IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Gowtham Balaji>
<02/12/2023>

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 Data 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:

The buildup for space race is building rapidly and one of the major contributors for this is SpaceX. SpaceX has revolutionized rocket launches by successfully landing the first stage of the rocket, so that it can be reused. Due to this, they are able to quote a cost of 62 million dollars for a rocket launch whereas other competitors are quoting upward of 162 million dollars. 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 SpaceX for a rocket launch.

Problems to find answers:

- What parameters influence the successful landing of the rocket?
- The relation between different parameters that determine the success rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful landing program?

Section 1

Methodology

Methodology

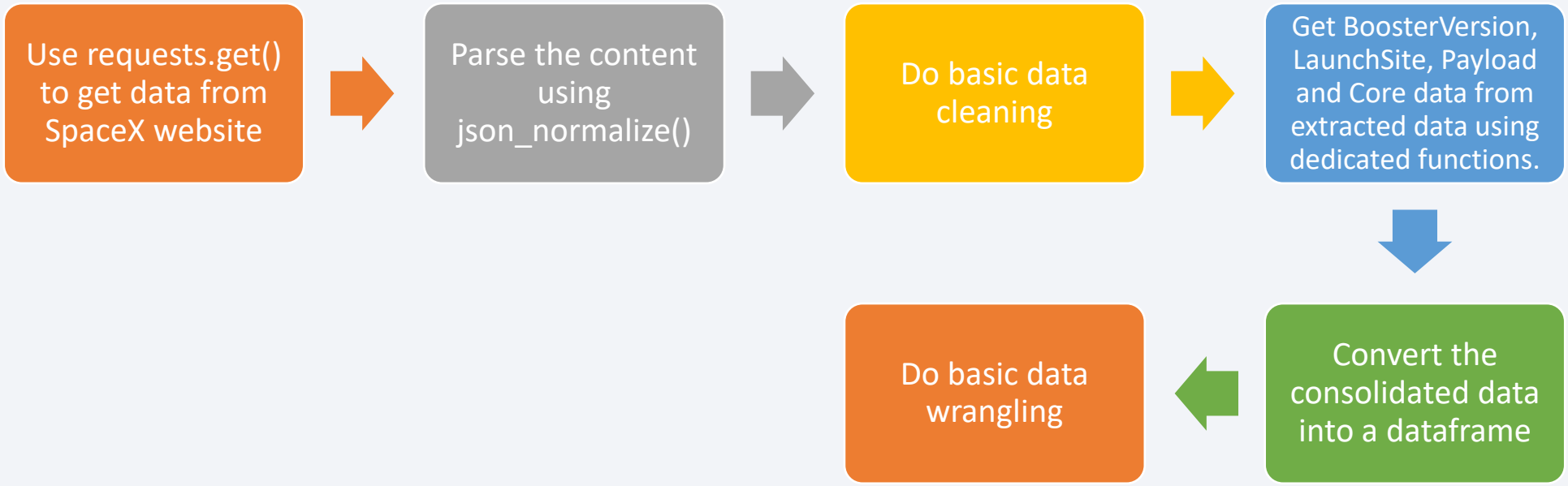
Executive Summary

- Data collection methodology:
  - Falcon 9 launch data is collected from the SpaceX launch data website using SpaceX API and webscraping from wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features, mainly for launch outcome.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different classification models were used and they were tuned to find the best hyper parameter for each model and were fitted to the data and their accuracy were evaluated.

Data Collection

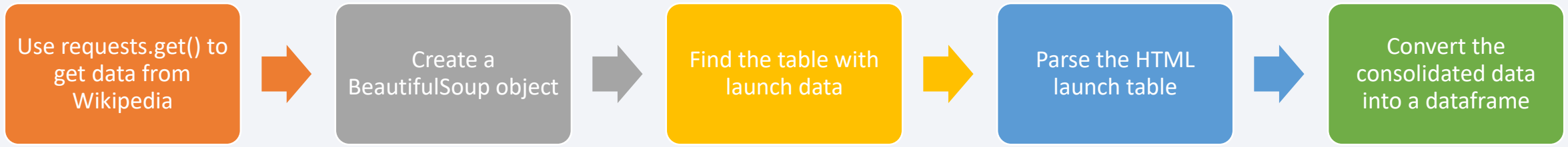
- Data collection was done using get request to the SpaceX API.
- Next, the response content was decoded as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
- Then the data was cleaned, missing values were filled where necessary.
- In addition, web scraping from Wikipedia was done for Falcon 9 launch records using BeautifulSoup.
- The aim was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for further analysis.

Data Collection – SpaceX API



- The GitHub link to the notebook is https://github.com/rgbalaji24/data_science_capstone/blob/045e8e298401c160ef2847e5bc6f0361e68a5000/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



- The GitHub link to the notebook is https://github.com/rgbalaji24/data_science_capstone/blob/045e8e298401c160ef2847e5bc6f0361e68a5000/jupyter-labs-webscraping.ipynb

Data Wrangling

- Exploratory data analysis was performed to determine the training labels.
- The number of launches at each site, and the number and occurrence of each orbits were calculated.
- Landing outcome label was created from outcome column
- The GitHub link to the notebook is https://github.com/rgbalaji24/data_science_capstone/blob/045e8e298401c160ef2847e5bc6f0361e68a5000/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Plotted Flight number against Pay load mass colored by outcome to see the relationship between them with outcome of the launch.
- Plotted Flight number against Launch site colored by outcome to see the outcome of launch in different launch sites.
- Plotted Pay load mass against Launch site colored by outcome to see the outcome of launch based on payload.
- Plotted a bar graph to see the outcome of launches intended for different orbits.
- Plotted different scatter plots of Flight number, orbit and pay load mass to see the relationship with launch outcome.
- Plotted the average success rate by timeline to see the success rate change over the years.
- The GitHub link to the notebook is https://github.com/rgbalaji24/data_science_capstone/blob/045e8e298401c160ef2847e5bc6f0361e68a5000/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010 06 04 and 2017 03 20, in descending order

GitHub Link to notebook:
https://github.com/rgbalaji24/data_science_capstone/blob/d203c51eff9f26c25bd04bbf4f2e7e10812437fb/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Added markers for each launch site and colored them according to launch success or failure.
- Each launch site is marked with a circle marker to easily identify them when visualizing.
- Added a marker cluster for each launch to show how many launches were performed from them.
- Calculated distance between launch site to its proximities to answer following questions.
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?

- GitHub Link to notebook: https://github.com/rgbalaji24/data_science_capstone/blob/b4006a3e2107464e99b8fa0a6dc3cc5eacc7c7f2/lab_jupyter_launch_site_location.jupyterlite.ipynb

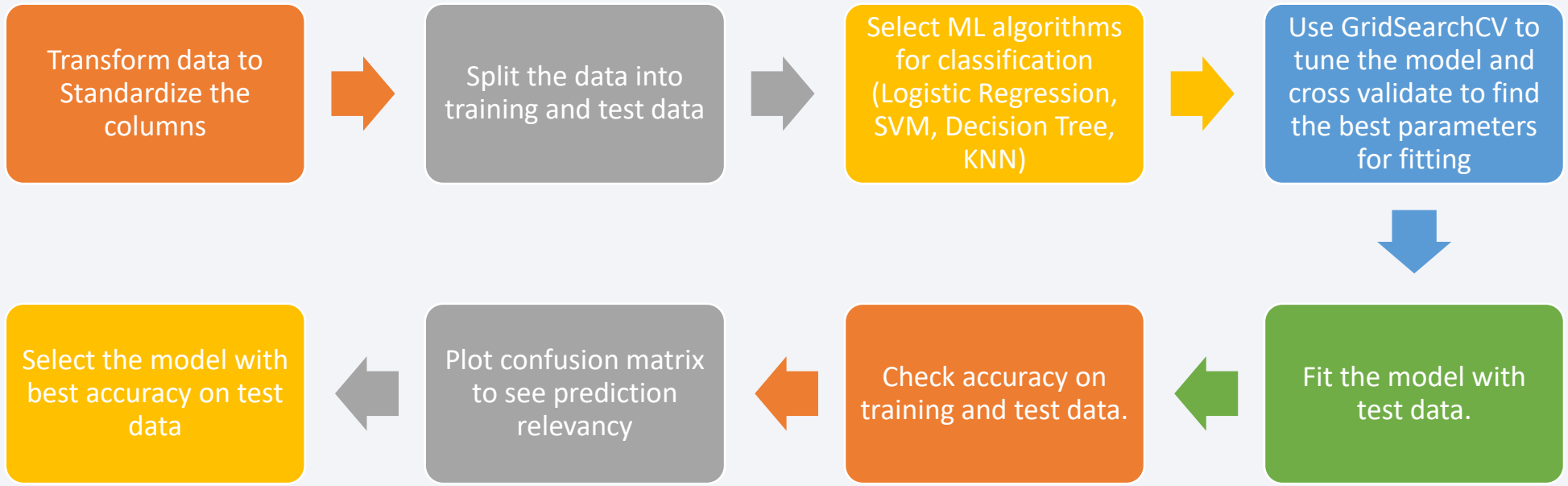
Build a Dashboard with Plotly Dash

- Created a dashboard using Plotly Dash to visualize the following.
  - Analyze the success rate of launch in different launch sites → To see which launch site has highest success rate
  - Check the correlation between Payload mass and launch outcome for different booster versions with a filter for payload range → To see which booster versions and payload mass had better success on launches.

