IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Georgios Bakirtzis
01/03/2025

Outline

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

Executive Summary

- The Methodologies used to Predict Falcon 9 first stage landing success to help estimate launch costs and inform competitive bidding include:
  - Data Collection (API & Web Scraping)
  - Data Wrangling
  - Exploratory Data Analysis (EDA) with Visualization & SQL
  - Feature Engineering
  - Interactive Mapping (Folium)
  - Dashboard Analysis (Plotly Dash)
  - Predictive Analysis (Classification Models)
- Summary of all results: A robust model and interactive tools offering actionable insights into launch success determinants.

Introduction

- Background:
  - SpaceX’s innovative reuse strategy for the Falcon 9 first stage drives cost savings.
- Problem Statement:
  - Predicting landing success is critical for estimating launch costs and improving reliability.
- Scope:
  - Data collected via SpaceX API and web scraping (Wikipedia snapshot).
  - Analysis through EDA, SQL, interactive mapping, dashboard development, and machine learning.

Section 1

Methodology

Methodology

Executive Summary

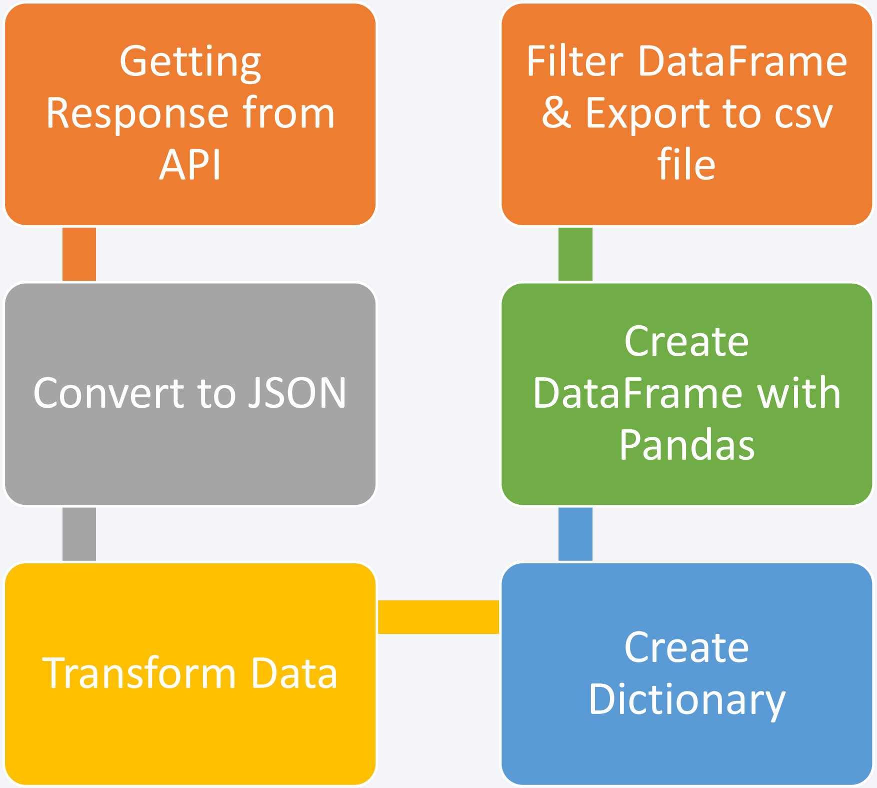
- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - How to build, tune, evaluate classification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

Data Collection

- Process:
  - Used GET requests to retrieve SpaceX launch data from a static JSON URL.
  - JSON response normalized into a Pandas DataFrame.
- Key Points:

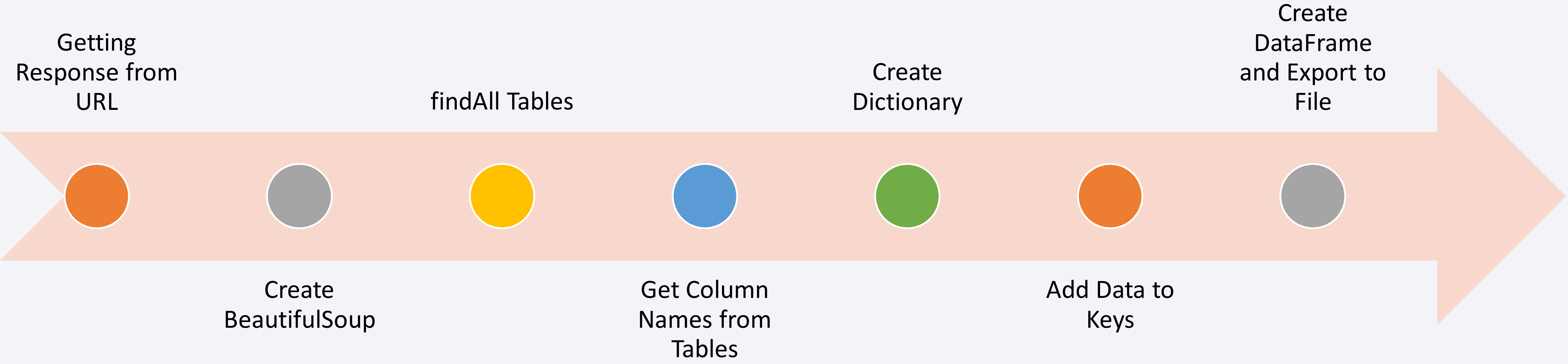
  Extracted features include Booster Version, Payload Mass & Orbit, Launch Site details, and Core information.
- GitHub URL: https://github.com/mpaki22/ADSC.git

Data Collection – SpaceX API



GitHub URL: https://github.com/mpaki22/ADSC.git

Data Collection - Scraping



GitHub URL: https://github.com/mpaki22/ADSC.git

Data Wrangling

- Steps Performed:
  - Filtered out rows with multiple cores or payloads.
  - Converted date strings to datetime objects and restricted launch dates.
  - Replaced missing values (e.g., PayloadMass) with computed means.
- Outcome:
  - A clean, structured dataset focused on Falcon 9 launches.

