Winning Space Race with Data Science

Yifan Yao
July 8, 2024

Outline

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

Executive Summary

- Method: To better understand and predict the outcome of the Falcon 9 launches, the launch data was collected using SpaceX Rest API and webscraped from WikiPedia. After cleaning the data, the exploratory data analysis was conducted to examine the successful rate of each launch, considering the flight number, pay load mass, orbit, launch site location etc. Multiple model were applied to find the best one for predicting the successful rate.
- Results: The most accurate model is support vector machines (SVM)

Introduction

- Background: As one of the most successful commercial space company, SpaceX reduce the cost of each launch by using reusable rockets. Its model Falcon 9 is one of such reusable rocket.
- Questions to be answered: to understand the cost of each launch, one key question is whether each launch can be successful or not which determined if the rocket can be reused and if the cost can be reduced.

Section 1

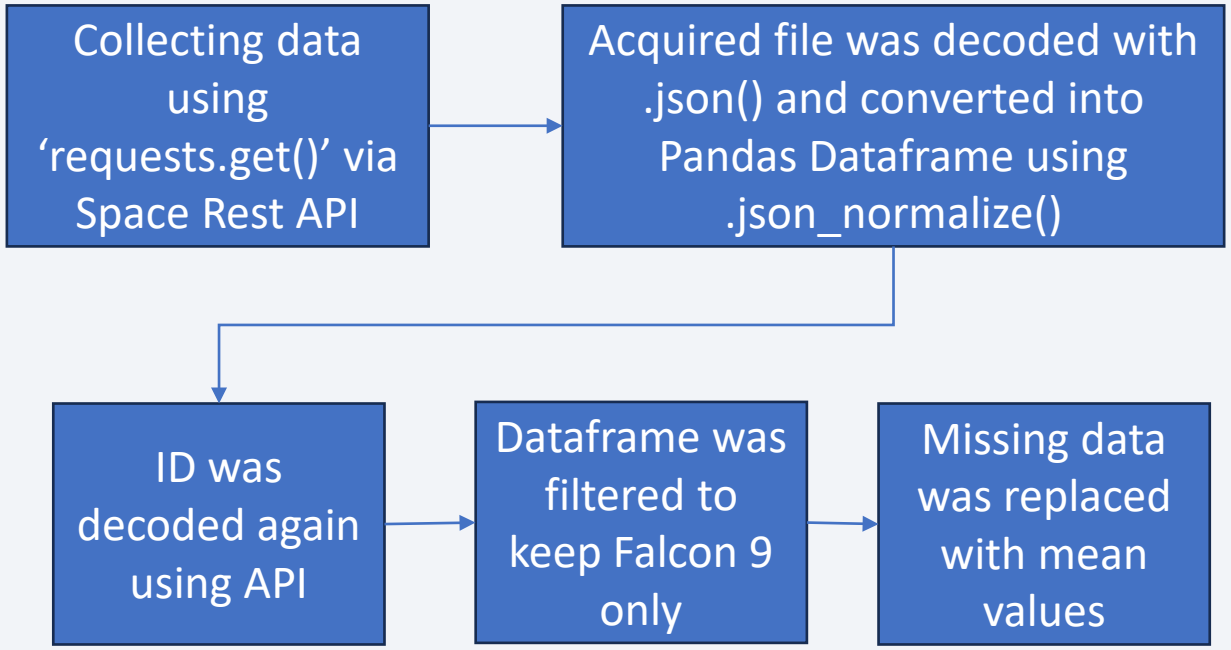
Methodology

Methodology

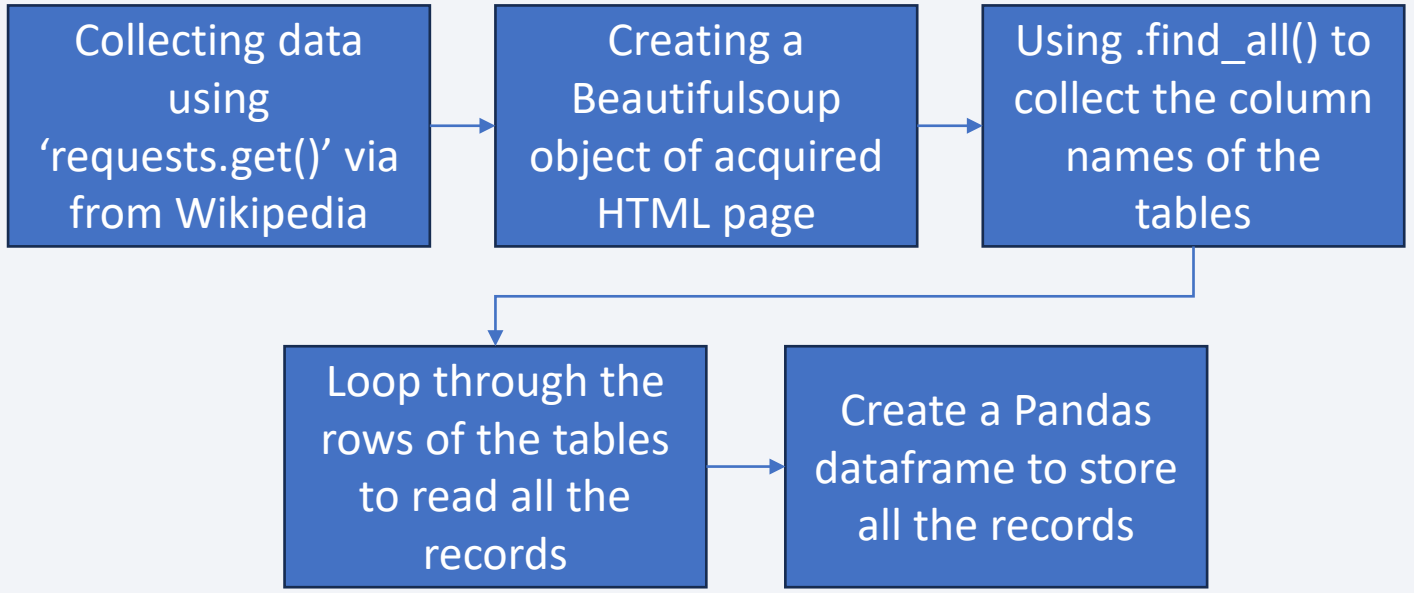
Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - WikiPedia
- Perform data wrangling
  - Missing data was replaced with the mean value of the variable.
  - Success/failure was dummy coded as 1/0
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Multiple models were tested: logistic regression, SVN, Tree classifier, KNN
  - Grid search was used to find optimal hyper-parameters
  - Jaccard score, F1 score, accuracy was used to evaluate models

