IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rohit Raghu Kumar
06-05-2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a dashboard with Plotly Dash
- Predictive analysis using machine learning (classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics
- Predictive analysis

Introduction

Project background and context

SpaceX, a successful company of the commercial space age advertises Falcon9 rocket launches on its website with a cost of 62 million dollars the other providers charge upwards of 165 million dollars each. Much of the savings is because unlike other providers, SpaceX's Falcon9 can reuse the 1st stage.

Problems you want to find answers

If we can determine whether the first stage will land, then we can determine the cost of a launch. We will train a machine learning model using the data from previous launches of the Falcon9 rocket which is publicly available to predict if the first stage of the SpaceX will land or not.

Section 1

Methodology

Methodology

Executive Summary

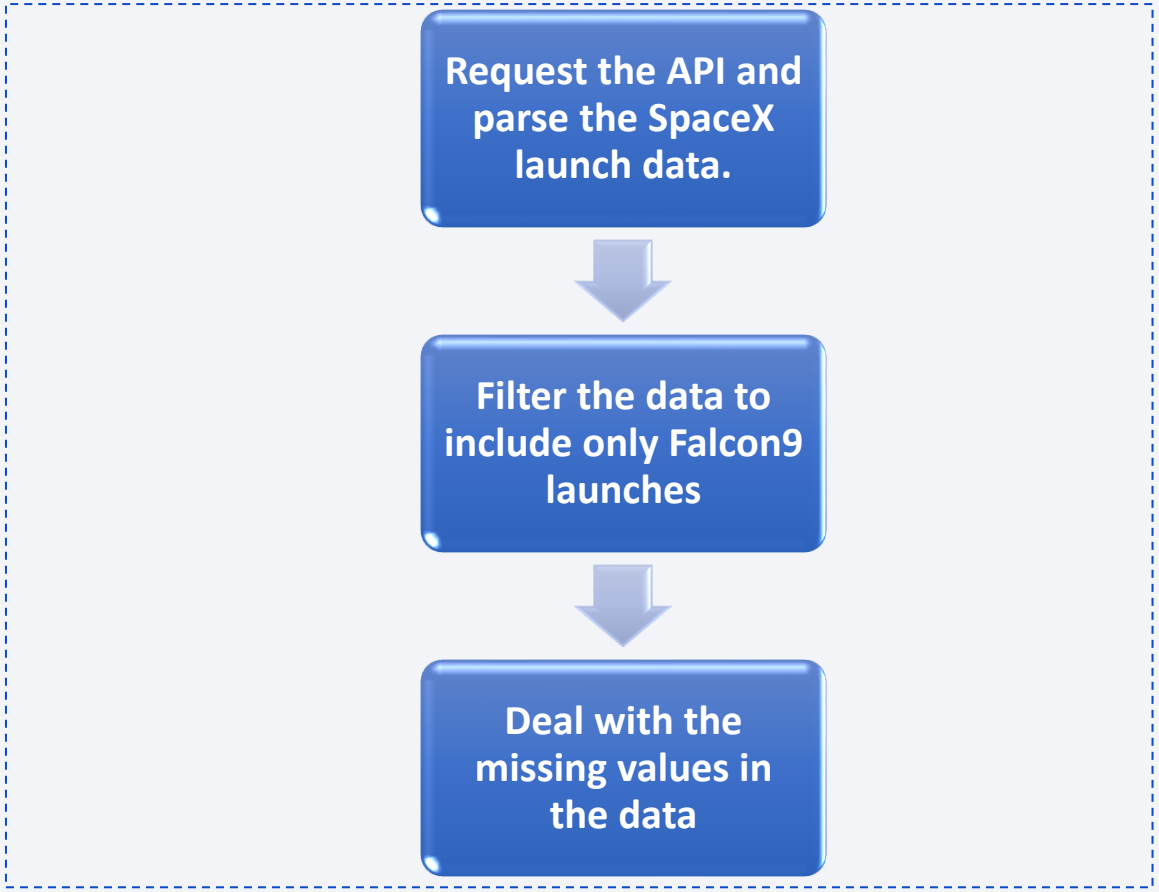
- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources:
    - SpaceX API
    - Web Scraping from Wikipedia
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided into training and testing data sets and evaluated by four different combinations of parameters.

Data Collection

- The data sets were collected from the following two sources using web scraping techniques.
    - SpaceX API (https://api.spacexdata.com/v4/rockets)
    - Wikipedia(https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

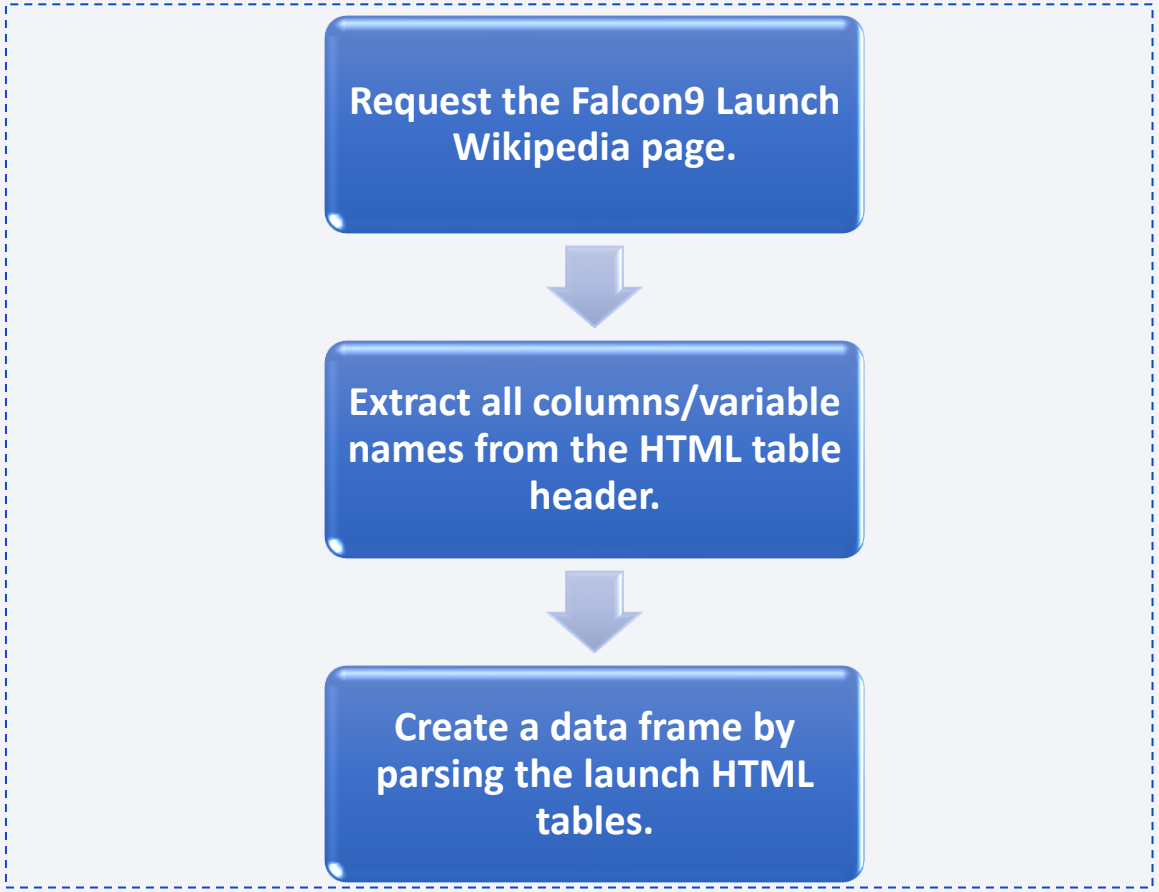
- SpaceX offers a public API from where data can be obtained and then used.
- This API was used according to the flowchart beside and then data is persisted.



