IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Allison Mueller
11-10-2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using SpaceX API
  - Data collection using webscraping
  - Data wrangling
  - Exploratory data analysis (EDA) with SQL
  - Exploratory data analysis (EDA) with data visualization
  - Interactive map using folium
  - Interactive dashboard using plotly dash
  - Machine learning predictions
- Summary of all results
  - EDA results
  - Screenshots of interactive components
  - Machine learning prediction results

Introduction

- Project background and context
  - Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. In this lab, you will create a machine learning pipeline to predict if the first stage will land.
- Problems you want to find answers
  - What factors determine that the Falcon 9 launch will successfully land?
  - What is the interaction between features and how does it contribute to successful landings?
  - What are the optimal conditions for ensuring a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API and webscraping from Wikipedia
- Perform data wrangling
  - Data was cleaned and one-hot encoding was applied to categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

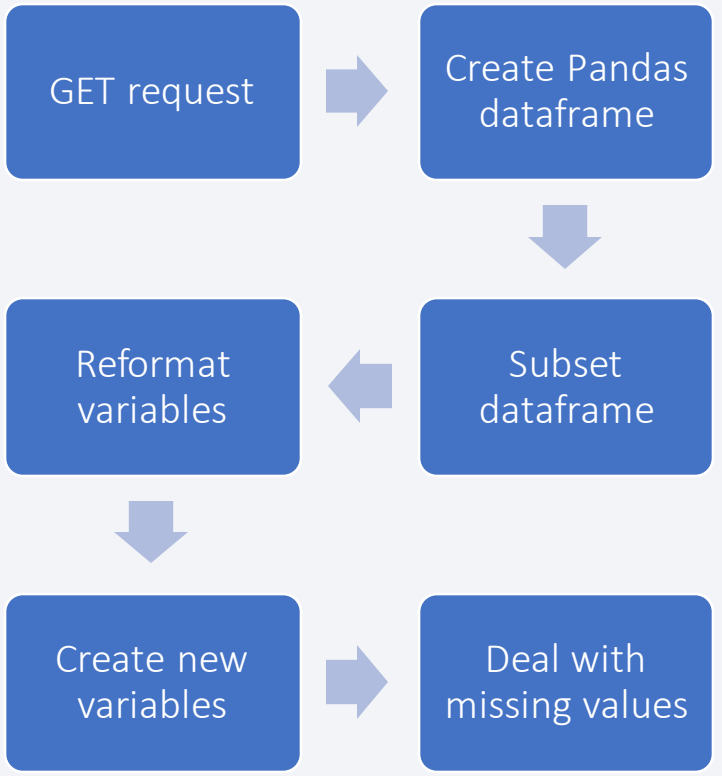
Data Collection

Data was collected from two sources:

- SpaceX API
  - Requested and parsed SpaceX launch data using the GET request
  - Decode response object using .json() and .json_normalize() to convert to a Pandas dataframe
  - Subset data, format variables, and create new variables for analysis
- Webscraping
  - Request the Falcon9 Launch Wiki page from its URL using HTTP GET request
  - Extract all column and value names from the HTML table header
  - Create a dataframe by parsing the launch HTML tables

Data Collection – SpaceX API

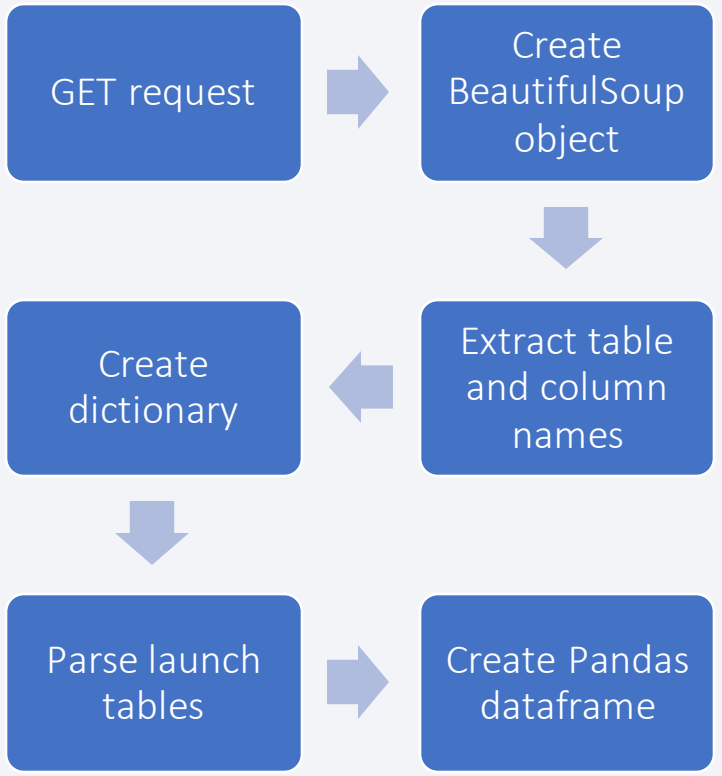
- Request and parse data using a GET reuqest
- Use .json() and .json_normalize() to create Pandas dataframe
- Subset data to limit dataframe to necessary variables
- Create new variables for analysis
- Replace missing values to complete dataset



SpaceX API Notebook

Data Collection - Scraping

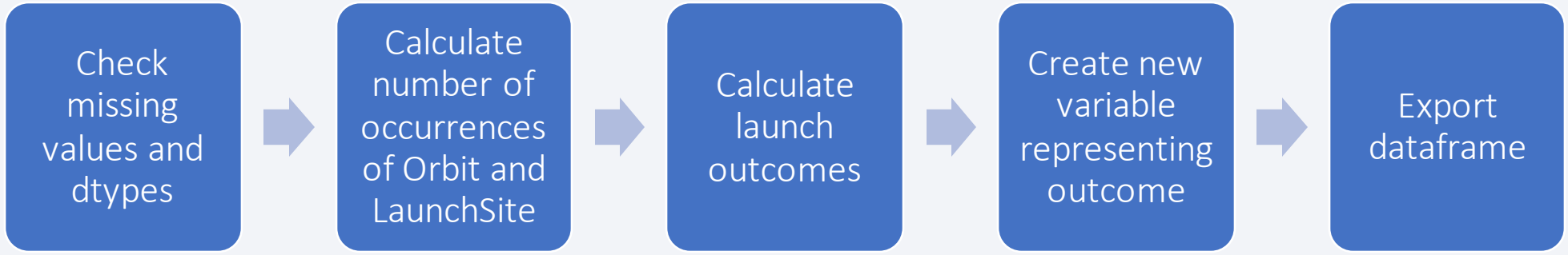
- Perform GET method to request Falcon 9 Launch HTML page
- Create BeautifulSoup object to parse HTML data
- Extract table and variable names from soup object
- Create empty dictionary and fill with parsed launch variables
- Convert to Pandas dataframe



Webscraping Notebook

Data Wrangling

- Check percentage of missing values and data types for each variable
- Use .value_counts() on `LaunchSite`, `Orbit`, and `Outcome`
- Create new variable `Class` to represent the outcome of each launch, 0=bad, 1=good
- Use .mean() on `Class` to get success rate



Data Wrangling Notebook

EDA with Data Visualization

- Scatterplot
  - Used to visualize the relationship between a numeric variable and a categorical variable
    - FlightNumber and LaunchSite
    - PayloadMass and LaunchSite
    - FlightNumber and Orbit
    - PayloadMass and Orbit
- Bar chart
  - Used to compare distribution of data across categorical variables
    - Succes rate (`Class`) for each Orbit type
- Line chart
  - Used to visualize date over a period of time
    - Success rate (`Class`) over time (`Date')

EDA with Data Visualization Notebook

EDA with SQL

- SQL Queries Performed:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch site names begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass using a subquery
9. List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
10. Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

EDA Using SQL Notebook

Build an Interactive Map with Folium

- Added four circles and markers to the map to show the four different launch sites across the United States.
- Added a marker cluster to all four locations to show successful and failed launches from each site.
- Added a line from VAFB SLC-4E to the closest coastline to show the distance
- Added a line from VAFB SLC-4E to the nearest city to show how close it is to other places.

