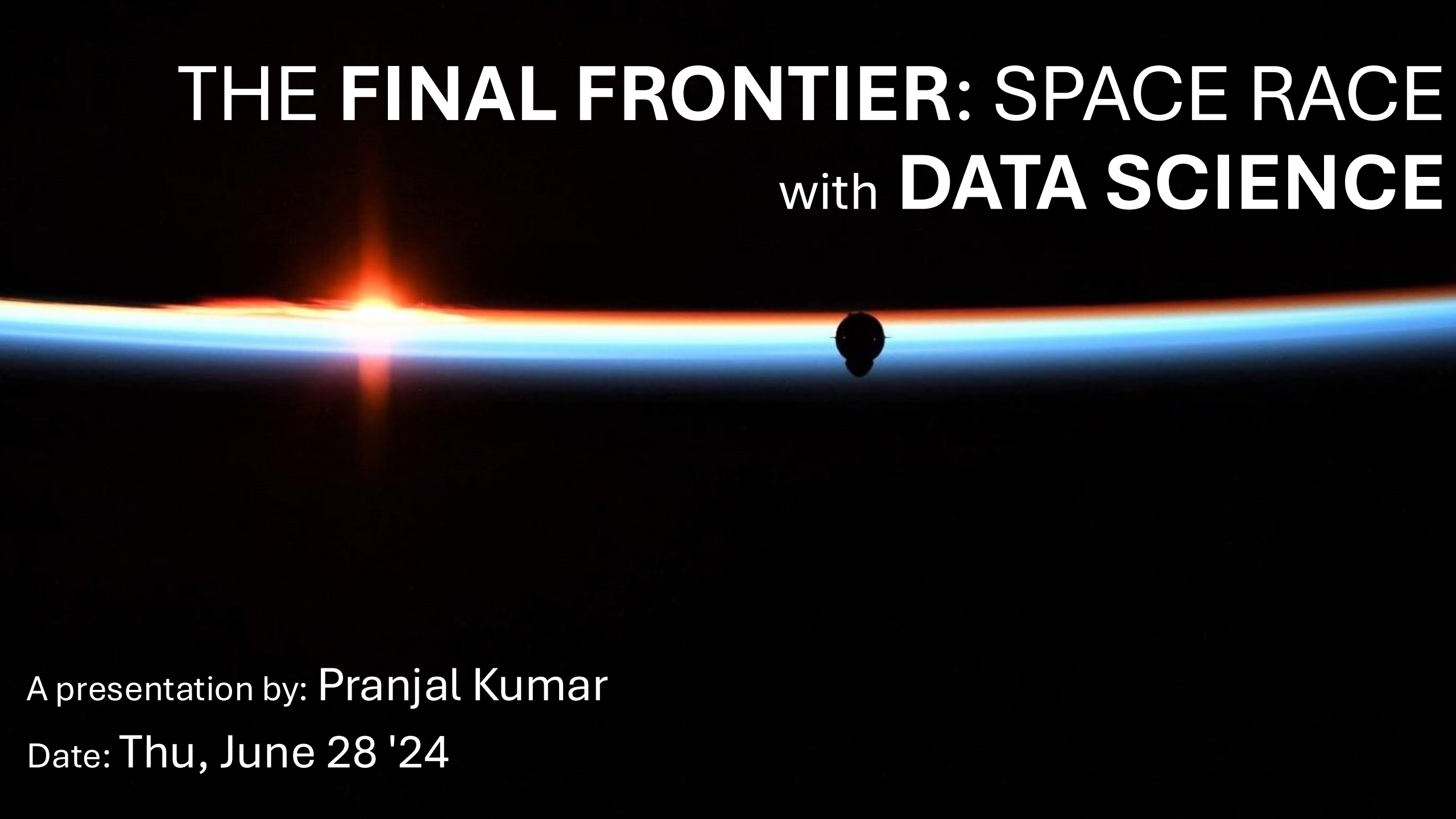
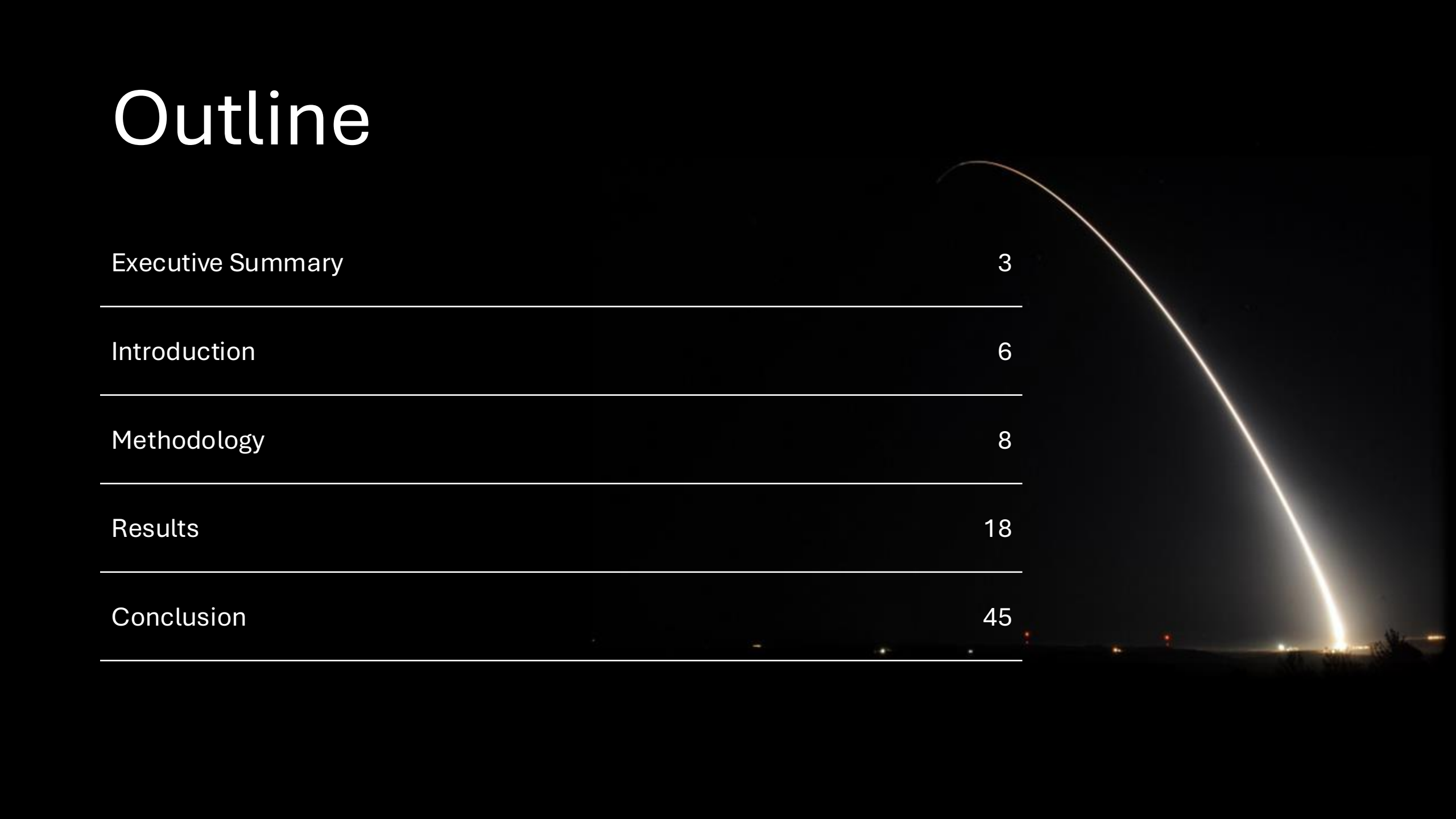


THE **FINAL FRONTIER**: SPACE RACE with **DATA SCIENCE**

A presentation by: Pranjal Kumar

Date: Thu, June 28 '24

Outline



Executive Summary

Summary of Methodologies:

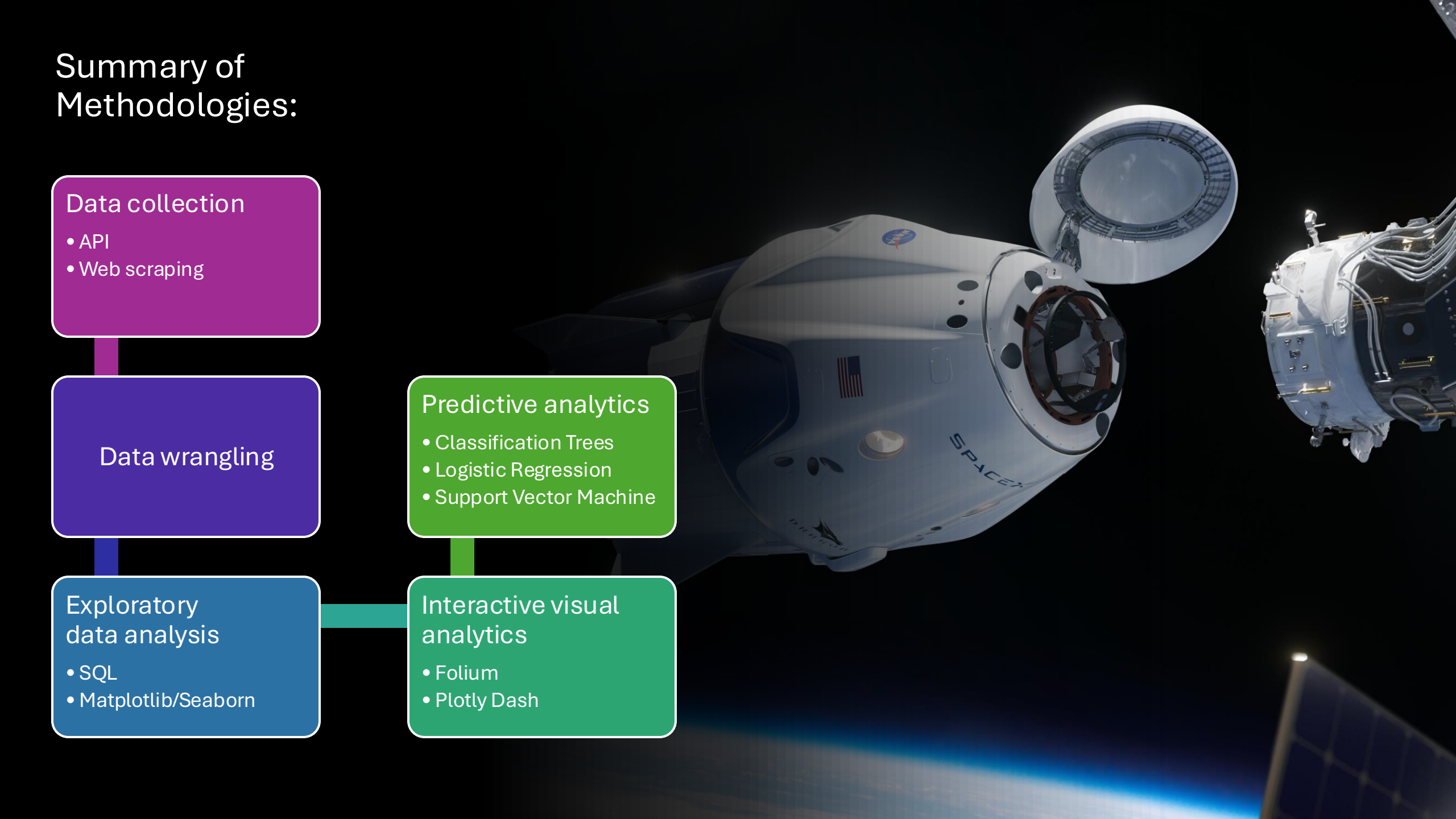
Data collection
- API
- Web scraping

Data wrangling

Exploratory data analysis
- SQL
- Matplotlib/Seaborn

Predictive analytics
- Classification Trees
- Logistic Regression
- Support Vector Machine

