The Economics of Space Launch with Data Science

Ailleen A. Gonzales

August 8, 2023

Outline

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

Executive Summary

Summary of methodologies

- Data collection from an API and Wikipedia page
- Data Wrangling to assign a classification variable to label the outcome of the launch
- Exploratory data analysis (EDA) using visualization and SQL to visualize the relationships of the different attributes of the data
- Interactive visual analytics using Folium to generate an interactive map of the launch sites, and Plotly to create various charts for analysis
- Perform predictive analysis using classification models

Summary of all results

- KSC LC-39A launch site has the highest success rate. Orbit types ES-L1, GEO, HEO, and SSO have the highest success rates.
- The success rate of the launches has an over-all upward trend yearly. The rockets in launch sites use Pay Load Mass (kg) not greater than 8000. There are few launches using heavy Pay Load Mass (kg).
- All launch sites are close in proximity to the equator , thus, the rockets get added advantage of earth's rotational speed. All the launch sites were also built away from populated areas and close to coastlines to ensure the safety of major population centers.
- All methods performed well on the test data with the Decision Tree method having the highest accuracy.

Introduction

Project background

- In the new space age, companies are making commercial space exploration affordable for everyone.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars. One reason SpaceX can do this is the rocket launches are relatively inexpensive. Other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can recover the first stage.
- For this project, a new rocket company, SpaceY, would like to compete with SpaceX. By gathering information about SpaceX and creating dashboards, it can be determined if SpaceX can reuse the first stage.

Problems to find answers

- Which site has the highest launch success rate?
- Find relationships between attributes using exploratory data analysis
- Find some geographical patterns about launch sites
- Find the method that performs best using test data


Methodology

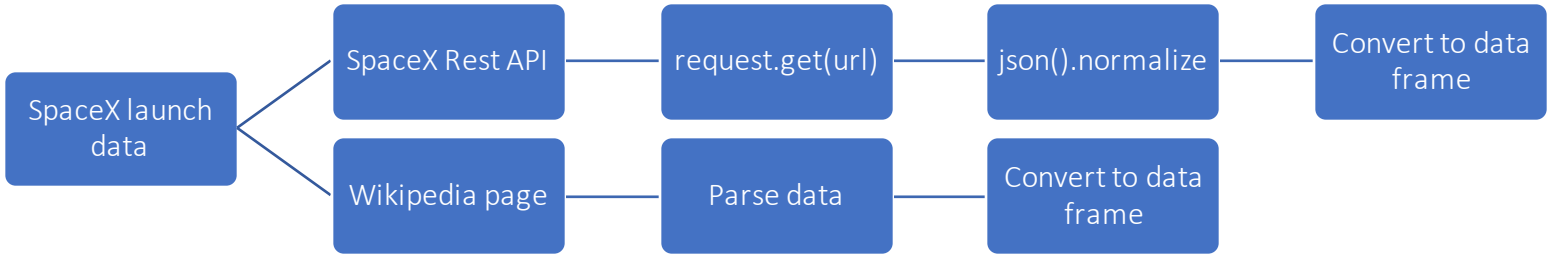
Methodology

- **Data collection**
  - The SpaceX launch data is gathered from an API, specifically SpaceX REST API and web scraping to collect Falcon 9 historical launch records from a Wikipedia page
- **Perform data wrangling**
  - Performed Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - Exploratory Data Analysis was performed and the Training Labels were determined. Then the best hyperparameter for SVM, Decision Trees, KNN and Logistic Regression was calculated. The method that performs best using test data was analyzed.

Data Collection

- The SpaceX launch data is gathered from an API, specifically SpaceX REST API (URL)
  - https://api.spacexdata.com/v4/launches/past)

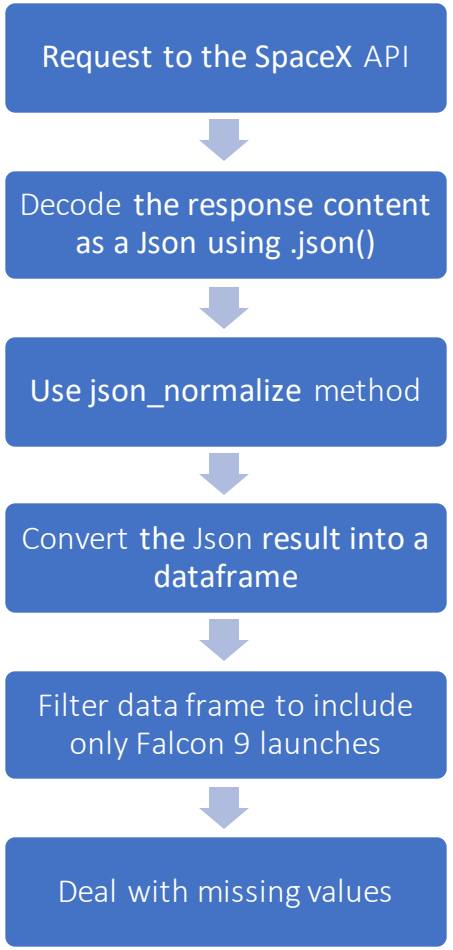
- Another popular data source for obtaining Falcon 9 launch data is web scraping related Wikipedia pages
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



Data Collection – SpaceX API

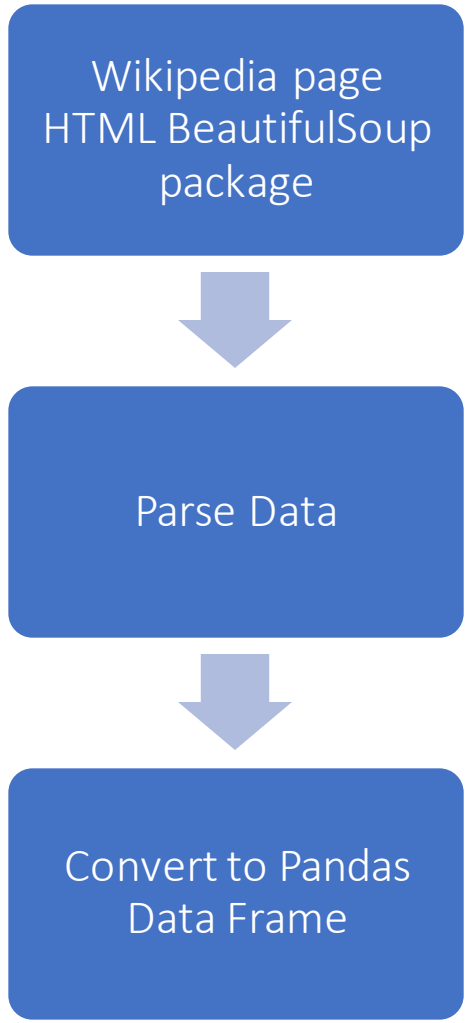
Data collection with SpaceX REST calls

- Collect and make sure the data is in the correct format from an API
- URL = "https://api.spacexdata.com/v4/launches/past"
- Request to the SpaceX API
- Parse the SpaceX launch data using the GET request
- Use json_normalize method to convert the Json result into a data frame
- Clean the requested data
- Filter the data frame to only include Falcon 9 launches
- Dealing with Missing Values
- https://github.com/Ilyn0402/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb



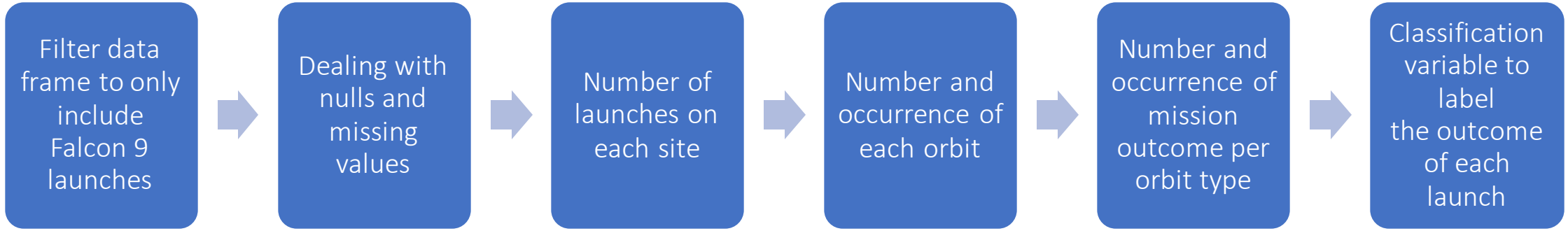
Data Collection - Scraping

- Performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Extract Falcon 9 launch records HTML table from Wikipedia.
- Parse the table and convert it into a Pandas data frame.
- https://github.com/Ilyn0402/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Perform Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models. Then convert those outcomes into Training Labels with 1, means the booster successfully landed and 0 means it was unsuccessful.



