IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bulent Tiryaki
26 October 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Space X Falcon 9 First Stage Landing Prediction - Data Collection using SpaceX API
  - Space X Falcon 9 First Stage Landing Prediction - Data Collection with Web Scraping from Wikipedia
  - Space X Falcon 9 First Stage Landing Prediction - Data Wrangling
  - Space X Falcon 9 First Stage Landing Prediction - Exploratory Data Analysis using SQL
  - Space X Falcon 9 First Stage Landing Prediction - Exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
  - Space X Falcon 9 First Stage Landing Prediction - Launch Sites Locations Analysis with Folium
  - Space X Falcon 9 First Stage Landing Prediction - Building a Dashboard Application with Plotly Dash
  - Space X Falcon 9 First Stage Landing Prediction - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis (EDA) Results
  - Interactive Visual Analytics and Dashboards
  - Predictive Analysis (Classification)

Introduction

- Project background and context
  - SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010.
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
  - In this project, the Data Science methodology, including data collection, data wrangling, exploratory data analysis, data visualization, model development, model evaluation, and reporting the results to stakeholders, was fully implemented to predict if the Falcon 9 first stage will land successfully.
  - The ultimate purpose was to help startups intending to compete with SpaceX make more informed bids against SpaceX for a rocket launch.
- Problems to find answers for
  - Finding out how the factors such as payload mass, orbit type, the location and proximities of a launch site affect the launch success rate.
  - Establishing predictive models using various machine learning algorithms using most influential factors
  - Finding out the predictive model that performs best.

Section 1

Methodology

Methodology

Executive Summary

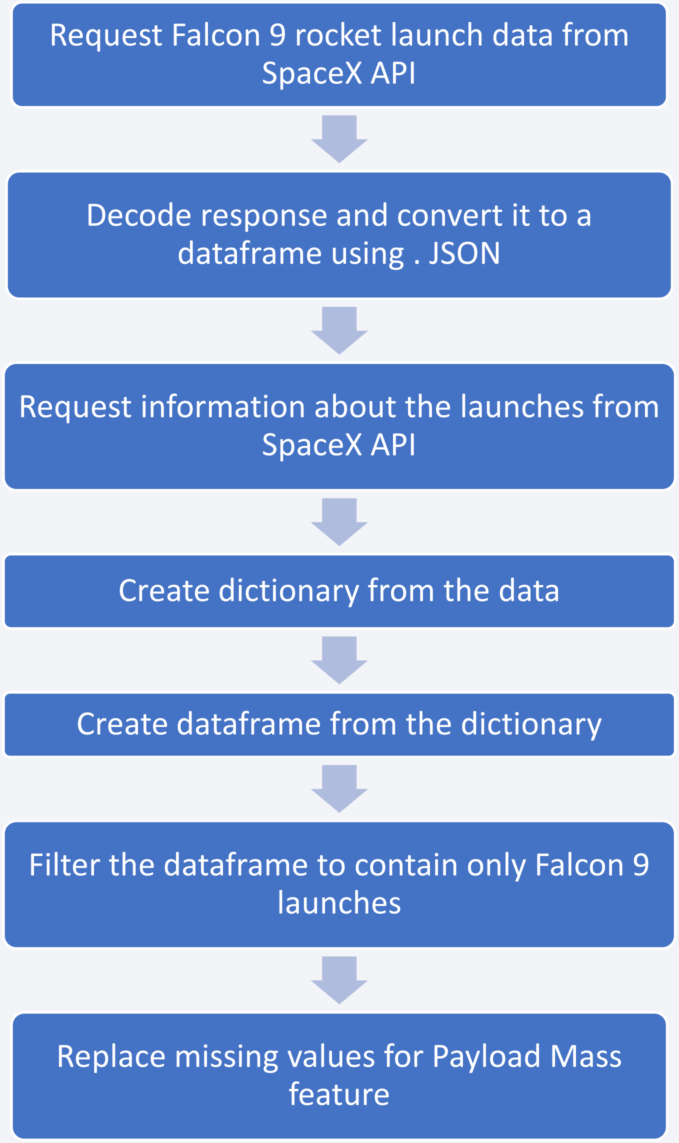
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was first collected using SpaceX API calls by making a get request to the SpaceX API. This was done by first defining a series helper functions that would help in the use of the API to extract information using identification numbers in the launch data and then requesting rocket launch data from the SpaceX API URL.
- Finally to make the requested JSON results more consistent, the SpaceX launch data was requested and parsed using the GET request and then decoded the response content as a JSON result which was then converted into a Pandas data frame that was filtered to contain only Falcon 9 launches.
- Web scraping was also implemented to collect Falcon 9 launch records from a Wikipedia page.
- Using BeautifulSoup and requests libraries, first, Space X Falcon 9 launch HTML table was extracted records from the Wikipedia page.
- Table was then parsed and converted it into a Pandas data frame.

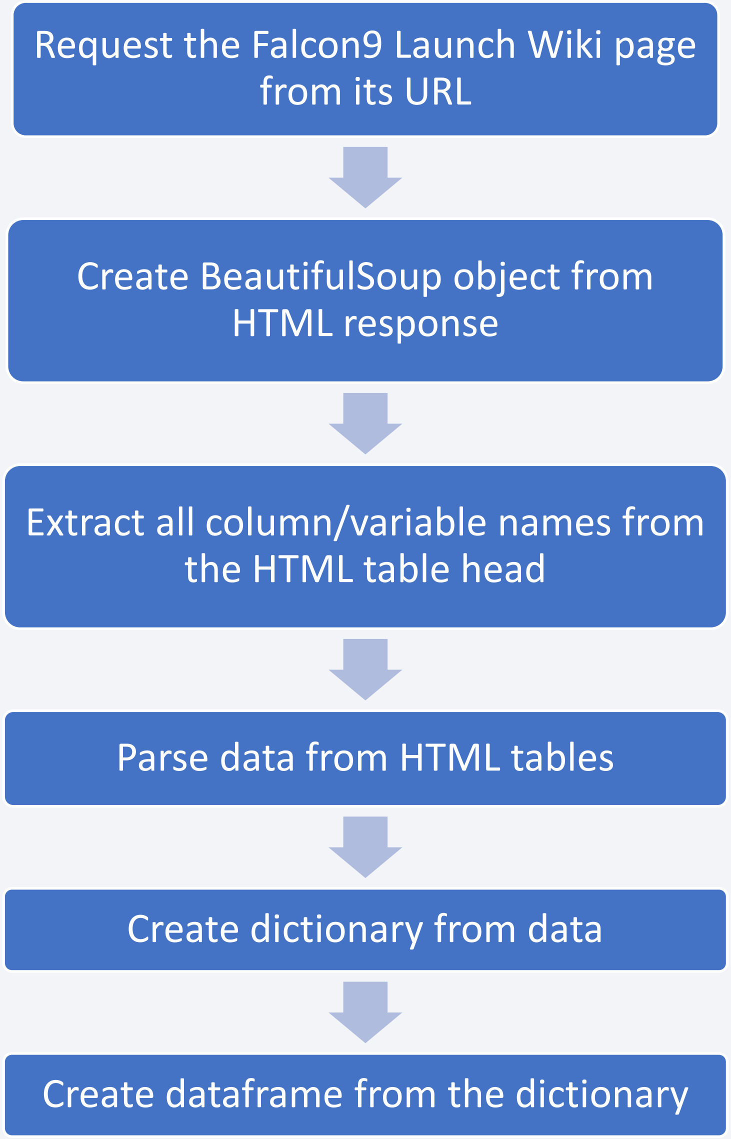
Data Collection – SpaceX API

- Data collected using SpaceX API calls by making a get request to the SpaceX API.
- SpaceX launch data was then requested and parsed using the GET request.
- Response content was decoded as a JSON result that was then converted into a Pandas data frame.
- Dataframe was filtered to get Falcon 9 launches and missing values for Payload Mass attribute was replaced by the mean value.
- Resulting dataframe was exported into a .CSV file.
- Flowchart of SpaceX API calls is given on the right.
- Click here to go to the GitHub URL of the completed SpaceX API calls notebook.



