IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nischala Nagisetty
Jan 19,2026

Outline

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

Executive Summary

- SpaceX is widely regarded as one of the top aerospace companies in the world due to major advancements in rocket reusability and launch efficiency. While traditional space companies charge around 165 million dollars per launch, SpaceX provides the same service for approximately 62 million dollars, which has attracted major clients such as NASA.
- This report assumes the role of a data scientist working for a new competitor named SpaceY that aims to enter the commercial orbital launch market founded by Elon Musk. The company seeks to evaluate whether it can realistically compete with SpaceX and gain a share of the market using data-driven insights.
- The analysis focuses on collecting information about SpaceX, conducting exploratory data analytics, building predictive machine learning models, and developing visual dashboards to support technical and business decision making for the SpaceY leadership team.

Introduction

- Since 1957, countries have competed to expand beyond Earth through satellite launches and space exploration, but these efforts have required extremely high capital where an average orbital rocket launch costs around 165 million dollars. SpaceX has changed this landscape by reducing launch costs to approximately 60 million dollars through advanced reusable rocket technology that allows the first stage to safely return.
- The major driver behind this cost reduction is SpaceX's ability to successfully land the first stage of the rocket, allowing it to be reused instead of discarded. Understanding which factors contribute to landing success is critical for any competitor aiming to enter the market.
- This report examines the attributes that influence landing outcomes, including payload mass, launch site, mission orbit, and operational characteristics, by applying data science techniques such as exploratory analysis, visualization, and machine learning using the IBM Data Science methodology.

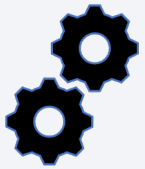
Section 1

Methodology

Methodology

1. Data collection methodology

The dataset was obtained using the SpaceX REST API and web scraping from Wikipedia pages.



2. Data wrangling

Preprocessing was performed using pandas and NumPy, including one-hot encoding, column filtering, normalization, and standardization.

3. Exploratory data analysis (EDA)

Visualization and data querying were conducted using seaborn, matplotlib, and SQL.

4. Interactive visual analytics

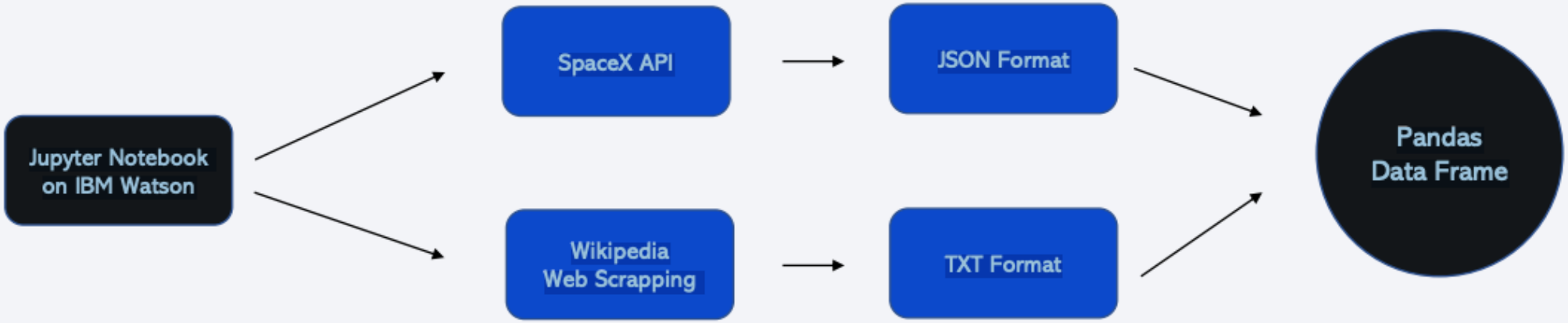
Interactive visualizations were developed using Folium and Plotly Dash.

5. Predictive analysis using classification models

- Split the dataset into train and test sets
- Performed hyperparameter tuning using Grid Search
- Selected the optimal model for deployment

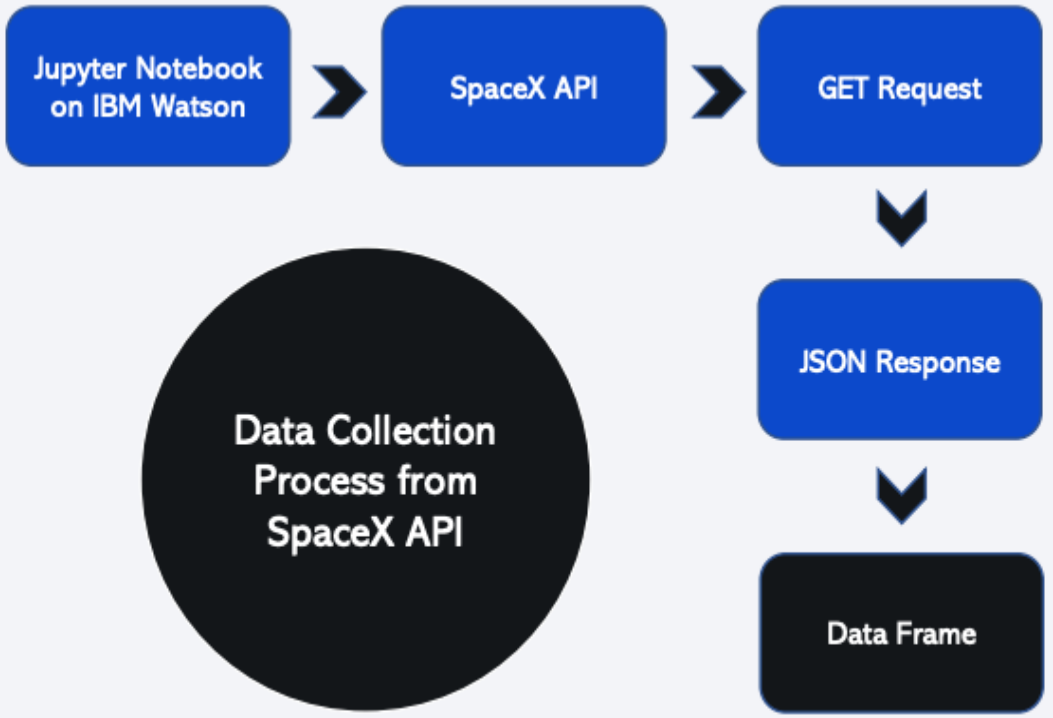
Data Collection

- We have collected the data from two main sources:
  - SpaceX API: Open Source REST API for launch, rocket, core, capsule, starlink, launchpad, and landing pad data.
  - Wikipedia: is a free online encyclopedia, created and edited by volunteers around the world and hosted by the Wikimedia Foundation



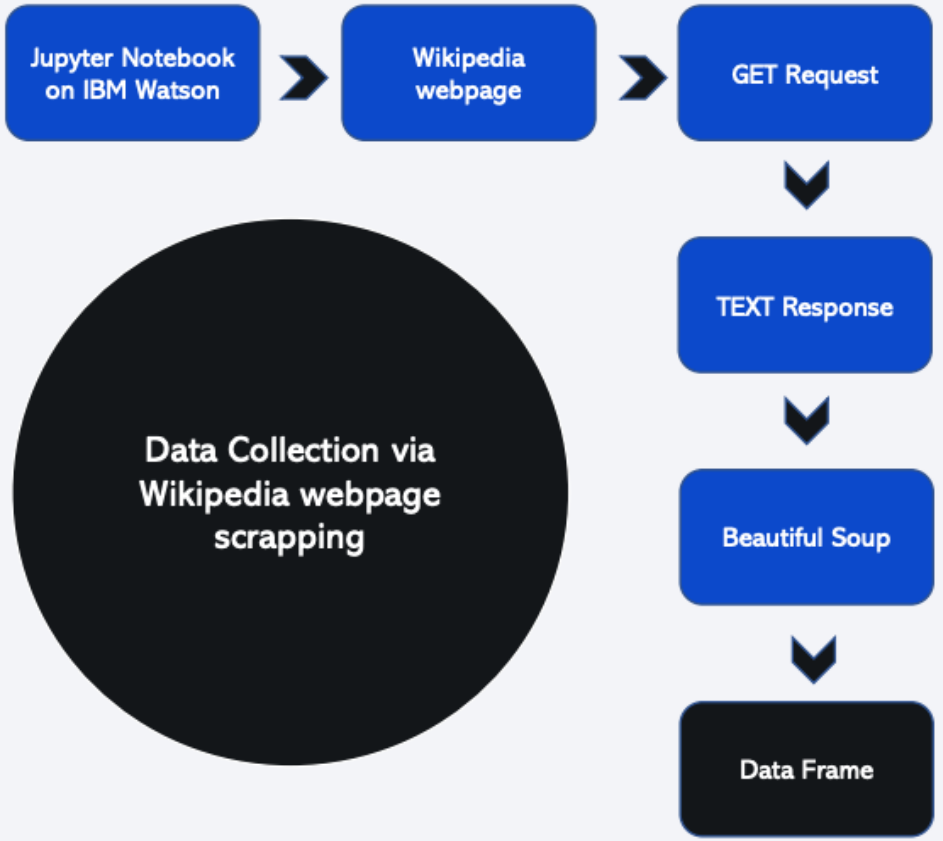
Data Collection – SpaceX API

- We started the data collection from SpaceX API by importing the required libraries such as pandas, NumPy and Request, then we established a URL GET request, this request is raised as JSON file to be finally converted to a data frame through choosing the required information like the geospatial info, rocket type, orbit, flight number and more.
- GitHub URL of the completed SpaceX API calls notebook here



