IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alan Gaugler
25/01/23

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection performed by web scraping the SpaceX API.
  - Data Wrangling of collected data
  - Exploratory Data Analysis of data to gain insights using SQL
  - Interactive Visual Analytics and creating maps with Folium
  - Dashboard created with Plotly Dash
  - Machine Learning to model success rate of future launches
- Summary of all results
  - Interactive analytics are demonstrated.
  - EDA determined which features are most important to predict launch success rate
  - Several models were developed to predict the success rate of landing the first stage of the rocket. The most successful model is a decision tree model with an accuracy of 94.44%

Introduction

- Project background and context

  SpaceX is pioneering space travel and is the most economical company in the space tourism industry. SpaceX states that launching the Falcon 9 rocket costs 62 million US dollars which is far cheaper than its competitors which can cost 165 million dollars or more. The large cost reductions are because SpaceX can reuse the first stage of the rocket in successive launches. The goal of this project is to develop a machine learning model that can accurately predict the landing success rate of the first stage rocket. This is fundamental in calculating the total costs for a competitor entering the space tourism market and being competitive.

- Problems you want to find answers
  - What are the key variables that affect the landing success rate.
  - What machine learning model most accurately predicts the landing success rate.
  - What can be done to further improve the success rate and thereby reduce costs.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from two sources:
    - The SpaceX API: https://api.spacexdata.com/v4/rockets
    - Web scraping Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Removal of unnecessary variableshttps://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Dealing with missing values
  - One Hot encoding to on non-numeric variables for binary classification.

Methodology

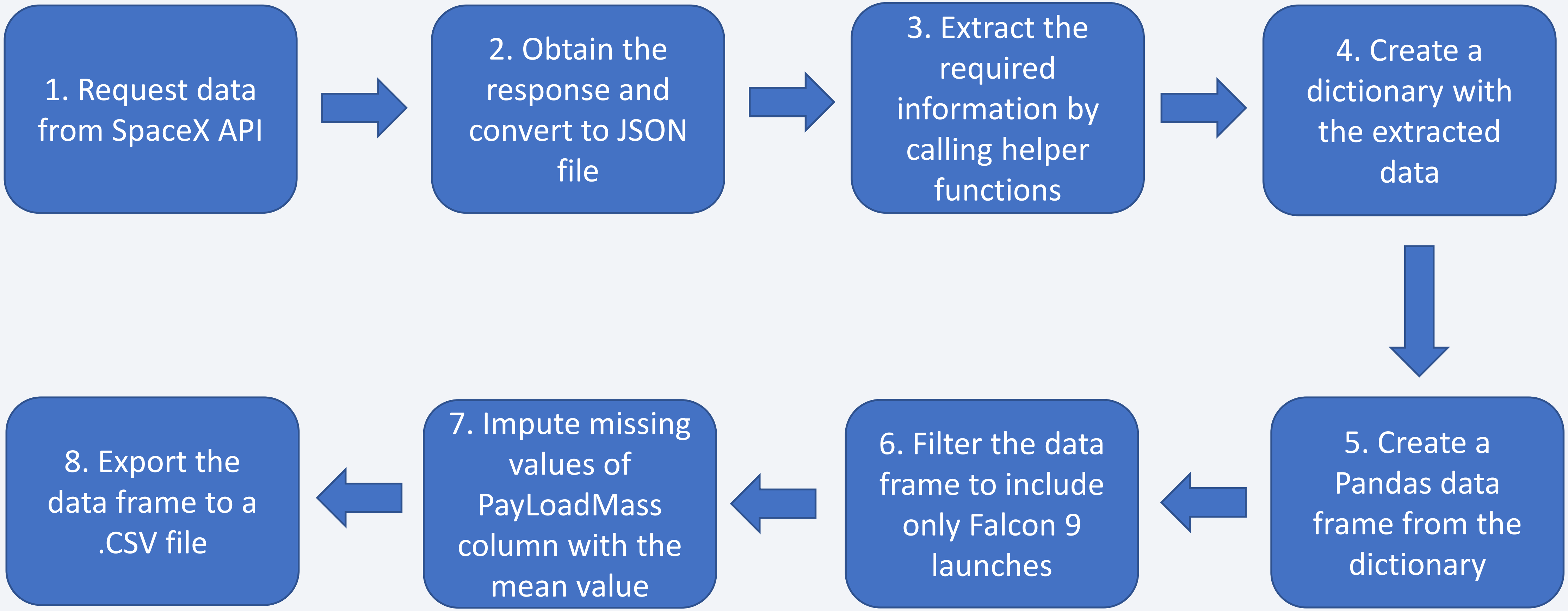
Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was scraped from the SpaceX API and Wikipedia
  - Relevant features were filtered and normalized.
  - The data was split into training and test sets (80/20 ratio).
  - Four classification algorithms were used to predict the 1st stage landing success rate.
  - The models' were evaluated and the best performing model was selected.

Data Collection

- The data required for this project were scraped from two sources:
  - The SpaceX API: https://api.spacexdata.com/v4/rockets
  - Web scraping Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Both sources were needed to obtain all the relevant information required for this project.
- The processes involved are described in more detail on the following pages.

Data Collection – SpaceX API



1. Request data from SpaceX API

2. Obtain the response and convert to JSON file

3. Extract the required information by calling helper functions

4. Create a dictionary with the extracted data

5. Create a Pandas data frame from the dictionary

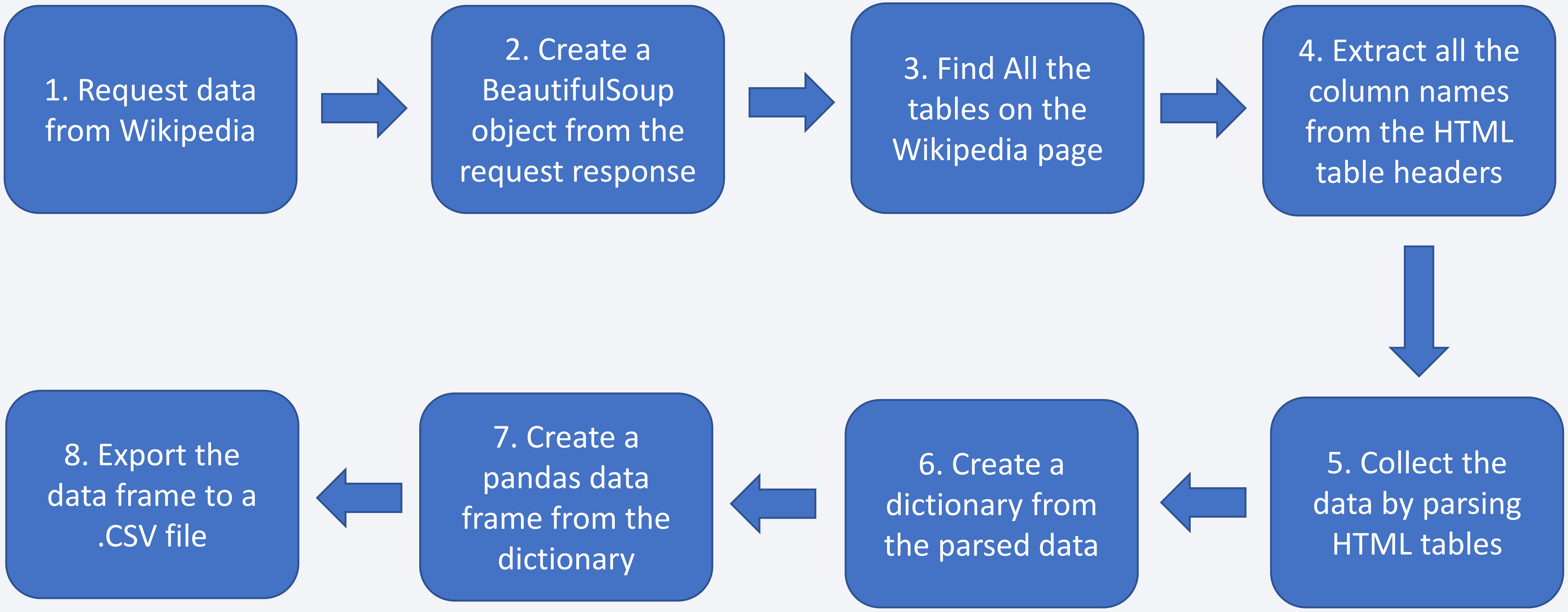
6. Filter the data frame to include only Falcon 9 launches

7. Impute missing values of PayLoadMass column with the mean value

8. Export the data frame to a .CSV file

Data Collection API - GitHub

Data Collection - Scraping



1. Request data from Wikipedia

2. Create a BeautifulSoup object from the request response

3. Find All the tables on the Wikipedia page

4. Extract all the column names from the HTML table headers

5. Collect the data by parsing HTML tables

6. Create a dictionary from the parsed data

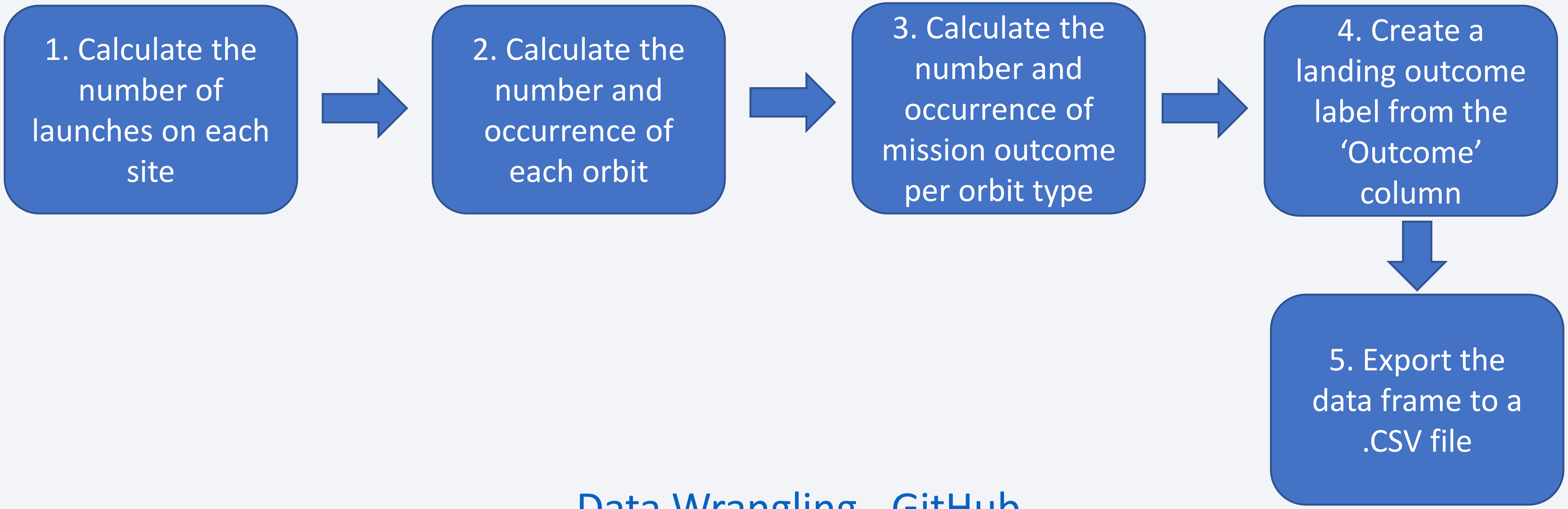
7. Create a pandas data frame from the dictionary

8. Export the data frame to a .CSV file

Data Collection with Web Scraping - GitHub

Data Wrangling

The target variable we want to obtain is whether the mission outcome was a successful landing or not. If the ‘Outcome’ column begins with True then the mission was successful, if it was False then the mission failed. A new target variable called ‘Class’ was created where ‘1’ represents a successful landing and ‘0’ represents a failure.



