IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Nemes
01/05/24

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Used SpaceX API calls and web scraping to gather data on Falcon 9 first stage launch data
  - Used Pandas to perform data wrangling, cleaning the data to be used
  - Performed EDA, investing the features available to understand the effect on first stage landing success
  - Created an interactive Folium map and dashboard as utilities to investigate properties of the launches
  - Created multiple predictive models and tested their performance against each other
- Summary of all results
  - Created many visualizations of the Falcon 9 first stage landing results
  - All the predictive models performed well, with the decision tree performing slightly better than the rest
  - The predictive models all suffered from a noticeable false positive rate

Introduction

- SpaceX performs launches of rockets, and attempts to land the first stage so as to re-use and to save costs

- Not all of these launches are successful, and the data of the launches can be used to investigate what features are linked to a successful launch

- Using the data this investigation explores using predictive modeling to see if we can determine if a launch will be successful given its parameters

Section 1

Methodology

Methodology

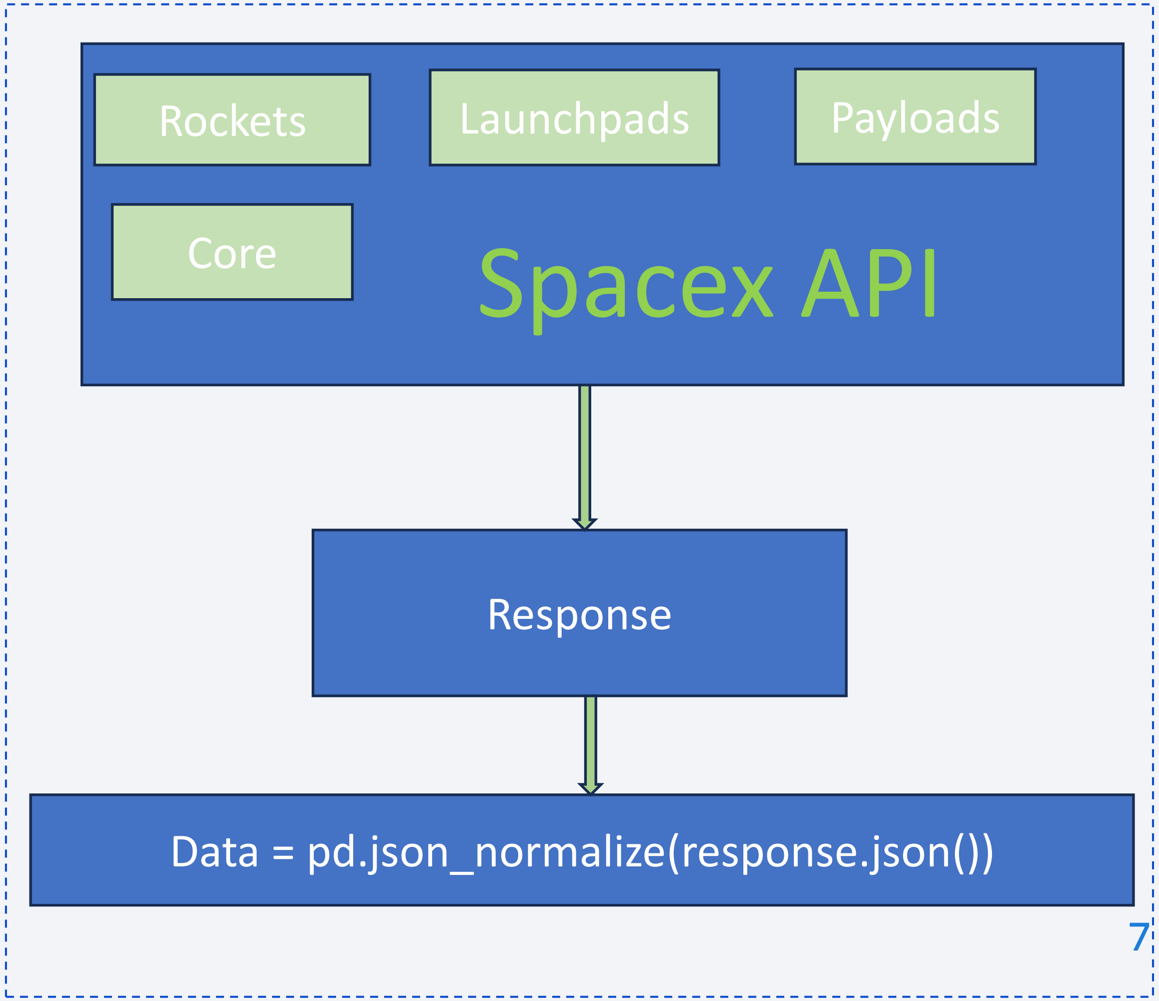
Executive Summary

- Data collection methodology:
  - Collected Data using spacex API calls
  - Collected Data using web scraping
- Perform data wrangling
  - Processed and cleaned data using PANDAS, such as imputing null values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built four predictive models, tested, and validated

Data Collection – SpaceX API

- Data collected with SpaceX REST calls

- Github Link: https://github.com/dbnemes2/IBM-Data-Science-Capstone-Project/blob/main/1_spacex_data_collection_api.ipynb

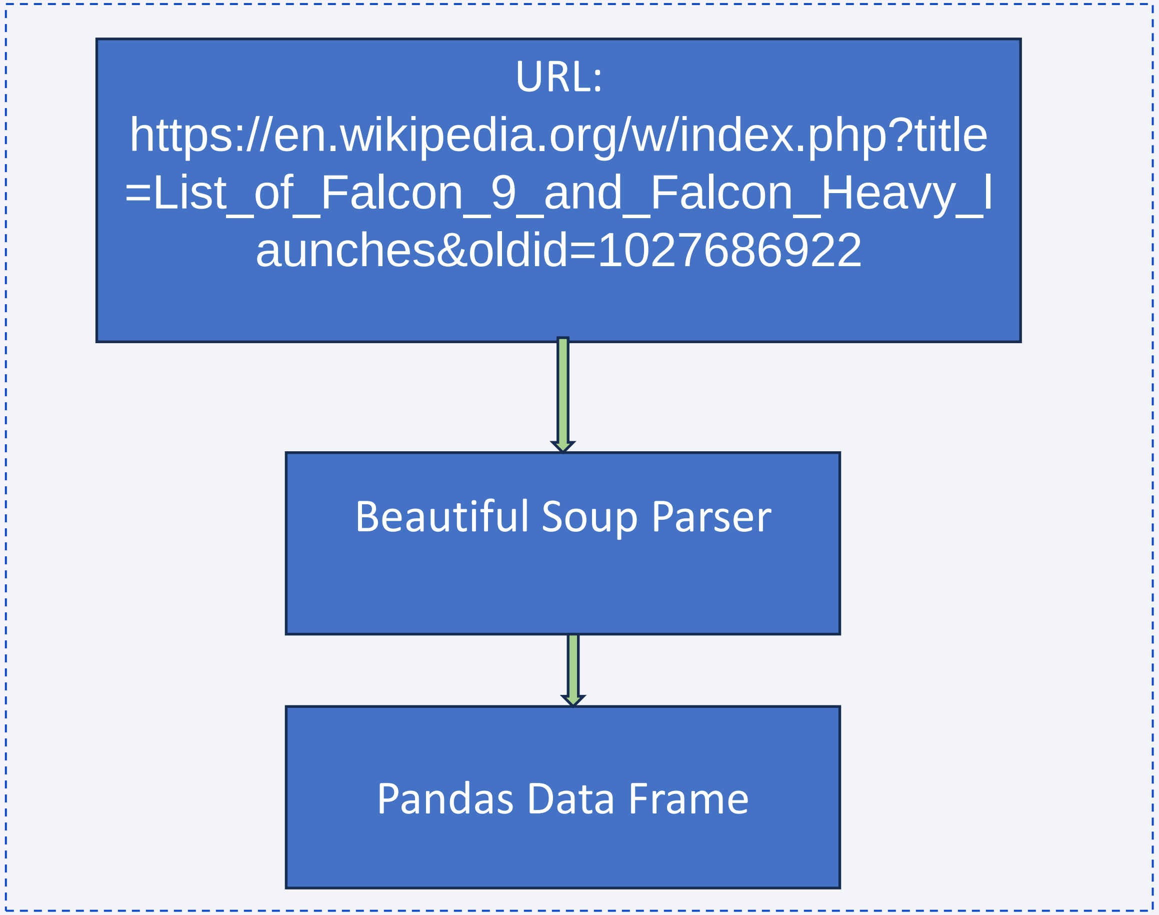


Data = pd.json_normalize(response.json())

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Github Link: https://github.com/dbnemes2/IBM-Data-Science-Capstone-Project/blob/main/2_spacex_data_collecton_webscraping.ipynb



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

Beautiful Soup Parser

Pandas Data Frame

Data Wrangling

- Initial data has missing entries, non-uniform names, values in undesirable units etc…

- Using Pandas the data is cleaned and missing values are imputed or dropped depending.

