IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eric Tang
November 13, 2024

Outline

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

Executive Summary

- Methodologies
  - We use the SpaceX API to extract information about their rocket launch missions
  - Based on data acquired via JSON and SQL requests, we visualize success rate trends and create a prediction model
  - We use Folium and Plotly to visualize launch sites and success rates with respect to each site and payload size
- Results Summary
  - We have created a model to predict the success rate of a Falcon 9 mission based on many input parameters with 83% test accuracy
  - We have found that the KSC LC-39A launch site has the highest succewss rate (76.9%), but has no success with payloads over 5500 kg
  - The FT booster version has highest success rate, while the v1.1 booster has the lowest

Introduction

- The first stage of rocket launches is extremely costly. Landing of the first stage is important for saving future costs as the rocket can be reused for later launches.
- We are interested in determining which factors can be used to predict whether or not the first stage will land as well as identify important variables related to success rate in order to better understand what conditions will need to a successful landing.

Section 1

Methodology

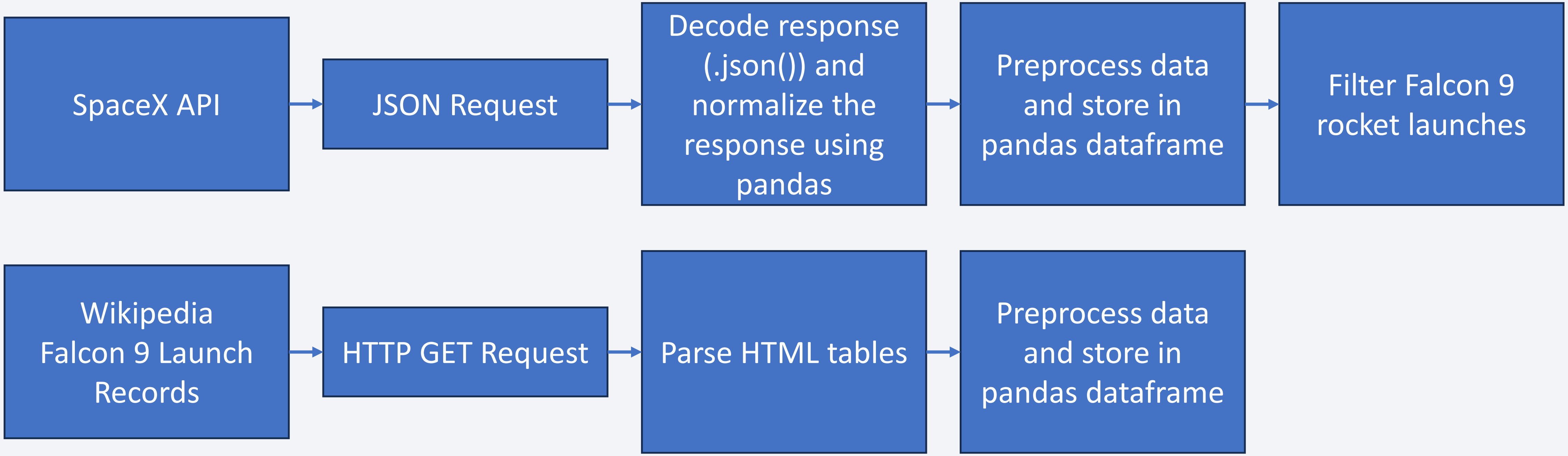
Methodology

Executive Summary

- Data collection methodology:
  - The data was collected via JSON and HTML requests from the SpaceX API and Wikipedia
- Perform data wrangling
  - The data was processed using pandas dataframes for ease of use
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We created multiple regression models and evaluated their performance on training and test data

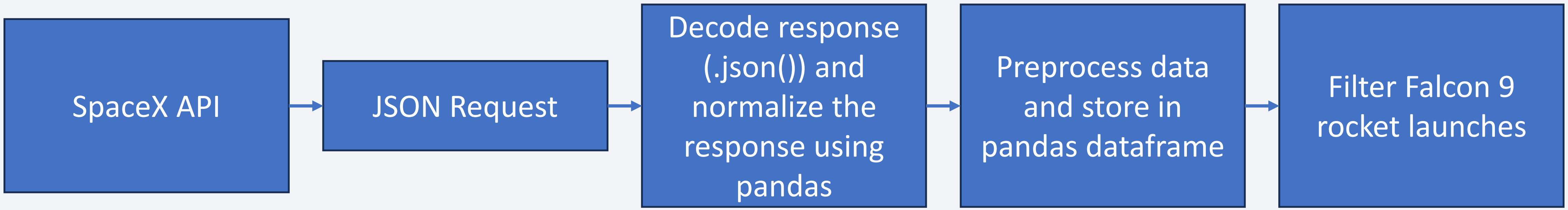
Data Collection

- The data was collected via JSON and HTML requests from the SpaceX API and Wikipedia



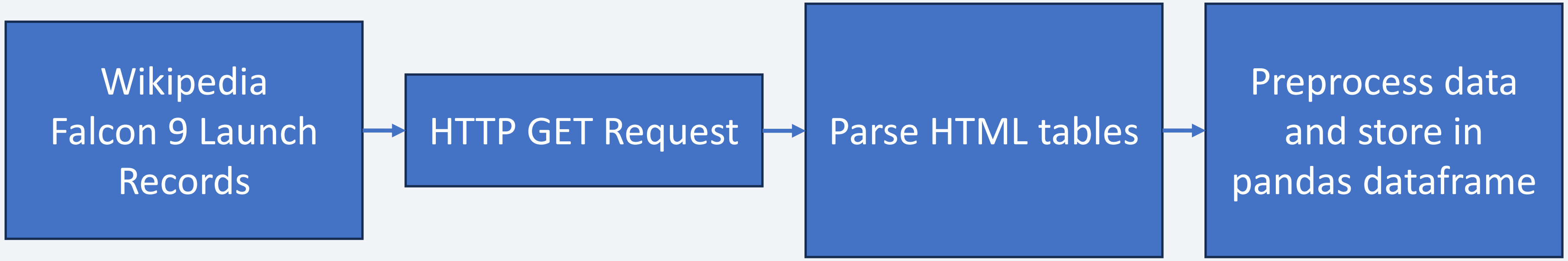
Data Collection – SpaceX API

GitHub URL



Data Collection - Scraping

GitHub URL



Data Wrangling

- In order to build a predictive model, we convert landing outcomes to binary results with either success or failure and add these values as a new ‘Class’ column in our dataset
- GitHub URL

Import CSV data generated in data collection step (Falcon 9 launches) → Convert landing outcomes to binary success (1) or failure (0) → Add binary outcome ‘Class’ to pandas dataframe → Export as CSV

EDA with Data Visualization

1. We looked at the relationship between flight number and launch site with respect to success. This showed that the KSC LC 39A had the highest success rate.

