IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mayur V

03-11-2023

Outline

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

Executive Summary

- Summary of methodologies

  - Data Collection.

  - Data Wrangling .

  - Exploratory Data Analysis using SQL.

  - Exploratory Data Analysis using Data Visualization.

  - Interactive Launch Site Location Analysis using Folium.

  - Interactive Dashboard with Plotly Dash.

  - Machine Learning Prediction (Supervised Learning – Classification).

- Summary of all results

  - Performed Exploratory Data Analysis and extracted insights.

  - Built interactive dashboards to view data.

  - Built Machine Learning models with very good predictive capabilities.

Introduction

- Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars while other providers mention costs 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

- Does the rate of successful landings increase over the years?

- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

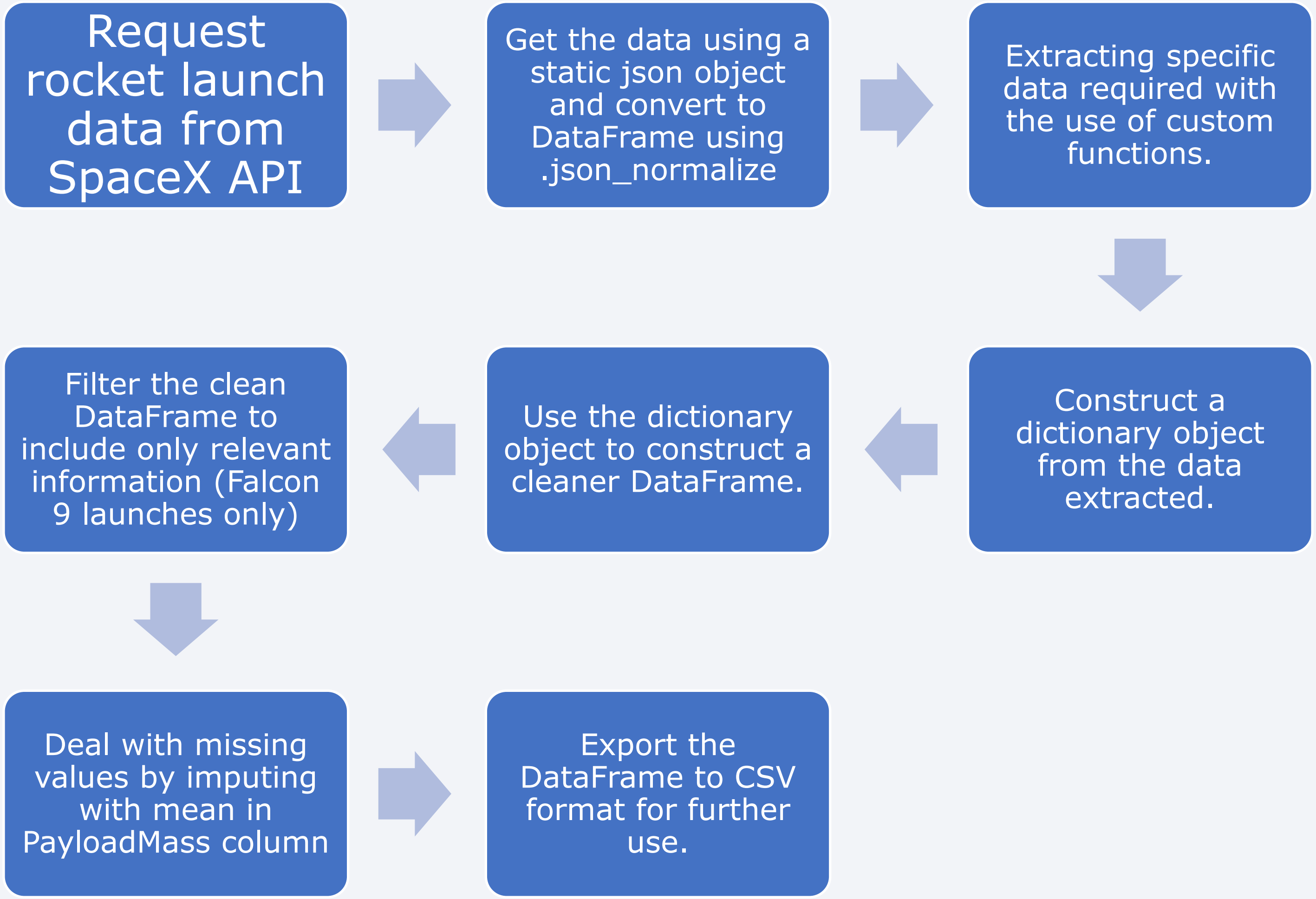
- Data collection methodology:
  - Data was collected using two approaches – SpaceX REST API and Webscraping (using BeautifulSoup).
- Perform data wrangling
  - Dealt with missing values by imputing appropriate statistical means, defined class labels based on landing outcome. Performed OneHot Encoding on categorical 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
  - First, readied the data for ML models by applying StandardScalar.
  - Split the data into test and train parts in accordance with holdout method.
  - Built various models like Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier.
  - Used GridSearchCV to aid in and find the best hyperparameter values.

Data Collection

- Describe how data sets were collected. The datasets were collected in two ways – SpaceX REST API and Web scraping the Wikipedia page of SpaceX launches.
- Both methods were used in order to ensure **completeness** of the dataset.
- The first method involved the use of the SpaceX API to **request** data on all SpaceX launches in a raw form. This raw data was **cleaned** and **parsed** to extract relevant finer details and build a clean DataFrame/dataset containing relevant launch data (only SpaceX Falcon 9 launched).
- The second method involved the use of the requests module, to **webscrape** data from the Wikipedia page, **isolate** the required table with launch data, **parsing** the raw HTML table to build a clean DataFrame/dataset.

Data Collection – SpaceX API

- Request were sent to the SpaceX API for launch data.
- In order to get consistent JSON results, a static response object was used and the obtained data was converted to a DataFrame using the .json_normalize() method.
- This was filtered and key information was extracted to dictionary items, which was used to build a cleaner DataFrame.
- Finally, the DataFrame was filtered to include only Falcon 9 Launch data. Missing data was imputed with the appropriate statistical mean. This was then exported to be used for the project.



Request rocket launch data from SpaceX API

Get the data using a static json object and convert to DataFrame using .json_normalize

Extracting specific data required with the use of custom functions.

Construct a dictionary object from the data extracted.

Use the dictionary object to construct a cleaner DataFrame.

Filter the clean DataFrame to include only relevant information (Falcon 9 launches only)

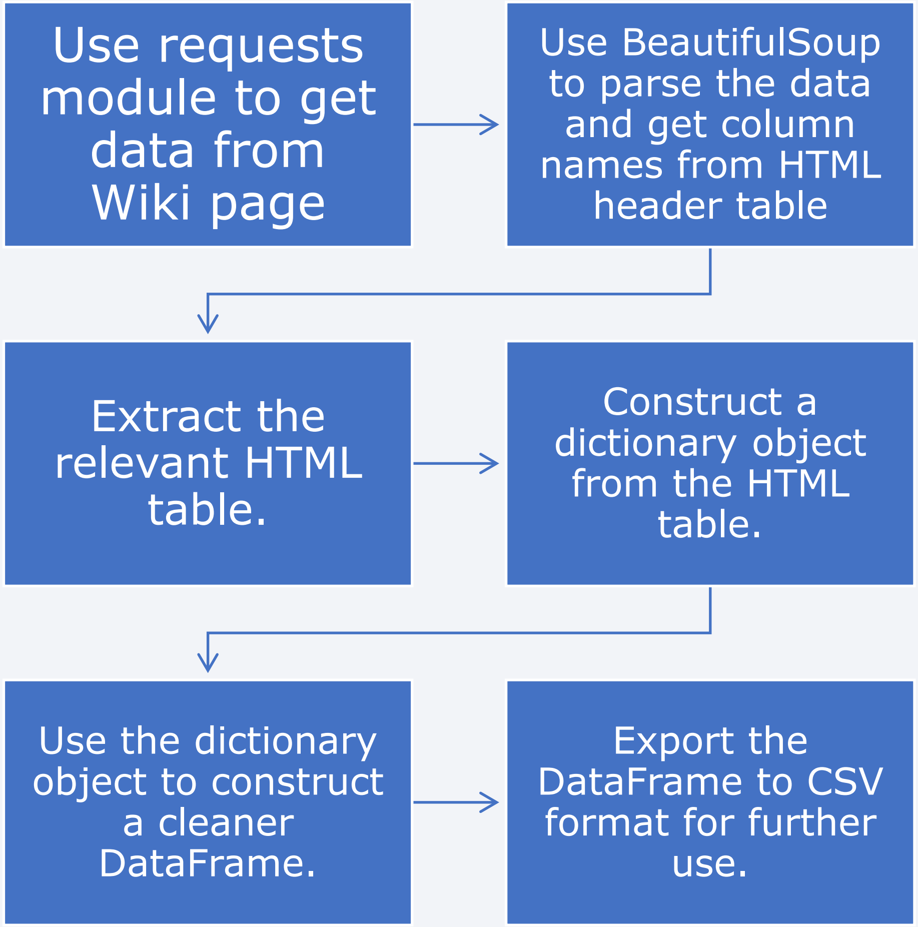
Deal with missing values by imputing with mean in PayloadMass column

Export the DataFrame to CSV format for further use.

URL: SpaceX Data Collection using API

Data Collection - Scraping

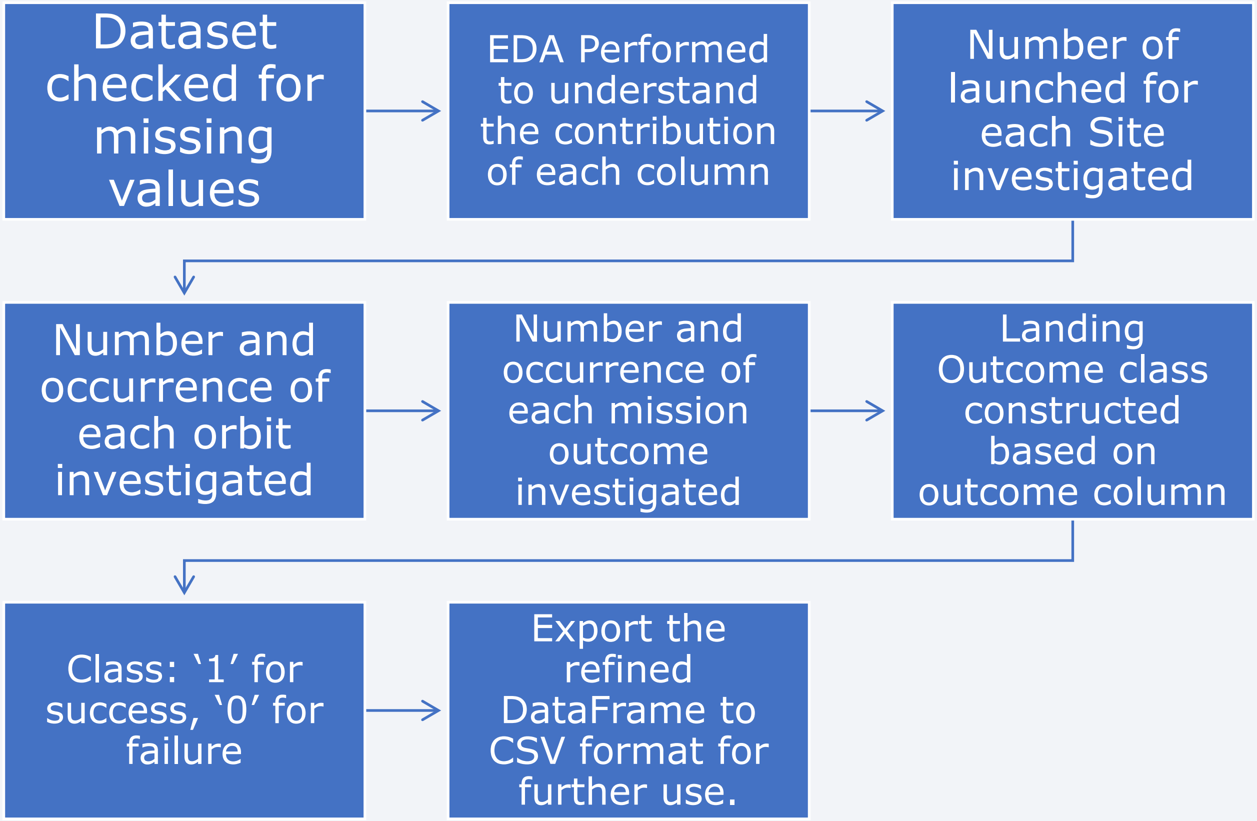
- Use the request module to get data from the Wikipedia page.
- Use BeautifulSoup to parse the obtained information and column names from HTML table header.
- After extracting the HTML tables, parse them to build a dictionary. This dictionary is then used to build the dataframe.
- Finally, export the dataset so that it can be used for the project.


Use requests module to get data from Wiki page
Use BeautifulSoup to parse the data and get column names from HTML header table
Extract the relevant HTML table.
Construct a dictionary object from the HTML table.
Use the dictionary object to construct a cleaner DataFrame.
Export the DataFrame to CSV format for further use.

URL: SpaceX Data Collection using Webscraping

Data Wrangling

- First, the dataset was checked for **missing** values in each column.
- **Exploratory Data Analysis (EDA)** done as stated below
- The number of launches for each site, number & occurrence of each orbit and the number & occurrence of mission outcomes were **investigated**.
- A landing outcome label (**feature construction**) was created based on the outcome column values.
- The refined DataFrame was exported to a CSV for further use.



Dataset checked for missing values

EDA Performed to understand the contribution of each column

Number of launched for each Site investigated

Number and occurrence of each orbit investigated

Number and occurrence of each mission outcome investigated

Landing Outcome class constructed based on outcome column

Class: ‘1’ for success, ‘0’ for failure

Export the refined DataFrame to CSV format for further use.