- Github Link: https://github.com/dbnemes2/IBM-Data-Science-Capstone-Project/blob/main/3_data_wrangling.ipynb

EDA with Data Visualization

- Initial Exploratory Data Analysis involving plotting values and investigating relationships

- The features primarily investigated were the flight number, the launch site, mass of payload and their effect on the outcome of the launch (success or failure)

- Github Link: https://github.com/dbnemes2/IBM-Data-Science-Capstone-Project/blob/main/4_EDA_visualization.ipynb

EDA with SQL

- The data was transformed from a Pandas dataframe and converted into a SQL database and various queries were made to investigate the data

- The distinct launching site names, the total and average payloads, amount of successful launches and more was explored

- Github URL: https://github.com/dbnemes2/IBM-Data-Science-Capstone-Project/blob/main/5_EDA_SQL.ipynb

Interactive Map with Folium

- Interactive map which shows the location of the launch sites, including extra information such as success and failures for each site, and the distance to nearby features

- From this map we can see that the launch sites are fine to be built near the coast and railroads, but are usually built with some distance from populated areas such as cities

- GitHub URL: https://github.com/dbnemes2/IBM-Data-Science-Capstone-Project/blob/main/6_FOLIUM_interactive_map.ipynb

Build a Dashboard with Plotly Dash

- A dashboard with Plotly Dash was built which allows users to select a site and show a pie chart which shows fraction of successes, and also a scatter plot with payload mass (kg) vs Class (success or failure)

- These two plots with the user input allows for the investigation of the effect of properties on what is linked to a successful launch

- GitHub URL: https://github.com/dbnemes2/IBM-Data-Science-Capstone-Project/blob/main/7_dashboard.py

Predictive Analysis (Classification)

- Using the features of the data investigated, four models were built easily with the scikit-learn package

- The data was split into training and testing partitions, then models were made with the training data and validated with the testing data

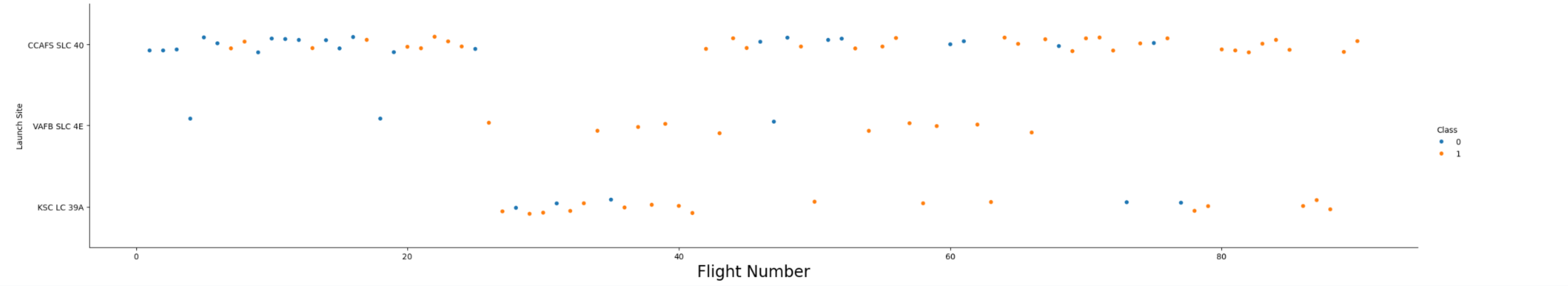
- The performance of the models were compared based on their accuracy and best score values

- GitHub: https://github.com/dbnemes2/IBM-Data-Science-Capstone-Project/blob/main/8_predictive_models.ipynb

Section 2

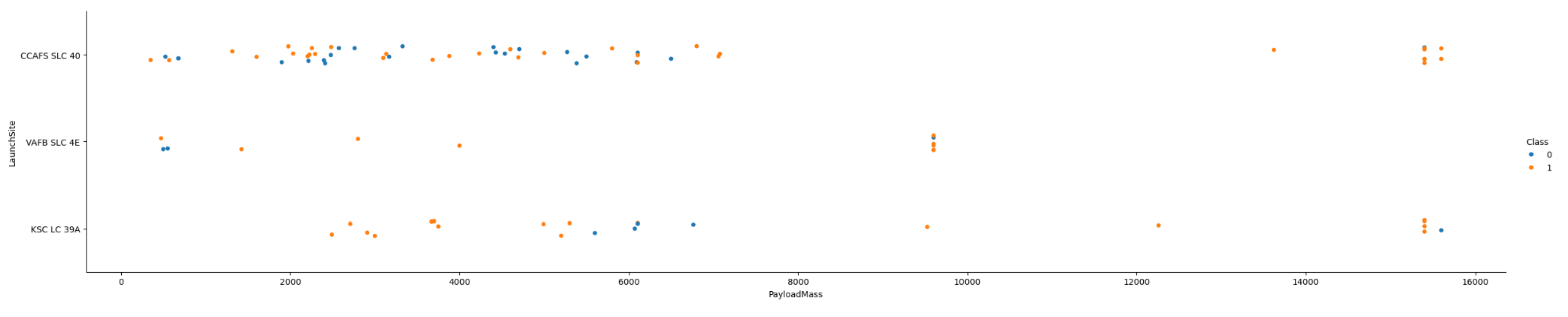
Insights drawn from EDA

Flight Number vs. Launch Site



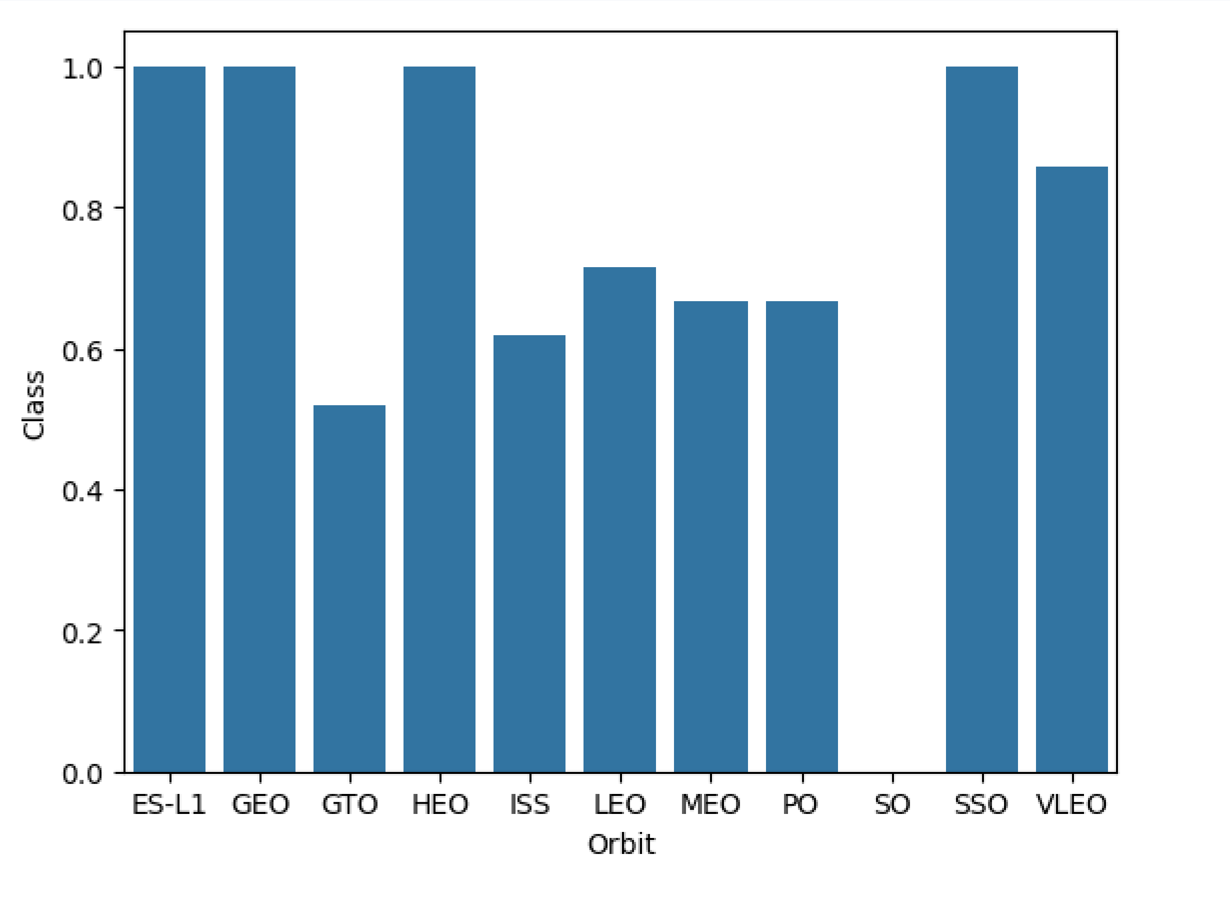
- Simple scatter plot showing the launch site used for each flight
- Points colored based on successful launch (orange) and failed launch (blue)
- Can see simple trend of more successful launches as time goes on

Payload vs. Launch Site



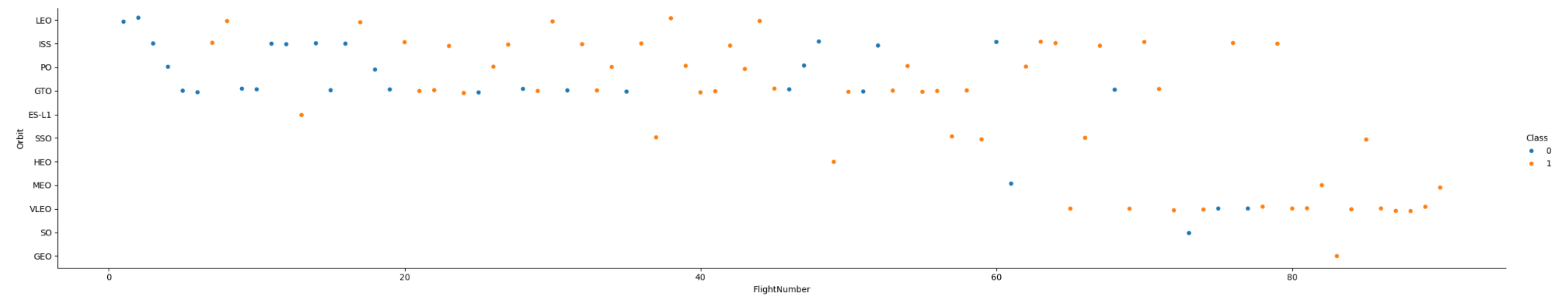
- Scatter plot of Payload Mass (kg) vs. Launch Site
- The VAF8 SLC 4E site doesn't have any payloads more massive than 10000 kg
- The heaviest payloads tend to be more successful
- The CCAFS SLC 40 site has the most amount of small payloads