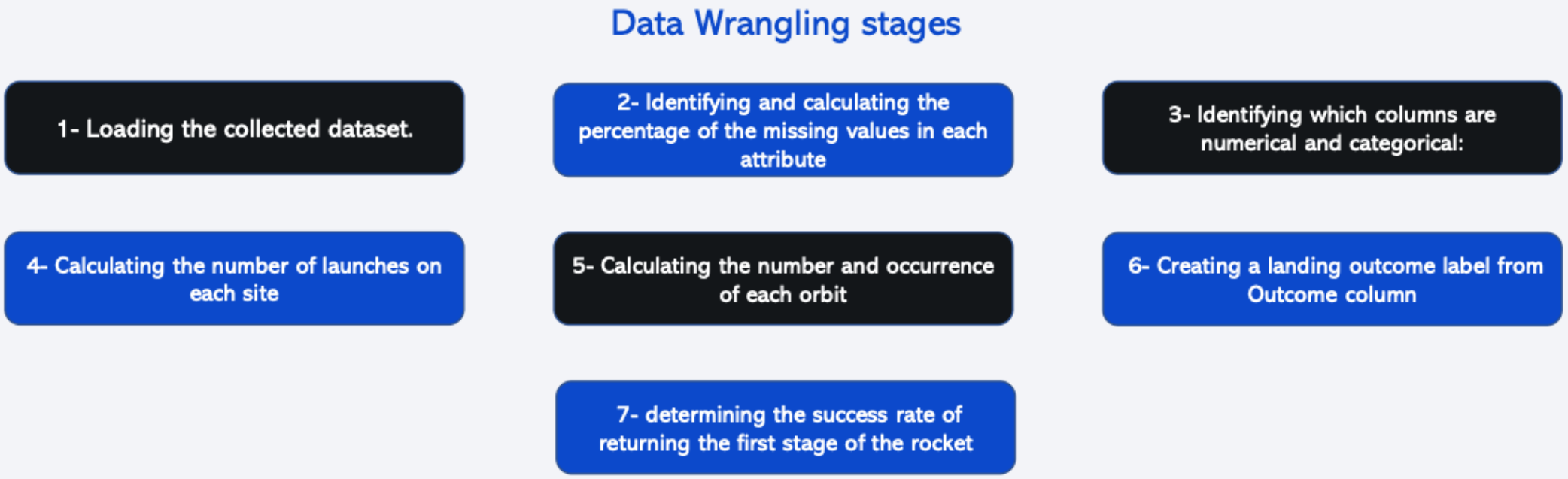
Data Collection - Scraping

- As we have done before we start by importing the required Python libraries beautiful soup and request to perform our task, and this time, we have used a webpage on Wikipedia called “Space X Falcon 9 First Stage Landing Prediction” as a data source, then we initialized an HTTP Get Request and the response was as a text format, then we used the beautiful soup library to extract the tables and columns effectively from the text response to be converted later to a pandas' data frame.
- GitHub URL of the completed web scraping task notebook here



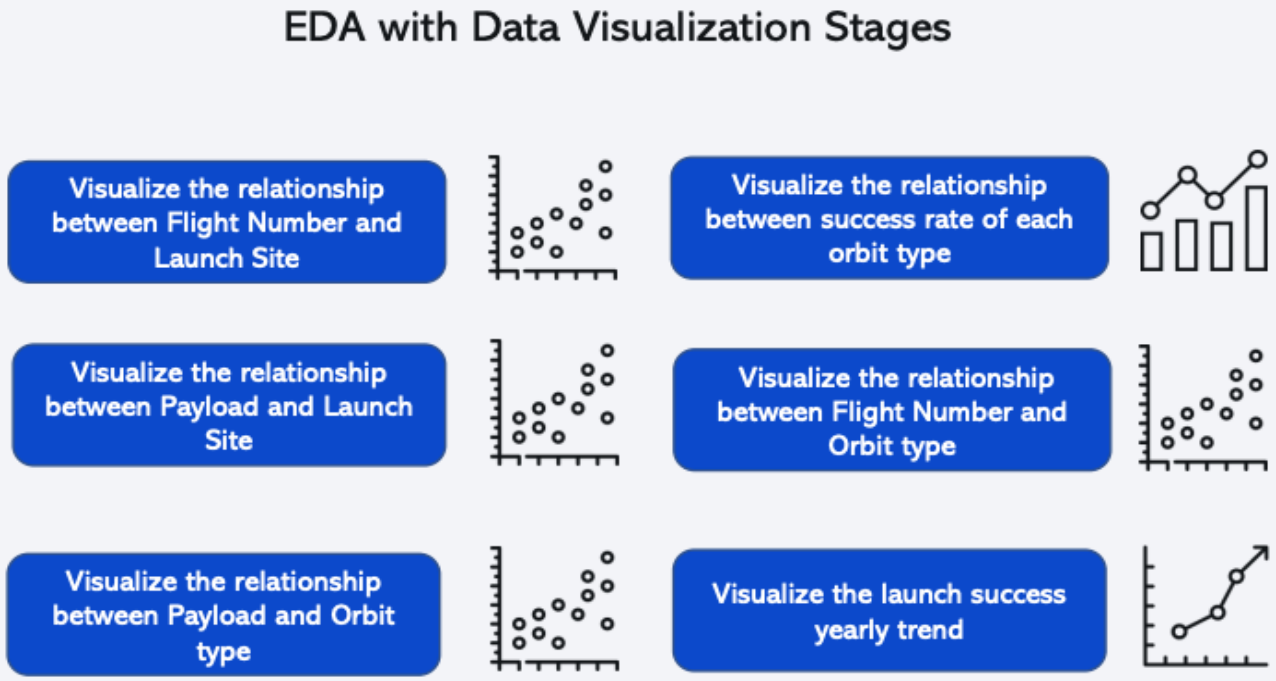
Data Wrangling

- In this stage we started by importing pandas and NumPy, loading our collected data in the previous stage to perform our exploratory data analysis which aimed to clean the data and choose the valid features for training a machine learning model.
- GitHub URL of the completed Data Wrangling Notebook here



EDA with Data Visualization

- In this stage we completed our EDA process through finding the correlation between the features and the target using different visualization tools via seaborn and matplotlib furthermore we have performed feature engineering by converting categorical features into dummy values.
- GitHub URL of the completed EDA with data visualization notebook [here]



EDA with SQL

- Summarized on the right is SQL used to complete the EDA on this dataset
- The GitHub URL of the completed EDA with SQL notebook is here

- 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).
- 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. Use a subquery.
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for the 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.

Build an Interactive Map with Folium

- In this stage we used folium library to represent our work as geospatial data by drawing markers circles and lines on an interactive map.
- GitHub URL of the completed interactive map with Folium map notebook, Here

We started our interactive map by drawing 4 circles on 4 different sites belongs to Falcon 9 rockets lunches have the following information:

Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610746

- We put markers to on the same sites to represent the successful/failed first stage of rockets return using marker objects
- Finally, we calculated the distances between the launch site (CCAFS LC40) to its proximities 1-the closest city, 2-coastline, and 3-highway. Then we drew polylines to represent these distances using PolyLine object.

