IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Panthanan Namwanta
1/02/2025

Outline

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

Executive Summary

- Summary of methodologies
    - Data collection through API
    - Data collection with Web Scarping
    - Data wrangling
    - EDA with SQL
    - EDA with Data visualization
    - Interactive Visual Analytics with Folium and Plotly Dash
    - Machine Learning Prediction
- Summary of all results
    - Exploratory Data Analysis result
    - Interactive analytics in screenshots
    - Predictive Analytics result

Introduction

- Project background and context
  - SpaceX's Falcon 9, a partially reusable, medium-lift launch vehicle, revolutionizes space access with its cost-effectiveness. Its reusable first stage drastically reduces launch costs, advertised at $62 million, significantly undercutting competitors. The ability of the first stage to land after separation is critical to this cost advantage. Predicting whether this first stage will successfully land is, therefore, crucial for cost estimation. Competitors could use this prediction to offer more competitive bids. This project aims to create a machine learning model predicting Falcon 9 first stage landing success using historical launch data, contributing to more accurate cost predictions in the space industry.
- Problems you want to find answers
  - What factors influence the rocket's ability to land successfully?
  - What patterns or data behaviors correlate with a higher success rate?
  - What is the most effective method for predicting future launch outcomes?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and Web Scraping from Wikipedia.
- Perform data wrangling
  - Data wrangling involved exploratory data analysis to engineer a target variable for supervised learning, with the goal of predicting landing success.
- 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 – SpaceX API

Request data from SpaceX API

Decode response and **normalize** data to Pandas dataframe

Extract data that relate to launches success

Filtered data to only include Falcon9

Handling missing values of payload mass with its mean

- https://github.com/PanTa7r/IBM_data_science_capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Request Falcon9 data from Wikipedia → **Create a BeautifulSoup object** from the HTML response → **Extract all column** names from the HTML table header → **Create a dictionary** from column names → **Create a dataframe** from dictionary

- https://github.com/PanTa7r/IBM_data_science_capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Perform EDA to identify missing values and columns types

Calculate:

- the number of launches on each site
- the number and occurrence of each orbit
- the number and occurrence of mission outcome of the orbits

Create a landing outcome label from Outcome column as binary

- https://github.com/PanTa7r/IBM_data_science_capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Scatter Chart: To view relationship between variable

- Flight Number vs Payload
- Flight Number vs Launch Site
- Flight number vs orbit type
- Payload Mass vs Launch Site
- Payload Mass and Orbit type

Bar Chart: To show comparisons among discrete categories

- Success rate of each orbit type

Line chart: To display trends over time

- Yearly success trends

- https://github.com/PanTa7r/IBM_data_science_capstone/blob/main/EDA-with-visualization.ipynb

EDA with SQL

- We applied EDA with SQL to get insight from the data.
  - ❑The name of unique launch sites in the space mission
  - ❑The total payload mass carried by boosters launched by NASA(CRS)
  - ❑The average payload mass carried by booster version F9 v1.1
  - ❑The total number of successful and failure mission outcomes
  - ❑The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - ❑The names of the booster versions which have carried the maximum payload mass.
  - ❑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
- https://github.com/PanTa7r/IBM_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Marker** is used to pinpoint specific locations on a map.
- **Circle** is used to add a highlighted circle area with a text label on specific coordinate.
- **Lines** is used to draw a distance between places.
- **Marker Clustering** combines multiple markers in proximity into a single cluster icon, improving visual organization and map readability.
- I added marker, circle, Lines, marker clustering to visualize where rockets were launched.
- https://github.com/PanTa7r/IBM_data_science_capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dropdown list with launch sites
  - Allow user to select launch sites to perform analysis
- Pie chart showing success percentage
  - Allow user to see successful and unsuccessful rate by each launch sites as percentage.
- Slider of payload mass range
  - Allow user to select payload mass range
- Scatter chart showing payload mass vs. success class by booster version
  - Allow user the see the correlation between payload and launch success
- https://github.com/PanTa7r/IBM_data_science_capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Create Numpy array from class column → Standardize the data → Split data to train and test dataset → Perform machine learning on logistic regression, SVM, decision tree, KNN → Calculate accuracy, Jaccard score , and F1_score on each model → Summarize the best performing model

- https://github.com/PanTa7r/IBM_data_science_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5_cuML.ipynb

Results

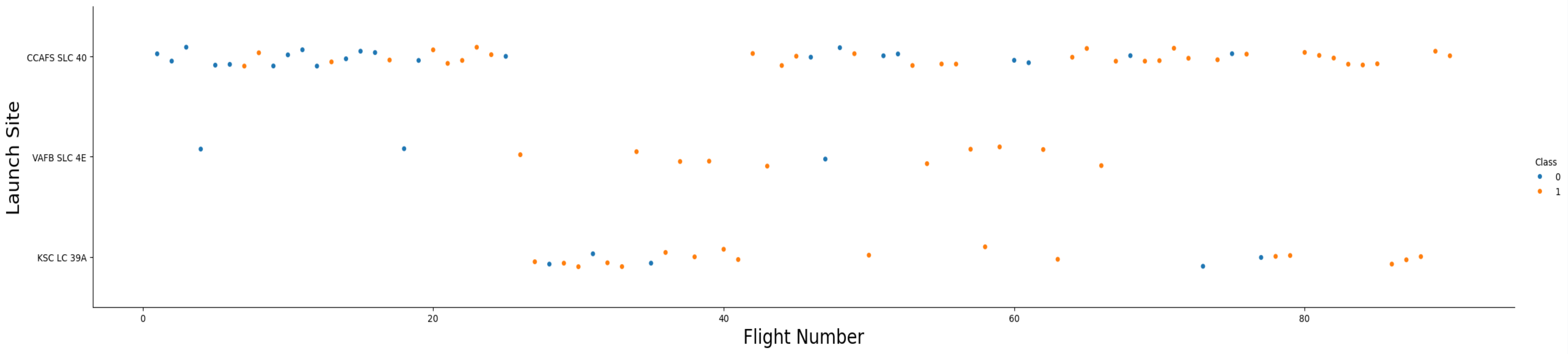
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

