IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fouad Atmeh

10-Jan-2024

Outline

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

Executive Summary

With space travel becoming more and more affordable, many companies are investing in space travel such as Virgin Galactic, Rocket Lab, Blue Origin and SpaceX.

SpaceX is providing the lowest price for sending an aircraft to the space compared to their competition, where the aim of this report is to try and predict the cost of a SpaceX flight to help with the decision to bid against them or not to bid.

- Multiple methodologies where used, where the Decision Tree methodology provided the best results.

Introduction

- Project background and context

  With space travel becoming more and more affordable, many companies are investing in space travel such as Virgin Galactic, Rocket Lab, Blue Origin and SpaceX.

  SpaceX is providing the lowest price for sending an aircraft to the space compared to their competition, where the aim of this report is to try and predict the cost of a SpaceX flight to help with the decision to bid against them or not to bid.

- Problems to find answers for:
  - Since the major cost of the launch is directly relevant to the return of the first stage rocket, we want to understand if it will successfully return or not
  - What inputs most impact the outcome of the returning of the first stage rocket.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through the SpaceX API and Wikipedia pages, Details are in the data collection slides.
- Perform data wrangling
  - Data was fairly clean, so little wrangling was done as described in the data wrangling slide.
- 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:

  1. The SpaceX API : This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

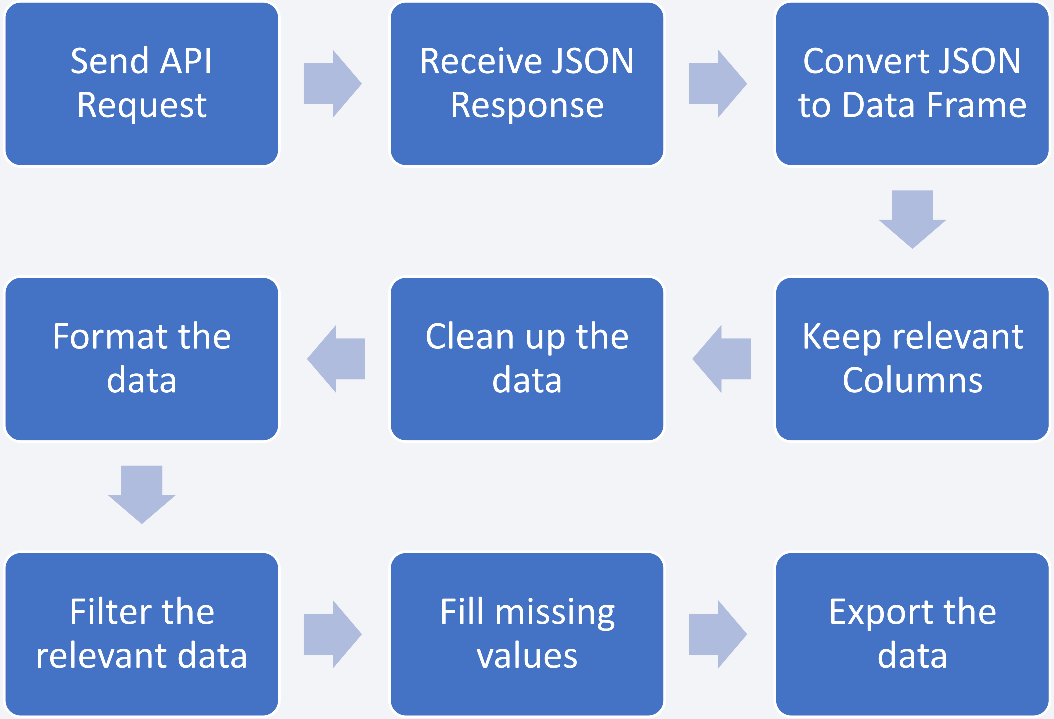
  2. Wikipedia's list of Falcon 9 and Falcon Heavy launches. The other data source for obtaining Falcon 9 Launch historical launch data, collected through web scraping of related Wiki pages.

- Data collection from each source is detailed in the following slides.

Data Collection – SpaceX API

- SpaceX API contains information about the Launches
- We will collect the data from this API, process it, and add it to a Data Frame according to the diagram on the right.

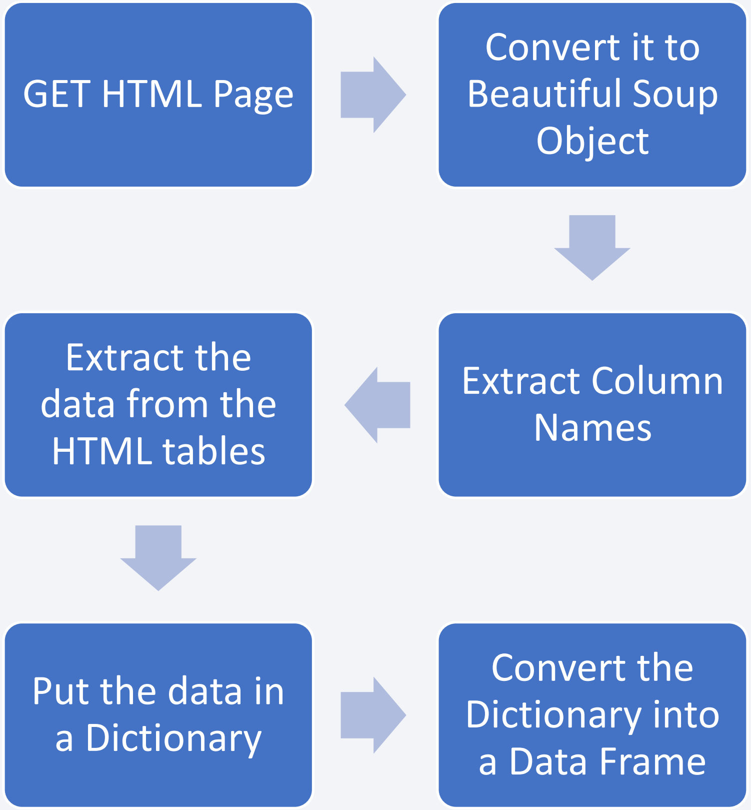
  - GitHub URL SpaceX API calls notebook: [Applied-DS-Capstone/W1-1 jupyter-labs-spacex-data-collection-api.ipynb at main · fouadatmeh/Applied-DS-Capstone · GitHub]()



Data Collection - Scraping

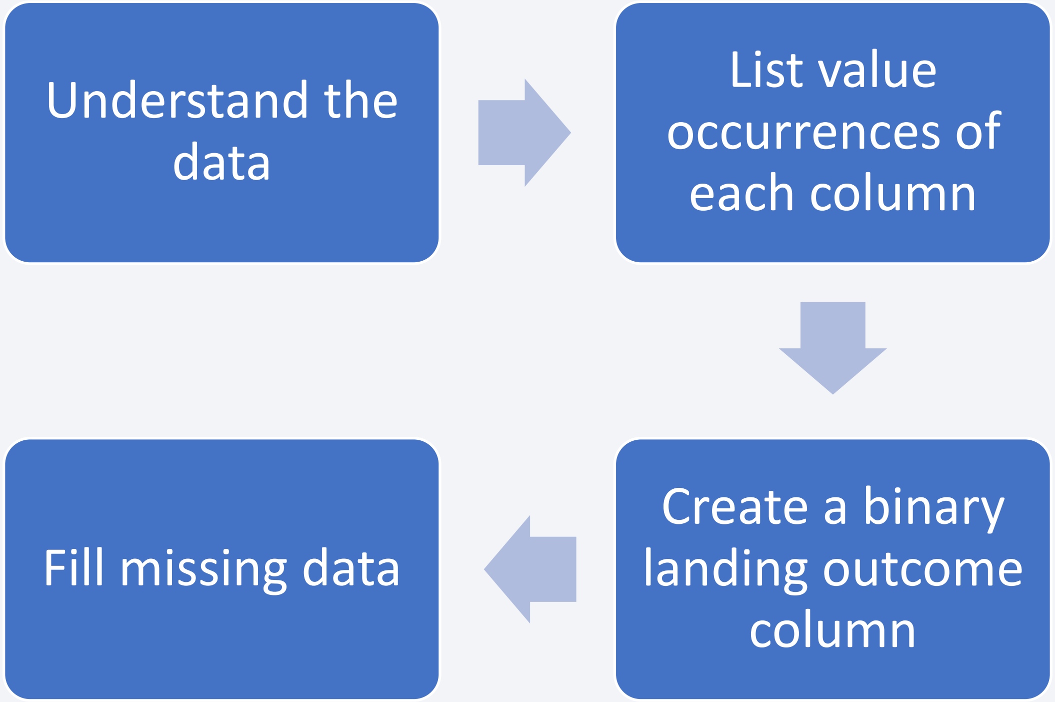
- Wikipedia contains historical Launch data
- We will collect data from it using Beautiful Soup for web scarping and put it in a data frame

