IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gaurav Ganesh Amin
July 14, 2025

Outline

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

Executive Summary

- Project Objective: This project aims to predict the success of Falcon 9 first stage landings using comprehensive data analysis and machine learning.
- Data Collection & Preparation: Data regarding SpaceX rocket launches was collected through web scraping from SpaceX API endpoints and Wikipedia tables. The data underwent wrangling to handle missing values and transform the outcome into a binary Class label (1 for successful, 0 for unsuccessful) for machine learning.
- Exploratory Data Analysis (EDA):
  - EDA using visualizations (Seaborn, Folium, Plotly Dash) and SQL queries revealed key trends and relationships.
  - Key findings include an overall high mission success rate (100 successes vs. 1 failure) , and an increasing success probability with higher flight numbers and over time (2013-2020 trend).
  - Launch site significantly impacts success, with VAFB SLC-4E and KSC LC-39A showing higher success rates. Orbit type and payload mass also influence success, with certain orbits (e.g., ES-L1, GEO, HEO, SSO) showing 100% success and heavier payloads increasing success for specific orbits (e.g., Polar, LEO, ISS).
- Predictive Modeling: Supervised machine learning models (Logistic Regression, SVM, Decision Tree, KNN) were developed, with hyperparameters optimized using GridSearchCV. All models achieved a consistent classification accuracy of approximately 83.33% in predicting landing outcomes, demonstrating their robust performance.

Introduction

- Project background and context
  - SpaceX's Falcon 9: Focus on the Falcon 9 rocket and its significance in space launches.
  - Cost-Effectiveness: Highlight SpaceX's competitive pricing ($62 million) compared to other providers (upwards of $165 million).
  - First Stage Reusability: Emphasize that a significant portion of these savings comes from the ability to reuse the first stage.
- Problems you want to find answers
  - Landing Prediction: Can we accurately predict whether the Falcon 9 first stage will land successfully?
  - Cost Determination: If successful landing can be predicted, how does this impact the determination of launch costs?
  - Competitive Bidding: How can this predictive capability be leveraged by alternate companies bidding against SpaceX for rocket launch contracts?

Section 1

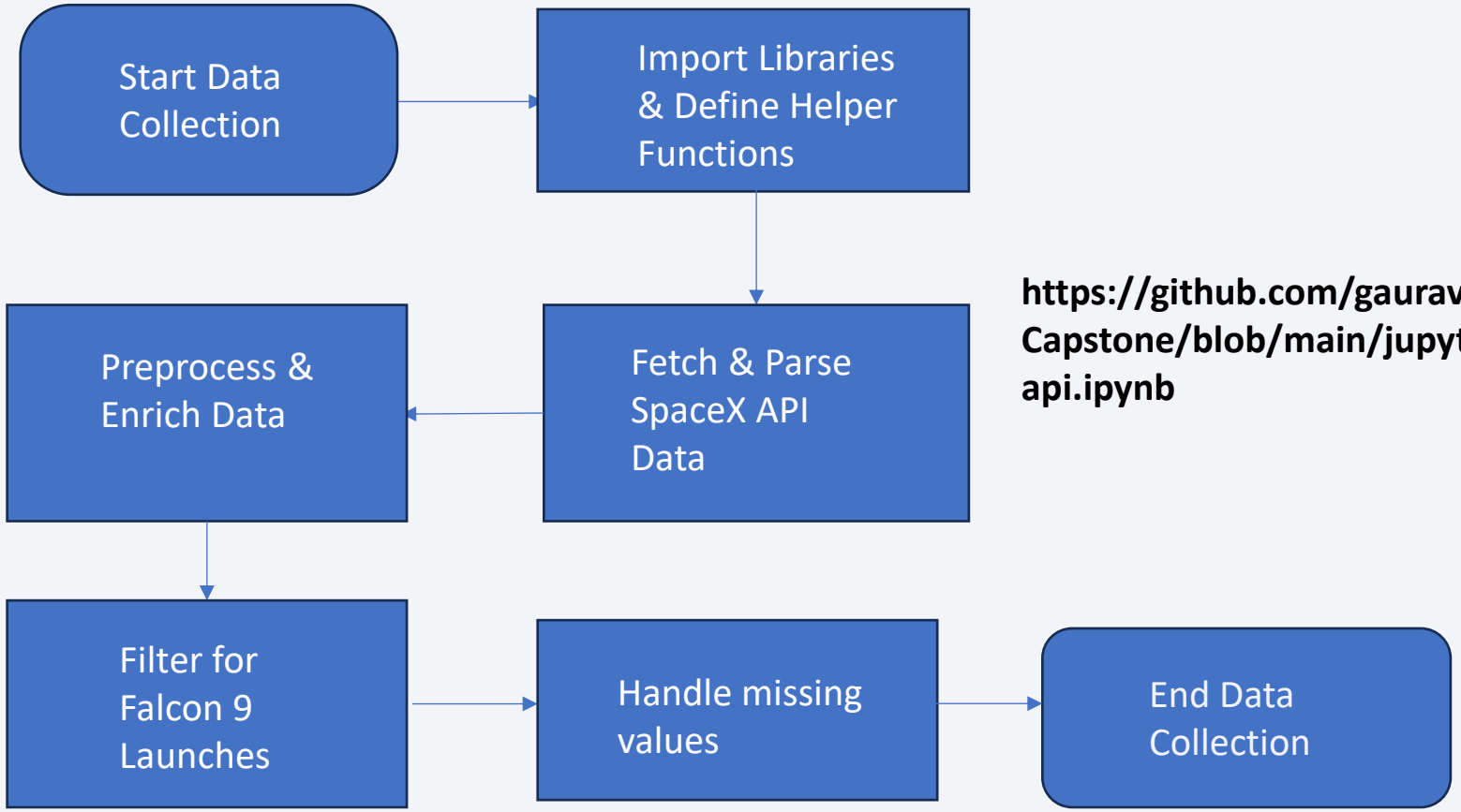
Methodology

Methodology

Executive Summary

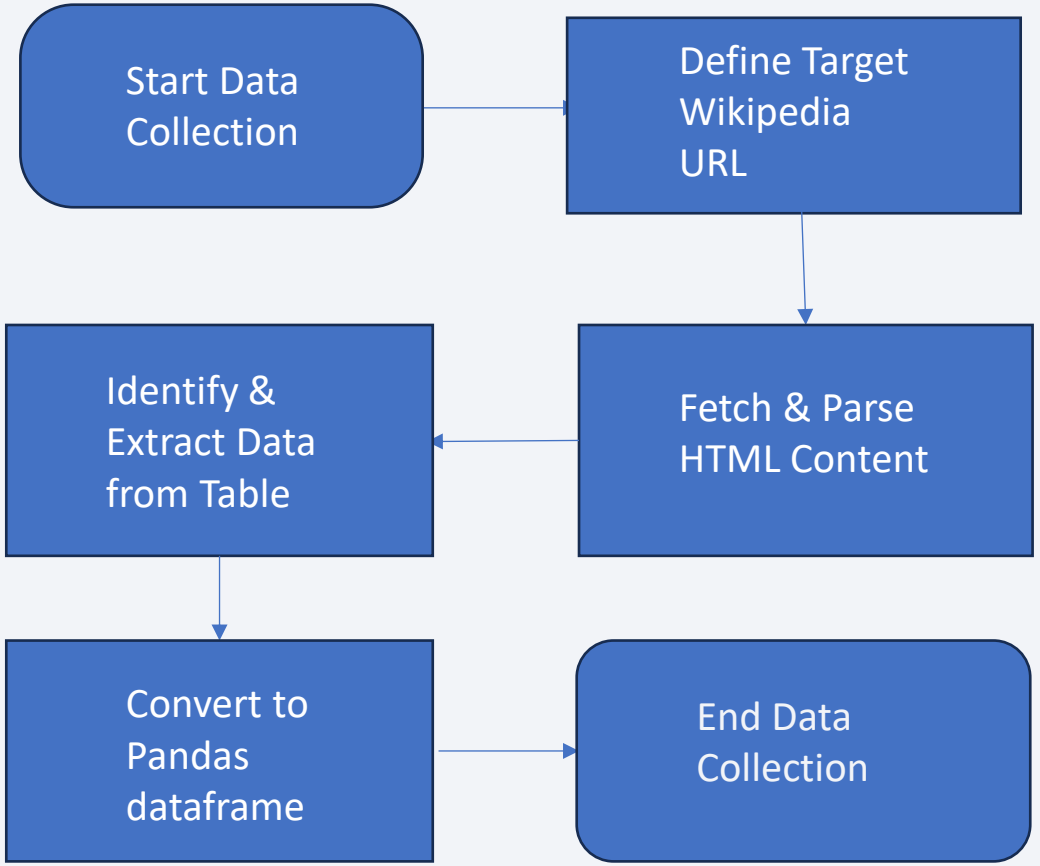
- Data collection methodology:
  - Collected SpaceX launch data from SpaceX REST API endpoint. Also, used Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.
- Perform data wrangling
  - The data wrangling process involved loading the Space X dataset, identifying missing values, and transforming the Outcome column into a binary Class label (1 for successful landing, 0 for unsuccessful landing) to prepare the data for supervised machine learning 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
  - Standardize data, split it into training and test sets, then optimize hyperparameters for algorithms like Logistic Regression, SVM, Decision Tree, and KNN using GridSearchCV with cross-validation on the training data, and finally assess performance on the test set using the score method and a confusion matrix.

