

Winning Space Race with Data Science

Mohamed Ali
February 24, 2025



Outline

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



Executive Summary

- Today, SpaceX is considered one of the best companies over the world in terms of space rockets launches, due to the rocket science advancements that achieved in making space missions more affordable and practical, generally space rockets companies offer a rocket launch with a cost of 165 million dollars, whereas SpaceX offers the same service for only 62 million dollars which considered a huge savings led organizations such NASA to sign contracts with SpaceX.
- In this report, I am taking the role of a data scientist working for a new rocket company, called **SpaceY** that would like to compete with SpaceX founded by Billionaire industrialist Elon Musk. my job is to use data science instead of rocket science to discover the possibility of competing with SpaceX. I am doing this by gathering information about SpaceX, performing data analytics, modeling ML algorithms and creating dashboards for my team.

Introduction

- Since 1957, the countries around the world are competing to expand beyond Earth, whether through satellites launches or space exploration, these missions require huge amounts of money where a launch of one space rocket costs in average 165 million dollars, a company like Space X changes the equation by reducing this amount of money massively to only 60 million dollars due to its unique and advanced technologies in returning the first stage of rocket structure.
- As mentioned above one of the key factor of SpaceX's success in the space race is the first stage of rockets return through a safe landing, from this point we will discover together what are the main attributes or variables that control these successful landings, through asking questions about the nature of the launch process , the payload mass of rocket, launch location, rocket orbit, and more by analyzing and visualizing these information through the data science methodology provided by IBM.

Section 1

Methodology

Methodology



1-Data collection Methodology:

The data was collected using SpaceX rest API in addition of using data web scraping on Wikipedia webpages.



2-Perform datawrangling:

The data was preprocessed using Pandas and NumPy, some of main technique are used: OneHotencoding, unnecessary columns removal, data normalization and standardization.



3-Perform exploratory data analysis (EDA)

Using libraries such as seaborn and matplotlib for visualization and SQL for data queries.



5-Perform predictive analysis using classification models

- Starting with splitting the data into train and test sets
- Identifying the best algorithm and parameters through hyperparameters tuning using Grid Search
- Adopting the best algorithm and parameters for the purpose of model deployment.

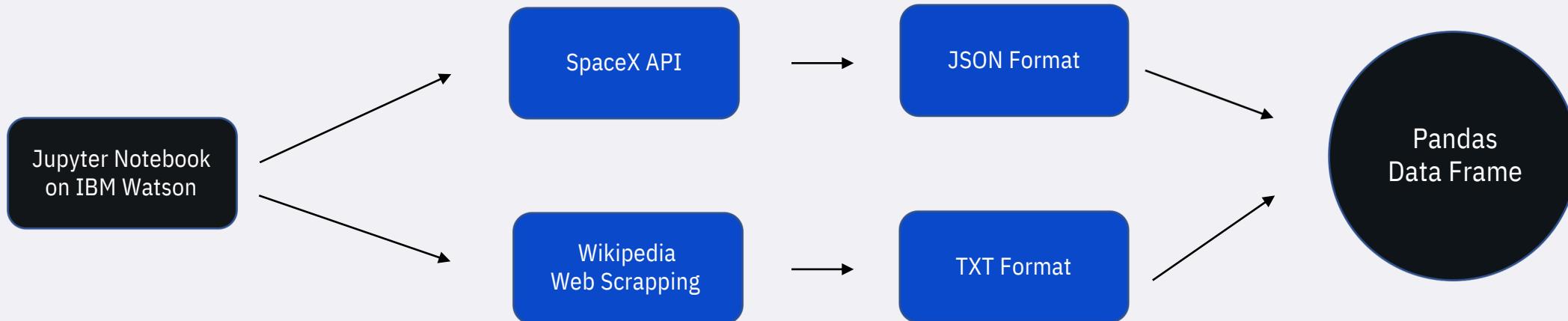


Data Collection

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

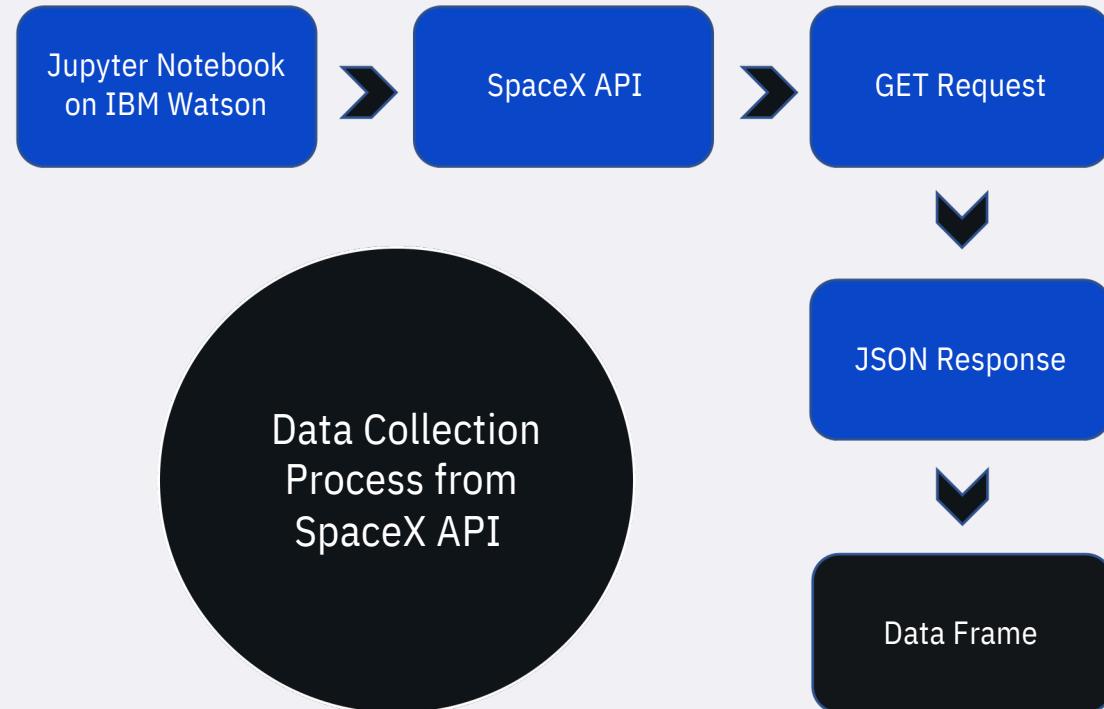
The Process of data collection:





