Winning Space Race with Data Science

JingZeng Xie
2024 Oct 26

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection and Data Wrangling
  - Interactive Visual Analytics
  - EDA with Visualization and SQL
  - Interactive Map with Folium and Plotly Dashboard
  - Model Development and Predictive Analysis
- **Summary of all results**
  - Data Processing for Analysis and Modeling
  - Explore Patterns in Data through EDA with SQL
  - Visualization Data with Folium and Plotly Dashboard
  - Use Accuracy Metrics to Evaluate Classification Models

Introduction

- **Background**
  Space exploration has typically involved high costs, limiting participation to a few organizations. However, SpaceX has significantly disrupted the industry by introducing cost-effective solutions. Notably, the Falcon 9 rocket features reusable first-stage technology, allowing multiple launches with the same hardware and significantly lowering launch costs.
- **Problem**
  This capstone project aims to investigate the factors that contribute to the successful landing of SpaceX's Falcon 9 first stage through exploratory data analysis and predictive modeling. The goal is to develop a classification model that can accurately predict landing outcomes using key observations and predictors.

Section 1

Methodology

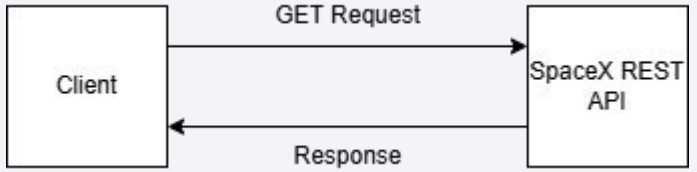
Methodology

Executive Summary

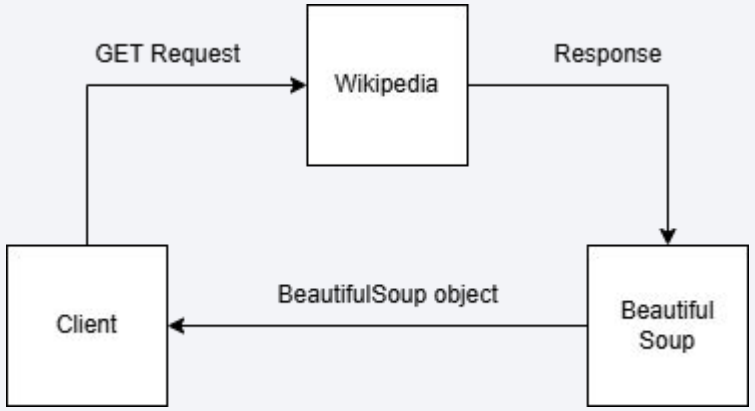
- Data collection methodology:
  - Web Scraping with SpaceX REST API
- Perform data wrangling
  - Missing Value Handling, One Hot Encoding, Target Variable Generate
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, K-NN and Decision Tree
  - Accuracy Metrics to Evaluate Train and Test Dataset

Data Collection

- Data collected from:
  - SpaceX REST API



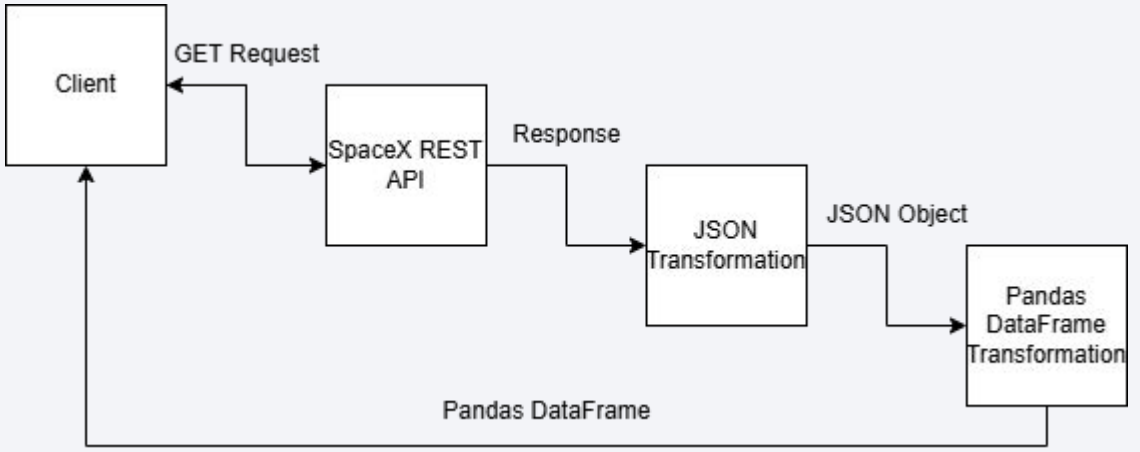
  - Wikipedia



Data Collection – SpaceX API

1. Send a GET request to the SpaceX REST API
2. Get the API response from SpaceX REST API
3. Parse the API response as a JSON object
4. Convert the JSON data into a Pandas DataFrame

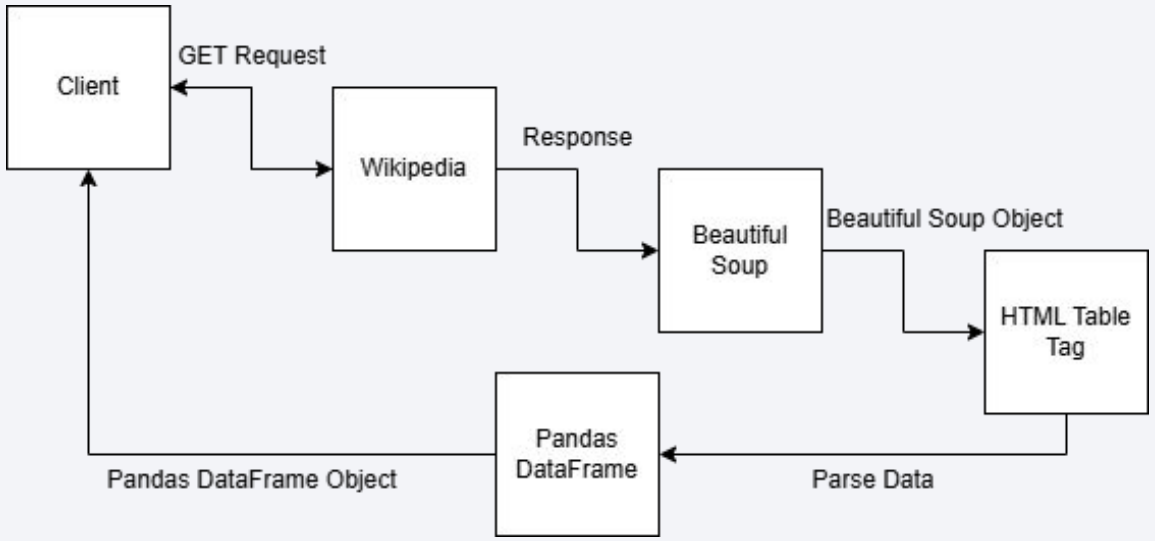
The Notebook Link in GitHub Please Click Here



Data Collection - Scraping

1. Send a GET request to the Wikipedia URL
2. Get the API response from Wikipedia URL
3. Convert the API response as a Beautiful Soup Object
4. Parse contents of HTML table tag from Beautiful Soup Object
5. Store the data into a Pandas DataFrame

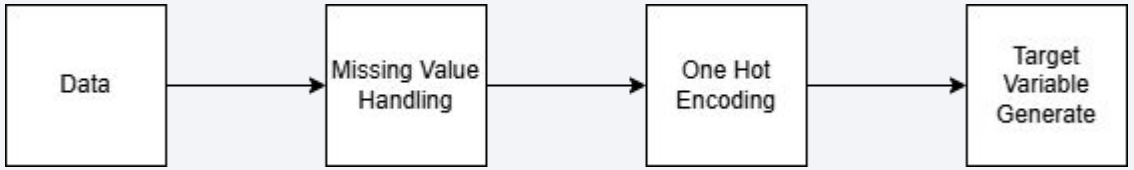
The Notebook Link in GitHub Please Click Here



Data Wrangling

1. Detect duplicate values and perform imputation
2. Convert categorical variables to numerical variables
3. Identify patterns and generate target variables by specifying features

The Notebook Link in GitHub Please Click Here



EDA with Data Visualization

- Categorical Plot
  - Categorical plots let us see how different categories (like launch sites or orbits) affect the outcome, such as landing success. They help identify patterns or differences between groups, making it easy to spot trends.
- Bar Plot
  - Bar plots are great for comparing the success rate or count across specific groups, like different orbits. They clearly show which categories perform better or worse, which is useful for comparing results.
- Line Plot
  - Line plots show changes over time, such as how the landing success rate has shifted by year. They're ideal for spotting trends and seeing if things are getting better, worse, or staying the same.

