IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roberto García Guzmán
2022-09-17

Outline

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

Executive Summary

- The data for this study was obtained from two sources: SpaceX API and Web Scrapping. This data was filtered and processed with different machine learning algorithms in order to obtain the best prediction on weather the Falcon 9 first stage will land or not: Logistic Regression, SVM, Decision Trees, KNN.
- We found that using Decision Trees is the best method to predict the landing of Falcon 9 first stage, getting a 83% accuracy.

Introduction

- SpaceX has significantly reduced their costs by reusing their rockets.
- As competitors, being able to predict whether the rocket will land or not will help us to offer more competitive options.
- The aim of this project, was to find a way to predict if the rocket will land or not
- Also, we tried to find which variables (payload, launch site, etc.) are the most important to determine the success of a landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Wikipedia
- Perform data wrangling
  - Label standardization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logreg, SVM, Decision Trees, KNN trained and tested on different datasets.

Data Collection

Two main sources of information were used:

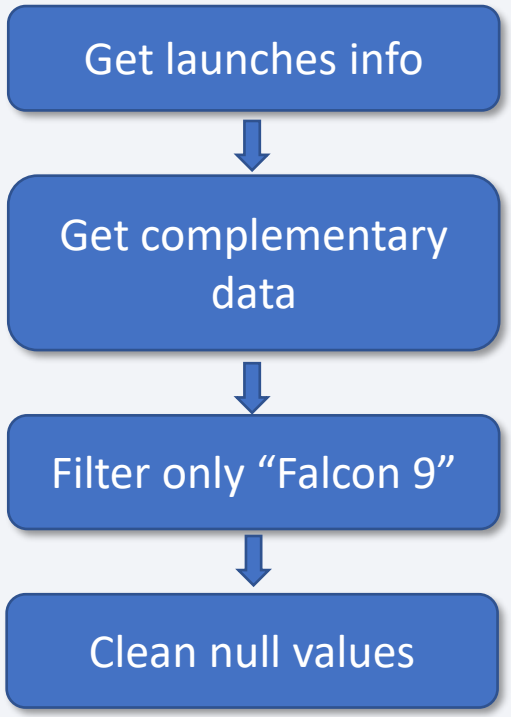
- SpaceX API: Provided by SpaceX, contains information about the last launches the company has made. (https://api.spacexdata.com/v4/rockets/)

- Wikipedia: It has a short history of launches and their outcomes; we collected some data using Web Scraping. (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- Provides different methods to obtain data from launches, rockets and missions.
- Data was collected and cleaned to store only needed values.

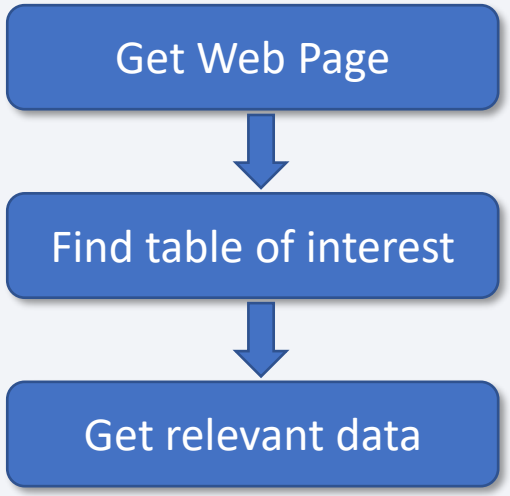
- Source code: https://github.com/Rivert97/Applied_Data_Science_Capstone/blob/main/Data%20Collection%20API.ipynb



Data Collection - Scraping

- Wikipedia provides history on Falcon 9 launches.
- Data cleaning must be done since it's meant to be read by humans, not machines.

- Source code: https://github.com/Rivert97/Applied_Data_Science_Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Basic statistics were obtained (Number of launch sites, Orbits, Outcomes).
- Since the outcomes were not suitable for processing, all outcomes were categorized in Successful or Failure, regardless of the landing place.

- Source code: https://github.com/Rivert97/Applied_Data_Science_Capstone/blob/main/EDA.ipynb

EDA with Data Visualization

- Scatter point charts were used to find relationships between the landing outcome and the following:
  - Flight number vs Launch Site
  - Payload vs Launch Site
  - Orbit
  - Flight number vs Orbit
  - Payload vs Orbit
  - Year trend
- Source code: https://github.com/Rivert97/Applied_Data_Science_Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

SQL queries were performed to get the following data:

- Unique launch sites.
- Launch site starting with 'ACC'.
- Total payload launched from NASA (CRS).
- Average payload carries by booster F9 v1.1
- First successful landing in ground pad.
- Boosters successfully landed on drone ships with payload between 4000 and 6000.
- Number of successful and failure missions.
- Boosters who carries max payload.
- Failed landings in drone ships in 2015.
- Outcome summary between 2010-06-04 and 2017-03-20.
- Source code: https://github.com/Rivert97/Applied_Data_Science_Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, Circles and Clusters were used in the maps.
  - Markers were used to indicate the launch sites
  - Circles indicated special interest areas.
  - Clusters grouped several relevant events occurred in the map.
  - Lines were used to highlight distance between points.

- Source code: https://github.com/Rivert97/Applied_Data_Science_Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Two graphs were used to show the information:
  - Pie chart: To display the landing outcome on each launch site.
  - Scatter point chart: To show the outcome for different payloads and boosters.
- Pie charts has a control to select different launch sites.
- Scatter plot has a selector to change the range of payloads of interest.

- Source code: https://github.com/Rivert97/Applied_Data_Science_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

