IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nikki de Vries
November 6, 2024

Table of Contents

Github Repository: https://github.com/NikkideVries/IBM-SpaceX-Capstone

Executive Summary

- This project aimed to investigate and find drivers that influenced a successful first-stage landing of the SpaceX Falcon 9 rocket.

- To get to this goal, SpaceX mission data was obtained through a public API and launch data from web scrapping Wikipedia. The data was then cleaned, processed, and transformed for Exploratory Data Analysis (EDA) using Python and SQL. Visualizations such as graphs, maps, and interactive dashboards to find drivers.
- Machine learning classification models, logistic regression, support vector machines, decision trees, and K-nearest-neighbors were analyzed for their training and test accuracy for predicting launch success.

- Based on the results from the analysis:
  - The factors that influenced the successful first-stage landing are time, payload mass, and desired orbit
  - Using these parameters, each classification model had an accuracy of 83.3%, predicting the first-stage rocket booster landing successfully. The decision tree model had the best train accuracy.

Introduction

- SpaceX 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 SpaceX can reuse the first stage.
- As a data scientist, if I determine a successful first-stage landing for the Falcon 9 rocket, I can determine the cost of the launch. This information will help SpaceX prevent alternative companies from bidding against a SpaceX launch, allowing SpaceX to continue its Falcon 9 launches and continue innovating.
- To find the answers to predict a successful first-stage launch, I need to know:
  - Which Factors determine if the Falcon 9 Launch will land successfully?
  - Are there interactions between features that influence this success rate?
  - Are there additional insights I can gather from the data that may also benefit cost?

Section 1

Methodology

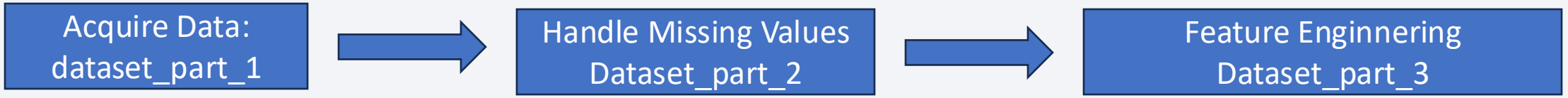
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through the SpaceX API and web scrapping Wikipedia
- Perform data wrangling
  - Missing values were replaced, and a landing outcome feature was created.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Normalized data and then performed a train test split for four classification models: logistic regression, support vector machines, decision trees, and k-nearest-neighbors.

Data Collection Overview

- API
  - Acquired API data through a get request to the SpaceX API.
  - Cleaned the requested data

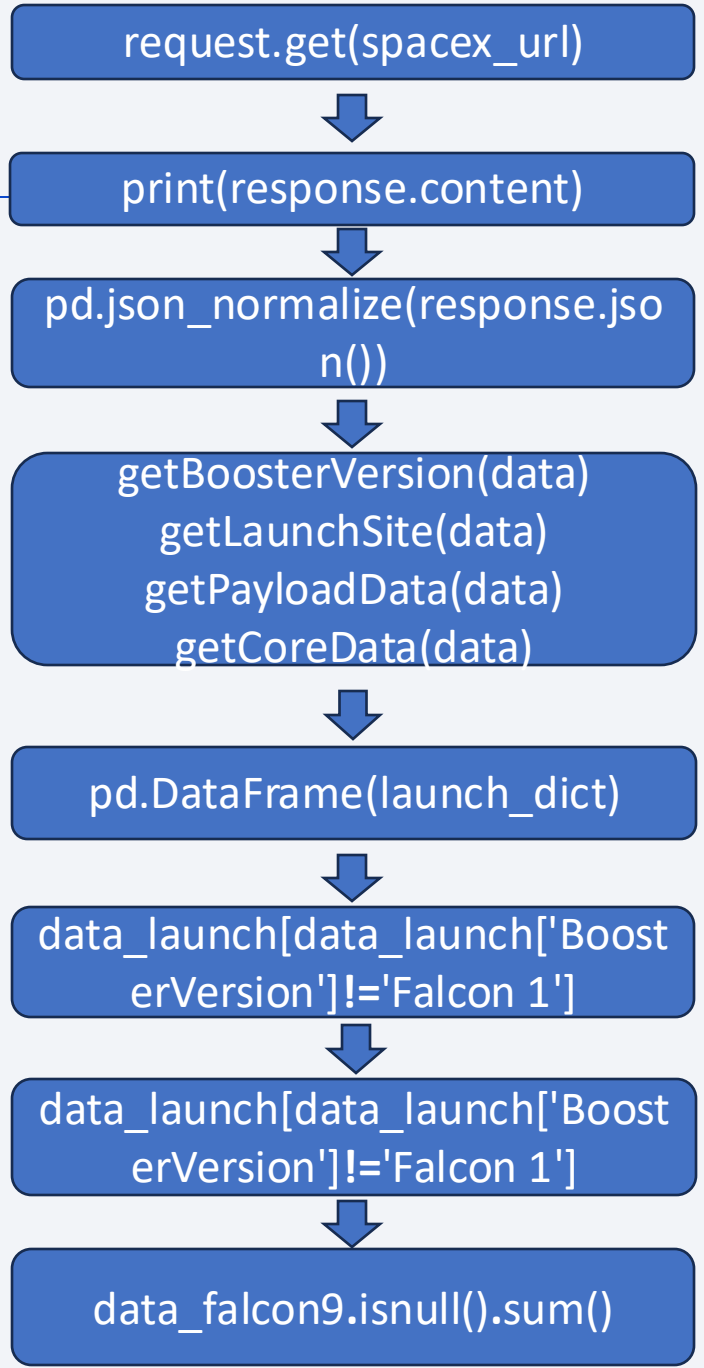


- Web scraping
  - Extracted Falcon 9 launch records HTML table from Wikipedia
  - Parsed the table and converted it into a pandas data frame

Data Collection – SpaceX API

1. Request rocket launch data.
   - Print data to check the content of the response
2. Decode the response content as a JSON
   - Normalizing the data to convert the json result into a data frame
3. Use API to request and replace column information
   - Take a subset of the data frame to keep the desired features
   - Create request functions for rocket data, launchpad data, payload data, core data
4. Create data frame
   - Combined data into a dictionary to be converted to a dataframe
5. Filter the data to only include Falcon 9 launches
6. Deal with missing Values
7. Save to CSV

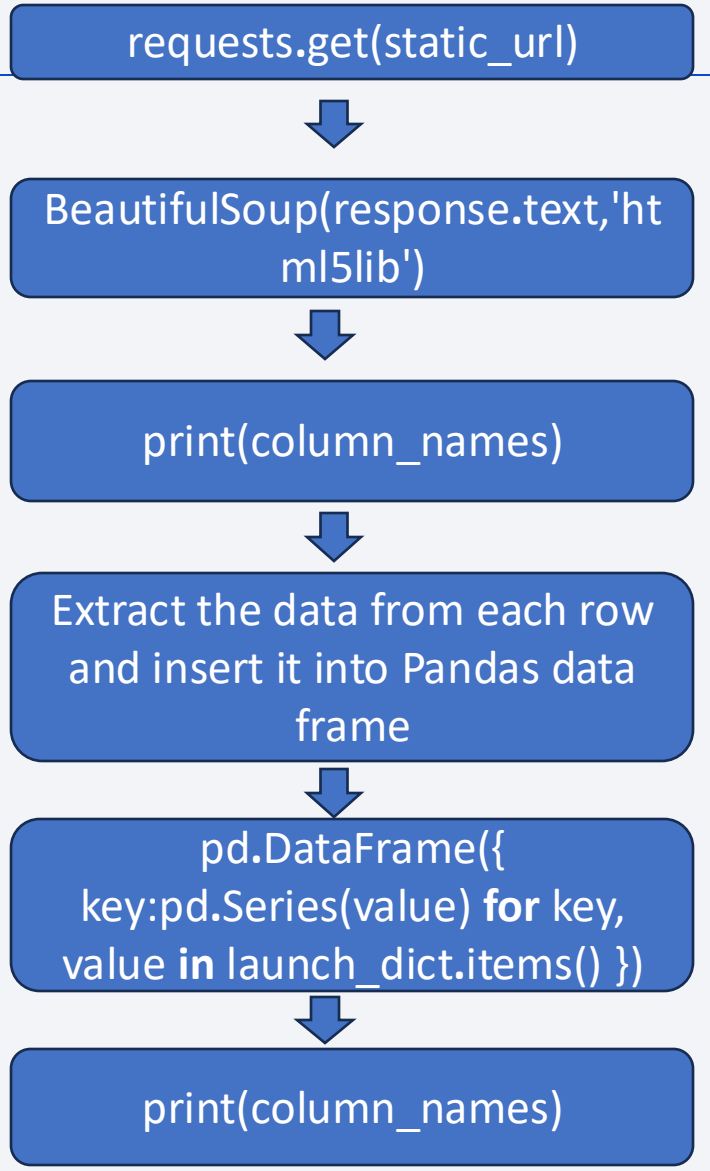
Github URL: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/1Data_Collection_API.ipynb



Data Collection – Web Scraping

1. Request the Falcon 9 launch data from the Wiki page URL
2. Use BeautifulSoup to parse the content
3. Extract the columns and variables from the HTML
4. Create data frame by parsing the launch HTML tables
5. Export to CSV