The GitHub link to the notebook file:

https://github.com/rgbalaji24/data_science_capstone/blob/93a1080df747519365d9fbca90d4c2d3a83ac008/spacex_dash_app.py

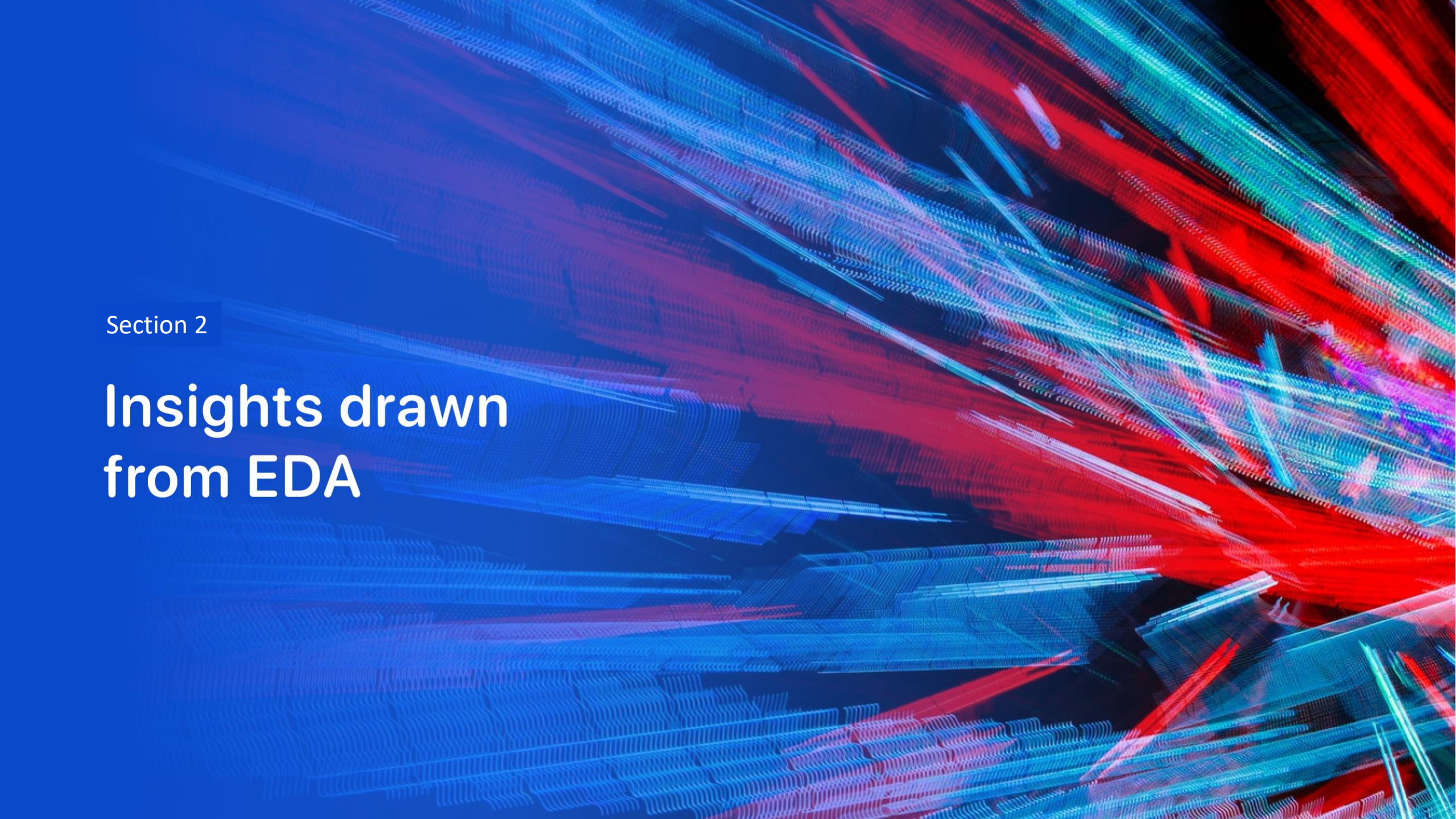
Predictive Analysis (Classification)



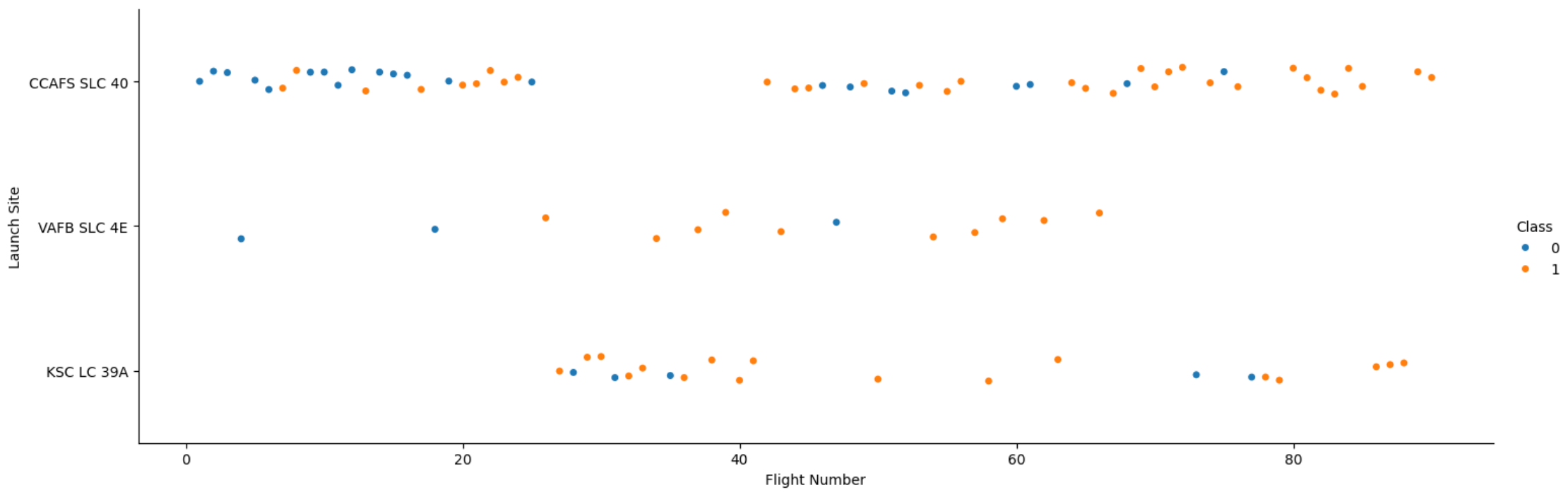
GitHub Link to notebook:
https://github.com/rgbalaji24/data_science_capstone/blob/5ab9696fc2a7725085fd7dc17671a70c171f7d62/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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


