IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jing Yu
01/25/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection and wrangling methodology
  - EDA and interactive visual analytics methodology
  - Predictive analysis methodology
- Summary of all results
  - EDA with visualization results
  - EDA with SQL results
  - Interactive map with Folium
  - Plotly Dash
  - Predictive analysis results

Introduction

- Project background and context
  - The commercial space age is here, companies are making space travel affordable for everyone. The most successful is SpaceX as its rocket launches are relatively inexpensive. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage.
  - As data scientists, we are going to analyze the Falcon 9 reuse of SpaceX and evaluate if a new rocket company, Space Y, could bid against SpaceX and compete.

- Problems you want to find answers
  - What are the key features to determine the price of each launch?
  - What influence the success rate of the Falcon 9 launch?
  - Where is the best launch site to achieve the largest success?
  - Could we predict the launch success by machine learning method?

Methodology

Executive Summary

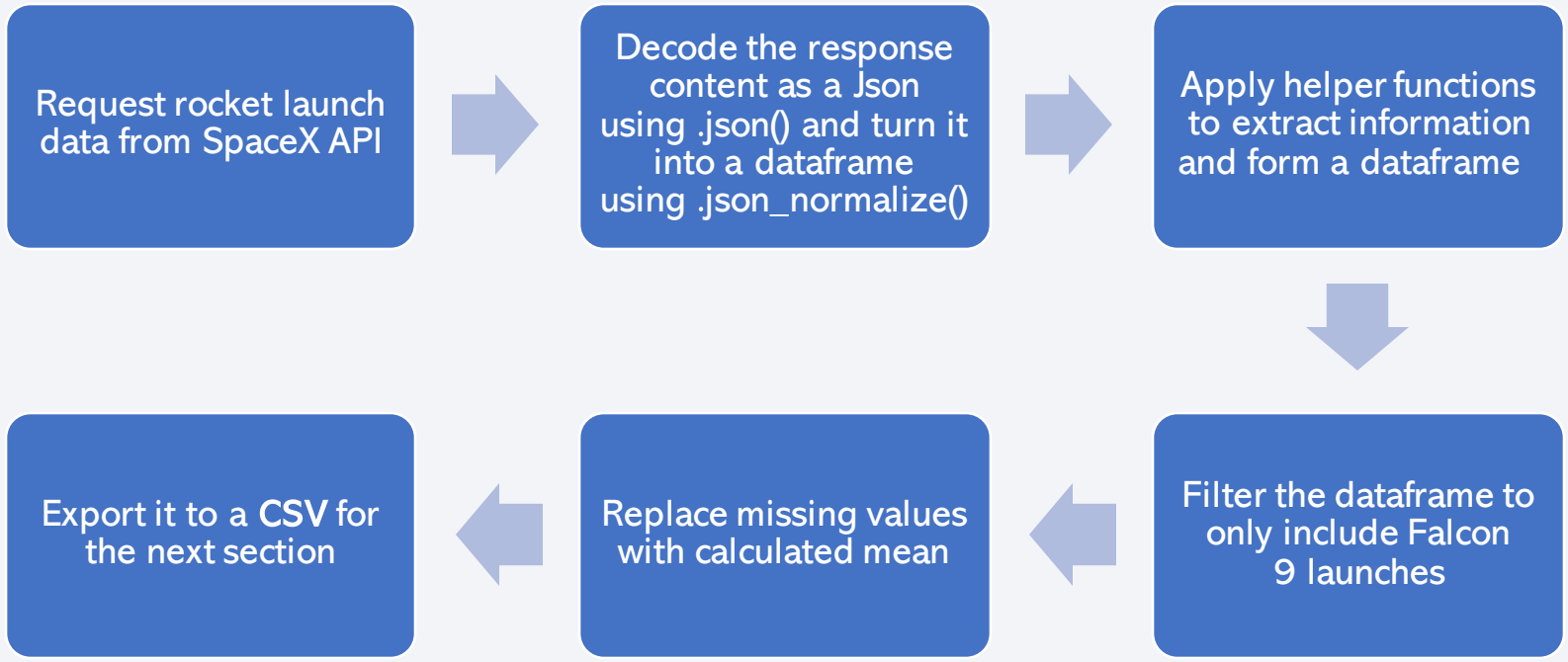
- Data collection methodology:
  - Request from SpaceX API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Filter data to only include Falcon 9 launches
  - Cleanse data to deal with the missing values
  - Determine training labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Transform data set and split into training and testing data
  - Train different models to find the best parameters and best score

Section 1

Methodology

Data Collection – SpaceX API

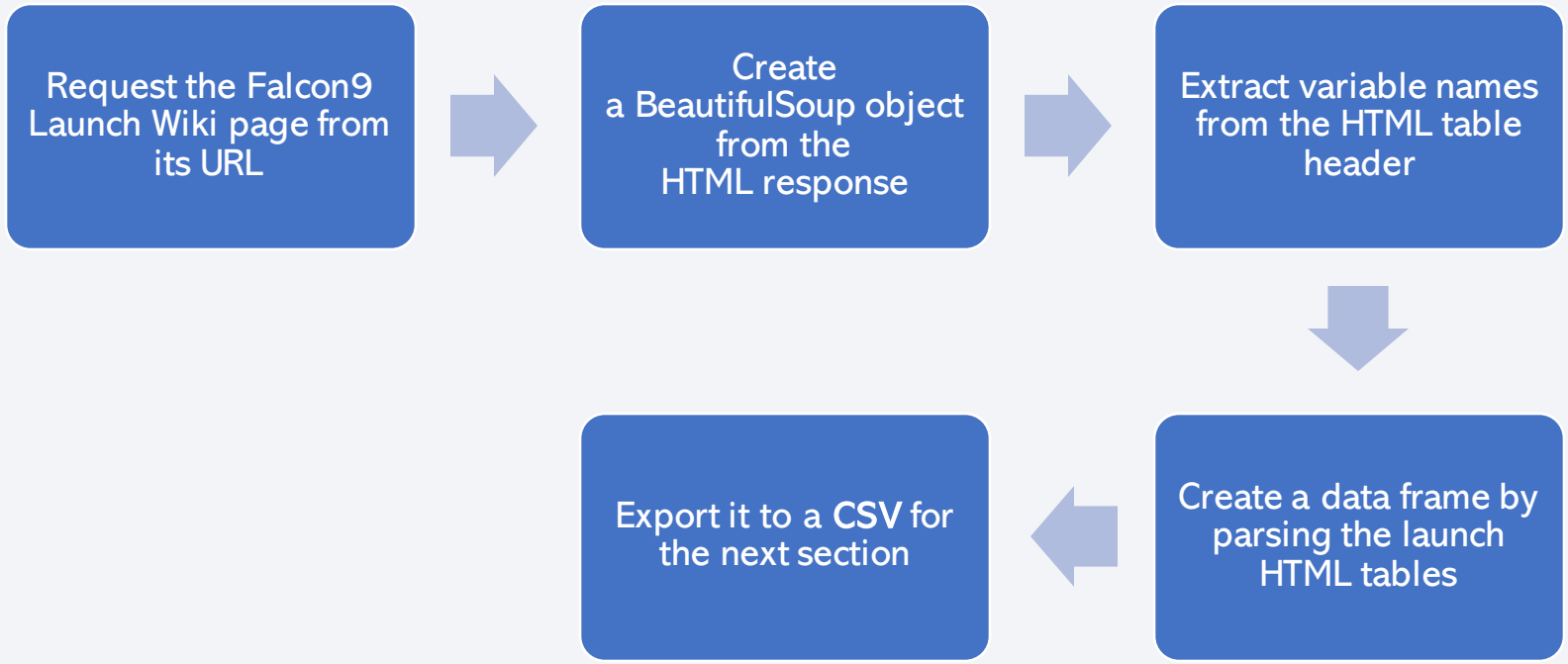
- Data collected from SpaceX API (https://api.spacexdata.com/v4/launches/past)
- Request and parse the SpaceX launch data from API_call_spacex_api.json



Applied-Data-Science-Capstone/Data Collection with API.ipynb at master · rowlland/Applied-Data-Science-Capstone · GitHub

