WINNING SPACE RACE WITH DATA SCIENCE

VIRAJ NOORITHAYA

MARCH 5TH 2023

OUTLINE

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

EXECUTIVE SUMMARY

- In this capstone project, we will predict whether the SpaceX Falcon 9 first stage will land successfully. If we can determine if the first stage will land, we can determine the cost of a launch. This will be achieved with the use of different machine learning classification algorithms.
- The methodology followed includes Data Collection through API and Web Scraping, Data Wrangling and Preprocessing, Exploratory Data Analysis, Data Visualization and finally, Machine Learning Prediction.
- During our investigation, the results of our analysis indicate that there are some features of rocket launches which can be good indicators of a successful launch
- In the end we conclude that the even though all Algorithms show the same accuracy, Logistic regression due to its interpretability might be the best machine learning algorithm to for this problem

INTRODUCTION

- SpaceX prides itself in being able to reuse the first stage of a rocket launch so much so that they advertise on their website that their rocket launches cost 62 million while other provides cost upward 165 million.
- Much of these savings are down to the first stage's reusability. If we can determine if the first stage will land, we can determine the cost of a launch.
- The main goal of this capstone project is to predict whether the Falcon 9 first stage will land successfully
- This information can be used if an alternate company wants to an informed bid against SpaceX for a rocket launch

METHODOLOGY

SECTION 1

METHODOLOGY

- Data collection methodology
  - Data was collected using SpaceX API and Web Scraping from Wikipedia
- Perform data wrangling
  - Data was cleaned, irrelevant columns were removed, one-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Scatter Graphs, Bar Graphs to show relationships between variables to show patterns of data.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Visually examine impact of different payloads and lunch sites on outcome
- Perform predictive analysis using classification models
  - Create a machine learning pipeline to predict if the first stage will land given the data.
  - Train the best performing model to make accurate predictions.

DATA COLLECTION

- Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes

- In order to predict launch outcome, we collected data in the following 2 ways:
  - Using SpaceX APIs using Requests library
  - Using Web Scraping using Beautiful Soup library

- Information from these 2 sources were then transformed into dataframes, cleaned, consolidated, filtered and exported as flat files for easier exploration and predictions

DATA COLLECTION – SPACEX API

GET request to SpaceX API to acquire launch related information

Parse JSON response and convert to a Pandas DataFrame

Select required columns and clean values

Get core, booster, launch site and payload related information using related APIs through auxiliary functions

Create a consolidated dataset by merging data from all API requests

Filter data to include only Falcon 9 information

Check missing values in columns and impute using average value

Notebook Link

DATA COLLECTION - SCRAPING

Request the Falcon 9 Launch Wiki page through its URL → Create a BeautifulSoup object from the HTML response → Extract all column/variable names from the HTML table header

↓

Create an empty dictionary with keys from the extracted column names → Fill up the dictionary with launch records extracted from table rows → Convert the dictionary into a CSV dataset

Notebook Link

DATA WRANGLING

Exploratory Data Analysis (EDA) was performed to find some patterns in the data and determine what would be the label for training supervised models.

We calculated the number of launches at each site

The number and occurence of mission outcome per orbit type was then calculated

We then created a landing outcome label from outcome column

The resulting dataset was then exported as CSV

Notebook Link

EDA WITH DATA VISUALIZATION

- Data visualization helps us understand data by curating it into a form that's easier to understand, highlighting the trends and outliers. The following types of charts were used in the visualization of the data.
- Scatter plots
  - Scatter plots were used to represent the relationship between two variables
  - Different sets of features were compared such as Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type and Payload vs. Orbit Type
- Bar chart
  - Bar charts were used makes it easy to compare values between multiple groups. The X axis represents a category, and the Y axis represents a discrete value.
  - Bar charts were used to compare the Success Rate for different Orbit Types
- Line chart
  - Line charts are useful for showing data trends over time
  - A line chart was used to show Success Rate over a certain number of Years

Notebook Link

EDA WITH SQL

We loaded the SpaceX dataset into a database and explored the data using SQL commands

Summary of SQL queries that were used for EDA:

- Names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Records showing the month names, failure landing outcomes in drone ship ,booster versions, launch site for year 2015
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Notebook Link

BUILD AN INTERACTIVE MAP WITH FOLIUM

- We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site
- In order to mark the success/failed launches for each site, marker clusters were used on the map
  - Red represents rocket launch failures while Green represents the successes
  - Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate
- Folium Markers were then used to show important landmarks nearest to SpaceX launch sites like railways, highways, cities and coastlines
- We calculated the distances between a launch site to its nearest landmarks and Polylines were used to connect them

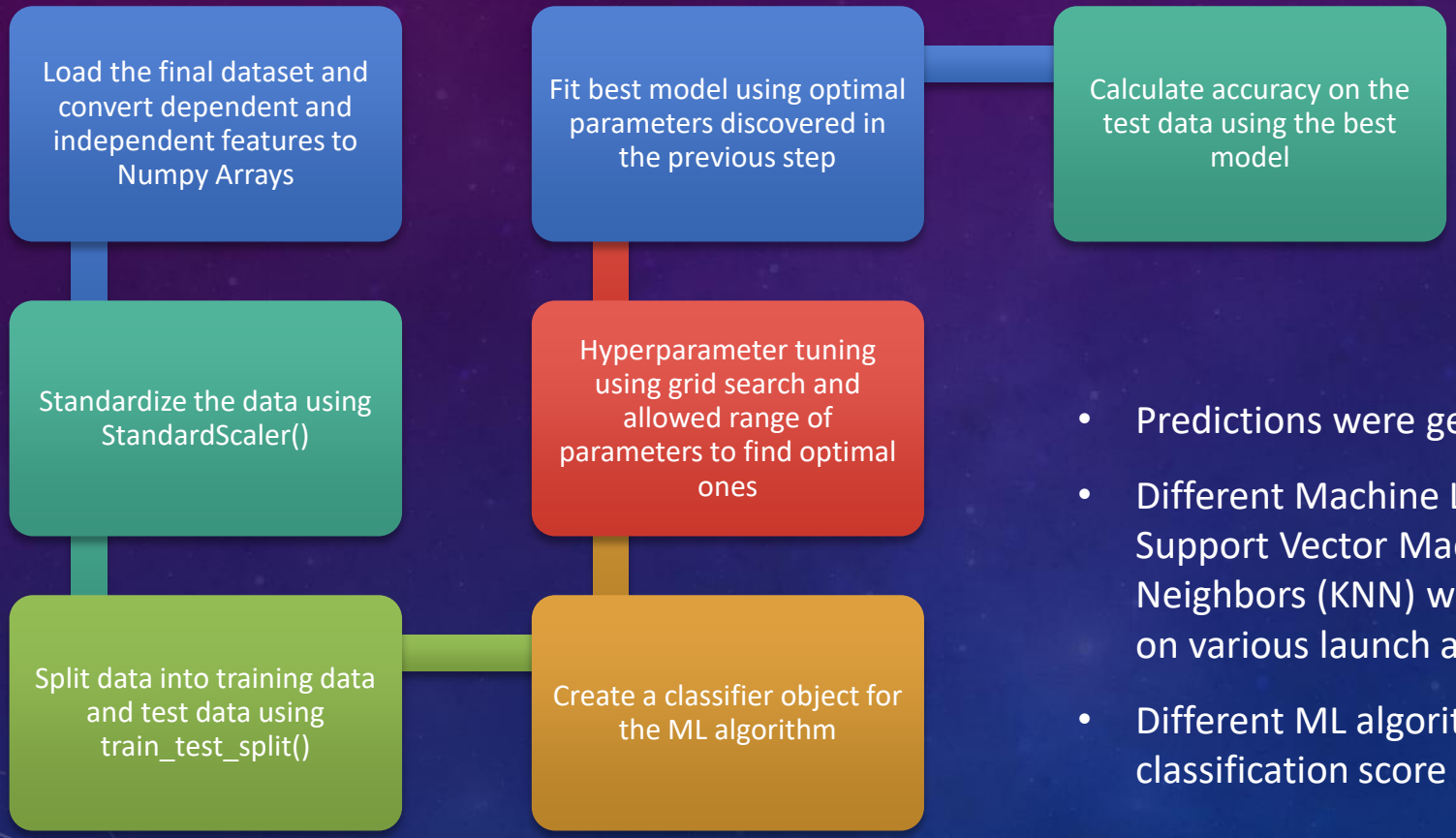
Notebook Link

BUILD A DASHBOARD WITH PLOTLY DASH

- We built an interactive dashboard using Plotly Dash
- Pie chart showing the total launches by site:
  - A drop down exists to select launch sites and tailor the chart accordingly. In the absence of any selection, the pie chart is displayed for all sites.
  - This chart is useful as you can visualize the distribution of landing outcomes across all launch sites or show the success rate of launches on individual sites
- Scatter chart showing the relationship with Outcome and Payload Mass (Kg) for the different booster version
  - In addition to launch site selection, a slider was added to select the range of payload mass. In the absence of any selection, the scatter chart is displayed for all sites and payload masses.
  - This chart is useful as you can visualize how different variables affect the landing outcomes

