IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Khoi Tran Anh
Sep 30, 2023

Outline

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

Executive Summary

- Summary of methodologies:
  - Using web scraper to collect data
  - Exploratory Data Analysis (EDA)
  - Machine learning
- Summary of all results
  - Collecting data from public resource successfully
  - Using EDA and machine learning help us see some significant feature of the data and develop some model to predict the success rate.

Introduction

- The objective is to help the new company Space Y to compete with Space X using Data Science
- Wanted answers:
  - Estimation the total cost for launches, by predicting the successful landing of the first stage of the rockets.
  - Decide which launch place has the highest success rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was collected from:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - Web scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcomes label based on outcome data after summarizing and analyzing some features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - Using processed data (normalized) to develop some predictive models, then evaluated the accuracy of each model.

Data Collection

- Data Source:
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - Web scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection - SpaceX API

Request API	Filter data	Dealing with missing value
Request API and parse the SpaceX launch data	Remove all unnecessary data, only keep those relative to the Falcon 9's launches	Replace missing value by the mean value of the selected column

Source code:
https://github.com/KhoiTranAnh/AppliedDataScienceCapstoneCoursera/blob/main/1.%20jupyter-labs-spacex-data-collection-api.ipynb

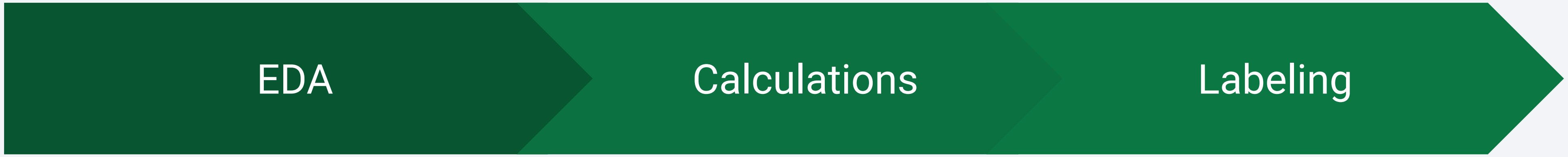
Data Collection - Scraping

Request the Wiki page from its URL	Extract data from the table header	Create a data frame
Create a BeautifulSoup object from the HTTP we just request from the given URL.	Iterate through the <th> elements and apply the provided function to extract column name.	Parsing through the launch HTML tables to create a data frame

Source code:
https://github.com/KhoiTranAnh/AppliedDataScienceCapstoneCoursera/blob/main/2.%20jupyter-labs-webscraping.ipynb

Data Wrangling

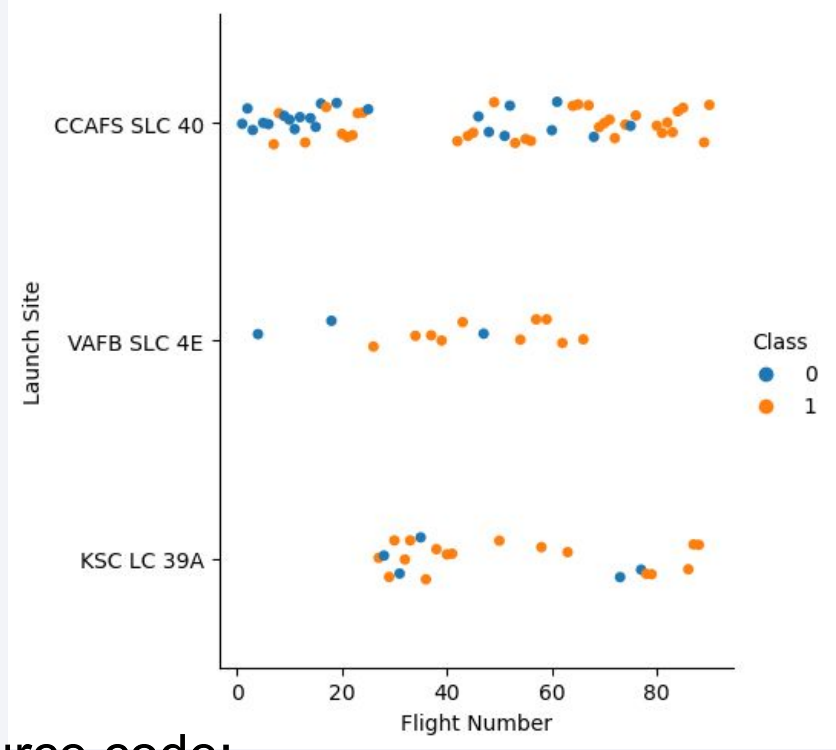
- Perform Exploratory Data Analysis (EDA) on the dataset
- Then doing some calculations on the dataset:
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurence of mission outcome of the orbits
- Finally create a landing outcome label from Outcome column

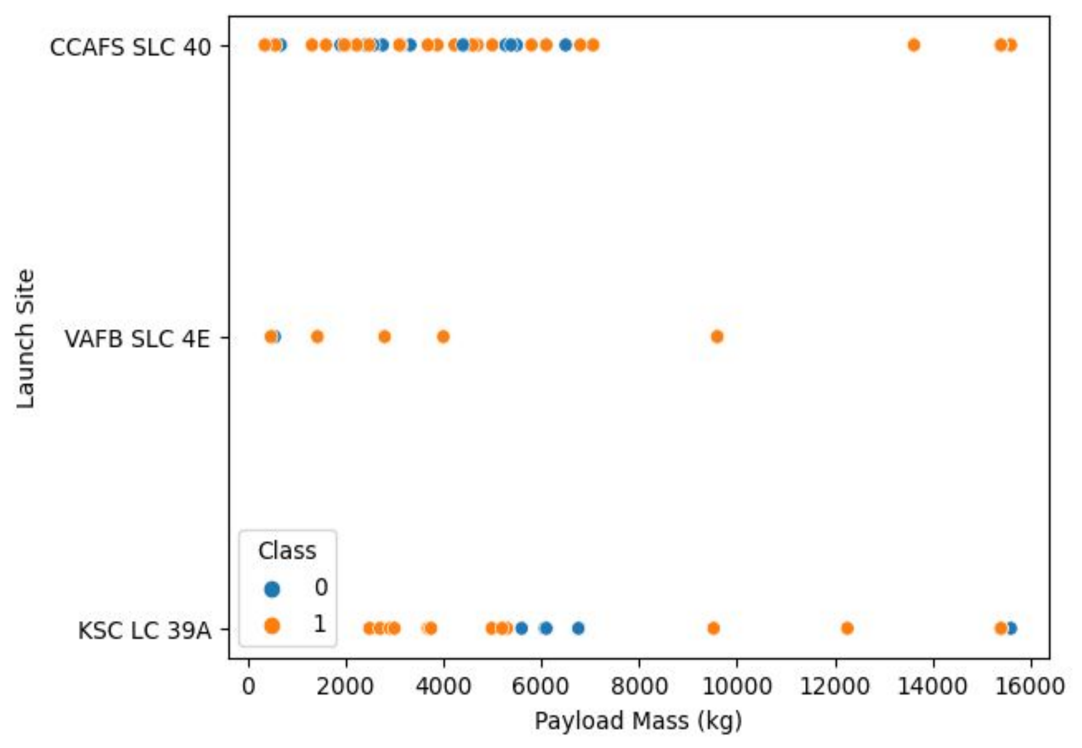


Source code:
https://github.com/KhoiTranAnh/AppliedDataScienceCapstoneCoursera/blob/main/3.%20labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- I use scatterplots and barplots to visualize the relationship between pair of features: FlightNumber vs. PayloadMass and Payload vs. Launch Site





Source code:
https://github.com/KhoiTranAnh/AppliedDataScienceCapstoneCoursera/blob/main/4.%20IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- I used these following SQL queries:
  - The select command to display the column I chose from the database
  - The where command to select the correct data I wanted to display
  - Some built-in functions like avg, min, count
  - The substr function to get “Month”
  - Group by to group data, and Order by to display in descending order.

