IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dilan Ruparell
06/03/2023

Outline

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

Executive Summary

Summary of methodologies:

1. Used Web Scrapping along with the SpaceX API to collect data.
2. Data wrangling and EDA (exploratory data analysis) techniques to visualize and create insights including an interactive dashboard.
3. Applied Machine learning techniques to the data.

Summary of all results:

1. Successfully was able to collect the required data.
2. Able to explore the data and visualize it to determine which features best indicate successful launch.
3. Successfully used ML techniques to find the best predictive features.

Introduction

Project background and context:

- SpaceY is a new company looking to enter the commercial space industry and views SpaceX as it's most successful competitor. Using Data analytic techniques, I needed to establish what makes SpaceX so successful and how SpaceY can succeed.

Problems I found answers to:

- The key to SpaceX's success was that they could reuse elements from there launch if it went successfully. This allowed them to keep their costs down, thus being able to predict if a launch would be successful would highly effect costs. It was then important to identify the features that make a successful launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was requested from the SpaceX API
  - Webscraped Falcon 9 launch data from Wikipedia
- Perform data wrangling:
  - Outcomes were classified based on the whether it was a positive or negative outcome with a 1 or 0
- 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 split into a training and testing set. 4 different ML models were trained on the training set then scored on the test set.

Data Collection

- Data was collected from the SpaceX API:

  ttps://api.spacexdata.com/v4/rockets/

- Falcon 9 data was webscrapped from Wikipedia

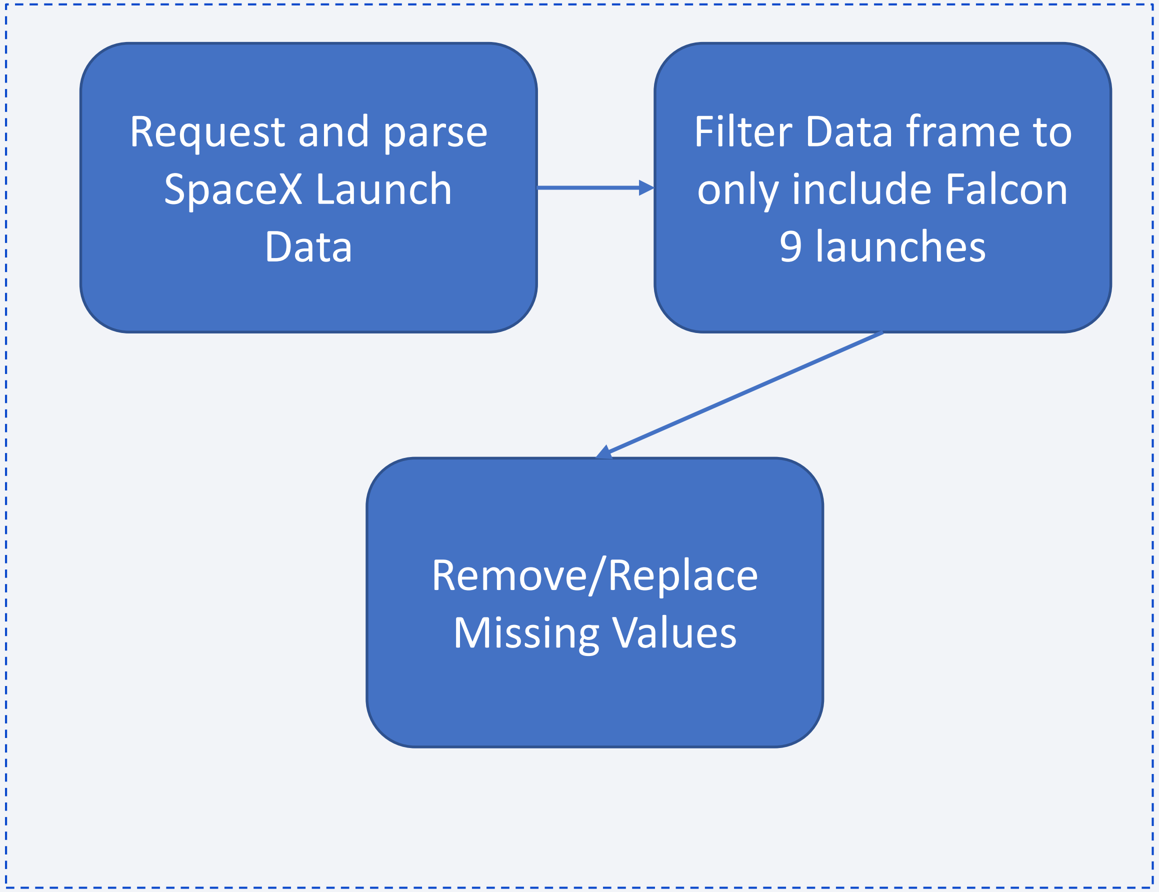
  https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

- The data was collected from SpaceX's API using a get request, then processed as shown in the flowchart.

Github:
https://github.com/DRuparell/ADSCP/blob/master/Data%20Collection%20API%20Lab.ipynb

It does not seem to load on the preview but is visible if you press open with github.dev

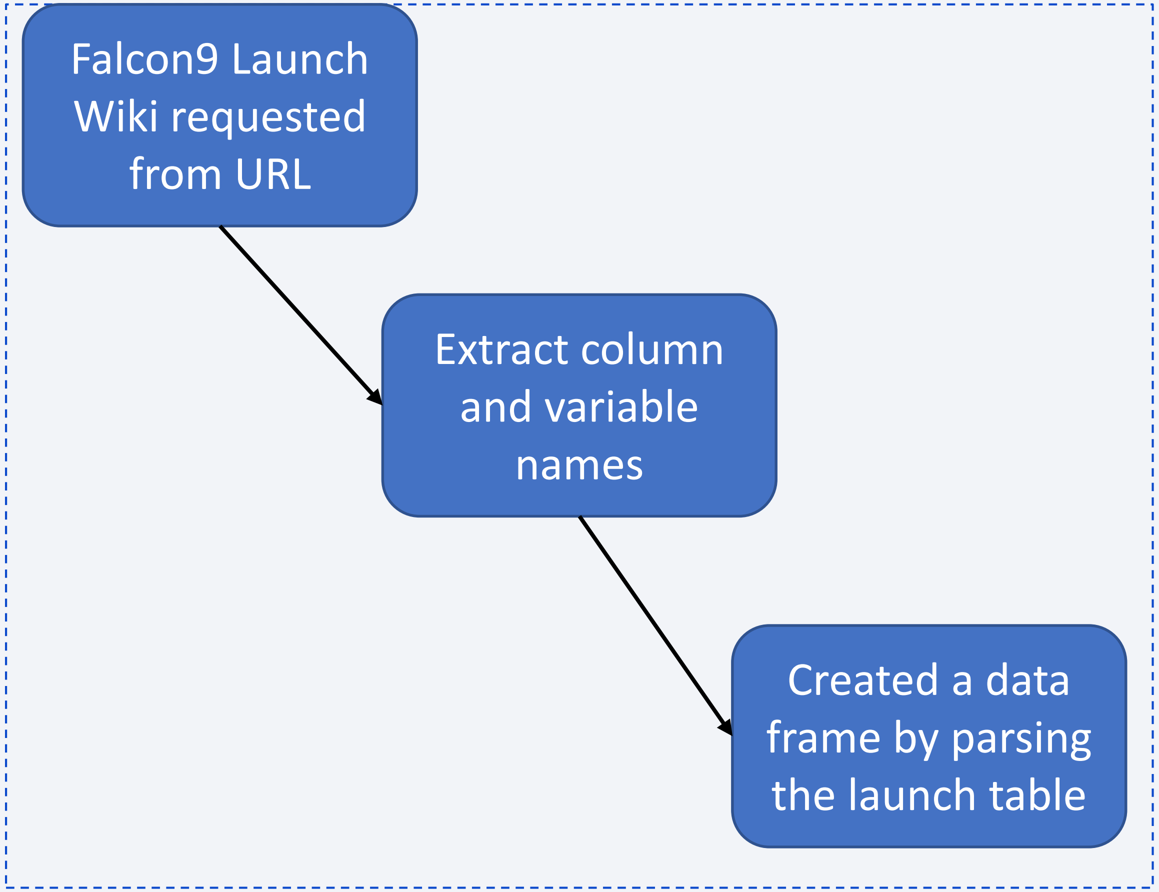

Request and parse SpaceX Launch Data
Filter Data frame to only include Falcon 9 launches
Remove/Replace Missing Values

Data Collection - Scraping

- The Falcon9 SpaceX data was readily available on Wikipedia. Using the web scrapping steps shown in the flow chart this data was collected.

Github: https://github.com/DRuparell/ADSCP/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

It does not seem to load on the preview but is visible if you press open with github.dev



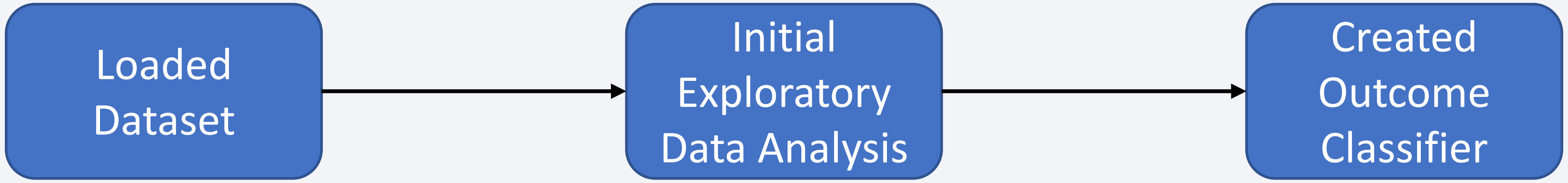
Falcon9 Launch Wiki requested from URL

Extract column and variable names

Created a data frame by parsing the launch table

Data Wrangling

- I explored the data by calculating the number of launches at each site and the number of occurrences of each orbit.
- Created a landing outcome column classifying an outcome as either good or bad.



Github: https://github.com/DRuparell/ADSCP/blob/master/EDA%20Lab.ipynb It does not seem to load on the preview but is visible if you press open with github.dev

EDA with Data Visualization

- To get a preliminary understanding of the relationship between variables I made scatter plots comparing:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Flight Number and Orbit Type
  - Payload and Orbit Type
- Bar charts showing success rate of each orbit type and a line chart showing the yearly success trend were created.

