Winning Space Race with Data Science

<Goh Qiu Le>
<19/05/23>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis (EDA) with Data Visualisation
  - Exploratory Data Analysis with SQL
  - Interactive Map with Folium
  - Dashboard with Plotly Dash
  - Machine Learning – predictive analysis
- Summary of all results
  - EDA results
  - Interactive maps and dashboard
  - Machine Learning prediction

Introduction

- Project background and context
  - In a bid to compete with SpaceX, Space Y wishes to determine the price of each launch, and also to determine if SpaceX will reuse the first stage.

- Problems you want to find answers
  - How do variables, such as payload mass and launch sites, interplay and affect the success of first stage landing?
  - What are the most desirable conditions that allow Space X to achieve the most optimal landing success rate?

Section 1

Methodology

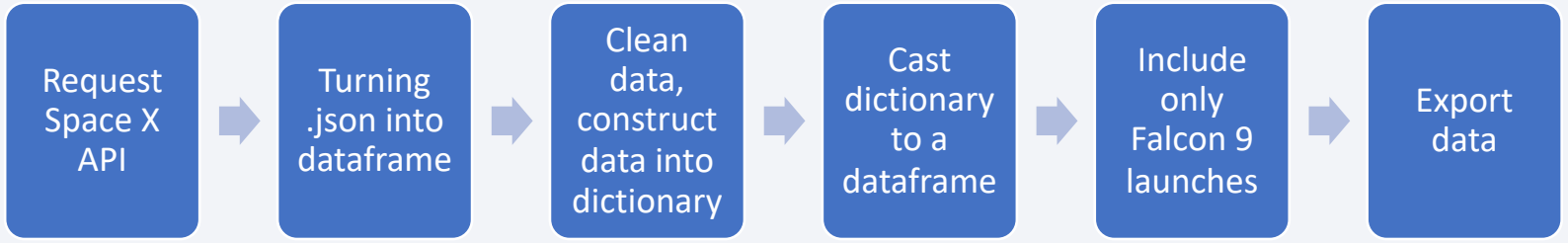
Methodology

Executive Summary

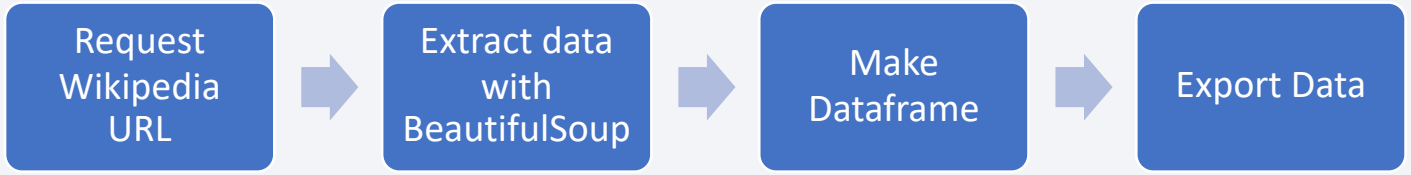
- Data collection methodology:
  - SpaceX Rest API
  - WebScraping
- Perform data wrangling
  - Classified landings as successful or otherwise via one hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was obtained from Rest SpaceX API and webscrapping Wikipedia.
  - SpaceX API - https://api.spacexdata.com/v4/payloads/
  - Wikipedia - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



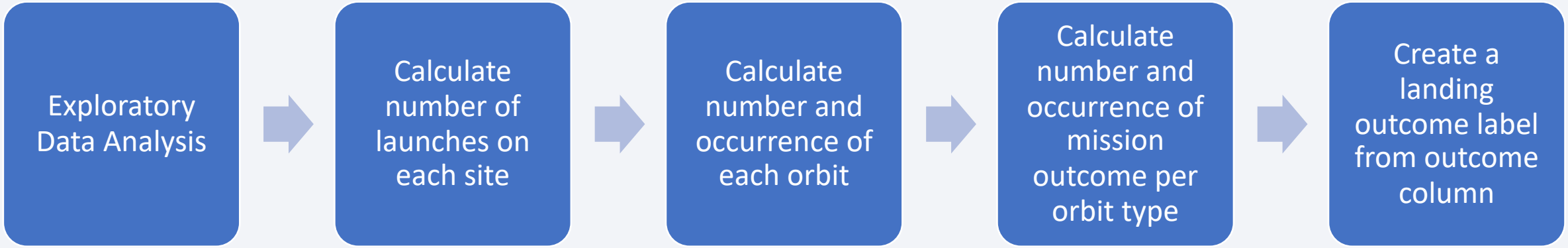
https://github.com/gohqiule/IBM-Data-Science-Capstone-Project/blob/6df67aa460388953b661078581b8299da21a3ad2/Data%20Collection.ipynb



https://github.com/gohqiule/IBM-Data-Science-Capstone-Project/blob/6df67aa460388953b661078581b8299da21a3ad2/Data%270Webscraping.ipynb

Data Wrangling

- In the data set, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.
- Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.



EDA with Data Visualization

- Scatterplots and barplots were used to visualize the relationship between pair of features
    - Flight Number vs Payload Mass, Flight Number vs Launch Site, Launch Site vs Payload Mass, Orbit Type vs Flight Number, Payload and Orbit Type
- Line charts were used to visualise the launch success yearly trend

https://github.com/gohqiule/IBM-Data-Science-Capstone-Project/blob/6df67aa460388953b661078581b8299da21a3ad2/Data%20Visualisation.ipynb

EDA with SQL

- SQL queries performed:
  - Displaying names of 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 the average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass > 4000, < 6000kg
  - 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 failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking the count of landing outcomes (such as Failure (drone shiop) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium

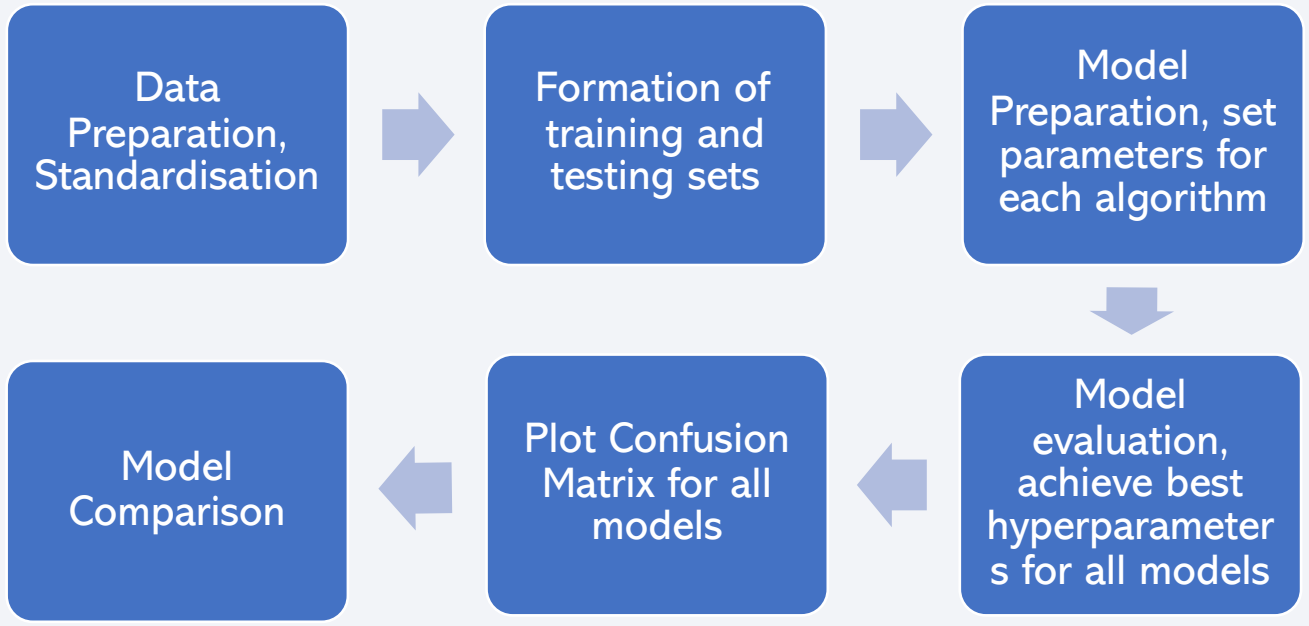
- Markers, circles, lines and marker clusters were employed
- These allow us to better understand the rationale behind launch sites, and also to visualise and map successful landings to their relative locations.

