IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fernando A. Scelzo
20 Nov 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection via API and Web Scraping
  - Data Wrangling
  - EDA with SQL and visuals
  - Maps and interactive Dashboards presentation
  - Machine Learning prediction model building
- Summary of all results
  - Some hypothesis were drawn
  - The optimal model(s) were found

Introduction

Companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX are making space travel more accessible and affordable. **SpaceX** stands out for its achievements, including sending spacecraft to the International Space Station and launching manned missions to space.

The **Falcon 9 rocket** consists of two stages, with the first stage doing most of the work and being significantly larger and more expensive than the second stage. SpaceX's ability to **reuse the first stage of the Falcon 9** is a major factor in reducing launch costs.

In this project, we will act as a data scientist for a new company, **Space Y**, aiming to compete with SpaceX.

Our task involves gathering **data on SpaceX**, creating dashboards, and training a machine learning model to **predict whether the first stage of the Falcon 9 will be reused**.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

Data collected via SPACEX API and Web Scrapping from Wikipedia

Perform data wrangling

Data was analysed to determine the best Training Labels

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

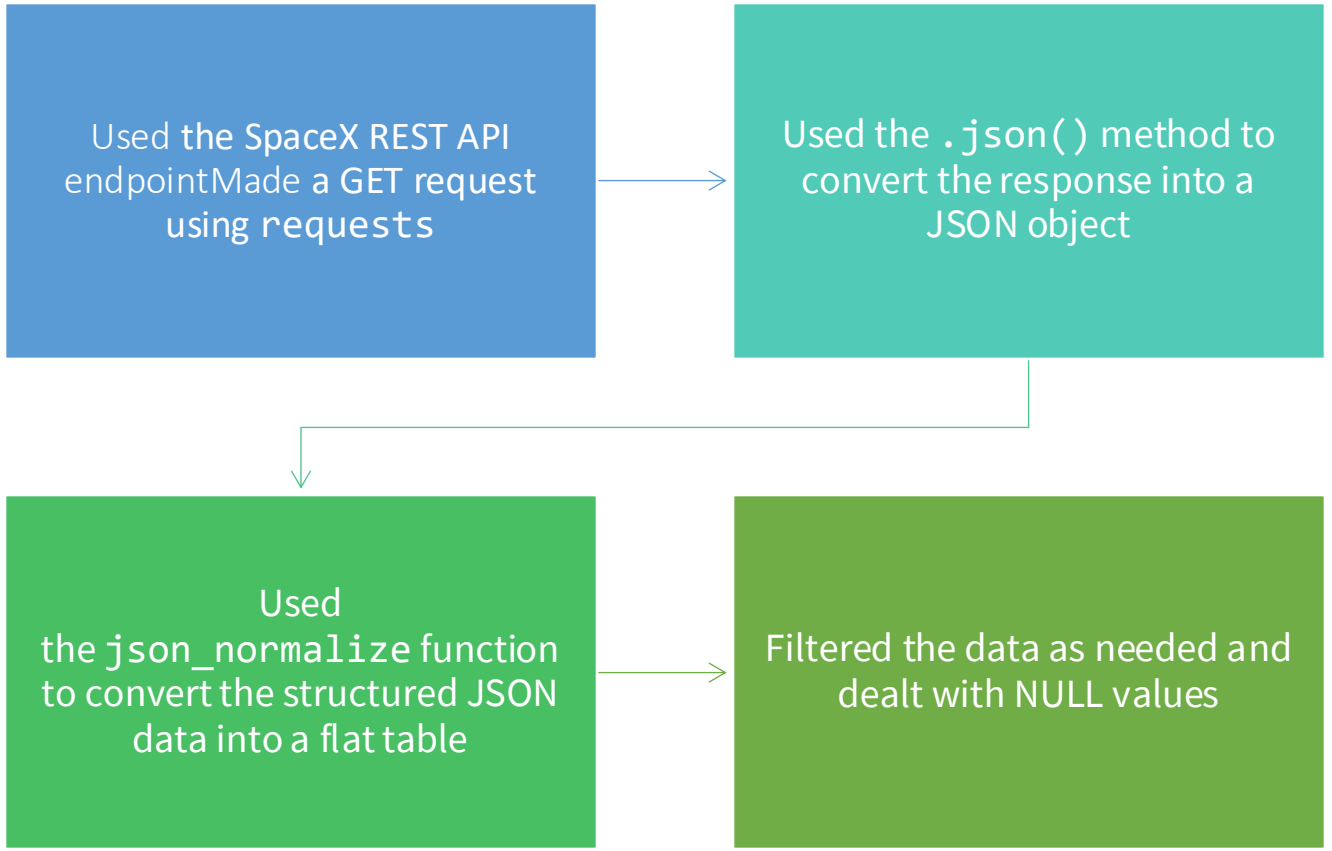
- API Data Collection:

  - **Using the SpaceX REST API**: I gathered launch data by performing GET requests to specific endpoints of the SpaceX REST API.
  - **Data Retrieval**: The responses, which were in JSON format, were then converted into a DataFrame using the `json_normalize` function from the pandas library.
  - **Handle Data Issues**: I filtered data as needed (e.g., removing unwanted launches) and dealt with NULL values by calculating means or using other methods to clean the dataset.

- Web Scraping:

  - **Utilizing BeautifulSoup**: I employed the BeautifulSoup package in Python to scrape HTML tables from related Wikipedia pages.
  - **Data Parsing and Conversion**: The scraped data was parsed and subsequently converted into a Pandas DataFrame for further analysis.

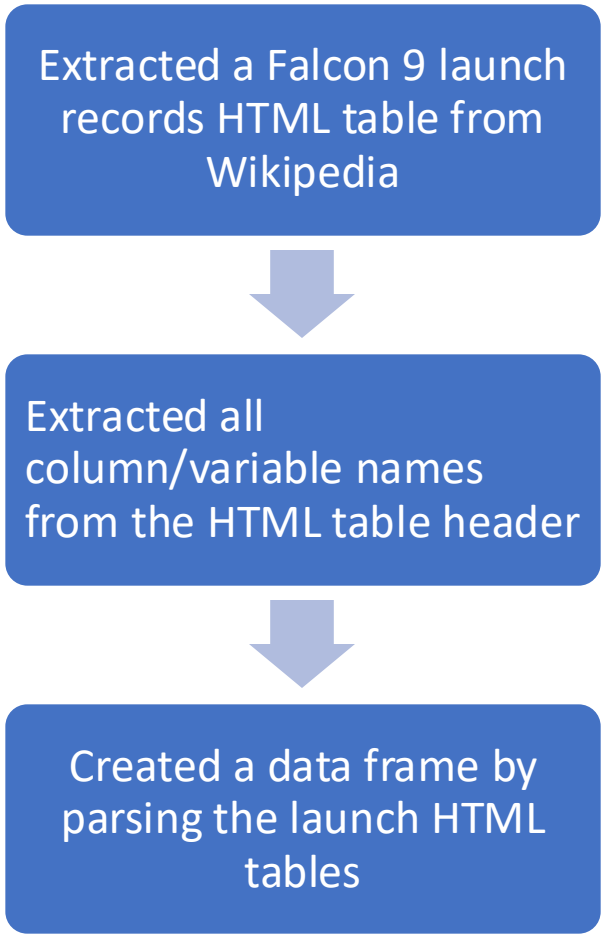
Data Collection – SpaceX API



- GitHub URL SpaceX API calls notebook

Data Collection - Scraping

- Web scraped Falcon 9 launch records with BeautifulSoup



- GitHub URL of web scraping notebook

Data Wrangling

Attributes Reviewed

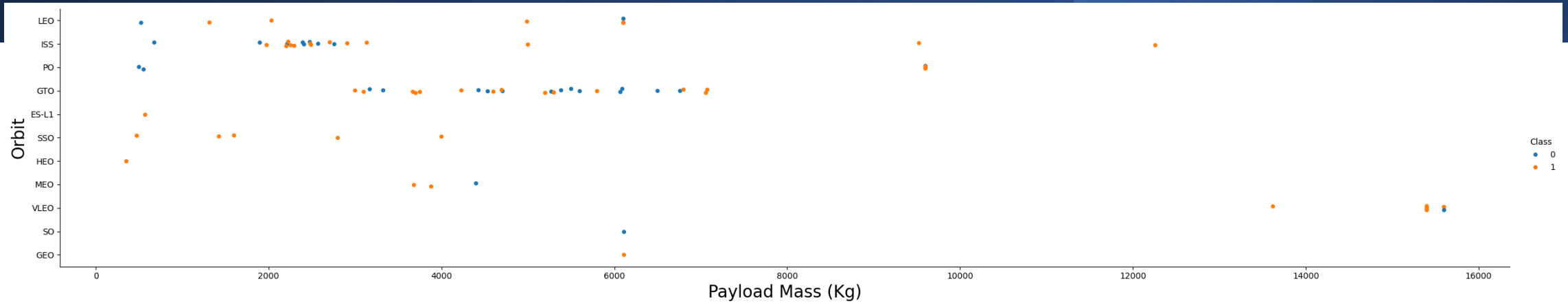
- Number of launches for each site
- Number and occurrence of each orbit
- Number and occurence of mission outcome of the orbits

Outcome Classification

- Outcome converted into binary classification (0 or 1)
- Landing outcome label created

