Winning Space Race with Data Science

Nozhin Azarpanah
August 2023

Outline

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

Executive Summary

Background and Objectives:

- Predicting the likelihood of the first stage rocket landing successfully, to reduce the cost of launches.
- Determining vital parameters and conditions to successfully recover first stage after use.

Methodology:

- Obtain launch data from SpaceX and Wikipedia
- Conduct exploratory data analysis
- Present the results

Results:

- Exploratory data analysis results
- Interactive analytics results
- Predictive analytics results

Introduction

- Project background and context:

  SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; While 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 successfully, 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 be answered:

  - What factors determine if the rocket will land successfully?
  - What is the interaction amongst various features that determine the success rate of a successful landing?
  - What operating conditions need to be in place to ensure a successful landing program?

Section 1

Methodology

Methodology

- Data collection
  - Used SPACEX REST API
  - Conducted web scraping on Wikipedia’s page about SPACEX launches
- Data wrangling
  - Converted HTML tables to a usable data frame
  - Replaced missing values
  - Conducted one-hot encoding to categorical feature
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Standardized data, split data into training/test data, and optimized models

Data Collection

- Data was collected using two different approaches:
  - The first approach was to use the SPACEX REST API and make a GET request for receiving the data.
  - The second approach was to web-scrape public websites for the data. For that purpose, we used Wikipedia's SPACEX website where data was presented in HTML tables.
- We performed requests on SPACEX API endpoints to get the data and saved them in a Pandas data frame.
- We web-scraped Falcon 9 launch records from Wikipedia using Python's BeautifulSoup package to extract them as an HTML table, parse the table and convert it to a Pandas data frame.

Data Collection - SpaceX API

- We got data from SPACEX API using GET request.
- We performed data cleaning and formatting and converted the data from JSON format to a data frame.

Notebook's GitHub link:

- https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/master/Data%20Collection%20using%20API.ipynb

Get launches' data from https://api.spacexdata.com/v4/launches/past

Convert JSON into Pandas data frame using json_normalize()

Clean the data and fill the missing values

Data Collection - Scraping

- We extracted Falcon 9 launch records in an HTML table from Wikipedia.
- We parsed the table and converted it into a Pandas data frame.

Notebook’s GitHub link:

- https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/master/Data%20Collection%20using%20Web%20Scraping.ipynb

Perform HTTP GET method to get the webpage

↓

Use ‘BeautifulSoup’ to parse and find tables

↓

Extract column names and create a dictionary with columns as keys and parse BeautifulSoup object webpage to fill the dictionary with table data

↓

Create data frame from dictionary

Data Wrangling

- In the data wrangling stage, we:
  - Found rows with missing values and replaced them.
  - Identified numerical and categorical columns and created dummy variables for the categorical ones.
  - Created a new column, named "Class" which represents the outcome of each launch. If the value is zero, the first stage did not land successfully, and one means the first stage landed successfully.

Notebook's GitHub link:

- https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- We created different charts to gain some insight about our data:
  - We created scatter plots to determine the relationship between different attributes of the data frame.
  - We created bar charts and line charts to determine the success rates.

Notebook’s GitHub link:

- https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/master/Exploratory%20Data%20Analysis%20with%20Visualization.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a SQLite3 database.
- We ran queries to provide more insight from the data.

Notebook's GitHub link:

- https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/master/Exploratory%20Data%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

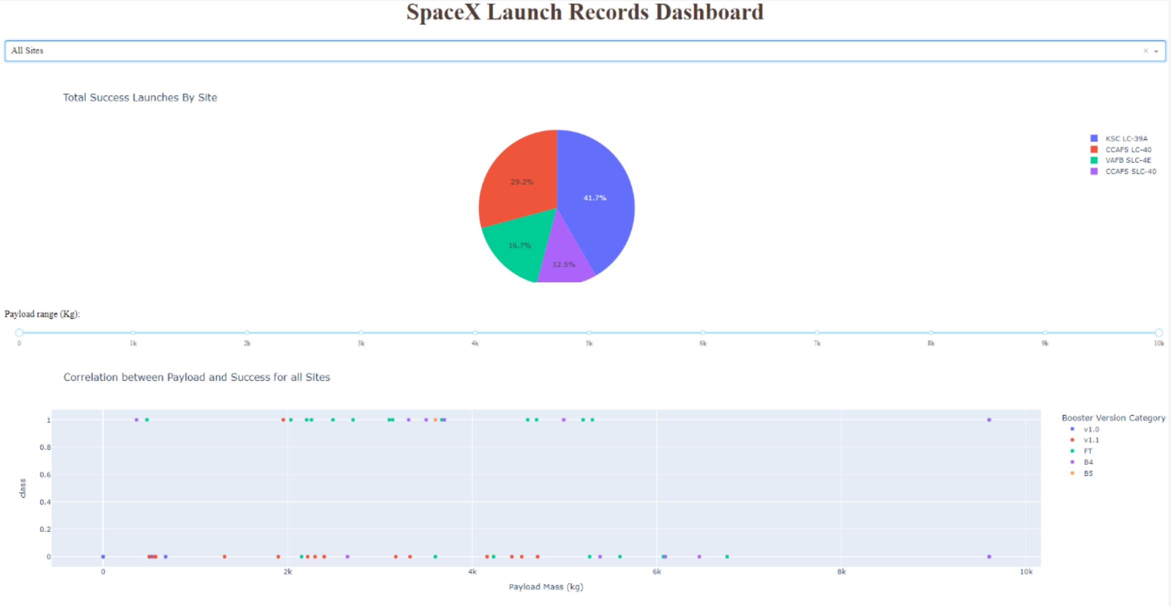
- We created a Folium map with:
    - Circles and markers to indicate launch site locations.
    - Markers to indicate all launches at respective locations (green=success, red=failure), to highlight success rate and number of launches per site.
    - Markers to indicate closest coastline, railway, highway, and city to each site. Lines were drawn to indicate distance and to determine how close these landmarks are to each site.

