IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jisha Augustine
09-10-2004

Outline

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

Executive Summary

Summary of methodologies

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of all results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

- Project background and context

  SpaceX offers Falcon 9 rocket launches at a significantly lower cost than other providers, with a price of $62 million versus $165 million or more from competitors. A critical factor in SpaceX's cost efficiency is its ability to reuse the rocket's first stage. Consequently, predicting whether the first stage will successfully land is crucial to determining the overall cost of a launch. This prediction can be valuable for other companies competing with SpaceX for rocket launch contracts. In this capstone project, we aim to predict the successful landing of the Falcon 9 first stage by using public data and machine learning models.

- Problems you want to find answers
  - What factors will determine the successful landing of the rocket?
  - How do the relationship between different features determine the success rate?
  - What are the conditions that needs to be set for successful landing?
  - Is it possible to predict successful landing?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- SpaceX launch data gathered from an API, specifically the SpaceX REST API.
- Another popular data source for obtaining Falcon 9 Launch data is web scraping-related Wiki pages

- Perform data wrangling
  - Used the method .value_counts() on the column Outcome to determine the number of landing_outcomes. Created a landing outcome label from Outcome column.

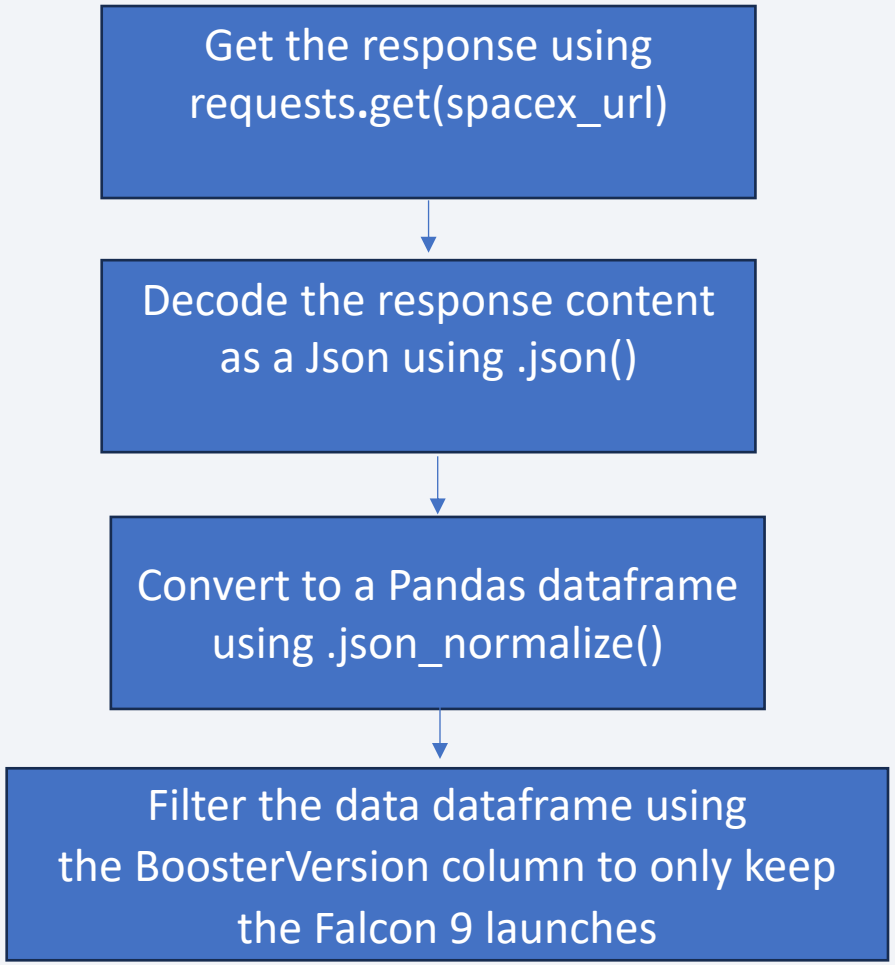
Methodology

- Perform exploratory data analysis (EDA) using visualisation and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Split the dataset into train and validation sets
    - Used Logistic regression classifier, SVM, Decision tree and KNN for prediction
    - GridSearchCV is used for hyperparameter tuning

Data Collection

- SpaceX launch data gathered from an API, specifically the SpaceX REST API.
  - Performed a get request using the requests library to obtain the launch data, which we will use to get the data from the API.
  - The response will be in the form of a JSON; using the json_normalize function, the JSON is converted into a data frame.
  - Filter the Falcon 9 launches data
- Another popular data source for obtaining Falcon 9 Launch data is web scraping-related Wiki pages
  - Python BeautifulSoup package is used to web scrape some HTML tables that contain valuable Falcon 9 launch records

Data Collection – SpaceX API

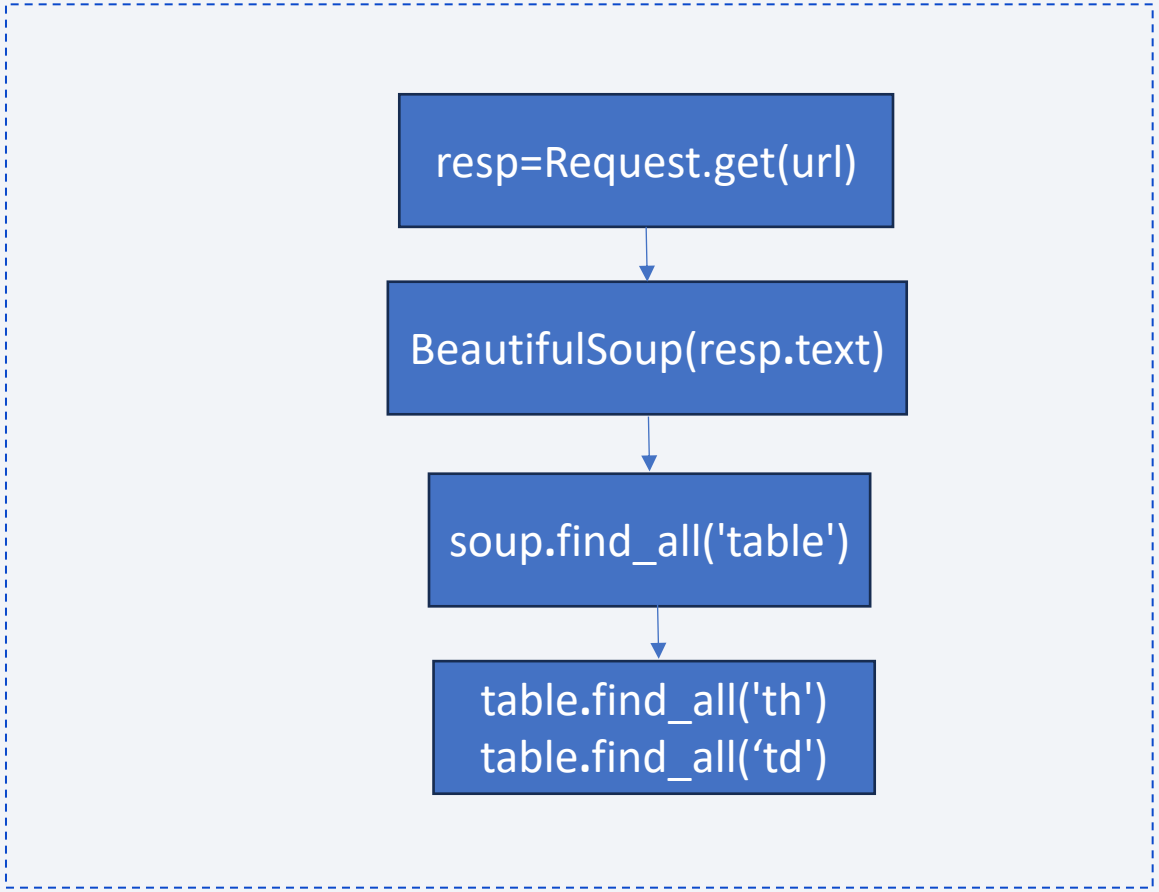


GitHub URL: https://github.com/jishaaugustine/testreponew/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

1. Request the Falcon9 Launch Wiki page from its URL
2. Create a BeautifulSoup object from the HTML response
3. Extract all column/variable names from the HTML table header
4. Create a data frame by parsing the launch HTML tables

GitHub URL:
https://github.com/jishaaugustine/testreponew/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- The number of launches on each site is calculated using the method value_counts() on the column LaunchSite.
- The number and occurrence of each orbit is determined using the method .value_counts() in the column Orbit.
- Used the method .value_counts() on the column Outcome to determine the number of landing_outcomes.
- Created a set of outcomes where the second stage did not land successfully
- Created a landing outcome label from Outcome

GitHub URL: https://github.com/jishaaugustine/testreponew/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

EDA with Data Visualization

- Scatter plot of FlightNumber vs. PayloadMass and overlay the outcome of the launch – As the flight number increases, the first stage is more likely to land successfully
- Scatter plot between FlightNumber and LaunchSite - As the flight number increases, the first stage is more likely to land successfully
- Scatter plot of launch sites and their payload mass - For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000)
- Bar chart for the success rate of each orbit – ES-L1,SSO,HEO and GEO have highest success rate

EDA with Data Visualization

- Scatter plot between FlightNumber and Orbit type - in the LEO orbit, success related to the number of flights. In the GTO orbit, no relationship between flight number and success
- Scatter plot between between Payload Mass and Orbit type - With heavy payloads the successful landing are more for Polar, LEO and ISS.
- Line chart of Year Vs average success rate - sucess rate since 2013 kept increasing till 2020

GitHub URL:
https://github.com/jishaaugustine/testreponew/blob/main/edadataviz.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.
- We applied EDA with SQL to get insight from the data. We wrote queries to get
  - The names of the unique launch sites in the space mission
  - Five records where launch sites begin with the string 'CCA'
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in the ground pad was achieved

EDA with SQL

- The list of names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- The names of the booster_versions which have carried the maximum payload mass
- The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranking 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.