Data Collection –SpaceX API

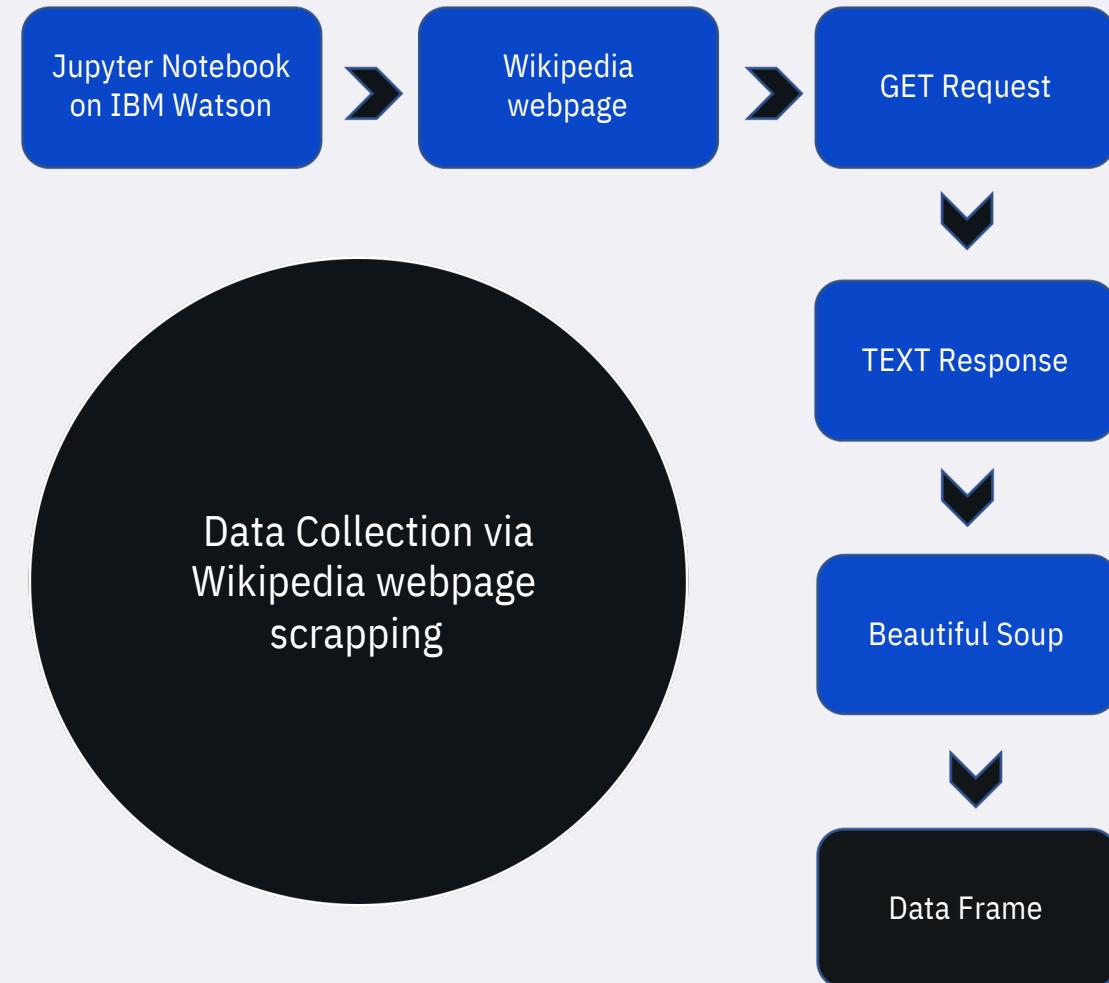
- We started the data collection from SpcaeX API by importing the required libraries such as pandas, NumPy and Request, then we established a URL GET request, this request is raised as JSON file to be finally converted to a data frame through choosing the required information like the geospatial info,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





Data Collection -Scraping

- As we have done before we start by ~~library beautiful soup and requests~~ 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.





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.

Data Wrangling stages

1-Loading the collected dataset.

2-Identifying and calculating the percentage of the missing values in each attribute

3-Identifying which columns are numerical and categorical:

4-Calculating the number of launches on each site

5-Calculating the number and occurrence of each orbit

6-Creating a landing outcome label from Outcome column

7-determining the success rate of returning the first stage of the rocket

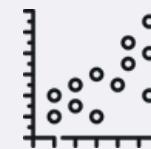


EDA with Data Visualization

- 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.

EDA with Data Visualization Stages

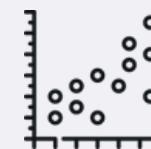
Visualize the relationship between Flight Number and Launch Site



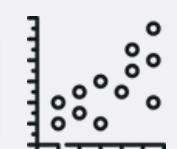
Visualize the relationship between success rate of each orbit type



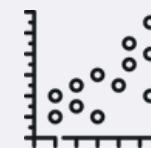
Visualize the relationship between Payload and Launch Site



Visualize the relationship between Flight Number and Orbit type



Visualize the relationship between Payload and Orbit type



Visualize the launch success yearly trend





EDA withSQL

- In this stage we used SQL quires as shown on the right to complete our EDA on our collected dataset
-

SQL Quires

- 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.

- We started our interactive map by drawing 4 circles on 4 different sites belonging to Falcon 9 rocket launches 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 LC- 40) 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

Building an Interactive 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

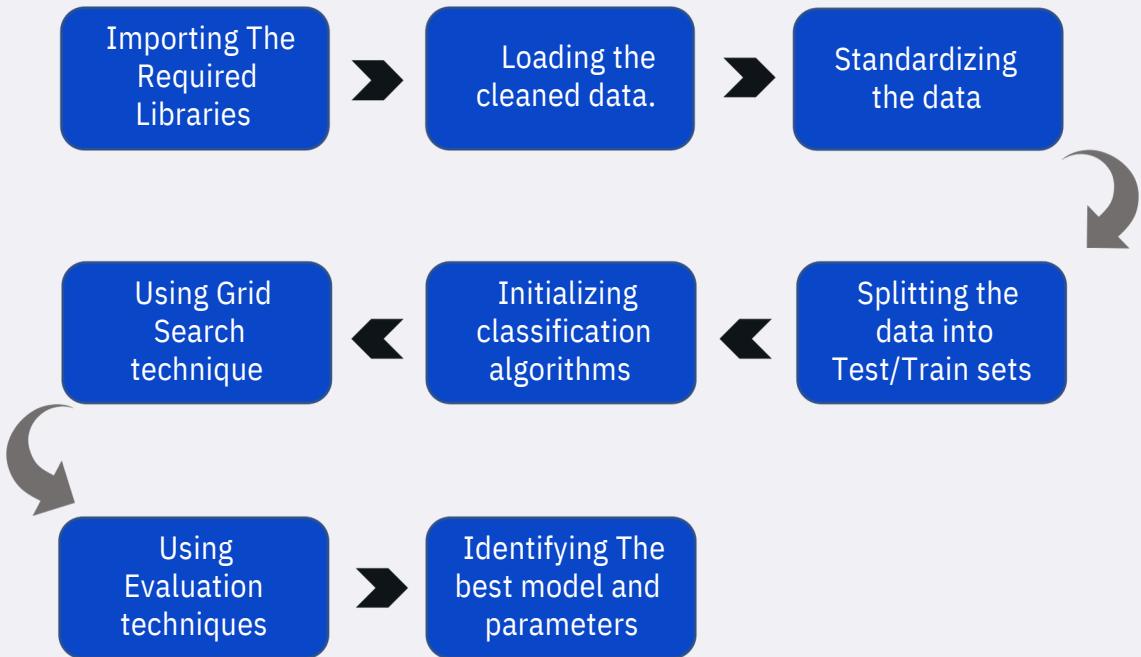


Predictive Analysis (Classification)

Machine Learning Stages:

- 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:
 - Logistic Regression (LR)
 - Support Vector Machine (SVM)
 - Decision Tree (DT)
 - K nearest neighbors (KNN)
- 6-Using GridSearchtechniquetofind 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 Pipelines