URL: SpaceX Data Wrangling

EDA with Data Visualization

- Many insightful charts/plots were created – FlightNumber vs PayloadMass, FlightNumber vs LaunchSite, PayloadMass vs LaunchSite, SuccessRate vs OrbitType, FlightNumber vs OrbitType, PayloadMass vs OrbitType, Launch Success Yearly trend. Many interesting insights were derived from the plots mentioned.
- Scatterplots showed relationships between a given variable and the target variable. A strong relationship meant that the given variable could be used in the ML model.
- Bar plots were used to compare contribution/relationship across categories.
- Yearly success rate plot showed success increasing consistently from the year 2013 to 2020.

URL: SpaceX EDA Data Visualization

EDA with SQL

- First, the number of unique Launch Sites were extracted.
- First five records containing Launch Sites starting with the string “CCA” were extracted.
- Total value of Payload Mass carried by boosters (launched) by NASA (CRS) was calculated and displayed.
- The average Payload Mass carried by the booster version F9 v1.1 was calculated.
- Date of the first successful landing in ground pad extracted.
- Names of the boosters having success in drone ship landing and Payload Mass between 4000 and 6000 extracted.
- Total number of successful and failed missioned summarized.
- Names of booster that carried maximum Payload Mass extracted.
- Records with failed drone ship landing outcome in 2015 extracted along with some other details.
- Count of each type of Landing Outcome between 2010-06-04 and 2017-03-20 ranked in descending order.

URL: SpaceX EDA with SQL using sqllite

Build an Interactive Map with Folium

- Added 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.
  - Added the same for all Launch Sites using their respective latitude and longitude coordinates.
  - Added coloured markers for landing outcomes for each Launch Site using MarkerCluster to identify Sites with relatively higher success rate. Success is green and Failure is red.
- Distance between a Launch Site and its proximities:
  - Added colored lines between the Launch Site and proximities around it like the railway, highway, coastline and closest city.

URL: SpaceX Launch Sites Locations Analysis with Folium

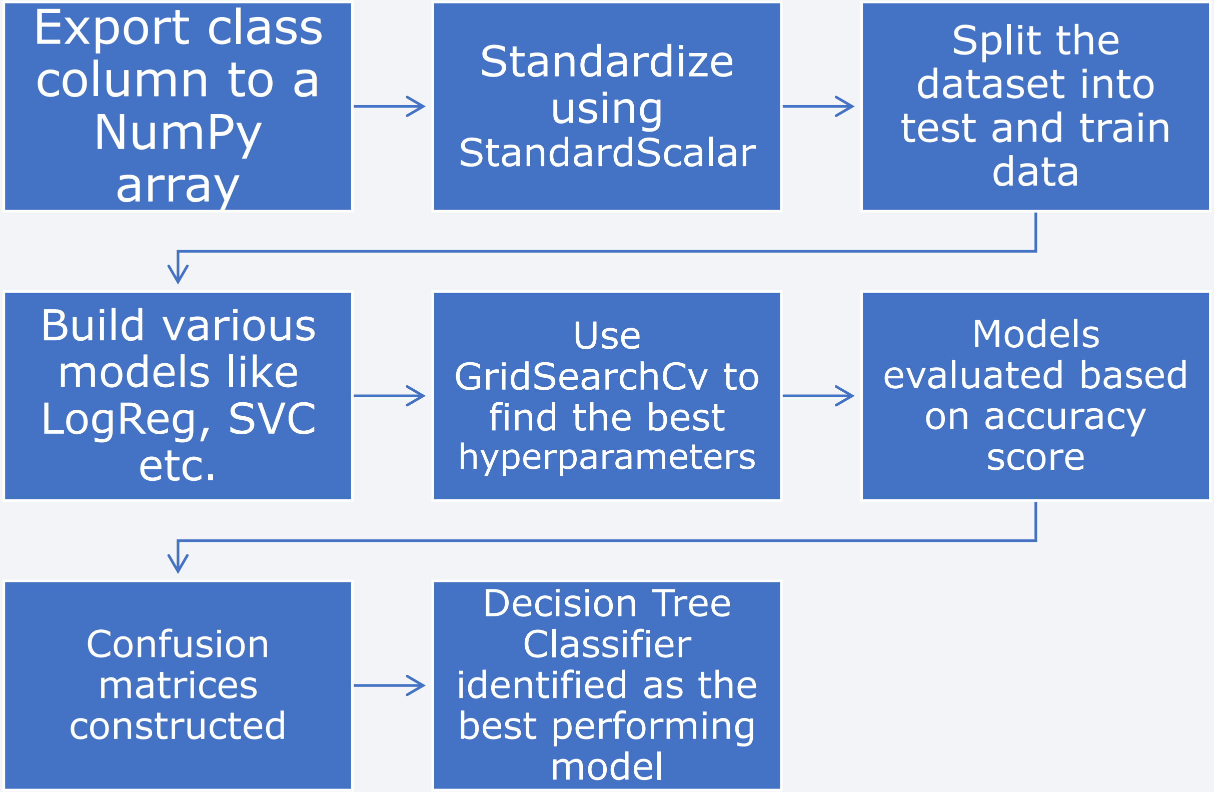
Build a Dashboard with Plotly Dash

- Launch Sites Dropdown:
  - A dropdown to select from the Launch Sites was added.
- Pie Chart showing information on Landing outcome for Launch Sites (ALL / Certain Site):
  - Pie chart showing the number of successful and unsuccessful landing outcomes for a selected Launch Site is displayed. The Launch Sites dropdown controls the output.
- Payload Mass Range Slider:
  - A range slider of the Payload Mass was created with values between the 0 and 10000, in order to select a certain Payload range.
- Scatterplot of Payload Mass vs Class for different booster versions:
  - A scatterplot showing Payload Mass vs class colored by different booster version added.

URL: Interactive Dashboard for SpaceX Launch Site

Predictive Analysis (Classification)

- The class column (dependent/target variable) of the DataFrame is **exported** to a NumPy array.
- The independent/predictor variables are **standardized** using StandardScalar.
- The datasets are split using train_test_split with a test size of 20%.
- GridSearchCV was used to build and find the best hyperparameters of various ML models like Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, kNN Classifier.
- Confusion matrix constructed. Models evaluated based on model score.



Export class column to a NumPy array
Standardize using StandardScalar
Split the dataset into test and train data
Build various models like LogReg, SVC etc.
Use GridSearchCv to find the best hyperparameters
Models evaluated based on accuracy score
Confusion matrices constructed
Decision Tree Classifier identified as the best performing model

URL: SpaceX Machine Learning Prediction

Results

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

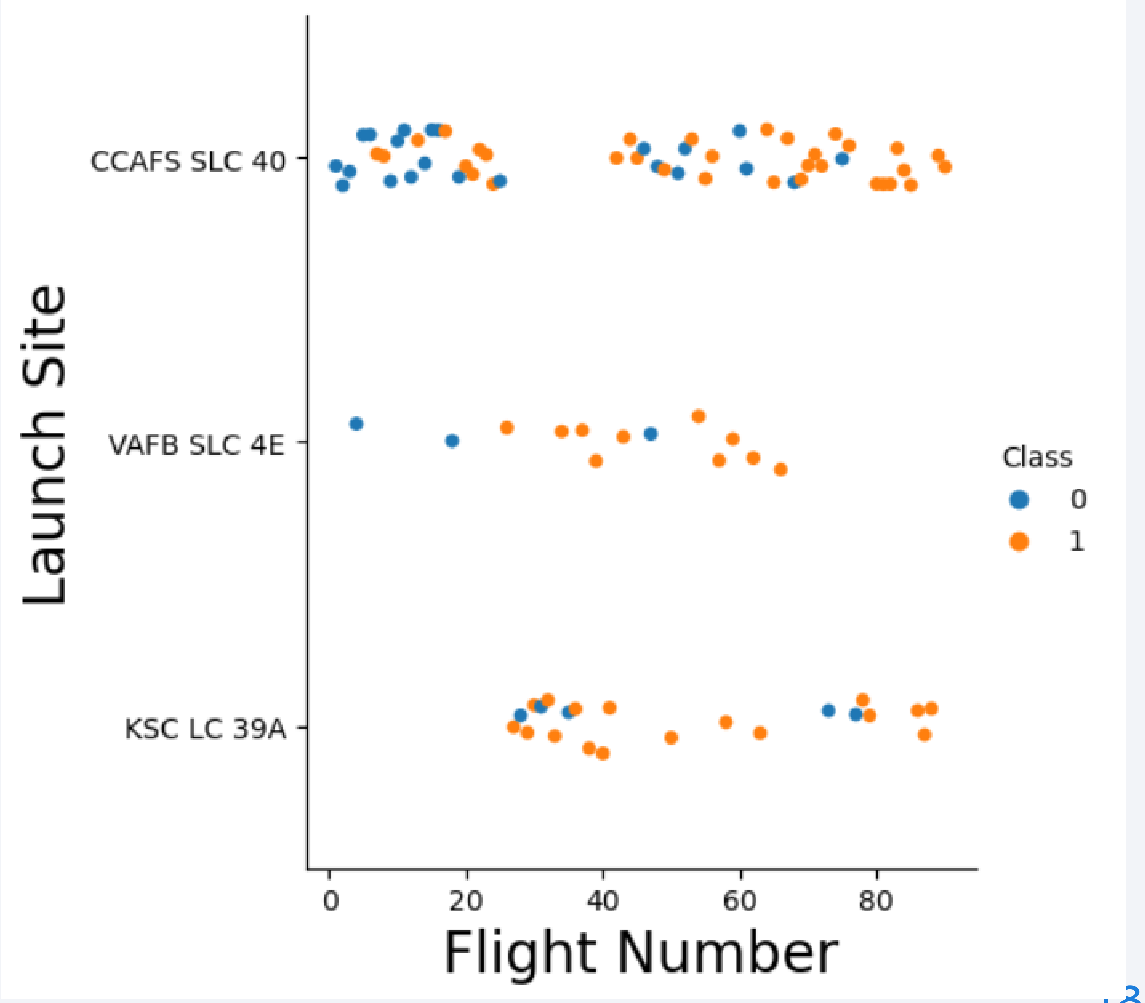
(IN SUBSEQUENT SLIDES)

Section 2

Insights drawn from EDA

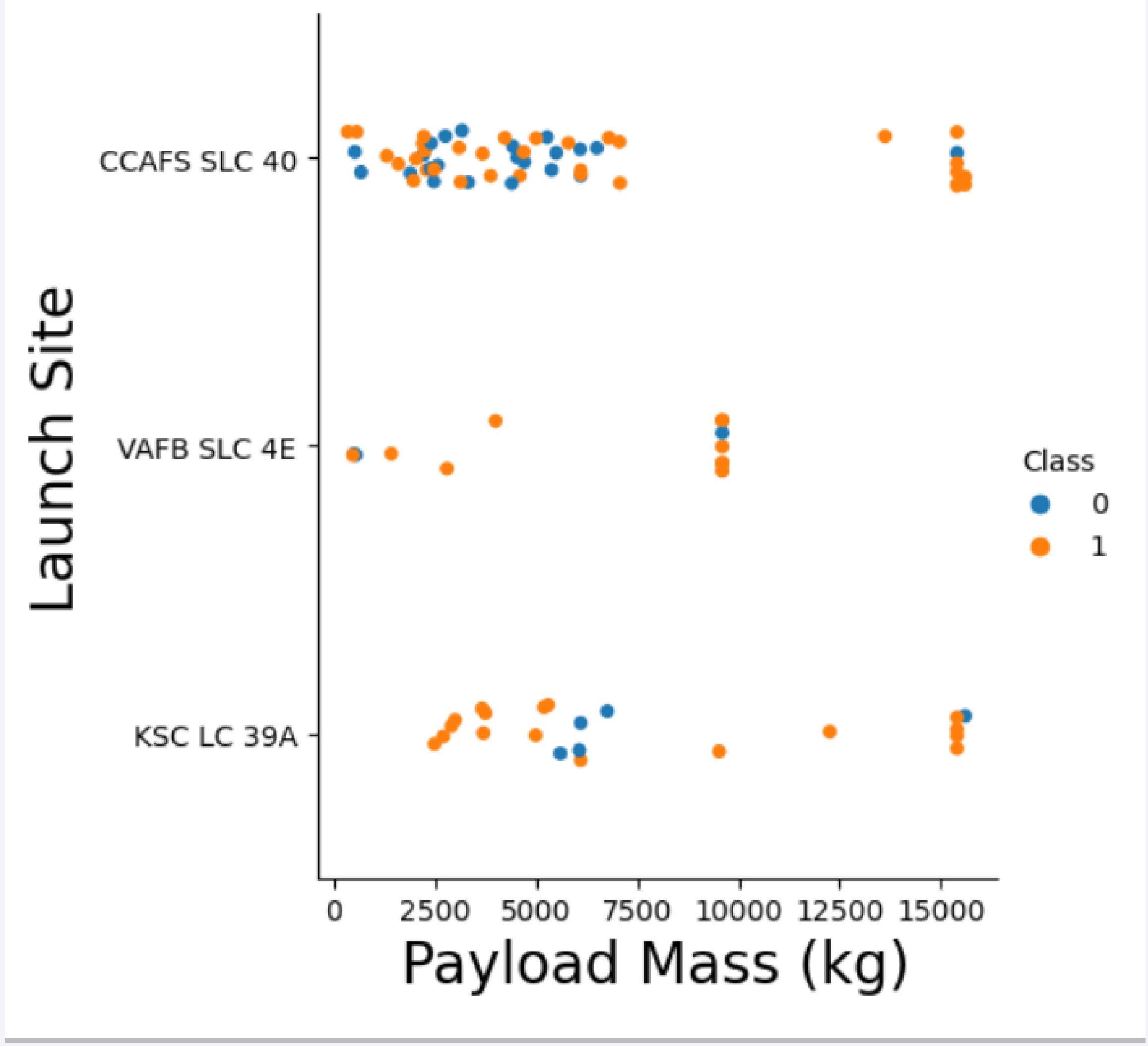
Flight Number vs. Launch Site

- The Launch Site vs Flight Number scatterplot is shown here.
- A majority of the early flights/launches resulted in unsuccessful landing outcomes
- The CCAFS SLC 40 seems to be a preferred choice as half of all launches happened here.
- However, the VAFB SLV 4E and KSC LC 39A boast of a higher success rate as compared to CCAFS SLC 40.
- Most of the recent flights/launches resulted in successful landing outcomes.



