IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

NAHEEM FARRI

JUNE 28TH, 2024

Outline

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

Executive Summary

- Data collection methodology:
  - Requested data from the API
  - Cleaned by filtering and removing the null values
- Perform data wrangling
  - Scraped launch data using "BeautifulSoup"
  - Created a dataframe
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Explored data using magic sql - "WHERE" clause, "GROUP BY" function, etc

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
  - Plotted landing points on a map with folium
  - Created a dashboard with Plotly Dash
- Perform predictive analysis using classification models
  - Trained and tested data
  - Fit them with classification models
  - Found the accuracy of predictions

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.
- In this case we are a company "SpaceY" that is hoping to bid successfully against SpaceX
- We are tasked to determine if the first stage will land which will determine the cost of a launch.

Section 1

Methodology

Methodology

- We made use of the SpaceX API to get launch data for rocket launches
- We proceeded to sort the data we wanted, cleaned the data and removed null values
- These help us get the data we later use for analysis and visualizations
- We can then get the most successful launch sites and use this as the site for our launches
- This will reduce the cost of launching from sites that produce unsuccessful launches

Data Collection – SpaceX API

Main Process:

- Request data from the launch API.

Sub Processes:

- Request rocket data from launch API
- Request launchpad data from launch API
- Request payload data from launch API
- Request rocket data from launch API

- **URL:**
- https://github.com/NAHEEM765/ADS_Capstone/blob/main/01-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Main Process:

Web scraping launch records from html table on Wikipedia

Sub Processes:

- Request the Falcon9 Launch Wiki page from its URL using get request
- Get the desired data using the BeautifulSoup package
- Creating a dataframe by parsing the HTML launch tables in the soup file

URL:

https://github.com/NAHEEM765/ADS_Capstone/blob/main/01-2-jupyter-labs-webscraping.ipynb

Data Wrangling

- MAIN: Exploratory data analysis, Determine Training labels
- Calculate number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome each orbit
- Create landing outcome labels

- GitHub URL:

https://github.com/NAHEEM765/ADS_Capstone/blob/main/02-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Visualize the relationship between Flight Number and Launch Site, scatterplot
- Visualize the relationship between Payload and Launch Site, scatterplot
- Visualize the relationship between success rate of each orbit type, bar chart
- Visualize the relationship between FlightNumber and Orbit type, scatterplot
- Visualize the relationship between Payload and Orbit type, scatterplot
- Visualize the launch success yearly trend, line plot
- Features Engineering
- Create dummy variables to categorical columns

EDA with Data Visualization

- GitHub URL:

https://github.com/NAHEEM765/ADS_Capstone/blob/main/04%20Exploring_and_preparing_data.ipynb

EDA with SQL

- 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 succesful landing outcome in ground pad was acheived.
- 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. Use a subquery
- 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.

EDA with SQL

- GitHub URL:
- https://github.com/NAHEEM765/ADS_Capstone/blob/main/03-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Interactive Map with Folium

Added Objects to maps:

- Markers: Marker1, Marker2, Marker3, Marker4, marker_cluster
- Circles: circle, circle1, circle2, circle3, circle4
- Lines: lines

Purpose of added those objects:

- The markers were added to put the names of the stations on the map(VAFB SLC-40, CCAFS LC-40, CCAFS SLC-40, KSC LC-39A)
- The circles were used to add a dot on the exact point the locations were
- The lines were used to show the distances between the points
- Marker cluster: Used to show the number of launches in each site, on clicking show the number of successful(green) launches and the number of unsuccessful(red) launches

Interactive Map with Folium - Link

GitHub URL:

https://github.com/NAHEEM765/ADS_Capstone/blob/main/06-lab_jupyter_launch_site_location.ipynb

Dashboard with Plotly Dash

Graphs/Interactions Added:

- Pie Chart, Scatterplot with grid, Range Slider

Added those plots and interactions:

- Pie chart - displays successful launches vs unsuccessful launches for launch sites
- Scatterplot: Plots the class (0 or 1) vs payload mass
- Range Slider: To change the range of payload masses

Dashboard with Plotly Dash - Link

- GitHub URL :

https://github.com/NAHEEM765/ADS_Capstone/blob/main/06-spacex_dash_app.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- Created an array of the column class
- Standardized and then transformed the model
- Trained and tested the data
- Used GridSearchCV with logistic Regression, Support Vector Machine, Decision Tree Classifier, KNN objects to find the best parameters, and the best score within each model
- Created a confusion matrix to find out the problem with each model
- Compared the methods using their scores and get the one that performs the best

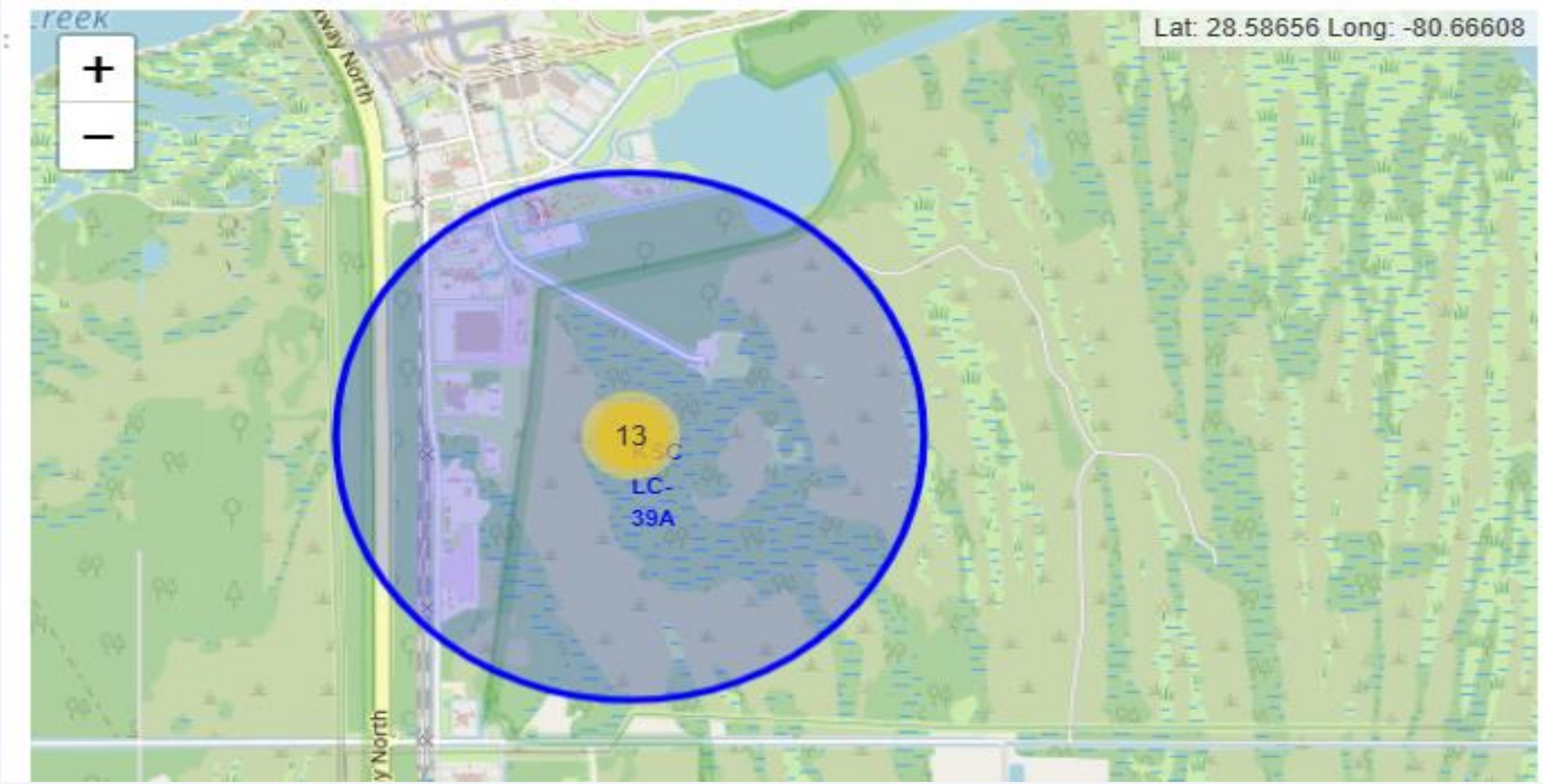
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results - Exploratory Data Analysis

