IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

CHIA JAN FENG
14 August 2024

Outline

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

Executive Summary

This project involved collecting and analyzing SpaceX launch data to predict the success of the Falcon 9's first stage landing, a key factor in reducing launch costs. Data was gathered via API requests and web scraping, followed by data wrangling and exploratory analysis using visualization tools and SQL queries. The analysis identified critical factors such as payload mass, launch site, and orbit type that influence landing success.

A predictive model was developed and fine-tuned using GridSearchCV, with the Decision Tree Classifier achieving the highest accuracy of 87.68%. The results provide actionable insights for optimizing rocket launch operations and reducing costs.

Introduction

- Project background and context

The commercial space age is now a reality, with companies like SpaceX leading the way in making space travel more affordable. SpaceX's Falcon 9 rocket is notable for its ability to reuse the first stage, significantly reducing launch costs compared to competitors. While other providers charge up to $165 million per launch, SpaceX offers launches for $62 million, thanks to this reusability.

- Problem Statement:

1. Can the first stage of Falcon 9 be successfully reused?
2. How to determine the price of a SpaceX rocket launch?
3. What factors influence the successful landing of the first stage?

Section 1

Methodology

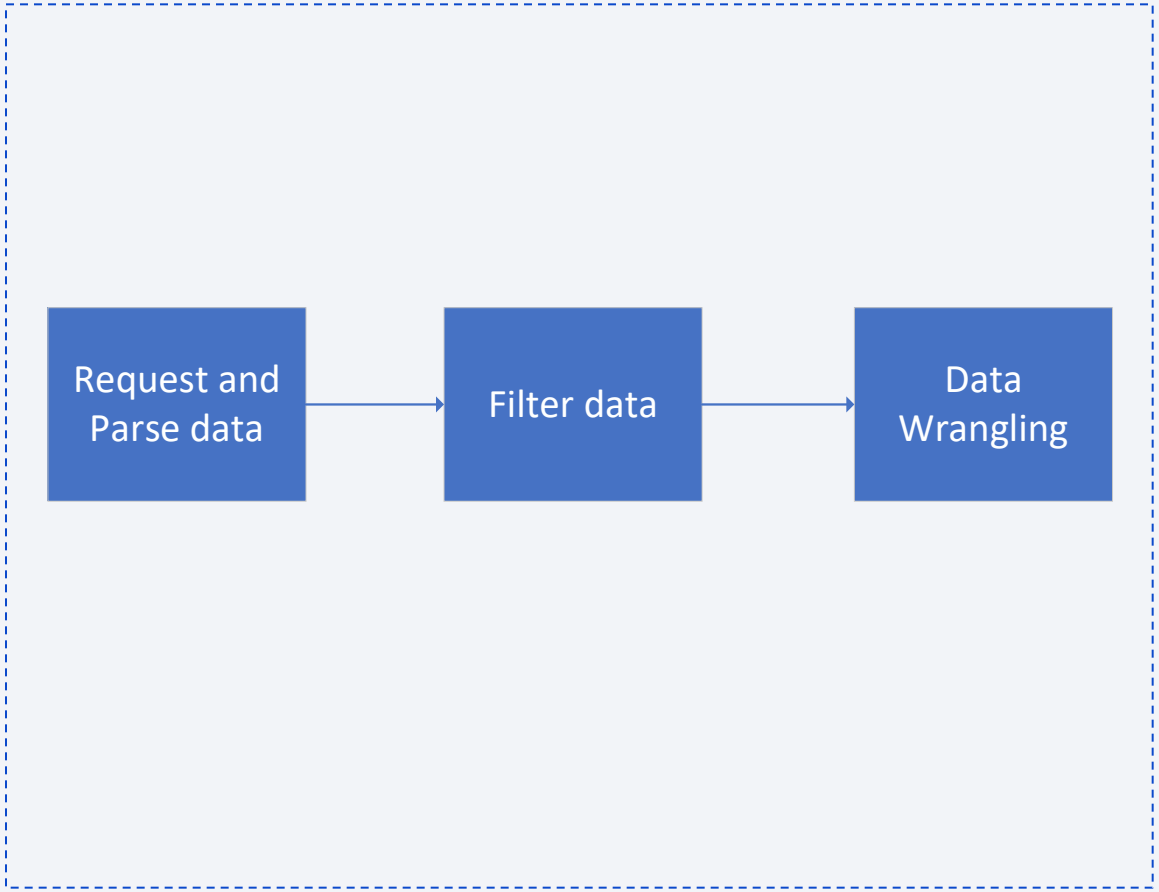
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by using API and Web Scraping using BeautifulSoup library
- Perform data wrangling
  - 'Class' column is added to show quantitatively if the landing is success (1) or failure (0).
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification model is built by creating an object and pass the data and parameters to the object, the model is then tuned by using GridSearchCV. Finally, the model with best performing parameters are evaluated based on its accuracy.

Data Collection – SpaceX API

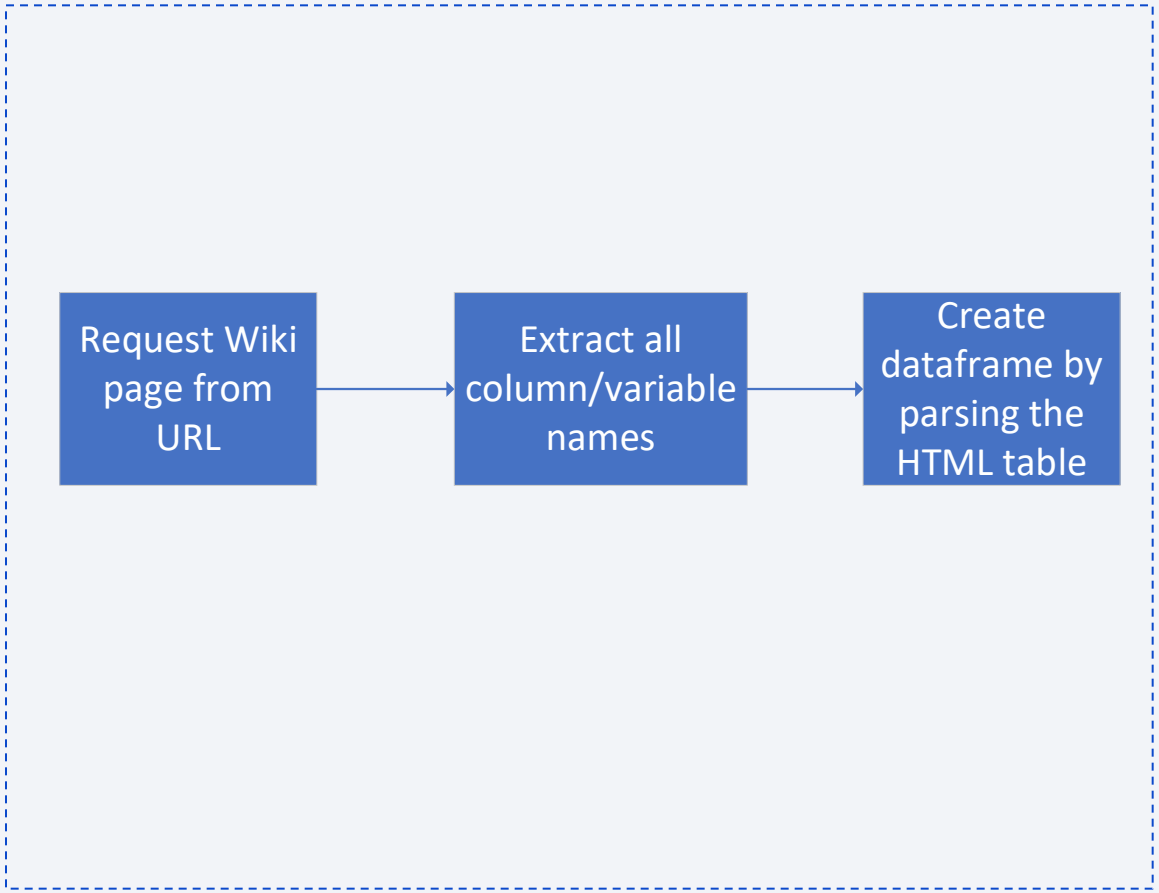
- Procedure:
  1. Request and parse SpaceX launch data using GET request
  2. Filter the data to only include ‘Falcon 9’ launches.
  3. Data wrangling and dealing with missing values



- GitHub URL:
  SpaceX_Project/Data_Collection_API.py at main · StunMan123/SpaceX_Project (github.com)

