IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Regina Hornung
December 26, 2023

Outline

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

Executive Summary

- The research seeks to predict whether the Falcon 9 first stage will land successfully, thereby allowing the SpaceX program to reuse the first stage in a future rocket launch.
- Summary of methodologies
  - Collect data with SpaceX API
  - Wrangle Data with success/fail variable from Falcon 9 launch records
  - Explore the data in terms of launch site, orbit, payload mass, version booster, and booster landing
  - Analyze the data with SQL to calculate statistics on number of successes and failures
  - Explore launch site statistics and locations
  - Visualize successful launch data
  - Build models using logistic regression, support vector machine, decision tree, and K nearest neighbor
- Summary of all results
  - All models performed with similar results, with the decision tree slightly outperforming the other models.

Introduction

- Background
  - We seek to predict if the Falcon 9 first stage will land successfully. SpaceX claims their Falcon 9 rocket launches at an expense of 62 million dollars; other providers show a cost of up to 165 million dollars each. The difference in cost is due to that fact that SpaceX can reuse the first stage. Therefore if we can predict if the first stage will land, we can determine the expense of a launch.
- Problems to Explore
  - How much do launch site location, orbits, payload mass, and number of flights affect landing success?
  - Have the rate of successful landings improved over time?
  - What is the best model to predict for successful landings?

Section 1

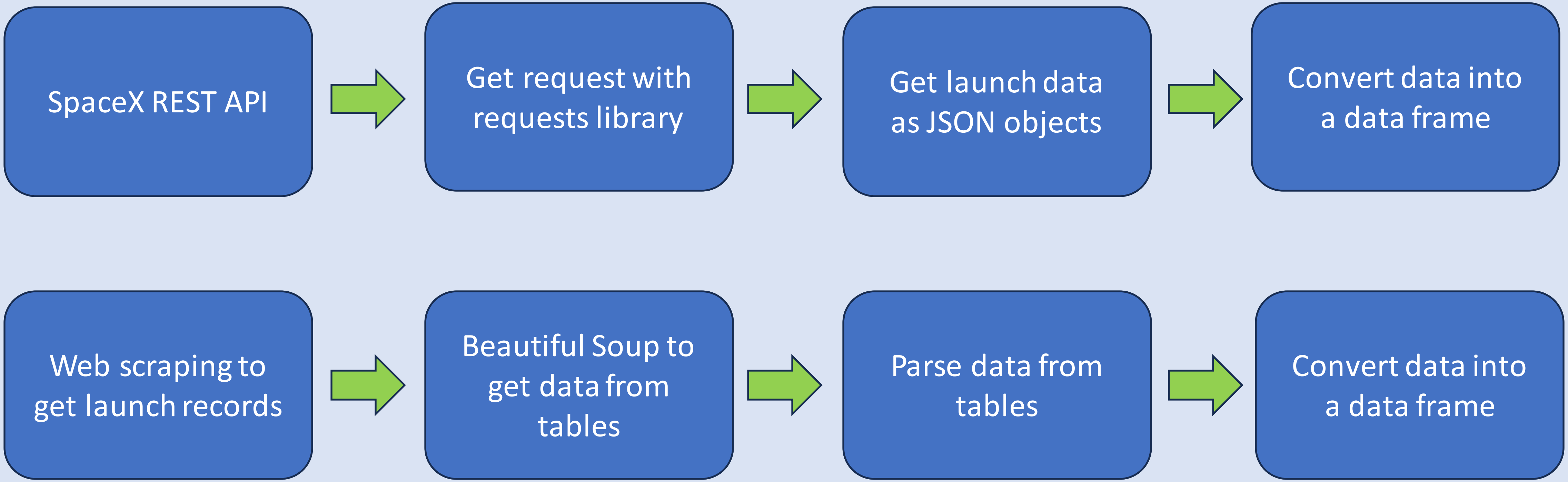
Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - SpaceX API and web scraping techniques
- Perform data wrangling
  - Filter the data and handle missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tune and evaluate models to find the best models and parameters

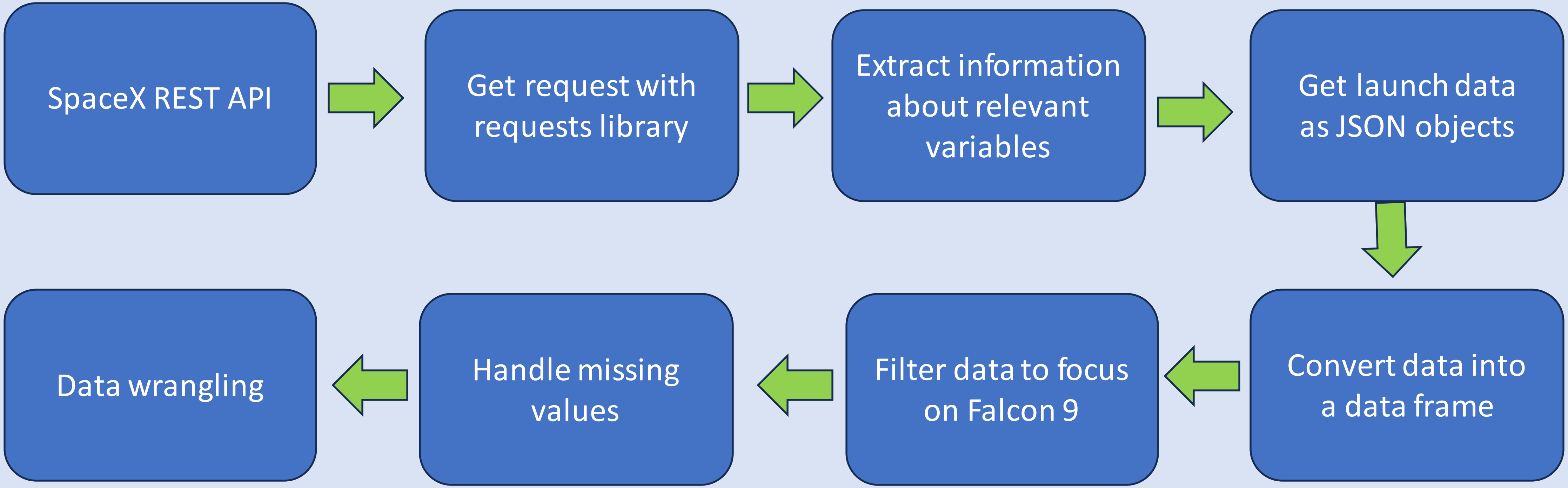
Data Collection

The data was retrieved from SpaceX REST API and by using web scraping techniques on wiki pages.



Data Collection – SpaceX API

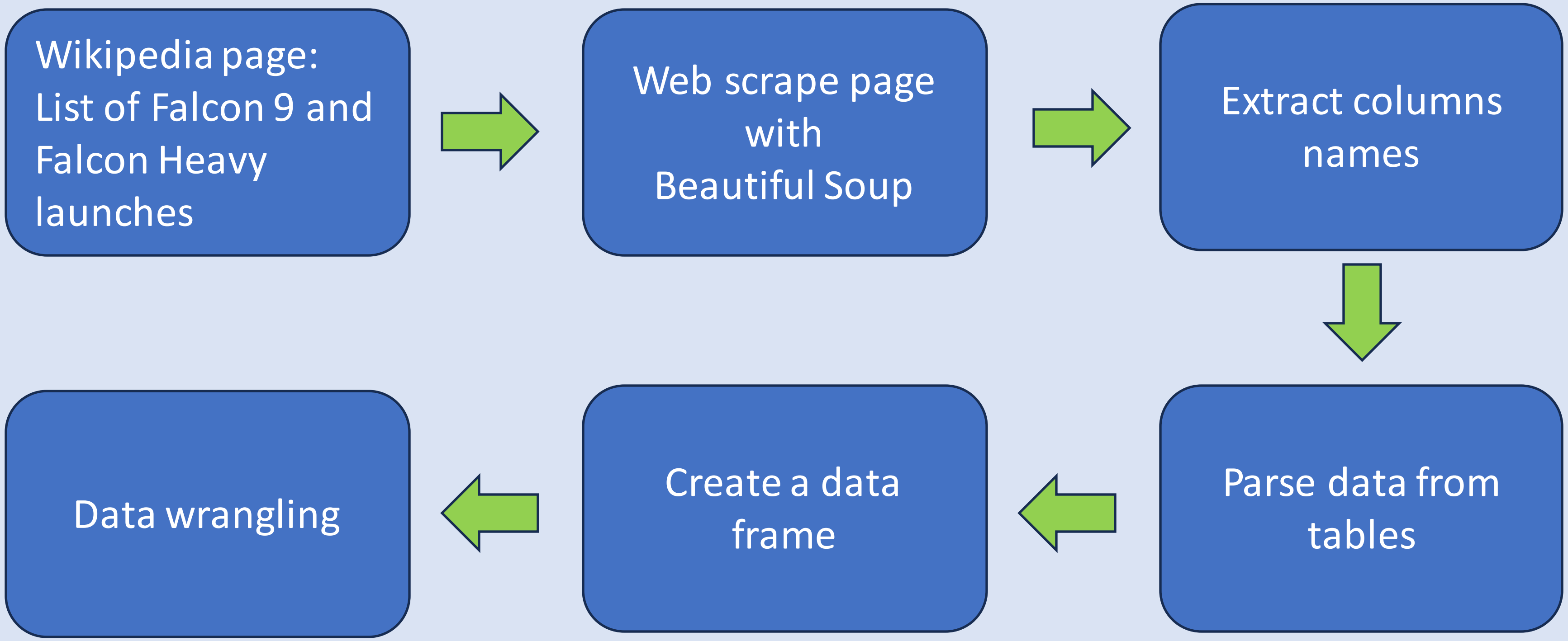
`Use API to collect data add ensure that it is in the desired format.`