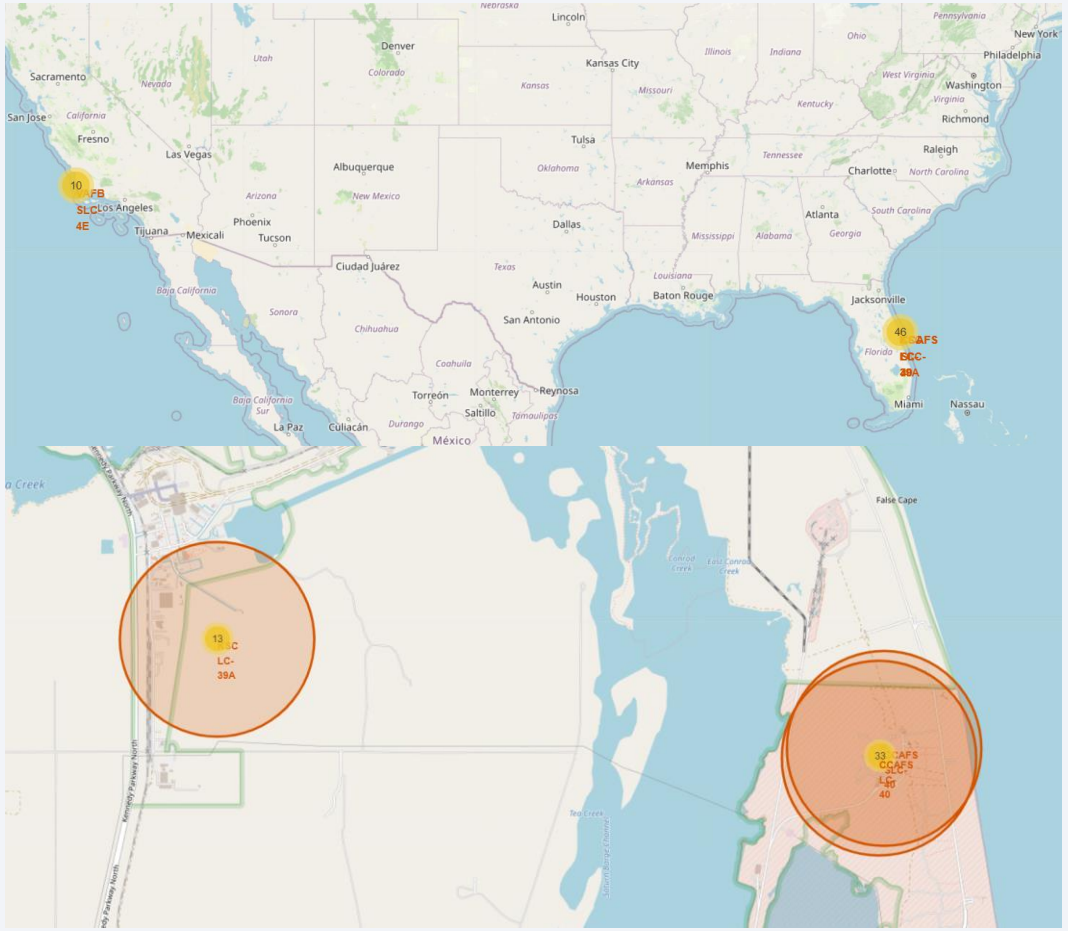
- Once the data was filtered the following steps were made for each classification model:
  - Normalization.
  - Split dataset into train and test.
  - Train the model with different parameters (Grid search) using the training data.
  - Evaluate model with test data.
- The previous process was made for the following models:
  - Logistic regression, Support Vector Machines, Decision Trees, K-Nearest Neighbors.
- Source code: https://github.com/Rivert97/Applied_Data_Science_Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results:
  - 4 different launch sites are used by SpaceX.
  - The most frequent orbit is GTO.
  - The most successful landings are in drone ships.
  - SpaceX has a success landing rate of 66%.
  - VAFB-SLC launch site has no rockets launched for payload greater than 10000.
  - When rockets are launched to the following orbits the success rate is 100%: ES-L1, GEO, HEO, SSO.
  - Success rate since 2013 increased constantly till 2020.

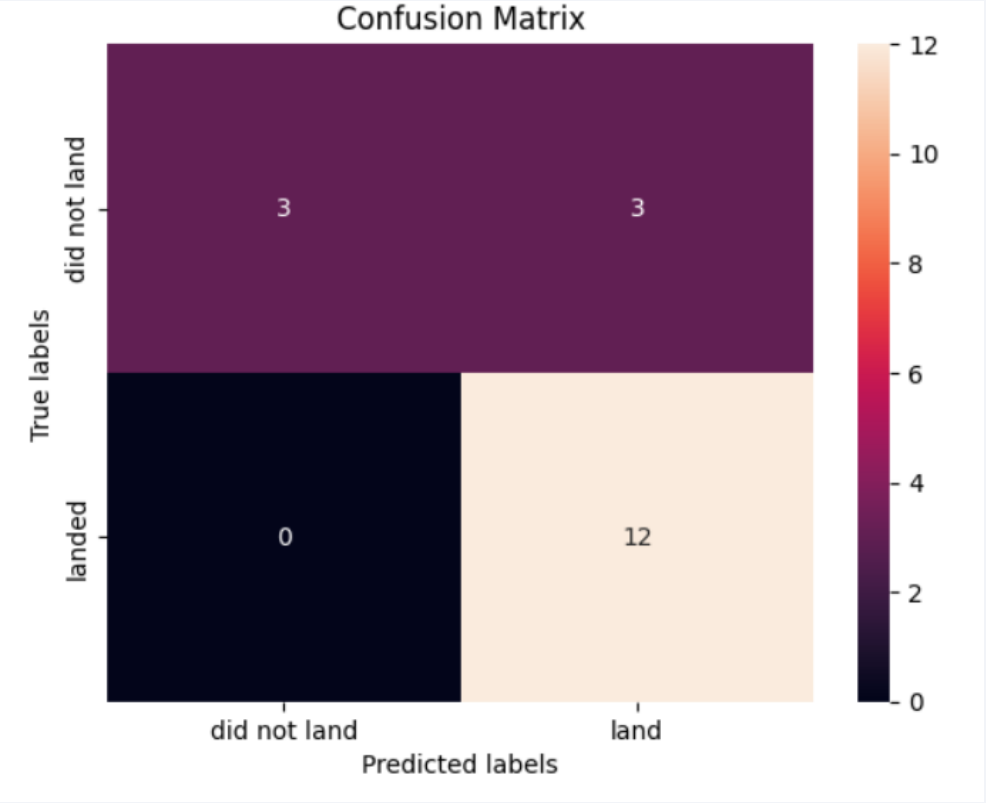
Results (continuation)

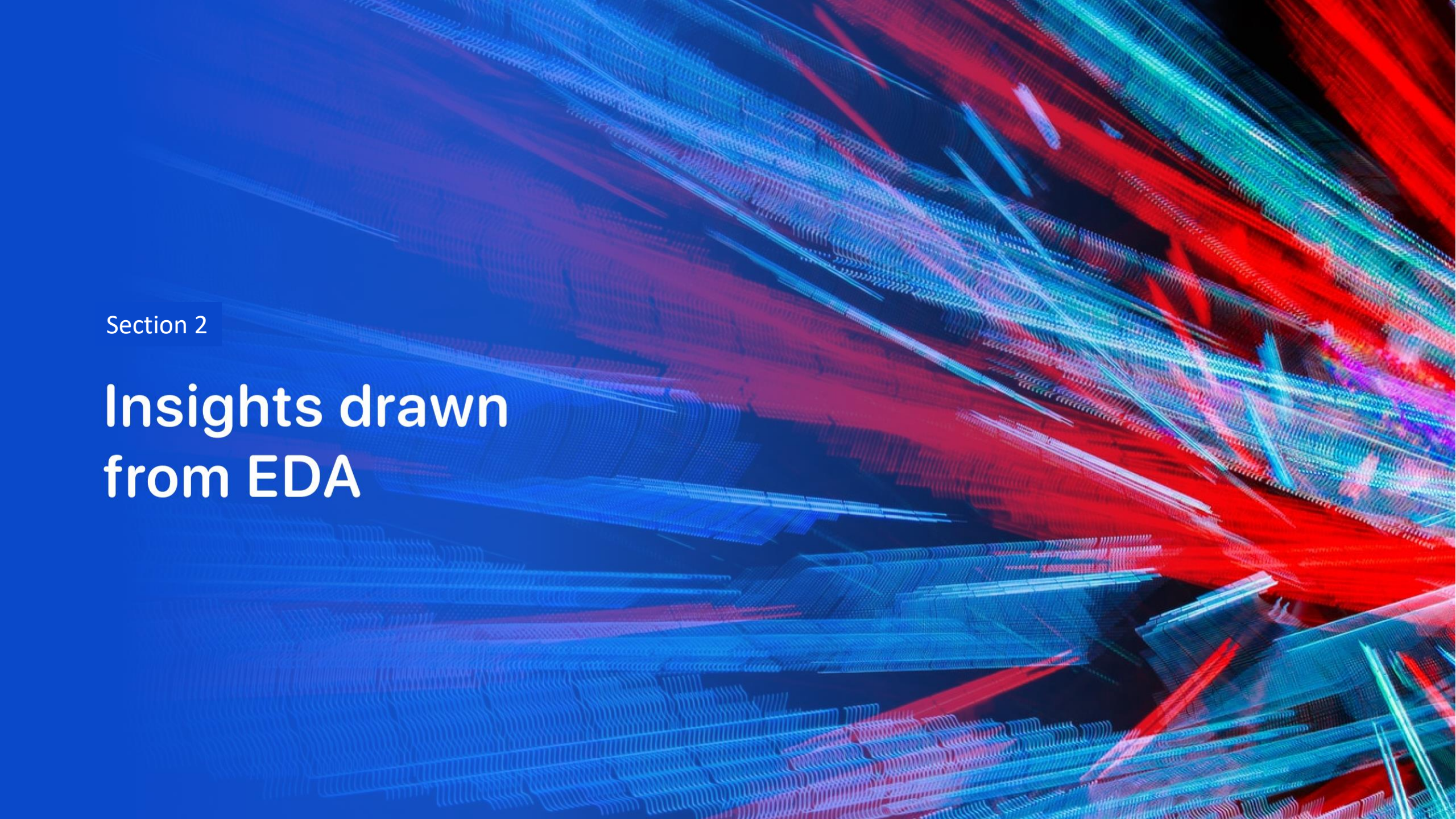
- Most launches are made from the east coast.
- Two launch sites are located close to each other.