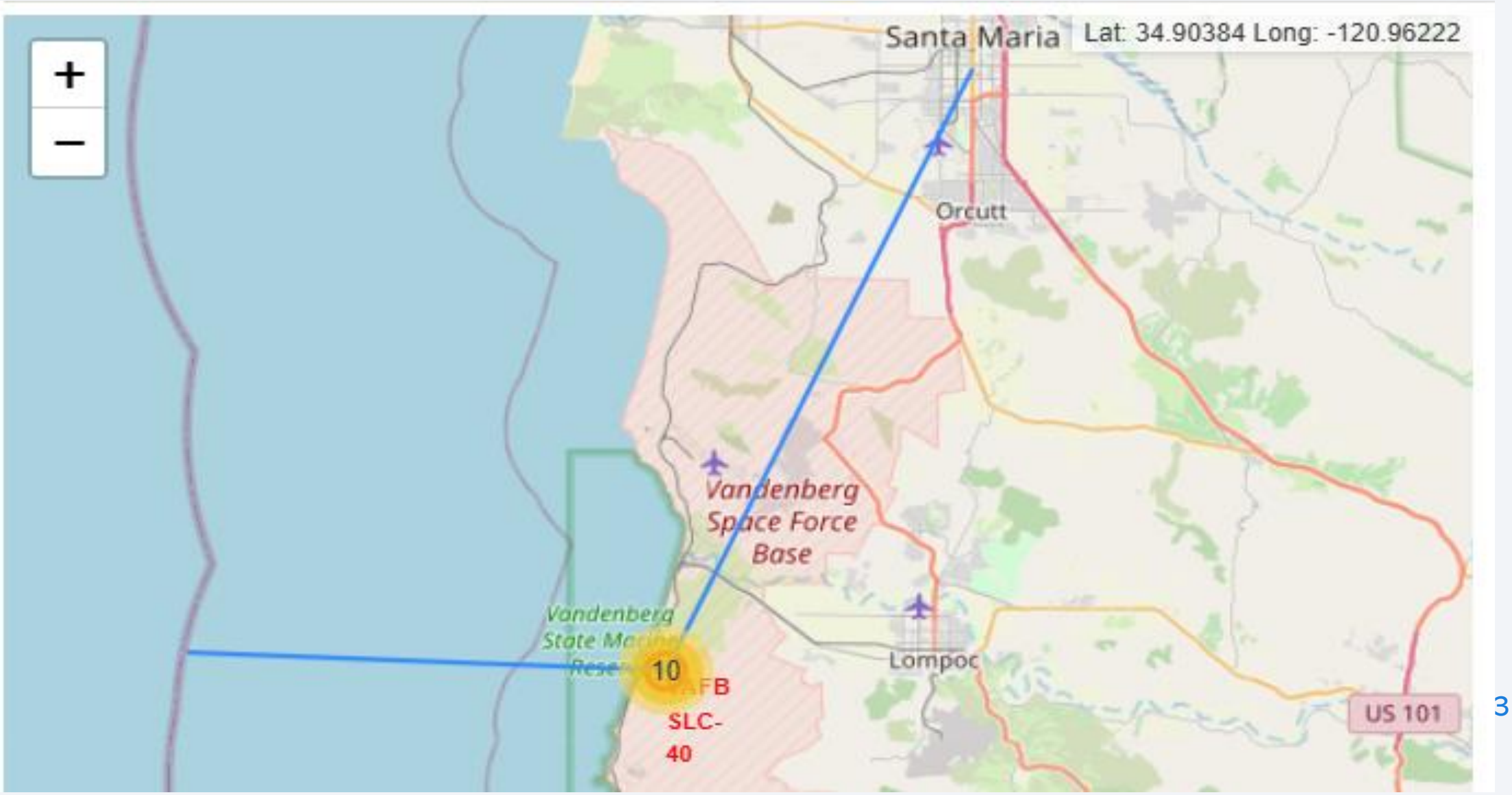
- Exploratory data analysis results
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- *#Orbits ES-L1, GEO, HEO and SSO are the most successful orbits (100%)*
- in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- the sucess rate since 2013 kept increasing till 2020