Number of Launches on each site → Number & Occurrence of each Orbit → Number & Occurrence of Mission Outcome → Create a Landing Outcome Label from Outcome Column → Export to CSV file

GitHub URL: https://github.com/mpaki22/ADSC.git

EDA with Data Visualization

- Techniques:
  - Plotted various charts (scatter plots, bar charts, line charts) to reveal trends.
- Key Analyses:
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Success Rate by Orbit Type
  - Yearly Launch Success Trends

GitHub URL: https://github.com/mpaki22/ADSC.git

EDA with SQL

- SQL queries performed in the Capstone Project:
  - 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)
  - Display average payload mass carried by booster version F9 v1.1
  - 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.
  - 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.
  - 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

- Launch Site Markers → Plotted all launch sites with clickable labels.
- NASA Johnson Space Center Reference Point → Added a blue circle marker.
- Mouse Position Tracker → Displays real-time latitude/longitude coordinates when hovering over the map.
- Clustered Markers → Grouped nearby launches for better visualization.
- Success/Failure Indicators → Used different marker colors/icons for launch success/failure.

GitHub URL: https://github.com/mpaki22/ADSC.git

Build a Dashboard with Plotly Dash

- Launch Site Dropdown → Select a specific site or view all sites.
- Success-Pie Chart → Displays success vs. failure counts for selected sites.
- Payload Range Slider → Filters launches based on payload mass.
- Scatter Plot (Payload vs. Success) → Identifies success patterns by booster version.
- Dynamic Updates → Graphs update instantly based on user selections.

GitHub URL: https://github.com/mpaki22/ADSC.git

Predictive Analysis (Classification)

- Objective:
  - Build classification models to predict landing success
- .Process:
  - Data split into training/testing sets, feature selection, and one-hot encoded inputs.
  - Multiple models trained and tuned.
- Evaluation Metrics:
  - Accuracy, confusion matrix.