Notebook’s GitHub link:

- https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with that includes:
  - Pie chart of total success launches by launch site, as well as success/failure rate of each site (by selecting specific site)
  - Scatter plot of launch success by payload mass, booster version, and specified site (if applicable) - controlled by payload mass slider and site selection



Notebook's GitHub link:

- https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/master/Interactive%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

- We separated data into dependent (landing outcome) and independent (all other parameters) variables.
- We transformed the data into standardized form.
- We split the dataset into train (80% of data) and test (20% of data) sets.
- We created different machine learning models including KNN, SVM, Classification Tree and Logistic Regression.
- We used GridSearchCV to find the best parameters for each model.
- We analyzed the accuracy with our test set based on score and confusion matrix, and the best performing model was found to be Decision Tree.

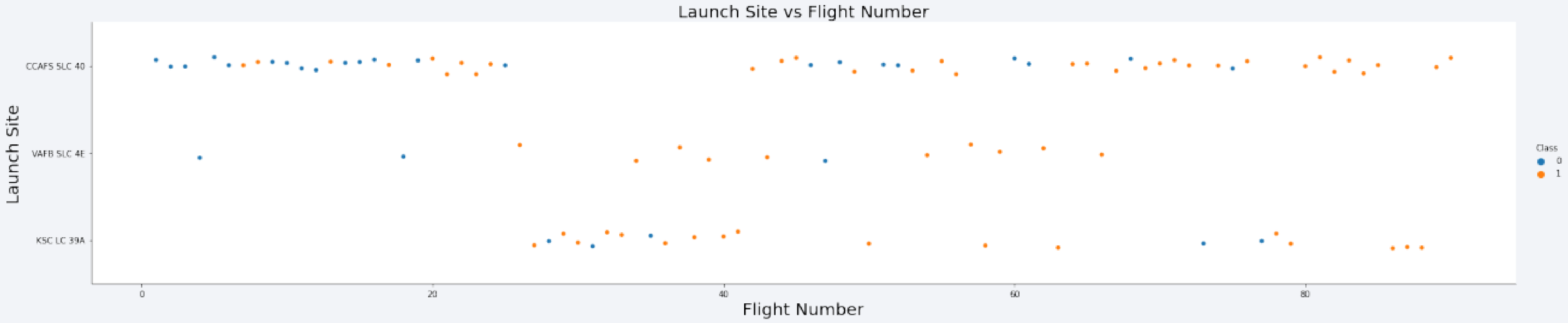
Notebook's GitHub link:

- https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/master/Machine%20Learning%20Prediction.ipynb

Section 2

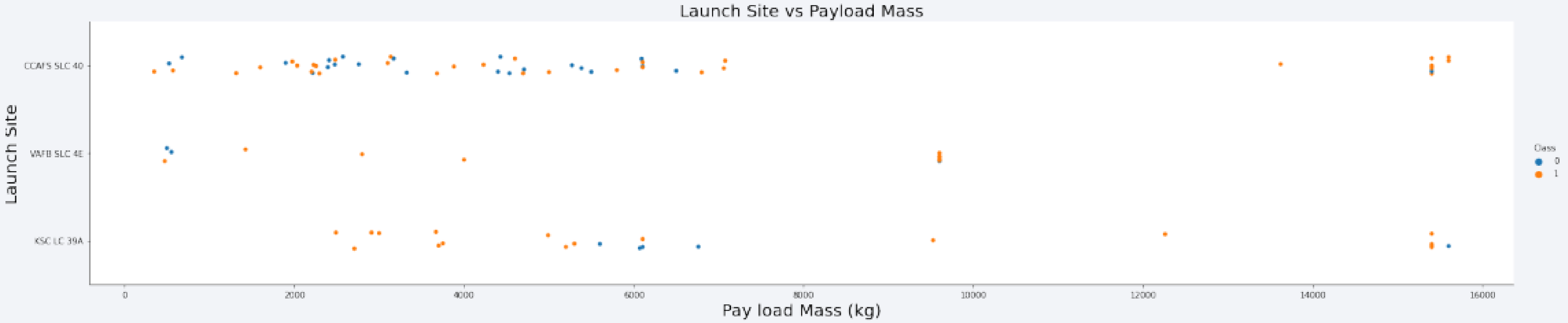
Insights drawn from EDA

Flight Number vs. Launch Site



- As the flight number increases, success rate increases.
- In KSC LC 39A launch site, the flights began later compared to other launch sites.
- The launch site with the highest number of launches is CCAFS 5LC 40.

