IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rakshith G
23/11/2023

Outline

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

Executive Summary

- Methodology
  - Data Collection and Data Wrangling
  - Exploratory Data Analysis using SQL and Visualization
  - Interactive Visual Analytics
  - Predictive Analysis using Machine Learning Models
- Results
  - Valuable data was obtained from publicly available sources and cleaned and analyzed to extract meaningful patterns
  - Machine Learning models were trained using the clean data to predict if the first stage of Falcon 9 will land successfully

Introduction

Space Y, founded by billionaire industrialist Allon Mask, is attempting to compete with SpaceX, in the race to commercialize safe space travel.

Taking up the role of a data scientist for the new rocket company, the main objective is to estimate the price of each launch using data science methodologies.

- Project Objectives:
  - To predict if the first stage of the rocket will land successfully
  - To find the optimal launch sites for lift-off

Section 1

Methodology

Methodology

Executive Summary

- Data Collection and Data Wrangling:
  - Data is obtained from publicly available sources – using RESTful API and by Webscraping
  - The data obtained is converted into a dataframe and cleaned using Data Wrangling techniques using Python code
- Exploratory Data Analysis (EDA):
  - EDA is then performed on the clean data using Python visualization and SQL to obtain actionable insights from the data

Methodology

Executive Summary

- Interactive Visual Analytics:
  - Folium is used to build an interactive map to analyze the launch site proximity
  - Plotly Dash is used to build interactive dashboard that contains pie charts and scatter plots to analyze data
- Predictive Analysis:
  - The data is normalized and split into training data and test data and evaluated by different classification models
  - Machine Learning models are trained to predict if the first stage of Falcon 9 will land successfully

Data Collection

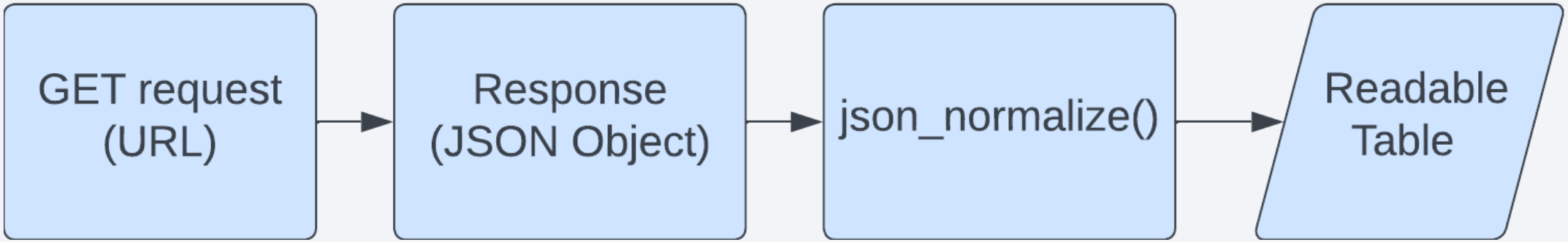
The data sets were collected from:

- SpaceX RESTful API - https://api.spacexdata.com/v4/launches/past
- Wikipedia, using Python web scraping techniques - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- The URL is followed to target a specific endpoint of the API
- A get request is performed using the requests library to obtain the launch data. The response will be in the form of a list of JSON objects
- json_normalize() function is then used to “normalize” the structured JSON data into a table

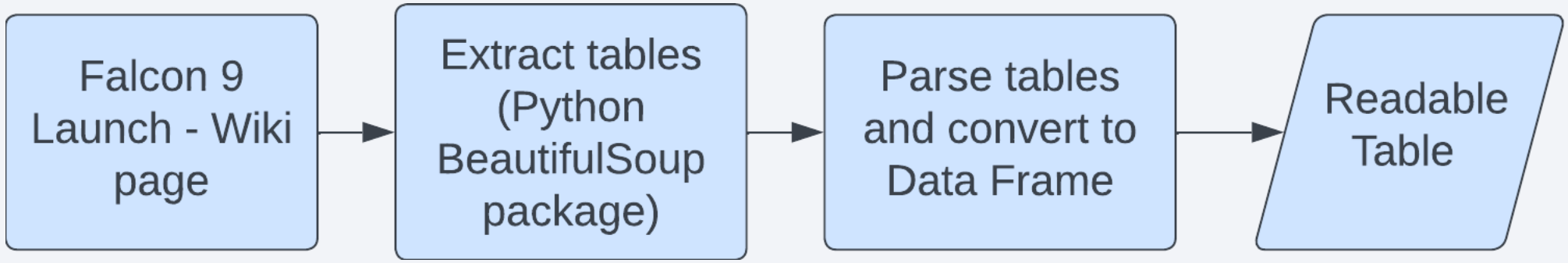
https://github.com/rkshthg/Capstone-Space-Race/blob/main/spacex-data-collection-api.ipynb



Data Collection - Scraping

- Python BeautifulSoup package is deployed to web scrape the HTML tables that contain Falcon 9 launch records
- The data from those tables are parsed and converted into a Pandas data frame
- After obtaining the data, some basic cleaning activities are performed, like – handling NULL values

https://github.com/rkshthg/Capstone-Space-Race/blob/main/webscraping.ipynb



Data Wrangling

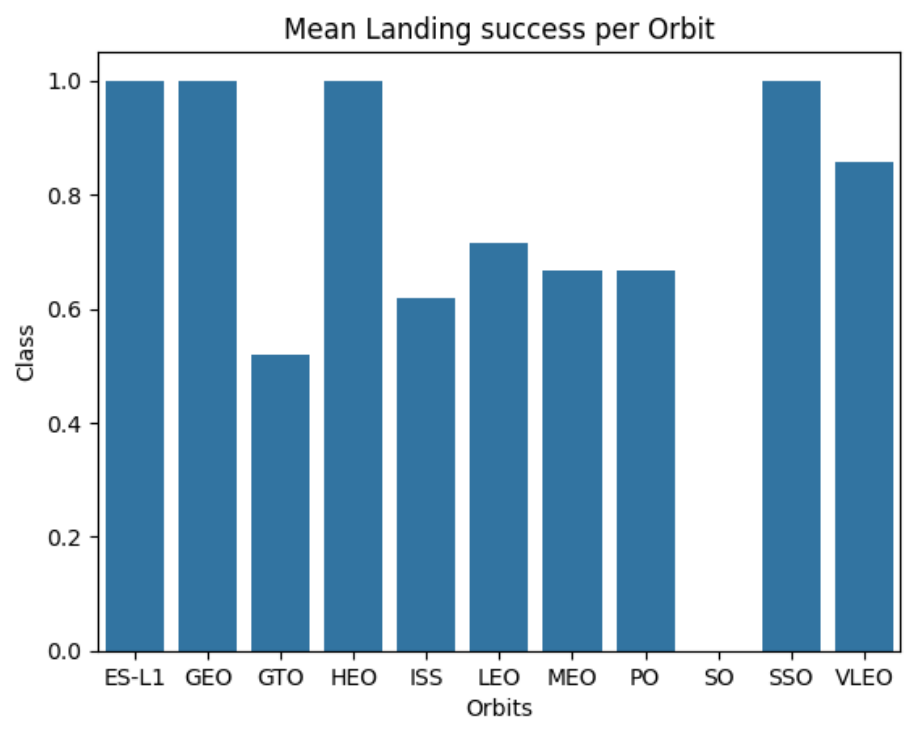
- In this step, the numerical and categorical columns are identified
- Some basic data analysis is performed to obtain quick summaries
- Landing outcomes are then labelled, which will be used to predict the successful landings

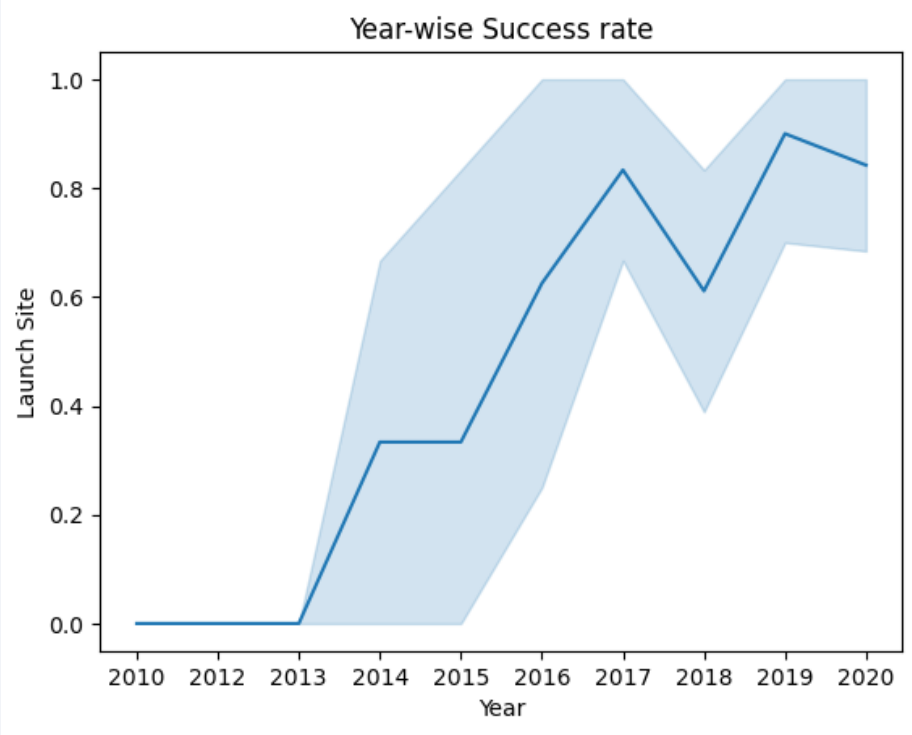
https://github.com/rkshthg/Capstone-Space-Race/blob/main/spacex-Data%20wrangling.ipynb

