IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Guo Siong Chou
29 June 2023

Outline

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

Executive Summary

- Methodologies
  - Performed data collection and wrangling on SpaceX launch data of Falcon 9
  - Data analysis with visualisations, SQL queries, Folium and Plotly dashboard
  - Perform classification predictive analysis with Logistic Regression, SVM, Decision Tree, KNN

- Results
  - All four algorithms practically have the same accuracy
  - More evaluation required, evaluate based on desired use case

Introduction

- On its website, SpaceX promotes Falcon 9 rocket launches at a price of $62 million, whereas other providers charge over $165 million for each launch. The significant cost savings are primarily attributed to SpaceX's ability to reuse the first stage of the rocket.
- Consequently, by ascertaining whether the first stage will successfully land, we can calculate the total expense of a launch.
- This information becomes valuable when an alternative company intends to compete with SpaceX for a rocket launch contract.
- A machine learning pipeline was created to predict if the first stage would land successfully.

Section 1

Methodology

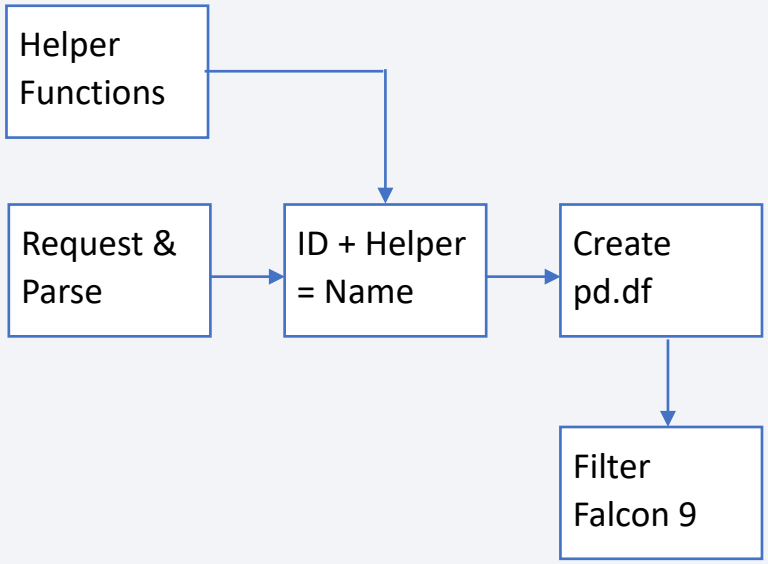
Methodology

Executive Summary

- Data collection methodology:
  - Data collected through SpaceX API and Web Scraped from Wiki
- Perform data wrangling:
  - Filter for Falcon 9 data, dealt with missing values
- Perform exploratory data analysis (EDA) using visualization and SQL:
  - Perform prelim exploration on the data within python notebook
- Perform interactive visual analytics using Folium and Plotly Dash:
  - Data visualisation and analysis of visual representation of launch data.
- Perform predictive analysis using classification models:
  - Normalise data, train_test_split(), GridSearchCV(), Train model, evaluate models

Data Collection – SpaceX API

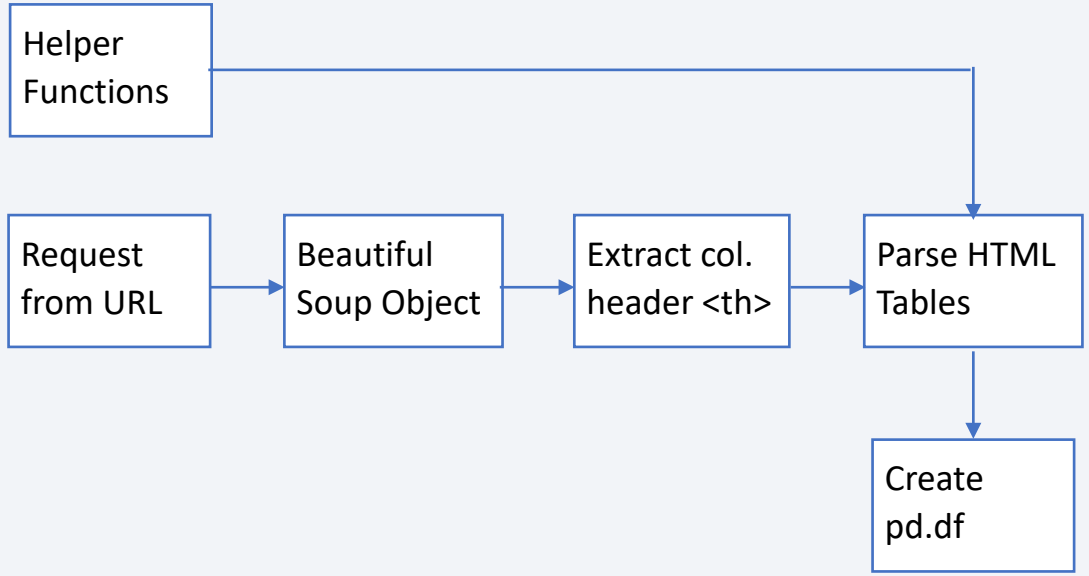
- Define helper API functions
- Request & Parse launch data from SpaceX API
- Using helper function API to extract information using ID numbers in launch data
- Create Pandas df by combining extracted columns into dict



GitHub URL: https://github.com/guosiong/Applied-Data-Science-Capstone/blob/14c49af85bb4a5c32d27b313690151d8c557888b/Data%20Collection%20API.ipynb

Data Collection – Scraping

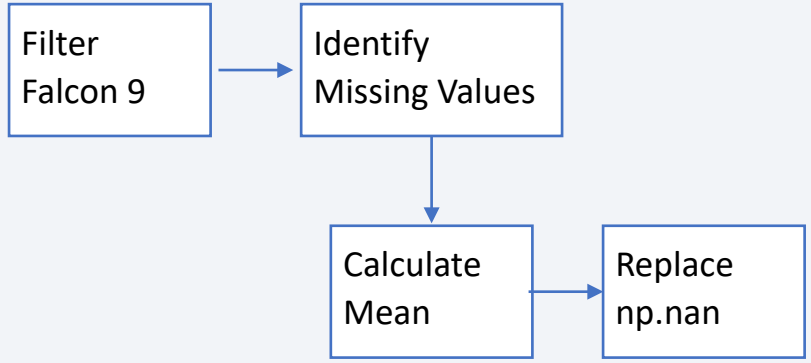
- Define helper functions
- Request from Wiki URL
- BeautifulSoup object from HTML response
- Extract column header from HTML <th> elements
- Parse HTML tables with helper functions
- Create pd dataframe



GitHub URL: https://github.com/guosiong/Applied-Data-Science-Capstone/blob/14c49af85bb4a5c32d27b313690151d8c557888b/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Filter include Falcon 9 data only
- Identify variable(s) with missing values that must be dealt with
- Fill missing values with mean
  - Calculate mean of PayloadMass
  - Replace np.nan values with mean



GitHub URL: https://github.com/guosiong/Applied-Data-Science-Capstone/blob/14c49af85bb4a5c32d27b313690151d8c557888b/Data%20Wrangling.ipynb

EDA with Data Visualization

- Scatter Plot to see how variables would affect the launch outcome
  - FlightNumber vs. PayloadMass
  - FlightNumber vs. LaunchSite
  - Payload vs. LaunchSite
- Bar chart to visualise success rate of each orbit type
  - Success Rate vs. Orbit type
- Scatter plot to see any relation between variables
  - FlightNumber vs. Orbit type
  - Payload vs. Orbit type
- Line chart to visualise the launch success yearly trend

GitHub URL: https://github.com/guosiong/Applied-Data-Science-Capstone/blob/14c49af85bb4a5c32d27b313690151d8c557888b/EDA%20with%20Visualization.ipynb

EDA with SQL

- SQL queries performed
  - Names of the unique launch sites in the space mission
  - 5 launch site names string beginning with ‘CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date of first successful landing in ground pad
  - Names of the boosters with successful landing in drone ship having payload between 4000 to 6000
  - List of successful and failure mission outcome
  - Names of the booster_version that carried the maximum payload mass
  - Records of month names, failure landing outcomes in drone ship, booster versions, launch site, for the months in year 2015
  - Ranked count of successful landing outcomes between date 04-06-2010 to 20-03-2017

GitHub URL: https://github.com/guosiong/Applied-Data-Science-Capstone/blob/14c49af85bb4a5c32d27b313690151d8c557888b/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Map objects added
  - Added markers to mark the 4 launch sites with each names.
  - Circle around the launch sites, encircling the features in its proximities.
  - Used clusters to show the number of launches within the sites, zooming in to show the numbers for respective sites, specifically for the East coast.
  - Added polyline to mark the distance measured to the interested features in the proximities of the launch sites.