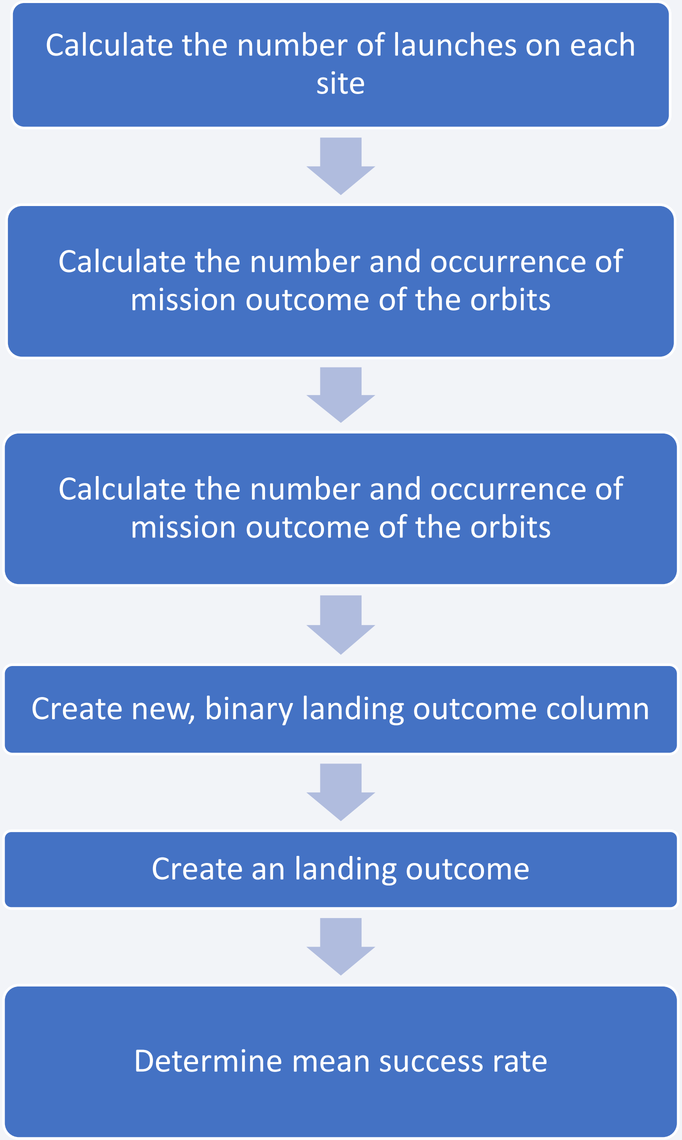
Data Collection – Web Scraping

- Falcon 9 historical launch records were collected from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches via web scraping
- requests and BeautifulSoup libraries were used for web scraping.
- Falcon 9 launch records data was first extracted from HTML table of the Wikipedia page,
- A data frame was then created from HTML table by parsing.
- Resulting dataframe was exported into a .CSV file.
- Flowchart of web scraping process is given on the right.
- Click [here]() to go to the completed web scraping notebook on GitHub .



Data Wrangling

- Exploratory Data Analysis was performed to find some patterns in the data and determine what would be the label for training supervised models.
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.
- Those outcomes were converted into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful as a result of Data Wrangling efforts.
- Resulting dataframe was exported into a .CSV file.
- Flowchart of data wrangling process is given on the right.
- Click here to go to the completed data wrangling related notebook on GitHub.



EDA with Data Visualization

- Exploratory Data Analysis and Feature Engineering were performed using Pandas and Matplotlib in order to
  - Visualize the relationship between Flight Number and Launch Site in a scatter plot
  - Visualize the relationship between Payload Mass and Launch Site in a scatter plot
  - Visualize the relationship between Success rate of each Orbit type in a bar plot
  - Visualize the relationship between Flight Number and Orbit type in a scatter plot
  - Visualize the relationship between Payload Mass and Orbit type in a scatter plot
  - Visualize the launch success yearly trend in a line plot
- Also, dummy variables were created to categorical columns and all numeric columns were cast to float64.
- Resulting dataframe was exported into a .CSV file.

EDA with Data Visualization (continued)

- Scatter plots were used because a scatter plot visualizes the relationship between two continuous variables, displaying individual data points as dots on a two-dimensional plane, allowing for the examination of patterns, clusters, and correlations.
- A bar plot represents categorical data with rectangular bars, A bar plot is also an effective way of representing data where the height of each bar corresponds to the value of a specific category, making it suitable for comparing values across different categories.
- A line plot displays the relationship between two continuous variables over a continuous interval, showing the trend or pattern of the data. It is a basic type of chart common in many fields. They are best suited for trend-based visualizations of data over a period of time.
- Click here to go to the completed EDA with data visualization notebook on GitHub.

EDA with SQL

- SQL queries were performed after establishing a connection with the database and then downloading it to gather information about the dataset.
  - 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 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
- Click [here]() to go to the completed EDA with SQL notebook on GitHub.

Build an Interactive Map with Folium

- The launch success rate may depend on many factors such as payload mass, orbit type along with others.
- It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- Finding an optimal location for building a launch site involves many factors and could help discover some of the factors by analysing the existing launch site locations.
- In the previous exploratory data analysis sections, SpaceX launch dataset has been visualised using matplotlib and seaborn and some preliminary correlations have been discovered between the launch site and success rates.
- In this section, more interactive visual analytics was performed using Folium to further investigate on the relationship between the launch site and success rates.
- Results are summarized in next two slides.
- Click here to go to the completed interactive map with Folium map on GitHub.

Build an Interactive Map with Folium (continued)

- All launch sites were marked on a map and each site's location was added on a map using site's latitude and longitude coordinates. A folium Map object was created, with an initial center location to be NASA Johnson Space Center at Houston, Texas. A blue circle was created at NASA Johnson Space Center's coordinate with a popup label showing its name.
- A circle and a marker for each launch site were created and added on the site map to be able to explore the map by zoom-in/out the marked areas and see if all launch sites are in proximity to the Equator line all launch sites are in very close proximity to the coast.
- For each launch site, a Circle object was added based on its coordinate (Lat, Long) values. In addition, Launch site name was added as a popup label.
- Colored markers of successful (green) and unsuccessful (red) launches at each launch site were created to show which launch sites have high success rates. A launch only happens in one of the four launch sites, which means many launch records will have the exact same coordinate. Marker clusters can be a good way to simplify a map containing many markers having the same coordinate.
- A new column in spacex_df dataframe called marker_color was also created to store the marker colors based on the class value.

Build an Interactive Map with Folium (continued)

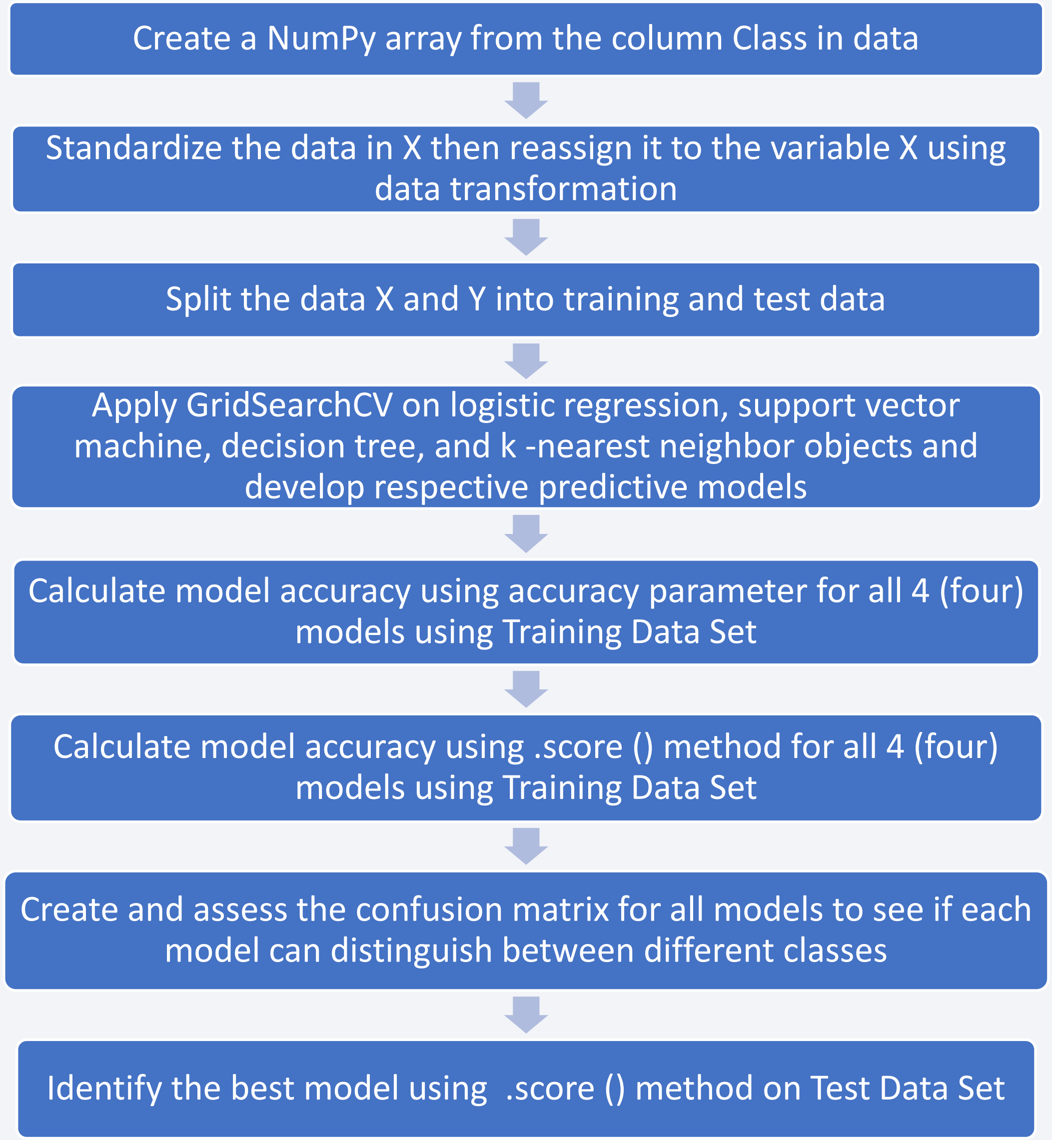
- For each launch result in spacex_df data frame, a folium.Marker was added to marker_Cluster to be able to easily identify which launch sites have relatively high success rates.
- A MousePosition on the map was added to get the coordinates (Latitude and Longitude) for a mouse over a point on the map so that one can easily find the coordinates of any points of interests (such as railway) while you are exploring the map and calculate the distance to the launch site.
- A point on the closest coastline was marked down using MousePosition and calculate the distance between the coastline point and the launch site. For this, a folium.Marker was created and added on the selected closest coastline point on the map and then the distance between coastline point and launch site was displayed using the icon property.
- Blue `folium.PolyLine` objects were created using the selected coordinates and the launch site coordinate to show distance between launch site CCAFS SLC40 and its selected proximities on the map.
- After plotting distance lines to the proximities, it has been found out that launch site CCAFS SLC40 is in close proximity to the coastlines and highways.