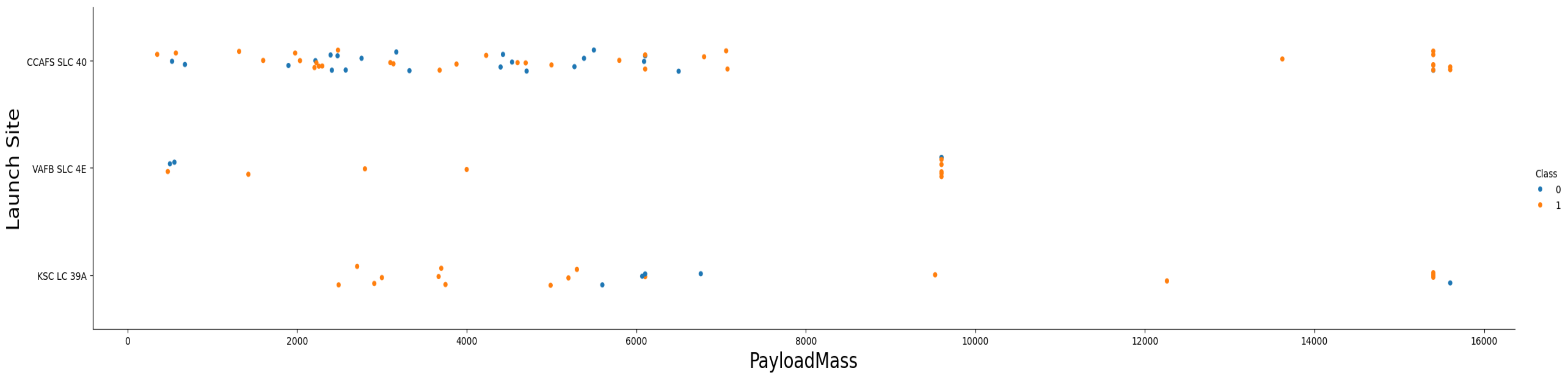
Flight Number vs. Launch Site

- Class 0 = Bad outcome (Unsuccessful launch or unsuccessful landing)
- Class 1 = Good outcome (Successful launch and successful landing)
- The plot indicates a trend of improving landing success rates with increasing flight number.



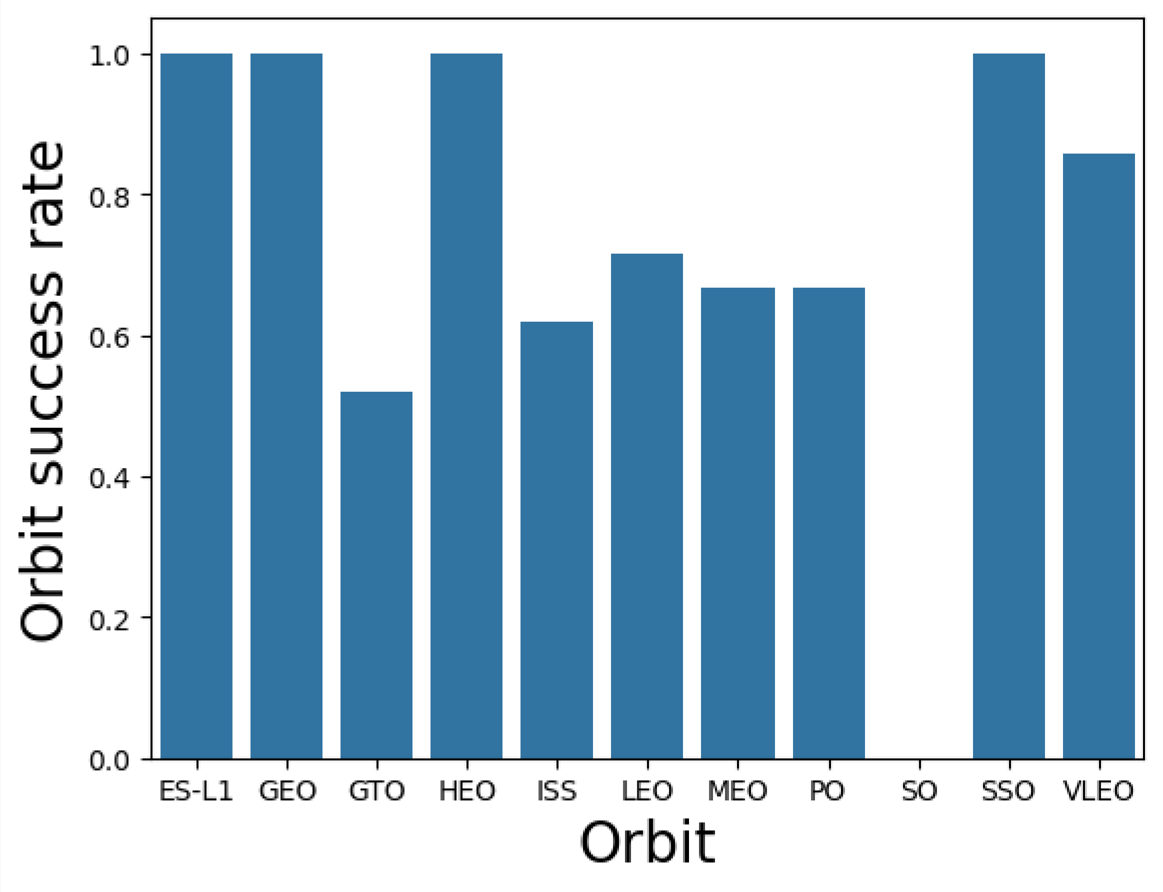
Payload vs. Launch Site

- The plot show that payload mass over 10,000kg were operated on CCAFS SLC40 and KSC LC 39A.
- Across all launch sites, there isn't a strong, clear trend of payload mass directly determining the outcome.



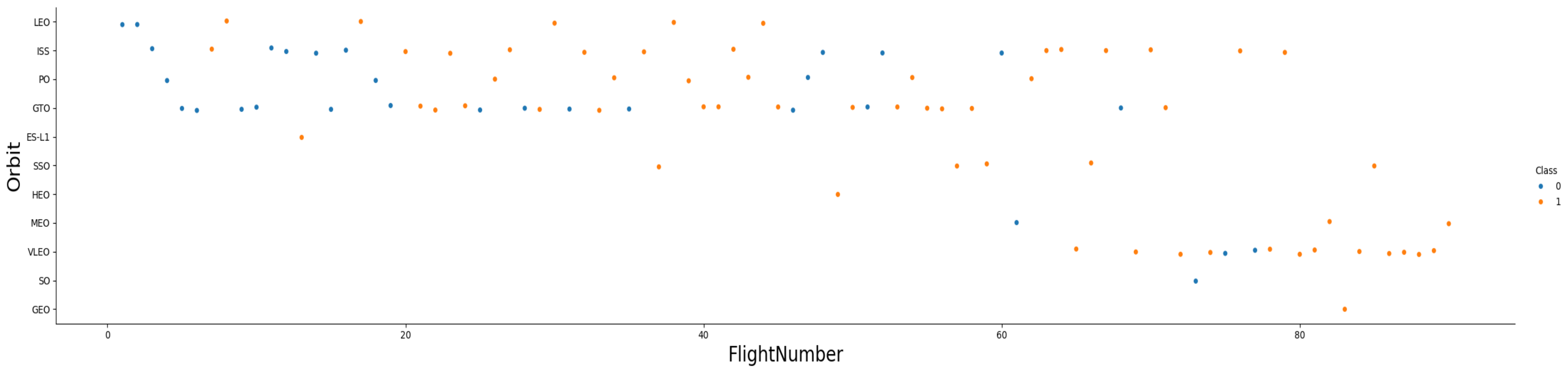
Success Rate vs. Orbit Type

- Some orbits (e.g., ES-L1, GEO, HEO, SSO) have a 100% success rate
- VLEO has a high success rate, but slightly less than 1.0.
- GTO has a relatively low success rate compared to others.
- ISS, LEO, MEO, and PO have moderate success rates, with values between 0.6 and 0.8
- SO (Solar Orbit) has 0 success rate.