Data Collection – SpaceX API



https://github.com/gauravganeshamin/IBM_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

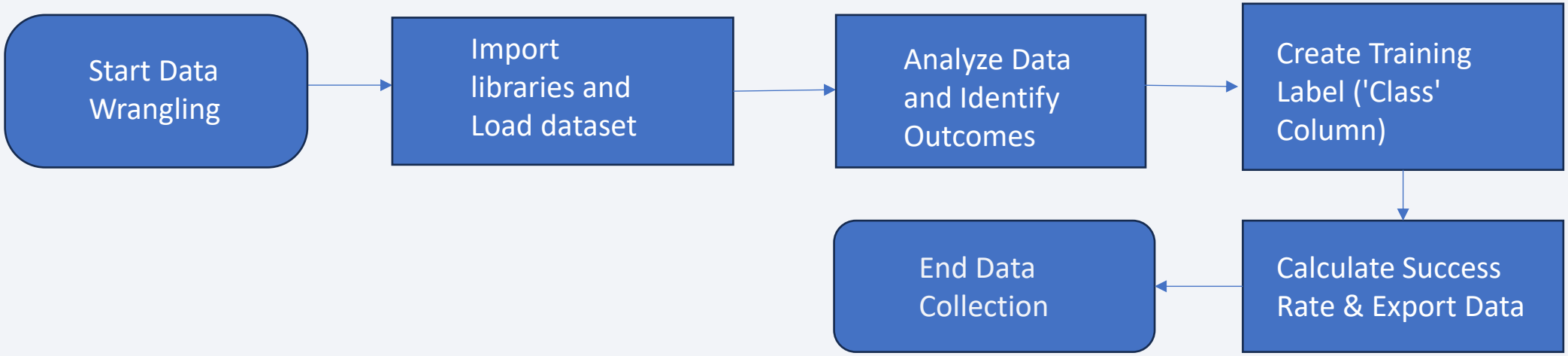
Data Collection - Scraping



https://github.com/gauravganeshamin/IBM_Data_Science_Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

https://github.com/gauravganeshamin/IBM_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- This lab primarily used Seaborn's catplot to visualize the relationships between FlightNumber, PayloadMass, LaunchSite, and Orbit with the Class (landing outcome).
- These scatter point plots were chosen to discern how continuous variables (Flight Number, Payload Mass) and categorical variables (Launch Site, Orbit) individually and interactively affect the success of a Falcon 9 first stage landing.
- Additionally, a bar chart was employed to show the success rate for each orbit type, providing a clear comparison of landing performance across different orbital destinations.
- Finally, a line plot was generated to illustrate the yearly trend in launch success rates, effectively demonstrating the improvement in landing success over time. Collectively, these charts facilitated exploratory data analysis to uncover patterns and identify important variables for future prediction models.

URL - **https://github.com/gauravganeshamin/IBM_Data_Science_Capstone/blob/main/edadataviz.ipynb**

EDA with SQL

- The SQL queries performed on the SPACEXTABLE dataset aimed to extract key insights from SpaceX launch records. These queries covered a range of analytical tasks, including displaying unique launch sites and filtering records based on specific launch site patterns.
- They calculated total and average payload masses for missions by specific customers (NASA CRS) and for particular booster versions (F9 v1.1).
- Temporal analyses included identifying the earliest date of a successful landing outcome and listing details of failed drone ship landings in 2015, extracting month names from dates.
- Furthermore, queries were used to identify booster versions that carried the maximum payload mass and to rank landing outcomes by their frequency within a defined date range, providing a comprehensive overview of mission performance and trends.

URL - **https://github.com/gauravganeshamin/IBM_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb**

Build an Interactive Map with Folium

Key map objects created and added include:

- folium.Map: This served as the foundational canvas, initialized at a central coordinate to provide the global context for plotting launch sites.
- folium.Circle and folium.map.Marker (with DivIcon): These were used together for each launch site. The Circle visually highlights the area of each site, while the Marker with a DivIcon displays the launch site's name directly on the map, providing immediate identification.
- folium.plugins.MarkerCluster: This was crucial for managing the numerous launch records that often share the same geographical coordinates. It groups overlapping markers into a clickable cluster, de-cluttering the map and enhancing user experience, especially when zoomed out.
- Color-coded folium.map.Markers (with folium.Icon): Individual launch outcomes were marked with these. Green markers indicated successful launches (class=1), while red markers represented failed launches (class=0). These were added to the MarkerCluster to visually denote the success/failure rate at each site at a glance.
- folium.plugins.MousePosition: This interactive tool was added to allow users to obtain precise latitude and longitude coordinates by hovering their mouse over any point on the map. This functionality was essential for identifying the exact locations of nearby geographical features like coastlines, railways, highways, and cities.
- Distance-displaying folium.map.Markers (with DivIcon) and folium.PolyLine: After calculating distances between launch sites and identified proximity points (e.g., coastline), dedicated Markers were created to display these distances as text labels on the map. Corresponding PolyLine objects were drawn to visually connect the launch site to these proximity points, illustrating the measured distances and helping to analyze the geographical factors influencing launch site selection.

URL - **https://github.com/gauravganeshamin/IBM_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb**

Build a Dashboard with Plotly Dash

The dashboard incorporates the following interactive elements and corresponding plots:

- Launch Site Dropdown (site-dropdown): This dropdown allows users to select a specific launch site or view data for all sites. Its inclusion enables users to filter the data interactively, focusing on the performance and characteristics of individual launch locations.
- Pie Chart (success-pie-chart): This chart initially displays the total successful launch counts for all sites. When a specific launch site is selected from the dropdown, it dynamically updates to show the proportion of successful versus failed launches for that particular site. A pie chart is used here to provide a clear, immediate visualization of the success rate distribution.
- Payload Range Slider (payload-slider): This slider allows users to define a minimum and maximum payload mass (in kg) range. This control is vital for investigating how launch success might be influenced by the payload's weight.
- Scatter Chart (success-payload-scatter-chart): This plot visualizes the correlation between payload mass and launch success (class), with different booster versions distinguished by color. The scatter chart dynamically updates based on both the selected launch site from the dropdown and the chosen payload range from the slider. It's used to identify potential relationships between payload characteristics, booster type, and the likelihood of a successful landing.

These plots and interactions were chosen to facilitate exploratory data analysis on SpaceX launch data. The dropdown and slider allow users to drill down into specific subsets of data, while the pie chart and scatter plot offer immediate visual insights into success rates and underlying correlations, respectively. The interactive nature empowers users to discover patterns and dependencies between launch variables and outcomes efficiently.

URL - **https://github.com/gauravganeshamin/IBM_Data_Science_Capstone/blob/main/spacex-dash-app.py**

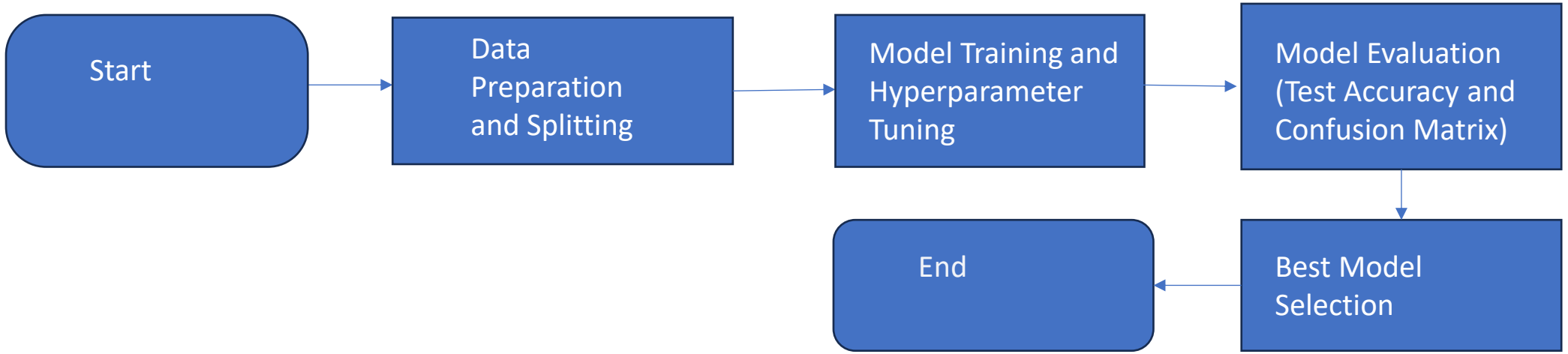
Predictive Analysis (Classification)

To summarize the process of building, evaluating, and improving classification models:

- Data Preparation: The first step involved preparing the data by extracting the target variable Y (landing success) and standardizing the feature set X using StandardScaler. This ensures all features contribute equally to the model.
- Data Splitting: The prepared data was then split into training and testing sets using train_test_split with an 80/20 ratio and a random_state of 2. This allowed for unbiased evaluation of the models on unseen data.
- Model Training and Hyperparameter Tuning: For each of the four classification algorithms (Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors), GridSearchCV was used with 10-fold cross-validation (cv=10) to systematically search for the optimal hyperparameters. This process involved fitting the models to the training data and identifying the best parameter combinations that maximized the accuracy on the validation sets.
- Model Evaluation: After finding the best hyperparameters for each model, their performance was evaluated on the unseen test data using the score method. Additionally, confusion matrices were plotted for each model to provide a detailed breakdown of true positives, true negatives, false positives, and false negatives, offering insights into the types of errors each model made.
- Best Model Selection: Finally, the test accuracies of all four optimized models were compared to determine which method performed best. The model with the highest accuracy on the test data was identified as the best performing classification model for predicting Falcon 9 first stage landings.