Results - CCAFS LC-40 and CCAFS LC-40

Results – KSC LC-39A - Visual Analytics



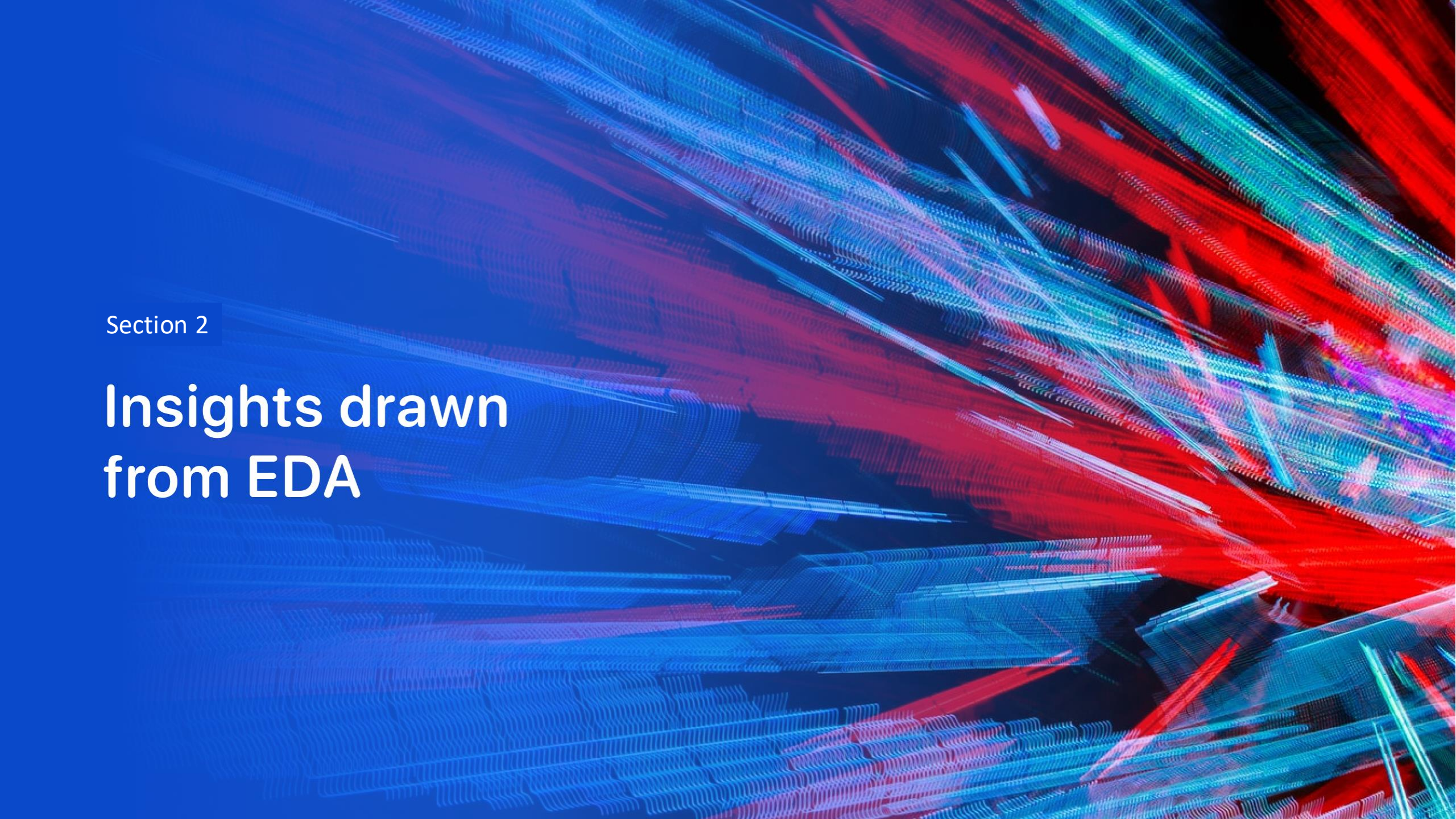
Results - VAFB SLC-40 -Visual Analytics



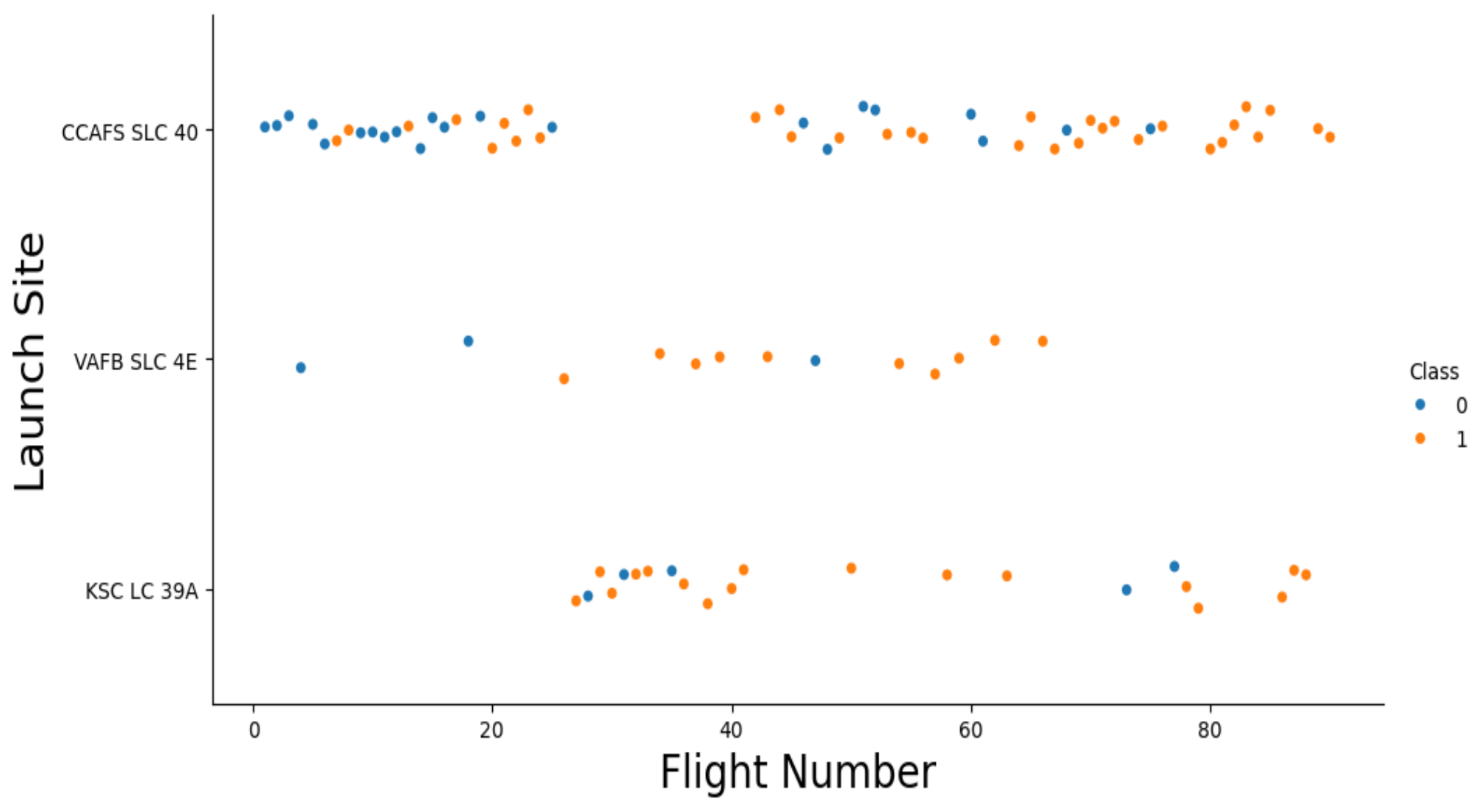
Results - Predictive Machine Learning Analysis

Predictive analysis results

- Best Hyperparameter, Accuracy on test data:
- Lr: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}, 0.8333333333333334
- Svm: `{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}, 0.8333333333333334`
- Decision Tree: `{'criterion': 'entropy', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'}, 0.7222222222222222`
- KNN: `{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}, 0.8333333333333334`


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site - Scatterplot



Flight Number vs. Launch Site - Details

From the scatterplot of Flight Number Vs. Launch Site

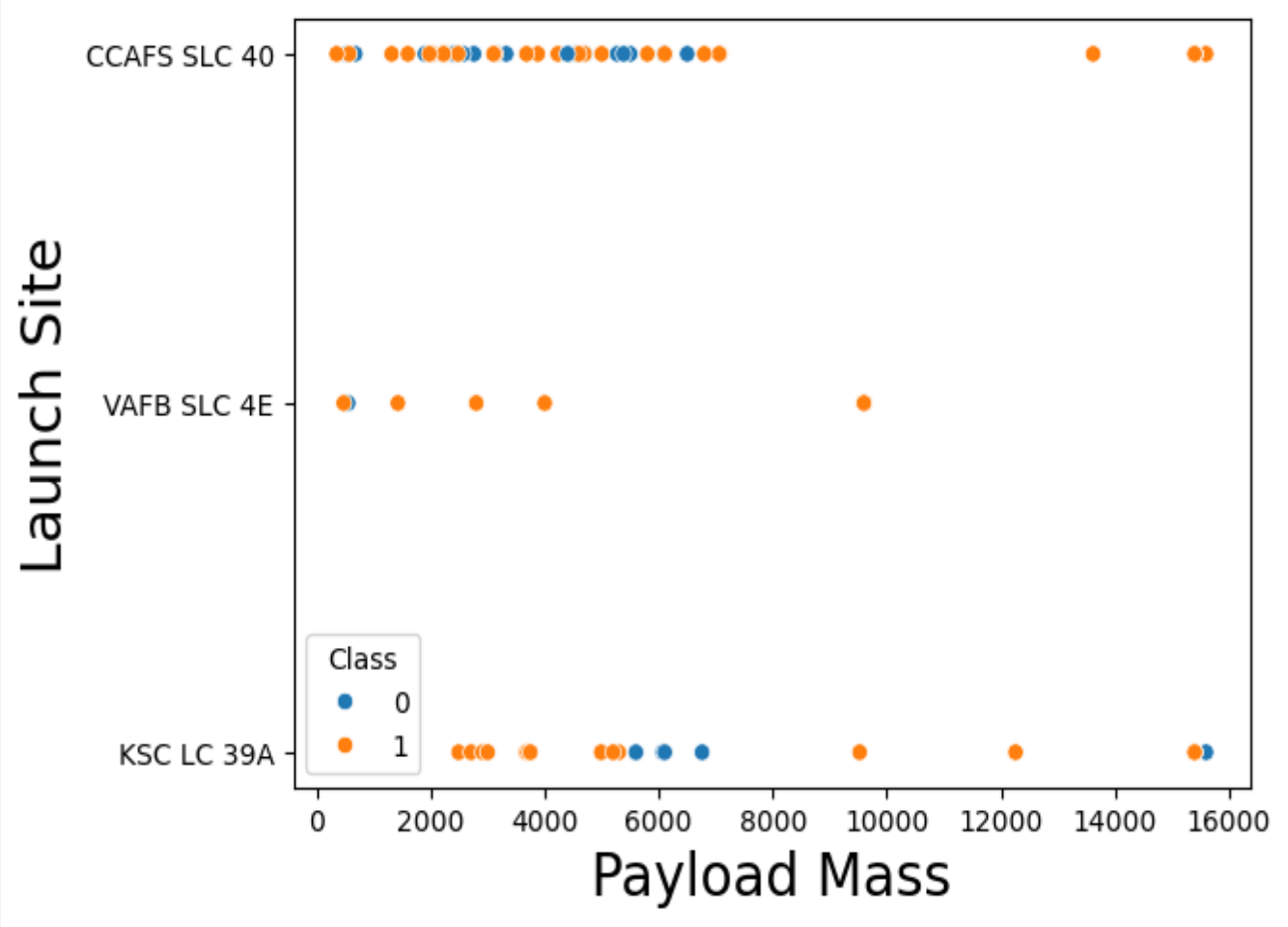
- With the blue dots indicating the class of Successful launches (0)
- With the orange dots indicating the class of Unsuccessful launches (1)
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Payload vs. Launch Site

From the scatterplot of Payload Mass Vs. Launch Site :

- With the blue dots indicating the class of Successful launches (0)
- With the orange dots indicating the class of Unsuccessful launches (1)

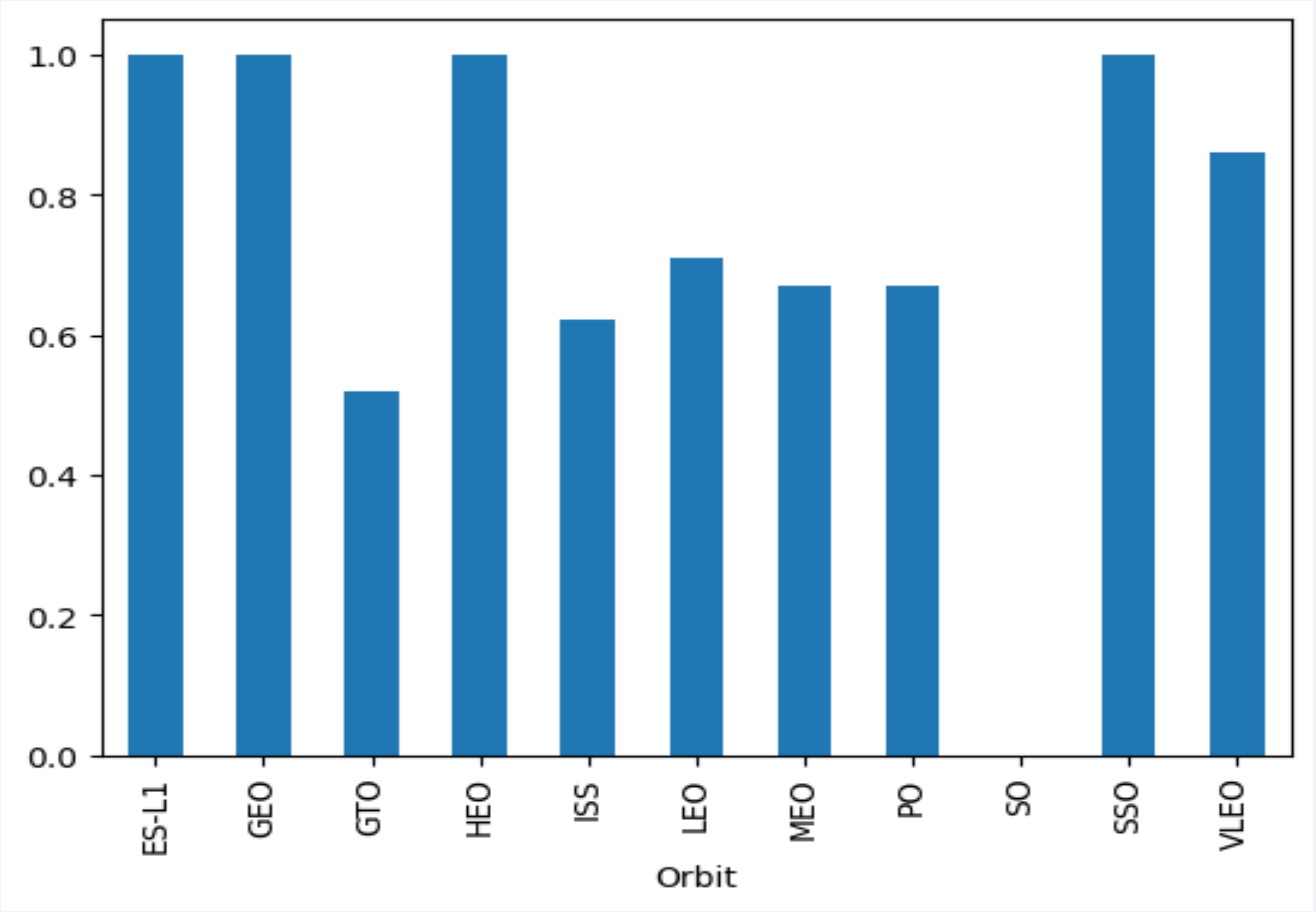
- The launch site VAFB SLC-40 does not use payload masses above 10000kg with mostly successful launches(80%)



