IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Diego L. Garcia Tinajero Morales
February 26th, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection API with Webscraping
  - Data Wrangling
  - Exploratory data analysis (EDA) using visualization and SQL
  - Perform exploratory data analysis (EDA) using visualization and SQL
  - Perform interactive visual analytics using Folium and Plotly Dash
  - Perform predictive analysis using classification models

- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context

SpaceX has emerged as a groundbreaking player in the commercial space industry, revolutionizing space travel by making it more accessible and cost-effective. Central to their success is the Falcon 9 rocket, prominently featured on their website with a price tag of 62 million dollars per launch. This stands in stark contrast to traditional providers whose launches can cost upwards of 165 million dollars each. The key to SpaceX's cost efficiency lies in their ability to reuse the first stage of their rockets, a feat that significantly reduces expenses.

Understanding the pivotal role of the first stage in cost reduction, our project aims to leverage public data and machine learning models to predict whether SpaceX will successfully land and thus be able to reuse the first stage. By doing so, we seek to provide insights into the factors influencing the success of first stage landings, thereby contributing to a deeper understanding of SpaceX's operational dynamics and cost-saving strategies.

- Problems you want to find answers

1. **Impact of Variables:** How do factors such as payload mass, launch site, number of flights, and orbits influence the success of the first stage landing? Analyzing these variables will shed light on the complexities involved in achieving successful landings and inform future decision-making processes.
2. **Temporal Trends:** Does the rate of successful landings show an increasing trend over the years? By examining historical data, we aim to discern patterns and trends that may reveal improvements in SpaceX's landing capabilities over time.
3. **Algorithm Selection:** What is the most suitable algorithm for binary classification in this scenario? By evaluating different machine learning algorithms, we seek to identify the model that yields the highest accuracy in predicting first stage landing outcomes, thereby enhancing the reliability of our predictions.

By addressing these questions, our project endeavors to provide valuable insights into the factors driving the success of first stage landings and the broader implications for SpaceX's cost-effective approach to space travel.

Section 1

Methodology

Methodology

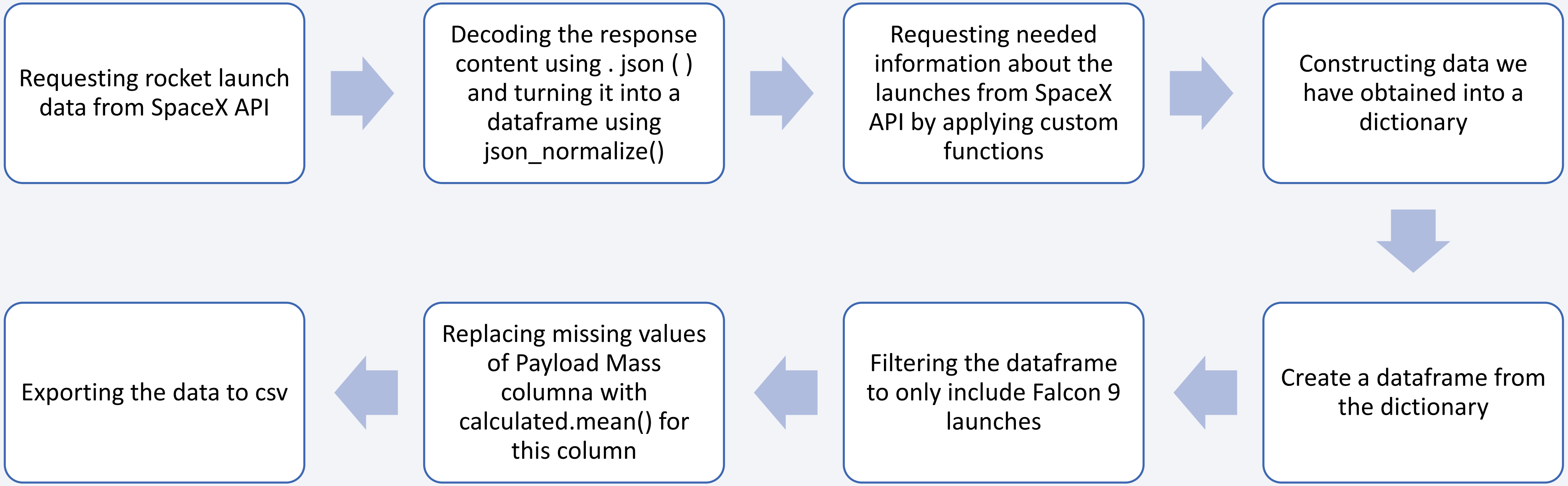
Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

To collect and process the data, the project involves performing GET requests using the requests library in Python to obtain JSON-formatted data from the API. This JSON data, representing a list of launch objects, is then normalized into a flat table format using the ‘json_normalize’ function from the Pandas library. Additionally, the project incorporates web scraping techniques using the BeautifulSoup package to extract Falcon 9 launch records from HTML tables on this Wikipedia page.

Data Collection – SpaceX API



Requesting rocket launch data from SpaceX API

Decoding the response content using . json ( ) and turning it into a dataframe using json_normalize()

Requesting needed information about the launches from SpaceX API by applying custom functions

Constructing data we have obtained into a dictionary

Create a dataframe from the dictionary

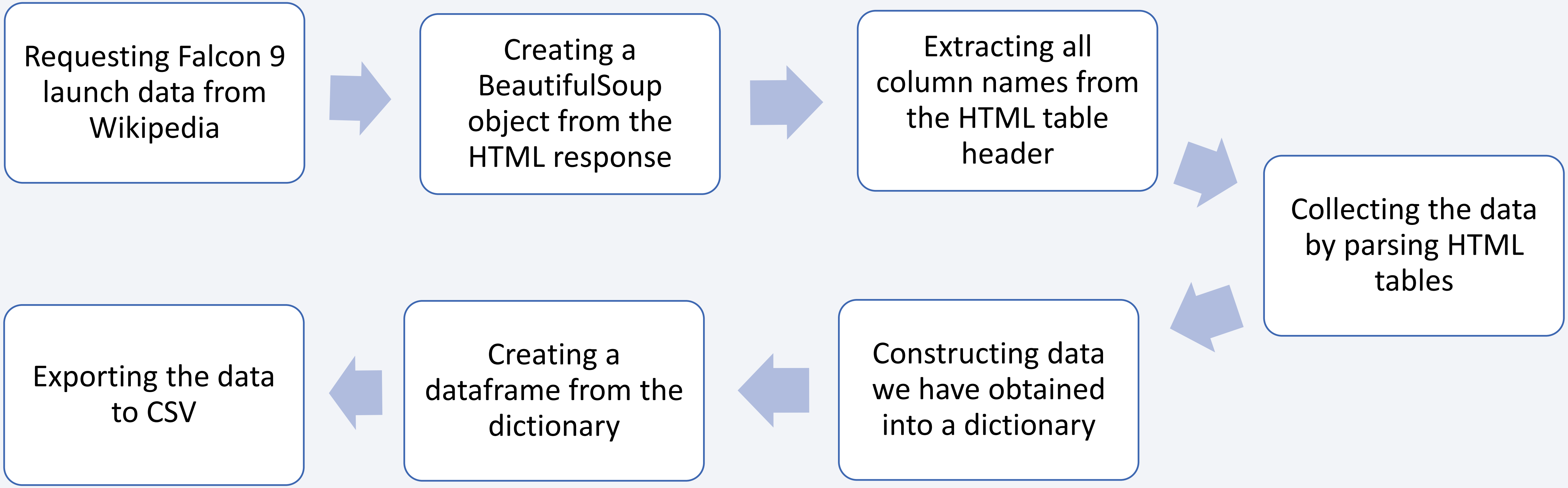
Filtering the dataframe to only include Falcon 9 launches

Replacing missing values of Payload Mass columna with calculated.mean() for this column

