SPACEX

coursera IBM

Falcon 9 Launch Analysis

Proteek Chatterjee

13th August 2025

C O N T E N T

EXECUTIVE SUMMARY

Methodologies:

- Data Collections
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Interactive Map with Folium
- Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Results:

- Exploratory Data Analysis Results
- Predictive Analysis Results

INTRODUCTION

Background:

SpaceX's Falcon 9 rocket reusability is a major factor in reducing the cost of space missions. Being able to reuse the first stage of the rocket saves millions of dollars, which is a key differentiator in SpaceX's competitive edge. The company's ability to predict whether a first-stage landing will succeed is vital for cost forecasting and operational planning.

To predict if SpaceX will reuse the first stage, we need to consider several factors such as payload mass, launch site, number of flights, and orbits. These variables are likely to affect the success of the landing, either directly or indirectly.

Questions to be Answered:

How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

Each of these variables could have a different influence on the probability of success for a landing.

Does the rate of successful landings increase over the years?

Given SpaceX's continuous improvements in rocket technology, it would be expected that the success rate would improve over time as more data is collected, and lessons are learned from previous missions.

What is the best algorithm that can be used for binary classification in this case?

The problem is essentially a binary classification problem: whether the first stage will land successfully or not (i.e., "success" or "failure").

METHODOLOGY

Data Collections

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Data Wrangling

- Filter the Data
- Deal Missing Values
- One Hot Encoding

Exploratory Data Analysis using Data Visualization and SQL

Interactive Visual Analytics using Folium and Plotly Dash

Predictive Analysis (Classification)

Building, evaluating classification models to find the best model.

DATA COLLECTION

The dataset was compiled using a **hybrid approach** combining:
API Requests – Data was retrieved from the official **SpaceX REST API**, which provided detailed technical and operational parameters of each launch.

Web Scraping – Additional details were extracted from the **SpaceX Wikipedia page**, specifically from the launch history table, to supplement missing attributes from the API.

This two-source approach ensured **completeness and accuracy**, allowing for a more comprehensive launch analysis.

Columns obtained via SpaceX REST API:
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Columns obtained via Wikipedia Web Scraping:
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

DATA COLLECTION- SPACEX API

Requesting rocket launch data from SpaceX API
↓
Decoding the response using .json() and converting to DataFrame using .json_normalize()
↓
Requesting needed launch information via custom functions
↓
Constructing obtained data into a dictionary
↓
Creating a DataFrame from the dictionary
↓
Filtering the DataFrame to only include Falcon 9 launches
↓
Replacing missing Payload Mass values with the column's mean
↓
Exporting the cleaned data to CSV

GitHub Link:
[Data Collection API]()

DATA COLLECTION - WEB SCRAPING

Requesting Falcon 9 launch data from Wikipedia
↓
Creating a BeautifulSoup object from the HTML response
↓
Extracting column names from the HTML table header
↓
Collecting the data by parsing HTML tables
↓
Constructing obtained data into a dictionary
↓
Creating a DataFrame from the dictionary
↓
Exporting the data to CSV

GitHub Link:
Data Collection with Web Scrapping

DATA WRANGLING

Landing Outcome Classification

The dataset contains multiple cases where the booster did not land successfully. Landing outcomes are categorized based on both the landing location and success status:

- **True Ocean** – Booster successfully landed in a designated ocean region.
- **False Ocean** – Attempted ocean landing, but unsuccessful.
- **True RTLS** – Booster successfully landed at a ground pad (*Return to Launch Site*).
- **False RTLS** – Attempted ground pad landing, but unsuccessful.
- **True ASDS** – Booster successfully landed on a drone ship (*Autonomous Spaceport Drone Ship*).
- **False ASDS** – Attempted drone ship landing, but unsuccessful.

For modeling purposes, these categorical outcomes were converted into **binary training labels**:

- 1 – Successful landing.
- 0 – Unsuccessful landing.

GitHub Link:
[Data Wrangling]

Perform exploratory data analysis and determine training labels

↓

Calculate the number of launches at each site

↓

Calculate the number and occurrence of each orbit type

↓

Calculate the number and occurrence of mission outcomes per orbit type

↓

Create a landing outcome label from the Outcome column

↓

Export the data to CSV

EDA WITH DATA VISUALIZATION

To explore relationships and trends in the dataset, several charts were plotted:

- **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**
- **Success Rate Yearly Trend**

Scatter plots were used to visualize relationships between continuous variables, helping to identify potential correlations that could be valuable for machine learning models.
Bar charts were employed to compare discrete categories, illustrating the relationship between specific categories and their corresponding measured values.
Line charts were utilized to display time-series trends, such as changes in success rates over the years.

GitHub Link:
[EDA with Data Visualization]

E D A W I T H S Q L

A series of SQL queries were executed to analyze and extract specific insights from the dataset:

1. Retrieve unique launch site names from all recorded space missions.
2. Display five records where launch site names begin with the string "CCA".
3. Calculate the total payload mass carried by boosters launched under the NASA (CRS) program.
4. Compute the average payload mass carried by booster version *F9 v1.1*.
5. Identify the date of the first successful ground pad landing outcome.
6. List boosters with successful drone ship landings carrying payloads between 4000 and 6000 kg.
7. Count total successful and failed mission outcomes.
8. Identify booster versions that have carried the maximum payload mass.
9. Retrieve failed drone ship landings in 2015, including booster versions and launch site names.
10. Rank landing outcomes (e.g., *Failure (drone ship)*, *Success (ground pad)*) by count between 2010-06-04 and 2017-03-20, in descending order.