Success Rate vs. Orbit Type

From the bar plot of Success Rate Vs. Orbit Type:

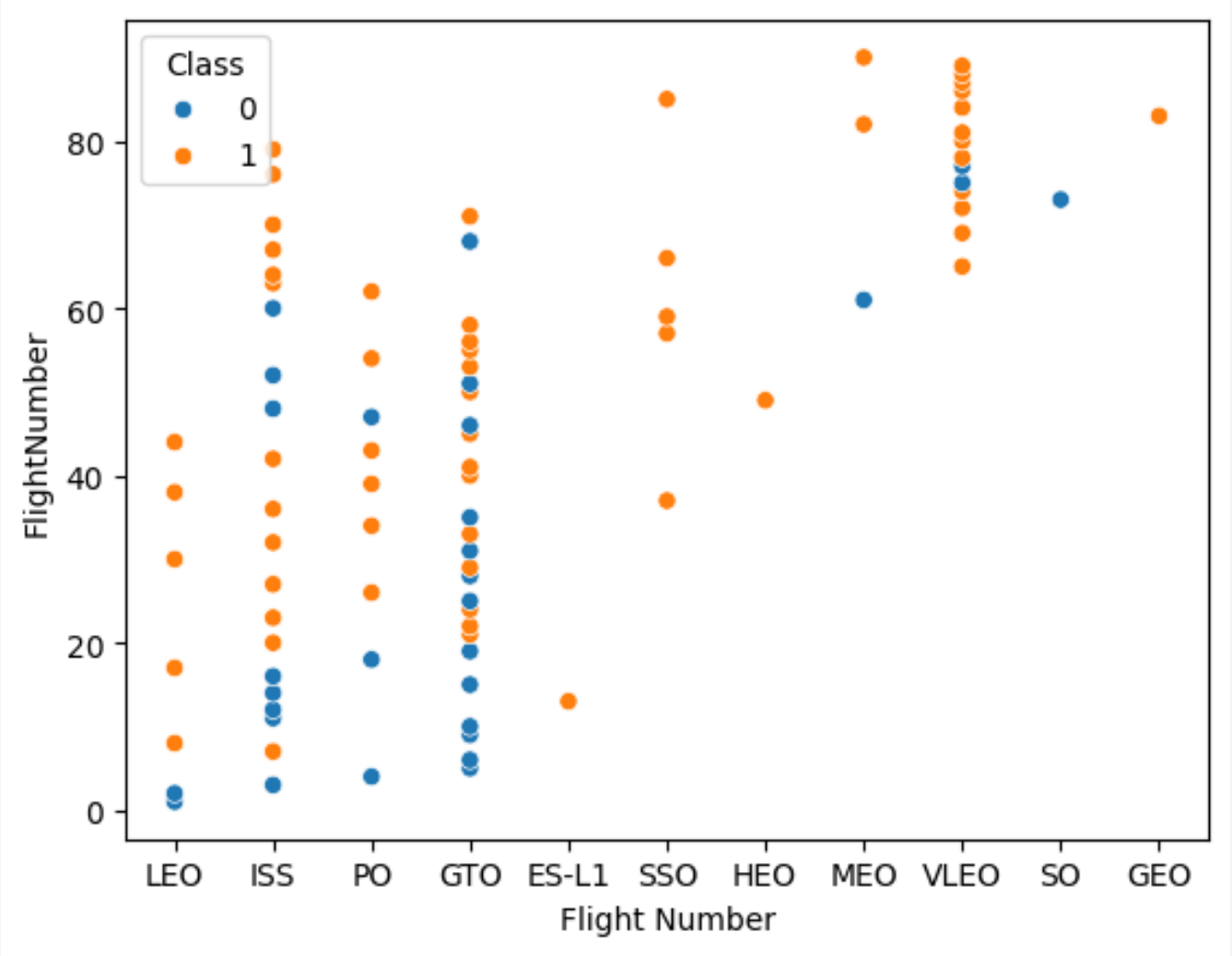
- The success rates (between 0 and 1) on the vertical axis
- The Orbits on the horizontal axis
- Orbits ES-L1, GEO, HEO and SSO are the most successful orbits (100%)
- All have 1 lauch each except "SSO" with 5 launches



Flight Number vs. Orbit Type

From the plot of Flight number vs. Orbit type:

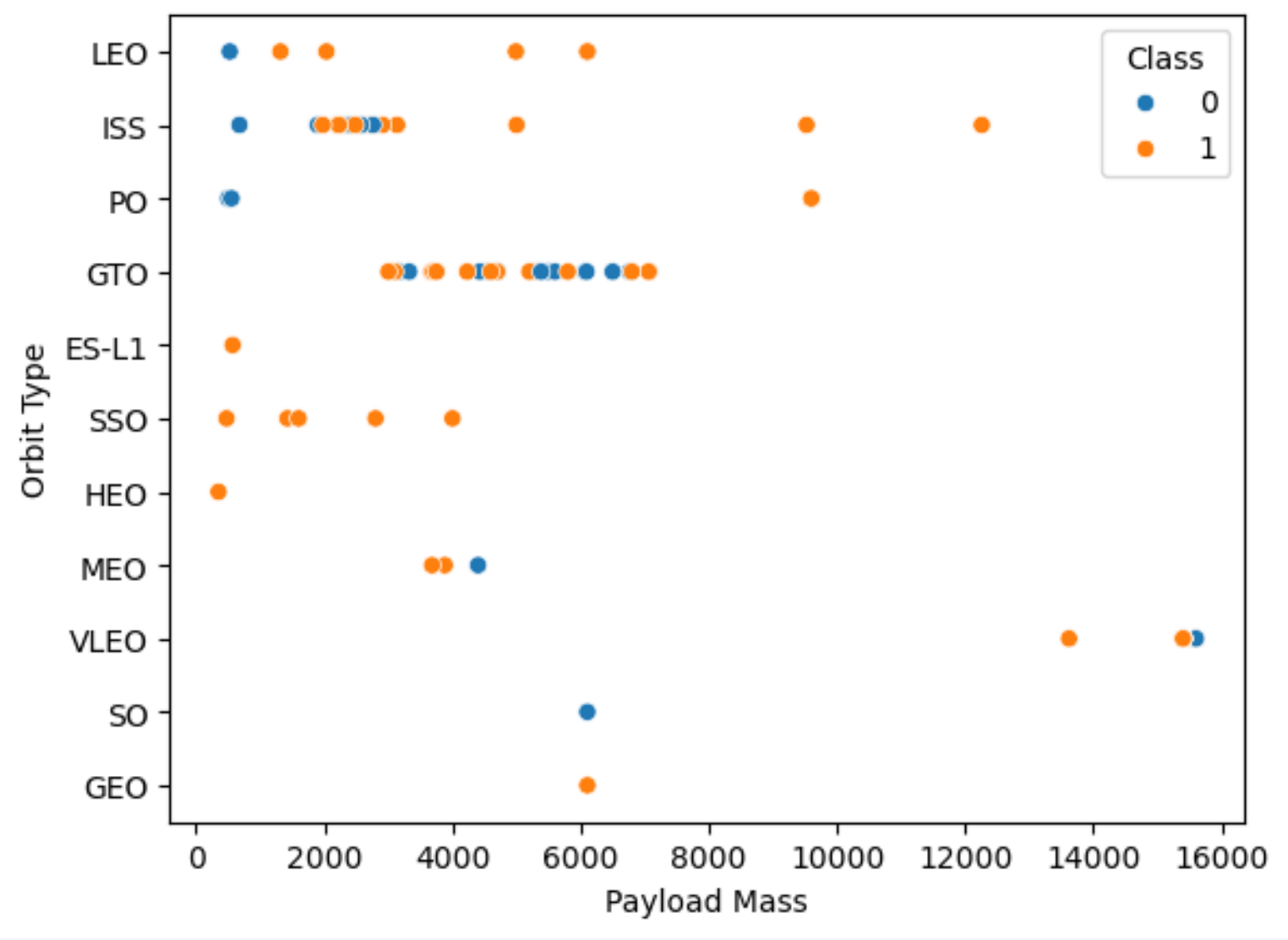
- In the LEO orbit the Success appears related to the number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

