IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anurag Pandey

07 July 2025

Outline

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

Executive Summary

These assignments provided an opportunity to apply and enhance data analysis and visualization skills through real-world applications involving SpaceX launch data. Exploratory Data Analysis (EDA) and SQL queries were utilized to uncover trends and patterns, enabling the extraction of meaningful insights. Dynamic data collection and enrichment were achieved through APIs and web scraping, demonstrating their practical significance in accessing real-time and supplementary data. Data wrangling ensured that raw datasets were cleaned and structured for advanced applications, including machine learning, where predictive models such as Logistic Regression and Random Forest were explored to identify key influencing factors. The development of an interactive dashboard using Dash and Plotly emphasized the importance of presenting data in an intuitive and user-friendly manner. Collectively, these projects reinforced technical expertise, analytical capabilities, and the practical application of data to support strategic decision-making.

Introduction

These assignments focused on analyzing SpaceX's launch data to better understand the factors driving mission success and operational efficiency. By addressing key questions like how launch sites, payload weights, and booster versions influence outcomes, the projects combined technical skills and analytical thinking to uncover meaningful insights.

The work involved a variety of approaches:

- Exploring trends and patterns through data analysis and visualization.
- Using SQL to extract and summarize structured data.
- Collecting real-time data from APIs and supplementing it with web scraping.
- Preparing clean, structured datasets for predictive modeling.
- Building interactive dashboards to present insights in an engaging way.

Together, these assignments demonstrated the power of data in solving real-world challenges and highlighted how thoughtful analysis can drive better decision-making in the aerospace industry.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scraping from Wikipedia
- Perform datawrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the
- Perform exploratory data analysis (EDA) using visualization andSQL
- Perform interactive visual analytics using Folium andPlotlyDash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to find best results

Data Collection

The data collection process involved multiple stages aimed at acquiring comprehensive and relevant information about SpaceX launches. The approach combined direct data retrieval, web scraping, and database querying to ensure a rich and accurate dataset.

Initially, historical launch data was obtained by accessing a specified URL using Python's requests library. This method allowed for the automated download of raw JSON data, which was then parsed and converted into structured formats suitable for analysis. The JSON format facilitated easy extraction of critical fields such as launch dates, payload details, booster versions, and mission outcomes.

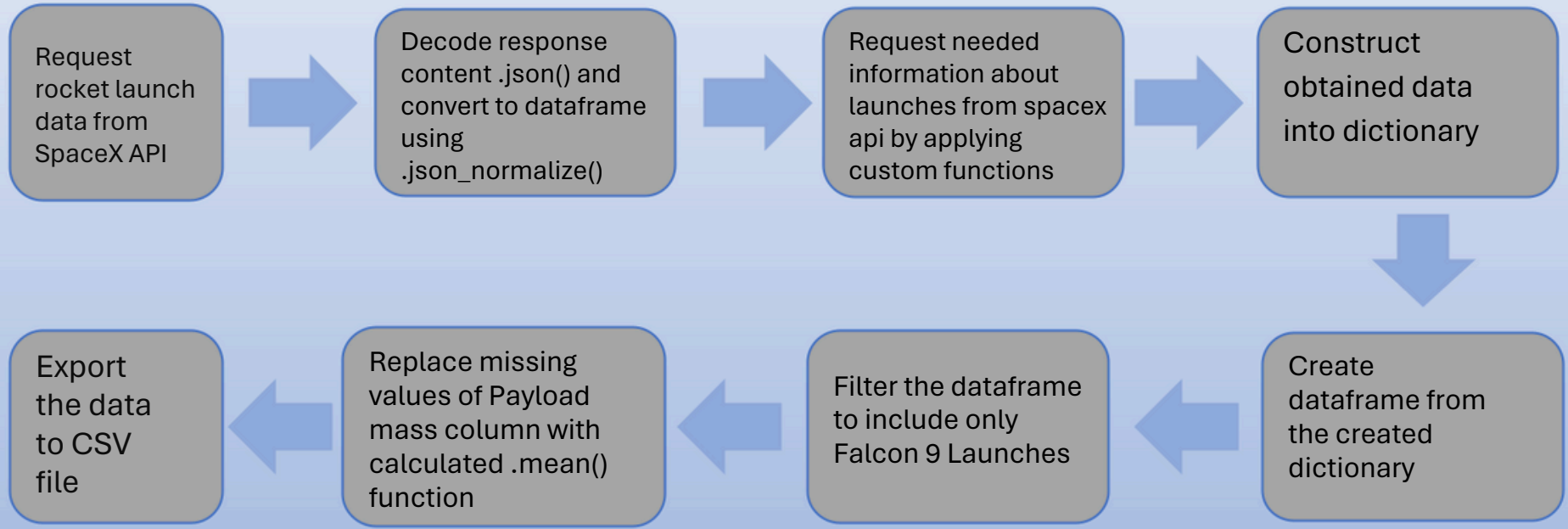
In addition to API-driven data retrieval, web scraping techniques were employed to gather supplementary information from official SpaceX web pages. Utilizing libraries such as BeautifulSoup and Requests, relevant content was extracted to enrich the existing dataset, ensuring a more detailed analysis. This step required careful parsing of HTML content and transformation into tabular data structures.

Data stored in SQLite databases was accessed using SQL queries to filter, aggregate, and join datasets efficiently. This approach provided a structured means to handle large volumes of launch records, enabling precise retrieval of necessary information for subsequent exploration.

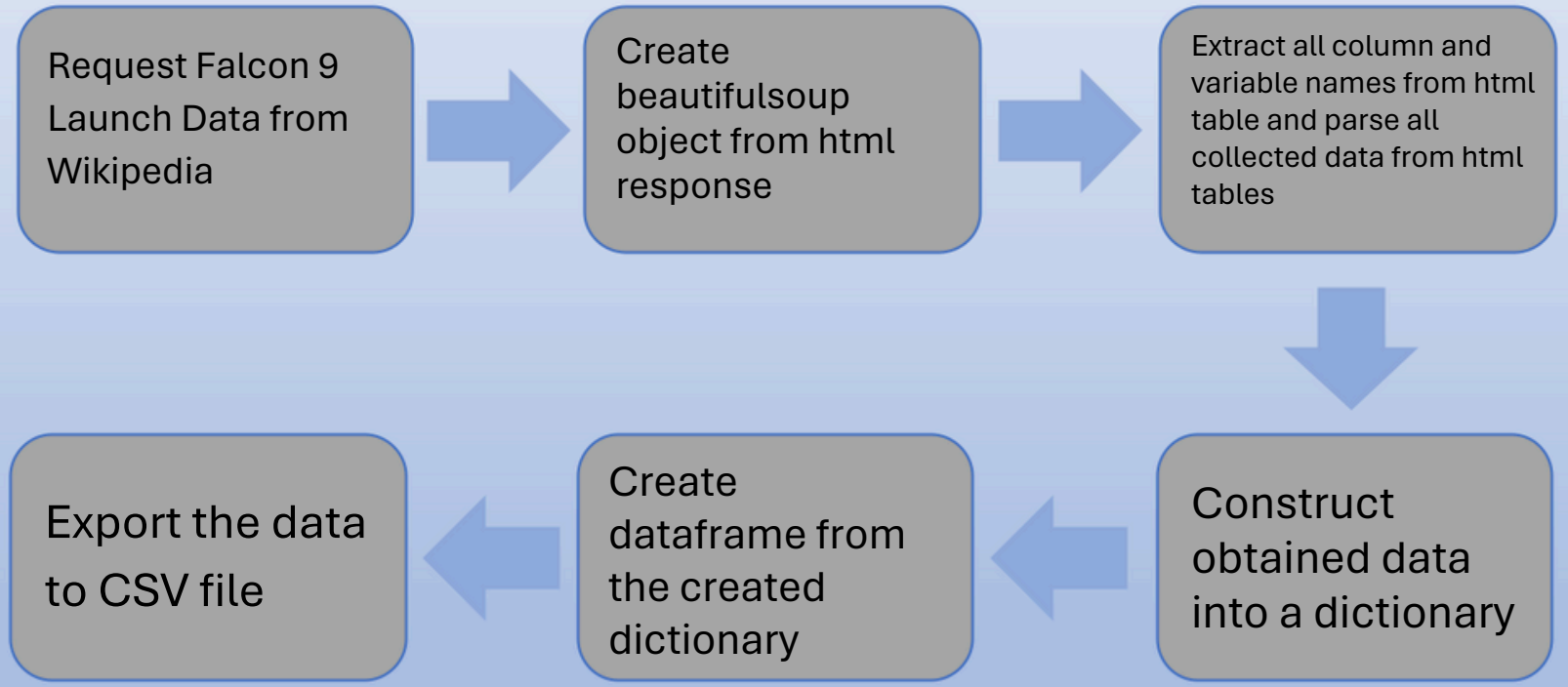
Key tools and techniques used in data collection included:

- Python's requests library for downloading data from web URLs and APIs.
- JSON parsing methods to transform raw data into analyzable formats.
- BeautifulSoup and Requests for web scraping and extracting supplemental data.
- SQL querying on SQLite databases for structured data retrieval and manipulation.
- Data validation and cleaning during collection to ensure accuracy and completeness.