GitHub URL: https://github.com/mpaki22/ADSC.git

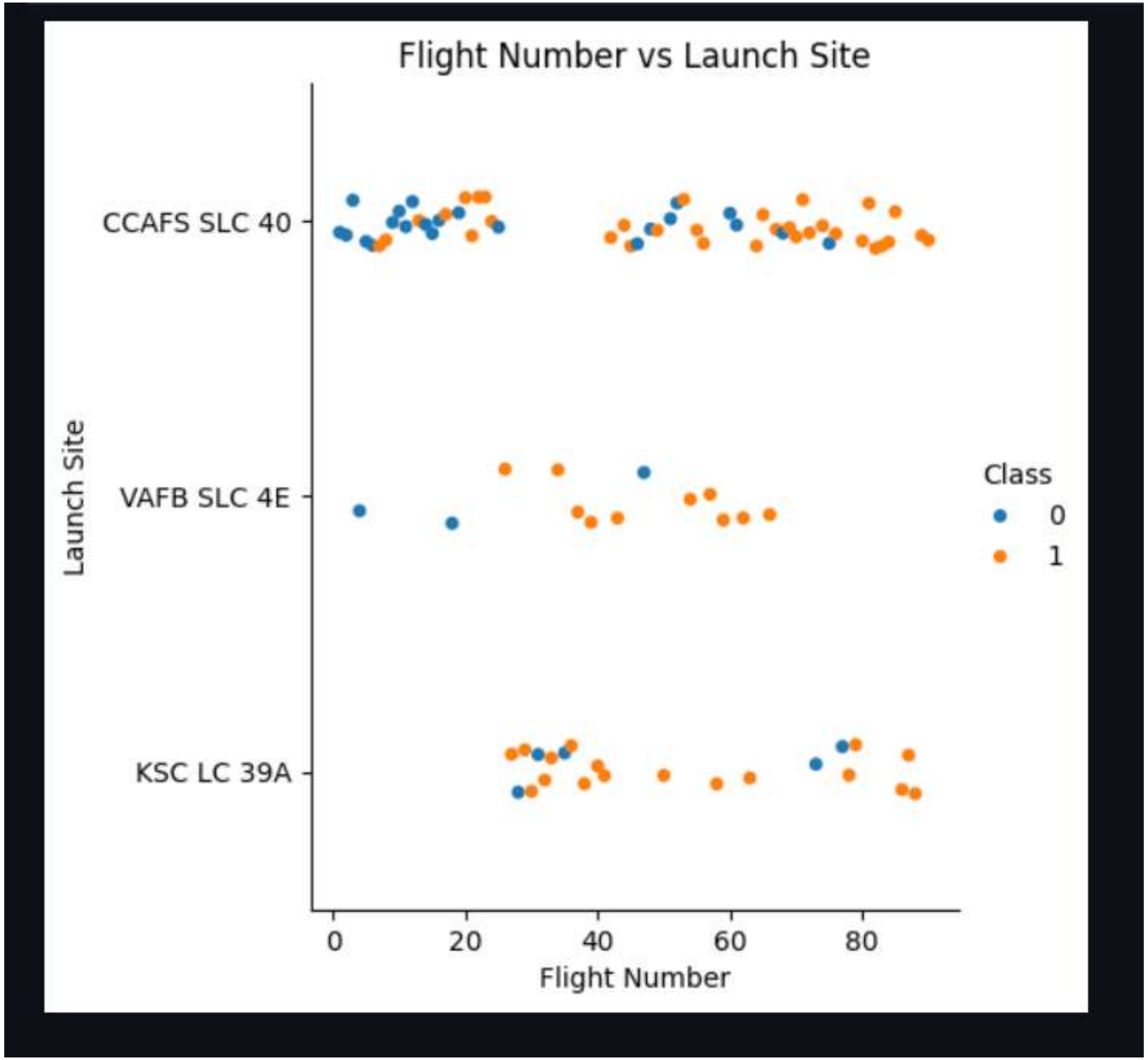
Results

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

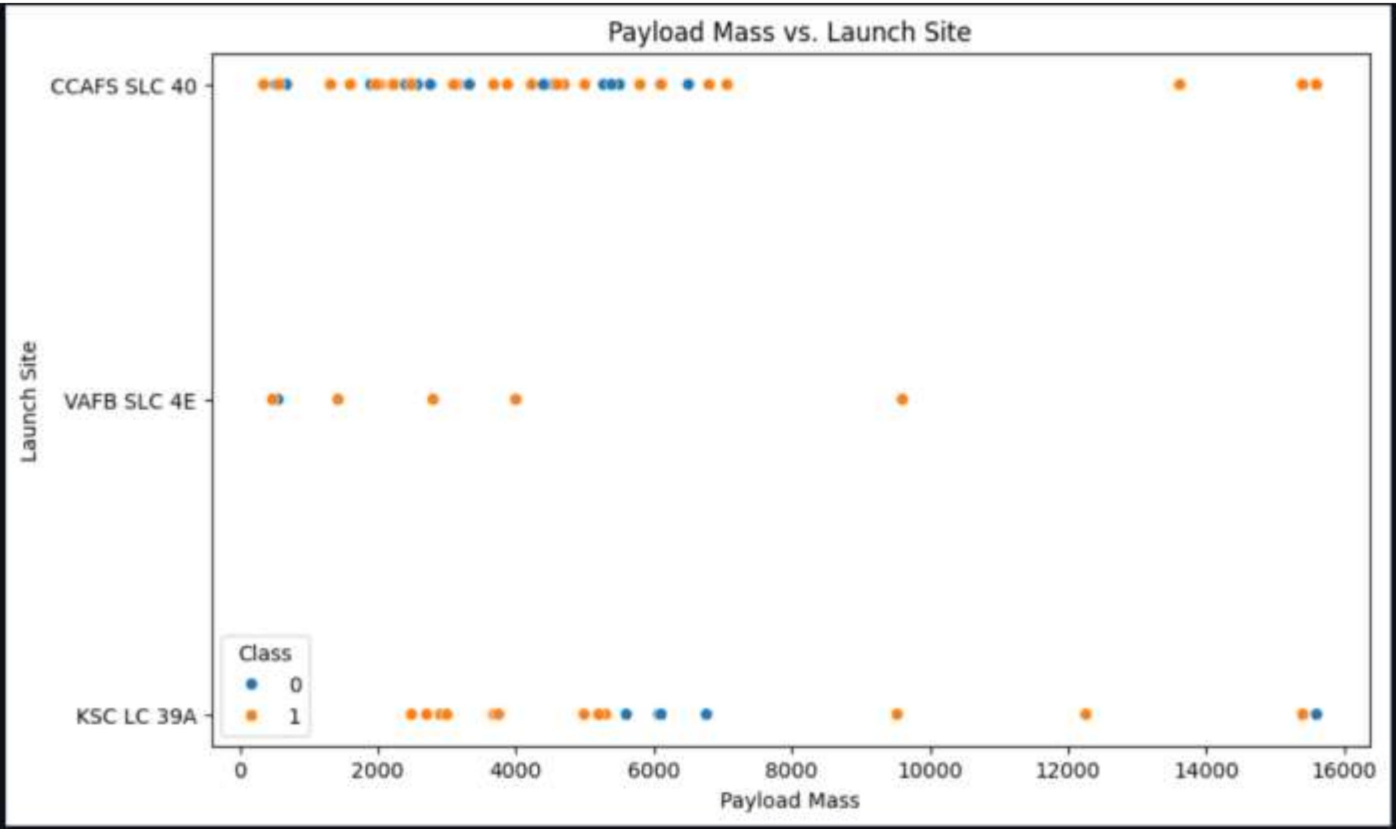
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

