IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hoang Hien Anh
14 February 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies:

We follow the below steps to determine predictive model:

✓ Data collection

✓ Perform data wrangling

✓ Perform exploratory data analysis

✓ Perform interactive visual analytics

✓ Perform predictive analysis

- Summary of all results: 4 models (Logistic regression, Support vector machine, Decision tree classifier, K nearest neighbors) are suitable to predict the success of the first stage

Introduction

- Project background and context
  - ✓ SpaceX advertises Falcon 9 rocket launches on its website with a surprisingly low cost.
  - ✓ The reason for this low cost is because SpaceX can reuse the first stage.
  - ✓ Therefore, we can determine if the first stage will land, we can determine the cost of a launch.

- Problems you want to find answers
  - ✓ We want to predict if the Falcon 9 first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

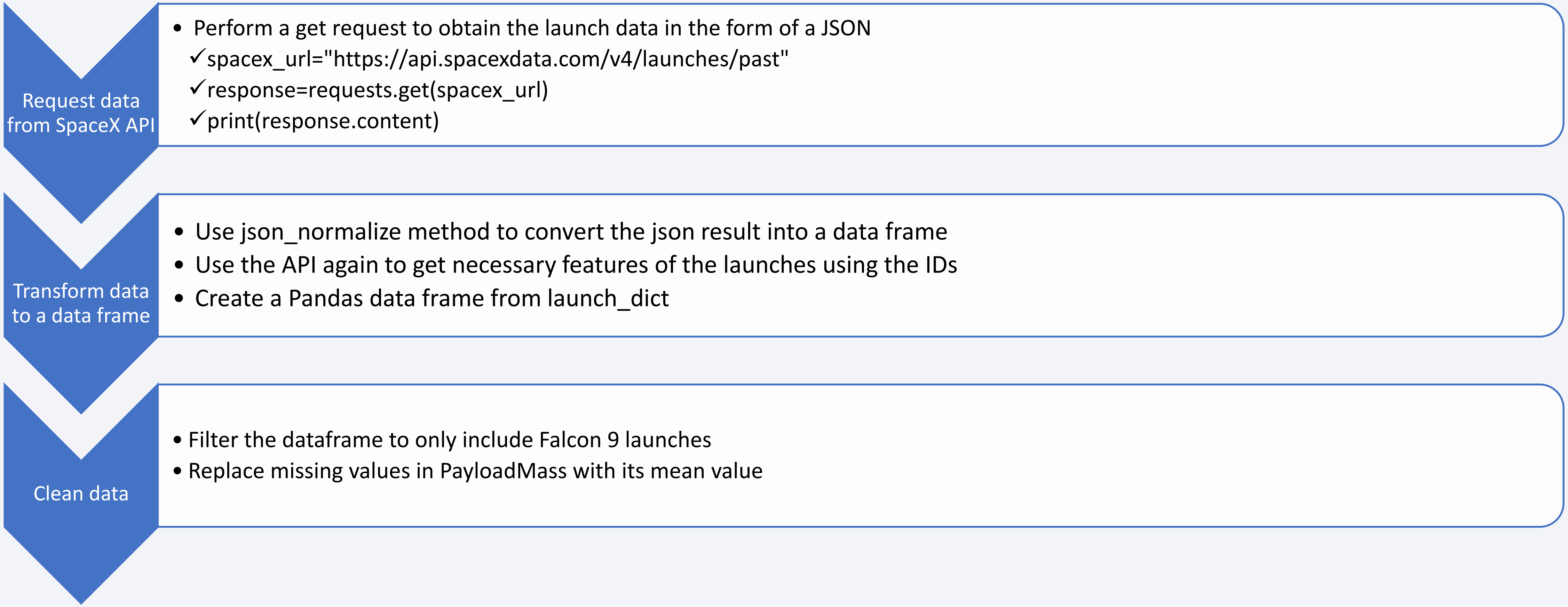
- Data collection methodology: We want to collect data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome. We can collect data using two ways:
  - SpaceX API
  - Web Scraping
- Perform data wrangling
  - Examine the data to see which attributes can be used to determine if the first stage can be reused
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Load data and standardize the data
  - Split the data into training and test data
  - Fit the data using different models
  - Using accuracy score and confusion matrix to determine the best model

Data Collection – SpaceX API



Request data from SpaceX API

- Perform a get request to obtain the launch data in the form of a JSON

✓spacex_url="https://api.spacexdata.com/v4/launches/past"

✓response=requests.get(spacex_url)

✓print(response.content)

Transform data to a data frame

- Use json_normalize method to convert the json result into a data frame
- Use the API again to get necessary features of the launches using the IDs
- Create a Pandas data frame from launch_dict

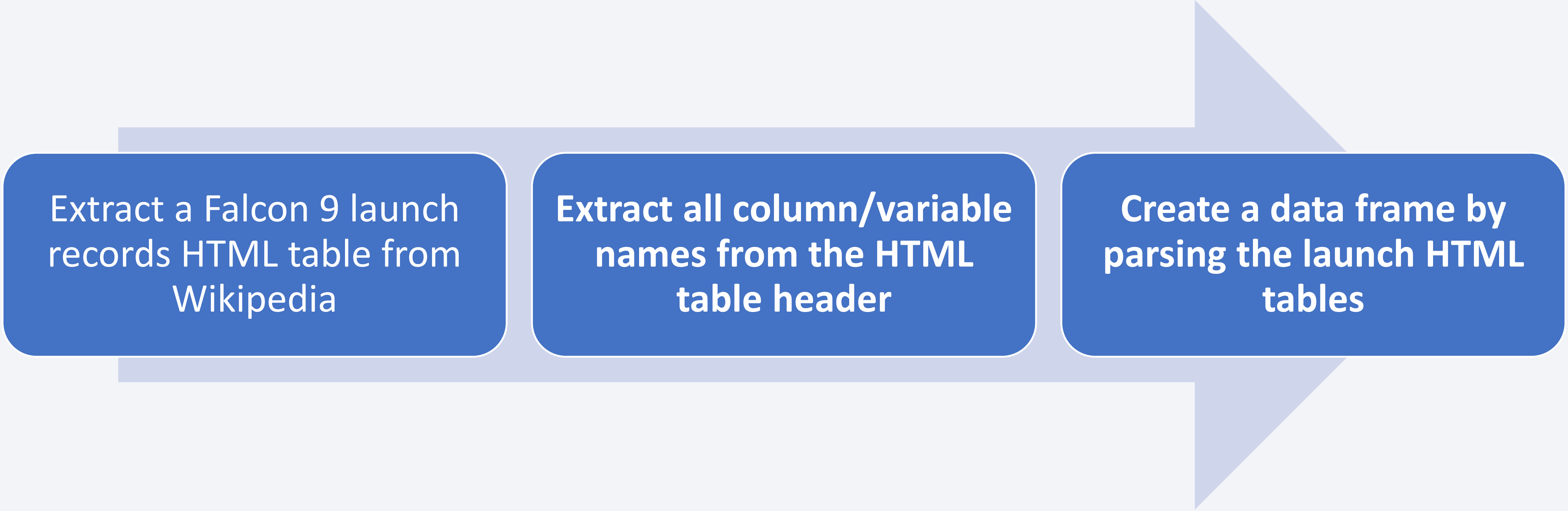
Clean data

- Filter the dataframe to only include Falcon 9 launches
- Replace missing values in PayloadMass with its mean value

GitHub URL:

https://github.com/hienanhhoang/Coursera_IBM_Data-Science/blob/Capstone-Project/Data%20collection_API.ipynb

