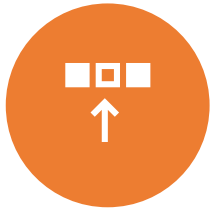
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eve Glenn Sandoval
10/07/2023

Outline



Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

Summary of methodologies

Data Collection through API

Data Collection with Web Scraping

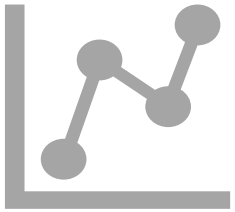
Data Wrangling

Exploratory Data Analysis with SQL

Exploratory Data Analysis with Visualization

Interactive Visual Analytics with Folium

Machine Learning Prediction



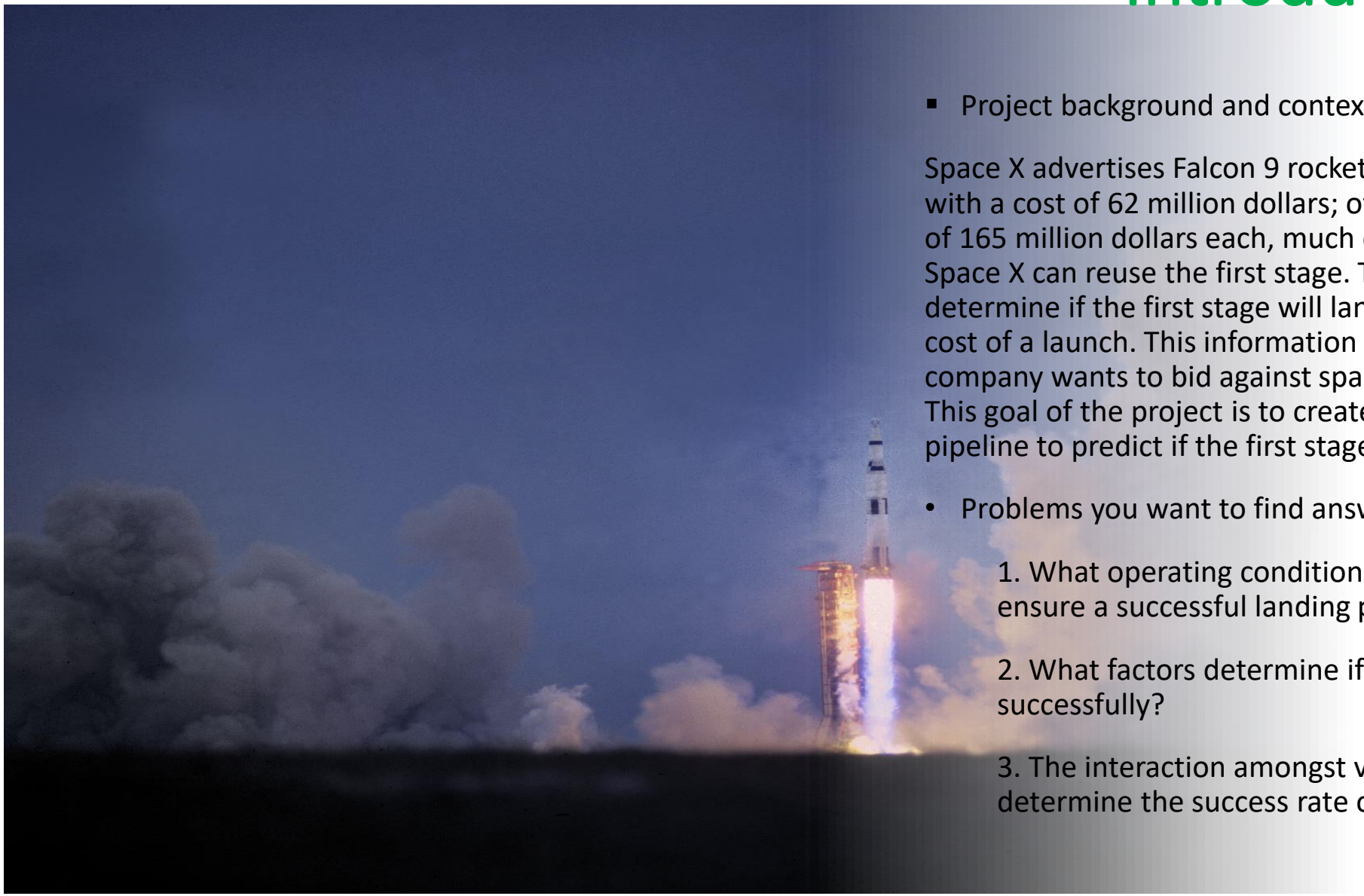
Summary of all results

Exploratory Data Analysis result

Interactive analytics in screenshots

Predictive Analytics result

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. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers

  1. What operating conditions needs to be in place to ensure a successful landing program.

  2. What factors determine if the rocket will land successfully?

  3. The interaction amongst various features that determine the success rate of a successful landing.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

Data was collected using SpaceX API and web scraping from Wikipedia.

Perform data wrangling

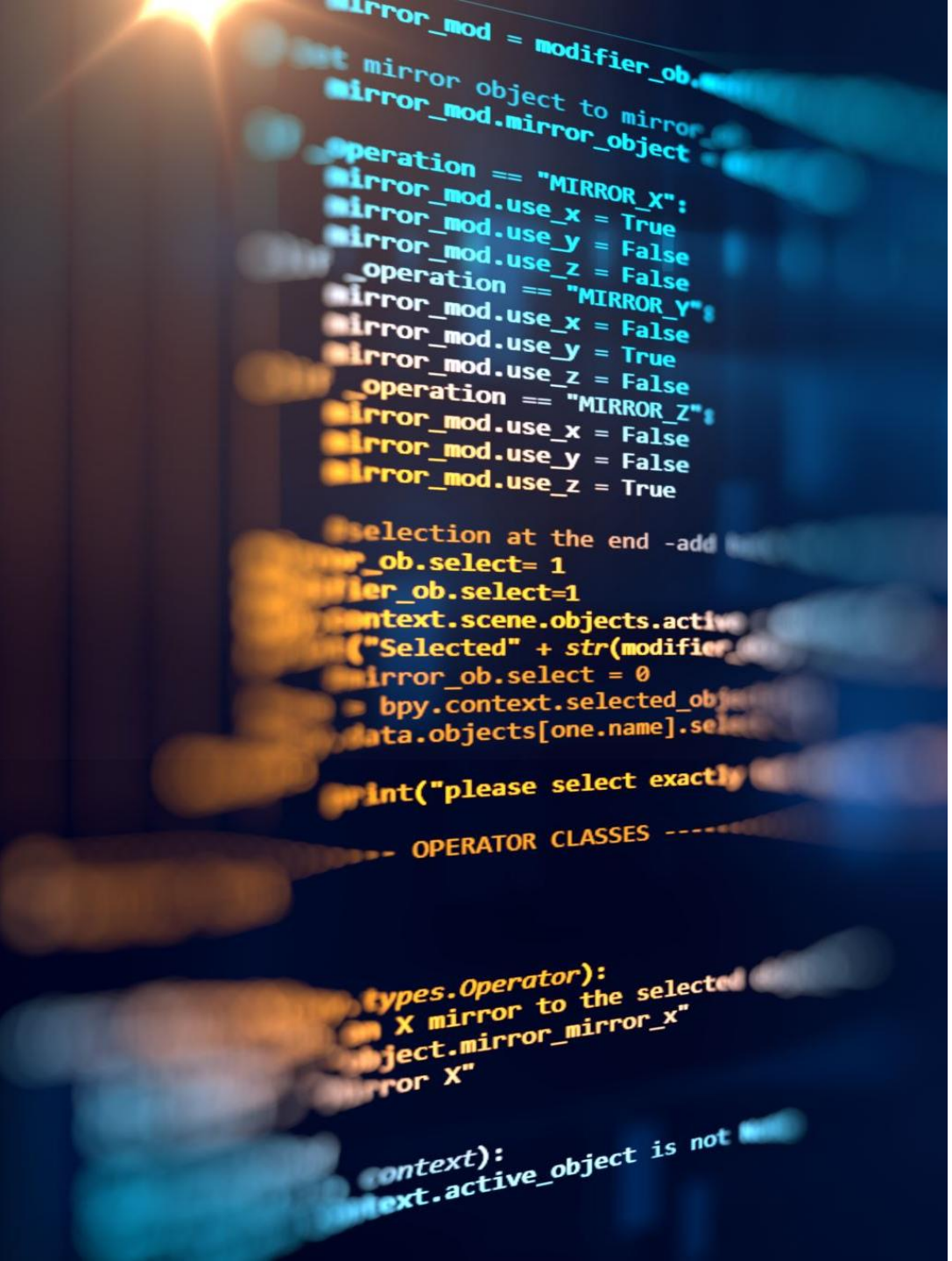
One-hot encoding was applied to categorical features

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 collection process involved a combination of API requests from SpaceX REST

API and Web Scraping data from a table in SpaceX's Wikipedia entry.