- Source Code: https://github.com/Rohitraghu11/Applied_Data_Science_Capstone/blob/5a659b29b2e2c6fd4ce6f338b6bccefc0da2fb03/jupyter-labs-spacex-data-collection-api-%5BPart-1%5D.ipynb

Data Collection - Scraping

- Data from past SpaceX launches can be obtained from Wikipedia.
- The data is downloaded from Wikipedia in accordance with the flowchart and the records are stored in a HTML table.
- Source Code: https://github.com/Rohitraghu11/Applied_Data_Science_Capstone/blob/5a659b29b2e2c6fd4ce6f338b6bccefc0da2fb03/jupyter-labs-webscraping-%5BPart-2%5D.ipynb



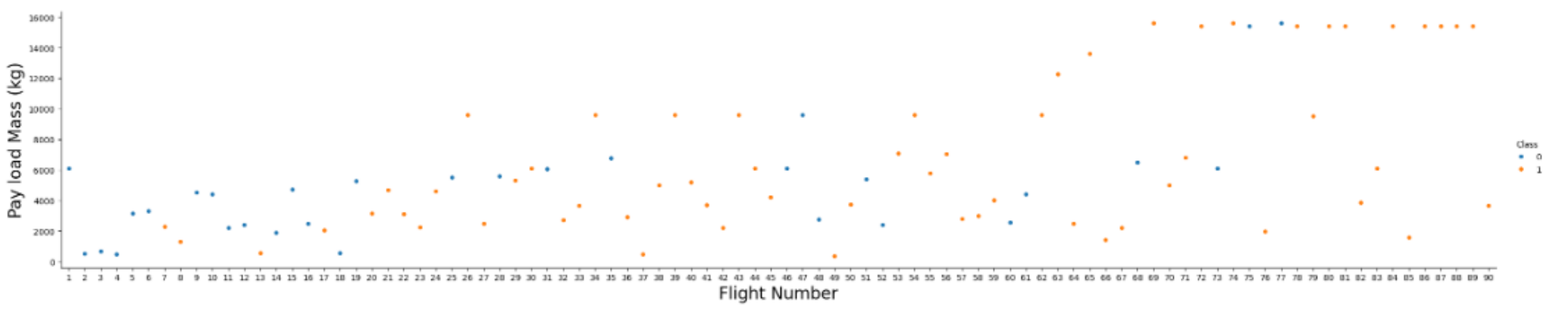
Data Wrangling

- We performed some Exploratory Data Analysis on the data set.
- Then the summaries of launches per site, occurrences of each orbit and occurrences of mission outcome per orbit were calculated.
- Finally, the landing outcome was created from Outcome column.

- Source Code: https://github.com/Rohitraghu11/Applied_Data_Science_Capstone/blob/5a659b29b2e2c6fd4ce6f338b6bccefc0da2fb03/labs-jupyter-spacex-Data%20wrangling-%5BPart-3%5D.ipynb

EDA with Data Visualization

- To explore the data, scatterplots and bar plots were used to visualize the relationship between the pairs of features.
    - Flight Number vs Launch Site, Pay Load Mass vs Launch Site, Orbit type vs Success Rate, Flight Number vs Orbit type, Pay Load Mass vs Orbit type, Year vs Success rate.



- Source Code: https://github.com/Rohitraghu11/Applied_Data_Science_Capstone/blob/5a659b29b2e2c6fd4ce6f338b6bccefc0da2fb03/edadataviz-%5BPart-5%5D.ipynb

EDA with SQL

- The following SQL queries were performed:
  - The names of unique launch sites in the space mission.
  - Display five 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 successful landing outcome in ground pad was achieved.
  - List the names of 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 booster versions which have carried the maximum payload mass using a subquery.
  - List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch sites for months in the 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.
- Source Code: https://github.com/Rohitraghu11/Applied_Data_Science_Capstone/blob/5a659b29b2e2c6fd4ce6f338b6bccefc0da2fb03/jupyter-labs-eda-sql-coursera_sqllite-%5BPart-4%5D.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used, and an interactive map was created using Folium maps.
    - Markers indicate points like launch sites.
    - Circles indicate the area around specific coordinates, for example NASA Johnson Space Center.
    - Marker clusters indicate group of events in each coordinate for example launches in a launch site.
    - Lines are used to indicate the distances between two coordinates.

- Source Code: https://github.com/Rohitraghu11/Applied_Data_Science_Capstone/blob/5a659b29b2e2c6fd4ce6f338b6bccefc0da2fb03/lab_jupyter_launch_site_location-%5BPart-6%5D.ipynb

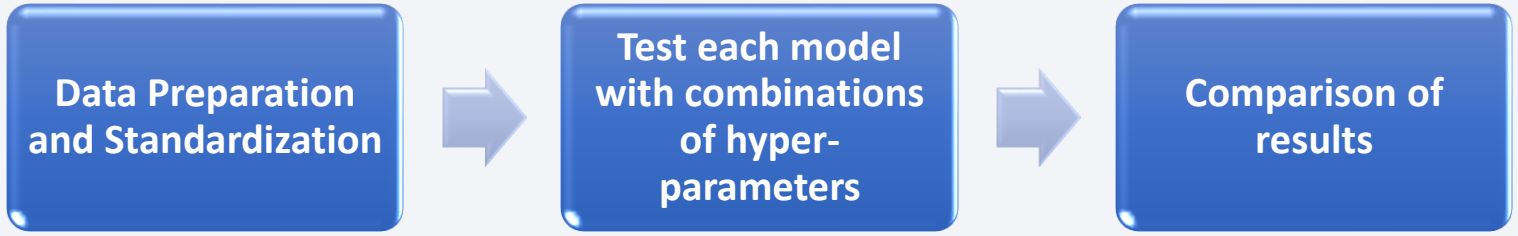
Build a Dashboard with Plotly Dash

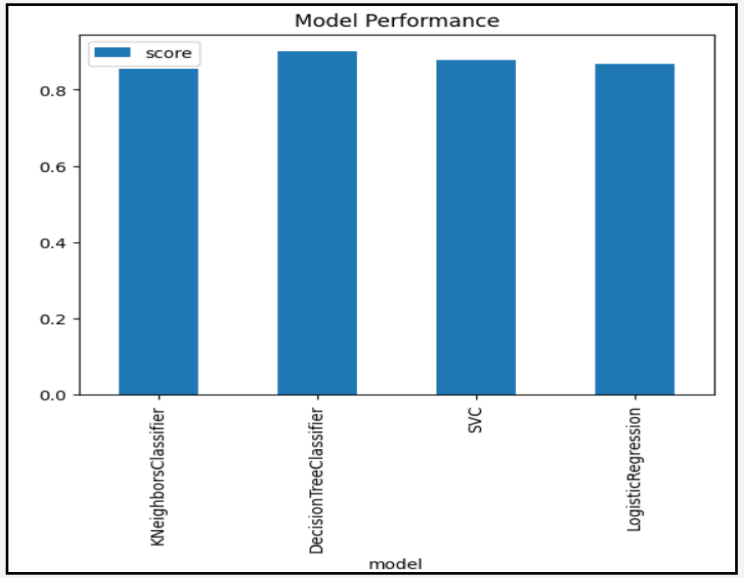
- The following plots and graphs were used to visualize the data.
  - Percentage of launches by site.
  - Payload range.
- This combination enabled quick analysis of the relation between payload and launch sites, which helped to identify the best place to launch according to payloads.

- Source Code: https://github.com/Rohitraghu11/Applied_Data_Science_Capstone/blob/5a659b29b2e2c6fd4ce6f338b6bccefc0da2fb03/spacex_dash_app-%5BPart-7%5D.py

Predictive Analysis (Classification)

- Four classification models namely, Logistic regression, K Nearest Neighbours, Support Vector Machine and Decision Tree model were built and compared.