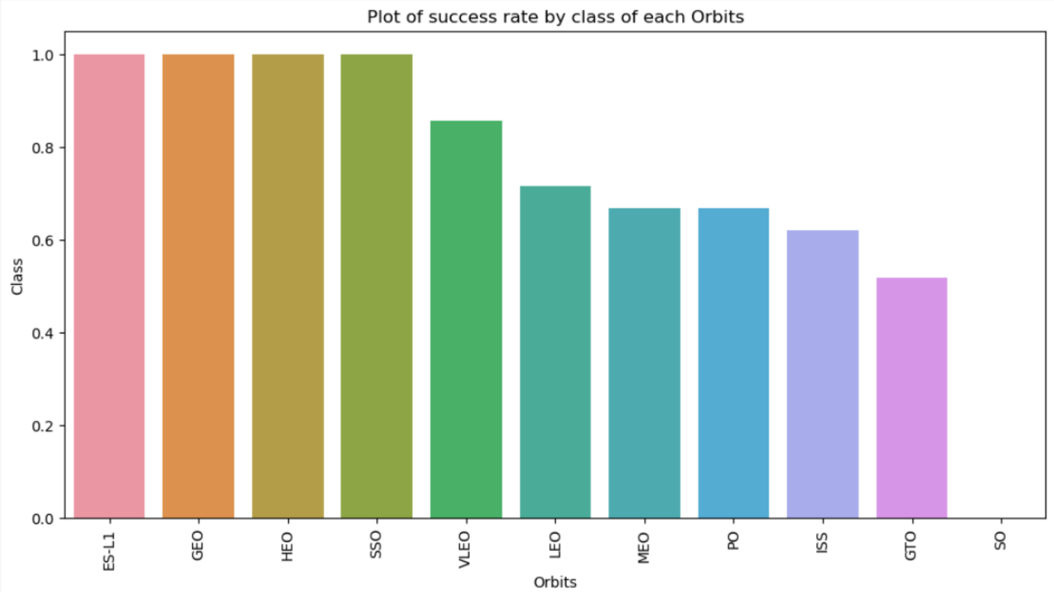
Payload vs. Launch Site



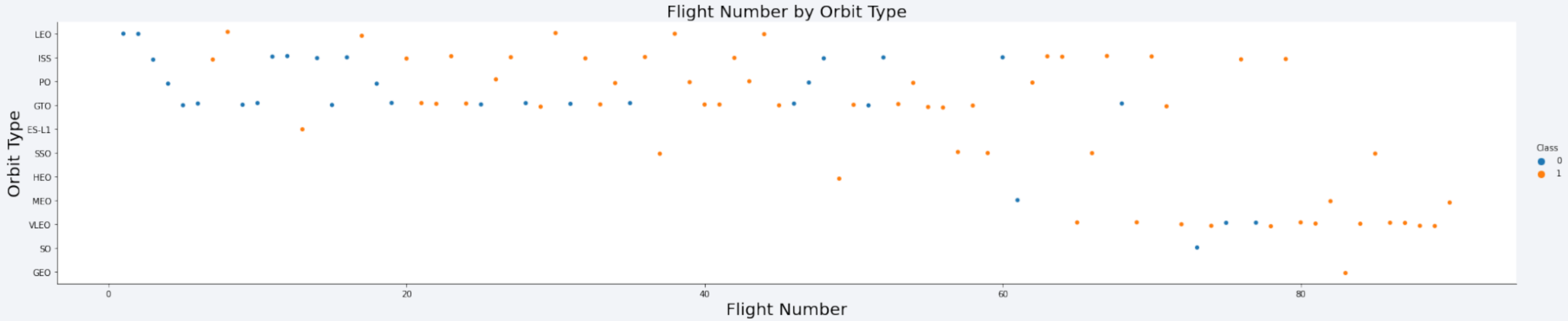
- In VAFB SLC 4E launch site, there are no rockets launched with heavy payload mass (>10000 kg).
- Most rockets have launched from CCAFS SLC 40 launch site.
- Rockets launched with higher payload mass (>10000 kg) generally yielded higher success rates.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO yielded 100% success rate.
- GTO yielded 51.8% success rate (2nd lowest).
- SO yielded 0% success rate (lowest).

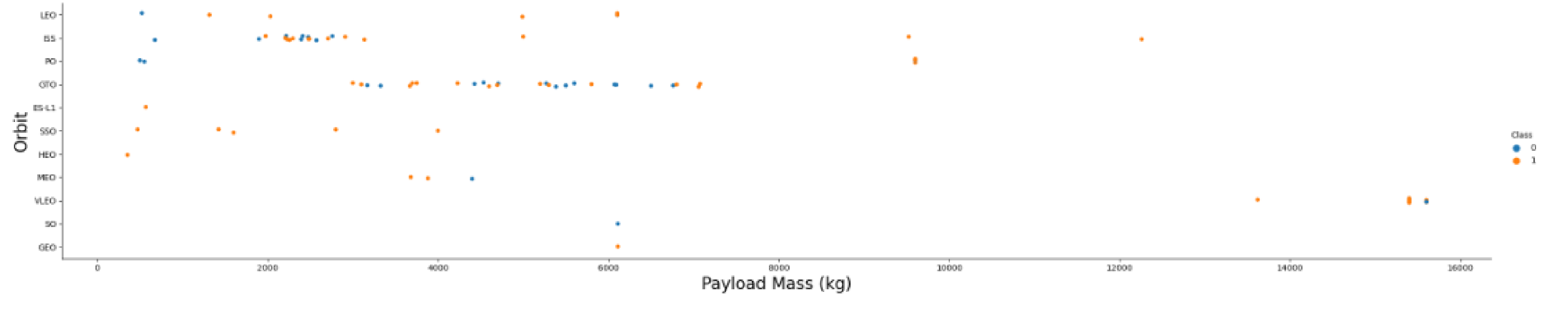


Flight Number vs. Orbit Type



- LEO occurs at earlier flights; HEO, MEO, VLEO, SO, and GEO occur at later flights (higher flight numbers).

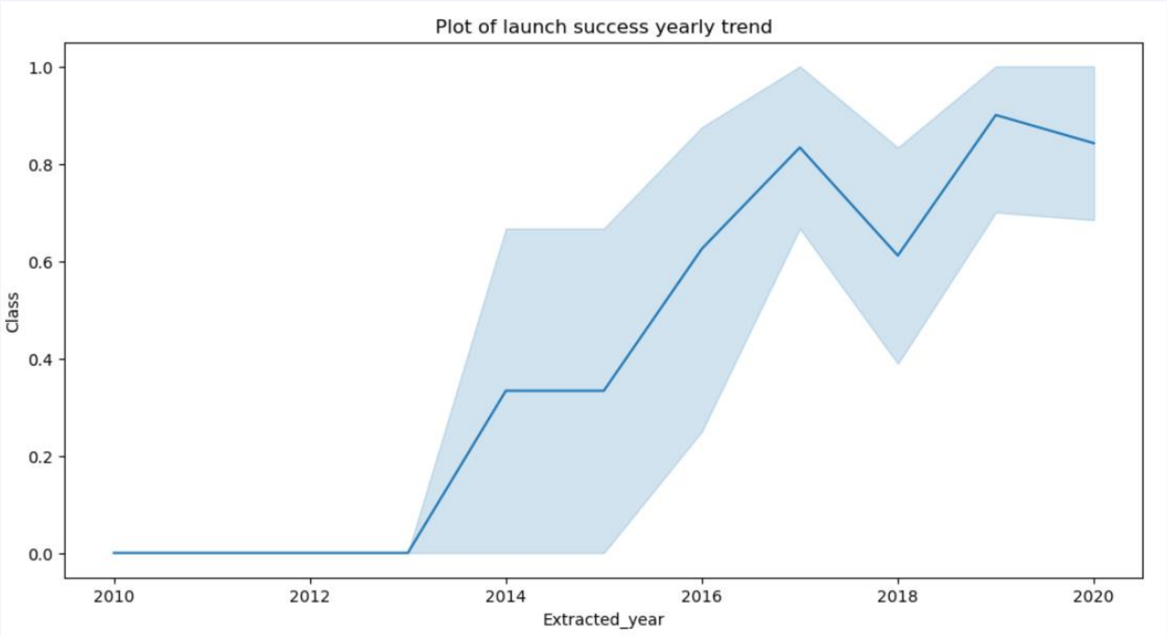
Payload vs. Orbit Type



- For LEO, ISS, PO, heavier payload mass tend to yield more successful launches.

