IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Janani RC
1/8/2024

Outline

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

Executive Summary

- Summary of methodologies
  - ✓ Data Collection
  - ✓ Data Wrangling
  - ✓ Exploratory Data Analysis with SQL
  - ✓ Exploratory Data Analysis with Data Visualization
  - ✓ Interactive map with Folium
  - ✓ Dashboard with Ploty Dash
  - ✓ Machine learning(Predictive analysis)
- Summary of all results
  - ✓ Exploratory Data Analysis results
  - ✓ Interactive analytics demo in screenshot
  - ✓ Predictive analysis results

Introduction

- Project background and context

  SpaceX has revolutionized the commercial space industry by making space travel more affordable. The company offers Falcon 9 rocket launches for $62 million on its website, significantly lower than the over $165 million charged by other providers. This cost reduction is largely due to SpaceX's ability to reuse the first stage of their rockets. By predicting whether the first stage will successfully land, we can estimate the cost of a launch. Utilizing publicly available data and machine learning models, our goal is to forecast the likelihood of SpaceX reusing the first stage.

- Problems you want to find answers

  ✓ How do factors like payload mass, launch site, number of flights, and orbits influence the success of the first stage landing?

  ✓ Has the success rate of first stage landings improved over the years ?

  ✓ What is the most effective algorithm for binary classification in this context?

Section 1

Methodology

Methodology

Executive Summary

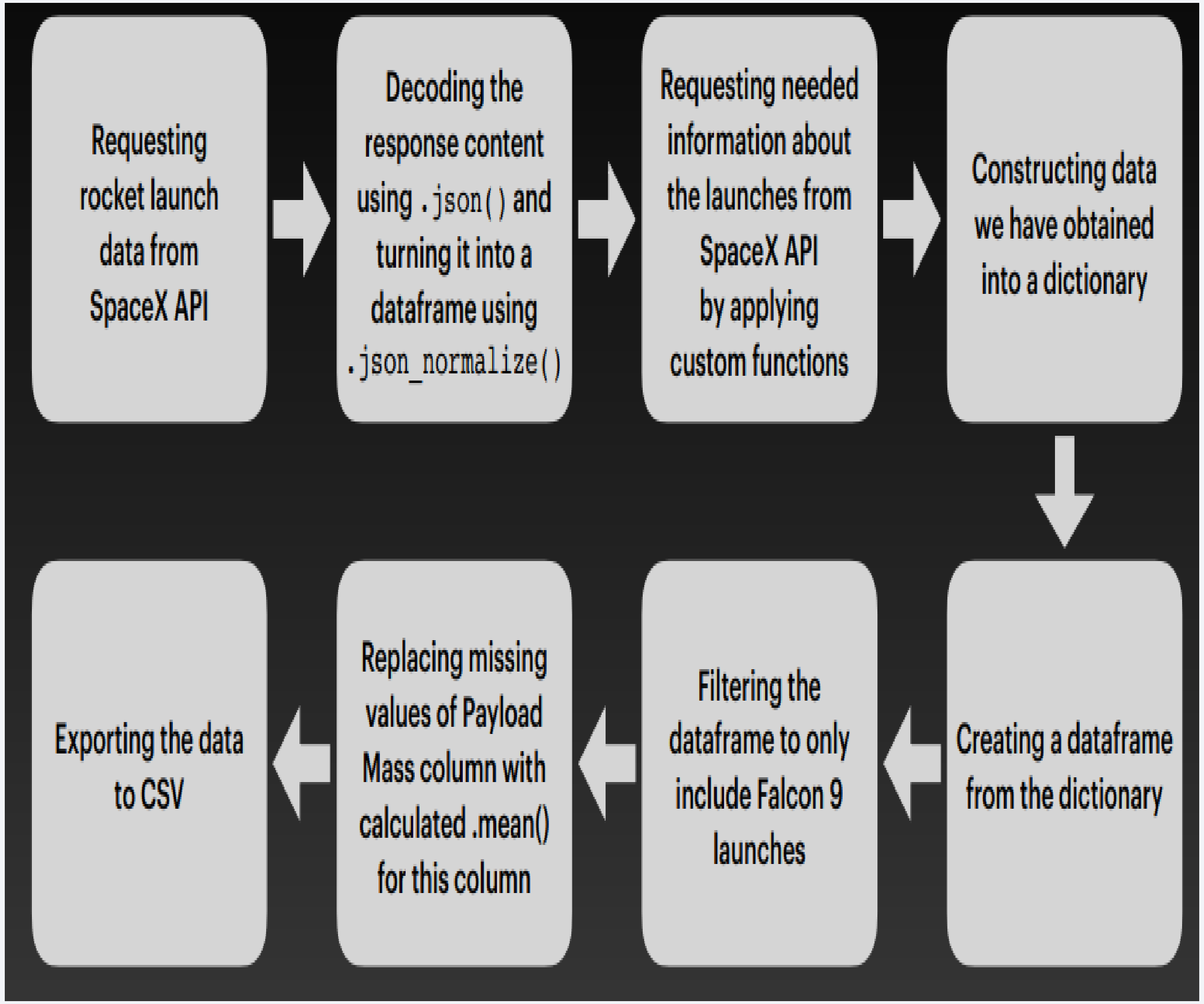
- Data collection methodology:
  - Using Space Rest API and Web scrapping
- Perform data wrangling
  - Filtering the data
  - Finding missing values from table
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build,tuning and evaluating of classification models to ensure the best results had been obtained.

Data Collection

- The data collection process involved a combination of API requests from the SpaceX REST API and web scraping from a table on SpaceX's Wikipedia page to gather comprehensive information about the launches for detailed analysis. Data columns obtained from the SpaceX REST API include: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude. Data columns obtained from Wikipedia web scraping include: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

