IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Siddhartha Parmar
04 February 2024

Outline

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

Executive Summary

Methodologies:

- For data collection, get request was performed on SpaceX API to collect entire dataset.
- Further, all available methodologies were implemented like data scraping, data wrangling and cleaning to obtain relevant fields for performing EDA
- Exploratory Data Analysis using SQL and Visualisations was carried out

Summary of all results

- Results depict that launches made from Kennedy Space Center (KSC LC Launch Site) and particularly for FT booster version have a very high success rate followed by Cape Canaveral (CCAFS)

Introduction

- Project pertains to performing Data Science operations and analysis on launch and landing data of SpaceX Falcon9 rockets.

- Main objective is to predict the launch success rates across their 4 launch sites, depending on various parameters like payload mass, orbit type, flight number etc.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

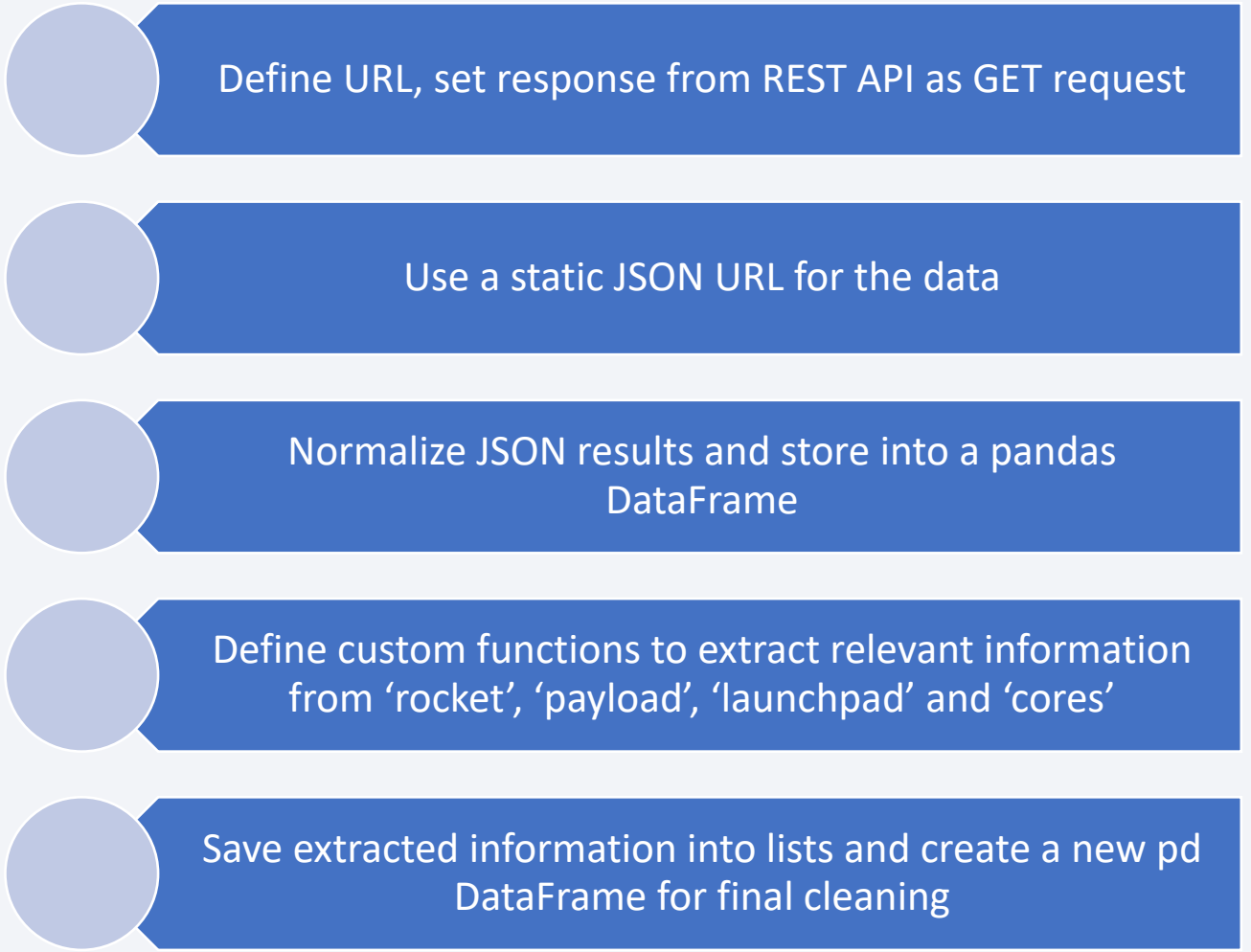
Data Collection

- Data sets were collected using GET request on SpaceX REST API
- Relevant information from various columns i.e. rocket, launchpad, payload, core was obtained for further cleaning and wrangling

- Process flow:

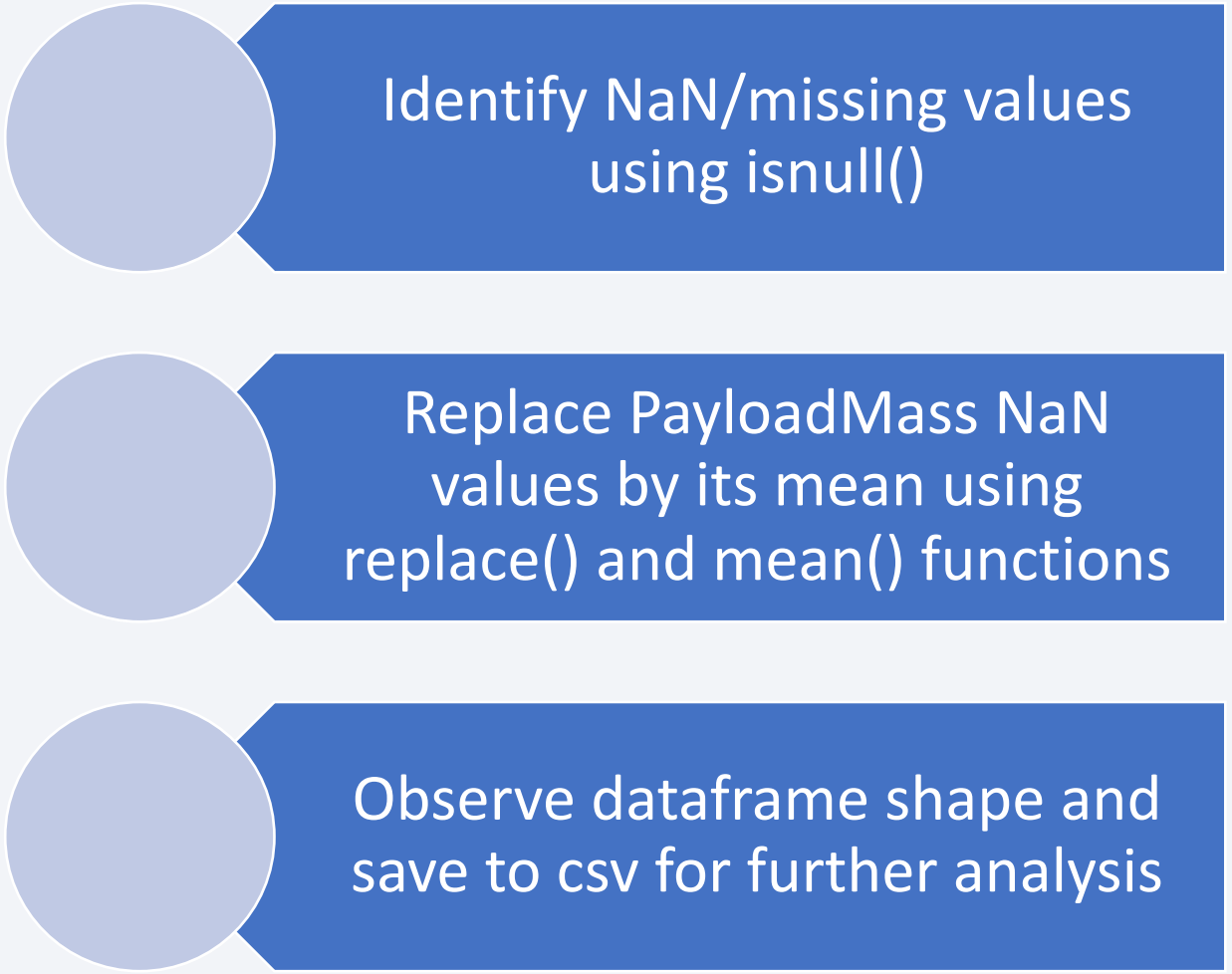
Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/rednax403791/CourseraAssessmentFinalCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



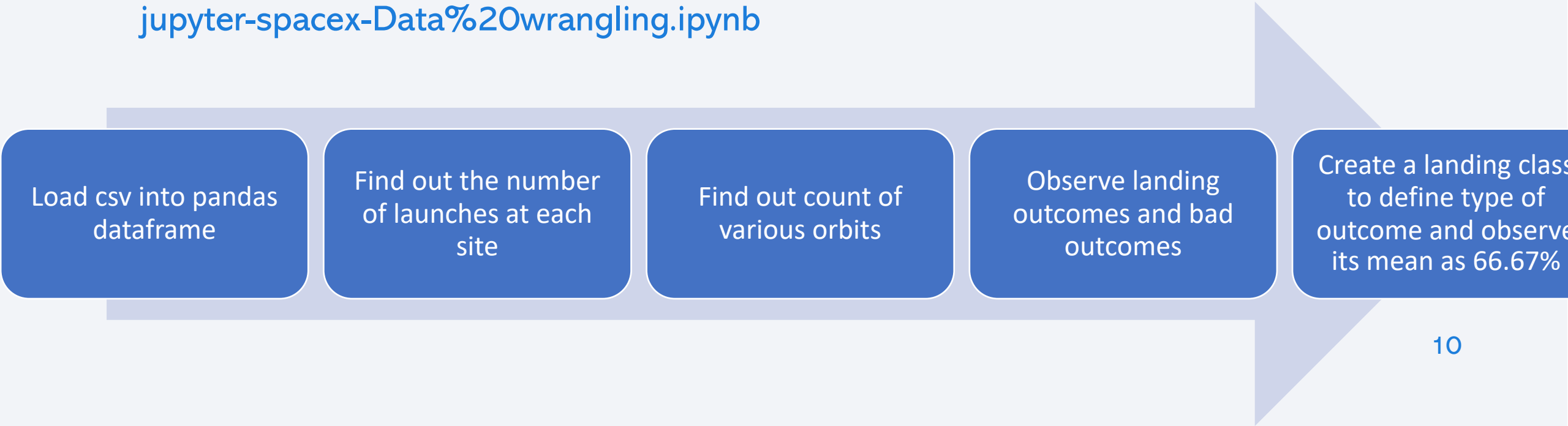
Data Collection - Scraping

- Missing values in the dataframe were identified using isnull() function
- NaN values in PayloadMass were replaced with the column mean
- LandingPad NaN values were maintained for data accuracy
- DataFrame shape was observed to be (90, 17)
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/rednax403791/CourseraAssessmentFinalCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Wrangling

