IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kwame Owusu Baah
January 13, 2023

Outline

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

Executive Summary

The data was collected from SpaceX Rest API and Webscraping from Falcon 9 and Falcon Heavy Launches Records from Wikipedia. The data was then loaded into Pandas data frame. The data frame was used to create a label column called “class”, which classified the successful landings.

The data was explored using combination of SQL queries, Visualizations, Folium Maps and Dashboards. All categorical variables were changed to binary using one hot coding. The data was then standardized, and GridSearchCV was used to find the best parameters for the machine learning models. The accuracy score for all the parameter models were visualized to find the best performing one.

The machine learning models used were: Logistic Regression, Decision Tree Classifier, Support Vector Machine (SVL) and K-Nearest Neighbors (KNN). After applying all the models, Decision Tree performed best with an accuracy of about 89%. All the models predicted successful landings.

Introduction

Background:

The Commercial Space age is here, and many companies are making the space travel affordable for everyone. SpaceX is the most successful space travel company among them.

The reason SpaceX has been able to achieve this is because their rocket launches are inexpensive. SpaceX advertises Falcon 9 rocket launches on its website at a cost of $62 million, while the cost for other providers is about $165 million each. Much of the savings SpaceX makes is due to their ability to reuse the first stage.

SpaceY wants to compete with SpaceX and therefore we want to know the price of each rocket launch.

Problem:

Our aim is to understand the variables that are responsible for having a successful landing of launched rockets. We intend to do this by gathering information about SpaceX and creating dashboards to better understand the data.

A machine learning model will be trained, and public information will be used to predict if the first stage will land successfully, and this will indicate if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX API https://api.spacexdata.com/v4/launches/past and Webscraping from List of Falcon 9 and Falcon Heavy launches - Wikipedia
- Perform data wrangling
  - Converting the launch outcomes into training labels with the Booster Landing being "successfully landed" or "unsuccessfully landed."
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use GridSearchCV to tune the machine learning models to find the best parameter.

Data Collection

The data collection process involved the combination of API request from SpaceX API https://api.spacexdata.com/v4/launches/past and Webscraping data from SpaceX Falcon 9 and Falcon Heavy Launch records that was obtained from Wikipedia. https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

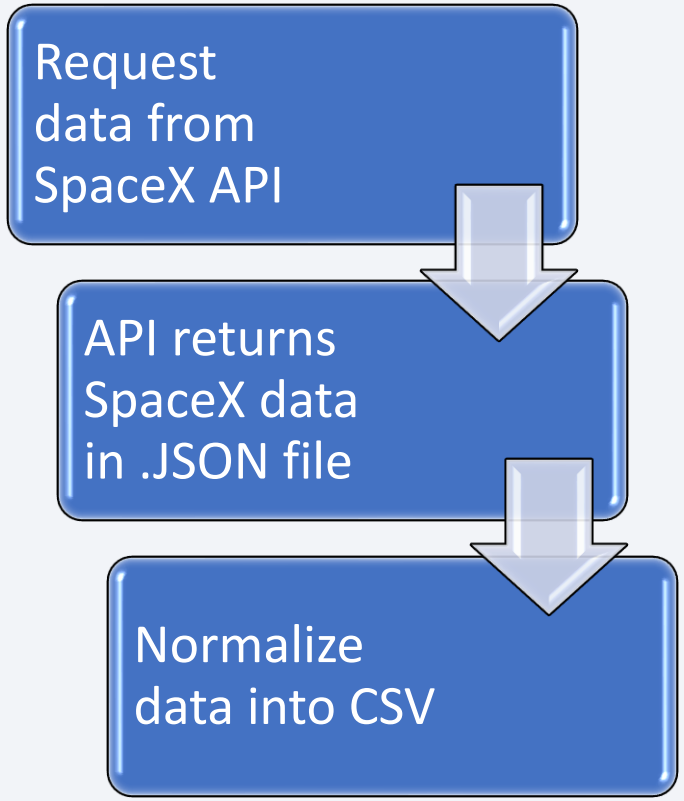
SpaceX API Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Web Scrape Data Columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

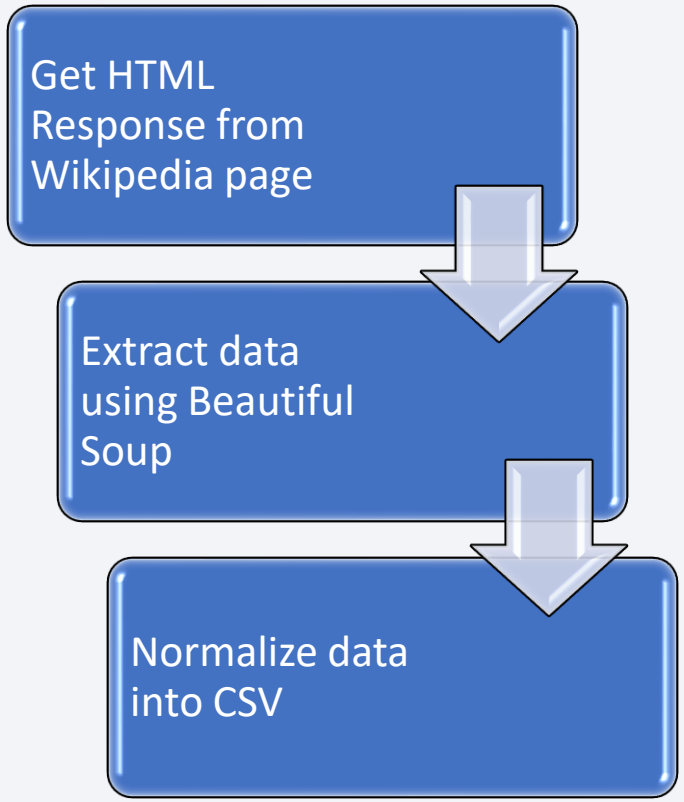
- The data was collected from a SpaceX public API where data is obtained and used.
- The data collected was taken through the process depicted in the flowchart.

GitHub: SpaceX API



Data Collection - Scraping

- SpaceX launches data was obtained from Falcon 9 and Falcon Heavy launches page on Wikipedia.
- The data was downloaded from Wikipedia using the process in the flowchart.



GitHub: SpaceX Web Scraping

Data Wrangling

Landing outcomes were converted into training labels with “1” meaning the booster successfully landed, and “0” meaning the booster unsuccessfully landed.

A new training label column called ”class” was created. A value of 1 was assigned if the first stage landed successfully, and a value of 0 was assigned if the first stage did not land successfully.

True Ocean, True RTLS, True ASDS set to 1

False Ocean, False RTLS, False ASDS, None None, None ASDS set to 0

GitHub: First Stage Landing Prediction

EDA with Data Visualization

Exploratory Data Analysis was performed to visualize the relationships between pairs of variables and trends over a period.

- Scatter Plots were used to show if there is a relationship between 2 variables and to see if one variable will influence the another.

Example: FlightNumber vs PayloadMass, FlightNumber vs LaunchSite, LaunchSite vs PayloadMass, FlightNumber vs Orbit Type and PayloadMass vs Orbit Type

- Bar Charts made it easy to compare datasets between multiple groups immediately and helped to visualize any relationships between variables.

Example: Success Rate vs Orbit Type

- Line Chart was used to show the success rate over a period of time.

Example: Year vs Success Rate

GitHub: EDA with Data Visualization

EDA with SQL

A dataset was loaded into a Table created that was named SPACEXTBL in IBM DB2 Database. The following SQL queries were performed:

- 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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

GitHub: EDA with SQL

Build an Interactive Map with Folium

Map Objects such as Markers, Circles, Lines and Mark Clusters were used with Folium Map.