Data Collection - Scraping



Extract a Falcon 9 launch records HTML table from Wikipedia

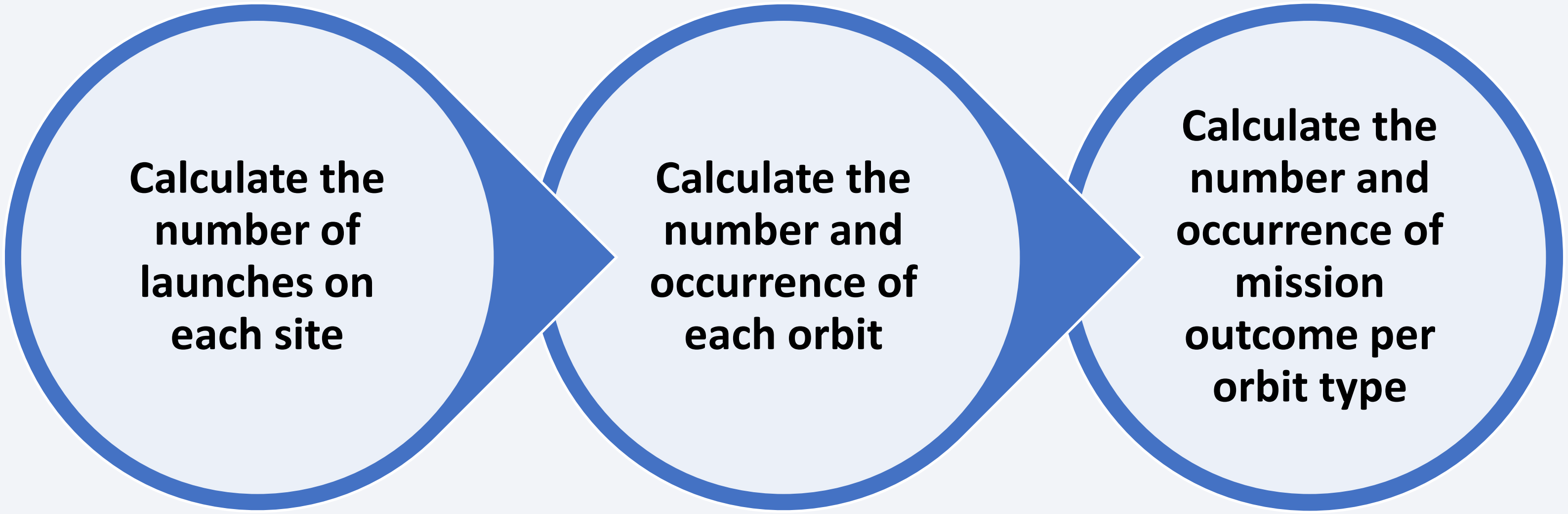
Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

GitHub URL:

https://github.com/hienanhhoang/Coursera_IBM_Data-Science/blob/Capstone-Project/Data%20collection_.Web%20Scraping.ipynb

Data Wrangling



Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

GitHub URL:

https://github.com/hienanhhoang/Coursera_IBM_Data-Science/blob/Capstone-Project/Data%20Wrangling.ipynb

EDA with Data Visualization

- In order to see how features of the launches affect the launch outcome, the following charts were plotted:
  - ✓ The relationship between Flight Number and Payload
  - ✓ The relationship between Flight Number and Launch Site
  - ✓ The relationship between Payload and Launch Site
  - ✓ The relationship between success rate of each orbit type
  - ✓ The relationship between Flight Number and Orbit type
  - ✓ The relationship between Payload and Orbit type
  - ✓ The launch success yearly trend

GitHub URL: https://github.com/hienanhhoang/Coursera_IBM_Data-Science/blob/Capstone-Project/EDA%20with%20Visualization.ipynb

EDA with SQL

- 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 which have success in drone ship and have payload mass greater than 4000 but less than 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 for in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

GitHub URL: https://github.com/hienanhhoang/Coursera_IBM_Data-Science/blob/Capstone-Project/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- All launch sites are marked on the map to visualize their locations
- The launch outcomes for each site are added to see which sites have high success rates
- The lines visualizing the distances between a launch site to its proximities (coastline, railway, highway, city) are added to see the conditions surrounding a launch site

GitHub URL: https://github.com/hienanhhoang/Coursera_IBM_Data-Science/blob/Capstone-Project/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

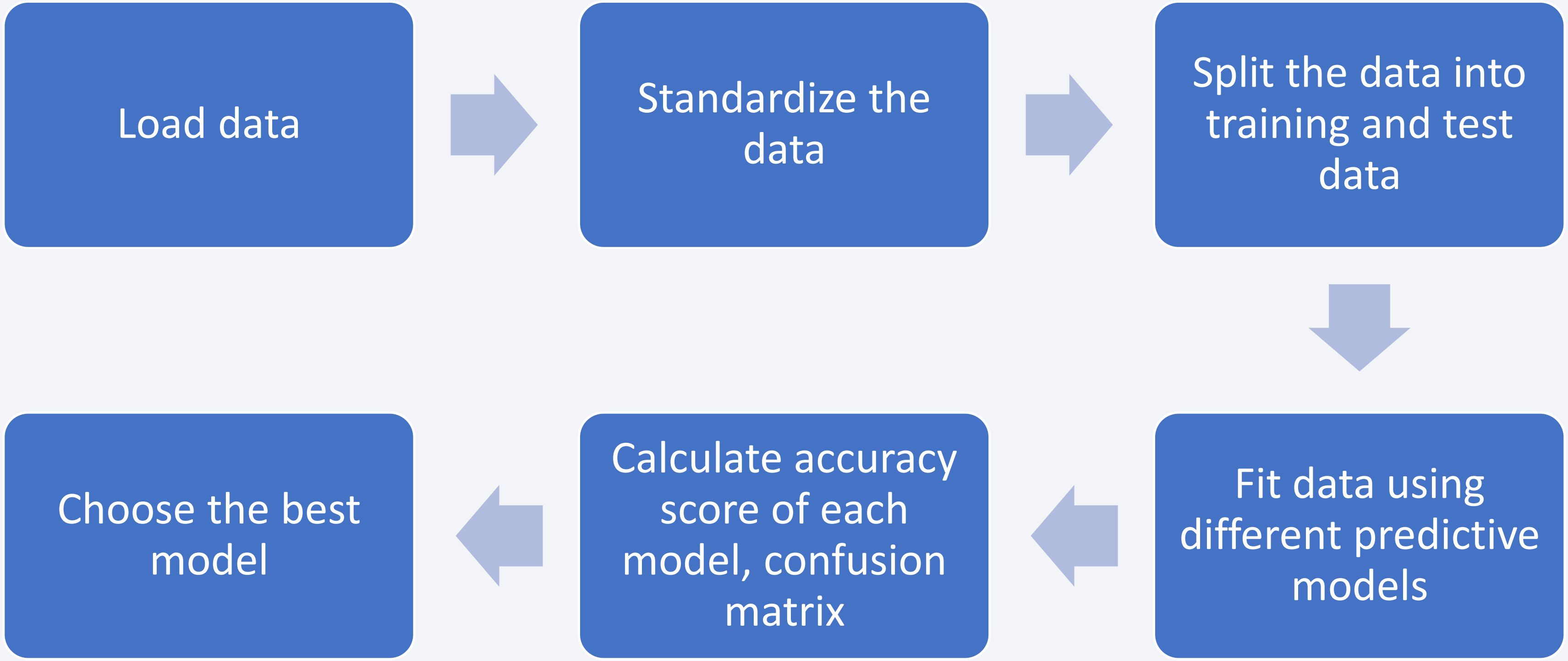
Build a Dashboard with Plotly Dash

In order to see which launch site, payload range, and booster version has the highest success rates, these plots are rendered:

- Total Successful Launches for All Sites and for each site
- Scatter Plot of Payload vs. Launch Outcome with the point color set to the booster version

GitHub URL: https://github.com/hienanhhoang/Coursera_IBM_Data-Science/blob/Capstone-Project/Dashboard.ipynb

Predictive Analysis (Classification)



