IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pedro Cardoso
25/02/2023

Outline

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

Executive Summary

- **Summary of methodologies**
- Data Collection API
- Data Collection with Web Scraping
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Interactive Visual Analytics with Folium
- Machine Learning Prediction
- **Summary of all results**
- Exploratory Data Analysis results
- Interative maps and dashboard
- Predictive results

Introduction

- **Project background and context**
- The objective of this project is to predict if the Falcon 9 first stage will successfully land. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. By determining if the stage will land, we can determine the cost of a launch.
- **Problems you want to find answers**
- What differs between a successful and failed landing?
- What are the effects of each relationship of the rocket's variables on the success or failure of the landing?
- What are the conditions which will allow SpaceX to achieve the best landing success rate?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- 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
  - How to build, tune, evaluate classification models

Data Collection

- **The Datasets are collected from Rest SpaceX API and webscrapping Wikipedia**

These are the steps:

1. SpaceX Rest API call
2. API returns JSON file
3. Make Dataframe from JSON
4. Clear Data and export it

- **The information obtained by the webscrapping of Wikipedia are launches, landing, payload information**
- These are the steps:

1. Get HTML response from Wikipedia
2. Extract Data with BeautifulSoup
3. Make Dataframe
4. Export data

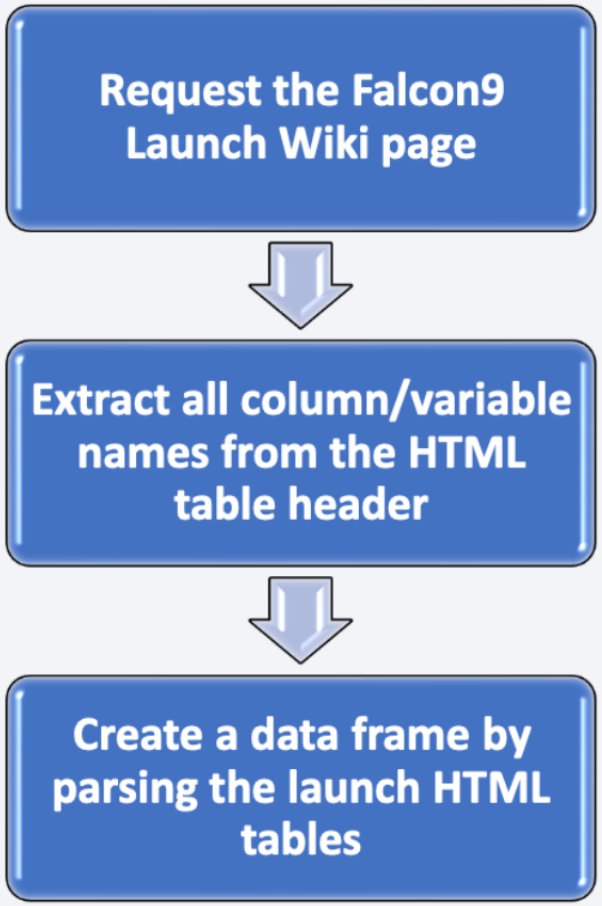
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then be used;
- This API was used according to the flow chart beside and then data is persisted.

https://github.com/PDCasC/Data-Science-Project/blob/master/Data Collection.ipynb

Data Collection - Web Scraping

- Data from SpaceX can also be obtained from Wikipedia;
- Data is downloaded from Wikipedia according to the flowchart and then persisted.

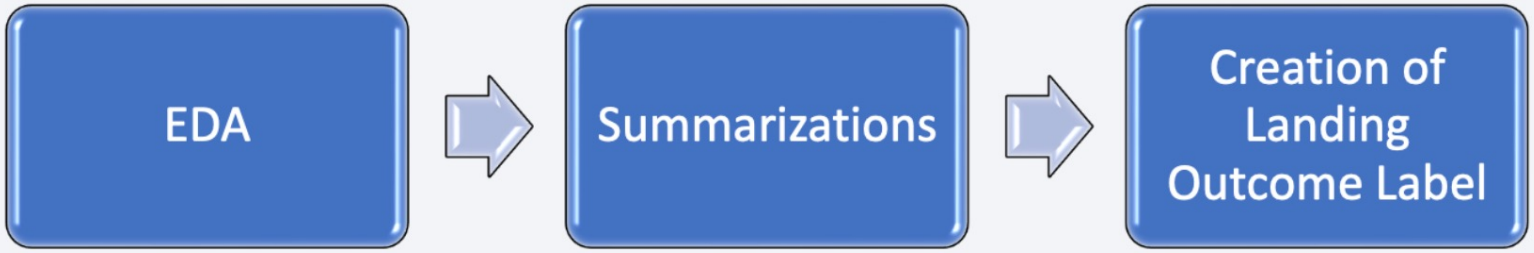


https://github.com/PDCasC/Data-Science-Project/blob/master/Data Collection with Web Scraping.ipynb

Data Wrangling

- EDA - Exploratory Data Analysis was performed on the dataset
- The summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type was calculated
- The landing outcome label was created from Outcome Column

https://github.com/PDCasC/Data-Science-Project/blob/master/Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatter Graphs
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Orbit vs. Flight Number
  - Payload vs. Orbit Type
  - Orbit vs. Payload Mass

Scatter plots show relationship between variables. This relationship is called the correlation.

- Bar Graph
  - Success rate vs. Orbit

Bar graphs show the relationship between numeric and categoric variables.

- Line Graph
  - Success rate vs. Year

Line graphs show data variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data.