2. Next we looked at payload vs. launch site, which showed high success at CCAFS SLC 40 for higher payloads (> 12000 kg) and high success at VAFB SLC 4E for lower payload (< 10000 kg)

3. We plotted the average success rate for each orbit type and number of flights per orbit type to determine which orbit type is best

4. We also looked in payload vs. orbit to determine which orbit is best for different payload sizes

[GitHub URL]

EDA with SQL

1. Read SpaceX CSV data in pandas dataframe
2. Convert dataframe to SQL database
3. SQL: Display unique launch sites and total payload mass carried by CRS boosters
4. SQL: Display average payload mass carried by booster version F9 v1.1
5. SQL: Find the first successful ground pad landing date
6. SQL: List boosters with successful drone ship landings with payloads between 4000 and 6000 kg
7. SQL: List the total number of successful and failure mission outcomes
8. SQL: Find the booster version used for the maximum payload launch
9. SQL: List drone ship landing failures in 2015
10. SQL: Count the number of landing outcomes between 2010-06-04 and 2017-03-20

GitHub URL

Build an Interactive Map with Folium

- Mark all launch sites on the map to see whether geographical factor affect launch site locations
- Create marker clusters for each launch site to show success and failures
- Calculate the distance between a launch site and the coastline/highway to understand more about launch site location

[GitHub URL]

Build a Dashboard with Plotly Dash

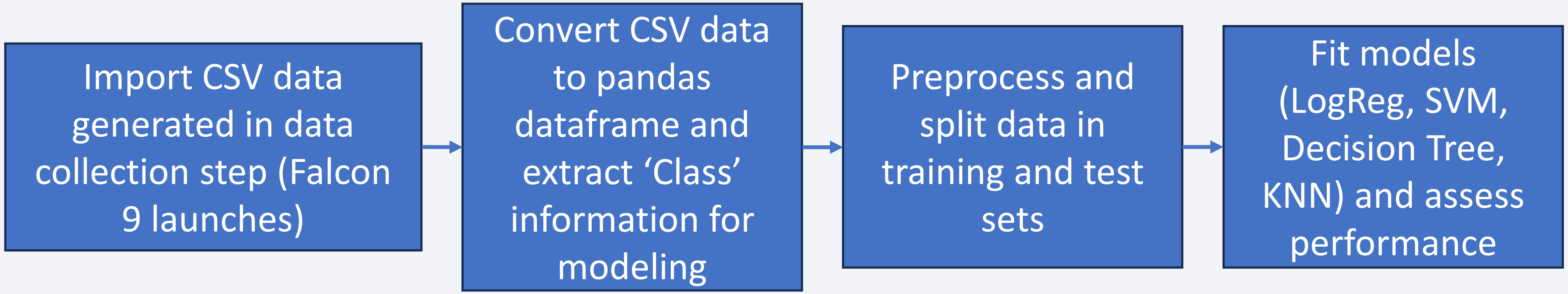
- We created a pie chart to show the distribution of launch successes for each site
- We also added a scatterplot and slider for payload range in order to see how payload and booster version affects launch success for each site

GitHub URL

Predictive Analysis (Classification)

- We extract Falcon 9 launch data by reading a CSV file into a pandas dataframe
- For our target variable, we extract the launch outcome ‘Class’ variable and split data into training and test sets
- We explore a number of models and train each to see which model generates the best results for predicting launch outcome

GitHub URL



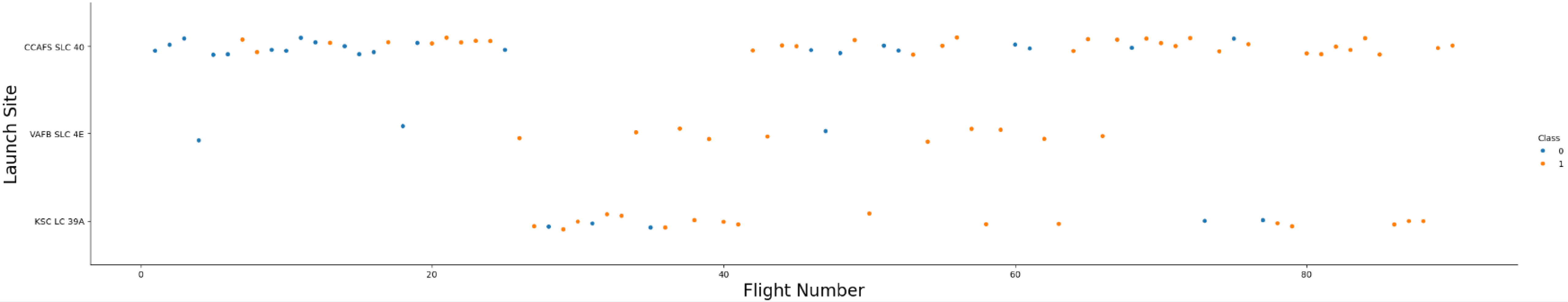
Results

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

Section 2

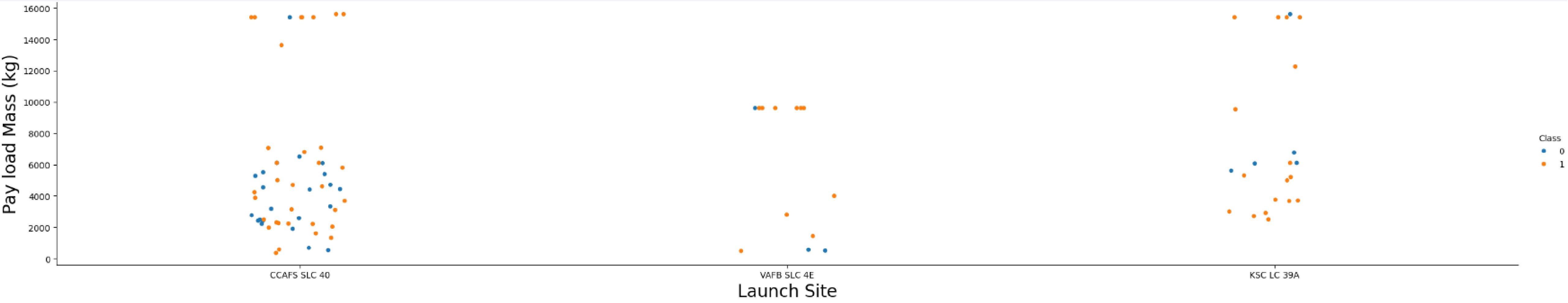
Insights drawn from EDA

Flight Number vs. Launch Site



- Here we can see most of the flights occur at CCAFS SLC 40 launch site
- Fewer flights occur at VAFB SLC 4E, but they have a high success rate