Launch Site Analysis Using Folium Notebook

Build a Dashboard with Plotly Dash

- Added a pie chart to show ratio of successful launches for all sites combined
- Created dropdowns to view ratio of successful launches for each individual site
- Added a scatterplot to show the correlation between payload and success by booster type for all sites and each individual site

Link to Dash

Link to notebook

Predictive Analysis (Classification)

- Created four machine learning models using different methods:
  - Logistic Regression, SVM, Decision Tree, K Nearest Neighbors
- Split data into training and testing sets
- Fit model to training data and determine best parameters
- Determine accuracy score using test data

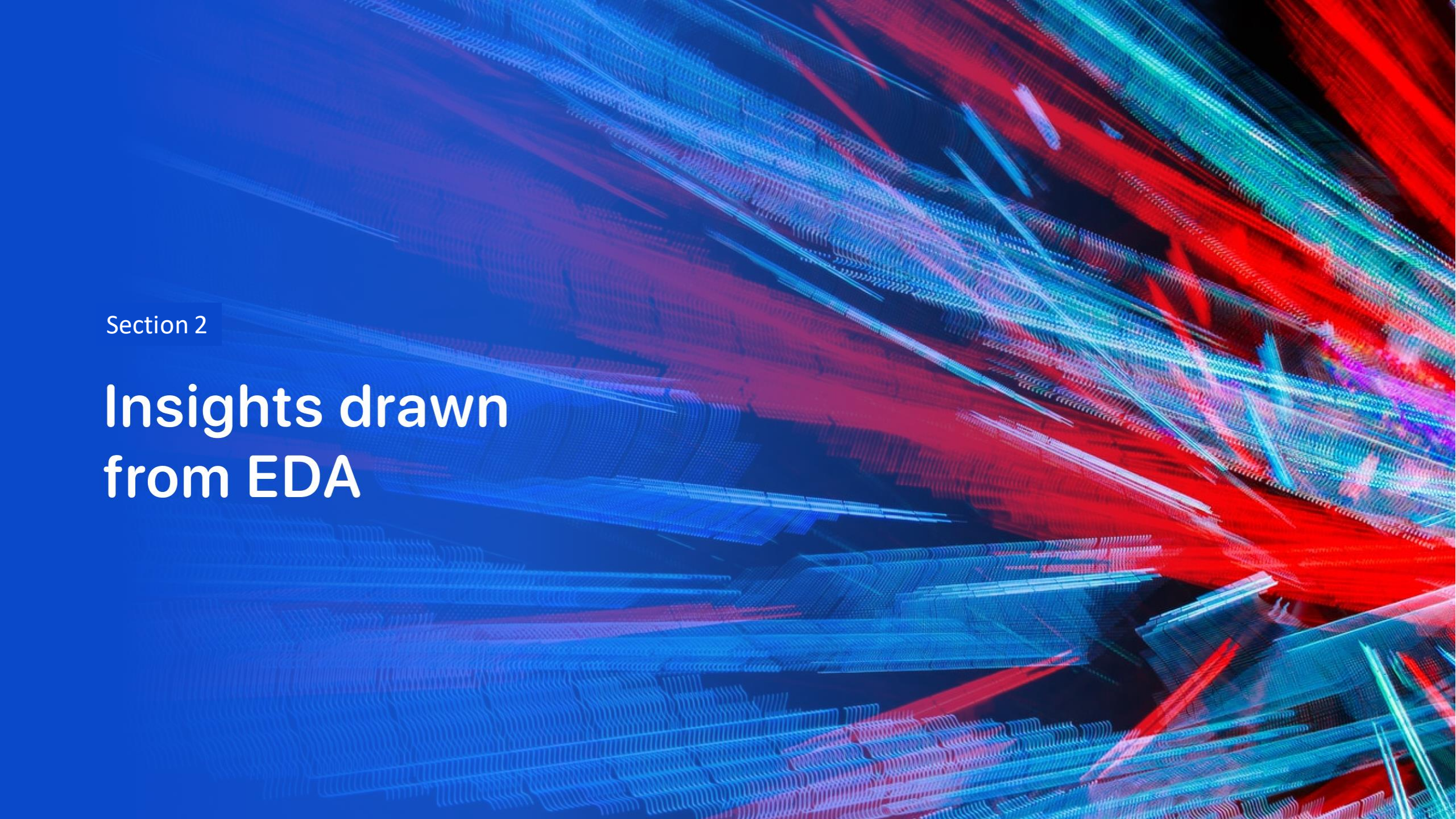
Machine Learning Model Notebook

Results

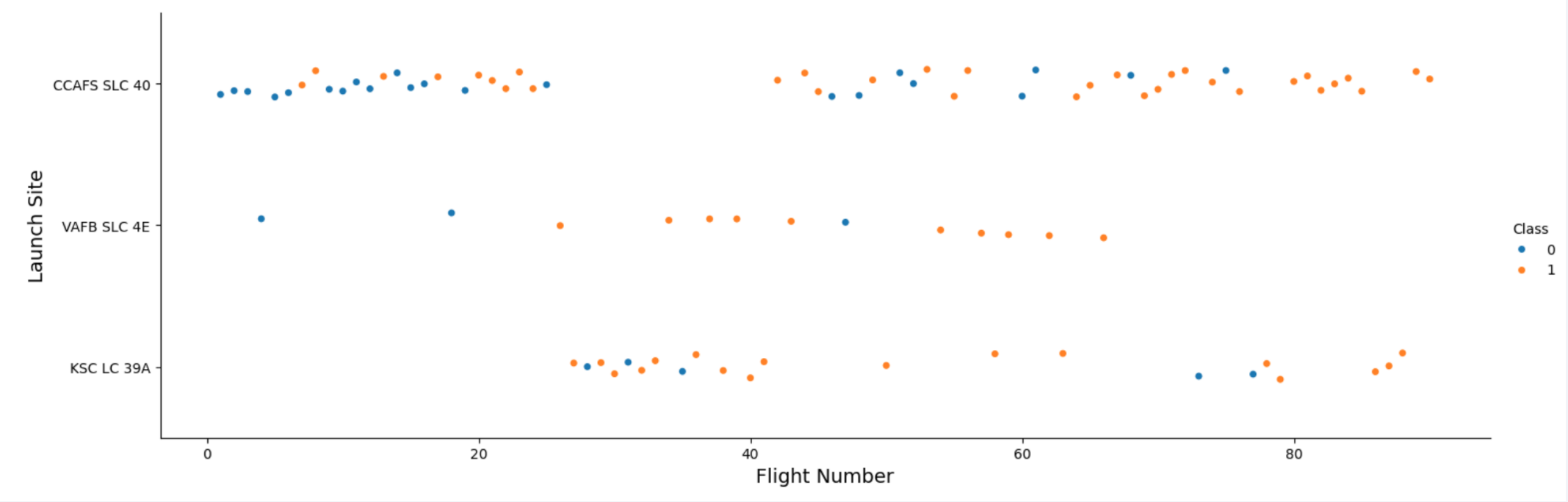
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