GitHub URL: https://github.com/guosiong/Applied-Data-Science-Capstone/blob/14c49af85bb4a5c32d27b313690151d8c557888b/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

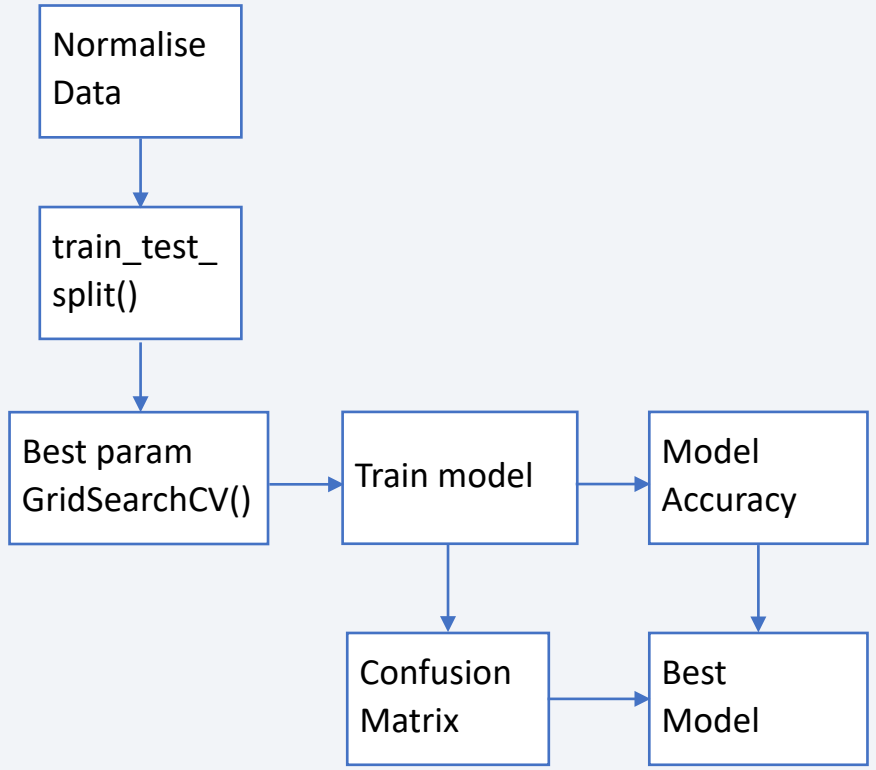
Build a Dashboard with Plotly Dash

- Pie chart of launch success ratio of all sites and success/failure ratio of each site.
  - Drop down to select “All Sites” or individual launch site to visualise with the pie chart.
  - For visualising proportion of successful launch from each launch site.
  - For visualising success/failure launch ratio at each launch site.
- Scatter plot of launch outcome vs. Payload.
  - Marker coloured in each booster versions.
  - Range slider for selecting payload range to display in the scatter plot.
  - For visualising payload range and booster version relation with launch outcome.

GitHub URL: https://github.com/guosiong/Applied-Data-Science-Capstone/blob/98d1d790eef8437f9cf68022bad134383033dac6/Build%20a%20Dashboard%20Application%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

- Normalise data to a similar scale for performance and training stability of the ML model.
- Split the data with train_test_split() function
- For each of the four ML model (Logistic Regression, SVM, Decision Tree, KNN), the below steps are performed:
  - Finding the best parameters with GridSearchCV()
  - Train model with best parameters
  - Calculate the model accuracy
  - Plot the confusion matrix
- Determine the best performing model based on the accuracy



GitHub URL: https://github.com/guosiong/Applied-Data-Science-Capstone/blob/0baa19bec3a1c27ee71065c621f7716c6b3683ea/Machine%20Learning%20Prediction.ipynb