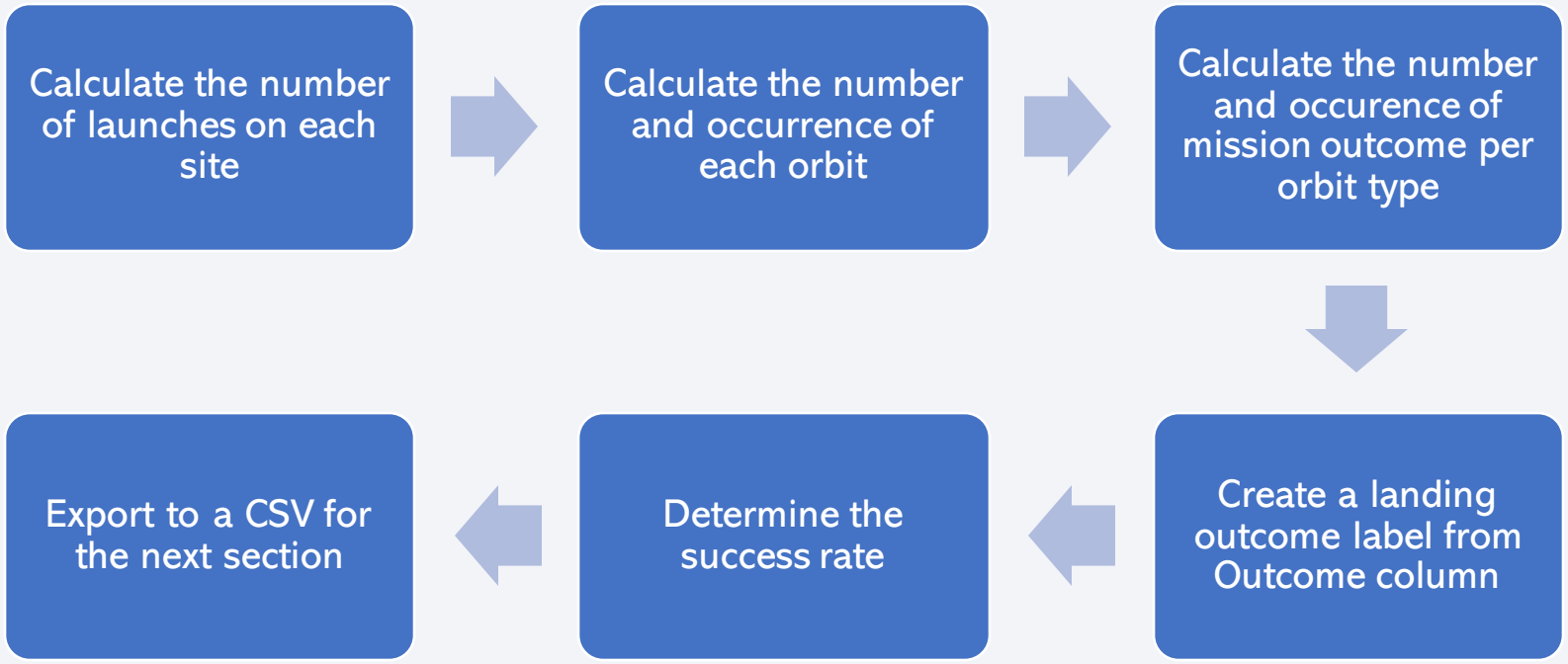
Data Collection - Scraping

- Perform web scraping to collect Falcon 9 historical launch records from Wikipedia
- (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)



Data Wrangling

- Perform exploratory data analysis for the data collected from API
- Determine training Labels



Applied-Data-Science-Capstone/Data Wrangling.ipynb at master · rowlland/Applied-Data-Science-Capstone · GitHub

EDA with Data Visualization

- Use scatter chart to plot out:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Flight Number vs. Orbit type
  - Payload vs. Orbit type
- Create a bar chart to visualize:
  - Success rate of each orbit type
- Plot a line chart to visualize:
  - Launch success yearly trend

Applied-Data-Science-Capstone/EDA with Visualization.ipynb at master · rowlland/Applied-Data-Science-Capstone · GitHub

EDA with SQL

- Execute SQL queries to:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters success in drone ship and have payload mass 4000 to 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names in 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 descending

Applied-Data-Science-Capstone/EDA with SQL.ipynb at master · rowlland/Applied-Data-Science-Capstone · GitHub

Build an Interactive Map with Folium

- Perform more interactive visual analytics using Folium:
  - Create a folium Map object
  - Use folium.Circle and folium.Marker to add each launch site on the site map
  - Use MarkerCluster to cluster the success(green)/failed(red) launches for each site on the map
  - Mark down proximity points using MousePosition and calculate the distance between the points and the launch site
  - Draw a PolyLine between a launch site to the selected proximity point

Applied-Data-Science-Capstone/Interactive Visual Analytics with Folium.ipynb at master · rowlland/Applied-Data-Science-Capstone · GitHub

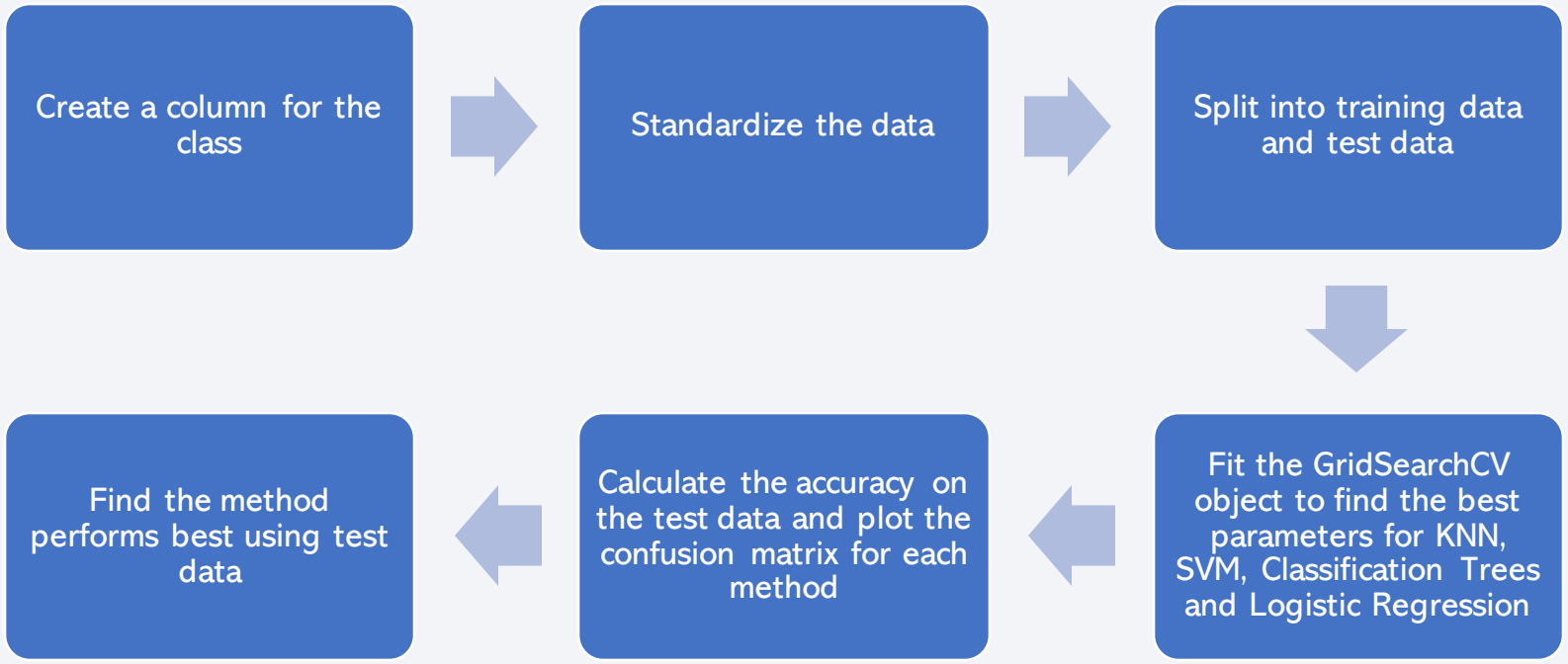
Build a Dashboard with Plotly Dash

- Build a Plotly Dash application to perform interactive visual analytics on SpaceX launch data in real-time
  - Create a dropdown menu to let us select different launch sites
  - Render a pie chart based on selected site to visualize launch success rate
  - Add a Range Slider to easily select different payload range
  - Plot a scatter chart to visually observe how payload may be correlated with mission outcomes for selected site(s)

Applied-Data-Science-Capstone/spacex_dash_app.py at master · rowlland/Applied-Data-Science-Capstone · GitHub

Predictive Analysis (Classification)

- Create a machine learning pipeline to predict if the first stage will land



Applied-Data-Science-Capstone/Machine Learning Prediction.ipynb at master · rowlland/Applied-Data-Science-Capstone · GitHub

