IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andy Hoang

17 May 2025

Outline

- Executive Summary
- Introduction
- Applied Methodology
- Results
- Conclusion

Executive Summary

Summary of Applied Methods

- Data Collection
- Data Wrangling
- Exploratory Data Analysis & Data Visualisation
- Exploratory Data Analysis with SQL
- Interactive Map with Folium
- Dashboard with Plotly Dash
- Predictive Analysis (Machine Learning)

Findings (results)

- Data Analysis Results
- Analytics Demo (Screenshots)
- Predictive Analysis Results

Introduction

SpaceX Background

SpaceX is founded by Elon Musk in 2002. Its Falcon 9 rocket, known for its reusable first stage, has revolutionised the industry by landing boosters for reuse, achieving a cost of $62 million per launch compared to competitors' $165 million.

SpaceX has accomplished historic milestones, including being the first private company to return a spacecraft from low-earth orbit in 2010.

This project analyse SpaceX launch data to predict first-stage landing success, aiding cost estimation for competitive bids.

Queries to be solved in this project

1. How do variables; payload mass, amount of lights, launch site and orbit affect the success/ fail rate of first stage landing.

2. Whether rate of successful landing improve.

3. What algorithm can be applied for binary classification.

Section 1

Methodology

Methodology Executive Summary

Data collection Methodology

- SpaceX Rest API
- Web Scrapping Wikipedia

Performed Data Wrangling

- Filter Data
- Dealing with Missing Values
- One Hot Encoding for Binary Classification

Deployment of Data Analysis

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - How to build, tune, evaluate classification models

Data Collection

Data collection process involved a dual data approach, and this included the use of 1) API request through SpaceX REST API which provided structured and relevant data and 2) Deployment of web-scraping to extract historical and additional information from SpaceX Wikipedia.

Through SpaceX REST API, data columns obtained included;

- Flight Number, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude and Latidue.

Through use of Web Scrapping SpaceX Wikepedia, data columns obtained included;

- FlightNo, LaunchSite, Payload, PayloadMass, Orbit, Customer, VersionBooster, LaunchOutcome, BoosterLanding, Date and Time.

Data Collection – SpaceX API

GitHub Link: 01 Data Collection API

- Request Rocket Launch data (SpaceX API)
- Decode Response with .json() and converse to dataframe .json_normise()
- Request Data – Launch from SpaceX API
- Construct Data with Dictionary
- Create Dataframe by call Dictionary
- Filter Dataframe only for Falcon 9 Launches
- Replace Missing Values of Payload Mass Column
- Explort Data to CSV

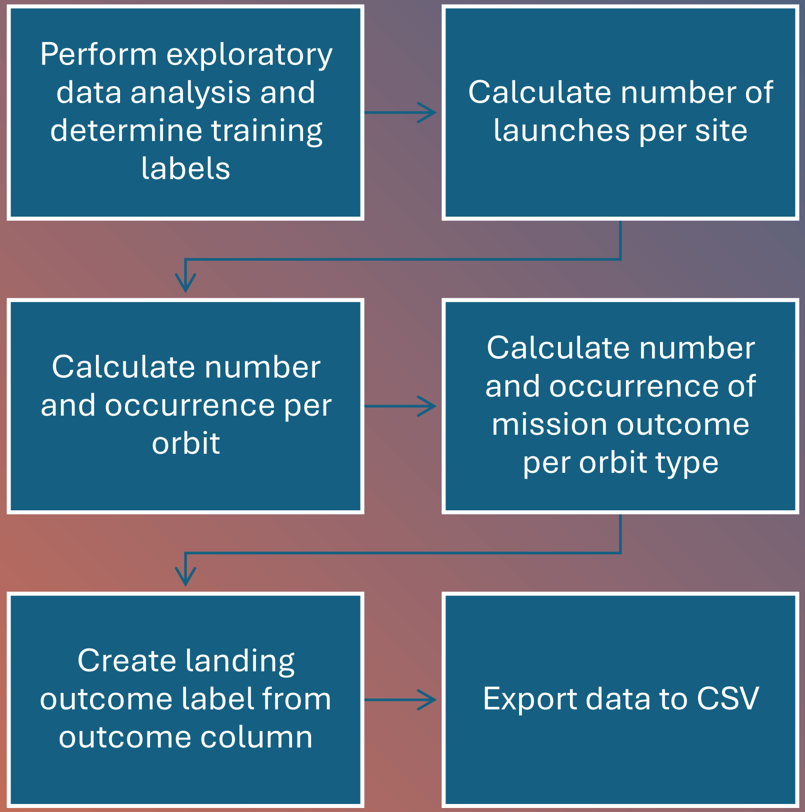
Data Collection – Web Scrapping

[GitHub Link: 02 Data Collection with Web Scraping]

- Request Falcon 9 Data Wikipedia
- Execute BeautifulSoup Object from HTML
- Extract column names from HTML Table Header
- Call parsing HTML tables
- Construct Data obtained into Dictionary
- Create DataFrame from Dictionary
- Export Data to CSV

Data Wrangling

[GitHub Link: 03 Data Wrangling]



The data wrangling process found multiple Falcon 9 booster landing outcomes, which highlights the challenges of reusability. True Ocean, True RTLS, and True ASDS described successful landings in the ocean, on a ground pad, or on a drone ship, respectively, while False Ocean, False RTLS, and False ASDS indicated failed attempts at these locations. Outcomes like None ASDS and None None reflected missions with no landing attempt, often from early Falcon 9 flights. To prepare the data for machine learning, these outcomes were simplified into binary labels: '1' for successful landings (True Ocean, True RTLS, True ASDS) and '0' for unsuccessful or non-attempted landings (False Ocean, False RTLS, False ASDS, None ASDS, None None).

EDA with Data Visualisation

[GitHub Link: 04 EDA with Data Visualisation]

The EDA process utilised various charts to uncover patterns, including Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, and Success Rate Yearly Trends.

- Scatter plots revealed relationships between variables, identifying potential features for machine learning models.
- Bar charts compared discrete categories, highlighting differences and relationships between specific groups and measured values.
- Line charts tracked trends over time, providing insights into the dataset's evolution.