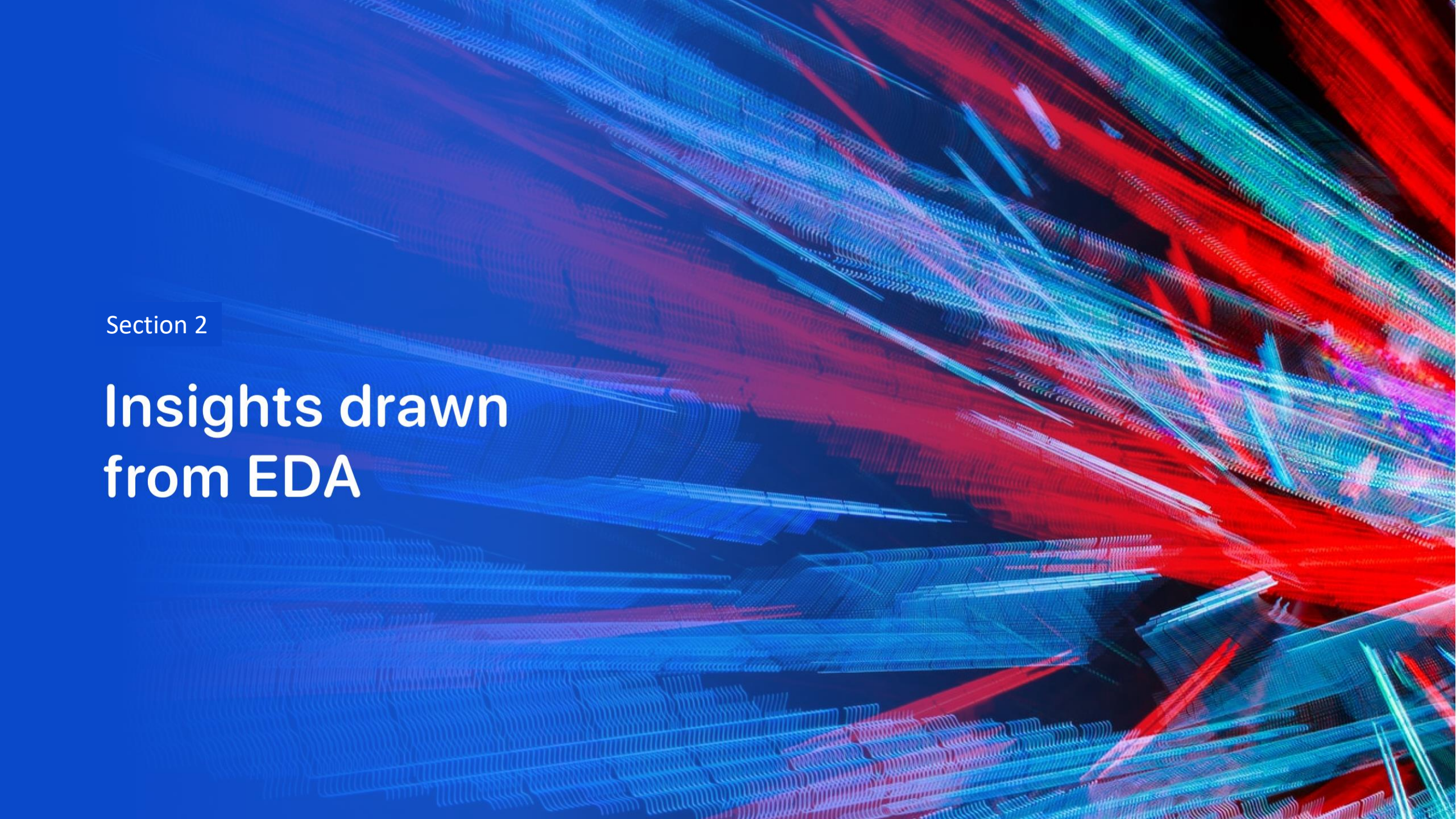


- Source Code: https://github.com/Rohitraghu11/Applied_Data_Science_Capstone/blob/5a659b29b2e2c6fd4ce6f338b6bccefc0da2fb03/SpaceX_Machine%20Learning%20Prediction_Part_5-%5BPart-8%5D.ipynb

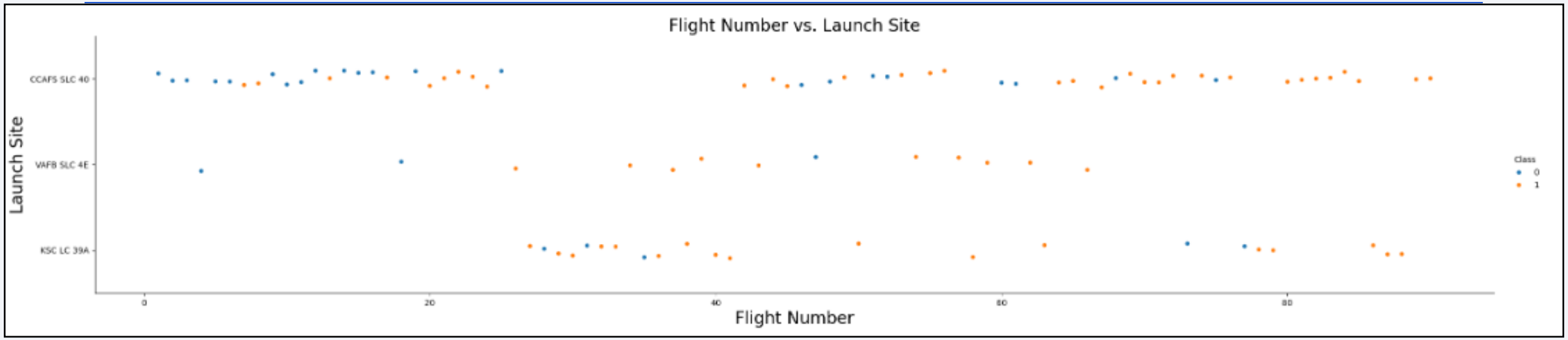
Results

- Exploratory data analysis results
  - The first successful landing outcome happened in December 2015 five years after the first launch.
  - The success rate of SpaceX launches has increased over time.
  - Low weighted payloads perform better than high weighted payloads.
- Interactive analytics demo in screenshots
  - Location of launch appears to be a key contributing factor to the success of missions.
  - It was possible to identify that launch sites are in safe places far away from suburban areas like near the sea and they have good logistical infrastructure around it.
  - Most launches happen at the launch sites in the East coast.
- Predictive analysis results
  - Decision tree model is the best in terms of prediction accuracy for this dataset.

For more information refer to slides from 18 to 45.

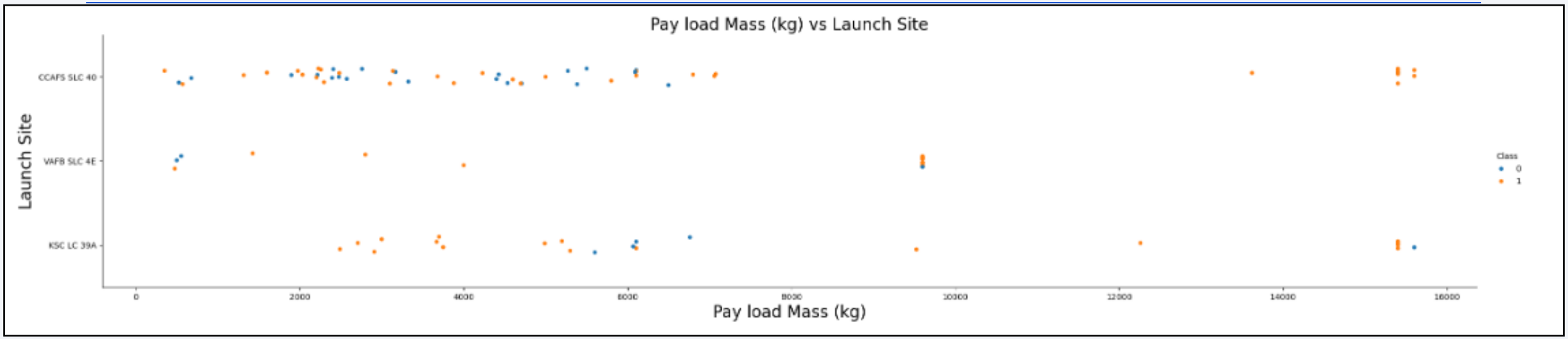

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the plot the best launch site nowadays is CCAF5 SLC 40 where most recent launches have been successful.
- Most launches took off from CCAF5 SLC 40 launch site.
- It is also seen that the general success rate has improved over time.