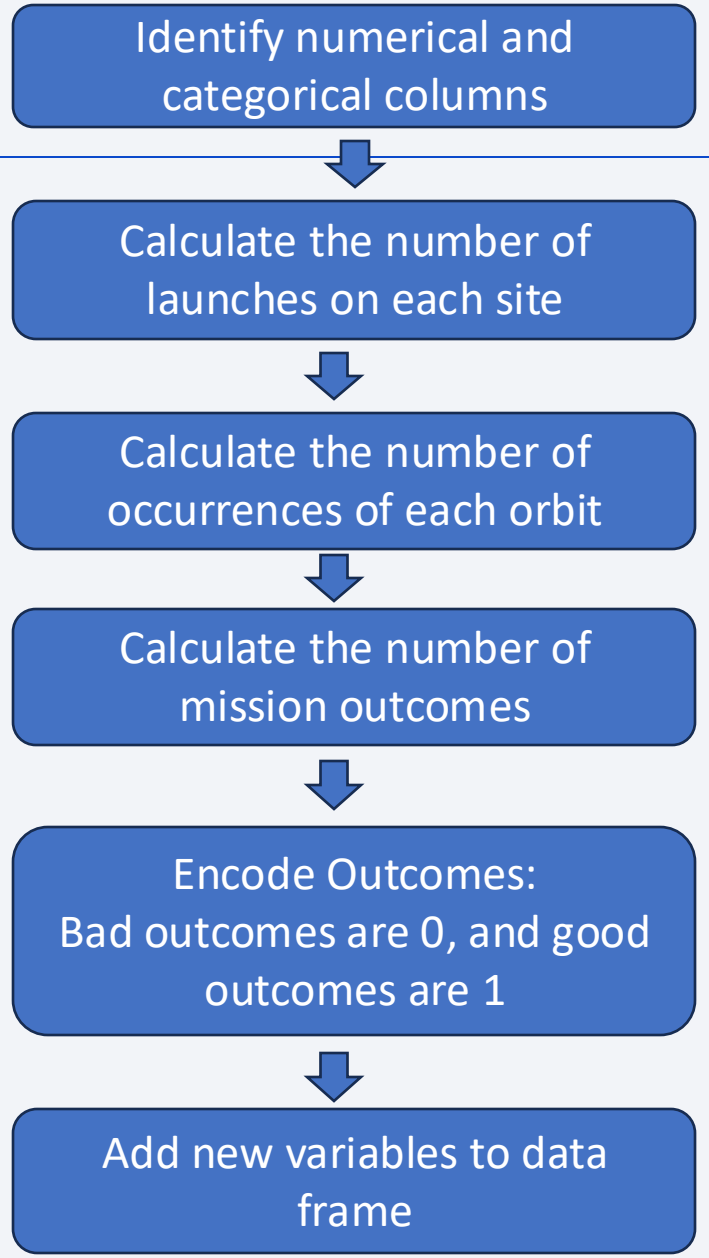
Git Hub: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/2Data_Collection_with_Web_Scrapping.ipynb



Data Wrangling

- Completed some initial exploratory data analysis to see what was happening in the data.
- Figured out from the data what mission outcome was either a failure to land or a successful landing.
- Based on that information, created a new classification variable, “class”. 0 indicates a bad outcome while 1 indicates a good outcome.
  - Bad outcomes are failures to land
  - Good outcomes are successful landings

Github: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/3Data_Wrangling.ipynb



EDA with Data Visualization

- Exploratory Data Analysis was completed by using Matplotlib visualizations. These visualizations were catplots, scatterplots, bar graphs, and line graphs. Specifically:
- Visualizations were color-coded by class to see successes and failures to land.
- The Features were engineered so categorical variables became numerical to allow for further visualization.

GITHUB: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/5EDA_with_Visualizations.ipynb

- Cat plots for:
  - Flight Number and Launch Site
  - Payload Mass and Launch Site
- Bar plot for:
  - Success rate and Orbit Type
- Scatter Plot for:
  - Flight Number and Orbit Type
  - Payload and Orbit Type
- Line Graph for:
  - Year and Success Rate

EDA with SQL

- Displaced the names of the unique launch sites in the space mission. **SELECT DISTINCT Launch_site FROM SPACEXTBL**;
- Displayed 5 records where launch sites begins with the string 'CCA'. **SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;**
- Displayed the total payload mass carried by boosters launched by NASA. **SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'**
- Display average payload mass carried by booster version F9 v1.1. **SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'**
- List the date when the first succesful landing outcome in ground pad was acheived. **SELECT MIN(Date) AS EarliestDate FROM SPACEXTBL WHERE Landing_Outcome LIKE '%(ground pad)' AND Mission_Outcome = 'Success';**
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000. **SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE '%(drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;**
- List the total number of successful and failure mission outcomes. **SELECT CASE WHEN Mission_Outcome LIKE '%success%' THEN 'Successful' ELSE 'Failures' END AS OutcomeCategory, COUNT(*) AS TotalCount FROM SPACEXTBL GROUP BY OutcomeCategory;**
- List the names of the booster_versions which have carried the maximum payload mass. **SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);**
- 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. **SELECT substr(Date, 6,2), Mission_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,0,5)='2015'**
- Rank the count of landing outcomes in descending order. **SELECT Landing_Outcome, COUNT(*) AS OutcomeCount FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY OutcomeCount DESC;**

Github: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/4EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- An interactive map was created in folium to allow for different Exploratory Analyses and to answer questions visually.

Github: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/6Interactive_Visuals_with_Folium.ipynb

- Markers:
  - Indicate launch site points to assist with identification.
- Circles:
  - Indicated highlighted regions around launch sites to help locate areas visually.
- Marker Clusters:
  - Specific launch sites in circles to help with identification
- Lines:
  - Show the distance between the coordinates

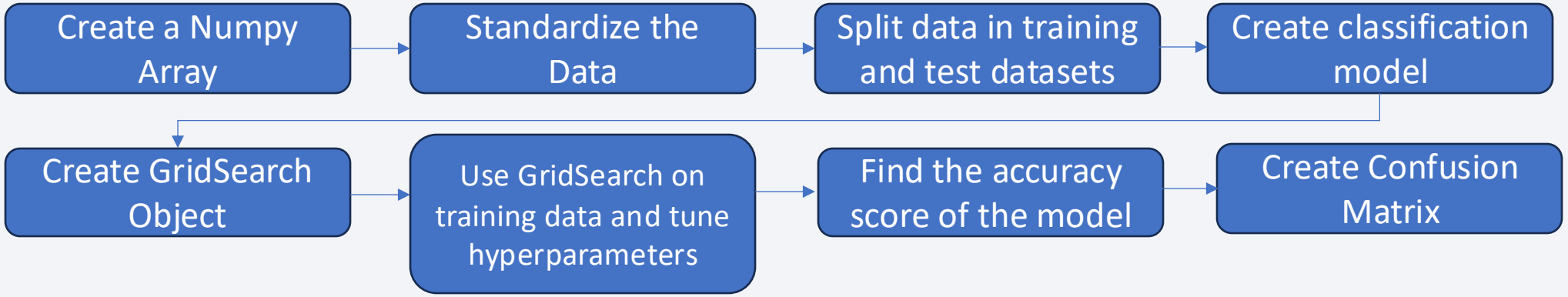
Build a Dashboard with Plotly Dash

- Using a Python interactive dashboard, Plotly Dash, an interactive dashboard was created to allow stakeholders to investigate the impacts of the launch site, payload mass, and booster type on the outcomes of the Falcon9 launch.
- Graphs that were created and that could be changed through the use of a dropdown menu were:
- A Pie chat:
    - To show which sites launched the most, and by site, what was the percentage of successful launches.
- Scatter plot:
    - To show the relationship between play load ranges and selected sites on successful launches.

Github: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/7Interactive_Dashboard_with_Ploty.py

Predictive Analysis (Classification)

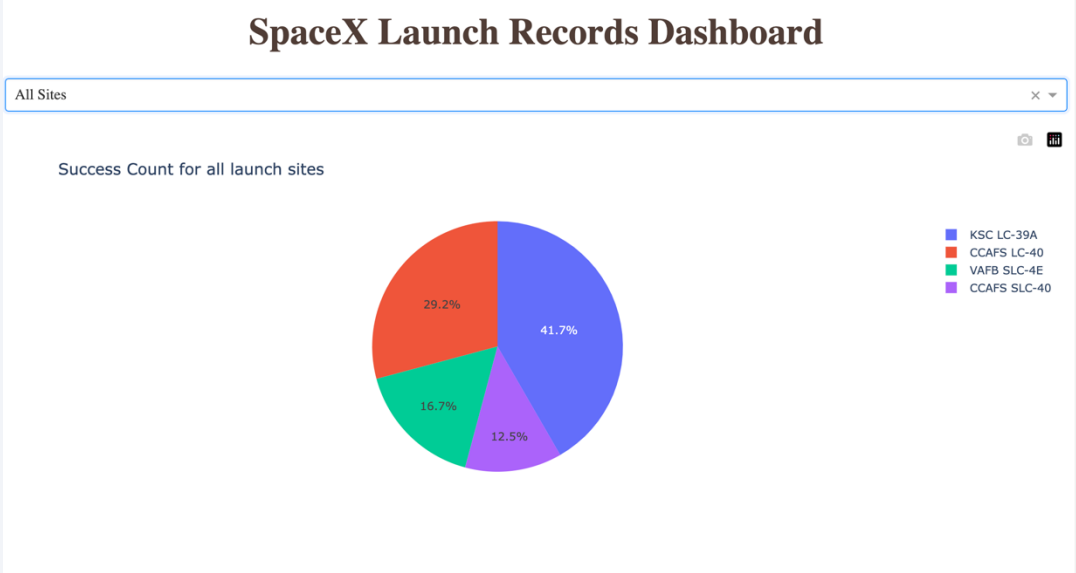
- Following the flow chart, four classification four models were used: Support Vector Machine, Decision Tree, Logistic Regression, and K-nearest neighbors.
- Accuracy was used to determine the best model. All models performed the same on the test data set with an accuracy of 83.3%. However, the model with the best training accuracy was the Decision Tree model with an accuracy of 87.5%.
- Github: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/8Machine_Learning_Prediction.ipynb