Build a Dashboard with Plotly Dash

1- We added a dropdown list to enable Launch Site selection including the following options:

All Sites, CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A

2- we added a pie chart to show the total successful launches count for all sites

3- we added a slider to select payload which ranges from 0 -10000

4- finally we added a scatter chart to show the correlation between payload and launch success

- GitHub URL of the completed Building an Interactive Dashboard with Plotly Dash notebook, Here

Predictive Analysis (Classification)

Machine Learning Steps

1- Importing the required libraries.

2- Loading the cleaned data.

3- Standardizing the data to prevent the bias.

4- splitting the data into 20% for testing data and 80% training data.

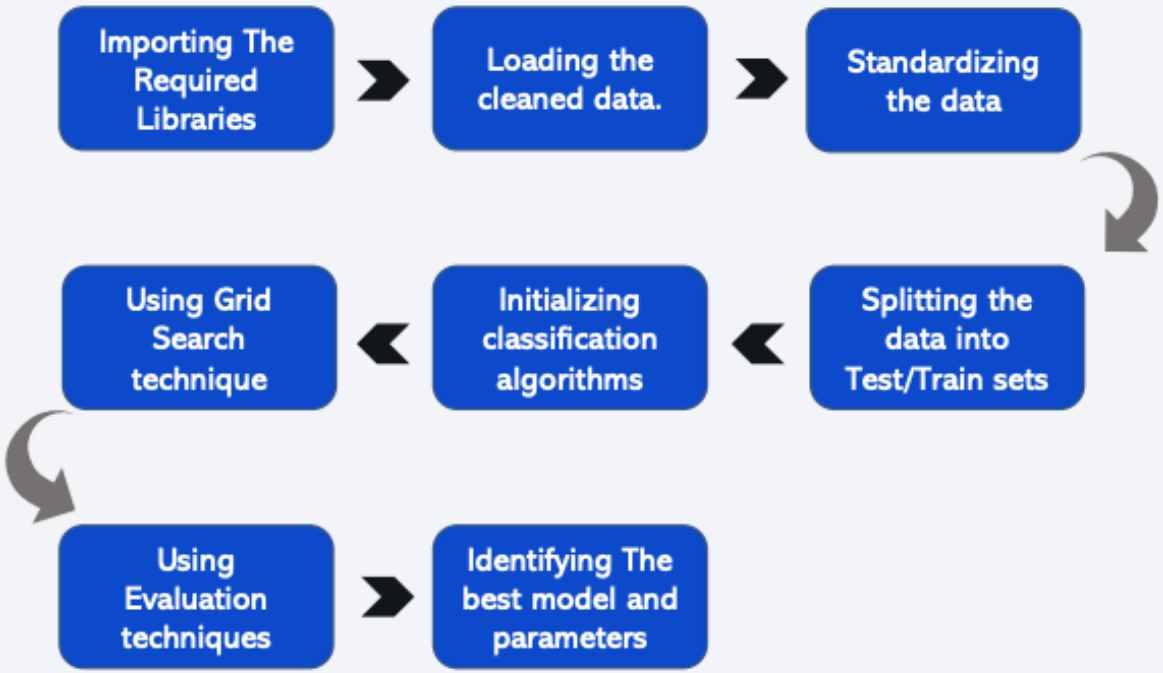
5- Initializing 4 different classification algorithms:

- o Logistic Regression (LR)
- o Support Vector Machine (SVM)
- o Decision Tree (DT)
- o K nearest neighbors (KNN)

6- Using Grid Search technique to find the best parameters

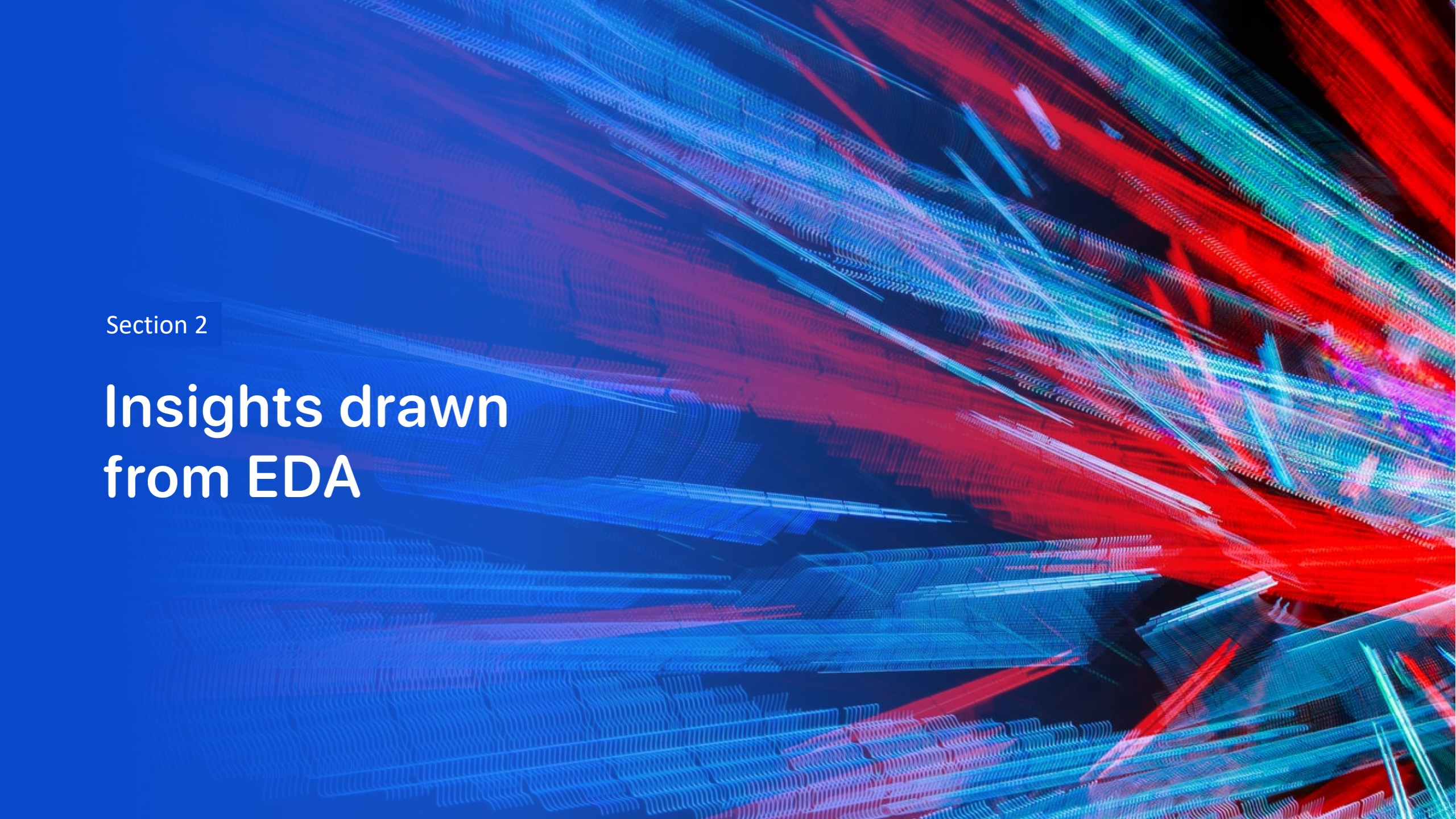
7- Using Evaluation techniques including, Confusion matrix , F1 score, Jaccard Score for the purpose of using the best model among the algorithms above.