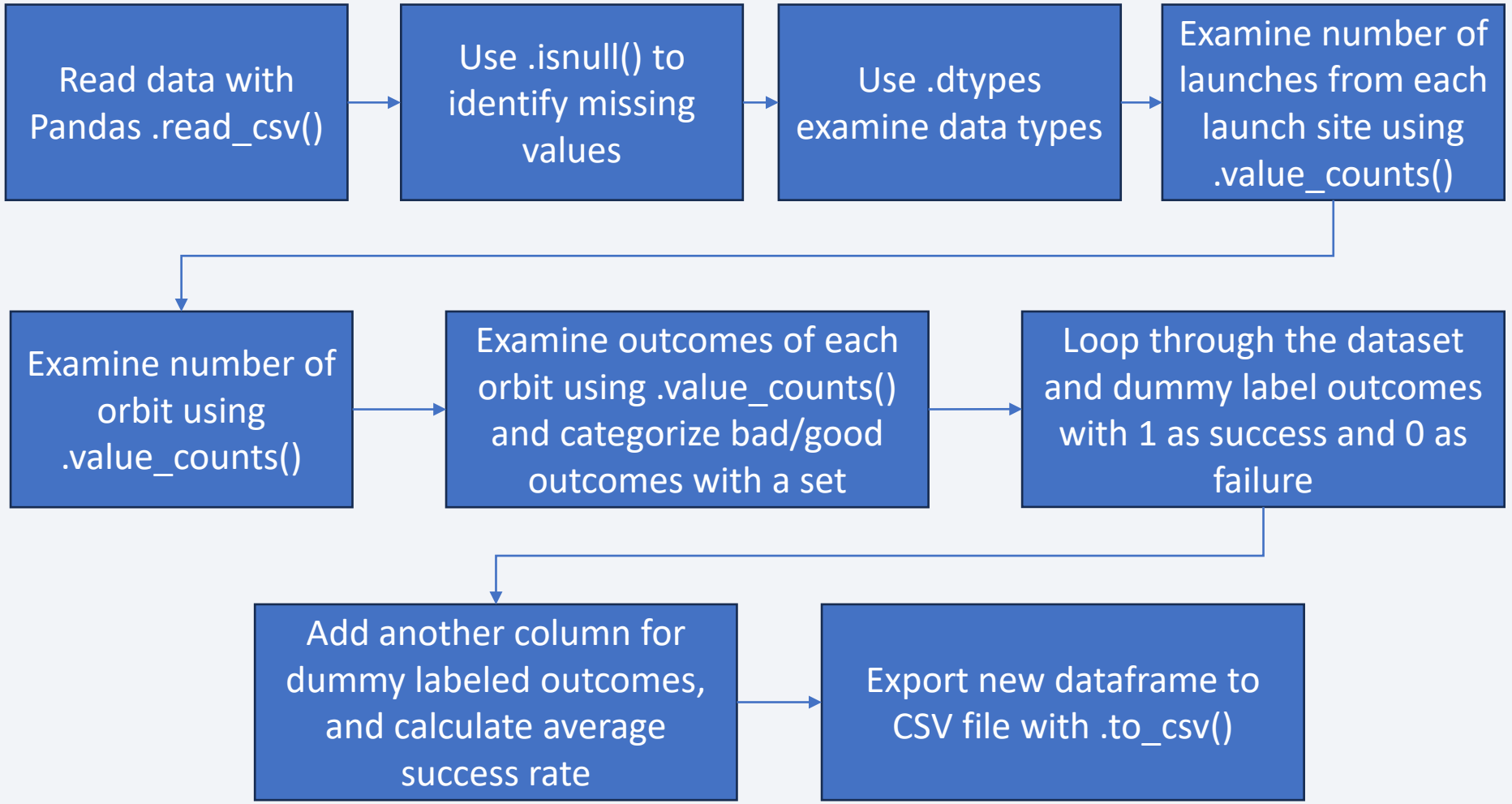
Data Collection – SpaceX API (Github link)



Data Collection – Scraping (Github link)



Data Wrangling (Github link)



EDA with Data Visualization (Github link)

Plot type	Application	Reason
Scatter plot	Flight number vs. pay load mass with outcomes Flight number vs. launch sites with outcomes Pay load mass vs. launch sites with outcomes Flight number vs. orbit with outcomes Pay load mass vs. orbit with outcomes	To explore relationship between different variables
Bar plot	Success rate of different orbits	To compare the measurements of different levels of the variable
Line plot	Success rate over years	To show changes over time

EDA with SQL (Github link)

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin 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 succesful 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. Use a subquery
- 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

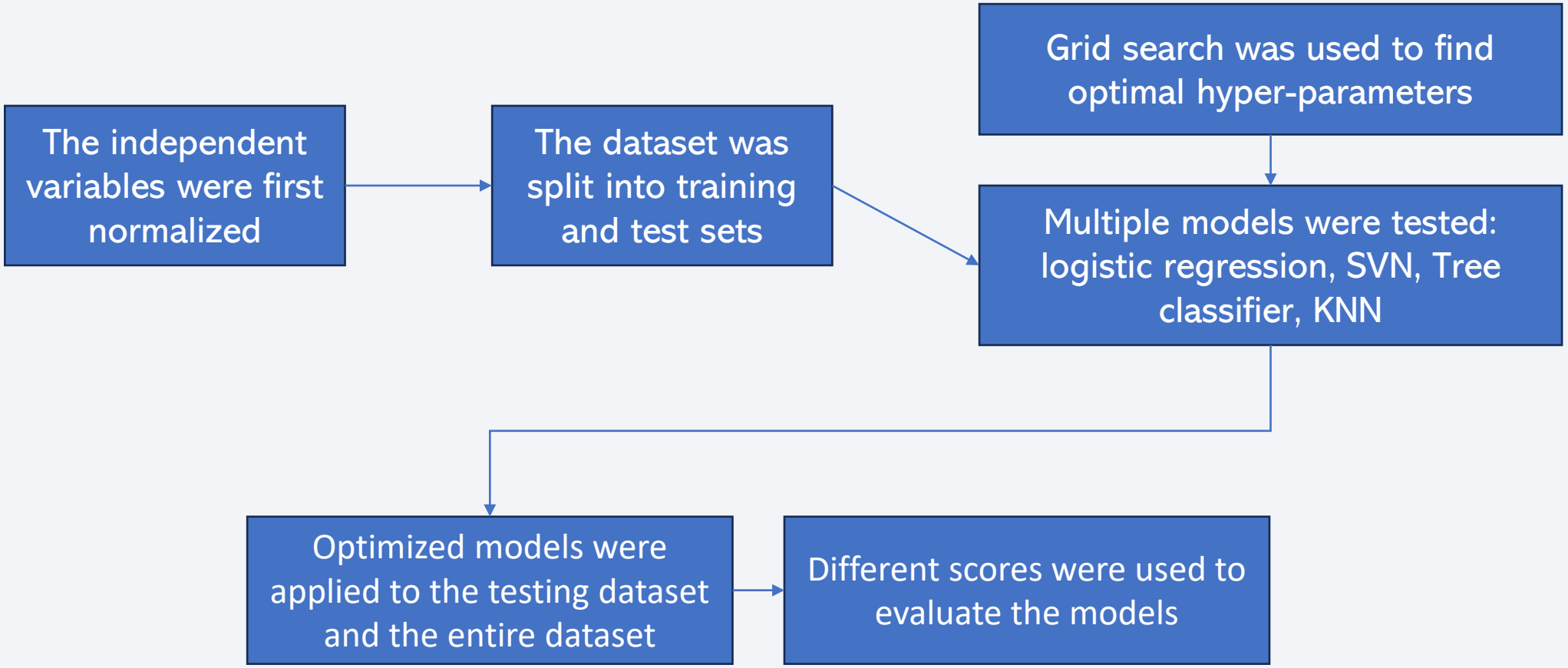
Build an Interactive Map with Folium (Github link)