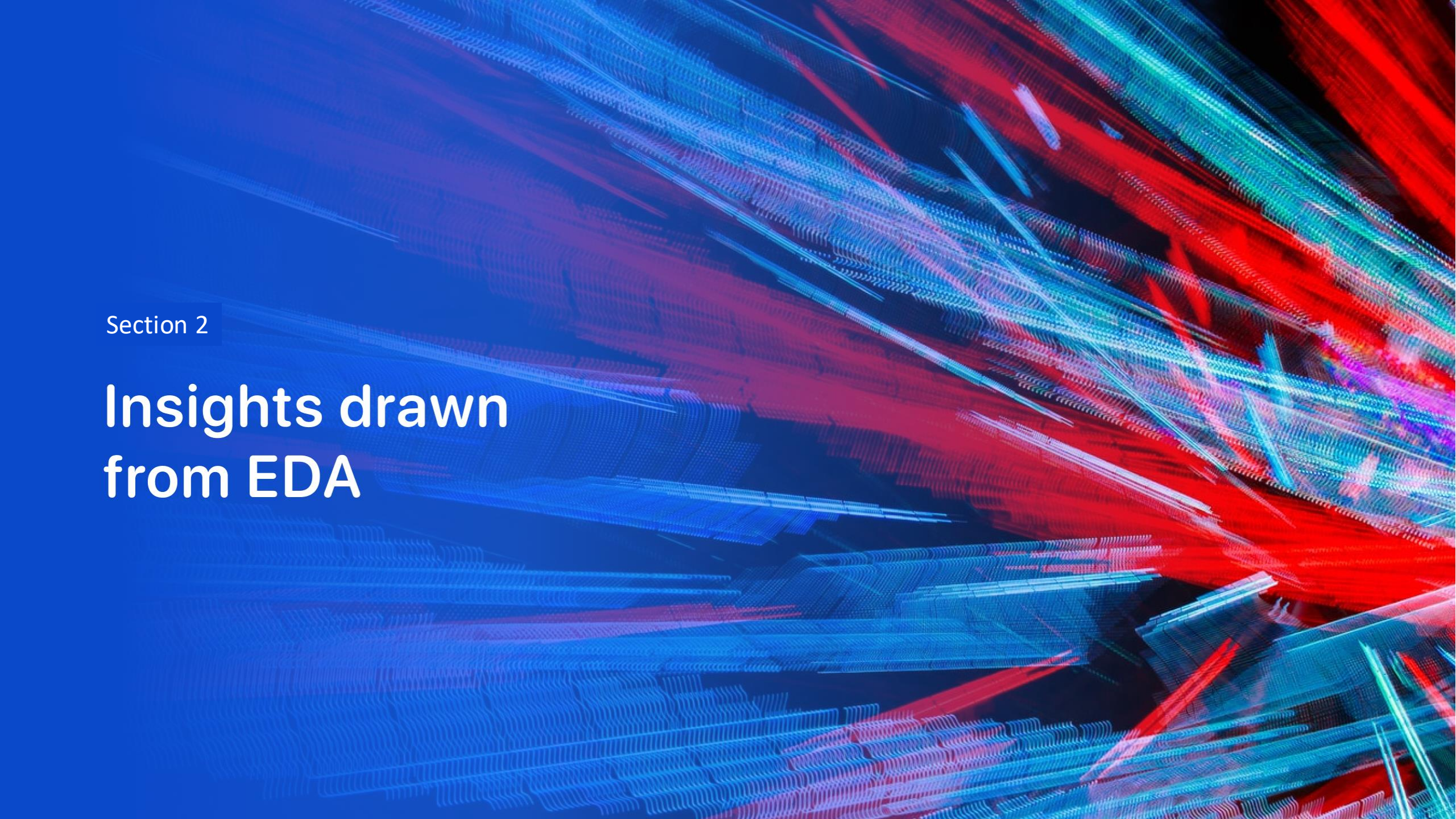


Results

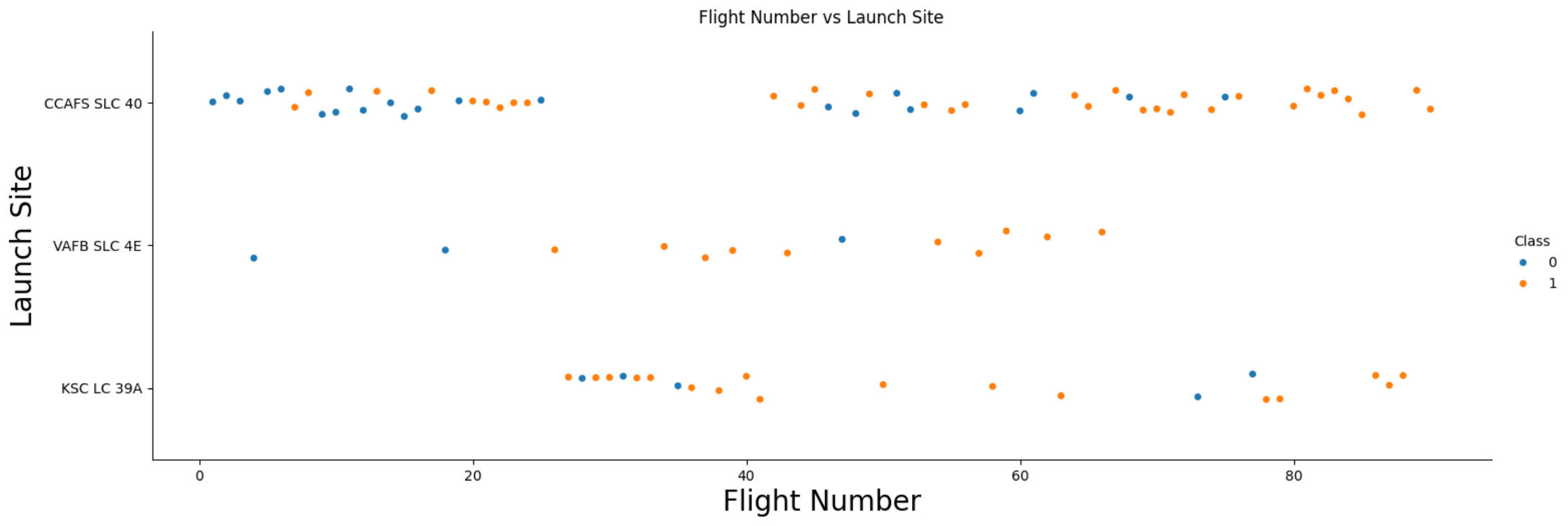
- Exploratory data analysis results
  - Launch success rate increases over time, with a large increase after 2013
  - The KSC site has the highest percentage of successful landings
  - Heavier payloads tend to lead to more landing failures

- Predictive analysis results
  - The decision tree model has the best training accuracy at 87.5%
  - All models perform the same on the test data with an accuracy of 83.3%.



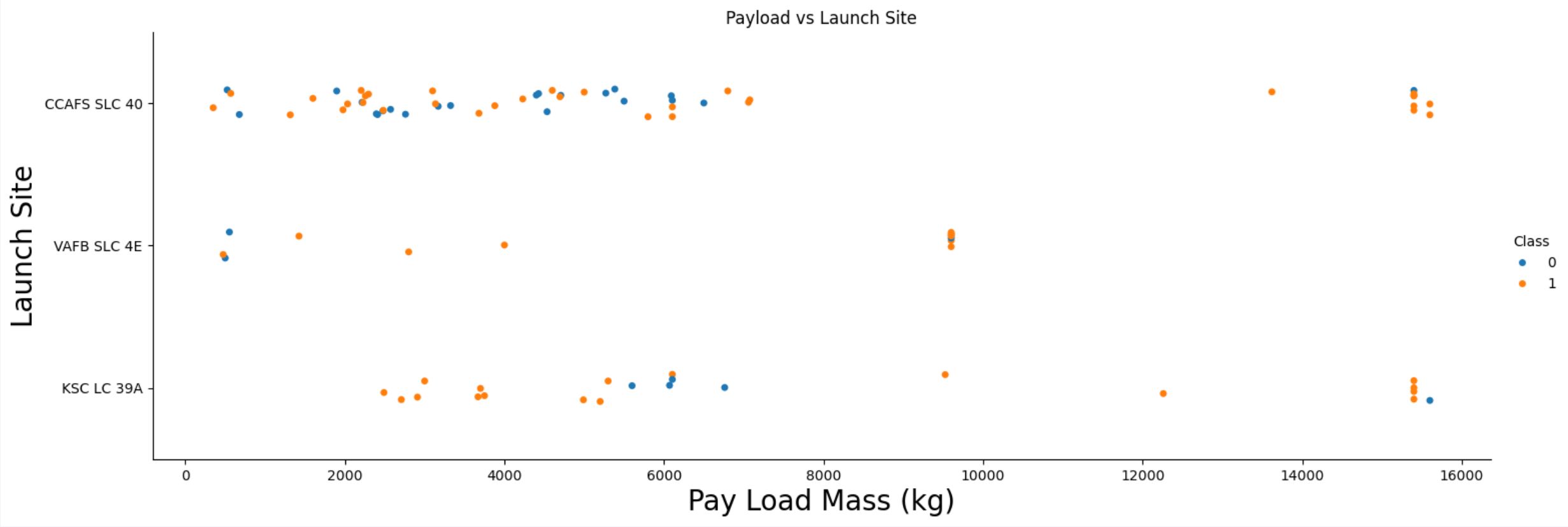

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Blue class indicates landing failure while Orange class indicates landing success.
- As the flight number increases across all launch site, there are more successful landings.

