IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gustavo Reis
February 10, 2022

Outline

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

Executive Summary

- Data was collected using a Web API and Web Scraping methods;
- A data wrangling step was executed to transform raw data into a useful format;
- Data visualization resources were created, including graphs and an interactive dashboard;
- Data models were trained to determine which best fits the data;
- The model with best accuracy was selected to determine to a given launch if the first stage of the rocket is likely to be recovered or lost.

Introduction

- Private space exploration is getting traction.
- The cost of the launches is decreasing mainly thanks to the reuse of the first stage of rockets.
- Determining if the first stage can be recovered or not has a great influence in determining the cost of the mission.
- We want to determine if the first stage of the rocket of a given SpaceX launch is likely to be recovered or lost, to estimate the cost to SpaceX and help a competitor to outbid its price.

Section 1

Methodology

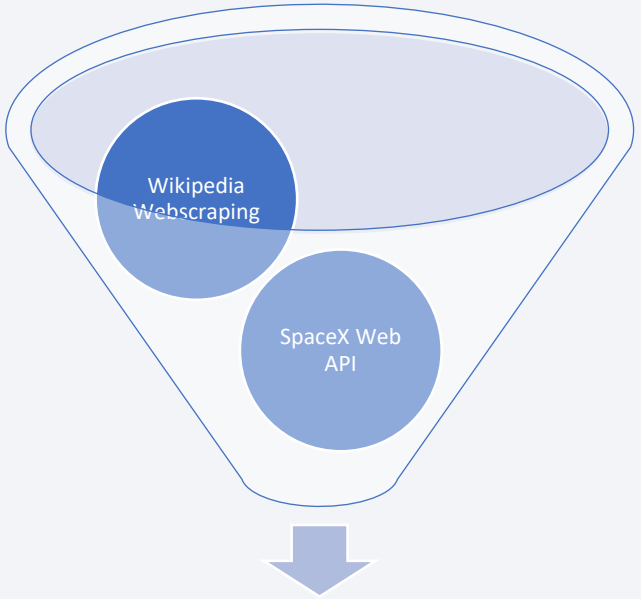
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using an API provided by SpaceX and web scraping the Wikipedia page
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was fit to four models: KNN, SVM, Decision Trees and Logistic Regression models. The best hyperparameters were found, the accuracy of each model evaluated, and the best model was selected.

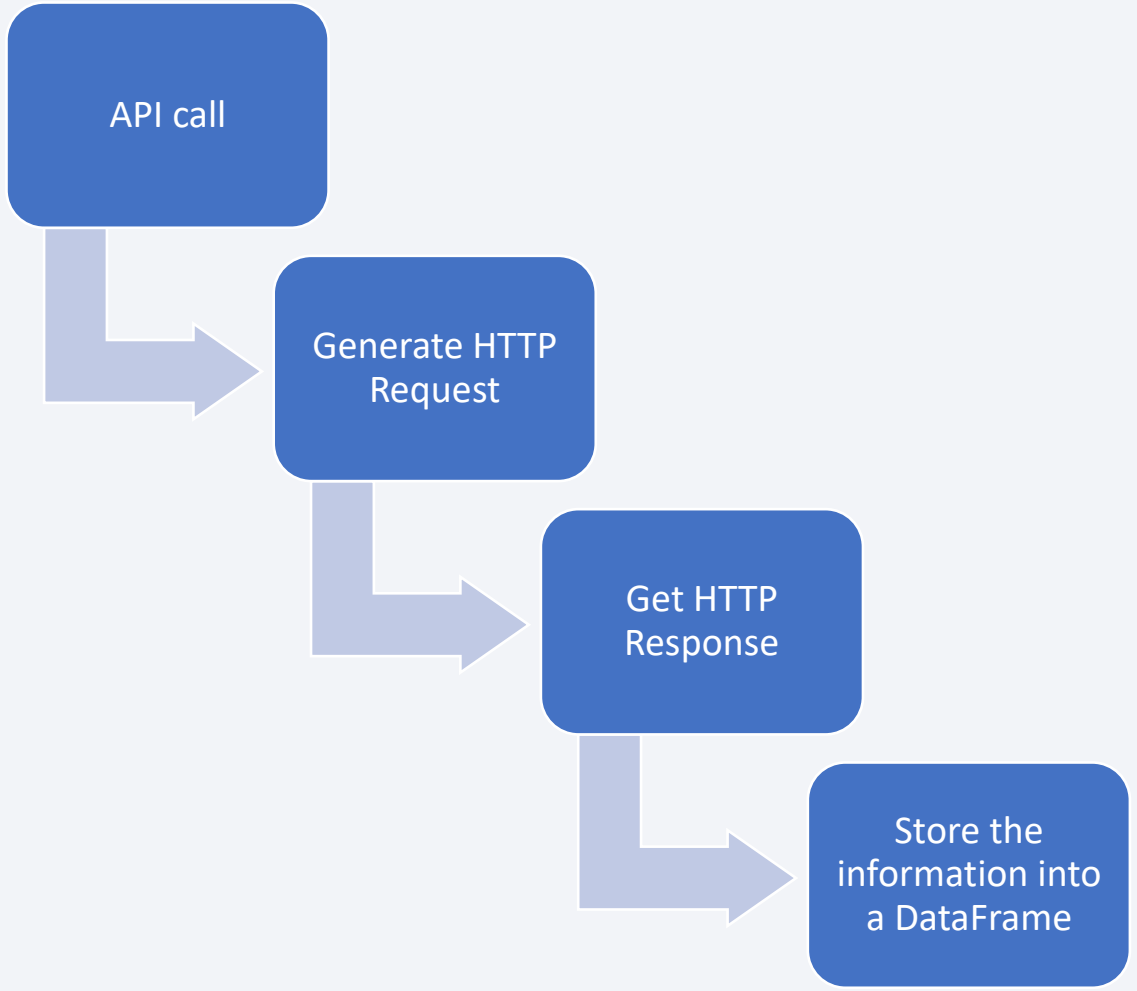
Data Collection

- Data was collected using an API provided by SpaceX and web scraping the Wikipedia page



Data Collection – SpaceX API

- API calls were used to retrieve information about each SpaceX launch and store them into a Pandas DataFrame.
- Predefine functions were used to retrieve specific information.