Interactive visual analytics
- Folium
- Plotly Dash



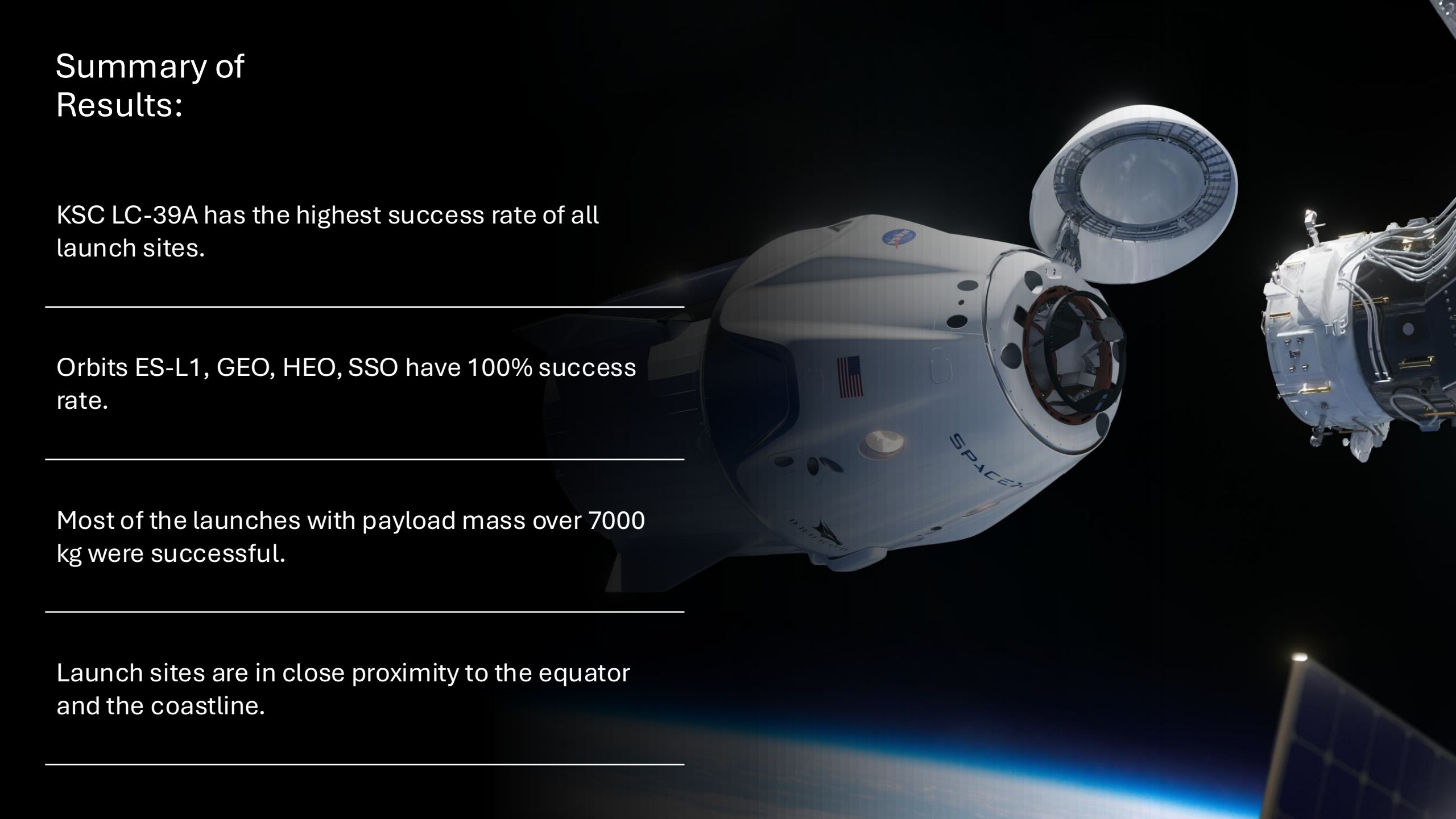
Summary of Results:

KSC LC-39A has the highest success rate of all launch sites.

Orbits ES-L1, GEO, HEO, SSO have 100% success rate.

Most of the launches with payload mass over 7000 kg were successful.

Launch sites are in close proximity to the equator and the coastline.



Introduction

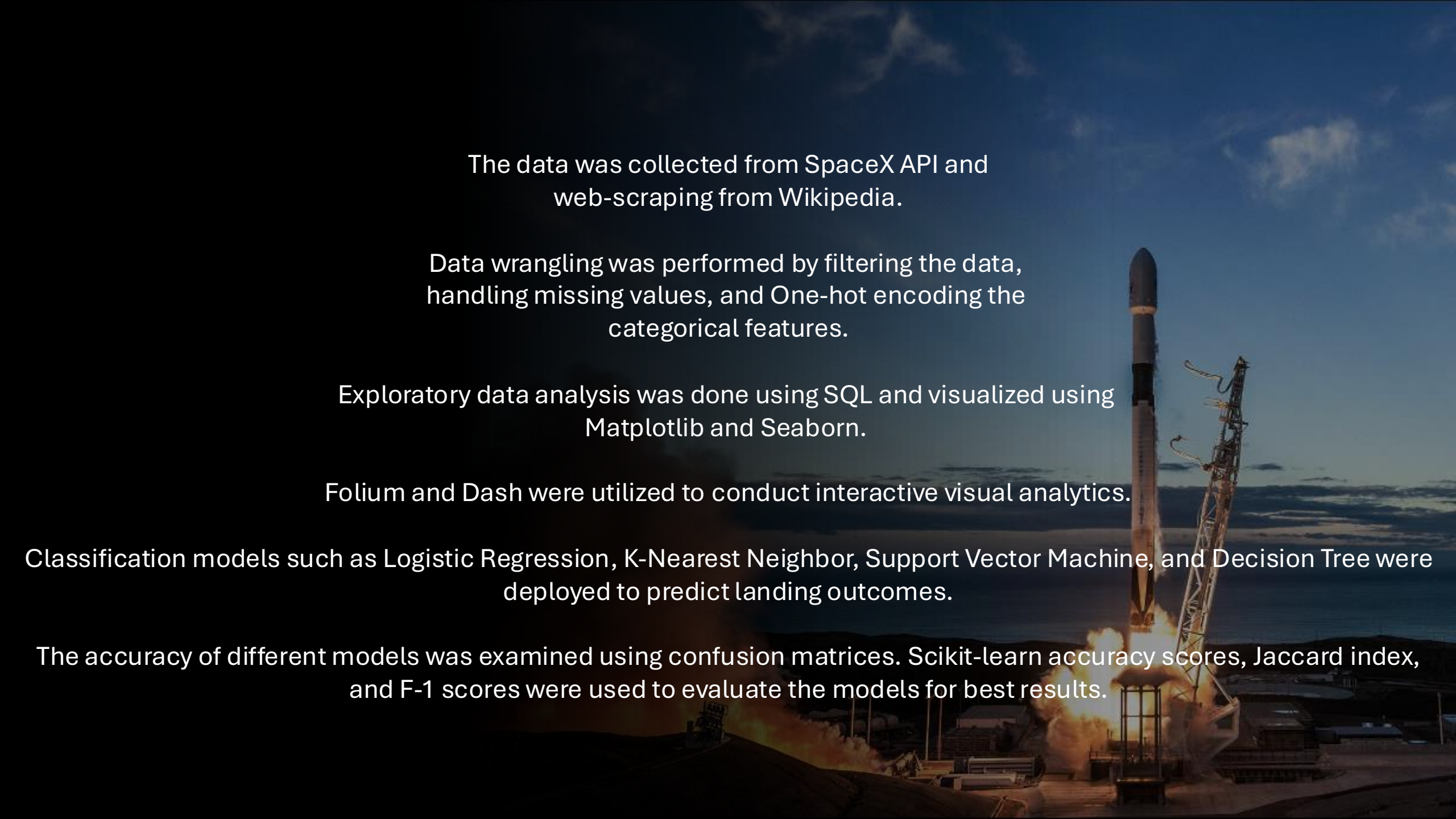
Project Background

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this report, we will use public data and machine learning to make predictions on whether SpaceX can successfully land their first stage for re-use.

Questions to be Answered

- How do factors like payload mass, launch site, number of flights, and orbit affect the success of the first stage landing?
- Does the rate of success increase over time?
- What is the best predictive model for binary classification for this case?

Methodology



The data was collected from SpaceX API and web-scraping from Wikipedia.

Data wrangling was performed by filtering the data, handling missing values, and One-hot encoding the categorical features.

Exploratory data analysis was done using SQL and visualized using Matplotlib and Seaborn.

Folium and Dash were utilized to conduct interactive visual analytics.

Classification models such as Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Decision Tree were deployed to predict landing outcomes.

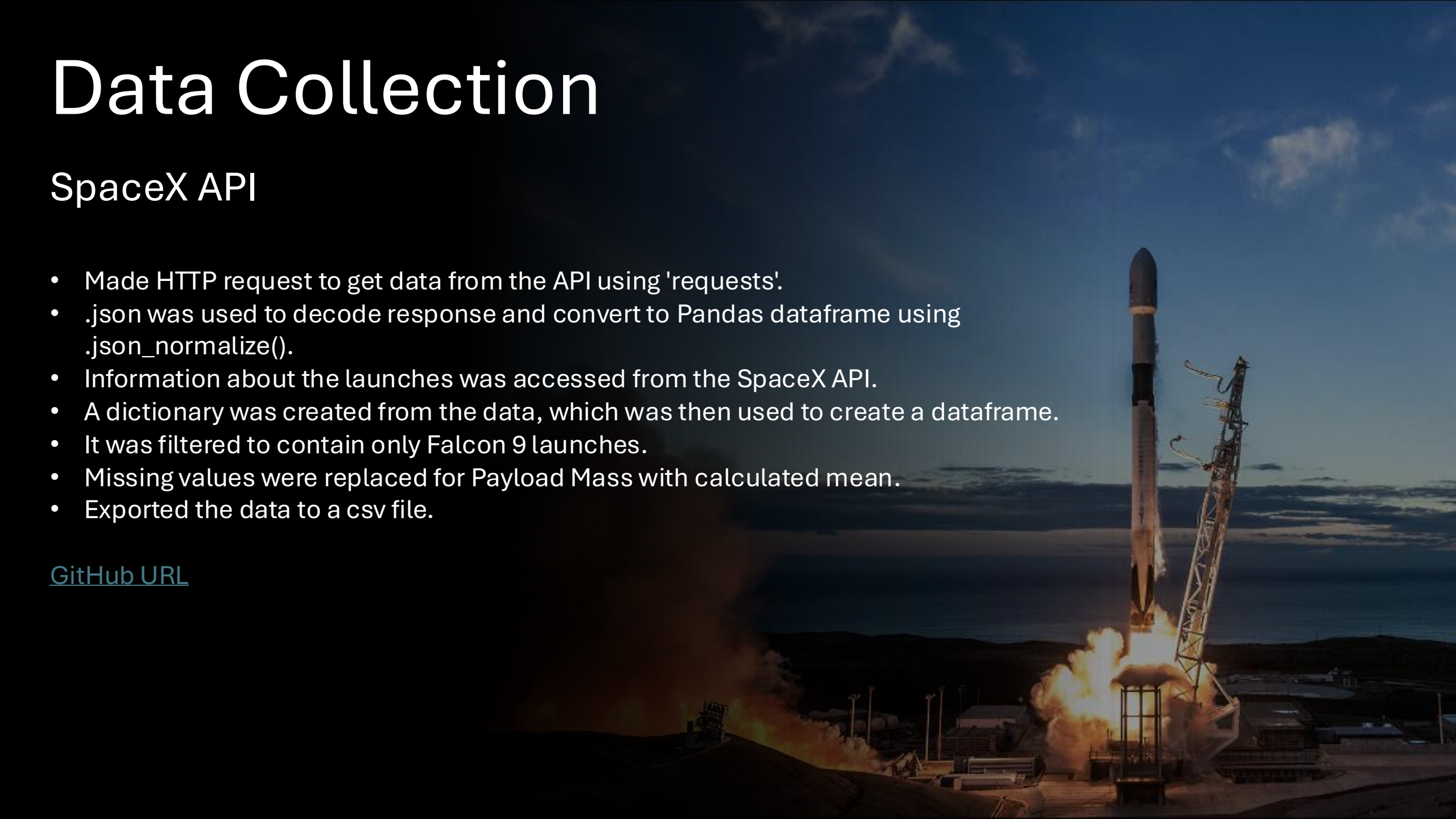
The accuracy of different models was examined using confusion matrices. Scikit-learn accuracy scores, Jaccard index, and F-1 scores were used to evaluate the models for best results.

Data Collection

SpaceX API

- Made HTTP request to get data from the API using 'requests'.
- .json was used to decode response and convert to Pandas dataframe using .json_normalize().
- Information about the launches was accessed from the SpaceX API.
- A dictionary was created from the data, which was then used to create a dataframe.
- It was filtered to contain only Falcon 9 launches.
- Missing values were replaced for Payload Mass with calculated mean.
- Exported the data to a csv file.

