Data Science Capstone Project: Space X

Olha Dmytriieva
October 13, 2024

Outline

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

Executive Summary

Summary of methodologies

- Data collection (API, Webscraping)
- Data wrangling
- EDA with SQL
- EDA with data visualization
- Interactive maps with Folium
- Creating a Dashboard with Plotly
- Predictive analysis, classification

Summary of all results

- Visual analysis of data
- Result of a predictive model
- Selecting the best model for a predictive analysis

Introduction

Project background and context

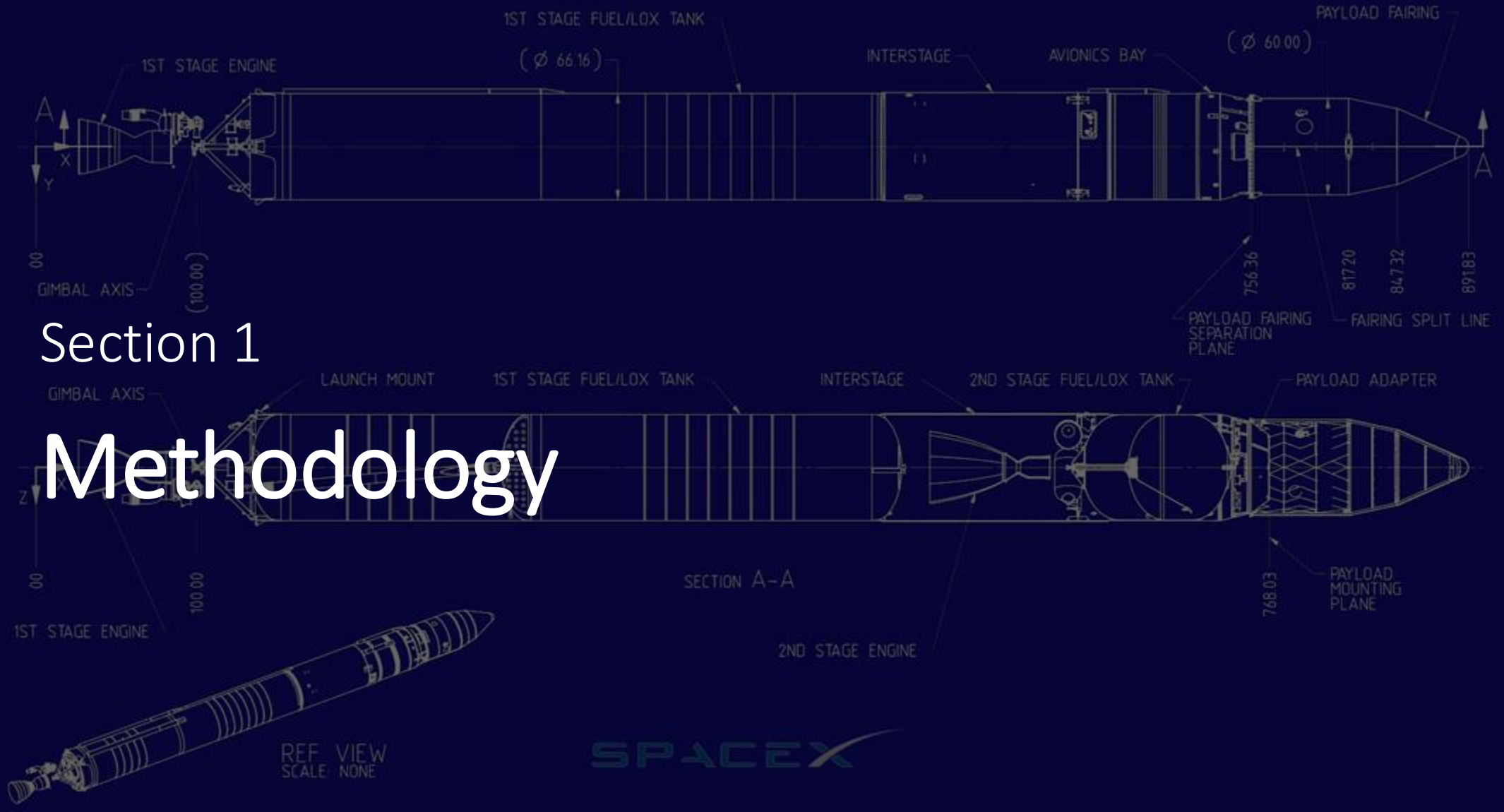
SpaceX offers the opportunity to launch rockets for $62 million, while launches from other companies exceed this amount several times. The main reason for such savings is the reuse of the first stage. At the same time, the successful landing of the first stage does not happen every time, and for the correct launch cost, it is necessary to take into account the percentage of successful and unsuccessful landings.

Problems you want to find answers

- What percentage of first stage landings are successful?
- What affects a successful landing?
- Does the success of the first stage landing depend on the landing site?

Section 1

Methodology



Methodology

- Data collection methodology:
  - SpaceX Rest API and web scraping from Wikipedia
- Perform data wrangling:
  - One hot encoding data fields for Machine Learning and dropping irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL:
  - Scatter plots and bar charts to visualize the data and extract meaningful patterns to guide the modeling process
  - Creating and execute SQL queries to select and sort data

Methodology

- Perform interactive visual analytics:
  - Creating interactive maps to visualize launch sites with Folium
  - Creating a dashboard with Plotly
- Perform predictive analysis using classification models:
  - Model creation and evaluation

Data collection

The data collection was done in two stages: launch data from the API (information about rockets, payload, launch pad, flight number, date, etc.) and using Web scraping (it contains additional information, landing results of stages, etc.).

You can see the notebooks at the links:
Collection API
Web Scraping

```
spacex_url="https://api.spacexdata.com/v4/launches/past
```

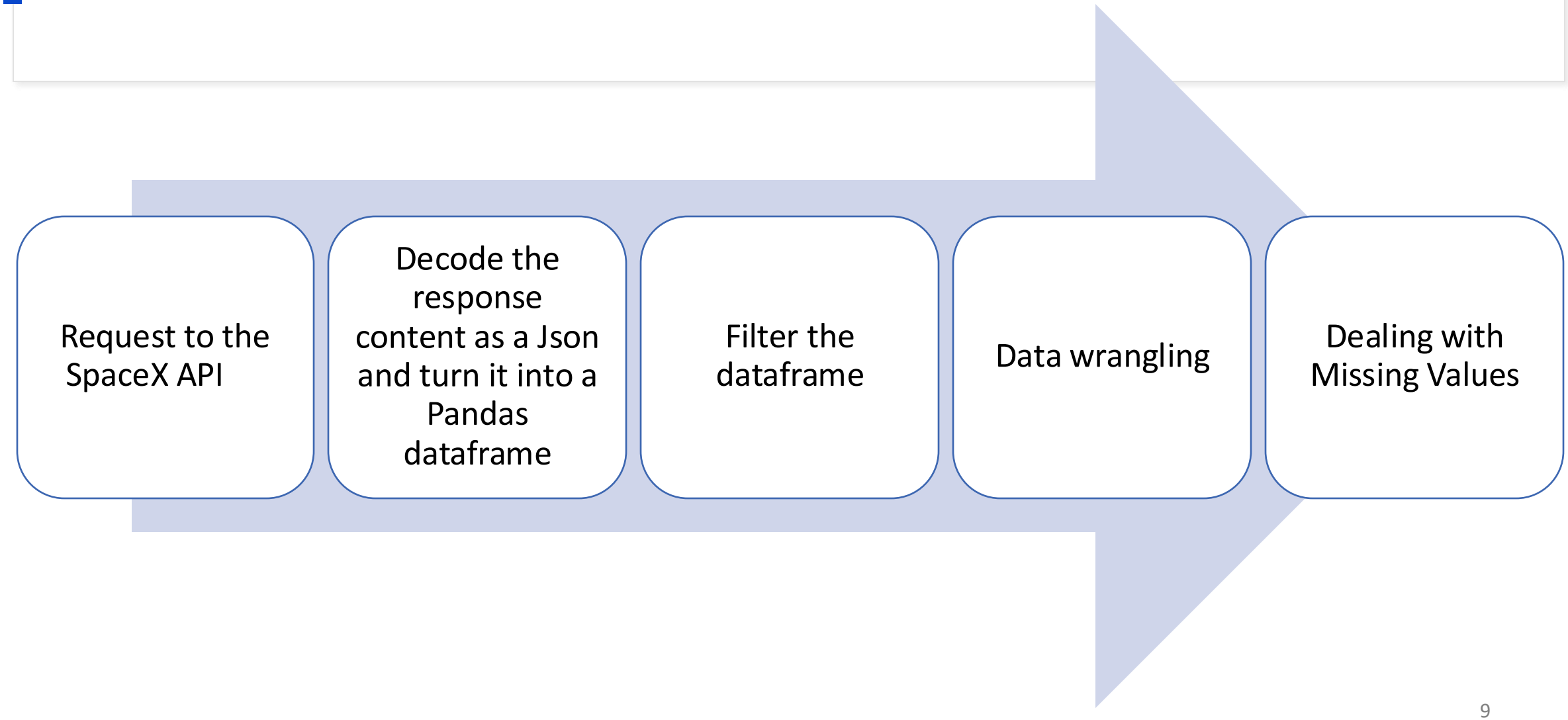
```
response = requests.get(spacex_url)
```

```
static_url = "https://en.wikipedia.org/w/index.php?title
static_url
```

```
response = requests.get(static_url).text
```

```
soup = BeautifulSoup(response,"html.parser")
```

Data collection: SpaceX API



Data collection: Web scraping

Request data from Wikipedia web page

Extract all variable names from the HTML table header

Create a data frame by parsing an HTML table

Data Wrangling

The next step was exploratory data analysis to find patterns in the data, determine which variables would be used for machine learning models. It was necessary to analyze various cases of successful landings and those cases where the landing was unsuccessful.