- Circles for launch sites with labels of launch site names to localize each launch sites
- Markers for outcomes (green as success and red as failure) to find out which launch site has higher success rate
- Lines and distance markers of nearest coastline, city, railway, and highway to launch site CCAFS LC-40 to show the proximities of the launch site

Build a Dashboard with Plotly Dash (Github link)

- A drop-down menu to select different launch site
- A pie chart showing success rate of a select launch site or all launch sites
- A range slider to select pay load mass
- A scatter plot of a select launch site or all launch sites showing relationship of pay load mass and success rate with labels of booster versions
- Drop-down menu and slider are interactions to select levels of variable need to be demonstrated. Pie chart is to show the percentage of success rate. Scatter plot is to show the relationship between variables

Predictive Analysis (Classification) (Github link)



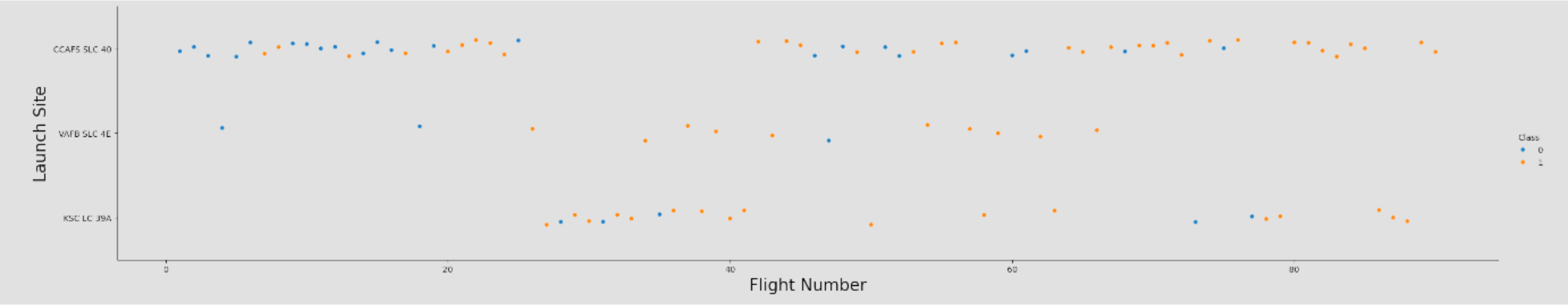
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

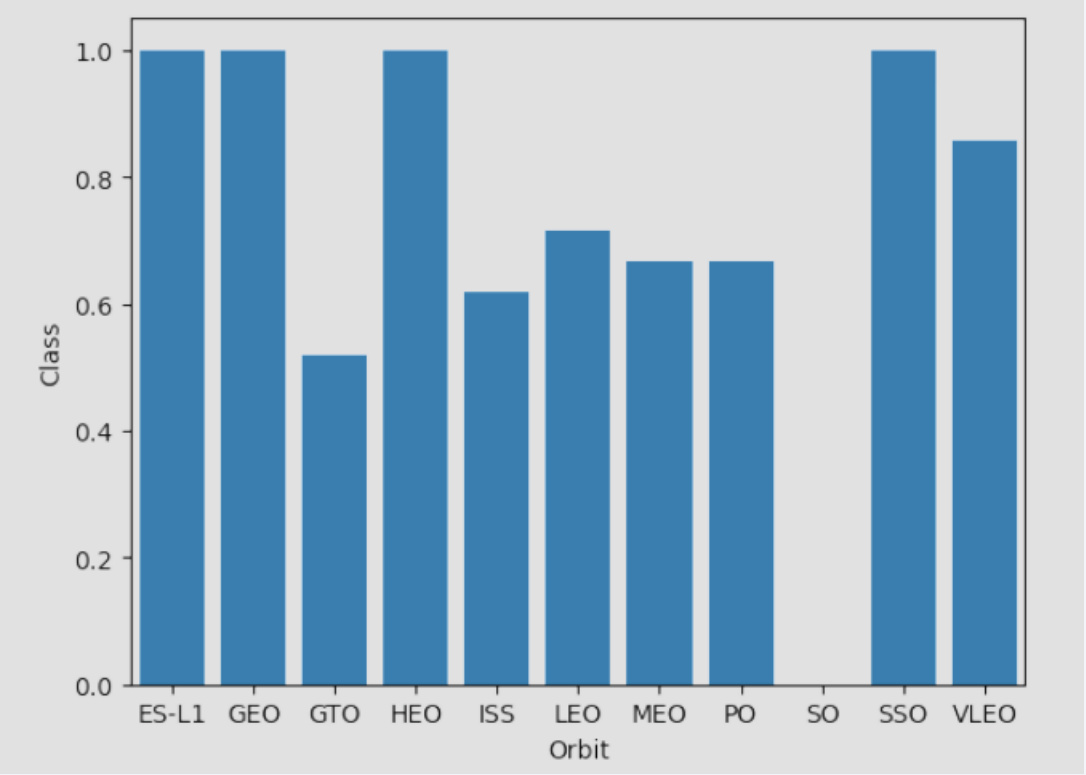


- Between flight number ~23 – 41, there was no launch from CCAFS SLC40;
- Most flights with larger number (>42) are from CCAFS SLC 40
- Flights have larger numbers have higher successful rate

Payload vs. Launch Site

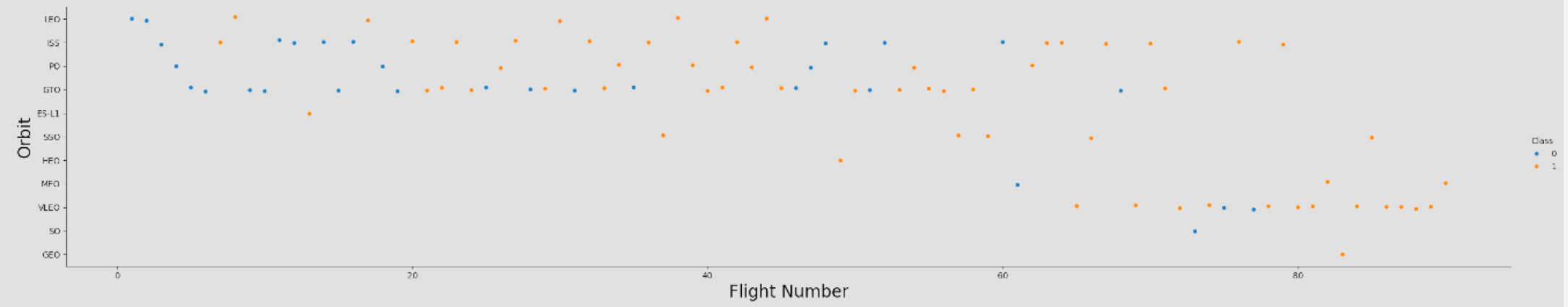
- From CCAFS SLC 40, launches with pay load mass above 7000 have 100% successful rate
- There in no launches from WAFB SLC 4E of launches with pay load mass above 10000
- KSC LC 39A has launches of a wide range of pay load mass. Launches with higher and lower end of the play load mass have relatively higher successful rate.