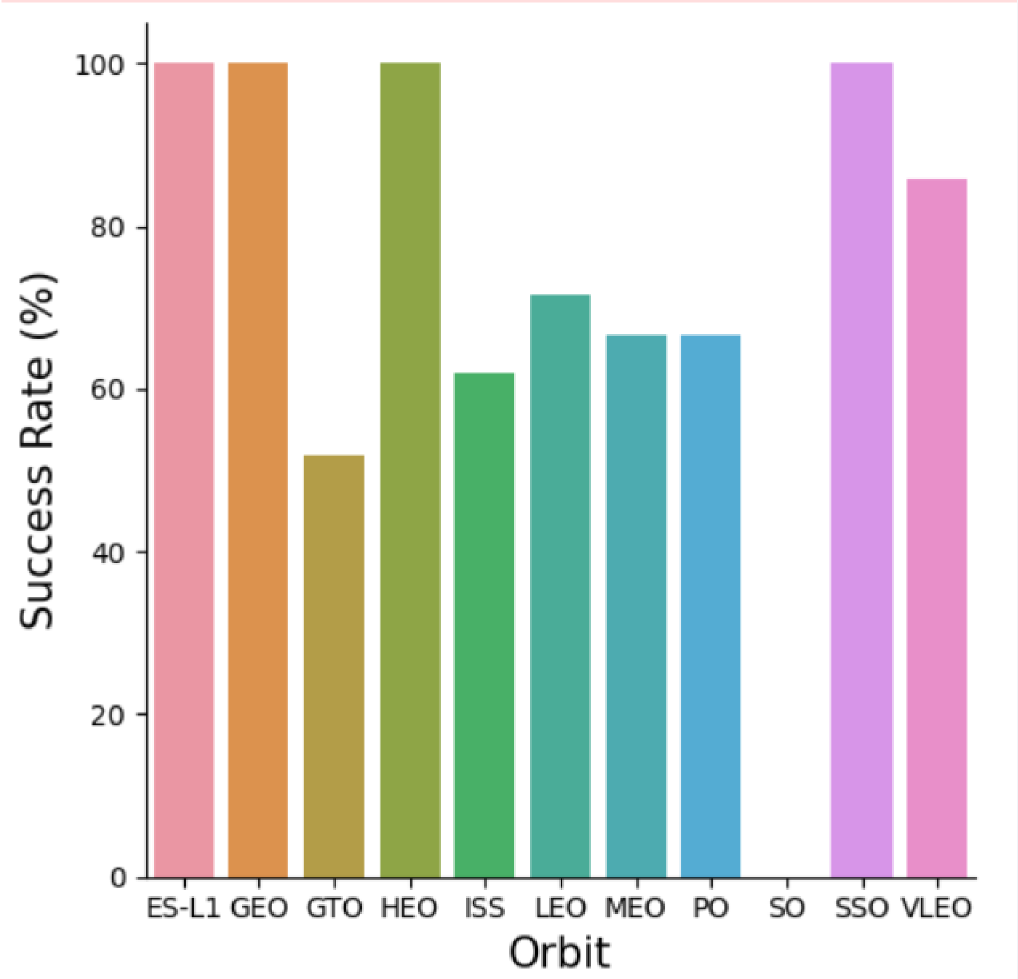
Payload vs. Launch Site

- A scatterplot of Payload vs. Launch Site has been presented here.
- The CCAFS SLC 40 Launch Site is predominantly used to launch low and high Payload capacities.
- VAFB SC 4E is used for moderate Payload launches.
- Low Payload missions tend to have a higher failure rate, whereas high Payload ones tend to have a higher success rate.
- KSC LC 39A has a 100% success rate for Payload launches below 5000kg.



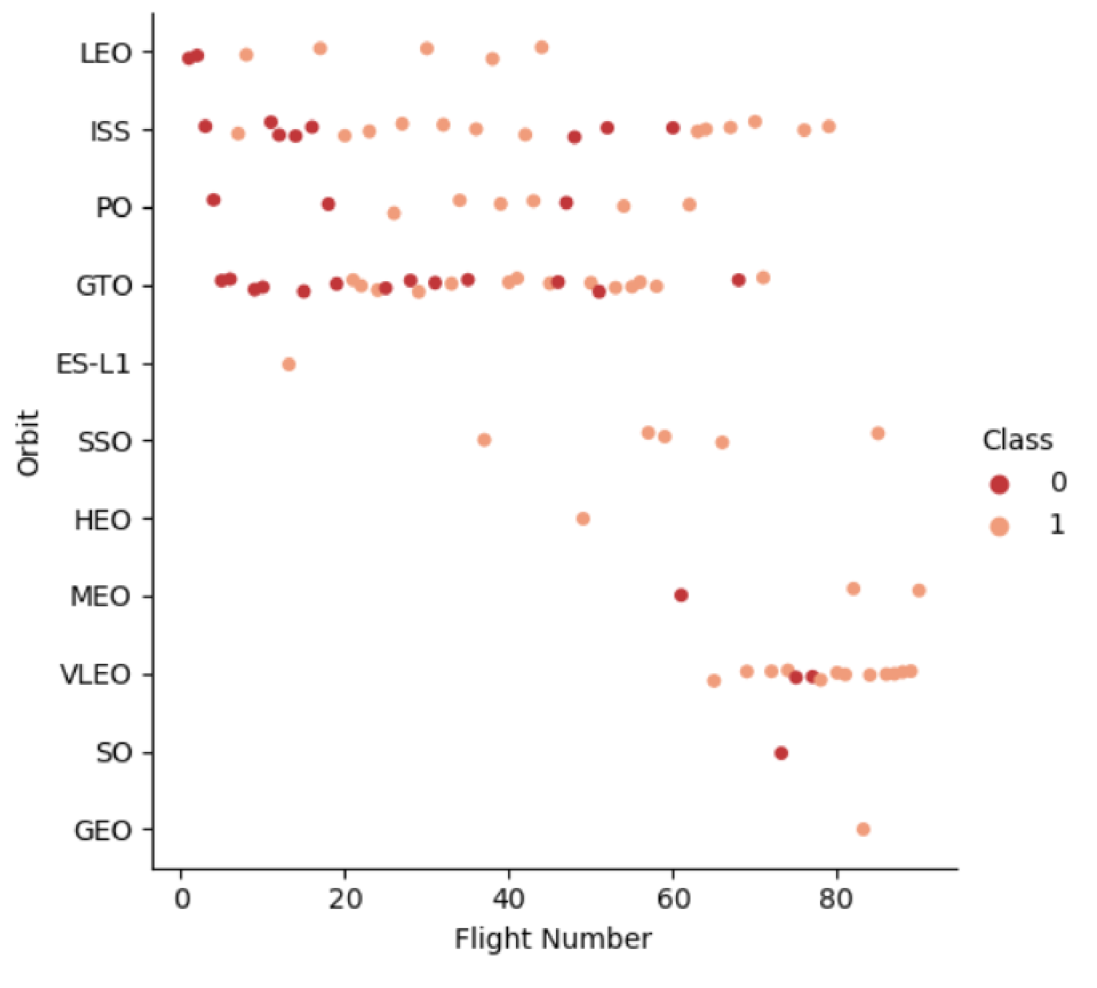
Success Rate vs. Orbit Type

- A bar chart for the success rate of each orbit type has been presented here.
- There are many types of orbits used in the missions.
- Some orbits' missions have a landing success rate of 100%, namely, ES-L1, GEO, HEO, SSO.
- While other have a landing success rate ranging from 50-70%, namely, GTO, ISS, LEO, MEO, PO, VLEO.
- Missions with the orbit SO used have a 0% success rate.



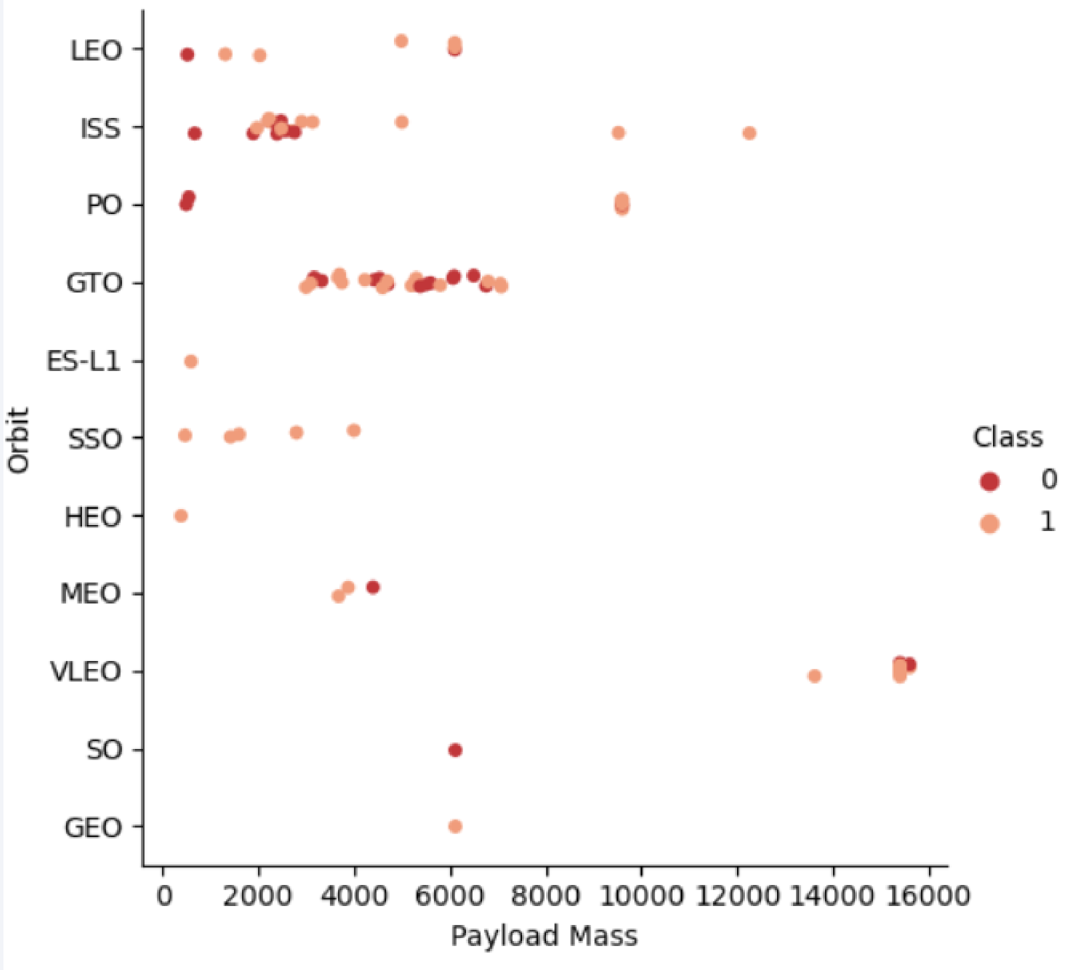
Flight Number vs. Orbit Type

- A scatter point of Flight number vs. Orbit type has been presented here.
- Early launches tended have certain orbits, namely, LEO, ISS, PO, GTO. They had a roughly 60% success rate.
- Recent launches have orbits like the SSO, HEO, MEO, VLEO, SO, GEO.
- All mission with ES-L1, SSO, GEO were successful.
- Additionally, missions with orbit VLEO have a high success rate.
- The orbits ES-L1 and GEO are used less with each being used in only one launch each.



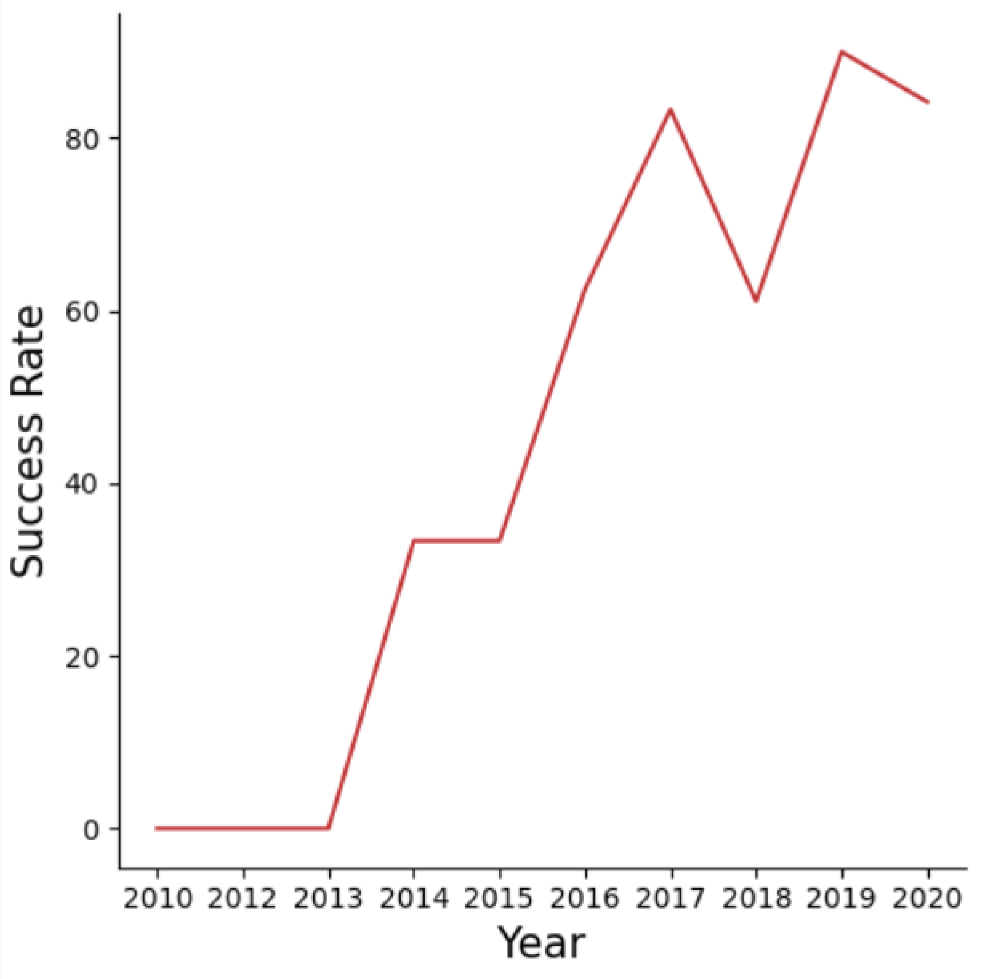
Payload vs. Orbit Type

- A scatter point of payload vs. orbit type has been presented here.
- Most orbits have a lower Payload Mass missions.
- High Payloads are used for missions with orbits VLEO, PO, ISS.
- Higher Payload missions have relatively higher success rates.
- Orbits ES-L1, SSO, HEO have low Payloads with 100% landing success rate.