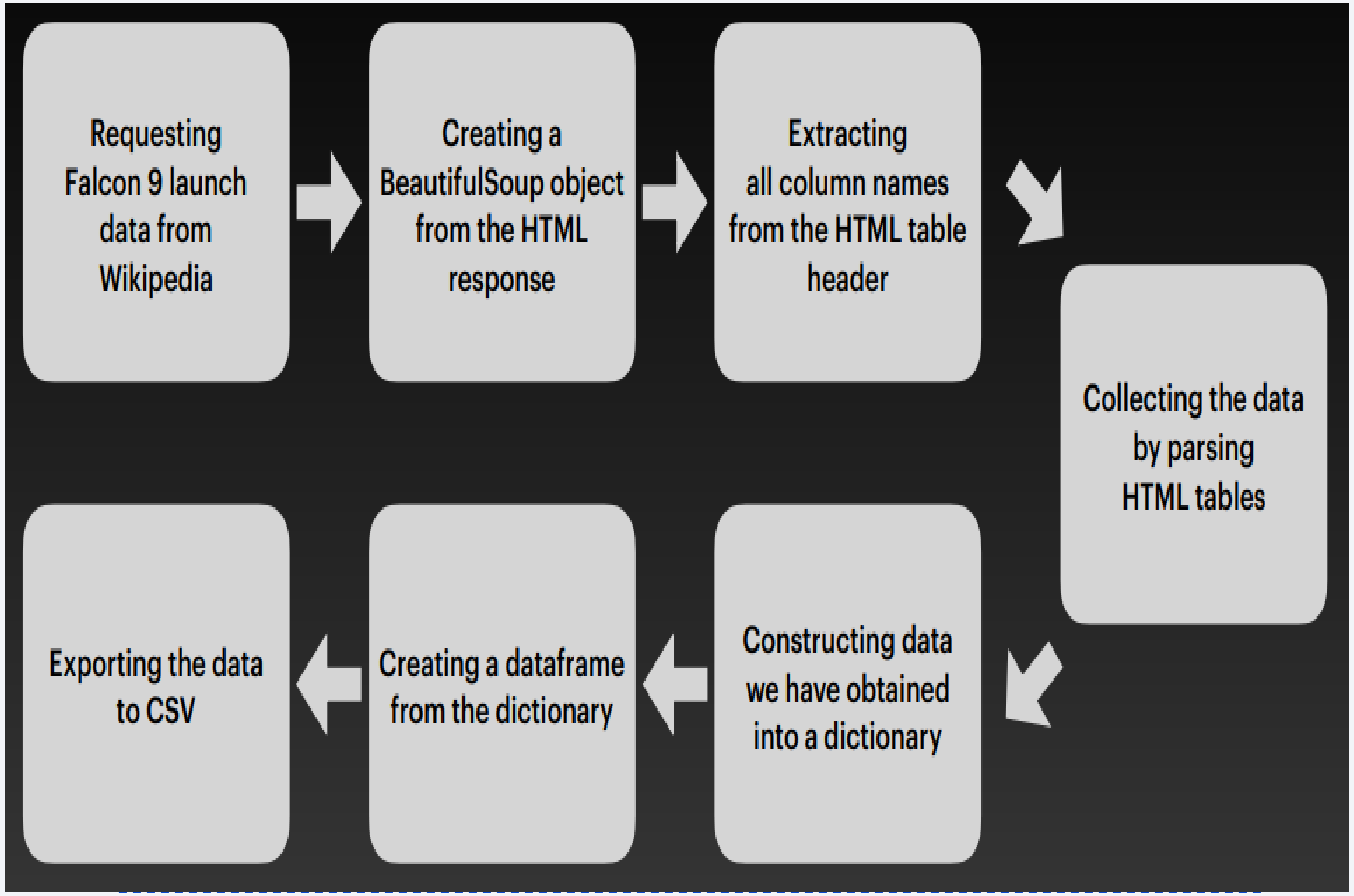
Data Collection – SpaceX API

- https://github.com/Janani-Ironic69/Applied-Data-Science-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



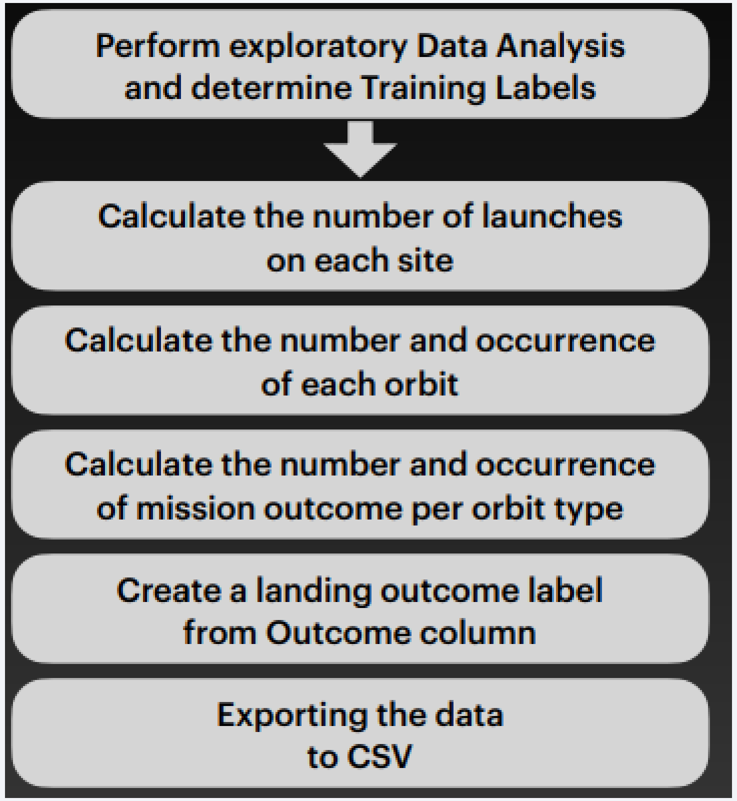
Data Collection - Scraping

https://github.com/Janani-Ironic69/Applied-Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- In the dataset, there are various instances where the booster did not land successfully. Sometimes, a landing was attempted but failed due to an accident. For instance, "True Ocean" indicates the mission outcome was a successful landing in a specific region of the ocean, while "False Ocean" signifies the mission outcome was an unsuccessful landing in the same region. Similarly, "True RTLS" denotes the mission outcome was a successful landing on a ground pad, whereas "False RTLS" indicates an unsuccessful landing on a ground pad. "True ASDS" means the mission outcome was a successful landing on a drone ship, and "False ASDS" means the mission outcome was an unsuccessful landing on a drone ship. We primarily convert these outcomes into training labels, where "1" represents a successful booster landing, and "0" indicates an unsuccessful attempt.



https://github.com/Janani-Ironic69/Applied-Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts were created as follows: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, and Success Rate Yearly Trend.
- Scatter plots were used to display the relationships between variables. If there is a relationship, these plots could be useful for machine learning models.
- Bar charts were used to compare different discrete categories, aiming to illustrate the relationship between the categories being compared and a measured value.
- Line charts were used to depict trends in data over time (time series).

https://github.com/Janani-Ironic69/Applied-Data-Science-Capstone-Project/blob/main/edadataviz.ipynb

EDA with SQL

- SQL queries were executed to:
- Display the names of the distinct launch sites used in the space missions.
- Show 5 records where the launch sites start with the string 'CCA'.
- Show the total payload mass carried by boosters launched by NASA (CRS).
- Show the average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing on a ground pad was achieved.
- List the names of boosters that have successfully landed on a drone ship and carried a payload mass greater than 4000 but less than 6000.
- Show the total number of successful and failed mission outcomes.
- List the names of the booster versions that have carried the maximum payload mass.
- Show the failed landing outcomes on drone ships, their booster versions, and launch site names for the months in the year 2015.
- Rank the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.