- https://github.com/Ilyn0402/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- **SCATTER POINT CHART**
  - Visualize the relationship between Flight Number and Launch Site
  - Visualize the relationship between Payload and Launch Site
  - Visualize the relationship between Flight Number and Orbit type
  - Visualize the relationship between Payload and Orbit type
- **BAR CHART**
  - Visualize the relationship between success rate of each orbit type
- **LINE CHART**
  - Visualize the launch success yearly trend
- https://github.com/Ilyn0402/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

SQL Queries

- 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 by using a subquery
- 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
- 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
- https://github.com/llyn0402/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Objects created and added to folium map

- Map object, as initial center location to be NASA Johnson Space Center at Houston, Texas.
- Folium.Circle, to add a highlighted circle area with a text label on a specific coordinate based on its latitude and longitude coordinate values
- Folium.Circle and Folium.Marker for each launch site on the site map
- MarkerCluster() create markers for all launch records. If a launch was successful (class=1), a green marker was used and if a launch failed, a red marker (class=0) was assigned. From the color-labeled markers in marker clusters, launch sites that have relatively high success rates can be easily identified.
- MousePosition was added to get coordinates for a mouse over a point on the map. As such, while exploring the map, the coordinates of any points of interests (such as railway) can be found easily. For example, using MousePosition, the distance between the coastline point and the launch site can be calculated.
- PolyLine object was used to draw between a launch site to its closest city, railway, highway, coastline, etc. MousePosition was used to find the coordinates on the map first.
- https://nbviewer.org/github/llyn0402/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

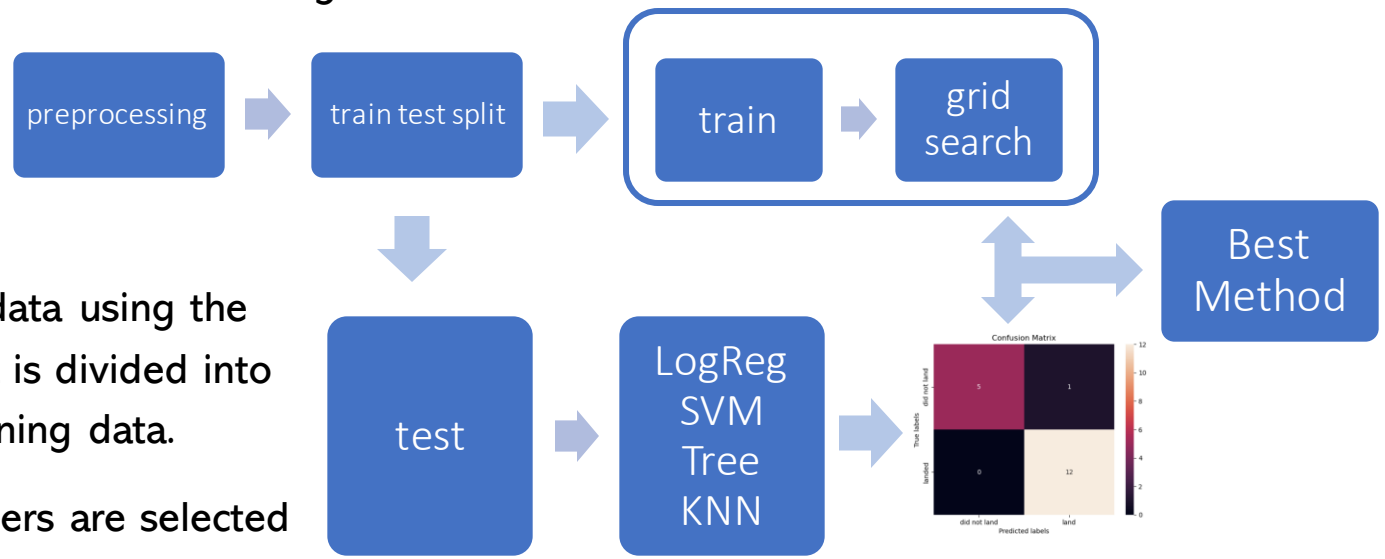
Plots, graphs and interactions added to the dashboard

- Add a Launch Site Drop-down Input Component to select different launch sites.
- Add a callback function to render pie chart based on selected site dropdown. The general idea of this callback function is to get the selected launch site from site-dropdown and render a pie chart visualizing launch success counts.
- Add a Range Slider to Select Payload. The range slider interact with the pie chart and scatter point chart to find if variable payload is correlated to mission outcome. Different payload ranges are easily selected to see if some visual patterns can be identified.
- Add a callback function to render the scatter point plot with the x axis to be the payload and the y axis to be the launch outcome (i.e., class column). So as to visually observe how payload may be correlated with mission outcomes for selected site(s). In addition, color-label was added to the Booster version on each scatter point so as to observe mission outcomes with different boosters.
- https://github.com/llyn0402/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Dashboard%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

Find the best performing classification model

- Create a machine learning pipeline to predict if the first stage will land given the data.
- Create a column for the class.
- Standardize the data.
- Split the data into training data and test data using the function train_test_split. The training data is divided into validation data, a second set used for training data.
- The models are trained and hyperparameters are selected using the function GridSearchCV.
- Find the best hyperparameter for SVM, Decision Trees, KNN and Logistic Regression.
- Find the method that performs best using test data.



- https://github.com/Ilyn0402/IBM-Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

Exploratory data analysis results

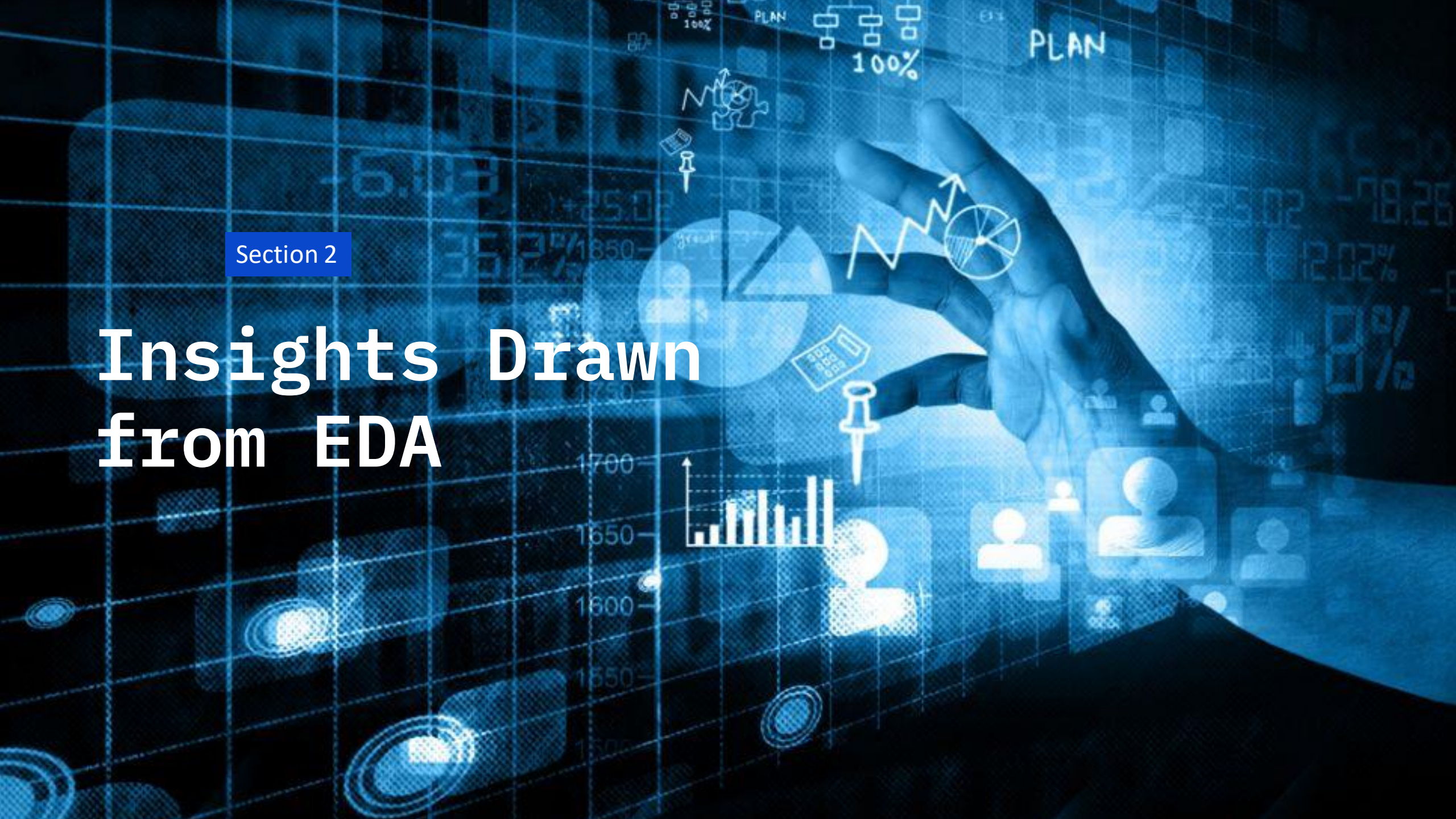
- The rockets launch in launch sites have payload mass (kg) not greater than 10000.
- The success rate of the launches has an over-all upward trend yearly.
- Orbit types ES-L1, GEO, HEO, and SSO have the highest success rates.
- Orbit type LEO was used during the earlier launch attempts. Orbit types ISS, PO, GTO were used more extensively on the continuous launch attempts. Later, continuous launch attempts were done on the VLEO orbit with good success rate.