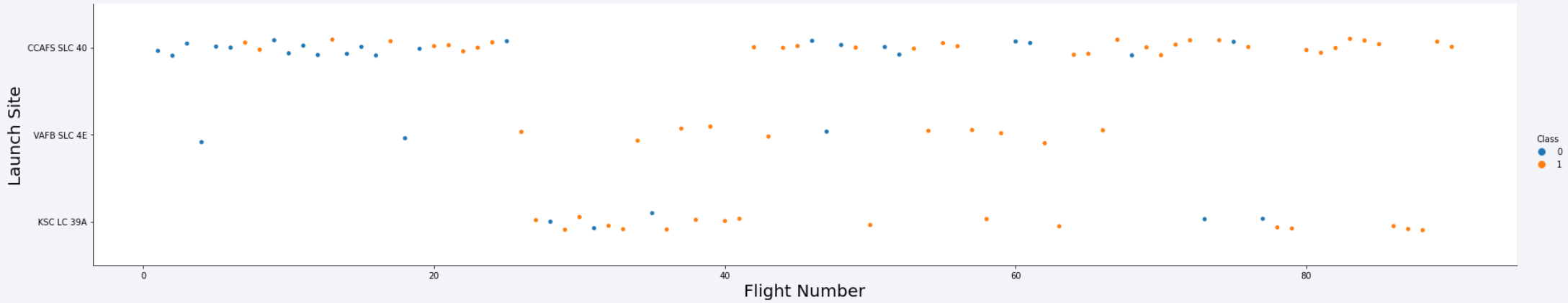
Results

- The analysis showed that Decision Tree is the model with highest accuracy with 83% accuracy.



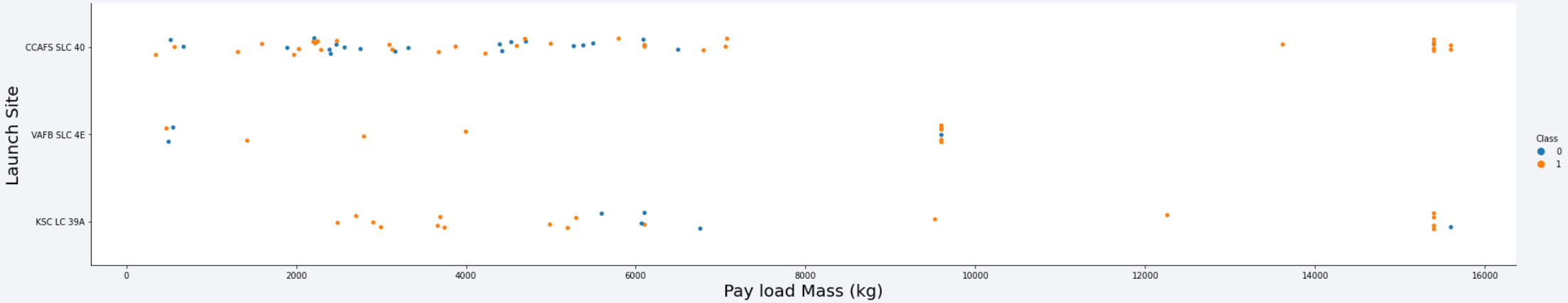

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SUC 40 was the first launch site used.
- An improvement on success can be seen through time.
- Last 13 flights were successfully landed

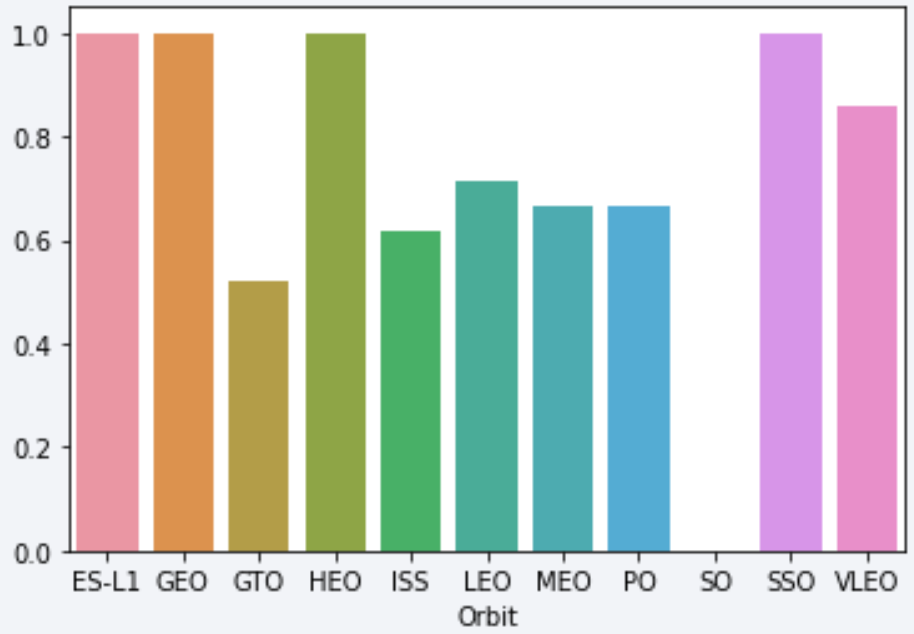
Payload vs. Launch Site



- There is almost 100% success rate for payload higher tan 8000.