- Primary objective is to ascertain training labels for modelling later
- Identify booster landings into 1 if successful and 0 if unsuccessful and assign it to a variable
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/rednax403791/CourseraAssessmentFinalCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Categorical, Scatter, Bar and Line plots are used to identify relationships between key features from the launch data as per below details:

- Categorical Plot: FlightNumber vs PayloadMass
- Categorical Plot: FlightNumber vs LaunchSite
- Scatter Plot: PayloadMass vs LaunchSite
- Bar Plot: Class vs Orbit (for orbit-wise success rate)
- Scatter Plot: FlightNumber vs Orbit
- Scatter Plot: PayloadMass vs Orbit
- Line Plot: Date vs Class (for date-wise success rate)

GitHub URL of completed EDA with data visualization notebook:
https://github.com/rednax403791/CourseraAssessmentFinalCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL (1 of 2)

After downloading datasets and connecting database, following SQL codes were implemented:

- %sql create table SPACEXTABLE as select * from SPACEXTBL where Date is not null
- %sql select Launch_Site from SPACEXTBL group by Launch_site
- %sql select * from SPACEXTBL where Launch_Site like "CCA%" limit 5
- %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like "NASA (CRS)”
- %sql SELECT avg(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version like "F9 v1.1”
- %sql select min(Date) from SPACEXTBL where Landing_Outcome like "Success (ground pad)”
- %sql select Booster_Version from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999

EDA with SQL (2 of 2)

- %sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME
- %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
- %sql SELECT substr(Date,6,2) as Month, Date, Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND substr(Date,0,5) = "2015"
- %sql SELECT LANDING_OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY COUNT_LAUNCHES DESC;

GitHub URL of completed EDA with SQL notebook:
https://github.com/rednax403791/CourseraAssessmentFinalCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Circles and Markers were created to locate Launch Sites
- Markers were colored differently based on success or failure of launch
- Mouse position object was defined to find coordinates of points on the map
- CalculateDistance function was defined to find distance between two points
- Closest highway, railway and city were observed using above function

- GitHub URL of completed interactive map with Folium map: https://github.com/rednax403791/CourseraAssessmentFinalCapstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- A dropdown menu is added showing all LaunchSites. Based on this selection different charts to be shown below it
- A Pie Chart is added to show success and failure for selected LaunchSite
- Scatterplot is added to show success and failure for various Payload Mass for each Booster Version
- A Slider is added to filter the Scatterplot based on different Payload Mass values

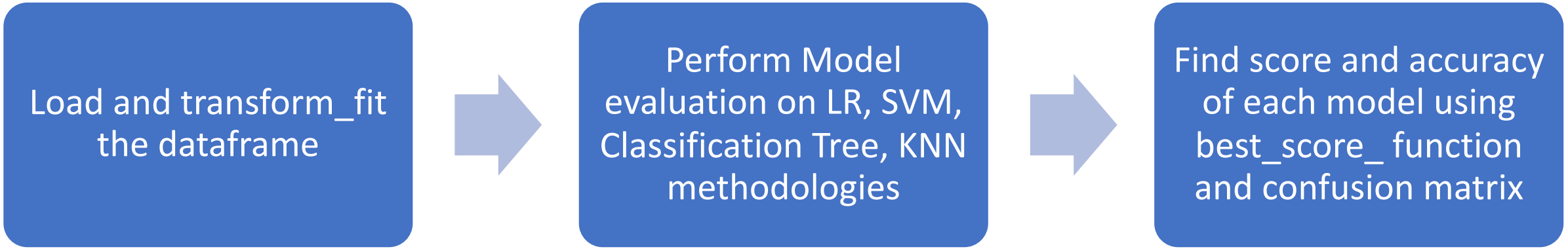
- GitHub URL of completed Plotly Dash lab: https://github.com/rednax403791/CourseraAssessmentFinalCapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification) – 1 of 2

- Load the data into DataFrame and transform-fit it
- Model evaluation:
    - Create Logistic Regression object using GridSearchCV for training datasets
    - Display best parameters using best_params_ and best score using best_score_ parameters
    - Calculate test dataset accuracy and observe confusion matrix
- Do the same for SVM, Classification Tree and KNN objects to find their accuracy and confusion matrix
- Understand the best method based on scores

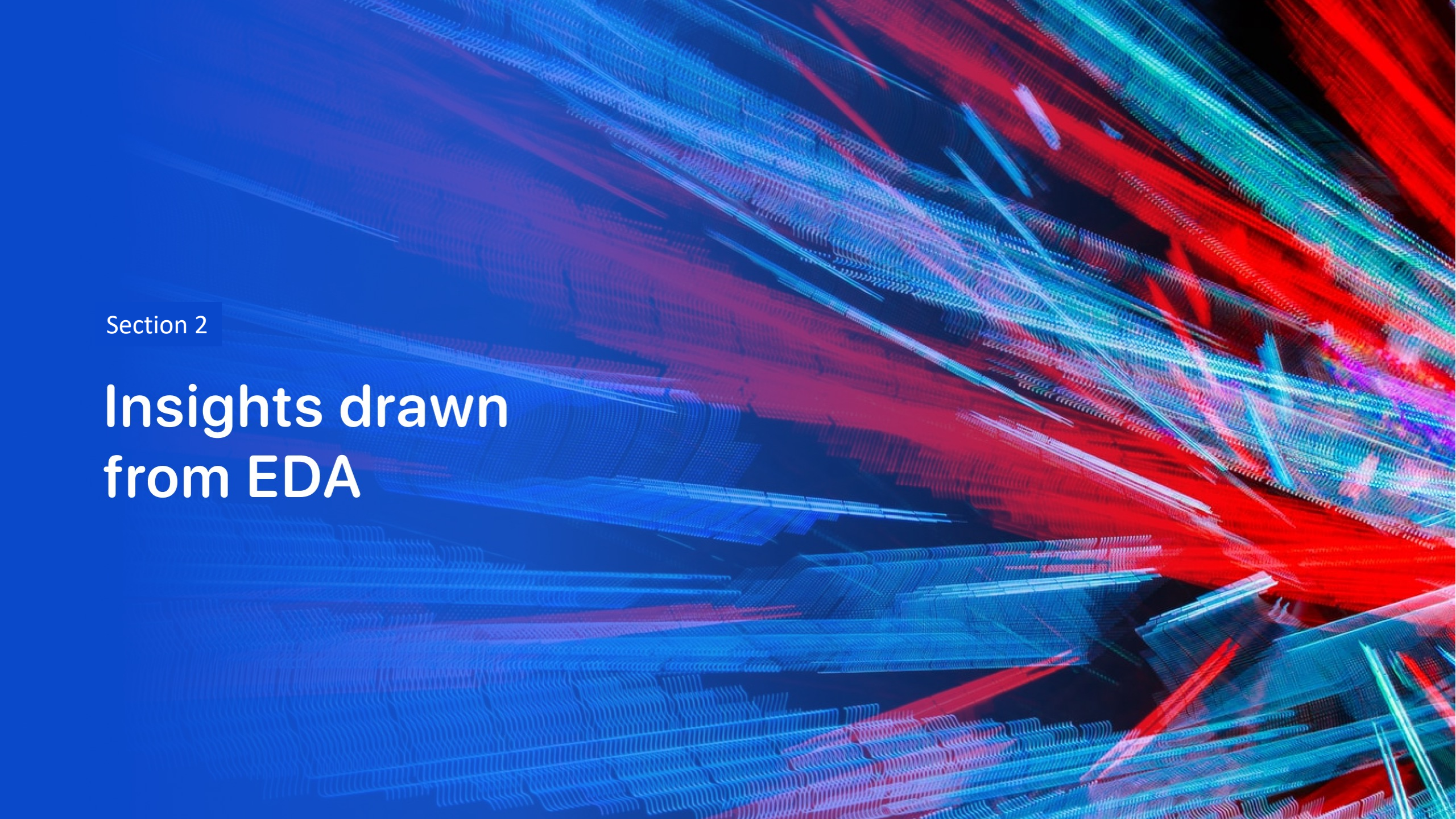
Predictive Analysis (Classification) – 2 of 2

- GitHub URL of completed predictive analysis lab: https://github.com/rednax403791/CourseraAssessmentFinalCapstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