Results



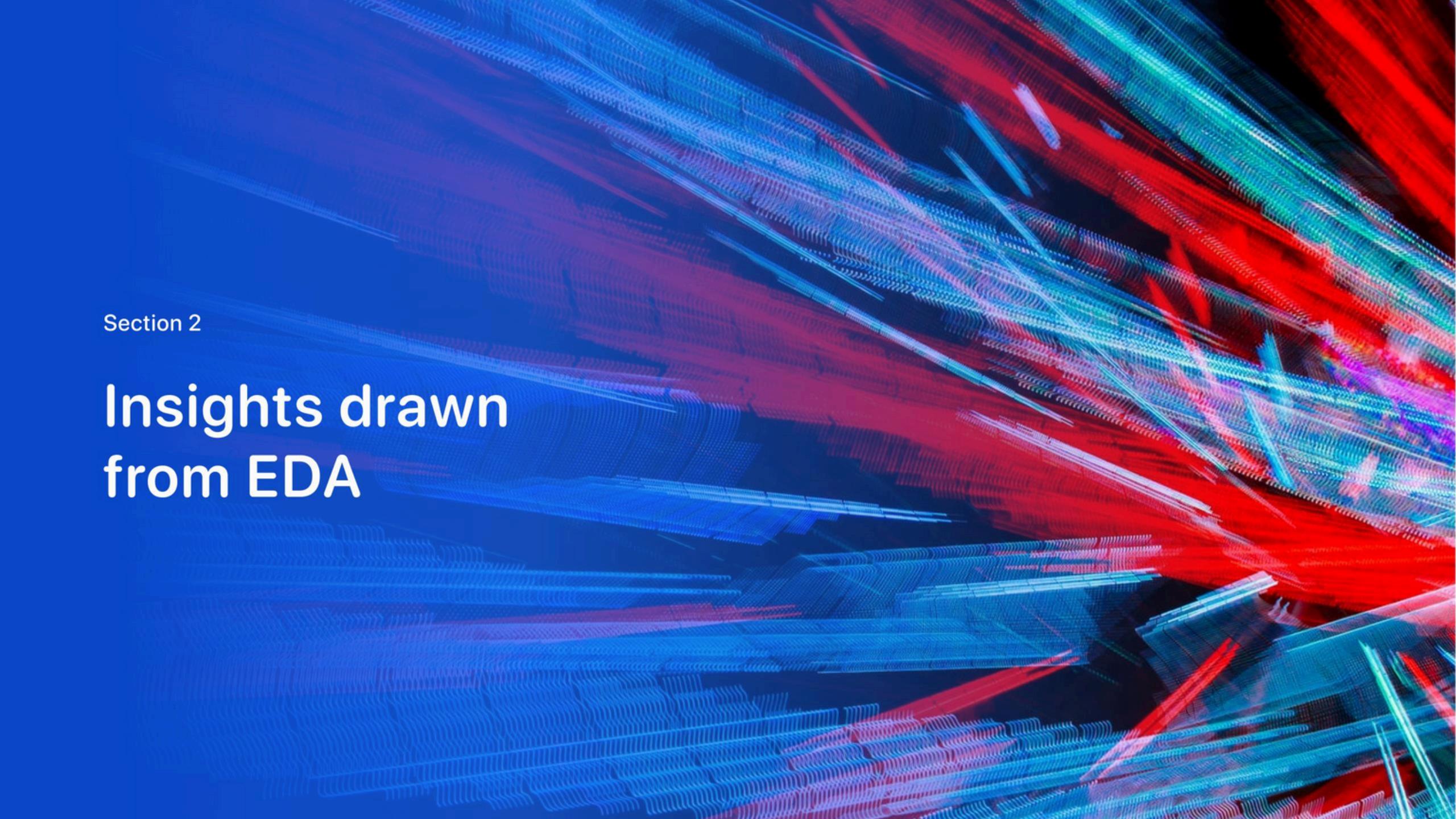
Exploratory data
analysis results



Interactive analytics
demo in screenshots



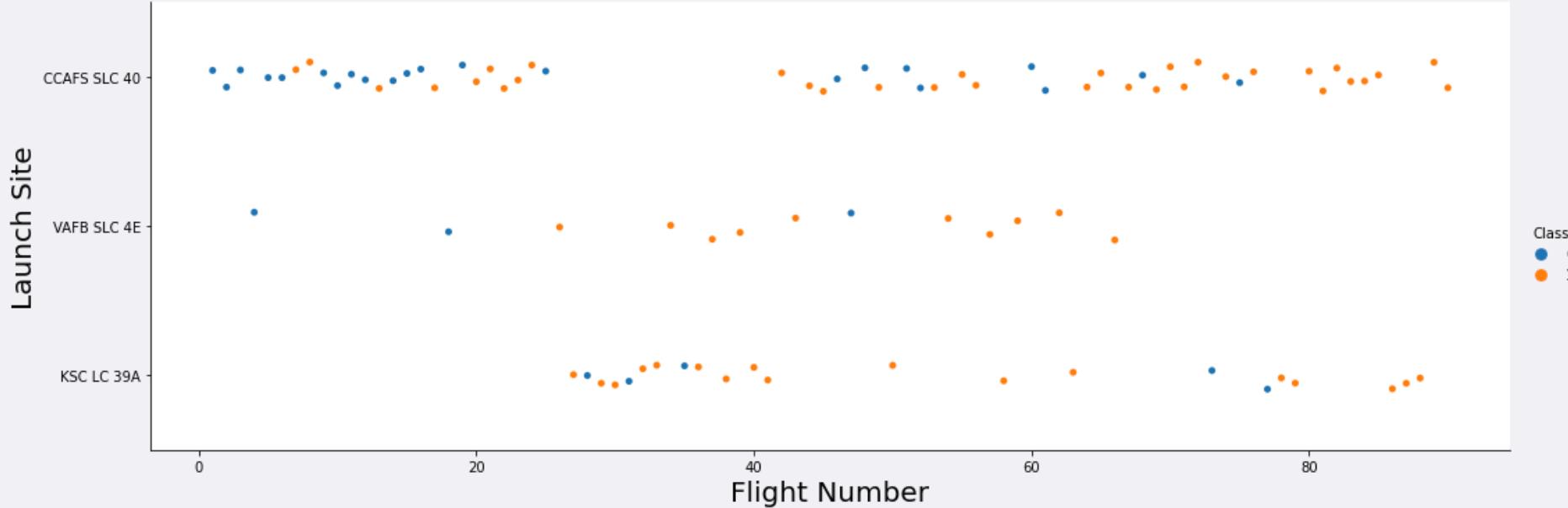
Predictive analysis
results

The background of the slide features a dynamic, abstract pattern of glowing particles. These particles are arranged in numerous thin, wavy lines that curve and twist across the frame. The colors of these lines are primarily shades of blue, red, and green, with some purple and white highlights. The overall effect is reminiscent of a digital or quantum simulation, suggesting complex data flow or energy distribution.