Data Collection - Scraping

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

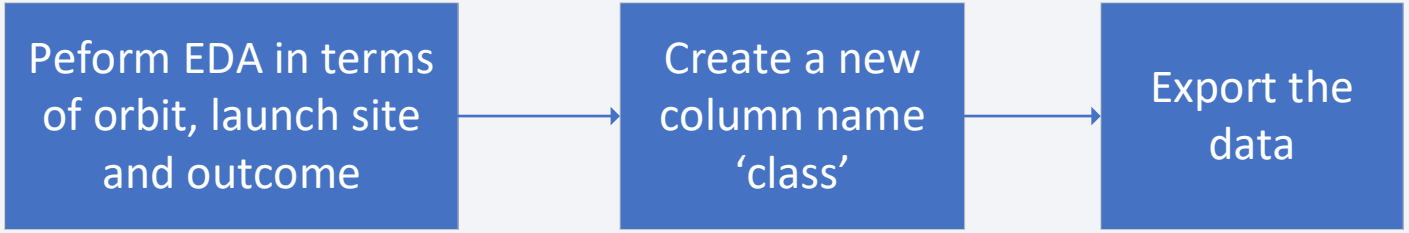


- GitHub URL:

  SpaceX_Project/Data_Collection_BeautifulSoup_wikipedia.py at main · StunMan123/SpaceX_Project (github.com)

Data Wrangling

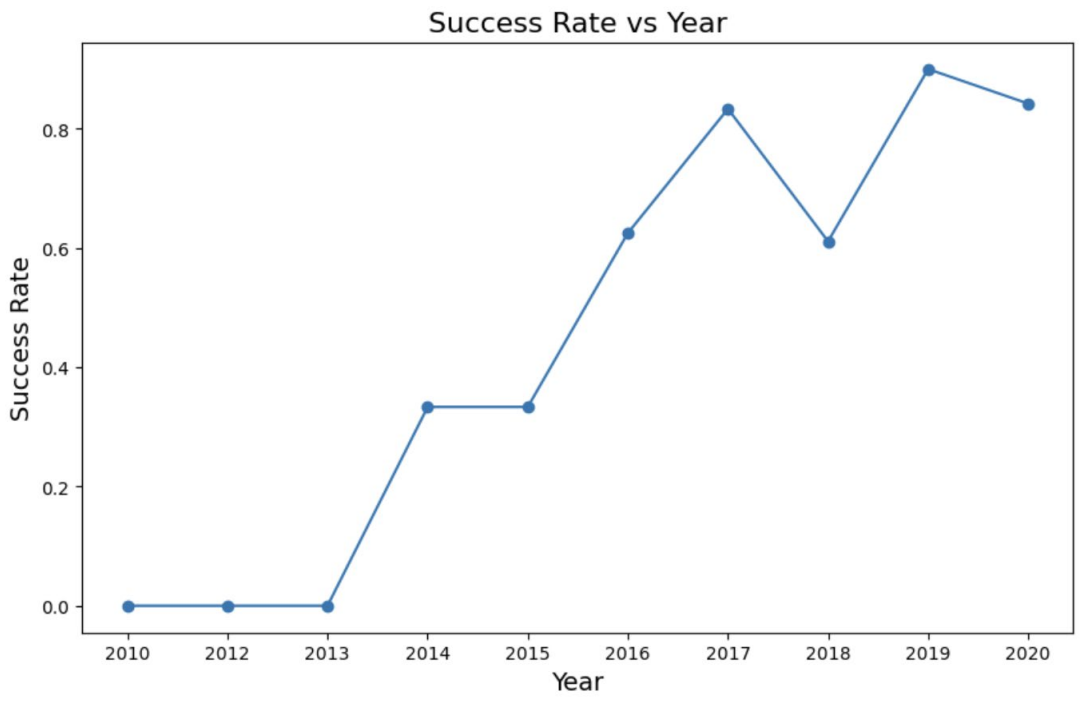
- Perform Exploratory Data Analysis (EDA):
  1. Calculate number and occurrence of each orbit
  2. Calculate number of launches on each site
  3. Calculate number and occurrence of mission outcome of the orbits
  4. Calculate landing outcome from outcome column
- All of those are to create quantitative data of how orbit, launch site will affect the landing outcome.



- GitHub URL: SpaceX_Project/Data_Wrangling_typical_pd_np.py at main · StunMan123/SpaceX_Project (github.com)

EDA with Data Visualization

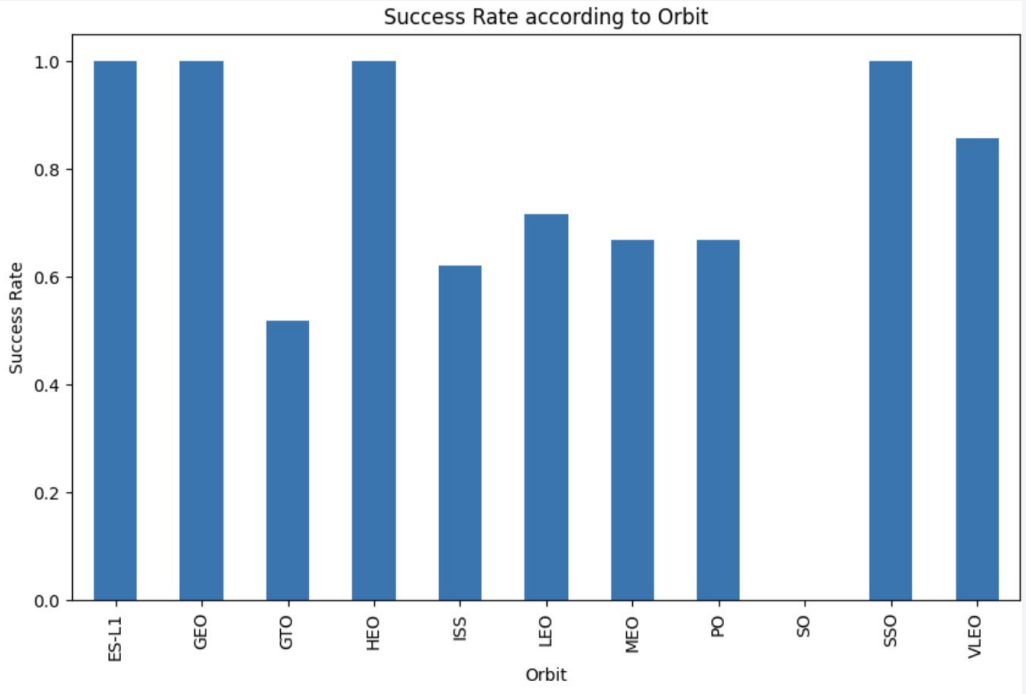
- Line plot: To show trend of success rate over the year



- GitHub URL: SpaceX_Project/EDA_pd_matplotlib.py at main · StunMan123/SpaceX_Project (github.com)

EDA with Data Visualization

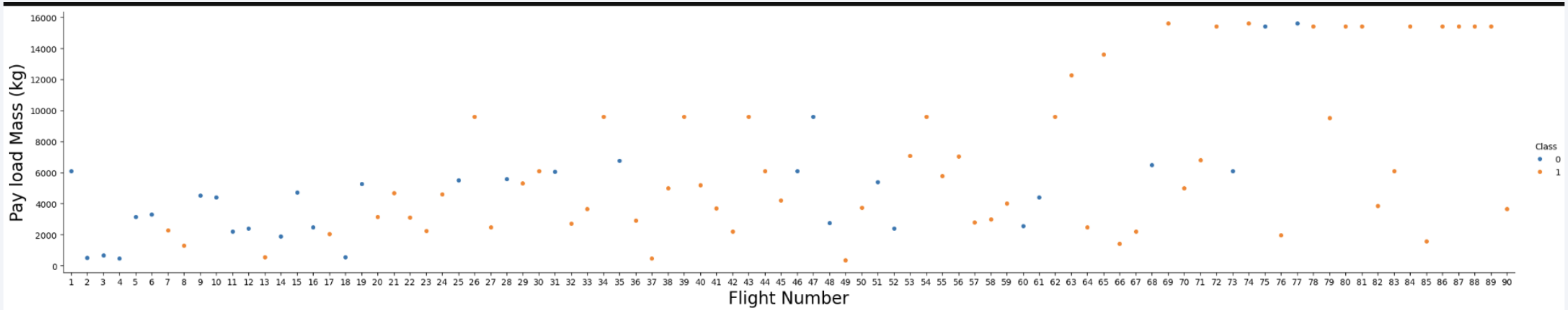
- Bar chart: To show success rate of each orbit.



- GitHub URL: [SpaceX_Project/EDA_pd_matplotlib.py at main · StunMan123/SpaceX_Project (github.com)]

EDA with Data Visualization

- Scatter plot: To relationship between payload mass and flight number and how they affect success rate.