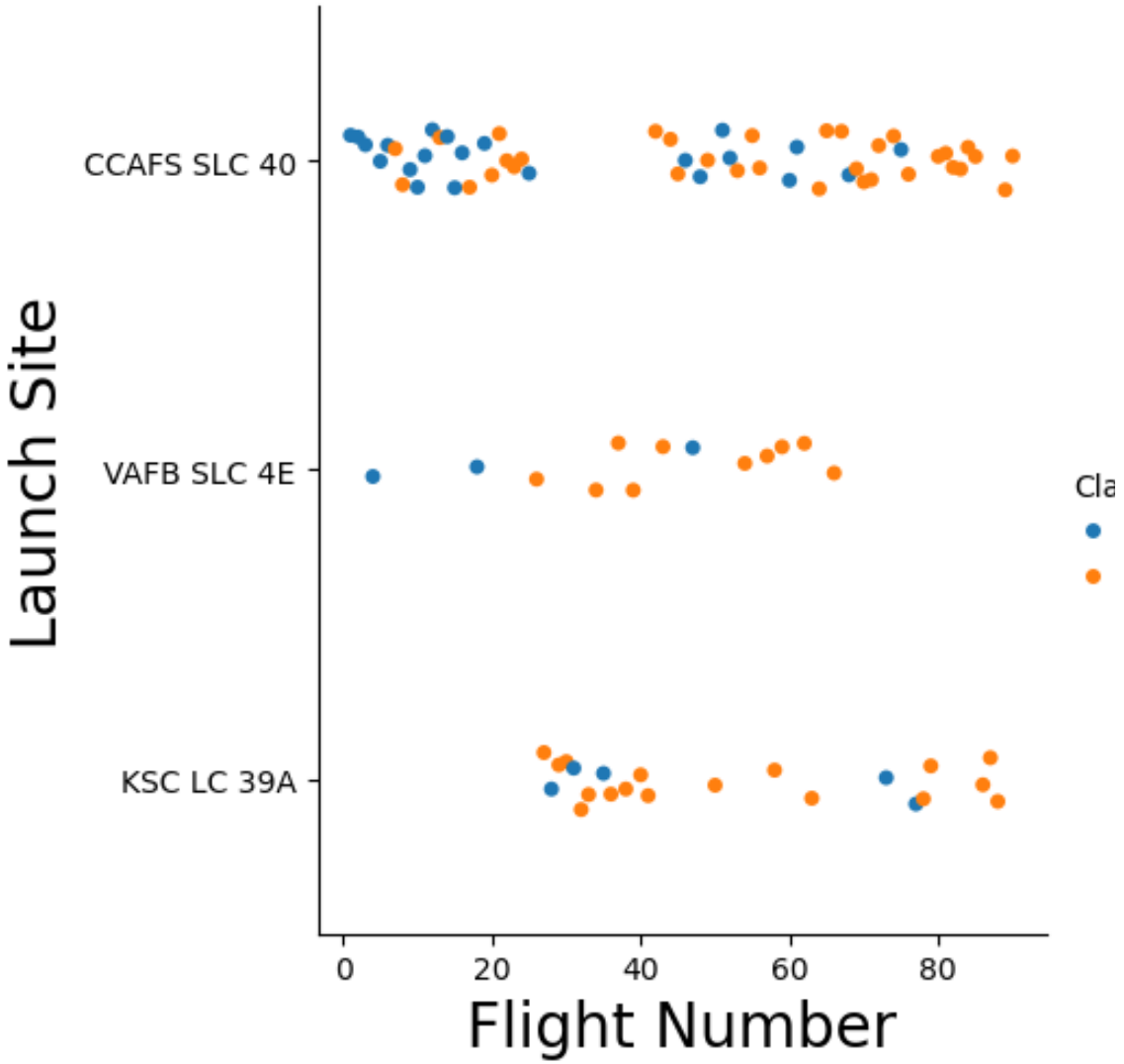
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

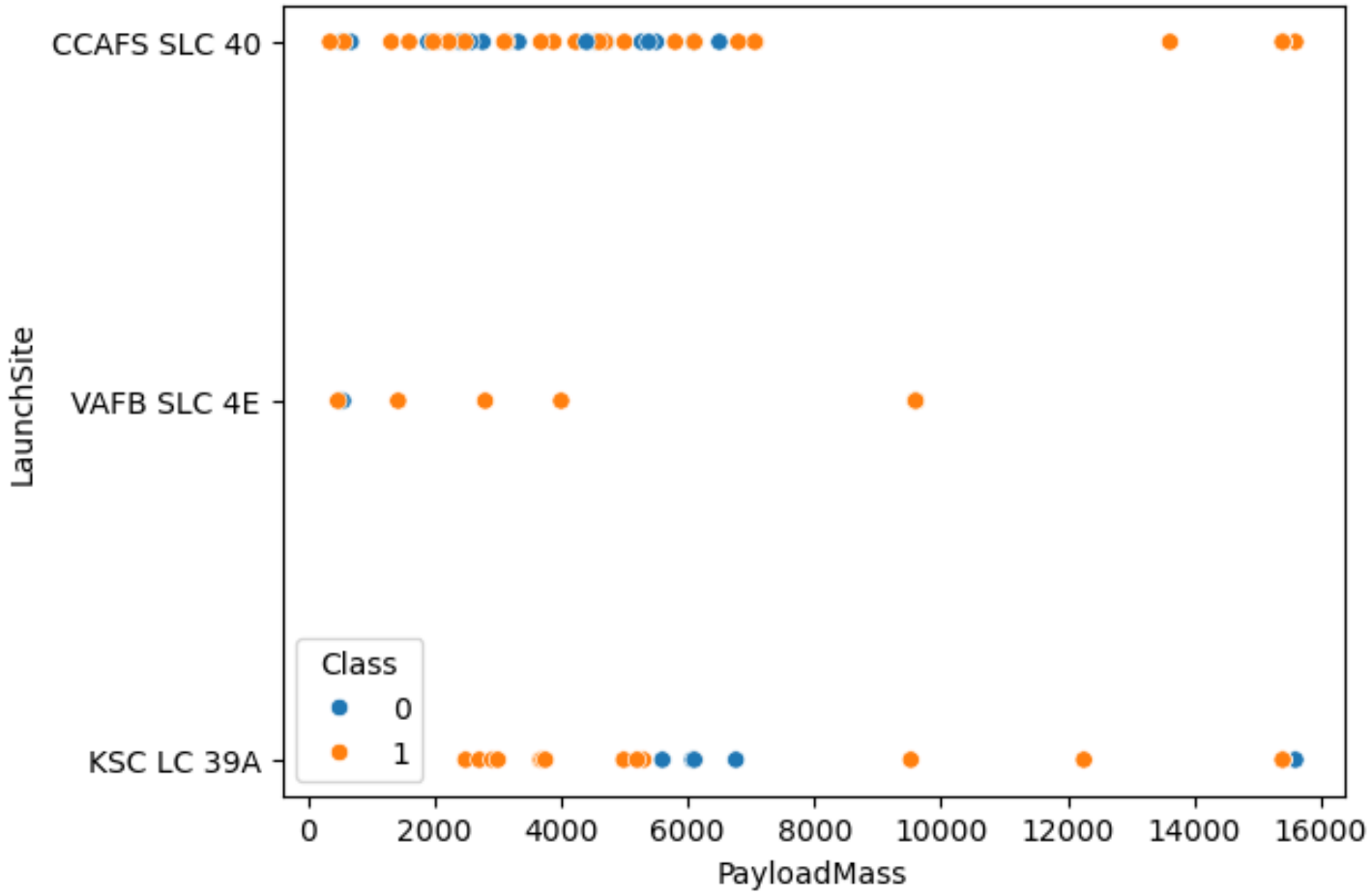
- Scatter plot of Flight Number vs. Launch Site
- Observations:
  - CCAFS SLC 40 has a high number of launches
  - VAFB and KSC have better success rates than CCAFS
  - With increasing FlightNumber, the success rate is improving for each Launch Site



Payload vs. Launch Site

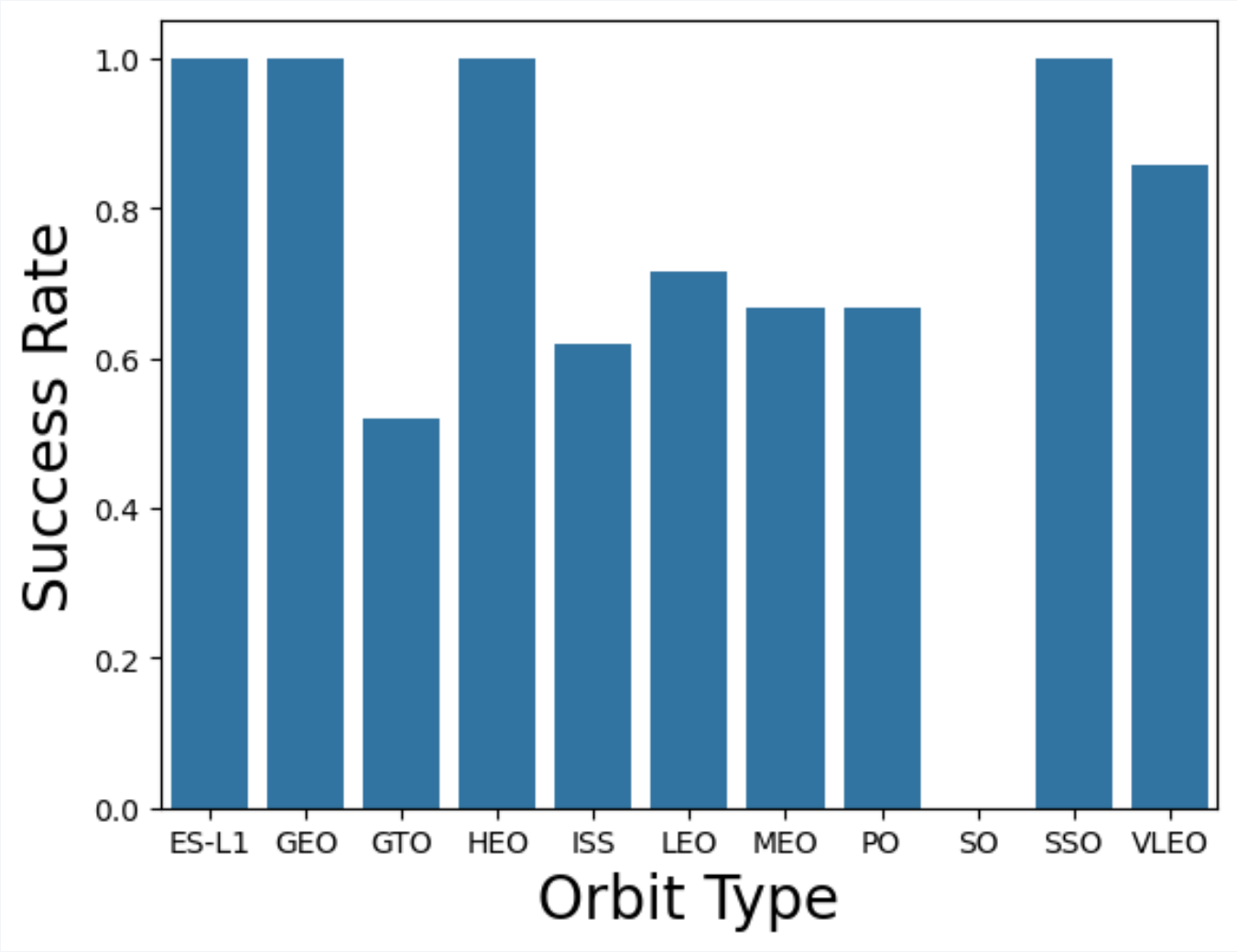
Observations:

- At VAFB launch site, no launches of payload mass beyond 10000 take place
- Only one heavy launch has been unsuccessful which was from KSC LC site



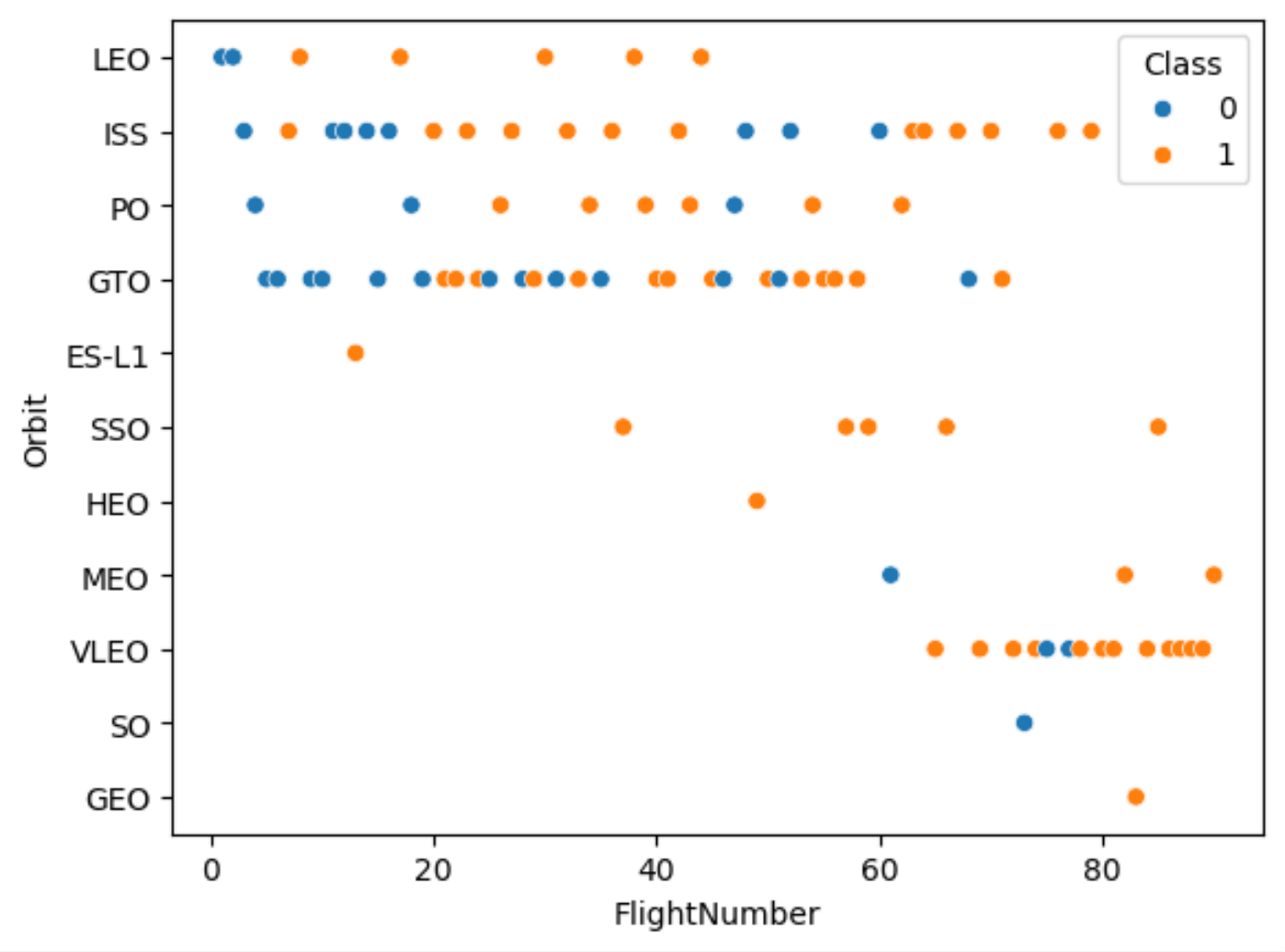
Success Rate vs. Orbit Type

- Observations:
  - ES-L1, GEO, HEO and SSO orbits have highest success rates
  - These are followed by VLEO, LEO and others at reducing success rates



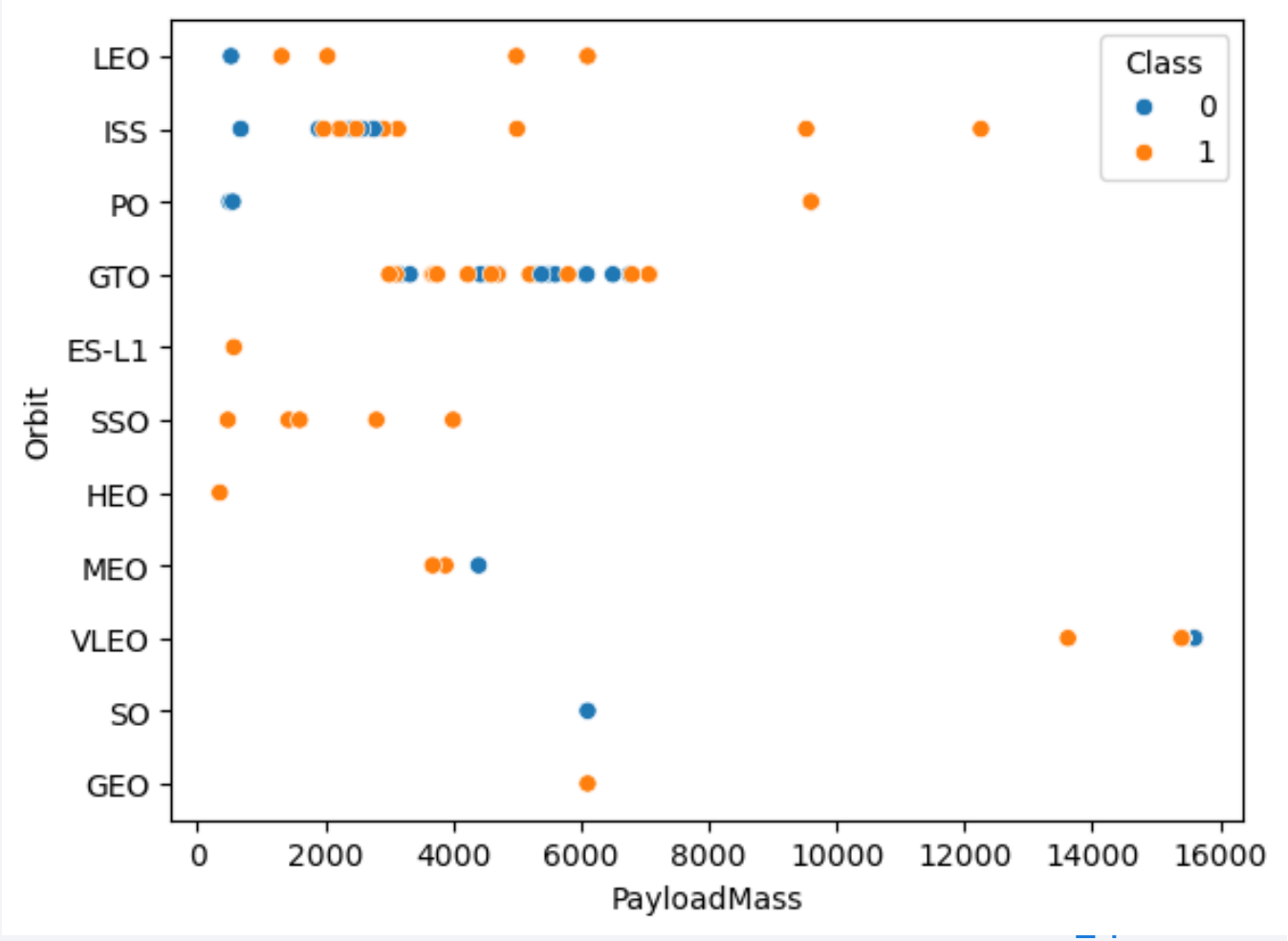
Flight Number vs. Orbit Type

- Observations:
  - LEO orbit success rates increase with the flight number
  - GTO orbit launches have mixed responses on success rates when seen with flight numbers