GitHub URL: https://github.com/jishaaugustine/testreponew/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marked all launch sites on a map
  - Created map object
  - Used folium.Circle to add a highlighted circle area with a text label on a specific coordinate
  - Added folium.Circle and folium.Marker for each launch site on the site map
- Marked the success/failed launches for each site on the map
  - Created MarkerCluster object
  - Added a folium.Marker to marker_cluster

Build an Interactive Map with Folium

- Calculated the distances between a launch site to its proximities
  - Added a MousePosition on the map to get coordinate for a mouse over a point on the map
  - Marked down a point on the closest coastline using MousePosition and calculated the distance between the coastline point and the launch site
  - Created a PolyLine between a launch site to the selected coastline point
  - Created a marker with distance to a closest city, railway and highway.
  - Drew line between the marker to the launch site

GitHub URL: https://github.com/jishaaugustine/testreponew/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection
- Added a callback function for `site-dropdown` as input, `success-pie-chart` as output
  - If all sites is selected, displayed a pie chart to show the total successful launches count for all sites
  - Else, Displayed total success launches for selected site
- Added a slider to select payload range

Build a Dashboard with Plotly Dash

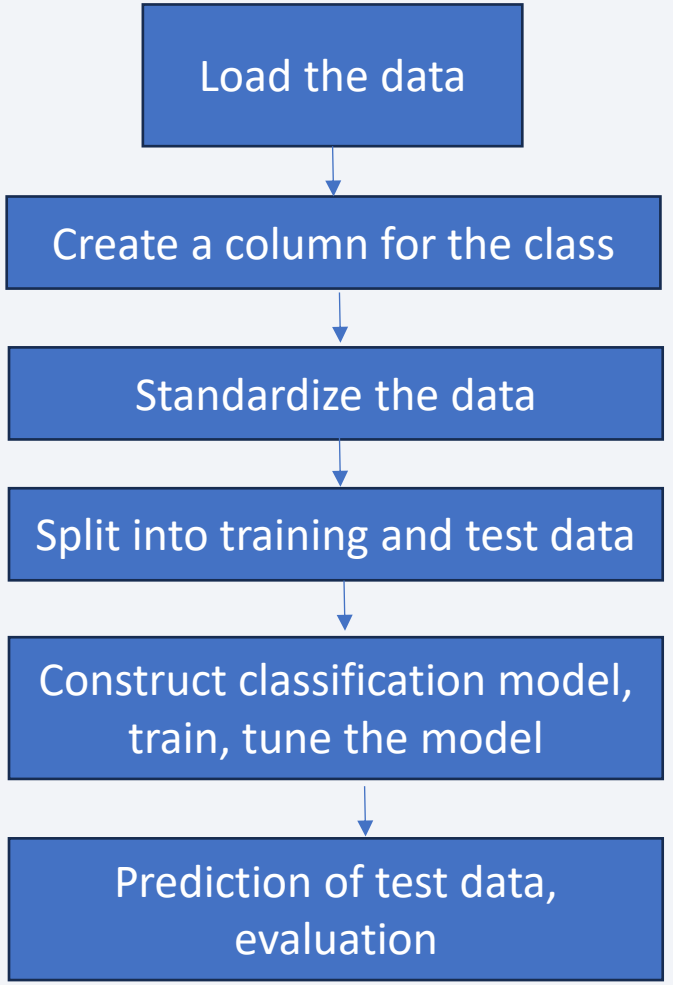
- Added a callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output
    - Added a scatter chart to show the correlation between payload and launch success

GitHub URL:
https://github.com/jishaaugustine/testreponew/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

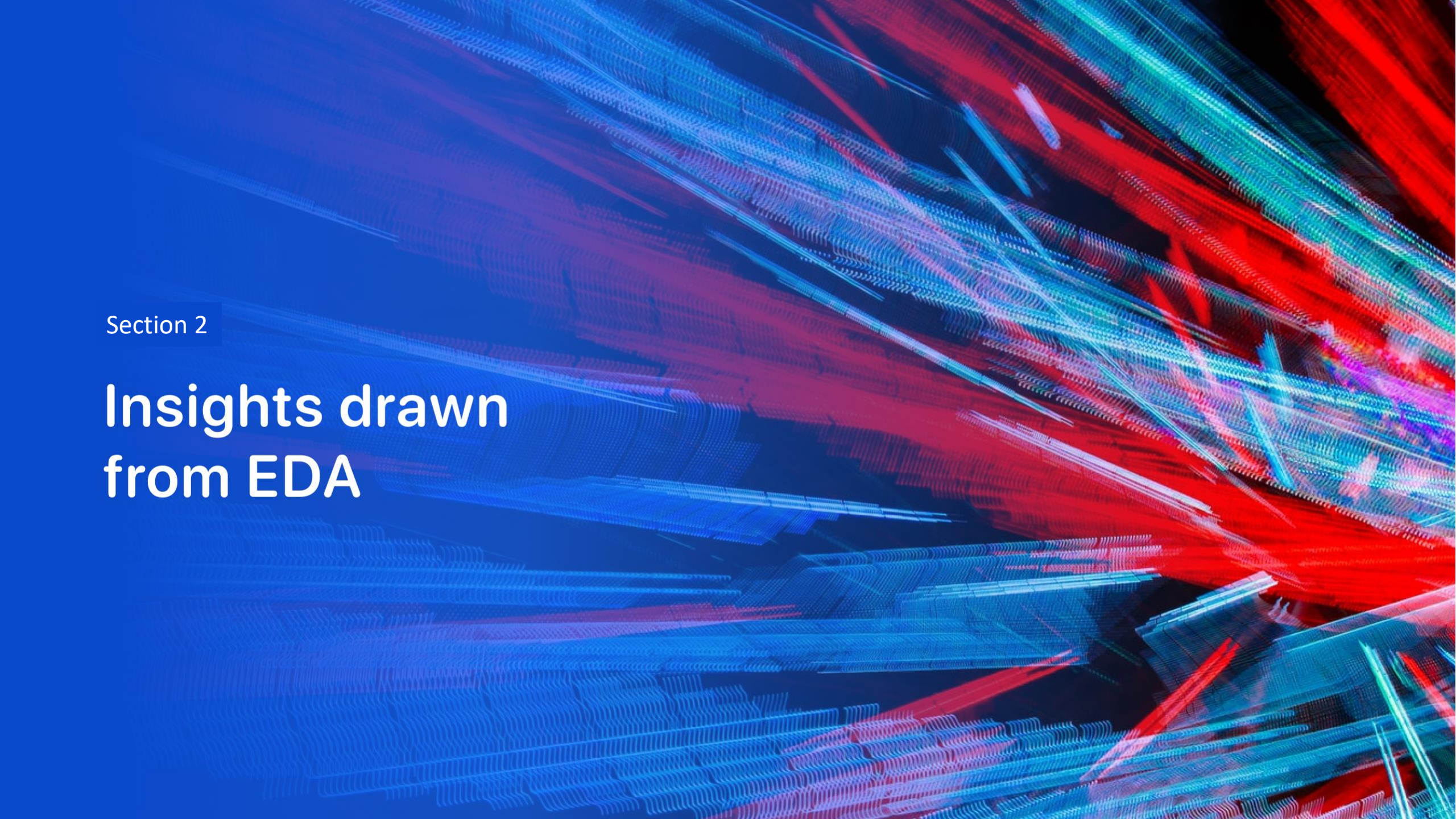
- Loaded the data
- Created class column
- Standardize data using Standardscaler()
- Splitted data into training and test set
- Constructed 4 classification models- LogisticRegression,SVM,Decision tree and KNN
- Train the model using training data
- Performed hyperparameter tuning using GridSearchCV()
- The best model is used to predict test data class
- Used accuracy and confusion metrics for evaluation
- Compare the results of each classifiers

GitHub URL: https://github.com/jishaaugustine/testreponew/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