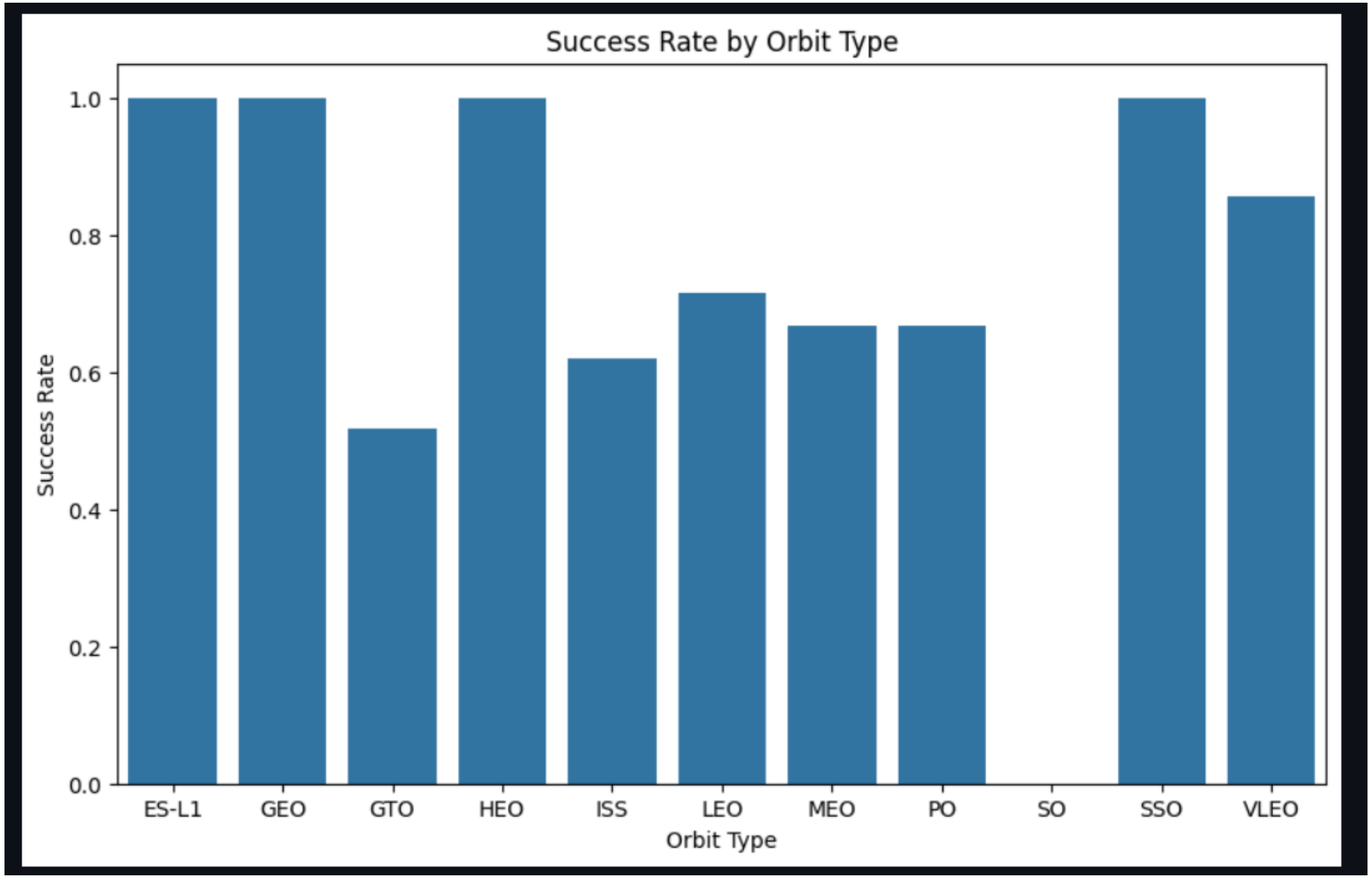


Payload vs. Launch Site

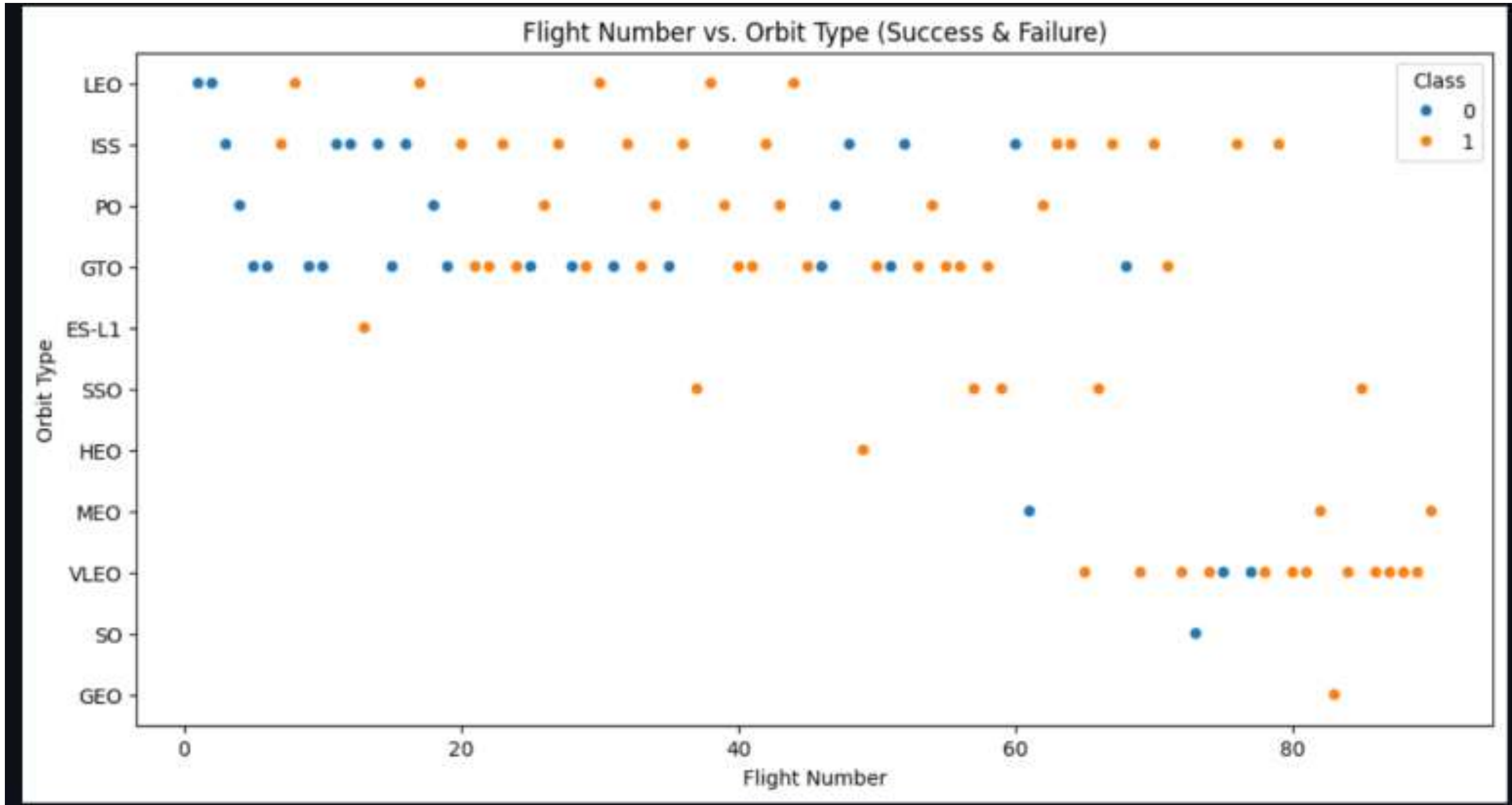


Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

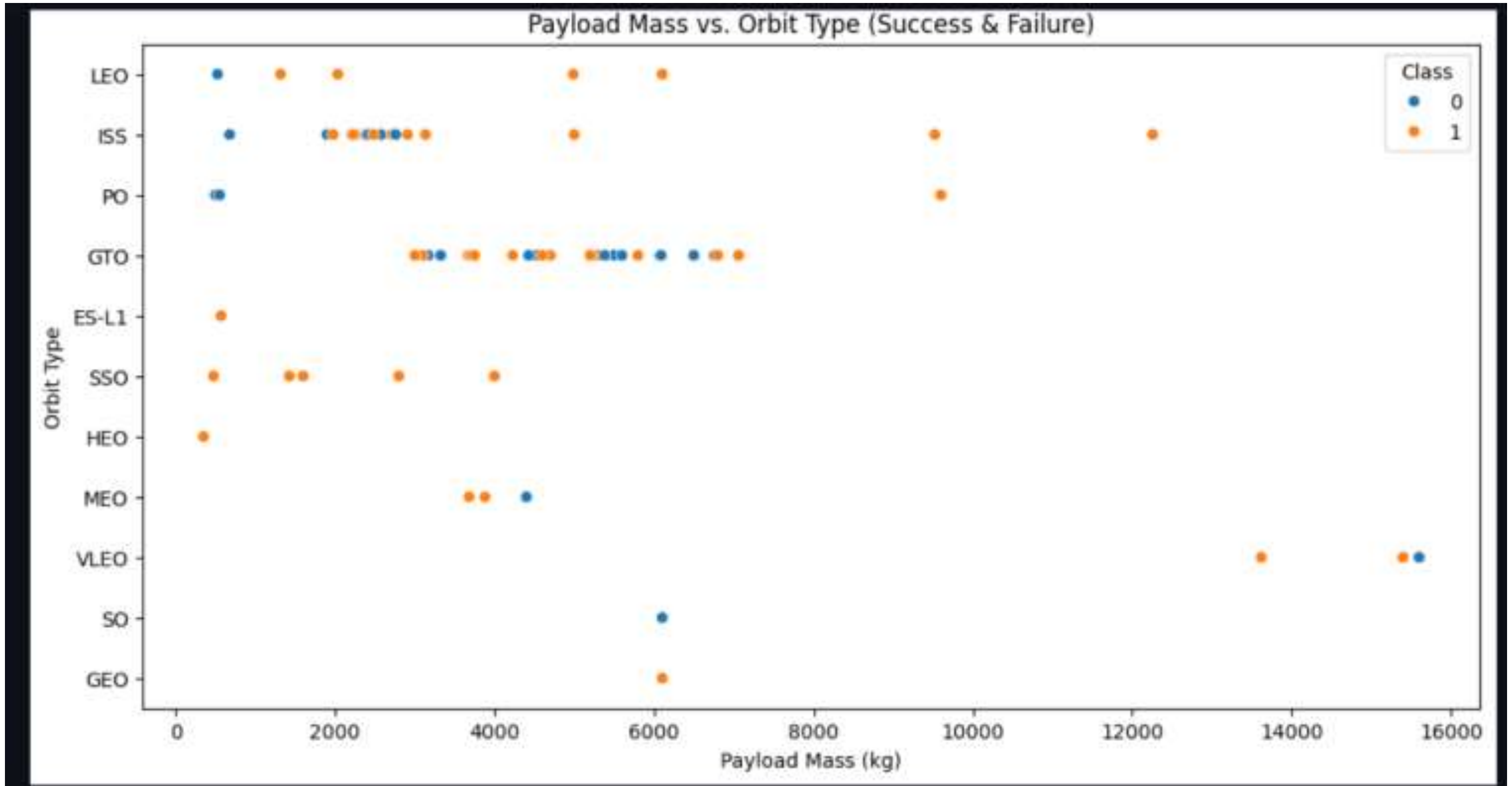


Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

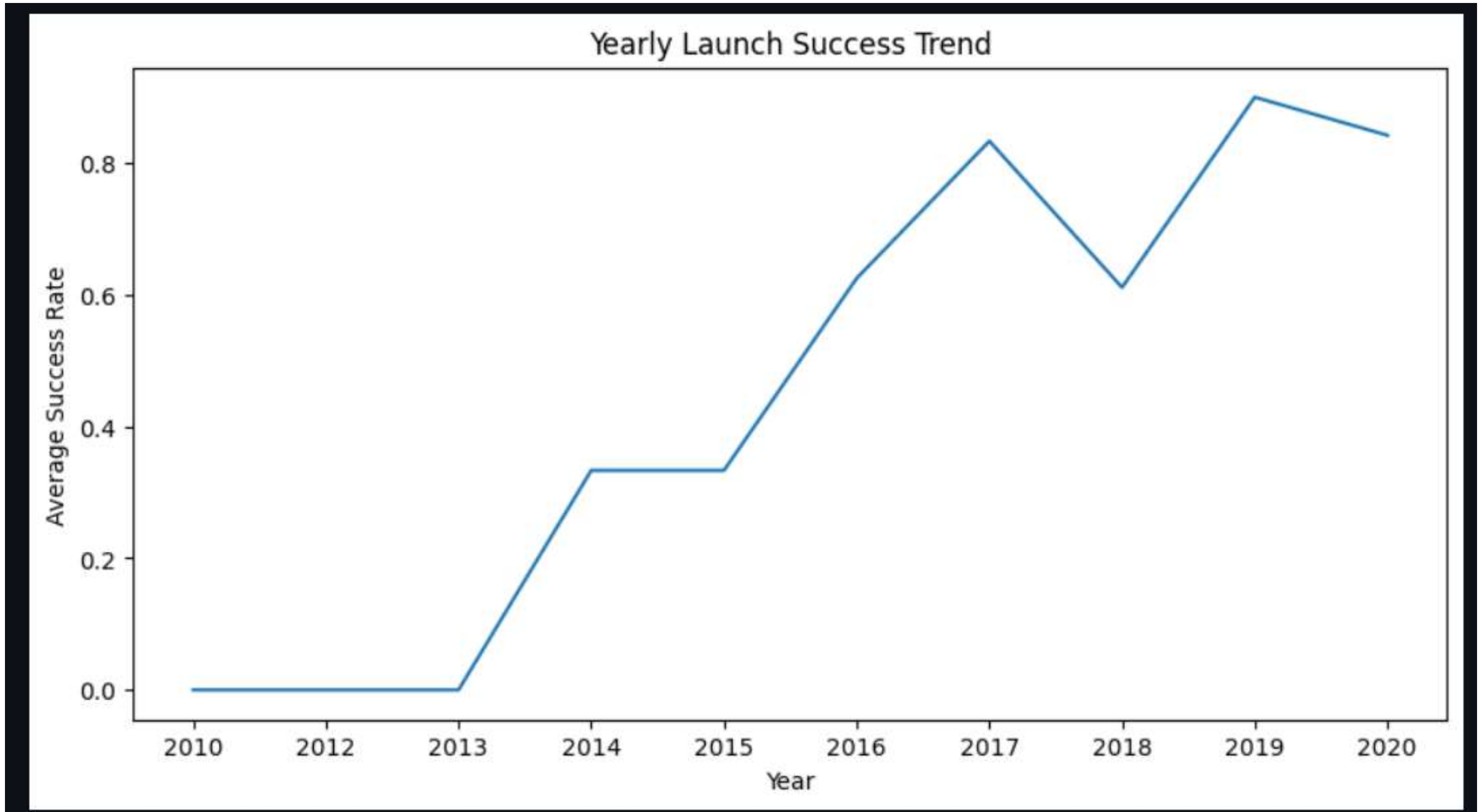
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
%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'