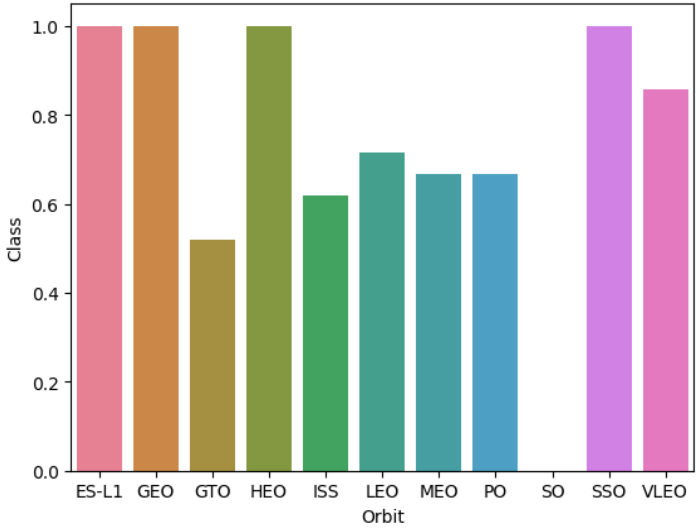
GitHub URL

EDA with Data Visualization

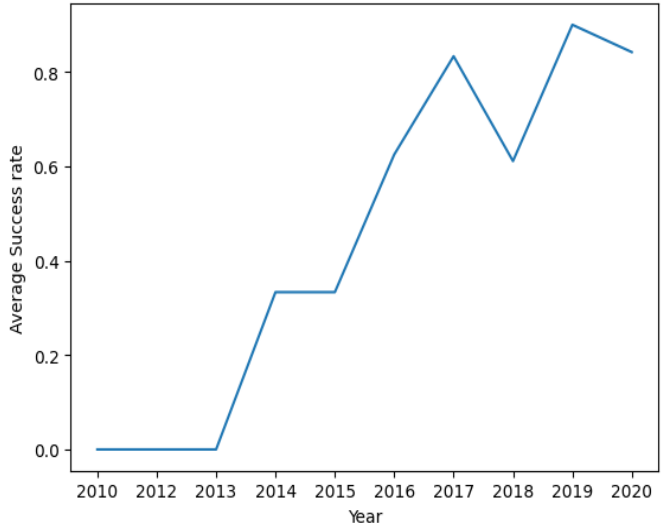


Payload vs. Orbit scatter plot

Success rate per Orbit type



Average launch success rate per year line chart



[GitHub URL]

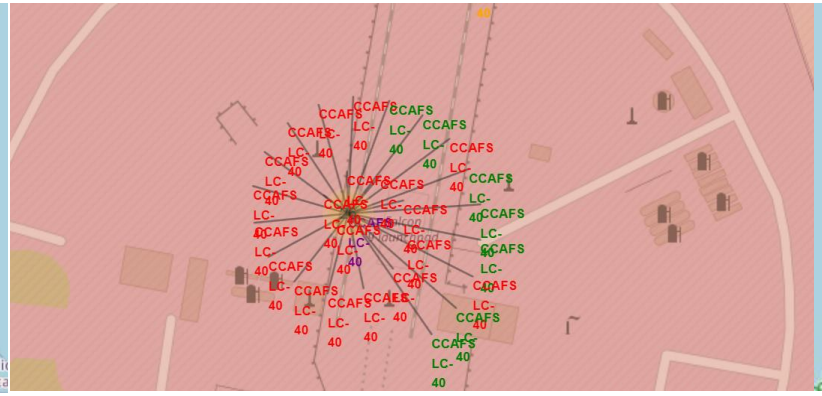
EDA with SQL

- EDA with SQL queries performed:
  - Display the names of the **unique launch site**s in the space mission
  - Display **5 records** where launch sites begin, for example, with the string 'CCA'
  - Display the **total payload mass** carried by boosters launched by NASA (CRS)
  - Display **average payload** mass carried by booster version F9 v1.1
  - List the **date when the first successful landing** outcome in ground pad was achieved
  - List the names of the **boosters which have success in drone ship** and have **payload mass greater than 4000 but less than 6000**
  - List the **total number** of successful and failure **mission outcomes**
  - List the names of the booster_versions which have carried the maximum payload mass
  - List the records which will display the **month names**, failure landing_outcomes in drone ship, booster versions, launch_site for the months **in year 2015**
  - **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

GitHub URL

Build an Interactive Map with Folium

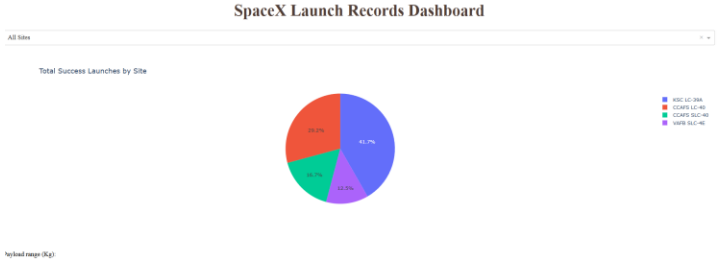
- Using the Python Folium library:
  - I generated interactive maps, plotted coordinates, and marked clusters by writing Python code, to show launch sites
  - I built an interactive map to mark success/failed launches for each site
  - I analyzed the launch site proximity and calculated distances on an interactive map

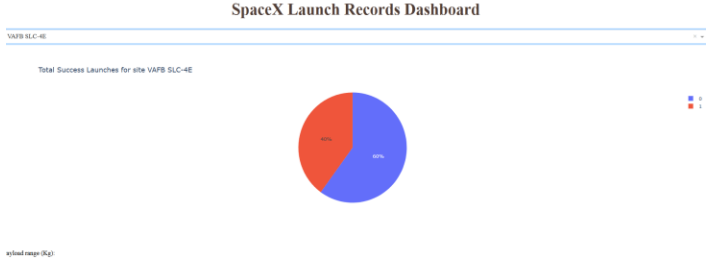


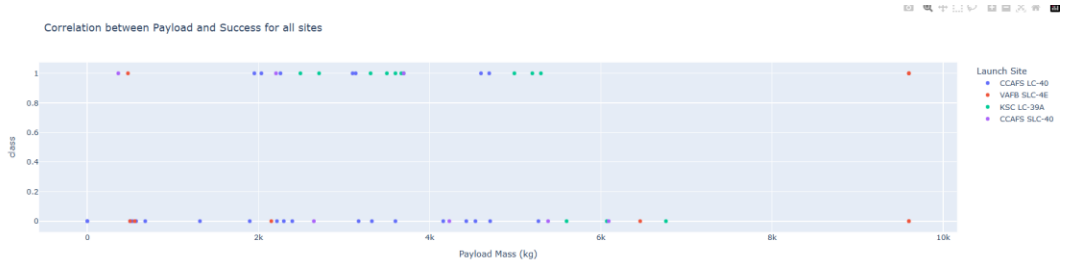
GitHub URL

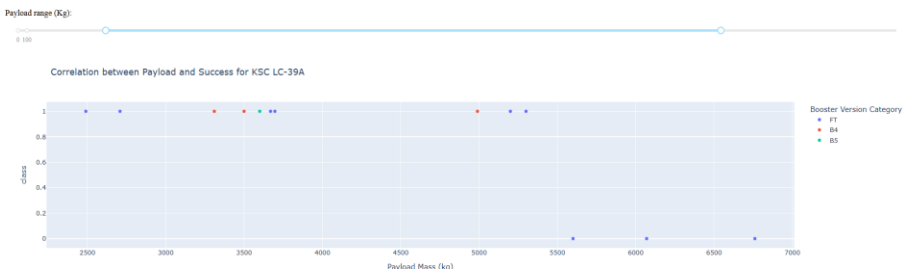
Build a Dashboard with Plotly Dash

- I built an Interactive Dashboard with Ploty Dash:
  - Added a Launch Site Drop-down Input Component
  - Added a callback function to render a pie chart based on selected site dropdown
  - Added a Range Slider to Select Payload
  - Added a callback function to render the Success Payload scatter plot



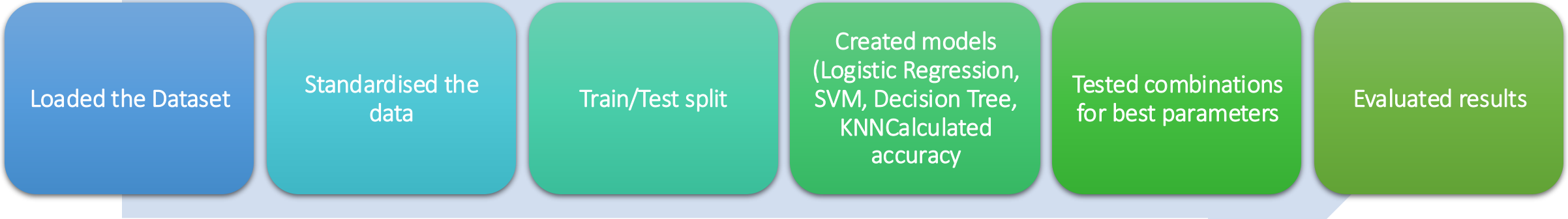






[GitHub URL]

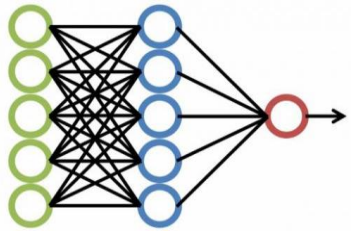
Predictive Analysis (Classification)