The Notebook Link in GitHub Please Click Here

EDA with SQL

- Distinct Statement
- String Matching (%)
- Functions as SUM, COUNT, AVG, MAX, MIN, etc.
- Multiple Conditions as OR and AND
- Group By Statement
- Range Conditions as BETWEEN
- Sorting Statement as ORDER

The Notebook Link in GitHub Please Click Here

Build an Interactive Map with Folium

- Circles: Represent launch sites on the map
- Markers: Display the names of launch sites as pop-ups
- Marker Cluster: Show individual launch records for each site, color-coded by launch outcome
- Mouse Position: Display latitude and longitude information based on mouse location on the map
- Lines: Connect launch sites to nearby landmarks, like coastlines, highways, railroads, or major cities

The Notebook Link in GitHub Please Click Here

Build a Dashboard with Plotly Dash

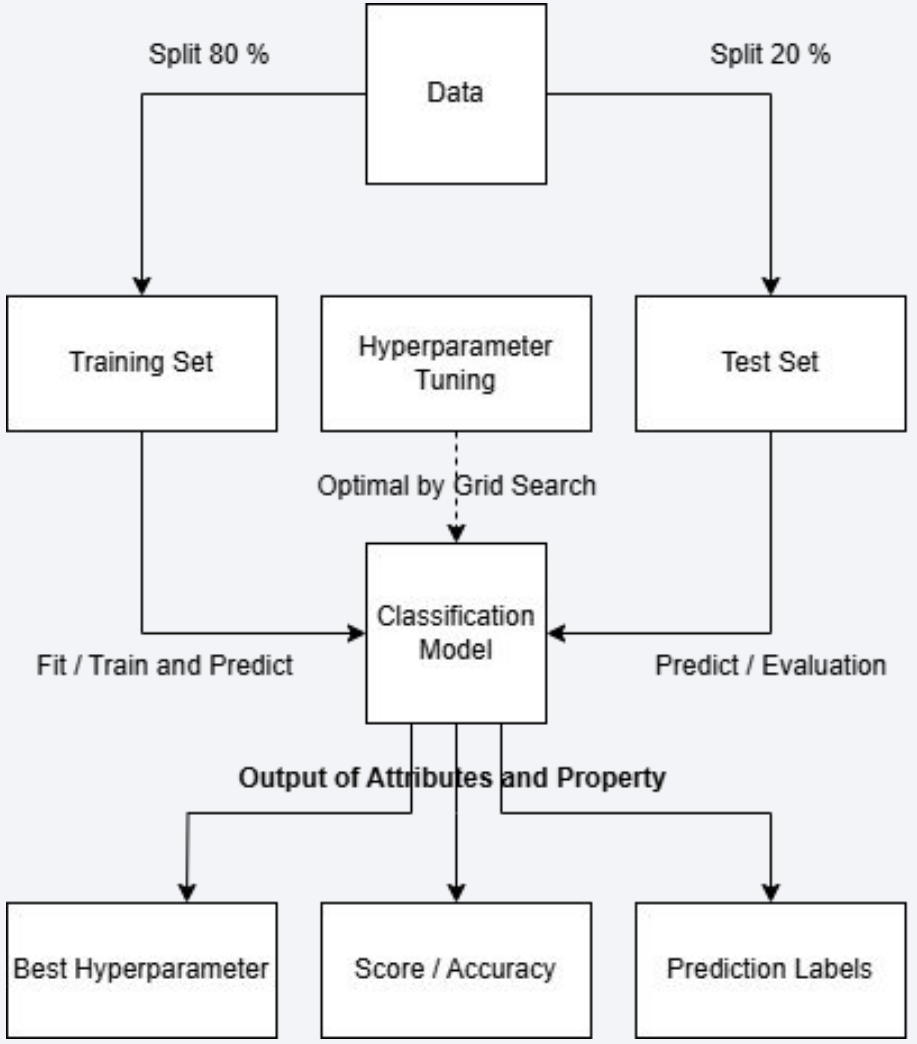
- Pie Chart
  - A pie chart shows how parts make up a whole. In EDA, it helps us see the proportion of categories, like how many landings were successful vs. unsuccessful.
- Categorical Plot
  - Categorical plots help us compare different groups. For example, they show how launch sites or orbits impact landing outcomes, making it easier to spot patterns and differences.
- Range Slider
  - A range slider lets us focus on a specific range of data, like a certain time period. It's useful for exploring changes over time without needing to look at everything all at once.

The Notebook Link in GitHub Please Click Here

Predictive Analysis (Classification)

1. Split dataset into Training and Test sets, which 80% and 20%.
2. Hyperparameter Tuning with Optimal parameter by Grid Search.
3. Use Training set to fitting the Classifier model.
4. Use Test set to Generate the Prediction Label/Target during Classifier model after training.