Task 2

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

```
%sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

```
 * 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

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

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

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 "DATE" FROM SPACEXTBL WHERE "LANDING_OUTCOME" LIKE '%Success (ground pad)%'
```

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

DATE
2015-12-22

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 "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING_OUTCOME" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MAS
```

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

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, (SELECT COUNT("MISSION_OUTC
```

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

SUCCESS	FAILURE
100	1

Boosters Carried Maximum Payload

Task 8

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_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

```
 * 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

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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr("DATE", 6, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL WHERE "LANDING_OUTCOME" = 'Failure (drone ship)' ar
```

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

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

Task 10

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.

```
%sql SELECT "LANDING_OUTCOME", COUNT(*) AS "COUNT" FROM SPACEXTBL WHERE "DATE" BETWEEN '2010-06-04' AND '2017-03-30' GROUP BY "LANDING_OUTCOME
```

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

Landing_Outcome	COUNT
No attempt	10
Success (drone ship)	6
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

Launch Sites on the map

- Launch sites are distributed in the coastline of the United States, with most of them on the east side.

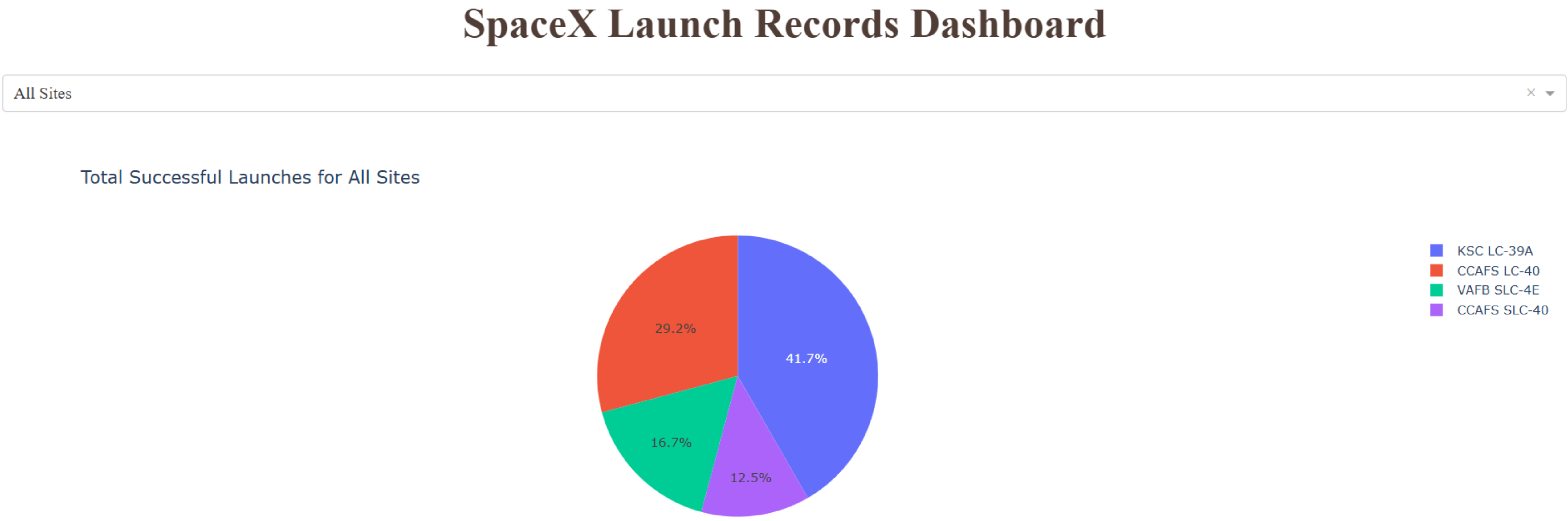
Success/Failed launches examples

- Green marker represents successful launches.
- Red marker represents unsuccessful launches.

Distance between a launch site to its proximities