From the scatterplot of payload vs. orbit type:

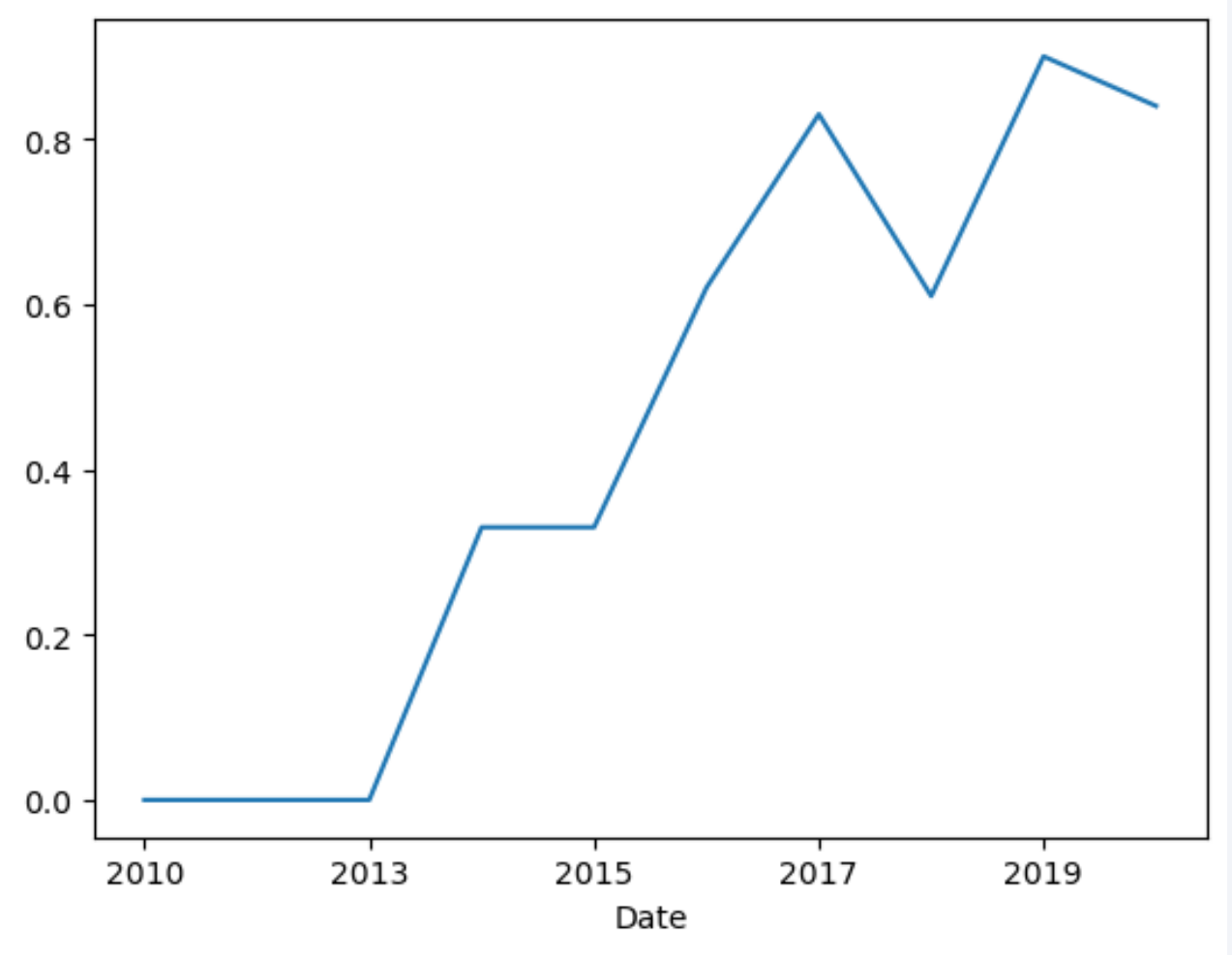
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

From the line chart of yearly average success rate:

- No change up until 2013
- The success rates keep going up until from then increases



All Launch Site Names

- **SQL QUERY:**

"%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE"

- **RESULT OF QUERY:**
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-40

These are the names of the launch without any repitition

Launch Site Names Begin with 'CCA'

SQL QUERY:

"%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5"

RESULT OF QUERY:

[15]:

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

- **SQL QUERY:**

"%sql SELECT SUM('PAYLOAD_MASS__KG_') FROM SPACEXTABLE WHERE 'Customer' = 'NASA (CRS)' "

- **QUERY RESULT:**

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- **SQL QUERY:**

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

- **QUERY RESULT:**

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- **SQL QUERY**:

' %sql SELECT min("Date") as "First Successful Landing Date" FROM SPACEXTABLE '

- **QUERY RESULT:**

Successful Drone Ship Landing with Payload between 4000 and 6000

- **SQL QUERY:**

' %sql SELECT ("Booster_Version") FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000 '

- **QUERY RESULT:**

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

Total Number of Successful and Failure Mission Outcomes

- **SQL QUERY:**

' %sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as "Number of Outcomes" FROM SPACEXTABLE GROUP BY ("Mission_Outcome") '

- **QUERY RESULT:**

Mission_Outcome	Number of Outcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- **SQL QUERY:**

' %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTABLE) '

- **QUERY RESULT:**

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

- **SQL QUERY:**

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

- **QUERY RESULT:**

Month	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

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

- **SQL QUERY:**

' %sql SELECT "Landing_Outcome", COUNT("Landing_Outcome") as "Outcome occurences" FROM SPACEXTABLE GROUP BY "Landing_Outcome" ORDER BY "Outcome occurences" DESC '

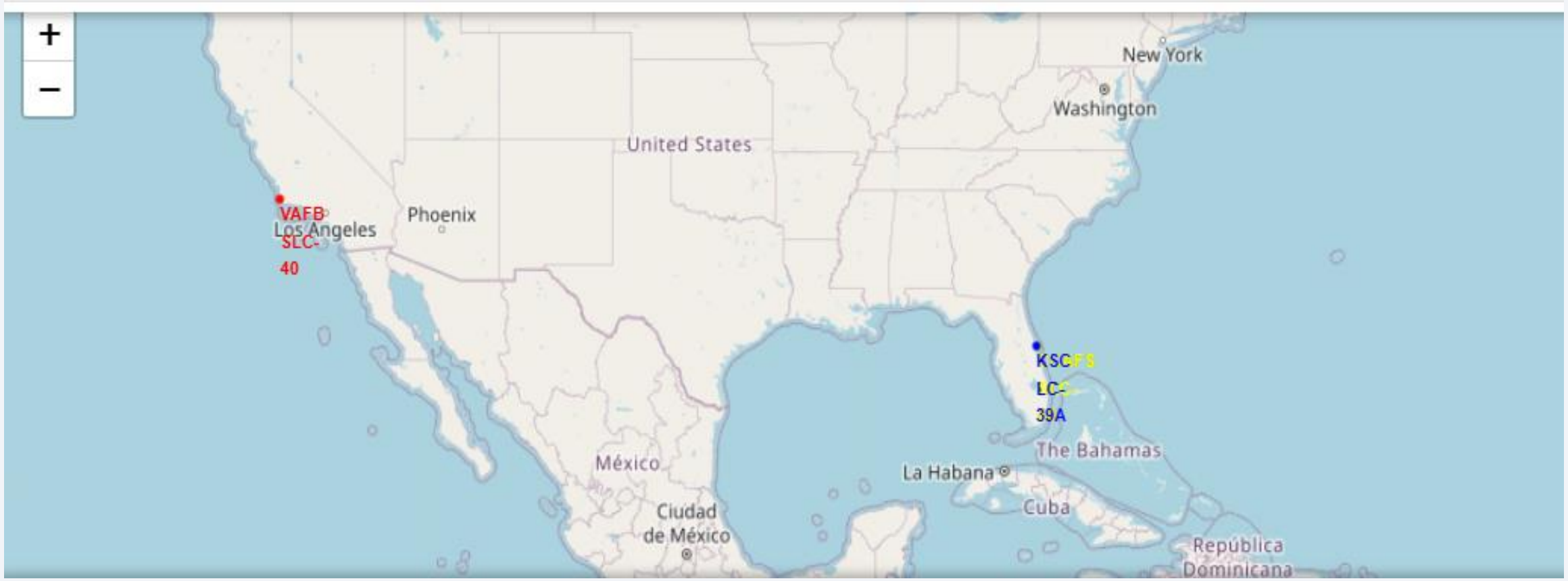
- **QUERY RESULT:**

Landing_Outcome	Outcome occurences
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