EDA with SQL

[GitHub Link: 05 EDA with SQL]

SQL Queries Performed

- SELECT DISTINCT revealed the variety of launch sites in the Launch_Site column.
- Used LIKE 'CCA%' and LIMIT 5 to showcase five missions from launch sites starting with 'CCA'.
- SUM calculated the total PAYLOAD_MASS__KG_ for NASA (CRS) missions with a WHERE clause for Customer = 'NASA (CRS)'.
- AVG determined the average PAYLOAD_MASS__KG_ for the F9 v1.1 booster using a WHERE condition.
- MIN(Date) pinpointed the earliest successful ground pad landing with WHERE Landing_Outcome = 'Success (ground pad)'.
- WHERE conditions identified Booster_Version for successful drone ship landings with PAYLOAD_MASS__KG_ between 4000 and 6000.
- COUNT(*) and GROUP BY Mission_Outcome tallied successful and failed mission results.
- A subquery with MAX highlighted Booster_Version carrying the maximum PAYLOAD_MASS__KG_.
- SUBSTR extracted the month and year to list Landing_Outcome, Booster_Version, and Launch_Site for 2015 drone ship failures.
- COUNT(*), GROUP BY, and ORDER BY ranked Landing_Outcome frequencies between 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

[GitHub Link: 06 Build Interactive Visual Analytics with Folium]

Markers for Launch Sites:

- Map Objects (circles for launch sites) - added folium circle objects at each launch site using their lat and long coordinates. Visual circles makes it easy to spot their geographical positions.
- Markers for Launch Sites - placed folium marker at launch site coordinates and labels showing the site names in orange for identification. Markers with site names ensured viewers could quickly identify each launch site without confusion.
- Success/Failure Markers - added markers for each launch using green for successful launches (class=1) and red for failed ones (class=0). Colour coded markers helps with visualisation.
- Distance Line to Coastline - drew a blue polyline between a launch site (CCAFS SLC-40) and a nearby coastline point. The polyline to the coastline calculated and displayed the distance, which helps to know how close the launch sites are to coastal areas.

Build a Dashboard with Plotly Dash

[GitHub Link: 08 Build a Dashboard with Plotly Dash]

Plots and Interactions Added to the Dashboard:

- Added a dropdown to choose a launch site, with options for 'All Sites' - CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40. The dropdown enables pick a launch site or view all sites, so this makes it easy to focus on specific locations.
- Pie Chart for Success Rates included to show success rates, updated dynamically via a callback. The pie chart shows the proportion of successful launches.
- Range Slider for Payload Mass included to filter launches by payload mass (0 to 10,000 kg, step 1,000 kg), initialised with the dataset's min and max payload values. The slider allows users to narrow down launches by payload mass.
- Scatter Plot for Payload vs. Success included showing a scatter plot (success-payload-scatter-chart) of payload mass vs. success and updated via a callback. The scatter plot displays payload mass vs. success with colour.

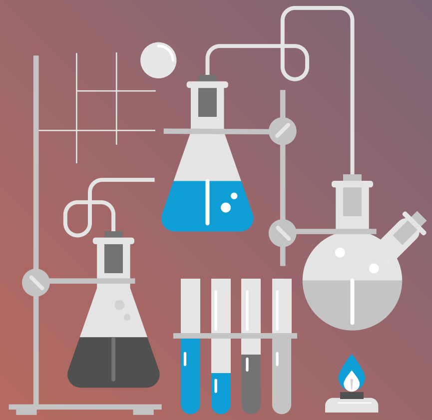
Predictive Analysis (classification)

[GitHub Link: 07 Machine Learning Prediction]

1. Load Data
2. Extract Labels (Y = data['Class’]), then Standardise Features X with StandardScaler
3. Split Data with 80% training, 20% testing via train_test_split, then train models on Logistic Regression, SVM, Decision Tree and KNN
4. Evaluate Models Accuracy scores and confusion matrices, then tune Hyperparameters with GridSearchCV for each model.
5. Re-evaluate tuned models for accuracy scores and compare, then visualise results
6. Select Best Model with SVM with accuracy 0.833.

Results

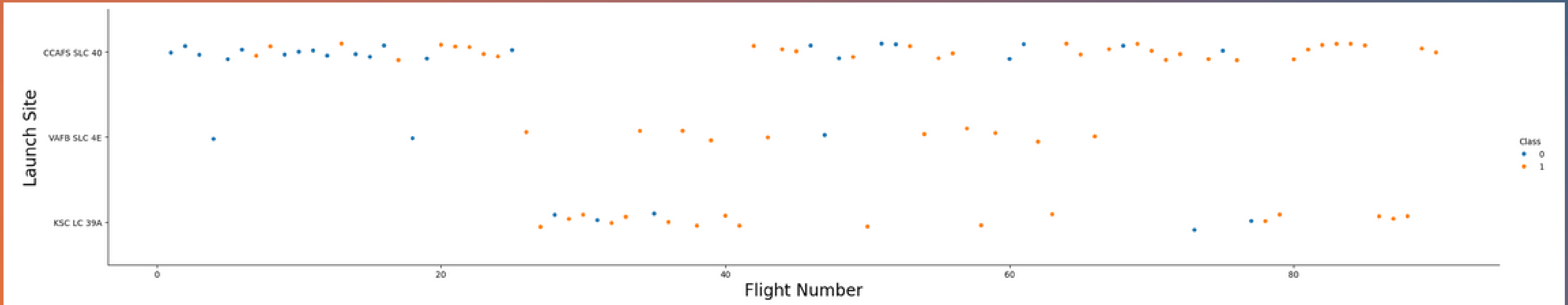
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

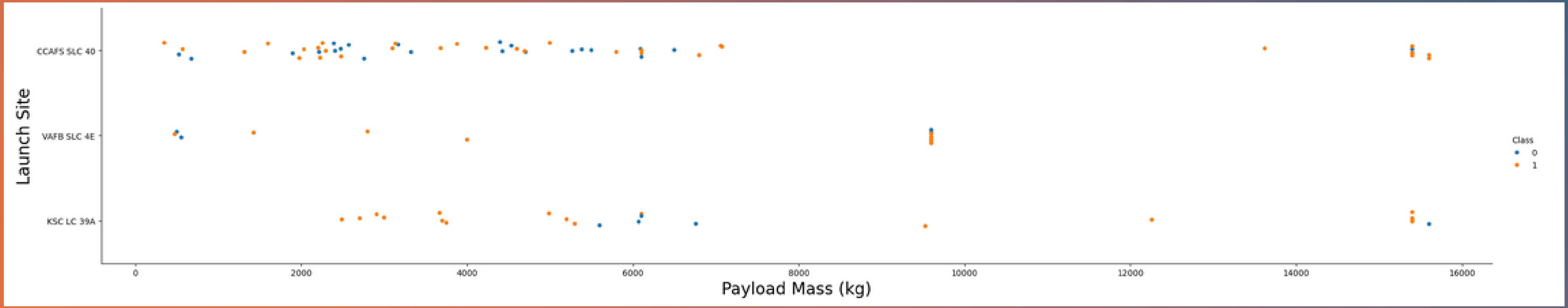
Flight Number vs. Launch Site



Explanation:

- Earlier flights experienced more failures with later lights having more successes.
- CCAFS SLC 40 appears to account for half of all launches.
- VAFB SLC 4E and KSC LC39A outlines higher success rates compared to CCAFS SLC 40.
- Pressumbly each new launch increases success.