Notebook Link

PREDICTIVE ANALYSIS (CLASSIFICATION)



Notebook Link

- Predictions were generated using Scikit-learn
- Different Machine Learning algorithms such as Logistic Regression, Support Vector Machine (SVM), Decision Trees, K-Nearest Neighbors (KNN) were used to predict the launch outcome based on various launch and rocket related features
- Different ML algorithms were then compared using the classification score to find the best technique

RESULTS

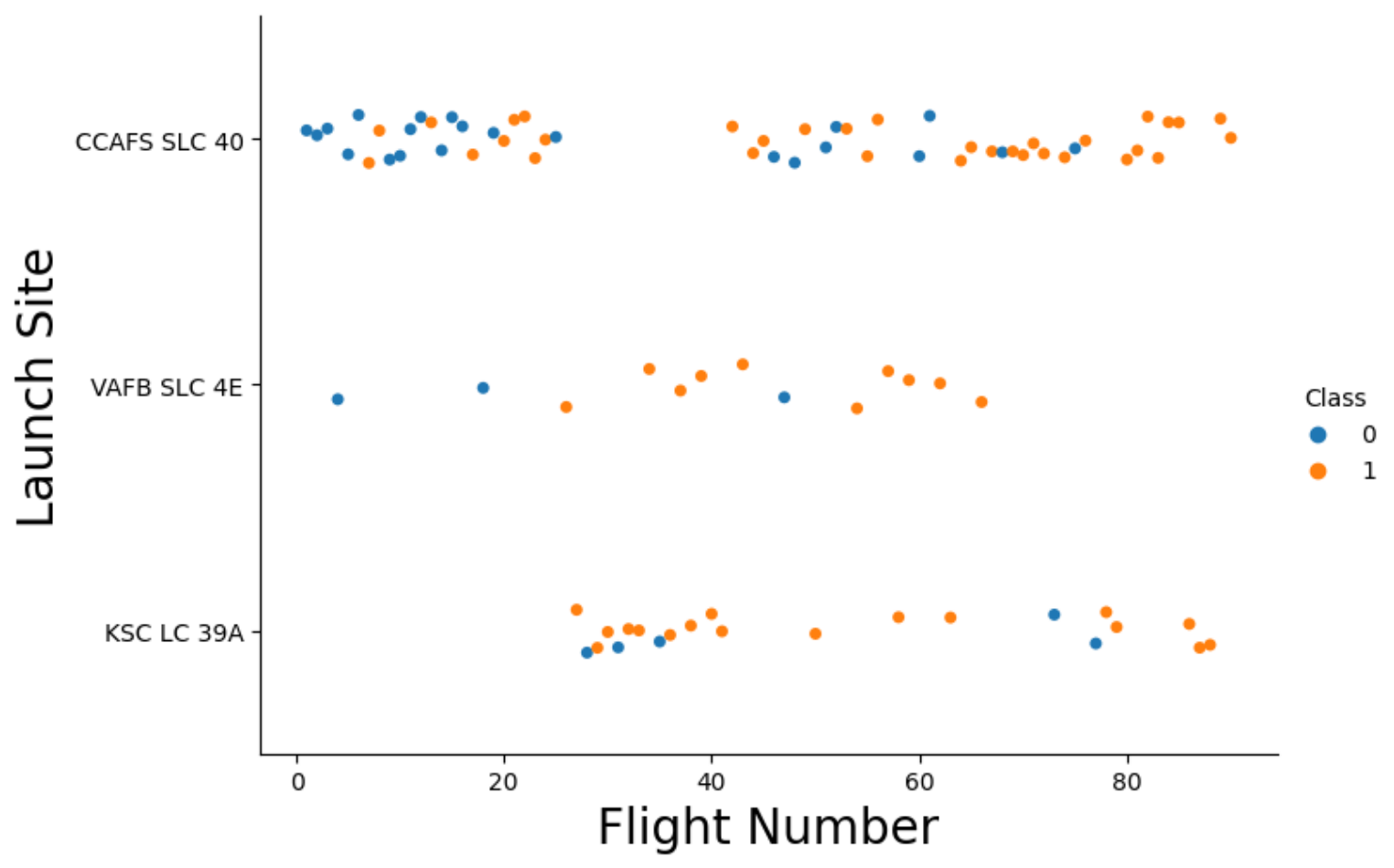
- EDA
  - Successful landing outcomes are positively correlated with number of flights
  - With heavy payloads the successful landing rate usually decreases but still good enough for Polar, LEO and ISS orbits
  - Successful landing outcomes have had a significant increase since the year 2013
- Interactive analytics demo in screenshots
  - All launch sites are located near the coastline away from populated areas, in order to save fuel and boosters and decrease any adverse effect due to crashes
  - Furthermore, the sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- Predictive analysis results
  - The machine learning models that were built, were able to predict the landing success of rockets with an accuracy score of 83.33%

INSIGHTS DRAWN FROM EDA

SECTION 2

Viraj Noorithaya

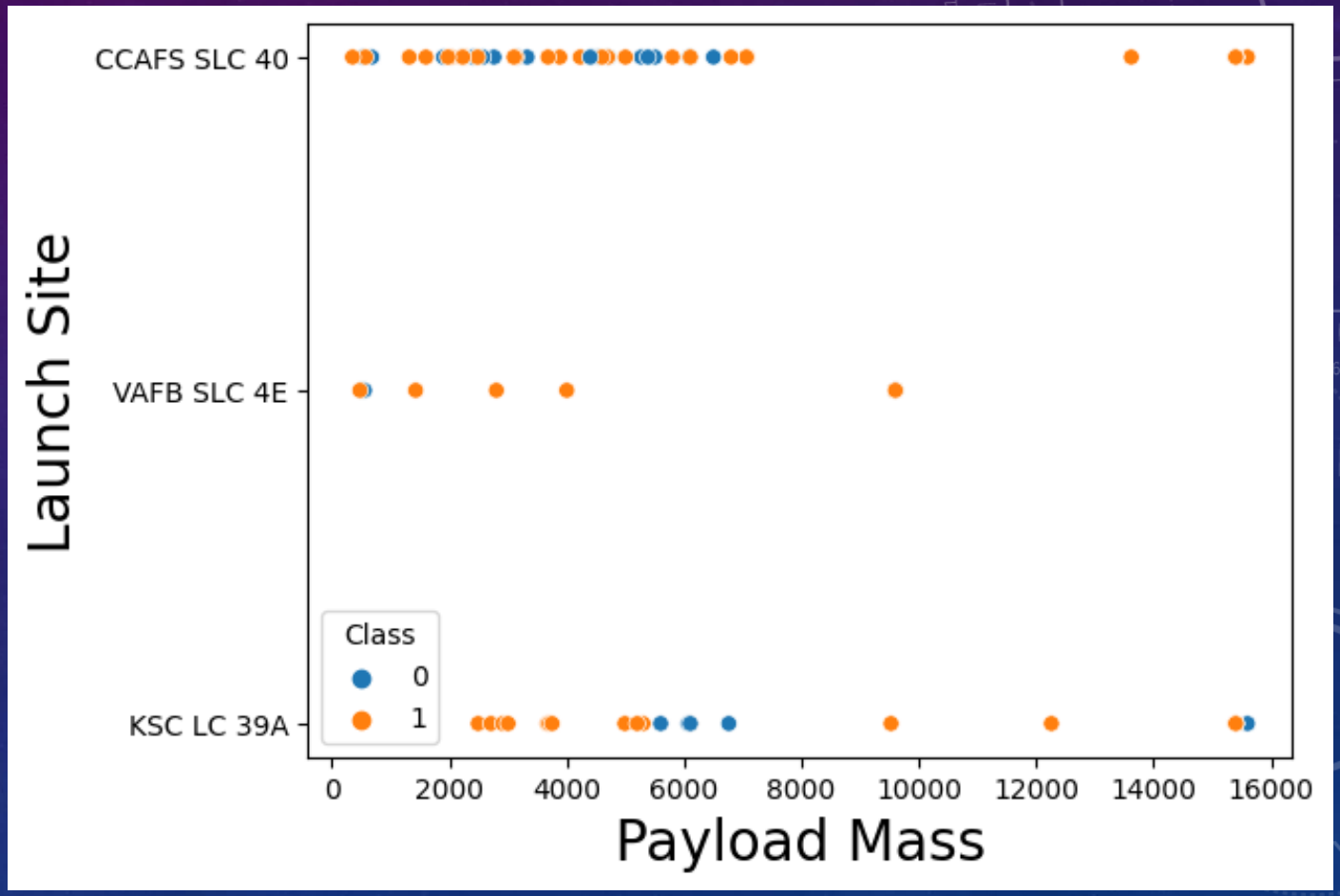
FLIGHT NUMBER VS. LAUNCH SITE



- We observe that the success rate increased as the number of flights increased
- The blue dots represent the successful launches while the orange dot represent unsuccessful launches.
- There seems to be an increase in successful flights after the 40th launch