- Markers were used to show points on the map, such as all launch sites.
- Circles were used to highlight areas around specific coordinates on the map, like NASA Johnson Space Center at Houston.
- Mark Clusters were used to simplify many markers on the map with the same coordinates, like many launches happening at the same coordinate or launch site.
- Lines were used to show the distances between coordinates, such as launch site and its proximities

GitHub: Interactive Map with Folium

IBM Cloud: Interactive Map with Folium

Build a Dashboard with Plotly Dash

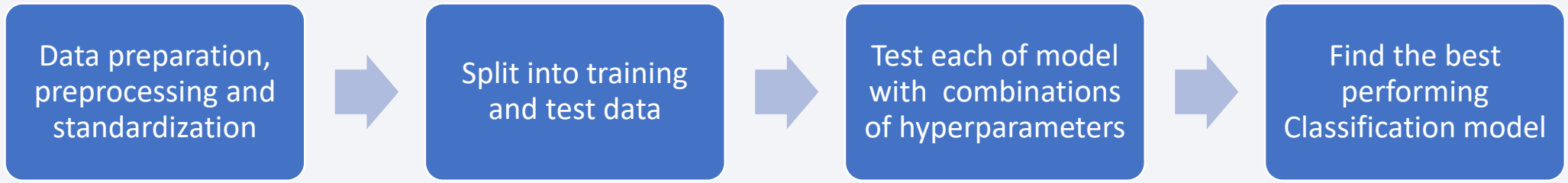
The Dashboard contains a Pie Chart and a Scatter Plot which helped to identify the best place to launch according to the payload mass.

- The Pie Chart was used to show the distribution of successful landings across all launch sites. There is a drop-down option which shows each individual launch site and their success rates.
- The Scatter Plot shows these inputs: all launch sites and individual launch sites, Booster Version Category and PayloadMass (kg) on a slider between 0 and 10,000 kg.
- The Scatter Plot shows how the success rate varies across different Launch Sites, Payload Mass and booster version categories.

GitHub: Interactivity Dashboard (code only)

Predictive Analysis (Classification)

- The data was prepared, preprocessed and standardized.
- The data was then split into training and testing data using the function train_test_split. The training data was divided into validation data, and a second set was used for training data on the algorithm.
- The model was trained and performed using Grid Search, enabling us to find the hyperparameters that allow a given algorithm to perform best. Using the best hyperparameter values, we determined the model with the best accuracy using the training data.
- The data was tested using Logistic Regression, Support Vector Machines (SVM), Decision Tree Classifier and K-Nearest Neighbors (KNN).
- Confusion Matrix was used to demonstrate prediction accuracy.



GitHub: [Machine Learning Predictive Analysis]

Results

Exploratory data analysis results

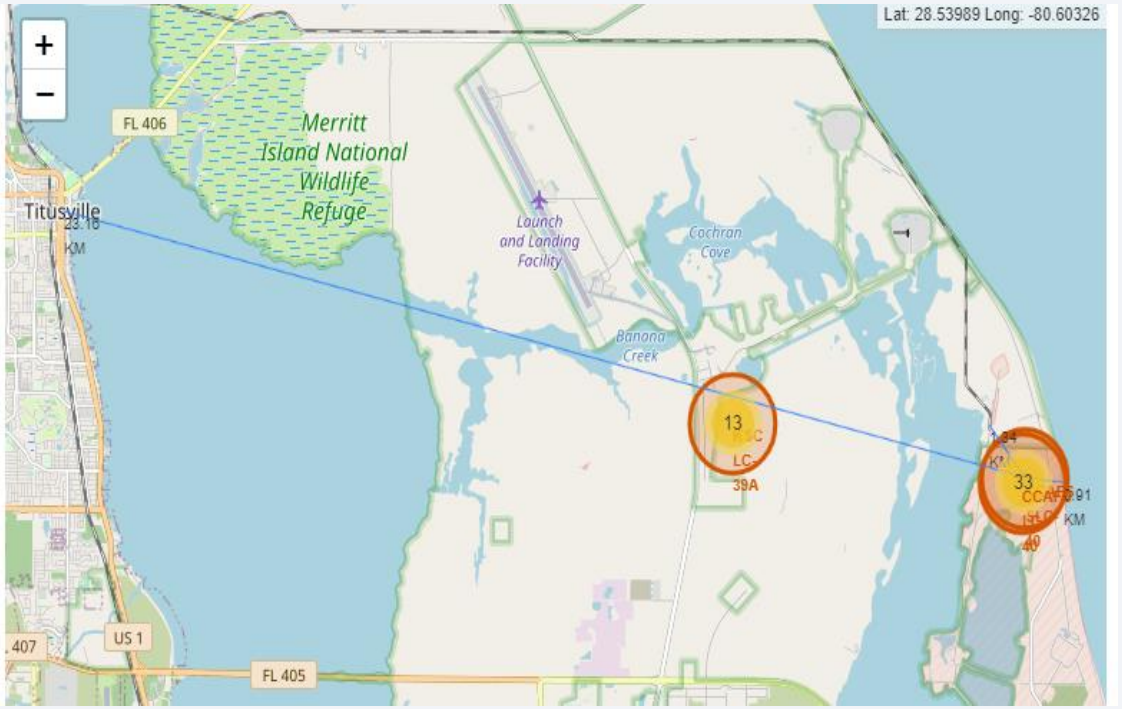
- SpaceX uses 4 different launch sites
- The first launches were done by Space X on 06/04/2010
- The average payload mass of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2015, five years after the first launch in 2010
- Many Falcon 9 booster versions were successful at landing in drone ships having payload mass 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 better as the years passed.

Results

Interactive analytics demo in screenshots

With Interactive Analytics, we were able to find that most of SpaceX launches take place in East Coast launch sites in Florida.

Also, the proximity of launch sites to the sea makes the launches safe. There are adequate infrastructures closer to these launch sites, such as access to railways and highways which make it easy to meet any logistical needs, including the transportation of personnel.