Section 2
Insights drawn from EDA

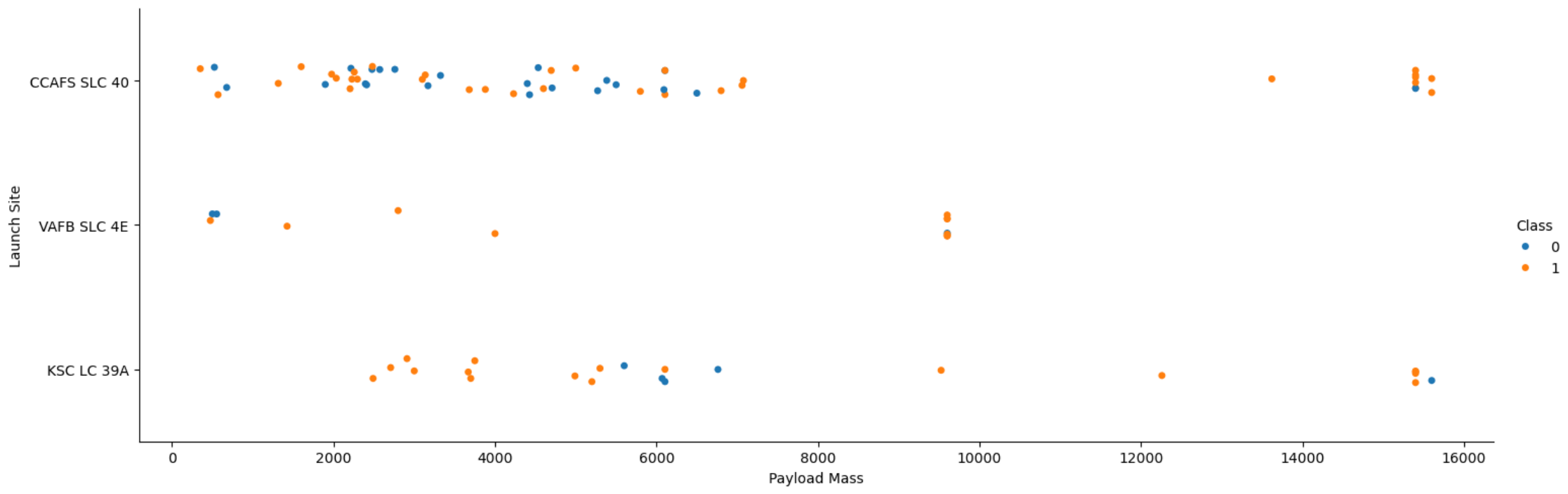
Flight Number vs. Launch Site



Observations:

- As the flight number increases, the number of successful landing increases.
- When more landings are attempted at a launch site, more successful landings happen.

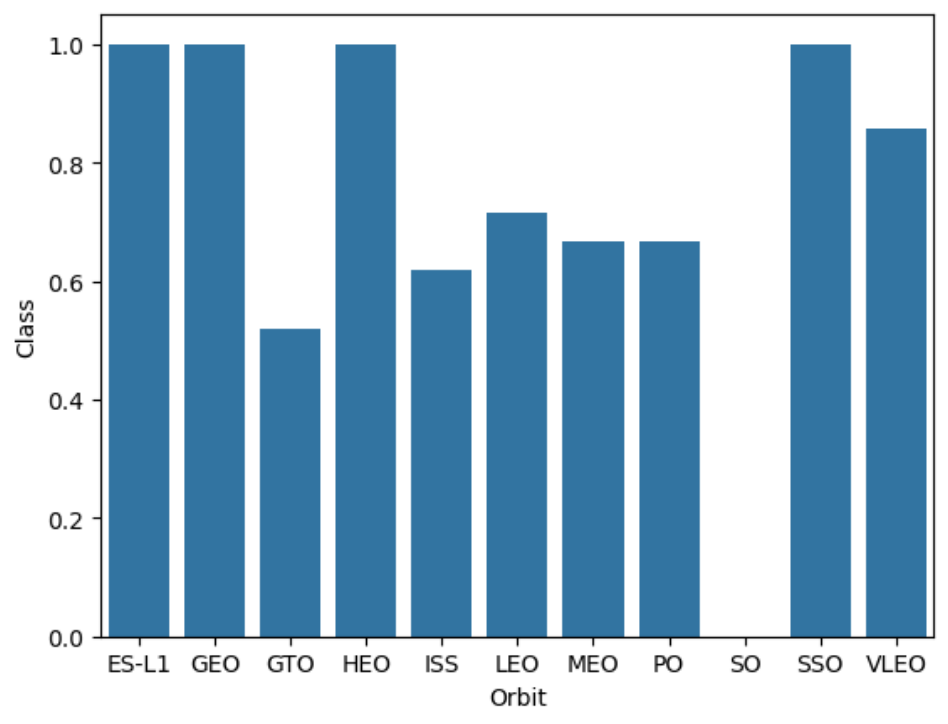
Payload vs. Launch Site



Observations:

- At VAFB SLC 4E launch site, there are no launches with payload more than 10000 kg.
- At KSC LC 39A site, more launches and success happens at lower payload range, while at VAFB SLC 4E more launches happen at mid payload range.
- At CCAFS SLC 40, as the payload increases, probability of successful landing increases.

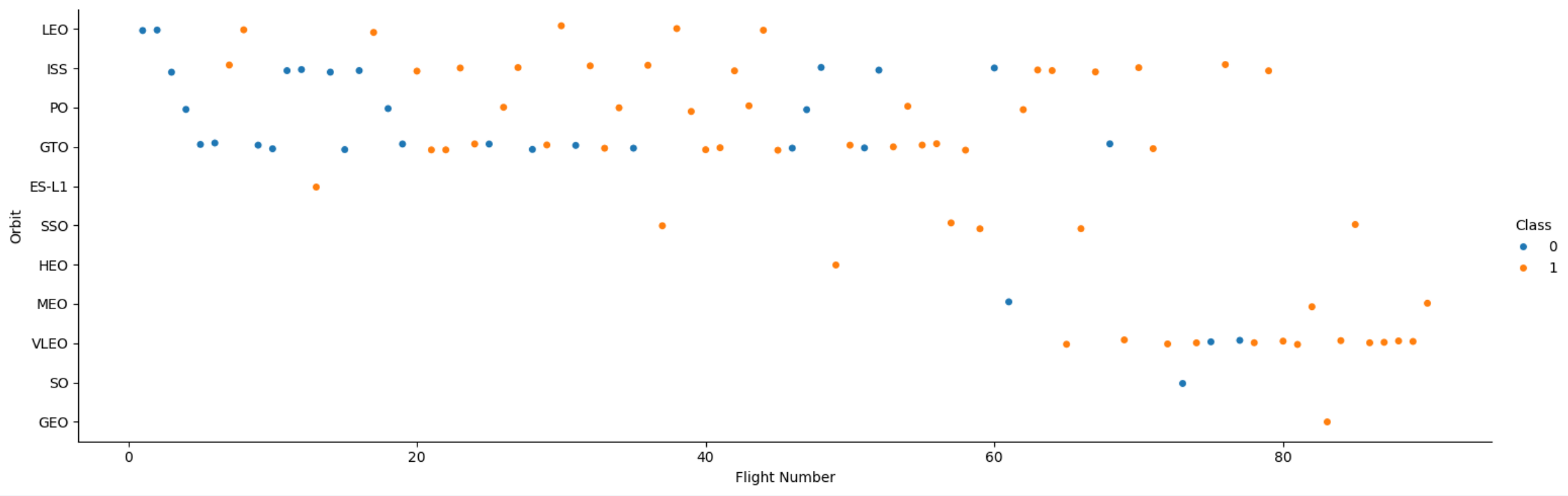
Success Rate vs. Orbit Type



Observations:

- Success rate is the highest for the orbits ES-L1, GEO, HEO and SSO

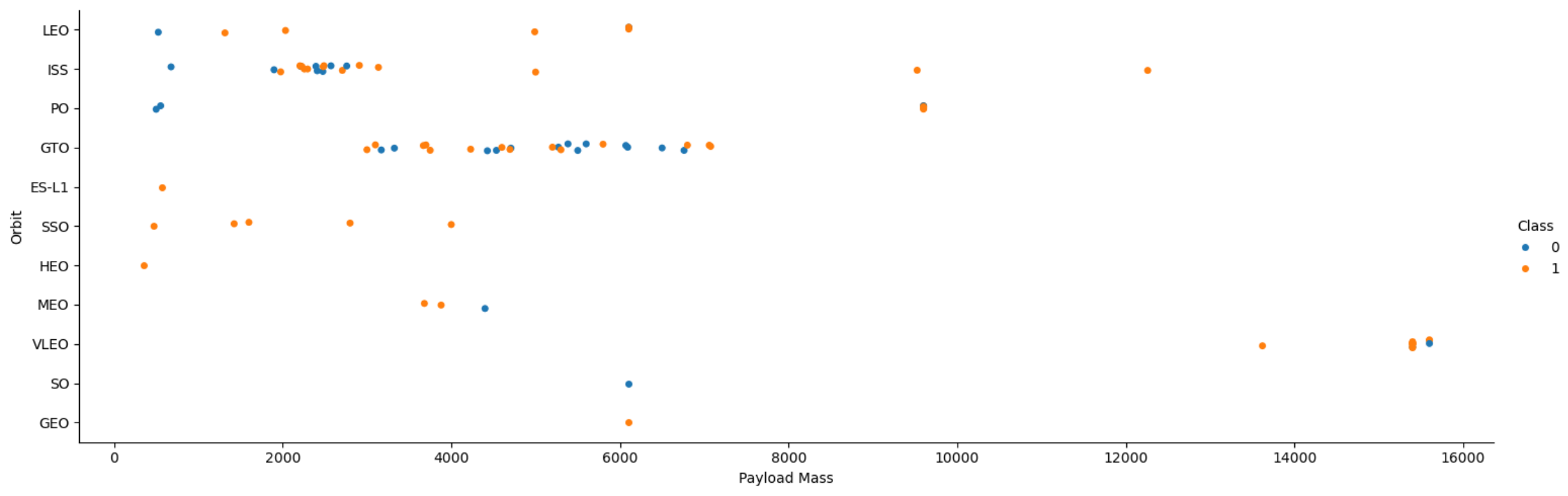
Flight Number vs. Orbit Type



Observations:

- For LEO, probability of success is high as flight number increases, there is no direct correlation for other orbits with flight number.

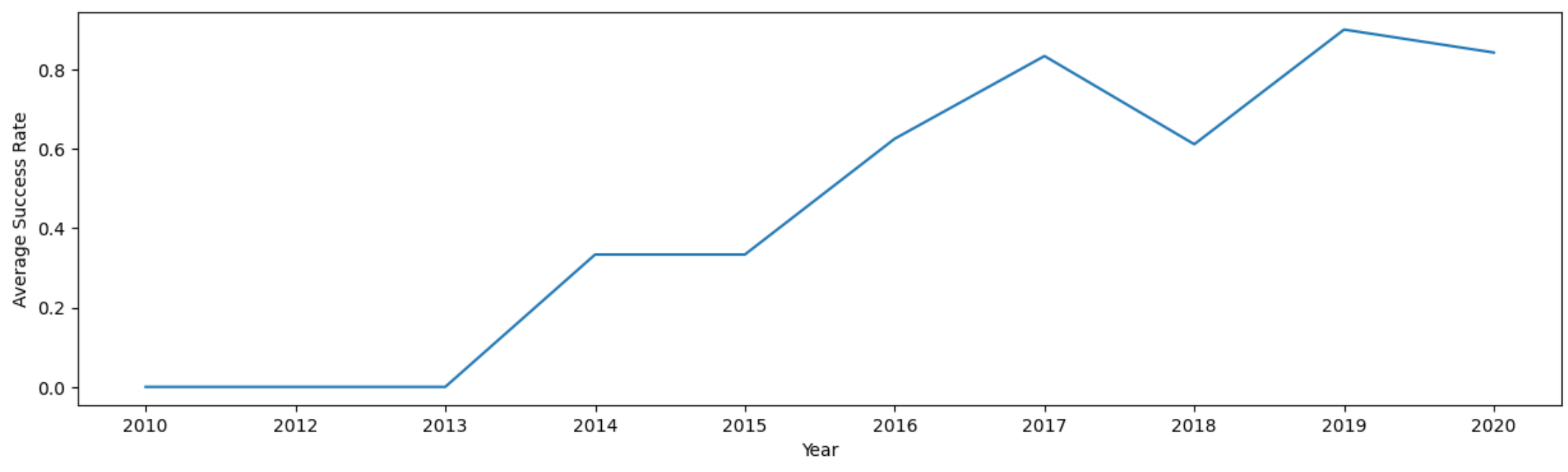
Payload vs. Orbit Type



Observations:

- For SSO, has good success rate for lower payload launches.
- LEO, ISS and PO tend to have better success at higher payloads.
- Other orbits do not have clear relationship with payload mass.

Launch Success Yearly Trend



Observations:

- Till 2013, there are no successful landings.
- After 2013, the probability of successful landings kept increasing till 2020, barring the year 2018.

All Launch Site Names

Task 1

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

[9]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

[9]:

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

- There are a total of 4 unique launch sites.
- Used the DISTINCT keyword to show only unique launch sites.

Launch Site Names Begin with 'CCA'

Task 2

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

```
[10]: %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE ("CCA%") LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

[10]:

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

- Used the LIKE keyword with % wild card to get sites starting with “CCA”.
- Used the LIMIT keyword to show only 5 records.

Total Payload Mass

Task 3

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

```
[11]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTBL WHERE Customer = "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

[11]:

total_payload_mass
45596

- Calculated total payload mass using SUM aggregation.
- Filtered only the boosters launched by “NASA (CRS)” in WHERE clause.
- Total payload mass launched by NASA (CRS) is 45596 kg.

Average Payload Mass by F9 v1.1

Task 4

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

```
[12]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass FROM SPACEXTBL WHERE Booster_Version LIKE("F9 v1.1%")
```

```
 * sqlite:///my_data1.db
Done.
```

[12]:

avg_payload_mass
2534.6666666666665

- Calculated the average payload mass carried by booster version F9 v1.1
- Filtered only the booster version “F9 v1.1%” in WHERE clause using LIKE keyword.
- The average payload mass carried by booster version F9 v1.1 is 2534.67 kg.

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

[13]:
```
%sql SELECT MIN(Date) AS first_successful_landing_date FROM SPACEXTBL WHERE Landing_Outcome = "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

[13]:

first_successful_landing_date
2015-12-22

- Found the date of the first successful landing outcome on ground pad using MIN function.
- Filtered only the successful landing outcome on ground pad in WHERE clause.
- The date of the first successful landing outcome on ground pad is 22/12/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

[14]:
```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

```
 * sqlite:///my_data1.db
Done.
```

[14]:

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

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- In WHERE clause, filtered the successful landing outcome on drone ship and the range of payload mass to be between 4000 and 6000 kg.
- Total 4 boosters had successful landing on drone ship.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

[15]:
```
%sql SELECT Mission_Outcome, COUNT(1) AS nb_of_outcomes FROM SPACEXTBL GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

[15]:

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

- Calculated the total number of successful and failure mission outcomes
- Used GROUP BY to aggregate on Mission_Outcome and counted the total number of events for each outcome.
- Total 100 successful outcomes and 1 failure outcome is observed.

Boosters Carried Maximum Payload

Task 8

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

```
[16]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

[16]:

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

- The names of the booster which have carried the maximum payload mass are listed
- Used a subquery to filter only the missions with maximum payload mass.

2015 Launch Records

Task 9

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.

[17]:
```
%sql SELECT SUBSTR(DATE,6,2) AS month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome LIKE("%Failure%") AND SUBSTR(DATE,0,5) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

[17]:

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

- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Used SUBSTR() function to get only the month from Date column and then filtered the failed landing outcome using LIKE and filtered the records from the year 2015.

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

Task 10

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.