EDA with Data Visualization

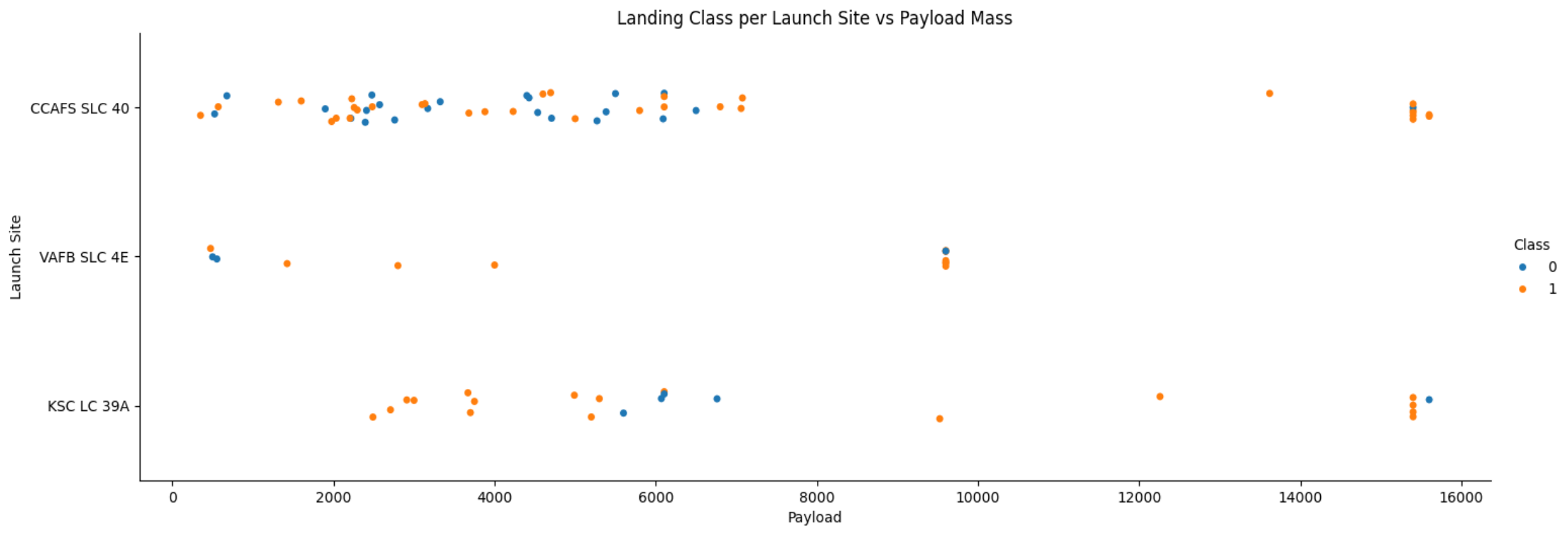
- The relationships between various pairs of features were explored using scatter plots and bar plots

https://github.com/rkshthg/Capstone-Space-Race/blob/main/eda-dataviz.ipynb.ipynb





EDA with Data Visualization



EDA with SQL

- Through this step, the Spacex data set is thoroughly understood using SQL queries.
- Valuable information is obtain, such as-
  - Names of the various Launch Sites
  - Total and Average Payload Mass carried by specific boosters
  - Date of the first successful landing outcome
  - Names of the boosters which have success in drone ship landings
  - Total number of successful and failed mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Ranks of count of landing outcomes

https://github.com/rkshthg/Capstone-Space-Race/blob/main/eda-sql.ipynb

Build an Interactive Map with Folium

- Circles, Markers, Marker Clusters and Lines were plotted using Folium
- Circles – to highlight the area around specific coordinates
- Markers – to indicate coordinates, like launch sites
- Marker Clusters – to indicate a group of events in a coordinate, like launches at one particular launch site
- Lines – to represent distances between coordinates

https://github.com/rkshthg/Capstone-Space-Race/blob/main/folium.ipynb

Build a Dashboard with Plotly Dash

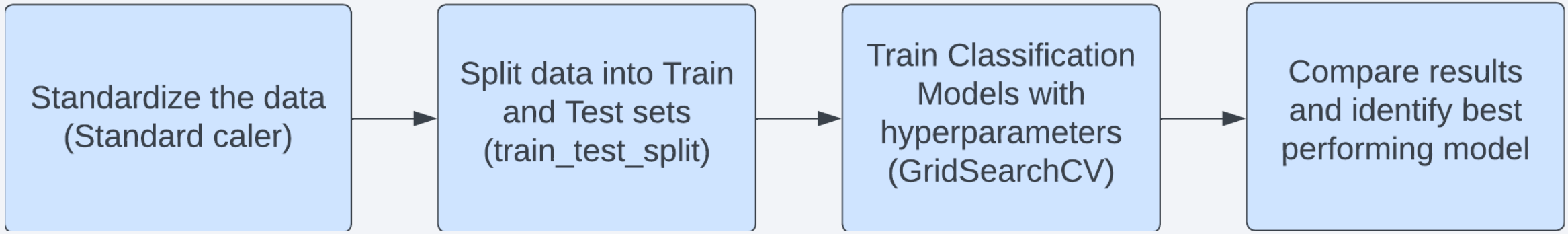
- The interactive dashboard consists of a dropdown menu to select Launch Sites and a range slider to filter Payload Mass.
  - By selecting a Launch Site, a pie chart of percentage of launch results is plotted
  - In combination with the selected Launch Site and Payload Mass, a scatter chart is plotted
- The combination allows quick analysis to identify which is the best launch site for a certain payload mass

https://github.com/rkshthg/Capstone-Space-Race/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The data is first standardized and then split into training and test sets
- Four classification models, Logistic Regression, SVM, Decision Tree, K-Nearest Neighbors, are trained using the training data with a combination of hyperparameters
- The four models are compared to identify which has the best accuracy

https://github.com/rkshthg/Capstone-Space-Race/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

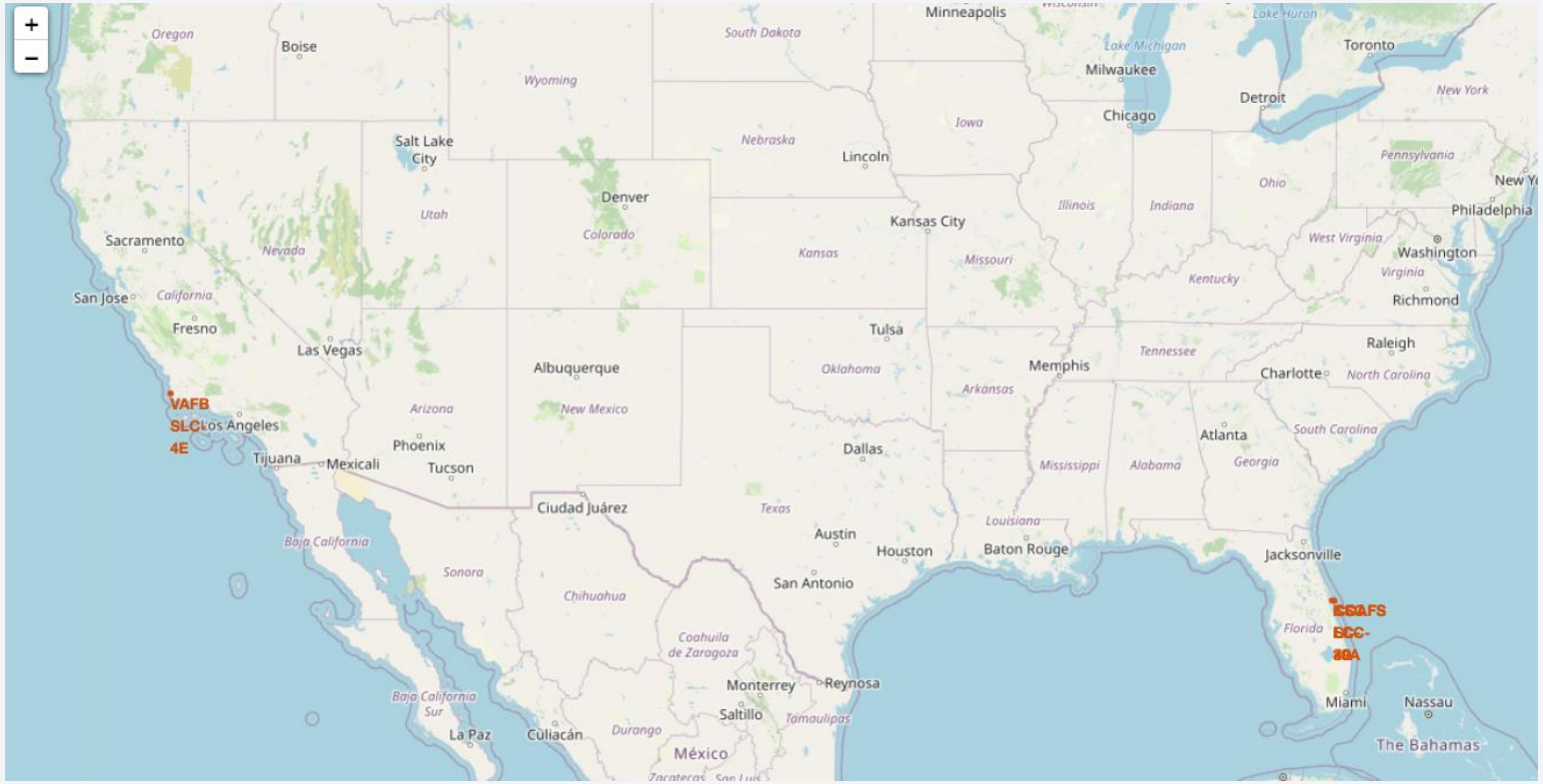


Results

- Exploratory Data Analysis:
  - SpaceX uses 4 distinct Launch Sites
  - Total of 45596 Kg of Payload has been launched by NASA
  - Average of 2534.66 Kg of Payload is carried by F9 v1.1% Booster Version
  - The first successful landing was on 2015-12-22
  - Boosters F9 FT B1022, B1026, B1021.2 and B1031.2 have success in drone ship landings
  - Nearly 100% of mission outcomes were successful
  - Number of landing outcomes have improved over time
  - Launch Site VAFB SLC 4E has seen most success