GitHub URL

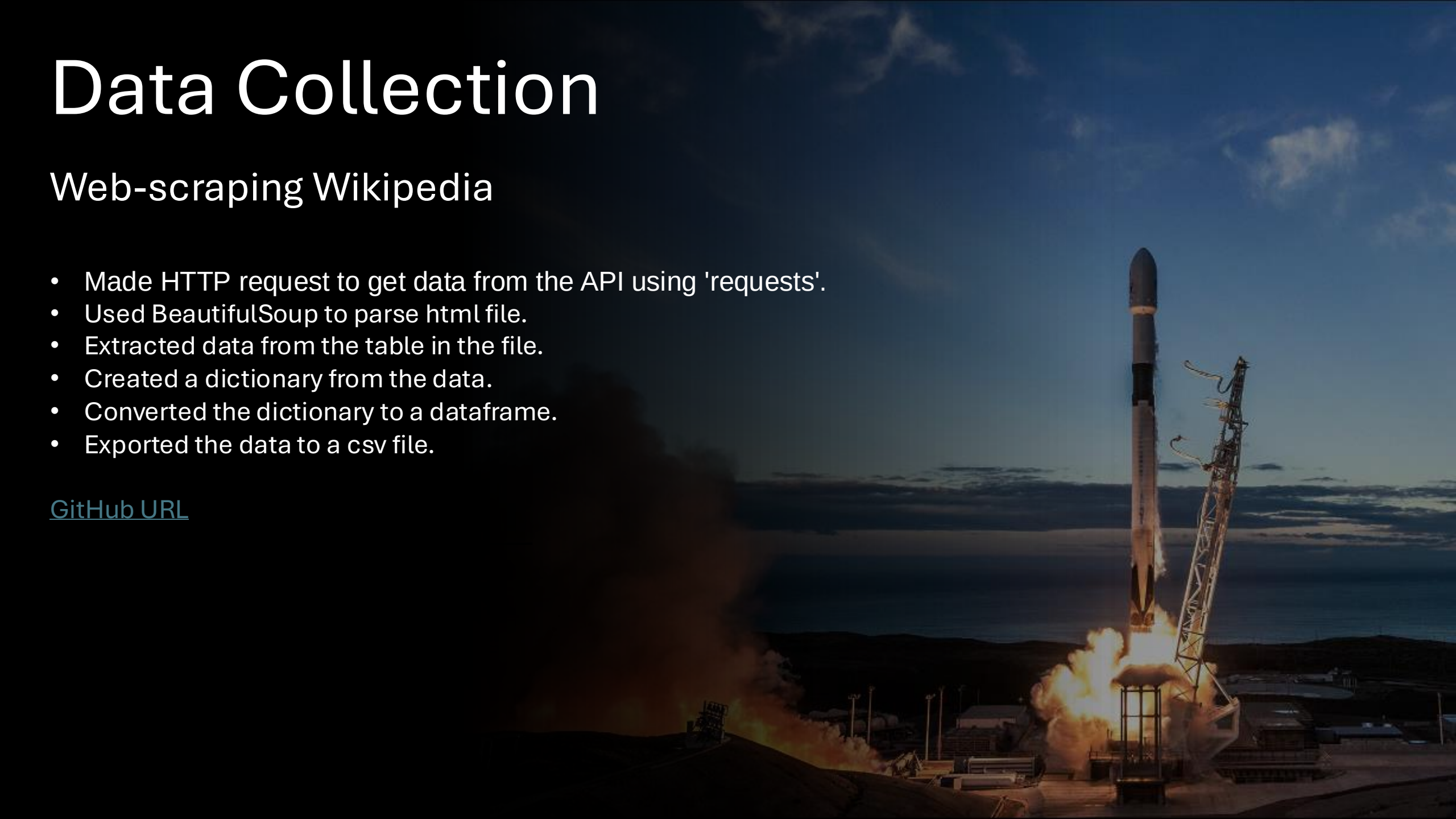


Data Collection

Web-scraping Wikipedia

- Made HTTP request to get data from the API using 'requests'.
- Used BeautifulSoup to parse html file.
- Extracted data from the table in the file.
- Created a dictionary from the data.
- Converted the dictionary to a dataframe.
- Exported the data to a csv file.

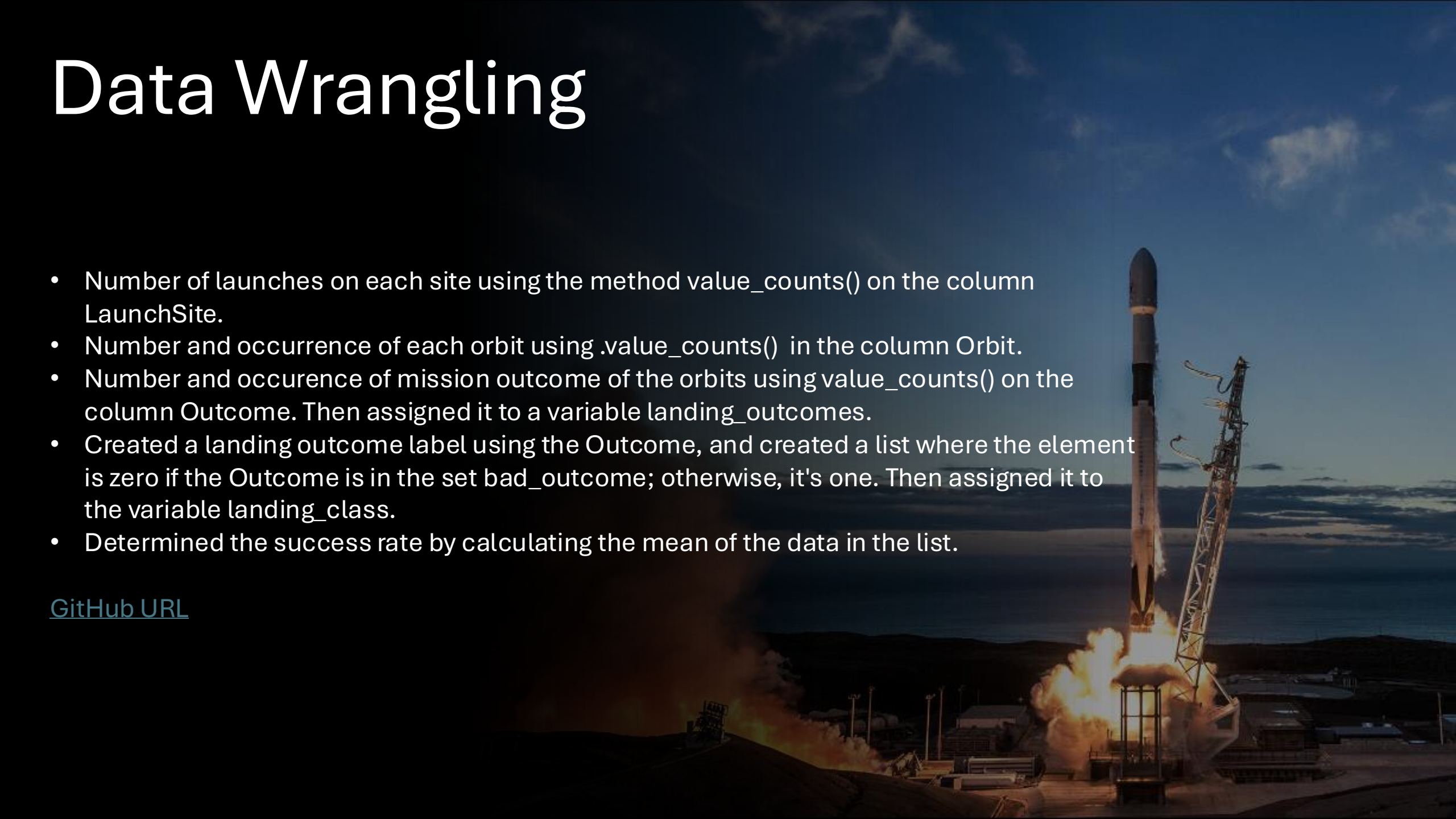
GitHub URL



Data Wrangling

- Number of launches on each site using the method value_counts() on the column LaunchSite.
- Number and occurrence of each orbit using .value_counts() in the column Orbit.
- Number and occurence of mission outcome of the orbits using value_counts() on the column Outcome. Then assigned it to a variable landing_outcomes.
- Created a landing outcome label using the Outcome, and created a list where the element is zero if the Outcome is in the set bad_outcome; otherwise, it's one. Then assigned it to the variable landing_class.
- Determined the success rate by calculating the mean of the data in the list.

GitHub URL



EDA

SQL

Displayed:

- The names of the unique launch sites.
- Records where launch sites begin with the ‘CCA'.
- Total payload mass carried by NASA (CRS) boosters.
- Average payload mass carried by boosterF9 v1.1

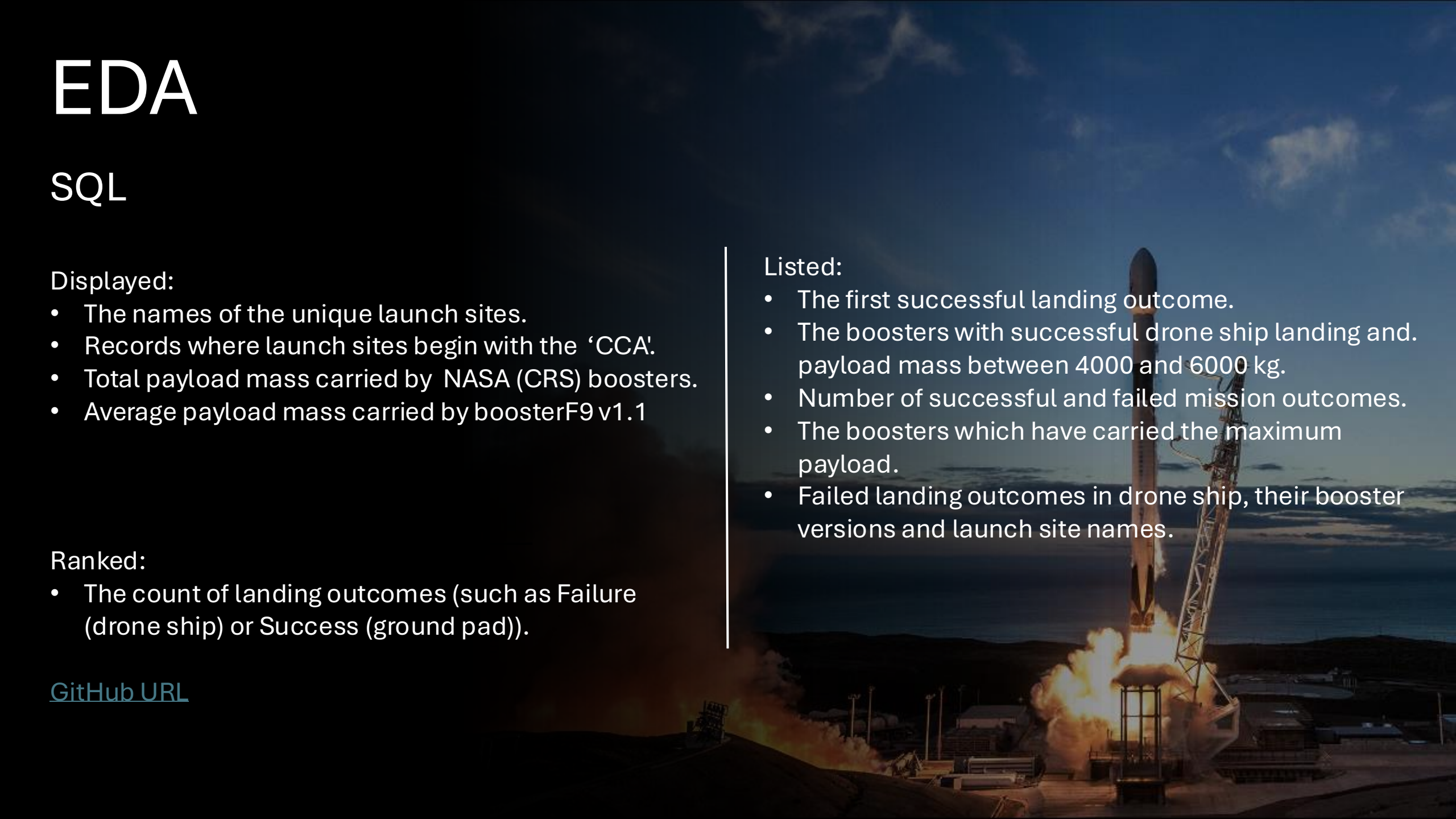
Ranked:

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)).

Listed:

- The first successful landing outcome.
- The boosters with successful drone ship landing and. payload mass between 4000 and 6000 kg.
- Number of successful and failed mission outcomes.
- The boosters which have carried the maximum payload.
- Failed landing outcomes in drone ship, their booster versions and launch site names.

GitHub URL



EDA

Pyplot, Seaborn

Plots:

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload Mass vs Launch Site
- Orbit Type vs Success Rate
- Flight Number vs Orbit Type
- Payload Mass vs Orbit Type
- Yearly trend for launch success

Scatter plots were chosen to show the relationship between variables which could be used in machine learning model.
Bar charts show comparisons among discrete categories such as Orbit Type.
A line chart was used to show trends in data over time (time series).

