IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Badr DRIDAKH
07 Aug 2024

Outline

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

Executive Summary

Methodologies:

- Data for this project was sourced from two main channels: a GET request to the SpaceX API and web scraping.
- The data underwent processing and wrangling to extract insights, followed by exploratory data analysis through visualizations.
- Interactive visual analysis was conducted using Folium and Plotly Dash.
- Predictive analysis was executed using classification models, which were assessed on both training and testing datasets to ensure high accuracy.

Results:

- It was found that heavier payloads negatively impact launch success, whereas higher flight numbers and longer durations improve success rates.
- This information can be utilized to predict whether SpaceX will reuse its first stage.

Introduction

The commercial space age has arrived, making space travel more affordable.

- Virgin Galactic: Provides suborbital spaceflights | Rocket Lab: Specializes in small satellite launches | Blue Origin: Manufactures sub-orbital and orbital reusable rockets
- SpaceX: Notable achievements include:
  - ✓ Sending spacecraft to the International Space Station
  - ✓ Developing Starlink, a satellite internet constellation
  - ✓ Conducting manned space missions
- SpaceX's rocket launches are relatively inexpensive, with Falcon 9 launches costing $62 million compared to other providers' costs of $165 million or more

Much of SpaceX's cost savings come from reusing the first stage of their rockets. Thus, predicting the landing success of the first stage can help determine the overall launch cost.

Project Objectives:

- Determine the price of each SpaceX launch
- Gather and analyze information about SpaceX
- Create dashboards to present data and insights to your team
- Predict whether SpaceX will reuse the first stage of their rockets
- Train a machine learning model using public information to forecast the landing success of the first stage

Section 1

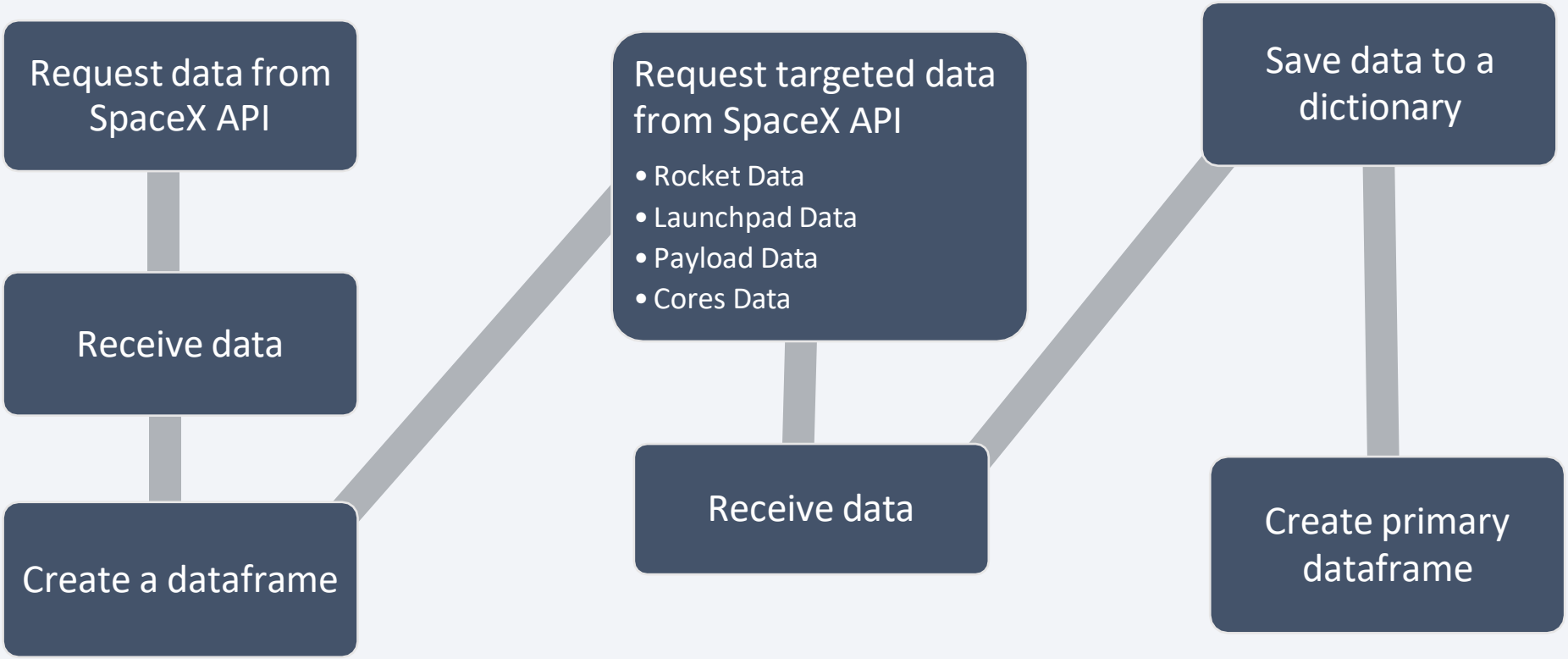
Methodology

Methodology – Executive Summary

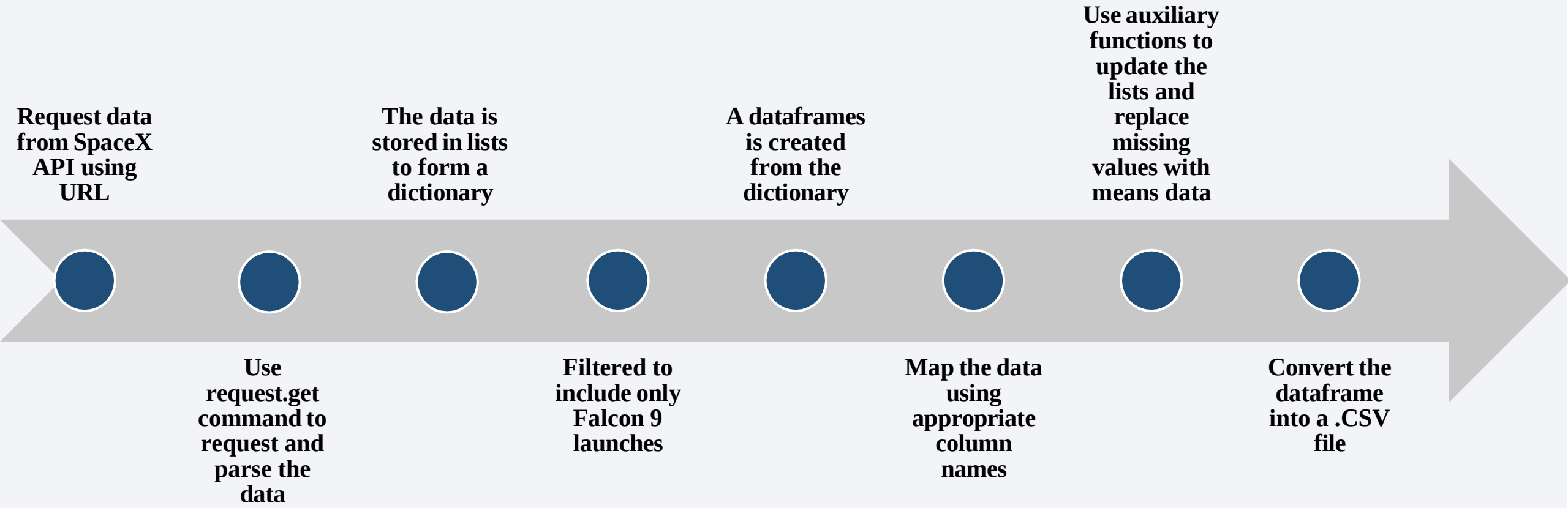
- Data collection methodology:
  - GET request to the SpaceX API and Web Scraping of launch information from Wikipedia
- Data Wrangling:
  - The data was filtered for actionable information and wrangled to enable the draw of conclusions
- Performed Exploratory Data Analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models

Data Collection

Launch data was retrieved from the SpaceX API using the GET command (requests.get). The response was then converted into a pandas dataframe. Then additional sub-requests were made to obtain further information about the columns stored in the dataframe. Using helper functions, the responses were stored in a dictionary which was convert it into a new dataframe to serve as our primary dataset.