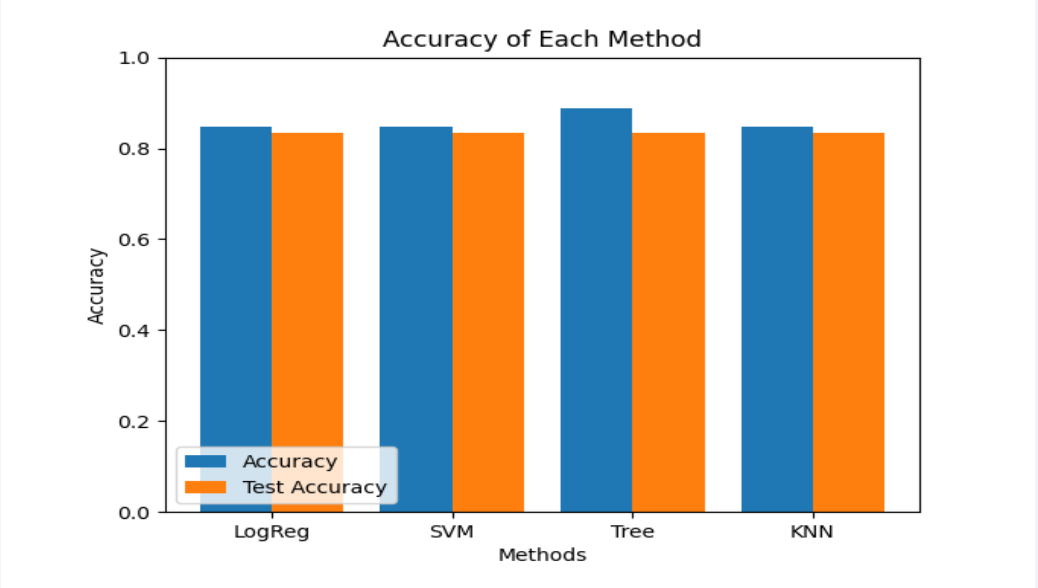


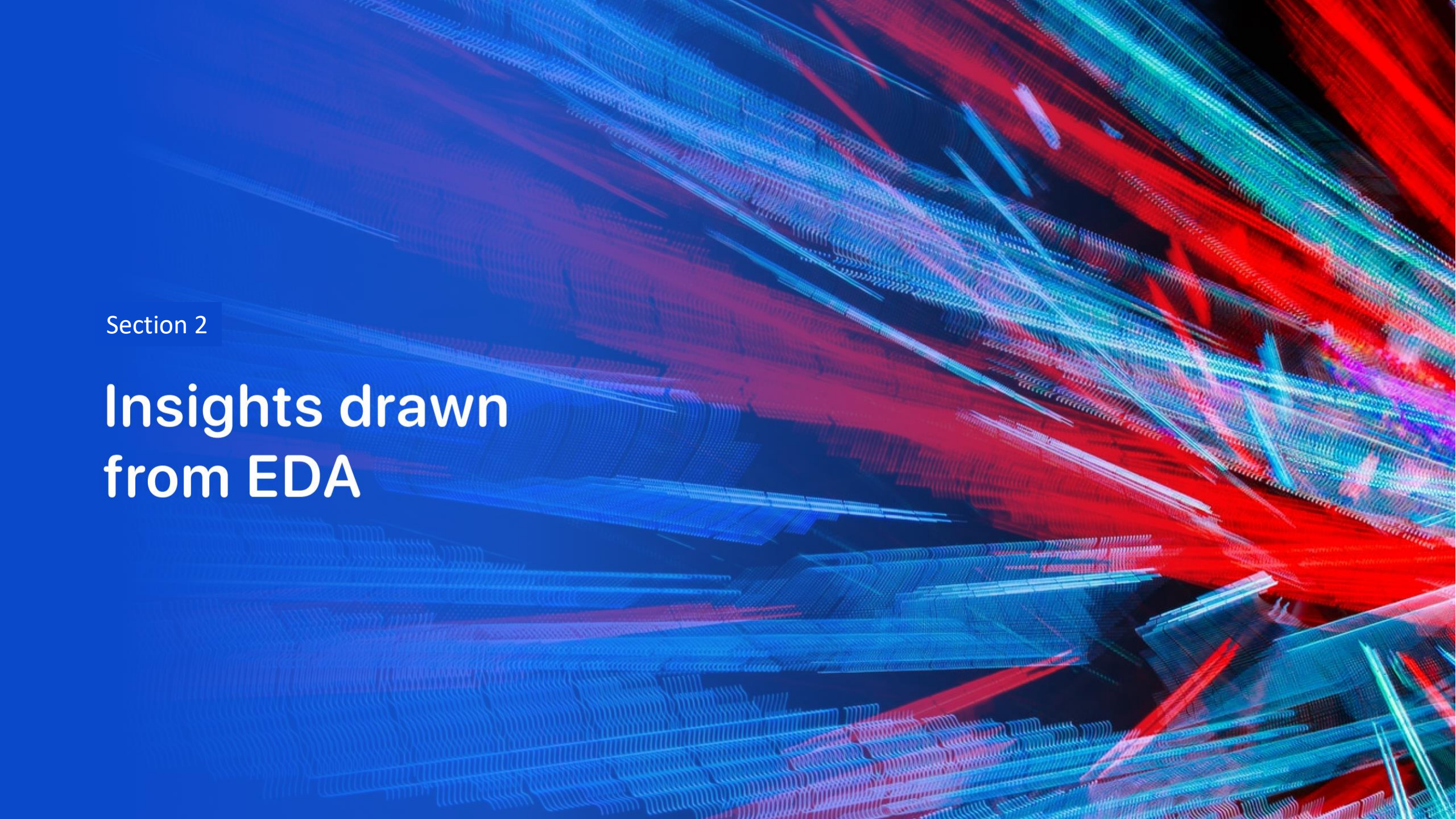
Results

Predictive Analysis results

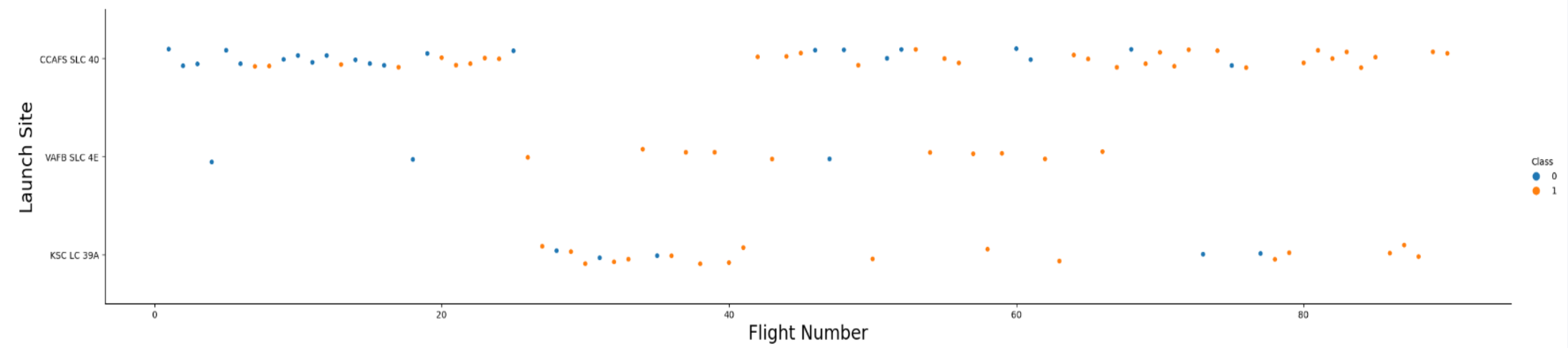
The predictive analysis showed that the Decision Tree Classifier is the best model to predict successful landings with an accuracy score of 88.9%, as compared to the other models.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.88929	0.83333
KNN	0.84821	0.83333




Section 2
Insights drawn from EDA

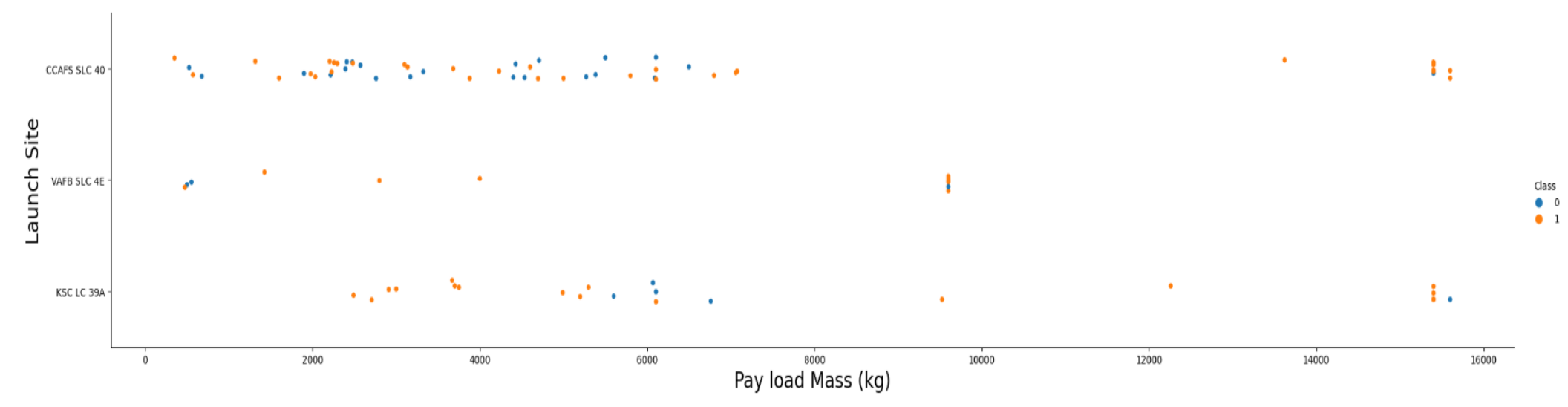
Flight Number vs. Launch Site



Class 0 (blue) indicates unsuccessful launch and Class 1 (orange) indicates successful launch.