Source code:
https://github.com/KhoiTranAnh/AppliedDataScienceCapstoneCoursera/blob/main/5.%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers are used to highlight launch sites;
  - Circles indicate highlighted areas around specific coordinates
  - Marker clusters show groups of events in each coordinate
  - Lines are used to indicate distances between two coordinates.

Source code:
https://github.com/KhoiTranAnh/AppliedDataScienceCapstoneCoursera/blob/main/6.%20lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Source code:
https://github.com/KhoiTranAnh/AppliedDataScienceCapstoneCoursera/blob/main/7.%20spacex_dash_app.py

Predictive Analysis (Classification)

- Based on the instructions, I’ve built and compared 4 models:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors
- The steps are represented in the following flow chart:

Source code:
https://github.com/KhoiTranAnh/AppliedDataScienceCapstoneCoursera/blob/main/8.%20SpaceX-Machine-Learning-Prediction-p5.ipynb

Results

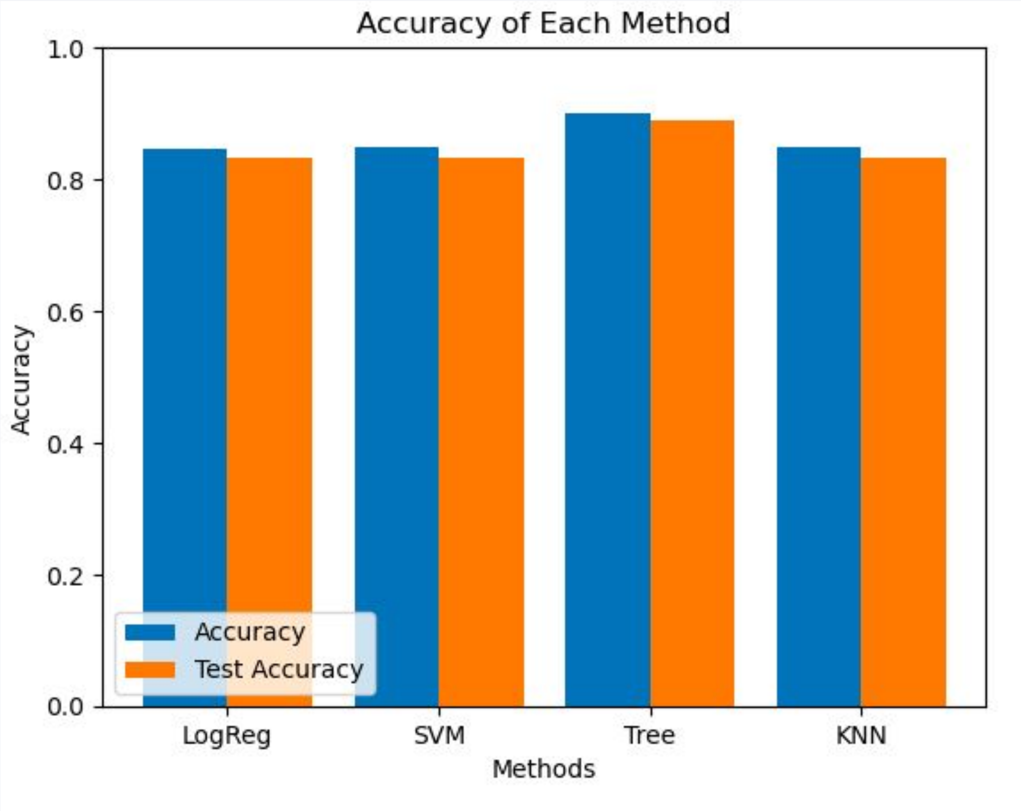
- Exploratory data analysis results:
  - There are 4 different launch sites used by Space X
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 five years after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

Results

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.

Results

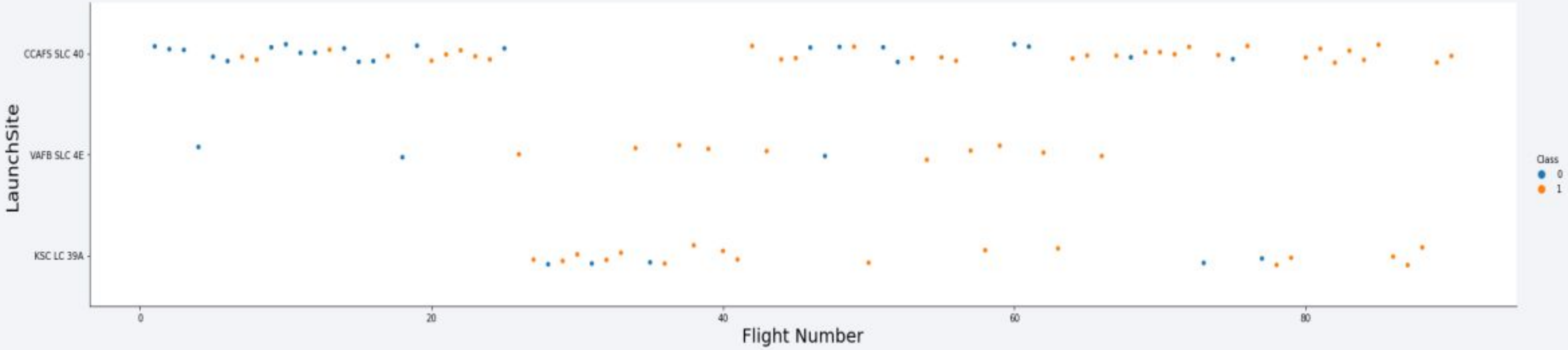
- Decision Tree Classifier is the best model to predict successful landings