  - GitHub URL of the web scraping notebook : [Applied-DS-Capstone/W1-2 jupyter-labs-webscraping.ipynb at main · fouadatmeh/Applied-DS-Capstone · GitHub]


GET HTML Page
Convert it to Beautiful Soup Object
Extract Column Names
Extract the data from the HTML tables
Put the data in a Dictionary
Convert the Dictionary into a Data Frame

Data Wrangling

- Some exploratory data analysis was done to understand the data
- Since the data is clean, little wrangling was needed as follows:
    - Simplify the Outcome column to either a 1 or 0
        - 1 for successful landing and 0 for unsuccessful landing
    - Fill missing PayloadMass values with the mean


Understand the data

List value occurrences of each column

Create a binary landing outcome column

Fill missing data

- GitHub URL of data wrangling notebook: [Applied-DS-Capstone/W1-3-labs-jupyter-spacex-Data wrangling.ipynb at main · fouadatmeh/Applied-DS-Capstone · GitHub]()

EDA with Data Visualization

- The following charts were plotted, Further Details are in section 2
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Success Rates of Different Orbits
  - Flight Number vs Orbit
  - Payload Mass vs Orbit
  - Success Rates of different Dates

- GitHub URL of EDA visualization notebook:Applied-DS-Capstone/W2-2 jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb at main · fouadatmeh/Applied-DS-Capstone · GitHub

EDA with SQL

- The following SQL Queries were done on the database to analyze the data, further details are in Section 2:
  - 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.
  - 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.
  - 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.
- GitHub URL of EDA with SQL notebook: Applied-DS-Capstone/W2-1 jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · fouadatmeh/Applied-DS-Capstone · GitHub

Build an Interactive Map with Folium

- I have added the following items to the folium map, for each site location:
  - Circle around the launch site, with a red color
  - Popup label for each site, showing the site name
  - Markers for each launch, with a green color for successful launches, and red color for unsuccessful ones
  - Lines showing distances from nearest facilities, namely: Coastline, Railroads and Highways.

- GitHub URL of interactive map notebook: Applied-DS-Capstone/W3-1 lab_jupyter_launch_site_location.jupyterlite.ipynb at main · fouadatmeh/Applied-DS-Capstone (github.com)

Build a Dashboard with Plotly Dash

We have added two main elements of the dashboard:

- Launch records dashboard: showing the success rate for all sites in the form of a pie chart
  - A selection box allows to drill down per site and see the success rate for each site.
- Payload vs Launch outcome: showing the success rate per Payload Mass in the form of a scatter plot
  - A range slider allows to narrow the payload mass selection by selecting a minimum value and a maximum value.
- GitHub URL of the Plotly Dash code: Applied-DS-Capstone/spacex_dash_app.py at main · fouadatmeh/Applied-DS-Capstone (github.com)

Predictive Analysis (Classification)

- Multiple classification models were tested, namely:
  - Logistics Regression
  - Support Vector Machine
  - Decision Tree
  - k Nearest Neighbors
- Each model was tested using the following methodology:
  - Data was split into test and training data
  - Model was trained using the training data
  - Multiple hyperparameters were tested, and the one with the best result was selected

- The decision Tree model provided the most accurate model.

- GitHub URL of predictive analysis Notebook. Applied-DS-Capstone/W4_1 SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

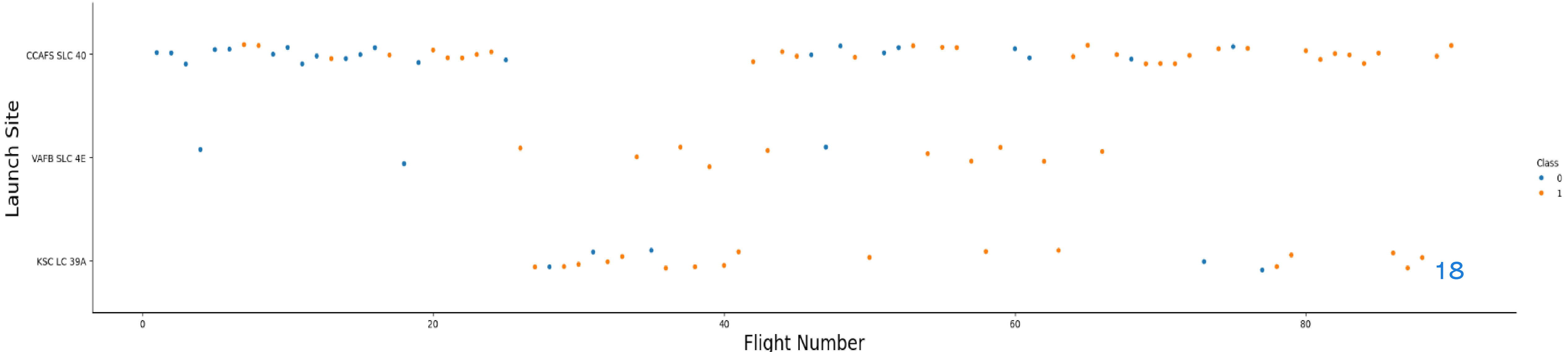
- Please check the following slides in Section two for the details of the analysis that was done.

Section 2

Insights drawn from EDA

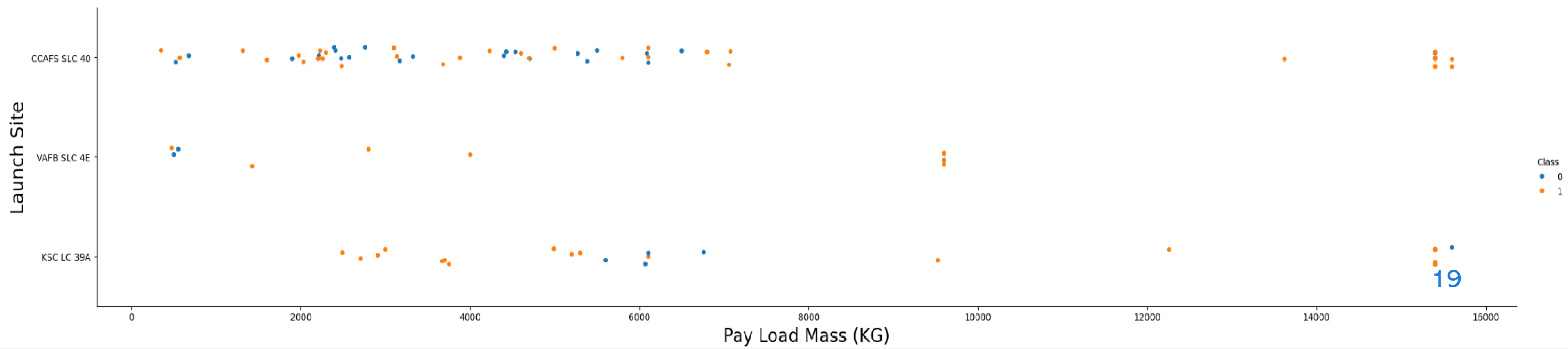
Flight Number vs. Launch Site

- Scatter plot comparing Flight Number (represents the continuous launch attempts) vs the Launch Site with success as a color overlay (blue fail and orange success)
- Outcomes:
  - As the flight number increases, the first stage is more likely to land successfully.
  - Different sites have different success rates