Build a Dashboard with Plotly Dash

- An interactive dashboard application titled "SpaceX Launch Records Dashboard" was built to analyse launch records interactively with Plotly Dash by:
    - Adding a Launch site dropdown input component to enable Launch Site selection
    - Adding a callback function to render success-pie-chart based on selected site dropdown to show the total successful launches count for all sites.
    - Adding a range slider to Select Payload to allow user to select payload mass range.
    - Adding a callback function to render the success-payload-scatter-chart scatter plot to allow user to see the relationship between Payload Mass and Launch Success
- A pie chart is a circular graphic that displays numeric proportions by dividing a circle (or pie) into proportional slices. It represents data in a simple and easy-to-understand format, it is easy to draw providing clean and clear data visualization allowing immediate analysis.
- Reasons why a scatter plot has been used are explained in previous Slide #12.
- Click here to go to the source code of the Dashboard on GitHub.

Predictive Analysis (Classification)

- A machine learning pipeline has been created to predict if the first stage of Falcon 9 will land successfully given the data from the preceding data analyses.
- For this purpose, data was split into training and test data sets to find the best hyperparameter for Support Vector Machine, Classification Trees, and Logistic Regression methods, and then the method that performs best using test data was found.
- Flowchart of machine learning process is given on the right.
- Click [here]() to go to the completed predictive analysis notebook on GitHub.



Results

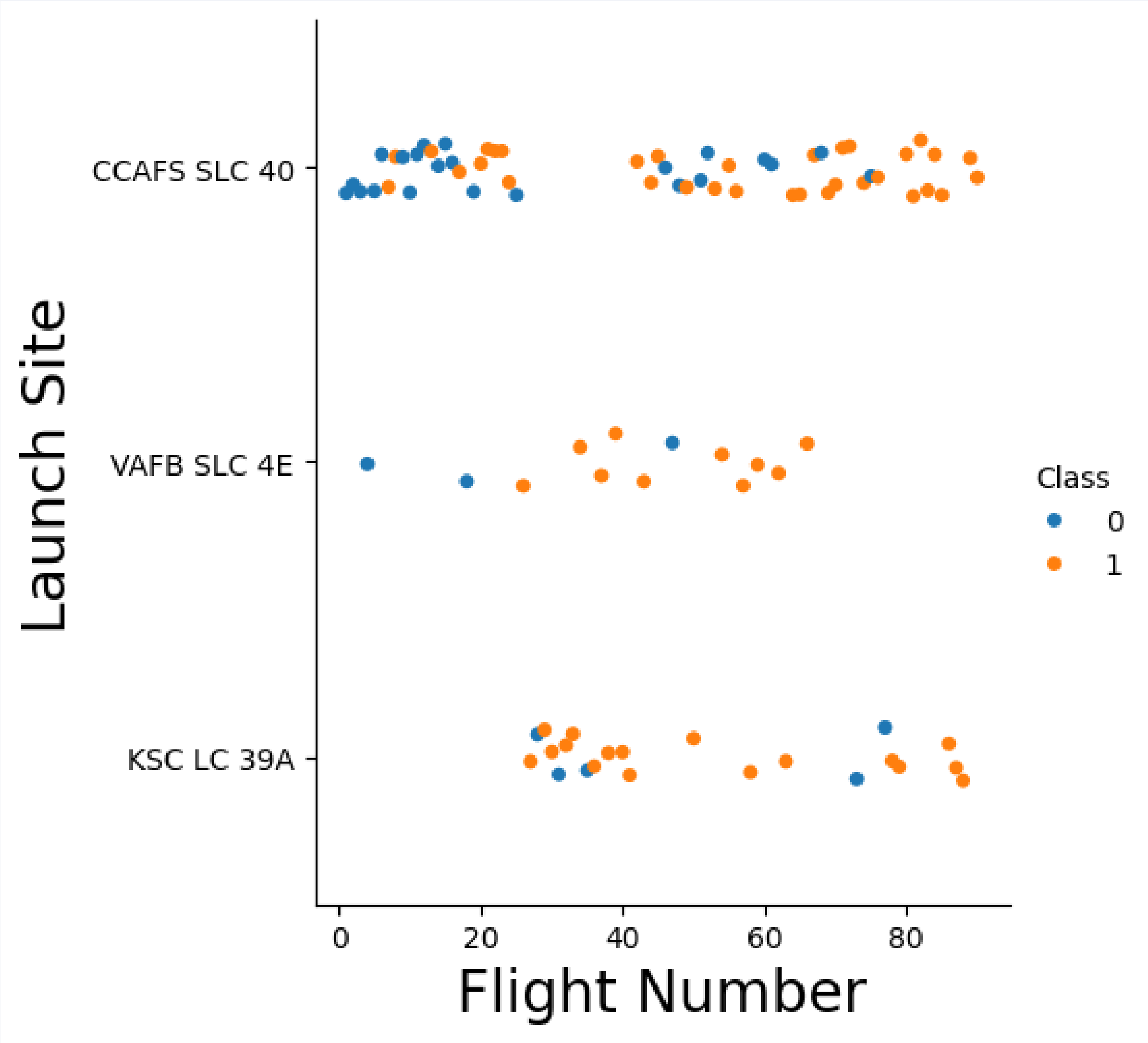
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

