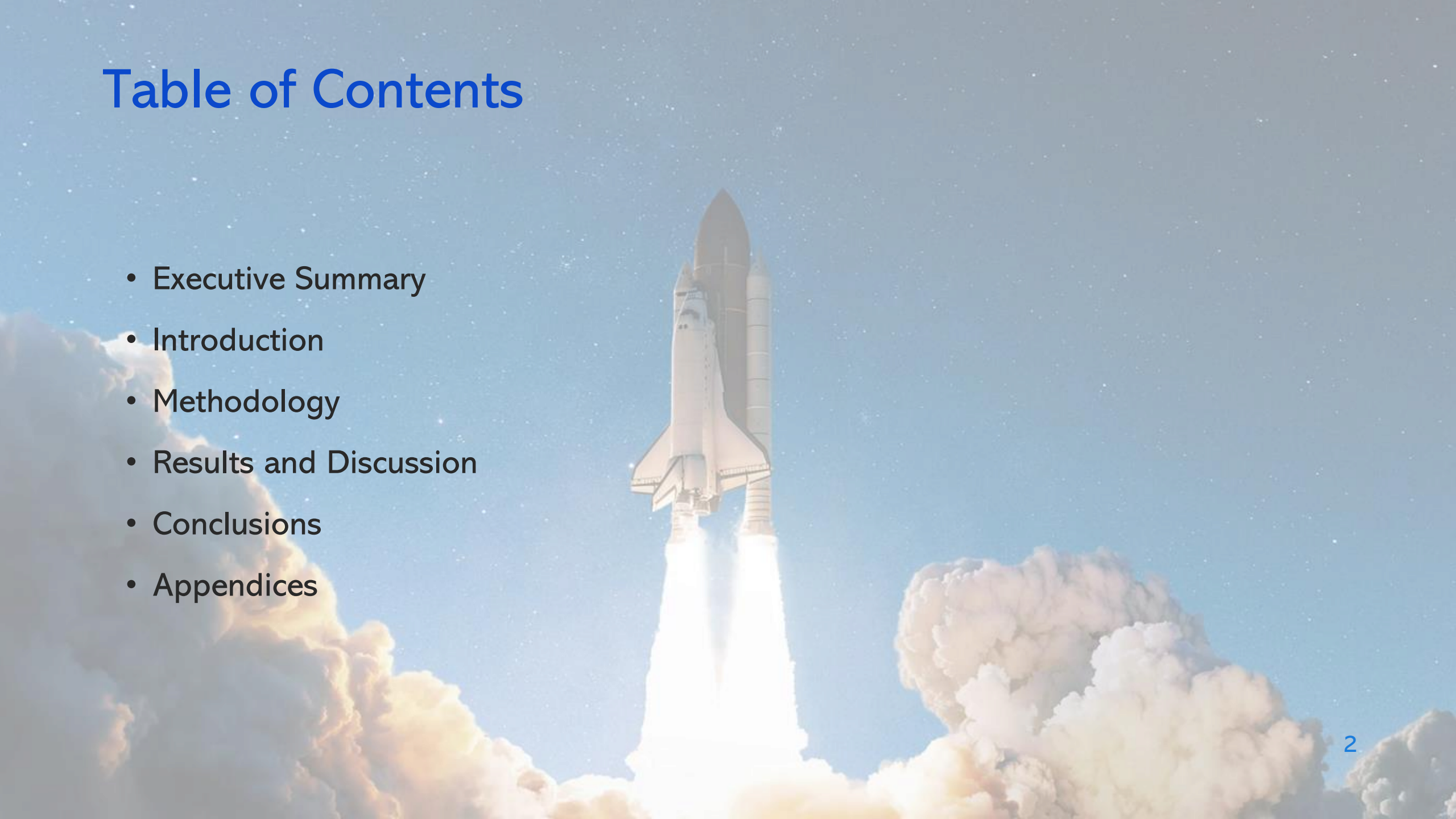
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kevin (Yiren) Mao
May 9th, 2024

Table of Contents



- Executive Summary
- Introduction
- Methodology
- Results and Discussion
- Conclusions
- Appendices

Executive Summary

- Methodologies Used:
  - Data collection done through APIs and web scraping
  - Data wrangling
  - Exploratory data analysis done with SQL and visualizations
  - Interactive map exploration
  - Interactive dashboard showcasing key visuals
  - Classification models
- Summary of all results
  - More recent flights show higher success rate
  - The site with the highest success rate is KSC LC-39A at 76.9%
  - The logistic regression model is best suited for classifying whether the booster will land or not since it has good accuracy and also gives a probability with its prediction

Introduction

- The objective is to use SpaceX's launch data to determine how SpaceY's launches should be performed to ensure the highest chance of success
- Key Questions:
  - What factors affect the chances that the first stage is successfully recovered?
  - Can we predict with high probability if the first stage lands successfully?

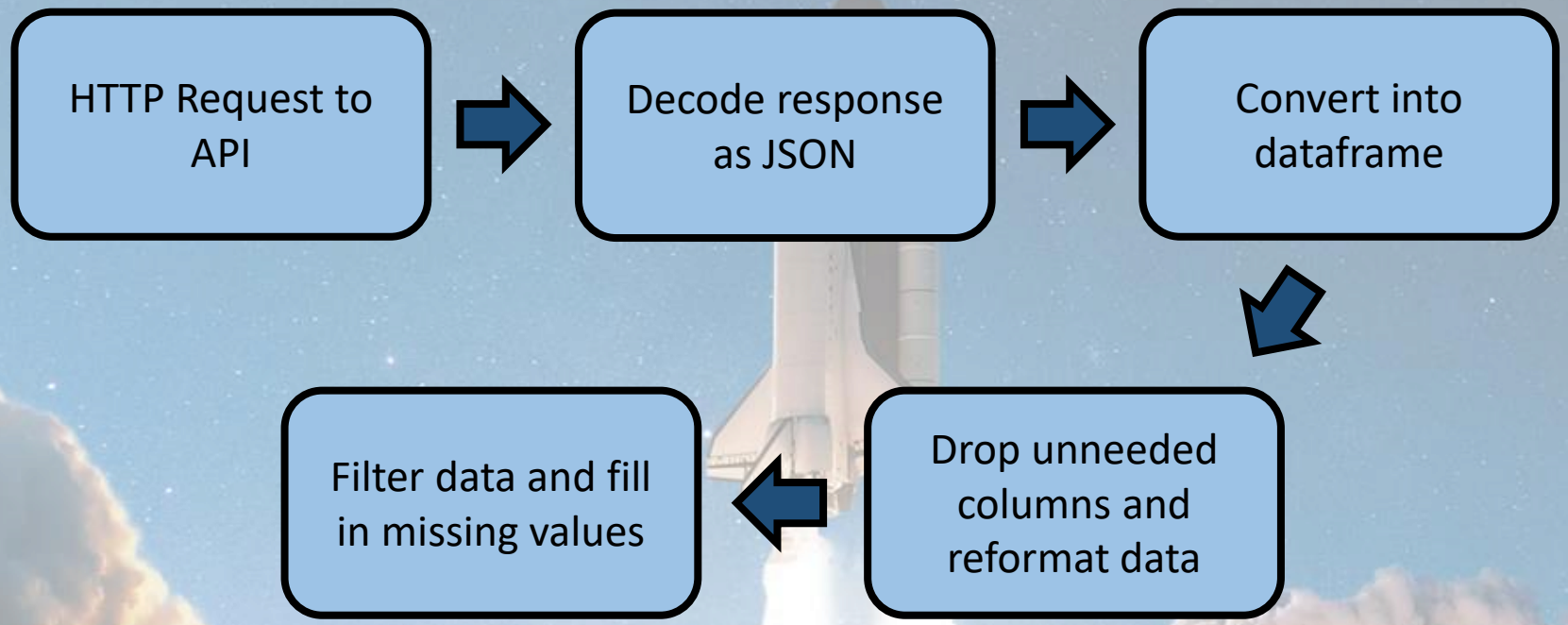
Section 1

Methodology

Methodology

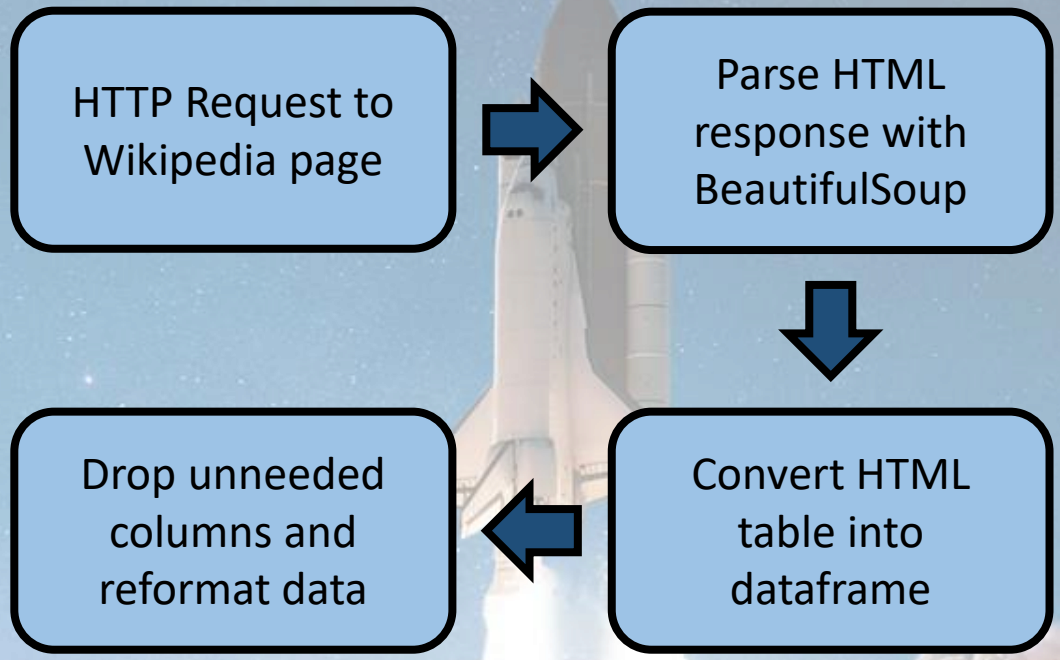
- Data collection methodology:
  - SpaceX REST API
  - Web scrapping using the SpaceX Wikipedia Page
- Perform data wrangling
  - Data filtered and cleaned using Python
  - Categorical data converted to numerical data through one hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained logistic regression, SVM, decision tree and k-nearest neighbors models and compared them against each other