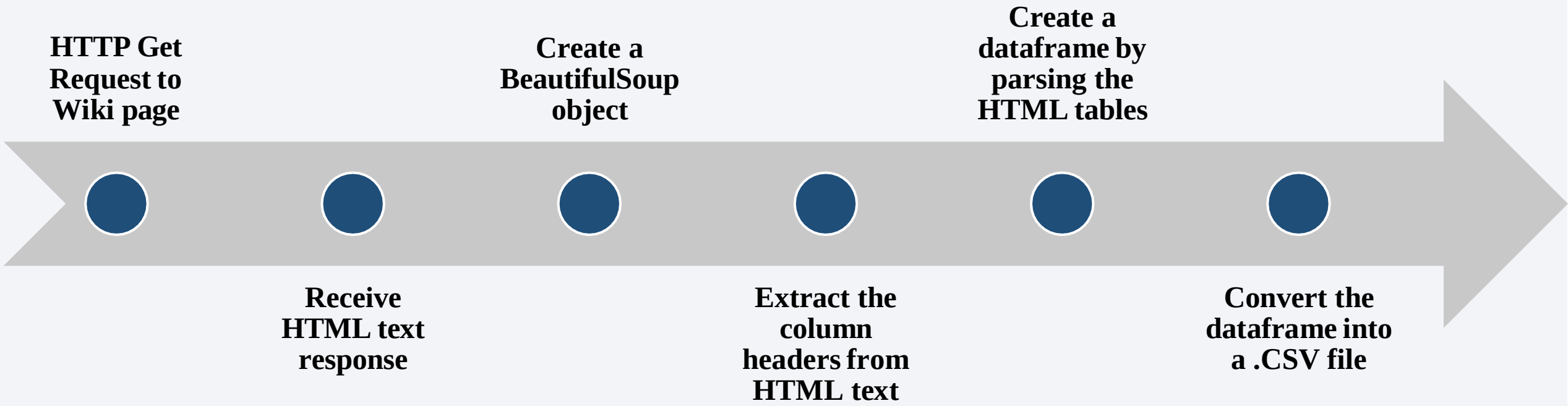


Data Collection - SpaceX API



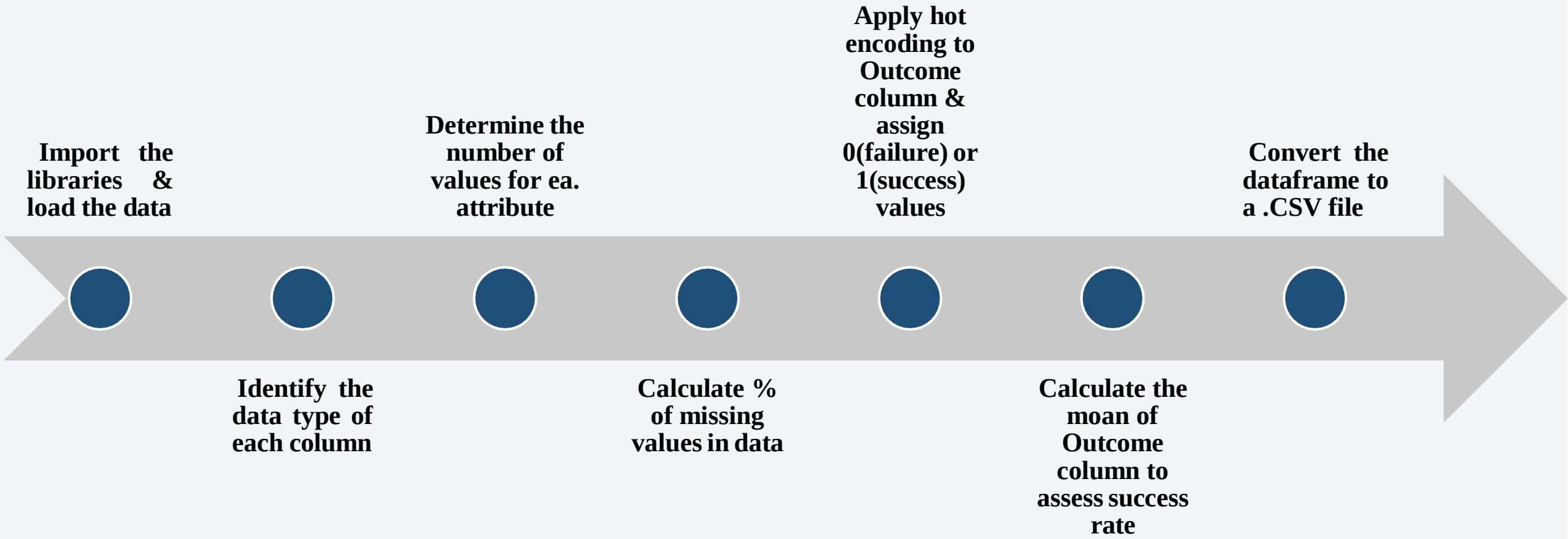
https://github.com/badrdridakh/Winning-Space-Race-With-Data-Science

Data Collection - Scraping



https://github.com/badrdridakh/Winning-Space-Race-With-Data-Science

Data Wrangling



https://github.com/badrdridakh/Winning-Space-Race-With-Data-Science

EDA with Data Visualization

- To get a better understanding of the relationship between variables, the data was visualized with scatter plots, bar chart, and line chart.
- Relationships Assessed:
  - Pay load mass against the Flight number
  - Lunch site against the Flight number
  - Lunch site against the Pay load mass
  - Orbit type against Class success rate
  - Flight number against Orbit type
  - Orbit type against the Pay load mass
  - Launch success yearly trend

Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

EDA with SQL

- To gain further insight into the datasets, the following SQL queries were performed:
  - Displayed the names of the unique launch sites in the space mission
  - Displayed 5 records where launch sites begin with the string 'CCA'
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1
  - Listed the date when the first successful landing outcome in ground pad was achieved
  - Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listed the total number of successful and failure mission outcomes
  - Listed the names of the booster versions which have carried the maximum payload mass
  - Listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Ranked 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

Add the GitHub URL of your completed EDA with SQL notebook, as an external reference

Build an Interactive Map with Folium

To visualize the launch sites better an interactive map was created with the following objects:

- Markers:
  - Purpose: Visualize each launch site on the map.
  - Reason: Helps in identifying the locations of the launch sites and their proximity to significant geographical features like the equator.
- Circles:
  - Purpose: Highlight the site locations.
  - Reason: Enhances visualization by making the launch sites more prominent on the map.
- Colored Markers:
  - Purpose: Indicate the outcome of each launch.
  - Reason: Different colors (green for success, red for failure) help in quickly identifying which sites have higher success rates.
- Mouse Pointer:
  - Purpose: Facilitate the retrieval of longitude and latitude coordinates.
  - Reason: Improves the ease of getting precise geographical data for the launch sites.
- Lines (Polylines):
  - Purpose: Visualize the distance between launch sites and significant structures like railways.
  - Reason: Assists in analyzing the proximities and logistical considerations of the launch sites.
- Initial Center Location:
  - Purpose: Center the map around NASA Johnson Space Center, Houston, Texas.
  - Reason: Provides a starting point for the map, making it easier to navigate and explore the launch sites.

These objects collectively enhance the map's functionality and visualization, aiding in the analysis of launch site locations and their success rates.

Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- A dashboard was created for users to perform interactive visual analytics on SpaceX launch data in real-time.
- The dashboard application contained input components such as a dropdown list to interact with a pie chart and a range slider to interact with a scatter point chart.
  - The dropdown list allows users to see visual analytics by launch site via a pie chart depicting the total success launch rate by site
  - The ranger slider enabled the users to visualize the correlation between payload range (Kg) in relation to a payload masses vs. launch outcome scatter plot

Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

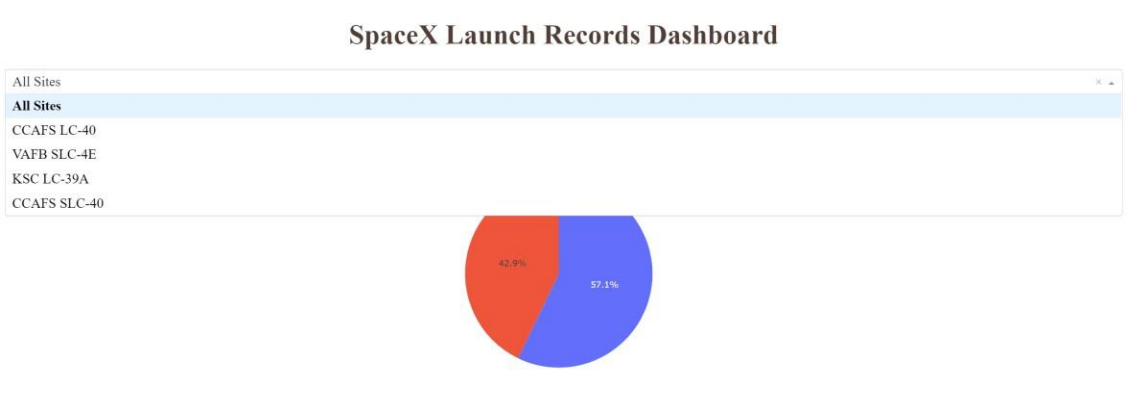
The classification model was built, evaluated, improved, and the best performing model was found:

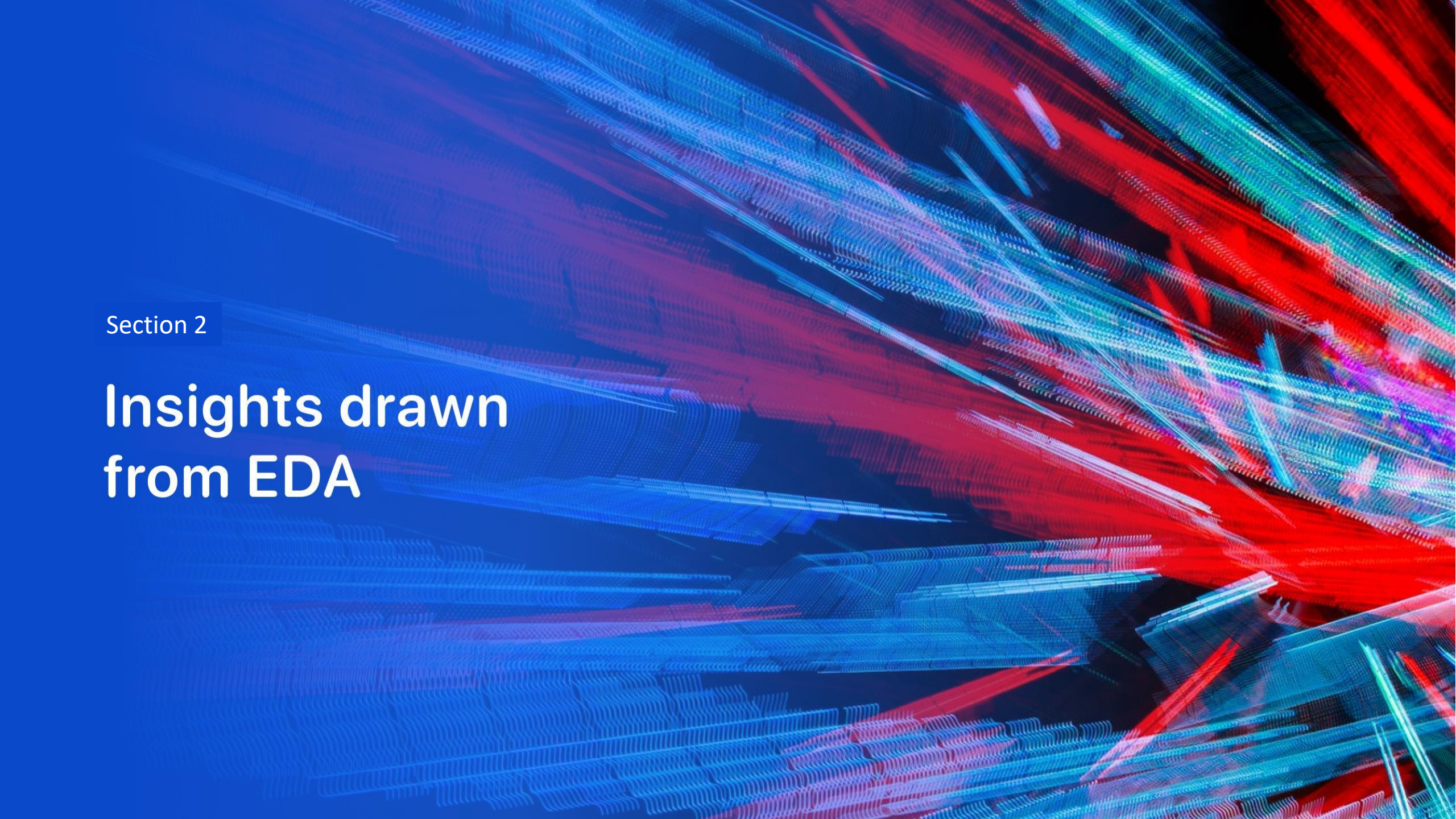
- Model Creation:
  - An object was created for each machine learning algorithm: Logistic Regression, Support Vector Machine, Decision Trees, and K-Nearest Neighbor.
- Parameter Tuning:
  - A GridSearchCV object was created for each model to perform hyperparameter tuning.
  - The GridSearchCV object was fitted to find the best parameters from a predefined dictionary of parameters.
- Evaluation:
  - The best parameters for each model were displayed.
  - The accuracy of each model on the validation data was calculated.
- Data Preparation:
  - The 'Class' column was used as the label for the predictor.
  - The data was normalized and then split into training and test sets.
  - The training data was further divided into a validation set for model evaluation.
- Model Selection:
  - After fitting the GridSearchCV object to each algorithm, the most accurate model was chosen based on its performance on the validation data.