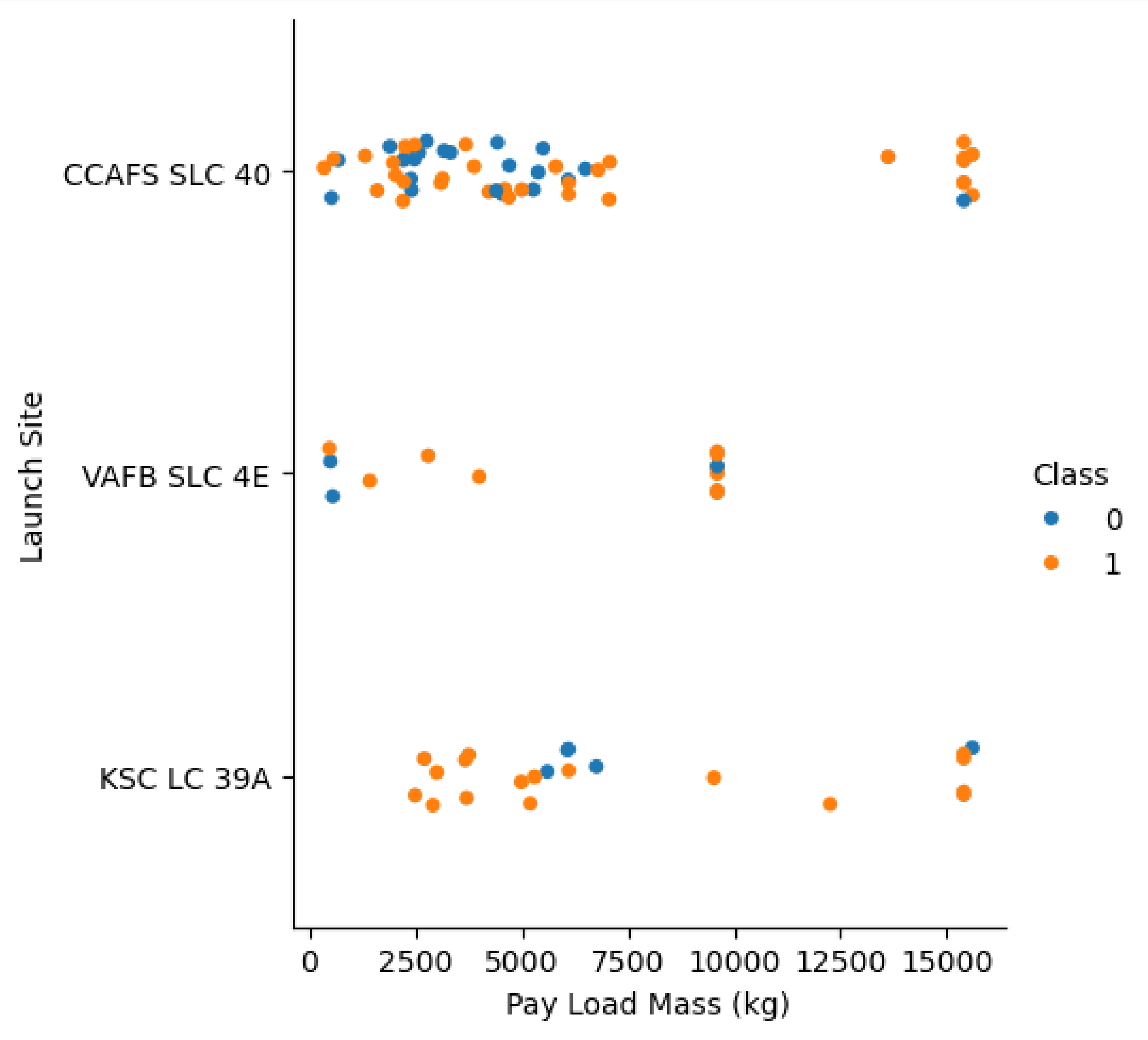
Flight Number vs. Launch Site

- As the Flight Number increases for each of the 3 (three) Launch Sites, Success Rate increases.



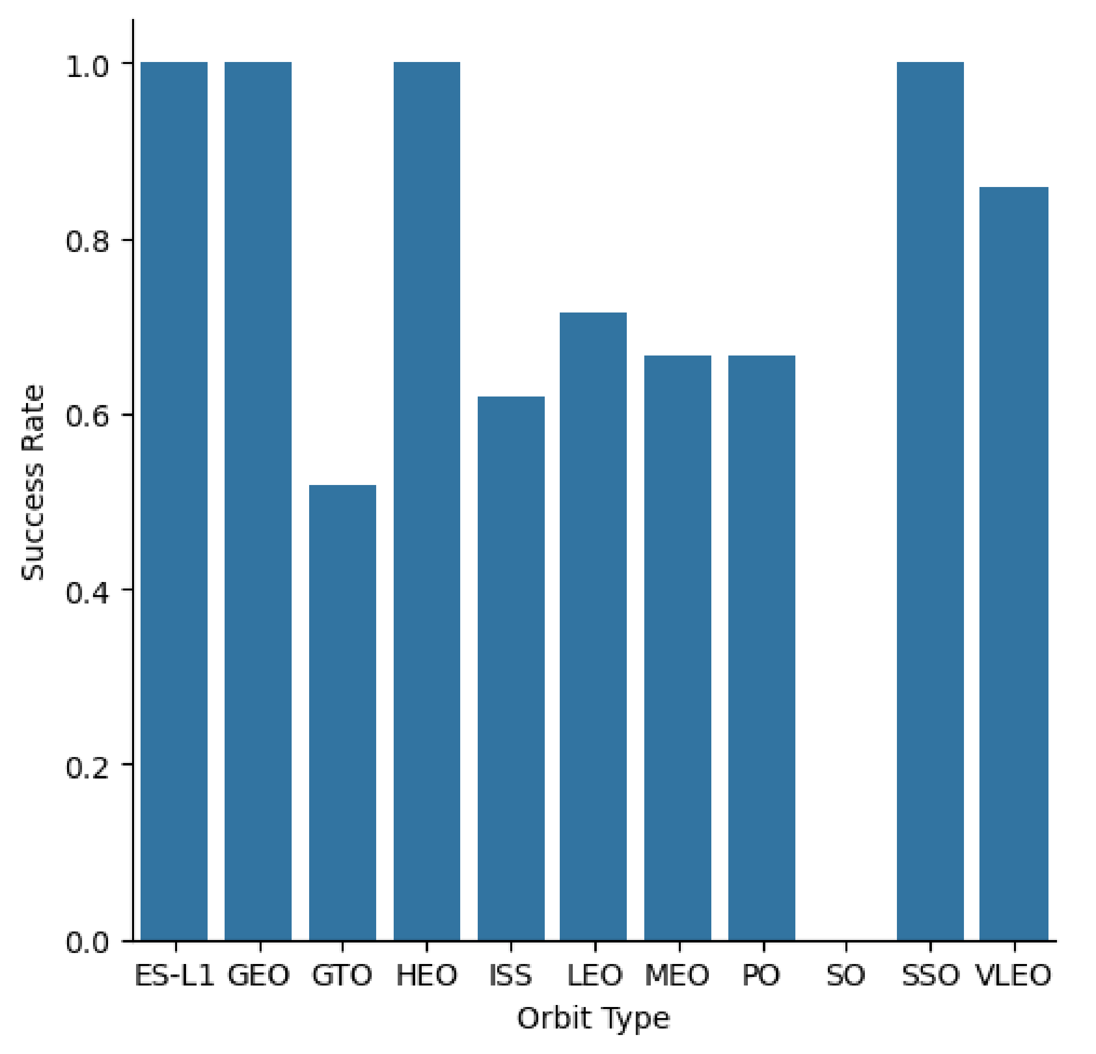
Payload vs. Launch Site

- For the VAFB-SLC Launch Site, there are no rockets launched for heavy payload mass (greater than 10,000 kg).



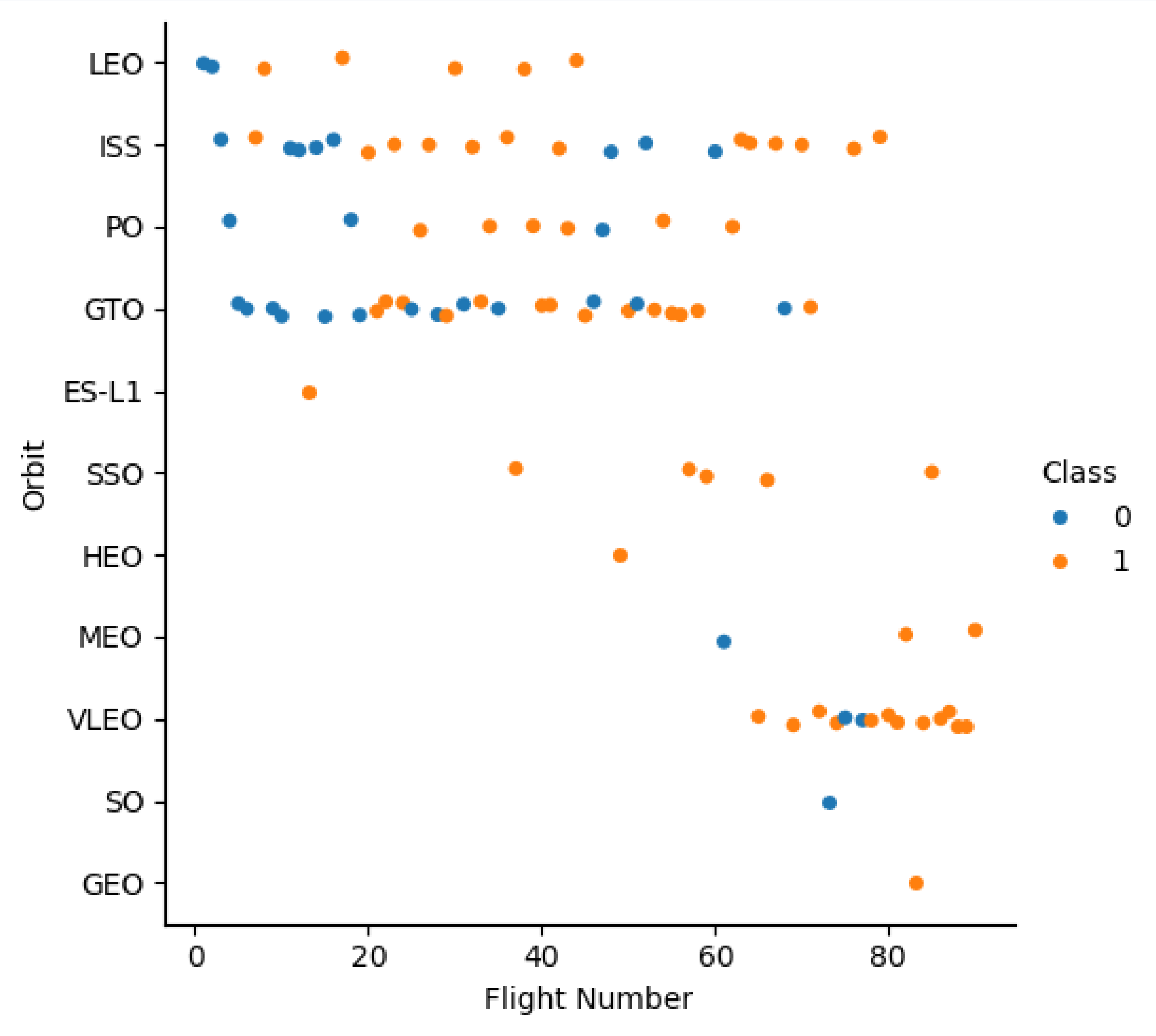
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, and SSO have the highest success with 100% Success Rate.