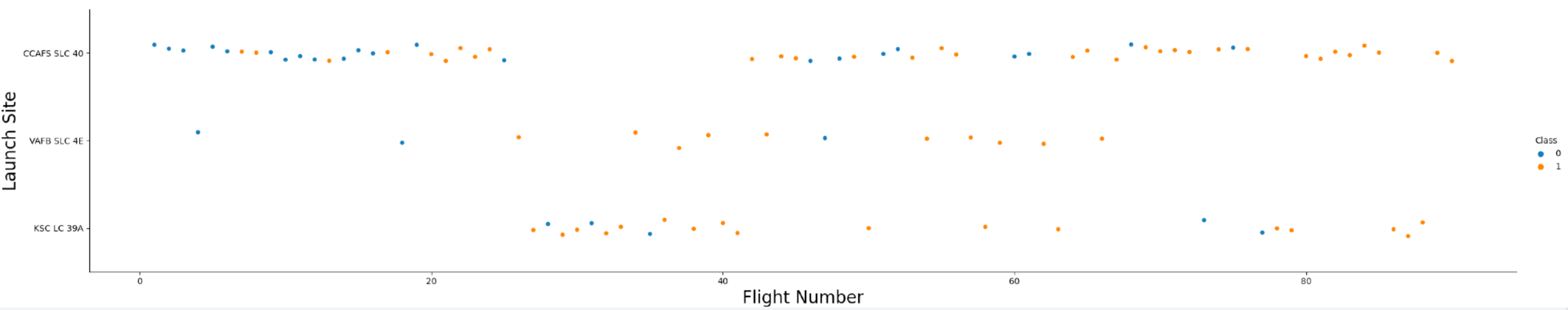
Results

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

Section 2

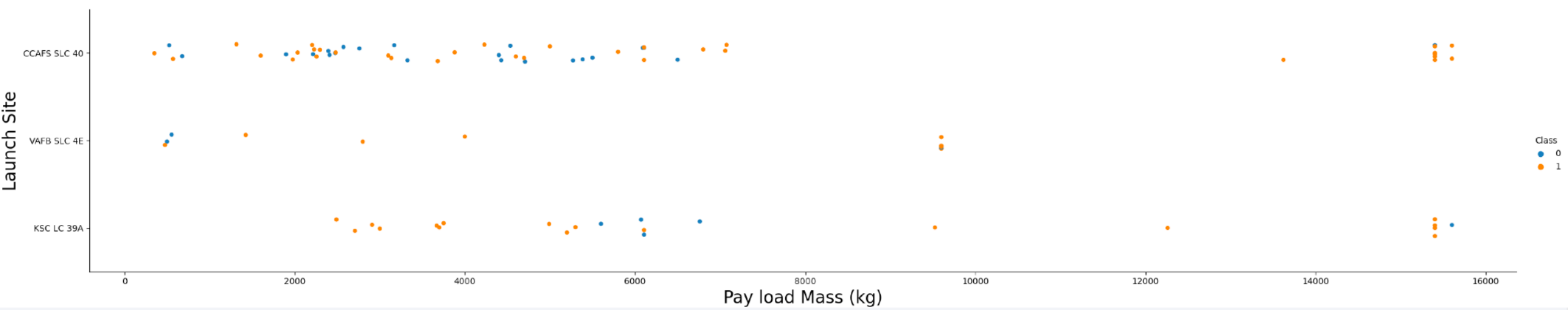
Insights drawn from EDA

Flight Number vs. Launch Site



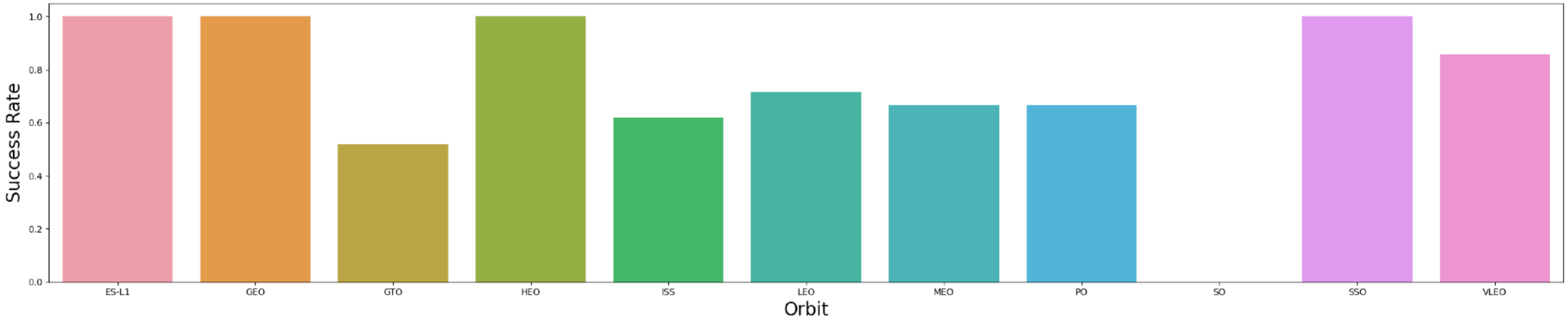
- As the flight number increases the probability of successful landing increases across launch sites.
- Launch were predominantly from CCAFS SLC 40, except around flight 25 to 40 was done at KSC LC 39A.
- Launch frequency from VAFB SLC 4E was sporadic.

Payload vs. Launch Site



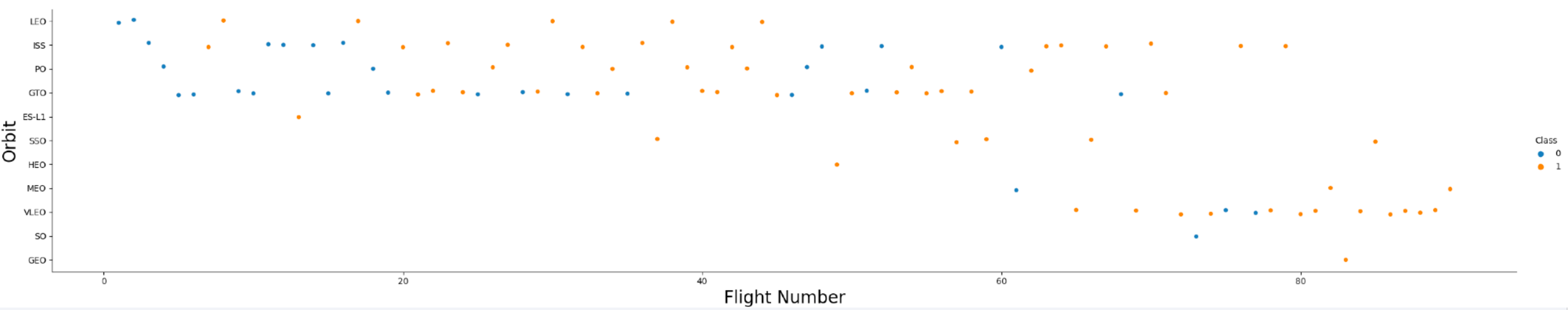
- Most launches are from the CCAFS SLC 40 launch site.
- Higher the payload mass, higher probability of successful landing.
- Seen from the scatter, the payload mass is in two distinct clusters, under 8000 and above 15000.

Success Rate vs. Orbit Type



- Launches into the ES-L1, GEO, HEO and SSO orbit relates to an almost 100% success rate.

Flight Number vs. Orbit Type

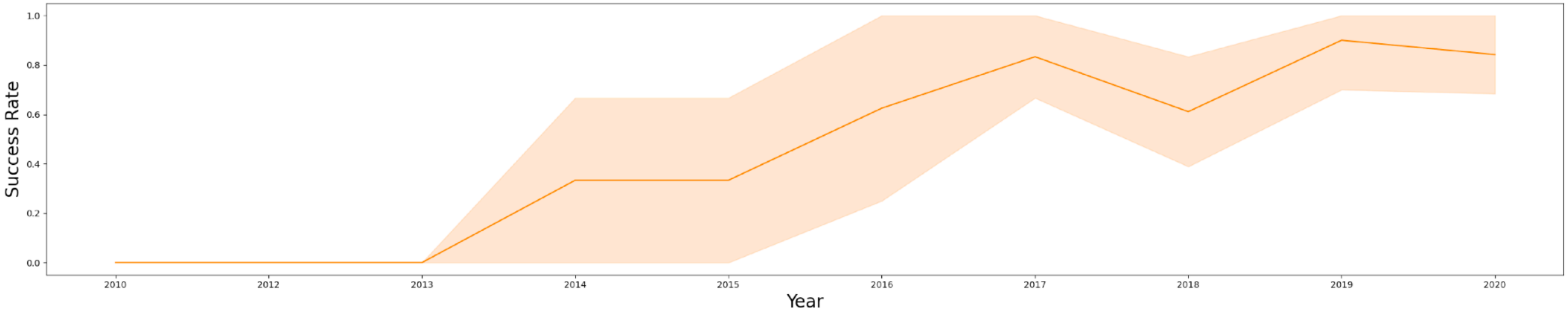


- The success rate across launches into various orbits improves with the increase of flight numbers
- Large number of earlier flights are to the LEO, ISS, PO and GTO orbits, while later flights are more focused to VLEO orbit.

Payload vs. Orbit Type

- A significant number of launches are to the ISS and GTO.
- Payloads to ISS have a spread of between 2000 to 3000kg.
- Payloads to GTO have a spread of between 3000 to 7000kg.

Launch Success Yearly Trend



- The success rate of launches increases over the years.
- The range of success rate narrows down over the years.

All Launch Site Names

- This query result shows the list of launch sites names.

Out[26]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- This query result shows 5 records where launch sites begin with `CCA`.

Out[27]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fai
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fai
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- This query result shows the total payload carried by boosters from NASA.
- 45,596kg total payload

Out[28]:

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- This query result shows the average payload mass carried by booster version F9 v1.1.

Out[29]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- This query shows the date of the first successful landing outcome on ground pad.

Out[31]:

MIN(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query result list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Out[32]:

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

Total Number of Successful and Failure Mission Outcomes

- This query results shows the total number of successful and failure mission outcomes. 98 success, 2 failure.

Out[41]:

Mission_Outcome	COUNT(Mission_Outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- This query result list the names of the booster which have carried the maximum payload mass.

Out[43]:

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

- This query result list the failed landing_outcomes in drone ship, with detail of their booster versions, and launch site names and month for the year 2015.

Out[68]:

Year	Month	Landing_Outcome	Booster_Version	Launch_Site
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
2015	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40

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

- This query result ranks 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.

Out[72]:

Landing_Outcome	Quantity
Success	20
Success (drone ship)	8
Success (ground pad)	7

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites

- The launch sites are located in close proximity to the equator line.
- Launch sites are on the coastal regions, west and east coast.

