IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ramon Avalos
November 7th, 2022

Outline

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

Executive Summary

- In this capstone, we predict whether the first stage of the Falcon 9 Space X rocket will land successfully. This will be determined by using various machine learning classification algorithms on past launch data.
- The methodology used includes data collection, data wrangling, exploratory data analysis, data visualization, and lastly, machine learning prediction.
- In our analysis, it is determined that various parameters of rocket launches have a strong correlation with the success rate of launches. It was determined that a decision tree classifier, was the best machine learning classifier to determine whether a mission will be a success or failure.

Introduction

- Unlike other rocket makers, Space X's Falcon 9 first stage can be recovered. Sometimes the first stage can't be successfully recovered based on data such as payload, orbit, and launch site. Our main goal in this project is to determine whether the first stage of the Falcon 9 will land successfully. If we can determine the reusability of the rocket, it can save Space X millions on. This can give Space X an advantages on bidding for future launches for space missions.
- The Question: With a given set of launch parameters, will the first stage of the Falcon 9 land successfully?

Section 1

Methodology

Methodology

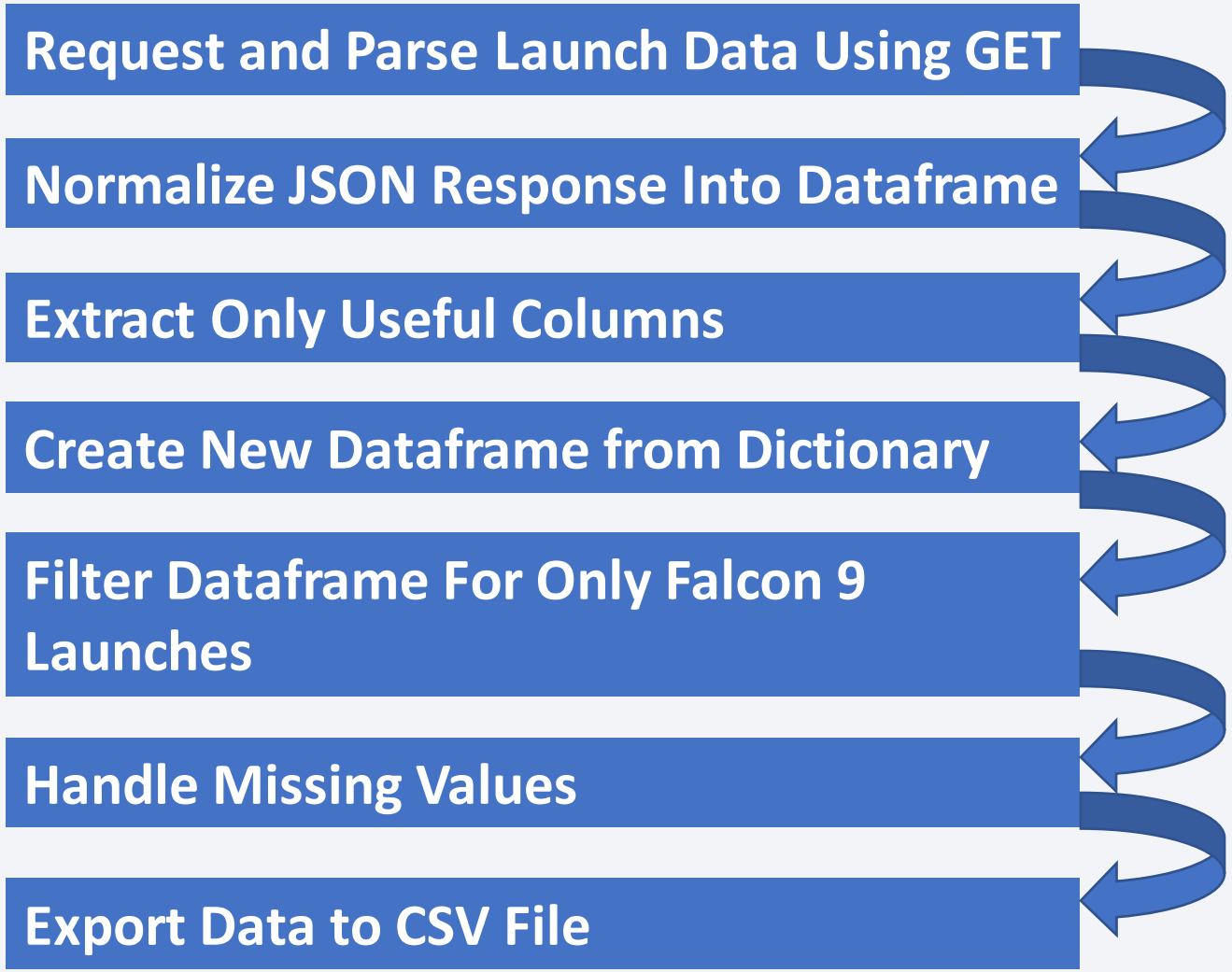
Executive Summary

- Data collection methodology:
  - Data was collected with two methods: requesting data from Space X's API and web scraping launch data from Wikipedia.
- Perform data wrangling
  - After collecting data, data wrangling was performed to transform and clean the data using the pandas library in Python.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - For our predictive analysis, four machine learning classification models were trained, tuned and evaluated to find the most accurate. The models used were logistic regression, support vector machines, k-nearest neighbor, and decision tree classifier.

Data Collection

- We first collected data from Space X's API's. To do this, a GET request was used to request and parse the launch data. Then the JASON response was normalized into a pandas dataframe. We then extracted only useful columns. A new pandas dataframe was created from a dictionary that included the relevant paramters. Then we filtered the dataframe to only include Falcon 9 launches. We then handled the missing values by finding the mean payload mass and using the mean to replace np.nan values in that column. The data was then save as a CSV file.
- Secondly, we collected data with web scrapping. To do this, we first requested launch data from Falcon 9 Wikipedia page. We then extracted all of the columns from the HTML table header. A dataframe was then created by parsing the launch HTML tables. The data was then saved as a CSV file.