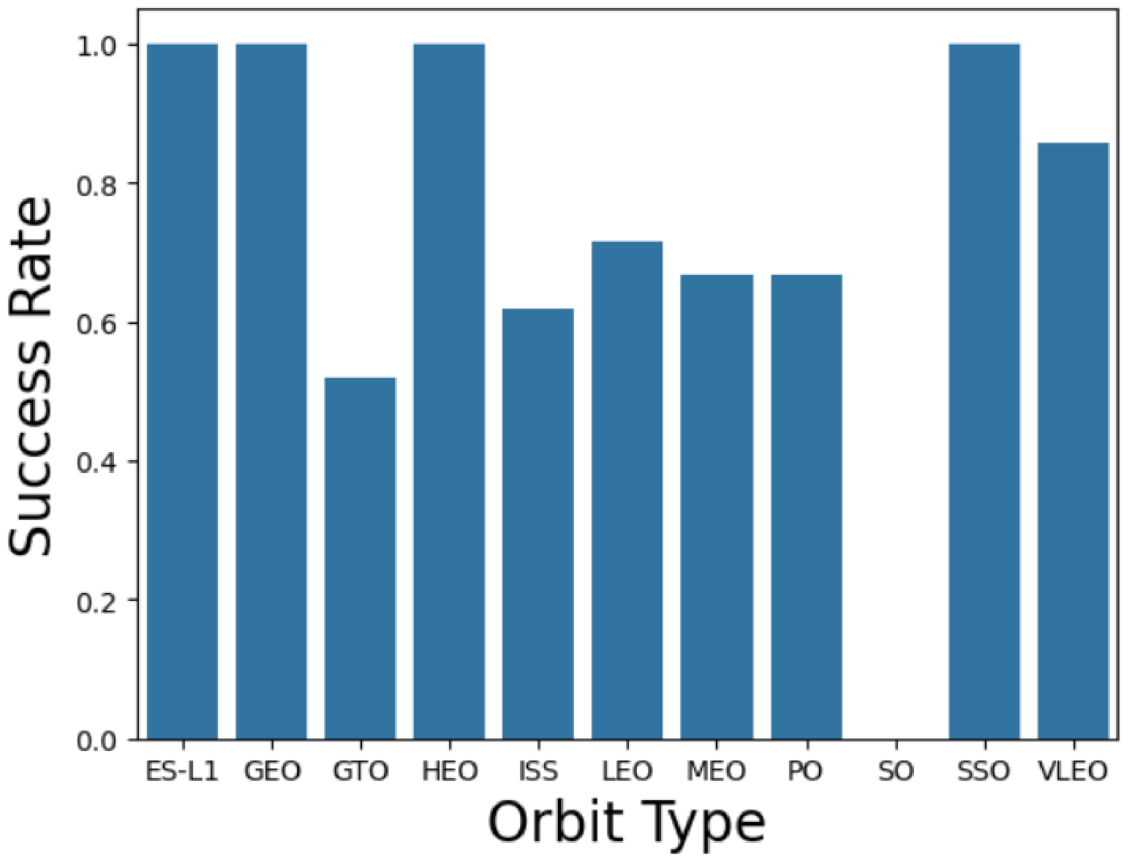
Payload vs. Launch Site



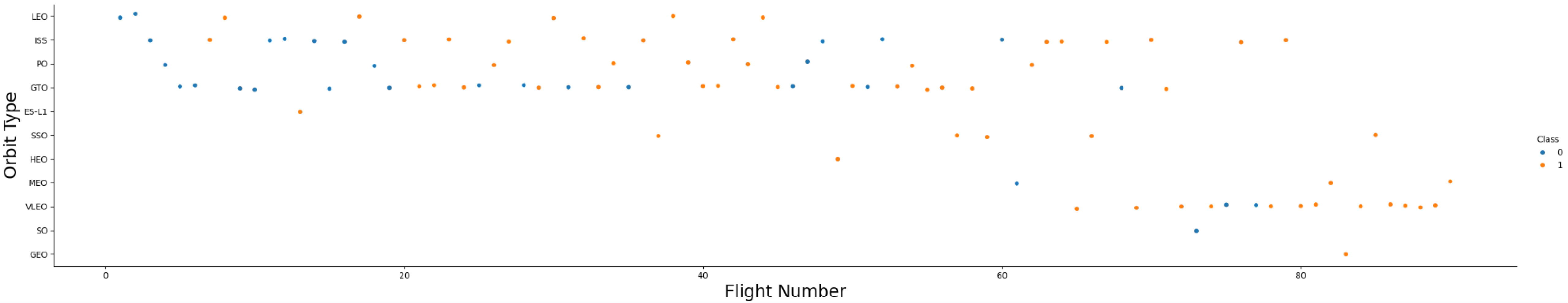
- At CCAFS SLC 40, there is a high success rate for payloads > 12000 kg
- At VAFB SLC 4E, there is a high success rate for 10000 kg payloads
- At KSC LC 39A, there is a relatively high success rate regardless of payload mass

Success Rate vs. Orbit Type

- We can see that the SO orbit type has the lowest success rate, whereas ES-L1, GEO, HEO, and SSO all have high success rates

