IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christopher Roth
03.01.2025

Outline

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

Methodology

Executive Summary

- Data collection methodology:
  - Data collection using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One-Hot encode categorical 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
  - How to build, tune, evaluate classification models

Introduction

Project background and context

- Space X 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 Space X 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 space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

Problems we want to find answers for

- What determines if a rocket will land successfully?
- The interaction amongs various features that determine the success rate of a successful landing.
- What operating conditions need to be in place to ensure a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One-Hot encode categorical 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
  - How to build, tune, evaluate classification models

Data Collection

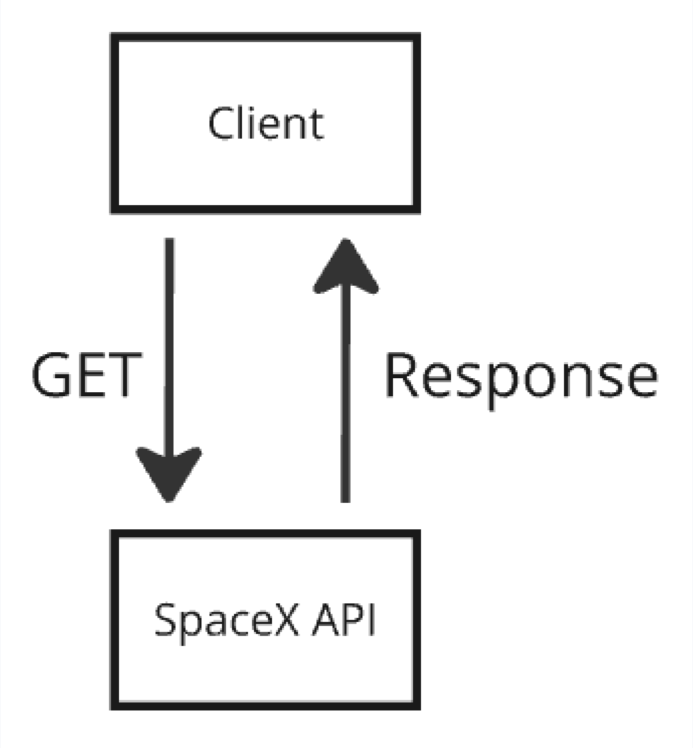
- Data was collected using the SpaceX REST API calls

- Additionally, Wikipedia was scraped for Information about Falcon 9 Launches

Data Collection – SpaceX API

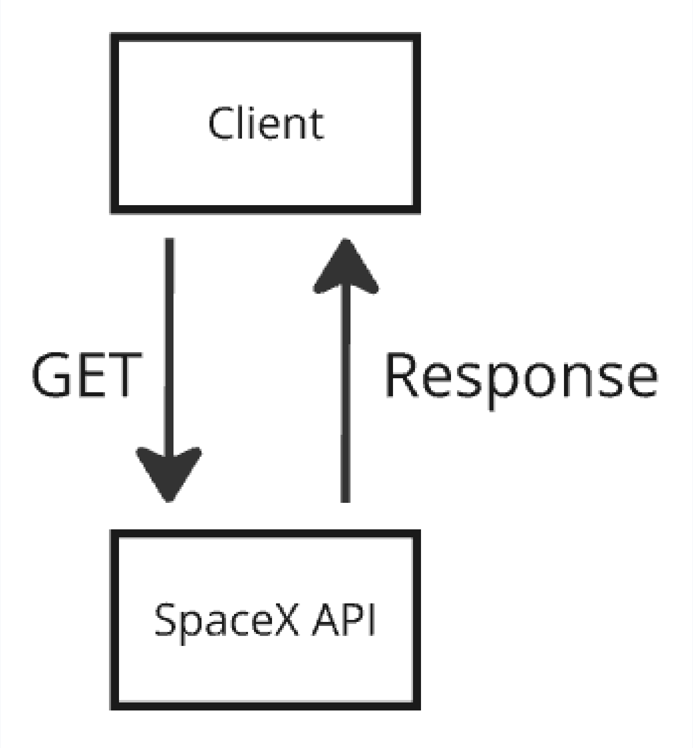
- Using the SpaceX REST API we requested data using GET calls
- Afterwards, the received json file was transformed into a pandas dataframe

- Reference: https://github.com/chrsr01/applied_capstone_dsspecialization/blob/main/Applied_DS_Capstone_1_DataCollectionAPI.ipynb



Data Collection - Scraping

- First, the respective Wikipedia page was scraped
- Using Beautifulsoup the relevant tables and their contents were extracted
- Reference: https://github.com/chrsr01/applied_capstone_dsspecialization/blob/main/Applied_DS_Capstone_1_DataCollectionScraping.ipynb



Data Wrangling

- The first step of data wrangling involved basic analysis of the distribution of specific columns, including LaunchSite and Orbit.

- In a second step, the target was transformed using a set of bad outcomes.

- Reference: https://github.com/chrsr01/applied_capstone_dsspecialization/blob/main/Applied_DS_Capstone_1_DataWrangling.ipynb

EDA with Data Visualization

- During the Data Visualization stage the following relationships were visualized:
  - Between Flight Number and Launch Site
  - Between Payload Mass and Launch Size
  - Between Success Rate and Orbit Type
  - Between Flight Number and Orbit Type
  - Between Payload Mass and Orbit Type
  - Yearly Trend of Launch Success
- These plots should help to visually identify possible relationships between the respective variables.

Reference: https://github.com/chrsr01/applied_capstone_dsspecialization/blob/main/Applied_DS_Capstone_2_EDA_Pandas_Matplotlib.ipynb

EDA with SQL

- In the EDA with SQL stage the following queries were performed:
  - Names of unique launch sites
  - Total payload mass by boosters launched by NASA
  - Average payload of specific booster versions
  - Date of first successful landing
  - Names of boosters which have success and have a payload between 4000 and 6000kg
  - Total number of successful and failure mission outcomes
  - Names of booster versions which have carried maximum payload mass
  - …

Reference: https://github.com/chrsr01/applied_capstone_dsspecialization/blob/main/Applied_DS_Capstone_2_EDA_SQL.ipynp

Build an Interactive Map with Folium

- In the next step, an interactive map was build using Folium.
- The map included markers for the specific landing spots and lines with distances to the nearest coastline, railway, highway and city.
- It also included marker clusters for each mission, indicating if the mission was successful (green) or failed (red)

Reference: https://github.com/chrsr01/applied_capstone_dsspecialization/blob/main/Applied_DS_Capstone_3_Interactive_Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

- As final part of the Interactive Visualization Stage a plotly dash application was build.
- The user gets the option to plot pie charts for the mission outcome of specific Landing Sites.
- Additionally, he can plot a scatterplot highlighting the relationship of the total tayload and the mission outcome.