https://github.com/Janani-Ironic69/Applied-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers for all Launch Sites:
- A marker was added with a circle, popup label, and text label for NASA Johnson Space Center, using its latitude and longitude coordinates as the starting location.
- Markers with circles, popup labels, and text labels were added for all launch sites, using their latitude and longitude coordinates to display their geographical locations and proximity to the equator and coasts.
- Colored Markers for Launch Outcomes at Each Launch Site:
- Colored markers were added to indicate successful (green) and failed (red) launches, using marker clusters to identify which launch sites have higher success rates.
- Distances Between a Launch Site and Its Proximities:
- Colored lines were added to show the distances between the launch site KSC LC-39A (as an example) and its proximities, such as railway, highway, coastline, and the closest city.

https://github.com/Janani-Ironic69/Applied-Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
- Included a dropdown list to allow the selection of Launch Sites.
- Pie Chart Displaying Successful Launches (All Sites/Specific Site):
- Added a pie chart to visualize the total number of successful launches across all sites and the Success vs. Failure counts for a selected site, if applicable.
- Payload Mass Range Slider:
- Implemented a slider to choose the range of Payload Mass.
- Scatter Chart of Payload Mass vs. Success Rate for Different Booster Versions:
- Added a scatter chart to illustrate the relationship between Payload Mass and Launch Success Rate.

https://github.com/Janani-Ironic69/Applied-Data-Science-Capstone-Project/blob/main/spacex_dash_app.py

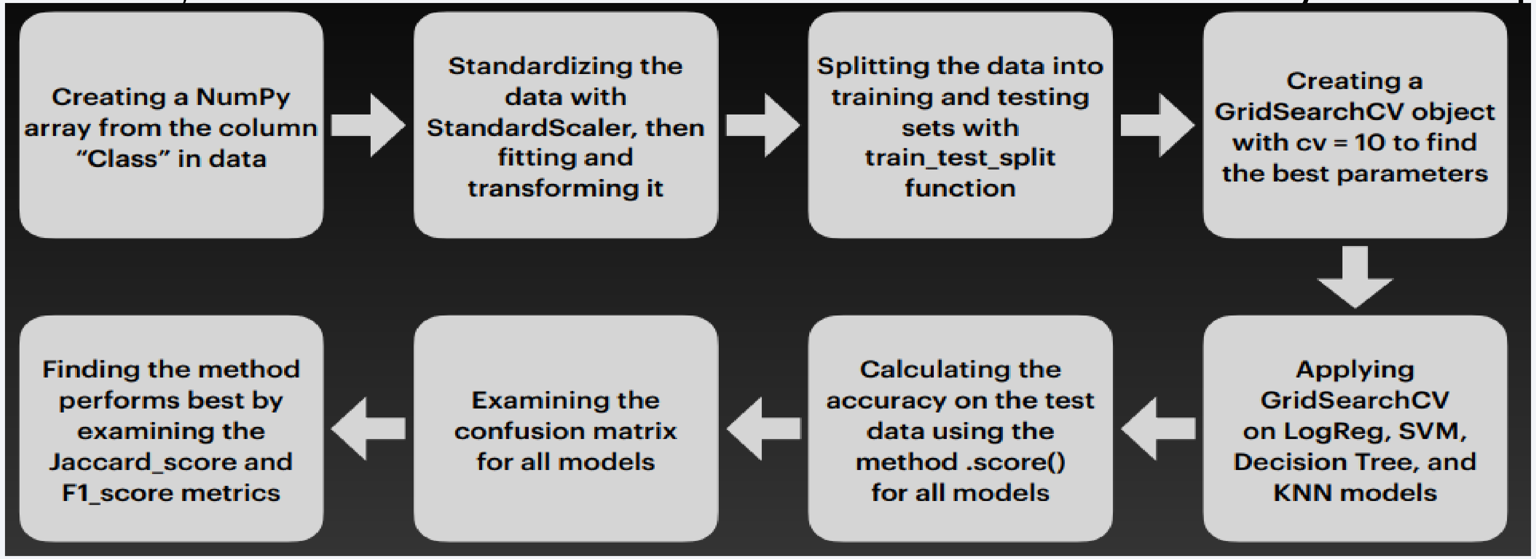
Predictive Analysis (Classification)

1. Data Preparation: Extract the "Class" column as a NumPy array, then standardize the data using StandardScaler.

2. Data Splitting: Divide the dataset into training and testing sets using train_test_split.

3. Hyperparameter Tuning: Use GridSearchCV with 10-fold cross-validation to optimize parameters for models including Logistic Regression, SVM, Decision Tree, and KNN.

4. Model Evaluation: Assess model performance on the test set using accuracy, confusion matrices, and metrics like Jaccard score and F1 score to identify the best-performing model.



https://github.com/Janani-Ironic69/Applied-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%5B1%5D.ipynb

Results

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

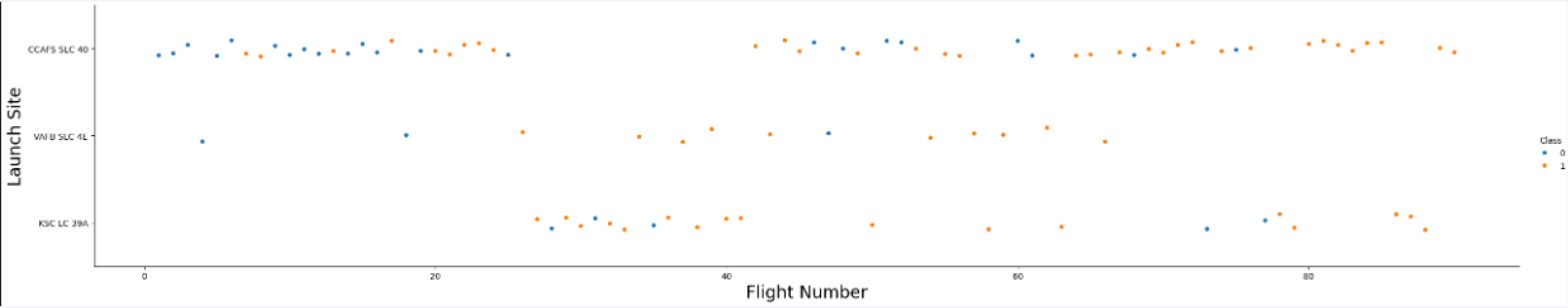
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Explaination:

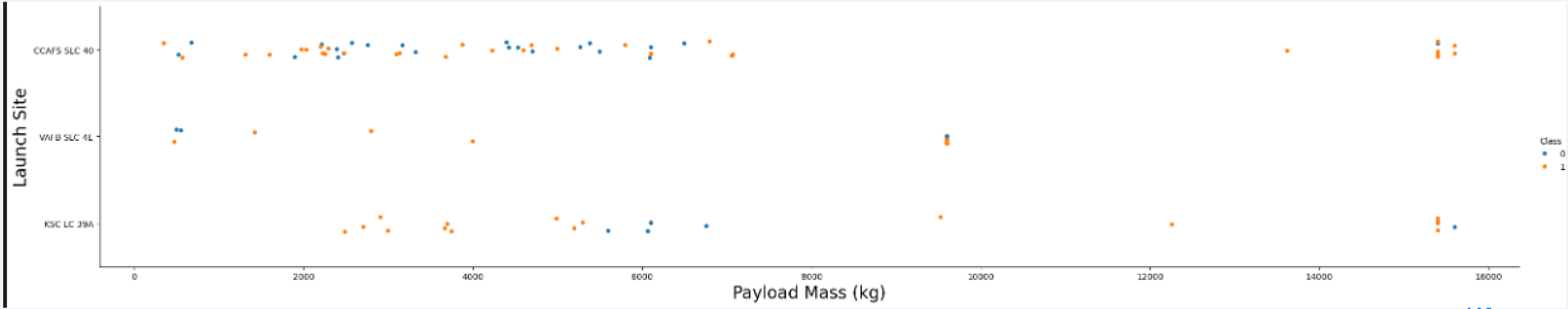
- Initially, flights were unsuccessful, but recent launches have consistently been successful.
- The CCAFS SLC 40 launch site accounts for approximately half of all launches.
- Both VAFB SLC 4E and KSC LC 39A exhibit higher success rates.
- It's likely that the probability of success increases with each new launch.



Payload vs. Launch Site

Explaination:

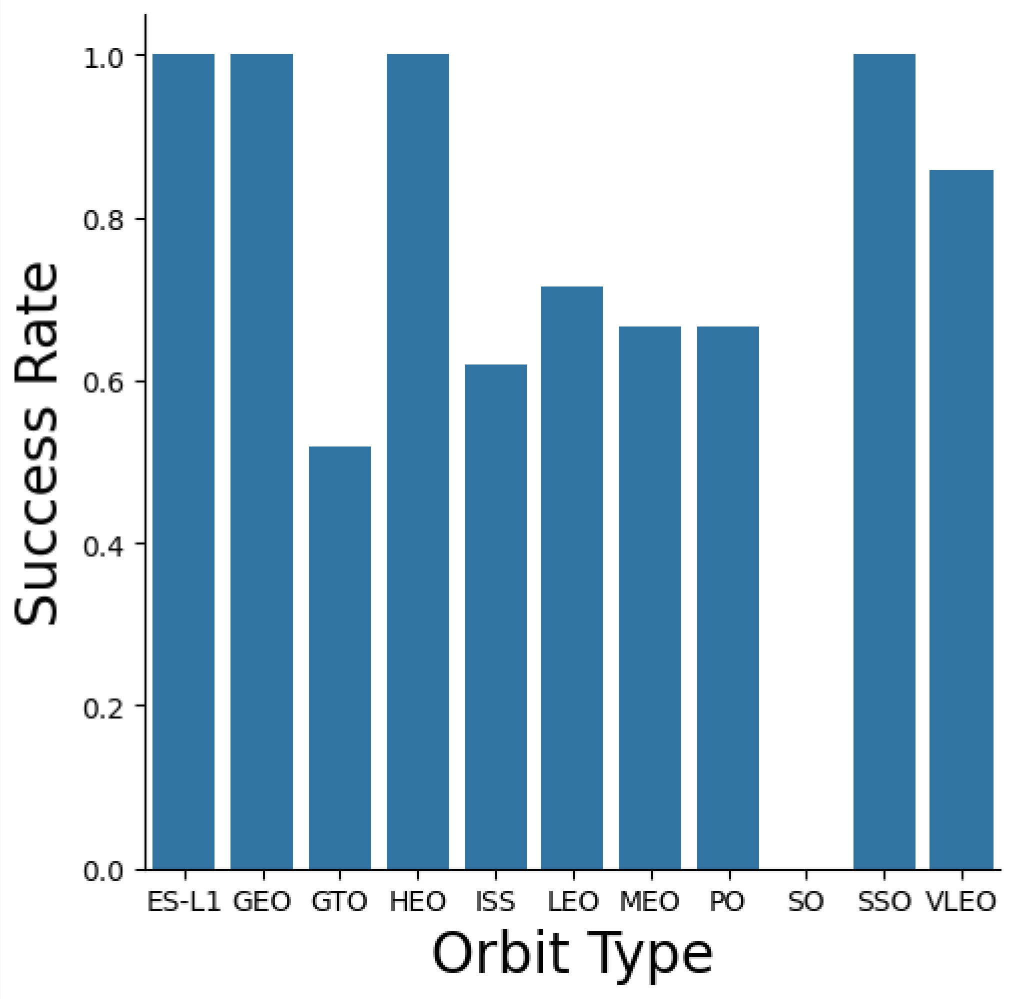
- A greater payload mass at any launch site correlates with an increased success rate.
- Launches carrying payloads exceeding 7000 kg generally achieved success.
- The KSC LC 39A site boasts a perfect success rate for payloads weighing under 5500 kg as well.



Success Rate vs. Orbit Type

Explaination:

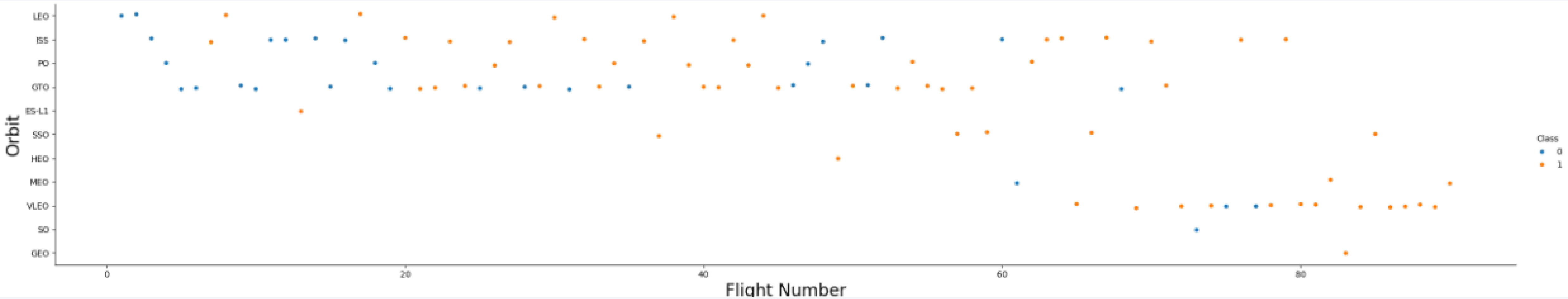
- Orbits with complete success rates include ES-L1, GEO, HEO, and SSO.
- The SO orbit has a 0% success rate.
- Orbits with success rates ranging from 50% to 85% are GTO, ISS, LEO, MEO, and PO.