- GitHub URL: [SpaceX_Project/EDA_pd_matplotlib.py at main · StunMan123/SpaceX_Project (github.com)]

EDA with SQL

SQL queries were performed to extract:

1. Names of unique launch sites
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

- GitHub URL: SpaceX_Project/EDA_SQL.py at main · StunMan123/SpaceX_Project (github.com)

EDA with SQL

SQL queries were performed to extract:

7. List the total number of successful and failure mission outcomes

8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

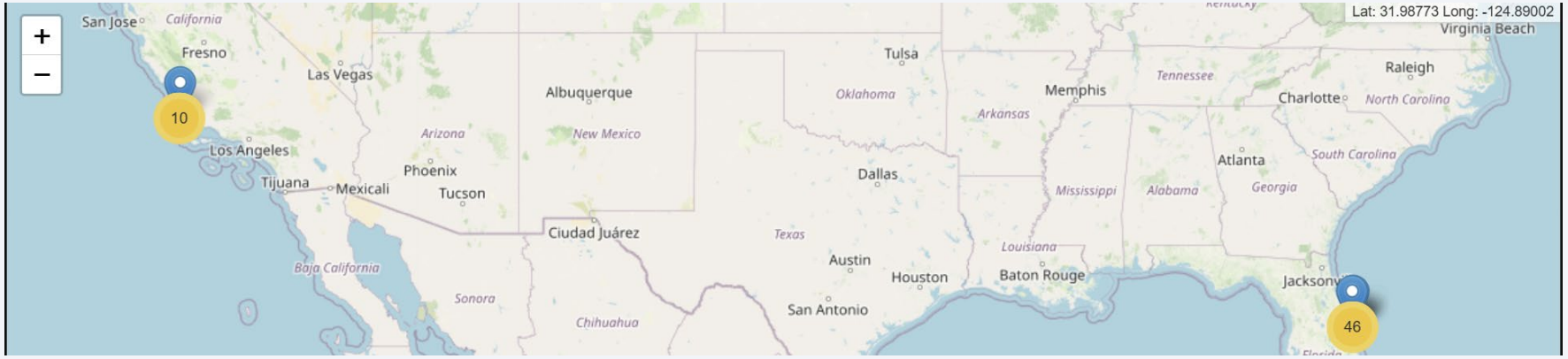
9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

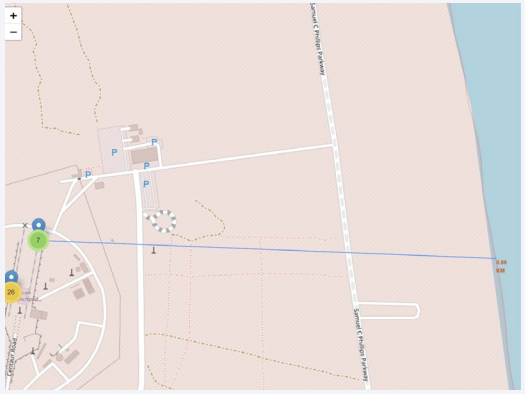
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- GitHub URL: SpaceX_Project/EDA_SQL.py at main · StunMan123/SpaceX_Project (github.com)

Build an Interactive Map with Folium

- markers, clusters, lines, mouse position were created and added to a folium map.
    1. Markers: To show the position and name of the launch site.
    2. Clusters: To make the visualization of occurrence of launch in launch site concise and clear.
    3. Mouse position: Allow user to know the coordinate by hover the mouse over the location.
    4. Lines: Know the distance between the launch site and closest coastline.





- GitHub URL: [SpaceX_Project/Interactive_Folium_Map.py at main · StunMan123/SpaceX_Project (github.com)]

Build a Dashboard with Plotly Dash

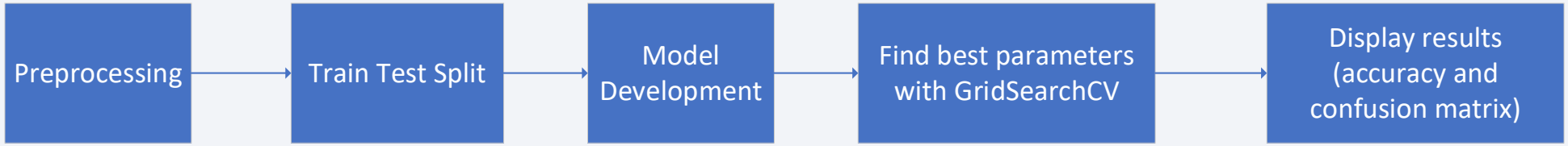
- Site dropdown, payload slider, pie-chart and scatter plot have added to a dashboard:
    1. Site dropdown: Allow user to specify the launch site
    2. Payload slider: Allow user to specify the range of payload mass of the rocket
    3. Pie-chart: Visualize the success rate based on the launch site user choose
    4. Scatter plot: Visualize the correlation between payload mass and success rate based on which launch site the user have chosen

- GitHub URL: SpaceX_Project/Interactive_Plotly.py at main · StunMan123/SpaceX_Project (github.com)

Predictive Analysis (Classification)

Predictive Analysis:

1. Preprocessing, it is to remove mean and scaling to unit variance, thus making data easier to analyse.
2. Split the data into train and test data.
3. Create machine learning model object
4. Find the best parameters for the model by using GridSearchCV
5. Display the result and confusion matrix of the model using the best parameters.



- GitHub URL: SpaceX_Project/Predictive_Analysis.py at main · StunMan123/SpaceX_Project (github.com)

Results

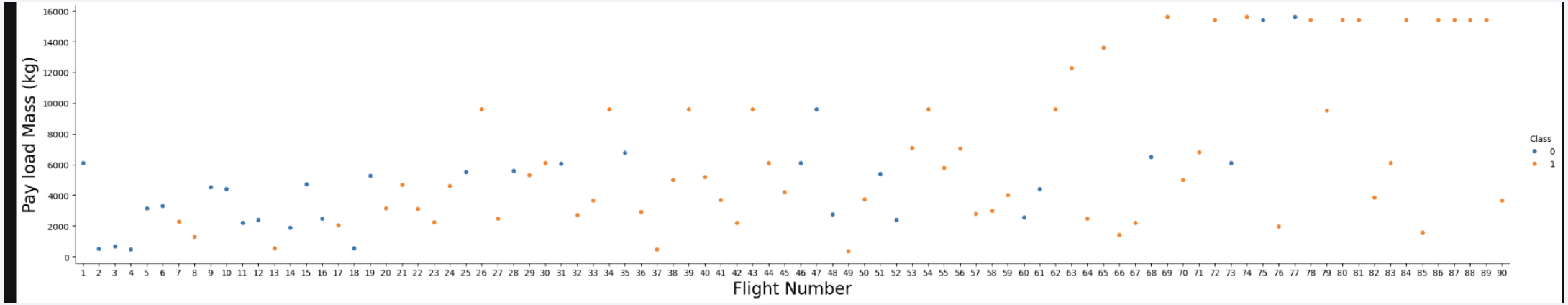
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

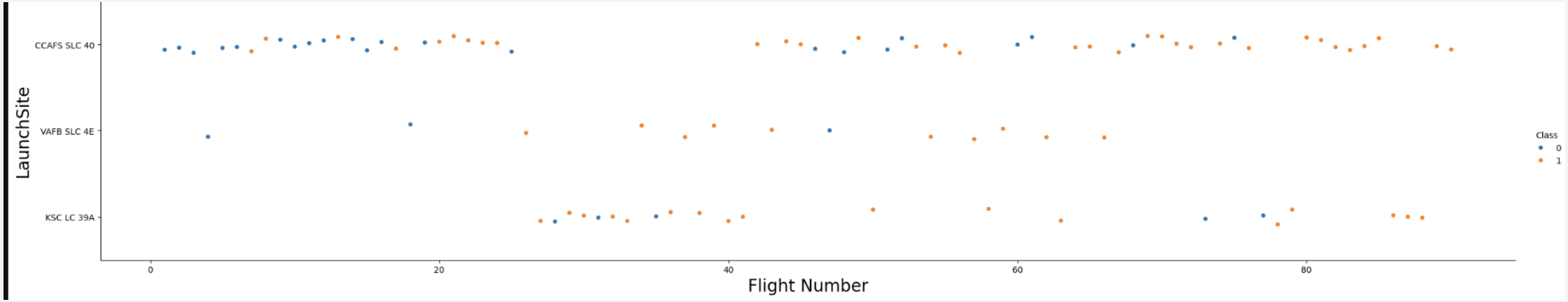
- Scatter plot of Flight Number vs. Launch Site



As flight number and payload mass increases, the success rate also increases. This is because the flight number is corresponded to the amount of data collection and technology advancement of SpaceX. So, as technology advanced and enough data to study, SpaceX rocket can accommodate higher payload and with higher success rate.