Data Collection – SpaceX API

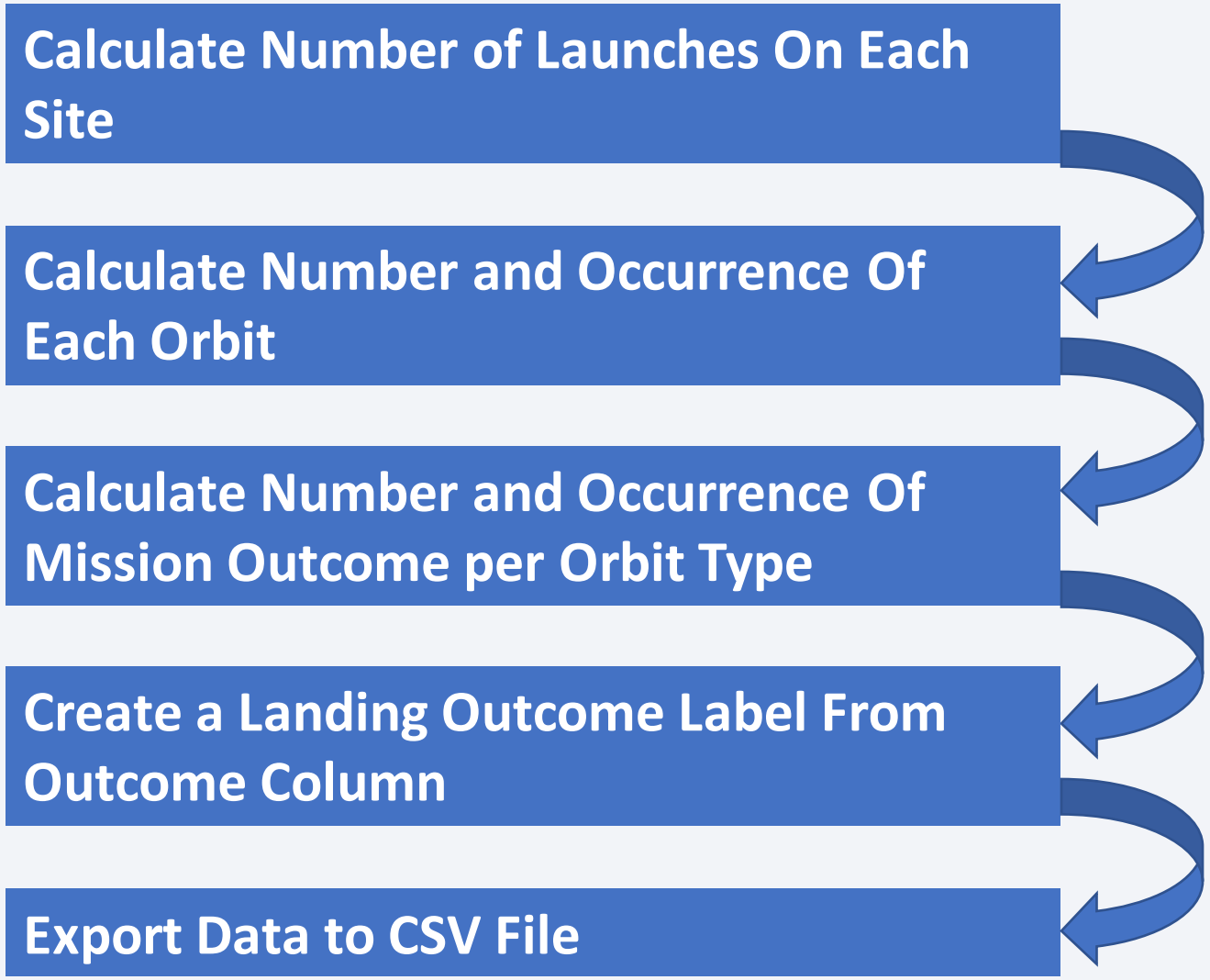


- GitHub URL: Data Collection

Data Collection - Scraping

- GitHub URL: Web Scrapping

Data Wrangling



- GitHub URL: Data Wrangling

EDA with Data Visualization

- Scatter Plots: Used to show relationship between two variables. The parameters compared includes Flight Number vs Launch Site, Payload vs Launch Site, Flight Number vs Orbit Type, and Payload vs Orbit Type.

- Bar Chart: Used to compare the success rates for different orbit types.

- Line Chart: Used to show success rate over a certain number of years.

- Github URL: EDA with Data Visualization

EDA with SQL

SQL Queries Performed:

- Display the names of the unique launch sites in the space mission, 5 records where launch sites begin with the string 'CCA, the total payload mass carried by boosters launched by NASA (CRS) and the average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved, the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000, the total number of successful and failure mission outcomes, the names of the booster versions which have carried the maximum payload mass using a subquery, and the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

GitHub URL: EDA with SQL

- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

- Marker objects, marker clusters, and line objects were created and added to a folium map. Marker objects were used to show launch locations. Other markers were used to show failed or successful launches at each launch site with the use of marker clusters. Line objects were used to show the distance between a launch site and nearby structures or locations.
- These objects were used to answer four basic questions:
  - Are launch sites in close proximity to railways? - Yes
  - Are launch sites in close proximity to highways? - Yes
  - Are launch sites in close proximity to coastline? - Yes
  - Do launch sites keep certain distance away from cities? - Yes

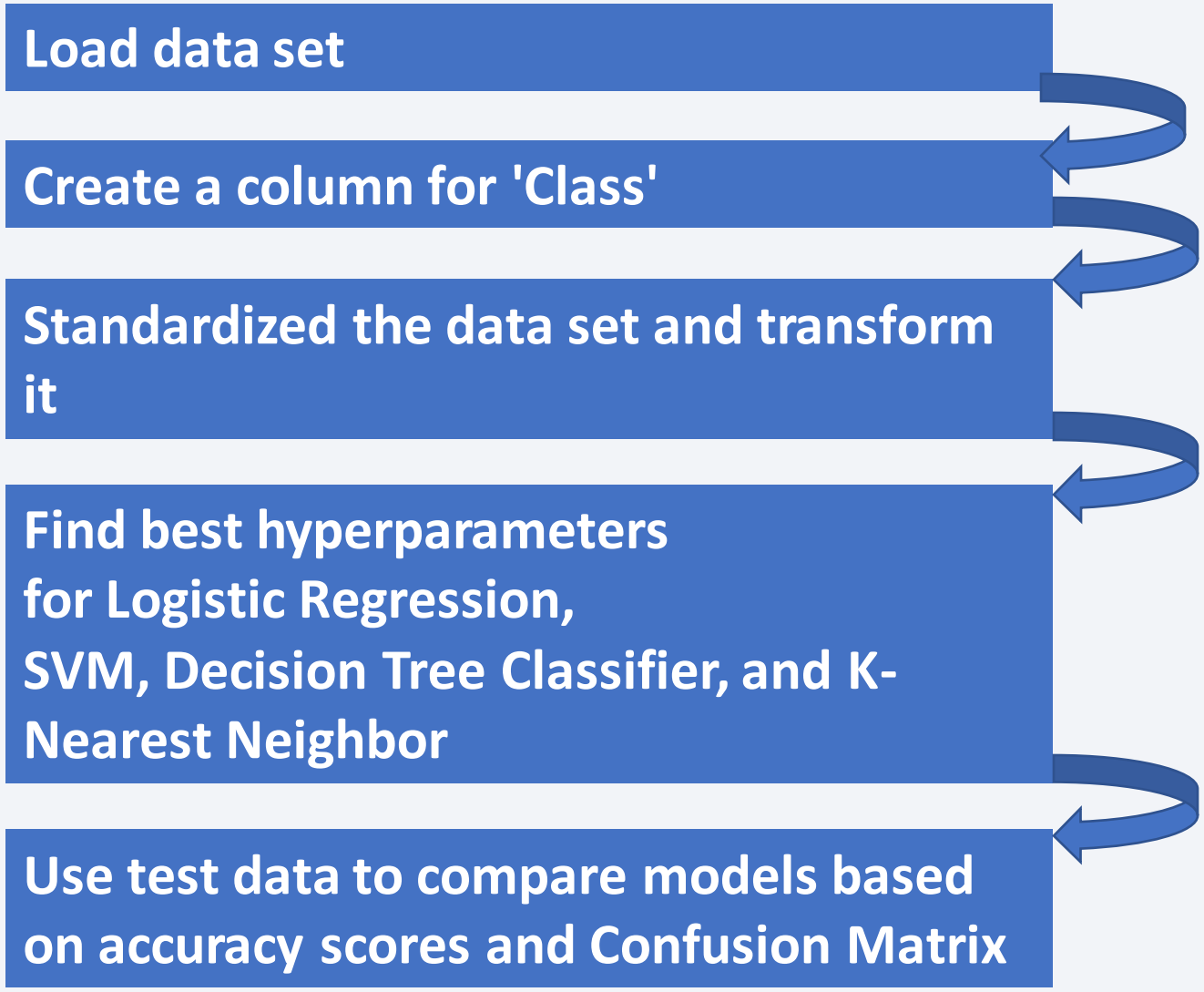
GitHub URL: Interactive Map with Folium

Build a Dashboard with Plotly Dash

- The dashboard has dropdown menu that lets you choose to show the successful launches for all sites or for each individual site with a pie graph. A slider is used to choose payload range. A scatterplot was used to show the relationship between payload and launch success.

GitHub URL: Space X Dashboard

Predictive Analysis (Classification)