The Notebook Link in GitHub Please Click Here

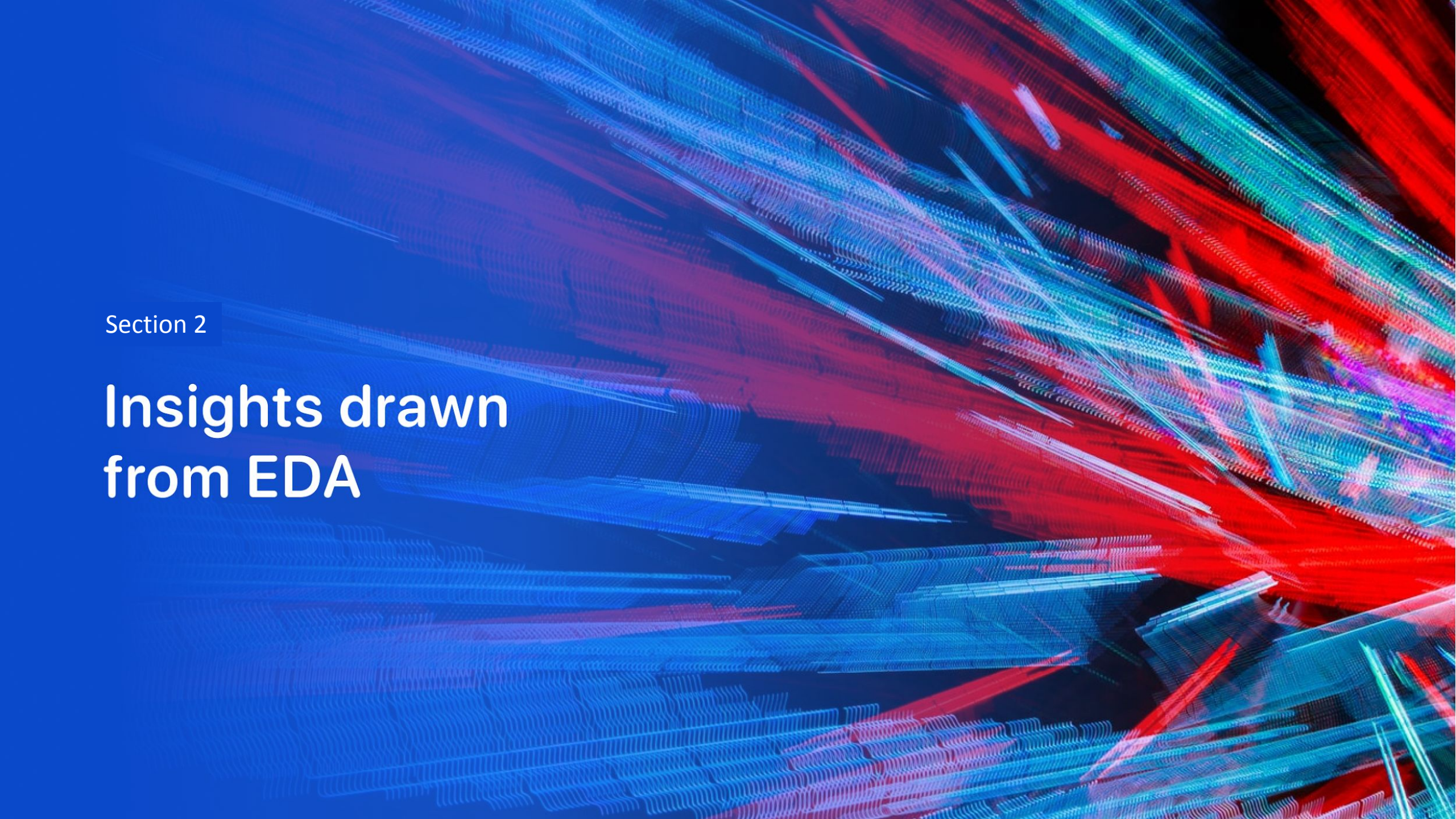


Results

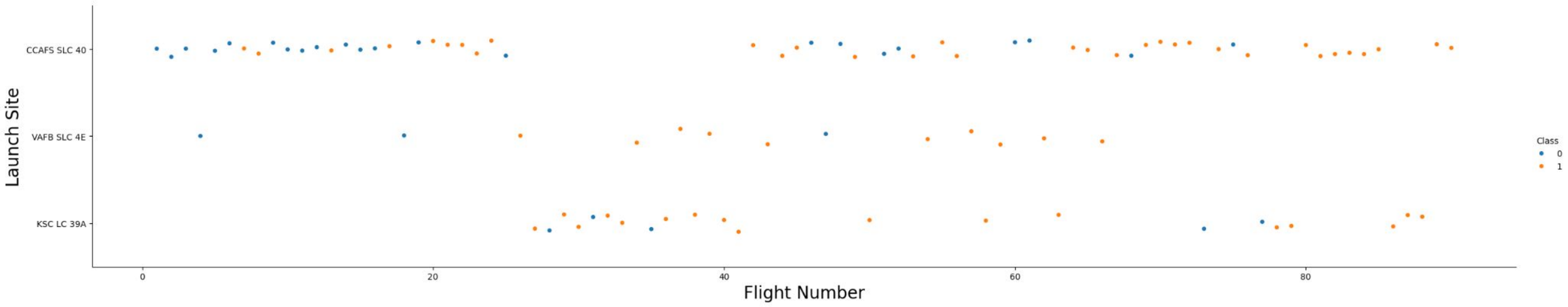
- Exploratory data analysis results
- Interactive analytics demo in screenshots
  - Interactive Visual Analytics with Folium lab
  - Interactive Dashboard with Ploty Dash
- Predictive analysis results

Section 2

Insights drawn from EDA

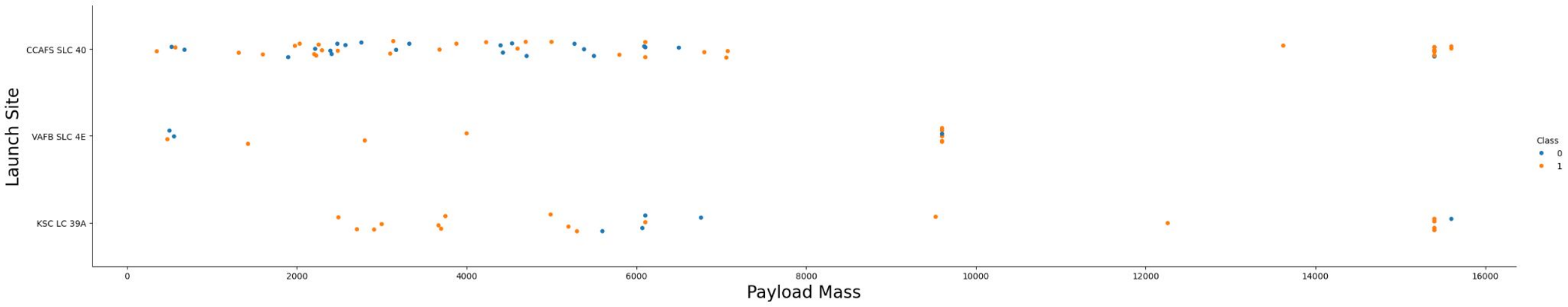


Flight Number vs. Launch Site



From the graph we can observe that the probability of successful landing increases with the number of flights. And the number of flights at CCAFS SLC 40 and KSC LC 39A far exceeds that of VAFB SLC 4E.

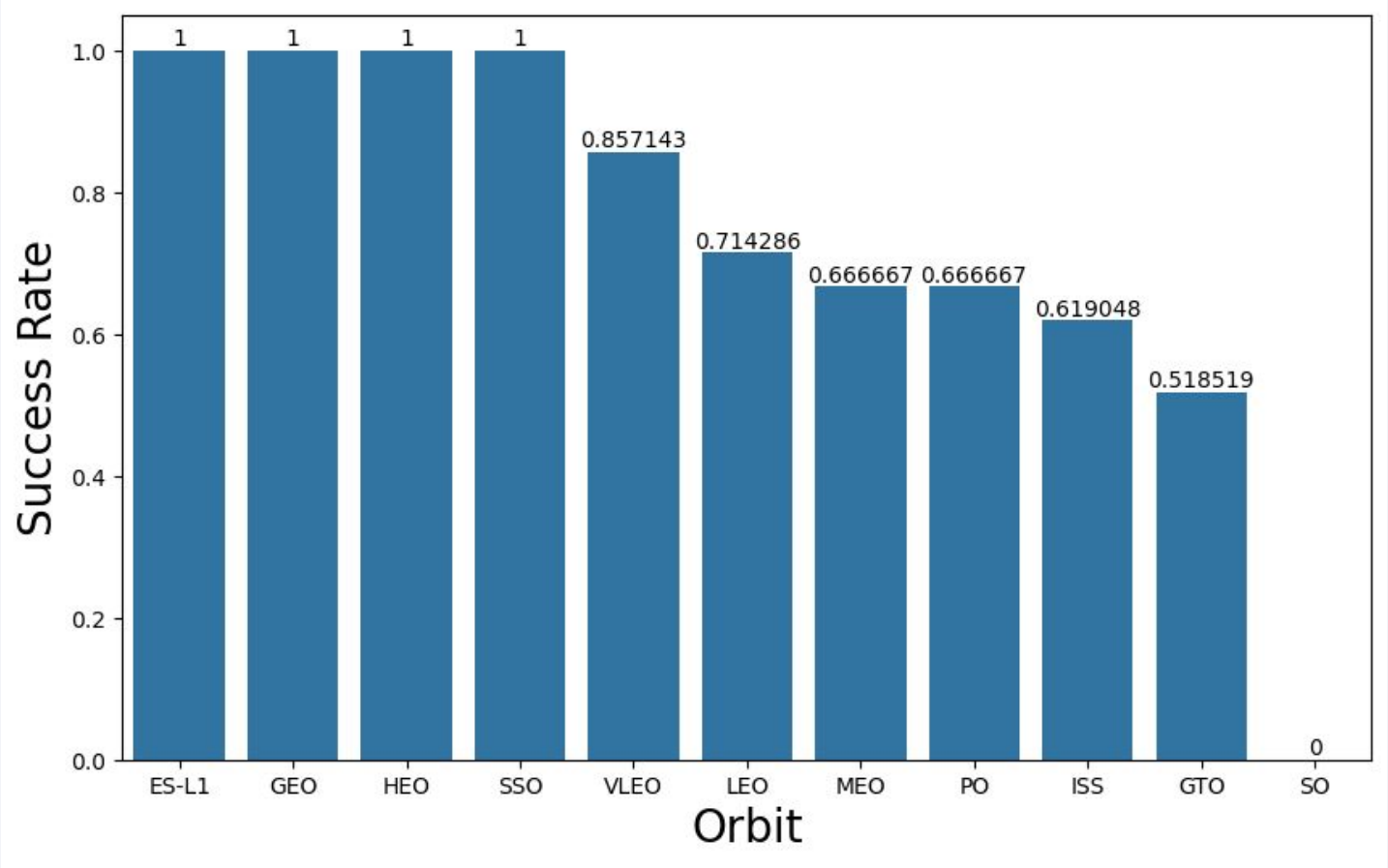
Payload vs. Launch Site



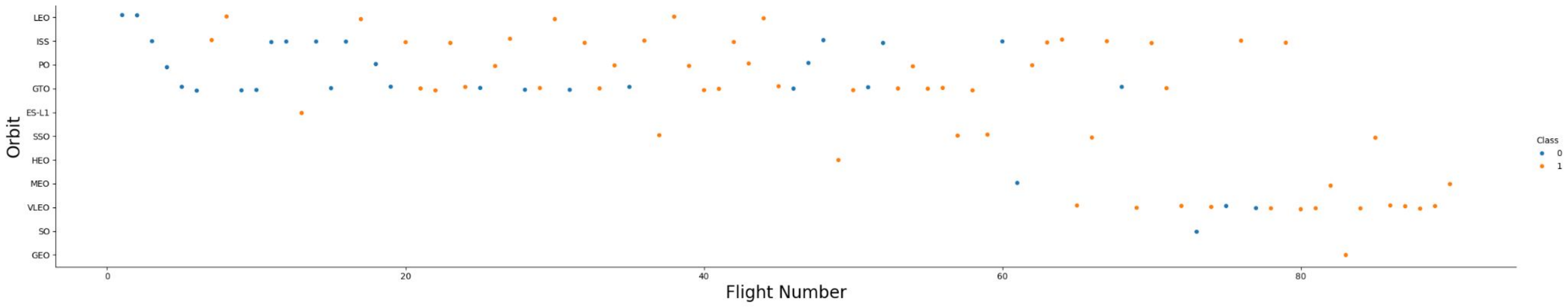
From the figure, we can observe that the successful landing rate is the highest in the payload mass range above 8000.