Success Rate vs. Orbit Type



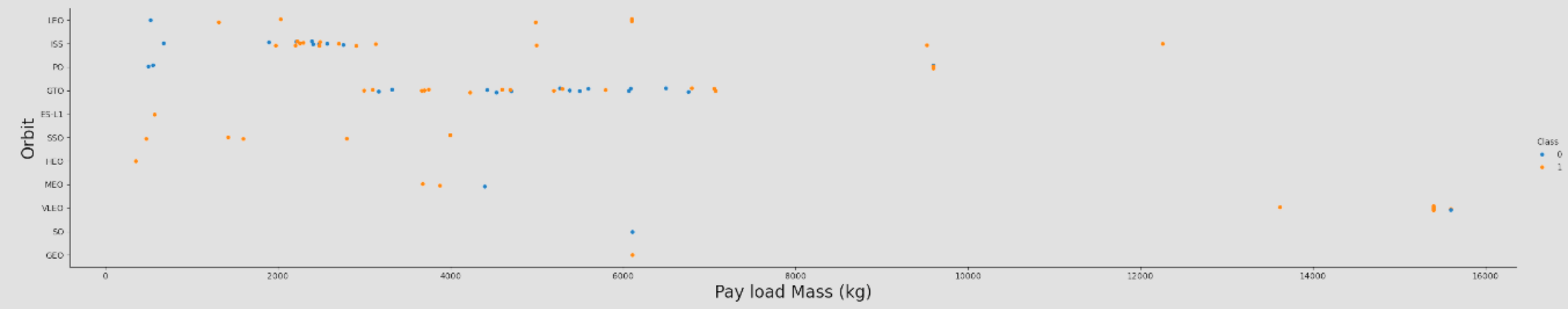
The orbits ES-L1, GEO, HEO, SSO have the highest successful rate.

Flight Number vs. Orbit Type

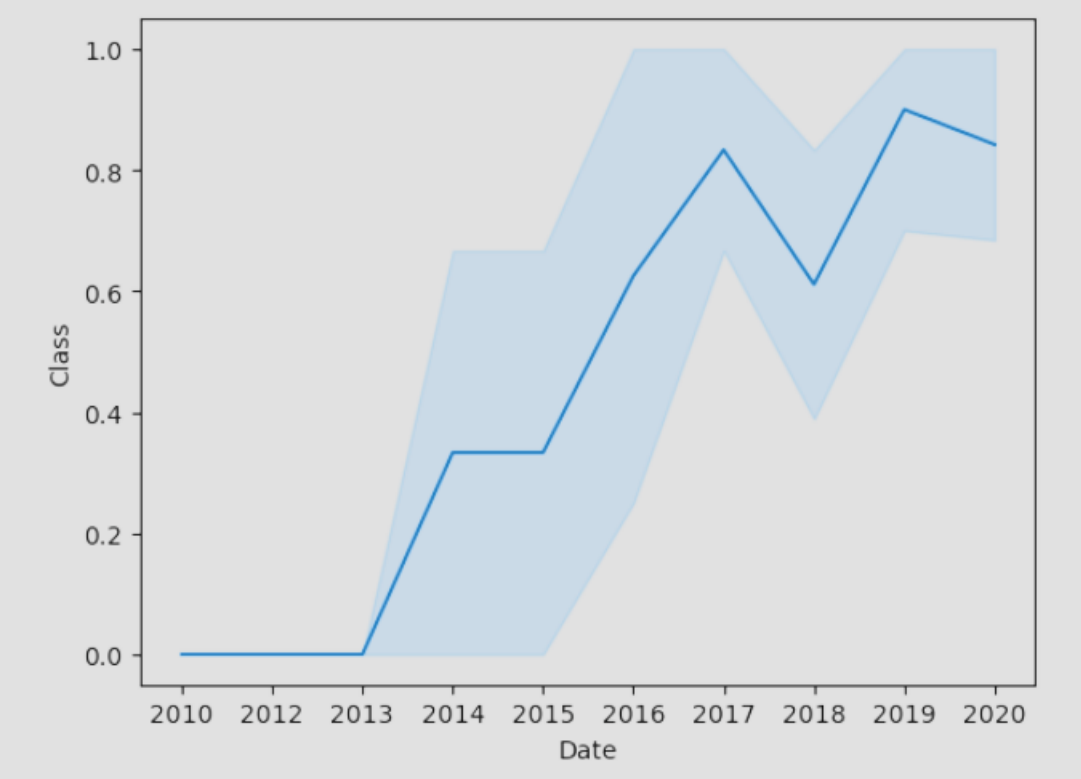


- SSO, HFO, MFO, VLEO, SO, GED only have larger flight numbers.
- For orbit IFO, larger flight number seems to have higher successful rate.

Payload vs. Orbit Type



Launch Success Yearly Trend



The successful rate of launch is increasing over the years.

All Launch Site Names

```
%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

There are 4 distinct launch sites identified.

Launch Site Names Begin with '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
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

First five records with the launch site names beginning with ‘CCA’.

Total Payload Mass

```
%sql select SUM("PAYLOAD_MASS__KG_") from "SPACEXTBL" where "Customer" like '%NASA%'
```

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

SUM("PAYLOAD_MASS__KG_")
107010

Total payload mass carried by boosters launched by NASA

Average Payload Mass by F9 v1.1

```
%sql select AVG("PAYLOAD_MASS__KG_") from "SPACEXTBL" where "Booster_Version"="F9 v1.1"
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql select MIN("Date") from "SPACEXTBL" where "Landing_Outcome" = "Success (ground pad)"
```

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

MIN("Date")
2015-12-22

The dates of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select "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

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql select count("Landing_Outcome") from "SPACEXTBL" where "Landing_Outcome" like "Success%"
```

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

count("Landing_Outcome")
61

```
%sql select count("Landing_Outcome") from "SPACEXTBL" where "Landing_Outcome" like "Failure%"
```

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

count("Landing_Outcome")
10

The total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select "Booster_Version" from "SPACEXTBL" where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from "SPACEXTBL")
```

```
 * 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

The names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%sql select substr("Date", 6,2) as "Month", "Landing_Outcome", "Booster_Version", "Launch_Site" from "SPACEXTBL" where  substr(Date,0,5)='2015' and "Landing_Outcome"
```

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

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

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%sql select "Landing_Outcome", count("Landing_Outcome") as "Num_count" from "SPACEXTBL" where "Date" between '2010-06-04' and '2017-03-20' group by "Landing_Outcome"
```