https://github.com/gohqiule/IBM-Data-Science-Capstone-Project/blob/c82c76119325db8516f34cc9ad96a42b623c6adb/Launch%20Sites%20Locations.ipynb

Build a Dashboard with Plotly Dash

- Dashboard consists
  - Launch Sites Dropdown List
  - Pie Chart (Success Launches)
  - Range Slider (Payload Mass)
  - Scatter Chart (Payload Mass vs Success Rate for different Booster versions)
- Allows quick analysis of correlation between variables

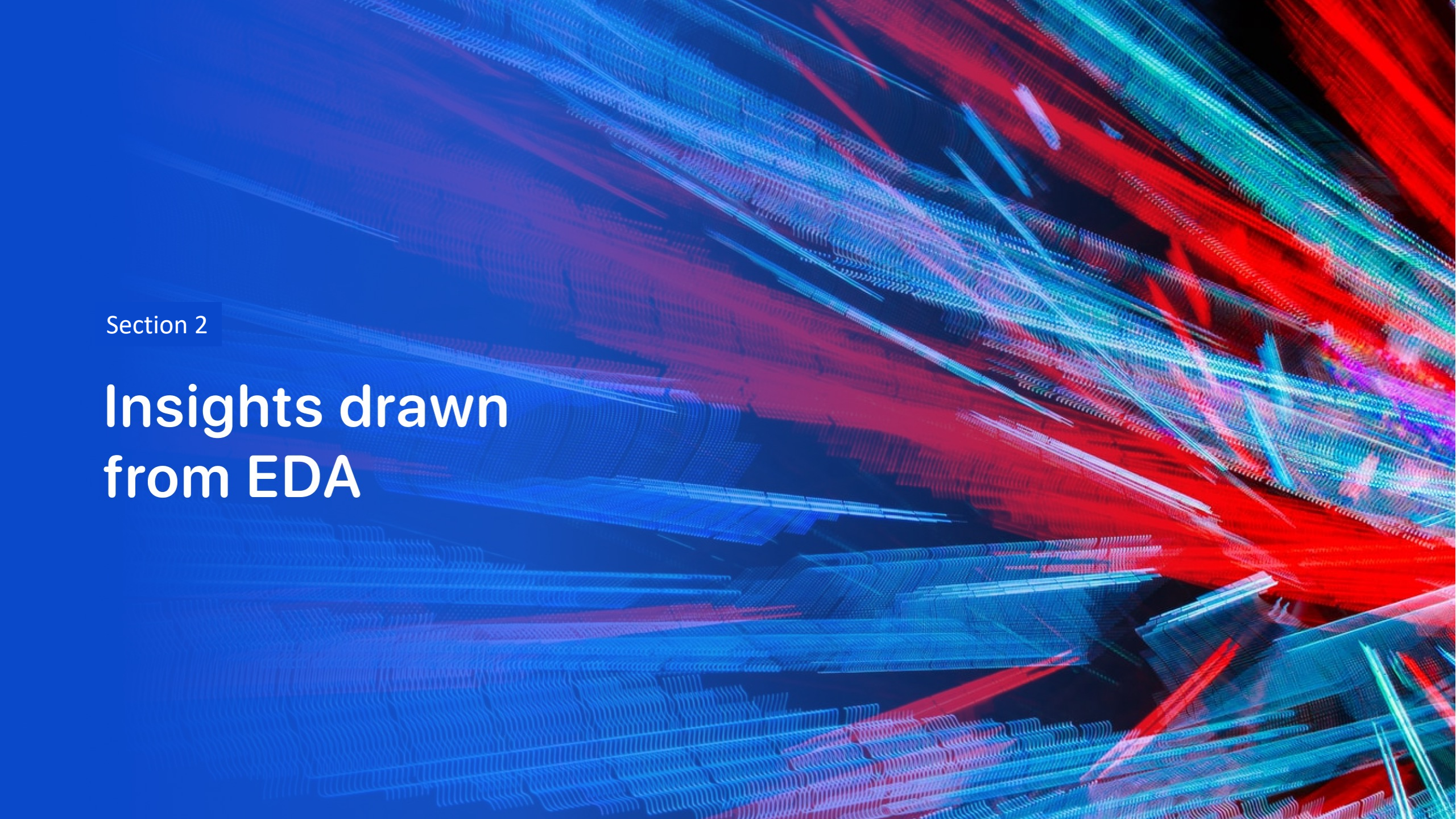
Predictive Analysis (Classification)



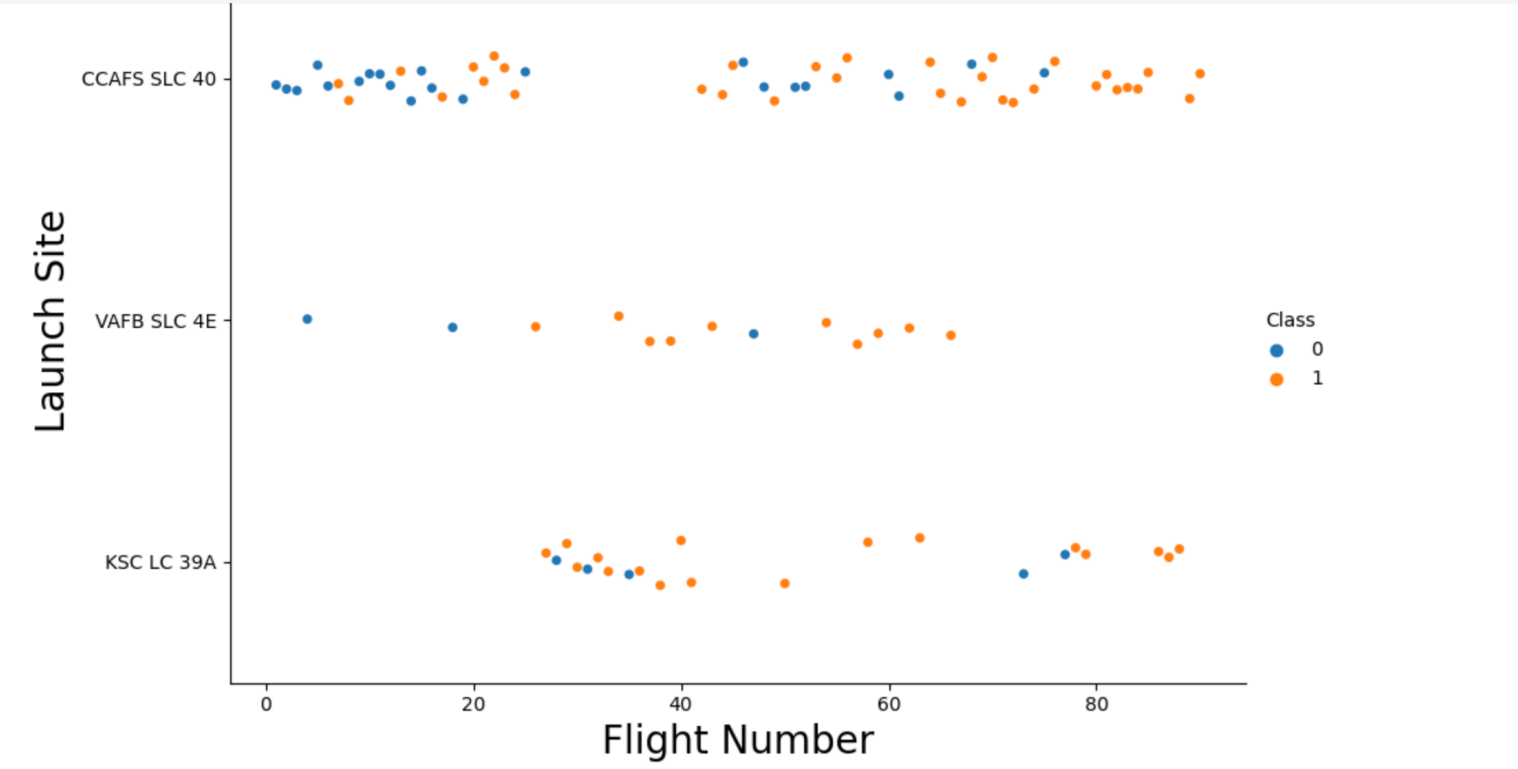
https://github.com/gohqiule/IBM-Data-Science-Capstone-Project/blob/c82c76119325db8516f34cc9ad96a42b623c6adb/Machine%20Learning.ipynb