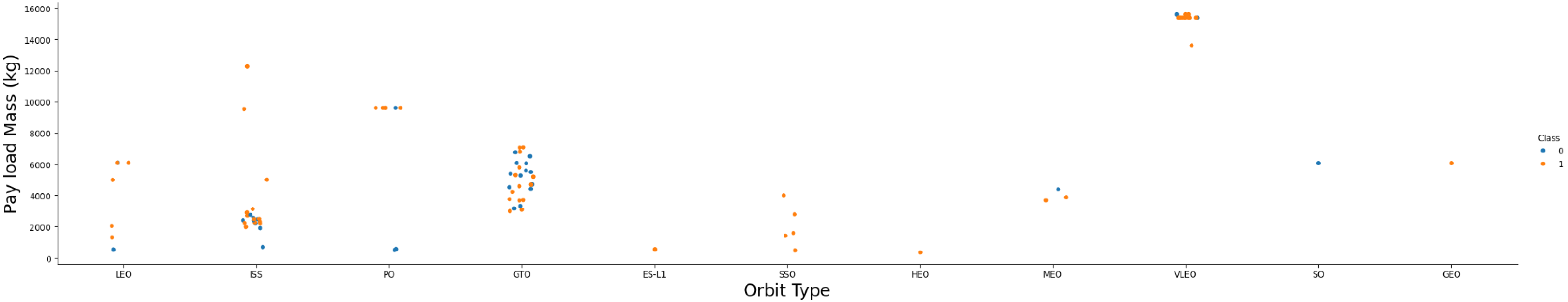


Flight Number vs. Orbit Type



- We can see that the ES-L1, GEO, HEO, and SSO orbit types despite having high success rate each have very few number of flights

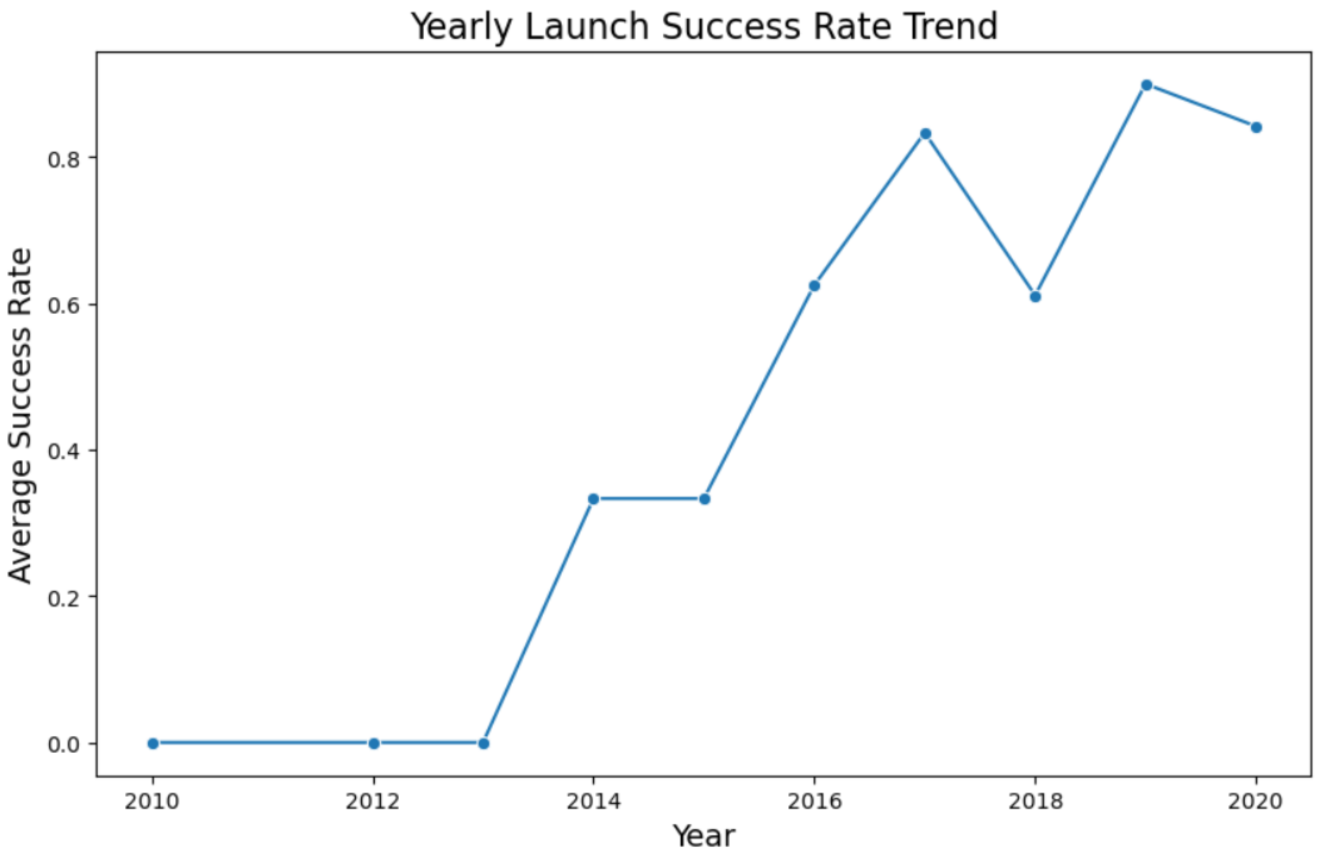
Payload vs. Orbit Type



- We can see that for high payloads > 12000 kg, a VLEO orbit has a high success rate
- For low payloads < 6000 kg, an SSO orbit has a high success rate
- For intermediate payloads of 10000 kg, a PO orbit has a high success rate

Launch Success Yearly Trend

- We can see that initially launches were unsuccessful, but have started to become more successful as time goes on
- This makes sense due to having more launch data and more technological development



All Launch Site Names

- We query the unique launch sites from the database

```
%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

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

- We can search for the ‘CCA’ substring to find 5 records

Total Payload Mass

- We can sum the payload mass for all boosters from CRS by using a substring search

```
%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where "Payload" like '%CRS%'

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

sum(PAYLOAD_MASS__KG_)
111268

Average Payload Mass by F9 v1.1

- By searching for the F9 v1.1 substring, we can find and average all corresponding payload masses

```
%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" like '%F9%v1.1%'

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

avg(PAYLOAD_MASS__KG_)
6138.287128712871

First Successful Ground Landing Date

- We can use the minimum function to find the least recent date for a successful ground pad landing

```
%sql select min("Date") from SPACEXTBL where "Landing_Outcome" like '%success%ground%pad%'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select "Payload" from SPACEXTBL where "Landing_Outcome" like '%success%drone%ship%' and "PAYLOAD_MASS__KG_" between 4000 and 6000
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

- We can use two conditions to search for successful drone ship landings and payload masses between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

- We can use the COUNT function to count the number of success or failure missions by group data by the mission outcome

```
%sql select "Mission_Outcome",Count(*) from SPACEXTBL group by "Mission_Outcome"
```

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

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

Boosters Carried Maximum Payload

- By using a subquery, we can isolate the data that has the maximum payload from the database and then select the booster version

```
%sql select "Booster_Version" from (select *, max("PAYLOAD_MASS__KG_") from SPACEXTBL)

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

Booster_Version
F9 B5 B1048.4

2015 Launch Records

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

