- For ES-L1, GEO, HEO, SSO, success rates are 100%
- SO orbit has the worst success rate (0%)

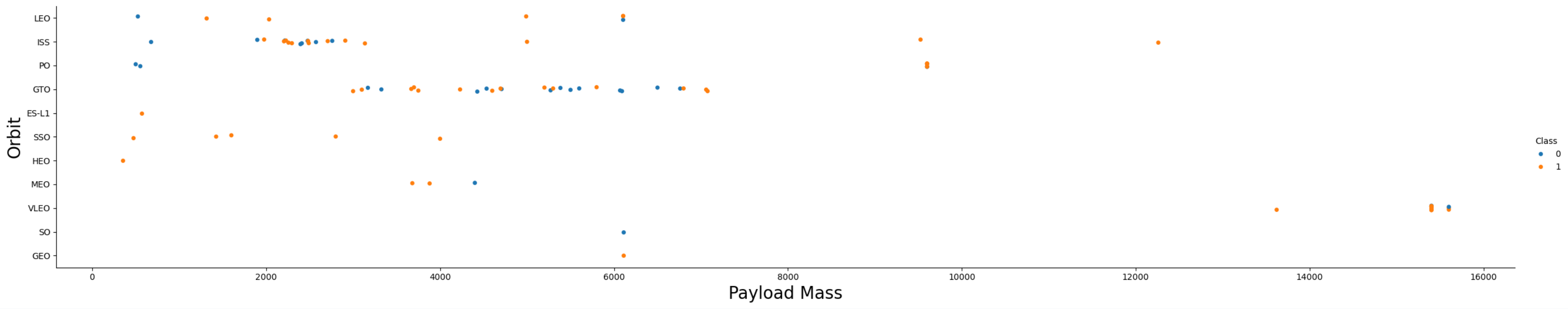


Flight Number vs. Orbit Type



- 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.

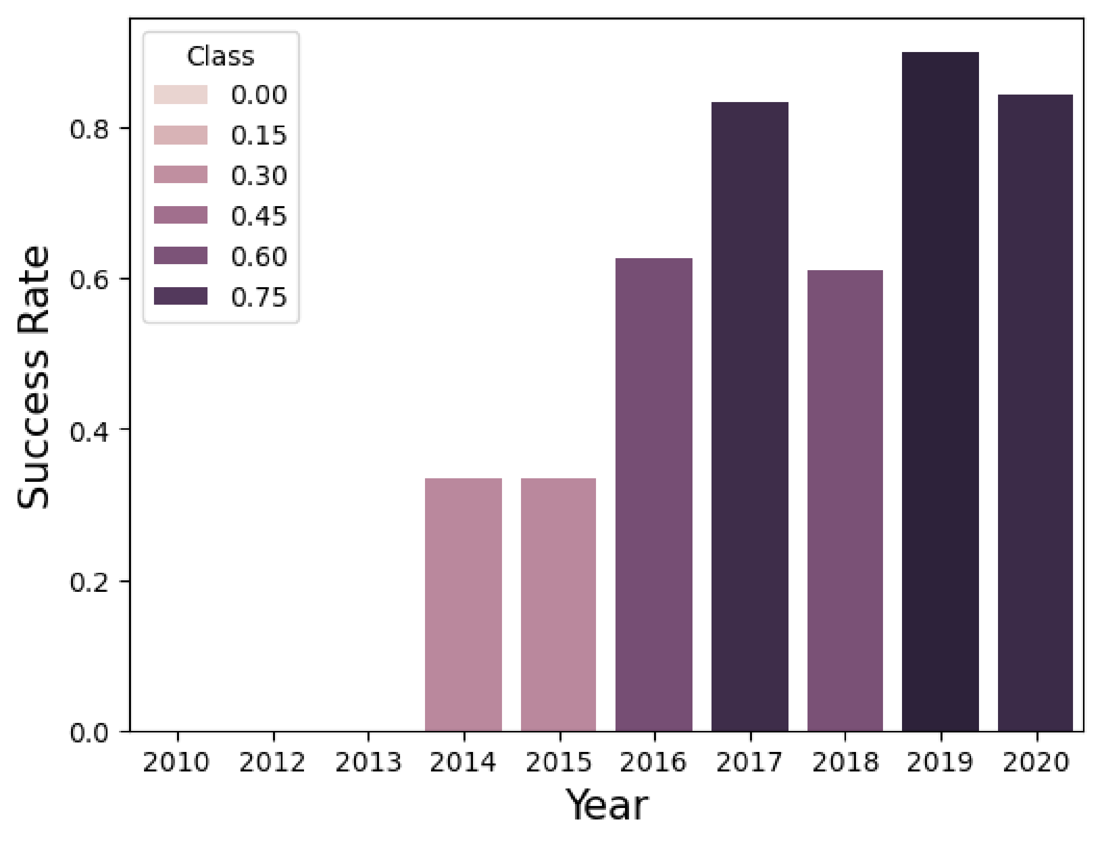
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for PO, VLEO 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

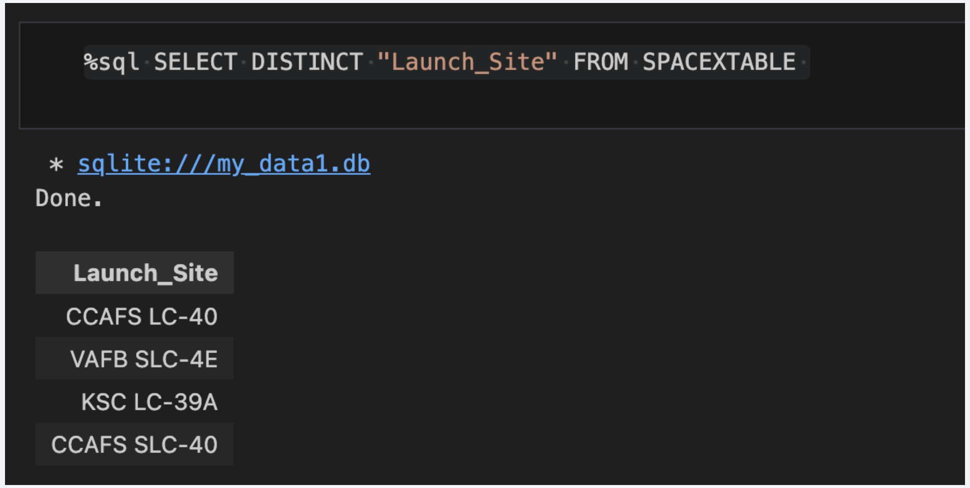


All Launch Site Names

- Find the names of the unique launch sites

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

- Present your query result with a short explanation here:

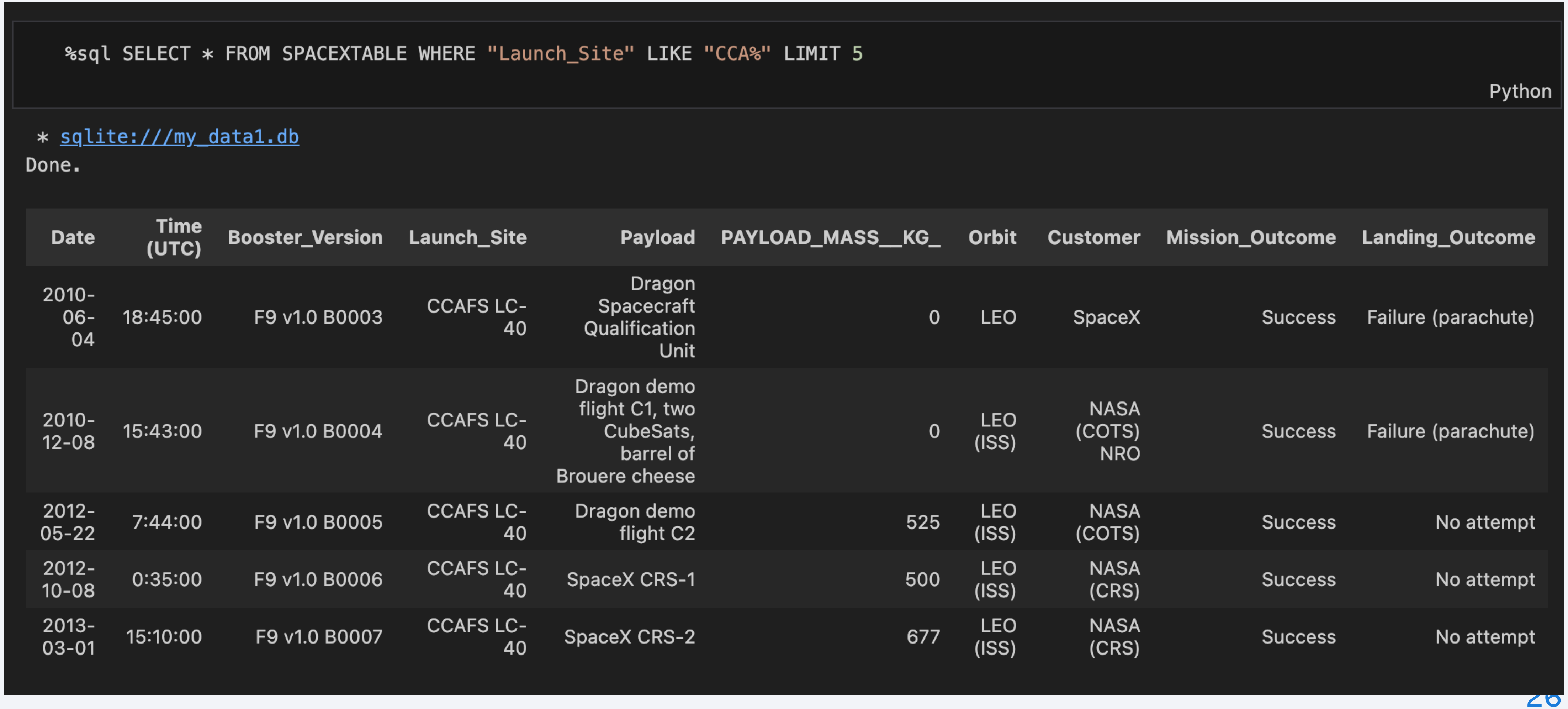


```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

```
 * 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'



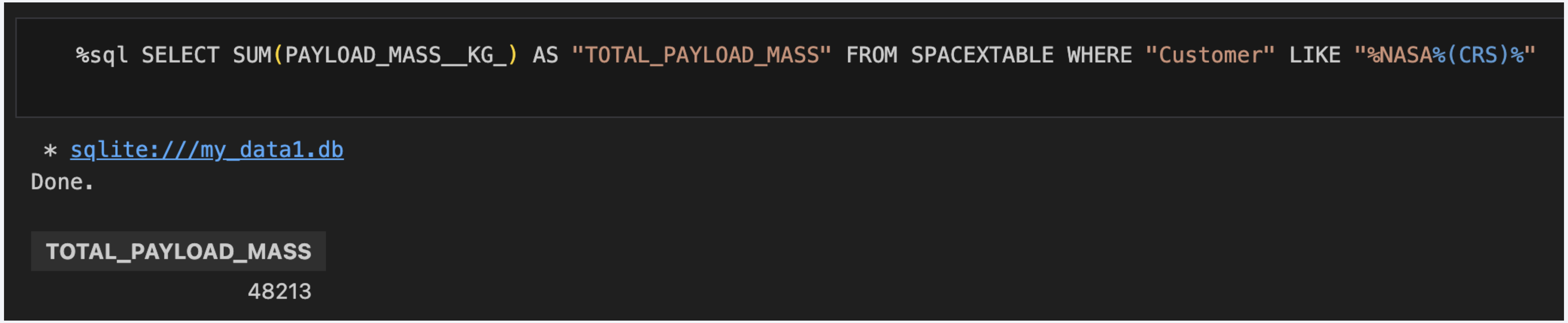
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
```

```
 * 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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA: 48213



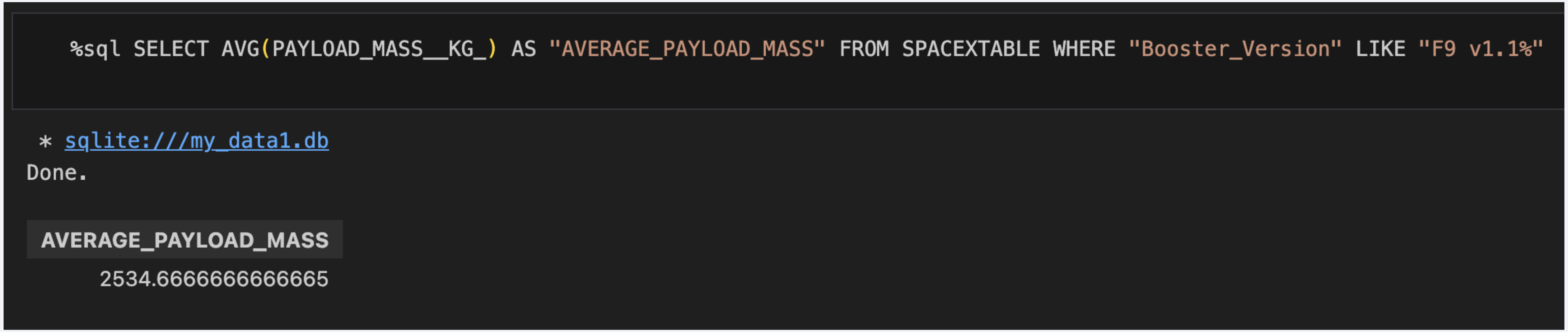
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "TOTAL_PAYLOAD_MASS" FROM SPACEXTABLE WHERE "Customer" LIKE "%NASA%(CRS)%"
```

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

TOTAL_PAYLOAD_MASS
48213

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here



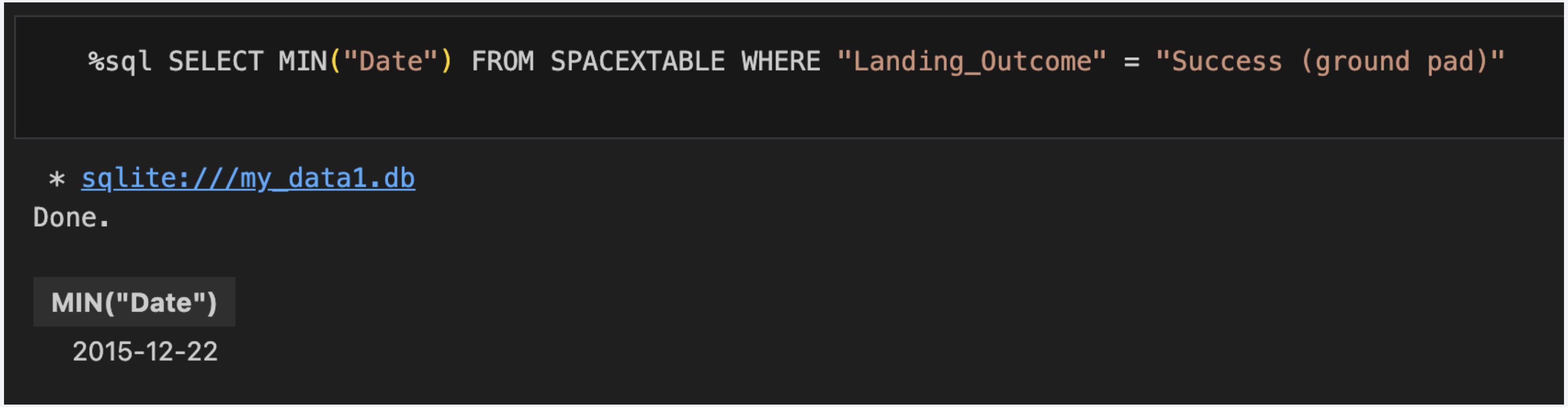
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "AVERAGE_PAYLOAD_MASS" FROM SPACEXTABLE WHERE "Booster_Version" LIKE "F9 v1.1%"
```