Results

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

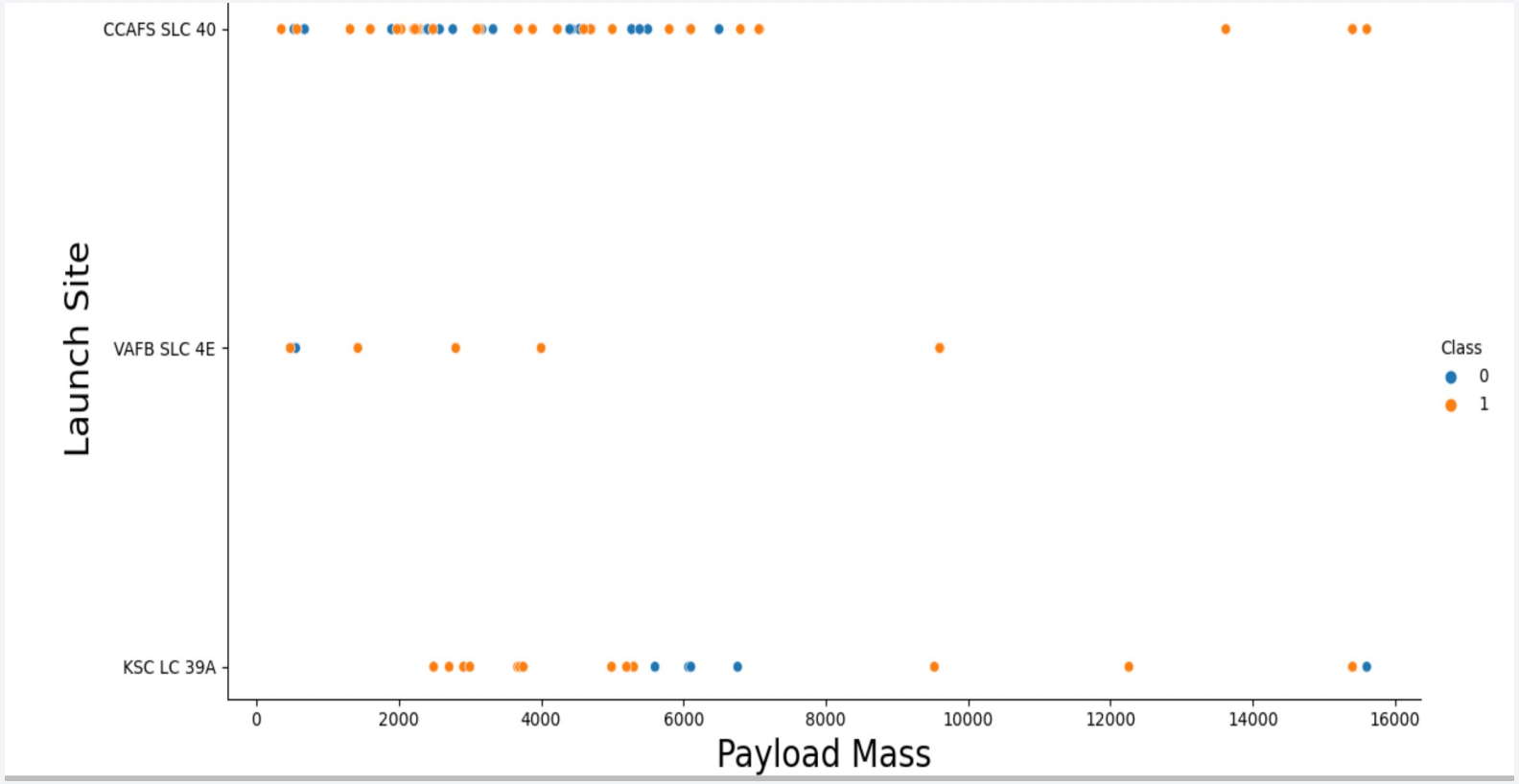

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



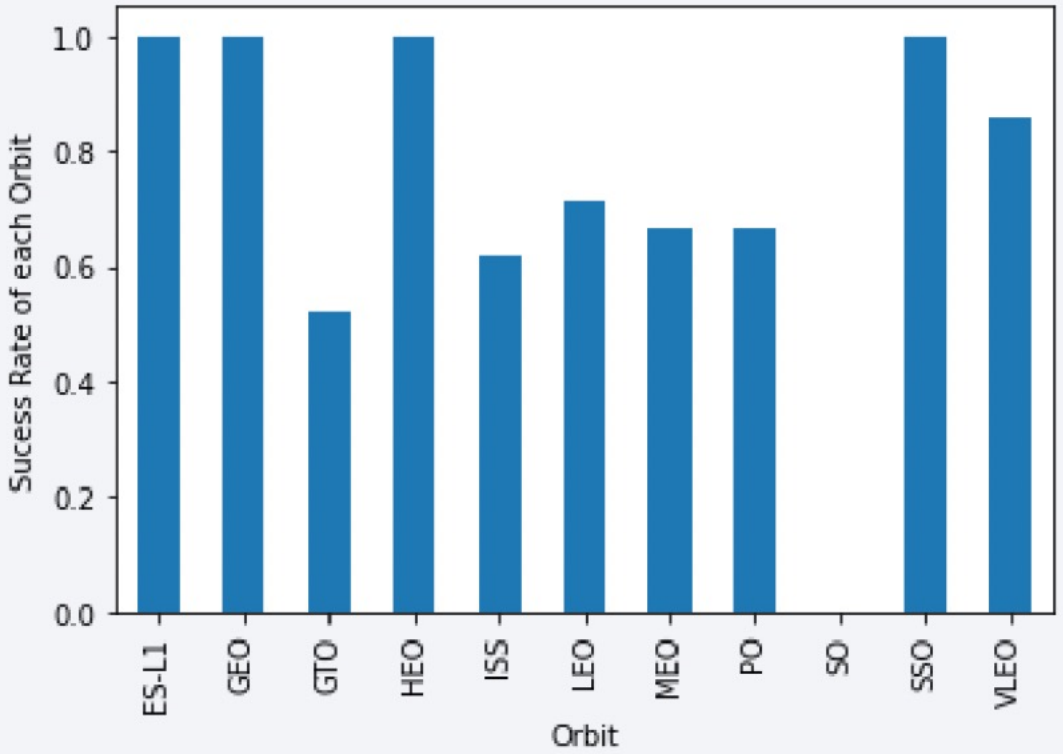
Success rate increases over time across all launch sites.

Payload vs. Launch Site



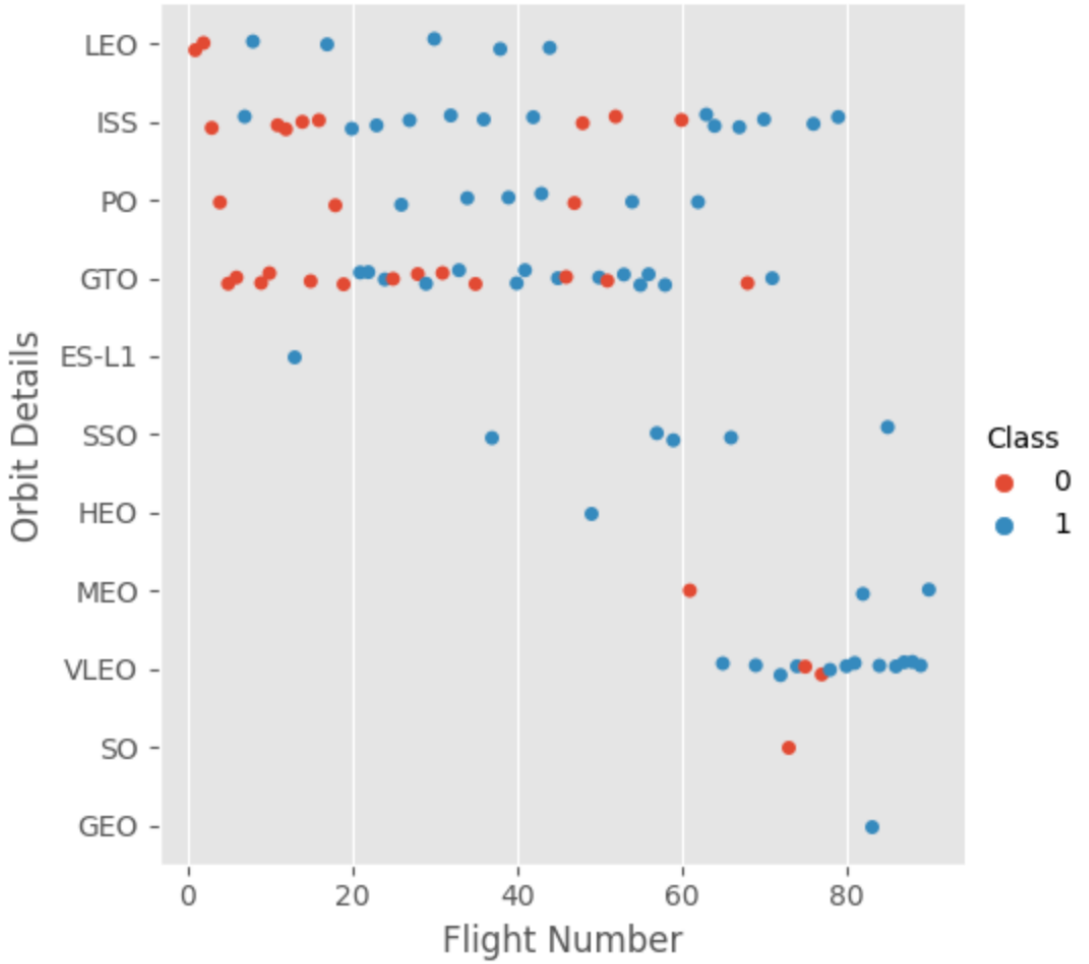
In general, the possibility of success increases with payload mass.