```
 * 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
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Here, we can list information by searching for launches in the year 2015 whose outcome was a drone ship landing failure

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

```
%sql select "Landing_Outcome",Count(*) from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by "Landing_Outcome" order by Count(*) desc
```

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

Landing_Outcome	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

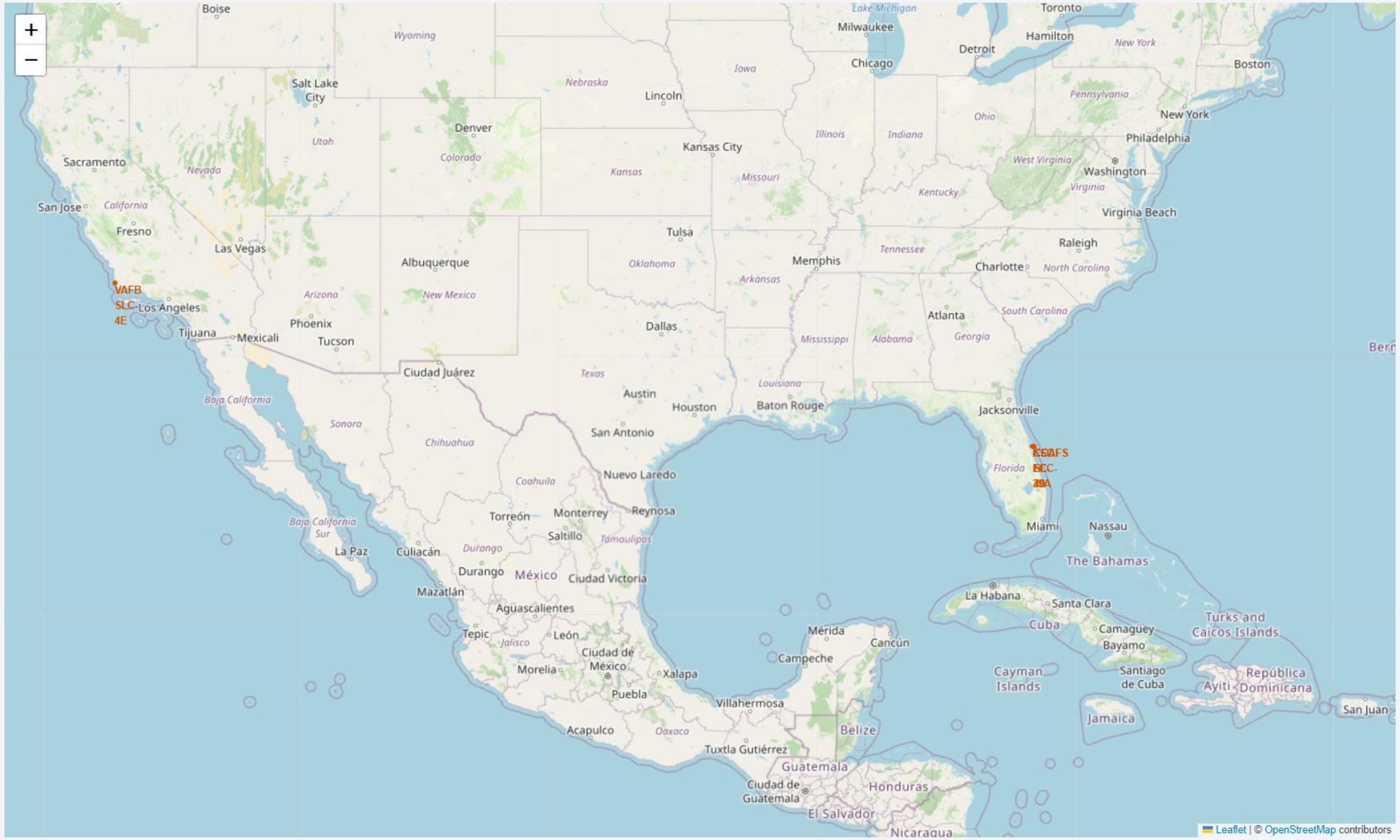
- Here we can identify and count the number of each outcome by group the data by landing outcome and filtering results between 2010-06-04 and 2017-03-20 in descending order