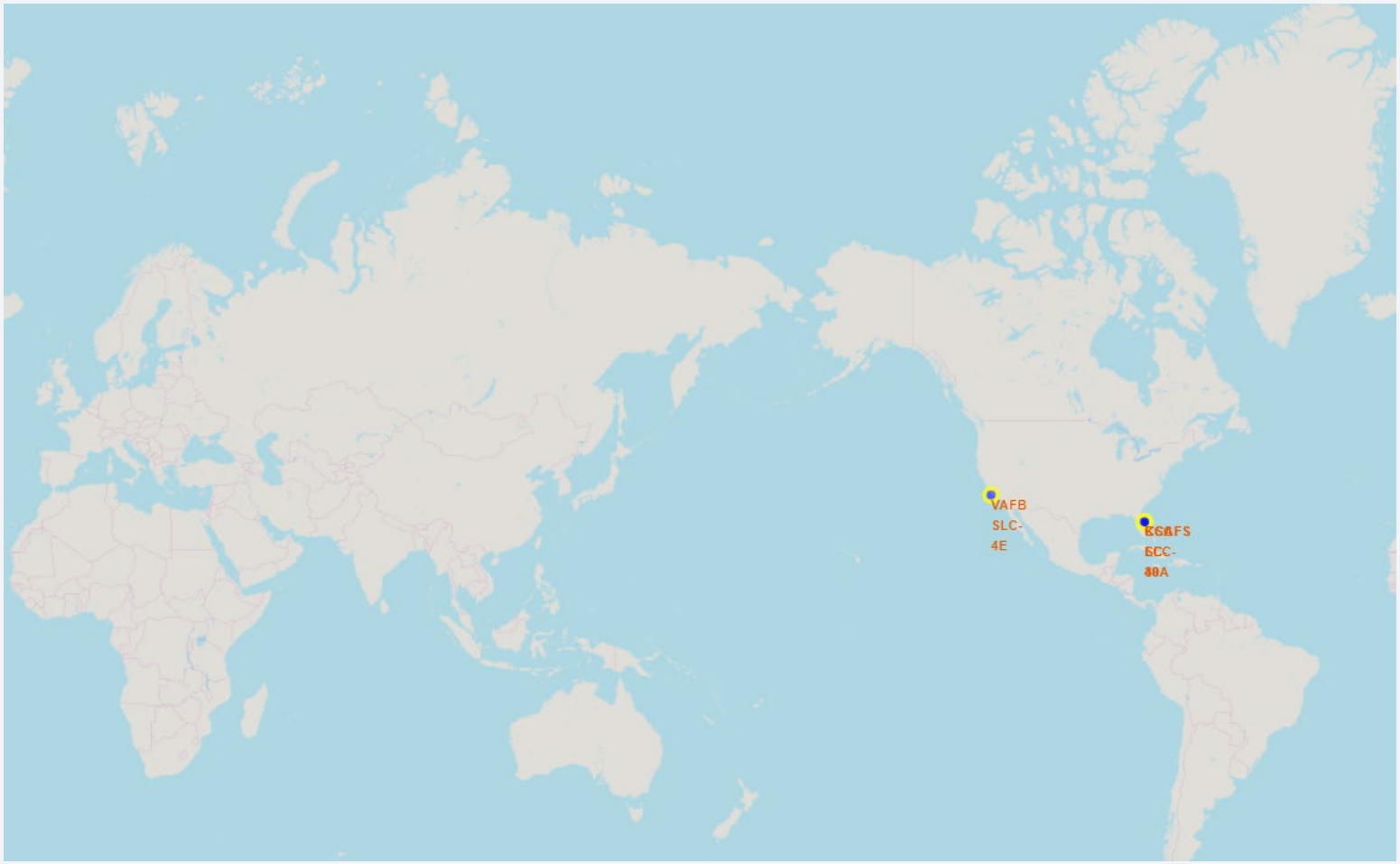
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Num_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

