IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lydia Gyenes
17.04.2025.

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- **Project background and context**
  SpaceX stands out as the leading company in the era of commercial space exploration, having made space missions significantly more cost-effective. The Falcon 9 rocket launches are listed on the company's website at a price of 62 million dollars, whereas other providers typically charge upwards of 165 million dollars. A major factor in these cost savings is SpaceX's ability to reuse the rocket's first stage. Consequently, if we can predict whether the first stage will successfully land, we can also estimate the overall launch cost. Using publicly available data and machine learning techniques, this project aims to predict the reusability of SpaceX's first-stage rockets.
- **Questions to be answered**
  - How do features like payload mass, launch site, number of prior flights, and orbit type influence the likelihood of first-stage landing success?
  - Has the success rate of landings improved over time?
  - Which machine learning algorithm performs best for binary classification in this scenario?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

The data for this project was collected using two complementary methods: API requests to the SpaceX REST API and web scraping from a table on SpaceX's Wikipedia page. Both approaches were necessary to obtain complete and detailed information about the rocket launches, allowing for a more thorough analysis.

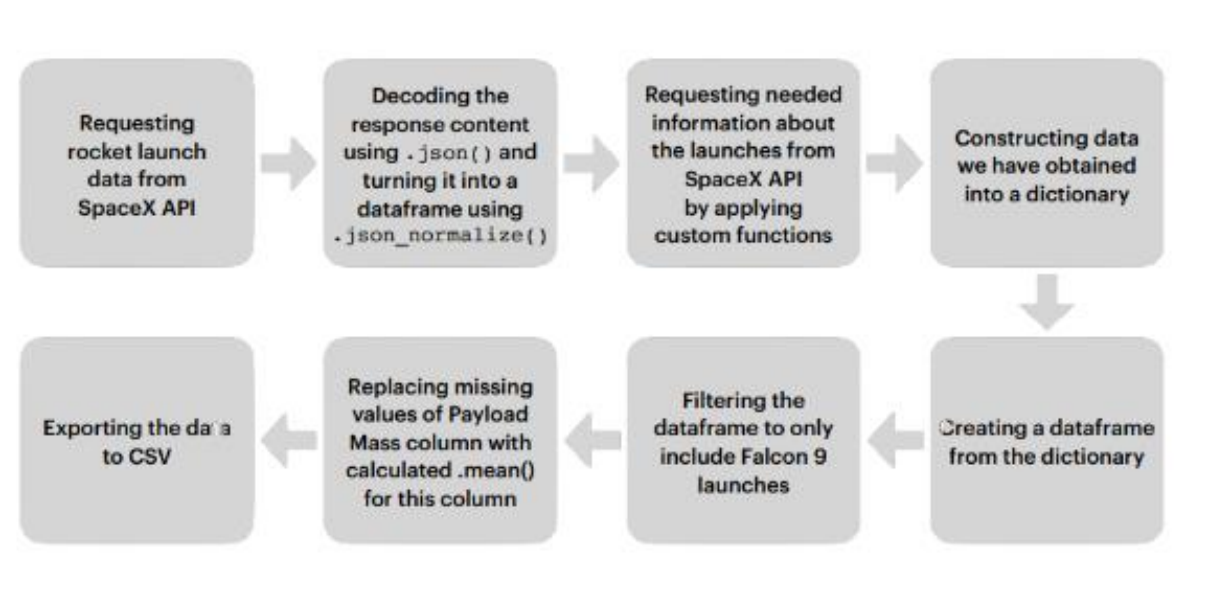
From the SpaceX REST API, we gathered the following data fields: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

In addition, web scraping from Wikipedia provided us with further data columns such as Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API

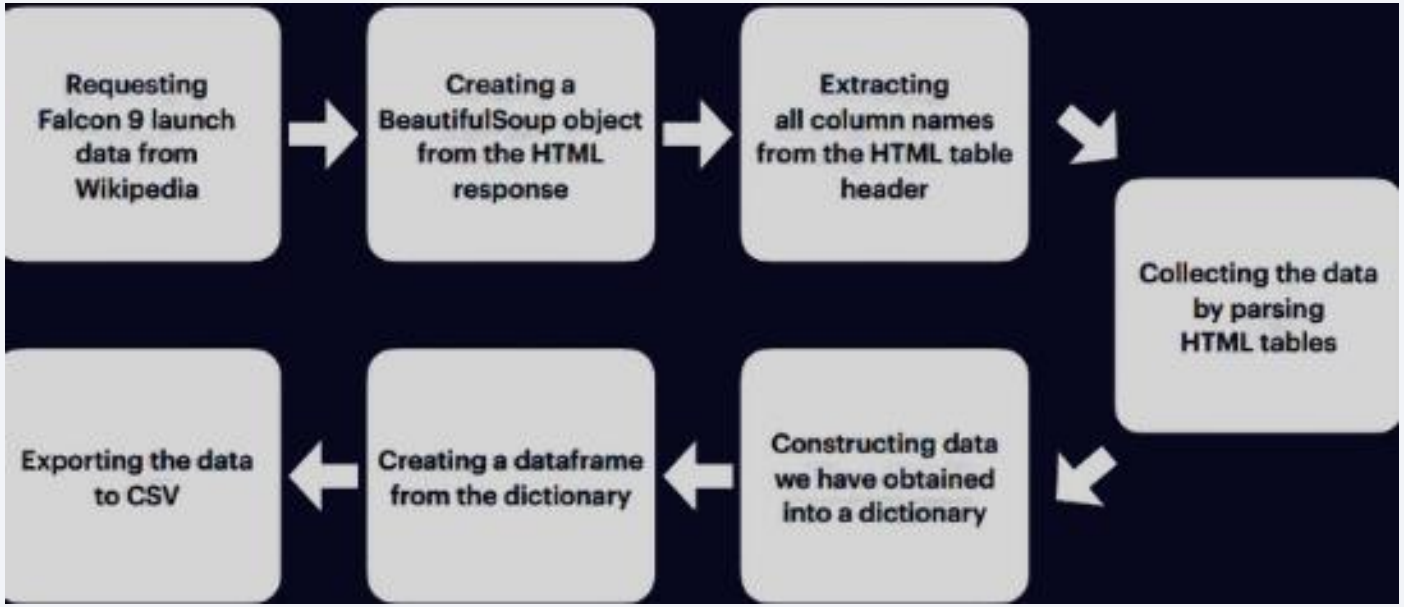
- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/Lydiagyenes/IBM-Applied-Data-Science/blob/main/Data%20Collection%20API.ipynb