Success Rate vs. Orbit Type

From the figure we can observe that the successful landing rate is the highest among ES-L1, GEO, HEO and SSO orbits.

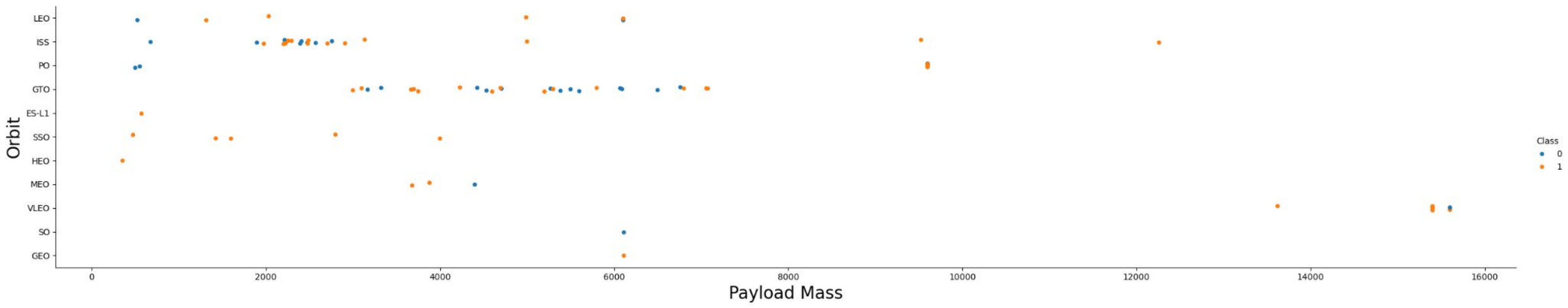


Flight Number vs. Orbit Type



From the figure, we can see that the number of flights increases, it can be clearly seen that the successful landing rate is also increasing. But the samples of ES-L1, HEO, SO and GEO orbits are all 1, which lacks credibility.

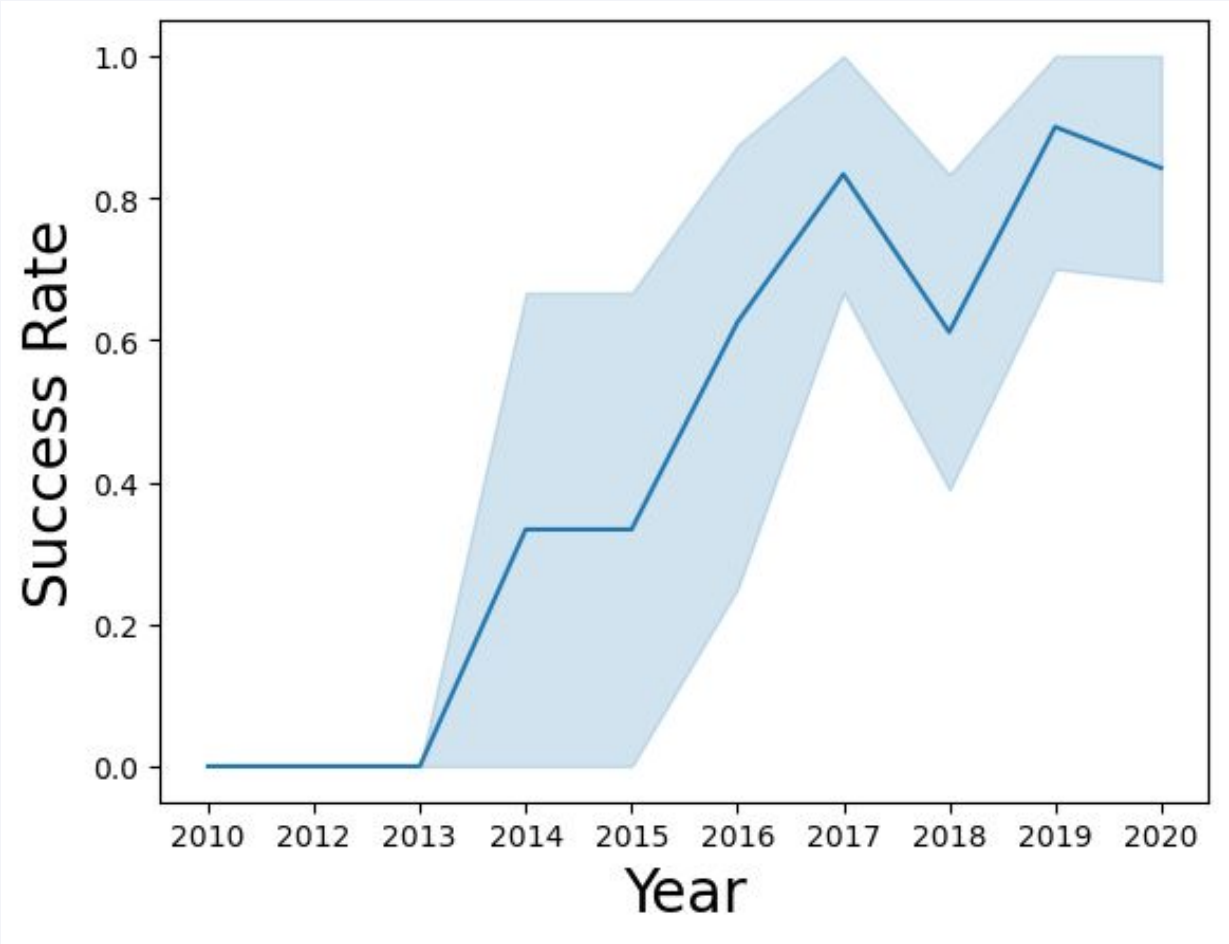
Payload vs. Orbit Type



From the figure, we can observe that the successful landing rate is the highest in the payload mass range above 8000 for ISS, PO and VLEO orbits.

Launch Success Yearly Trend

From the figure, we can observe that as the years increase, it is obvious that the successful landing rate is also increasing.



All Launch Site Names

The unique launch sites in the space mission:

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

Task 1

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

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

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

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

Launch Site Names Begin with 'CCA'

Task 2

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

```
[ ] %%sql
    SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5;
```

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

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

There are 5 records of launch sites starting with "CCA".

Total Payload Mass

Task 3

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

```
[ ]  %%sql
     SELECT  Customer,  SUM(PAYLOAD_MASS__KG_)  AS  "total  payload  mass"  FROM  SPACEXTABLE  WHERE  Customer  LIKE  "%NASA  (CRS)%";
```

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

Customer	total payload mass
NASA (CRS)	48213

The total payload mass carried by boosters launched by NASA (CRS) is 48213.

Average Payload Mass by F9 v1.1

Task 4

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

```
[ ] %%sql
    SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) AS "average payload mass" FROM SPACEXTABLE WHERE Booster_Version LIKE "F9 v1.1%";
```

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

Booster_Version	average payload mass
F9 v1.1 B1003	2534.6666666666665

The average payload mass carried by booster version F9 v1.1 is 2534.666666666665

First Successful Ground Landing Date

```
[ ] %%sql
    SELECT MIN(DATE), Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome LIKE "Success (ground pad)";
```

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

MIN(DATE)	Landing_Outcome
2015-12-22	Success (ground pad)

The first successful landing was in 2015-12-22.

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

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

There are 4 records of successful drone ship landing with payload between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[ ] %%sql
    SELECT Mission_Outcome, COUNT(Mission_Outcome) AS "number of mission outcomes" FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

Mission_Outcome	number of mission outcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

There are 100 records of successful landing.

Boosters Carried Maximum Payload

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

- 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

Task 8

List the names of the booster_versions which have carried the maximum p

```
[ ] %%sql
    SELECT  Booster_Version,  PAYLOAD_MASS__KG_  FROM  SPACEXTABLE  WHERE
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

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.