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

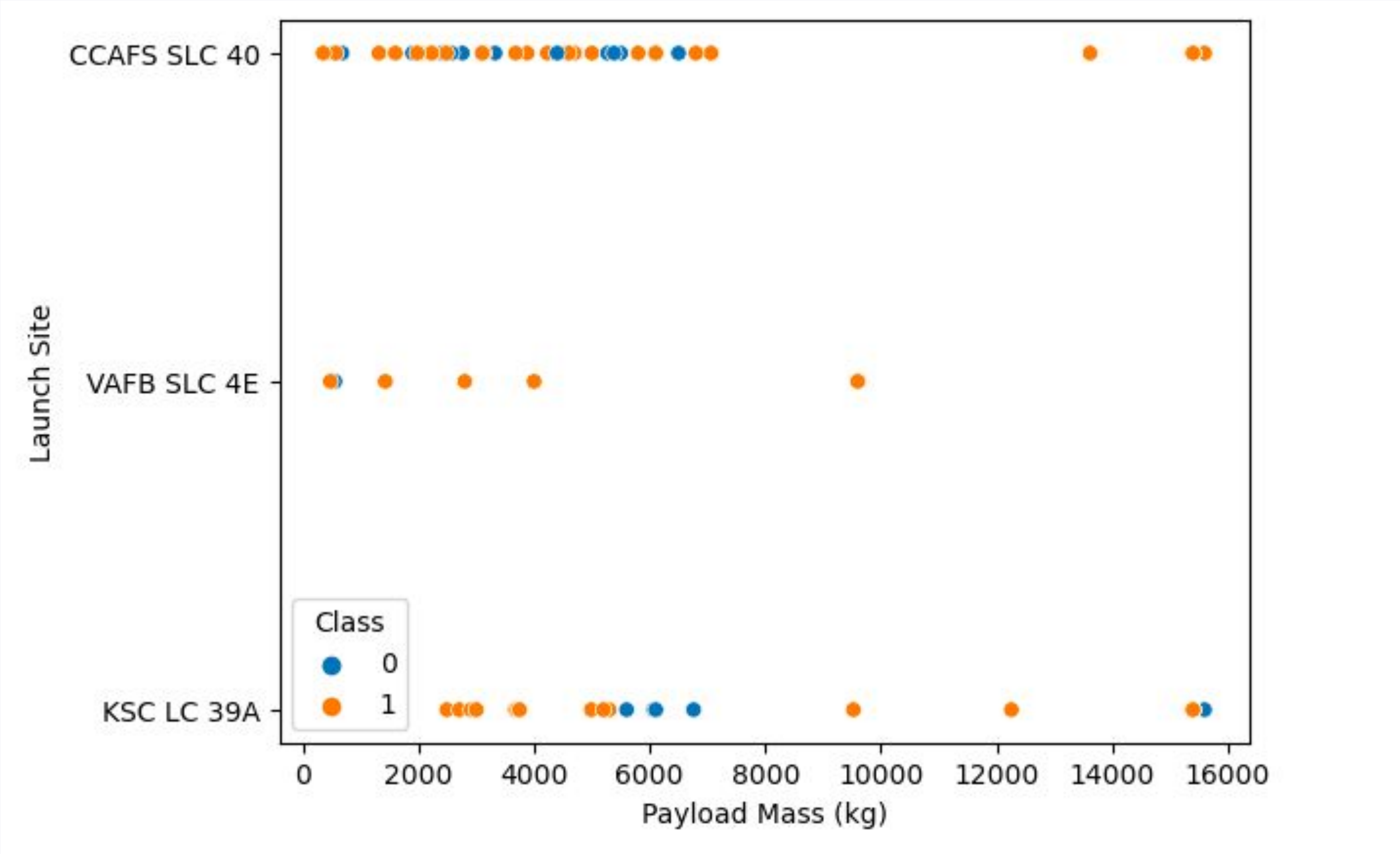


- According to the plot above, CCAF5 SLC 40 shows the best success rate
- In second place is VAFB SLC 4E and third place is KSC LC 39A

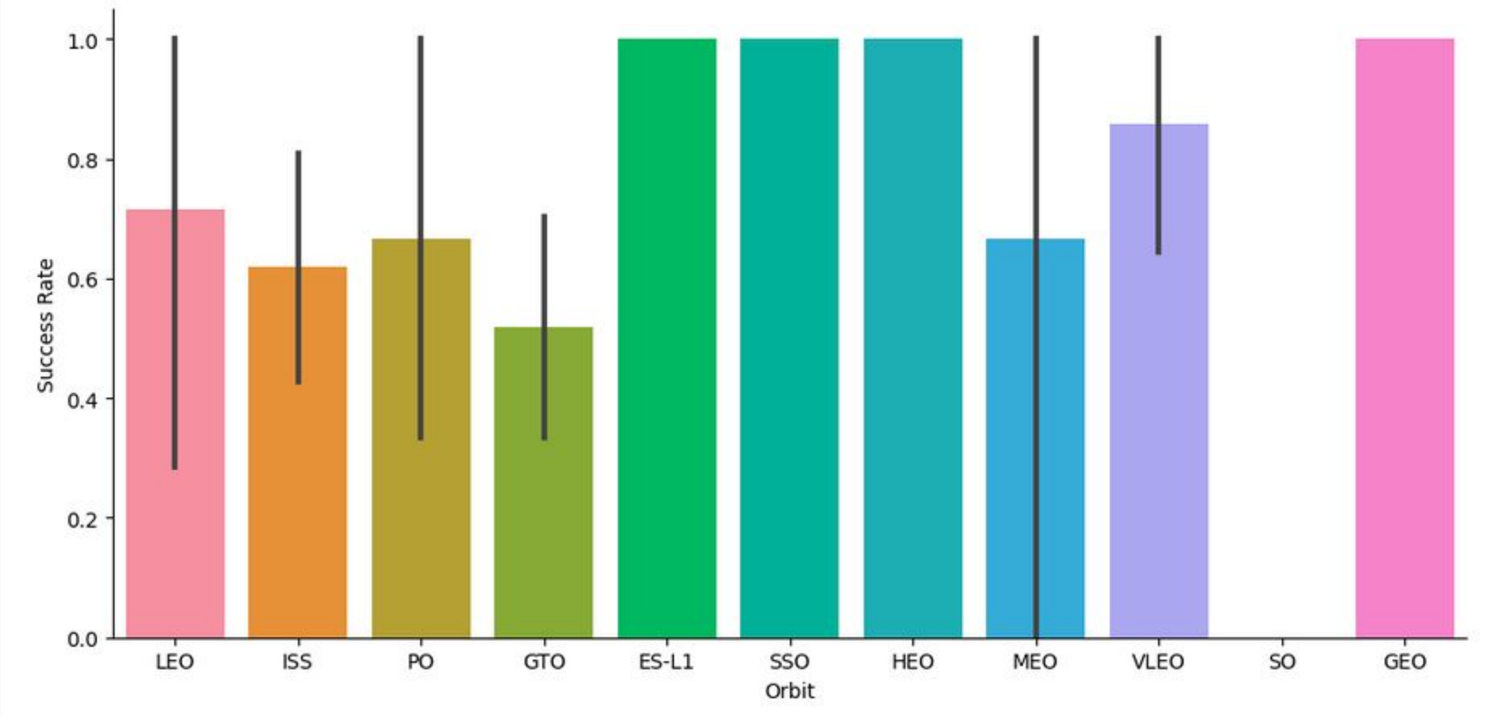
Payload vs. Launch Site

For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000).