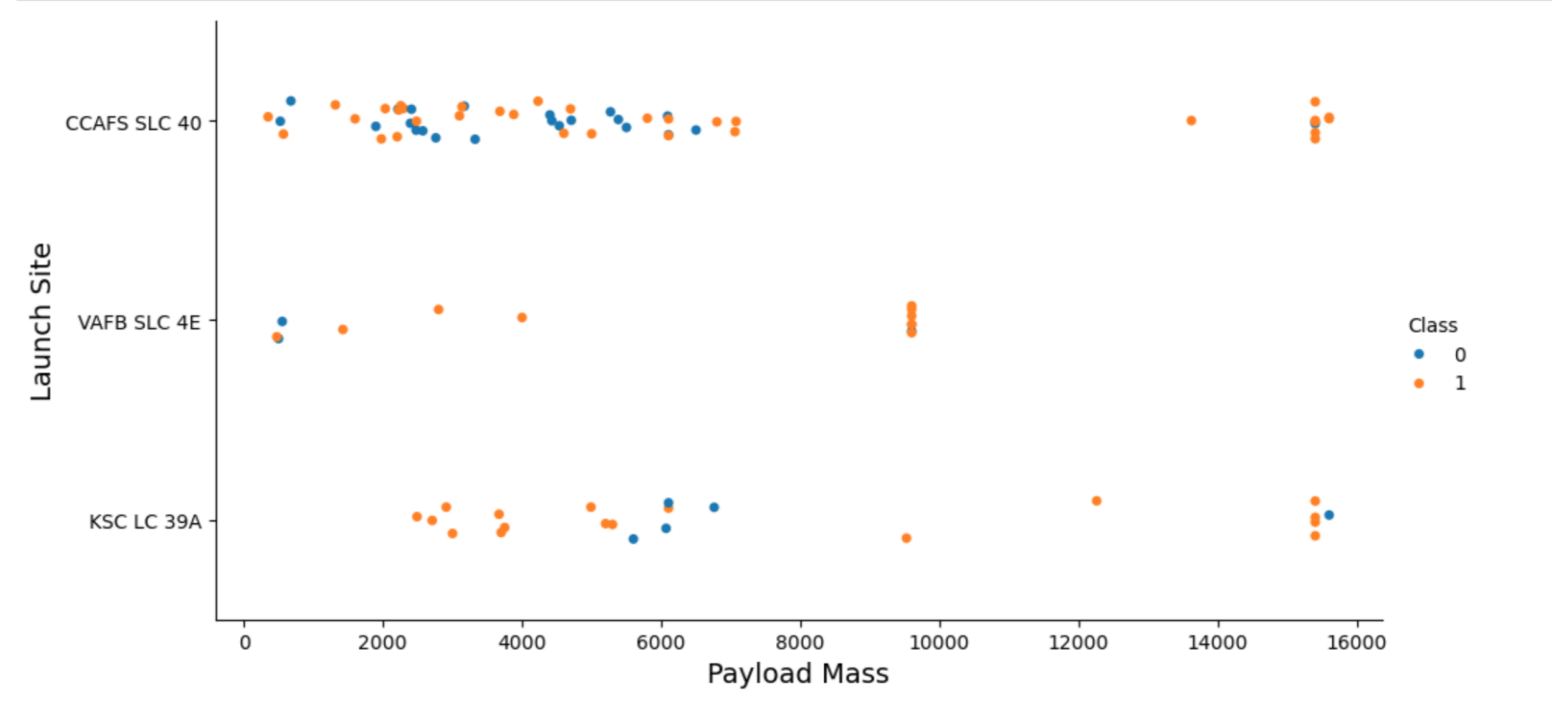


Flight Number vs. Launch Site



As the number of flights increases, there is a greater amount of successful launches for all three sites.

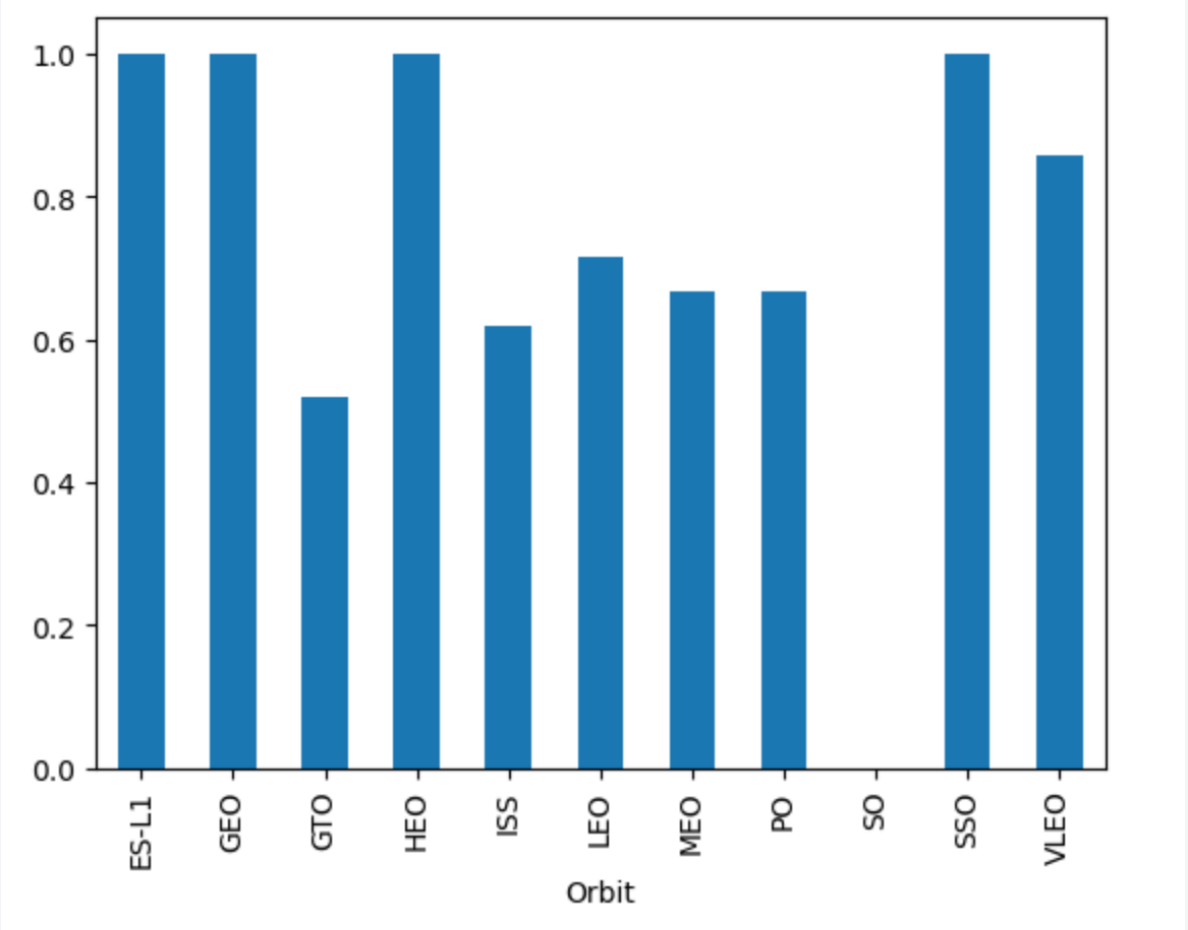
Payload vs. Launch Site



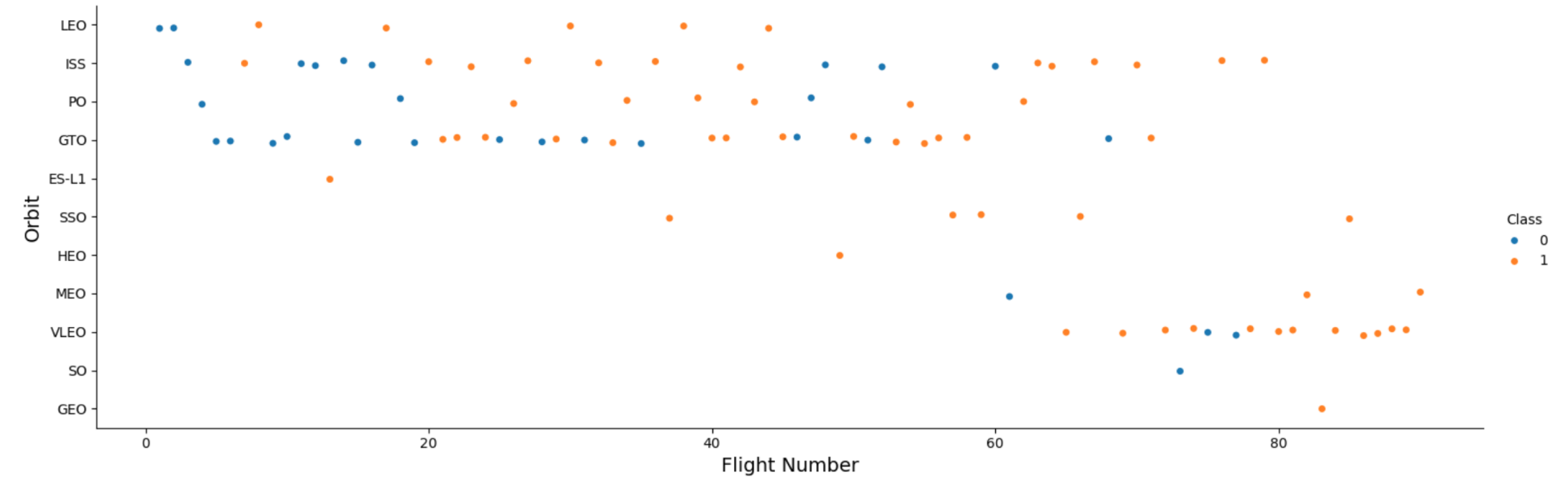
- Success rate is greatest for launches with the highest payload for launch site CCAFS SLC-40
- For site VAFB SLC-4E, there are no launches with a payload greater than 10,000 kg

