Winning Space Race with Data Science

Humphry Tlou

Outline

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

Executive Summary

- Data acquisition and preparation
- Data cleaning and preprocessing
- Exploratory analysis with visual data representation
- Exploratory analysis using SQL techniques
- Creating interactive maps using Folium
- Developing dashboards with Plotly Dash
- Predictive modelling and classification analysis

Introduction

SpaceX has revolutionized the commercial space industry by making space travel more affordable. The company offers Falcon 9 rocket launches at a cost of $62 million, significantly lower than competitors who charge over $165 million. This cost efficiency is largely due to SpaceX's ability to reuse the rocket's first stage. By using public data and machine learning models, we aim to predict whether the first stage of a Falcon 9 rocket will be reusable, which directly impacts the launch cost.

Key questions to address:

- How do factors like payload mass, launch site, number of flights, and orbit type influence the success of the first stage landing?
- Has the success rate of landings improved over the years?
- Which algorithm is most effective for binary classification in this scenario?

Section 1

Methodology

Methodology

Executive Summary

Data Collection Methodology:

- Retrieved data using the SpaceX REST API.
- Collected additional data through web scraping from Wikipedia.

Data Wrangling Steps:

- Filtered and cleaned the dataset.
- Addressed missing values.
- Applied One-Hot Encoding to transform the data for binary classification.

Analysis and Visualization:

- Conducted exploratory data analysis (EDA) with visualizations and SQL.
- Created interactive visualizations using Folium and Plotly Dash.

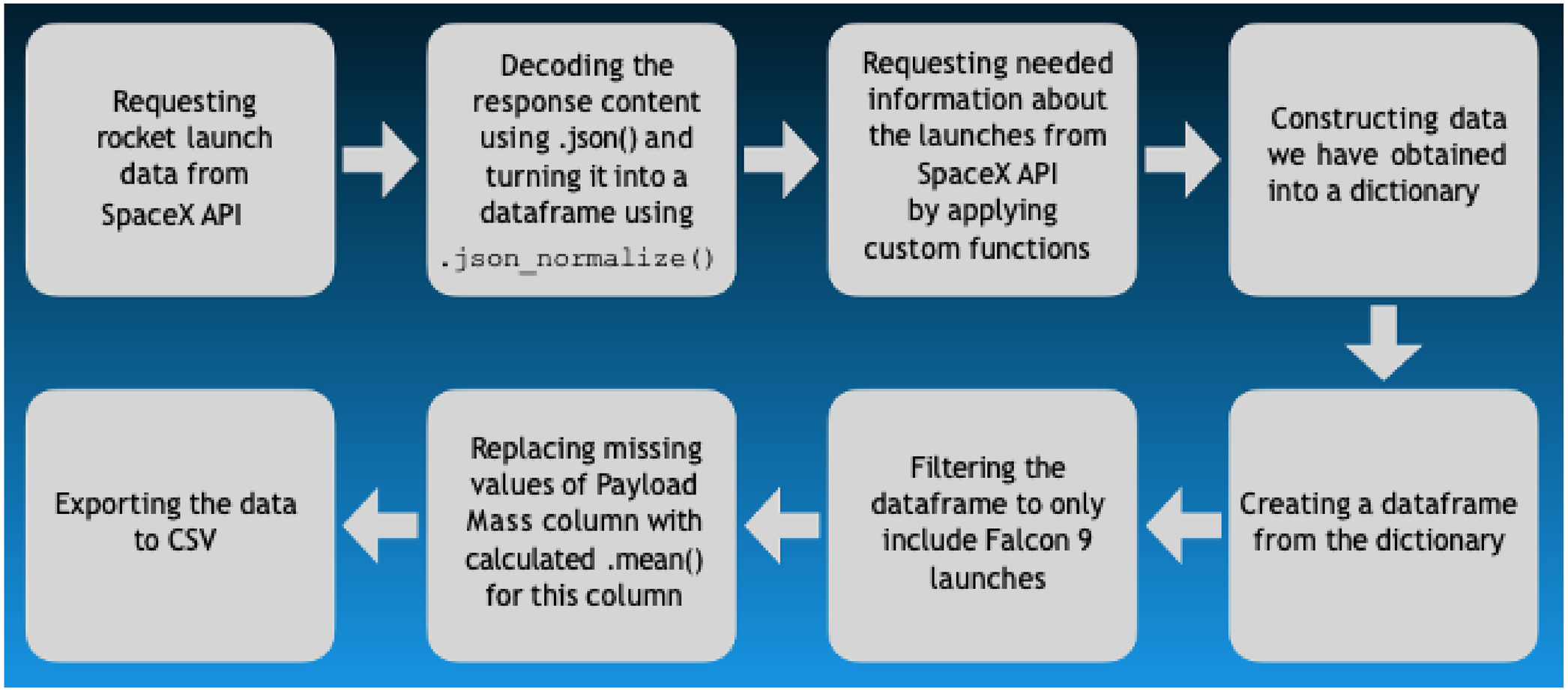
Predictive Analysis:

- Built, tuned, and evaluated classification models.
- Optimized models to ensure the highest accuracy and performance.