Interactive analytics demo in screenshots

- KSC LC-39A launch site has the highest total success.
- Payload range from 3000 kg to 6000 kg has the highest success rate. Payload range higher than 6000 kg does not provide enough data to predict a high success rate.
- FT Booster Version has the highest success rate.
- All launch sites are close in proximity to the equator
- All the launch sites were also built away from populated areas and close to coastlines to ensure the safety of major population centers.

Predictive analysis results

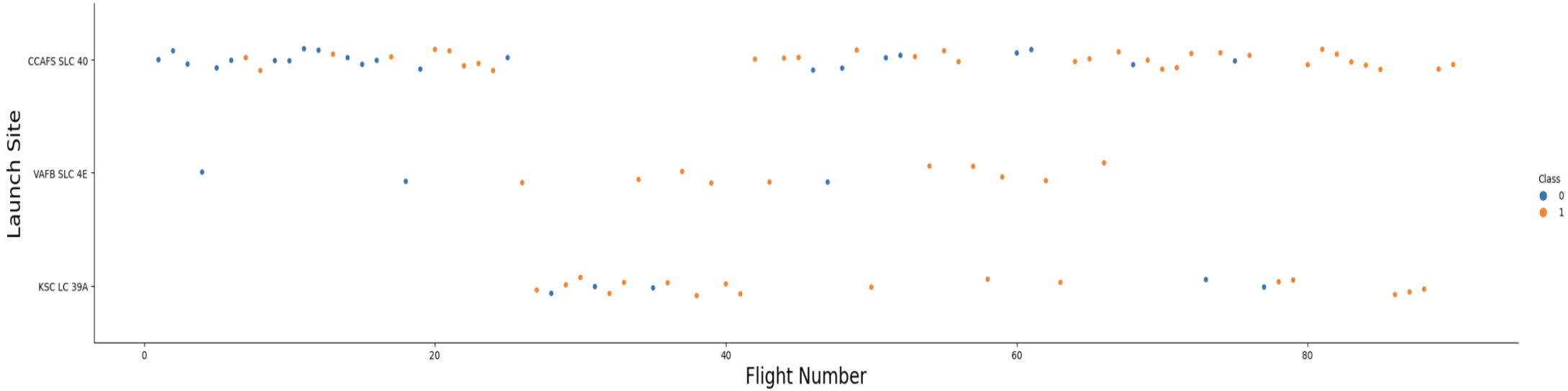
- Confusion Matrix of Decision Tree Classifier is the method that performed best. It proves its accuracy by showing the big numbers of True Positives (TP) and True Negatives (TN).



Section 2

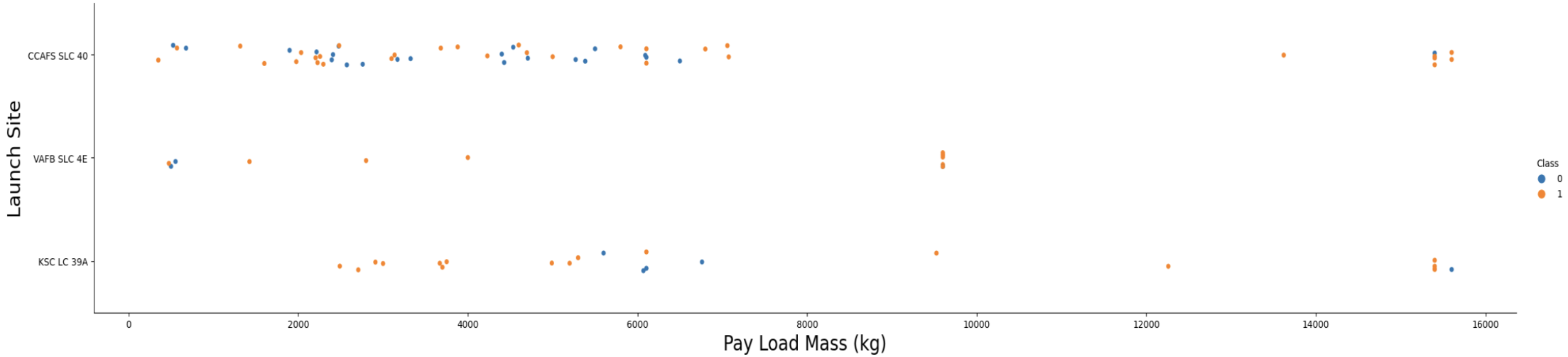
Insights Drawn from EDA

Flight Number vs. Launch Site



- The continuous launch attempts made possible for the first stage more likely to land successfully.
- The success rate of the Launch Sites also increases as the flight number increases over time.

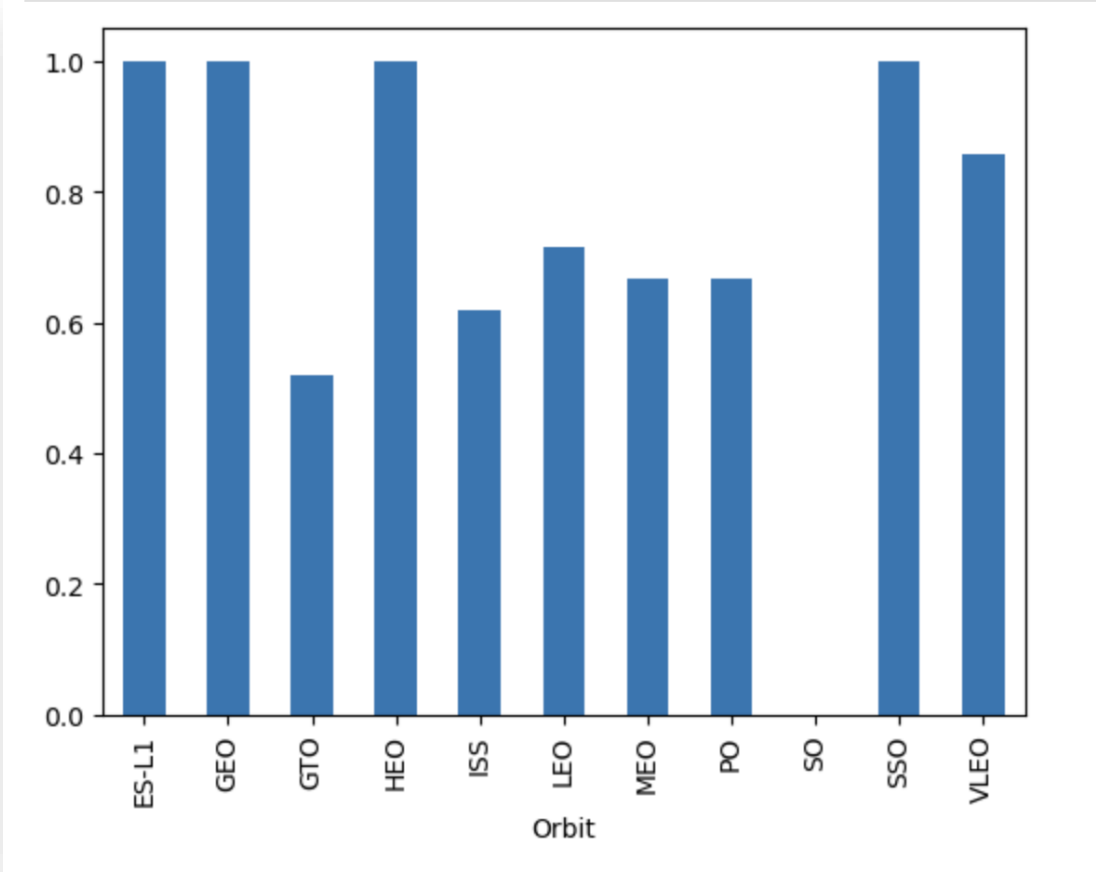
Payload vs. Launch Site



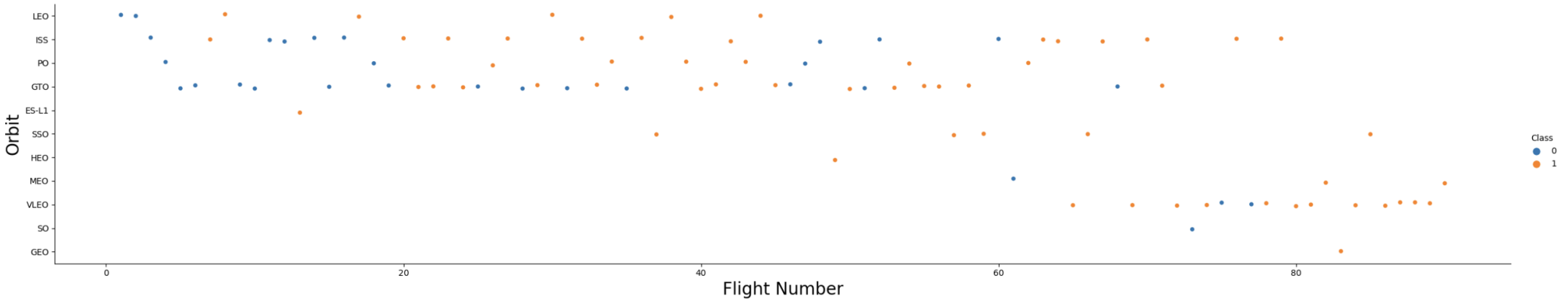
- The rockets in launch sites use Pay Load Mass (kg) not greater than 8000. There are few launches using heavy Pay Load Mass (kg).
- VAFB-SLC launch site did not launch rockets for payload mass greater than 10000.
- There is not enough data to predict the outcome of the launch for the launch sites with massive payload though results show high success rate.