Exporting the data to csv

Git Hub URL: Data Collection API

Data Collection - Scraping



Requesting Falcon 9 launch data from Wikipedia

Creating a BeautifulSoup object from the HTML response

Extracting all column names from the HTML table header

Collecting the data by parsing HTML tables

Constructing data we have obtained into a dictionary

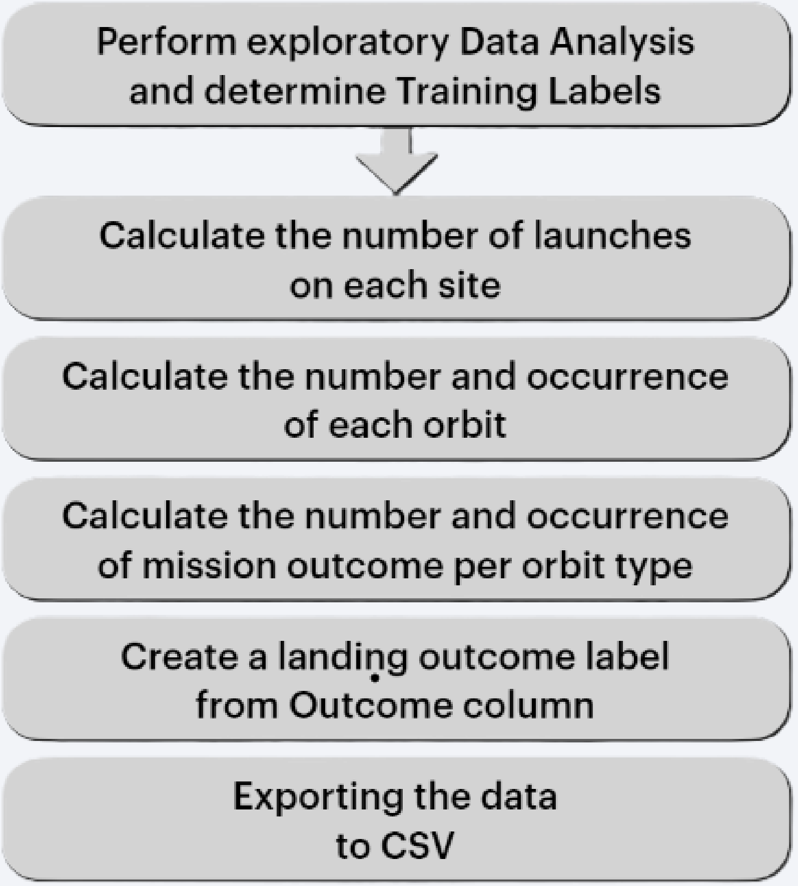
Creating a dataframe from the dictionary

Exporting the data to CSV

Git Hub URL: Data Collection with Web Scraping

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.



Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

Git Hub URL: Data Wrangling

EDA with Data Visualization

In our exploratory data analysis (EDA), we utilized various visualization techniques to gain insights into the data. Here are the types of charts used:

- Scatter Plots:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Flight Number vs. Orbit Type
  - Payload Mass vs. Orbit Type
- Bar Chart:
  - Orbit Type vs. Success Rate
- Line Chart:
  - Launch Success Rate Yearly Trend

Git Hub URL: EDA with Data Visualization

EDA with SQL

Performed SQL queries:

- 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 succesful landing outcome in ground pad was acheived.
- 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.
- 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.
- 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.

Git Hub URL: EDA with SQL

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

Git Hub URL: Interactive Visual Analytics with Folium

Predictive Analysis (Classification)

1. Load and Prepare Data

1. Load data
2. Extract 'Class' column as Y
3. Standardize feature variables as X
4. Split data into train and test sets

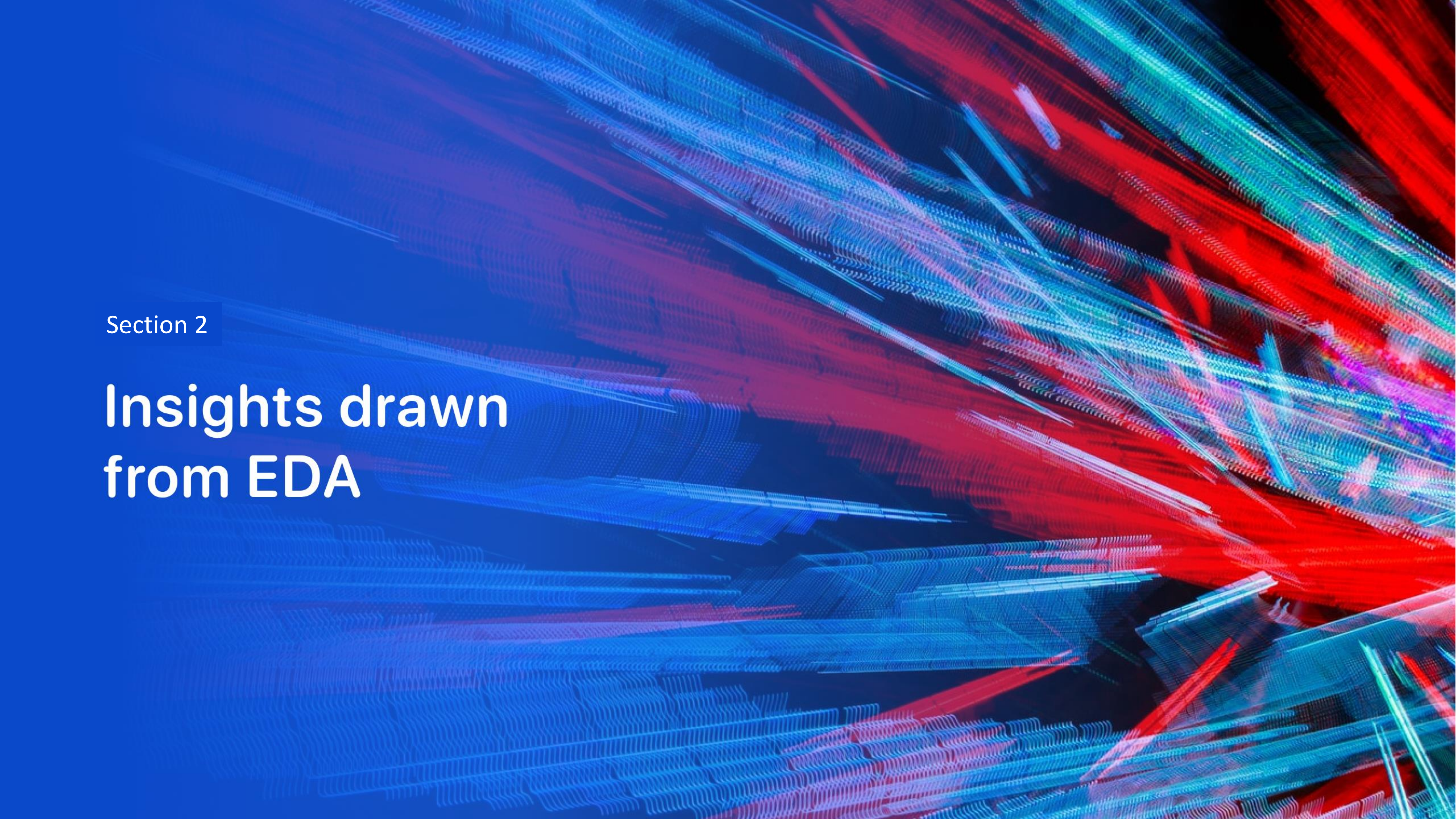
2. Model Selection and Evaluation

1. Iterate over each machine learning algorithm:
    1. Logistic Regression
    2. Support Vector Machine (SVM)
    3. Decision Tree
    4. K Nearest Neighbors (KNN)
2. For each algorithm:
    1. GridSearchCV for best parameters
    2. Train with best parameters
    3. Evaluate accuracy
    4. Plot confusion matrix
    5. Calculate Jaccard score, F1 score
3. Compare model performance metrics
4. Determine best-performing model

Git Hub URL: Machine Learning Prediction

Results

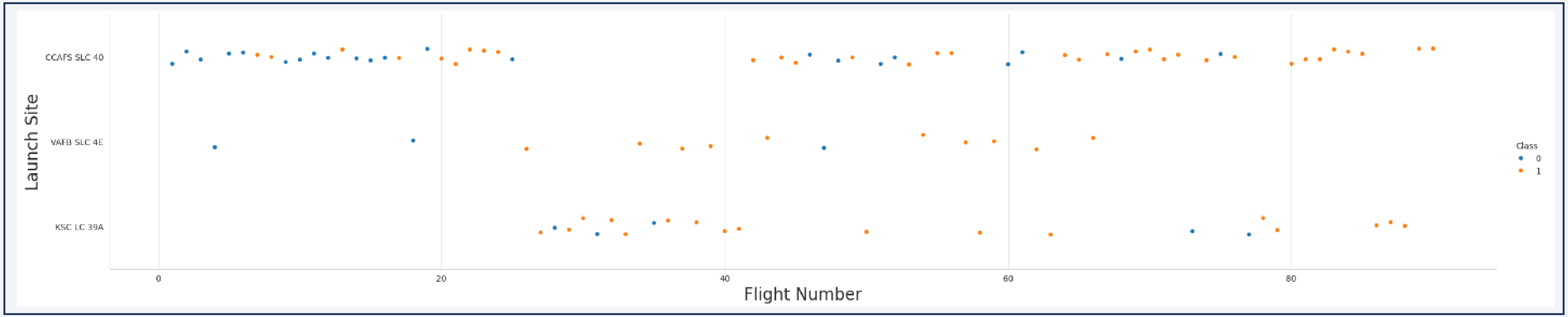
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