Data Collection – SpaceX API



Source code

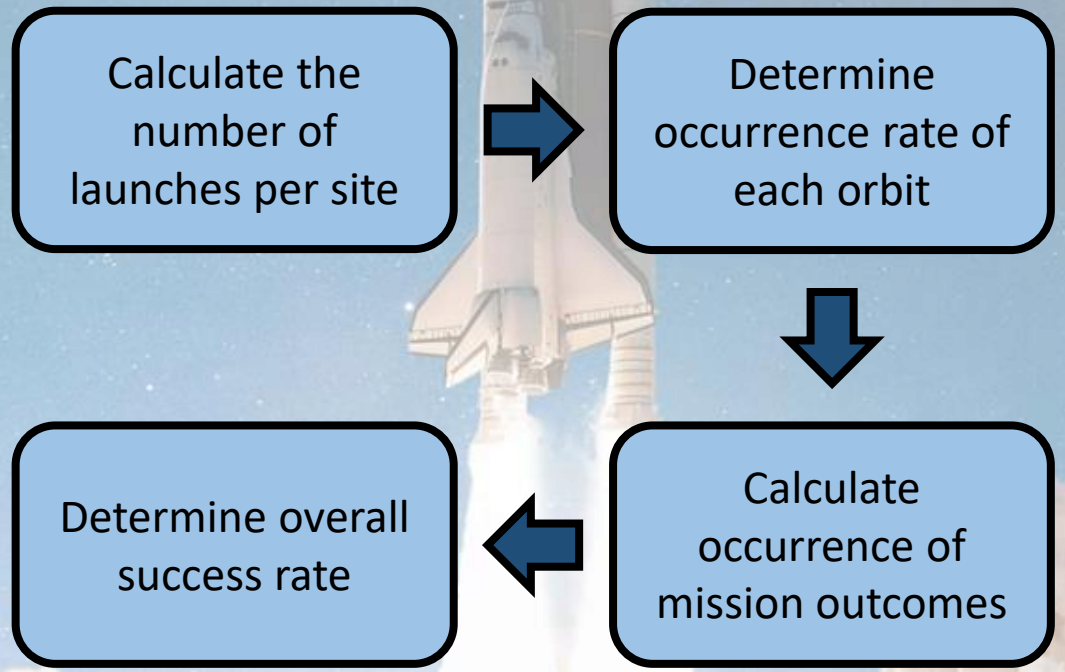
Data Collection - Scraping



Source code

Data Wrangling

- Data is cleaned and validated to ensure data meets our standards in terms of accuracy, consistency and completeness



Source code

EDA with Data Visualization

- Charts plotted:
  - Scatter plots: Flight No. vs. Payload Mass, Flight No. vs. Launch Site, Payload Mass vs. Launch Site, Flight No. vs. Orbit, Payload Mass vs. Orbit
  - Bar plot: Success Rate per Orbit
  - Line plot: Success Rate over Time
- The scatter plots and line plots allow us to observe the relationship and possible trends between two variables
- The bar plot allows us to directly compare the success rates of orbits side-by-side

Source code

EDA with SQL

- SQL queries were used to extract the following information:
  - Launch sites
  - Total payload mass carried by NASA (CRS)
  - Average payload mass carried by the F9 v1.1
  - Date of first successful landing outcome
  - Boosters with success in drone ship while having a payload mass between 4000 and 6000 kg
  - Total number of successful and failed missions
  - Booster versions that have carried the maximum payload mass
  - Failures in drone ship in the year 2015
  - Landing outcomes between June 2010 and March 2017

Source code

Build an Interactive Map with Folium

- An interactive map showing the launch sites was created using Folium
- Circles highlight the launch site areas
- Markers were placed on each launch site with green markers indicating successful launches and red markers indicating failures
- Lines were drawn from the launch site to the nearest coastline, railway, highway and city along with its distance to indicate the launch site's proximity to these objects

Source code

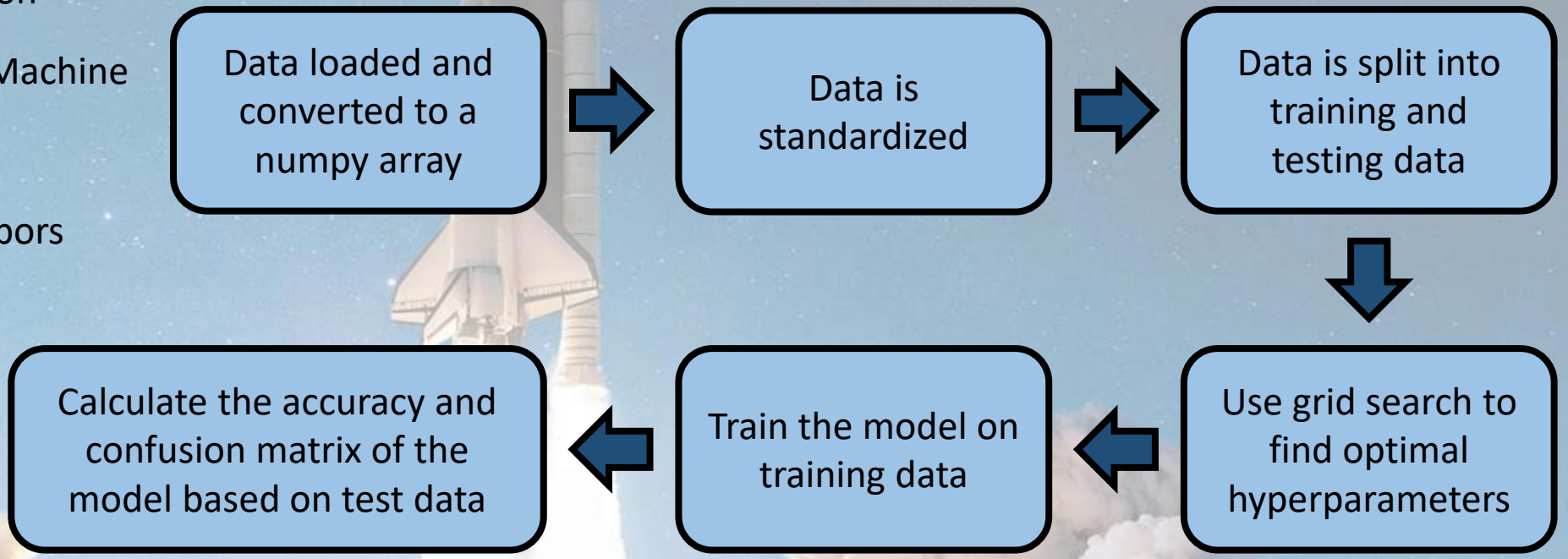
Build a Dashboard with Plotly Dash

- The dashboard has two plots
- The pie chart displays the number of successful launches at the selected site
  - This makes it easy to compare sites to each other
- The scatter plot shows how success rate correlates with payload mass and booster version
  - This plot allows for easy identification of trends based on payload mass and booster version

Source code

Predictive Analysis (Classification)

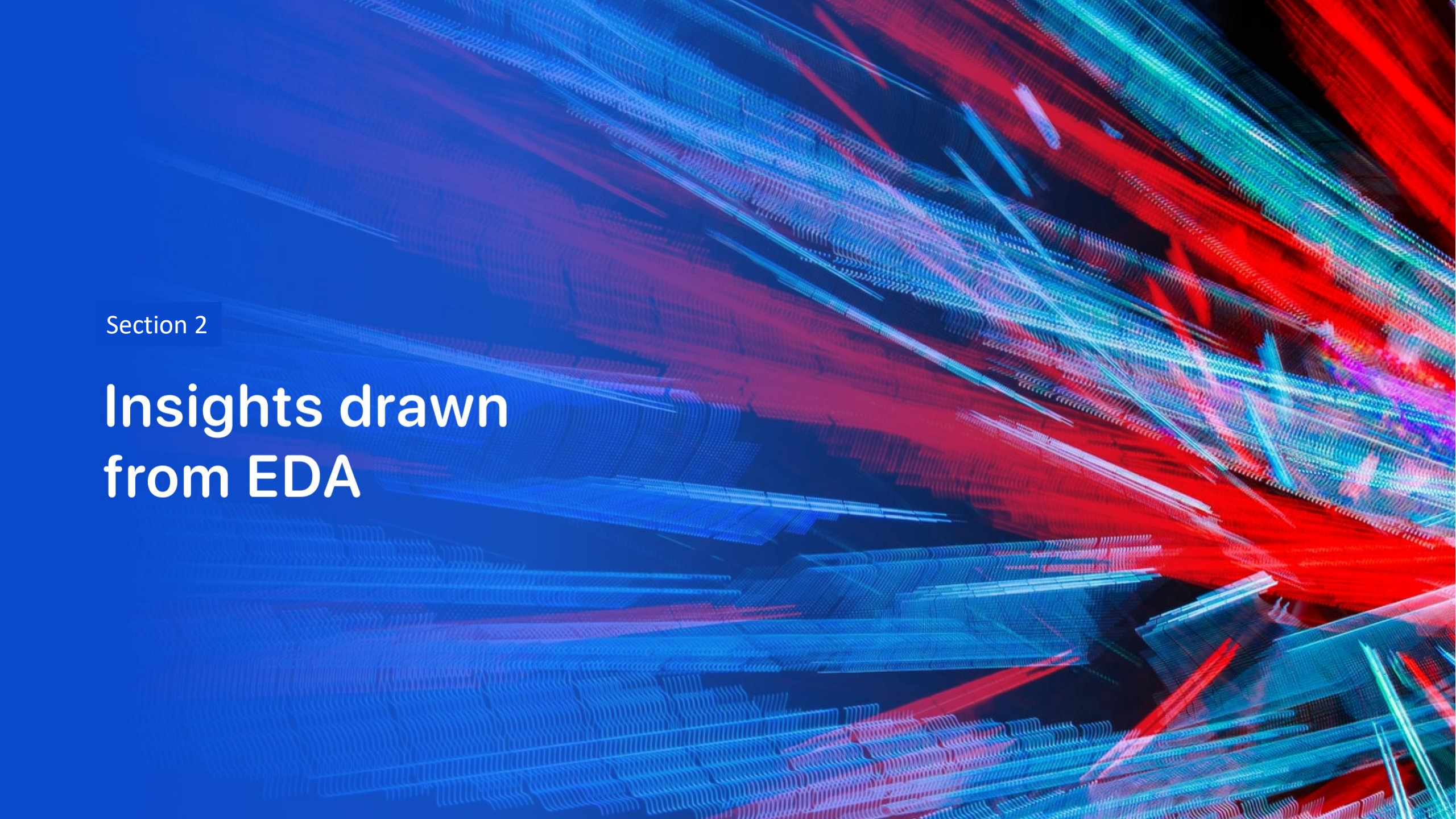
- The following flow chart was performed for four different model types:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - k-Nearest Neighbors



