IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ralph Mueller
6/27/2023

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection with SpaceX API
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Interactive Map with Folium
  - Dashboard with Plotly
  - Predictive Analysis
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project Background and Context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars
  - SpaceX can save money can reuse the first stage.
  - If we can determine if the first stage will land, we can determine the cost of a launch.
- Problems you want to find answers
  - Find the factors that predict a successful launch of the Falcon 9 rocket

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX Web API as well as with webscraping from Wikipedia
- Perform data wrangling
  - Data was cleaned and one hot encoding was used for Machine Learning
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, KNN, and Decision Tree models were trained and evaluated

Data Collection

- The data was collected from the Space x API available at https://api.spacexdata.com/v4/launches/past
- Additional data for rocket version, payload, launch site, cores were retrieved via helper functions
- Furthermore, Wikipedia was scraped using Beautiful Soup to retrieve Launch Data for the Falcon 9 rocket

Data Collection – SpaceX API

Launch Data was obtained from:
https://api.spacexdata.com/v4/launches/past

Create Pandas DF with pd.json_normalize(data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc']] and clean data

Retrieve Additional Data for BoosterVersion, LaunchSite, Payload, and CoreData, Append to a dictionary and then convert it to a Pandas Data Frame.

Filter Data Frame to only include Falcon 9 launches, fill missing payload values with mean

GitHub Url:
https://github.com/ralph240574/datascience-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from
static_url =
"https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
And create Beautiful Soup Object

↓

Extract all column/variable names from the HTML table header

↓

Create a data frame by parsing the launch HTML tables

↓

Save Data Frame as CSV

GitHub URL: https://github.com/ralph240574/datascience-capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome per orbit type

Create a landing outcome label from Outcome column

GitHub: https://github.com/ralph240574/datascience-capstone/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Catplot to visualize the relationship between flight number and Launch Site, flight numbers are plotted on the x-axis and y-axis shows the lanch site. As the flight numbers are increasing the success rate tends to increase for most launch sites
- Catplot to visualize the relationship between payload and launch site, payload is plotted on the x-axis and launch sites are plotted on the y–axis
- Barchart to visualize the relationship between success rate of each orbit type, with x- axis showing orbit type and y-axis showing success rate
- Catplot to visualize the relationship between FlightNumber and Orbit type, with flight number on x-axis and orbit type on y-axis
- GitHub URL: https://github.com/ralph240574/datascience-capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization

- Catplot to visualize the relationship between Payload and Orbit type, with Payload on x-axis and orbit type on y-axis
- Linechart to visualize the launch success yearly trend,
- GitHub URL: https://github.com/ralph240574/datascience-capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL queries performed:
  - Unique launch sites: "**select distinct** LAUNCH_SITE **from** SPACEX"
  - 5 records where launch sites begin with the string 'CCA': **"select * from SPACEX where launch_site LIKE 'CCA%' limit 5"**
  - Total payload mass carried by boosters launched by NASA (CRS): **"select sum(payload_mass__kg_) as total_payload_mass_kg from SPACEX where customer = 'NASA (CRS)'"**
  - Average payload mass carried by booster version F9 v1.1: **"select avg(payload_mass__kg_) as avg_payload_mass_kg from SPACEX where booster_version LIKE 'F9 v1.1%'"**
  - Date when the first succesful landing outcome in ground pad was achieved: **"select DATE from SPACEX where "Landing _Outcome" LIKE 'Success (ground pad)%' order by date asc limit 1"**
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000: **"select date, booster_version, payload, payload_mass__kg_ from SPACEX where payload_mass__kg_ between 4000 and 6000 and "Landing _Outcome" LIKE 'Success (drone ship)%'"**

EDA with SQL

- SQL queries performed:

  - Total number of successful and failure mission outcomes: **"select count(*), mission_outcome from SPACEX group by mission_outcome"**
  - Names of the booster_versions which have carried the maximum payload mass: **"select * from SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX)"**
  - 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: "**select** monthname(DATE) **as month**, "Landing _Outcome" **from** SPACEX **where** DATE **LIKE** '2015%' **and** "Landing _Outcome" **LIKE** 'Failure (drone ship)%'"
  - Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order: "**select count**(*) **as count**,"Landing _Outcome" **from** SPACEX **where** date **between** '2010-06-04' **and** '2017-03-20' **and** "Landing _Outcome" **LIKE** 'Success%' **group by** "Landing _Outcome" **order by count desc"**

GitHub URL https://github.com/ralph240574/datascience-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Launch sites were marked with Markes and Circles.
- MarkerCluster Object and Markers are used to visualize successful and failed Launches
- Distances of LaunchSites to proximities were visualized using PolyLines

- GitHub URL: https://github.com/ralph240574/datascience-capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Piechart with with launches by certain launch sites, the drop down menu allows to pick individual sites or all
- Scatter Plot showing relationship between payload and and outcome for different booster versions, slider allows to filter by pay load
- GitHub URL: https://github.com/ralph240574/datascience-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data was loaded using Pandas and scaled
- The response Y was created using Numpy
- Dataset was split into training set and test set
- Logistic Regression, SVM, Decision Tree Classification, and KNN was performed using GridSearch
- GitHub URL: https://github.com/ralph240574/datascience-capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