Heavy payload mass has very high success rate



Success Rate vs. Orbit Type

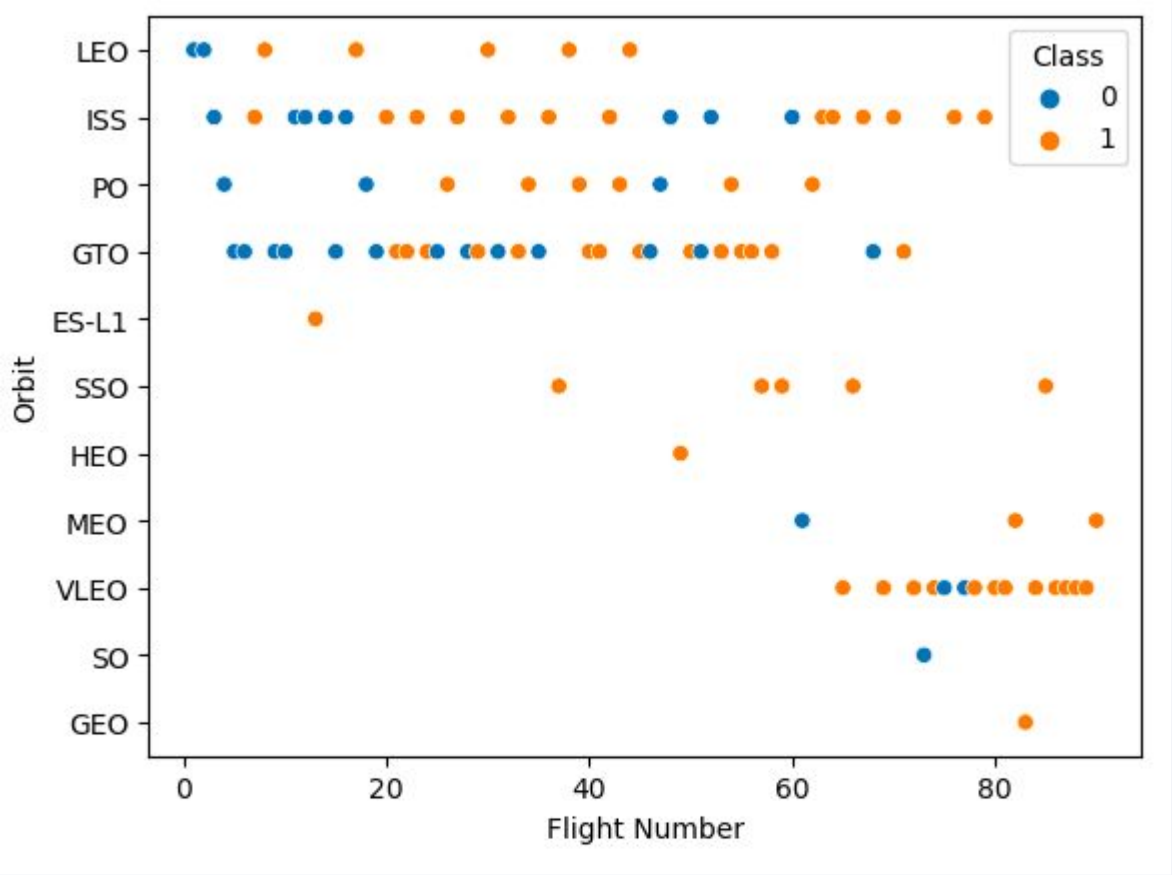


The orbits that has the highest success rate are ES-L1, SSO, HEO, and GEO

SO is the only orbit that has success rate of 0%

Flight Number vs. Orbit Type

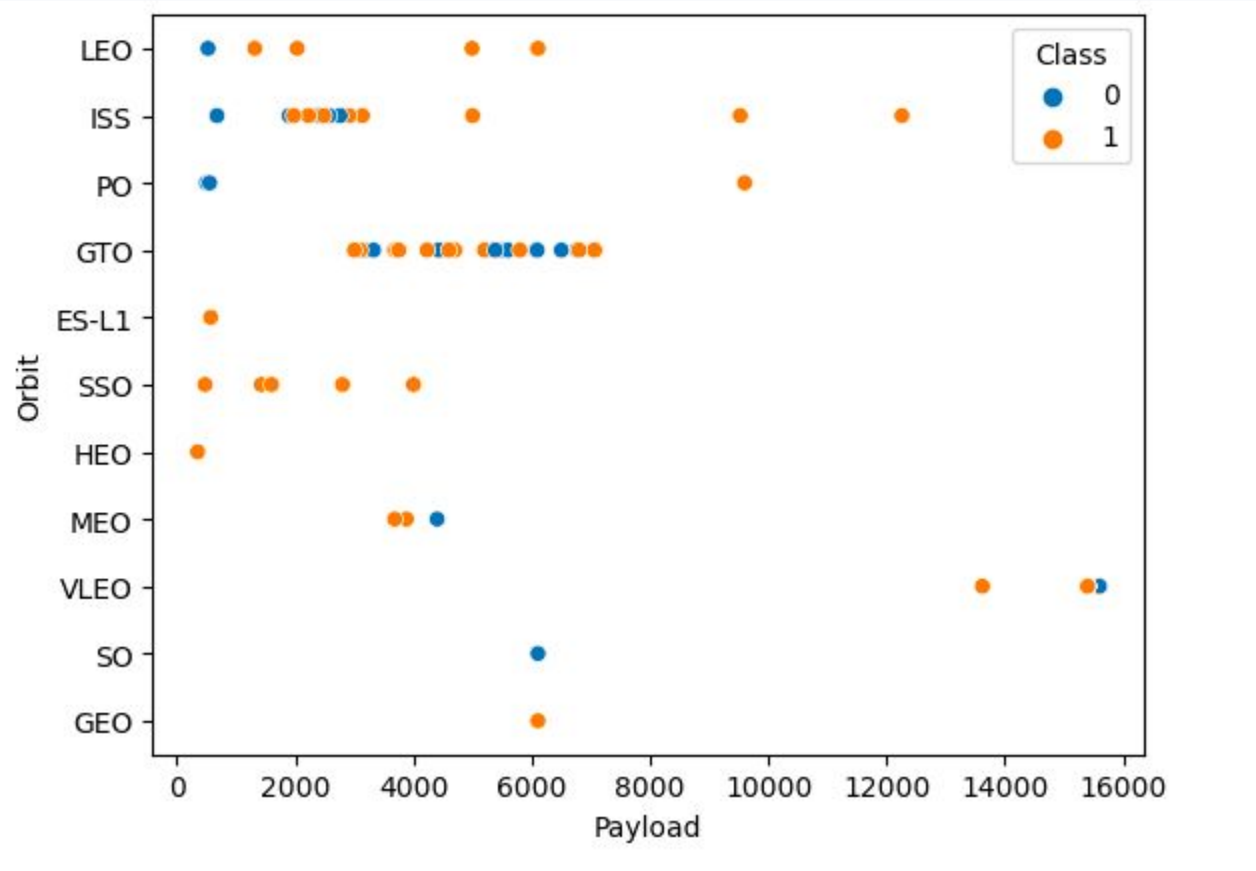
As Flight Number increase, success rate also increase in all orbit