SpaceX Launch Sites

West coast

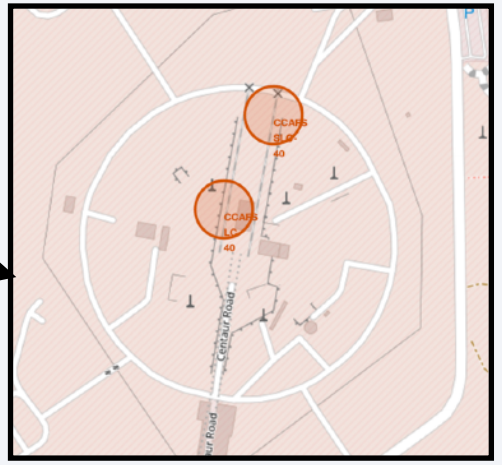
Vandenberg Space Force Base

- VAFB SLC-4E

East coast

- KSC LC-39A

- CCAFS LC-40
- CCAFS SLC-40



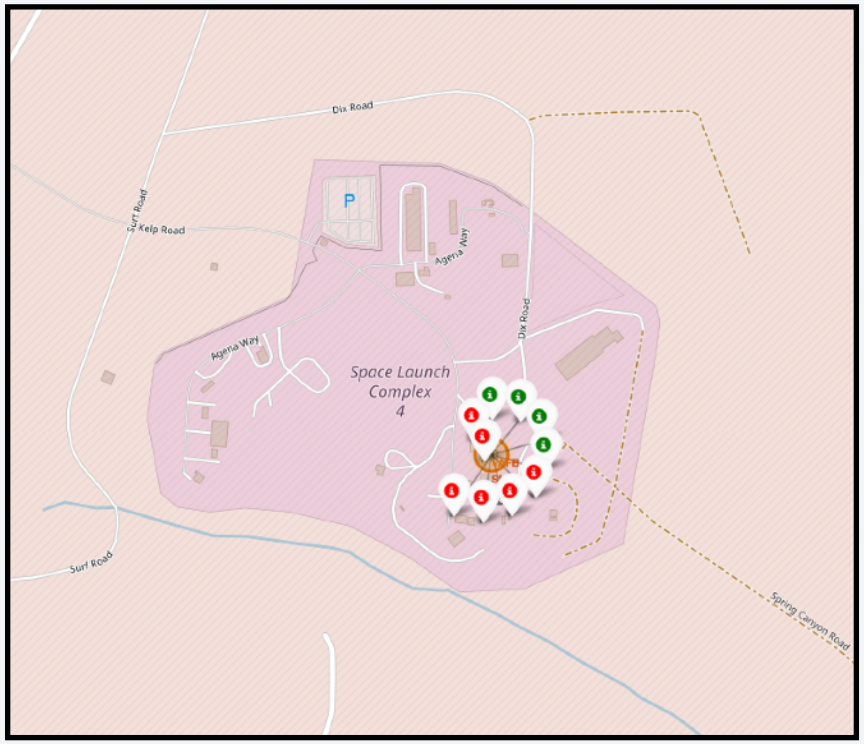
SpaceX Success and Failed Launches

The number of launches on West and East coast launch sites.

SpaceX Success and Failed Launches

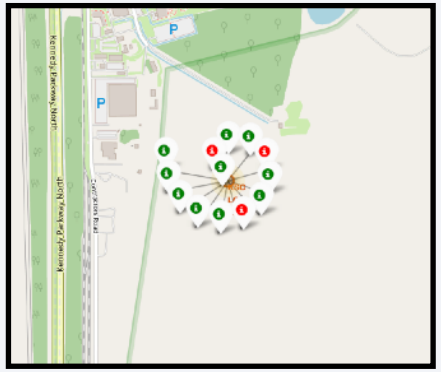
West coast

- VAFB SLC-4E



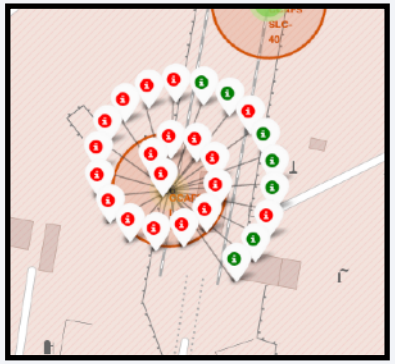
East coast

- KSC LC-39A
- This launch site has relatively higher number of successful launches

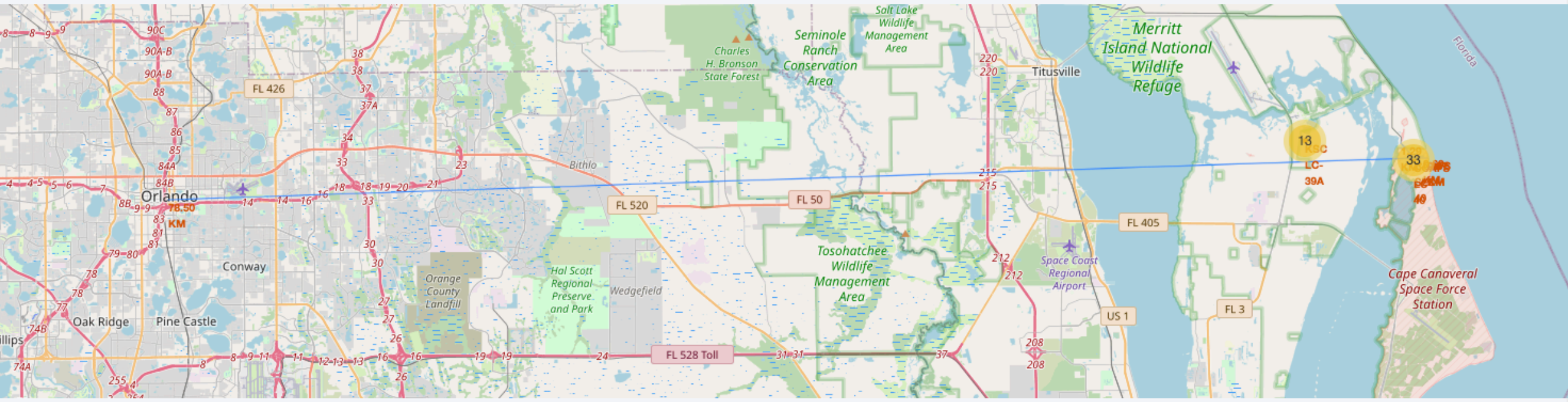


- CCAFS SLC-40

- CCAFS LC-40

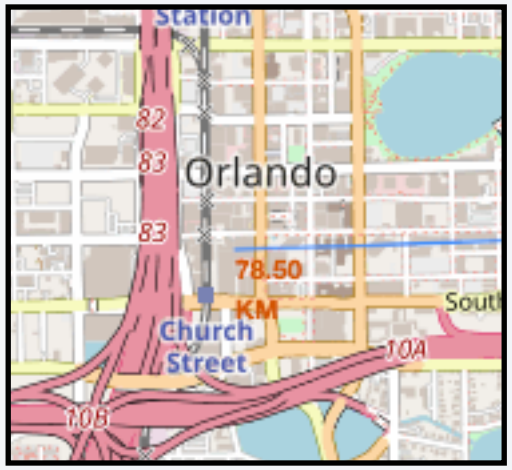


Distance Between Launch Sites and its Proximities



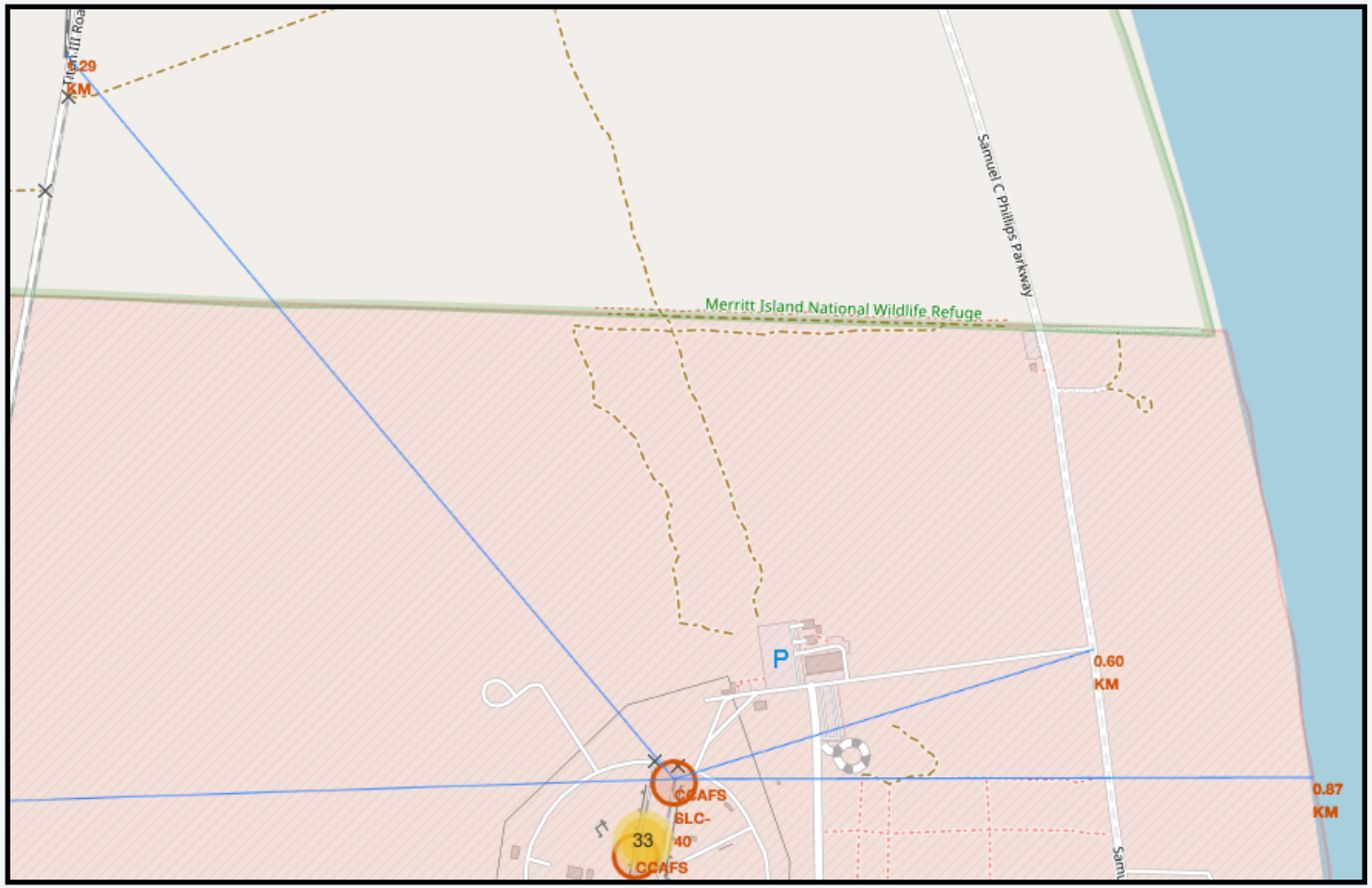
The East coast Launch sites are approximately 78.5KM away from the nearest city, Orlando.