- GitHub URL of the completed SpaceX API calls notebook
- https://github.com/ghornung/1-SpaceX-Capstone.git

Data Collection – Scraping

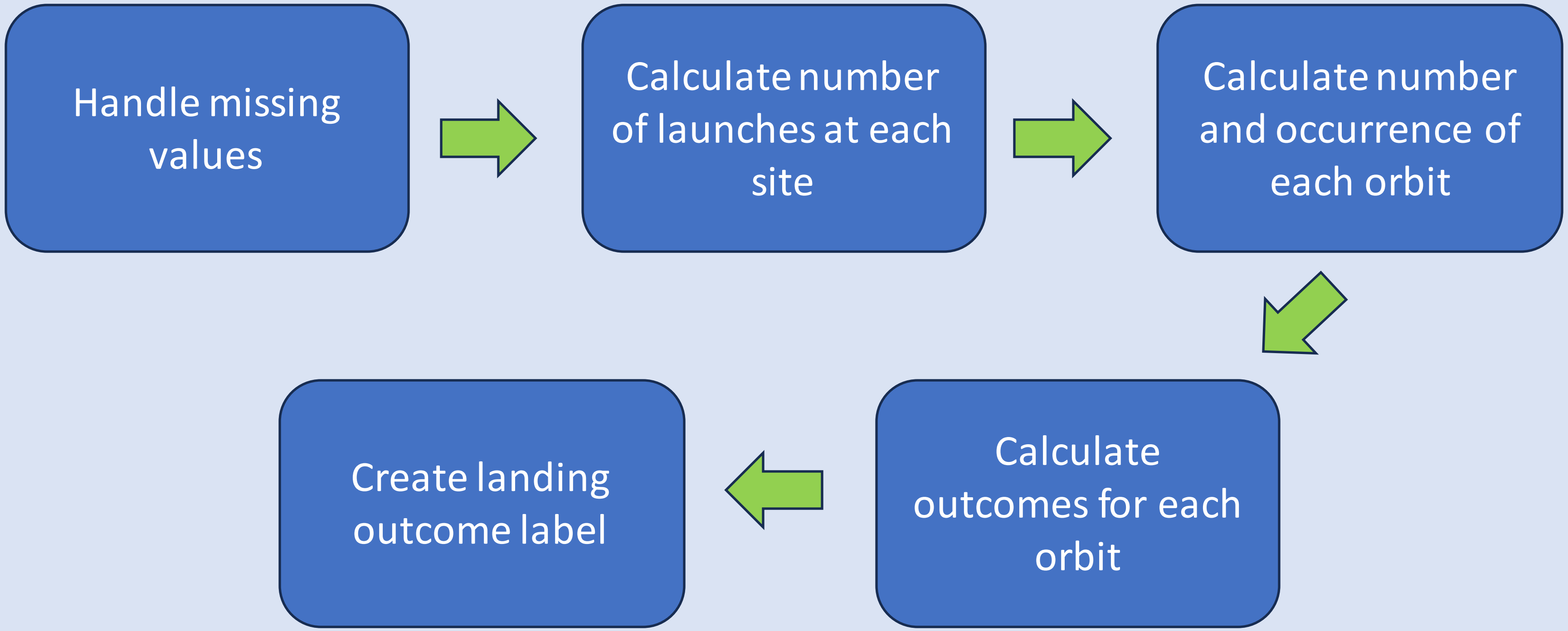
`Use web scraping to collect Falcon 9 launch data from wiki pages.`



- GitHub URL of the completed web scraping notebook
- https://github.com/ghornung/2-SpaceX-Capstone.git

Data Wrangling

`We find patterns in the data and determine the labels for training supervised models.`



- GitHub URL of completed data wrangling related notebook
- https://github.com/ghornung/3-SpaceX-Capstone-.git

EDA with Data Visualization

- Charts plotted:
  - Flight number vs payload mass
  - Launch site vs flight number
  - Payload mass vs launch stie
  - Orbit vs class
  - Orbit vs flight number
  - Payload mass vs orbit
  - Success rate by year

- GitHub URL of completed EDA with data visualization notebook
- https://github.com/ghornung/5-SpaceX-Capstone.git

EDA with SQL

- SQL queries
  - Names of the unique launch sites
  - Launch sites that begin with CCA
  - Total payload mass carried by boosters
  - Average payload mass carried by boosters
  - Date of first successful landing on ground pad
  - Boosters which have had success in drone ship
  - Total number of successes and failures
  - Names of boosters with maximum payload mass
  - Rank count of landing outcomes in descending order
- GitHub URL completed EDA with SQL notebook
- https://github.com/ghornung/4-SpaceX-Capstone.git

Build an Interactive Map with Folium

- Summary of map objects created and added to folium map
  - `folium.Circle` - adds a highlighted circle area with a text label on a specific coordinate
  - `folium.Marker` - marks each launch site on the map
  - `MarkerCluster` object – marks several objects with the same coordinates
  - `MousePosition` - gets coordinate for a mouse over a point on the map.
  - `PolyLine` – indicates distance between a launch site to a selected coastline point or to another point

- GitHub URL completed interactive map with Folium map
- https://github.com/ghornung/6-SpaceX-Capstone.git

Build a Dashboard with Plotly Dash

- Summary of plots/graphs and interactions added to a dashboard
  - This dashboard contains a dropdown list and slider to interact with pie charts and scatter point charts.
    - The drop down menu allows four different launch sites to be selected
    - A pie chart is rendered based on the launch site selected and the number of successful launches
    - A slider is included to identify patterns in different payloads.

- GitHub URL of completed Plotly Dash
- https://github.com/ghornung/7-SpaceX-Capstone.git

Predictive Analysis (Classification)

- Summary of building, evaluating, improving, and finding the best performing classification model

NumPy array from the column `Class` in `data`

Standardize the data in X

Split the data X and Y into training and test data

Create a logistic regression object

Compare the accuracies to find the best model

Create a k nearest neighbors object

Create a decision tree classifier object

Create a support vector machine object

- GitHub URL of completed predictive analysis lab
- https://github.com/ghornung/8-SpaceX-Capstone.git

Results

Exploratory data analysis results

- Launch success rate has improved over time.
- KSC LC 39A is the landing site with the highest success rate.
- Orbit types ES-L1, GEO, HEO, and SSO have a 100% success rate

Interactive analytics demo in screenshots

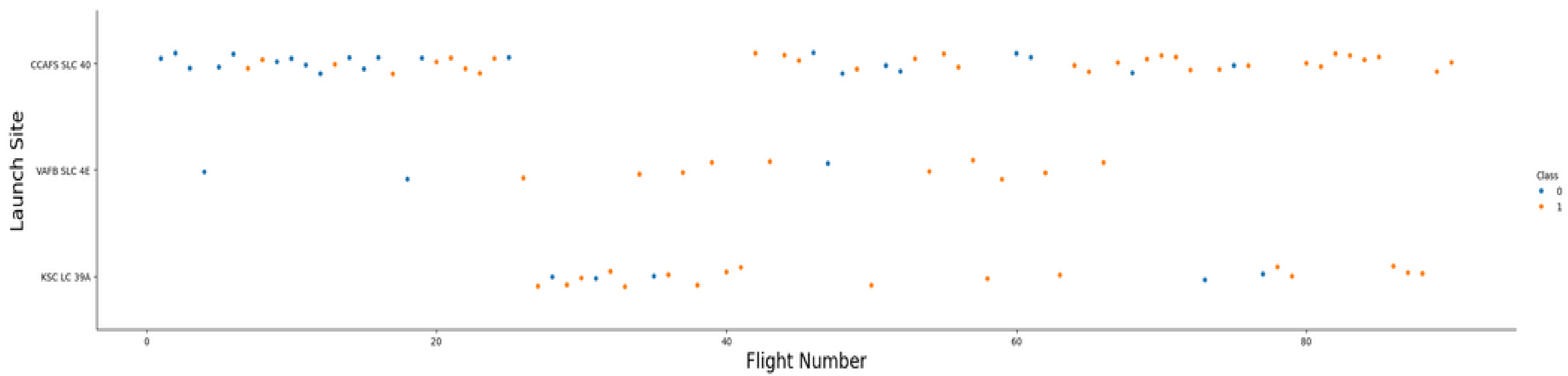
- Launch site tend to be located relatively near the equator and close to the coast.