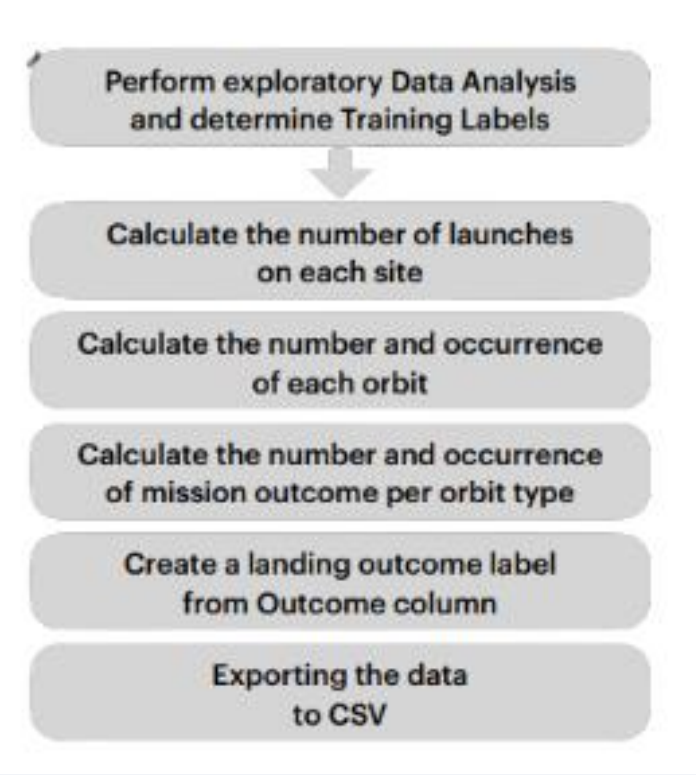
Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- https://github.com/Lydiagyenes/IBM-Applied-Data-Science/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Within the dataset, there are several different scenarios where the booster either landed successfully or failed to do so. In some cases, a landing was attempted but did not succeed due to accidents or technical issues. For instance, a label such as True Ocean indicates that the booster successfully landed in a designated ocean area, while False Ocean means the attempt to land in the ocean failed. Similarly, True RTLS (Return to Launch Site) represents a successful landing on a ground pad, whereas False RTLS indicates a failed ground pad landing. True ASDS refers to a successful landing on a drone ship (Autonomous Spaceport Drone Ship), and False ASDS means the booster did not successfully land on the drone ship.To simplify these outcomes for machine learning purposes, we converted them into binary training labels: a value of “1” represents a successful landing, and a value of “0” indicates an unsuccessful landing.



- https://github.com/Lydiagyenes/IBM-Applied-Data-Science/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- Several charts were created to explore relationships and trends in the dataset, 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 a yearly trend of the Success Rate.
- **Scatter plots** were used to visualize the relationships between continuous variables. When a clear relationship is observed, these plots can provide insights that are valuable for building machine learning models.
- **Bar charts** were utilized to compare values across discrete categories. These charts help highlight the connection between specific categories and the values being measured.
- **Line charts** were employed to illustrate trends over time, making them particularly useful for analyzing changes and patterns in time series data, such as the success rate of booster landings over the years.
- https://github.com/Lydiagyenes/IBM-Applied-Data-Science/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- A series of SQL queries were executed to extract specific insights from the dataset:
- Retrieved the names of all unique launch sites involved in space missions.
- Displayed five records where the launch site names start with the string "CCA".
- Calculated the total payload mass carried by boosters launched under NASA's CRS missions.
- Computed the average payload mass delivered by boosters of the version F9 v1.1.
- Identified the date when the first successful landing on a ground pad occurred.
- Listed the names of boosters that successfully landed on a drone ship and carried a payload mass between 4000 and 6000 kg.
- Counted the total number of successful and failed mission outcomes.
- Determined which booster versions carried the highest payload mass.
- Retrieved records of failed drone ship landings in 2015, including the booster versions and corresponding launch site names.
- Ranked the frequency of different landing outcomes (e.g., "Failure (drone ship)", "Success (ground pad)") between June 4, 2010, and March 20, 2017, in descending order.

- https://github.com/Lydiagyenes/IBM-Applied-Data-Science/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

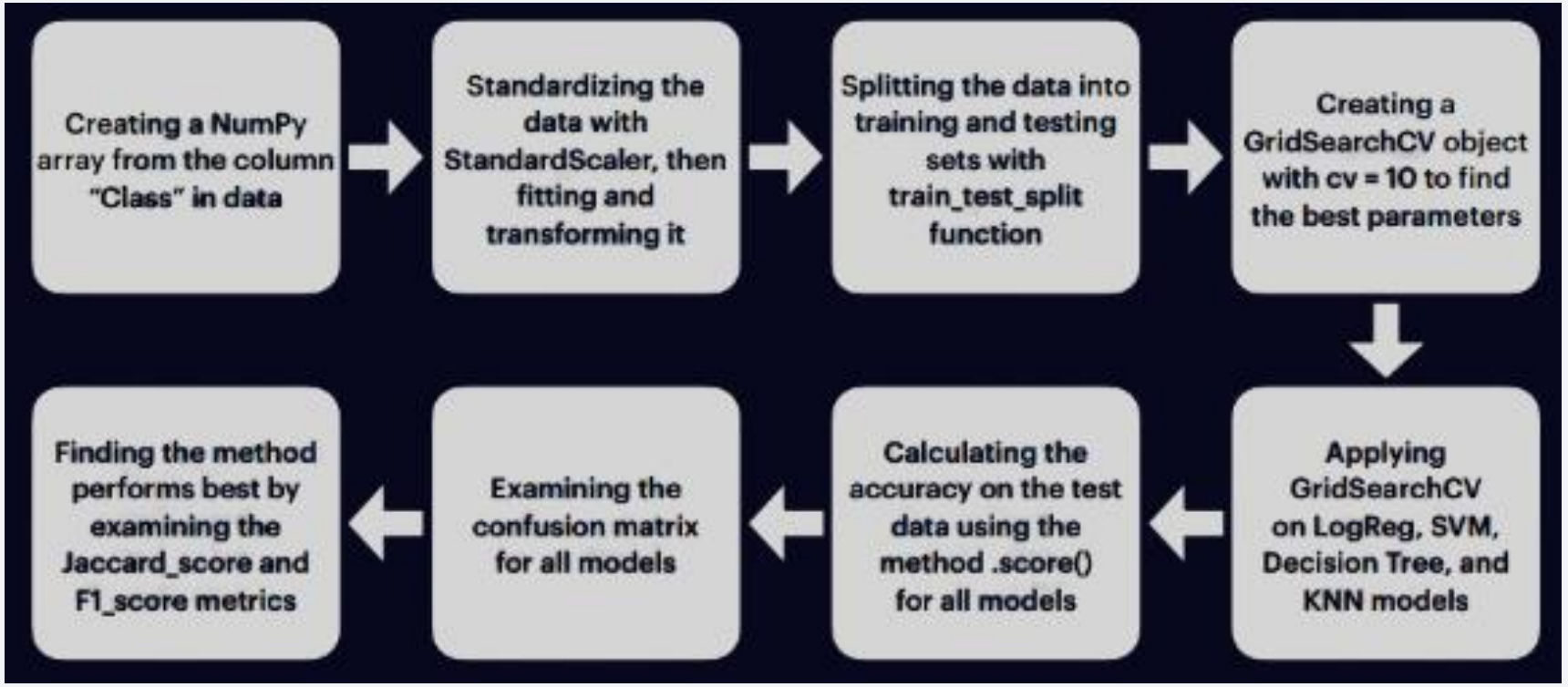
- **GitHub URL: Interactive Visual Analytics with Folium**
  The following steps were taken to create an interactive visual representation of launch sites and their outcomes using Folium:
- **Markers for Launch Sites:**
- A marker was added for **NASA Johnson Space Center**, including a circle, popup label, and text label, using its latitude and longitude coordinates as the starting point.
- Markers with circles, popup labels, and text labels were also added for **all launch sites**, showing their geographical locations along with their proximity to the Equator and coastlines.
- **Colored Markers for Launch Outcomes:**
- Colored markers were implemented to represent the **launch outcomes**: **Green** for successful launches and **Red** for failed ones. A marker cluster was used to visualize which launch sites had relatively high success rates.
- **Distances Between Launch Sites and Proximities:**
- Colored lines were drawn to illustrate the **distances** between a selected launch site, such as **KSC LC-39A**, and nearby points of interest, such as the railway, highway, coastline, and the closest city.
- https://github.com/Lydiagyenes/IBM-Applied-Data-Science/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List:**
- A **dropdown list** was implemented to allow users to select a specific launch site.
- **Pie Chart for Success Launches (All Sites/Certain Site):**
- A **pie chart** was added to display the total count of successful launches across all sites. When a specific launch site is selected, the chart updates to show the **Success vs. Failure** counts for that site.
- **Slider for Payload Mass Range:**
- A **slider** was integrated to enable users to select a range of payload masses.
- **Scatter Chart of Payload Mass vs. Success Rate for Different Booster Versions:**
- A **scatter chart** was added to illustrate the relationship between **payload mass** and **launch success**, highlighting the differences across various booster versions.
- https://github.com/Lydiagyenes/IBM-Applied-Data-Science/blob/main/spacex_dash_app.py