This process ensured that the best performing classification model was identified through systematic evaluation and improvement.

Results

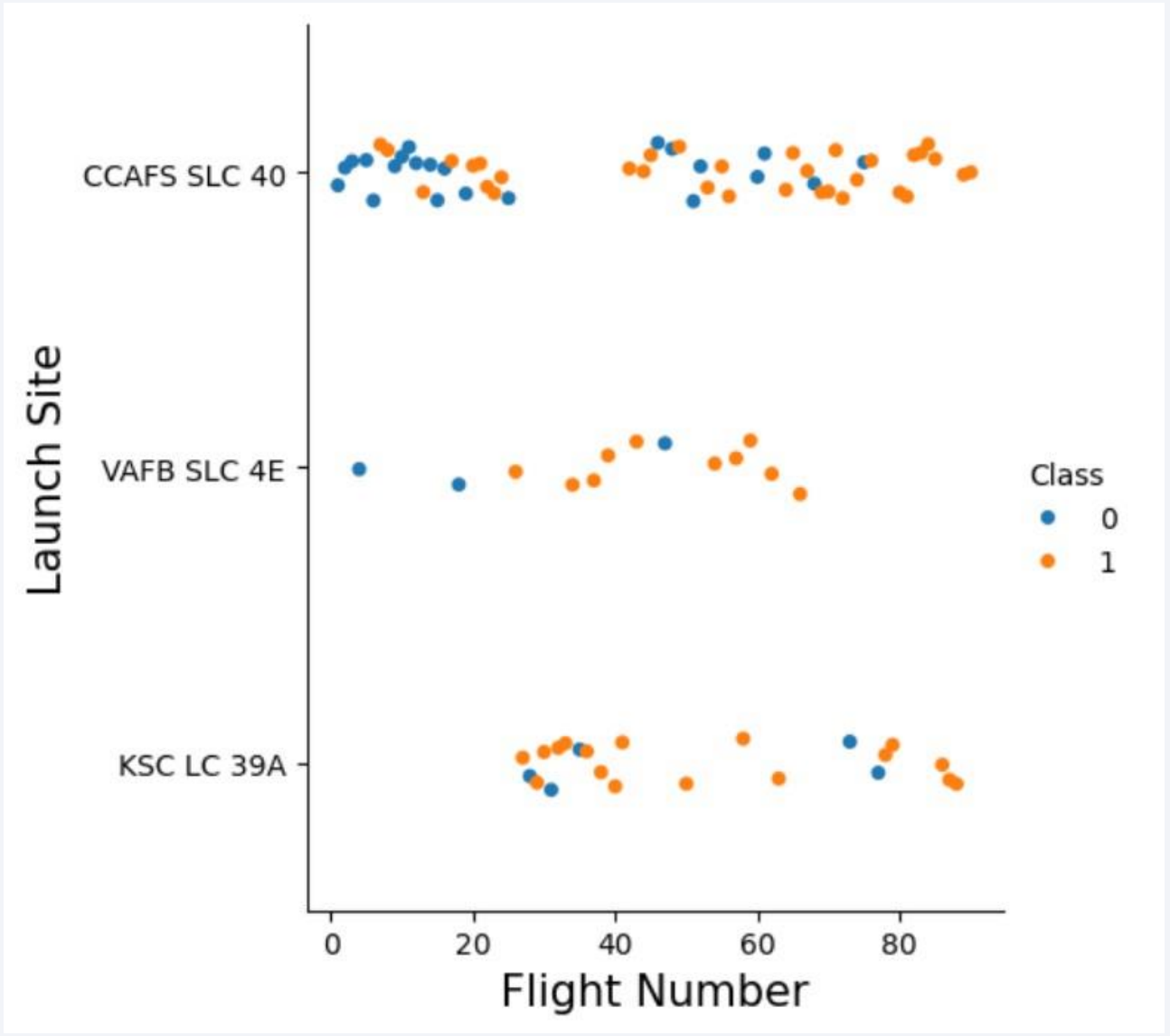
- Exploratory data analysis results:
  - Launch success rate see a steady increase after 2013
  - The orbit type and launch site have a direct correlation to the outcome of the launches
- Predictive analysis results:
  - The decision tree model demonstrated the highest accuracy at 0.9
- Interactive analytics demo in screenshots




Section 2
Insights drawn from EDA

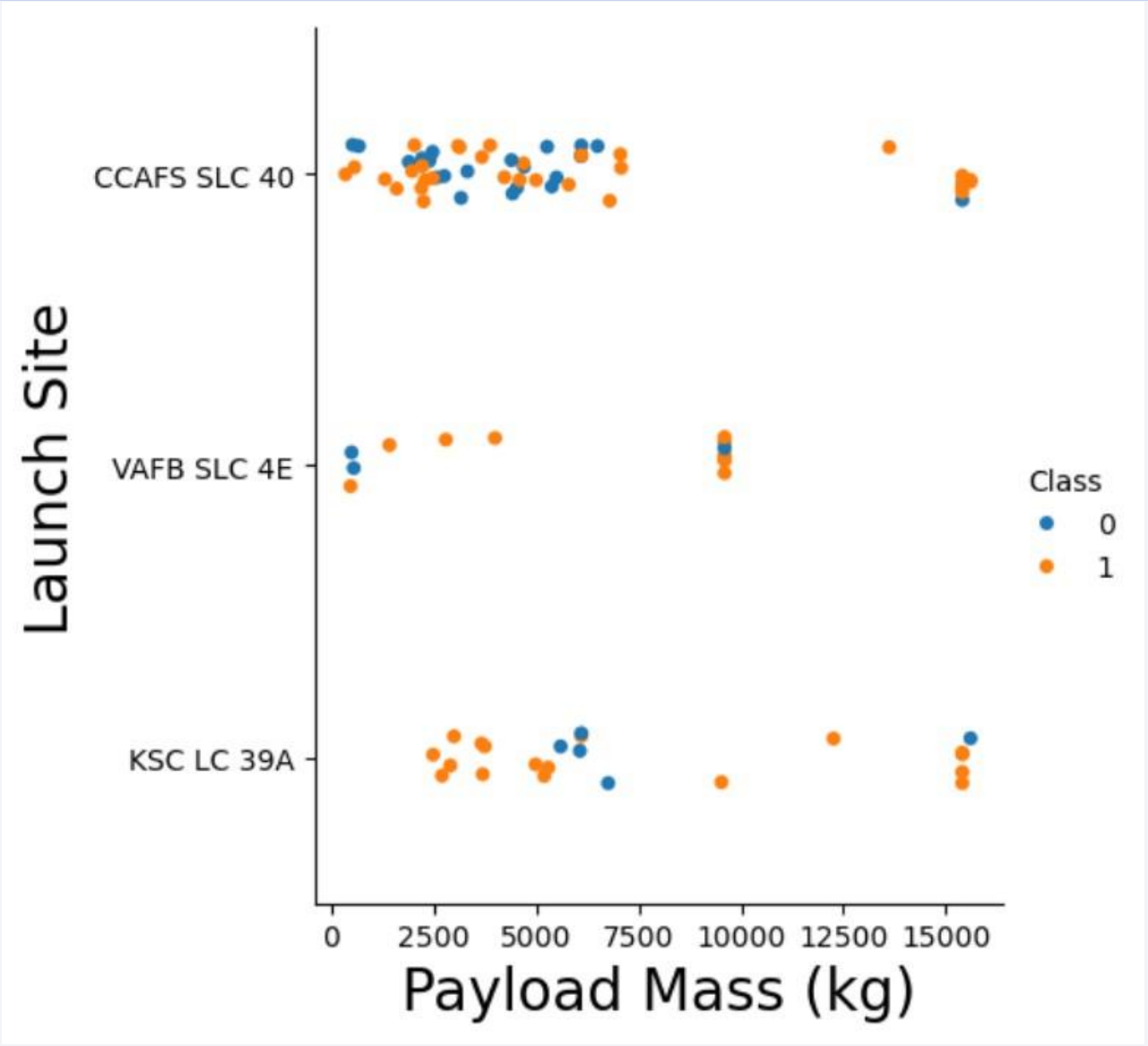
Flight Number vs. Launch Site

- CCAFS SLC 40 shows consistent success in flight number range ~40-90 but had a higher fail rate at the ~0-30 range
- VAFB SLC 4E has success in the range of ~20-60 flight number (note: has least data points)
- KSC LC 39A demonstrates a high success to failure ratio in the flight number rage ~26-90