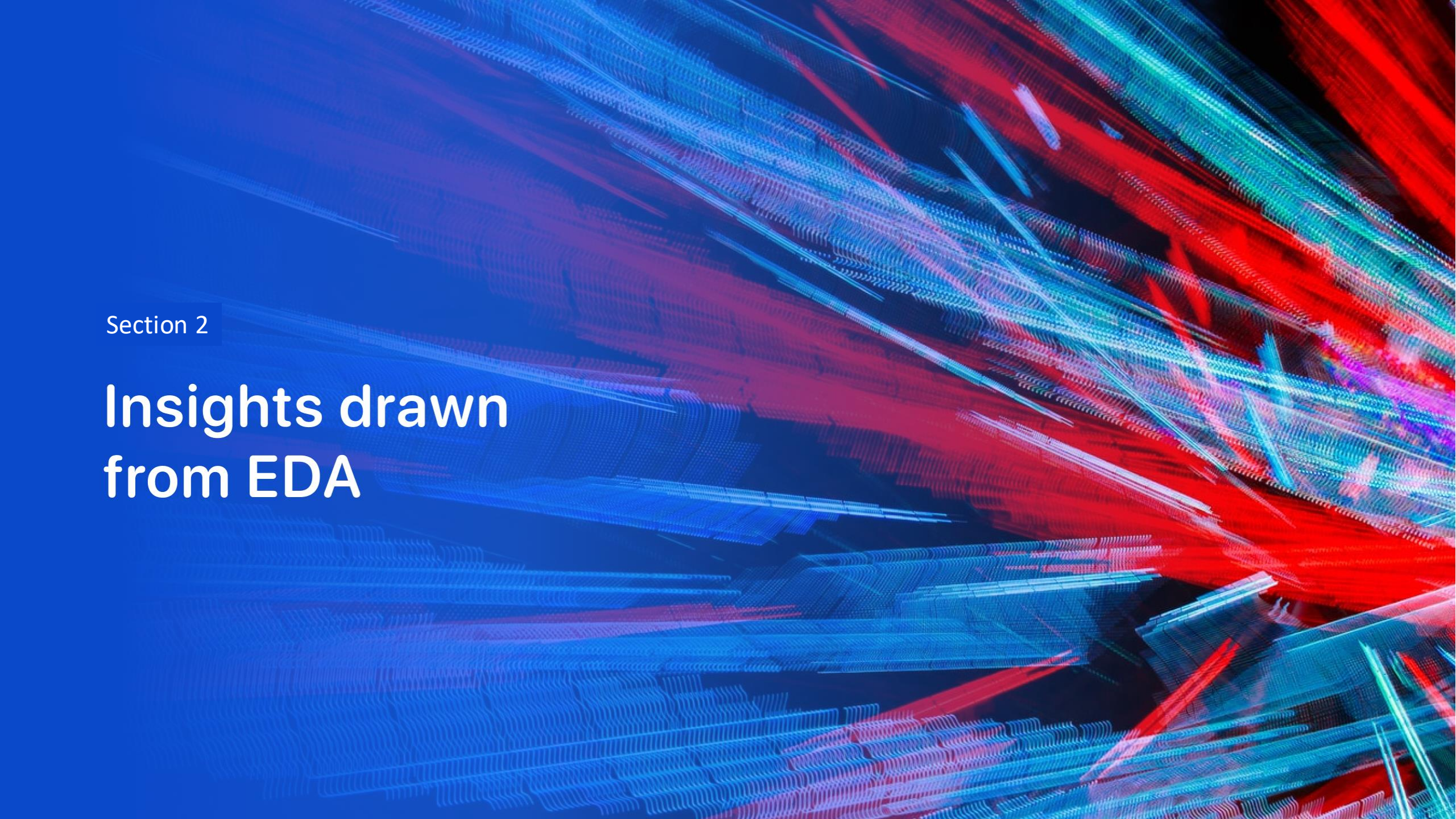
- The data was loaded and standardised, then
- split into training data and test data
- The best Hyperparameter was found for Logistic Regression, KNN, Classification Trees, and SVM models
- It was evaluated what method(s) performed best

GitHub URL

Results

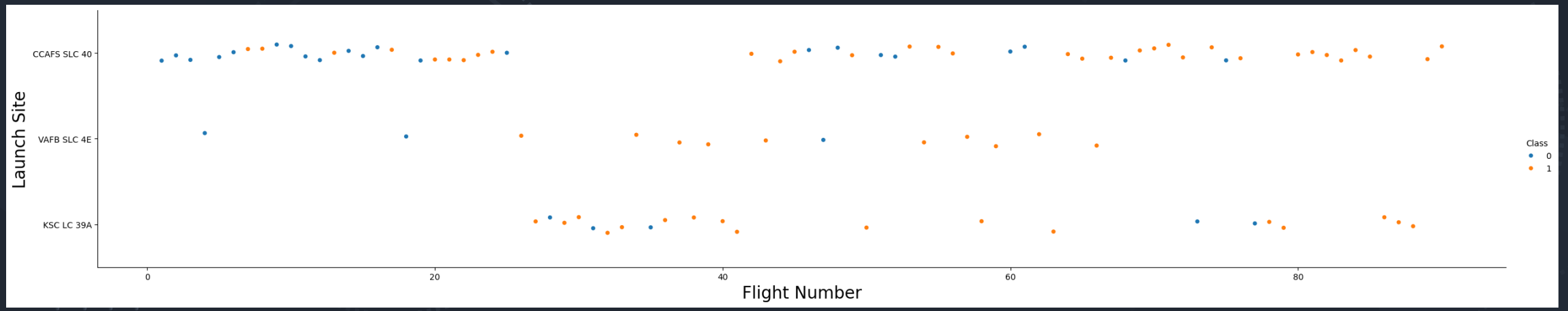


- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

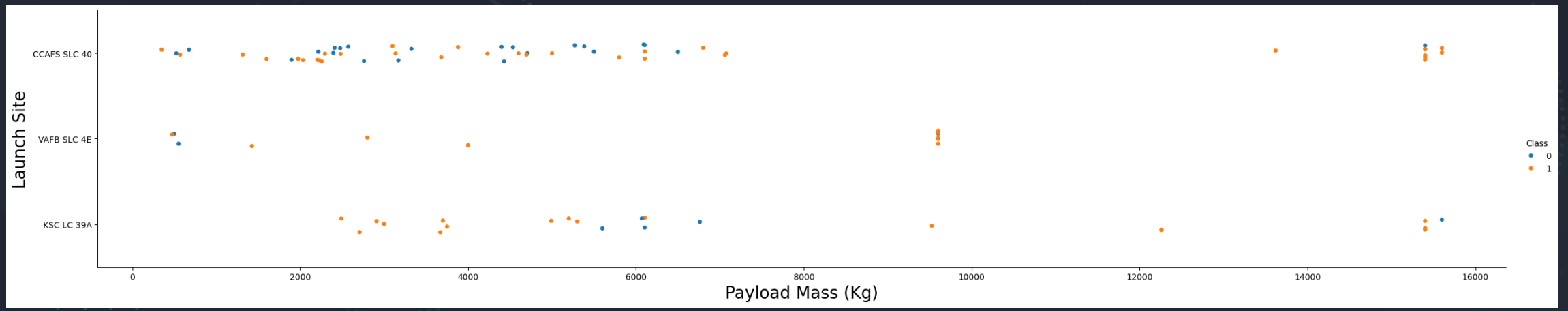
Flight Number vs. Launch Site

- The majority of launches occur at CCAFS SLC 40
- There's no significant pattern of success/failure as regards site or trend over time



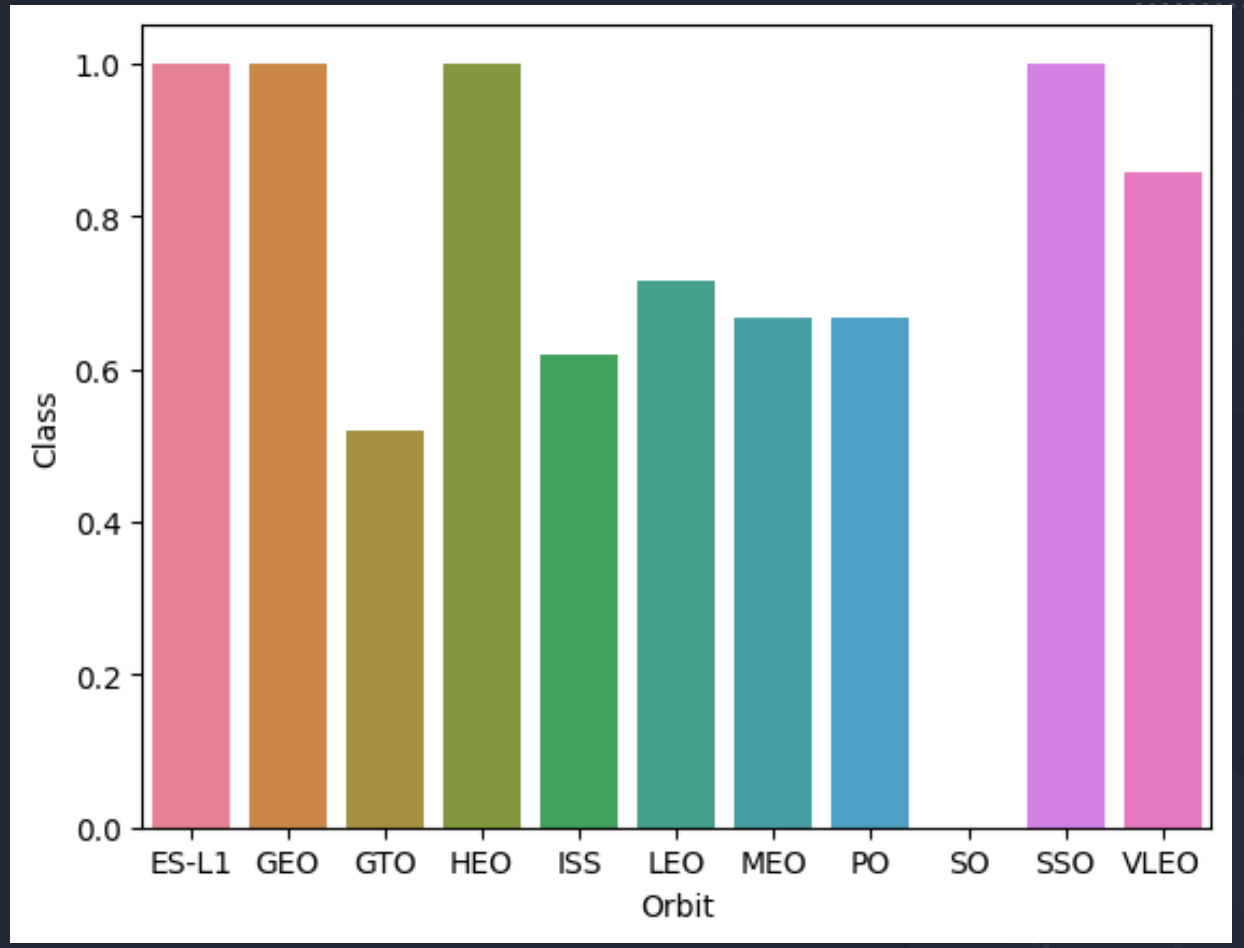
Payload vs. Launch Site

- VAFB SLC 4E has fewer launches and generally with smaller payloads
- It would appear that heavier payloads have a higher chance of success, but there may be other factors involved