Load data

Standardize the data

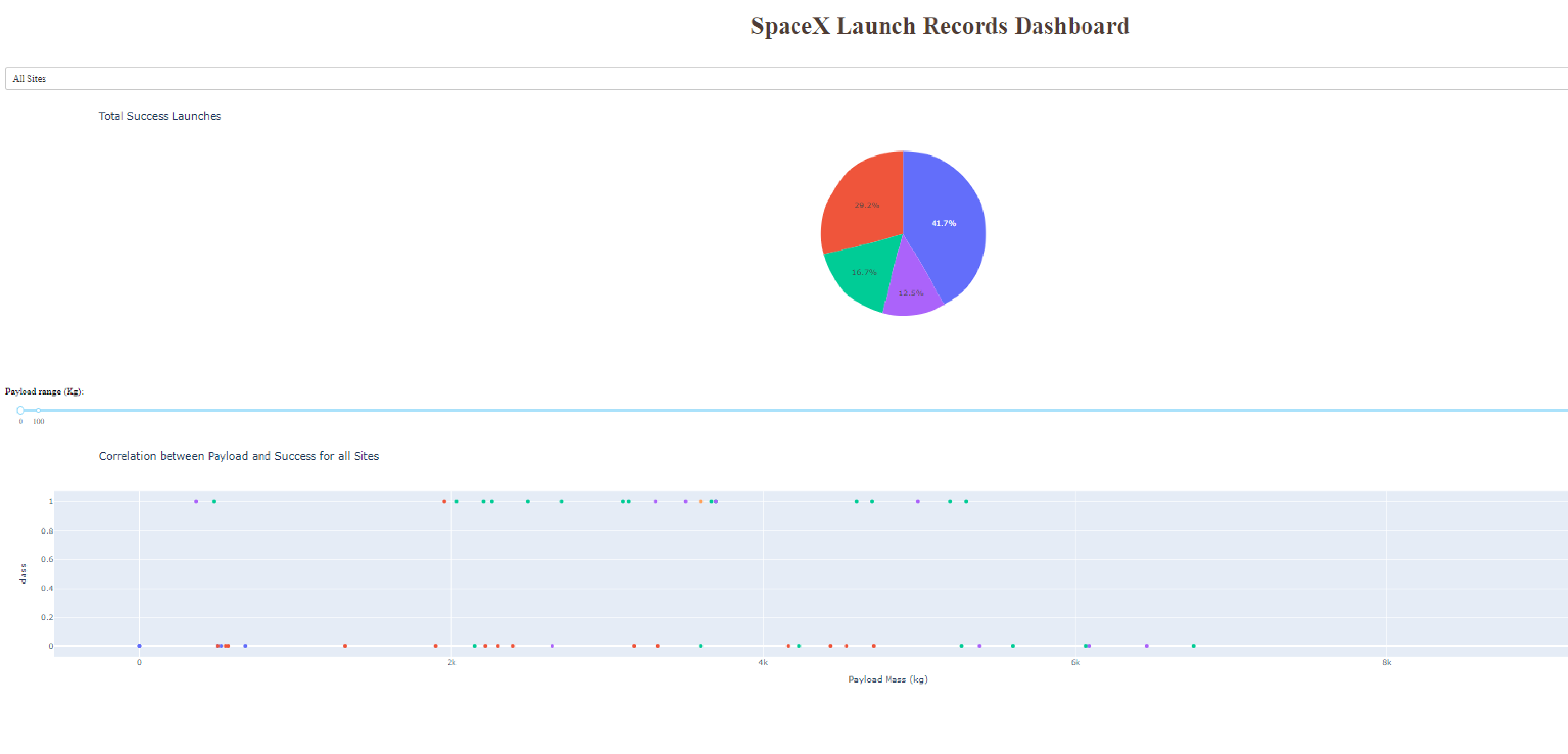
Split the data into training and test data

Fit data using different predictive models

Calculate accuracy score of each model, confusion matrix

Choose the best model

GitHub URL: https://github.com/hienanhhoang/Coursera_IBM_Data-Science/blob/Capstone-Project/Machine%20Learning%20Prediction.ipynb



SpaceX Launch Records Dashboard

All Sites

Total Success Launches

Payload range (Kg):

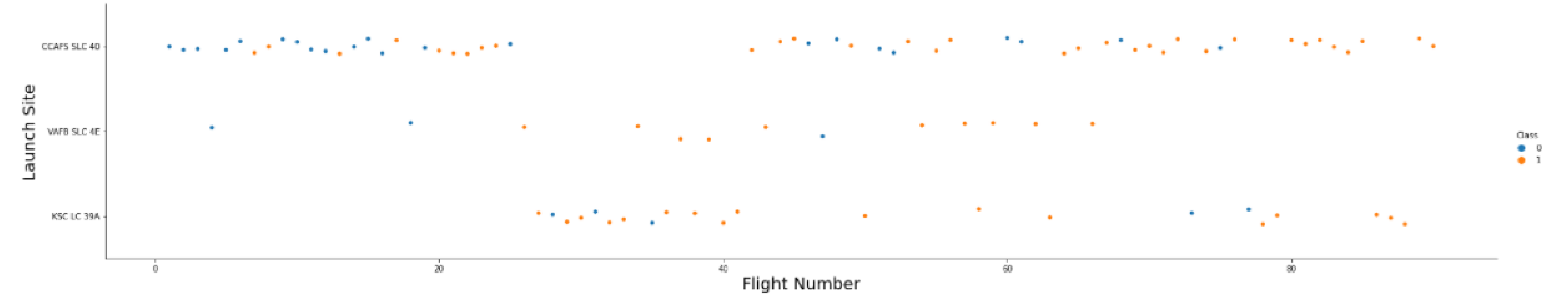
Correlation between Payload and Success for all Sites

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

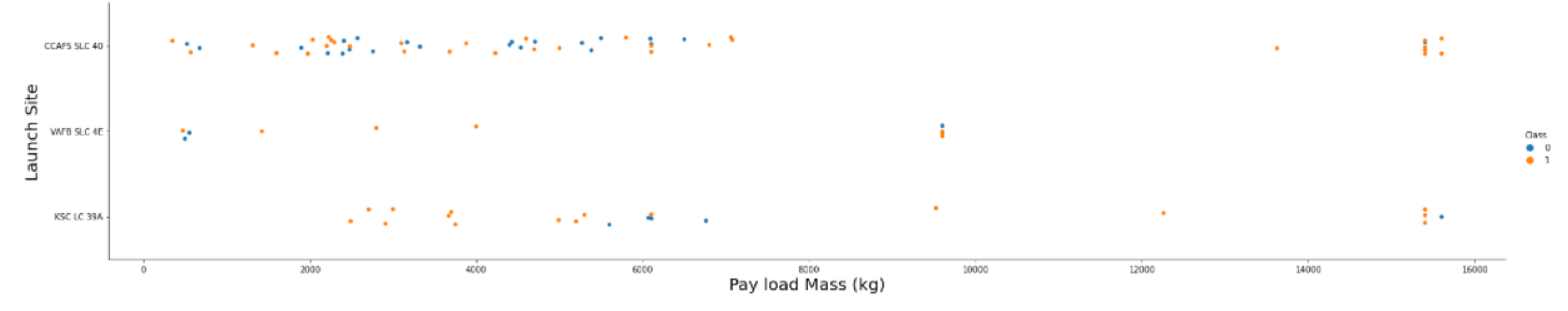
- Scatter plot of Flight Number vs. Launch Site



- For all three launch sites, as the flight number increases, the first stage is more likely to land successfully
- Launch site CCAFS SLC 40 has the highest number of launches.

Payload vs. Launch Site

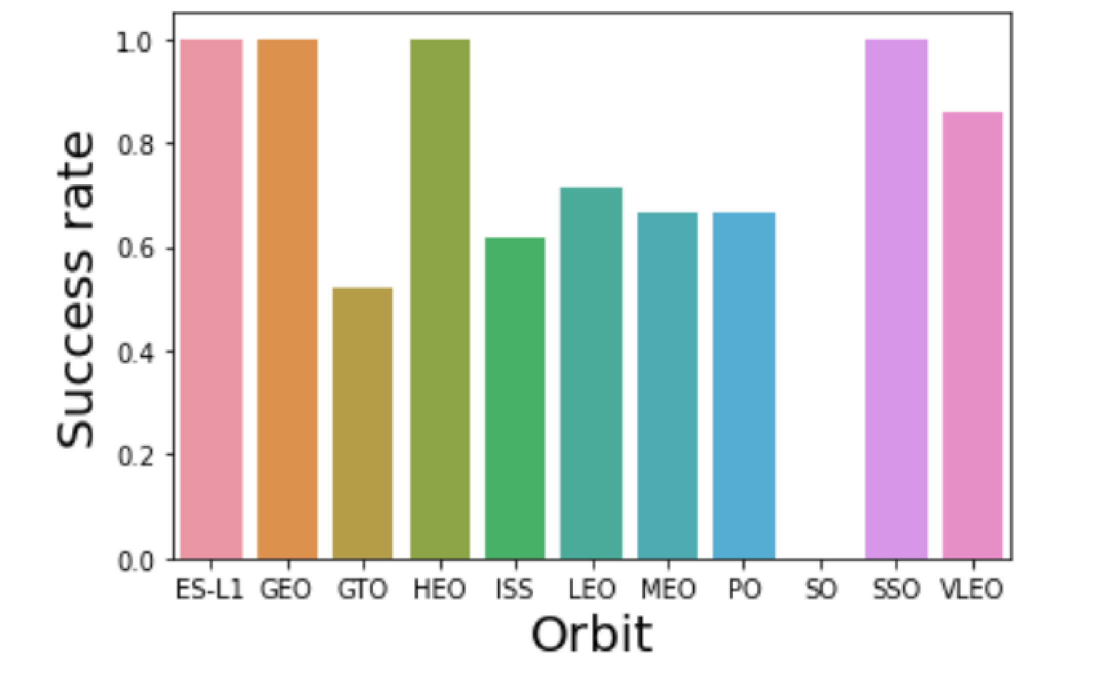
- Scatter plot of Payload vs. Launch Site