Data Collection

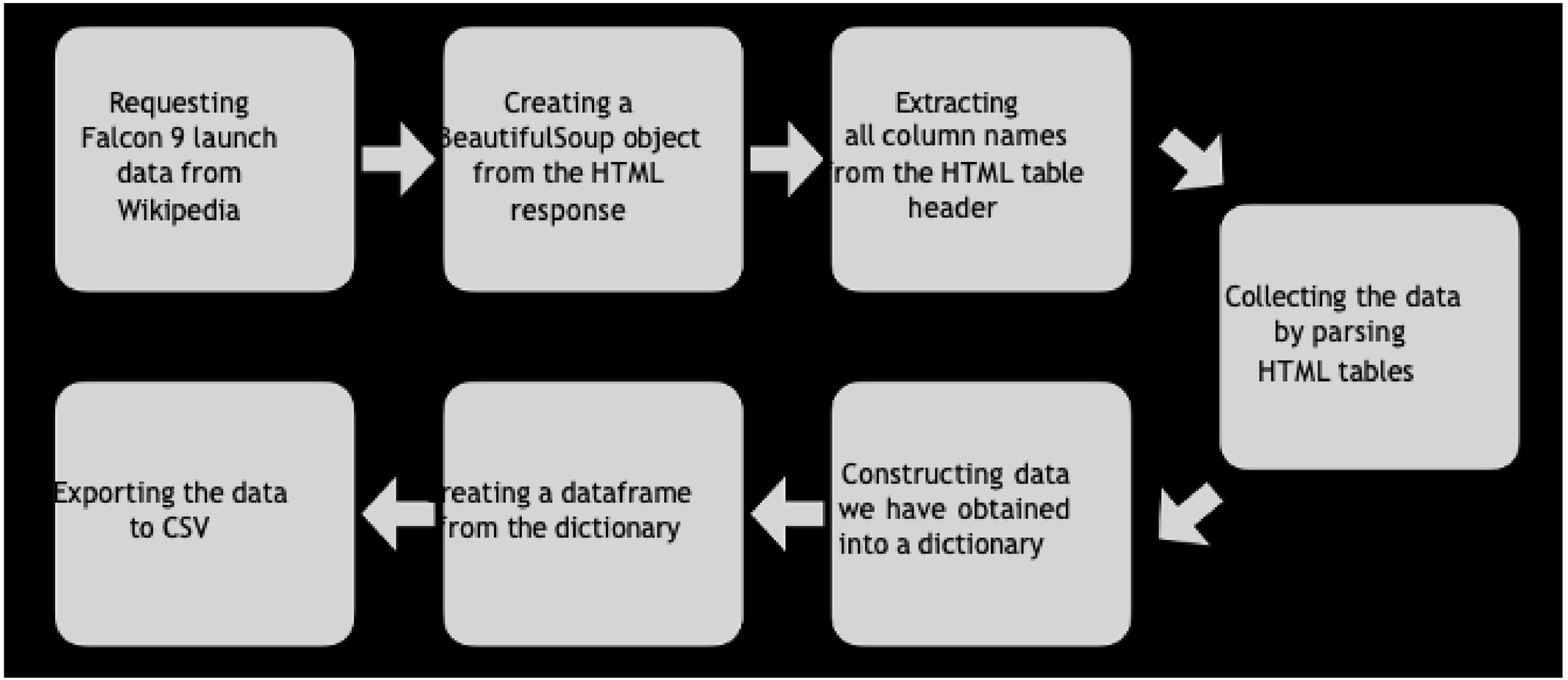
- **Data Collection Process:**
  The process involved combining data from the SpaceX REST API with web-scraped data from a Wikipedia table about SpaceX launches. Both methods were essential to gather comprehensive information for detailed analysis.
- **Data Columns Obtained via SpaceX REST API:**
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- **Data Columns Obtained via Wikipedia Web Scraping:**
- Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time