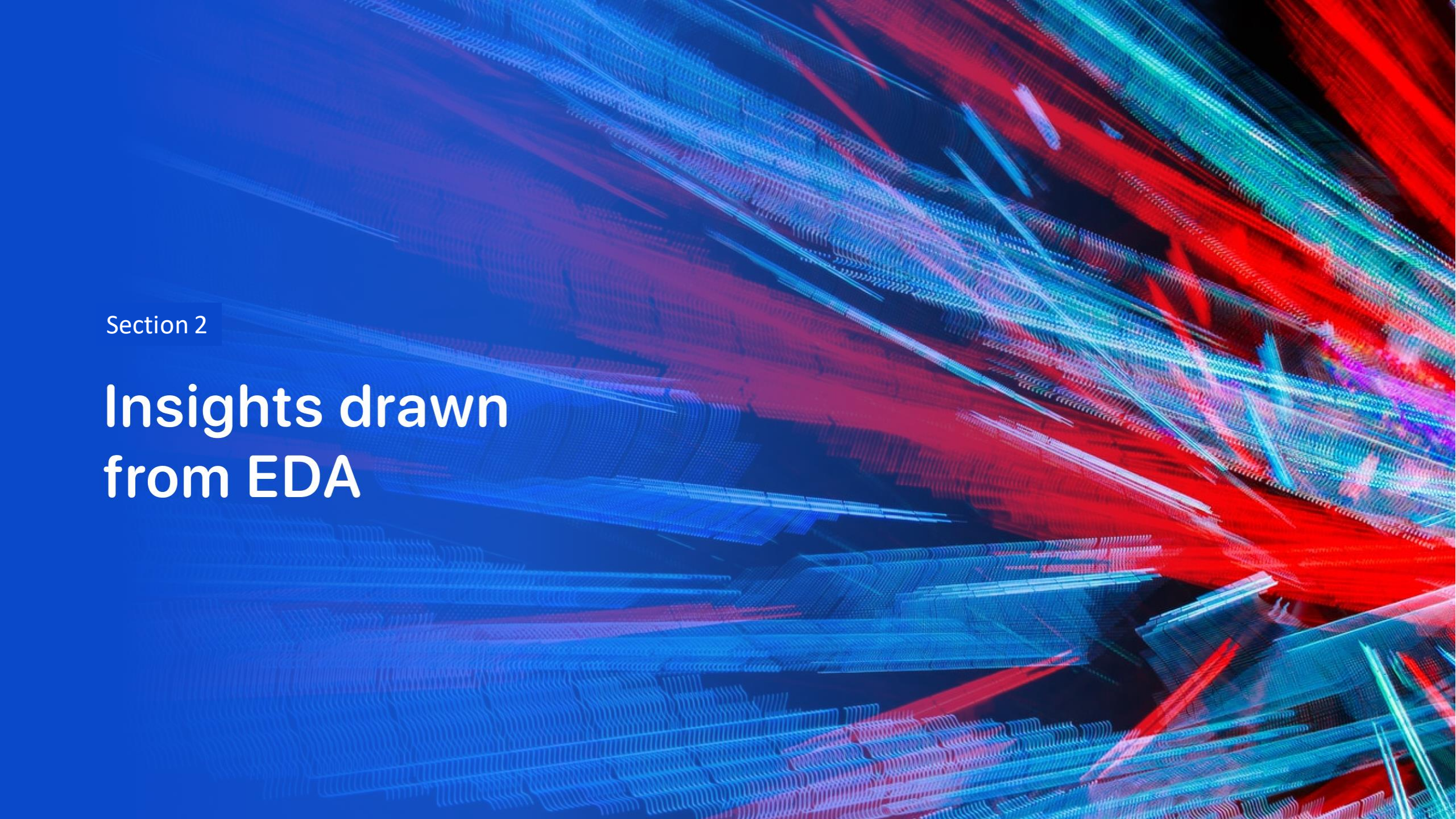
- GitHub URL: Predictive Analysis

Results

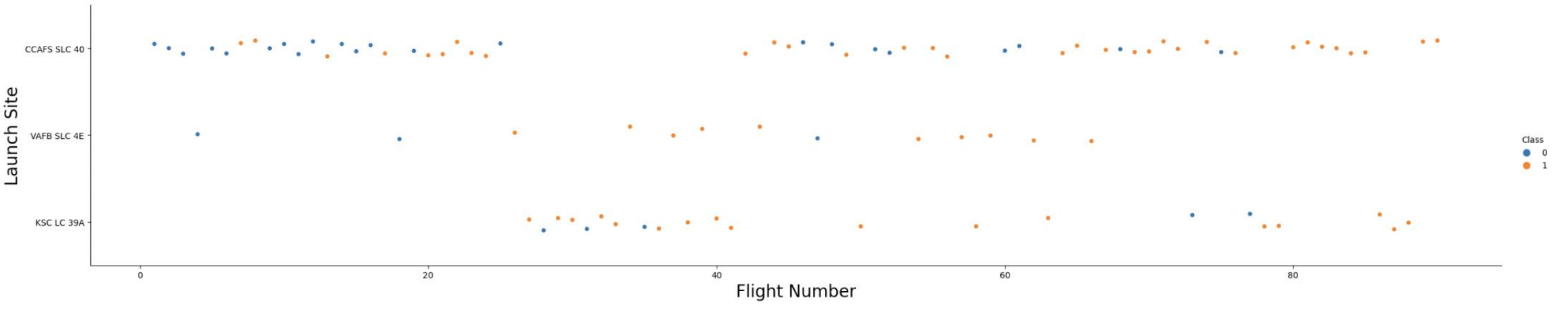
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

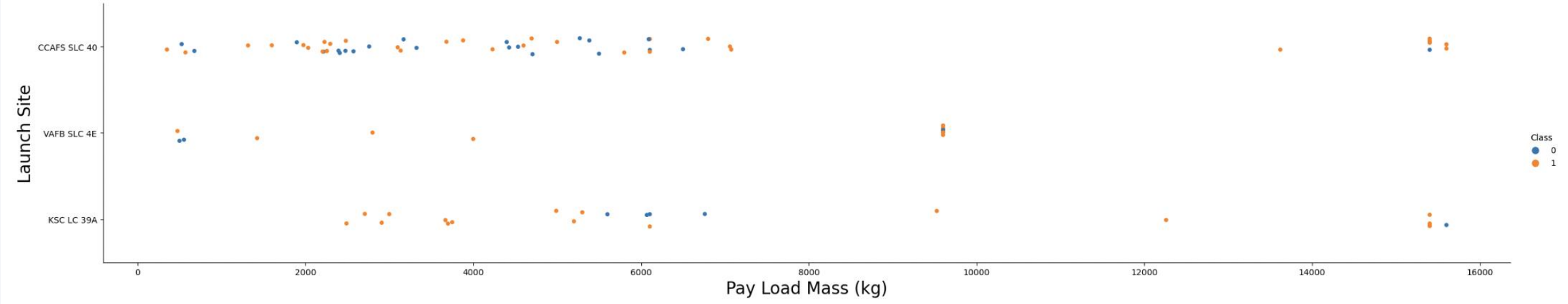


Flight Number vs. Launch Site



- For this scatterplot, orange dots represent a successful launch while a blue dot represents a failure.
- This figure shows that the success rate of launches increased as more flights were made.