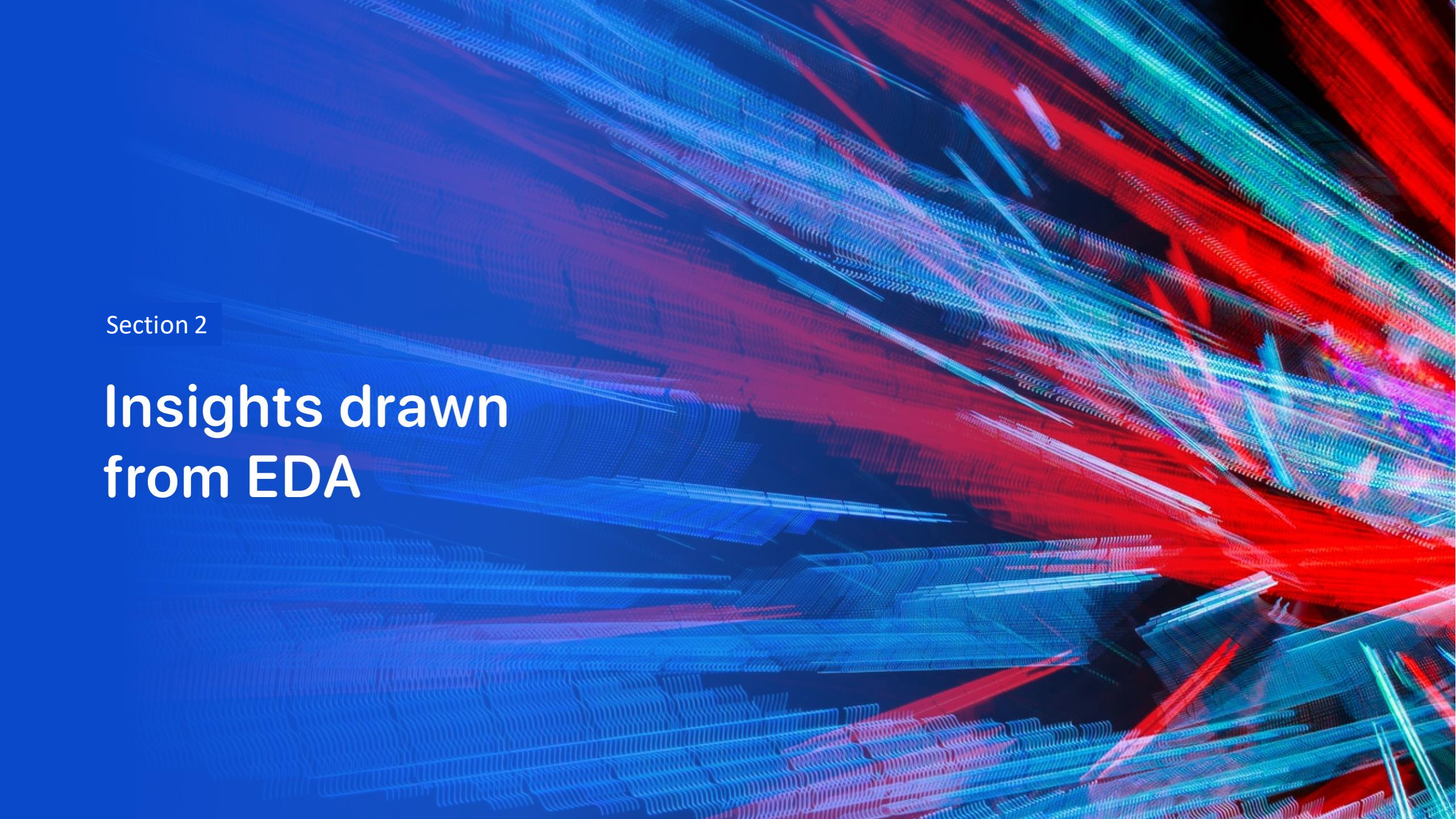
Results

Executive Summary

- Insights drawn from EDA
- Launch Sites Proximity Analysis
- Build a Dashboard with Plotly Dash
- Predictive Analysis (classification)

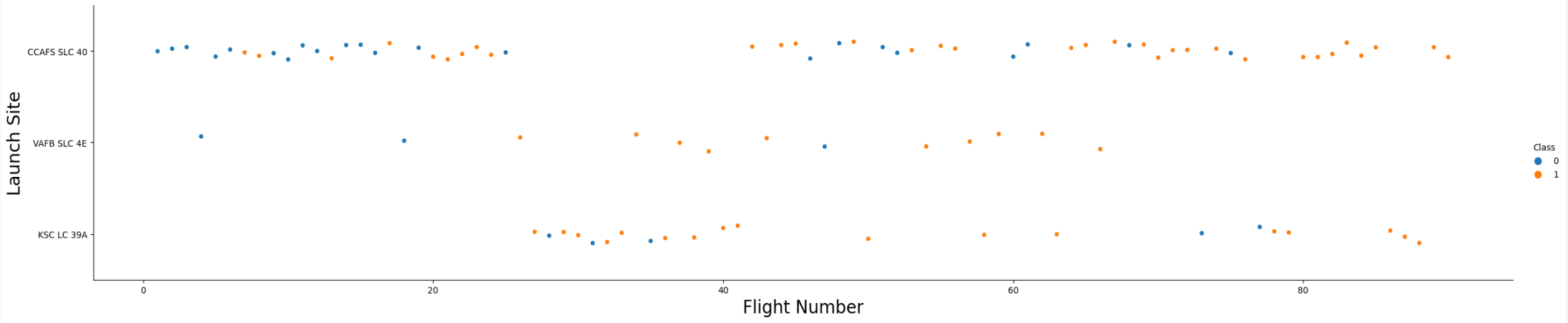
Section 2

Insights drawn from EDA



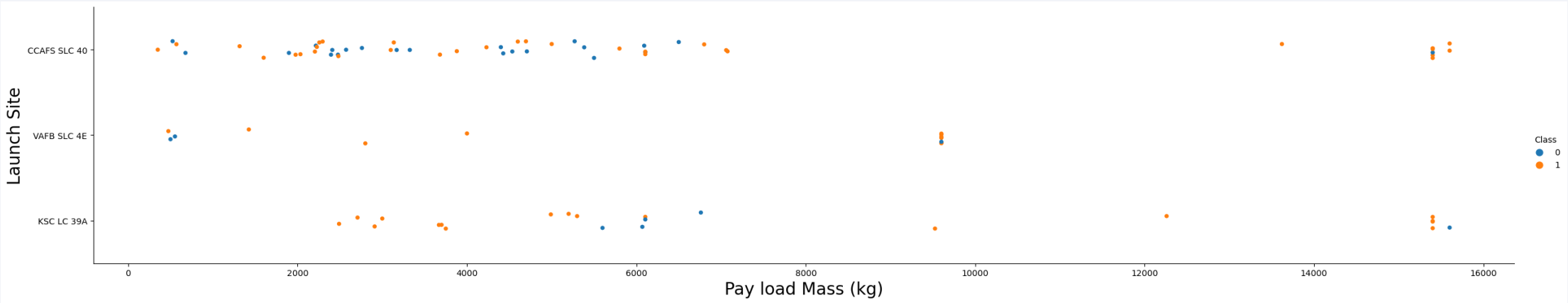
Flight Number vs. Launch Site

- Explanations:
  - Most of the rockets were launched at CCAFS SLC 40. However, its success rate is lower.
  - As the flight number increases, the first stage is more likely to land successfully.



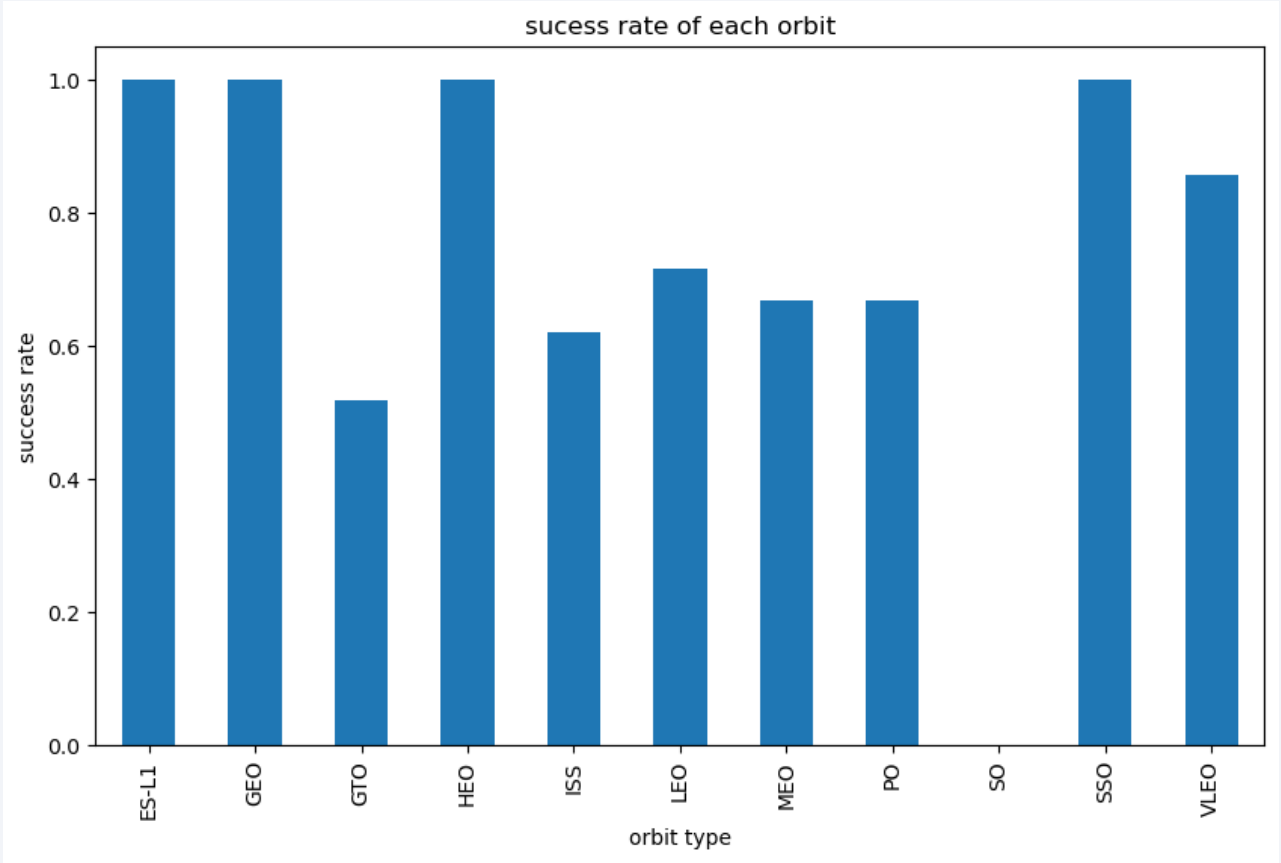
Payload vs. Launch Site

- Explanations:
  - Most of the rockets were launched with payload mass under 8,000 kg.
  - No rockets launched for heavy payload mass(greater than 10,000) at VAFB-SLC.