GitHub Link:
[EDA with SQL]

INTERACTIVE MAP WITH FOLIUM

Markers of All Launch Sites
Placed a marker with a circle, popup label, and text label for **NASA Johnson Space Center** using its latitude and longitude coordinates as the initial reference location.
Added markers with circles, popup labels, and text labels for **all launch sites**, displaying their geographical positions and illustrating their proximity to the Equator and nearby coastlines.

Coloured Markers for Launch Outcomes
Utilized **Marker Clusters** to plot coloured markers representing launch outcomes — green for successful launches and red for failed launches — enabling visual identification of launch sites with relatively high or low success rates.

Distance Mapping to Nearby Features
Drew coloured lines to indicate distances between the **KSC LC-39A** launch site (as an example) and nearby infrastructure, including railway lines, highways, coastlines, and the closest city.

GitHub Link:
[Interactive Map with Folium]()

DASHBOARD WITH PLOTLY DASH

Launch Sites Dropdown List
Implemented a dropdown menu to allow users to select a specific launch site for analysis.

Success Launches Pie Chart
Created a pie chart that displays the total number of successful launches across all sites when no selection is made. When a specific launch site is selected, the chart updates to show **Success vs. Failure** counts for that site.

Payload Mass Range Slider
Added an interactive slider to filter launches based on a selected payload mass range.

Payload Mass vs. Success Rate Scatter Chart
Developed a scatter plot illustrating the relationship between payload mass and launch success rates, categorized by booster versions.

GitHub Link:
SpaceX Dash App

P R E D I C T I V E   A N A L Y S I S

Creating a NumPy array from the column "Class" in data
↓
Standardizing the data with StandardScaler (fit and transform)
↓
Splitting the data into training and testing sets using train_test_split
↓
Creating a GridSearchCV object with cv=10 to find best parameters
↓
Applying GridSearchCV on:
- Logistic Regression (LogReg)
- Support Vector Machine (SVM)
- Decision Tree
- K-Nearest Neighbors (KNN)

↓
Calculating model accuracy using .score() on test data
↓
Examining the confusion matrix for all models
↓
Finding the best model using F1-score

GitHub Link:
[Machine Learning Prediction]