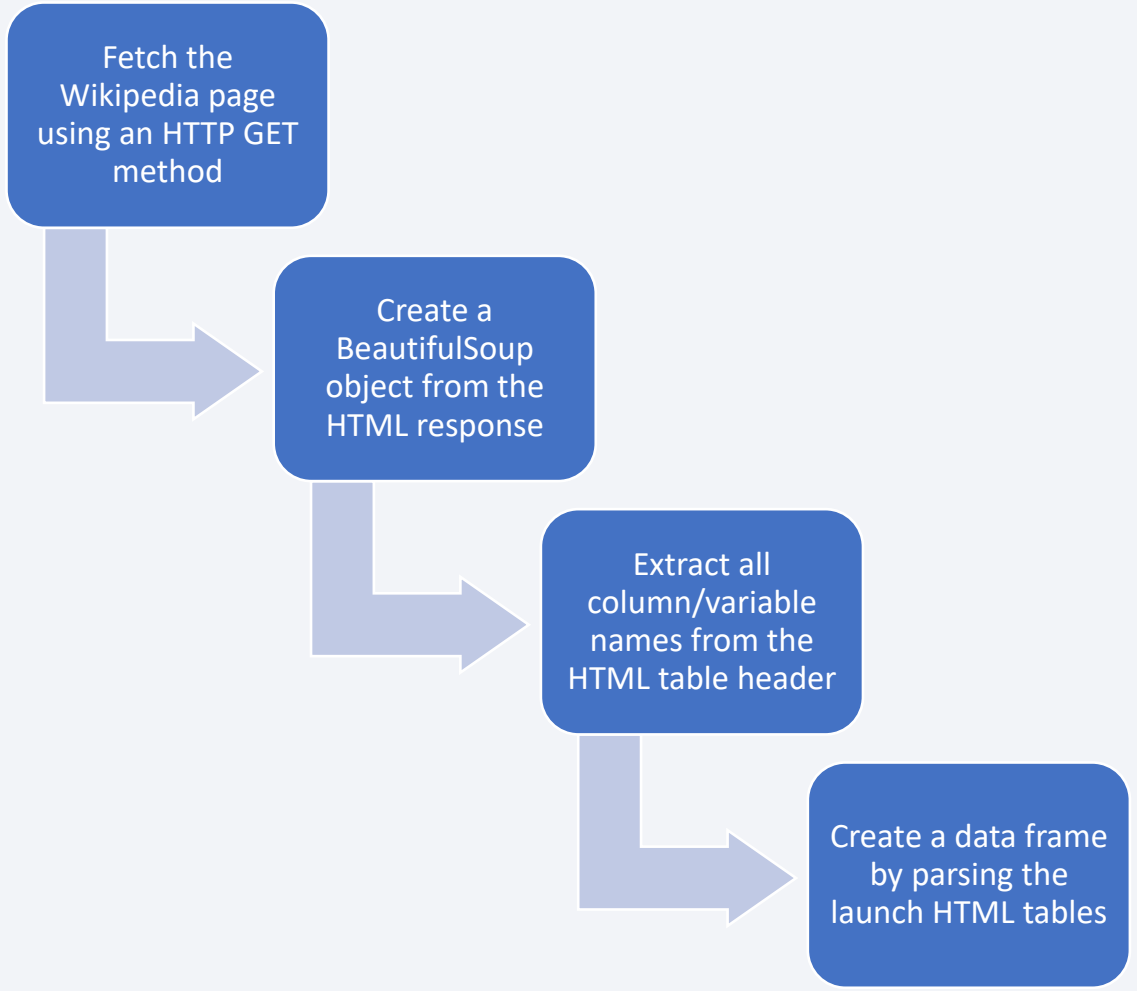


link to the notebook: https://github.com/reisgr/space-y/blob/master/Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

The webscaping process consisted of the following steps:

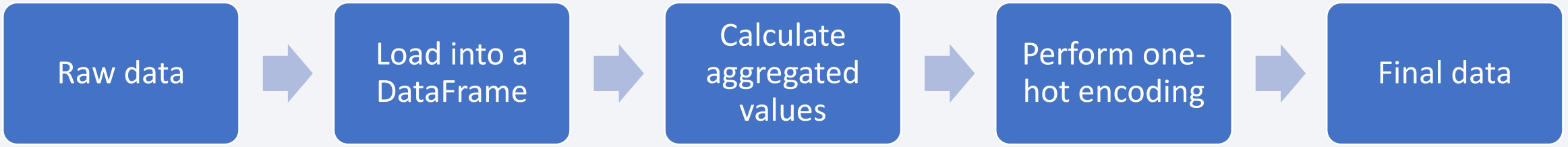
- Download the Wikipedia page containing the data
- Parsing it through a BeautifulSoup object
- Extract the HTML tables and from them extract the relevant data



Link to the notebook: https://github.com/reisgr/space-y/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Data were loaded into a Pandas Dataframe and a few steps where performed:
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type
  - Encode the data into a one-hot encoding scheme to train the models



Link to the notebook: https://github.com/reisgr/space-y/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

EDA with Data Visualization

- Charts plotted:
  - Scatter Plots
    - Flight Number x Launch Site: To show how success rate changed in a give site as de number of launches increased;
    - Payload x Launch Site: To determine if the payload launched from a given site would affect the success rate;
    - Flight Number x Orbit type: To determine if success rate changed as de number of launches increased to a given orbit;
    - Payload x Orbit type: To see if the weight of a payload would influence the success rate when launching to a given orbit
  - Bar chart: Showing the probability of a successful launch based on the target orbit;
  - Line plot: Showing the yearly evolution of success rate.

link to the notebook: https://github.com/reisgr/space-y/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

SQL queries performed:

- Display the names of the unique launch sites in the space mission

```
select distinct(launch_site) from SPACEXDATASET
```

- Display 5 records where launch sites begin with the string 'CCA'

```
select * from SPACEXDATASET
where launch_site like 'CCA%'
limit 5
```

- Display the total payload mass carried by boosters launched by NASA (CRS)

```
select sum(payload_mass__kg_) as Total_payload_mass
from SPACEXDATASET
```

EDA with SQL

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

```
select avg (payload_mass__kg_) from SPACEXDATASET
where booster_version like '%F9 v1.1%'
```

- List the date when the first successful landing outcome in ground pad was achieved

```
select min(DATE) from SPACEXDATASET
```

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
select booster_version from SPACEXDATASET
where payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
```

- List the total number of successful and failure mission outcomes

```
select mission_outcome, count(mission_outcome) from
SPACEXDATASET
group by mission_outcome
```

EDA with SQL

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
select booster_version from SPACEXDATASET
where payload_mass__kg_ = (select max(payload_mass__kg_)
from SPACEXDATASET)
```

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

```
select landing__outcome, booster_version, launch_site from
SPACEXDATASET
where year(date) = 2015 and landing__outcome like
'%Fail%drone%'
```

EDA with SQL

- 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

```
select landing__outcome, count(landing__outcome) from
SPACEXDATASET
where date > '2010-06-04' and date < '2017-03-20'
group by landing__outcome
order by count(landing__outcome) desc
```

link to the notebook: https://github.com/reisgr/space-y/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Circles were created to pinpoint each individual launch in the map;
- Lines were drawn to show the distance to points of interest: coastlines, railroads and highways.

link to the notebook: https://github.com/reisgr/space-y/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20.ipynb

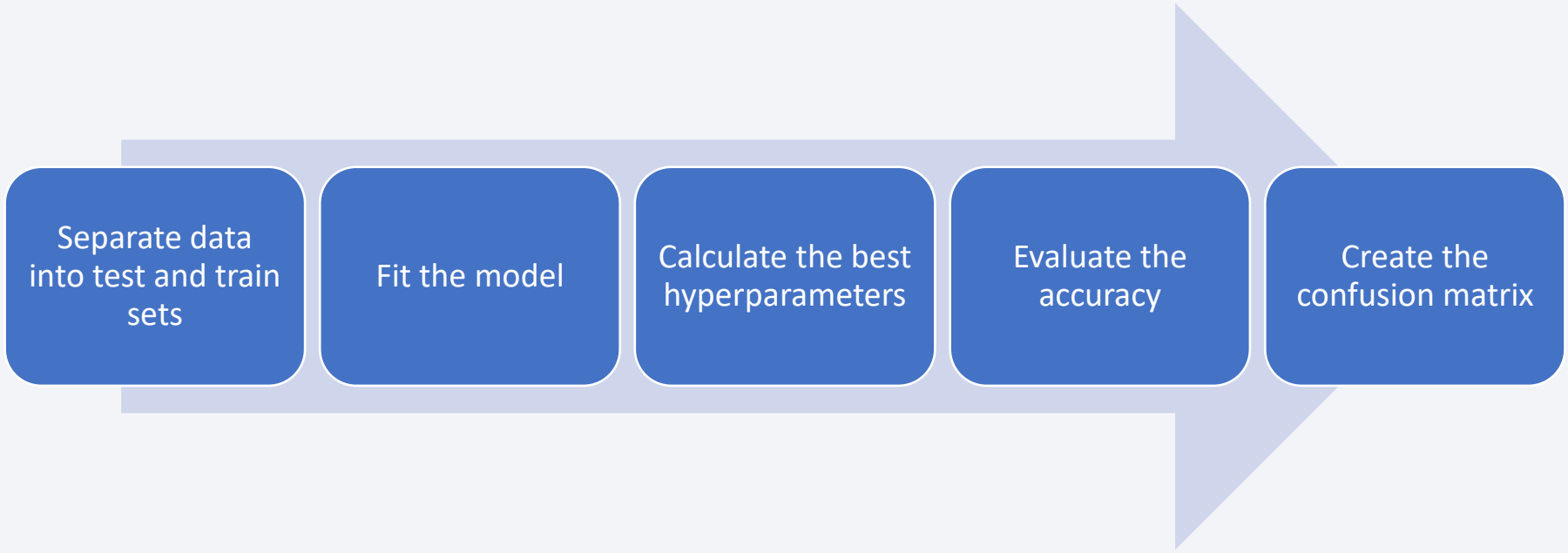
Build a Dashboard with Plotly Dash

- The dashboard has two parts, a pie chart and a scatter plot;
- The pie chart shows successful and unsuccessful launches for a given launch site;
  - It's useful to visualize the proportion of successful launches and evaluate if a site has a good record of missions;
- The scatter plot shows each launch with respect to whether it was successful or not in the y-axis and the payload mass in the x-axis. The color of the point identifies the booster version used for the launch.
  - This allows to visualize each individual launch witch respect to payload mass and booster and see which booster was used.
  - It is also possible to filter for specific launch sites using the drop-down menu at the top of the page
  - There is also a range slider that allows for filtering the payload mass, narrowing down the number of results shown.