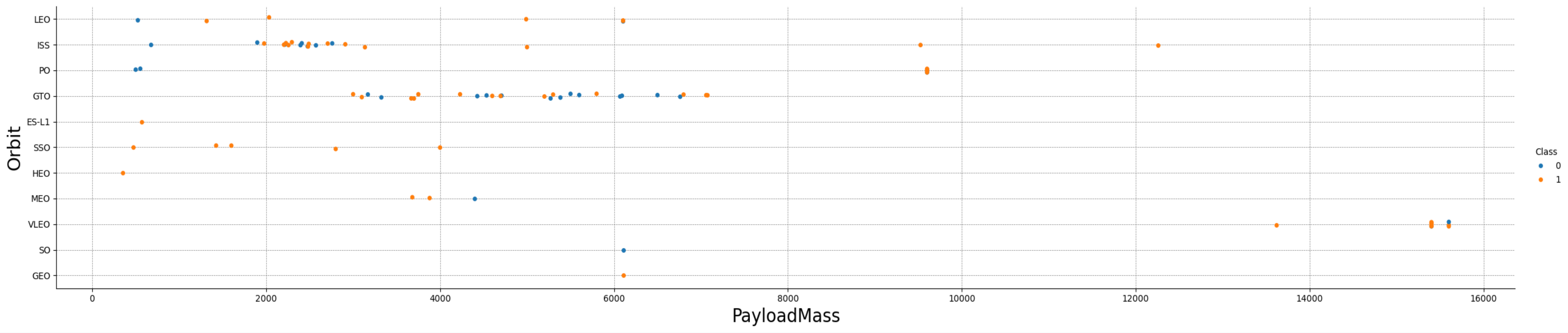
Flight Number vs. Orbit Type

- LEO orbit show success increases over time.
- GTO and ISS orbits show no relationship between flight number and success.



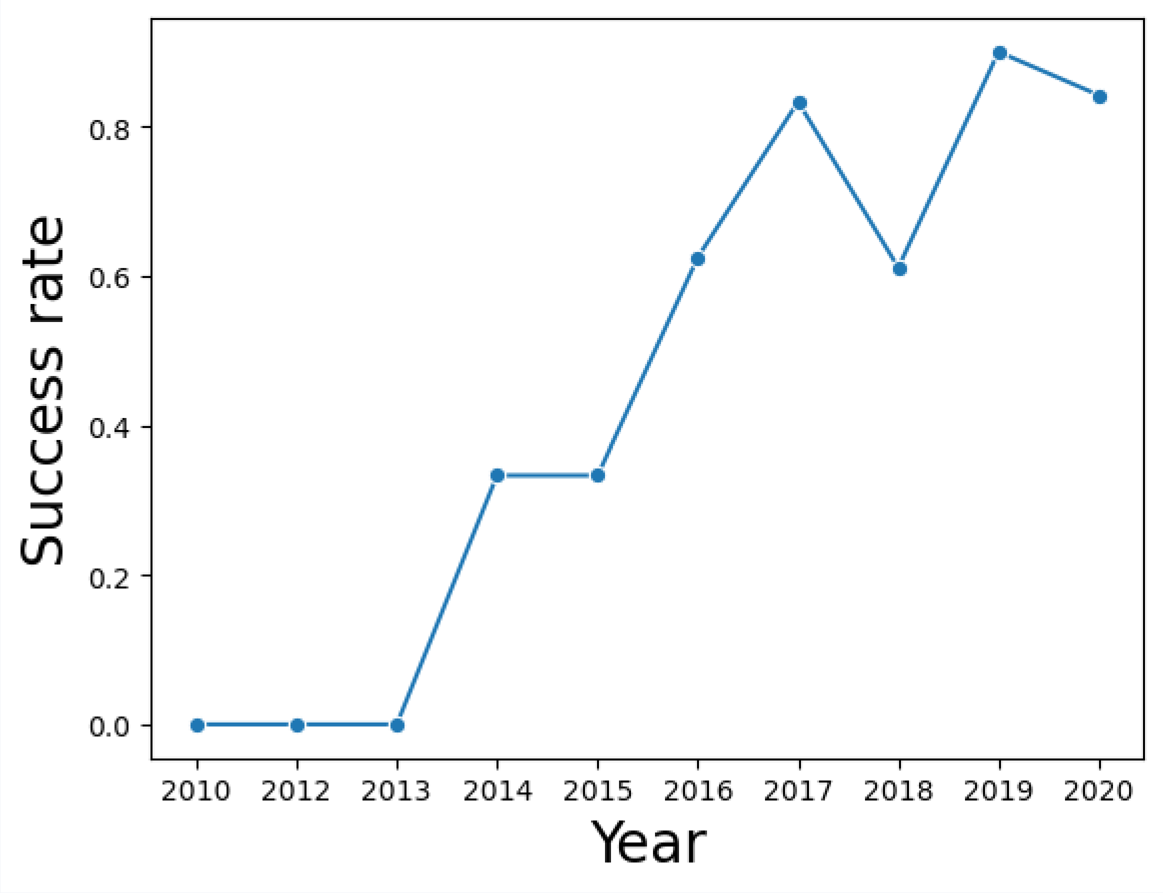
Payload vs. Orbit Type

- **Heavier payloads** the likelihood of a successful landing appears to be higher for PO, LEO, ISS. This suggest these orbits **may have more favourable conditions** or optimized launch systems for successful landing.
- **GTO orbit** presents a mixed outcome. It indicates that **factors beyond payload mass,** may influence the landing success rate.



Launch Success Yearly Trend

- The success rate since 2013 keep increasing till 2020.
- There is a short downtrend in year 2018. It indicate that there must be some experiment that unsuccessful during that time.