Launch Success Yearly Trend

- Success rate has increased over time, with the exceptions of 2018 and 2020.
- The peak of success rate was in 2019.



All Launch Site Names

The query command to find all launch site names from the database:

```
SELECT DISTINCT Launch_Site
FROM SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

The query command to show the first 5 records where launch sites begin with 'CCA':

```
SELECT *
FROM SPACEXTBL
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	L
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

The total payload mass was found from the query:

```
SELECT SUM(Payload_Mass__KG_) AS TotalPayloadMass
FROM SPACEXTBL
WHERE Customer LIKE 'NASA (CRS)'
```

TotalPayloadMass
45596.0

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 was calculated from the query:

```
SELECT AVG(Payload_Mass__KG_) AS Avg_PayloadMass
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1'
```

Avg_PayloadMass
2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad was found with the query:

```
SELECT MIN(Date) AS FirstSuccessfullLandingDate
FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Success (ground pad)'
```

FirstSuccessfullLandingDate
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

The query command for listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 6000 is:

```
SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
    AND Payload_Mass__KG_ > 4000
    AND Payload_Mass__KG_ < 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 query command for the total number of successful and failure mission outcomes:

```
SELECT
   COUNT(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 END) AS
SuccessOutcome,
   COUNT(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 END) AS
FailureOutcome
FROM SPACEXTBL;
```

SuccessOutcome	FailureOutcome
100	1

Boosters Carried Maximum Payload

Listing the names of the boosters that have carried the maximum payload mass can be done using the query:

```
SELECT Booster_Version
        FROM SPACEXTBL
        WHERE Payload_Mass__KG_ = (
                        SELECT MAX(Payload_Mass__KG_)
                        FROM SPACEXTBL
                        )
        ORDER BY Booster_Version
```

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

2015 Launch Records

The records of 2015 with their months, failure landing outcomes in drone ship, booster versions and launch sites was found using the query:

```
SELECT Booster_Version, Launch_Site, Landing_Outcome, substr(Date, 4, 2) AS Month
FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Failure (drone ship)'
AND substr(Date,7,4)='2015'
```

Booster_Version	Launch_Site	Landing_Outcome	Month
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	10
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	04

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

Lastly, we found the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order using the query:

```
SELECT Landing_Outcome, COUNT(Landing_Outcome)
FROM SPACEXTBL
WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017'
AND Landing_Outcome LIKE 'Success%'
GROUP BY Landing_Outcome
ORDER BY COUNT(Landing_Outcome) DESC
```

Landing_Outcome	COUNT(Landing_Outcome)
Success	20
Success (drone ship)	8
Success (ground pad)	7

Section 3

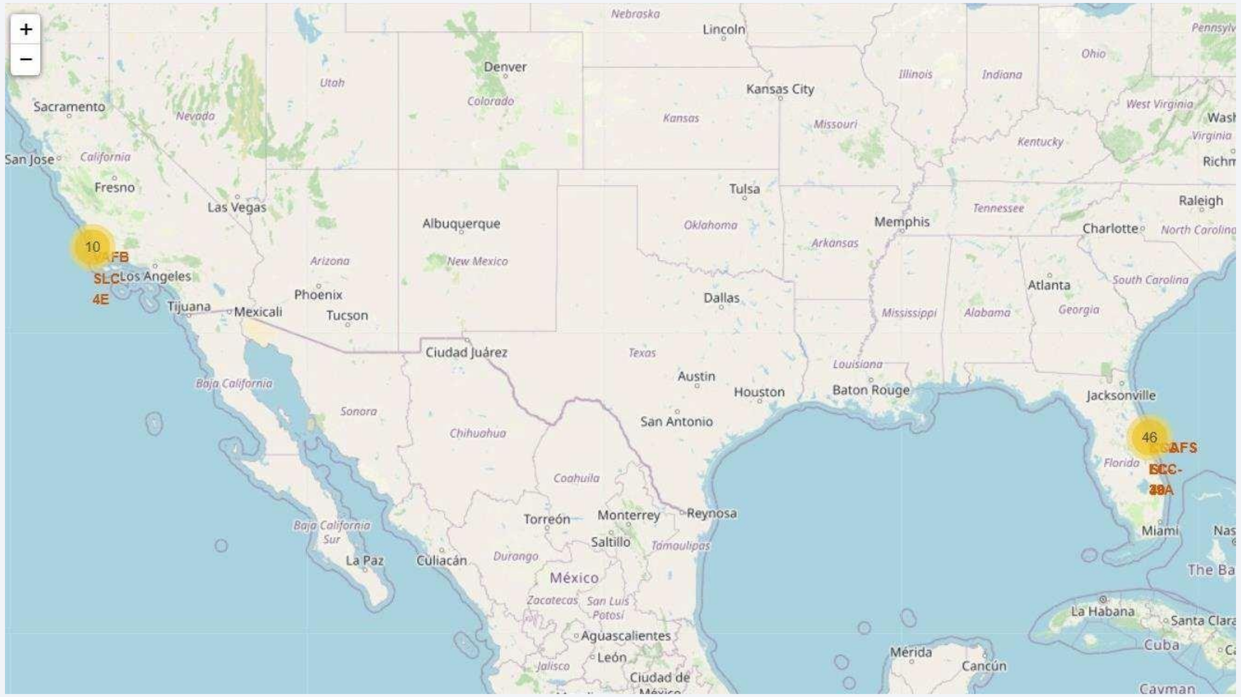
Launch Sites Proximities Analysis

Launch Site on the map using Folium

- This screenshot depicts all launch sites on a Folium map.
- All the launch sites are in the coasts of the United States of America.

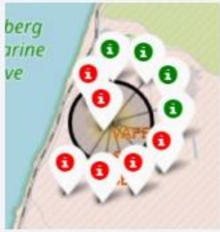
Number of launches for each site

- This screenshot depicts the number of launches for each site.
- Most launches took place in Cape Canaveral sites.



Showing the result of launches with Markers

This screenshot depicts the results of launches in each site with green and red markers.