Success Rate vs. Orbit Type



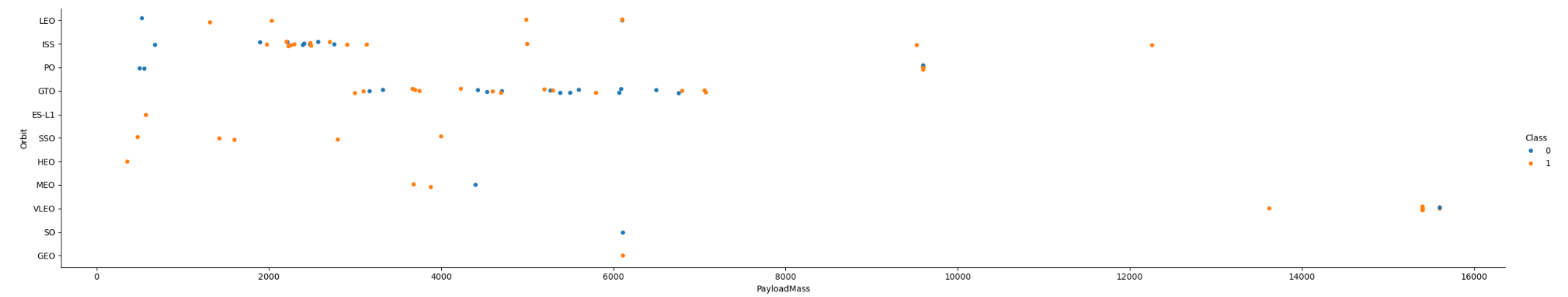
- Success rate of launches for different orbit types
- The y-axis (Class) is the fraction of successful launches
- Multiple orbits have only successes

Flight Number vs. Orbit Type



- Scatter plot of orbit type vs flight number
- Can see that the zero-success rate from SO orbit isn’t significant as it only has one entry
- The VLEO orbit had high success rate, but it was only attempted for later flight numbers

Payload vs. Orbit Type



- Scatter plot of orbit type vs payload mass (kg)

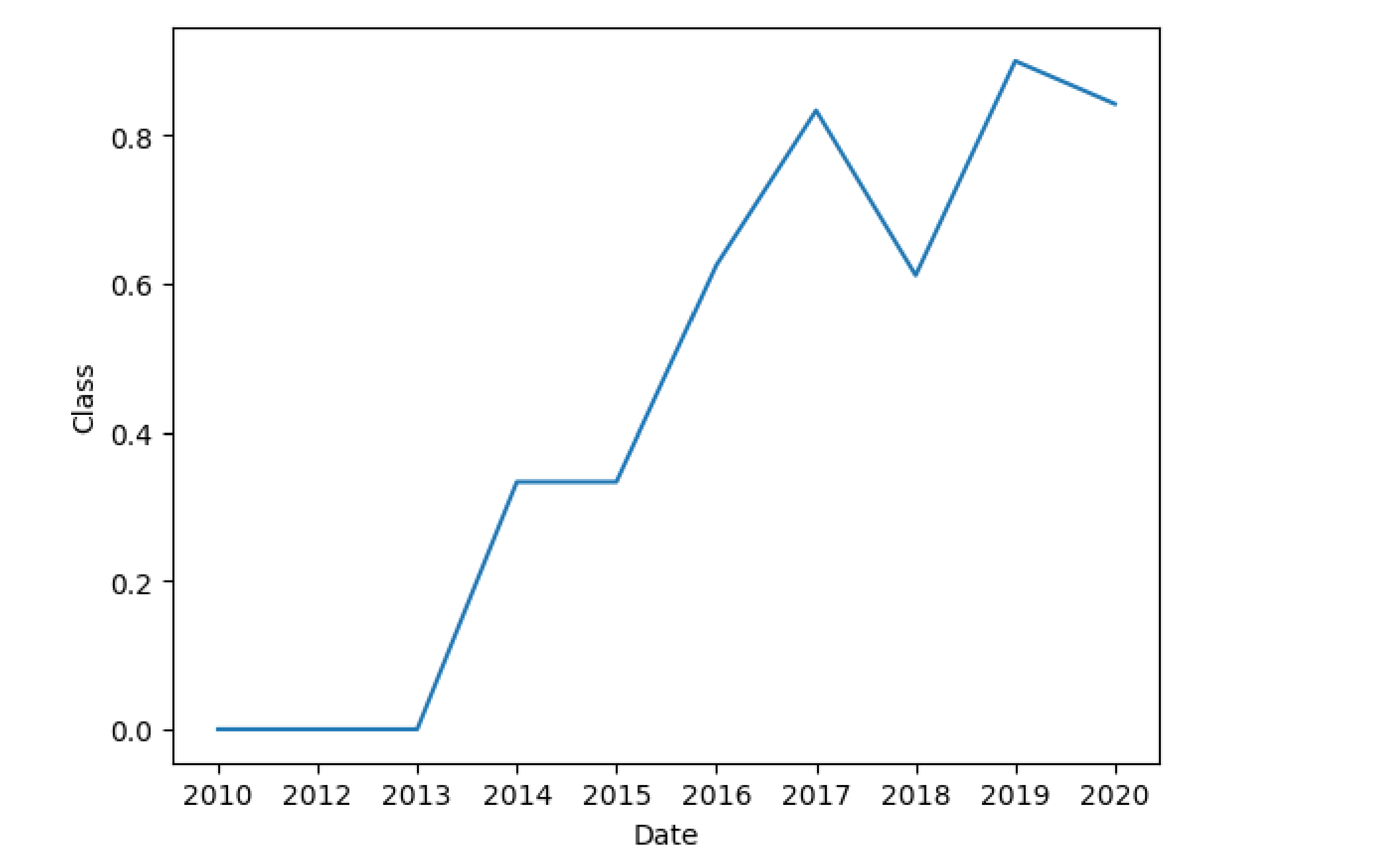
- The GTO orbit has a more confined range of payload masses than most of the other orbits

Launch Success Yearly Trend

- Line plot of the success rate (shown as Class) for launches for each year

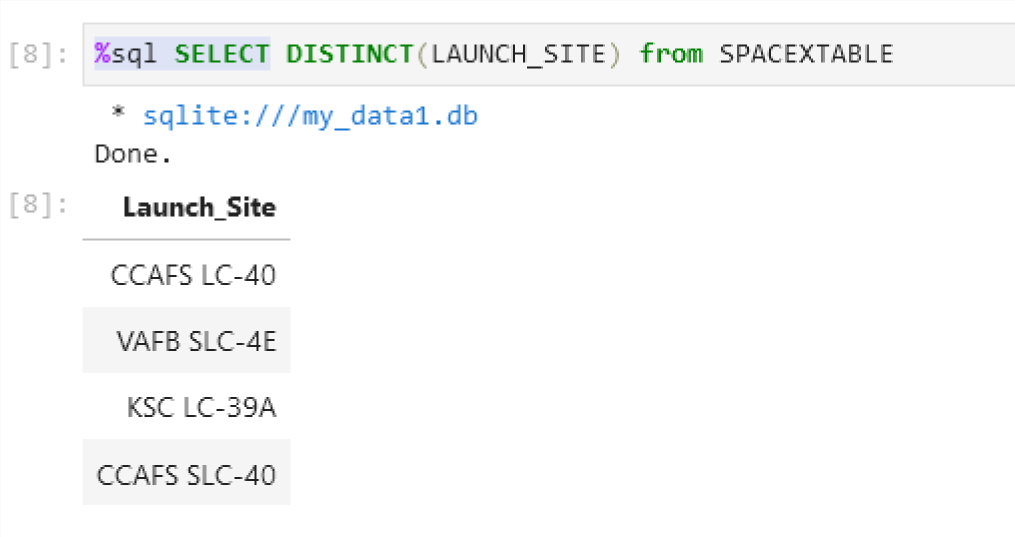
- As time went on, spacex had a higher rate of successful launches

- Expected as they improved over time



All Launch Site Names

- Using SQL all launch site names were found
- Used DISTINCT keyword to make sure not to have duplicates



```
[8]: %sql SELECT DISTINCT(LAUNCH_SITE) from SPACEXTABLE
```

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

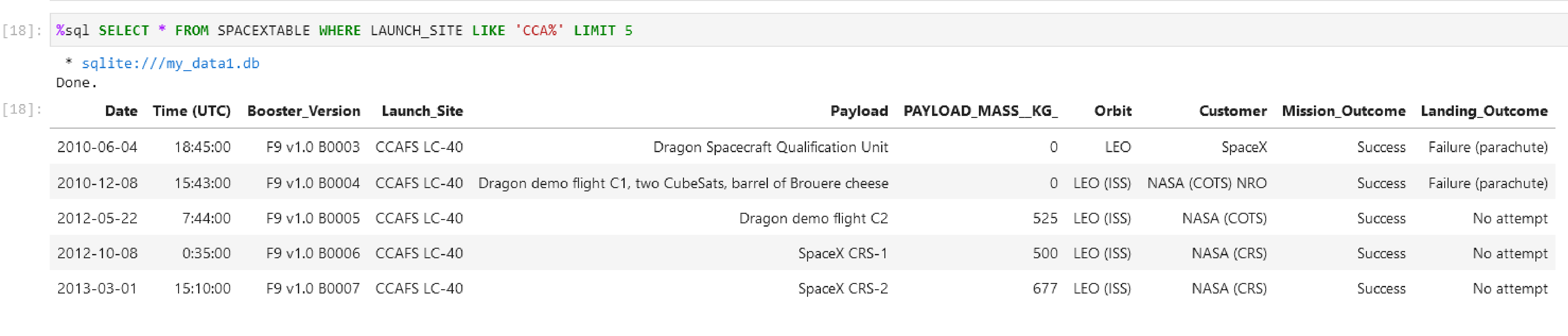
[8]:

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

Launch Site Names Begin with 'CCA'