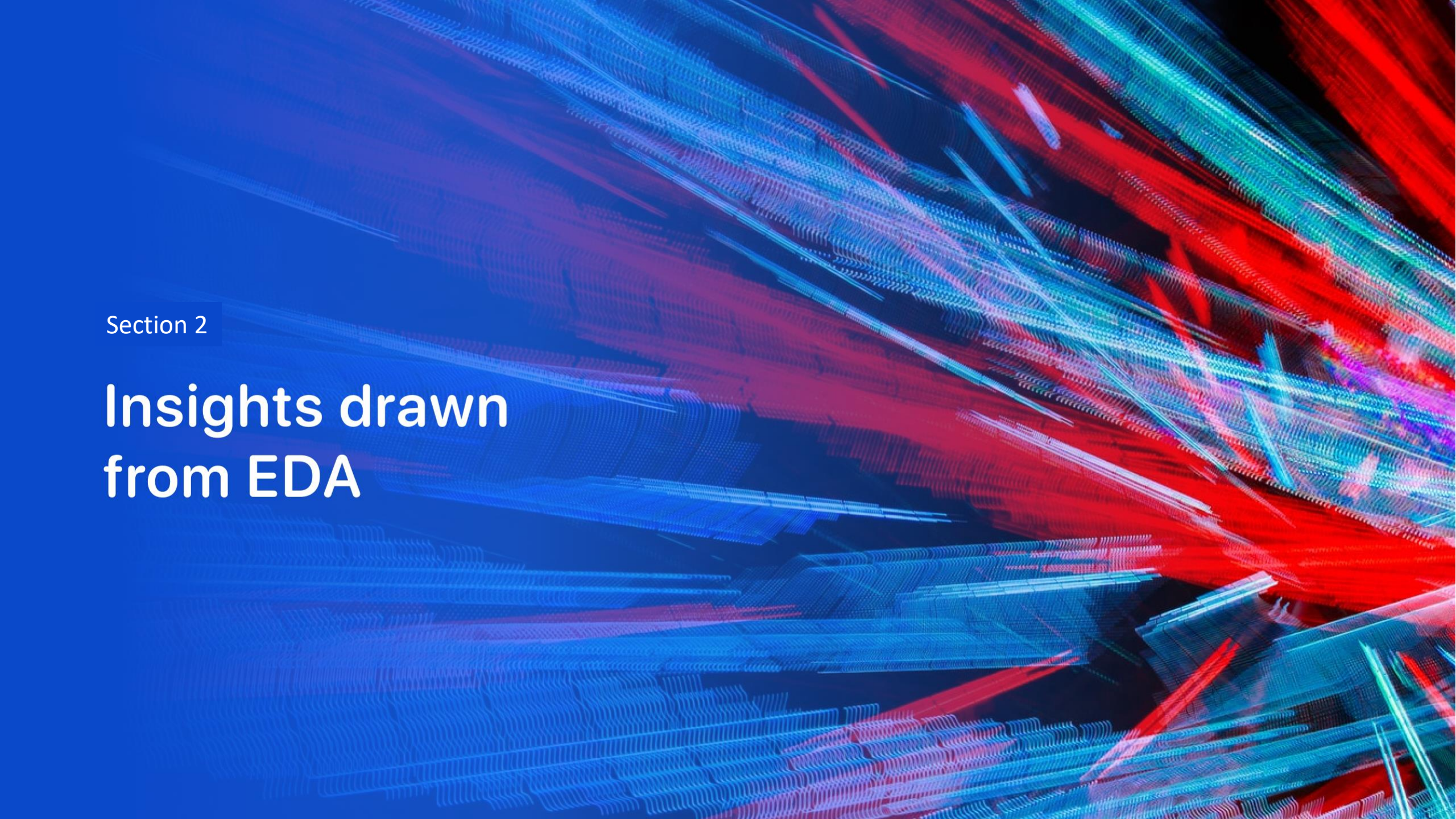
Predictive Analysis (Classification)

- https://github.com/Lydiagyenes/IBM-Applied-Data-Science/blob/main/Machine%20Learning%20Prediction.ipynb