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

AVERAGE_PAYLOAD_MASS
2534.666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here



```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = "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

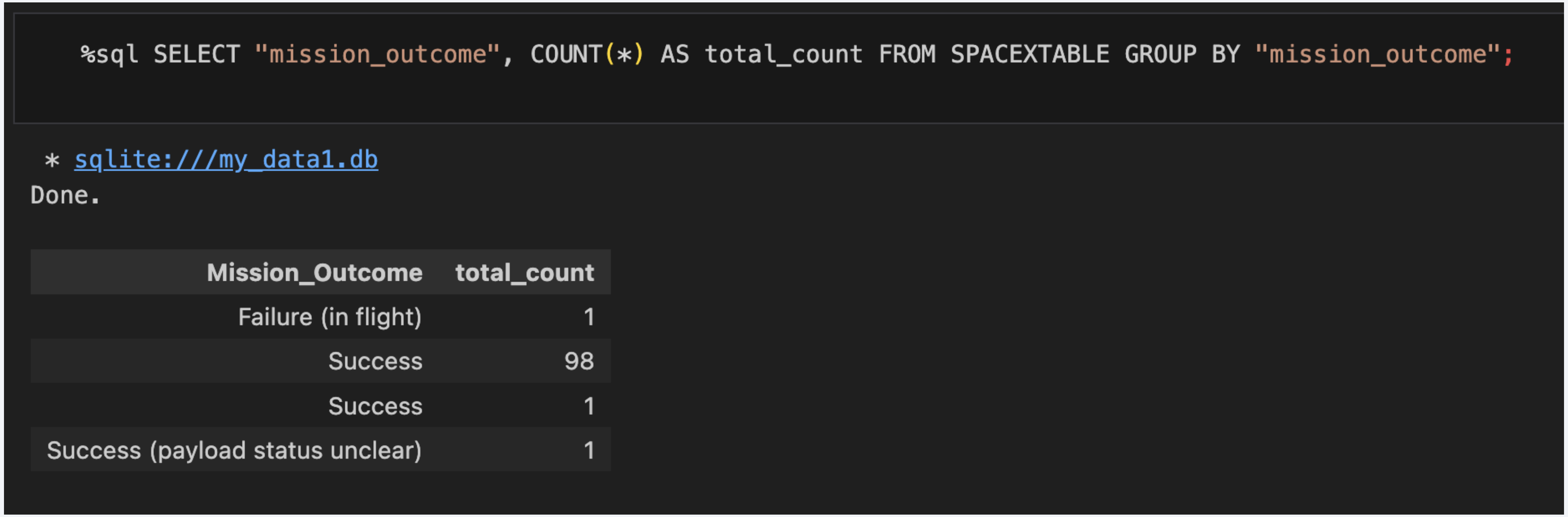
  F9 FT B1022
  F9 FT B1026
  F9 FT B1021.2
  F9 FT B1031.2

- Present your query result with a short explanation here:

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (drone ship)" AND
("PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000)
```

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here



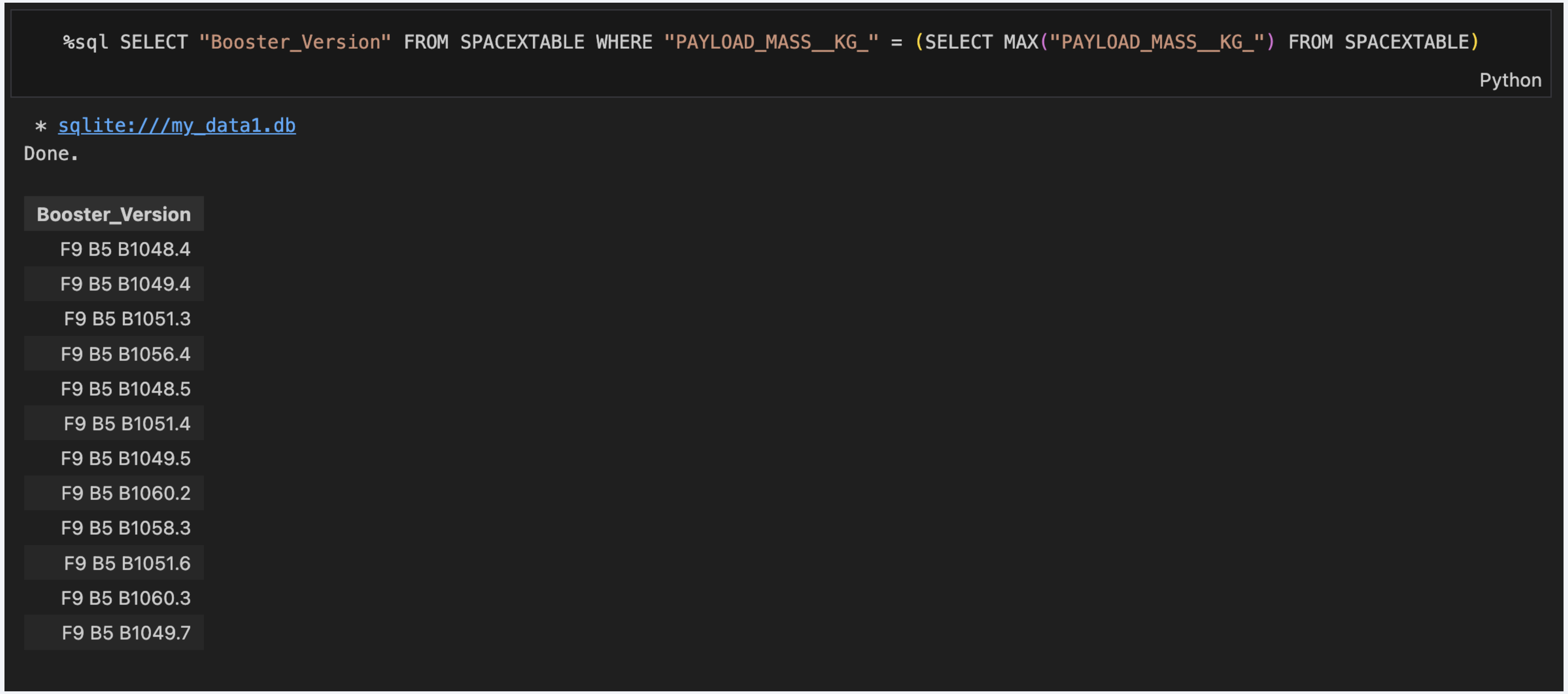
```
%sql SELECT "mission_outcome", COUNT(*) AS total_count FROM SPACEXTABLE GROUP BY "mission_outcome";
```