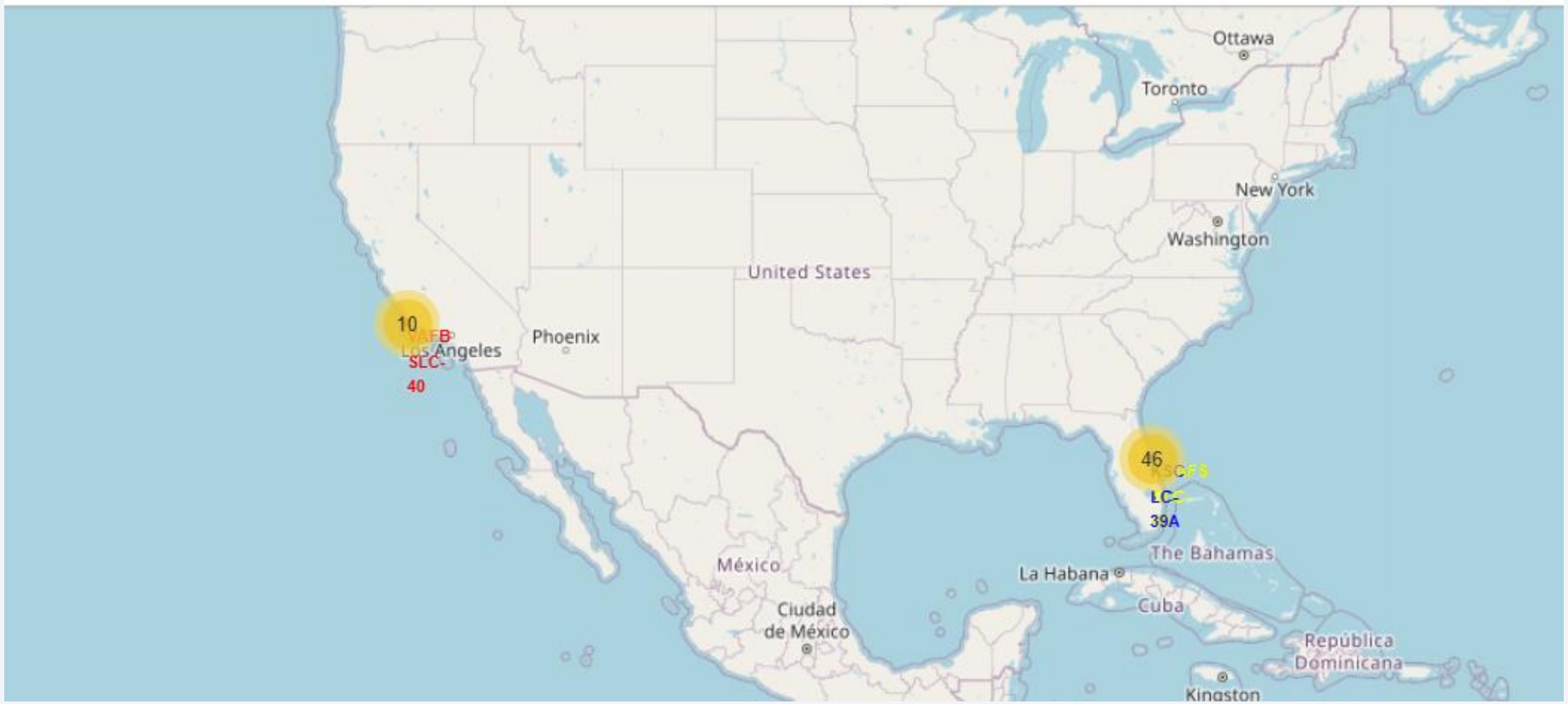
Section 3

Launch Sites Proximities Analysis

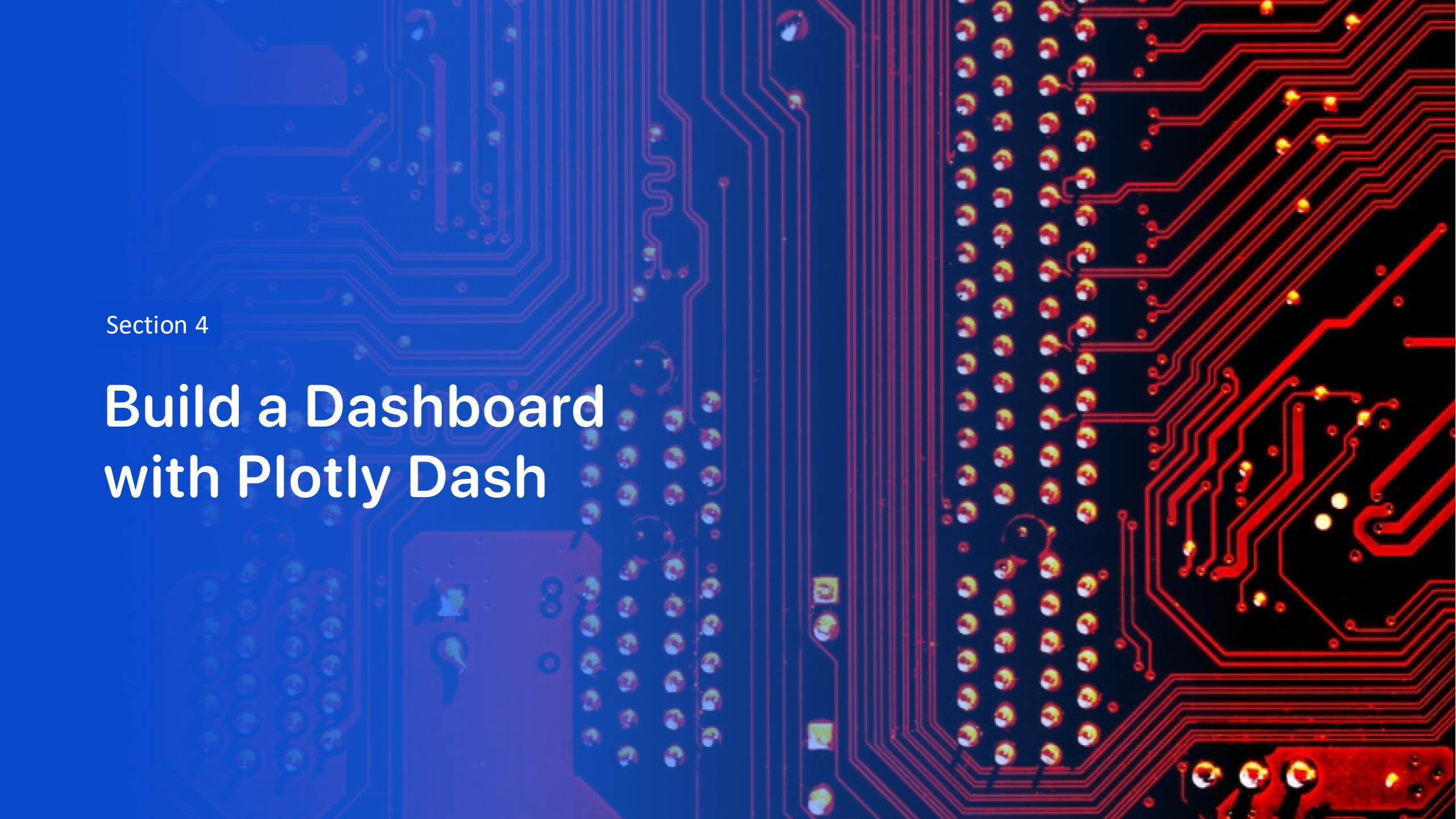
Map Of Launch sites with markers



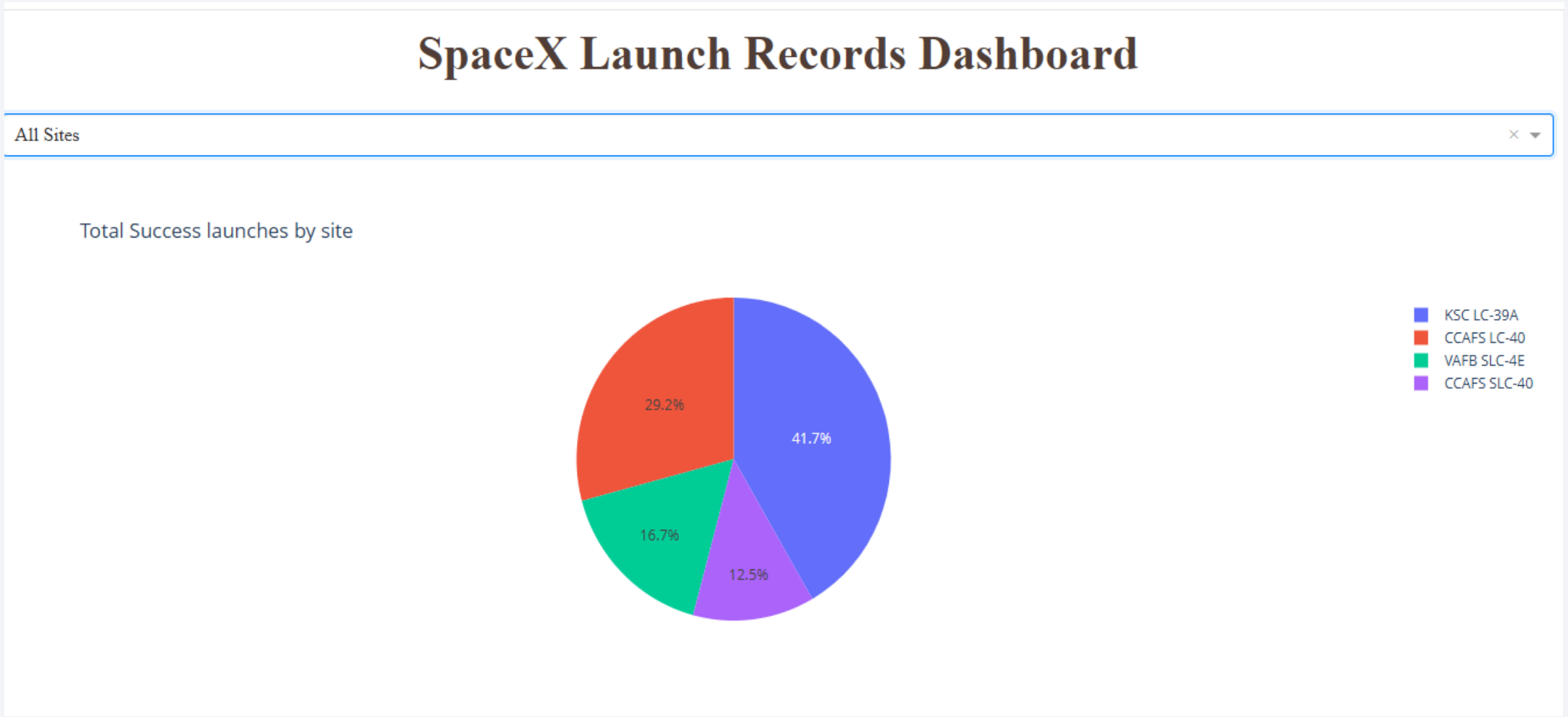
Map of Launch Site with Labeled Launch Outcomes



Proximity of Launch site to nearby locations


Section 4
Build a Dashboard with Plotly Dash

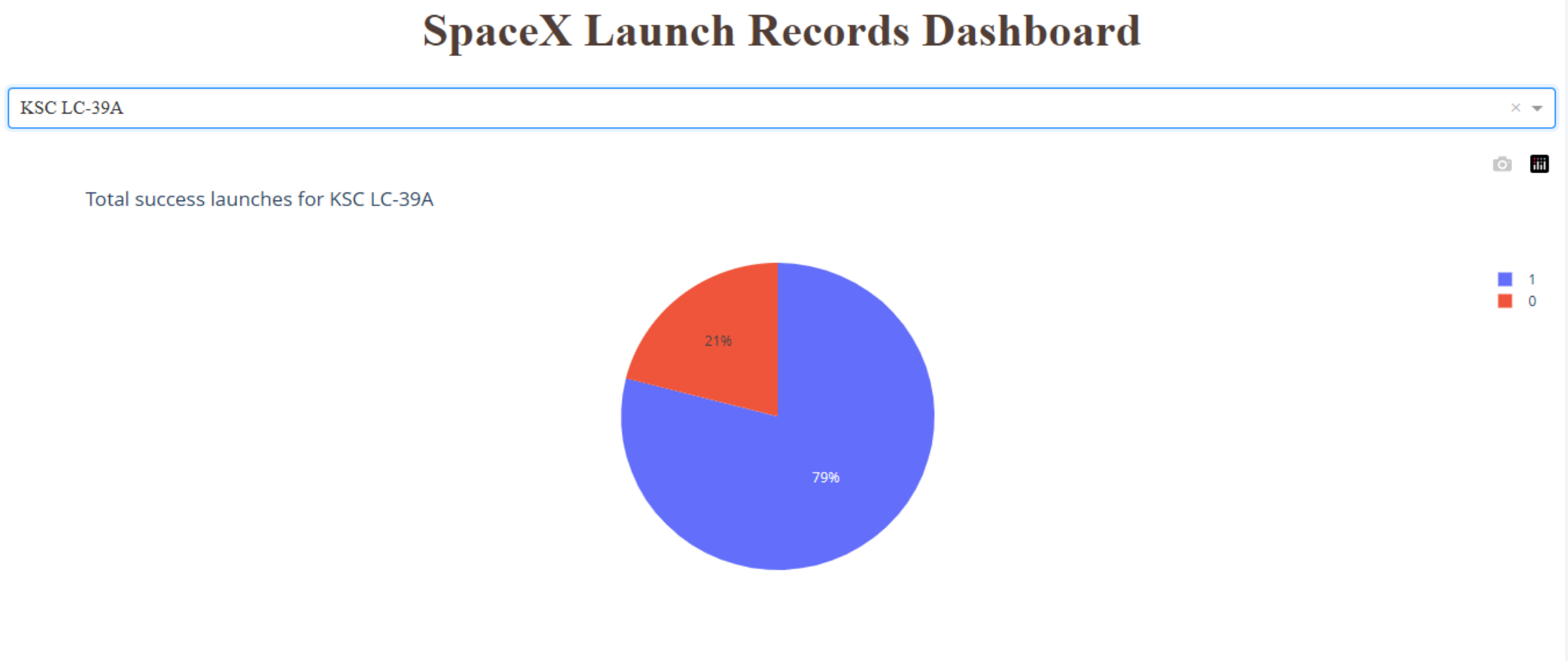