- The scatter plot above shows that most successful launches happened at CCAFS SLC 40, where most of the launches took place.
- Also, it can be noted from the plot that the general success rate of the launches improved over time as the number of flights increased.
- Comparing the success rates of launch sites, the plot showed that site CCAFS SLC 40 had the most success rate followed by site VAFB SLC 4E. The least success rate was recorded at KSC LC 39A.

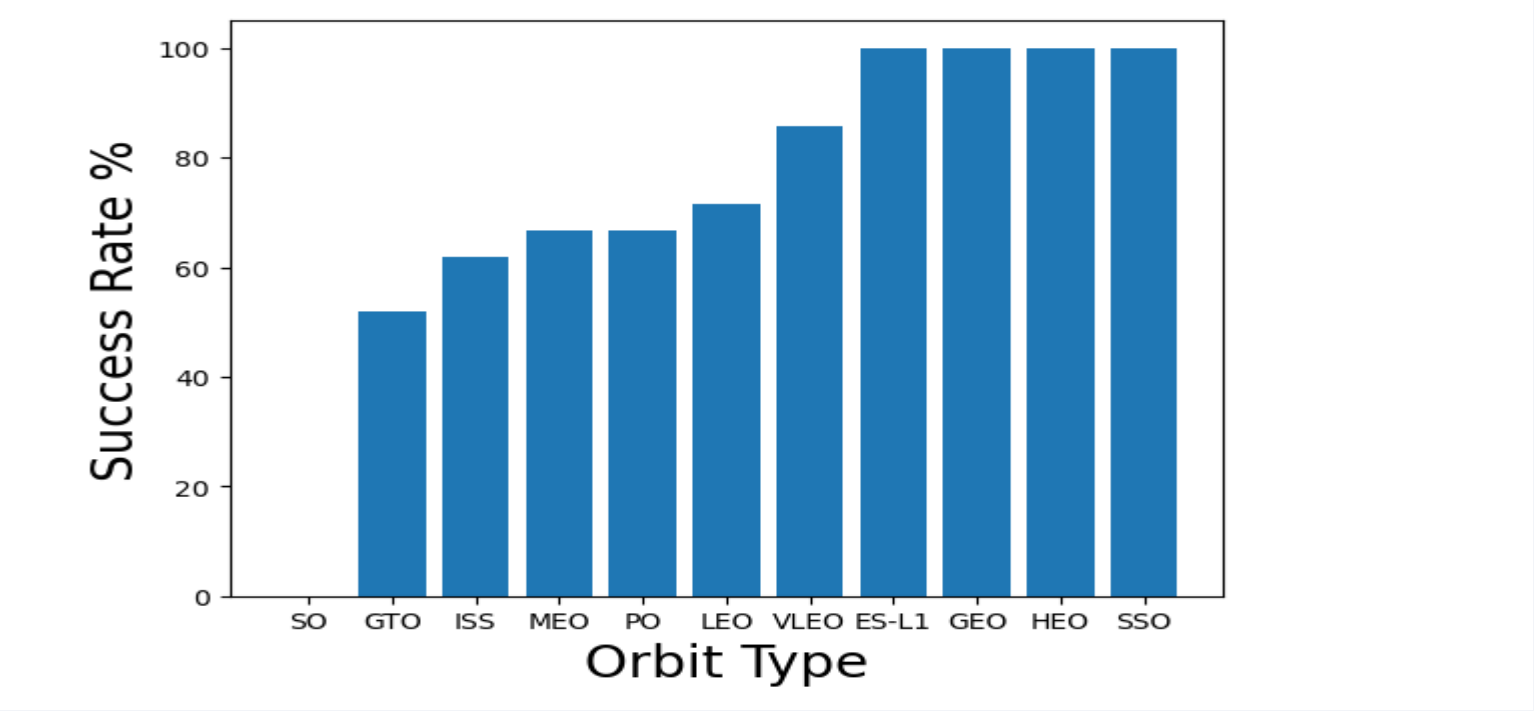
Payload vs. Launch Site



Class 0 (blue) indicates unsuccessful launch and Class 1 (orange) indicates successful launch.

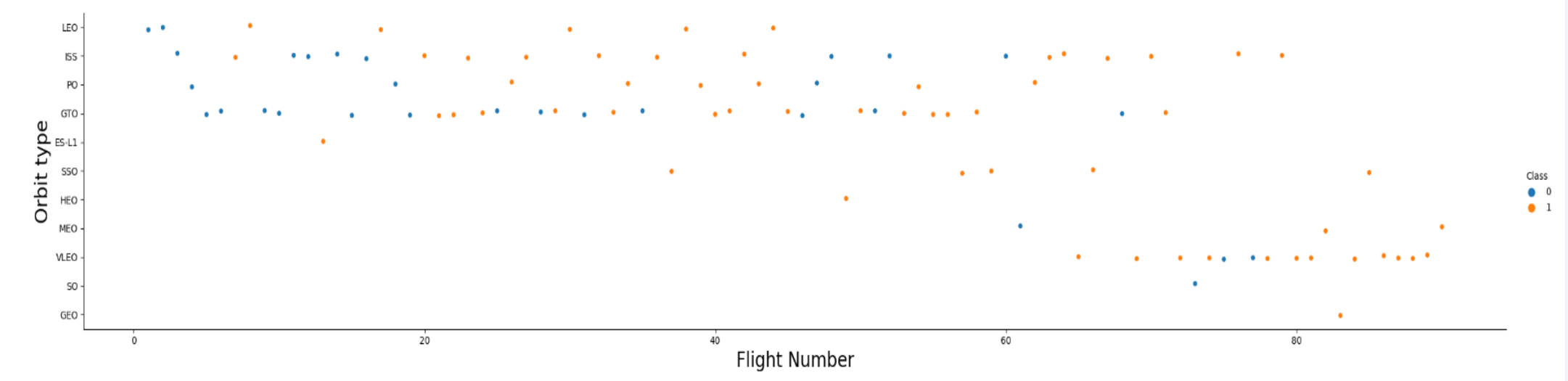
- Most of the launches appear to fall below the payload mass of 8,000kg.
- Launches with payload mass of over 9,000kg had a good success rate.
- From the plot, launches with payload mass of over 12,000kg usually happened at launch sites CCAFS SLC 40 and KSC LC 39A and were mostly successful.

Success Rate vs. Orbit Type



- Orbit **ES-L1, GEO, HEO** and **SSO** had the highest success rates of 100%.
- Orbit SO had the lowest success rate of 0%, which was a failure in a single attempt.
- The rest of the Orbit types had success rates from 50% to 80%.