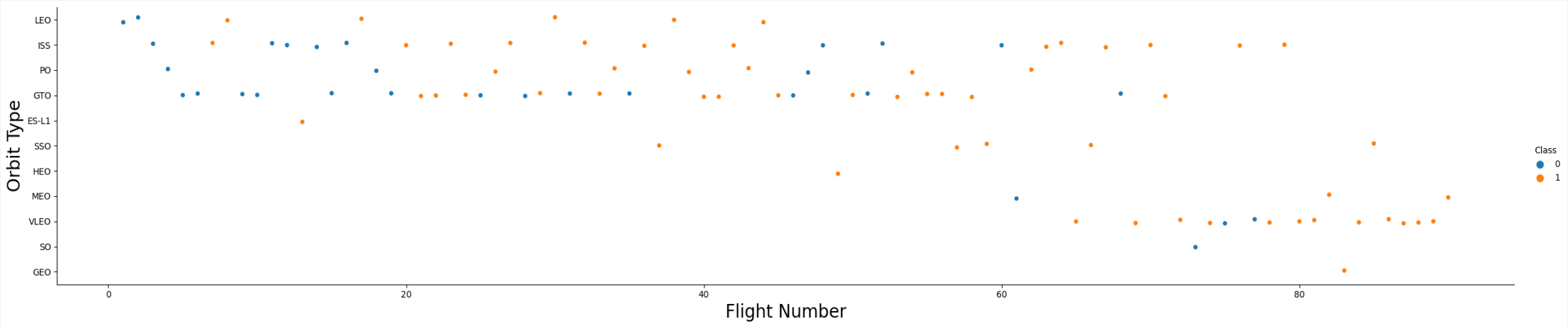
Success Rate vs. Orbit Type

- Explanations:
  - In the ES-L1, GEO, HEO, SSO orbits, the success rates are 100%.
  - The SO orbit has the lowest success rate, 0%.



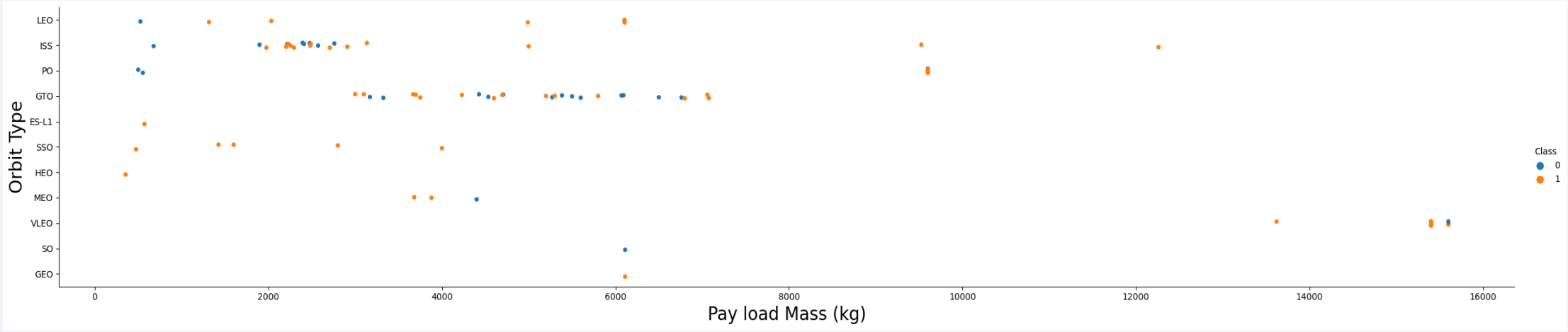
Flight Number vs. Orbit Type

- Explanations:
  - In the LEO orbit the success appears related to the number of flights
  - There seems to be no relationship between flight number when in GTO orbit.



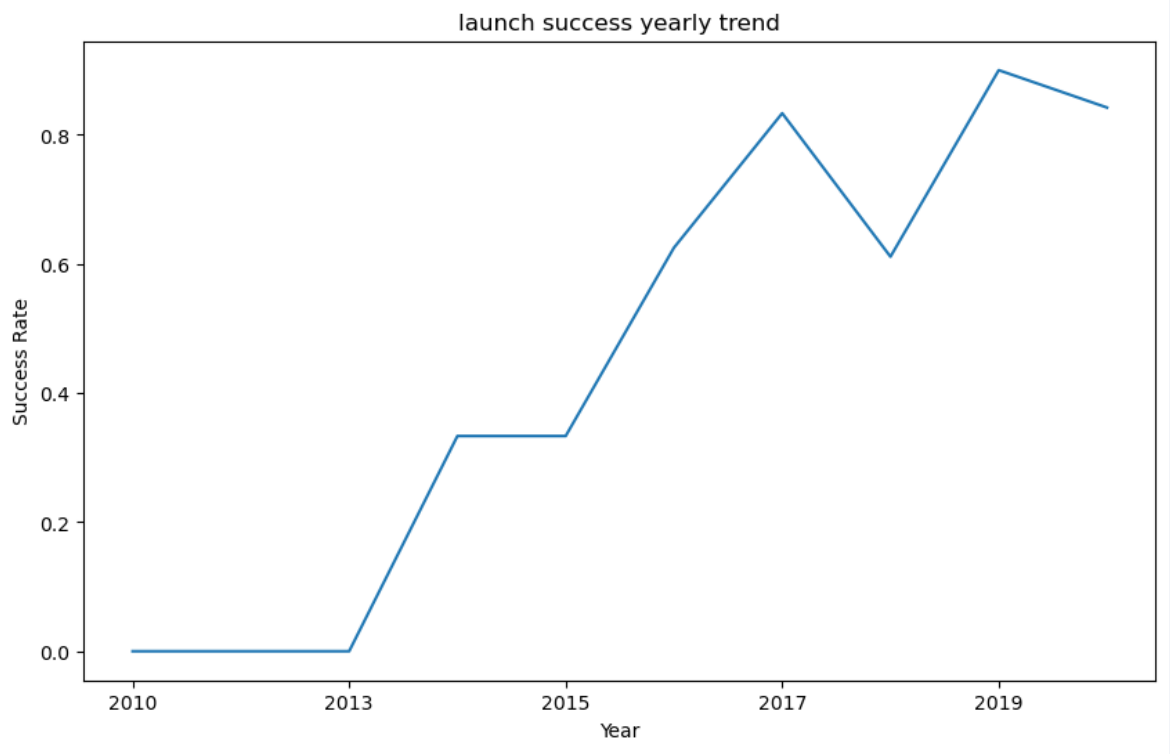
Payload vs. Orbit Type

- Explanations:
  - With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
  - For GTO we cannot distinguish this relationship well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- Explanations:
  - The success rate since 2013 kept increasing till 2020.



All Launch Site Names

- The names of the unique launch sites

```
%sql select distinct LAUNCH_SITE from spacex;
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%sql select * from spacex where LAUNCH_SITE like 'CCA%' limit 5;
```

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

- The total payload carried by boosters from NASA (CRS): 48,213 KG

```
%sql select sum(payload_mass__kg_) as total from spacex where customer like '%NASA%CRS%';
```

total	customer
45596	NASA (CRS)
2617	NASA (CRS), Kacific 1

total
48213

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1: 2,928 KG

```
%sql select avg(payload_mass__kg_) as avg_payload from spacex where booster_version = 'F9 v1.1';
```

avg_payload
2928

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad:
  - 2015-12-22

```
%sql select min(date) as Date from spacex where LANDING__OUTCOME = 'Success (ground pad)';
```

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select BOOSTER_VERSION from spacex where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - Success 100
  - Failure 1

```
%sql select MISSION_OUTCOME, count(*) as Total from spacex group by MISSION_OUTCOME;
```

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

Boosters Carried Maximum Payload

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

```
%sql select booster_version from spacex
     where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex);
```

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

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

```
%sql select booster_version, launch_site, landing__outcome from spacex
      where landing__outcome = 'Failure (drone ship)' and year(date) = 2015;
```