I had to use both of these data collection methods in order to get complete

information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite,

Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount,

Serial, Longitude, Latitude

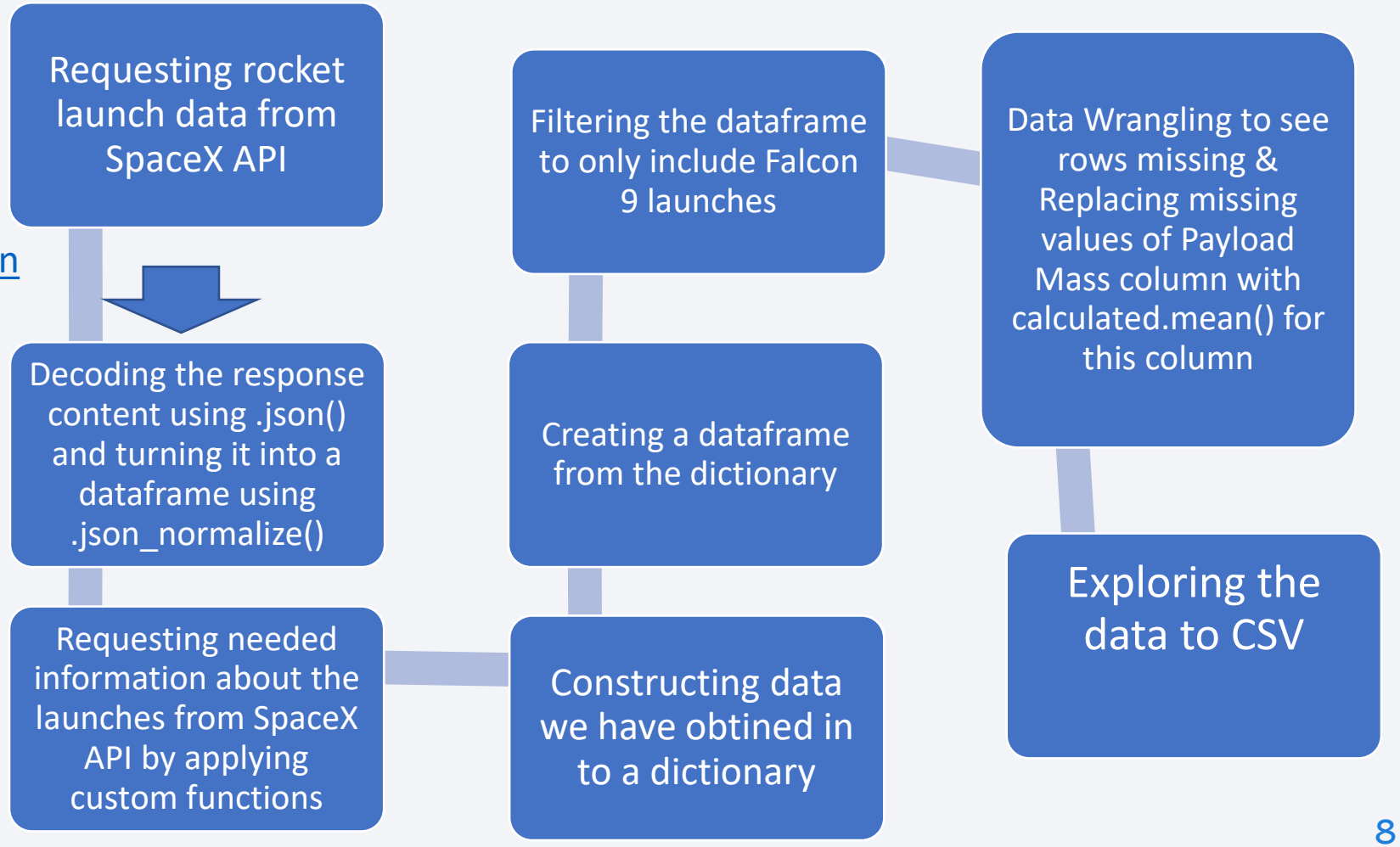
Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch

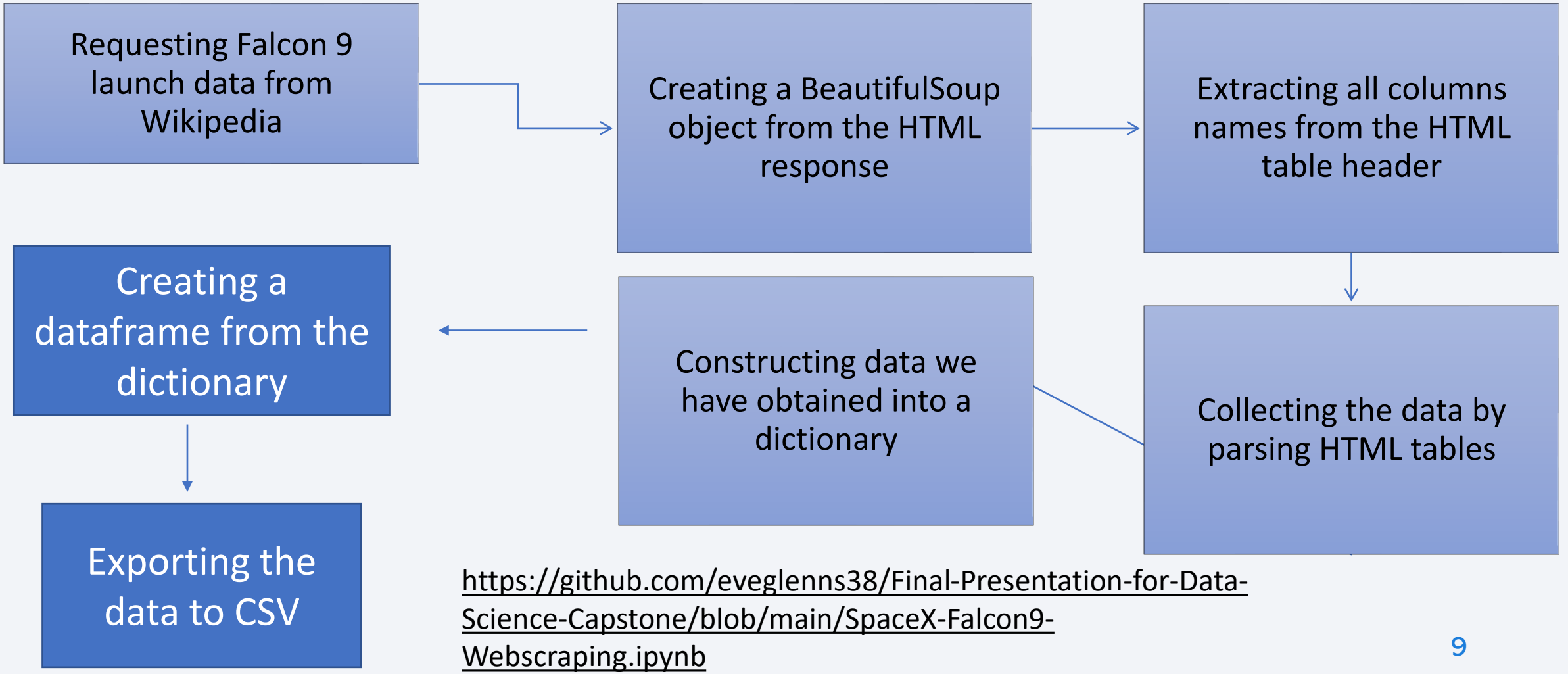
outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

https://github.com/eveglenns38/Final-Presentation-for-Data-Science-Capstone/blob/main/spacex-data-collection-api.ipynb

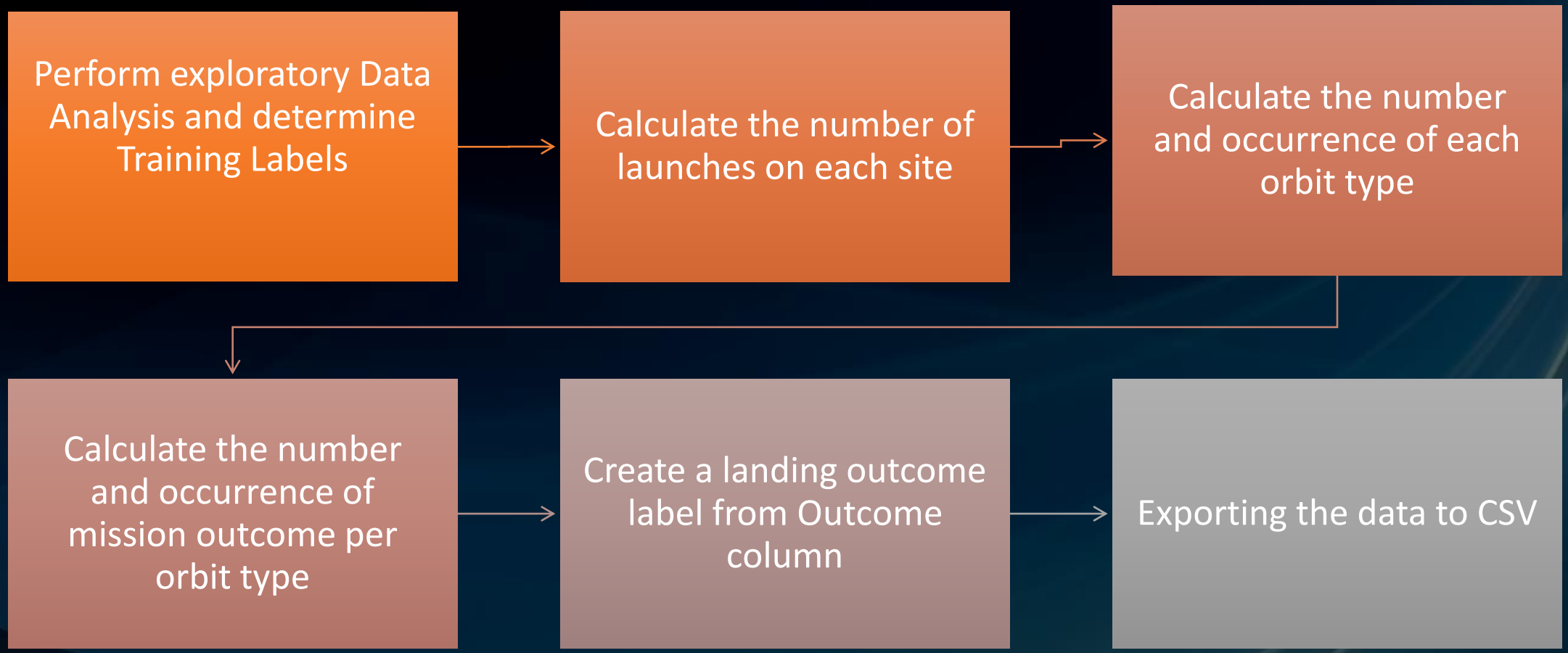


Data Collection - Scraping



https://github.com/eveglenns38/Final-Presentation-for-Data-Science-Capstone/blob/main/SpaceX-Falcon9-Webscraping.ipynb

Data Wrangling



https://github.com/eveglenns38/Final-Presentation-for-Data-Science-Capstone/blob/main/Space%20X%20Falcon%209%20Data%20Wrangling.ipynb

EDA with Data Visualization

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

Scatter plots show the relationship between variables.