Payload vs. Launch Site

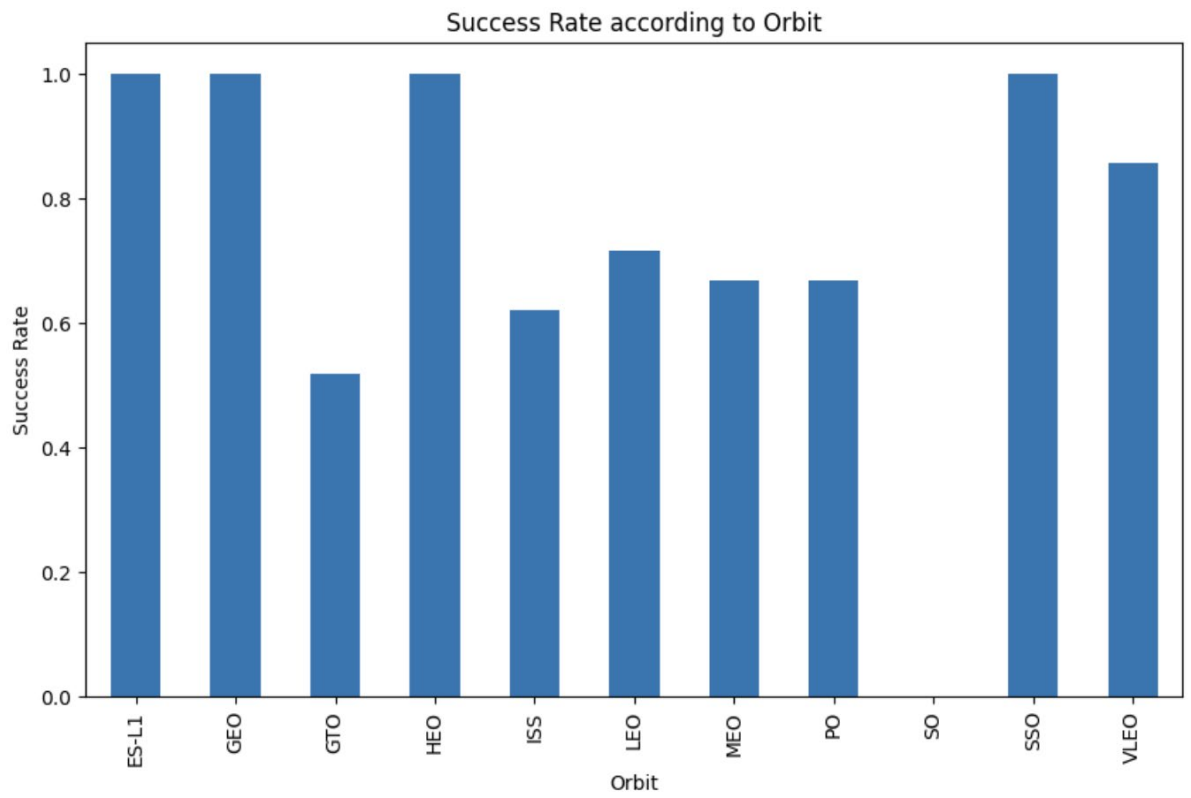
Scatter plot of Payload vs. Launch Site



Launch Site CCAFS SLC 40 has the highest success rate

Success Rate vs. Orbit Type

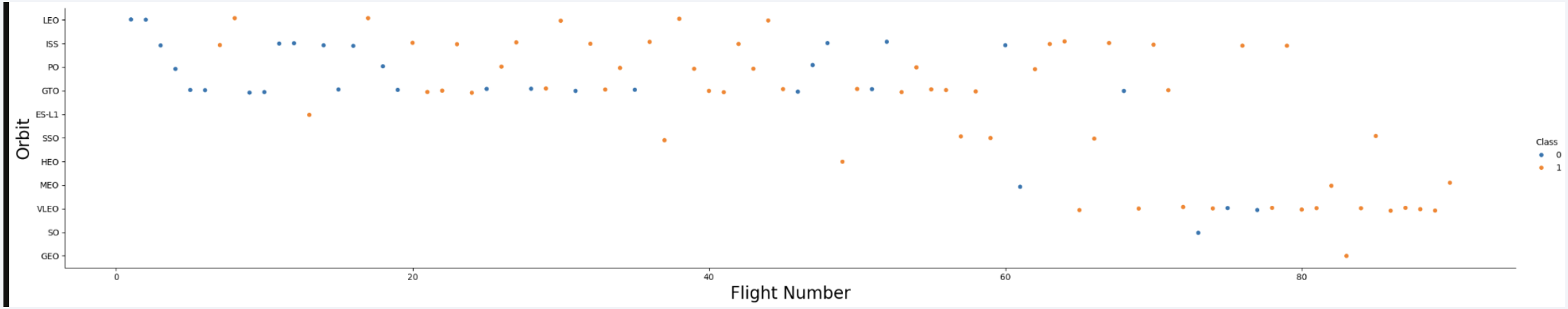
Bar chart for the success rate of each orbit type



Orbit ES-L1, GEO, HEO, SSO has the highest success rate, while SO has 0 success rate.

Flight Number vs. Orbit Type

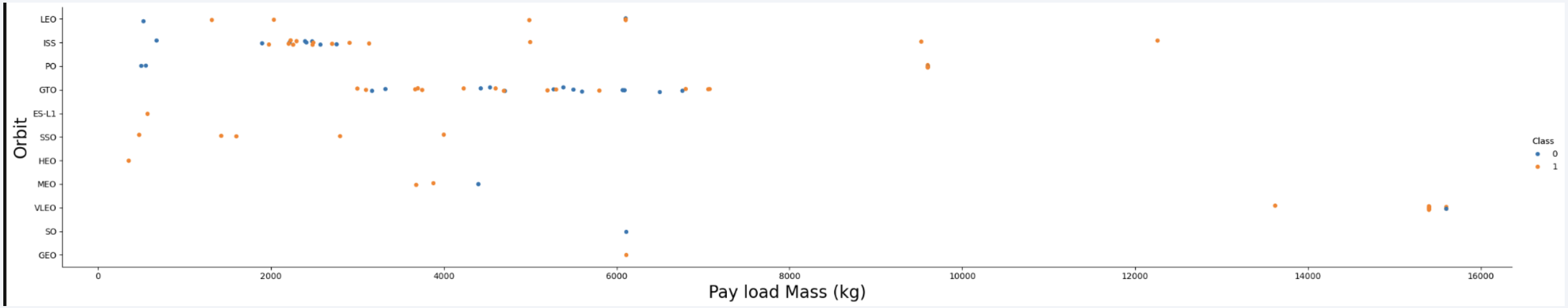
Scatter point of Flight number vs. Orbit type