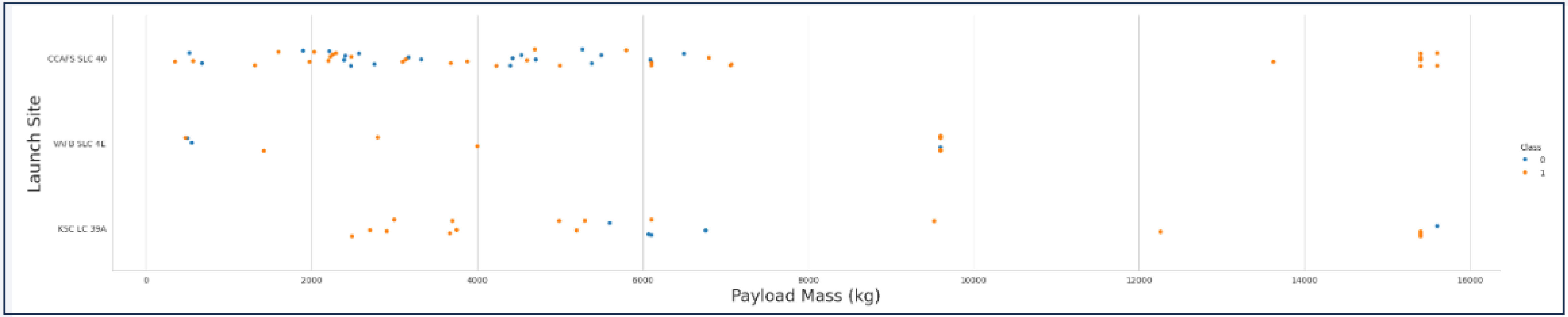
Flight Number vs. Launch Site



Explanation:

- It can be assumed that each new launch has a higher rate of success
- The CCAFS SLC 40 launch site site has about a half of all launches
- VAFB SLC 4E and KSC LC 39 have higher success rate

Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

Success Rate vs. Orbit Type

Explanation:

- Orbits with 100% success rate:

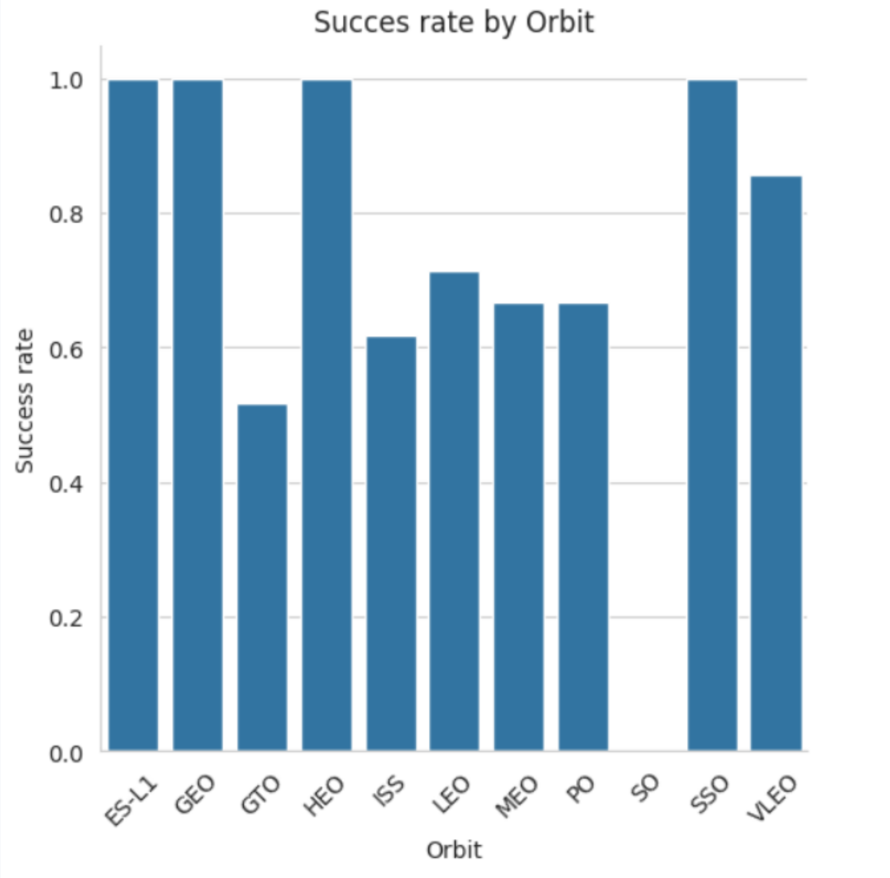
  ES-L1,GEO,HEO,SSO

- Orbits with 0% success rate:

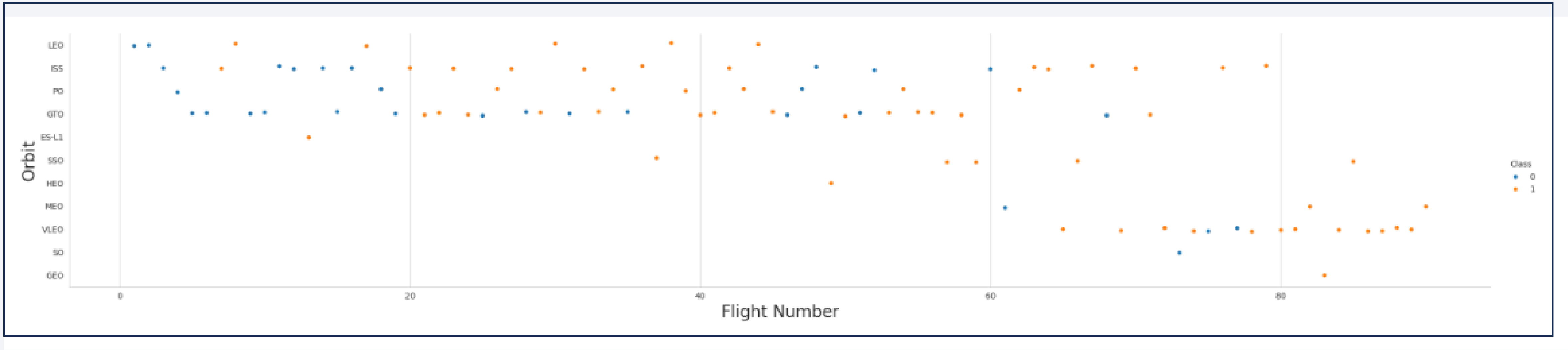
  SO

- Orbits with success rate between 50-5 and 85%:

  GTO,ISS,LEO,MEO, PO



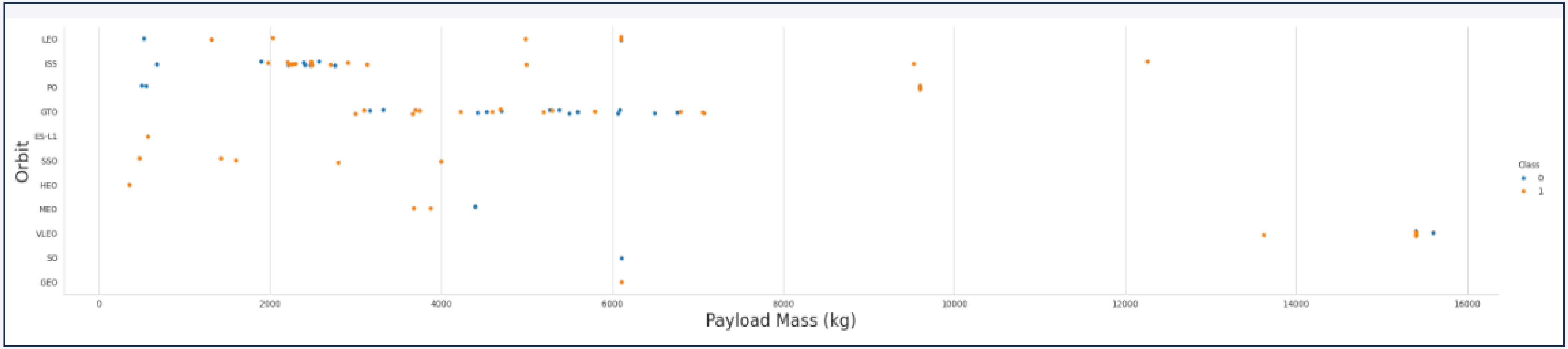
Flight Number vs. Orbit Type



Explanation:

- 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



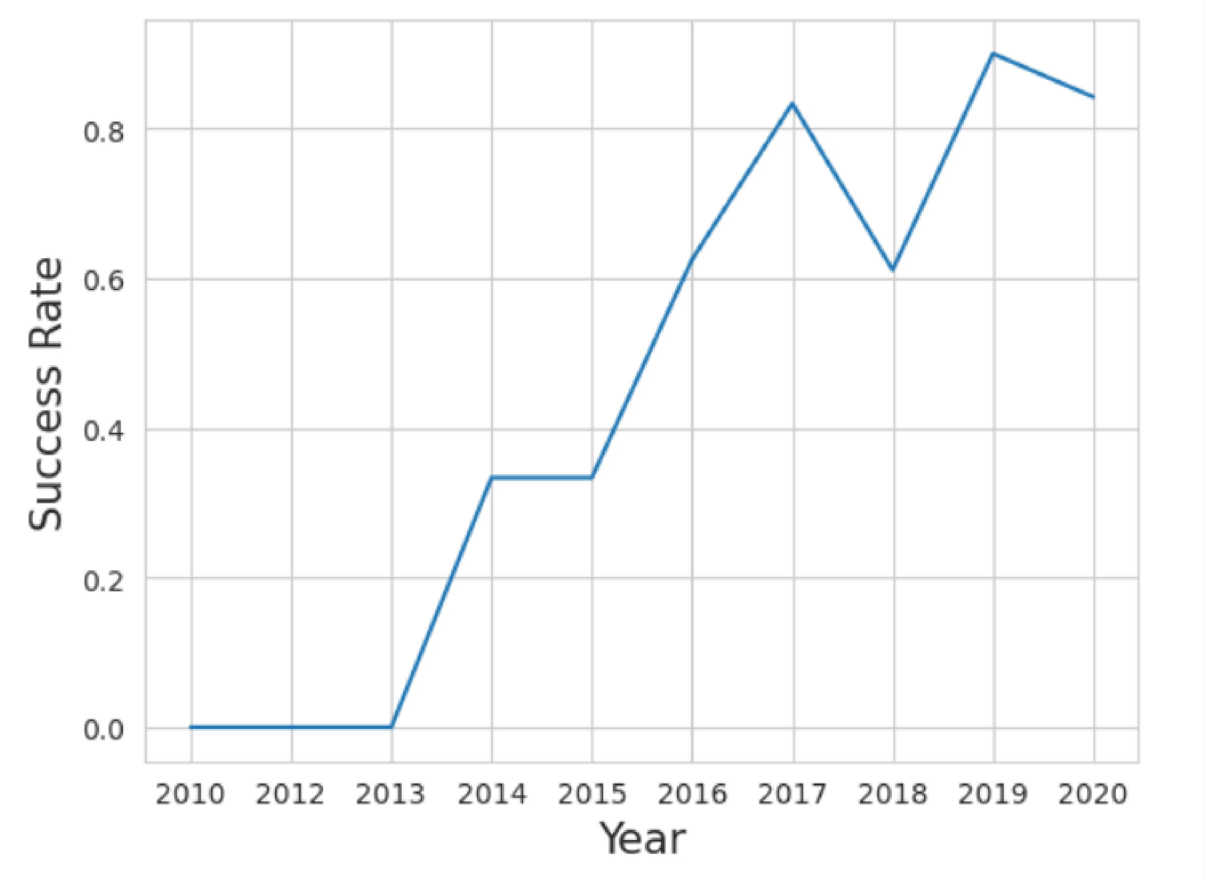
Explanation:

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits

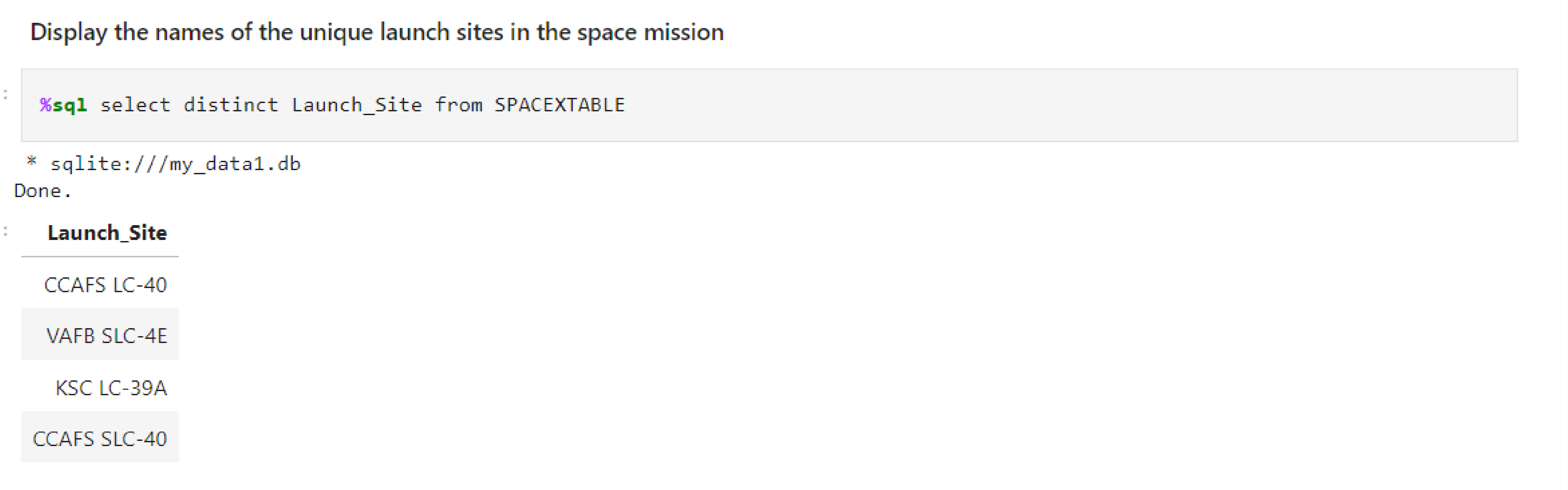
Launch Success Yearly Trend

Explanation:

The success rate since 2013 kept increasing till 2020



All Launch Site Names



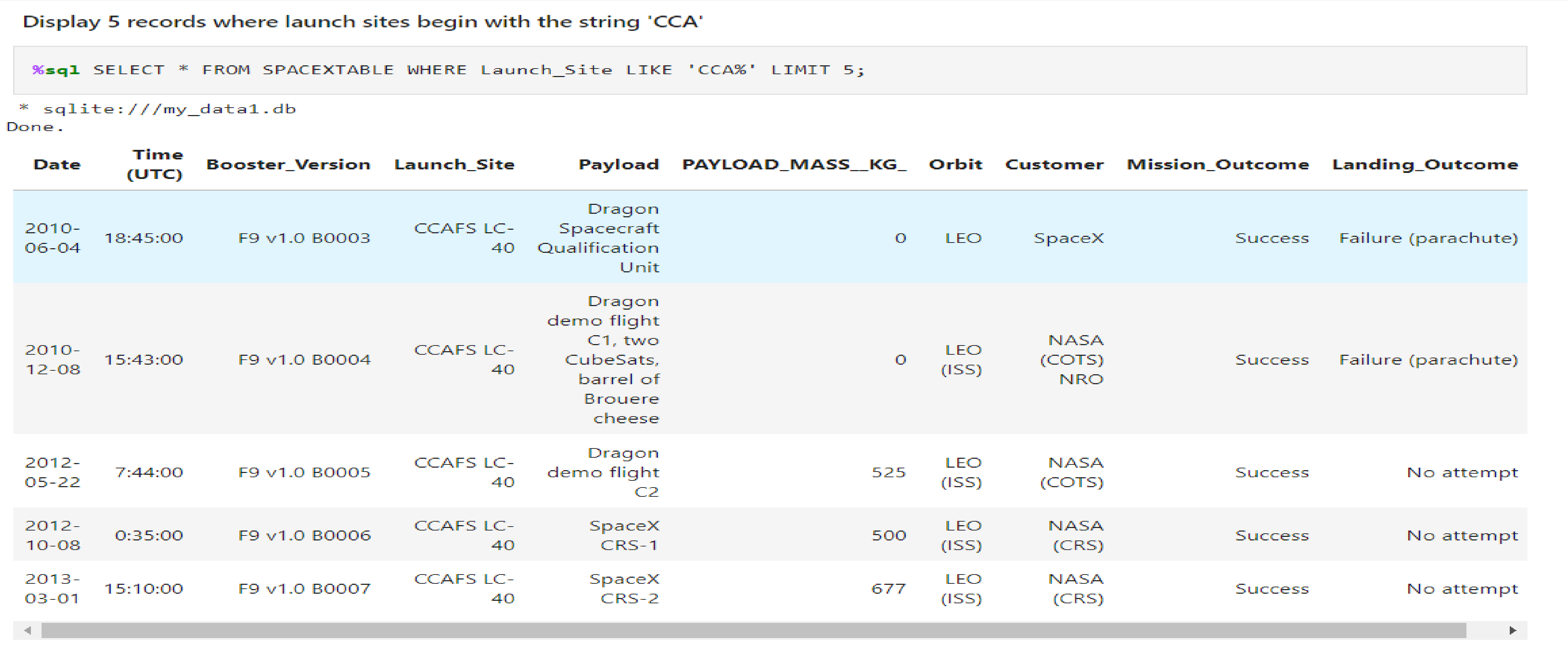
Display the names of the unique launch sites in the space mission

```
%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'



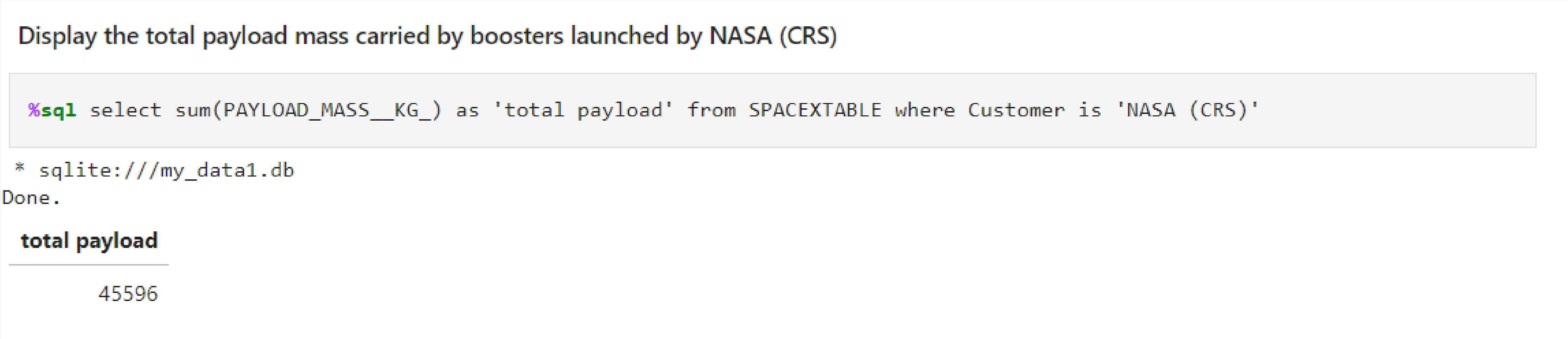
Display 5 records where launch sites begin with the string 'CCA'