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

https://github.com/PDCasC/Data-Science-Project/blob/master/EDA%20with%20SQL%20Lab.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
  - Lines are used to indicate distances between two coordinates.

https://github.com/PDCasC/Data-Science-Project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

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

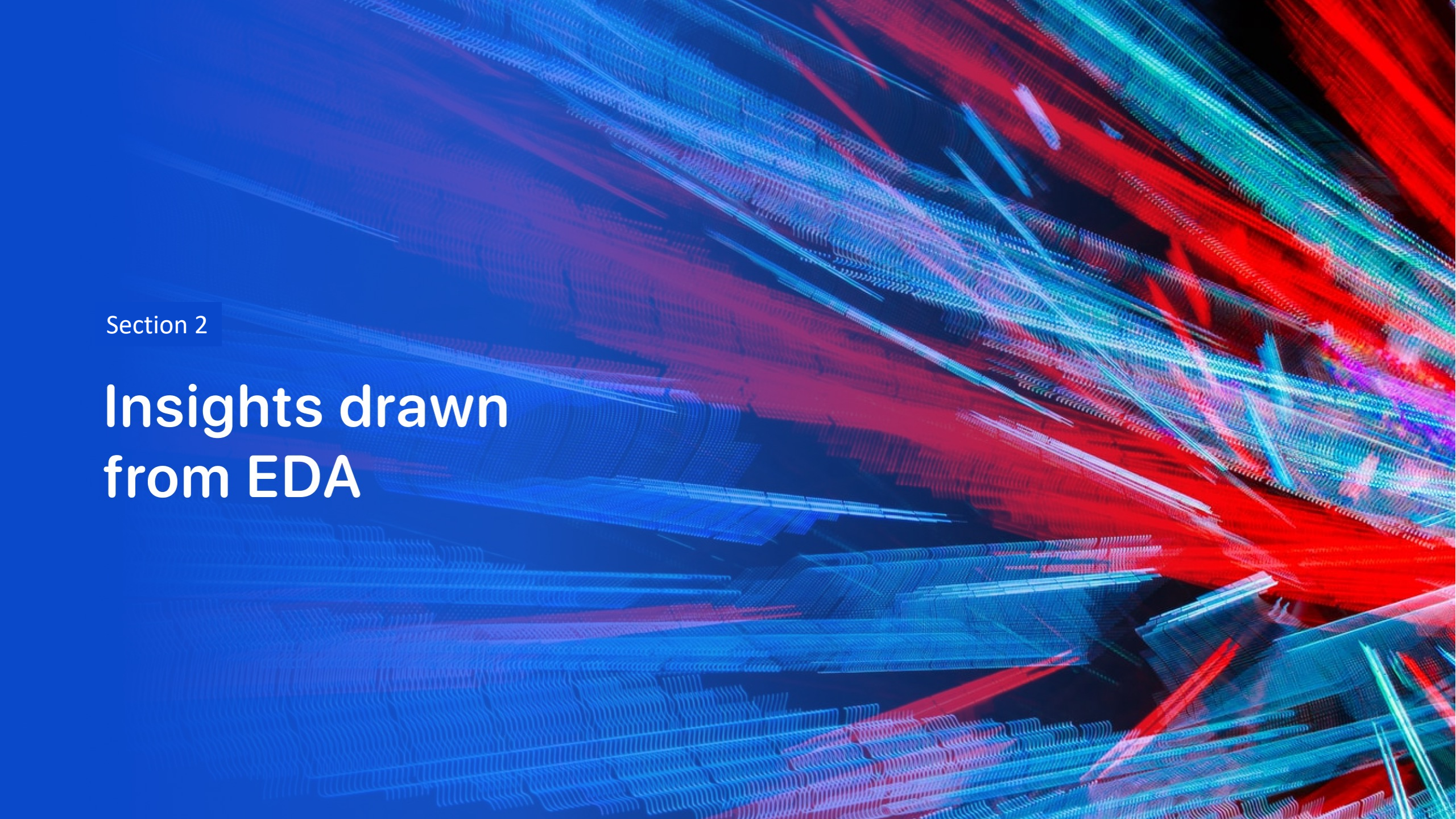
Predictive Analysis (Classification)

- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model comparison
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen (see Notebook for result)

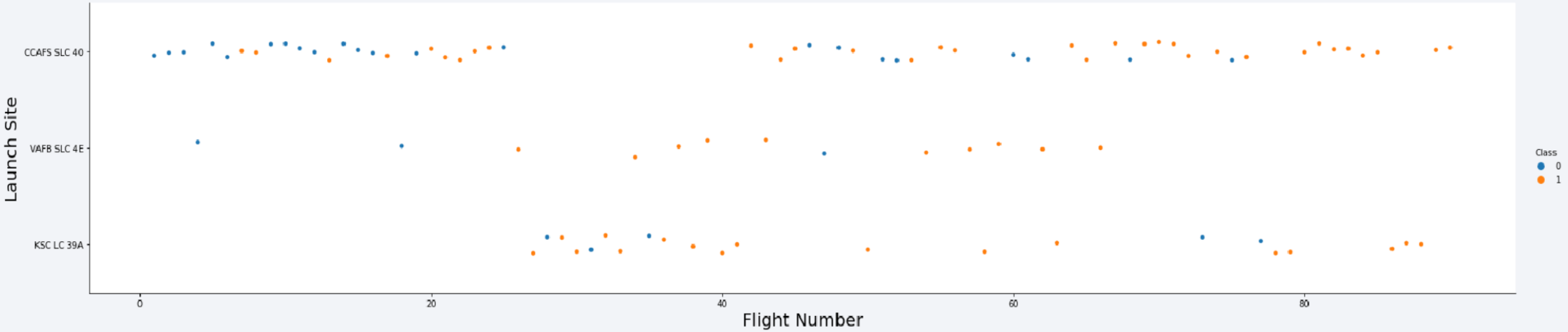
https://github.com/PDCasC/Data-Science-Project/blob/master/Machine%20Learning%20Prediction.ipynb