Launch Success For All Sites - Pie Chart



Launch Success For All Sites - Pie Chart

- The launch site KSC LC-39A has the most successful number of launches (41.7%)
- The launch site with the least number of launches is CCAFS SLC-40 (12.5%)

Launch Site with the highest launch success rate KSC LC-39A (41.7%)

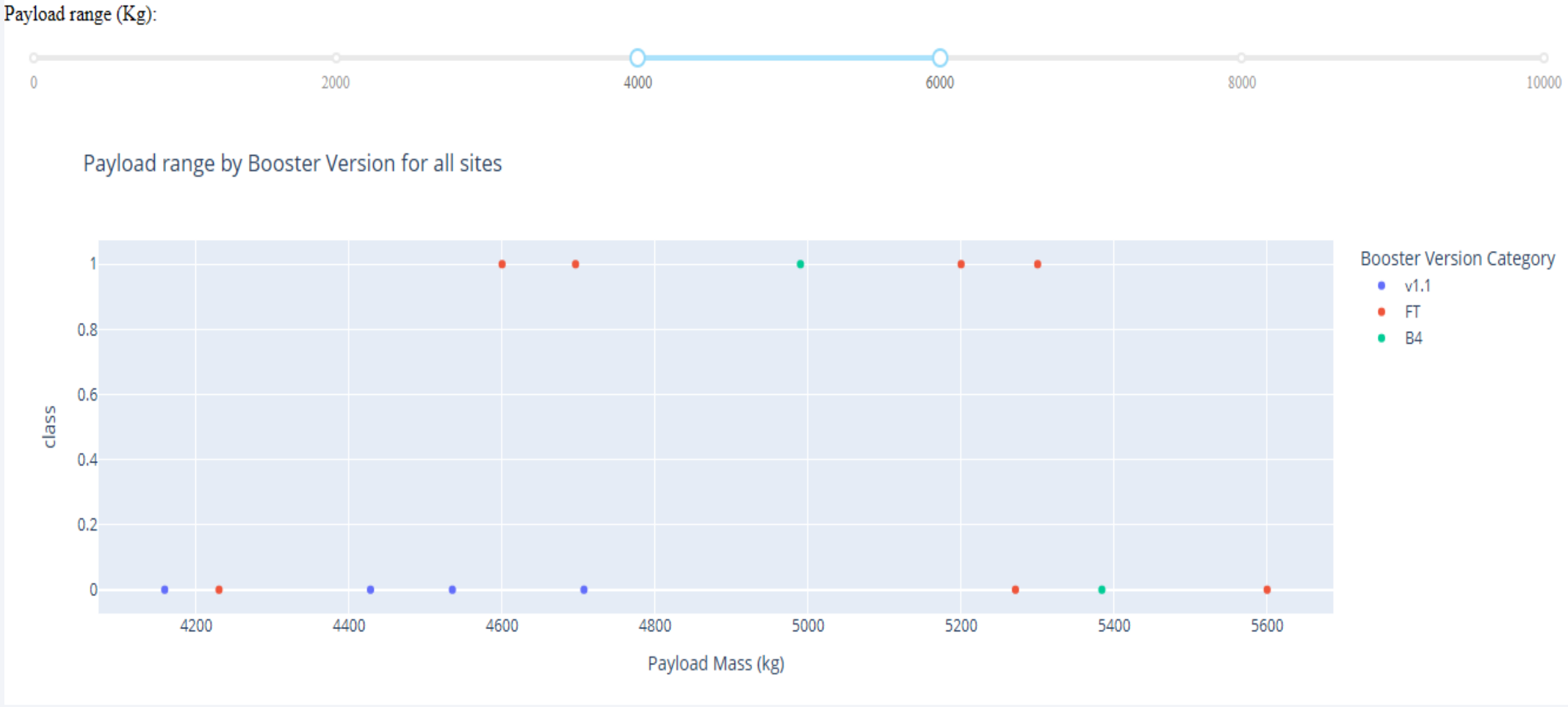


Launch Site with the highest launch success rate KSC LC-39A (41.7%)