Success Rate vs. Orbit Type

- Orbit types ES-L1, GEO, HEO, and SSO have the highest success rates.
- GTO orbit has the lowest success rate.
- SO orbit has not enough data to show success rate.

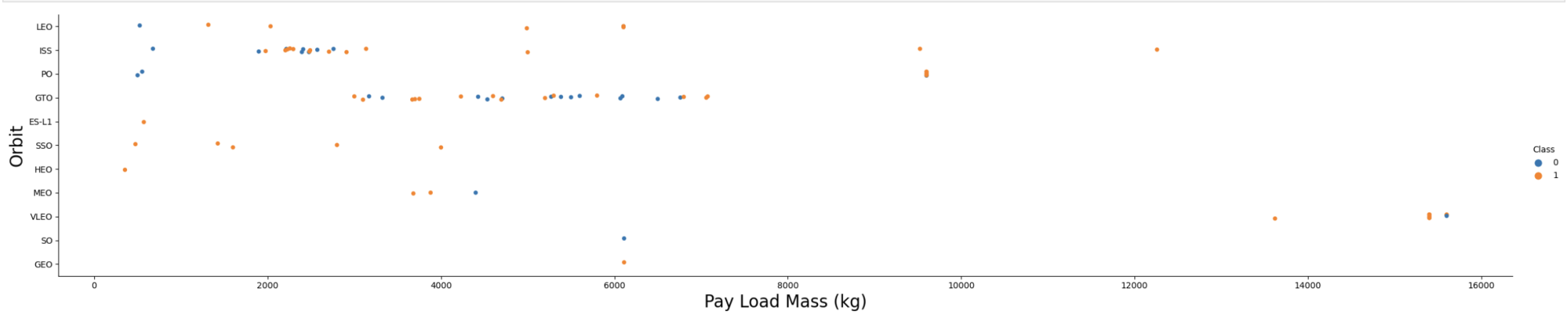


Flight Number vs. Orbit Type



- Orbit type LEO was used during the earlier launch attempts.
- Orbit types ISS, PO, GTO were used more extensively on the continuous launch attempts.
- Later, continuous launch attempts were done on the VLEO orbit with good success rate.

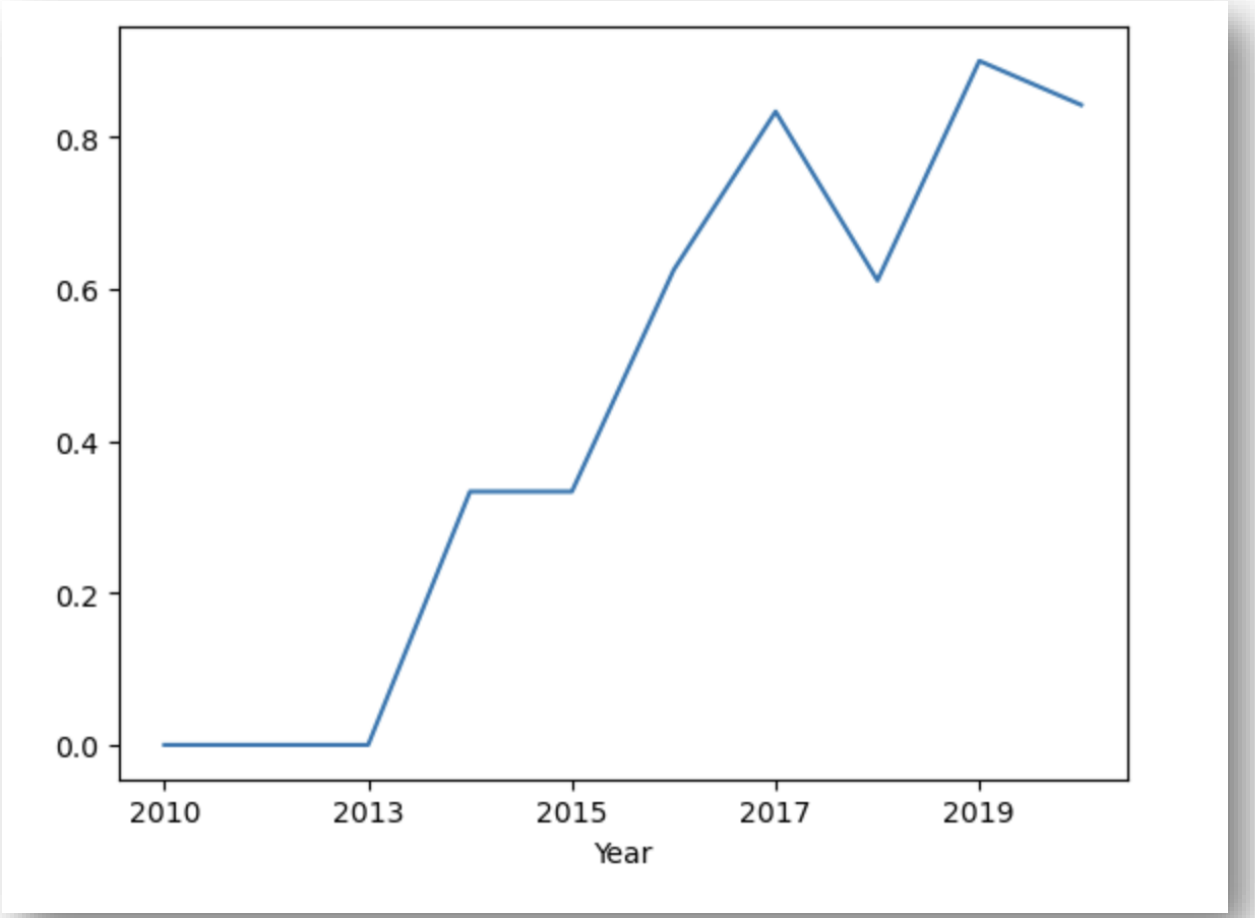
Payload vs. Orbit Type



- All orbit types except VLEO, have success rate with payload mass not greater than 10000 kg.
- Orbits ISS and GTO both do not show distinguished success rate.

Launch Success Yearly Trend

- There is an upward trend in the average launch success rate.
- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

- The SELECT DISTINCT statement is used to display the names of the unique launch sites in the space mission.

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster Version	Launch Site	Payload	PAYLOAD MASS (KG)	Orbit	Customer	Mission Outcome	Landing Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The LIKE operator is used in a WHERE clause to search for launch sites that begin with the string 'CCA'.

Total Payload Mass

- The SUM() function returns the total sum of payload mass carried by boosters launched by NASA (CRS)

Total Payload Mass (kg)
111268

Average Payload Mass by F9 v1.1

- The AVG() function returns the average value of payload mass carried by booster version F9 v1.1.

Average Payload Mass (kg)
2928.4

First Successful Ground Landing Date

- The MIN() function returns the smallest value of the date when the first successful landing outcome in ground pad was achieved, December 12, 2015.

Success (Ground Pad)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The SELECT DISTINCT statement is used to return only the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 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

Mission Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The COUNT() function returns the number of rows that matches the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

- The MAX() function is used in the subquery to return the largest value of the names of the booster versions which have carried the maximum payload mass.

Booster Version	
F9 B5 B1048.4	F9 B5 B1049.5
F9 B5 B1049.4	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1058.3
F9 B5 B1056.4	F9 B5 B1051.6
F9 B5 B1048.5	F9 B5 B1060.3
F9 B5 B1051.4	F9 B5 B1049.7

2015 Launch Records

Booster Version	Launch Site	Landing Outcome	Month
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	10
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	04

- The **SUBSTR()** function is used to extract the months in the year 2015 to list the records which will display the booster versions and launch site with failure landing outcomes in drone ship.