1. Calculate the number of launches on each site

2. Calculate the number and occurrence of each orbit

3. Calculate the number and occurrence of mission outcome per orbit type

4. Create a landing outcome label from the ‘Outcome’ column

5. Export the data frame to a .CSV file

Data Wrangling - GitHub

EDA with Data Visualization

Scatter Plots – show if there is a correlation between two variables:

1. Payload Mass vs Flight Number: Launch success rate clearly increases with time.
2. Launch Site vs Flight Number: CCAFS SLC 40 has lower success rate than other sites.
3. Payload vs Launch Site: All launches in VAFB SLC 4E < 10,000 kg.
4. Orbit vs Flight Number: Least successful orbit is GTO.
5. Payload vs Orbit Type: Least successful orbit is GTO, most successful is SSO.

Bar Charts – display comparisons of relations between continuous variable and discrete categories:

1. Mean Launch Success per Orbit Type: Displays which orbit types are more or least successful.

Timeline Charts – display how a variable varies over a period of time:

1. Launch Success Rate vs Year: There is a clear increase in launch success rate from 2010 (0%) to 2020 (over 80%).

[EDA with Data Visualization - GitHub]

EDA with SQL

- **The Following SQL queries were performed:**
  - 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 successful landing outcome in ground pad was achieved.
  - 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. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes (Failure or Success) between the date 2010-06-04 and 2017-03-20, in descending order.

EDA with SQL - GitHub

Build an Interactive Map with Folium

The Folium map object was created. This is initially centered on the NASA Johnson Space Center in Houston, TX.

Map Objects Created:

- Red circles, text labels and popup labels for all four SpaceX launch sites. These display where the launch sites are located.
- Clusters of points in close proximity so that the number of information points can be observed in one area when zoomed out.
- Coloured markers at each launch site representing each flight number. The coloured pins represent a successful landing (green) and a failed landing (red). This shows the success rate at each launch site.
- Markers and lines showing the distance from a launch site to nearby locations such as as the ocean or a populated centre. These distances demonstrate if the launch site is sufficiently far from a population or infrastructure, incase a landing strays from its target.

Interactive Visual Analytics with Folium - GitHub

Build a Dashboard with Plotly Dash

The dashboard has four components:

- The dropdown allows the user to select which launch site or all launch sites from which to view the information.
- The pie chart enables the user to view the percentage of successful launches for all sites or the success to failed percentage for individual launch sites.
- The slider can be used to select a range of the payload mass for each launch to be displayed.
- The scatter chart shows the relationship of the launch outcome to payload mass.

[SpaceX Dash App - GitHub]

Predictive Analysis (Classification)

1. Data Preparation

- Load the dataset into a Pandas data frame.
- Select the features to be used in the models.
- Standardize the data.
- Split the data into training and test sets.

2. Model Preparation

- Select the machine learning algorithms.
- Select the hyperparameters for the grid search with CV.
- Fit the model to the training set and obtain the optimum parameter settings.

3. Model Evaluation

- Determine the models' prediction accuracy on the test set.
- Create and evaluate the confusion matrix.
- Fit the model to the training set and obtain the optimum parameter settings.
- If required go back to Step 1 and select new features to be used in the models.

4. Model Evaluation

- Select the model with the highest prediction accuracy and use it in for predicting the landing outcomes.
- As more data becomes available, add this to the dataset and retrain the model.

Machine Learning Prediction - GitHub

Results

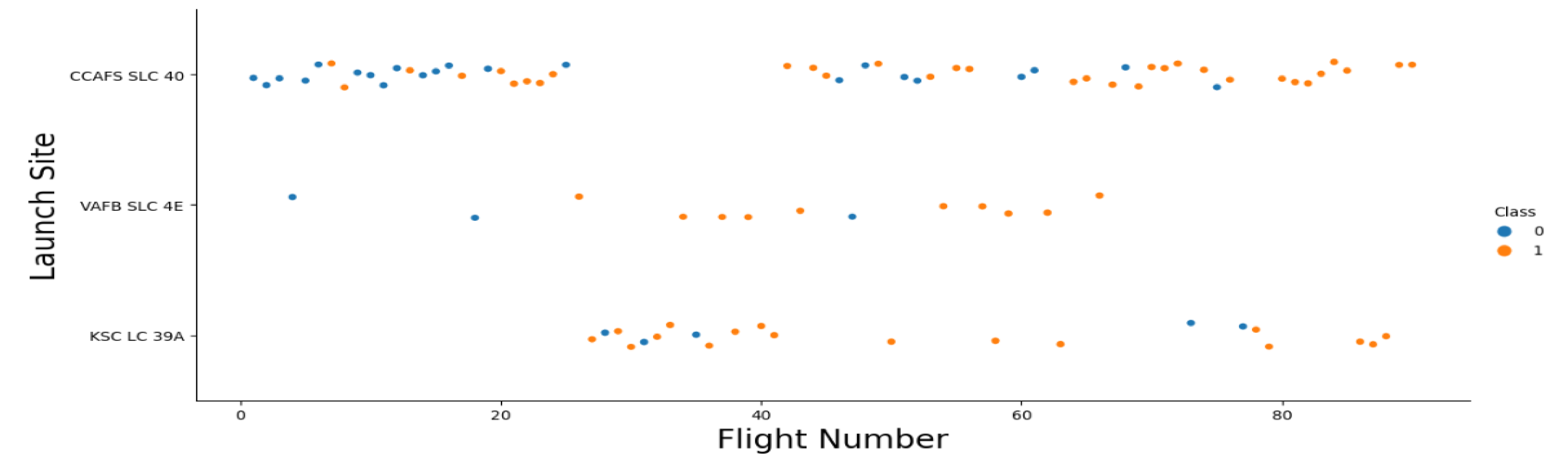
The results are categorized into three main sections:

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

Section 2

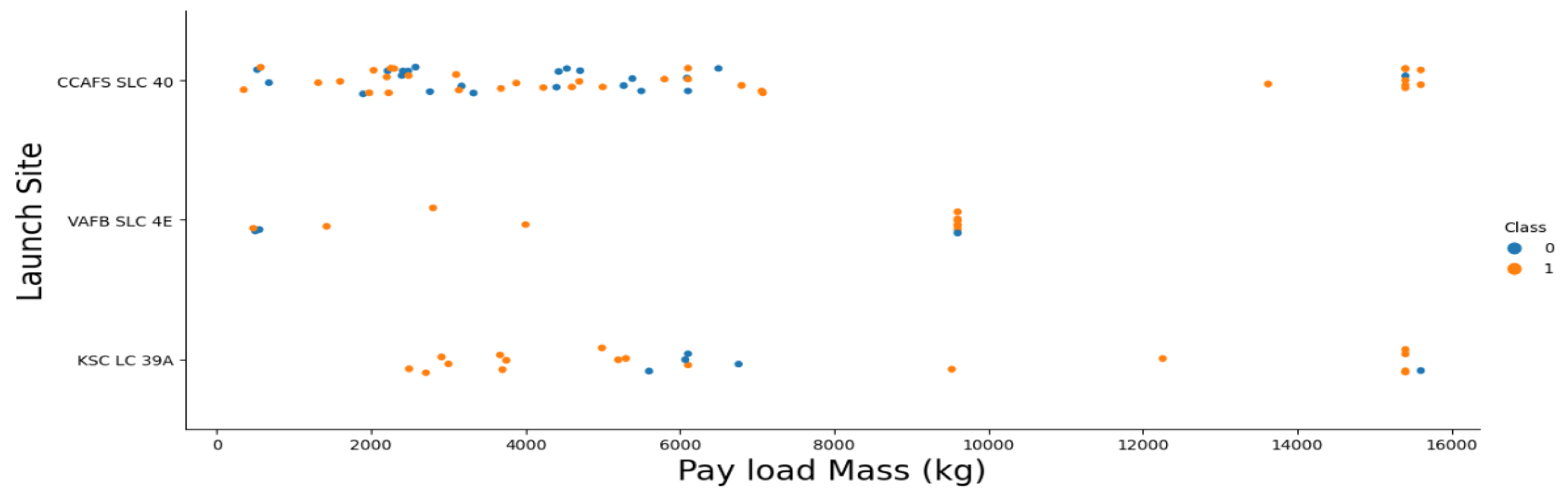
Insights drawn from EDA

Flight Number vs. Launch Site



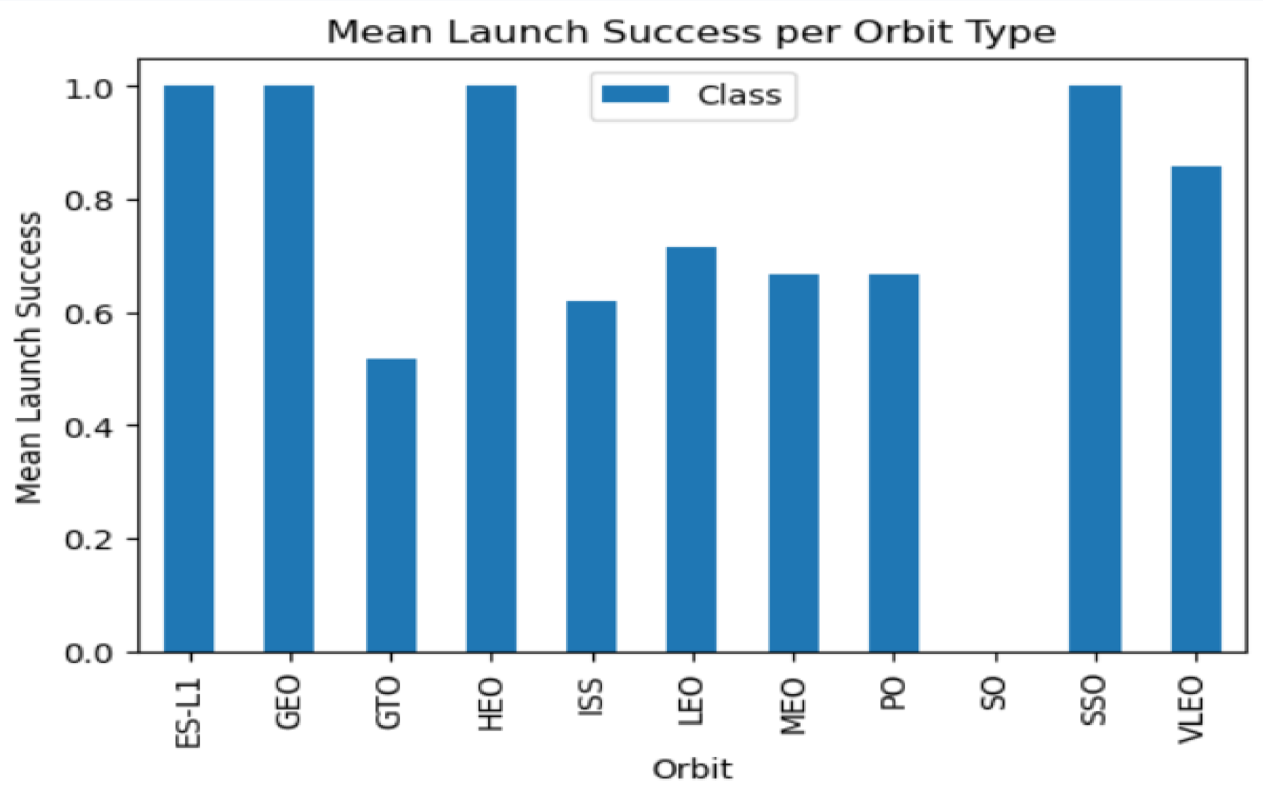
- Launch success rate has clearly improved significantly over time.
- Launch site CCAFS SLC 40 has had the lowest success rate of the three sites, particularly early on.
- Launch site CCAFS SLC 40 is the busiest with approximately 50% of launches.