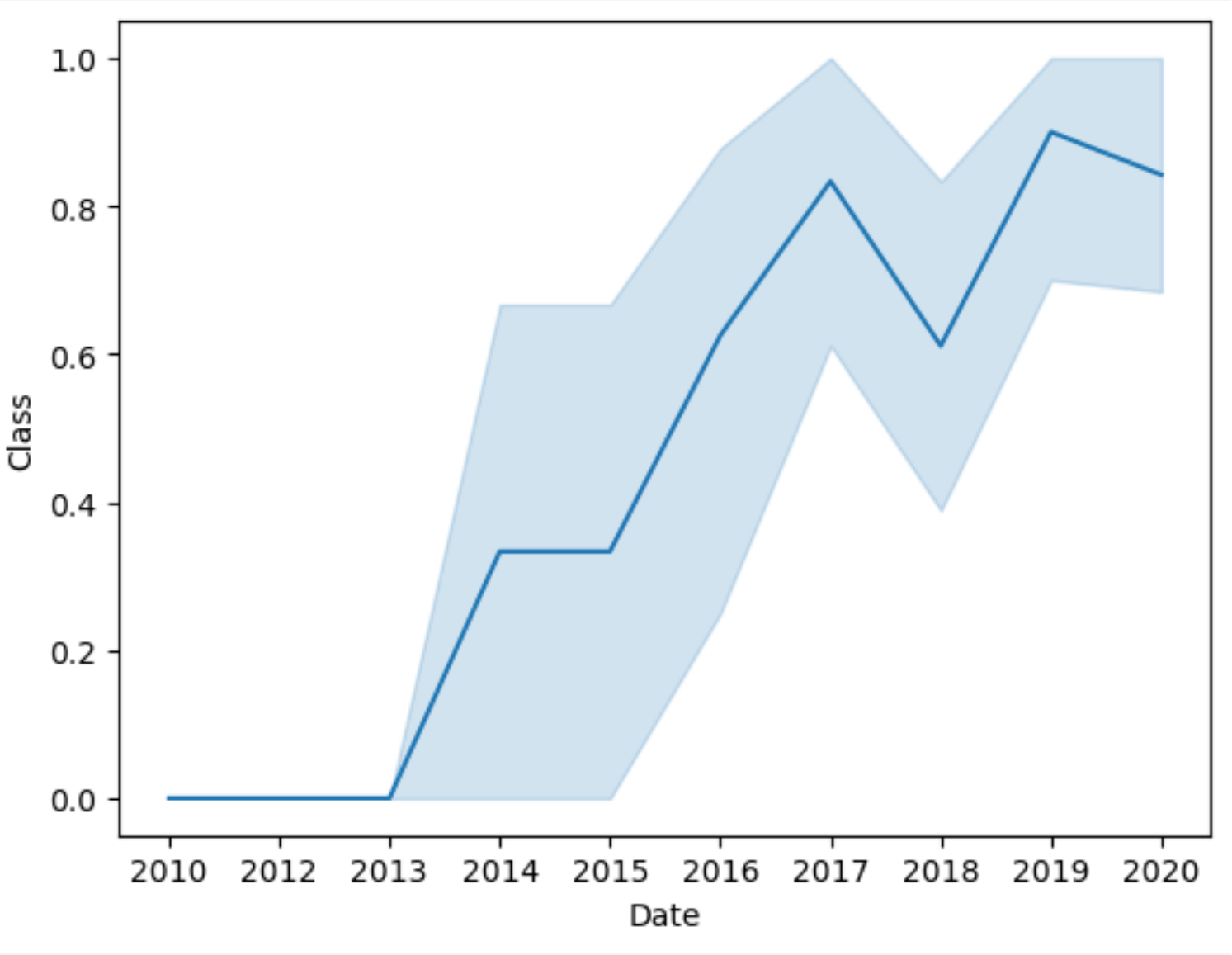
Payload vs. Orbit Type

- Observation:
  - LEO, ISS and PO have better success rates when payload mass increases
  - GTO payload masses range between 2000 and 8000 only



Launch Success Yearly Trend

- Observations:
  - Launch success has increased since the year 2013
  - There was a considerable decline in success in 2018 but it increased in 2019
  - Launch success peaked in 2019



All Launch Site Names

- Find the names of the unique launch sites

%sql select Launch_Site from SPACEXTBL group by Launch_site

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

- Using SQL we have queried to show Launch_Site from the table, grouping the results by Launch_Site, hence showing only unique values

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

%sql select * from SPACEXTBL where Launch_Site like "CCA%" limit 5

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

- We have queried to display all fields of first 5 records where the Launch_Site feature starts with CCA, essentially to look like CCA….

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like "NASA (CRS)"

sum(PAYLOAD_MASS__KG_)
45596

- We have queried to show the sum of payload mass for records where the Customer column is “NASA (CRS)”

Average Payload Mass by F9 v1.1

- Calculate the total payload carried by boosters from NASA

%sql SELECT avg(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version like "F9 v1.1"

avg(PAYLOAD_MASS__KG_)
2928.4

- We have queried to show the average Payload mass of those records whose booster version is F9 v1.1

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

%sql select min(Date) from SPACEXTBL where Landing_Outcome like "Success (ground pad)"

min(Date)
2015-12-22

- Result is shown using the min(Date) function for records where Landing_Outcome is “Success (groud pad)”

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

%sql select Booster_Version from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999

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

- Using between keyword we have identified and displayed records between particular payload numbers, and also with Landing_Outcome being “Success (drone ship)”

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

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

- We have grouped by Mission Outcome to display only two columns from the table to show count of Mission Outcomes