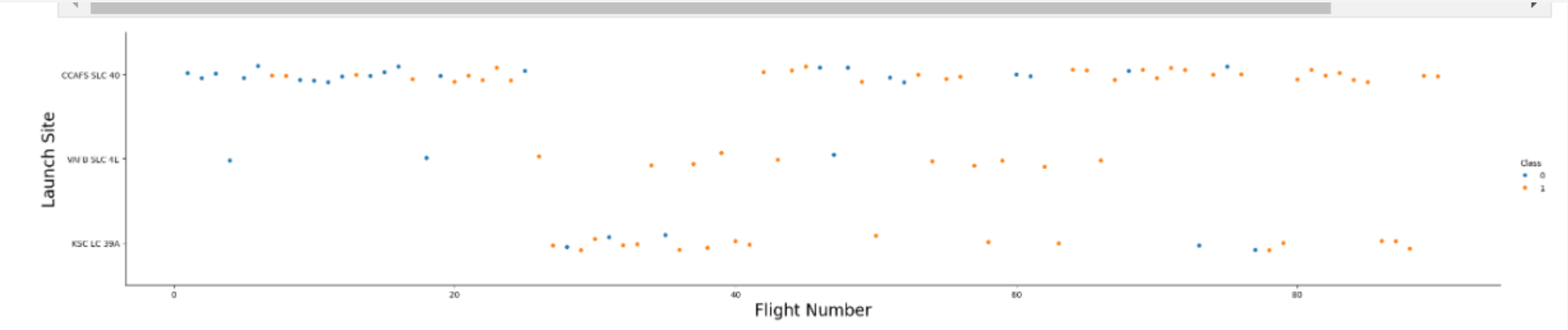


Results

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

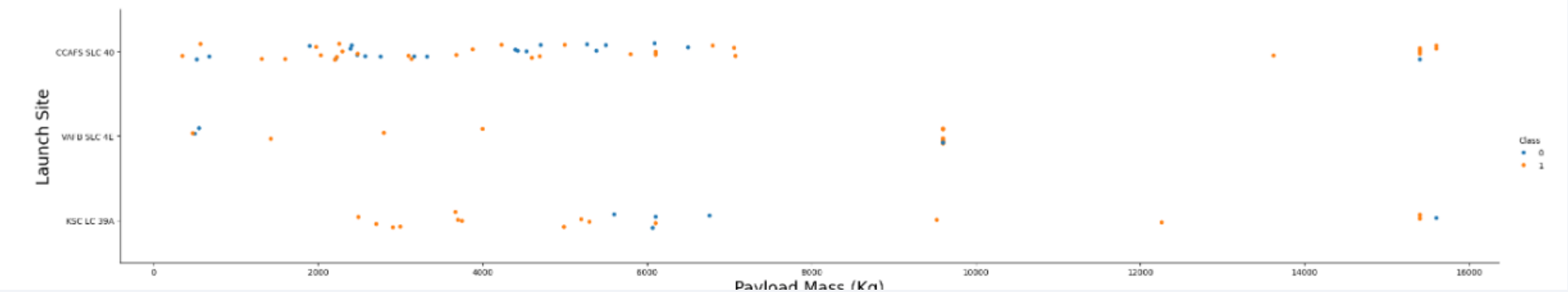

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



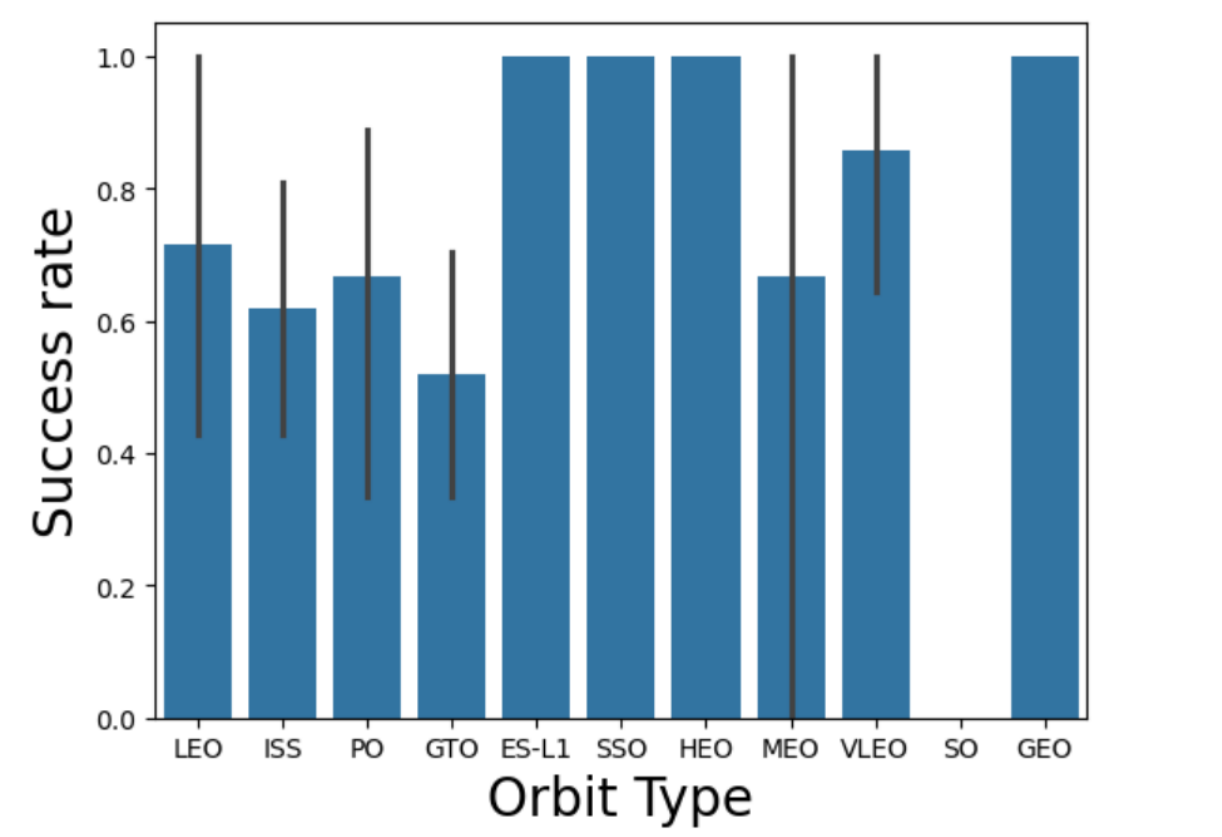
- For all the Launch sites, when flight number increases the first stage is more likely to land successfully.

Payload vs. Launch Site



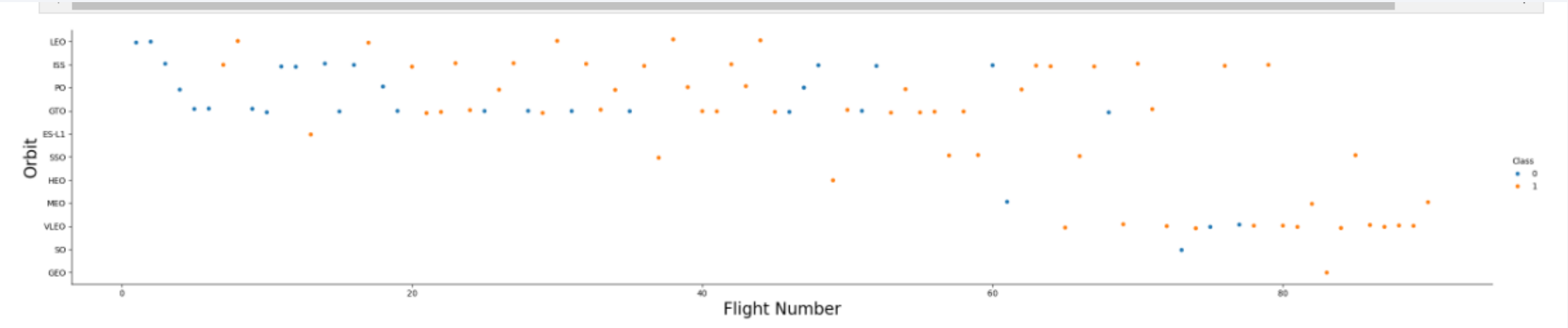
- For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000).
- For the KLC-LC launch site, the success rate is high for light payload mass

Success Rate vs. Orbit Type



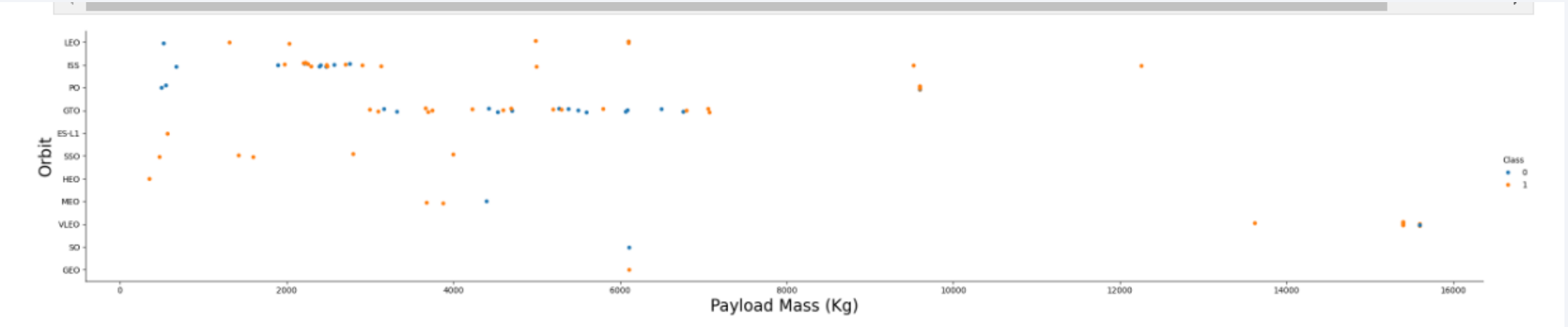
- The orbit types eS-L1, SSO, HEO and GEO have a high success rate
- SO doesn't show any success rate