Section 2

Insights drawn from EDA

Flight Number vs. LaunchSite

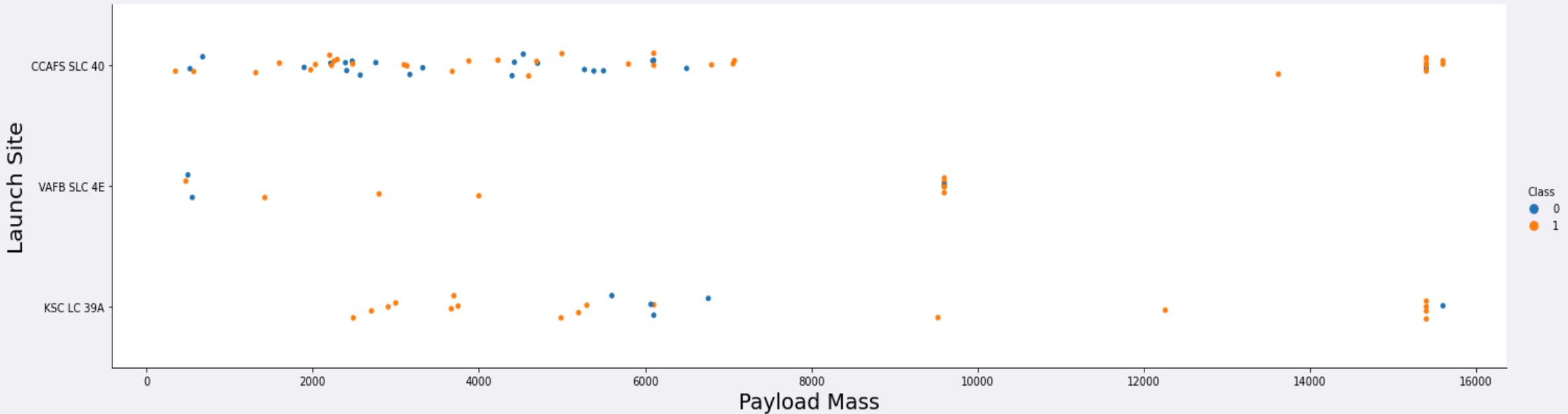


1-CCAFS SLC 40 : is the most usable site for launching SpaceX's rockets and it has 55 trials, 33 of them are successful and 22 of them are failed # 60% success rate

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

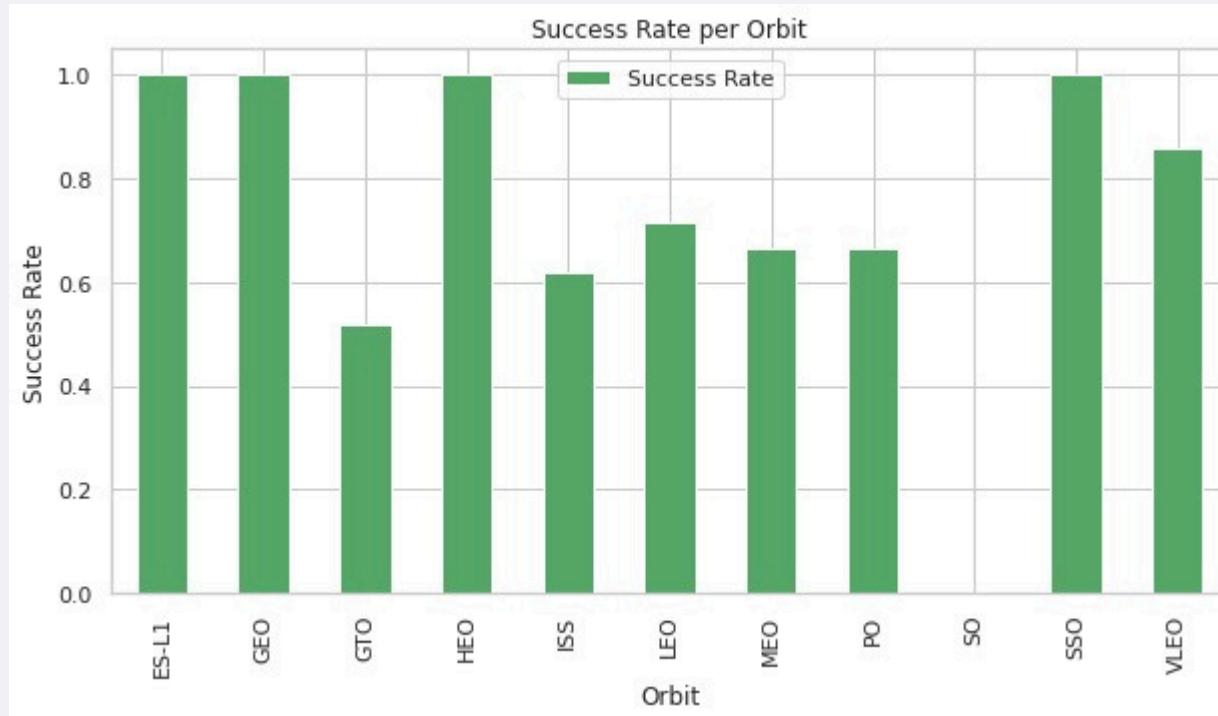
Payload vs. Launch Site



According to the plot above:

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

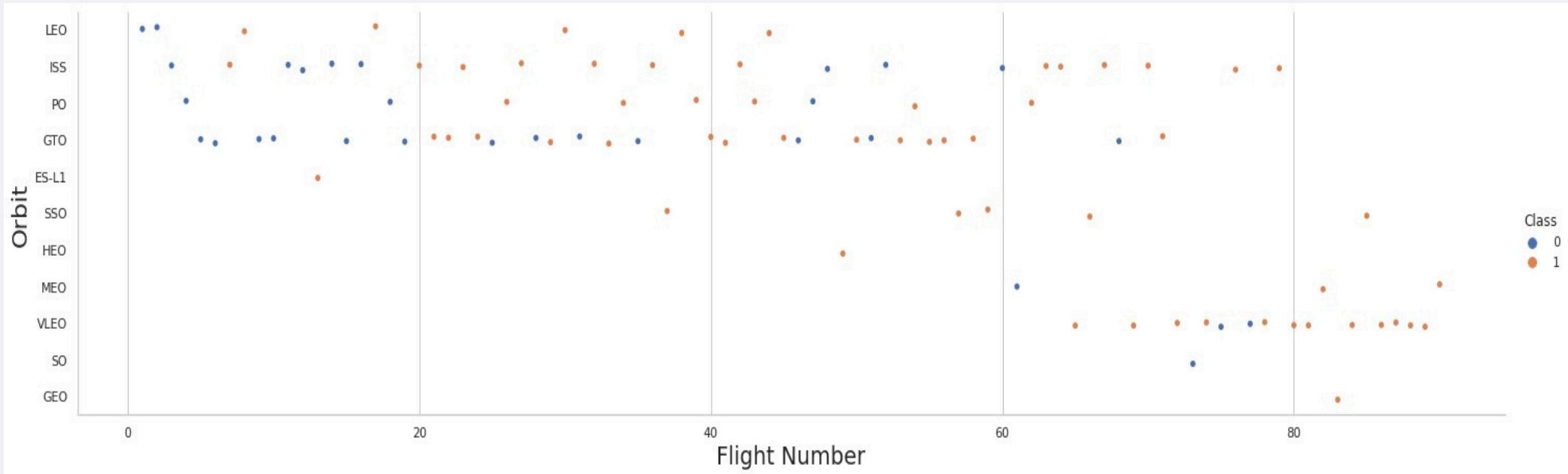
Success Rate vs. Orbit Type



According to the bar plot :