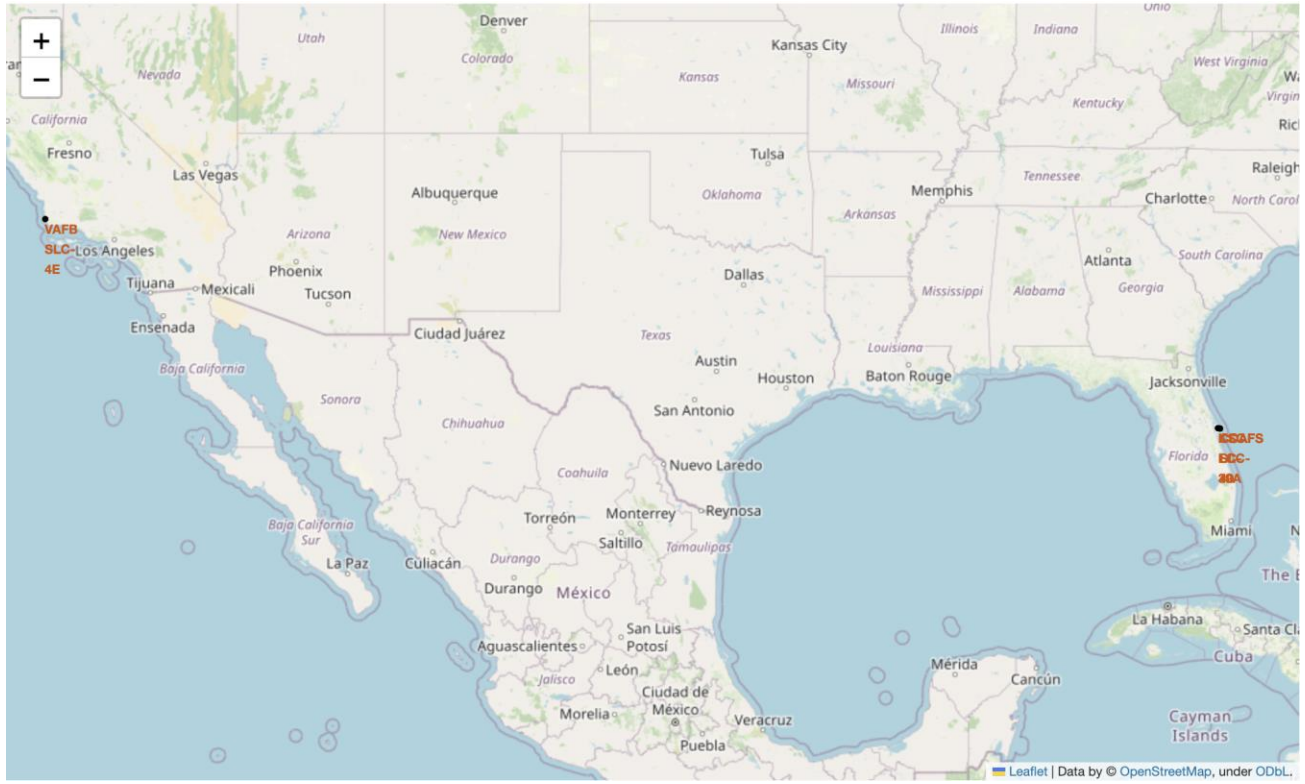
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The GROUP BY statement is used with aggregate function COUNT() to rank the landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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

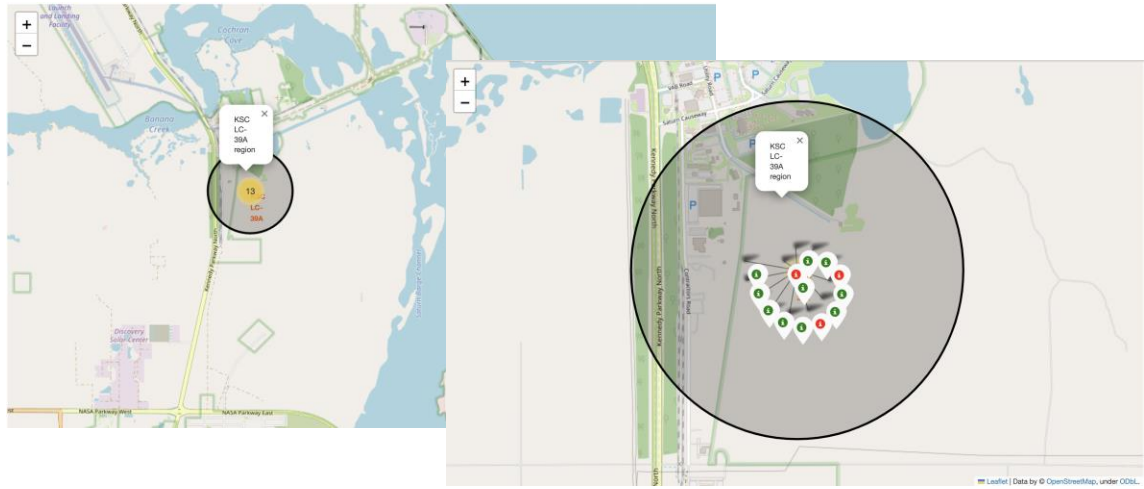
Launch Sites Proximities Analysis

All Marked Launch Sites on the Map



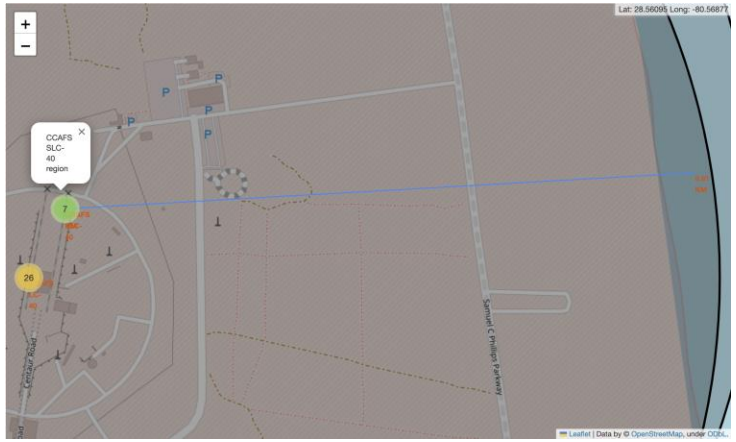
- All launch sites are in proximity to the Equator line which makes them more efficient.
- All launch sites are in very close proximity to the coast which would minimize the risk of debris dropping to densely populated areas.

Launch Outcome for Launch Site on the Map

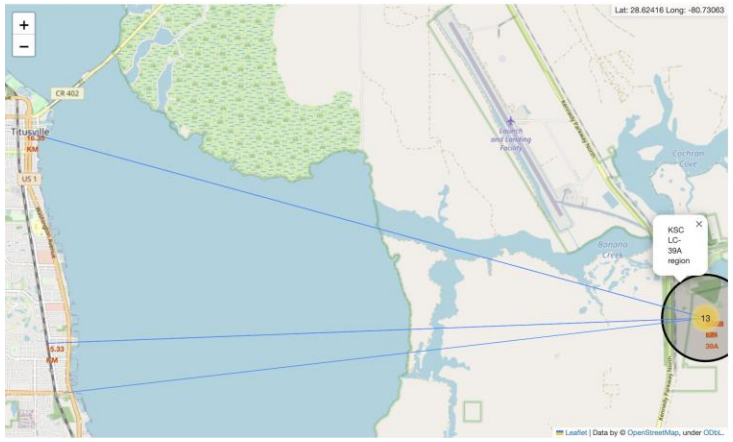


- The map is enhanced by adding the launch outcomes for each site, and see which sites have high success rates.
- Markers are created for all launch records. If a launch was successful, a green marker is used and a red marker for a failed launch.
- From the color-labeled markers in marker clusters, launch sites with relatively high success rates would easily be identified.

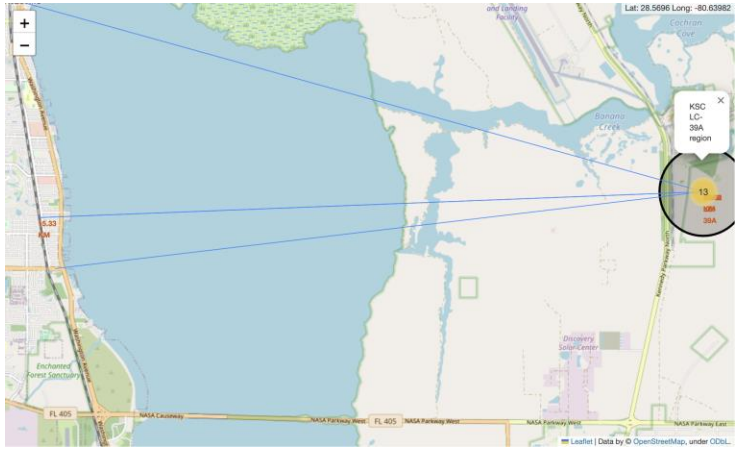
Proximities of Launch Sites



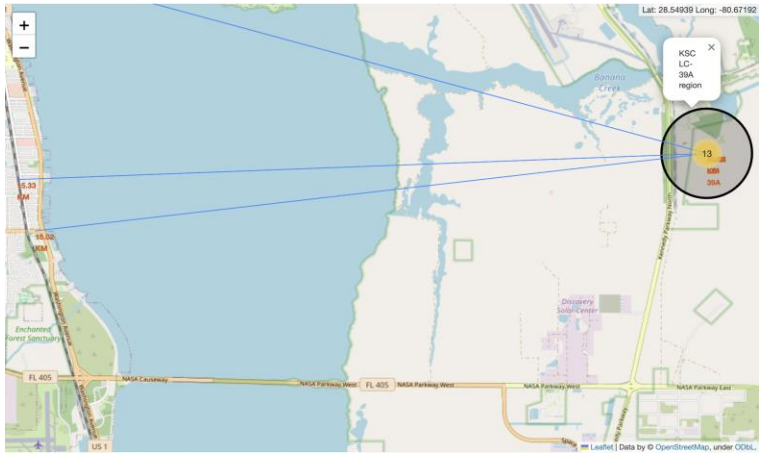
CCAFS SLC-40 to closest coastline at .91 KM