Github: https://github.com/DRuparell/ADSCP/blob/master/EDA%20with%20Data%20Visualisation.ipynb It does not seem to load on the preview but is visible if you press open with github.dev

EDA with SQL

- I looked at all the distinct launch sites and the head of records where the launch site began with ‘CCA’
- Calculated the sum of payload mass carried by boosters launched by NASA as well as the average carried by booster version F9 v1.1
- Found the date of the first successful landing
- Found the name of the boosters with both a payload mass between 4000 and 6000 and have a success in drone ship
- Identified the number of successful and unsuccessful mission outcomes
- Used a subquery to find the booster versions carrying the maximum payload mass
- Listed the records of failed landing outcomes in drone ship for 2015
- Ranked the count of successful landing outcomes between 04/06/2010 – 20/03/2017

Github: https://github.com/DRuparell/ADSCP/blob/master/EDA%20SQL.ipynb It does not seem to load on the preview but is visible if you press open with github.dev

Build an Interactive Map with Folium

- Created a folium map with all the launch sites, identifying the successes and the failures. The map also included lines and distances to the nearest coast, city, railway and highway.
- This allowed me to better understand the choices in these locations and why they are good launch sites

Github: https://github.com/DRuparell/ADSCP/blob/master/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb It does not seem to load on the preview but is visible if you press open with github.dev

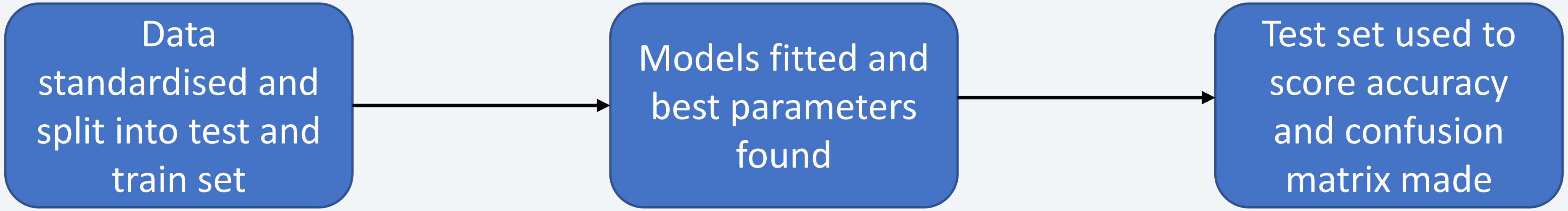
Build a Dashboard with Plotly Dash

- Created a Plotly dashboard with a drop-down box to select launch site, a pie chart defultly showing the success by launch site, when a site is selected it shows the percent of successful to unsuccessful launches. Below the pie chart was a scatter plot of payload mass against the outcomes of the launch with the points colour coded based on booster version.
- The dashboard is a very accessible way for anyone to assess which launch site is the most successful, understand if there is correlation between any of outcome, booster version or payload mass.

Github: https://github.com/DRuparell/ADSCP/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- The data was first standardized and split into a testing and training set. The following 4 models were fitted, and the best possible parameters were found: logistic regression, support vector machine, decision tree and k nearest neighborhoods. Confusion matrixes were then plotted for each along with using the test set to score the accuracy of the model.



Data standardised and split into test and train set

Models fitted and best parameters found

Test set used to score accuracy and confusion matrix made

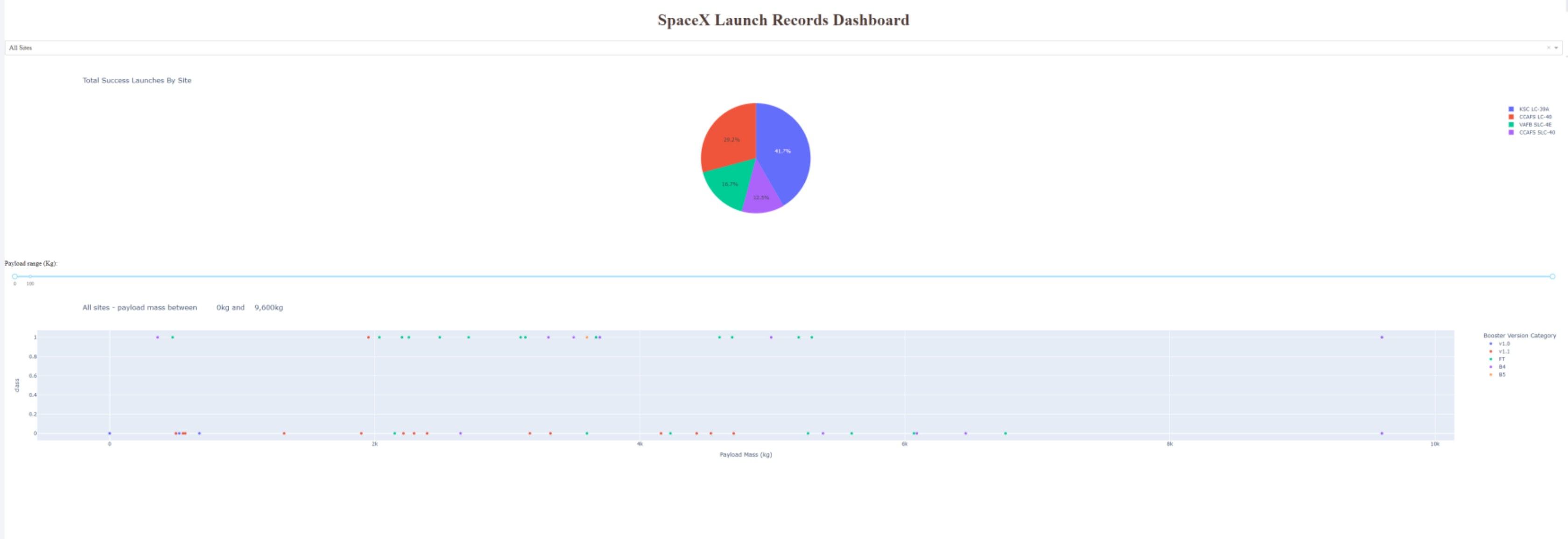
Github: https://github.com/DRuparell/ADSCP/blob/master/Machine%20Learning%20Analysis.ipynb It does not seem to load on the preview but is visible if you press open with github.dev

Results (1)

EDA Results:

- Different launch sites have different success rates
- Heavy payloads were more successful with Polar, LEO and ISS orbits
- As time has progressed on the average launch success has improved
- There were 4 different launch sites
- There was only 1 failure in mission outcomes

Interactive analytics demo:



SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

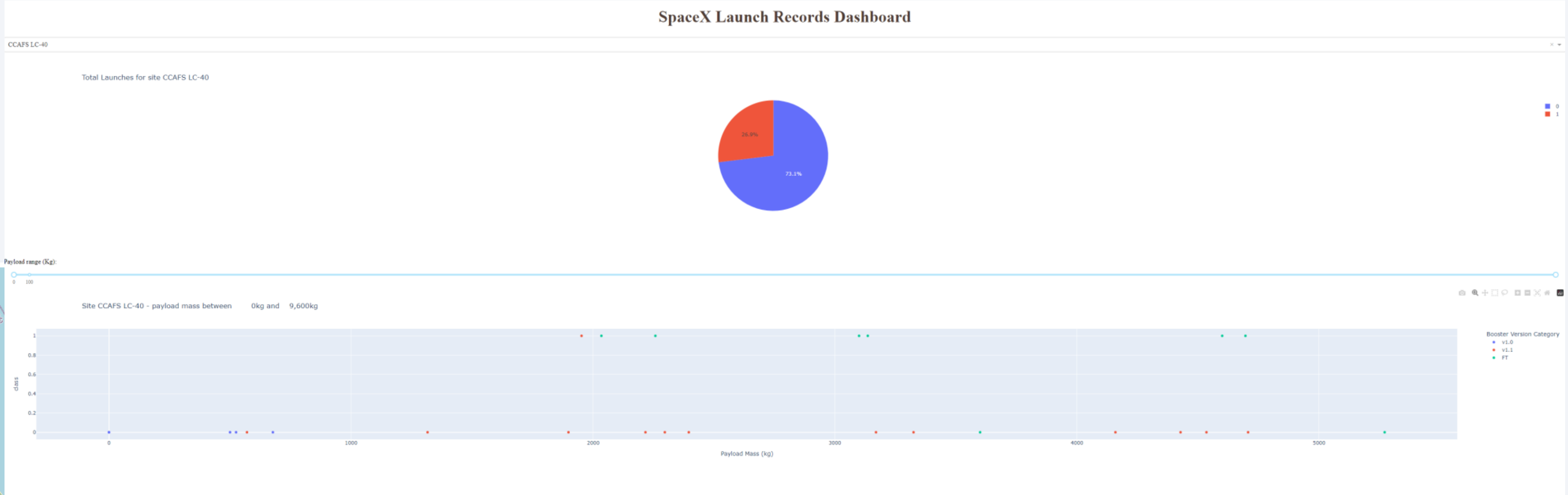
Payload range (Kg):

All sites - payload mass between 0kg and 9,600kg

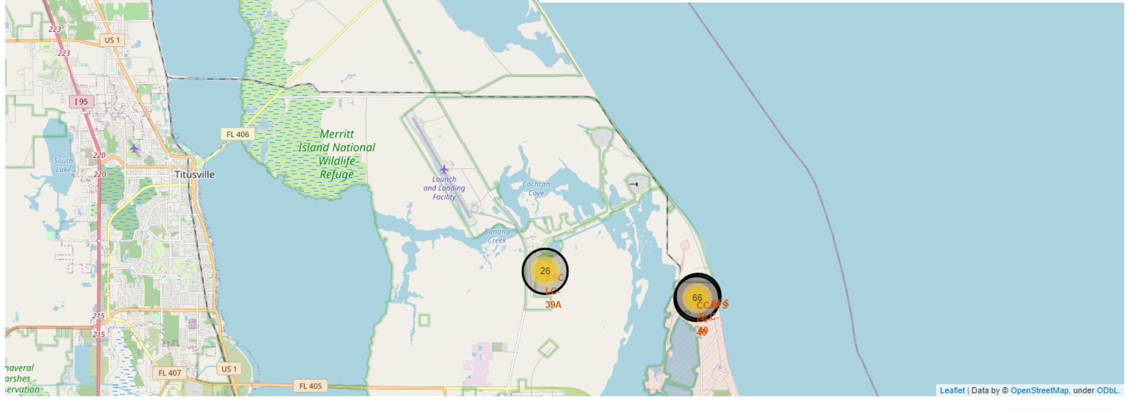
- This is the page the Plotly dashboard opens on

Results (2)

- The plotly dashboard when a launch site is selected



SpaceX Launch Records Dashboard



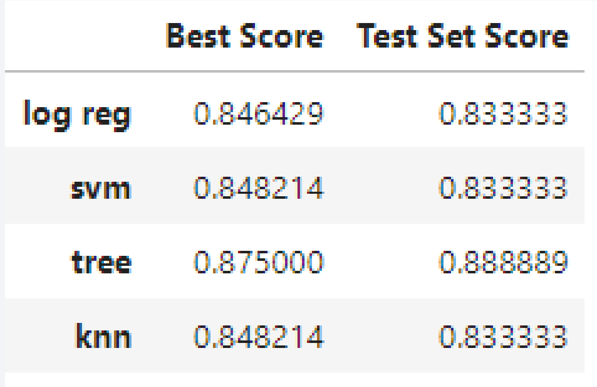
Scottsmoor
US 1
I 95
FL 406
Merritt Island National Wildlife Refuge
Titusville
Launch and Landing Facility
Cochran Cove
Banana Creek
26
39A
66
CCAFS
FL 407
FL 405
Leaflet | Data by © OpenStreetMap, under ODbL.