RESULTS

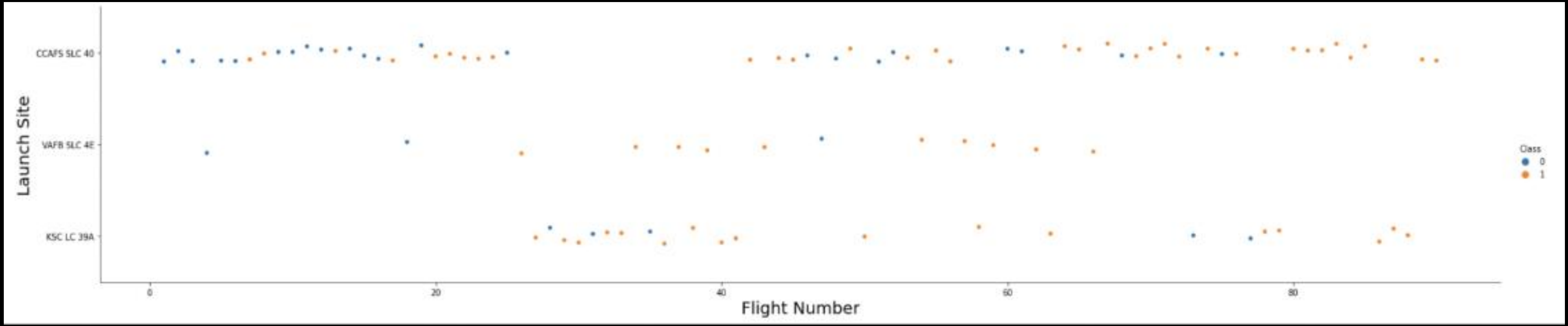
Exploratory Data Analysis Results

Interactive Analytics Results

Predictive Analysis Results

EXPLORATORY DATA ANALYTICS WITH VISUALIZATION

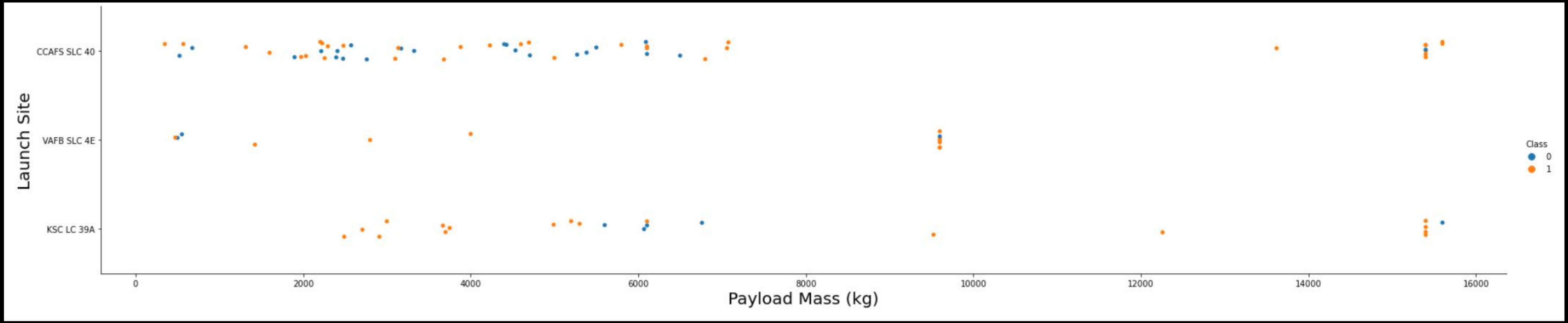
FLIGHT NUMBER VS LAUNCH SITE



Explanations:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

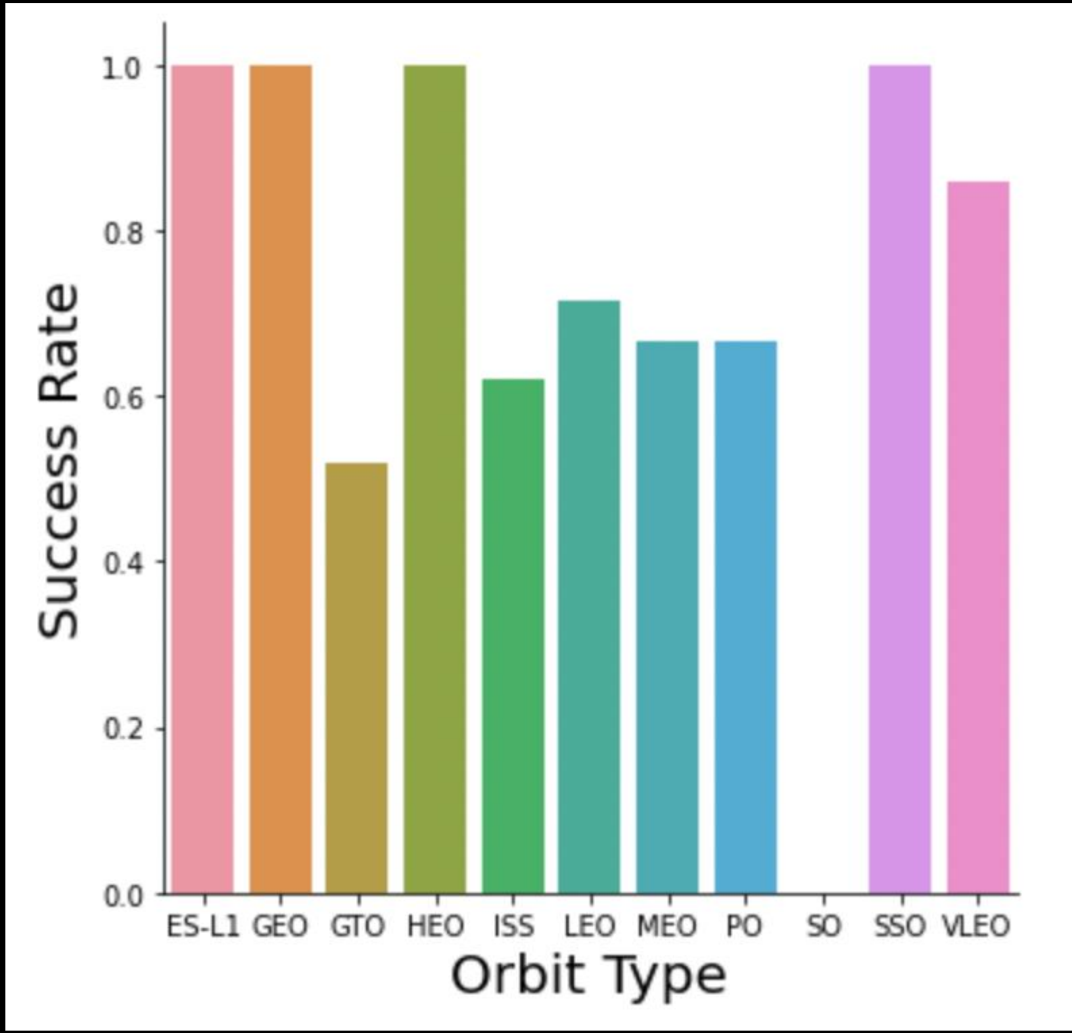
PAYLOAD VS LAUNCH SITE



Explanations:

- For every launch site the higher the payload mass, the higher the success rate.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

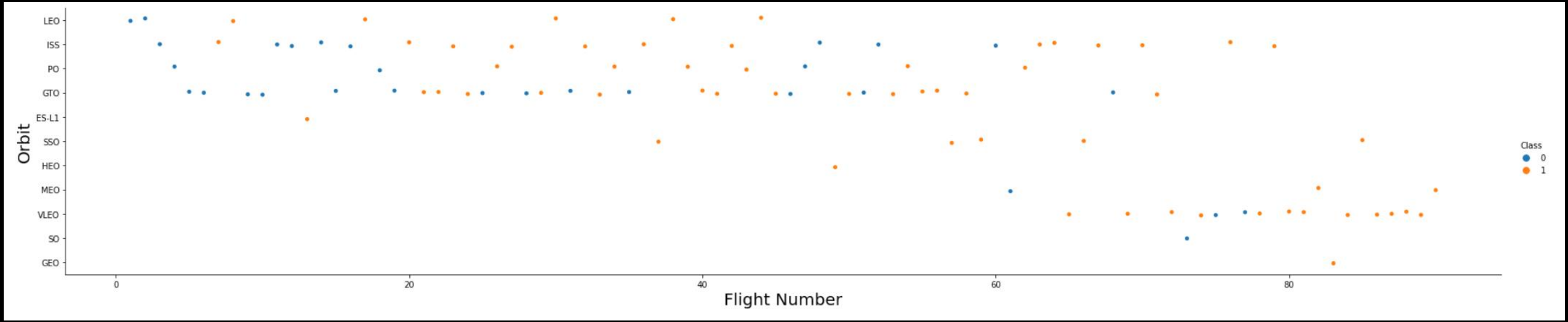
SUCCESS RATE VS ORBIT TYPE



Explanations:

- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: - SO
- Orbits with success rate between 50% and 85%: - GTO, ISS, LEO, MEO, PO, VLEO