```
[18]: %sql SELECT Landing_Outcome, COUNT(1) AS nb_of_outcomes FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY nb_of_outcomes DES
```

```
 * sqlite:///my_data1.db
Done.
```

[18]:

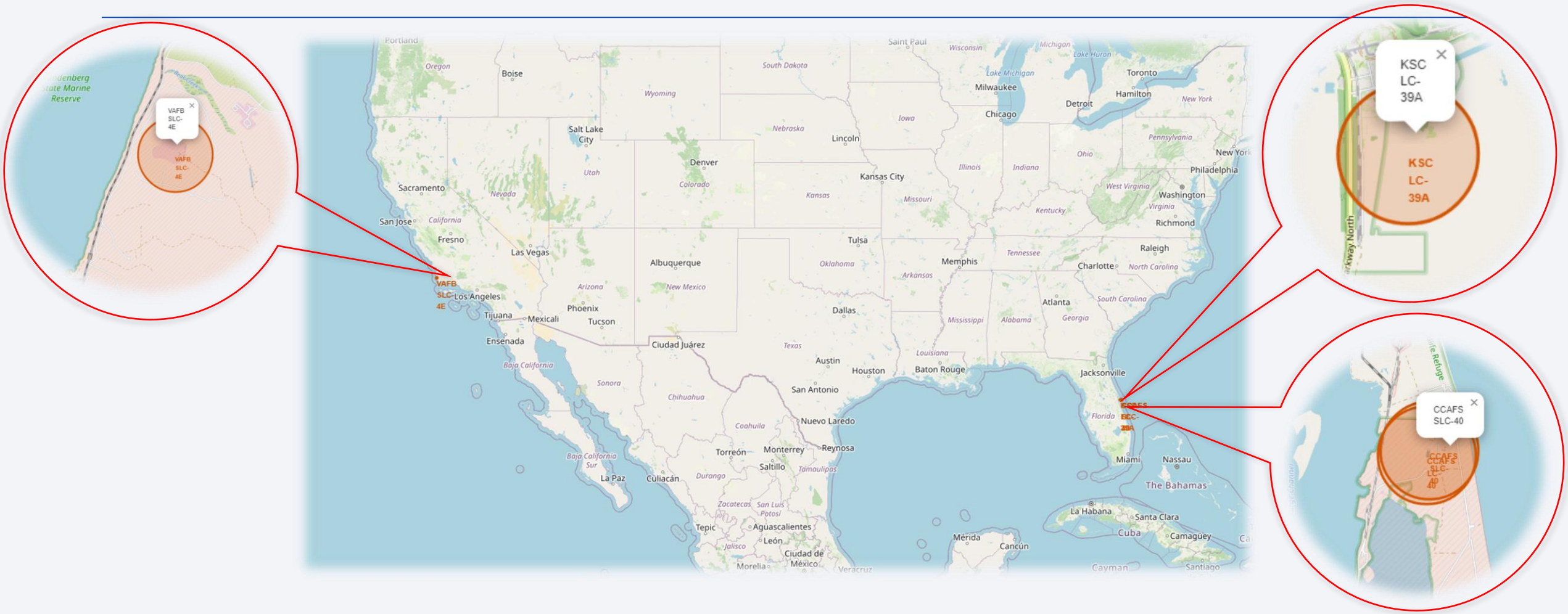
Landing_Outcome	nb_of_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- 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
- Filtered the Date column for required date range. Used GROUP BY on Landing_outcome to COUNT the nb_of_outcomes for each outcome and used ORDER BY on nb_of_outcomes with DESC order to rank the landing_outcome.

Section 3

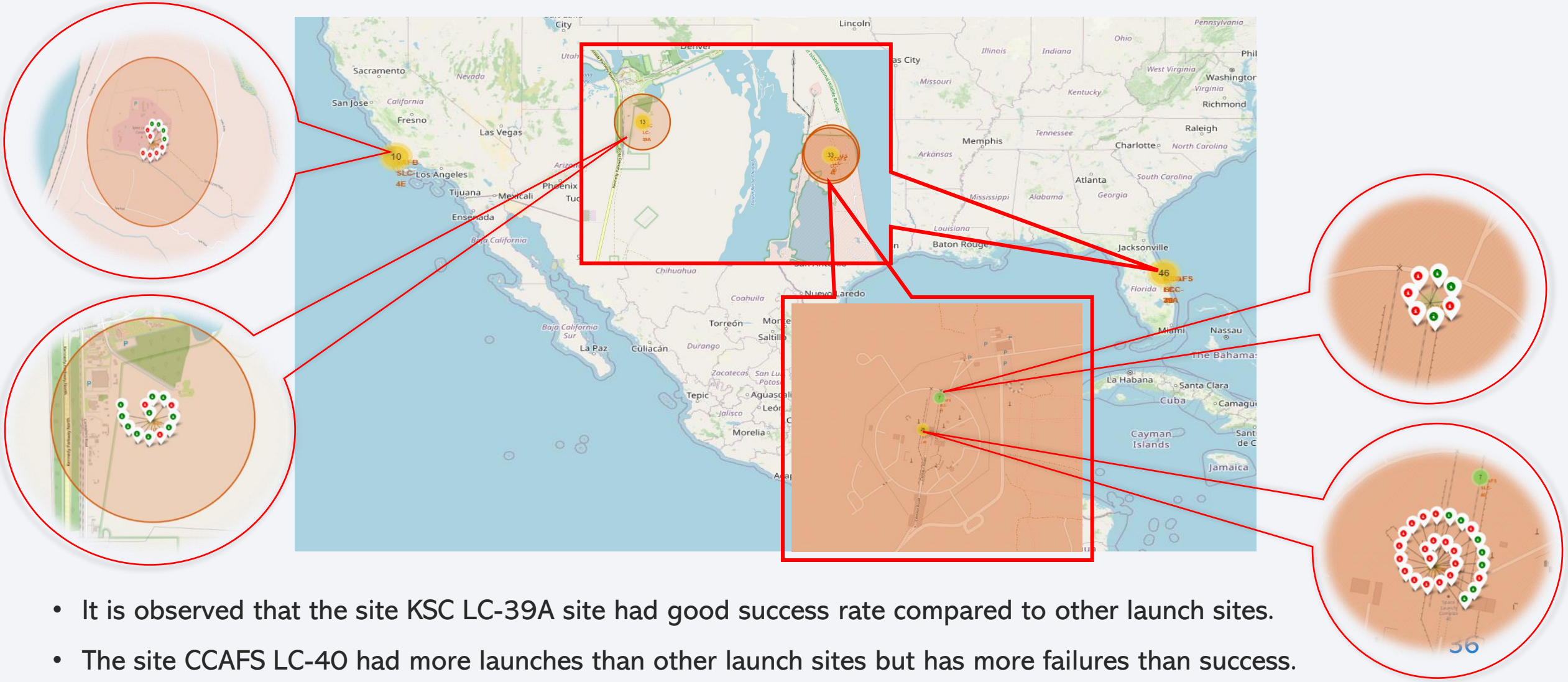
Launch Sites Proximities Analysis

Launch sites locations on global map



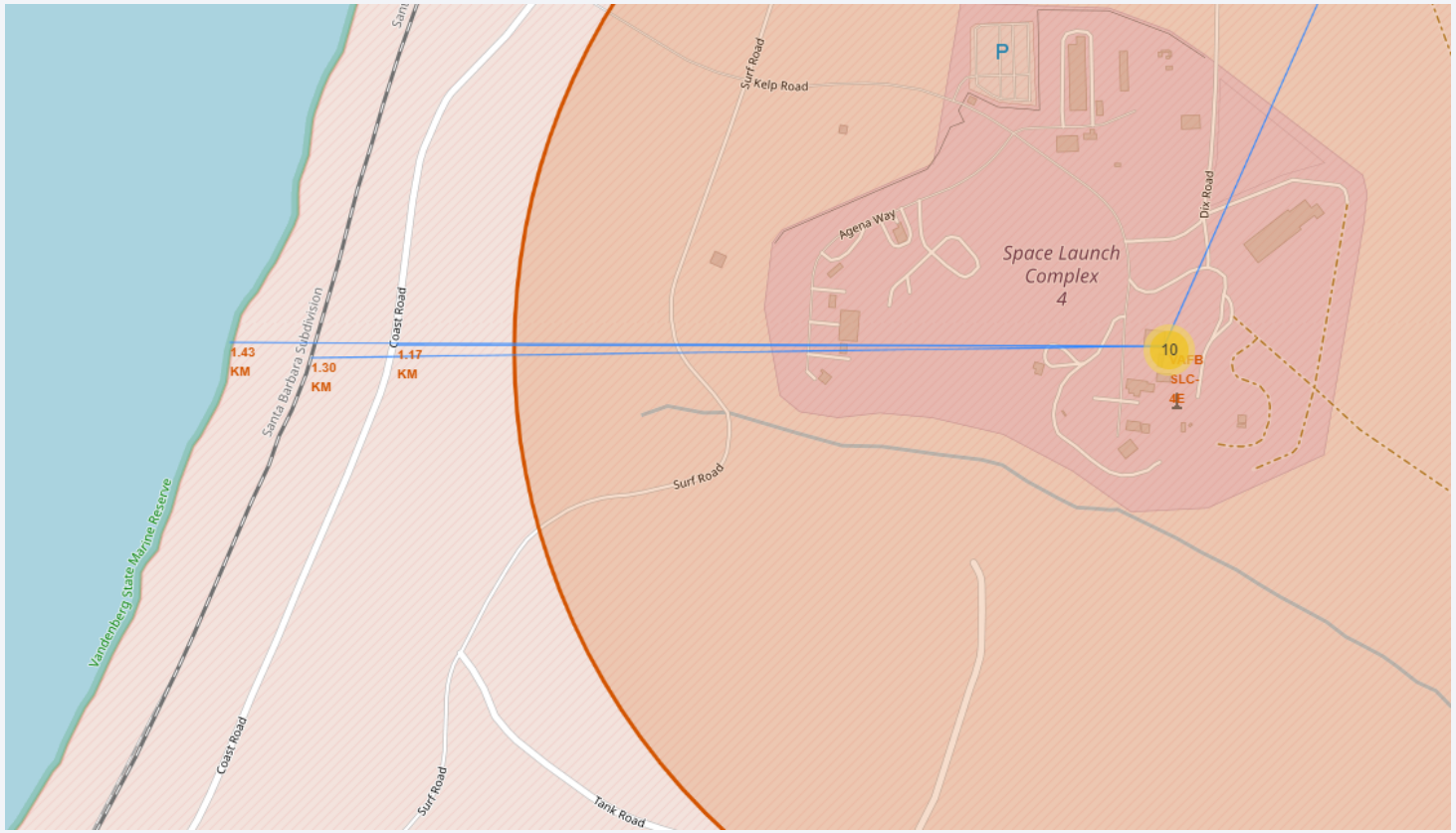
- All launch sites are located close to the ocean.
- 1 site is located on the western coast and remaining 3 sites are located In eastern coast

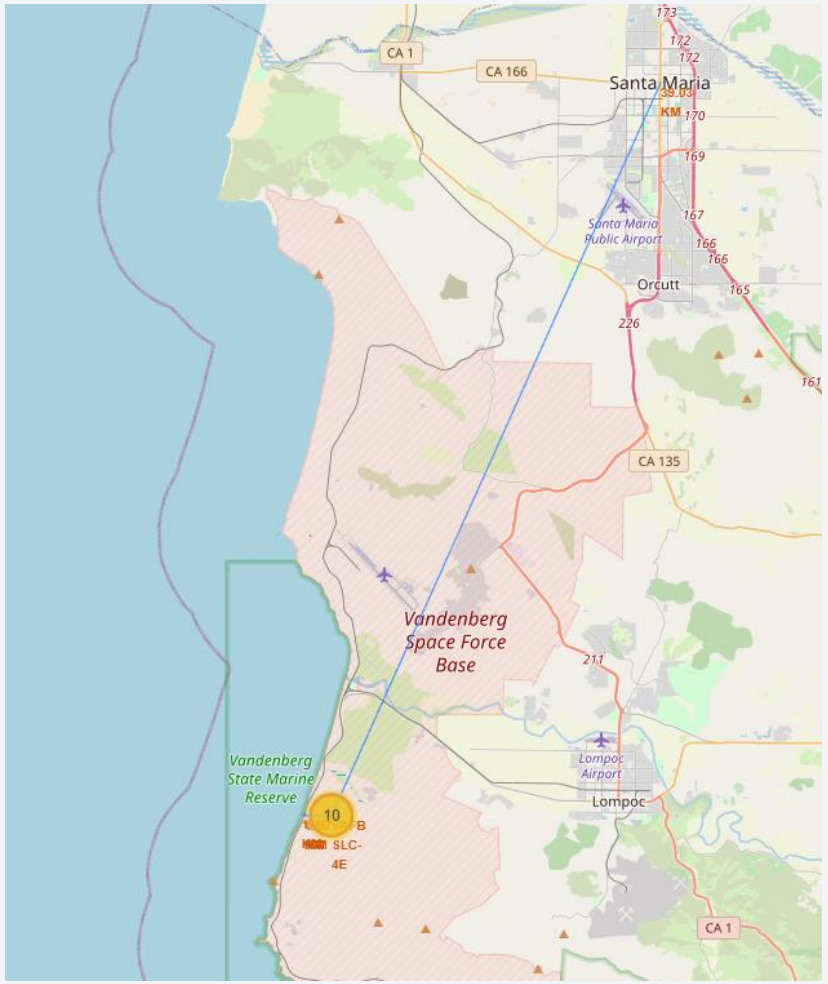