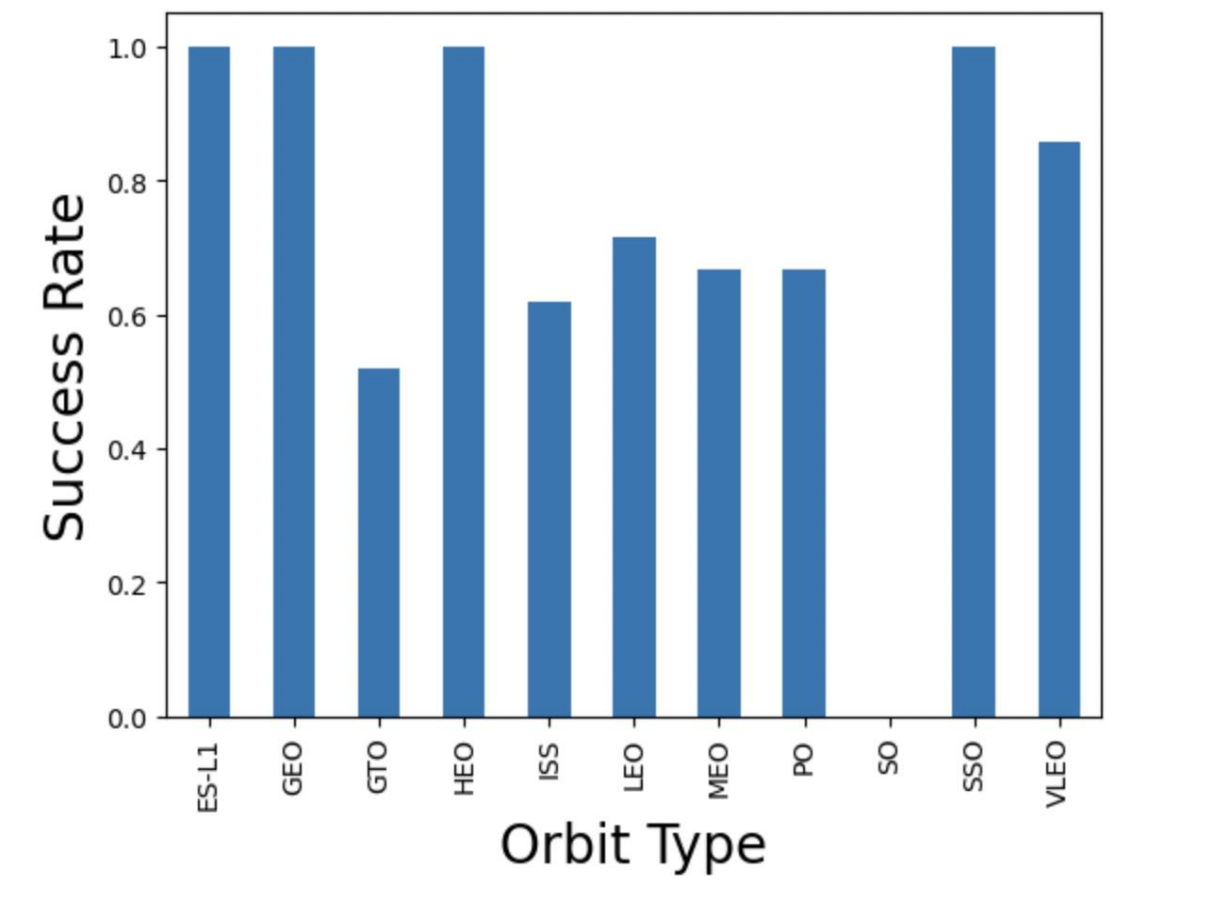
Payload vs. Launch Site



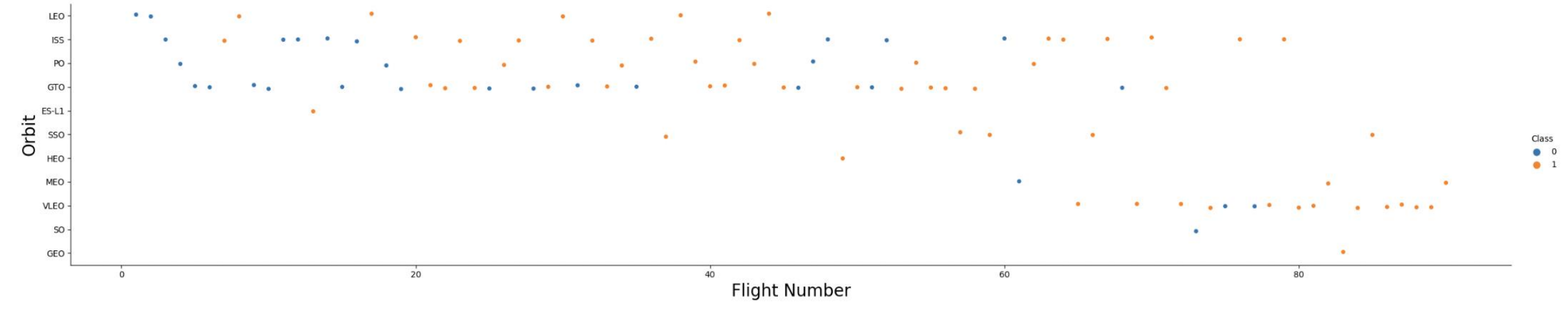
- For this scatterplot, orange dots represent a successful launch while a blue dot represents a failure.
- This figure shows that there doesn't seem to be a strong correlation between payload mass and launch success.

Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO, and SSO launch sites had a 100% success rate.
- The ISS, LEO, MEO, PO, and VLEO launch sites had success rates ranging from 60% and 80%.
- The SO launch site had no successful launches.

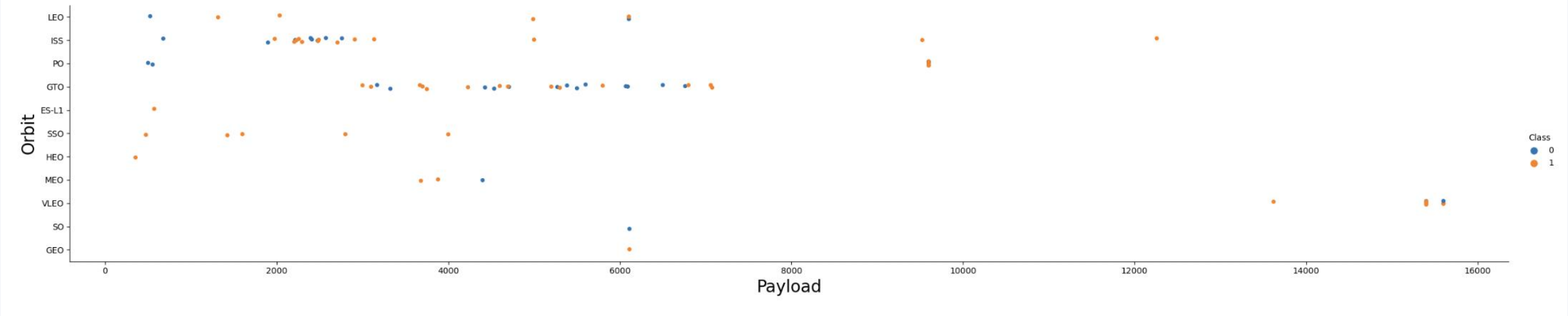


Flight Number vs. Orbit Type



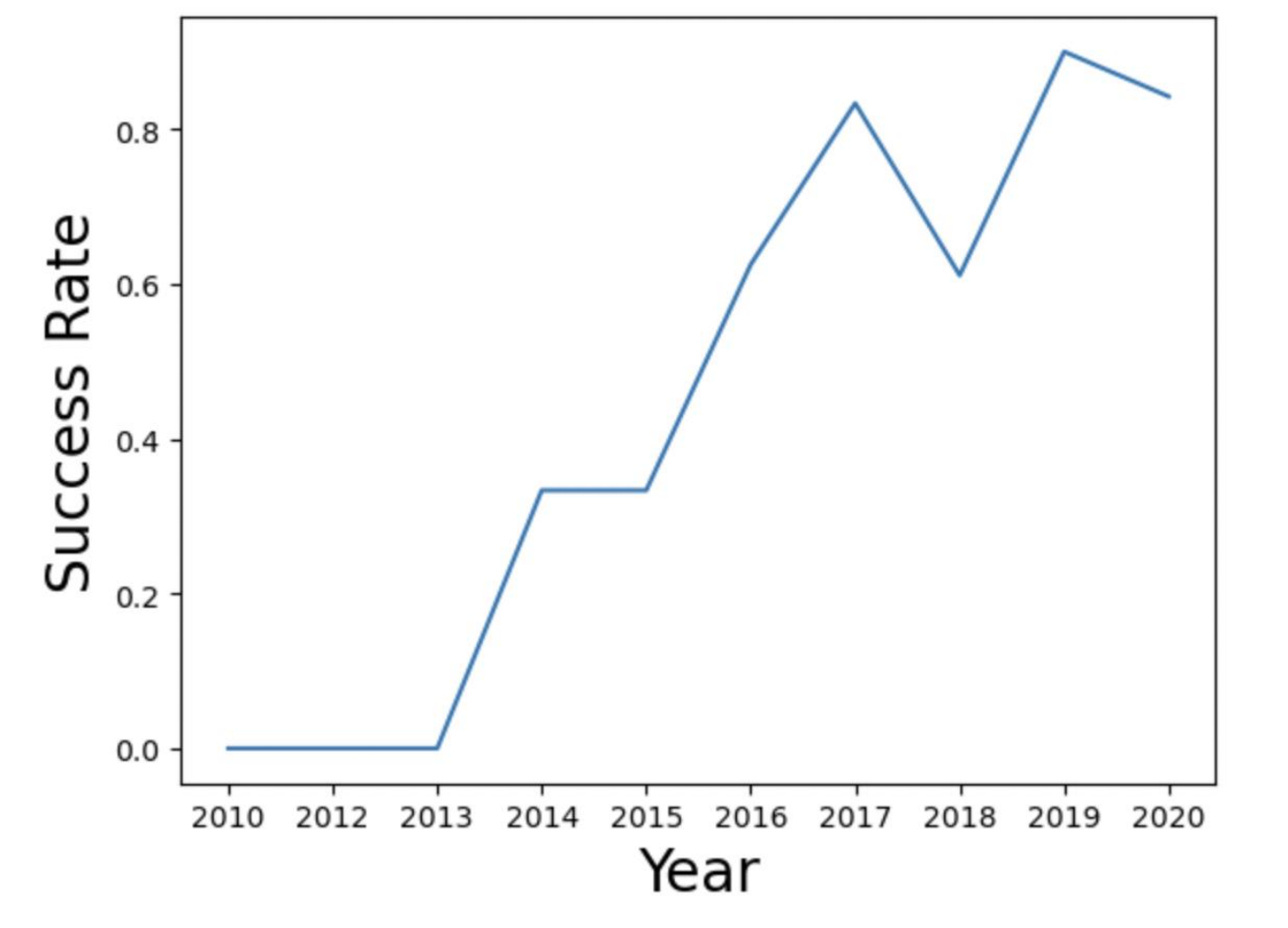
- For this scatterplot, orange dots represent a successful launch while a blue dot represents a failure.
- Flight numbers after the 40 mark seem to have more successful launches than earlier flights showing a strong correlation between flight number and success rate.