Source code

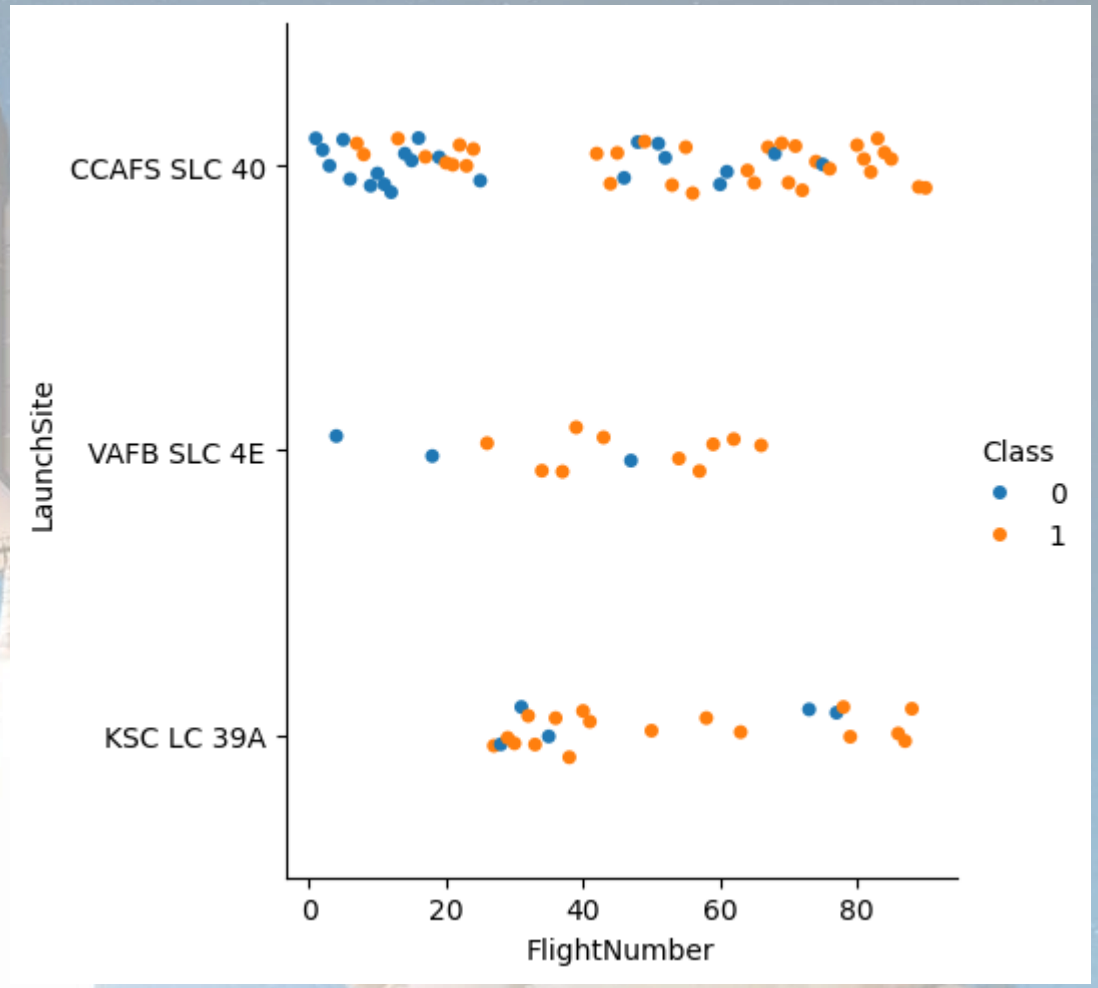
Results

- The results of the following will be shown in the following slides
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results


Section 2
Insights drawn from EDA

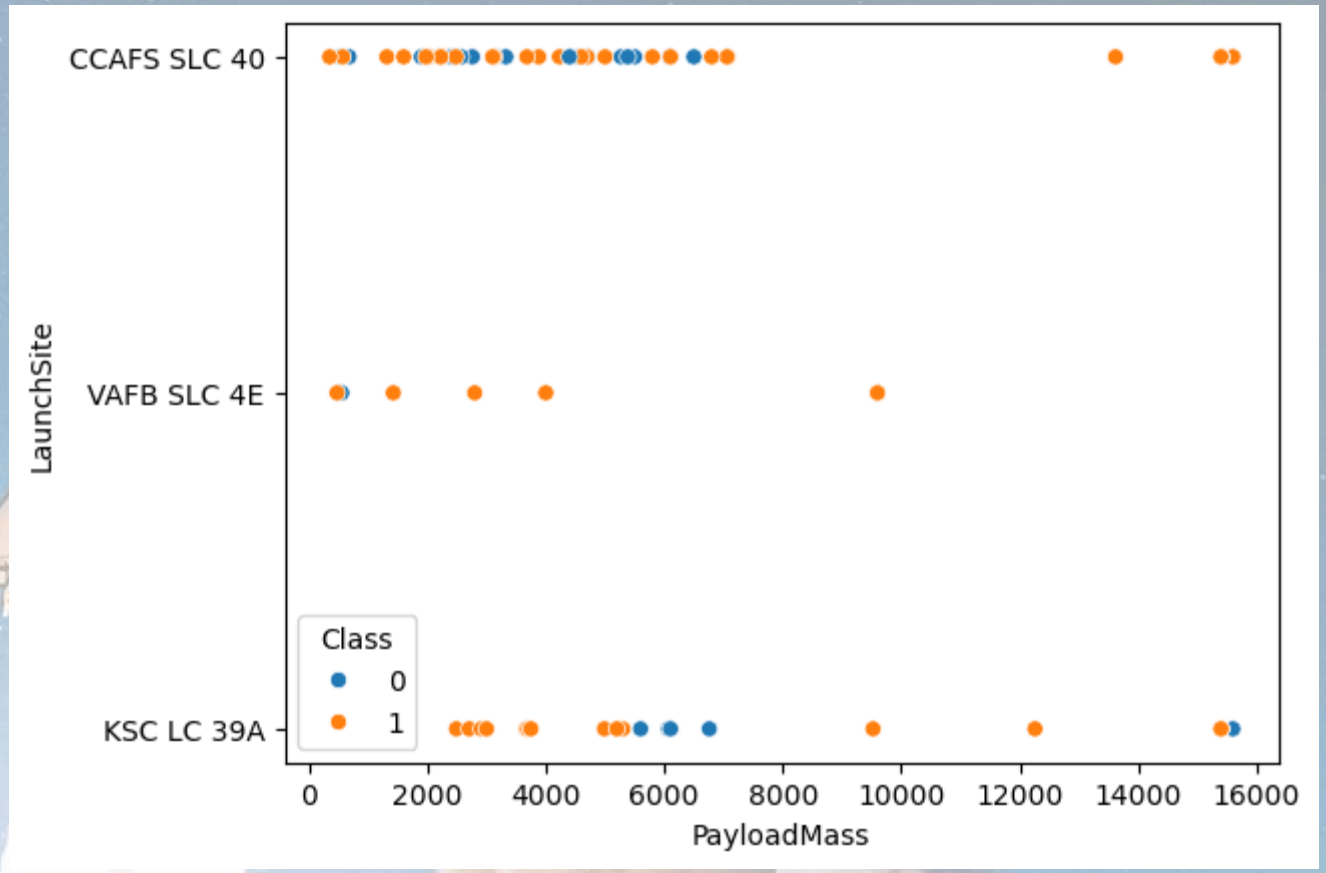
Flight Number vs. Launch Site

- Orange points indicate mission successes
- Blue points indicate mission failures
- More recent flights show higher success rate
- Initial flights at a given site have a lower success rate