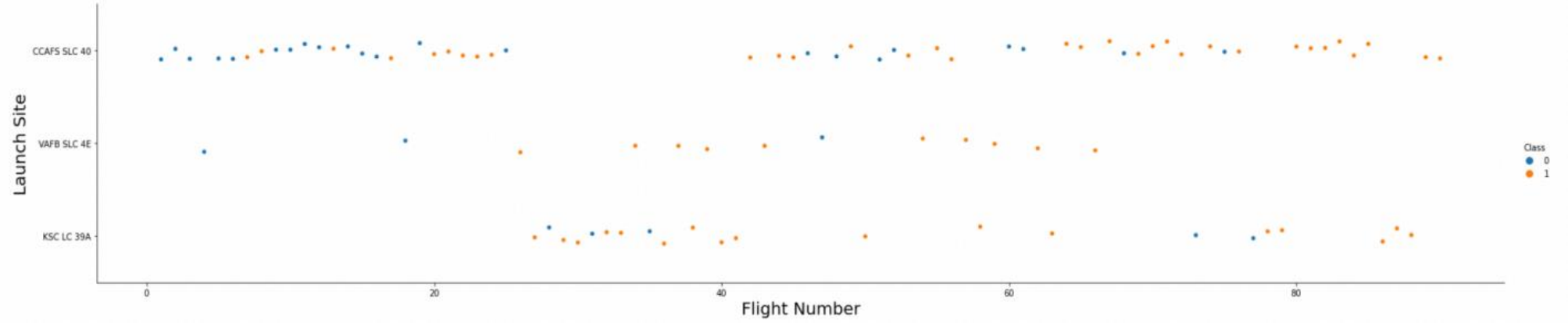


Results

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


Section 2
Insights drawn from EDA

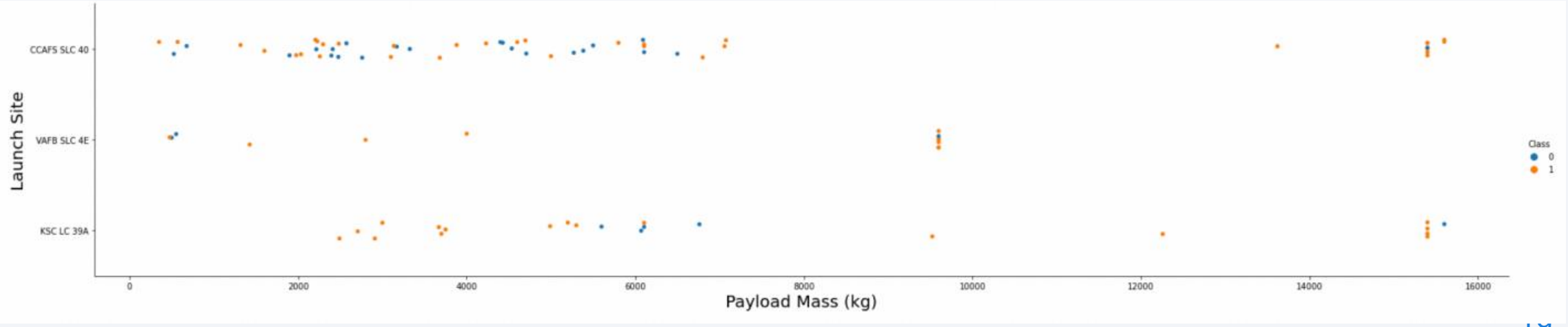
Flight Number vs. Launch Site



- Explanation:
  - 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.

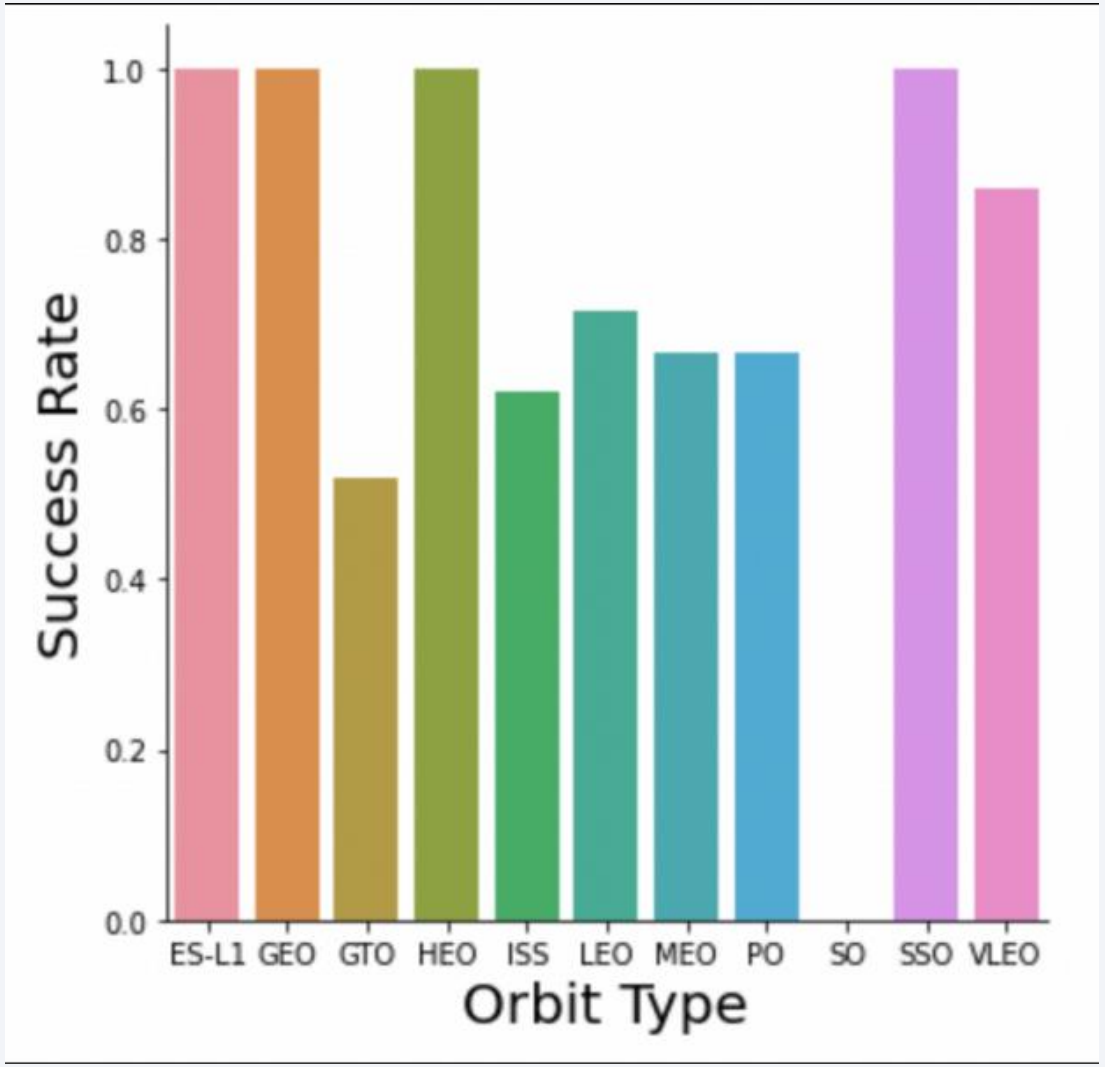
Payload vs. Launch Site

- Explanation: •
  - For every launch site the higher the payload mass, the higher the success rate.
  - Most of the launches with payload mass over 7000 kg were successful.
  - KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.