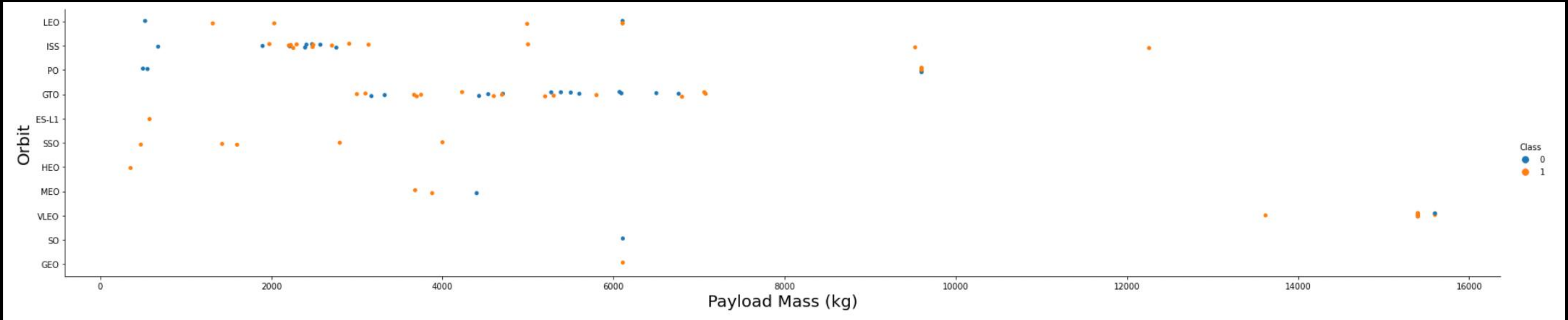
FLIGHT NUMBER VS ORBIT TYPE



Explanations:

- 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.

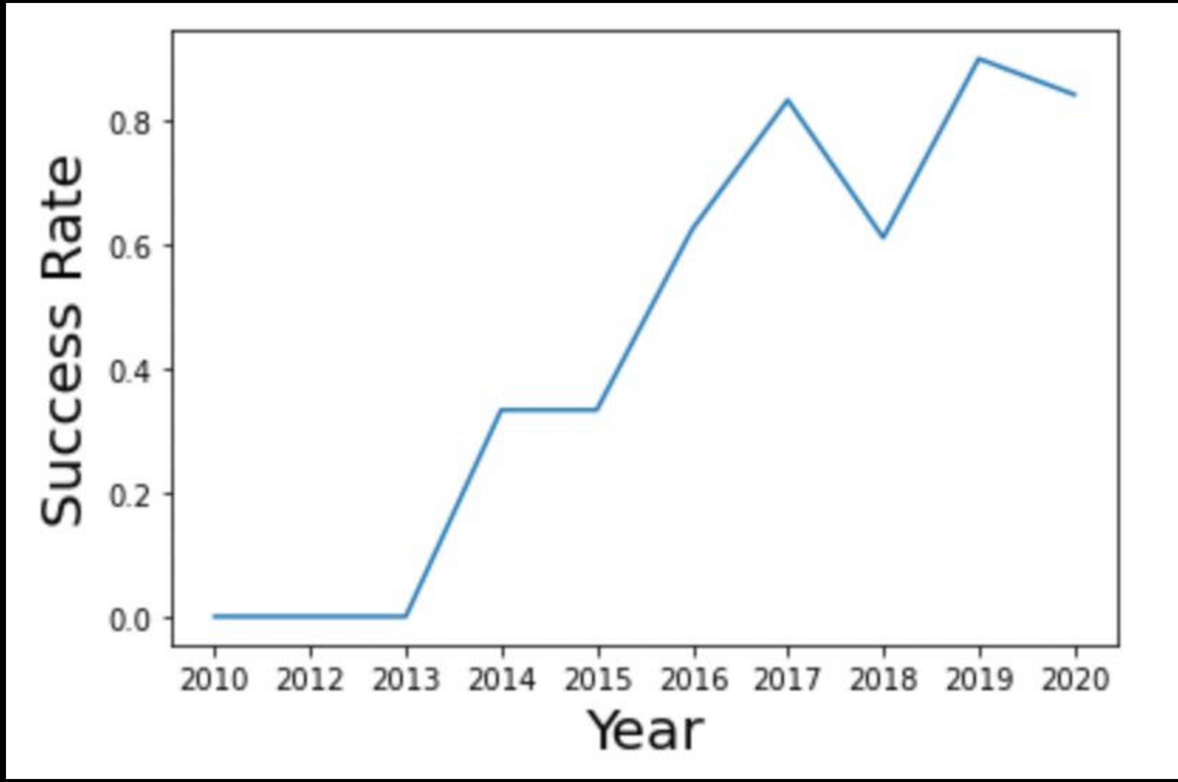
PAYLOAD MASS VS ORBIT TYPE



Explanations:

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

LAUNCH SUCCESS YEARLY TREND



Explanations:

- The success rate since 2013 kept increasing till 2020.

EXPLORATORY DATA ANALYTICS WITH SQL

ALL LAUNCH SITE NAMES

```
%sql select distinct launch_site from SPACEXTABLE;
```

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

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

Explanations:

Displaying the names of the unique launch sites in the space mission.

LAUNCH SITE NAMES BEGIN WITH `CCA`

```
%sql select * from SPACEXTABLE 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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanations:

Displaying 5 records where launch sites begin with the string 'CCA'.

TOTAL PAYLOAD MASS

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

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

total_payload_mass
45596

Explanations:

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

AVERAGE PAYLOAD MASS BY 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

Explanations:

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

FIRST SUCCESSFUL GROUND LANDING DATE

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

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

first_successful_landing
2015-12-22

Explanations:

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

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
%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.
```

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

Explanations:

Listing the 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

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

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

Explanations:
Listing the total number of successful and failure mission outcomes.