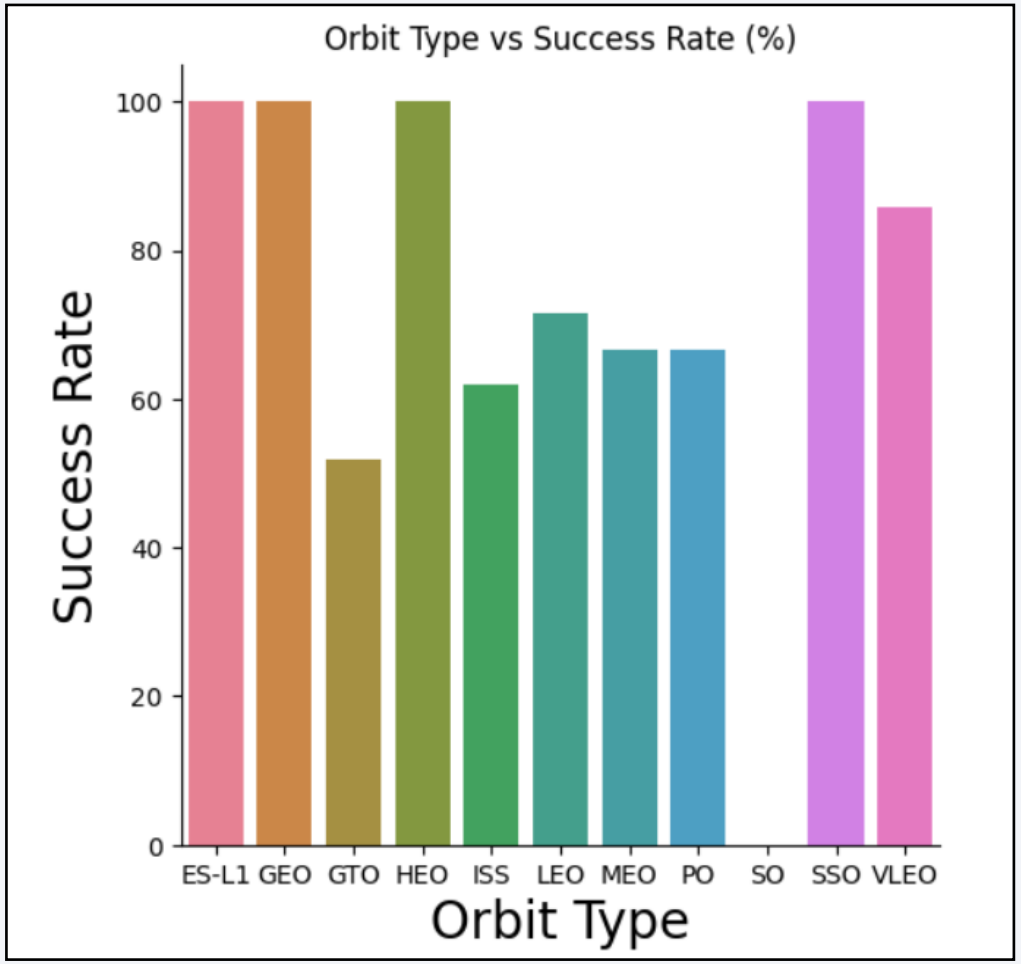
Payload vs. Launch Site



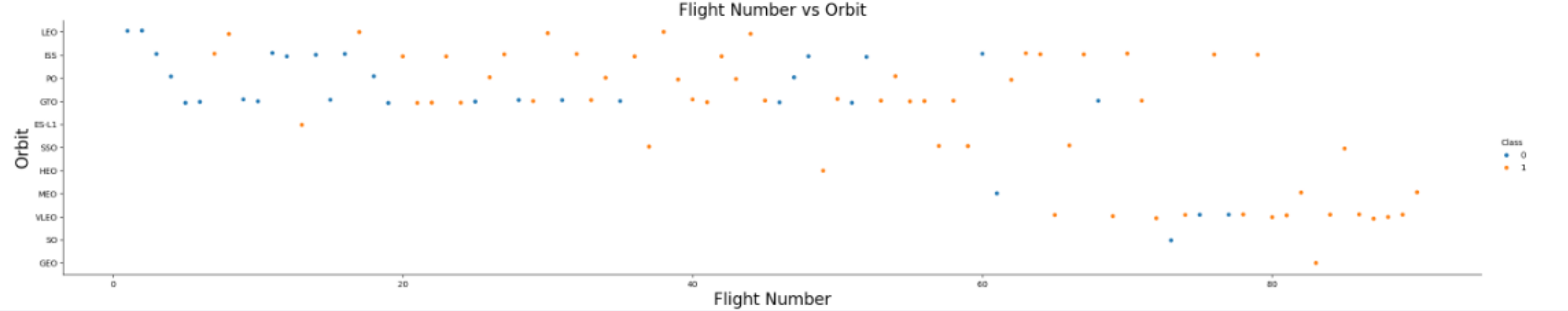
- VAFB SLC 4E launch site conducts launches with lower payloads.
- CCAF5 SLC 40 hosts a higher number of landings involving higher and lower payloads.
- Most payloads over 9000 kg have achieved successful outcomes.
- Payloads over 12000 kg seems to be possible only on CCAF5 SLC 40 AND KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbits achieved the highest success rate at 100%.
- These are followed by VLEO with a success rate >80% and LEO with a success rate >70%.

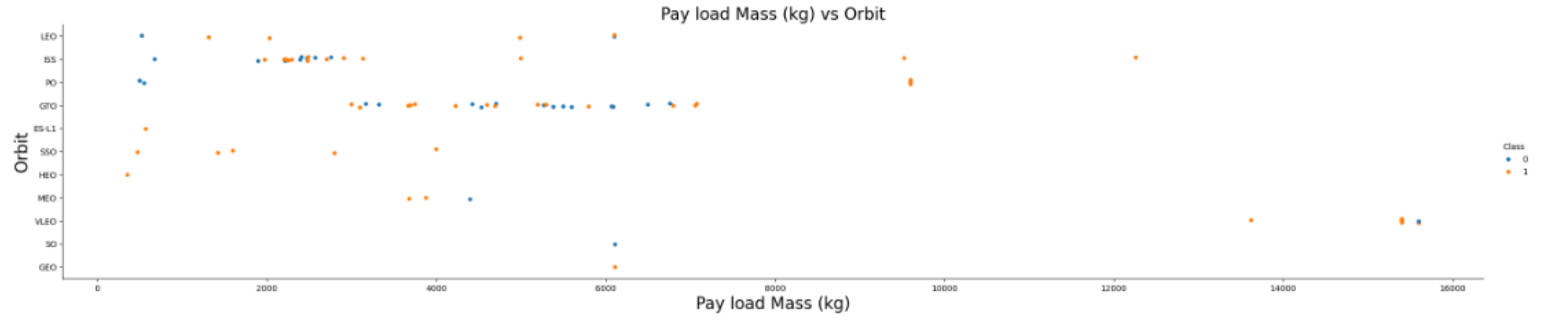


Flight Number vs. Orbit Type



- In recent years there has been a transition towards launching missions into Very Low Earth Orbits (VLEO) with a significantly high rate of success.
- While the GTO orbit experiences a low success rate, there appears to be no discernible relationship between the flight number and rate of success in this orbit.
- The success rate has improved over time for all orbits.

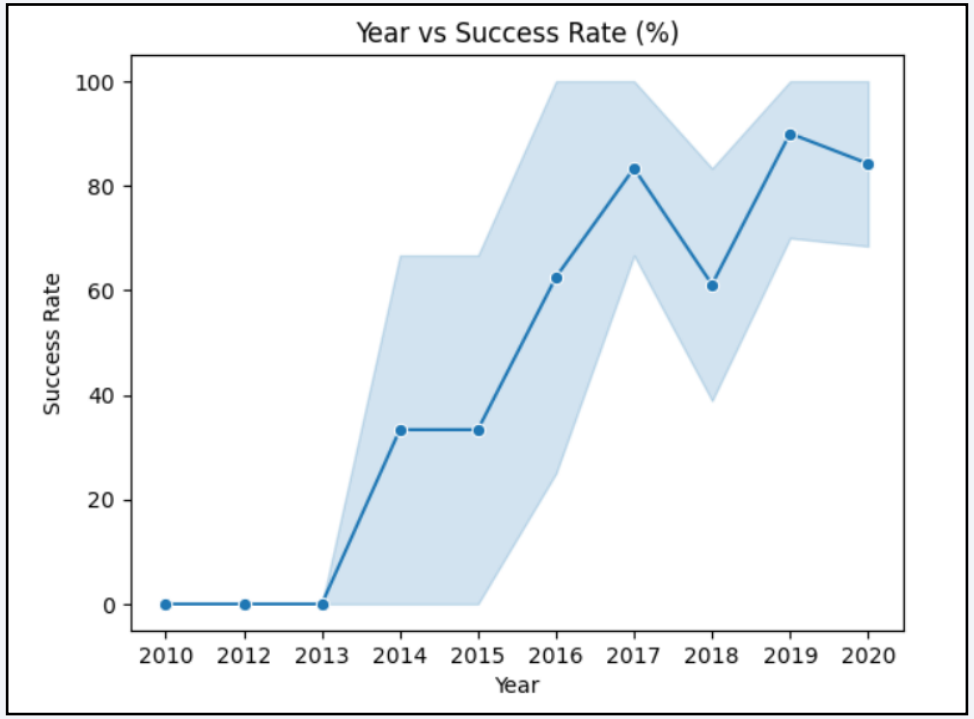
Payload vs. Orbit Type



- For heavy payloads the rate of success is higher for VLO and ISS orbits.
- In the case of GTO there is no apparent relationship between payload and rate of success.
- ISS orbit has the widest range of payload and a good rate of success.
- There are a few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- The rate of success has seen a notable rise since 2013 and continued to rise until 2020, possibly attributed to technological advancements.
- The first three years (2010 - 2013) seem to have a phase focused on adjustment, fine tuning and technological enhancement.



All Launch Site Names

- The names of the four launch sites in the mission are as follows:

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

- The associated query:

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

- This is obtained by selecting the unique occurrences of ‘Launch_Site’ values from the data set.

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`:

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

- The associated query:

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

- Here we can see five records of launch site names beginning with ‘CCA’.