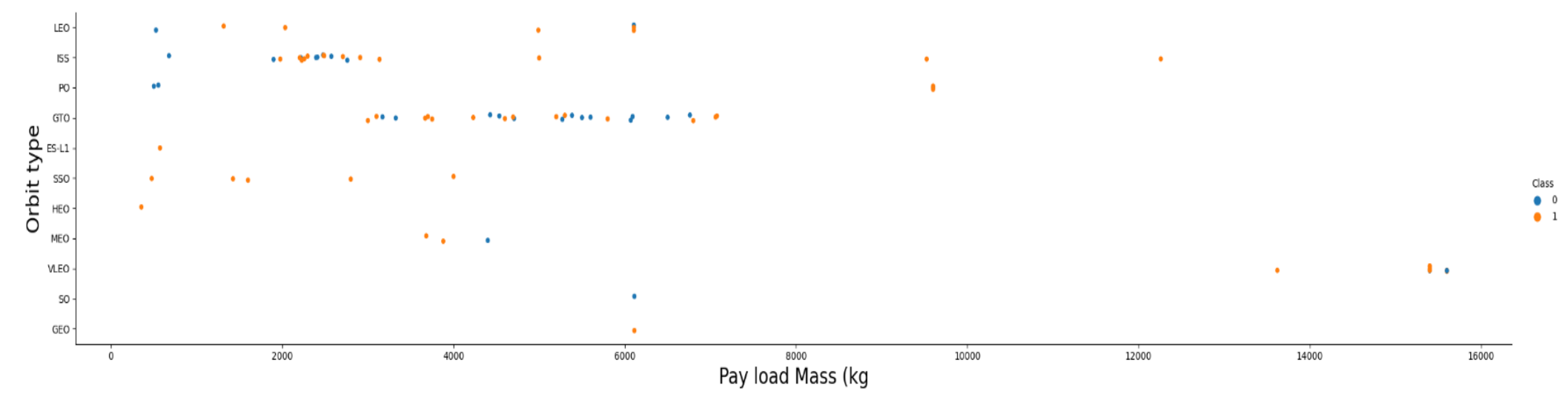
Flight Number vs. Orbit Type



Class 0 (blue) indicates unsuccessful launch and Class 1 (orange) indicates successful launch.

- The plot shows that the success rate or launch outcome improved at most orbit types as the number of flights at these orbit types increased.
- Orbit VLEO is shown to be used recently as its success rate keeps increasing with increased flight numbers.

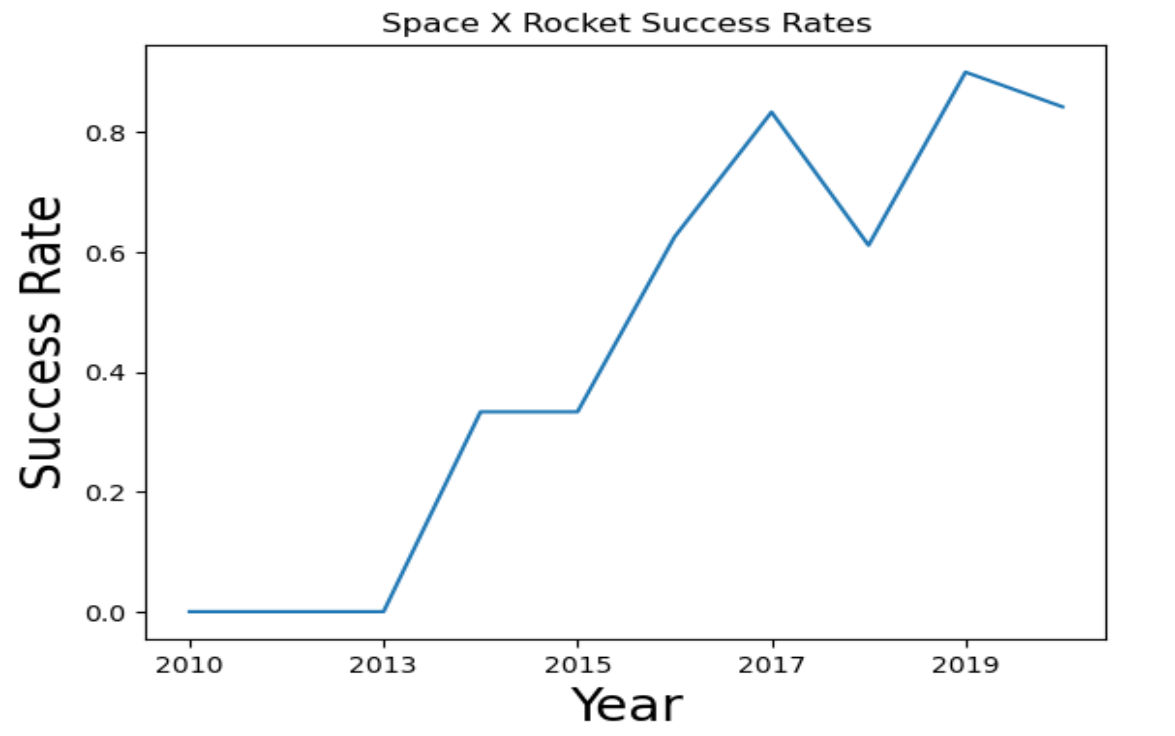
Payload vs. Orbit Type



Class 0 (blue) indicates unsuccessful launch and Class 1 (orange) indicates successful launch.

- From the plot, ISS orbit had good success rates as the payload mass increased.
- All launches from orbit SSO were successful as the payload mass increased from 0 – 4,000kg.
- As shown on the plot, orbit GTO has a mixture of successful and unsuccessful landing outcomes as the payload mass increases. This makes it difficult to see any relationship between payload mass and success rate.

Launch Success Yearly Trend



Success Rate scale with 0 as 0%, 0.4 as 40% and 0.8 as 80%

- The success rate continued to increase from 2013 until 2017.
- The success rate decreased in 2018 and increased again in 2019.
- A 0% success rate was recorded for the first 3 years.

All Launch Site Names

- From the data, there are four unique launch sites

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

- When Distinct launch sites were queried from SpaceX Table in the database, these four unique launch sites were obtained.
- CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- These are the 5 records where launch sites begin with `CCA`

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

- Above are the first five entries of records of launch sites in the database with names beginning with ‘CCA’

Total Payload Mass

- This is the total payload mass carried by boosters from NASA

total_payload_mass_kg
45596

- The total payload mass is the sum of all the payload masses where the customer is NASA (CRS)

Average Payload Mass by F9 v1.1

- This is the average payload mass carried by booster version F9 v1.1

average_payload_mass_kg
2928

- This average payload mass is the result obtained after filtering the data by booster version ‘F9 v1.1’ and finding the average payload mass.

First Successful Ground Landing Date

- This is the date of the first successful landing outcome on ground pad

first_successful_landing_date
2015-12-22

- This date is a result obtained after filtering the data by ‘Success (ground pad)’ and finding the minimum value for Date which showed this date (2015-12-22) as the first successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload Mass between 4000kg and 6000kg

- These are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000kg but less than 6000kg

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

- These are the four distinct booster versions after filtering the data by ‘Success (drone ship)’ and a payload mass between 4,000kg and 6,000kg

Total Number of Successful and Failure Mission Outcomes

- These are the total number of successful and failure mission outcomes

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- These are the results of counting the records of each mission outcome and grouping them by ‘mission outcome.’ The results show that SpaceX achieved 99% successful mission outcomes.

Boosters Carried Maximum Payload

- These are the names of the booster versions which have carried the maximum payload mass
- The booster versions carried the maximum payload mass of 15,600 kg

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- These are the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

landing_outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- These are the only two occurrence of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

- This is the ranking of all the landing outcomes between the date 2010-06-04 and 2017-03-20.
- From the results, there were 5 Failure (drone ship) landing outcomes and 5 Success (drone ship) landing outcomes in the given period.
- There were 3 successful landing outcomes on ground pad in the given period.

landing_outcome	total_number_of_landing_outcomes
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

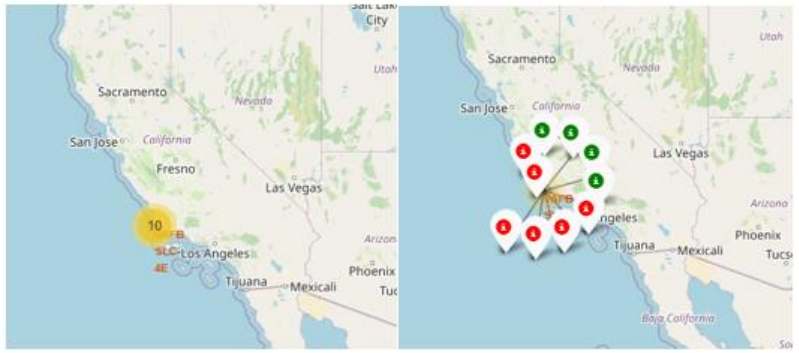
Section 3

Launch Sites Proximities Analysis

Locations of all Launch Sites

- The map shows all the launch sites locations in the United States: California in the West and Florida in the East.
- Also, it can be seen from the map that all the launch sites are located near the coast, most likely for safety reasons.