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

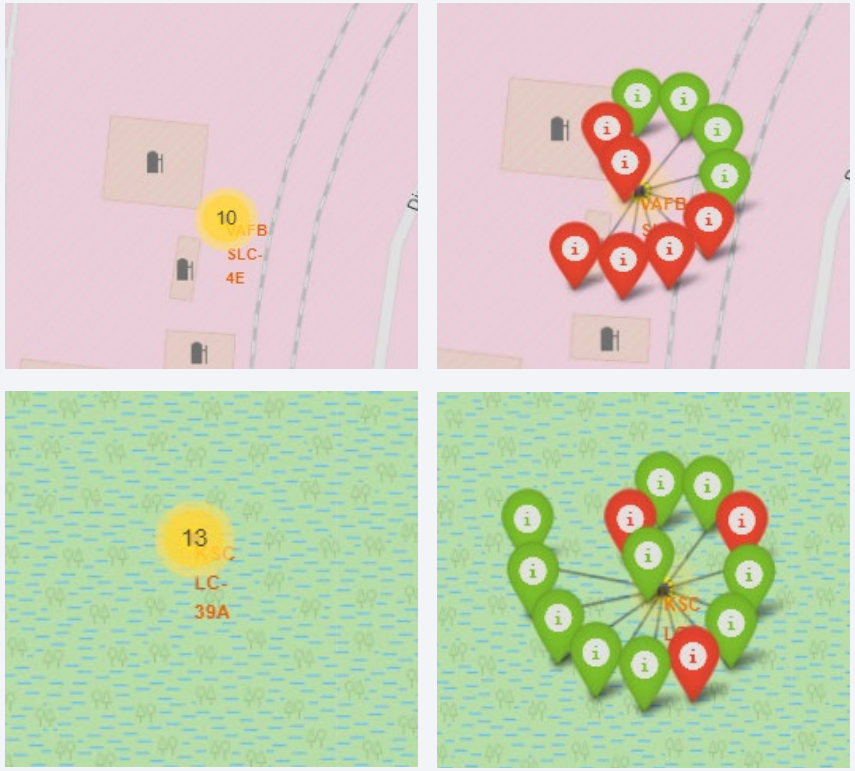

Section 3
Launch Sites
Proximities Analysis

Launch site locations



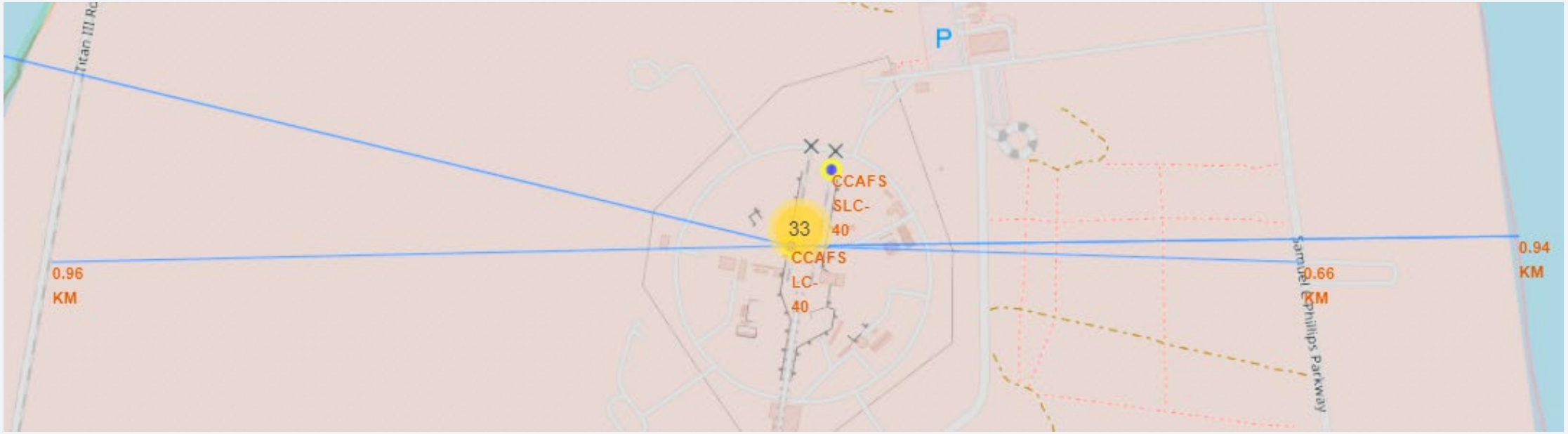
1 launch site is localized to the west coast; 3 launch sites are localized to the east coast.
Most launch sites are close to the equator line and the coastline.

Successful and Failed Launches

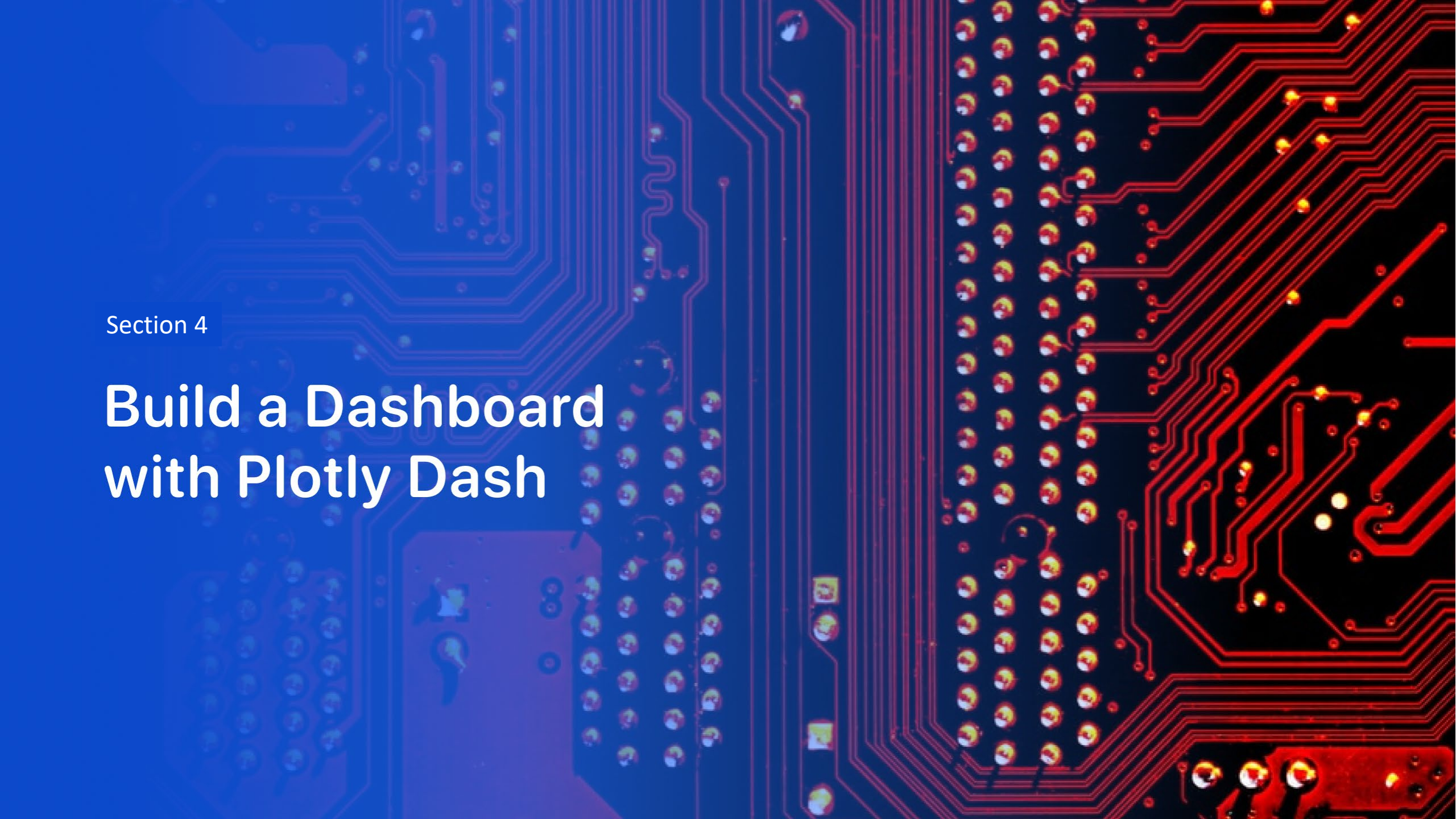


Successful and failed launches of each launch site (green as success, red as failure).
Launch site KSC LC-39A has the highest successful rate.