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

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

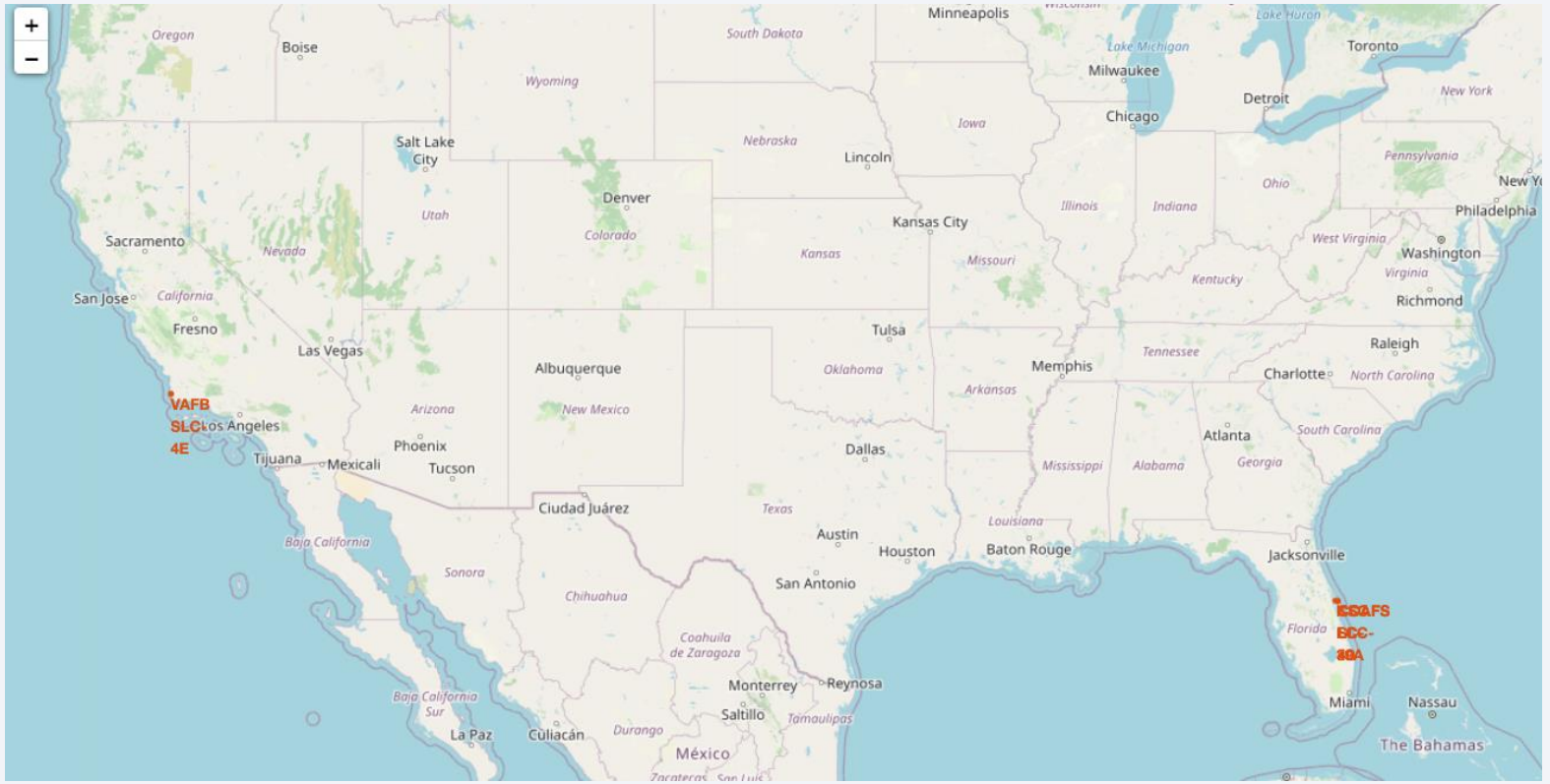
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

```
%sql select landing__outcome, count(*) as total from spacex
     where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by total
```


Section 3
Launch Sites
Proximities Analysis

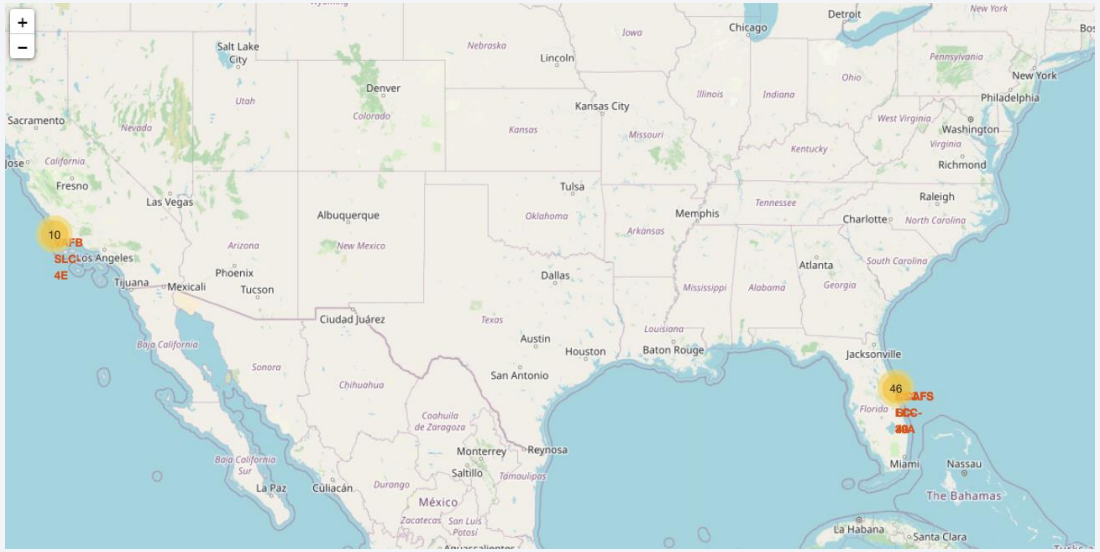
All Launch Sites on a Map

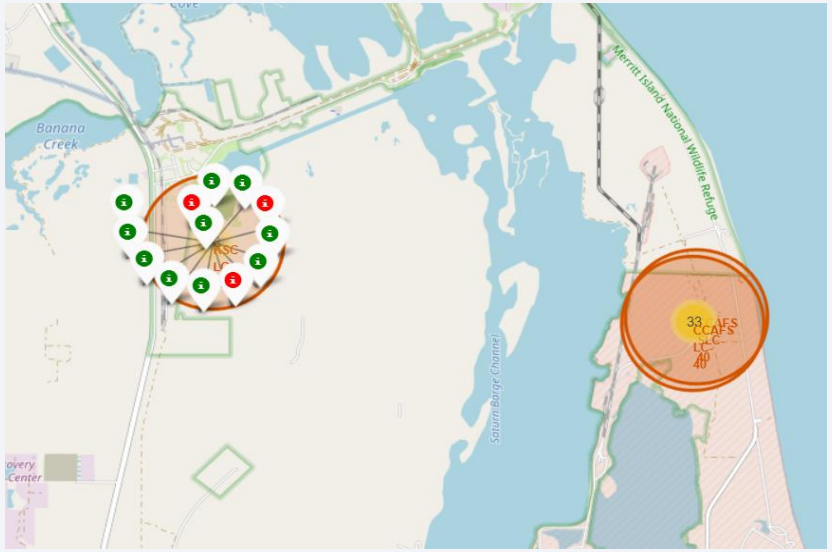
- All launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast.



Success/Failed Launches for Each Site on the Map

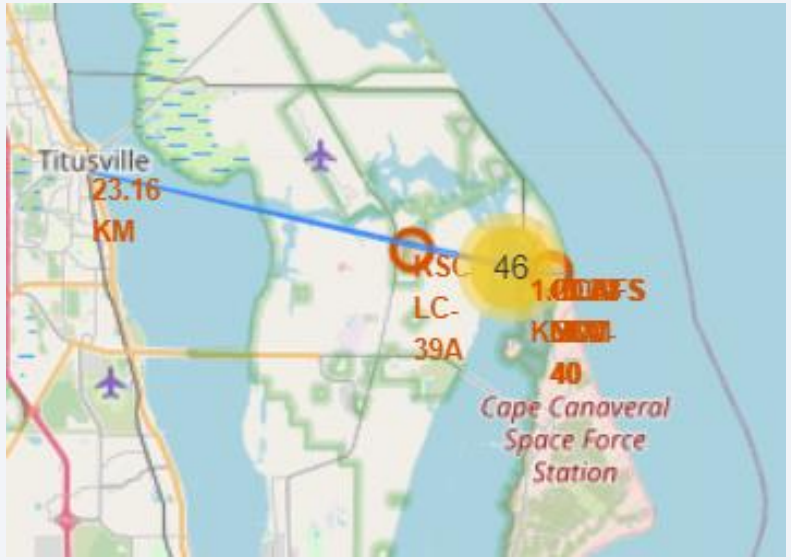
- The color-labeled markers in marker clusters: success as green, failure as red.
- We could easily identify:
  - CCAFS LC-40 have most launches
  - KSC LC-39A sites have relatively higher success rates