Reference: https://github.com/chrsr01/applied_capstone_dsspecialization/blob/main/Applied_DS_Capstone_3_PlotlyApp.py

Predictive Analysis (Classification)

- In the predictive analysis stage four models were evaluated:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - KNN
- Each model was hyperparameter tuned using a grid search
- The results of the best performing models were compared

Reference: https://github.com/chrsr01/applied_capstone_dsspecialization/blob/main/Applied_DS_Capstone_4_Predictive_Analytics.ipynb

Results

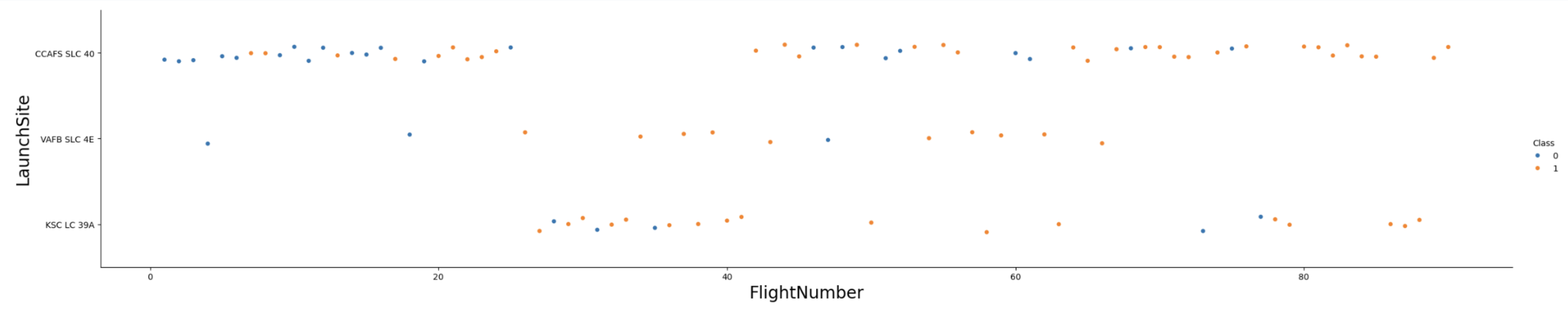
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

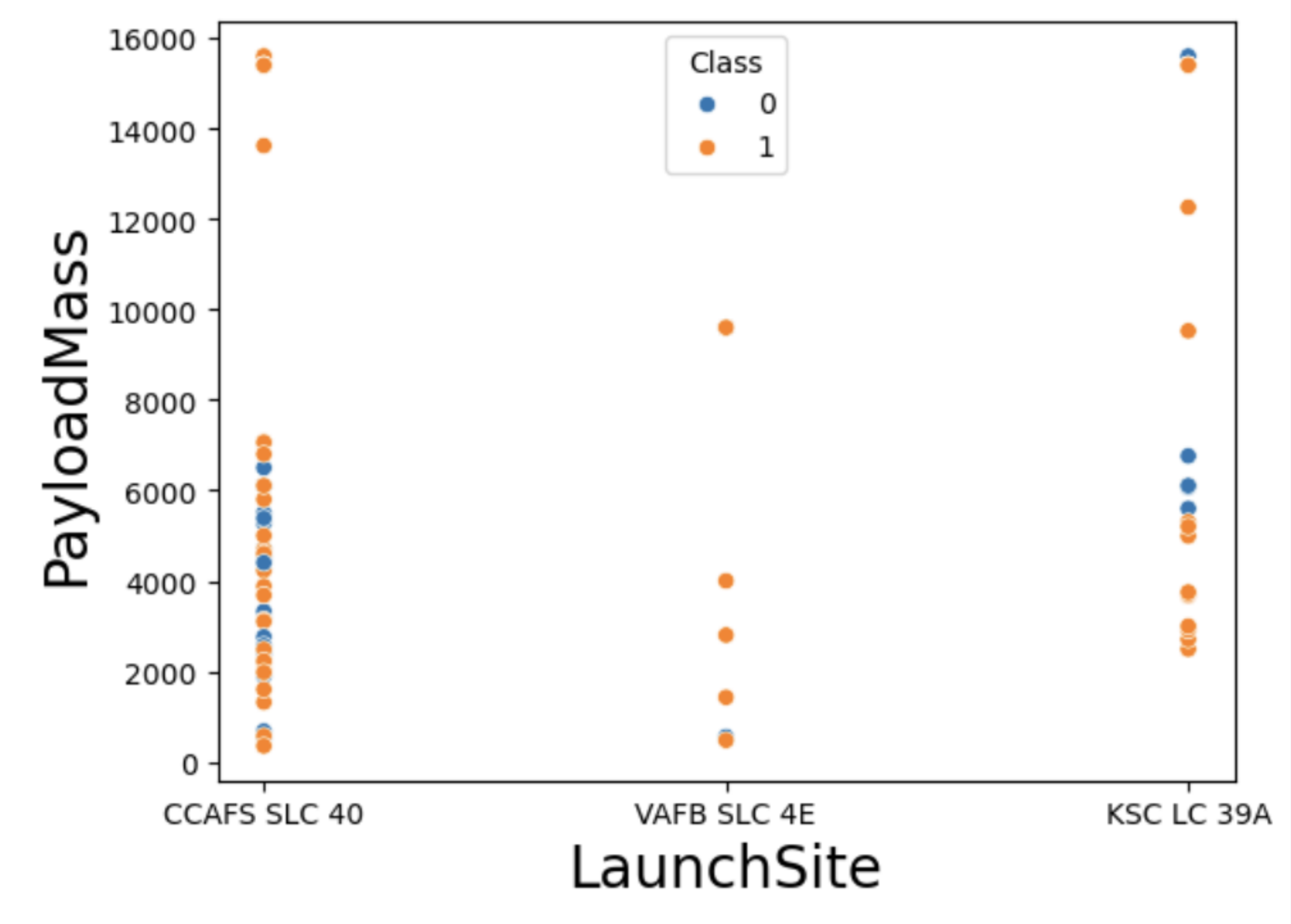
Flight Number vs. Launch Site

- **Insights**: The larger the flight number, the likelier it gets that a launch is successful