KSC LC-39A to closest city (Titusville) at 16.35 KM

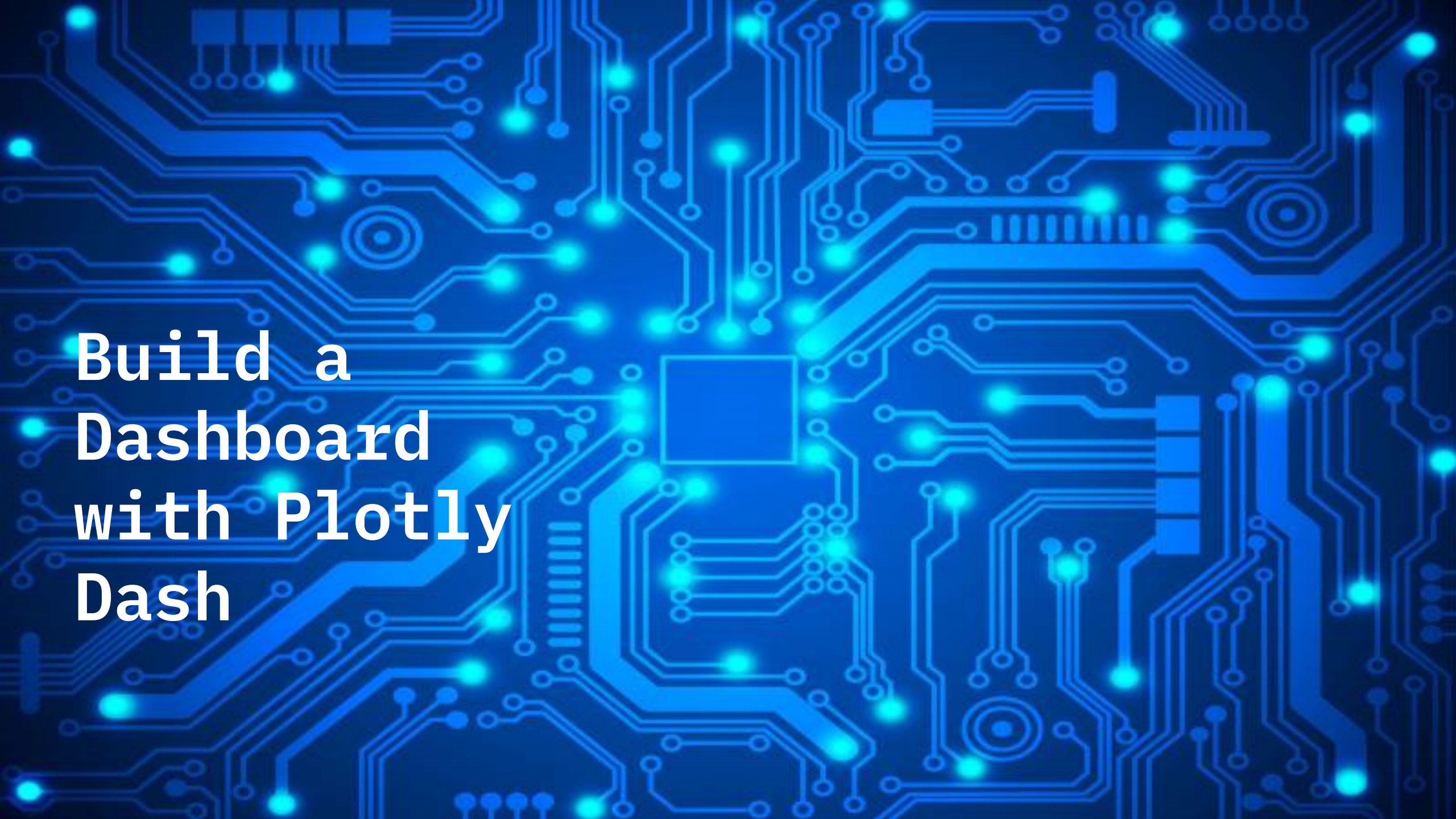


KSC LC-39A to closest railway (Florida Eastcoast Railway) at 15.33KM

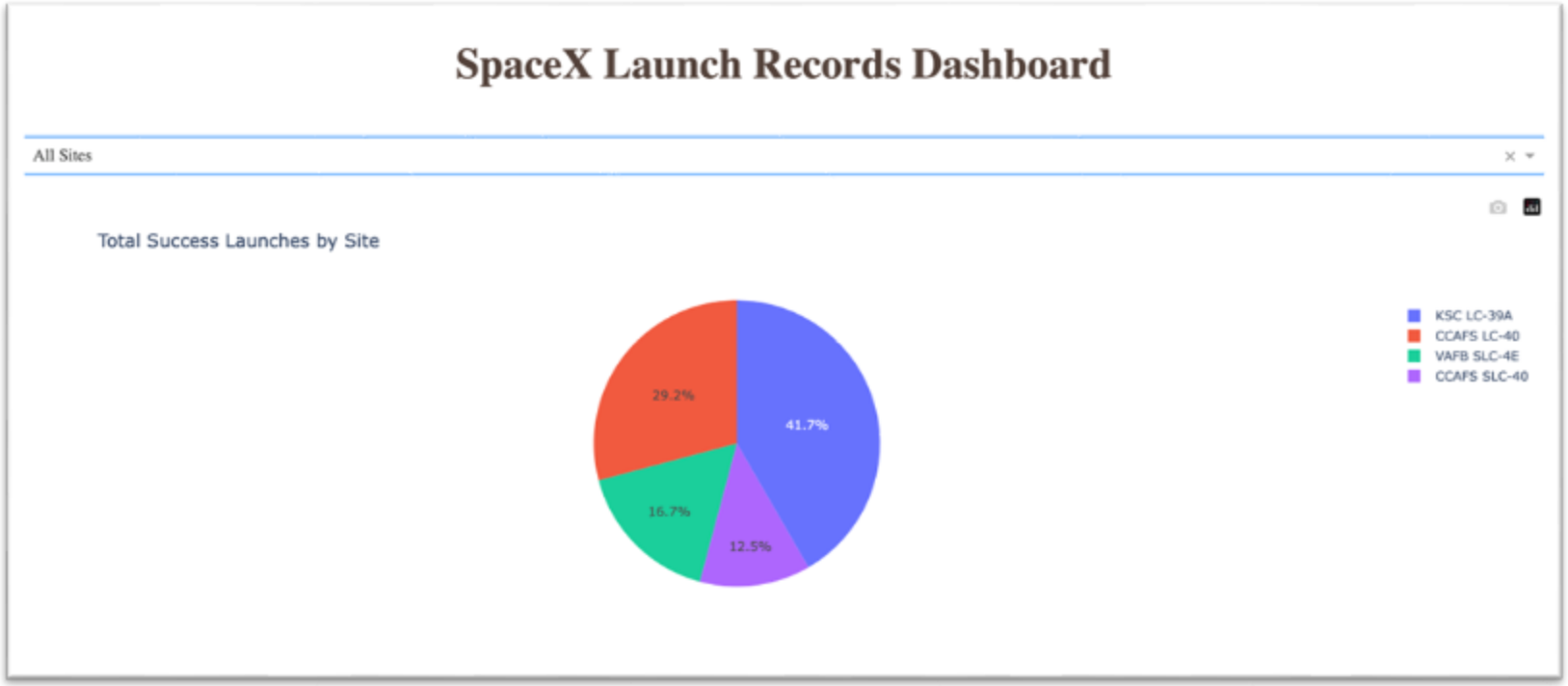


KSC LC-39A to closest highway (Washington Avenue) at 15.02 KM

- The Launch Sites are close to coastlines with CCAFS SLC-40 having a distance of .91 km from the coast.
- The Launch Sites are far enough from railways, highways and cities having an average distance of 15 km away.
- The Launch Sites are in close proximity to the coast and far from densely populated areas which makes it safe in case something goes wrong.