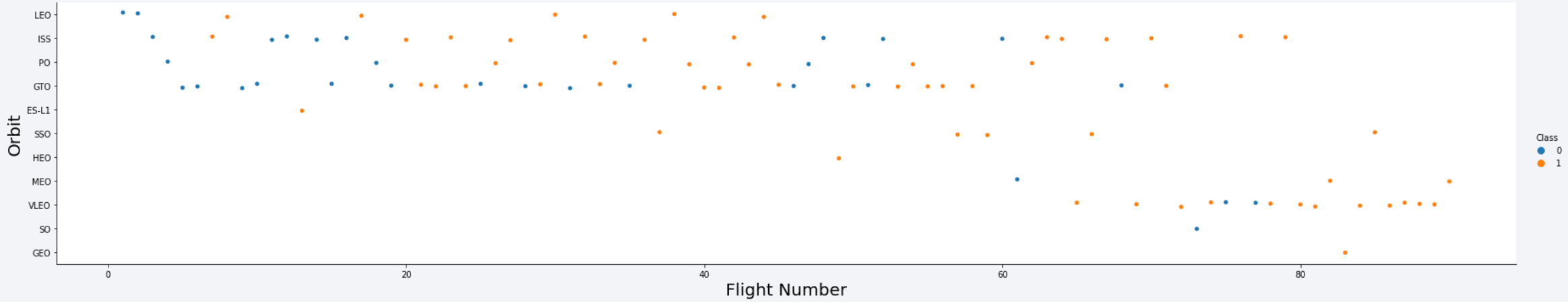
Success Rate vs. Orbit Type

- The orbits with 100% success is are:
  - ES-L1
  - GEO
  - HEO
  - SSO



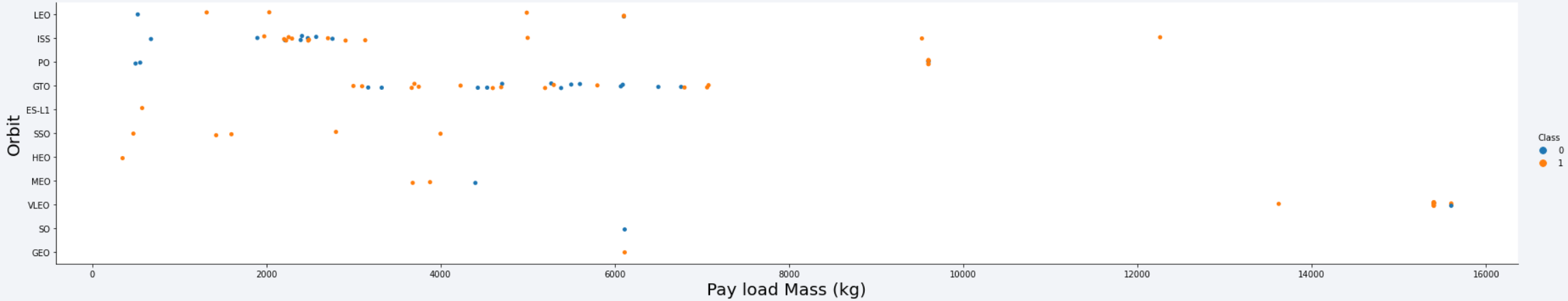
Flight Number vs. Orbit Type

- The most recent launches are for Low Orbits and have been successful



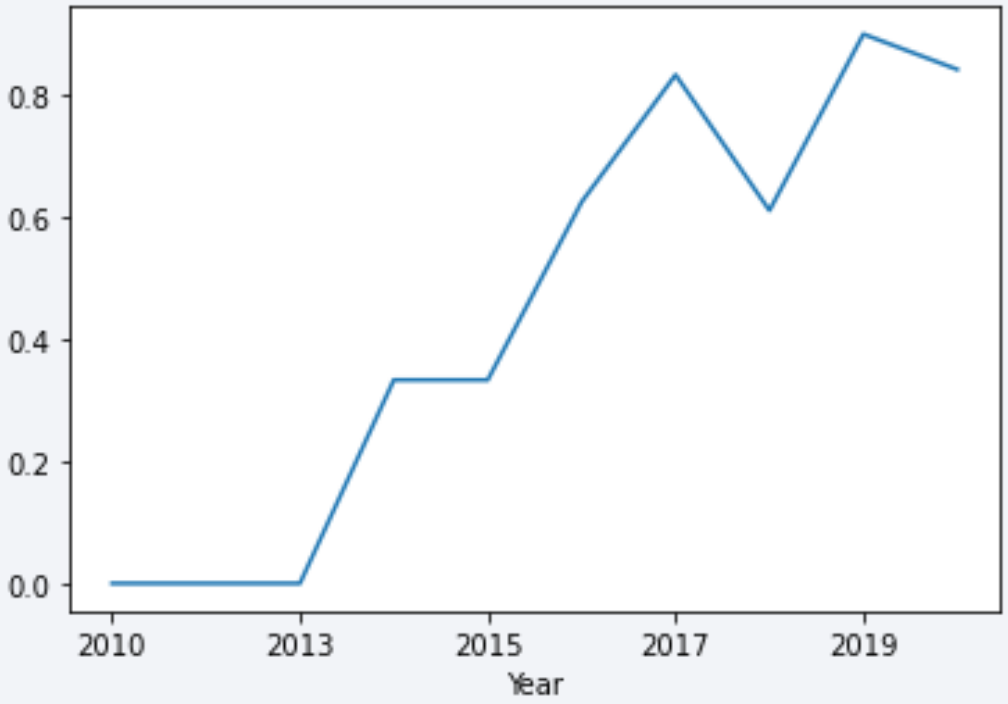
Payload vs. Orbit Type

- The highest payloads are deployed into the lowest orbits.



Launch Success Yearly Trend

- From 2013 the success rate has increased and reached almost 90% in 2019.



All Launch Site Names

- There are 4 unique launch sites in the dataset.

```
%sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

- The first launches from ‘CCA’ were failures

```
%sql SELECT * FROM SPACEXTBL 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 carries by the boosters is 45,496.

```
%sql SELECT SUM(payload_mass__kg_) AS total_payload FROM SPACEXTBL WHERE customer = 'NASA (CRS)'
```

total_payload
45596

Average Payload Mass by F9 v1.1

- The average payload mass for F9 v1.1 booster is 2,928.

```
%sql SELECT AVG(payload_mass__kg_) AS average_payload FROM SPACEXTBL WHERE booster_version = 'F9 v1.1'
```

average_payload
2928

First Successful Ground Landing Date

- The first successful landing in a Ground pad was on 2015-12-22.