BOOSTERS CARRIED MAXIMUM PAYLOAD

```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

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

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

Explanations:

Listing the names of the booster versions which have carried the maximum payload mass.

2015 LAUNCH RECORDS

```
%%sql
SELECT
    CASE substr(Date, 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS Month,
    Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE
    substr(Date, 1, 4) = '2015' AND
    Landing_Outcome LIKE '%Failure%' AND
    Landing_Outcome LIKE '%drone ship%';
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanations:

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

```
%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

Explanations:

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.

INTERACTIVE MAP WITH FOLIUM

ALL LAUNCH SITES' LOCATION MARKERS ON A GLOBAL MAP

Explanations:

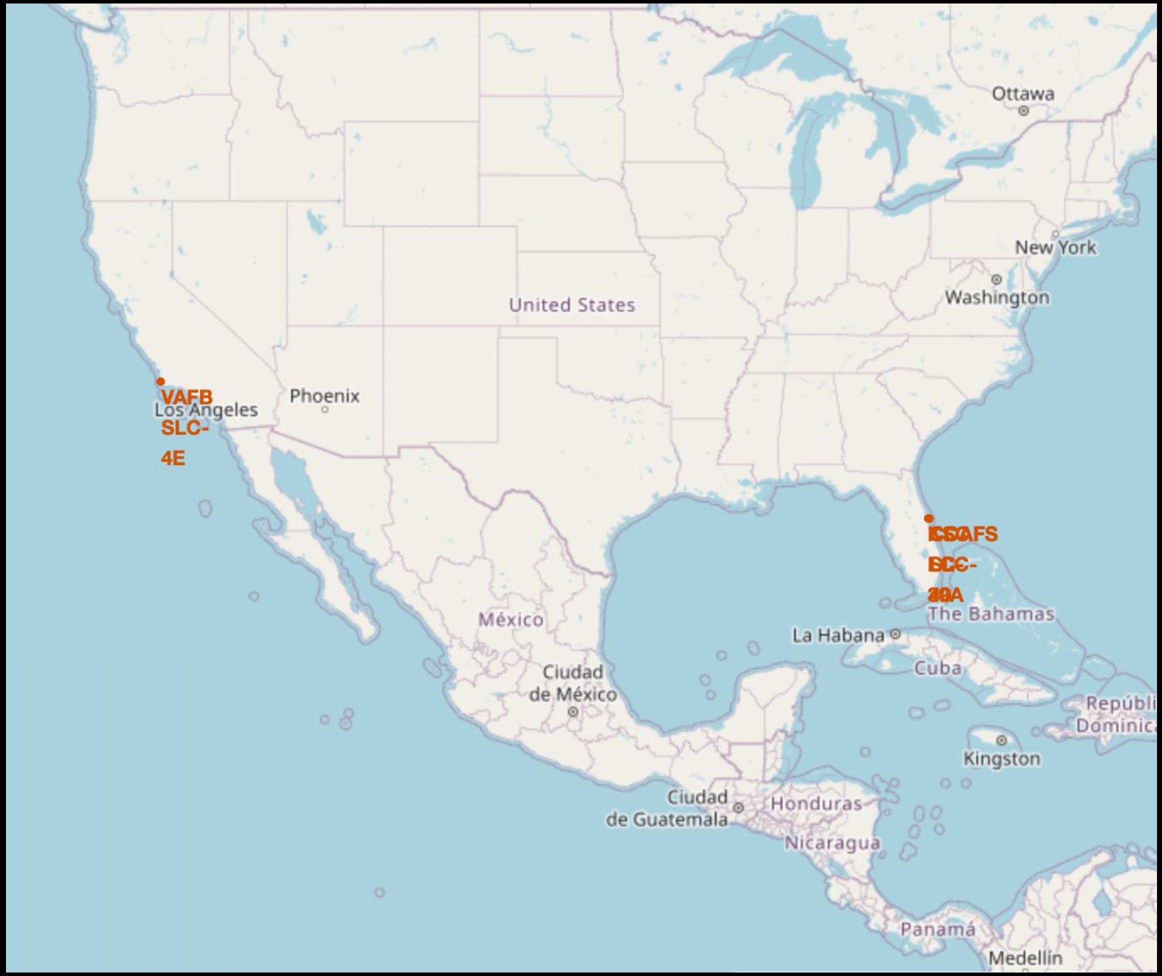
Proximity to the Equator:

- The Earth rotates fastest at the equator — approximately **1670 km/h**.
- A spacecraft launched from the equator inherits this velocity, giving it a **boost in speed** that helps it more easily reach and stay in orbit.

Proximity to the Coast:

- Most launch sites are located **near oceans or large bodies of water**.
- Rockets are typically launched **over the ocean** to:
  Minimize risk to human life in case of explosions or falling debris.

Coastal locations provide **large uninhabited areas** to safely manage failed launches or discarded rocket stages.



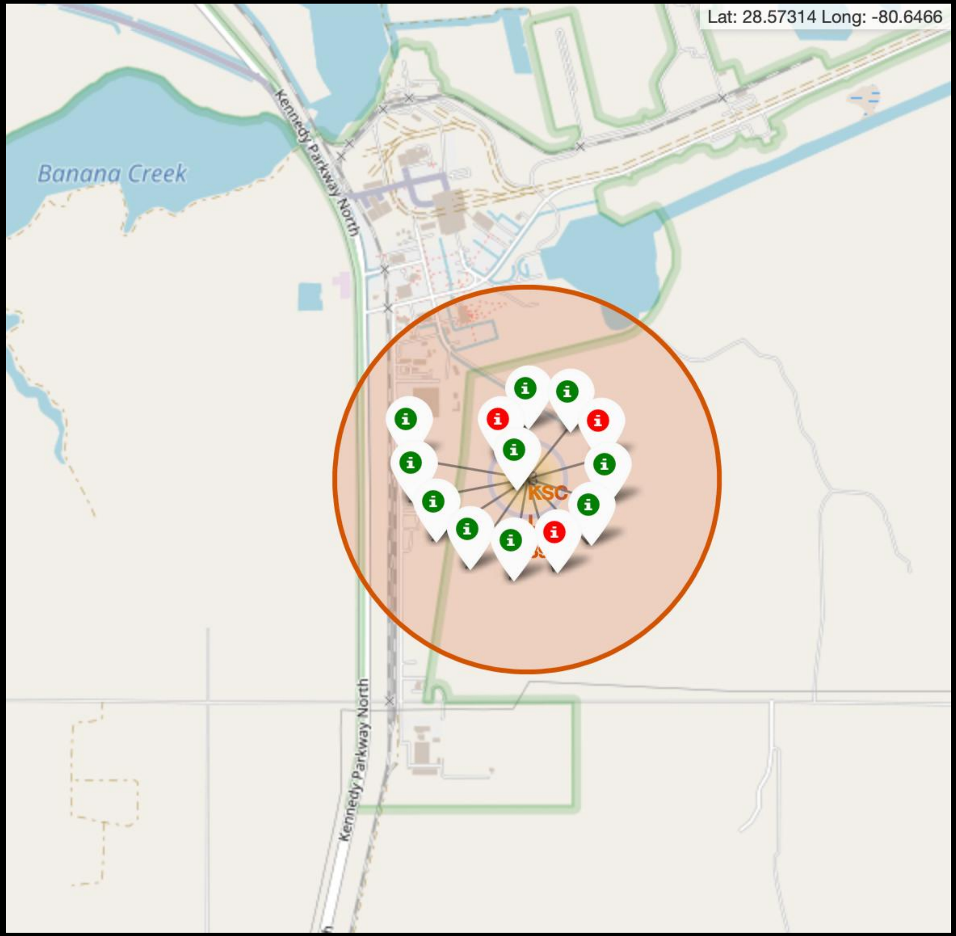
COLOUR-LABELED LAUNCH RECORDS ON THE MAP

Explanations:

From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

Green Marker = Successful Launch
Red Marker = Failed Launch

Launch Site KSC LC-39A has a very high Success Rate.