Payload vs. Launch Site



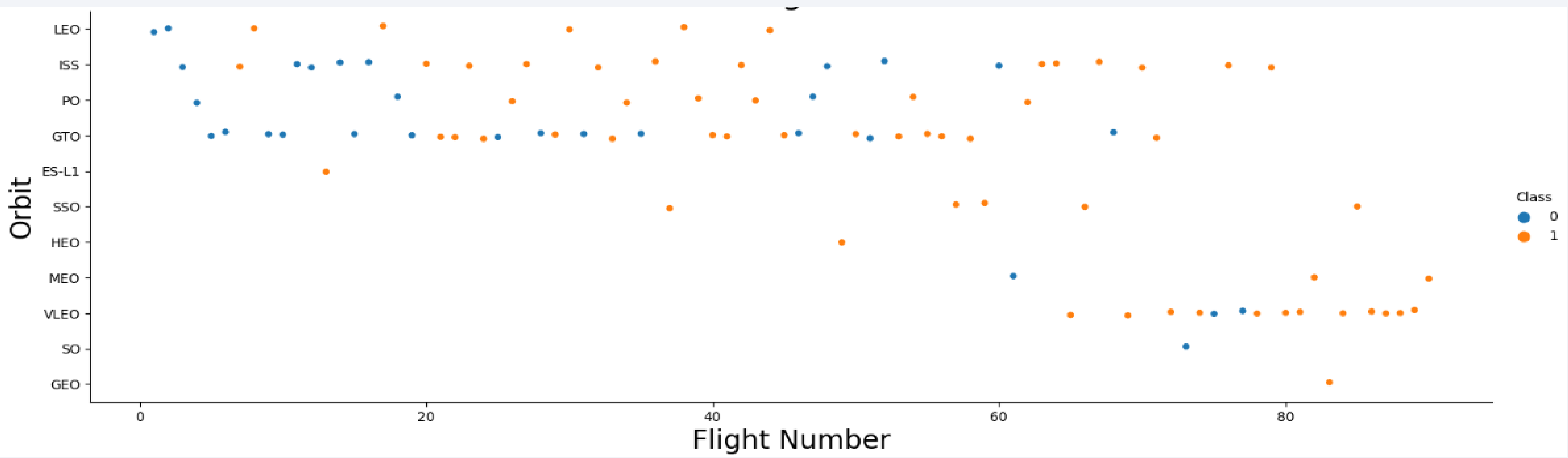
- Heavier payloads (over 8,000 kg) have a higher success rate.
- VAFB SLC 4E has not launched payloads greater than 10,000 kg.
- As noted previously, launch site CCAFS SLC 40 has had the lowest success rate of the three sites.

Success Rate vs. Orbit Type



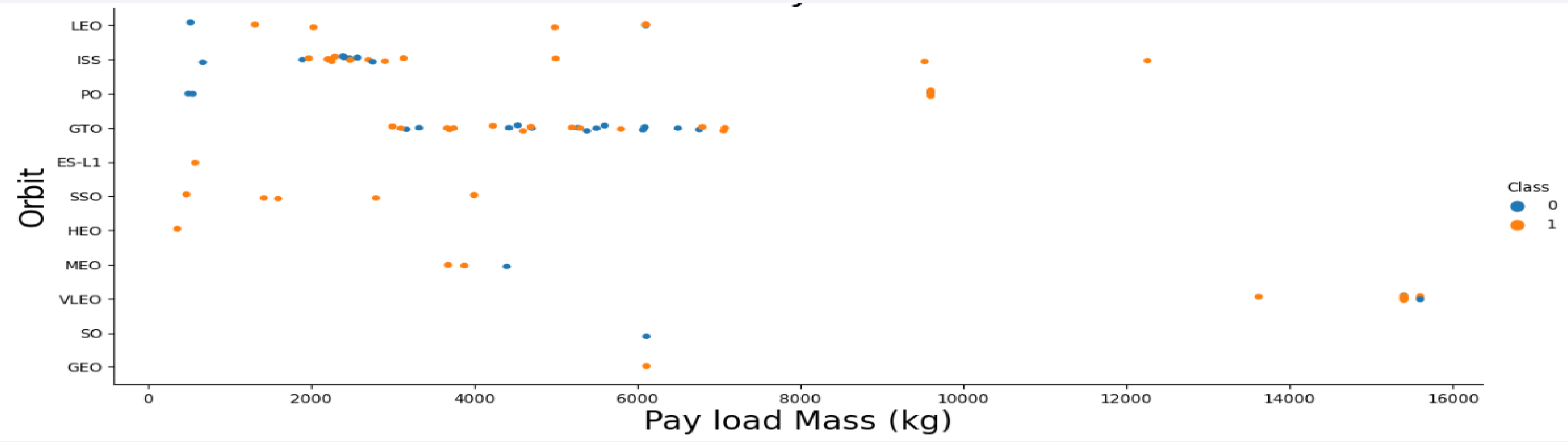
- Orbit types ES-L1, GEO, HEO and SSO are the most successful with a rate of 100%.
- Orbit type SO has a 0% success rate.
- Orbit type GTO is the next least successful with a rate of 50%.

Flight Number vs. Orbit Type



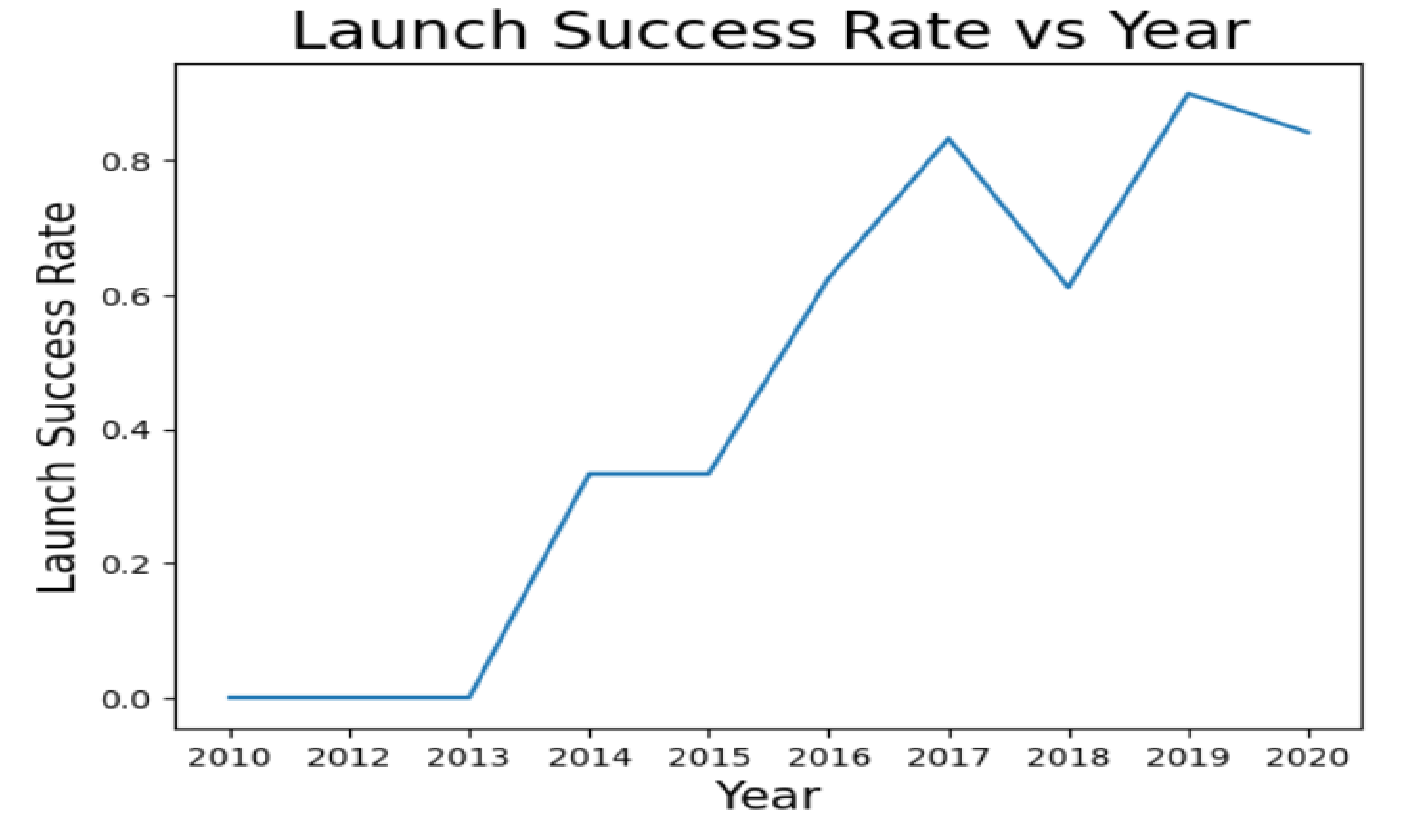
- Again, the success rate of launches increases over time, regardless of the orbit type.
- Orbit type LEO has not been conducted since Flight 43.
- Orbit type VLEO only commenced at flight 65.

Payload vs. Orbit Type



- Again, heavier payloads (over 8,000 kg) have a much higher success rate.
- GTO orbits are the most common with payloads between 2,500 and 7,500 kg.
- Lighter GTO payloads seem to be more successful.
- SO and GEO orbits have only had one attempt each.

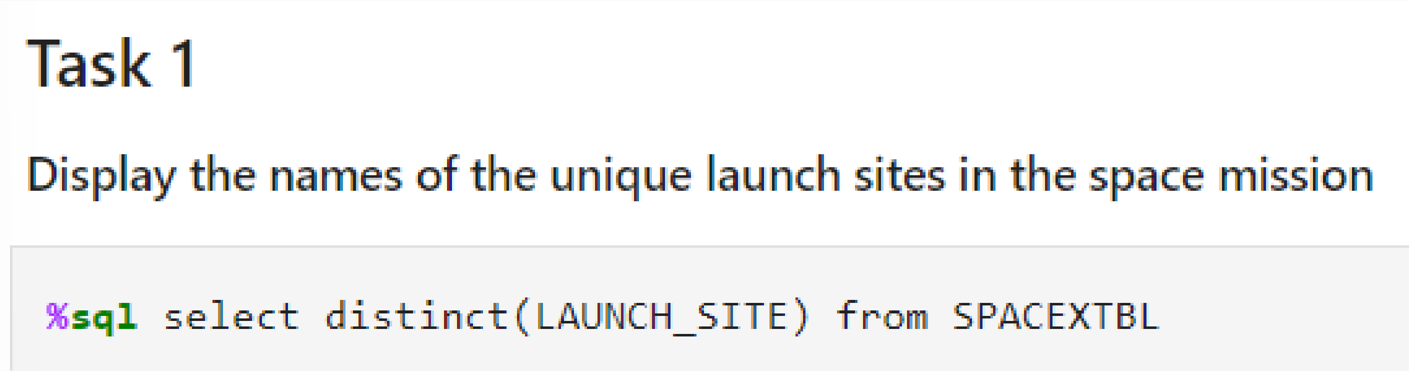
Launch Success Yearly Trend



- The success rate has increased significantly between 2013 (0%) and 2020 (80%).
- Before 2013 the success rate was 0%.