Color-labeled Launch Outcomes



Launch Site in California

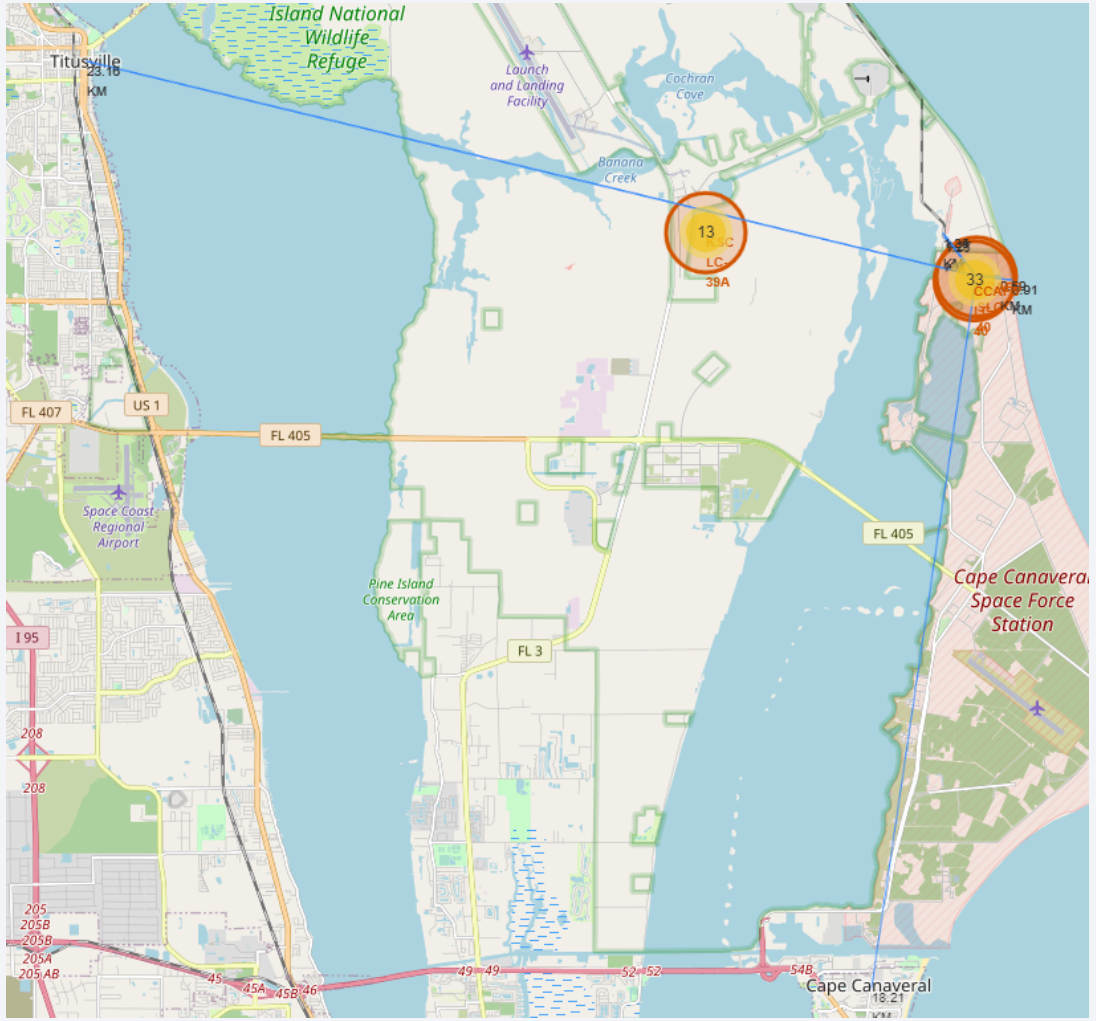
Launch Site in Florida

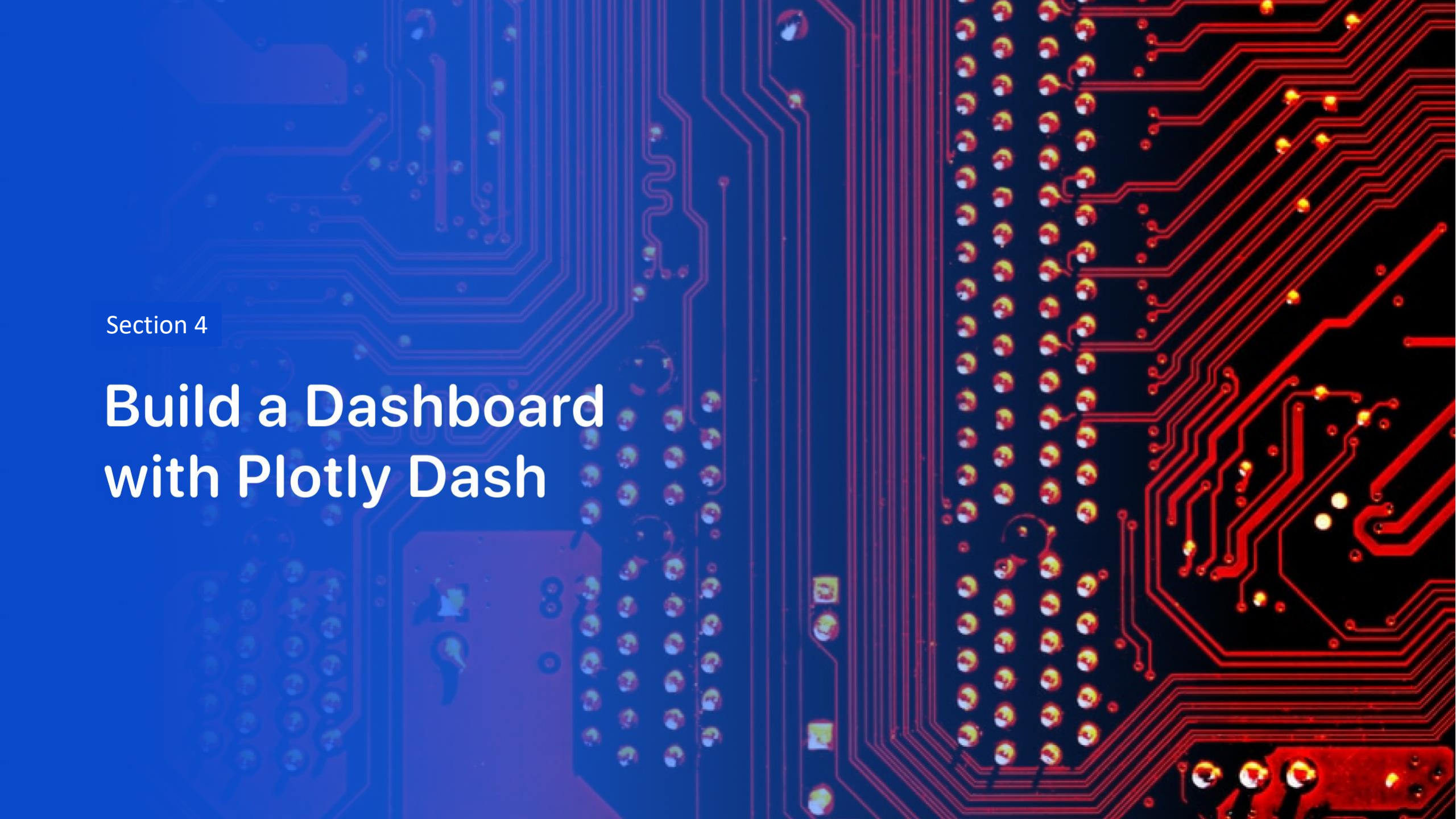
- The marker clusters on the map when clicked, displays the successful landings (green) and unsuccessful landings (red).
- The maps shown depict both successful and unsuccessful landing outcomes at launch sites in California and Florida.