Predictive analysis results

- The decision tree model is the best suited for the data set.

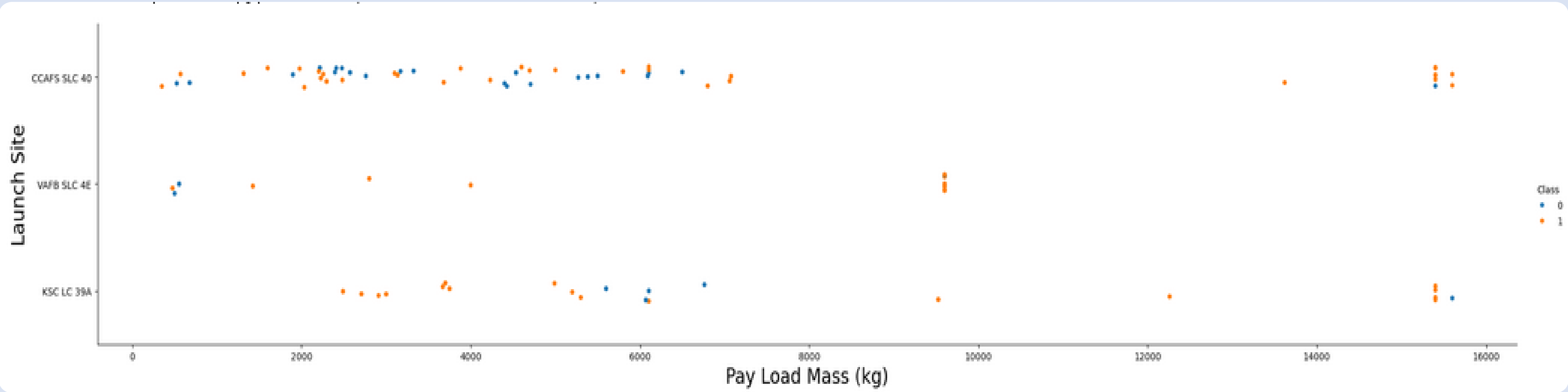
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

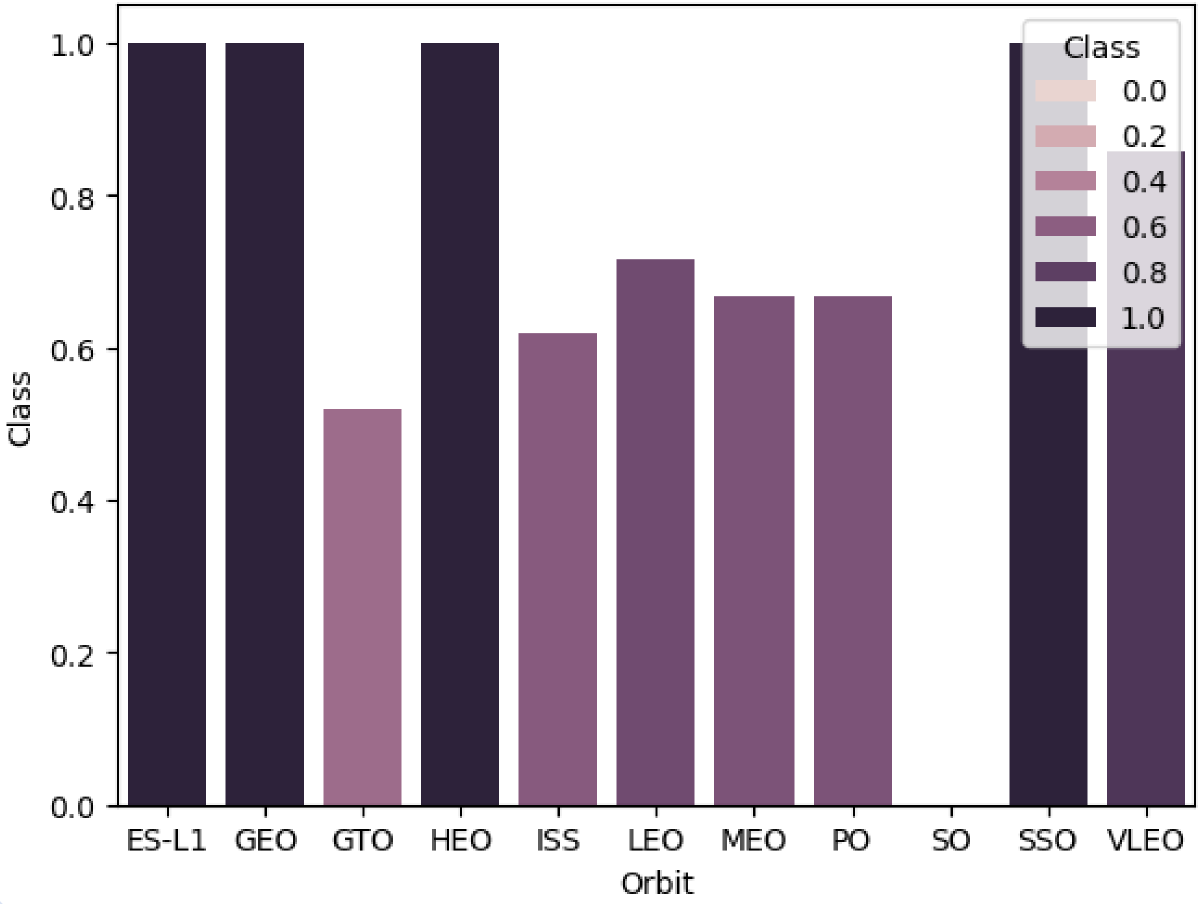
- Over time, the rate of successful launches has increased at all sites.
- Site CCAFS SLC-40 in particular shows a high successful landing rate as time has progressed.



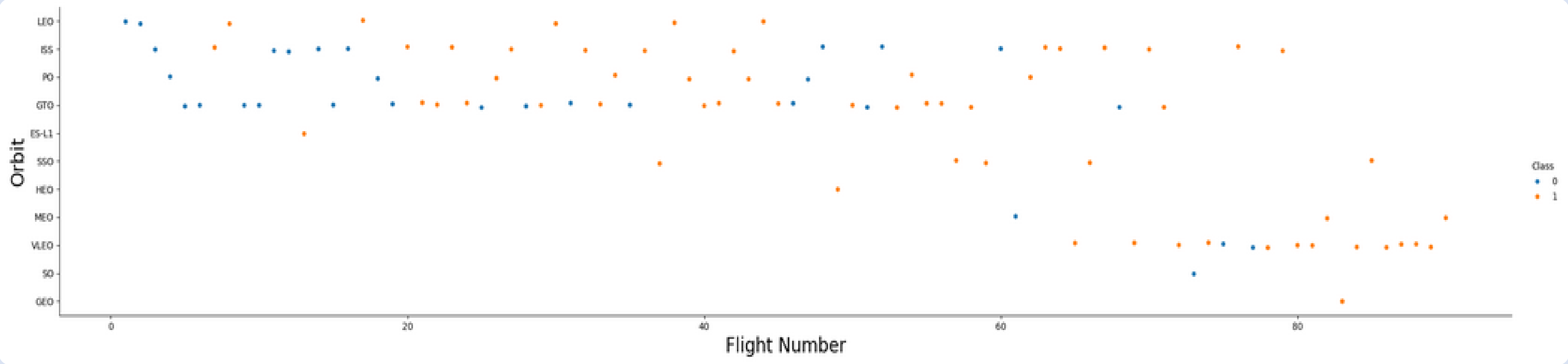
Payload vs. Launch Site

- Site VAFB SCL-4E did not have a launch with a payload mass greater than 10000kg.
- Site KSC LC-39A did not have a launch with a payload mass less than 2500kg.
- Site CCAFS SLC-40 did not have a launch with a payload mass between 7500 and 13000kg.