Payload vs. Orbit Type



- For this scatterplot, orange dots represent a successful launch while a blue dot represents a failure.
- The success rate in the LEO, ISS, PO, and SSO orbits increases as payload increases.
- This does not seem to be the case for GTO orbits.

Launch Success Yearly Trend



- The success rate of launches increases as time goes by which follows the strong correlation between flight number and success rate of launches in previous figures.
- However, there are drops in 2018 and 2020.

All Launch Site Names

```
%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

- Query to find the names of the unique launch sites.
- The four distinct launch sites are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

- Query to find 5 records where launch sites begin with `CCA`
- A LIMIT clause was used to return only 5 records and a LIKE clause to only include launch sites that started 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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass_nasa_crs_kg from SPACEXTBL where CUSTOMER = 'NASA (CRS)';
```

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

total_payload_mass_nasa_crs_kg
45596

- Query to calculate the total payload carried by boosters from NASA CRS.
- The SUM function was used to calculate the total payload mass for NASA (CRS) as 45596kg. It was then labeled AS total_payload_mass_nasa_crs_kg.

Average Payload Mass by F9 v1.1

- Query to calculate the average payload mass carried by booster version F9 v1.1
- The AVG function was used to calculate the average payload mass for F9 v1.1 as 2534.66.

```
%sql select avg(PAYLOAD_MASS__KG_) as total_payload_mass_f9v1_1_kg from SPACEXTBL where BOOSTER_VERSION like 'F9 v1.1%';
```

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

total_payload_mass_f9v1_1_kg
2534.6666666666665

First Successful Ground Landing Date

- Query to find the date of the first successful landing outcome on ground pad
- The MIN function was used to find the first successful landing outcome on ground pad which was in December 22nd, 2015.
- I had to convert Dates in SPACEXTBL to YYYY-MM-DD in my sqlite database before running this query.

```
%%sql select min(DATE) as first_successful_landing_outcome from SPACEXTBL
where "LANDING _OUTCOME" = 'Success (ground pad)';
```

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

first_successful_landing_outcome
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select BOOSTER_VERSION from SPACEXTBL where "LANDING _OUTCOME" = 'Success (drone ship)' and 4000 < PAYLOAD_MASS__KG_ < 6000;
```

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

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

- Query to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- The less than and greater than symbols were useful for to find a payload mass between 4000 and 6000.
- Many Booster version were listed to fit this description.

Total Number of Successful and Failure Mission Outcomes

- Query to calculate the total number of successful and failure mission outcomes
- Out of the 101 mission outcomes, only one failed when in flight. All others ended in a success one way or another.

```
%sql select MISSION_OUTCOME, count(*) as 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

```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

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

- Query to list the names of the boosters which have carried the maximum payload mass
- A subquery was used to compare the various PAYLOAD_MASS_KG_ to the max value in the column.
- 12 different booster versions were shown

2015 Launch Records

- Query to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- A substr function was used to extract the year from the Data variable.
- Two Booster versions returned: B1012 and B1015

```
%%sql select "LANDING _OUTCOME", BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL
where "LANDING _OUTCOME" = 'Failure (drone ship)' and substr(Date, 1, 4) = '2015';
```

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

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

- Query to 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
- Order by and desc were used to rank them sarting with the highest count to the lowest.
- 10 were labeled as 'no attempt' while 8 were labeled a success

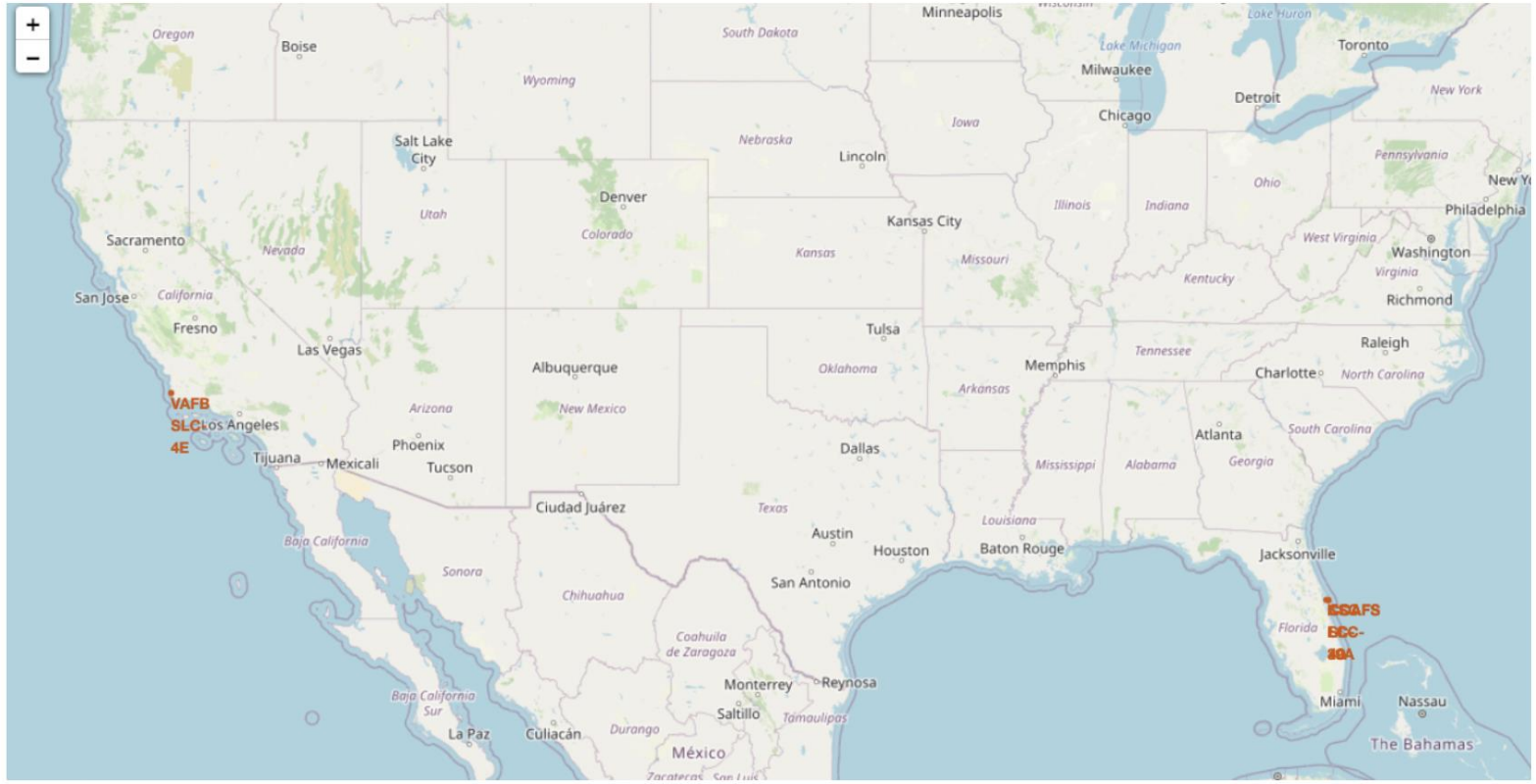
```
%%sql select "LANDING _OUTCOME", count(*) as count from SPACEXTBL
where DATE between '2010-06-04' and '2017-03-20' group by "LANDING _OUTCOME" order by count desc;
```