Results

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

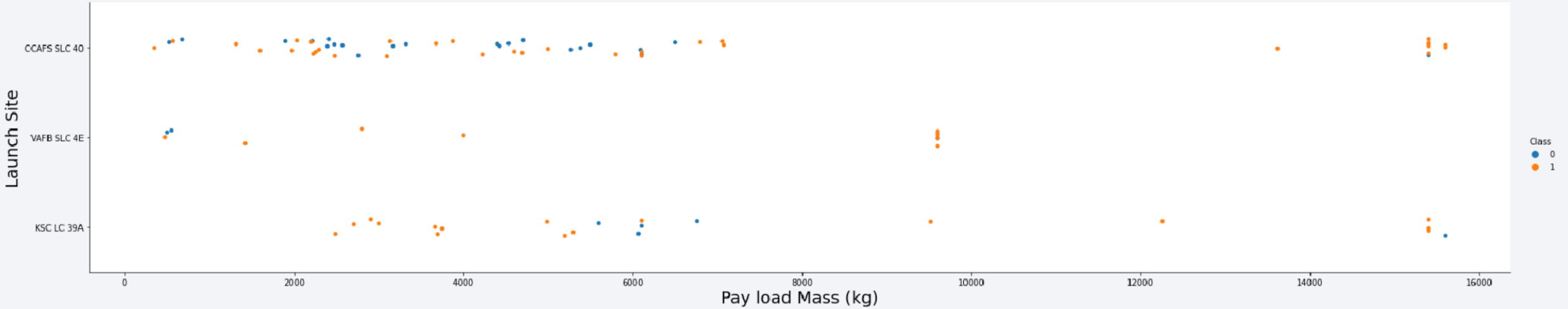

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



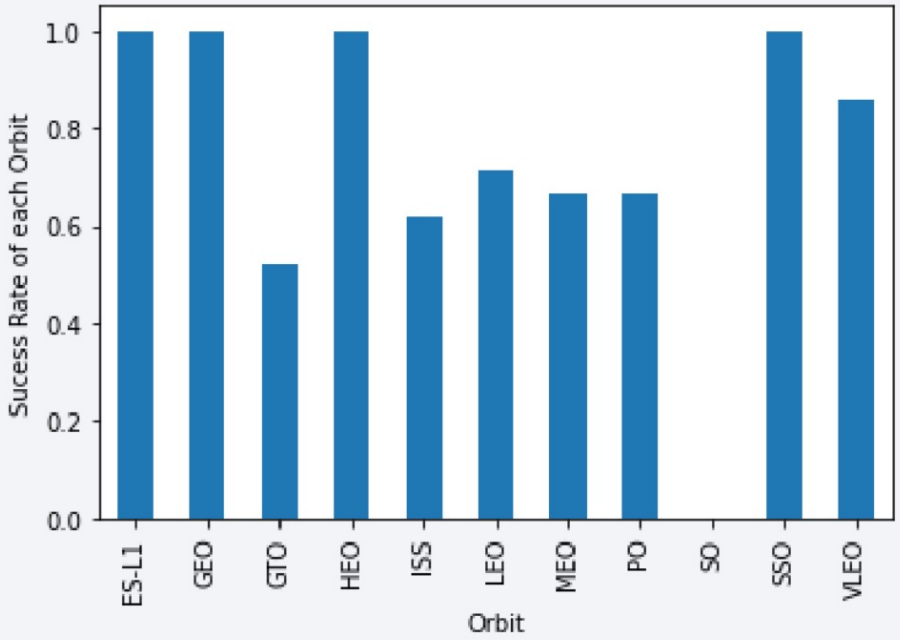
We observe that, for each site, the success rate is increasing.

Payload vs. Launch Site



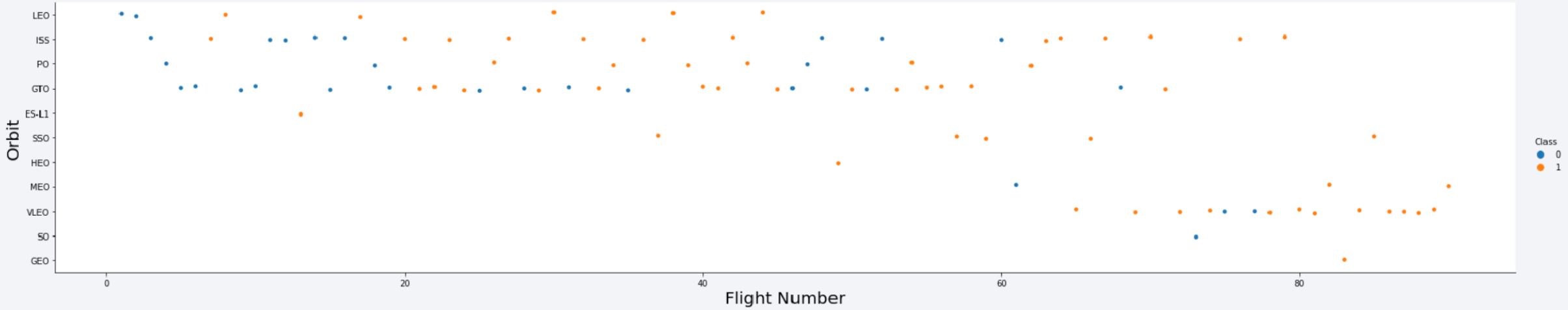
Depending on the launch site, a heavier payload may be a consideration for a successful landing. On the other hand, a too heavy payload can make a landing fail.

Success Rate vs. Orbit Type



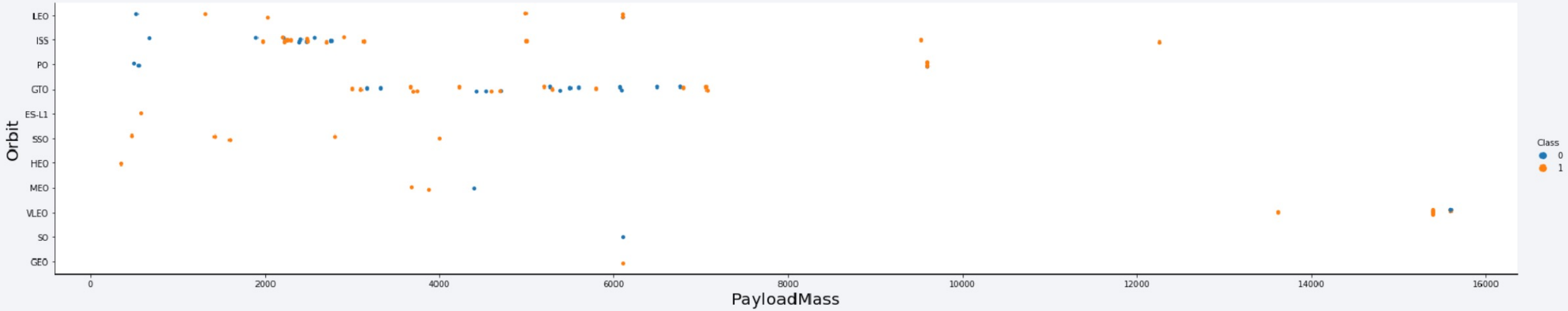
With this plot, we can see success rate for different orbit types. We note that ES-L1, GEO, HEO, SSO have the best success rate.