The main questions at this stage were:

- In which regions the percentage of successful landings is higher
- What type of orbit is the launch aimed at
- Number and frequency of missions when the rocket reached a certain orbit

A notebook on this topic can be found at the link

Data Wrangling

```
df['LaunchSite'].value_counts()
```

```
LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: count, dtype: int64
```

Calculate the number of launches on each site

```
df['Orbit'].value_counts()
```

```
Orbit
GTO       27
ISS       21
VLEO      14
PO         9
LEO        7
SSO        5
MEO        3
HEO        1
ES-L1      1
SO         1
GEO        1
Name: count, dtype: int64
```

Calculate the number and occurrence of each orbit

```
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
Outcome
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: count, dtype: int64
```

Calculate the number and occurence of mission outcome of the orbits

```
# landing_class = 0 if bad_outcome
landing_class = []
for key, value in df['Outcome'].items():
    if value in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
# landing_class = 1 otherwise
```

```
df['Class']=landing_class
df[['Class']].head(5)
```

Create a landing outcome label from Outcome column

EDA with Data Visualization

During the research analysis, visualizations were also created to better understand the relationship of variables. Below you can see graphs that allow you to evaluate the relationship of the following variables: payload mass, launch site, flight number, success rate, etc.

You can see the data analysis using visualization at the link

EDA with Data Visualization

Scatter plots were created for the relationship between the following variables:

- The relationship between Flight Number and Launch Site
- The relationship between Payload Mass and Launch Site
- The relationship between Flight Number and Orbit type
- The relationship between Payload Mass and Orbit type

Bar charts for the following:

- The relationship between success rate of each orbit type

Line chart to visualize success rate over time

EDA with SQL

During EDA, SQL was used to understand the data set. Using SQL queries, you can filter the data, get acquainted with the desired variables, etc.
Here is the code to connect to the database.

You can view the notebook at the link

```
!pip install sqlalchemy==1.4
```

```
!pip install ipython-sql
```

```
%load_ext sql
```

```
import csv, sqlite3

con = sqlite3.connect("my_data1.db")
cur = con.cursor()
```

```
!pip install -q pandas
```

```
%sql sqlite:///my_data1.db
```

```
import pandas as pd
df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.
df.to_sql("SPACEXTBL", con, if_exists='replace', index=False,method="
```

EDA with SQL

The following queries were used:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first succesful landing outcome in ground pad was acheived.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass
- Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Build an Interactive Map with Folium

Interactive maps allow you to track the location of launches. Finding the best place to build launches requires taking into account many factors, such as proximity to the coastline, proximity to the equator, etc. In this project, visual analytics was performed using Folium.

To work with Folium, you need to download the following packages (image on the right)

You can see the full notebook at the link

```
import piplite
await piplite.install(['folium'])
await piplite.install(['pandas'])
```

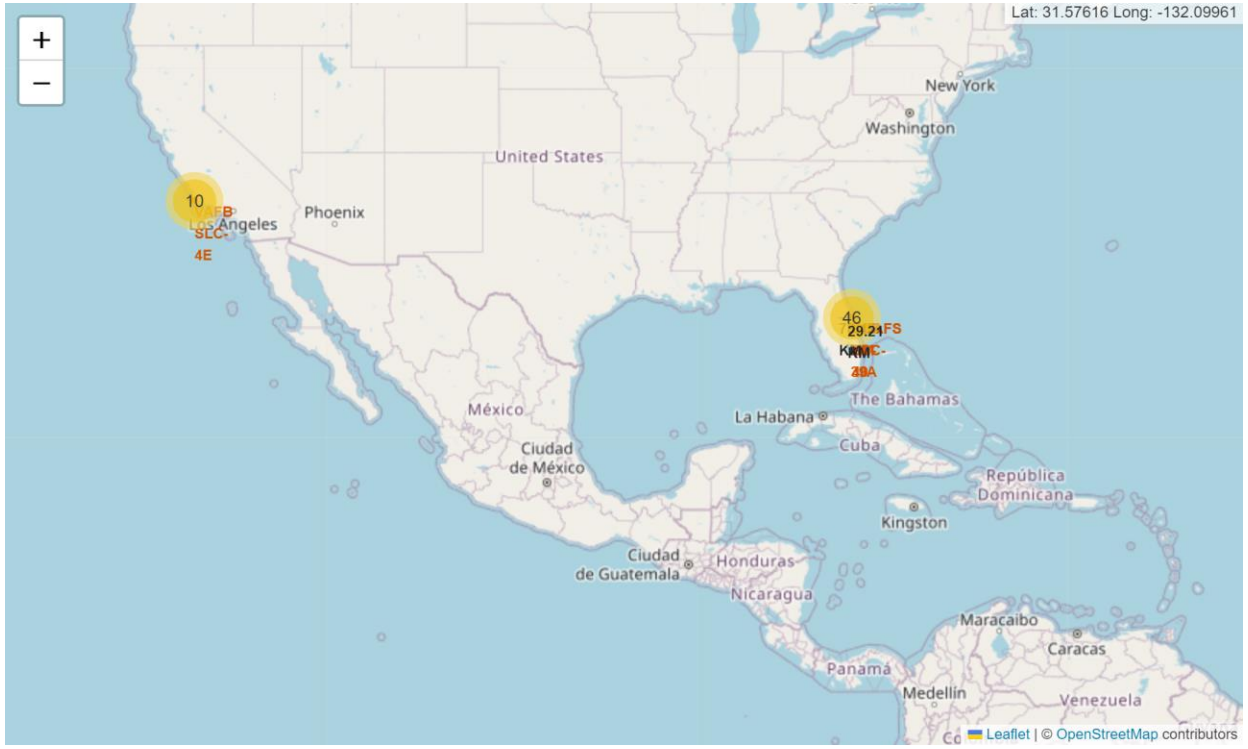
```
import folium
import pandas as pd
```

```
# Import folium MarkerCluster plugin
from folium.plugins import MarkerCluster
# Import folium MousePosition plugin
from folium.plugins import MousePosition
# Import folium DivIcon plugin
from folium.features import DivIcon
```

Build an Interactive Map with Folium

During the analysis of data using Folium, the following was done:

- Marking all launch sites on a map
- Marking the success/failed launches for each site on the map
- Calculating the distances between a launch site to its proximities



Build a Dashboard with Plotly Dash

Plotly Dash allows you to create an interactive application for users to perform visual analytics of startup data in real time.
The following tasks were completed during the creation of the application:

- Added a Launch Site Drop-down Input Component
- Added a callback function to render success-pie-chart based on selected site dropdown
- Added a Range Slider to Select Payload
- Added a callback function to render the success-payload-scatter-chart scatter plot

The application code can be found at the link

Predictive Analysis (Classification)