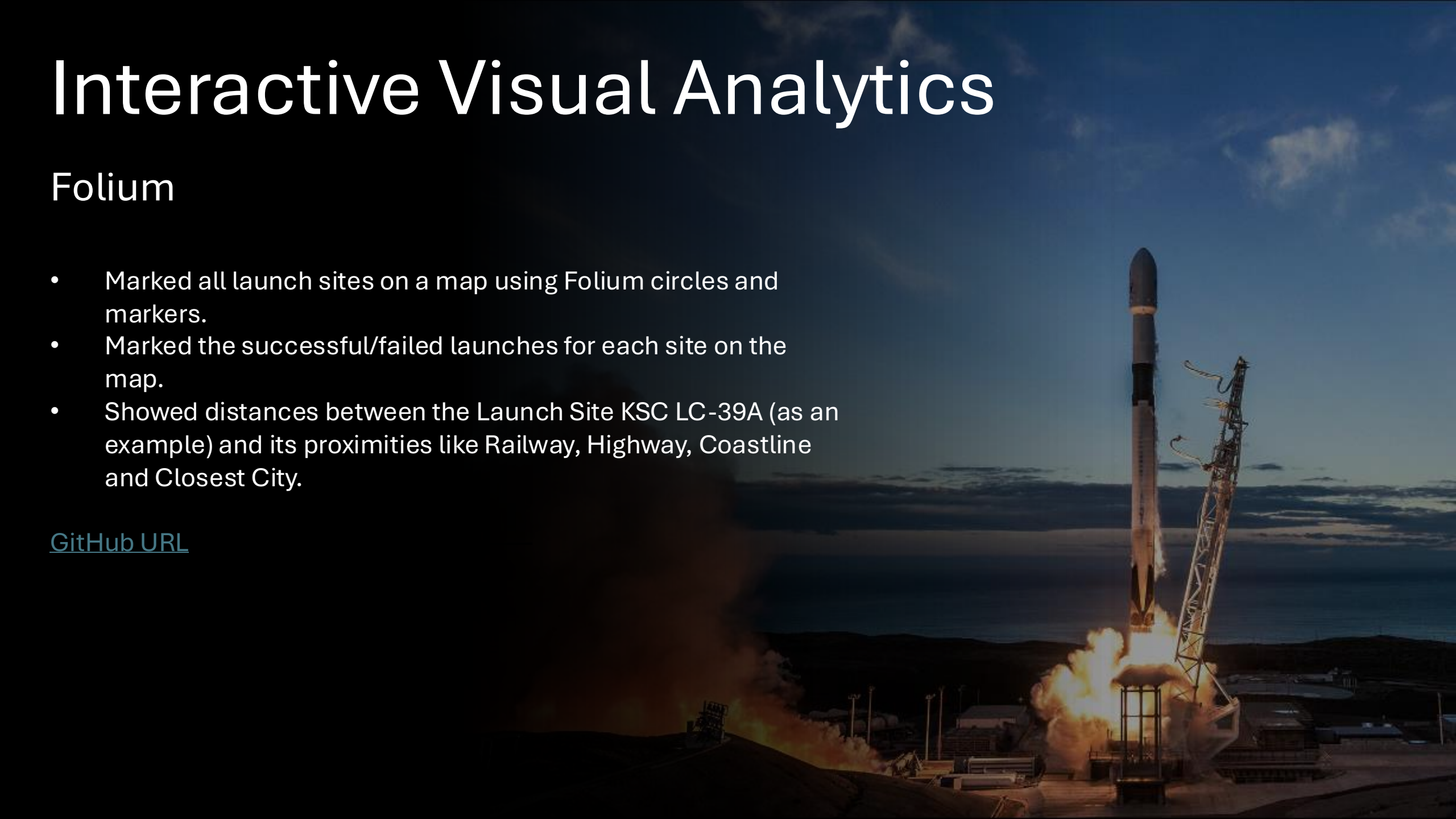
GitHub URL

Interactive Visual Analytics

Folium

- Marked all launch sites on a map using Folium circles and markers.
- Marked the successful/failed launches for each site on the map.
- Showed distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

GitHub URL

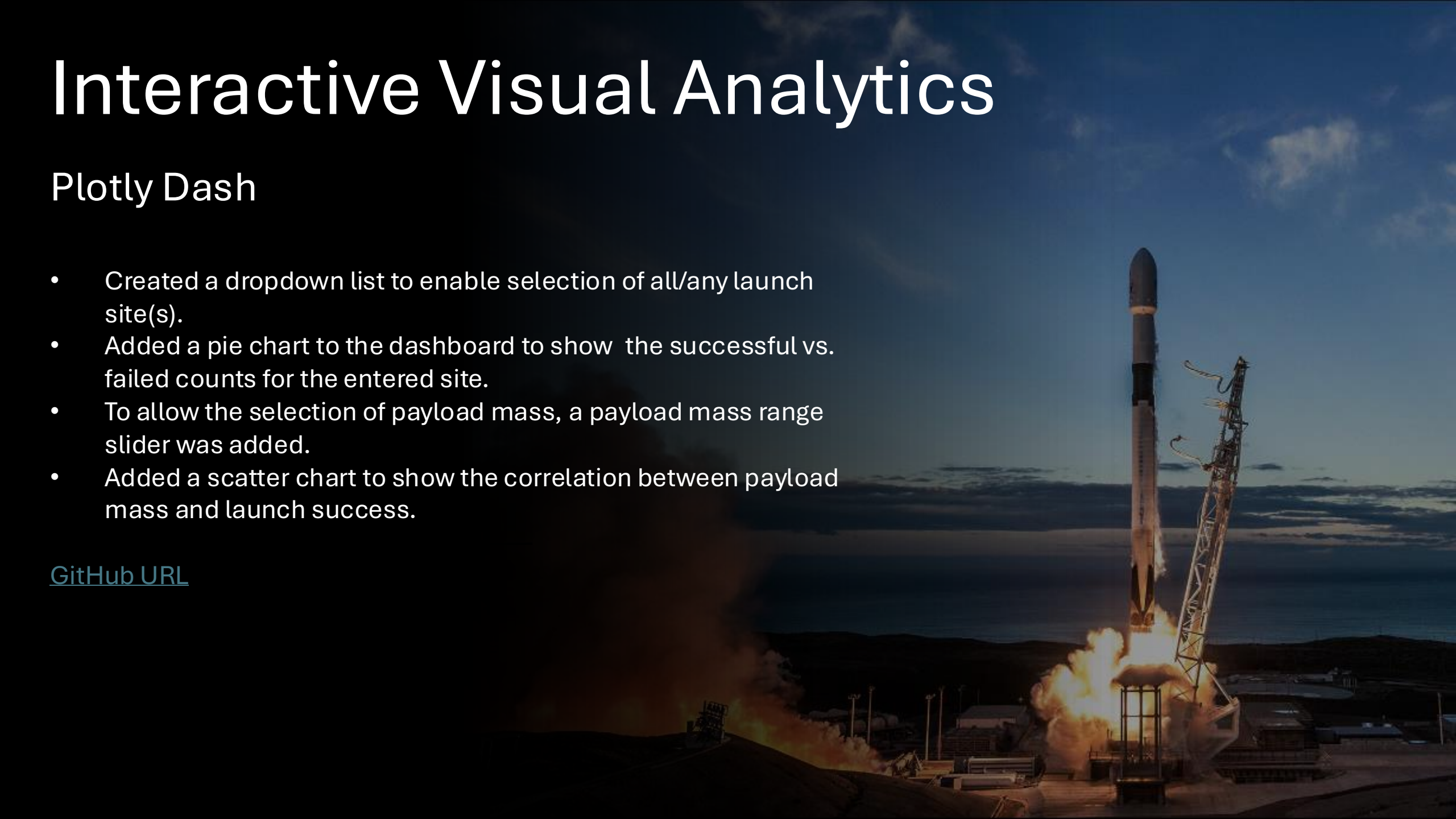


Interactive Visual Analytics

Plotly Dash

- Created a dropdown list to enable selection of all/any launch site(s).
- Added a pie chart to the dashboard to show the successful vs. failed counts for the entered site.
- To allow the selection of payload mass, a payload mass range slider was added.
- Added a scatter chart to show the correlation between payload mass and launch success.

GitHub URL



Predictive Analytics

LogReg, SVM, Tree, K-NN

- A NumPy array was created from the Class column.
- The data was standardized with StandardScaler. Data was fit and transformed.
- The data was split into training and testing sets.
- GridSearchCV cross-validation object was created for parameter optimization with cv=10.
- The object was applied to the following algorithms:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbor
- Confusion matrices for each algorithm were examined to determine that the major problem with them were false positives.
- Scikit-learn metrics (accuracy_score, jaccard_score, f1_score) were used to identify the best model.

GitHub URL

349
TOTAL LAUNCHES

306
TOTAL LANDINGS

280
TOTAL REFLIGHTS

Results

Summary

Exploratory Data Analysis

- Improved launch success over time.
- KSC LC-39A most successful among landing sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Visual Analytics

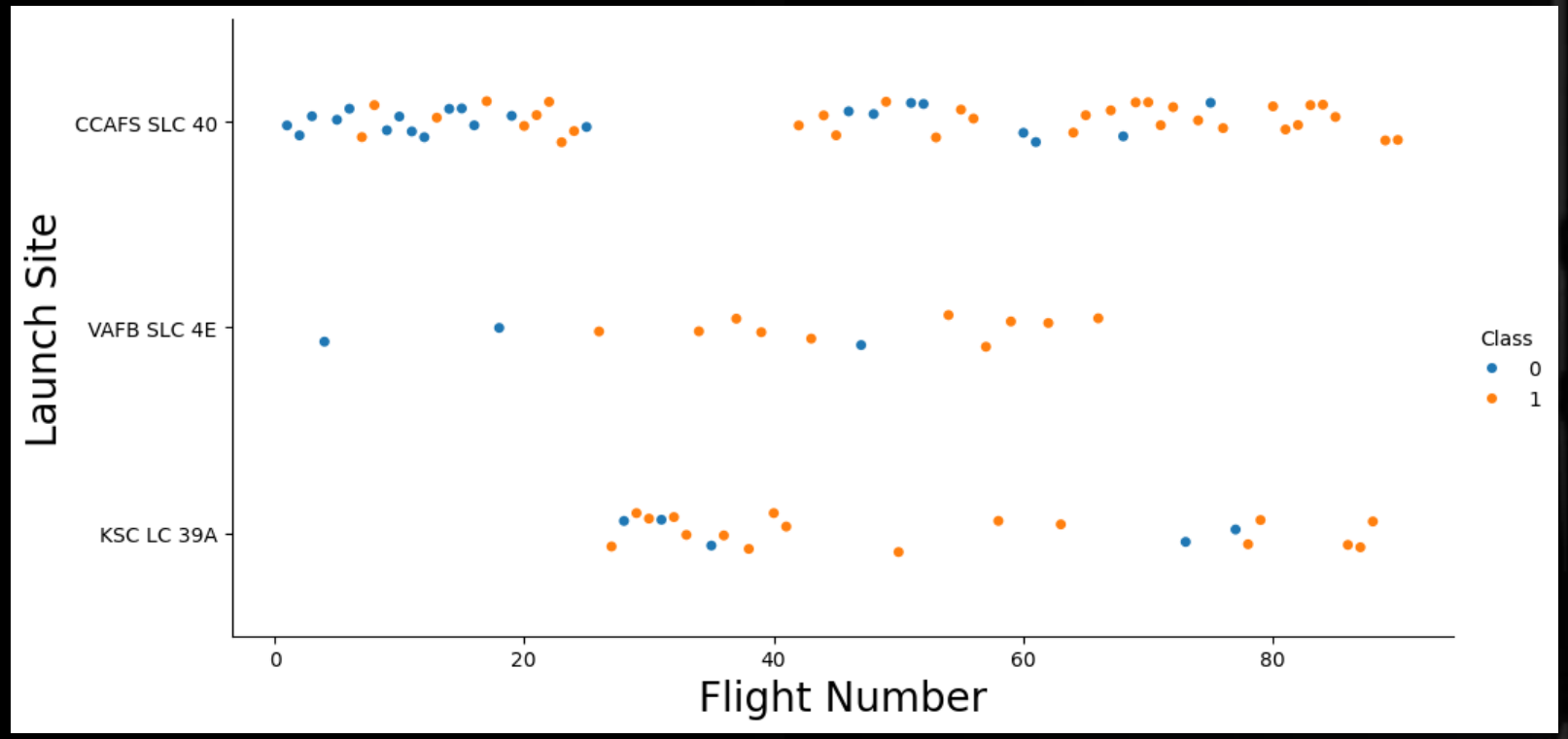
- Most launch sites are near the equator and close to coast.
- Launch sites are far away from anything a failed launch can damage (city, highway, railway), while still logistically feasible.

Predictive Analytics

- Decision Tree is the best predictive model for the dataset.