Link for the dashboard's python file: https://github.com/reisgr/space-y/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

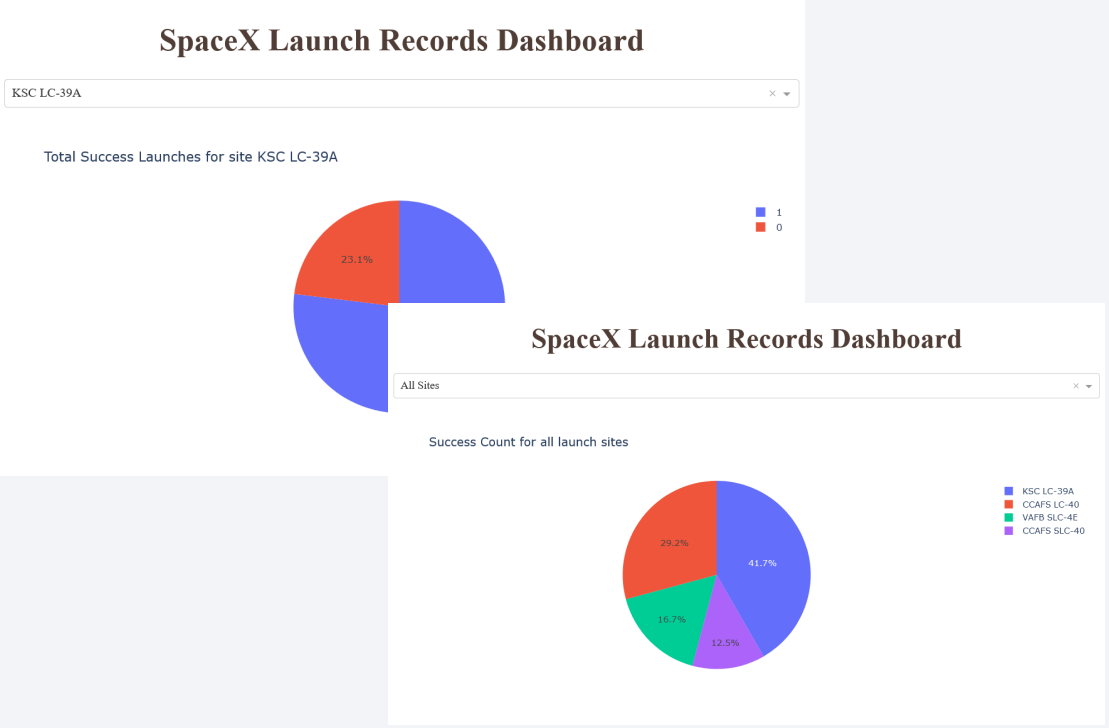
- The data was fit to four models: KNN, SVM, Decision Trees and Logistic Regression models. The best hyperparameters were found, the accuracy of each model evaluated, and the best model was selected.

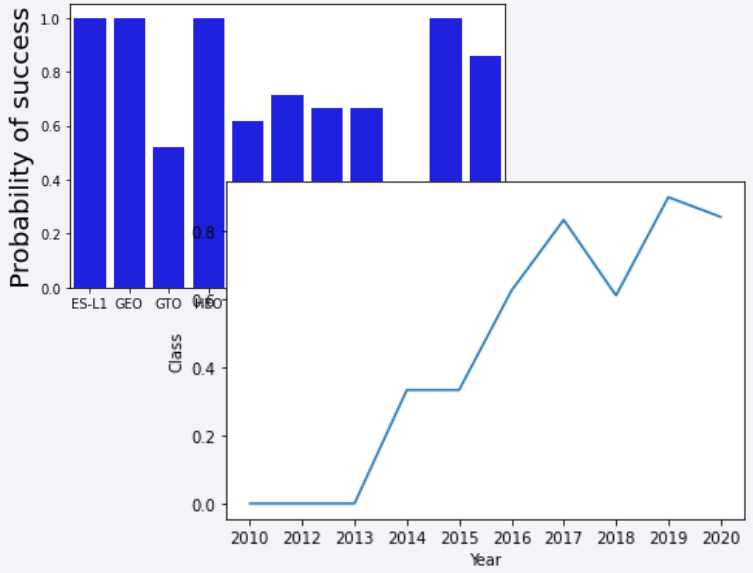


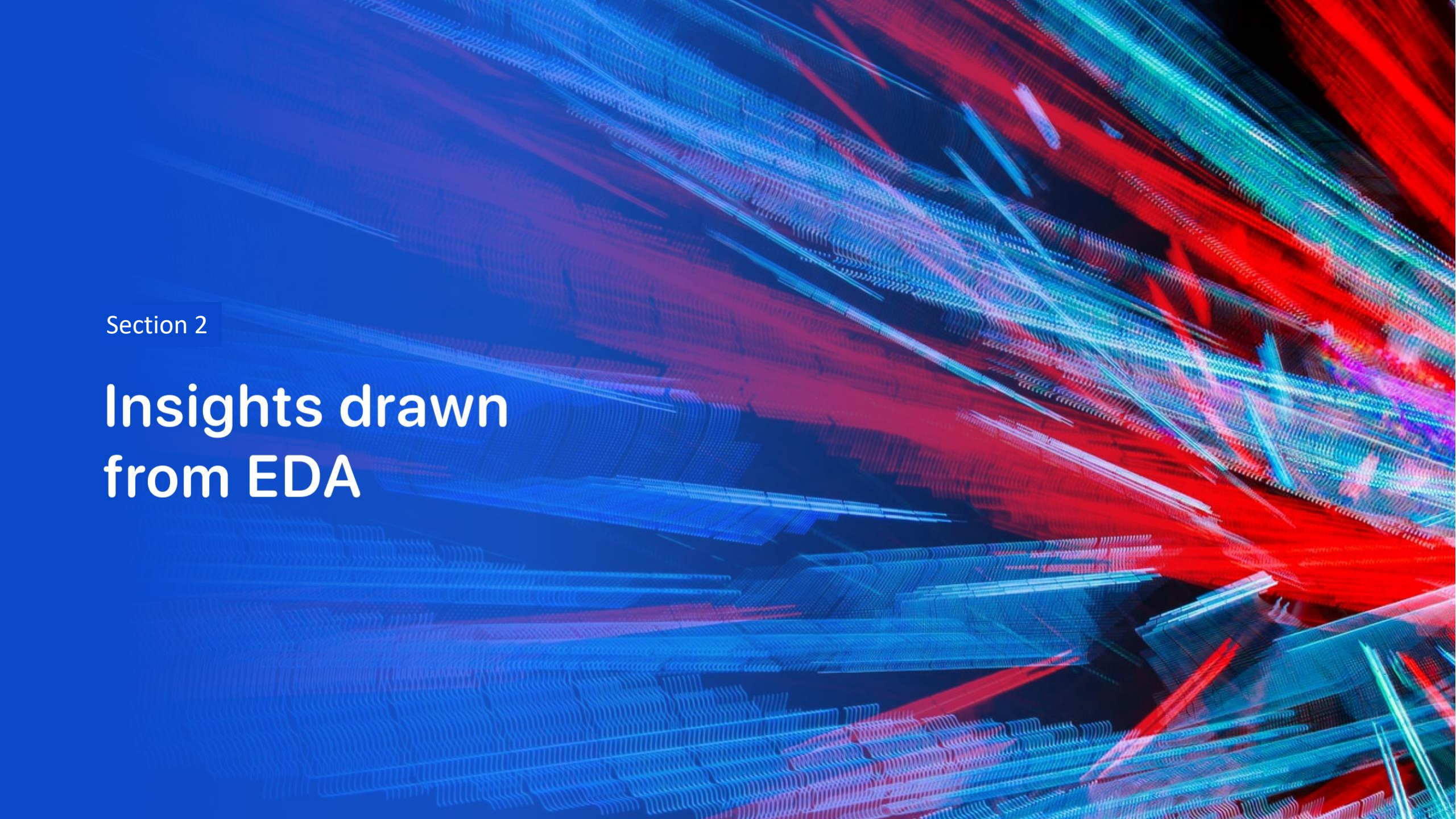
link to the notebook: https://github.com/reisgr/space-y/blob/master/Machine%20Learning%20Prediction%20.ipynb