```
 * 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

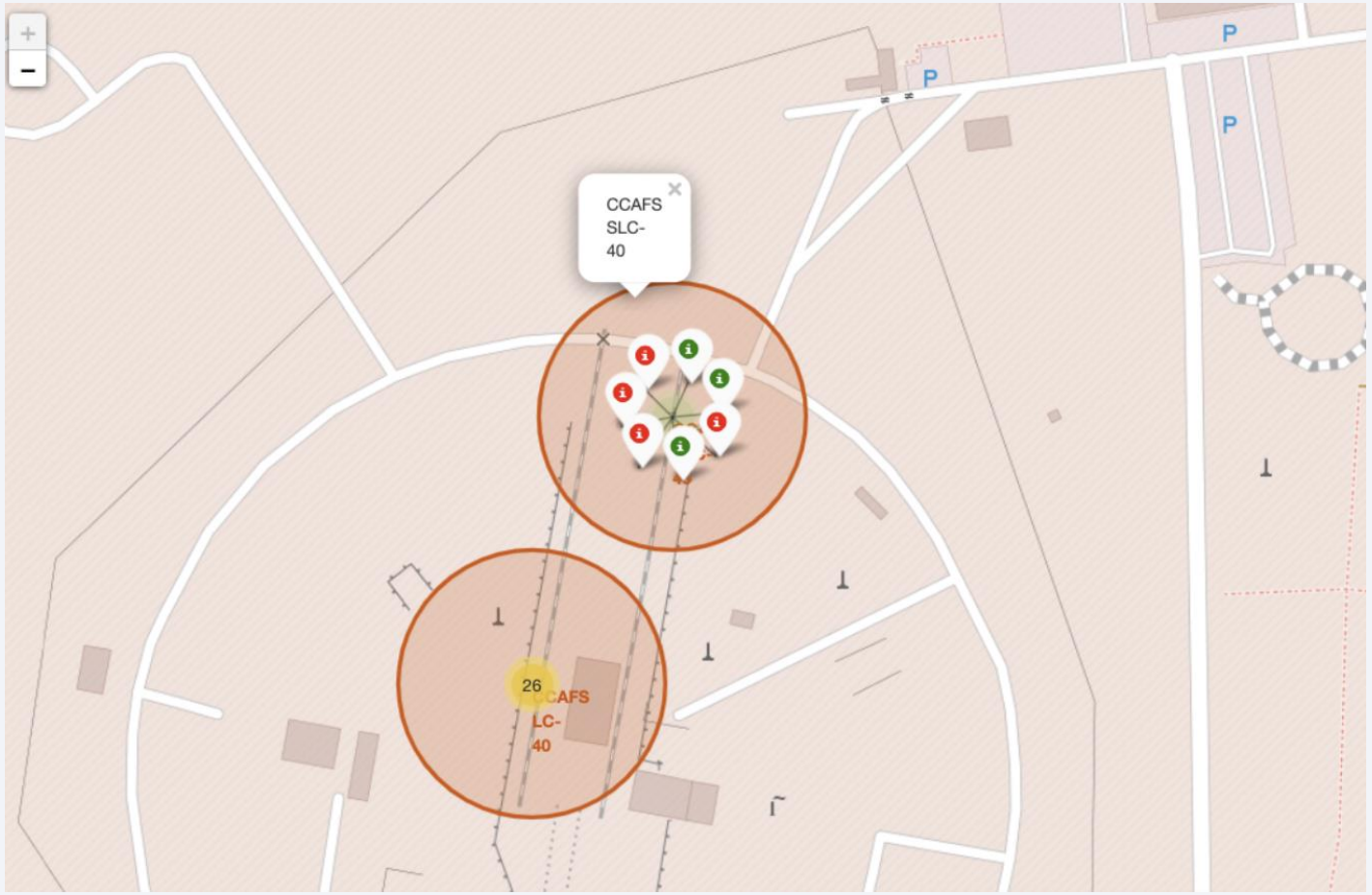

Section 3
Launch Sites
Proximities Analysis

Folium Map with Launch Site Locations



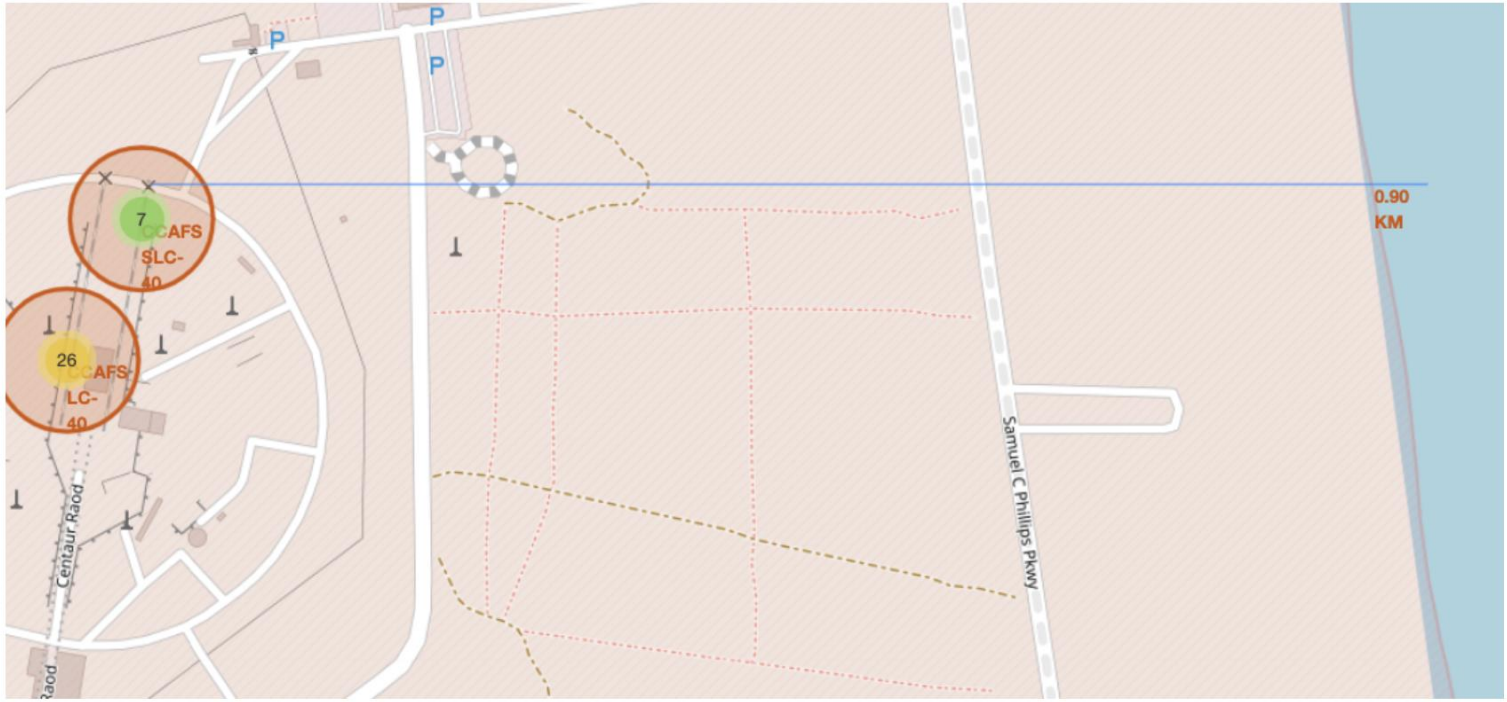
- This folium map reveals where the 4 main launch locations are located.
- Three launch sites are in Florida and another in South California.
- The two in Florida are CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A.
- The one in California is VAFB SLC 4E.