Total Payload Mass

- The total payload (kg) carried by boosters from NASA (CRS) is:

Total_Payload_Mass
45596

- The associated query:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

- The total payload mass is calculated by summing all payloads carried by boosters from NASA (CRS).

Average Payload Mass by F9 v1.1

- The average payload mass (kg) carried by booster version F9 v1.1:

Average_Payload_Mass
2928.4

- The associated query:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

- We have obtained this value by filtering the data by booster version above and calculating the average payload mass.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad:

First_Successful_Landing_Ground
2015-12-22

- The associated query:

```
%sql SELECT MIN(DATE) AS First_Successful_Landing_Ground FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

- The date of first successful landing on ground pad is obtained by filtering data by successful landing outcome on ground pad and getting the minimum value for date.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

- The associated query:

```
%%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000;
```

- By selecting the booster versions according to the filter mentioned above we get these 4 versions as the result.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes:

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

- The associated query:

```
%%sql SELECT Mission_Outcome, COUNT(*) AS Total_Count
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

- By grouping mission outcomes and counting the records for each group will give the desired result.

Boosters Carried Maximum Payload

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

- The associated query:

```
%%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

- These are the boosters that have carried maximum payload mass recorded in the data set obtained by using a subquery.

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are:

Month	Failure_Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The associated query:

```
%%sql SELECT
 CASE
        WHEN Date like '%-01-%' THEN 'January'
        WHEN Date like '%-02-%' THEN 'February'
        WHEN Date like '%-03-%' THEN 'March'
        WHEN Date like '%-04-%' THEN 'April'
        WHEN Date like '%-05-%' THEN 'May'
        WHEN Date like '%-06-%' THEN 'June'
        WHEN Date like '%-07-%' THEN 'July'
        WHEN Date like '%-08-%' THEN 'August'
        WHEN Date like '%-09-%' THEN 'September'
        WHEN Date like '%-10-%' THEN 'October'
        WHEN Date like '%-11-%' THEN 'November'
        WHEN Date like '%-12-%' THEN 'December'
    END AS Month,
    Landing_Outcome AS Failure_Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE Date like'%2015%' AND Landing_Outcome LIKE 'Failure (drone ship)';
```

- There are only two occurrences in the list.

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

- 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:
- As shown the highest count is attributed to 'No attempt' (10 times), followed by 'Success (drone ship)' and 'Failure (drone ship)' (5 times each) followed by 'Controlled' which occurs 3 times
- The associated query:

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

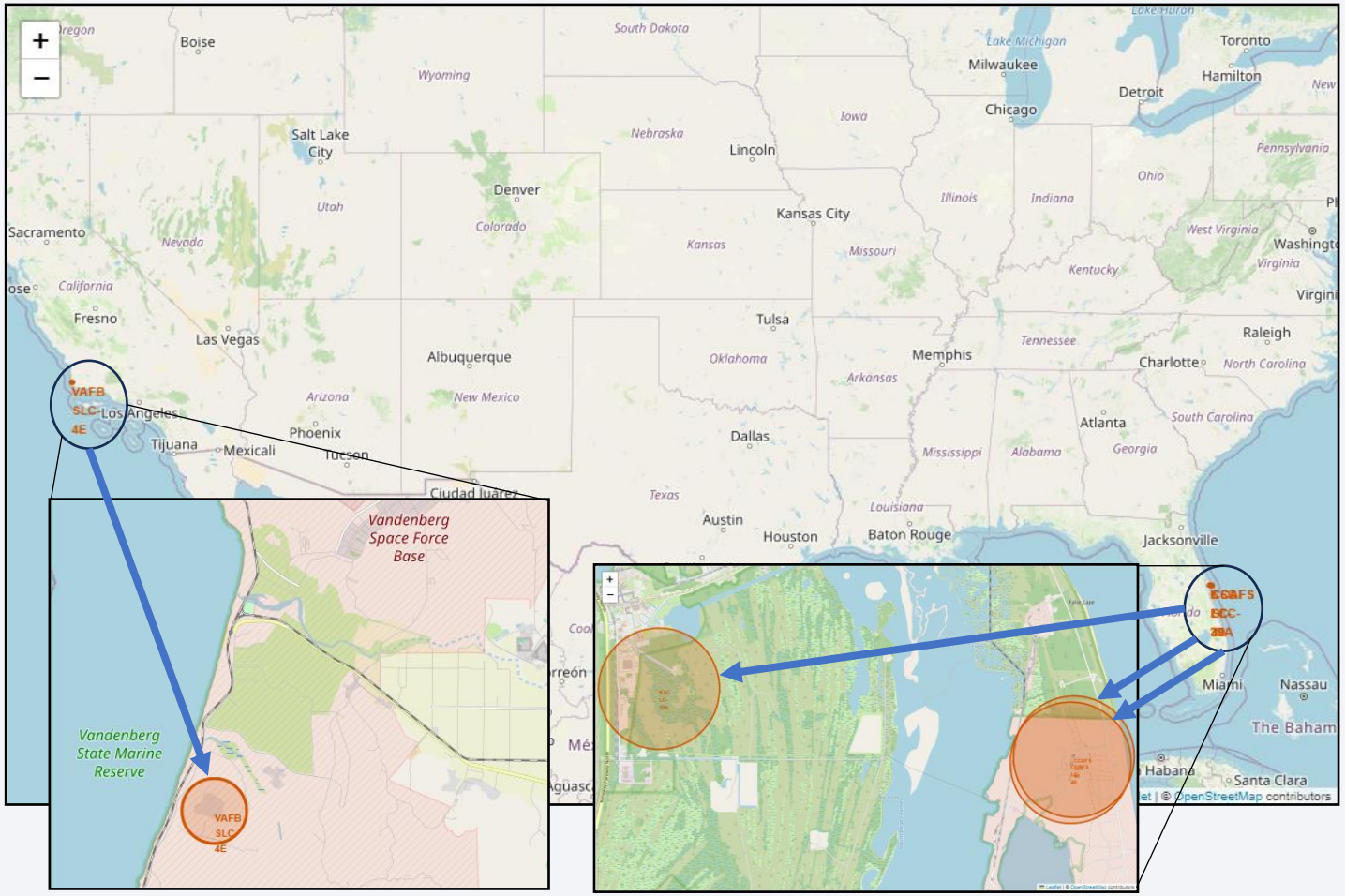
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

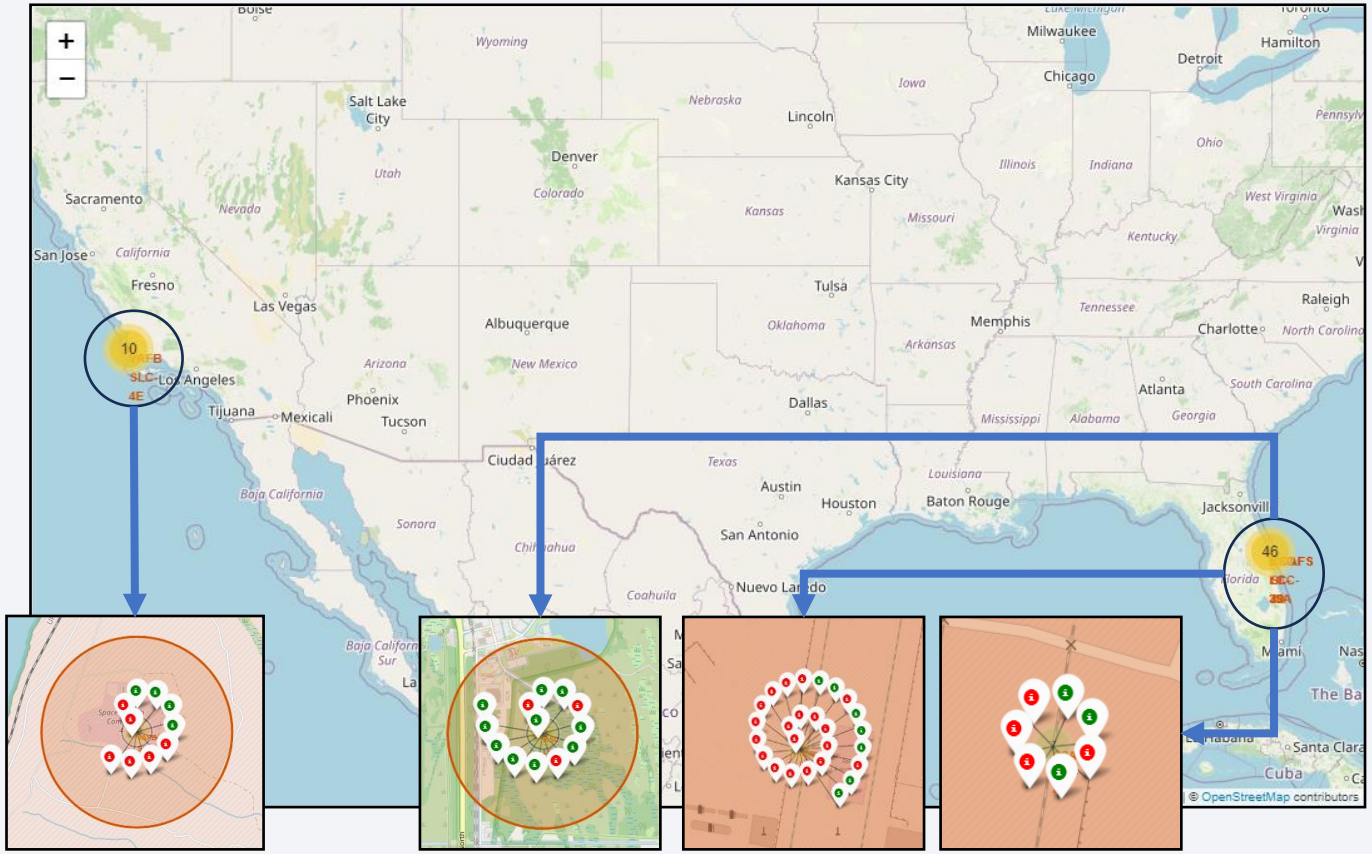
Folium Map Screenshot 1 – Launch Site Locations

- VAFB SLC-4E is situated near the western coastline, while KSC LC-39A, CCAF LC-40 and CCAF SLC-40 are positioned near the eastern coastline.
- Upon zooming in it seems that CCAF LC-40 and CCAF SLC-40 are situated near one-another.
- Launch sites are near the sea, probably for safety reasons but not far from roads and railroads.