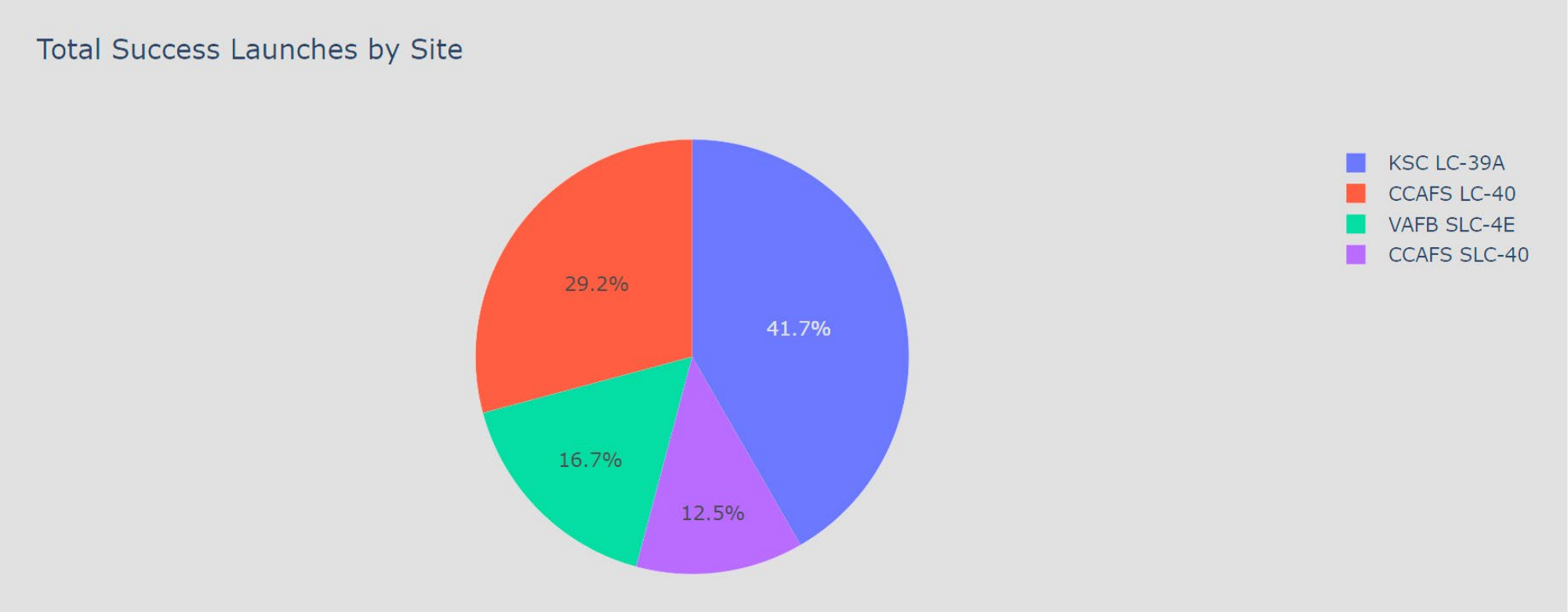
Launch site proximities



Distance of launch site CCAFS LC-40 to its nearest coastline (0.94km), railway (0.96km), highway (0.66km).

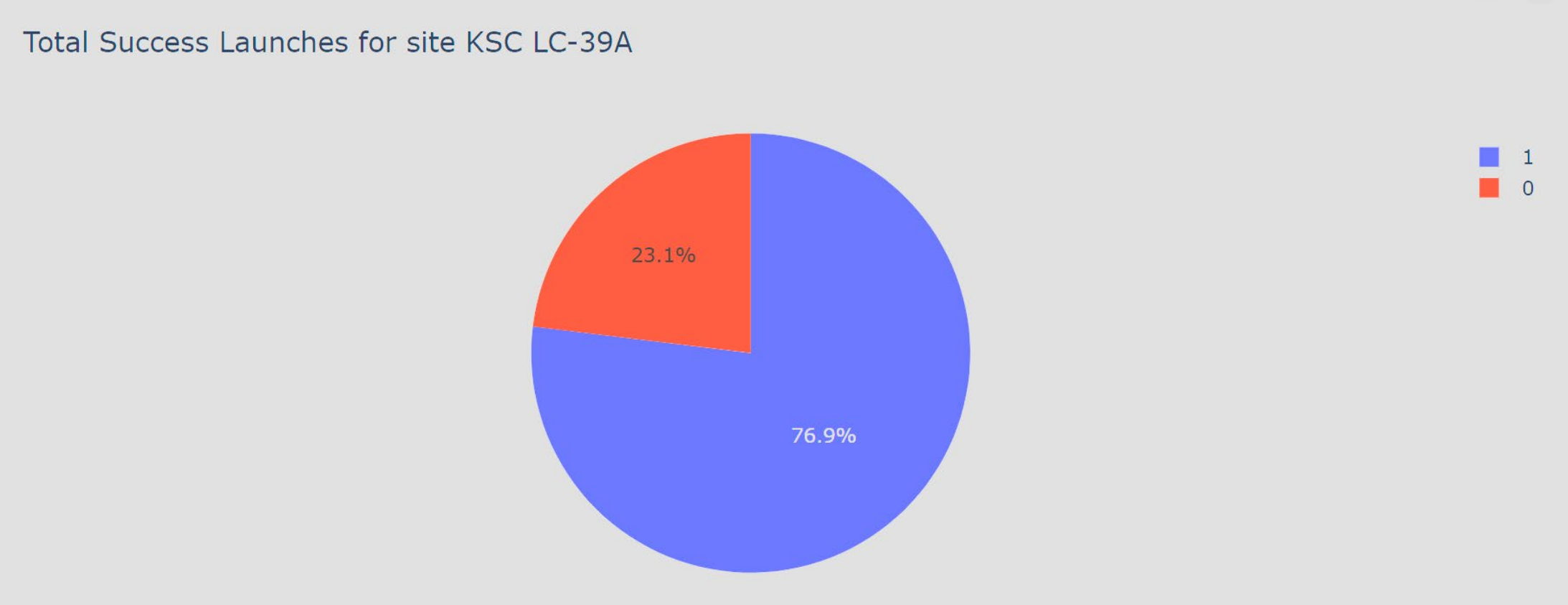

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches of All Sites



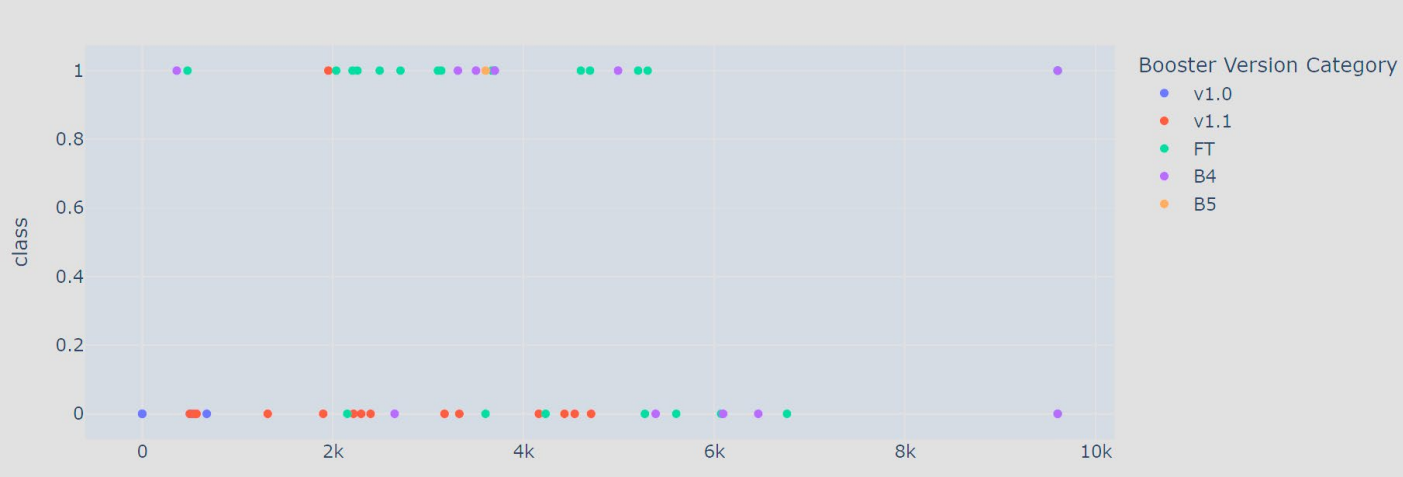
KSC LC-39A has the highest successful launch rate and CCAFS SLC-40 has the lowest.