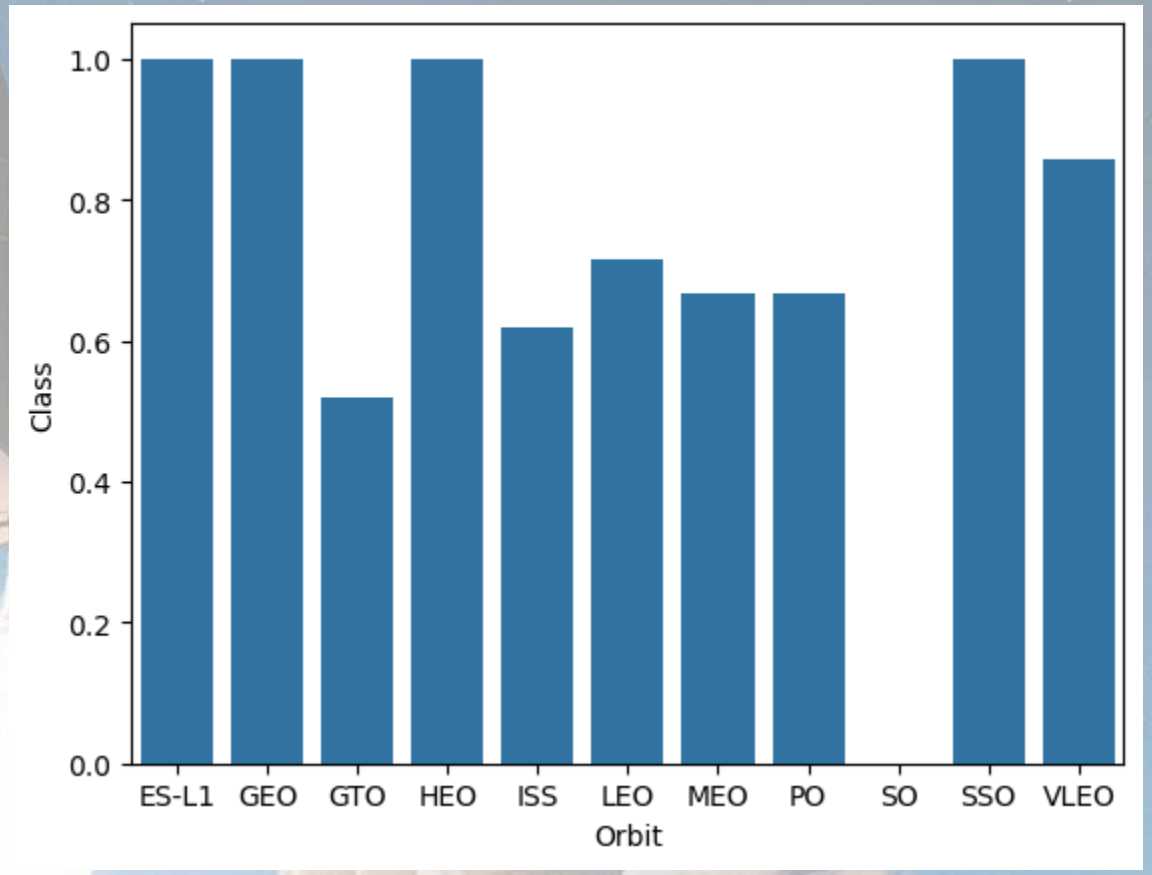
Payload vs. Launch Site

- Orange points indicate mission successes
- Blue points indicate mission failures
- Lower payload masses are more common
- Payload masses above 8000 kg have a high success rate
  - However, this could be due to small sample size



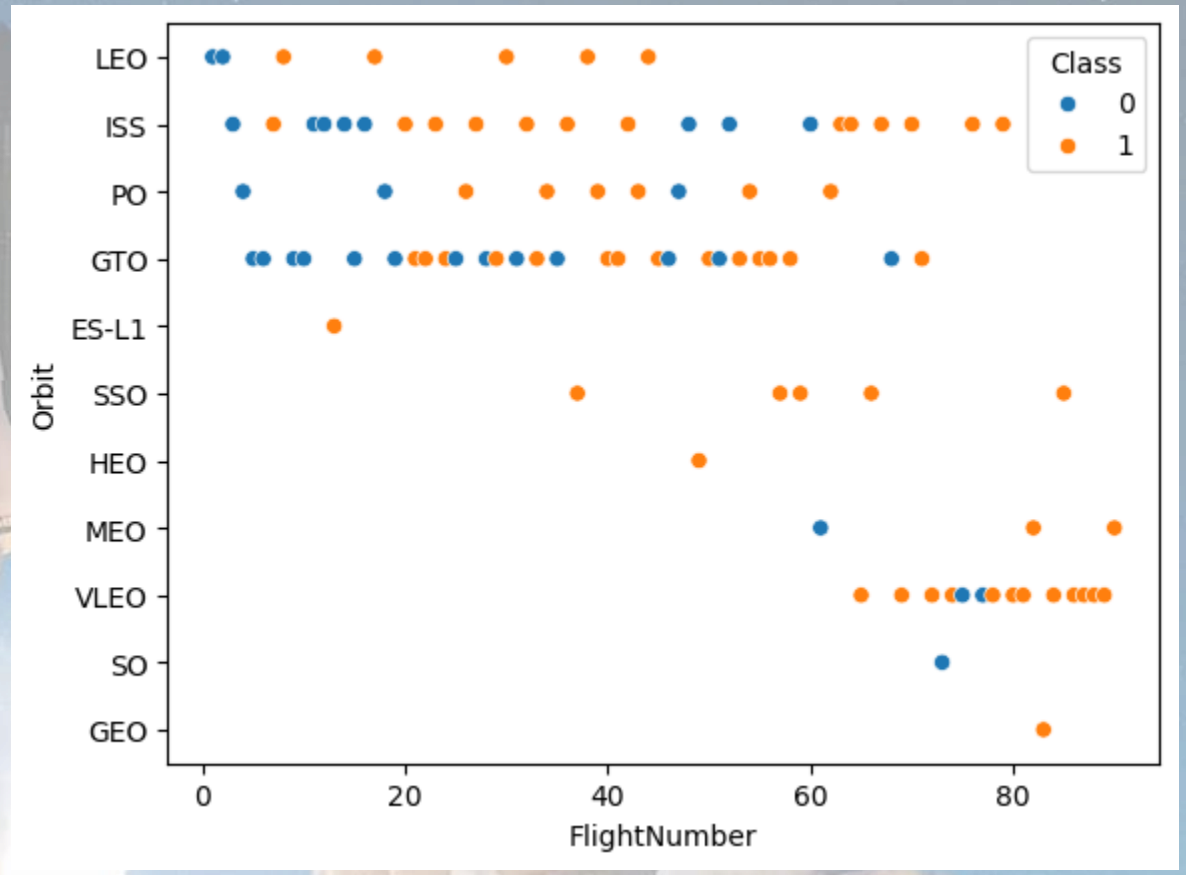
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO orbits have 100% success rate
- SO orbit has a 0% success rate
- The 5 orbits mentioned above have very low sample sizes so they are not conclusive
- VLEO has a success rate over 80% while having a sample size of 14



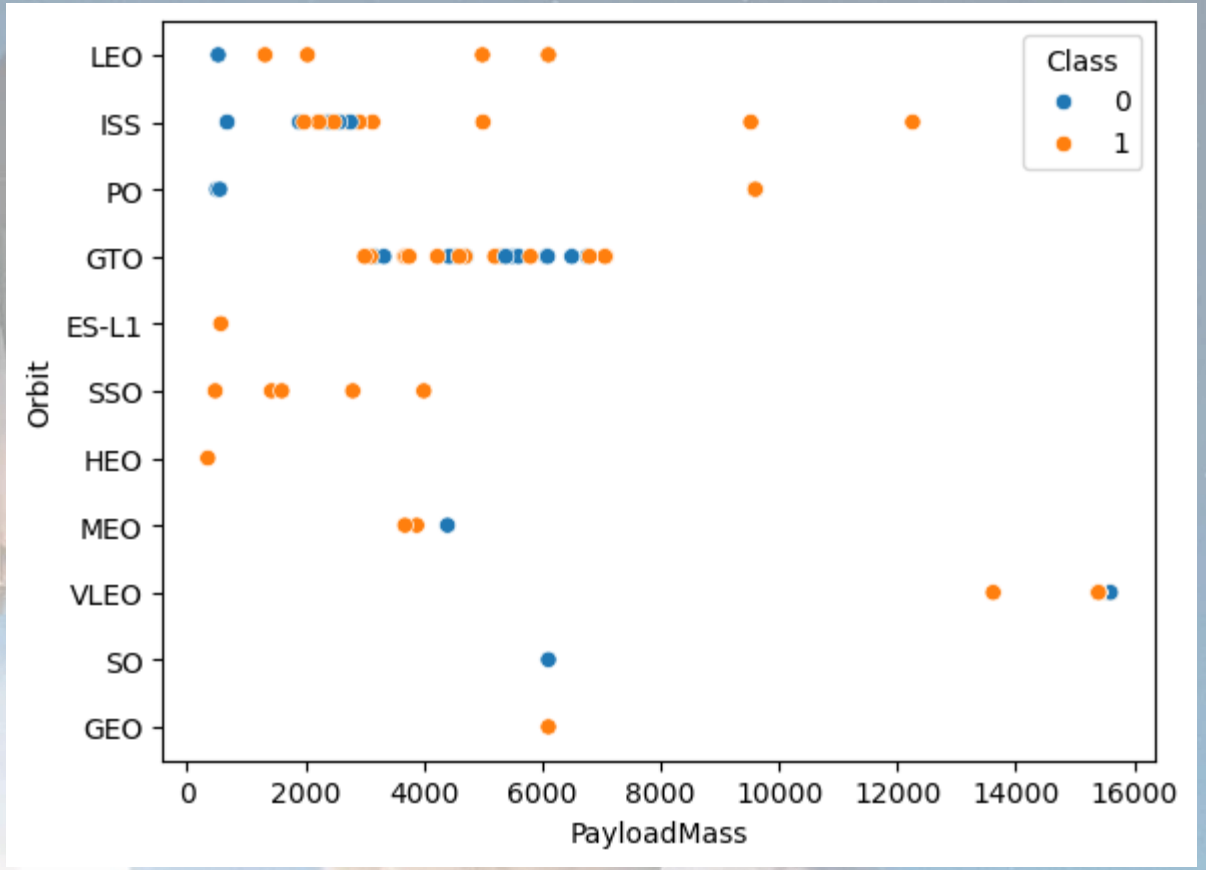
Flight Number vs. Orbit Type

- Orange points indicate mission successes
- Blue points indicate mission failures
- LEO, PO and MEO orbits have improved success rates over time
- ISS and GTO orbits have relatively consistent success rates over time