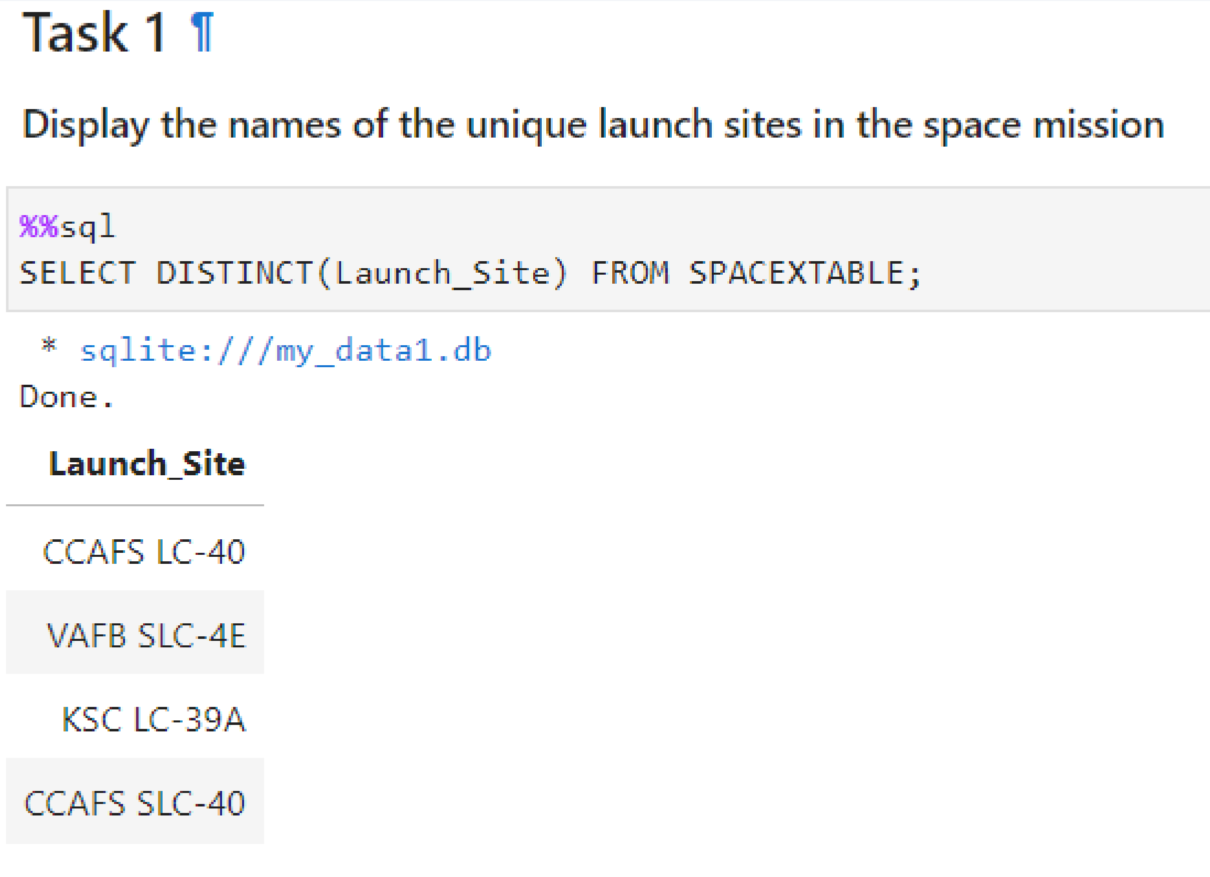


Launch Success Yearly Trend

- A line chart of yearly average success rate has been presented here.
- The Landing success has steadily increased over the years from 2013.
- 2018 saw a dip in the success rate.
- Overall, the success rate increased till the year 2020.



All Launch Site Names



Task 1 ¶

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

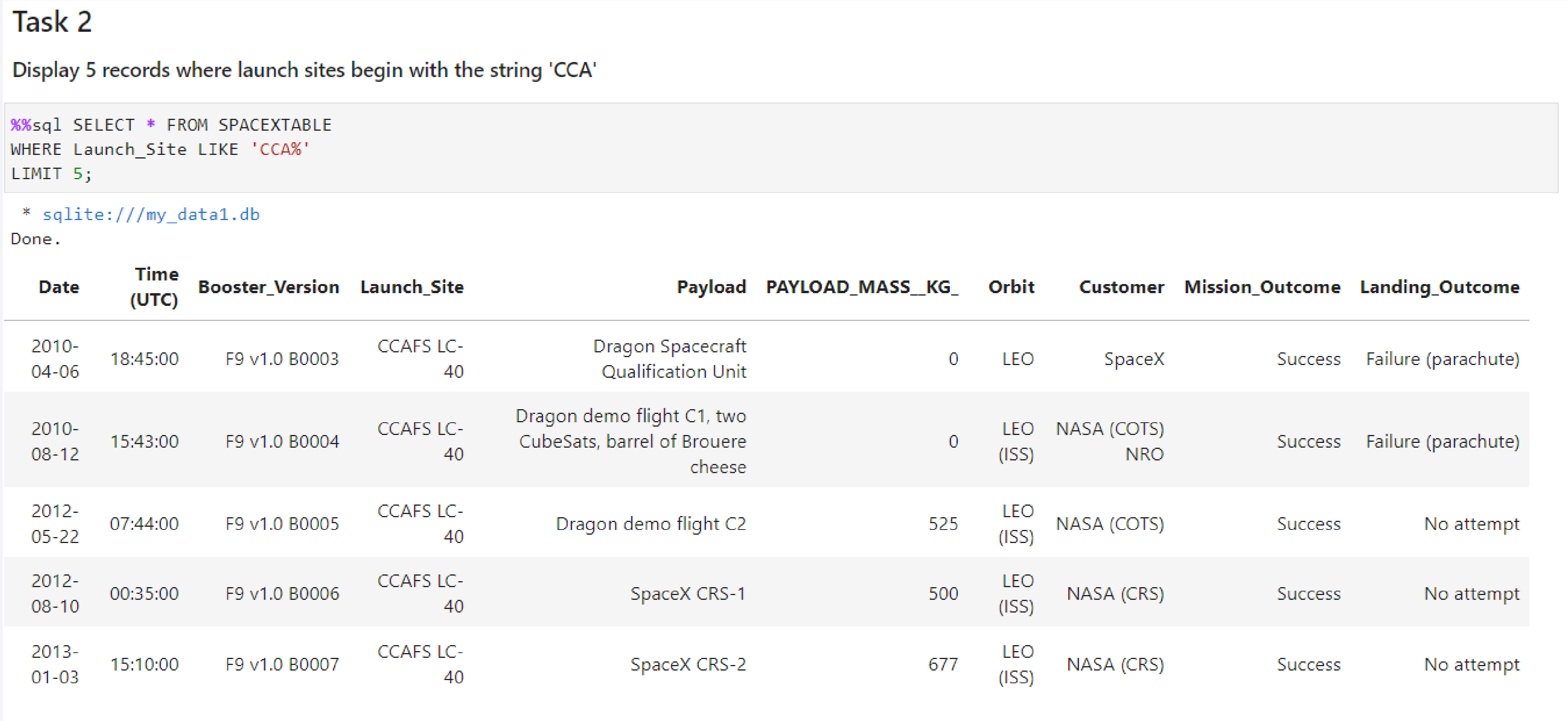
```
%%sql
SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE;
```

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

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

- Names of Launch Sites has been displayed.

Launch Site Names Begin with 'CCA'