EDA - Visualization

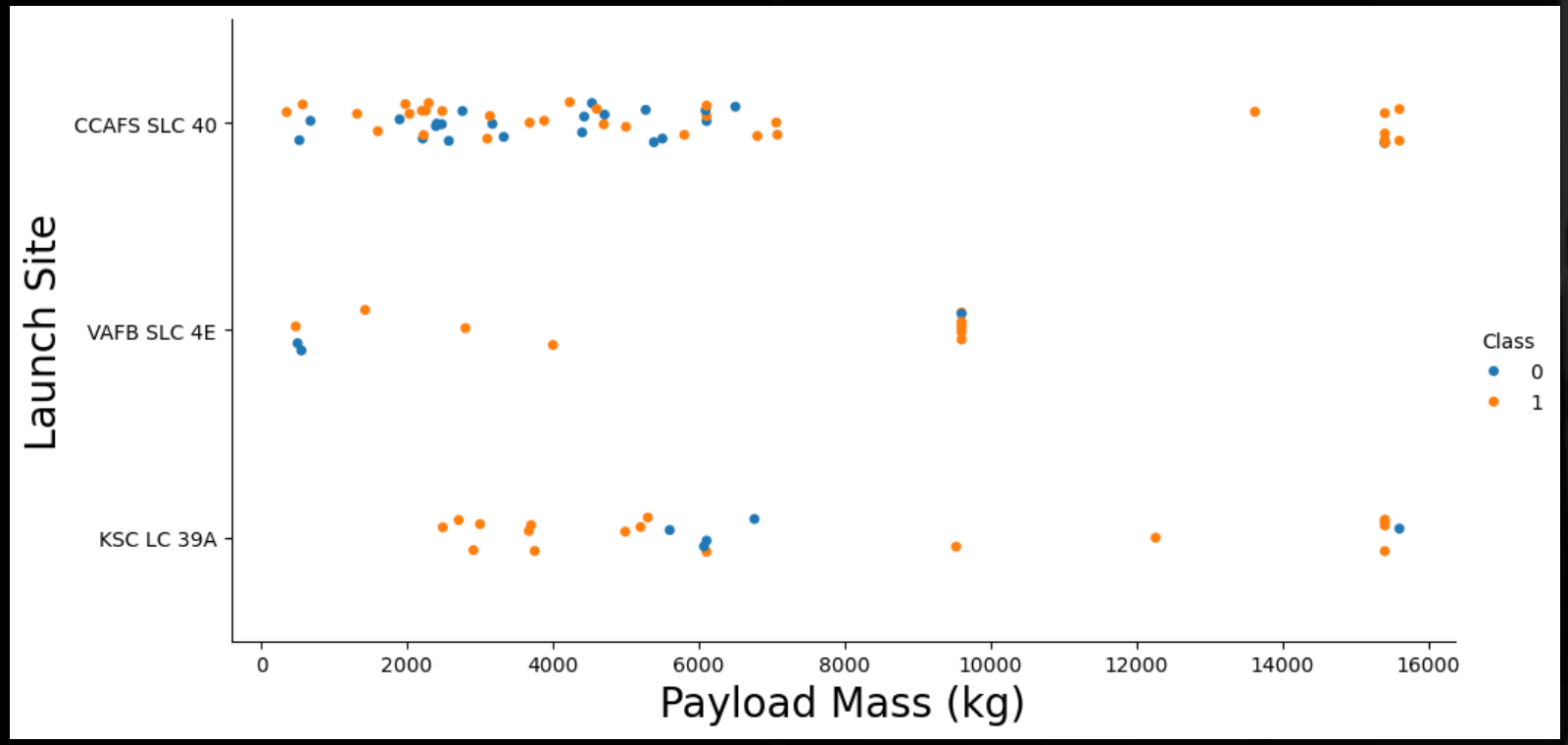
Flight Number vs Launch Site



Insights:

- Earlier flights had a lower success rate than later flights.
- Around half of all launches were from the launch site CCAFS SLC 40.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

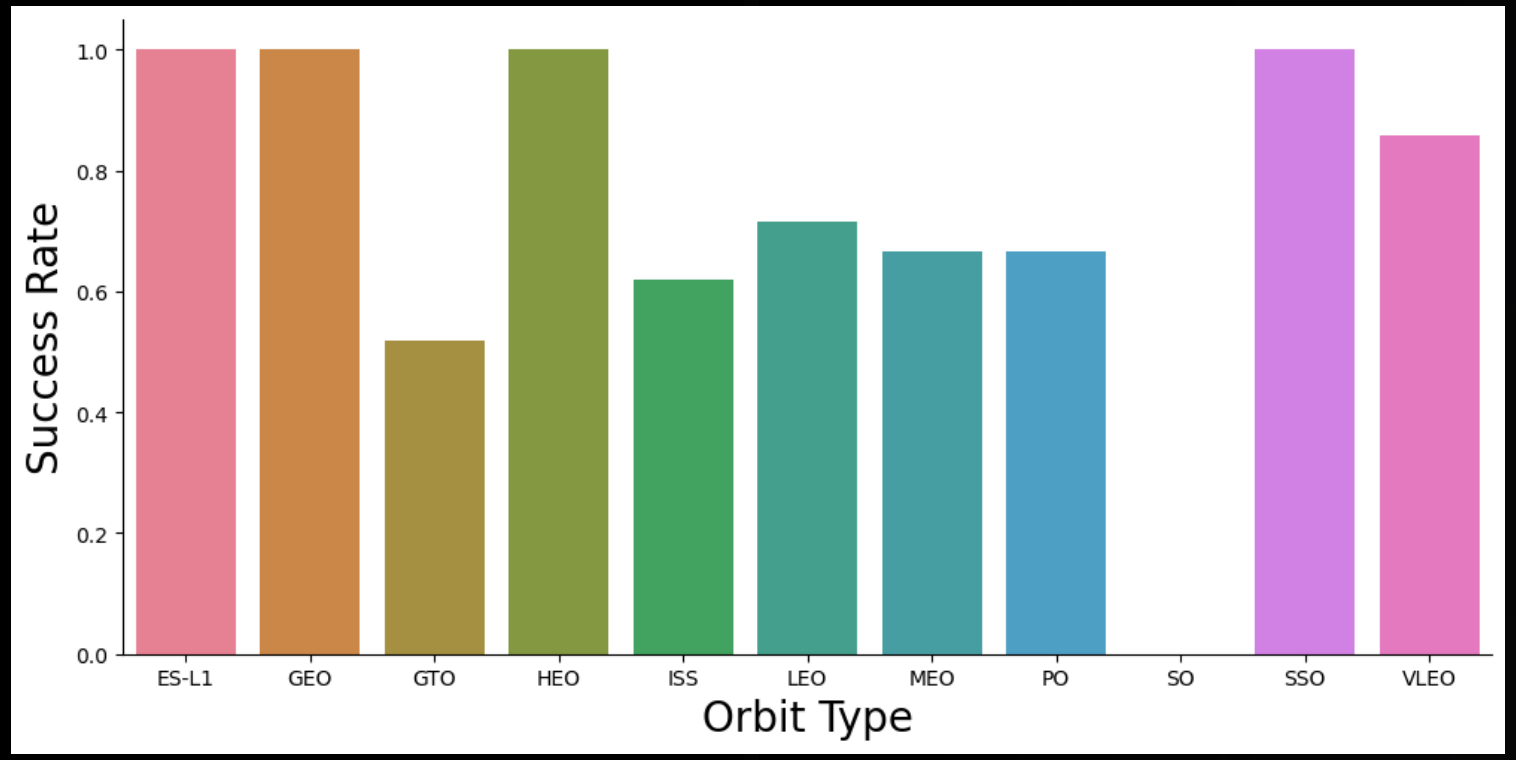
Payload Mass vs Launch Site



Insights:

- Higher the payload mass, higher the success rate.
- Launches with a payload greater than 7,000 kg mostly successful.
- KSC LC-39A has 100% success rate for payloads less than 5,500 kg.
- VAFB SKC-4E has not launched anything greater than 10,000 kg.

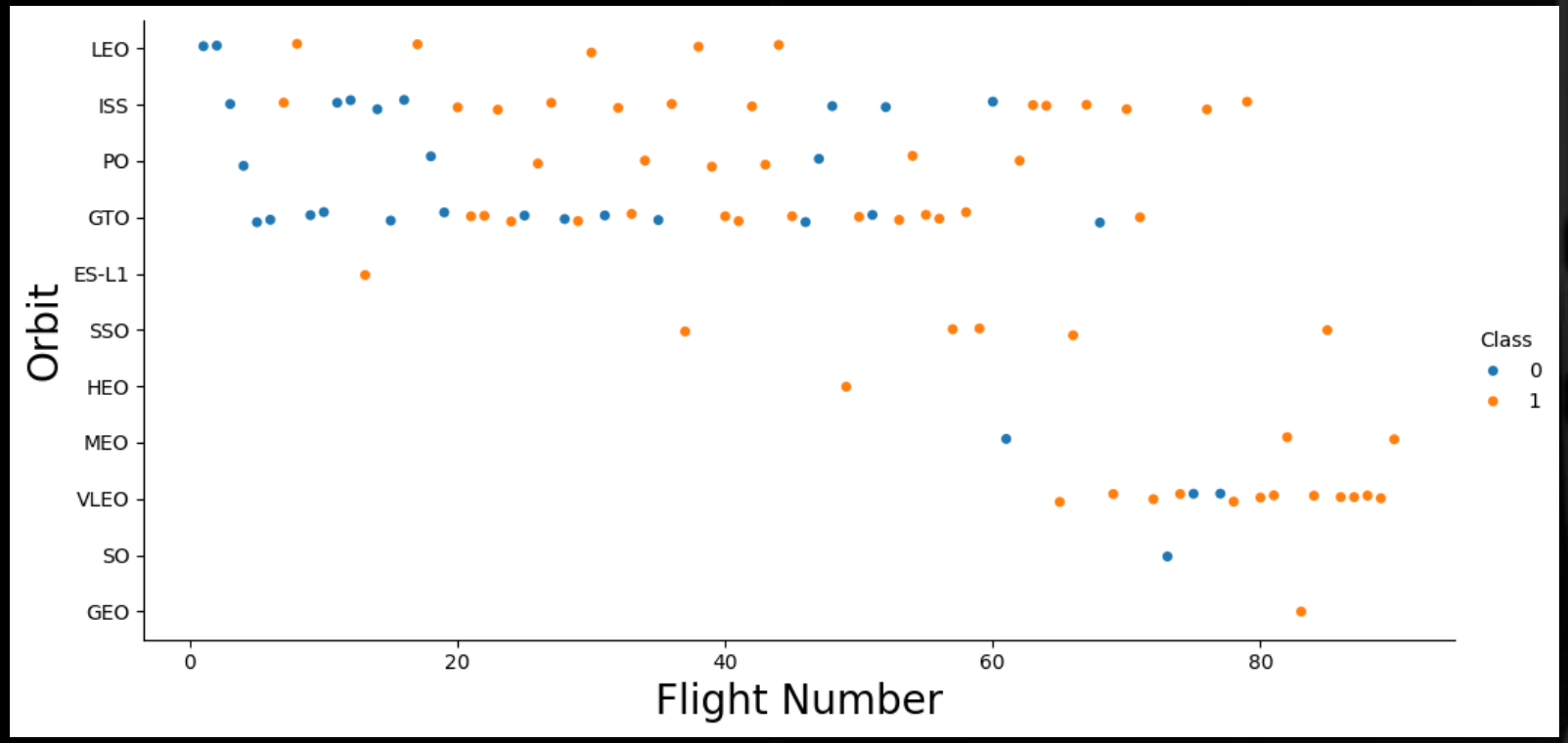
Success Rate by Orbit Type



Insights:

- Orbits with 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO
  - ISS
  - LEO
  - MEO
  - PO

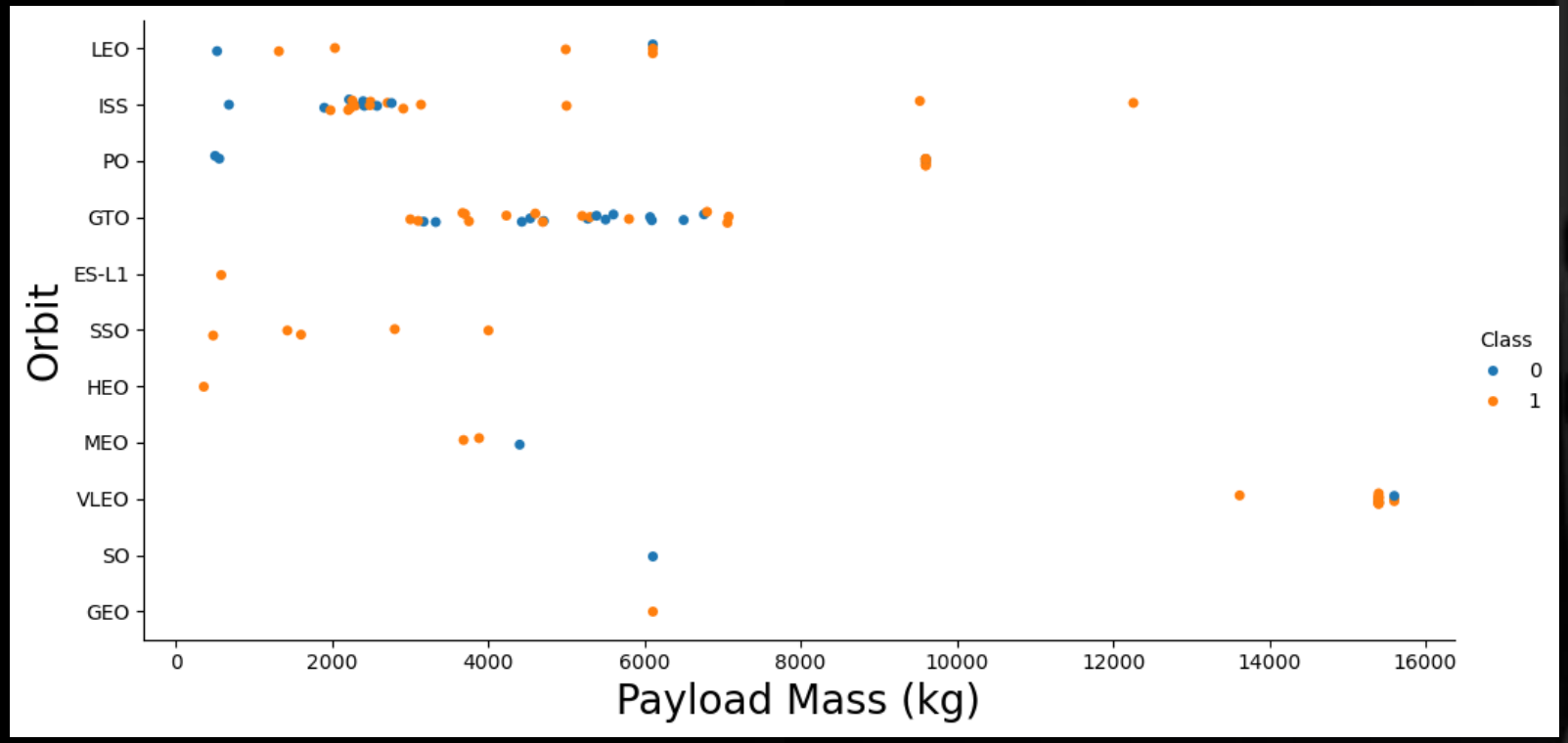
Flight Number vs Orbit Type



Insights:

- The success rate increased with the number of flights for each orbit.
- This relationship is highly apparent for the LEO orbit.
- The GTO orbit does not follow this trend.