- There are not many rockets launched for heavy payload mass (greater than 10000)
- It seems the more massive the payload, first stage is more likely to land successfully

Success Rate vs. Orbit Type

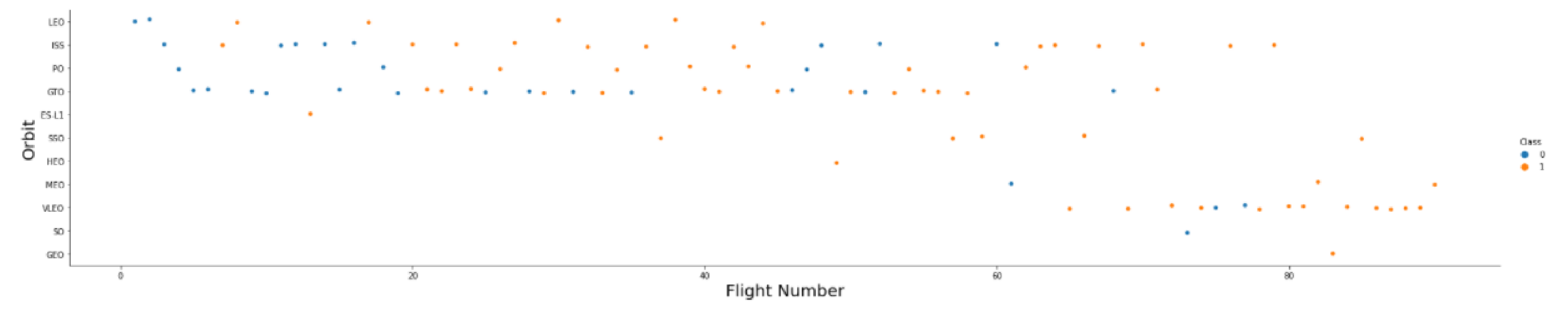
- Bar chart for the success rate of each orbit type



- ES-L1, GEO, HEO, and SSO have the highest success rate (=1), while SO's success rate is 0
- Other orbits' success rates hover around 0.5

Flight Number vs. Orbit Type

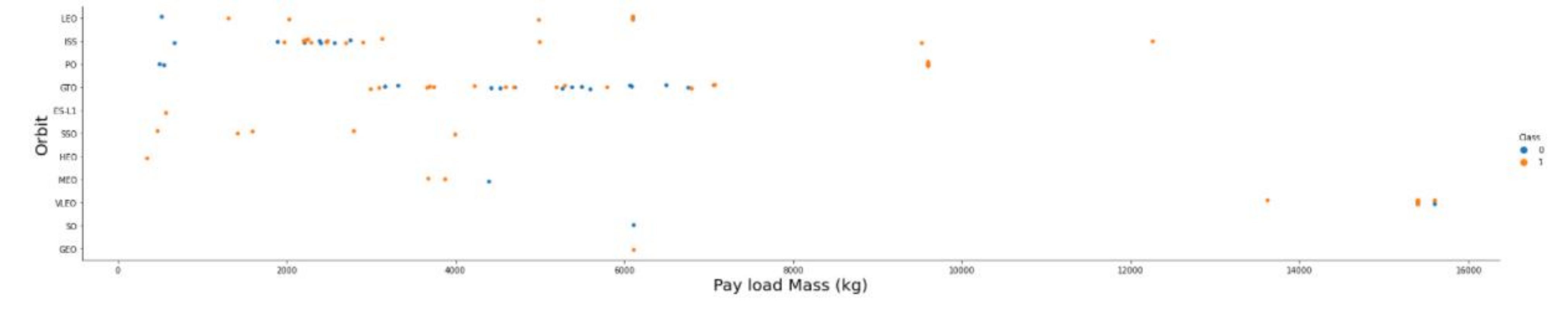
- Scatter point of Flight number vs. Orbit type



- There is only one launch in the orbits which have perfect success/unsuccess rate → not enough information to make predictions
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

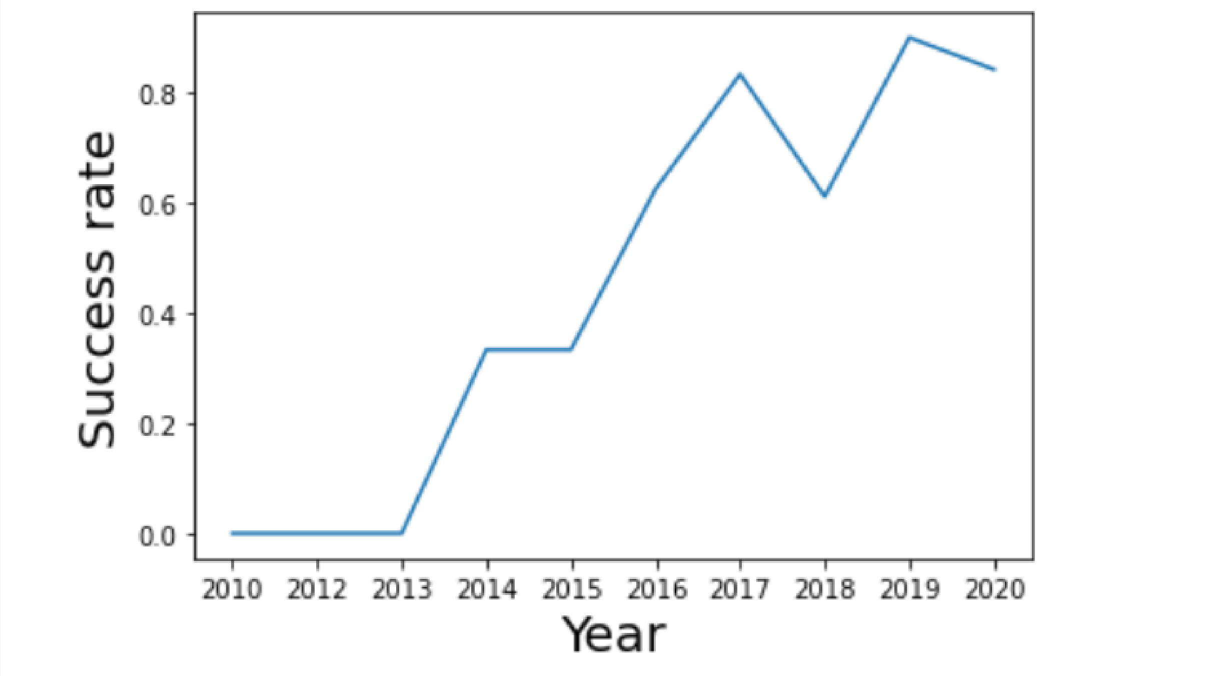
- Scatter point of payload vs. orbit type