Orbit VLEO has the highest success rate at higher flight number range

Payload vs. Orbit Type

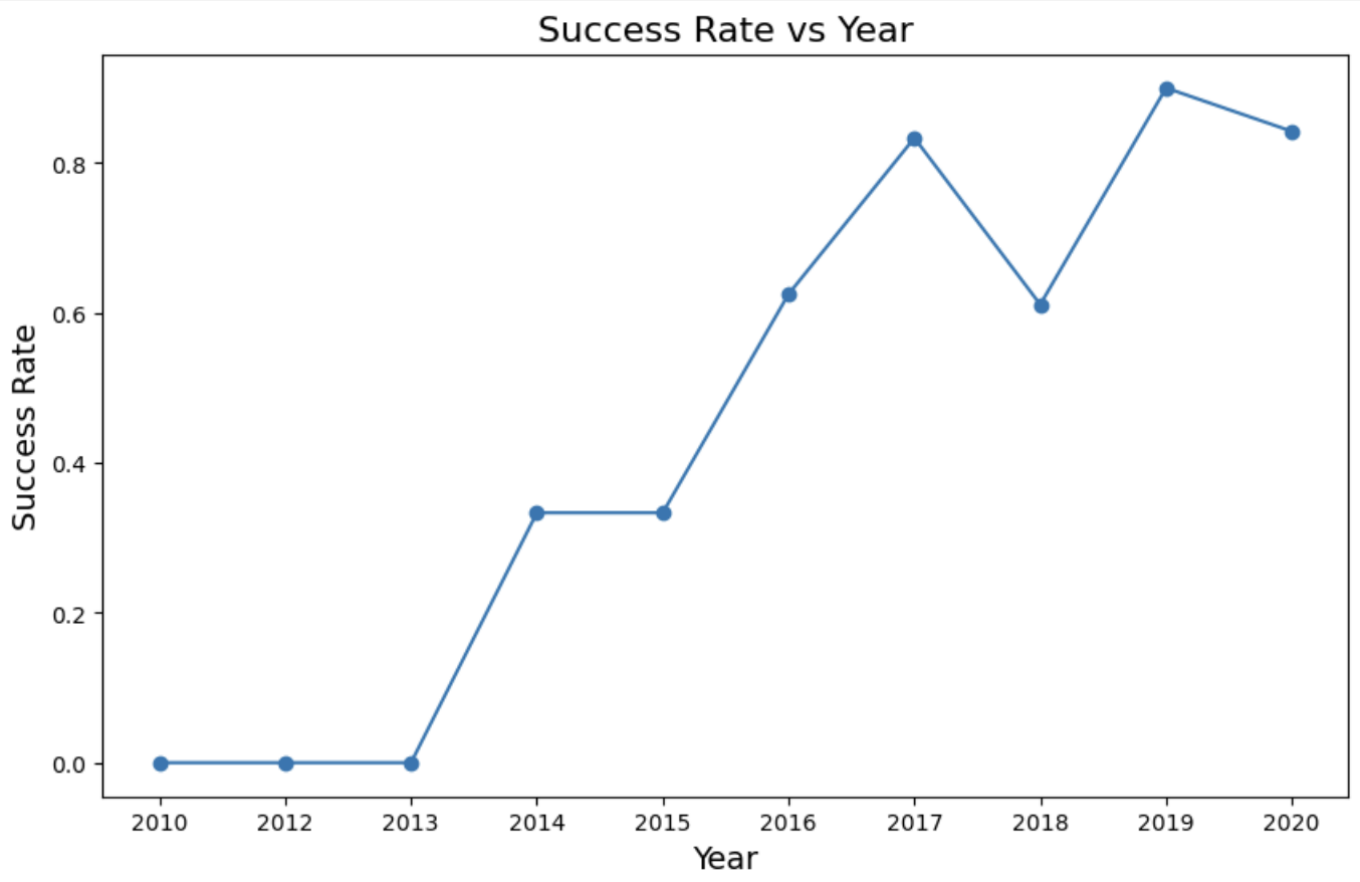
Scatter plot of payload vs. orbit type



- Orbit VLEO can accommodate more high pay load (15000 kg – 16000 kg)
- Orbit GTO better accommodate medium range pay load (3000 kg to 7000 kg)
- Orbit ESS better accommodate low range pay load (2000 kg to 4000 kg)

Launch Success Yearly Trend

Line chart of yearly average success rate



Success rate increases with year, meaning the technology of the SpaceX are increasingly safe and stable over the year to ensure high success rate.

All Launch Site Names

- Names of the unique launch sites:

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

- Query result:

Launch_Site	Frequency
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

Launch Site CCAFS SLC-40 is launch site with highest number of launches.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Query result:

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

The first 5 earliest record where launch sites begin with `CCA` were : ‘CCAFS LC – 40’

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Query result:

total payload mass
48213

NASA (CRS) booster has transported total payload mass of 48213 kg over the year

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Query result:

average payload mass
2534.67

F9 v1.1 booster has transported average payload mass of 2534.67 kg over the year

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Query result

Date of first successful landing
2015-12-22

The date of first successful landing on ground pad was at 22nd December 2022

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
- Query result

success boosters
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Only 4 type of boosters were successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
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
- Query result:

Landing_Outcome	Frequency
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Overall, number of success outcome is 61, number of failure outcome is 10.

Boosters Carried Maximum Payload

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

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

The booster which have carried the maximum payload mass are from F9 B5 series

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Query result:

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

Outcome with failure landing on drone ship were from launch site CCAFS LC-40 and with Booster F9 v1.1 series.

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

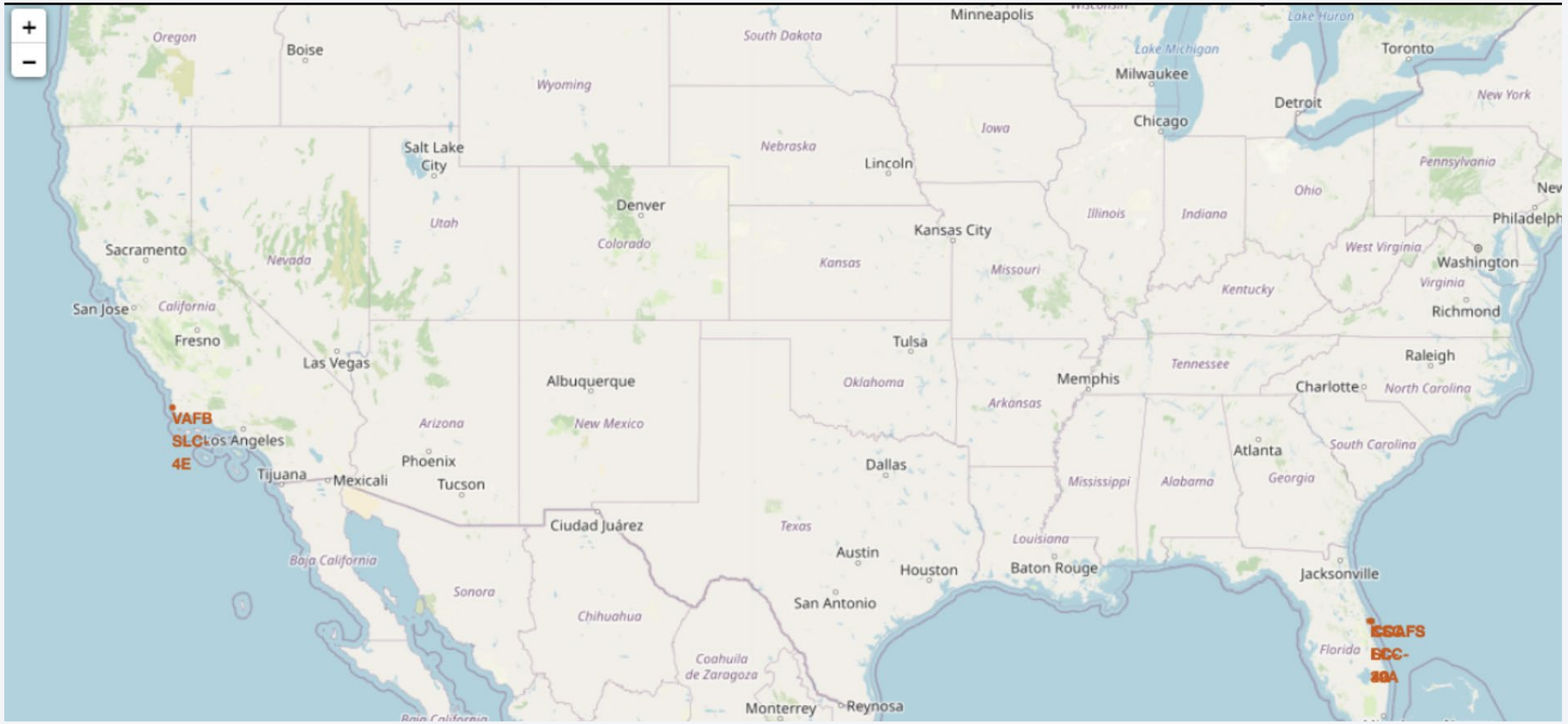
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Query result:

Landing_Outcome	Frequency
Failure (drone ship)	5
Success (ground pad)	3

There are 5 cases of failure landing on drone ship and 3 cases of success landing on ground pad.