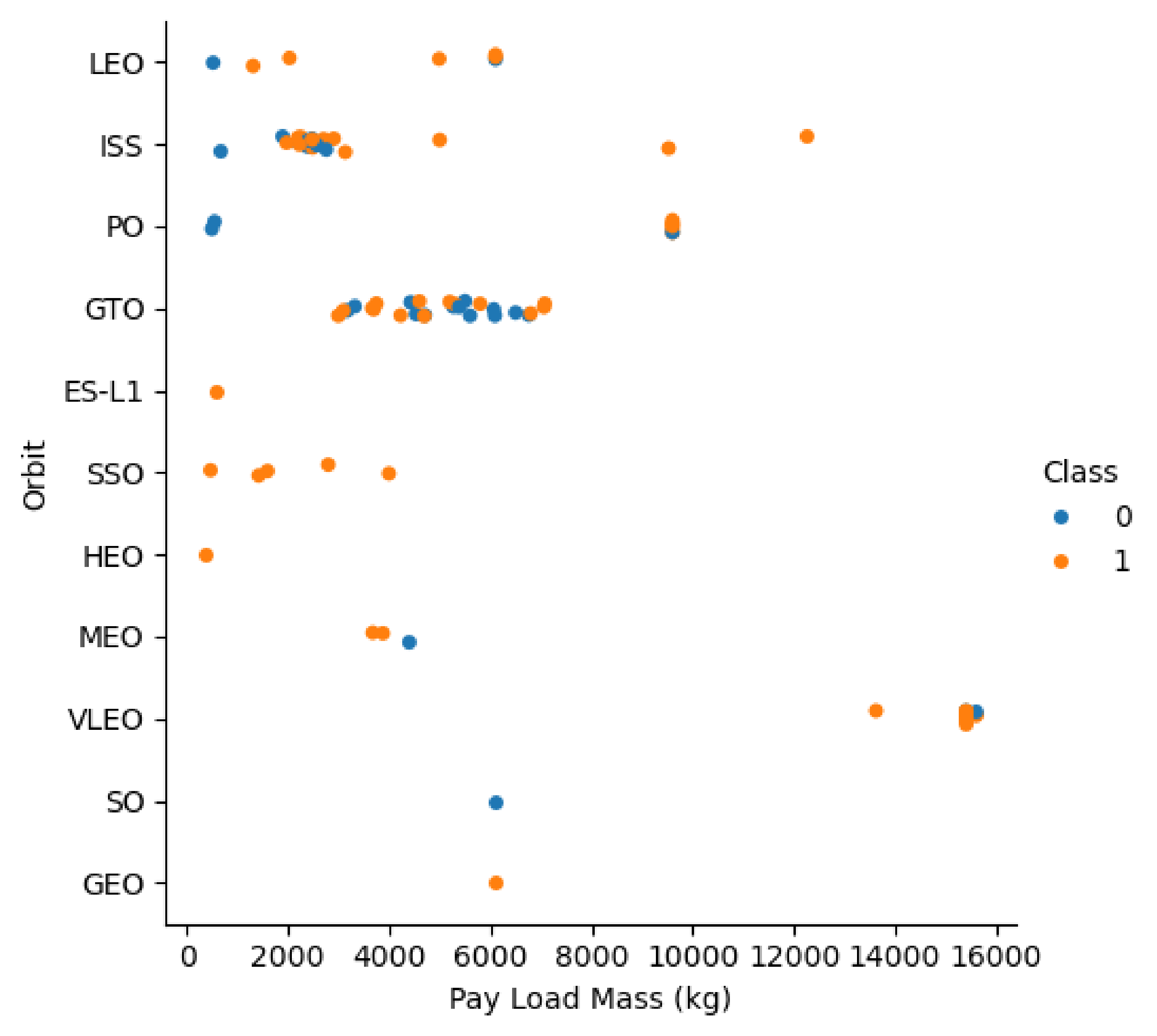
Flight Number vs. Orbit Type

- In the LEO orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



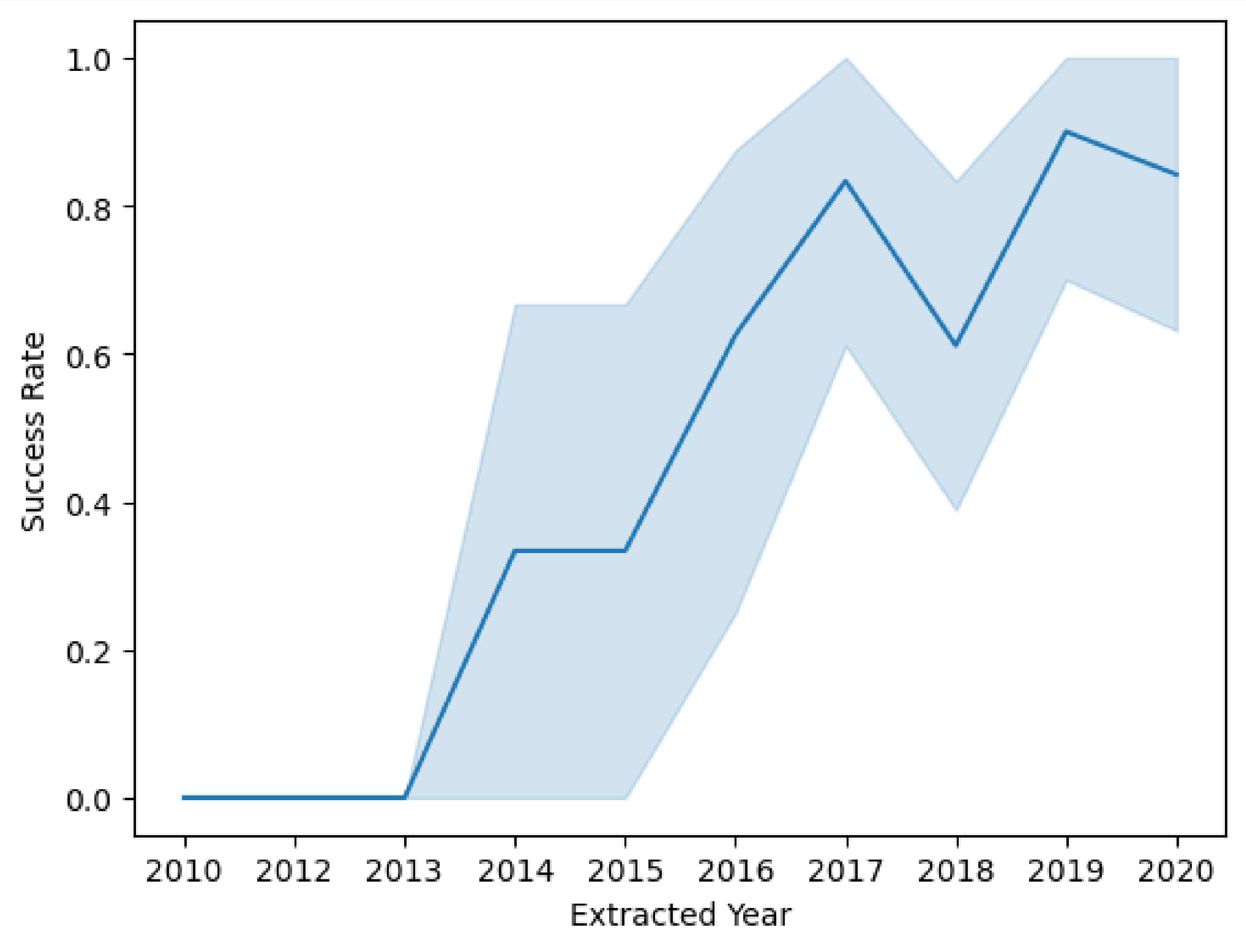
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO, and ISS orbits.
- For GTO orbit, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- The Success Rate since 2013 kept increasing until 2020.



All Launch Site Names

- Displayed the names of the unique launch sites in the space mission.
- Applied SELECT DISTINCT statement to retrieve and display only the unique launch sites from the Launch_Site column of the SPACEXTBL Table.

Task 1

Display the names of the unique launch sites in the space mission

```
[10]: %sql SELECT DISTINCT Launch_Site from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

[10]:

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

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

[11]:
```
%sql SELECT * from SPACEXTABLE where Launch_Site like 'CCA%' Limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

[11]:

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

- Displayed 5 records where launch sites begin with `CCA`.
- Applied LIKE operator with % wildcard in WHERE clause to retrieve and display a table of all records where launch sites begin with the string CCA.