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

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



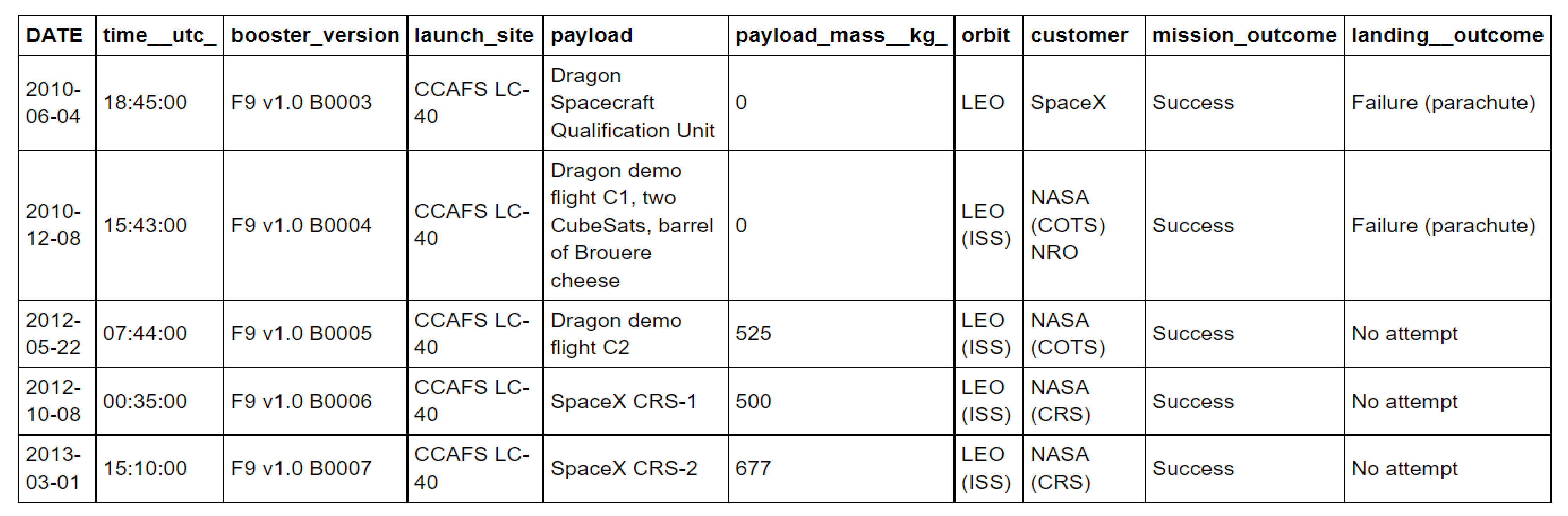
- The success rate is increasing since 2013 till 2017.
- The success rate fluctuates during 2017 – 2020.

All Launch Site Names

- Query: %sql select distinct(LAUNCH_SITE) from SPACEX
- There are 4 unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names 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

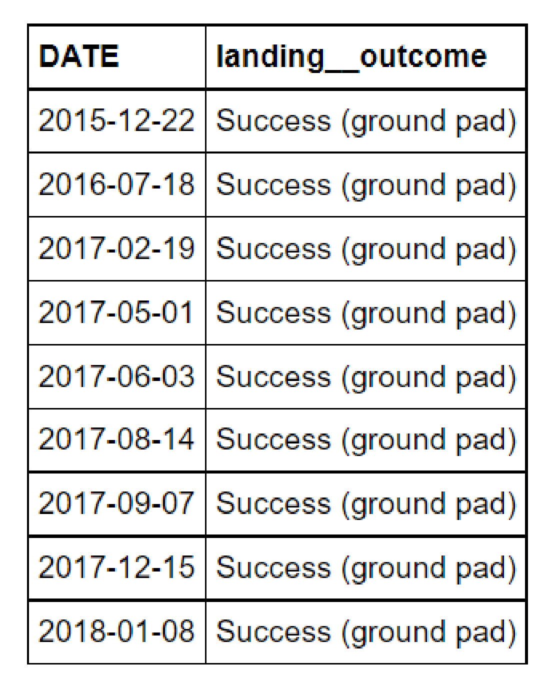
- %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEX where customer = 'NASA (CRS)’
- The total payload mass carried by boosters launched by NASA (CRS)：45596

Average Payload Mass by F9 v1.1

- %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEX where booster_version like '%F9 v1.1%'
- Average payload mass carried by booster version F9 v1.1: 2534

First Successful Ground Landing Date

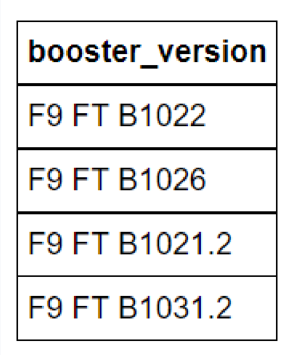
- %sql select date, landing__outcome from SPACEX where landing__outcome = 'Success (ground pad)' order by date



DATE	landing__outcome
2015-12-22	Success (ground pad)
2016-07-18	Success (ground pad)
2017-02-19	Success (ground pad)
2017-05-01	Success (ground pad)
2017-06-03	Success (ground pad)
2017-08-14	Success (ground pad)
2017-09-07	Success (ground pad)
2017-12-15	Success (ground pad)
2018-01-08	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select booster_version from SPACEX where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000



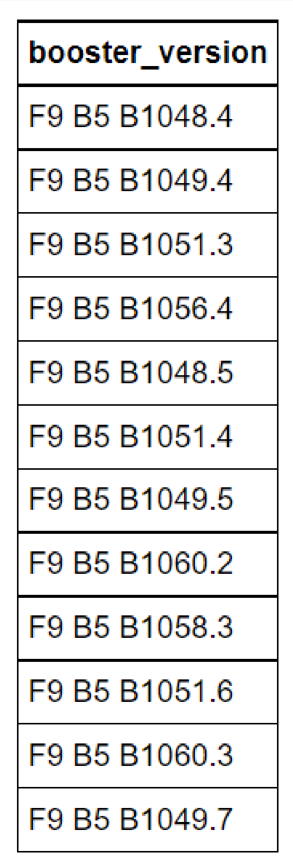
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- %sql select count(landing__outcome) from SPACEX where landing__outcome like '%Success%' or landing__outcome like '%Failure%’
- Total number of successful and failure mission outcomes：71

Boosters Carried Maximum Payload

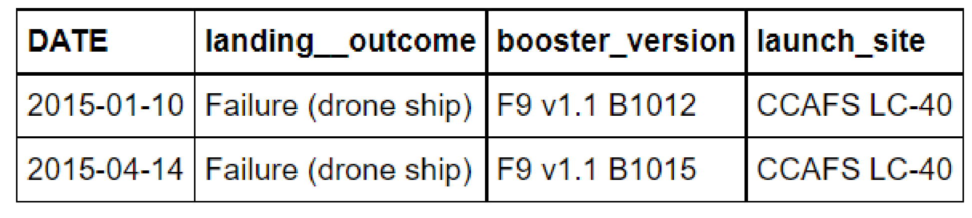
- %sql select booster_version from SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX)



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

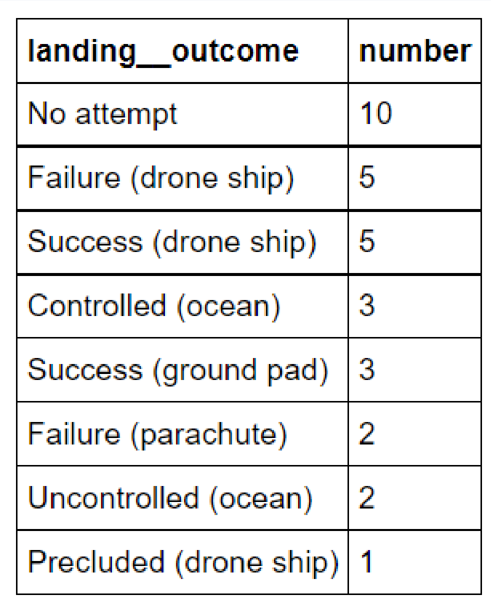
- %sql select DATE, landing__outcome, booster_version, launch_site from SPACEX where landing__outcome = 'Failure (drone ship)' and DATE like '%2015%’



DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- %sql select landing__outcome, count(landing__outcome) as number from SPACEX where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome order by number desc