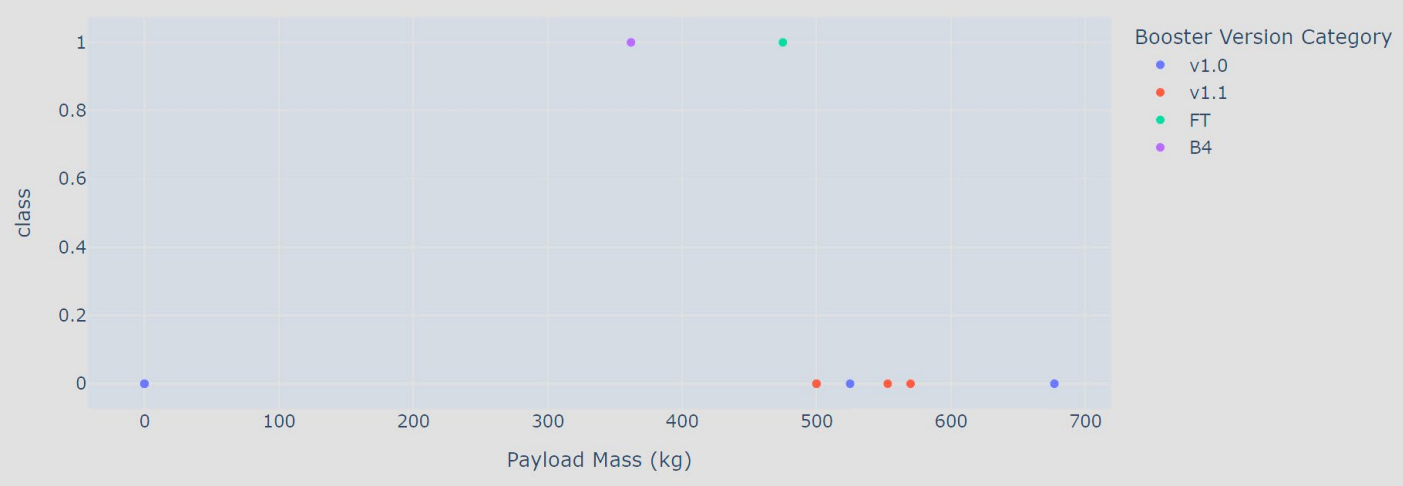
Success Launch Rate of KSC LC-39A



KSC LC-39A has the highest successful launch rate (10/13) of all sites.

Pay Load Mass vs. Successful Rate





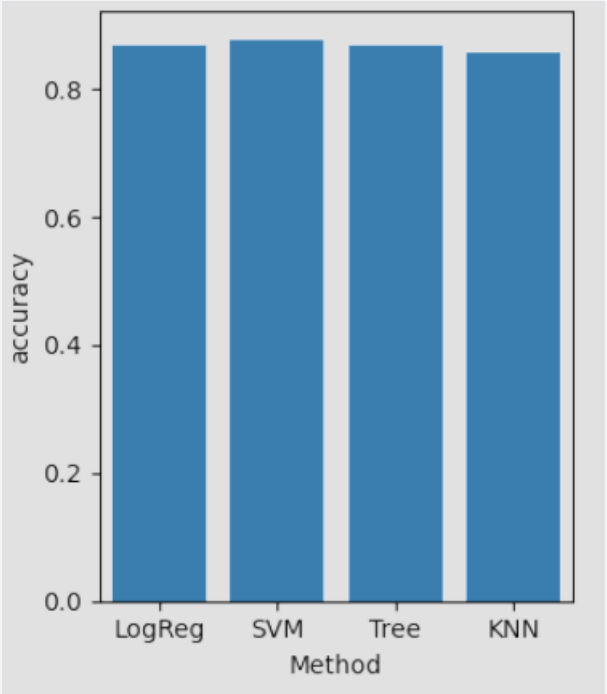
Relationship between pay load mass and successful rate with different booster versions

There is no pay load mass around 8K kg; No B5 booster was used of pay load mass under 700 kg. Booster v1.1 has the lowest successful rate.

Section 5

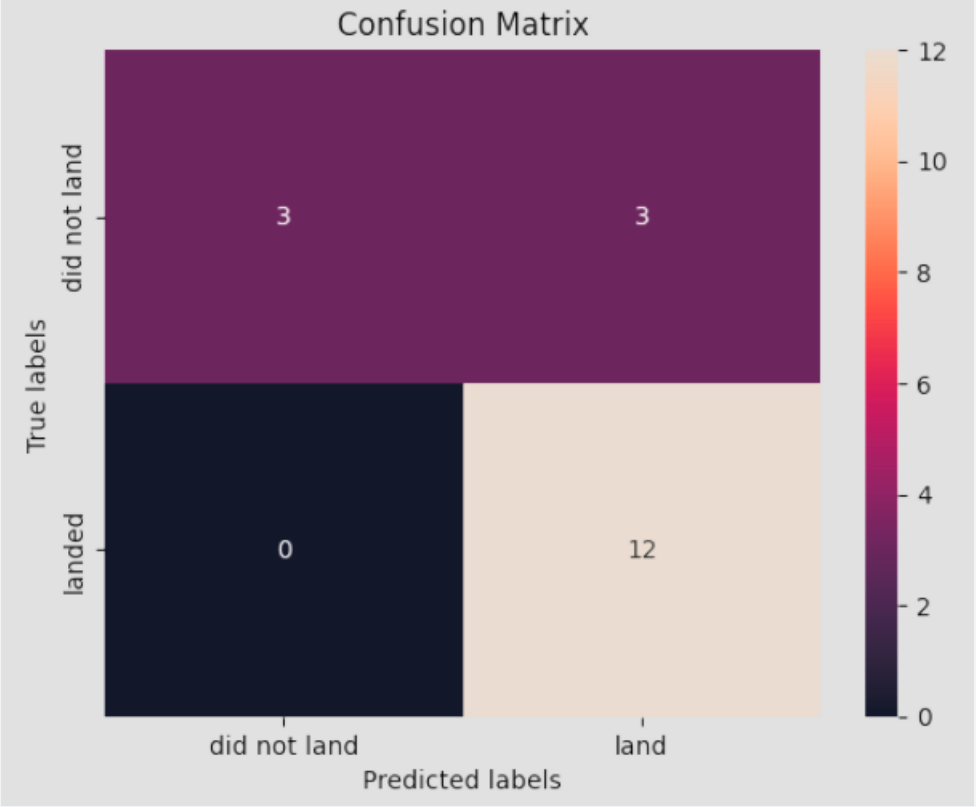
Predictive Analysis (Classification)

Classification Accuracy



The accuracy of all the models. The models were applied to the entire dataset. Based on the plot, SVM has the highest accuracy.

Confusion Matrix



The confusion matrix of the SMV model. The precision is 12/15=0.8, the recall is 12/12 = 1

Conclusions

- Launch site KSC LC-39A has the highest successful rate
- The successful rate of launch is increasing over the years
- Flights have larger numbers have higher successful rate
- The orbits ES-L1, GEO, HEO, SSO have the highest successful rate
- launches with higher pay load mass have higher successful rate
- SVM has the highest accuracy for predicting landing outcomes

Appendix

- Extracted dataset link

Thank you!