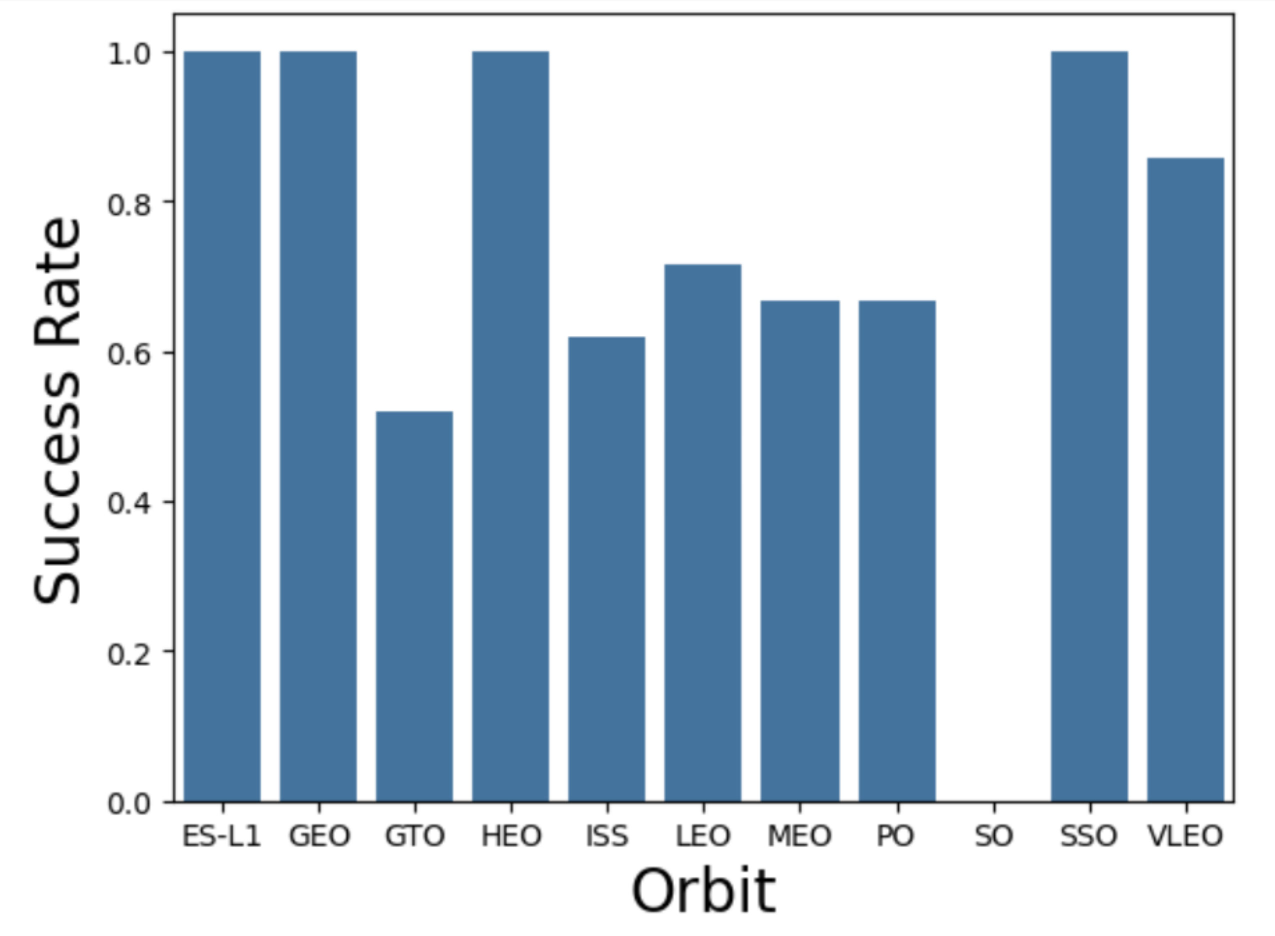
Payload vs. Launch Site

- **Insights:**
  - For Launch Site CCAFS SLC 40, the higher the payload mass, the likelier the mission is successful
  - For Launch Site KSC LC 39 A, there seems to be an interval of payload masses, where missions are not successful