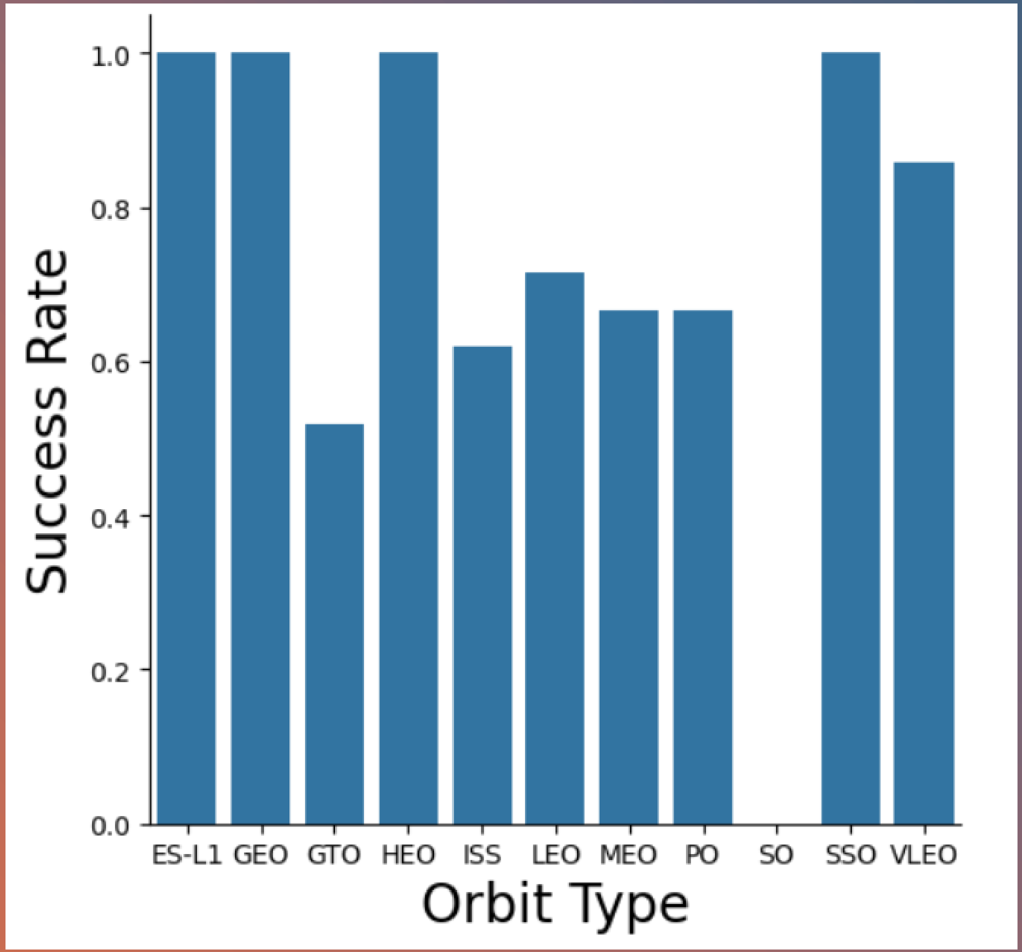
Payload vs. Launch Site



Explanation:

- Higher payloads more than 6000kg are correlated with higher success rates, but there may be nuance by sites.
- Most launches with payload mass of more than 7000kg were successfully.

Success Rate vs. Orbit Type



Explanation:

Orbit type with 100% success rate:

- EL-L1, GEO , HEO and SSO

Orbit type with 60% to 80% success rate:

- ISS, LEO, MEO and PO

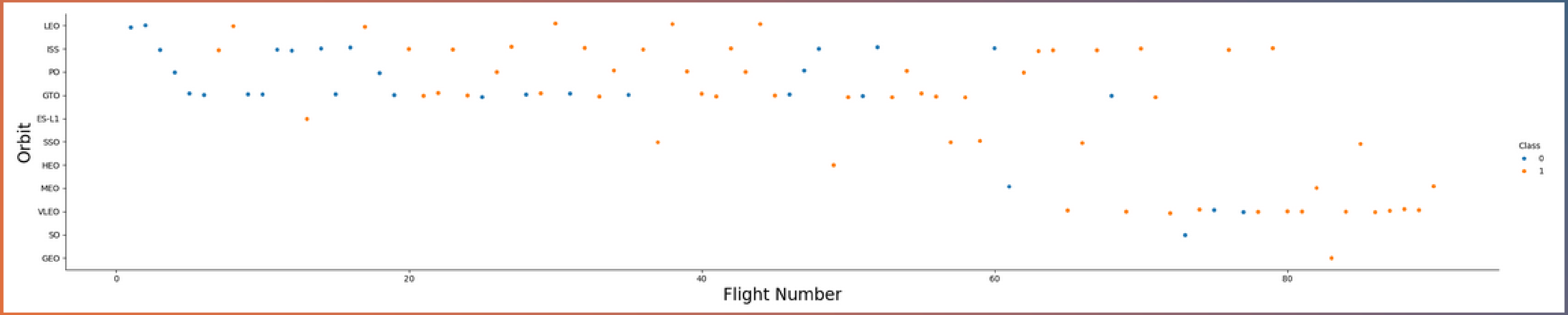
Orbit type with 80% to 90% success rate:

- VLEO

Orbit type with 0% success rate:

- SO

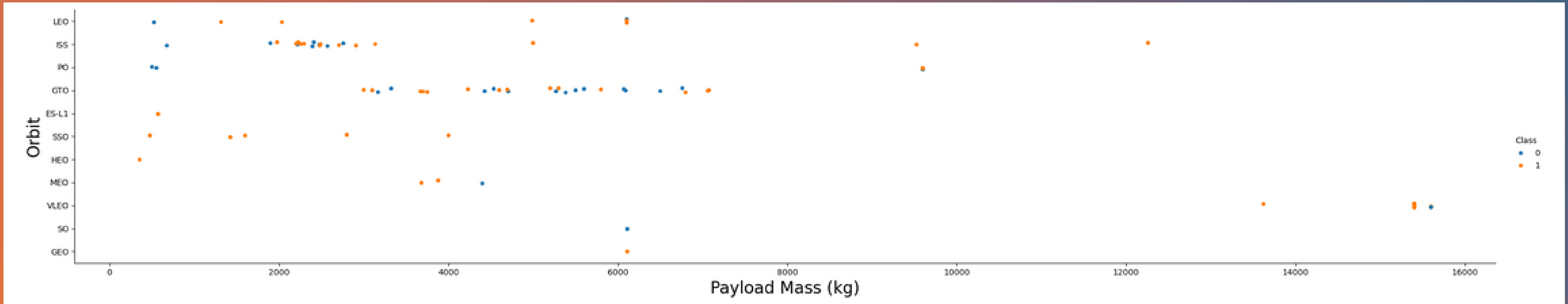
Flight Number vs. Orbit Type



Explanation:

- It would appear LEO Success Rate may be related to number of flights, but GTO orbit shows there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



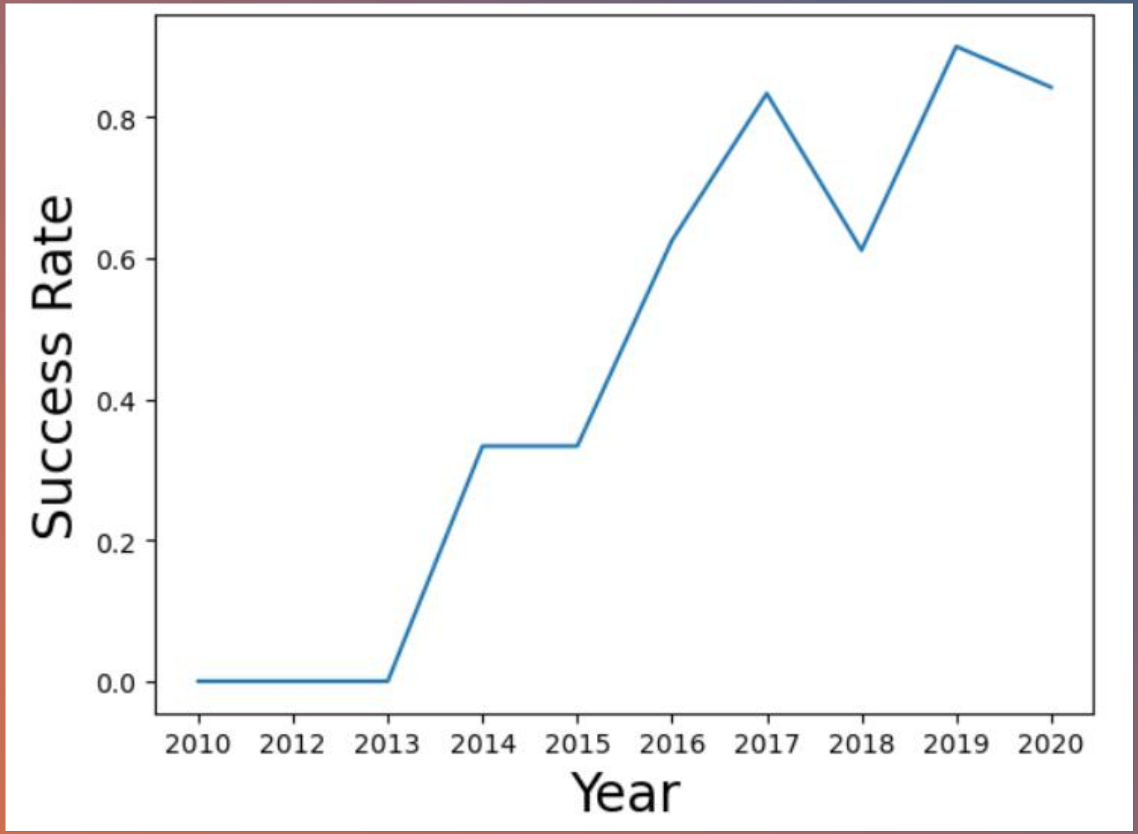
Explanation:

- It would appear heavy payloads have negative influence on GTO orbits and positive on GOT/ Polar Leo (ISS) orbits.

Launch Success Yearly Trend

Explanation:

- Graph shows mostly progression and success from 2013 through to 2020.



All Launch Site Names

Explanation:

- Pull from data base to display names of unique launch sites for space missions.

```
%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

Launch Site Names Begin with 'CCA'

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

[45]:
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

[45]:

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

Explanation:

- Pull from database to display 5 records where launch sites with keyword with CCA.

Total Payload Mass

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

```
[46]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

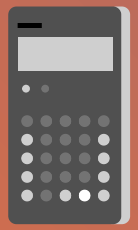
```
 * sqlite:///my_data1.db
Done.
```

[46]:

SUM("PAYLOAD_MASS__KG_")
45596

Explanation:

- Results display total payload mass carried by boosters launched by 'NASA CRS'.



Average Payload Mass by F9 v1.1

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

```
[47]: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

[47]:

AVG("PAYLOAD_MASS__KG_")
2928.4

Explanation:

- Results display average payload mass carried by booster version F9 v1.1.

• First Successful Ground Landing Date

```
[48]: %sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

[48]:

MIN("Date")
2015-12-22

Explanation:

- Results display date when first successful landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[49]: %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;

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

[49]:

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

Explanation:

- Results display name of boosters with success via drone ship and have payload mass more than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
[50]: %sql SELECT "Mission_Outcome", COUNT(*) as count FROM SPACEXTABLE GROUP BY "Mission_Outcome";

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

[50]:

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

Explanation:

- Results display number of successful/ failed missions.

Boosters Carried Maximum Payload

```
[51]: %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

[51]:

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

Explanation:

- Results display name of booster versions that have carried maximum payload mass.

2015 Launch Records

January

```
[52]: %sql SELECT SUBSTR("Date", 6, 2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND SUBSTR("Date", 0, 5) = '2015';

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

[52]:

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

April

Explanation:

- Results display failed outcome in drone ship, their booster versions and launch site name for months = 01 for January 2015 and 04 for April 2015.