All Launch Site Names

- This query will give us a list of all the different launch sites used by SpaceX, with no repeats.

```
[50]:
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE;

 * sqlite:///my_data1.db
Done.
[50]:
```

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

Launch Site Names Begin with 'CCA'

- This query will give us the complete details (all columns) for the first 5 SpaceX launches that occurred from launch sites starting with "CCA." This lets us examine data specifically related to Cape Canaveral

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

```
[51]: %sql SELECT * FROM SPACEXTABLE WHERE launch_site like 'CCA%' LIMIT 5;
```

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

[51]:

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

Total Payload Mass

- This query will return a single number: the total combined payload mass (in kilograms) for all SpaceX launches where the customer was "NASA (CRS)
- This tells us the total amount of cargo delivered under that specific contract.

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

```
[52]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE where Customer = 'NASA (CRS)';

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

[52]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- This query will return a single number representing the average payload mass (in kilograms) for all SpaceX launches that used the "F9 v1.1" booster.
- This gives us insight into the typical cargo capacity of this specific booster version.

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

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

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

[53]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- This query will return a single date – the earliest date in the dataset when SpaceX achieved a successful ground pad landing

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

Hint:Use min function

```
#%sql SELECT min(Date),Landing_Outcome FROM SPACEXTABLE where Landing_Outcome like 'Success%'; #For search thorght all success outcome.
#FOR MOST accurate base on given TASK
%sql SELECT min(Date) FROM SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query will return a table showing the booster version, payload mass, and landing outcome for launches that satisfy both conditions: a successful drone ship landing and a payload mass between 4000 and 6000 kg.
- This helps us analyze a specific subset of successful missions.

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

```
[55]: %sql SELECT Booster_Version,PAYLOAD_MASS__KG_,Landing_Outcome FROM SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and (PAYLOAD_MASS__KG_ >4000 and PAYLOAD_MASS__KG_ <6000);
 * sqlite:///my_data1.db
Done.
```

[55]:

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- This query will return a table with two columns:MISSION_OUTCOME: The different types of mission outcomes (e.g., "Success", "Failure").total_number: The number of launches that had each outcome.

```
[56]: %sql SELECT MISSION_OUTCOME, COUNT(*) AS total_number FROM SPACEXTABLE GROUP BY MISSION_OUTCOME;
 * sqlite:///my_data1.db
Done.
```

[56]:

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

Boosters Carried Maximum Payload

- This query will return a table showing the booster version(s) and the corresponding payload mass for the launch(es) that carried the heaviest payload.
- It essentially tells us which booster achieved this record.

```
[57]: %sql SELECT distinct(BOOSTER_VERSION),PAYLOAD_MASS__KG_\
FROM SPACEXTABLE\
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

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

[57]:

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

2015 Launch Records

- This query will return a table with the month, booster version, launch site, and landing outcome for launches that meet both conditions: they were failed drone ship landings and they occurred in 2015.
- This gives us a focused view of specific events in the dataset.