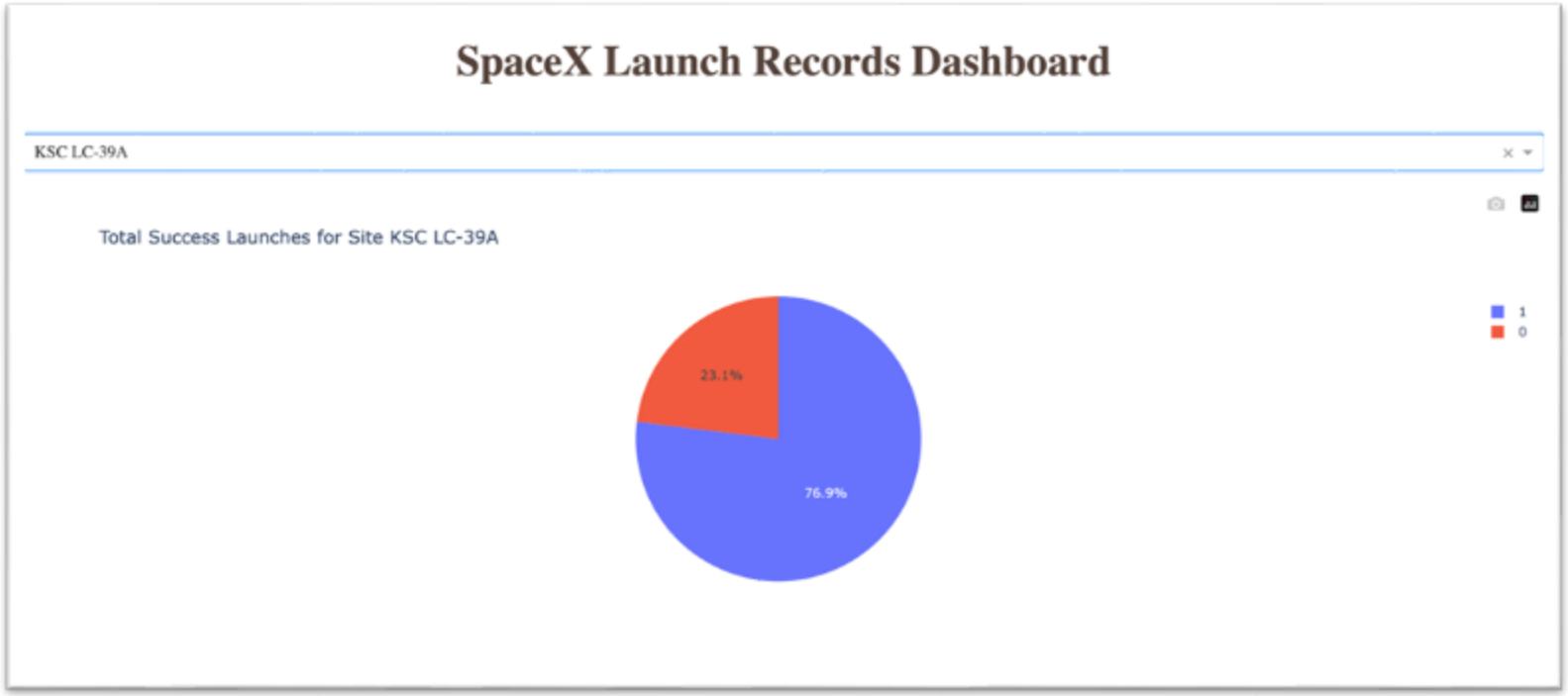

Build a
Dashboard
with Plotly
Dash

Total Success Launches by All Sites



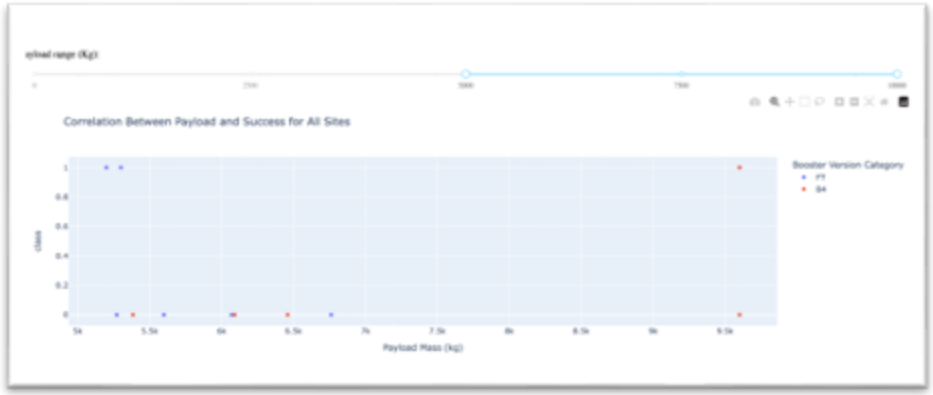
- The chart shows the percentage distribution of the total success launches by all sites.
- KSC LC-39A represented by the blue slice has the highest total success.

Launch Site with Highest Success Rate

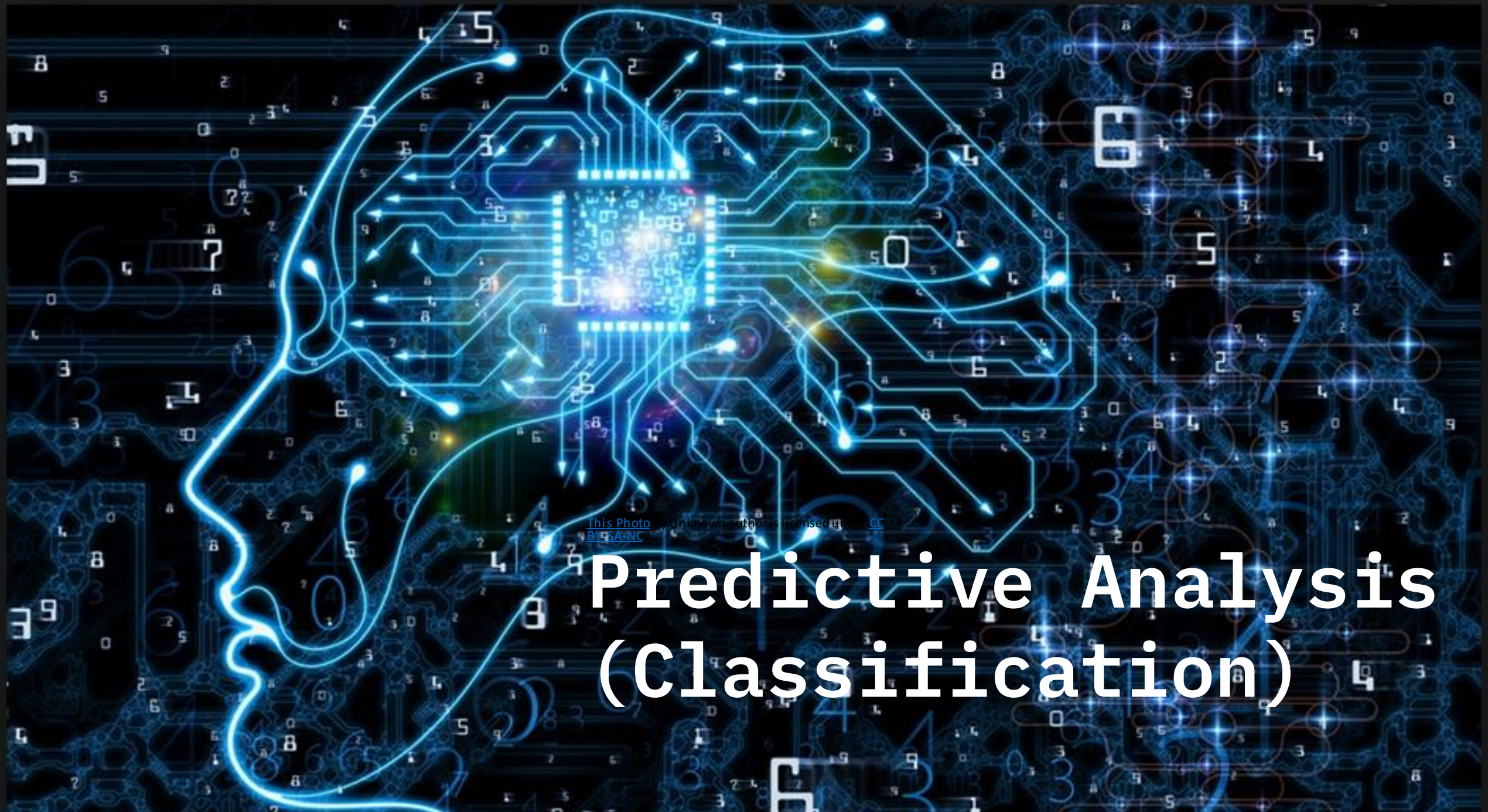


- Launch Site KSC LC-39A is the launch site with the highest success rate.
- Launch Site KSC LC-39A has a 76.9% success rate from all launches.