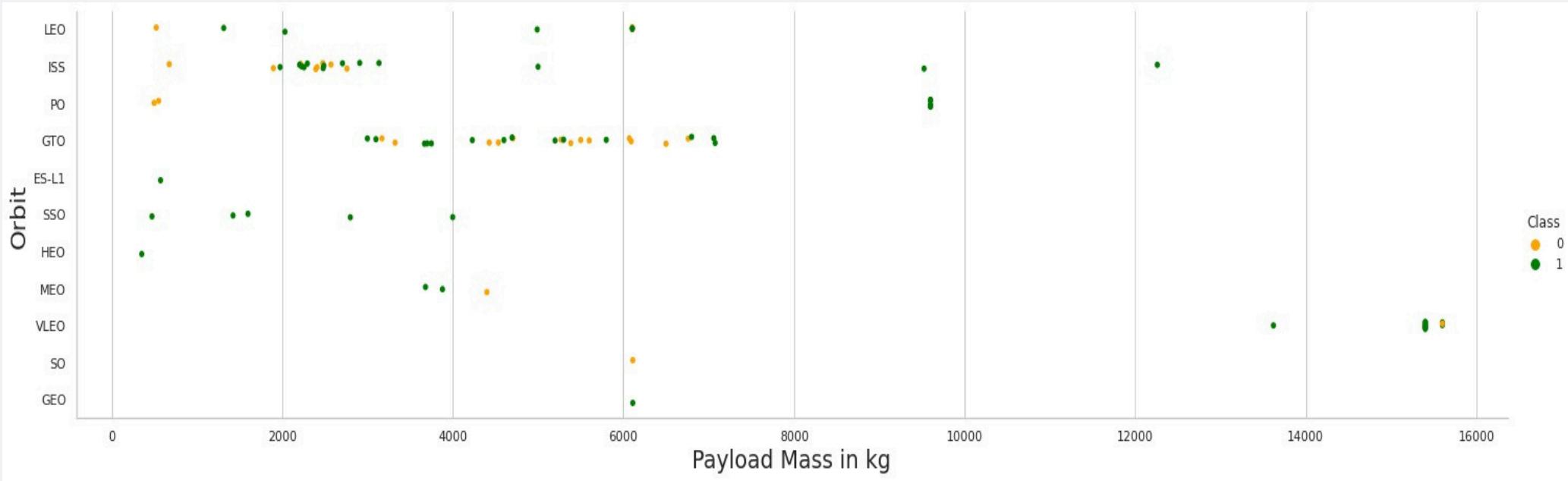
The best orbits in terms of successful first stage returns are ['ES-L1', 'GEO', HEO, SSO] Where the worst orbit is 'GTO' , therefore we need to understand why it is the worst to avoid the failure of first stage return.

Flight Number vs. Orbit Type



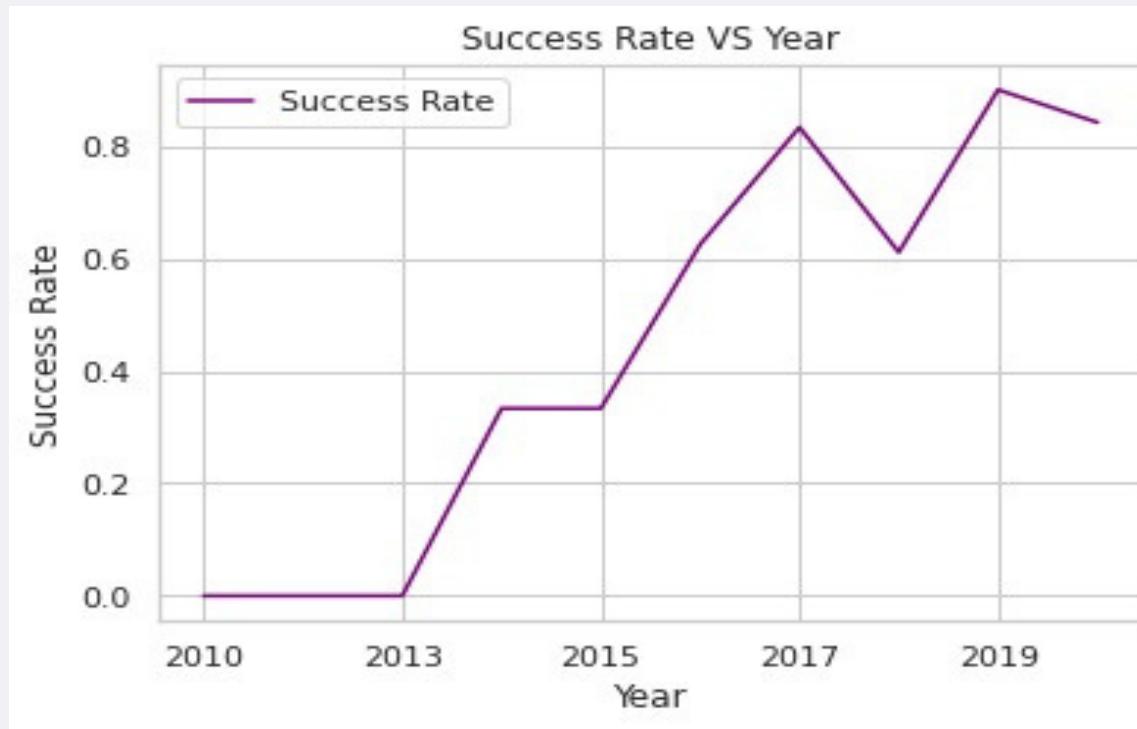
We can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. OrbitType



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

Launch Success Yearly Trend



We can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

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

As shown above we have [4 distinct sites](#) for rockets launches listed in the Table above.

Launch Site Names 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

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

The total amount of payload that moved to the outer space by NASA through SpaceX rockets equals to [45596 Kg](#) which is equal to [50.261 Us ton](#).

Average Payload Mass by F9 v1.1

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

The average payload mass carried by booster version F9 v1.1 is [2928 kg](#)

First Successful Ground Landing Date

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

Hint: Use min function

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

1
2015-12-22

Date of the first successful landing outcome on ground pad
was in [22-12-2015](#)

Successful Drone Ship Landing with Payload between 4000 and 6000

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

The boosters which have success in drone ship landing with payload between 4000 and 6000 kg are :

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

Total Number of Successful and Failure Mission Outcomes

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

We can see clearly that the success rate of mission outcomes is the most dominant we have only 1 failed mission while we have 99 successful ones.