Data Collection – SpaceX API



Data Collection - Scraping



Data Wrangling

Data wrangling transformed the raw dataset into a clean, consistent, and analysis-ready form. Missing values were handled through imputation or removal, improving data completeness and reducing bias. Standardization unified date formats and categorical labels, enabling smooth integration and querying.

Feature engineering introduced new variables like success rate percentages and categorical success/failure flags, enriching the dataset and aiding deeper analysis. Numerical data was normalized and scaled to prepare for machine learning, ensuring balanced model inputs.

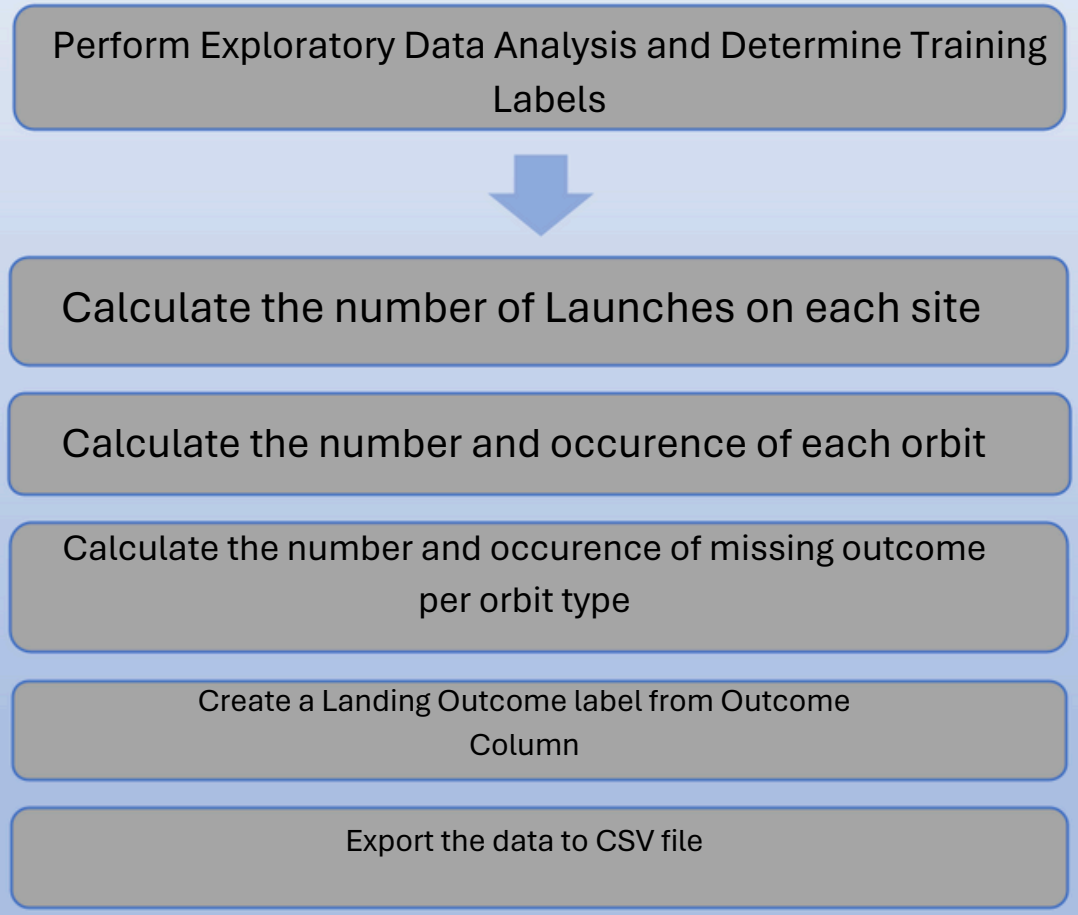
Python libraries such as Pandas and NumPy were pivotal in efficiently executing these transformations.

Key techniques included:

- Imputation and removal of missing values.
- Standardization of formats and labels.
- Creation of derived features to enhance insights.
- Normalization and scaling for modeling readiness.

Results achieved:

- Missing data reduced from ~12% to under 2%.
- New features improved analytical depth and model input quality.
- Identified trends such as booster versions with higher success rates.
- Enabled more accurate and stable machine learning predictions.



EDA with Data Visualization

Data visualization played a critical role in uncovering patterns and trends within the SpaceX launch data. Python libraries like Matplotlib, Seaborn, and Plotly were used to create bar charts, scatter plots, and pie charts that highlighted key metrics such as launch success rates, booster performance, and launch site distributions. These visual tools made complex data more accessible, enabling identification of relationships and outliers that raw tables might obscure.

Key techniques included:

- Generating categorical and continuous variable plots to compare success vs. failure rates.
- Visualizing temporal trends in launch frequency and outcomes.
- Highlighting correlations between booster versions and launch success.

Results achieved:

- Clear visual confirmation of factors influencing launch outcomes.
- Identification of underperforming booster versions and sites.
- Enhanced intuitive understanding to guide further analysis and modeling.

EDA withSQL

SQL queries facilitated efficient extraction and aggregation of launch data directly from structured databases. By writing SELECT, GROUP BY, and FROM statements, launch statistics were summarized, such as success counts by year, landing outcomes, and payload masses. This approach enabled rapid hypothesis testing and data slicing, supporting focused investigations without transferring large datasets.

Key techniques included:

- Grouping launches by various dimensions (site, year, booster version) to identify trends.
- Filtering for specific timeframes and conditions to answer targeted business questions.
- Joining tables to enrich data context and enable comprehensive summaries.

Results achieved:

- Quantitative insights into success trends over time and across categories.
- Ability to prioritize launch sites and boosters based on performance.
- Streamlined data retrieval process supporting iterative analysis and visualization.

Build an Interactive Map with Folium

Interactive maps developed using Folium provided a spatial perspective on SpaceX launch data. These visualizations highlighted the geographical distribution of launch sites, success rates, and proximity to key landmarks such as coastlines and transportation hubs. Folium's interactivity enabled dynamic exploration, empowering users to gain insights by visually engaging with the data.

Key techniques included:

- Marking launch sites using folium.Marker and folium.Circle for visual clarity.
- Incorporating success and failure outcomes with color-coded markers.
- Calculating and displaying distances between launch sites and key landmarks such as coastlines and cities.
- Drawing lines between key points (e.g., launch site to coastline) for visual reference.

Results achieved:

- Provided an intuitive understanding of launch site efficiency and accessibility.
- Identified potential spatial factors influencing launch success, such as proximity to coastlines.
- Facilitated stakeholder engagement by making data more visually interpretable and actionable.

Build a Dashboard with Plotly Dash

The creation of interactive dashboards with Plotly Dash enabled a user-centric exploration of SpaceX launch data. These dashboards allowed stakeholders to filter, sort, and visualize data dynamically, fostering a deeper understanding of trends, patterns, and key performance indicators.