Launch Sites and Their Proximities

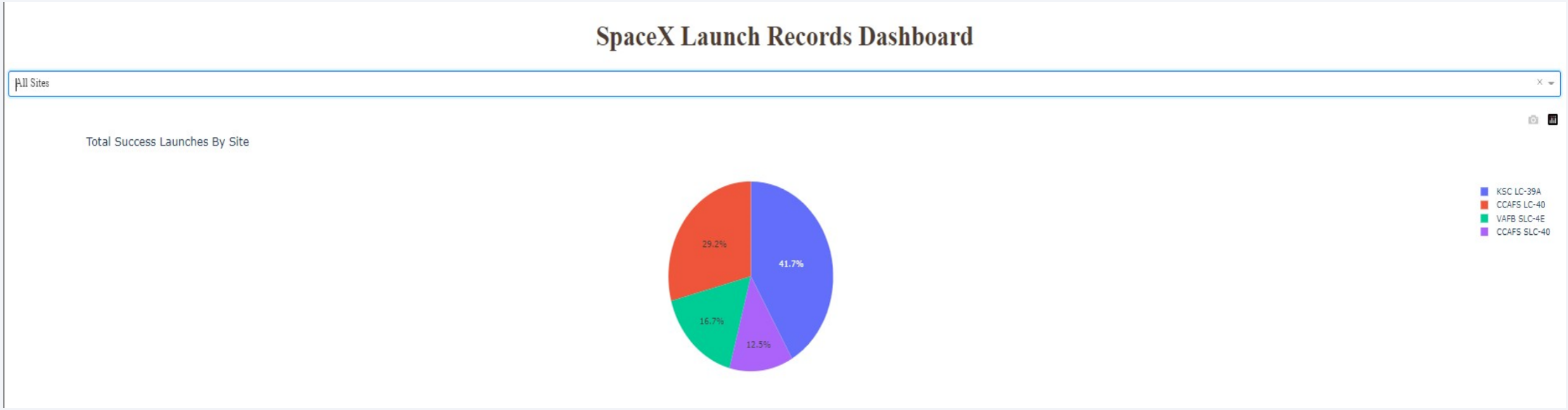
- Are launch sites in close proximity to railways? **Yes**
- Are launch sites in close proximity to highways? **Yes**
- Are launch sites in close proximity to coastline? **Yes**
- Do launch sites keep certain distance away from cities? **Yes**

- As shown on the map, the launch site is in close proximity to railways, highways and coastline, which makes it easy to meet any logistical needs, including the transportation of personnel and equipment.
- The launch sites as shown on the map keep certain distance away from cities to provide safety in case any malfunctions or unsafe landings pose a risk to nearby communities or cities.



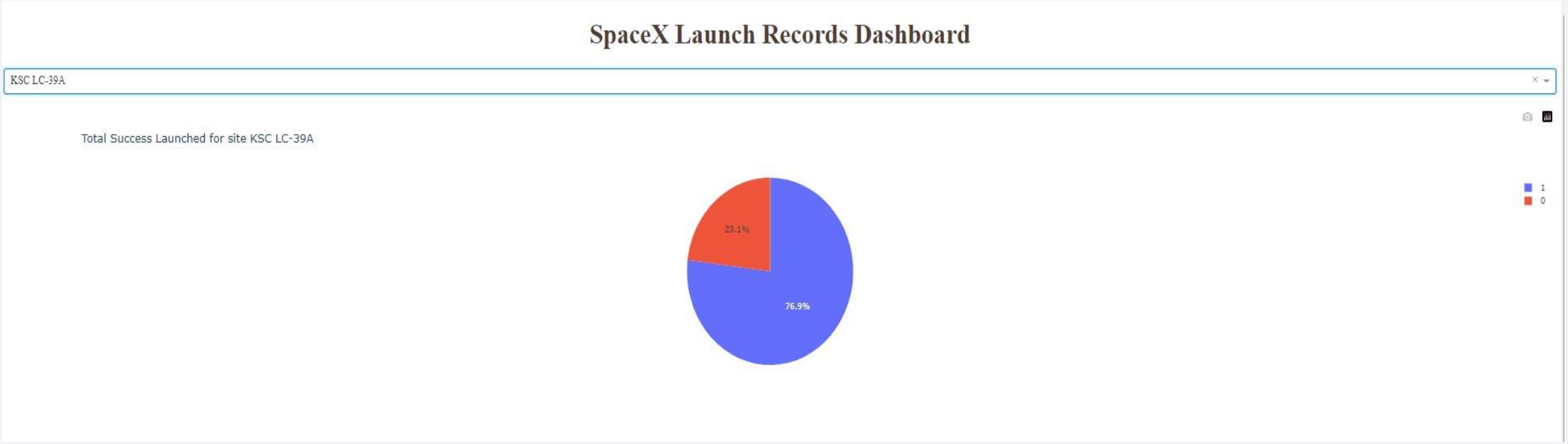

Section 4
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



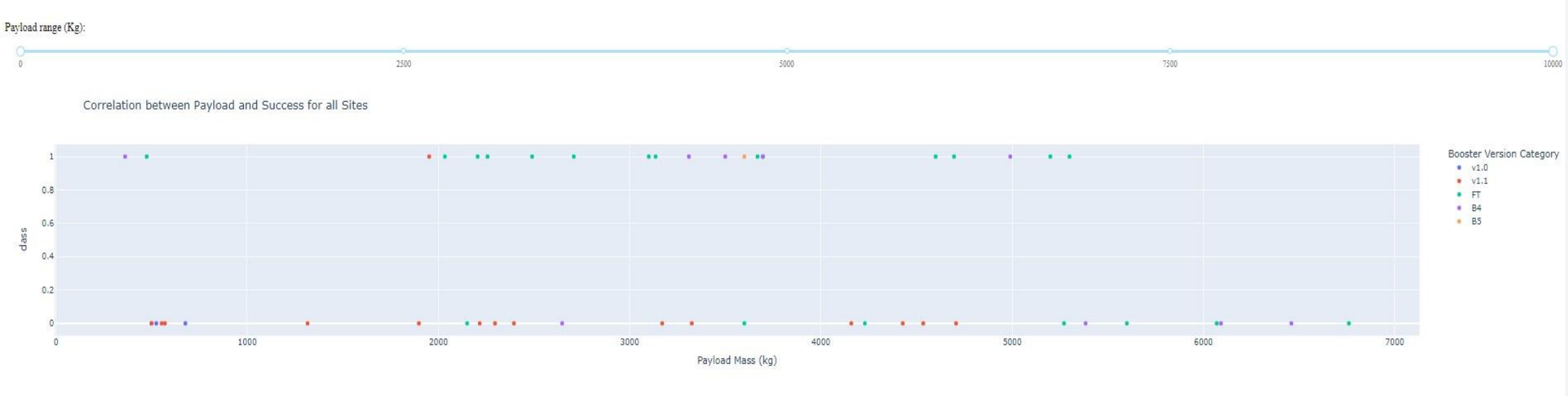
- The Pie Chart shows the distribution of successful landings across all launch sites.
- The KSLC- 39A site recorded the highest successful launches among all the launch sites.