Flight Number vs. Orbit Type



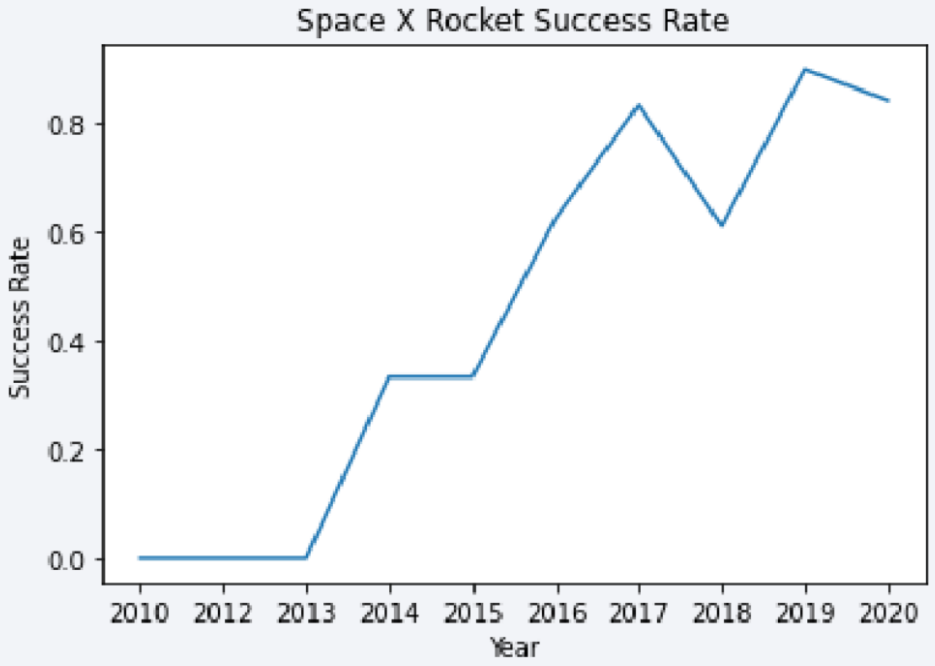
We notice that the success rate increases with the number of flights for the LEO orbit. For some orbits like GTO, there is no relation between the success rate and the number of flights. But we can suppose that the high success rate of some orbits like SSO or HEO is due to the knowledge learned during former launches for other orbits.

Payload vs. Orbit Type



The weight of the payloads can have a great influence on the success rate of the launches in certain orbits. For example, heavier payloads improve the success rate for the LEO orbit. Another finding is that decreasing the payload weight for a GTO orbit improves the success of a launch.

Launch Success Yearly Trend



Since 2013, we can see an increase in the Space X Rocket success rate.

Quick Note for EDA in SQL

- SQL in DB2 was not loading the data
- The following slides show the code that would get to the results

All Launch Site Names

```
%sql SELECT distinct launch_site from SPACEXDATASET;
```

Launch Site Names Begin with 'CCA'

```
%sql SELECT * from SPACEXDATASET WHERE launch_site like 'CCA%' LIMIT 5;
```

Total Payload Mass

```
%sql SELECT SUM(payload_mass__kg_) AS total_payload_mass FROM SPACEXDATASET WHERE customer = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(payload_mass__kg_) AS average_payload_mass FROM SPACEXDATASET WHERE booster_version like '%F9 v1.1%';
```

First Successful Ground Landing Date

```
%sql SELECT min(date) as first_successful_landing FROM SPACEXDATASET WHERE landing__outcome = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT booster_version FROM SPACEXDATASET WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ between 4000 and 6000;
```

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, COUNT(*) as total_number FROM SPACEXDATASET GROUP BY mission_outcome;
```

Boosters Carried Maximum Payload

```
%sql SELECT booster_version FROM SPACEXDATASET where payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET);
```

2015 Launch Records

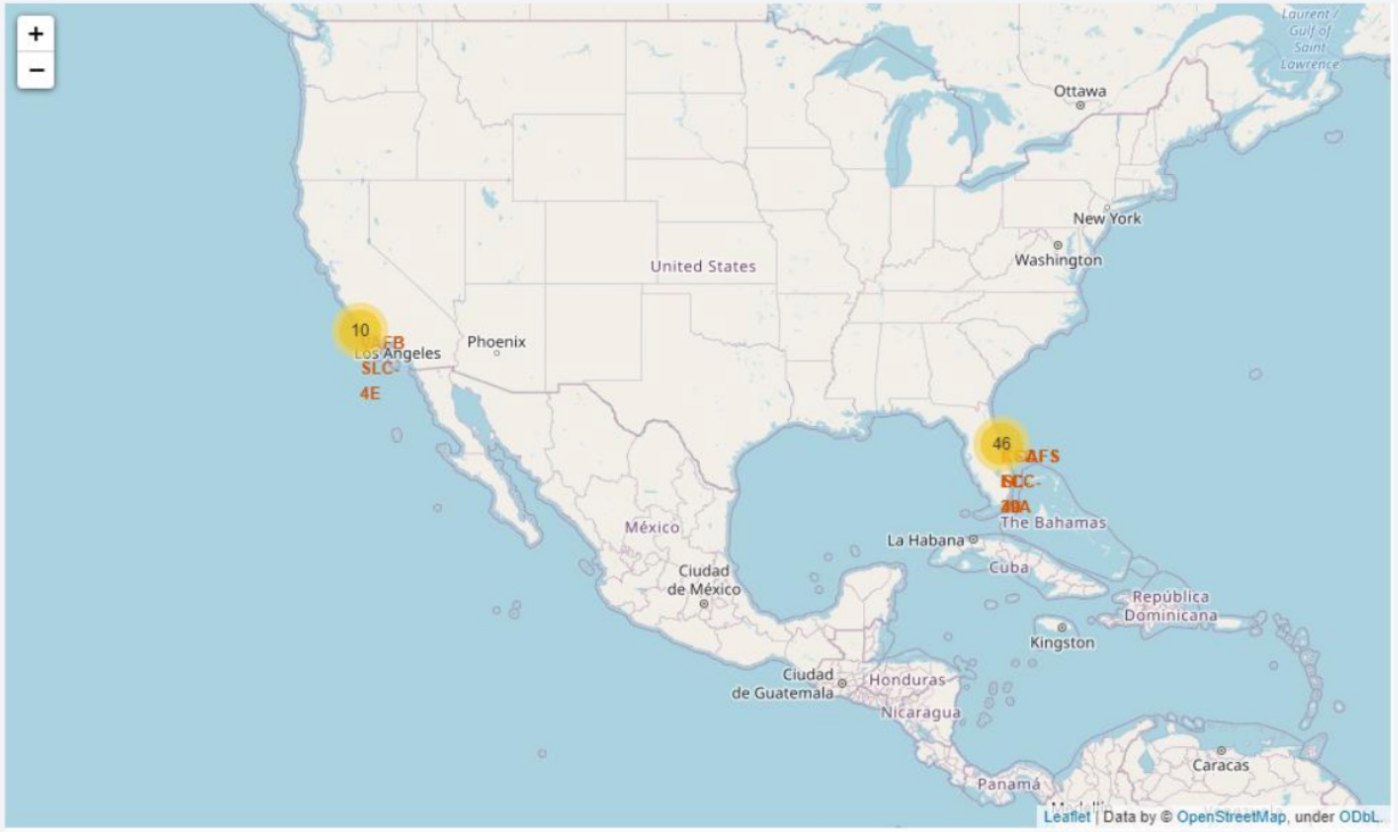
```
%%sql SELECT monthname(date) AS month, date, booster_version, launch_site, landing__outcome FROM SPACEXDATASET
      WHERE landing__outcome = 'Failure (drone ship)' AND year(date)=2015;
```

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