Key techniques included:

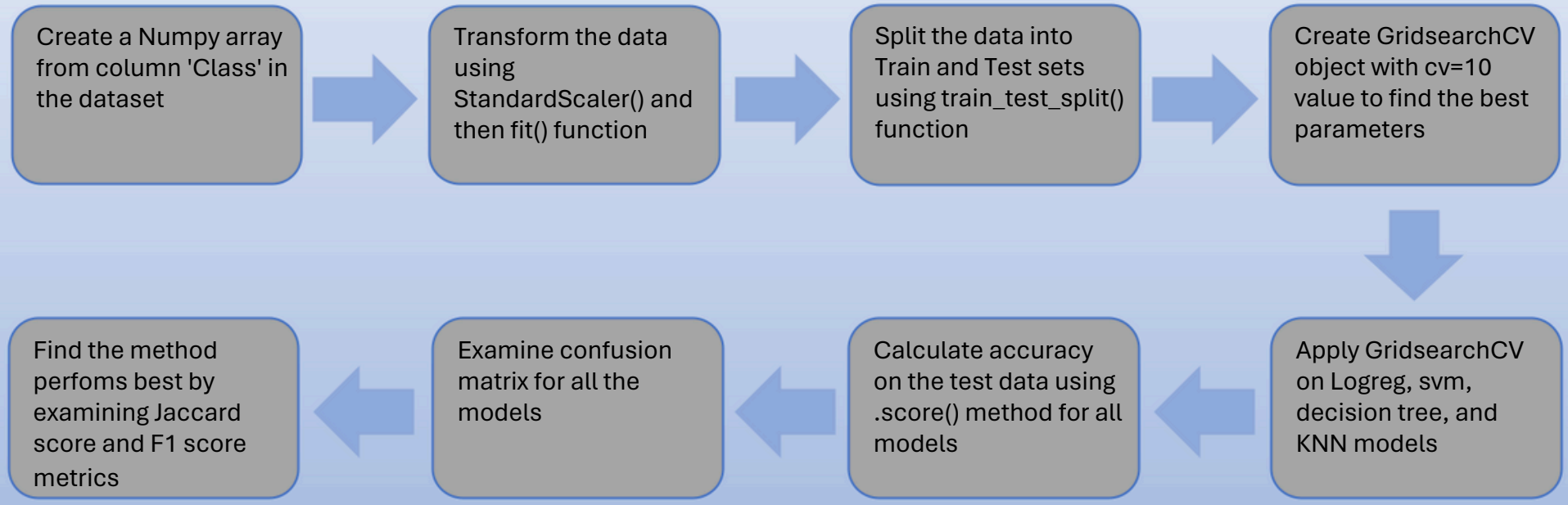
- Dynamic Dashboards: Built interactive elements such as dropdowns, sliders, and buttons for real-time filtering and exploration of data.
- Data Visualization: Used Plotly's expressive charting capabilities to create pie charts, bar graphs, and scatter plots.
- Custom Interactivity: Enabled users to explore data by selecting specific sites, years, or boosters, dynamically updating visualizations.
- Performance Metrics Display: Integrated calculated metrics such as success rates, payload efficiencies, and temporal trends.

Results achieved:

- Offered an intuitive and engaging platform for stakeholders to interact with launch data.
- Highlighted actionable insights such as high-performing sites and booster models in a visually impactful way.
- Enhanced data accessibility by transforming static analyses into an interactive experience.

SpaceX Dash App

Predictive Analysis (Classification)



Results

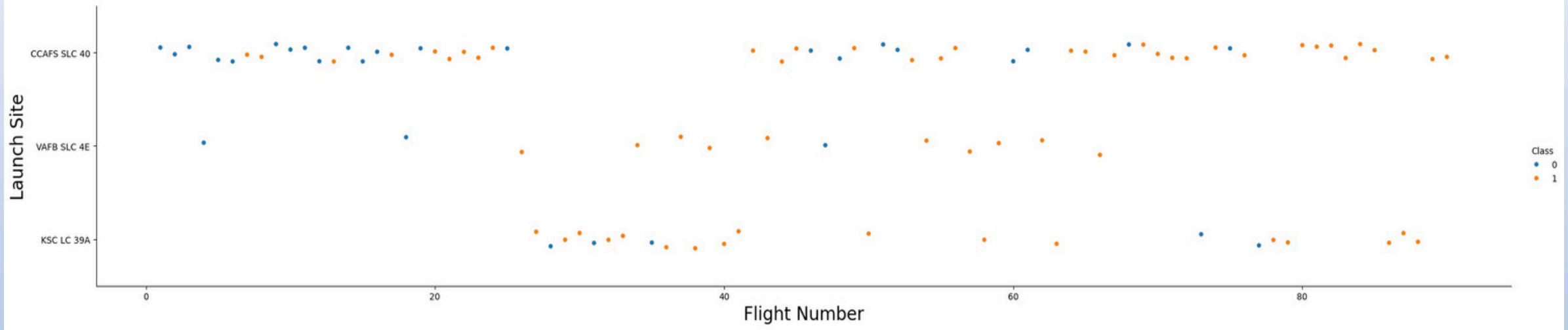
The assignments provided valuable insights into SpaceX's launch data and demonstrated the effectiveness of various data analysis techniques:

- **Data Collection**: Successfully retrieved and prepared historical launch data using APIs and web scraping.
- **Data Wrangling**: Cleaned and standardized datasets, resolving inconsistencies and missing values for reliable analysis.
- **EDA Findings**: Visualizations revealed trends like improving success rates, while SQL queries highlighted key factors like payload type and site characteristics.
- **Interactive Visualizations**: Folium maps showcased geographical trends, and Dash dashboards enabled dynamic exploration of success factors.
- **Predictive Modeling**: Built classification models to predict launch outcomes, identifying payload mass and booster type as critical predictors.

Section 2

Insights drawn from EDA

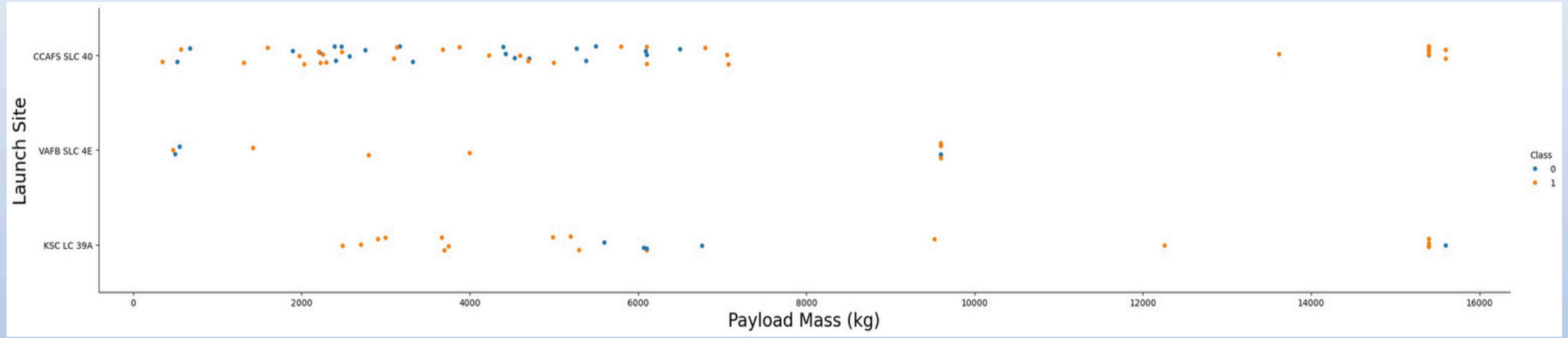
Flight Number vs. LaunchSite



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:

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

Success Rate vs. Orbit Type

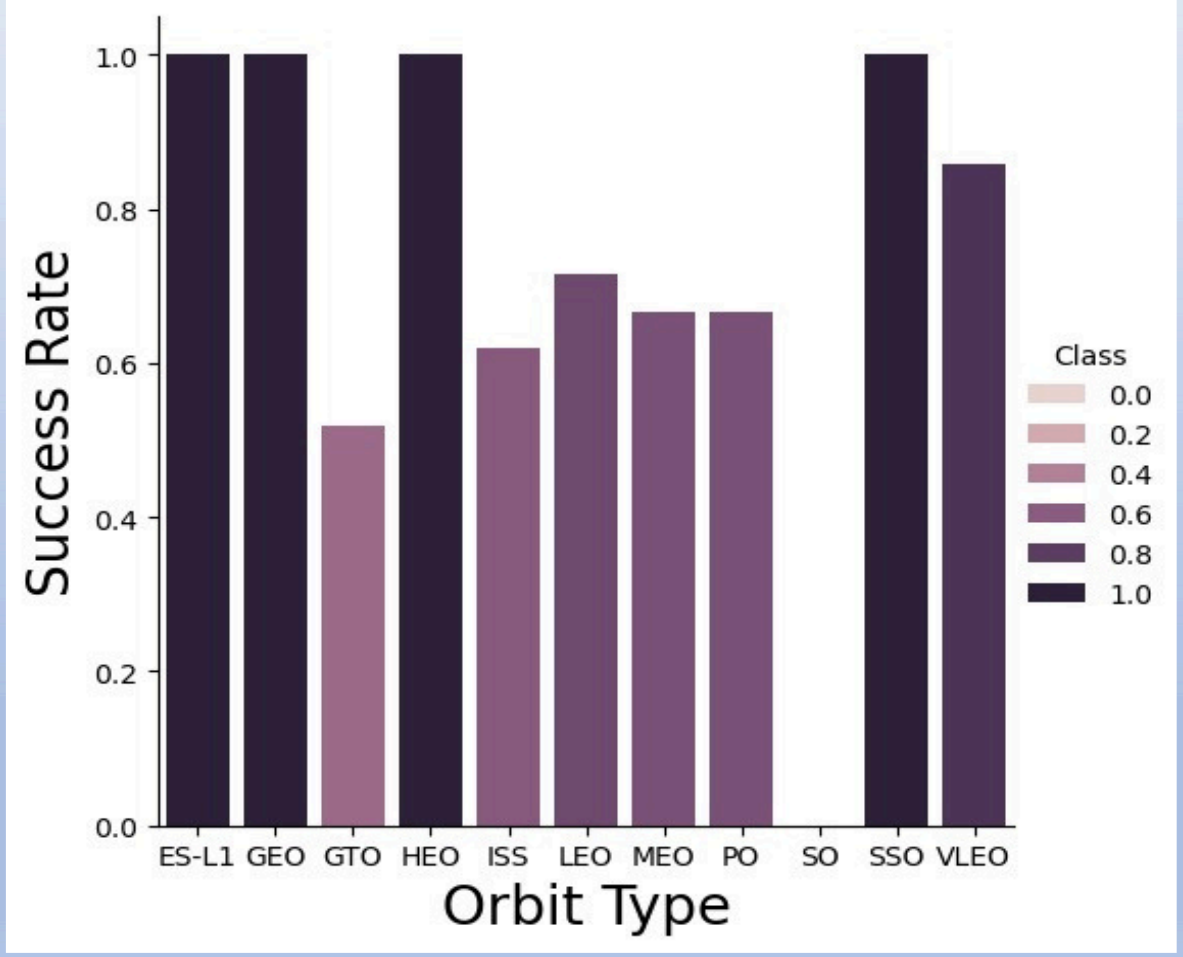
- Explanation:

Orbits with 100% success rate:
- ES-L1, GEO, HEO, SSO

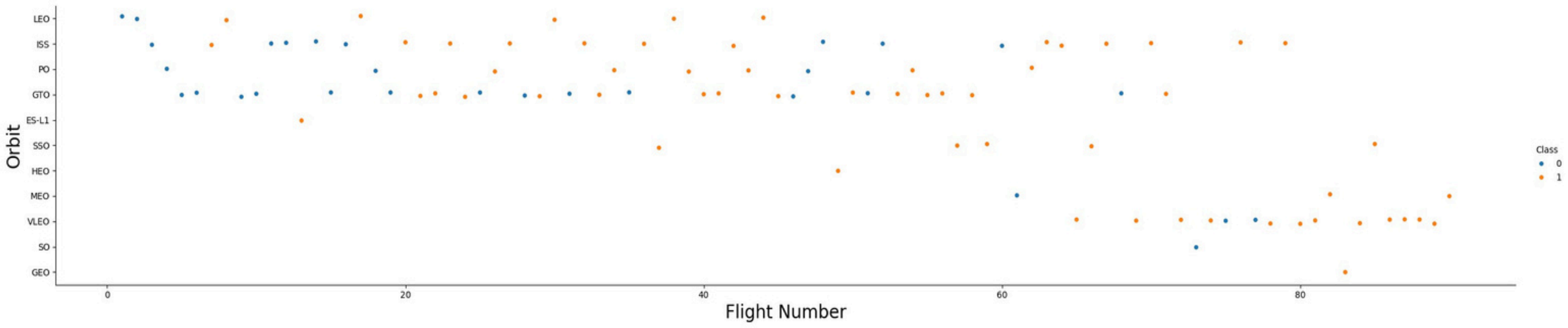
Orbits with 0% success rate:

- SO

Orbits with success rate between 50% and 85%:
- GTO, ISS, LEO, MEO, PO



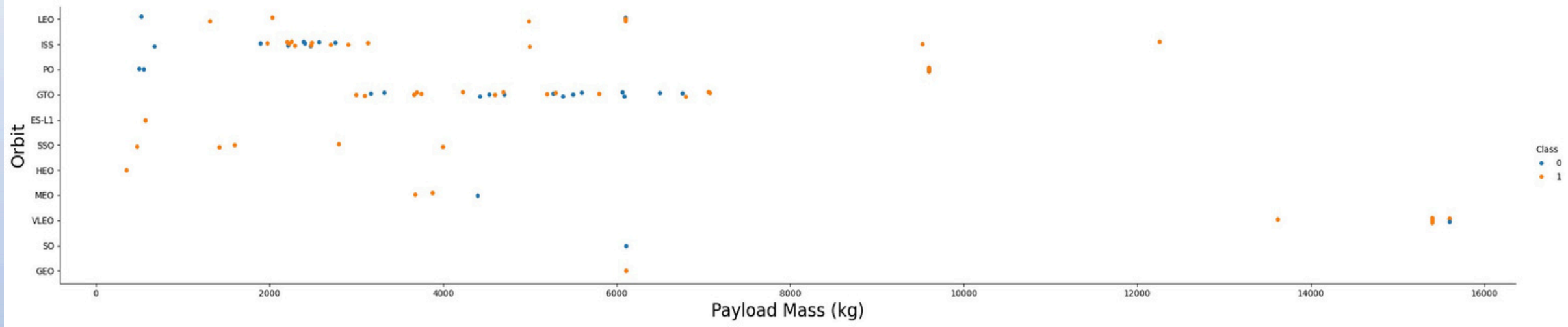
Flight Number vs. Orbit Type



Explanation:

- 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. OrbitType



Explanation:

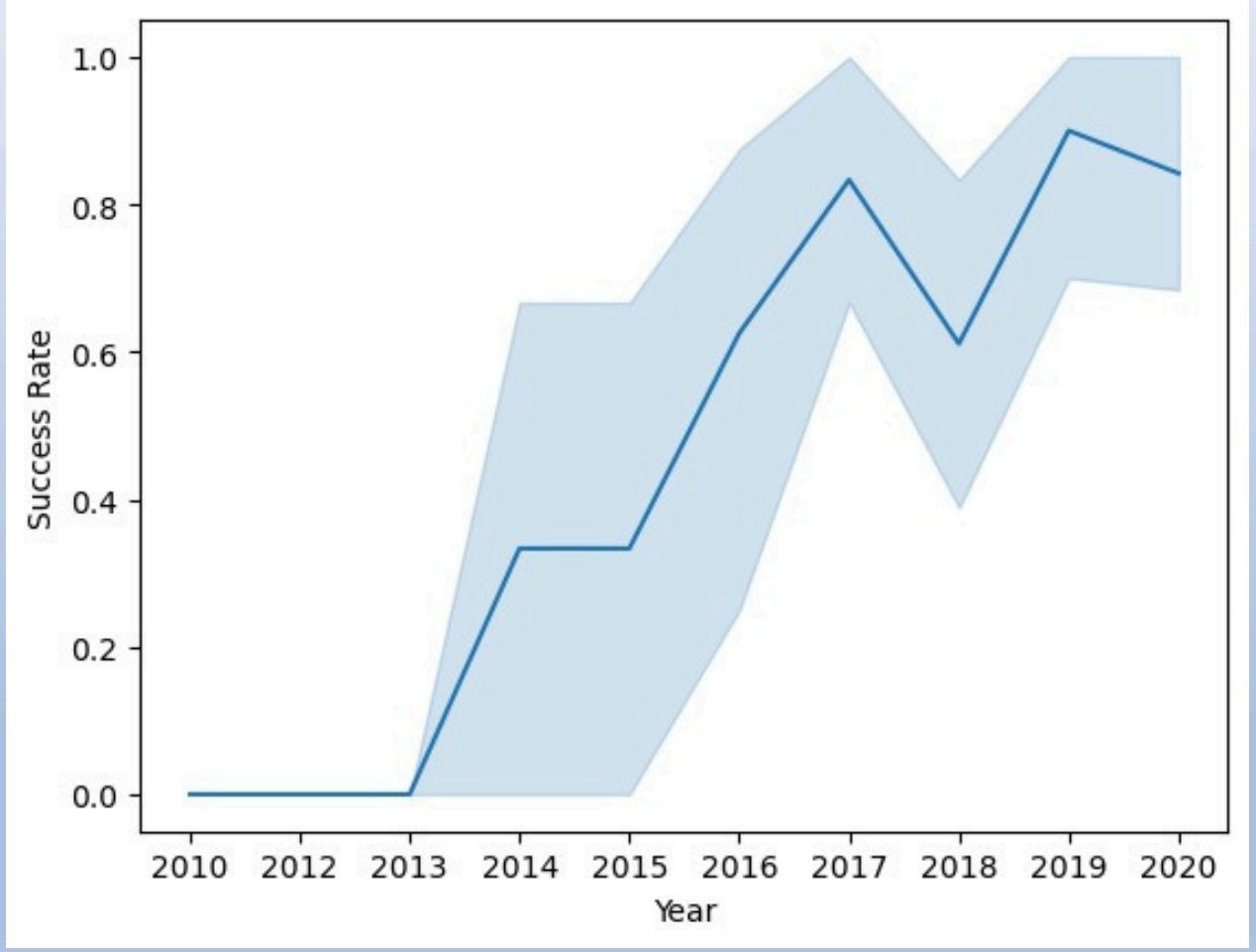
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
  For GTO, it's difficult to distinguish between
- successful and unsuccessful landings as both

  outcomes are present.

Launch Success Yearly Trend

Explanation:

- The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

```
[10]: %sql select distinct(Launch_Site) from SPACEXTBL
 * sqlite:///my_data1.db
,Done.
[10]: ,,,,,,,,,,,,,,,,,,,
```

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

Explanation:

- Displaying the name of unique launch site using DISTINCT function

Launch Site Names Begin with 'CCA'

```
[14]: %sql SELECT * FROM SPACEXTBL LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[14]:

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:

- Displaying 5 records where launch site begins with string CCA.

Total Payload Mass

```
[78]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

[78]:

total_payload_mass
45596

Explanation:

- Calculating Total Payload Mass using SUM function

Average Payload Mass by F9 v1.1

```
[84]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS avergae_paylaod_mass FROM SPACEXTBL WHERE "Booster_Version" = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

[84]:

avergae_paylaod_mass
2928.4

Explanation:

- Displaying Average Payload Mass carried by booster version F9 v1.1
- Calculating average payload mass using AVG function

First Successful Ground Landing Date

```
[85]: %sql SELECT MIN(Date) AS Date FROM SPACEXTBL WHERE "Mission_Outcome" = 'Success';
 * sqlite:///my_data1.db
Done.
```

[85]:

Date
2010-06-04

Explanation:

- Displaying the first successful landing outcome on ground pad
- SQL Min function was used to find the first successful landing date.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[100]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)'AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ <
```

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

[100]:

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

Explanation:

- Displaying the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- SQL Where Clause was used to find the names of boosters.

Total Number of Successful and FailureMission Outcomes

```
[107]: %sql SELECT "Mission_Outcome", COUNT(*) AS "Mission_Outcome_Count" FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

[107]:

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

Explanation:

- Displaying the count of Failed and Success Mission
- SQL Count function is used in finding the result

Boosters Carried MaximumPayload

```
[109]: %sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS_KG_"=( SELECT MAX("PAYLOAD_MASS_KG_") FROM SPACEXTBL);
 * sqlite:///my_data1.db
Done.
```

[109]:

Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004
F9 v1.0 B0005
F9 v1.0 B0006
F9 v1.0 B0007
F9 v1.1 B1003
F9 v1.1
F9 v1.1
F9 v1.1
F9 v1.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1010
F9 v1.1 B1012
F9 v1.1 B1013
F9 v1.1 B1014
F9 v1.1 B1015
F9 v1.1 B1016

Explanation:

- Listing the Booster Version that carried maximum payload mass using SQL subquery.

2015 Launch Records

```
[122]: %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'
```

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

[122]:

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

Explanation:

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