- Five records were found with launch site name beginning with `CCA`

- Used keyword ‘LIKE’ to find strings matching ‘CCA%’ with % being a wildcard

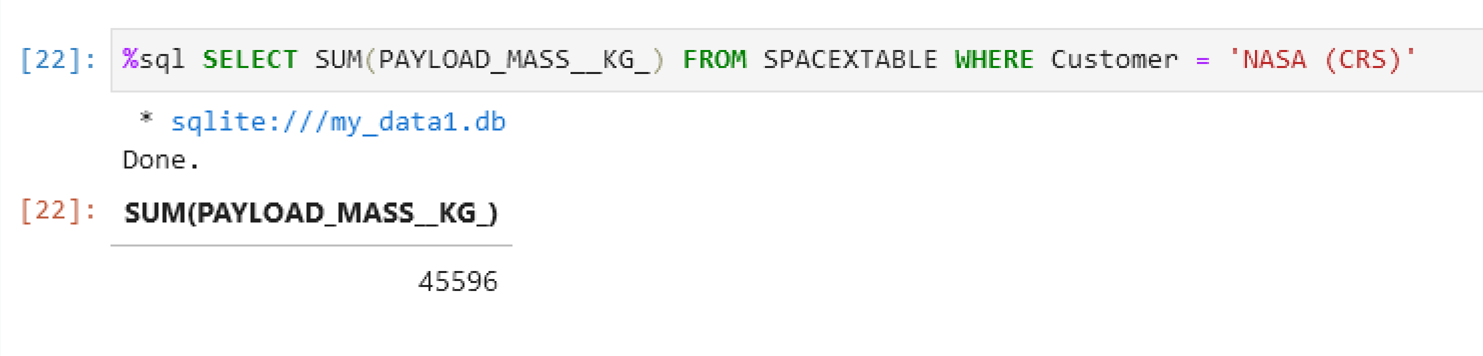


```
[18]: %sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

 * sqlite:///my_data1.db
Done.
[18]:
```

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

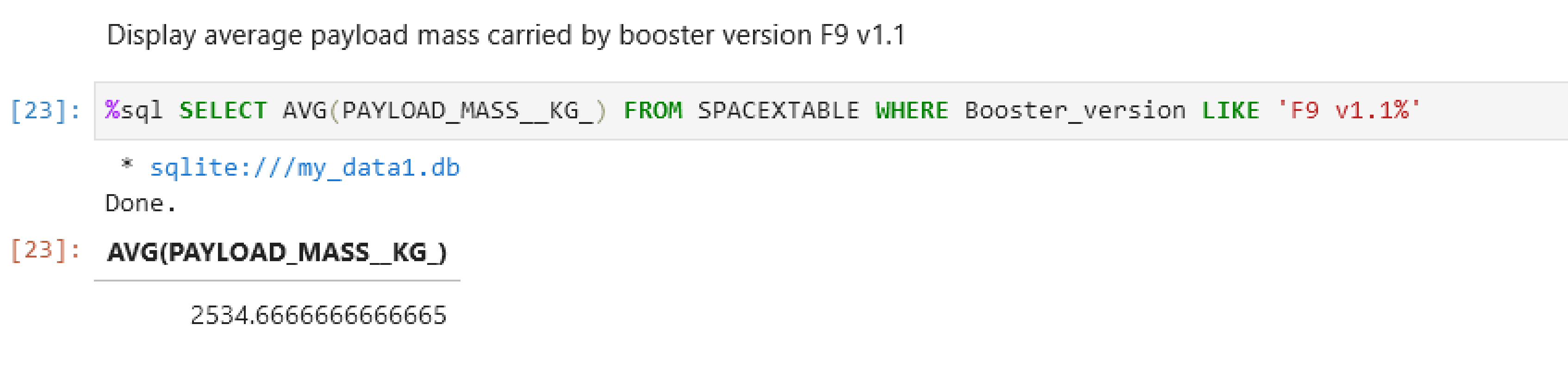
- Calculated the total payload mass (in kg) carried by boosters from NASA
- Used the SUM keyword combined with selecting Customer = ‘NASA (CRS)’



```
[22]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.

[22]: SUM(PAYLOAD_MASS__KG_)
                     45596
```

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1.1
- Used the AVG keyword combined with LIKE to select booster version F9 v1.1



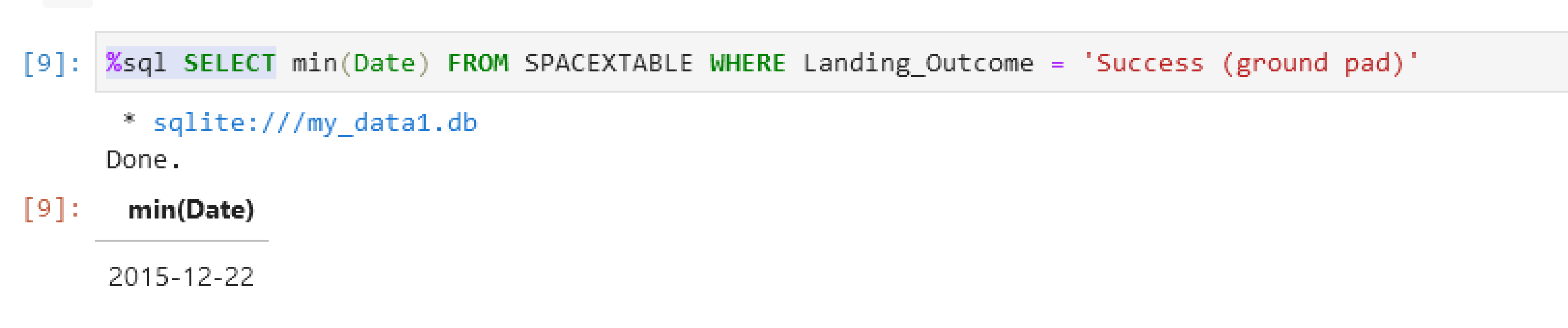
Display average payload mass carried by booster version F9 v1.1

```
[23]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_version LIKE 'F9 v1.1%'

 * sqlite:///my_data1.db
Done.

[23]: AVG(PAYLOAD_MASS__KG_)
              2534.6666666666665
```

First Successful Ground Landing Date

- Found the date of the first successful landing outcome on ground pad

- Used the keyword MIN while selecting on success (ground pad) landing outcomes

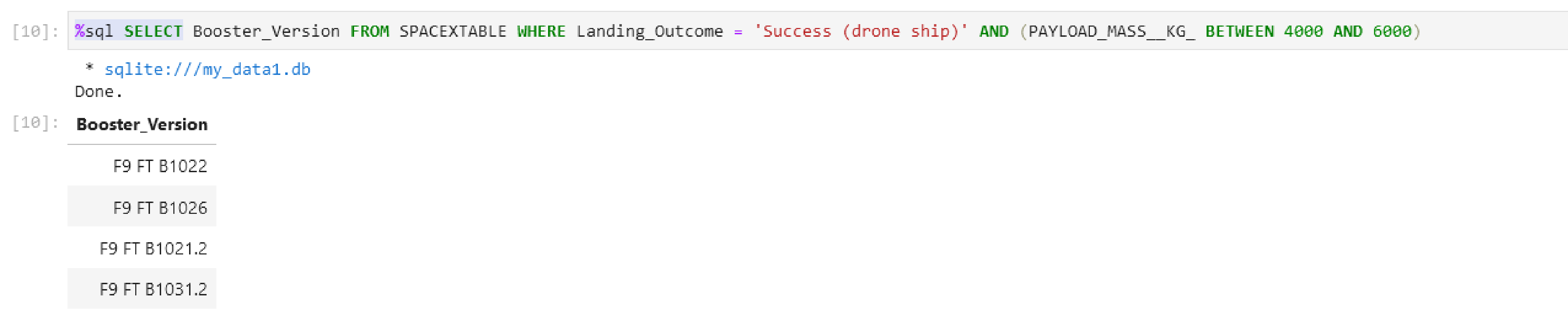


```
[9]: %sql SELECT min(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'

      * sqlite:///my_data1.db
     Done.

[9]:   min(Date)

      2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Used the keyword BETWEEN to select values in a range for PAYLOAD_MASS__KG



```
[10]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

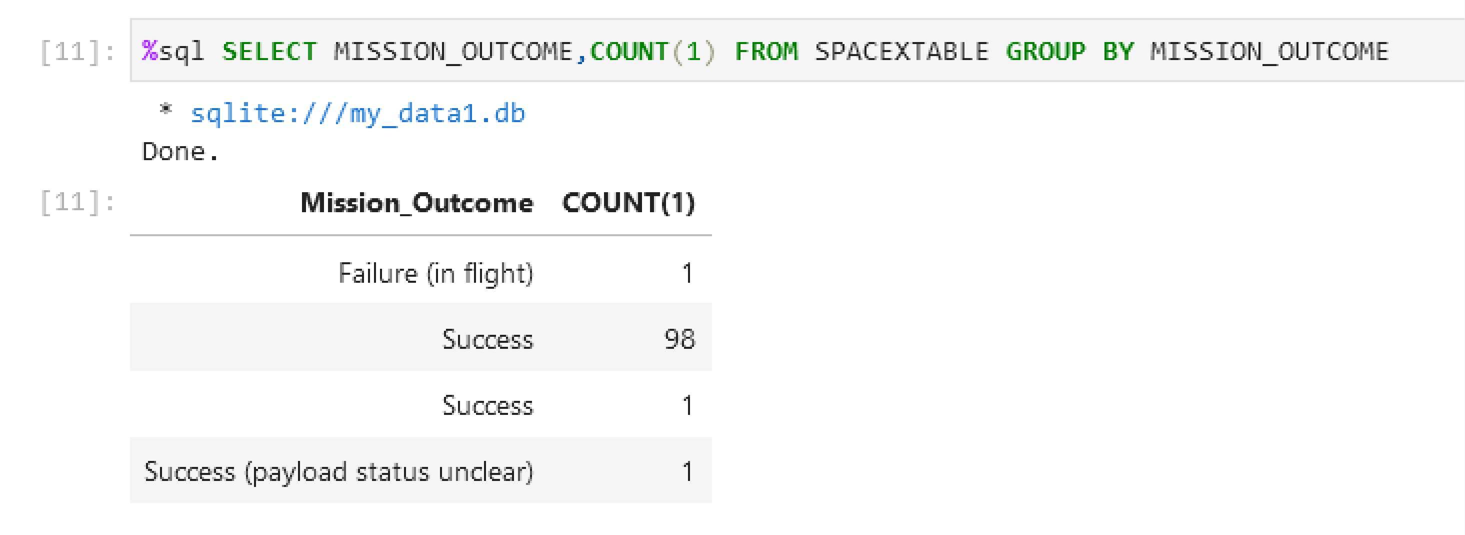
```
 * sqlite:///my_data1.db
Done.
```

[10]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failure mission outcomes
- Used keywords GROUPY BY and COUNT to get the number of outcomes for each type
- There are multiple Success types, not successes are the same and have different properties