Success Rate vs. Orbit Type

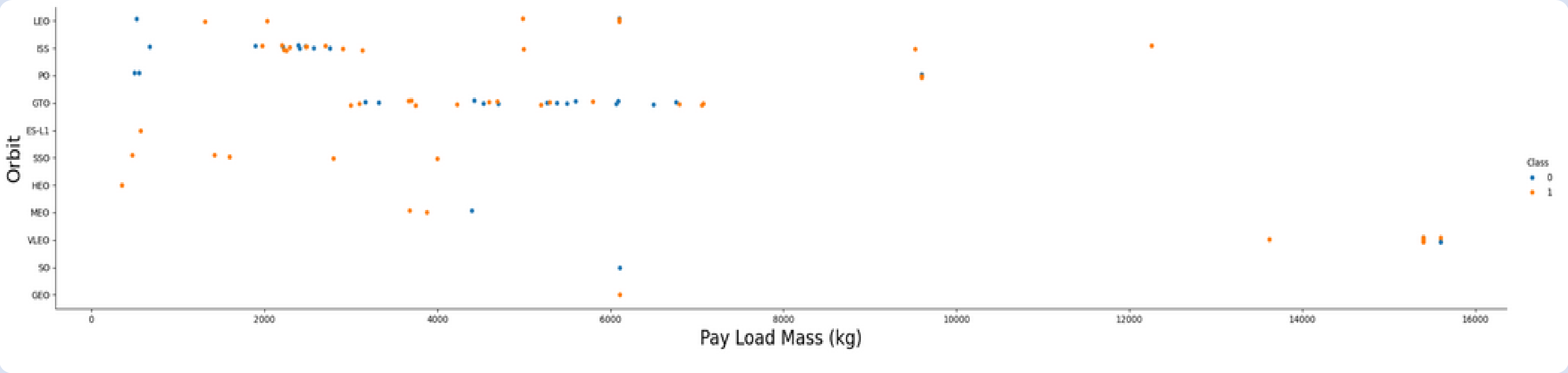


- Orbit types ES-L1, GEO, HEO, and SSO have the highest success rate at 100%.
- Orbit SO had a 0% success rate.



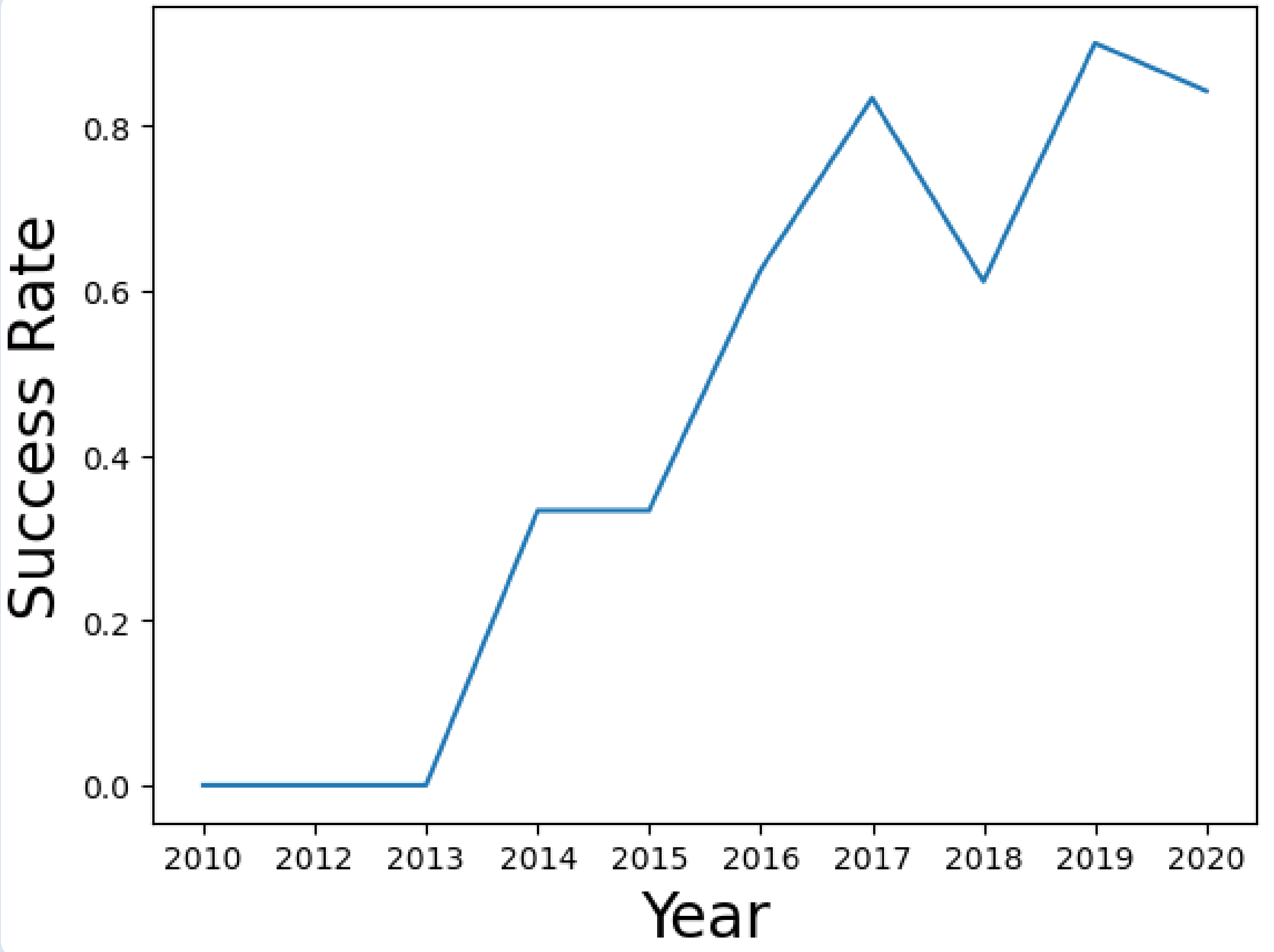
Flight Number vs. Orbit Type

- Orbits with a higher number of flight tend to have a higher success rate.
- An exception to this is orbit GTO.



Payload vs. Orbit Type

- Orbits LEO, ISS, and PO have done well with heavy payloads.
- Orbit GTO shows both positive and negative results with heavy payloads.



Launch Success Yearly Trend

- This chart clearly shows an increased success rate over time.

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

- There are four unique launch sites in this study.
- The distinct statement returns these four unique sites.

```
%sql select LAUNCH_SITE from SPACEXTABLE where (LAUNCH_SITE) like 'CCA%' LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Launch Site Names Begin with 'CCA'

- Like 'CCA%' LIMIT 5 returns 5 records where launch sites begin with `CCA`

```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTABLE;
```

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

payloadmass
619967

Total Payload Mass

- We calculate the total payload carried by boosters from NASA using the sum function.

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

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

avg(PAYLOAD_MASS__KG_)
2928.4

Average Payload Mass by F9 v1.1

- We calculate the average payload mass carried by booster version F9 v1.1 using the average function (avg).

```
%sql select DATE from SPACEXTABLE where (Landing_Outcome) like 'Success (ground pad)' LIMIT 1;
```

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

Date
2015-12-22

First Successful Ground Landing Date

- We find the dates of the first successful landing outcome on ground pad by selecting the first date which indicated a success and ground pad.

```
%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

Successful Drone Ship Landing with Payload between 4000 and 6000

- We can list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 with the use of the greater than and less than functions.

Boosters Carried Maximum Payload

- We can list the names of the booster which have carried the maximum payload mass by using the max function.

```
%sql select max(PAYLOAD_MASS__KG_) from SPACEXTABLE;
```

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

max(PAYLOAD_MASS__KG_)
15600

```
%sql select Booster_Version as boosterversion from SPACEXTABLE where PAYLOAD_MASS__KG_ = (15600);
```

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

boosterversion
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

```
%sql select count(Mission_Outcome) as successes from SPACEXTABLE where (Mission_Outcome) like 'Success%';
```

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

successes
100

```
%sql select count(Mission_Outcome) as failures from SPACEXTABLE where (Mission_Outcome) like 'Failure%';
```

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

failures
1

Total Number of Successful and Failure Mission Outcomes

- We calculate the total number of successful and failure mission outcomes with the use of the count function.

```
%sql select substr(Date, 6,2),Mission_Outcome,booster_version,launch_site from SPACEXTABLE where substr(Date,0,5)='2015';
```