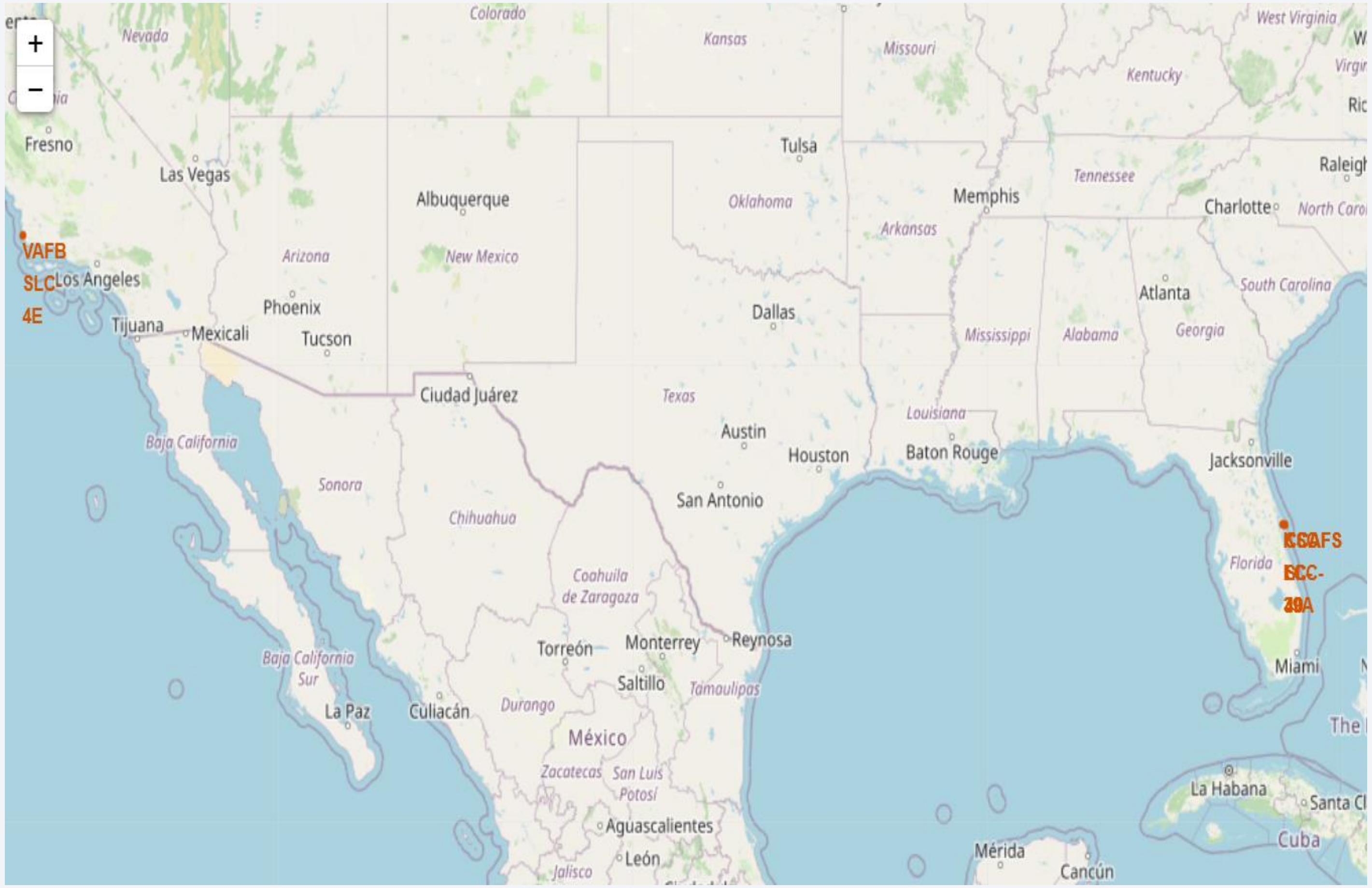
landing__outcome	number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

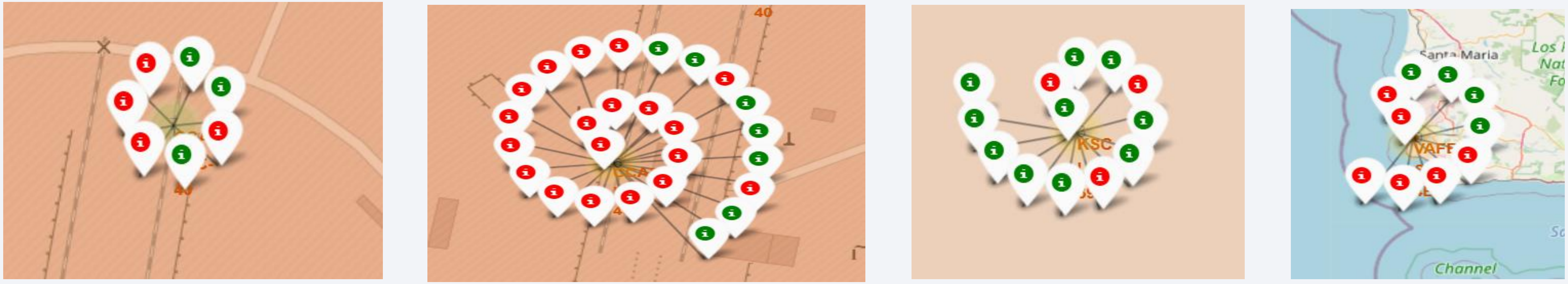
Launch Sites Proximities Analysis

Launch Sites’ Locations

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



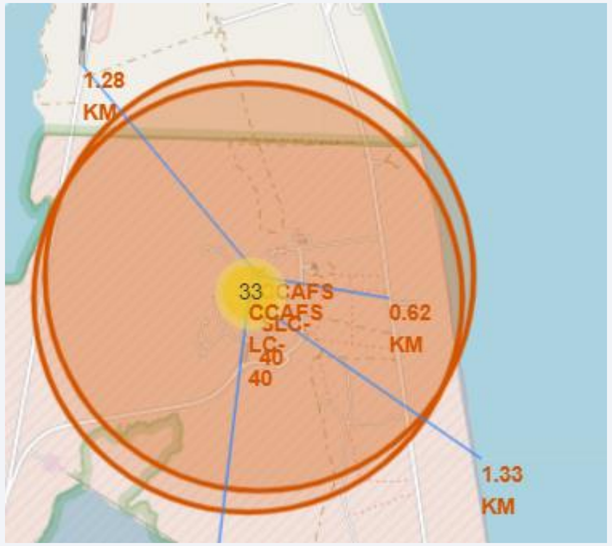
The success/failed launches for each site