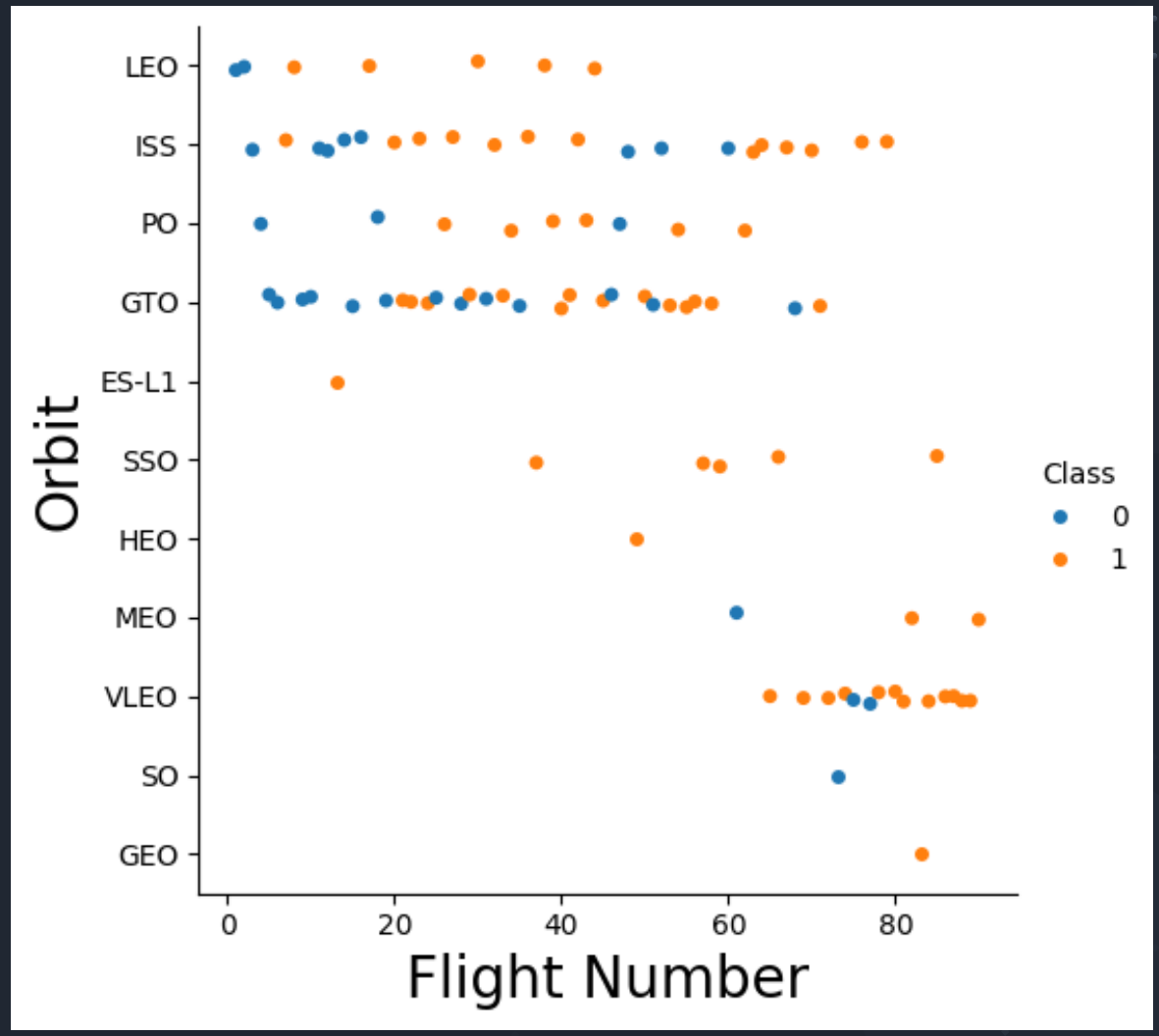
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have higher success rates
- GTO has the lowest one



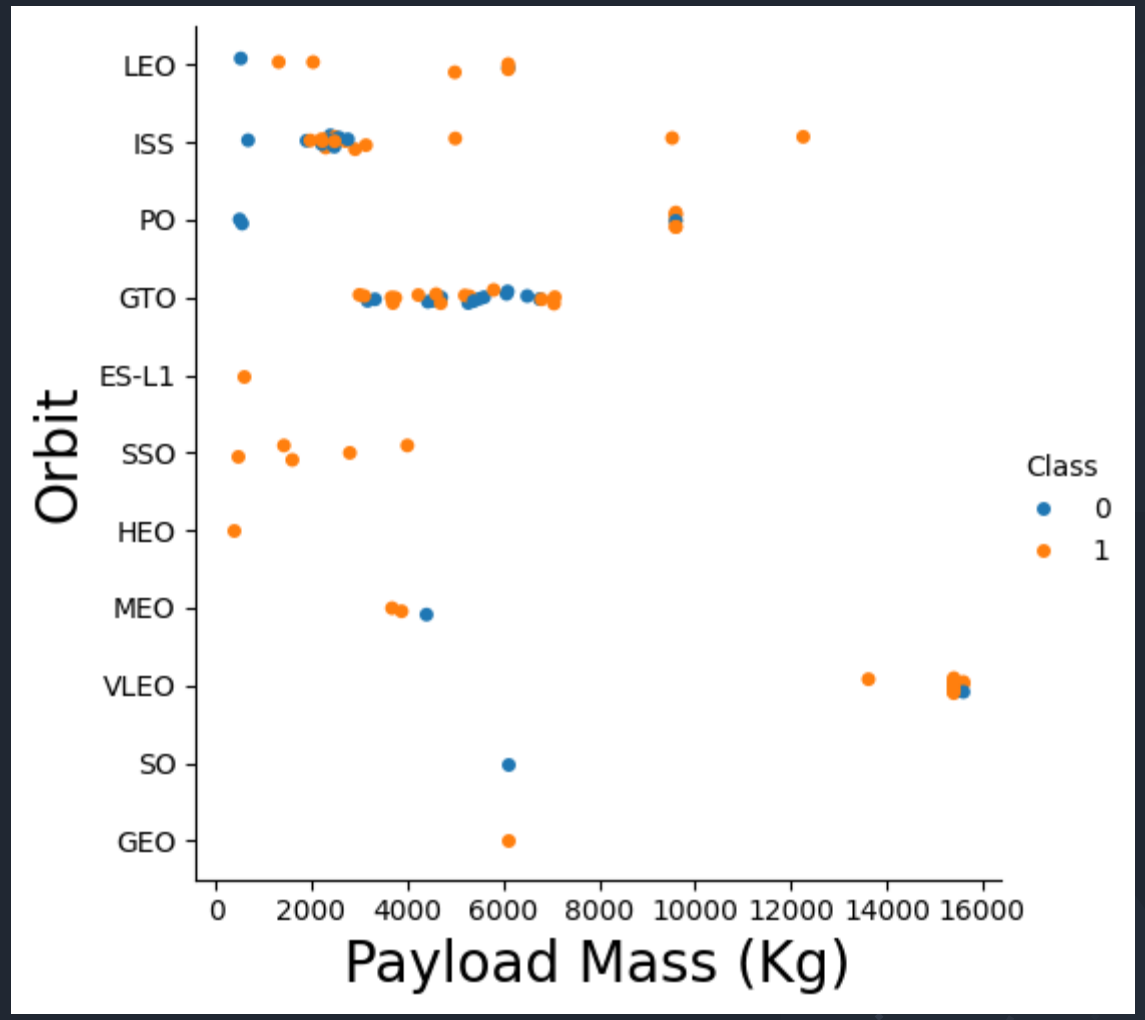
Flight Number vs. Orbit Type

- Most flights are concentrated in ISS (International Space Station), GTO (Geostationary Transfer Orbit), and VLEO (Very Low Earth Orbit).
- LEO, VLEO, SSO, and VLEO are generally more successful as flight number increases.
- There does not seem to be a relationship between Orbit and Flight Number in GTO and ISS.