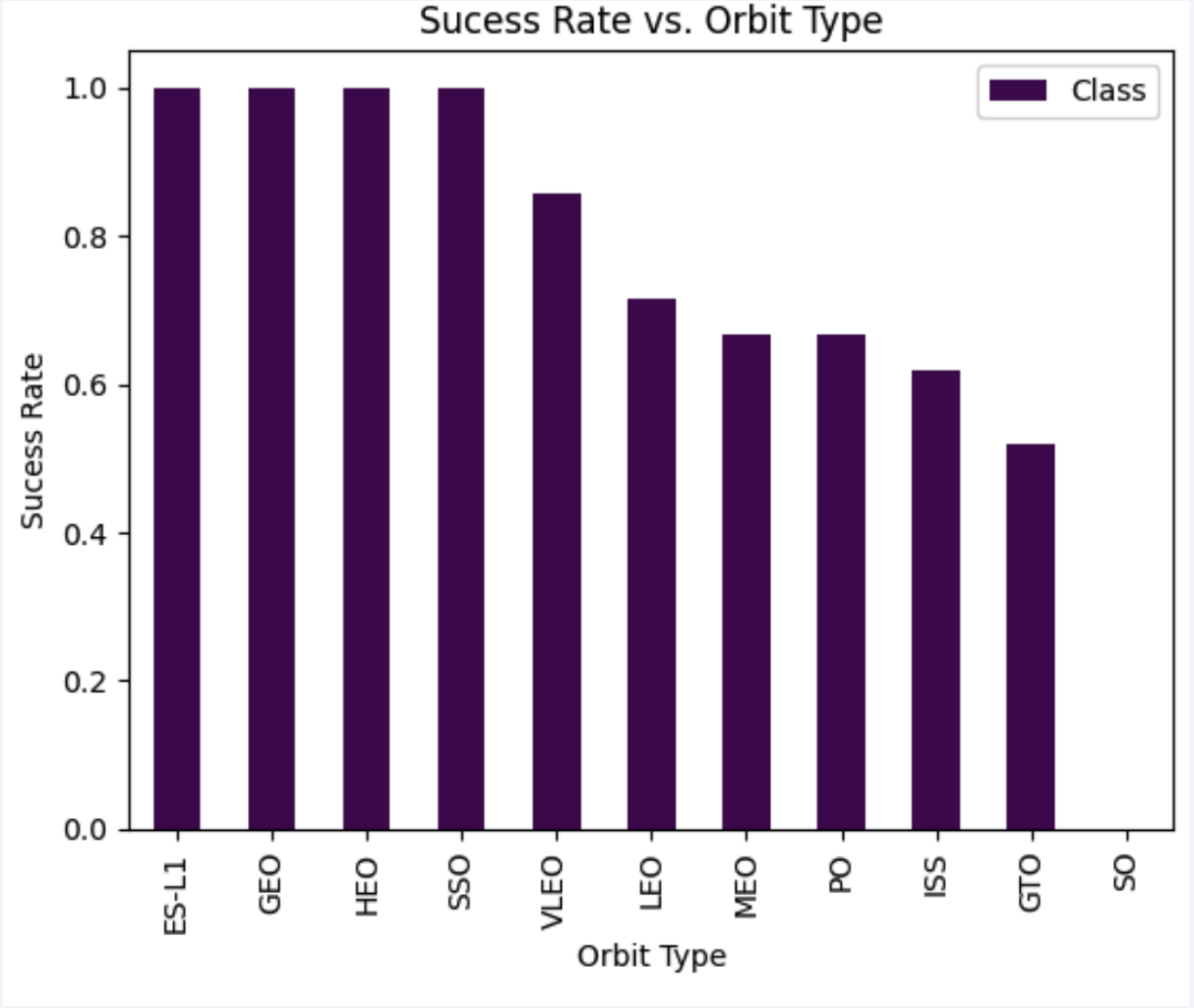
Payload vs. Launch Site



- The higher the payload mass for sites KSC and CCA have higher successful landings.

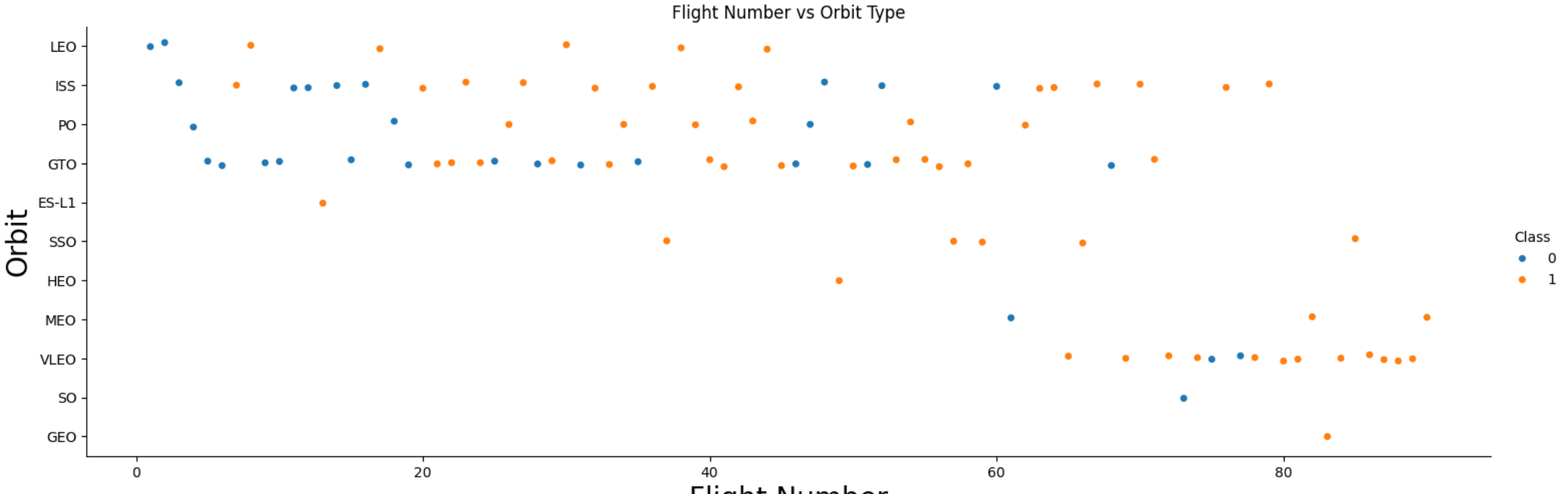
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbit types have the highest successful rates.
- GTO has the lowest success rate.



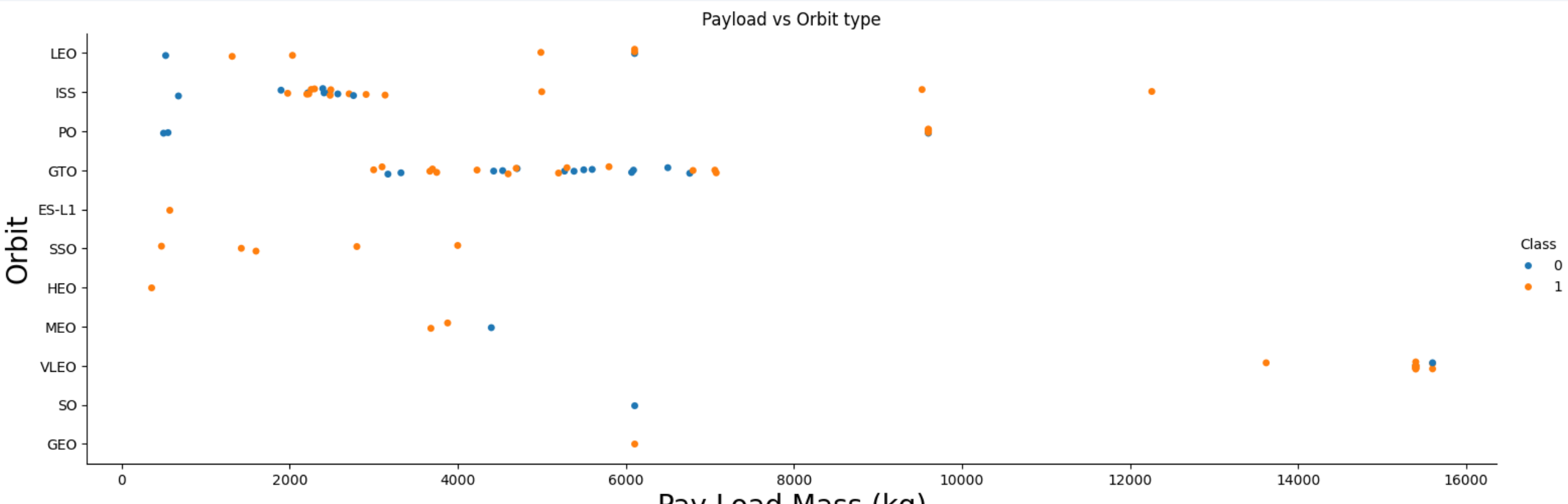
Flight Number vs. Orbit Type

- As the number of flights increases, the number of successful landings increases.
- Blue indicates a failed landing, while Orange indicates a successful landing.