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE landing__outcome = 'Success (ground pad)'
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters that have successfully landed on those conditions are:
  - F9 FT B1021.2
  - F9 FT B1031.2
  - F9 FT B1022
  - F9 FT B1026

```
%sql SELECT UNIQUE(booster_version) FROM SPACEXTBL WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_
```

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

Total Number of Successful and Failure Mission Outcomes

- Space X has a 99% mission success rate. Impresive.

```
%sql SELECT COUNT(*) AS number FROM SPACEXTBL WHERE mission_outcome = 'Success' UNION (SELECT COUNT(*) AS number FROM SPACEXTBL WHERE mission_outcome
```

number
1
99

Boosters Carried Maximum Payload

- The boosters that can carry the maximum payload are listed below.

```
%sql SELECT UNIQUE(booster_version) FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- The boosters that failed to land in 2015 are listed below

```
%sql SELECT booster_version, launch_site FROM SPACEXTBL WHERE landing__outcome like 'Failure%' AND DATE LIKE '2015%'
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- There is a high percentage of no attempst.

```
%sql SELECT landing__outcome, COUNT(*) AS number FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY nu
```

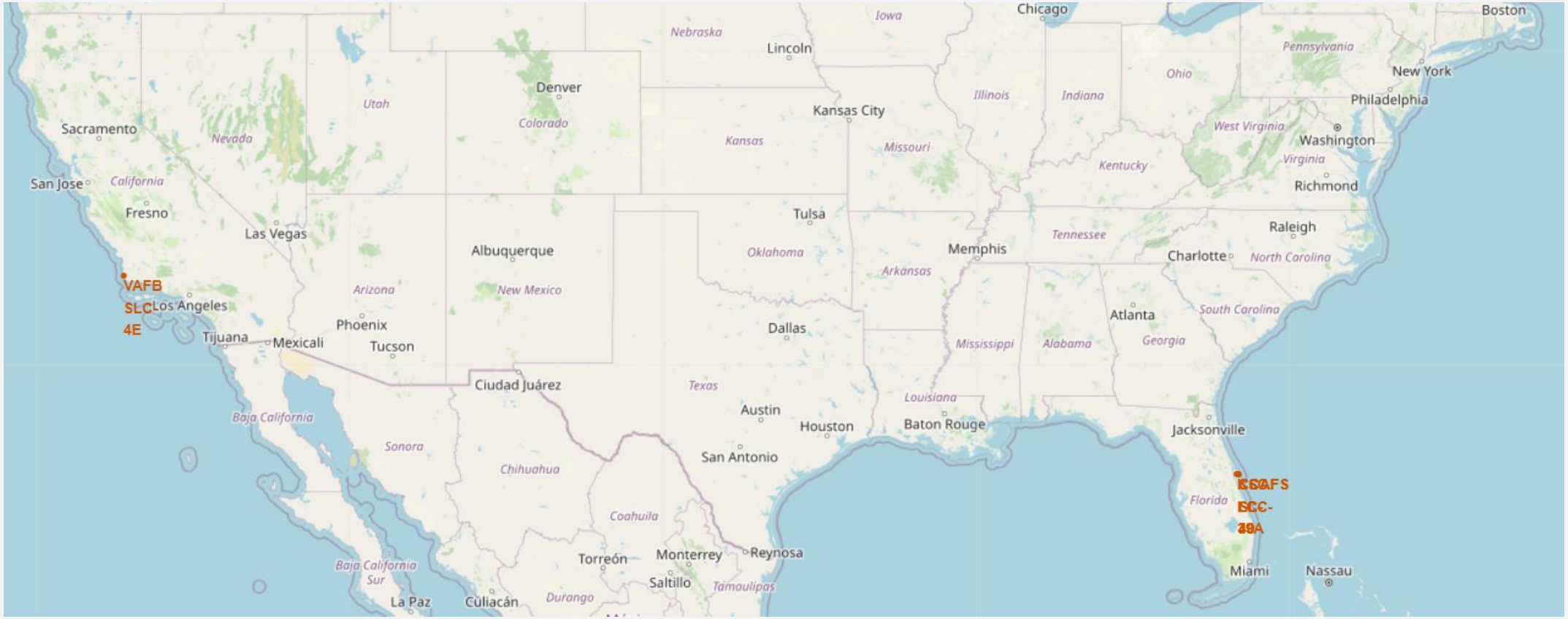
landing__outcome	number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

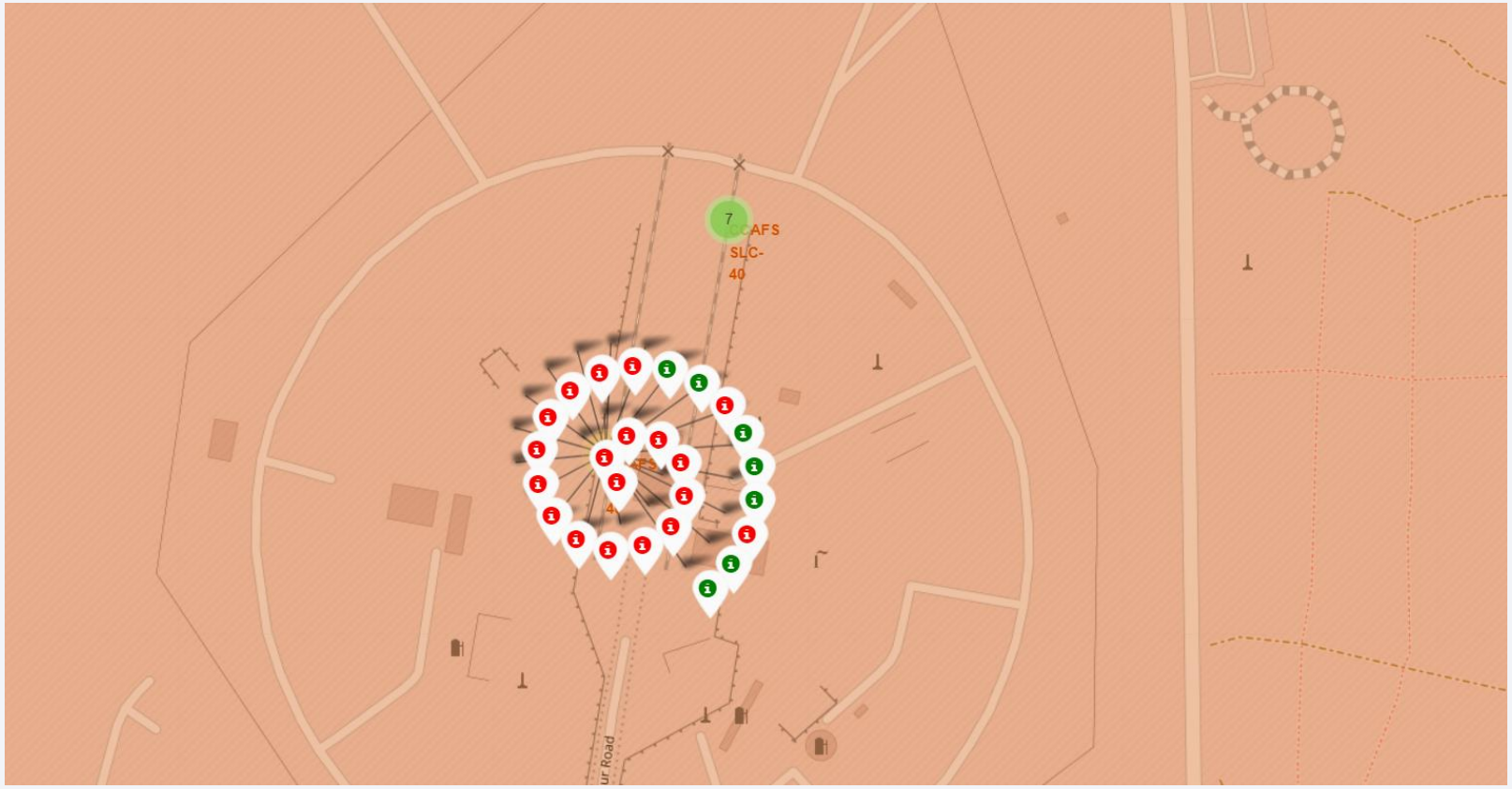
Launch sites map

- SpaceX has two regions where they perform their launches, each on each US coast.