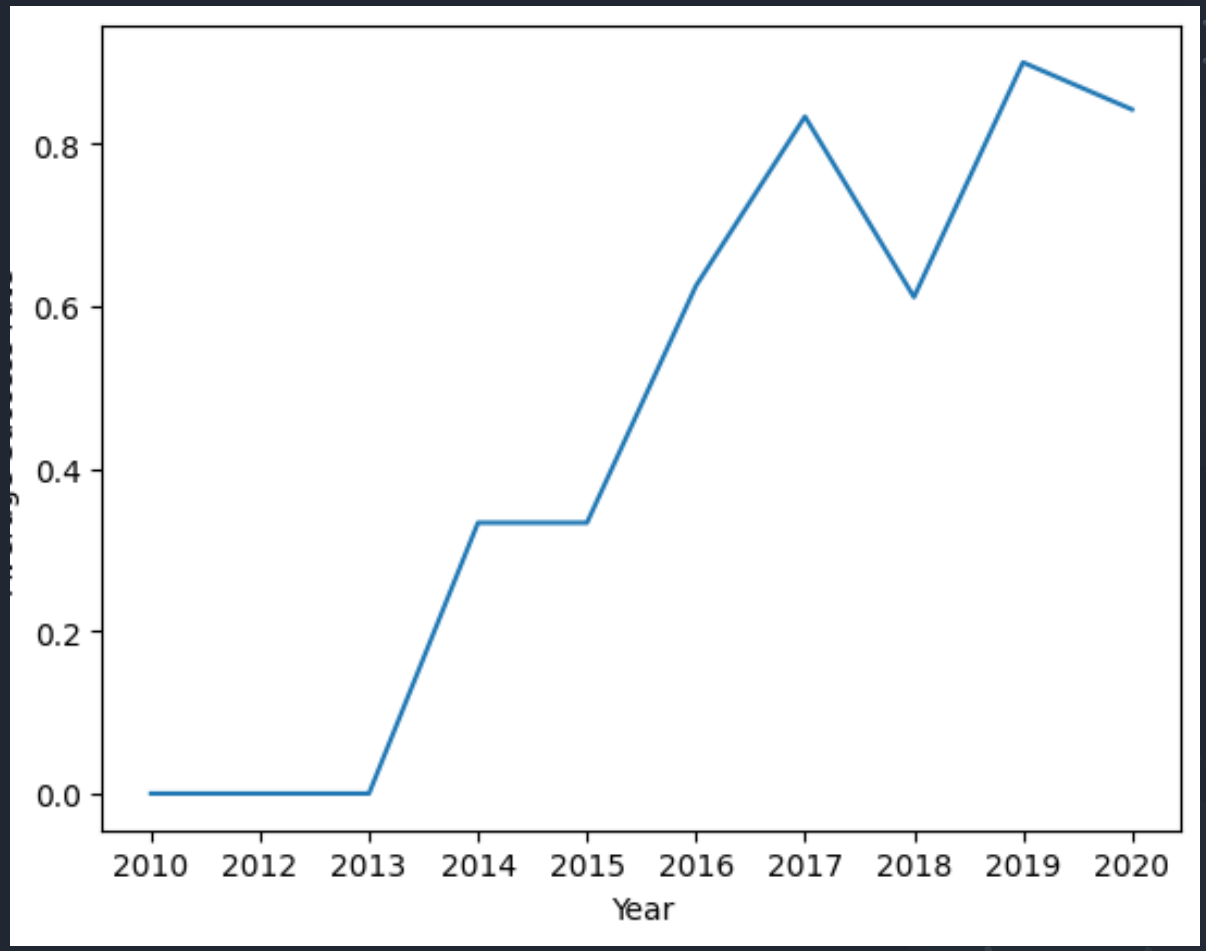
Payload vs. Orbit Type

- Generally speaking, launches with higher payload masses seem to have a higher success rate
- SSO seems more reliable for lower orbits (but this may be misleading due to fewer data points)
- GTO success may not be directly related to Payload Mass



Launch Success Yearly Trend

- Launch Success has continuously improved during 2013-2017 (with a small plateau during 2014).
- There has been a decline in success rate after 2017, but it has improved again in 2018.



All Launch Site Names

```
%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

Launch Site Names Begin with 'CCA'

```
%%sql SELECT * FROM SPACEXTABLE
WHERE Launch_Site LIKE "CCA%"
LIMIT 5
```

Python

```
 * 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

Total Payload Mass

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by 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_)
2534.6666666666665

First Successful Ground Landing Date

```
%%sql SELECT * from SPACEXTABLE
WHERE Landing_Outcome LIKE "%ground pad%"
ORDER BY Date ASC LIMIT 1
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT Booster_Version from SPACEXTABLE
WHERE Landing_Outcome LIKE "Success (drone ship)"
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 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

```
%%sql SELECT Mission_Outcome, SUM(1) AS Total_Missions
FROM SPACEXTABLE
GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

```
%%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

2015 Launch Records

```
%%sql SELECT SUBSTR(Date, 6,2) AS Month, Date, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = "Failure (drone ship)"
AND SUBSTR(Date, 0,5) = "2015"
```

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

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

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