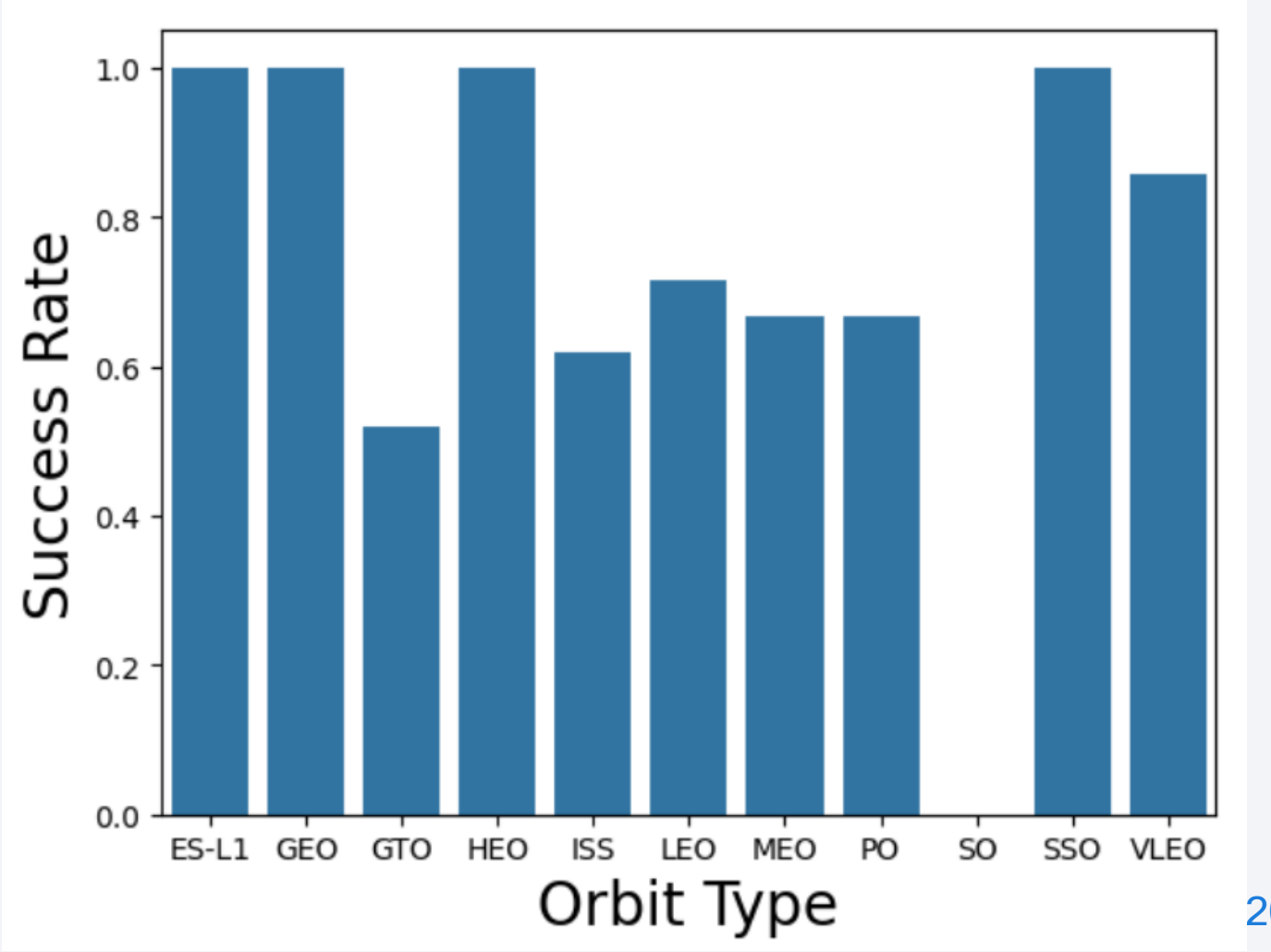
Payload vs. Launch Site

- CCAFS SLC 40 show a near equal success to fail ration in the payload mass range ~0-7500
- VAFB SLC 4E demonstrates high success rate at the 10K payload but seems to be unable to support higher payloads since no launches occurred
- KSC LC 39A has a high success to failure ratio in terms of payload



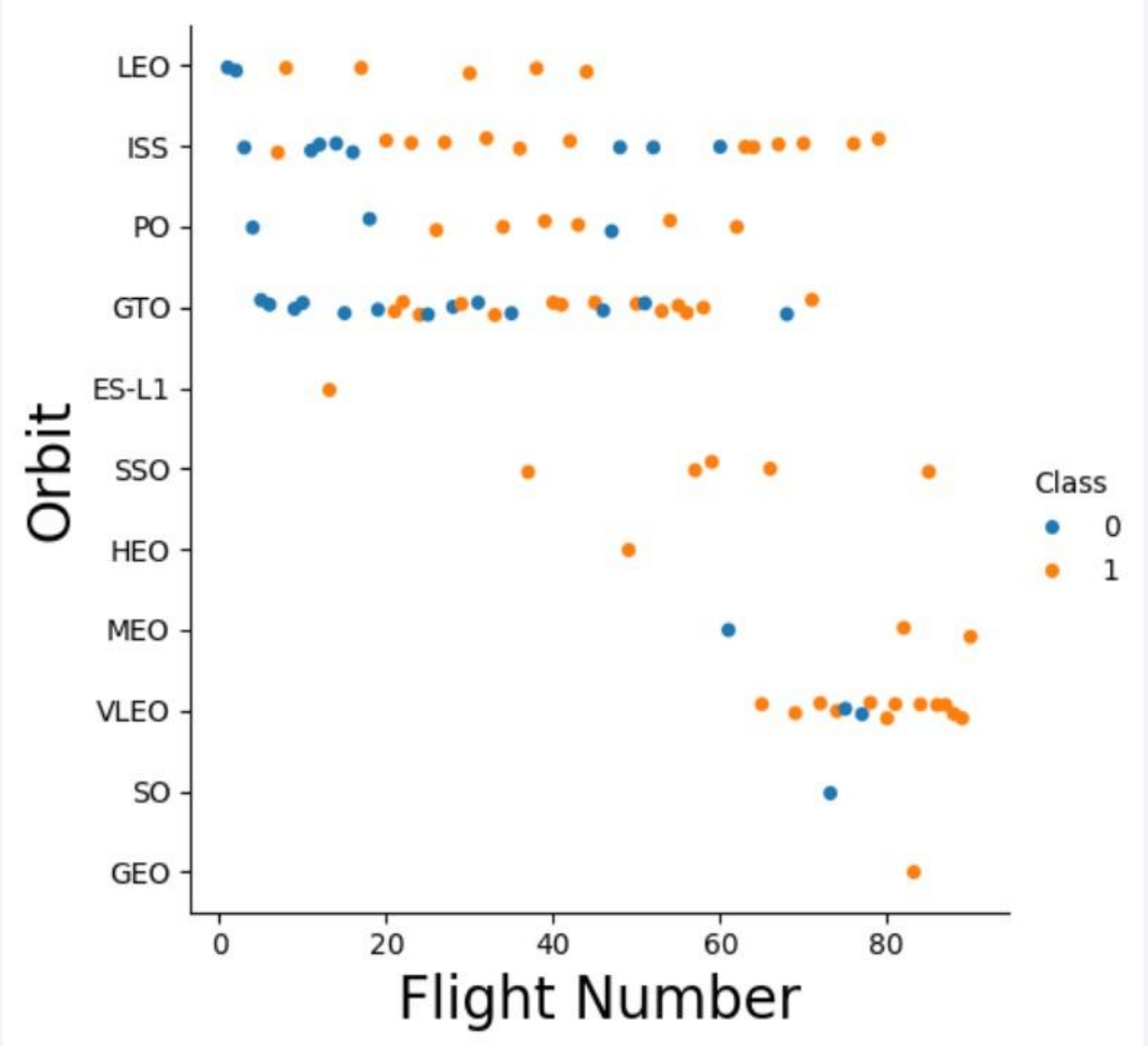
Success Rate vs. Orbit Type

- The bar chart shows that the orbits with the highest success rate are ES-L1, GEO, HEO, & SSO



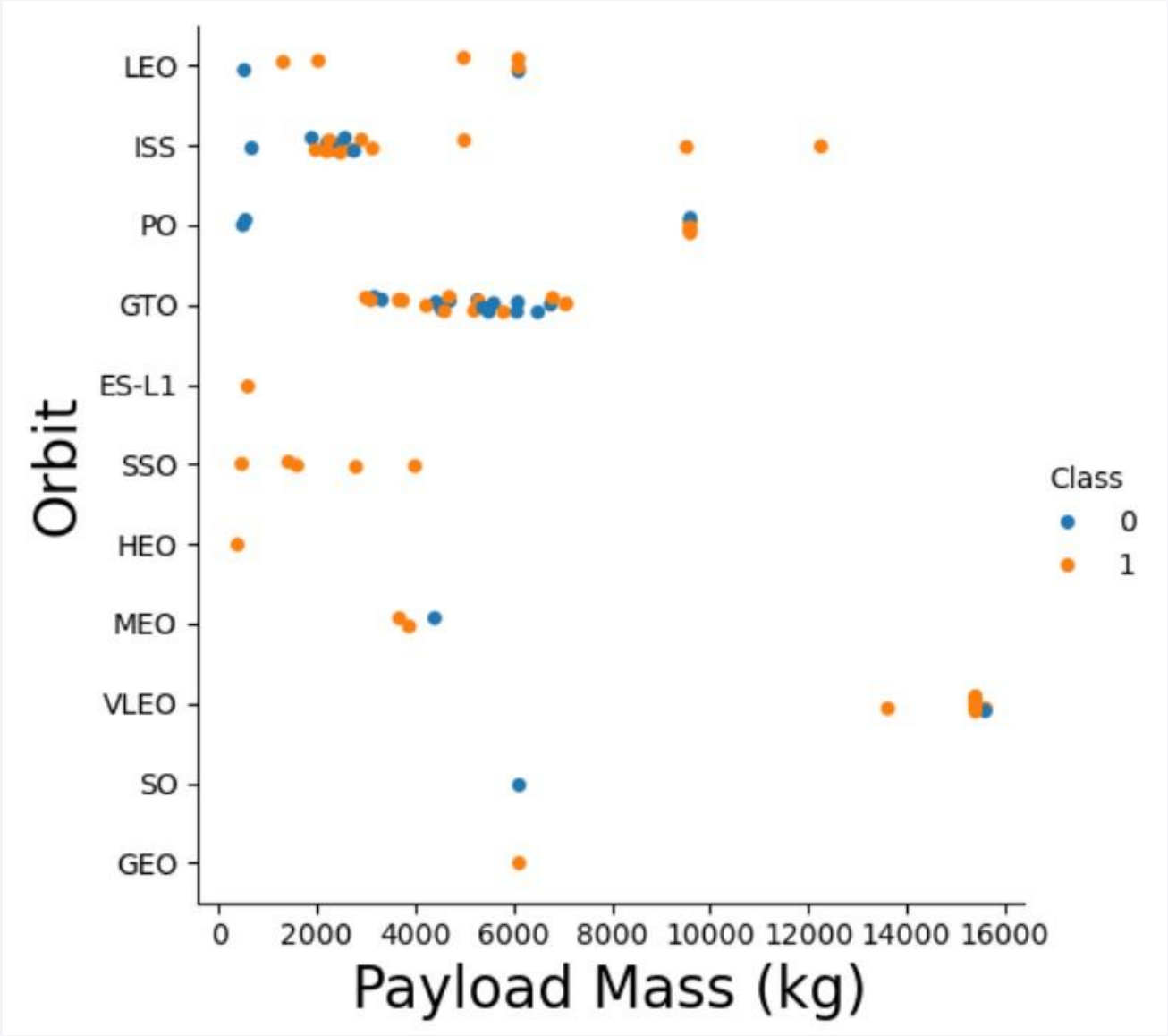
Flight Number vs. Orbit Type

- The LEO & PO orbit's success seems to be related to the number of flights
- The GTO orbit appears to have no relationship between flight number and success
- The , ES-L1, HEO, SO, & GEO orbit's lack data to decide success rate
- The SSO orbit is the only to orbits with a 100% success rate