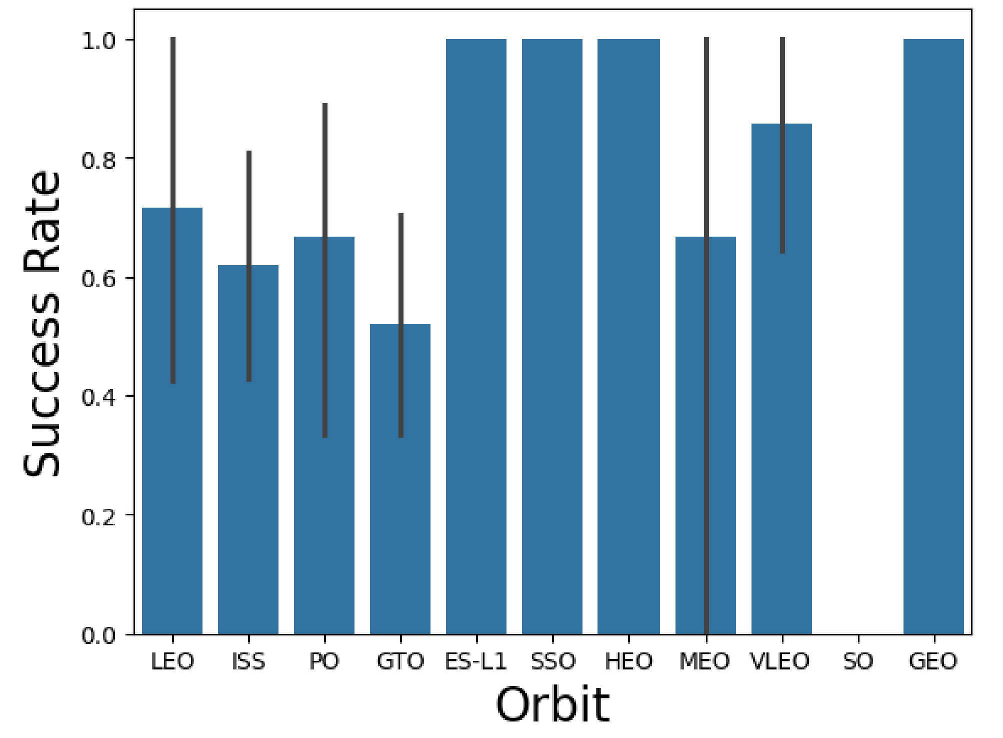
Payload vs. Launch Site

- Scatter plot comparing Payload Mass vs the Launch Site with success as a color overlay (blue fail and orange success)
- Outcomes:
  - Some sites don’t launch heavy payloads (greater than 10,000)
  - Different sites have different success rates



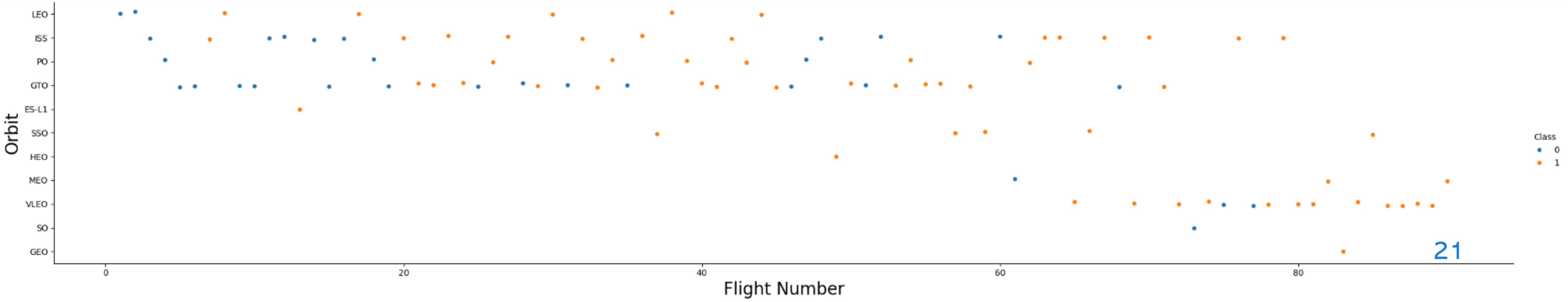
Success Rate vs. Orbit Type

- Bar showing Success Rate of Different Orbits
- Outcomes:
  - Different orbits have different success rates
  - Some orbits have a success rate of nearly 100%



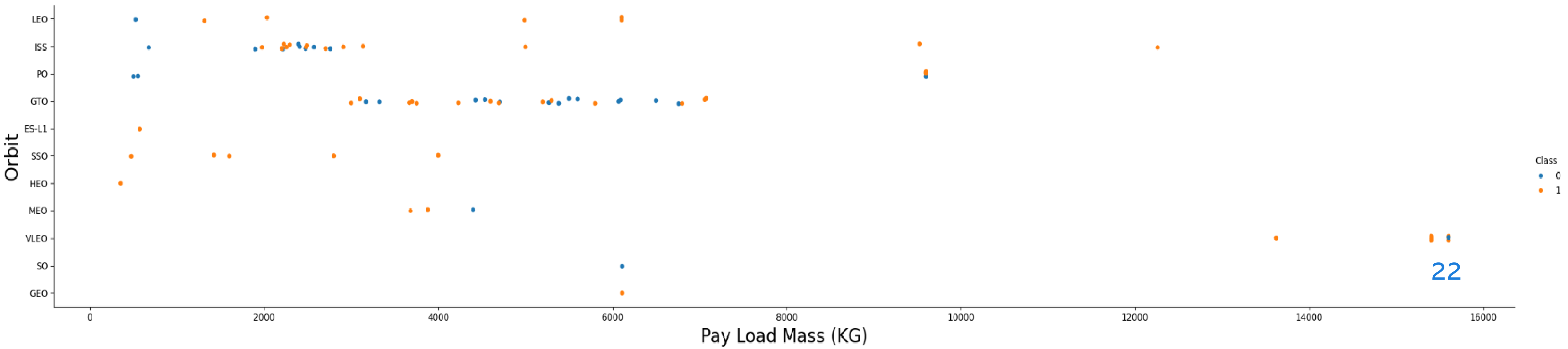
Flight Number vs. Orbit Type

- Scatter plot comparing Flight Number (represents the continuous launch attempts) vs the Orbit with success as a color overlay (blue fail and orange success)
- Outcomes:
  - Some orbits like Leo has a success rate related to flights Like LEO
  - Other orbits don't seem to have such a relation like GTO orbit.