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

Mission_Outcome	total_count
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
- Present your query result with a short explanation here



```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)
```

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

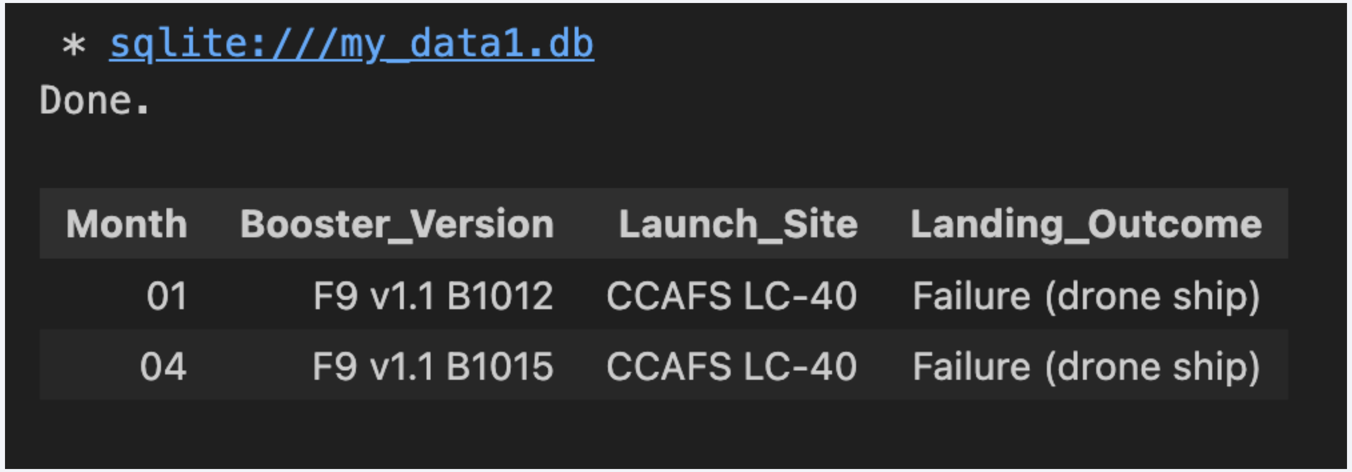
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

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", "Booster_Version", "Launch_Site", "Landing_Outcome" FROM
SPACEXTABLE WHERE "Landing_Outcome"="Failure (drone ship)" AND substr(Date,0,5)="2015"
```

- Present your query result with a short explanation here



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

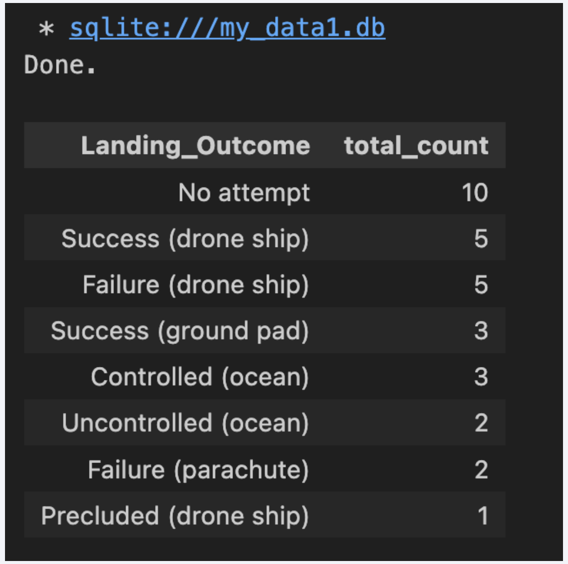
Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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(*) AS total_count FROM SPACEXTABLE WHERE "Date" BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY "landing_outcome" ORDER BY "total_count" DESC
```

- Present your query result with a short explanation here



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

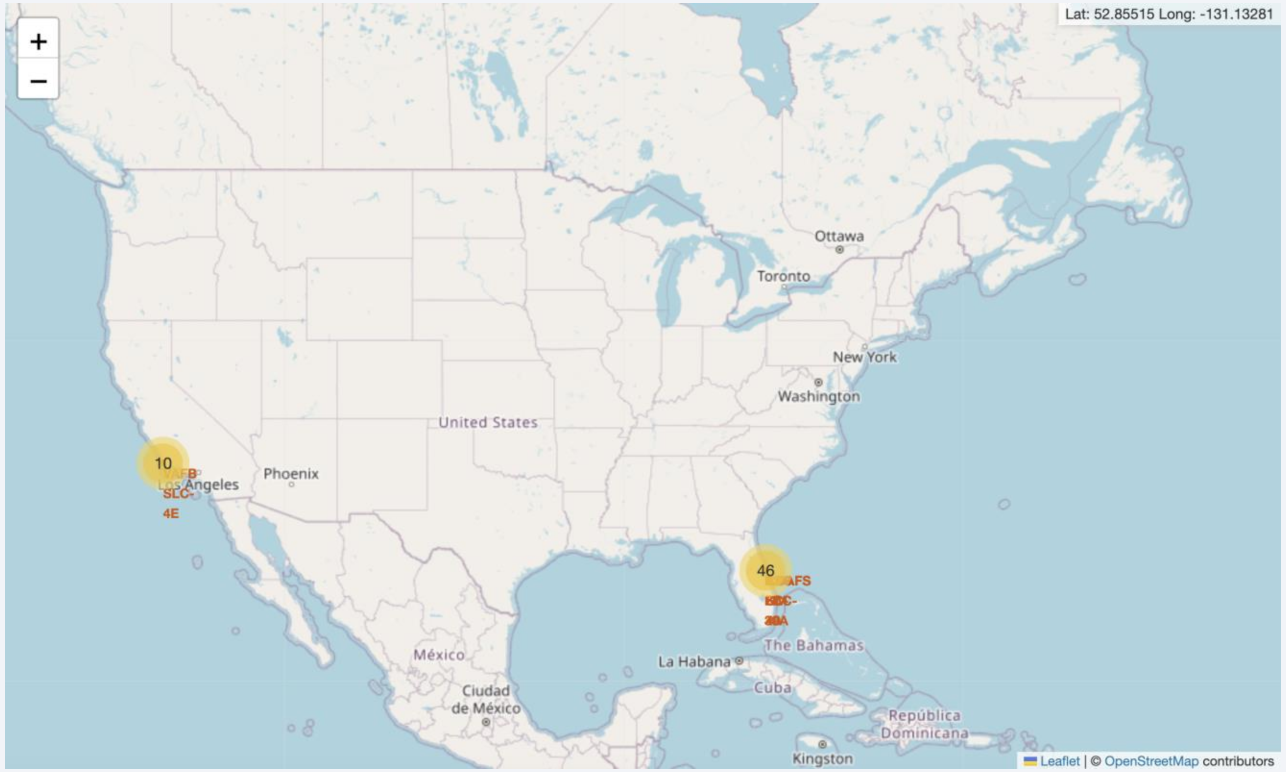
Landing_Outcome	total_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