```
%%sql SELECT landing__outcome, count(*) AS count_outcomes FROM SPACEXDATASET
      WHERE date between '2010-06-04' AND '2017-03-20'
      GROUP by landing__outcome
      ORDER by count_outcomes desc;
```

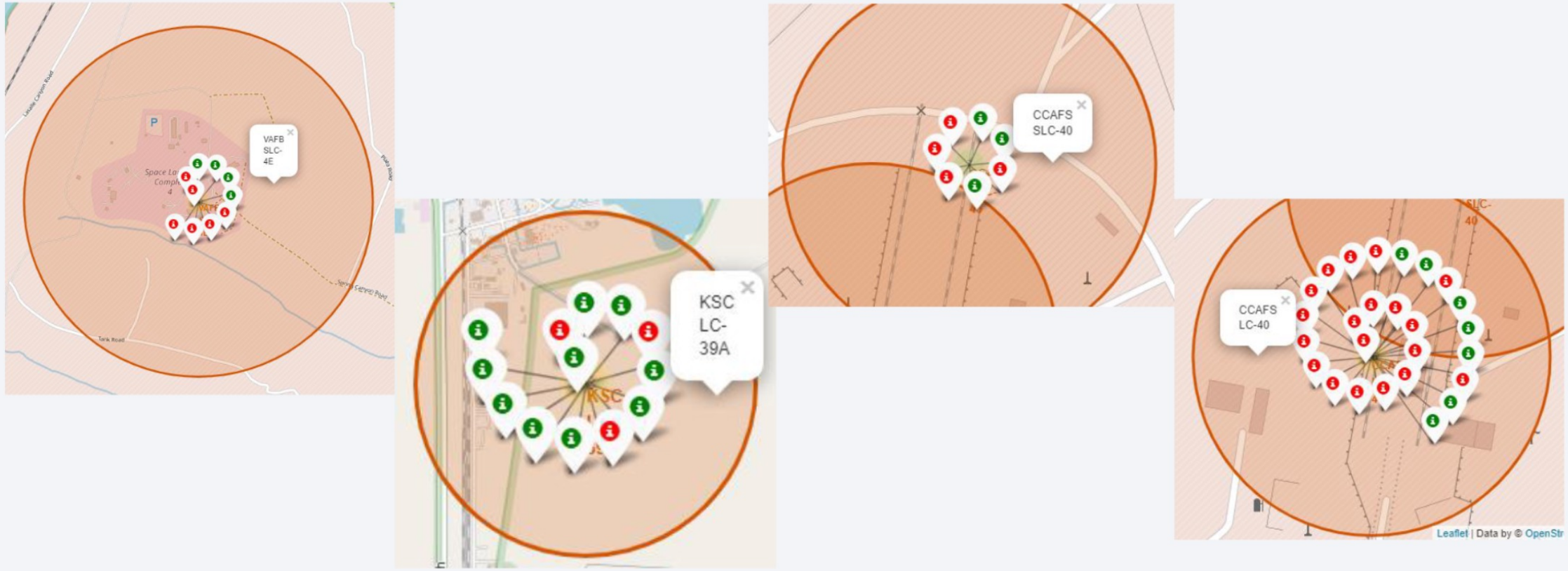

Section 3
Launch Sites
Proximities Analysis

Folium Map - Ground Stations



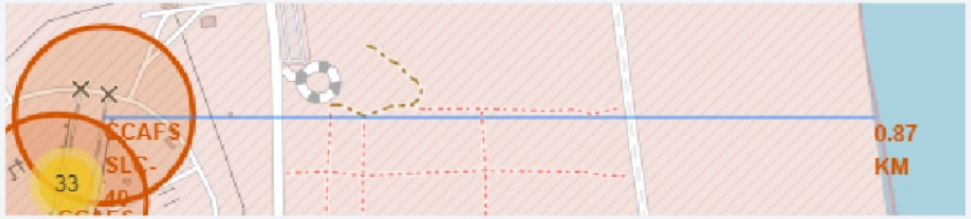
We see that Space X launch sites are located on the coast of the United States

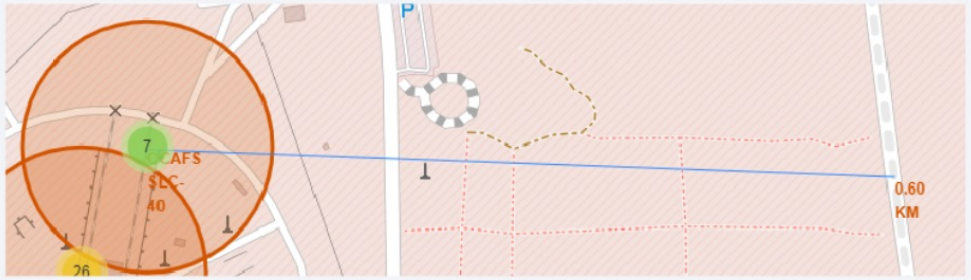
Folium Map – Color Labeled Markers



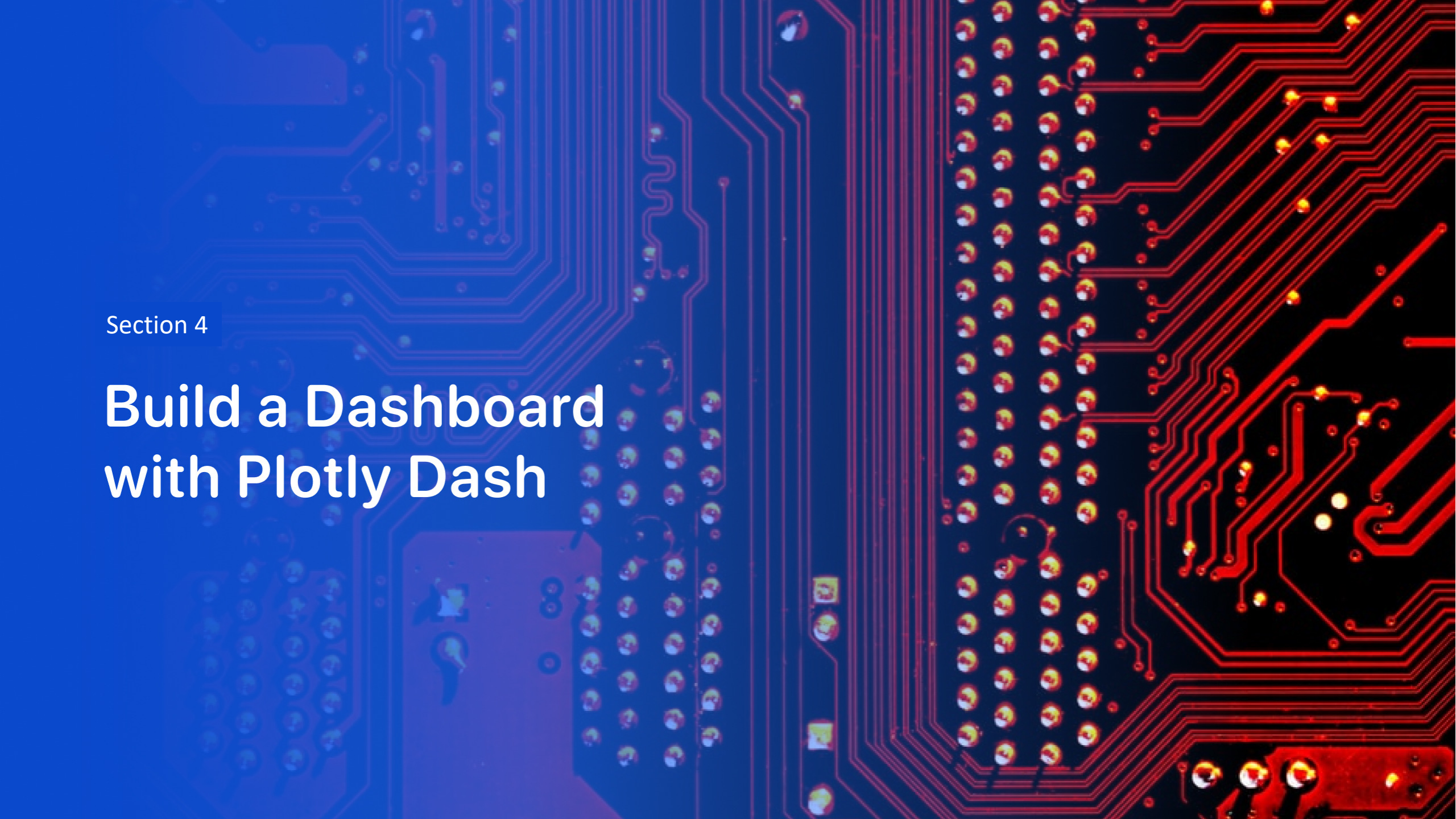
The green marker represents successful launches. Red marker represents unsuccessful launches. The KSC LC-39A has the higher success rate.