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

substr(Date, 6,2)	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

2015 Launch Records

- We can list the landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- SQLLite does not support month names. So we use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql select landing_outcome as "Landing Outcome", COUNT(landing_outcome) as "total" from SPACEXTABLE\
where date between '2010-06-04' and '2017-03-20'\
GROUP BY landing_outcome\
order by count(landing_outcome) desc;
```

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

Landing Outcome	total
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

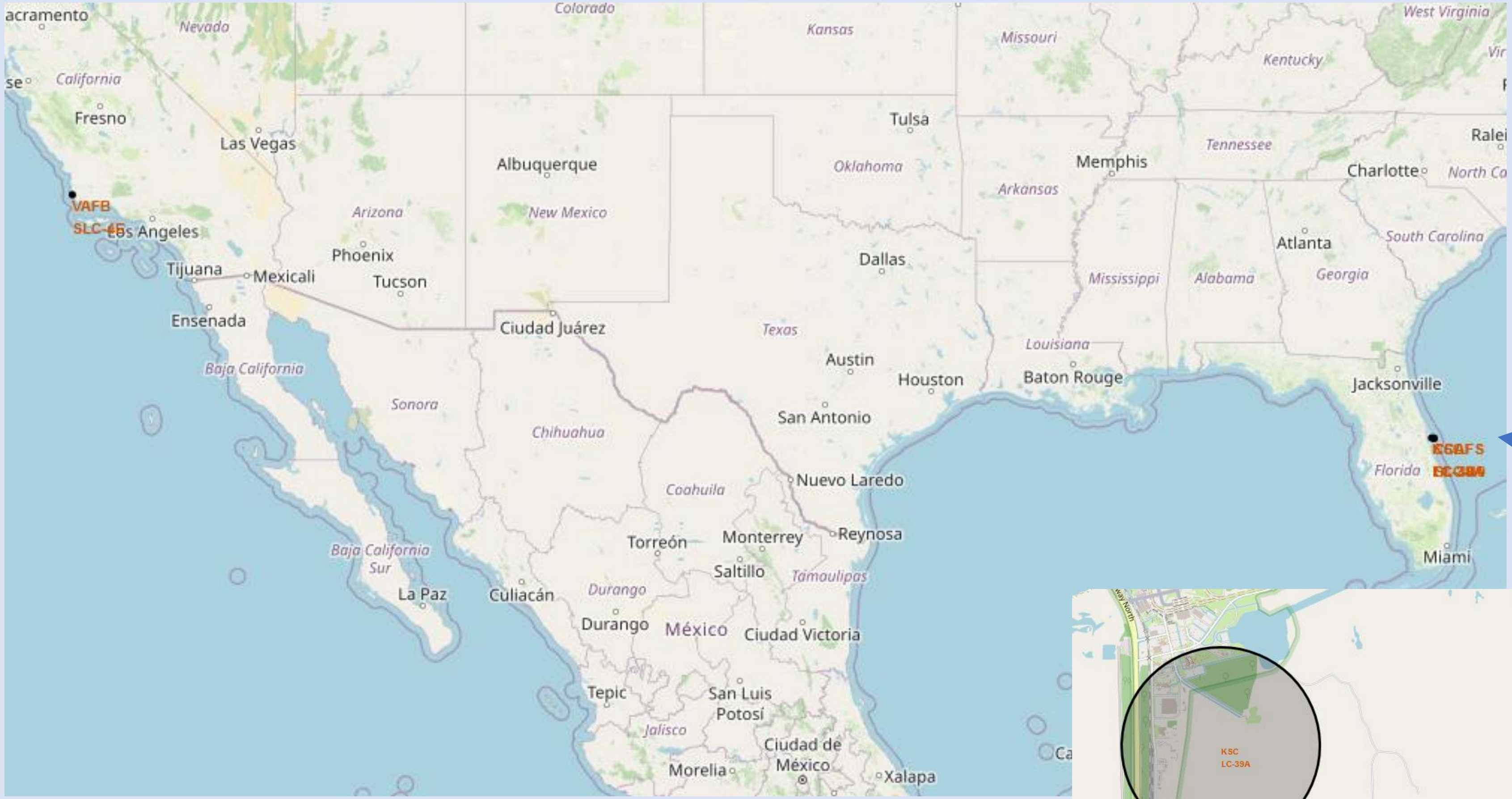