Folium Map Screenshot 2 – Success/Failed Launches for Launch Sites

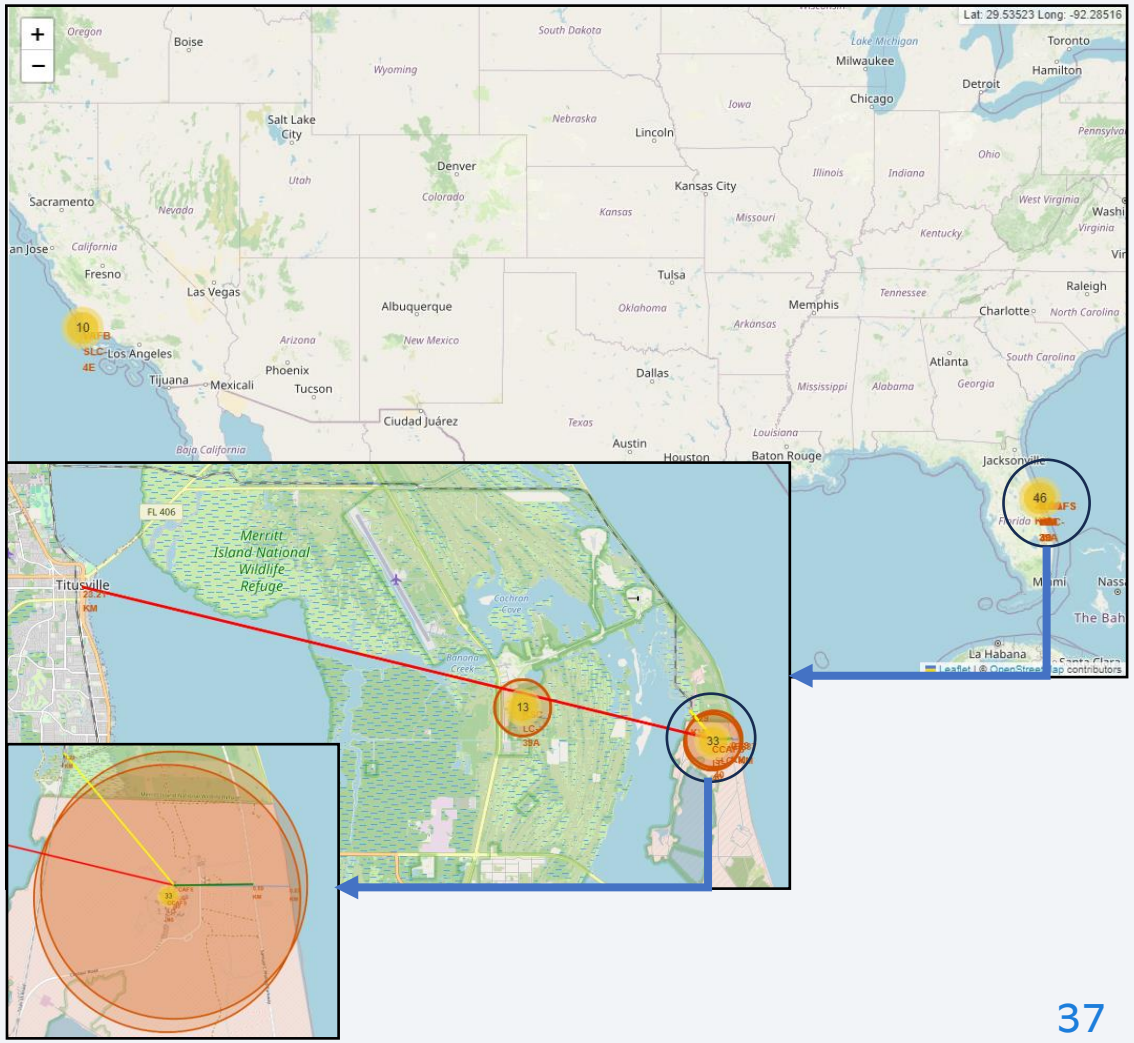
- Upon zooming into the map, we can see the success and failure markers which are green and red colored respectively for each site.
- It is observed that out of the 13 launches in KSC LC-39A 10 of them were successful while 3 missions failed which gives it the highest success rate of about 76.9%.

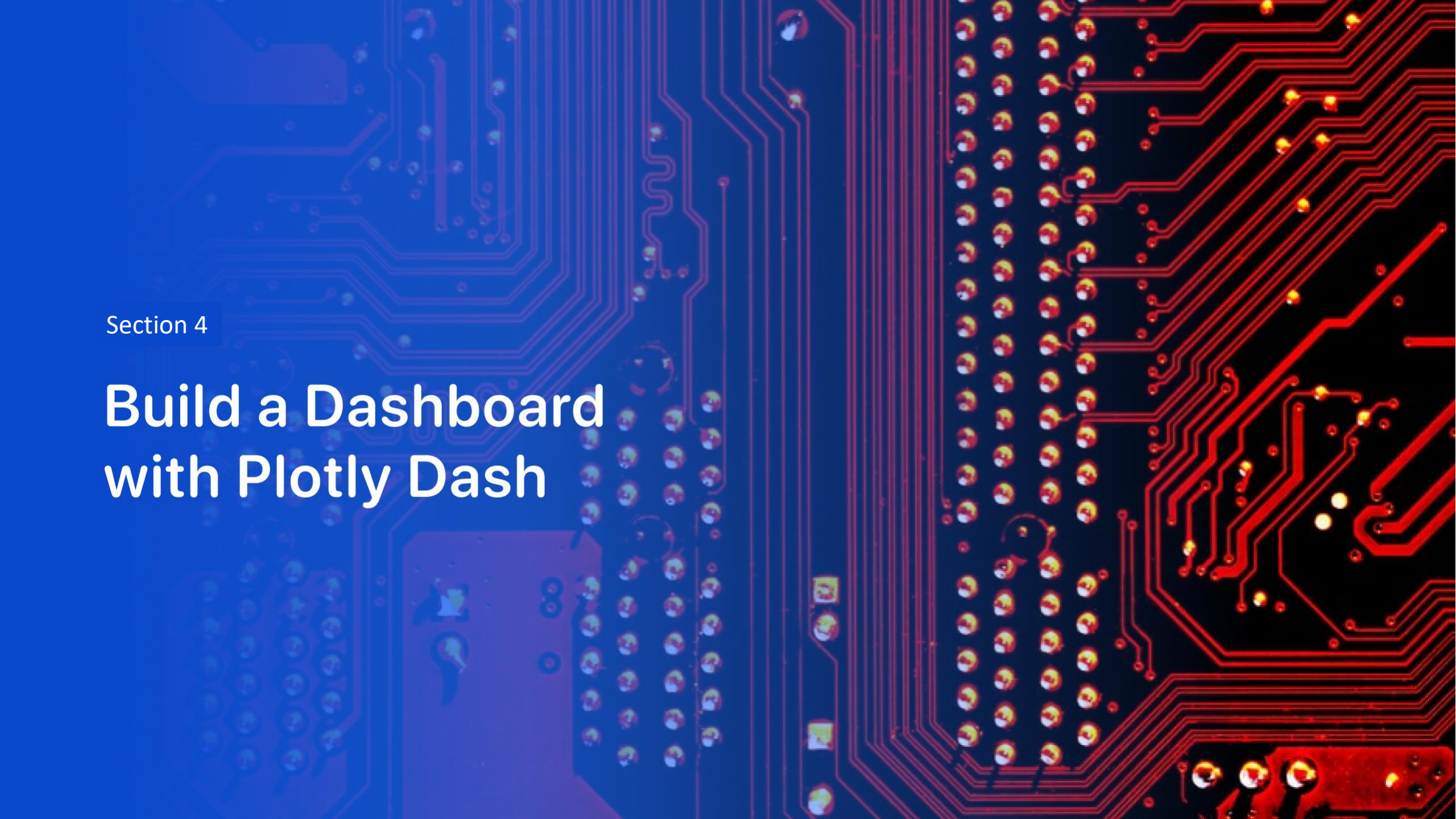


VAFB SLC-4E KSC LC-39A CCAFS LC-40 CCAFS SLC-40

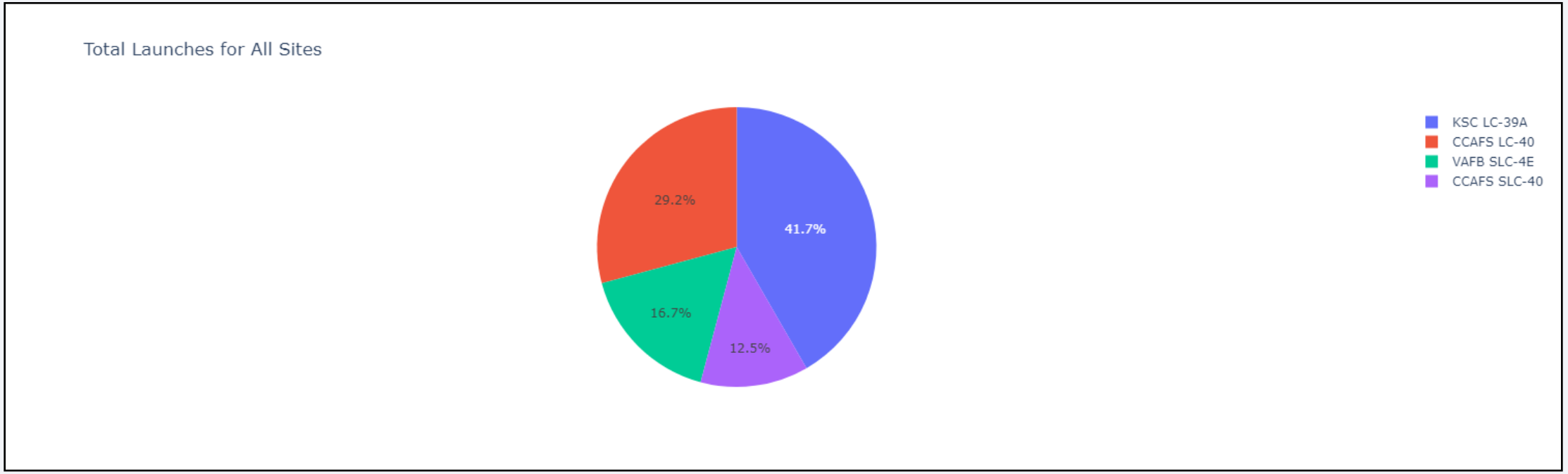
Folium Map Screenshot 3 – Proximity of Launch Site(s) to Other Areas

- From the map we can see that the launch site CCAFS SLC-40 is near the coastline (0.87 km), railroad (1.29) and highway (0.59 km).
- CCAFS SLC-40 is far away from the cities (23.21 km from Titusville).
- All launch sites are near the coastline near railroads and highways but not near cities for safety reasons so that the debris or fuel do not hit people.



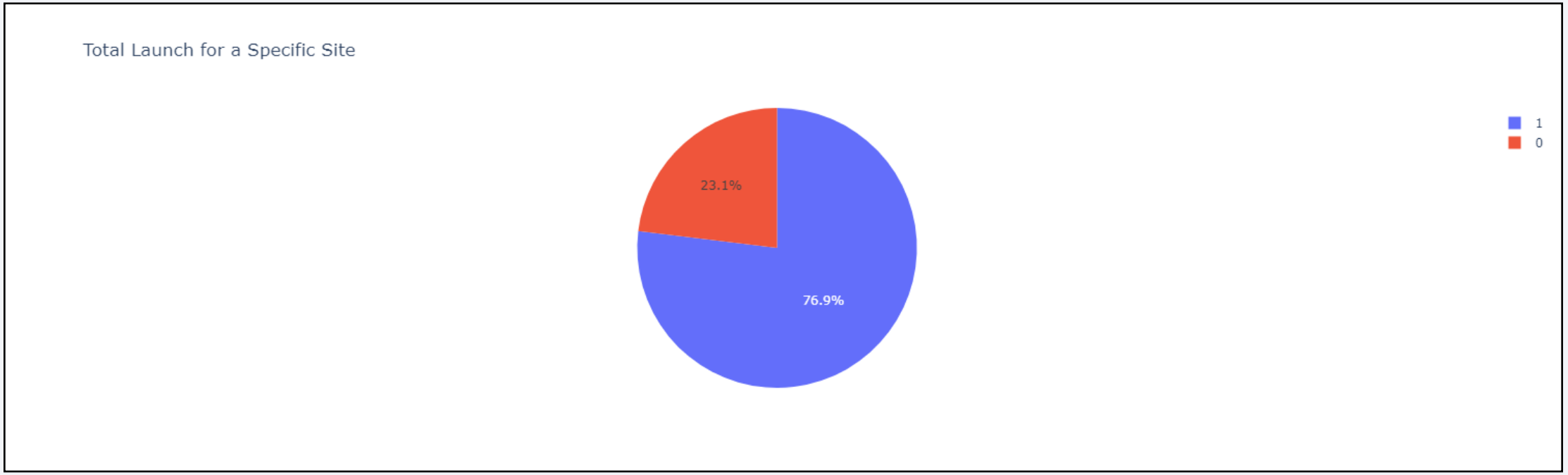

Section 4
Build a Dashboard with Plotly Dash

Dashboard Screenshot 1 – Total Success By All Sites



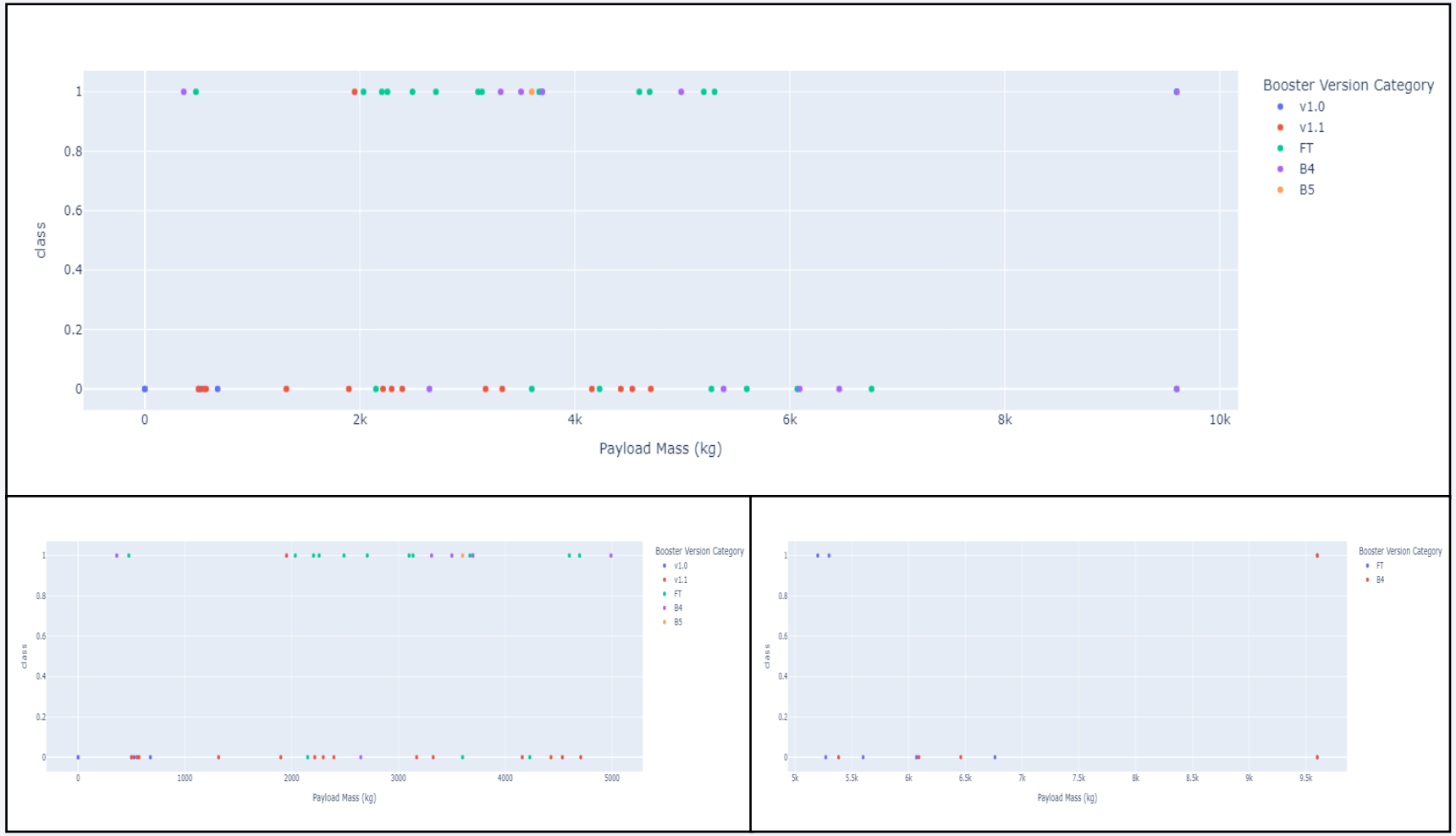
- The location of launch site appears to be an important contributing factor to success of missions.
- KSC LC-39A has the most successful launches compared to other sites.

Dashboard Screenshot 2 – Launch Site With Highest Launch Success Ratio



- Upon using the dropdown menu on the dashboard allows us to view single site launches.
- At KSC LC-39A, 76.9% of the launches resulted in success while 23.1% of the launches resulted in failure.

Dashboard Screenshot 3 – Payload vs Launch Outcome