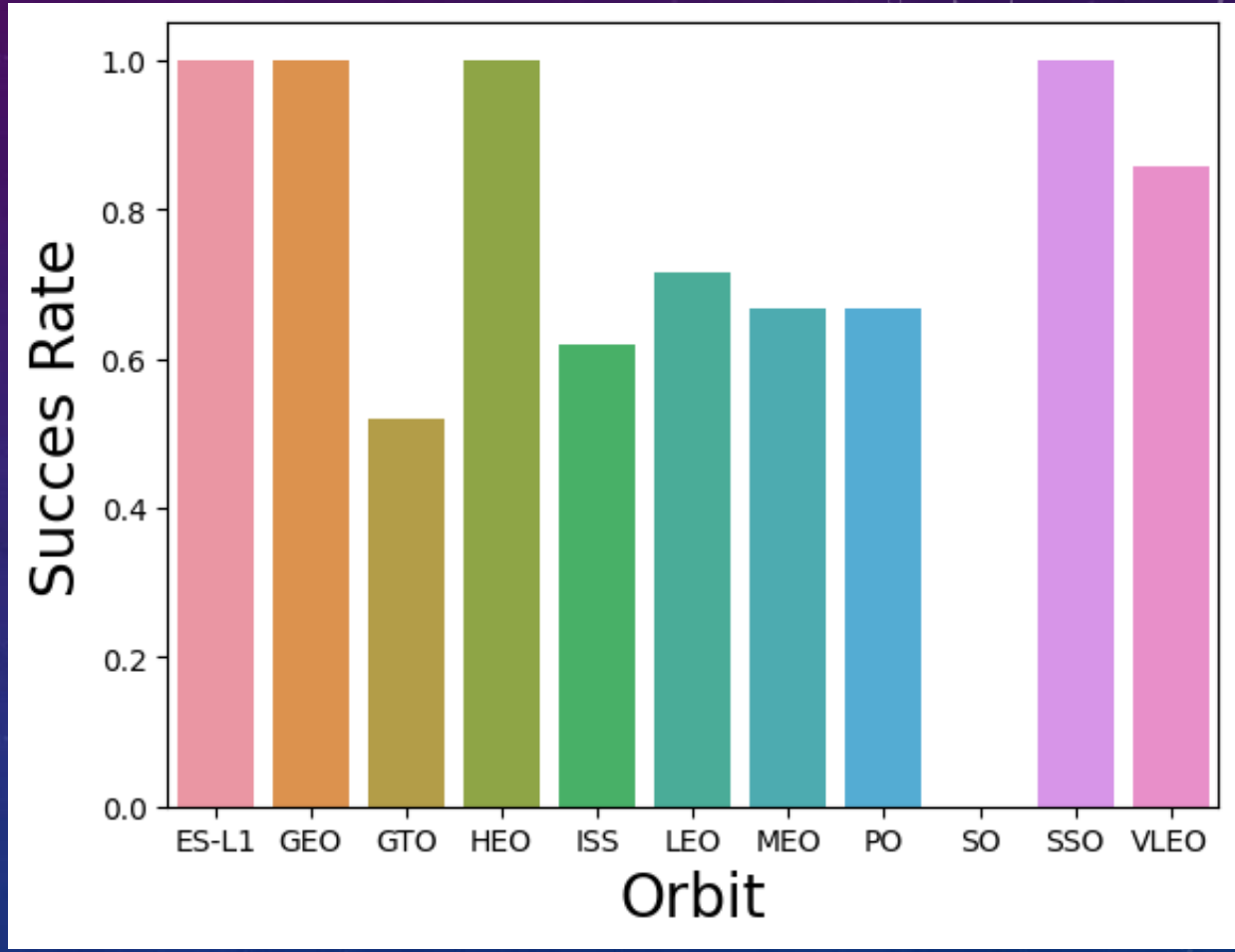
PAYLOAD VS. LAUNCH SITE

- The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass
- Due to weak correlation, there is no clear pattern to decide if the Launch Site is dependent on Payload mass for a success launch



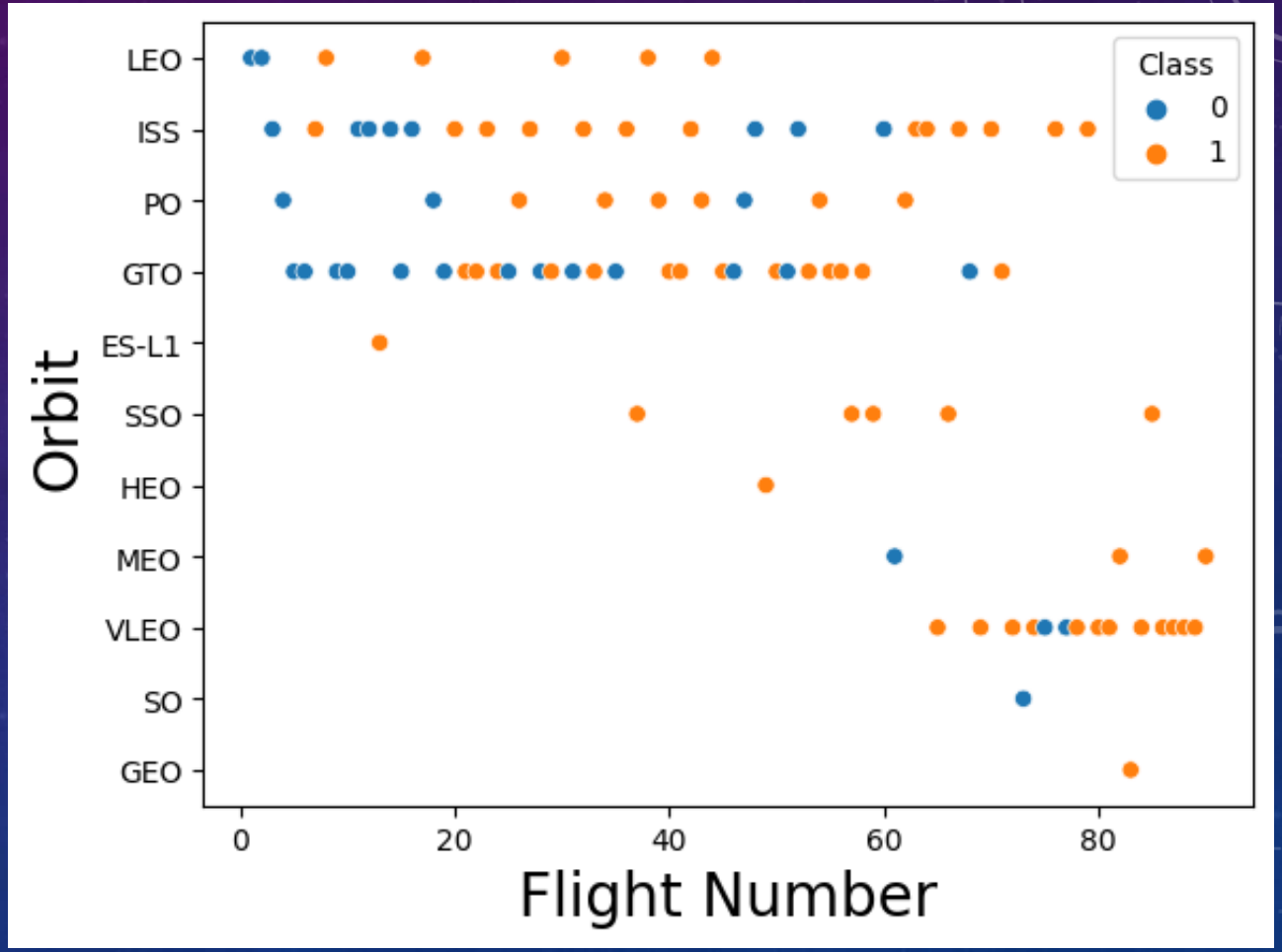
SUCCESS RATE VS. ORBIT TYPE

- Orbits SSO, HEO, GEO and ES-L1 have 100% success rates
- The SO orbit did not have any successful launches