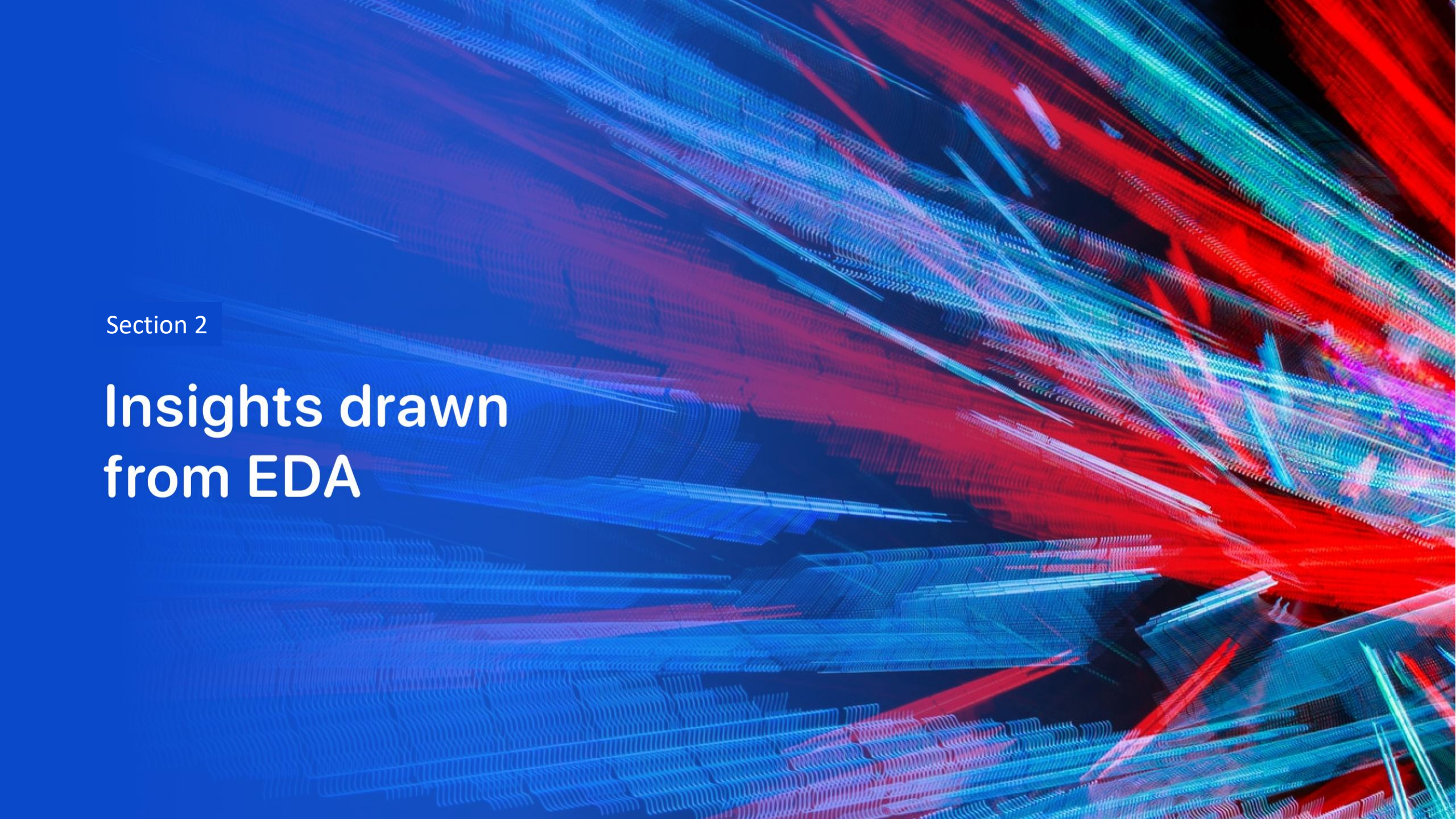
URL - **https://github.com/gauravganeshamin/IBM_Data_Science_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb**

Predictive Analysis (Classification)



Results

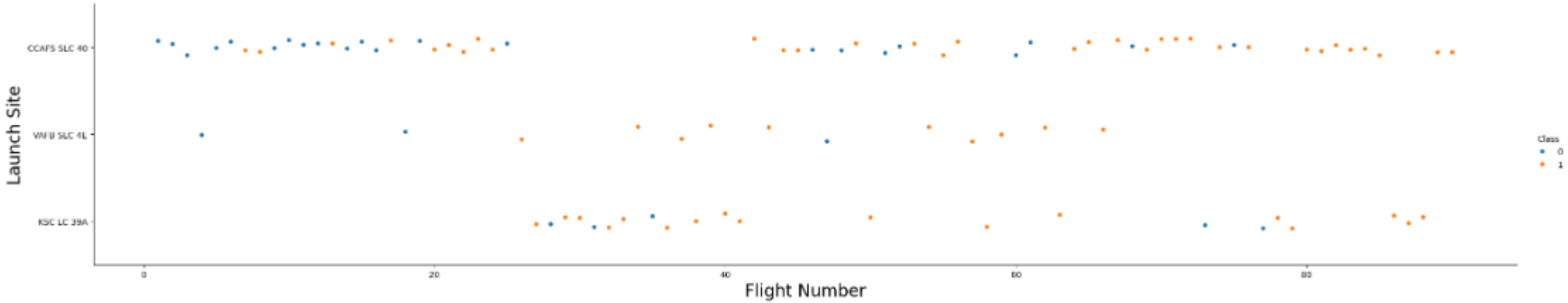
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

In [5]:

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

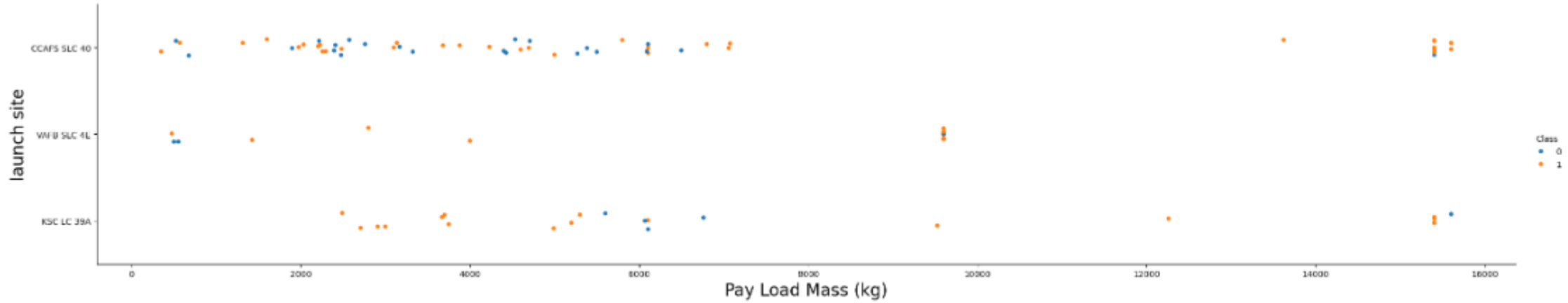


We see that as flight numbers increase, the probability of successful first stage increases irrespective of launch site.