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

```
[60]: query_result = %sql SELECT "Landing_Outcome", COUNT(*) as count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY count DESC;
df_result = query_result.DataFrame()
df_result

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

[60]:

	Landing_Outcome	count
0	No attempt	10
1	Success (drone ship)	5
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	2
7	Precluded (drone ship)	1

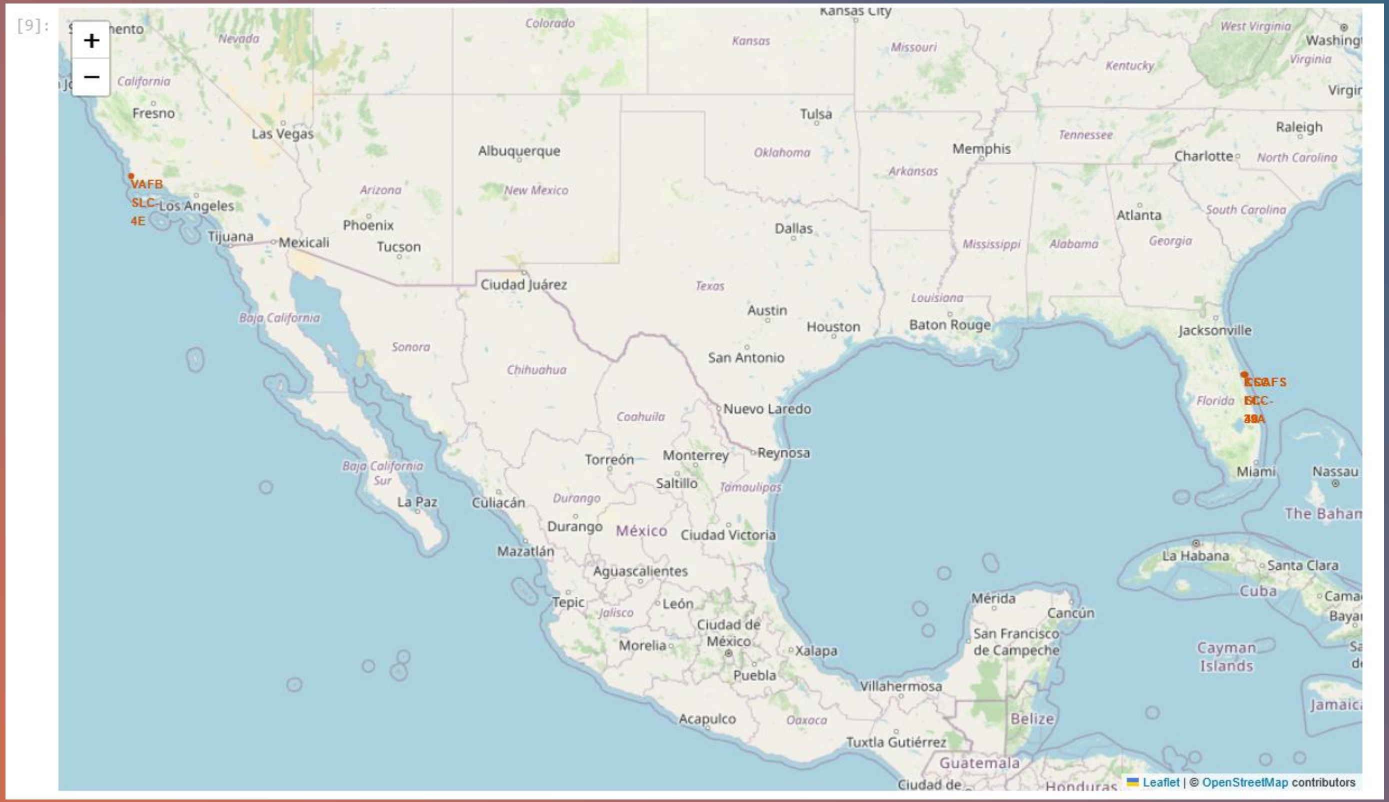
Explanation:

- Results display ranking of 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

Section 3

Launch Sites Proximities Analysis

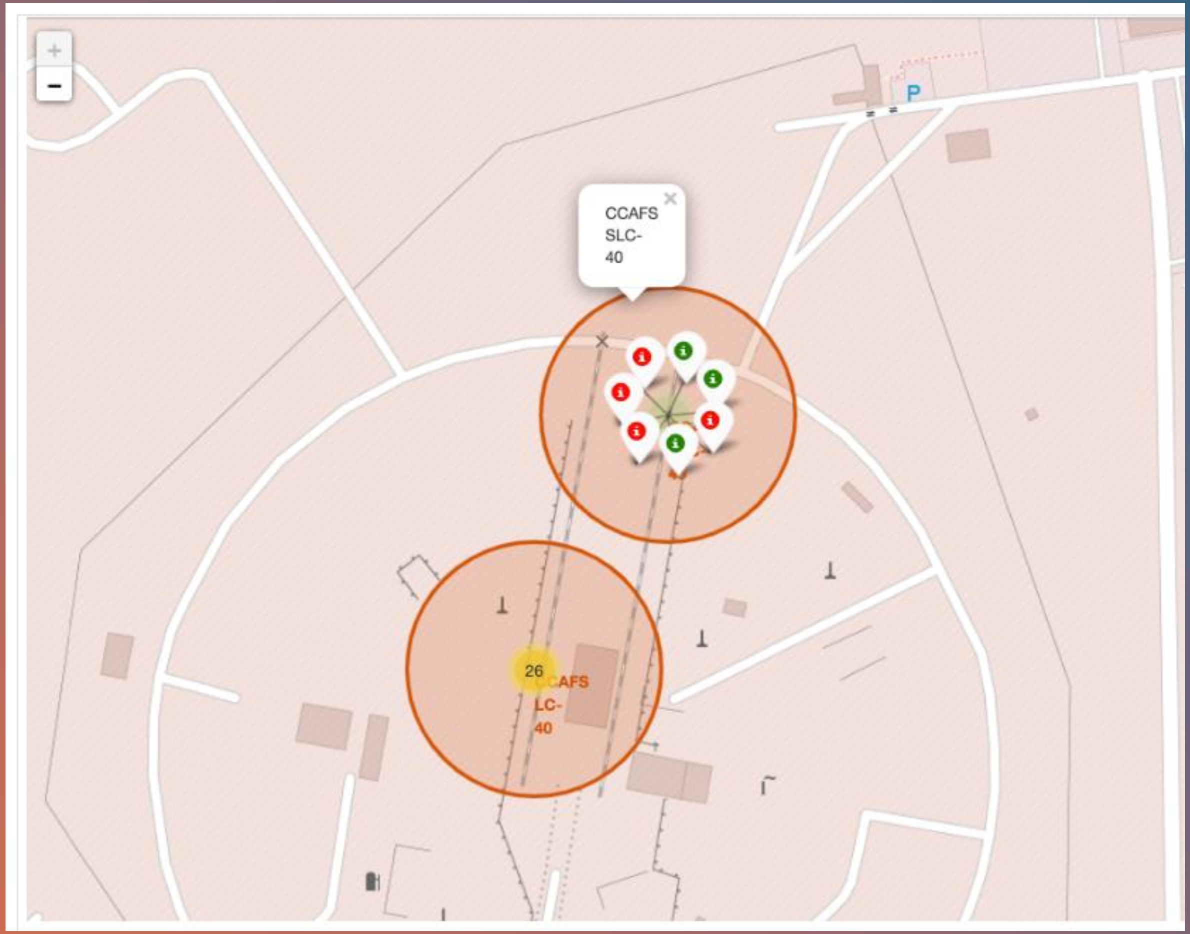
Launch Sites on Global Map



Explanation:

- Most of the launch sites are within proximity to the equator line. Most of the launch sites are in close proximity to the coast, and this infers risk mitigation, whereby having debris drop or explore towards the ocean is safer in comparison to in land.