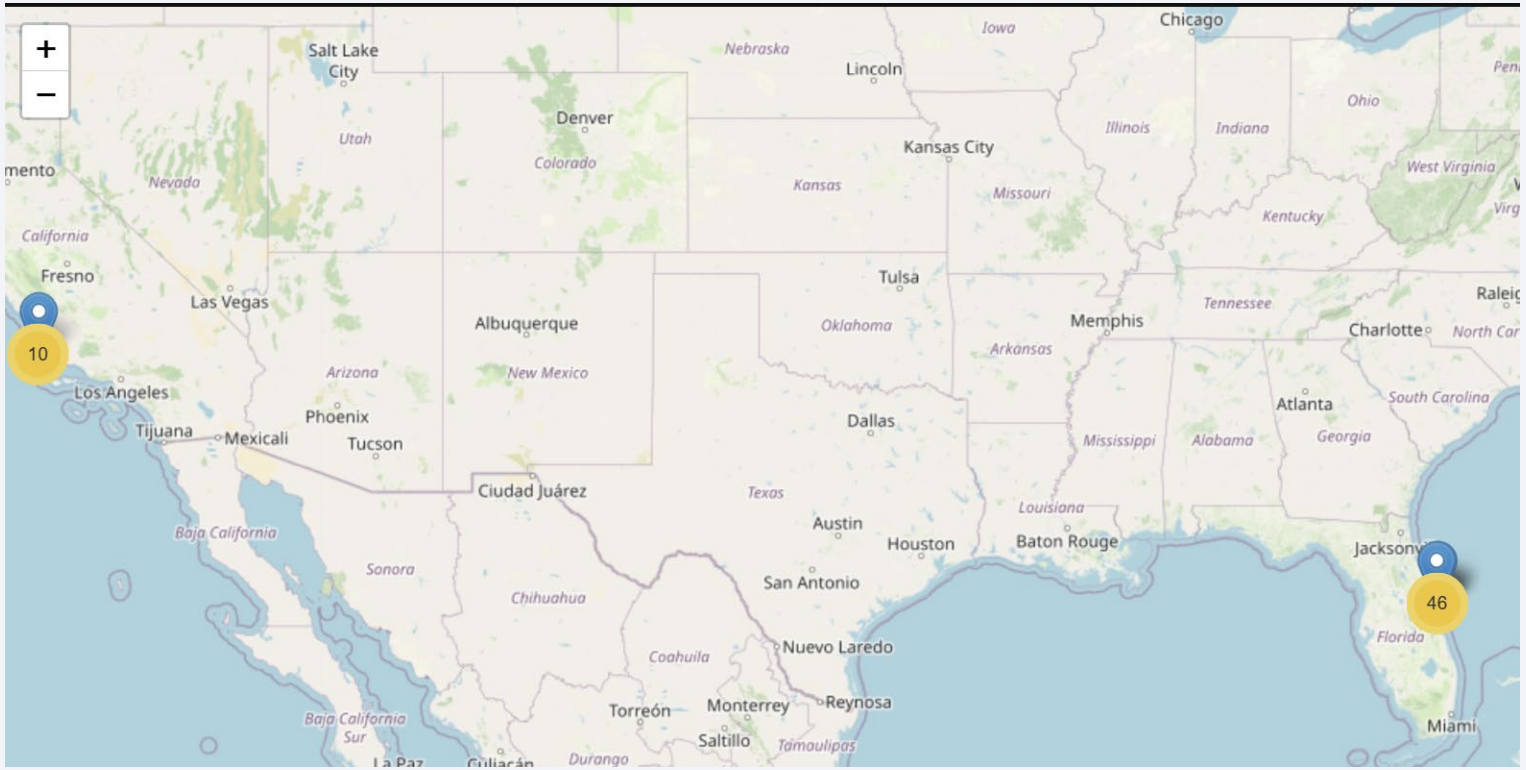

Section 3
Launch Sites
Proximities Analysis

All Launches Sites' location markers on global Map



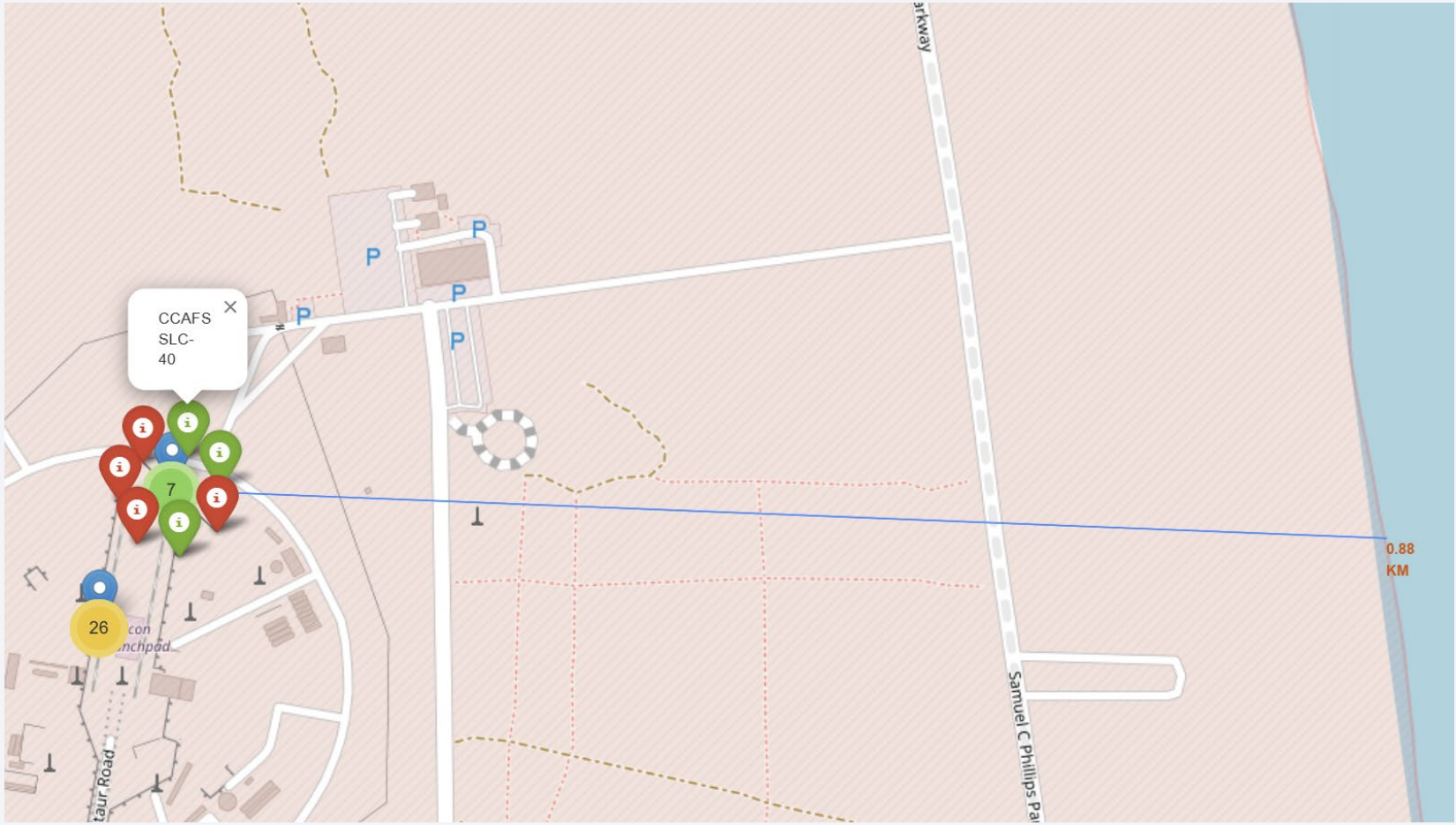
All the launches were in United States, focuses on west coast and Florida