Payload vs. Launch Site

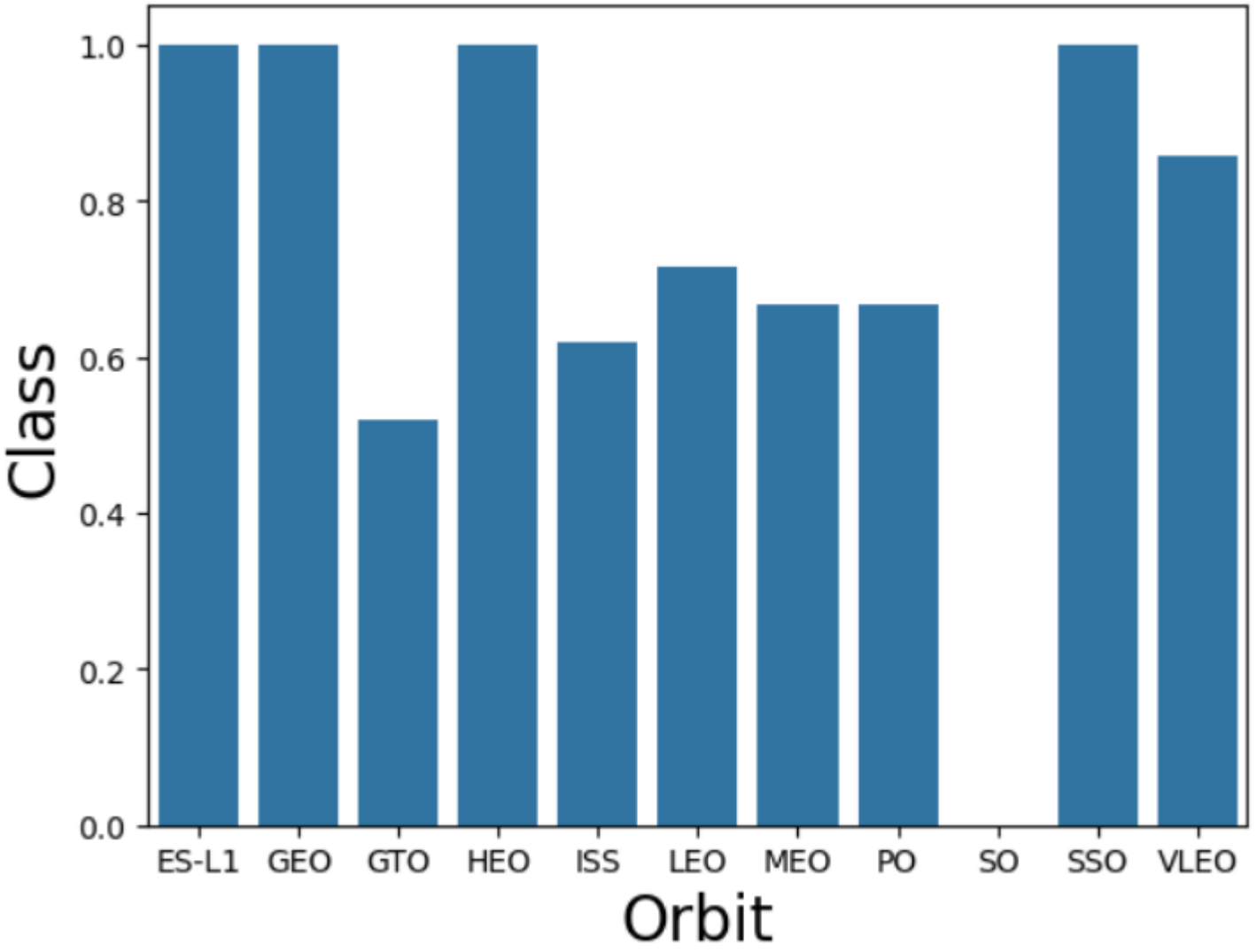
In [7]:

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the cla
sns.catplot(x="PayloadMass", y="LaunchSite", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay Load Mass (kg)", fontsize = 20)
plt.ylabel("launch site", fontsize = 20)
plt.show()
```



The success of the first stage is more in VAFB SLC-4E and KSC LC-39A as compared to other launch site irrespective of payload mass.

Success Rate vs. Orbit Type

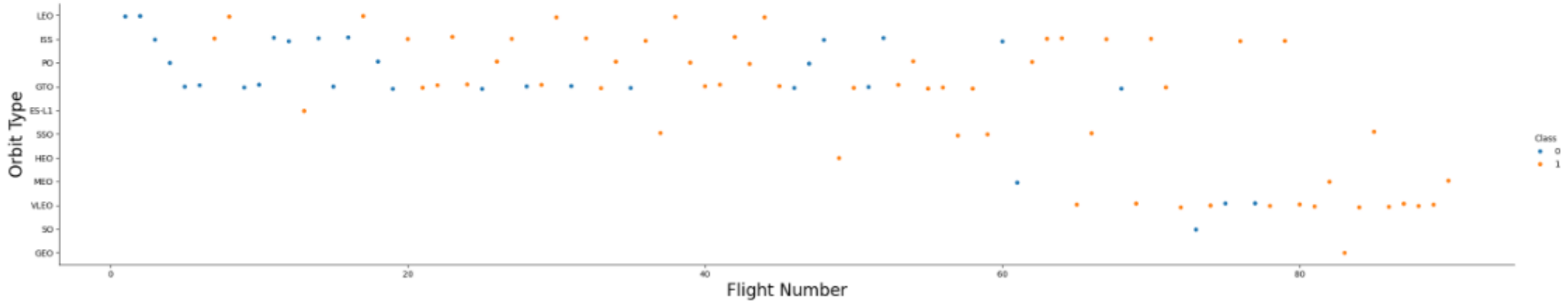


We can observe that different orbit types give different success rates. On one hand, there is 100% success rate when rocket is launched to orbits like ES-L1, GEO, HEO, SSO, while there is no absolutely chance of success when rocket is launched into SO orbit.

Flight Number vs. Orbit Type

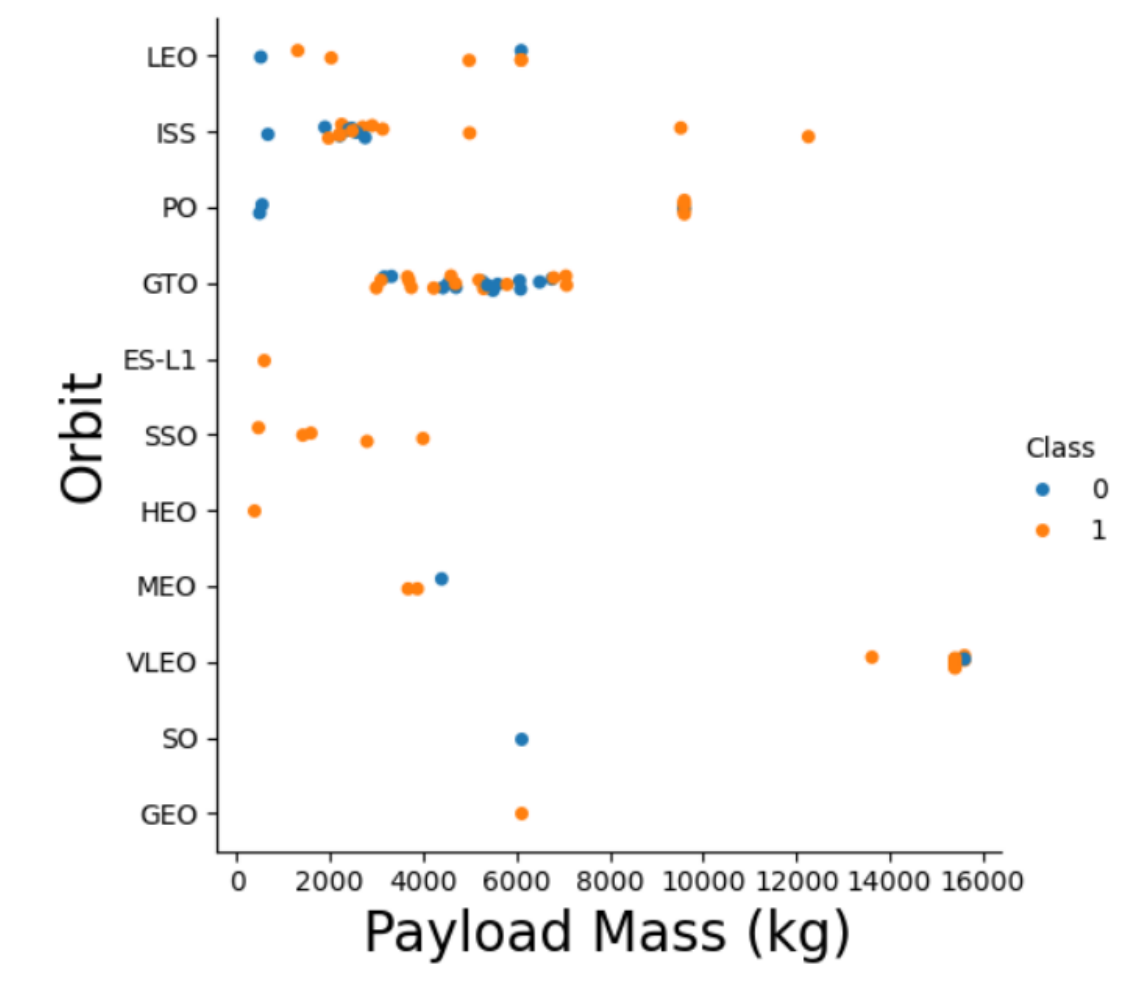
In [18]:

```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(x = 'FlightNumber', y = 'Orbit', hue = 'Class', data = df, aspect = 5)
plt.xlabel('Flight Number', fontsize = 20)
plt.ylabel('Orbit Type', fontsize = 20)
plt.show()
```



We can observe that for LEO Orbit, success increases as the number of flights increases. Conversely, for GTO orbit, there appears to be no relationship between success and flight number.

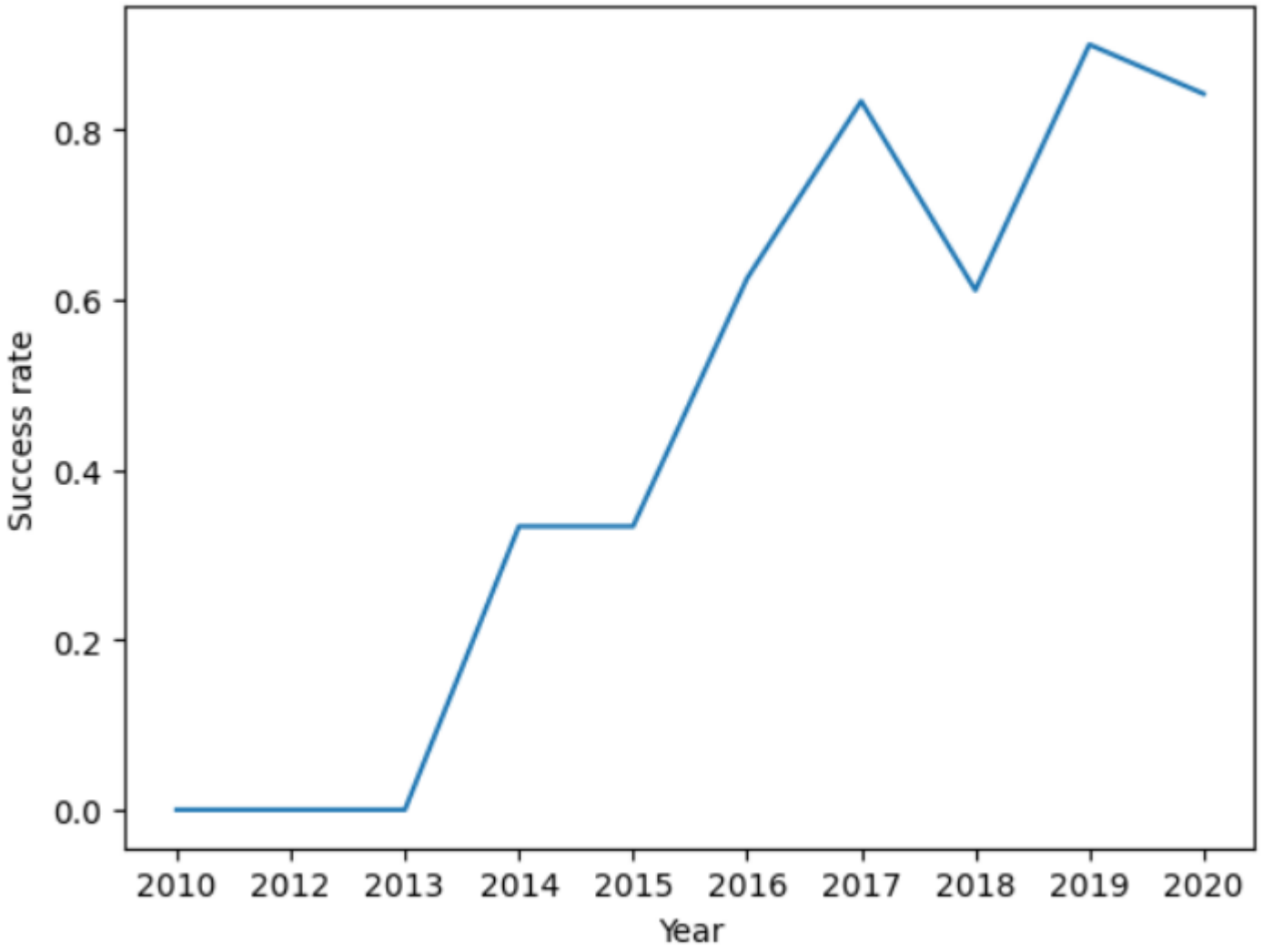
Payload vs. Orbit Type



We see that for heavier payloads, successful landing rates increases for Polar, LEO, ISS orbits.

However, for GTO, payload mass does not seem to have an impact on successful landing rate.

Launch Success Yearly Trend



We can see that the success rate increases overall from 2013 to 2020.

All Launch Site Names

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

```
In [10]: %sql select distinct launch_site from spacextable
```

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

Out[10]:

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

The query result displays the name of unique launch sites present in the SpaceX mission dataset.

Launch Site Names Begin with 'CCA'

In [11]:

```
%%sql
select * from spacextable
where launch_site like 'CCA%'
limit 5
```

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

Out[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

The query result displays the launch sites whose names begin with ‘CCA’.

Total Payload Mass

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

In [18]:
```
%%sql
select sum(payload_mass__kg_) from spacextable
where customer = 'NASA (CRS)'
```

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

Out[18]:

sum(payload_mass__kg_)
45596

The query result provides the total payload mass (45596 kg), which is carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

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

In [19]:
```
%%sql
select avg(payload_mass__kg_) from spacextable
where booster_version = 'F9 v1.1'
```

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

Out[19]:

avg(payload_mass__kg_)
2928.4

The query result gives the average payload mass (2928.4 kg) carried by booster version F9 v1.1.

First Successful Ground Landing Date

In [20]:

```
%%sql
select min(date) from spacextable
where landing_outcome like 'Success%'
```

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

Out[20]:

min(date)
2015-12-22

The query result gives the first date when ground landing was successful.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [21]:

```
%%sql
select booster_version from spacextable
where landing_outcome like '%Success (%drone ship)%'
and payload_mass__kg_ >4000 and payload_mass__kg_ <6000
```

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

Out[21]:

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

The query result gives the booster versions of rockets which landed successfully on drone ships and have payloads between 4000-6000 kg.

Total Number of Successful and Failure Mission Outcomes

In [29]:

```
%%sql
select
case
    when mission_outcome like '%Success%' then 'Success'
    when mission_outcome like '%Failure%' then 'Failure'
end as general_mission_outcome,
count(*) as cnt
from spacextable
group by general_mission_outcome
```

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

Out[29]:

general_mission_outcome	cnt
Failure	1
Success	100

The query result provides the number of successful and failed rocket launch missions.

Boosters Carried Maximum Payload

In [30]:
```
%%sql
select booster_version from spacextable
where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextable)
```

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

Out[30]:

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

The query result gives the booster versions of launched rockets which carried maximum payload.

2015 Launch Records

In [32]:

```
%%sql
select substr(date,6,2) as date_month, landing_outcome, booster_version, launch_site
from spacextable
where substr(date,0,5) = '2015'
and landing_outcome = 'Failure (drone ship)'
```

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

Out[32]:

date_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

The query result lists the booster versions of rockets launched during 2015 but failed to land successfully on drone ship, along with month of launch, landing outcome and launch site from which rocket was launched.

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

In [43]:

```
%%sql
with landing_outcome_occurence as
(
    select landing_outcome, count(landing_outcome) as cnt
    from spacextable
    where date between '2010-06-04' and '2017-03-20'
    group by landing_outcome
)
select landing_outcome, cnt,
dense_rank() over(order by cnt desc) as rnk
from landing_outcome_occurence
```

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

Out[43]:

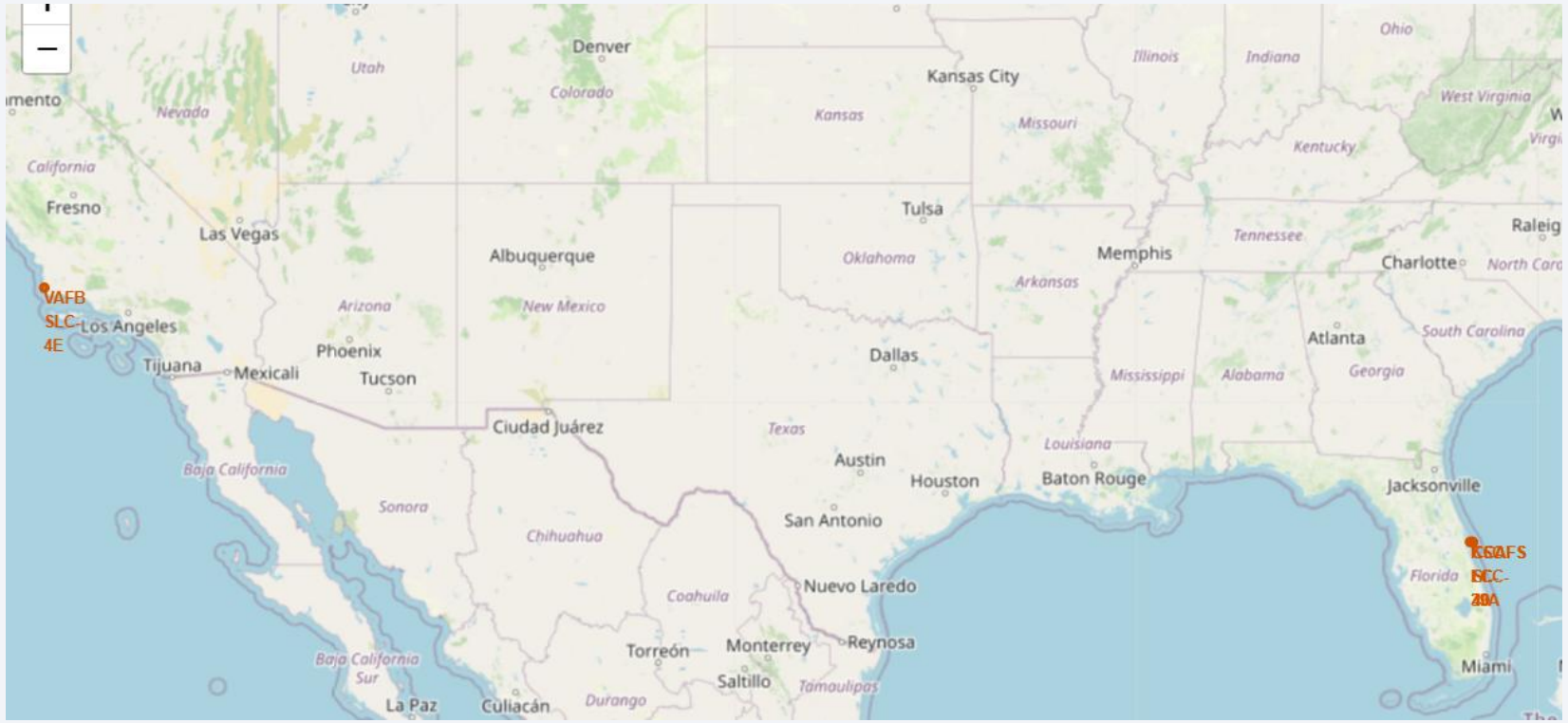
landing_outcome	cnt	rnk
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	3
Success (ground pad)	3	3
Failure (parachute)	2	4
Uncontrolled (ocean)	2	4
Precluded (drone ship)	1	5

This query result ranks landing outcomes for rockets launched between 2010-06-04 and 2017-03-20 based on the number of each landing outcome observed. The landing outcome observed the most was ranked first.

Section 3

Launch Sites Proximities Analysis

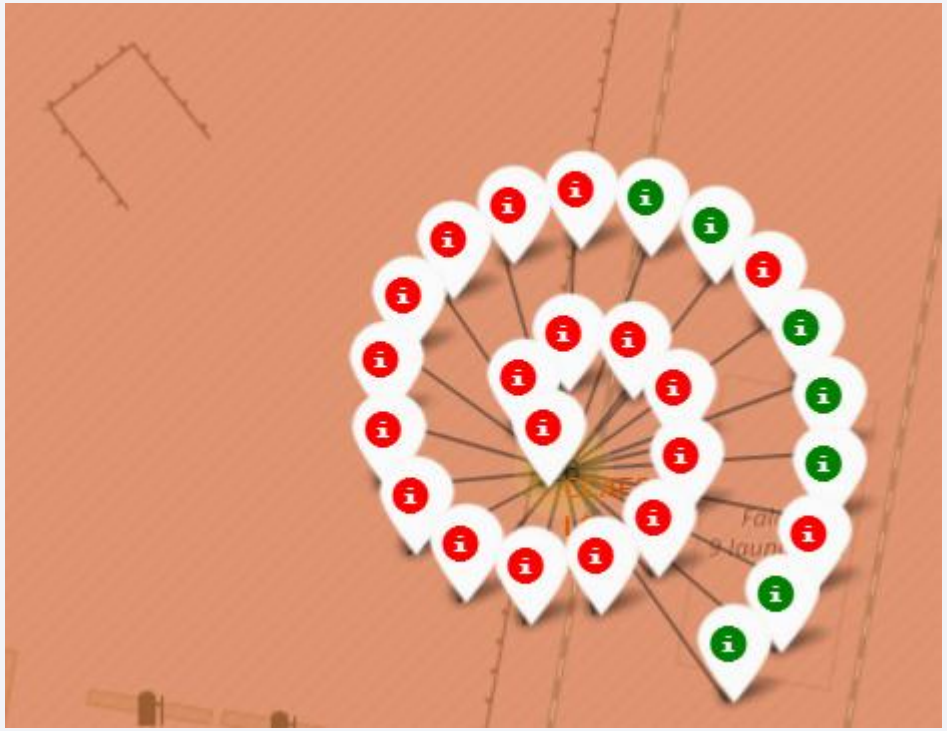
Location markers for launch sites



In this screenshot of the map, the locations of launch sites have been marked by red circles with names of launch sites.

The launch sites are located close to the coast, most of them located close to the East Coast.

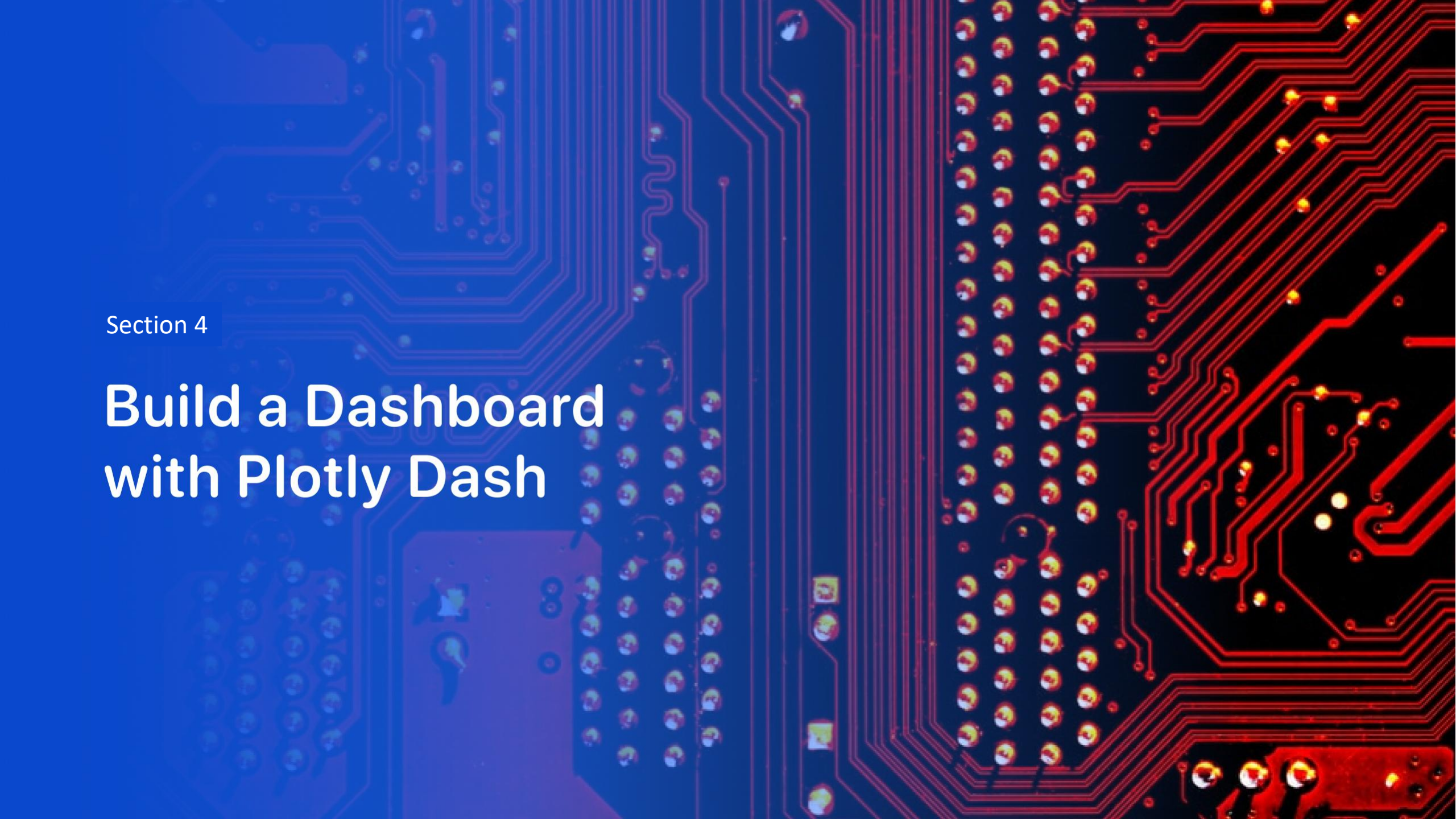
Landing outcome for each rocket launch



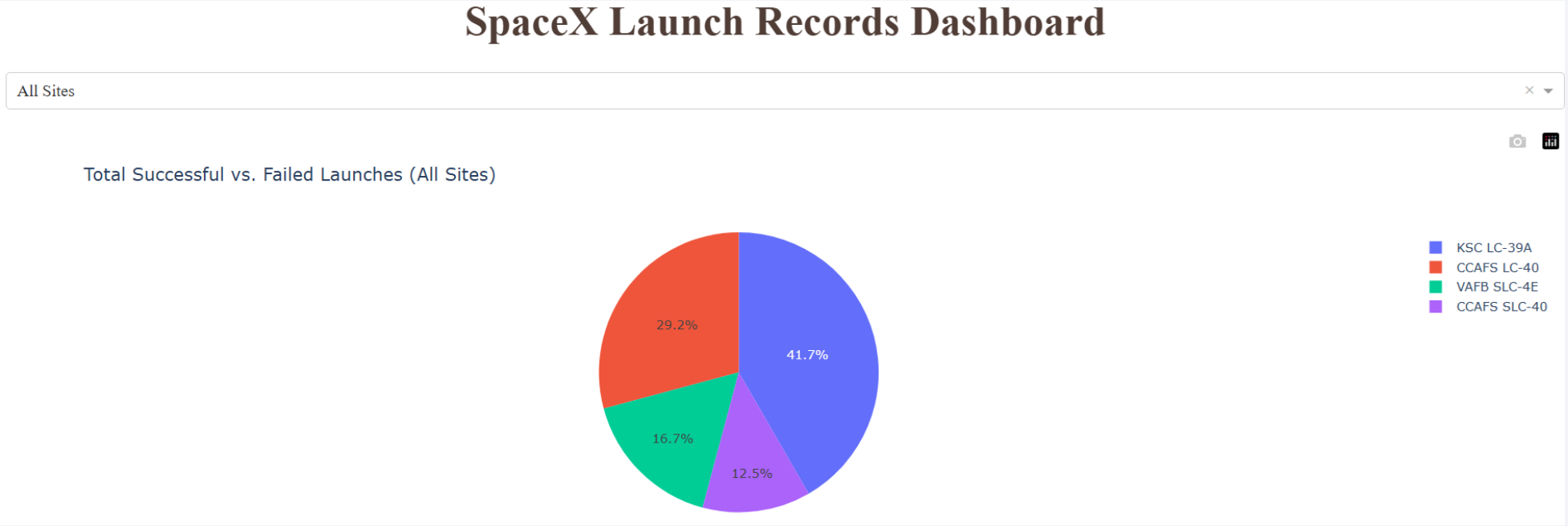
In this screenshot, the landing outcomes for each rocket launch in the SpaceX dataset have been marked, denoted by red for failed outcomes and green for successful outcomes.

Proximity of launch site

In this given screenshot, the proximity of launch site KSC LC-39 A to highway FL-405 has been displayed with a distance of 10.76 km.