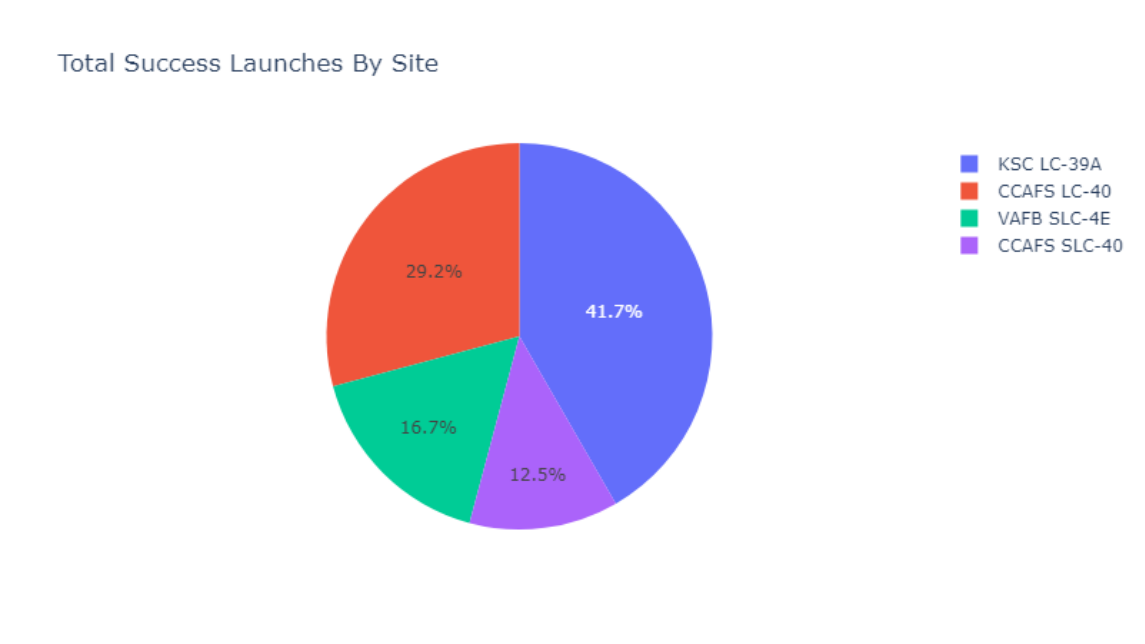
Results

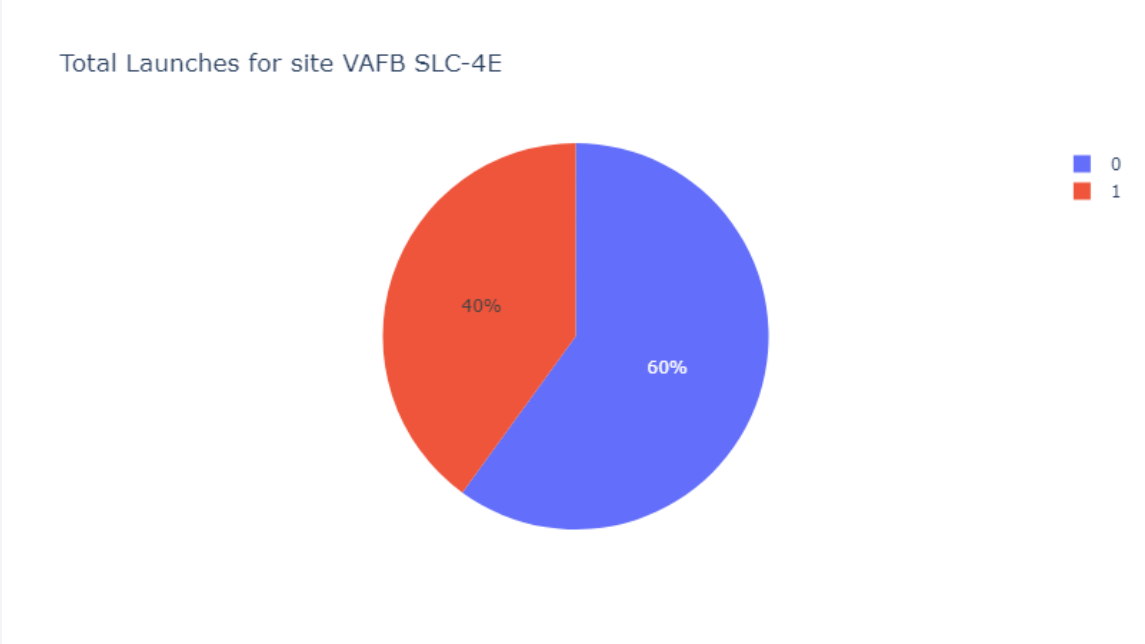
- Interactive Analysis:
  - All launch sites are situated near the sea
  - 75% of Launch Sites are situated on the Eastern Coast of USA



Results

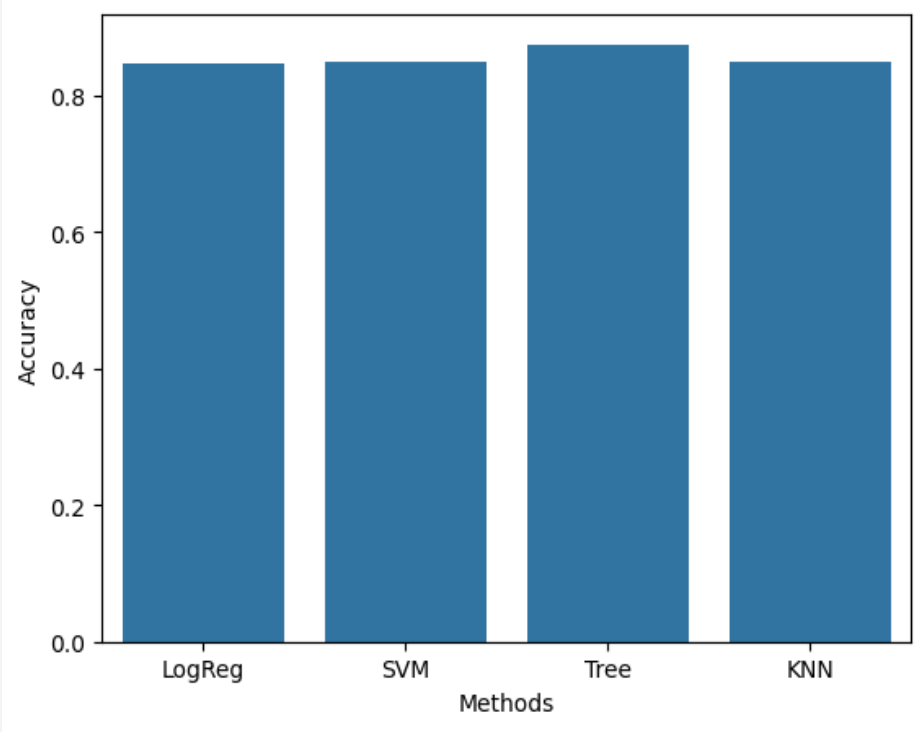
- Interactive Analysis:
  - KSC LC-39A boasts highest success rate with 41.7%
  - VAFB SLC-4E has a success rate of 60%