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

[12]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

[12]:

SUM(PAYLOAD_MASS__KG_)
45596

- Calculated the Total Payload Mass carried by boosters from NASA.
- Applied SELECT statement with SUM function to calculate, retrieve, and display the total sum of PAYLOAD_MASS__KG column for Customer NASA(CRS).

Average Payload Mass by F9 v1.1

- Calculated the Average Payload Mass carried by booster version F9 v1.1.
- Applied SELECT statement with AVG function to calculate, retrieve, and display the Average Payload Mass carried by booster version F9 v1.1.

Task 4

Display average payload mass carried by booster version F9 v1.1

```
[13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version='F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

[13]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Found the date of the first successful landing outcome on ground pad.
- Applied SELECT statement with MIN function to retrieve, calculate, and display the date when first successful landing outcome on ground pad.

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

[14]:
```
%sql SELECT Min(Date) from SPACEXTABLE where Landing_Outcome="Success (ground pad)";
```

```
 * sqlite:///my_data1.db
Done.
```

[14]:

Min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[15]:
```
%sql SELECT DISTINCT Booster_Version from SPACEXTABLE where Landing_Outcome="Success (drone ship)" and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000;
```

```
 * sqlite:///my_data1.db
Done.
```

[15]:

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

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4,000 kg but less than 6,000 kg.
- Applied SELECT DISTINCT statement to only list the unique names of boosters that had payload mass greater than 4,000 kg but less than 6,000 kg and had landing outcome of ‘Success (drone ship)’ in Landing_Outcome column.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

[111]:
```
%sql SELECT (SELECT Count(Mission_Outcome) from SPACEXTABLE where Mission_Outcome LIKE '%Success%') AS 'Successful Mission Outcomes', (SELECT Count (Mission_Outcome) from SPACEXTABLE w
```

```
 * sqlite:///my_data1.db
Done.
```

[111]:

Successful Mission Outcomes	Failure Mission Outcomes
100	1

- Calculated the Total Number of Successful and Failure Mission Outcomes.
- Applied COUNT function together with GROUP BY clause to calculate, retrieve, and display total number of missions outcomes.

Boosters Carried Maximum Payload

Task 8

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

```
[38]: %sql SELECT Booster_Version AS "Booster Versions Carried Max Payload Mass" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_"=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

```
 * sqlite:///my_data1.db
Done.
```

[38]:

Booster Versions Carried Max Payload Mass
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

- Listed the names of the booster versions which carried the Maximum Payload Mass.
- Applied a subquery that calculated and passed Max Payload Mass and then used that value to list all the booster versions that carried the Max Payload Mass.

2015 Launch Records

Task 9

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

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
[87]: %sql SELECT substr(Date,0,5) AS "Year", substr(Date, 6,2) AS "Month", "Booster_Version","Launch_Site", "Landing_Outcome" FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND "Landing_Out
```

```
 * sqlite:///my_data1.db
Done.
```

[87]:

Year	Month	Booster_Version	Launch_Site	Landing_Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Listed the failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015.
- Applied SUBSTR function with SELECT statement to retrieve the months and year 2015 from the Date column in SPACEXTABLE Table and COUNT function to count relevant ‘Failure (drone ship’) records in Landing_outcome column.

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

Task 10

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

[110]:
```
%sql SELECT Landing_Outcome, COUNT(*) AS "Count_of_Landing Outcomes" FROM SPACEXTABLE WHERE (Date BETWEEN '2010-06-04' AND '2017-03-20') GROUP BY Landing_Outcome ORDER BY "Count_of_Lan
```