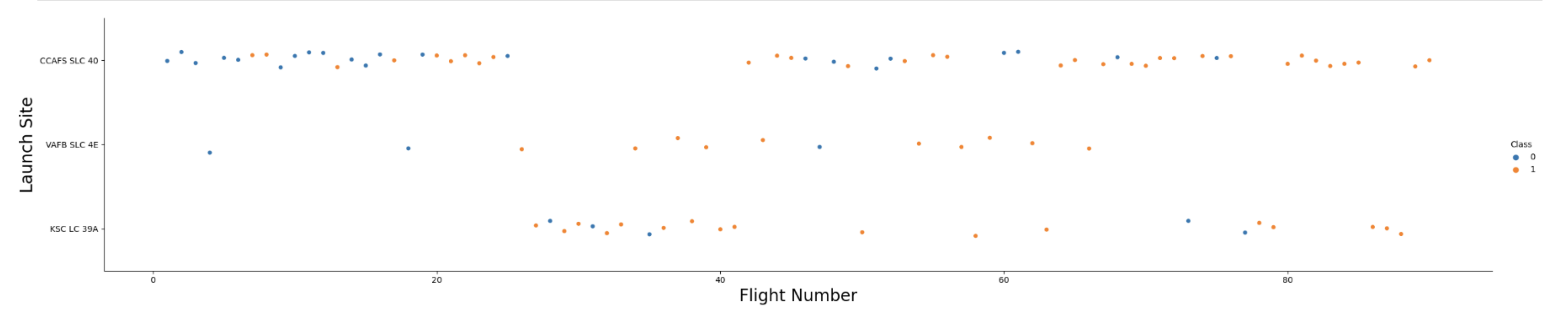
Results

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

Section 2

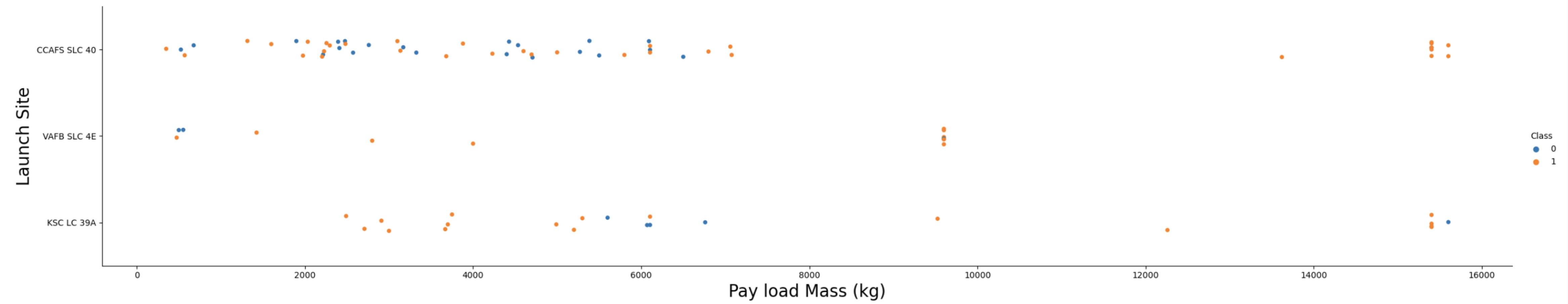
Insights drawn from EDA

Flight Number vs. Launch Site



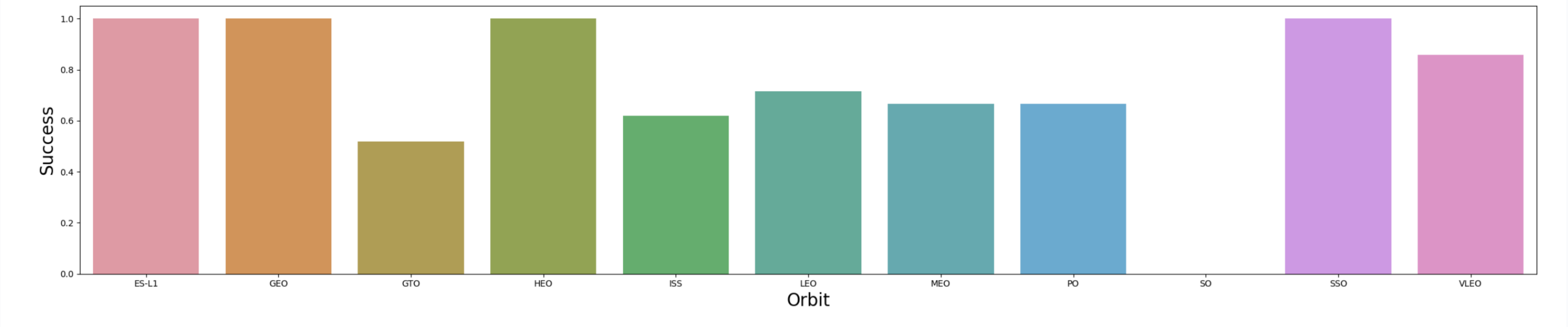
The plot shows that higher flight numbers have a higher chance of success

Payload vs. Launch Site



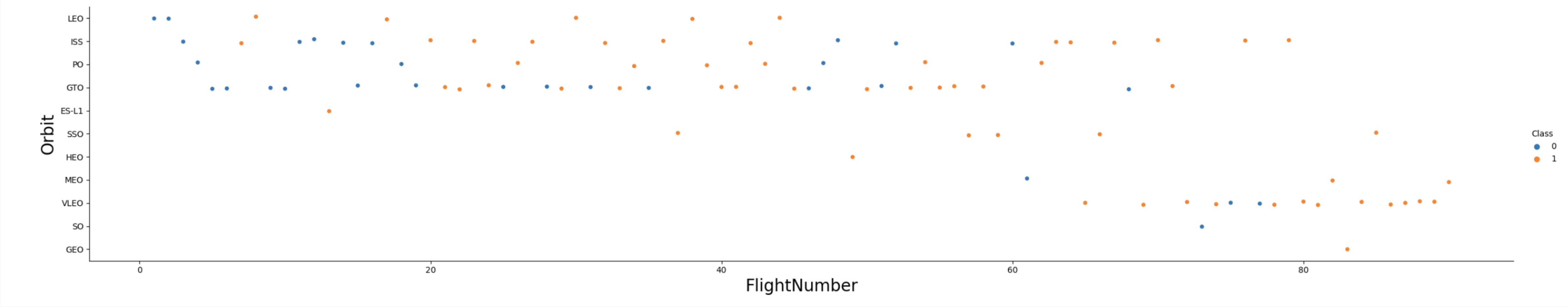
For Site CCAFS SLC 40 higher payloads have higher chance of success