- A look at the folium interactive map without the distances

Results (3)

Predictive Analysis Results:

- The decision tree method was the best for both Best Score and Test set score.
- The other 3 methods had the same Test set score however KNN and SVM had a better best score.

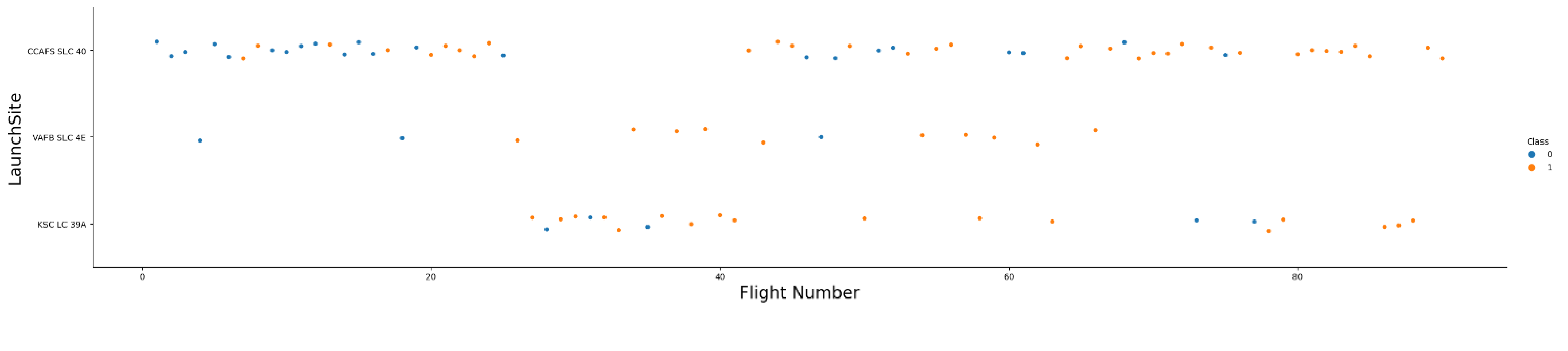


	Best Score	Test Set Score
log reg	0.846429	0.833333
svm	0.848214	0.833333
tree	0.875000	0.888889
knn	0.848214	0.833333

Section 2

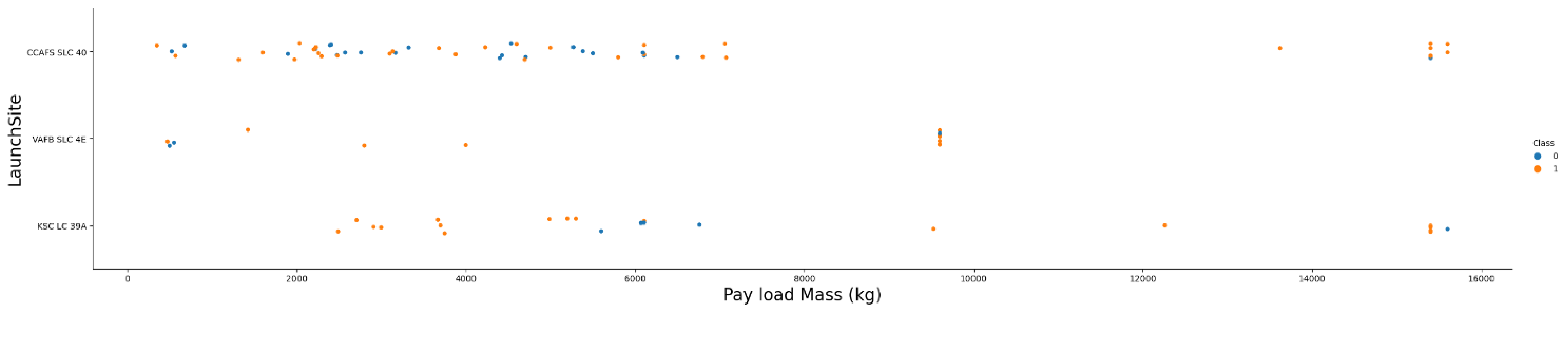
Insights drawn from EDA

Flight Number vs. Launch Site



- For CCAFS SLC 40 as the flight number increases the probability of an outcome being successful increases.
- This trend is not apparent for the other 2 launch sites.
- A lot of the later flights launched from CCAFS SLC 40 which would suggest it is now seen as the best launch site.

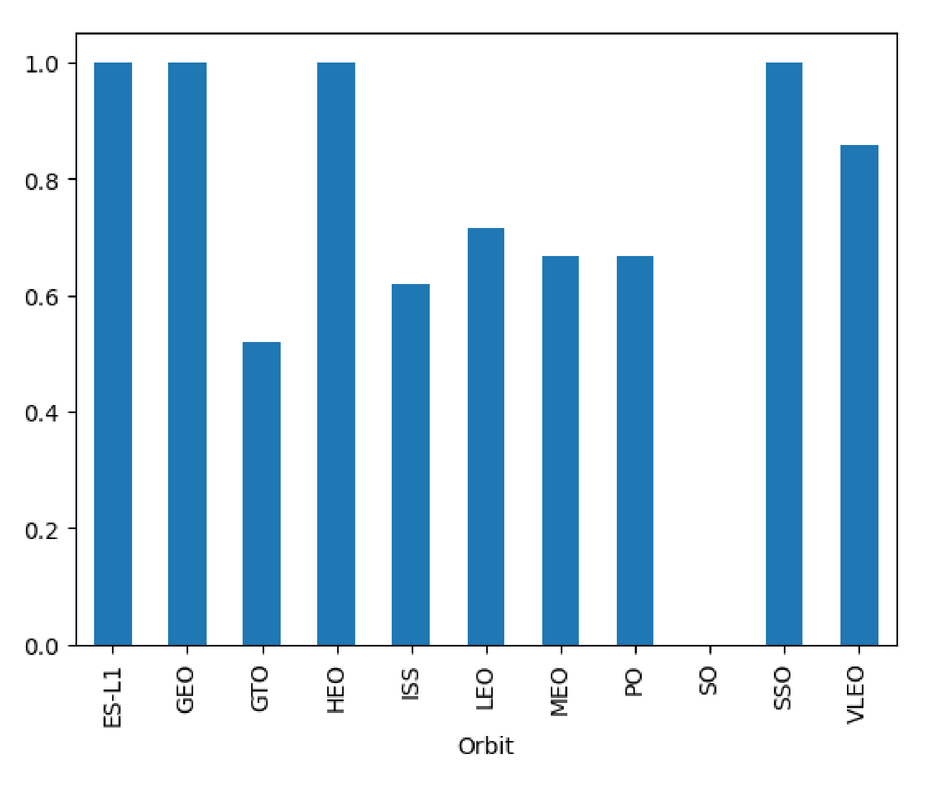
Payload vs. Launch Site



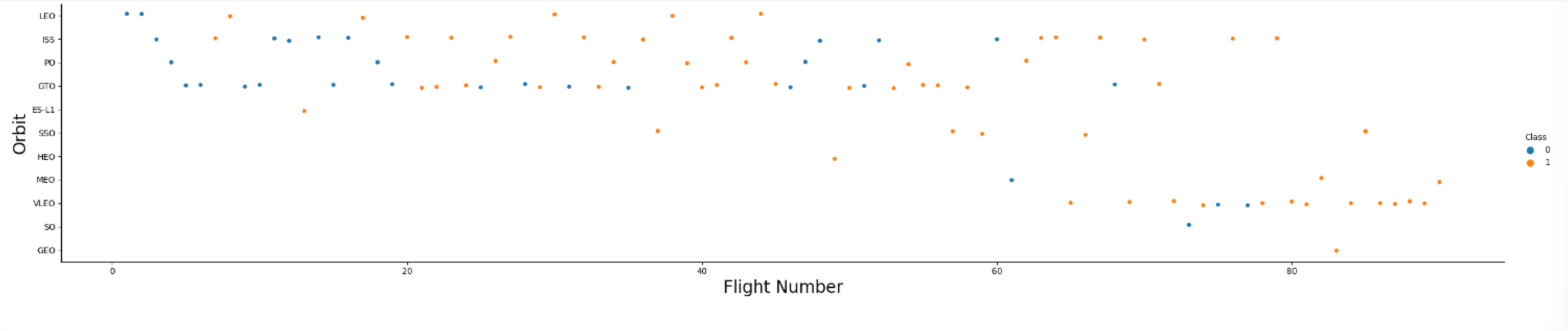
- For CCAFS SLC 40 and KSC LC 39A for payload masses over 8000kg the success rate is very high.
- For VAFB SLC the are no launches above 10,000kg, this could suggest launches are of higher pay loads are not possible from this site.