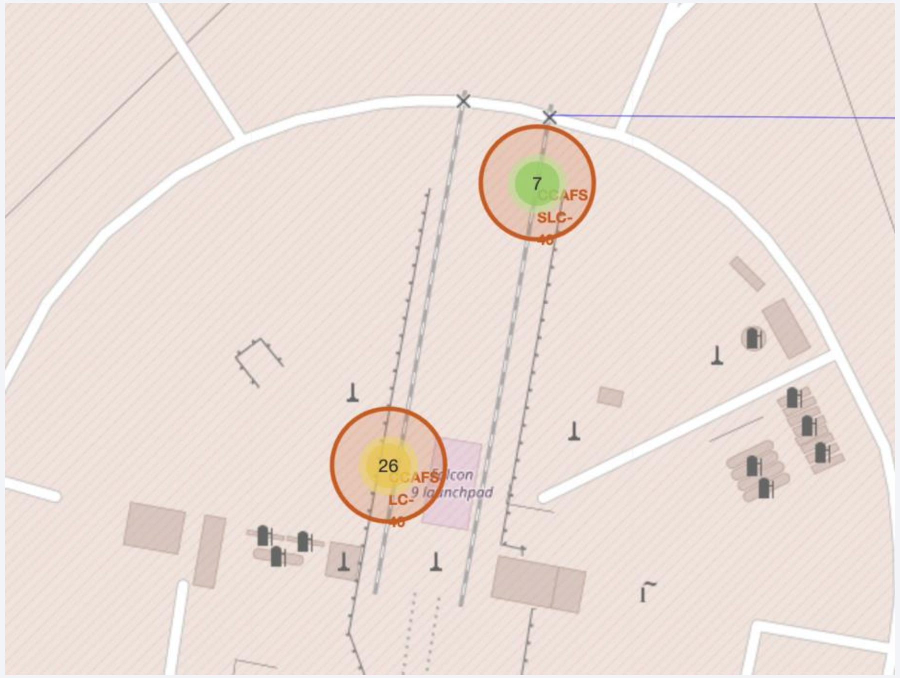
Launch Sites Proximities Analysis

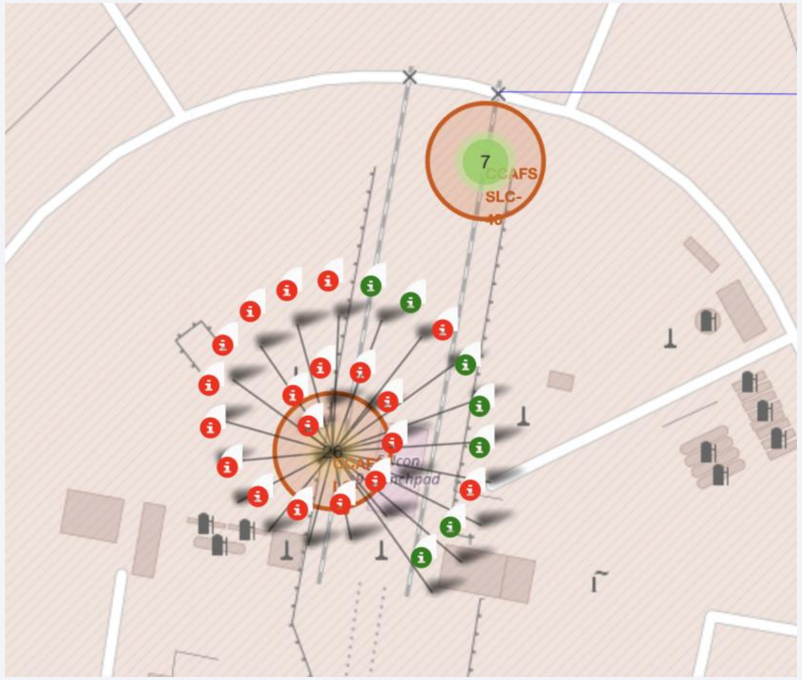
Overview of SpaceX launch sites

- There 2 main launch areas (one in west coast and one in east coast)