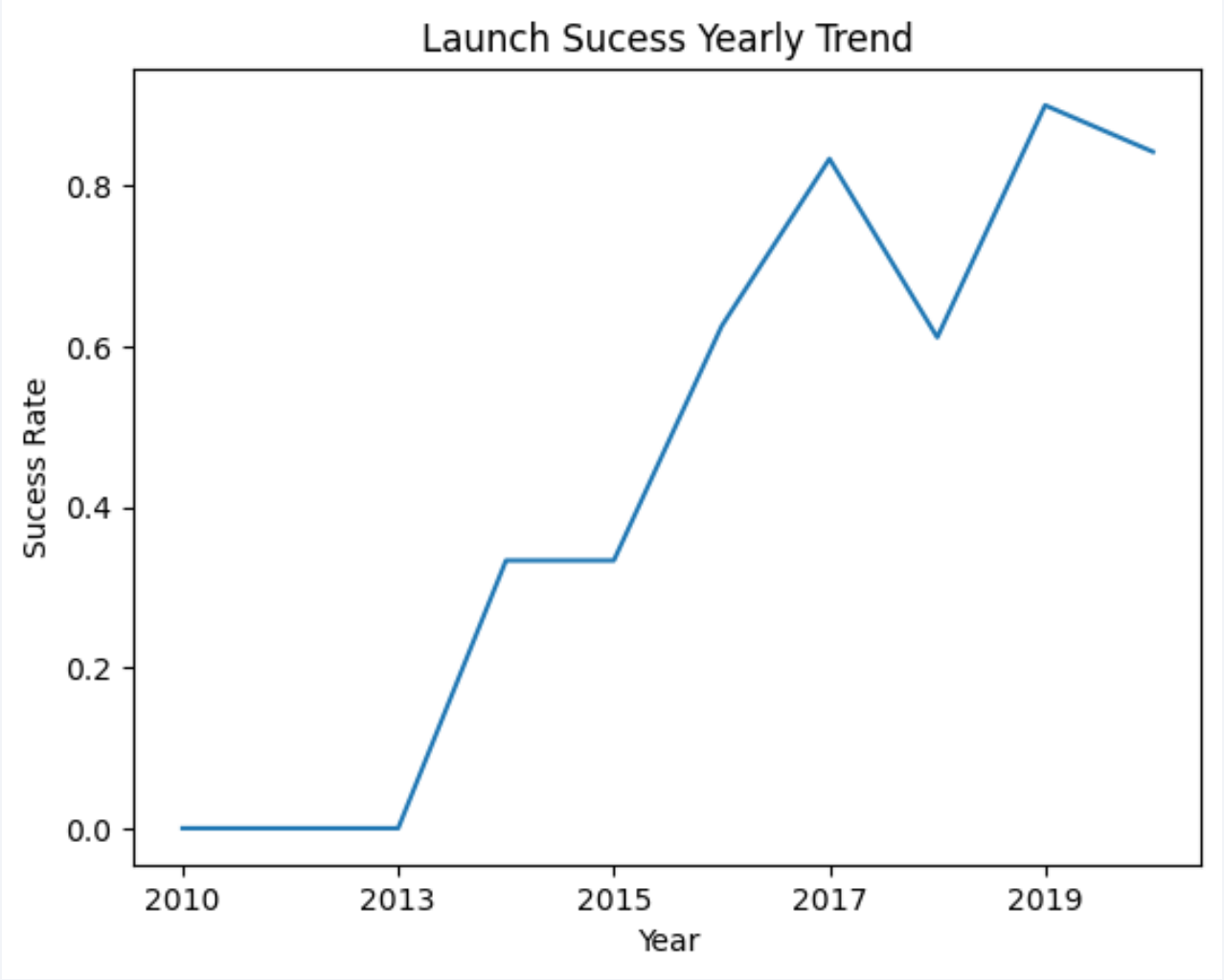
Payload vs. Orbit Type

- There is no noticeable trend for payload vs orbit type. I would further explore this using clusters.
- Blue indicates landing failure, while Orange indicates landing success.



Launch Success Yearly Trend

- After 2013, the success rate has steadily been increasing until 2018, where there was a dip.



All Launch Site Names

- Display of the unique launch sites
  - SELECT DISTINCT Launch_site FROM SPACEXTBL;
- SELECT DISTINCT will select the unique launch site names from the launch site column

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

Launch Site Names Begin with 'CCA'

- Display 5 records where the launch site begins with ’CCA’
  - SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- The command filters WHERE launch sites have names LIKE CCA, and limits it to 5 records

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 boosters launched by NASA
  - SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
- SUM, sums the payload values from the customers from NASA.

TotalPayloadMass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'
- The method involves AVG(averaging) the payload mass from the specific F9 v1.1 Booster

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - SELECT MIN(Date) AS EarliestDate FROM SPACEXTBL WHERE Landing_Outcome LIKE '%(ground pad)' AND Mission_Outcome = 'Success';
- The MIN command finds the most recent date from the landing outcomes that meet the ground pad a success parameters

EarliestDate
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE '%(drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
- The code asks that the answer is larger than 4000 and smaller than 6000. An alternative way to solve this to use the BETWEEN command.

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes:

```
SELECT
    CASE
        WHEN Mission_Outcome LIKE '%success%' THEN 'Successful'
        ELSE 'Failures'
    END AS OutcomeCategory,
    COUNT(*) AS TotalCount
FROM SPACEXTBL
GROUP BY OutcomeCategory;
```

OutcomeCategory	TotalCount
Failures	1
Successful	100

- The case statement checks if the mission outcome contains the string success, then labels it as outcome category (success, failure) and counts the total number of records in each category.

Boosters Carried Maximum Payload

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

```
SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
);
```

- The subquery finds the maximum payload mass from the table and then becomes filtered to find the booster version that has that mass.

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - SELECT Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure (drone ship)' AND Date BETWEEN '2015-01-01' AND '2015-12-31';
- Filter through the data to find landing outcomes like failure and between year of 2015.

Booster_Version	Launch_Site	Landing_Outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

```
SELECT Landing_Outcome, COUNT(*) AS OutcomeCount
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY OutcomeCount DESC;
```

- The where statement selects for the dates, group by groups by landing outcome, and order by put them in descending order.

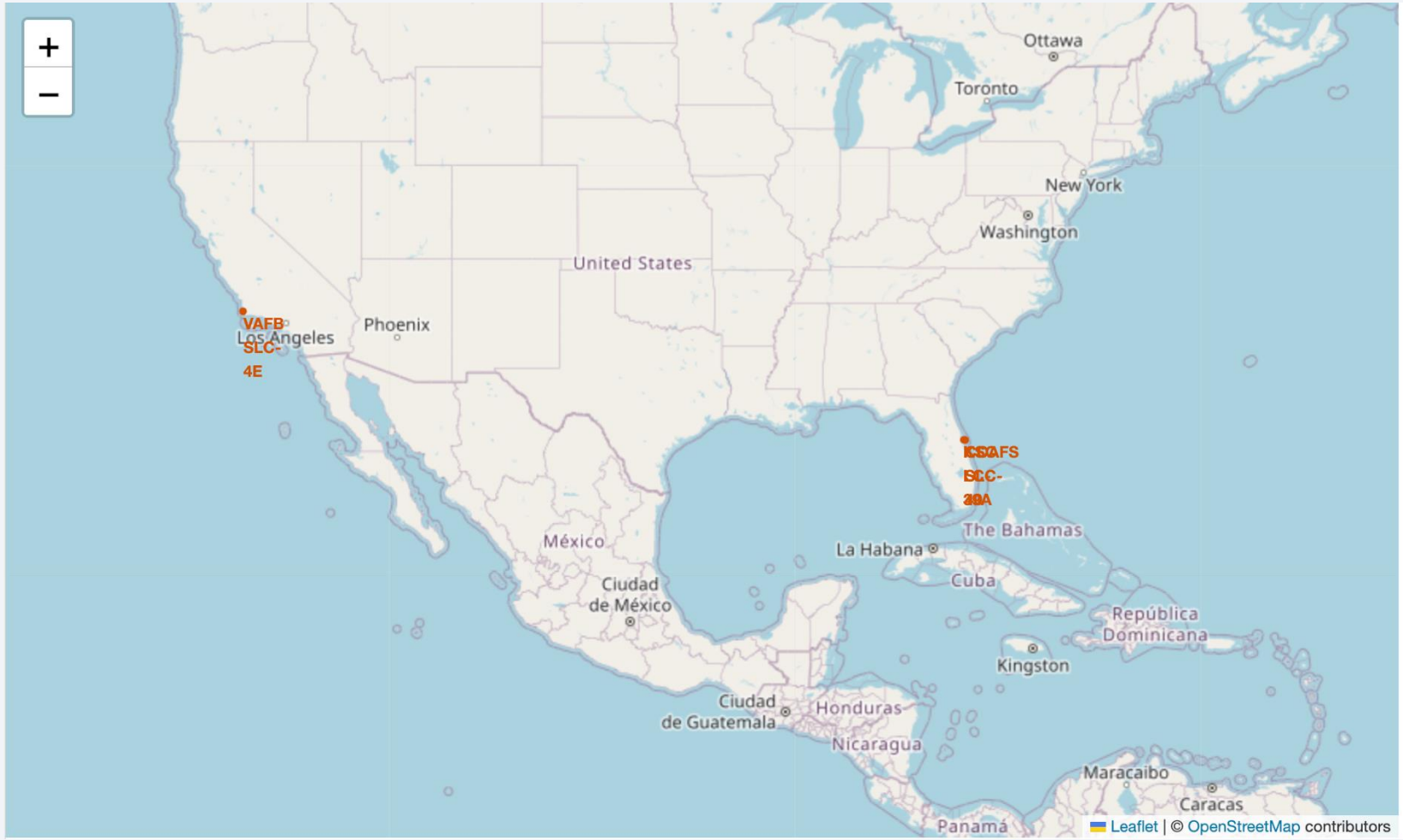
Landing_Outcome	OutcomeCount
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

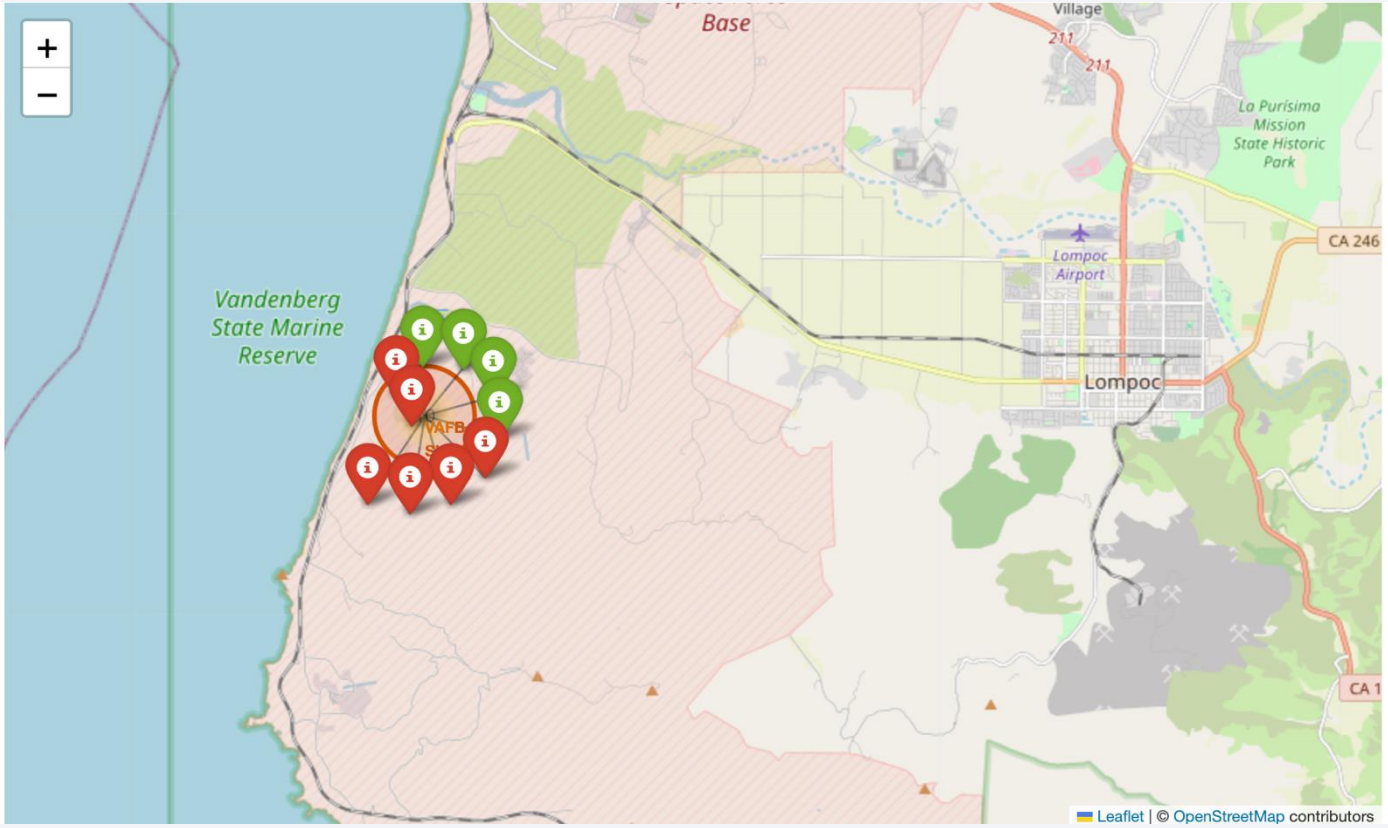
All Launch Sites

- Vanderberg Space Force Base (VAFB) is located in California.
- Cape Canaveral (CCA) and Kennedy Space Center (KSC) are both located in Florida.