Folium Map – Distances between CCAFS SLC-40 and its proximities

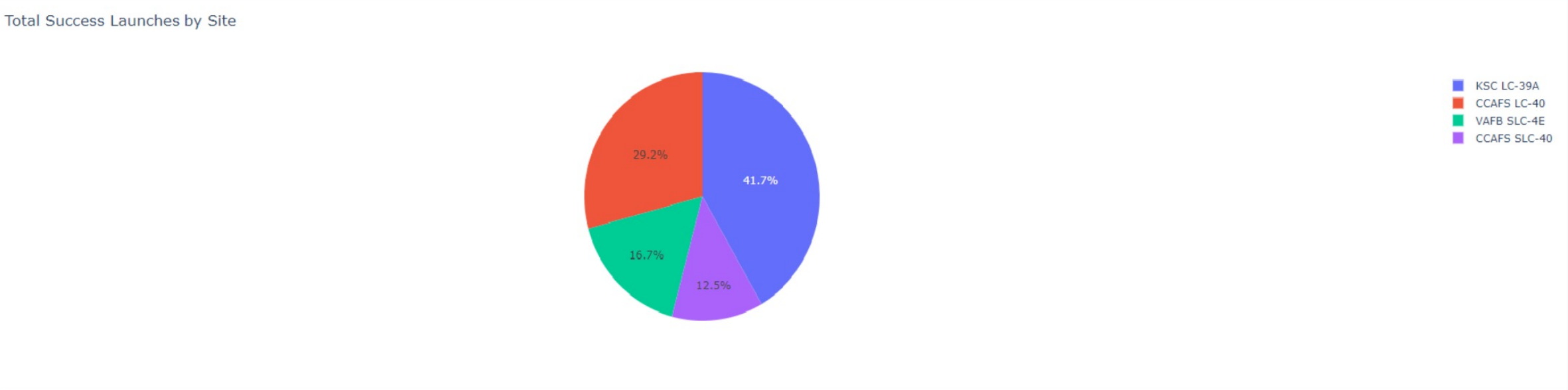




Is CCAFS SLC-40 in close proximity to railways ? Yes
Is CCAFS SLC-40 in close proximity to highways ? Yes
Is CCAFS SLC-40 in close proximity to coastline ? Yes
Do CCAFS SLC-40 keeps certain distance away from cities ? No

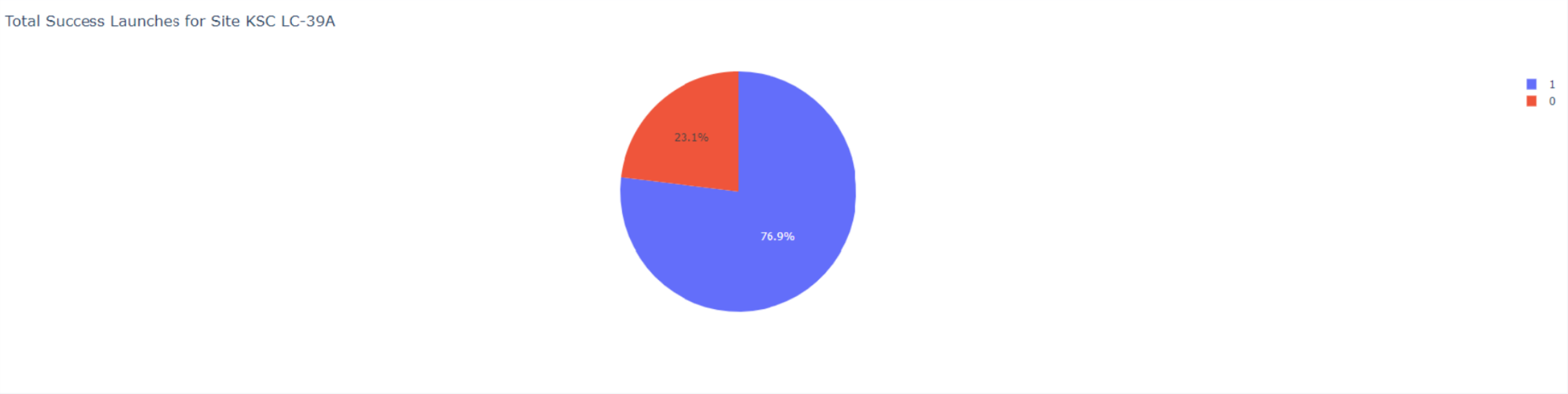

Section 4
Build a Dashboard with Plotly Dash

Dashboard – Total Success by Site



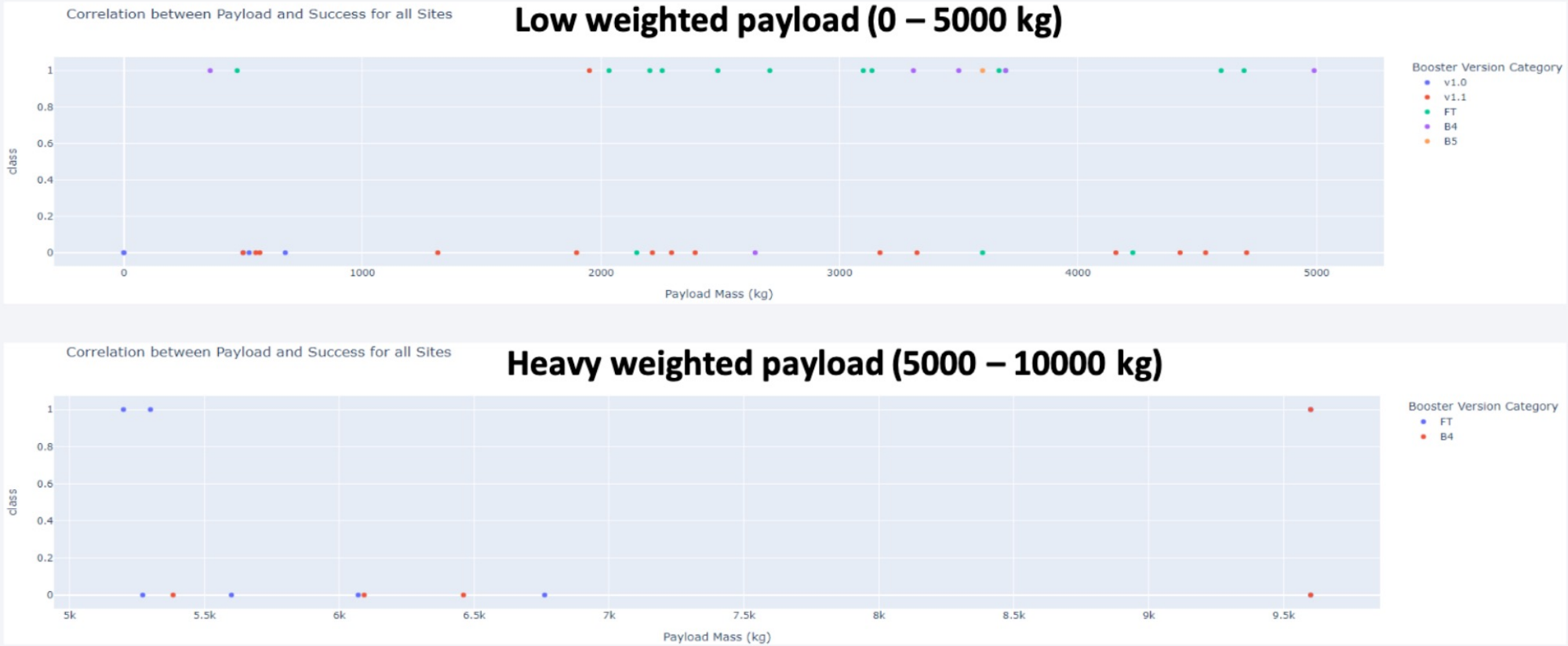
We see that KSC LC-39A has the best success rate of launches.

Dashboard – Total success launches for site KSC



We see that KSC LC-39A has achieved a 76.9% success rate while getting a 23.1% failure rate.