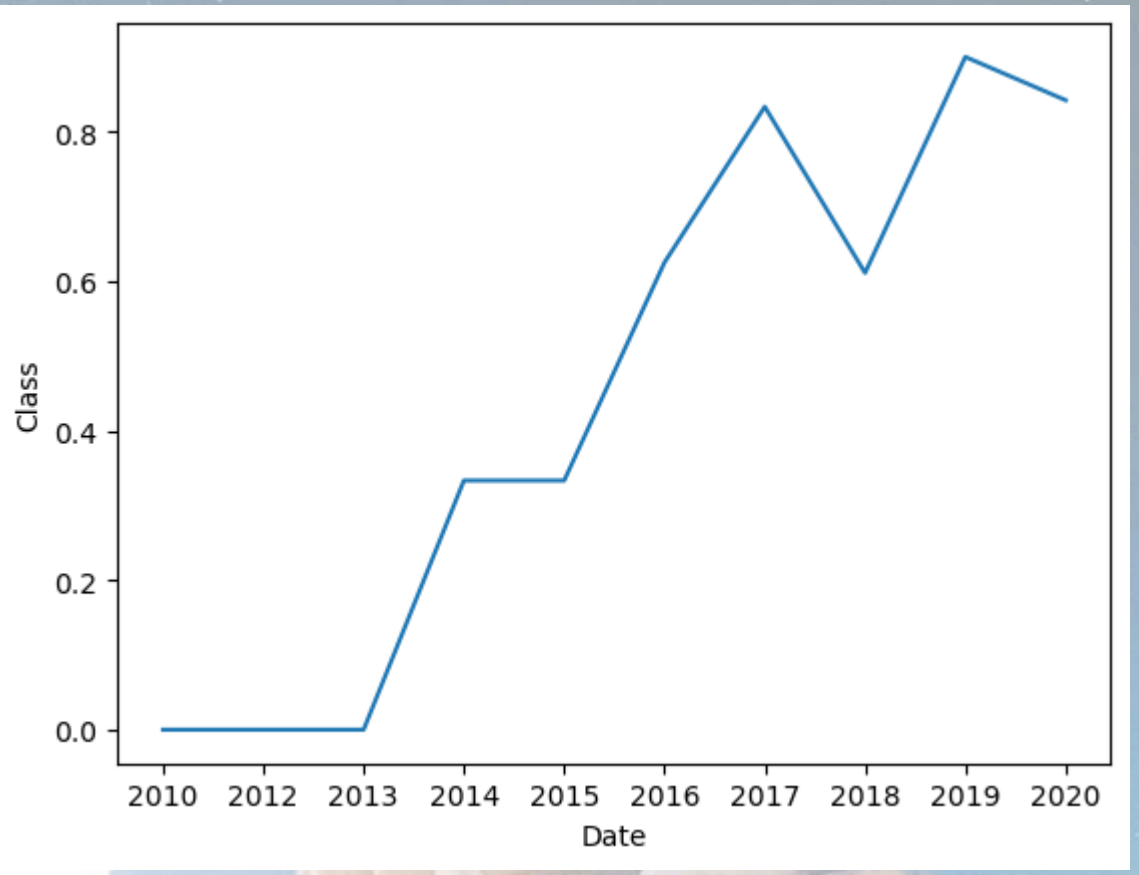
Payload vs. Orbit Type

- Orange points indicate mission successes
- Blue points indicate mission failures
- ISS orbit missions have better success rate with heavier payloads
- LEO, GTO, ES-L1, SSO, HEO, MEO, SO and GEO orbits seem to correlate with lighter payloads
- VLEO orbit missions seem to have heavier payloads



Launch Success Yearly Trend

- The success rate of missions increase over time



All Launch Site Names

- The following is the result of the query that retrieves the names of all unique launch sites:

```
%sql SELECT DISTINCT(launch_site) FROM SPACEXTABLE
```

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

Launch Site Names Begin with 'CCA'

- The following is the result of querying the first 5 data entries with launch sites that begin with ‘CCA’

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

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

Total Payload Mass

- The total payload mass carried by boosters for the customer NASA (CRS) was found to be 45,596 kg

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by the F9 v1.1 booster was found to be 2534.7 kg

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE '%F9 v1.1%'
```

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful ground pad landing was found to be December 22, 2015

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The following boosters have had a successful drone ship landing with a payload between 4000 kg and 6000 kg

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

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

Total Number of Successful and Failure Mission Outcomes

- The following query counts the number of mission successes and failures

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

Boosters Carried Maximum Payload

- The following query shows the booster versions that have carried the maximum payload mass

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

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 following query shows the launches that failed to land on the drone ship during the year 2015