Results

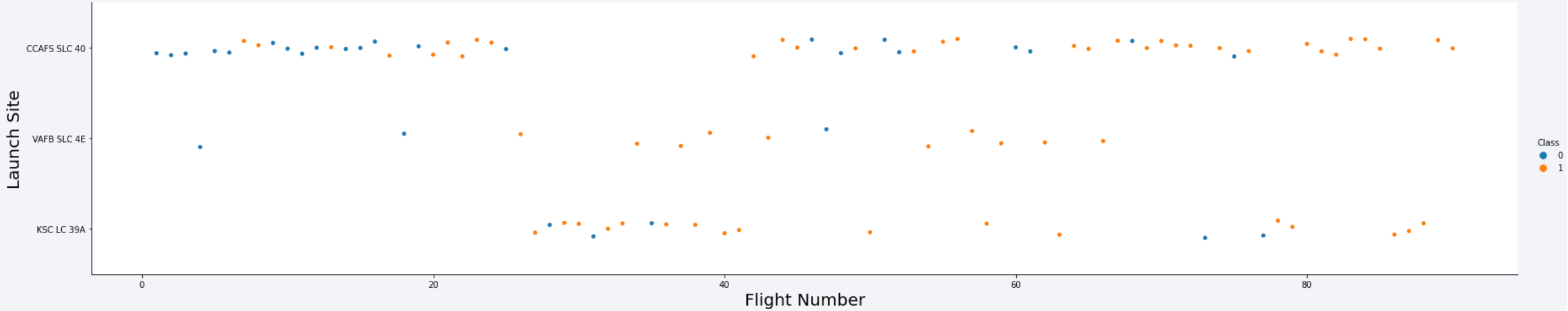
- Data was collected and transformed
- Graphical visualizations were created to help see the data
- A Decision Tree model was selected and trained to make predictions





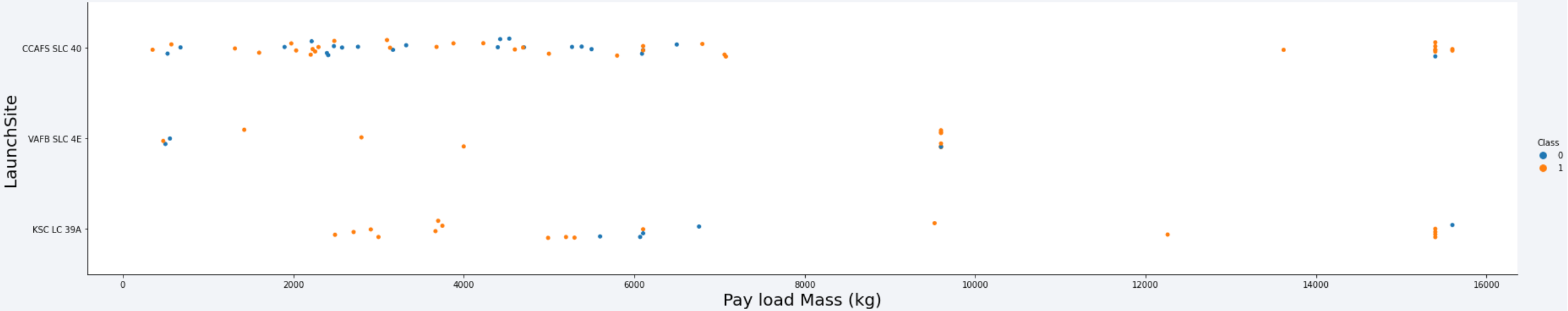

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



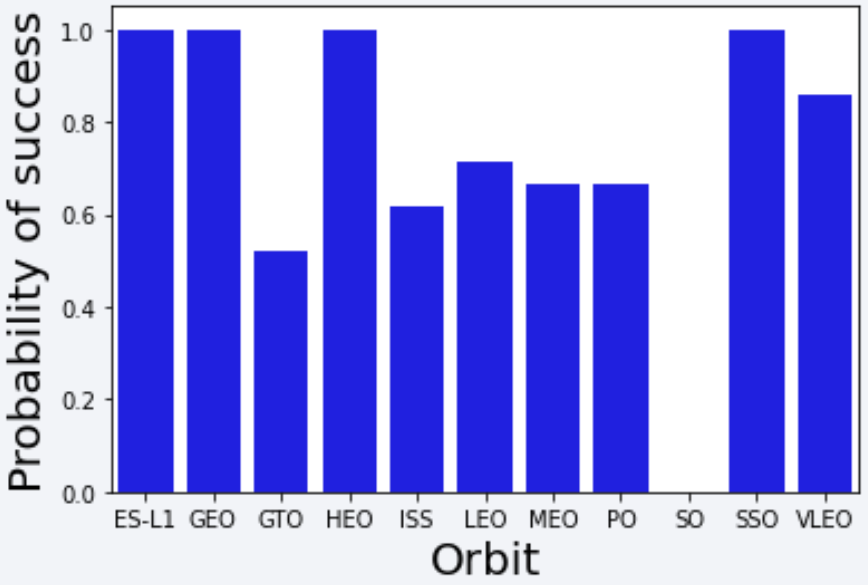
- In general, as the number of flights increases, the likelihood of success also increases;
- One exception is the launch site KSC LC 39A, where proportion of bad outcomes doesn't seem to decrease over time.

Payload vs. Launch Site



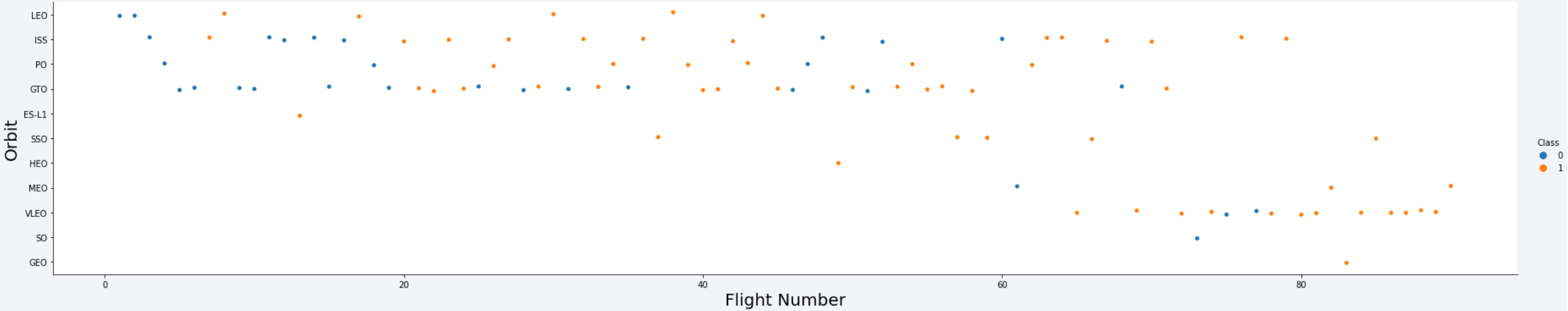
- Heavier payloads seems positively correlated to likelihood of success;
- For launch site KSC LC 39, light payloads (2,000 to 4,000 Kg) also indicates success;
- No heavy payloads (> 10,000 Kg) where launched from launch site VAFB-SLC.

Success Rate vs. Orbit Type



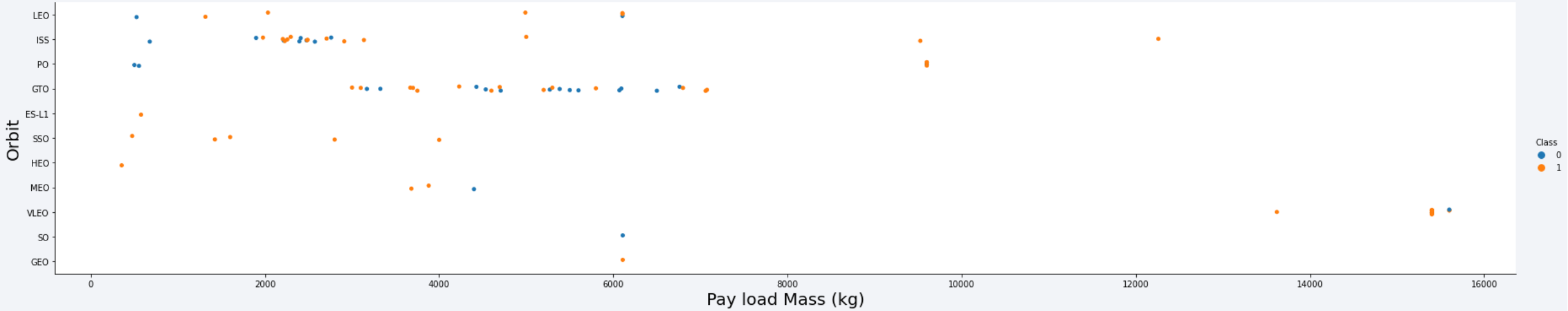
- Rockets launched to ES-L1, GEO, HEO and SSO have the highest probability of success (close to 100%);
- Rockets launched to GTO have the lowest probability of success (just above 50%).