Task 2

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

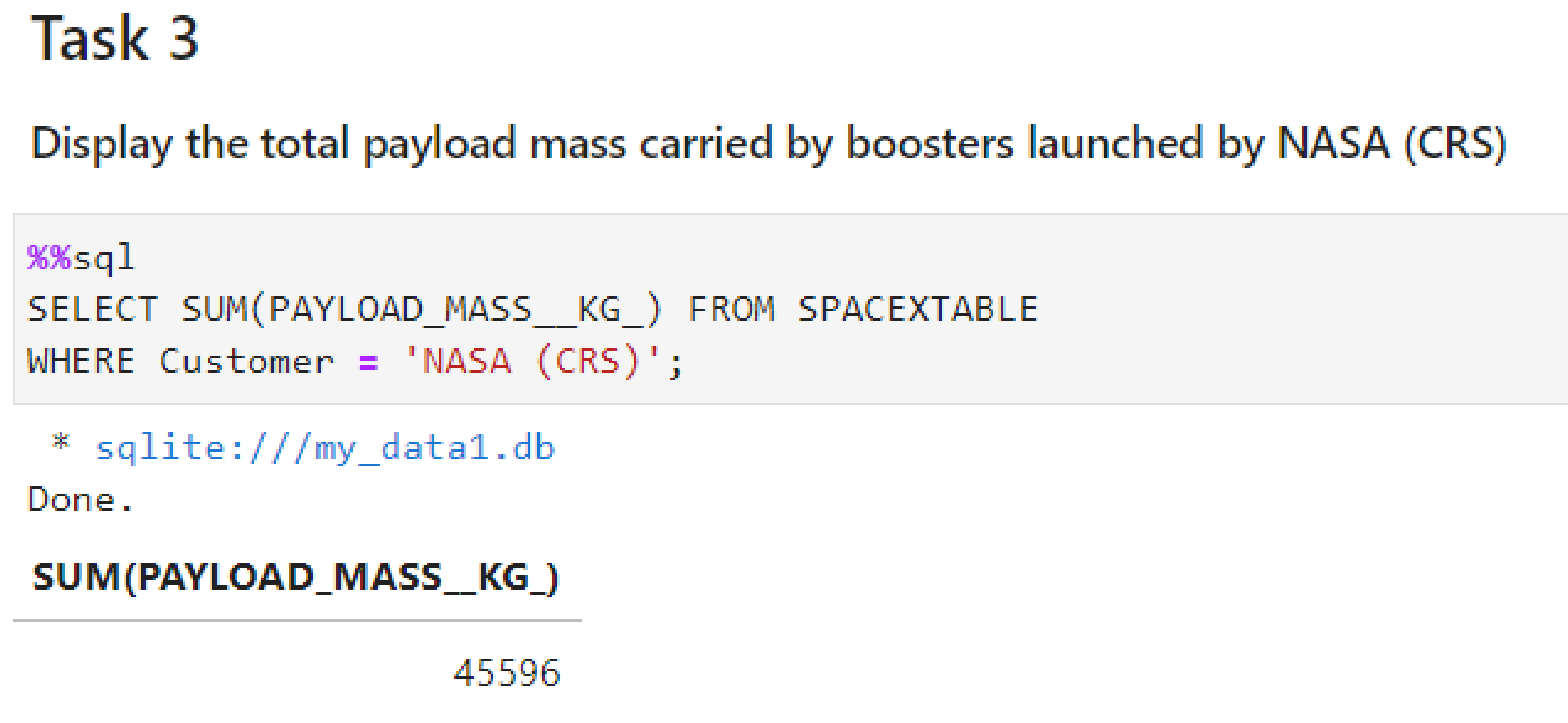
```
%%sql SELECT * FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	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

- 5 Records with Launch Sites starting with the string ‘CCA’ displayed

Total Payload Mass



Task 3

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

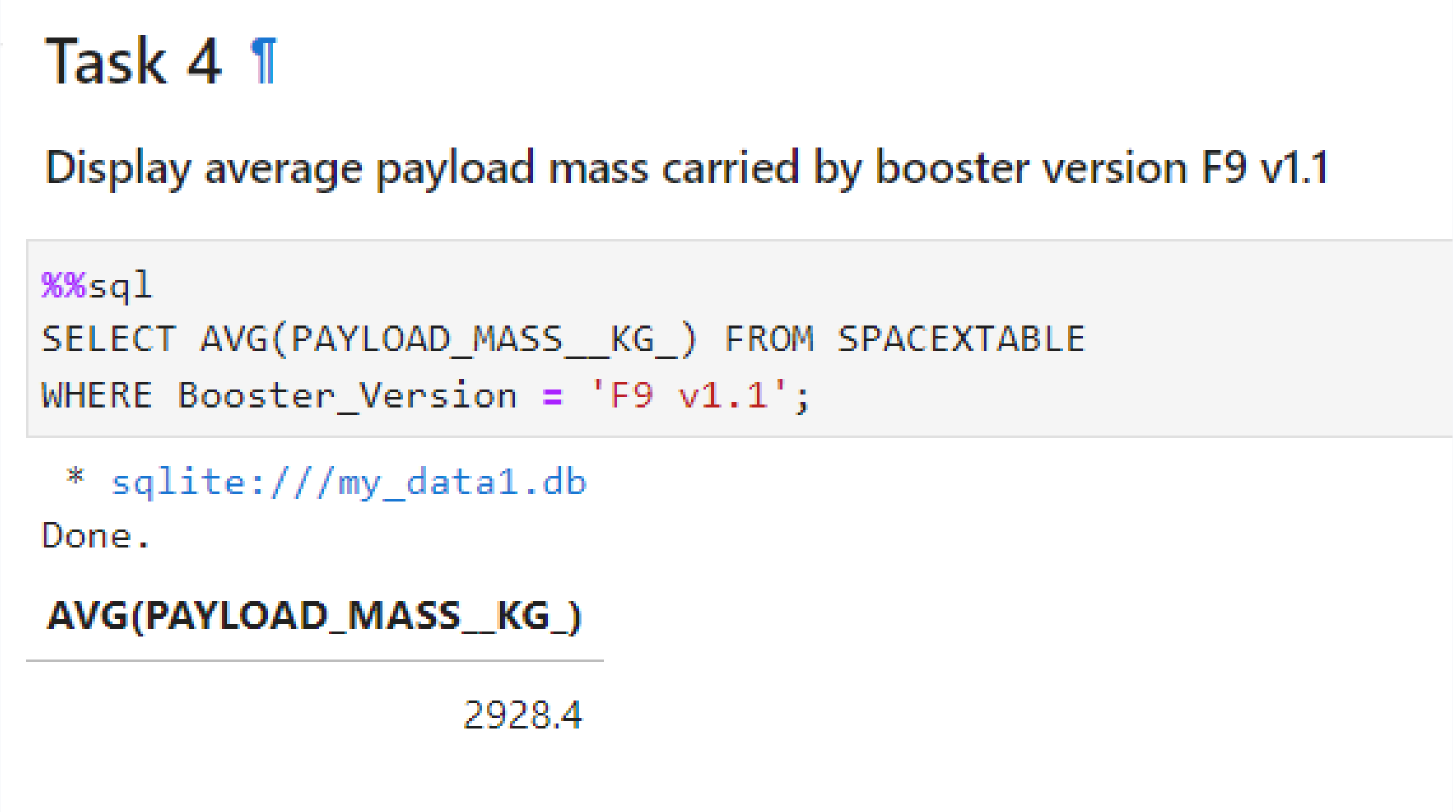
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- The total payload mass launched by NASA has been calculated and presented.

Average Payload Mass by F9 v1.1



Task 4 ¶

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

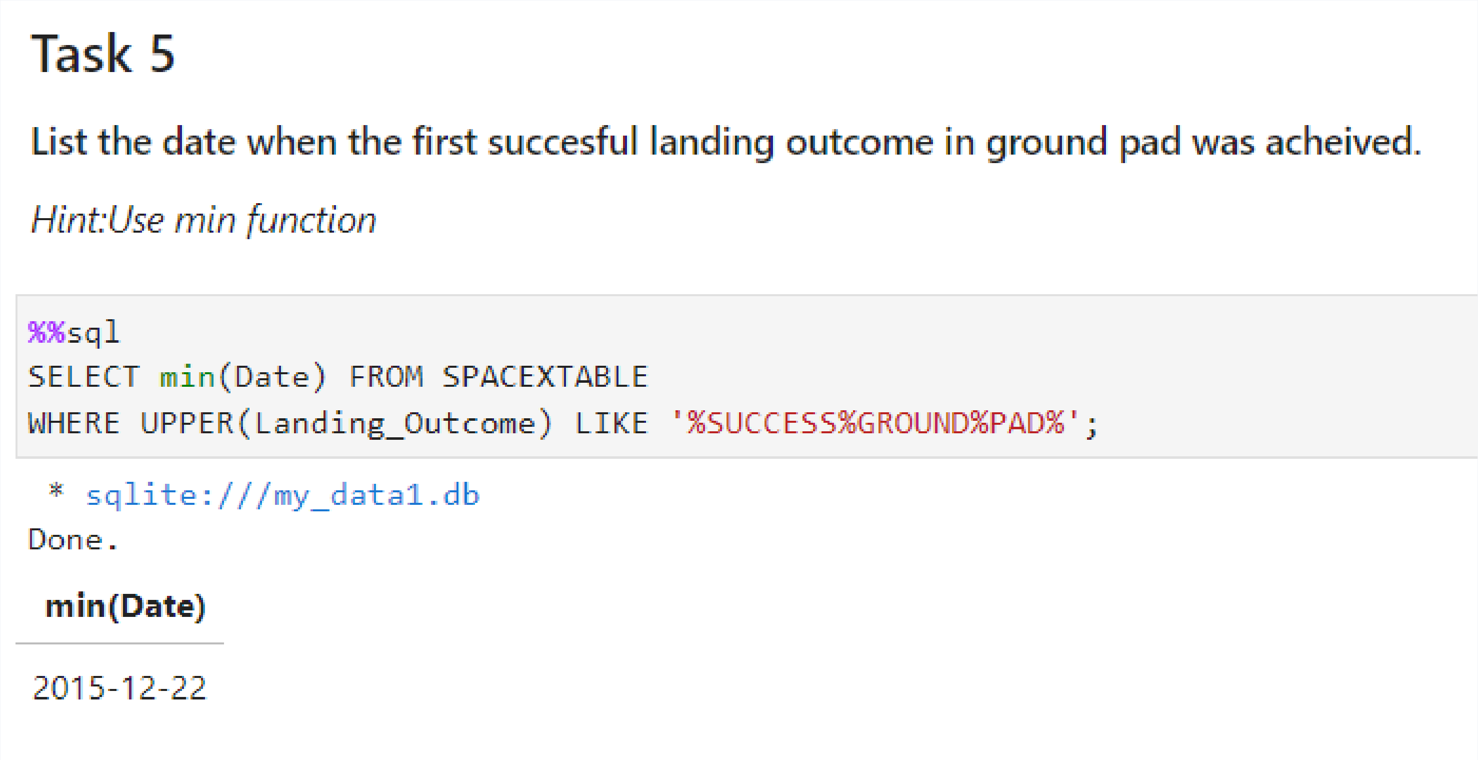
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average payload mass carried by booster version F9 v1.1 has been calculated and presented.

First Successful Ground Landing Date



Task 5

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

Hint:Use min function

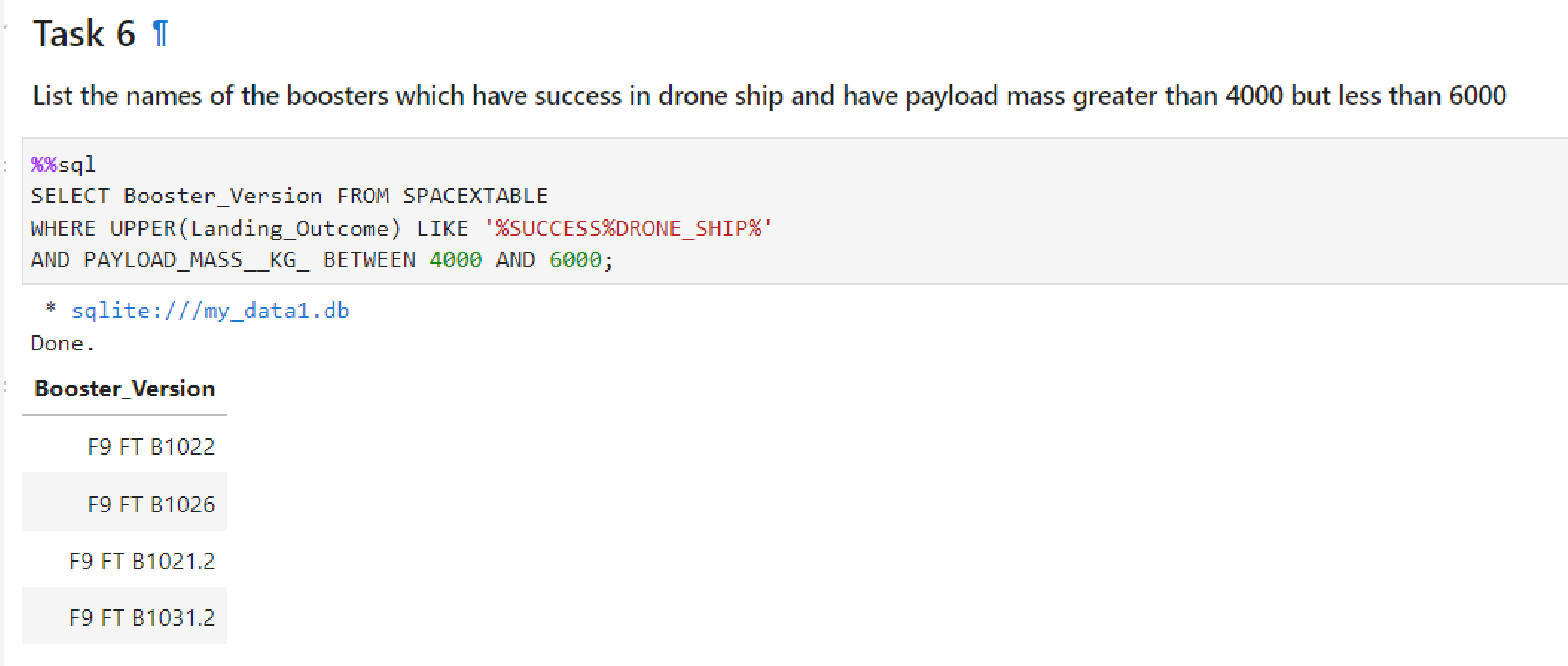
```
%%sql
SELECT min(Date) FROM SPACEXTABLE
WHERE UPPER(Landing_Outcome) LIKE '%SUCCESS%GROUND%PAD%';
```

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

min(Date)
2015-12-22

- Date of the first successful landing outcome on ground pad bas been extracted.

Successful Drone Ship Landing with Payload between 4000 and 6000



Task 6 ¶

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

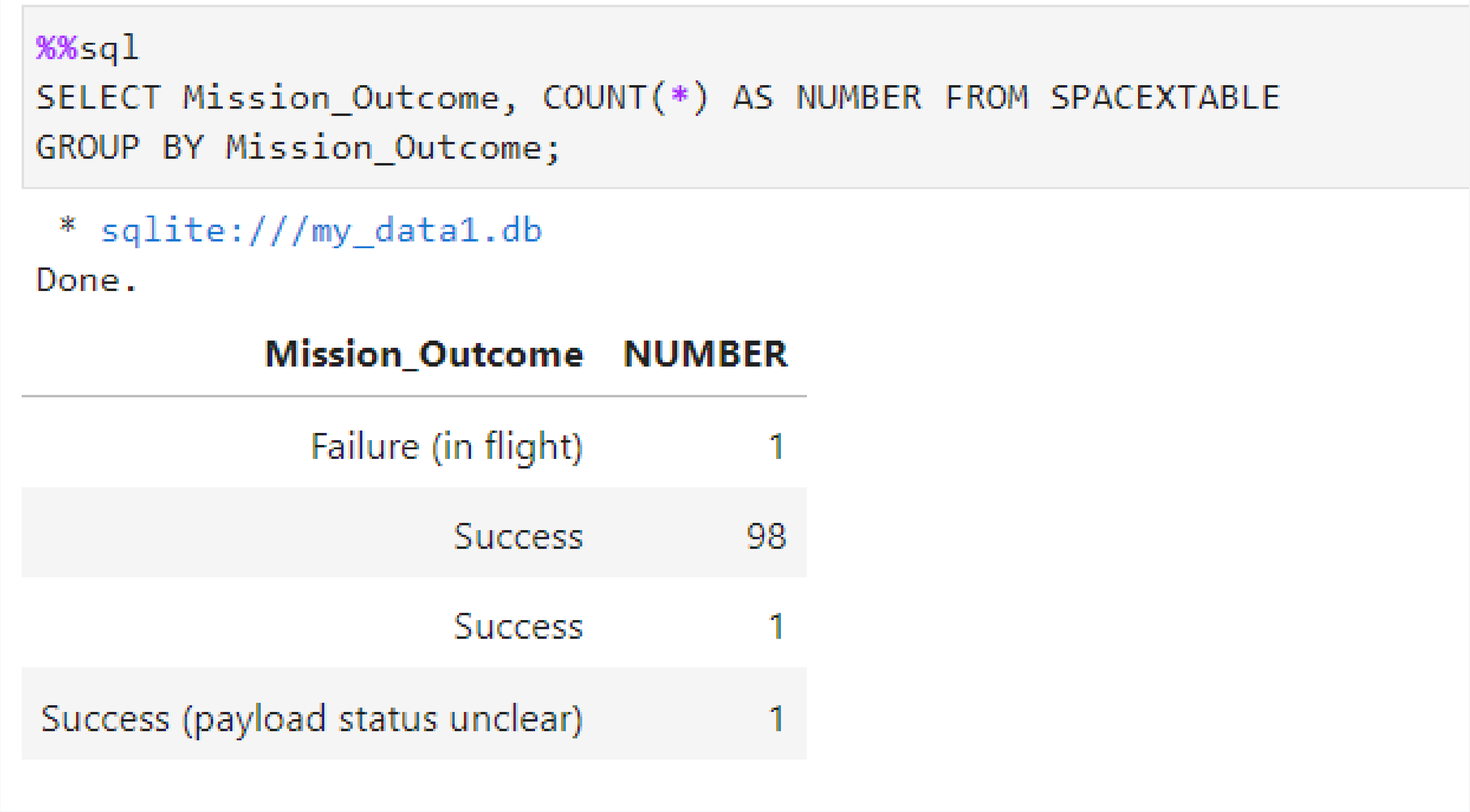
```
%%sql
SELECT Booster_Version FROM SPACEXTABLE
WHERE UPPER(Landing_Outcome) LIKE '%SUCCESS%DRONE_SHIP%'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

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

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 has been extracted and presented.