Distance Between Launch Sites and its Proximities

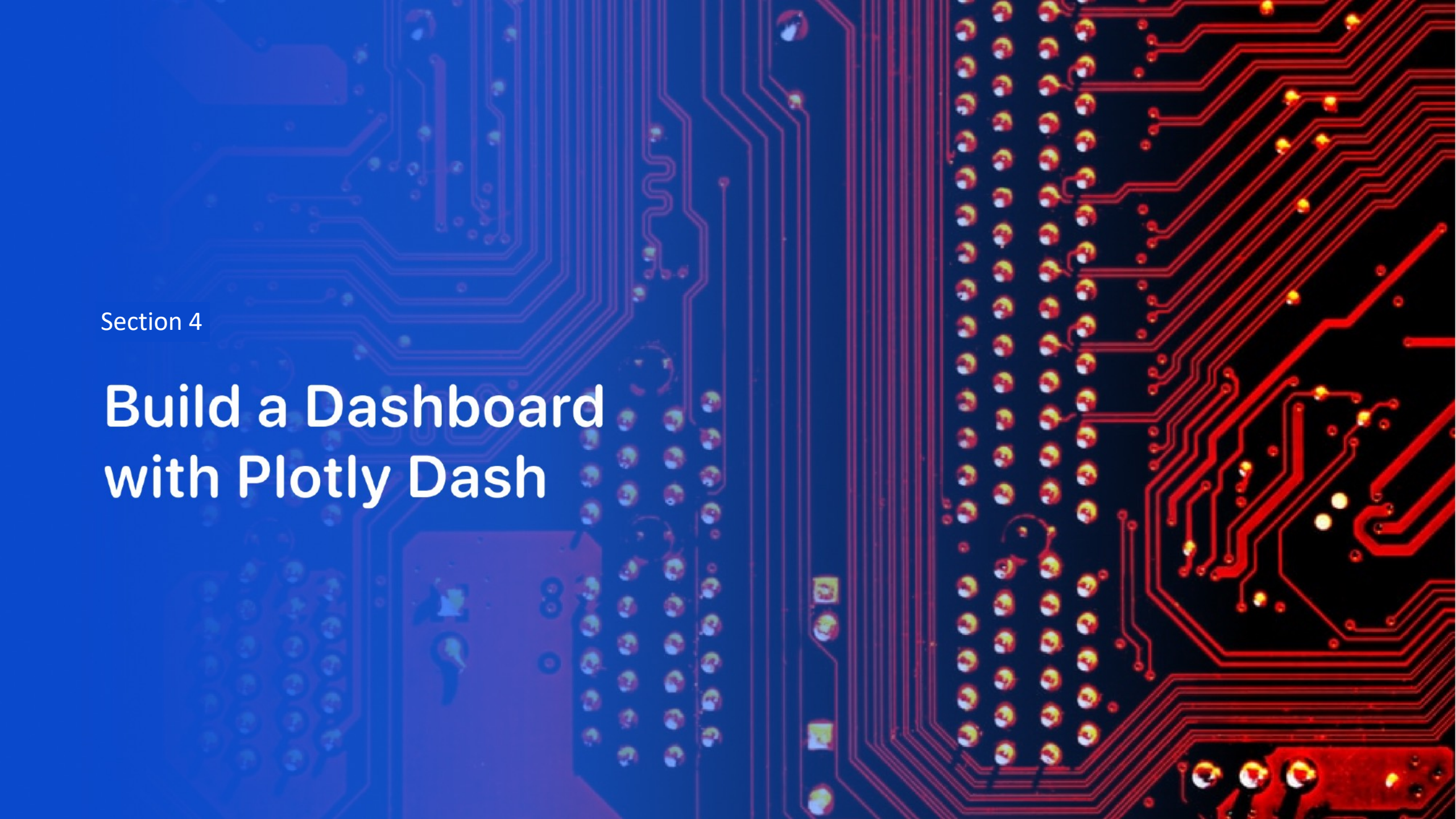


The East coast Launch sites are approximately 78.5KM away from the nearest city, Orlando.

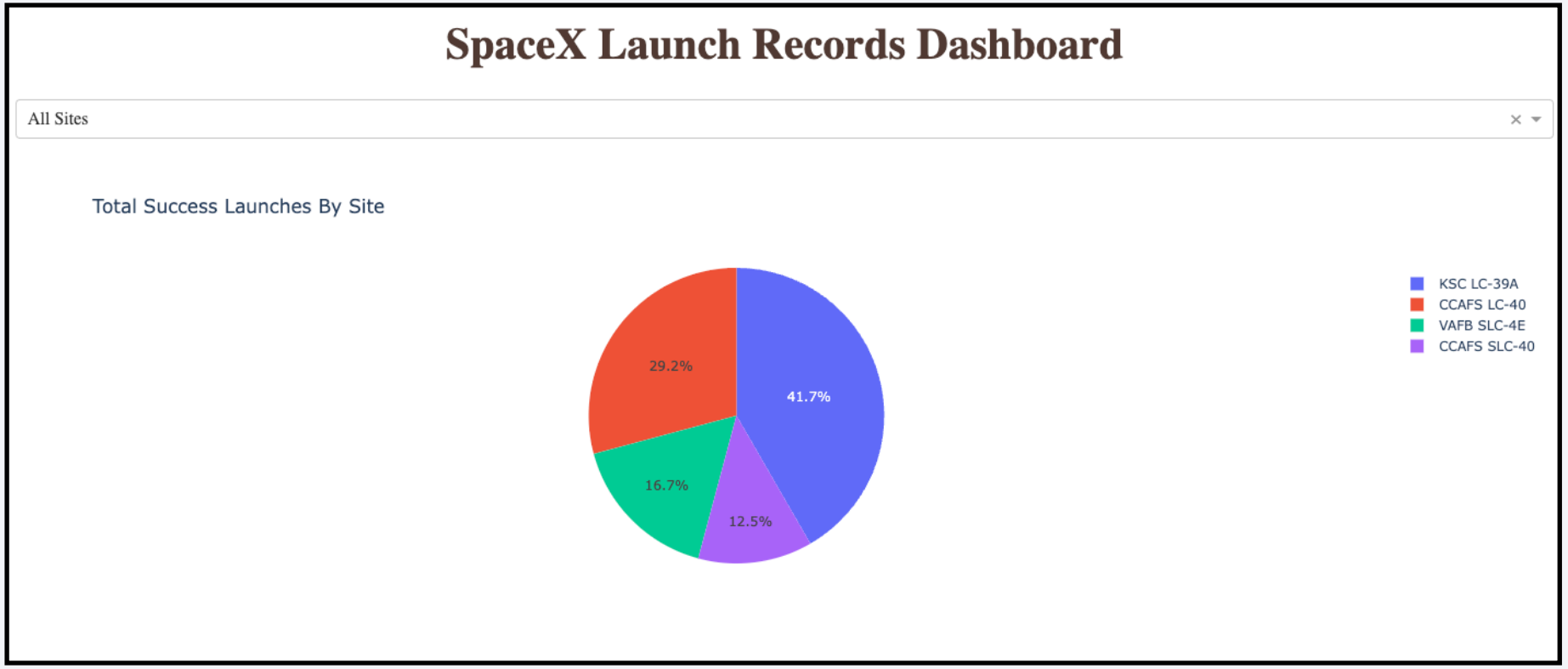
Distance Between Launch Sites and its Proximities



- The Launch sites are in close proximity from the nearest railway and highway, 0.6KM and 1.29Km respectively.
- The Launch sites are close to the coastline at 0.87KM away.