```
%%sql SELECT Landing_Outcome, COUNT(1) as TOTAL, Date
FROM SPACEXTABLE
GROUP BY Landing_Outcome
HAVING (Date BETWEEN "2010-06-04" AND "2017-03-20")
ORDER BY Total DESC
```

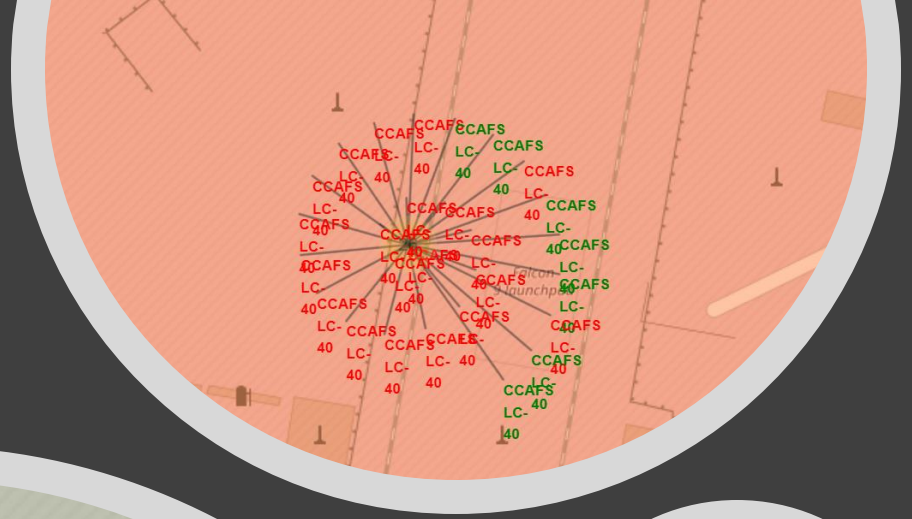
```
 * sqlite:///my_data1.db
Done.
```

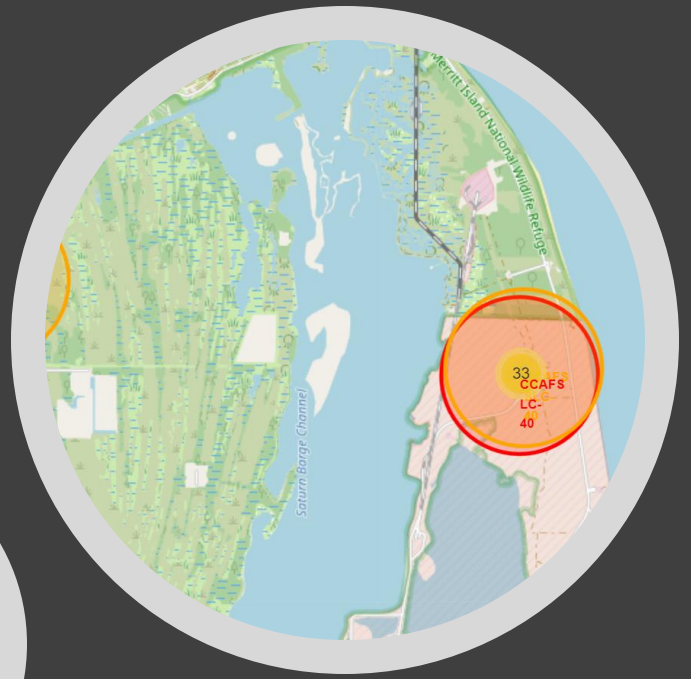
Landing_Outcome	TOTAL	Date
No attempt	21	2012-05-22
Success (drone ship)	14	2016-04-08
Success (ground pad)	9	2015-12-22
Failure (drone ship)	5	2015-01-10
Controlled (ocean)	5	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

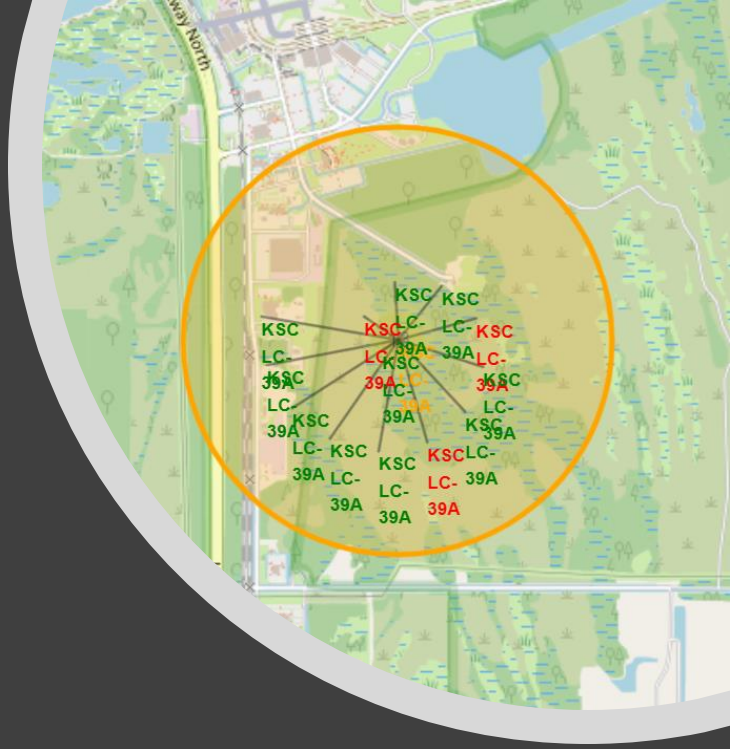
Section 3

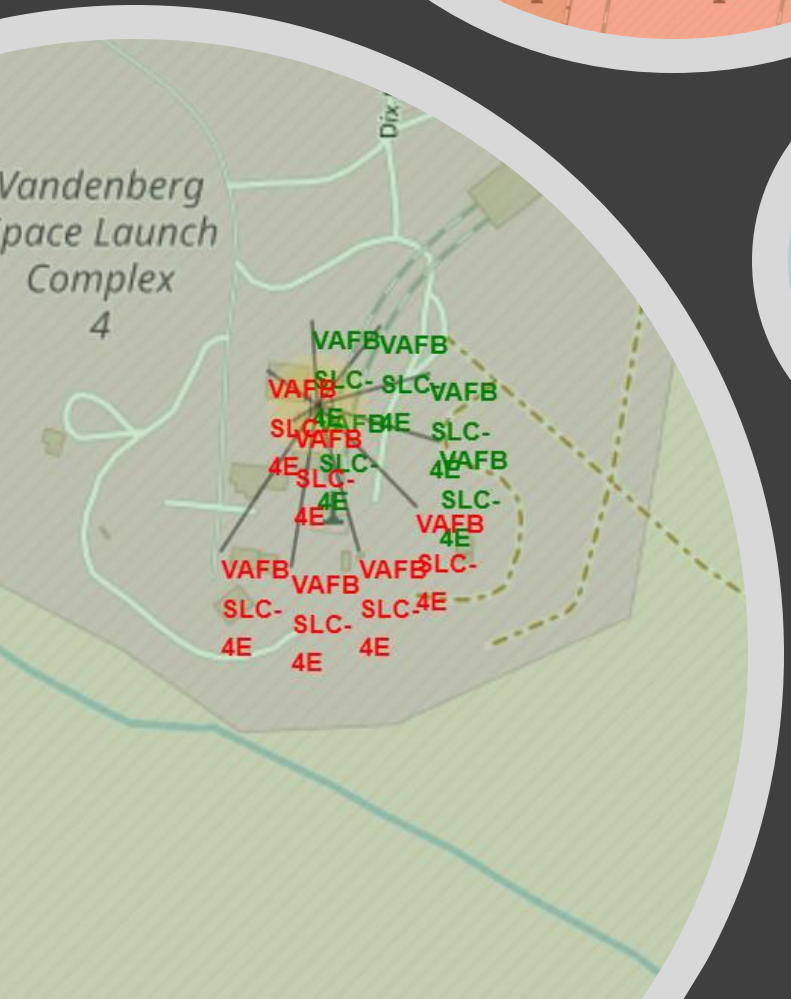
Launch Sites Proximities Analysis

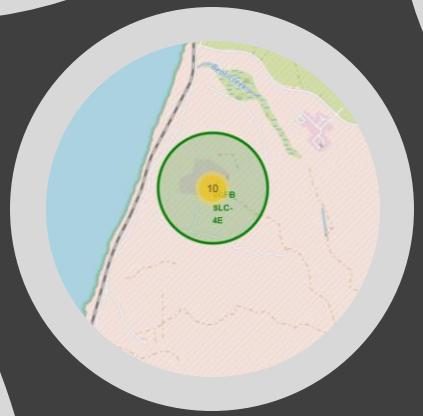
SpaceX launch locations



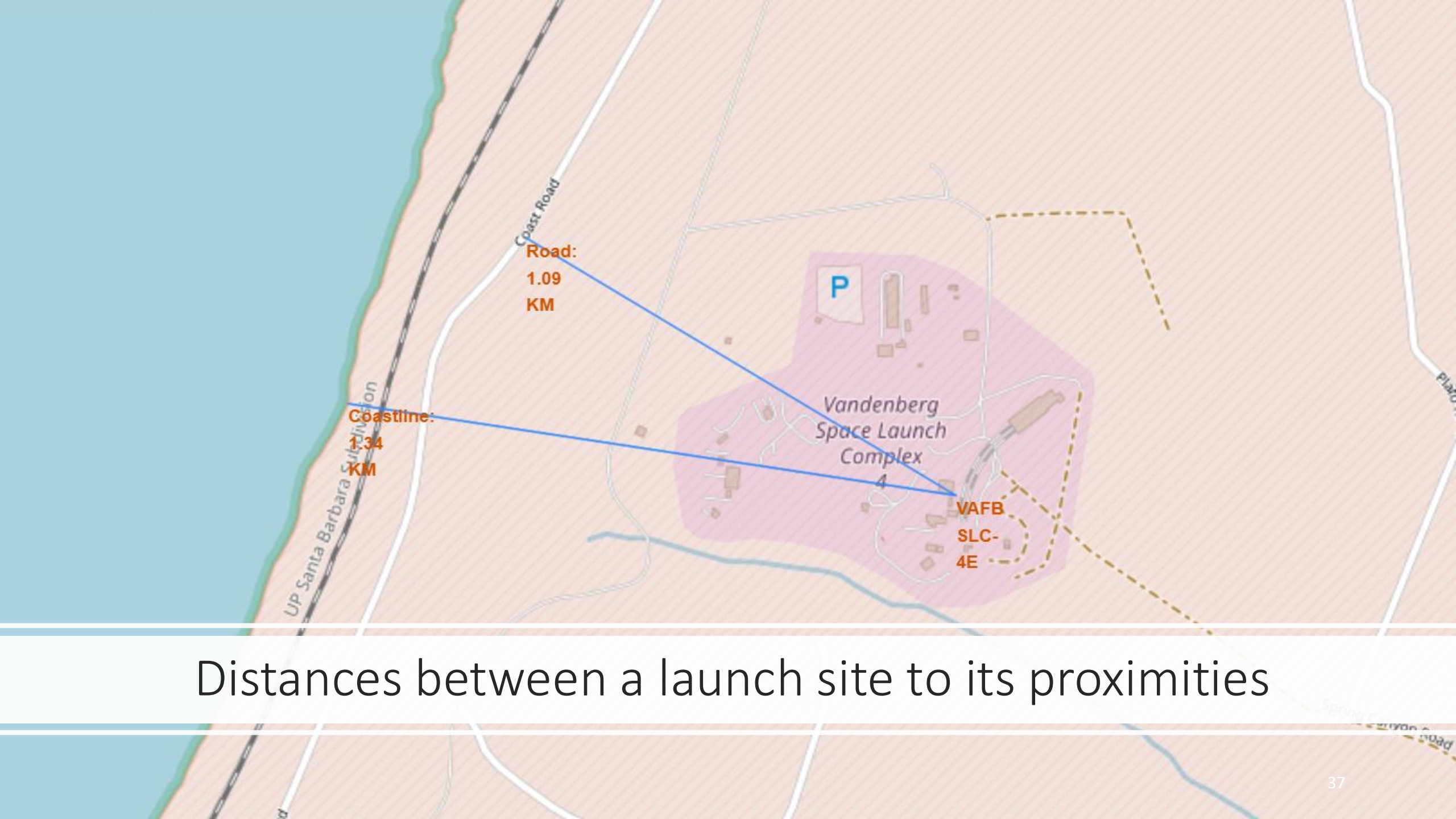




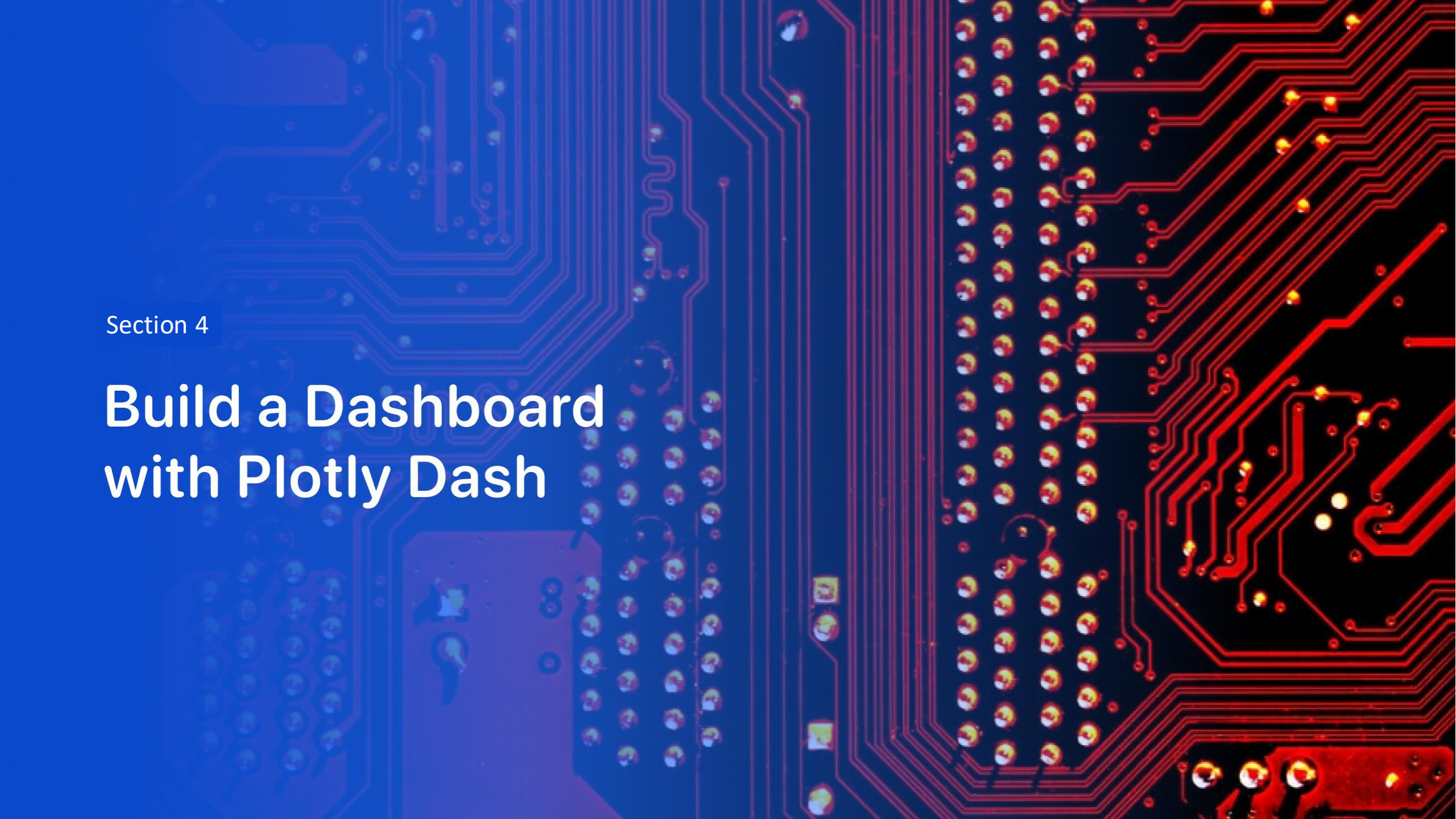