All Launch Site Names

The keyword ‘distinct’ will show all unique names in the table.

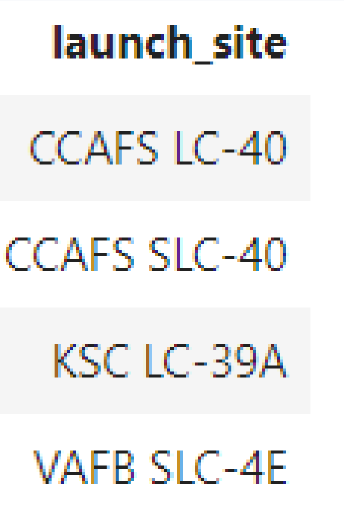


Task 1

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

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

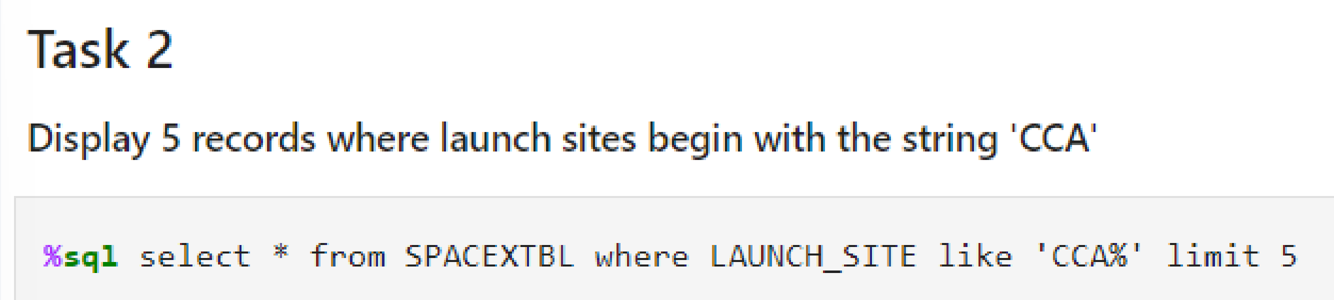
Result:



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

Launch Site Names Begin with 'CCA'

‘limit 5’ will show the 5 first rows where the launch site contains the string ‘CCA’.

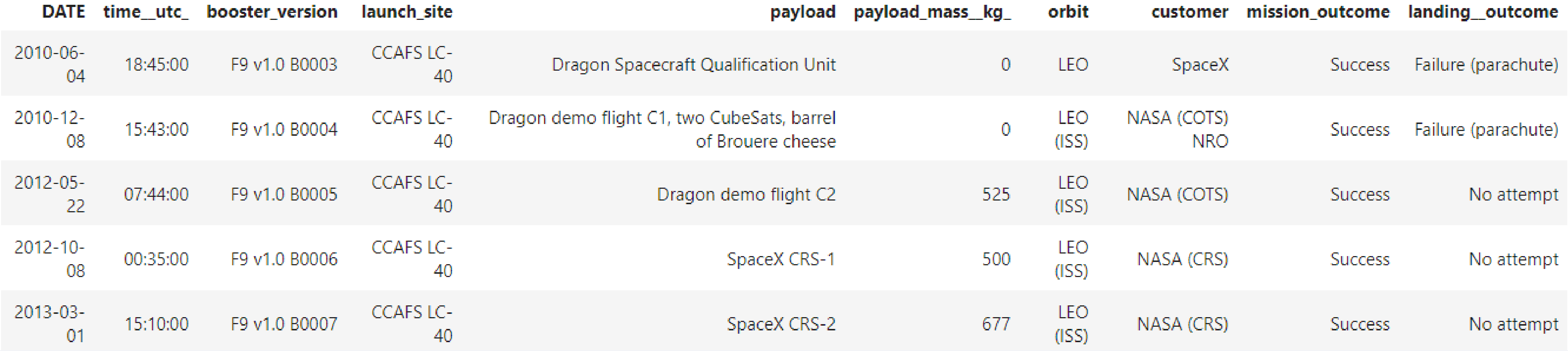


Task 2

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

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

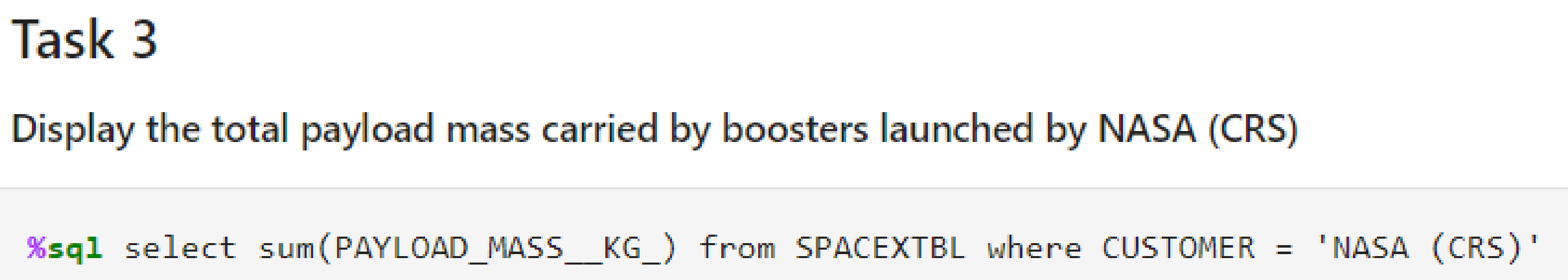
Result:



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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

‘Sum’ will add up the total payload mass from all entries. The total mass is 45,596 kg.

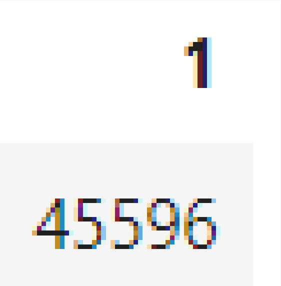


Task 3

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

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

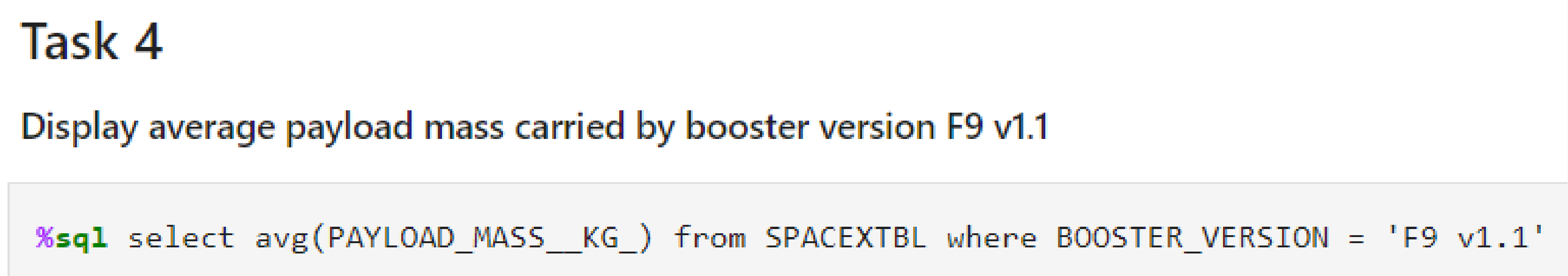
Result:



1
45596

Average Payload Mass by F9 v1.1

‘avg’ will calculate the average. The result is 2928.4 kg.

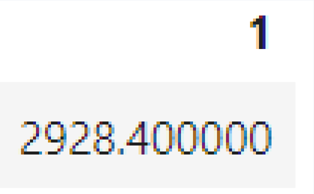


Task 4

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

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

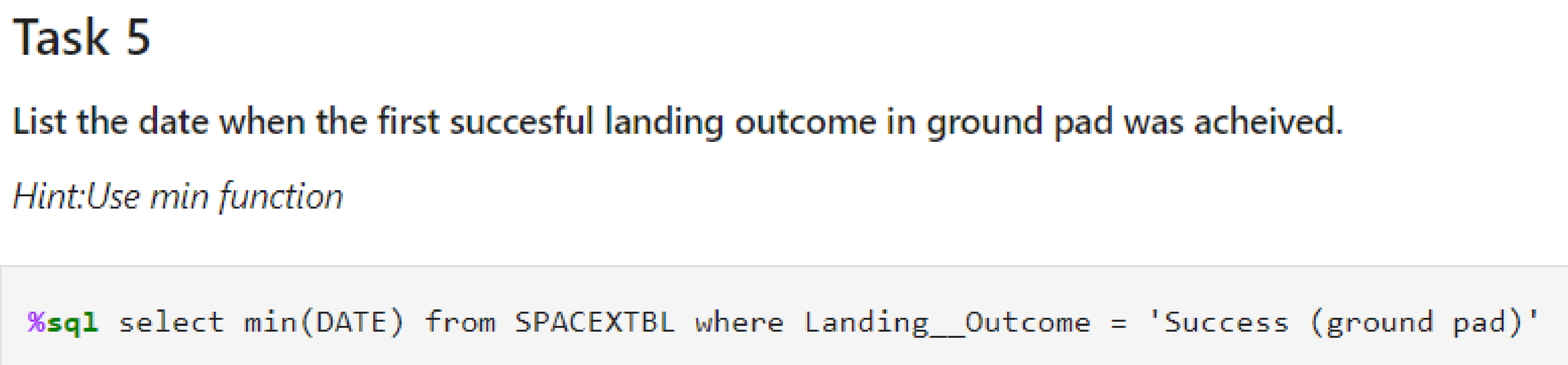
Result:



1
2928.400000

First Successful Ground Landing Date

‘min(DATE)’ will find the earliest date. This date was the 22nd of December, 2015.



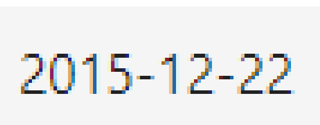
Task 5

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

Hint:Use min function

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

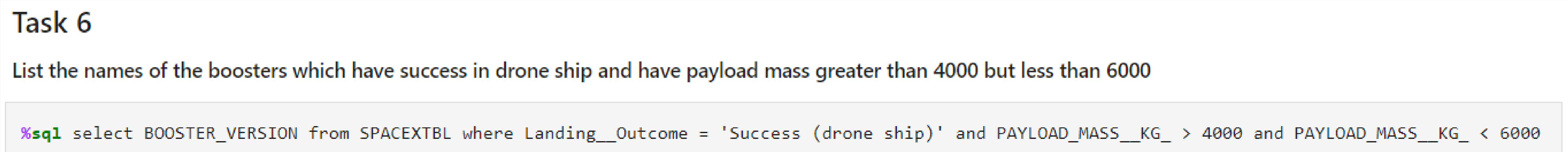
Result:



2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

‘where’ will filter for conditions where ‘Landing_Outcome’ is successful and ‘Payload_Mass’ are between 4,000 and 6,000 kg.