Flight Number vs. Orbit Type



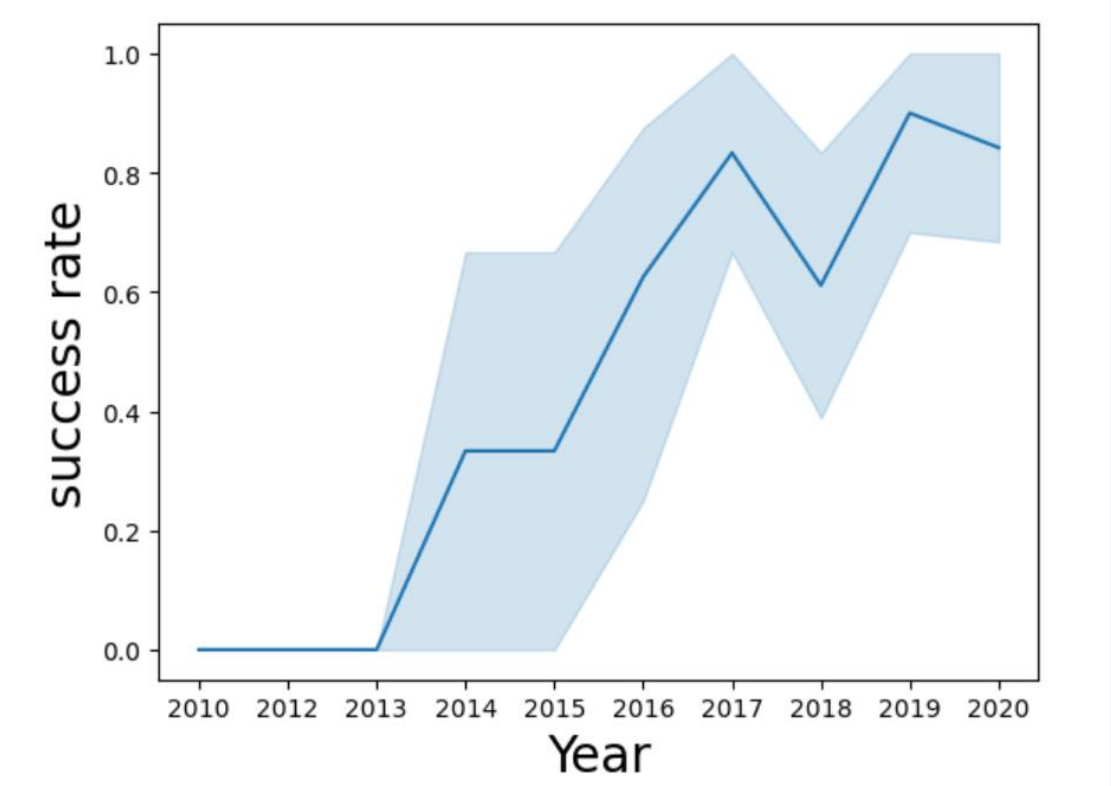
- In the LEO orbit, success seems to be related to the number of flights.
- In the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

- There are 4 unique Launch Site names

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	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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 were calculated as 45596 using SQL query.