Payload vs. Orbit Type

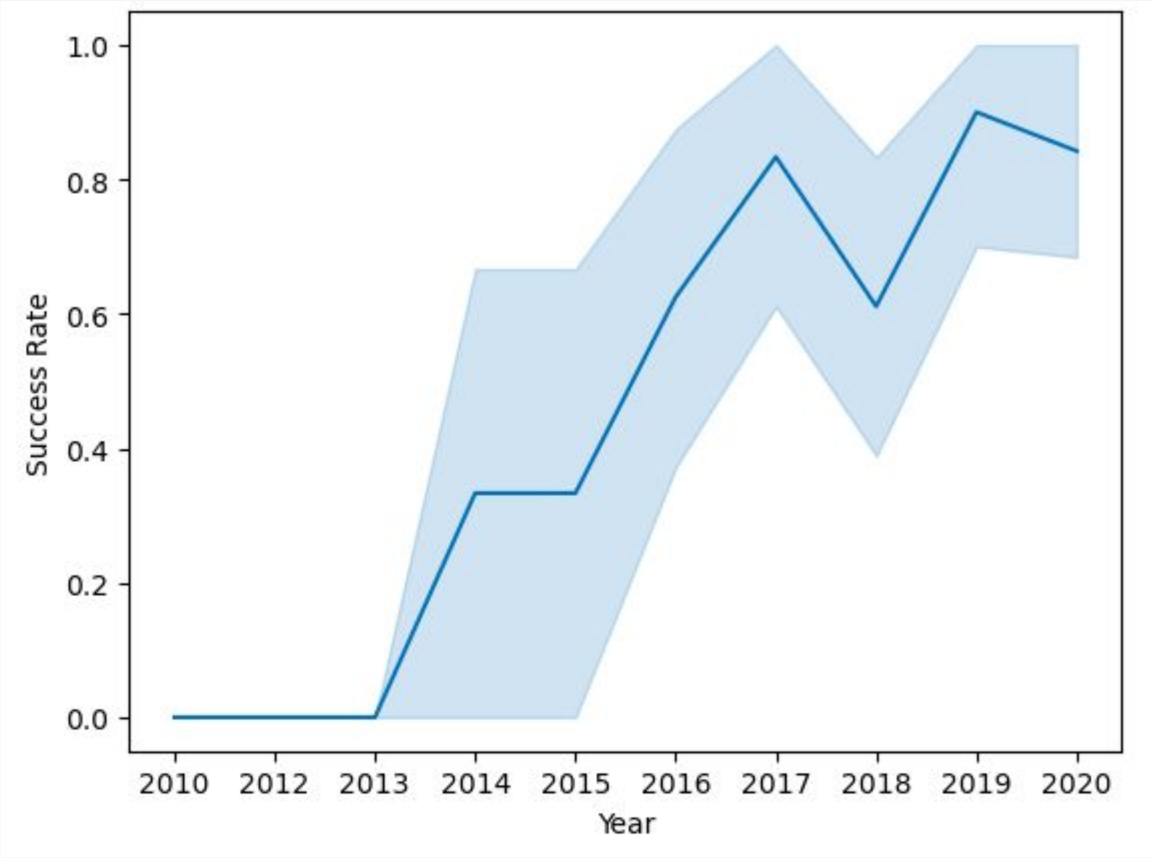
ISS has a wide range of Payload

GTO mainly focus in the Payload of the range around 3000 to 8000



Launch Success Yearly Trend

The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

There are four unique launch sites:

CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40

They are obtain by using the distinct function in the sql query

In [9]:

```
%%sql
select distinct("Launch_Site") from SPACEXTABLE
```

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

Out[9]:

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
2010-08-12	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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

The total payload carried by boosters from NASA are represent below, obtain by using the sum function in the sql query

```
In [15]:  %%sql
          select sum("PAYLOAD_MASS__KG_")
          FROM SPACEXTABLE
          where "Customer" = "NASA (CRS)"
          -- where "Customer" like "NASA (CRS)"

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

Out[15]:

sum("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

```
In [17]: %%sql
         select avg("PAYLOAD_MASS__KG_")
         from SPACEXTABLE
         where "Booster_Version" like "F9 v1.1%";

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

Out[17]:

avg("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

```
In [21]:
%%sql
select min("Date")
from SPACEXTABLE
where "Landing_Outcome" = "Success (ground pad)";

 * sqlite:///my_data1.db
Done.

Out[21]:
```

min("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [24]:

```
%%sql
select "Booster_version"
from SPACEXTABLE
where "Landing_Outcome" = "Success (drone ship)" and "Payload_mass__KG_" > 4000 and "Payload_mass__KG_" < 6000;
```

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

Out[24]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

In [33]:
```
%%sql
SELECT COUNT("Mission_Outcome")
FROM SPACEXTBL
WHERE "Mission_Outcome" LIKE 'Success%'
```

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

Out[33]:

COUNT("Mission_Outcome")
100

In [34]:
```
%%sql
SELECT COUNT("Mission_Outcome")
FROM SPACEXTBL
WHERE "Mission_Outcome" LIKE 'Fail%'
```

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

Out[34]:

COUNT("Mission_Outcome")
1

Boosters Carried Maximum Payload

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

Out[35]:

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

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

In [43]:
```
%%sql
select substr("Date", 6, 2) as "Month", "Booster_version", "Launch_site"
from SPACEXTABLE
where "Date" like "2015-%" and "Landing_outcome" = "Failure (drone ship)";
```

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

Out[43]:

Month	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- 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

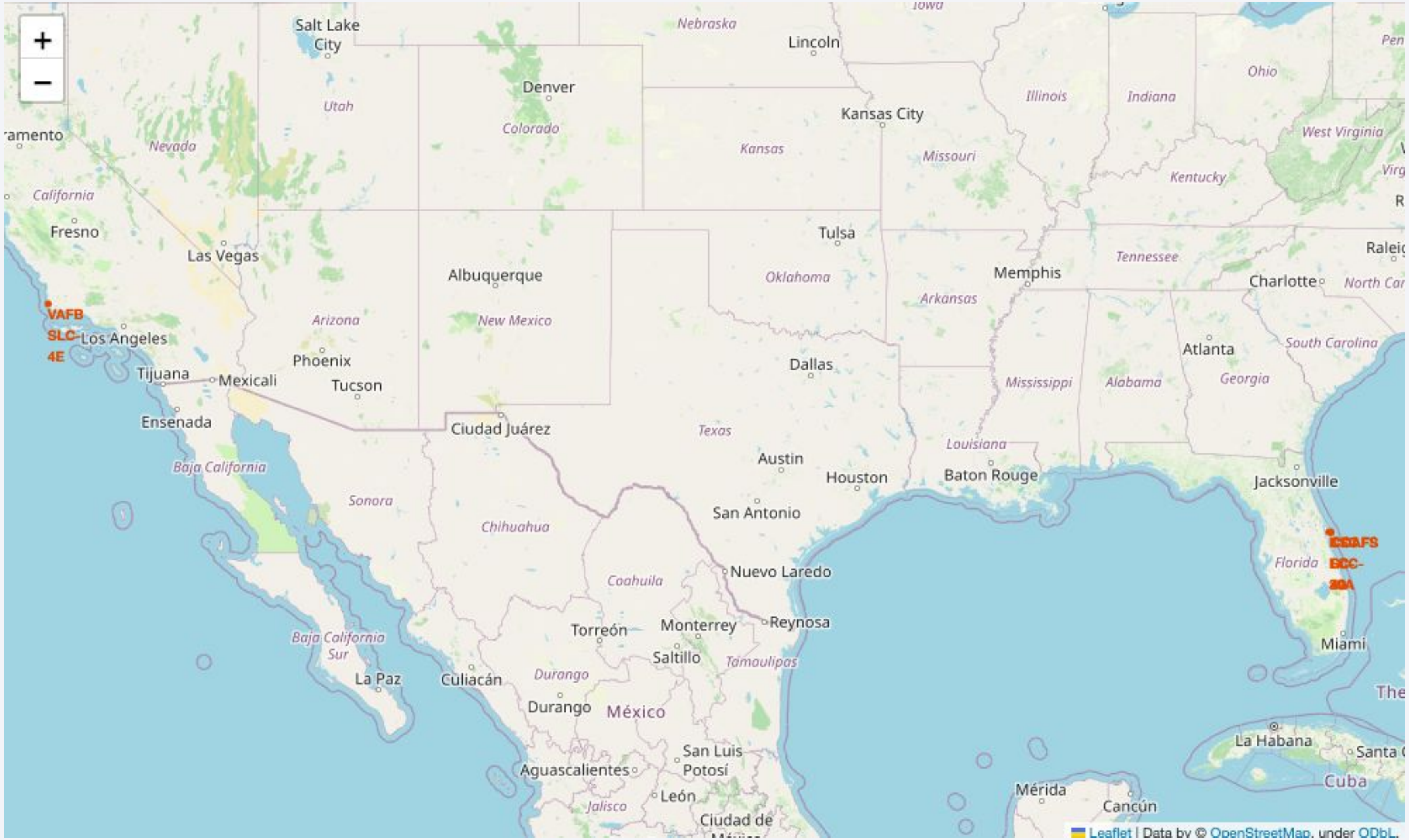
Out[50]:

Landing_Outcome	Total Number
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

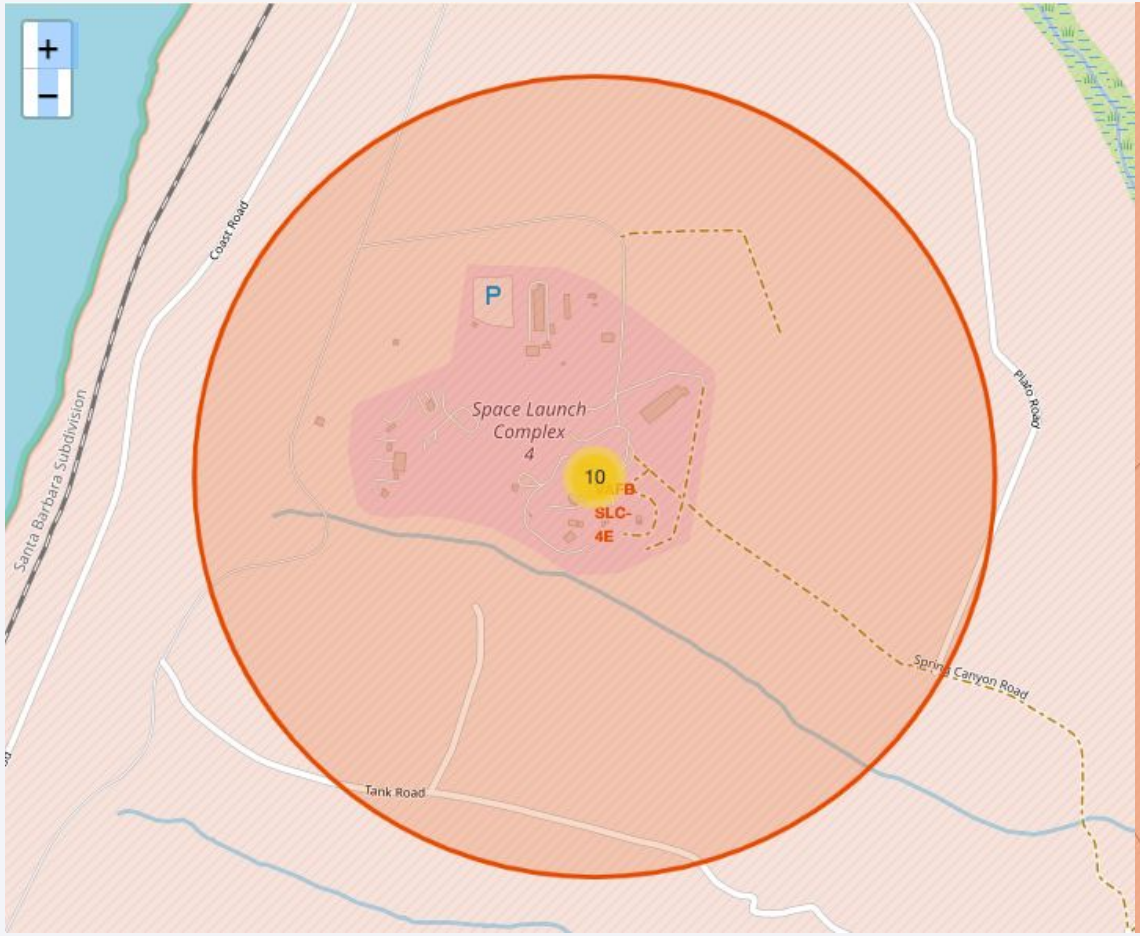
Section 3

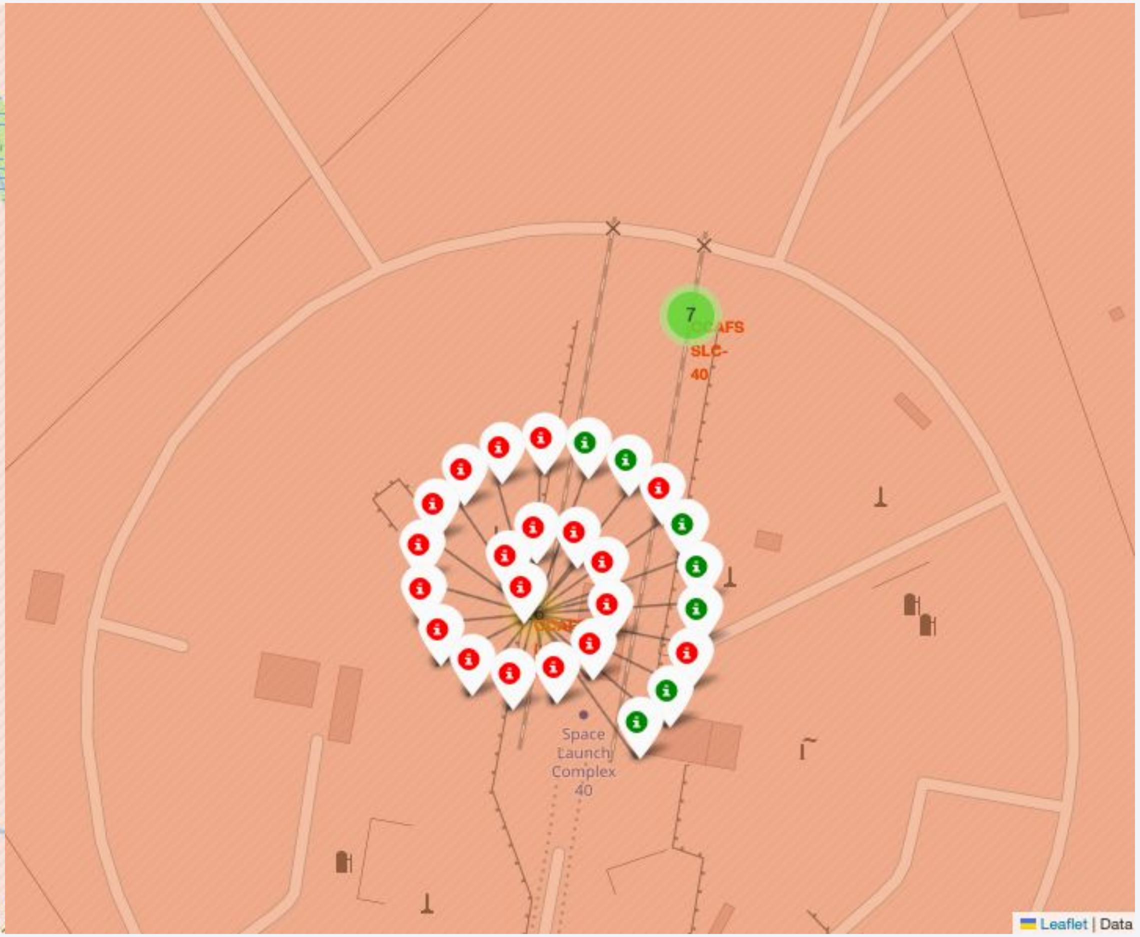
Launch Sites Proximities Analysis

All Launch Sites



Launch Outcomes by Site





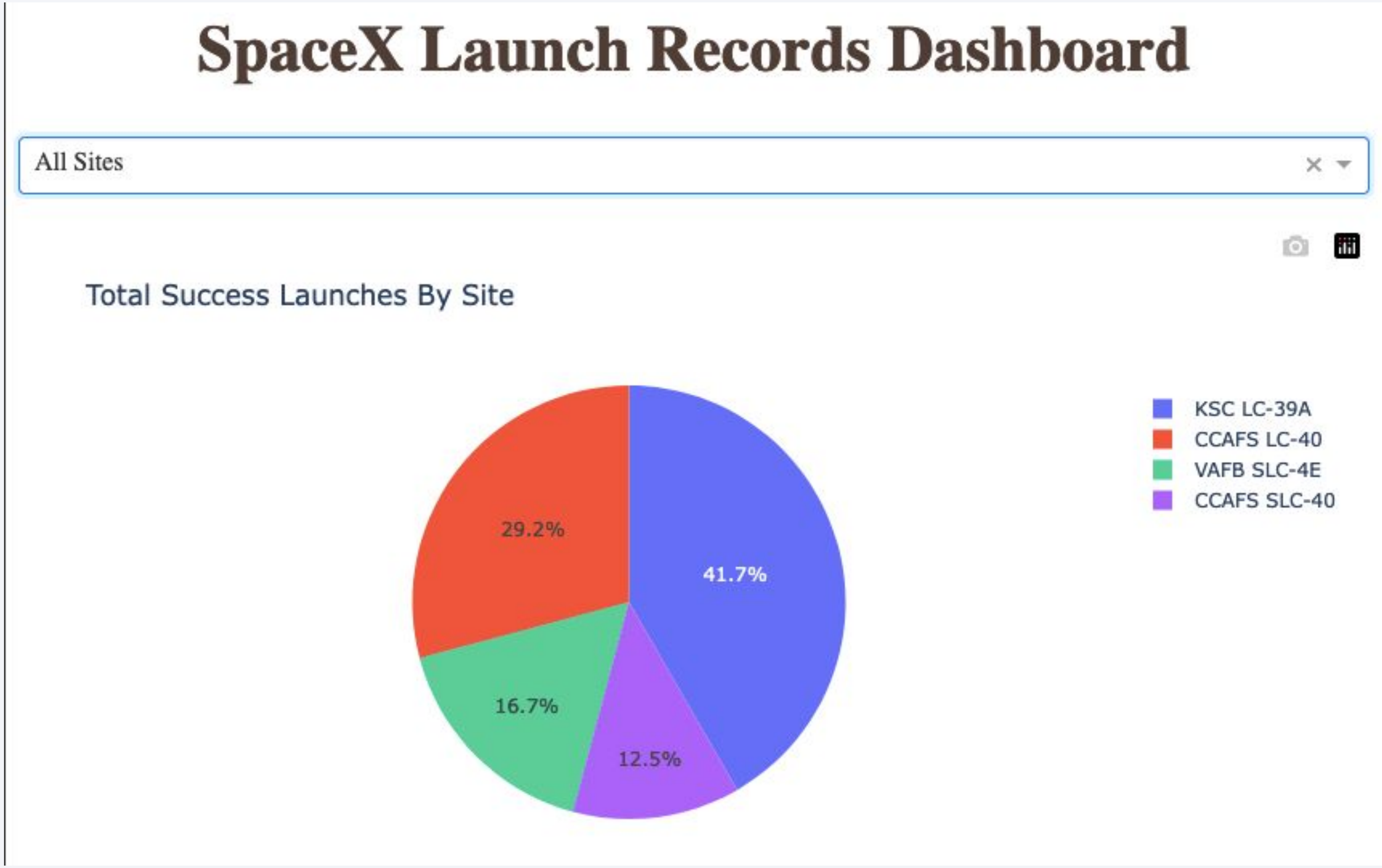