If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts show trends in data over time (time series).

https://github.com/eveglenns38/Final-Presentation-for-Data-Science-Capstone/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

https://github.com/eveglenns38/Final-Presentation-for-Data-Science-Capstone/blob/main/Exploratory%20Data%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

-Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

https://github.com/eveglenns38/Final-Presentation-for-Data-Science-Capstone/blob/main/Interactive_Visual_Analytic_with_Folium.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

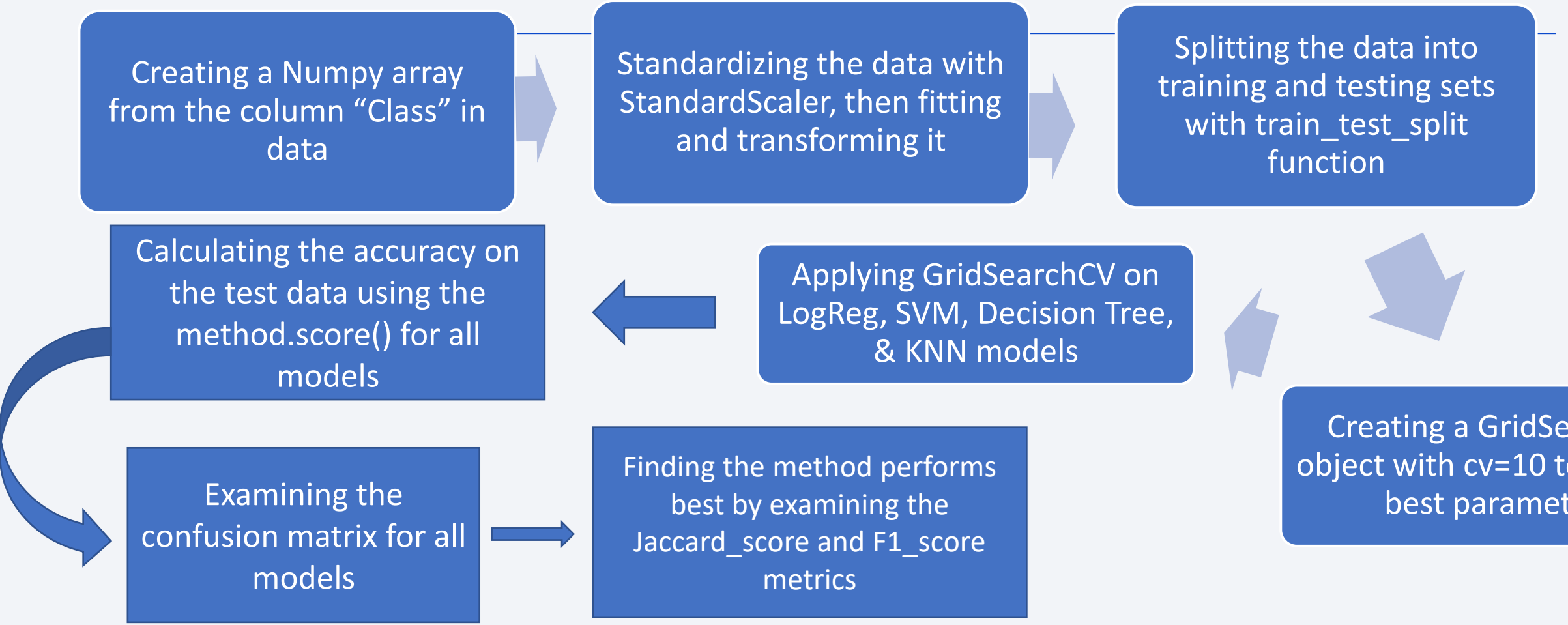
- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

https://github.com/eveglenns38/Final-Presentation-for-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/eveglenns38/Final-Presentation-for-Data-Science-Capstone/blob/main/SpaceX%20Falcon%209%20%20First%20Stage%20Landing%20Machine%20Learning%20Prediction.ipynb

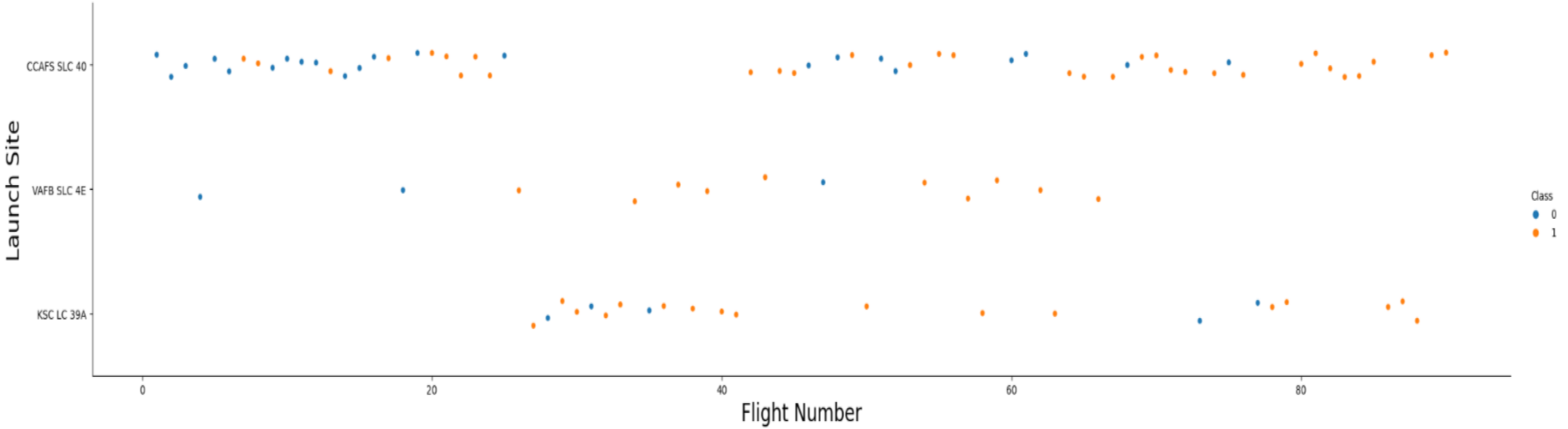
Results

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

Section 2: EDA with Visualization

Insights drawn from EDA

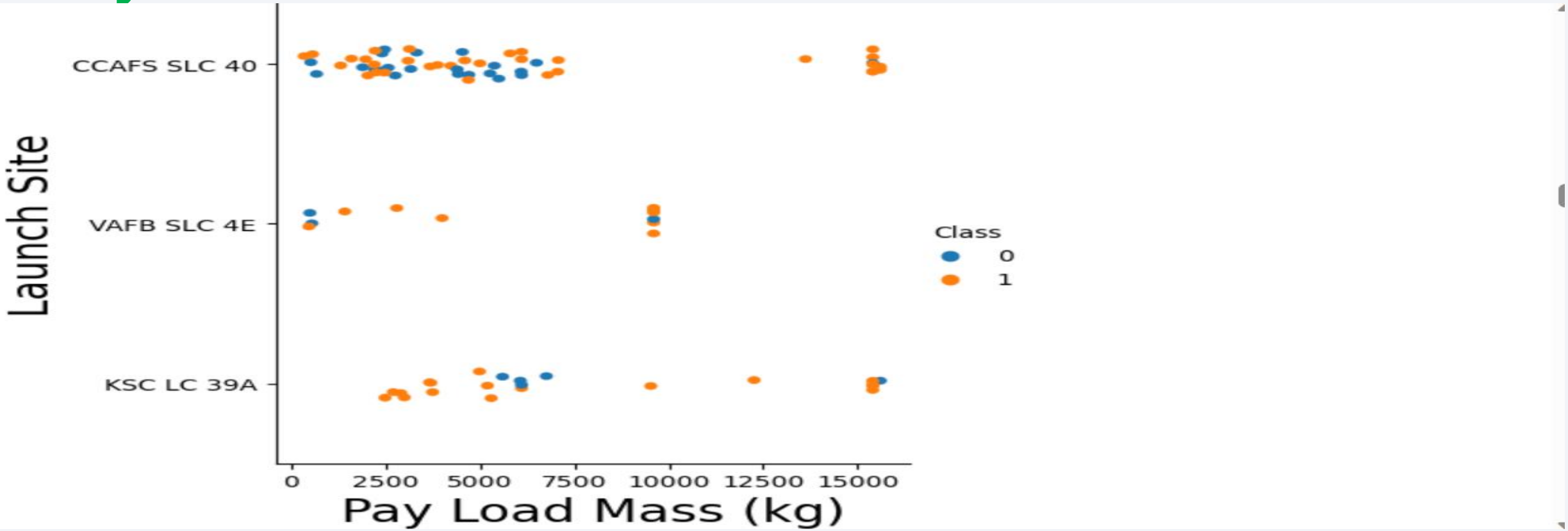
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type