```
%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



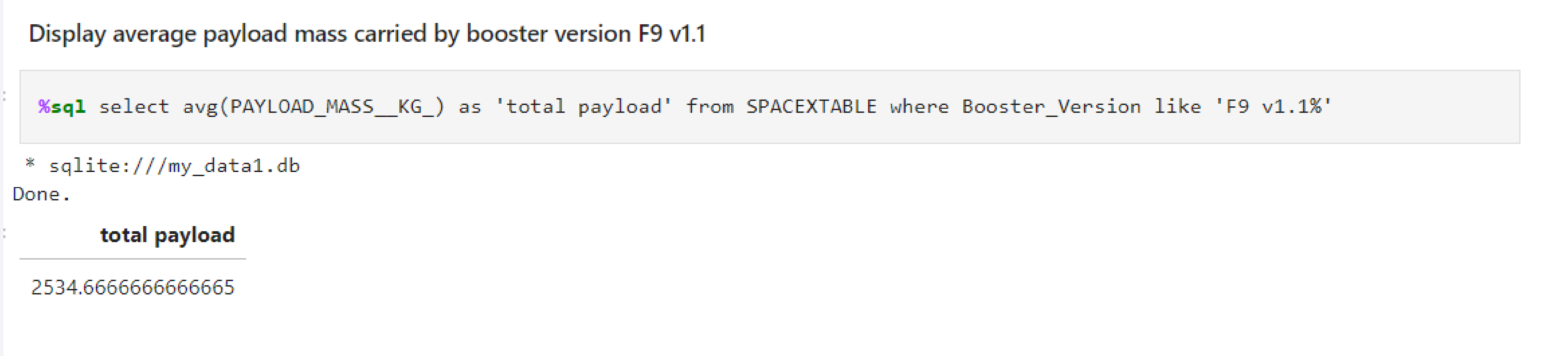
Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) as 'total payload' from SPACEXTABLE where Customer is 'NASA (CRS)'
```

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

total payload
45596

Average Payload Mass by F9 v1.1



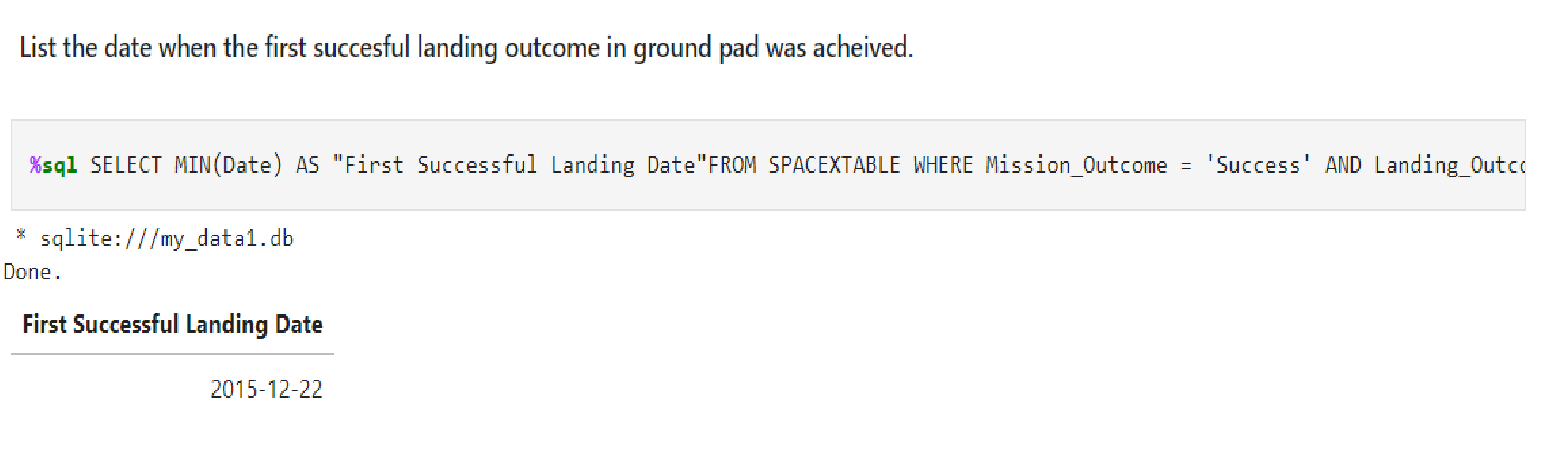
Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as 'total payload' from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

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

total payload
2534.6666666666665

First Successful Ground Landing Date



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

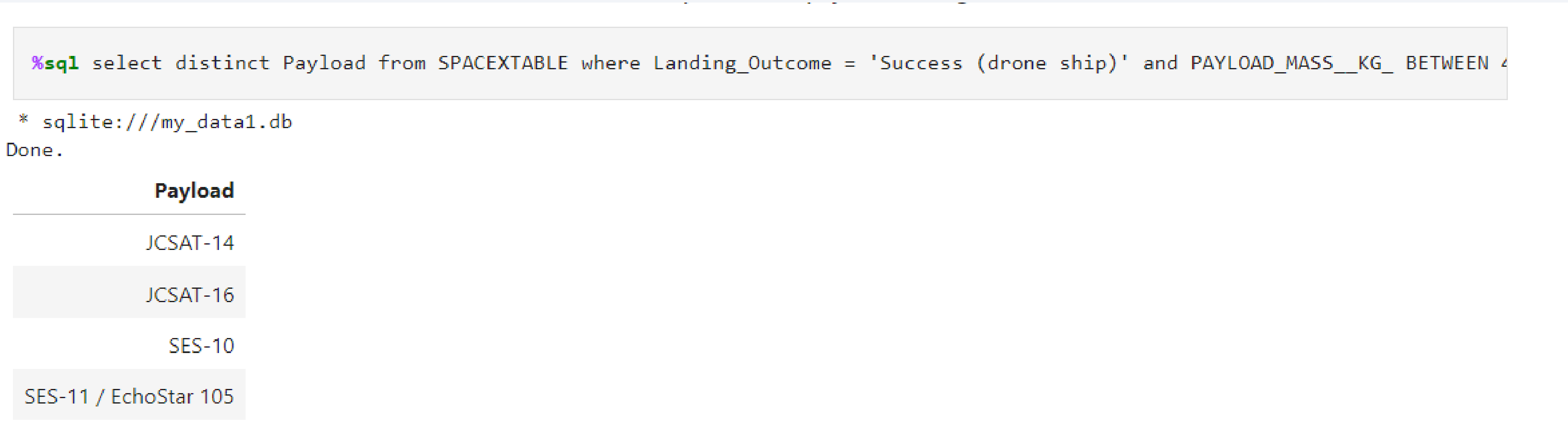
```
%sql SELECT MIN(Date) AS "First Successful Landing Date"FROM SPACEXTABLE WHERE Mission_Outcome = 'Success' AND Landing_Outc
```

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

First Successful Landing 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



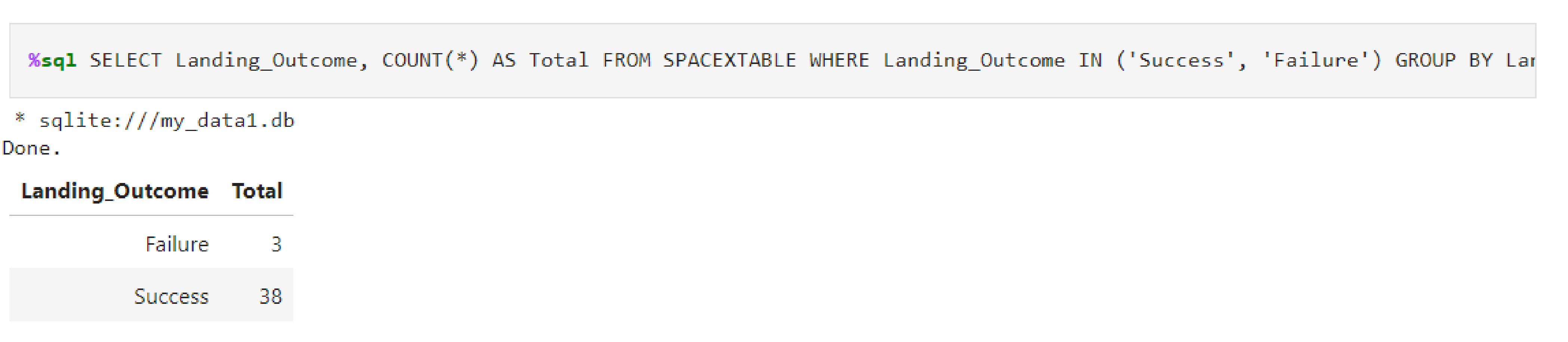
```
%sql select distinct Payload from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