```
%%sql SELECT STRFTIME('%m', Date) AS Month, Date, Booster_Version, Launch_Site, Landing_Outcome
      FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND  STRFTIME('%Y', Date) = '2015'
```

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)

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

- The following query presents the number of each landing outcome between 2010-06-04 and 2017-03-20 in descending order

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

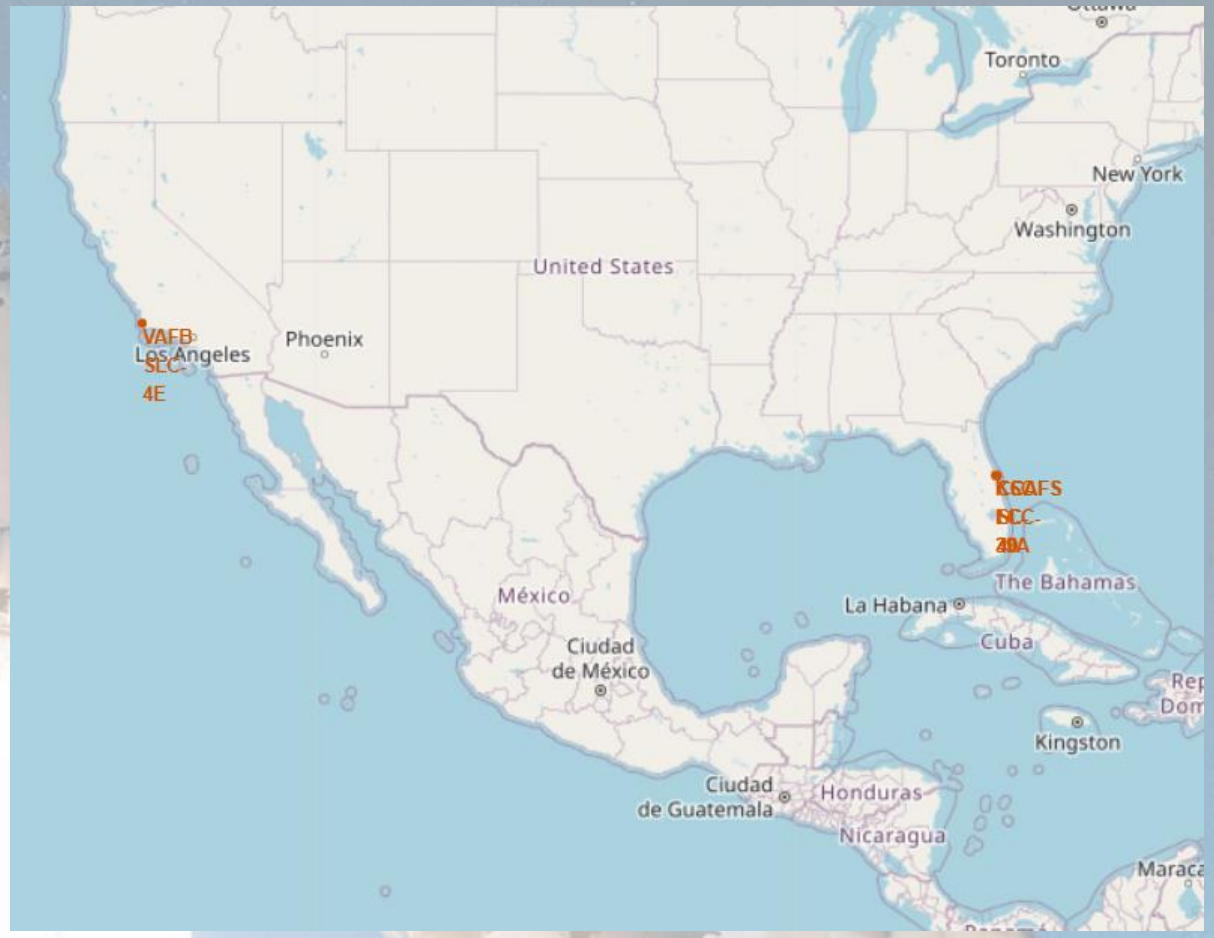
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

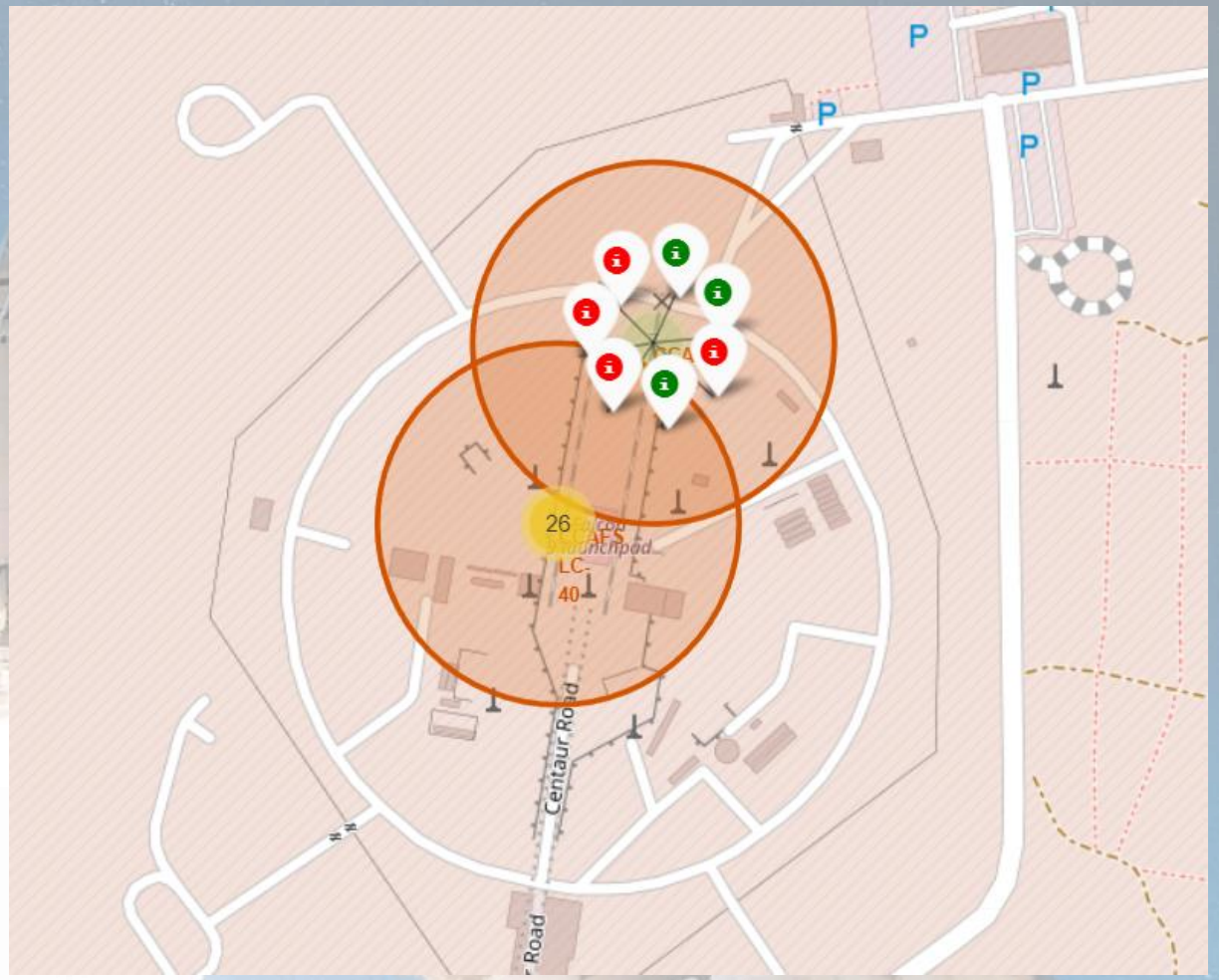
Folium Map of Launch Sites

- The interactive map generated by folium shows the launch sites are concentrated on two locations in the continental United States
- The locations are close to the ocean to reduce the possibility of accidents causing damage and harm to civilians and infrastructure



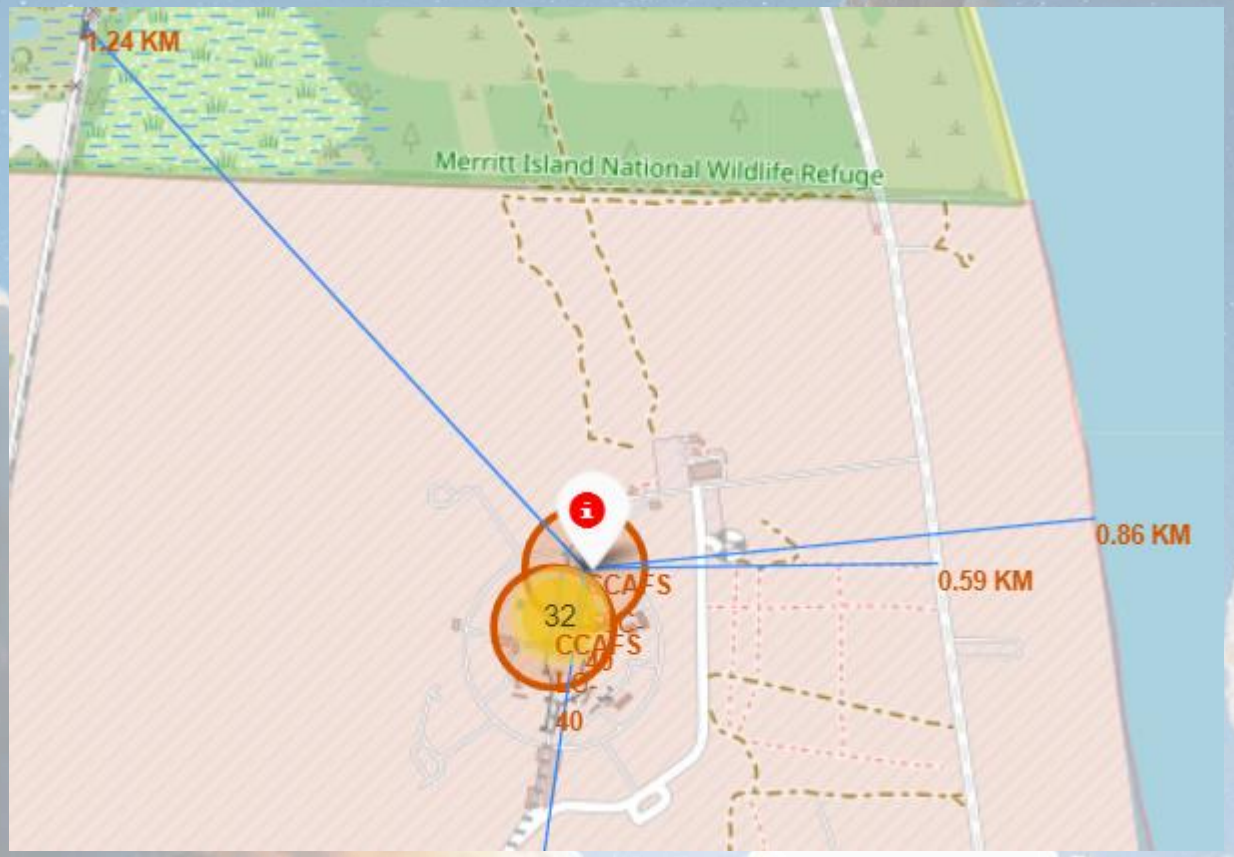
Folium Map of Successes and Failures

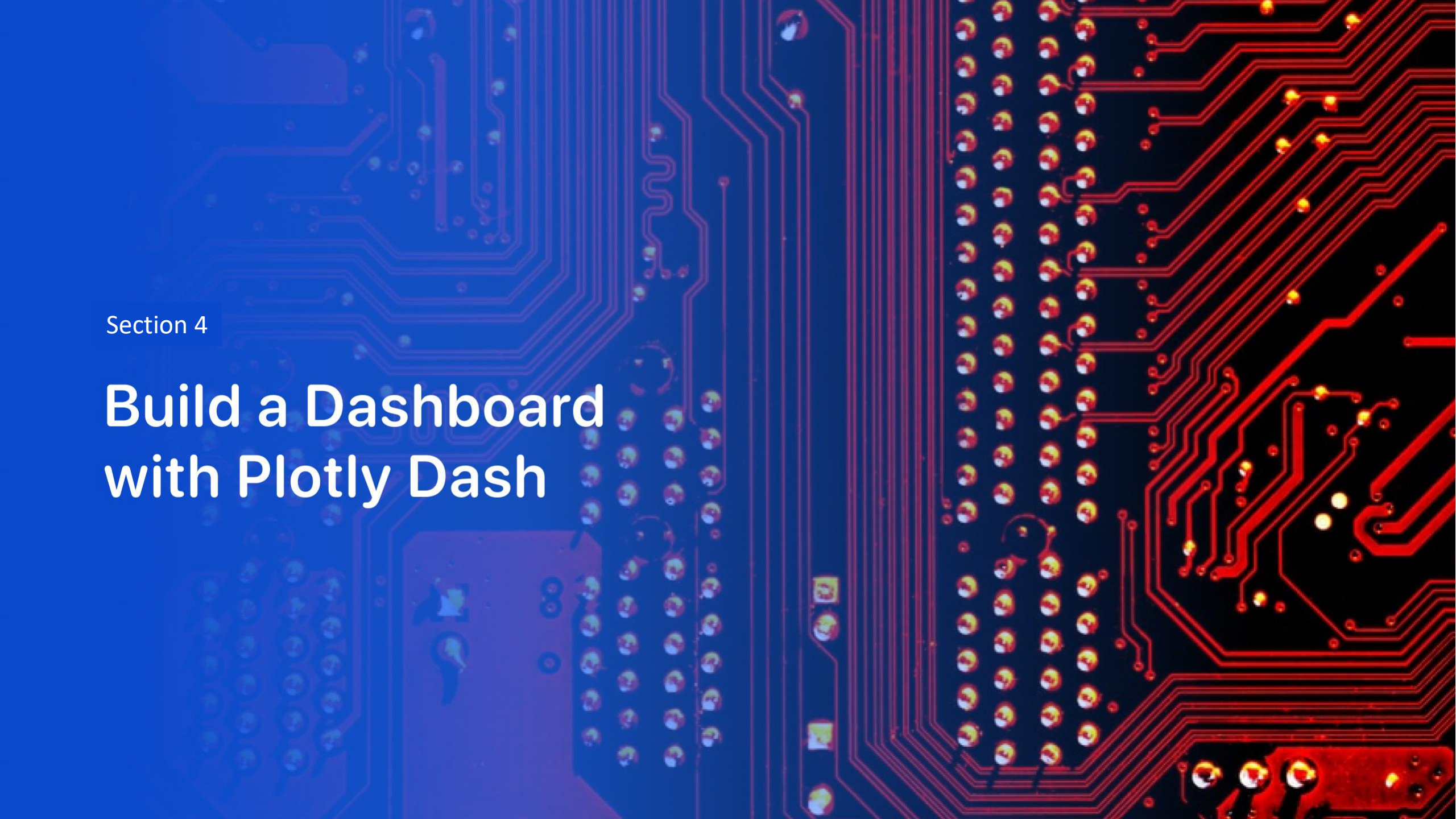
- This Folium map shows the number of launches at each site with markers