Payload Range with Highest Success Rate



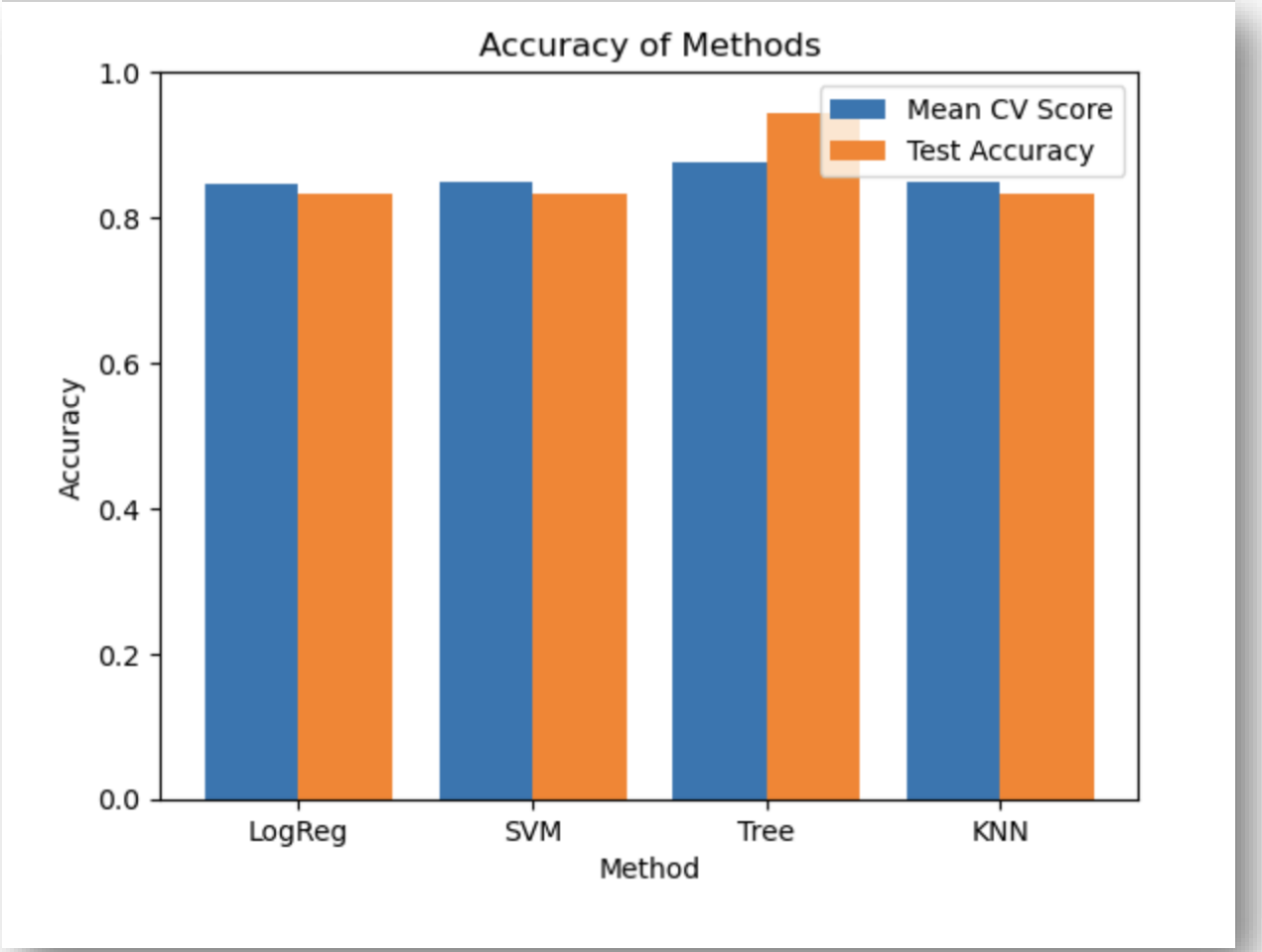
- Payload range from 3000 kg to 6000 kg has the highest success rate.
- FT Booster Version has the highest success rate.
- Payload range higher than 6000 kg does not provide enough data to predict a high success rate.



Predictive Analysis (Classification)

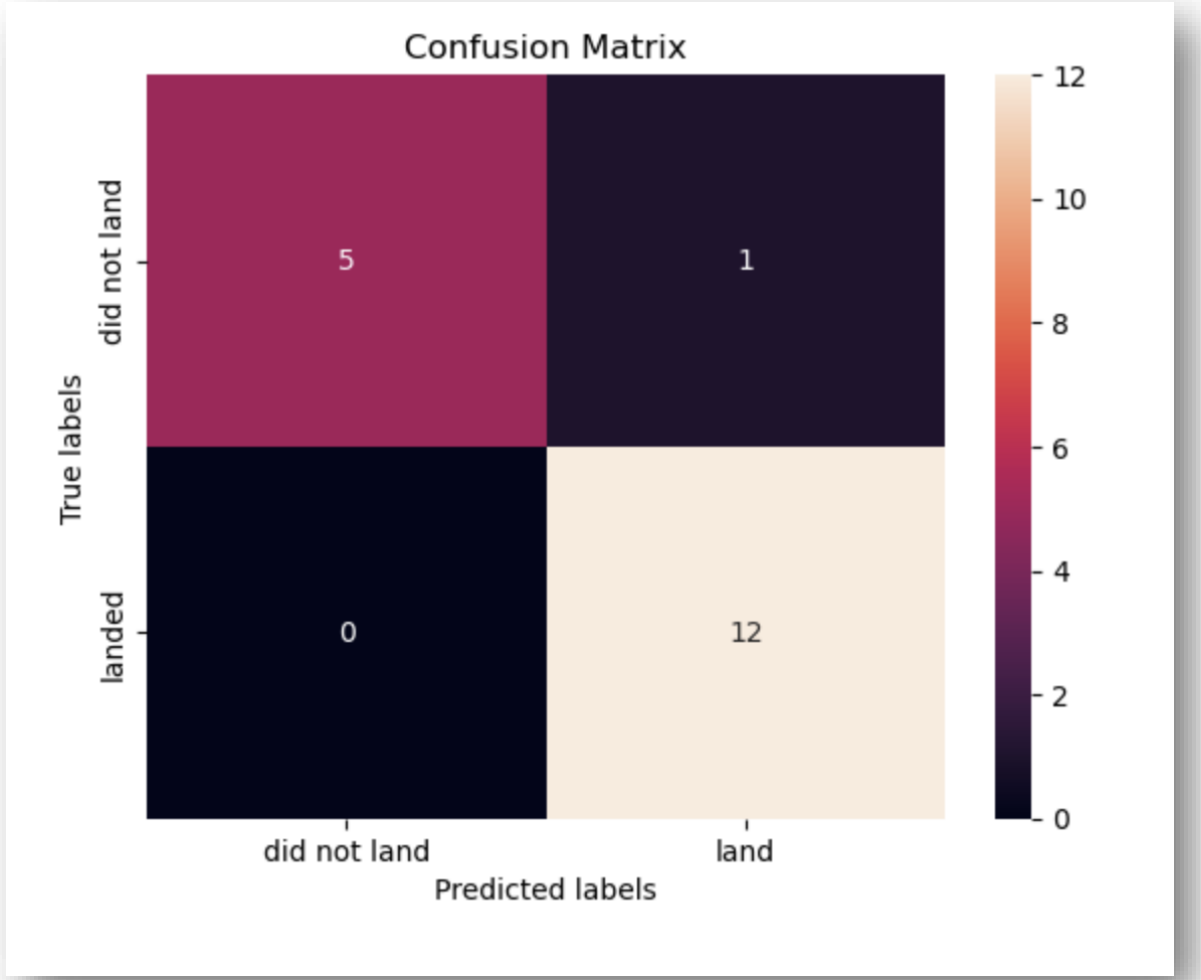
Classification Accuracy

- The bar chart shows the accuracy of the models with the Mean CV Score greater than 0.80 for all models with the Decision Tree model having the highest score.
- The Decision Tree model has the highest classification accuracy of 0.94.



Confusion Matrix

- Confusion Matrix of Decision Tree Classifier is the method that performed best. It proves its accuracy by showing the big numbers of True Positives (TP) and True Negatives (TN).



Conclusions

- The models performed with similar results on the test set with the decision tree model slightly outperforming the other models.
- Most of the launch sites are near the equator. Thus, the rockets get added advantage of earth's rotational speed which helps save the cost of putting in extra fuel and boosters.
- All the launch sites are close to the coast and away from populated areas to ensure that no components are shed over these areas in case something went wrong.
- Launch success increases over time.
- KSC LC-39A has the highest success rate among launch sites.
- Orbit types ES-L1, GEO, HEO, and SSO have a100% success rate.
- The rockets in launch sites use Pay Load Mass (kg) not greater than 8000. There are few launches using heavy Pay Load Mass (kg).

To summarize, SpaceY company should take into account the above points in the development of orbital launch systems that can be reused many times. Analysis showed that over the decade, SpaceX was able to facilitate successful landing outcomes for rapid reuse of launch vehicles. Since SpaceX first achieved a successful landing and recovery of a first stage in December 2015, reflights of refurbished first stages then became routine. Thus, SpaceX significantly reduced the cost of access to space, and change the increasingly competitive market in space launch services. Since SpaceX has developed technologies, the company is now developing the Starship system with a fully-reusable two-stage launch vehicle. The project intends to replace all of its other launch vehicles and spacecraft for satellite delivery and human transport.

Appendix

- Interactive Visual Analytics with Folium.ipynb
- File does not render on GitHub, copy the URL below to view file on nbviewer

  - https://nbviewer.org/github/Ilyn0402/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

  - nbviewer is a web application that lets you enter the URL of a Jupyter Notebook file on GitHub, then it renders that notebook as a static HTML web page
  - Go to https://nbviewer.jupyter.org/

- Reference reading consulted during this project

  - https://en.wikipedia.org/wiki/SpaceX_reusable_launch_system_development_program


Thank You!