VAFB SLC-4E


KSC LC-39A


CCAFS LC-40

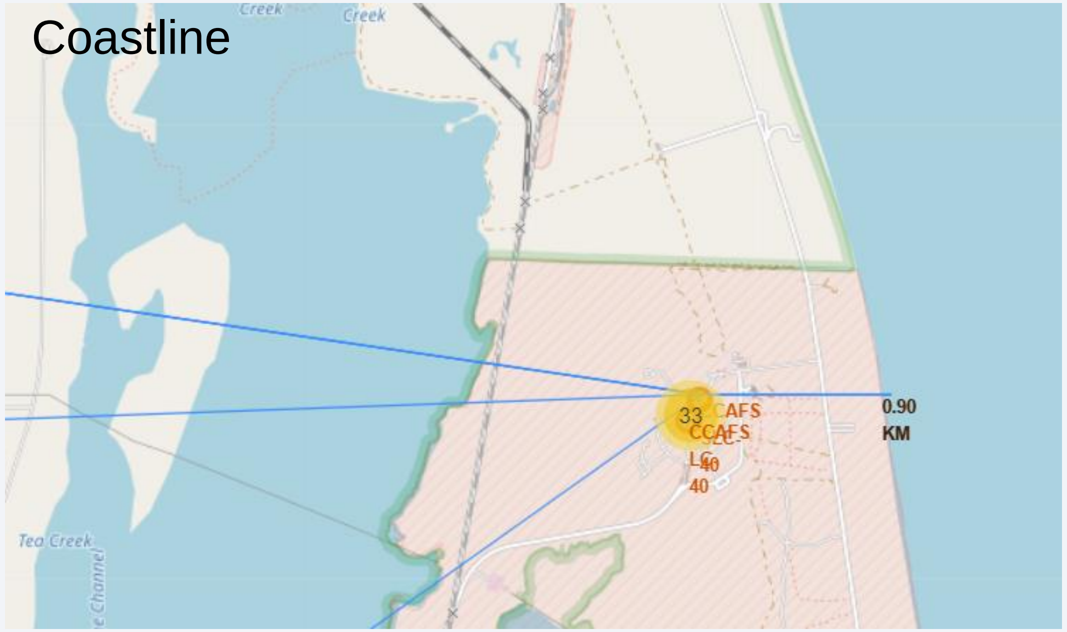

CCAFS SLC-40

- CCAFS yielded most flights (LC-40 and SLC-40).
- KSC, despite being close to CCAFS, yielded the highest success rate.

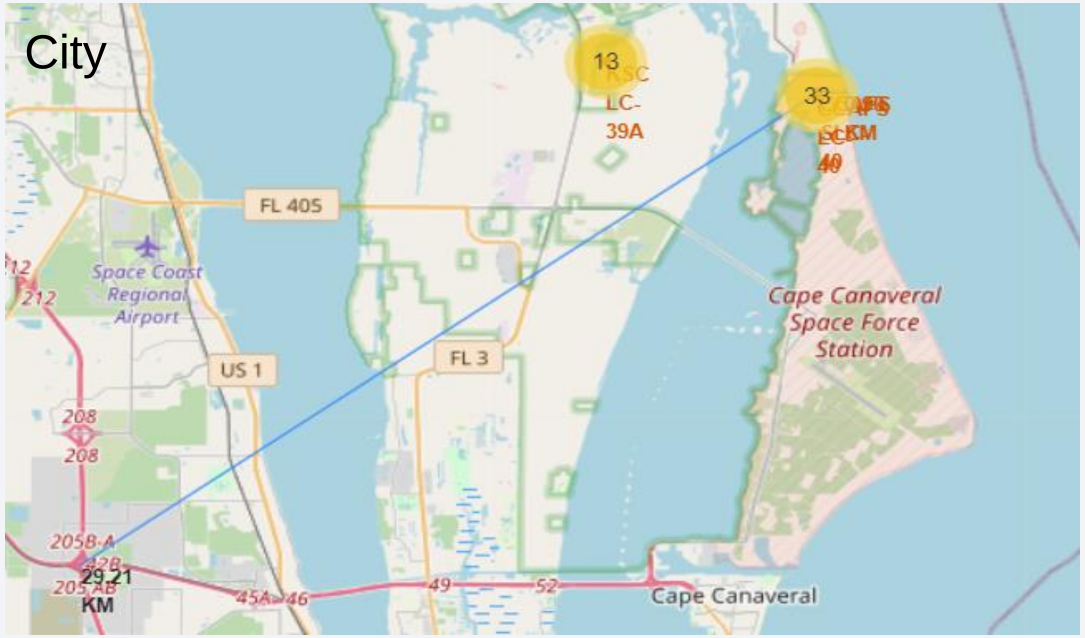
Launch Sites Distance to Landmarks

We calculated the distance from launch sites to the closest coastline, closest highway, closest city, and closest railway station.

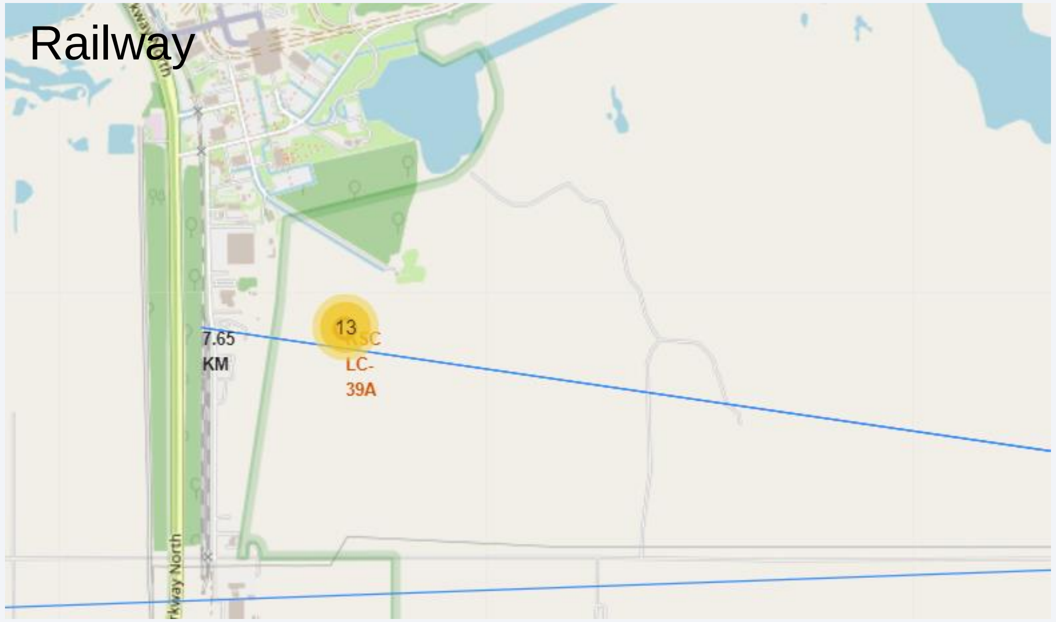
Coastline



City



Launch Sites Distance to Landmarks

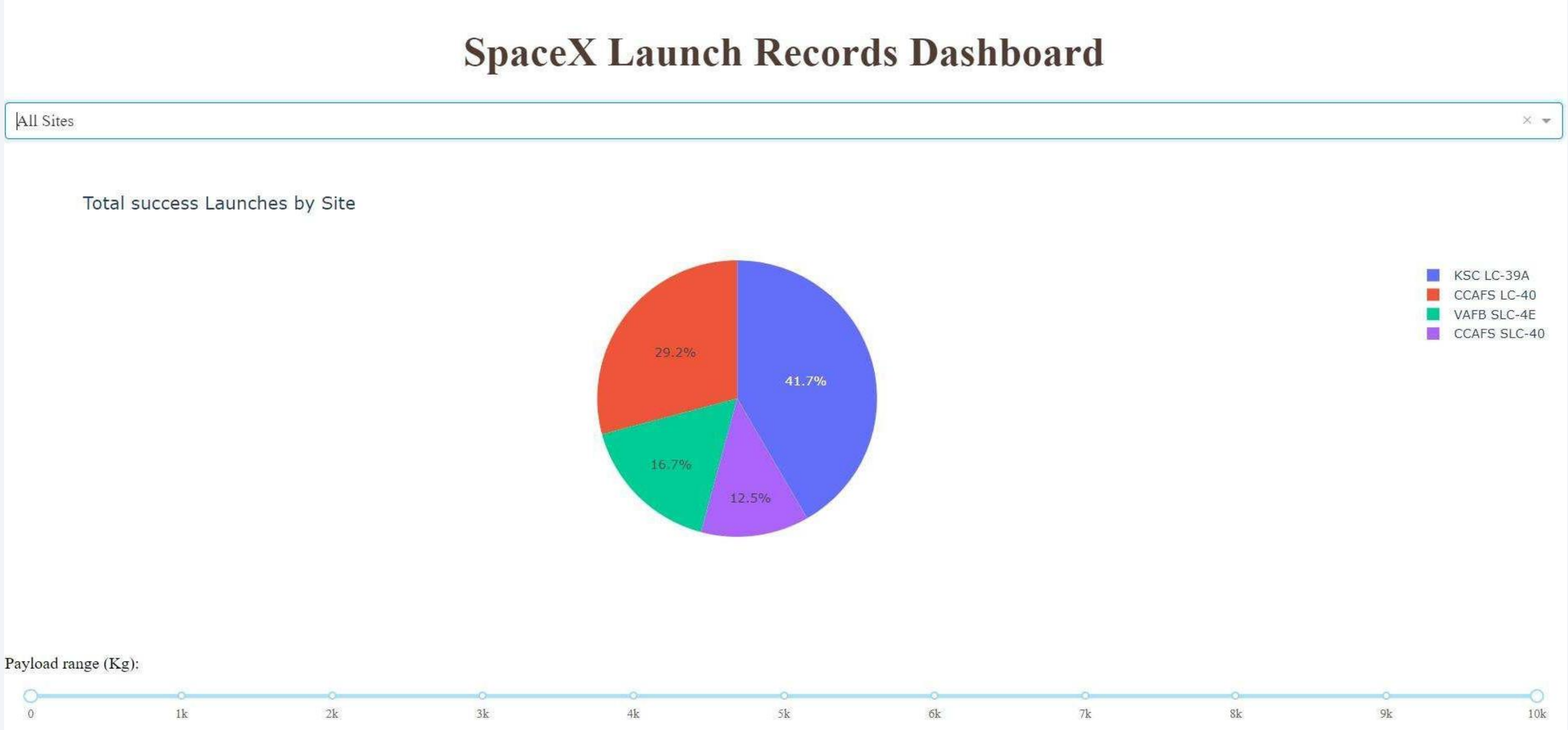


- Launch sites' distance to coastlines is short.
- Launch sites have more distance from cities.

Section 4

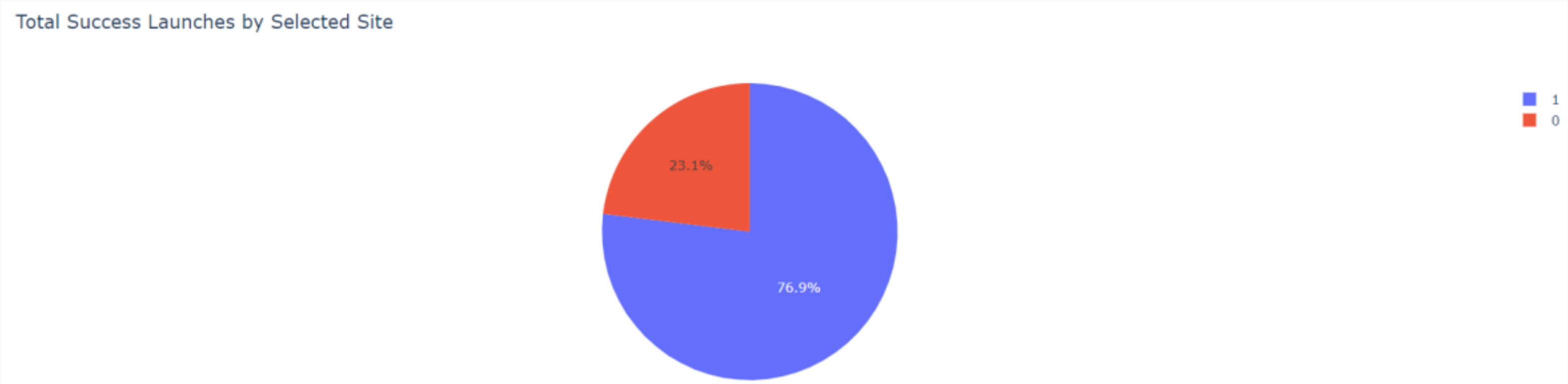
Build a Dashboard with Plotly Dash

Distribution of Successful Launches by Site



- KSC launch site yielded most successful launches.