Machine Learning Pipeline

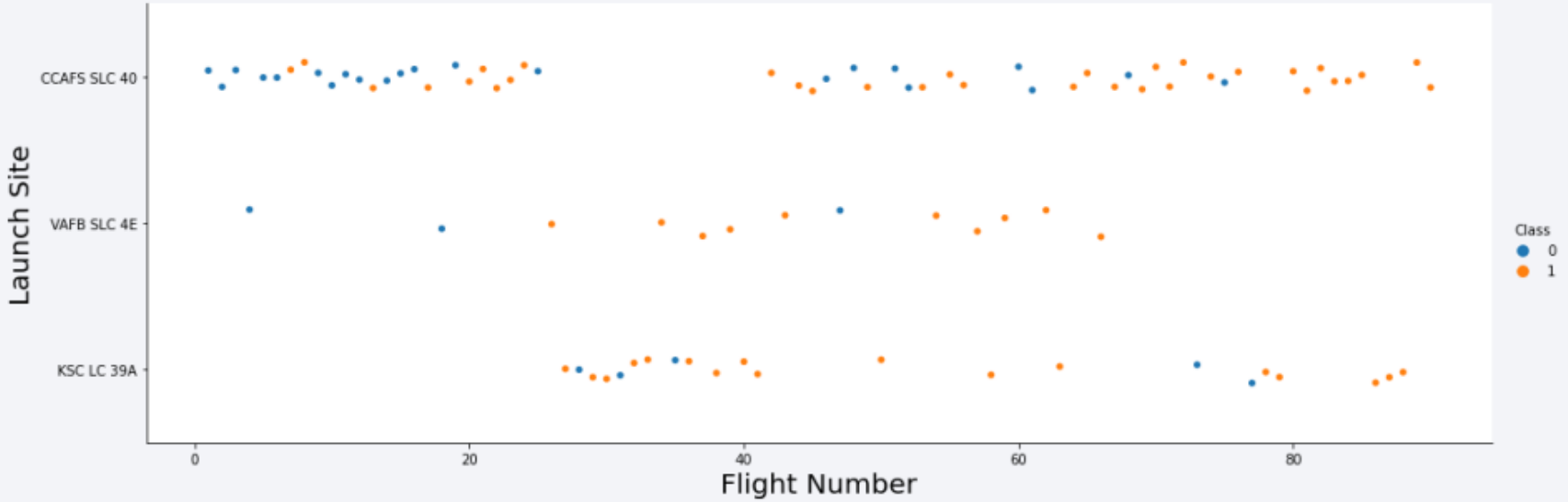


Results

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

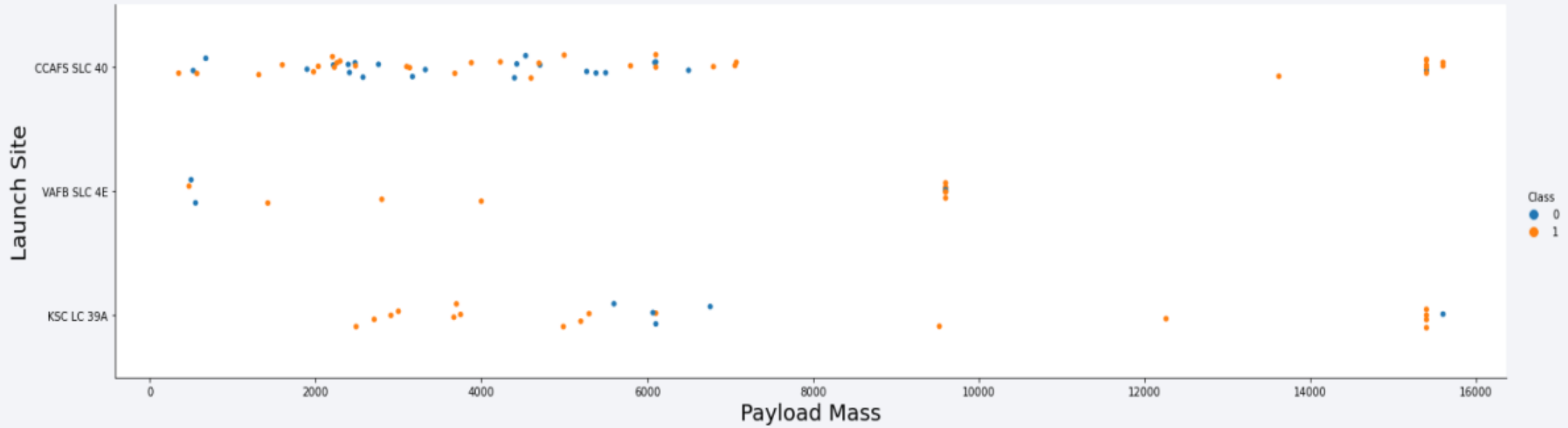

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- 1- CCAFS SLC 40 : is the most usable site for launching SpaceX's rockets and it has 55 trials, 33 of them are successful and 22 of them are failed # 60% success rate
- 2- VAFB SLC 4E : is the least usable site for launching SpaceX's rockets and it has 13 trials, 10 of them are successful and 03 of them are failed # 77% success rate
- 3- VAFB SLC 4E : is a moderate site in terms of launching SpaceX's rockets and it has 22 trials, 17 of them are successful and 05 of them are failed # 77% success rate

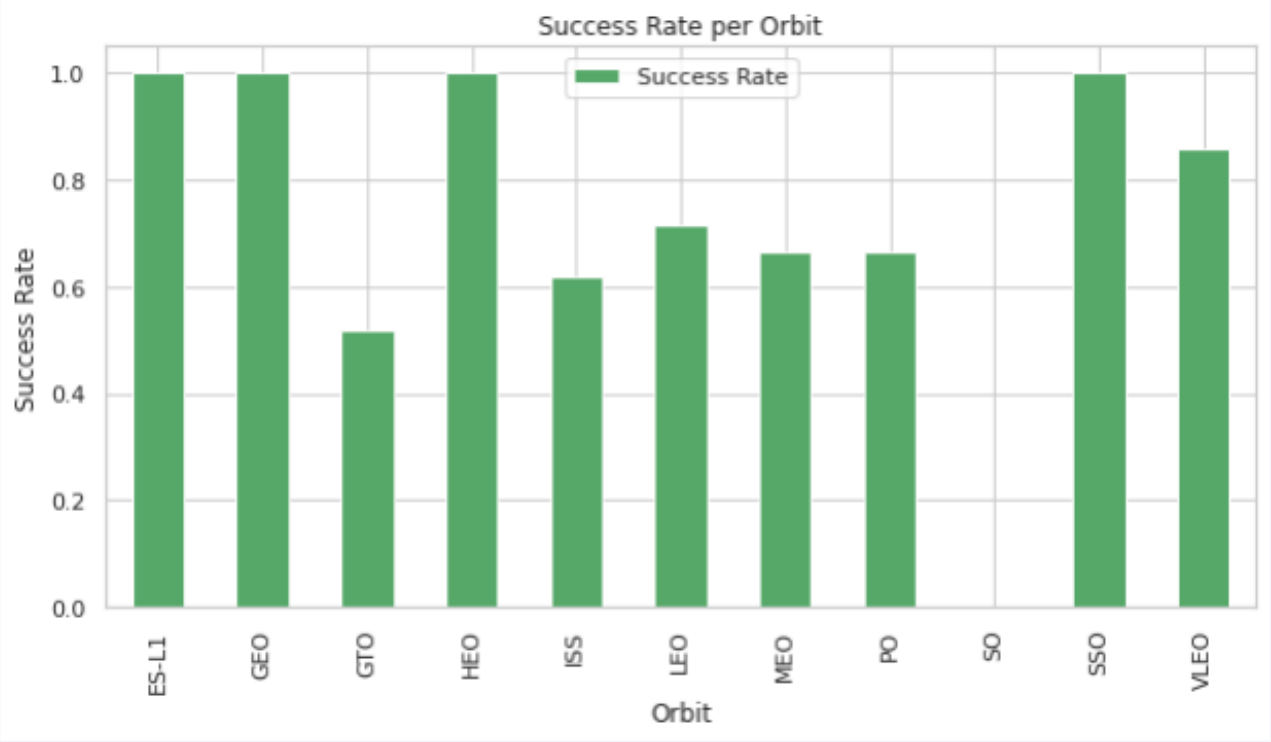
Payload vs. Launch Site



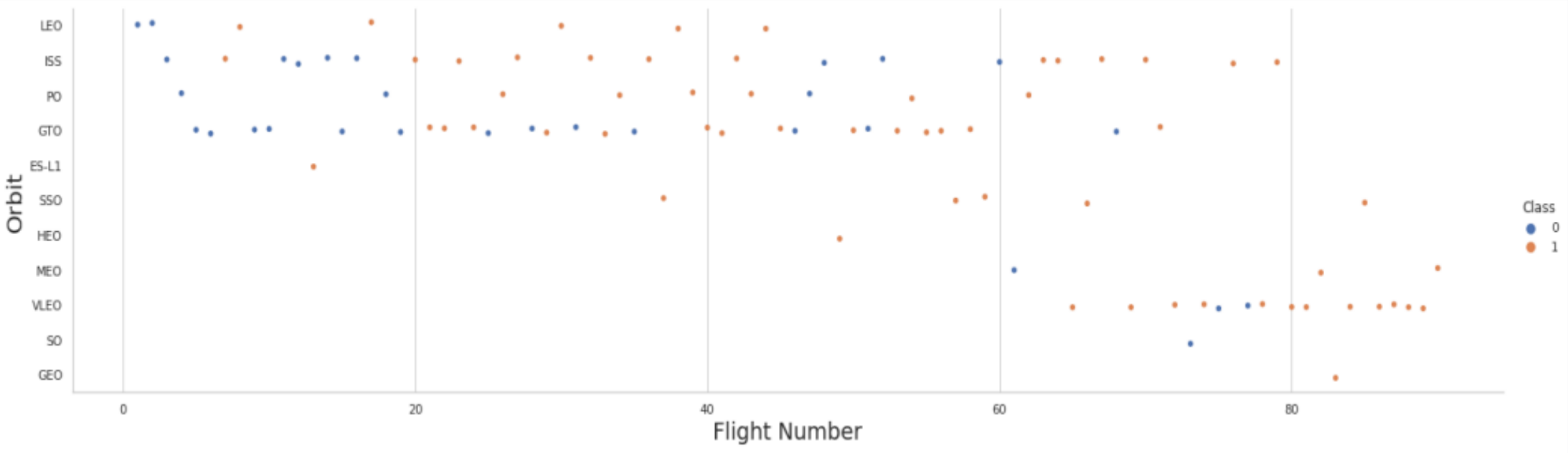
There is no strong relationship between the payload mass and the success of first stage return since there are approximately equal numbers of failed and successful trials.