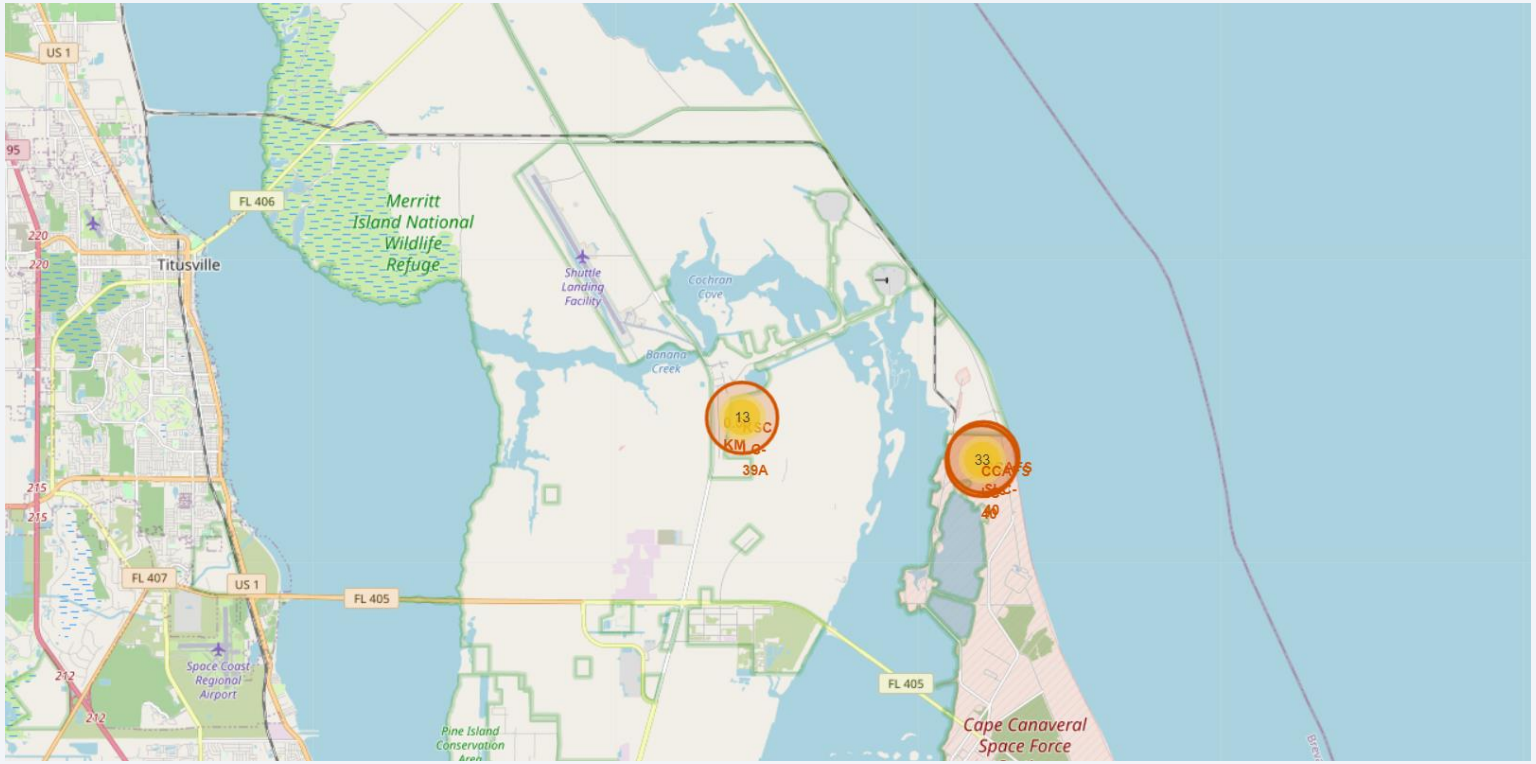
Success/Failure launch sites map

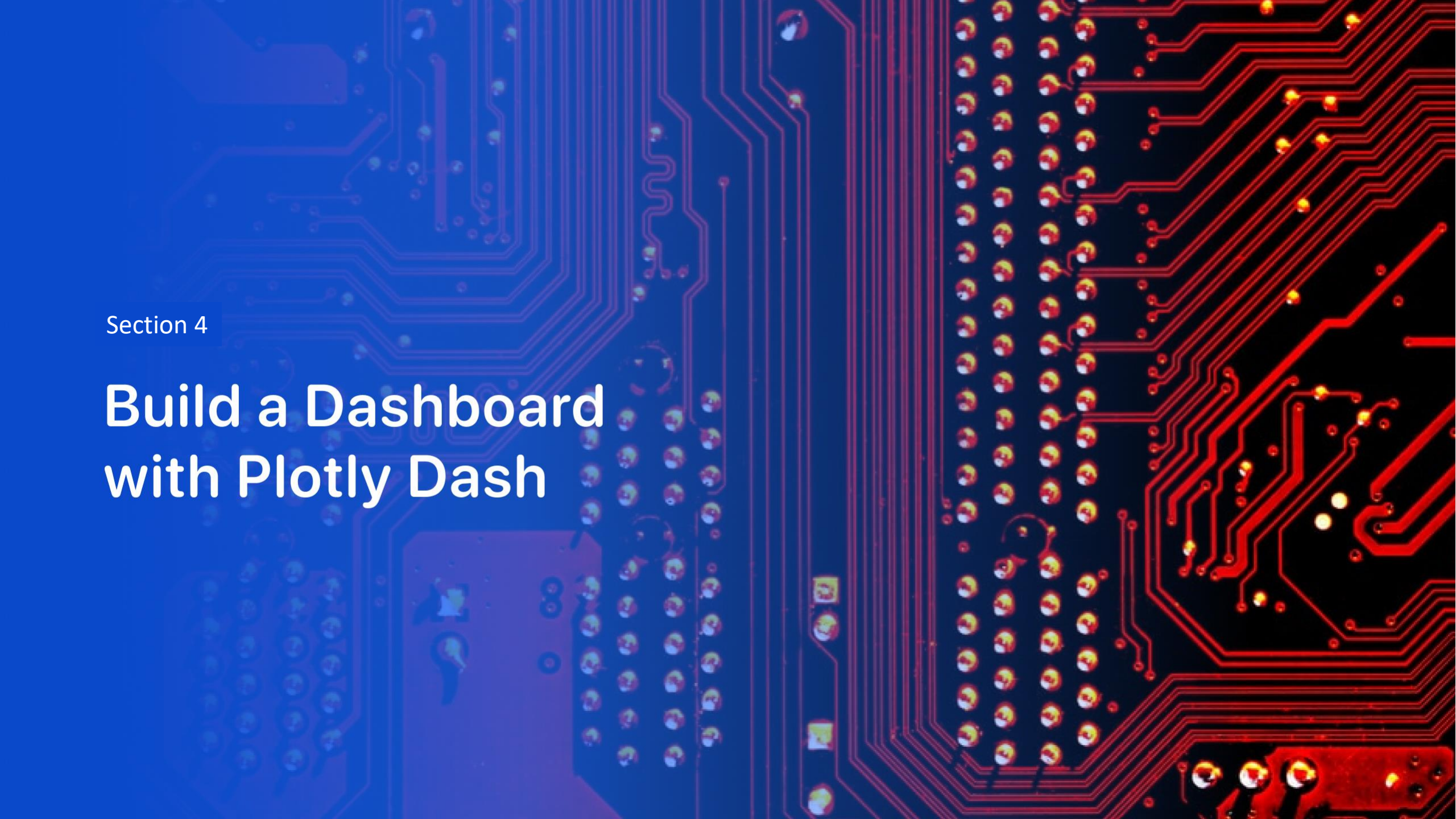
- In green are showed the launch sites of the successful landings