```
 * sqlite:///my_data1.db
Done.
```

[110]:

Landing_Outcome	Count_of_Landing Outcomes
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

Applied COUNT function and display

- Ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.
- Applied WHERE clause, COUNT function, and GROUP BY clause together to retrieve, count, and group the count of landing outcomes between the given dates, ORDER BY keyword to rank the result before SELECT statement to display the outcome.

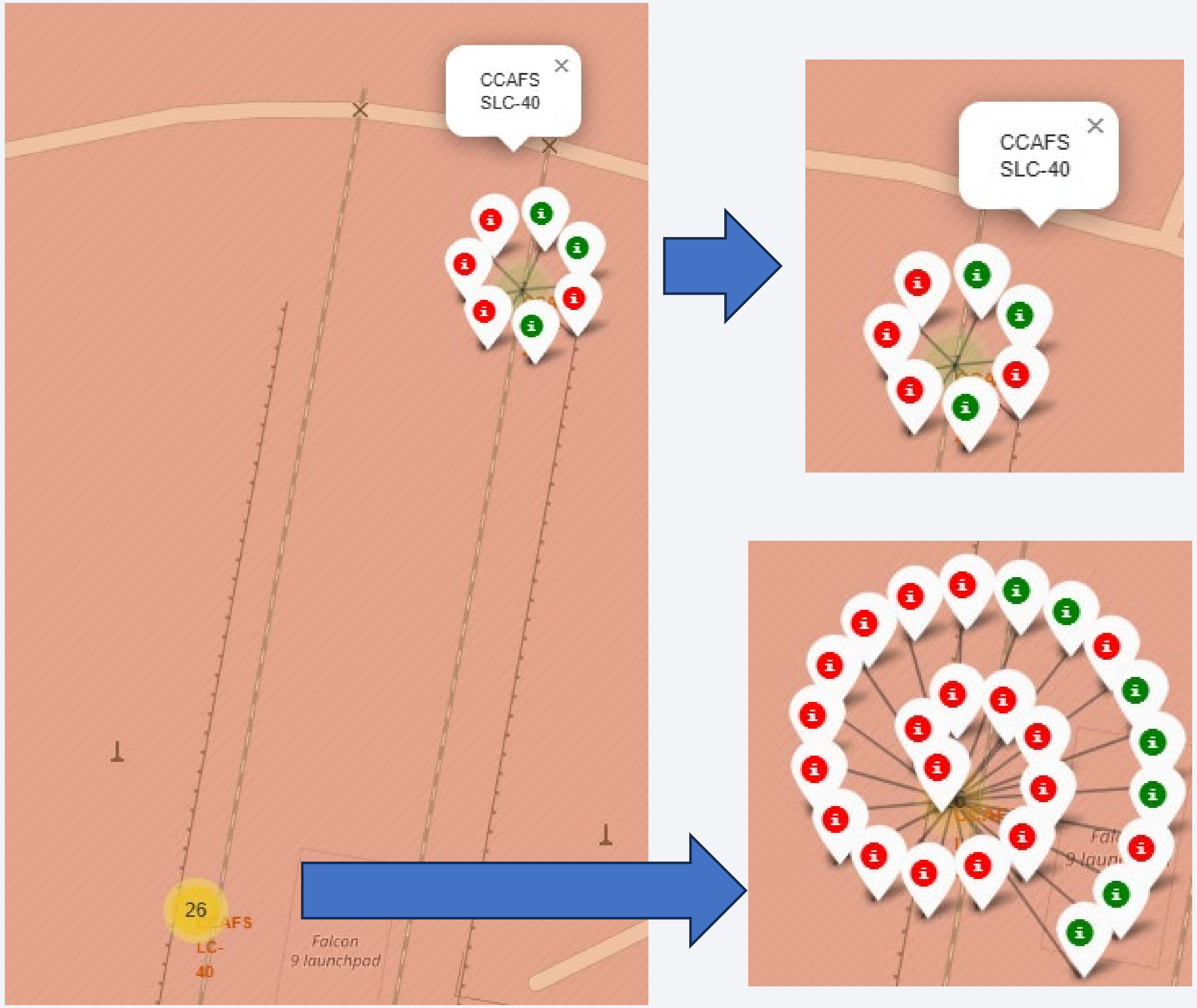
Section 3

Launch Sites Proximities Analysis

All Launch Sites

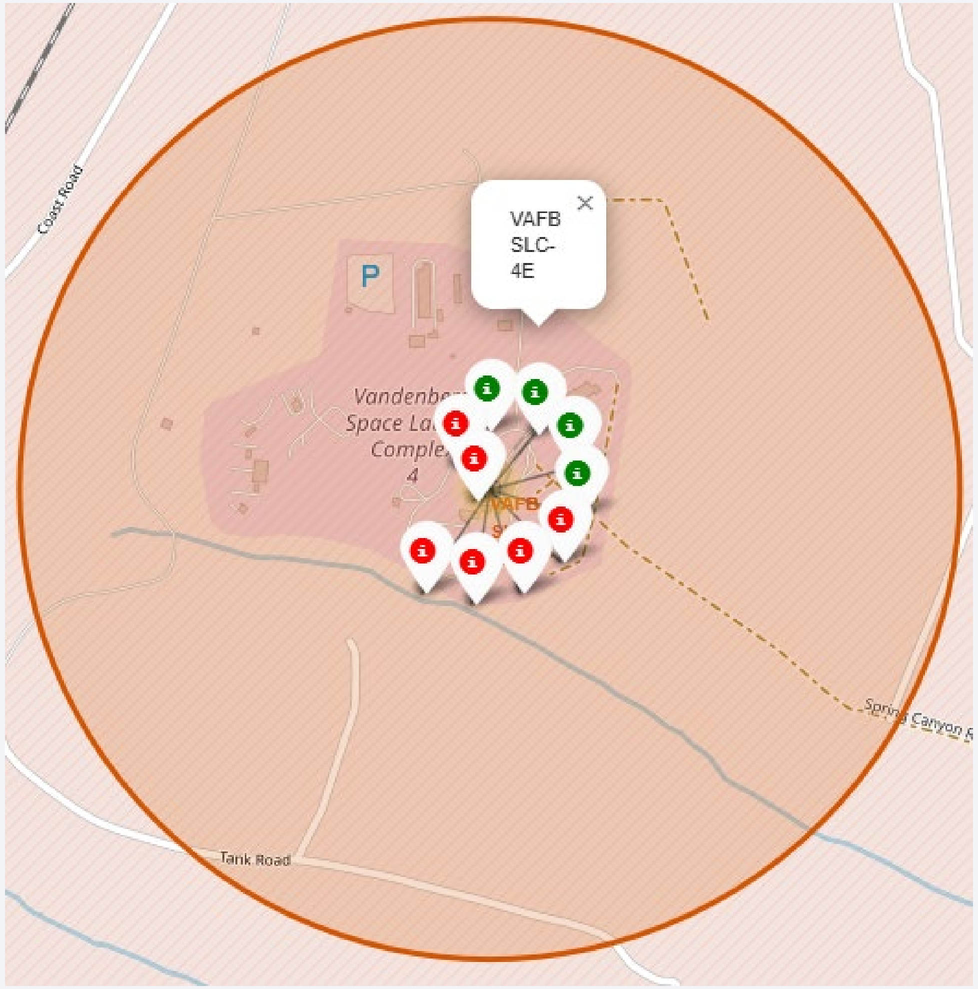
- All launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast.

Launch Outcomes (East Coat Locations)



- KSC LC-39A launch site has higher success rates than CCAFS LC-40 & CCAFS SLC-40 sites.

Launch Outcomes (West Coast Locations)



VAFB SLC-4E has lower success rates than KSC LC-39A launch site that is the most successful site in the East Coast.

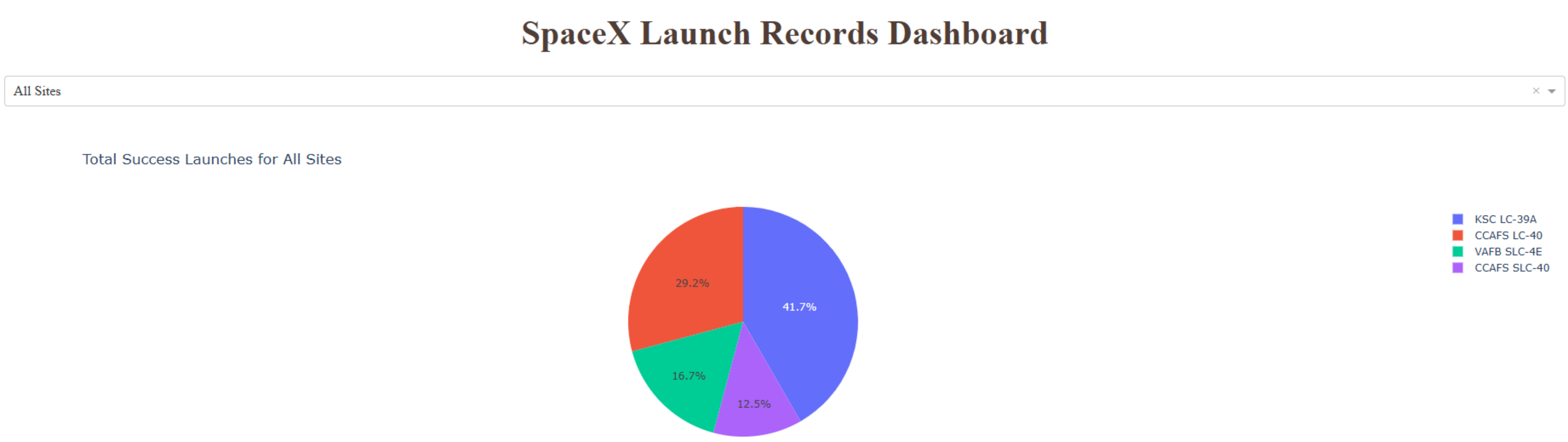
Distance of CCAFS SLC-40 Launch Site to Its Proximities

- CCAFS SLC-40 launch site is just 0.58 km away from the nearest highway.
- CCAFS SLC-40 launch site is just 0.85 km away from the nearest coastline.

Section 4

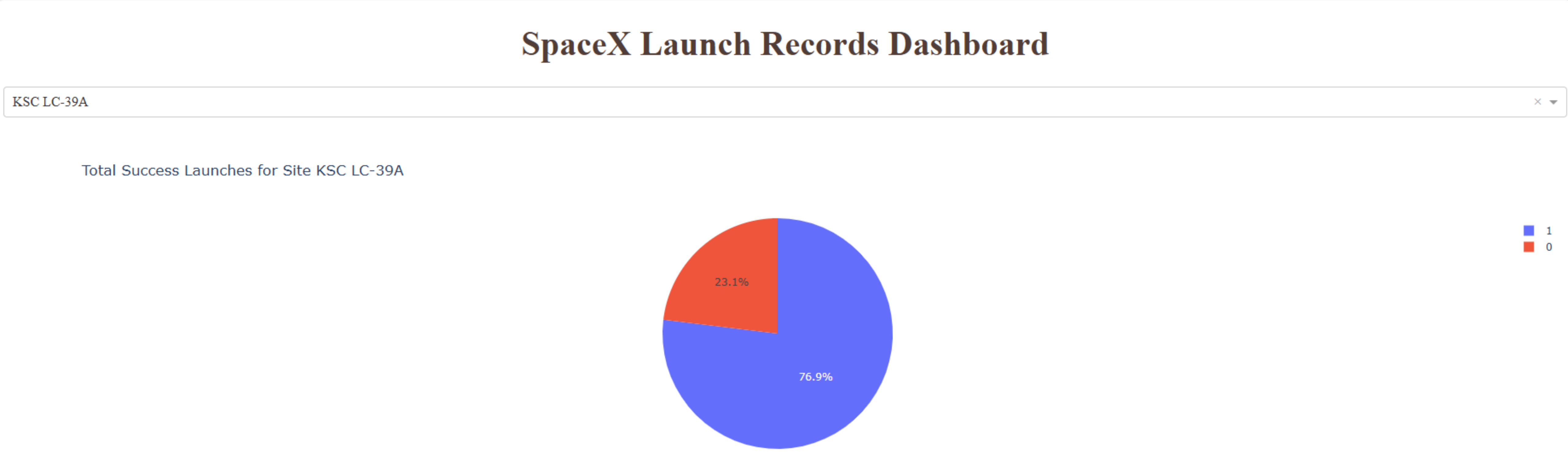
Build a Dashboard with Plotly Dash

Pie Chart for Launch Success Count for All Sites



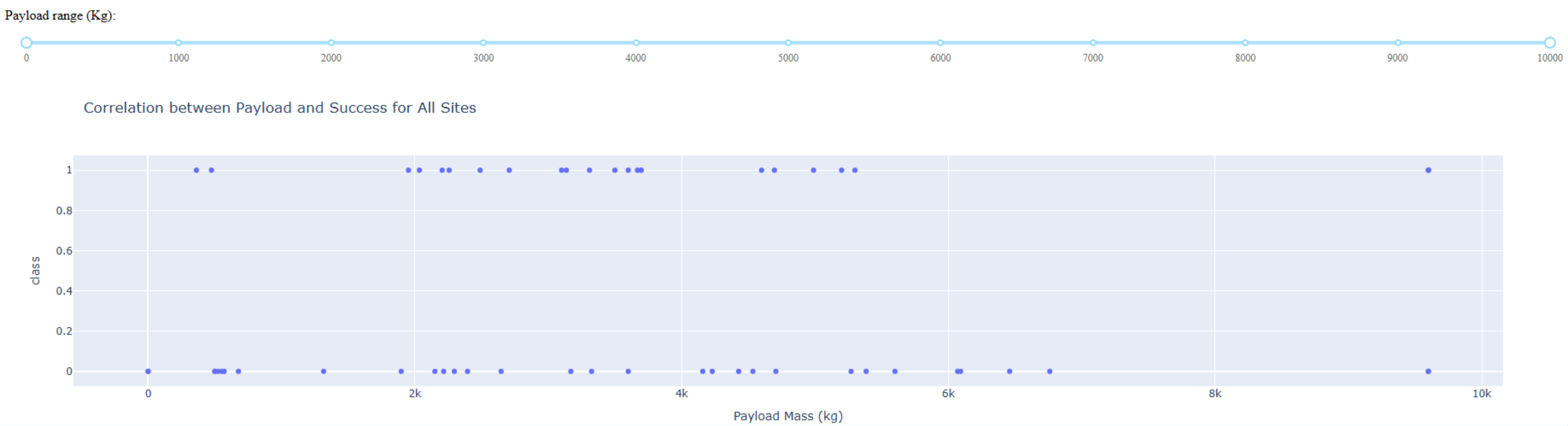
- KSC LC-39A launch site has the most successful launches amongst all launch sites.

Pie Chart for KSC LC-39A Launch Site