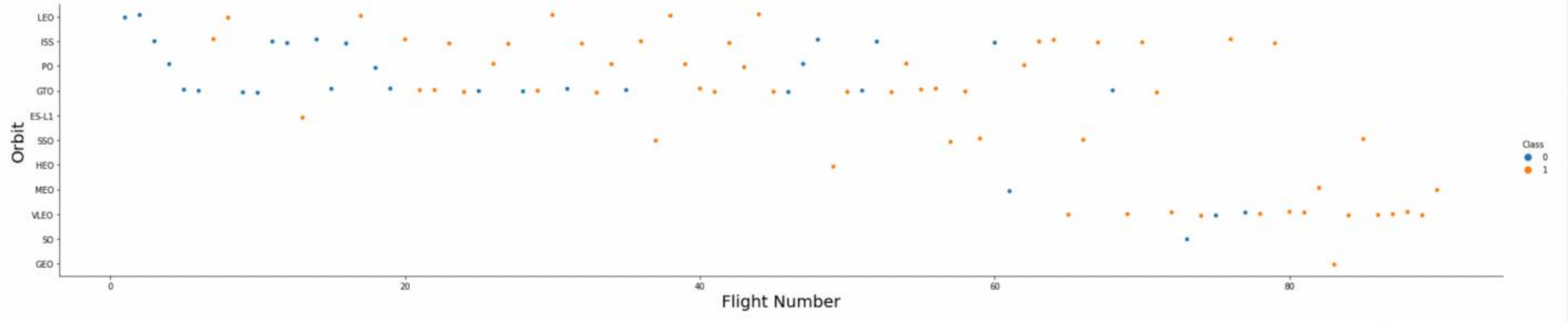
Success Rate vs. Orbit Type

- Explanation:
  - 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



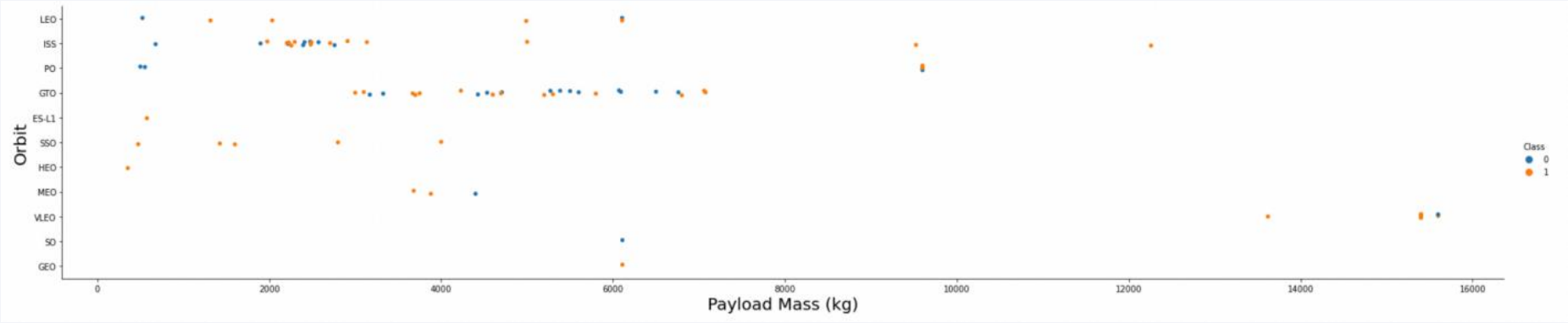
Flight Number vs. Orbit Type

- Explanation:
  - 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.