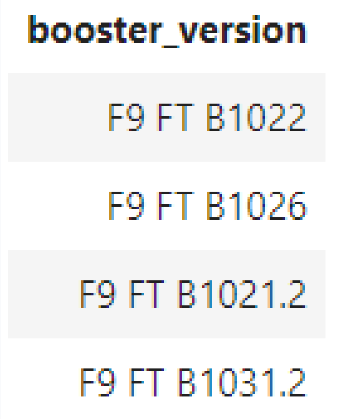


Task 6

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 SPACEXTBL where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

Result:

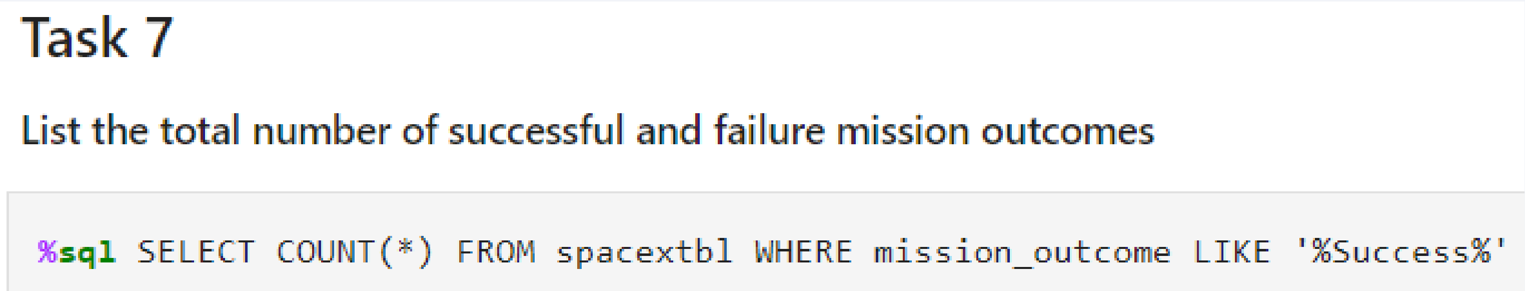


booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The ‘where’ clause will filter for mission_outcome containing the strings ‘Success’ or ‘Failure’.

Count(*) will count all entries where there ‘where’ clause is met.

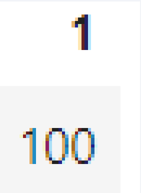


Task 7

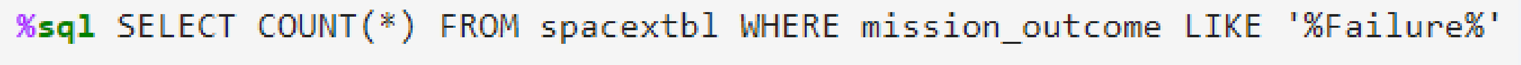
List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Success%'
```

Result:

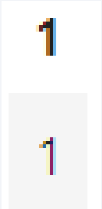


1
100



```
%sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Failure%'
```

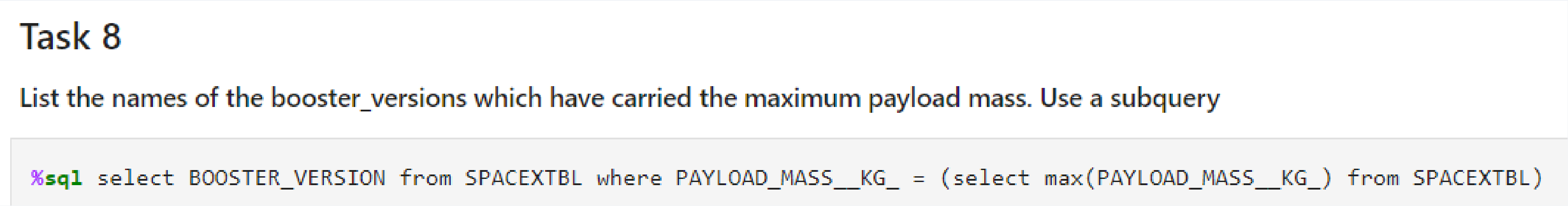
Result:



1
1

Boosters Carried Maximum Payload

The subquery select max(payload_mass_kg_) is required to determine the maximum payload to be used in the where clause.

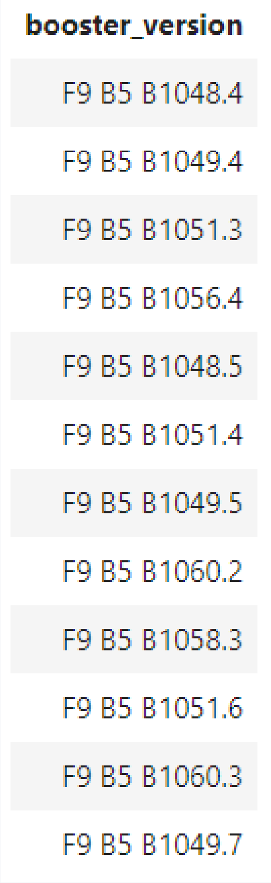


Task 8

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

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

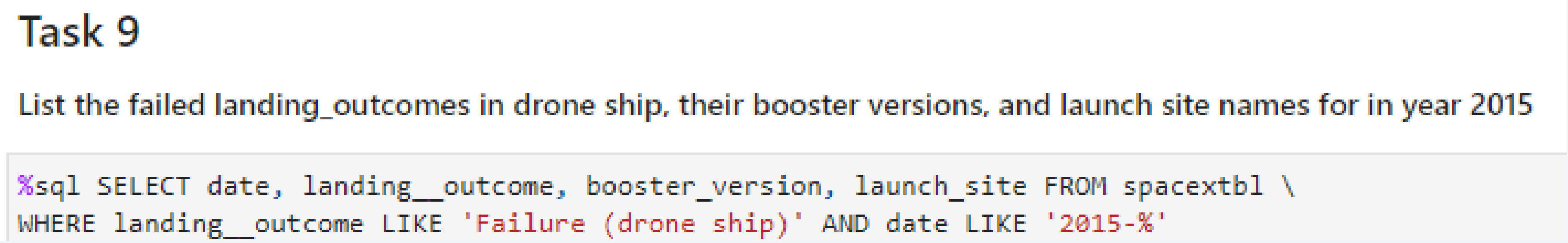
Result:



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

The ‘where’ clause will find all rows of ‘failure (drone ship)’ in the year 2015.

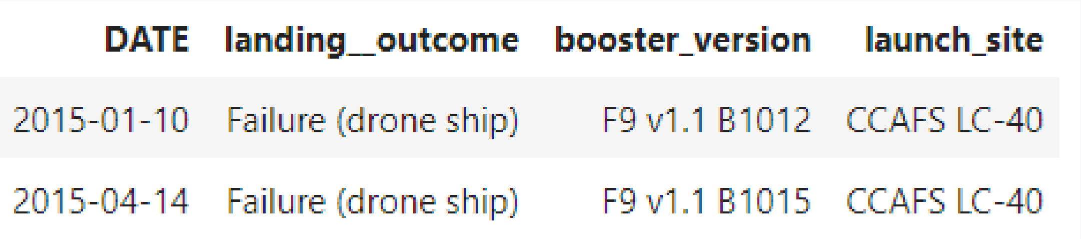


Task 9

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

```
%sql SELECT date, landing__outcome, booster_version, launch_site FROM spacextbl \
WHERE landing__outcome LIKE 'Failure (drone ship)' AND date LIKE '2015-%'
```

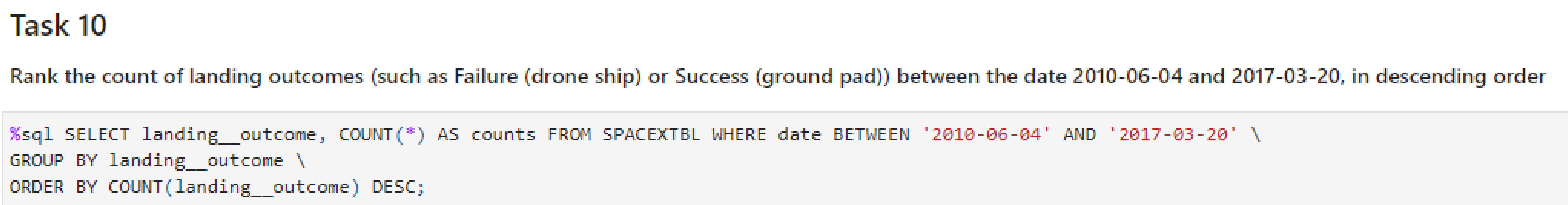
Result:



DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

The ‘landing_outcomes’ and the count of ‘landing_outcomes’ are selected and used with the where clause filter between the desired dates. The results are grouped by the landing outcome and sorted in descending order of the total counts.

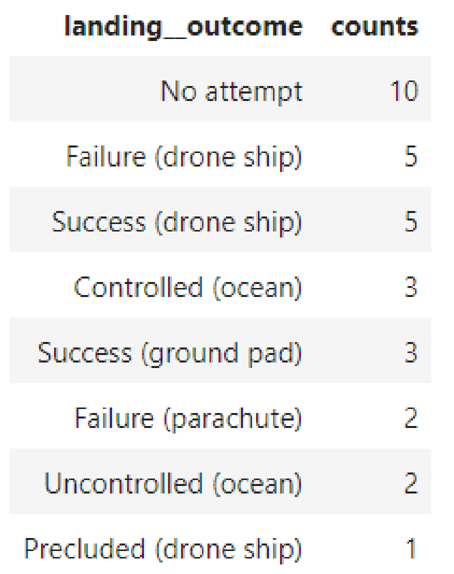


Task 10

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(*) AS counts FROM SPACEXTBL WHERE date BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY landing__outcome \
ORDER BY COUNT(landing__outcome) DESC;
```

Result:

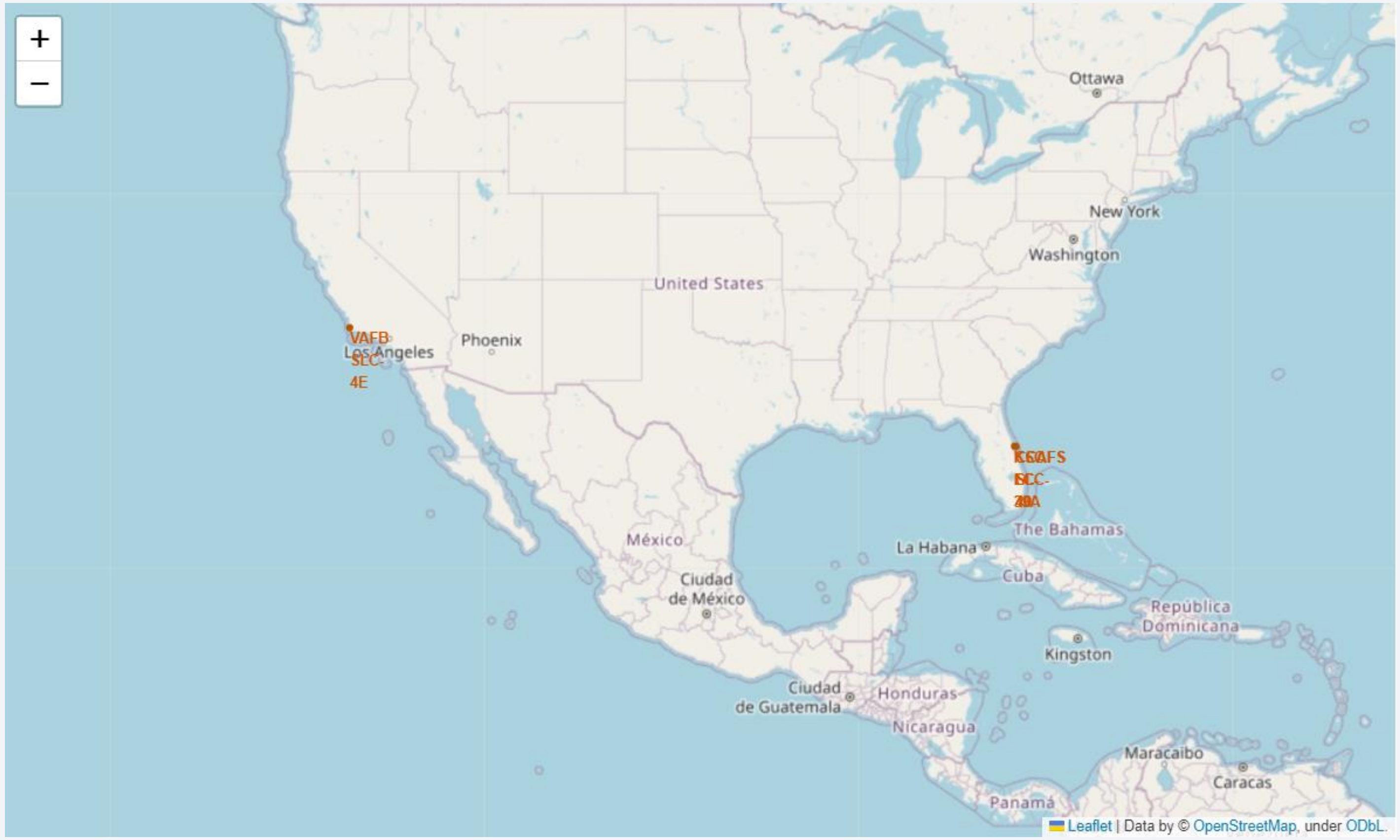


landing__outcome	counts
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

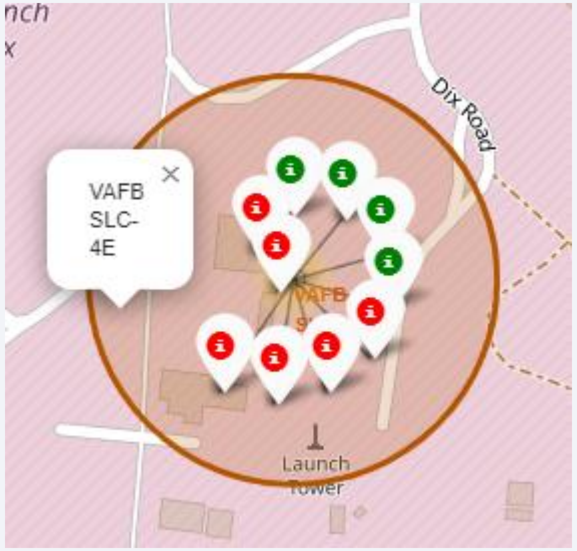
Launch Sites Proximities Analysis

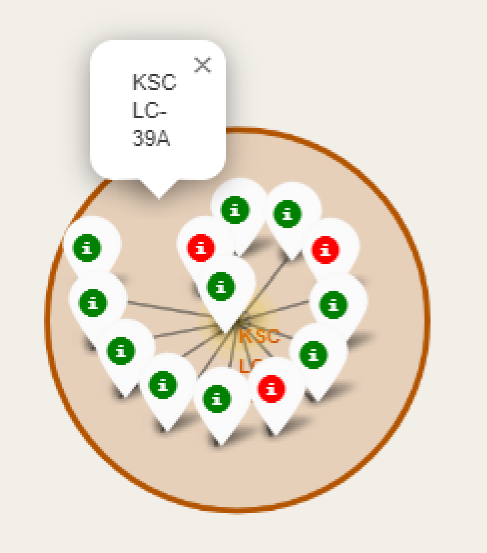
Folium Map – SpaceX Launch Site Locations

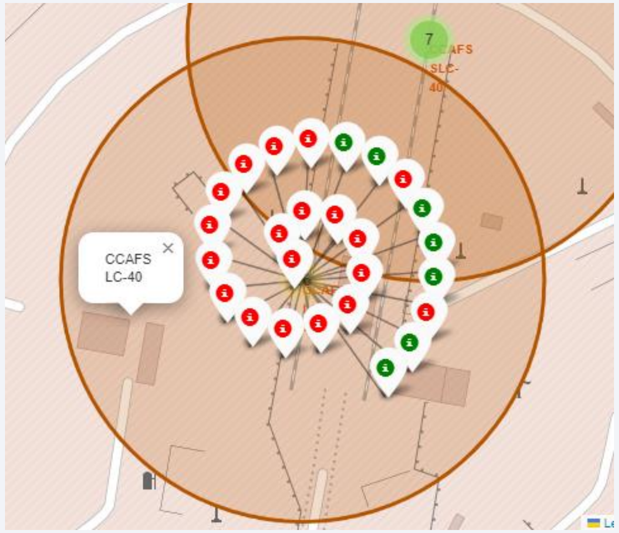


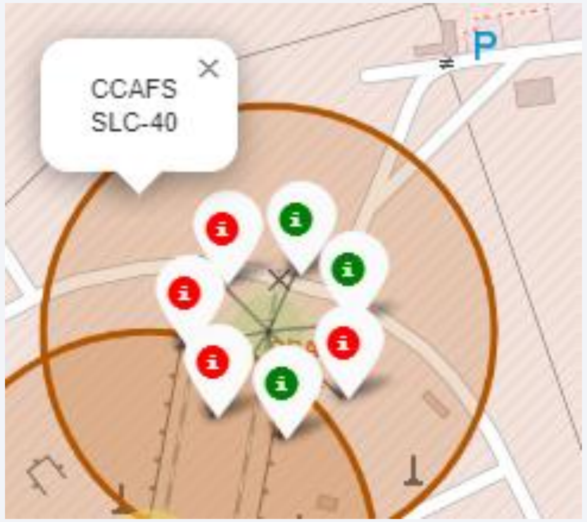
- One launch site is located in southern California close to Los Angeles.
- Three sites are located in Central Florida in very close proximity to each other.
- All launch sites are located in the South of the country (closer to the equator) making launches more fuel efficient.
- All launch sites are near the ocean for safety over populated areas and so that the first stage booster can be recovered from the water.

Launch Outcome Labels per Launch Site



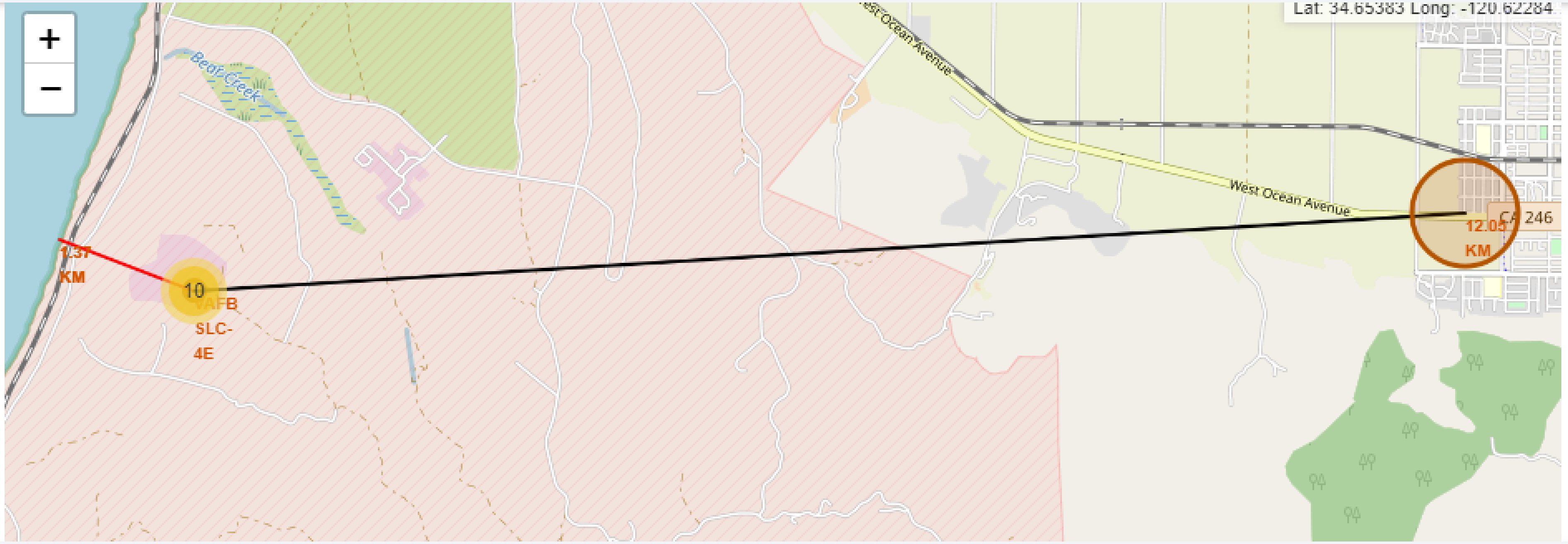






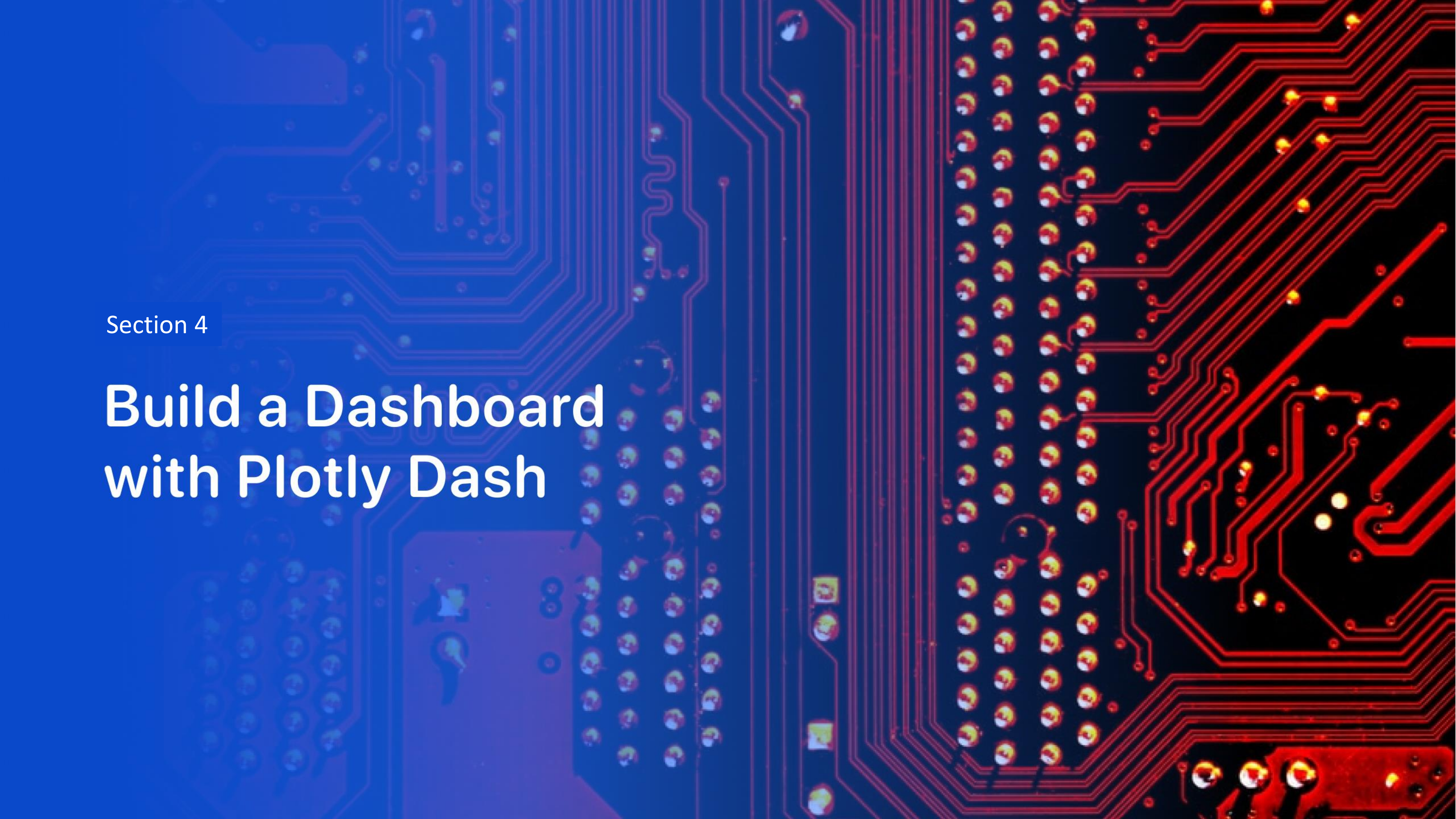
- Green markers indicate successful launch outcome.
- Red markers indicate failed launch outcome.
- KSC LC 39A has a high success rate.
- The three other launch sites have success rates below 50%.