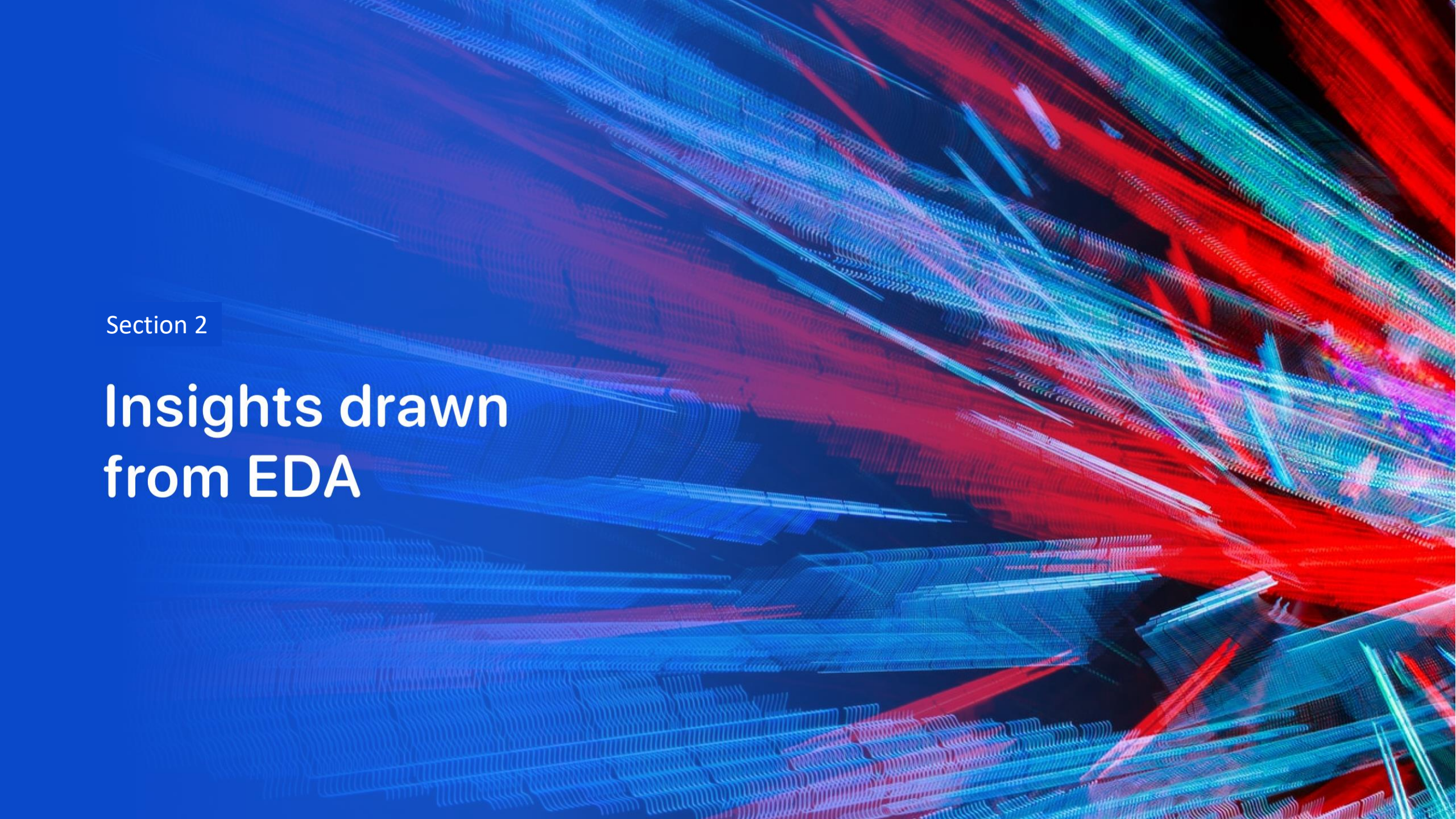




Results

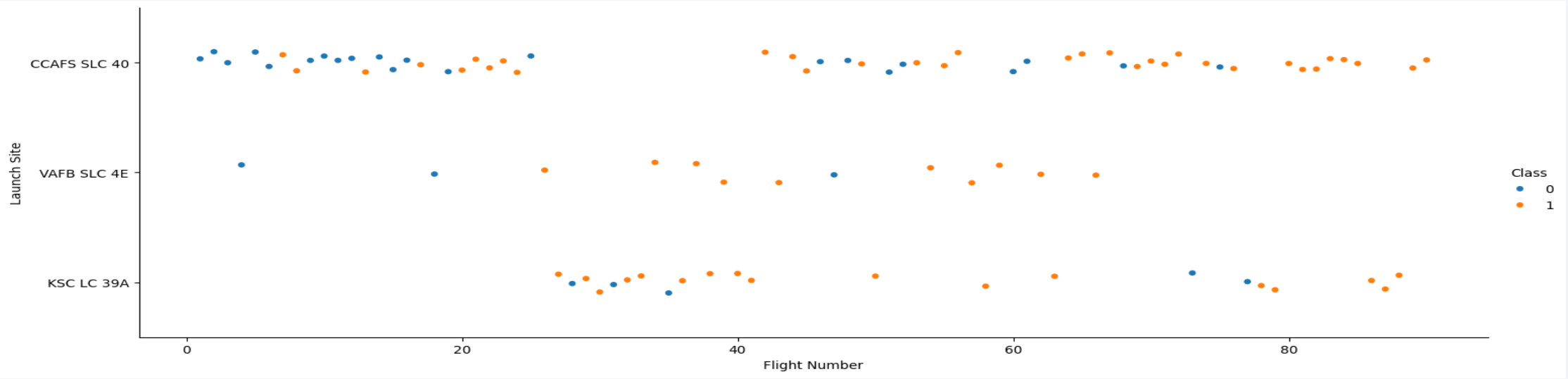
- Predictive Analysis:
  - Four classification models, Logistic Regression, SVM, Decision Tree, K-Nearest Neighbors, are trained using the training data with a combination of hyperparameters
  - On comparing the models, it is found that Decision Tree model shows the best accuracy




Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

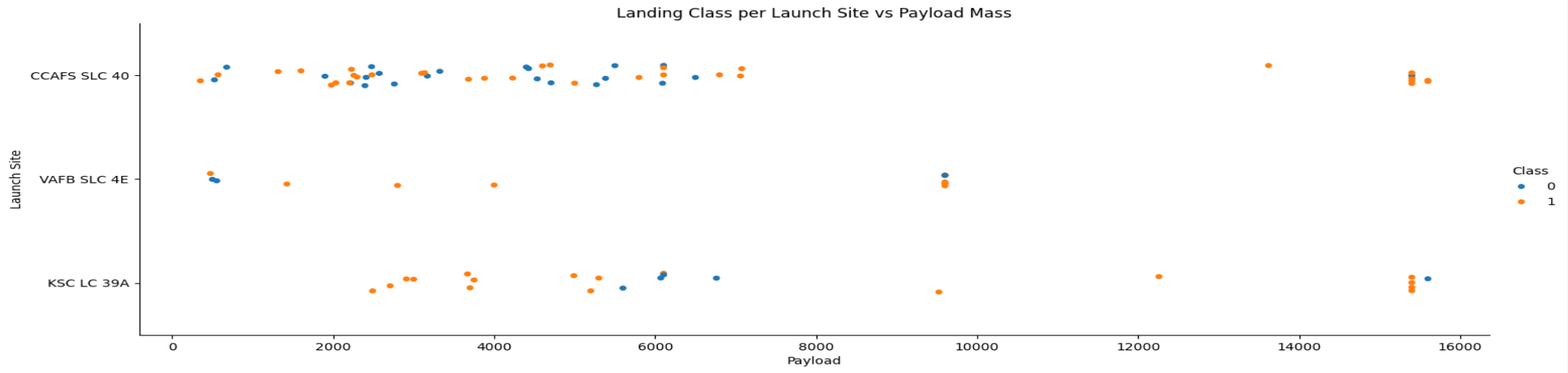
- Scatterplot of Flight Number vs. Launch Site



- CCAFS SLC 40 is the most used launch site
- The Launch Site with highest success rate is seen to be VAFB SLC 4E
- A trend of increasing success rate can be observed

Payload vs. Launch Site

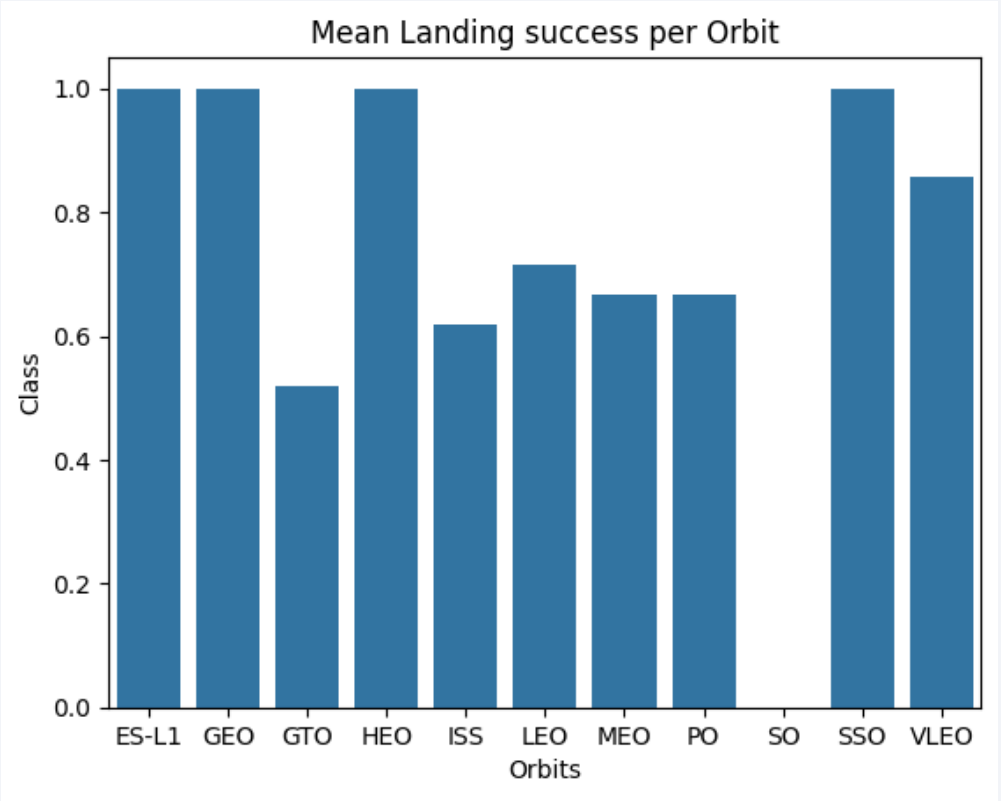
- Scatterplot of Payload vs. Launch Site