```
[58]: %sql SELECT SUBSTR(Date, 6, 2) AS month, BOOSTER_VERSION, LAUNCH_SITE,LANDING_OUTCOME FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND SUBSTR(Date, 0, 5) = '2015';
 * sqlite:///my_data1.db
Done.
```

[58]:

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

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

- This query will return a table with two columns:
  - LANDING_OUTCOME: The different landing outcome categories.
  - count_outcomes: The number of times each outcome occurred within the specified date range, sorted from most frequent to least frequent.

[62]:
```
%%sql
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS count_outcomes
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY count_outcomes DESC
```

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

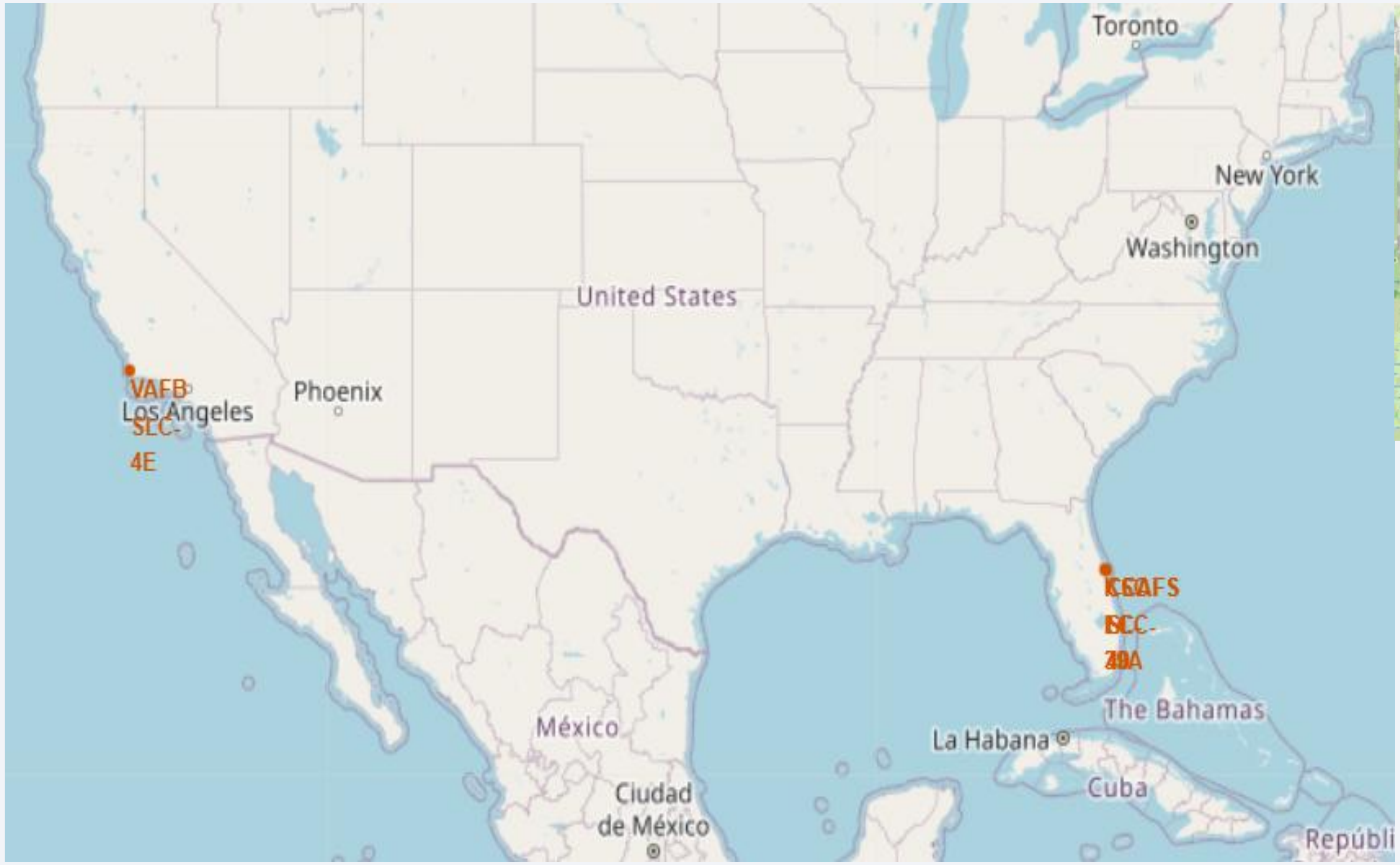
[62]:

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

Section 3

Launch Sites Proximities Analysis

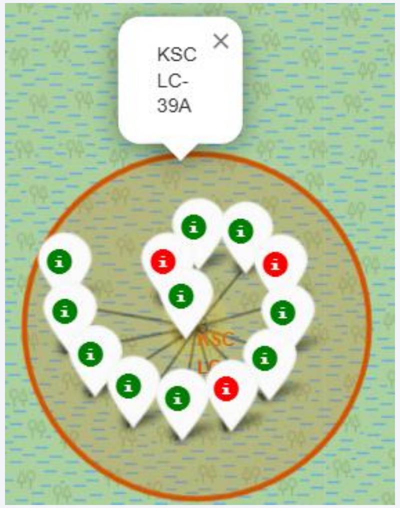
All launch site in SpaceX mission

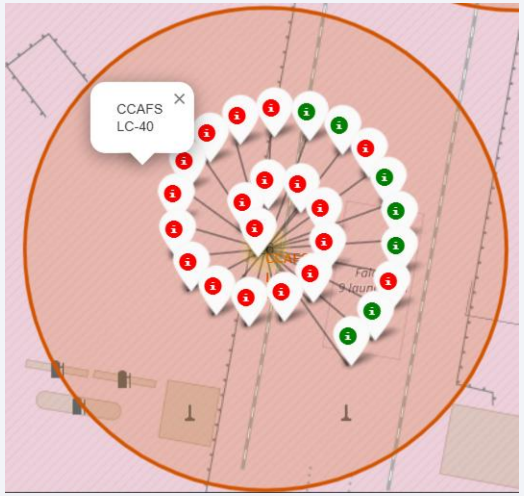