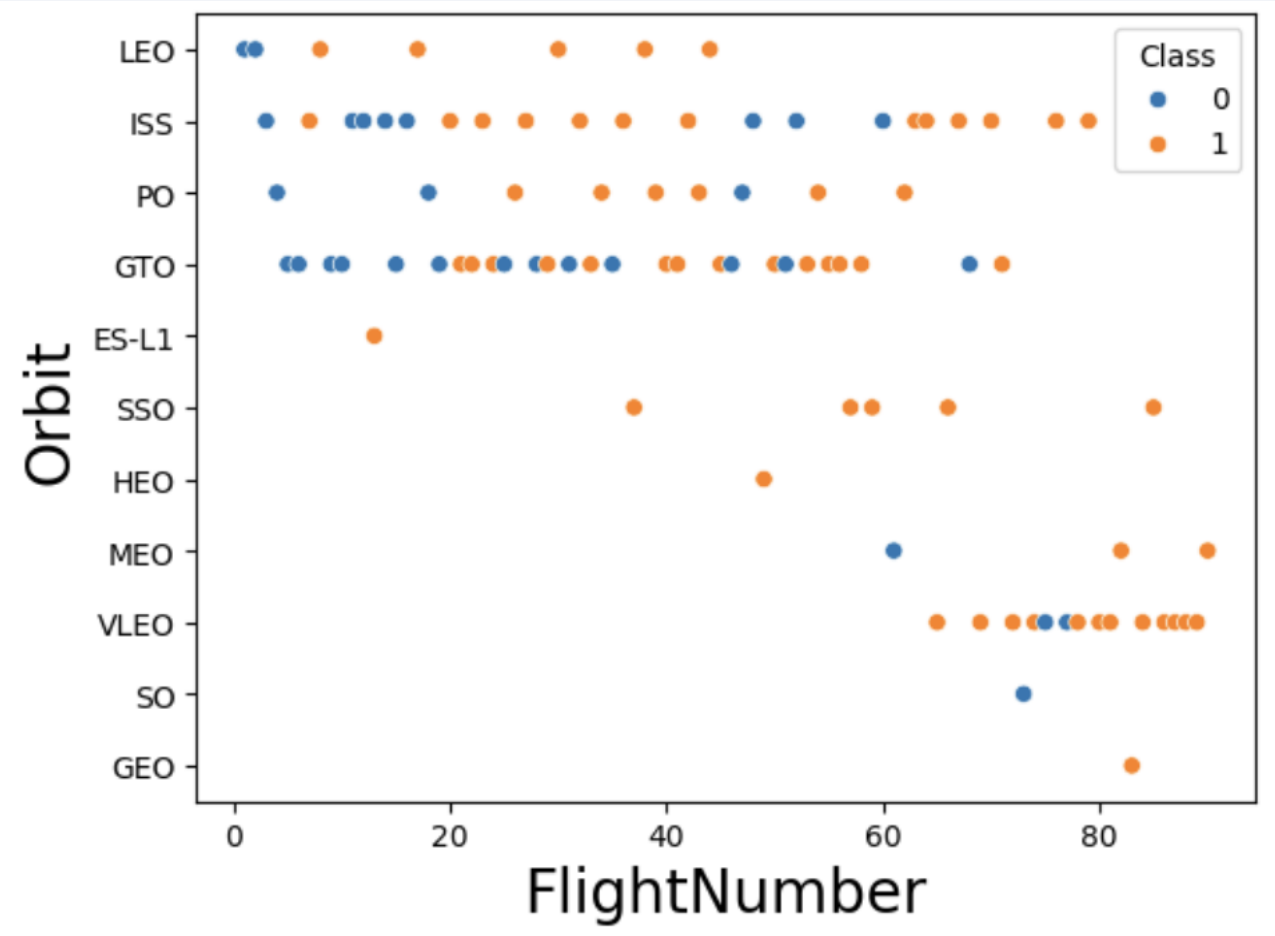
Success Rate vs. Orbit Type

- **Insights:**
  - ES-L1, GEO, SSO, HEO have perfect success rates.
  - There has been no successful mission to orbit SO



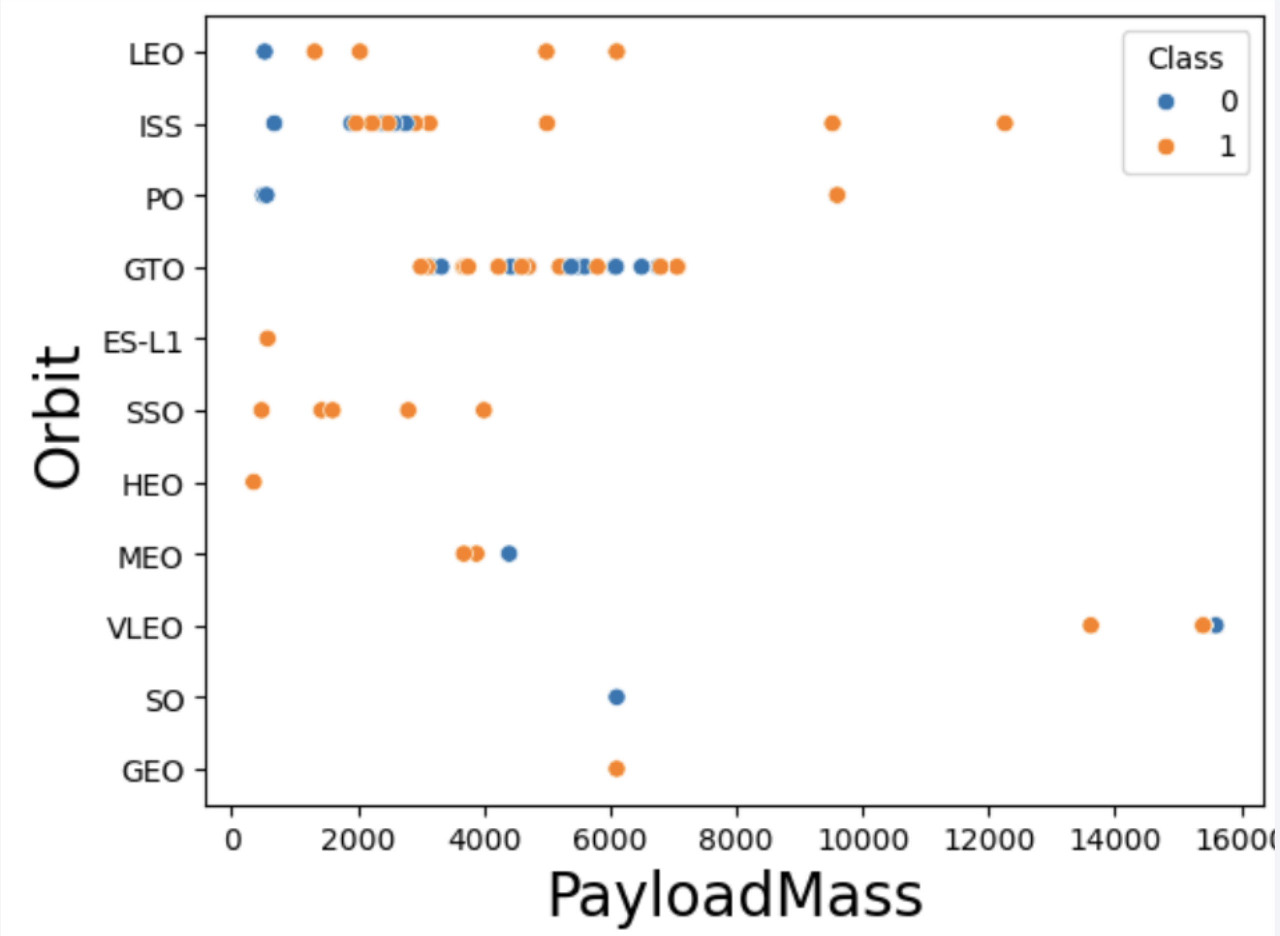
Flight Number vs. Orbit Type

- **Insights:**
  - With an increasing number of flights the missions for each orbit become more successful