Success Rate vs. Orbit Type

- The best orbits in terms of successful first stage returns are ['ES-L1', 'GEO', HEO, SSO]
- The worst orbit is 'GTO' , therefore we need to understand why it is the worst to avoid the failure of first stage return.

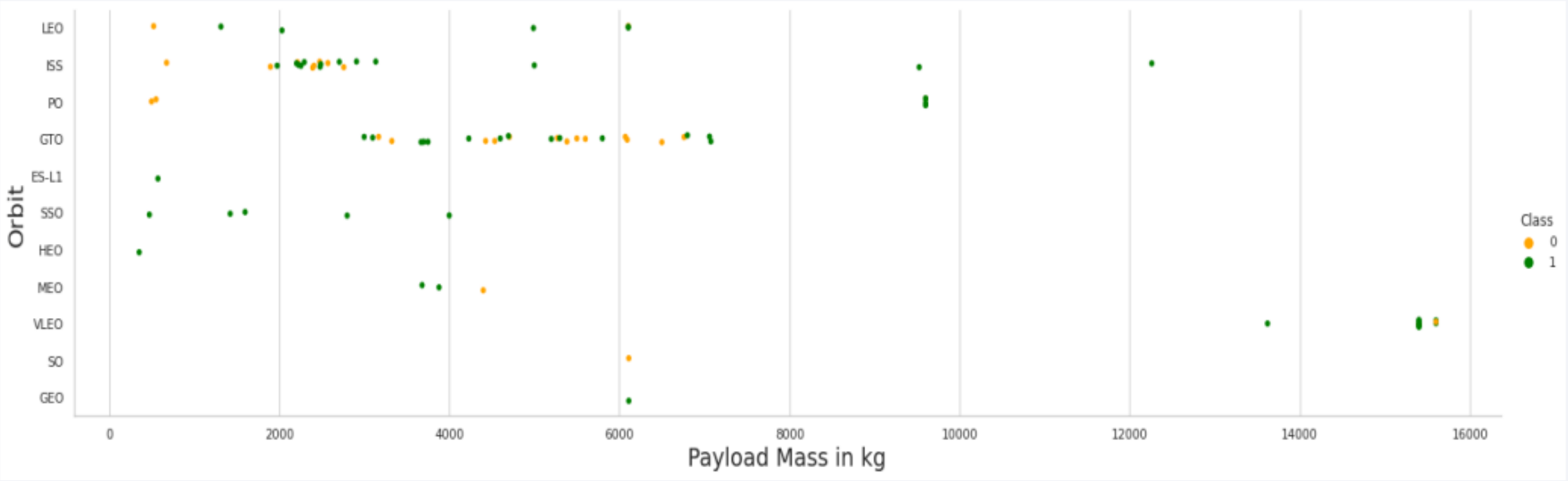


Flight Number vs. Orbit Type



In the LEO orbit, the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

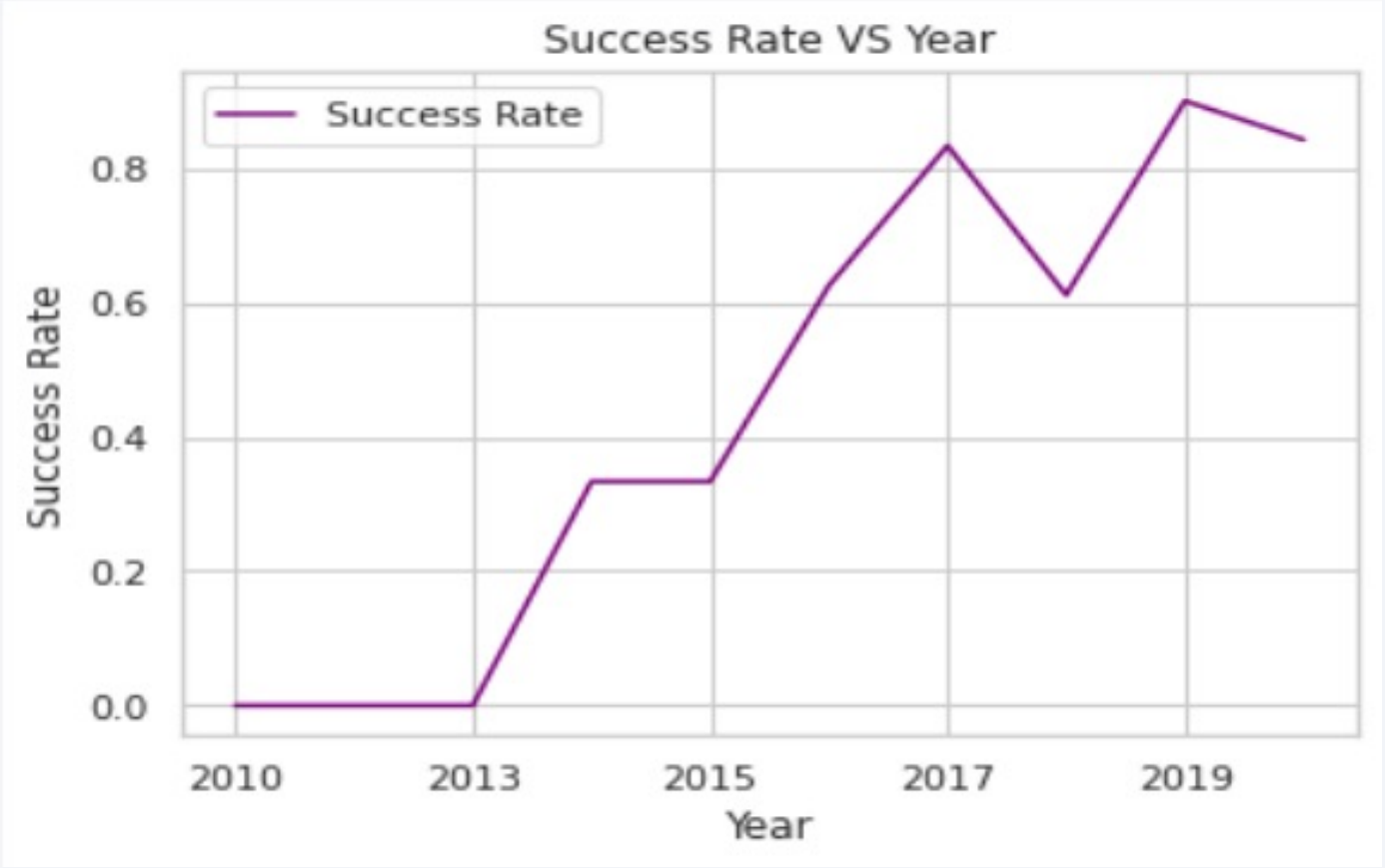
Payload vs. Orbit Type



Heavy Payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO orbits

Launch Success Yearly Trend

Observations: Success rate increased from 2013 to 2020



All Launch Site Names

The 4 Distinct sites for rocket launches:

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

```
%sql select distinct launch_site from SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

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

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

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcom
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total amount of payload that moved to the outer space by NASA through SpaceX rockets equals to 45596 Kg

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

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';
```

1
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928 kg

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

```
%sql select avg(payload_mass__kg_) as avg_mass_F9 from SPACEXTBL where booster_version = 'F9 v1.1'
```

avg_mass_f9
2928

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad is 2015-12-22

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

Hint:Use min function

```
%sql select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)'
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

1. F9 FT B1029.1
2. F9 FT B1036.1
3. F9 B4 B1041.1

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