Success Rate vs. Orbit Type



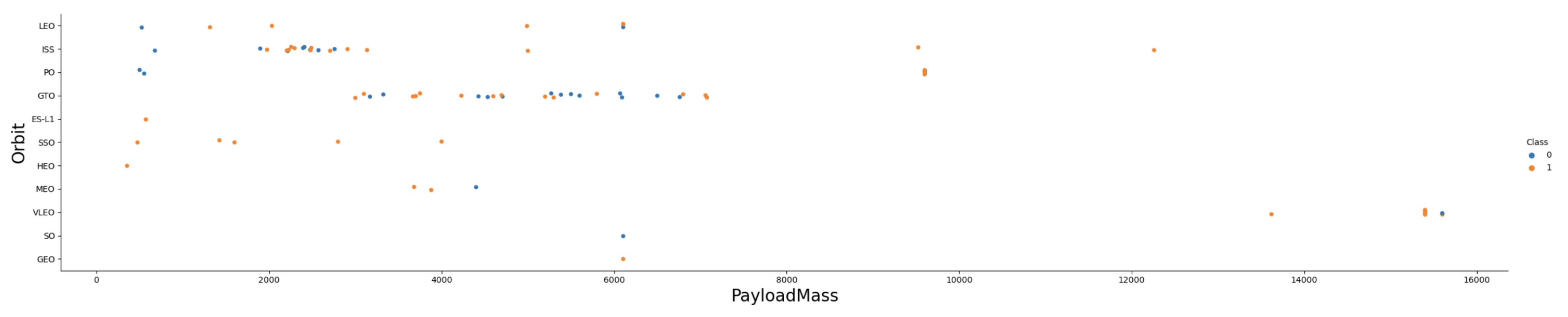
- ES-L1, GEO, HEO, SSO orbits had the highest success rates

Flight Number vs. Orbit Type



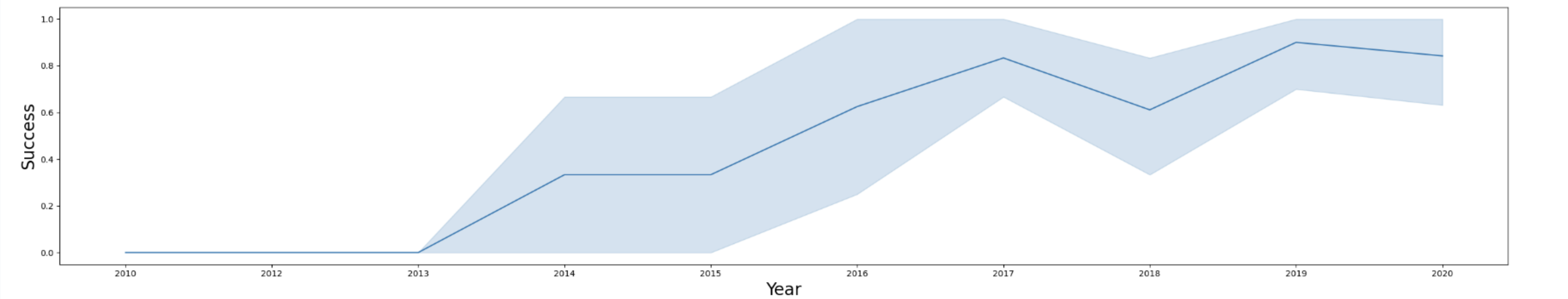
- For LEO the success rate is higher for higher flight numbers

Payload vs. Orbit Type



LEO, PO, and ISS have higher success for higher payloads

Launch Success Yearly Trend



Success rate is increasing for most years

All Launch Site Names

Use DISTINCT keyword to only show unique values

```
%sql Select DISTINCT Launch_Site from SPACEXTBL where Launch_Site != "None";
```

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

Launch Site Names Begin with 'CCA'

```
%sql 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	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

Using SUM function to calculate total payload

```
%sql Select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = "NASA (CRS)";
```

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

Using AVG function to calculate Average:

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Using MIN function and converting dateformat:

```
%sql Select Min(substr(Date,7,4)||"/"||substr(Date,1,2)||"/"||substr(Date,4,2)) AS
First from SPACEXTBL where Landing_Outcome = "Success (ground pad)";
```

First
2015/22/12

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql Select distinct Booster_Version from SPACEXTBL where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 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

%sql Select Count(Mission_Outcome) as missions from SPACEXTBL where Mission_Outcome like "Success%" or Mission_Outcome like "Failure%";

missions
101

Boosters Carried Maximum Payload

%sql Select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (Select max(PAYLOAD_MASS__KG_) from SPACEXTBL);

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,7,4) as year, strftime('%m',substr(Date, 4,2)) as month from SPACEXTBL where year = '2015'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	year	month
01/10/2015	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395.0	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)	2015	12
02/11/2015	23:03:00	F9 v1.1 B1013	CCAFS LC-40	DSCOVR	570.0	HEO	U.S. Air Force NASA NOAA	Success	Controlled (ocean)	2015	12
03/02/2015	3:50:00	F9 v1.1 B1014	CCAFS LC-40	ABS-3A Eutelsat 115 West B	4159.0	GTO	ABS Eutelsat	Success	No attempt	2015	11
14/04/2015	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898.0	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)	2015	11
27/04/2015	23:03:00	F9 v1.1 B1016	CCAFS LC-40	Turkmen 52 / MonacoSAT	4707.0	GTO	Turkmenistan National Space Agency	Success	No attempt	2015	11