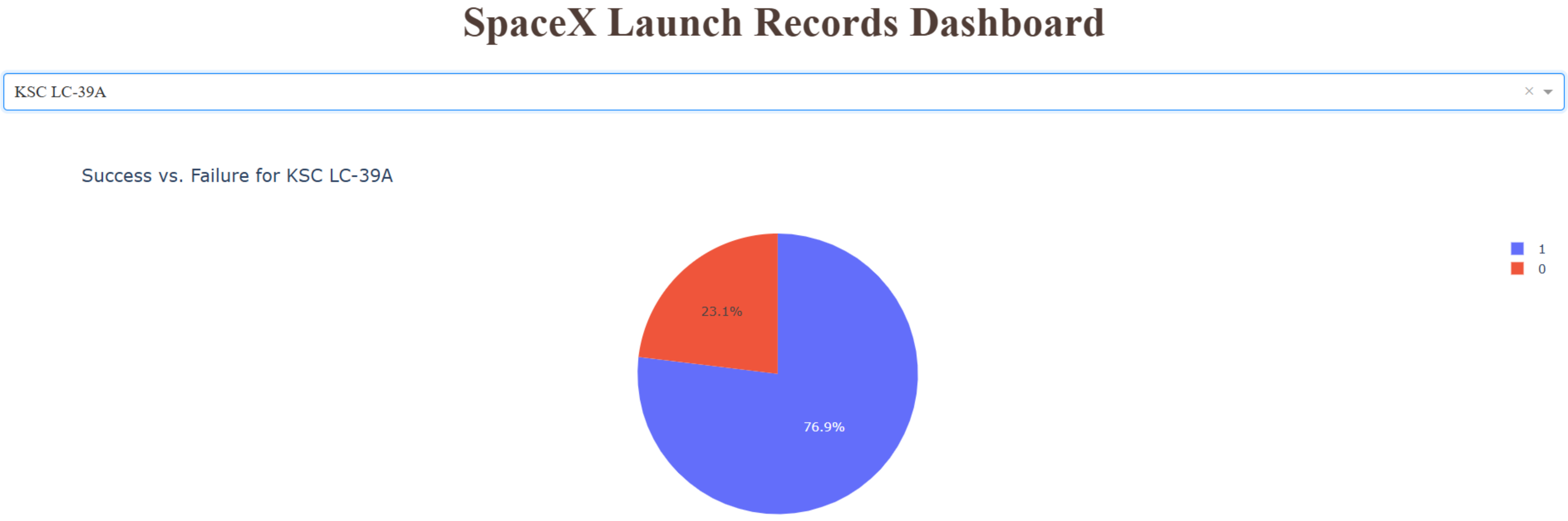
Section 4

Build a Dashboard with Plotly Dash



Total Success Launches (%) by Site

- KSC LC-39A has obviously the best percentage

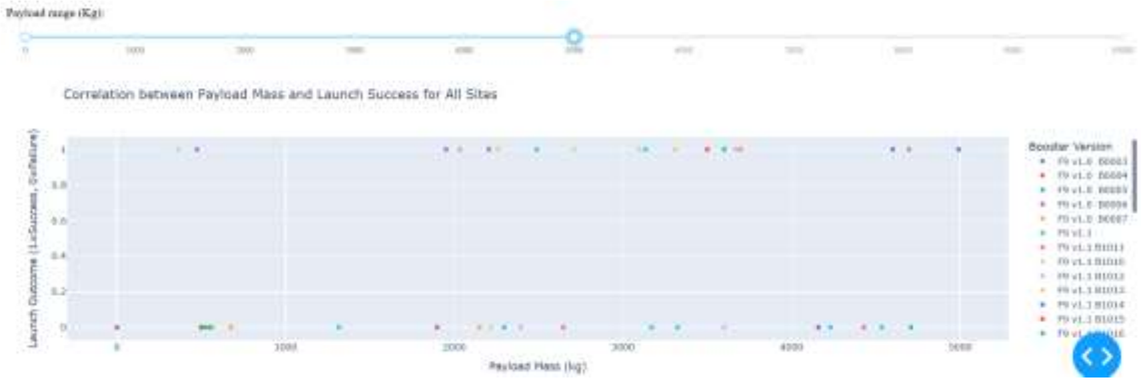


Success Launches for KSC LC-39A

- The success rate for the site is almost 77%

Correlation between payload and success for different payload ranges

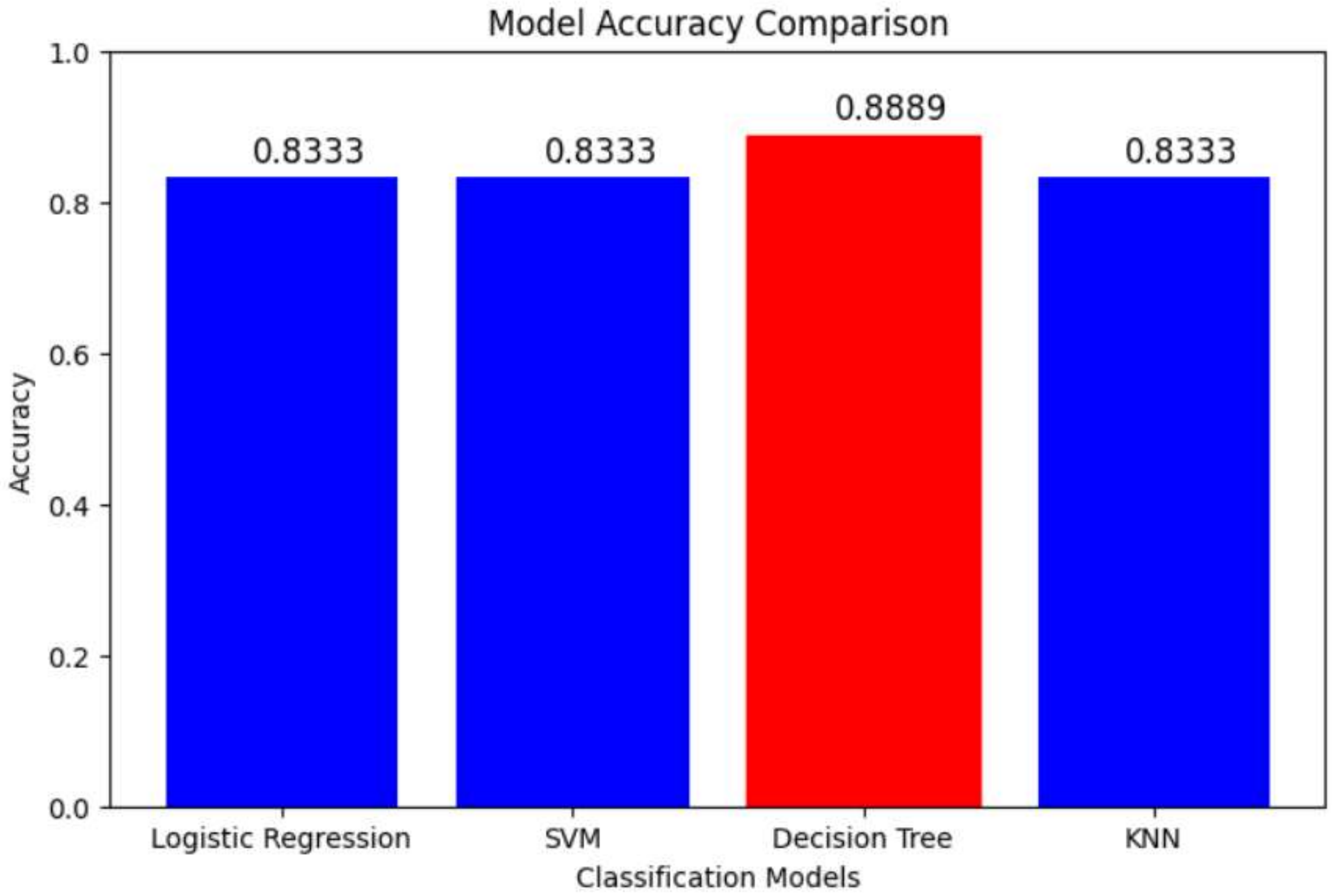
Low-mass payloads have better success rate than the large-mass payloads.



Section 5

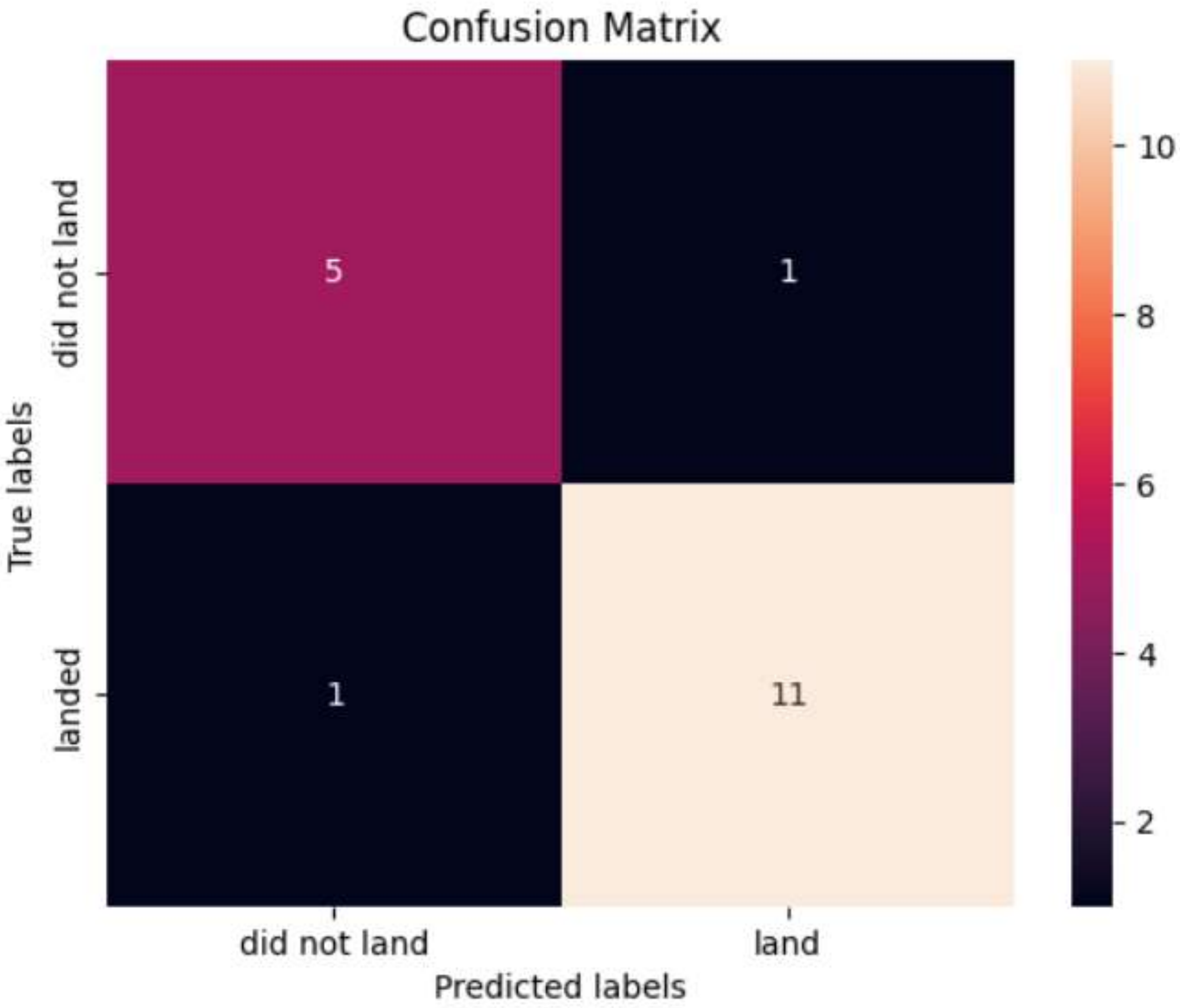
Predictive Analysis (Classification)

Classification Accuracy




```
Best performing model: Decision Tree with an accuracy of 0.8889
```

Confusion Matrix



Conclusions

- Key Takeaways:
  - Comprehensive analysis from data collection to predictive modeling.
  - Clear trends and operational insights extracted from SpaceX launch data.
  - Strong foundation for further research and real-world application.
- Final Thoughts:
  - The project demonstrates the value of integrated data science approaches in solving complex aerospace challenges.

Thank you!