```
%sql select booster_version from SPACEXTBL\
where (landing__outcome = 'Success (drone ship)' and (payload_mass__kg_ > 4000 and  payload_mass__kg_ > 6000));
```

booster_version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1

Total Number of Successful and Failure Mission Outcomes

We have only 1 failed mission while we have 99 successful ones.

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome
```

mission_outcome	counts
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The booster versions that carry the maximum payload starts with F9 B5 and ranges from B1048 up to B1060

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

```
%sql select distinct booster_version from SPACEXTBL\
where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXTBL);
```

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

2015 Launch Records

We have two failed landing in 2015 on a drone ship which both in the same site, CCAFS LC-40 and with same booster version F9 v1.1

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select landing__outcome, booster_version, launch_site from SPACEXTBL\
where (landing__outcome = 'Failure (drone ship)' and date like '2015%')
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%sql select landing__outcome, count(*) as counts_of_landing_outcomes from SPACEXTBL\
where DATE between '2010-06-04' and  '2017-03-20' group by landing__outcome\
order by count(landing__outcome) desc
```

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


Section 3
Launch Sites
Proximities Analysis

Folium Map: Launch Sites

- SpaceX selects coastal and low-latitude locations to minimize potential hazards and optimize launch conditions.
- All active launch sites are concentrated in California and Florida.

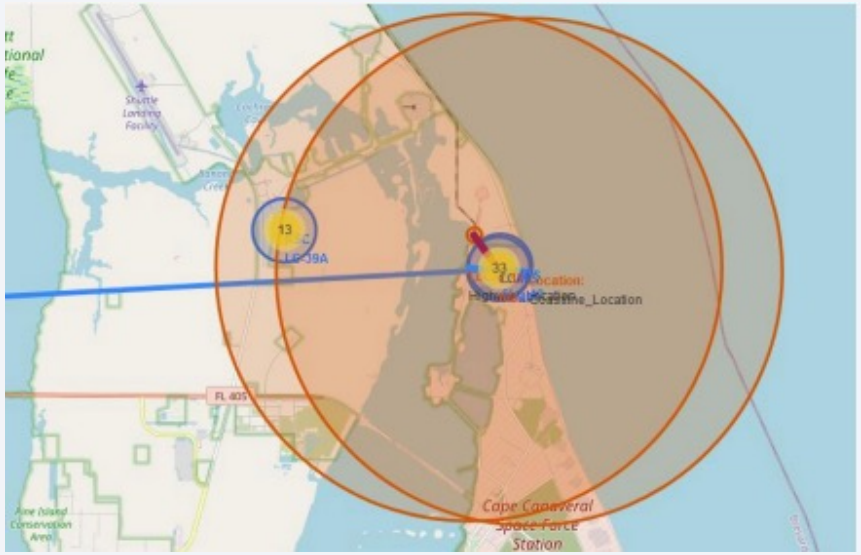
Folium Map: Success Rate for Each Launch Location