Boosters Carried Maximum Payload

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

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

2015 Launch Records

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

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

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
```

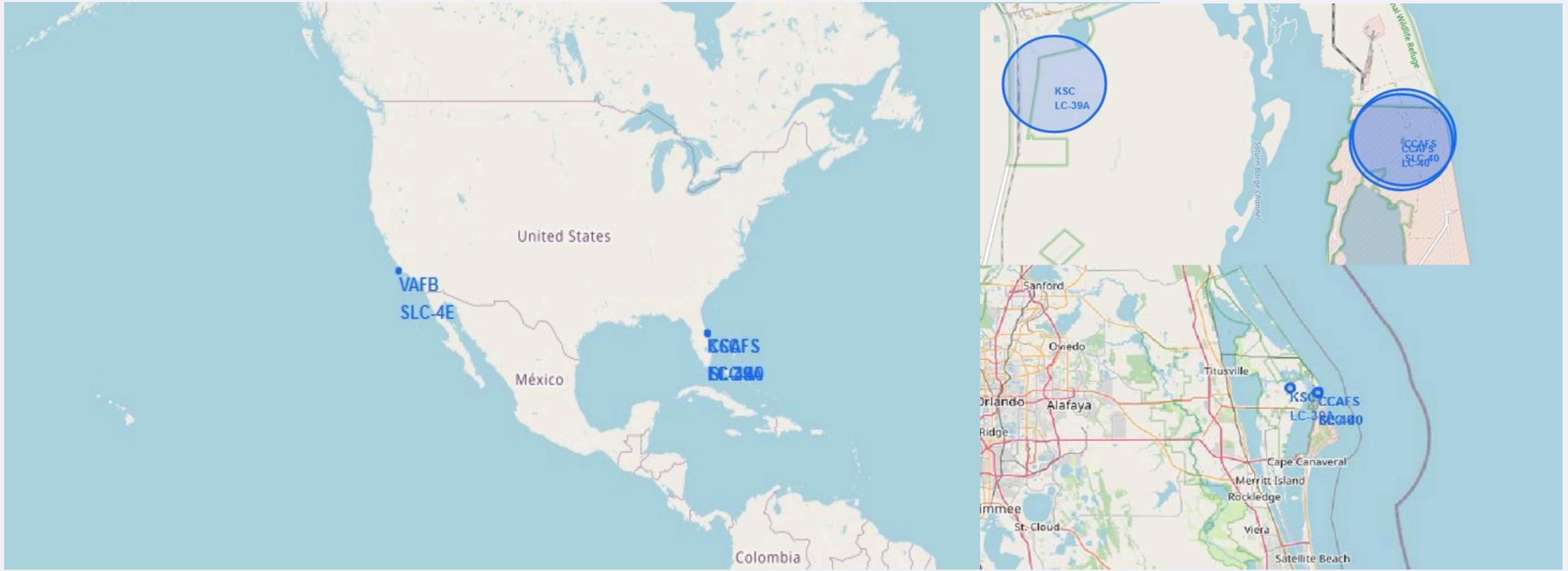
landing_outcome	counts_of_landing_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against the dark void of space. City lights are visible as numerous glowing yellow and white spots, primarily concentrated in the lower right quadrant where the United States appears. The atmosphere is visible as a thin blue layer, and the horizon shows the transition from the dark void to the blue of the atmosphere.

Section 4

Launch Sites Proximities Analysis

Folium Map: Launch Sites.



All site locations are near the coast and Equator line, SpaceX focuses on locations that are close to water and the zeroth latitude for the purpose of avoiding any undesired accidents.
The launch sites are distributed in two states California and Florida

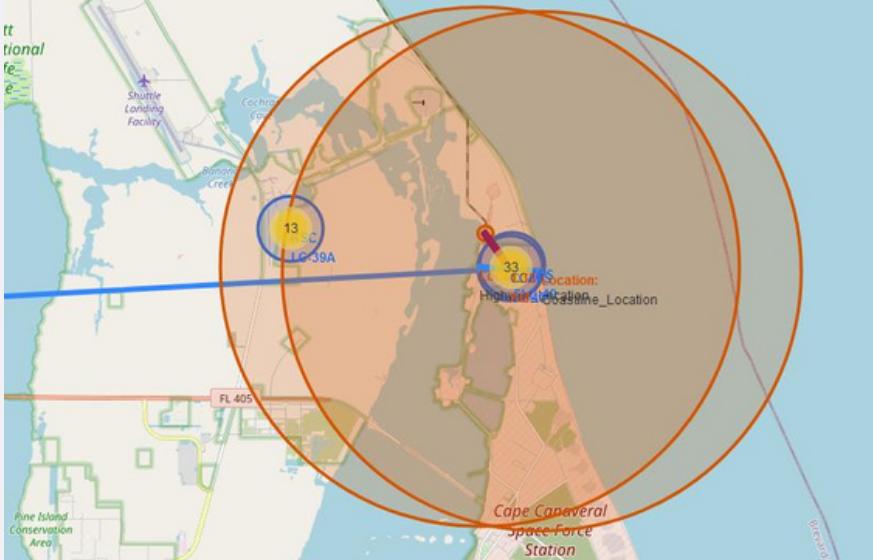
Folium Map: Success rate for each launch location



From the color-labeled markers in marker clusters, we can easily identify which launch sites have relatively high success rates. **GreenMarker** = **SuccessfulReturn**