Success Rate vs. Orbit Type

- The orbits with the highest success rate are ES-L1, GEO, HEO and SSO.
- VLEO performed just below these.
- The worst performing orbits were GTO, ISS, LEO, MEO, PO and SO which was especially bad as it had no successful landings

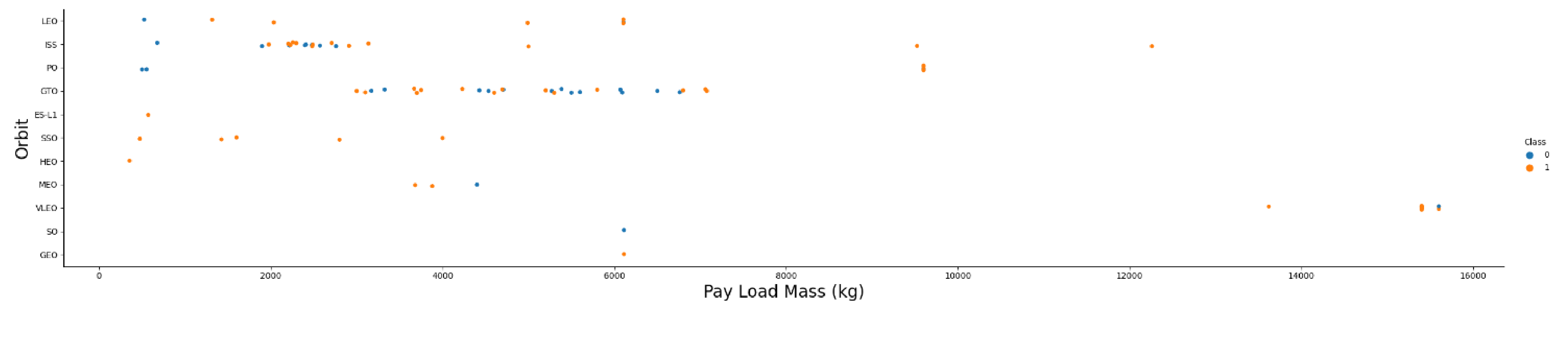


Flight Number vs. Orbit Type



- VLEO was not used until after flight 65 however has been used a lot since with a high success rate. This would indicate a high success rate.
- Success increase as flight number increases with only 5 failures after flight 60

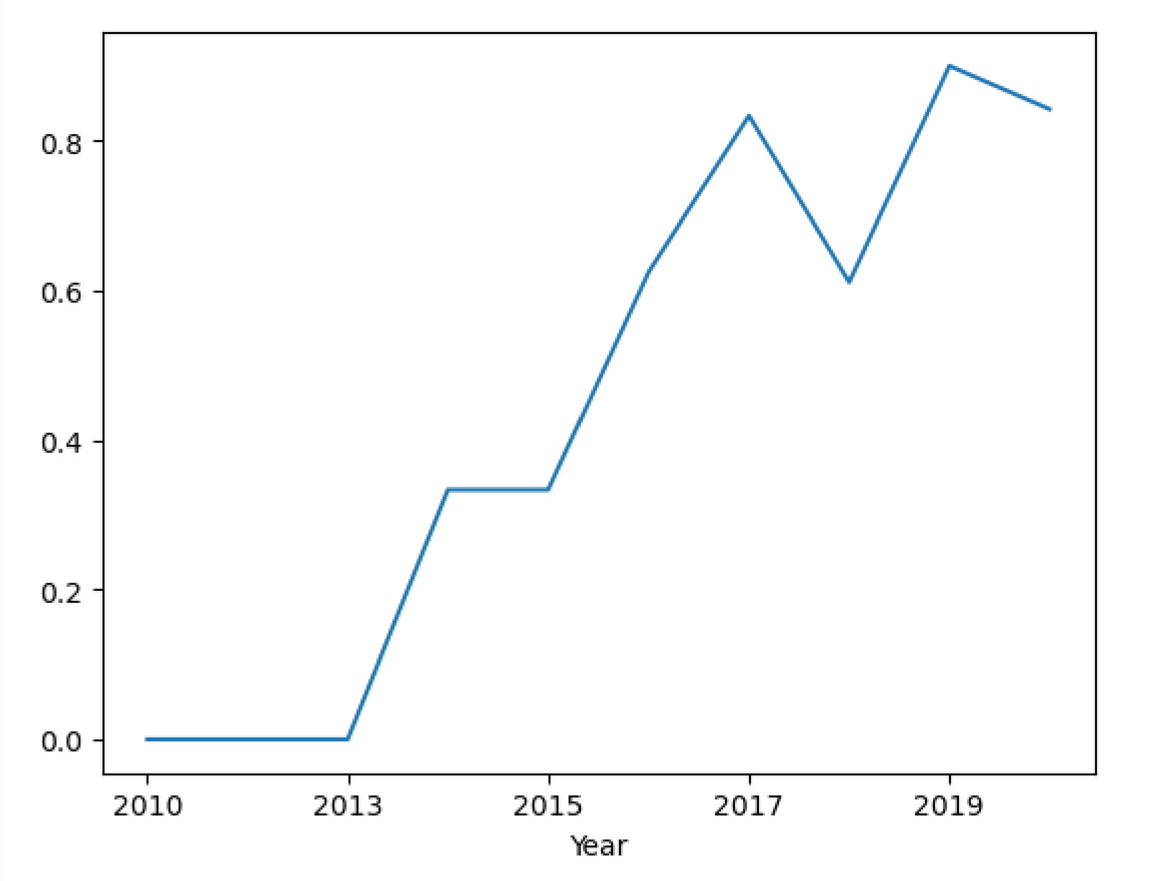
Payload vs. Orbit Type



- The highest payload mass(kg) are on the orbit VLEO. This could indicate that it is the only orbit that can handle these.
- GTO has little correlation between payload mass and success

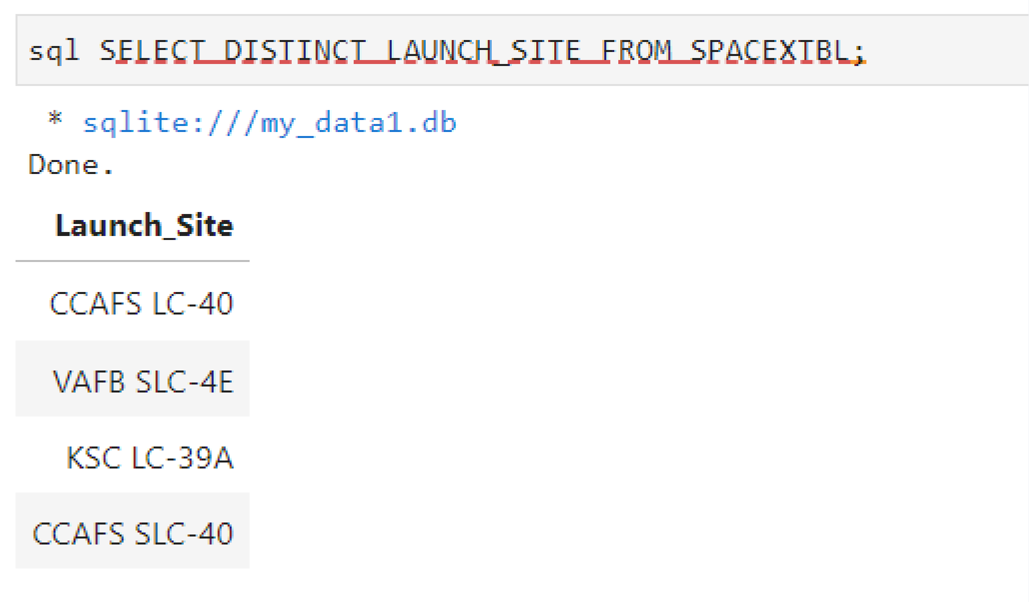
Launch Success Yearly Trend

- As time has progress the general trend is that the mean outcome has improved
- There is a big exception to this however in 2018
- Between 2015 and 2017 the success rate more than doubled



All Launch Site Names

- Selects all the distinct Launch Sites from the data set.
- There were 4 different Launch sites

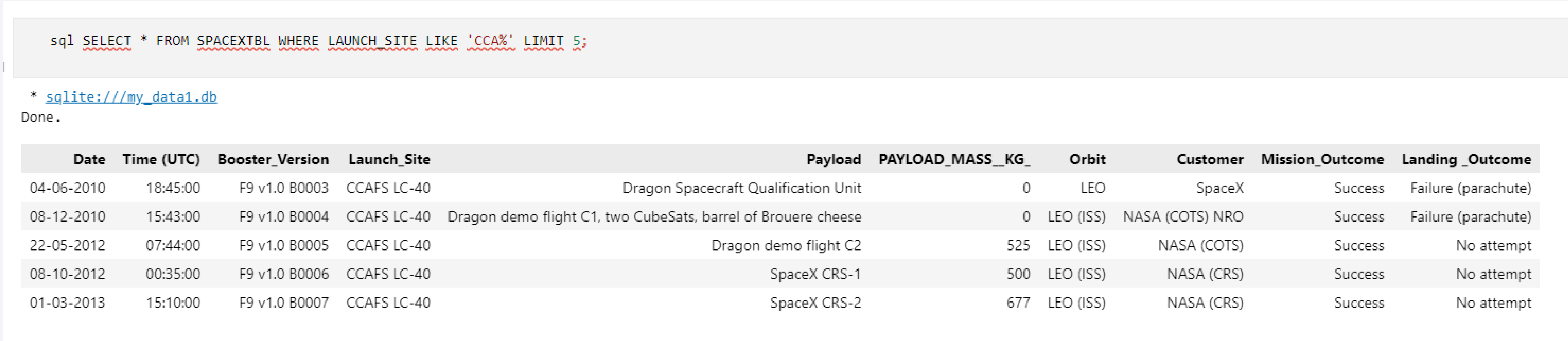


```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'



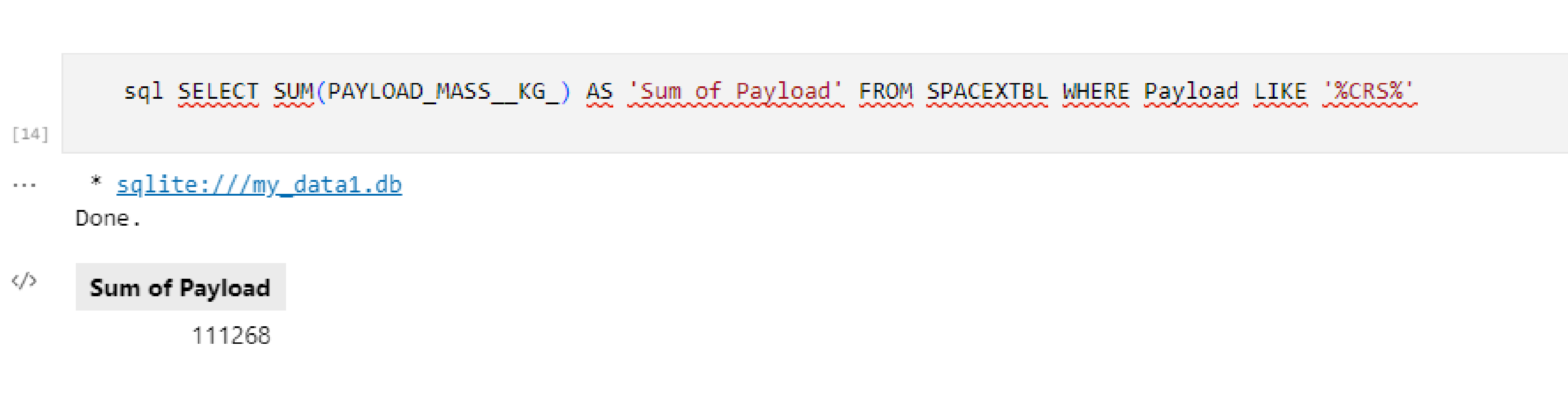
```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Selects all the fields from SPACEXTBL where the Launch Site contains ‘CCA’ and them limits it to the first 5 entries.
- This gives us an understanding of what fields are in the data set and the results within them