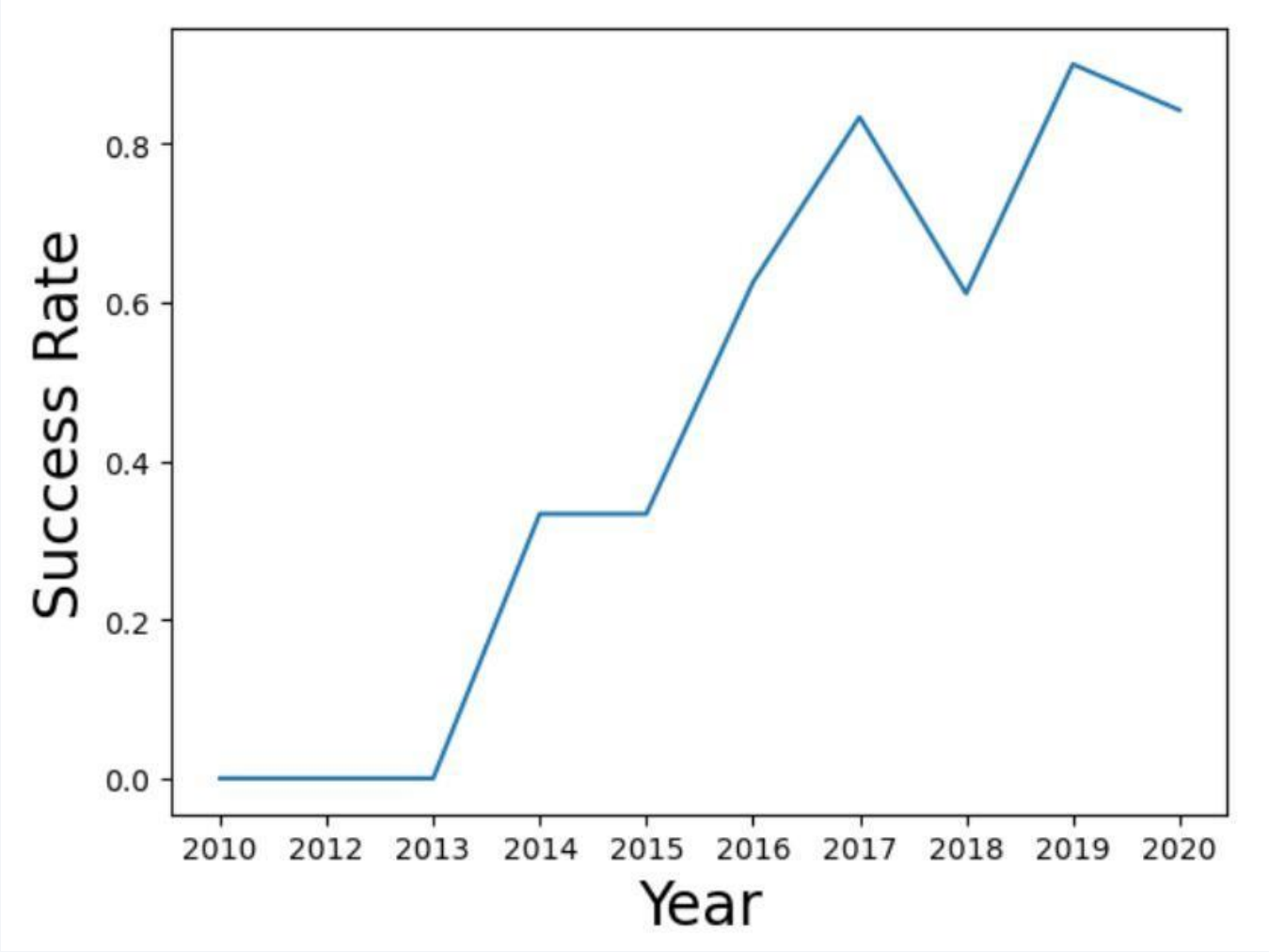
Payload vs. Orbit Type

- The ISS, PO, & VLEO orbits are the only orbits with success at payloads >~9K Kg
- The LEO, ISS, SSO, & MEO orbits demonstrate greater success rate at payloads <~8K Kg
- The GTO orbit again does not show a correlation in success when comparing orbit to payload mass



Launch Success Yearly Trend

- We can see that years 2010 to 2013 may represent the very early stages of development and testing with a 0% success rate
- From 2013 to 2020 a constant increase is observed with a notable regression between 2018 & 2019
- Year 2020 seems to depict a leveling in success rate, but further data is required to make a determination



All Launch Site Names

- The database was queried using SQL to extract the launch sites for ease of viewing

Display the names of the unique launch sites in the space mission

```
task1_query = "SELECT DISTINCT Launch_Site FROM SPACEXTABLE"
launch_sites = pd.read_sql(task1_query, con)
launch_sites
```

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
task2_query = "SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5"
cca_records = pd.read_sql(task2_query, con)
cca_records
```

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Above we have a depiction of a SQL query to find the 1st five records with ‘CCA’ in the launch site column

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
task3_query = "SELECT SUM(PAYLOAD_MASS__KG_) as TotalPayloadMass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%'"
nasa_crs_payload = pd.read_sql(task3_query, con)
nasa_crs_payload
```

	TotalPayloadMass
0	48213

- Above we see a SQL query to calculate the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
task4_query = "SELECT AVG(PAYLOAD_MASS__KG_) as AveragePayloadMass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'"
f9_v1_1_payload = pd.read_sql(task4_query, con)
f9_v1_1_payload
```

	AveragePayloadMass
0	2928.4

- Above we see a SQL query to calculate the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
task5_query = """
SELECT MIN(Date) as FirstSuccessfulLanding
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (ground pad)'
"""
first_successful_landing = pd.read_sql(task5_query, con)
first_successful_landing
```

	FirstSuccessfulLanding
0	2015-12-22

- Above we see a SQL query to find the date of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
task6_query = """
SELECT Booster_Version
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000
"""
successful_drone_ship_boosters = pd.read_sql(task6_query, con)
successful_drone_ship_boosters
```

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

- Above we see a SQL query to list the names of boosters which have successfully landed on drone ships and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

- Here we see a SQL query to calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes ¶

```
task7_query = """
SELECT
    CASE
        WHEN "Mission_Outcome" LIKE 'Success%' THEN 'Success'
        ELSE 'Failure'
    END as Outcome,
    COUNT(*) as Count
FROM SPACEXTABLE
GROUP BY Outcome
"""
mission_outcomes = pd.read_sql(task7_query, con)
mission_outcomes
```

	Outcome	Count
0	Failure	1
1	Success	100

Boosters Carried Maximum Payload

- Here we see a SQL query to list the names of the booster which have carried the maximum payload mass

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
task8_query = """
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
"""
max_payload_boosters = pd.read_sql(task8_query, con)
max_payload_boosters
```

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
task9_query = """
SELECT
    substr(Date, 6, 2) as Month,
    "Landing_Outcome",
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE 'Failure (drone ship)'
AND substr(Date, 0, 5) = '2015'
"""
failure_landing_2015 = pd.read_sql(task9_query, con)
failure_landing_2015
```

	Month	Landing_Outcome	Booster_Version	Launch_Site
0	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Above we see a SQL query to list the failed landing outcomes in drone ships, their booster versions, and launch site names for the year 2015

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.