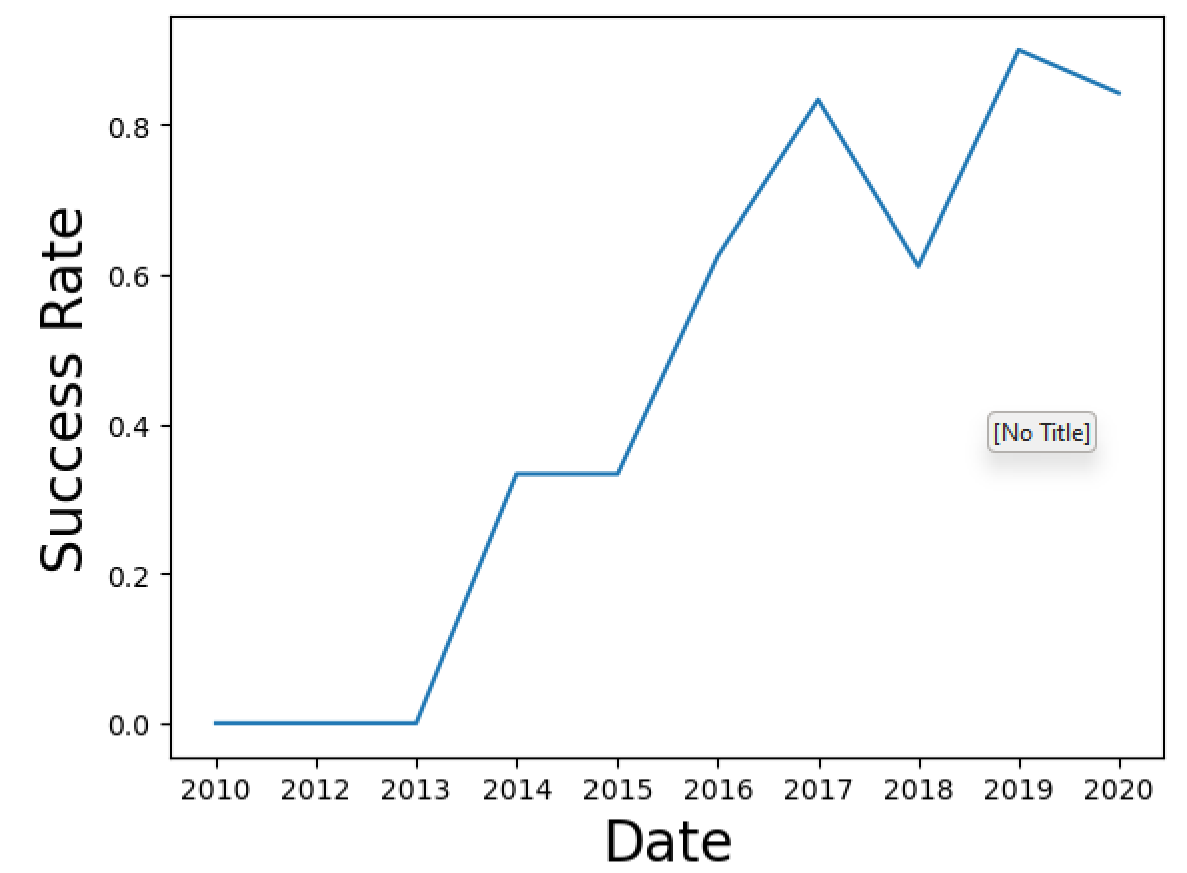
Payload vs. Orbit Type

- Scatter plot comparing Payload Mass vs the Orbit with success as a color overlay (blue fail and orange success)
- Outcomes:
  - For some orbits, heavy payloads have higher successful landing rate such as Polar, LEO and ISS.
  - For other orbits (such as GTO) don't seem to have such a relation



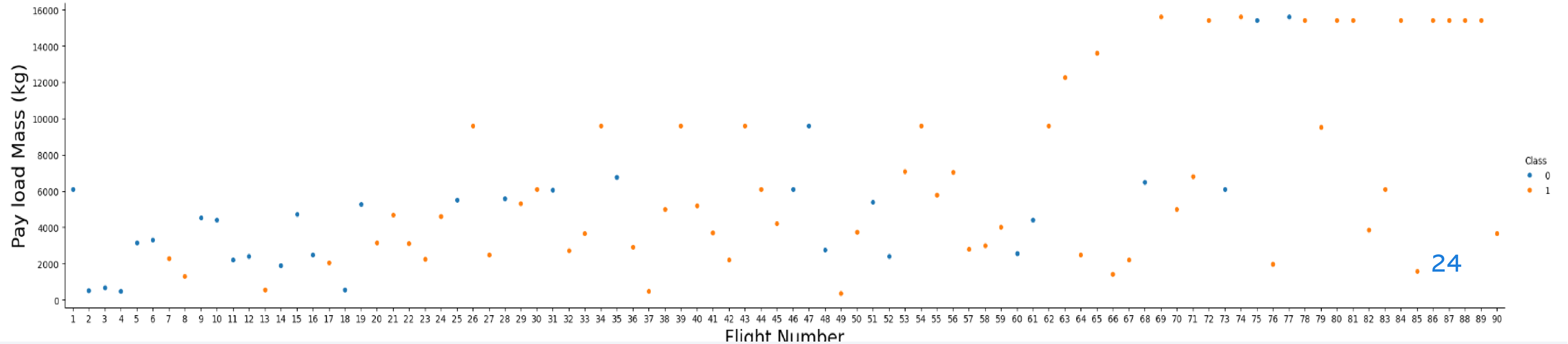
Launch Success Yearly Trend

- Line Chart Showing Success Rate of Different Dates
- Outcomes:
  - Success rate since 2013 kept increasing till 2020



Flight Number vs Payload Mass

- Scatter plot comparing Flight Number (represents the continuous launch attempts) vs the Payload with success as a color overlay (blue fail and orange success)
- Outcomes:
  - As the flight number increases, the first stage is more likely to land successfully.
  - Payload mass is also important; where the higher the payload, the less likely the first stage will return.



All Launch Site Names

- The names of the unique launch sites are

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

- The query used is :

```
%sql SELECT DISTINCT Launch_Site from SPACEXTABLE
```

Launch Site Names Begin with 'CCA'

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- The SQL Query used is:

```
%sql SELECT Launch_Site from SPACEXTABLE where Launch_Site LIKE 'CCA%' LIMIT 5
```

Total Payload Mass

- The total payload carried by boosters from NASA is:

45596 KG

- The SQL Query used is:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE CUSTOMER="NASA (CRS)"
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is:

2928.4 KG

- The SQL Query used is:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version = "F9 v1.1"
```

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is:

22-12-2015

- The SQL Query used is:

```
%sql SELECT Min(Date) from SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

- The SQL Query used is:

```
%sql SELECT Booster_Version from SPACEXTABLE WHERE  Landing_Outcome ='Success (drone ship)'
and PAYLOAD_MASS__KG_ Between 4000 and 6000
```

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes is:

100

- The total number of failure mission outcomes is:

1

- The SQL Query used are (respectively):

```
%sql SELECT Count(Mission_Outcome) from SPACEXTABLE WHERE Mission_Outcome!='Failure (in flight)’
```

```
%sql SELECT Count(Mission_Outcome) from SPACEXTABLE WHERE Mission_Outcome='Failure (in flight)'
```

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass is:

Booster_Version
F9 v1.1 B1018

- The SQL Query used is:

```
%sql SELECT Booster_Version from SPACEXTABLE WHERE Booster_Version=(SELECT
MAX(Booster_Version) from SPACEXTABLE )
```