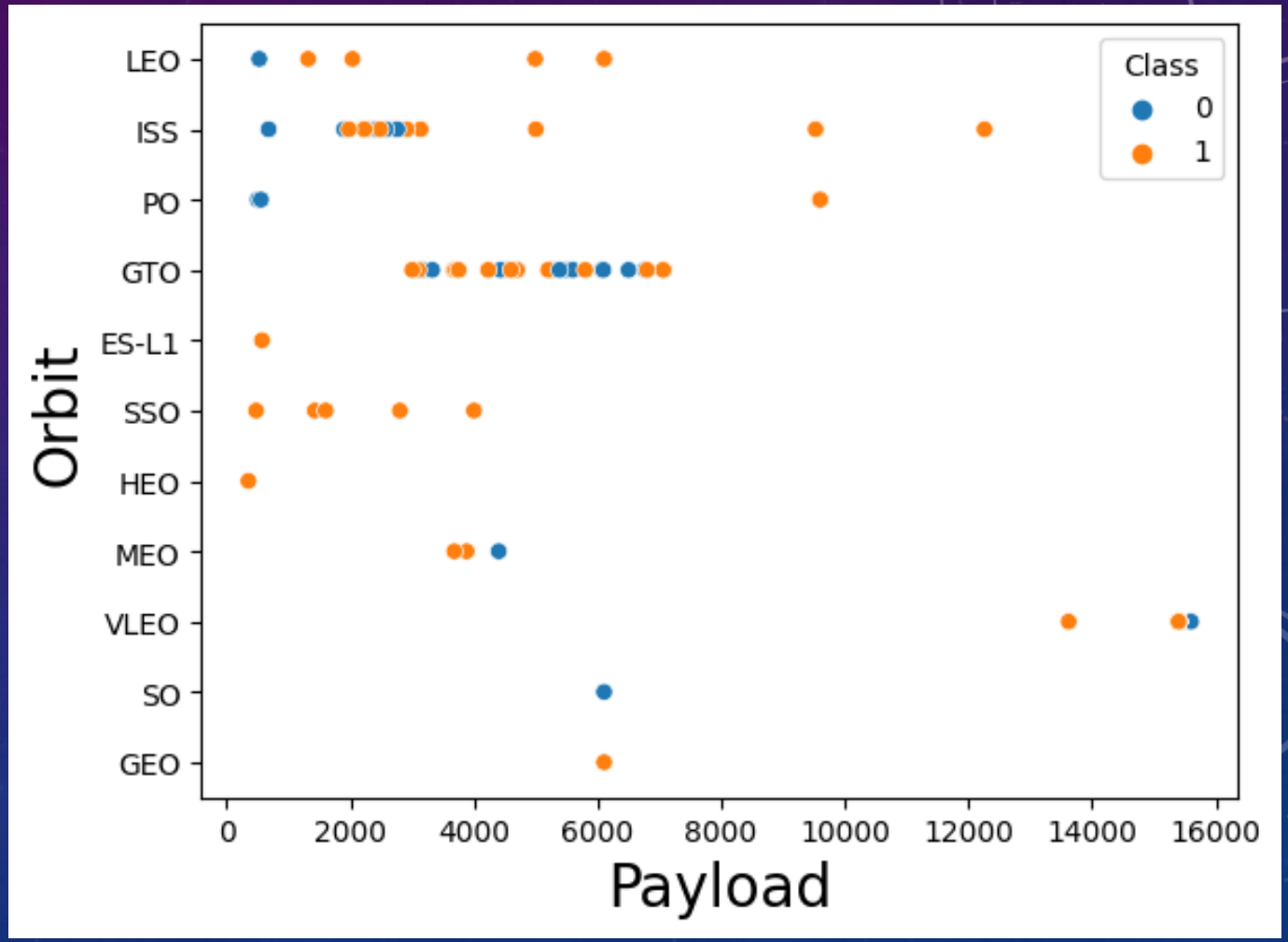
FLIGHT NUMBER VS. ORBIT TYPE

- In the LEO orbit, the launch success is positively correlated to the number of flights
- There seems to be no relationship between flight number in the GTO orbit
- The SSO orbit has a 100% success rate for all flights
- Flight outcomes seem to have improved for all Orbits after 40 launches



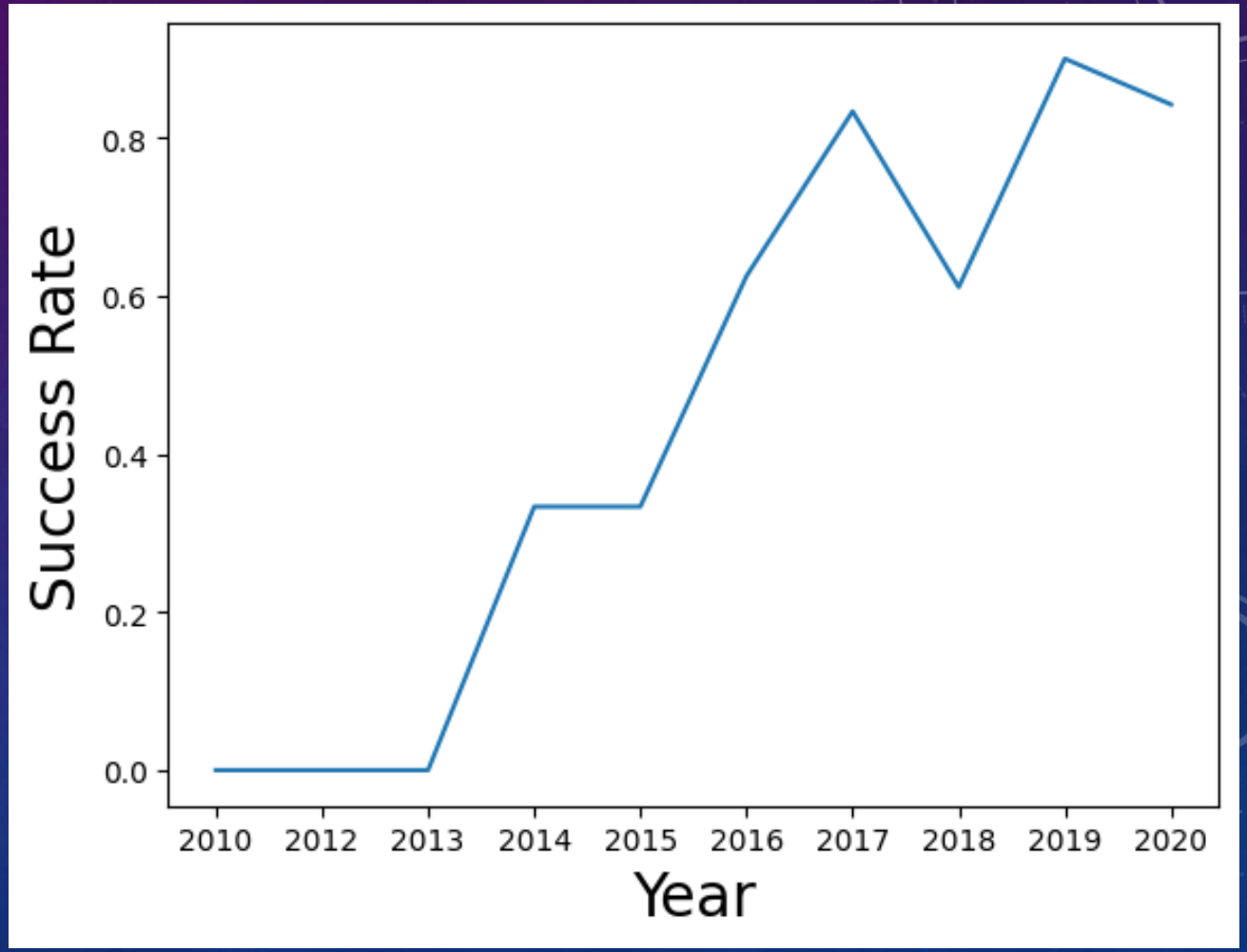
PAYLOAD VS. ORBIT TYPE

- Heavier payload has positive impact on LEO, ISS and P0 orbit
- However, it has negative impact on MEO and VLEO orbit
- GTO orbit seem to depict no relation between the 2 attributes
- SO, GEO and HEO orbit need more data points to see any pattern or trend



LAUNCH SUCCESS YEARLY TREND

- We see an increase in landing success rate as the years pass
- There is however a dip in 2018 as well as in 2020



ALL LAUNCH SITE NAMES

- The DISTINCT clause was used to return unique rows from the “launch_site” column using SQL
- The names of the 4 launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

```
%%sql

select distinct Launch_Site from SPACEXTBL
```

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

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

LAUNCH SITE NAMES BEGIN WITH 'CCA'

```
1  %%sql
2
3  select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

The LIMIT and LIKE clauses were used to display only the top five results where the "launch_site" name starts with 'CCA'

TOTAL PAYLOAD MASS

Task 3

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

```
%%sql

select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTBL where Customer = "NASA (CRS)"
```

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

total_payload_mass
45596

- The sum() function was used to the calculate the total payload carried by boosters from NASA (CRS) from the Payload Mass column
- Total payload mass is 45,596 kgs

AVERAGE PAYLOAD MASS BY F9 V1.1

Task 4

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

```
1  %%sql
2  select avg(PAYLOAD_MASS__KG_) as avg_payload_mass from SPACEXTBL
   where Booster_Version = "F9 v1.1"
```

Python

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

avg_payload_mass
2928.4

- The AVG() function was used to the calculate the average payload carried by booster version after WHERE clause was used to filter "F9 v1.1" booster versions

- The average payload mass carried by F9 v1.1 was 2928.4 kg.

FIRST SUCCESSFUL GROUND LANDING DATE

Task 5

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

Hint:Use min function

```
%%sql

select min(Date) as first_success_landing_date from SPACEXTBL where
"Landing _Outcome" = "Success (ground pad)";
```

✓ 0.1s Python

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

first_success_landing_date
01-05-2017

- The min(Date) function was used to find the date of the first successful landing outcome on ground pad
- The WHERE clause ensured that the results were filtered to match only when the 'landing_outcome' column is 'Success (ground pad)'

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

Task 6

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

```
%%sql

select distinct(Booster_Version) from SPACEXTBL where "Landing
_Outcome" = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000
and 6000;
```

✓ 0.2s Python

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

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

We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

Task 7

List the total number of successful and failure mission outcomes