```
task10_query = """
SELECT
    "Landing_Outcome",
    COUNT(*) as Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY Count DESC
"""
ranked_landing_outcomes = pd.read_sql(task10_query, con)
ranked_landing_outcomes
```

	Landing_Outcome	Count
0	No attempt	10
1	Success (drone ship)	5
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	2
7	Precluded (drone ship)	1

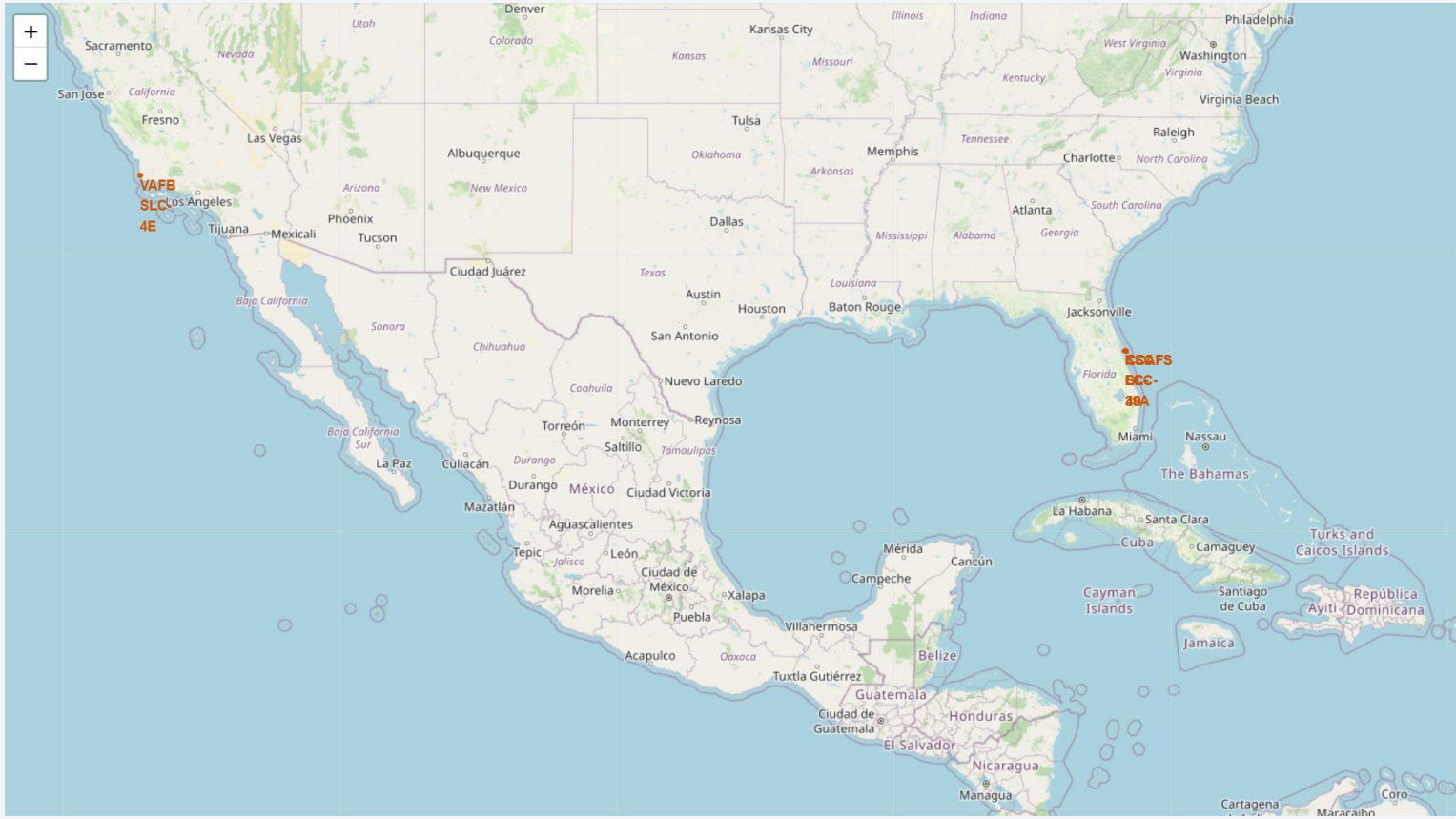
- Above we see a SQL query to rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the dates of 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

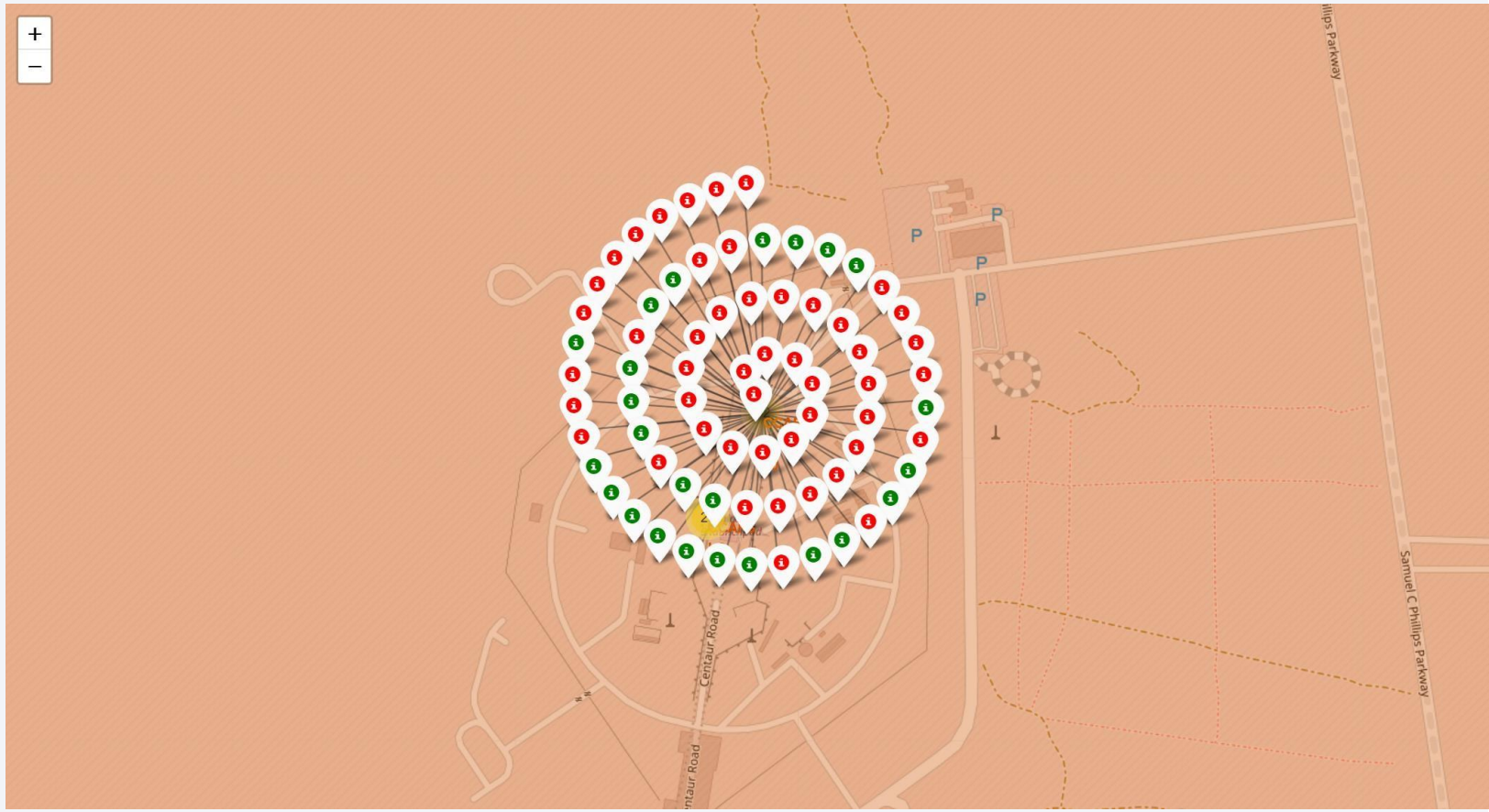
Interactive Global Launch Site Map

- This interactive map was created with Folium
- Utilizing the map, we can easily identify the launch site from a global scale
- Here we can see that the launch site are primarily located adjacent to coastlines for the U.S.



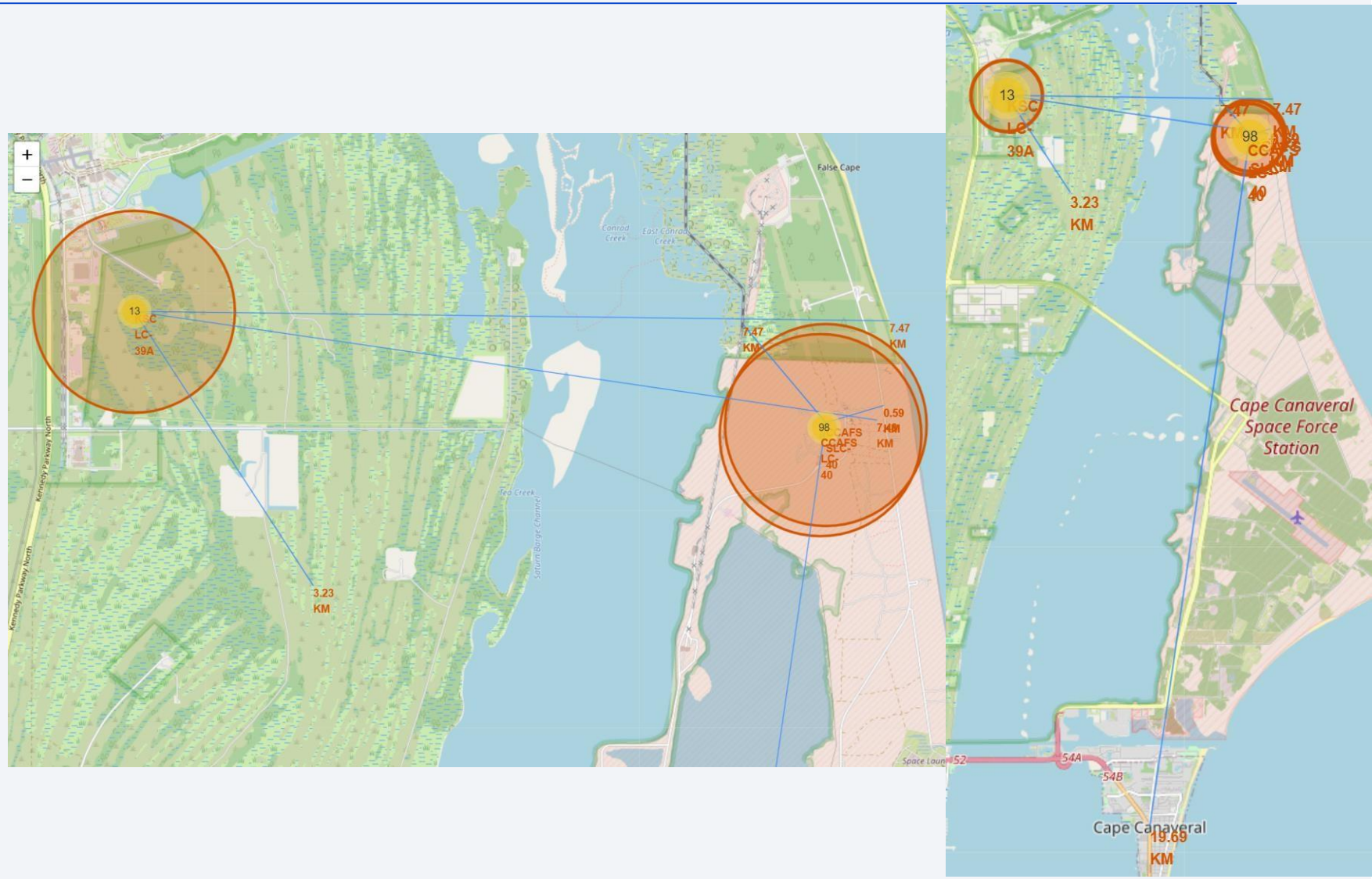
Interactive Global Launch Site Map – Color Labels

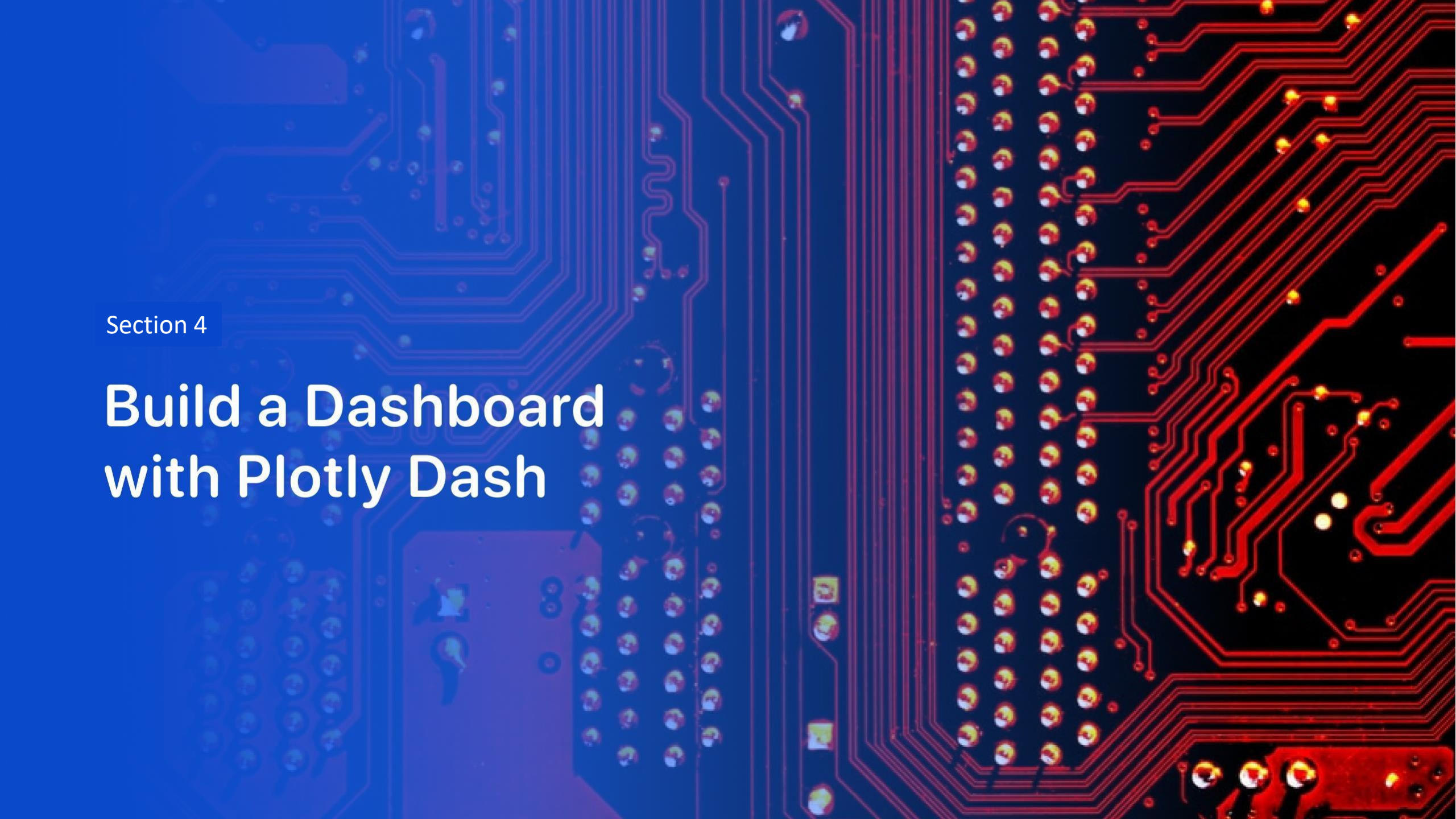
- Here we have a zoomed in view of launch site to illustrate the addition of color labels
- A simple green (success) and red (failure) scheme was applied to launches
- Viewers can quickly discern a result from the data via the visualization i.e., this site had more failures than successes



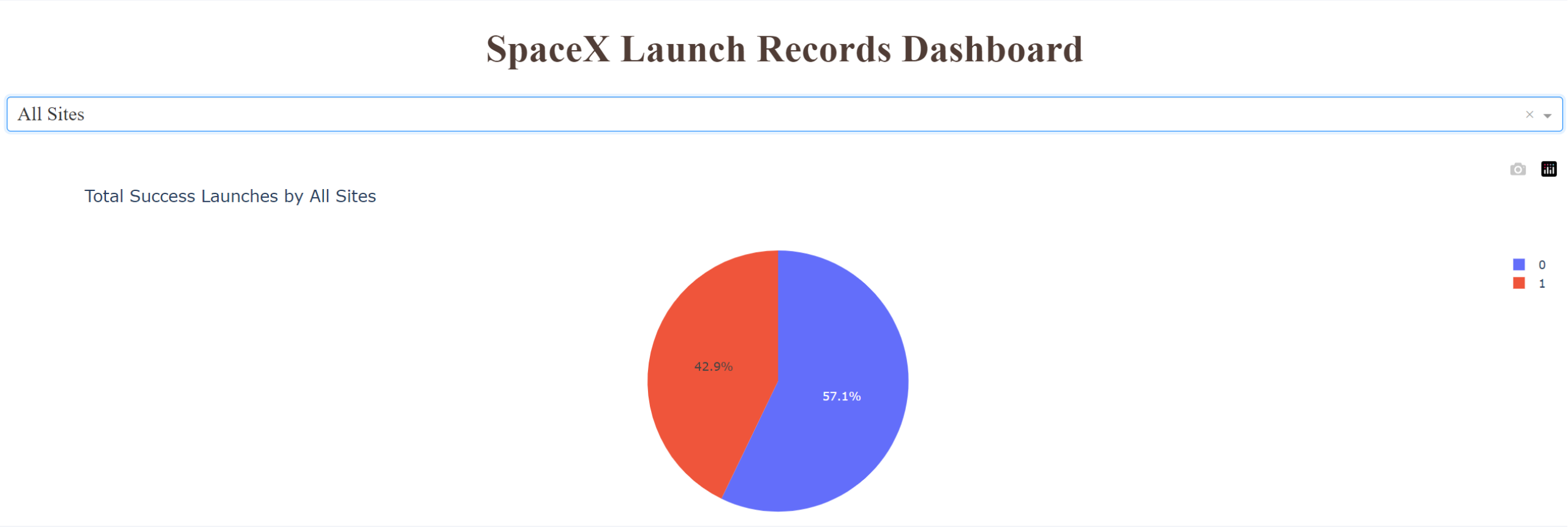
Interactive Global Launch Site Map – Distance Markers