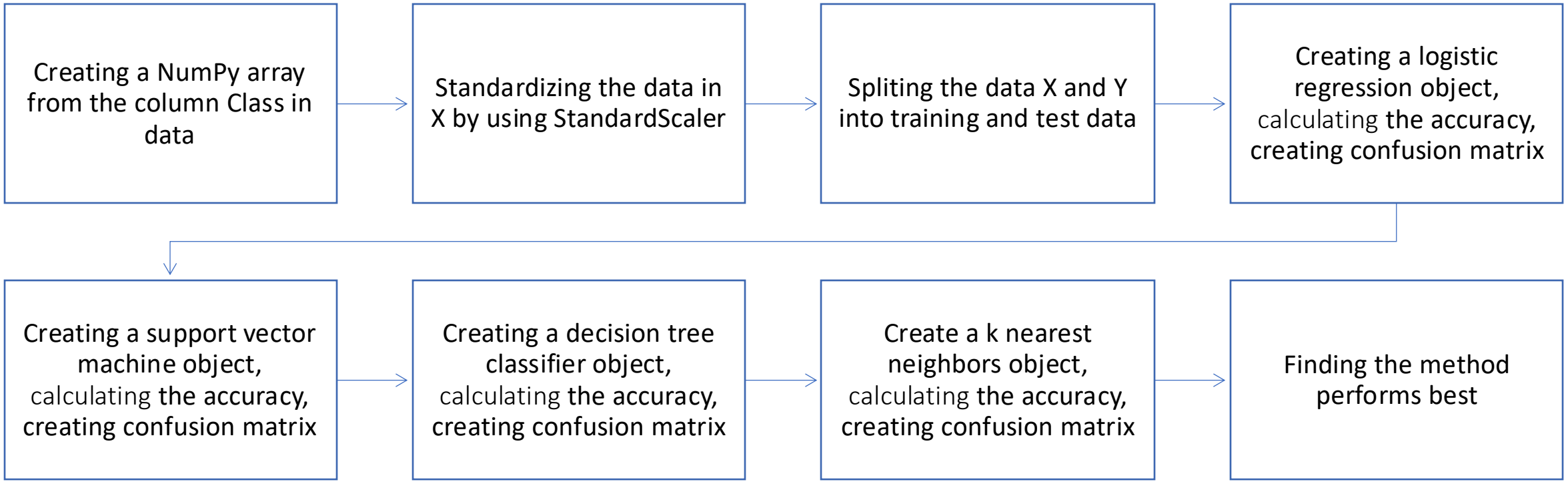
Machine learning was used to determine whether the Falcon 9 first stage landed successfully.

During the work on the machine learning models, the following was done:

- Data was divided into training and testing
- Various classification models were trained to identify the optimal model
- Optimized the Hyperparameter grid search
- Build a predictive model
- The application code can be found at the link

You can see the notebook at the link

Predictive Analysis (Classification)