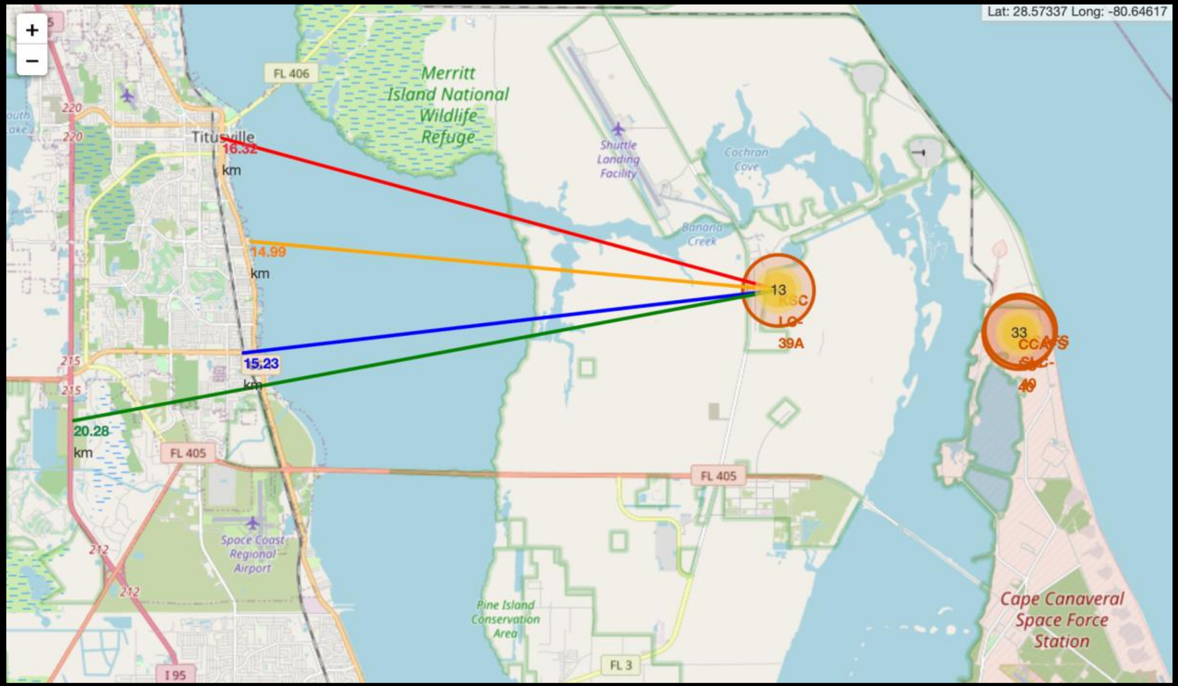


DISTANCE FROM THE LAUNCH SITE KSC LC-39A TO ITS PROXIMITIES

Explanations:

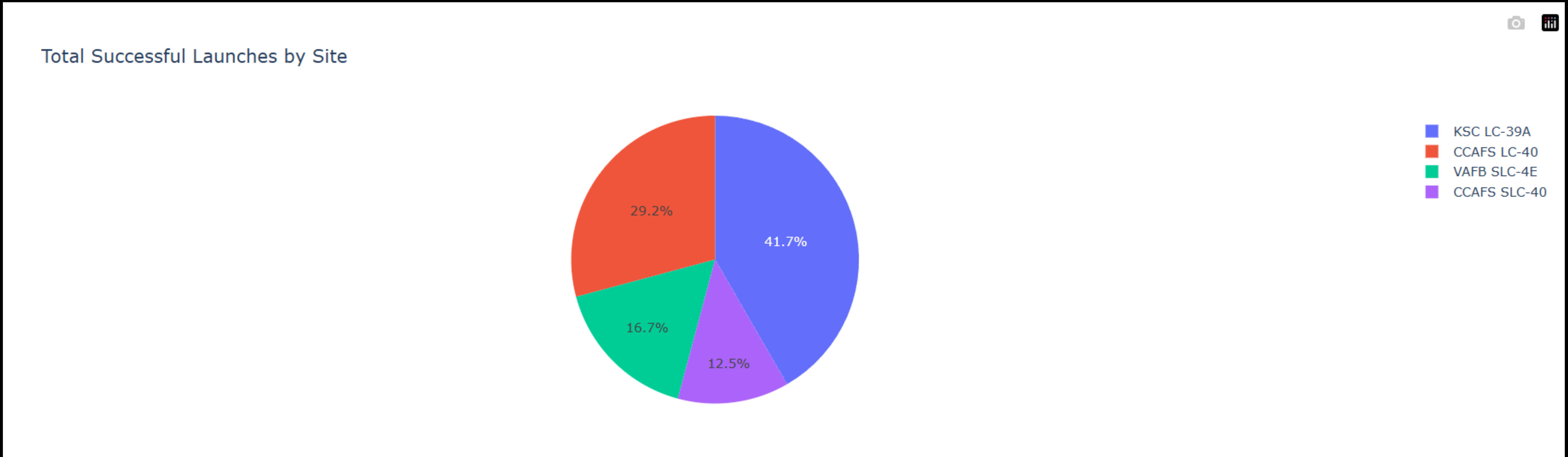
From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

- relatively close to railway (15.23 km)
- relatively close to highway (20.28 km)
- relatively close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relatively close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



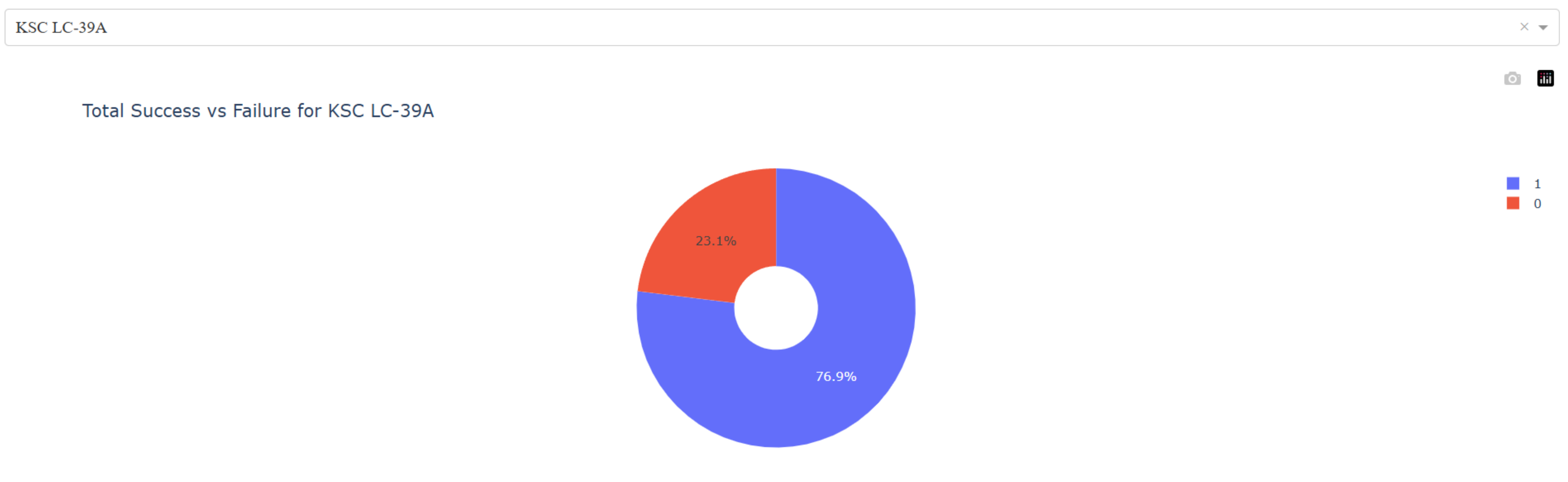
BUILD A DASHBOARD WITH PLOTLY DASH

LAUNCH SUCCESS COUNT FOR ALL SITES



Explanations:
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

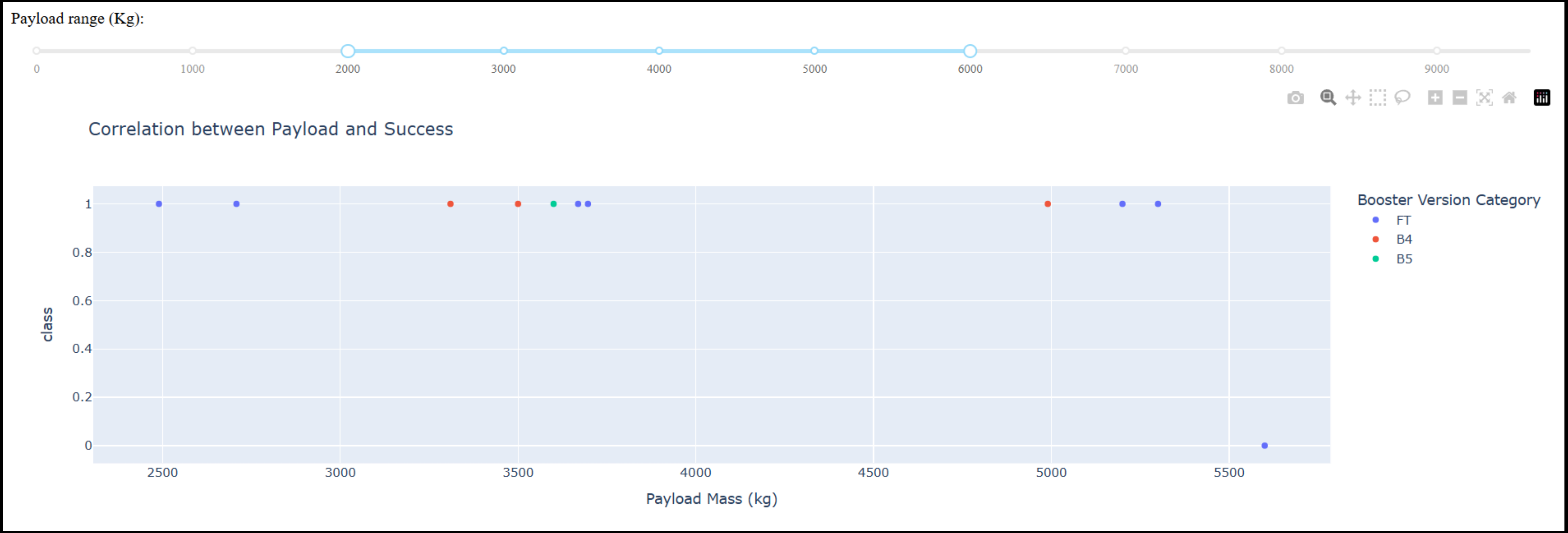
LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO



Explanations:

KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

ALL LAUNCH SITES' LOCATION MARKERS ON A GLOBAL MAP



Explanations:
The charts show that payloads between 2000 and 5500 kg have the highest success rate.

PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY

Explanations:

- The scores of Test Set doesn't help us determine which method performs the best
- The scores of whole dataset suggests that the Best model is SVM as it has both highest F1 score and Accuracy

For Test Set

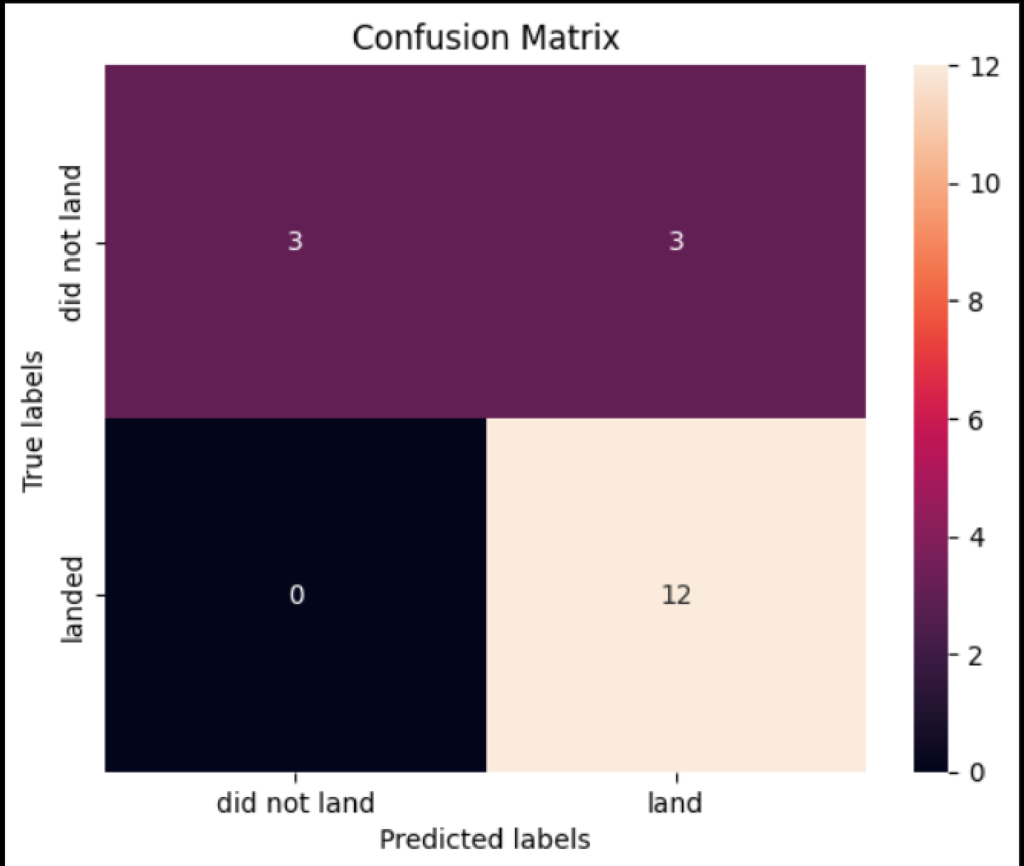
	LogReg	SVM	Tree	KNN
F1 Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

For Entire Set

	LogReg	SVM	Tree	KNN
F1 Score	0.909091	0.916031	0.900763	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

CONFUSION MATRIX

Explanations:
Examining the confusion matrix, we see that SVM can distinguish between the different classes. Also the major problem is false positives.



CONCLUSION

- Support Vector Machine (SVM) is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

APPENDIX

Thanks to:

The Instructors
IBM
Coursera