```
[ ] %%sql
    SELECT SUBSTR(Date, 6, 2) AS MONTH, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE SUBSTR(Date, 0, 5) = "2015" AND Landing_Outcome = "Failure (drone ship)";
```

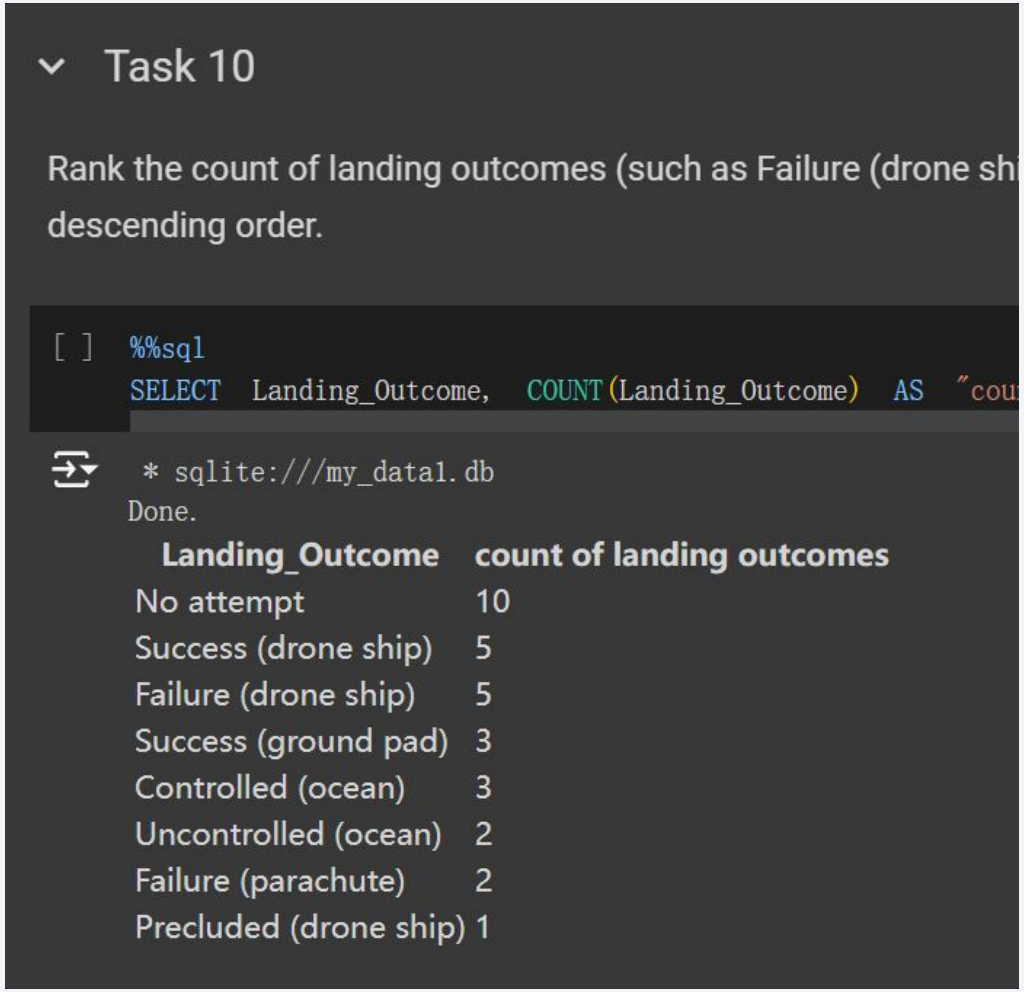
```
 * sqlite:///my_data1.db
Done.
```

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

There were two failed landings recorded in 2015, both at the CCAFS LC-40 launch site: the F9 v1.1 B1012 booster version in January and the F9 v1.1 B1015 booster version in April.

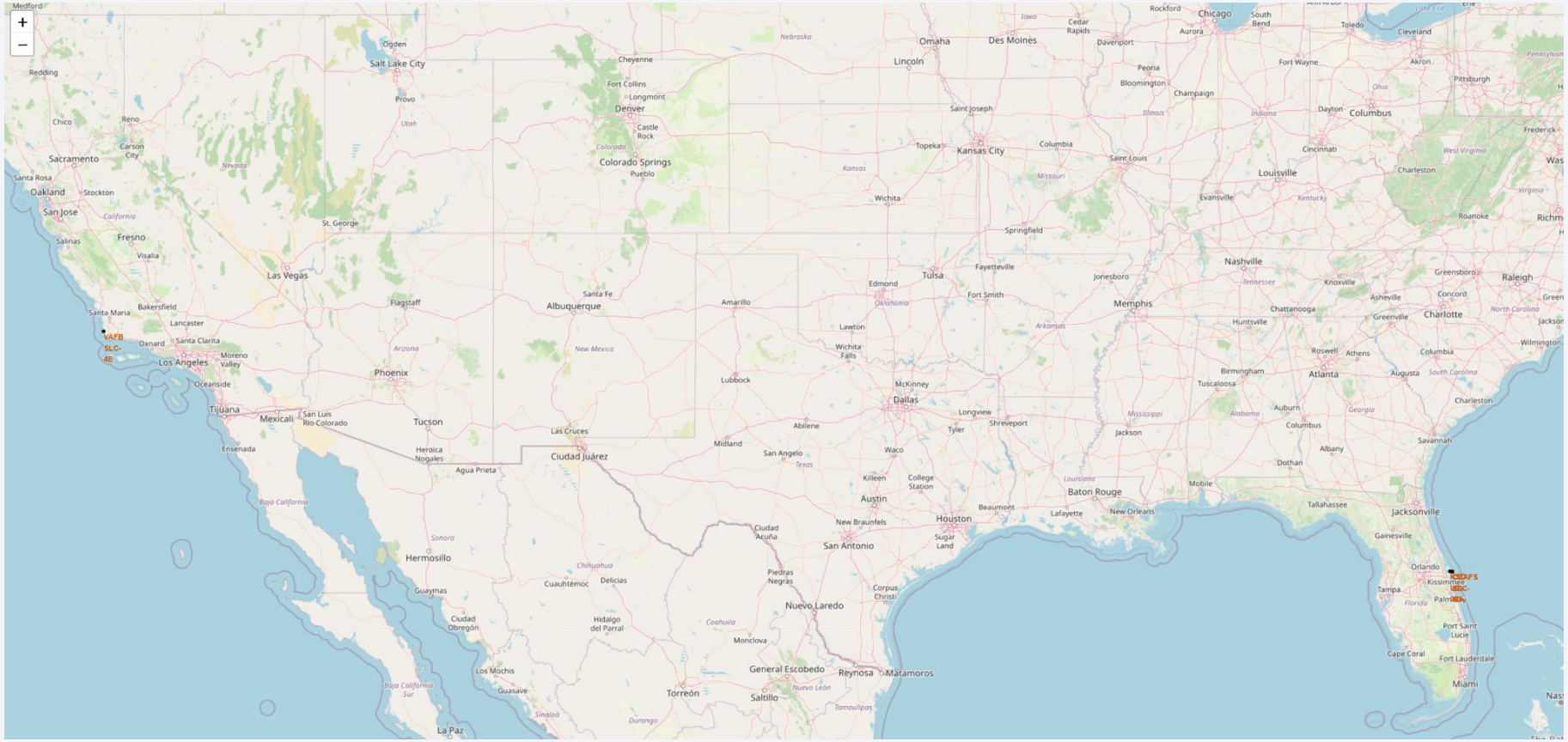
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

There are 8 types of landing outcomes between the dates 2010-06-04 and 2017-03-20.



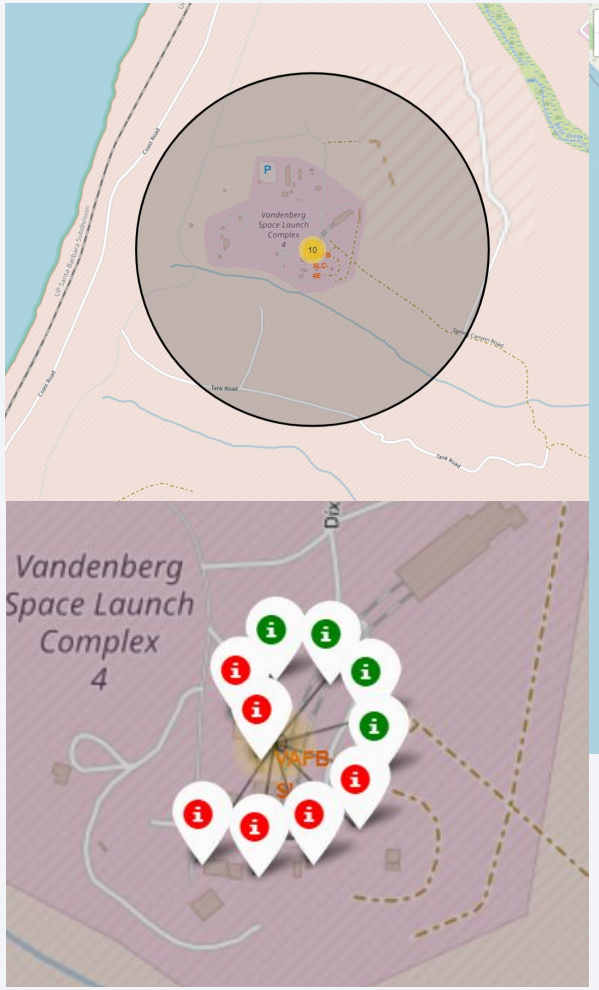

Section 3
Launch Sites