Flight Number vs. Orbit Type



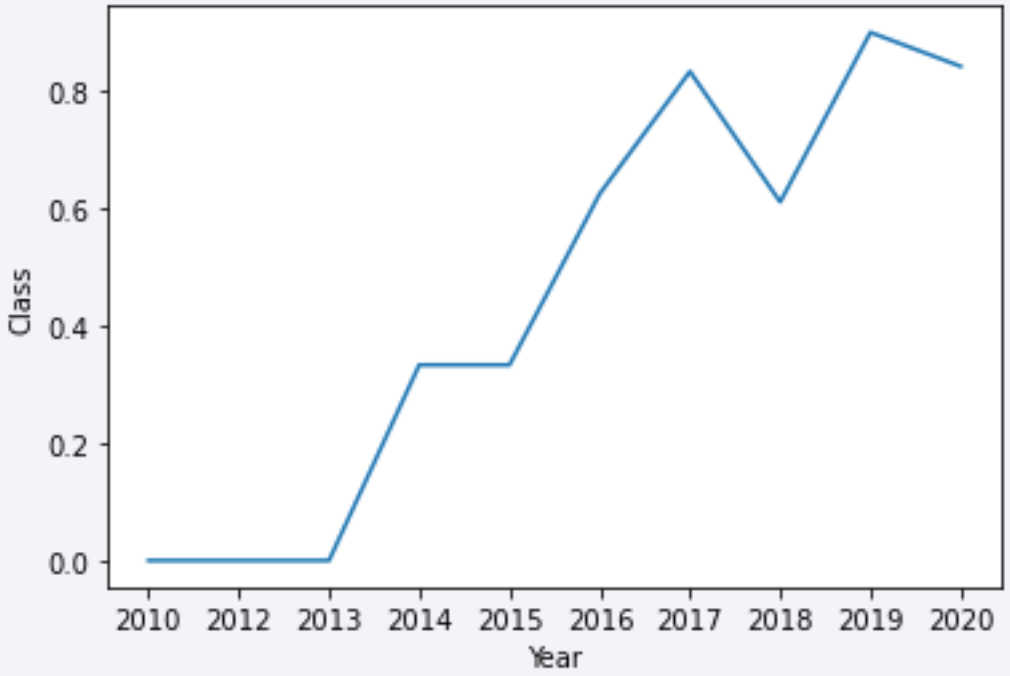
- LEO seems related to the number of flights;
- Other orbits doesn't appear to show any relationship to the number of flights.

Payload vs. Orbit Type



- Heavy payloads where delivered (mostly successfully) to Polar, Low Earth Orbit and the ISS orbit;
- For the Geostationary Transfer Orbit, there seems to be no correlation between success and payload mass

Launch Success Yearly Trend



- Starting 2013, and except for 2018, every year the success rate increases.

All Launch Site Names

- These are the names of all launch sites used by SpaceX obtained used a SQL query:

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

Launch Site Names Begin with 'CCA'

- These are the first five records where the launch_site column contains the substring “CCA%”

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

- This is the sum in kilograms of the payload mass from all launches obtained using the SQL SUM function:

Total_payload_mass
619967

- The total payload mass from all launches is approximately 620 tons

Average Payload Mass by F9 v1.1

- This is the average in kilograms of the payload mass from all launches using the F9 v1.1 rocket obtained using the SQL AVG function:

Avg_payload_mass
2534

- The average payload mass is about 2,500 Kg.

First Successful Ground Landing Date

- This is the date of the first successful landing outcome in a ground landing pad:

1st_landing_date
2010-06-04

- The first successful landing was on the 4th of June of 2010

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of the boosters which have success in drone ship landing and have payload mass greater than 4,000 but less than 6,000 Kg

booster_version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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

- Names of the booster which have carried the maximum payload mass
- All of them are version F9 B5

2015 Launch Records

- Failed landing outcomes in drone ship, booster versions, and launch site names for year 2015
- Both events happened using booster version F9 v1.1 launched from CCAFS CL-40

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

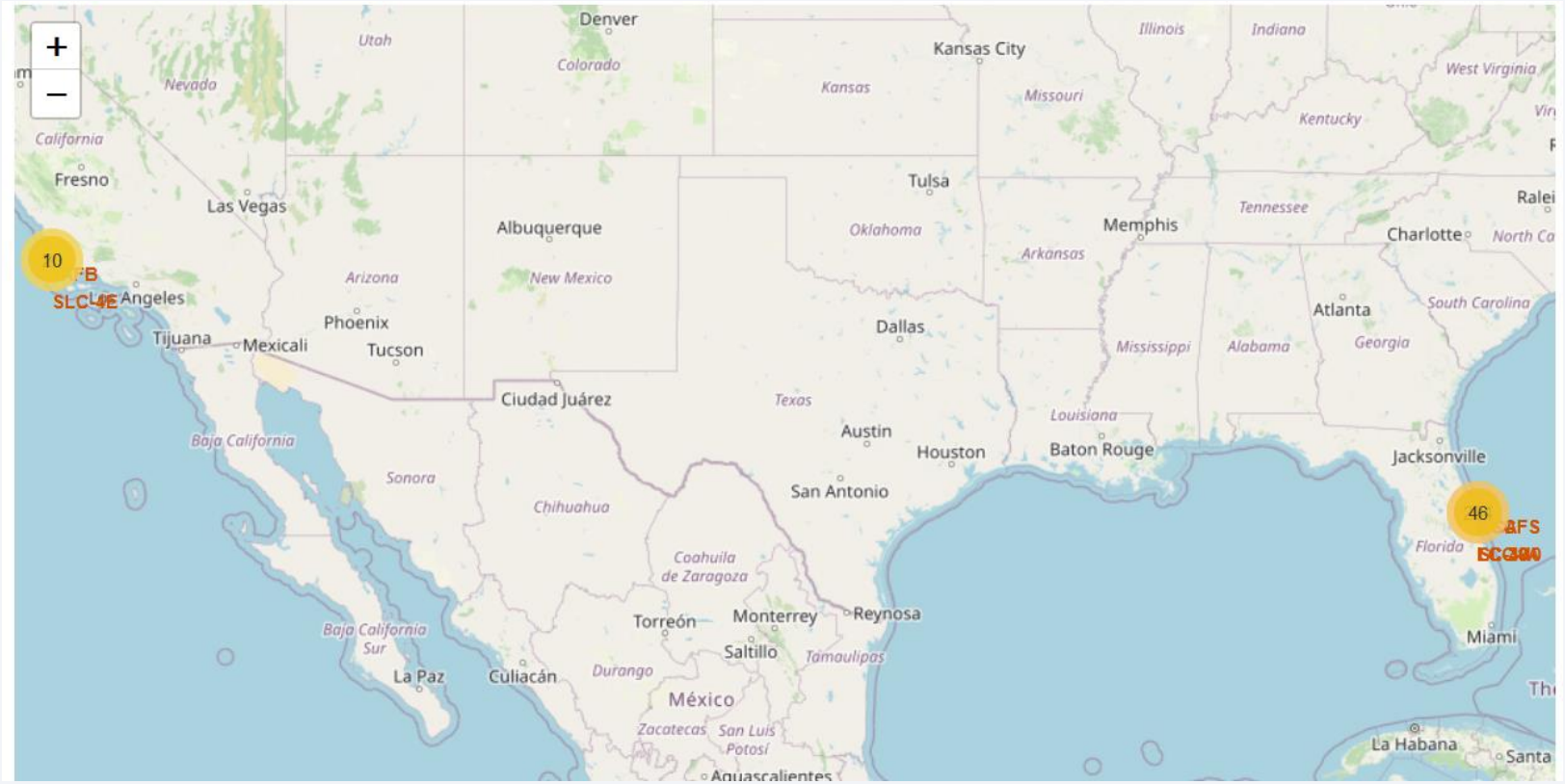
- Attempted landings account for about two thirds of missions, with a success rate close to 60%