Boosters Carried Maximum Payload

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

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE
PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM
SPACEXTBL)
```

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

- Using subquery we have displayed booster versions that have carried maximum payload mass

2015 Launch Records

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

%sql SELECT substr(Date,6,2) as Month, Date, Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND substr(Date,0,5) = "2015"

Month	Date	Landing_Outcome	Booster_Version	Launch_Site
01	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Launch records for failed drone ship landings during 2015 are queried here

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

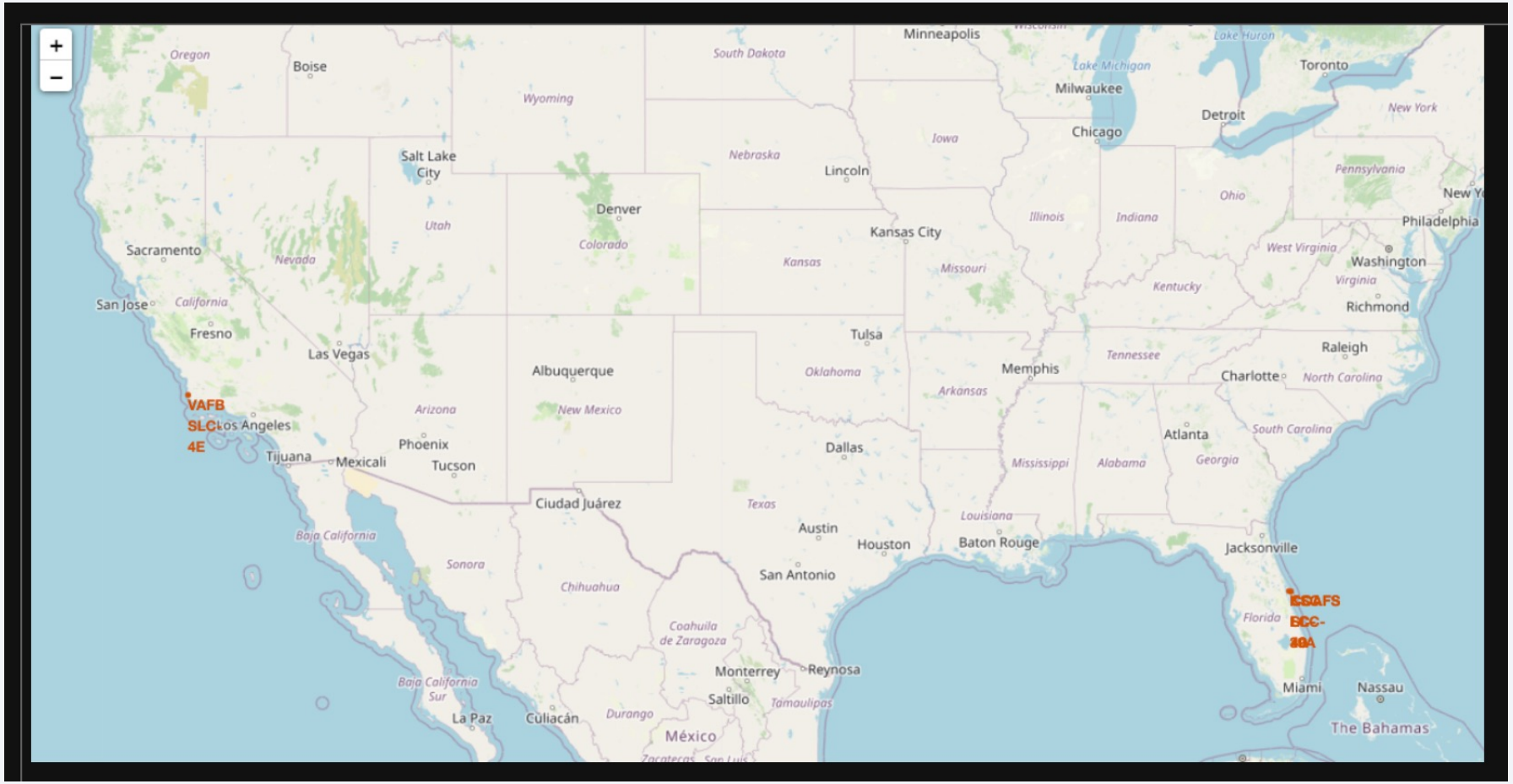
```
%sql SELECT LANDING_OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY COUNT_LAUNCHES DESC;
```

Landing_Outcome	COUNT_LAUNCHES
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