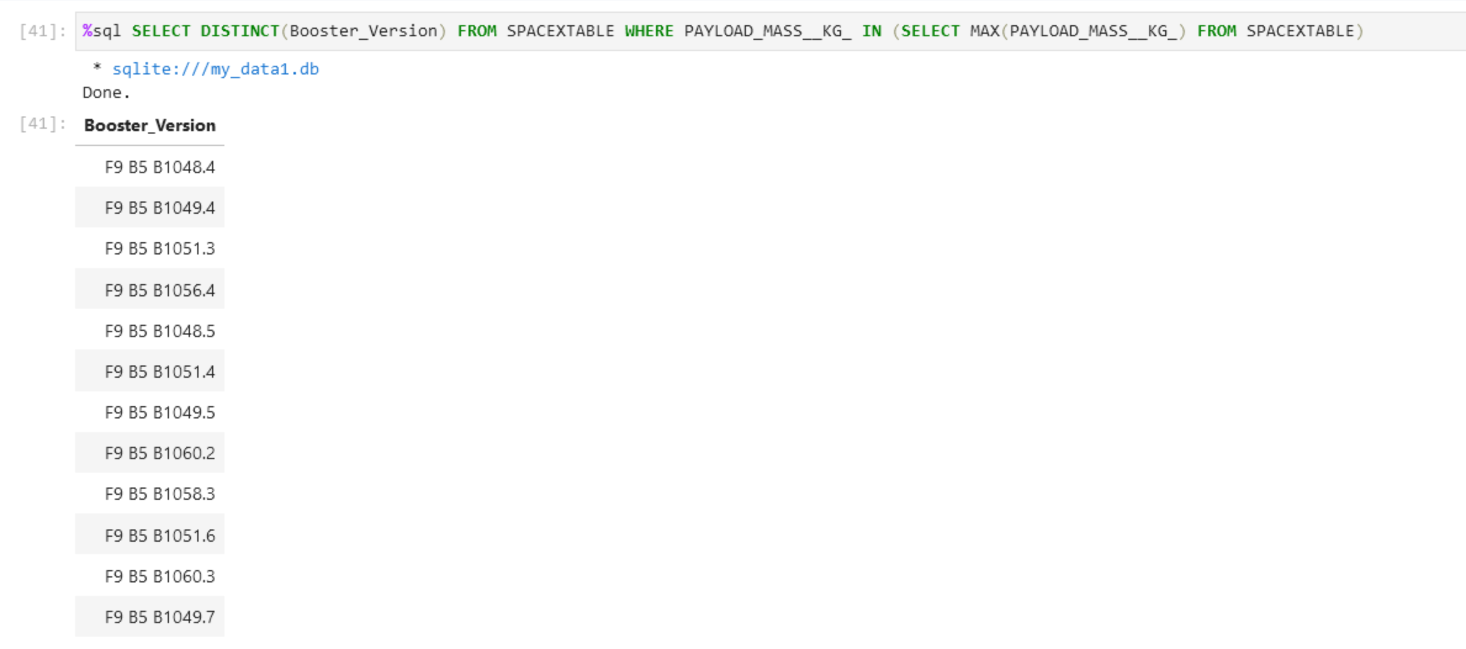
```
[11]: %sql SELECT MISSION_OUTCOME,COUNT(1) FROM SPACEXTABLE GROUP BY MISSION_OUTCOME

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

[11]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Used a subquery to find the max payload mass, then selected DISTINCT booster_versions which had matching payload mass to the max



```
[41]: %sql SELECT DISTINCT(Booster_Version) FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

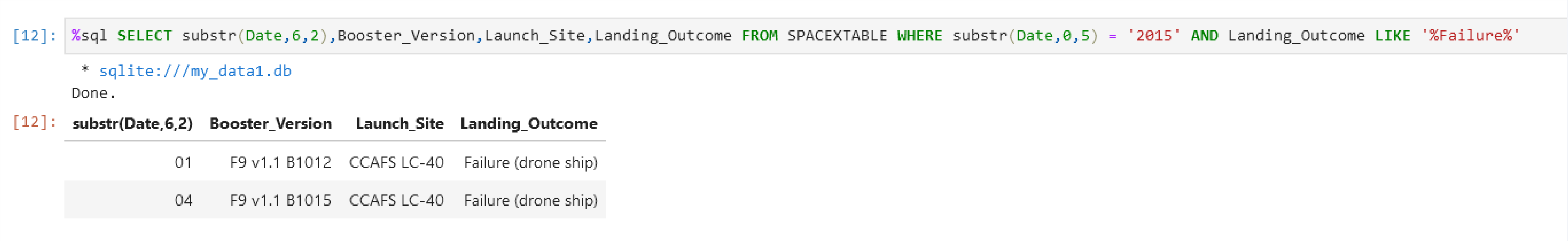
[41]:

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

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

- Used substr to select the month, and a combination of WHERE, AND, and LIKE



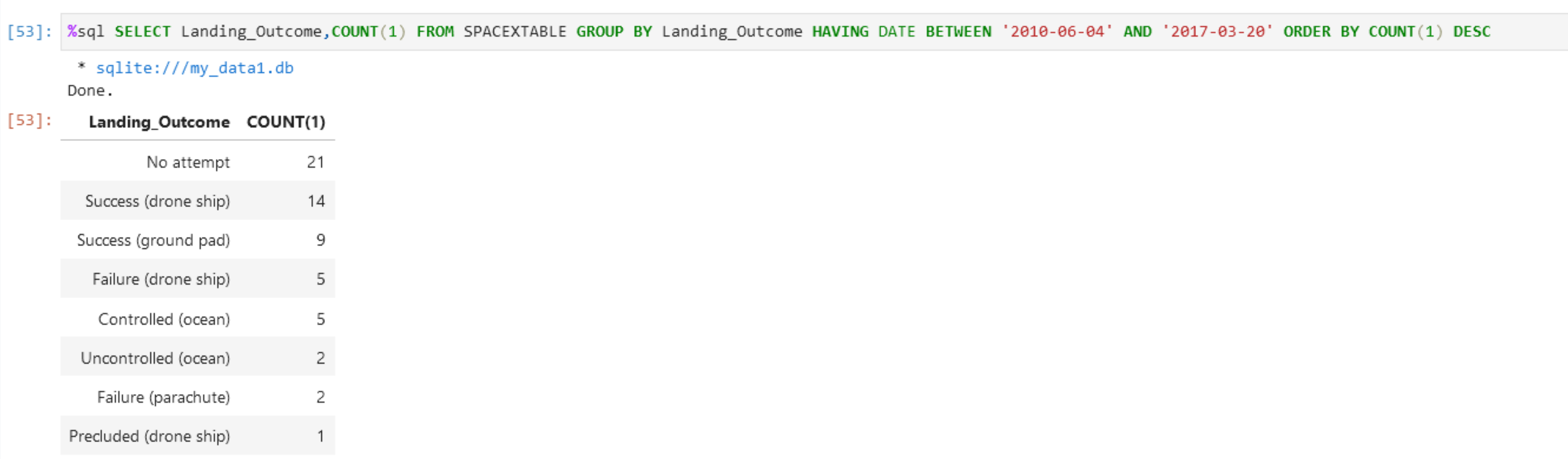
```
[12]: %sql SELECT substr(Date,6,2),Booster_Version,Launch_Site,Landing_Outcome FROM SPACEXTABLE WHERE substr(Date,0,5) = '2015' AND Landing_Outcome LIKE '%Failure%'

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

[12]:

substr(Date,6,2)	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Ranked 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
- Used keyword GROUPY BY to group by Landing_Outcome, and HAVING to select entries BETWEEN the selected dates.
- The most common outcome is “No attempt” at 21 counts



```
[53]: %sql SELECT Landing_Outcome,COUNT(1) FROM SPACEXTABLE GROUP BY Landing_Outcome HAVING DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY COUNT(1) DESC

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

[53]:

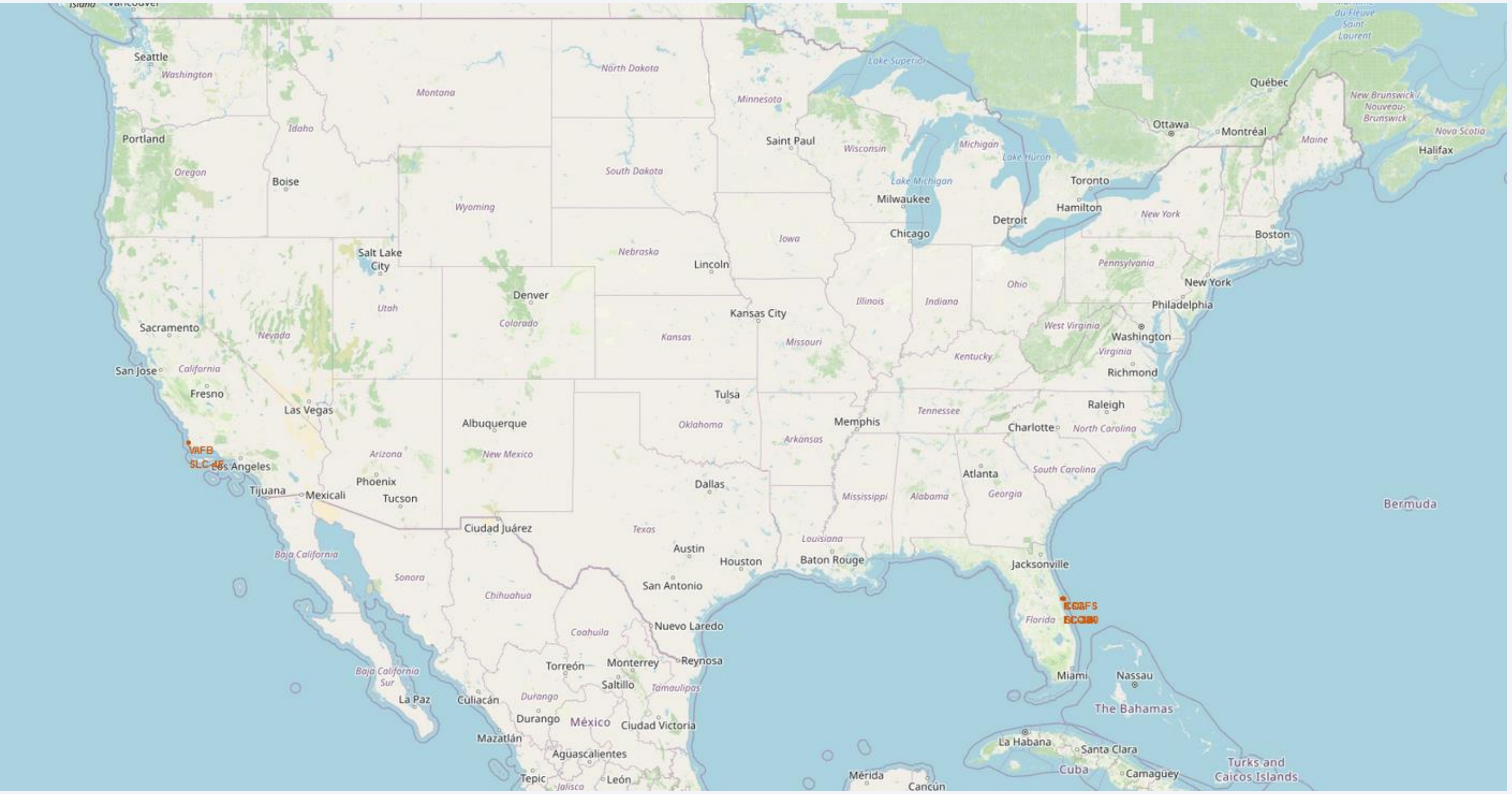
Landing_Outcome	COUNT(1)
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Site Locations in the United States

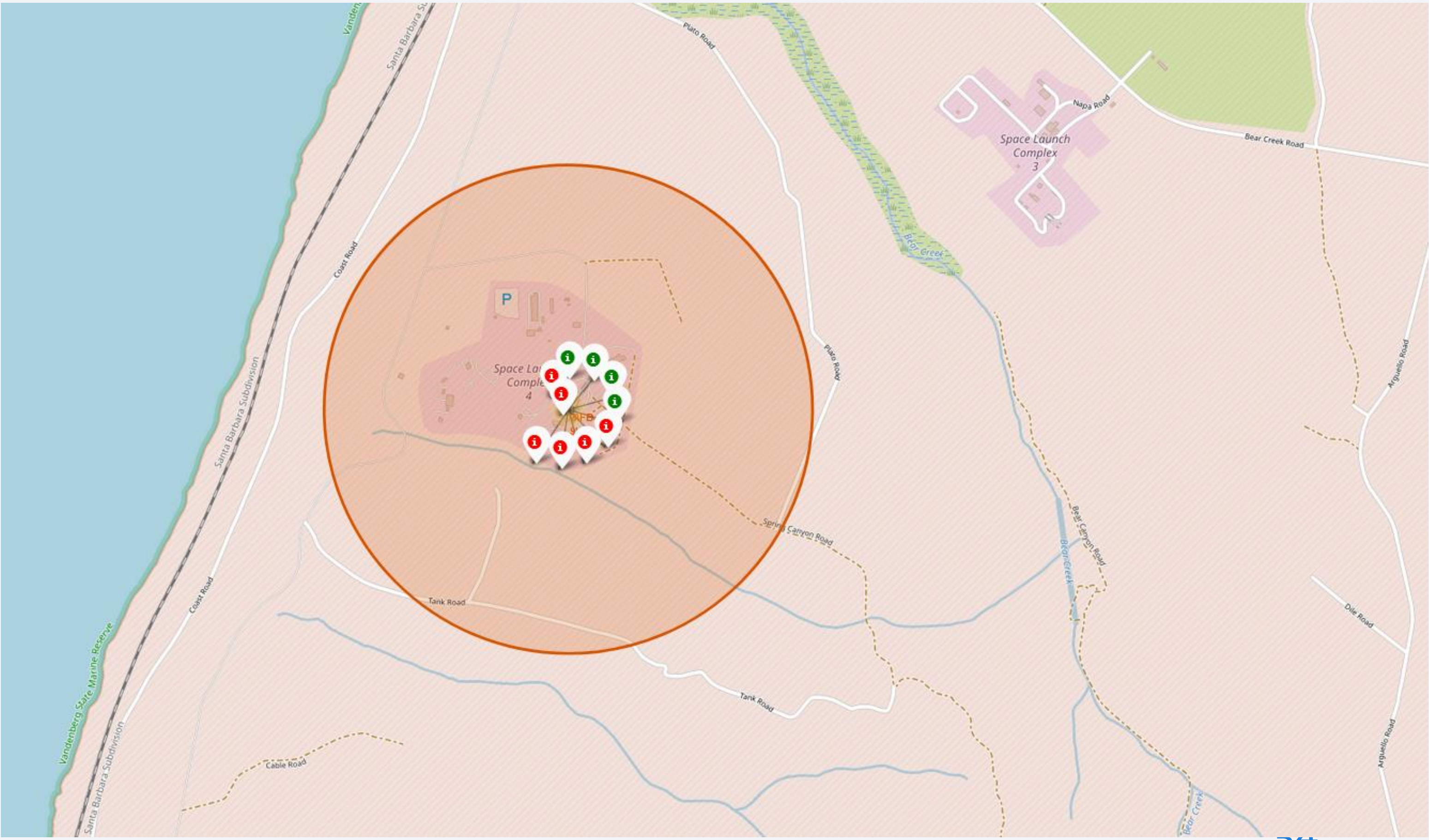
- Interactive map which showcases the locations of the launch sites
- Multiple are located in Florida, while a single one on the coast of California



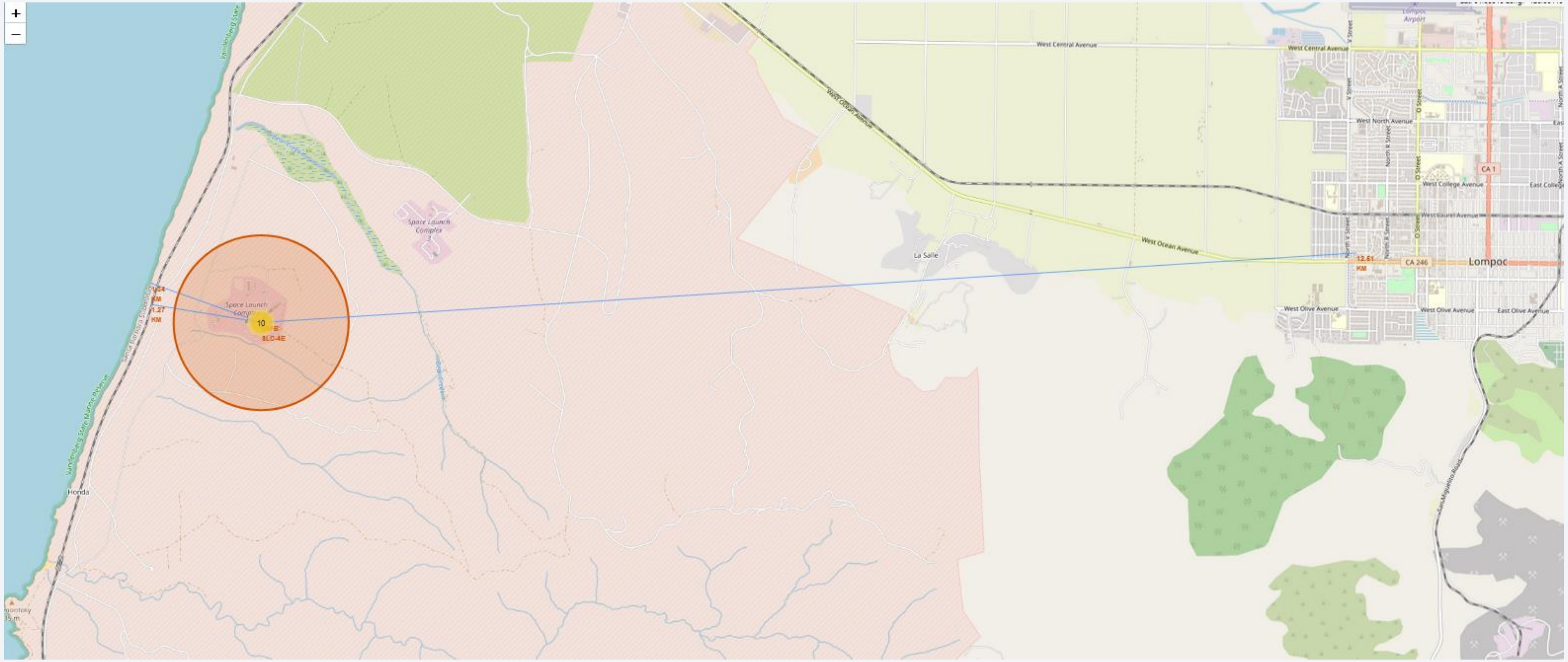
Marker Cluster featuring Success and Failures of Launches