Success Rate vs. Orbit Type



ES-L1, GEO, HEO, SSO have the highest success rates.

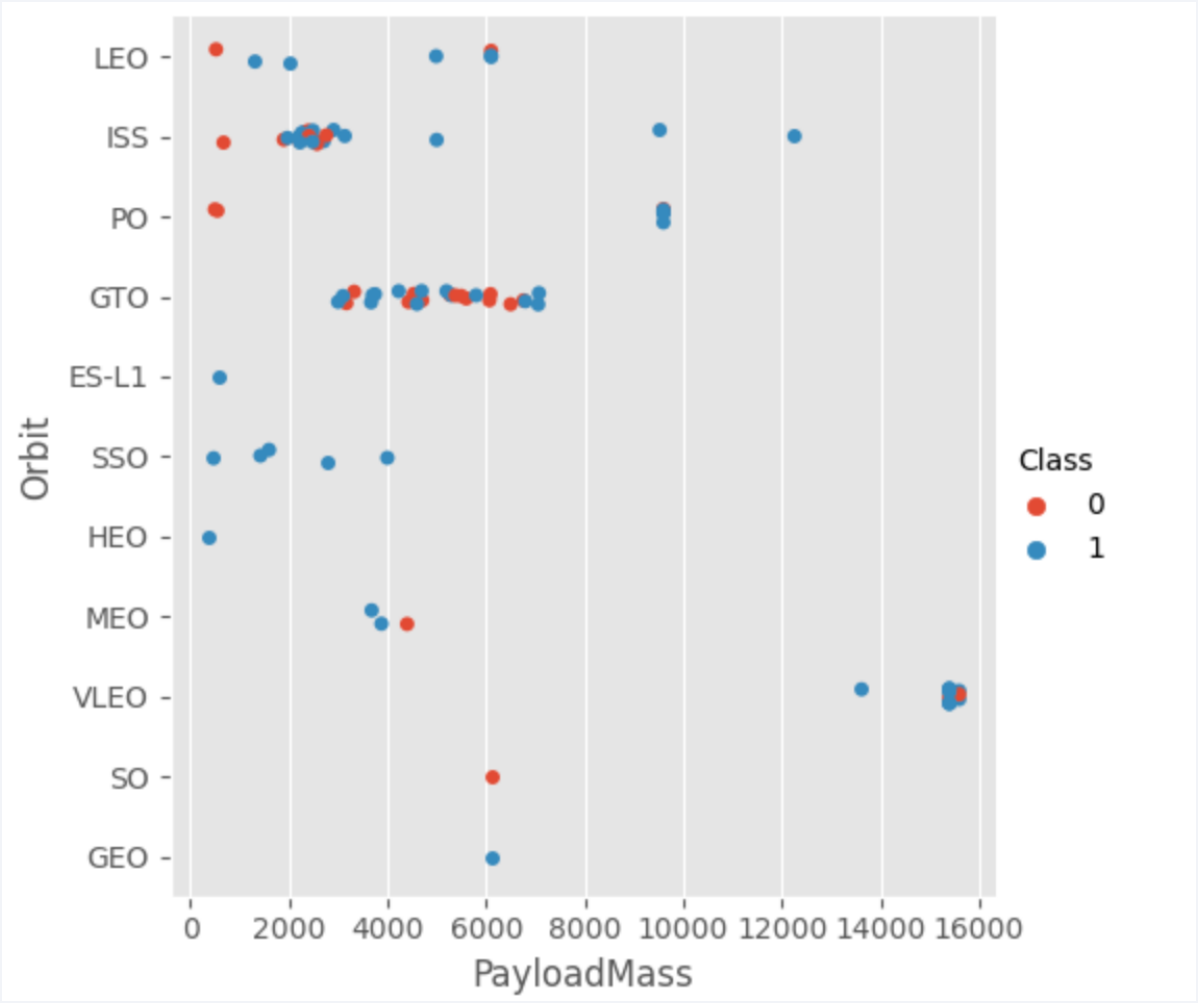
Flight Number vs. Orbit Type



It appears that for LEO, ISS, and perhaps PO, the success rate increases with the number of flights.

No such correlation is observed for GTO.

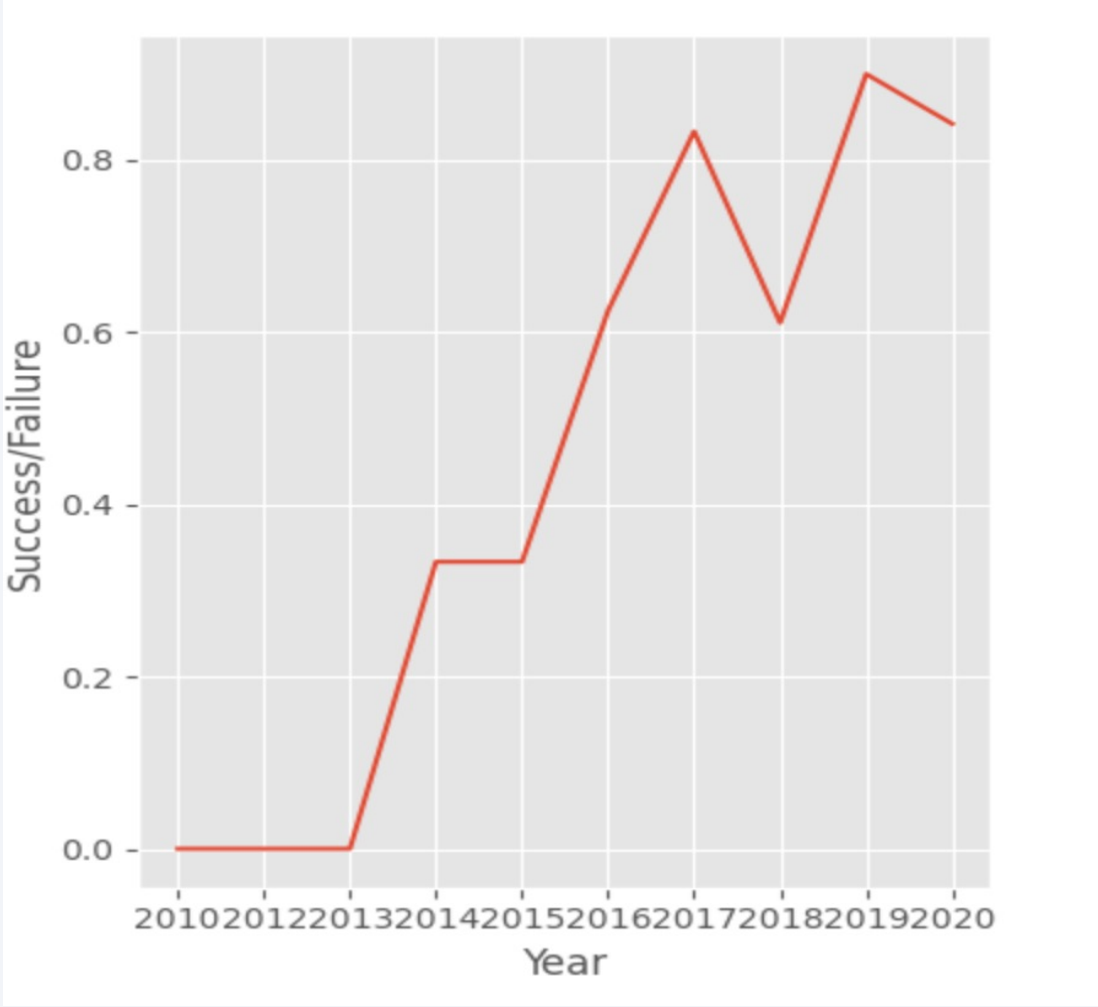
Payload vs. Orbit Type



The correlation between success rate and payload mass differs across orbits.