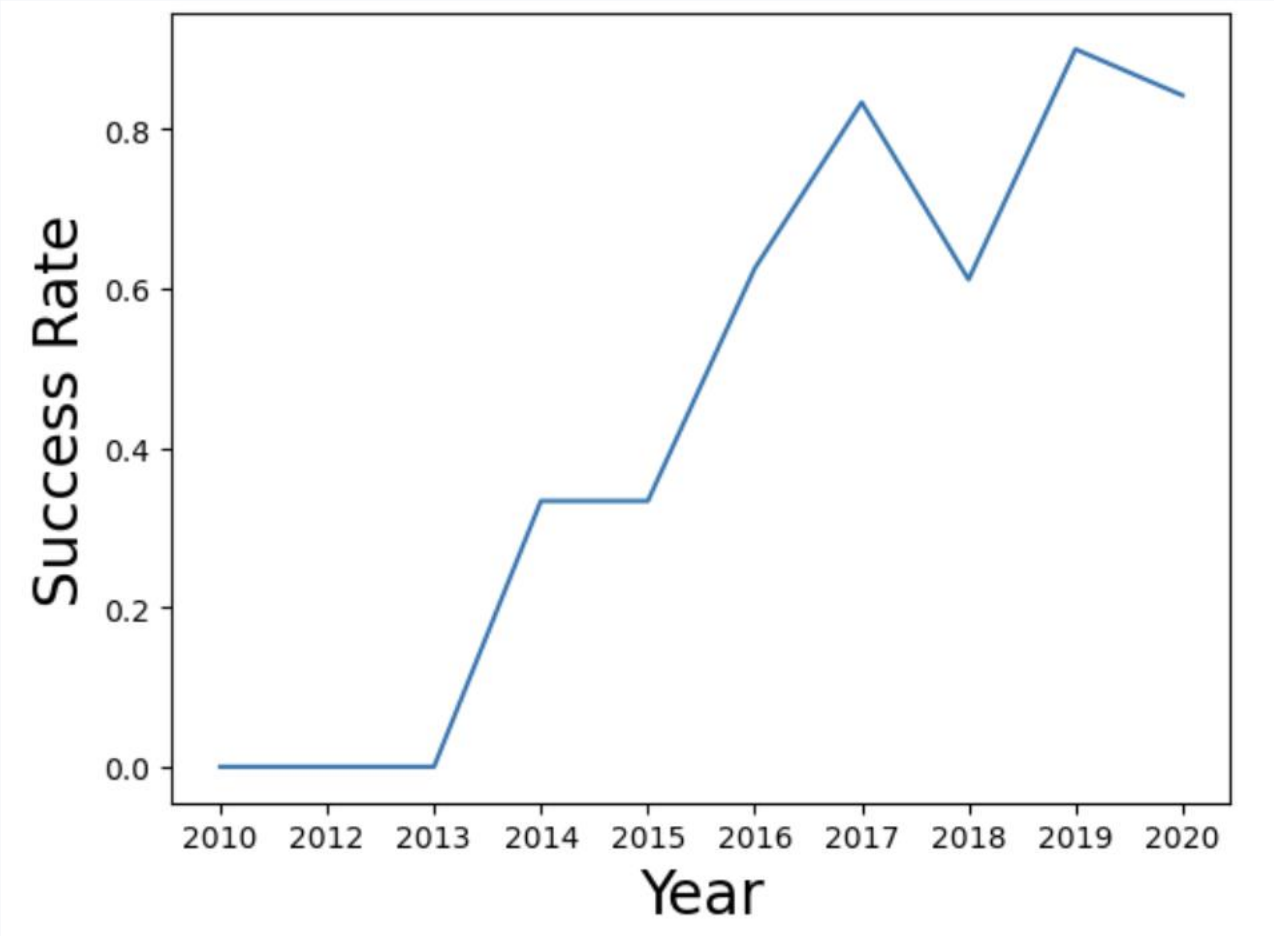
Payload vs. Orbit Type

- **Insights**:
  - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
  - However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- **Insights**:
    - the success rate since 2013 kept increasing till 2020



All Launch Site Names

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

```
%sql SELECT DISTINCT Launch_Site from SPACEXTBL
 * 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'

- The query shows 5 unordered records where the launch site begins with the string ‘CCA’

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

```
%sql SELECT * from SPACEXTBL WHERE Launch_Site like "CCA%" LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dra
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	

Total Payload Mass

- The total payload carried by rockets from NASA is 45596kg

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload from SPACEXTBL WHERE Customer = "NASA (CRS)"
```

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

Total_Payload
45596

Average Payload Mass by F9 v1.1

- The average payload carried by rockets with booster version F9 v1.1 was 2928.4 kg

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as Average_Payload from SPACEXTBL WHERE BOOSTER_VERSION = "F9 v1.1"
```

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

Average_Payload
2928.4

First Successful Ground Landing Date

- The first successful ground landing happened on 2015-12-22

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome = "Success (ground pad)"
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters that have successfully landed:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL 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

Total Number of Successful and Failure Mission Outcomes

- There were 100 successful and 1 unsuccessful missions

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

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

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

Boosters Carried Maximum Payload

- Using a sub-query that finds the highest payload mass value the booster versions were filtered

```
%sql SELECT DISTINCT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_= (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * 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

- Two Booster Versions have failed in the year 2015:
  - F9 v1.1 B1012
  - F9 v1.1 B1015