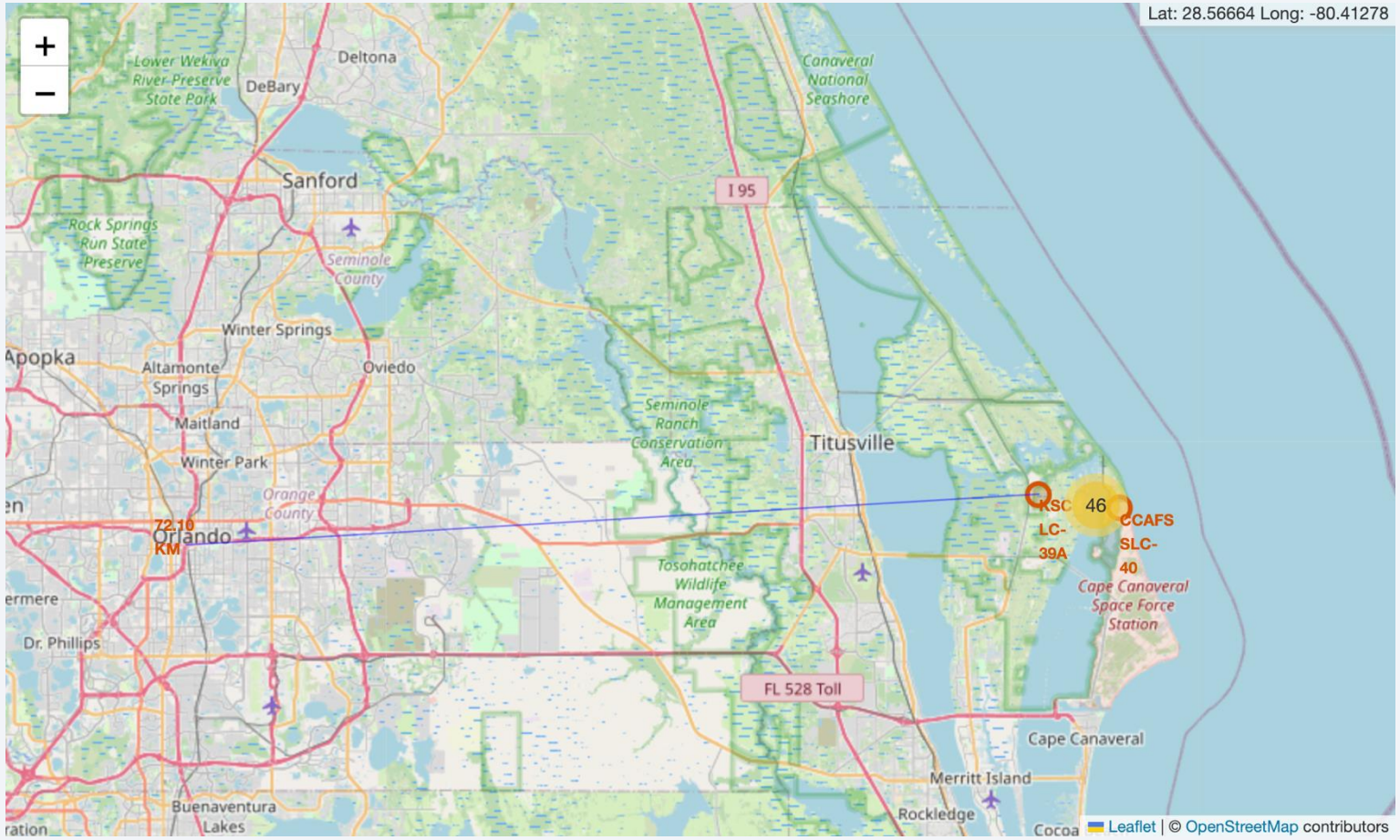
Successful Landings Within a Launch Site

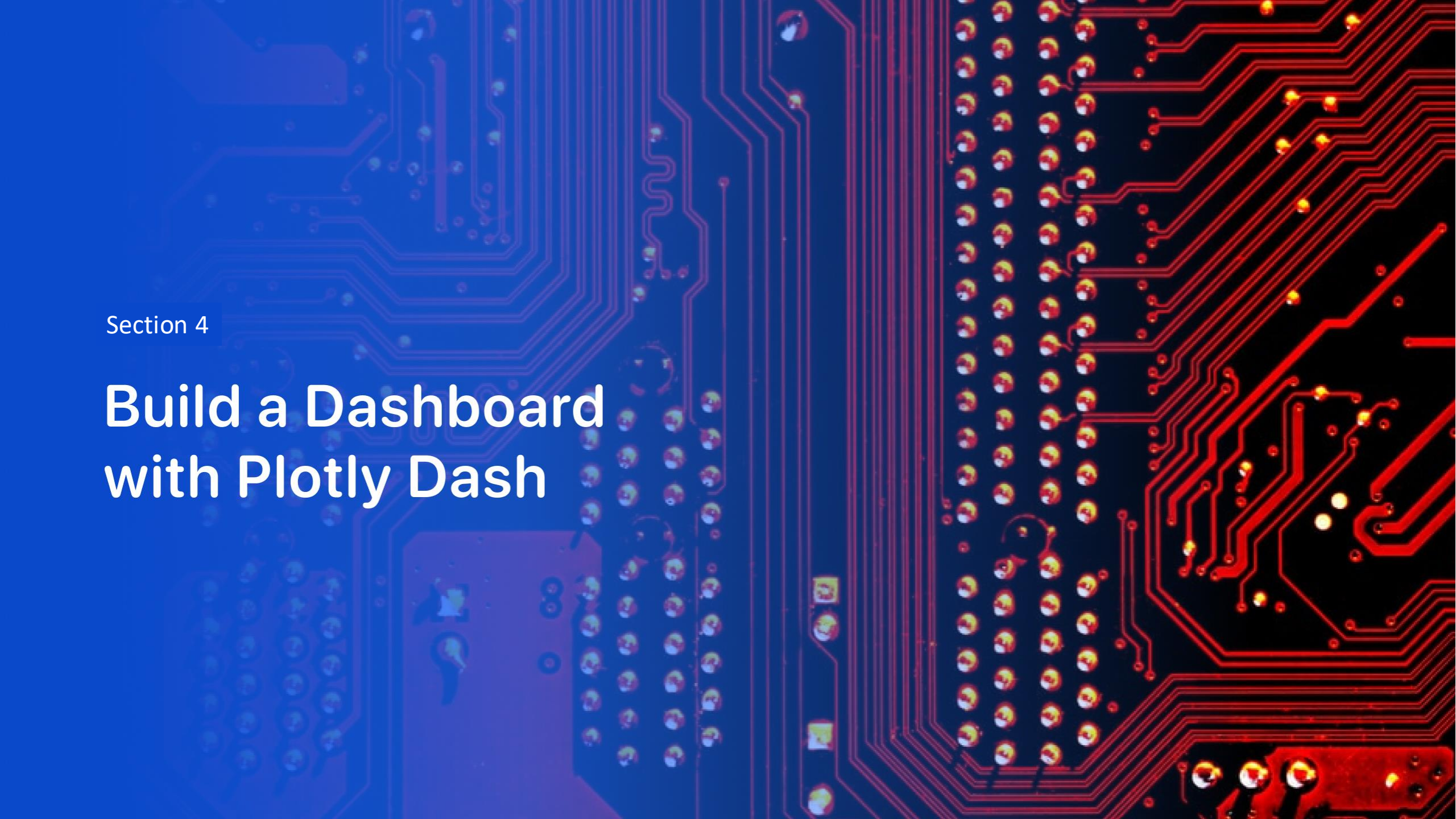
- A close-up view of the landing outcomes at Vandenberg Space Force Base.
- Red indicates an failed landing and green indicates a successful landing.