Total Payload Mass



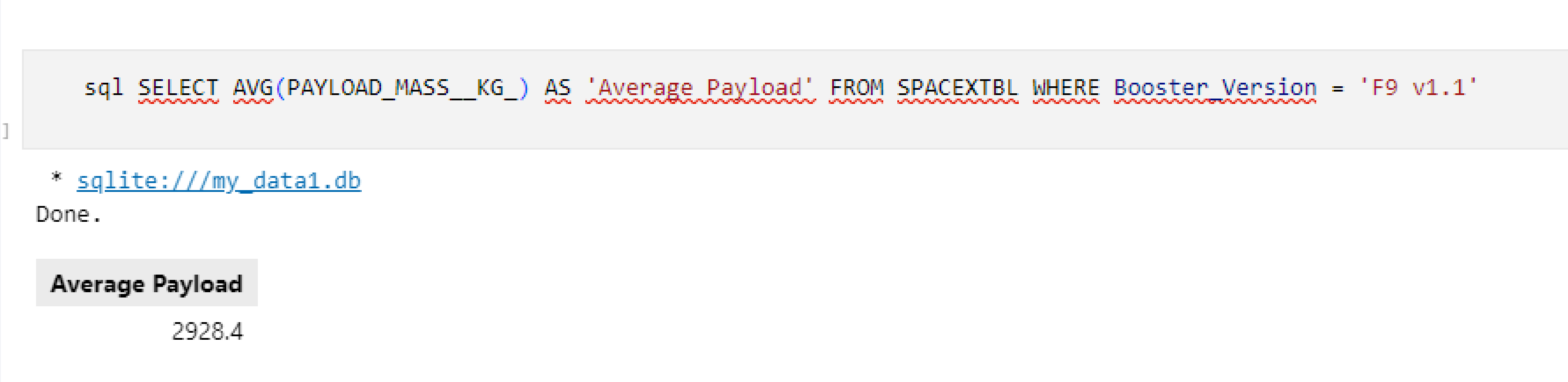
```
[14]   sql SELECT SUM(PAYLOAD_MASS__KG_) AS 'Sum of Payload' FROM SPACEXTBL WHERE Payload LIKE '%CRS%'

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

Sum of Payload
111268

- Selects the sum of the Payload Mass from SPACEXTBL where payload contains ‘CRS’ and names it ‘Sum of payload’

Average Payload Mass by F9 v1.1



```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS 'Average Payload' FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'
```

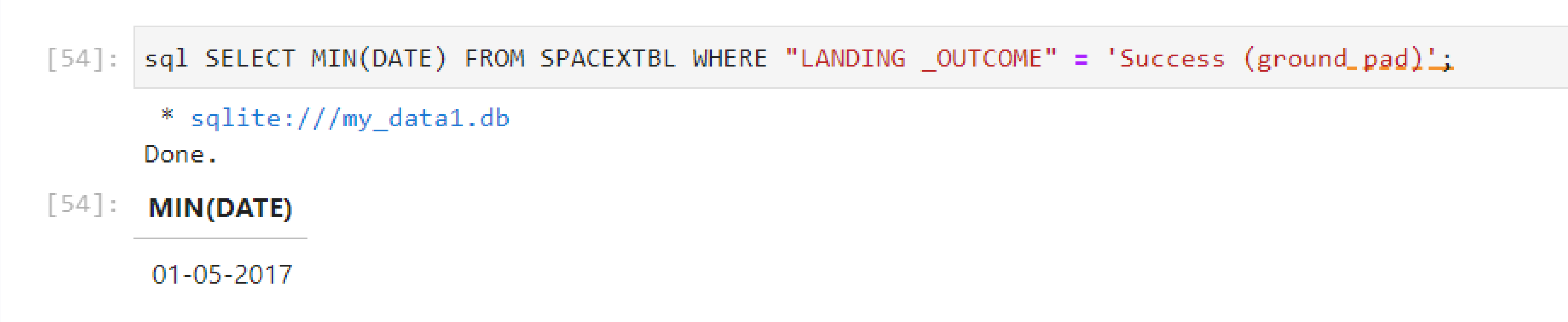
```
 * sqlite:///my_data1.db
Done.
```

Average Payload
2928.4

- Selects the average of Payload Mass from SPACEXTBL where Booster Version is ‘F9 v1.1’ and names it ‘Average Payload’

First Successful Ground Landing Date

- Selects the minimum date from the SpaceX table where the Landing Outcome is ‘Success (ground pad)’



```
[54]: sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (ground_pad)';
```

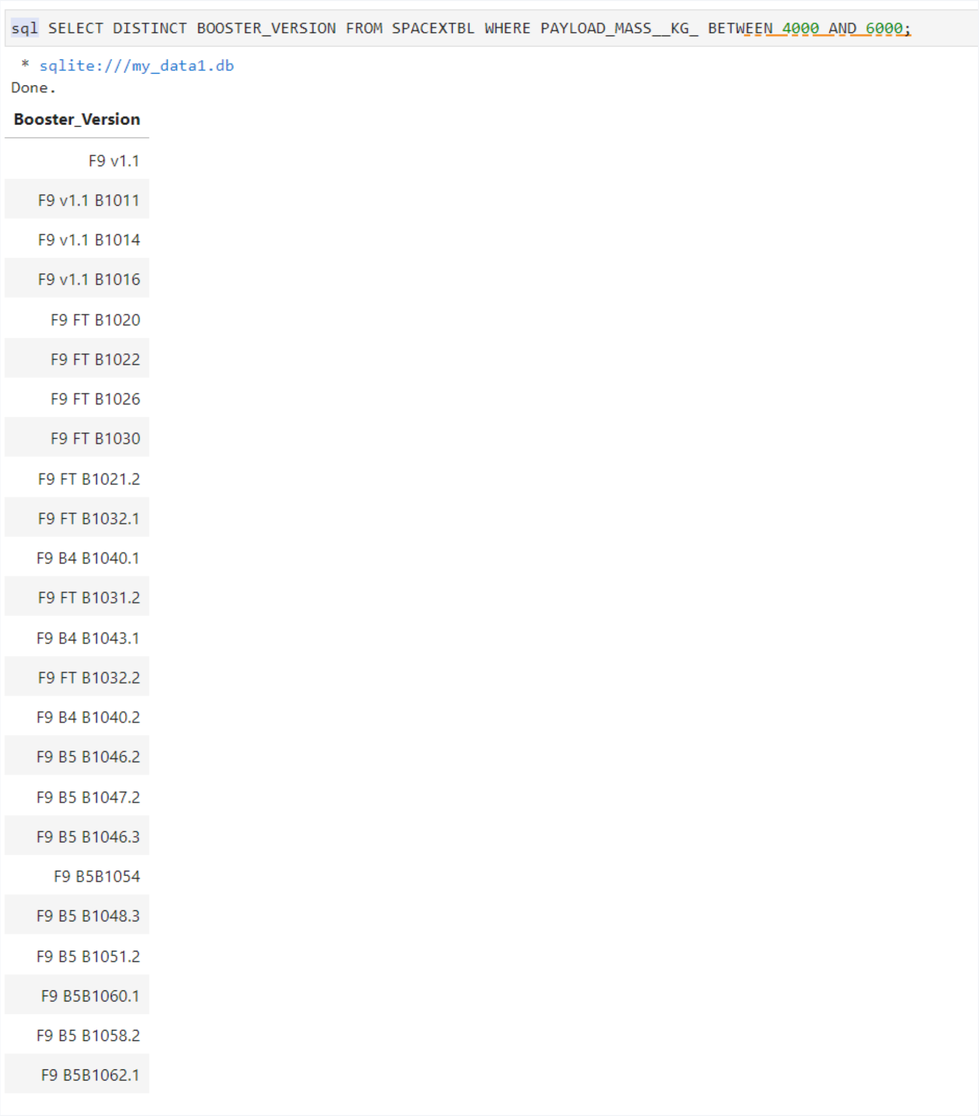
```
 * sqlite:///my_data1.db
Done.
```

[54]:

MIN(DATE)
01-05-2017

- The results show this was first achieved on 01/05/2017

Successful Drone Ship Landing with Payload between 4000 and 6000