Dashboard – Payload mass vs Outcome for all sites with different payload mass selected



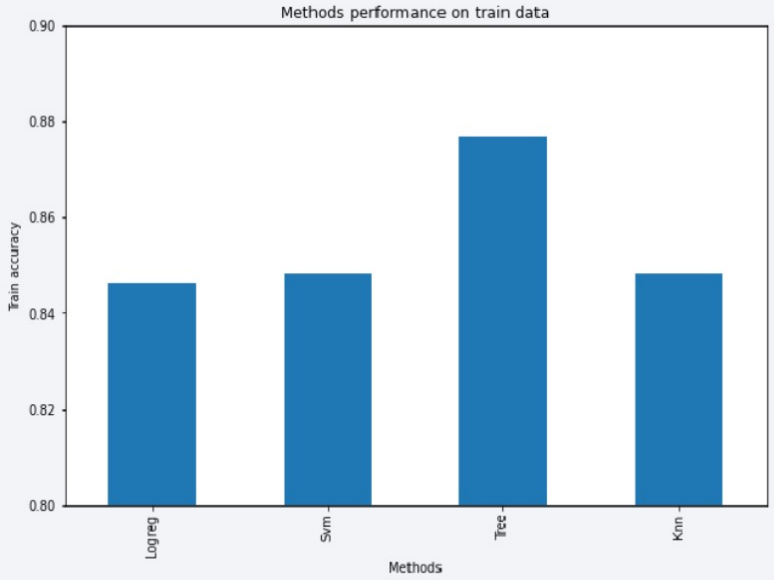
Low weighted payloads have a better success rate than the heavy weighted payloads.

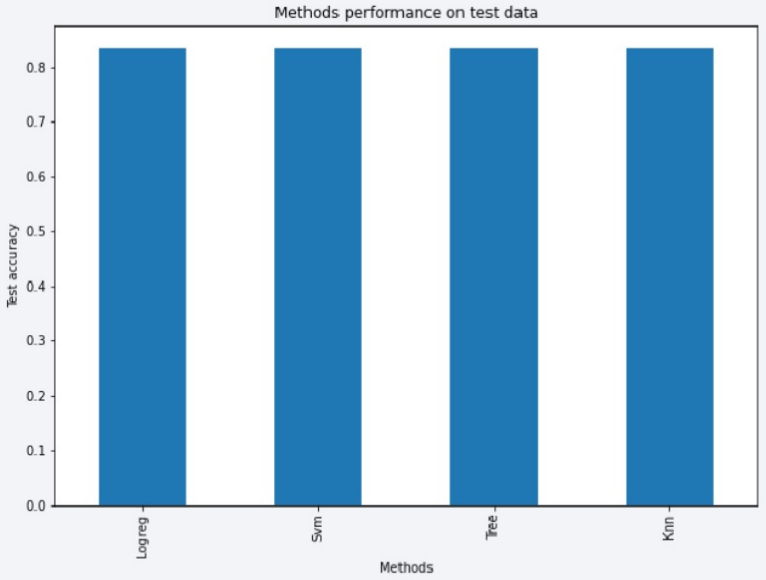
Section 5

Predictive Analysis (Classification)

Classification Accuracy

	Accuracy Train	Accuracy Test
Tree	0.876786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333





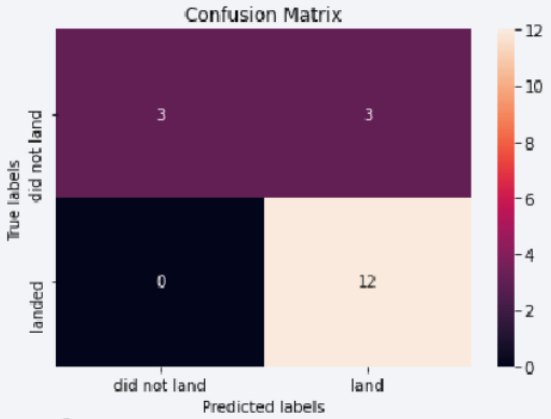
For accuracy test, all methods performed similar. **We could get more test data to decide between them. But if we really need to choose one right now, we would take the decision tree.**

Decision tree best parameters

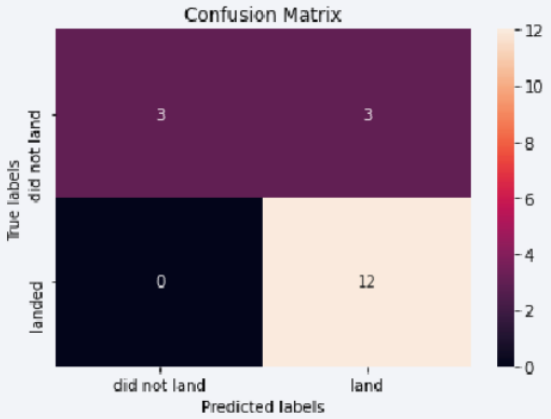
```