- Payloads with mass over 9000 Kg is observed to have a high success rate
- Payloads with mass over 10000 Kg are observed to be launched only from CCAFS SLC 40 and KSC LC 39A

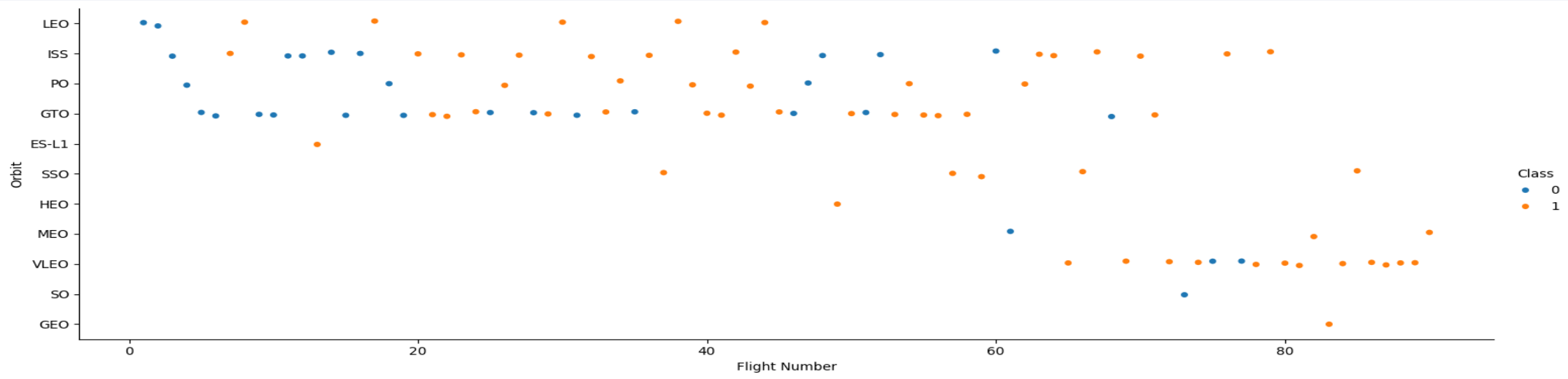
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
  - 100% success rates are seen for
    - ES-L1
    - GEO
    - HEO
    - SSO
  - VLEO orbit type shows a success rate of over 80%
  - SO orbit type has the highest failure rate of 100%



Flight Number vs. Orbit Type

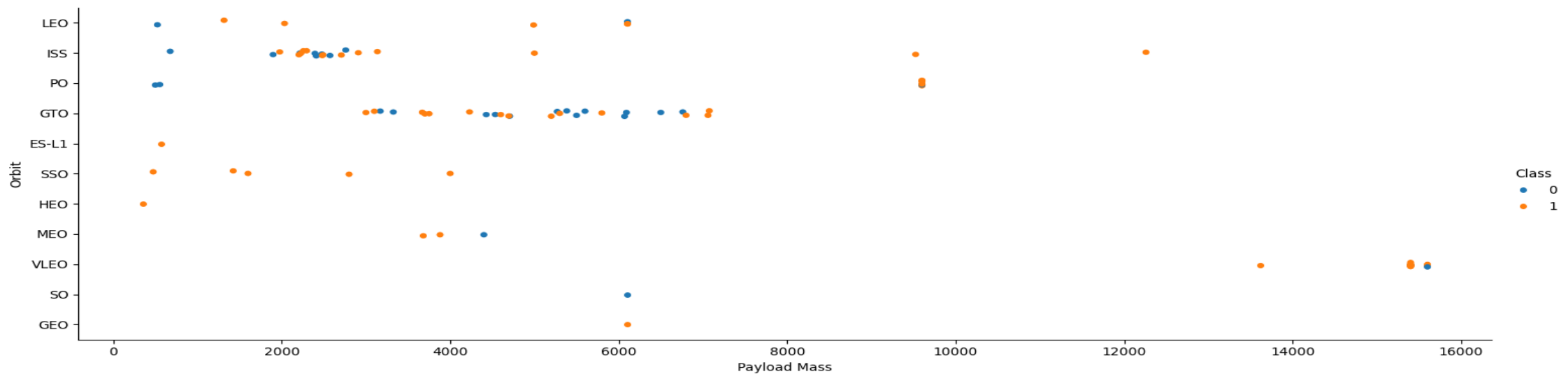
- Scatterplot of Flight number vs. Orbit type



- VLEO appears to be a new type, with increase in frequency of flights
- There appears to be an improvement in rate of success for all orbit types

Payload vs. Orbit Type

- Scatterplot of payload vs. orbit type



- ISS orbit appears to carry the largest range of payload mass
- Payloads of over 8000 Kg is observed to have an orbit type of VLEO, ISS and PO
- VLEO type is observed to be carrying the heaviest Payload Mass

Launch Success Yearly Trend

- Line chart of yearly average success rate
  - There is an observed upward trend in successful launches since 2013
  - Success rate dipped a small amount in 2018, before quickly recovering the following year
  - Rate of successful launches were highest in the year 2019

All Launch Site Names

- Unique launch sites
  - A simple SELECT query is used

```
select distinct "Launch_Site" from SPACEXTABLE
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
  - A simple select query using a wildcard match is used

select * from SPACEXTABLE 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	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

Total Payload Mass

- Total payload carried by boosters from NASA
  - An aggregation query is used, with a wildcard match

```
select sum("PAYLOAD_MASS__KG_") as "Total Payload" from SPACEXTABLE where
Customer='NASA (CRS)'
```

Total Payload
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
  - An aggregation query is used, with a wildcard match

```
select avg("PAYLOAD_MASS__KG_") as "Average Payload" from SPACEXTABLE where
"Booster_Version" like 'F9 v1.1%'
```

Average Payload
2534.6666666666665

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad
  - An aggregation query is used, with a string match

```
select min(Date) as “Date” from SPACEXTABLE where
"Landing_Outcome"='Success (ground pad)'
```

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - A select query is used, with string matching and logical conditions

```
select distinct "Booster_Version" from SPACEXTABLE
where "Landing_Outcome"='Success (drone ship)' and
"PAYLOAD_MASS__KG_">4000 and
"PAYLOAD_MASS__KG_"<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

- Total number of successful and failure mission outcomes
  - A select query and an aggregation query is used, with a group by function

    select "Mission_Outcome",count("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome"

Mission_Outcome	count("Mission_Outcome")
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
  - A select query is used with a sub-query

```
select "Booster_Version" from SPACEXTABLE where
"PAYLOAD_MASS__KG_"=(select max("PAYLOAD_MASS__KG_")
from SPACEXTABLE)
```

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

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - A substring query is used with a substring math and string match functions

select substr(Date, 6,2) as month, "Landing_Outcome", "Booster_Version", "Launch_Site“ from SPACEXTABLE where substr(Date,0,5)='2015' and "Landing_Outcome"='Failure (drone ship)'

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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
  - A select and an aggregation query is used, with date matching, group by and order by statements

```
select "Landing_Outcome", count(*) from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by "Landing_Outcome" order by count("Landing_Outcome") desc
```