• Colour Label Launch Records on Map



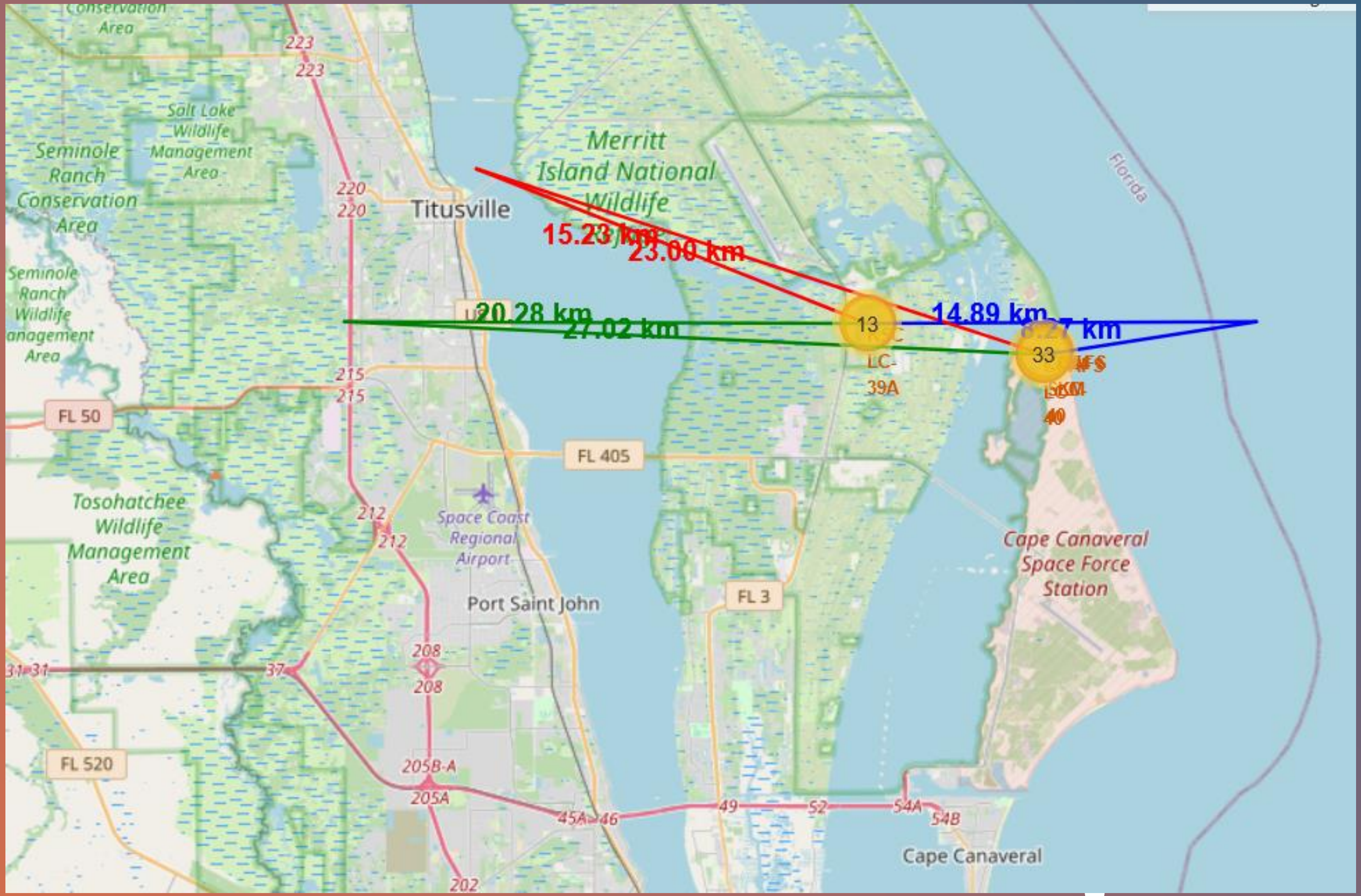
Explanation:

- Green markers outline launch sites that have greater success rate.
    - Green = Successful Launch
    - Red = Failed Launch

Distance from Launch Site CCAFS LC-40 to proxmities

Explanation:

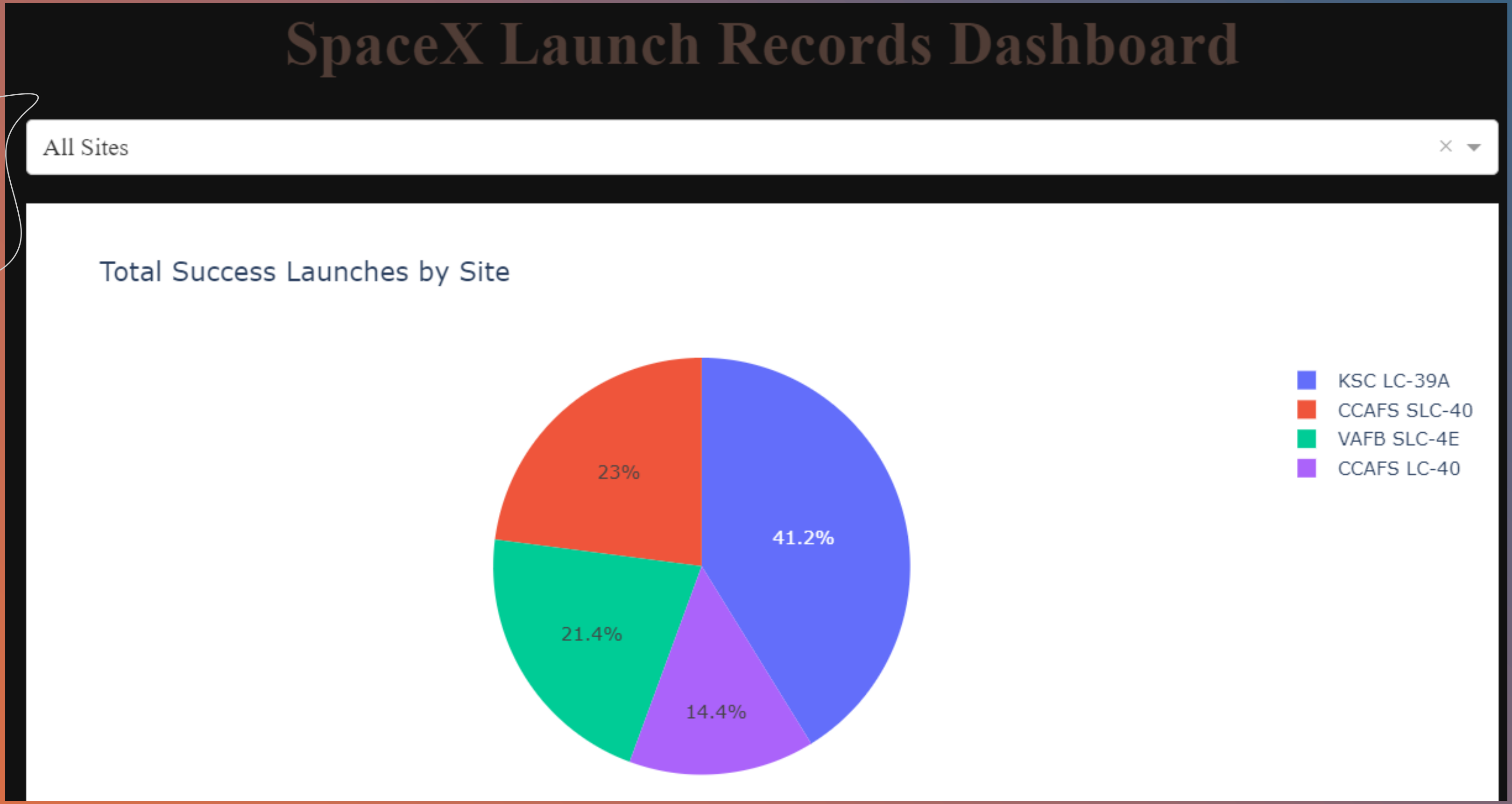
- Visual results show launch from shows the following:
  - Red line = 23 KM to nearest railway.
  - Green line = 27.02 KM to nearest
  - Blue line = 8.27 KM to nearest coastline.



Section 4

Build a Dashboard with Plotly Dash

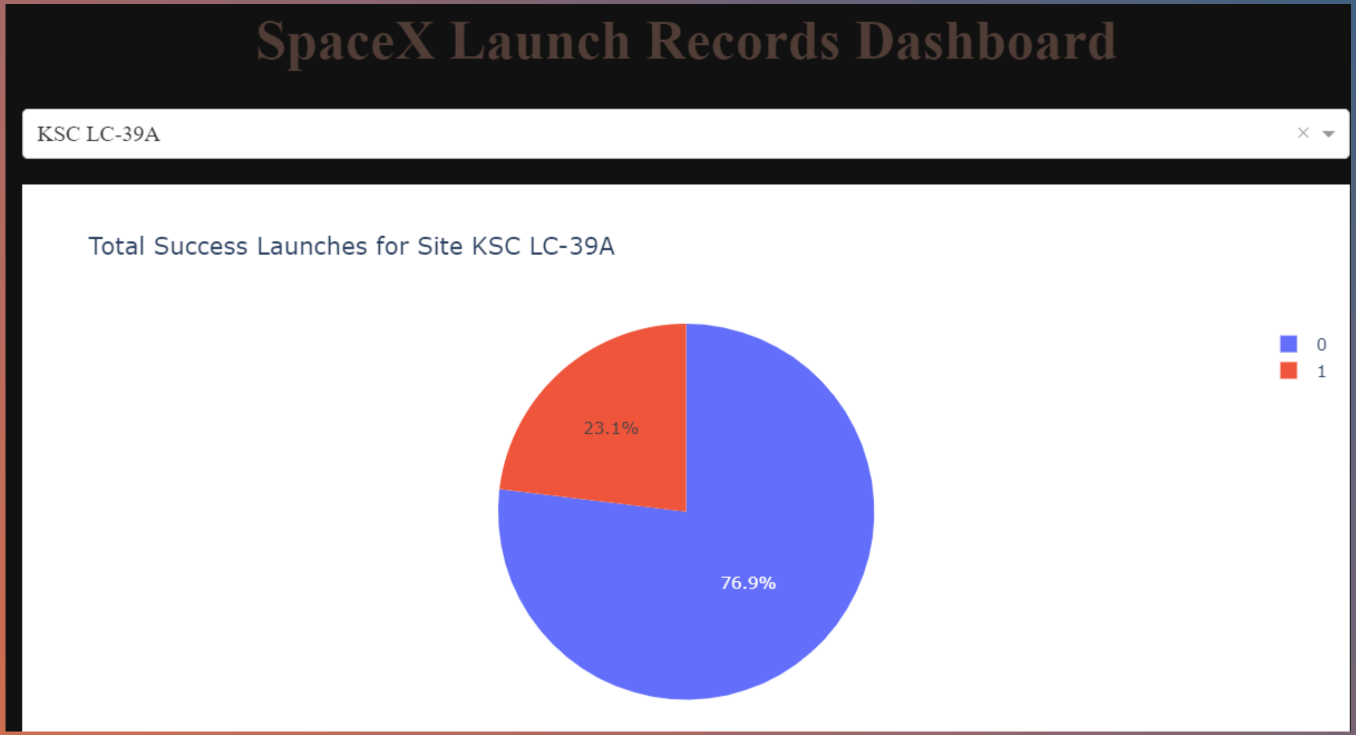
Launch Success Count – All Sites



Explanation:

- Chart outlines all sites with KSC LC-39A having the most successful launches.

Launch site with Highest Launch Ratio