For LEO, PO and ISS it could be argued that greater payload mass is correlated with higher success rate.

Launch Success Yearly Trend



The success-failure ratio generally trends upwards with time.

All Launch Site Names

Display the names of the unique launch sites in the space mission

[7]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

[7]:

Launch_Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with '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	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Task 3

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as PM_KG_TOTAL, Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)'
```

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

PM_KG_TOTAL	Customer
45596.0	NASA (CRS)

Average Payload Mass by F9 v1.1

Task 4

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

```
0]: %sql SELECT AVG(PAYLOAD_MASS__KG_) as PM_KG_AVG FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

0]:

PM_KG_AVG
2534.6666666666665

First Successful Ground Landing Date

Task 5 ¶

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

Hint:Use min function

[16]:
```
%sql select min(DATE) from SPACEXTBL;
```

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

[16]:

min(DATE)
01/06/2014

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

[21]:
```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Mission_Outcome = 'Success' AND PAYLOAD_MASS__KG_ between 4001 and 5999
```

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

[21]:

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
2]: %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) as Total FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

2]:

Mission_Outcome	Total
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
[39]: %sql SELECT "LANDING__OUTCOME", COUNT("LANDING__OUTCOME") AS Landing_Count FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010' and '20-03-2017' GROUP BY "MISSI
```

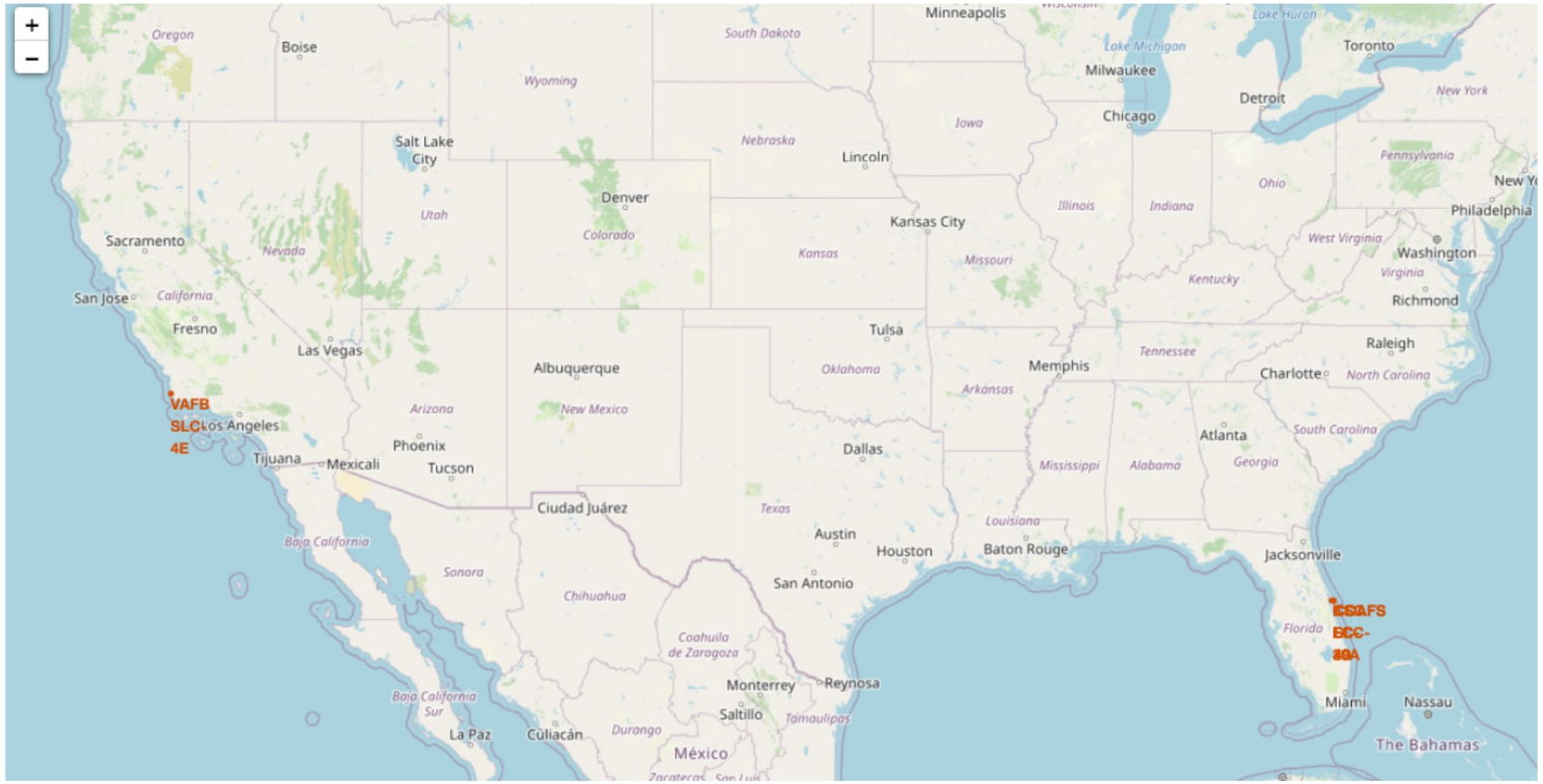
```
 * sqlite:///my_data1.db
Done.
```

[39]:

"LANDING__OUTCOME"	Landing_Count
LANDING__OUTCOME	56


Section 3
Launch Sites
Proximities Analysis