- Green markers indicate a successful launch
- Red markers indicate a failed launch



Folium Map of Launch Site Proximities

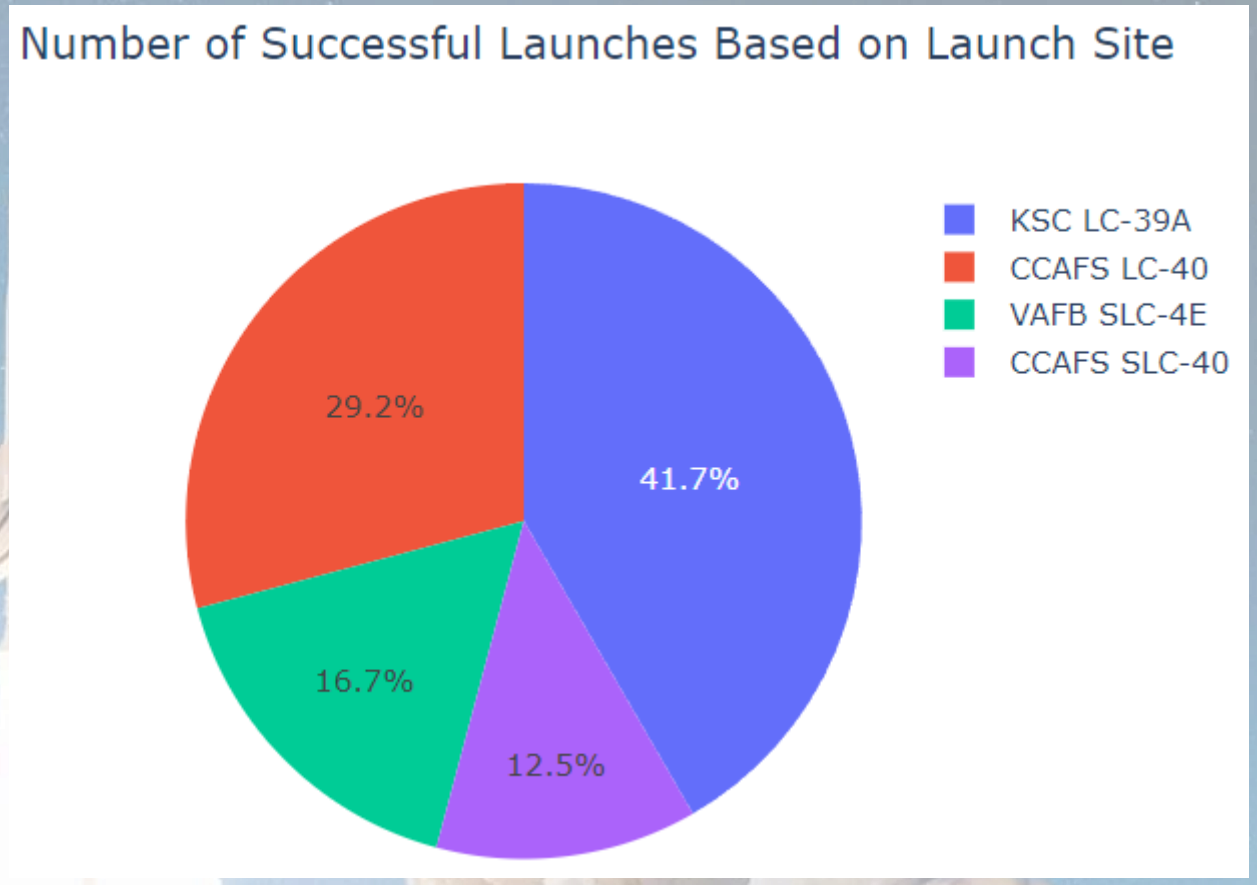
- This map shows the distances of the nearest coastline, highway, railroad and city to the CCAFS SLC-40 launch site




Section 4
Build a Dashboard with Plotly Dash

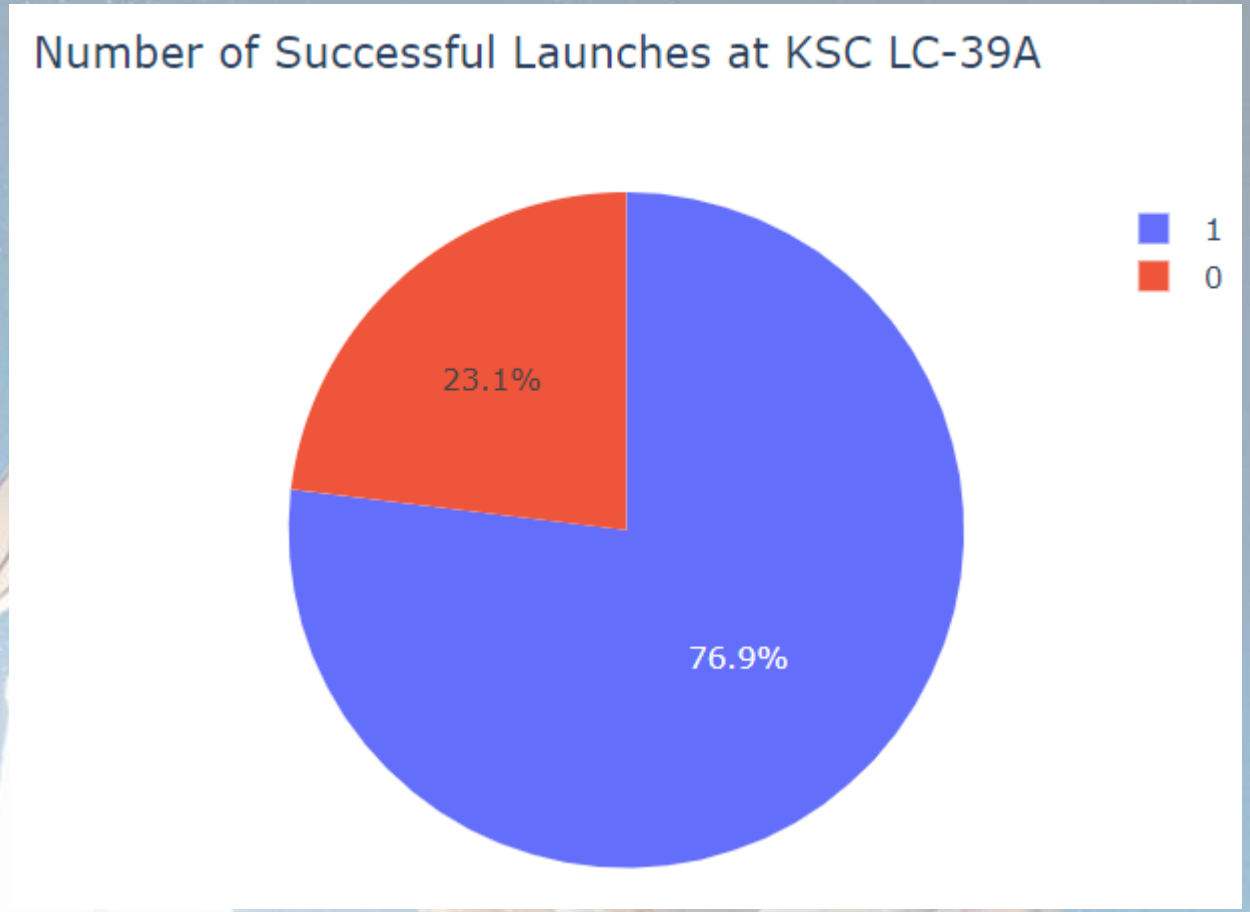
Dashboard Pie Chart of Successful Launches for All Sites

- The site with the highest number of successes is KSC LC-39A
- The site with the least number of successes is CCAFS SLC-40



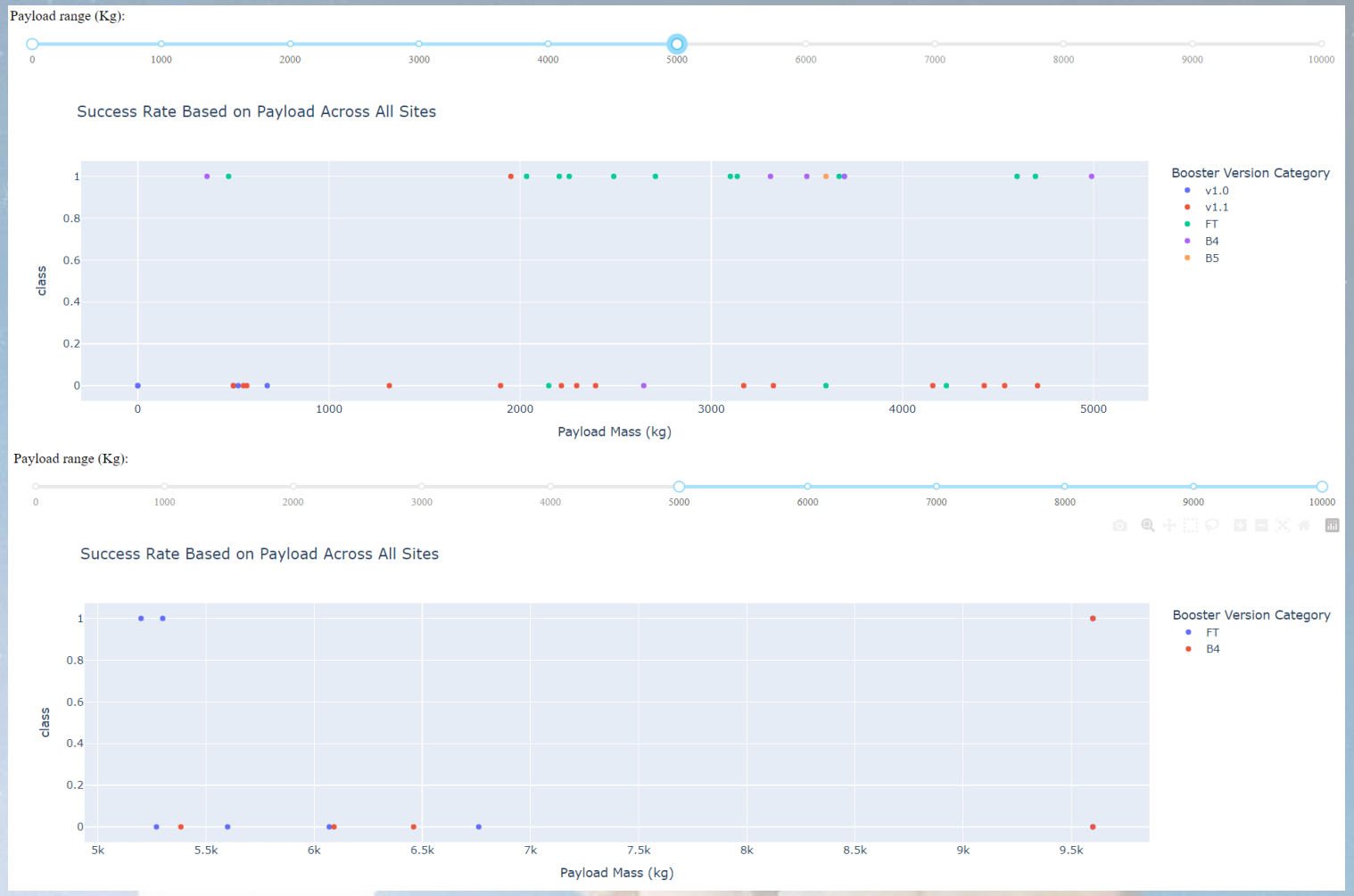
Dashboard Pie Chart of Success Rate at KSC LC-39A

- The site with the highest success rate is KSC LC-39A
- It's success rate is 76.9%



Dashboard Scatter Plot of Payload vs. Launch Outcome Across All Sltes

- FT boosters seem to have a high success rate with payloads under 5000 kg
- V1.1 boosters seem to have a low success rate with payloads under 5000 kg
- Only two booster version categories have had a payload over 5000 kg