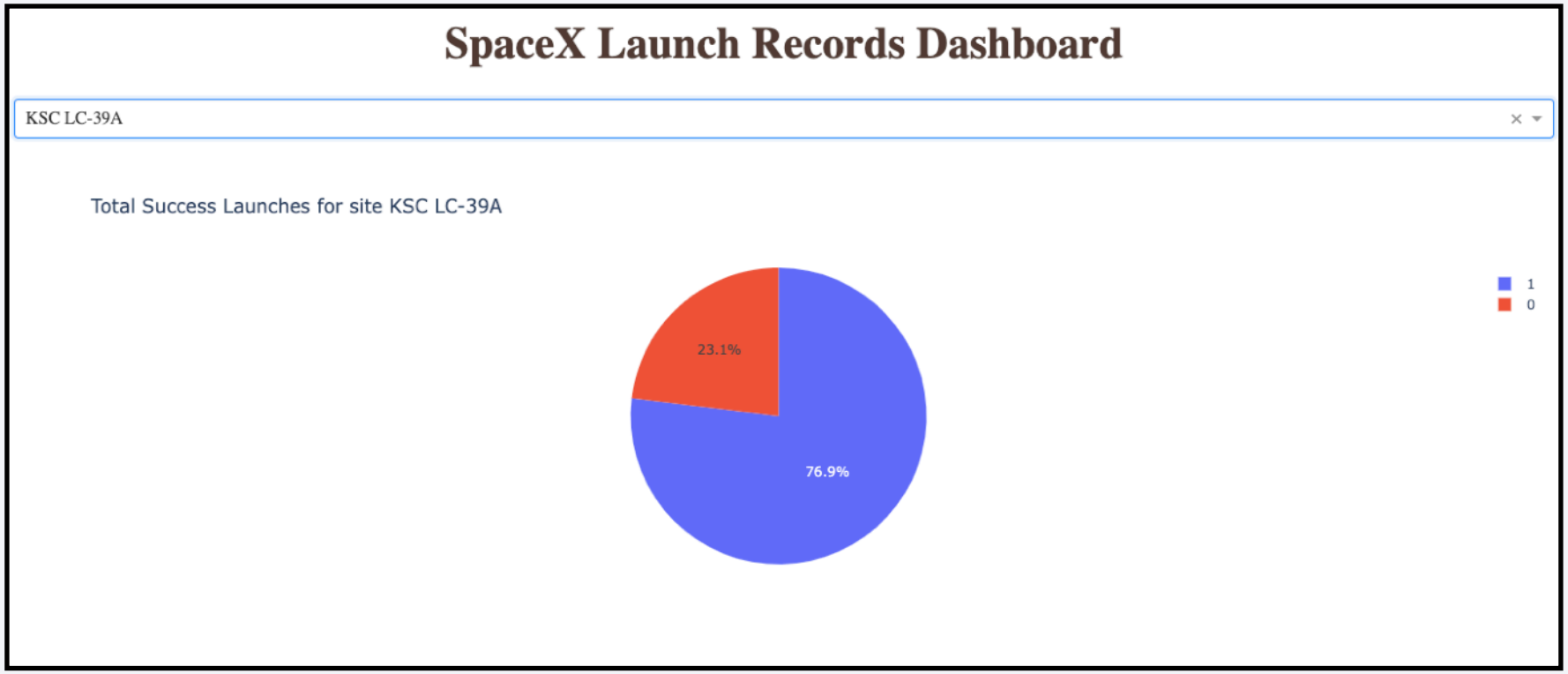

Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Success of All Sites



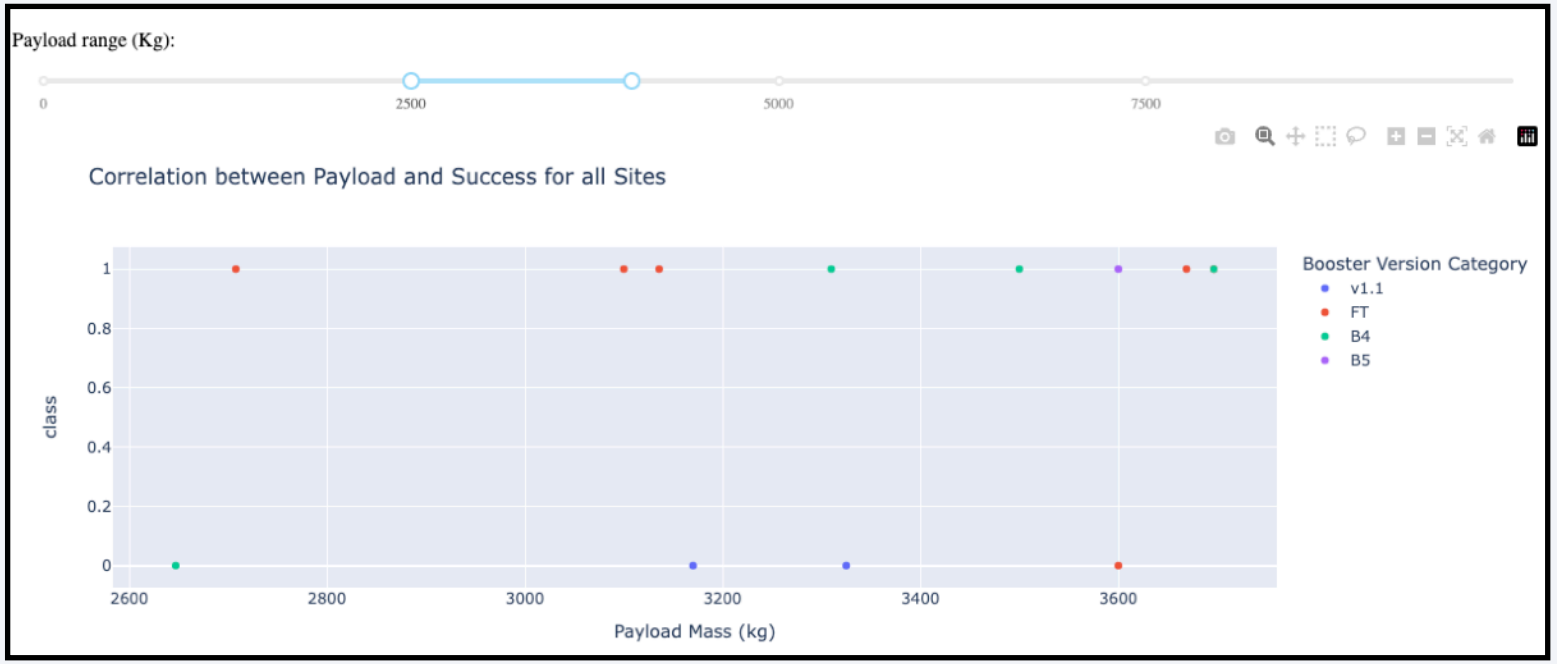
- This pie chart shows the launch success proportion of all sites.
- As shown in the chart, KSC LC-39A has the highest number of launch success compared to the rest.