Proximities Analysis

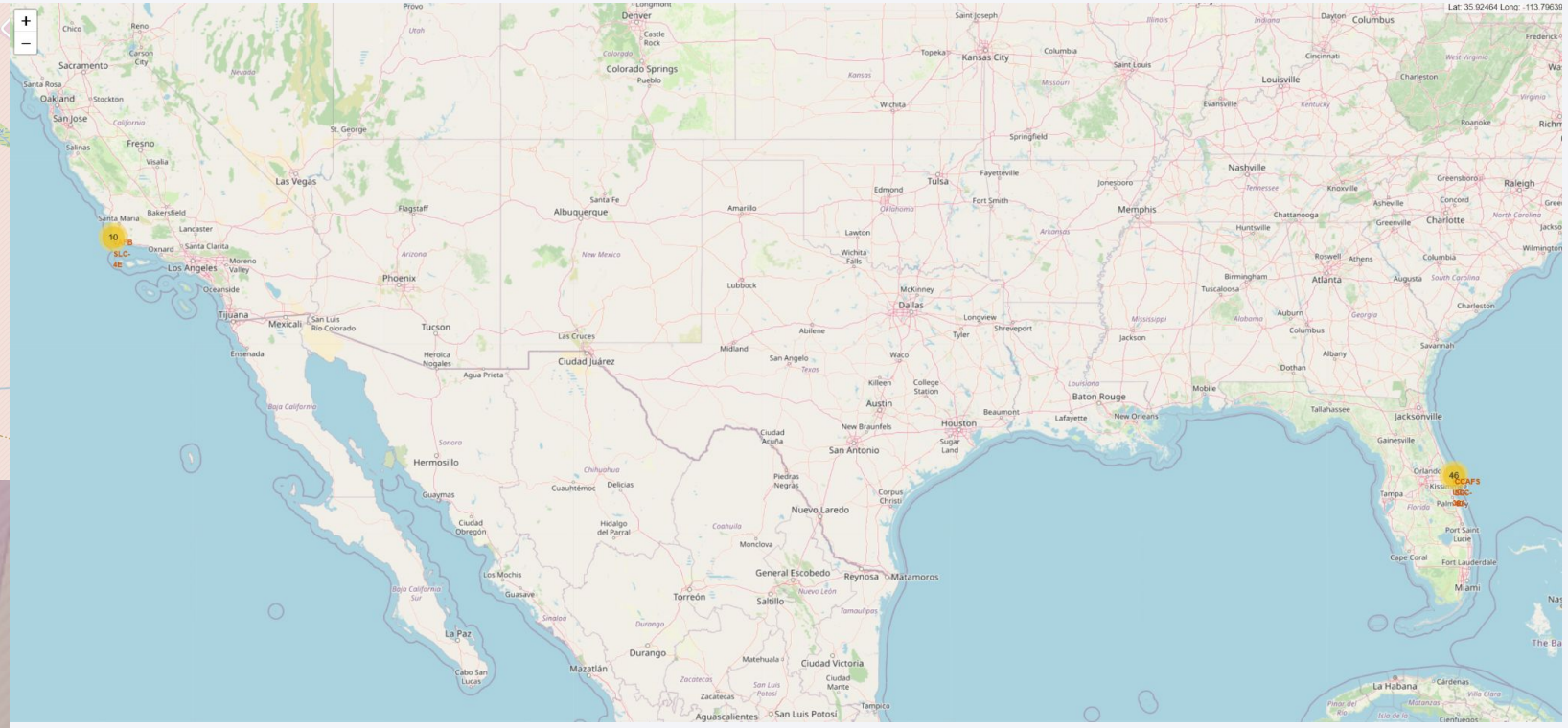
The Launch Site on the Site Map



West Coast: VAFB SLC-4E
East Coast: CCAFS LC-40, KSC LC-39A and CCAFS SLC-40

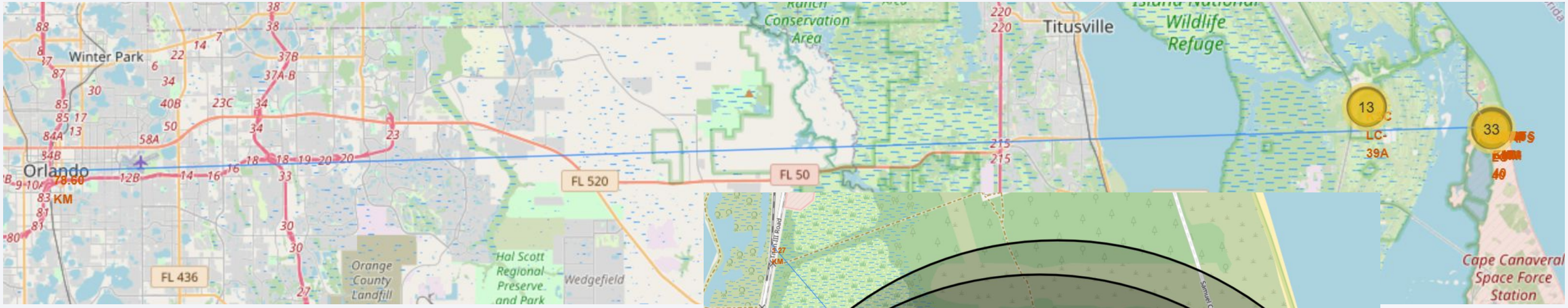
The Launch Result on the Site Map

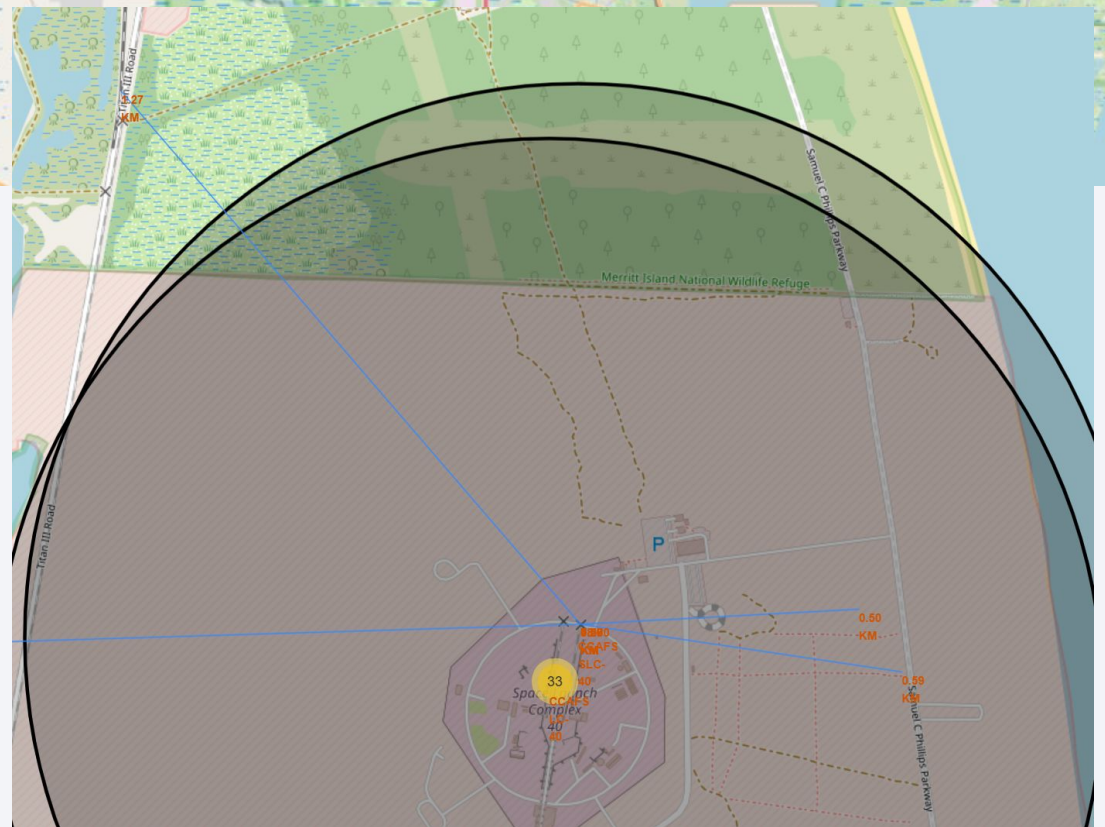




There are 10 records in west coast, and 46 records in east coast.

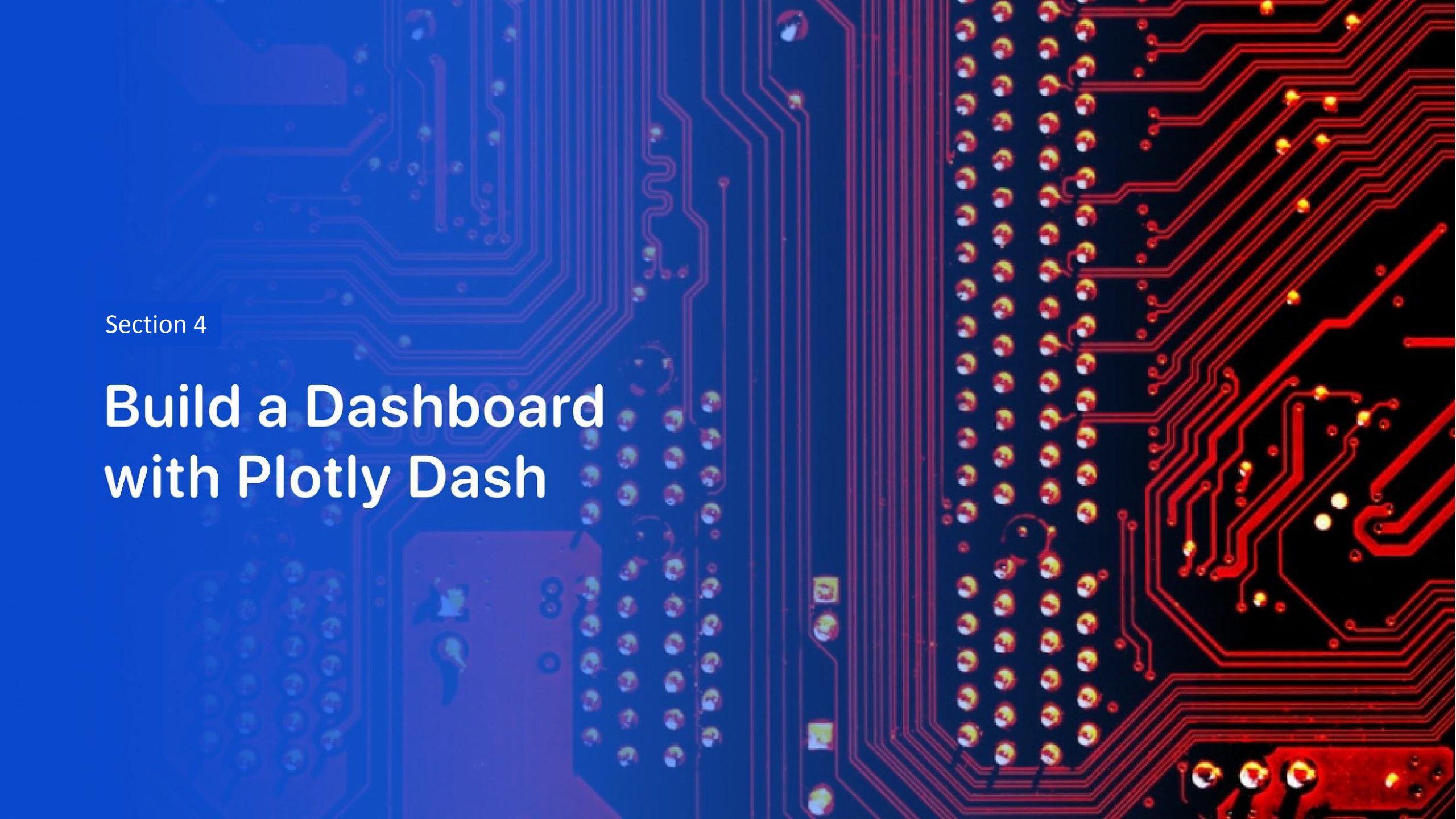
Near City, Railway, Highway and Coastline



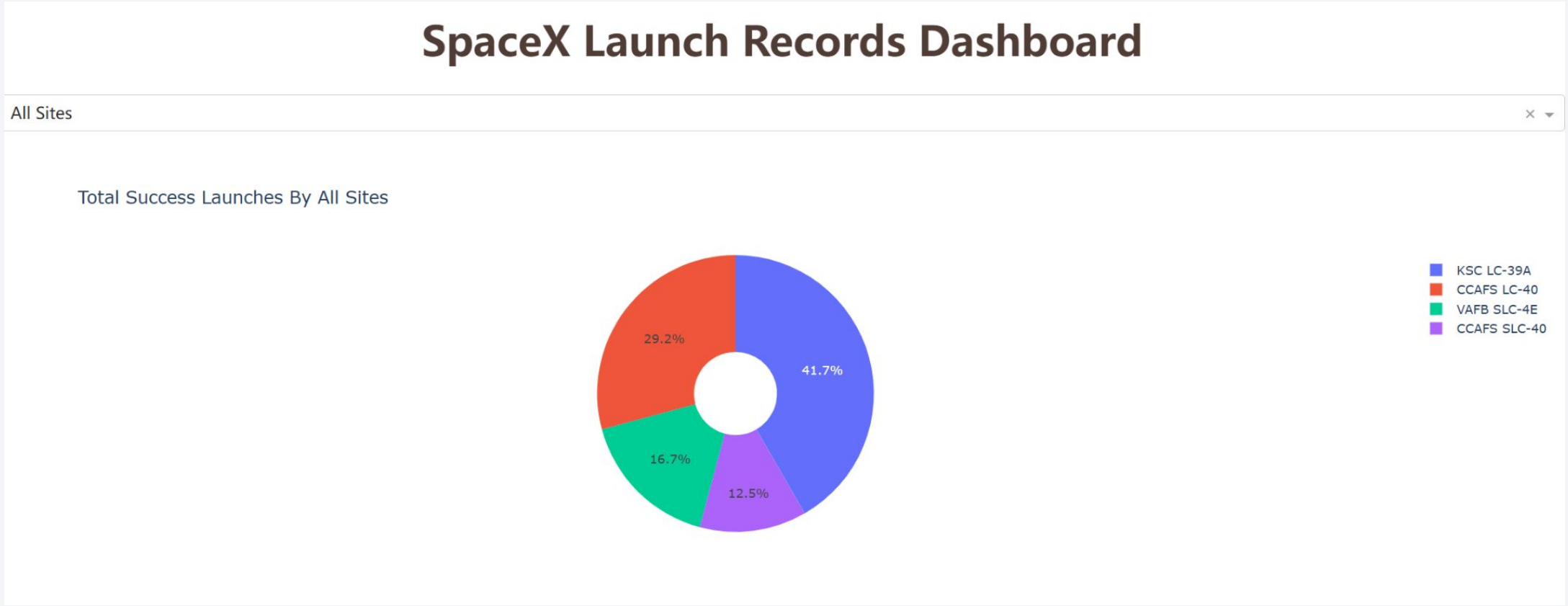


The distance between launch site to:

- Near City: 78.60 KM
- Railway: 0.59 KM
- Highway: 1.27 KM
- Coastline: 0.50 KM

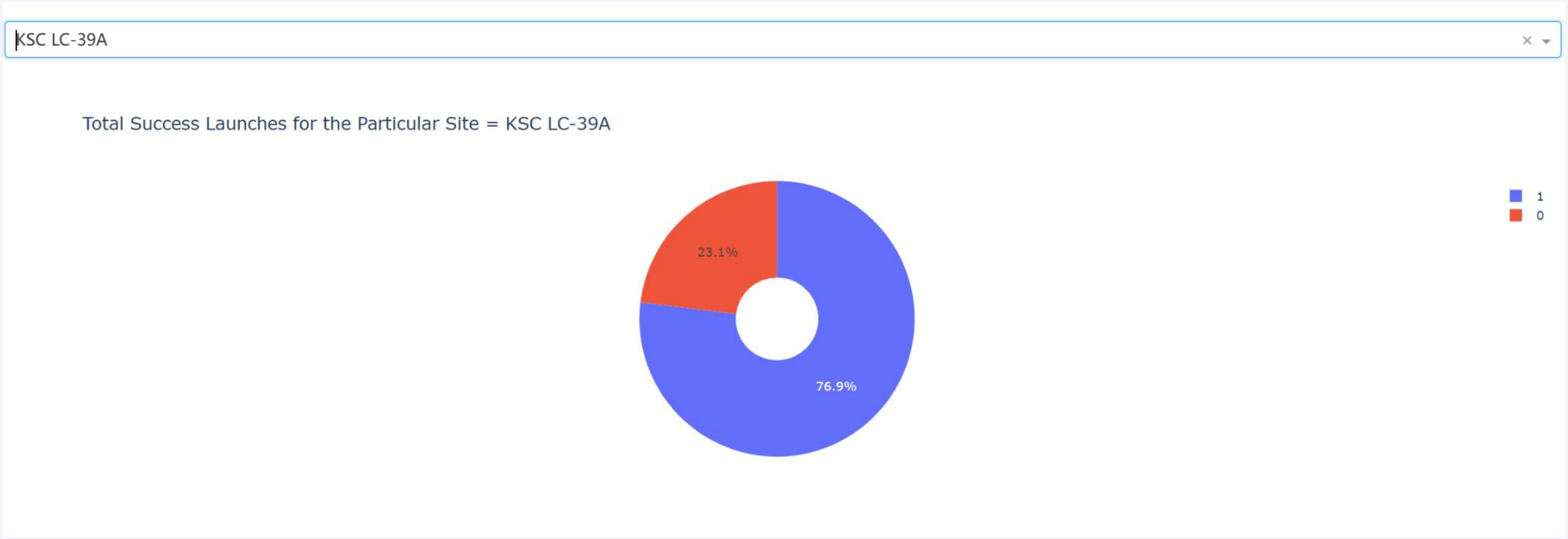

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



The success ratio:
KSC LC-39A: 41.7%
CCAFS LC-40: 29.2%