- We have queried Landing_Outcome and the count of launches between two date values, ordering the results by count of launches in descending order

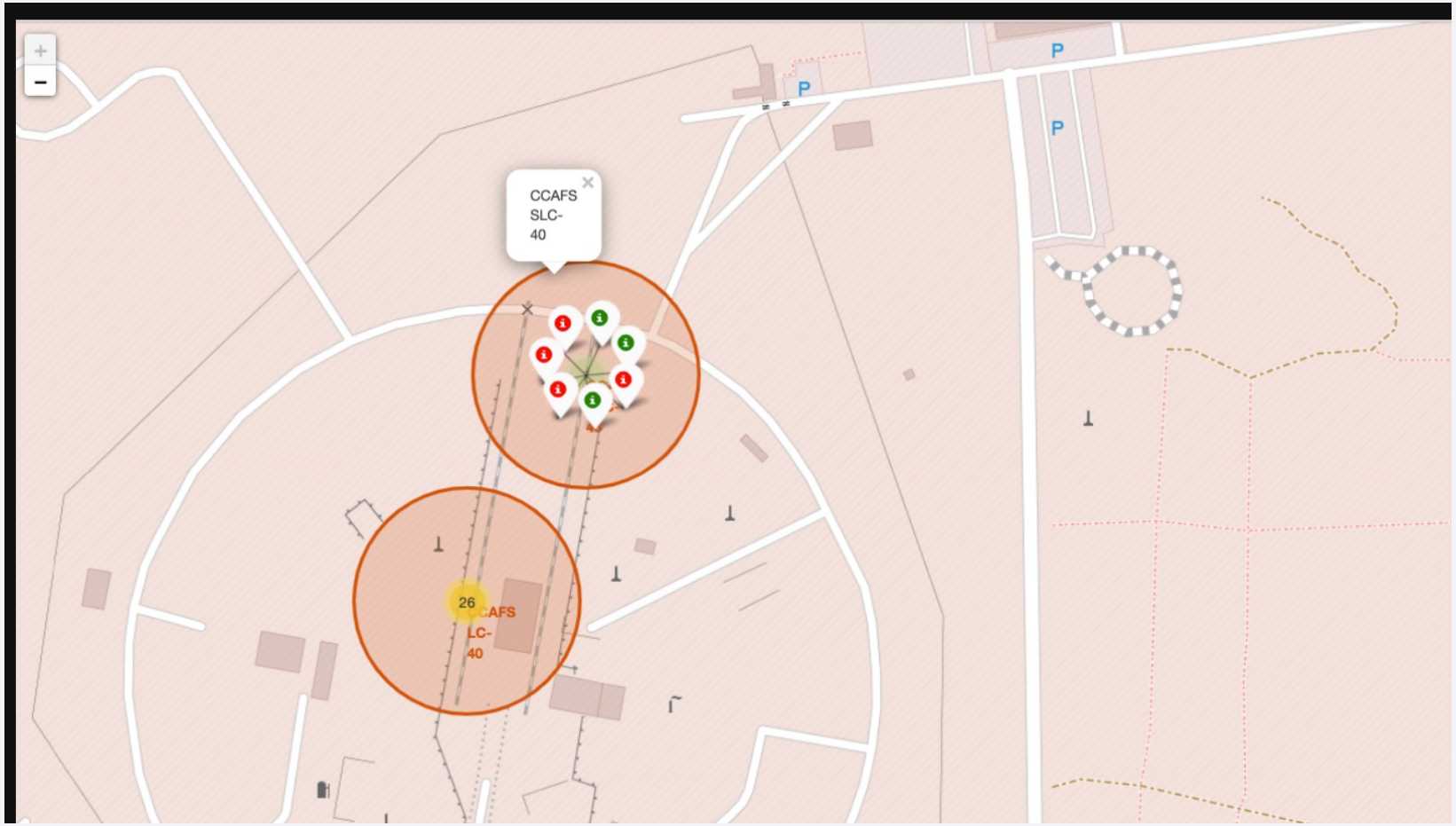

Section 3
Launch Sites
Proximities Analysis

All Launch Sites



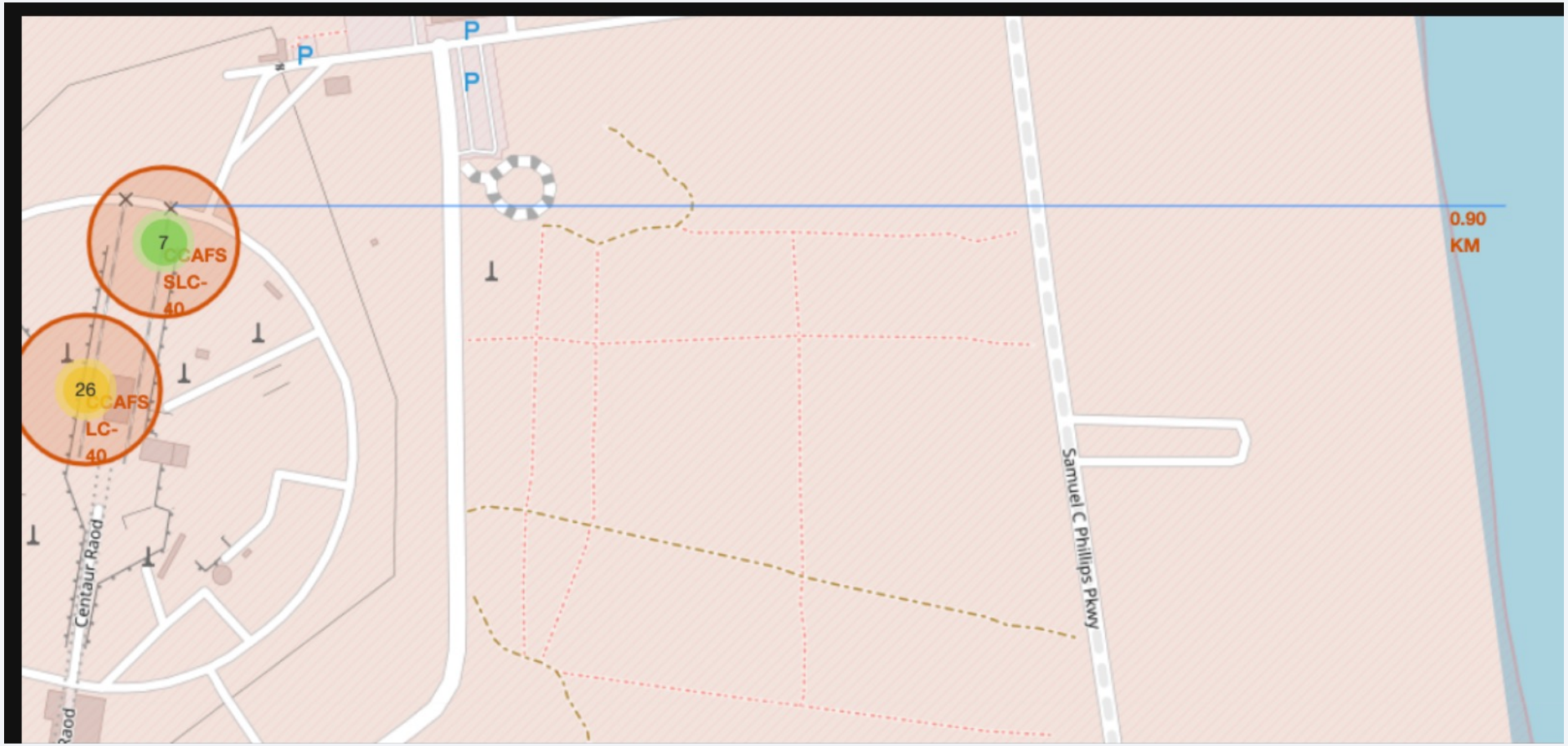
Map shows the 4 launch sites which have been marked using markers

Launch Outcomes

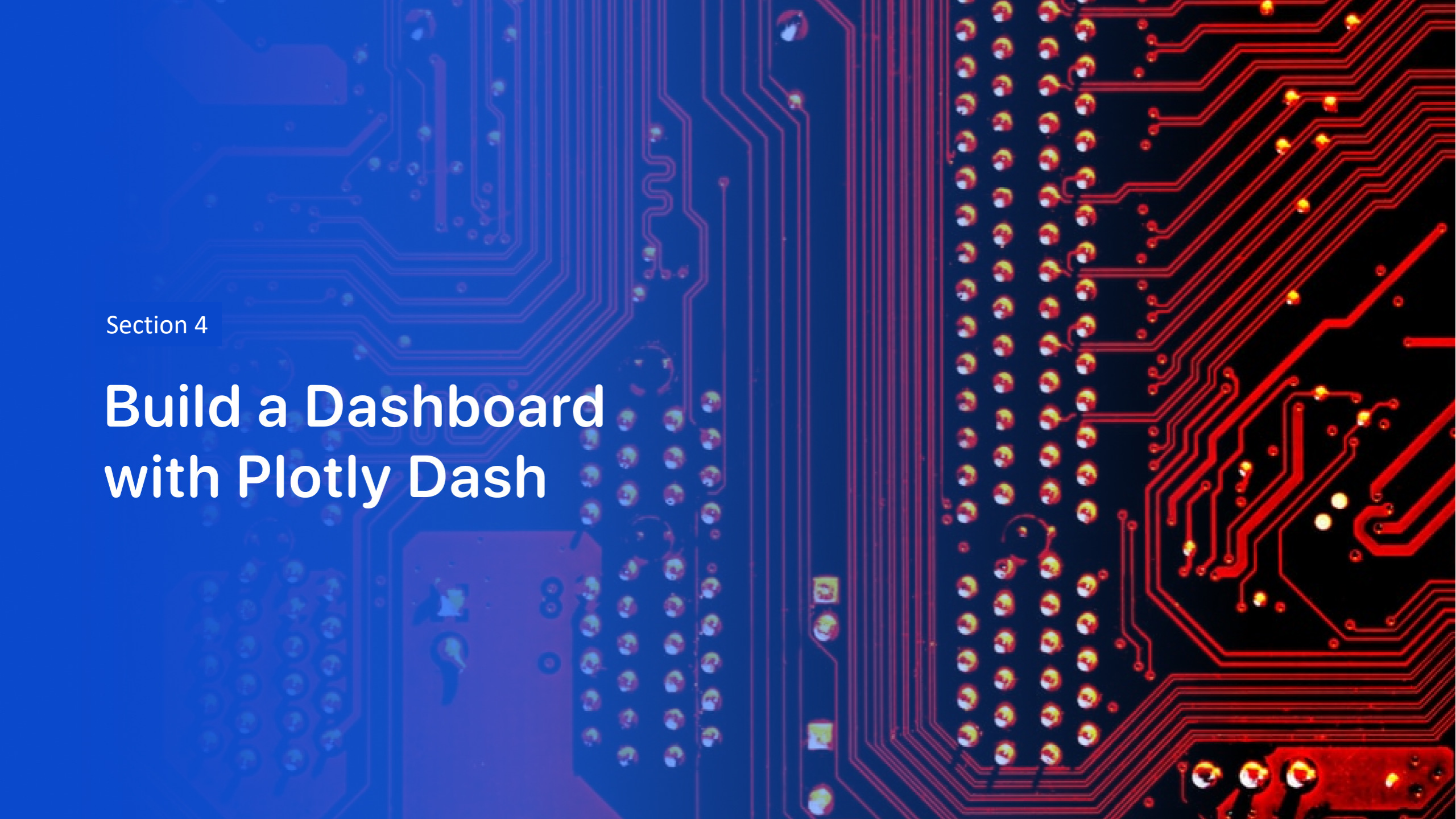


- Launch outcomes are marked in green (successful) or red (unsuccessful)

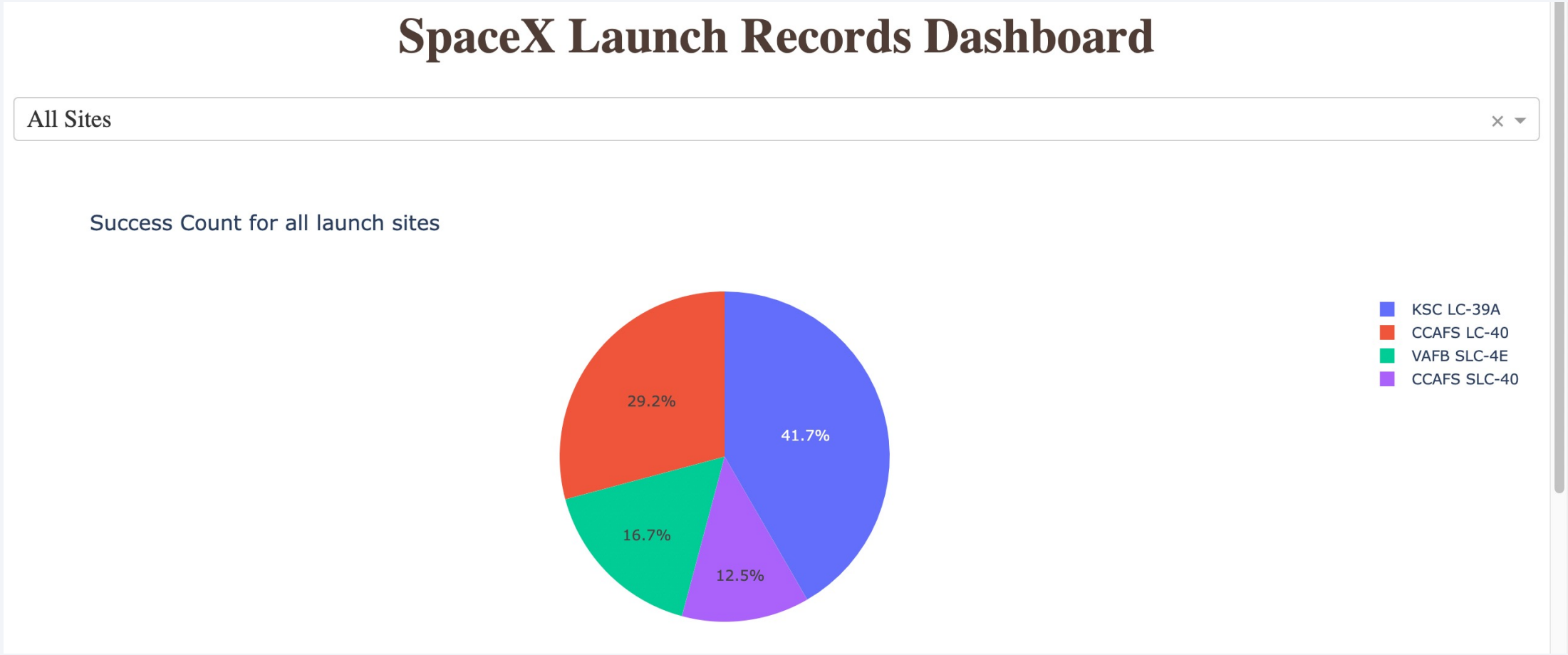
Launch Site proximity to Map elements



- Results show launch site proximity to nearest railway, highway and city

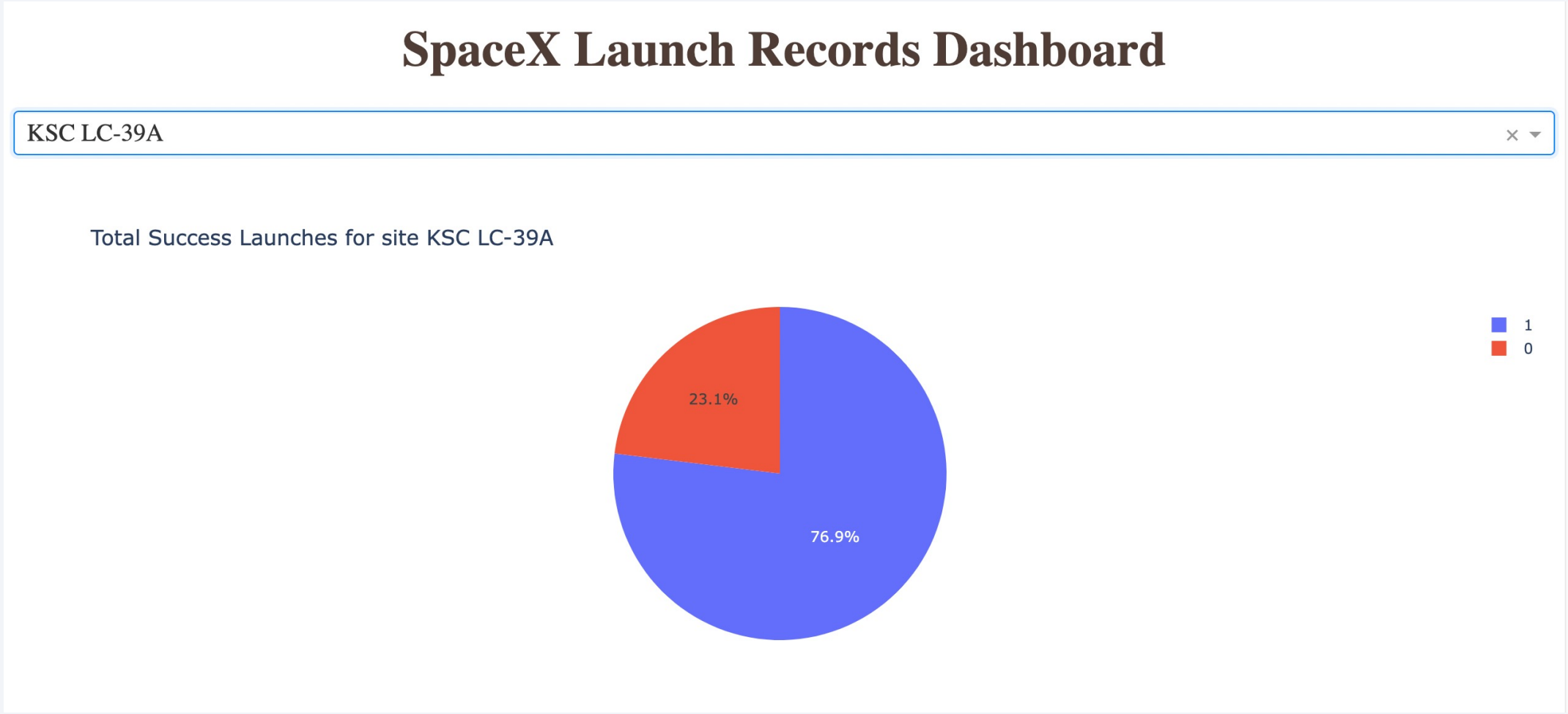

Section 4
Build a Dashboard with Plotly Dash

Launch Success for All Launch Sites



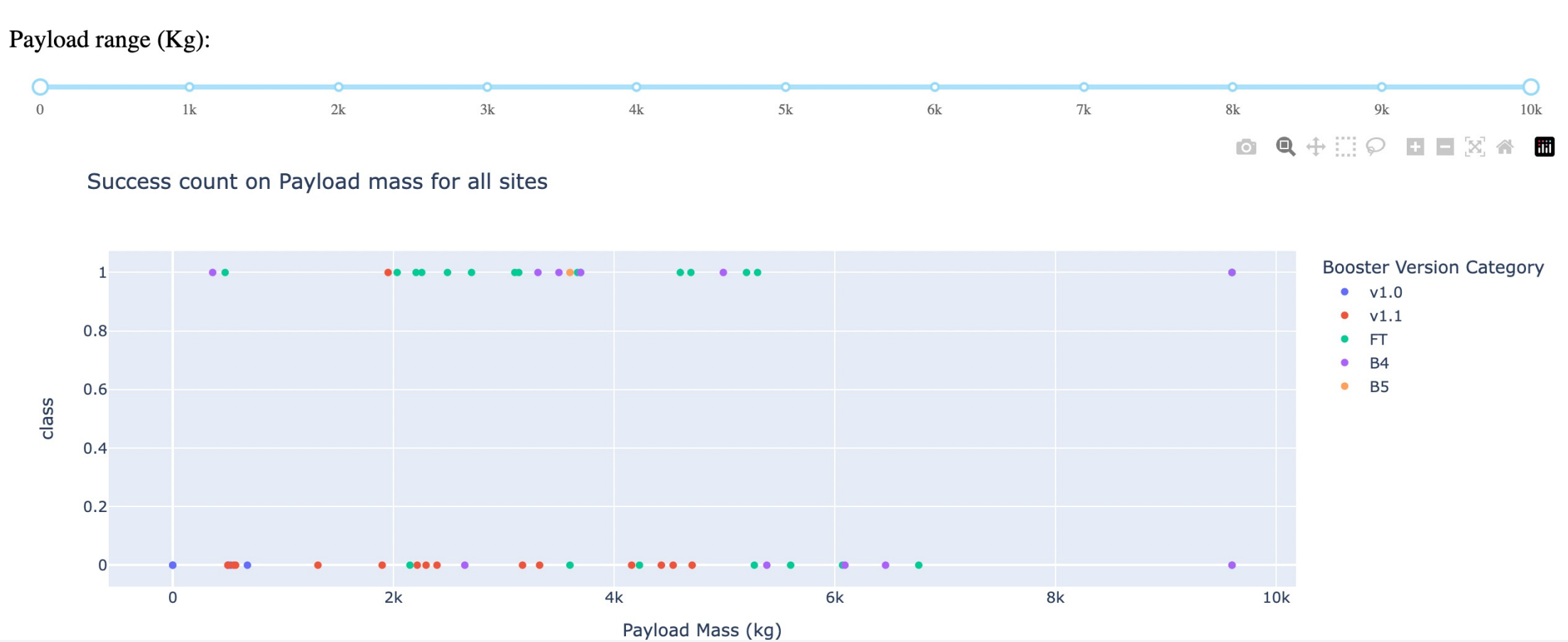
Pie chart shows launch success for all launch sites, with most successat KSC LC-39A followed by CCAFS LC-40

Launch Site with Highest Launch Success Ratio



Pie chart shows split between successful and unsuccessful launches at KSC LC 39A which has overall highest launch success