Flight Number vs. Orbit Type

Explaination:

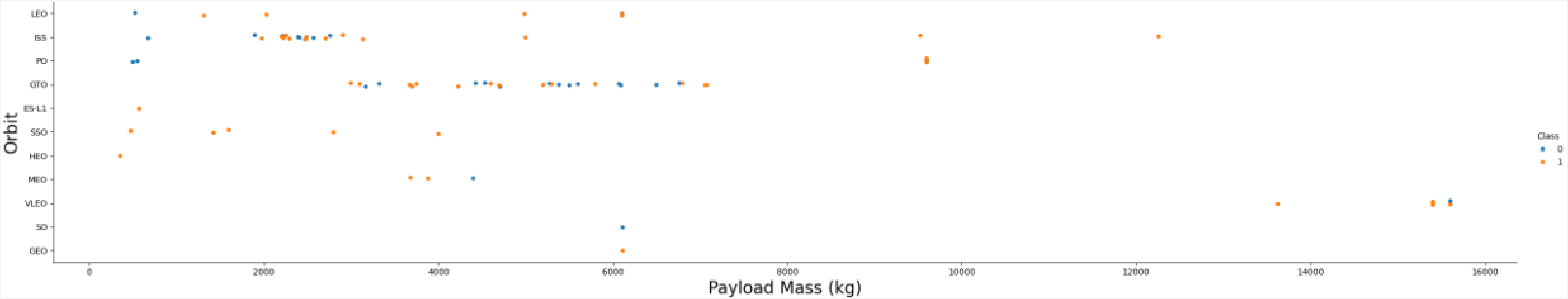
- In the LEO orbit, success seems to be linked to the number of flights.
- Conversely, in the GTO orbit, there appears to be no correlation between the number of flights and success.



Payload vs. Orbit Type

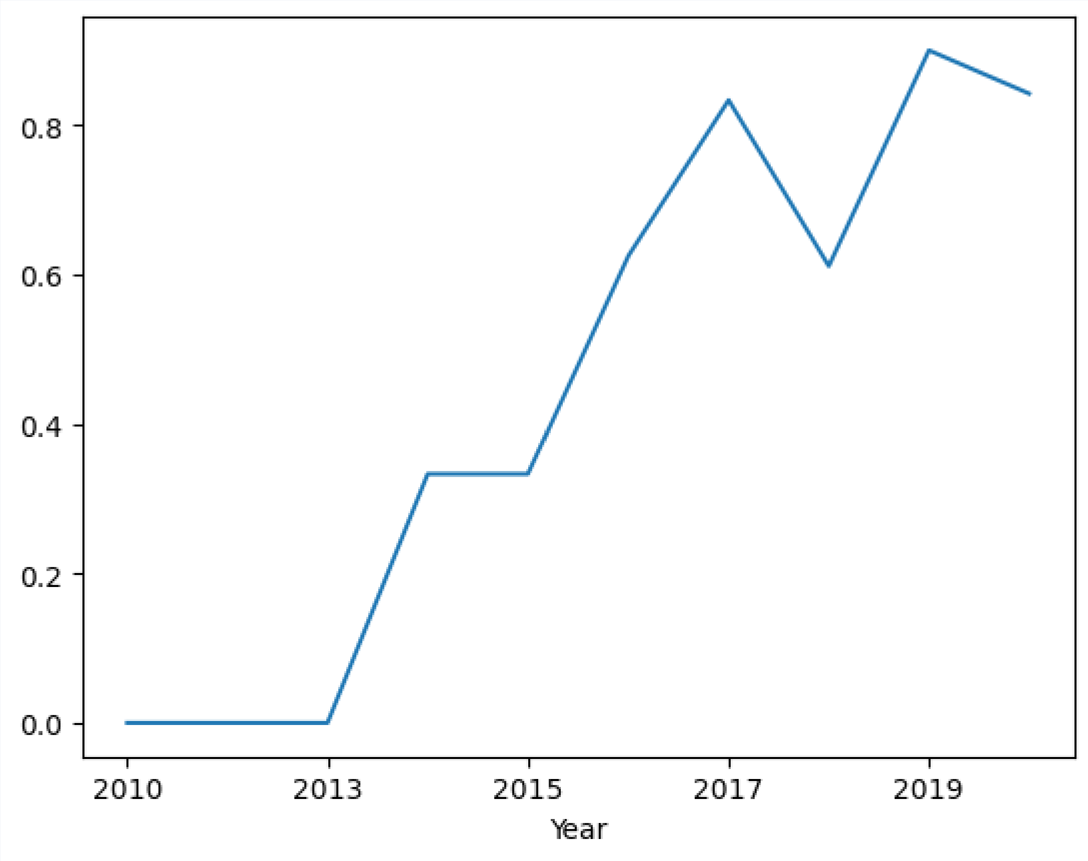
Explanation:

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- The success rate start increasing from 2013 and kept increasing ever since.



All Launch Site Names

- Represented all the unique launch site name from the spacex table

In [12]:
```
%sql select distinct launch_site from SPACEXTABLE;
```

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

Out[12]:

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

Launch Site Names Begin with 'CCA'

- Displayed 5 records where launch site began with the string like ‘CCA’

In [13]:
```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

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

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lar
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Fai
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	Fai
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Displayed total payload mass where boosters were launched by NASA

In [14]:
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

Out[14]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- Displayed the first successful landing date from the table whose outcome is success and retrieved the lowest date using min function.

```
In [17]: %sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';

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

Out[17]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed all the booster version which have successful landed in drone ship with payload in between 4000 and 6000

In [19]:

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

Out[19]:

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

Total Number of Successful and Failure Mission Outcomes

- Displayed the total number of successful and failure mission

```
In [20]: %sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;

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

Out[20]:

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

Boosters Carried Maximum Payload

- Displayed the booster version which carried the maximum load using a sub query which retrieved the max payload first.

In [22]:
```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

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

Out[22]:

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

- Listed all failed landing outcomes in drone ship,their booster version along with launch site for the months in year 2015

In [29]:
```
%%sql select substr(6,2,date) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
    where landing_outcome = 'Failure (drone ship)' and substr(date,0,5)='2015';
```

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

Out[29]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome
	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Ranking 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.