Explanation:

• Orbits with 100% success rate:

- ES-L1, GEO, HEO, SSO

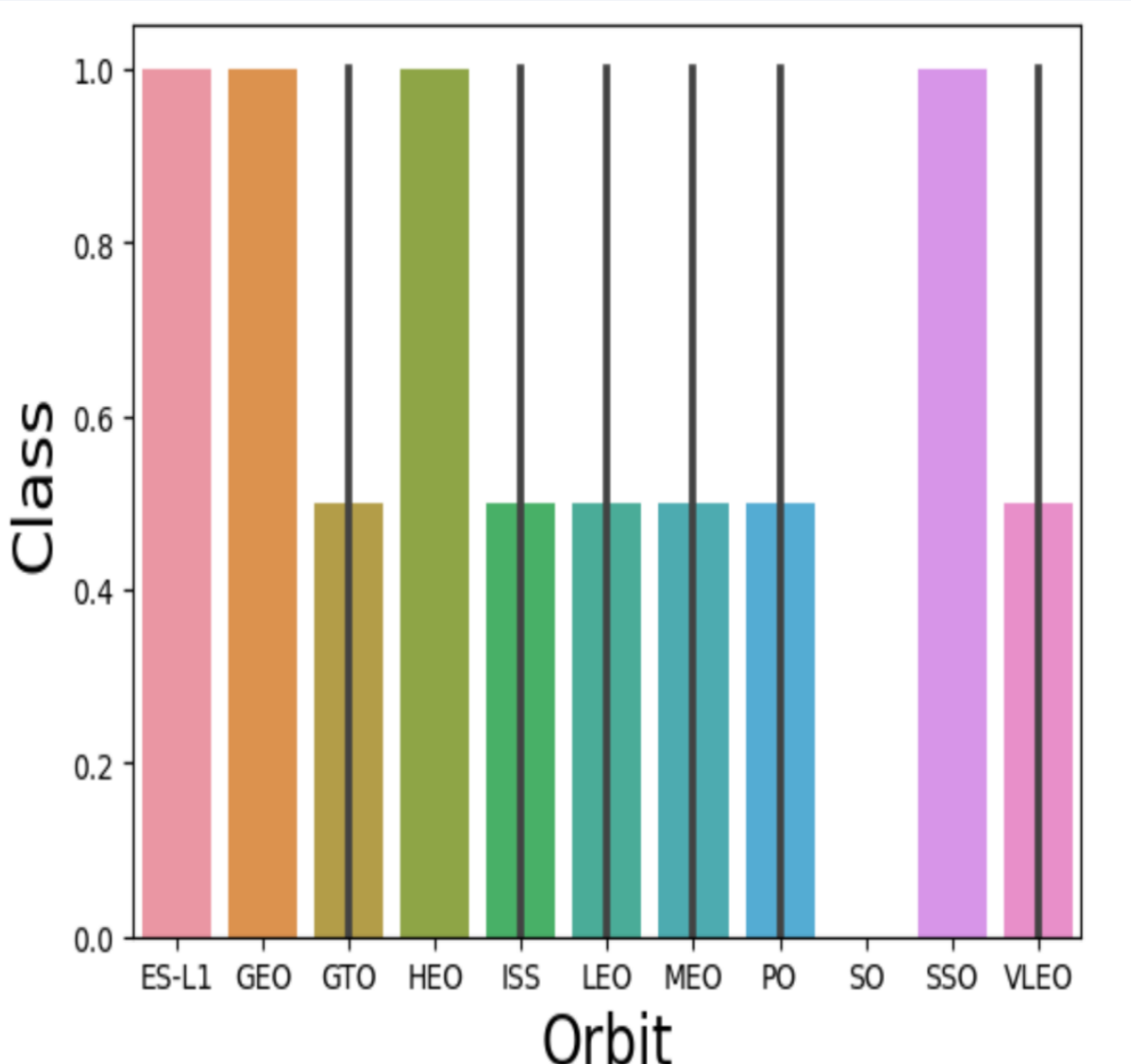
• Orbits with 0% success rate:

- SO

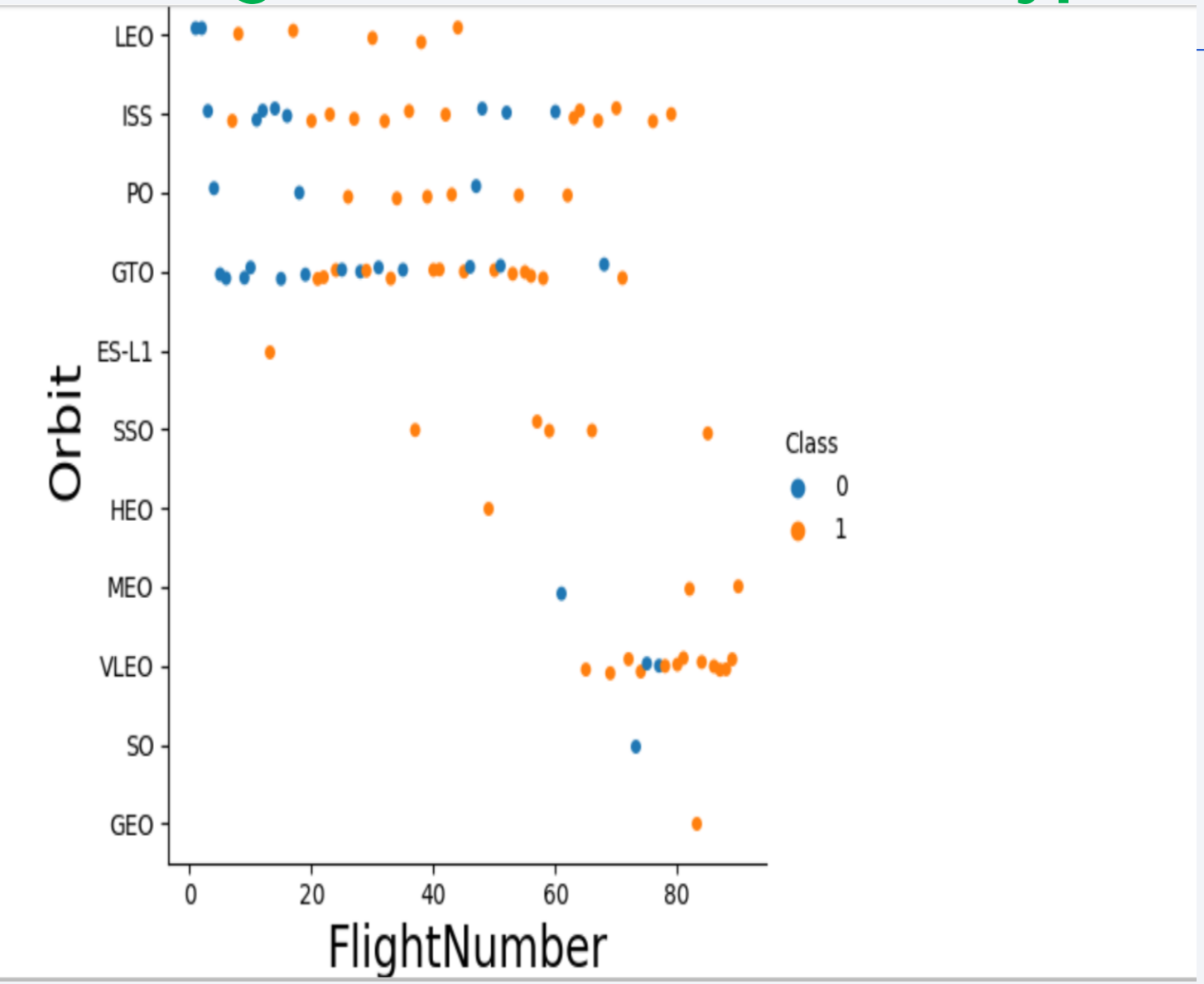
• Orbits with success rate

between 40% and 60%:

- GTO, ISS, LEO, MEO, PO



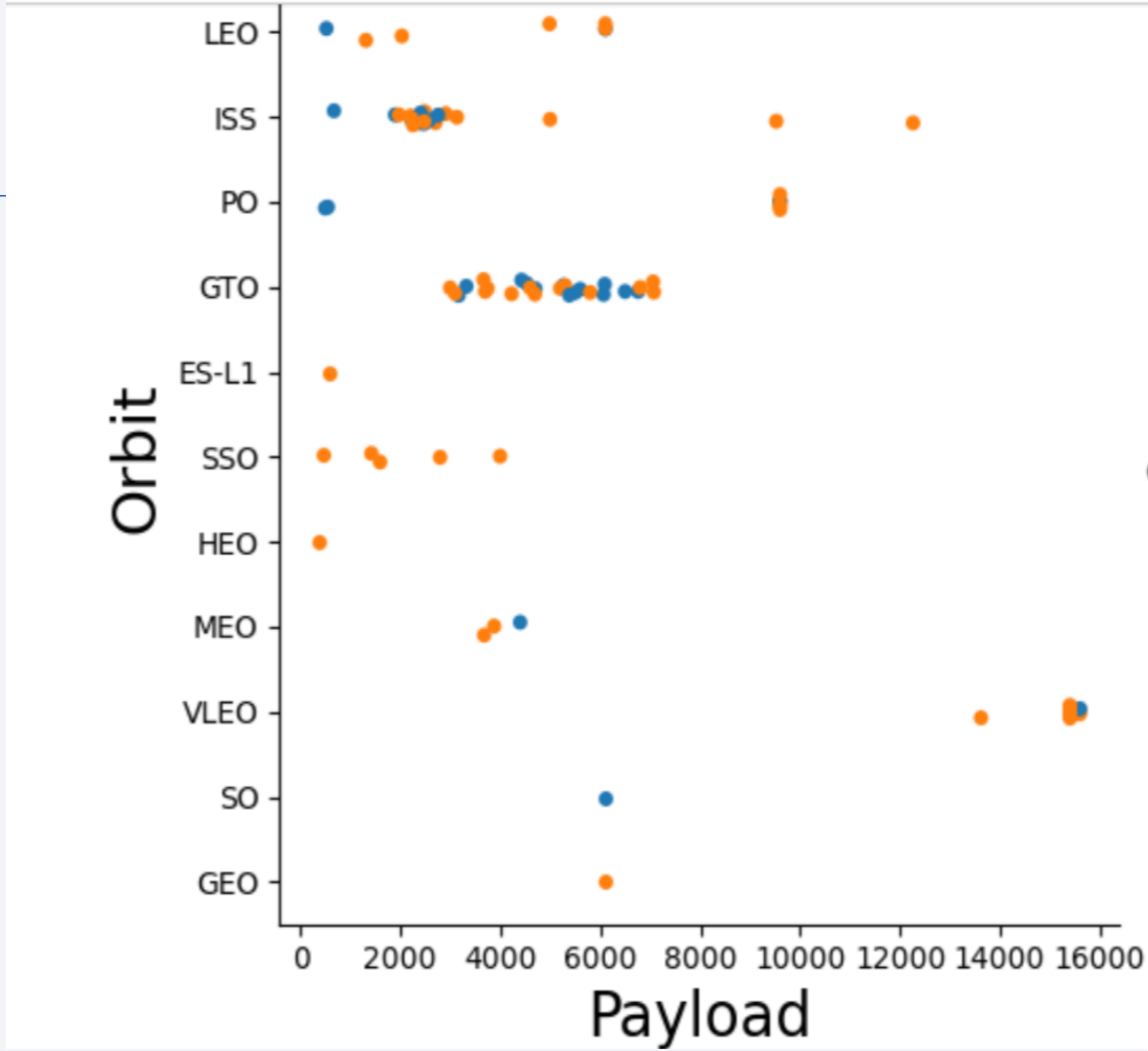
Flight Number vs. Orbit Type



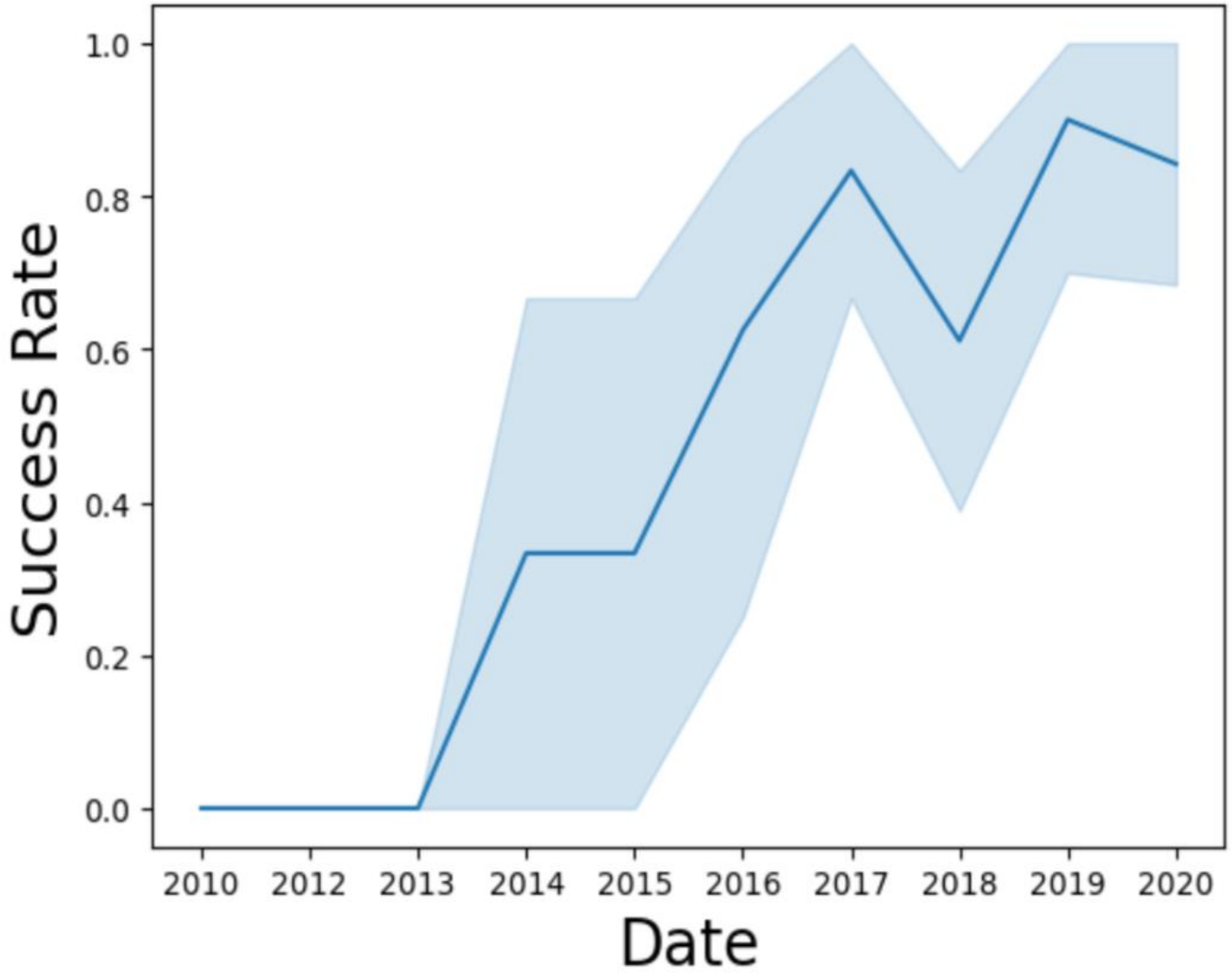
Explanation: In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

Explanation: Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend



Explanation: The success rate since 2013 kept increasing till 2020.


EDA with SQL

All Launch Site Names

[7]:

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

Explanation: Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation: Displaying 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

total_payload_mass
45596

Explanation: Displaying the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
In [10]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.

Out[10]: AVG(PAYLOAD_MASS__KG_)
         ----------------------
                        2928.4
```

Explanation: Displaying average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

first_successful_landing
2015-12-22

Explanation: Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

Out[9]:

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

Explanation: Listing the names of the boosters which have success in drone ship
and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Explanation: Listing the total number of successful and failure mission outcomes.

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