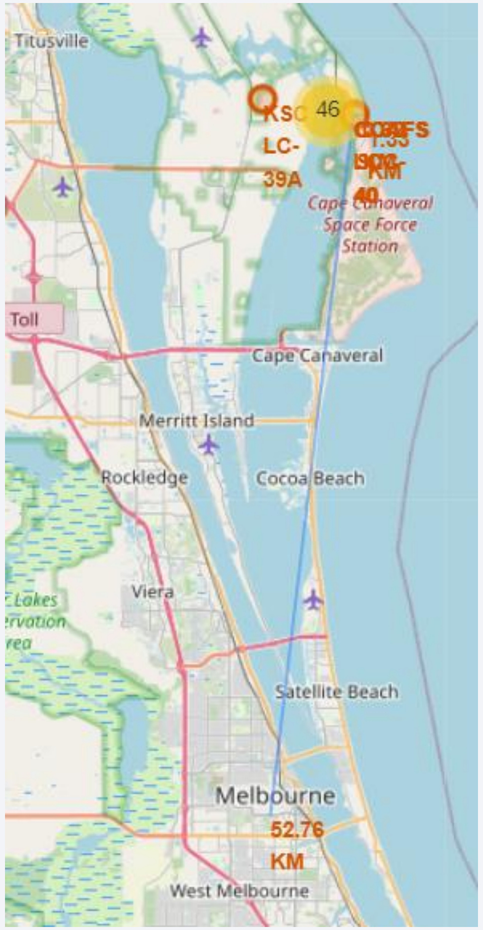


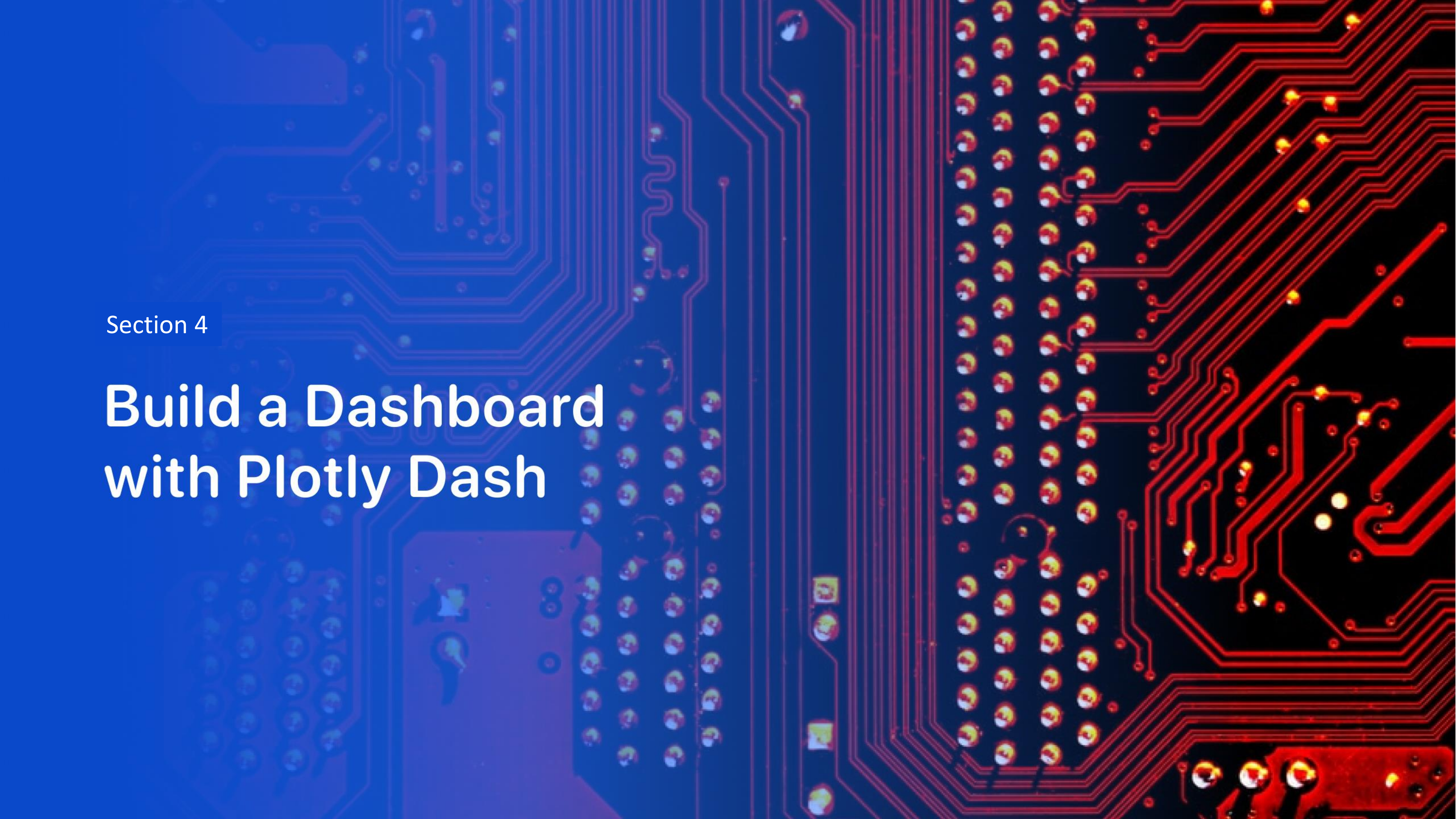
- Green color represents successful launch while red color represents unsuccessful launch.
- From the map, we can see that KSC LC-39A has the highest success rate

A launch site and its proximities

- Launch sites are close to railway, highway, and coastline.
- Launch sites are far from the cities.



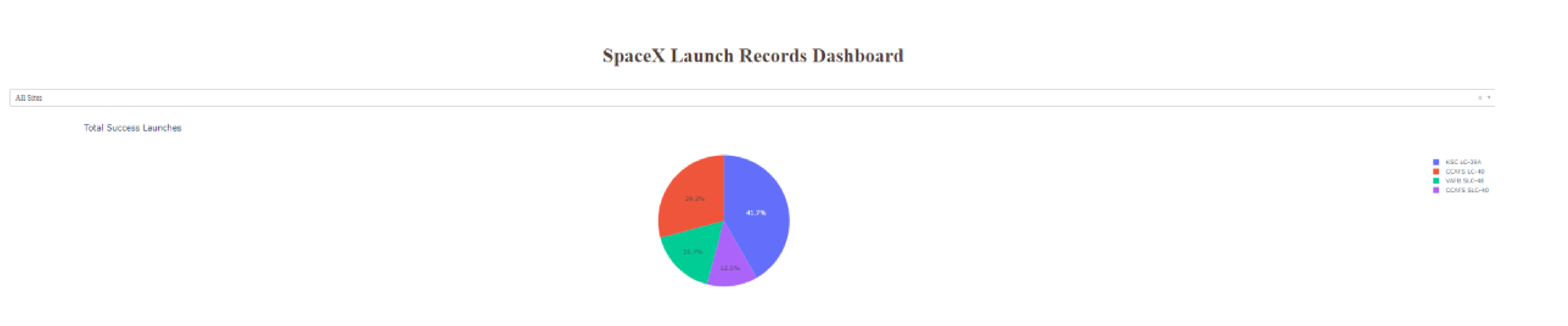




Section 4

Build a Dashboard with Plotly Dash

Successful Launches for All Sites



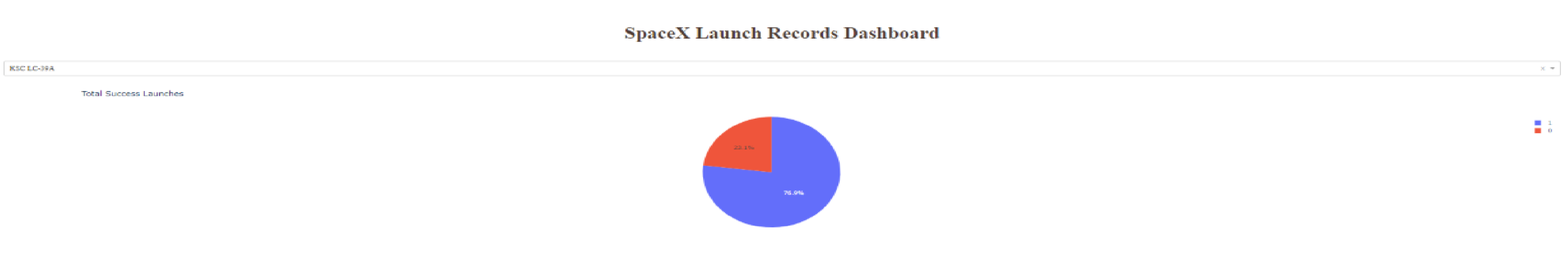
SpaceX Launch Records Dashboard

All Sites

Total Success Launches

KSC LC-39A has the highest success rate.