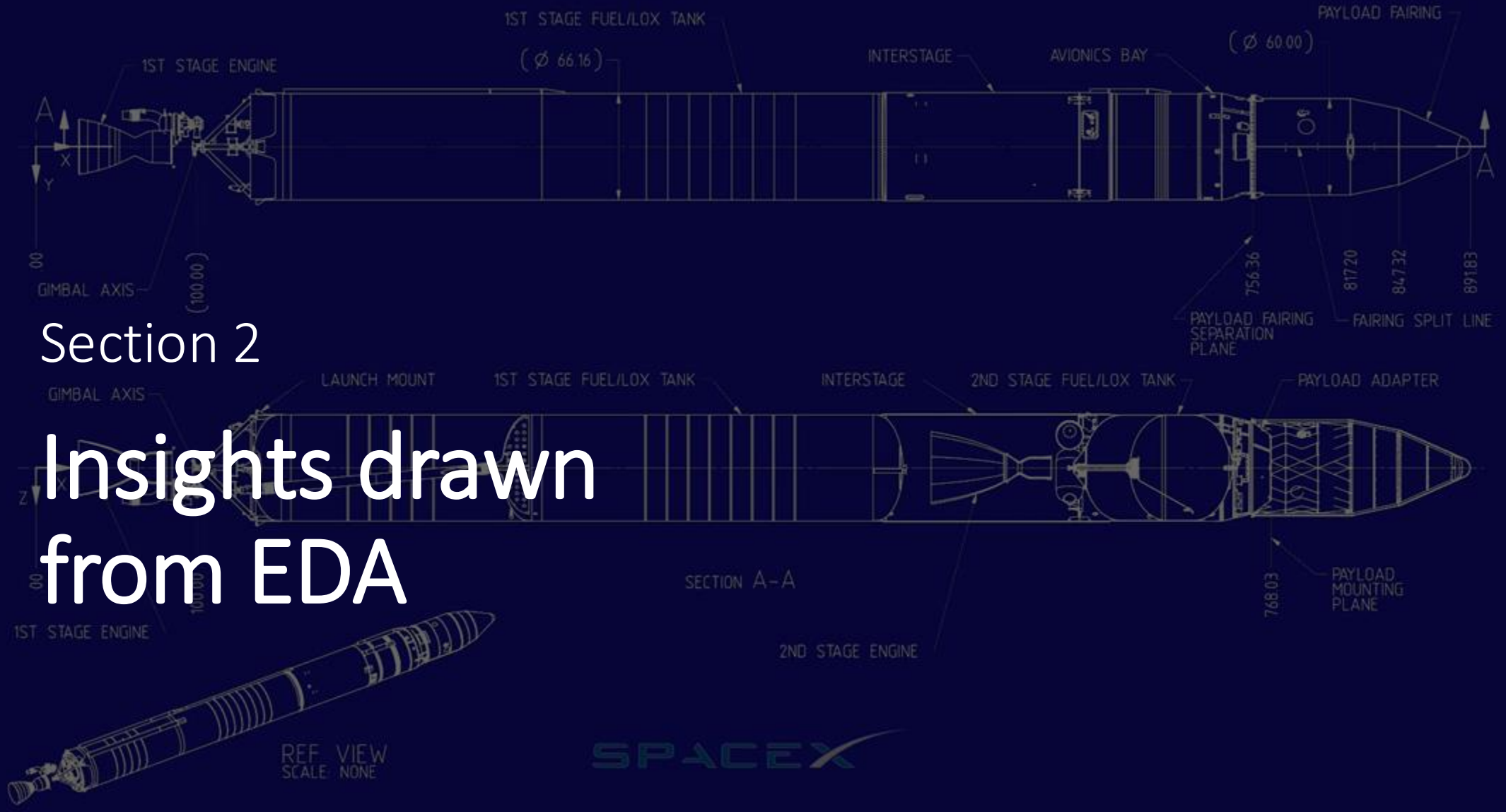


Results

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

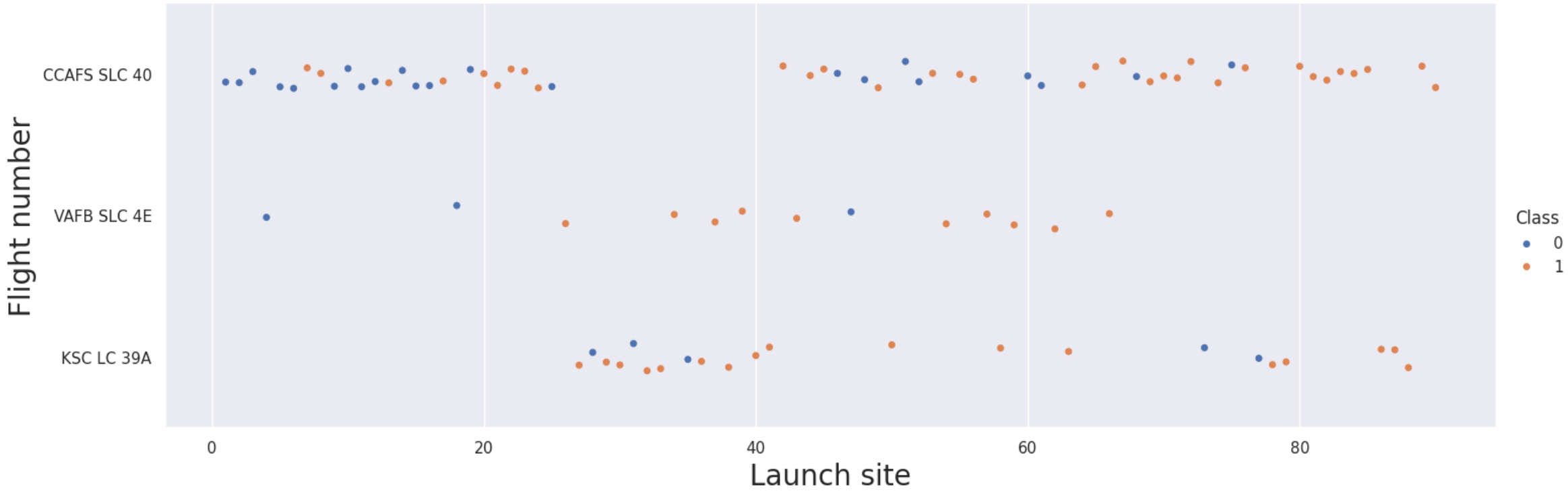
Section 2

Insights drawn from EDA



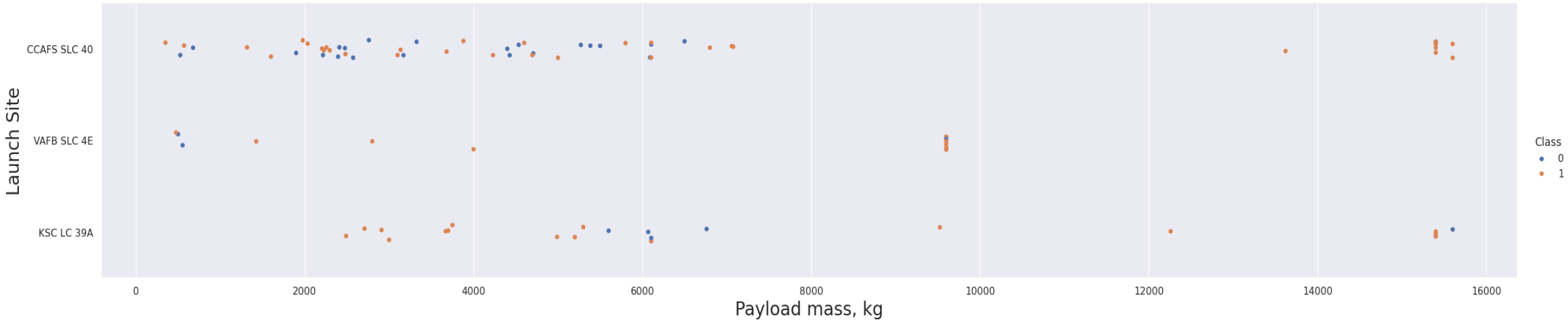
EDA with Data Visualization

The relationship between Flight Number and Launch Site



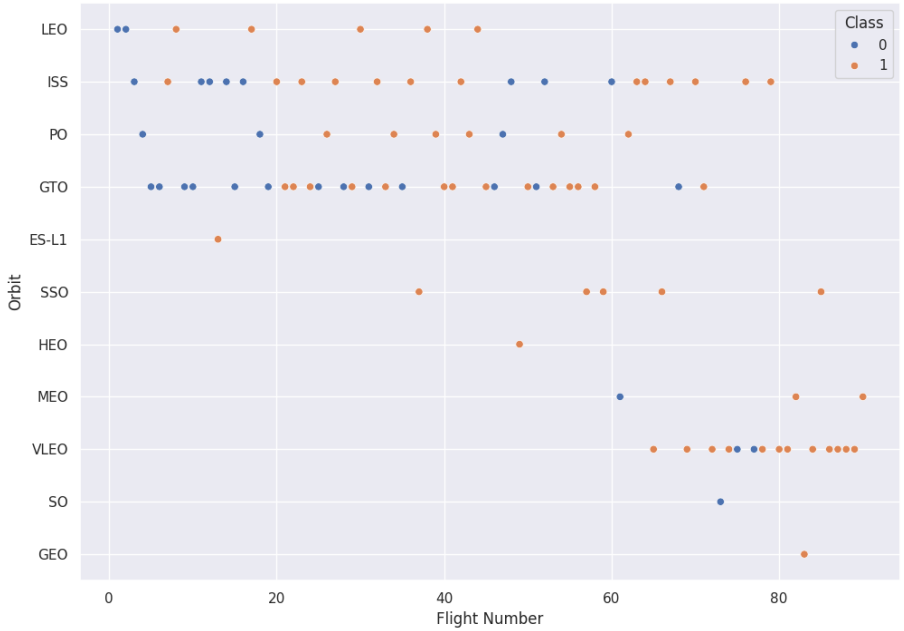
EDA with Data Visualization

The relationship between Payload Mass and Launch Site

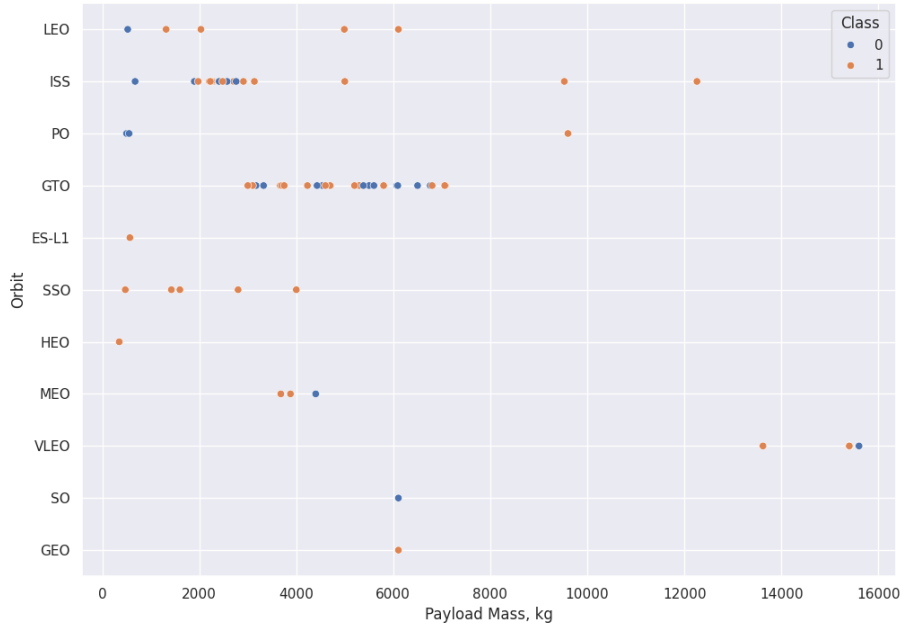


EDA with Data Visualization

The relationship between FlightNumber and Orbit type

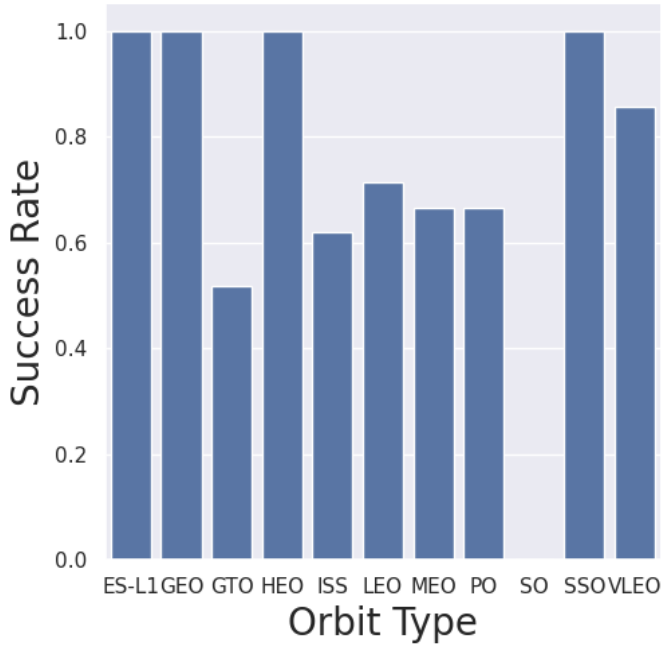


The relationship between Payload Mass and Orbit type

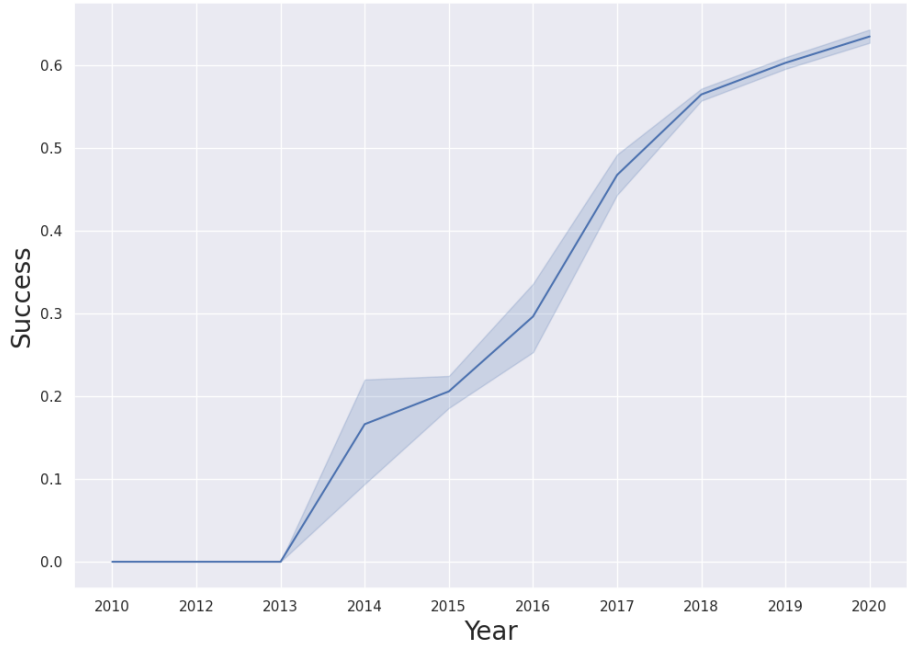


EDA with Data Visualization

The relationship between success rate of each orbit type



The launch success yearly trend



EDA with Data Visualization

Displaying 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

Displaying 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

EDA with Data Visualization

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

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

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

SUM(PAYLOAD_MASS__KG_)
45596

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

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

EDA with Data Visualization

Listing the date when the first succesful landing outcome in ground pad was acheived.

```
%sql SELECT MIN(DATE) \
FROM SPACEXTABLE \
WHERE Landing_Outcome \
LIKE 'Success (ground pad)'
```