Distance


Lat: 28.57433 Long: -80.64098
Contractors Road
Kennedy Parkway North
Kennedy Parkway North
0.55
KM
13
LC-
39A
Leaflet | Data by © OpenStreetMap, under ODbL

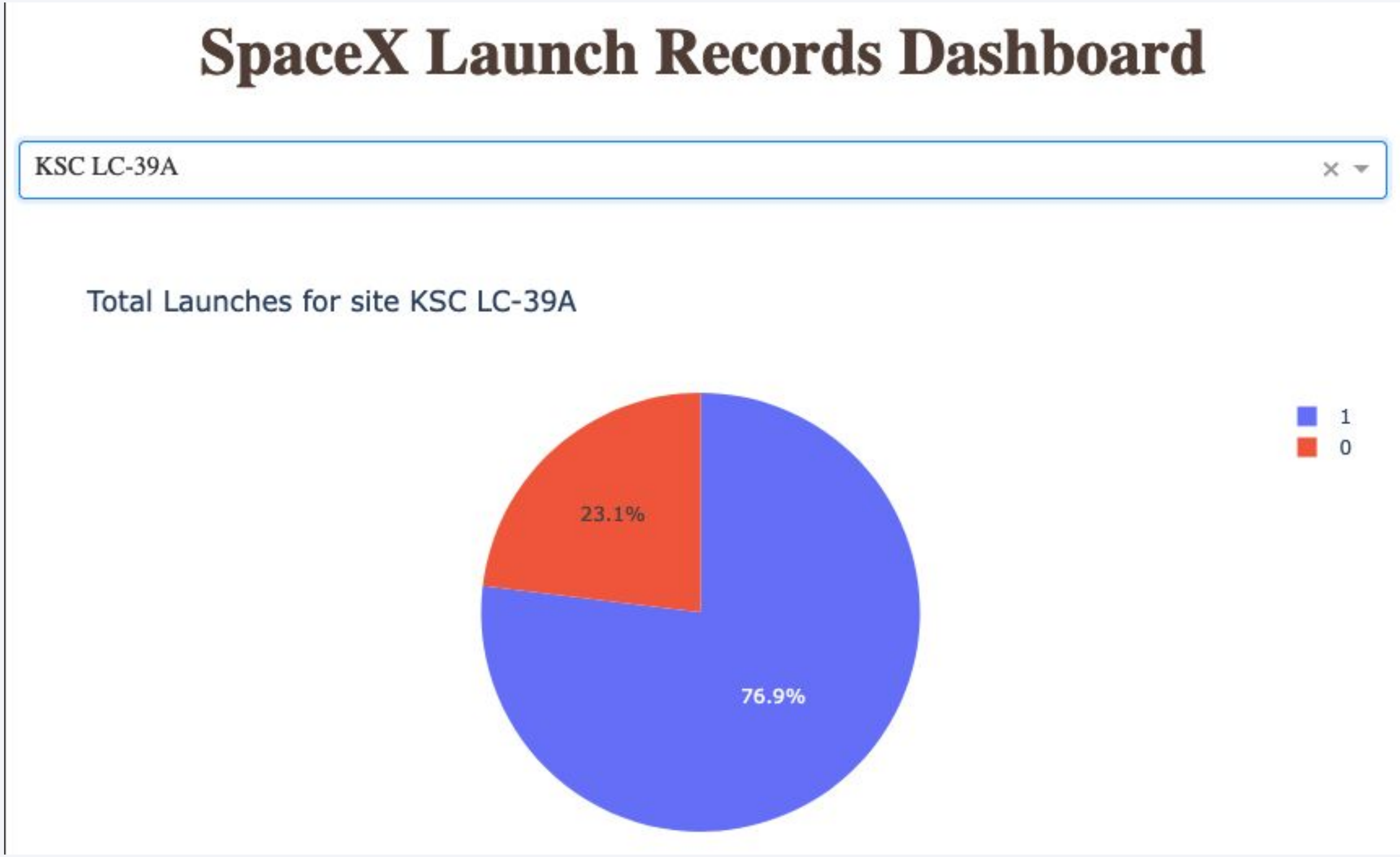
Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site