All Launches Sites' location markers on global Map with labeled Outcomes

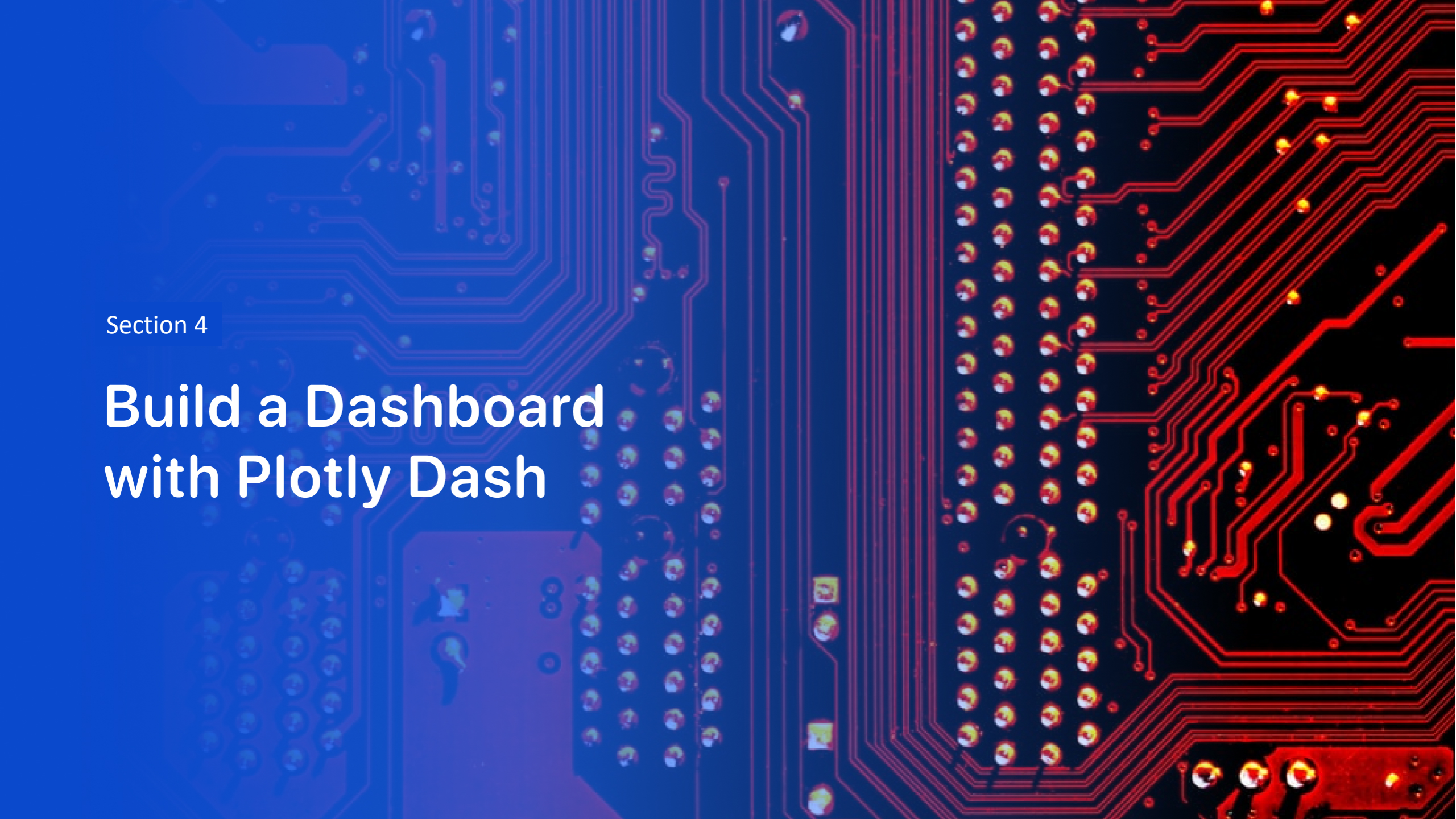


There are 10 launches on west coast and 46 launches on Florida

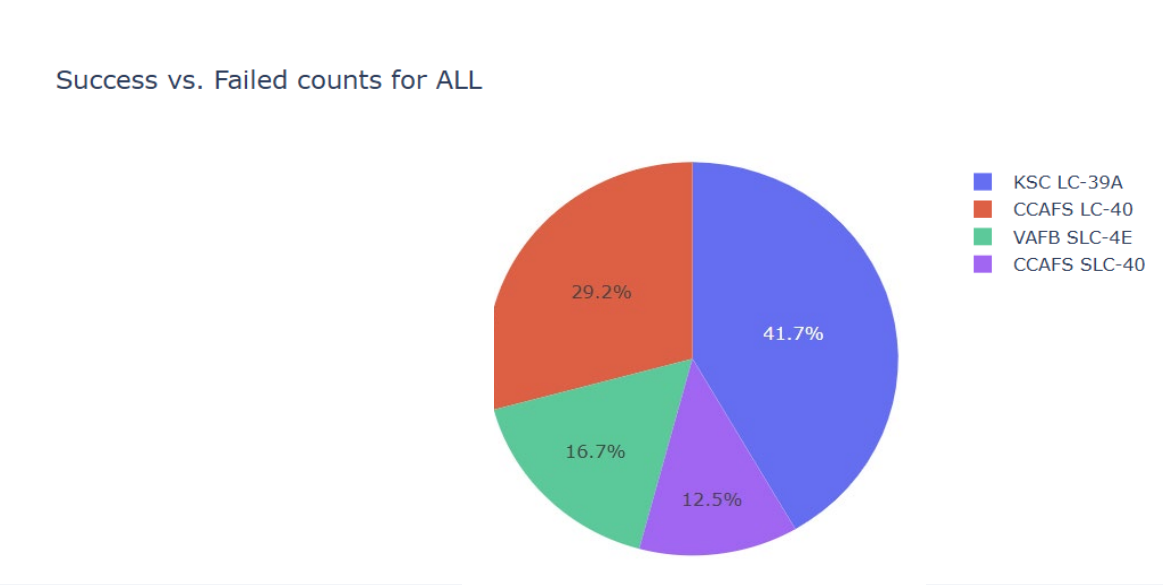
CCAFS SLC – 40 Launch Site with its proximity coastline



There is only 0.88 km from CCAFS SLC – 40 launch site to its nearest coastline

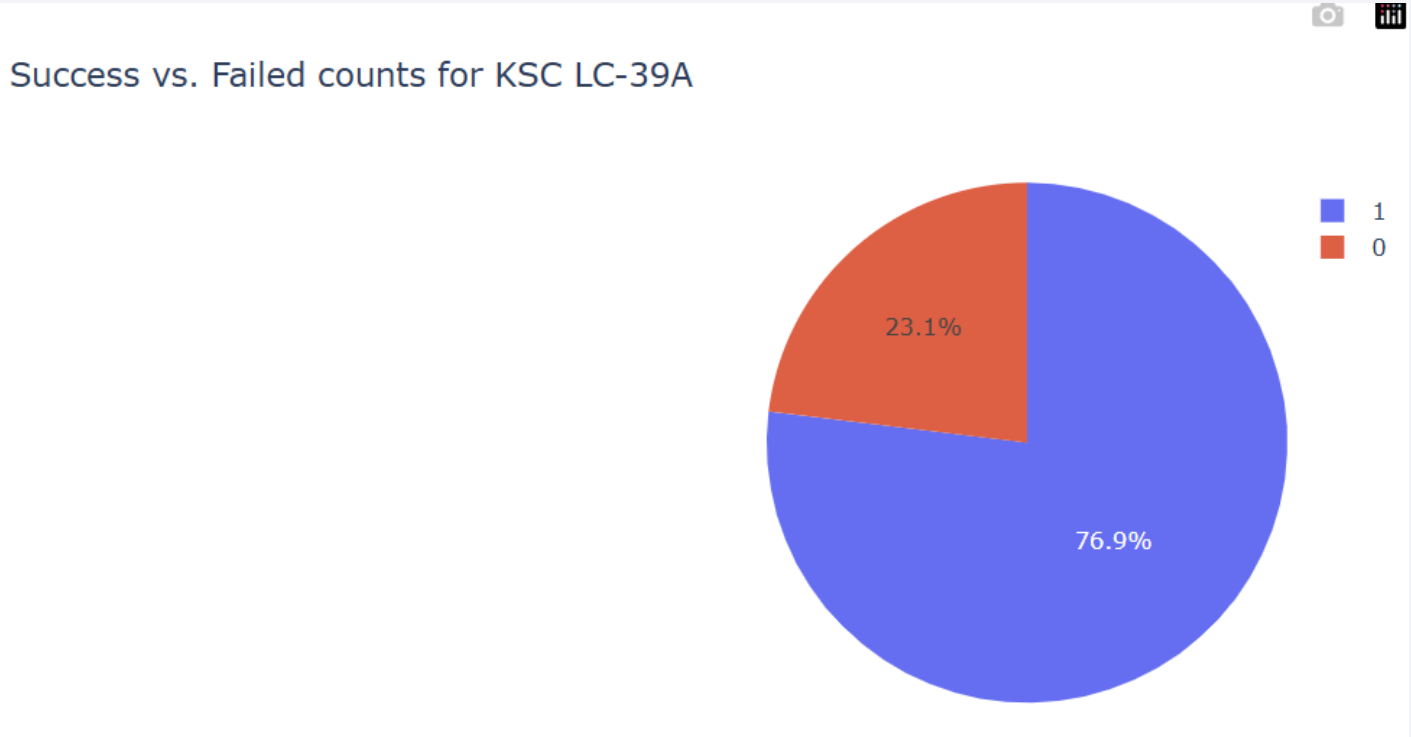

Section 4
Build a Dashboard with Plotly Dash

Pie Chart of SpaceX Launch Records For All Site



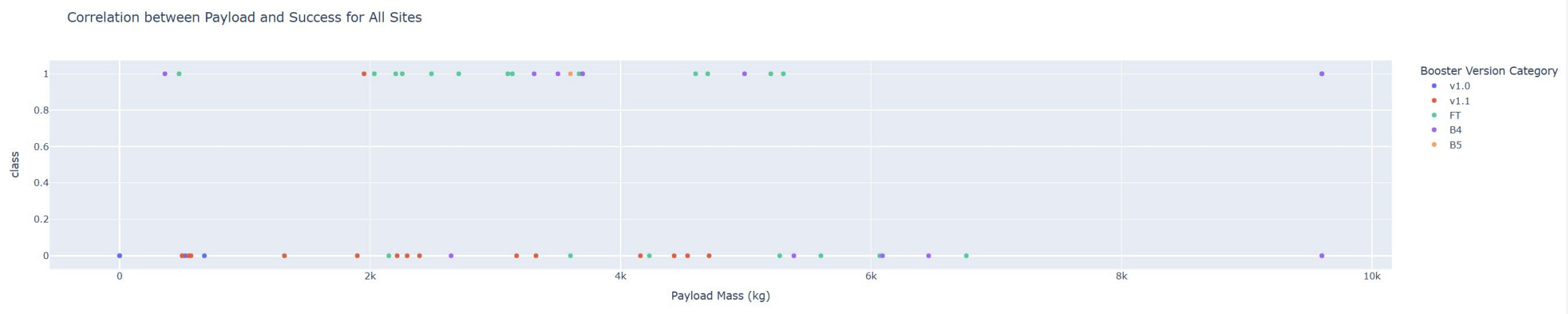
- KSC LC-39A is the launch site with most launches
- CCAFS SLC- 40 is the launch site with least launches