```
1  %%sql
2
3  select Mission_Outcome, count(*) as mission_count from SPACEXTBL
   group by Mission_Outcome;
```

✓ 0.1s Python

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

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

- The COUNT() function is used to count the number of occurrences of different mission outcomes with the help of the GROUPBY clause applied to the “mission_outcome” column

- There have been 99 successful mission outcomes out of 101 missions.

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), PAYLOAD_MASS__KG_ from SPACEXTBL where
PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL) order by
Booster_Version;
```

✓ 0.1s Python

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

The MAX() function was used in a subquery to retrieve a list of boosters which have carried the maximum payload mass

2015 LAUNCH RECORDS

Task 9

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

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql

select substr(Date, 4, 2) as month, "Landing _Outcome" as Landing_Outcome,
Booster_Version from SPACEXTBL where substr(Date, 7, 4) = "2015" and "Landing _Outcome"
= "Failure (drone ship)";
```

17] ✓ 0.2s Python

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

month	Landing_Outcome	Booster_Version
01	Failure (drone ship)	F9 v1.1 B1012
04	Failure (drone ship)	F9 v1.1 B1015

We used a combinations of the WHERE clause and SUBSTR function to filter for failed landing outcomes in drone ship, their booster versions, and landing outcome for year 2015

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql

select "Landing _Outcome" as Landing_Outcome, count(*) as count from SPACEXTBL where
Date between "04-06-2010" and "20-03-2017" and "Landing _Outcome" like "Success%" group
by "Landing _Outcome" order by count desc;
```

✓ 0.0s Python

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

Landing_Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	6

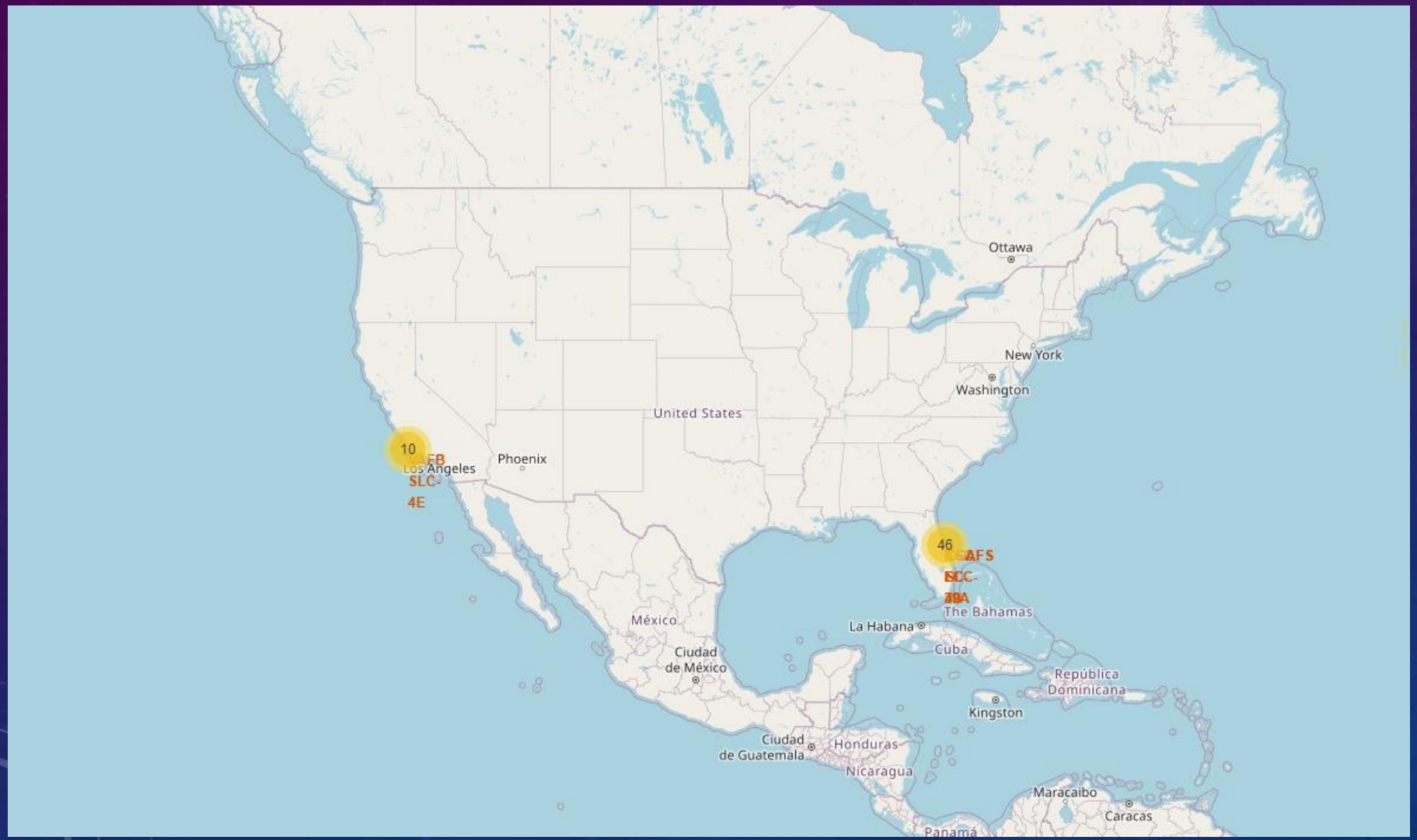
- Used WHERE clause to filter for successful landing outcomes BETWEEN 2010-06-04 to 2017-03-20
- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order