Success Rate vs. Orbit Type

- The orbits that have the highest success rates include:
  - ES-L1
  - GEO
  - HEO
  - SSO

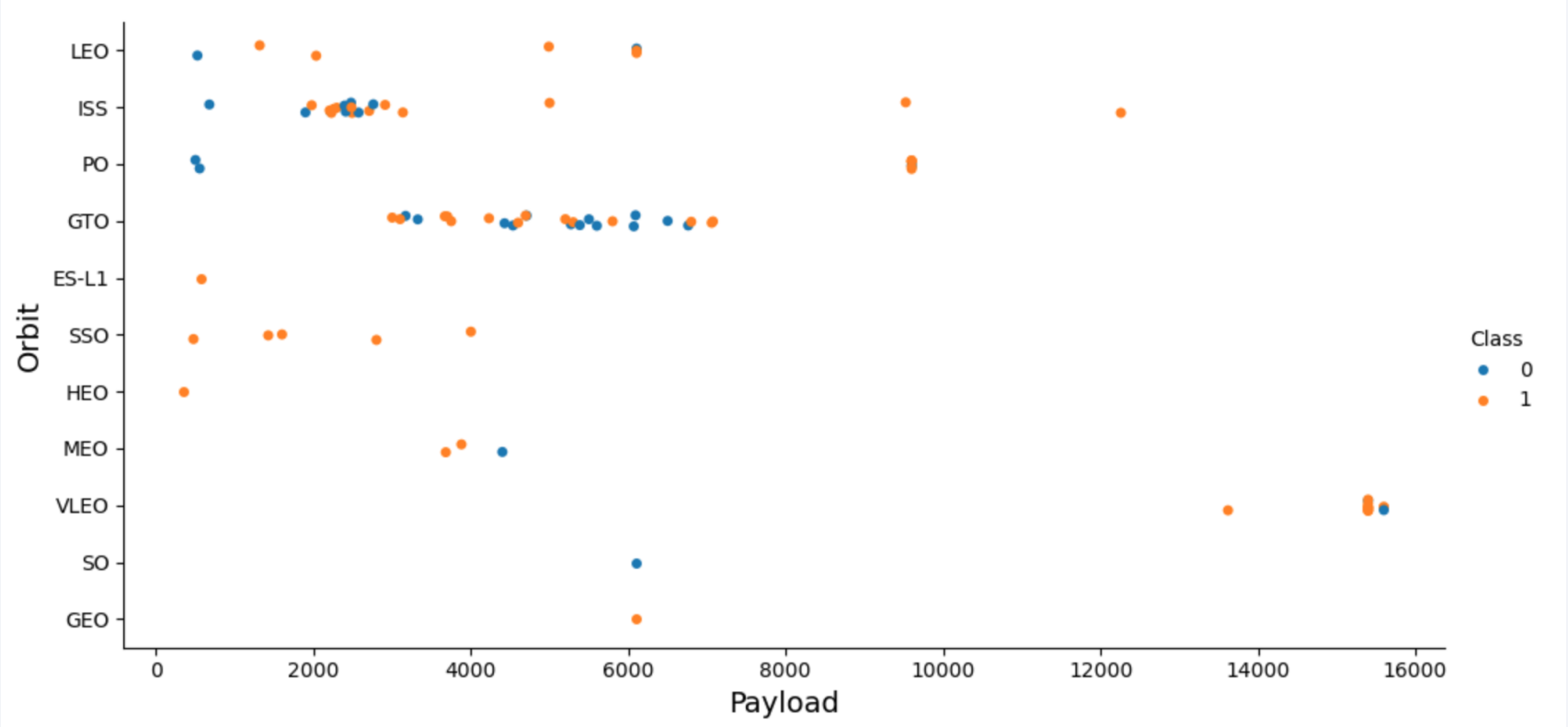


Flight Number vs. Orbit Type



- For orbit LEO, success appears to be related to the number of flights
- There appears to be no relationship between number of flights and GTO orbit

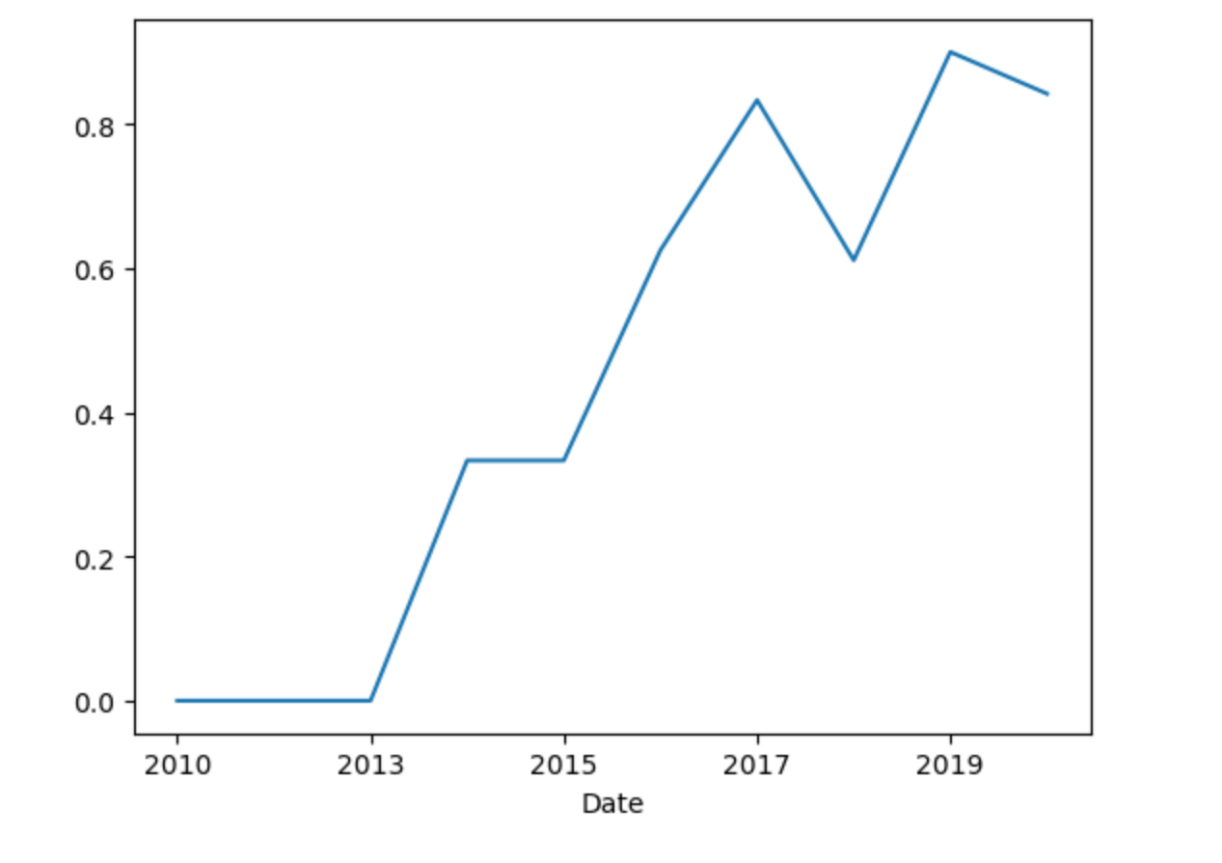
Payload vs. Orbit Type



- For heavier payloads, the orbits with the highest successful landing rate are PO, LEO, and ISS
- For GTO, it is not possible to tell if there are more successful or unsuccessful landings because both are present.