Folium Map with Launch Outcomes

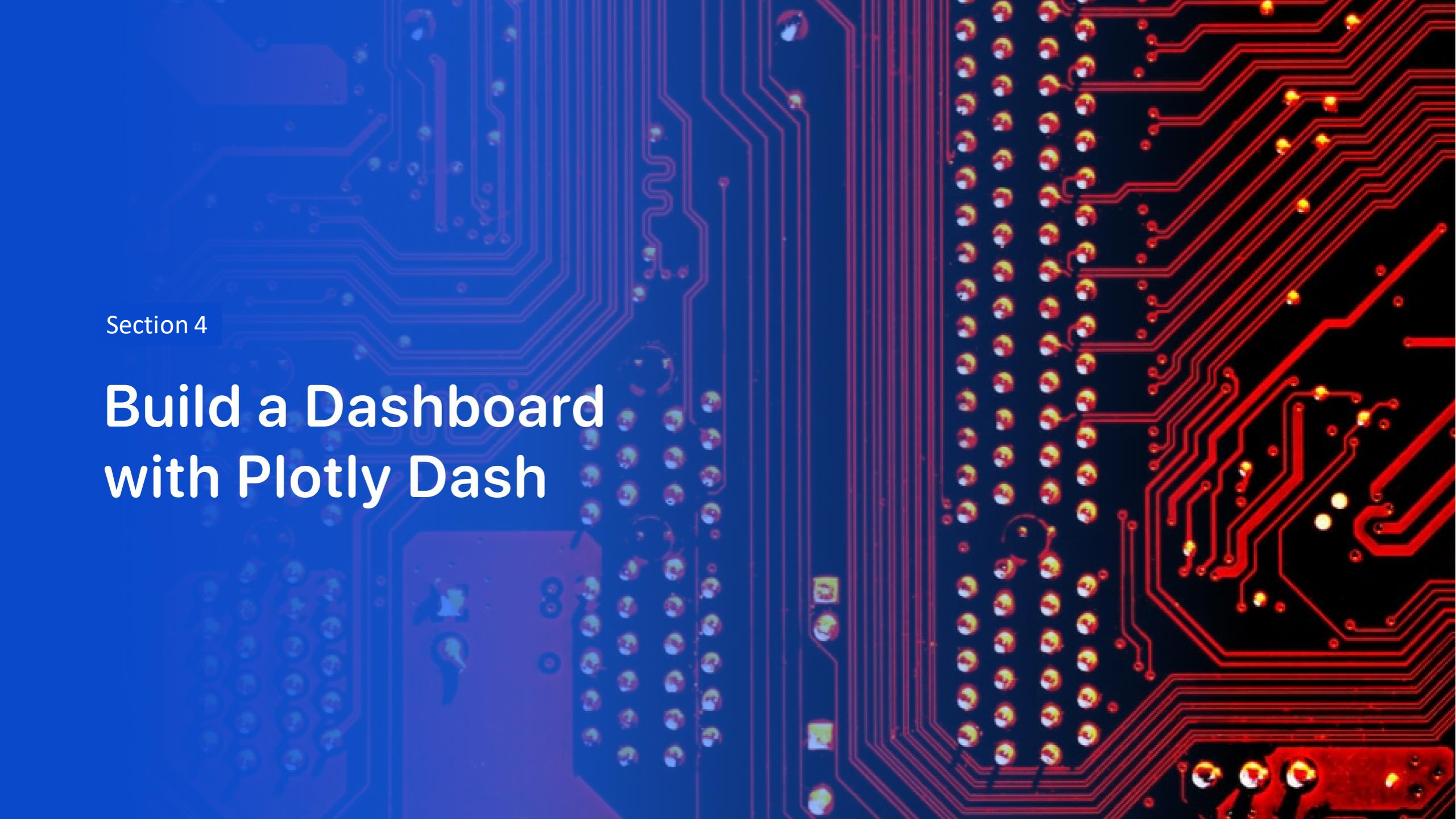


- This folium map reveals the landing outcomes by launch location.
- This particular image shows the launches that failed and succeeded at the CCAFS SLC-40.
- At this location, there were three successful launches and 4 that ended in failure.

Folium Map with Distance from Proximities

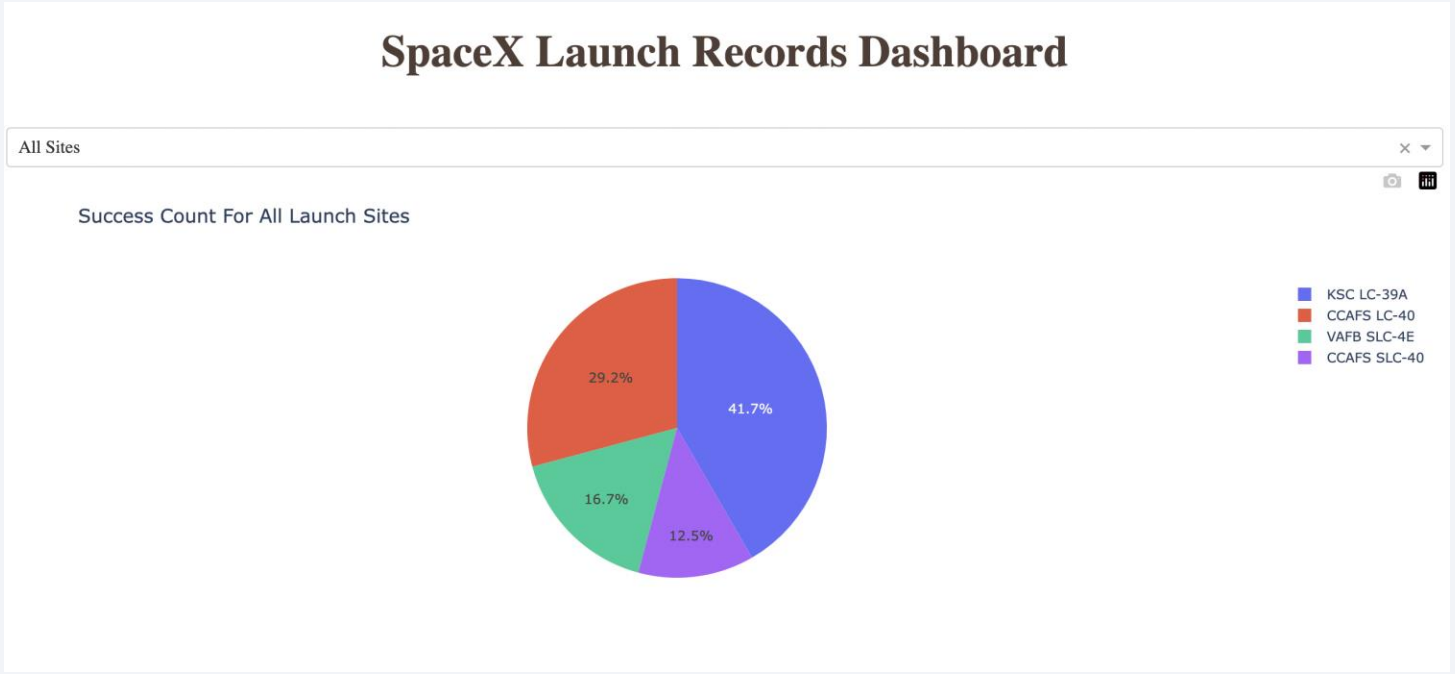


- This folium map reveals the distance between a launch sites and it's proximities.
- This particular image shows the distance of CCAFS SLC-40 from the coast which is about .9km away.
- There are launch sites relatively close to railways, highways, coastlines, and cities. For example, the CCAFS SLC-40 and CCAFL LC-40 launch locations are about less than 2 KM away from the NASA Railway, 7 KM from the Samuel C Phillips Parkway, less than 1 KM from the Atlantic Ocean coastline, and 50 KM away from Melbourne (which is not too far away). The same can be said about VAFB SLC-4E by looking at the map. It is near both Lompoc and Santa Maria. It is also located right next to a railway and the coast. There are no major highways nearby, however.