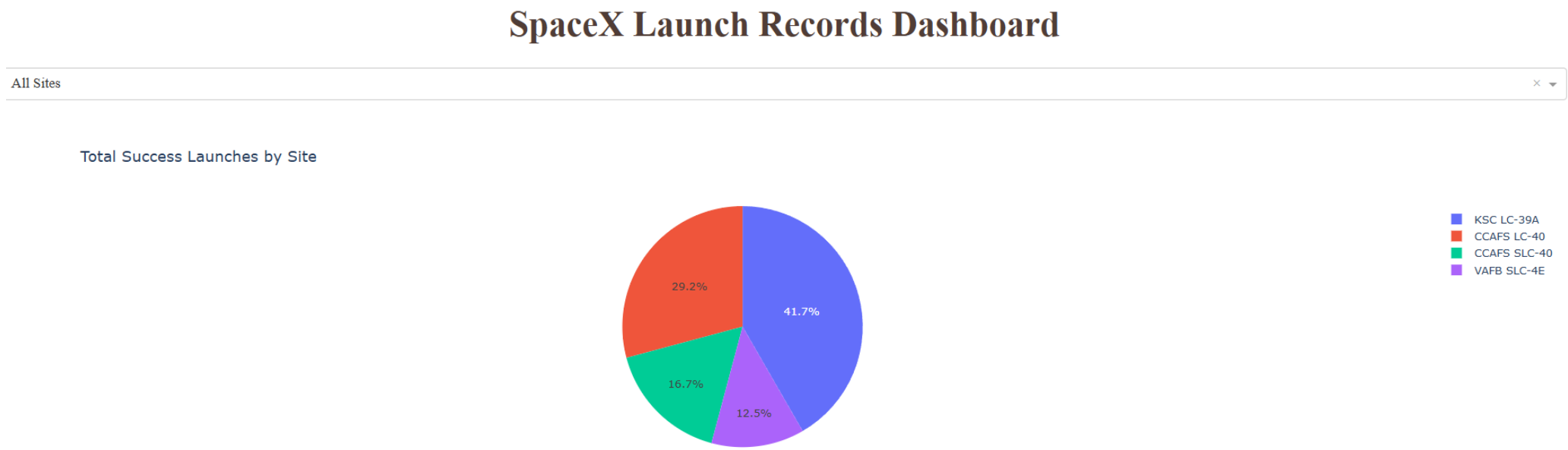


Success/failed launches for each site



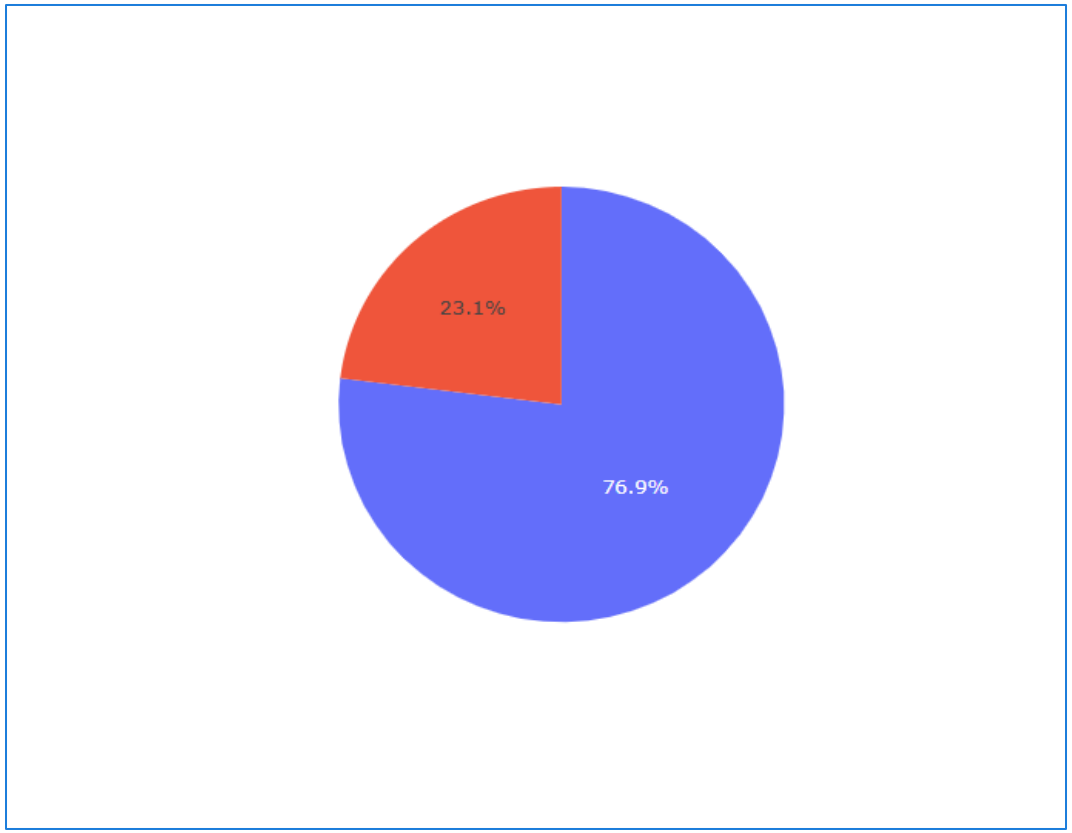
Distances between a launch site to its proximities


Section 4
Build a Dashboard with Plotly Dash



SpaceX Launch Records Dashboard: Pie chart

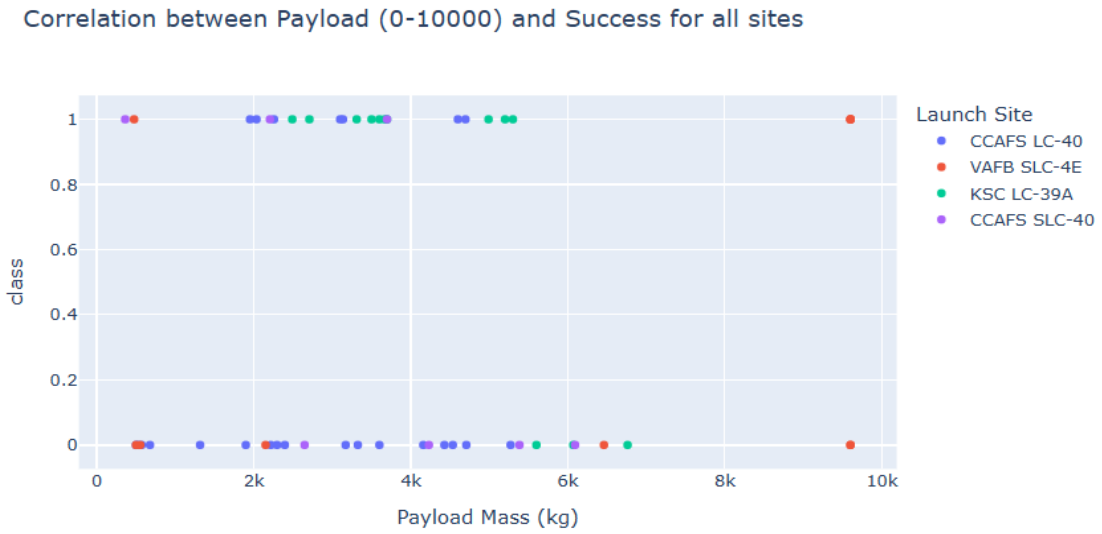
- Total successful launches distribution:
- KSC LC-39A has the highest percentage of launches (41.67%) and VAFB SLC-4E the lowest (12.50%)

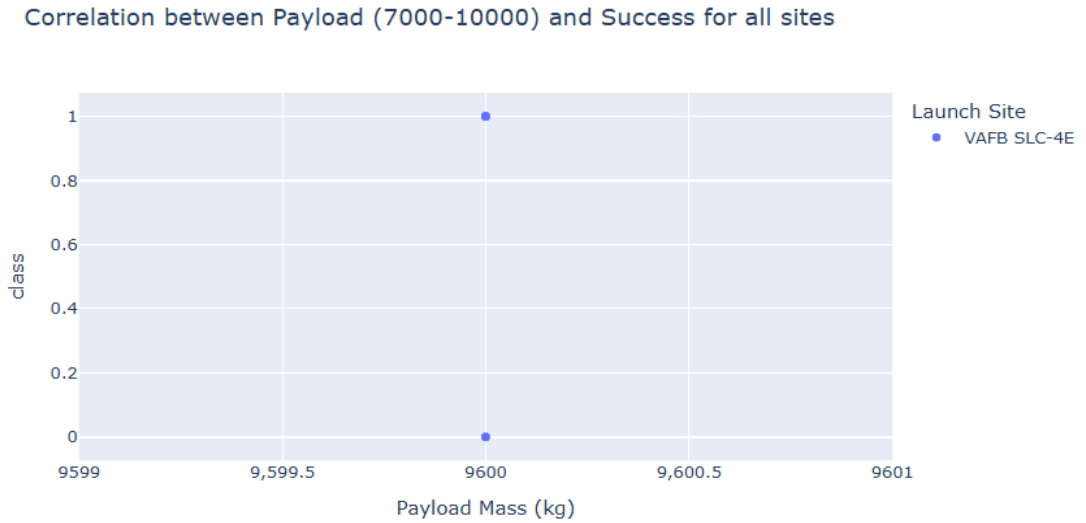


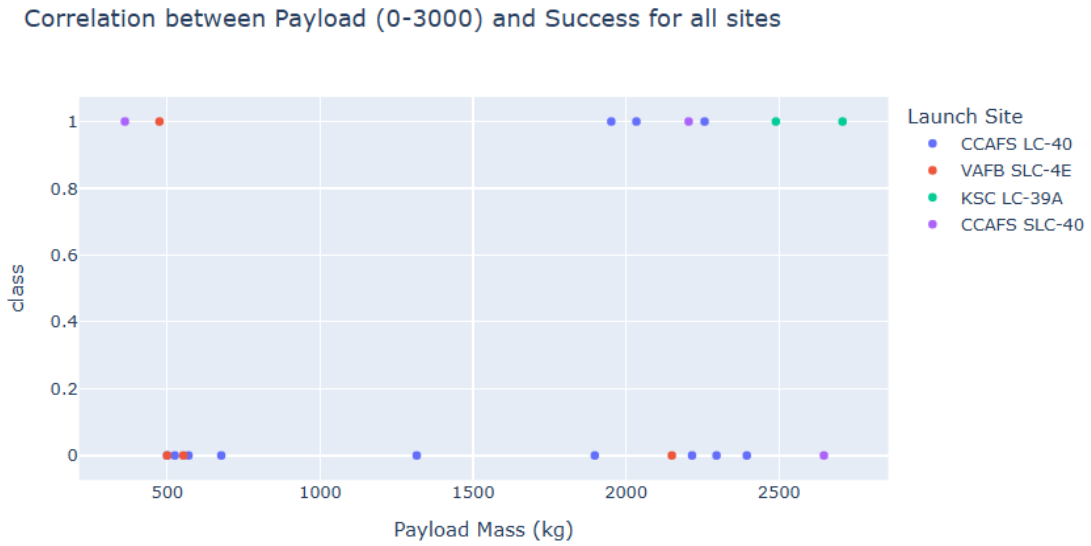
Total successful launches for site KSC LC-39A

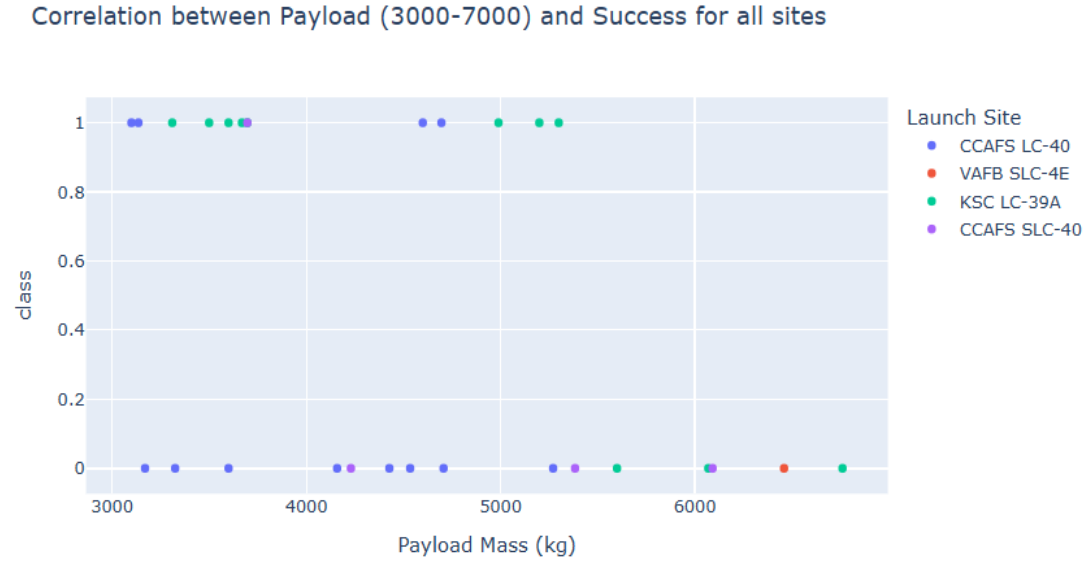
- Successful (77 %) vs failed launches (23%)

Correlation between different Payloads Success for all sites





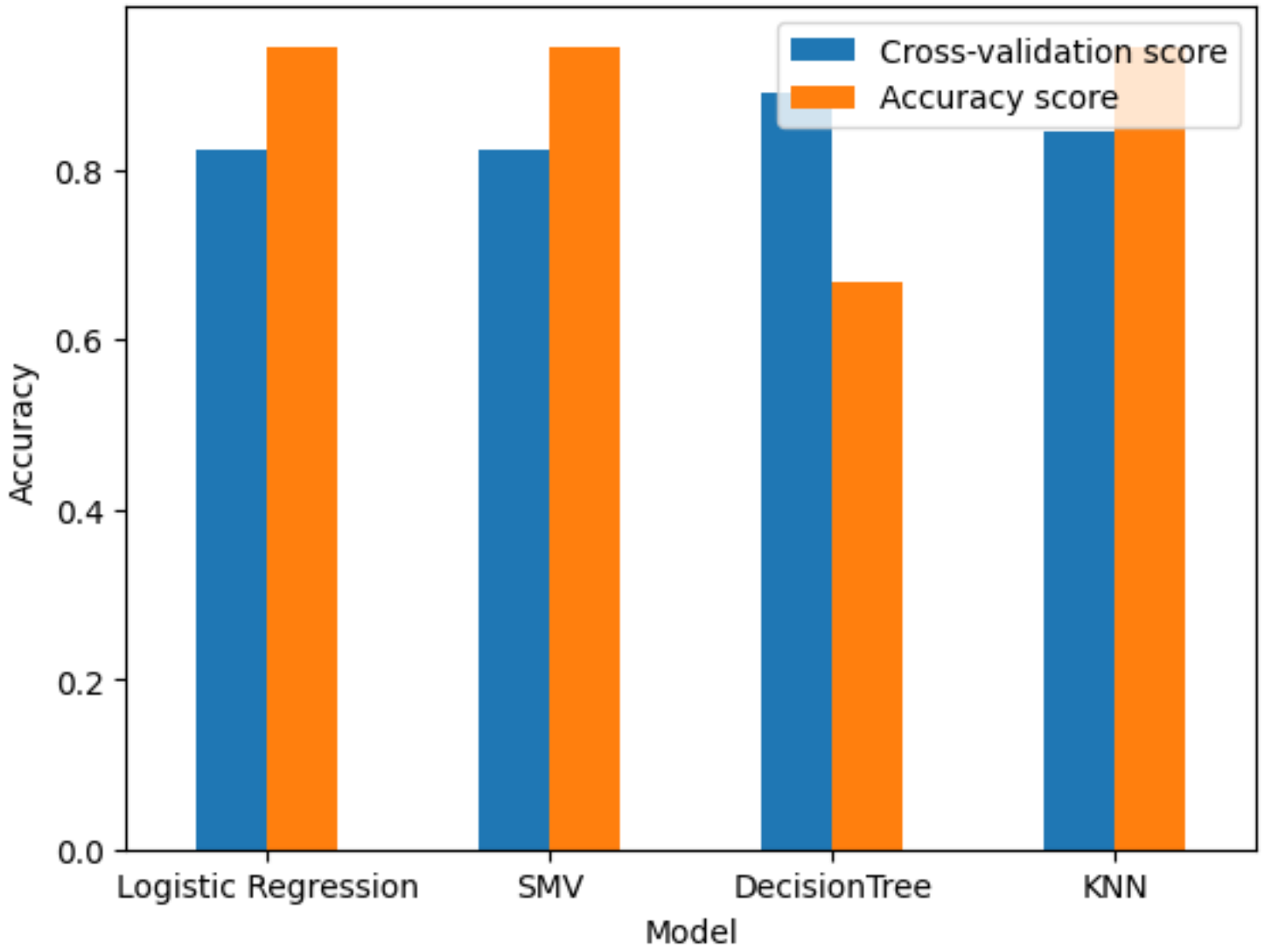