28/06/2015	14:21:00	F9 v1.1 B1018	CCAFS LC-40	SpaceX CRS-7	1952.0	LEO (ISS)	NASA (CRS)	Failure (in flight)	Precluded (drone ship)	2015	11
22/12/2015	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034.0	LEO	Orbcomm	Success	Success (ground pad)	2015	12

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

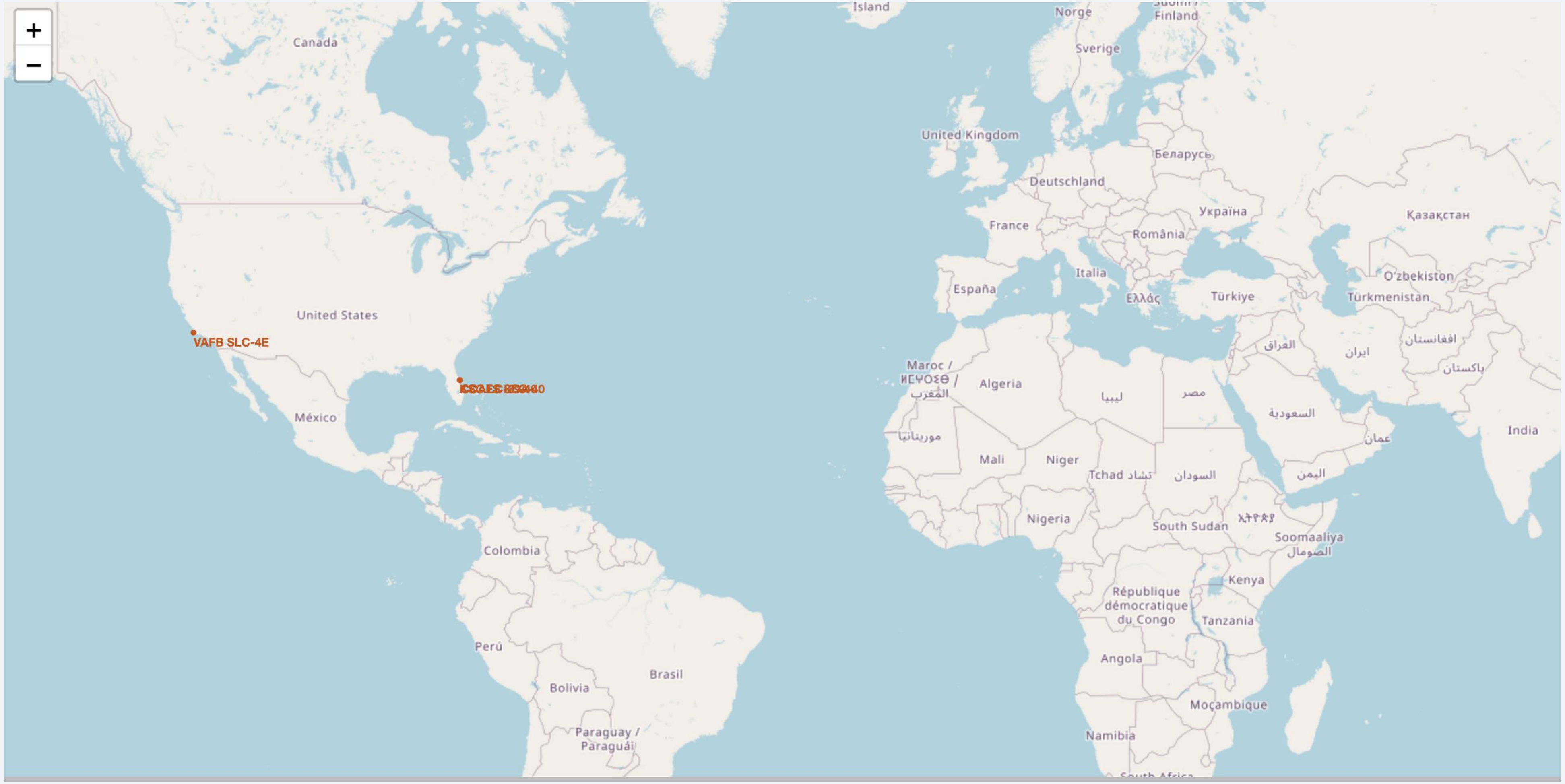
%sql Select Landing_Outcome, count(Landing_Outcome) from SPACEXTBL where landing_outcome like 'Success%' and substr(Date, 7,4) between '2010' and '2017' group by Landing_Outcome;

Landing_Outcome	count(Landing_Outcome)
Success (drone ship)	12
Success (ground pad)	8

Section 3

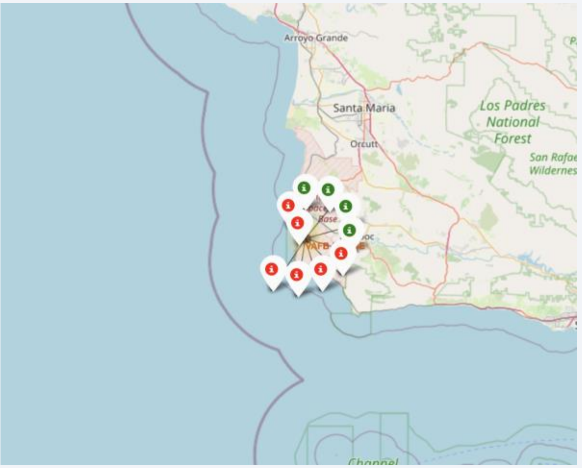
Launch Sites Proximities Analysis

SpaceX Launch Sites