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

MIN(DATE)
2015-12-22

Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT Booster_Version \
FROM SPACEXTABLE \
WHERE Landing_Outcome \
LIKE 'Success (ground pad)' \
AND PAYLOAD_MASS__KG_ > 400 \
AND PAYLOAD_MASS__KG_ < 6000
```

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

Booster_Version
F9 FT B1019
F9 FT B1025.1
F9 FT B1031.1
F9 FT B1032.1
F9 FT B1035.1
F9 B4 B1039.1
F9 B4 B1040.1
F9 FT B1035.2
F9 B4 B1043.1

EDA with Data Visualization

Listing the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(Mission_Outcome) \
FROM SPACEXTABLE WHERE Mission_Outcome \
LIKE "%Success%" UNION \
SELECT COUNT(Mission_Outcome) \
FROM SPACEXTABLE WHERE Mission_Outcome \
LIKE "%Failure%"
```

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

COUNT(Mission_Outcome)
1
100

Listing the names of the booster_versions which have carried the maximum payload mass

```
%sql SELECT Booster_Version \
FROM SPACEXTABLE \
WHERE PAYLOAD_MASS__KG_ == \
(SELECT PAYLOAD_MASS__KG_ \
FROM SPACEXTABLE \
ORDER BY PAYLOAD_MASS__KG_)
```

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

Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004

EDA with Data Visualization

Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

```
%sql SELECT SUBSTR(Date, 6, 2), Landing_Outcome, Booster_Version, Launch_Site \
FROM SPACEXTABLE \
WHERE Landing_Outcome LIKE '%Failure (drone ship)' AND \
(SELECT SUBSTR(Date, 0, 5) FROM SPACEXTABLE) \
ORDER BY SUBSTR(Date, 6, 2)
```

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

SUBSTR(Date, 6, 2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
01	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
03	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
06	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

EDA with Data Visualization

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

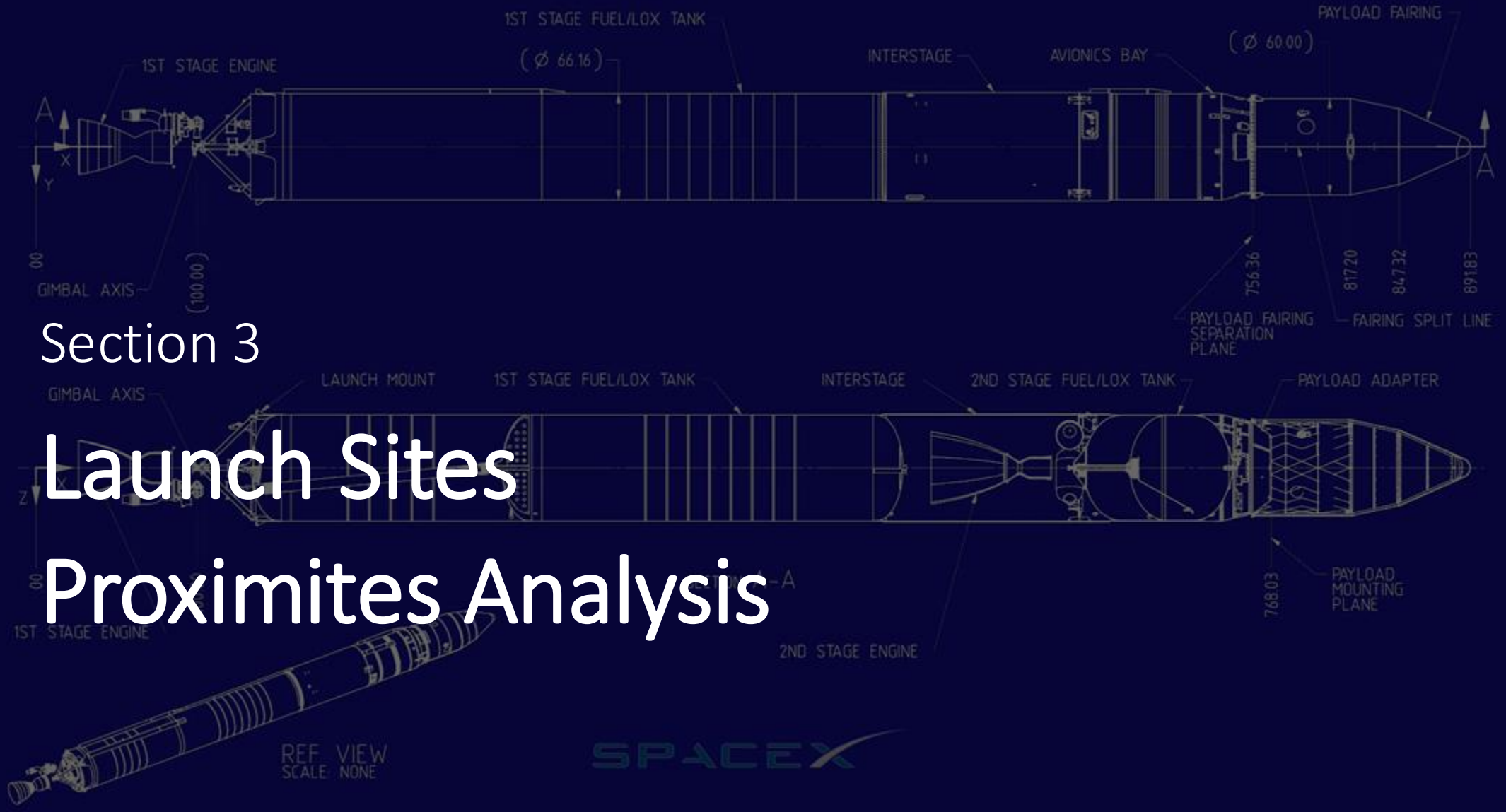
```
%sql SELECT DISTINCT Date, Landing_Outcome, COUNT(Landing_Outcome) \
FROM SPACEXTABLE \
    WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY Landing_Outcome \
ORDER BY COUNT(Landing_Outcome) DESC
```

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

Date	Landing_Outcome	COUNT(Landing_Outcome)
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