- Here we see the addition of distance markers and trace lines to the map
- This further enables the viewer to explore the data in a new way through the addition of this visualization
- The map can now be used to explore other levels of correlation presented by the data



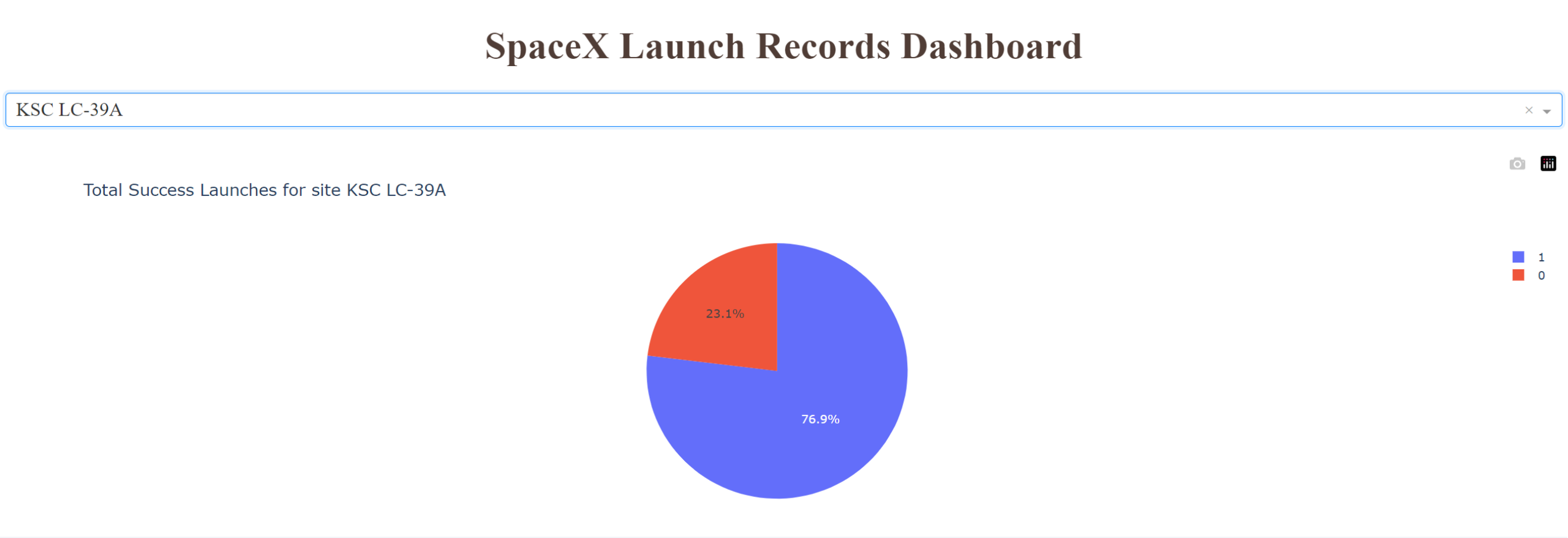

Section 4
Build a Dashboard with Plotly Dash

Interactive Dashboard - Pie Chart



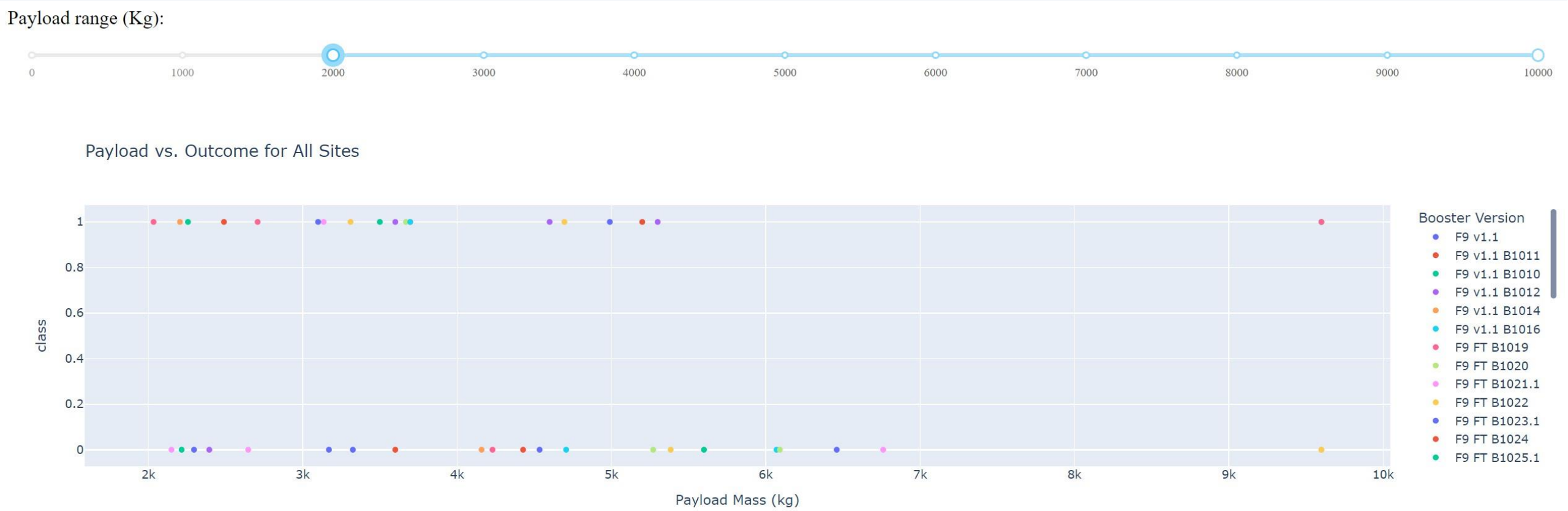
- The above image shows and interactive dashboard which provides a high-level view of the data to the viewer
- Here we can see the calculated success & failure % of all launch sites via pie chart

Interactive Dashboard - Dropdown Menu



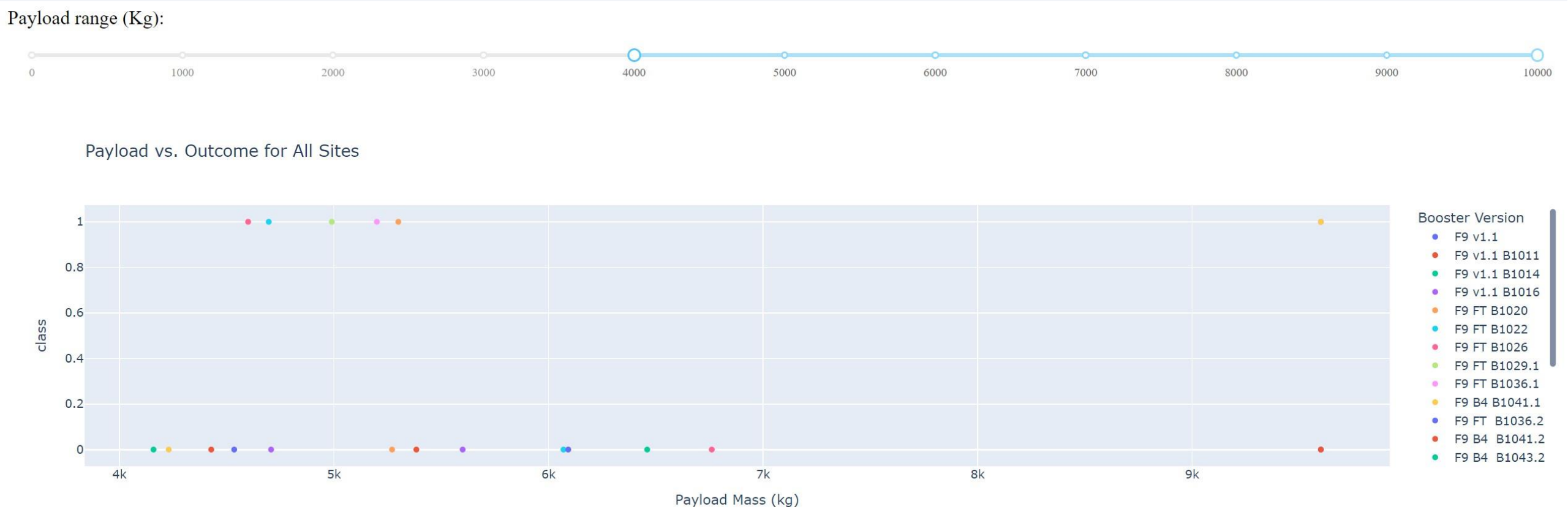
- Here we can see the calculated success & failure % of a specific launch site using the dropdown menu option for KSC LC-39A
- KSC LC had the highest success rate of 76.9%

Interactive Dashboard – Range Slider & Scatter Plot



- Shown here, is a payload range slide that is synced with the payload vs. outcome scatter plot to again enable quick interpretation of the data
- We can see that booster v1.1 had success in the range of 2K-6K Kg as an example

Interactive Dashboard – Range Slider & Scatter Plot



- Additional example depicting how changes to the slider bar affect the scatter plot

Interactive Dashboard - Range Slider & Scatter Plot

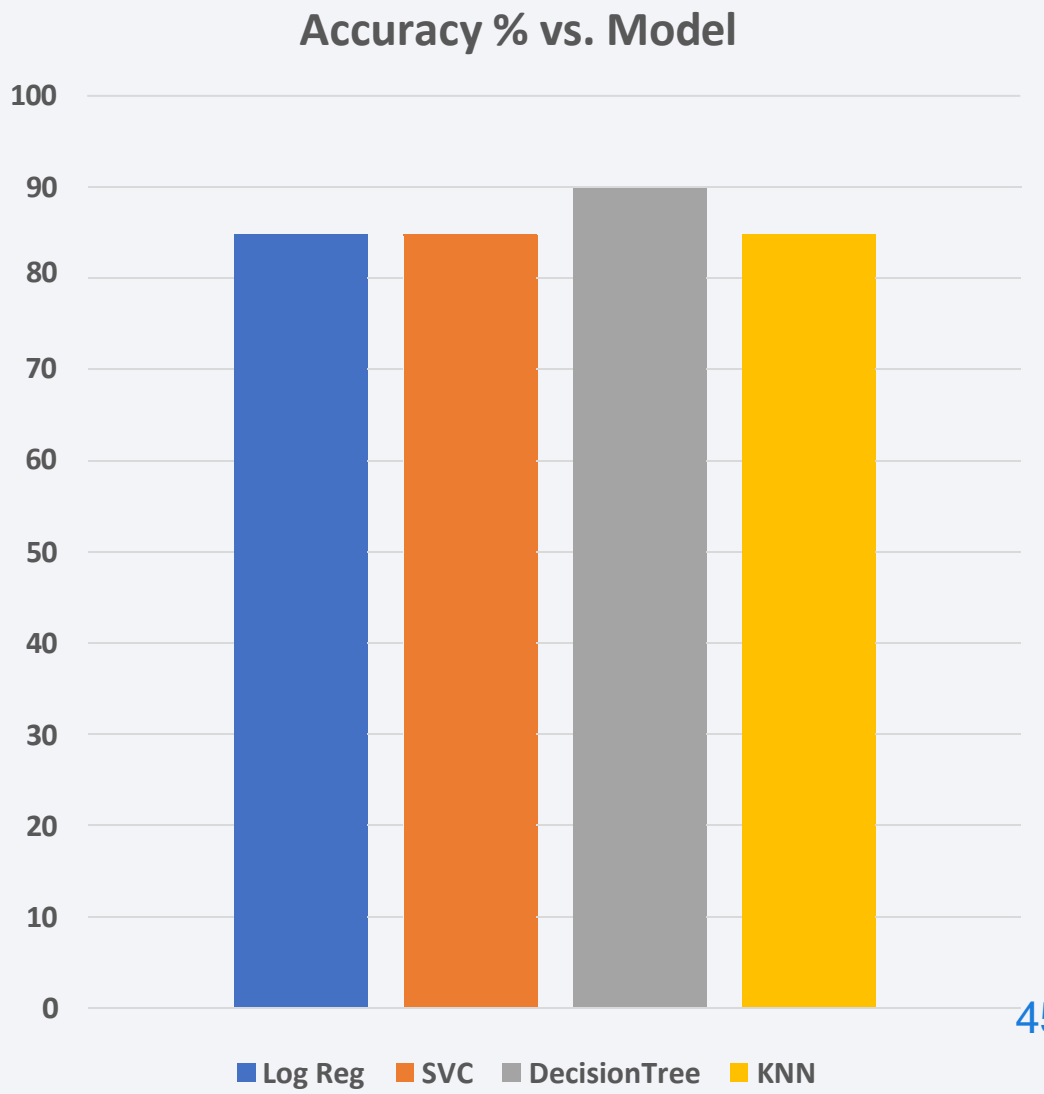