Green Marker = Successful Return

Red Marker = Failed Return

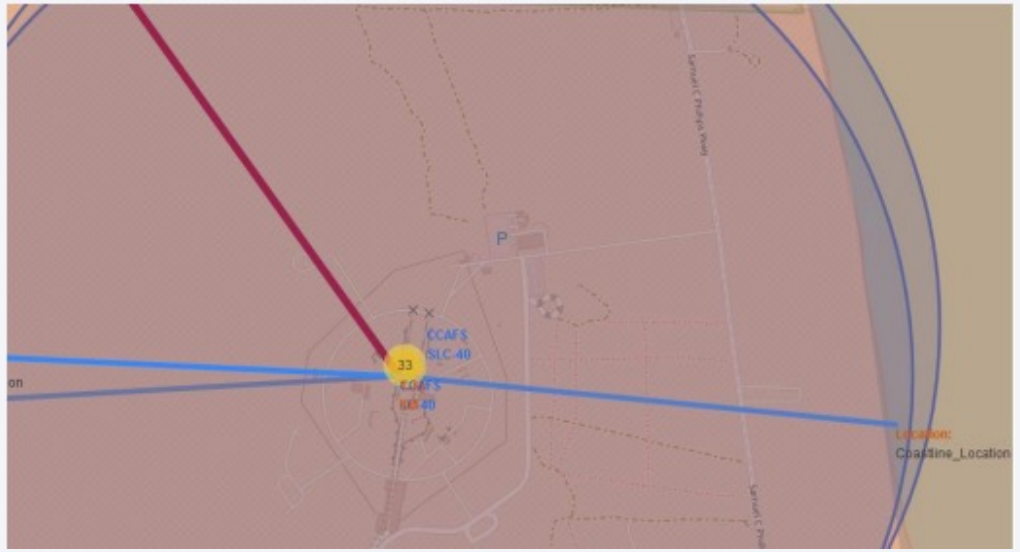
Folium Map: Closest Proximities to CCAFS LC-40

Proximities Coordinates

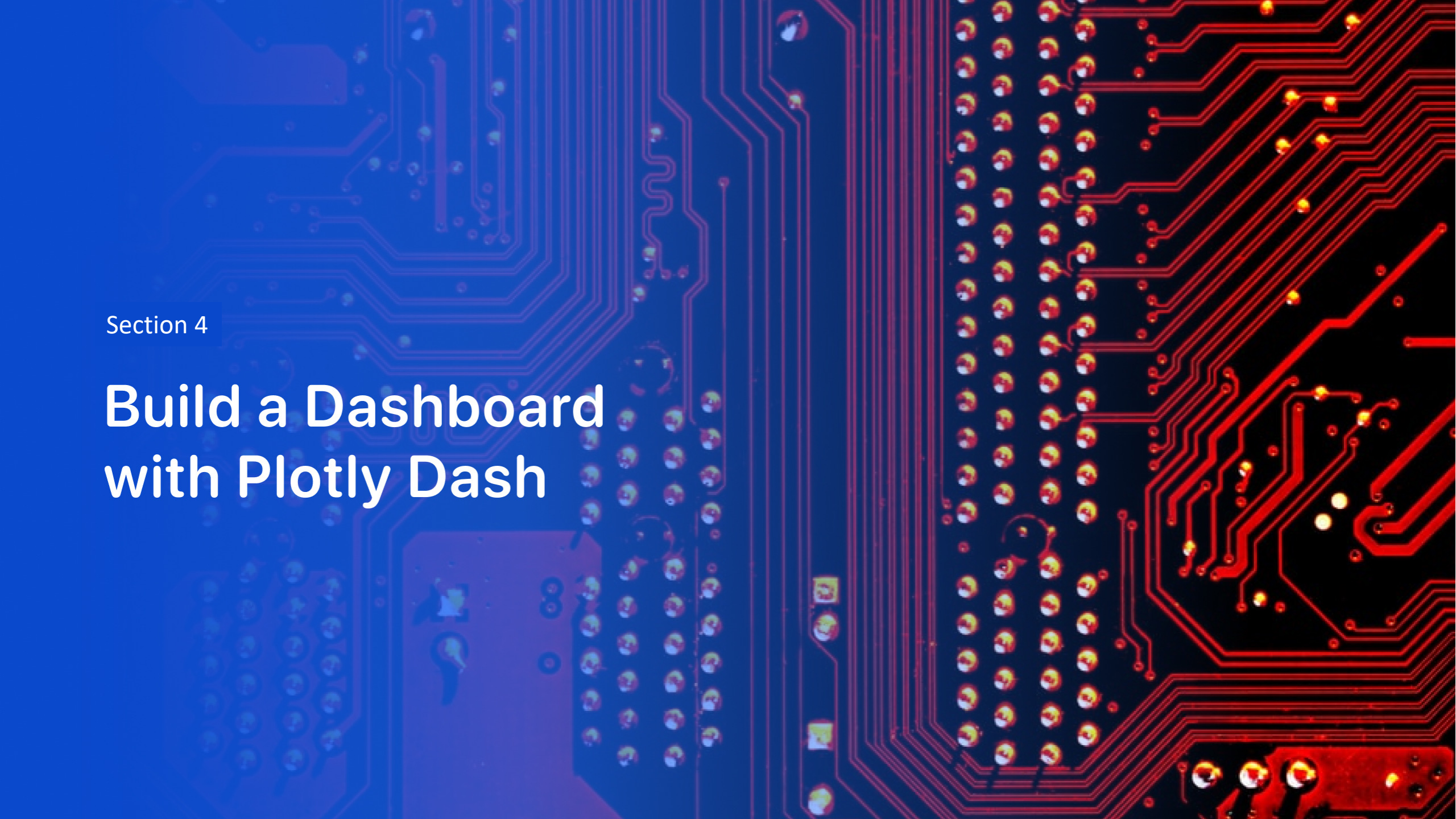


	Location	Lat	Long
0	Orlando_Location	28.52300	-81.38260
1	Coastline_Location	28.56146	-80.56746
2	Highway_Location	28.56270	-80.58703

Distances



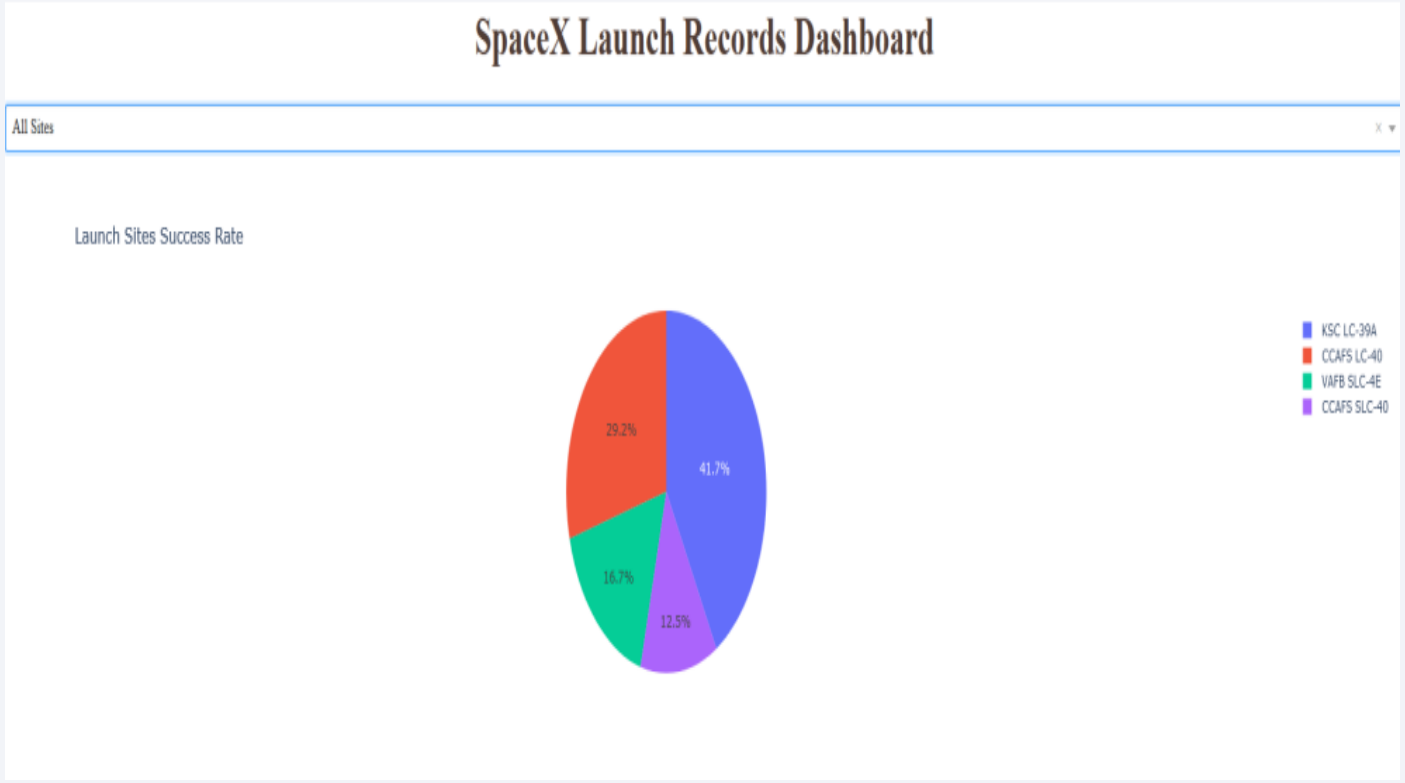
Orlando City Distance ≈ 78.8 Km
Coastline Distance ≈ 0.97 Km
Highway Distance ≈ 0.95Km


Section 4
Build a Dashboard with Plotly Dash

Dashboard: Launch Success Count(all sites)

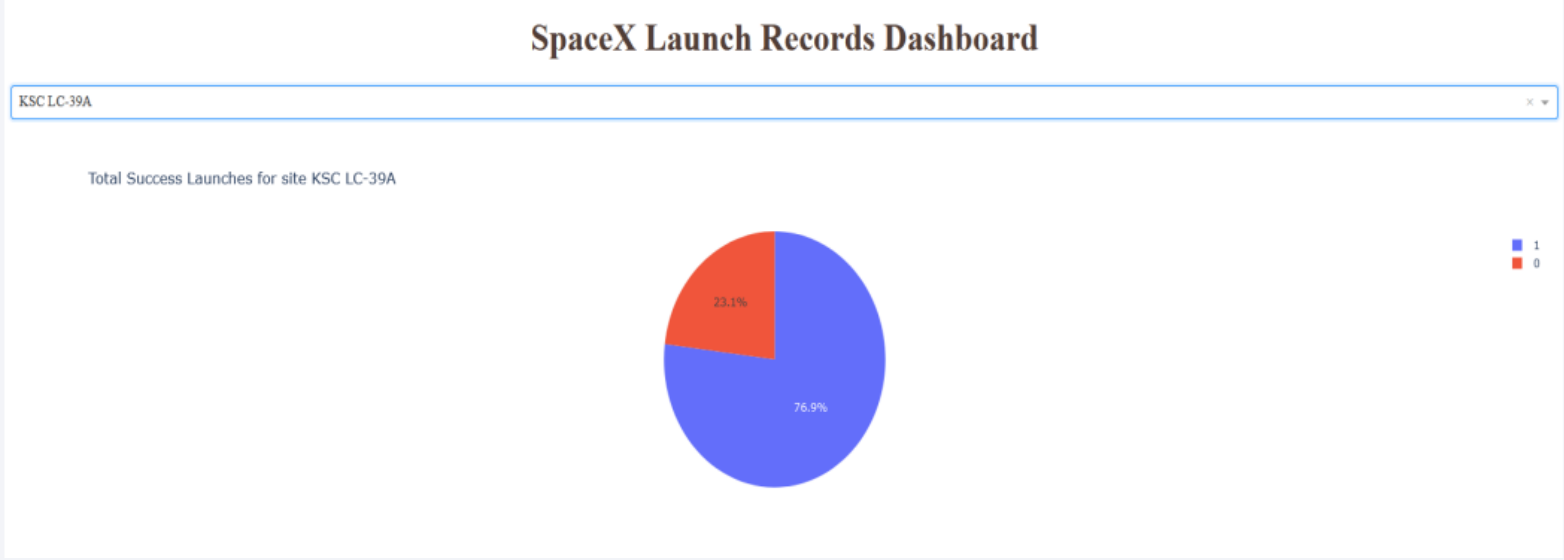
The graph shows the success percentage for every single site in terms of first stage return:

- The best site is KSC LC-39A with 41.7%.
- The least site for with only 12.5%. successful rocket launch: CCAFS SLC-40

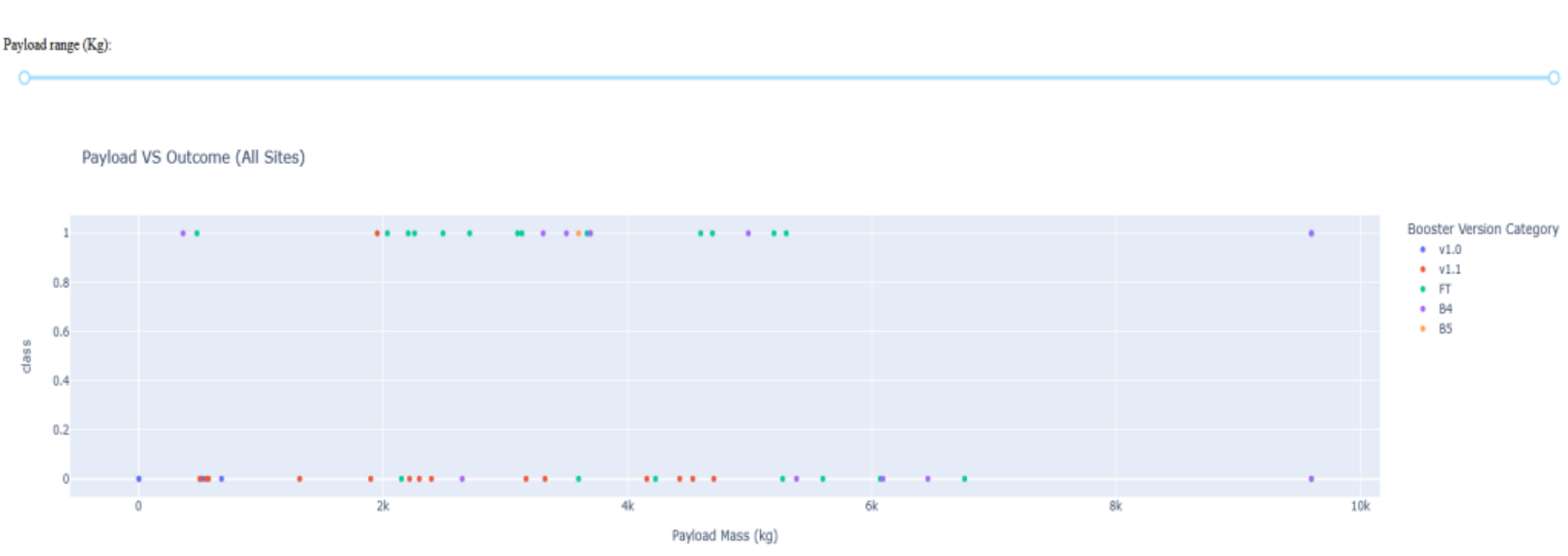


Dashboard: Launch success for KSC LC 39A

- Total Success Launches for site KSC LC-39
- KSC LC-39A with 76.9% successful missions
- KSC LC-39A with 23.1% Failed missions



Dashboard: Payload Mass vs Launch Outcome



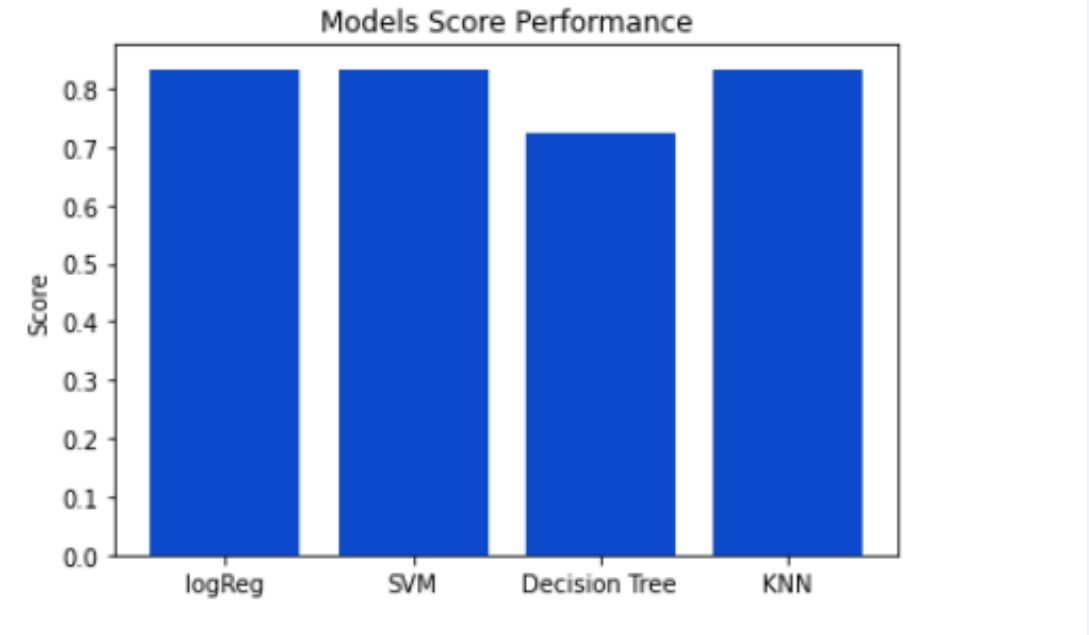
This interactive scatter plot illustrates the relationship between payload mass and launch outcome. We can observe that missions with payloads under 4,000 kg tend to have a higher success rate, particularly for certain booster versions.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistic Regression, SVM, and KNN produced identical results with a Jaccard score of 0.8.
- In contrast, the Decision Tree model showed the weakest performance among all models evaluated.

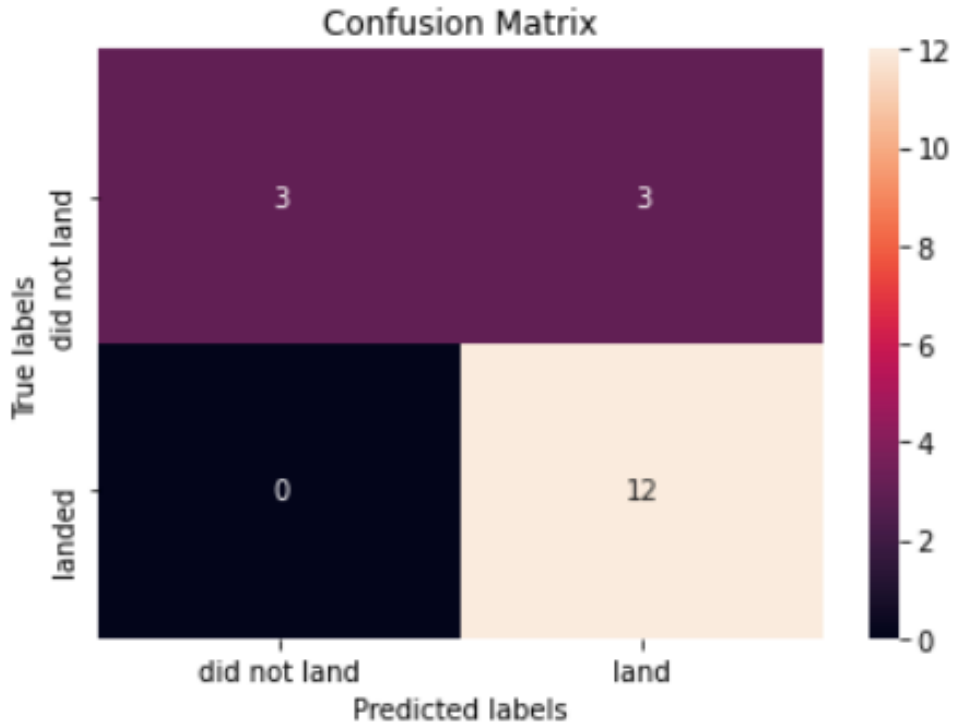


Confusion Matrix

Logistic Regression , SVM and KNN have the same confusion matrix and results:

- **True Positive = 12**
- **False Positive = 0**
- **True Negative = 3**
- **False Negative = 3**