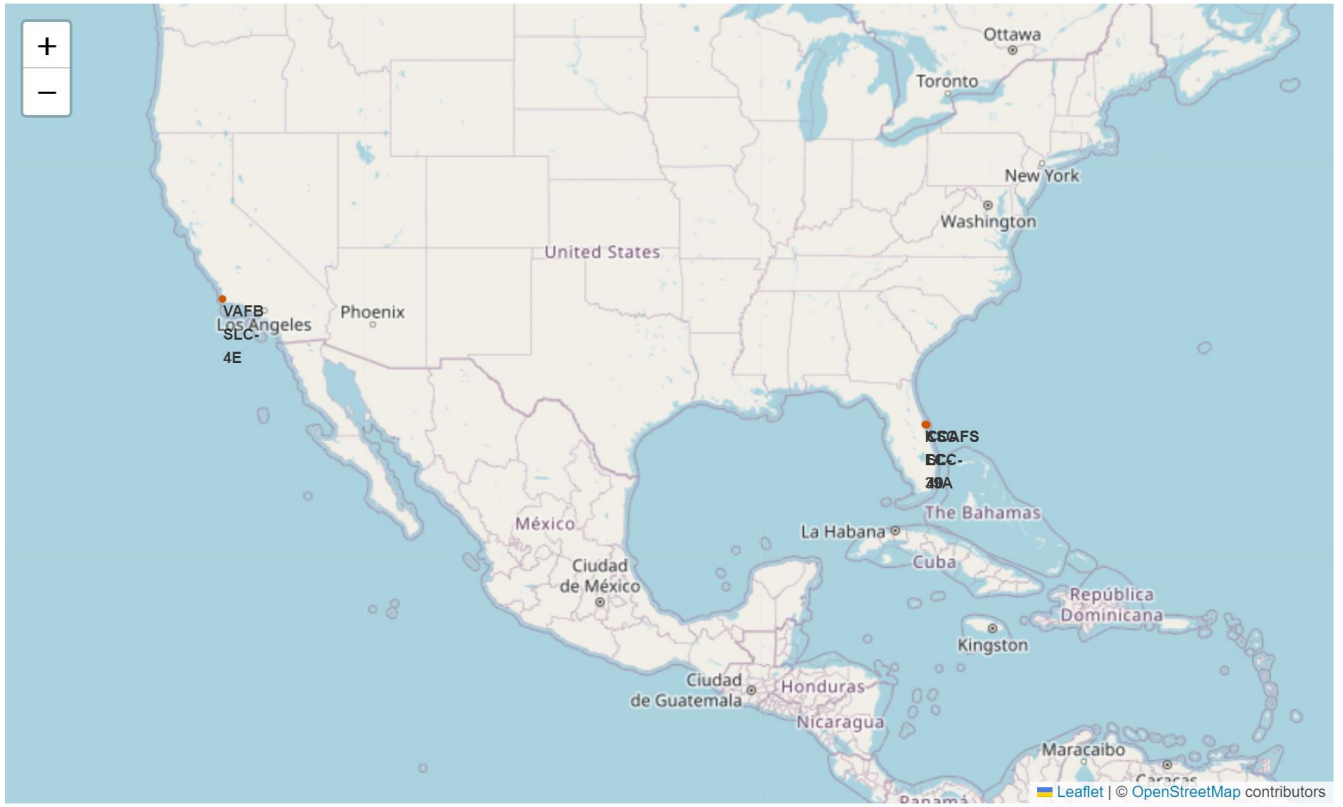
Section 3

Launch Sites Proximites Analysis



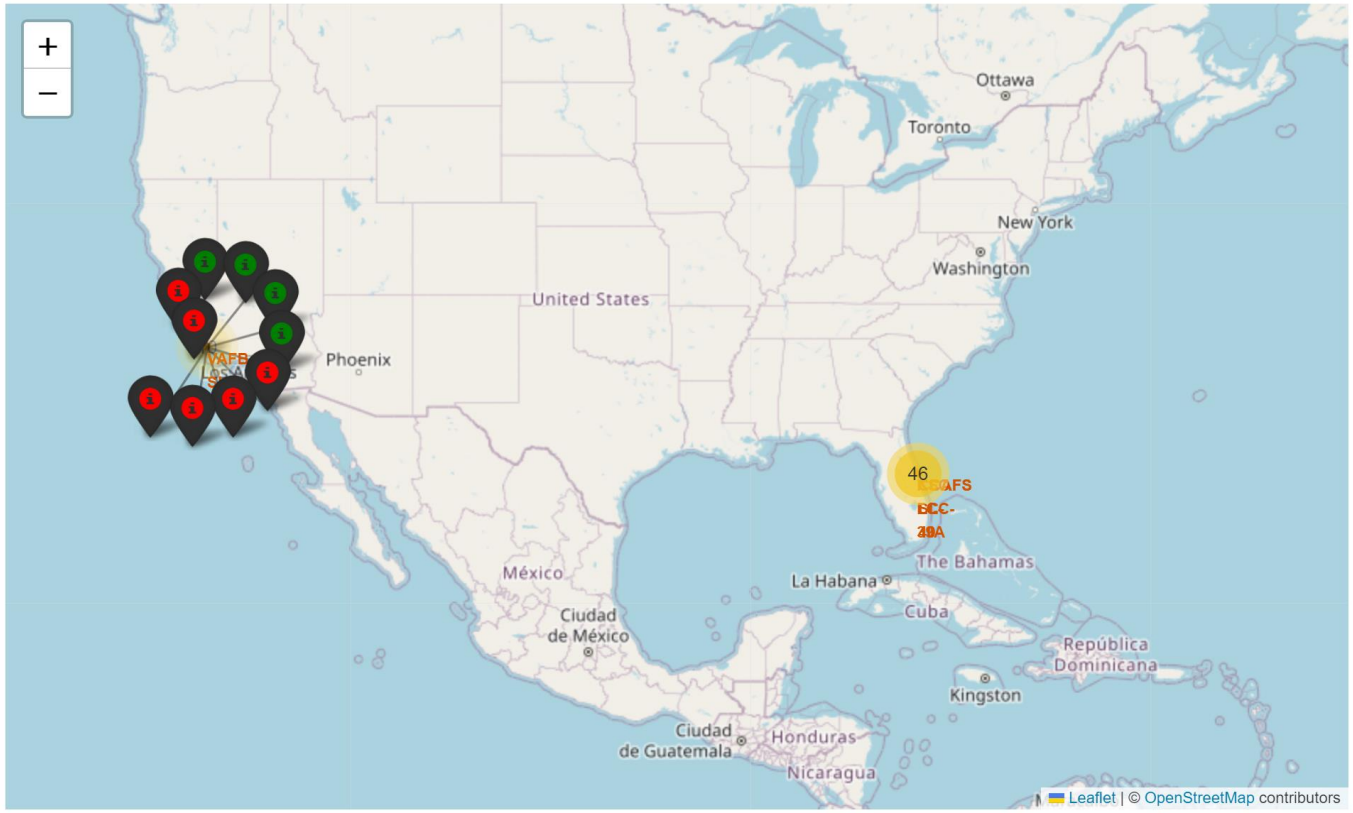
Folium map

Here is a screenshot of the Folium map with markers created for each launch pad.