Roads near launch sites

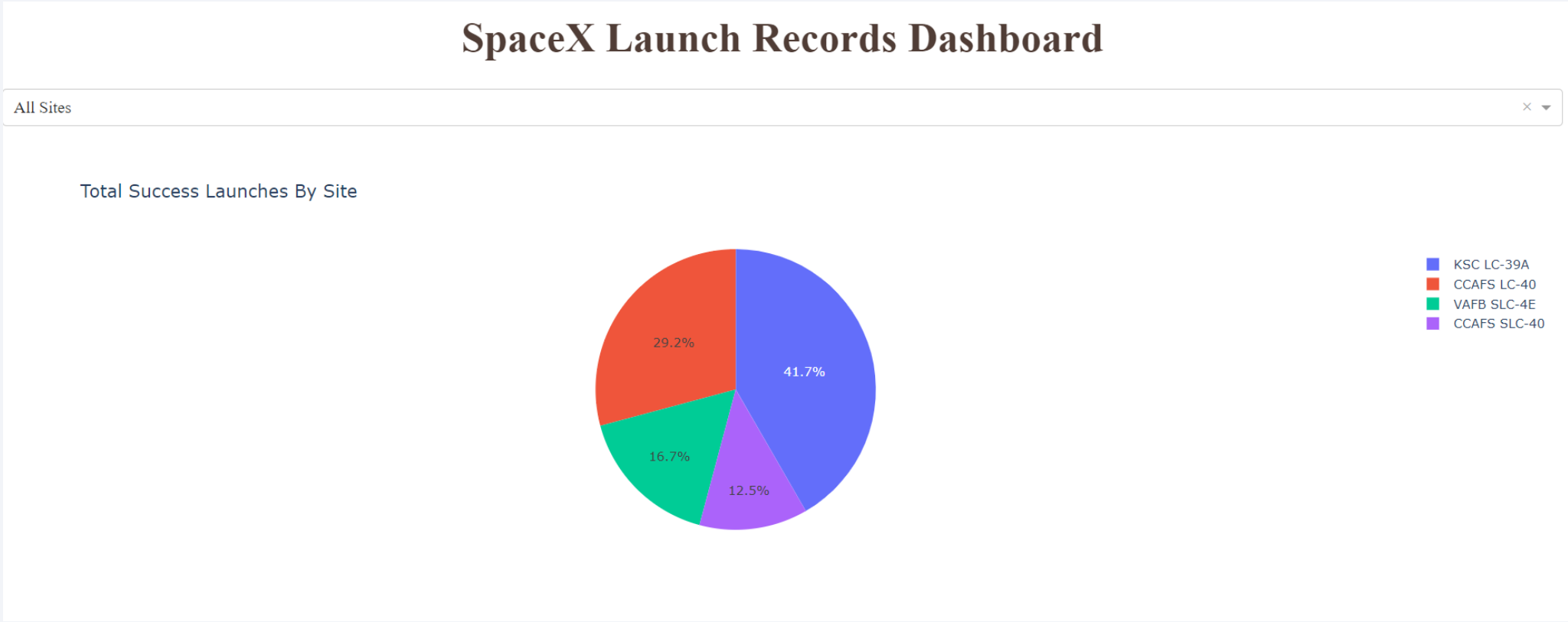
- The launch sites in the east coast are connected with railways even when there is no population nearby.




Section 4
Build a Dashboard with Plotly Dash

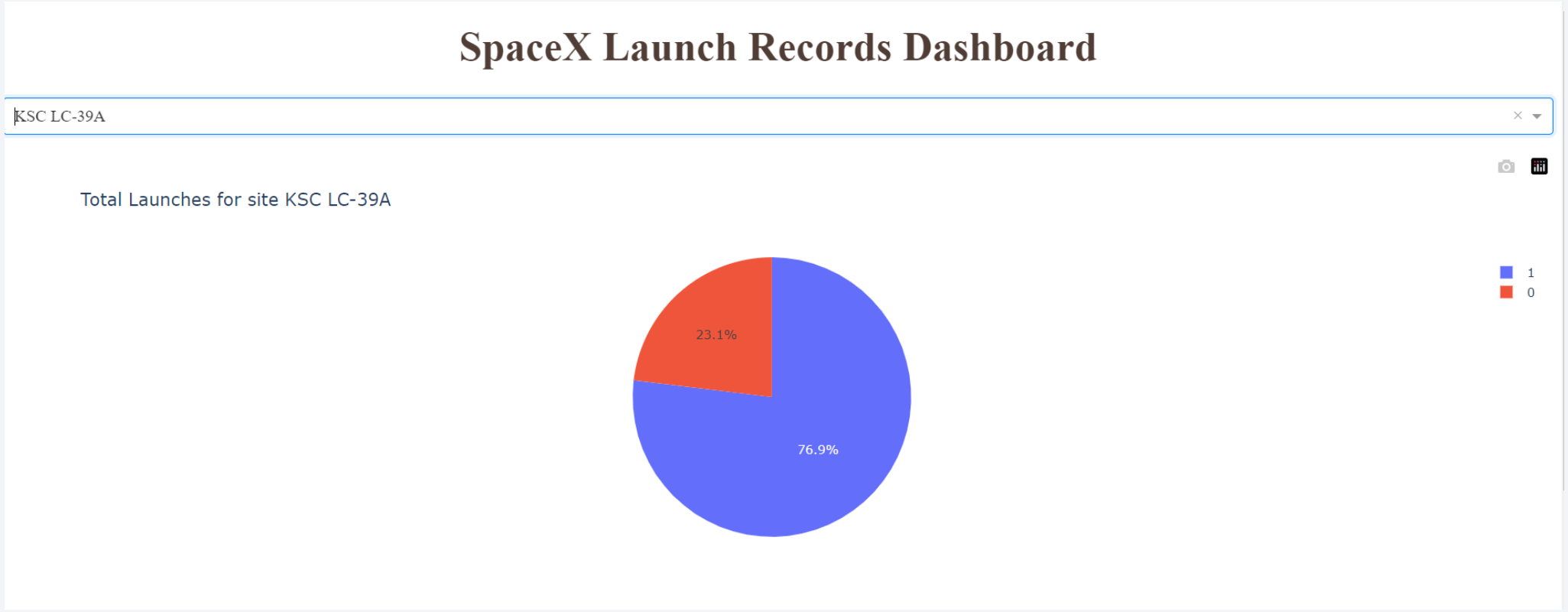
Success count for Launch site

- The KSC LC-39A has the highest launch success count o all the launch sites.
- CCAFS SLC-40 has the lowest success count.