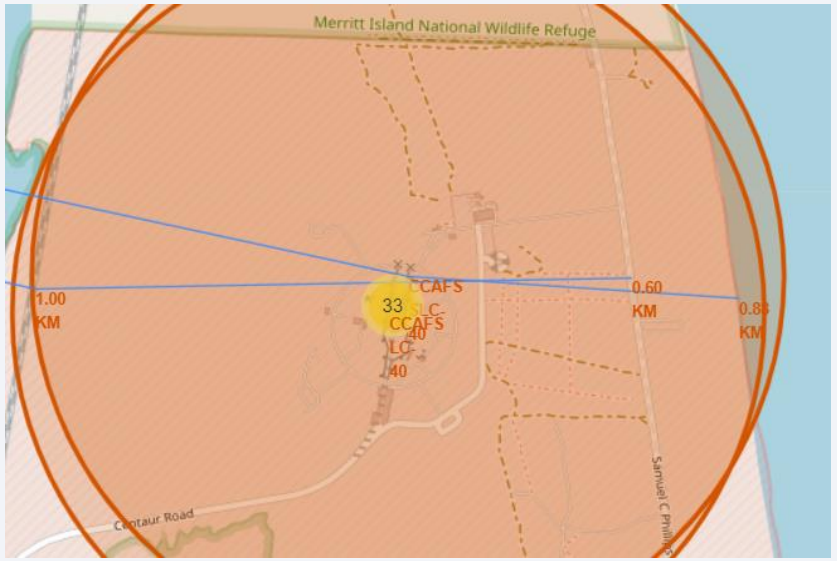


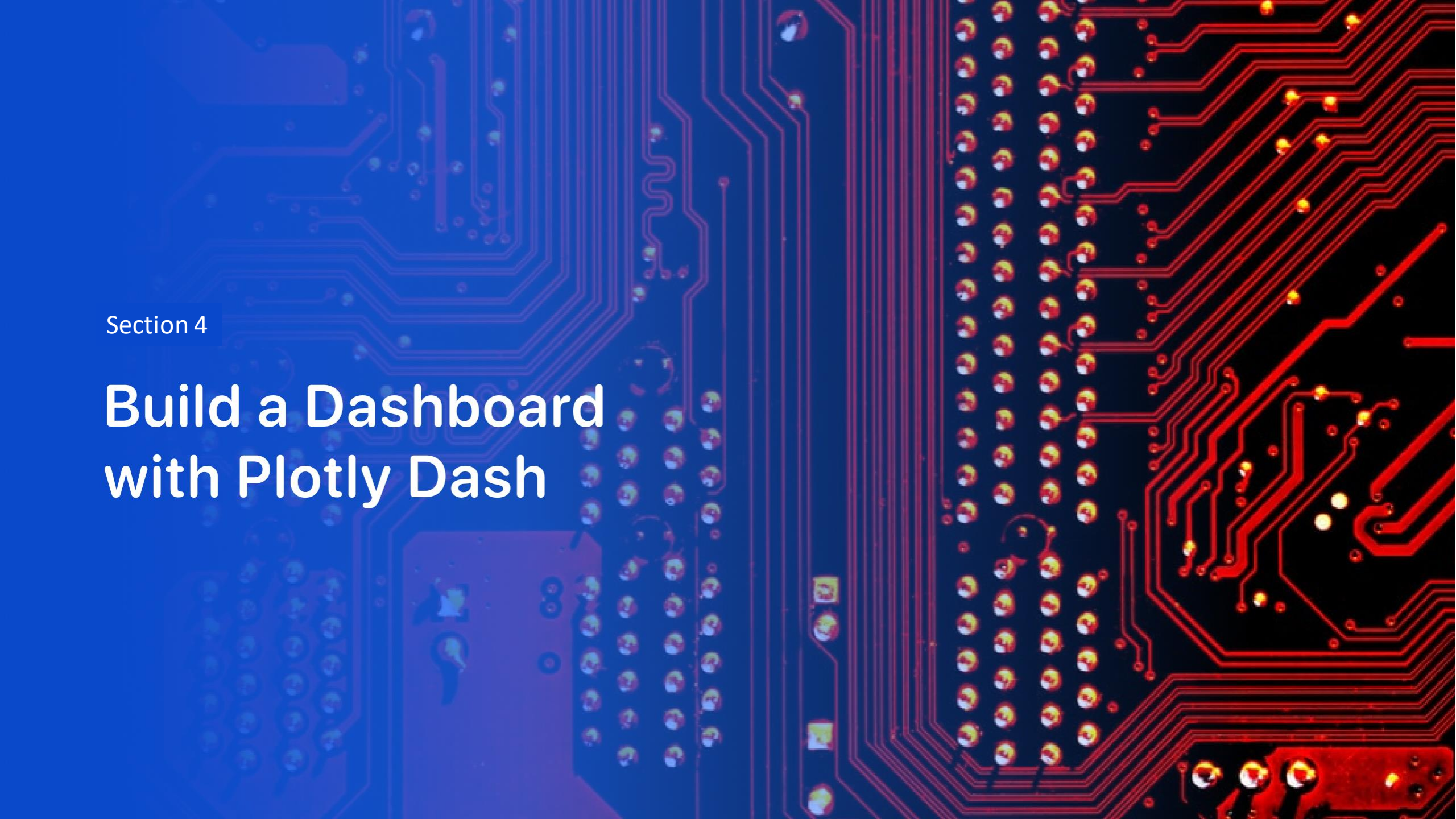


Distances Between a Launch Site to Proximities

- Select CCAFS SLC-40 as an example:
  - Its proximities such as railway, highway, coastline are within 1-km distance for convenient transportation
  - The closest city, Titusville, is 23.16 km away as failed launches could be dangerous in populated areas

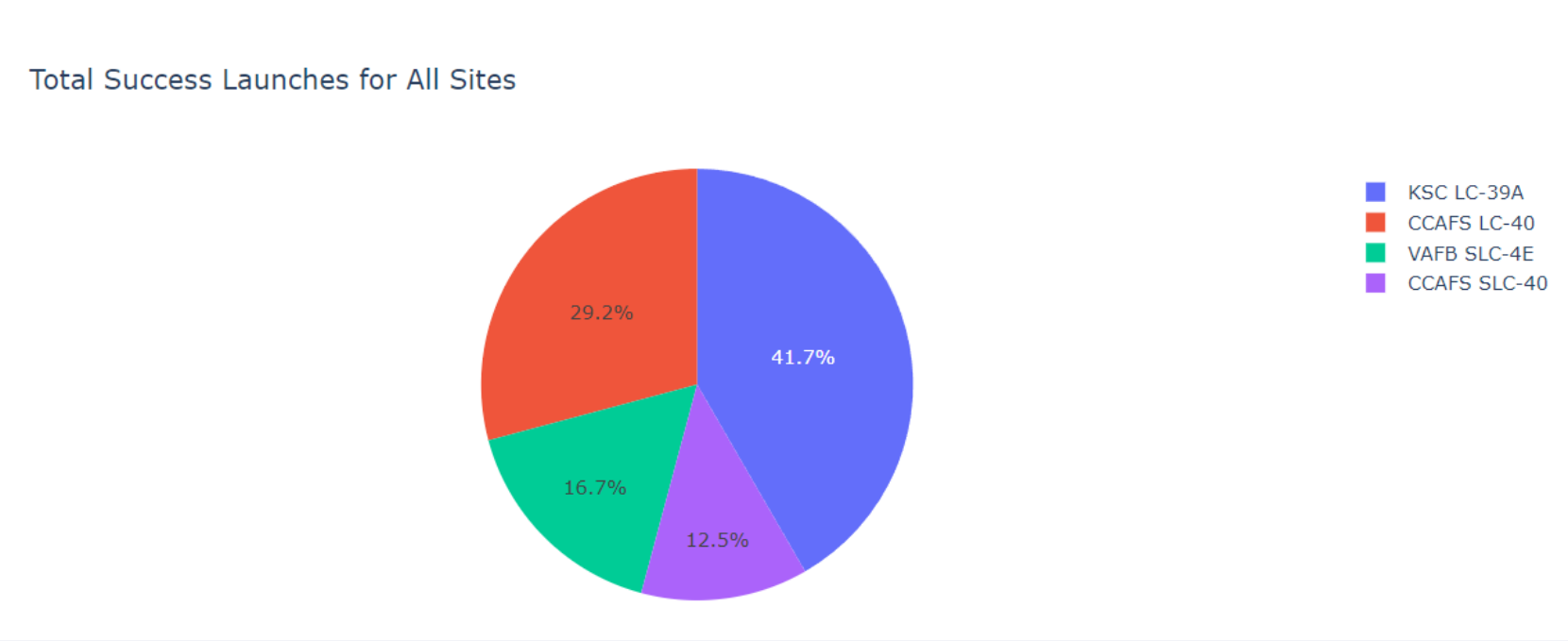





Section 4
Build a Dashboard with Plotly Dash

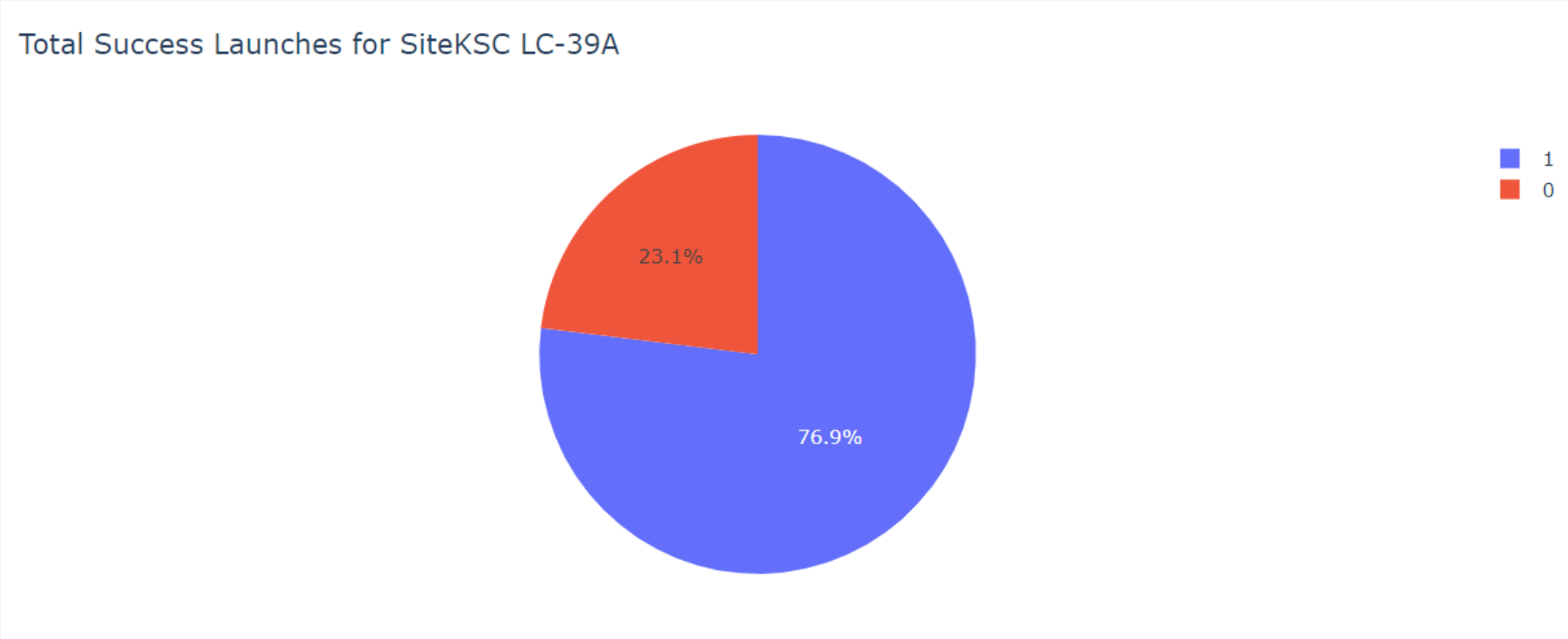
Total Success Launches for All Sites

- KSC LC-39A has the highest success rate among all sites
- CCAFS SLC-40 has the lowest success rate.