Launch Success Yearly Trend

- From 2010 to 2013, success rate remained constant
- Slight decrease in success in 2018
- Since 2013, success rate has been in increasing



All Launch Site Names

- There are four unique launch sites

```
]: %sql SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE2;
 * 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'

- The first 5 records where launch sites begin with `CCA`

```
%sql SELECT * FROM SPACEXTABLE2 WHERE "Launch_Site" LIKE "CCA%" LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload mass carried by NASA (CRS) was 45,596 kilograms

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS "Total Payload Mass" FROM SPACEXTABLE2 WHERE "Customer" = "NASA (CRS)";
```

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

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2,928.4 kilograms

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass" FROM SPACEXTABLE2 WHERE "Booster_Version" = "F9 v1.1";
```

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

Average Payload Mass
2928.4

First Successful Ground Landing Date

- The date of the first successful ground landing was December 22, 2015.

```
%sql SELECT MIN("Date") FROM SPACEXTABLE2 WHERE "Landing_Outcome" = "Success (ground pad)";
```

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

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters that successfully landed on a drone ship that had a payload mass between 4000 and 6000 kilograms were F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE2
WHERE "Landing_Outcome"="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

Total Number of Successful and Failure Mission Outcomes

- There were 100 successful mission outcomes, and one failed mission outcome.

```
%sql SELECT "Mission_Outcome", COUNT(*) AS "total" FROM SPACEXTABLE2 GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

- There were 12 boosters that carried the maximum payload

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE2
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

2015 Launch Records

- There were two failed drone ship landings in 2015, and both came from launch site CCAFS LC-40. The first was on January 10, 2015, and the second was on April 14, 2015.

```
%%sql
SELECT substr("Date", 6, 2) AS "Month", "Date", "Booster_Version", "Launch_Site"
FROM SPACEXTABLE2
WHERE substr("Date",0,5) LIKE '%2015' AND "Landing_Outcome" = "Failure (drone ship)";
```

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

Month	Date	Booster_Version	Launch_Site
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

- Between June 4, 2010, and March 20, 2017, the most three most common landing outcomes in descending order were "no attempt", success (drone ship), and failure (drone ship)

```
%%sql
SELECT "Landing_Outcome", COUNT("Landing_Outcome") AS "Rank"
FROM SPACEXTABLE2
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY "Rank" DESC;
```

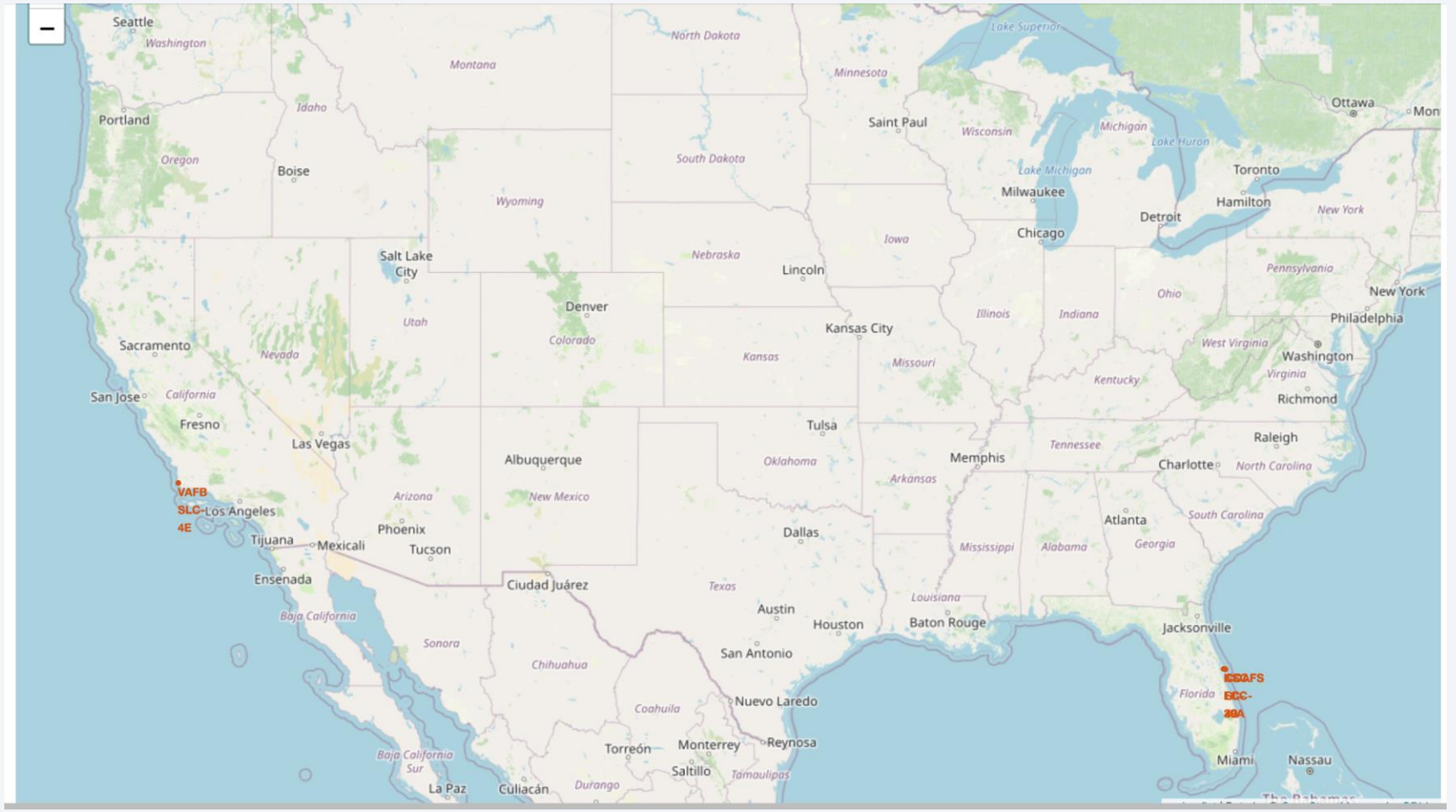
```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

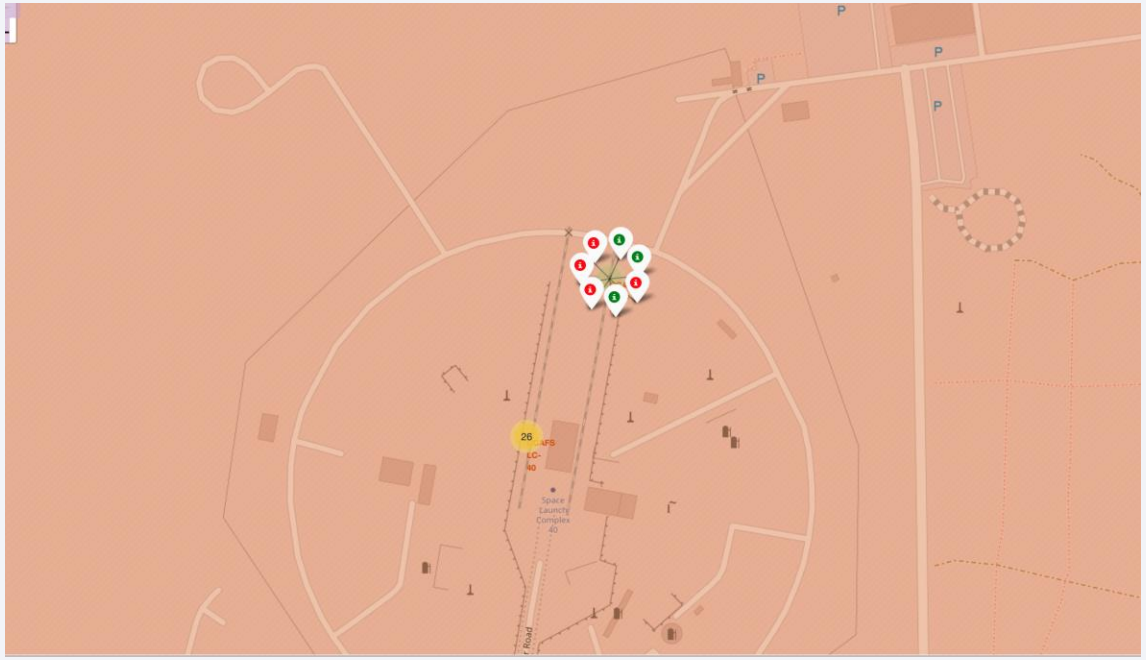
Launch Site Locations

- There is one launch site, VAFB SLC-4E, in California, and 3 launch sites in Florida.