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4

Booster_Version	average_payload_mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- The first successful landing outcome on ground pad is on 22nd December 2015

```
%sql select min(Date) from SPACEXTABLE  where Landing_Outcome like 'Success (Ground pad)'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 4 boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- There are 1 Failure and 100 Successful Mission Outcomes

Mission_Outcome	mission_outcome_count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- There are 12 boosters which have carried the maximum payload mass

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

- Two CCAFS LC-40 launch sites have failure landing_outcomes in drone ship for in year 2015

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

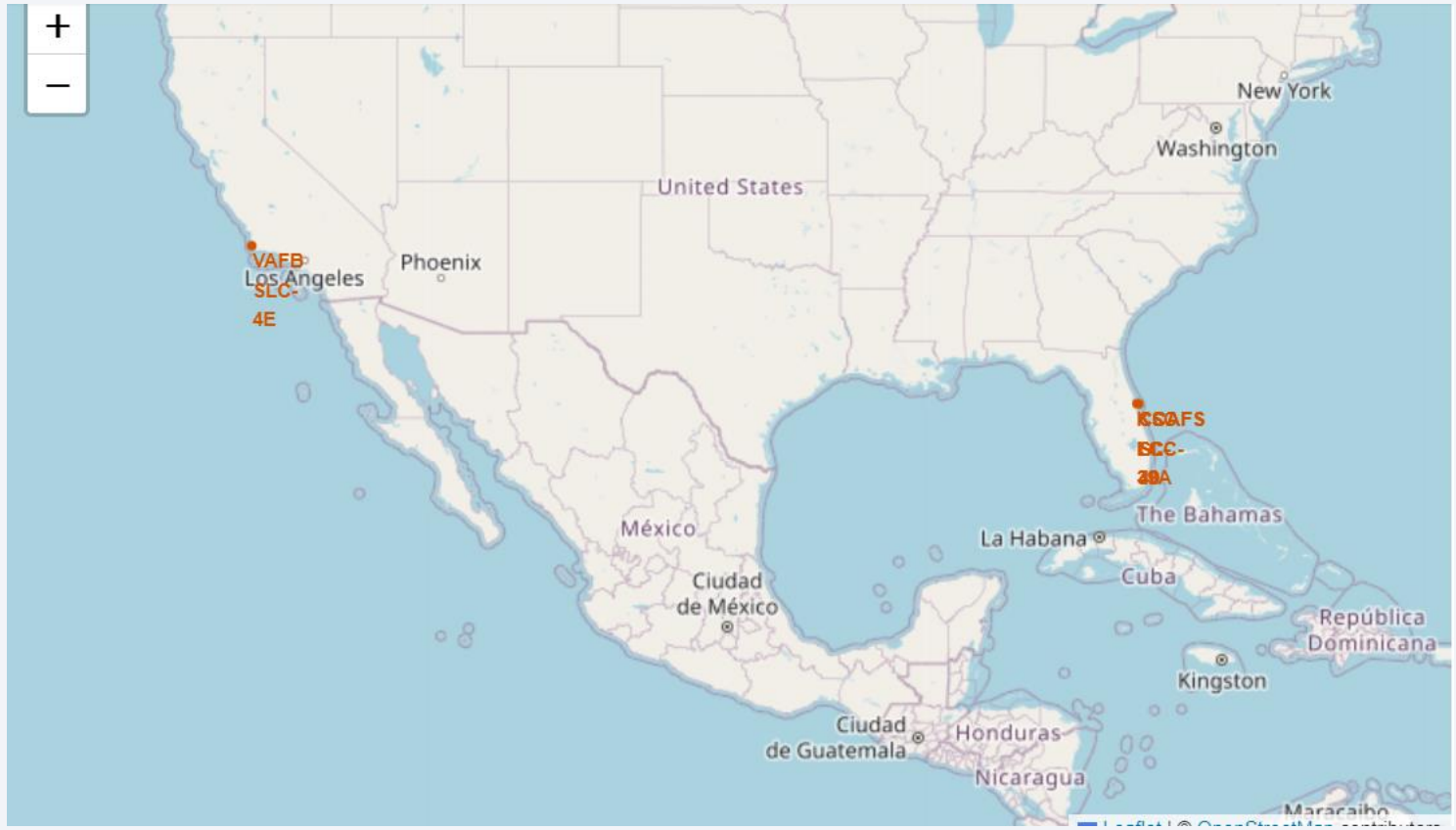
- 'Precluded (drone ship)' has the lowest rank and 'No attempt' has the highest rank.

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