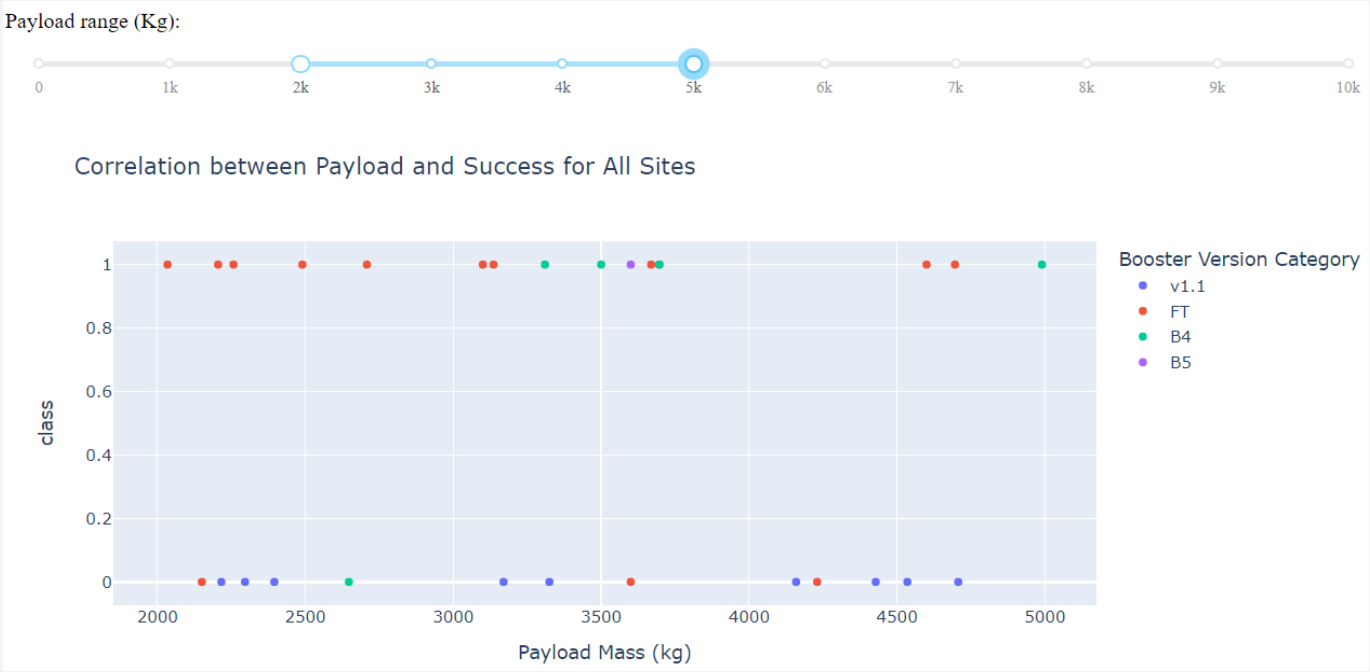
Launch Site with Highest Launch Success Ratio

- KSC LC-39A has the success rate of 76.9%.



Payload vs. Launch Outcome for All Sites

- Most launches are with payload mass between 2,000 to 5,000 KG
- FT booster has the highest success ratio within payload range of 2,000 to 5,000 KG
- Most launches with v1.1 booster are failed within payload range of 2,000 to 5,000 KG



Heavy Payload vs. Launch Outcome for All Sites

- With payload mass above 5,000 KG, the first stage is less likely to land successfully

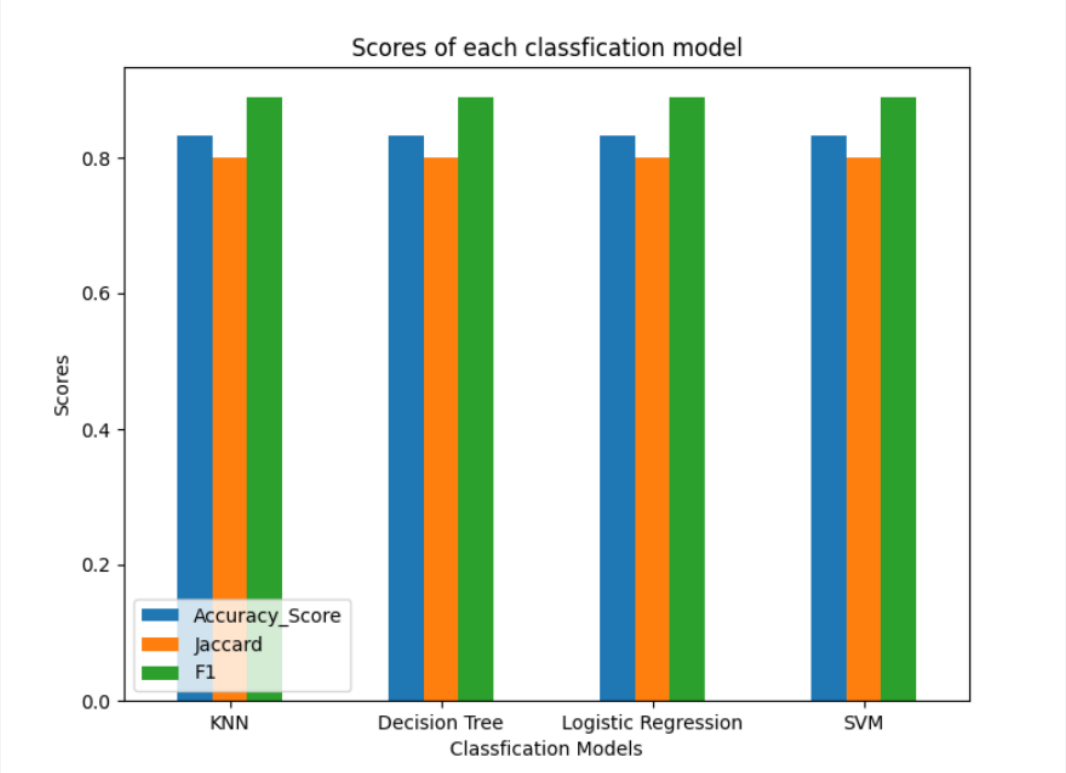
- Only FT and B4 booster launch with heavy load above 5,000 KG

Section 5

Predictive Analysis (Classification)

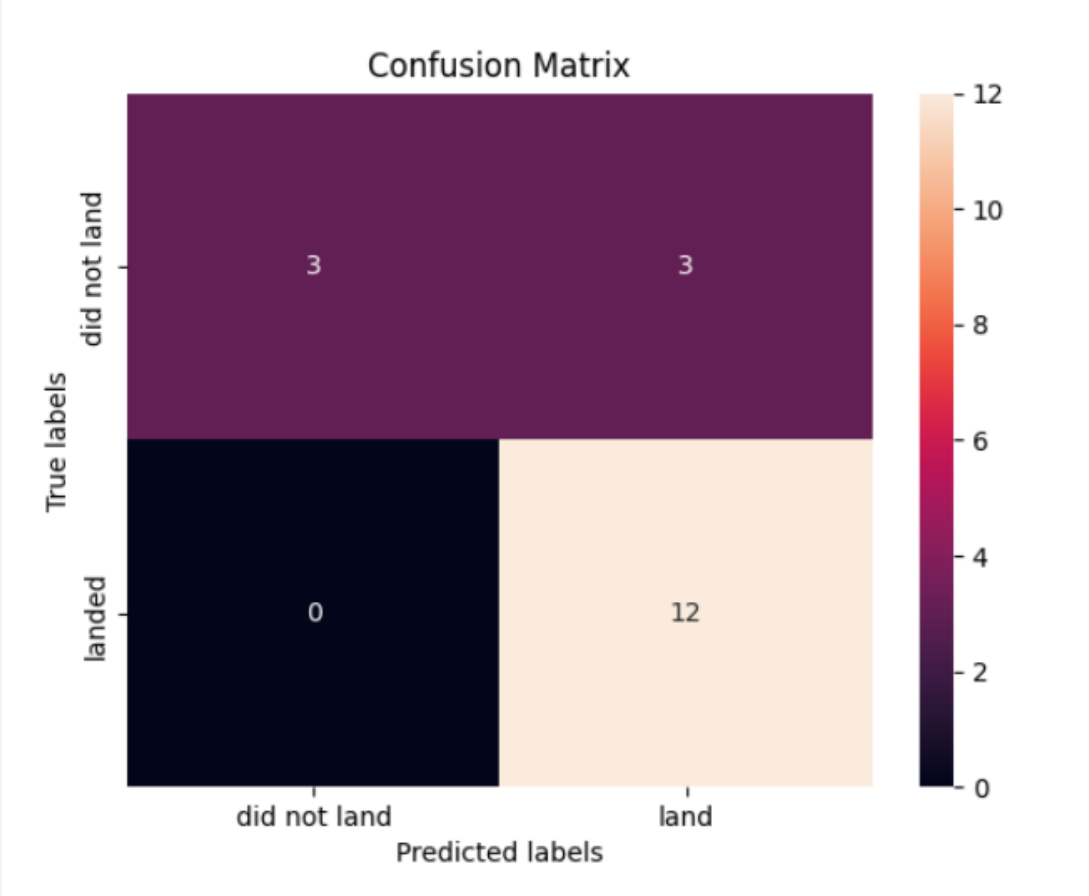
Classification Accuracy with test size 20%

- We split the data X and Y into training and test data. Set the parameter test_size to 0.2 and random_state to 2
- We train four classification models including KNN, Decision Tree, Logistic Regression and SVM based on the training data and get the scores based on the test data
- From the bar chart, we can hardly find a better method to predict the target 'Class'. One reason is that the test set has only 18 samples.