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

Section 3

Launch Sites Proximities Analysis

Launch Sites Location

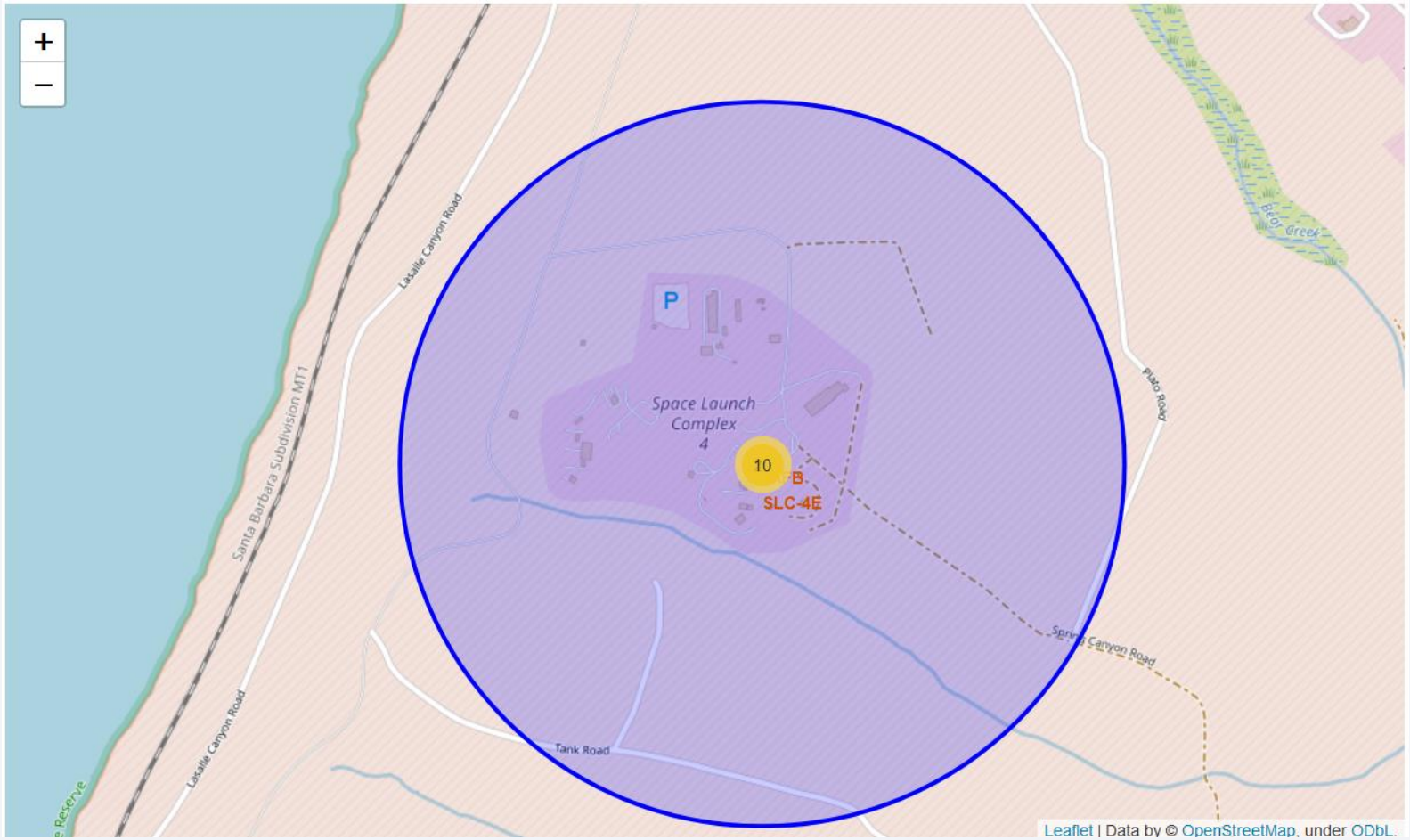


- Each circle represents a launch site with the number of launches from that site

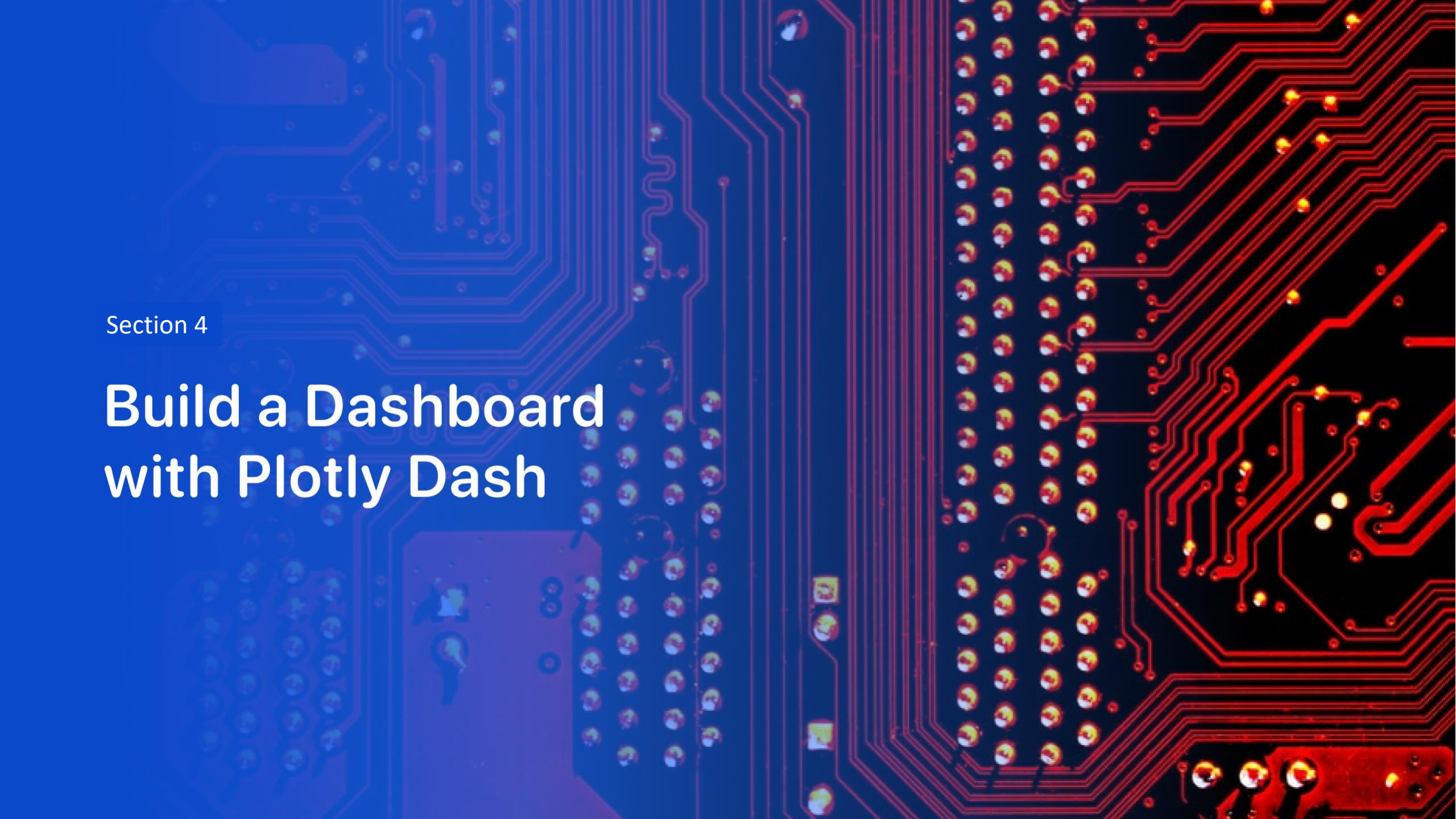
Individual launches

- When clicking the circle corresponding to a launch site, it expands and shows each individual launch color-coded to the outcome of the mission: green means success and red means failure.