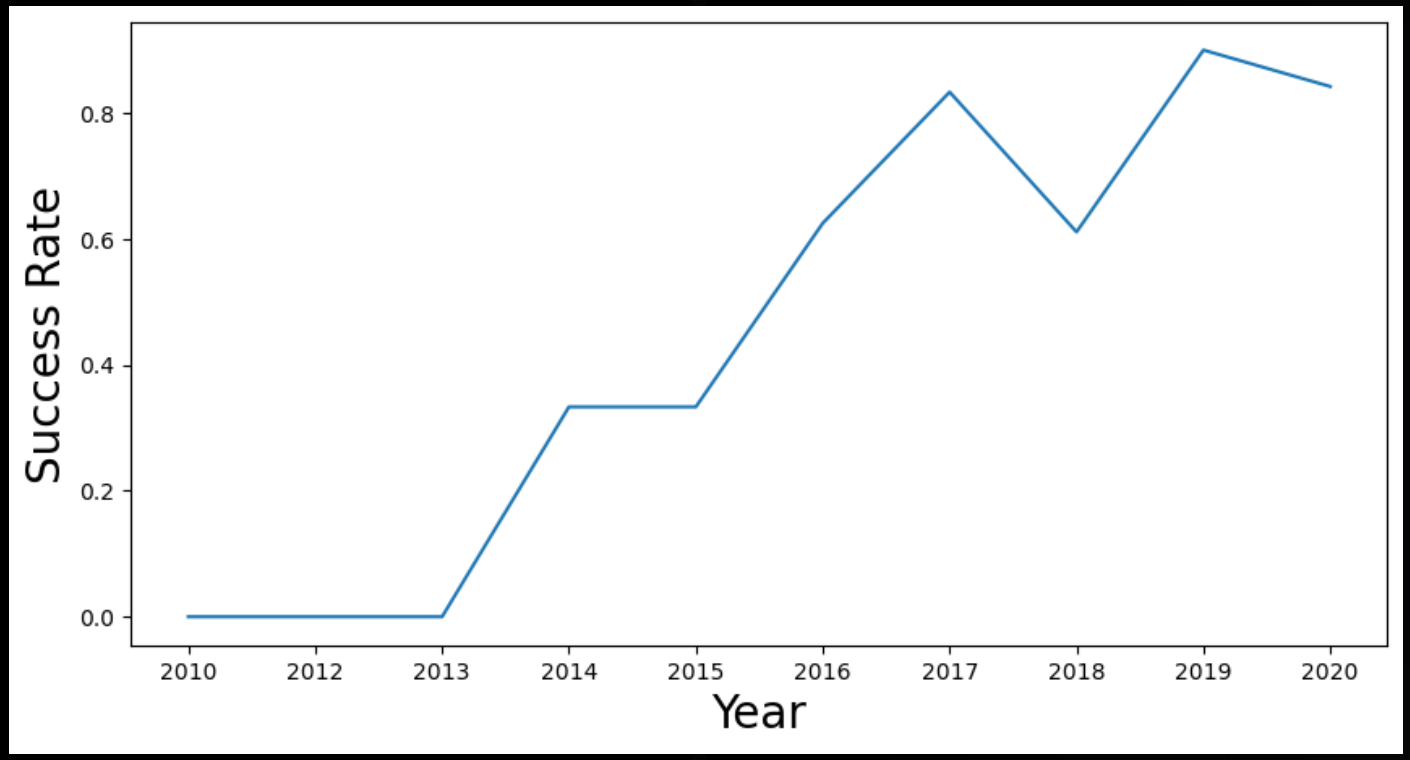
Payload Mass vs Orbit Type



Insights:

- The heavier payloads show better results with the LEO, ISS and PO orbits.
- The GTO orbit shows mixed success with heavier payloads.

Success Rate over Time



Insights:

- Success rate has improved from 2013-2017 and 2018-2019.
- Success rate decreased from 2017-2018 and from 2019-2020.
- Overall, the success rate has seen a positive trend since 2013.

EDA - SQL

All Launch Sites

```
%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

There are four distinct launch sites in the dataset: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

CCA Launch Sites

```
%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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (
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 (
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where "Customer" = 'NASA (CRS)';
```

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

total_payload_mass
45596

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where "Booster_Version" like '%F9 v1.
```

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

average_payload_mass
2534.6666666666665

The total payload for NASA (CRS) launches was 45,596 kg while booster F9 v1.1 carried an average payload of ~2,535 kg.

Landing & Mission Information

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)';
```

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

first_successful_landing
2015-12-22

```
%sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

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

Landing & Mission Information Contd.

```
%sql select "Mission_Outcome", count(*) as total_number from SPACEXTABLE group by "Mission_Outcome";
```

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

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

- The first successful landing was on 22/12/2015.
- Booster versions F9 FT B1022, B1026, B1021.2, B1031.2 landed successfully on drone ships.
- There were 99 successful mission outcomes, 1 failure in flight, and 1 success with unclear payload status.

Boosters

```
%sql select "Booster_Version" from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

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

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

List of boosters carrying maximum payload mass.

2015 Failed Launches

```
%%sql select substr(Date, 6,2) as month, date, "Booster_Version", "Launch_Site", "Landing_Outcome" from SPACEXTABLE
      where "Landing_Outcome" = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Failed drone ship landing outcomes in 2015 by month.

Landing Outcome Rankings Between 2010-06-04 and 2017-03-20

```
%%sql select "Landing_Outcome", count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by "Landing_Outcome"
      order by count_outcomes desc;
```

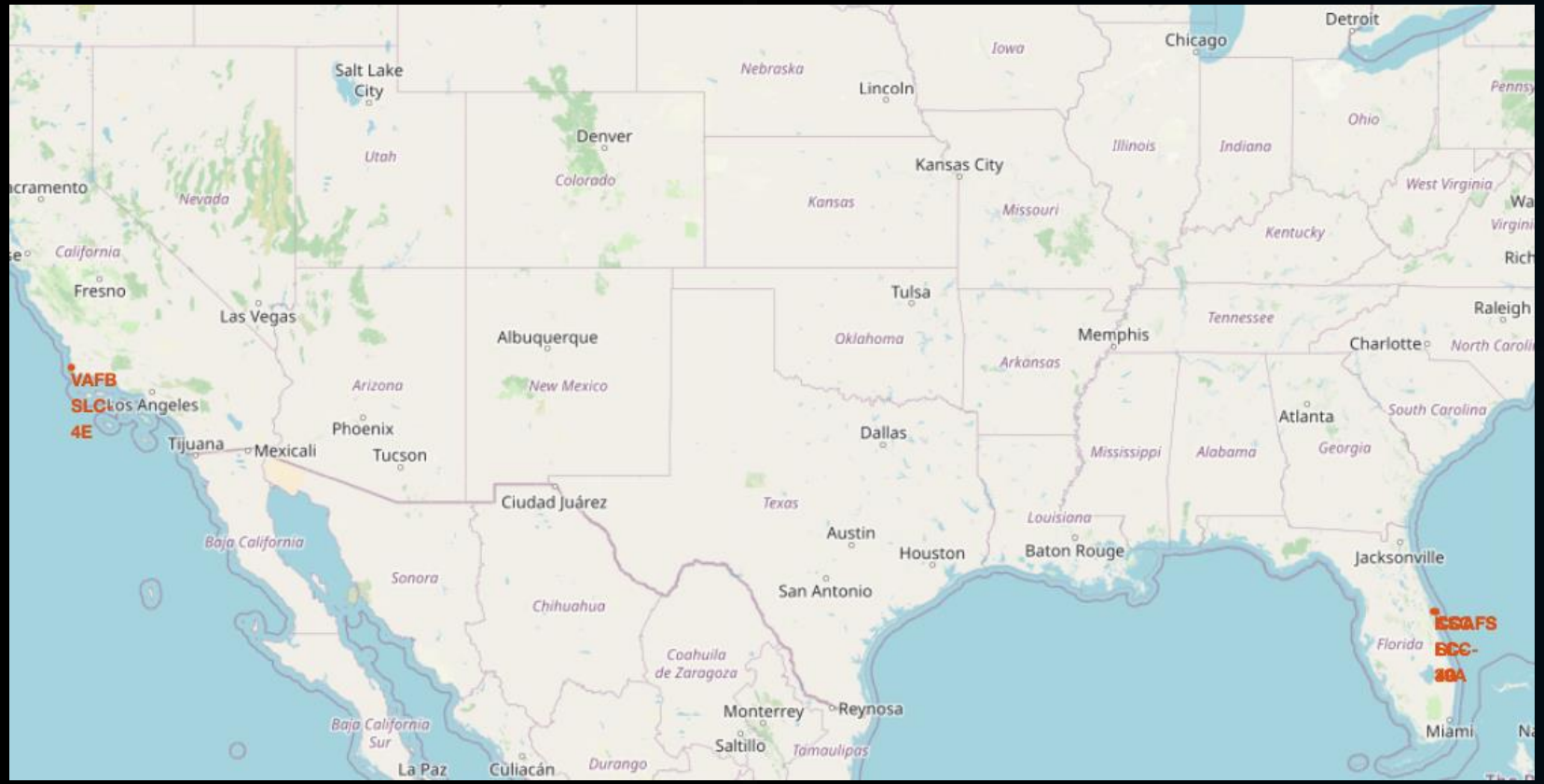
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_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

Launch Site Proximity Analysis

Launch Sites

All Launch Sites



Insights:

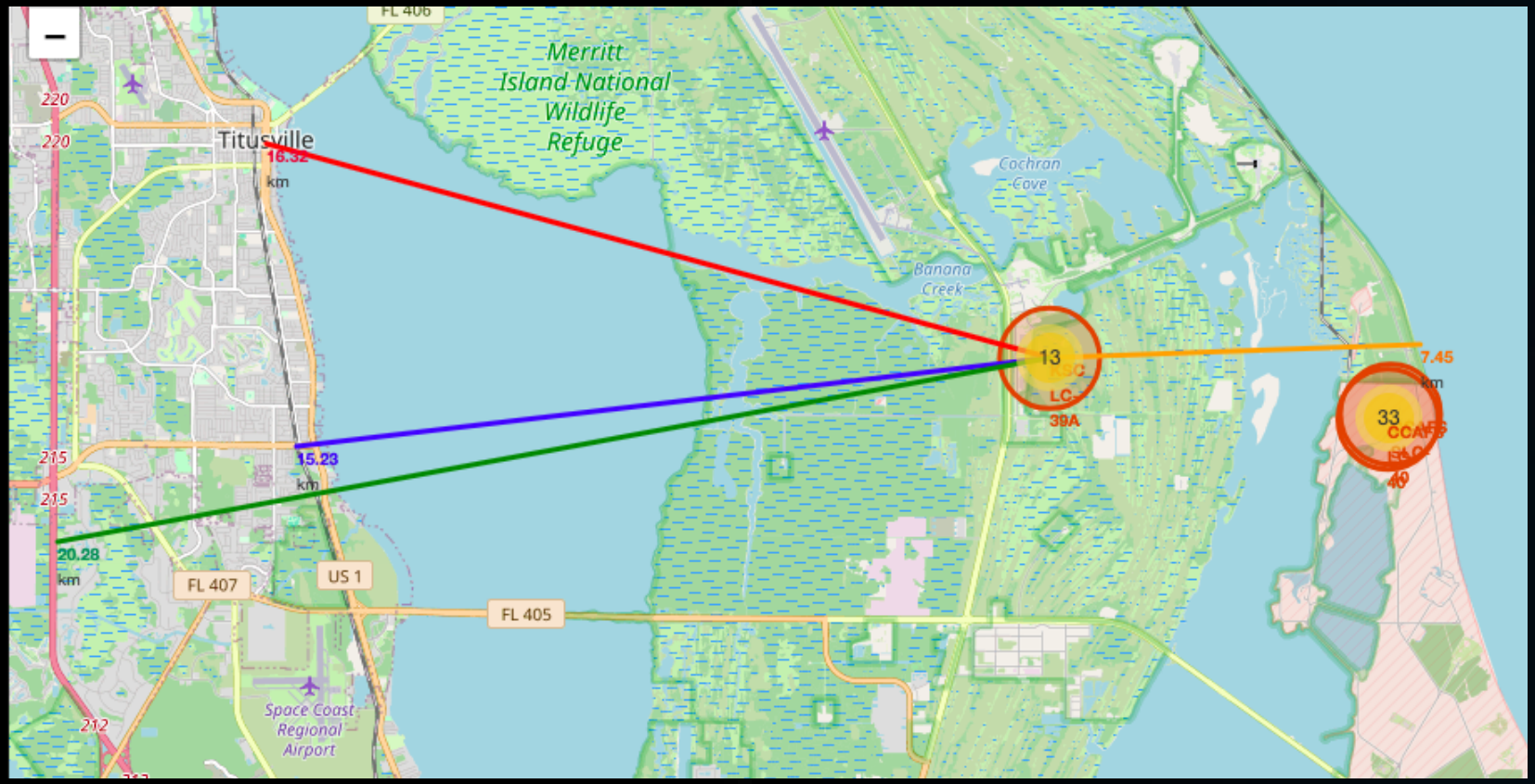
It is easier it is to launch the closer the launch site is to the equator. Rockets launched from sites near the equator get an additional natural boost - due to the rotation of the earth which helps to save fuel costs. They are also close to the coastline to ensure security of the launchpad as well as the safety of the people.

Launch Sites by Mission Outcome

Insights:

- Green markers are successful launches.
- Red markers are unsuccessful launches.
- Launch site KSC LC-39A has a 10/13 success rate (76.92%).

Distance to Proximities



Insights:

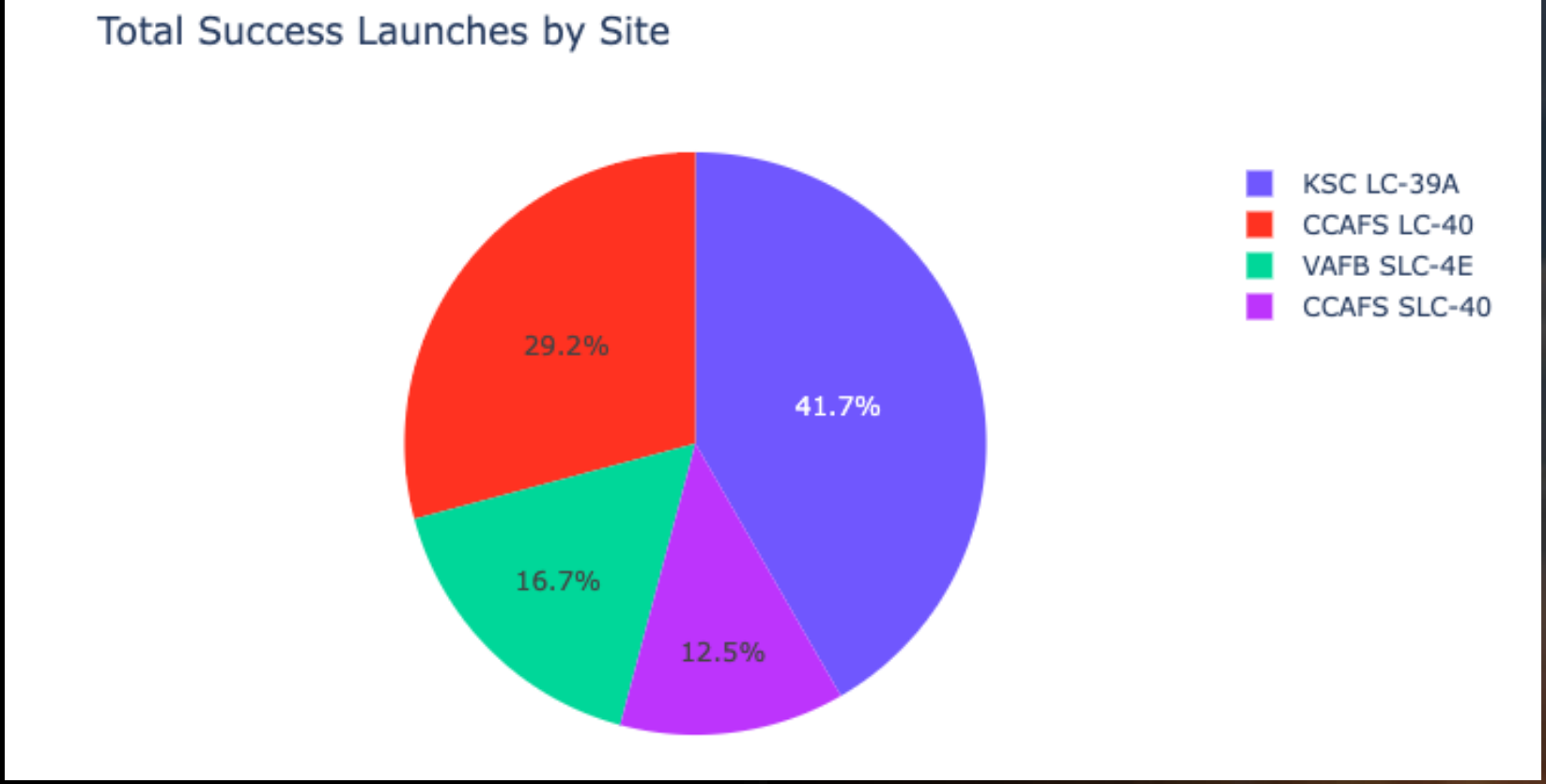
Launch site KSC LC-39A is:

- 15.23 km from nearest railway
- 20.28 km from nearest highway
- 7.45 km from nearest coastline
- 16.32 km from its closest city Titusville

Dashboard with Plotly

Launch Success

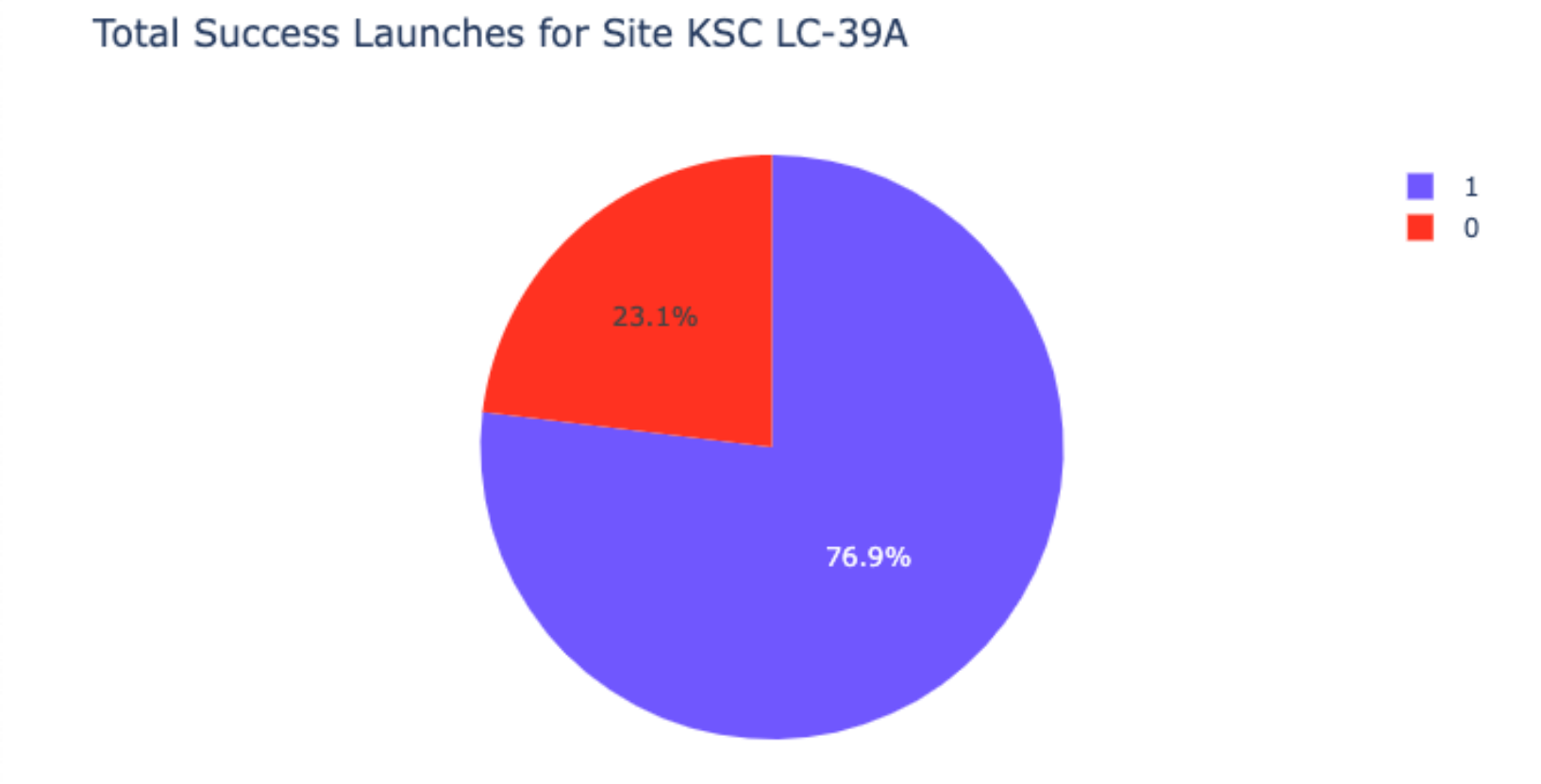
All Launch Sites



Insights:

KSC LC-39A has the most successful launches amongst launch sites (41.7%)

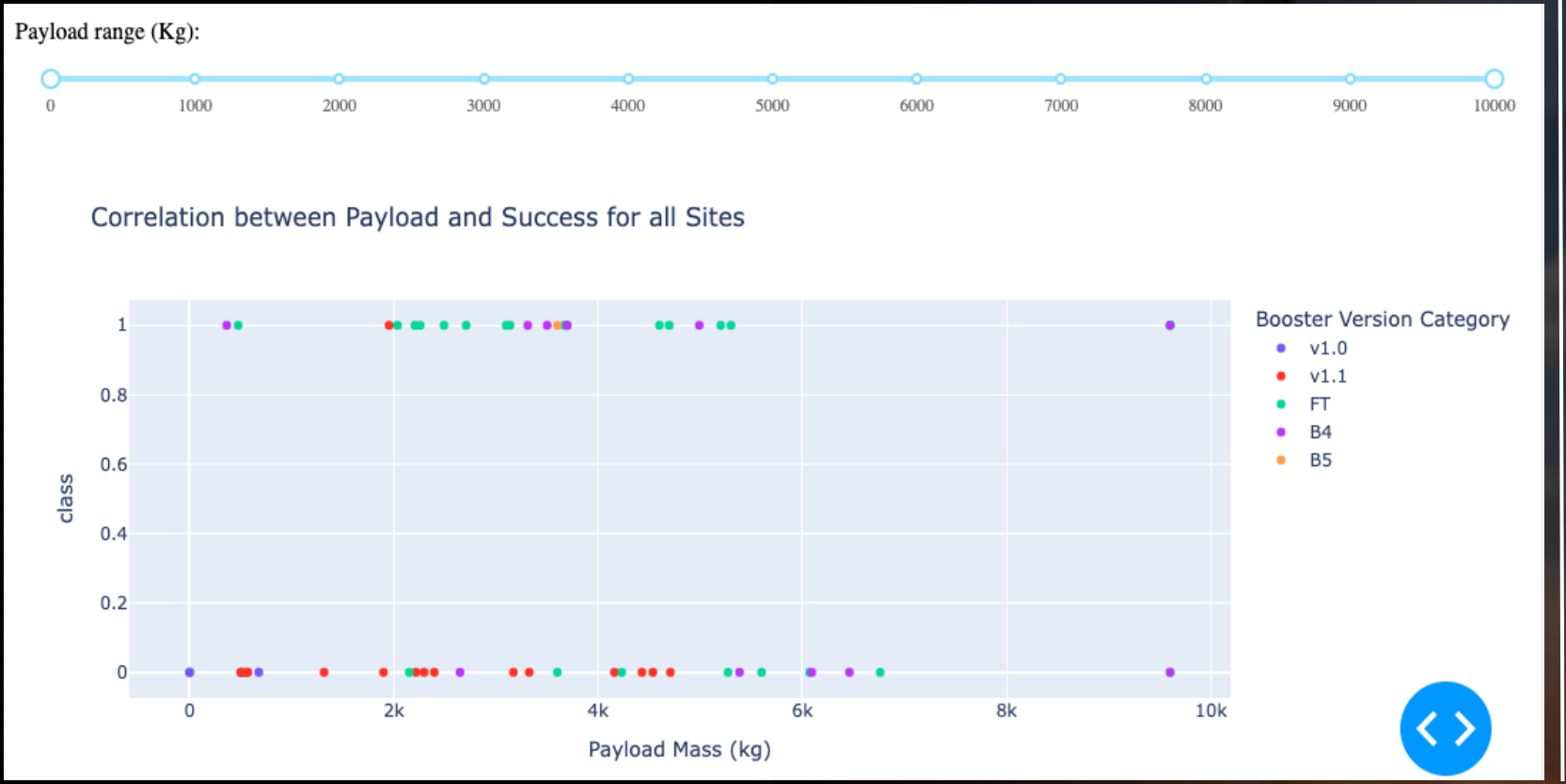
Highest Launch Success Ratio



Insights:

Amongst all the launches for site KSC LC-39A, 76.9% were a success and 23.1% were failed.