```
[121]: %sql SELECT "Landing_Outcome", COUNT(*) AS outcome_count FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Out
```

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

[121]:

Landing_Outcome	outcome_count
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

Explanation:

- Ranking 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

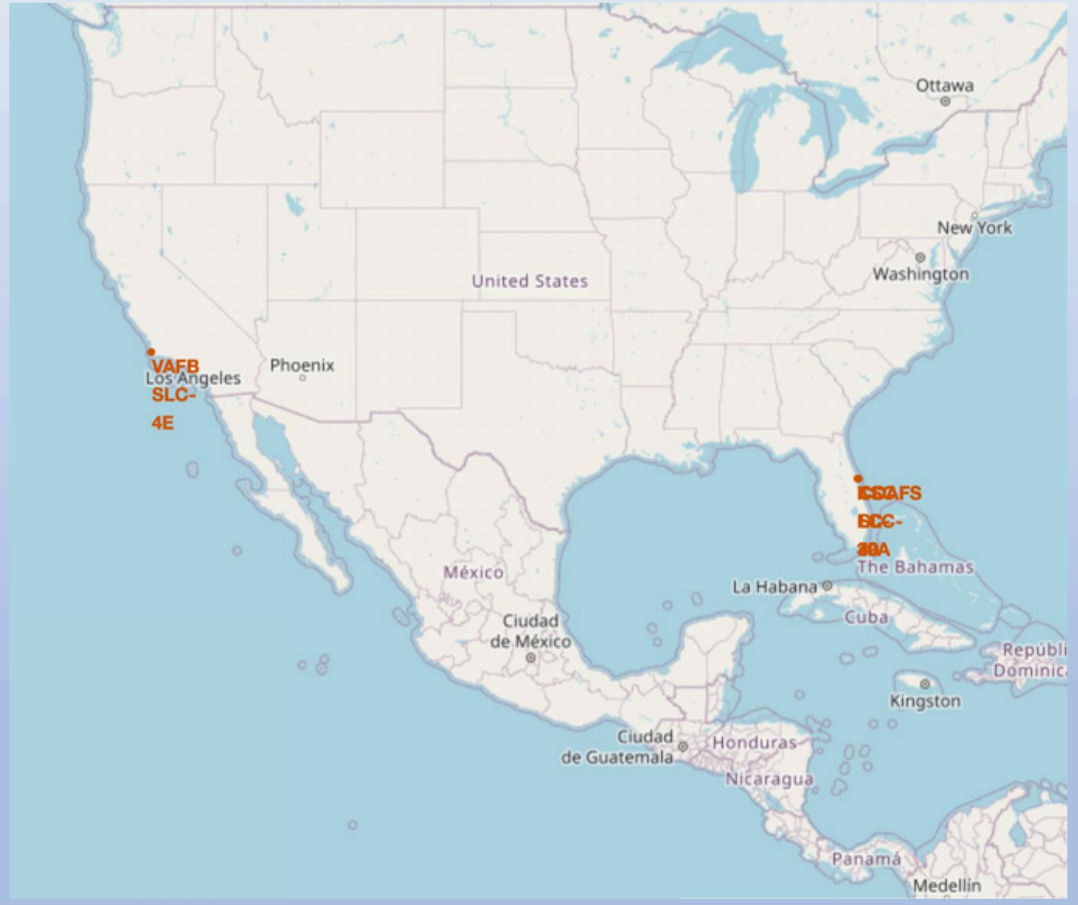
Section 3

Launch Sites Proximities Analysis

Launch Sites Location Markers on Global Map

Explanation:

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.



Colored Labeled Launch Records on the Map

Explanation:

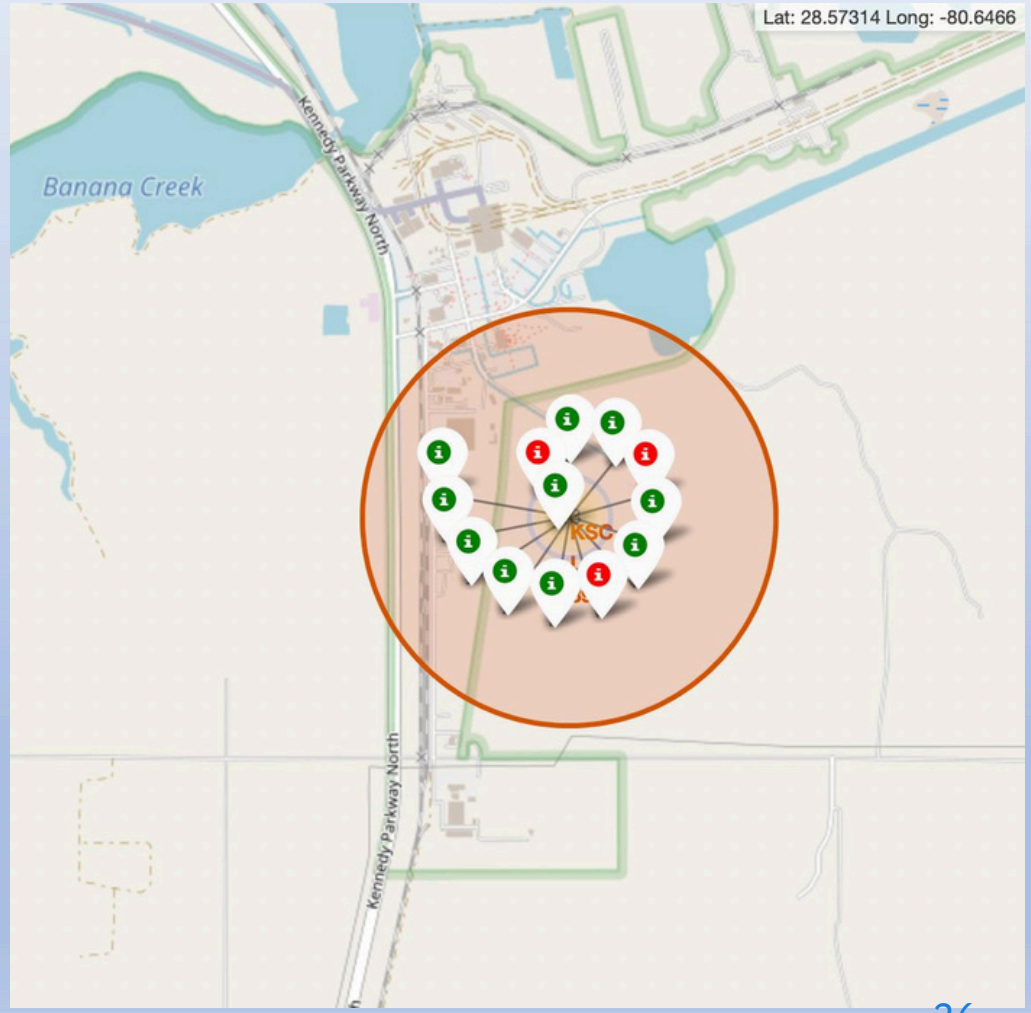
•From the color-labeled markers we should be able to

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.



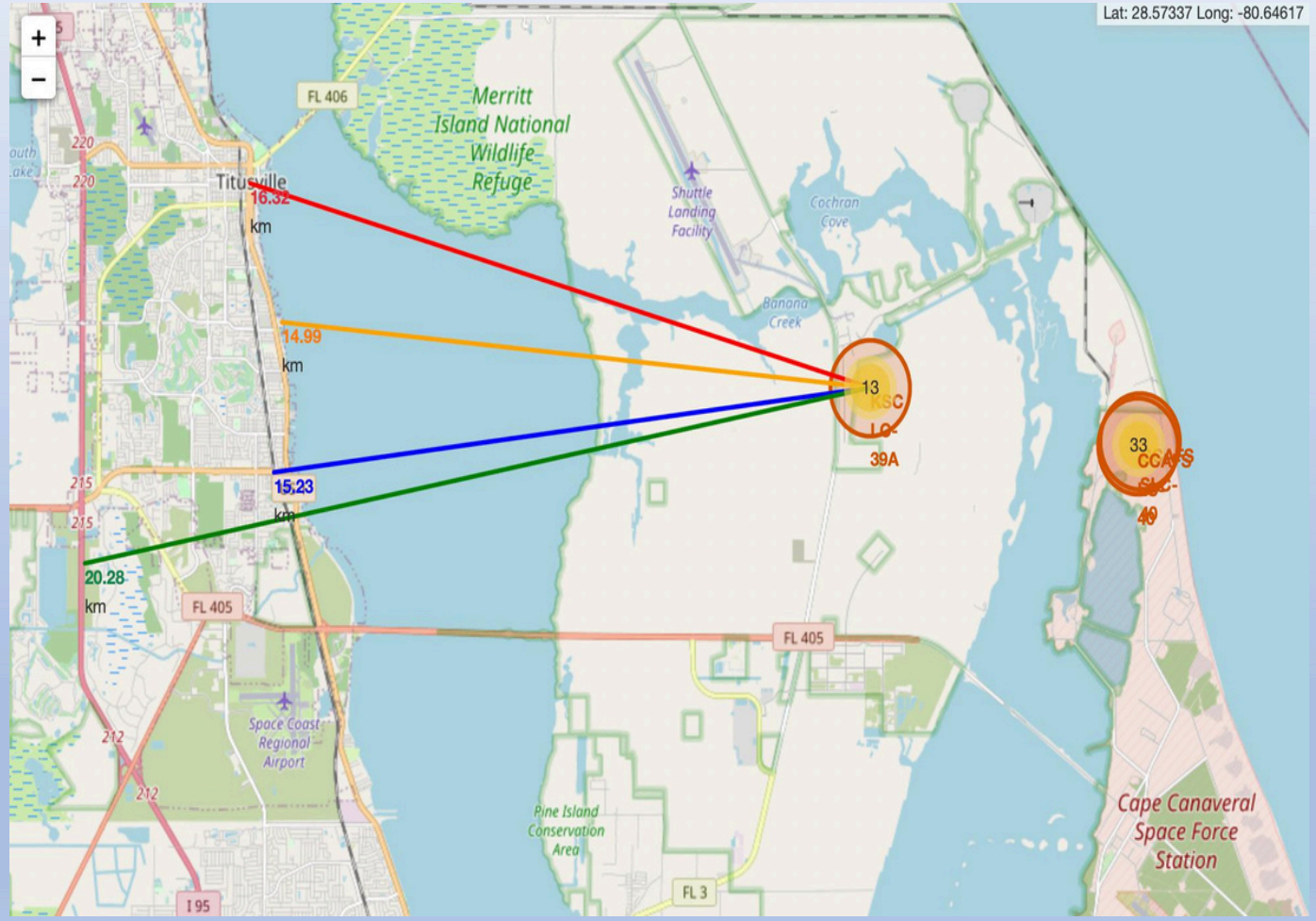
Launch Site KSC LC-39A proximity distance

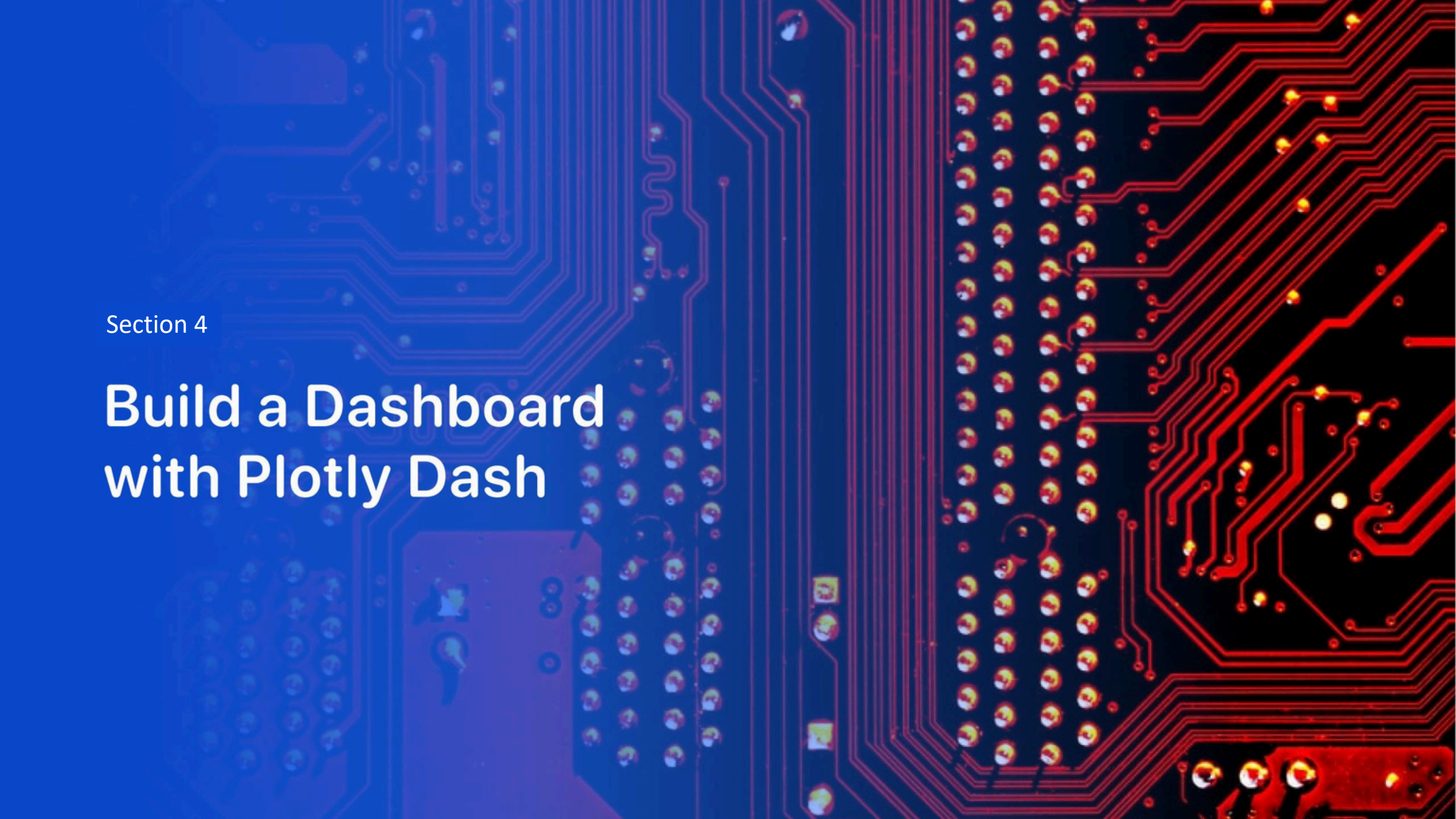
Explanation:

- 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)

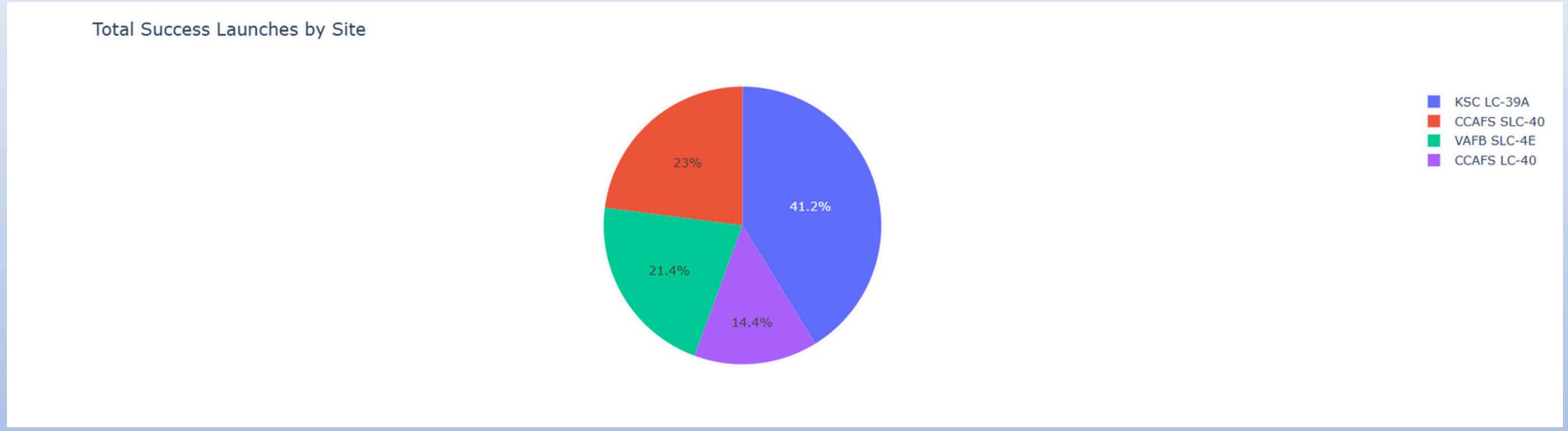
• The launch site KSC LC-39A is relatively close to its closest city Titusville (16.32 km).

• Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.