```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

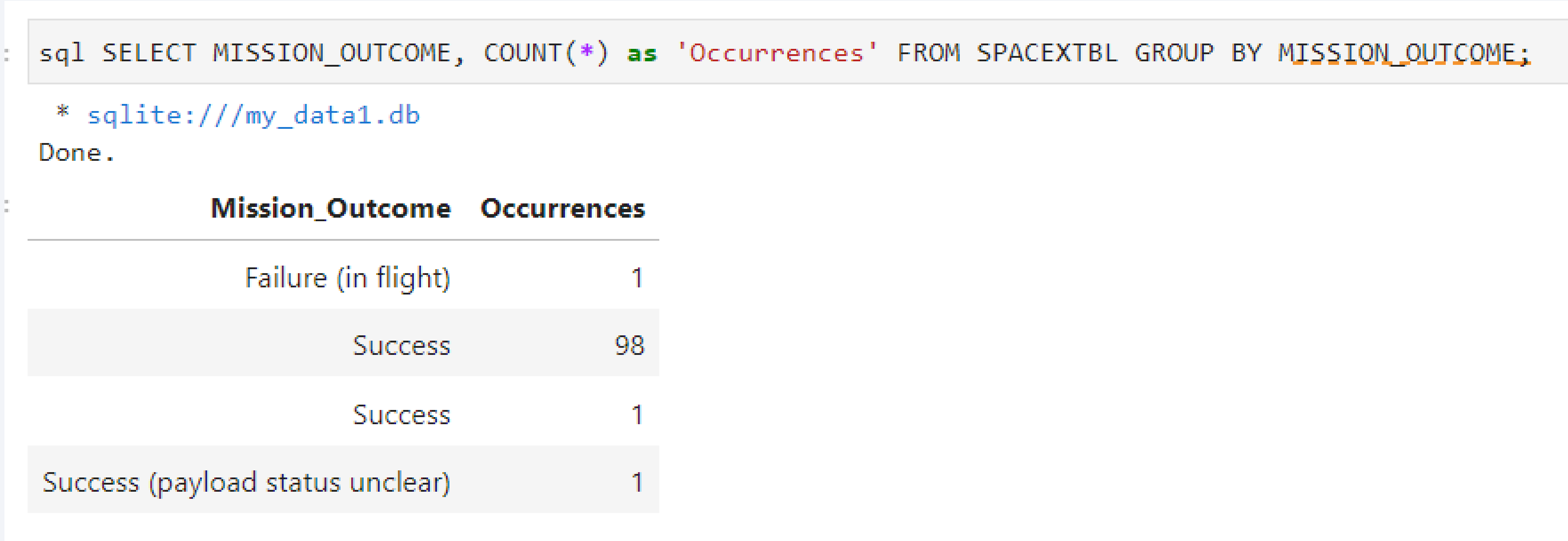
```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

- Selects all the distinct Booster Versions from the SpaceX table where the payload mass is between 4000kg and 6000kg
- There were 25 distinct values in this range

Total Number of Successful and Failure Mission Outcomes

- Selects Mission outcomes and counts the quantity of each outcome.
- As we can see only 1 mission had a failure in mission outcome. That is a failure rate of less than 1%

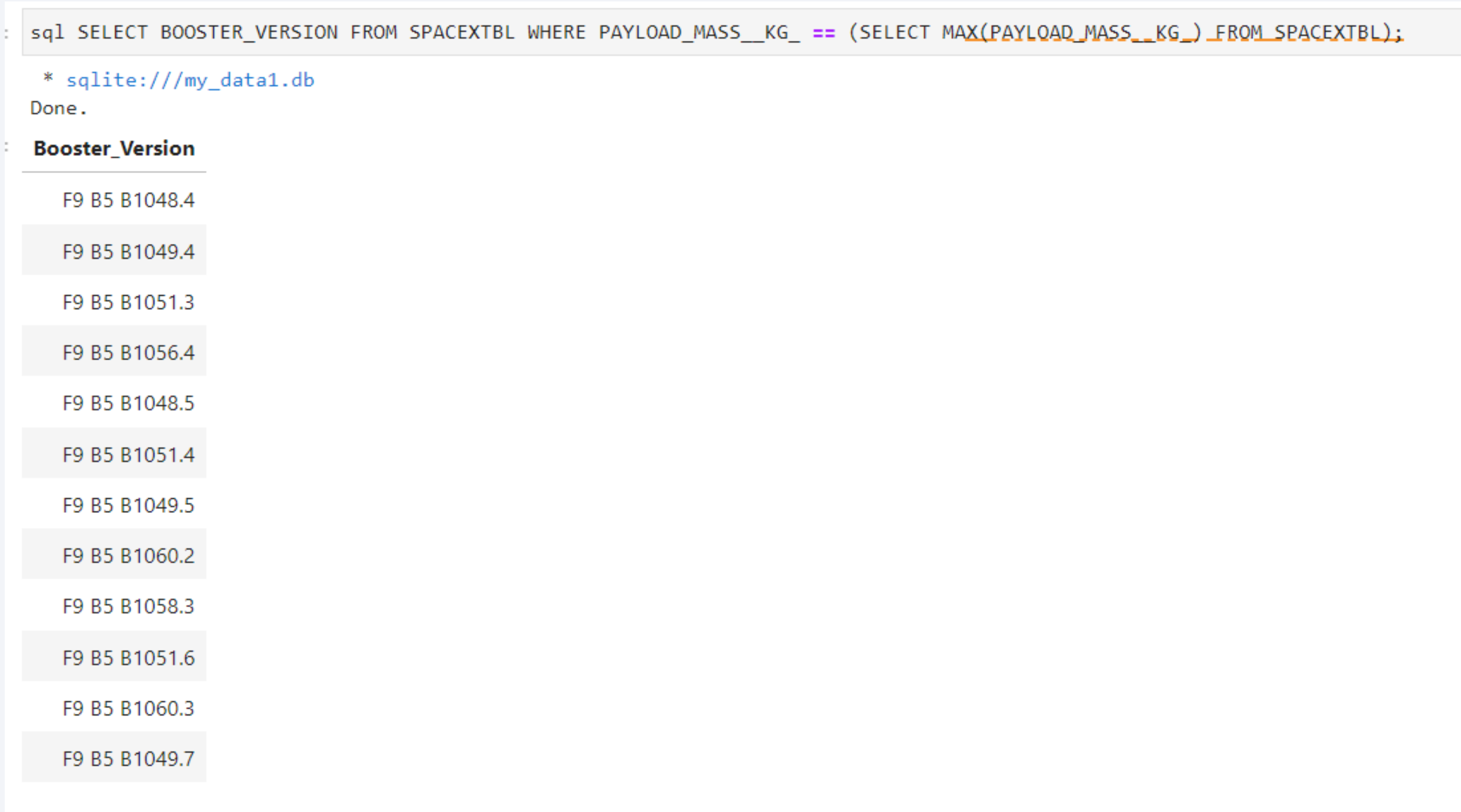


```
sql SELECT MISSION_OUTCOME, COUNT(*) as 'Occurrences' FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload



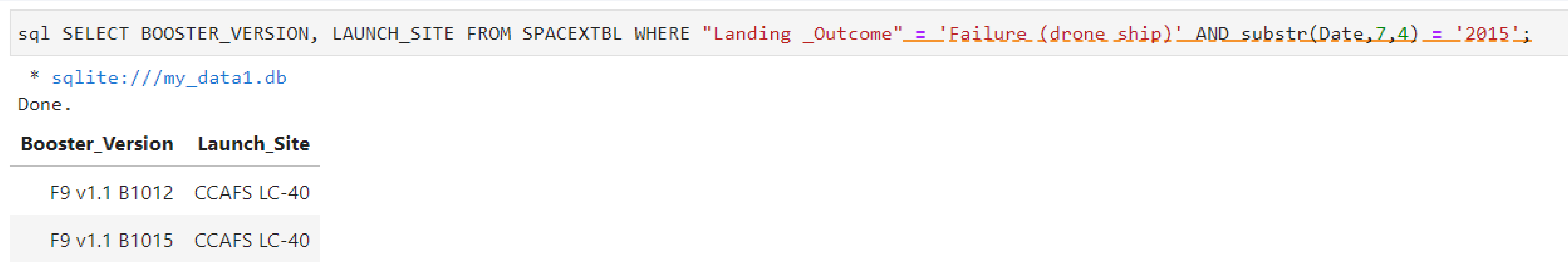
```
sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

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

- Select the booster versions from the Space X table where the payload mass is equal to its max value
- We observe 12 booster versions for the max payload mass

2015 Launch Records