```
%sql SELECT Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure (drone ship)' AND Date BETWEEN '2015-01-01' AND '2015-12-31'
```

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

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

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

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

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

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

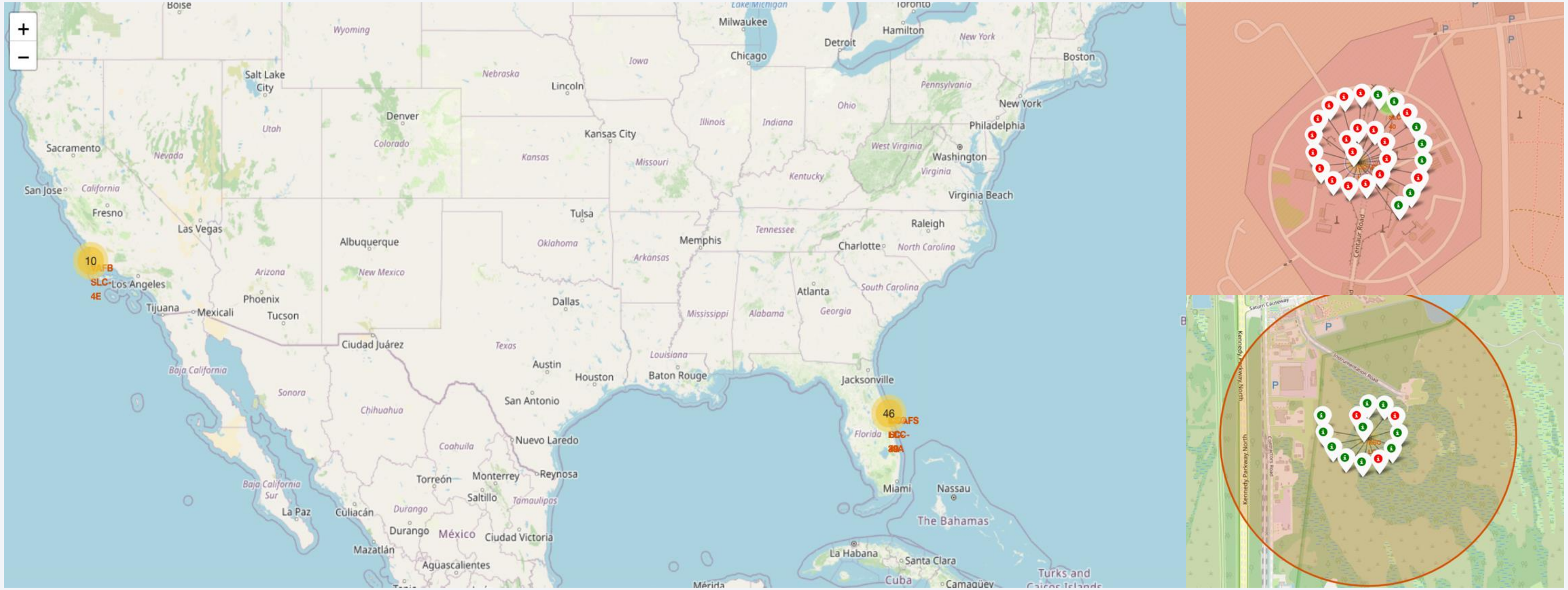
Section 3

Launch Sites Proximities Analysis

All launch sites

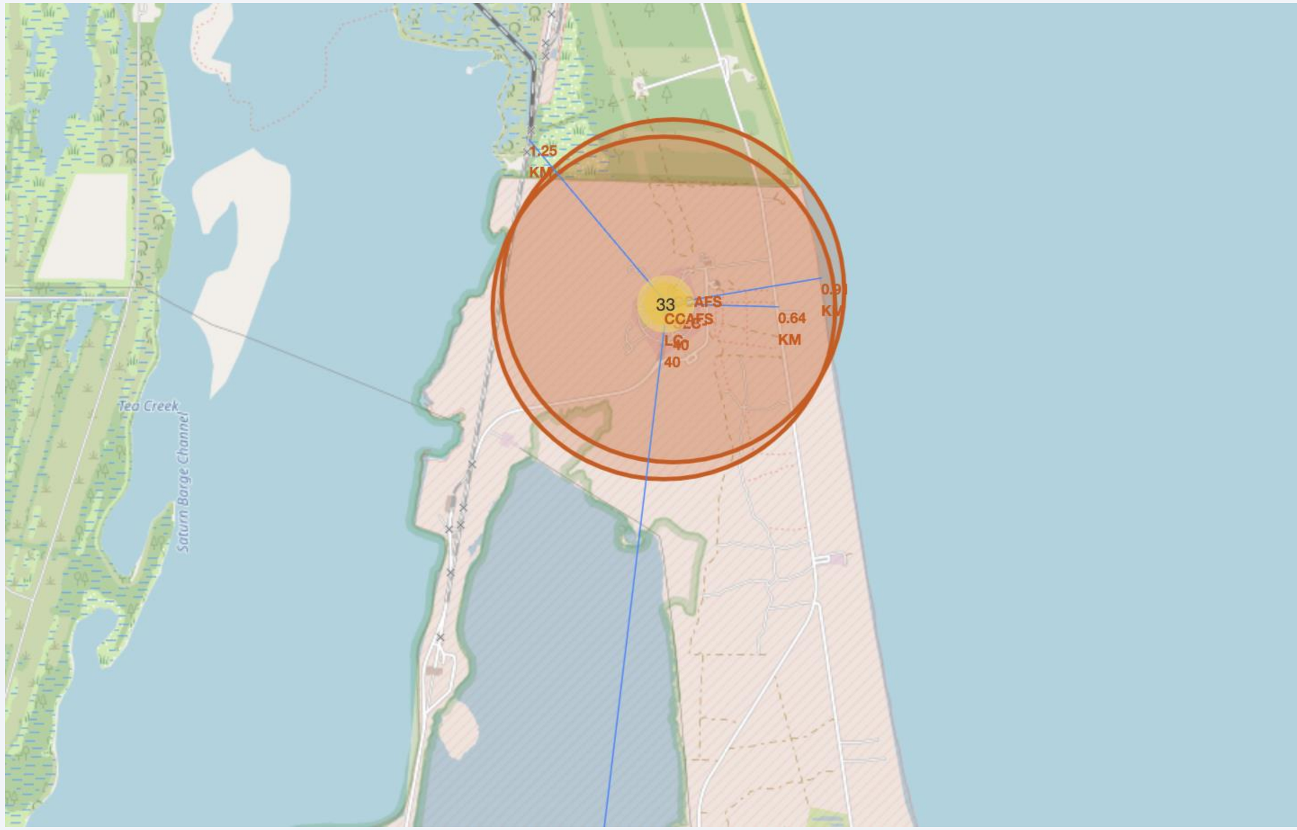
- The map shows red markers for the landing sites

Records of (un)successful landings at each Landing Site