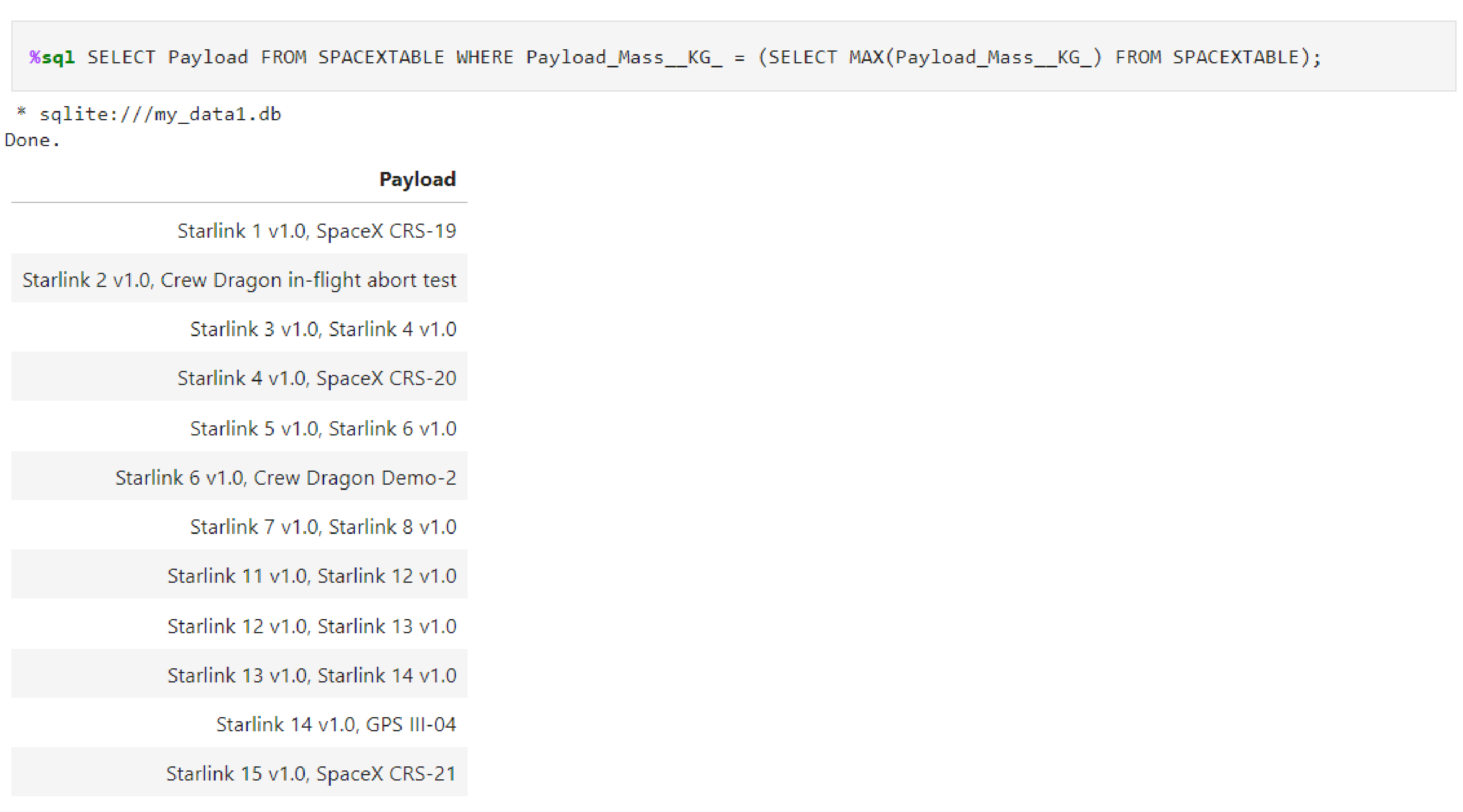
```
%sql SELECT Landing_Outcome, COUNT(*) AS Total FROM SPACEXTABLE WHERE Landing_Outcome IN ('Success', 'Failure') GROUP BY Lar
```

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

Landing_Outcome	Total
Failure	3
Success	38

Boosters Carried Maximum Payload

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



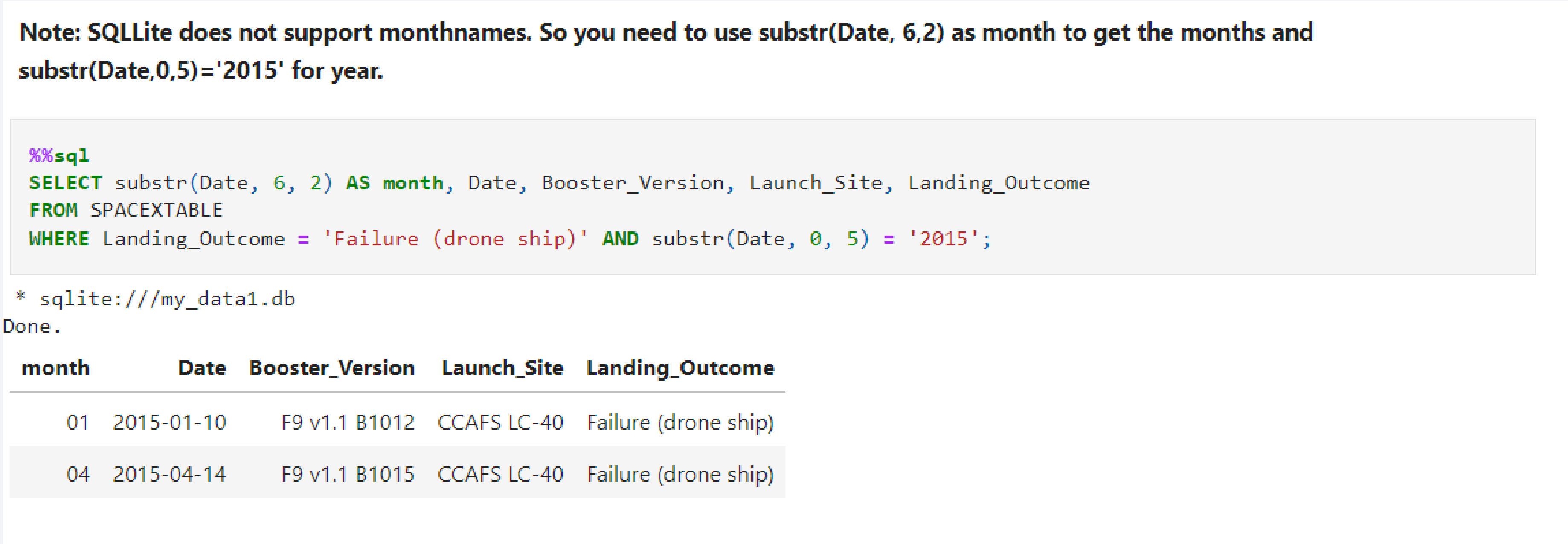
```
%sql SELECT Payload FROM SPACEXTABLE WHERE Payload_Mass__KG_ = (SELECT MAX(Payload_Mass__KG_) FROM SPACEXTABLE);
```

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

Payload
Starlink 1 v1.0, SpaceX CRS-19
Starlink 2 v1.0, Crew Dragon in-flight abort test
Starlink 3 v1.0, Starlink 4 v1.0
Starlink 4 v1.0, SpaceX CRS-20
Starlink 5 v1.0, Starlink 6 v1.0
Starlink 6 v1.0, Crew Dragon Demo-2
Starlink 7 v1.0, Starlink 8 v1.0
Starlink 11 v1.0, Starlink 12 v1.0
Starlink 12 v1.0, Starlink 13 v1.0
Starlink 13 v1.0, Starlink 14 v1.0
Starlink 14 v1.0, GPS III-04
Starlink 15 v1.0, SpaceX CRS-21