Section 3

Launch Sites Proximities Analysis

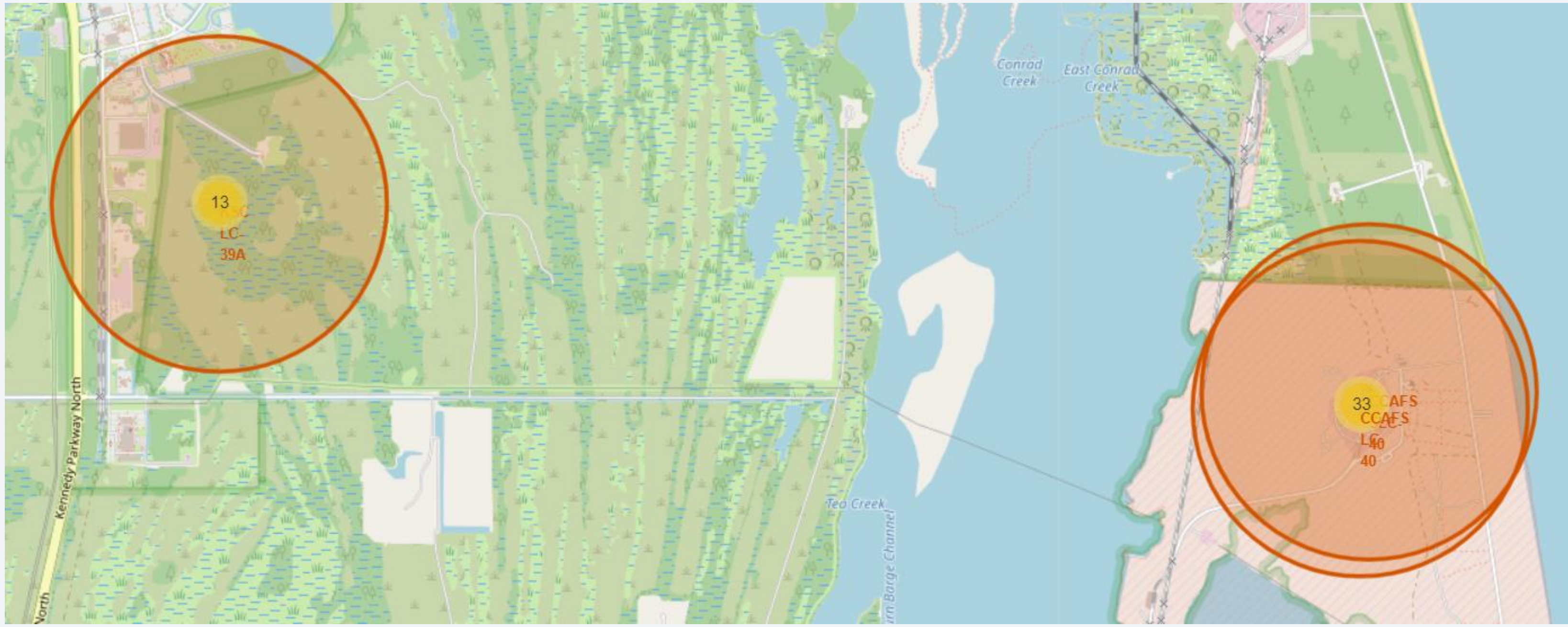
Folium: Launch Site Locations

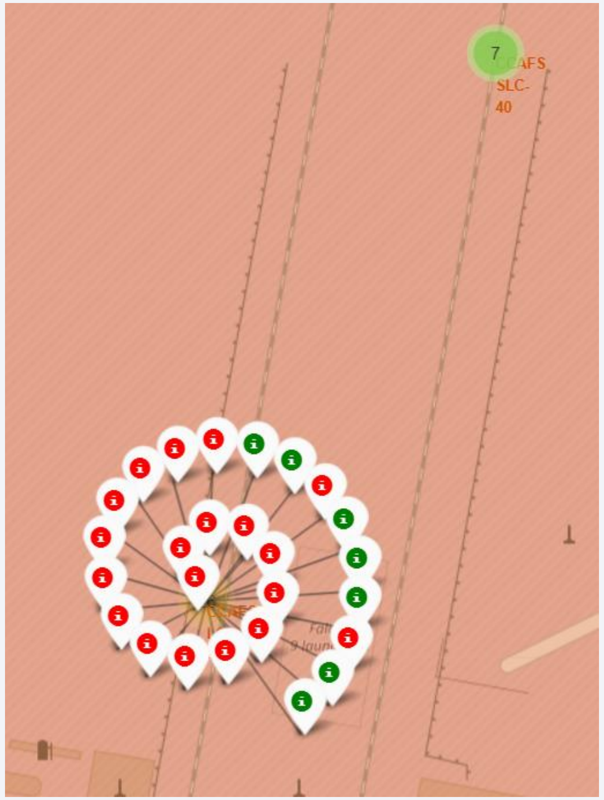
- Here, we can see launch sites are located extremely close to the coastline



Folium: Launch Site Flight Outcome

- We can easily visualize the total number of launches as well as the number of successes and failures from the interactive map





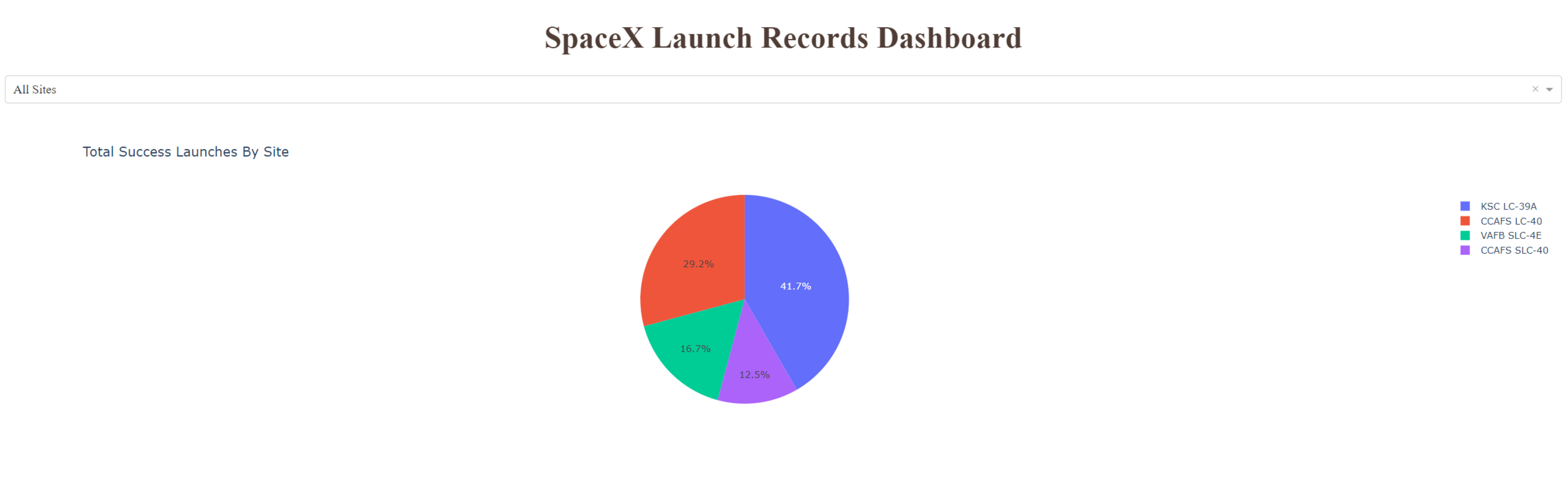
Folium: Launch Site Proximity

- Here we can see that launch sites are located very close to the coastline (likely in the case of mission failure) and close to major transportation routes for ease of accessing materials