LAUNCH SITE PROXIMITY ANALYSES

SECTION 3

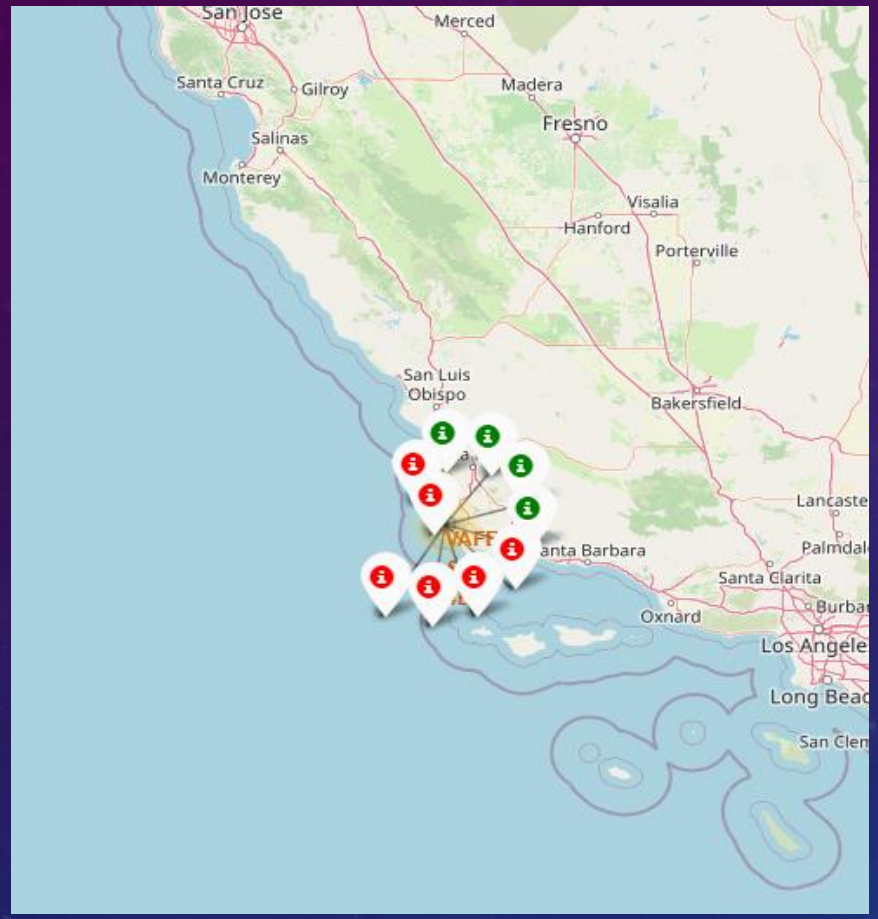
Viraj Noorithaya

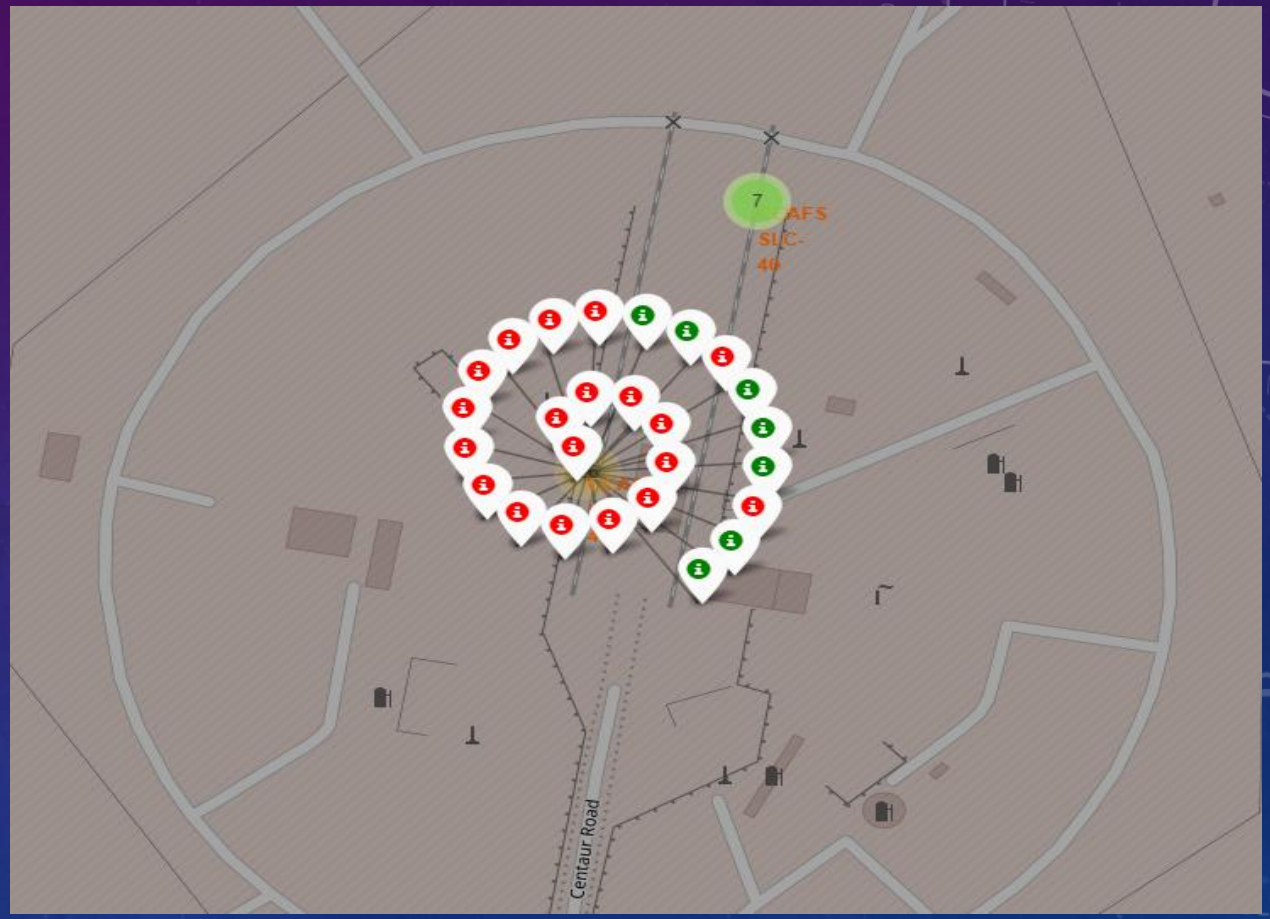
ALL LAUNCH SITES



- The yellow markers are indicators of where the locations of all the SpaceX launch sites are situated in the US.
- The launch sites have been strategically placed near the coast

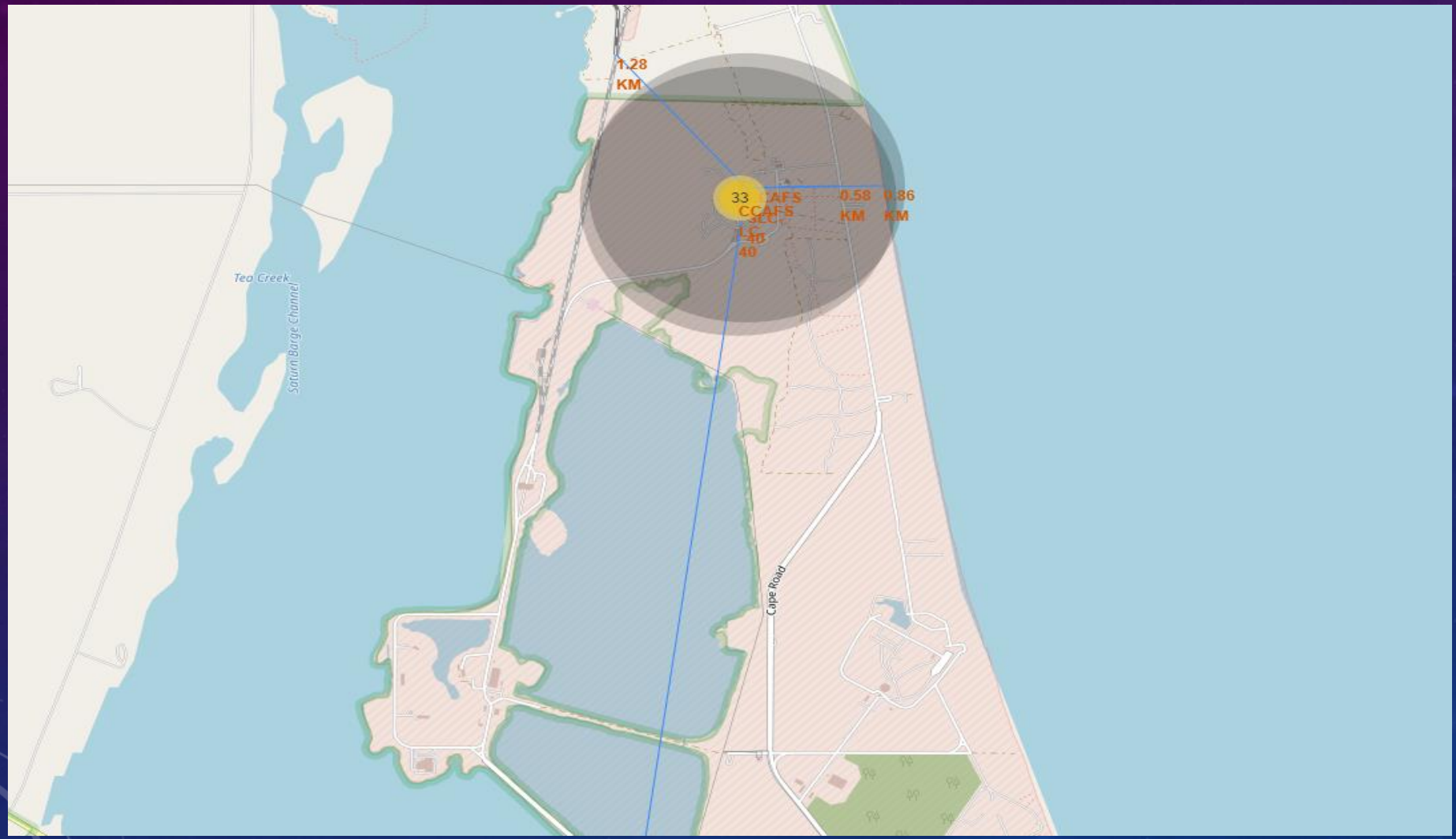
SUCCESS AND FAILURE OUTCOMES BY LAUNCH SITES





The successful launches are represented by a green marker while the red marker represents failed rocket launches

LAUNCH SITE PROXIMITIES



The generated map shows that the selected launch site is close to a highway for transportation of personnel and equipment

The launch site is also close to the coastlines for launch failure testing.