VAFB SLC-4E: 16.7%
CCAFS SLC-40: 12.5%

Launch Site with Highest Launch Success Ratio



The highest launch success ratio is KSC LC-39A, which is 76.9%, the other launch success ratio are CCAFS LC-40 with 73.1%, VAFB SLC-4E with 60% and CCAFS SLC-40 with 57.9%.

Payload vs. Launch Outcome for All Sites Between 2000 Kg to 8000 Kg

It can be observed from the figure that after the payload mass exceeds 6000, the failure landing rate increases significantly.

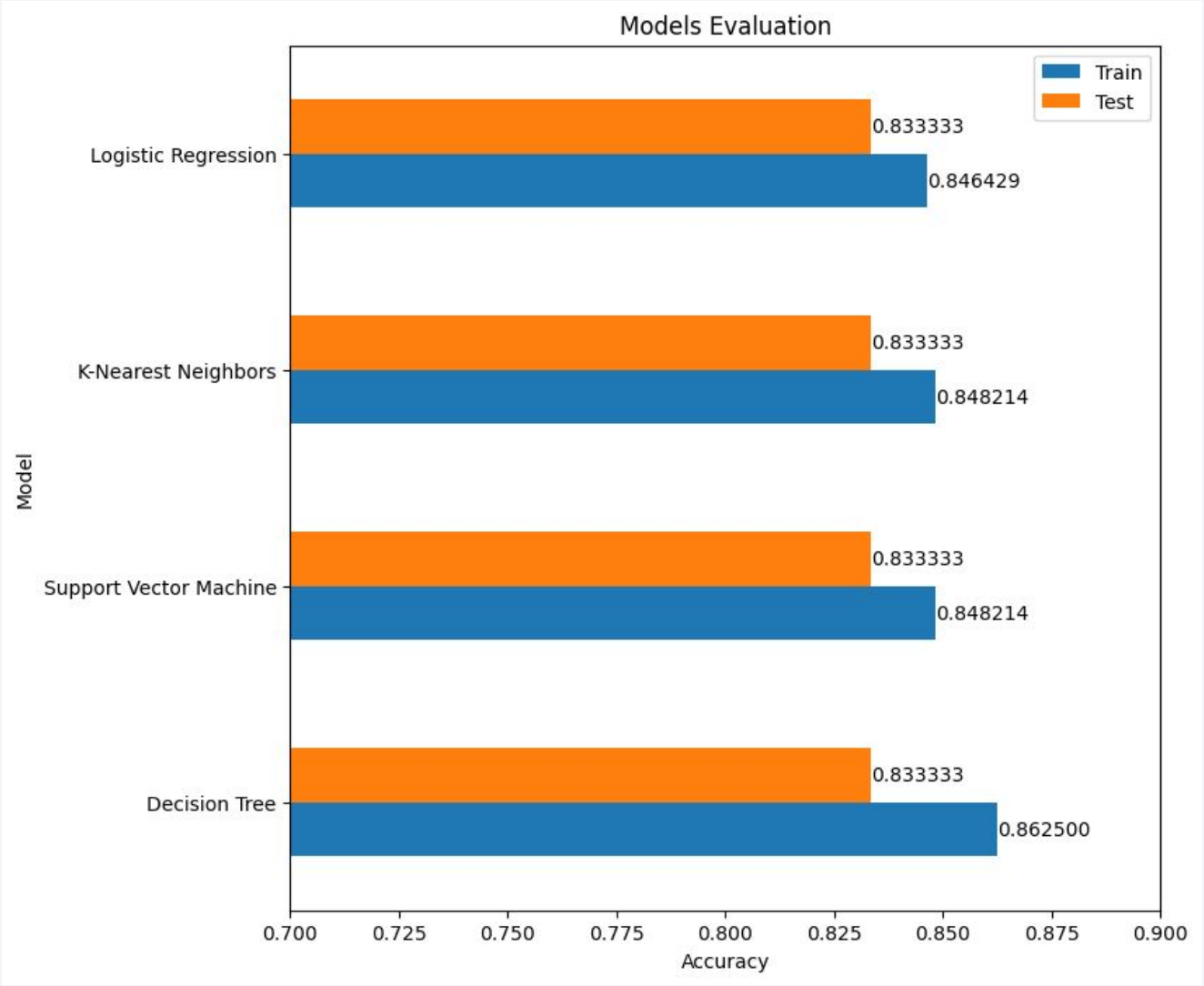
Section 5

Predictive Analysis (Classification)

Classification Accuracy

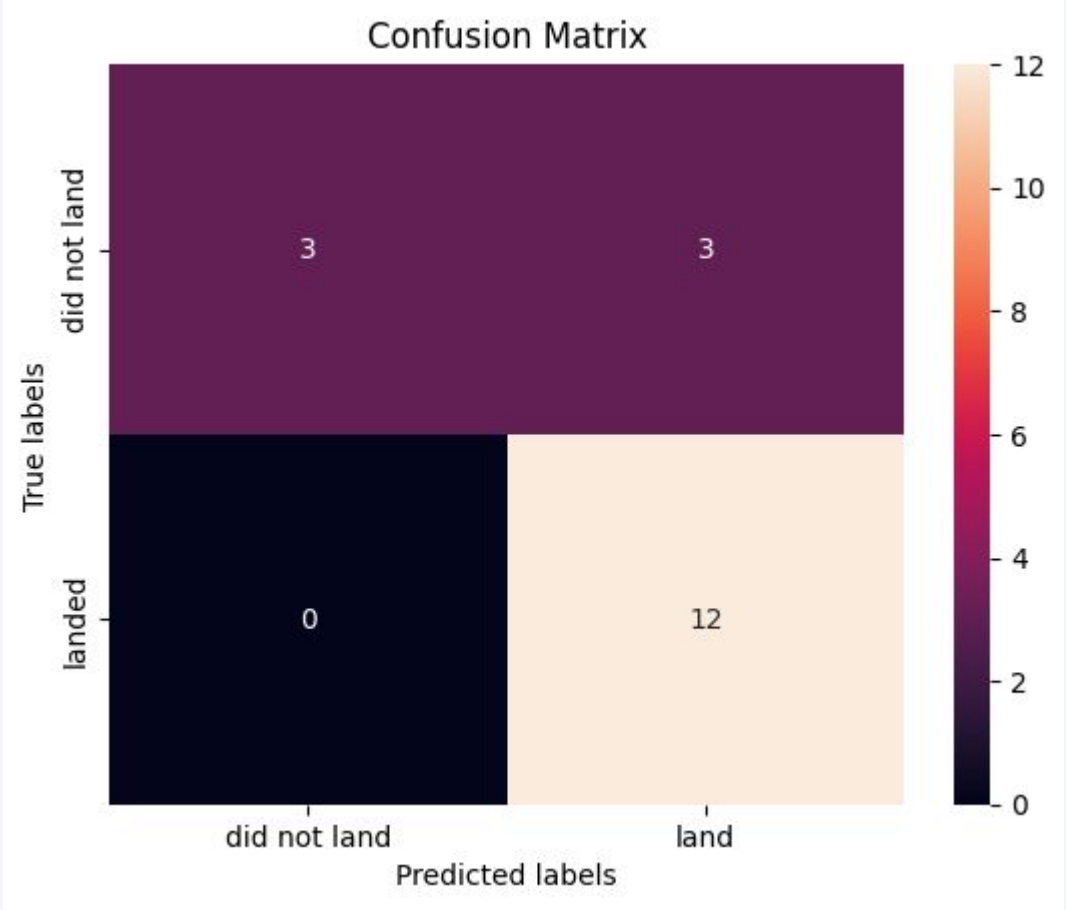
Since the performance of all models on Test set are the same, but the accuracy of Decision Tree model is the highest on Training set.

Therefore, the most performance model for this project is **Decision Tree**.



Confusion Matrix

There are 15 correct prediction out of 18 records on Test set. But 3 error/failed prediction when the landing is failed, but predict with successful.



Conclusions

- Launches with lower payload masses tend to have higher success rates compared to those with larger payloads.
- Over time, the success rate of launches has steadily improved.
- Among the launch sites, KSC LC-39A shows the highest success rate.
- The ES-L1, GEO, HEO, and SSO orbits achieve a 100% success rate.
- The Decision Tree model was selected as the most suitable model for this project.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Generating Maps with Python

Thank you!