Total Number of Successful and Failure Mission Outcomes



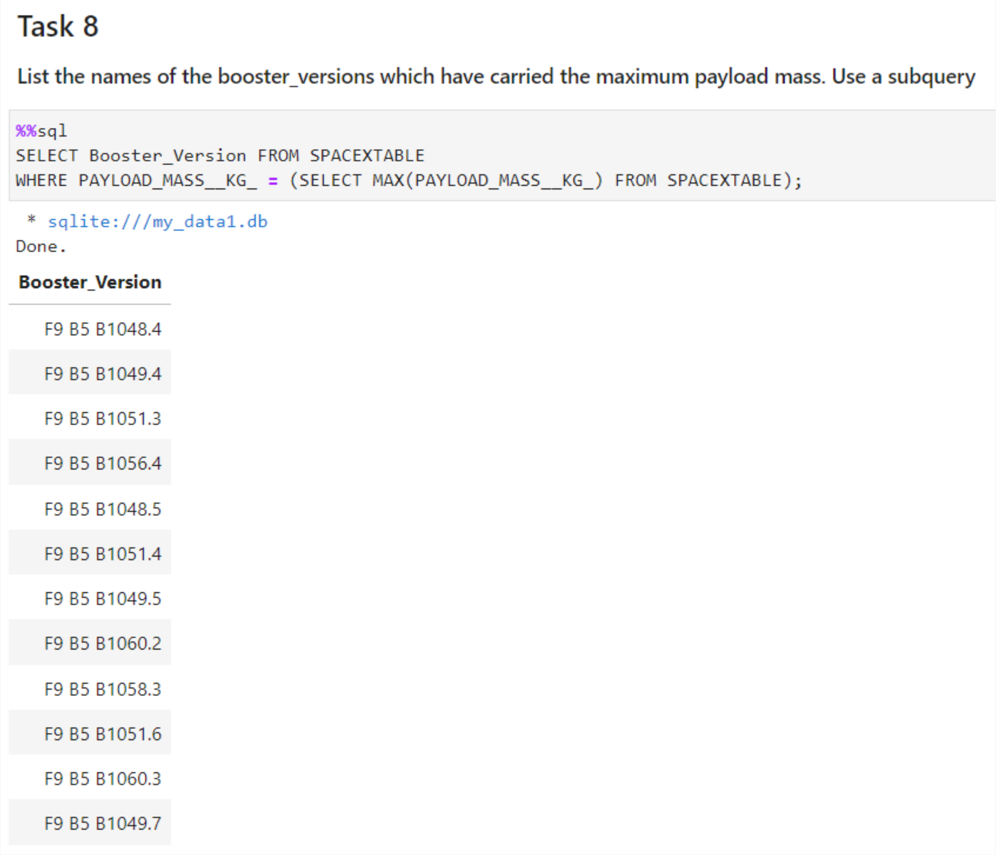
```
%%sql
SELECT Mission_Outcome, COUNT(*) AS NUMBER FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

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

- Total number of successful and failure mission outcomes and their count have been summarized and presented.

Boosters Carried Maximum Payload



Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

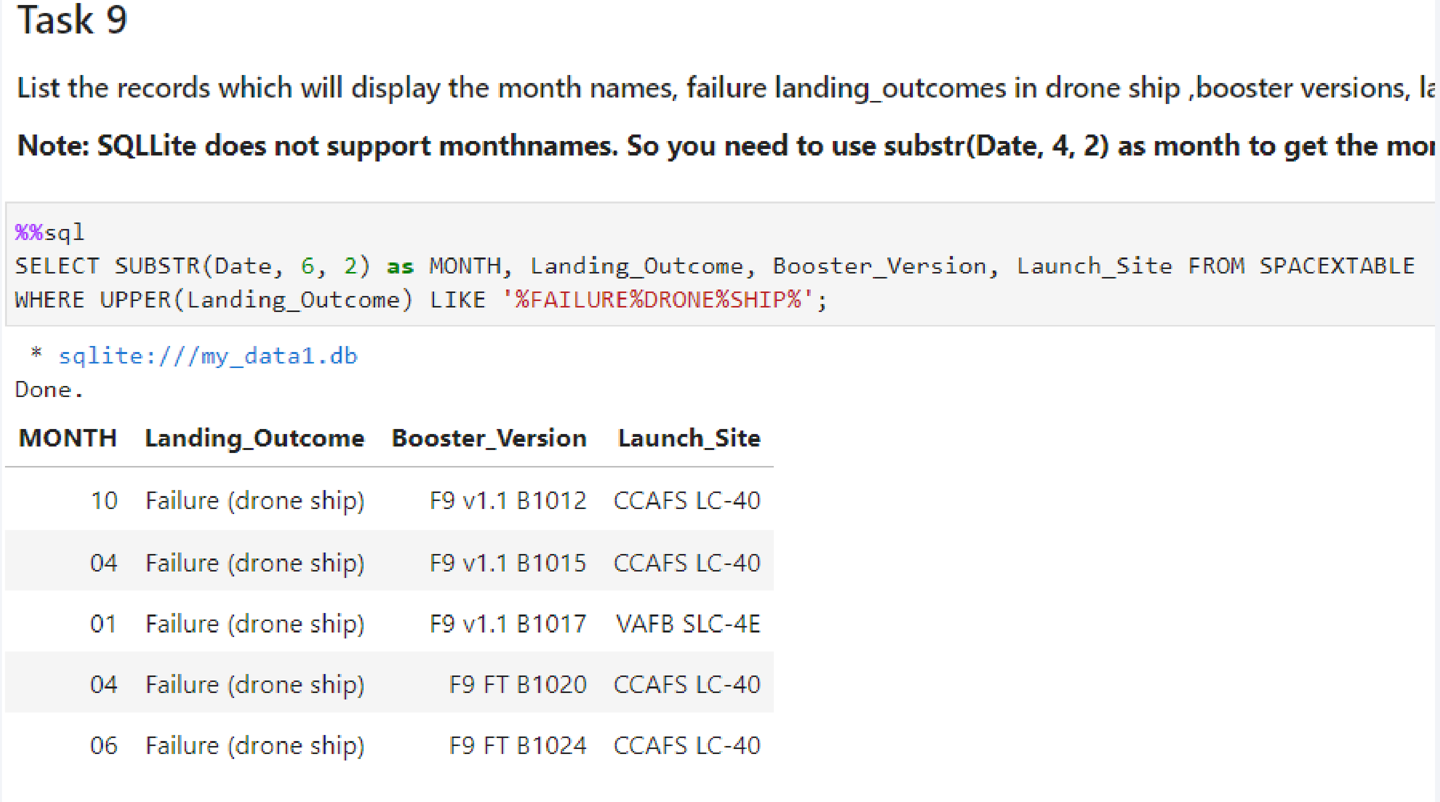
```
%%sql
SELECT Booster_Version FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

```
 * 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

- Names of the boosters which have carried the maximum payload mass have been extracted.

2015 Launch Records



Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, la

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the mor

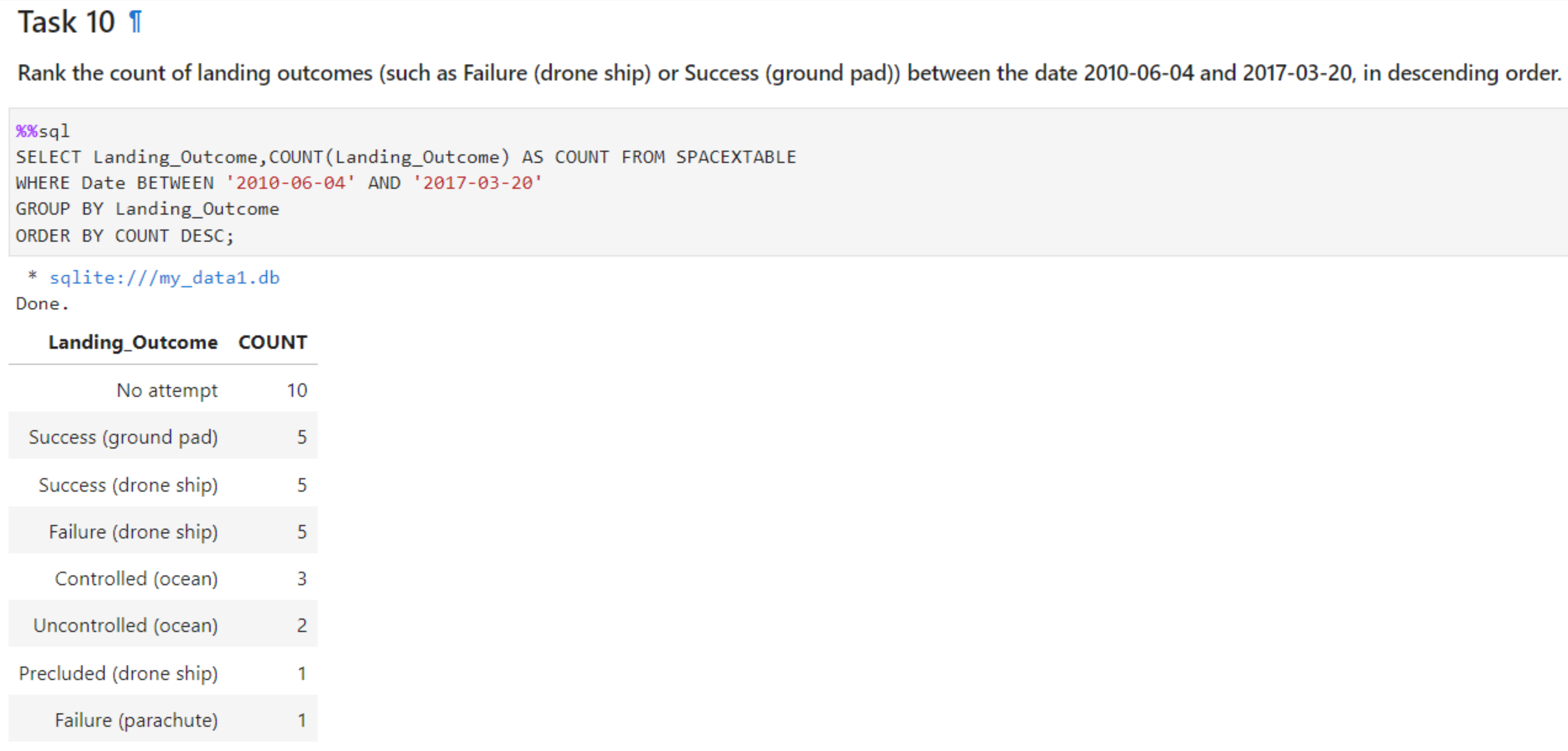
```
%%sql
SELECT SUBSTR(Date, 6, 2) as MONTH, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE UPPER(Landing_Outcome) LIKE '%FAILURE%DRONE%SHIP%';
```

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

MONTH	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
01	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
04	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
06	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 have been extracted and presented.

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



Task 10 ¶

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 FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT DESC;
```

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

Landing_Outcome	COUNT
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

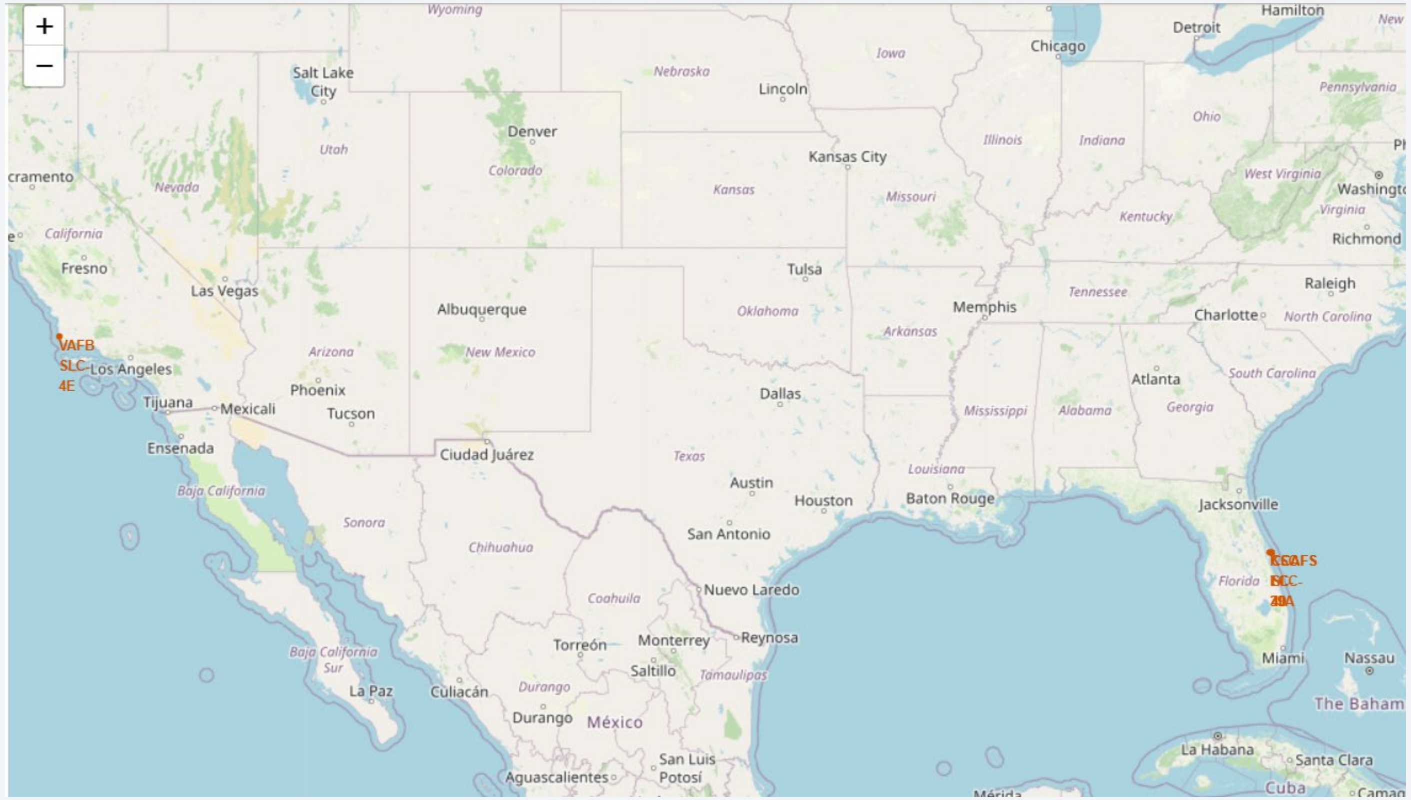
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 have been ranked, in descending order and presented.