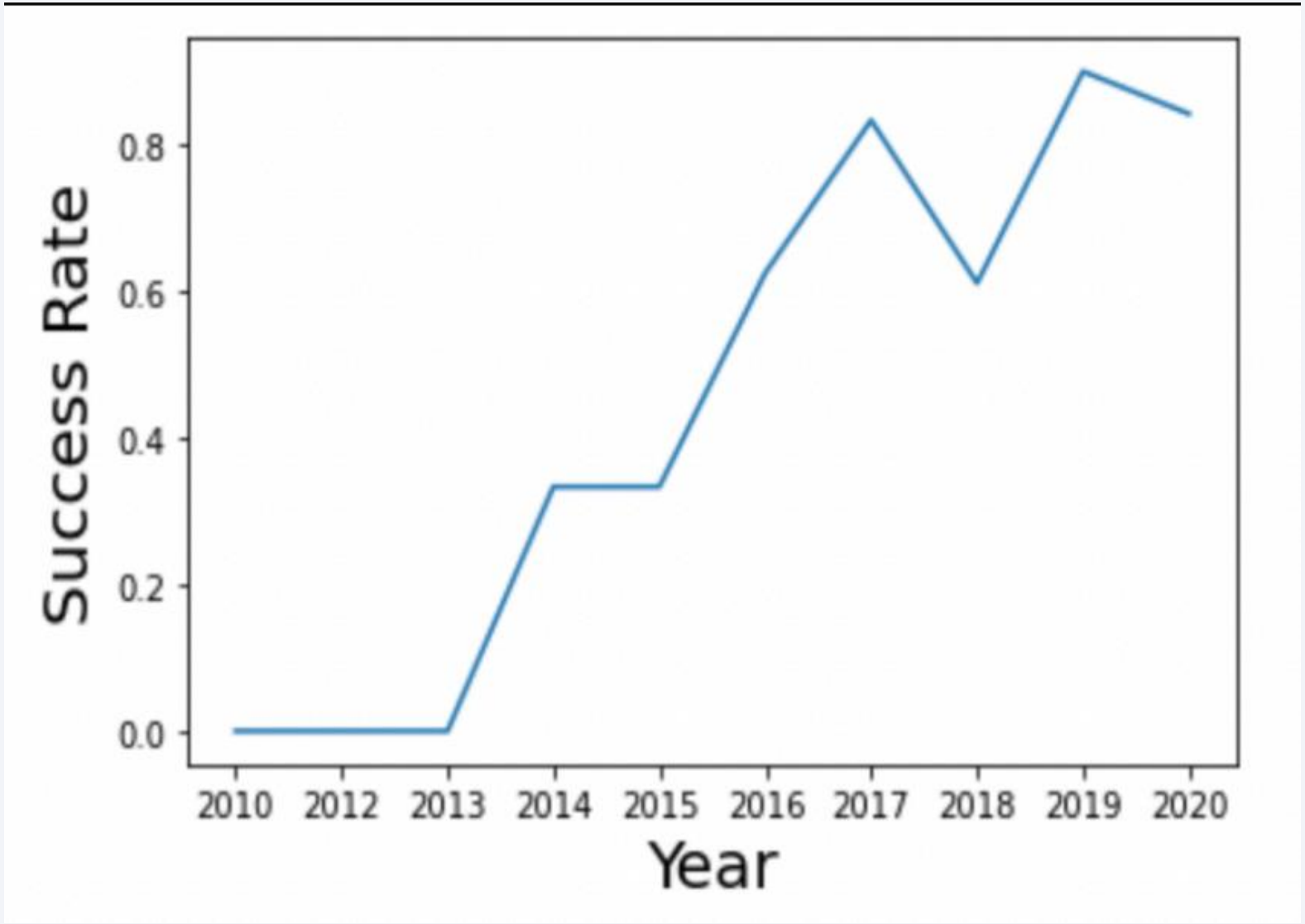
Payload Mass vs. Orbit Type

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



Launch Success Yearly Trend

- Explanation:
  - The success rate since 2013 kept increasing till 2020.



Launch Site Names Begin with 'CCA'

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

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

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

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[7]:
```

average_payload_mass
2534

First Successful Ground Landing Date

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

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[8]:
```