tuned hyperparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf':
4, 'min_samples_split': 2, 'splitter': 'random'}
```

Confusion Matrix

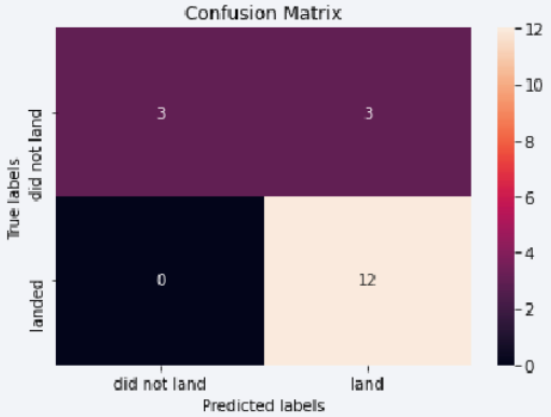
Logistic regression



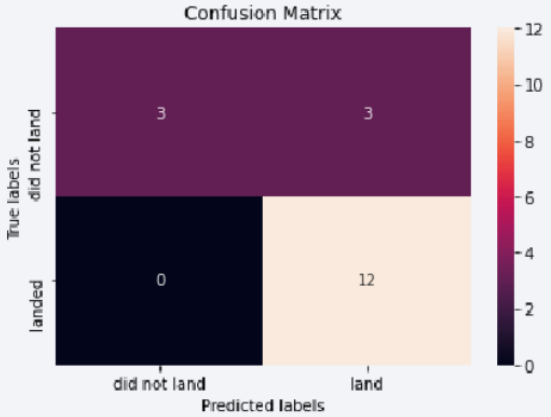
Decision Tree



kNN



SVM



As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models are false positives.

		Actual values 1	Actual values 0
Predicted values	1	TP	FP
	0	FN	TN

Conclusions

- Different data sources were analyzed, refining conclusions along the process.
- The best launch is KSC LC-39A.
- Decision Tree Classifier can be used to predict successful landings and increase profits.
- Low weighted payloads perform better than the heavier payloads.

Thank you!