```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND substr(Date,7,4) = '2015';
```

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Selects the columns booster version and launch site from the SpaceX table where the landing outcome is drone ship failures and the year is 2015
- As we can see there was only 2 outcomes with these conditions

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

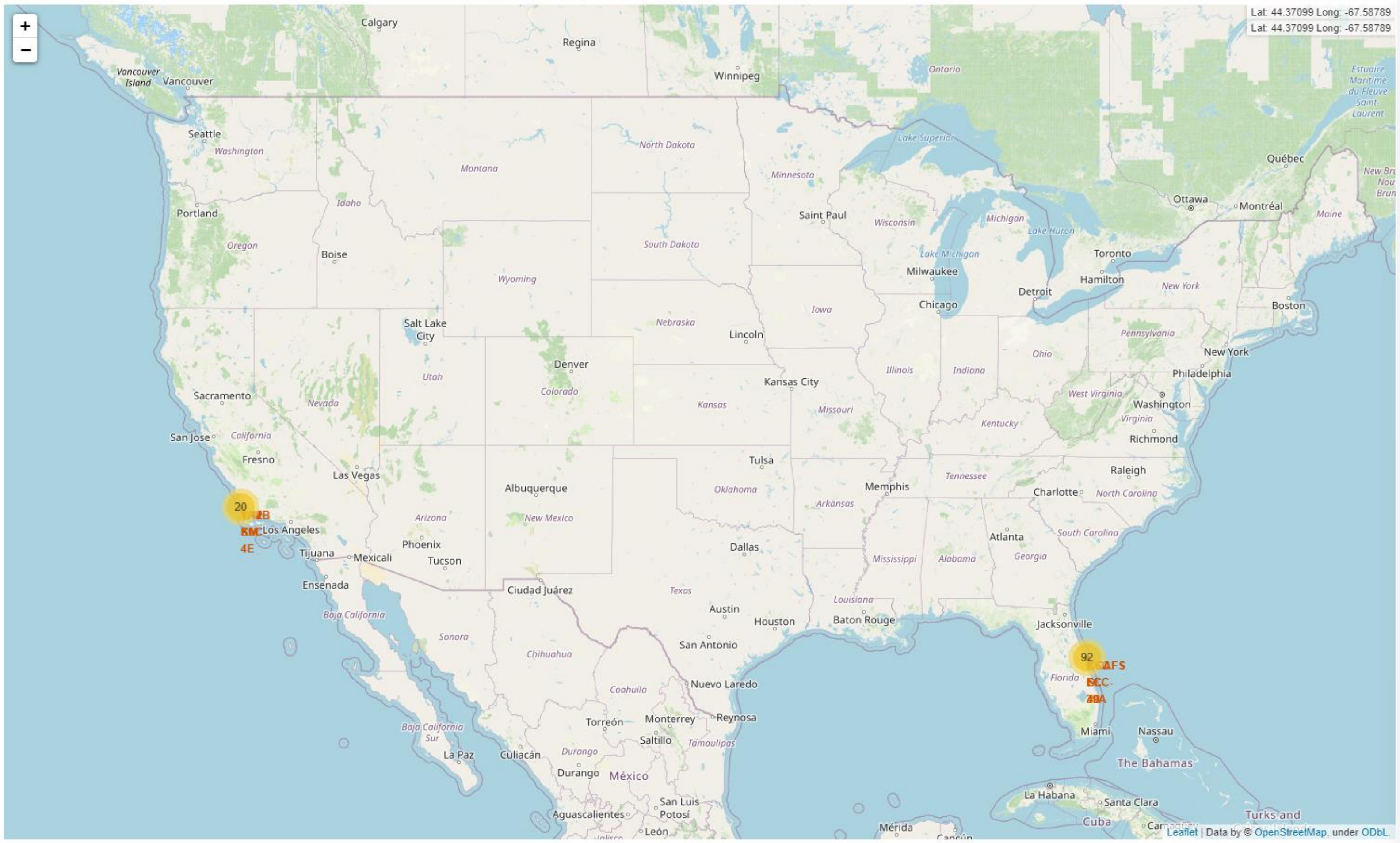
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- I was unable to complete this task

Section 3

Launch Sites Proximities Analysis

Folium Map US View

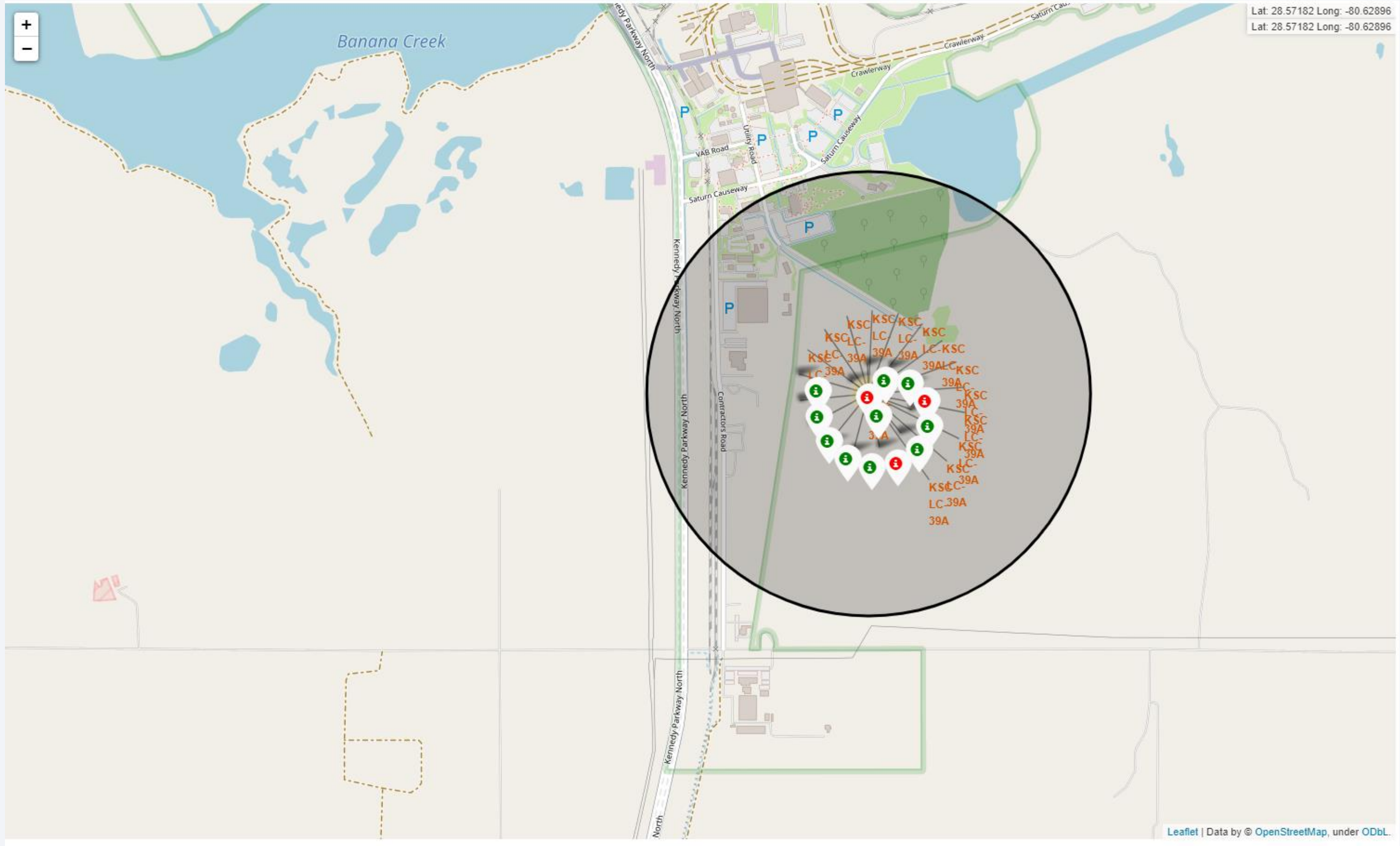


- From the view of the US we can see there were 92 launches on the southeast coast and 20 on the southwest

Exploring a Launch Site

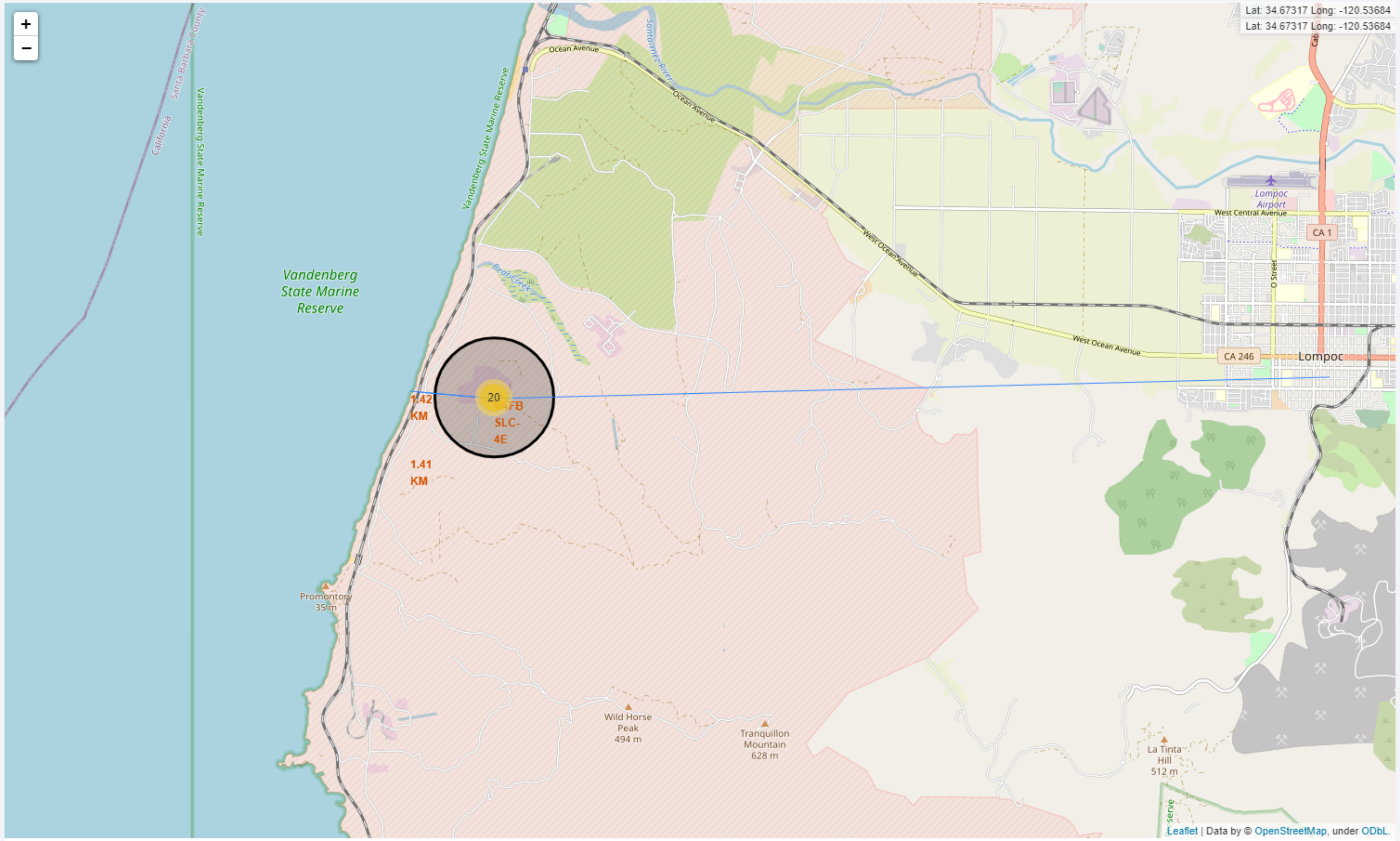
- When examining a specific launch site, we can see success and failure markers for launch outcomes.

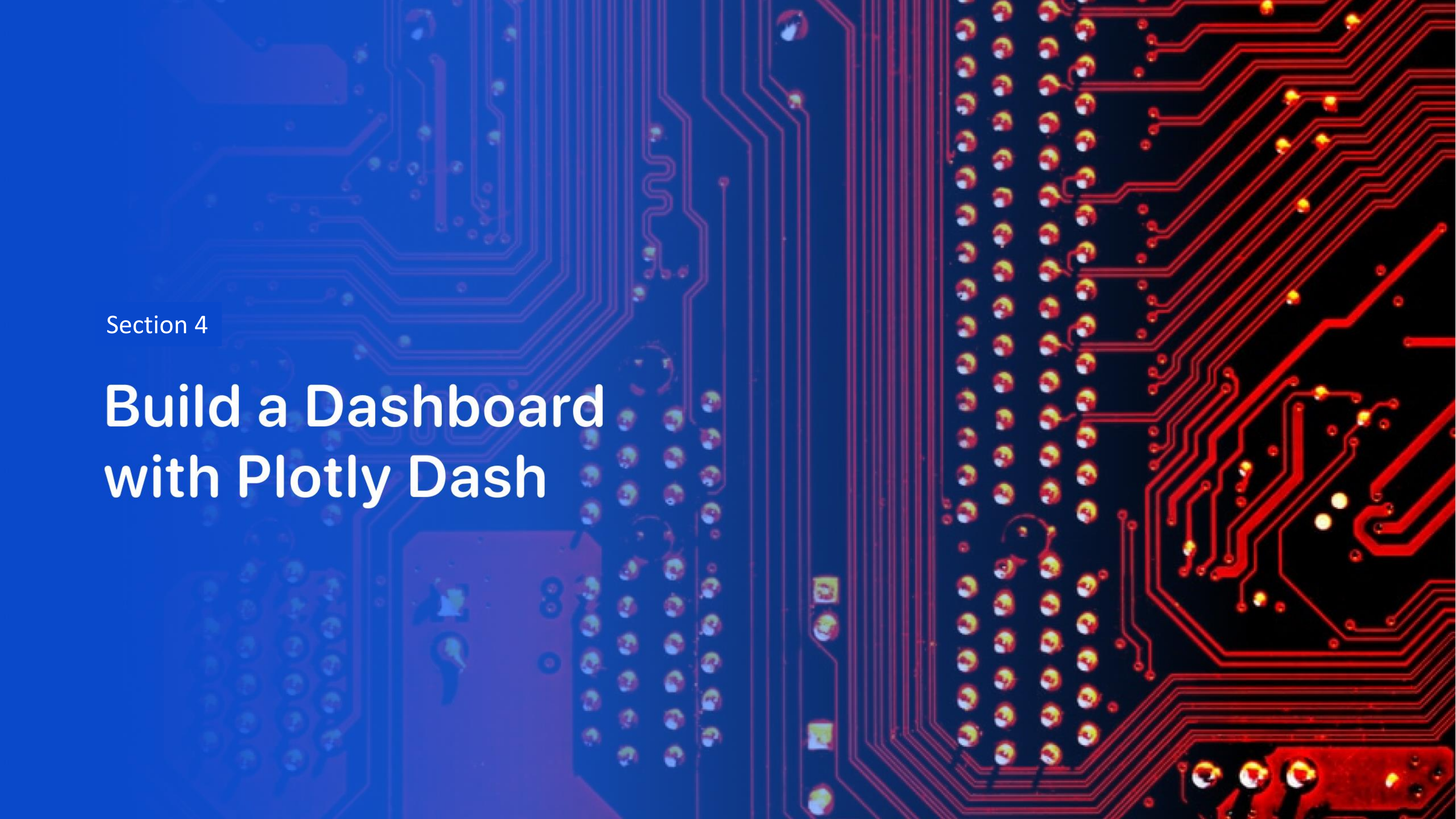
- Successes in green and failures in red



Launch Site Proximity

- There is a line between the launch site and highway and the railroad, but they are all in line with the line to the coastline.
- As we can see the distance to coastlines, railroad and highways is very close but much further to a city



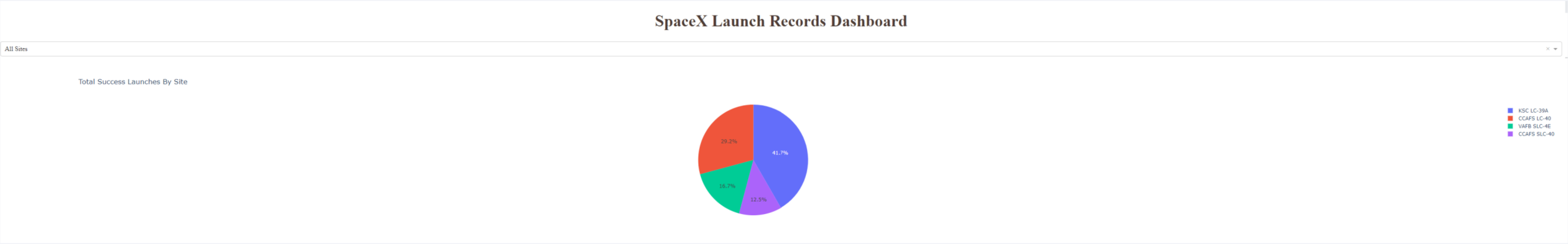


Section 4

Build a Dashboard with Plotly Dash

Pie Chart for all Launch Sites

- As we can see KSC LC-39A has the most success followed by CCAFS LC-40, VAFB SLC-4E then CCAFS SLC-40.

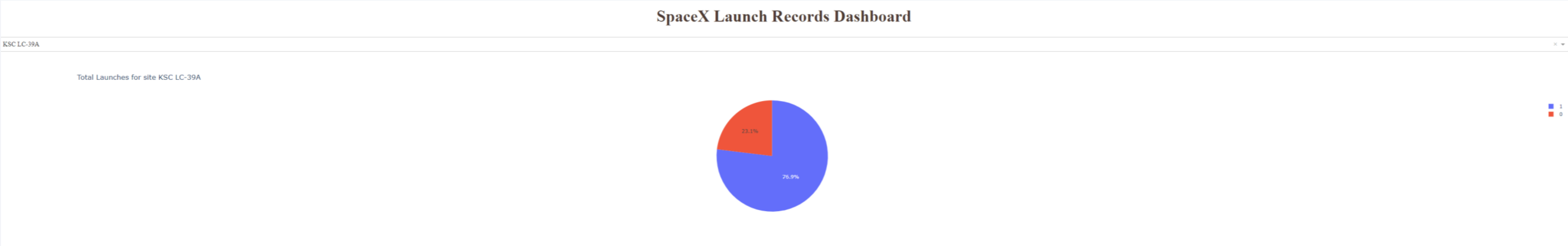


SpaceX Launch Records Dashboard

All Sites

Pie Chart for KSC LC-39A

- Below is the pie chart for KSC LC-39A, the site with the best success to failure ratio.