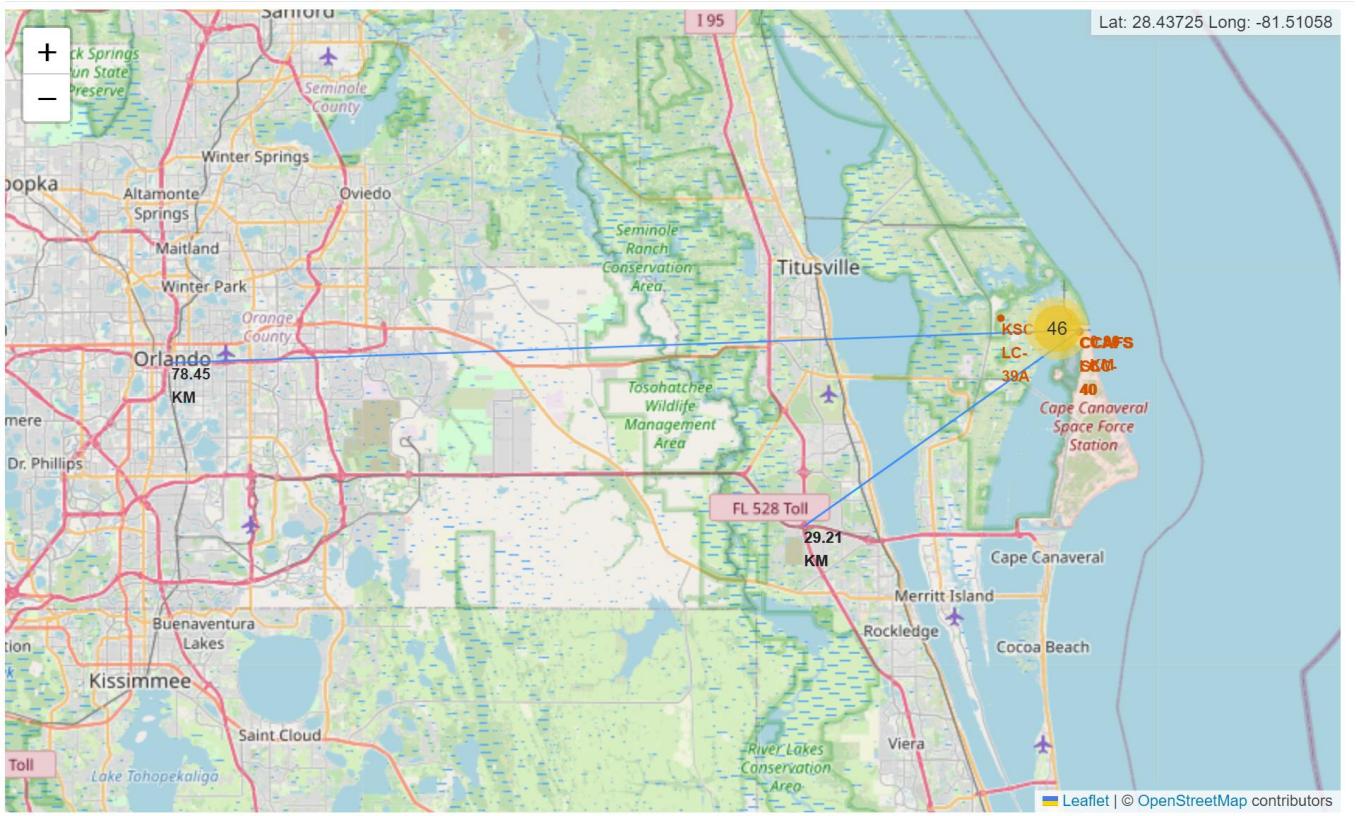
Folium map

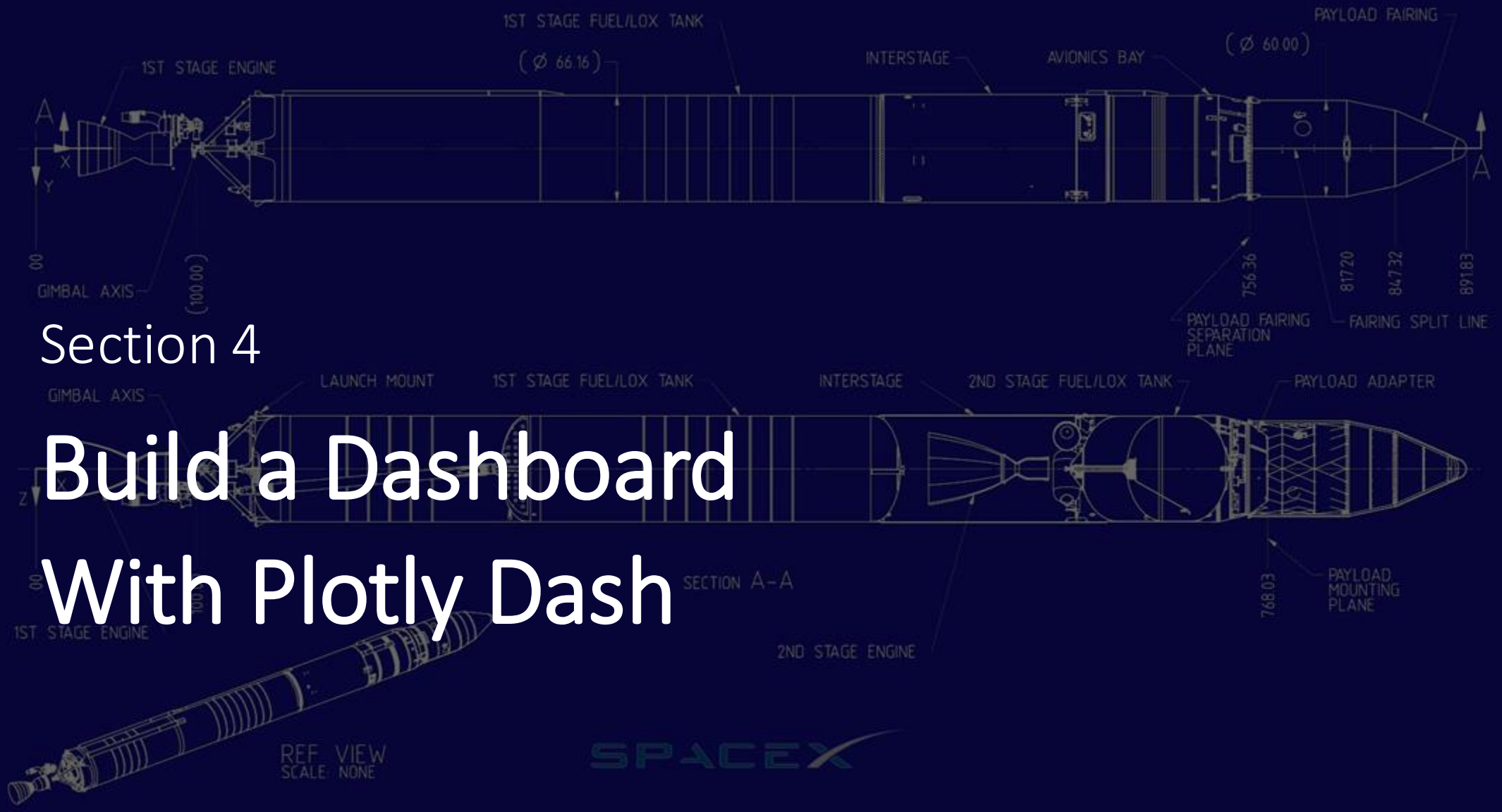
Colored markers have been created to indicate the success or failure of the landing.



Folium map

The calculated distance to the coast, railway, and nearest city has been added to the map.



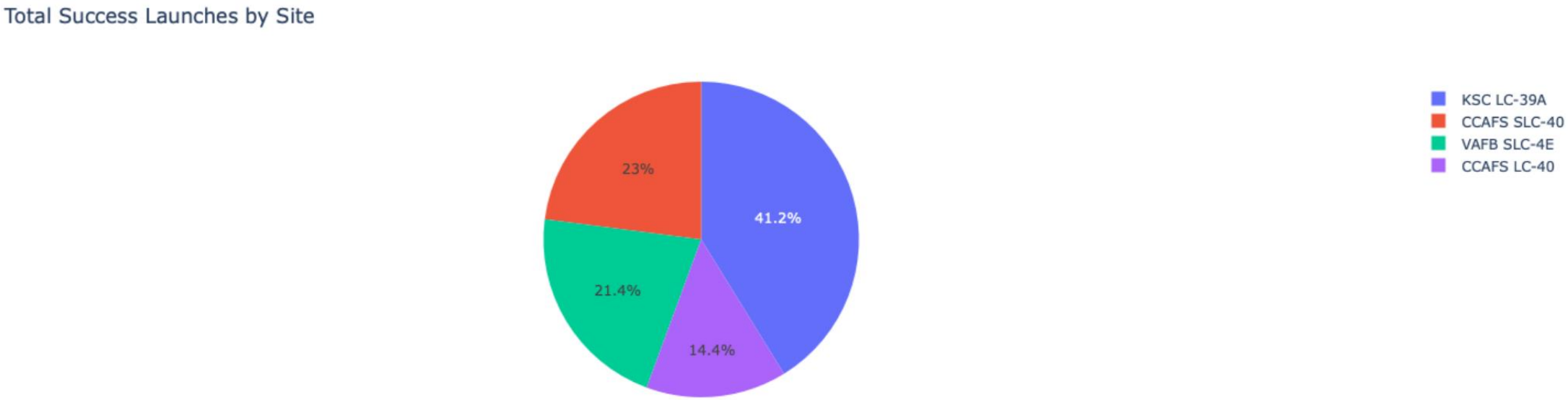


Section 4

Build a Dashboard With Plotly Dash

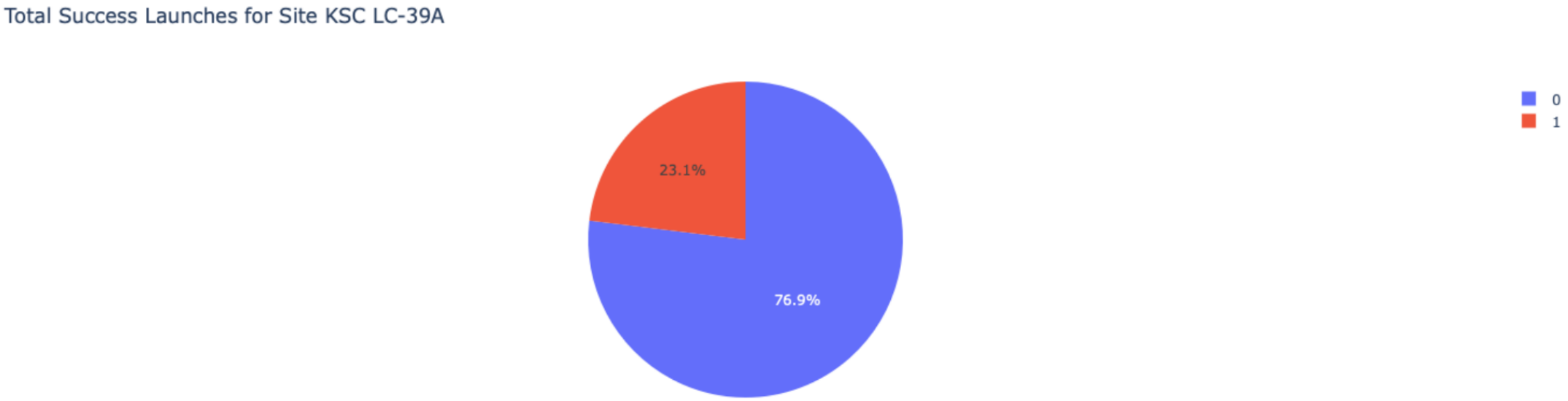
Successful launches

Below you can see successful rocket launches from different launch pads.