Success rate for KSC LC-39A

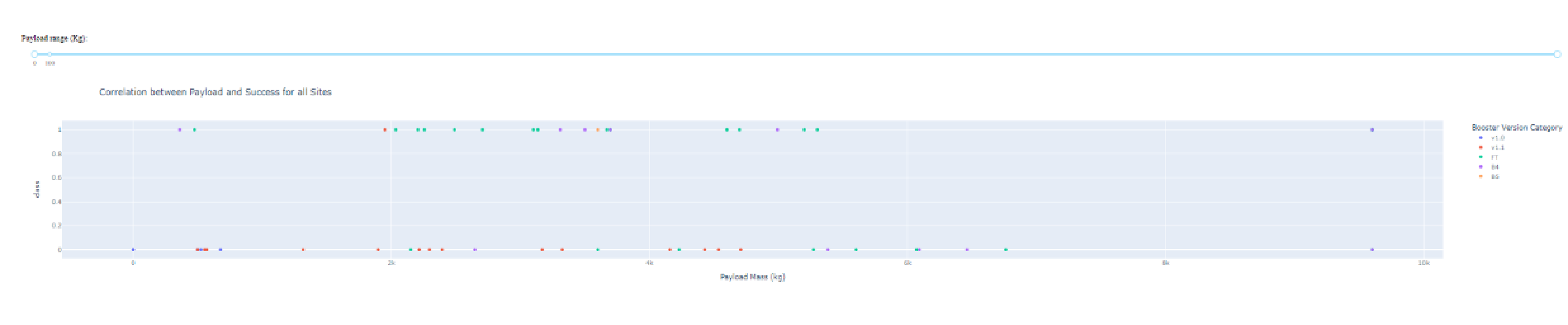


SpaceX Launch Records Dashboard

KSC LC-39A

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

Payload vs. Launch Outcome

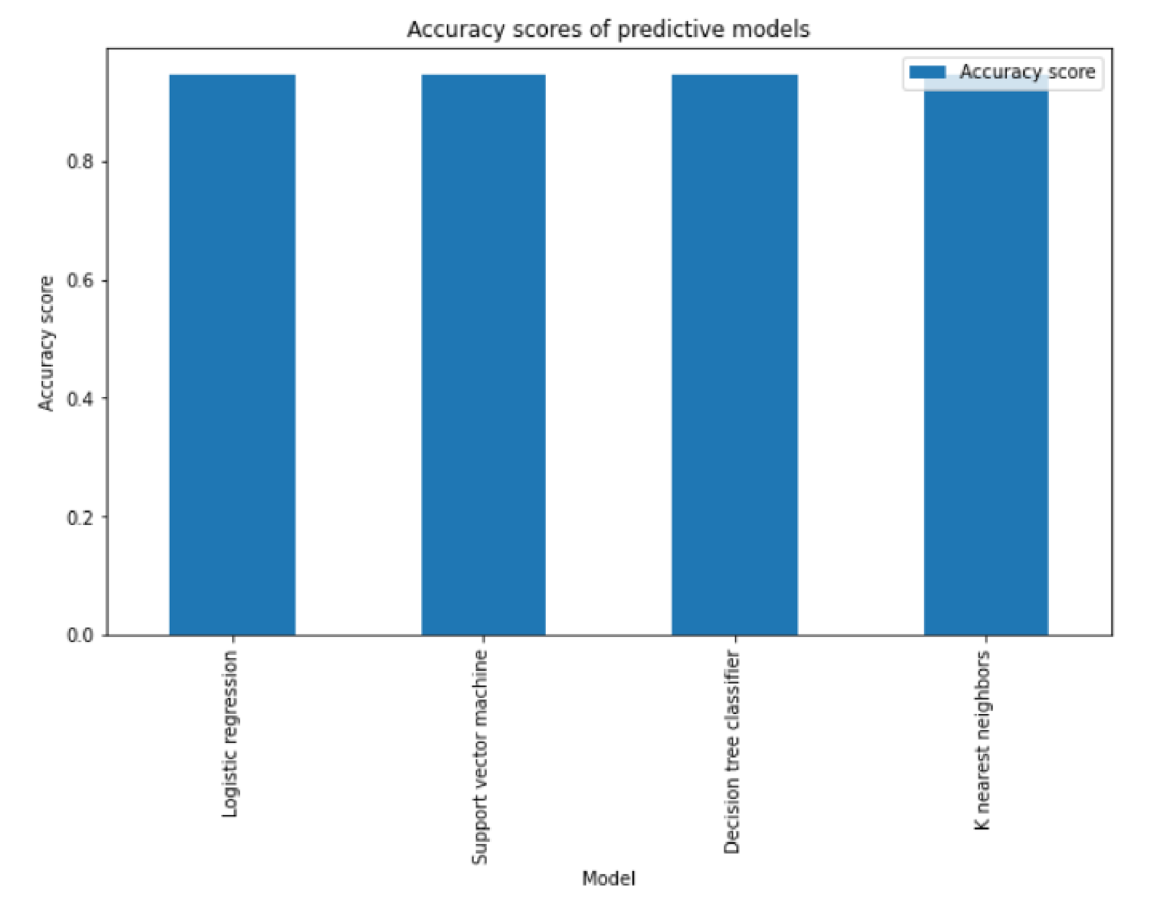


- Booster version FT has the highest success rate
- Payload range (2000 – 6000) has the highest success rate

Section 5

Predictive Analysis (Classification)

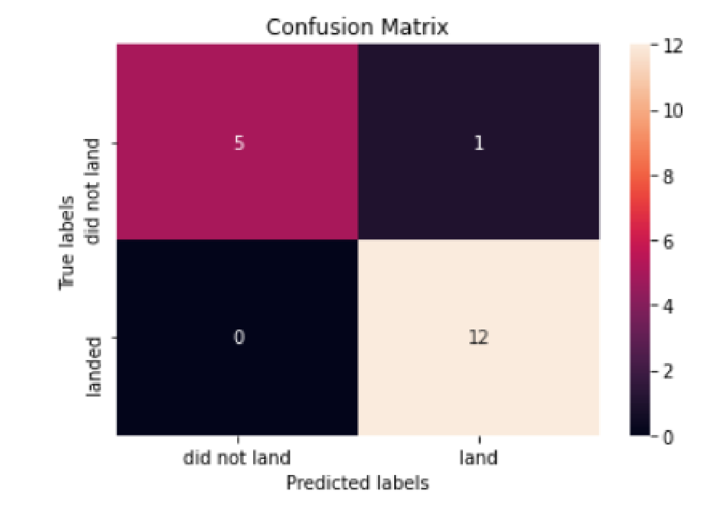
Classification Accuracy

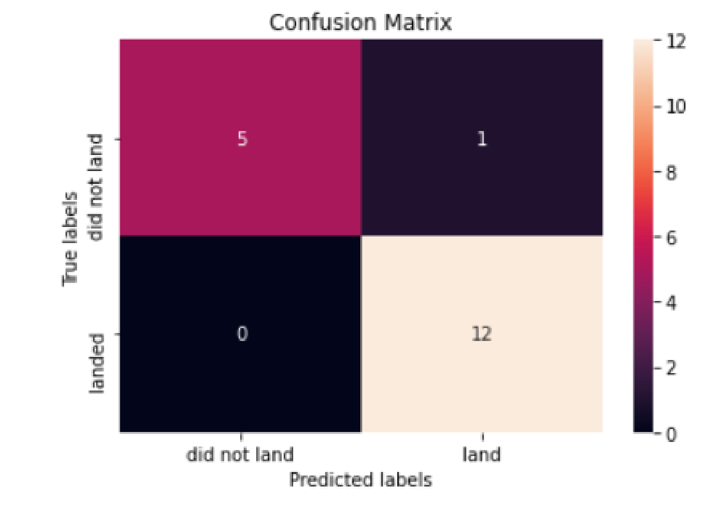


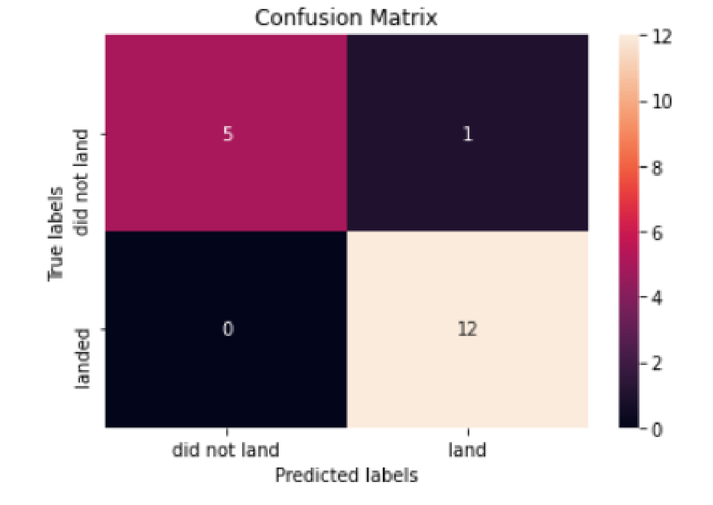
- All models have the same accuracy score.

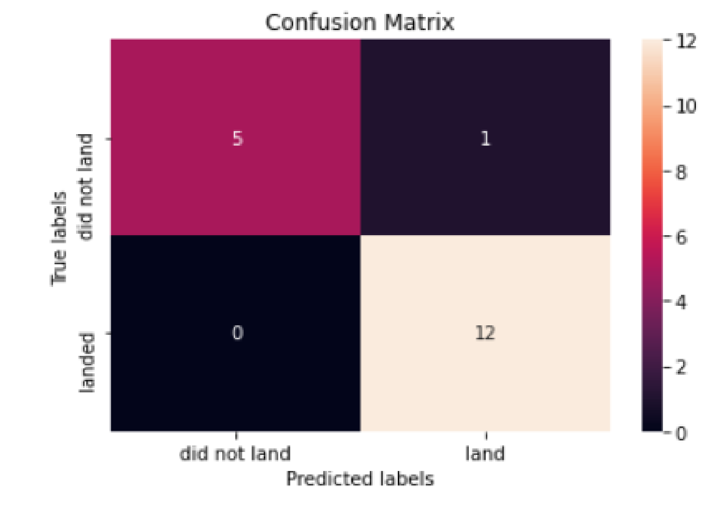
Confusion Matrix

- All models perform equally well.









Conclusions

- The following attributes can be used to predict the success of the first stage
  - ✓ Flight Number
  - ✓ Pay load
  - ✓ Launch Site
  - ✓ Orbit Type
  - ✓ Booster Version
- We can use all 4 models (Logistic regression, Support vector machine, Decision tree classifier, K nearest neighbors) to predict the success of the first stage

Thank you!