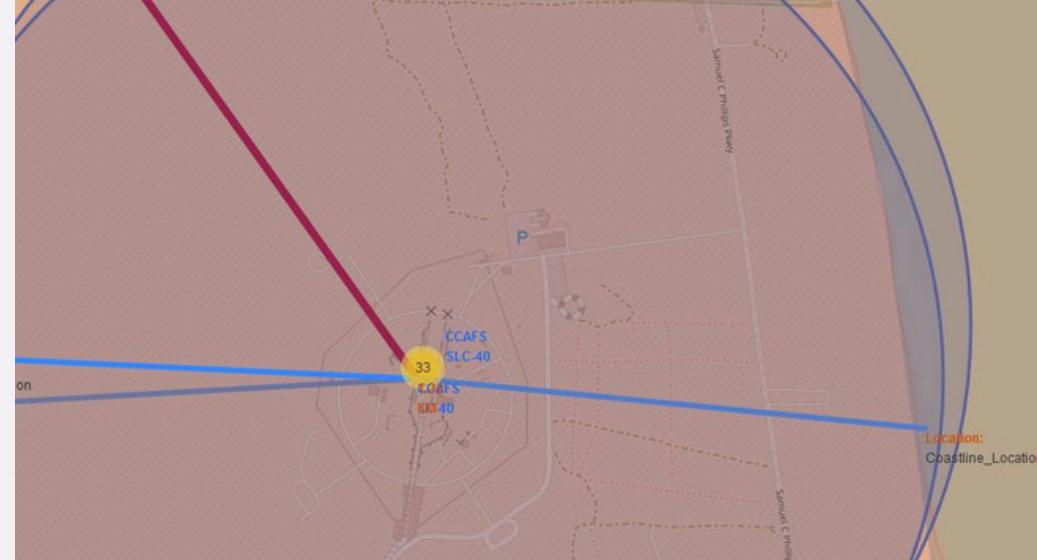
RedMarker = **FailedReturn**

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

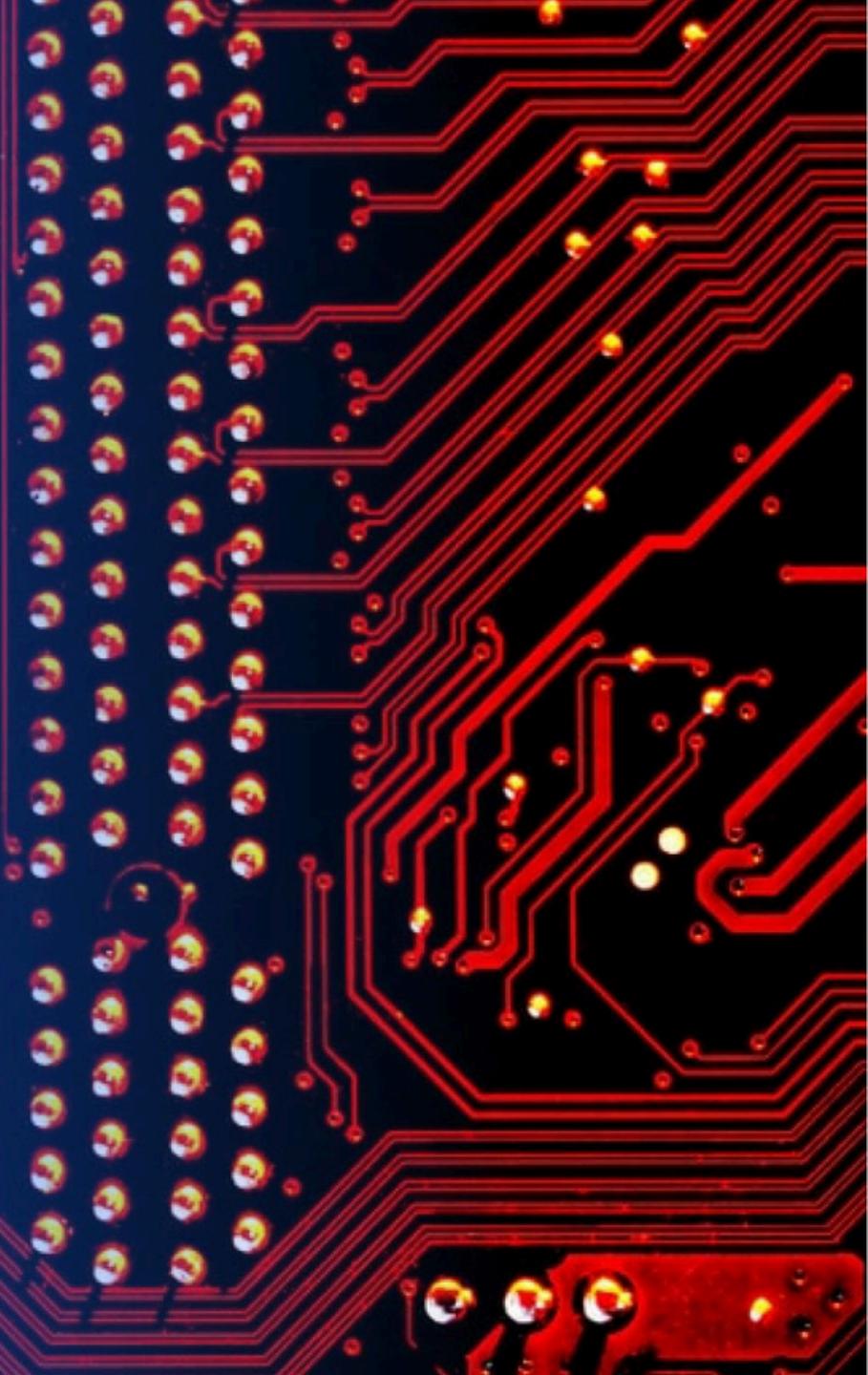


we calculated the distances between the launch site (CCAFS LC-40) to its proximities

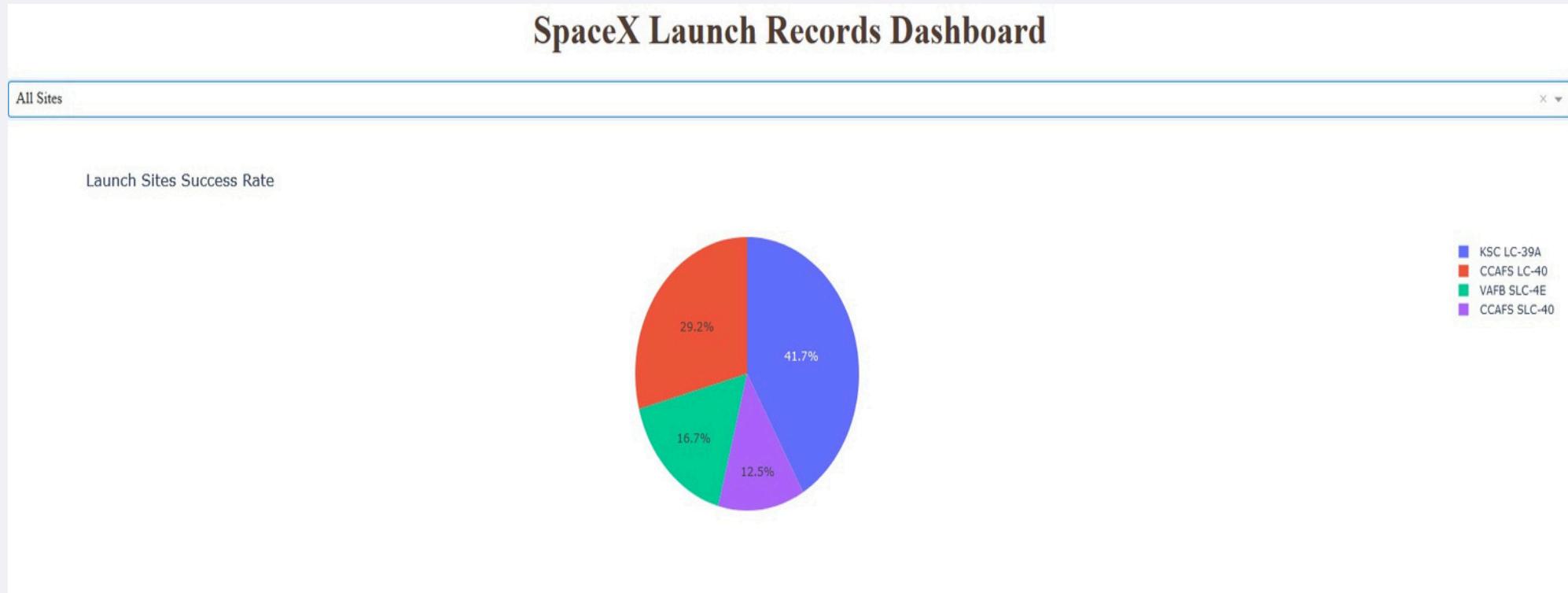
Orlando City Distance \approx 78.8 Km,
Coastline Distance \approx 0.97 Km,
Highway Distance \approx 0.95Km

Section 5

Build a Dashboard with Plotly Dash



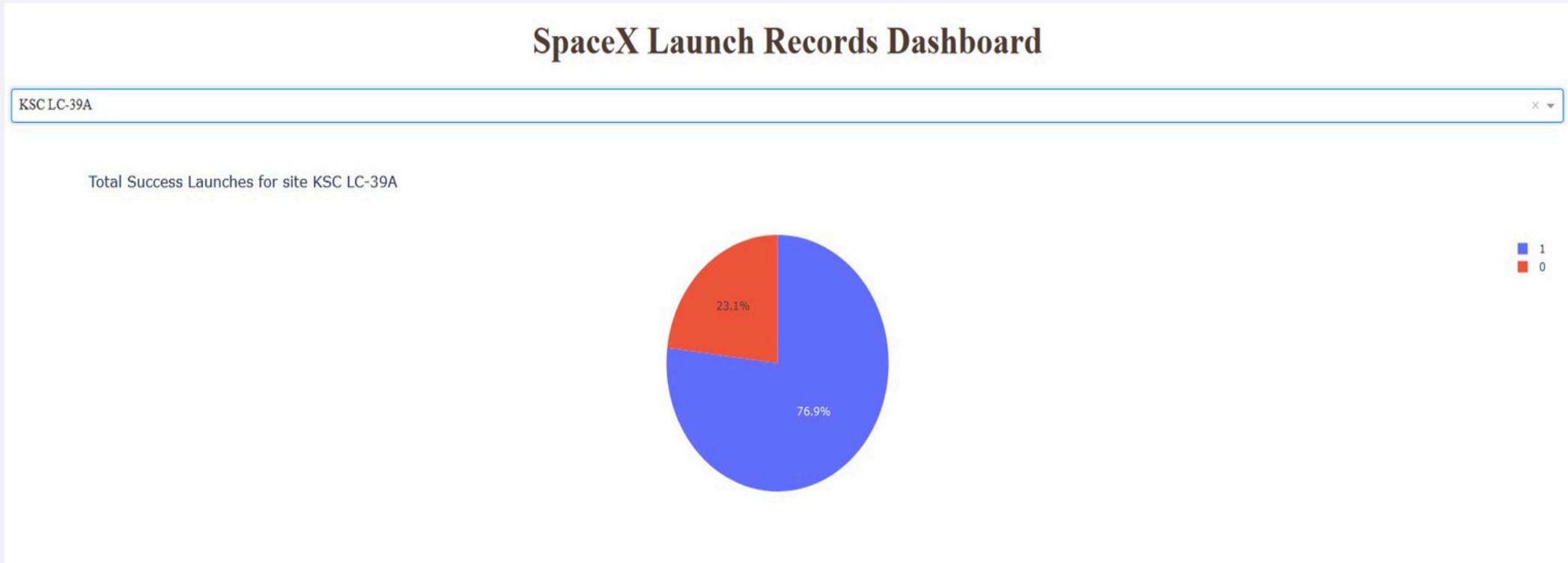
Dashboard: Launch success count for all sites



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% of total successful rocket launch among all sites
- The least site for successful rocket launch: CCAFS SLC-40 with only 12.5%.

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 Screenshot 3>



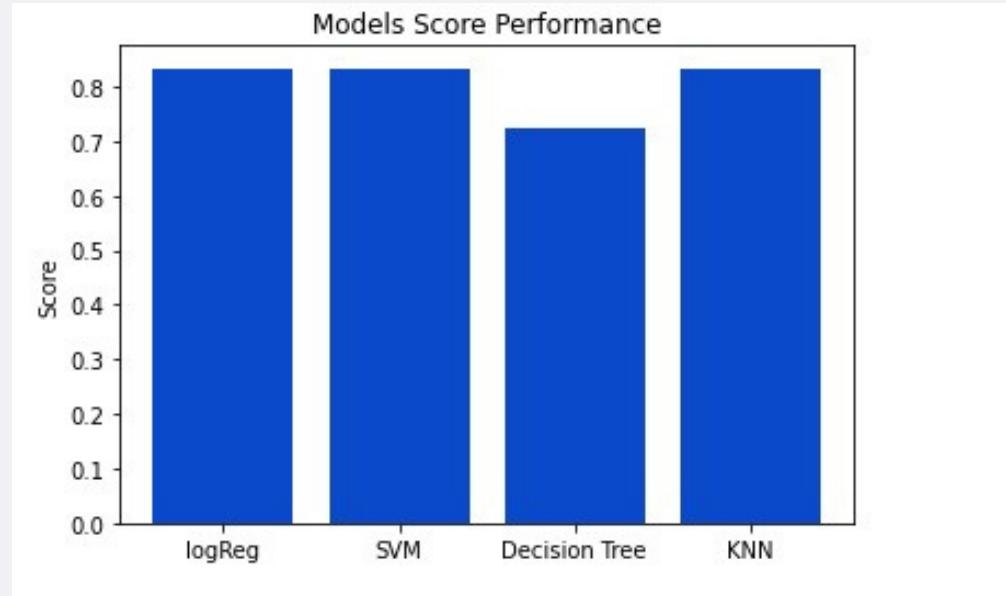
This interactive scatter plot shows the launch outcome based on the relationship between the payload mass and the final outcomes as we can infer that the success rate per booster version for example, the payload mass < 4000 kg more likely to be successful.