2015 Launch Records

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

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

- The SQL Query used is:

```
%sql SELECT substr(Date, 6,2),Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE Landing_Outcome ='Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

- Rank of 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:

Date	Landing_Outcome	count(Landing_Outcome)
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

- The SQL Query used is:

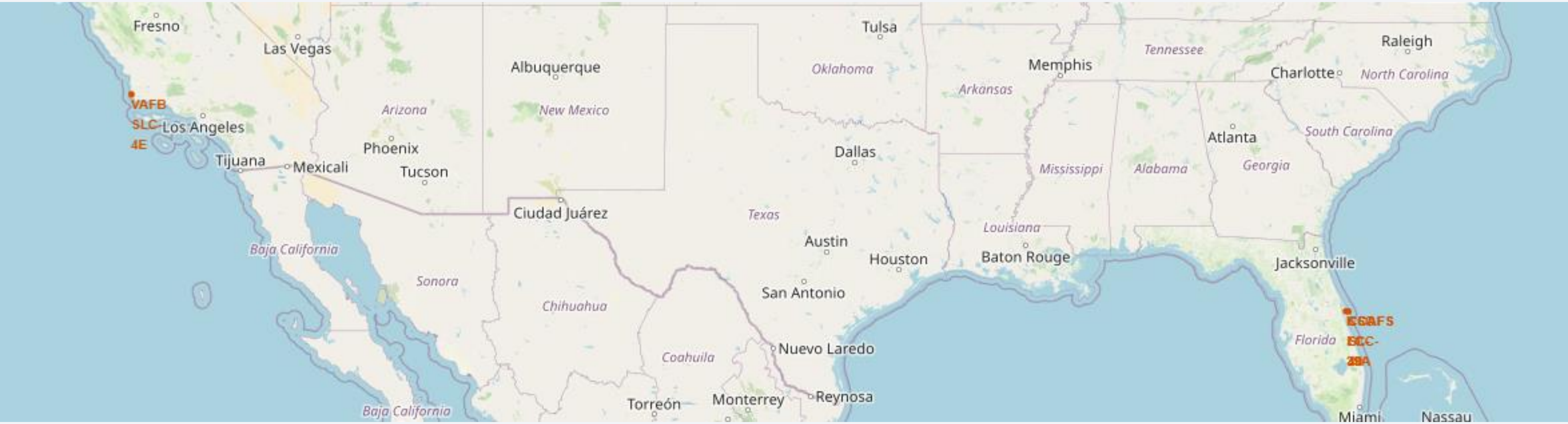
```
%sql SELECT Date, Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE \
  WHERE Date BETWEEN '2010-06-04' and '2017-03-20' \
  GROUP BY Landing_Outcome \
  ORDER by count(Landing_Outcome) DESC
```

Section 3

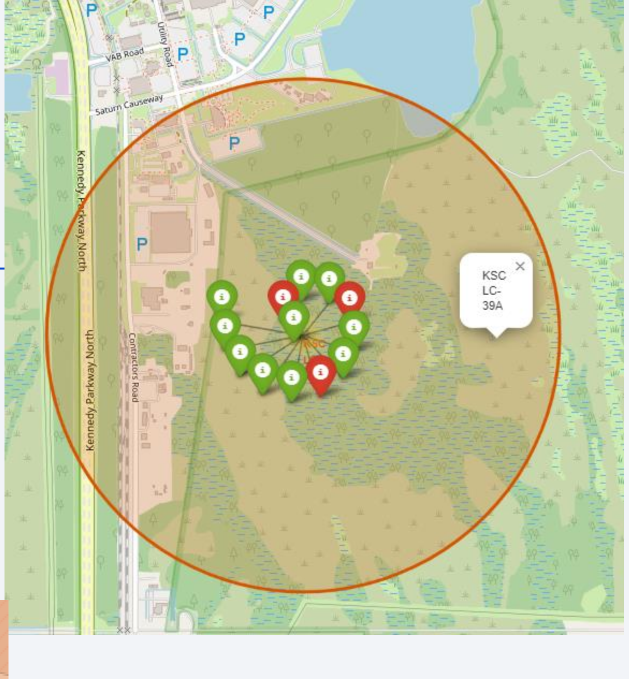
Launch Sites Proximities Analysis