Section 4

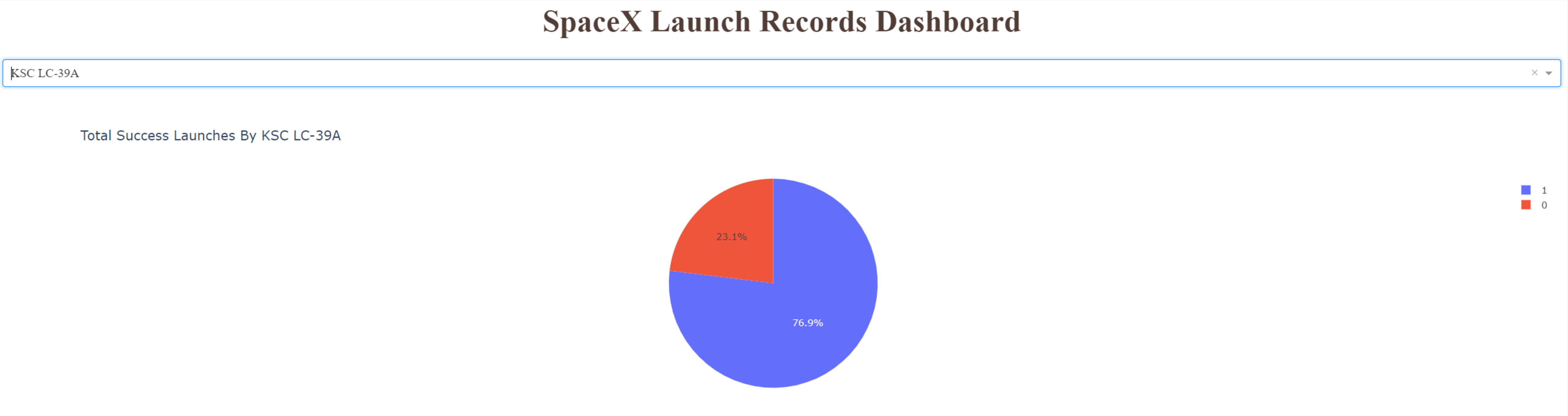
Build a Dashboard with Plotly Dash

Plotly Dash: Total Success Launches



- Here we can easily see that the KSC LC-39A sites has the highest number of successful flights amongst all sites

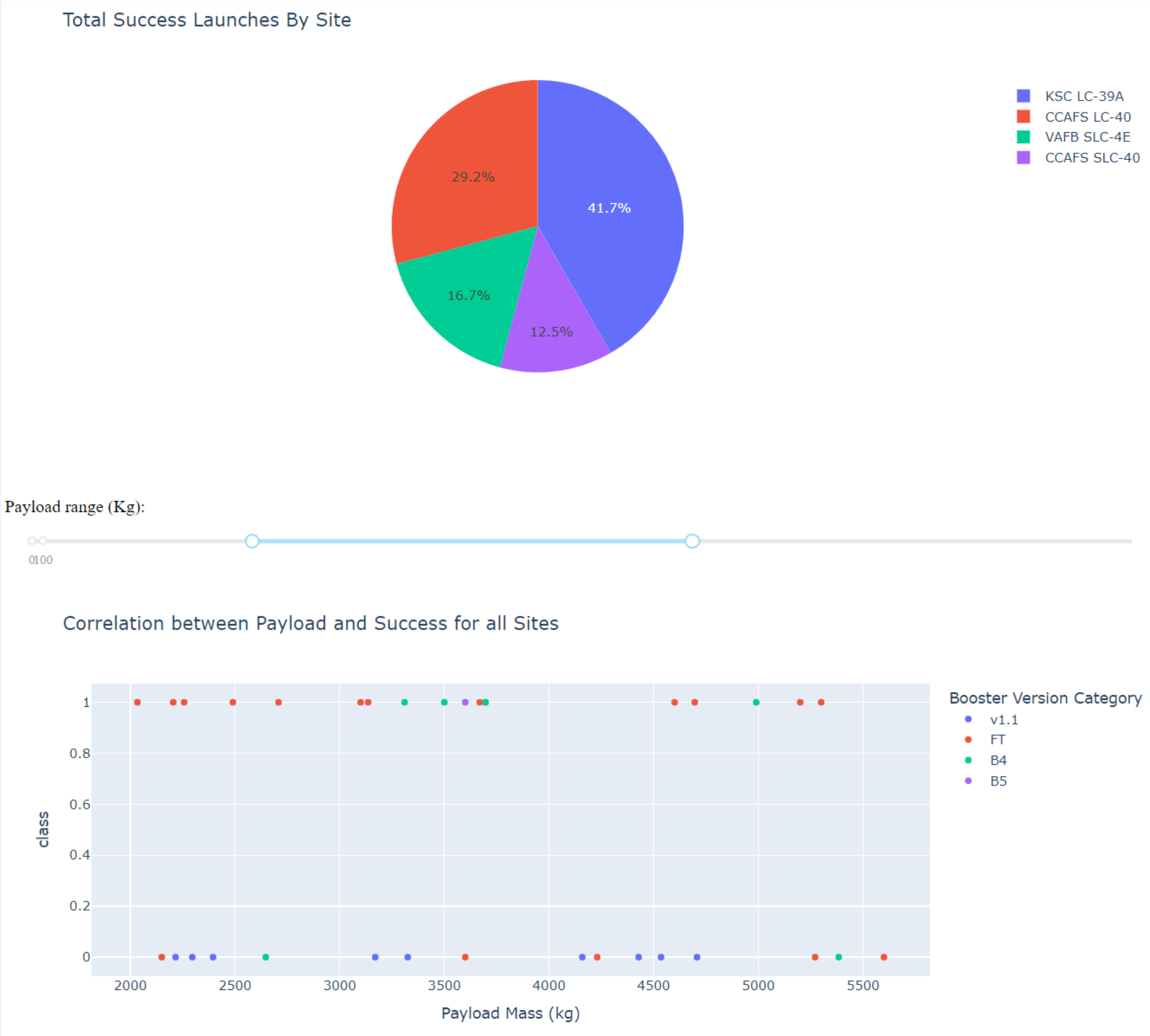
Plotly Dash: Total Success Launches by KSC LC-39A



- Specifically, the KSC LC-39A has a 77% success rate

Plotly Dash: Launch Success vs. Payload Mass

- We can see that many factors play a role in whether or not a launch is successful, including the payload mass as well as the type of booster