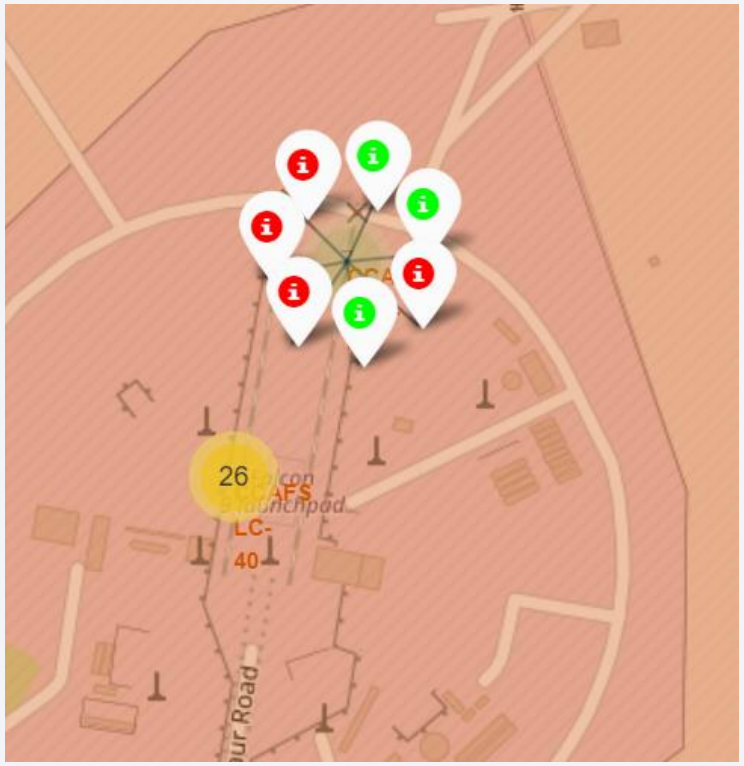
Launch Sites Proximities Analysis

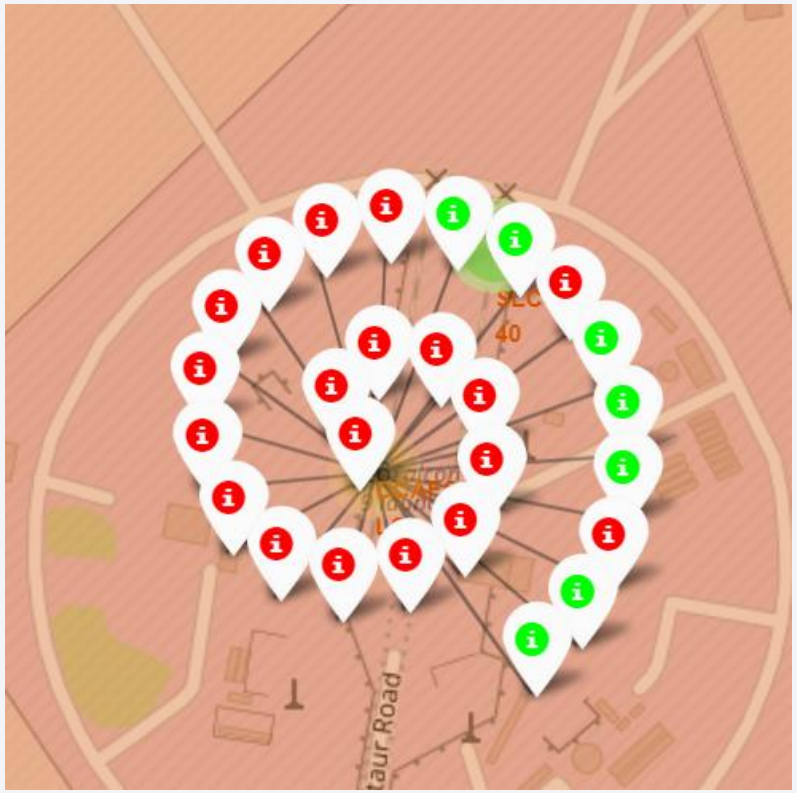
All launch sites



- The launch sites CCAFS LC-40 and CCAFS SLC-40 are marked as overlapping circles.
- VAFB SLC-4E is near Los Angels

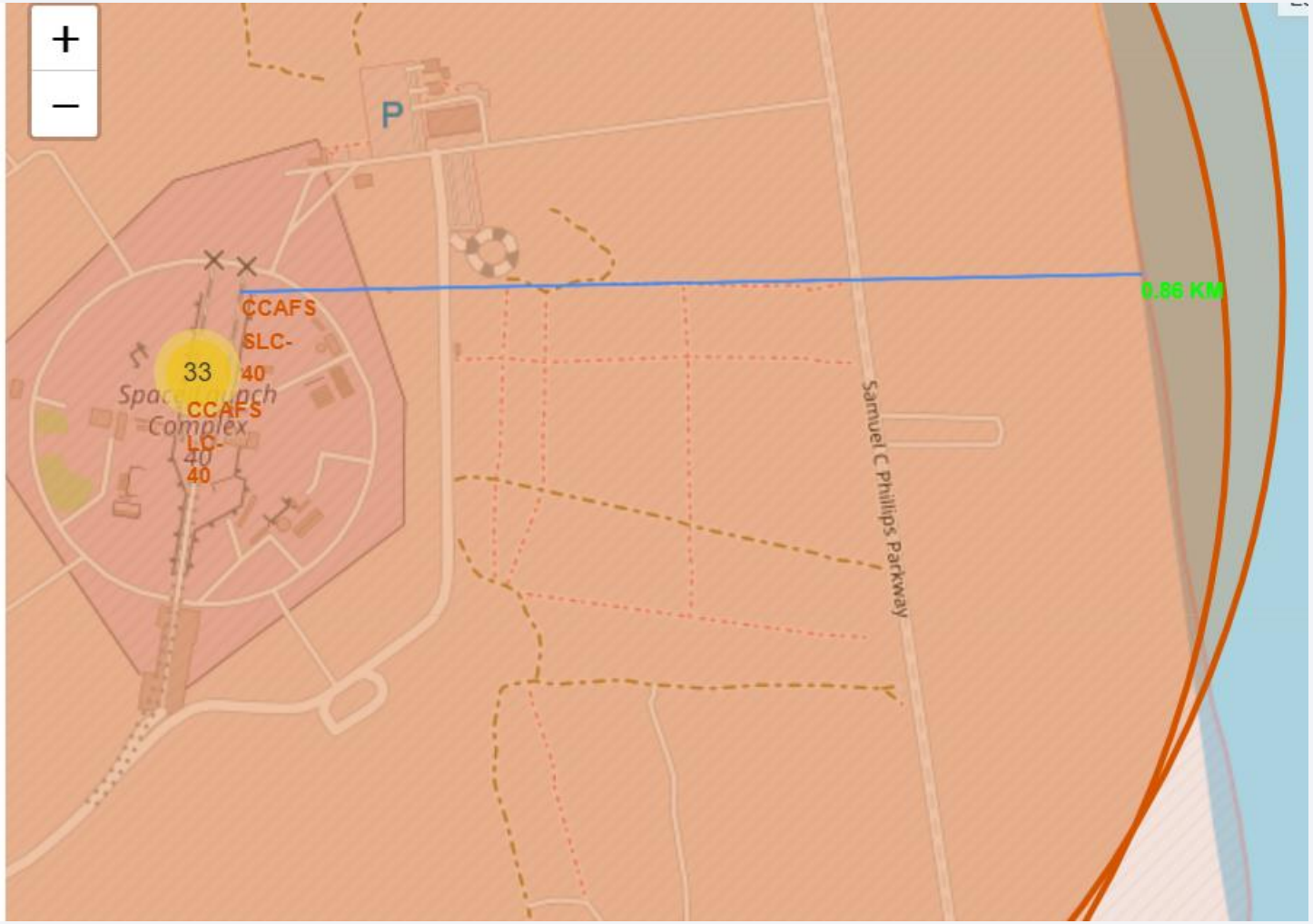
the success/failed launches for each site





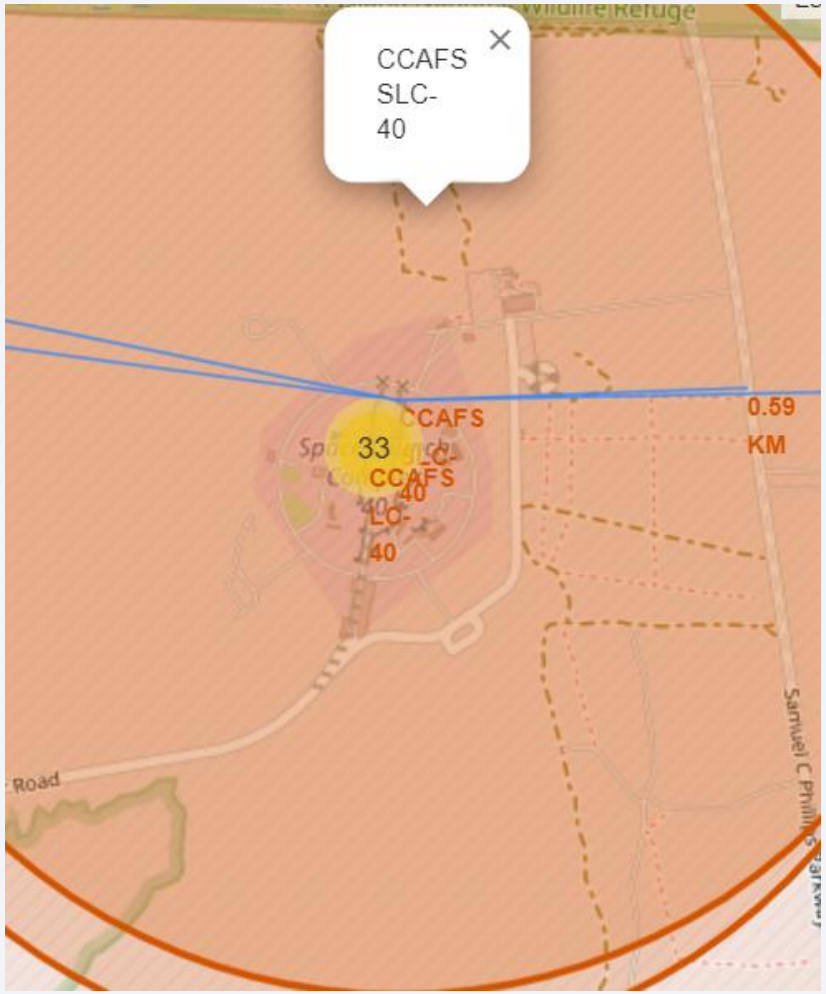
- For CCAFS SLC-40 there are 3 Success and 4 failed Launches
- The red indicates failed launches and the green indicates successful launches

The distances between CCAFS SLC-40 to coastline



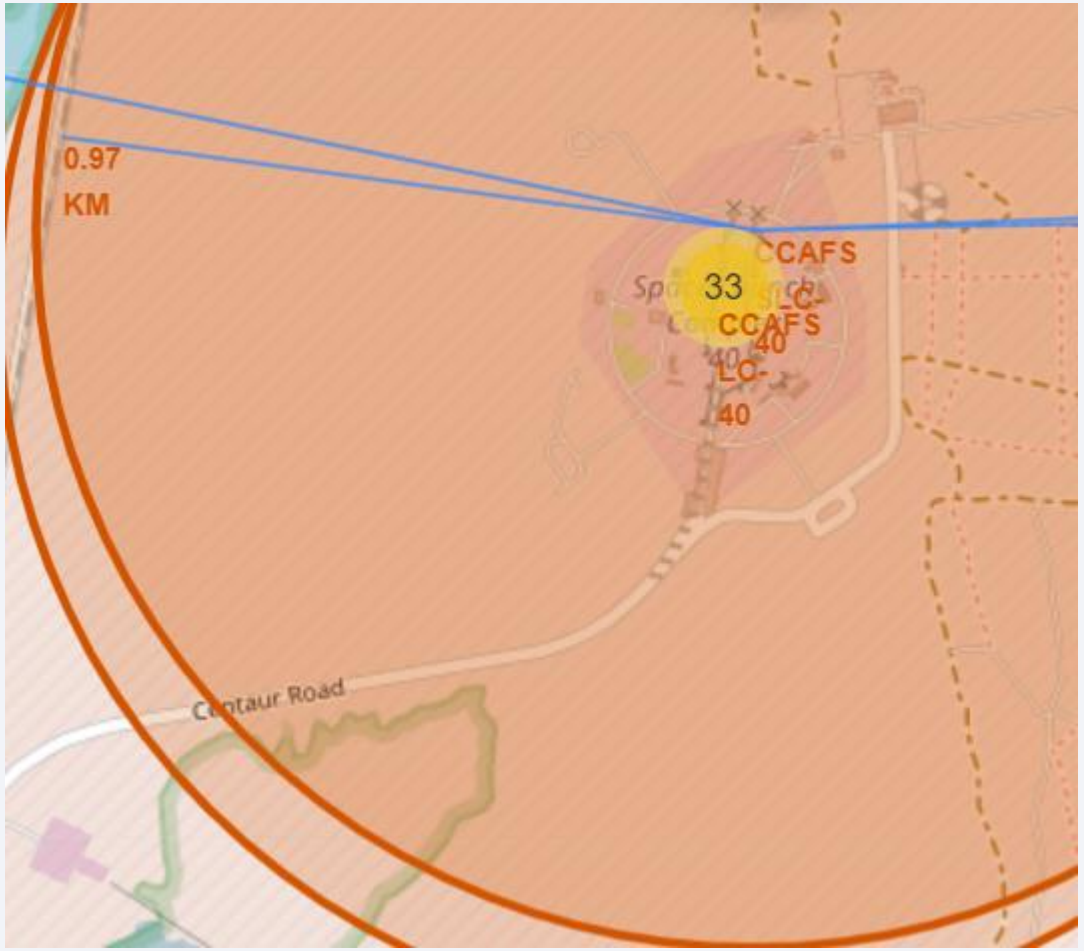
- The distance between CCAFS SLC-40 and nearest coastline is 0.86km

The distances between CCAFS SLC-40 to Highway



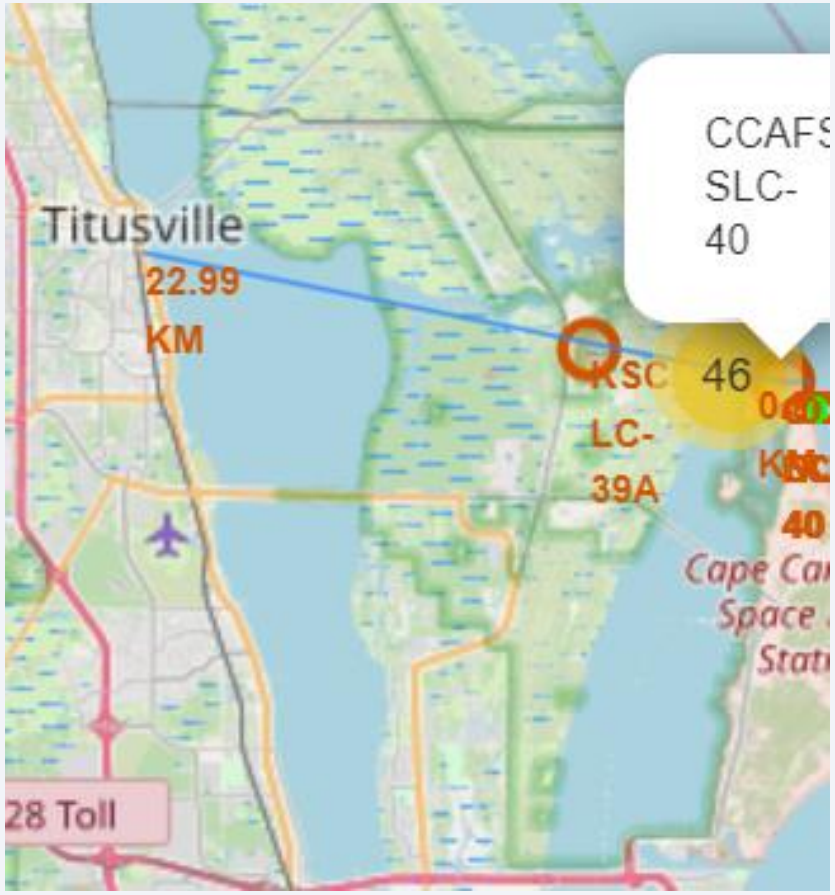
- The distance between CCAFS SLC-40 and nearest Highway is 0.59km

The distances between CCAFS SLC-40 to Railroad



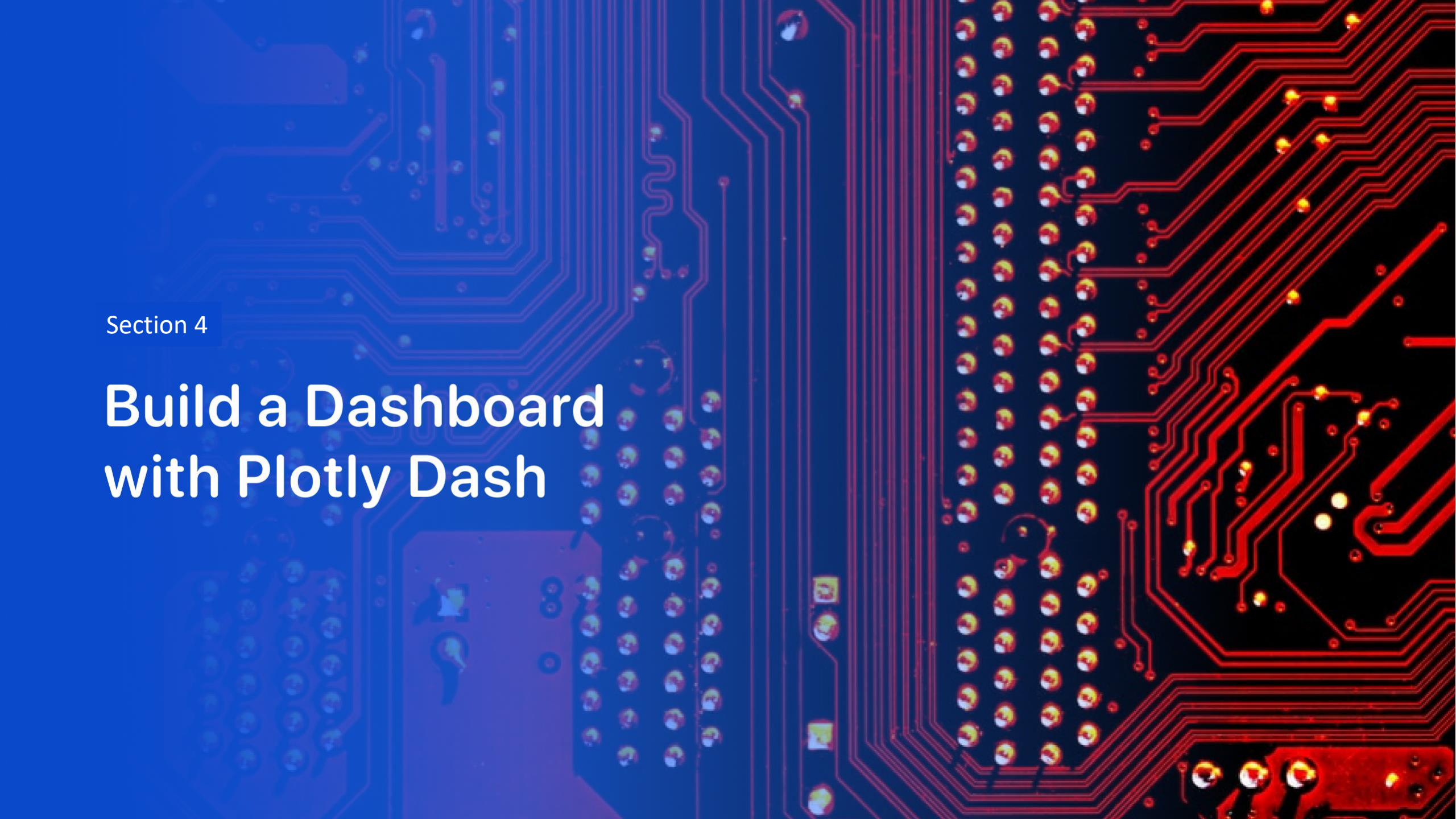
- The distance between CCAFS SLC-40 and nearest Railroad is 0.97km