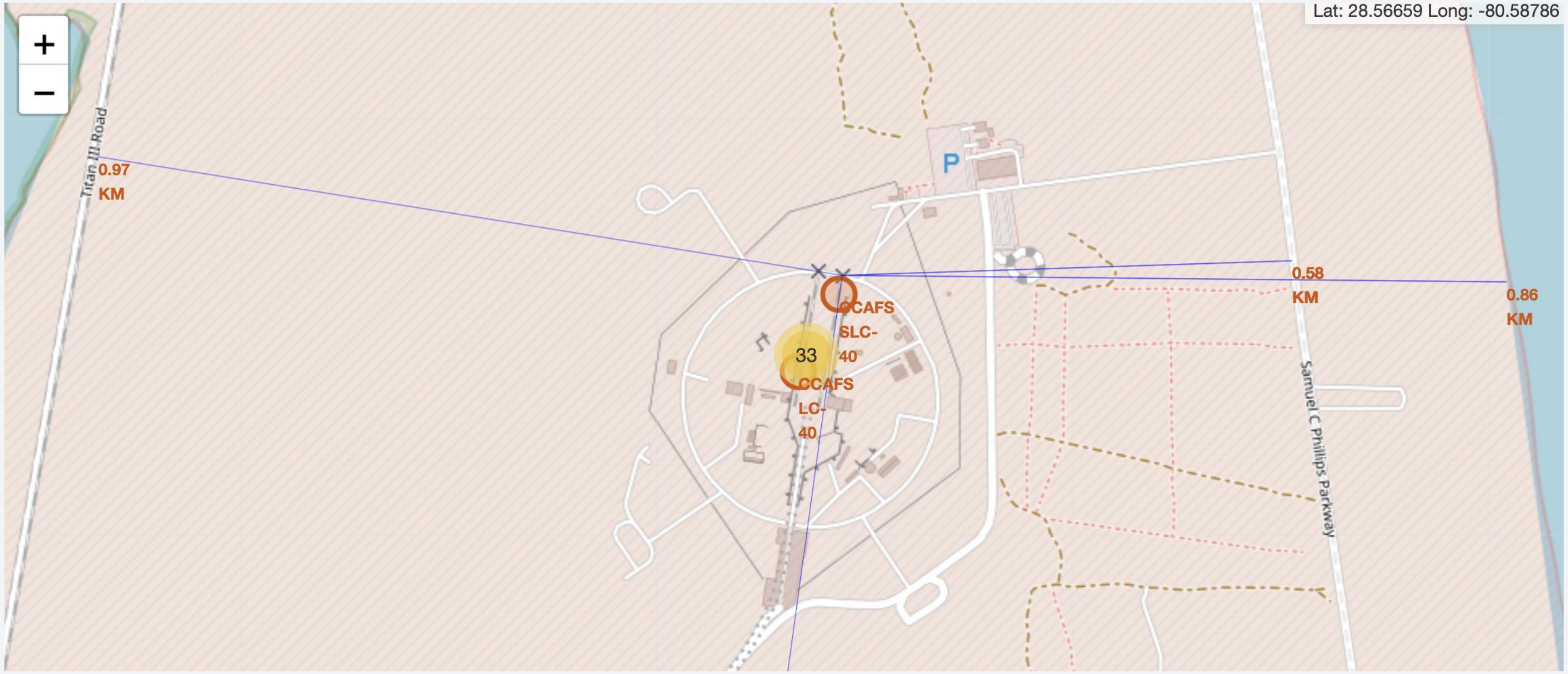
Success/Failed Launches For Each Site



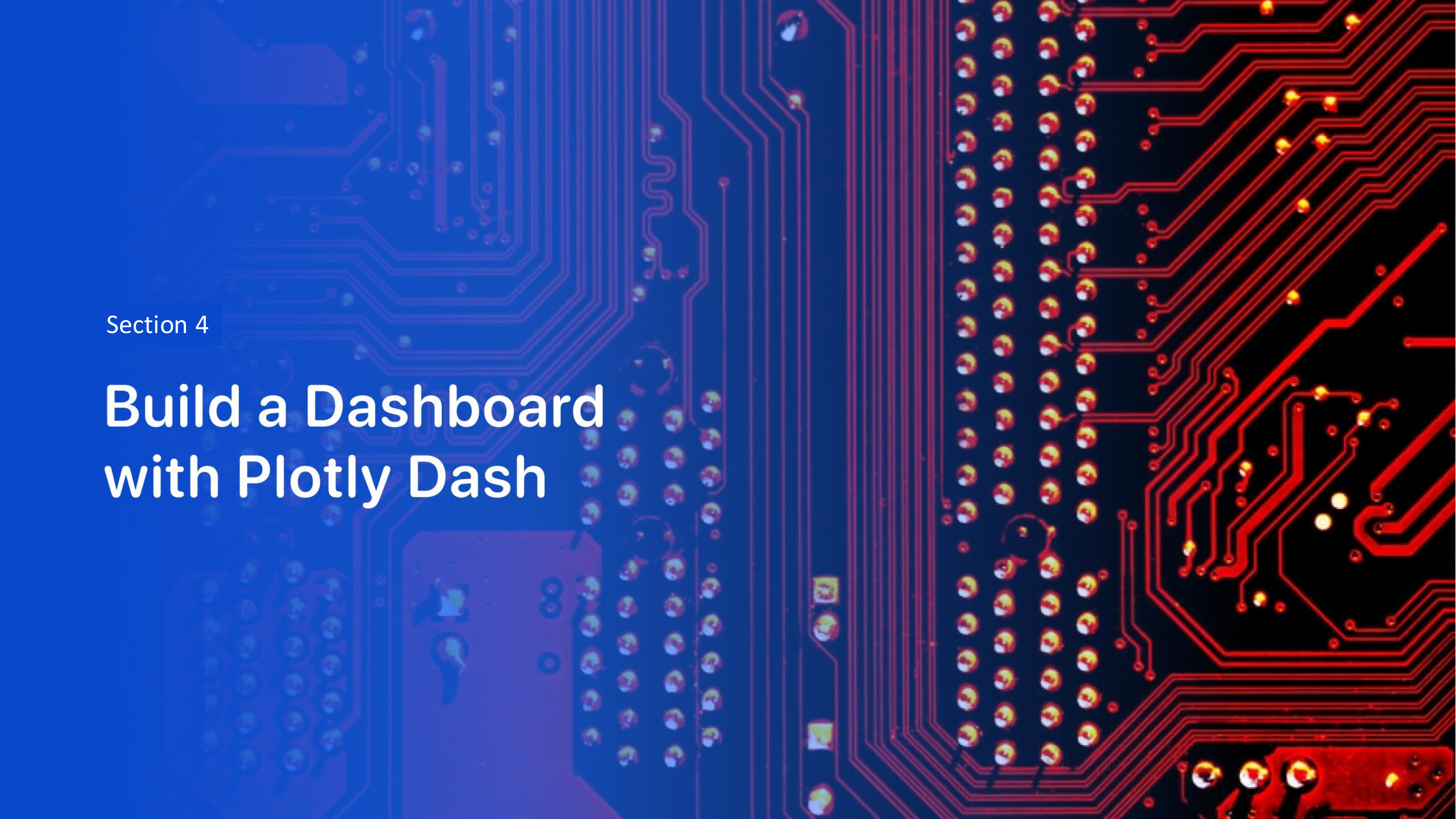


- The first map shows launch sites with total number of launches,
- The second show a green marker if a launch was successful and a red if a launch was failed.

Distance between a launch site to its proximities



Launch sites are near to railway, roads, highways and coastline. They are quite far from cities.