- Zoom in screenshot of Folium map on the VAFB-SLC-4E launch site in California

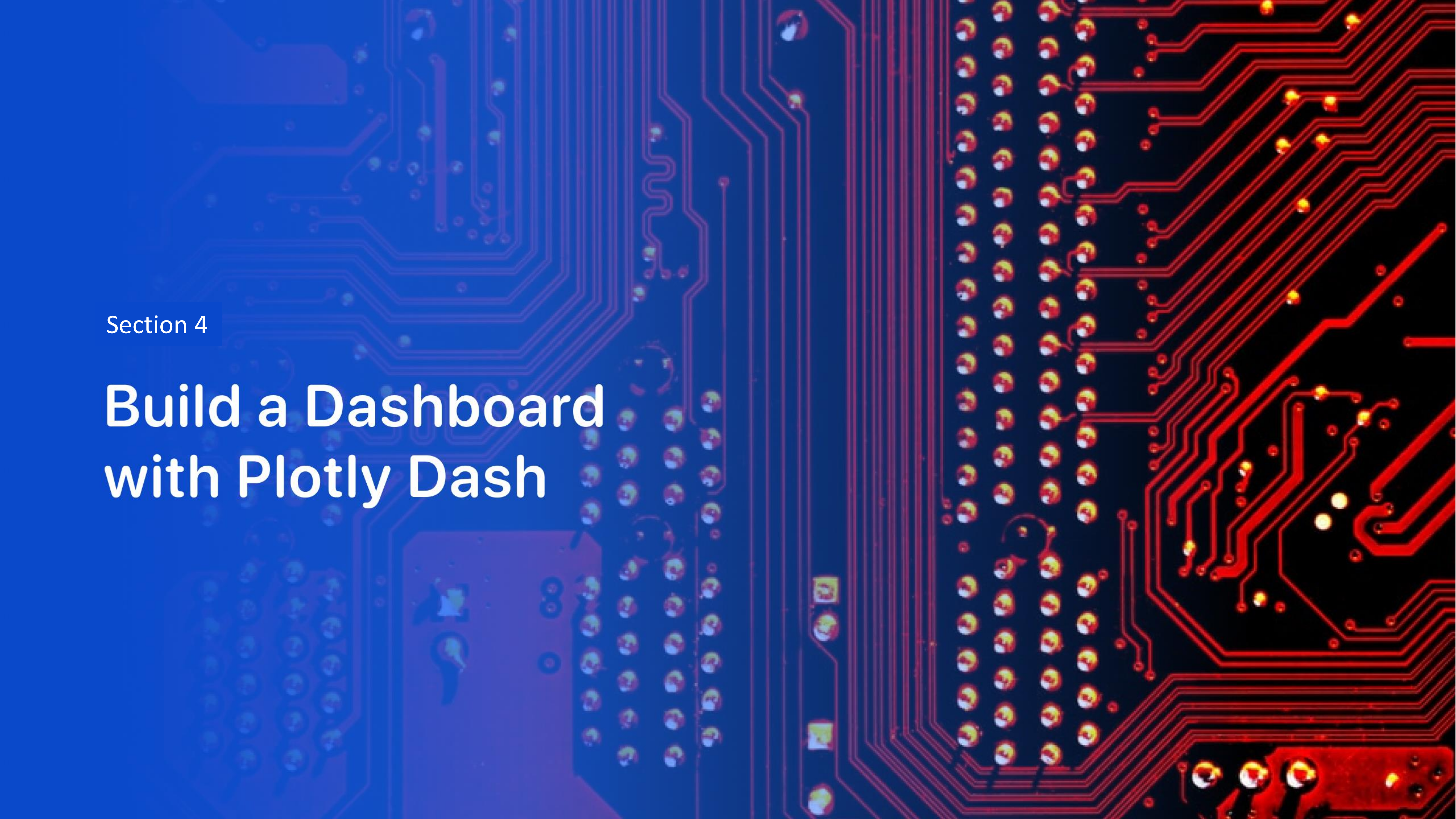
- Shown is a marker cluster expanded, showing the successes in green and the failures in red



Distance from Site to Coast, Railroad, and Nearest City



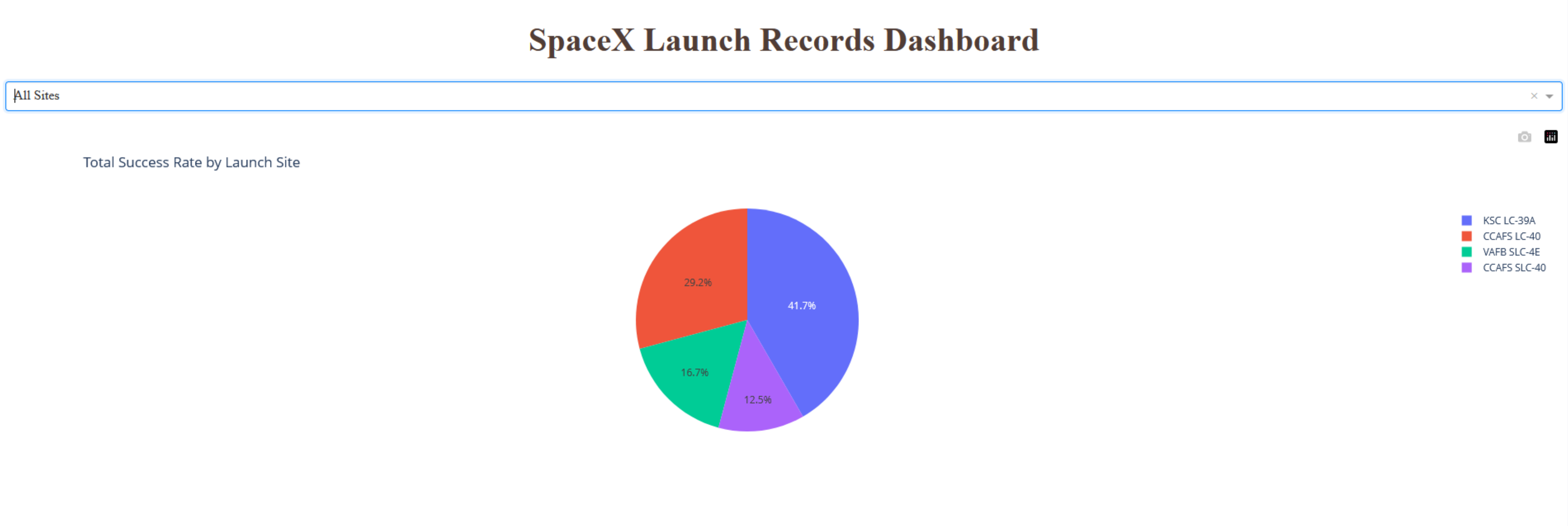
- On interactive map, drew lines to nearest coast (1.34 km), nearest railroad (1.27 km) and nearest city (12.61 km) Can see easier on interactive map
- Launch sites seem to be fine being located adjacent to the coast and railroads, but are distanced from populated towns/cities



Section 4

Build a Dashboard with Plotly Dash

Pie Chart from Dashboard

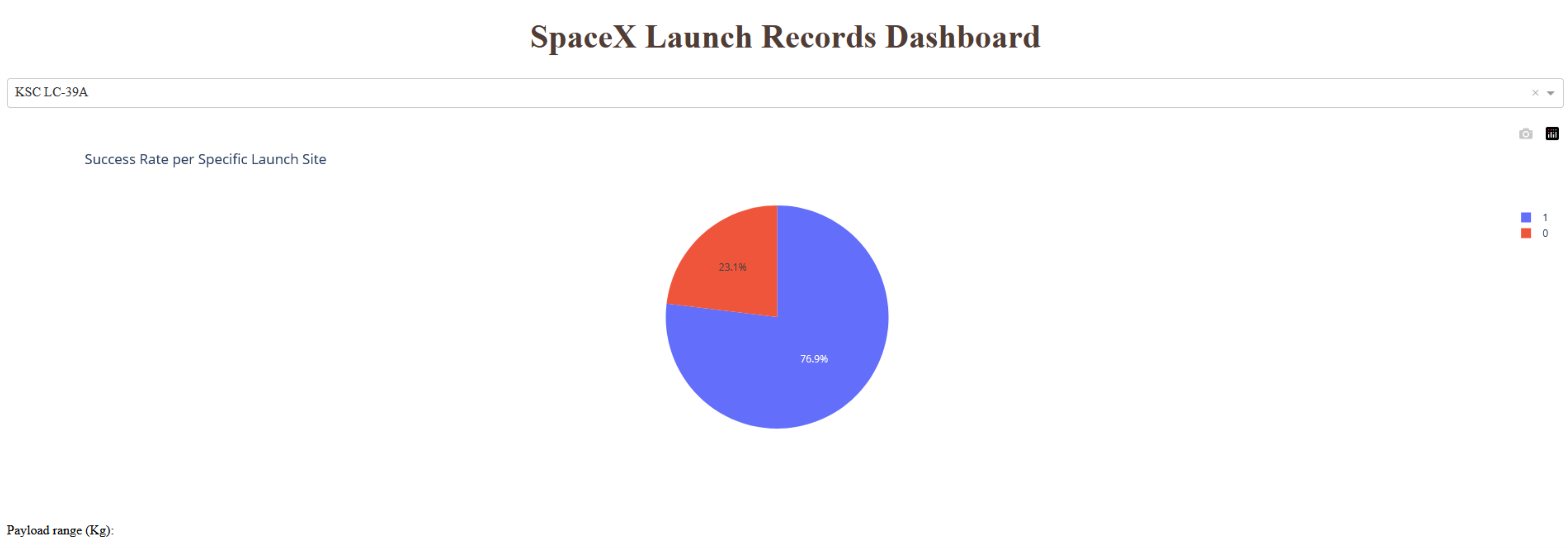


SpaceX Launch Records Dashboard

All Sites

- Screen shot from Dashboard showing a pie chart generated using the dropdown menu
- Pie chart shows the fraction of successful launches from each launch site