Launch Site with the Highest Launch Success



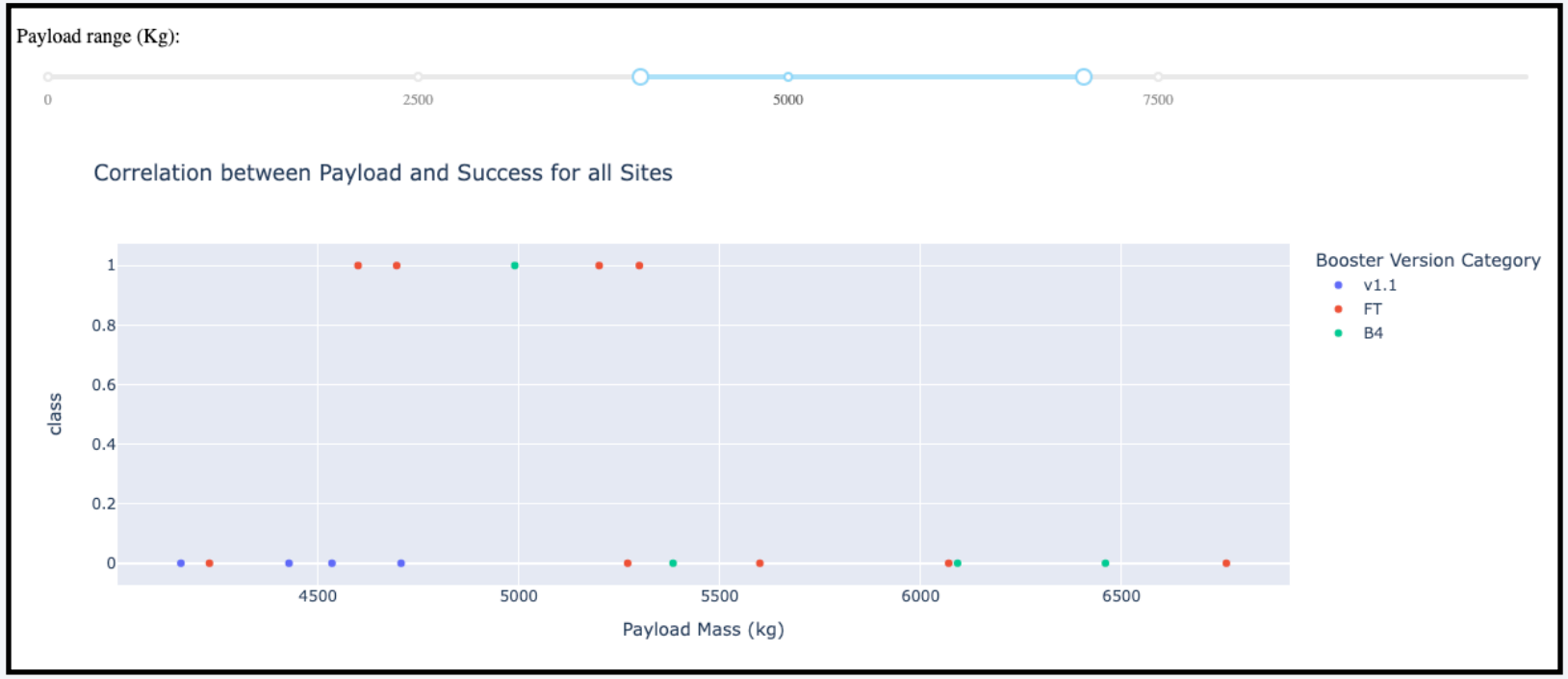
- The launch site with the highest launch success ratio KSC LC-39A is shown above.
- Ratio of 76.9% Success to 23.1% Failure.

Payload vs. Launch Outcome



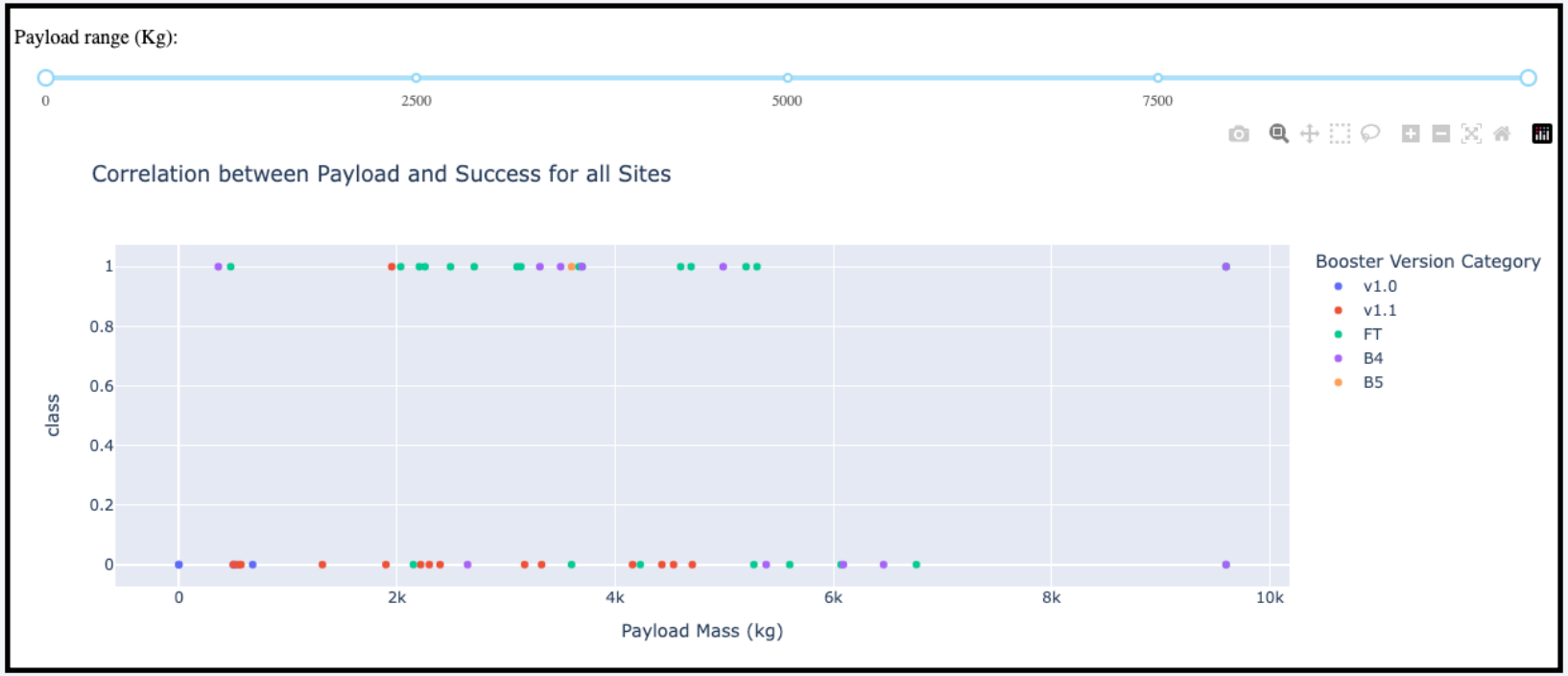
- Payload range between 2500kg and 4000kg has the highest launch success rate.

Payload vs. Launch Outcome



- Payload range between 4000kg and 7000kg has the lowest launch success rate.