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

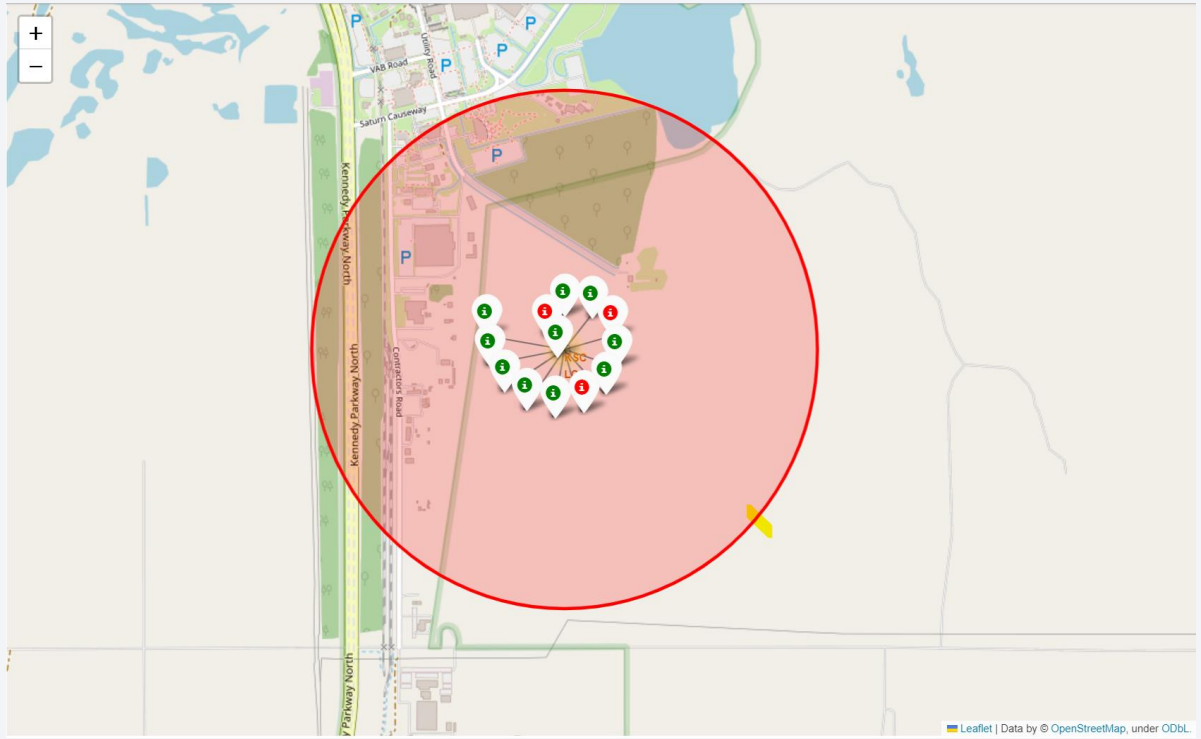
Section 3

Launch Sites Proximities Analysis

Launch Sites – Circle Markers

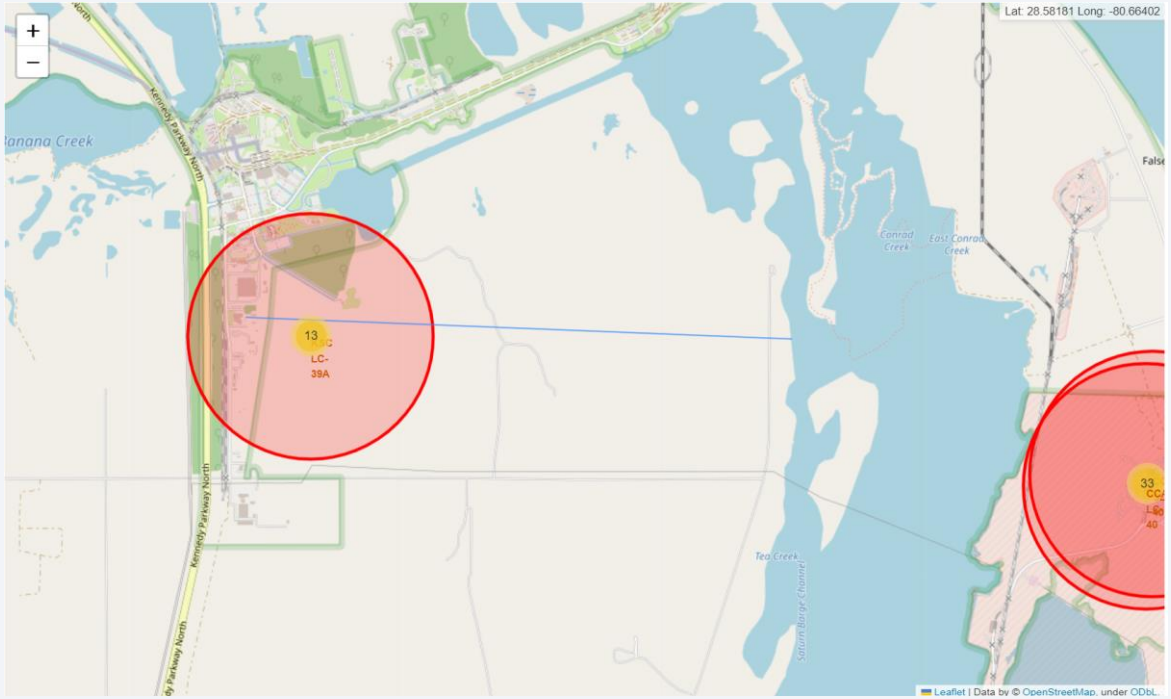
- The Yellow Circles seen on the map show the launch sites used by Space X
- It is observed that all launch sites are situated at the coasts
- 3 out of 4 launch sites are on the Eastern coast of USA

Success Rate of Launch Sites

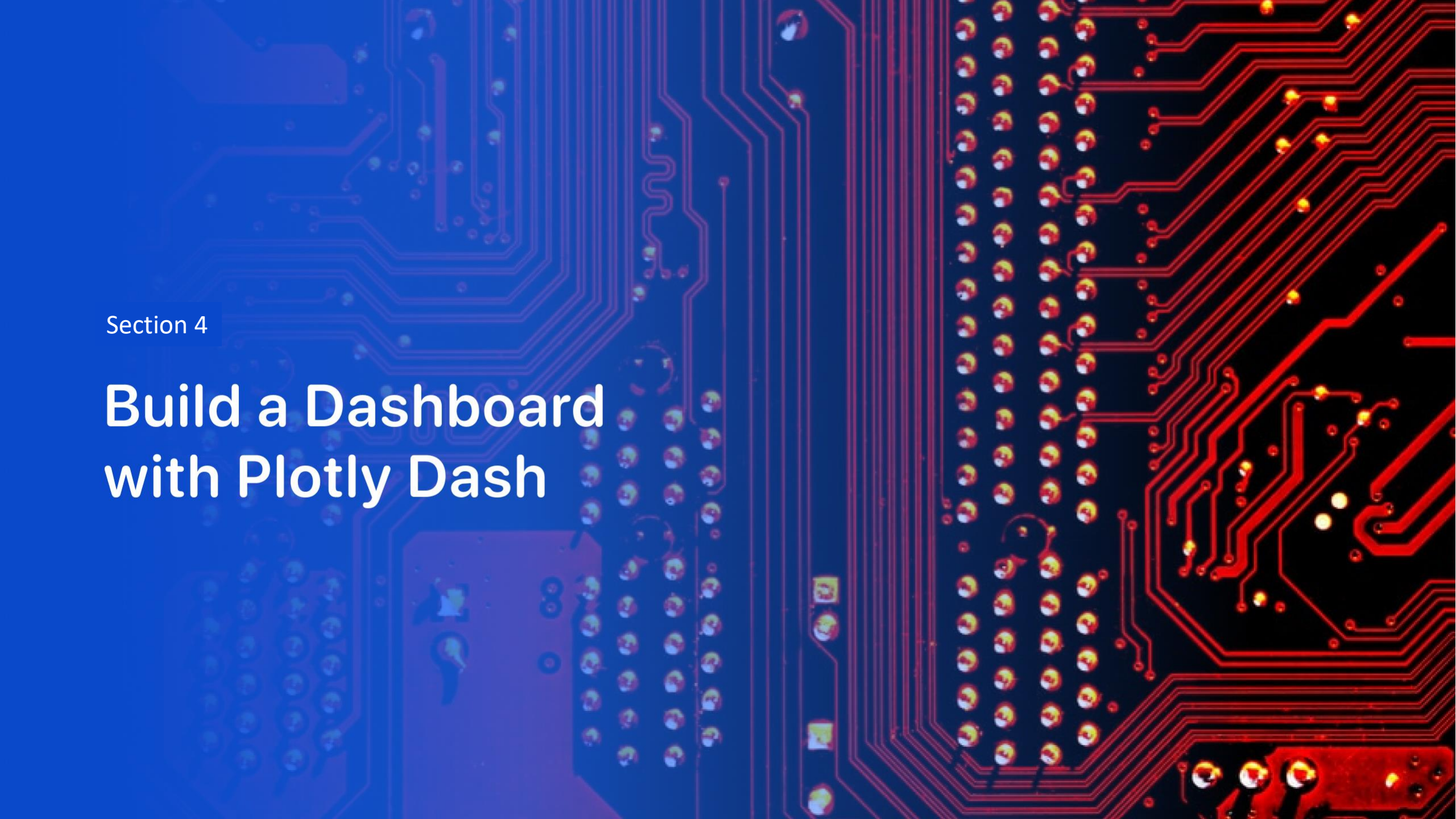


- The circle marker, when clicked, shows event markers showing the launches and the result of the landings.
- The GREEN markers show a successful landing, while the RED markers show failed landing