Success ratio on KSC LC-39A

- KSC LC-39A has a 79% success ratio.



Payload/Success

- Lowest payloads tend to have higher success ratio.
- Payloads < 5000 kg are always successful.

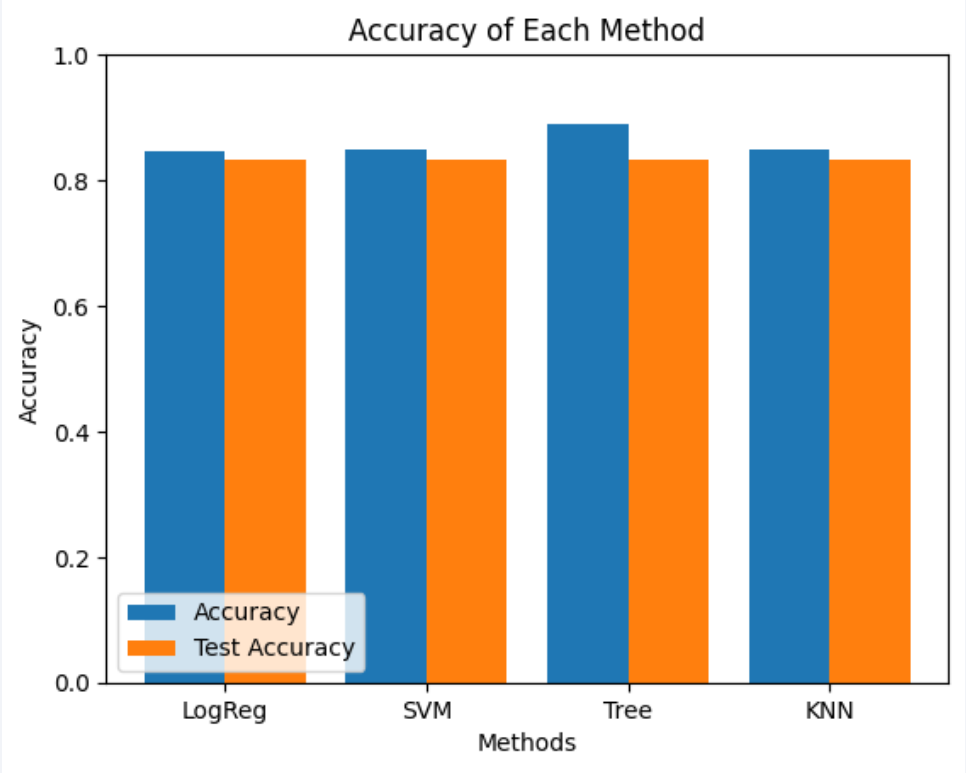
Section 5

Predictive Analysis (Classification)

Classification Accuracy

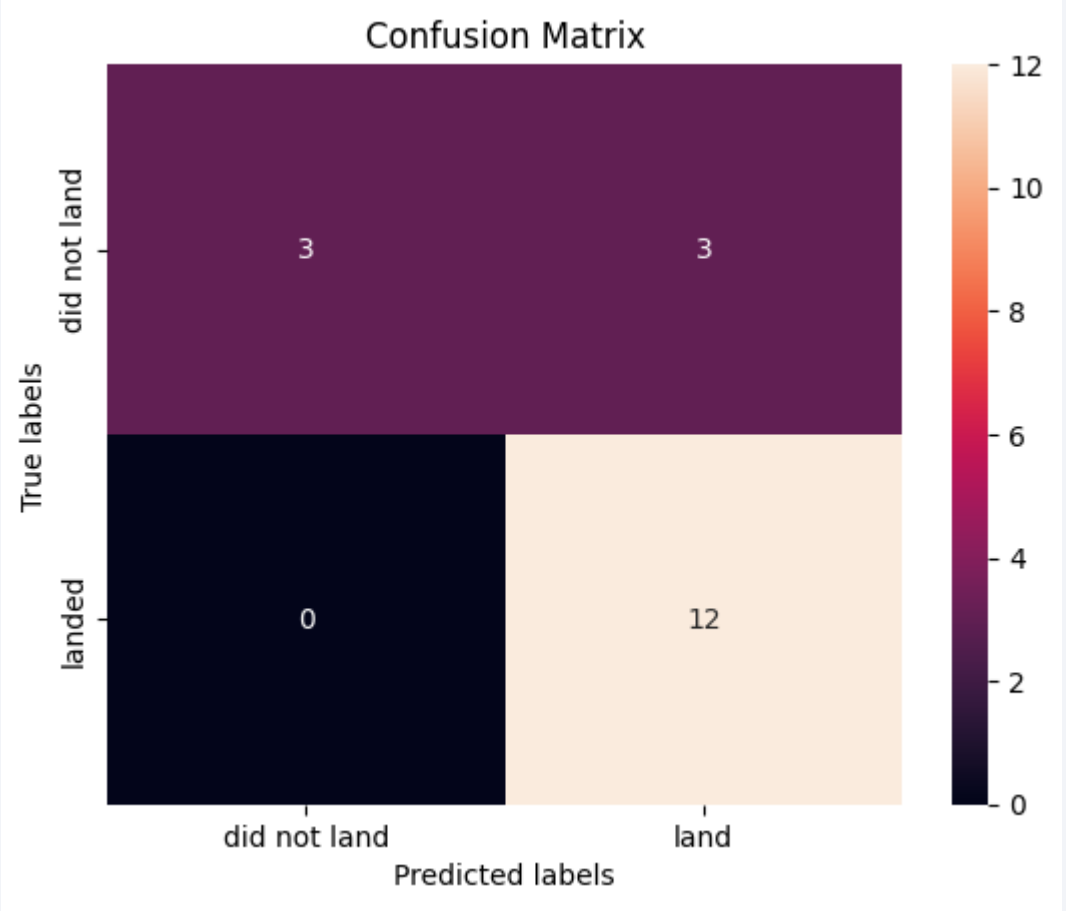
- Even though the decision tree has the highest accuracy for training, all the models perform the same on the testing dataset.

- Equal accuracy on testing can be caused by a bias in the testing set.



Confusion Matrix

- Decision tree is the best performing model in training.
- It correctly predict all cases on the successful samples, however, is very poor on the failure samples.



Conclusions

- It is possible to predict the outcome of a landing.
- The success rate increased over the years, this is probably because the learnt how to do it, whoever, this may affect the accuracy of the models.
- Despite Decision Tree being the best model on training, all the model performed the same on testing data.

Appendix

- Full GitHub repository in:
  https://github.com/Rivert97/Applied_Data_Science_Capstone

Thank you!