Launch Site Locations

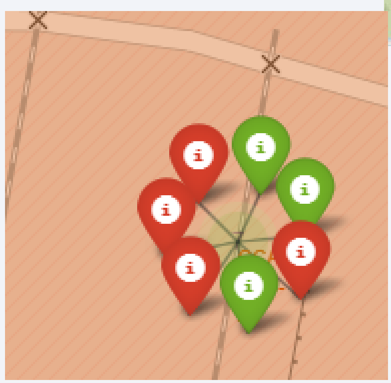
- Following map shows all the launch site locations on the map
- We see that:
  - all the launch sites are in the US
  - They are near the coastline (east and west)

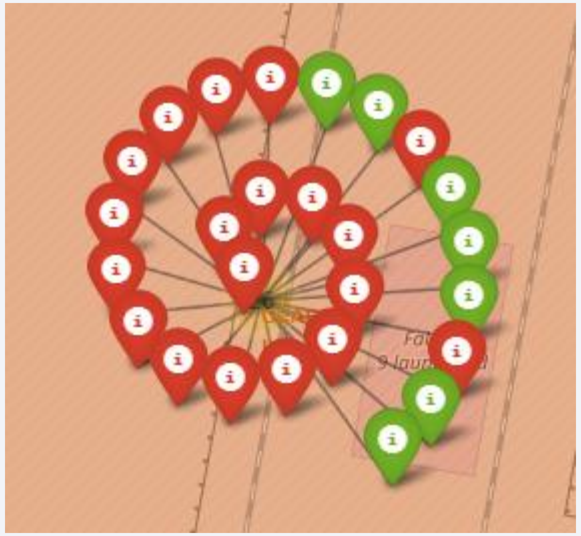


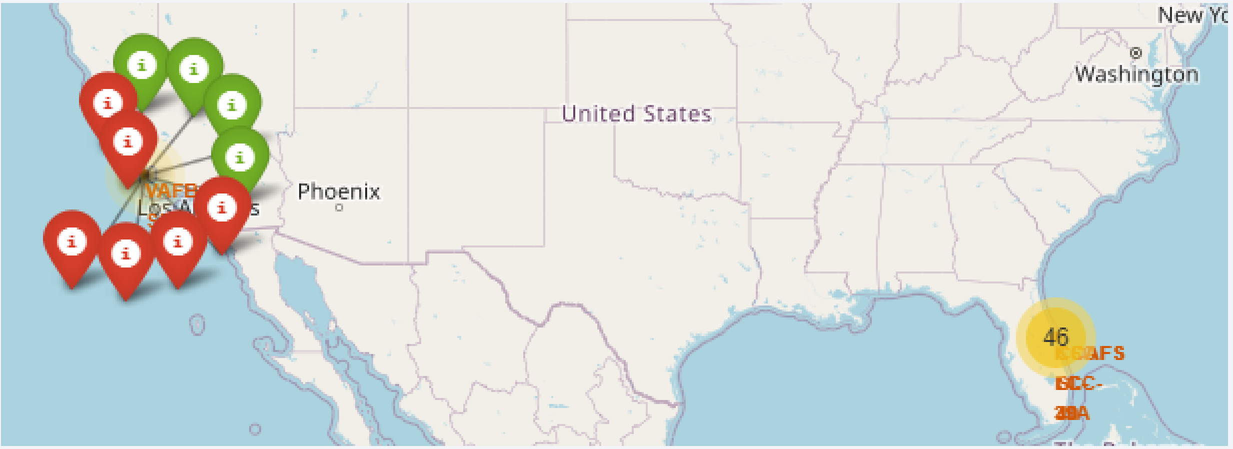
Launch Outcomes per Site

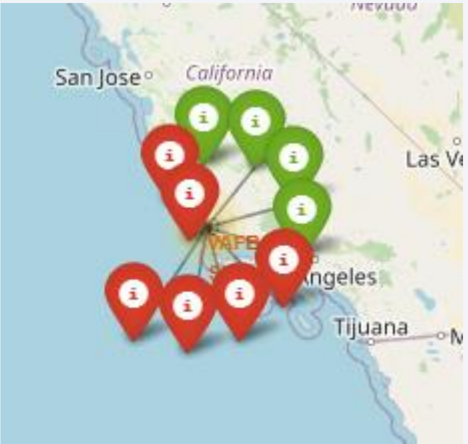


- We can see the different launch outcomes of each site by clicking on it as see below
- We can zoom in to the details of 4 launch sites



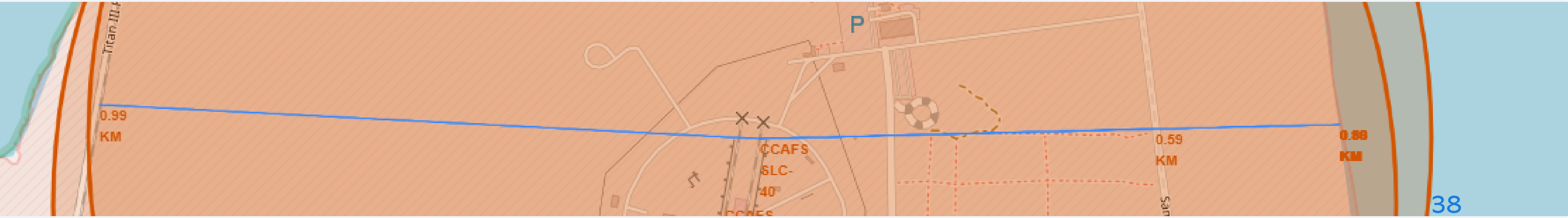


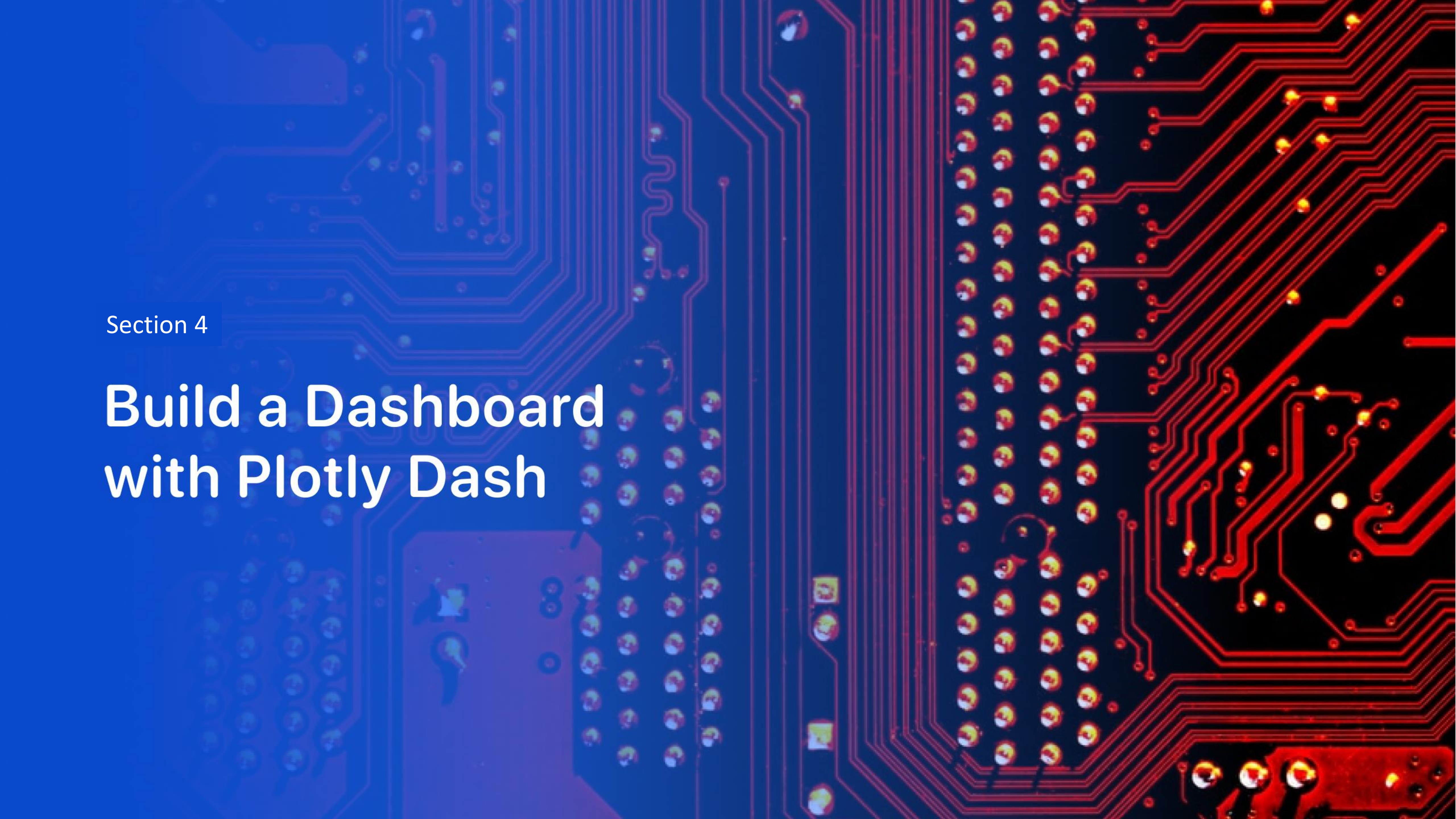




Distance to Coastline, Railway, and Highway

- Each launch site requires some facilities to be nearby such as Coastline, Railways and Highways.
- We explored the CCAFS SLC-40 site and found the following distances:
  - Coastline: 0.88 KM
  - Highway: 0.59 KM
  - Railway: 0.99 KM



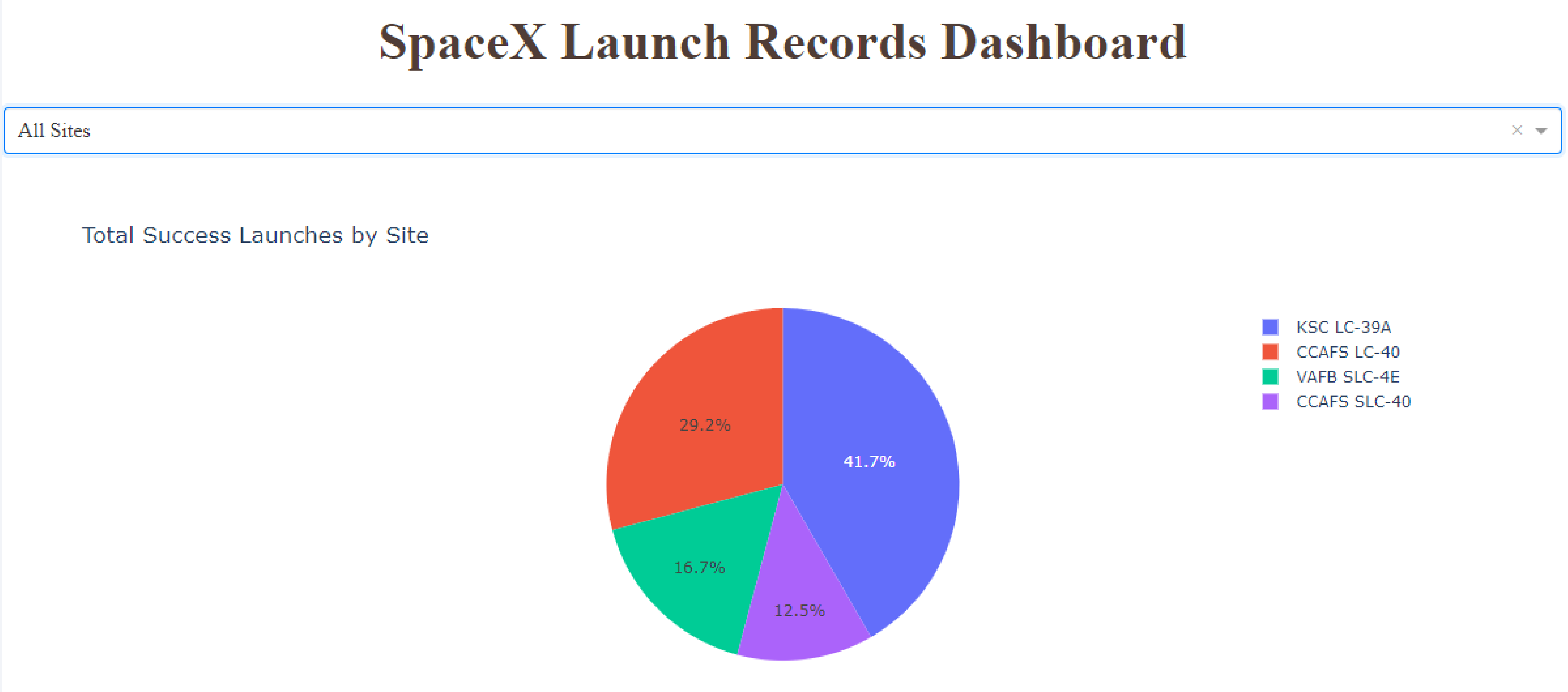


Section 4

Build a Dashboard with Plotly Dash

Launch Success Count Chart

- Dashboard screen showing success count for each launch site in a piechart

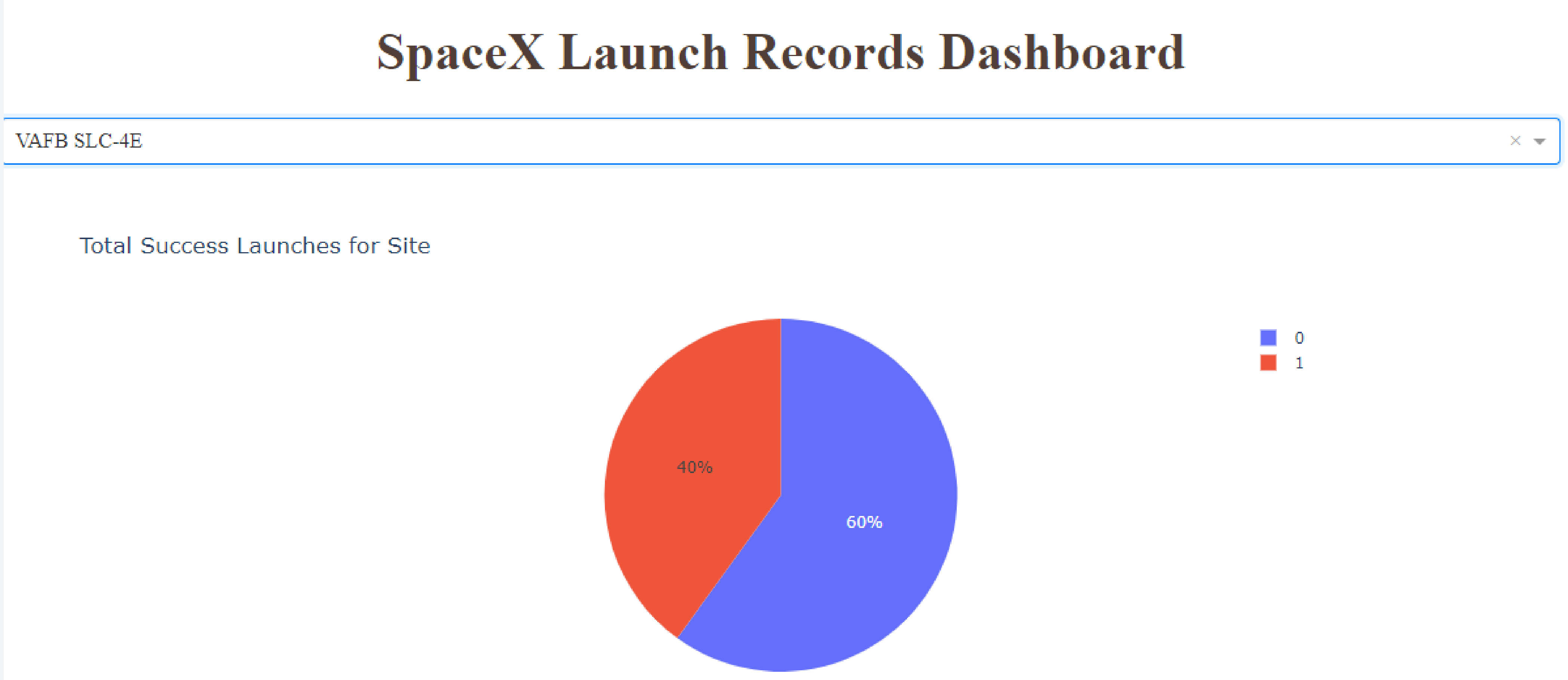


SpaceX Launch Records Dashboard

All Sites

Site with Highest Launch Success Rate

- Dashboard screen showing site with hightes launch success rate in a piechart



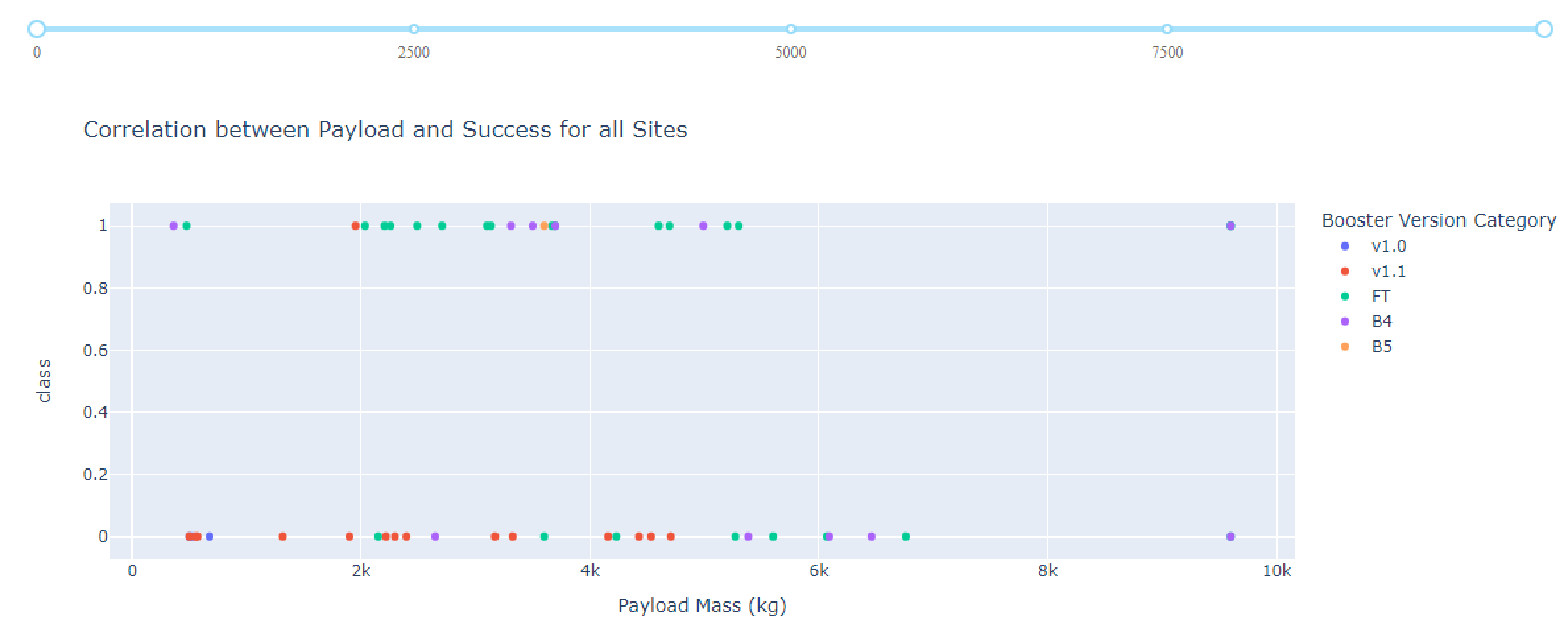
SpaceX Launch Records Dashboard

VAFB SLC-4E

Total Success Launches for Site

Payload vs Launch Outcome Scatter Plot

- Dashboard screen showing Payload vs. Launch Outcome scatter plot for all sites, with full payload range selected in the range slider
  - The FT booster version has the highest success rate on lower payload masses