Distance Site Proximities

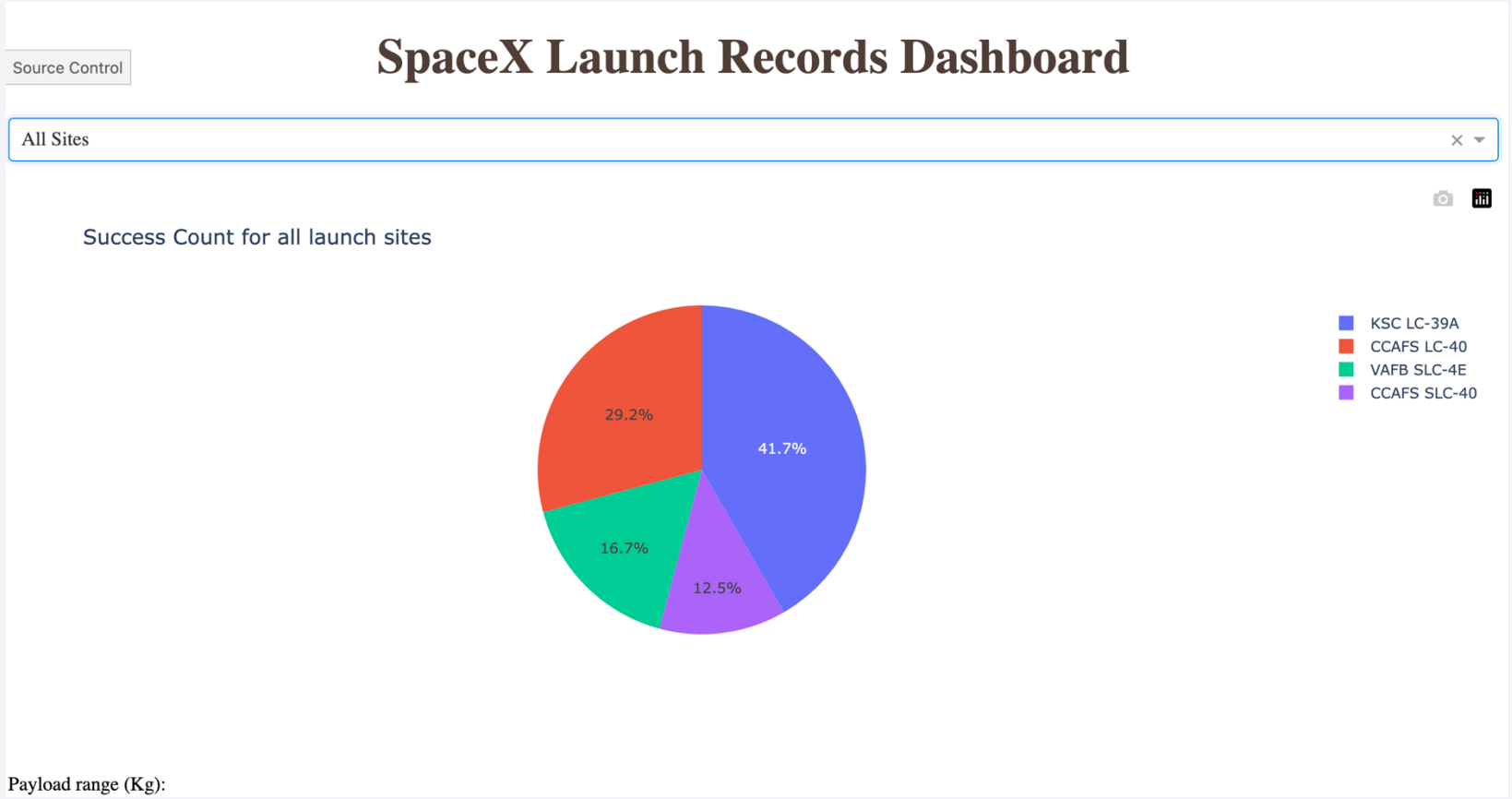
- Proximity of Kennedy Space Center (KSC) to Orlando.
- The distance is 72.10 Km.




Section 4
Build a Dashboard with Plotly Dash

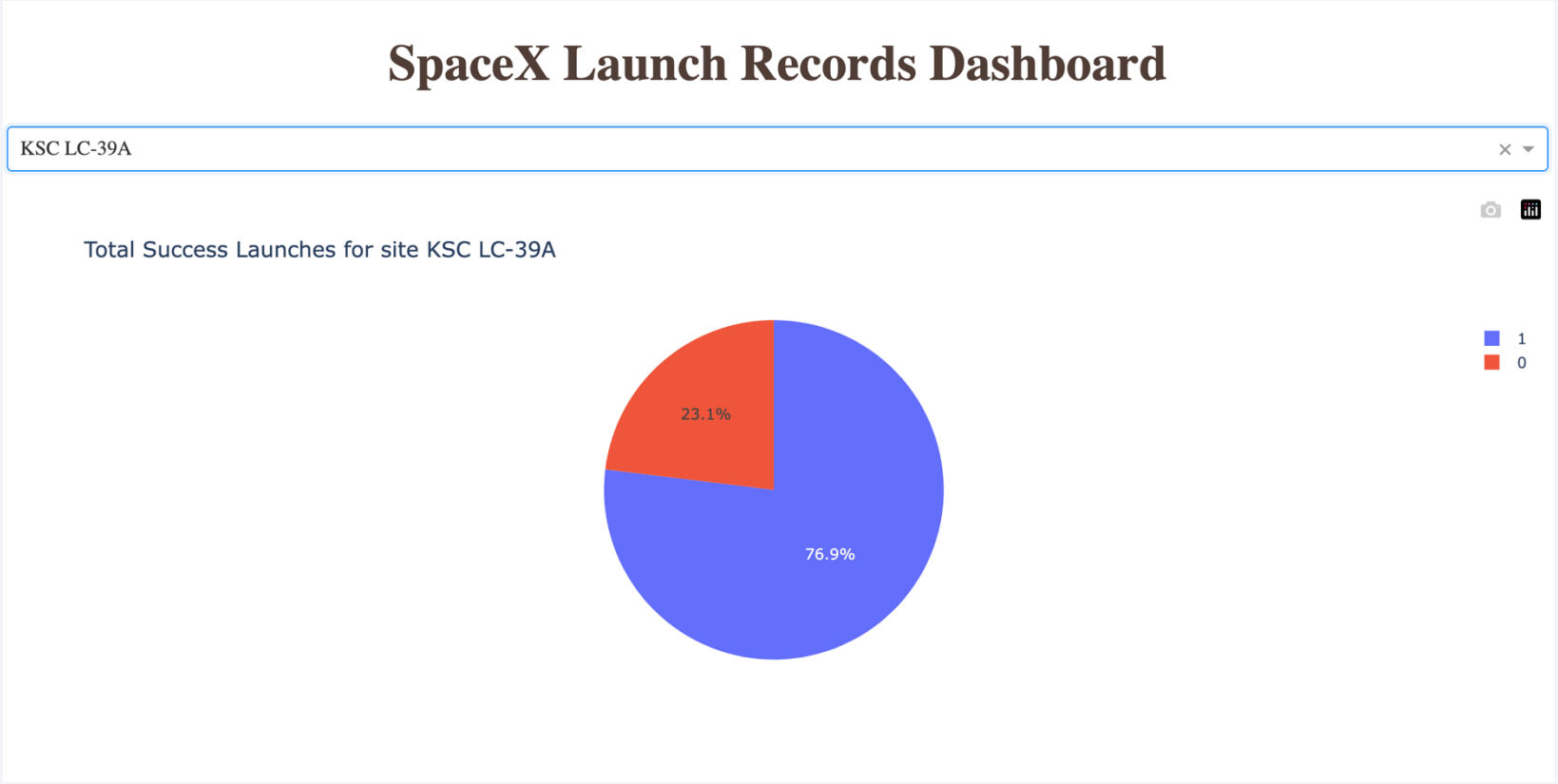
Launch Success for all Launch Locations

- Kennedy Space Center (KPC) has the highest percentage of successful landings.



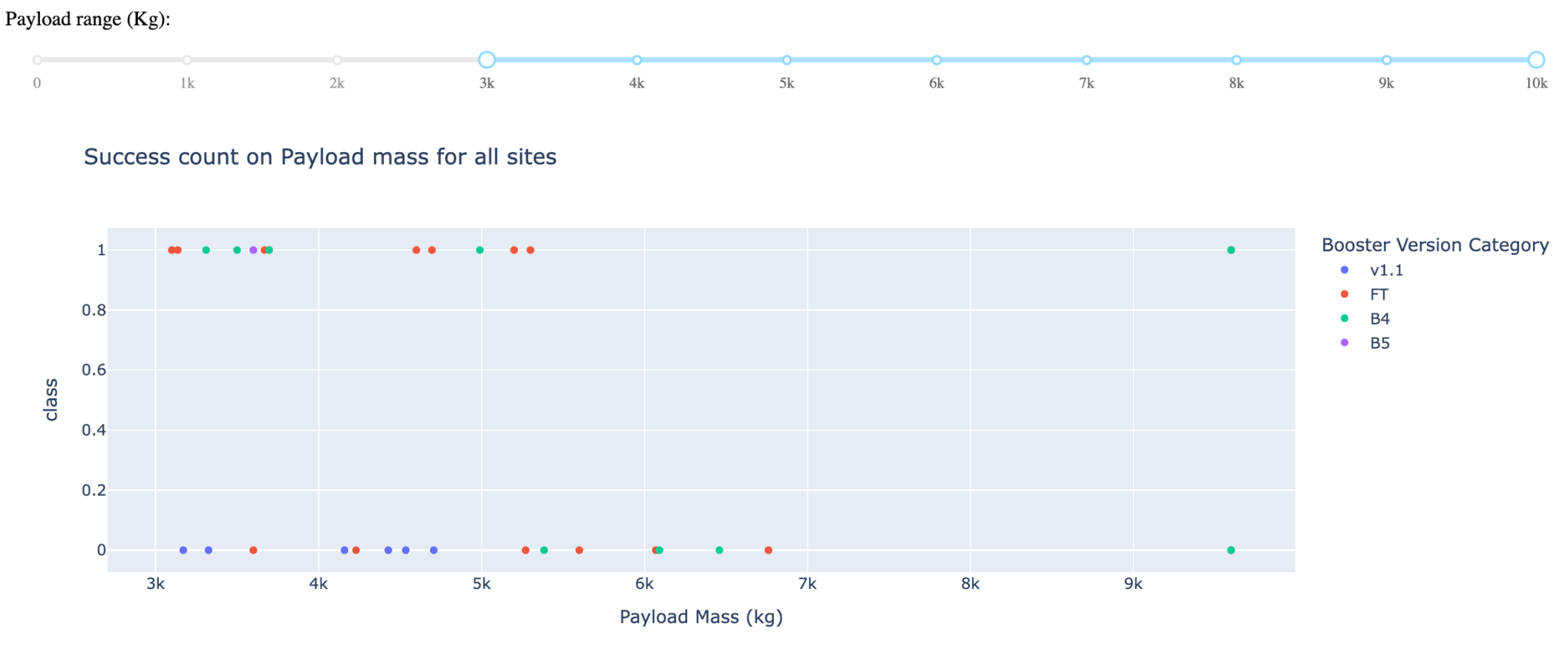
Success Rate for the Highest Success Location KSC

- The Kennedy Space Center Launch Pad has a successful landing for Falcon 9 mission at 76.9 percent.