Launch site with highest launch success ratio



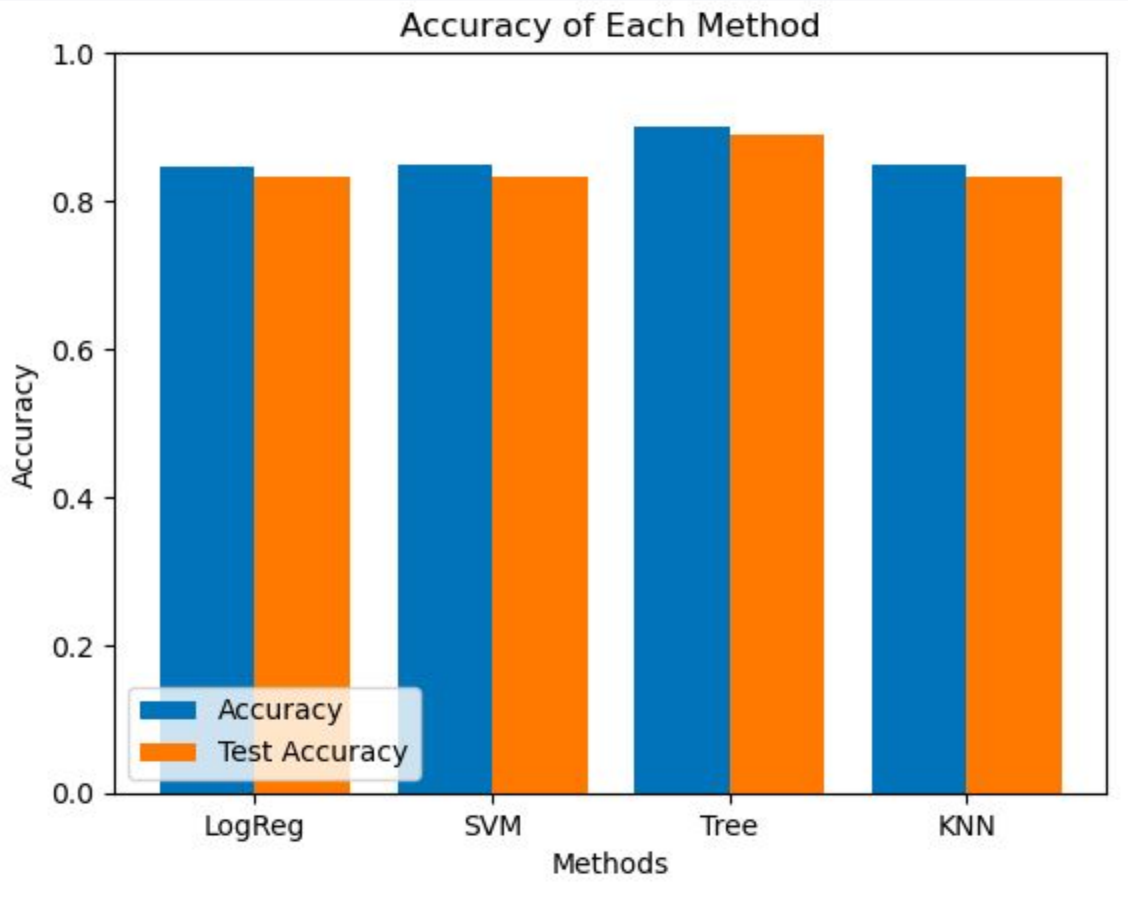
Payload vs. Launch Outcome scatter plot

Section 5

Predictive Analysis (Classification)

Classification Accuracy

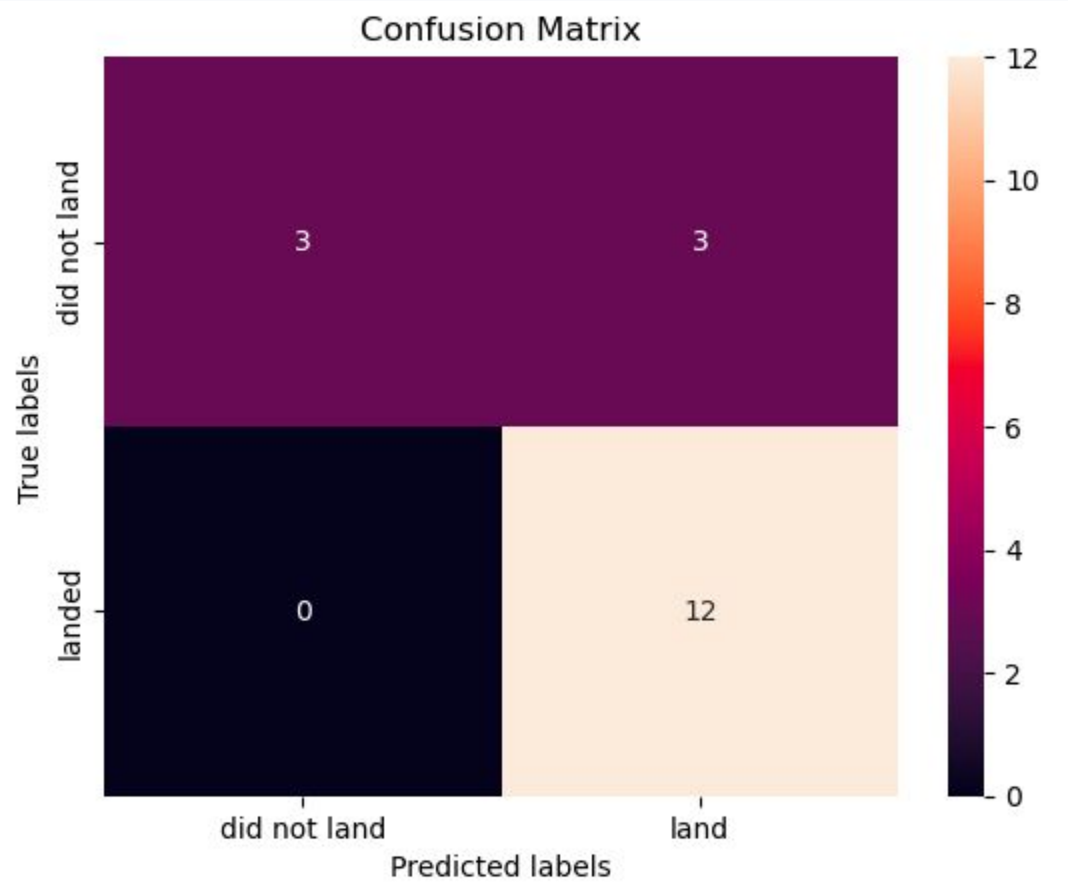
- A bar chart demonstrates accuracy for all built classification models
- Decision Tree Classifier is the model with the highest accuracy, at 87%



Confusion Matrix of Decision Tree Classifier

The confusion matrix beside shows the big numbers of true positive and true negative compared to the false ones.

Therefore, proving the selected method has the best output.



Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!