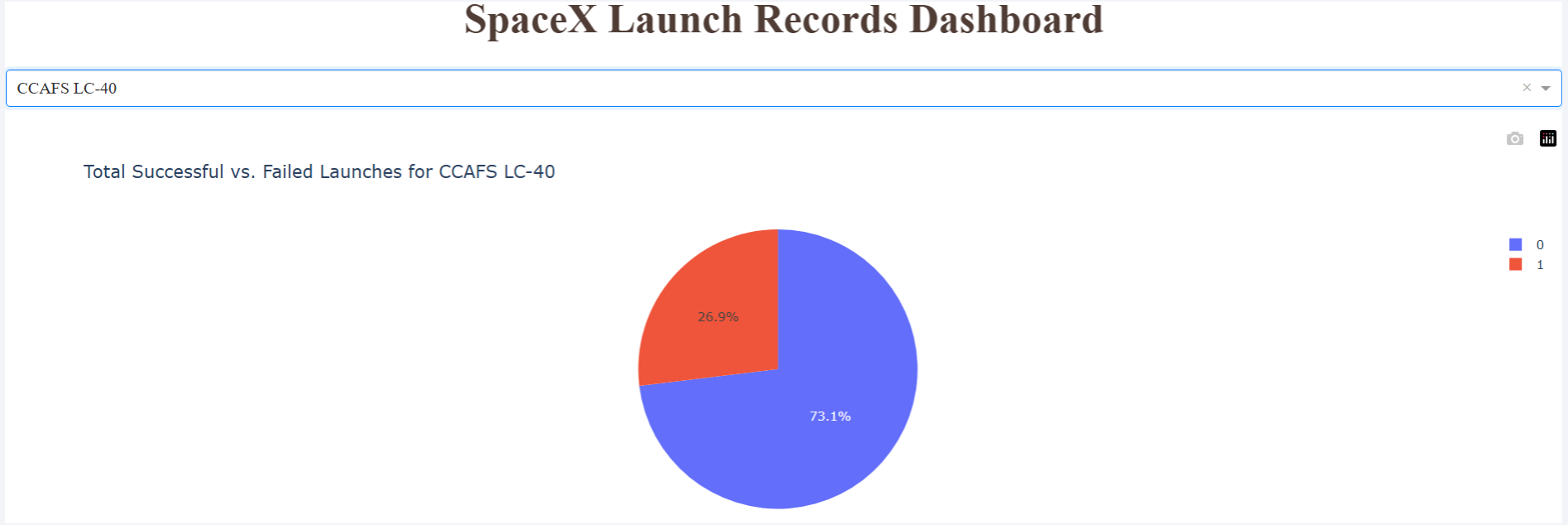

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all launch sites



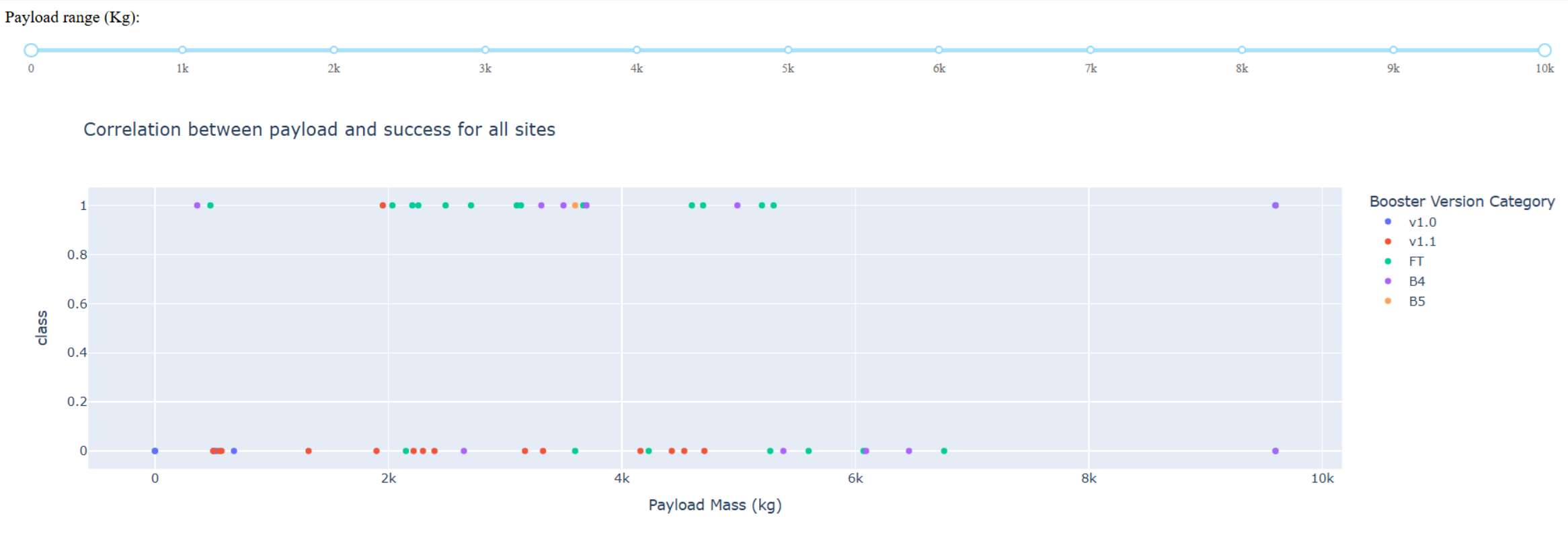
The above screenshot shows pie chart containing the success count for all launch sites.

Launch site with highest launch success ratio



The above screenshot shows the pie chart containing the success ratio for the launch site CCAFS LC-40.

Payload vs launch outcome scatter plot

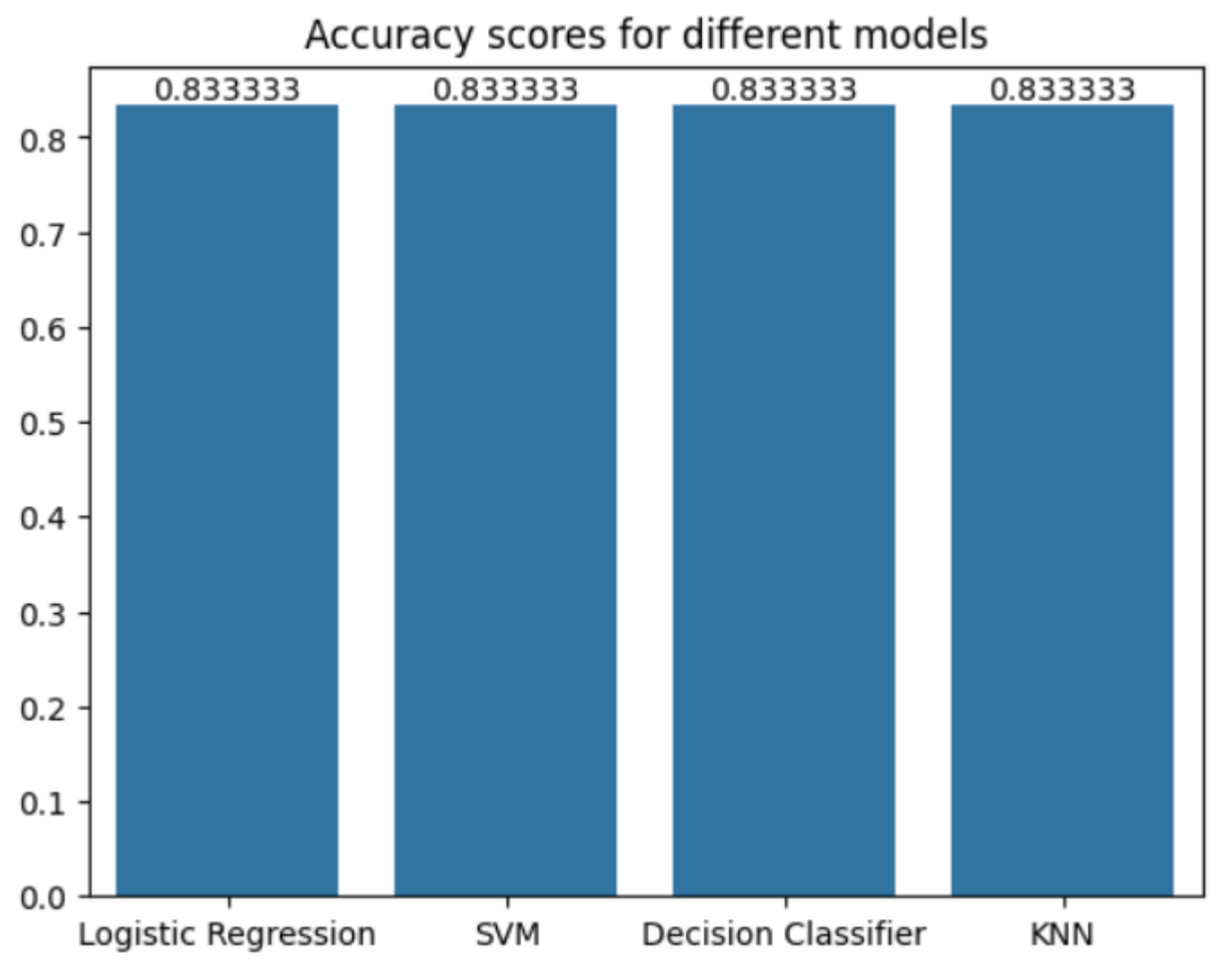


Booster version FT has the highest successful launch outcome, I believe that payload range of 0-10000 kg of payload will give more successful outcome.

Section 5

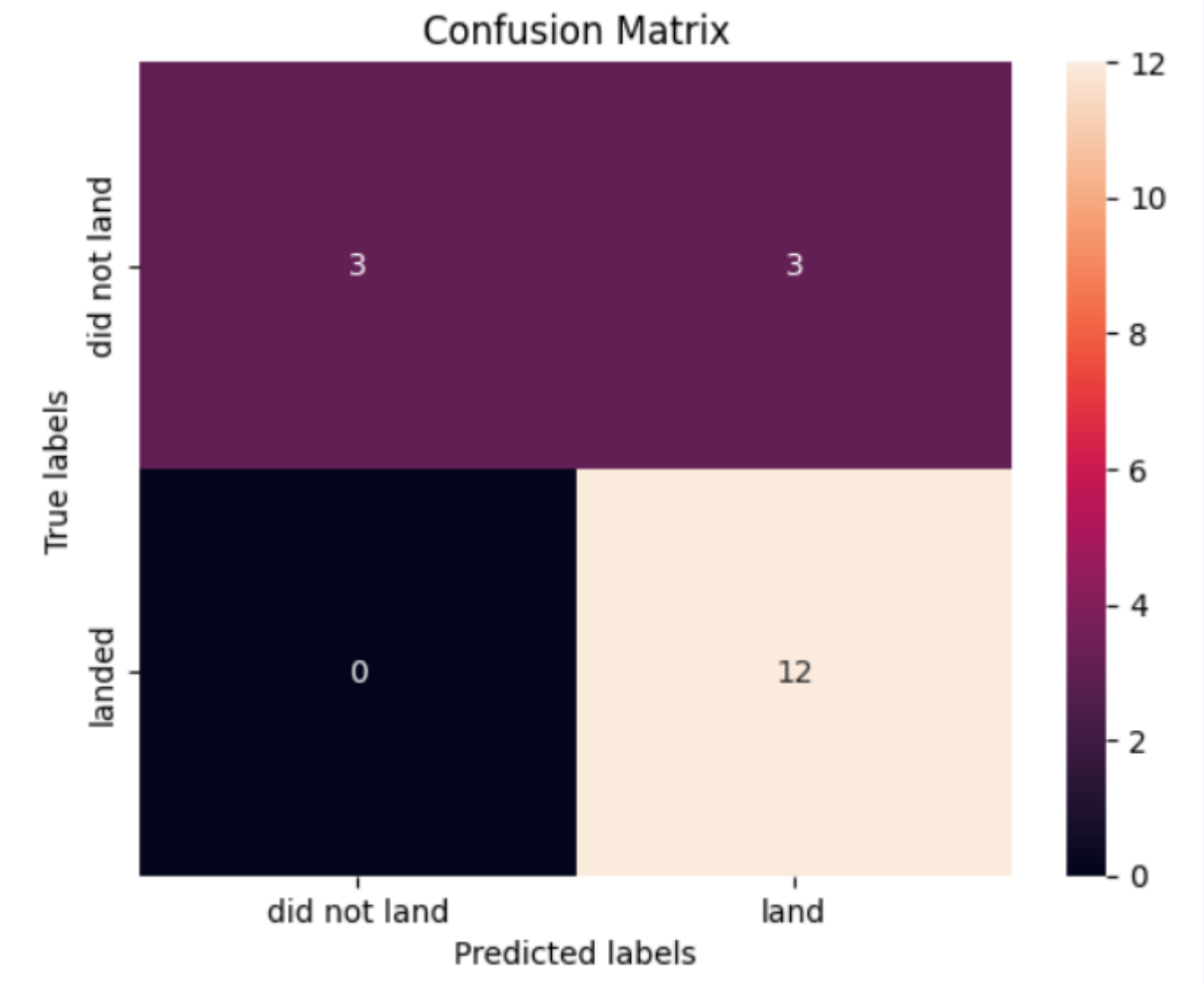
Predictive Analysis (Classification)

Classification Accuracy



From the bar plot, we can observe that all the models have the same classification accuracy.

Confusion Matrix



Since the classification accuracy is the same for all the models, the confusion matrix for all the models are also the same. We observe that for all the models, the model mostly correctly predicts whether the rocket will successfully land, and at times falsely predicts successful landing when landing has failed and vice versa.

Conclusions

- Comprehensive Data Analysis: The project successfully collected and prepared SpaceX launch data from both API endpoints and web scraping, followed by comprehensive exploratory data analysis using visualization and SQL queries.