Pie Chart from Dashboard Selecting a Specific Launch Site



SpaceX Launch Records Dashboard

KSC LC-39A

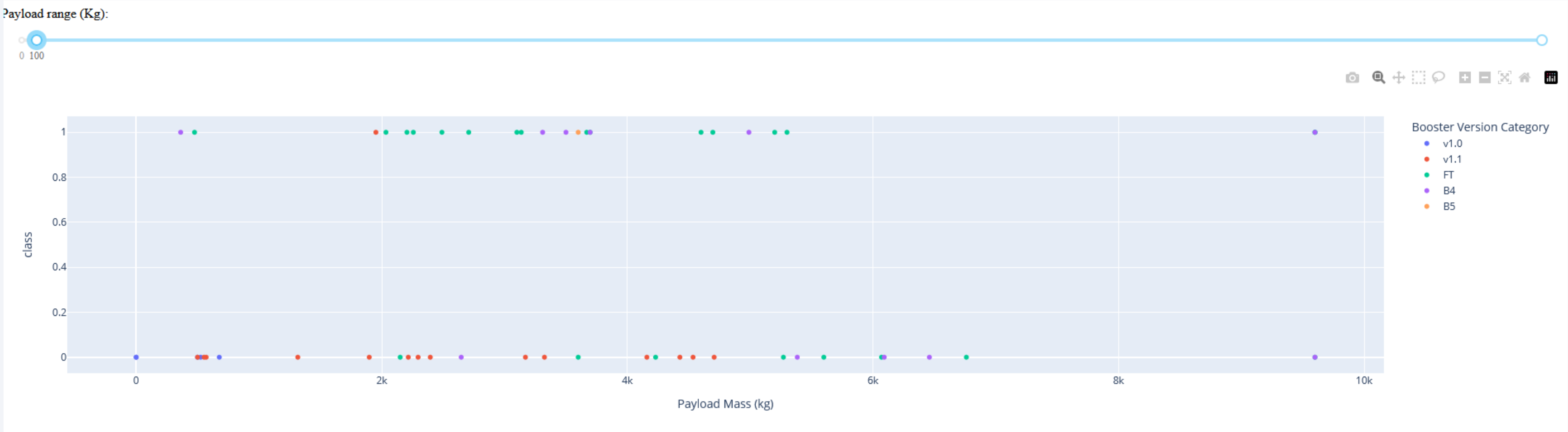
Success Rate per Specific Launch Site

Payload range (Kg):

- Screen shot from Dashboard selecting a single launch site (selecting site with the highest success rate)

- Pie chart now shows the percentage of success (blue) and failures (red)

Scatter Plot from Dashboard with Payload Mass Slider



- Scatter plot showing successes (class=1) and failures (class=0) for different payload masses and booster categories
- FT booster version has the highest rate of success
- The launches with payload masses between 2000kg and 6000kg have the highest success rate

Section 5

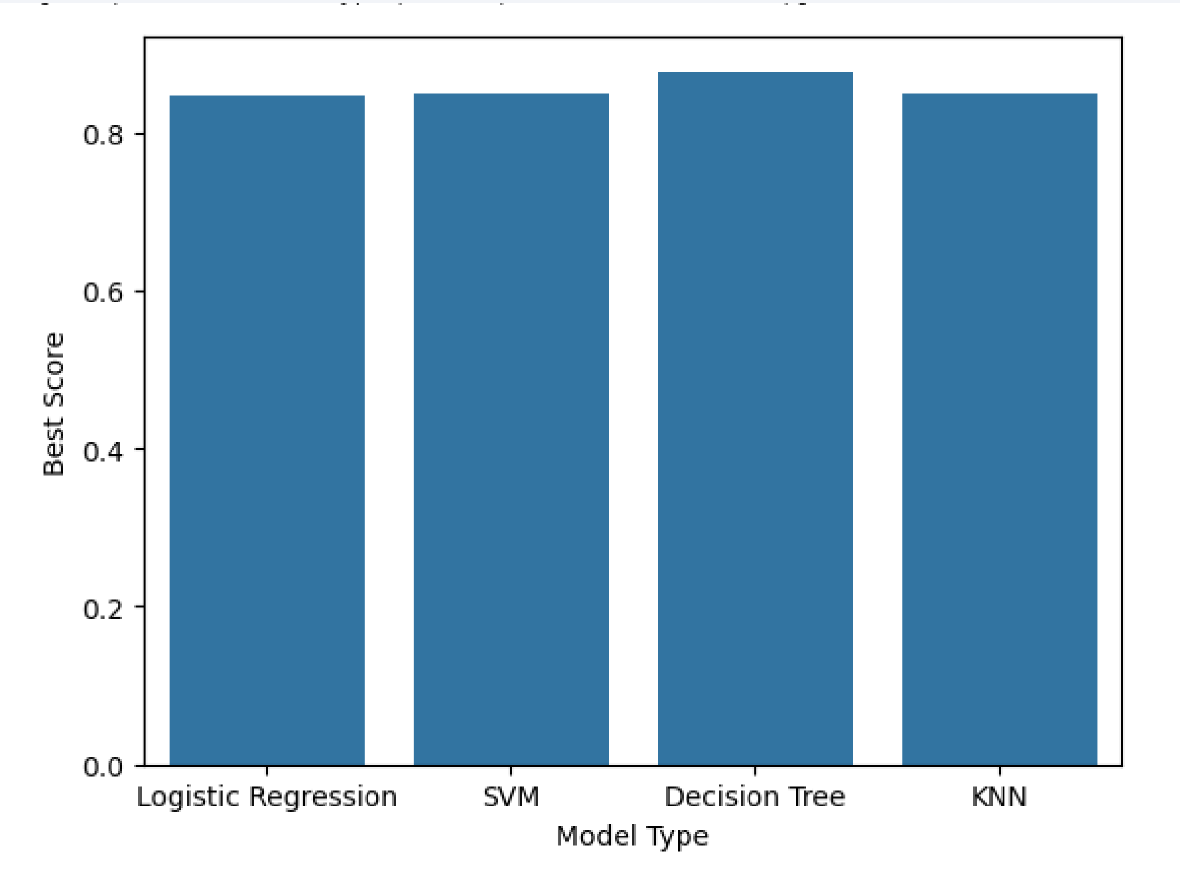
Predictive Analysis (Classification)

Classification Accuracy

- Four models were tested

- The best scores from each model was compared and is shown in the bar chart

- All models have similar performance, but Decision Tree had a slightly higher score



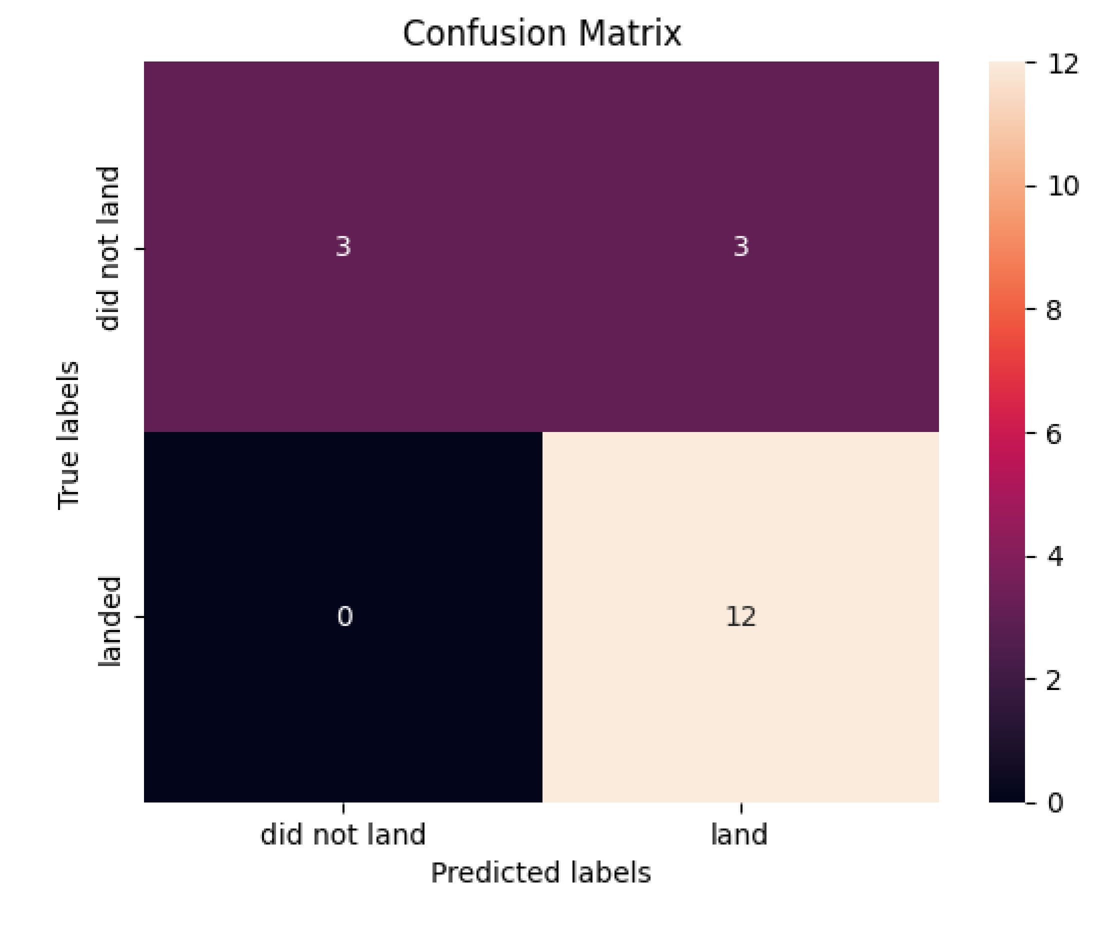
Confusion Matrix – Decision Tree

- Shown is the confusion matrix for the decision tree

- All four models had the same confusion matrix

- All true landed in the test sample was correctly predicted to be landed

- There are however false positives, some 'did not land' values in the test sample were predicted to have landed



Conclusions

- In this project the spacex data for the Falcon 9 first stage landing was explored and used to develop models to predict the success or failure based on input features

- All four models tested perform relatively well, and none were significantly better than the other, with the Decision Tree performing slightly better

- These models all had false positives, with the features used not completely sufficient to fully disentangle failed launches from successes

Thank you!