- All launches were conducted from the coasts of the United States. Three launch sites are located in Florida, and **all sites are strategically positioned as close to the Equator as possible to maximize the benefit of Earth's rotational speed**.

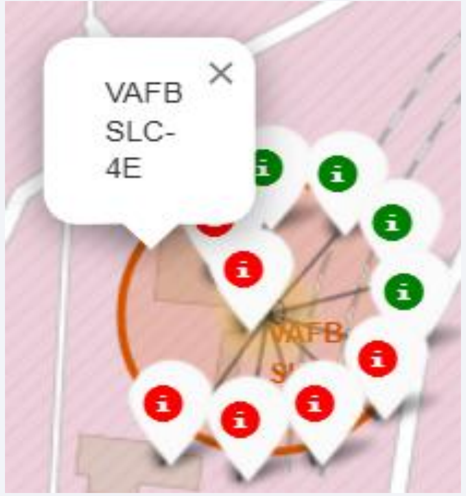
Launch marker

Florida launch sites



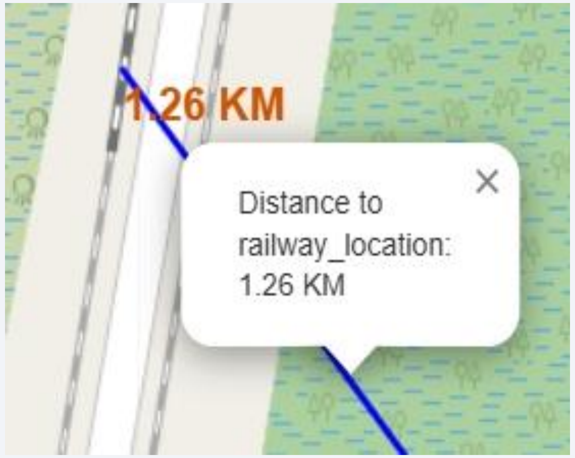


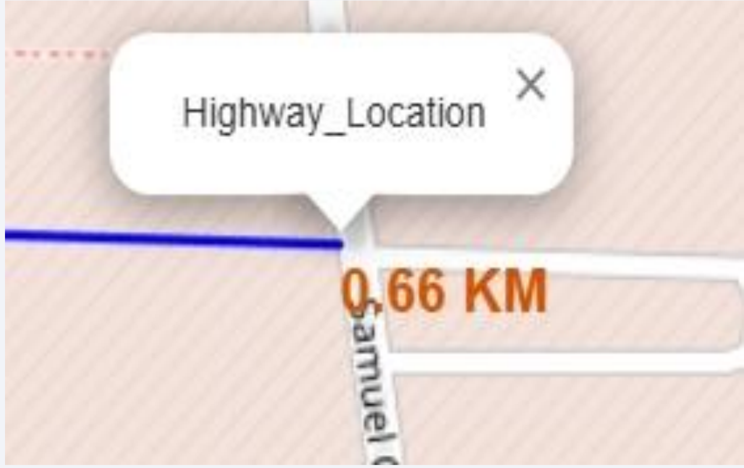
California launch sites

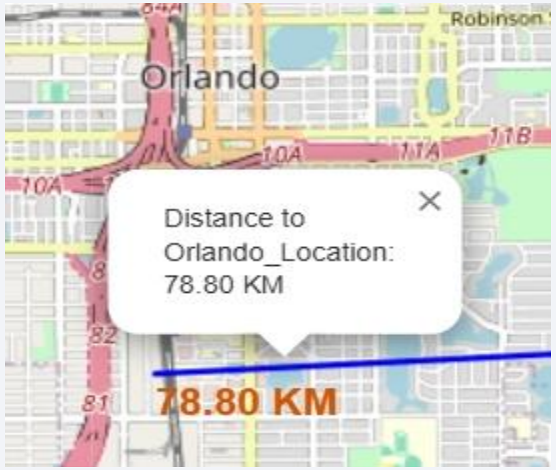


- Green marker shows Successful missions.
- Red marker show Failure missions.
- KSC LC-39A has the best track record for successful launches
- CCAFS LC-40 has had the most launches, but a lower percentage of them were successful.

Launch site distance to landmarks





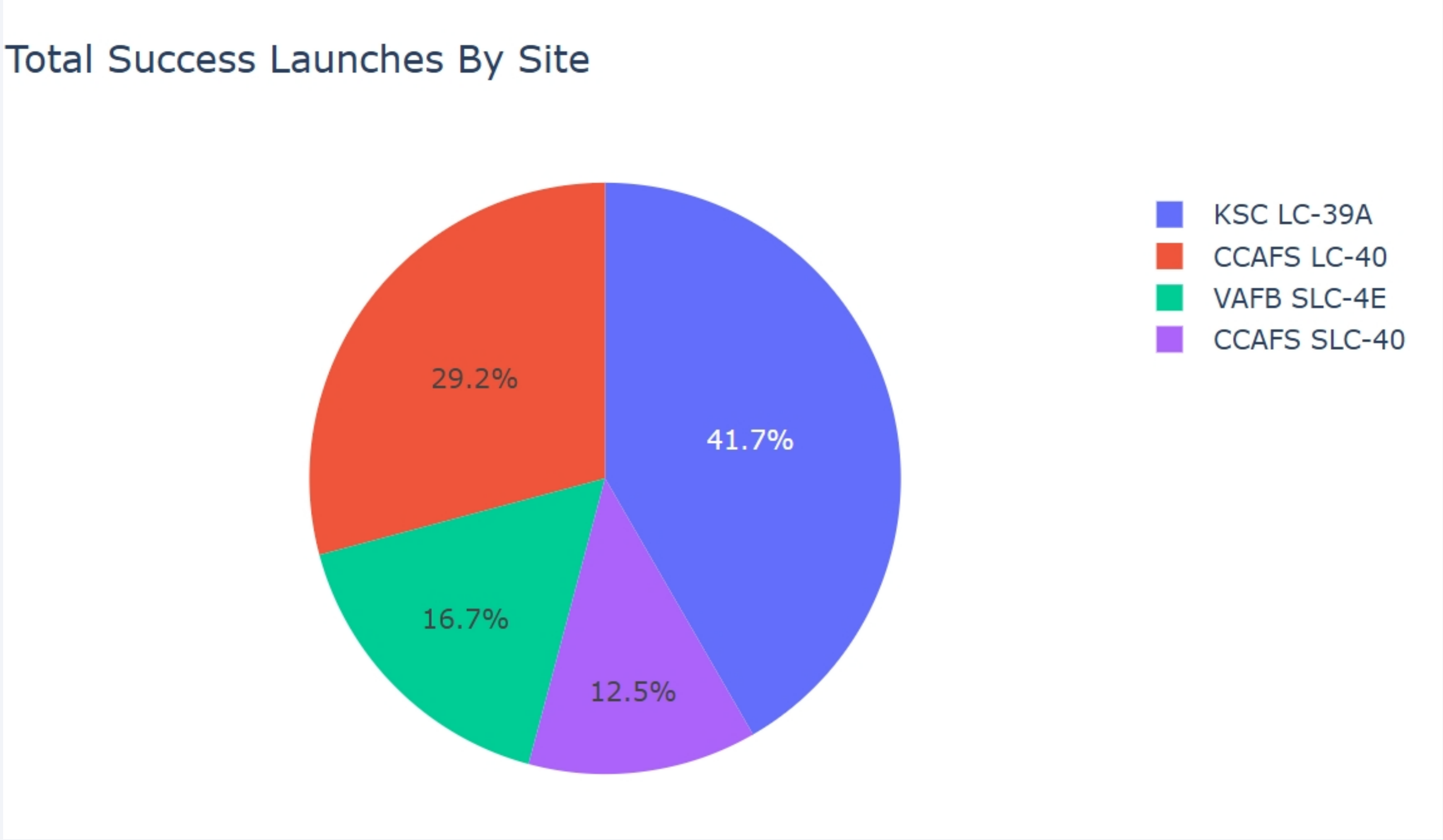


- Launch sites are closed to railways, highways and Coastline.
- The distance from launch site to cities is far enough for safety with the closest city 78.80 KM (Orlando)

Section 4

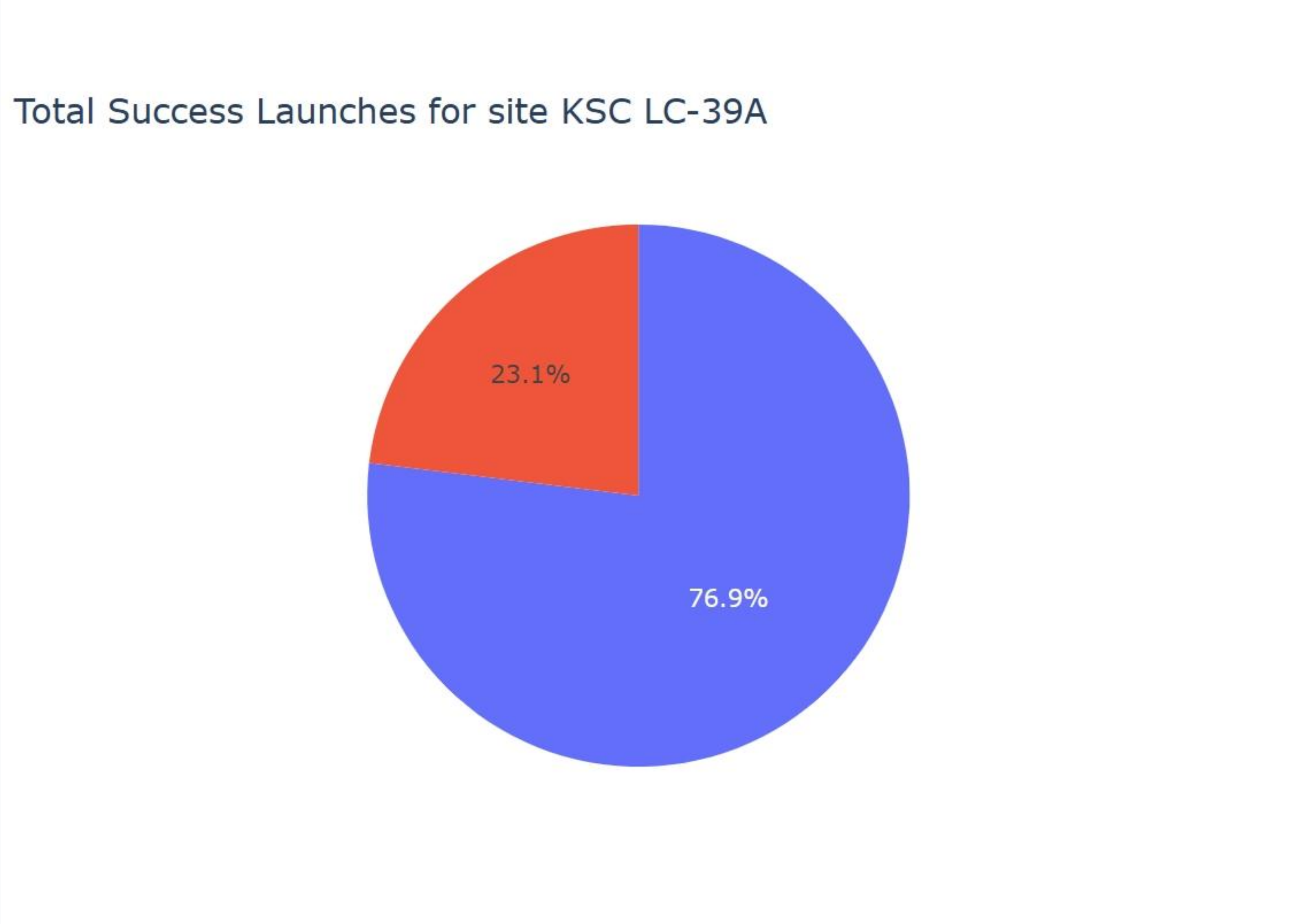
Build a Dashboard with Plotly Dash