The distances between CCAFS SLC-40 to City

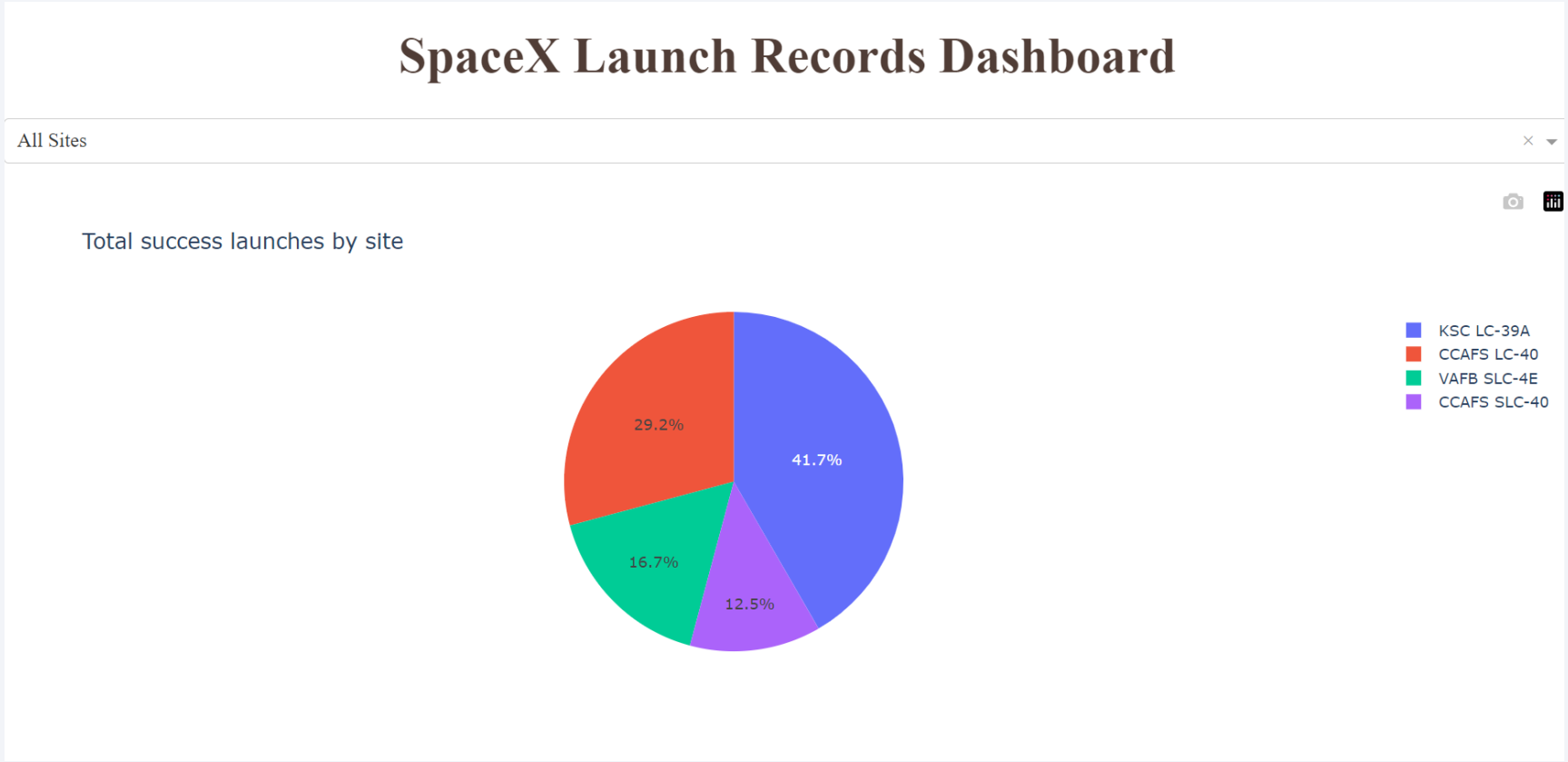


- The distance between CCAFS SLC-40 and nearest city is 22.99km

- 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? No

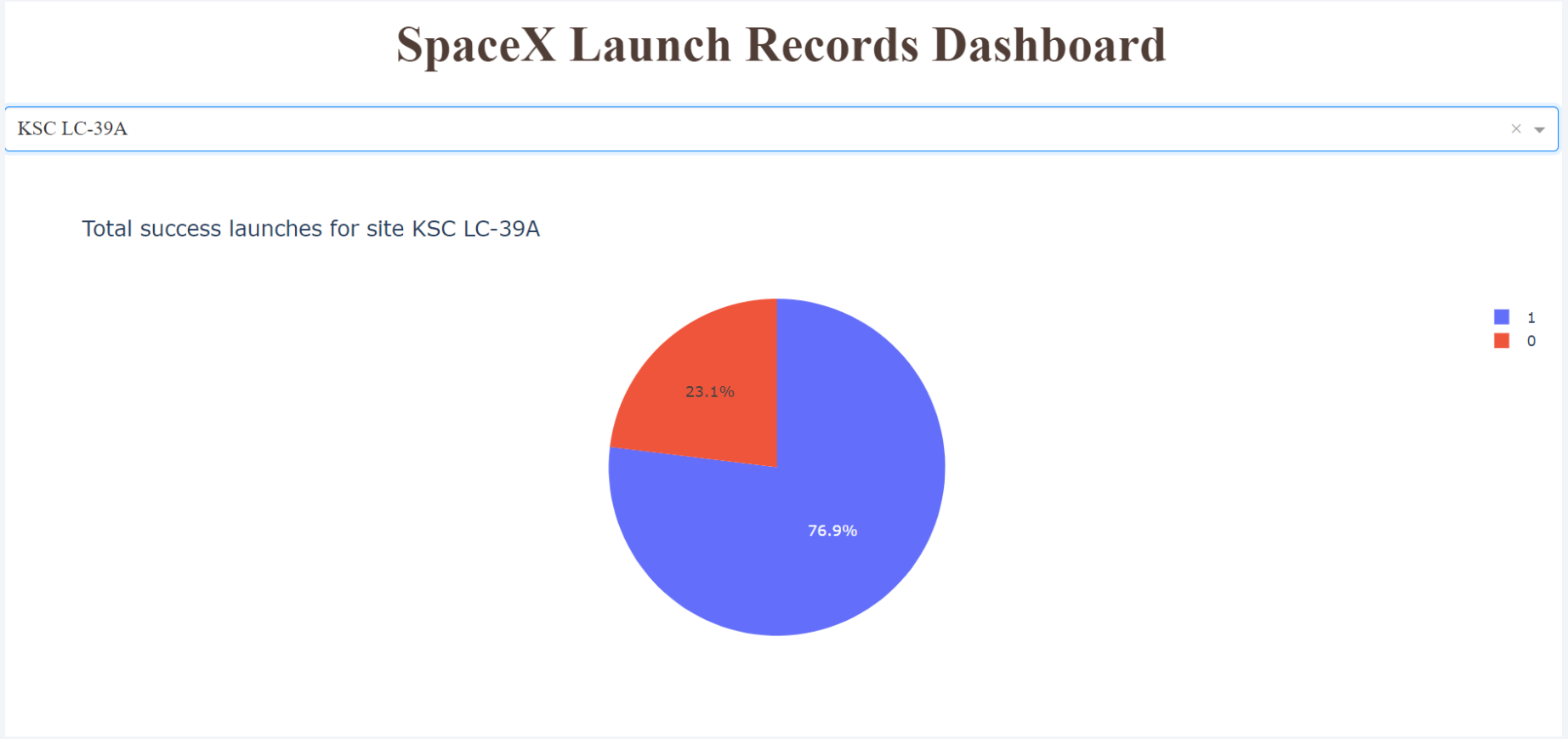

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



- The success rate is highest in launch site KSC LC-39A and lowest in CCAFS SLC-40

The launch site with highest launch success ratio

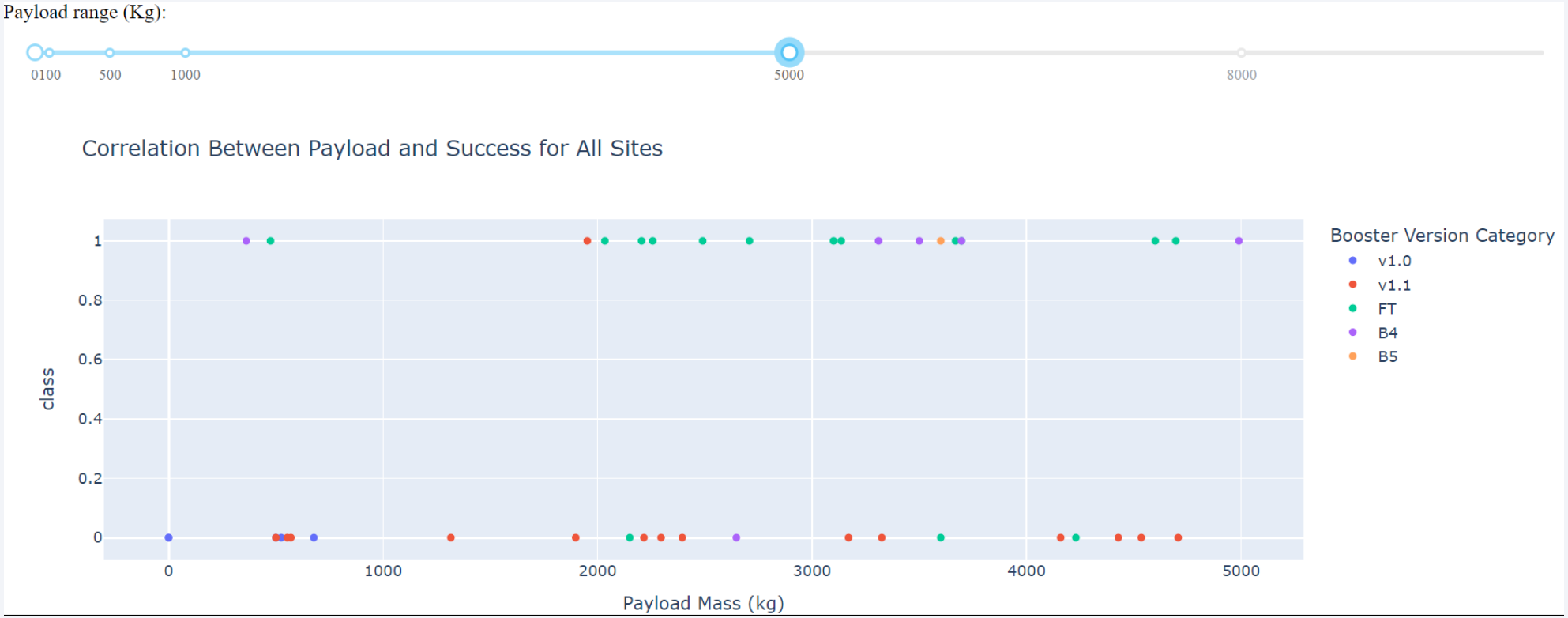


- The success launches for the site KSC LC-39 A is 76.9%

Payload vs. Launch Outcome scatter plot for all sites

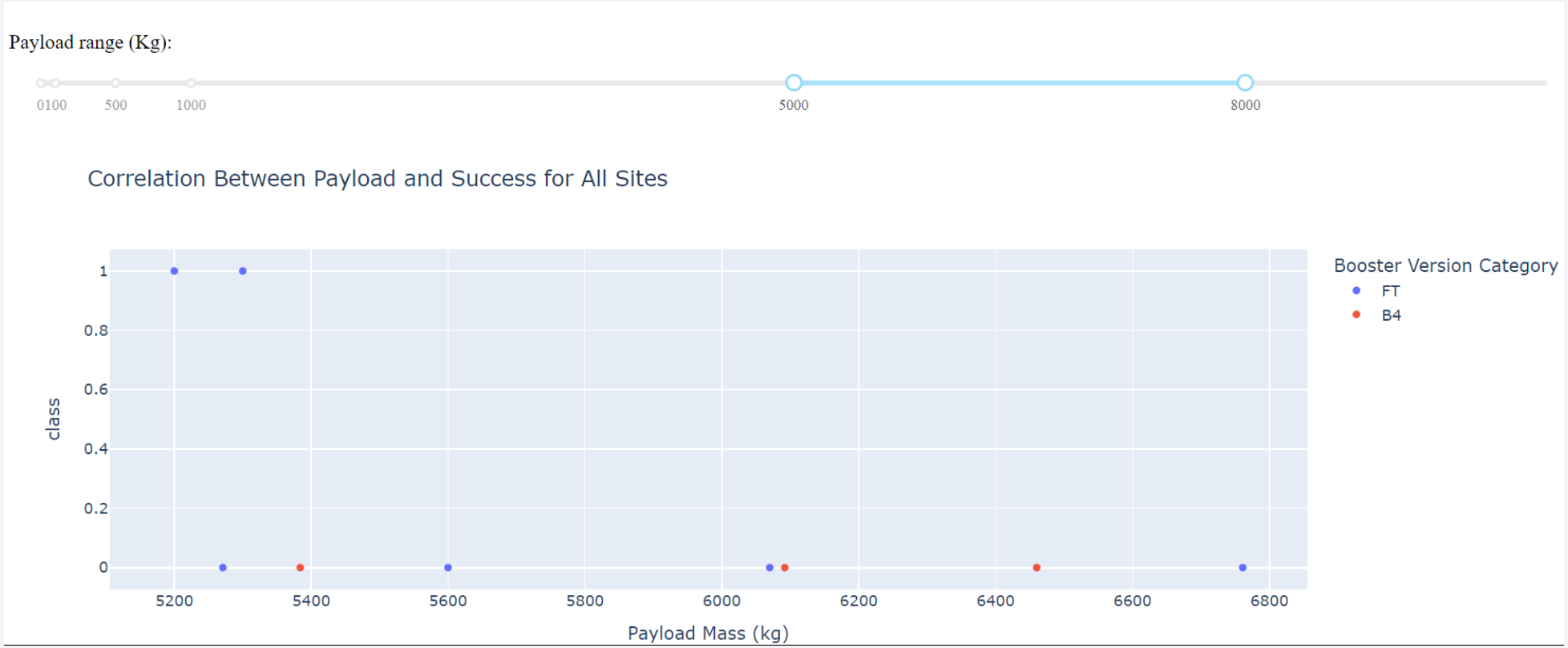
On low payload mass range(0-500) only booster version FT and B4 have success.

Payload vs. Launch Outcome scatter plot for all sites



On range 0-5000 V1.0 does not have any success

Payload vs. Launch Outcome scatter plot for all sites



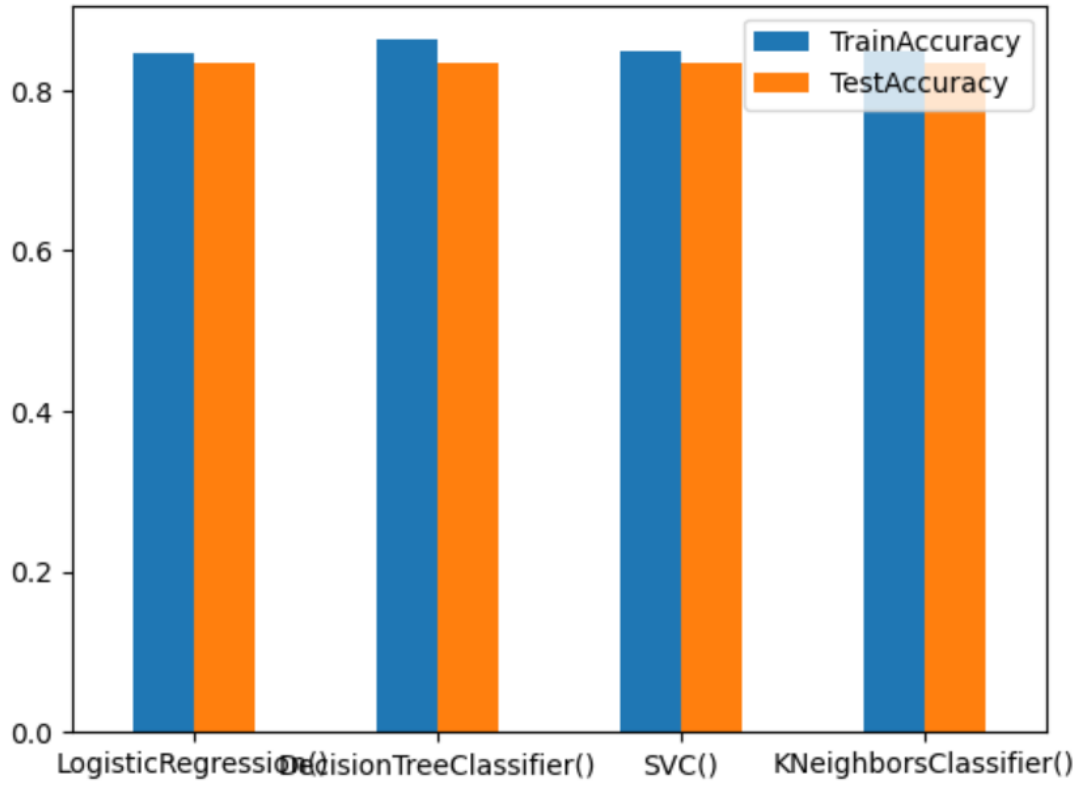
Between 5000 and 8000, only FT and B4 present

Section 5