Outcome results in each site on global map



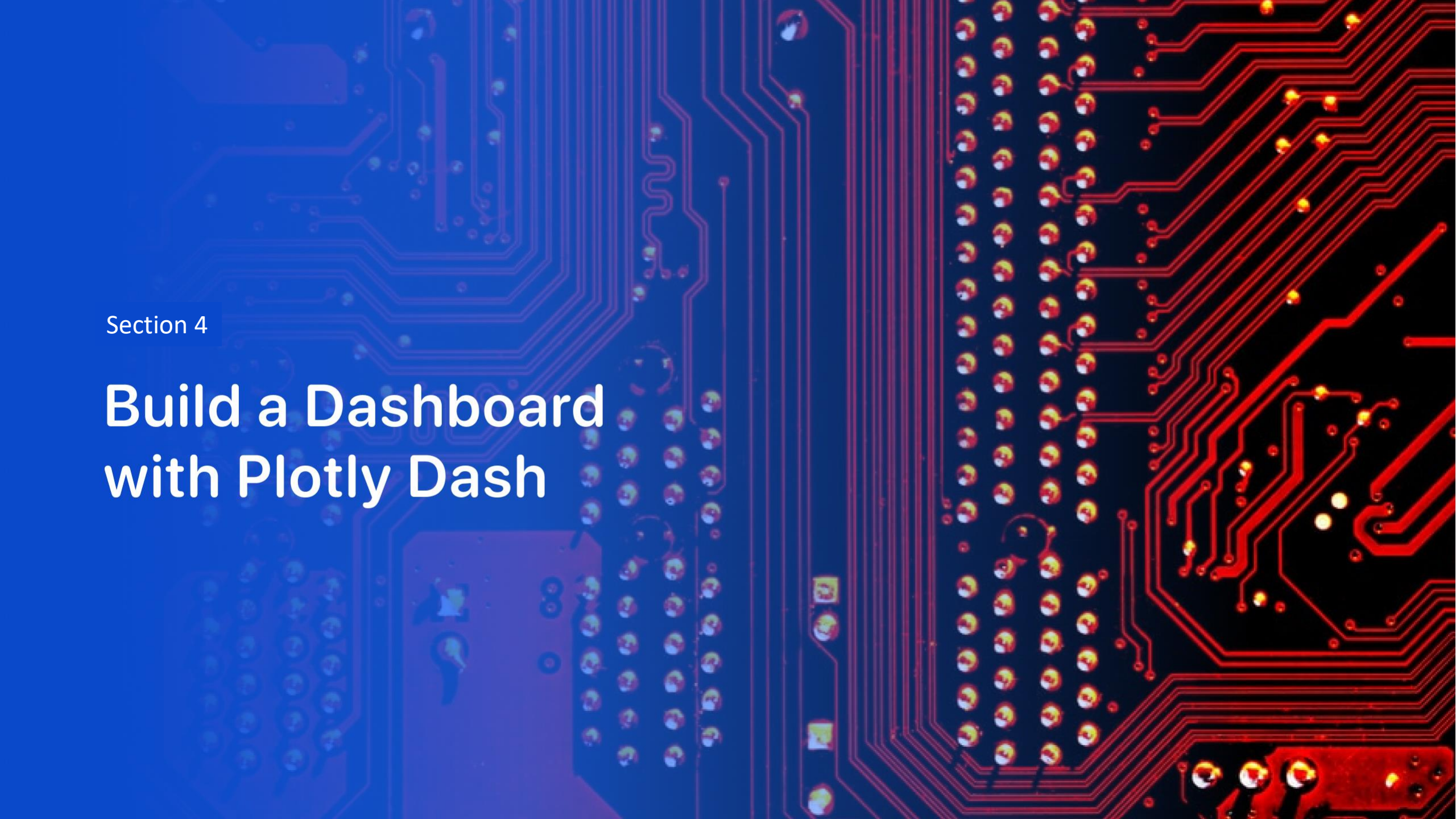
- It is observed that the site KSC LC-39A site had good success rate compared to other launch sites.
- The site CCAFS LC-40 had more launches than other launch sites but has more failures than success.

Launch site proximities

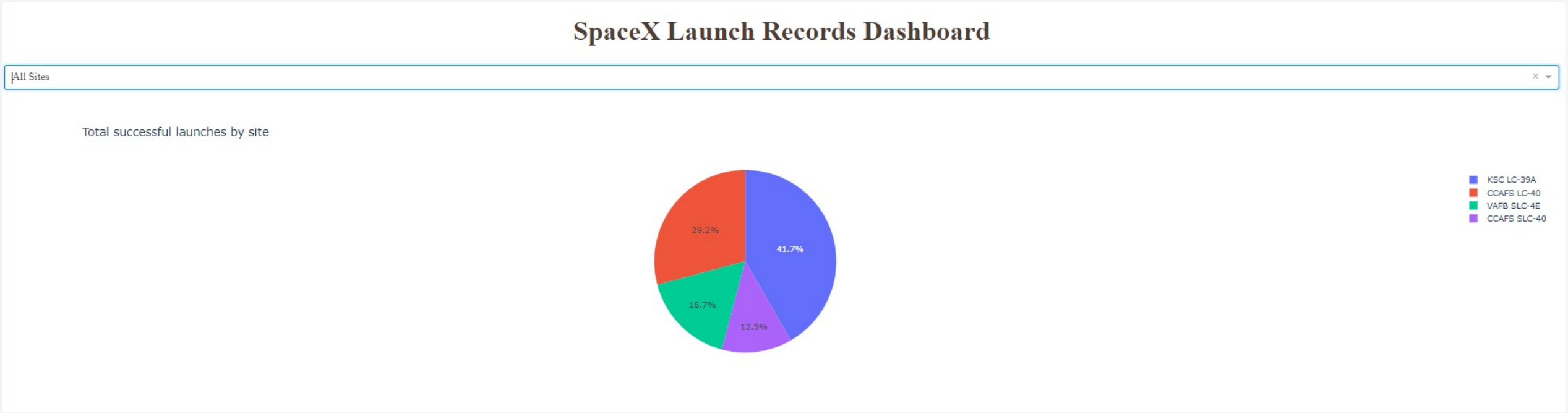




- As observed, the basic infrastructure is built around a launch site. The nearest coastline, highway and railline are within 2 km range.
- The nearest city Santa Maria is located approx. 40 km from the launch site considering safety measures.

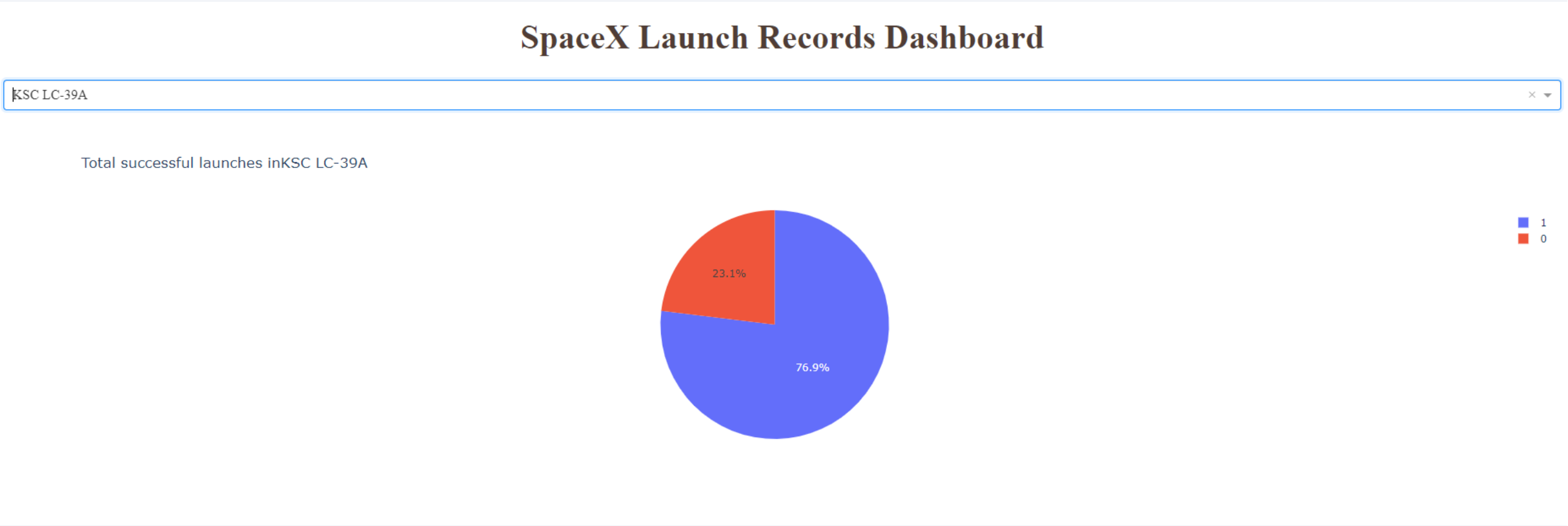

Section 4
Build a Dashboard with Plotly Dash

Distribution of successful launch in all sites



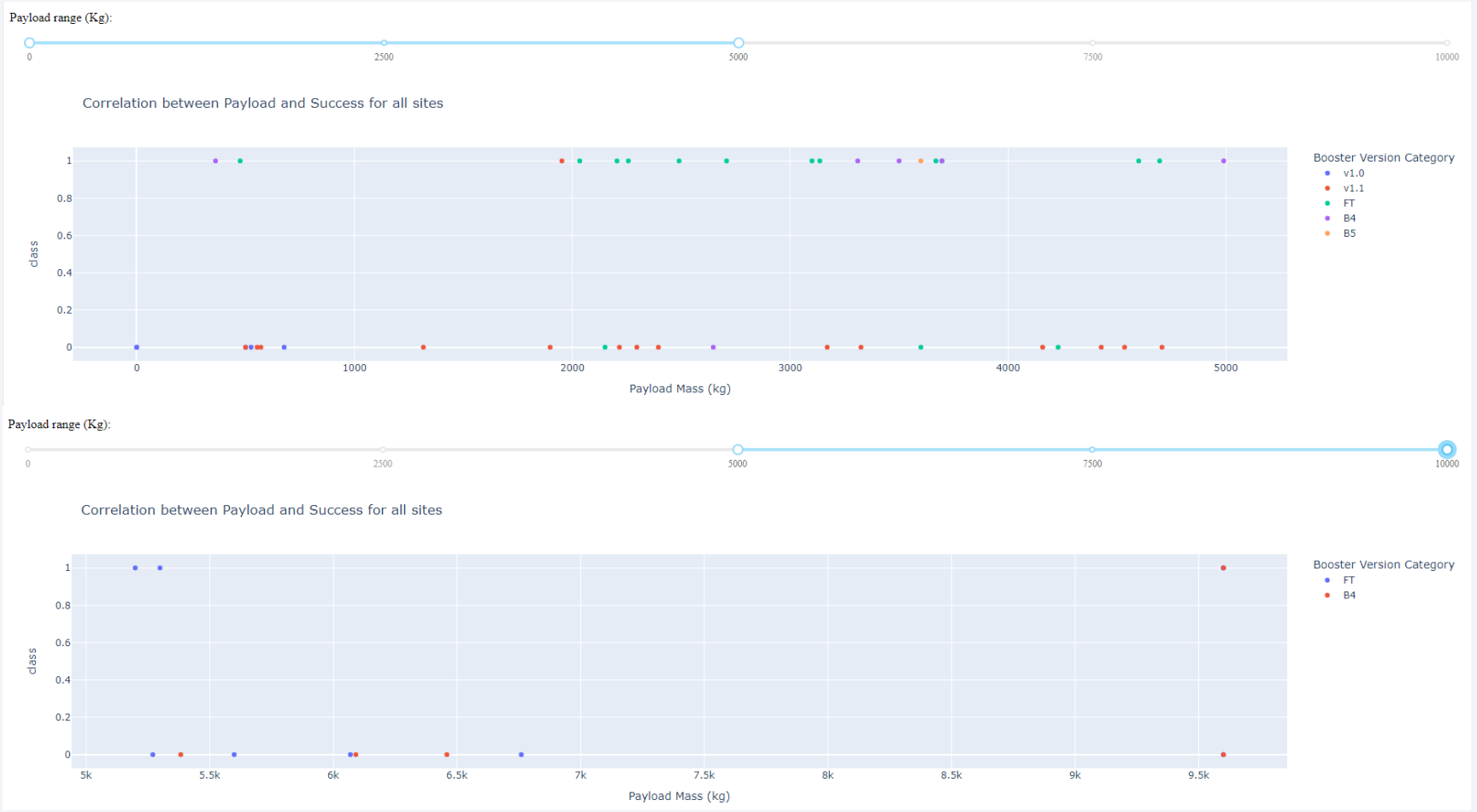
- Most successful launches is from the site KSC LC-39A with 41.7% of all launch successes, followed by CCAFS LC-40 with 29.2% approx.