Section 5

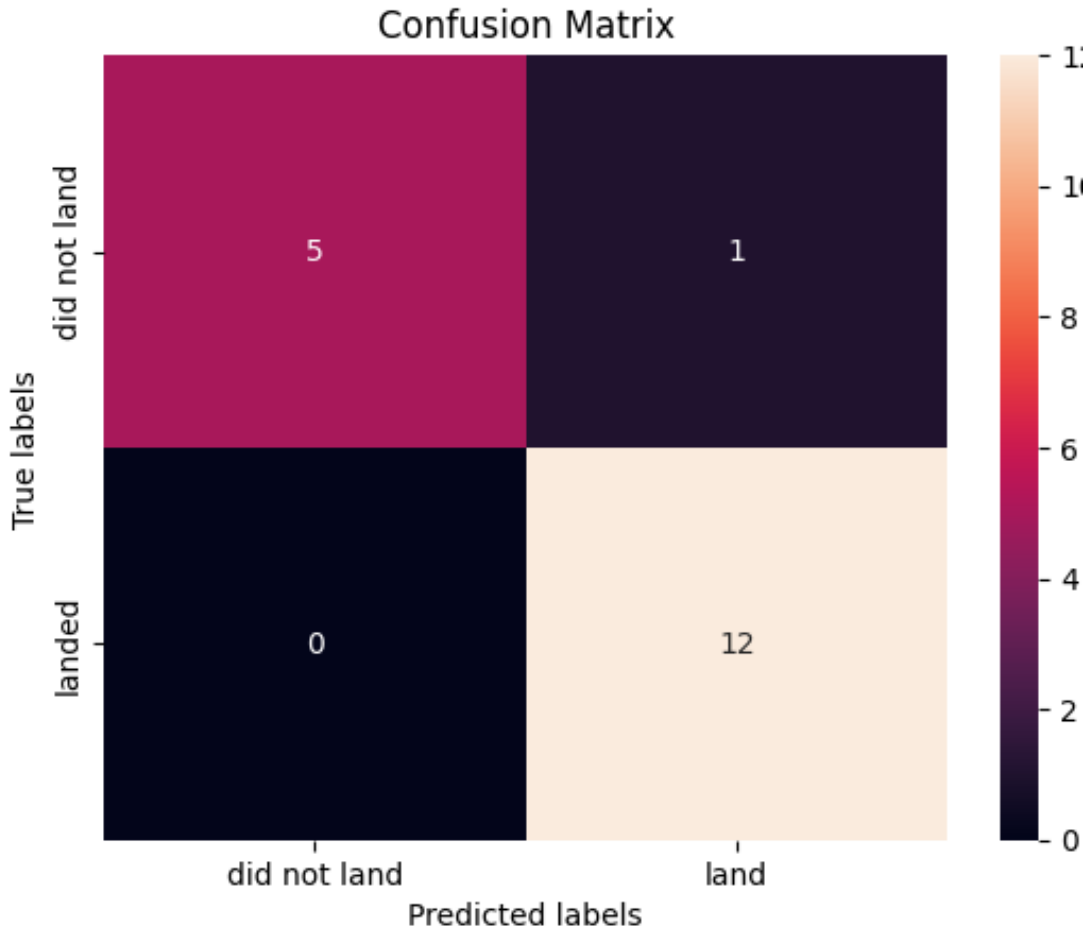
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- The best model(s) performed extremely well:
  - 12 successful landings were correctly classified (True positives), as well as 5 failed landings (True negatives)
  - Only 1 landing was incorrectly predicted as landed (True positive), and none as not landed.



Conclusions

- Launch success has continuously improved since 2013, with some setbacks after 2017, but still high.
- ES-L1, GEO, HEO, and SSO have higher success rates.
- KSC LC-39A has the highest percentage of launches.
- Higher payload masses tend to have a higher success rate (more data is needed to explain why).
- With the current data, **Logistic Regression, SMV, and KNN** perform equally well.

Appendix

Thank you!