- Additional example depicting how changes to the slider bar affect the scatter plot

Section 5

Predictive Analysis (Classification)

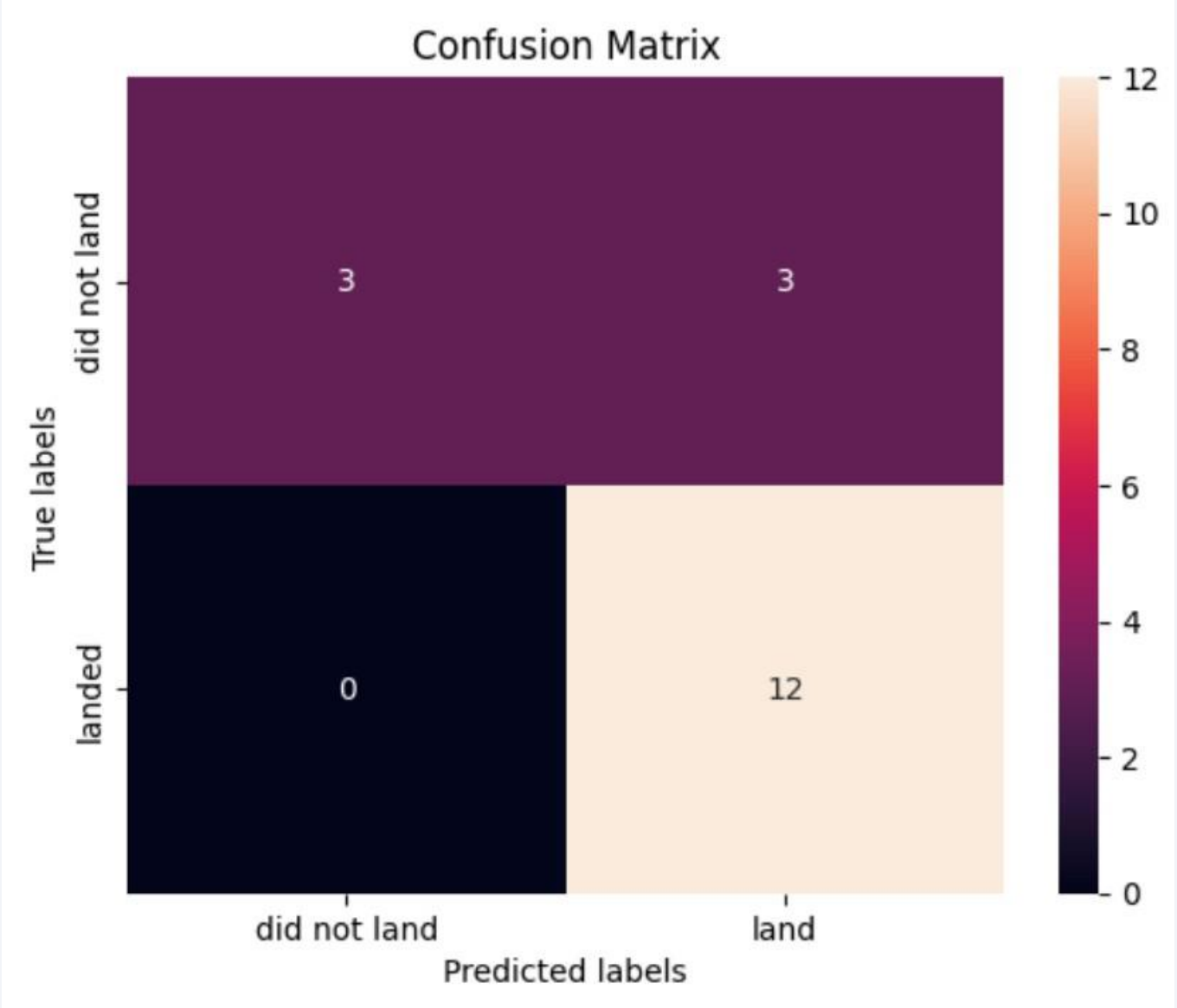
Classification Accuracy

- Through the use of classification accuracy, we can see that the Decision Tree model had the highest accuracy at 0.90



Confusion Matrix

- The confusion matrix for the decision tree model shows that 12 successful launches were predicted correctly
- Additionally, 3 false positive and 3 false negative were predicted by the model



Conclusions

- This dataset uncovered fascinating insights; the payload can influence the landing result of the launch based on the booster version used
- A higher flight number, booster v1.1, and launch site KSC LC-39A are key factors for the successful launch of Falcon9’s first stage. The orbit and payload mass have varying impacts on launch success depending on other factors
- KSC LC-39A boasts the highest number of successful launches, potentially affected by orbit, boosters, and payload mass
- Booster v1.1 shows numerous successful outcomes across a broad payload range
- The decision tree model is the best machine learning model to use due to its high accuracy of 0.90

Appendix

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

- Example of the dataframe used in the exploration and development of this slide deck report

Thank you!