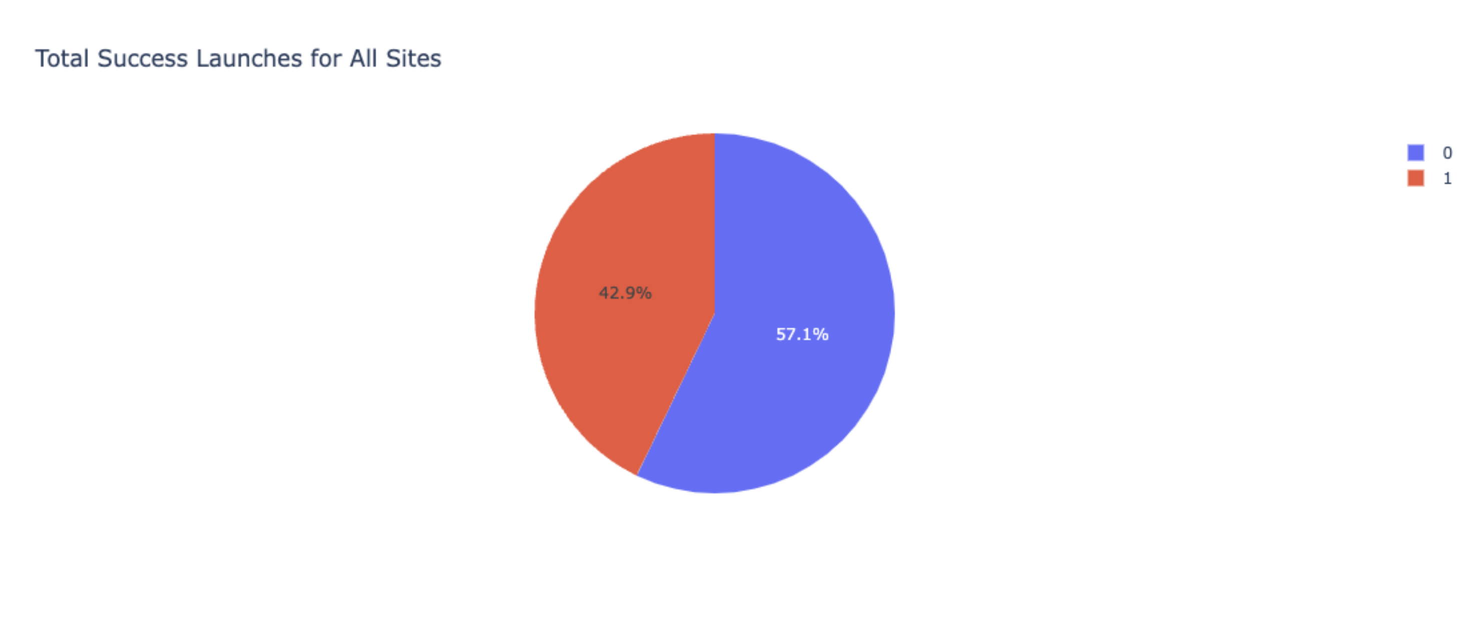


Section 4

Build a Dashboard with Plotly Dash

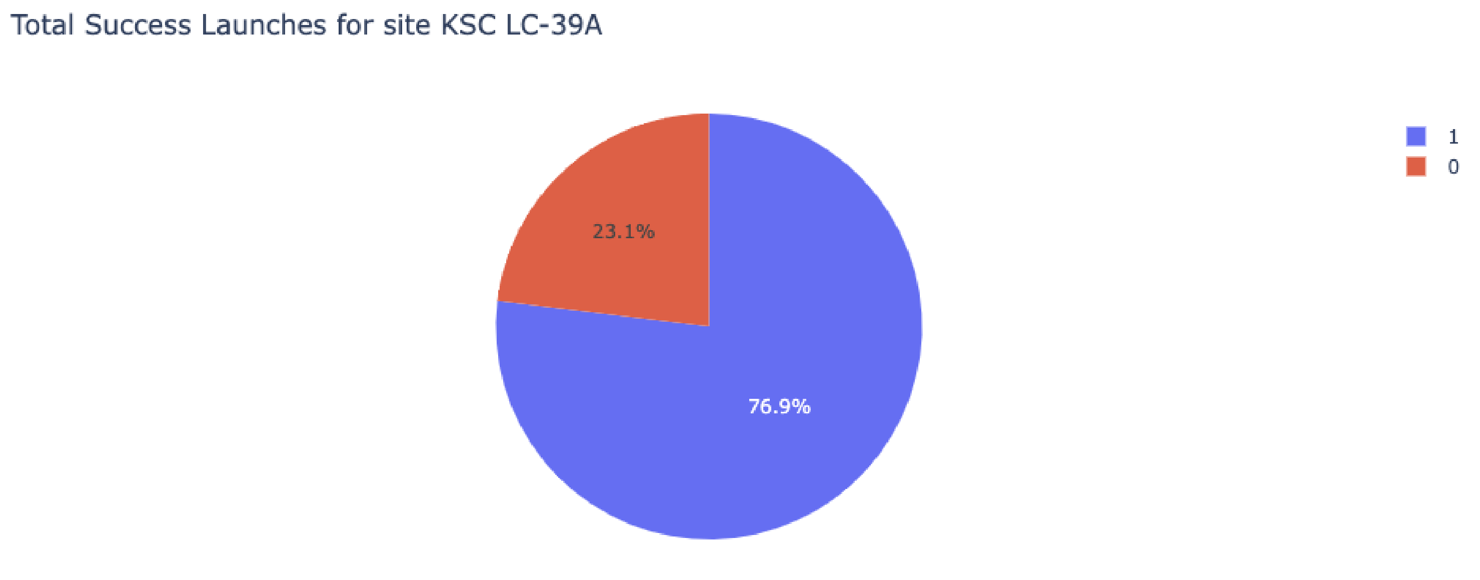
Total Success launches for all sites

- 42.9% of all launches were successful
- 57.1% of all launches were failed



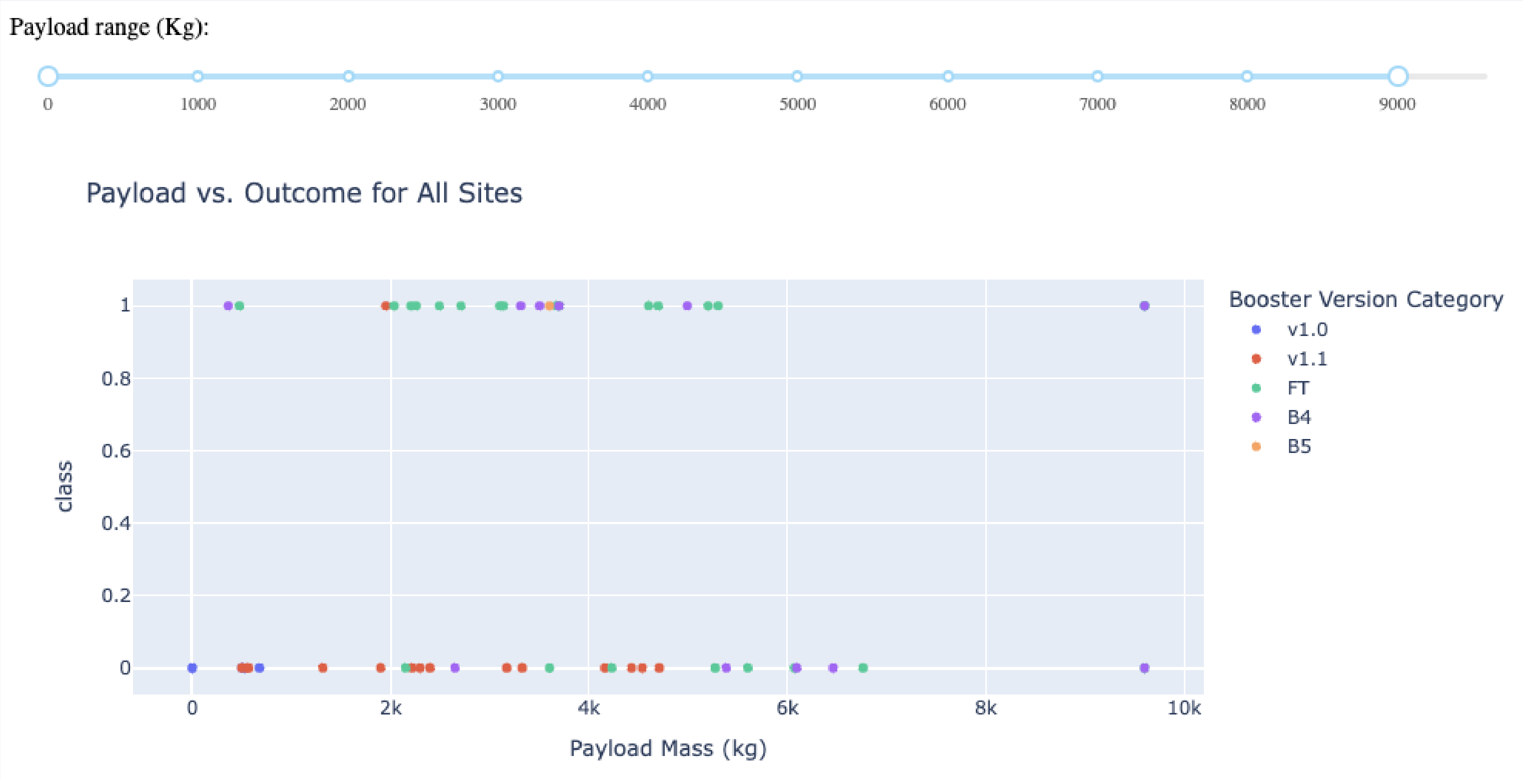
The launch site with highest launch success ratio

- 76.9% of all launches at KSC LC-39A were successful



Payload vs. Outcome for All Sites

- Most of the successful launches are light payload launches (less than 6000kg)
- The success rate will decrease if payload increases to more than 6000kg