Booster Version vs. Launch Outcome

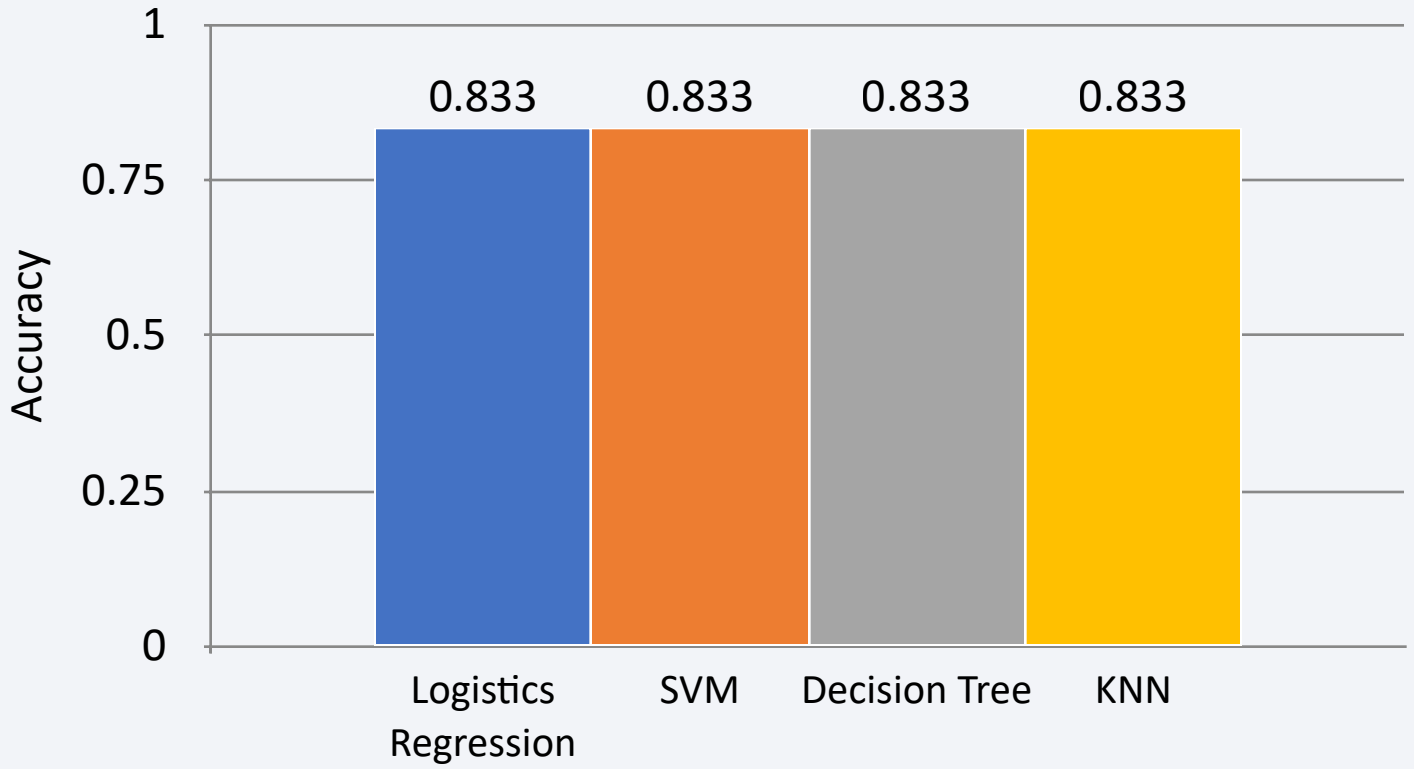


- FT version booster has the highest launch success rate.
- While V1.1 version booster has the lowest launch success rate.

Section 5

Predictive Analysis (Classification)

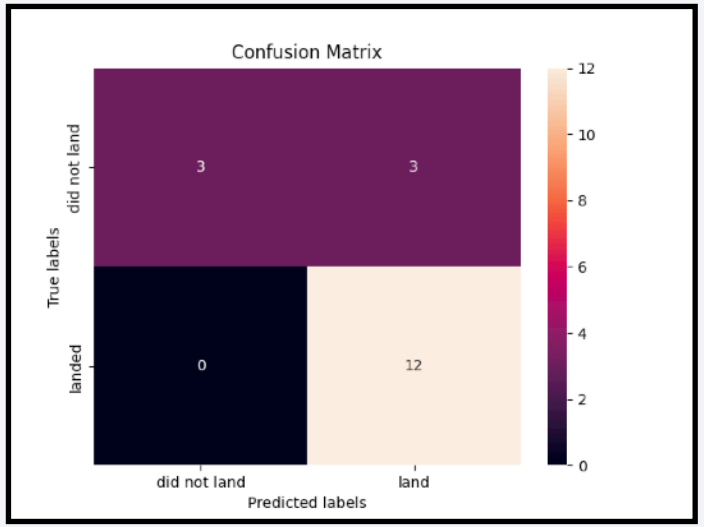
Classification Accuracy



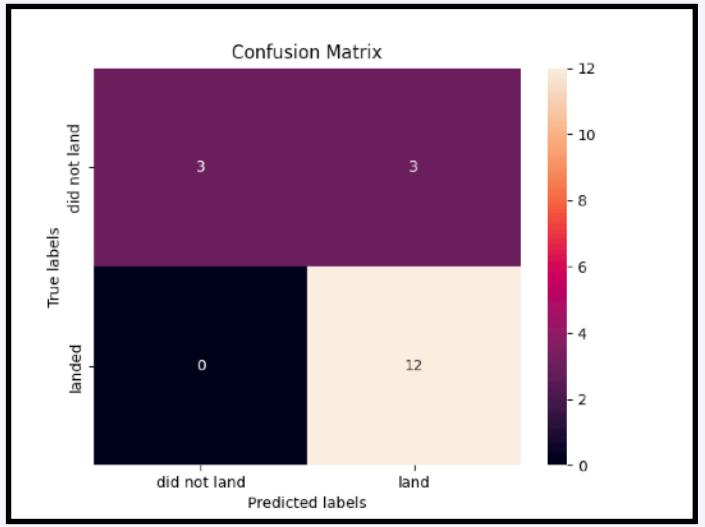
- Practically all these four algorithms have the same accuracy.

Confusion Matrix

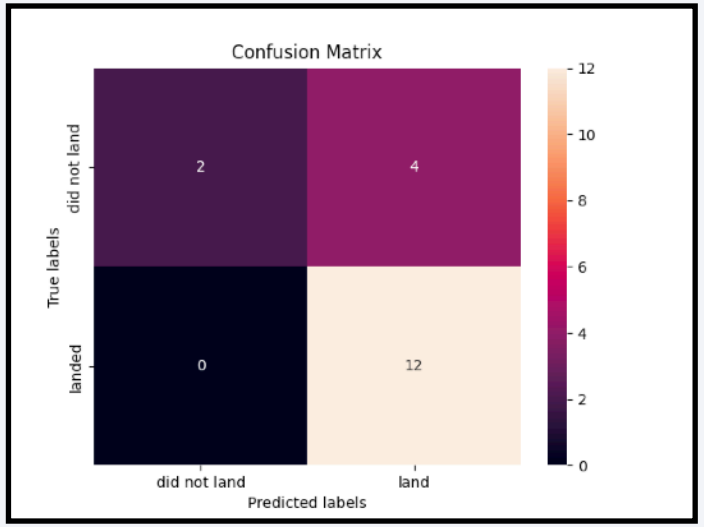
Logistics Regression



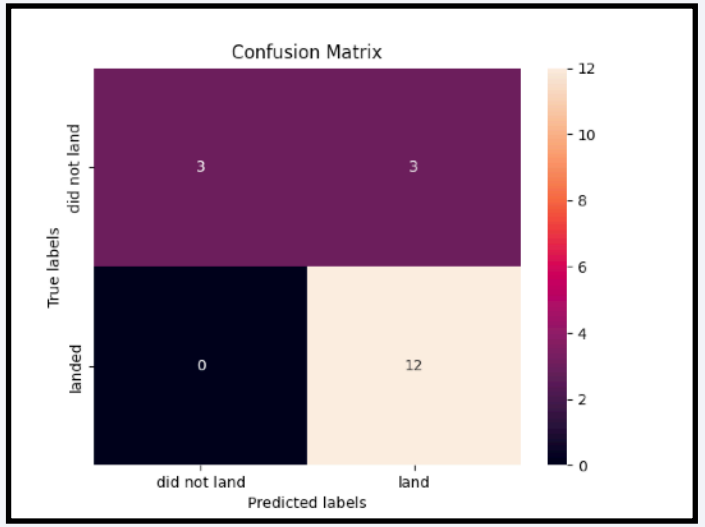
SVM



Decision Tree



KNN



- Practically all these four algorithms have the same accuracy.

Conclusions

- Data collection using SpaceX API and web scraping
- Data wrangling and dealing with missing values
- Explorative data analysis with visualisation and SQL queries
- Visual analysis with Folium and Plotly dashboard
  - Found geospatial relations, launch outcome ratios
- Classification predictive analysis with Logistic Regression, SVM, Decision tree, KNN
  - Found all four ML algorithms have the same accuracy
  - More evaluation required, evaluate based on desired use case

Appendix

GitHub repository: https://github.com/guosiong/Applied-Data-Science-Capstone.git

SpaceX launch data : https://github.com/guosiong/Applied-Data-Science-Capstone/blob/0baa19bec3a1c27ee71065c621f7716c6b3683ea/spacex_launch_dash_data.csv

Thank you!