Folium Map – Distances from VAFB SLC-4E to Proximities



Launch Site VAFB SLC-4E is in close proximity to the ocean (1.37 km) where the 1st stage can return land safely.

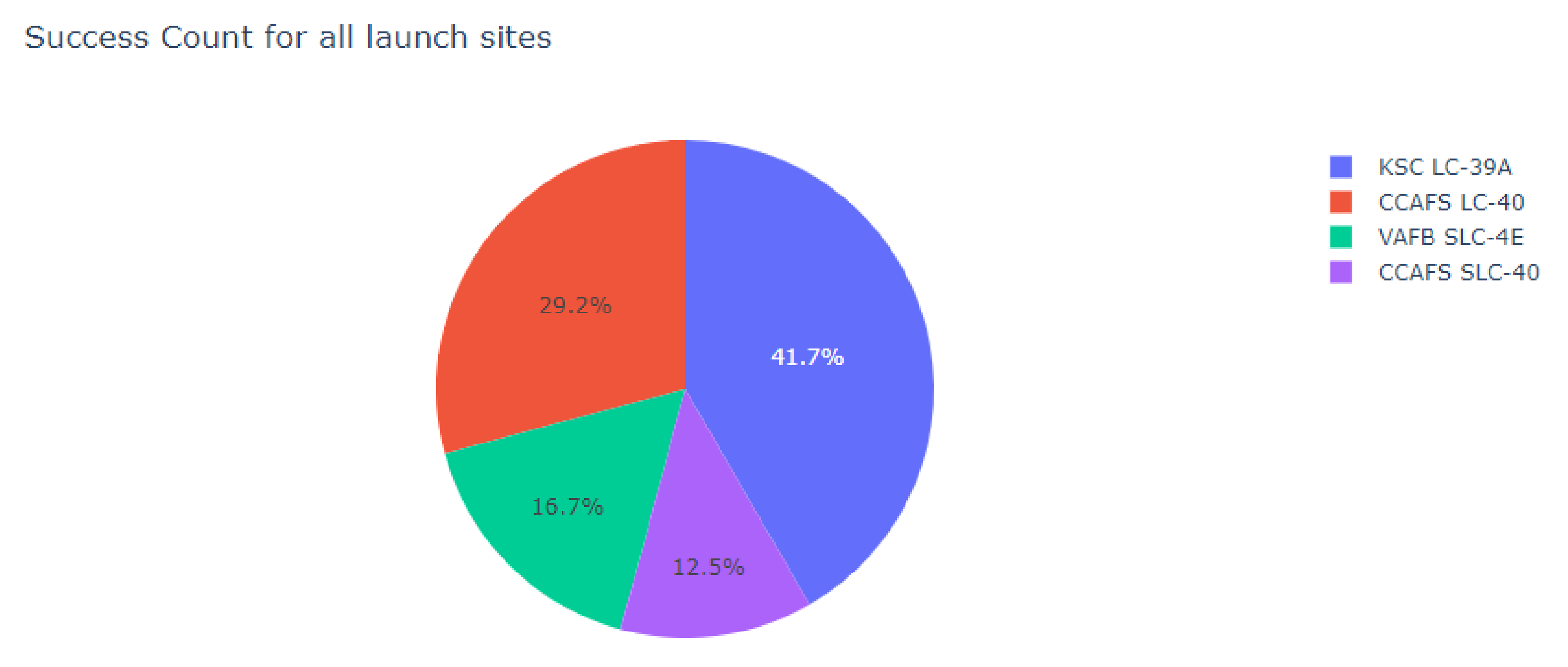
The launch site is a good distance (12.05 km) from the nearest populated area (Lompoc).



Section 4

Build a Dashboard with Plotly Dash

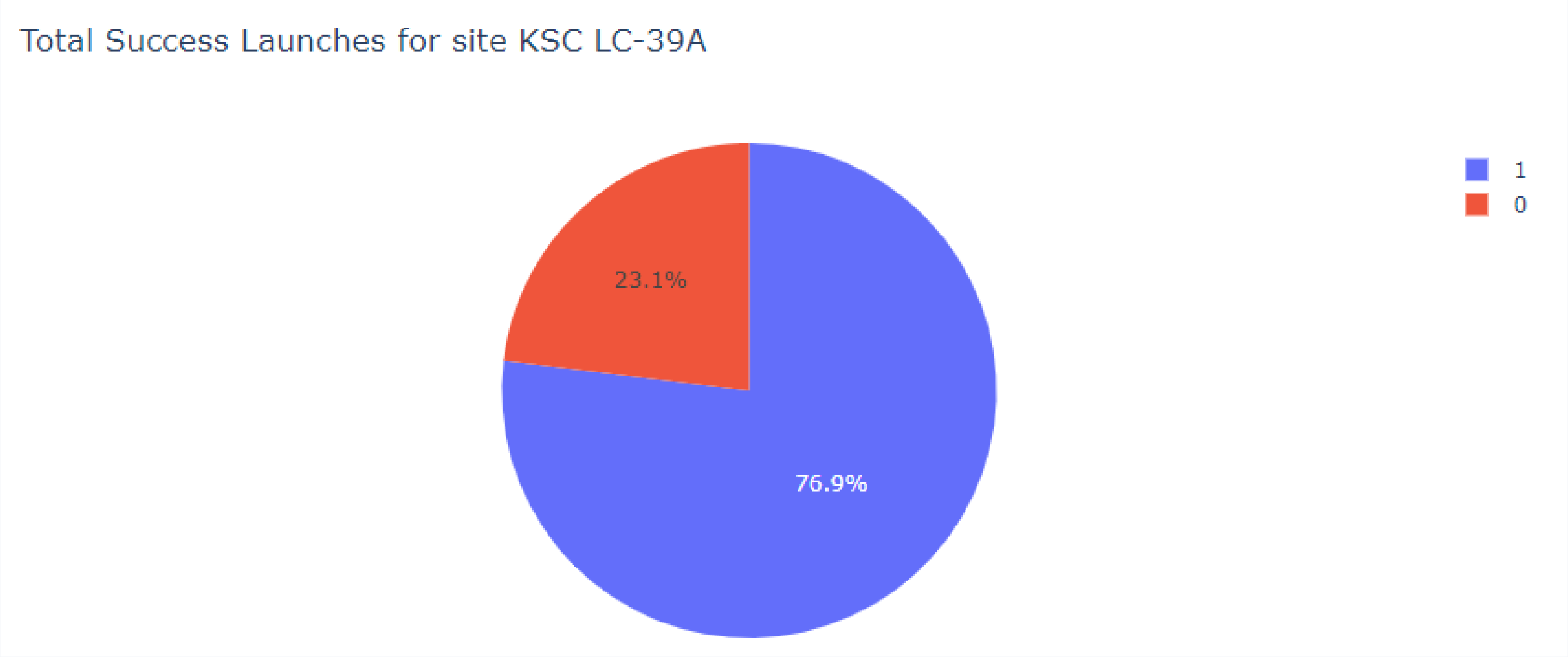
Successful Launches per Launch Site



This chart shows the percentage of successful launches by launch site.

KSC LC-39A has had the highest number of launches.