Launch Sites with Highest Launch Success Ratio



- KSLC- 39A had the highest launch success rate with 10 successful landings (76.9%) and 3 unsuccessful landings (23.1%)

Payload Mass vs Launch Outcome



- As shown on the dashboard, Booster Version FT with Payload mass under 6000kg had the most launch success rates.

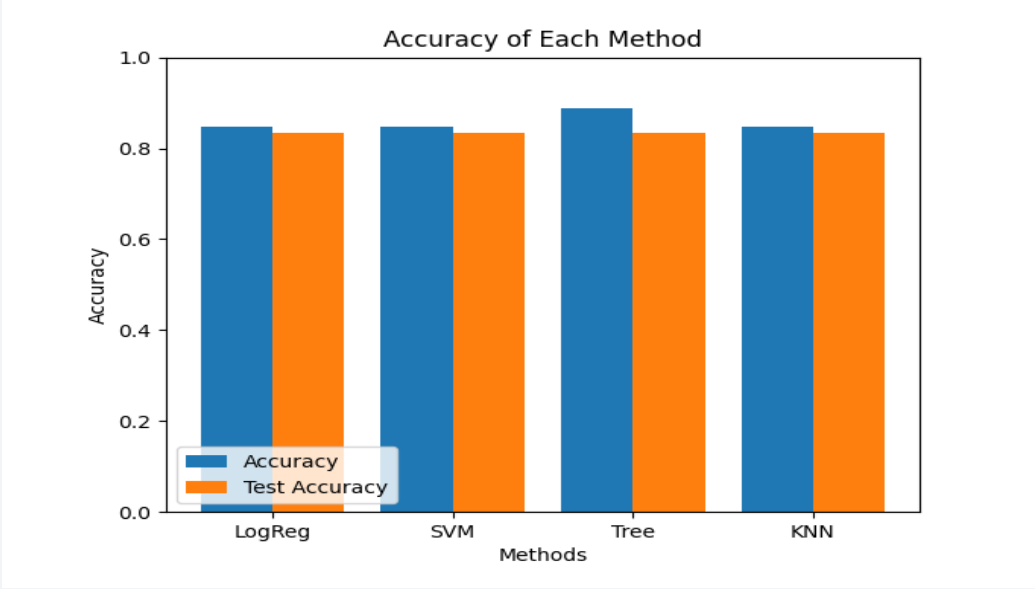
Section 5

Predictive Analysis (Classification)

Classification Accuracy

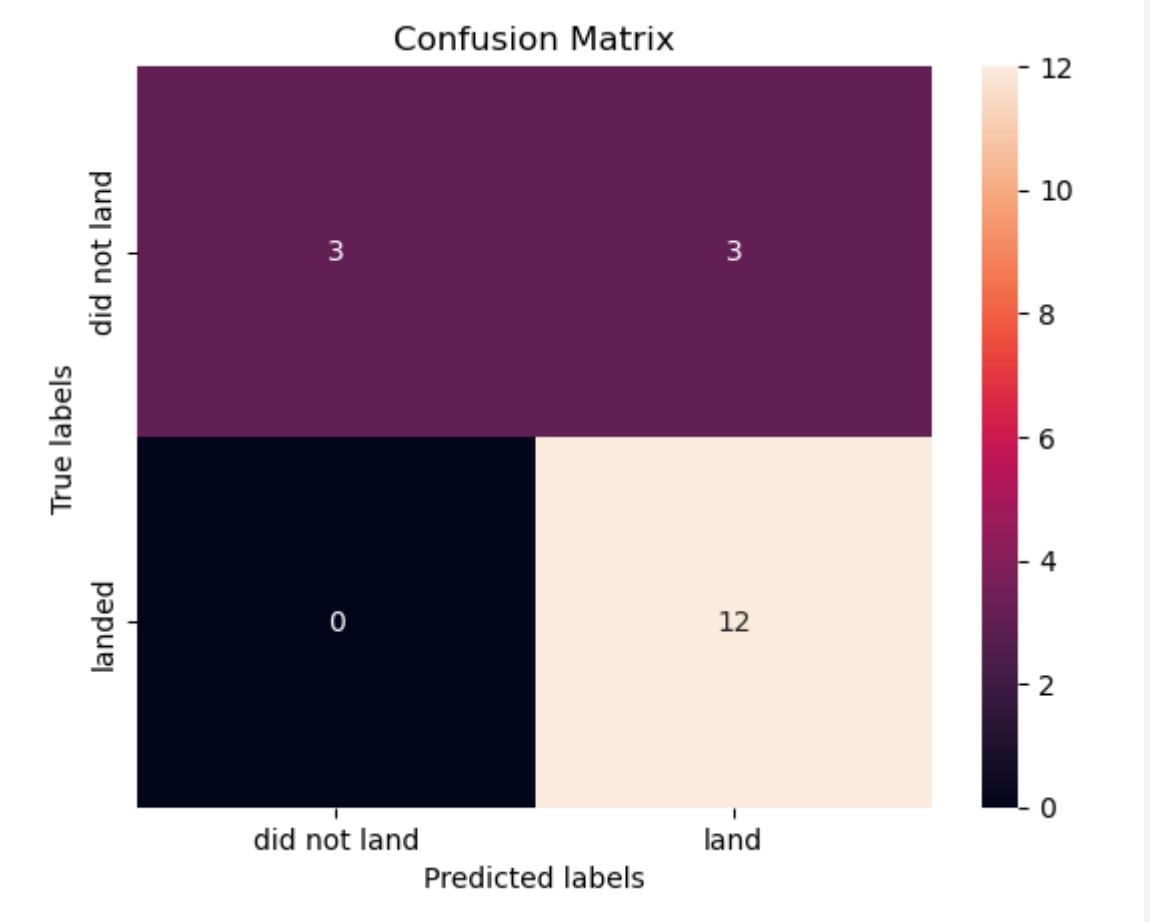
- In the test set, all four classification models had the same test accuracy of 83%.
- The model with the best classification accuracy was the Decision Tree with an accuracy of 88.9%.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.88929	0.83333
KNN	0.84821	0.83333



Confusion Matrix

- The confusion matrix is the same for all models because all models performed the same test set
- The models predicted 12 successful landings when the true label was successful and 3 unsuccessful landings when the true label was unsuccessful
- The model also predicted 3 successful landings when the label was unsuccessful.
- The Confusion Matrix confirmed the accuracy for the prediction of successful landings.



Conclusions

- Our task was to develop a machine learning model for SpaceY who plans to compete with SpaceX
- The model was to predict if the first stage will land successfully.
- The data used was from public SpaceX API and webscraping from SpaceX's Wikipedia page.
- The data sourced were used to create data labels, and then stored and analyzed.
- Dashboards were created for visualization to enable easy understanding of the data.
- The Machine Learning model was created and the Decision Tree Classifier had the highest accuracy of 88.9%, which predicts successful landings at Stage 1.
- SpaceY can use the Decision Tree model to predict with high accuracy that a launch will successfully land at Stage 1, which will save SpaceY money.

Appendix

- GitHub Repository: https://github.com/KWAMEOB/Applied-Data-Science-Capstone
- IBM Cloud: https://dataplatform.cloud.ibm.com/projects/872041fb-e350-4771-a85e-fde20ca04cee/assets?context=cpdaas

- Special Thanks to all the Instructors

https://www.coursera.org/professional-certificates/ibm-data-science#instructors

Thank you!