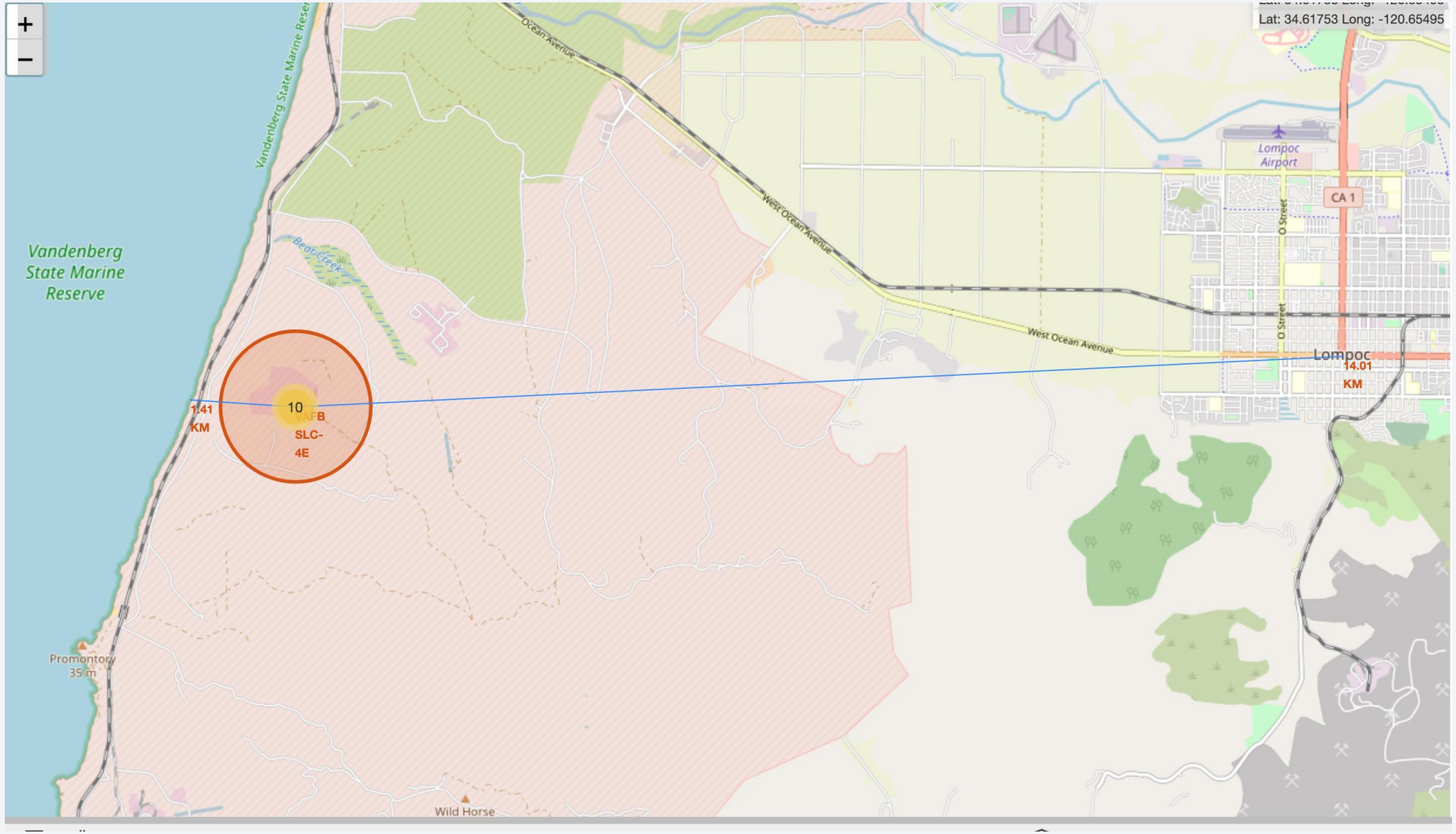
Launch Outcomes by Location

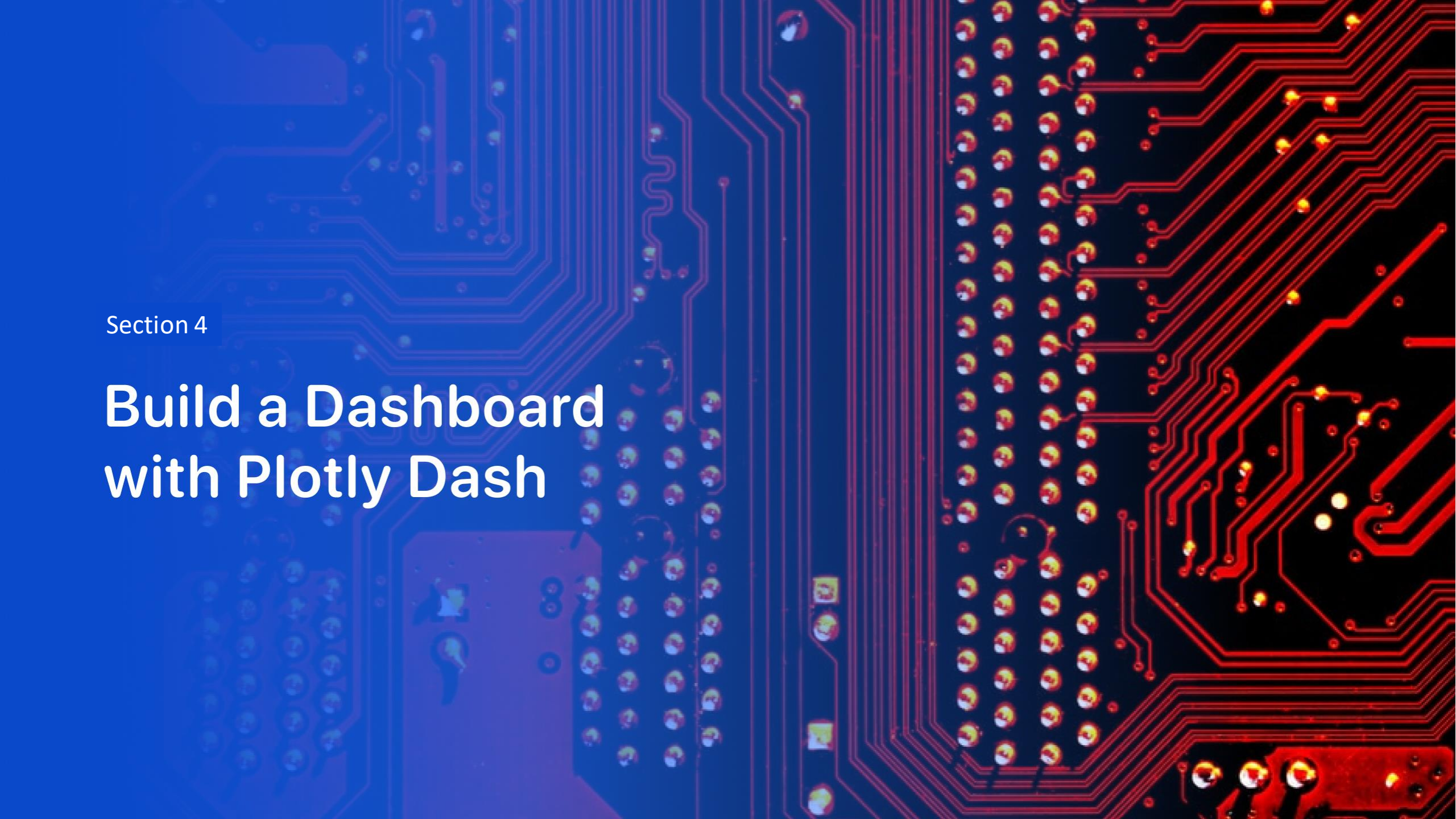
- Most of the launches occurred at the Florida sites, and the site in California had more unsuccessful outcomes than successful.