2015 Launch Records

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



Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

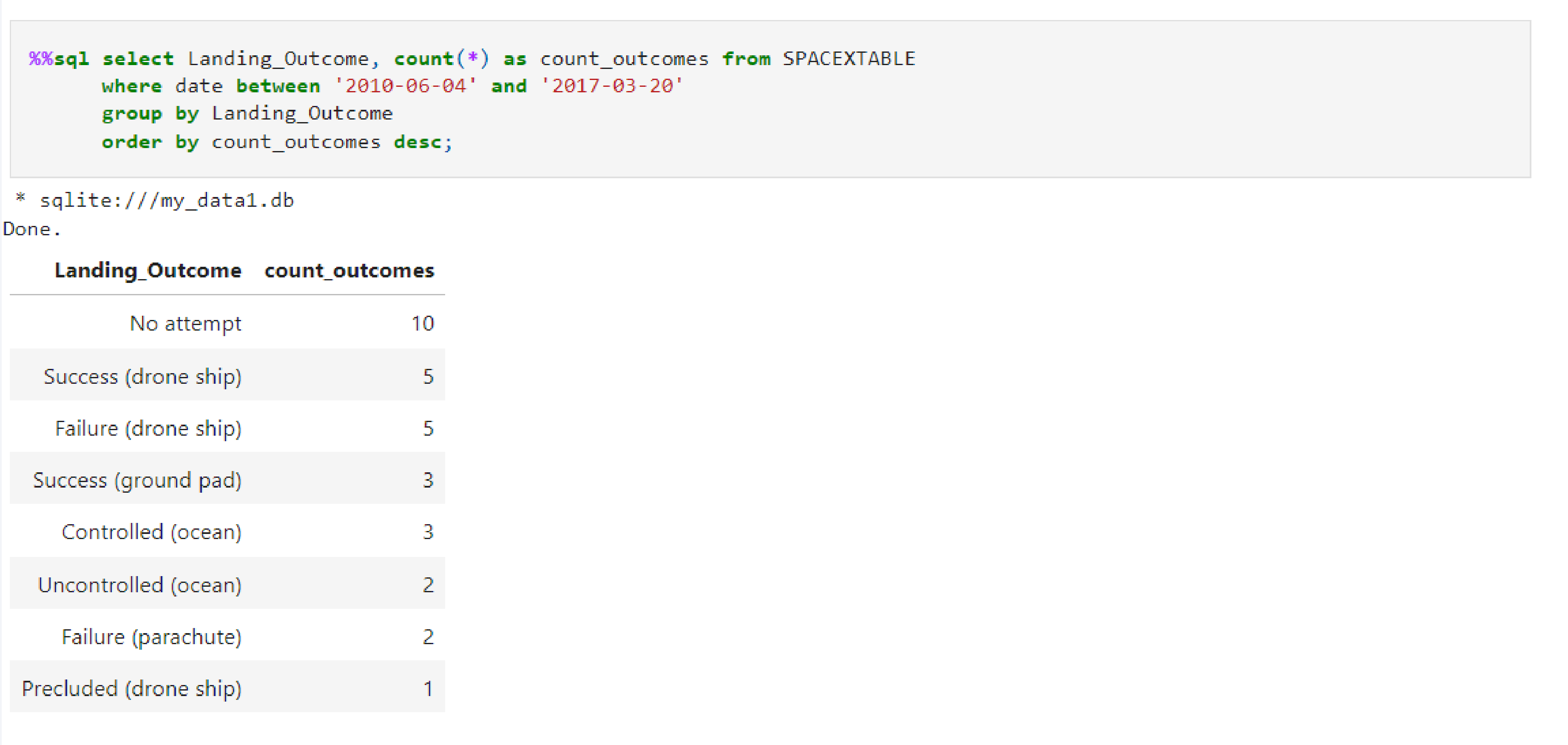
```
%%sql
SELECT substr(Date, 6, 2) AS month, Date, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date, 0, 5) = '2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	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 count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by Landing_Outcome
      order by count_outcomes desc;
```

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

Landing_Outcome	count_outcomes
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

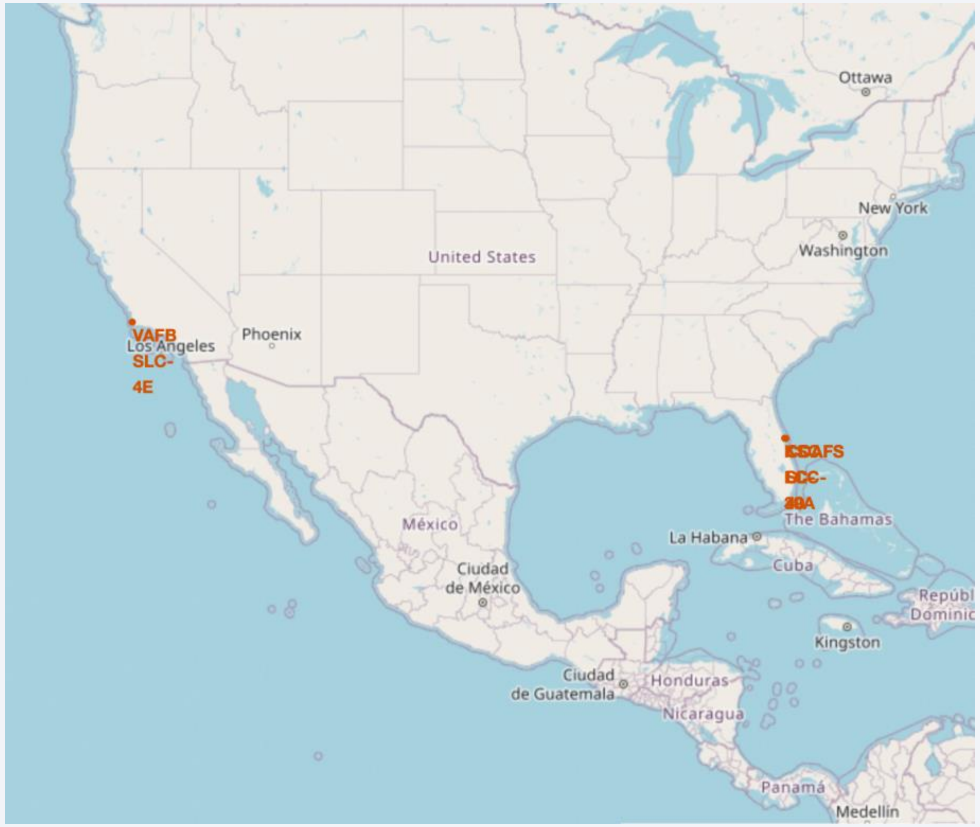
Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on a global map

- Most of Launch sites considered in this project are in proximity to the Equator line. Launch sites are made at the closest point possible to Equator line, because anything on the surface of the Earth at the equator is already moving at the maximum speed (1670 kilometers per hour). For example launching from the equator makes the spacecraft move almost 500 km/hour faster once it is launched compared half way to north pole.

- All launch sites considered in this project are in very close proximity to the coast While starting rockets towards the ocean we minimise the risk of having any debris dropping or exploding near people.