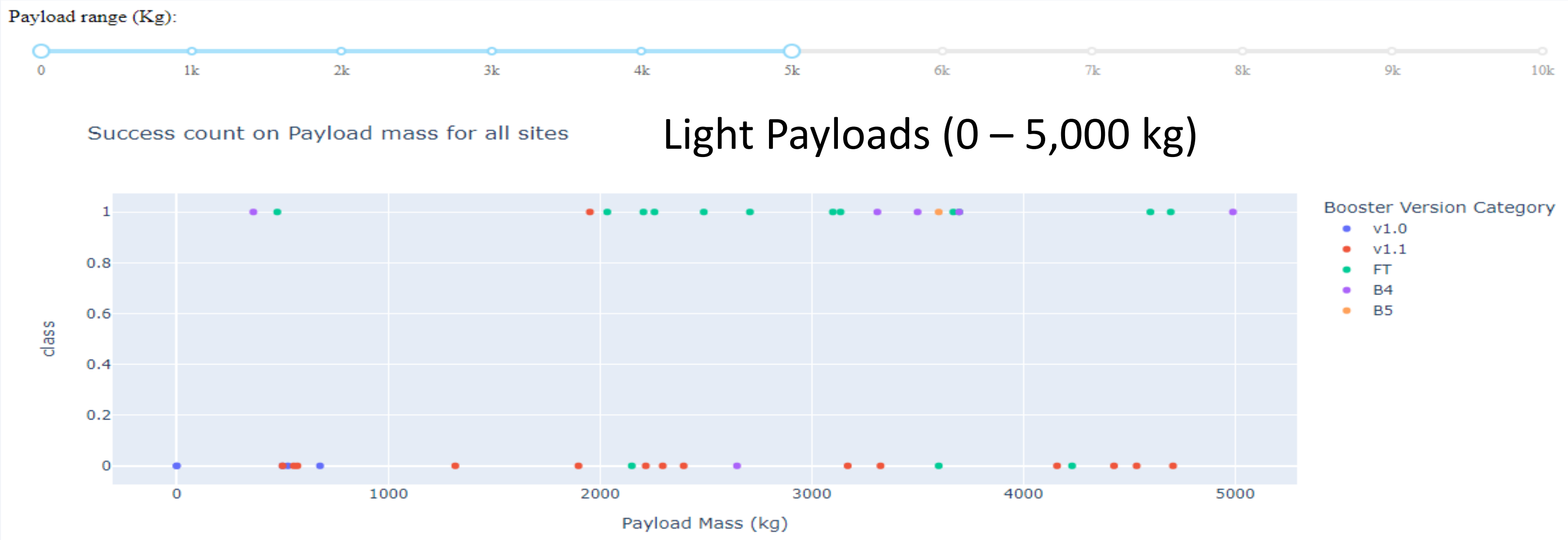
Launch Success Rate for KSC LC-39A



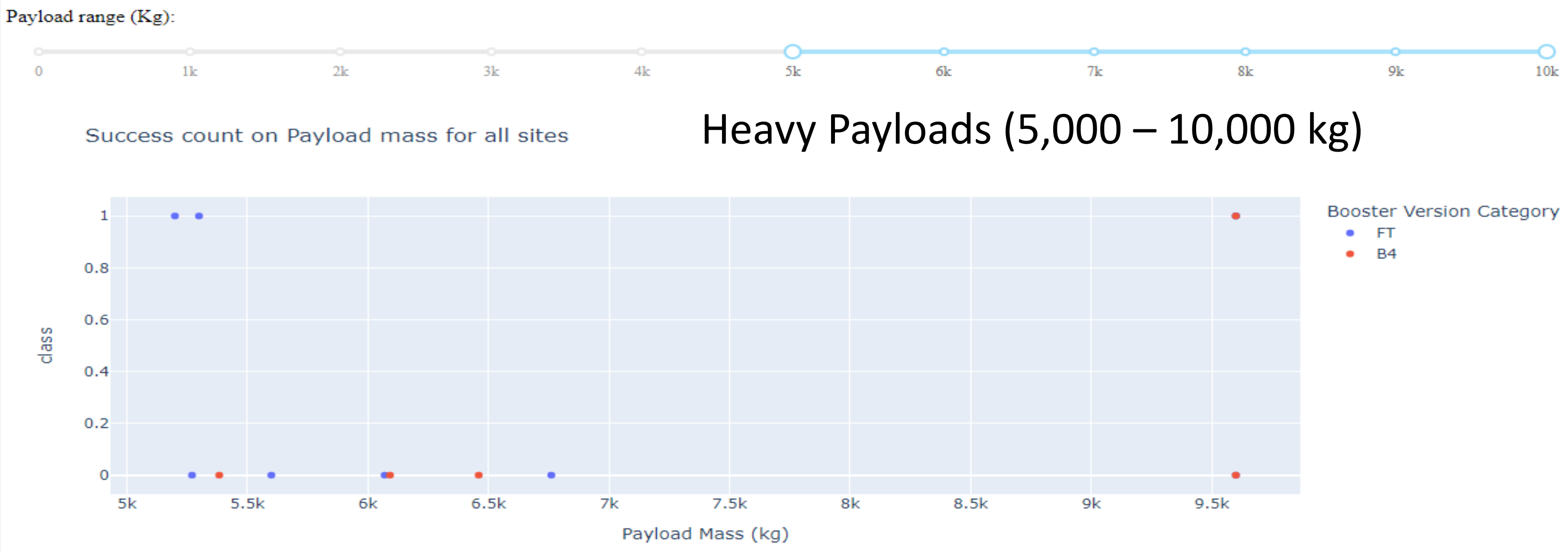
Launch site KSC LC-39A has the highest landing success rate of the four launch sites.

It has had 10 successful landings out of 13 launches, yielding a success rate (blue) of 76.9%.

Landing Success by Payload Mass for All Launch Sites



Light Payloads (0 – 5,000 kg)



Heavy Payloads (5,000 – 10,000 kg)

There are far more payloads weighing less than 5,000 kg than weighing more.

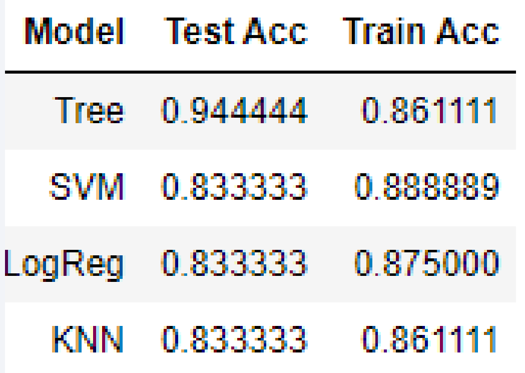
It can be observed that payloads weighing between 2,000 and 6,000 kg have the highest landing success rate.

Section 5

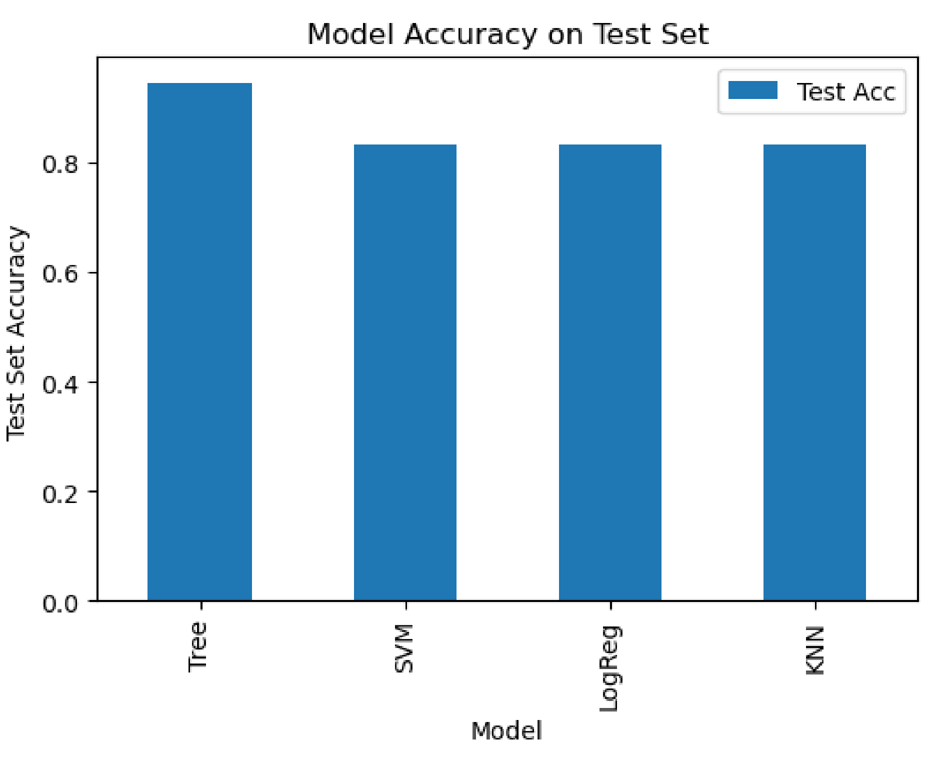
Predictive Analysis (Classification)

Classification Accuracy

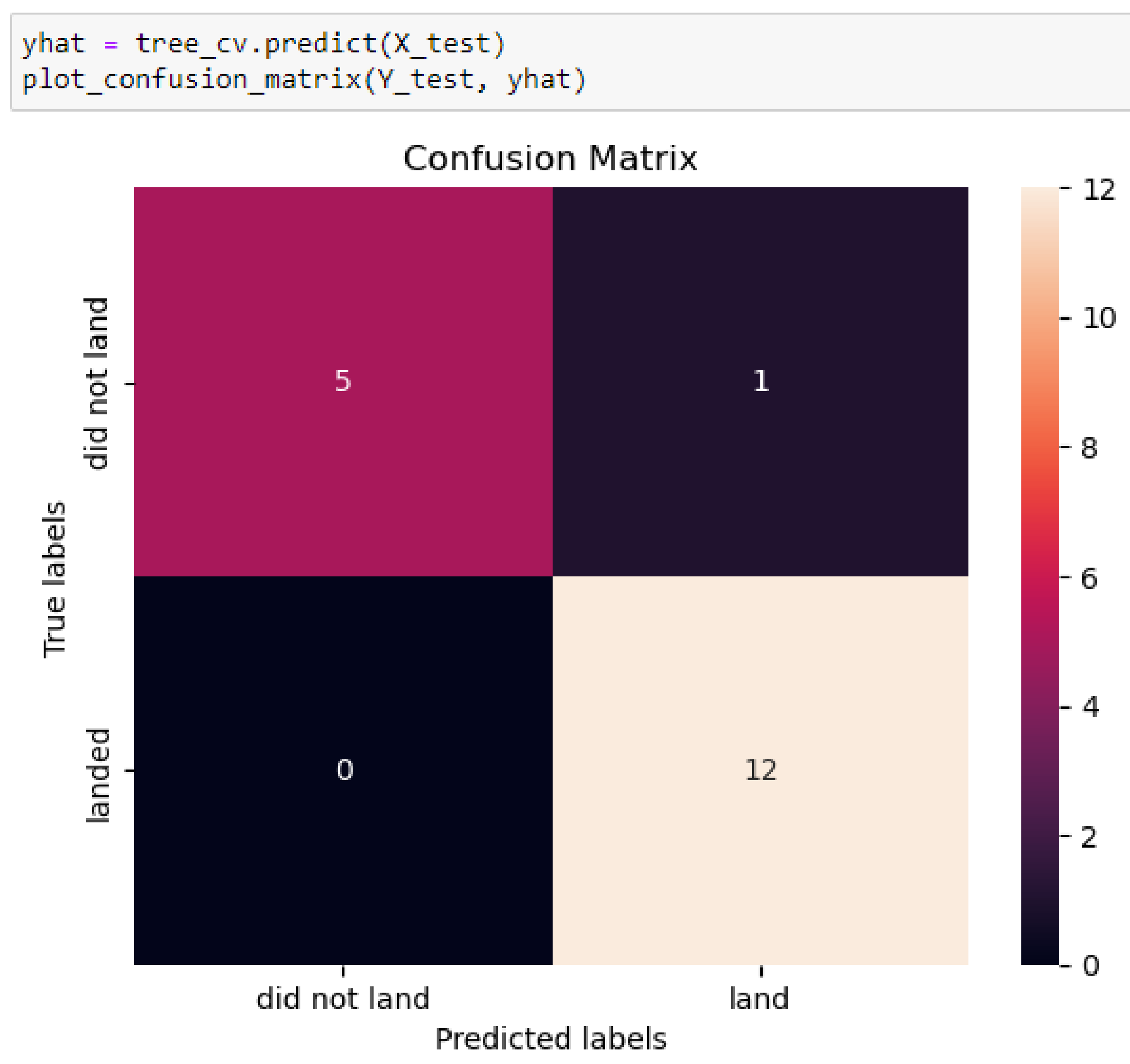
- The bar chart below visualizes the model accuracy for the four classification models. The Decision Tree model is the most accurate on the test set.
- Looking at the table, all four models had a similar prediction accuracy on the training data. The decision tree had the best prediction accuracy on the test set which is a better indicator of the model's performance as it is evaluating data that was not used to train the model.



Model	Test Acc	Train Acc
Tree	0.944444	0.861111
SVM	0.833333	0.888889
LogReg	0.833333	0.875000
KNN	0.833333	0.861111



Confusion Matrix



```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test, yhat)
```

	did not land	land
did not land	5	1
landed	0	12

- As noted in the previous slide, the decision tree algorithm is the most accurate with an accuracy rate of 94.44%. This is only 1 misclassified case out of 18 samples in the test set. The other models all had 3 misclassified predictions resulting in an accuracy of 83.33% in the test set.
- The misclassified prediction is a false positive where the predicted label is a successful landing but the True Label is actually a failed landing.

Conclusions

- There are many variables that can influence the success of a mission. The most important among these were how recent the launch was, the launch site, the payload mass and the orbit type.
- The landing success rate has clearly improved over time, due to the engineers improving the rocket hardware and processes.
- The launch site KSC LC-39A has the highest success rate of the four launch sites. The reason for this could not be obtained from the available dataset.
- Payloads weighing between 2,000 and 6,000 kg have the highest success rate.
- The orbit types with the highest success rates are ES-L1, GEO, HEO and SSO.
- A model using the decision tree classifier proved to be the most successful at predicting the landing outcome on the test dataset.

Acknowledgements

Thank you to IBM and the instructors for putting together and excellent and educational course. These instructors are shown on the link below:

IBM Data Science Professional Certificate Instructors

Thank you!