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Explanation:
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
        Done.
```

Out[9]:

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

Total Number of Successful and Failure Mission Outcomes

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

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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

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

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
         where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Explanation:
  - 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.

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
             where date between '2010-06-04' and '2017-03-20'
             group by landing__outcome
             order by count_outcomes desc;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

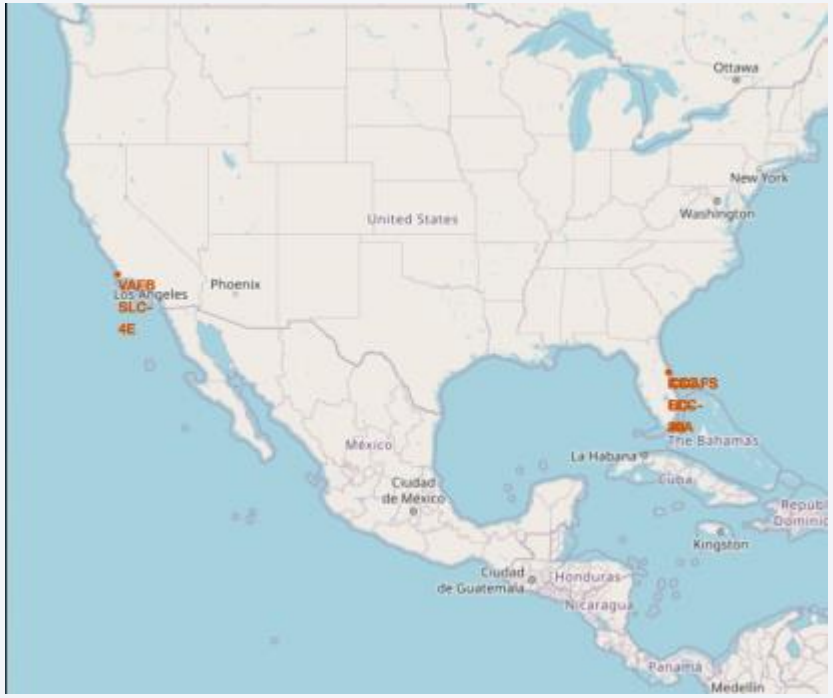
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Map of Launch Sites

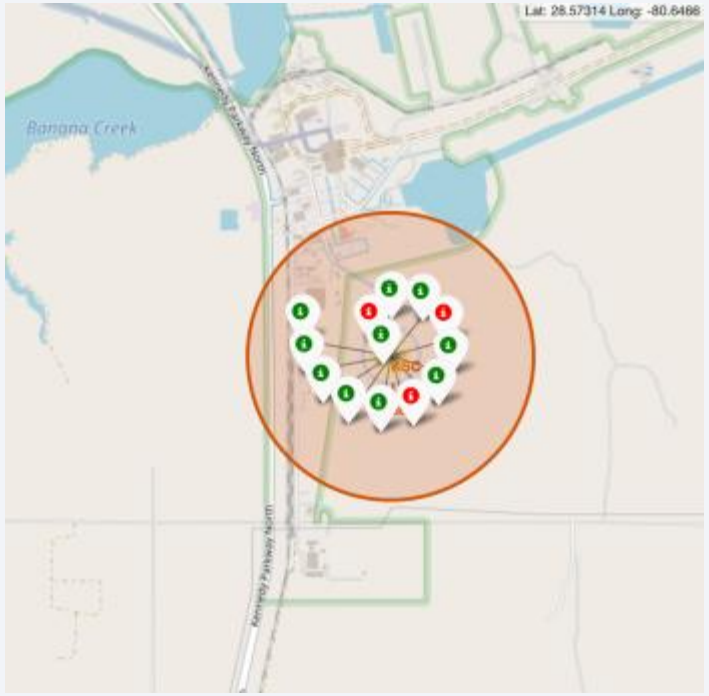
- Proximity to the Equator: Most launch sites are located close to the Equator. This is because the Earth’s surface moves fastest at the equator—about 1670 km/h—due to the planet's rotation. When a rocket is launched from the equator, it already carries this rotational speed thanks to inertia, which helps the spacecraft maintain a sufficient velocity to enter and stay in orbit more efficiently.
- Proximity to the Coast: Launch sites are also situated near coastlines. This strategic placement allows rockets to be launched over the ocean, minimizing risk to human populations in case of debris or explosions during launch. It provides a safer environment for both people and infrastructure.



Color-Labeled Launch Records

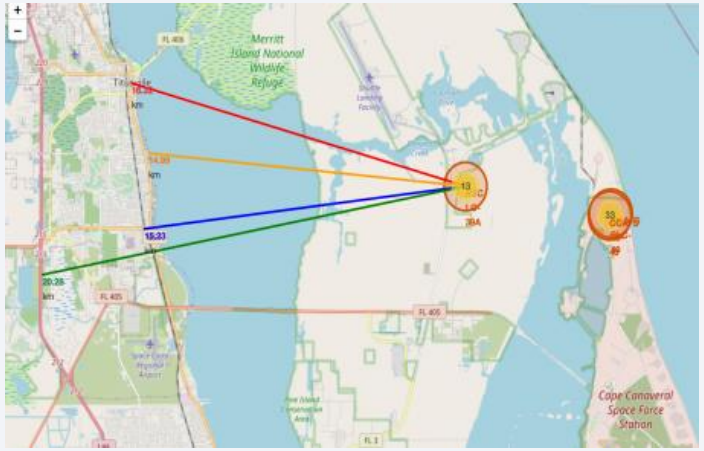
- Explanation:The color-coded markers make it easy to visually identify which launch sites have higher success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch

  From the map, it's clear that Launch Site KSC LC-39A has a very high success rate, indicated by the dominance of green markers in that area.