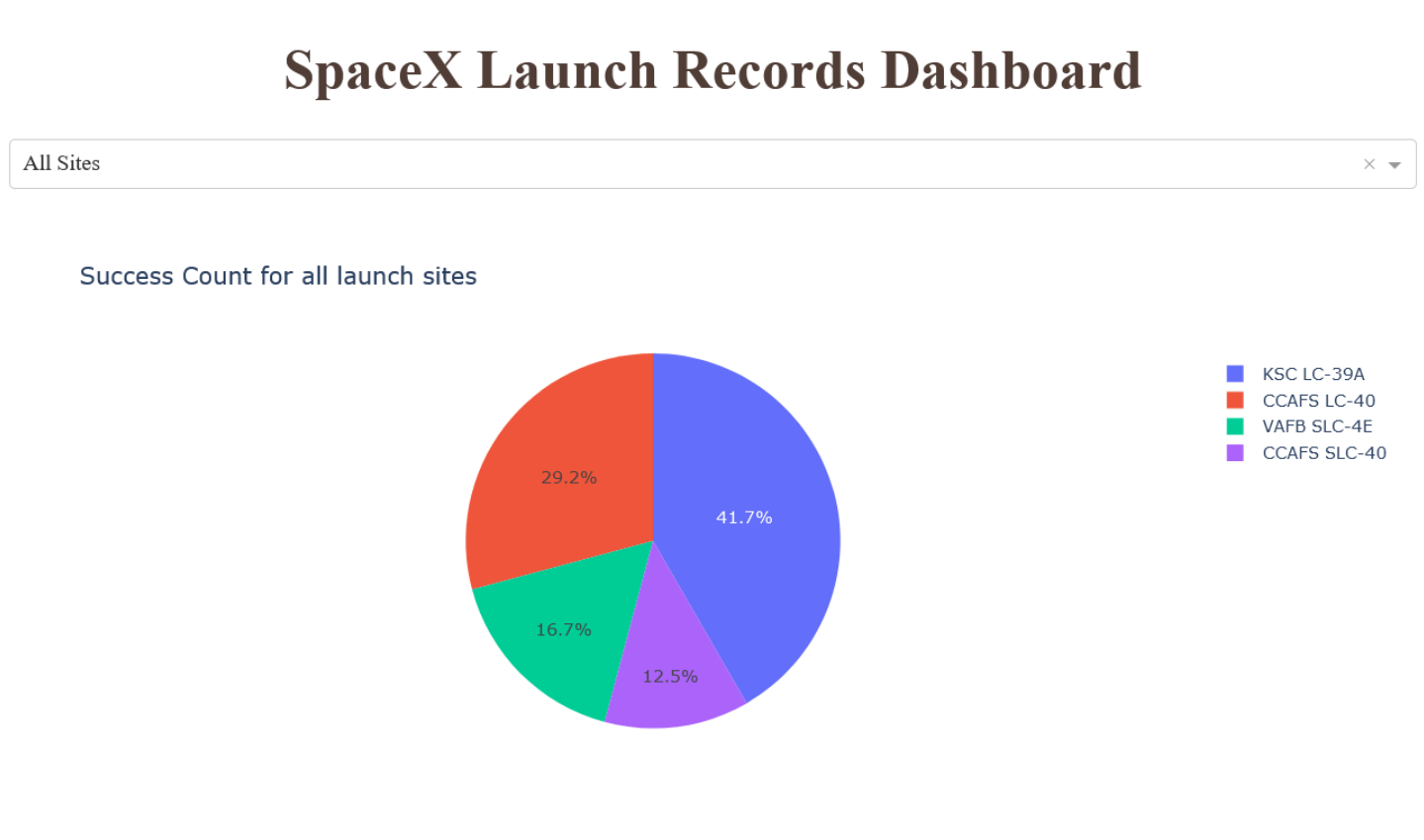
Proximity to nearby features



- We can see the proximity of a given launch site to highways, railways and coastlines.