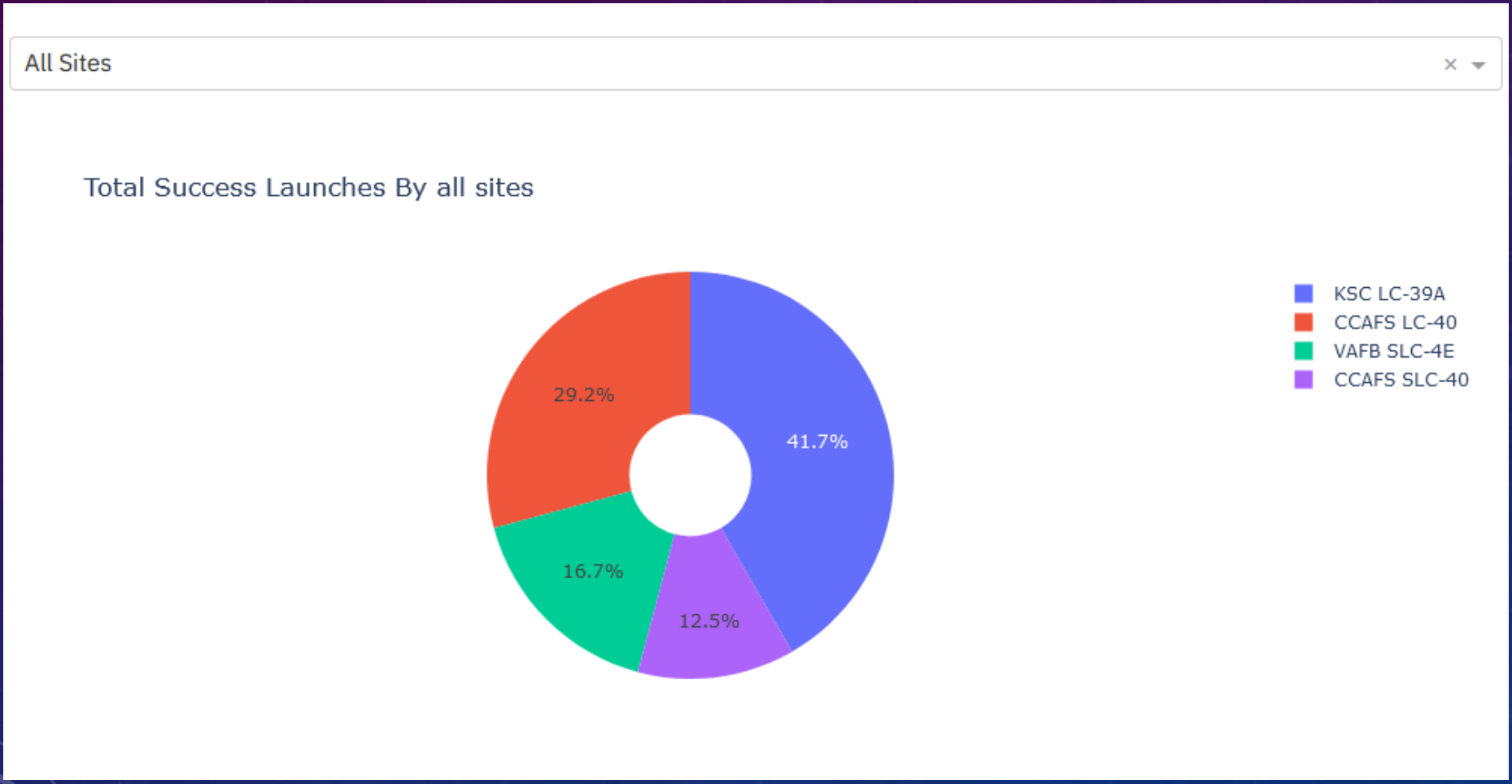
The launch sites also maintain a certain distance from the cities

BUILDING A DASHBOARD WITH PLOTLY DASH

SECTION 4

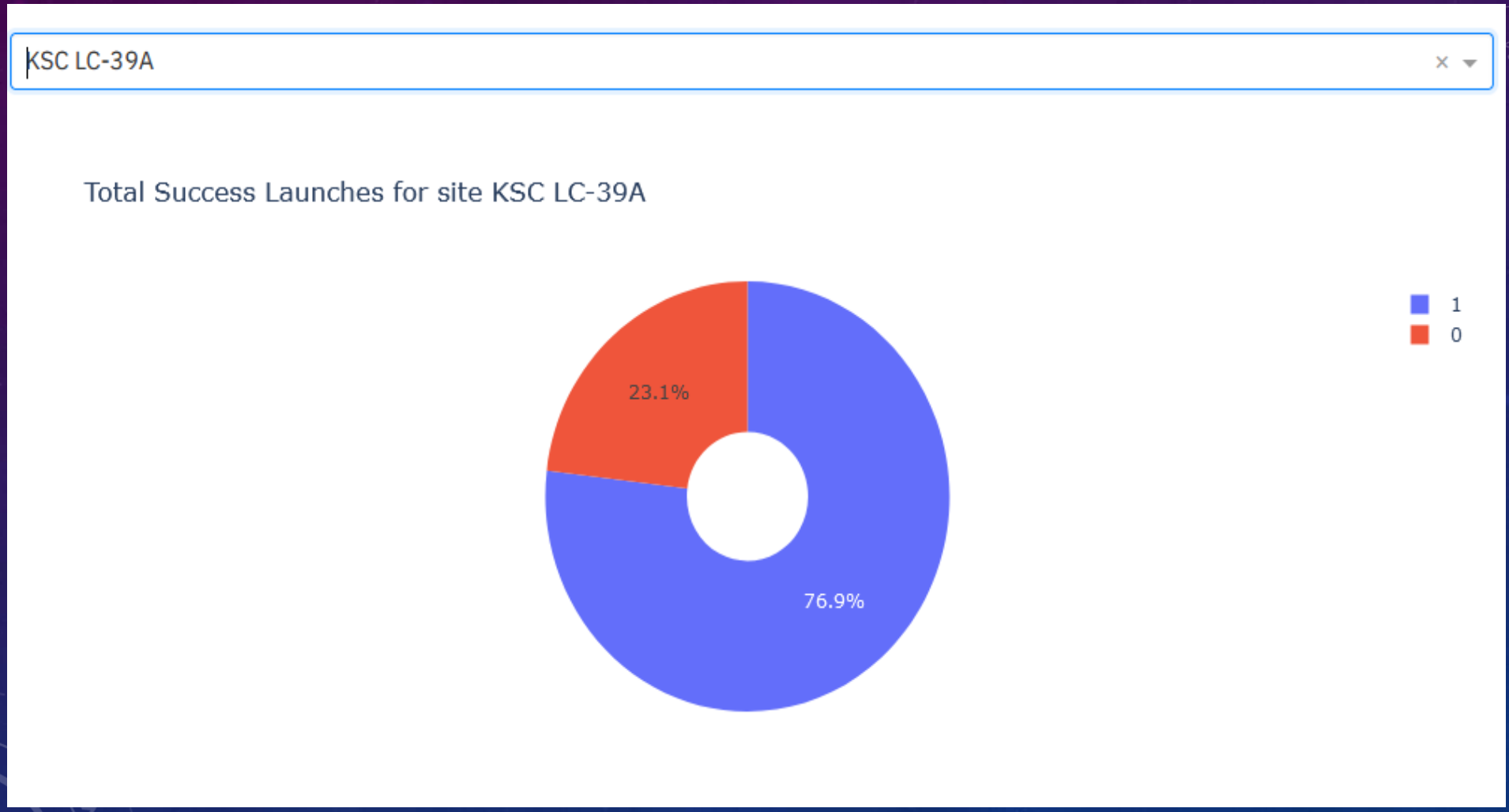
Viraj Noorithaya

TOTAL SUCCESSFUL LAUNCHES BY SITE



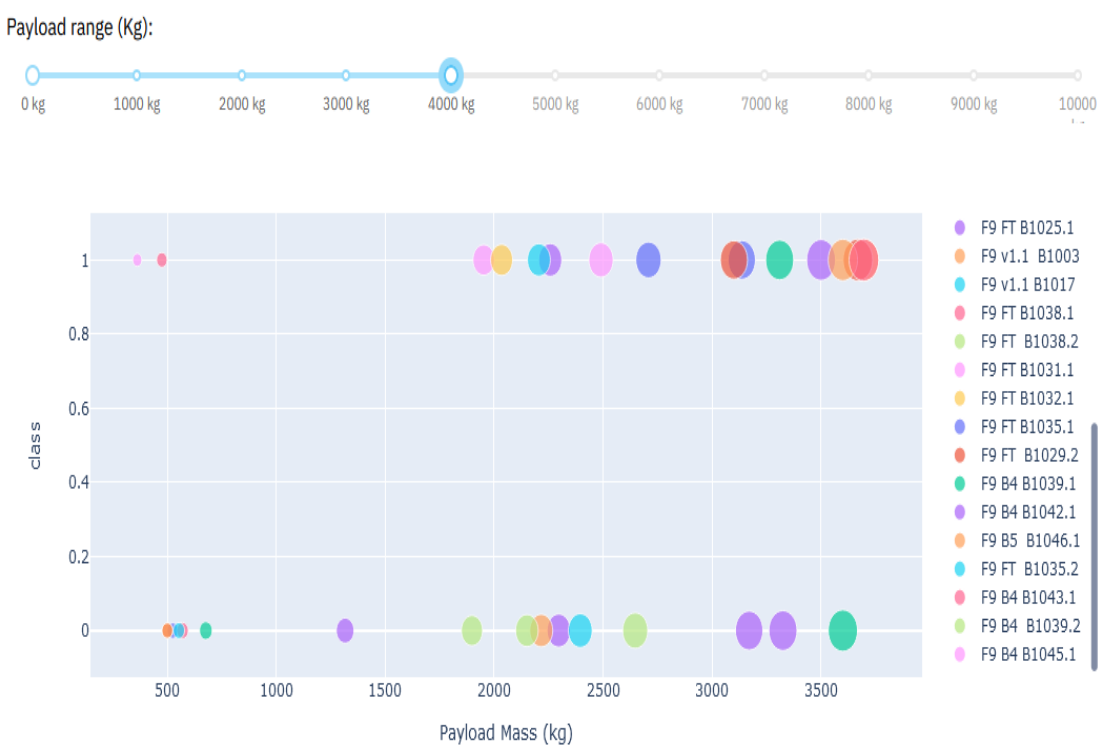
Site KSC LC-39A has the largest successful launches as well the highest launch success rate.