- With range slider we can observe the outcomes of both successful and failed launches for each booster version along with the corresponding payload carried by it.
- Success rate is higher for Lower (Lighter) payloads compared to Higher (Heavier) payloads.

Low Weighted Payload (0 - 5000 kg)

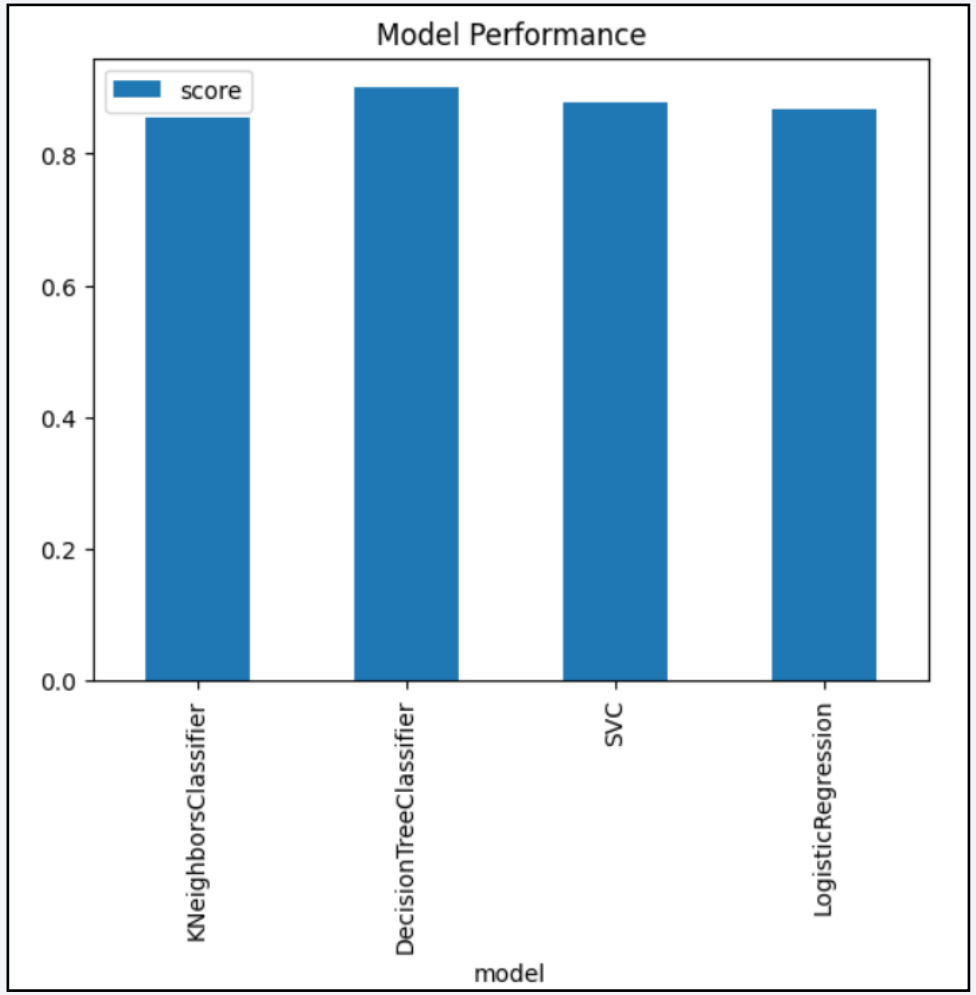
High Weighted Payload (5000 – 10000 kg)

Section 5

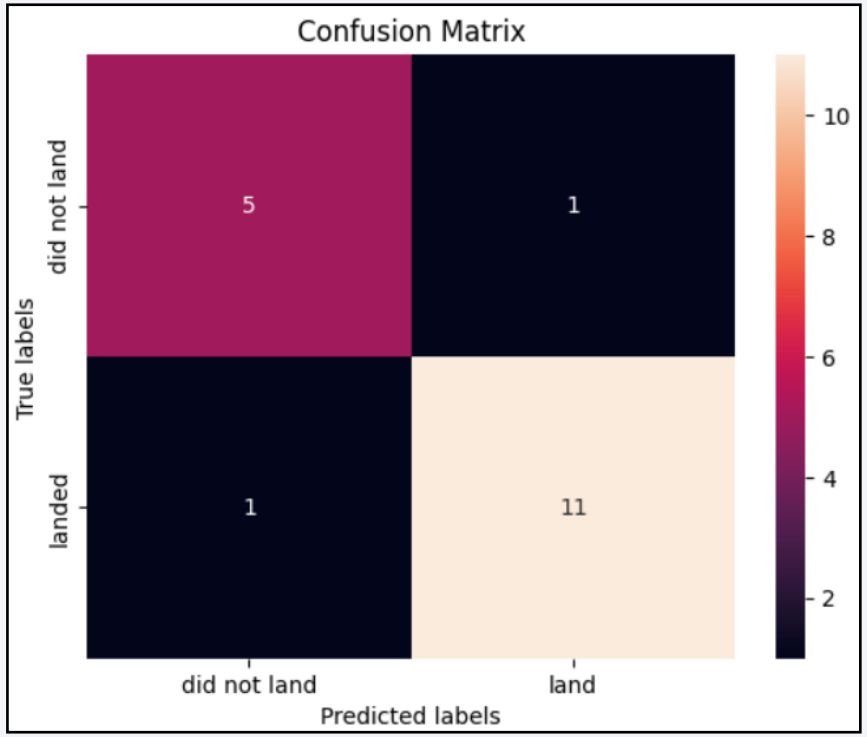
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies were plotted as shown.
- The decision tree model achieved the highest accuracy of around 88% out of the four models tested.



Confusion Matrix



- The best performing model was the decision tree model with an accuracy of 88%.
- The confusion matrix of the model shown above proves the accuracy of the decision tree model as it has high values of true positive and true negative values while having low values of false positive and false negative values.

Conclusions

- The success rates for SpaceX launches increased over time.
- KSC LC-39A has the most successful launches compared to other sites.
- Low weighted payloads have a higher success rate compared to high weighted payloads.
- ES-L1, GEO, HEO and SSO orbits achieved highest success rate.
- Decision tree model is the best in terms of prediction accuracy for this dataset.
- Through utilisation of available data and comprehensive analysis, rocket companies can pinpoint the most effective techniques for decreasing launch expenses. This approach ensures that the companies don’t lose clients and remain in the competitive market.

Appendix

- Folium maps do not show on GitHub so Jupyter nbviewer was used and screenshots of the maps were taken.

Thank you!