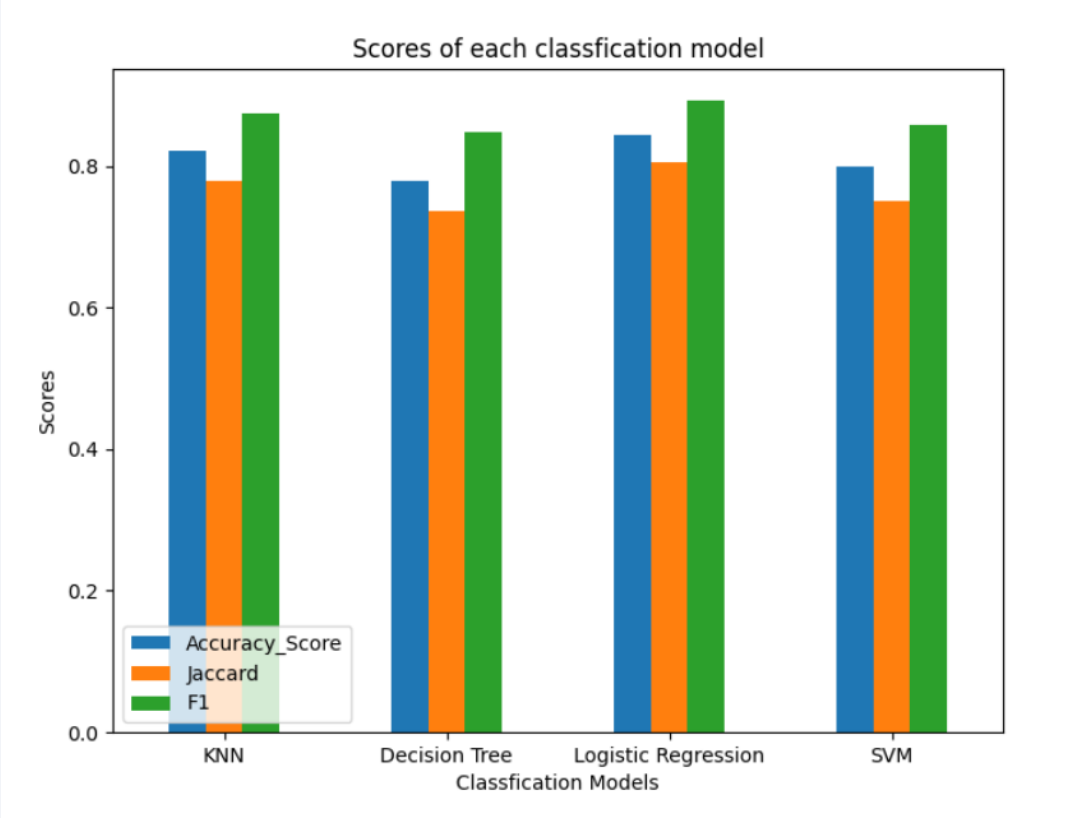
Confusion Matrix with test size 20%

- The built classification models have similar confusion matrix based on the test data
- From the confusion matrix, the major issue is the false positive which is the upper left part
- To find a better model, we need to expand the test data.



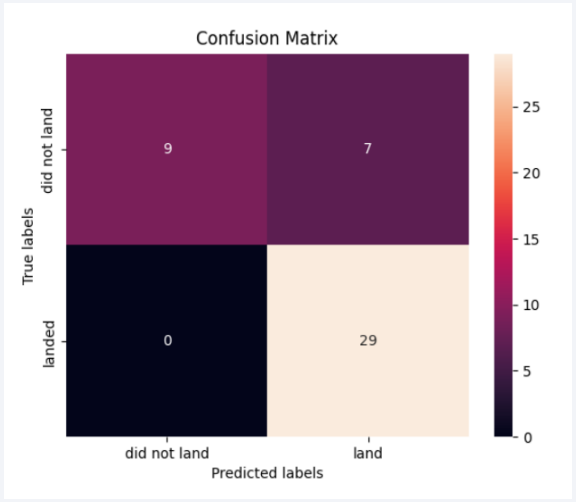
Classification Accuracy with test size 50%

- We reset the test_size to 0.5, keep the random_state as 2
- We train four classification models including KNN, Decision Tree, Logistic Regression and SVM based on the training data and get the scores based on the test data
- From the bar chart, we can find logistic Regression has all the scores higher than other models

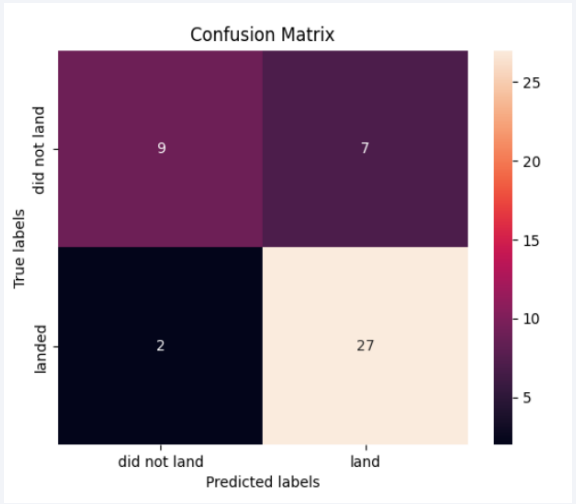


Confusion Matrix with test size 50%

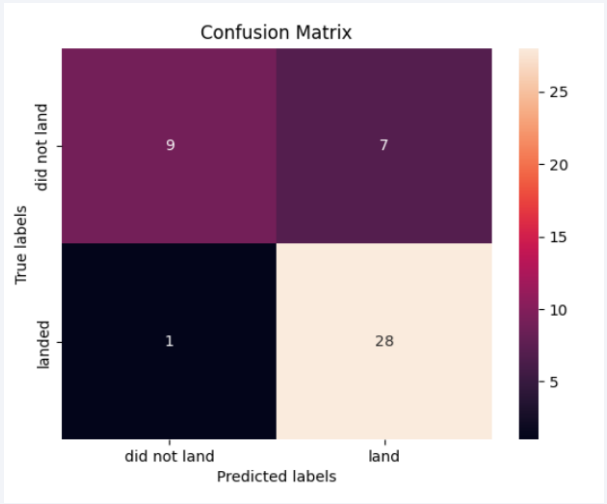
- The Logistic Regression model has a better confusion matrix based on the test data
- From the confusion matrix, the major issue is still the false positive part



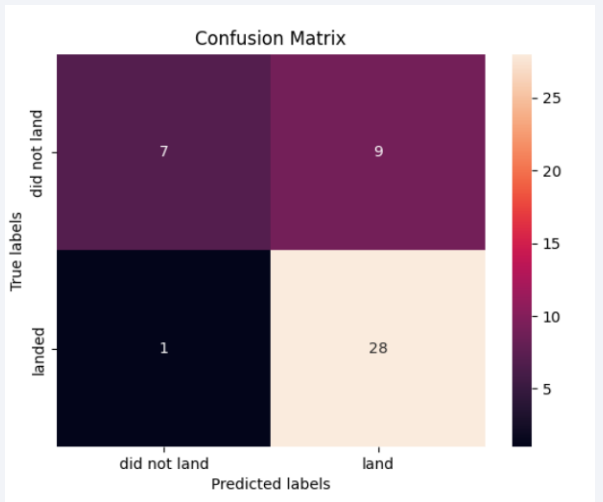
Logistic Regression



SVM



KNN



Decision Tree

Conclusions

- To determine the price of each launch, some key features are payload mass, orbit type, launch site, booster version
- The success rate is increasing based on the yearly trend
- With 2,000 to 5,000 KG payload, the first stage is more likely to land successfully
- KSC LC-39A has the highest success rate among all sites
- FT booster has the highest success rate
- ES-L1, GEO, HEO, SSO orbits have success rates of 100%
- The launch site should be close to highway, railway or coastline, away from populated areas
- Based on our data, Logistic Regression is a better classification model to predict the successful landing of Falcon 9

Appendix

- Dashboard:
  - https://github.com/rowlland/Applied-Data-Science-Capstone/tree/master/dash%20screenshot
- Folium maps:
  - https://github.com/rowlland/Applied-Data-Science-Capstone/tree/master/folium%20maps

Thank you!