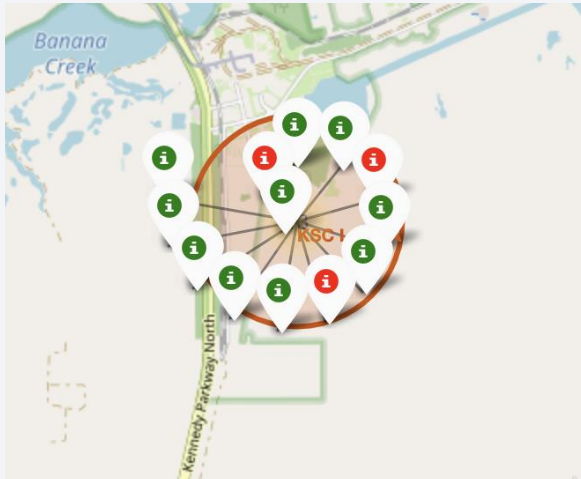


Success/Failures on Map

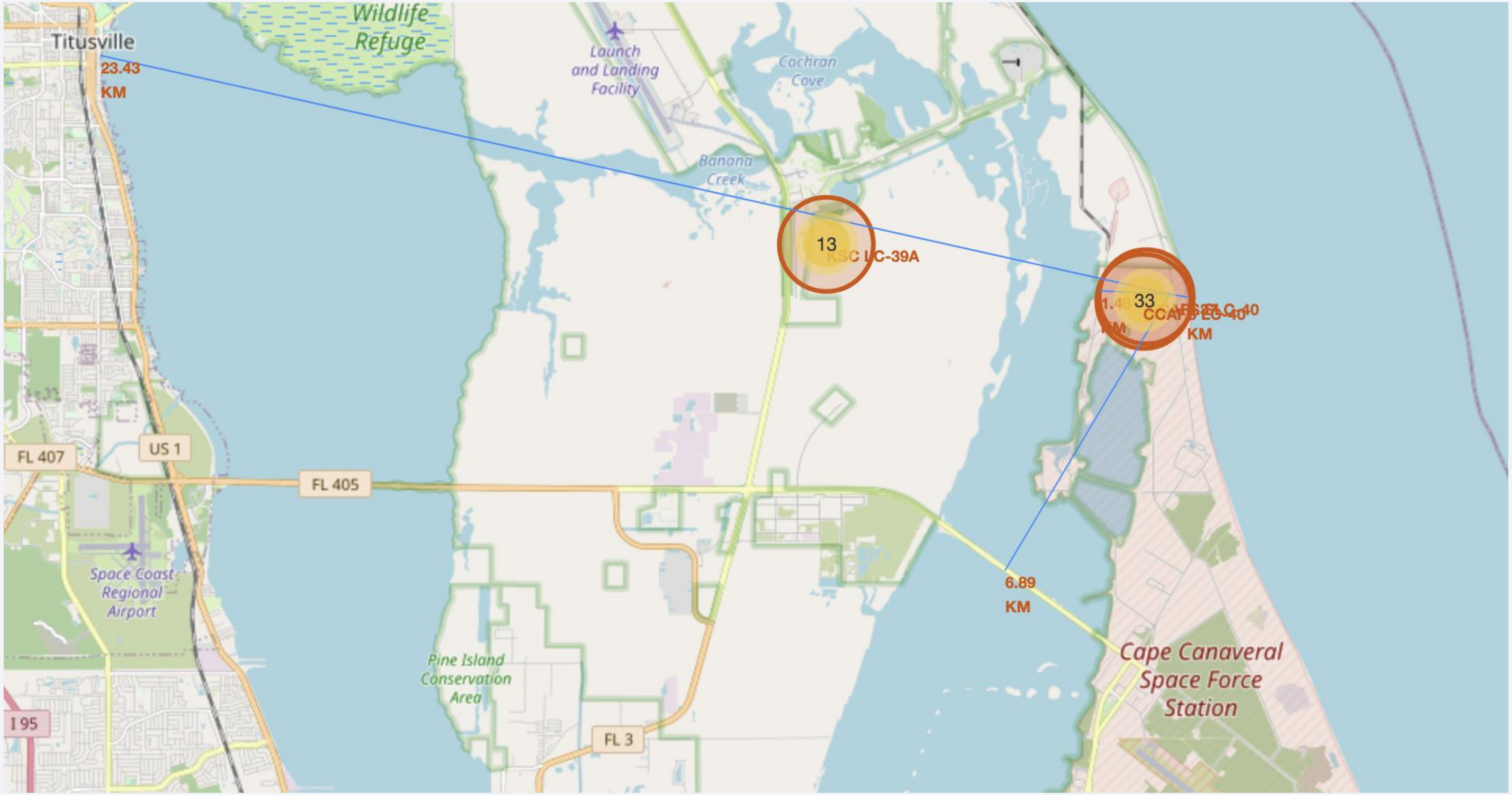
- California:



- Florida:



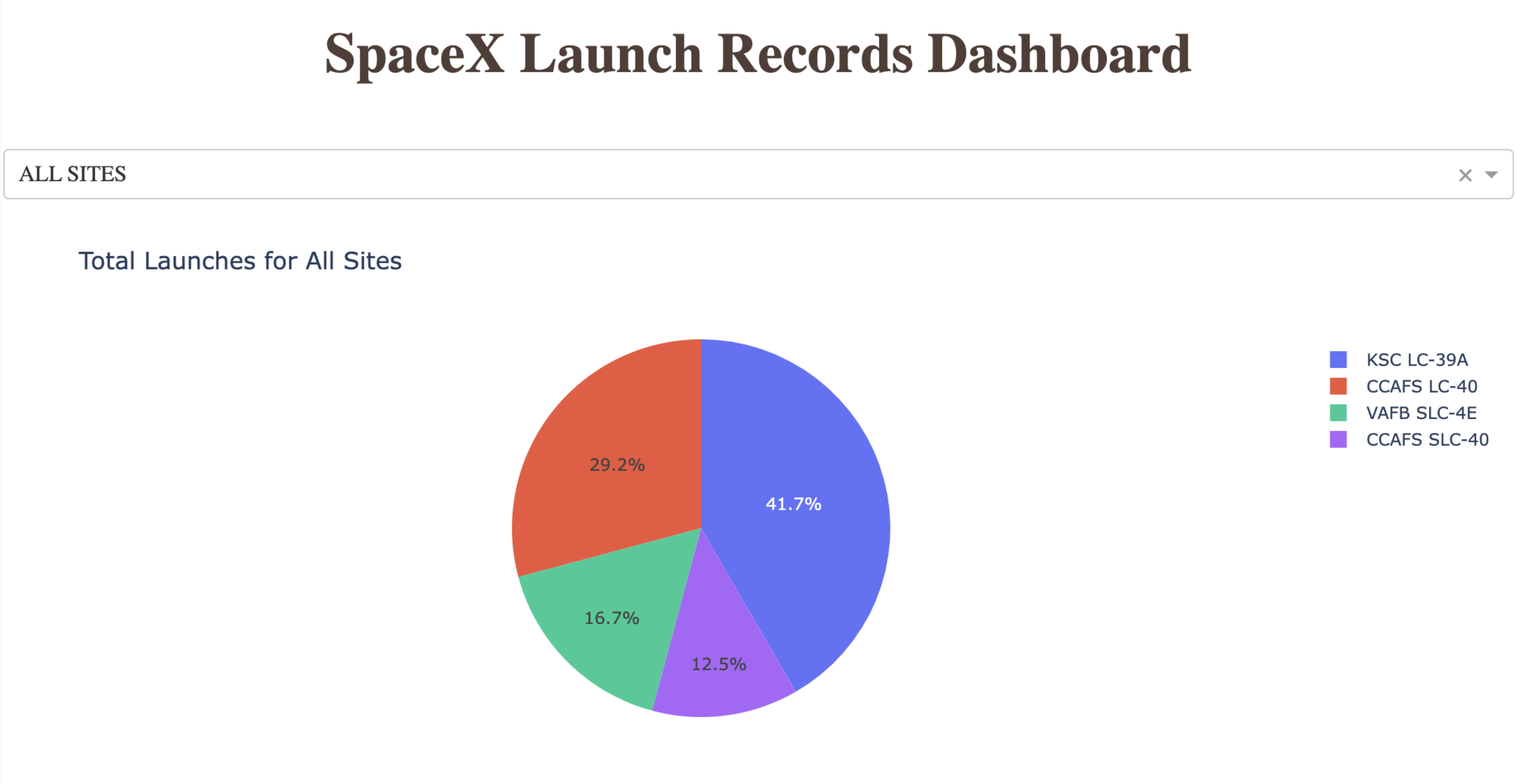
Distances to Proximities for CCAFS SLC-40



Section 4

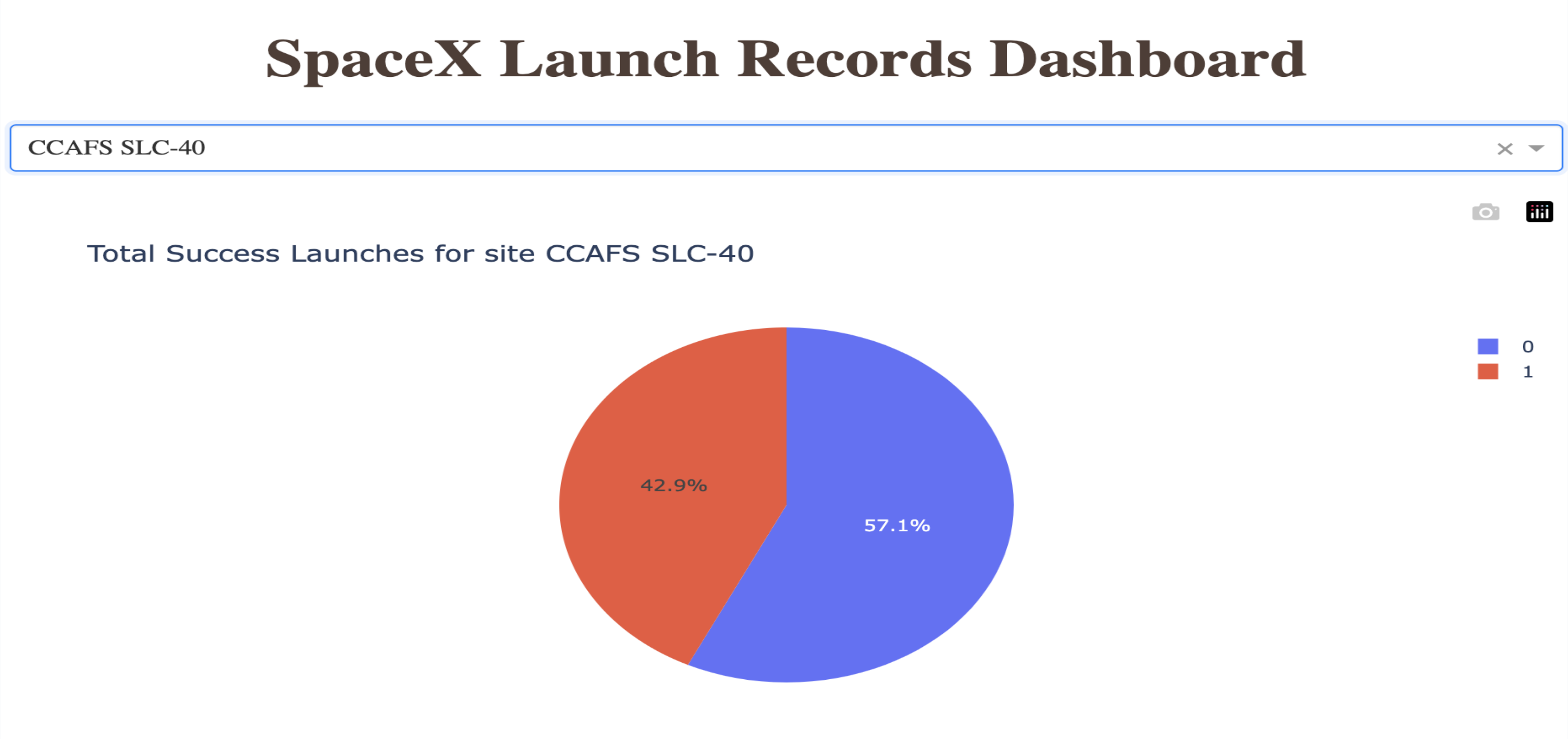
Build a Dashboard with Plotly Dash

Total Launches for all sites



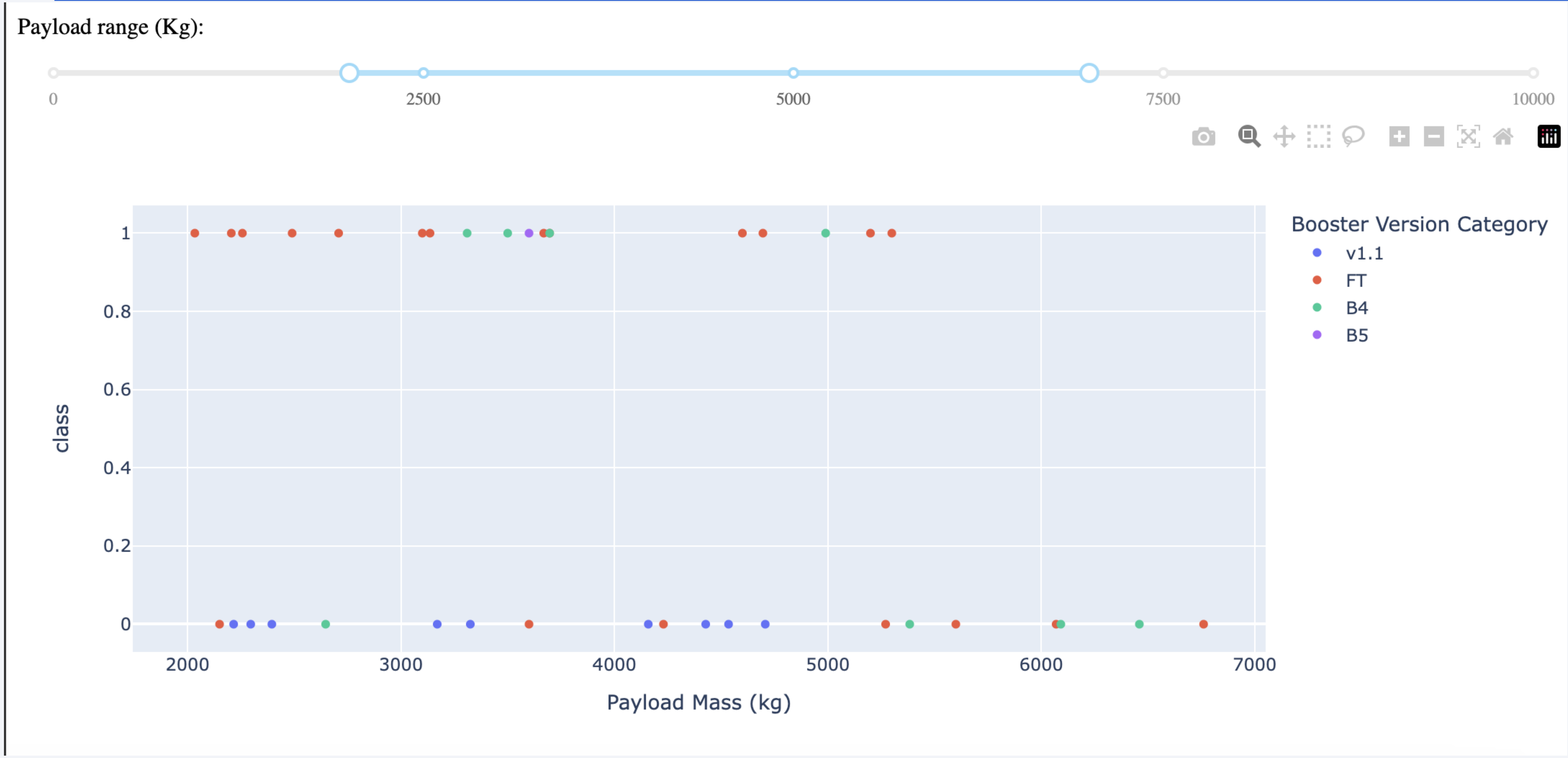
KSC LC-39A had most launches

Highest Success Rate Launch Site



CCAFS SLC-40 has the highest success rate

Payload vs Launch Outcome

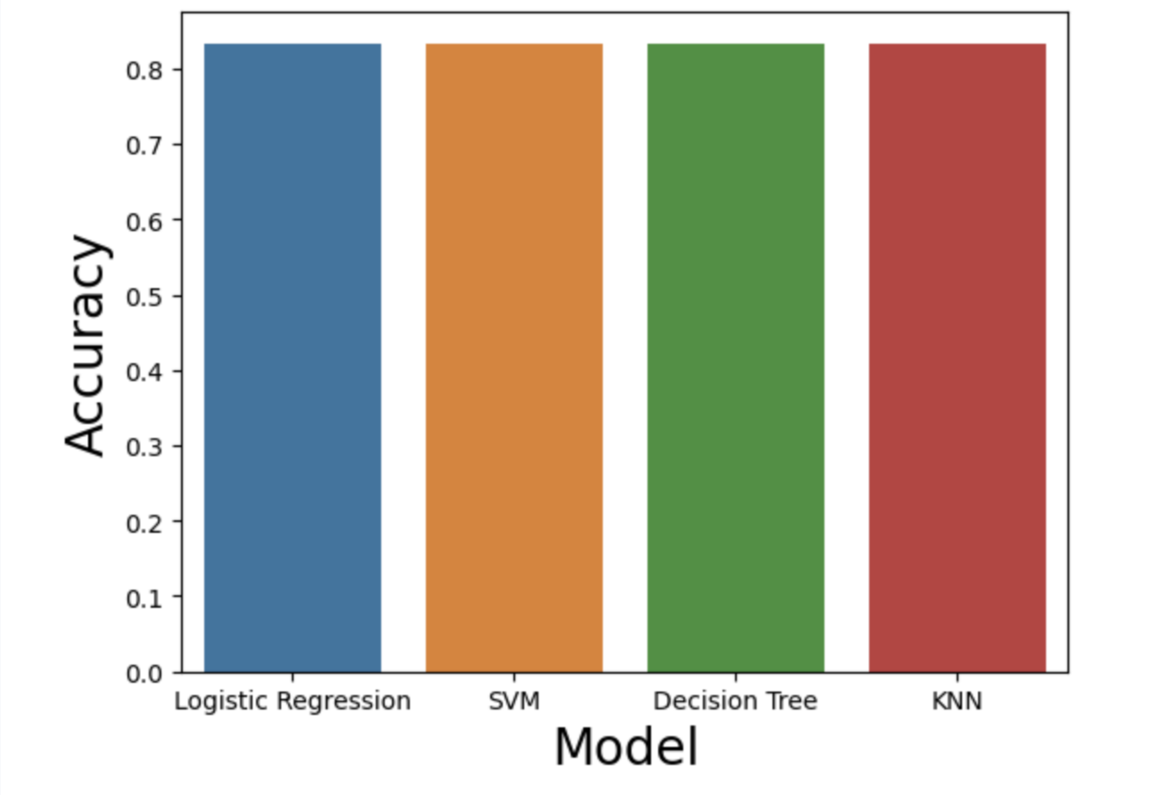


Booster Version FT has the most successful outcomes

Section 5

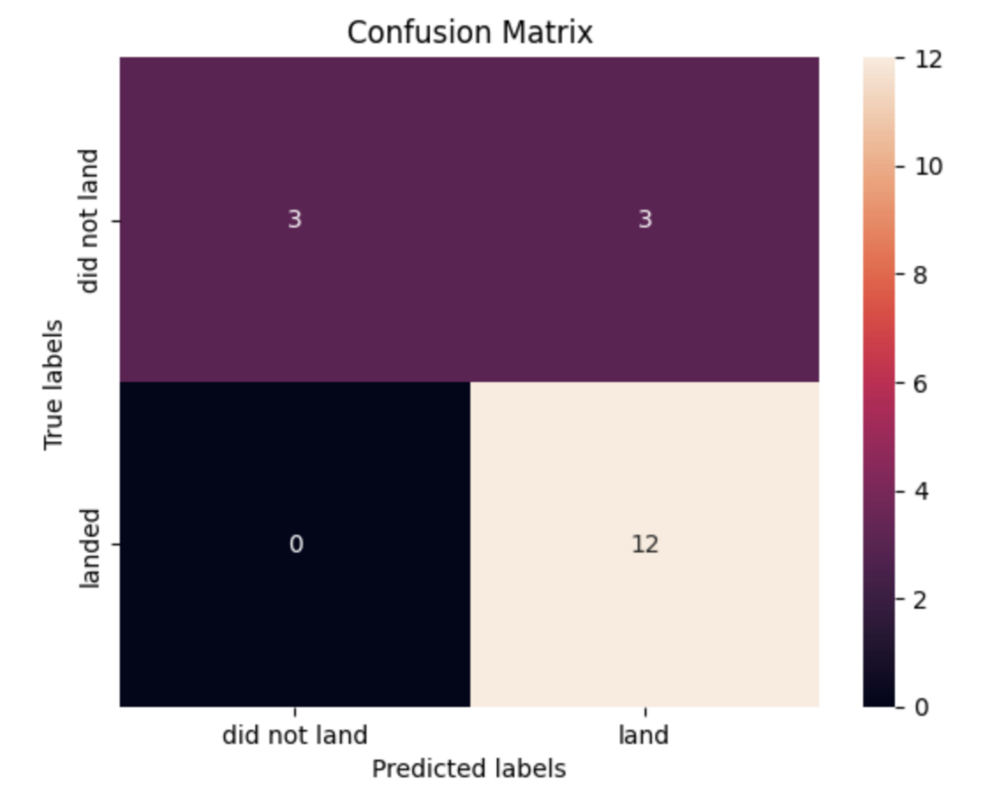
Predictive Analysis (Classification)

Classification Accuracy



- All Models have the same accuracy with regard to the give test data

Confusion Matrix



- KNN Model: out of 15 predicted landings 12 actually landed, out of 3 predicted unsuccessful landings all 3 did not land

Conclusions

- Success Rate improved over time starting in 2013, with a small decrease in 2018
- There is a learning effect as higher flight numbers are correlated with higher success rates. Higher flightvnumbers also means that there was more time to learn from previous failure
- Higher Payloads have higher success rates
- The Orbits GEO, SSO, ES-L1, HEO, VLEO have the highest success rates
- KSC-LC-39A has the most of the successful launches
- All Prediction Models showed the same accuracy with respect to the test data

Appendix

- https://github.com/ralph240574/datascience-capstone/

Thank you!