Payload vs. Launch Outcome Scatterplot for All Sites

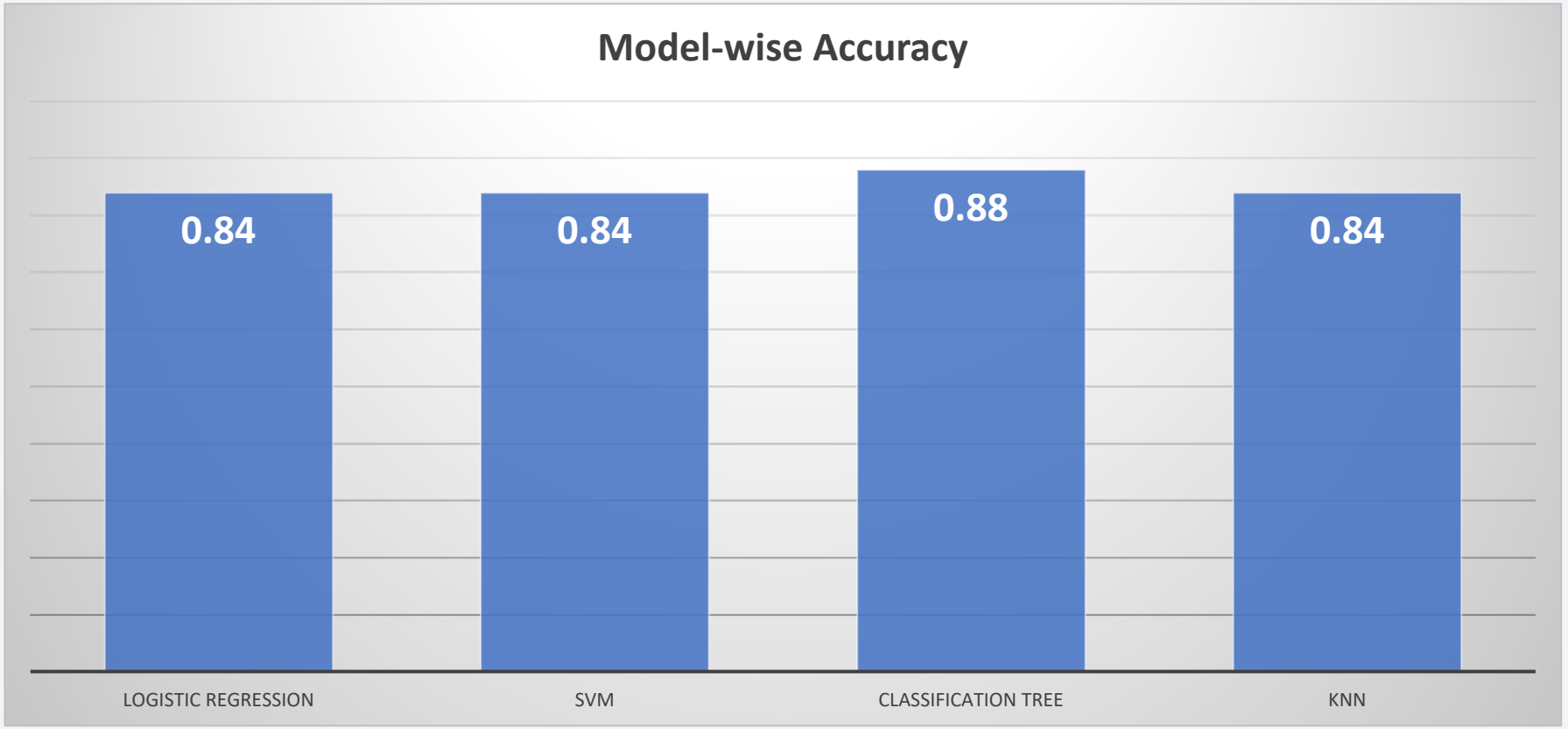


- Scatterplot shows that Booster version FT has the highest success rate
- Maximum success in all booster versions is obtained at payload mass range of 2000 to 4000

Section 5

Predictive Analysis (Classification)

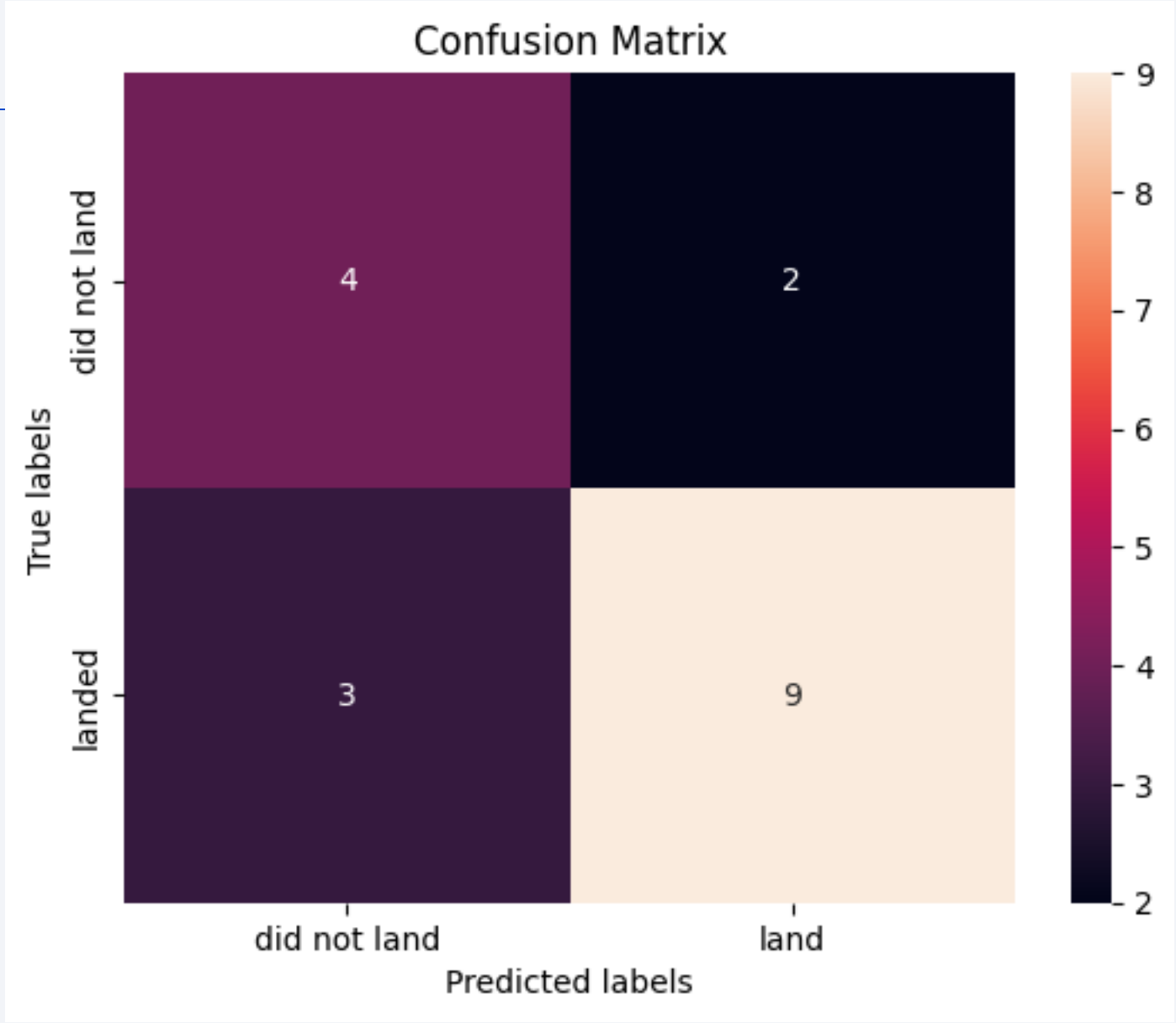
Classification Accuracy



- Practically all models have comparable accuracy, however Classification Tree (0.88) has slightly higher accuracy than others (0.84 each)

Confusion Matrix

- Confusion Matrix for Classification Tree model is shown here
- 4 True Positives and 9 True Negatives are indication of a good fit



Conclusions

- KSC LC Launch site has a high success rate at 41.7%
- Booster Version “FT” has a high success rate amongst all 5 categories
- Any of the 4 Classification models could be utilized, however preference could be given to Tree model

Appendix

- Github link to entire project:

https://github.com/rednax403791/CourseraAssessmentFinalCapstone/tree/main

Thank you!