```
In [14]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
 * sqlite:///my_data1.db
Done.
```

Out[14]:

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

Boosters Carried Maximum Payload

Explanation: Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

Out[12]:

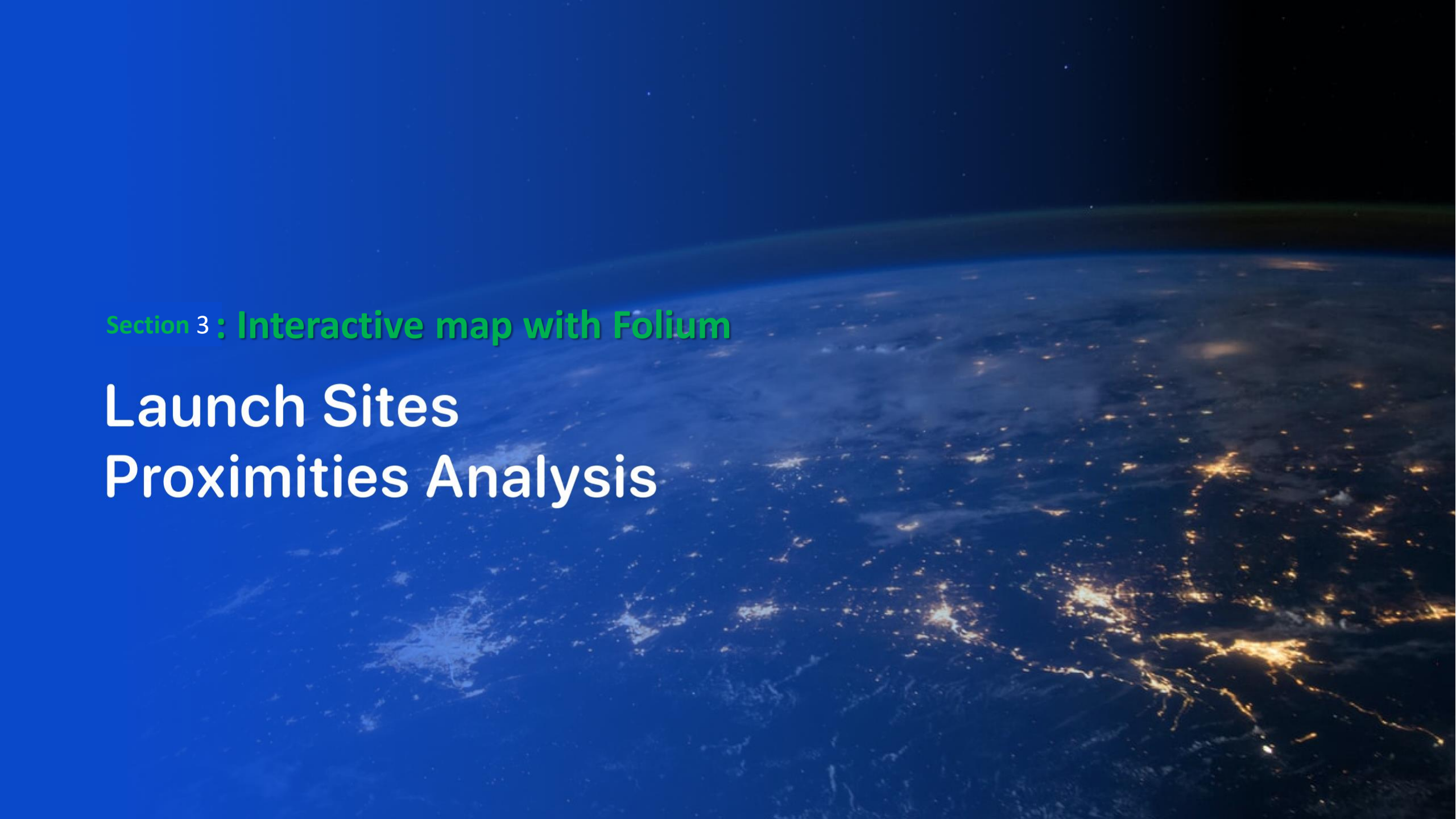
MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation: Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

Explanation: Ranking 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.

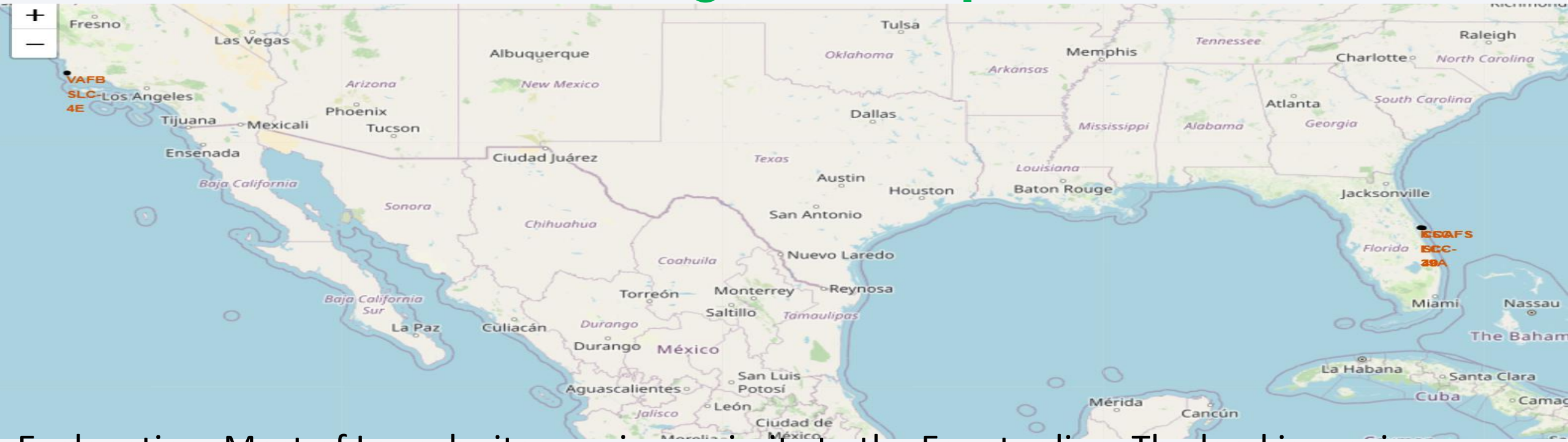
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1



Section 3 : Interactive map with Folium

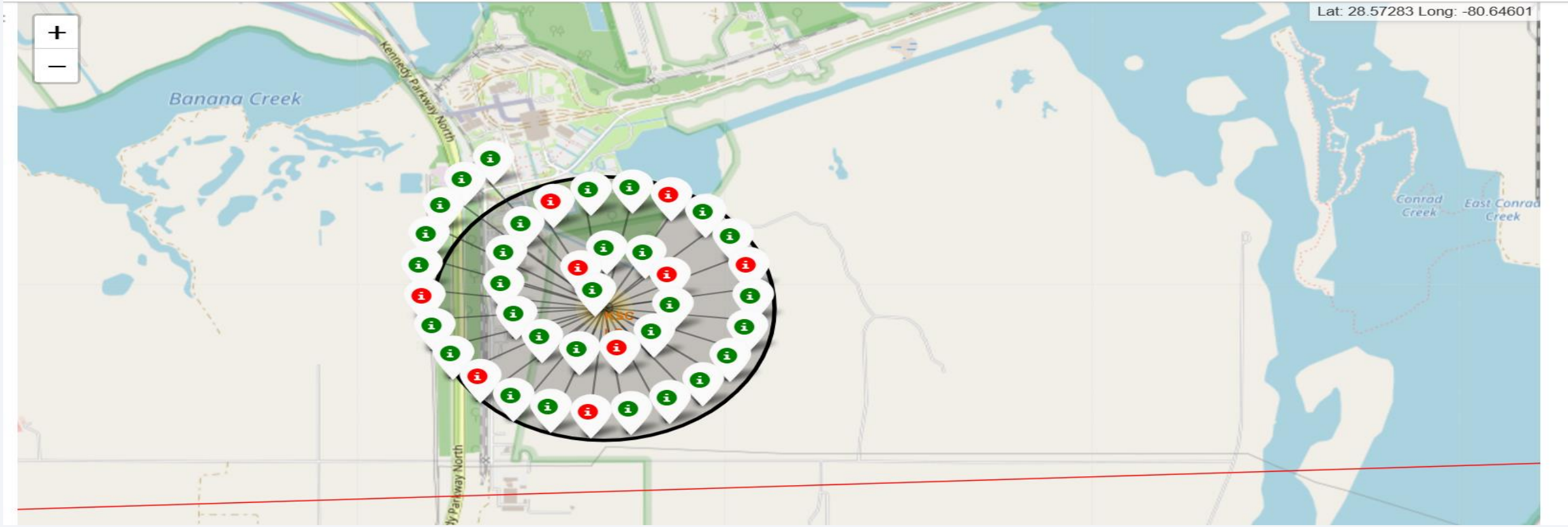
Launch Sites Proximities Analysis

All launch sites global map markers



Explanation: Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit. All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.

Colour-labeled launch records on the map