The background of the slide features a dynamic, abstract design. It consists of several thick, curved lines in shades of blue and yellow, creating a sense of motion and depth. The lines curve from the top right towards the bottom left, with some lines being more prominent than others. The overall effect is reminiscent of a tunnel or a high-speed train track.

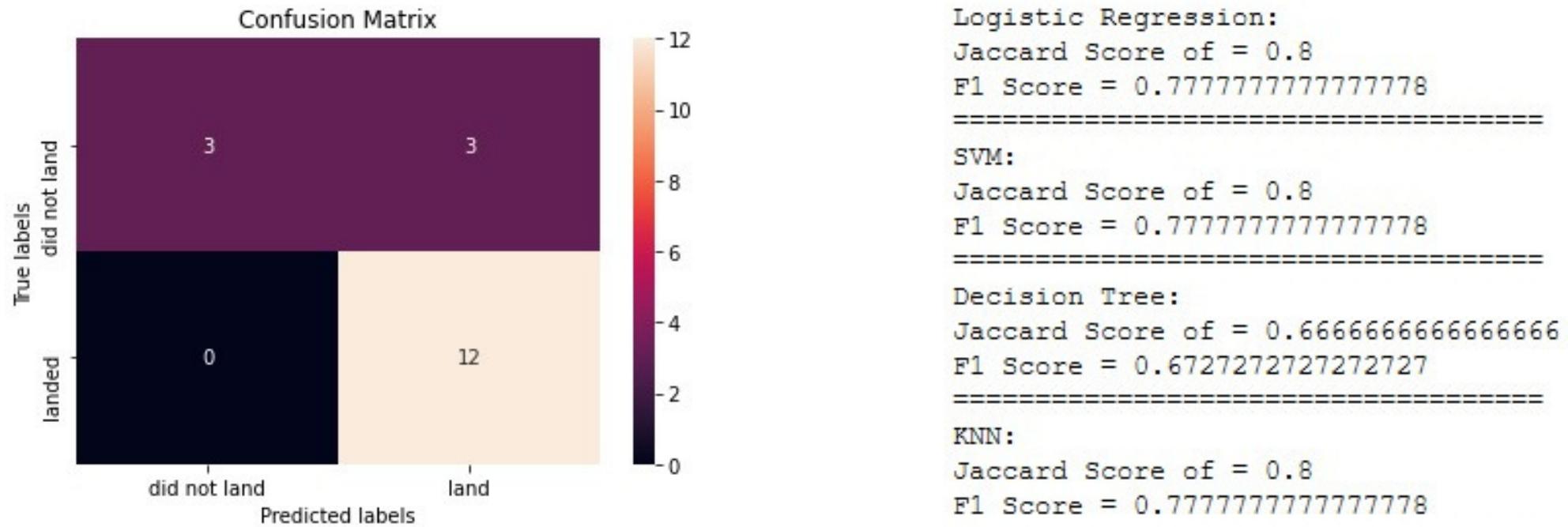
Section 6

Predictive Analysis (Classification)

Classification Accuracy



LogisticRegression, SVM and KNN has the same performance where the Jaccard Score is same : 0.8
Where Decision Tree has the worst performance compared to other models.



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

Conclusions



a successful first stage return, leads to huge savings in terms of rockets lunches cost



A wide range of attributes affects the possibility of a successful first stage return. In our model there were 83 attributes were taken into consideration.



SpaceX Falcon9 launch sites were all close to a highway, railway, and coastline proximities, which aimed in transportation cost-reduction



insight requires further investigation. SpaceX success rate increased with years, KSC LC -39A site is the highest in success rate.



Orbit and booster version affect success rate.

Appendix

- SpaceX API URL "[Click Here](#)"
- SpaceX Static Wikipedia URL "[Click Here](#)"
- SpaceX data used in ML training "[Click Here](#)"

Thank you!