```
Logistic Regression:
Jaccard Score of = 0.8
F1 Score = 0.777777777777778
====================================
SVM:
Jaccard Score of = 0.8
F1 Score = 0.777777777777778
====================================
Decision Tree:
Jaccard Score of = 0.6666666666666666
F1 Score = 0.672727272727272
====================================
KNN:
Jaccard Score of = 0.8
F1 Score = 0.777777777777778
====================================
```



Conclusions

- **Mission characteristics influence first stage success**
  Success varies with orbit type, payload mass, and booster version. Lighter to moderate payloads and orbits such as ES-L1, GEO, HEO, and SSO show higher return success, while GTO missions have notably lower performance due to increased payload demands.

- **Launch site selection supports safety and operational efficiency**
  All active SpaceX launch sites are coastal and located in Florida or California, close to transport infrastructure and water for safe ascent and recovery. These locations reduce logistical costs and mitigate risks during launch and return.

- **SpaceX performance improved significantly over time**
  From 2013 through 2020, first stage return outcomes show a steady rise in success rate, reflecting gains in engineering reliability, mission design, and vehicle reusability. Booster generations also evolved to support heavier payloads.

- **Predictive modeling confirms success is not random**
  Machine learning models using payload, orbit, launch site, and booster attributes achieved competitive performance. Logistic Regression, SVM, and KNN reached the same Jaccard score of 0.8, confirming that mission outcomes can be predicted from operational data.…

Appendix

- SpaceX API URL
- SpaceX Static Wikipedia URL
- SpaceX data used in ML training

Thank you!