Launch Site with the Highest Probability of Success



- The KSC LC 39A Launch Site also has the highest probability of success per launch
- 76.9% of all launches at the KSC LC 39A site landed successfully and 23.1% of all launches at the KSC LC 39A Site failed to land

Payload Mass against Launch outcome viewed by Booster Version

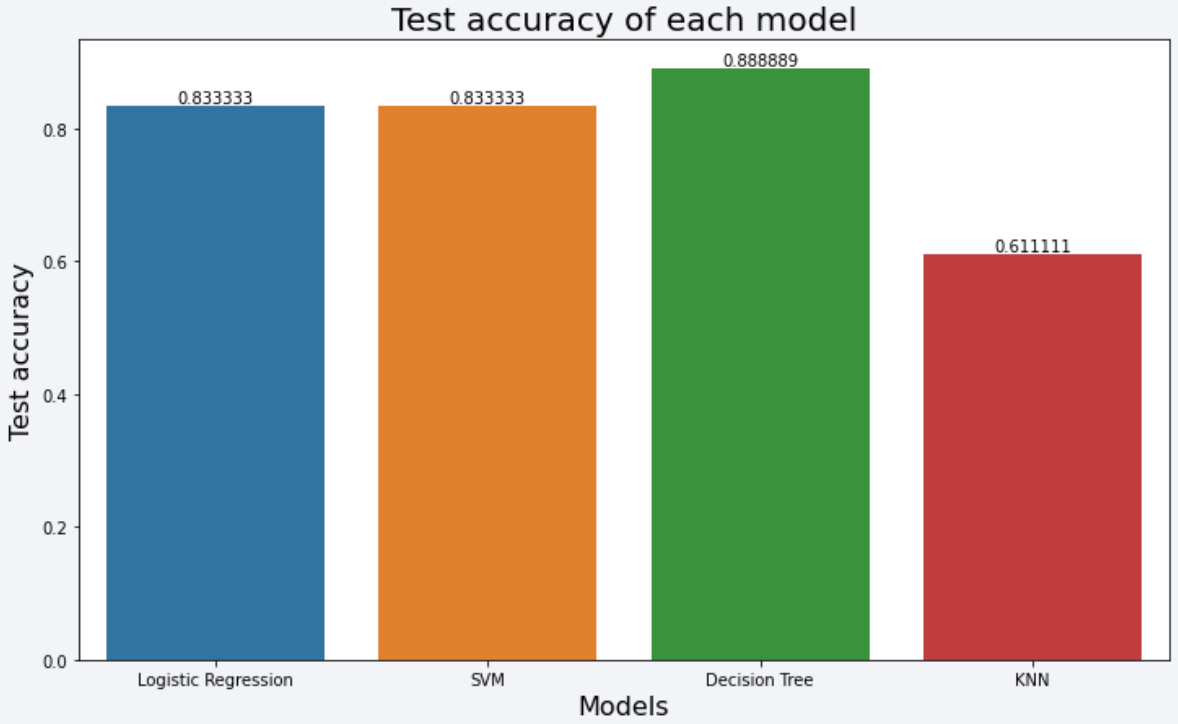


- Lower payload ranges (<5000) have better success rate compared to higher payload ranges (>5000).
- Booster version FT has better success rate than other boosters.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

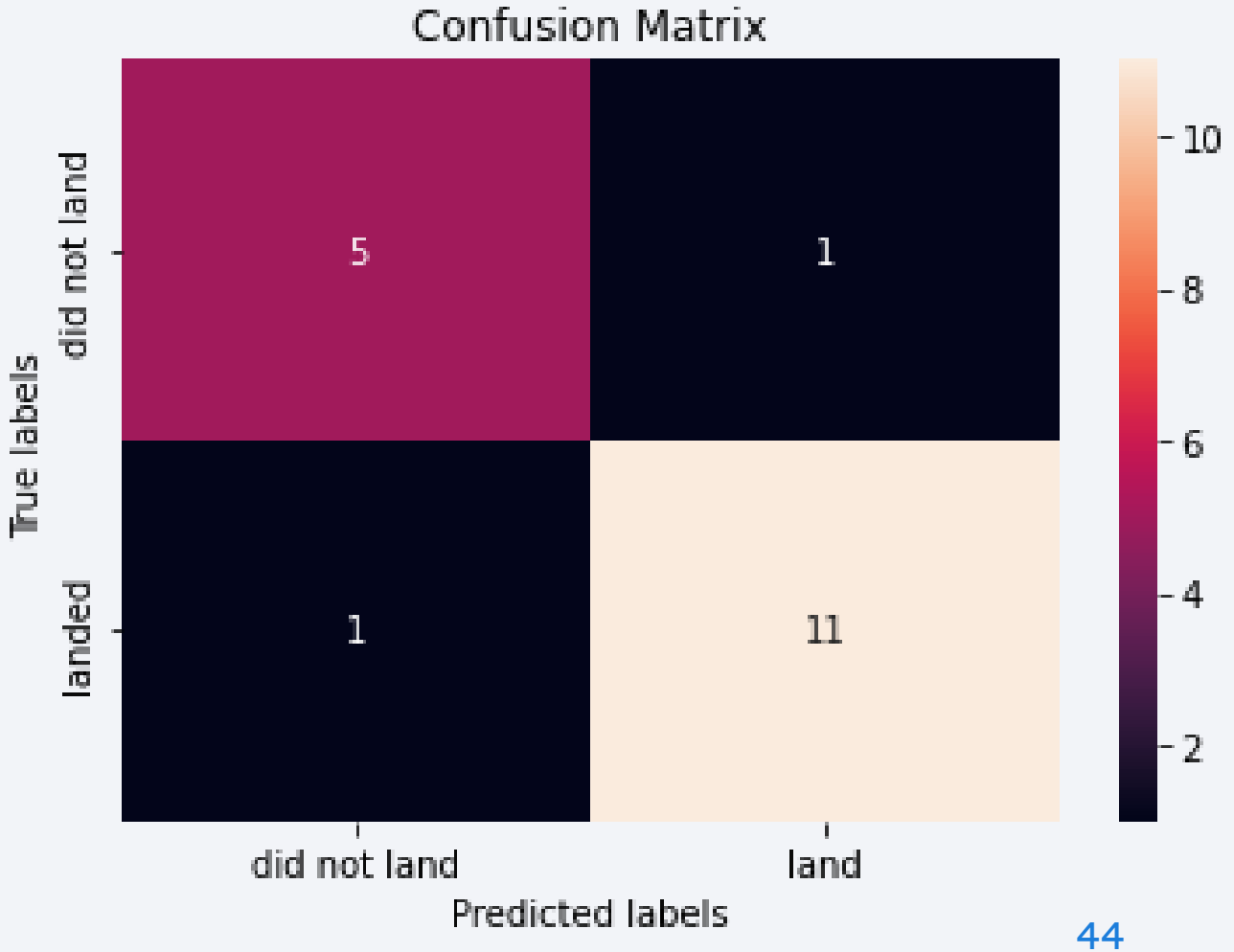


- As seen from the charts, the Decision Tree model has the best classification accuracy, both in training (87.5%) and testing (88.89%) data.

Confusion Matrix of Decision Tree model

- The confusion matrix of the decision tree model shows that the False positives and False negatives are less. Hence it is a fairly good model to use for prediction of landing outcomes.

- False positives = 5.5% (1/18)
- False negatives = 5.5% (1/18)



Conclusions

- SpaceX didn’t have a successful launch until 2013. Thereafter the success rate kept on increasing.
- Intended orbit for payloads had impact on outcome. Success rate is the highest for the orbits ES-L1, GEO, HEO and SSO.
- Basic infrastructure such as highways, rail-line are built around the launch sites. Launch sites are located fairly far from cities and closer to coastline.
- Launch site KSC LC-39A had the highest success rate for Falcon 9 rocket launches.
- Lower payload launches had higher success rate than higher payload launches.
- A decision tree model was built that can predict the launch outcome of a Falcon9 rocket launch, with an accuracy of 88.89% on test data and 87.5% on training data,

Thank you!