• The map displays each record of a mission as a green/red marker if the mission was successful/unsuccessful

Proximities to infrastructure



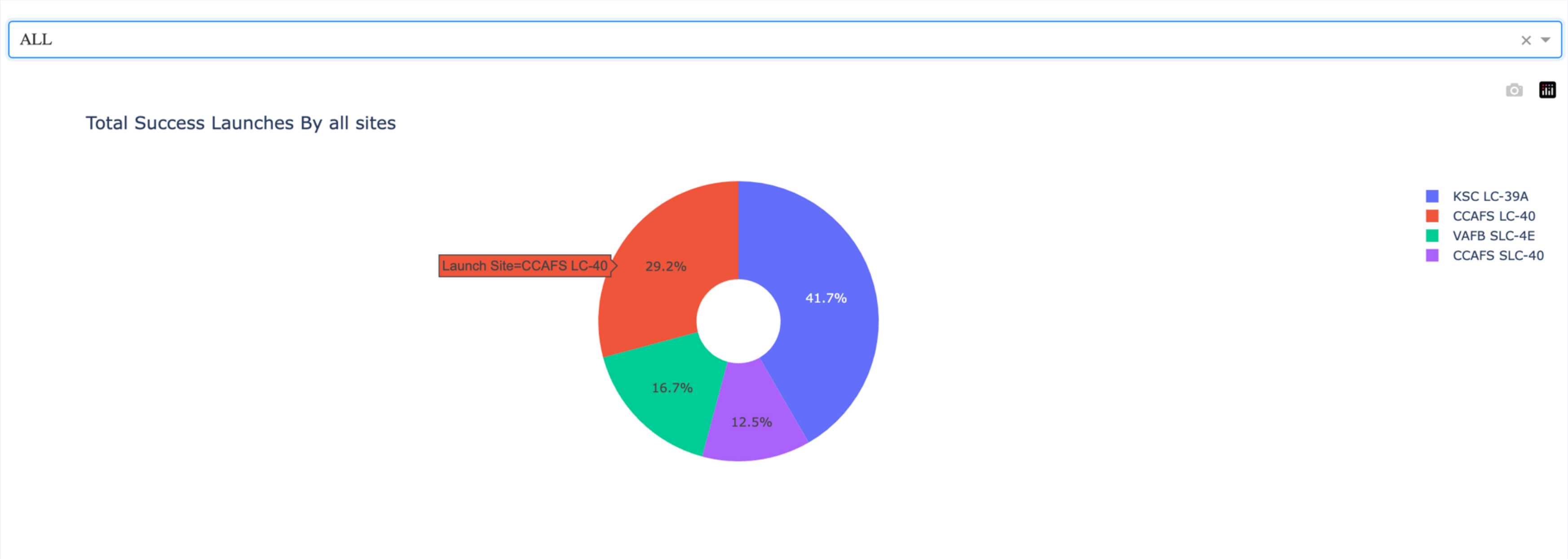
- The map displays the distances to the closest coastline, railway, highway and city.

Section 4

Build a Dashboard with Plotly Dash

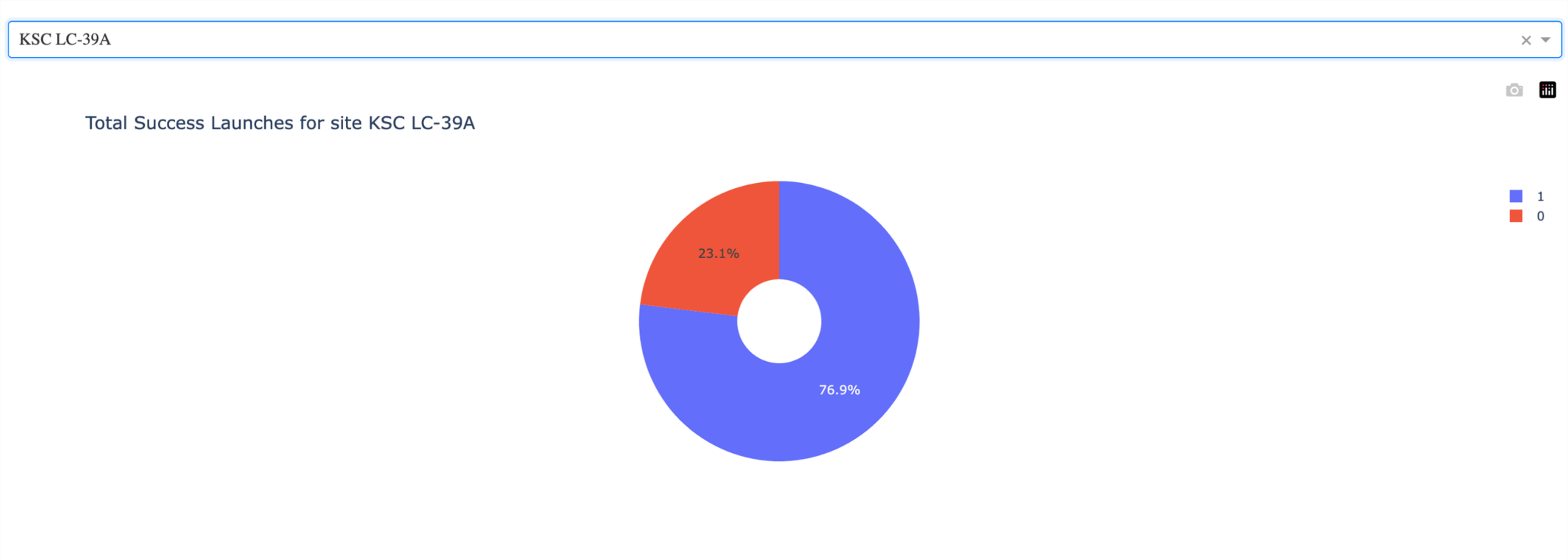
Dash: Total Success Launches

- Most successful launches came from site KSC LC-39A.



Dash: Total Success Launches for KSC LC-39A

- Site KSC LC-39A also has the highest success rate of launches.



Dash: Different Payload Ranges

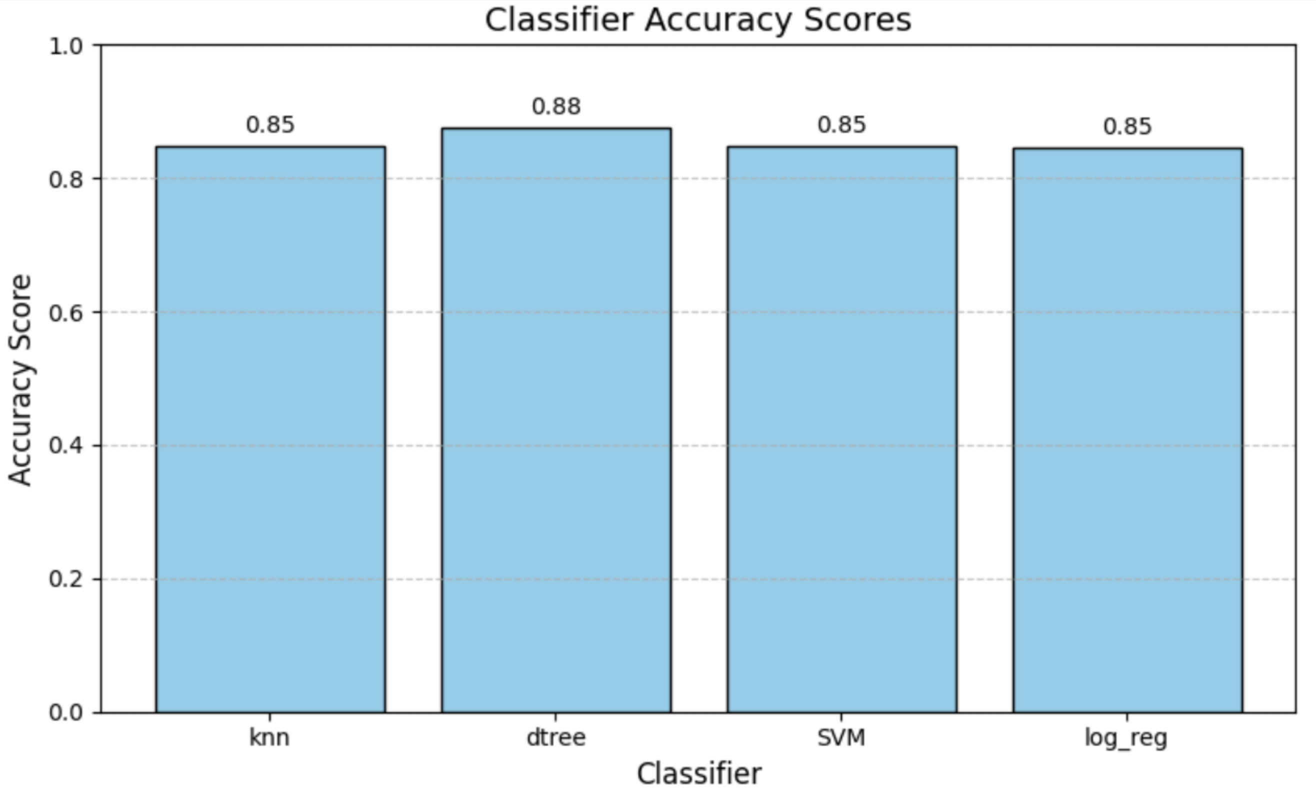
- For high payloads the success rates between the two booster categories seems to be even