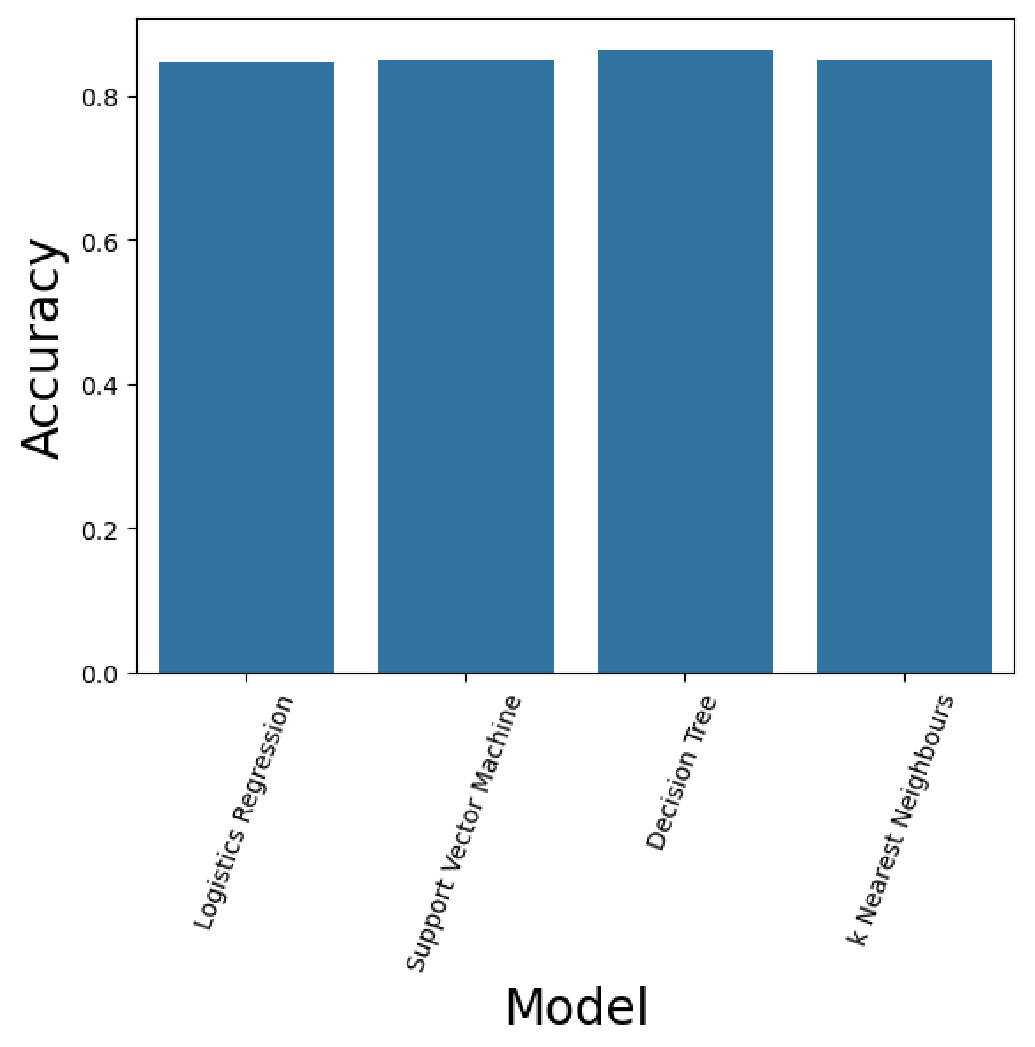


Section 5

Predictive Analysis (Classification)

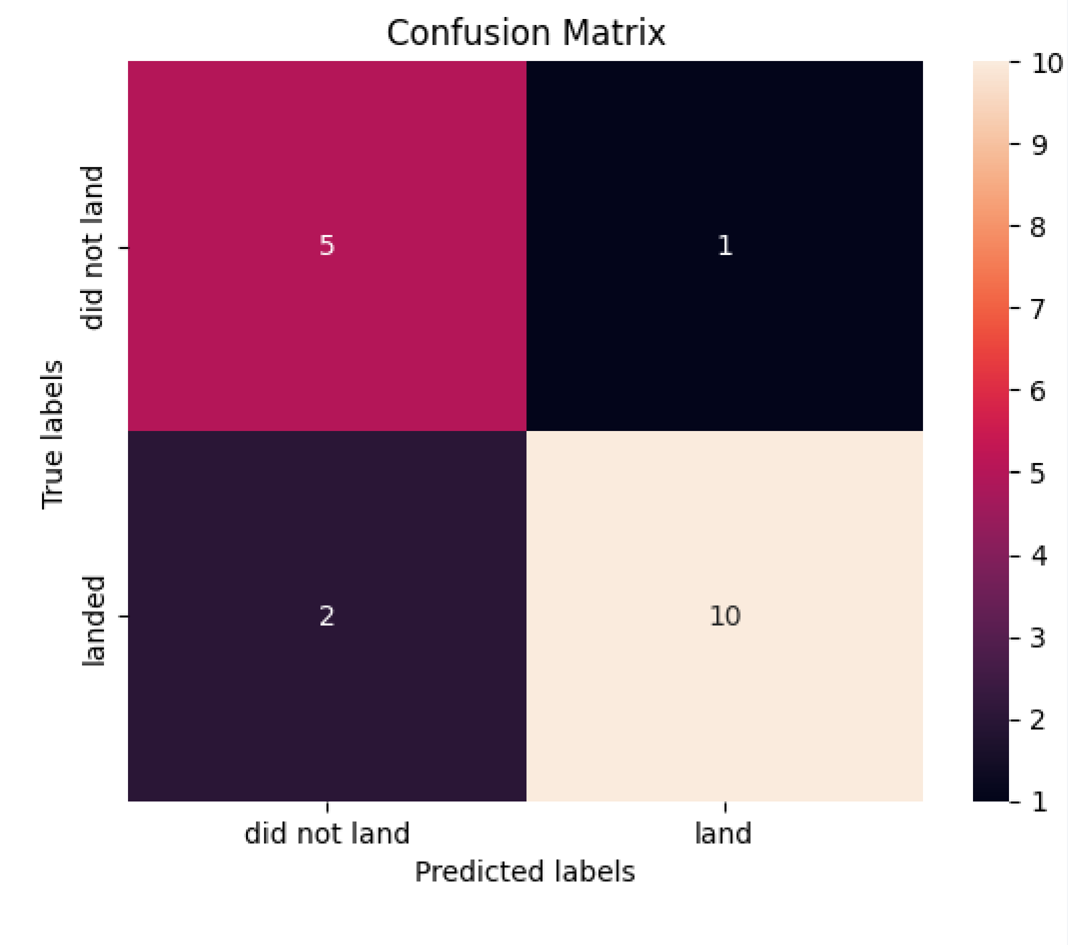
Classification Accuracy

- Model accuracy for the various models are plotted on the right
- Decision Tree method got the highest accuracy of :
  - 86429%



Confusion Matrix

- Following is the confusion matrix of the decision tree model



Conclusions

- Each classification model provides a different accuracy
- The following models were tested:
  - Logistics Regression
  - Support Vector Machine
  - Decision Tree
  - k Nearest Neighbors
- The decision Tree model provided the most accurate model.

Appendix

- All code, notebooks and other relevant information are included in the GitHub repository:
  - Applied-DS-Capstone (github.com)

Thank you!