- The success rate is 79%
- The failure rate is 21%, more than 3 times less than its success rate

Payload vs. Launch Outcome Scatter Plot of All Sites

Payload vs. Launch Outcome Scatter Plot of All Sites



Payload vs. Launch Outcome Scatter Plot of All Sites

Results of Visuals Analysis

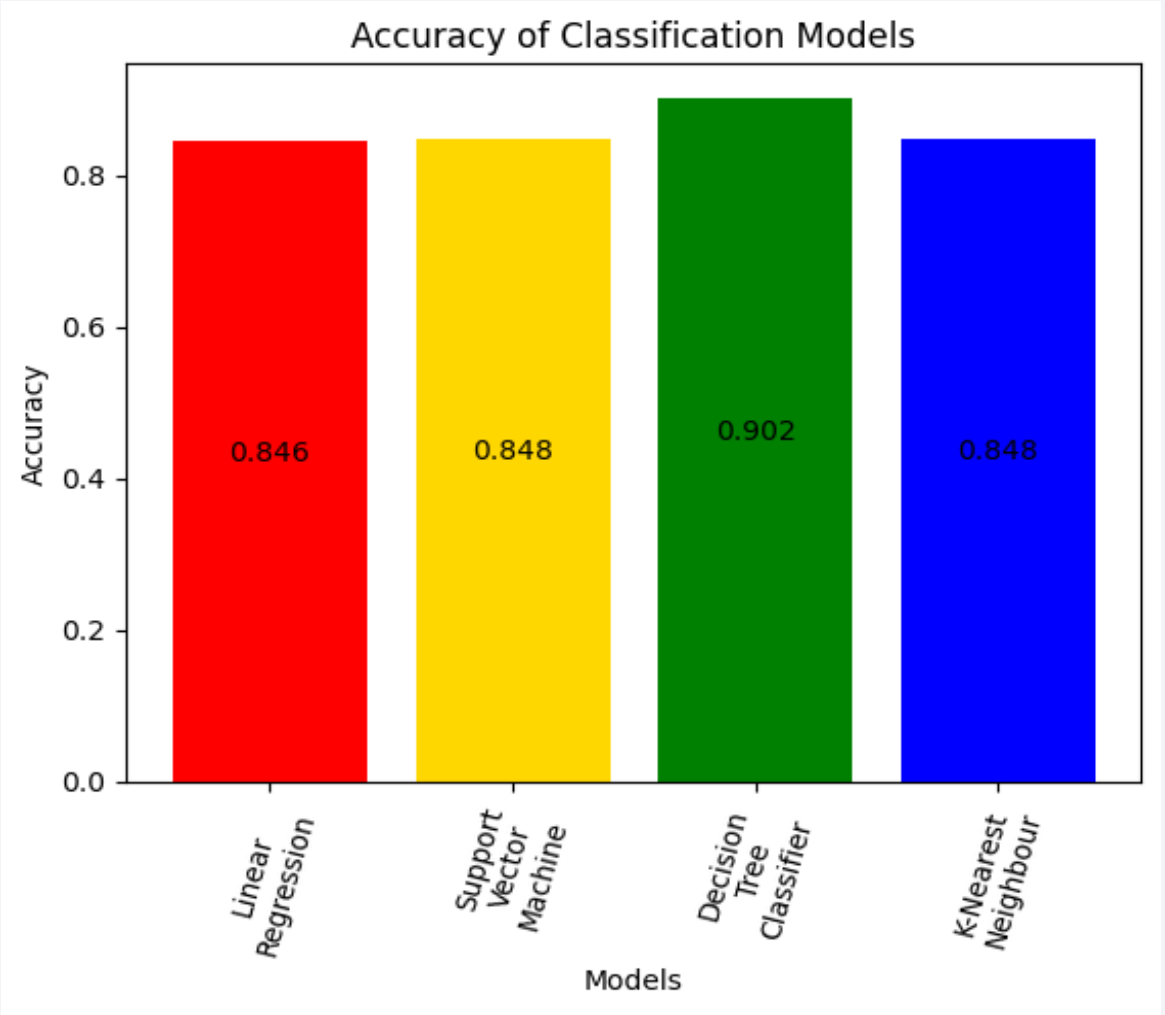
- The Payload Range with the highest success rate is 3000kg to 4000kg

Section 5

Predictive Analysis (Classification)

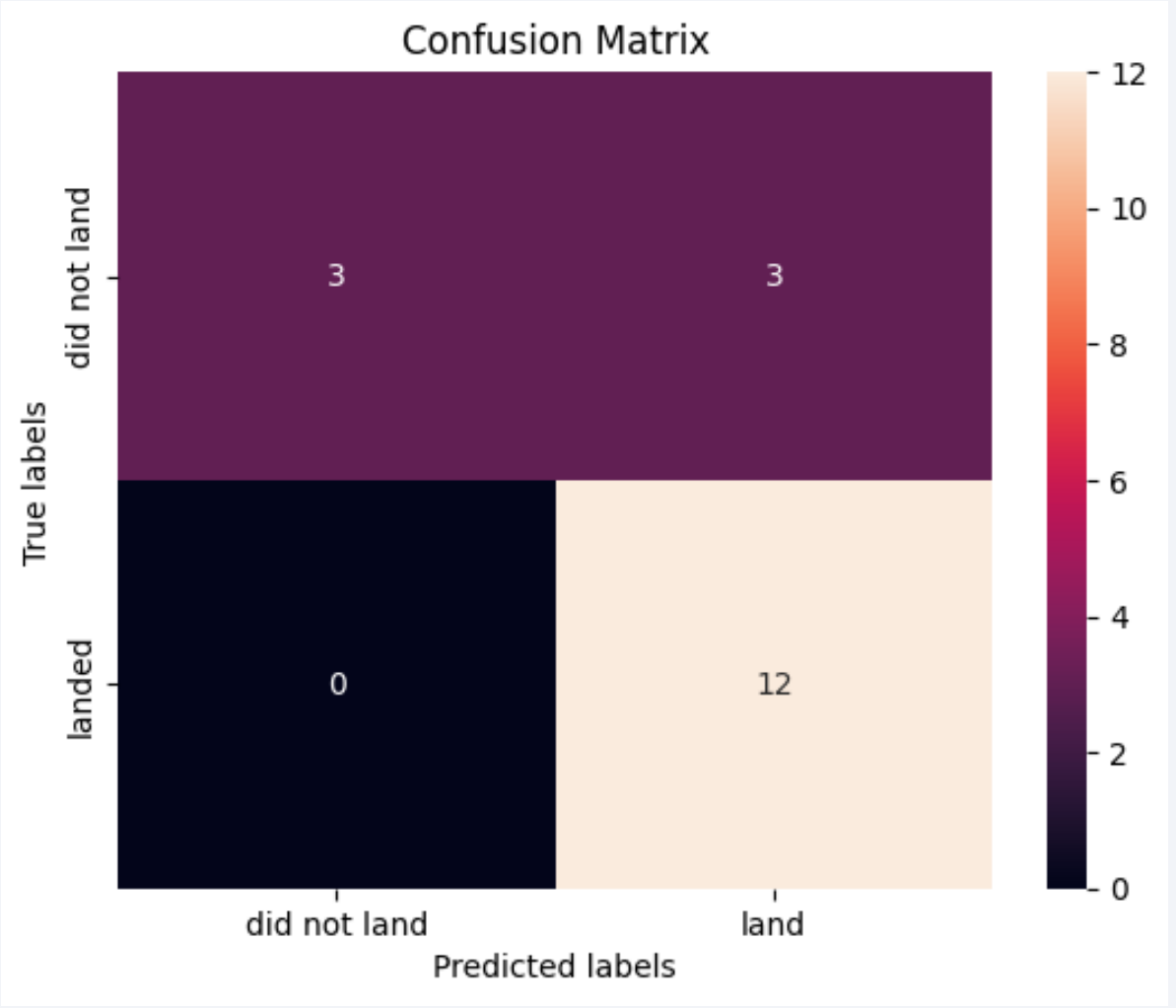
Classification Accuracy

- The Decision Tree Classifier model has the highest classification accuracy



Confusion Matrix

- The false positives are the main issue here
- This is because the predictions say it landed but it actually did not
- The false negative is zero, so we have no problems there



Conclusions

- The success rate of launches increases from 2013 onwards
- The most successful launch orbit is "SSO"
- "KSC LC-39A" is the launch site has the highest success rate
- It also has the most successful launches
- The payload range with the highest accuracy is 3000kg to 4000kg

Thank you!