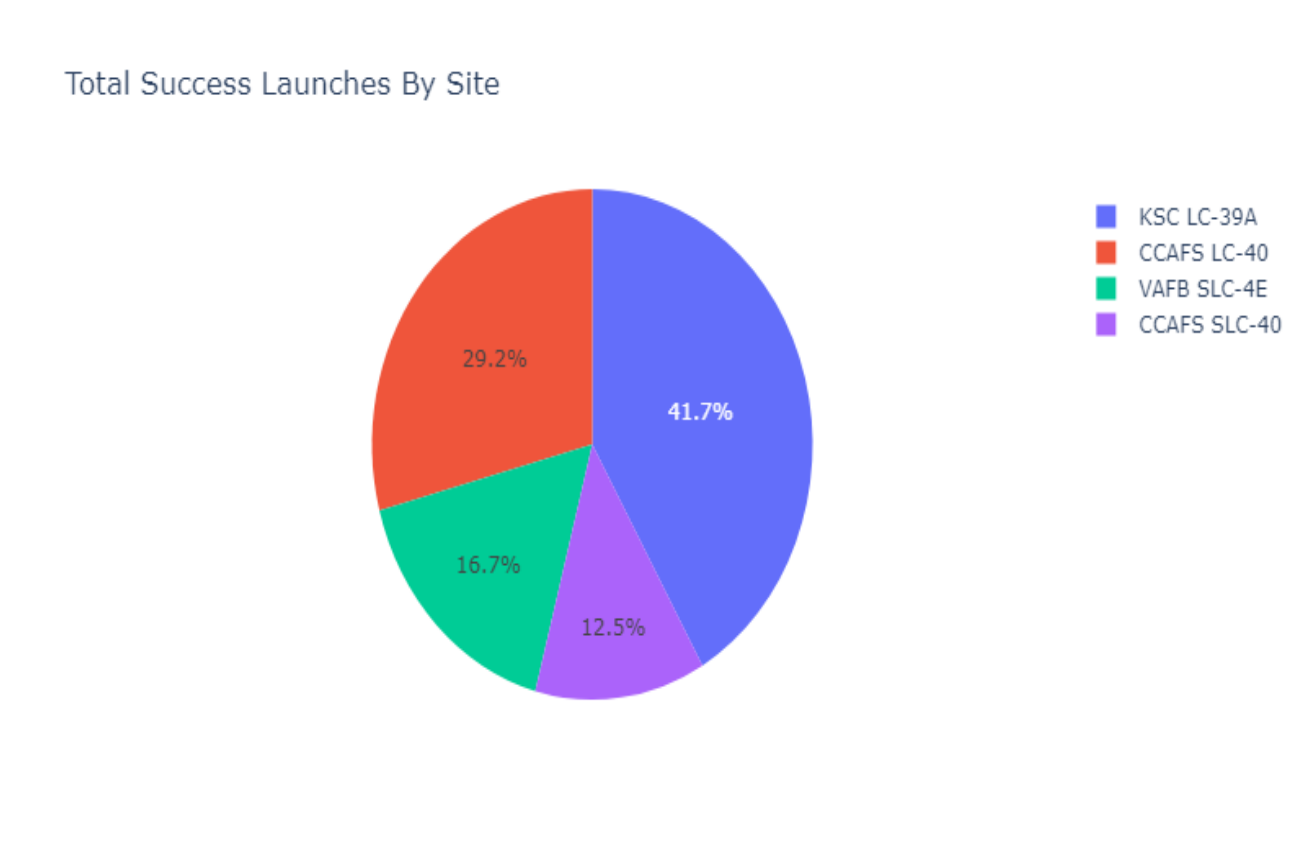
Proximity Map



- The BLUE line from the launch site to the coast shows the distance.

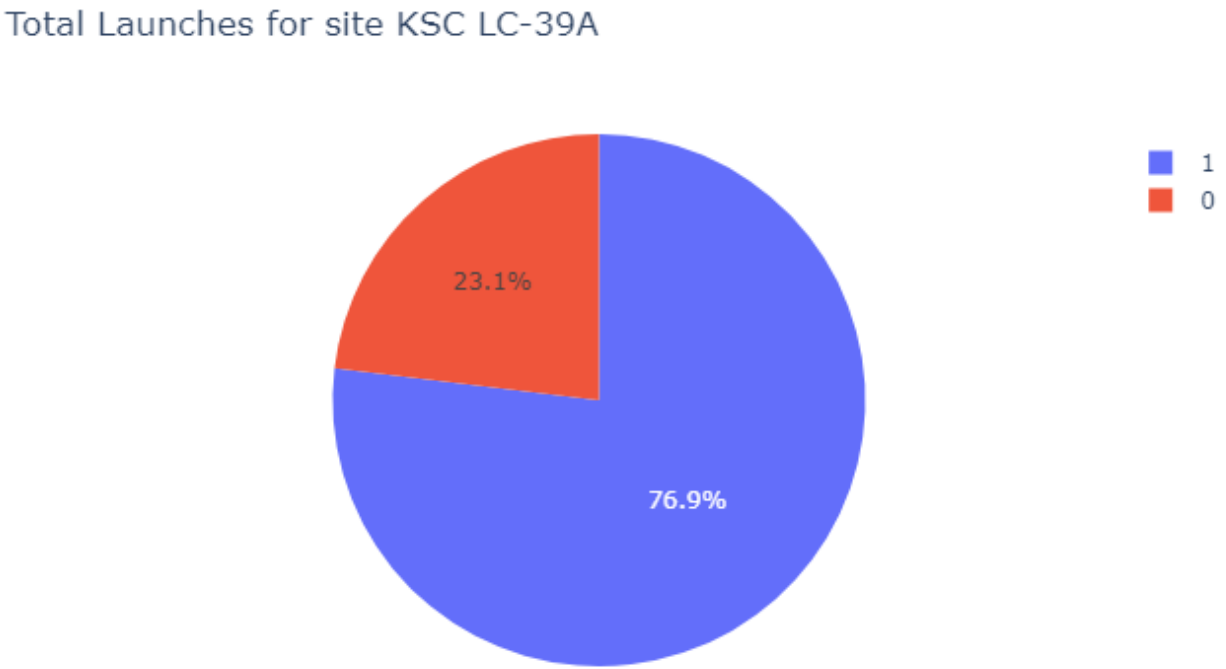

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



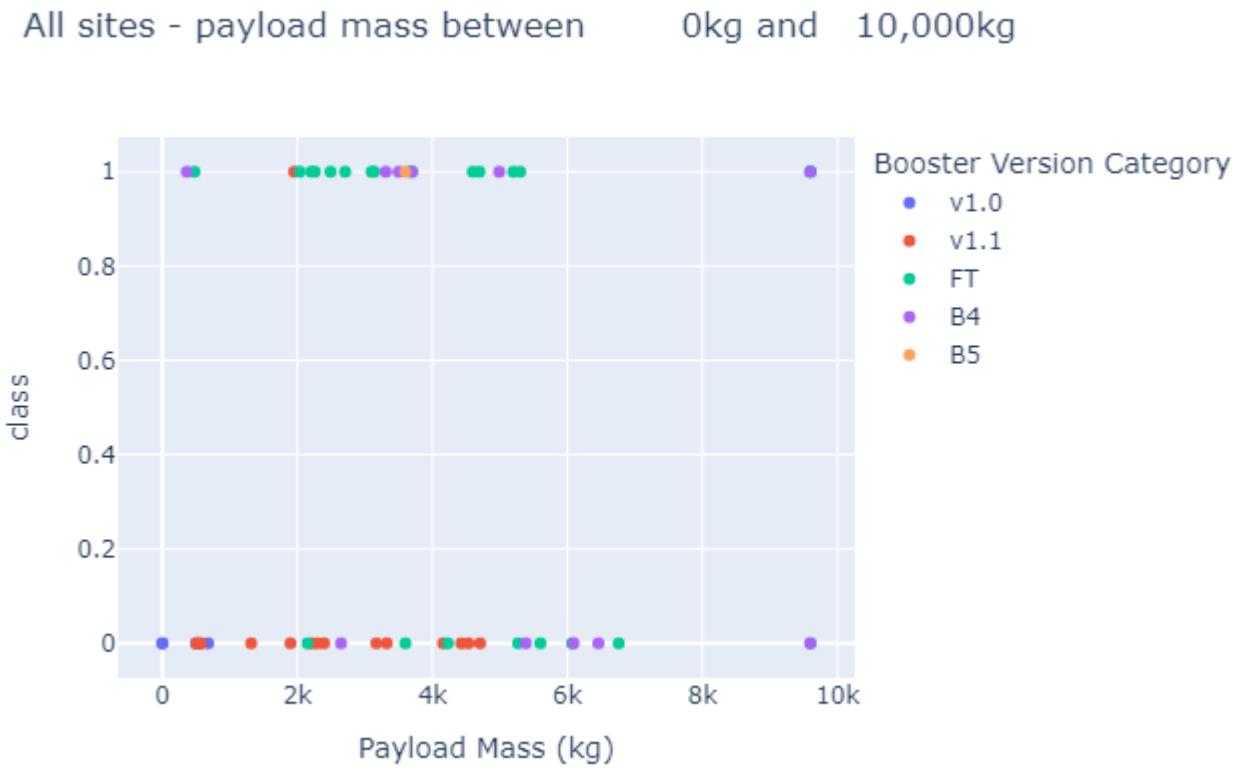
- From the pie chart, it is observed that the launch site KSC LC-39A has the highest rate of successful launches
- CCAFS SLC-40 has the lowest success rate compared to other launch sites

Total launches for site KSC LC-39A



- Launches made from site KSC LC-39A see the highest success rate, of 76.9%
- Only 23.2% of the launches have resulted in failed landings