Section 3

Launch Sites Proximities Analysis

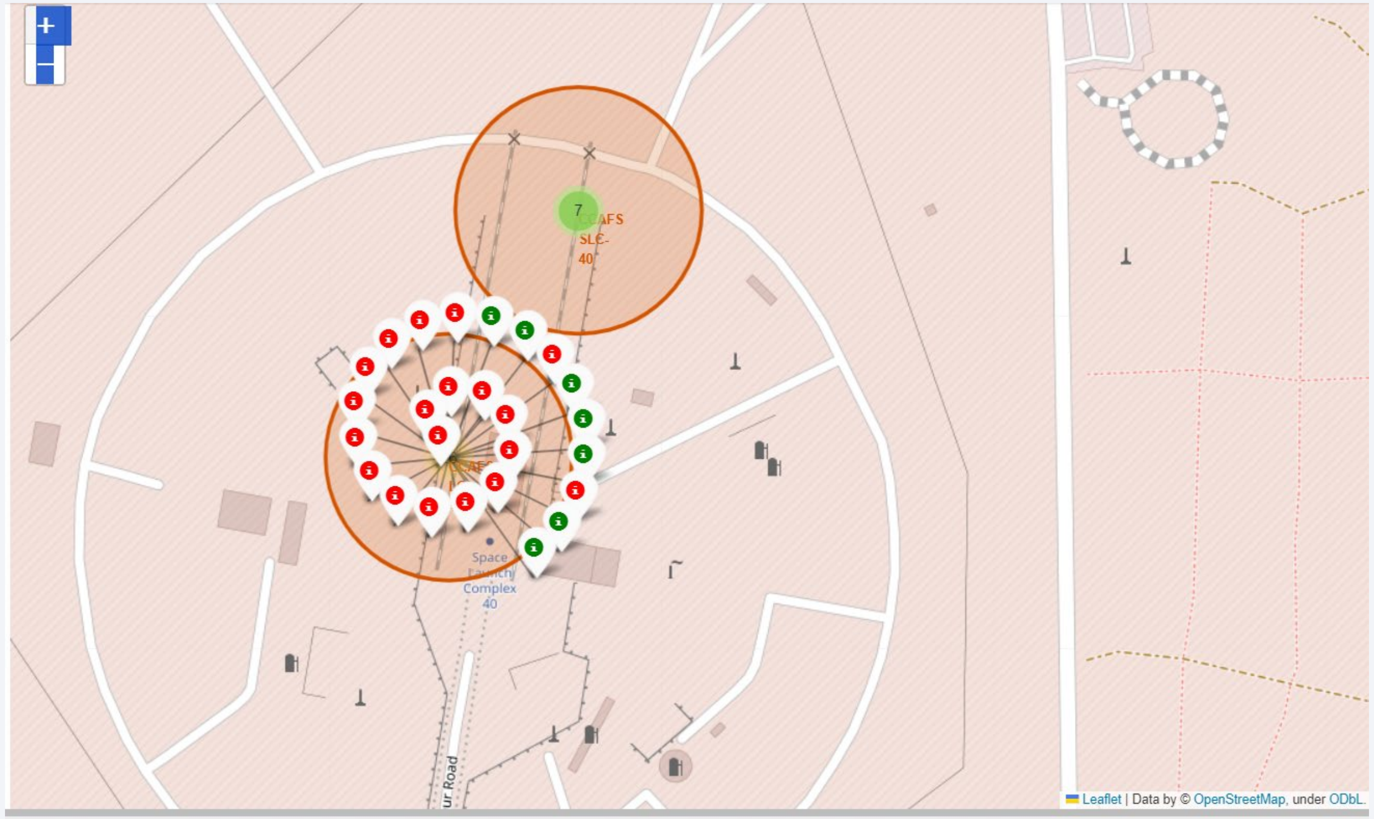
Launch Sites Location on Global Map:

- All the Launch Sites are closes to the Equator and the coastlines.
- Rockets launched near the Equator benefit from the inertia due to the rotation of the Earth on its own axis.
- Launch Sites near coastlines are preferred as the mission can be aborted with minimum damage in the ocean in case of an untoward incident.



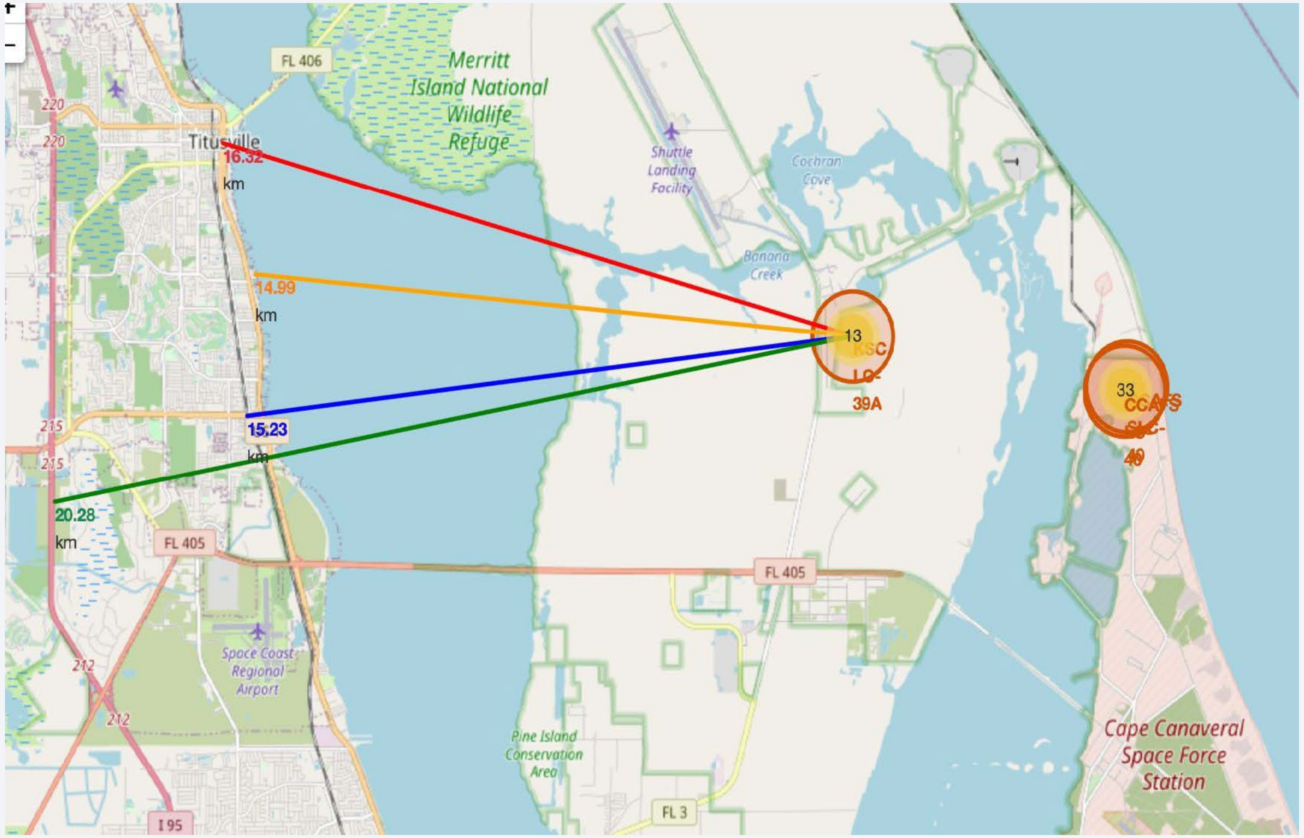
Color-based Markers for Launch Outcomes on Map:

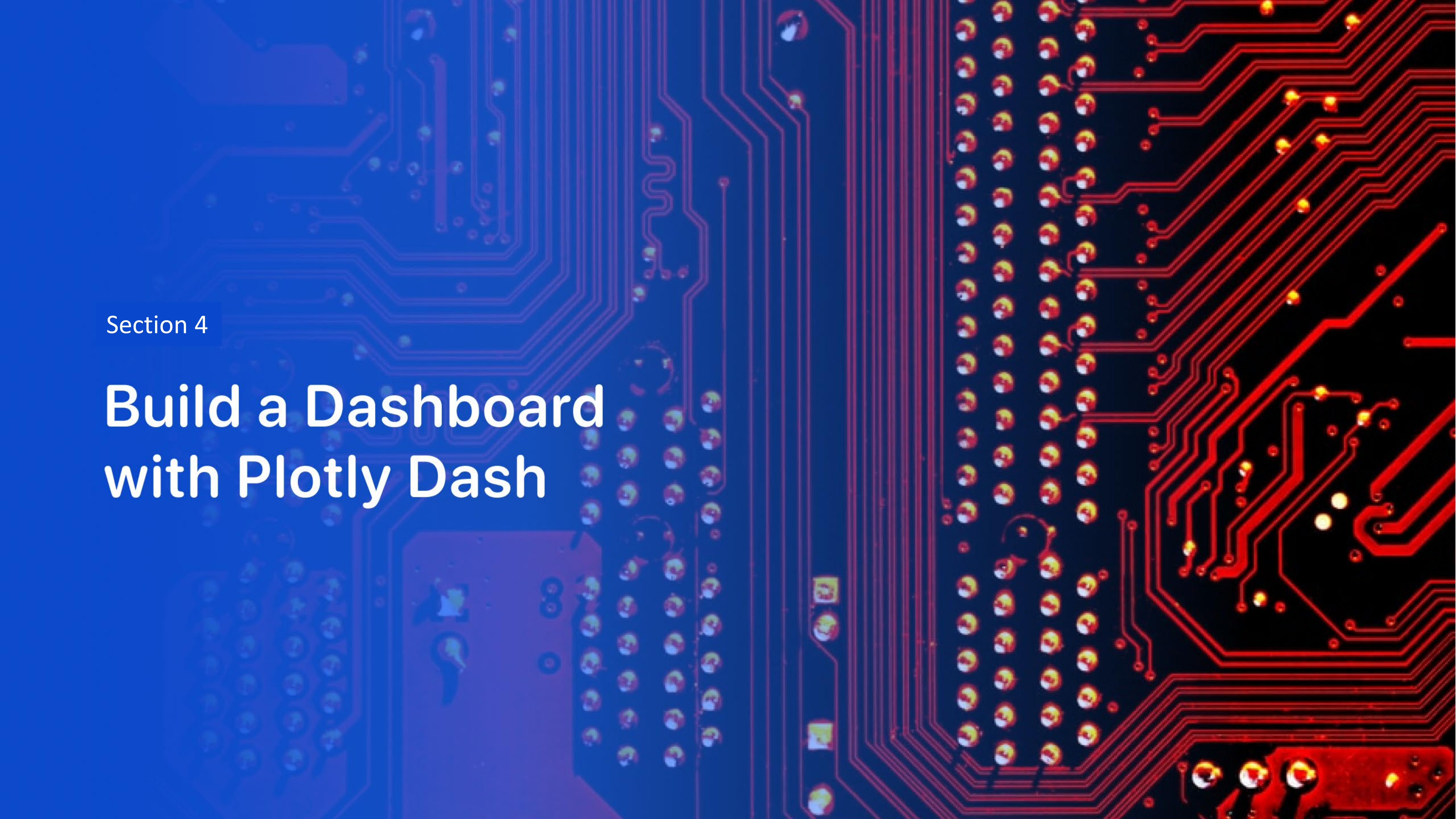
- Color-based Markers have been placed at each Launch Site to help easily visualize the success rate. This aids in identifying Launch Sites with relatively higher levels of success rates
- Green Marker signifies a Successful Launch, whereas Red signifies an Unsuccessful Launch.
- KSC LC-39A has the highest launch success rate, whereas CCAFS LC-40 has the lowest.



Distance from Launch Site to its proximities:

- Considering the Launch Site KSC LV-39A, we see that relatively it is –
  - 15.23km from the railway.
  - 20.28km from the highway.
  - 15km from the coastline
  - 16.32km from the nearest city Titusville.
- Location of the Launch Site is good as it is near various facilities and away from population centers.