Successful launches

Launch location with the highest success rate

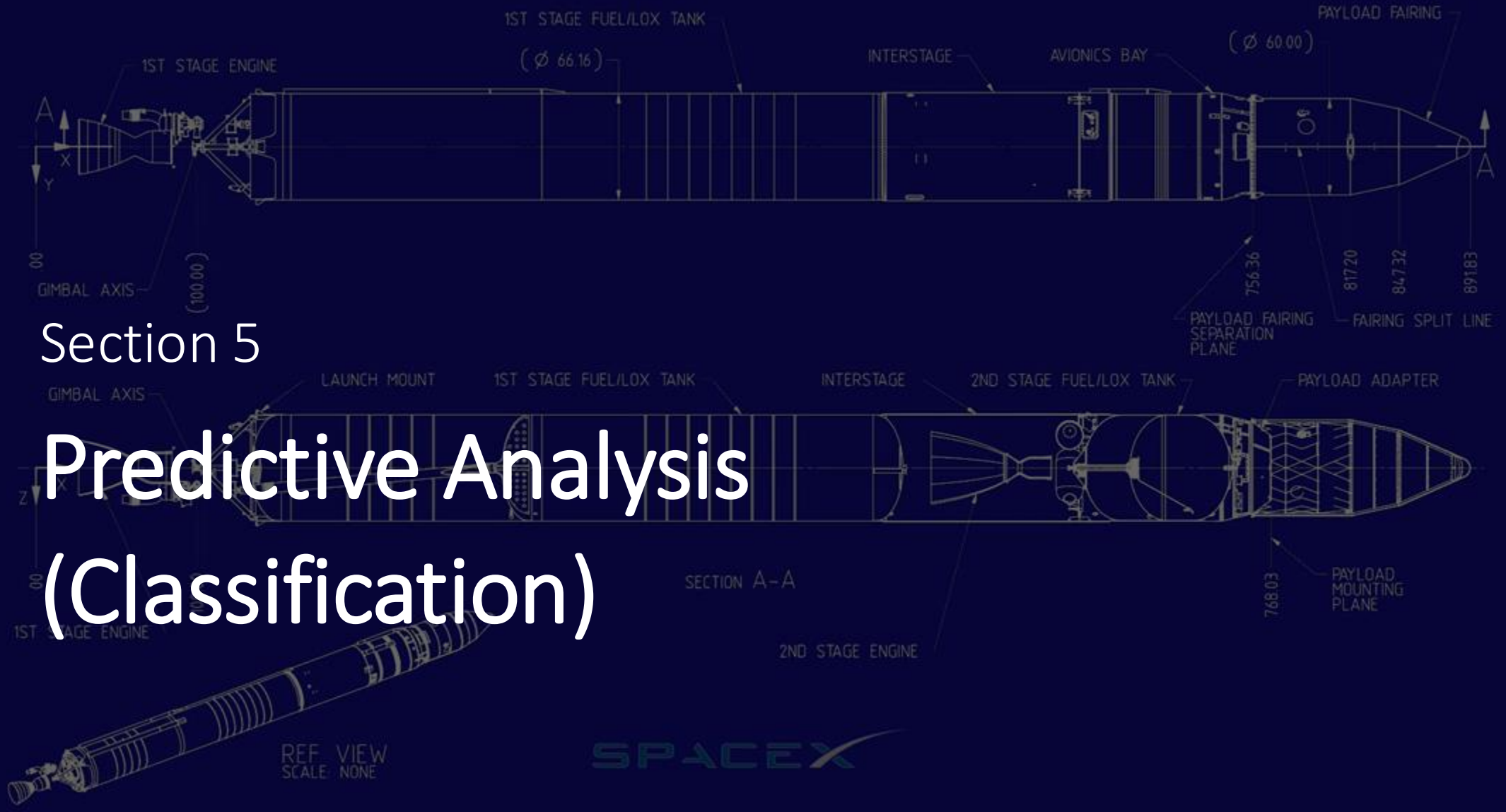


Payload vs. Launch Outcome scatter plot

With a payload of up to 5000, the FT has the greatest success.

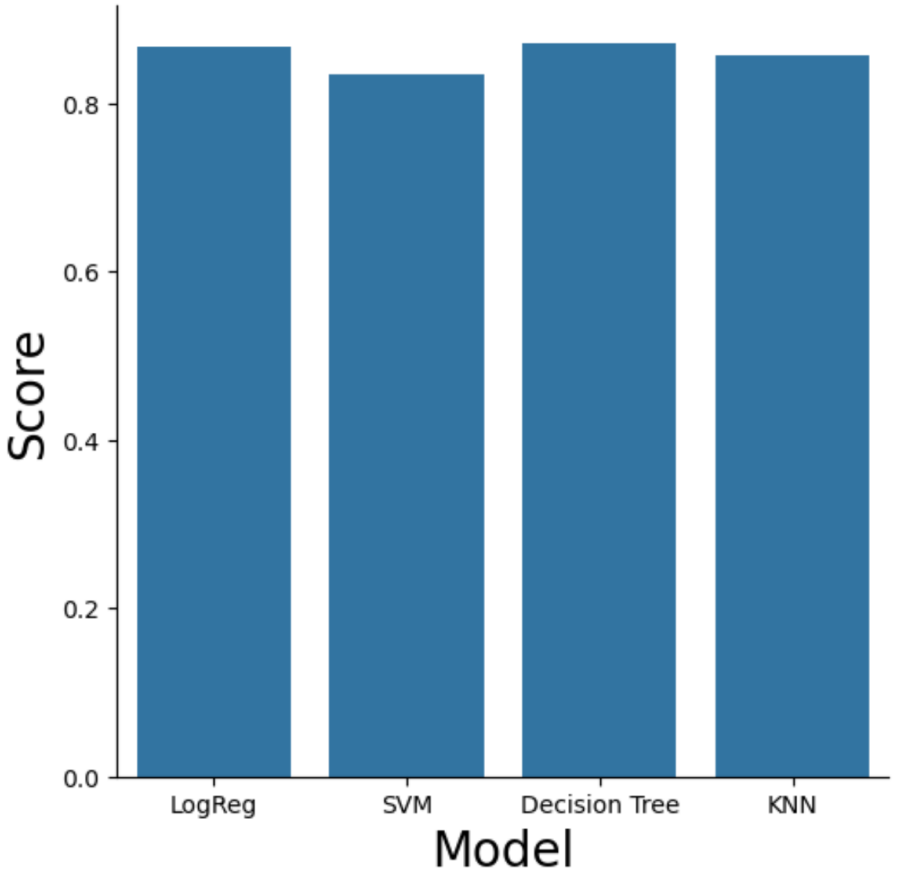
Section 5

Predictive Analysis (Classification)



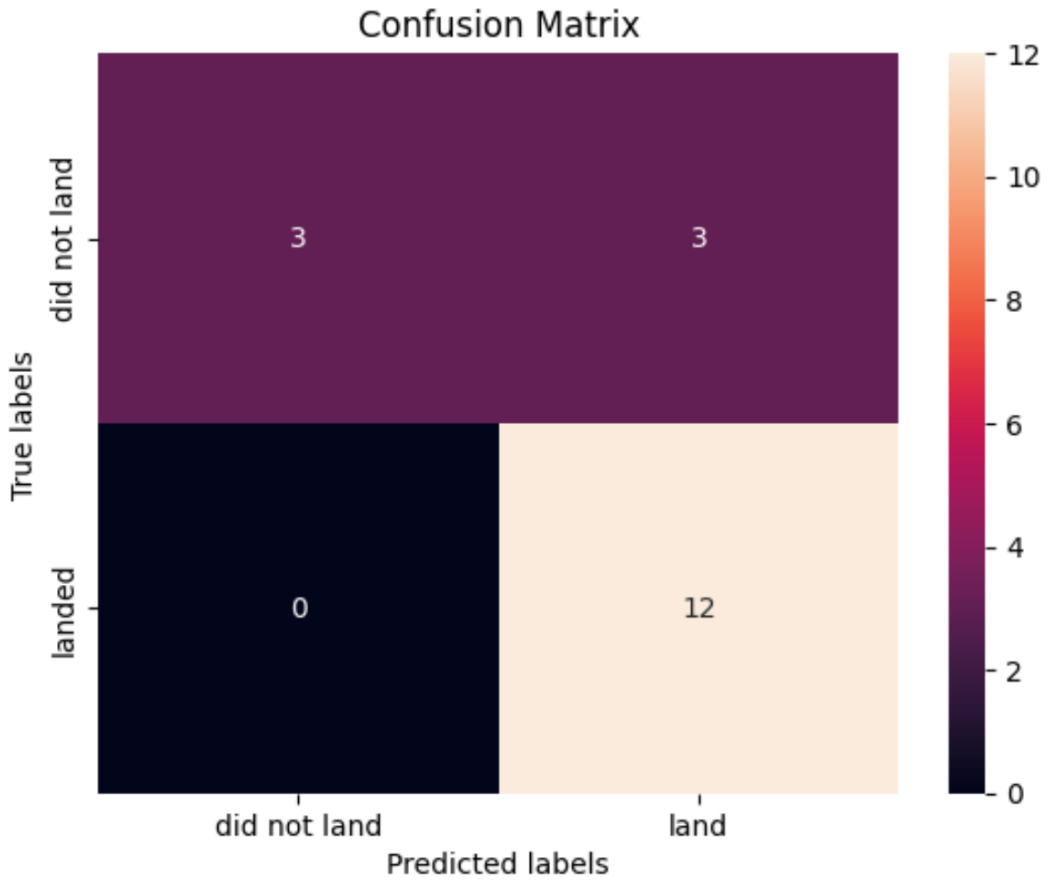
Classification Accuracy

During the project, the efficiency of each model was assessed. According to the results of the assessment, the Decision Tree has the highest accuracy.



Classification Accuracy

Examining the confusion matrix, we see that Decision Tree can distinguish between the different classes. We see that the problem is false positives. True Postive - 12 (True label is landed, Predicted label is also landed). False Postive - 3 (True label is not landed, Predicted label is landed).



Conclusions

- Payload has a strong impact on mission success
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- In addition, the amount of payload also has a strong correlation with the success of entering orbit
- Also an important factor for success is the launch site.
- KSC LC-39A is the site that has the highest percentage of successful missions

Thank you!