Explanation:

- Chart outlines KSC L-39A has the highest launch rate of 76.9%.

Payload vs. Launch Outcome scatter plot for all sites

Explanation:

- Scatter plot shows correlations between payload mass and launch sites, especially FT and B5 boosters holding highest success rates across payload ranges with 3000 – 6000 kg range being consistent.

Section 5

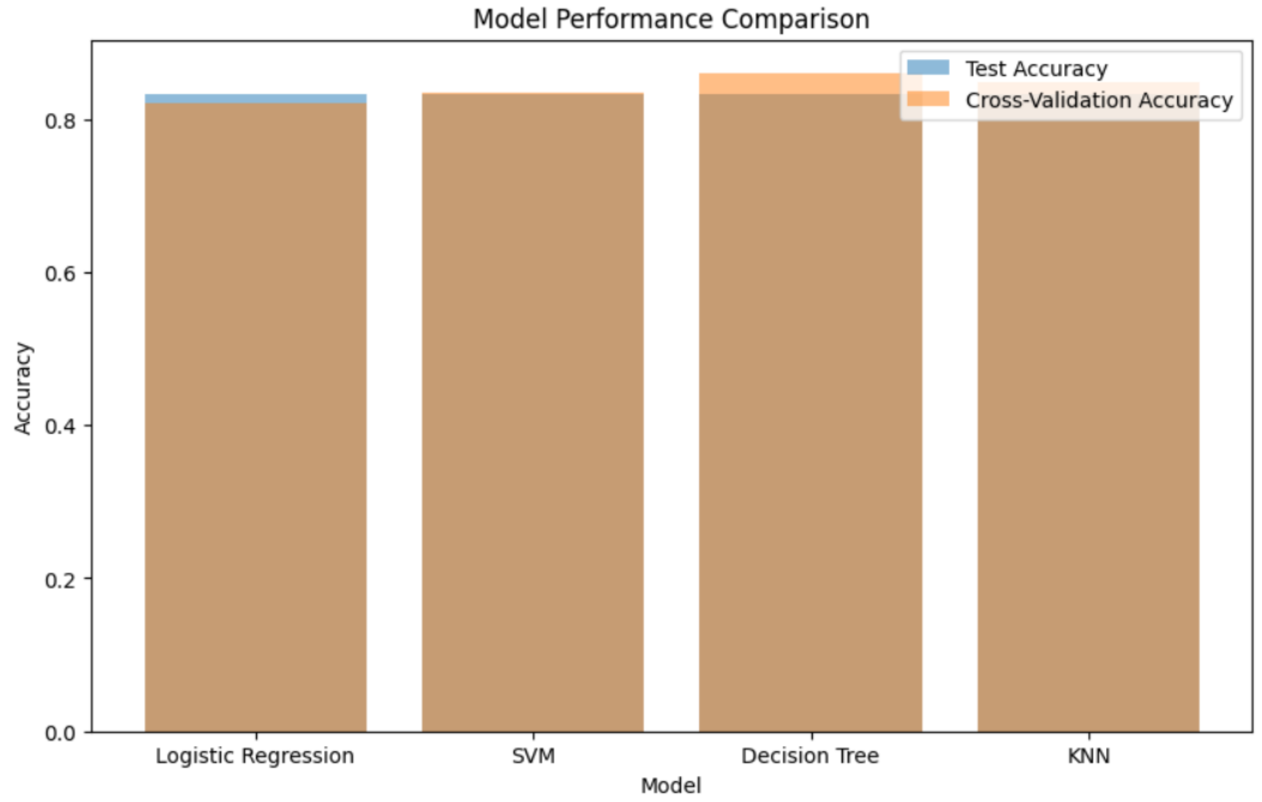
Predictive Analysis (Classification)

Classification Accuracy

Explanation:

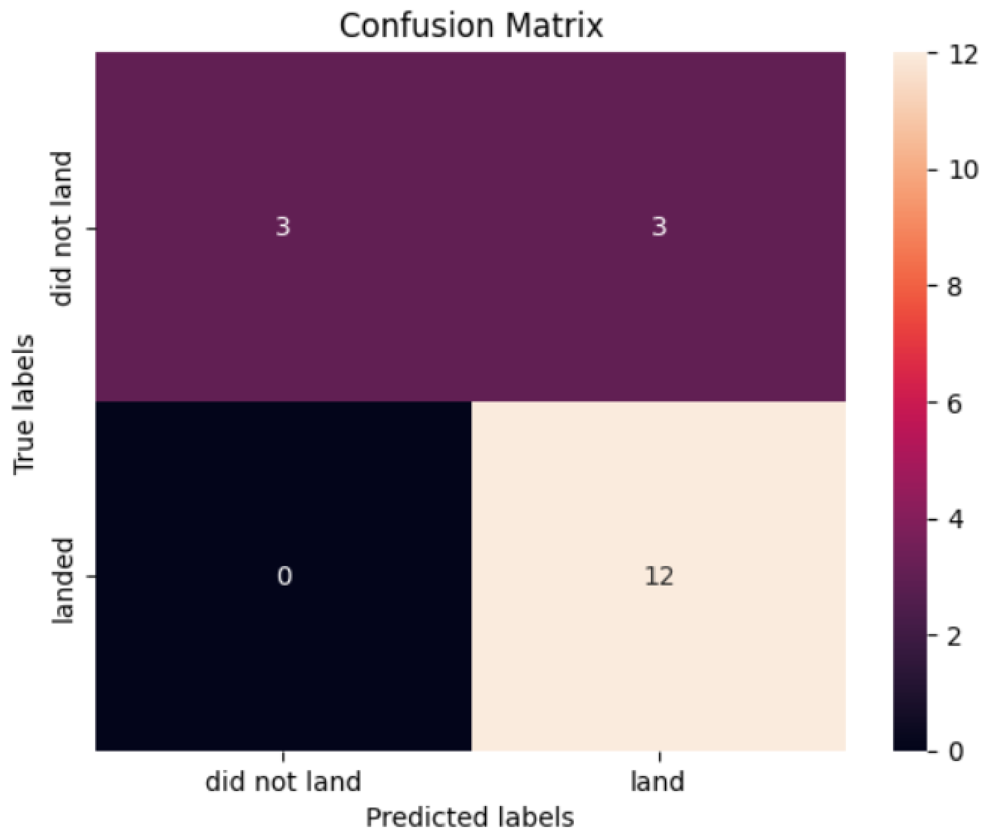
- The bar chart compares four models – 1) Logistic Regression, 2) SVM, 3) Decision Tree, and 4) KNN—using Test and Cross-Validation Accuracy. All models scored 0.8333, but the Decision Tree's higher Cross-Validation Accuracy (0.8607) and performance on the full dataset confirm it as the best model.

	Model	Cross-Validation Accuracy	Test Accuracy
0	Logistic Regression	0.8214	0.8333
1	SVM	0.8357	0.8333
2	Decision Tree	0.8607	0.8333
3	KNN	0.8482	0.8333



Confusion Matrix

```
[37]: yhat = knn_cv.predict(X_test)
      plot_confusion_matrix(Y_test,yhat)
```



Explanation:

- Given slide 43 shows the Decision Tree has the highest accuracy with = 0.8607, indicating it generalised better across the dataset.

Conclusions

- KSC LC-39A has the highest success rate = 76.9% with sites near the equator and coast for safety.
- Payloads over 6,000 kg show higher success with 3,000–6,000 kg range is most consistent, varying by orbit type.
- CCAFS LC-40 is 8.27 km from the coast, 23 km from a railway, and 27.02 km from a highway, aiding risk mitigation.
- Launch success improved from 2013 to 2020 which reflects technological advancements.
- Predictive tools enable better cost predictions.

Thank you!