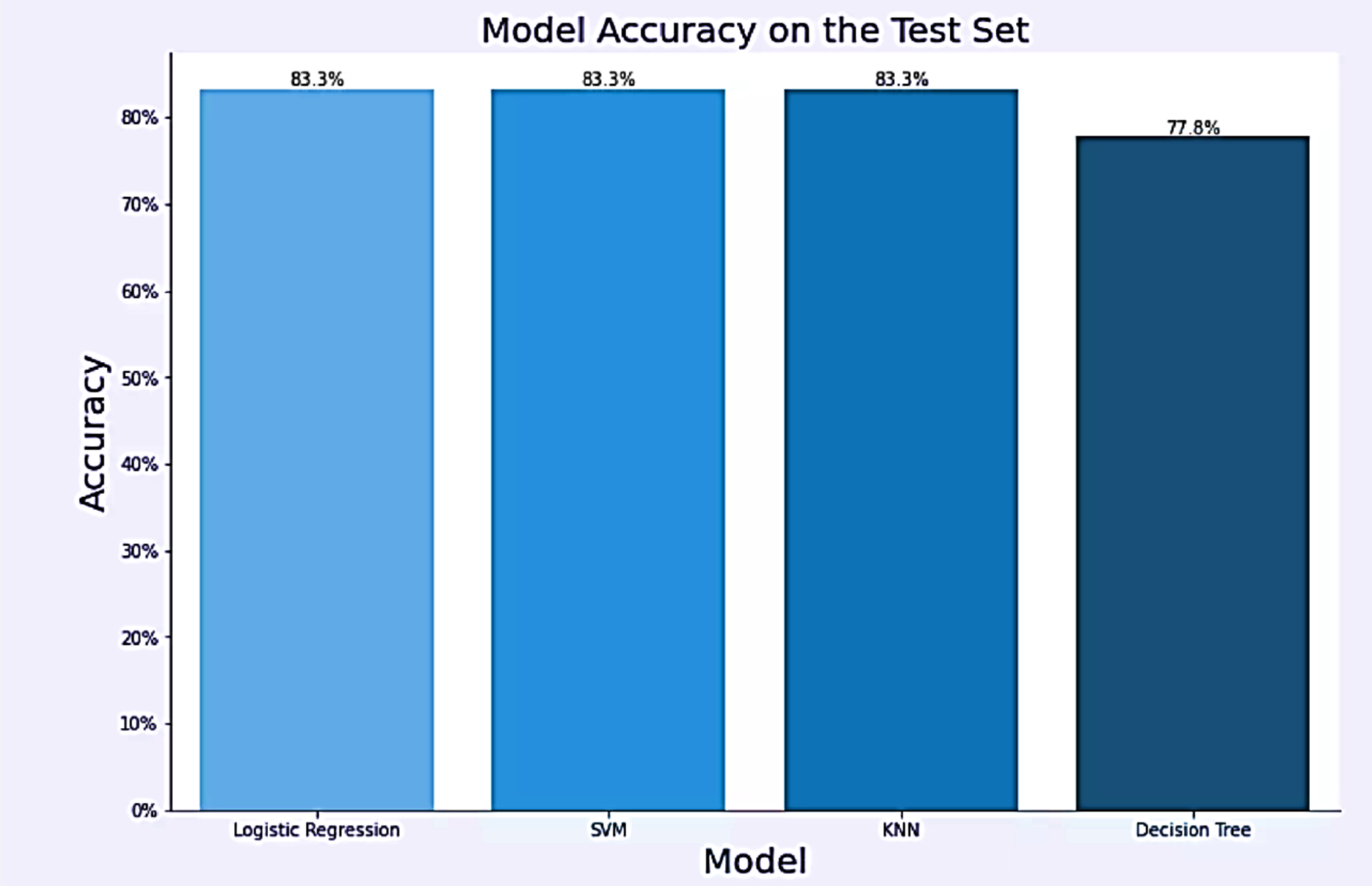


Section 5

Predictive Analysis (Classification)

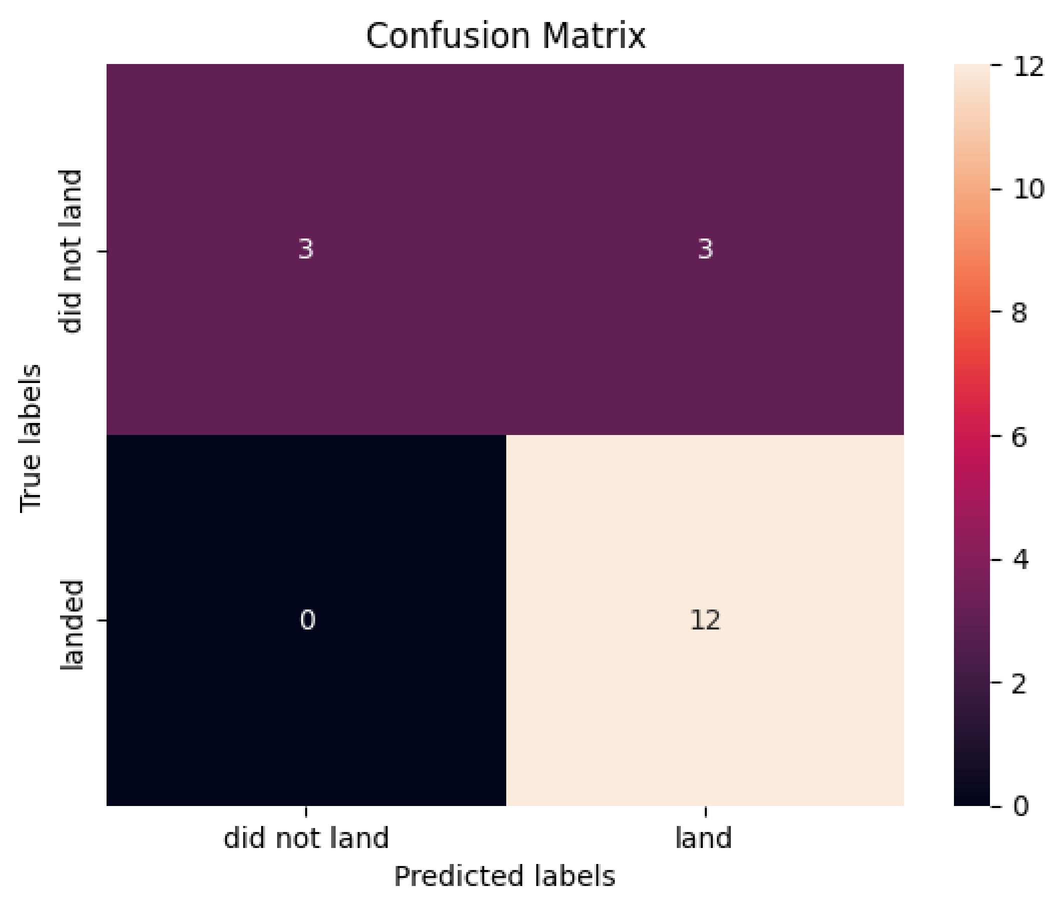
Classification Accuracy

- Logistic Regression, SVM and KNN model have the highest classification accuracy.



Confusion Matrix

- This is the confusion matrix of the best performing model (Logistic Regression, SVM and KNN). We can see that they predicted correctly all 12 true landed samples.



Conclusions

- The SVM, KNN and Logistic Regression model are the best in terms of prediction accuracy for this dataset.
- Light payload mass has better success rate than the heavy payload mass.
- The success rates of SPACEX launches gradually increases over the years.
- KSC LC-39A had the best success rate from all launch sites.
- Orbit GEO, HEO, SSO, ES L1 has the best success rate.

Appendix

- GitHub: Link

Thank you!