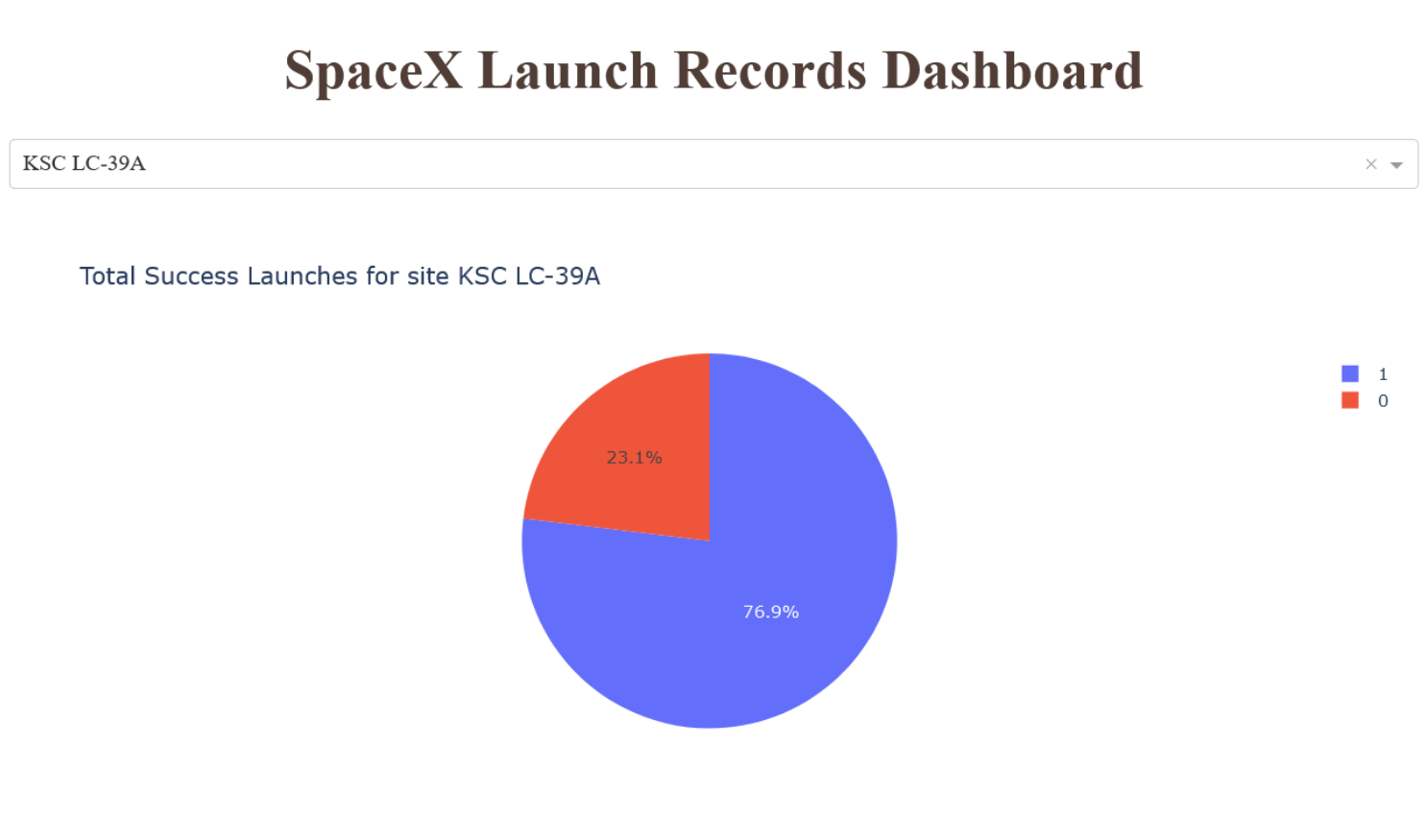

Section 4
Build a Dashboard with Plotly Dash

Success Count for all launch sites



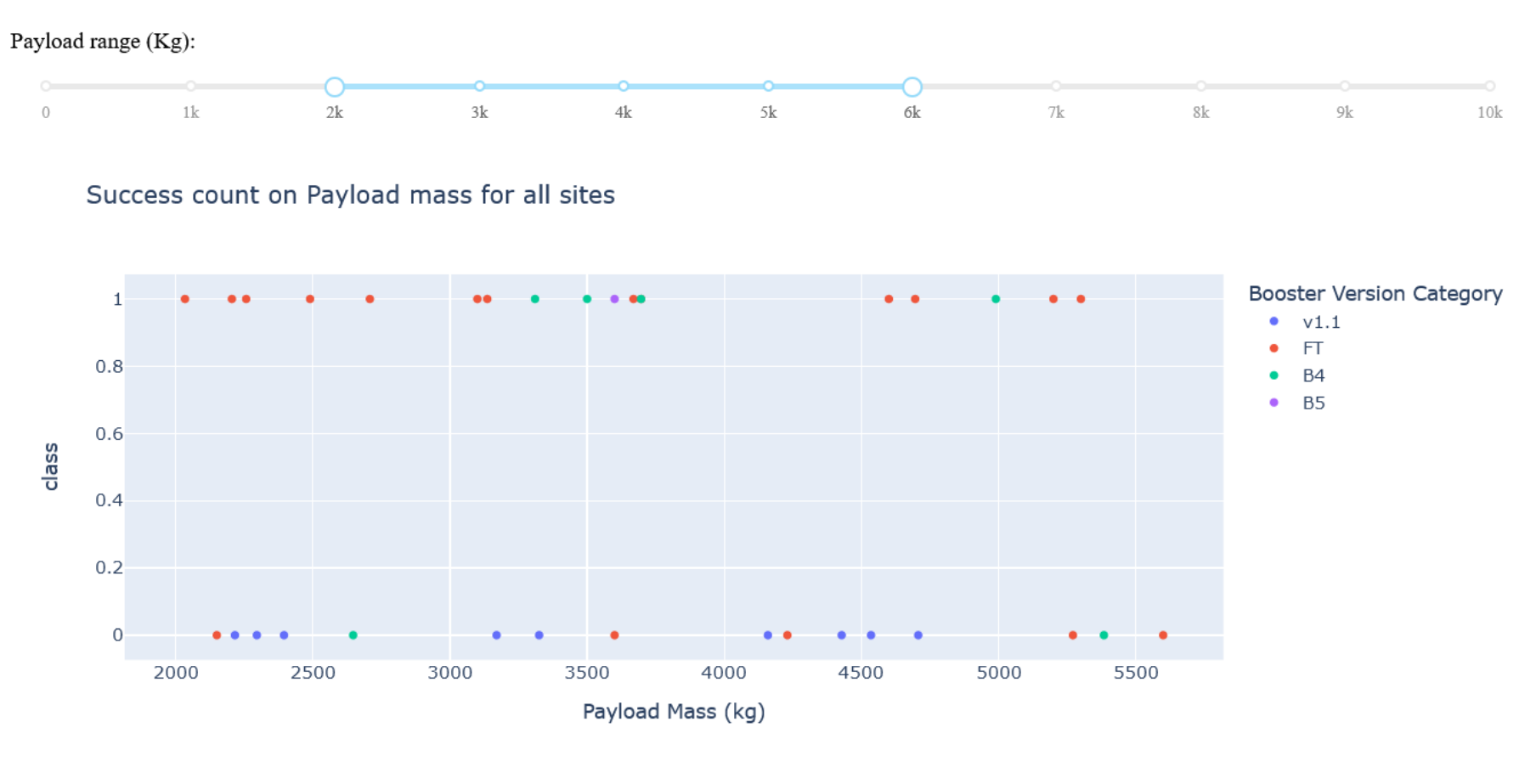
- The launch site that responds to most of the successful launches is site KCS LC-39A

Total Success Launches for site KSC LC-39A



- 76.9% of launches from site KSC LC-39A where successful, and 23.1% failed

Success count on Payload mass for all sites

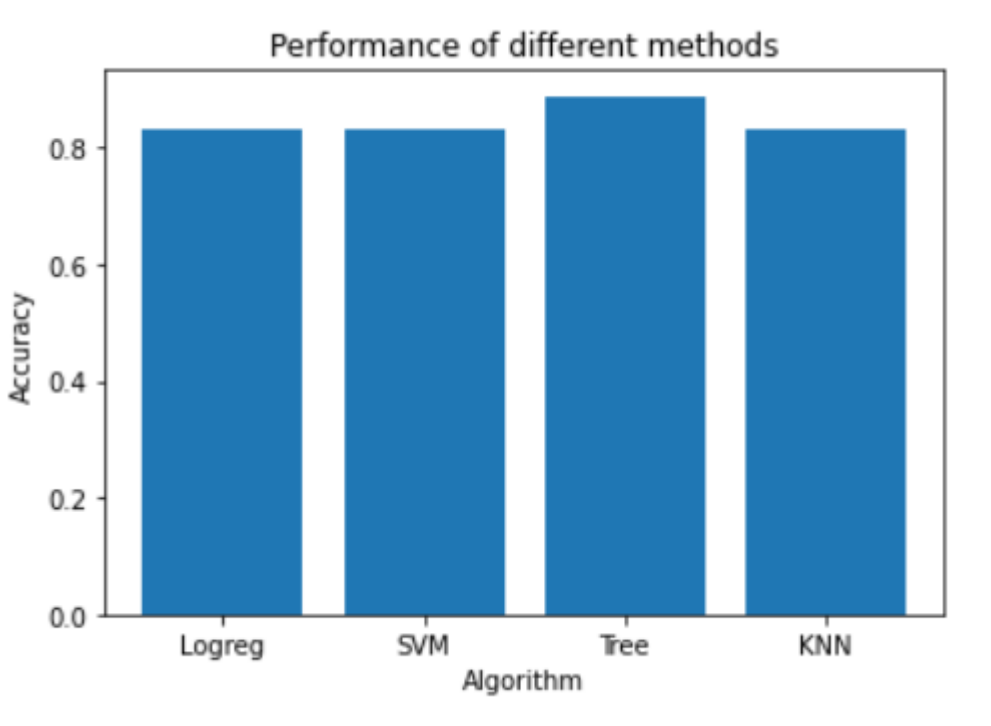


- We are visualizing the success count for payload masses between 2,000 and 6,000 Kg
- Most of the successes occurred when using booster FT, and most of the failures when using booster v1.1

Section 5

Predictive Analysis (Classification)

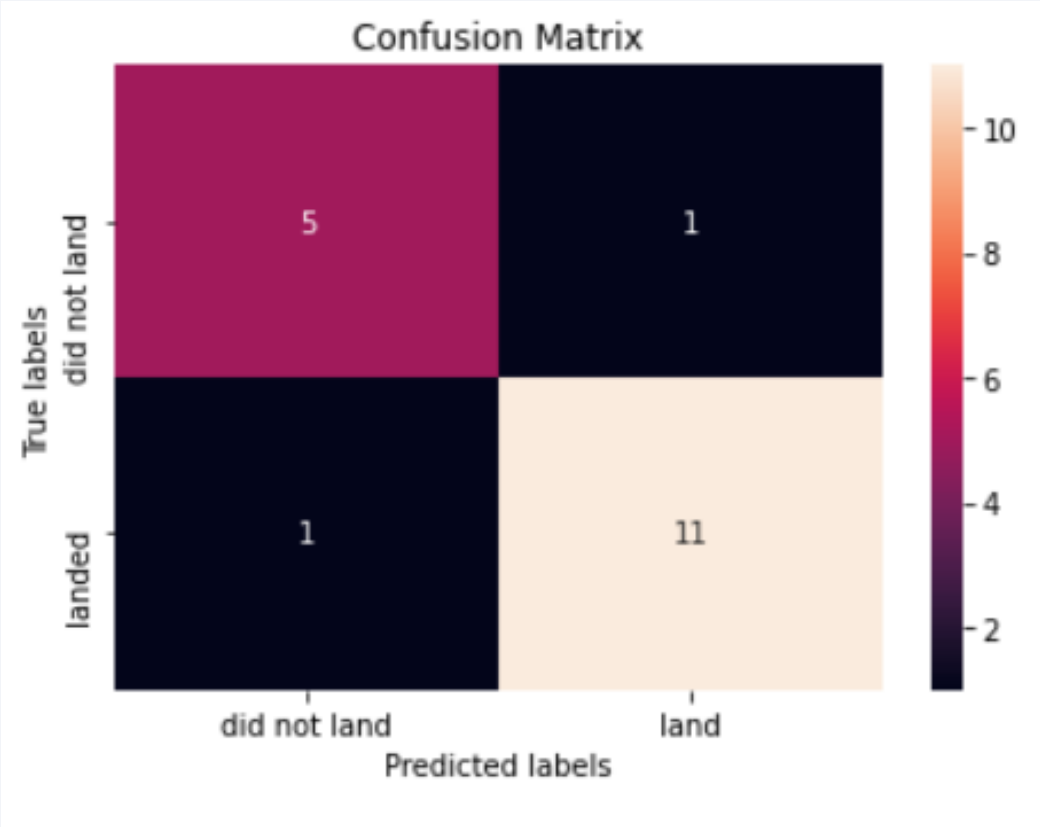
Classification Accuracy



Model	Accuracy
LogReg	0.833333
SVM	0.833333
Tree	0.888889
KNN	0.833333

- All methods show a high accuracy, with a slight advantage to the decision tree

Confusion Matrix



- The model correctly classify most of the data, with only one case of false positive e one of false negative

Conclusions

- We can predict the likelihood of the first stage of a rocket being recovered of lost;
- Some launch sites have a higher probability of recovery;
- Heavy payloads means a higher chance of success;
- Success rate is increasing over the years.

Thank you!