Success percentage by Launch sites



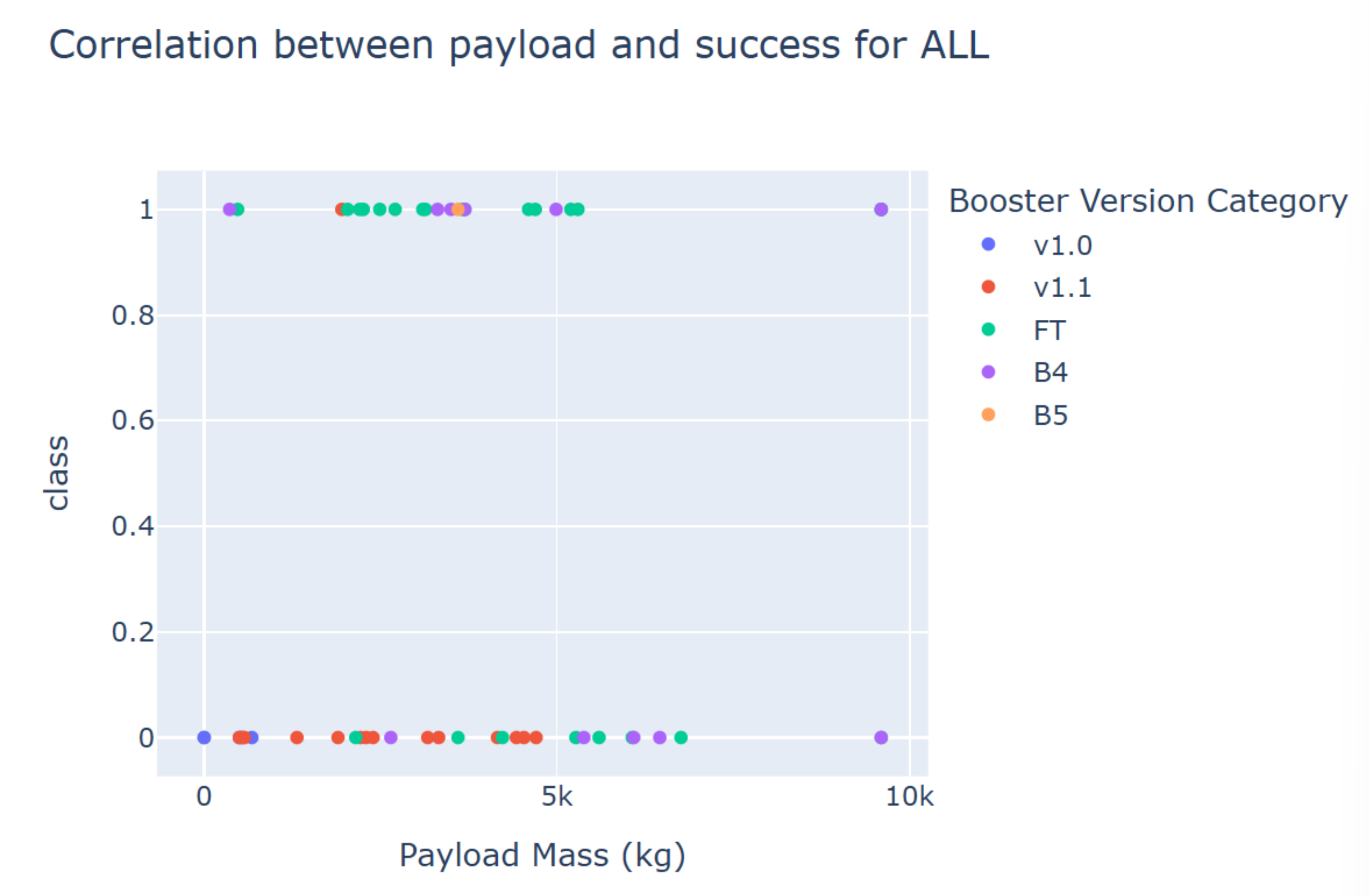
- Site KSC LC-39A had the most successful launches. It indicate that this site is the most suitable for experiments.
- Site CCAFS SLC-40 had the lowest successful launches.

The highest success rate launch site (KSC LC-39A)



- KSC LC-39A had the success rate of 76.9%.
- KSC LC-39A had the failure rate of 23.1%.

Impact of Payload Mass on Launch Success

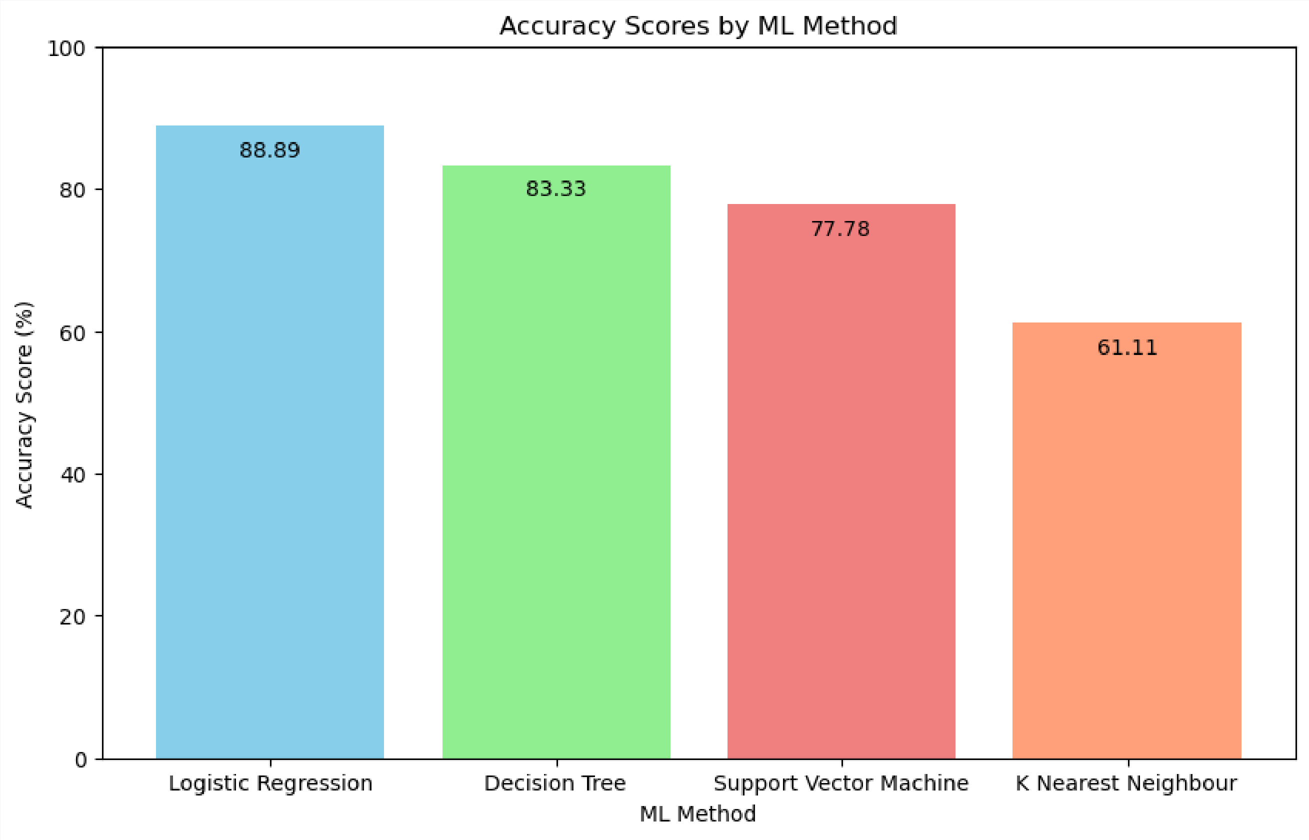


- The FT booster version had the highest success rate for payloads under 5,400 kg.
- Booster version v1.1 and v1.0 both had very low success rates.
- The B4 booster version was successful for payloads up to 5,000 kg, but failed for higher payloads.

Section 5

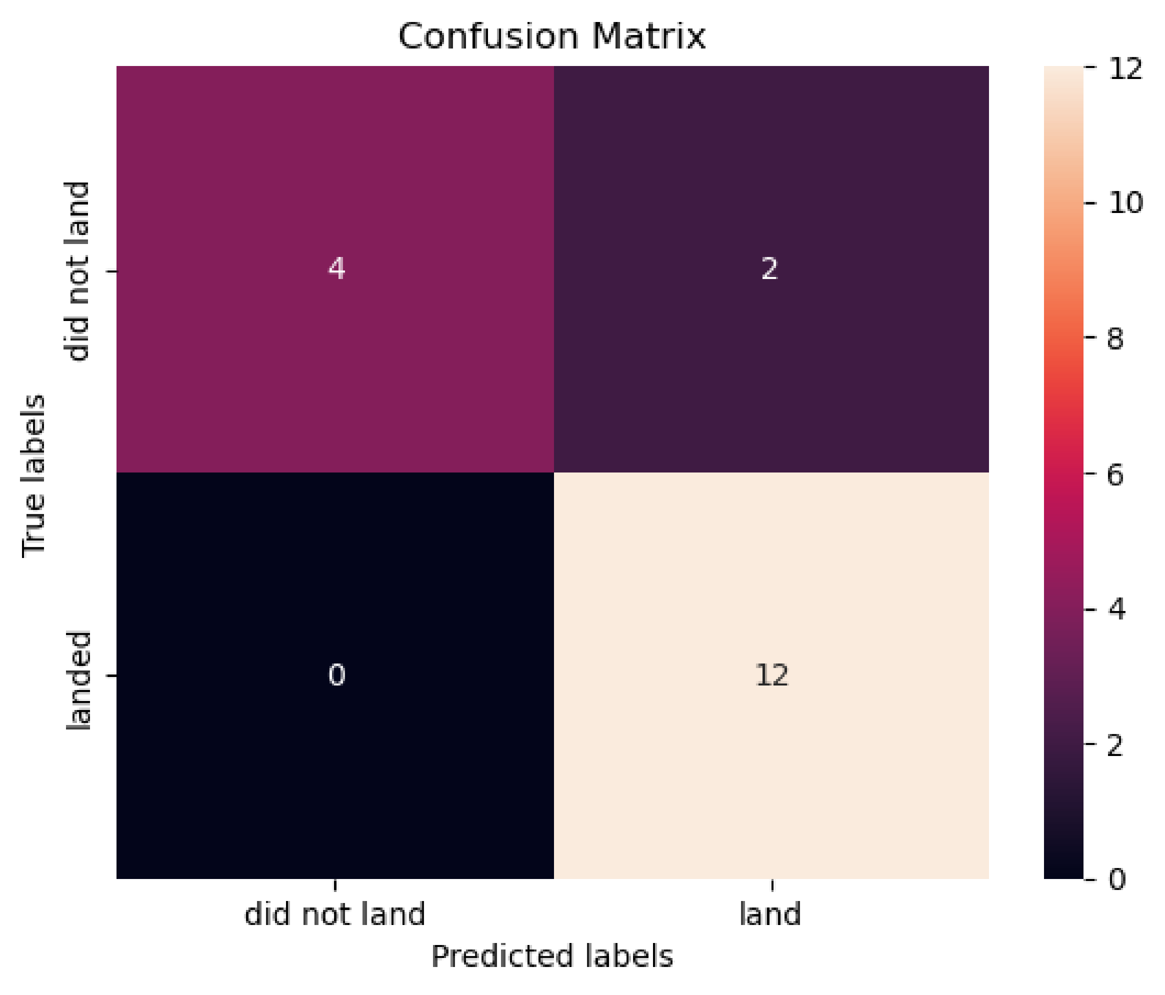
Predictive Analysis (Classification)

Classification Accuracy



"Logistic regression had the highest model accuracy"

Confusion Matrix of Logistic regression



- Accuracy: Measures the overall correctness of the model's predictions.
  - Calculated as: (TP + TN) / (TP + TN + FP + FN) = 89%
  - The model has a high accuracy of 89%.
- Precision: how many were actually correct?
  - Calculated as: TP / (TP + FP) = 86%
  - The model is 86% precise when it predicts a landing.
- Recall: how many did the model correctly predict?
  - Calculated as: TP / (TP + FN) = 100%
  - The model **captures all actual landings**.
- F1-Score: A combined measure of precision and recall
  - Calculated as: 2 * (Precision * Recall) / (Precision + Recall) = 92%
  - A strong F1-score of 92% indicates good balance between precision and recall.

Conclusions

We can conclude that:

- Launch Success Trend: Success rates significantly improved from 2013-2020, suggesting a link between experience and success.
- Top Launch Site: KSC LC-39A boasts the highest success rate among analyzed launch sites.
- Best Predictive Model: Logistic regression appears most effective, among tested models, for predicting future launch outcomes.
- Payload and Orbit Correlation: Heavier payloads correlate with higher success rates for PO, LEO, and ISS missions.

Thank you!