Pie Chart of SpaceX Launch Records For KSC LC -39A



KSC LC-39A is not only the launch site with most launches, it also has the highest success rate (76.9%)

Correlation between Payload and Success for All Sites

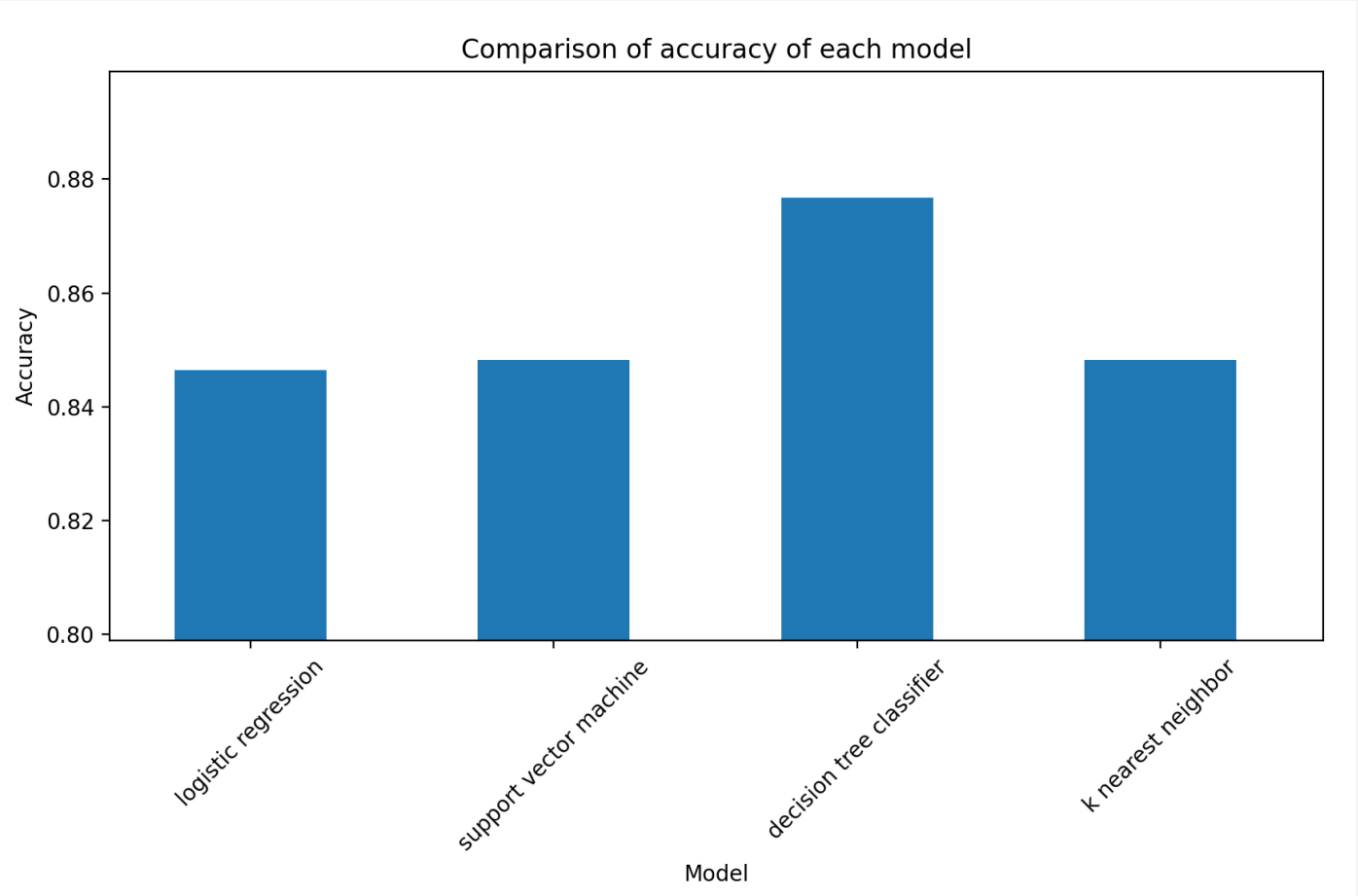


- For medium payload mass, Booster FT is the best because it has the highest success rate
- For high payload mass, it is preferrable to use Booster v.10
- For low payload mass, Booster v.10 and FT can be used, there is no single best choice

Section 5

Predictive Analysis (Classification)

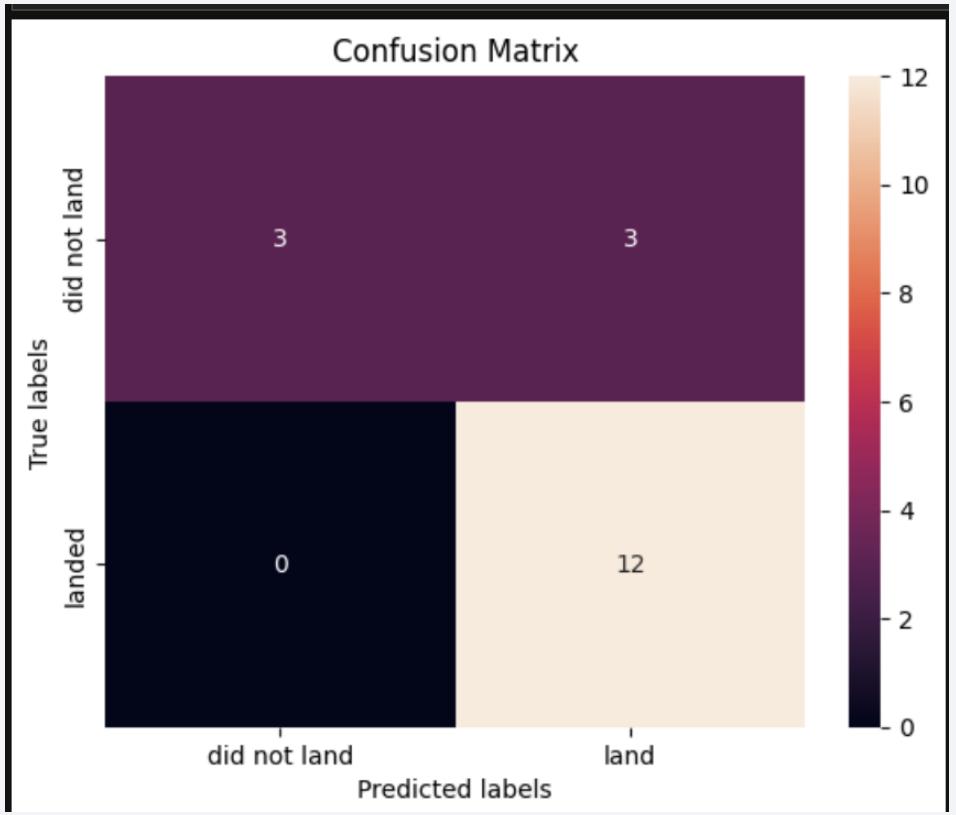
Classification Accuracy



Decision tree classifier has the highest accuracy (0.8768)

Confusion Matrix of Decision tree classifier

Confusion Matrix of Decision tree classifier



Examining the confusion matrix, we see that decision tree classifier can distinguish between the different classes. We see that the problem is false positives, which will cause wrong expectations and thus disasters

Overview:
True Postive - 12 (True label is landed, Predicted label is also landed)
False Postive - 3 (True label is not landed, Predicted label is landed)

Conclusions

Overall SpaceX has shown tremendous improvement over the years in terms of payload mass, success rate, and orbit, in which show the possibility of commercial space at affordable price and the contiuous improvement of SpaceX in terms of technology

1. Can the first stage of Falcon 9 be successfully reused?

Yes, the highest success rate that SpaceX achieved is 76.9%, thus, more than ¾ of the first stage of Falcon 9 can be reused.

2. How to determine the price of a SpaceX rocket launch?

By accounting for these 3 main factors : Payload Mass, Launch Sites, Orbit

3. What factors influence the successful landing of the first stage?

Payload Mass, Launch Sites, Orbit

Appendix

- Overall GitHub URL: StunMan123/SpaceX_Project: This project is for a data science courses (github.com)
- Plotly-Dash references:

dcc.Dropdown() => https://dash.plotly.com/dash-core-components/dropdown

dcc.RangeSlider() => https://dash.plotly.com/dash-core-components/rangeslider

pie chart => https://plotly.com/python/pie-charts/

scatter chart => https://plotly.com/python/line-and-scatter/

Thank you!