All Sites – Payload vs Class Scatterplot



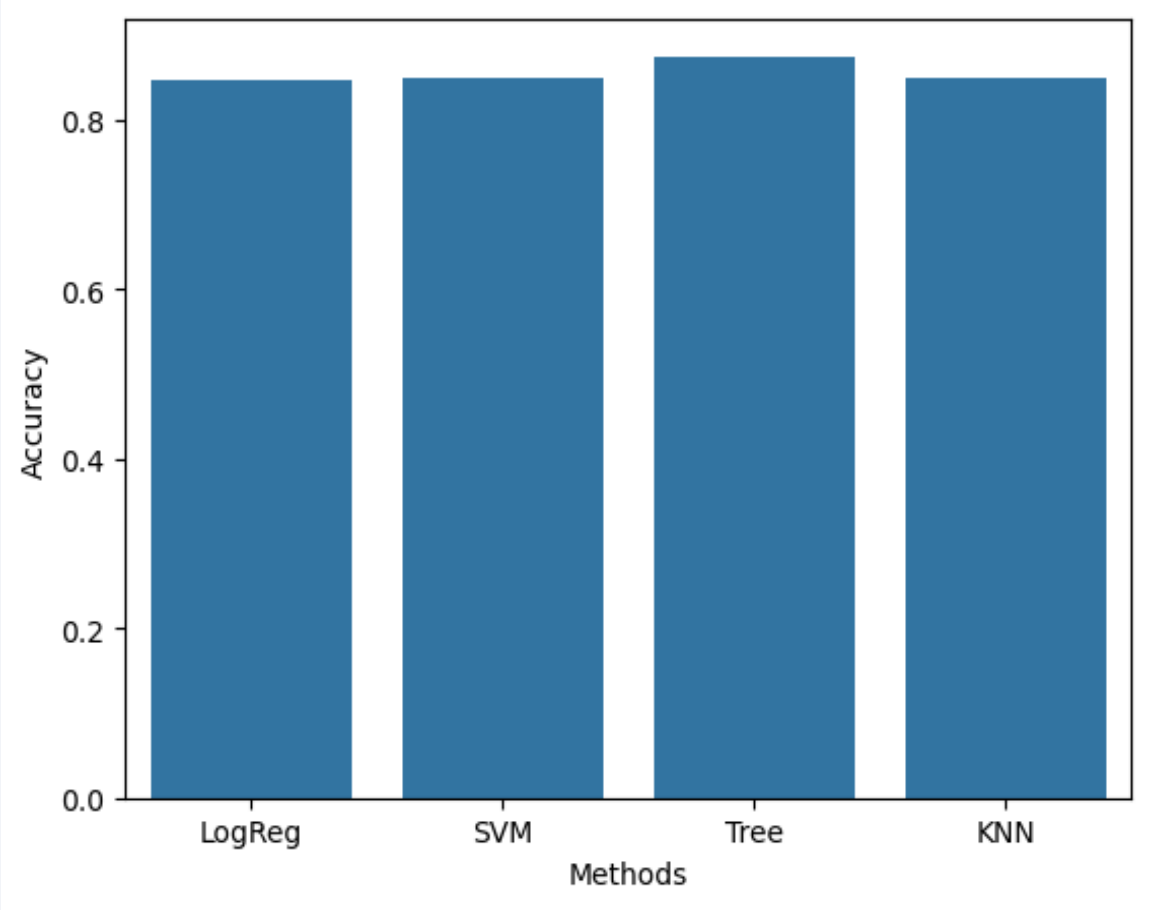
- As observed from the scatterplot, FT Boosters show a high success rate between 2000 Kg and 6000Kg
- Booster Version v1.1 show the lowest success rate at any Payload Mass

Section 5

Predictive Analysis (Classification)

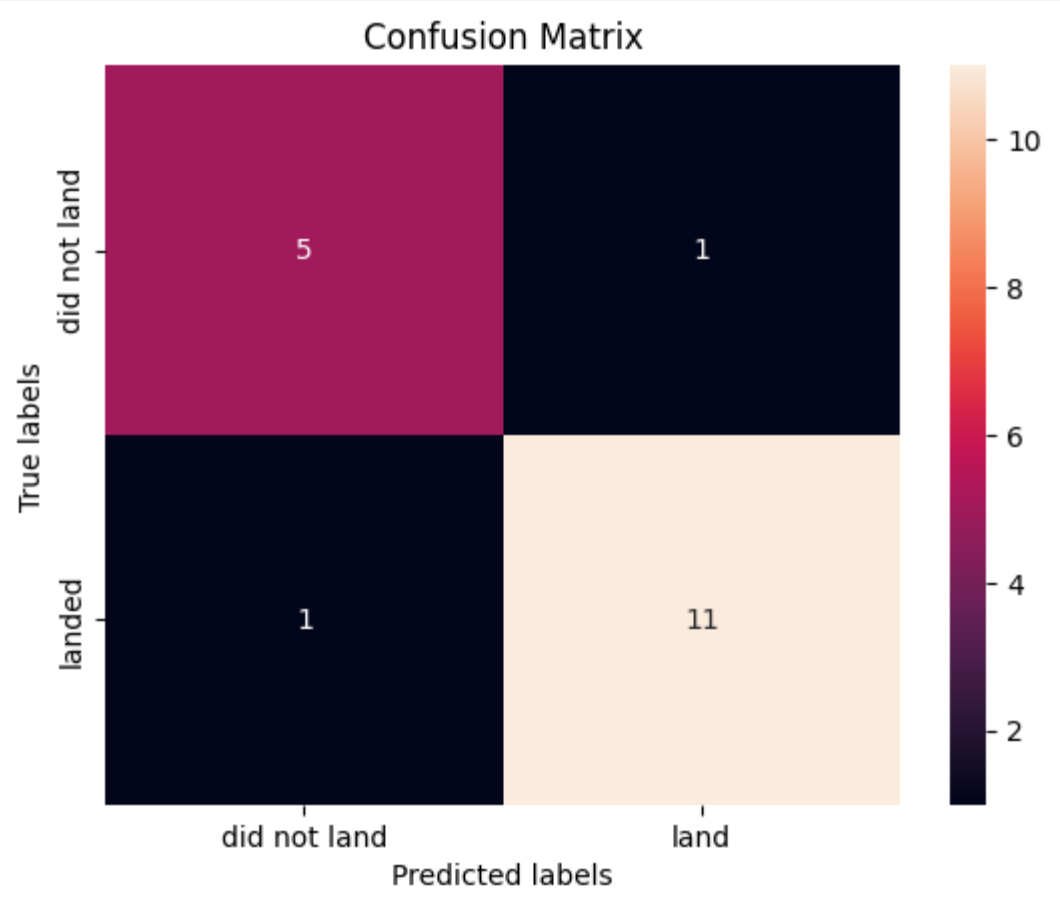
Classification Accuracy

- By visualizing the model accuracy for all built classification models, in a bar chart, the following observations can be made:
  - Decision Tree model boasts the highest accuracy
  - All models have a prediction accuracy of over 80%



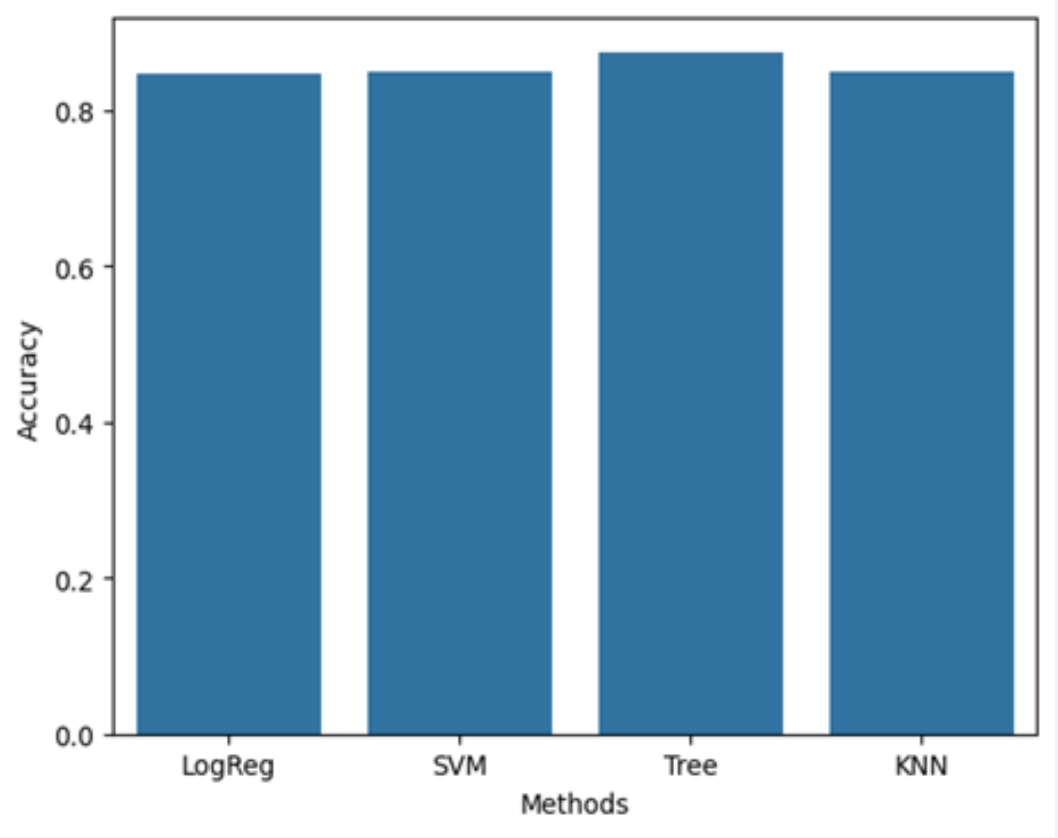
Confusion Matrix

- The confusion matrix of the best performing model – Decision Tree Classifier is shown in the adjoining figure
- Of the 6 unsuccessful results, the model predicted 5 (True Negative) correctly and misclassified 1 as successful (False Negative)
- Of the 12 successful results, the model predicted 11 (True Positive) correctly and misclassified 1 as successful (False Positive)



Conclusions

- The Decision Tree Classifier model is observed to be the best model to predict the successful landing, with an accuracy of 88.89%
- The model predicted 5 (True Negative) correctly
- The model misclassified 1 as successful (False Negative)
- The model predicted 11 (True Positive) correctly
- The model misclassified 1 as successful (False Positive)



Thank you!