- As we can see it had a launch success of 76.9%



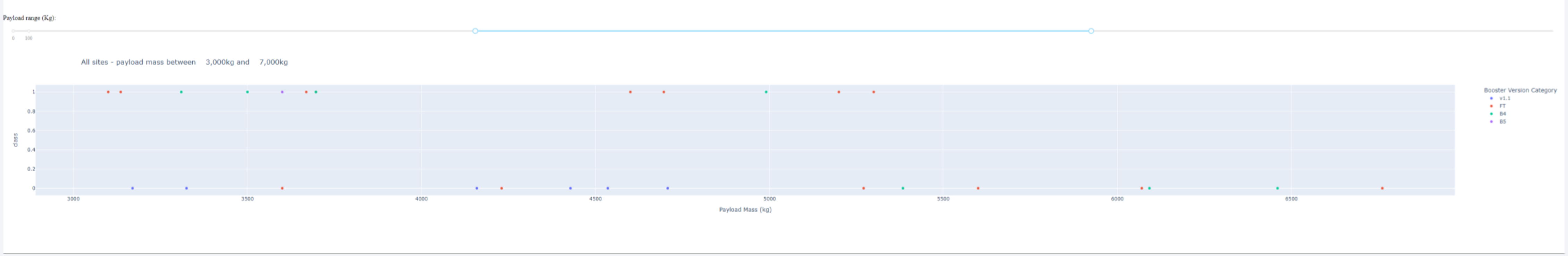
SpaceX Launch Records Dashboard

KSC LC-39A

Total Launches for site KSC LC-39A

Payload Mass (kg) Compared to Outcome

- Below is the table of payload mass compared to outcome with launch sites colour coded.
- We can also see it is filtered on payloads between 3000kg and 7000kg



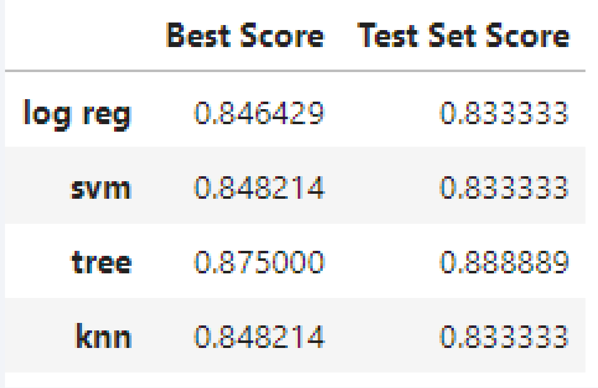
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- As we can see from the table on the right, the decision tree method had the best best score and the best test score. This means it performed best with the training set and the test set. Thus, we can conclude it was the most accurate method.
- The other 3 methods all had the same Test Score
- In terms of best score the other 3 methods ranked:

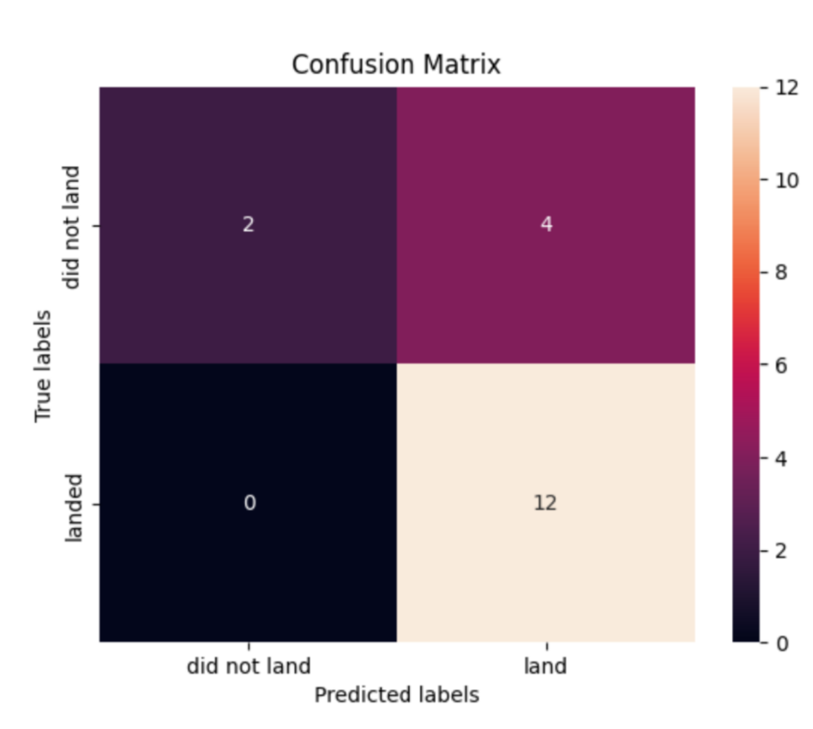
  1. KNN and SVM
  2. Log Reg



	Best Score	Test Set Score
log reg	0.846429	0.833333
svm	0.848214	0.833333
tree	0.875000	0.888889
knn	0.848214	0.833333

Confusion Matrix

- On the right is the confusion matrix for the tree model
- The model only got 4 outcomes wrong, and they were all for landed.
- Model was correct 14/18 occacions



Conclusions

- As time continued the success rate of launches increased so definitely expect improvement from SpaceY's initial launches.
- Heavier payloads were more successful in certain orbits; thus we can conclude the mass should affect what orbit to use
- The decision tree method was the best predictor so we should use that model when predicting SpaceY's success
- From the folium maps we can see that is best not to put a lunch site near a city

Appendix

- All notebooks with all code used are linked on the specific pages

Thank you!