- We rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order by using count and group by.

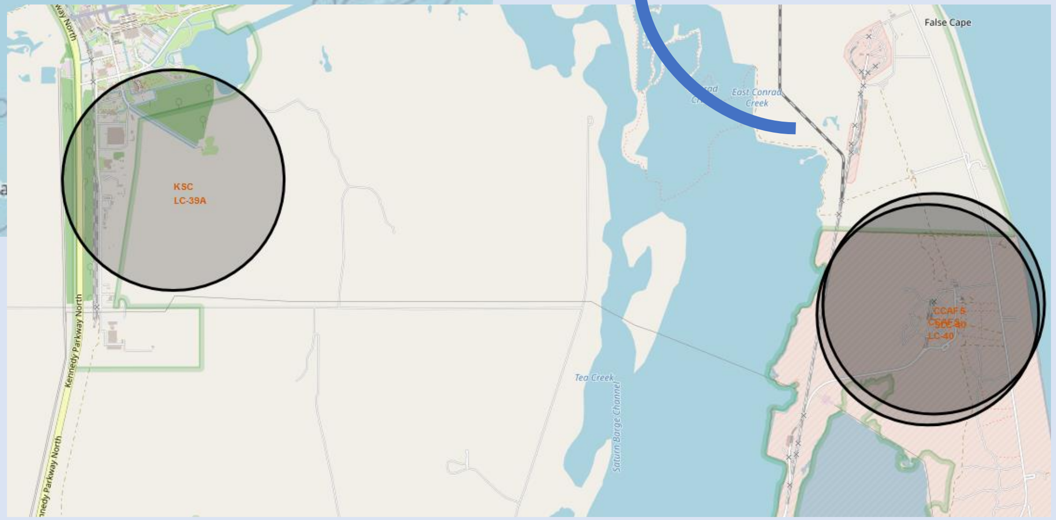
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

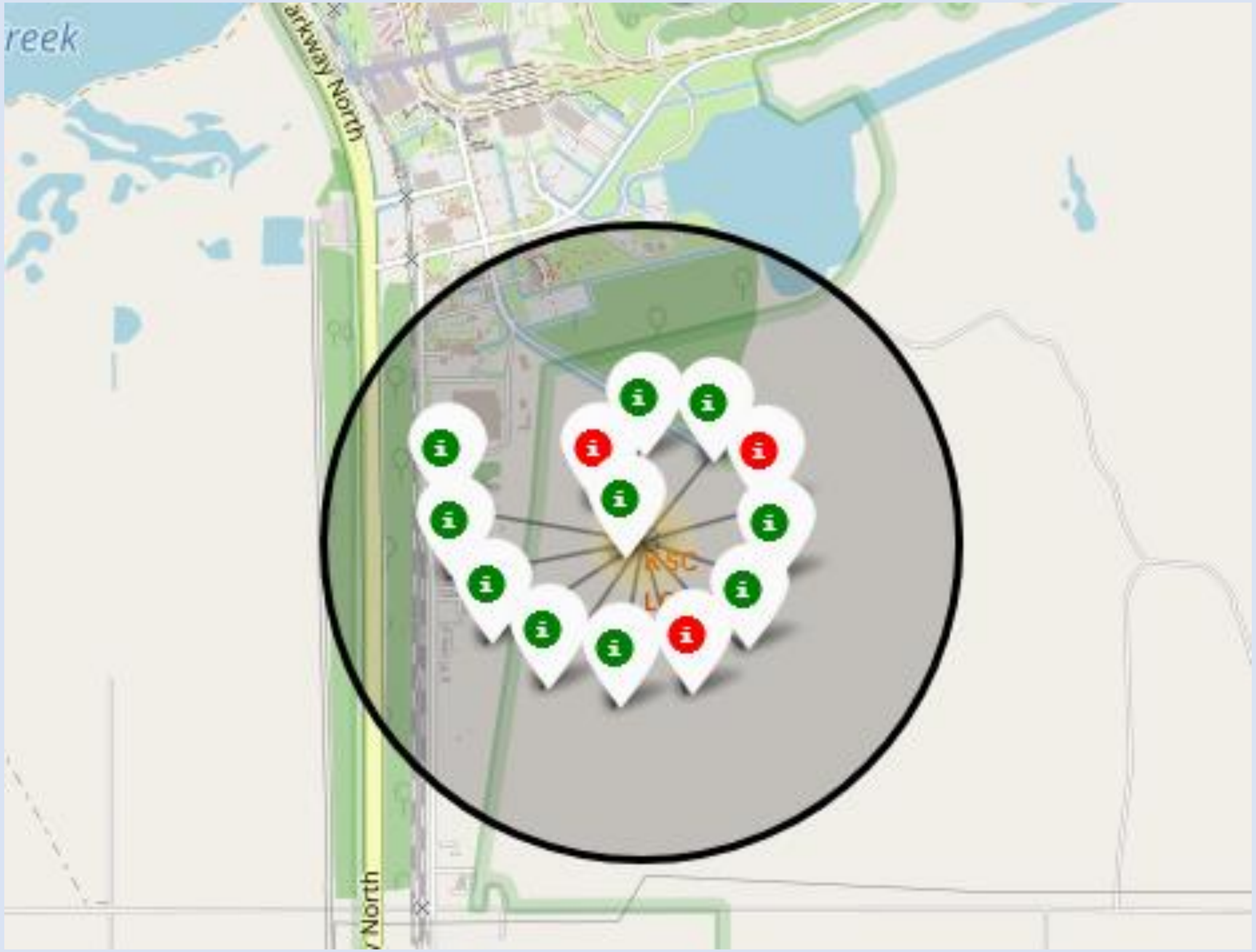
All Launch Sites



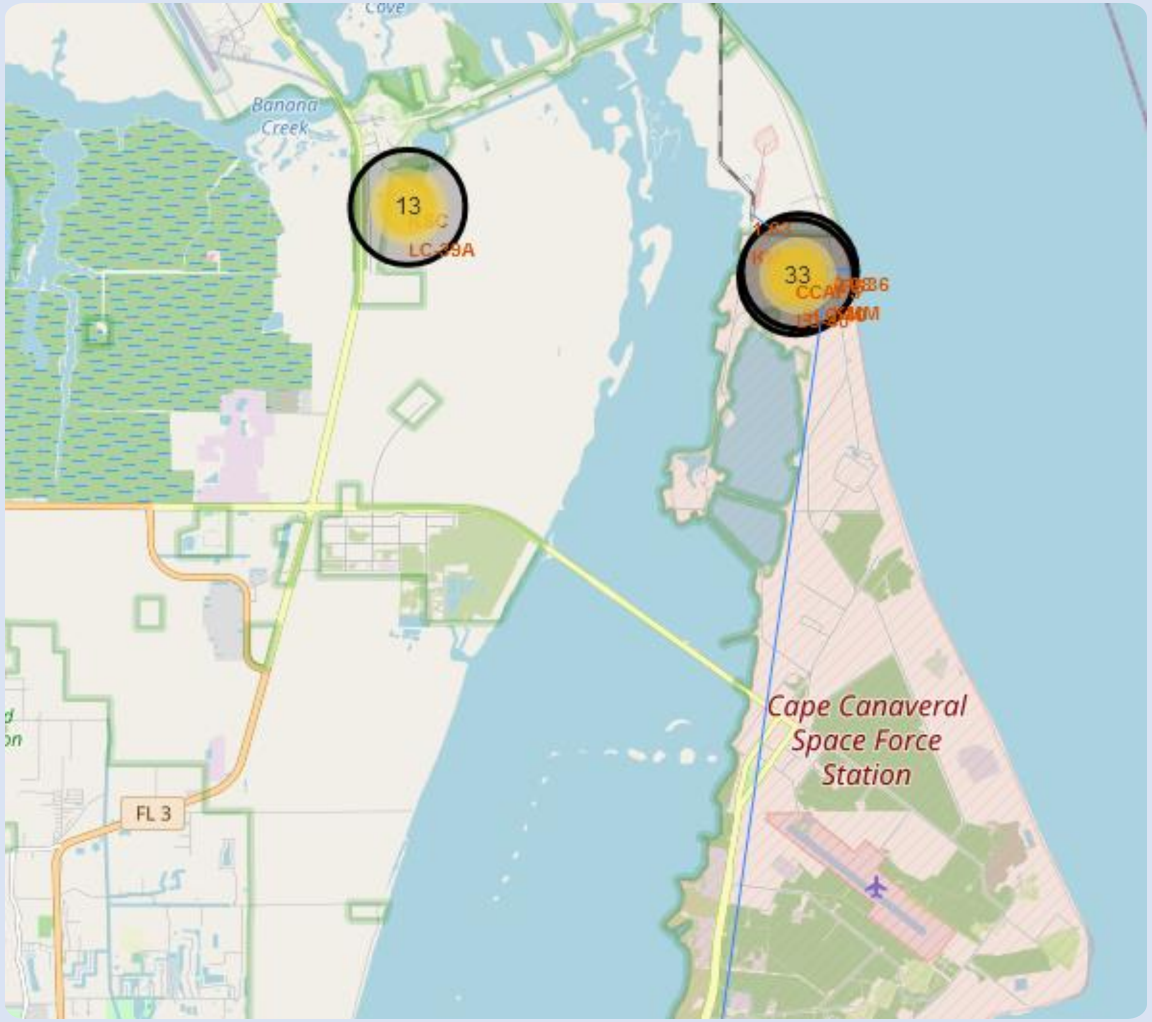


- One site is one the west coast, while the other three are on the east coast and very close together.

Landing Success/Failures by Site



- Cluster markers for each site indicate the number of successes and failures.

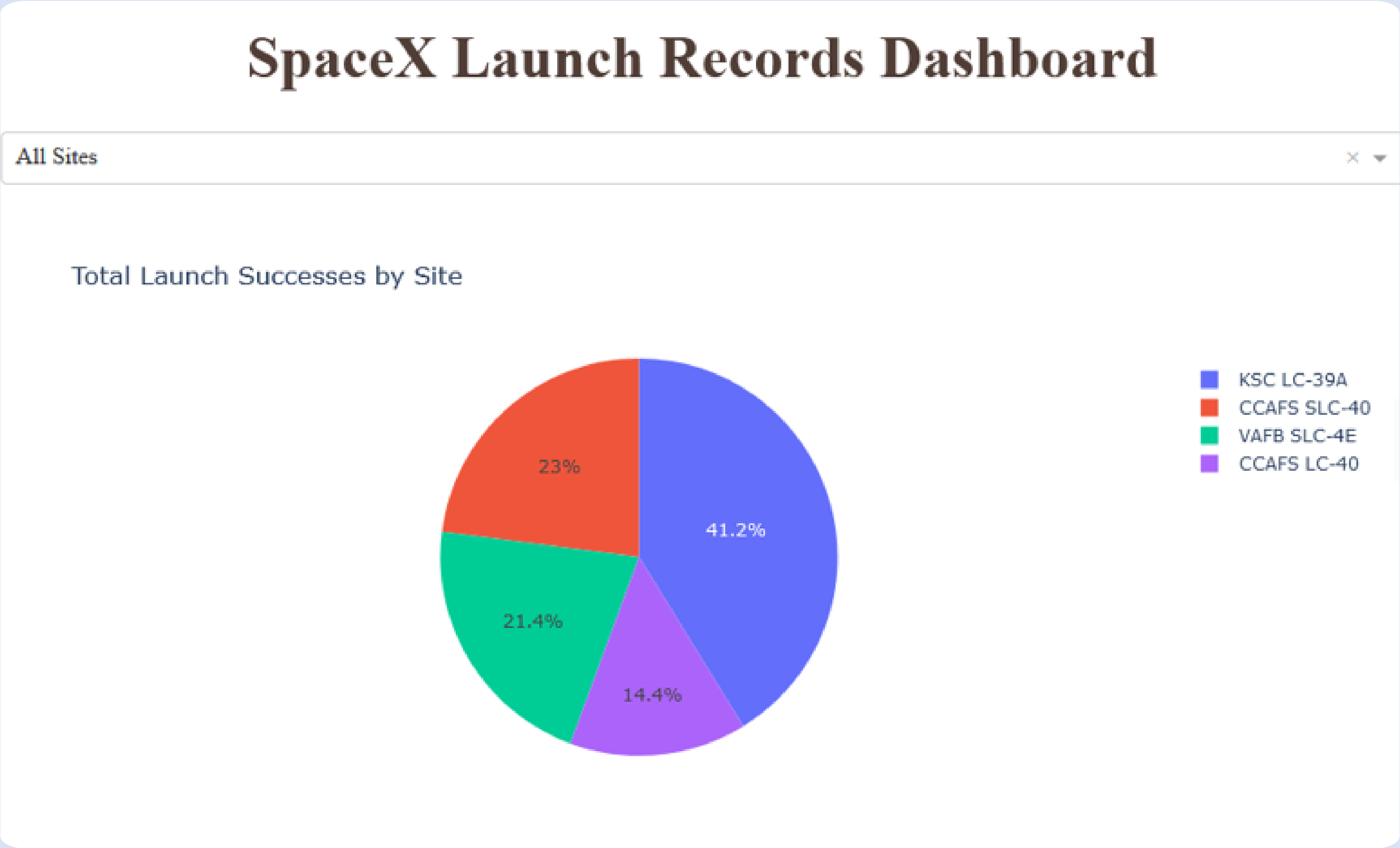


Calculating Distances from Launch Sites

- The launch sites shown are 13 and 33 km from the nearest railroad.

Section 4

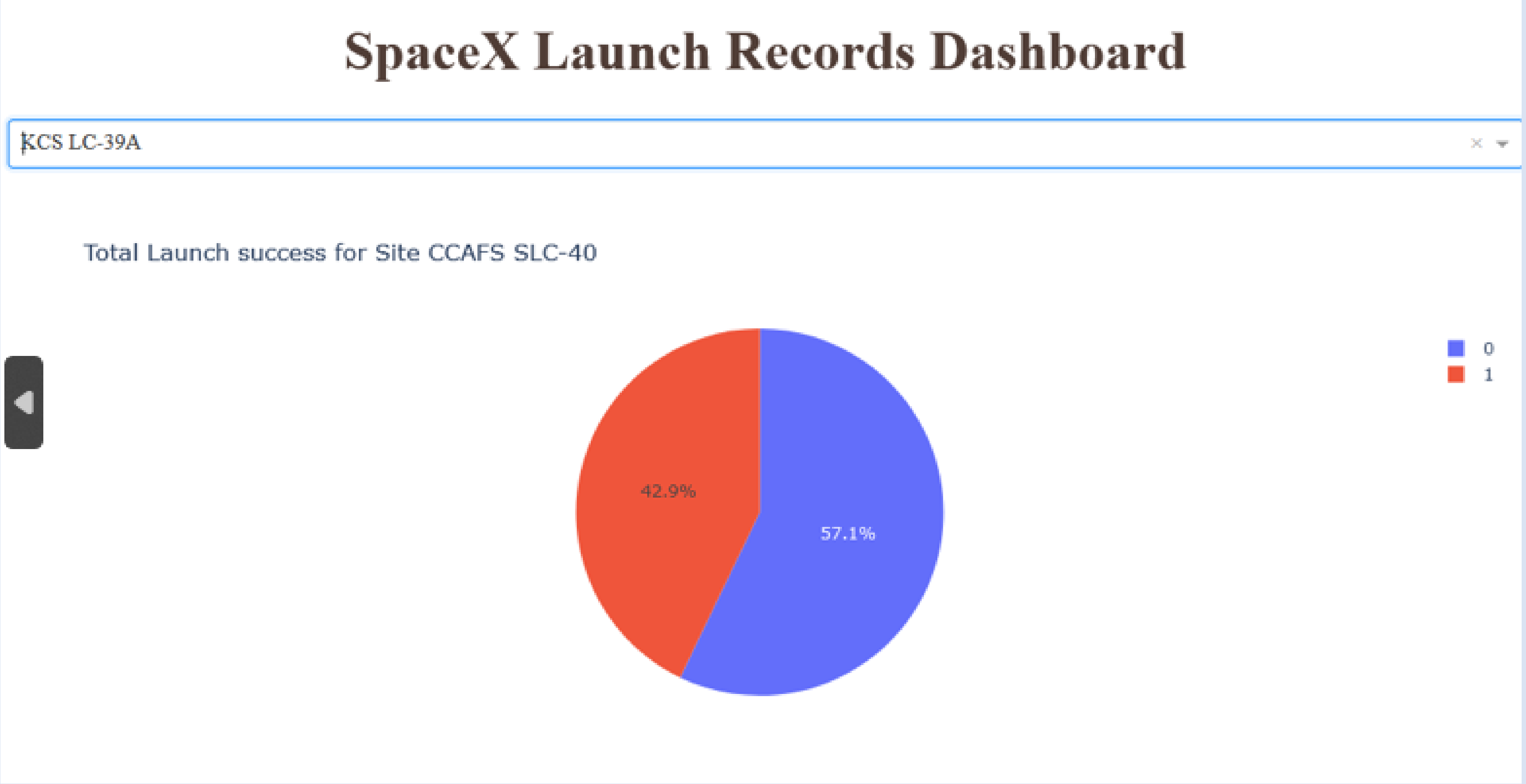