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

Landing_Outcome	count_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

Section 3

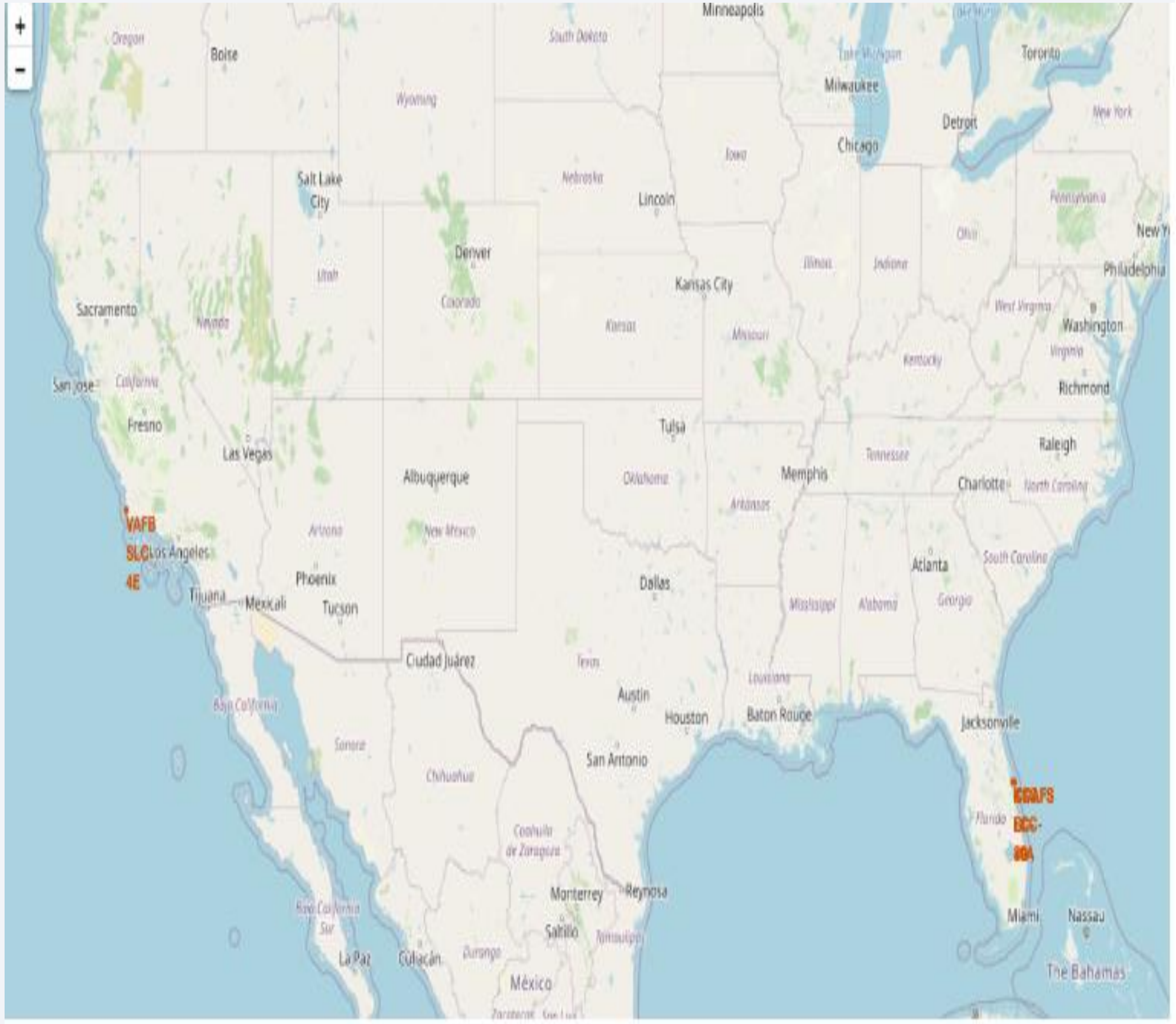
Launch Sites Proximities Analysis

Location of all launch sites on a world map

Explaination:-

•The majority of launch sites are located near the Equator. At the Equator, the Earth's surface moves faster than at any other latitude, with a speed of 1670 km/hour. This rotational velocity provides an initial boost due to inertia when a spacecraft is launched, helping it maintain the necessary speed to stay in orbit.

•All launch sites are situated close to coastal areas. Launching rockets towards the ocean minimizes the risk of debris or explosions impacting populated areas, enhancing safety.



Color labelling success and failure records on the map

Explaination:

- Color-coded markers are used to indicate the success rates of launch sites.
  - **Green Marker**: Represents a successful launch.
  - **Red Marker**: Represents a failed launch.

Geographical layout and proximity analysis of CCAFS SLC-40 and LC-40 Launch sites