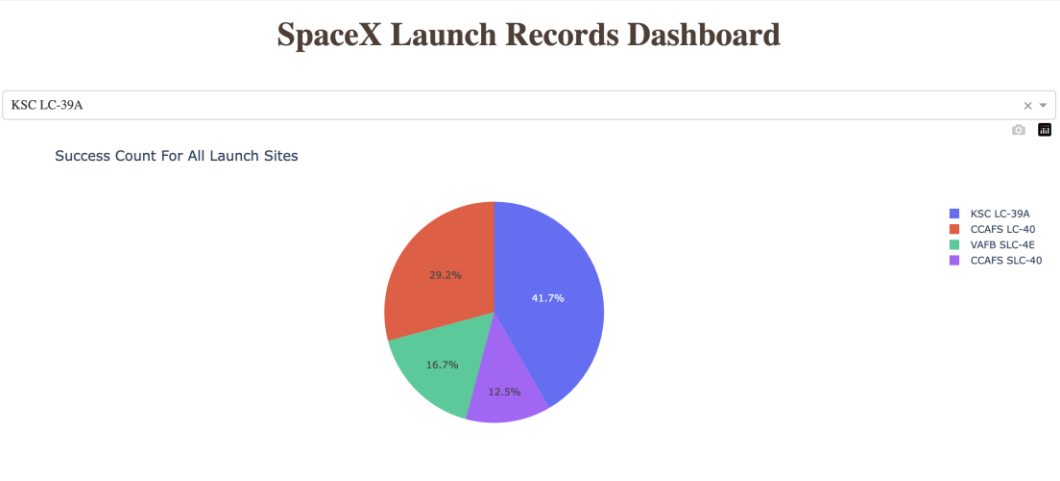

Section 4
Build a Dashboard with Plotly Dash

Success Count For All Launch Sites

- This pie chart shows the percentage of the launch success count for all sites.
- 41.7% of the launches were at the KSC LC-39A launch site, 29.2% of the launches were at the CCAFS LC-40 launch site, 16.7% of the launches were at VAFB SLC-4E the launch site, and 12.5% of the launches were at the CCAFS sLC-40 launch site.



Highest Launch Success Ratio



- Pie chart of the launch site with highest launch success ratio
- KSC LC-39A has the most successful launch success ratio.

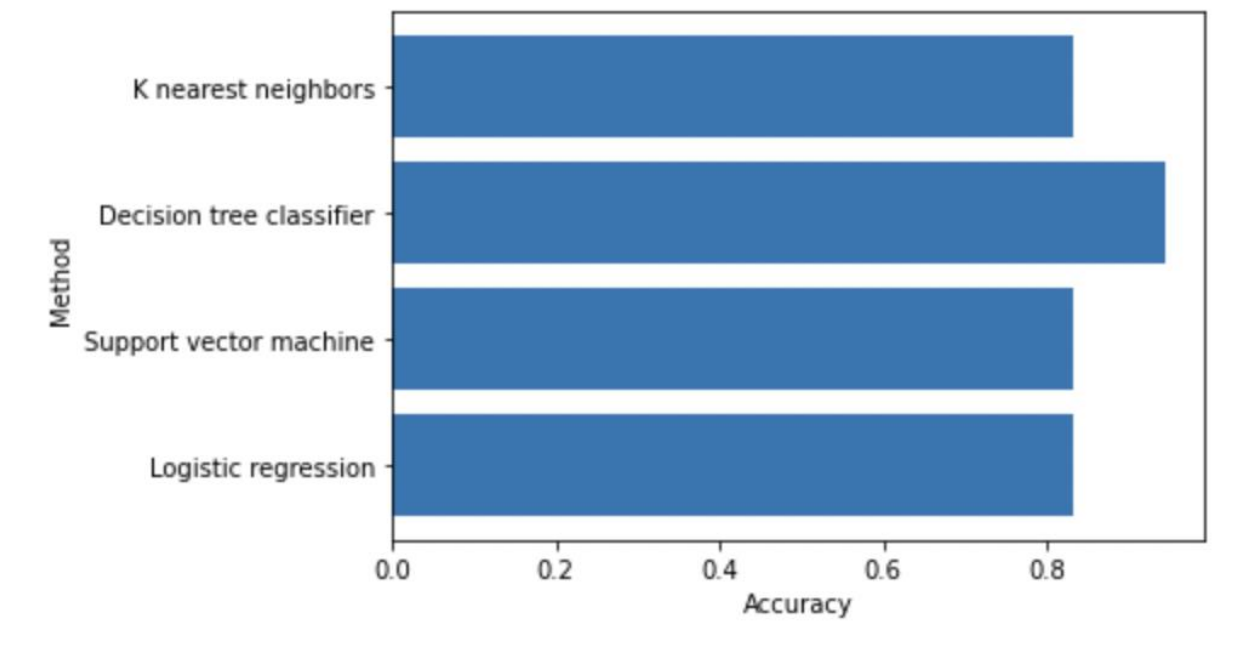
Payload vs Launch Outcome Scatter Plot

- These scatter plots show the Payload and Launch Outcome for all sites, with different payload selected in the range slider. Class shows a success with a value of 1 and a failure with a value of 0.
- All launches failed with booster version 1.0 and most failed with version 1.1. FT, B4, and B5 performed much better.
- The bigger the payload mass gets, the more likely the launch is to end in a failure.

Section 5

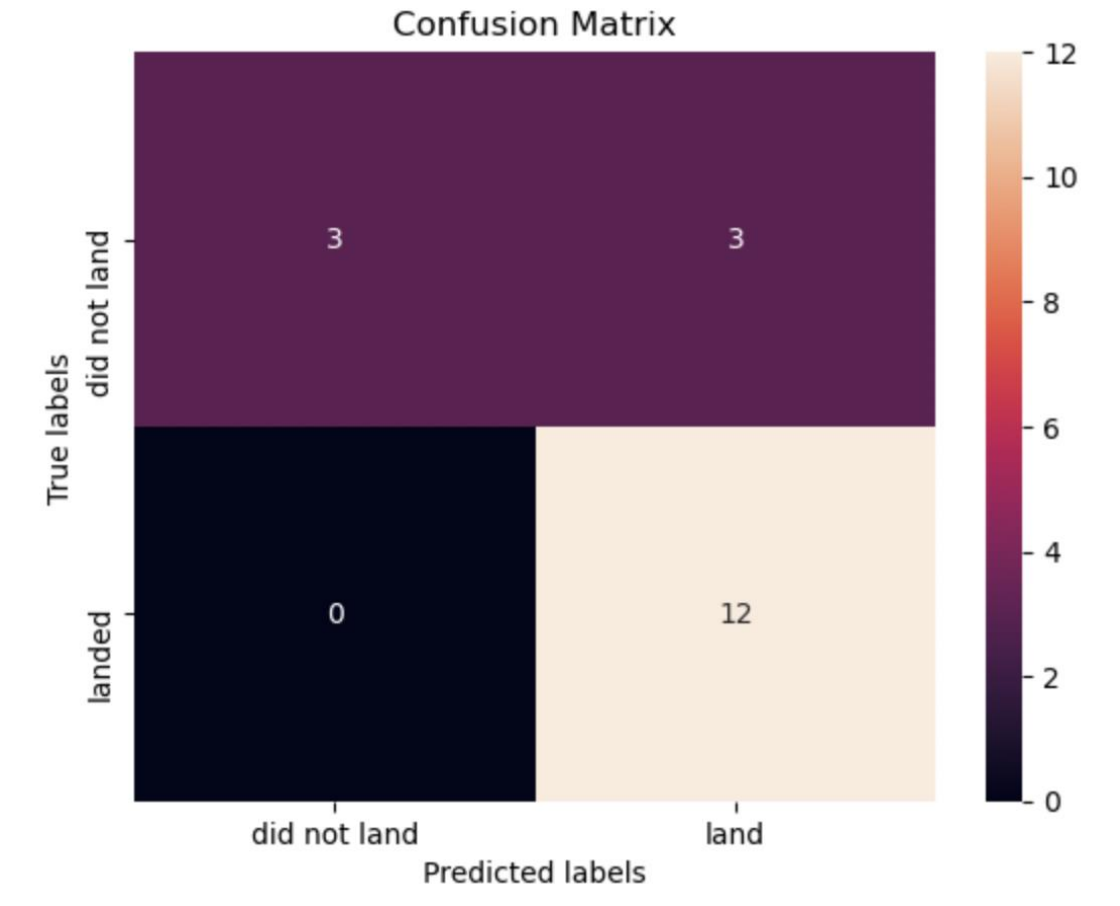
Predictive Analysis (Classification)

Classification Accuracy



- This bar chart visualizes the built model accuracy for all built classification models.
- The decision tree classifier had the highest classification accuracy with a 94%.

Confusion Matrix



- This image shows the confusion matrix for the decision tree classifier model.

- 12 predicted lands actually landed. 3 predicted lands did not land. 3 predicted failures did not land.

Conclusions

- Our exploratory data analysis revealed that the success rate of missions went up over time.
- Launch sites tend to be placed near major proximities including railways, highways, coastlines, and cities.
- There is a strong correlation between payload and launch outcome with mission failure more likely to occur the heavier the payload is.
- Our predictive analysis reveals that the Decision Tree Classifier Algorithm was the most accurate classification method.

Appendix

Data Sets

Thank you!