Predictive Analysis (Classification)

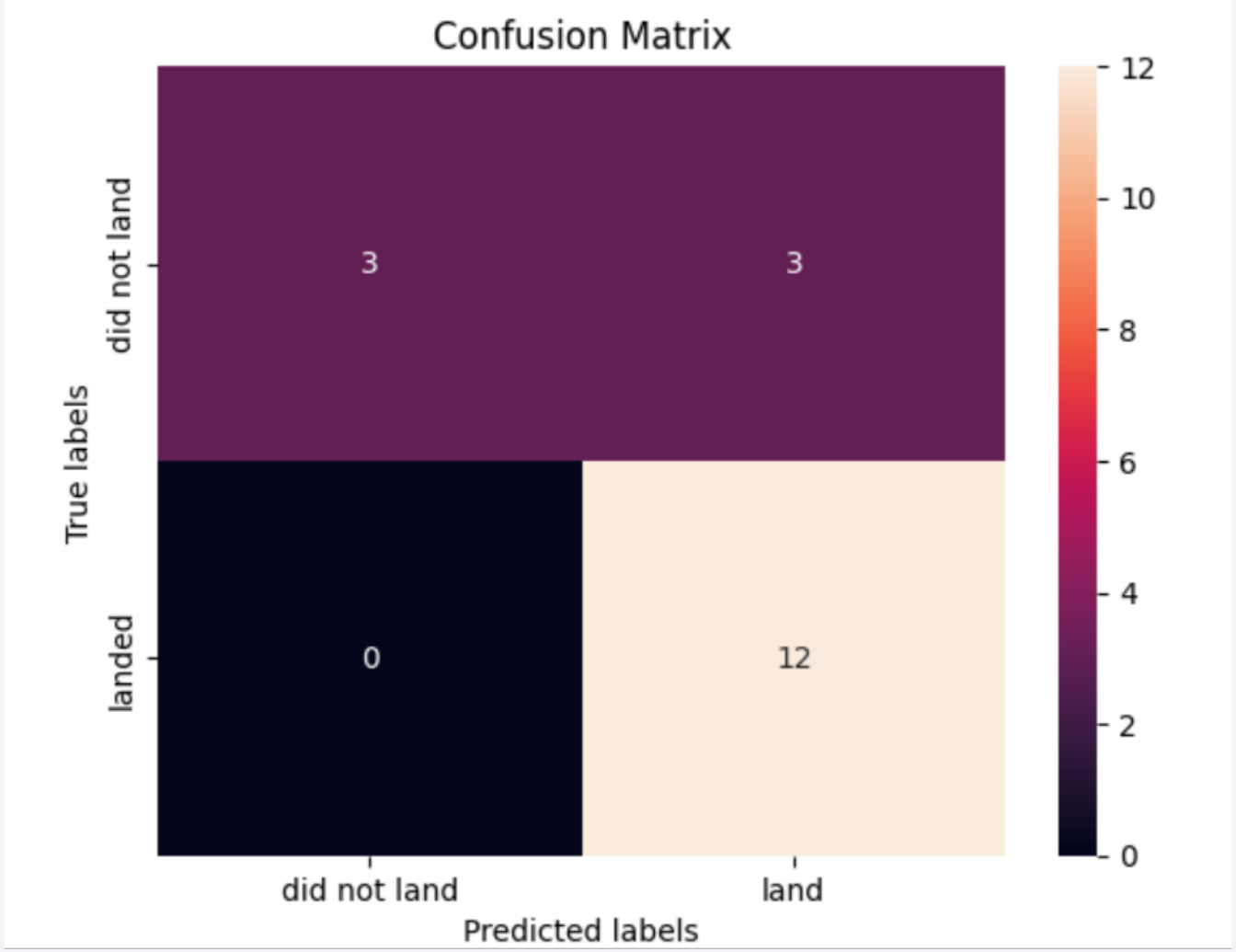
Classification Accuracy

```
accuracy of LogisticRegression() is: 0.8333333333333334
accuracy of DecisionTreeClassifier() is: 0.8333333333333334
accuracy of SVC() is: 0.8333333333333334
accuracy of KNeighborsClassifier() is: 0.8333333333333334
9]: <AxesSubplot:>
```



- All 4 Models have the same test accuracy- 83.33%
- Highest training accuracy is for the DecisionTree classifier -86.25%

Confusion Matrix



- All 4 models have same confusion matrix
- There are three false positives .i.e., unsuccessful landing marked as successful landing by the classifier.
- There are no false negatives

Conclusions

We can conclude that:

- The flight number at a launch site increases, the greater the success rate at a launch site.
- The orbit types eS-L1, SSO, HEO and GEO have a high success rate.
- The success rate since 2013 kept increasing till 2020
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task

Thank you!