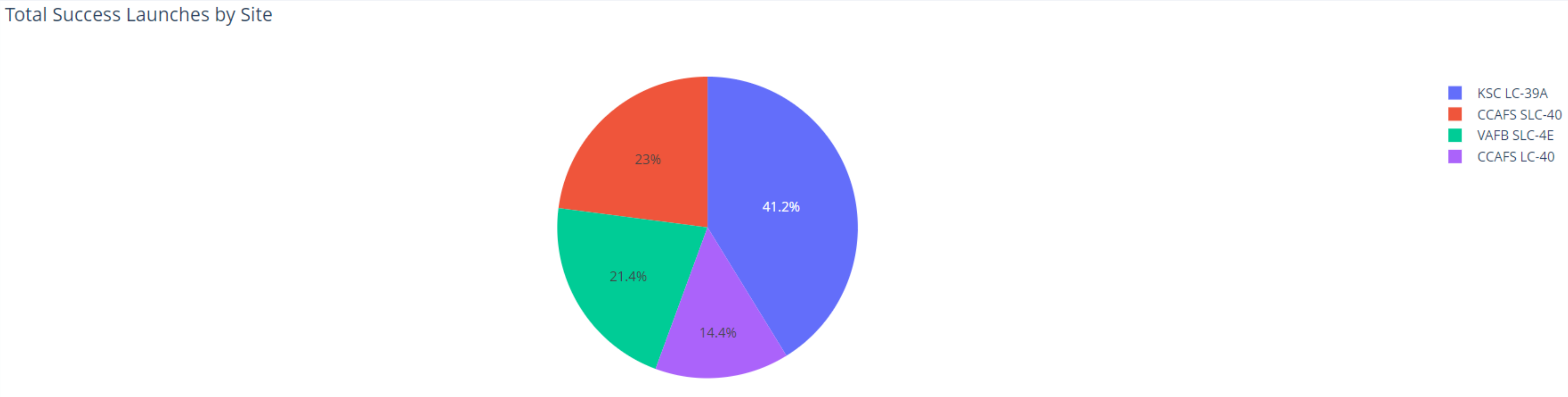




Section 4

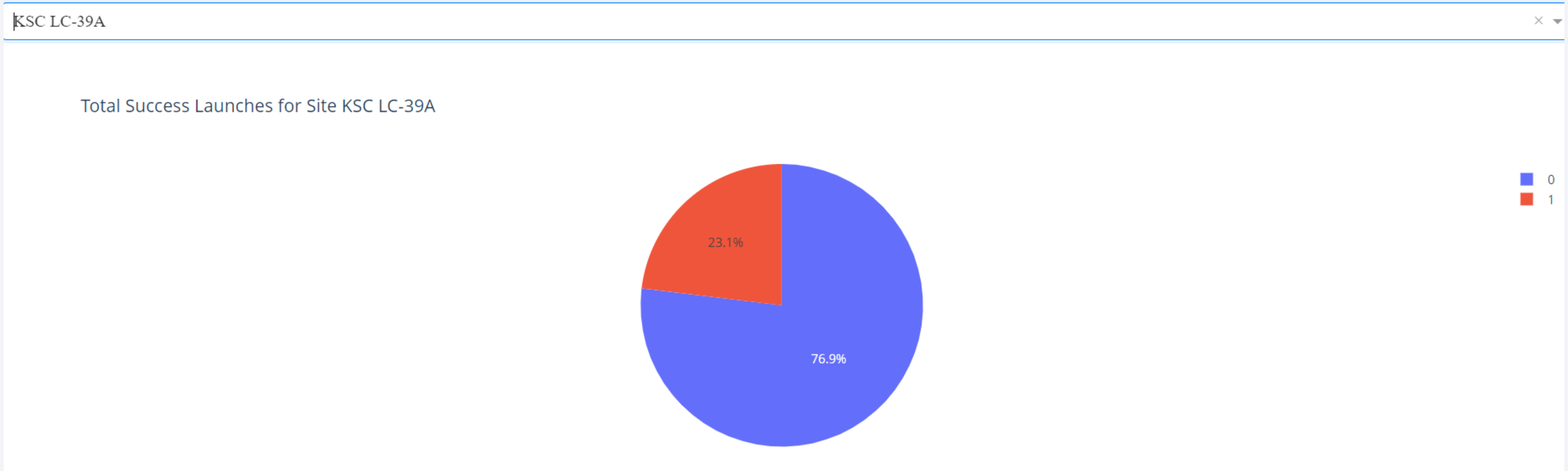
Build a Dashboard with Plotly Dash

Launch Success count for all Sites:



- Positive Launch Success outcomes for the Launch Sites have been presented here as a pie chart.
- It can be seen that the Launch Site KSC LC-39A has the highest share of successful launches.

Launch Site data with highest launch success ratio:

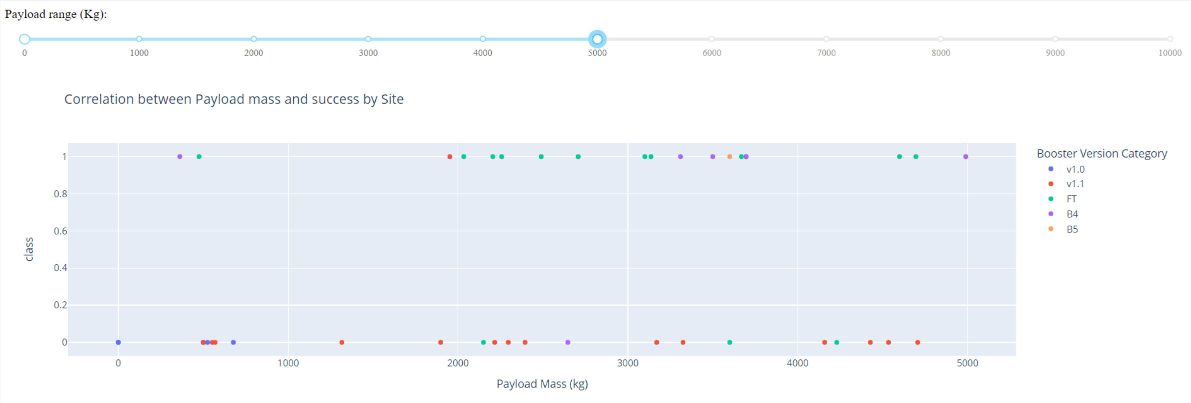


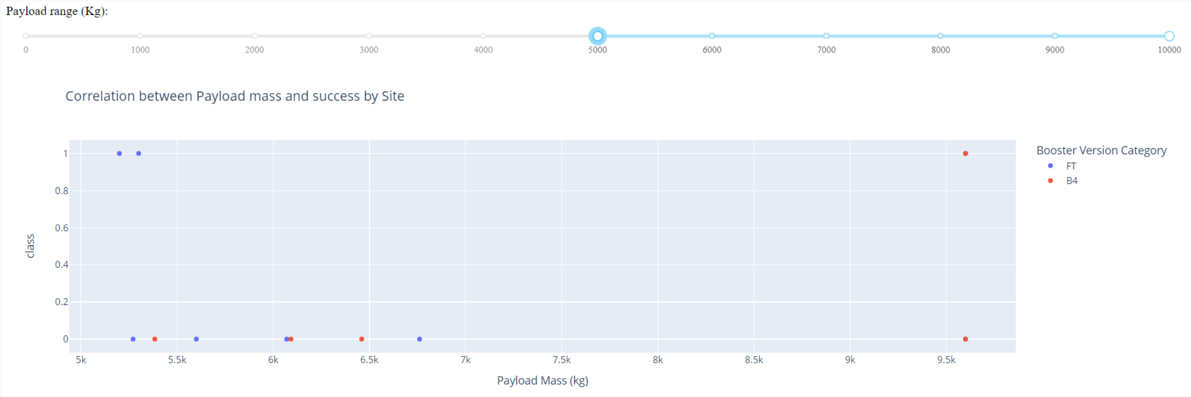
KSC LC-39A

Total Success Launches for Site KSC LC-39A

- KSC LC-39A is the Launch Site with the highest success rate.
- 76.9% of all launches from here were successful.

Payload Mass vs Launch Outcomes for all Sites:





- The charts show the Launch Outcomes vs Payload Mass for each booster version category.
- The Booster Version Category FT is associated with successful outcomes.

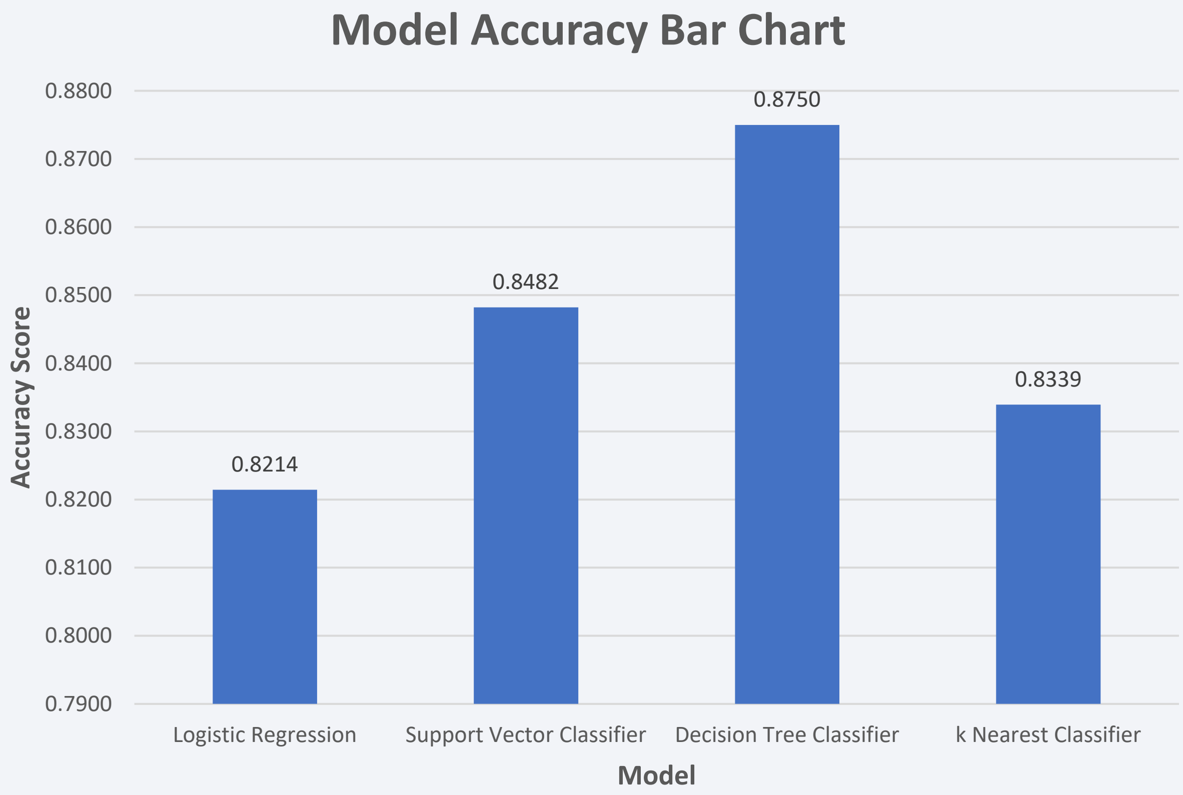
Section 5

Predictive Analysis (Classification)

Classification Accuracy

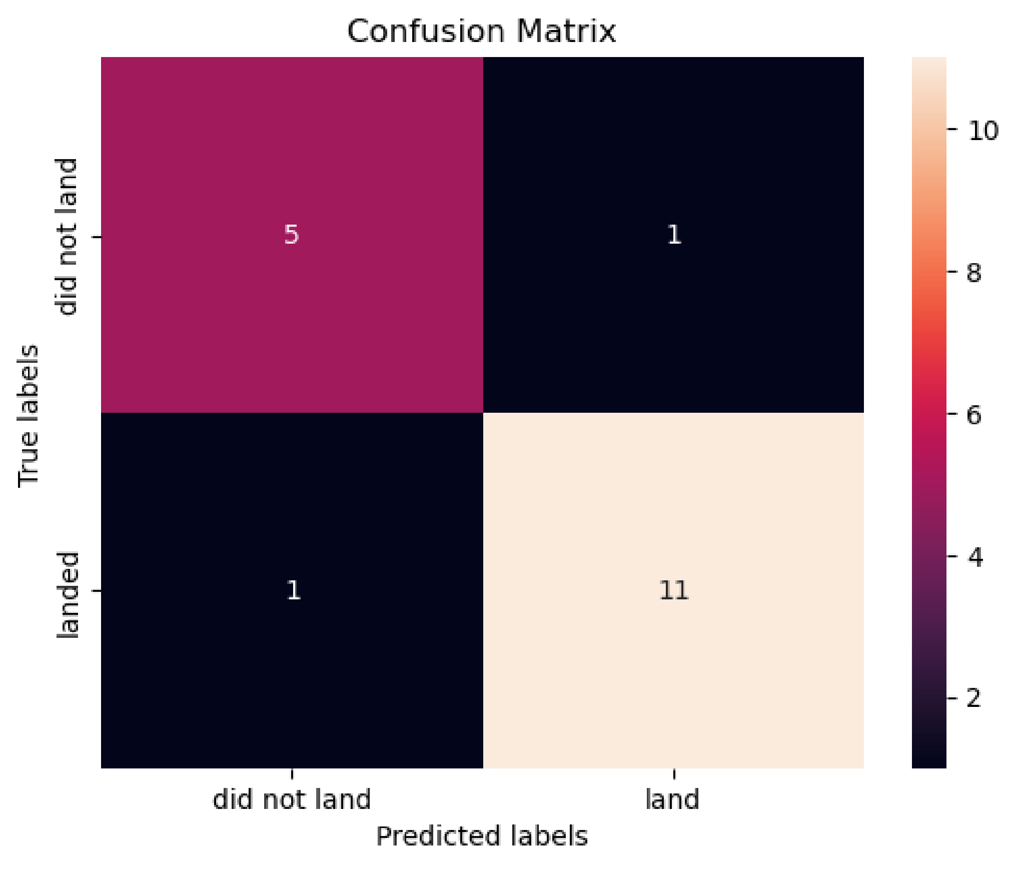
- Model accuracy for all built classification models has been presented in a bar chart here.

- Decision Tree Classifier has been found to be the best performing model with an accuracy of 0.875 or 87.5%.



Confusion Matrix

- The confusion matrix of the best performing model (Decision Tree Classifier) has been presented here.
- The model predicted 16 instances correctly and misclassified a single instance each as False Positive and a False Negative.



Conclusions

- Decision Tree Classifier is the best suited Model for prediction of this dataset.
- Higher Payload Mass launches are associated with a higher chance of success.
- All launches are done in proximity to the coastline and away from population centers.
- The success rate has consistently increased over time.
- KSC LC-39A is the Launch Site with the highest success rate.
- Missions with Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

Heartfelt Thanks to:

Instructors

IBM

Coursera

Thank you!