- KSC LC-39A launch site, being the most successful launch site, achieved a 76.9% success rate and 23.1% failure rate.

Payload vs. Launch Outcome Scatter Plot for All Sites



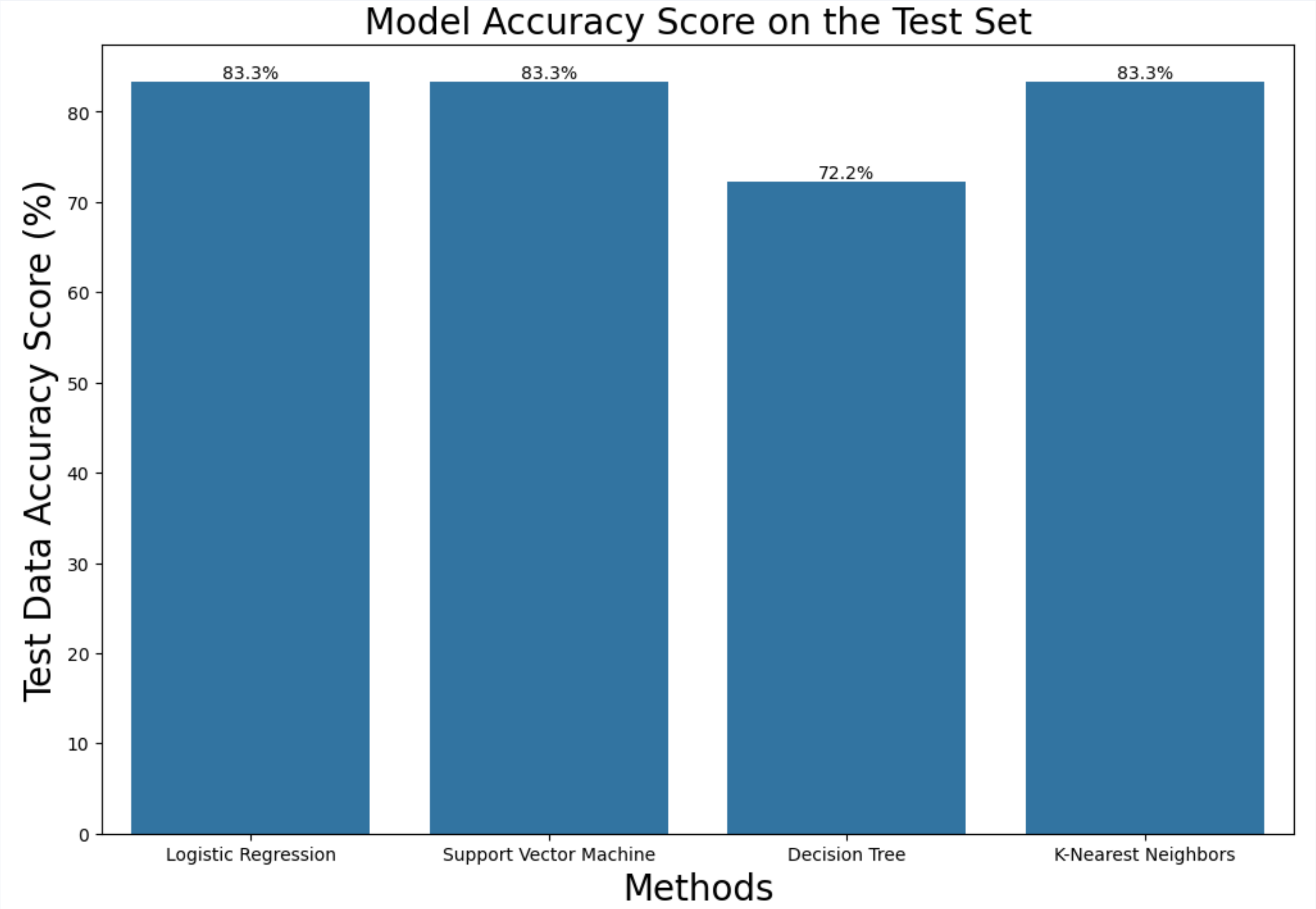
- Payloads between 1,000 kg and 6,000 kg have the highest success rate.

Section 5

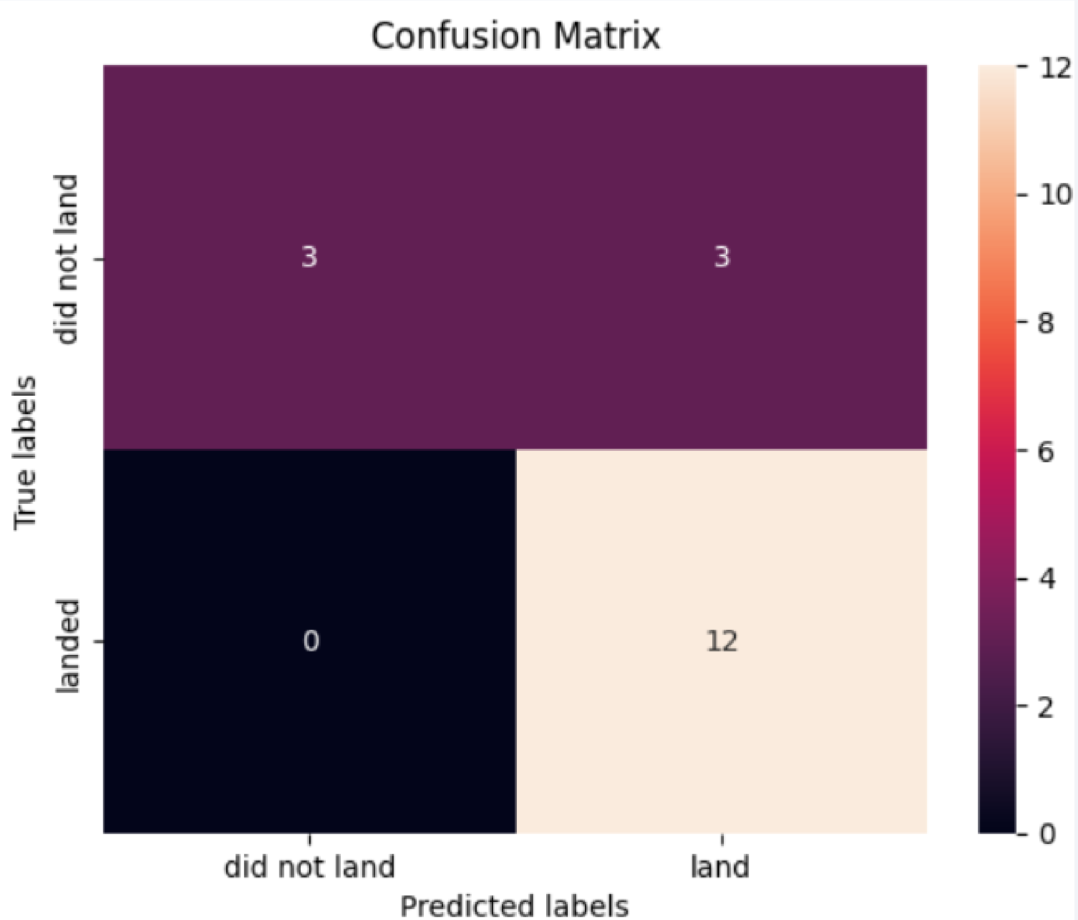
Predictive Analysis (Classification)

Classification Accuracy

- Logistic regression, support vector machine, and k-nearest neighbors models performed equally well with highest classification accuracy score of 83.33% on the Test Set.



Confusion Matrix



- All 3 successful classification models have the same confusion matrix shown on the left.
- Confusion matrix indicates that All 3 successful classification models can distinguish between the different classes.
- However, there is the same problem with false positives for all 3 models as follows:
  - True Positive - 12 (True label is landed, Predicted label is also landed.)
  - False Positive - 3 (True label is not landed, Predicted label is landed.)

Conclusions

- Logistic regression, support vector machine, and k-nearest neighbors methods can successfully predict if the Falcon 9 first stage will land successfully.
- KSC LC-39A launch site has the most successful launches amongst all sites.
- As the flight number increases, so does the success rate for VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40 launch sites.
- Payloads between 1,000 kg and 6,000 kg have the highest success rate.
- Orbits ES-L1, GEO, HEO, and SSO have the highest success rate.

Thank you!