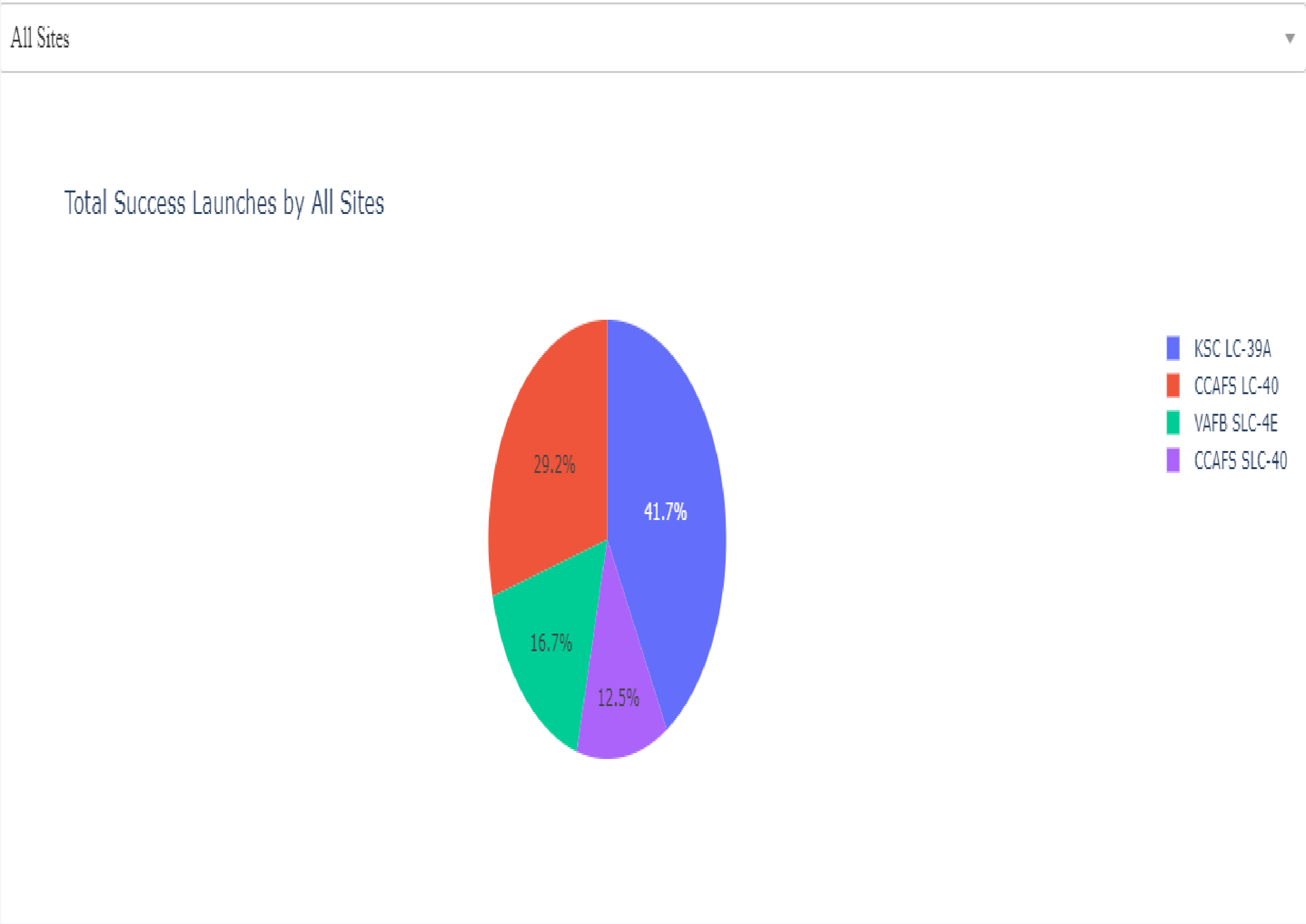
- The image shows a map highlighting the positions and proximities of two launch sites labeled CCAFS SLC-40 and LC-40. The following details are noted:
- **Green Marker (7)**: Likely represents successful launches from the CCAFS SLC-40 site.
- **Yellow Marker (26)**: Possibly indicates the total number of launches, with the difference in color potentially denoting different outcomes or statuses (such as successful or failed launches).
- The map also shows the locations relative to major nearby infrastructure:
- **Proximity to the Coast**: The coastline is marked 0.90 km from the identified locations.
- **Nearby Roadways**: Samuel C Phillips Pkwy is indicated nearby, suggesting accessibility and logistical considerations.
- The circles around the launch sites may represent zones of interest or safety areas, although their specific meanings aren't explicitly detailed in the image alone. The map provides a visual representation of the geographical layout and distances, useful for understanding the logistics and safety implications related to the launch activities at these sites.

Section 4

Build a Dashboard with Plotly Dash

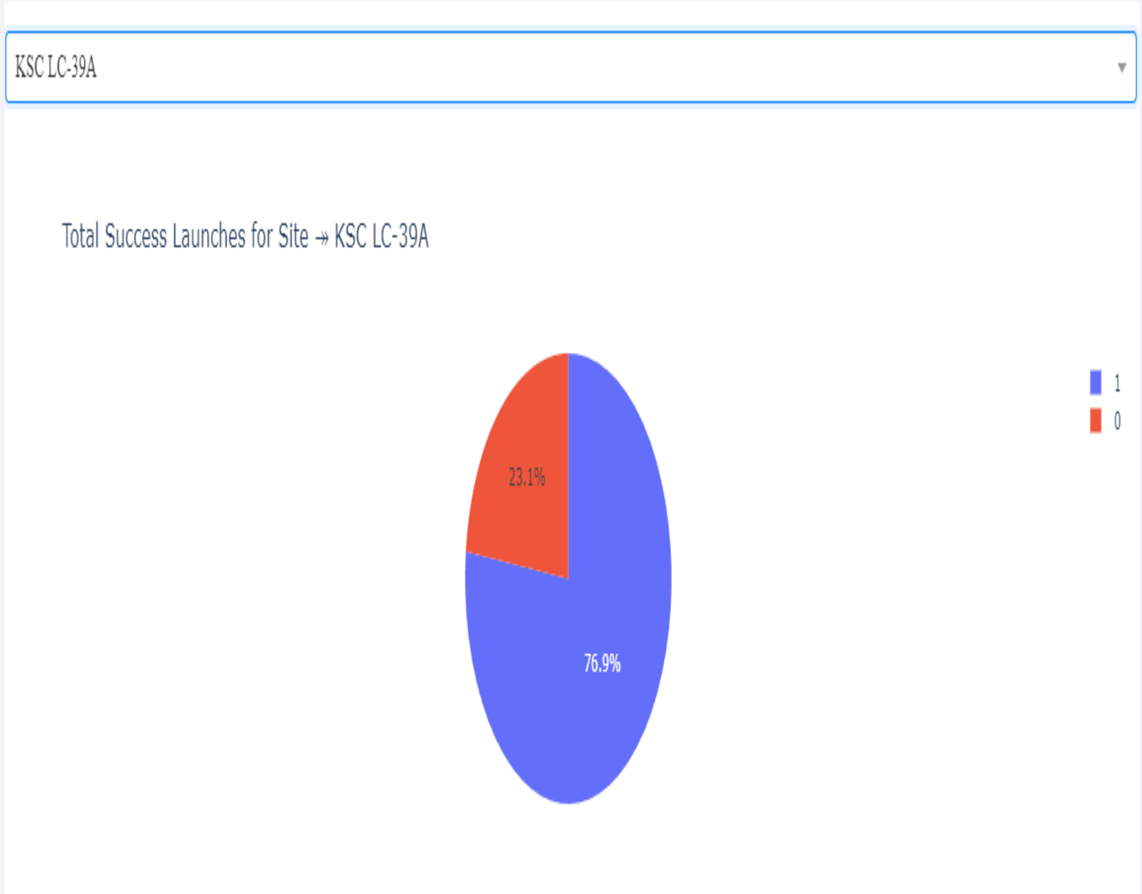
Total success launches by all sites

The pie chart shows the percentage distribution of total successful launches across four sites: KSC LC-39A (41.7%), CCAFS LC-40 (29.2%), VAFB SLC-4E (16.7%), and CCAFS SLC-40 (12.5%). The largest share of successful launches is attributed to KSC LC-39A.