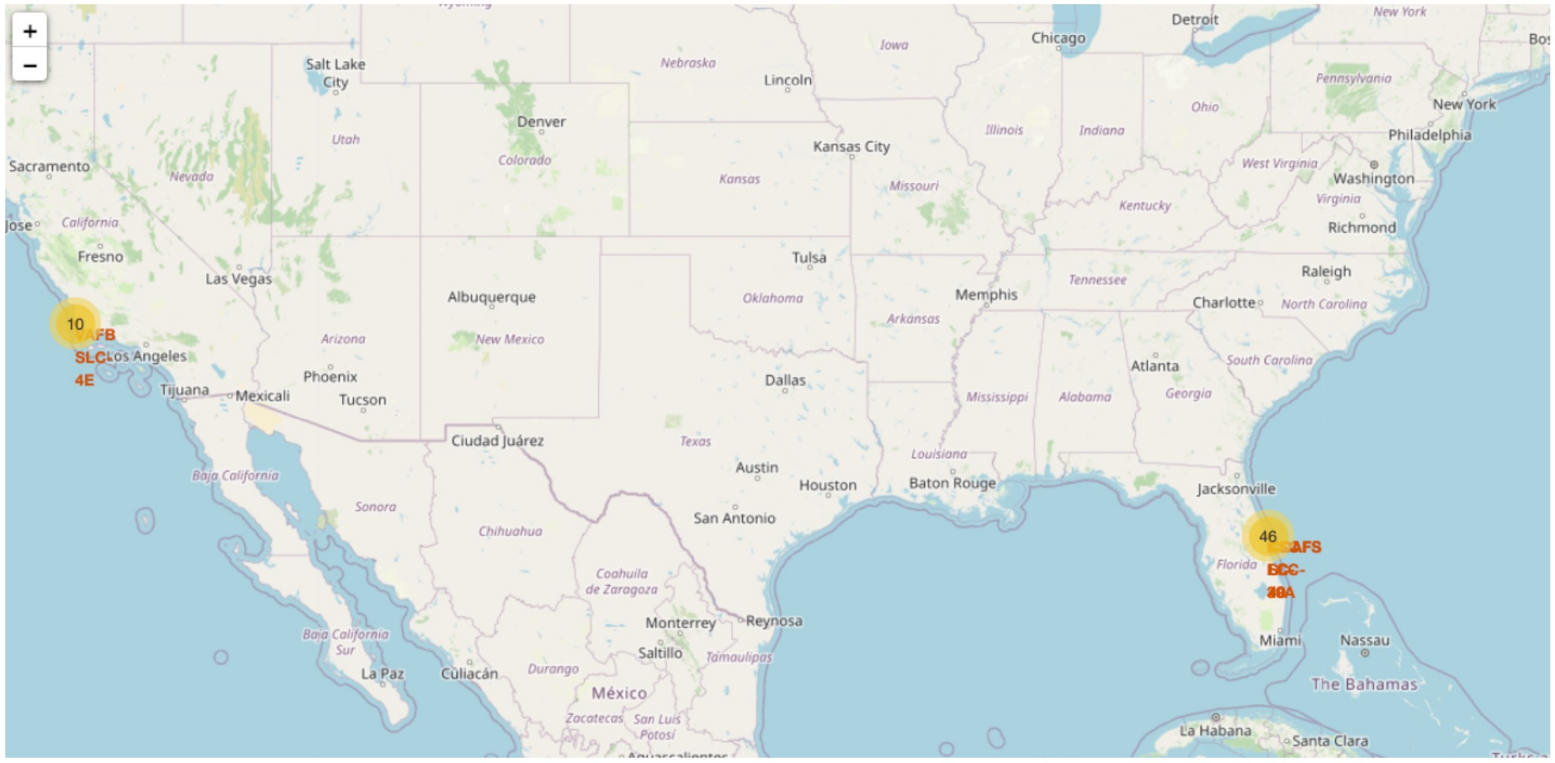
<Marked Launch Sites>



Most launch sites appear to be in proximity to the equator and in proximity to the coastal areas.

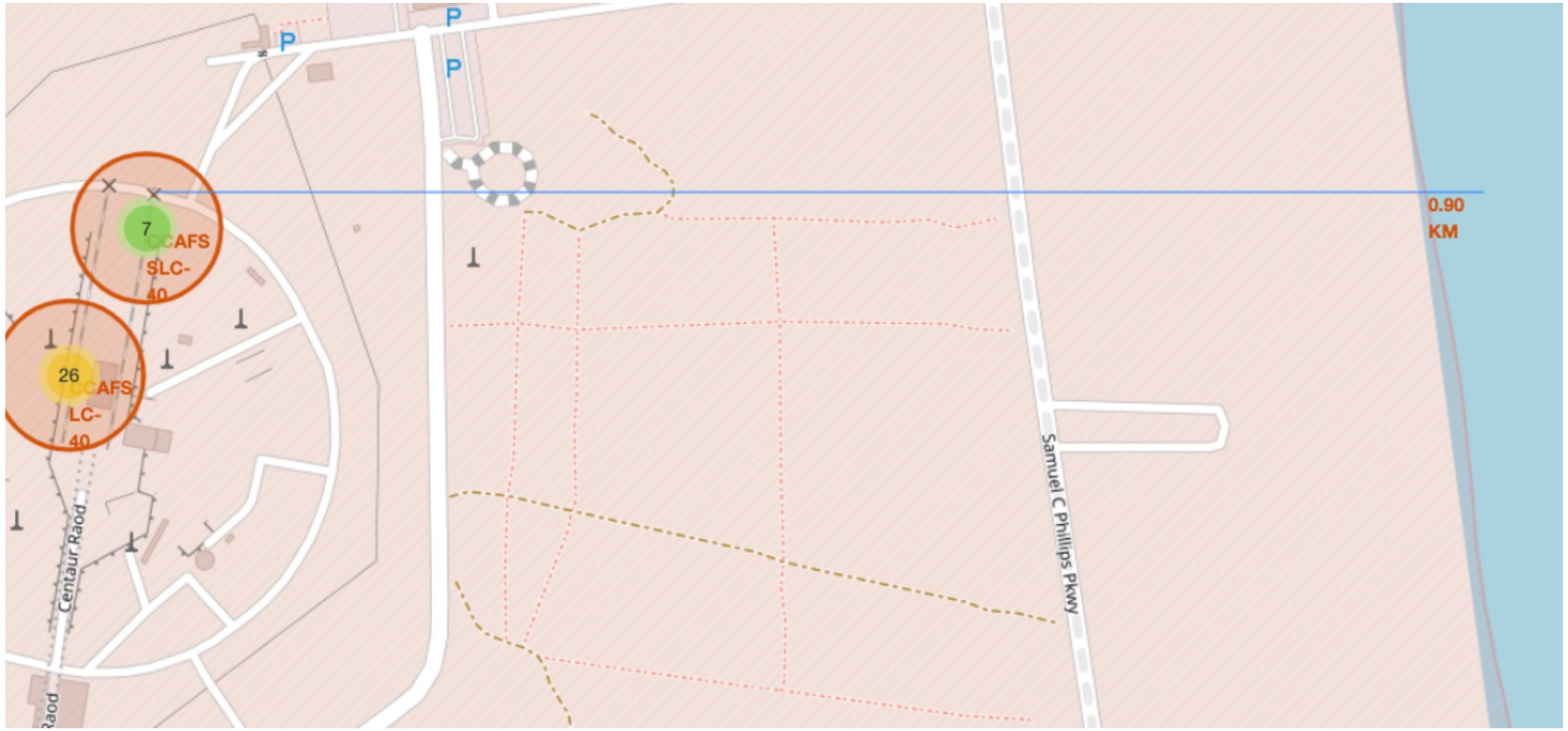
This may be explained by the fact that rockets launched from these sites can get an additional boost from Earth's substantial rotational speed, which then translates to cost savings from fuel and boosters.

<Success/Failure of each site>



The color-labeled markers in marker clusters allow us to to easily identify which launch sites have relatively high success rates.

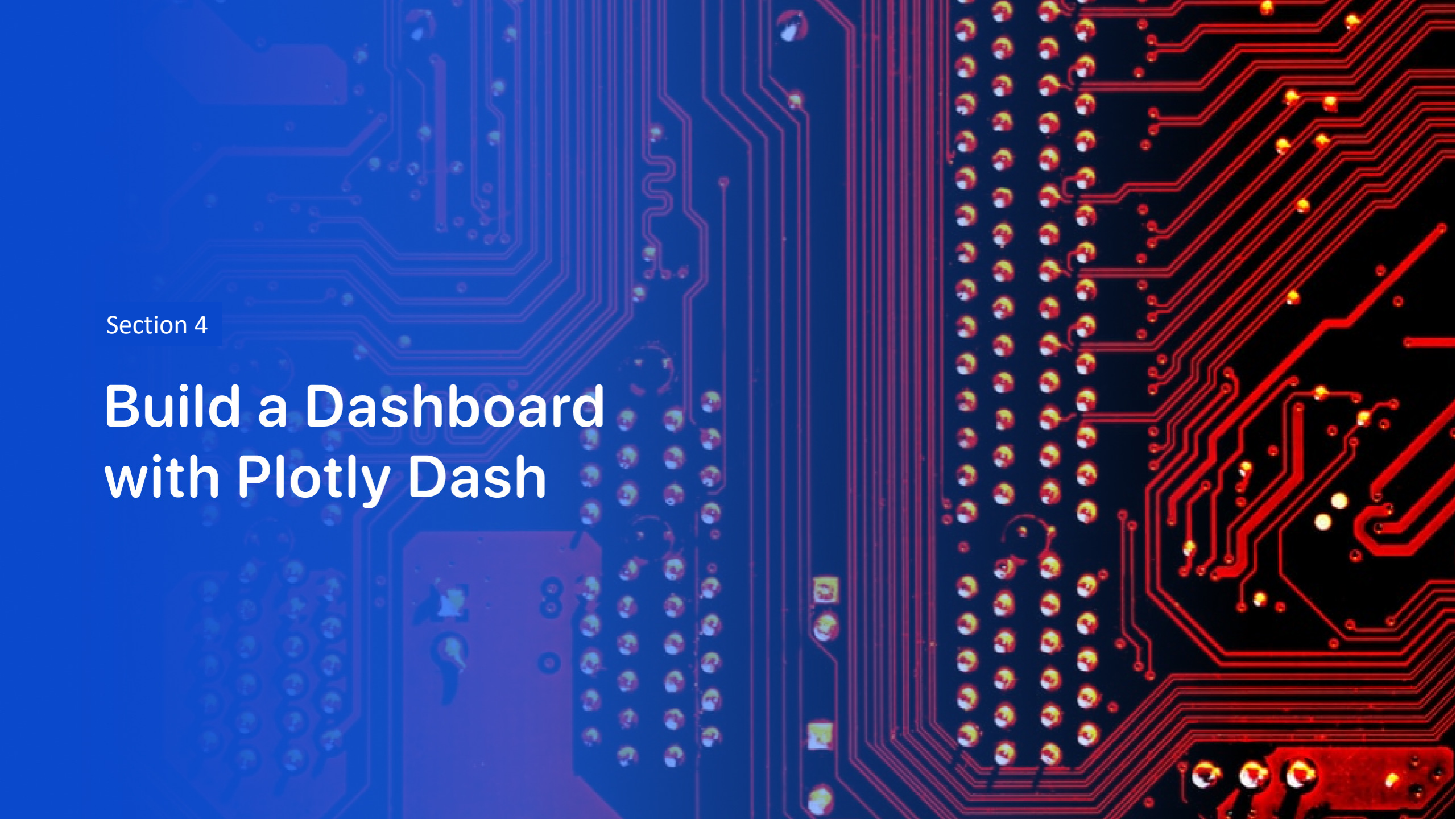
<Proximity of Launch sites to various structures>



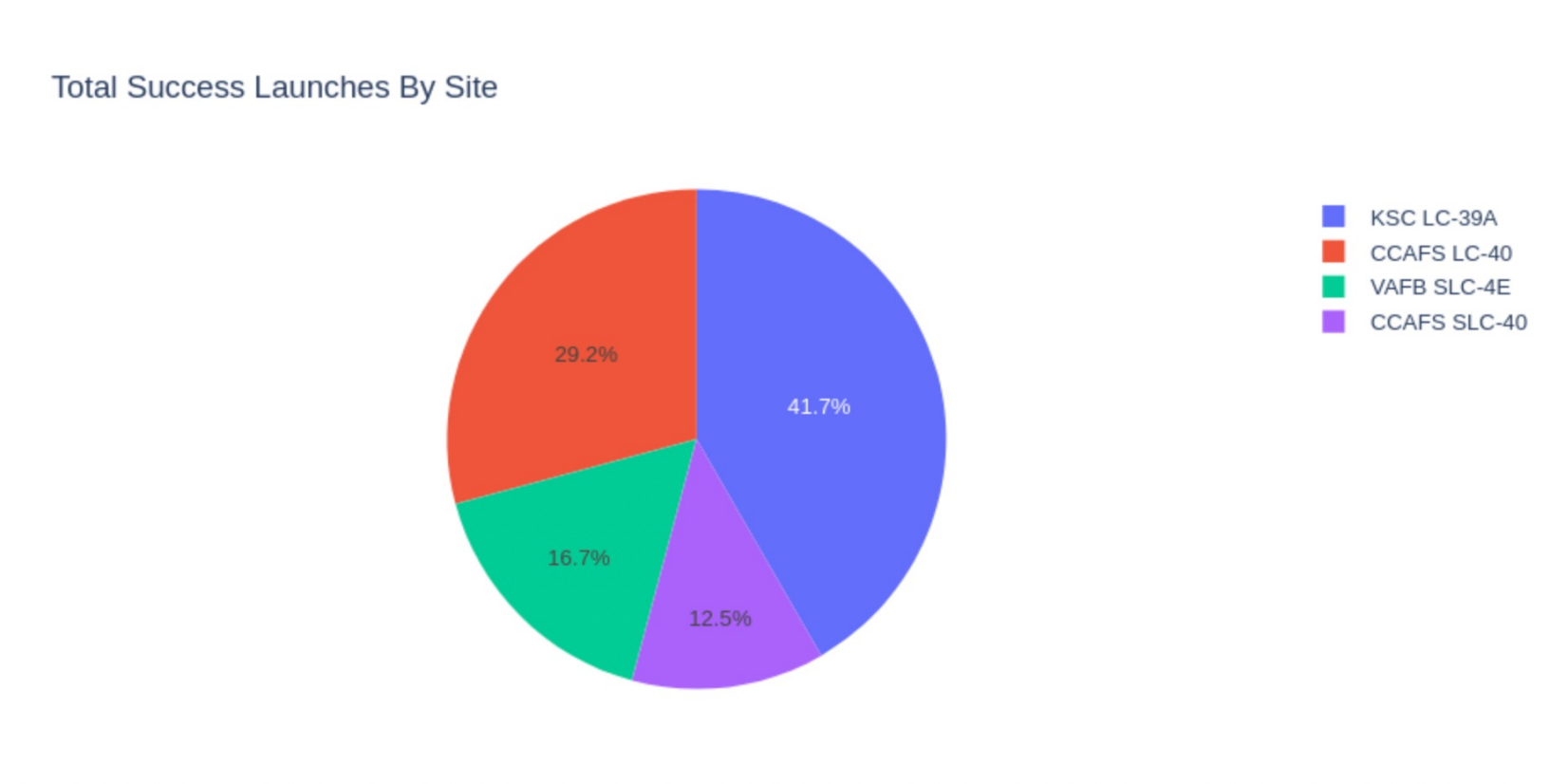
Launch sites are located close to equator for reasons previously mentinoed.

In close proximity to railways/highways to facilitate transport of heavy equipment.

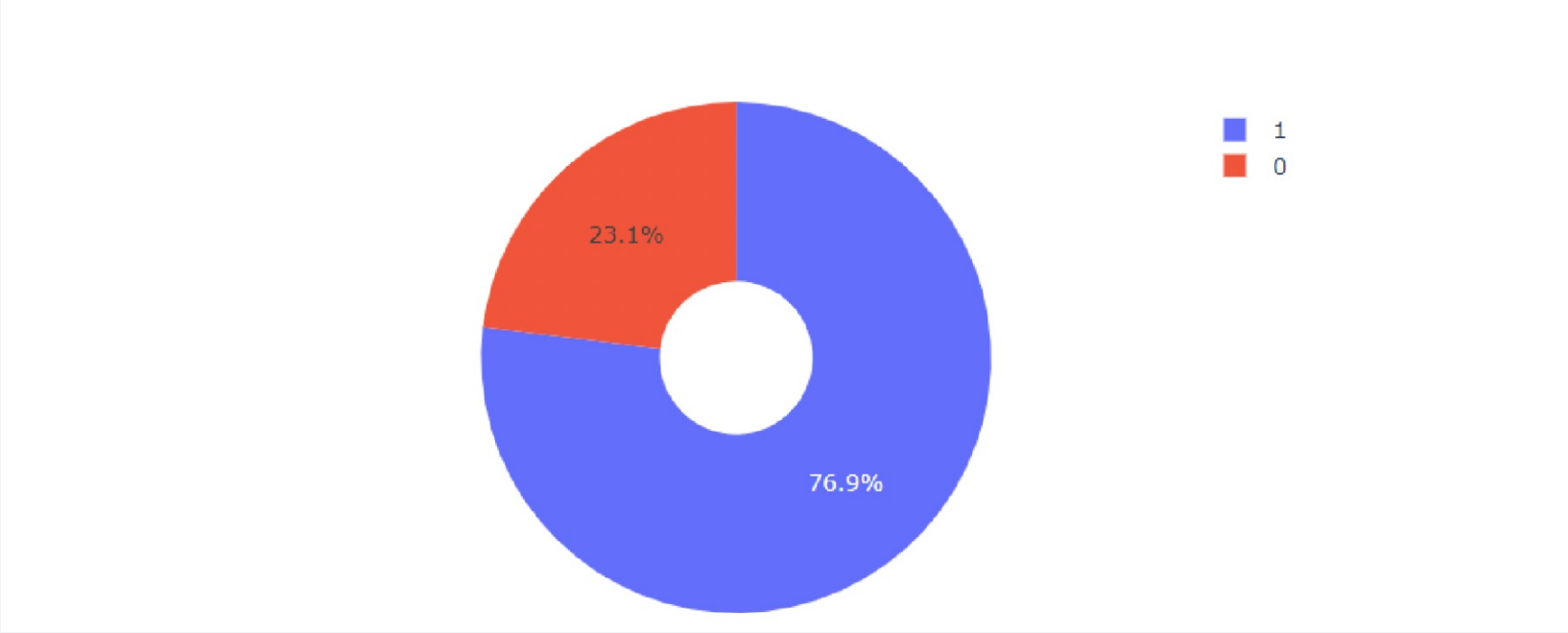
Likely situated further from cities to alleviate and mitigate risk to populations


Section 4
Build a Dashboard with Plotly Dash

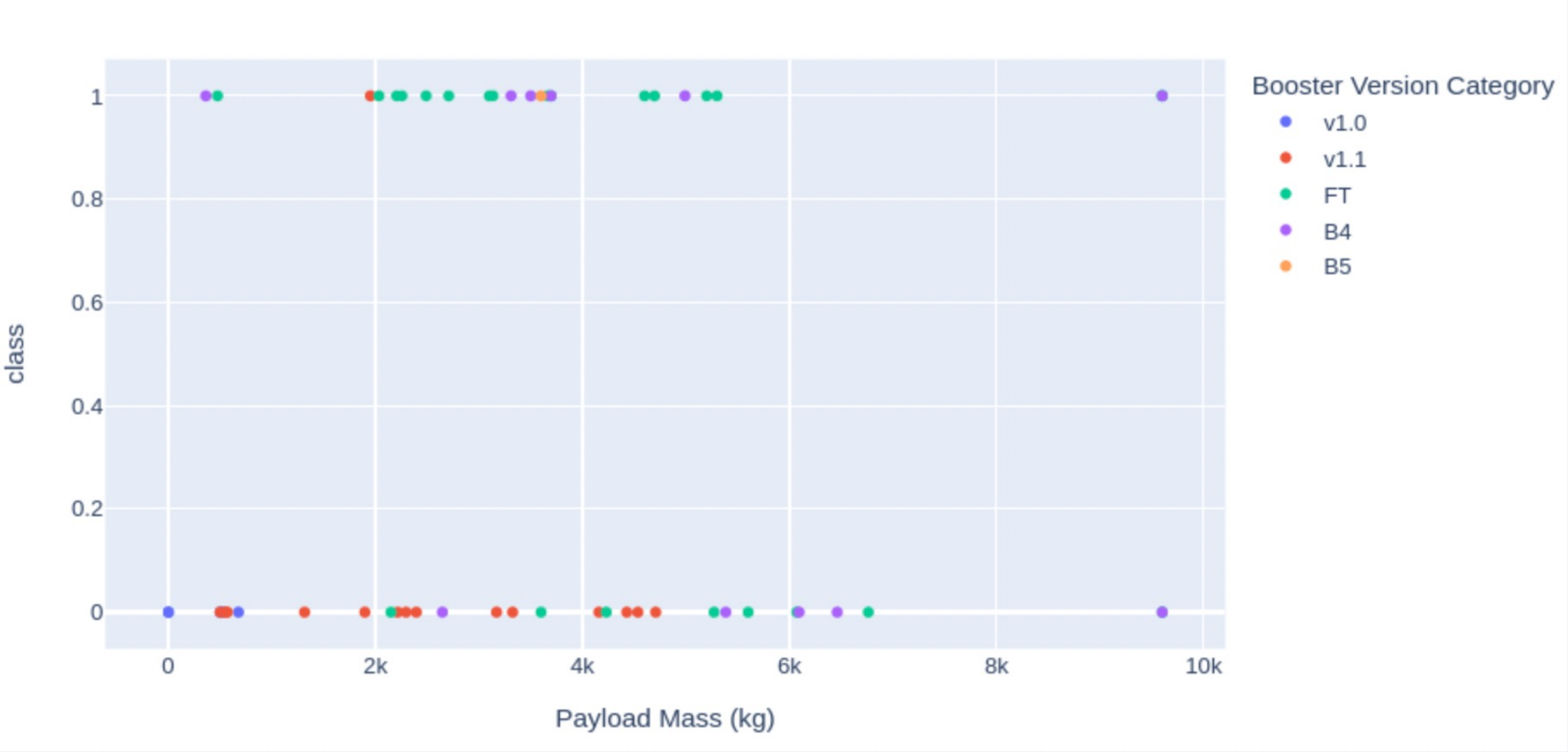