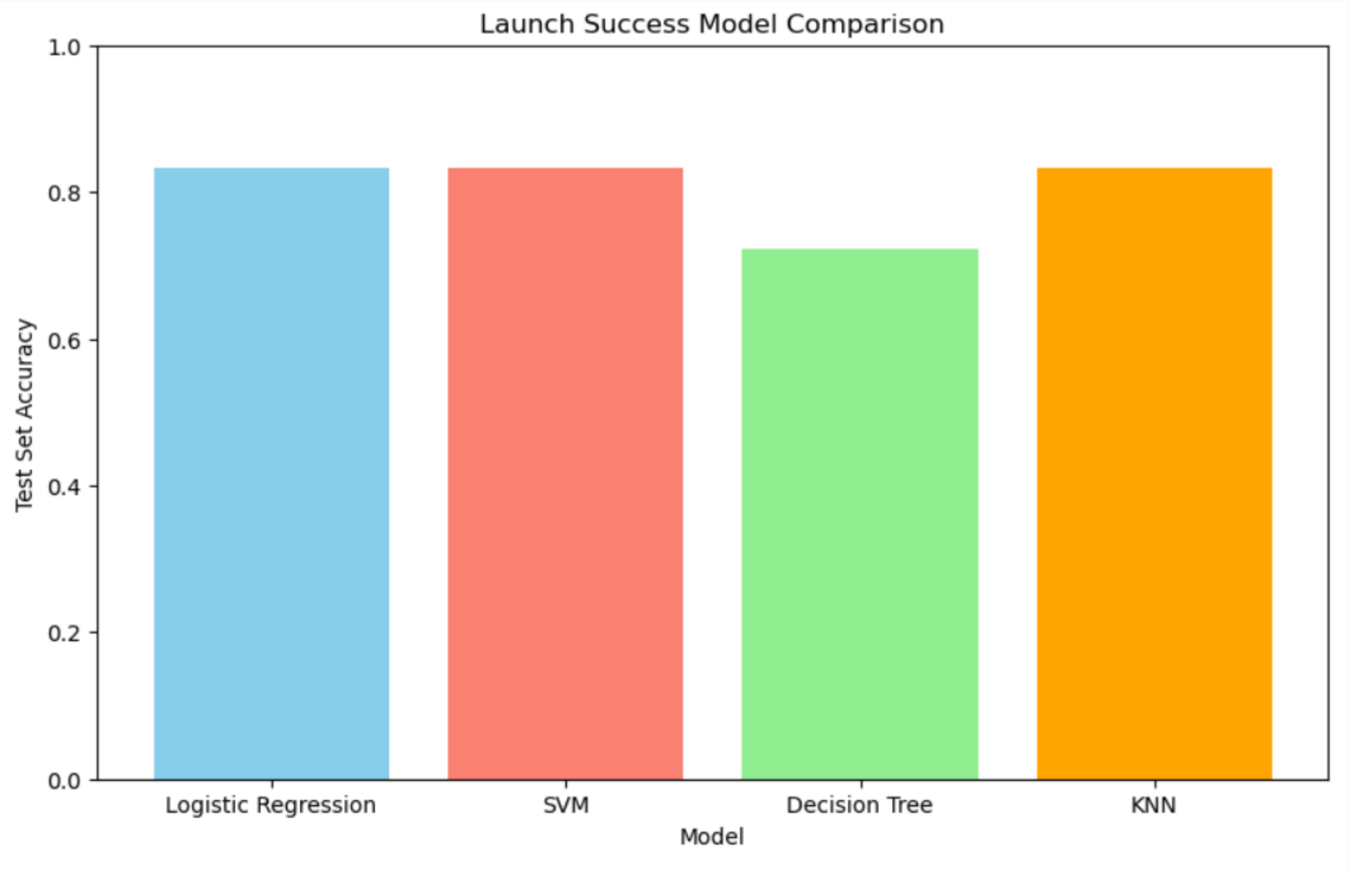


Section 5

Predictive Analysis (Classification)

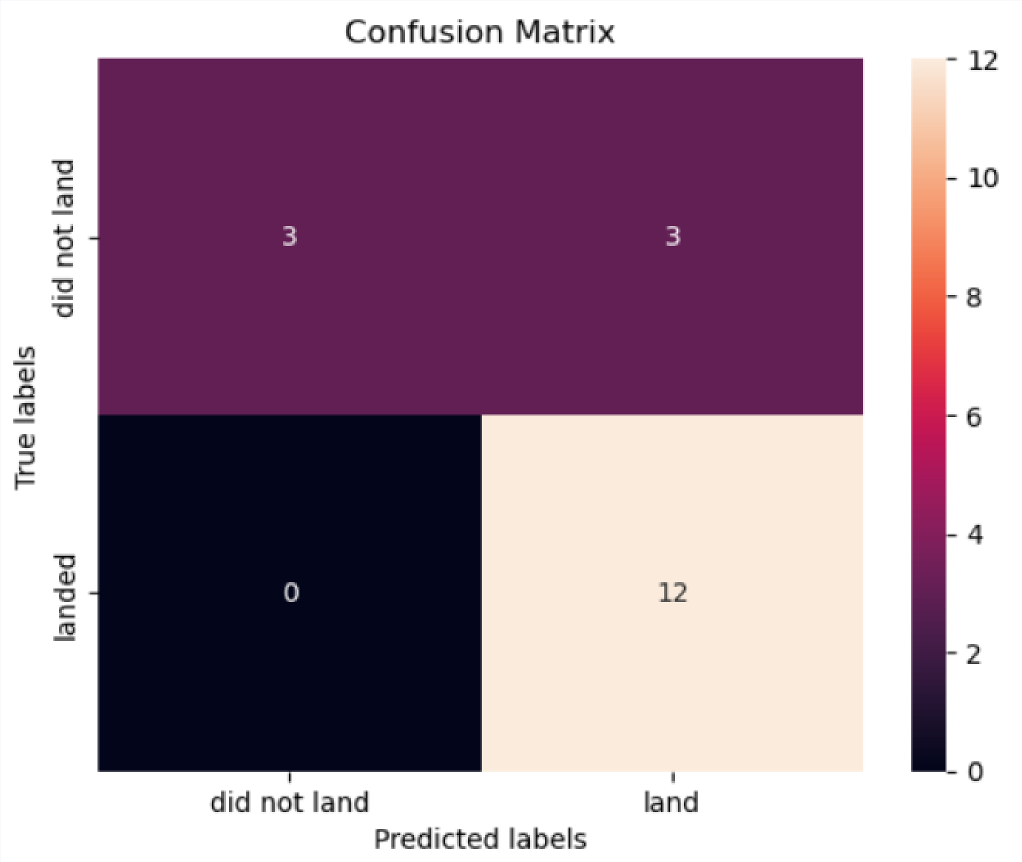
Classification Accuracy

- Based on the test set results, logistic regression, SVM, and KNN all perform the best (83% accuracy) while the decision tree classifier performed the worst (72% accuracy)



Confusion Matrix (KNN)

- For the KNN model, 15 test points were correctly predicted as successes or failures
- However, there were 3 false positives where the model predicted a success but the rocket did not actually land



Conclusions

- In conclusion, a number of factors (such as payload mass and booster version, launch site) determine whether or not a landing will be a success or failure
- By using a database of past launch information, we created a model that can predict whether or not a launch will be a success up to 83% accuracy

Appendix

- Project GitHub URL

Thank you!