Distance from Launch Site KSC LC-39A to Nearby Features

A visual analysis of the launch site KSC LC-39A highlights its proximity to several key locations.

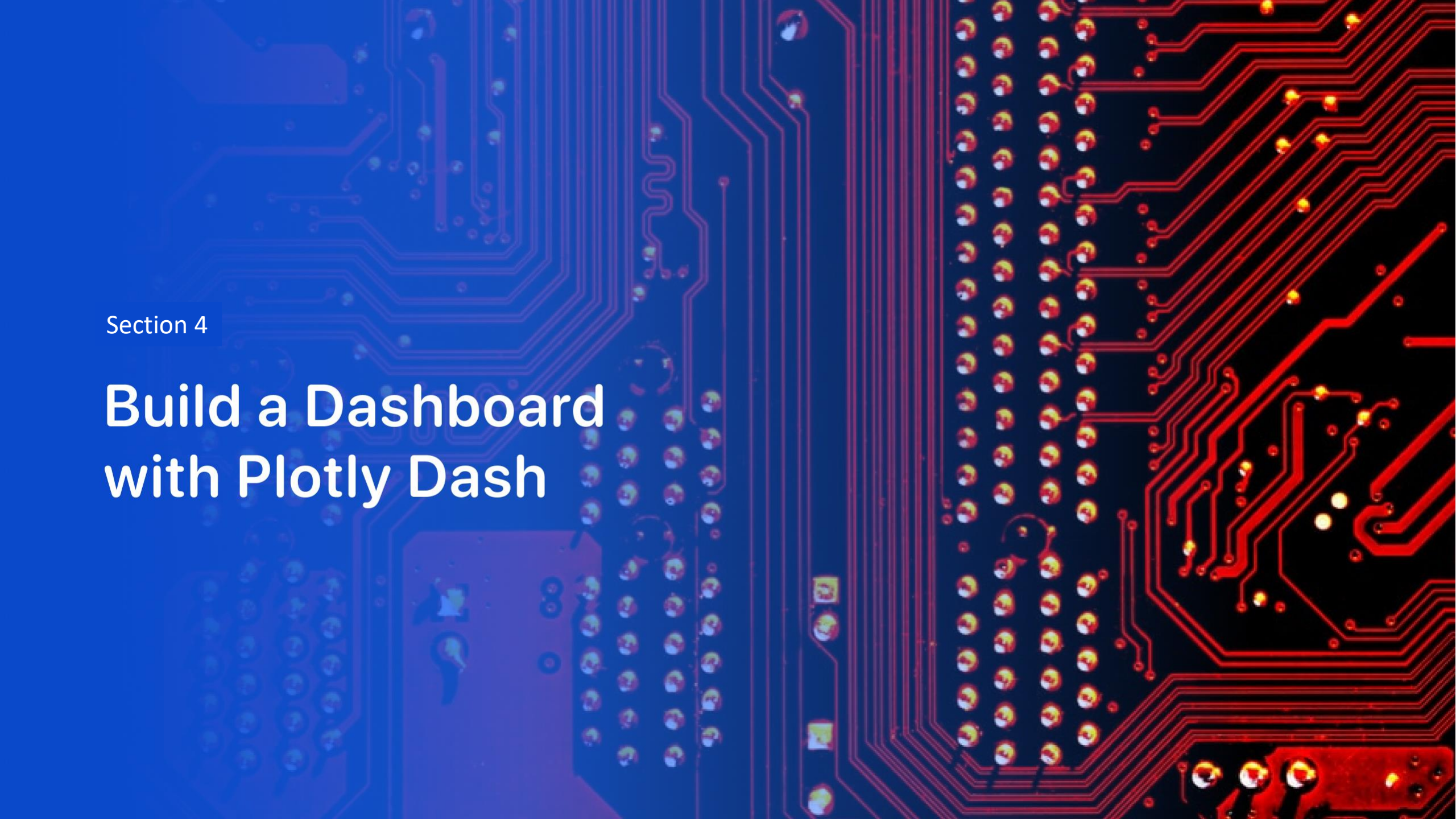


Explanation:

- The launch site is relatively close to:
- Railway – approximately 15.23 km
- Highway – approximately 20.28 km
- Coastline – approximately 14.99 km

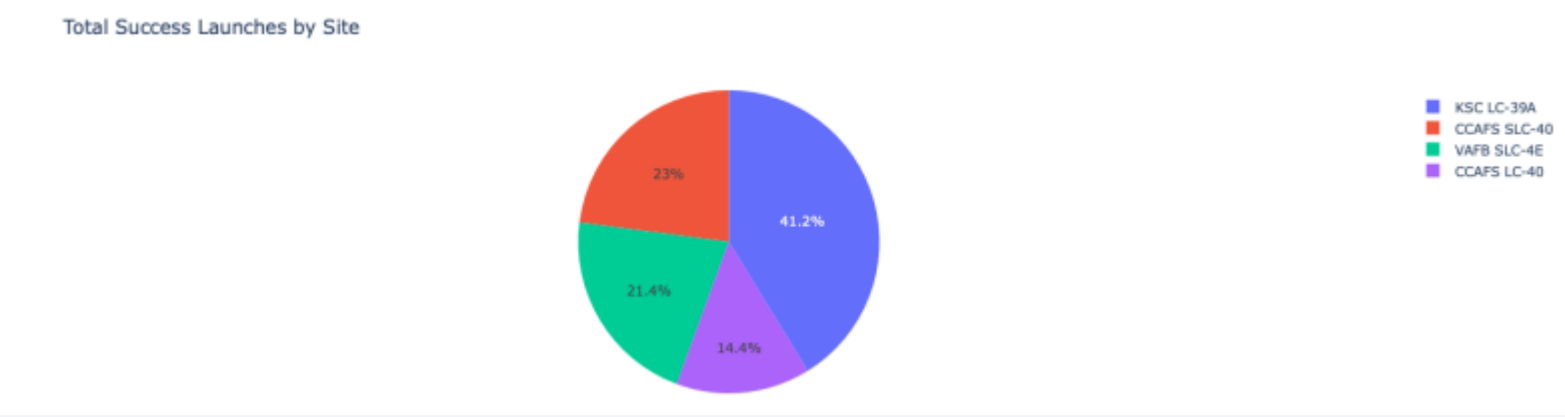
It is also fairly close to its nearest city, Titusville, which is about 16.32 km away.

Since rockets travel at extremely high speeds, even a failed launch could cover 15–20 km in just a few seconds, posing a potential risk to nearby populated areas.


Section 4
Build a Dashboard with Plotly Dash

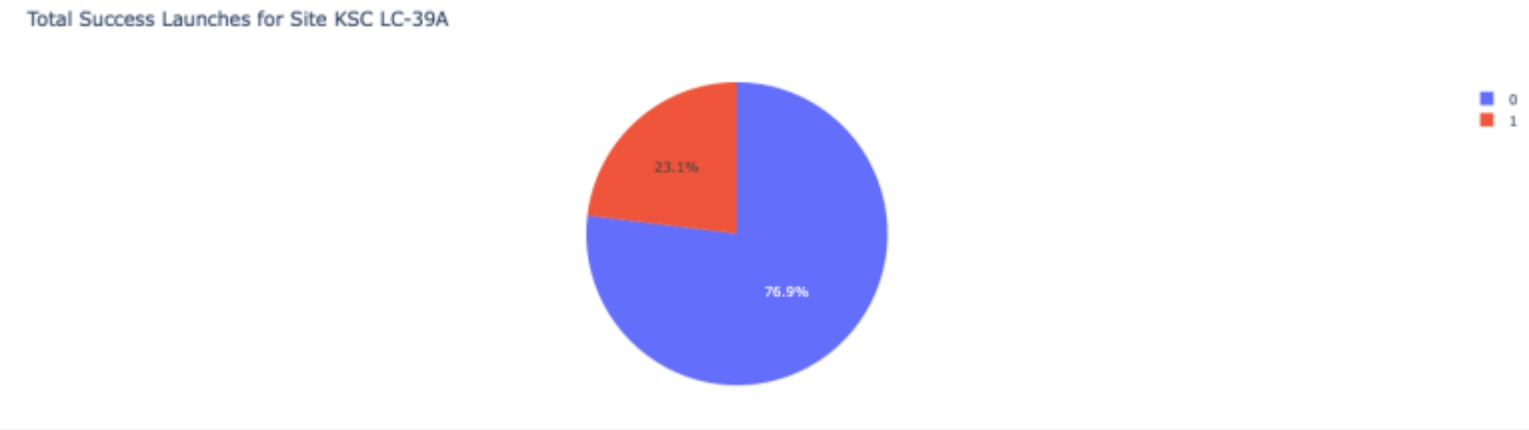
Launch success count

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



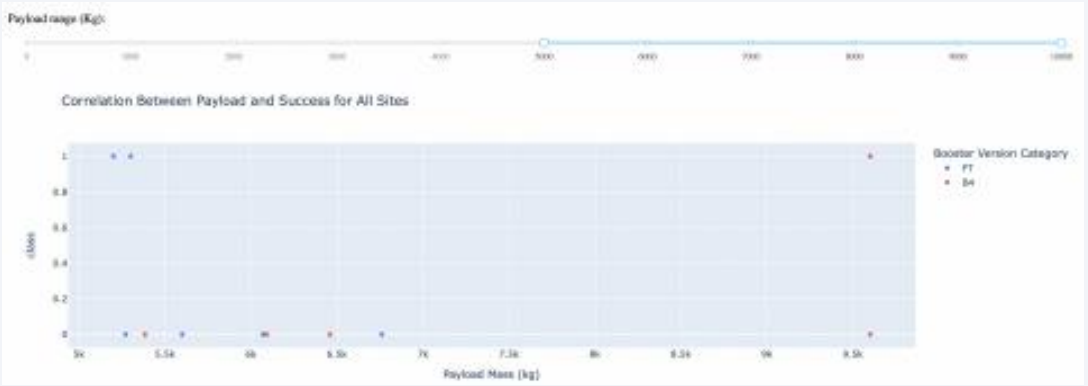
Launch site with highest launch success ratio

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



Payload vs. Launch Outcome scatter plot

- The charts show that payloads between 2000 and 5500 kg have the highest success rate



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

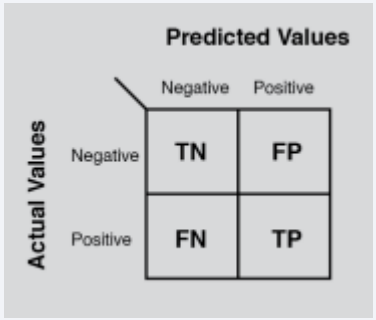
- Scores and Accuracy of the Entire Data Set

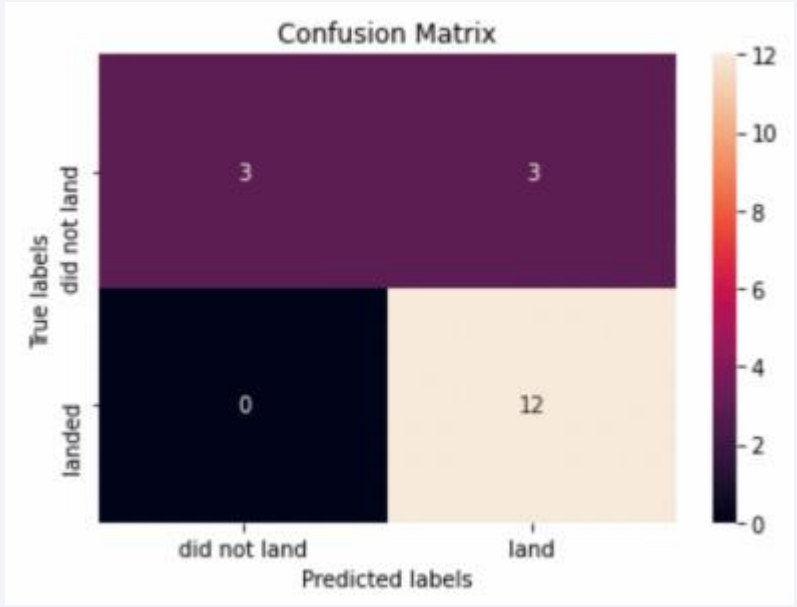
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.





Conclusions

- The Decision Tree Model proved to be the most effective algorithm for this dataset.
- Launches with lower payload mass tend to have higher success rates compared to those with heavier payloads.
- The majority of launch sites are located near the Equator, and all sites are positioned very close to coastlines, which supports both orbital efficiency and safety.
- The success rate of launches has steadily increased over the years, indicating improvements in technology and operations.
- Among all launch sites, KSC LC-39A demonstrated the highest launch success rate.
- Certain orbits—ES-L1, GEO, HEO, and SSO—achieved a 100% success rate, suggesting these missions are particularly well-executed or optimized.

Thank you!