KSC LC-39A pie chart i.e having highest launch rate

- KSC LC-39A has the highest launch success rate with 76.9 %



Payload VS Launch outcome

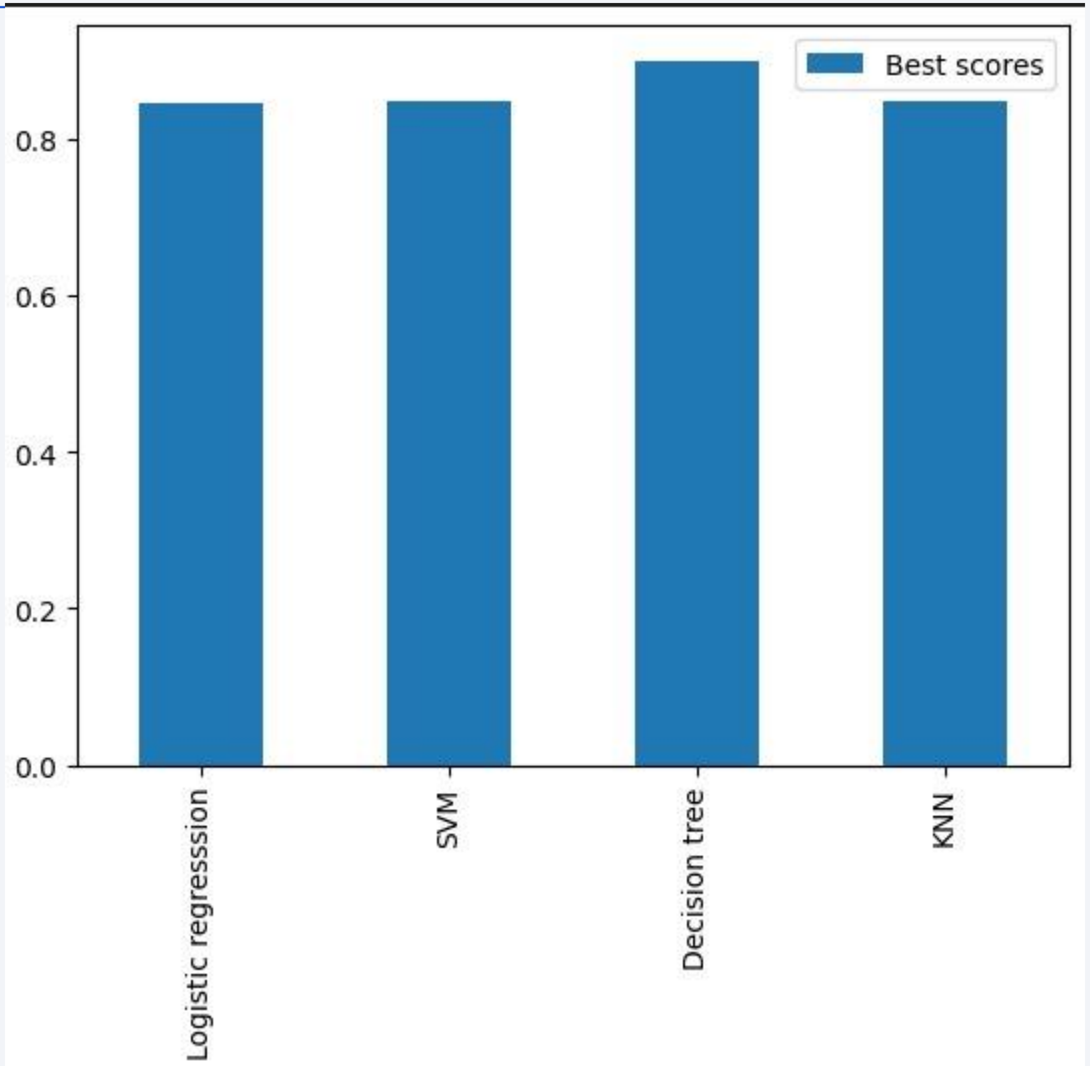
The chart shows the success rate compared to payload mass FT has high success rate with payload mass below 6000kg

Section 5

Predictive Analysis (Classification)

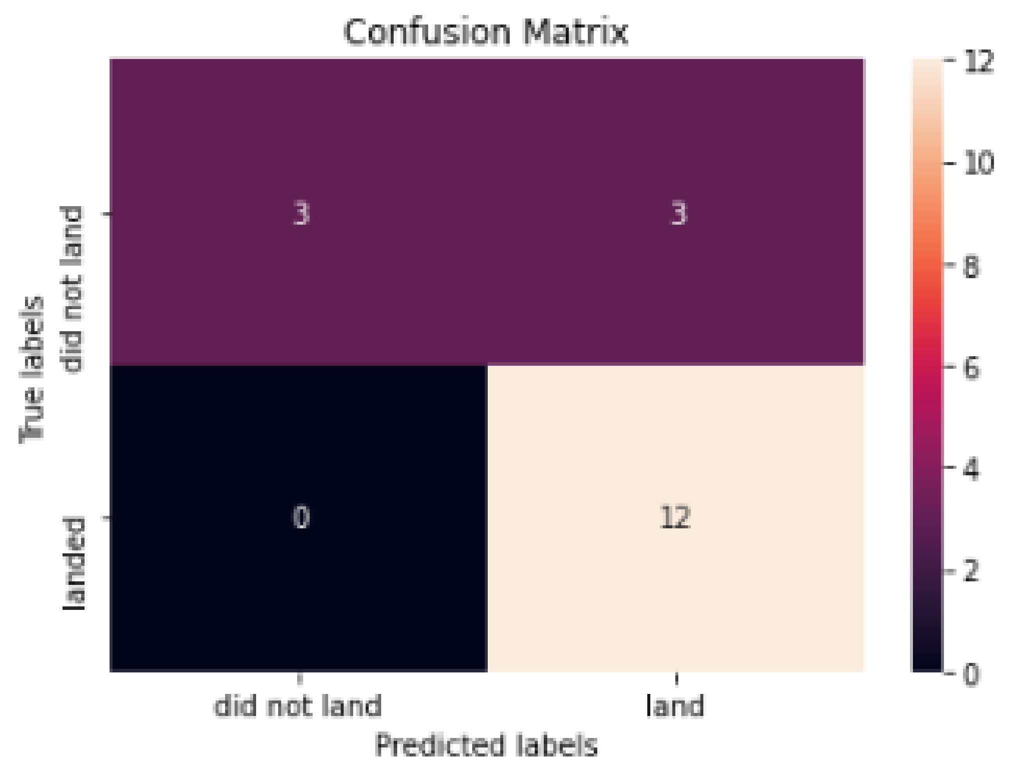
Classification Accuracy

- Visualized all the models via bar graph and took the observation that Decision tree has the highest accuracy



Confusion Matrix

- The confusion matrix of decision tree shows the highest accuracy with only problem as false positives.



Conclusions

- The Decision Tree Model has proven to be the most effective algorithm for this dataset.
- Launches carrying lighter payloads tend to yield better outcomes compared to those with heavier payloads.
- Most launch sites are located near the Equator and are all situated close to the coast.
- The success rate of launches has shown an upward trend over the years.
- KSC LC-39A boasts the highest success rate among all launch sites.
- The orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.

Thank you!