Build a Dashboard with Plotly Dash



Total Launch Successes by Site

- KSC LC-39A has the highest number of successful launches.

KSC LC-39A

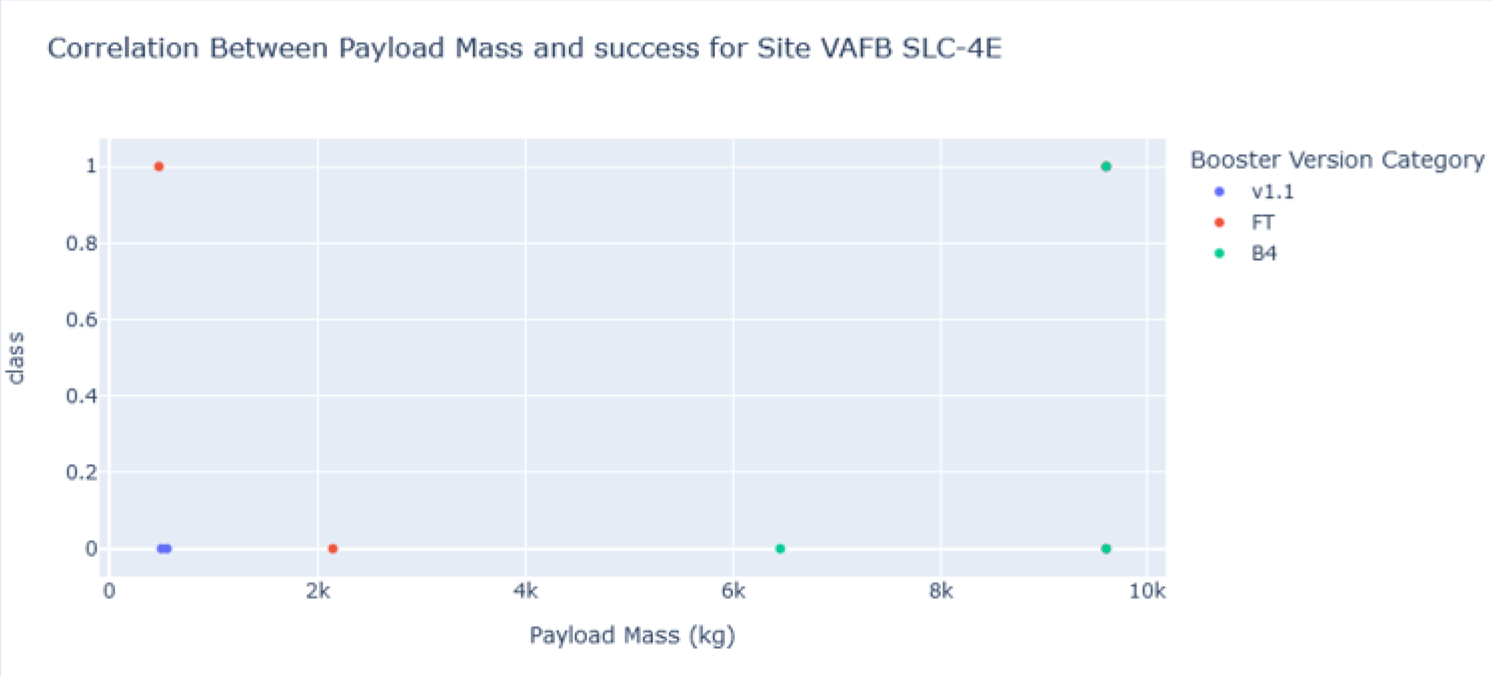


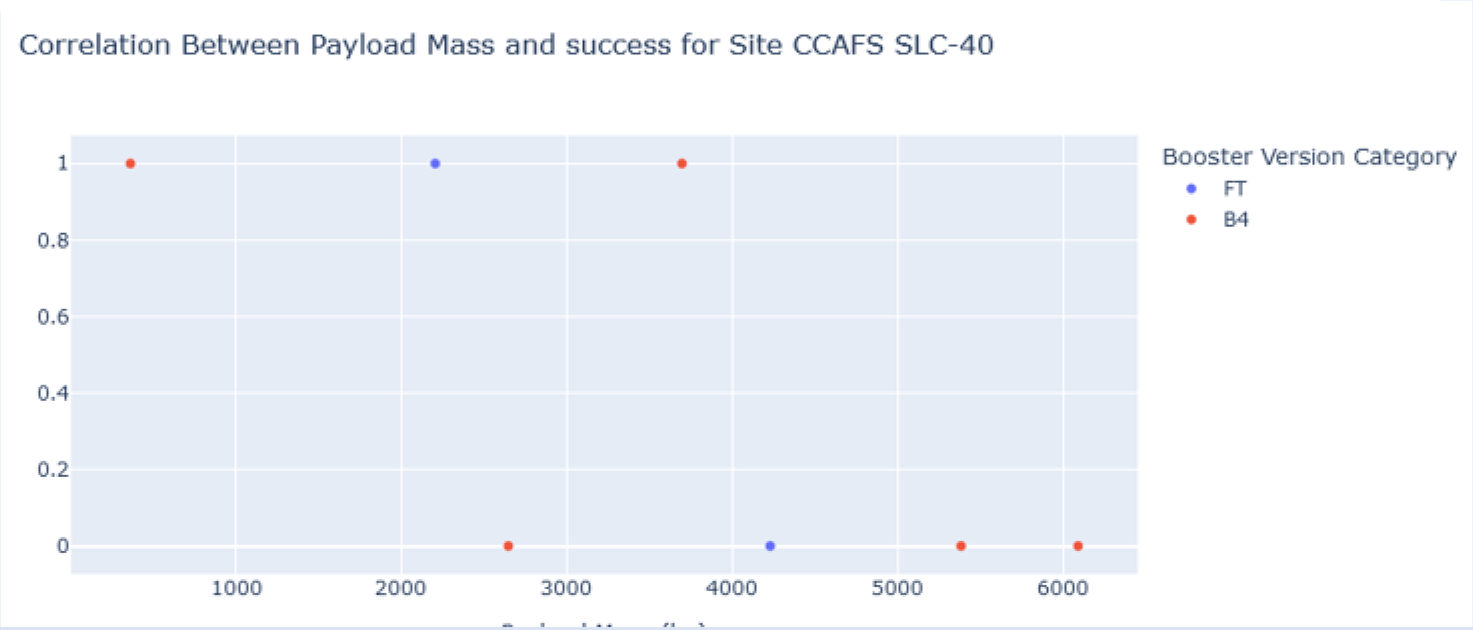
- KSC LC-39A has the highest number of successful launches.

Payload Mass vs Launch Outcome

- The majority of successful launches are in the 2500-500kg range.

Payload Mass vs Launch Outcome





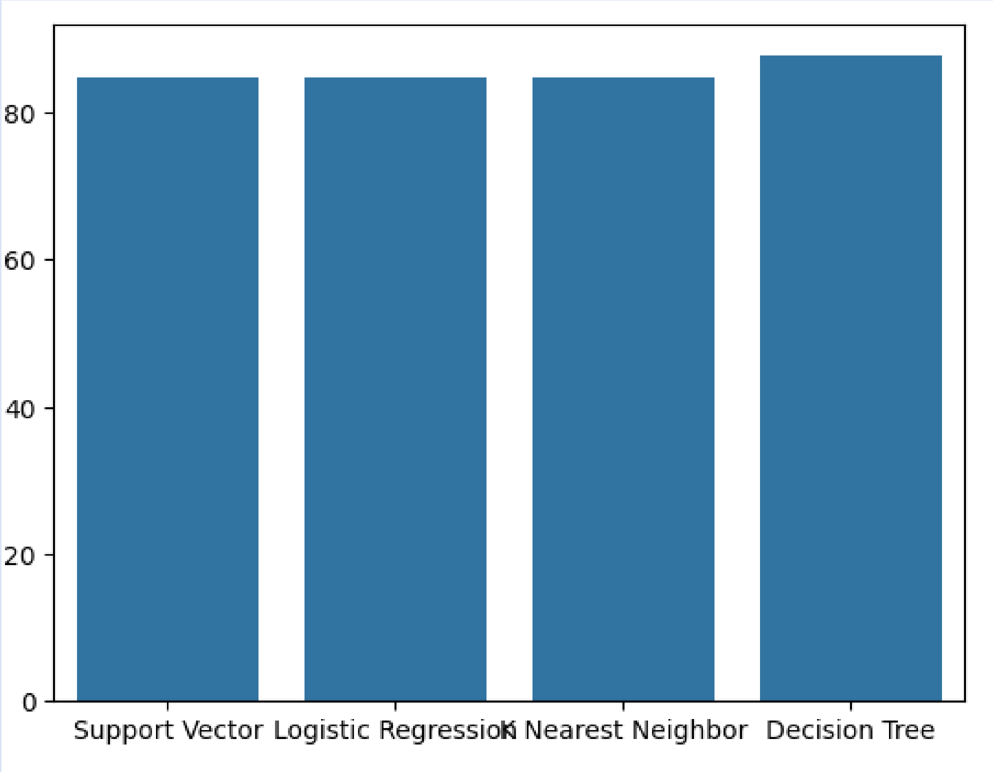
- The majority of successful launches are in the 2500-500kg range.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

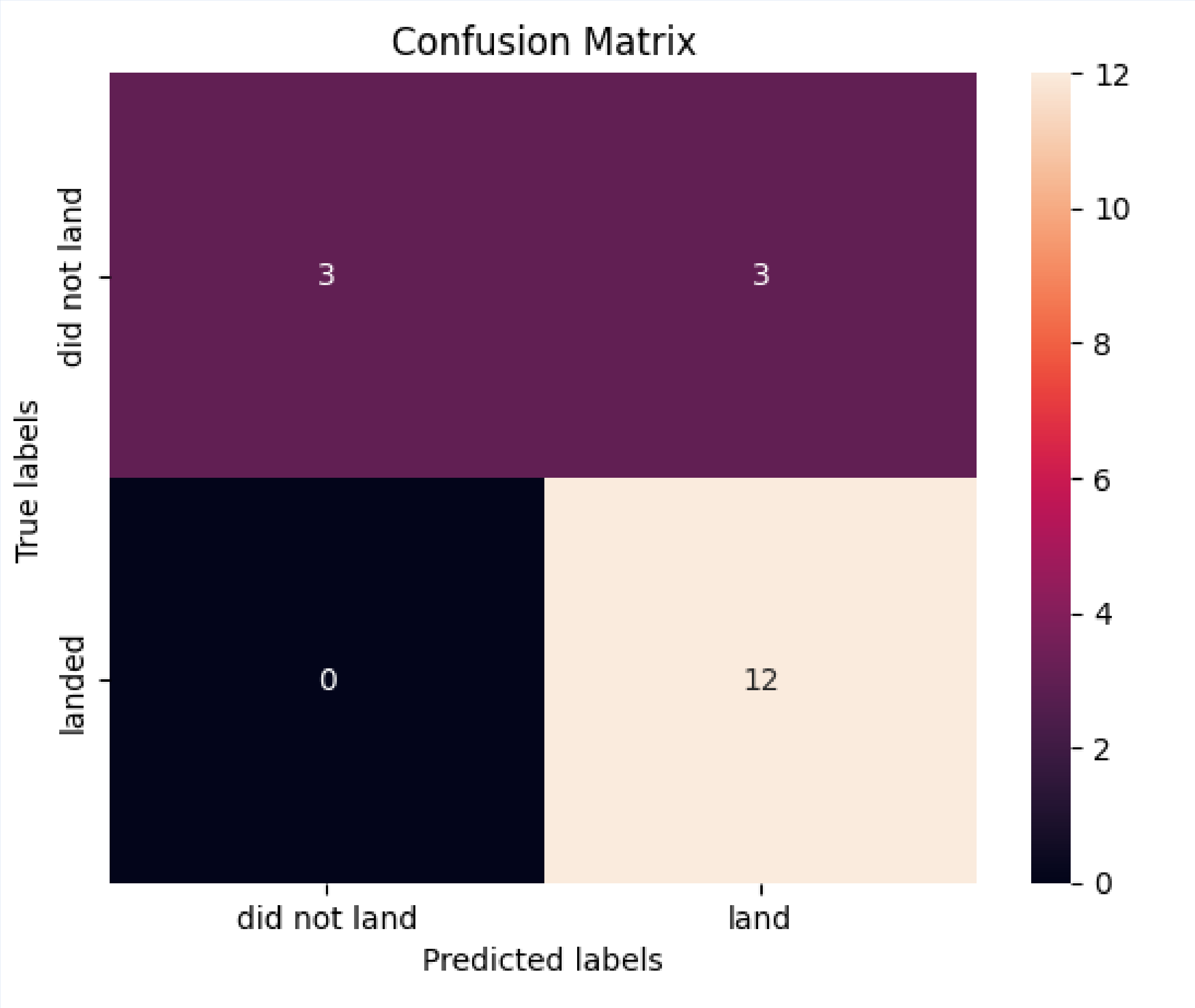
- As we see from the accuracy scores, the built model accuracy for all built classification models are very similar.
- However, the Decision Tree model has a slightly higher score, at 86.25% accuracy.



	ML Method	Accuracy Score %
0	Support Vector	84.821429
1	Logistic Regression	84.642857
2	K Nearest Neighbor	84.821429
3	Decision Tree	86.250000

Confusion Matrix

- The confusion matrix for all models is the same.



Conclusions

- Point 1: Launch success rate has improved over time.
- Point 2: KSC LC 39A is the landing site with the highest success rate.
- Point 3: Orbit types ES-L1, GEO, HEO, and SSO have a 100% success rate
- Point 4: Launch site tend to be located relatively near the equator and close to the coast.
- Point 5: The decision tree model is the best suited for the data set

Thank you!