LAUNCH SITE WITH THE HIGHEST LAUNCH SUCCESS RATIO



The KSLC-39A has the highest success rate with 76.9%

PAYLOAD MASS VS. LAUNCH SUCCESS FOR ALL SITES



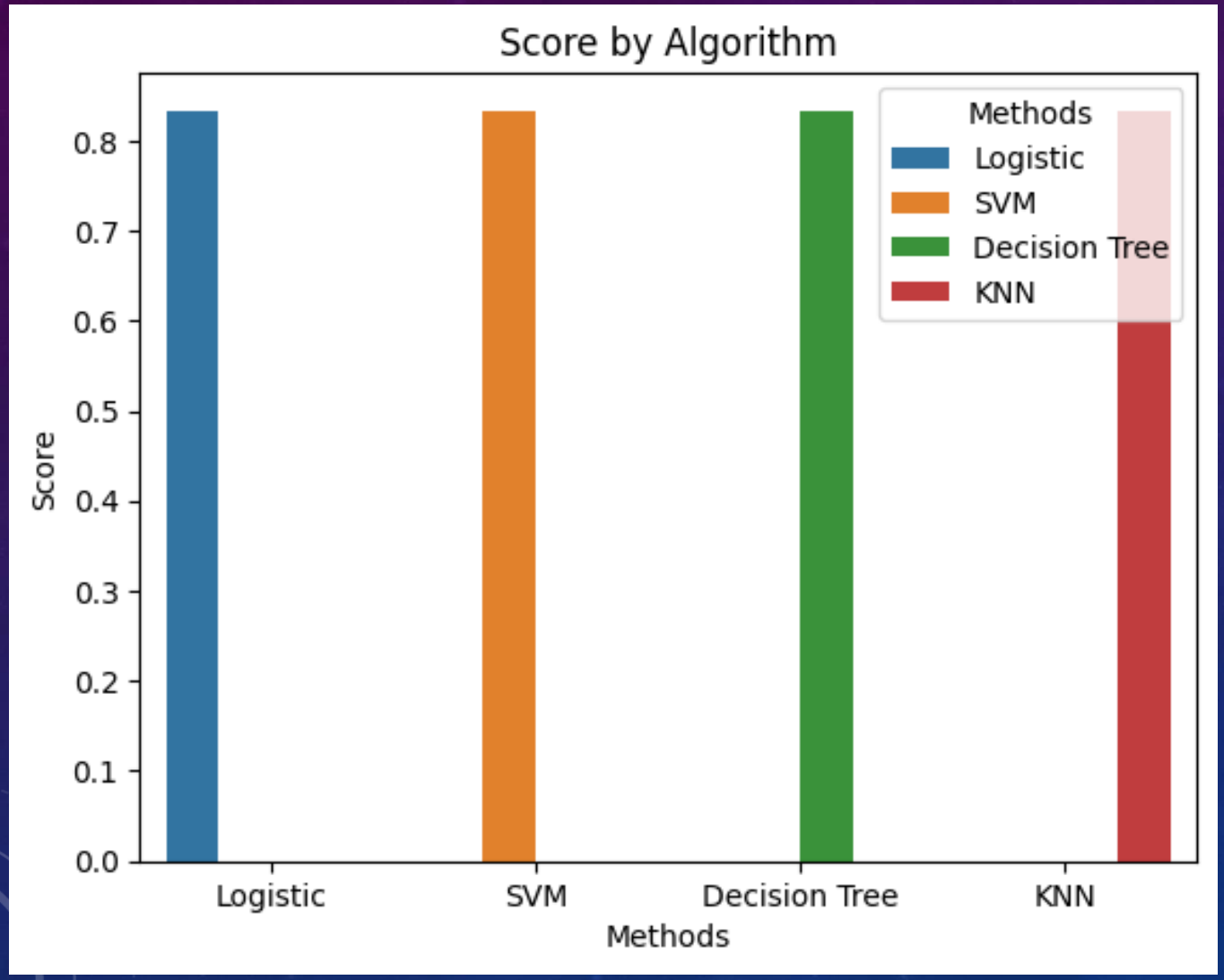
We observer that success for lower weight payloads (0k to 4k kgs) is higher than higher weight payloads (4k to 10k kgs)

PREDICTIVE ANALYSES (CLASSIFICATION)

SECTION 5

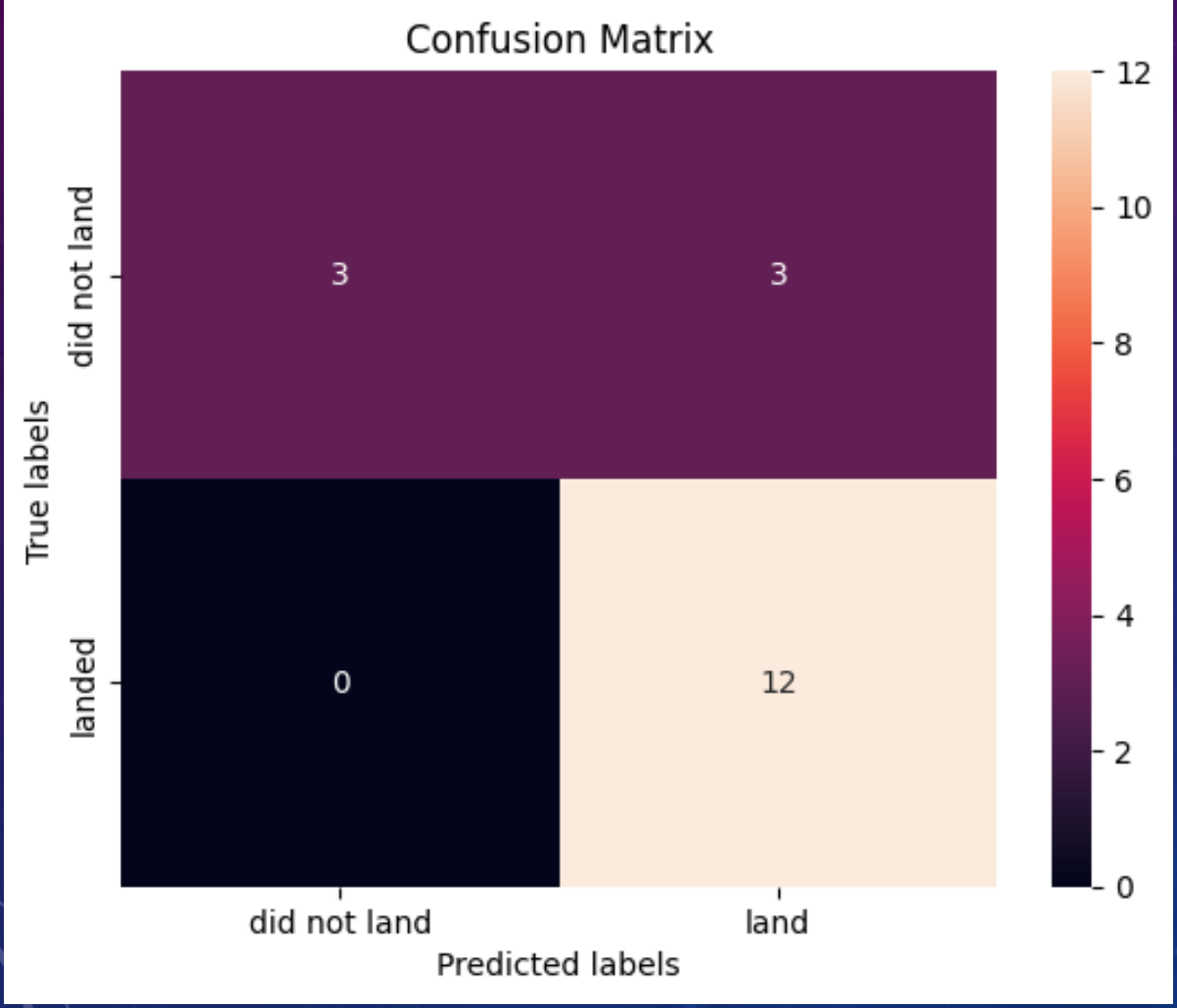
Viraj Noorithaya

CLASSIFICATION ACCURACY



- All 4 techniques showed the same accuracy score of 83.33%
- We decided to proceed with Logistic regression because of its simplicity interpreting model results with same accuracy when compared to other techniques

CONFUSION MATRIX



- The model predicted 12 successful landings correctly when the True label was successful (True Positive)
- The model also predicted 3 successful landings when the True label was unsuccessful landing (False Positive) which needs to be improved
- The model generally predicted successful landings

CONCLUSIONS

- In order to compete with SpaceX, it was crucial to analyze their data. Through this process, a general picture of their success methods was produced.
- All launch sites are located near the coastline away from populated areas, in order to save fuel and boosters and decrease any adverse effect due to crashes. Furthermore, the sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- Successful landing outcomes are positively correlated with number of flights. From 2013 onwards, the success rate of rocket landings significantly increased.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate. KSC LC-39A had the most successful launches of any sites.
- Low weighted payloads perform better than the heavier payloads
- The machine learning models that were built, were able to predict the landing success of rockets with an accuracy score of 83.33%

APPENDIX

- History of SpaceX
- Course Page
- Specialization Page
- Code Repository
- Library Documentation
  - Requests
  - Beautiful Soup
  - Numpy
  - Pandas
  - Scikit-learn

THANK YOU!