VAFB SLC-4E Proximity to Other Places

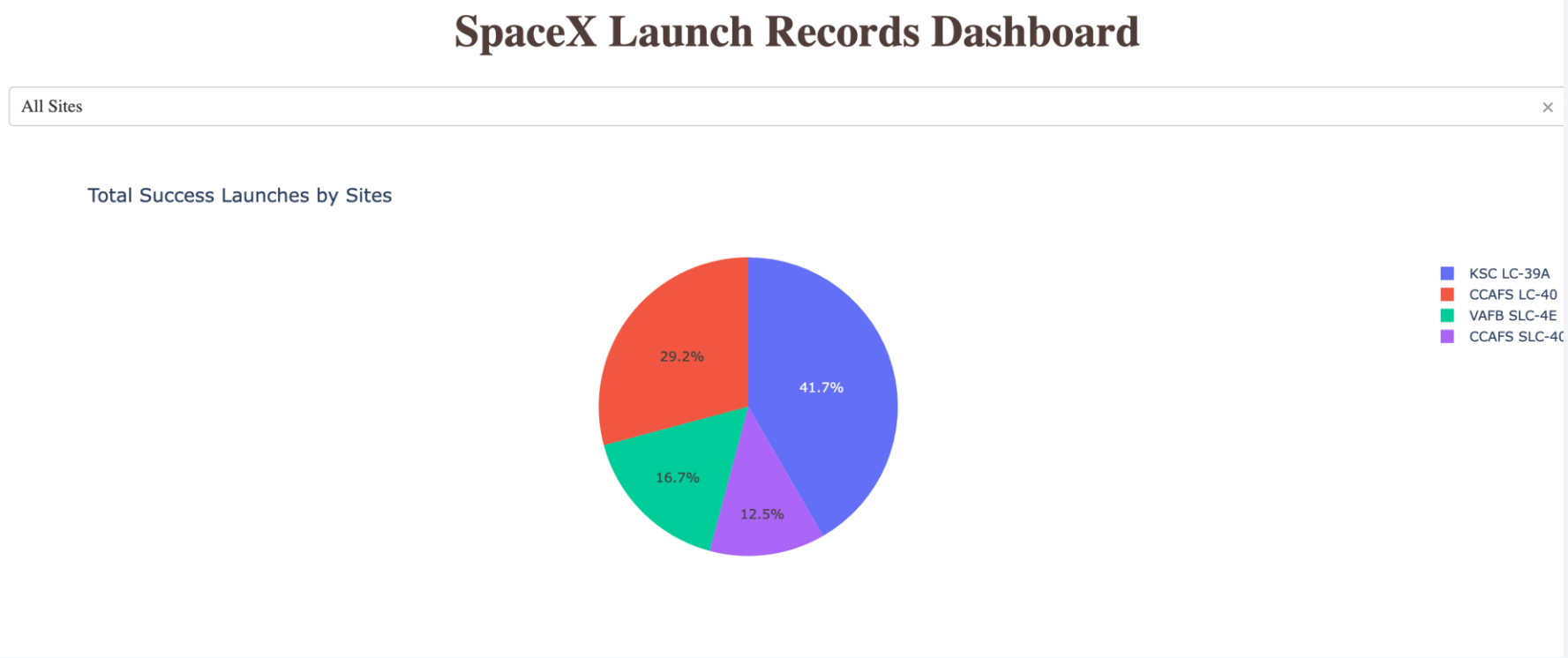
- VAFB SLC-4E is 1.41 KM from the nearest coast and 14.01 KM from the nearest city, Lompoc




Section 4
Build a Dashboard with Plotly Dash

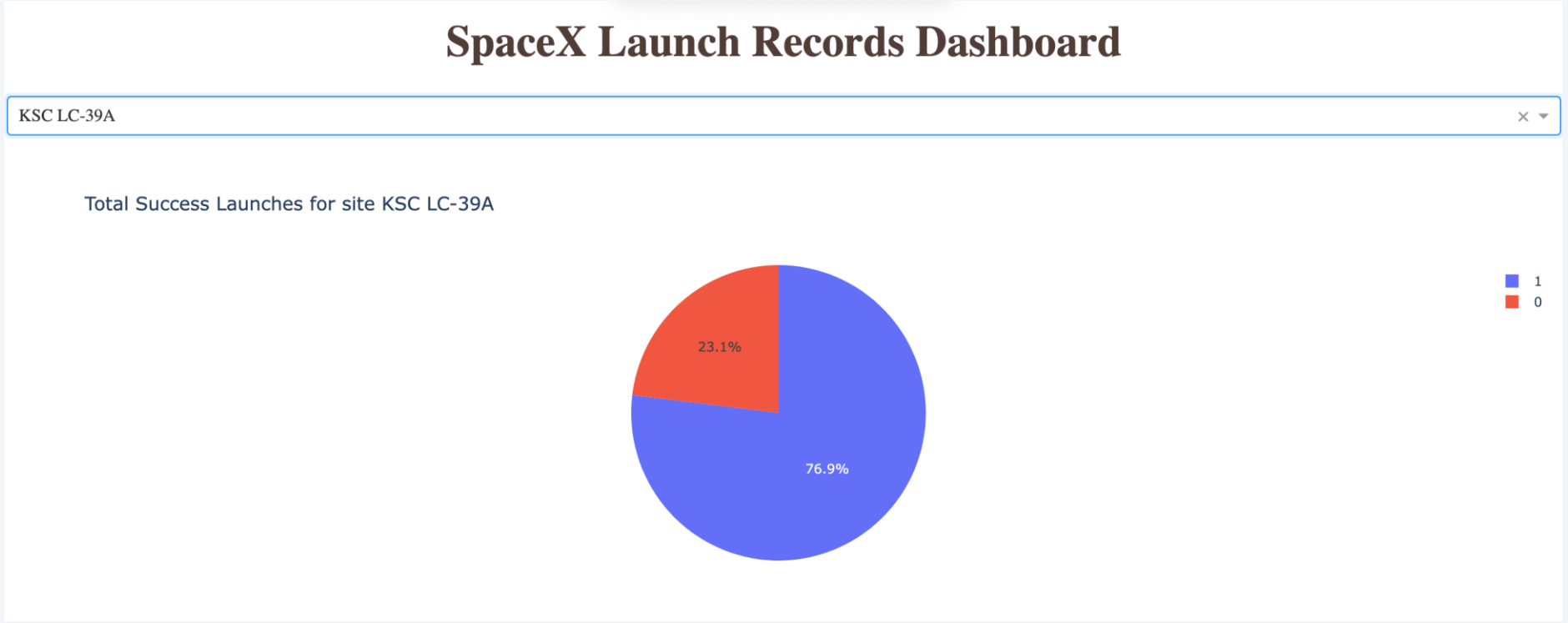
Success Count for All Launch Sites

- For all sites, the most successful was KSC LC-39A, and the second most successful was CCAFS LC-40



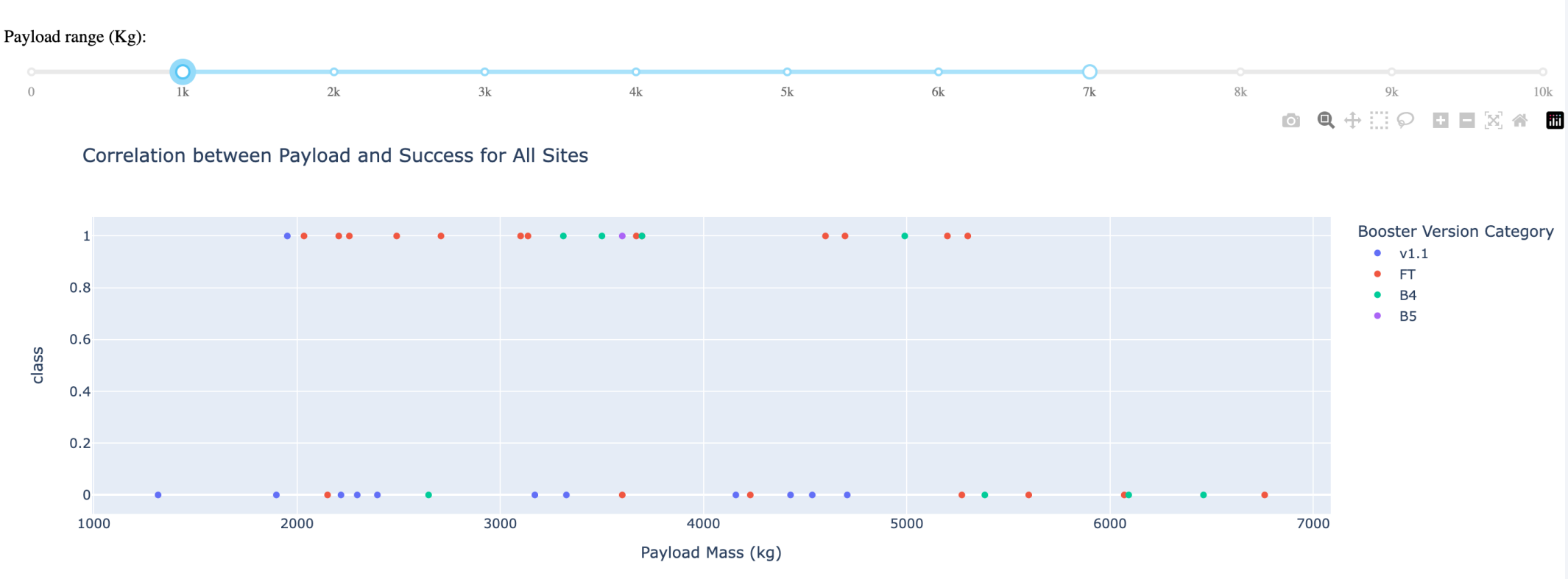
Most Successful Launch Site

- The most successful launch site was KSC LC-39A, with 76.9% (10 launches) of total launches being successful



Payload vs. Launch Outcome

- For the range between 1,000 kilograms and 7,000 kilograms, the most successful boosters were FT and B4 for all sites.