Explanation: From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- **Green Marker** = Successful Launch
- **Red Marker** = Failed Launch

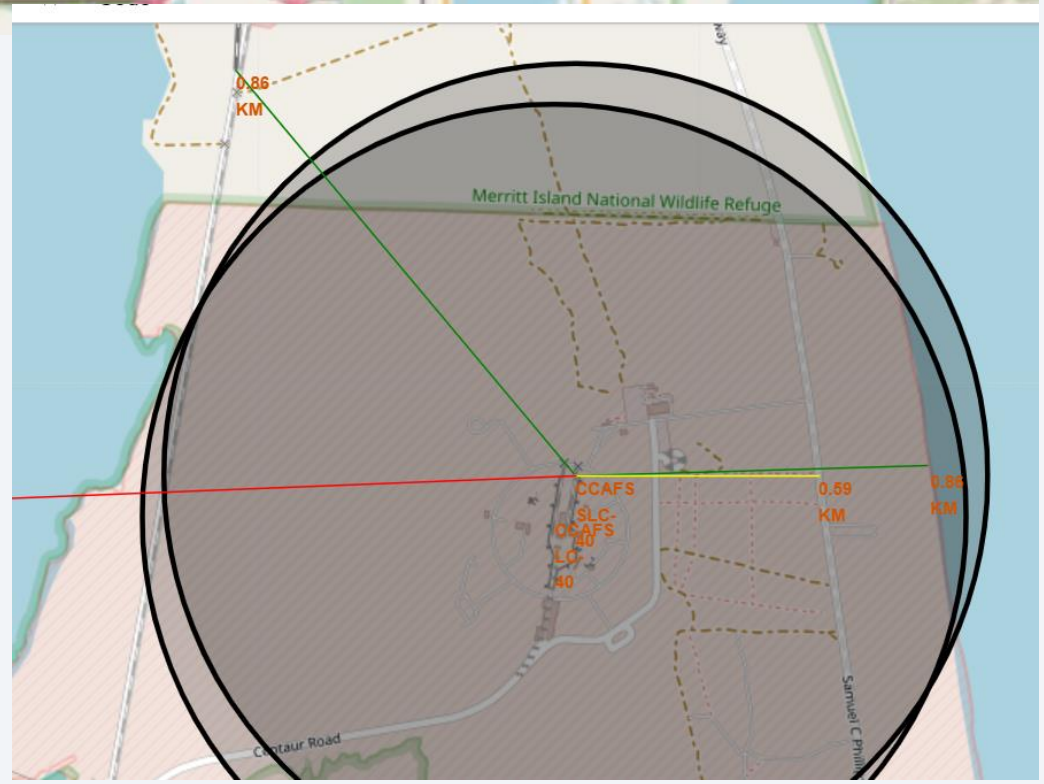
✓ Launch Site KSC LC-39A has a very high Success Rate.

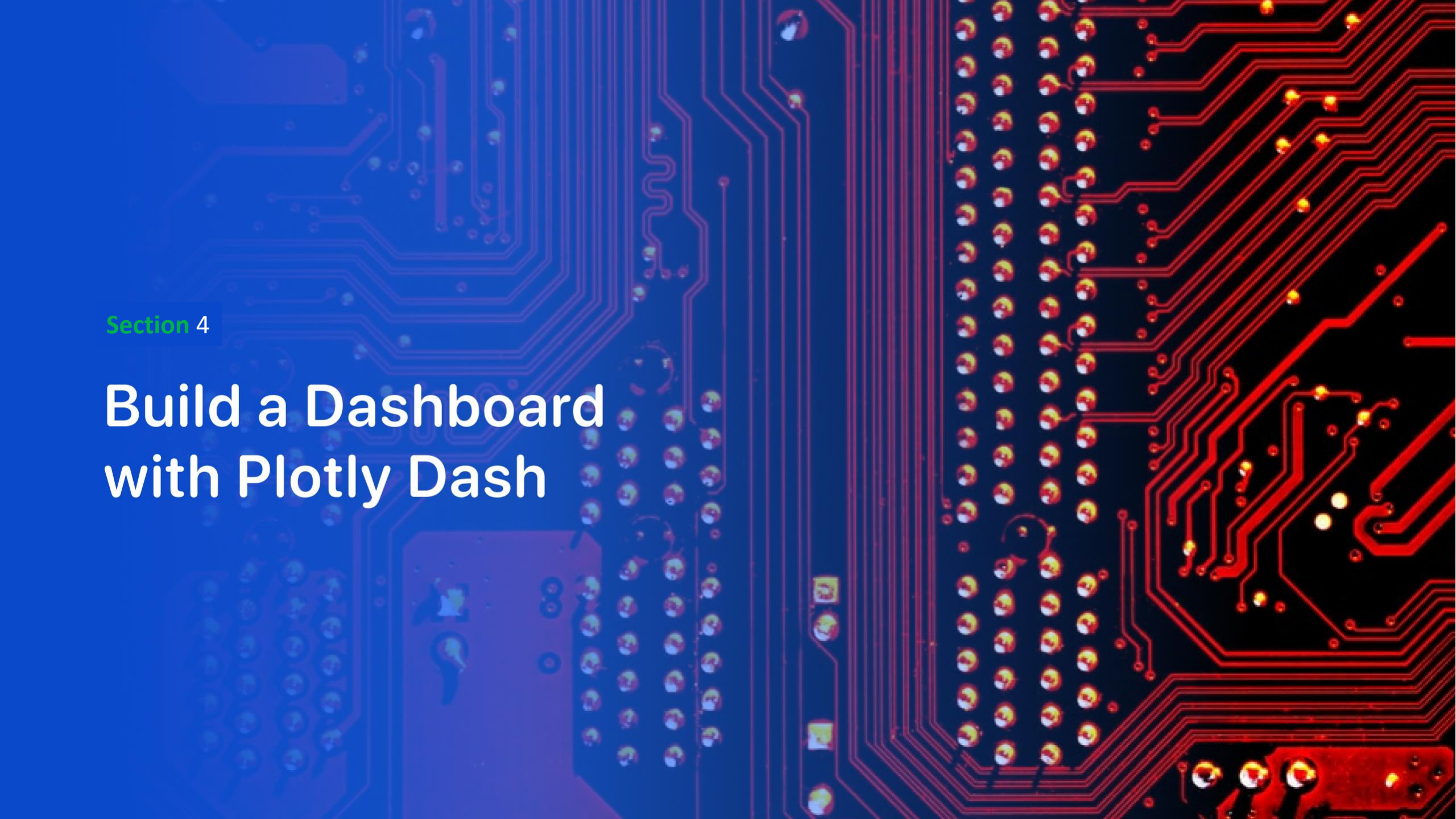
Distance from the launch site CCAFS CLC-40 to its proximities

Explanation: From the visual analysis of the launch site CCAFS CLC-40 we can clearly see that it is:

- relative close to railway (0.86 km)
- relative close to highway (0.59km)
- relative close to coastline (0.86 km)

Also the launch site CCAFS CLC-40 is (78.64 km) from Orlando City.

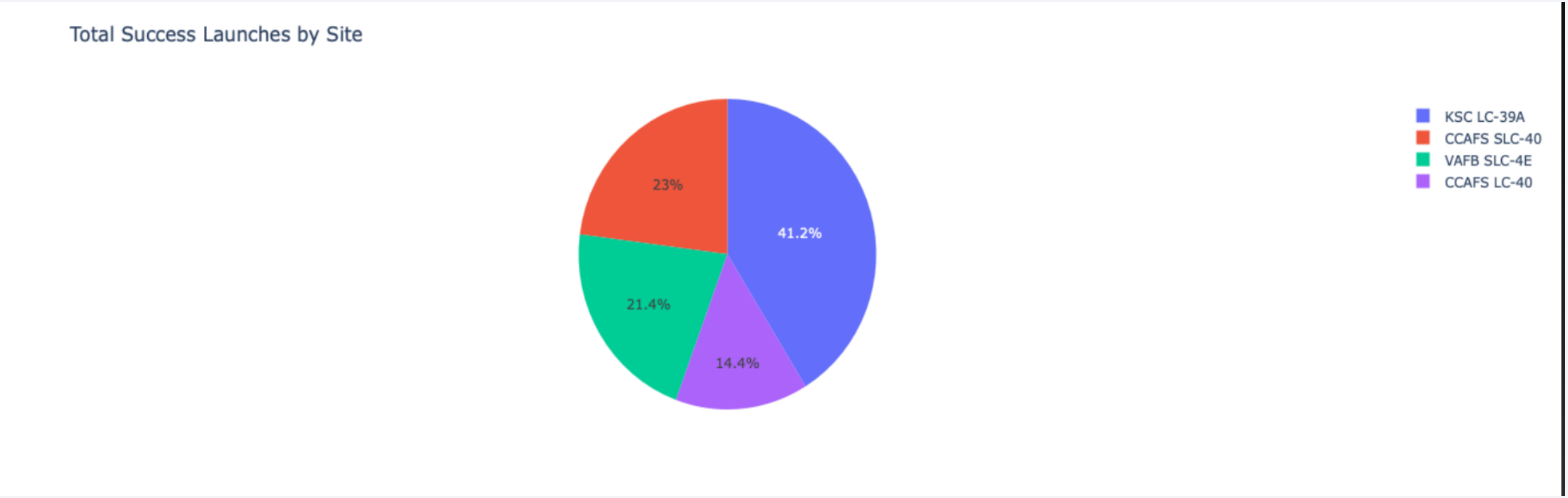




Section 4

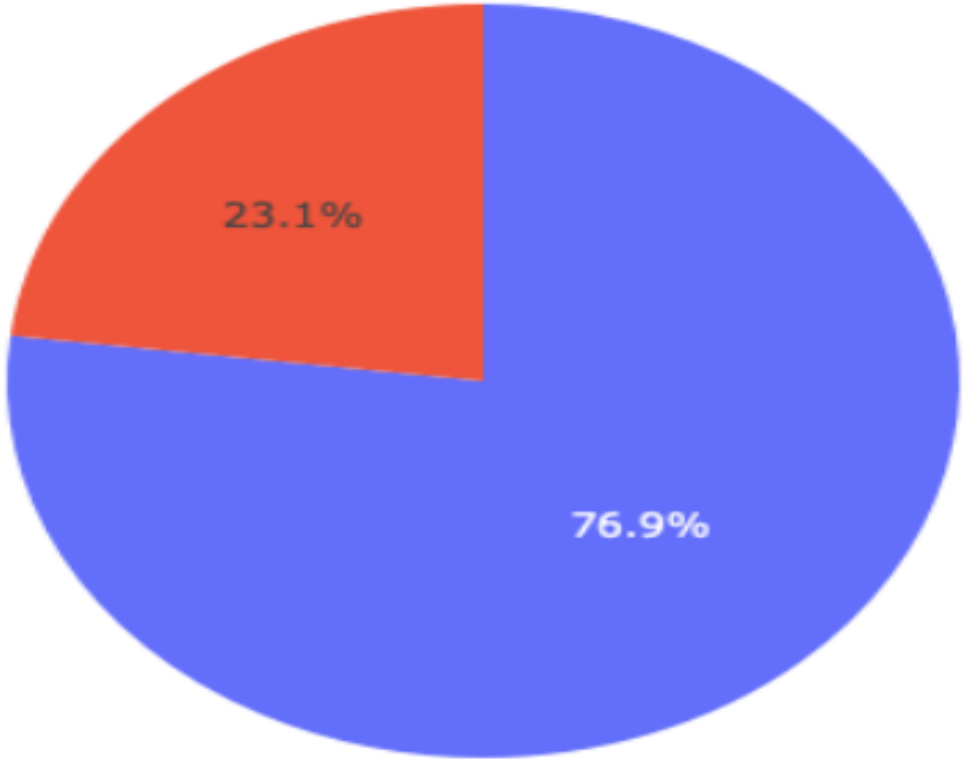
Build a Dashboard with Plotly Dash

Launch success count for all sites



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches and the CCAFS LC-40 has the least successful launches.

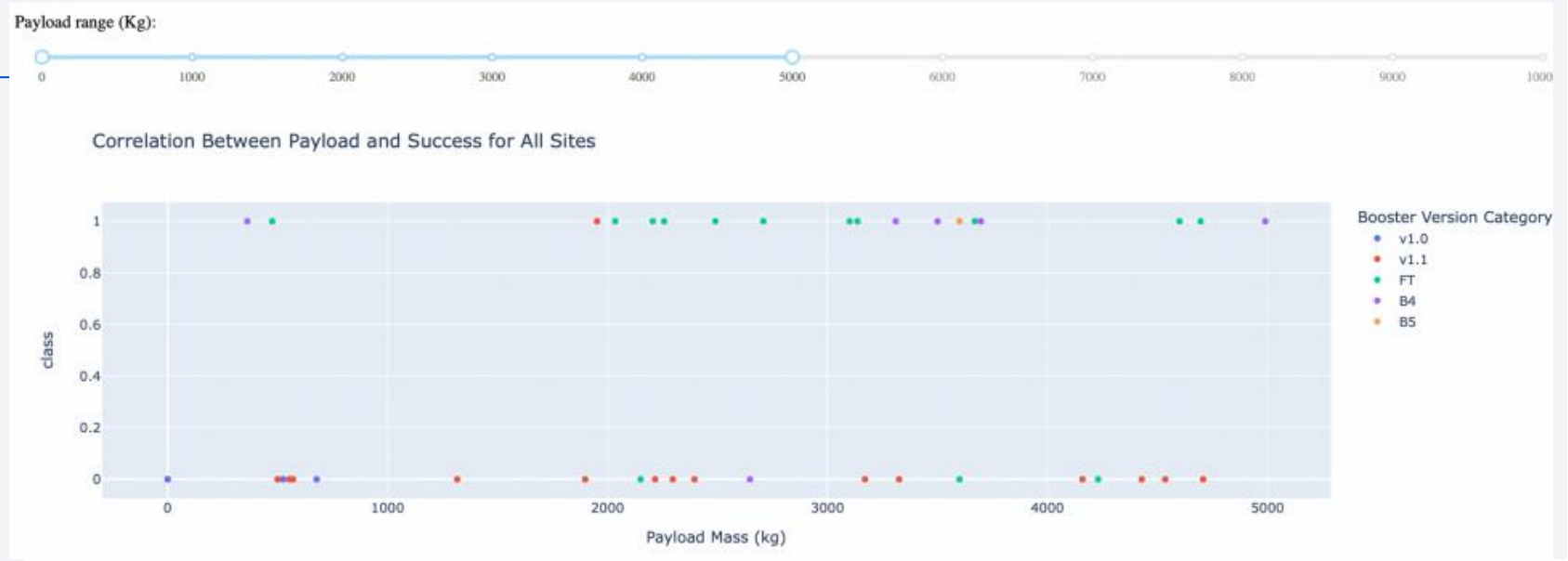