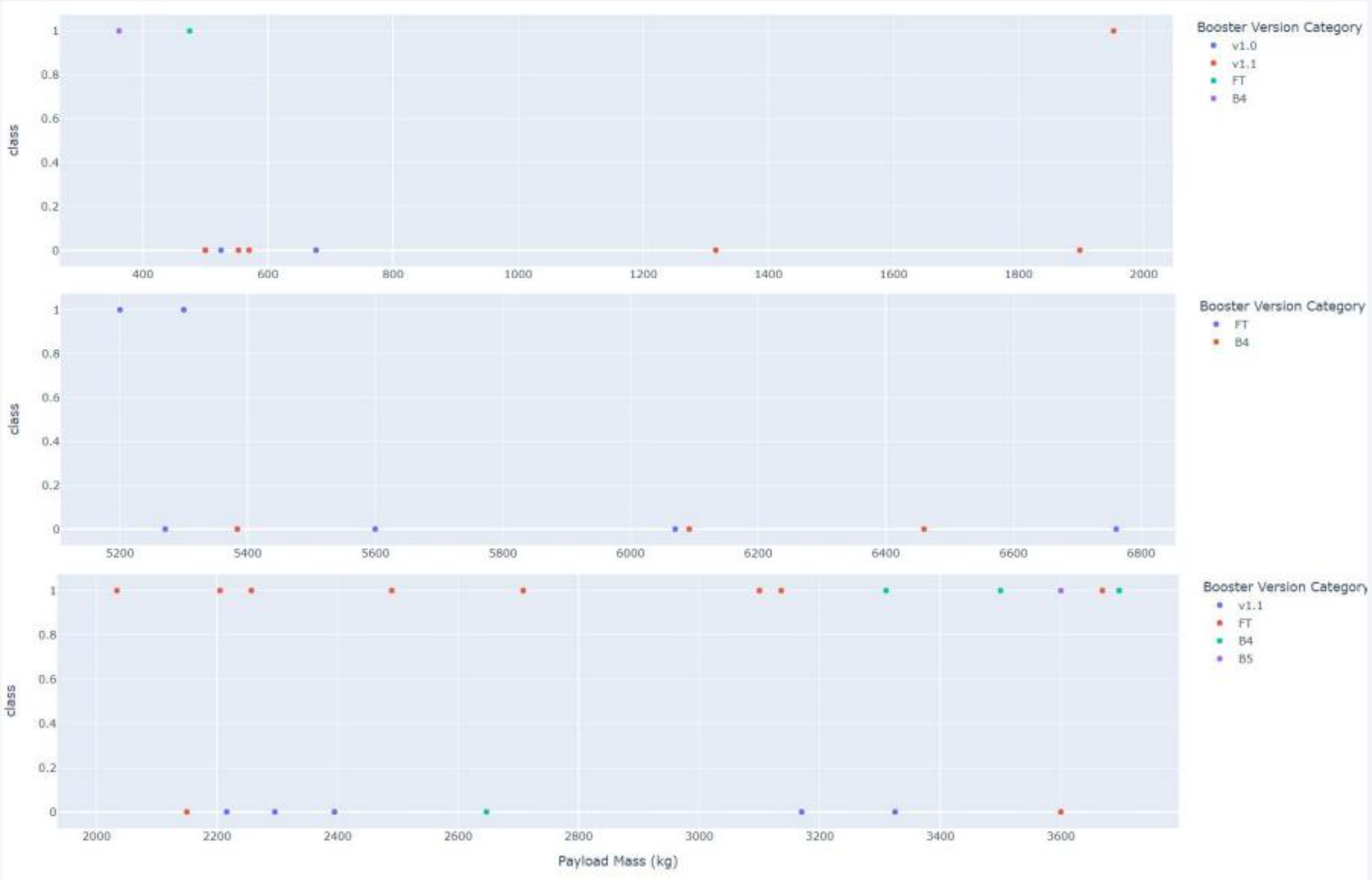
Success/Failure Rate at each Launch Site



- 76.9% of launches at KSC were successful, highest among all sites.
- All other sites yield a successful rate below 50%.

Scatter plot for Payload vs Launch Outcome

- Most successful launches had payload mass of 2000 - 4000 kg.
- V1.1 is least successful booster.
- FT is most successful booster.

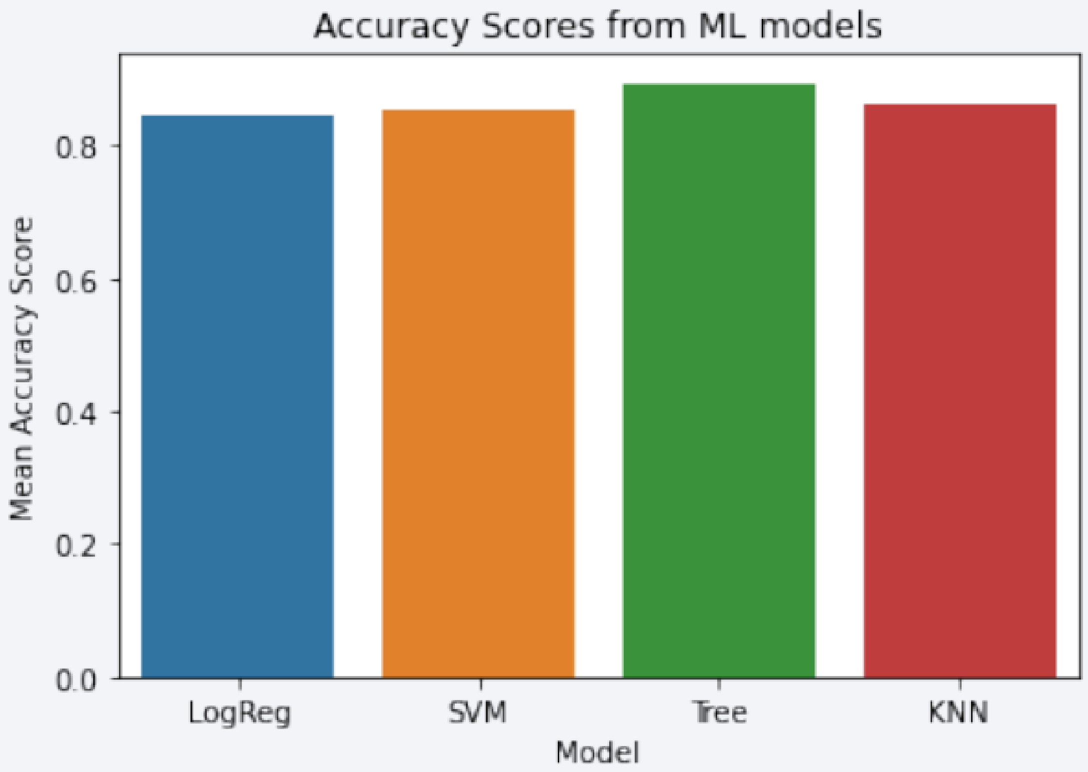


Section 5

Predictive Analysis (Classification)

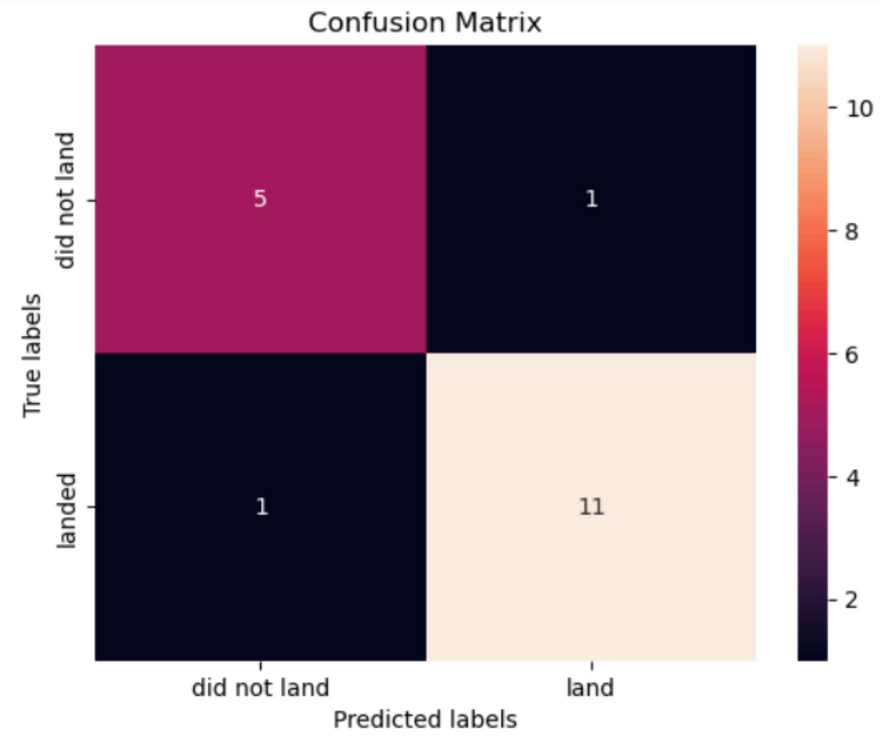
Classification Accuracy

Decision Tree yielded the highest average model accuracy among 4 machine learning algorithms.



Confusion Matrix of Decision Tree Classifier

There is only one false positive and one false negative in the confusion matrix of decision tree classifier.



Conclusions

- There is a positive correlation between number of flights and success rate as the success rate has improved over the years.
- Orbits ES-L1, GEO, HEO and SSO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- Heavier payload mass generally yielded greater success.
    - Different orbit types have specific ranges for success (low payload mass for LEO, SSO, MEO and high payload mass for VLEO)
- F9 FT boosters are best for mid-range payload masses and F9 B5 boosters are best for high-range payload masses.
    - Avoid earlier booster versions (F9 v1.0, F9 v1.1)
- Launch sites should be close to coastlines and far away from cities.
- Decision Tree model yields best classification, despite slight variations between iterations.

Appendix

- SPACEX API: https://github.com/NojinAp/SpaceX-Falcon-9-first-stage-Landing-Prediction/tree/master
- GitHub URL with all Notebooks: https://github.com/SakisHous/spacex-data-science

Thank you!