Launch Success by Payload Mass

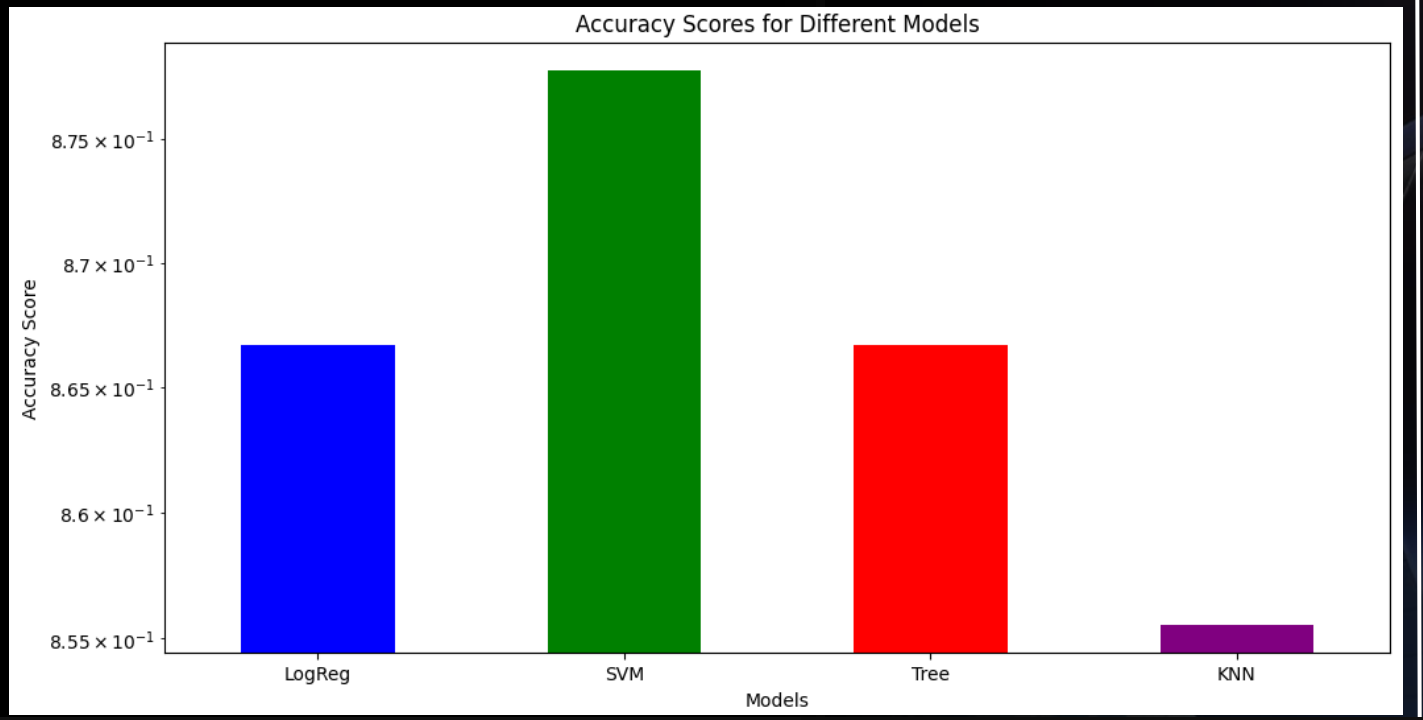


Insights:

Payloads between 2,000 kg and 5,000 kg have the highest success rate.

Predictive Analytics

Classification Accuracy



Insights:

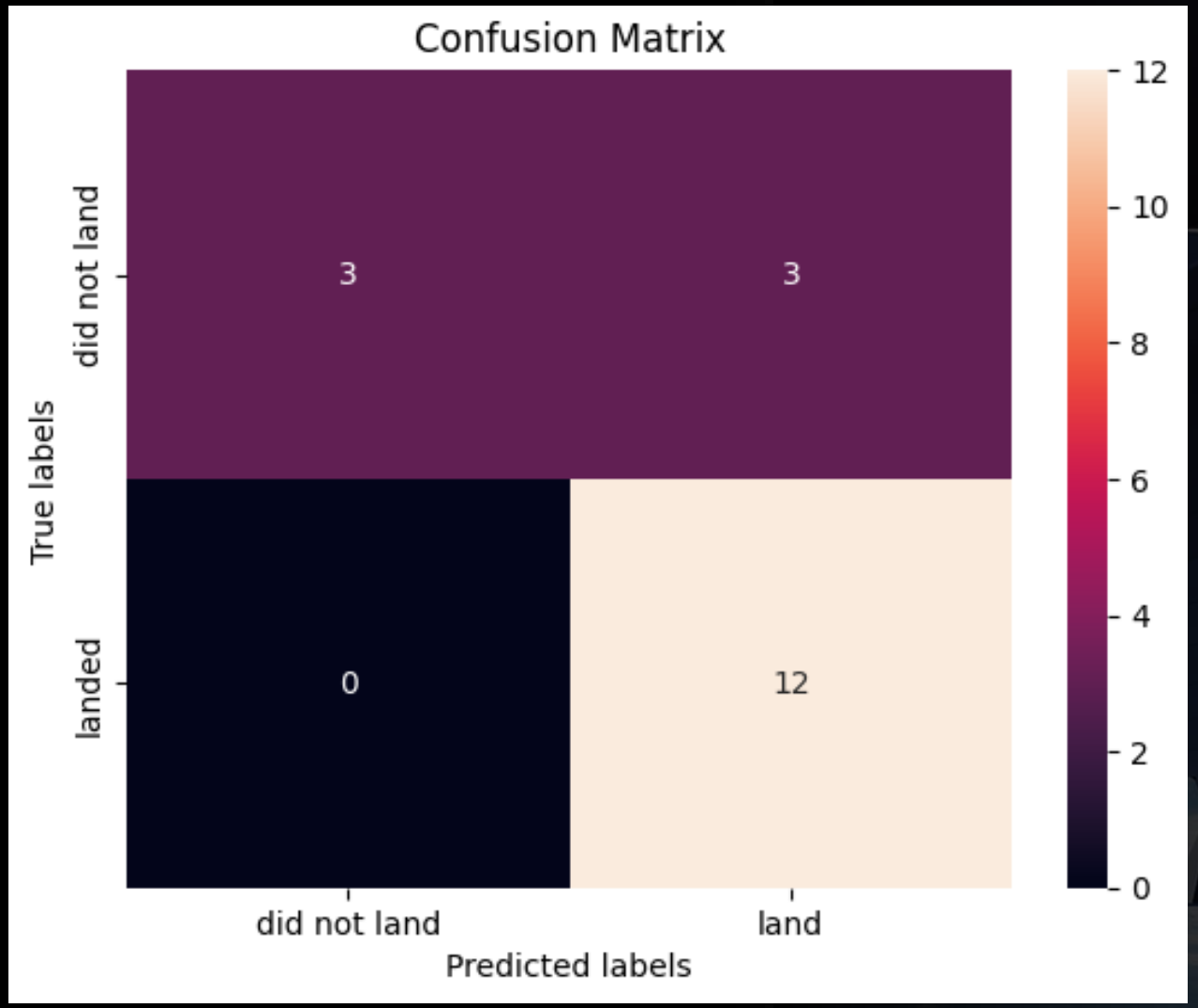
All the models performed about the same and had similar accuracy scores, likely due to a small dataset.

The Decision Tree model slightly outperformed the rest (best_score_)

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.833333	0.819444
F1_Score	0.909091	0.916031	0.909091	0.900763
Accuracy_Score	0.866667	0.877778	0.866667	0.855556

```
Best model is DecisionTree with a score of 0.8892857142857145
Best params is : {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt',
'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'}
```

Confusion Matrix



Insights:

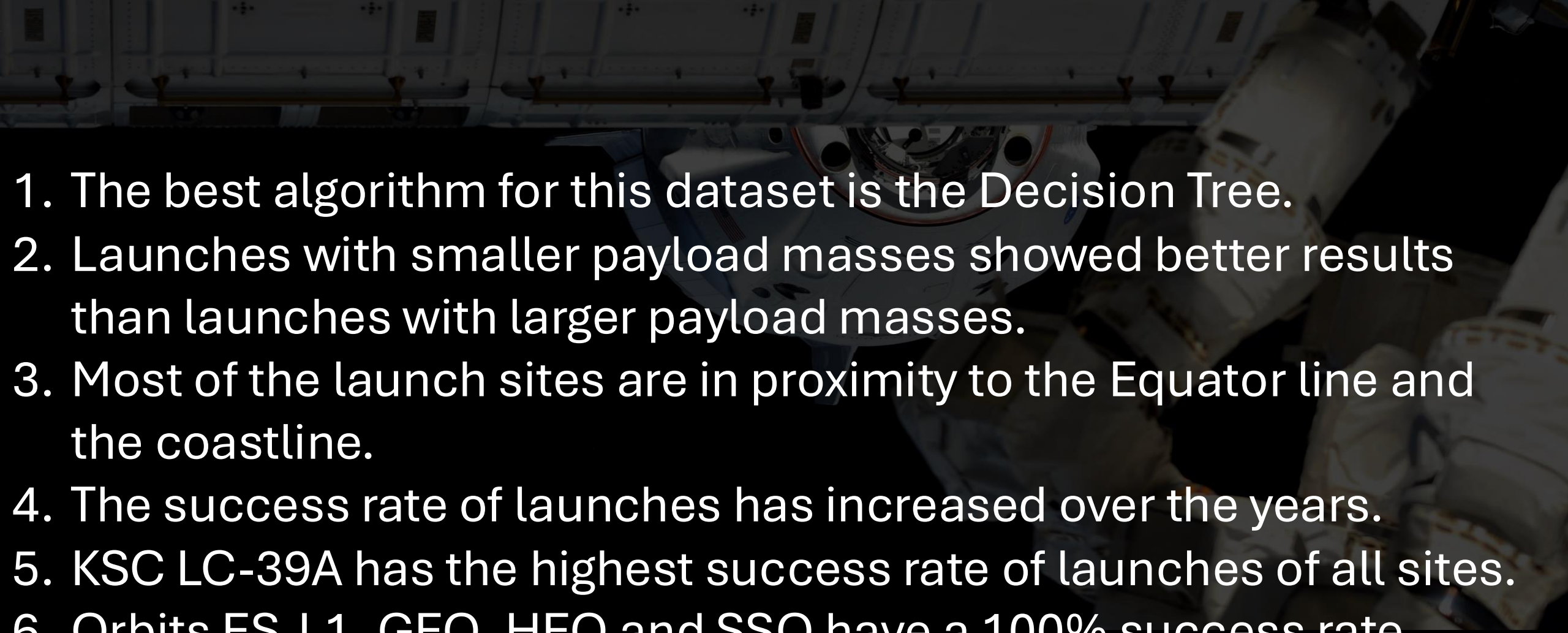
All the models produced identical confusion matrices. It can be seen that the models were able to classify the labels.

The outcomes were:

- 12 True positives
- 3 True negatives
- 3 False positives
- 0 False negatives

The major problem is false positives.

Conclusion



1. The best algorithm for this dataset is the Decision Tree.
2. Launches with smaller payload masses showed better results than launches with larger payload masses.
3. Most of the launch sites are in proximity to the Equator line and the coastline.
4. The success rate of launches has increased over the years.
5. KSC LC-39A has the highest success rate of launches of all sites.
6. Orbits ES-L1, GEO, HEO and SSO have a 100% success rate.


Ad
Astra.