Launch site with highest launch success ratio

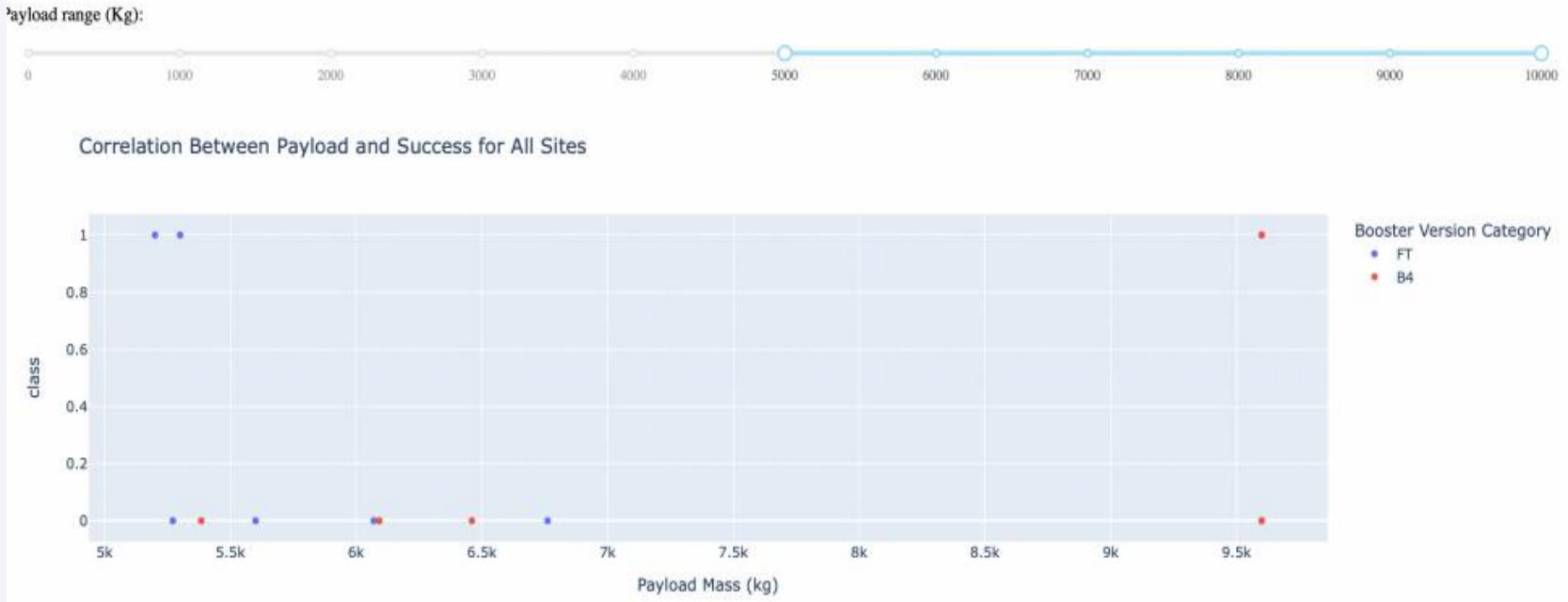


KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only failed landings.

Payload Mass vs. Launch Outcome for all sites

The charts show that payloads between 2000 and 5500 kg have the highest success rate.





Classification Accuracy

Explanation: The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Entire Test Set

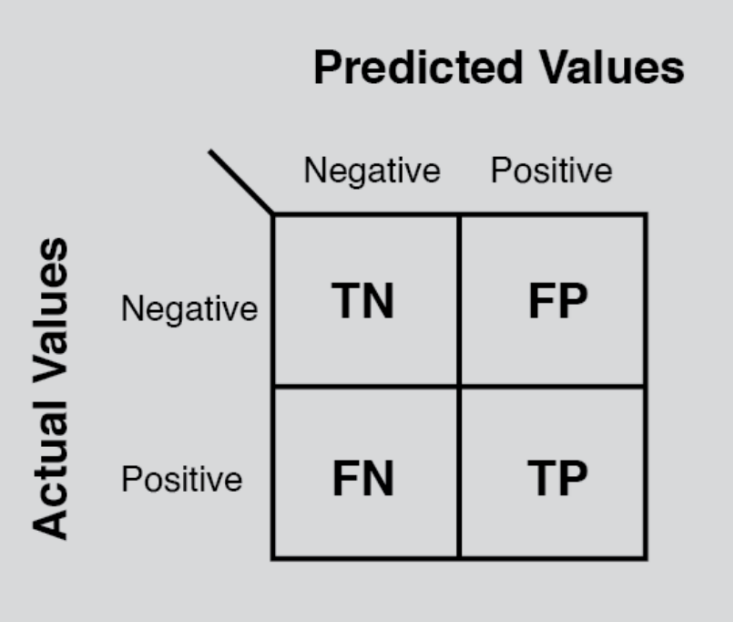
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

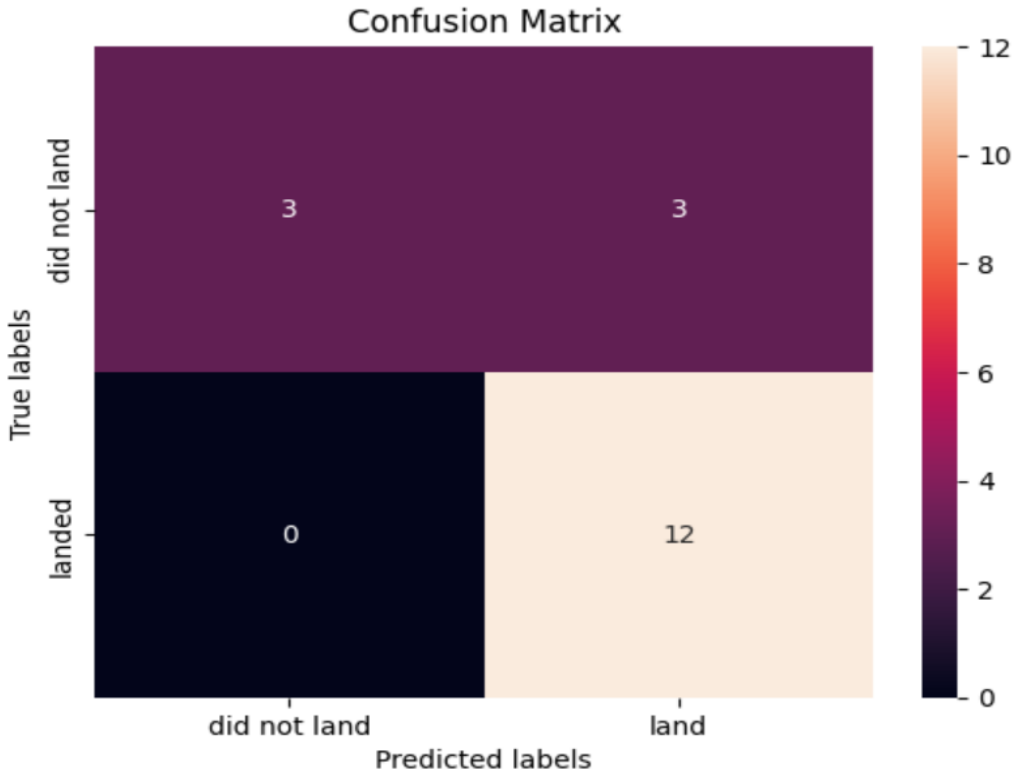
Section 5

Predictive Analysis (Classification)

Confusion Matrix

Explanation: Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.





Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- KSC LC-39A has the highest success rate of the launches from all the sites. Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

Special Thanks to:

Instructors, IBM & Coursera

Thank you!