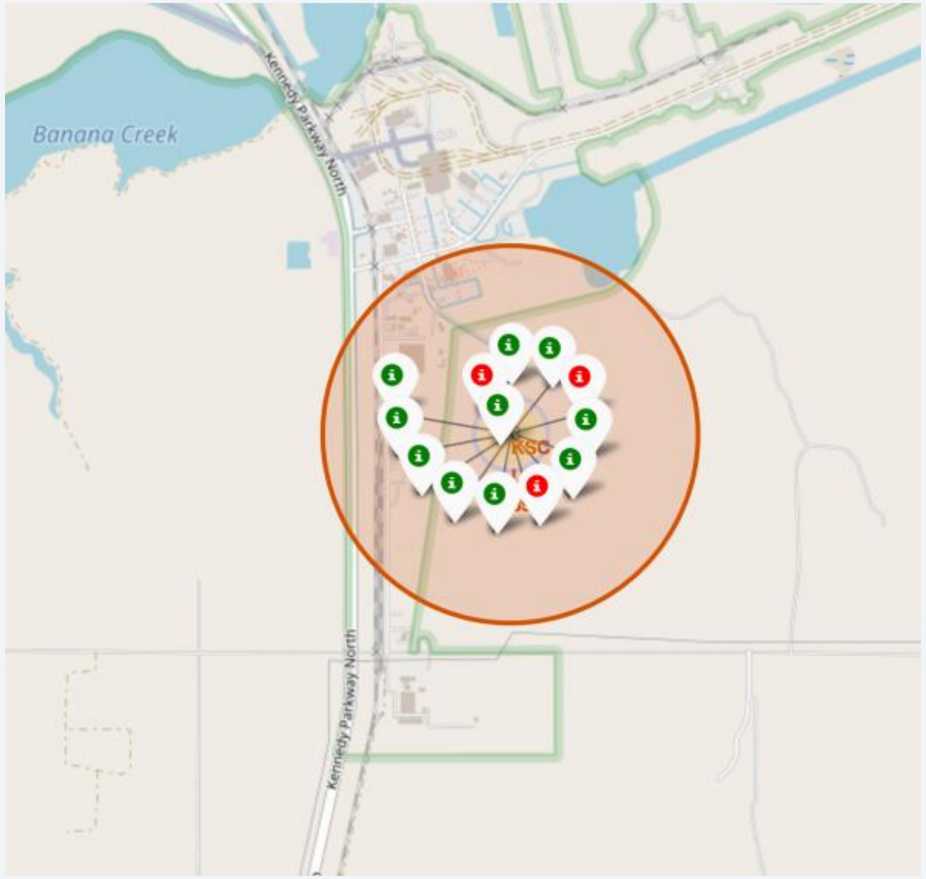
Colour-labeled launch records on the ma

From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

Green Marker = Successful Launch

Red Marker = Failed Launch

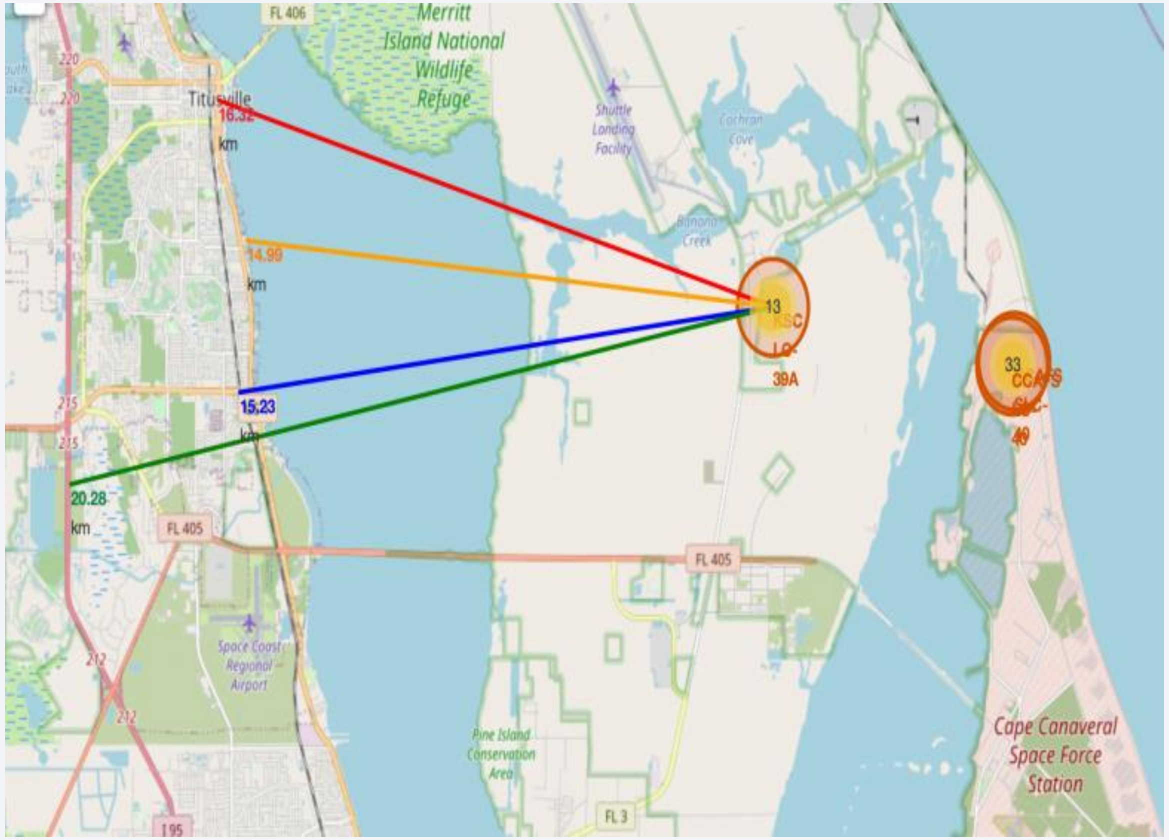
- Launch Site KSC LC-39A has a very high Success Rate.



Distance from the launch site KSC LC-39A to its proximities

After visually analyzing the launch site KSC LC-39A, it's evident that it has the following characteristics:

- •It is situated relatively close to a railway, approximately 15.23 km away.
- •It is also relatively close to a highway, approximately 20.28 km away.
- •Additionally, it is near the coastline, approximately 14.99 km away.
- •Furthermore, the launch site KSC LC-39A is relatively close to its nearest city, Titusville, approximately 16.32 km away.
- •Considering the high speed of a failed rocket, it can cover distances of 15-20 km in mere seconds. This poses a potential risk to populated areas.

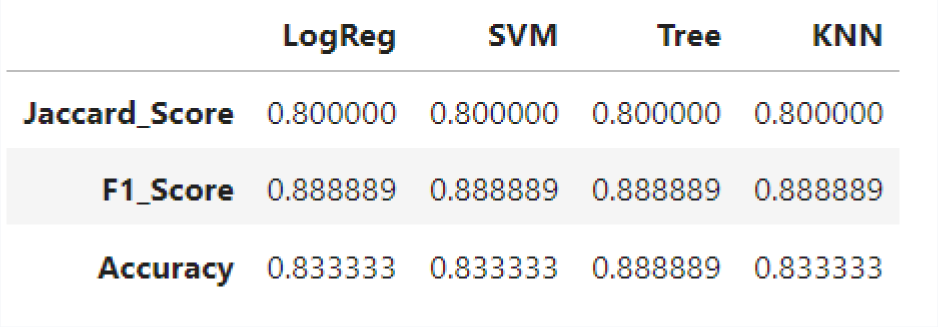


Section 5

Predictive Analysis (Classification)

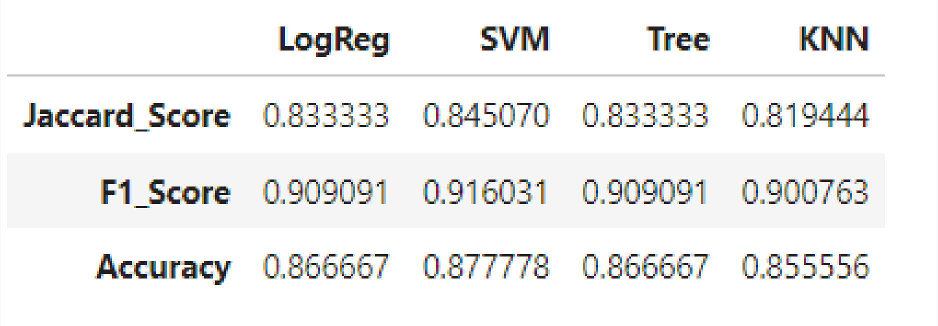
Classification Accuracy

Scores and Accuracy of the Test Set



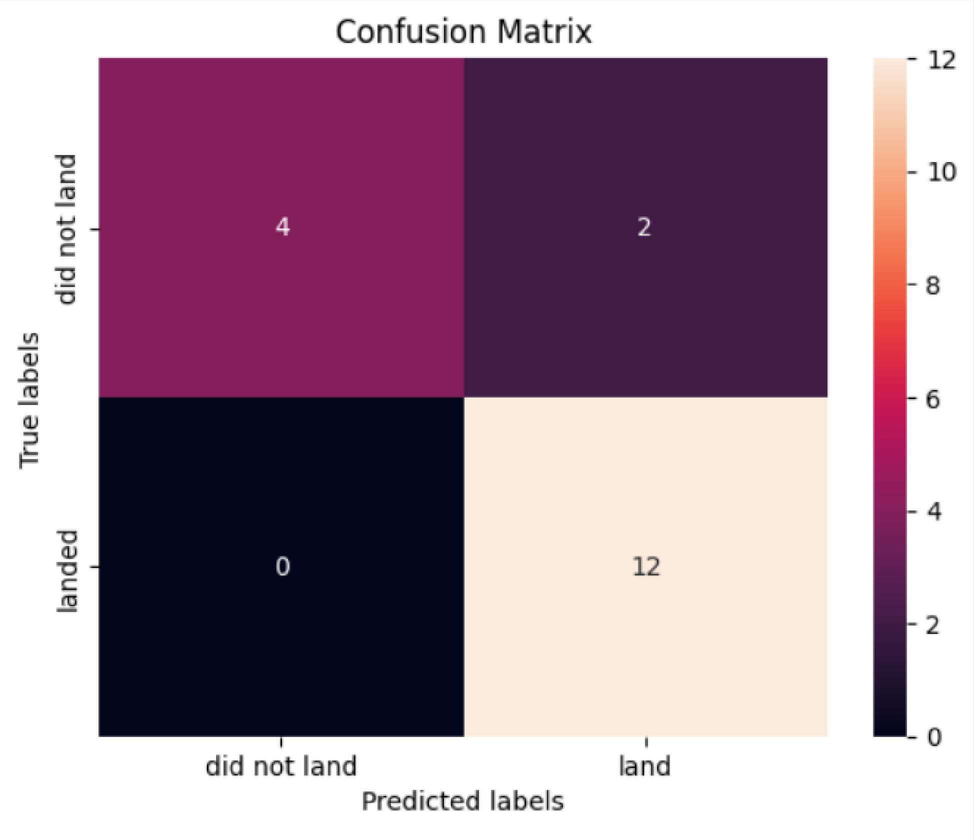
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

Scores and Accuracy of the Entire Data Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.833333	0.819444
F1_Score	0.909091	0.916031	0.909091	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

Confusion Matrix



Conclusions

- In conclusion, upon analyzing the results of the different machine learning models, we observe that SVM and Decision Tree show comparable performance in terms of Jaccard Score, F1 Score, and accuracy. However, SVM slightly outperforms the rest in terms of Jaccard and F1 Score, while Decision Tree has the highest accuracy. Although the difference in metrics among the models is minimal, these figures suggest that SVM might be the best choice for this specific dataset. However, it's important to consider other factors such as model interpretability and scalability when making final decisions about the model to use.

Thank you!