Section 4
Build a Dashboard with Plotly Dash

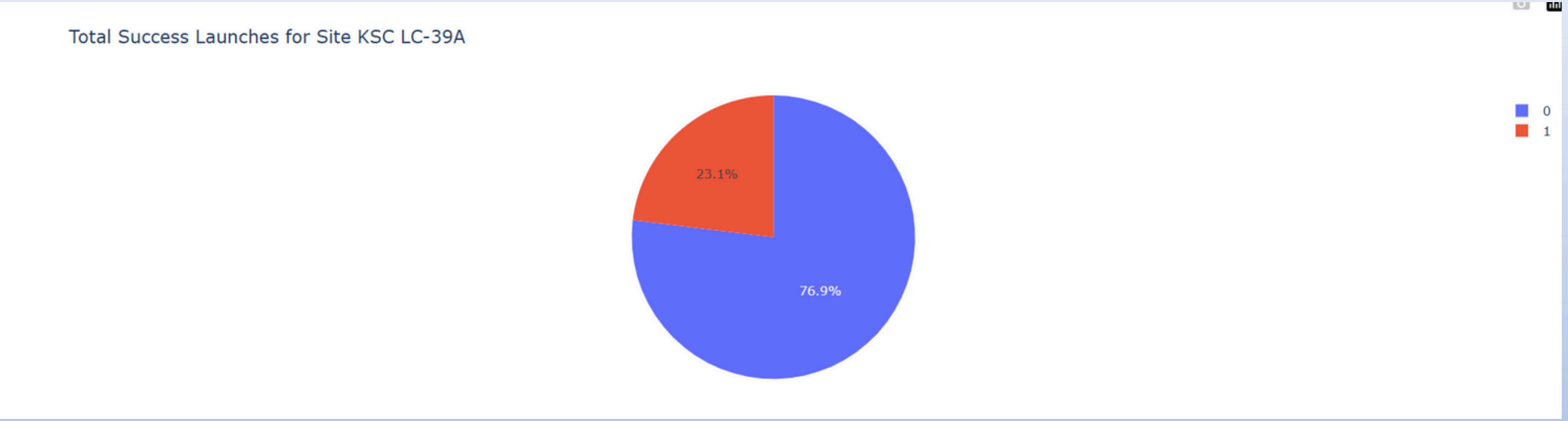
Success Rate for all Launch Sites



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



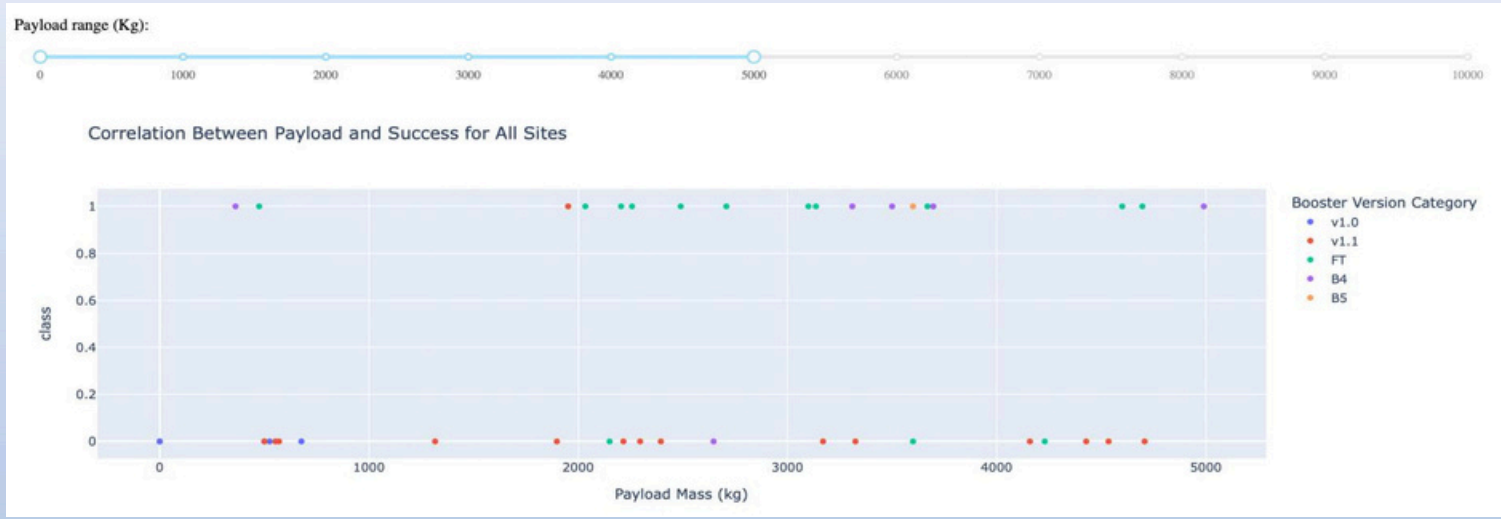
Explanation:

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

Pay Load Mass vs. Launch Outcome for All Sites

Explanation:

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



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- 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.

Scores and Accuracy of 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

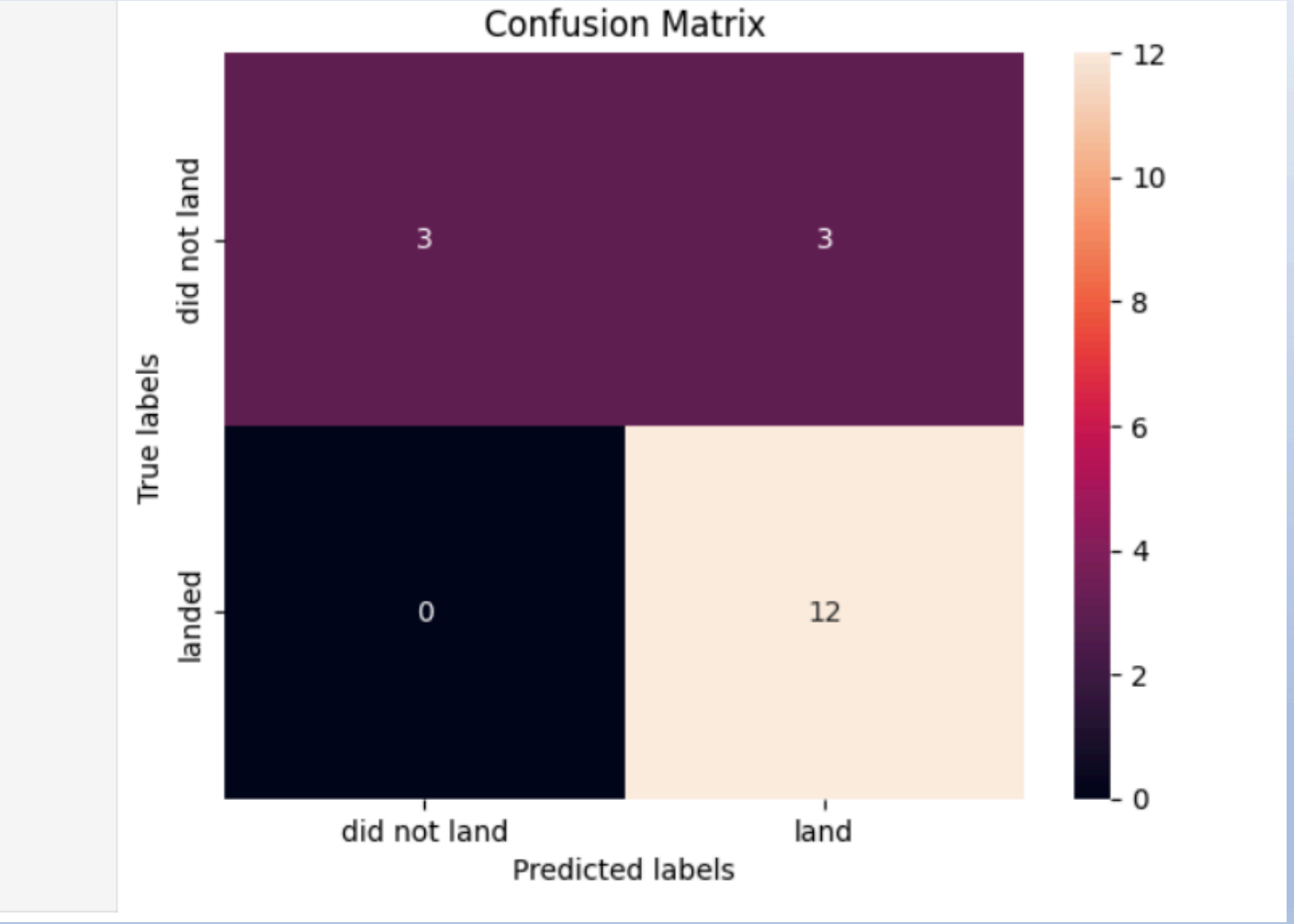
Scores and Accuracy of the data set

	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

Confusion Matrix

Explanation:

- 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 analysis of SpaceX's launch data provided a comprehensive understanding of the factors influencing the success and efficiency of space missions. Each phase of the analytical process—from data collection to predictive modeling—played a crucial role in delivering valuable insights.

The assignments demonstrated:

- The Importance of Data Preparation: Clean and structured data served as the foundation for all subsequent analyses, emphasizing the need for robust wrangling techniques.
- Insights through EDA: Exploratory Data Analysis, both visual and SQL-based, uncovered key patterns such as the impact of payload types, launch sites, and operational parameters on mission success rates.
- Interactive Tools for Stakeholder Engagement: Interactive visualizations using Folium and Dash simplified complex data, enabling users to explore trends dynamically and make informed decisions.
- Predictive Analysis for Strategic Planning: The development and fine-tuning of classification models provided accurate predictions of launch outcomes, helping to identify critical factors that could guide future mission planning.
-

This systematic approach to analyzing SpaceX's data demonstrated the power of integrating multiple methodologies and tools. The experience highlighted the potential of data analysis to optimize decision-making, improve operational strategies, and contribute to the broader understanding of space exploration challenges.

Appendix

Thank you!