Influence of Payload Mass on Success Rates

- The FT booster has the most overall Success.
- Lower payload masses seem to lead to a higher chance of successful landings.



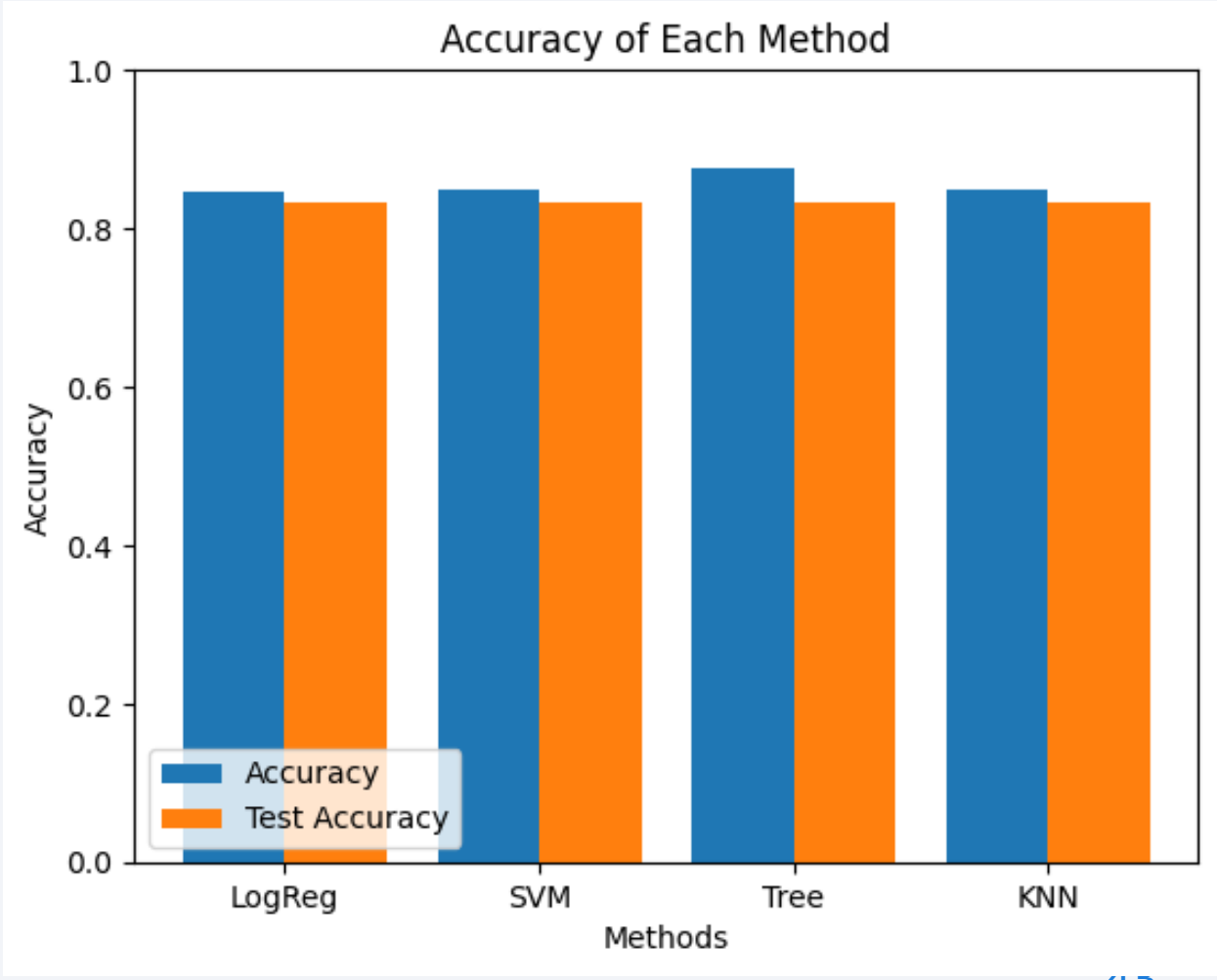
Section 5

Predictive Analysis (Classification)

Classification Accuracy

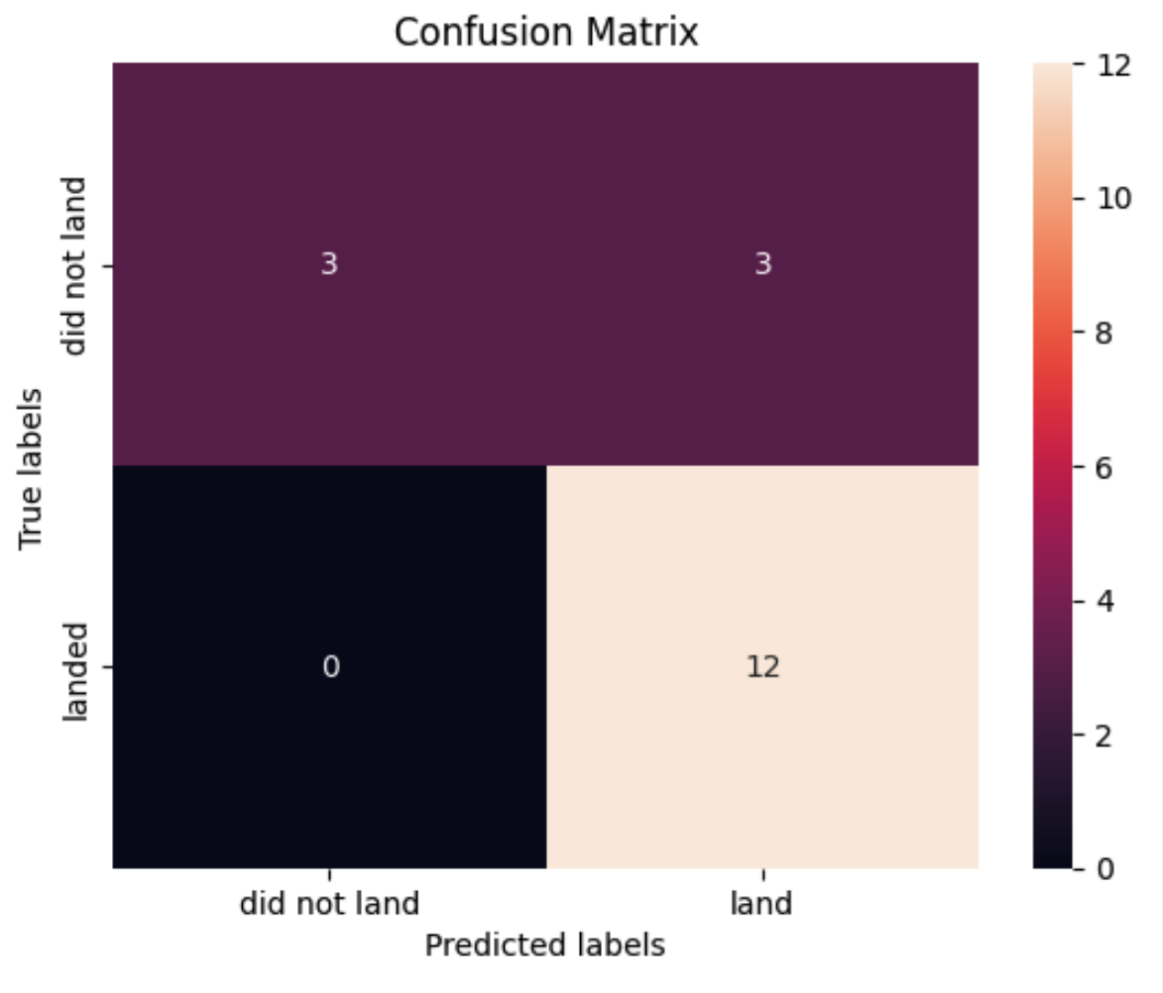
- The best training accuracy mode is the Decision tree , with a score of 87.5%.
- All models, however, had the same Test Accuracy at 83.3%.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.8333333333333334
SVM	0.84821	0.8333333333333334
Tree	0.875	0.8333333333333334
KNN	0.84821	0.8333333333333334



Confusion Matrix

- True Positive (TP): 12 cases where the decision tree correctly predicted: “landed.”
- True Negative (TN): 3 cases where the decision tree correctly predicted: “did not land”.
- False Positive (FP): 3 cases where the decision tree incorrectly predicted “landed” when it was “did not land”.
- False Negative (FN): 0 cases where the decision tree incorrectly predicted “did not land” when it was “landed”.
- The accuracy of this model was 83.3%



Conclusions

- Successful landing rates for Falcon 9 Rockets have increased over time. This could be due to better technological advancements.
- The most successful launch site is Kennedy Space Center (SPC), with a success rate of 76.9%.
- The best model is the Decision Tree, with a training accuracy of 87.5% and a test accuracy of 83.3%.

Appendix

- Coursera: https://www.coursera.org/professional-certificates/ibm-data-science
- Repository: https://github.com/NikkideVries/IBM-SpaceX-Capstone/tree/main
- GitHub links:
  - Collection API: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/1Data_Collection_API.ipynb
  - Collection Web Scrapping: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/2Data_Collection_with_Web_Scrapping.ipynb
  - Data Wrangling: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/3Data_Wrangling.ipynb
  - EDA with SQL: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/4EDA_with_SQL.ipynb
  - EDA With Visualizations: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/5EDA_with_Visualizations.ipynb
  - Visuals with Folium: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/6Interactive_Visuals_with_Folium.ipynb
  - Dashboard: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/7Interactive_Dashboard_with_Ploty.py
  - Machine Learning Predictions: https://github.com/NikkideVries/IBM-SpaceX-Capstone/blob/main/8Machine_Learning_Prediction.ipynb

Thank you!