Dashboard – Pie Chart (success count for all sites)



Dashboard – Pie Chart (KSC LC-39A success ratio)



Dashboard- Payload vs Launch Outcome Scatterplot



Section 5

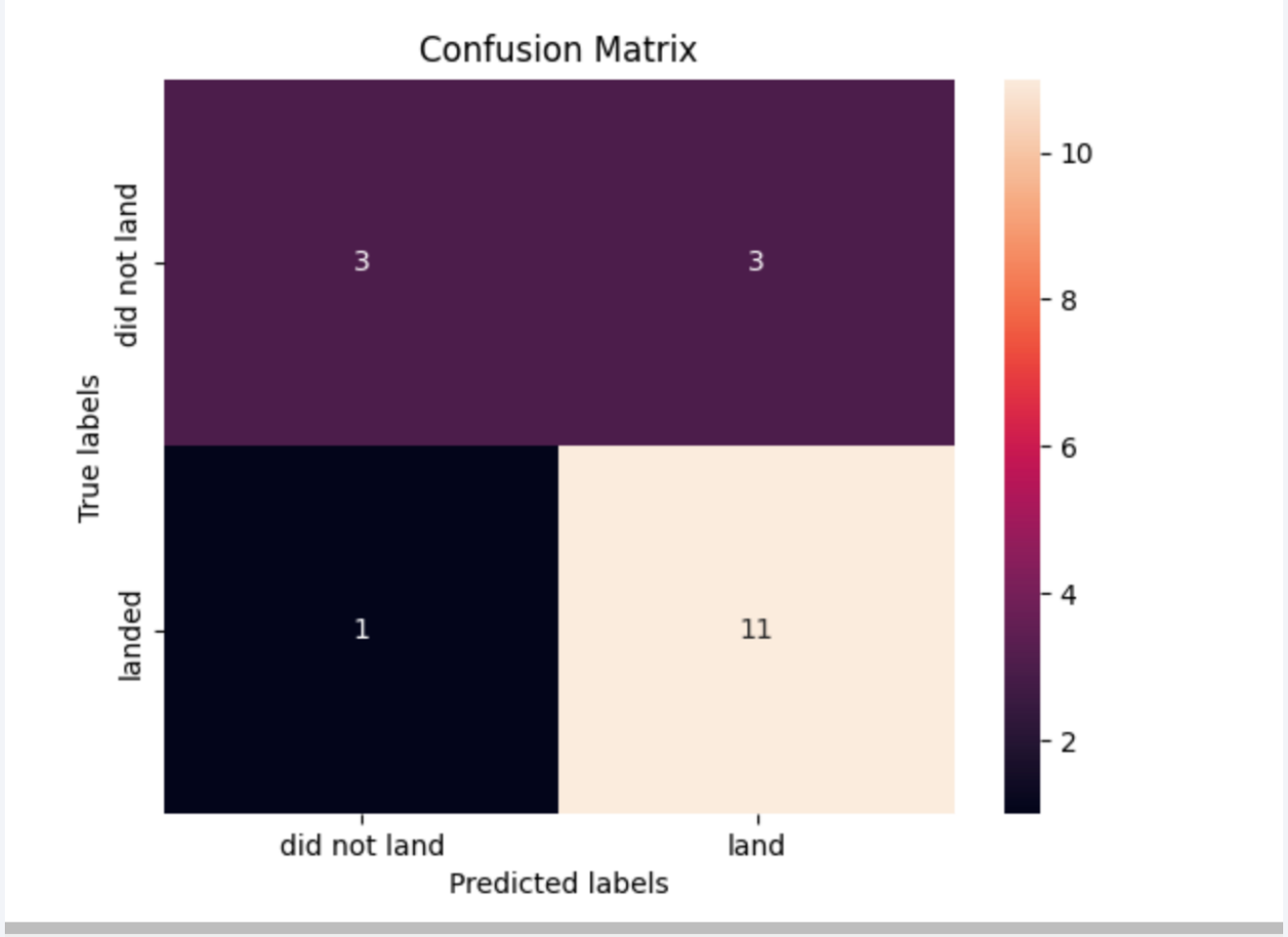
Predictive Analysis (Classification)

Classification Accuracy

```
Best Algorithm is Tree with a score of 0.8875
Best Params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10,
'splitter': 'best'}
```

The tree algorithm achieves the highest accuracy across all algorithms.

Confusion Matrix



The Tree algorithm is able to distinguish between the outcomes. Therein lies a pertinent issue of false positives.

Conclusions

- The success of a mission is contingent upon several factors, such as launch site, payload mass etc. The probability of success increases with time, presumably with experimentation and experience.
- Orbits with the best success rates are as follows: GEO, HEO, SSO, ES-L1.
- KSC LC-39A had the most successful launches across all the launch sites.
- The Tree classifier algorithm is the best machine learning model out of all the algorithms.

Thank you!