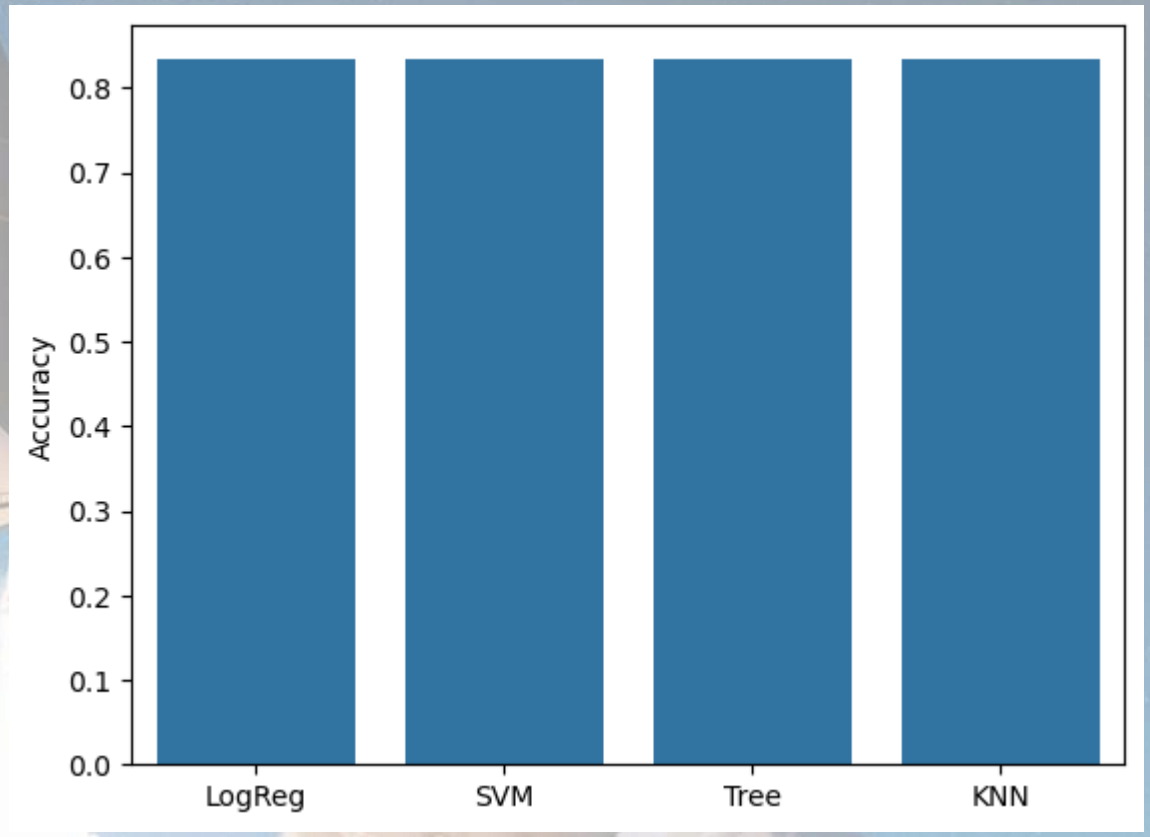


Section 5

Predictive Analysis (Classification)

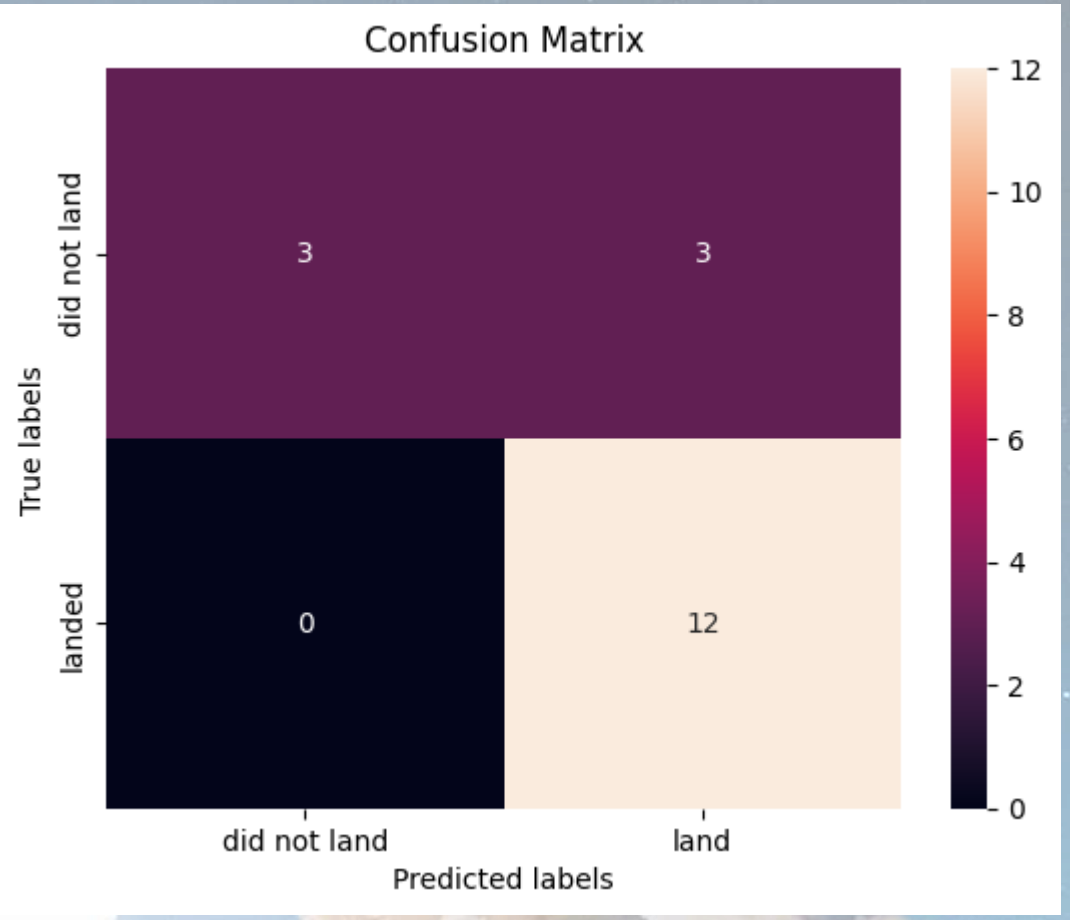
Classification Accuracy

- All four models that were trained had the same classification accuracy
- The logistic regression model was chosen as the best model since it not only predicts the class but the probability as well



Confusion Matrix for Logistic Regression Model

- The confusion matrix shows that the model is very good at predicting the outcome when the booster lands
- However, the model is much less accurate when the booster does not land



Conclusions

- More recent flights show higher success rate
- Lower payload masses are more common
- The VLEO orbit has a success rate over 80% while having a relatively large sample size
- LEO, PO and MEO orbits have improved success rates over time
- The site with the highest success rate is KSC LC-39A at 76.9%
- The logistic regression model is best suited for classifying whether the booster will land or not since it has good accuracy and also gives a probability with its prediction

Thank you!