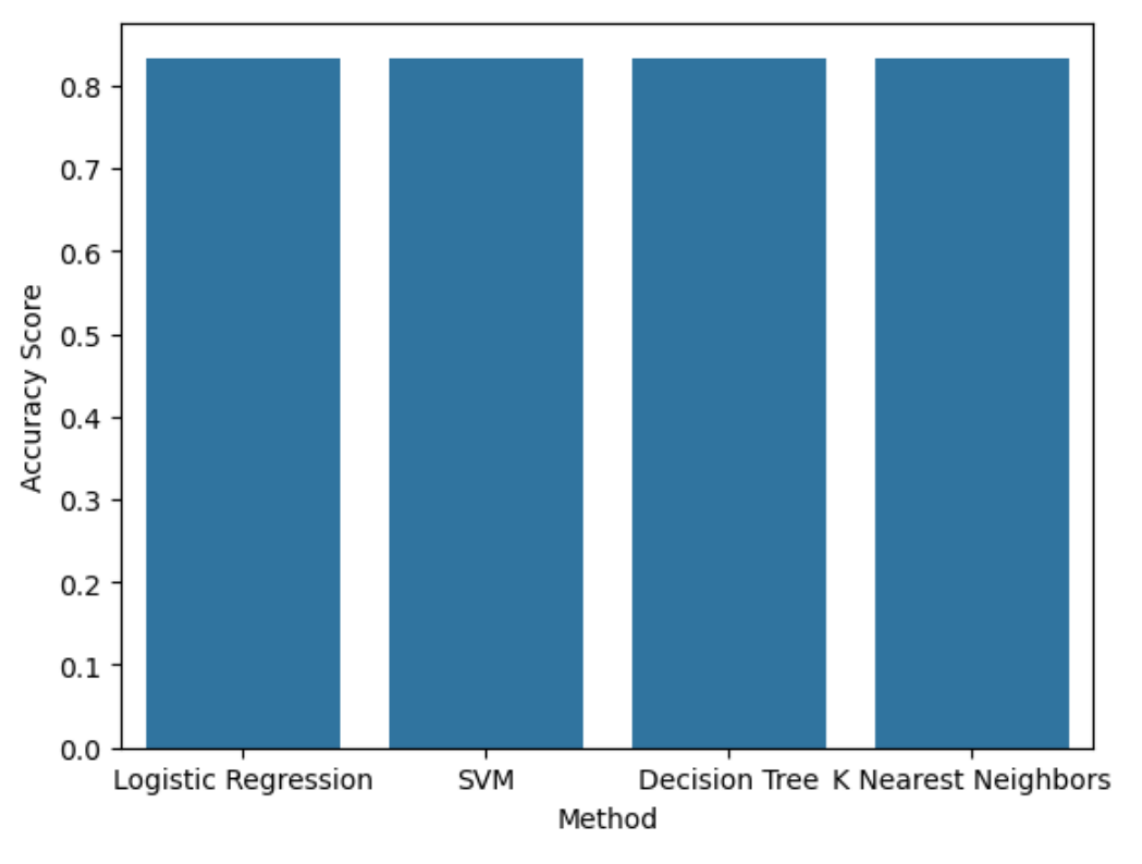


Section 5

Predictive Analysis (Classification)

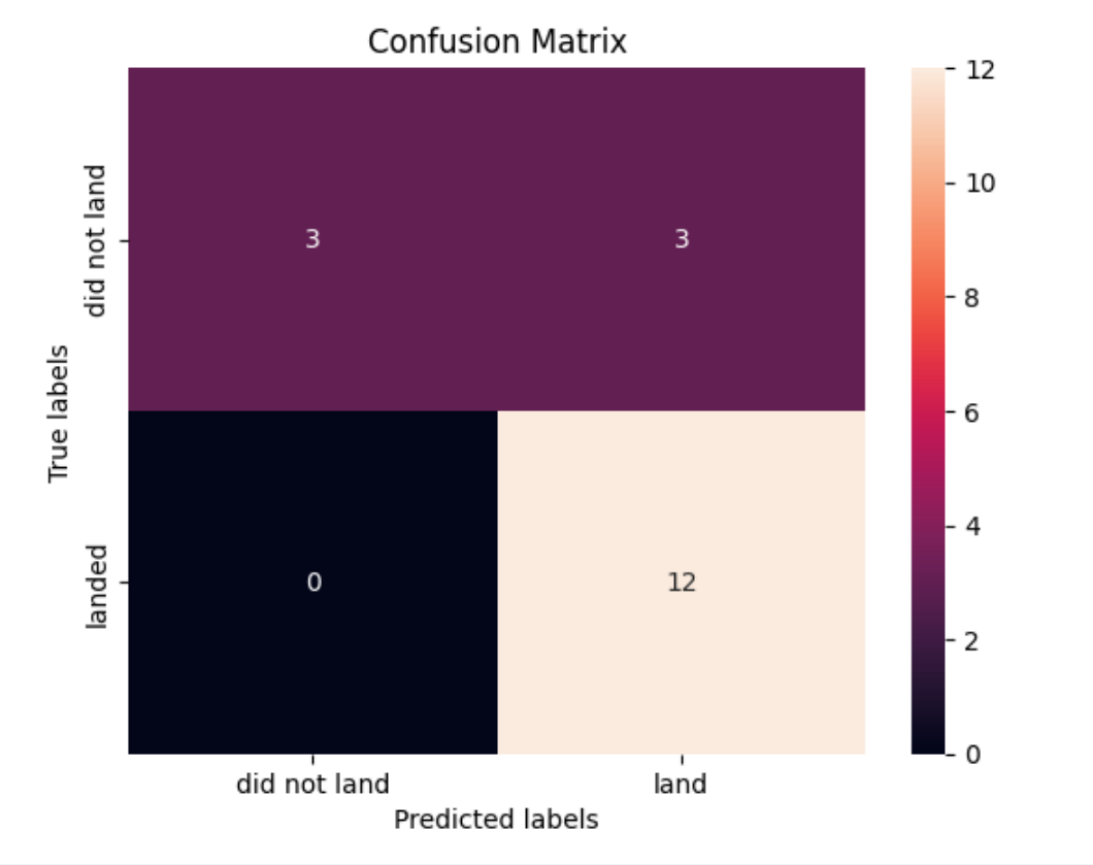
Classification Accuracy

- All four models had an accuracy score of approximately 83.3%



Confusion Matrix

- All models have the same confusion matrix



Conclusions

- Over time, especially since 2013, Space X has had increasingly more success with the launches of the Falcon 9 rocket.
- The site with the most successful launches was KSC LC-39A, we should further investigate the factors that contribute to their success
- All four machine learning models had virtually the same result with the testing data, these models should be refined further to improve their predictive abilities.

Appendix

- Data Set 1: Collecting Data with SpaceX API
  - https://github.com/allisonlmueller/spacex-ibm/blob/main/dataset_part_1.csv
- Data Set 2: Webscraping
  - https://github.com/allisonlmueller/spacex-ibm/blob/main/dataset_part_2.csv
- Data Set 3:
  - https://github.com/allisonlmueller/spacex-ibm/blob/main/dataset_part_3.csv
- Data Set 4: EDA using SQL
  - I had to change the date manually, so I created my own data set
  - https://github.com/allisonlmueller/spacex-ibm/blob/main/spacex_web_scraped.csv

Thank you!