- For low payloads it seems like the booster category FT has the highest success rate

Section 5

Predictive Analysis (Classification)

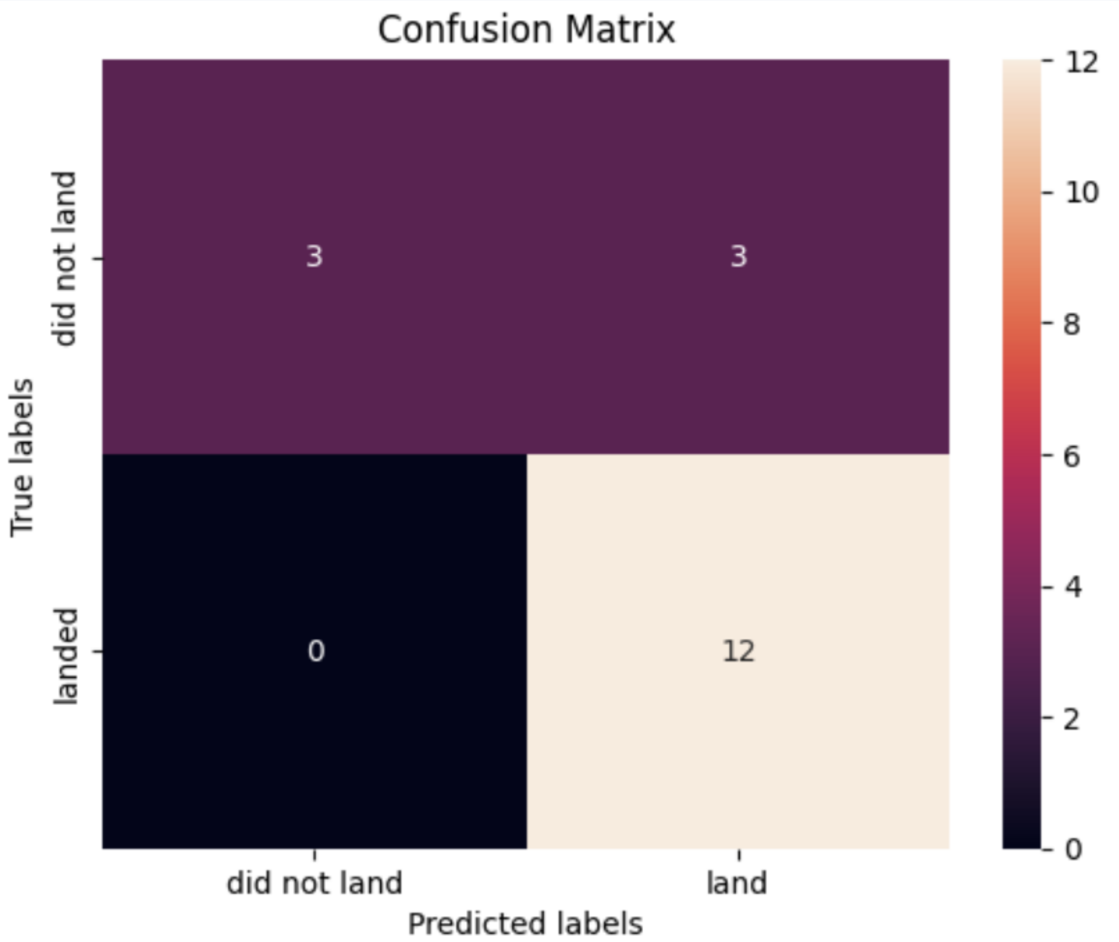
Classification Accuracy

- The decision tree achieved the highest classification accuracy score.



Confusion Matrix

- The confusion matrix shows how many samples the model did classify correctly and how many not.



Conclusions

- **Conclusion**:
  - The more flights from a launch sitethe higher the success rate at a launch site
  - Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
  - KSC LC-39A had the most successful launches of any sites.
  - The decision tree is the best performing classification algorithm for this task.

Thank you!