- Overwhelming Mission Success: SpaceX has achieved an exceptionally high mission success rate, with 100 successful launches compared to just 1 failure in the dataset.
- Factors Influencing Launch Success:
    - Temporal Improvement: The probability of successful first stage landing generally increases with higher flight numbers and shows an overall upward trend in success rate from 2013 to 2020.
    - Launch Site Dependence: Success rates vary significantly by launch site; VAFB SLC-4E and KSC LC-39A demonstrate higher success probabilities compared to other sites, irrespective of payload mass.
    - Orbit Type Impact: Different orbit types exhibit varying success rates, with 100% success for orbits like ES-L1, GEO, HEO, and SSO, and a notable absence of success for SO orbit.
    - Payload Mass Correlation: For heavier payloads, successful landing rates increase for Polar, LEO, and ISS orbits, while for GTO, payload mass does not appear to have a significant impact on success.
    - Booster Version Performance: The Booster version FT is associated with the highest successful launch outcomes.
- Strategic Launch Site Locations: Launch sites are strategically located near coastlines, predominantly on the East Coast, for safety and logistical advantages.
- Effective Predictive Modeling: Classification models, including Logistic Regression, SVM, Decision Tree, and KNN, all achieved a consistent accuracy of approximately 83.33% in predicting Falcon 9 first stage landing outcomes, demonstrating their predictive capability. The models generally predict successful landings correctly.

Thank you!