Data Collection – SpaceX API



GitHub URL

Data Collection - Scraping

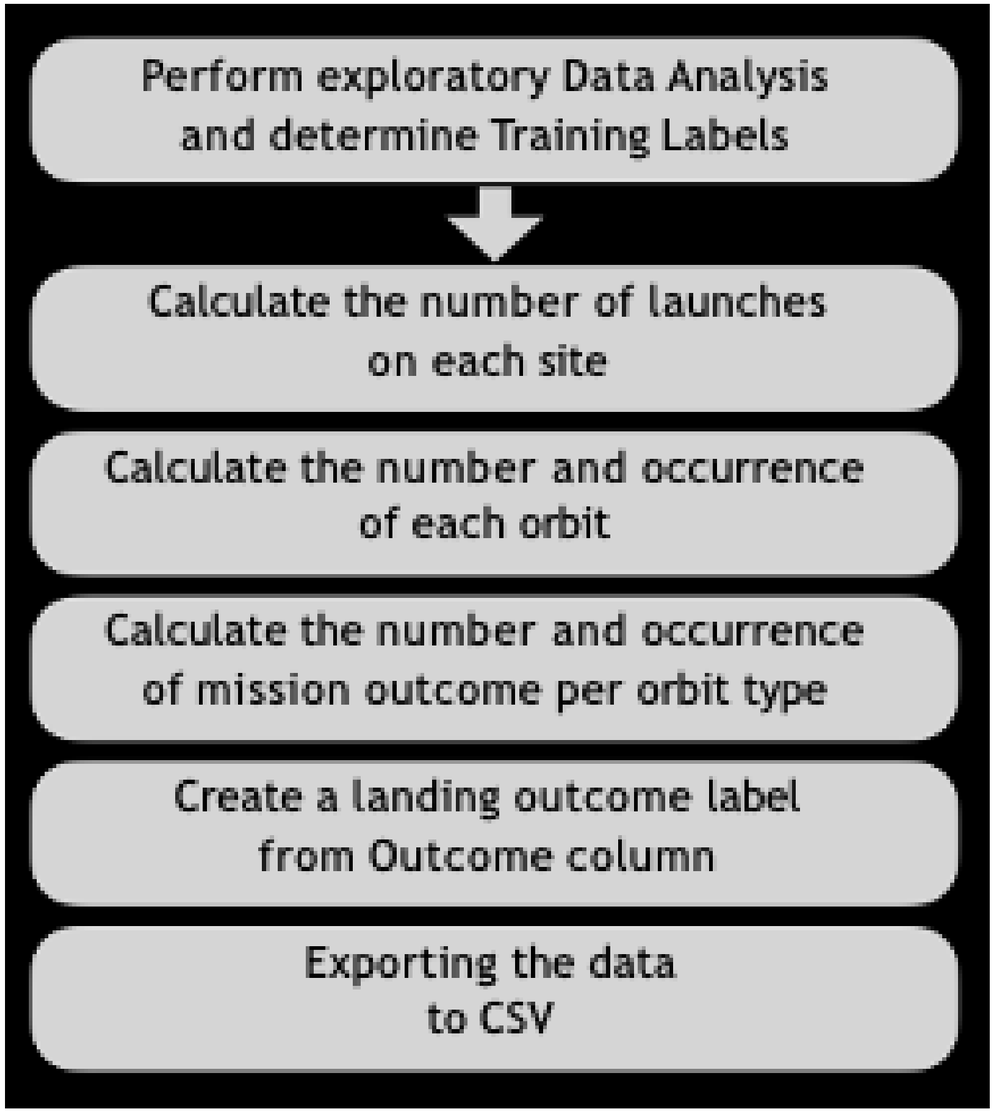


GitHub URL

Data Wrangling

- The dataset includes outcomes like "True Ocean" (successful ocean landing) and "False Ocean" (failed ocean landing), "True RTLS" (successful ground pad landing) and "False RTLS" (failed ground pad landing), and "True ASDS" (successful drone ship landing) and "False ASDS" (failed drone ship landing). These outcomes are converted into training labels: "1" for success and "0" for failure.

GitHub URL



EDA with Data Visualization

Charts Plotted:

- 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
- Success Rate Yearly Trend

Chart Insights:

- **Scatter Plots**: Reveal relationships between variables for potential use in machine learning models.
- **Bar Charts**: Compare discrete categories to measured values.
- **Line Charts**: Highlight trends over time (time series).

GitHub URL

EDA with SQL

- **SQL Queries Performed:**
- Identified unique launch site names.
- Retrieved 5 records with launch sites starting with 'CCA'.
- Calculated the total payload mass for boosters launched by NASA (CRS).
- Found the average payload mass carried by booster version F9 v1.1.
- Determined the date of the first successful ground pad landing.
- Listed boosters with successful drone ship landings carrying payloads between 4000 and 6000.
- Summarized total successful and failed mission outcomes.
- Identified booster versions with the maximum payload capacity.
- Listed failed drone ship landings, booster versions, and launch site names for 2015.
- Ranked landing outcomes (e.g., Success/Failure) between 2010-06-04 and 2017-03-20 in descending order.

GitHub URL

Build an Interactive Map with Folium

- **Markers of Launch Sites:**
- Added markers with circles, popup labels, and text labels for NASA Johnson Space Center using latitude and longitude coordinates as the starting location.
- Added markers with circles, popup labels, and text labels for all launch sites, showing their geographical locations and proximity to the Equator and coastlines.
- **Coloured Markers for Launch Outcomes:**
- Used green markers for successful launches and red markers for failed launches. Clustered markers to visualize sites with higher success rates.
- **Distances to Proximities:**
- Added colored lines to represent distances from the KSC LC-39A launch site to nearby features like railways, highways, coastlines, and the closest city.

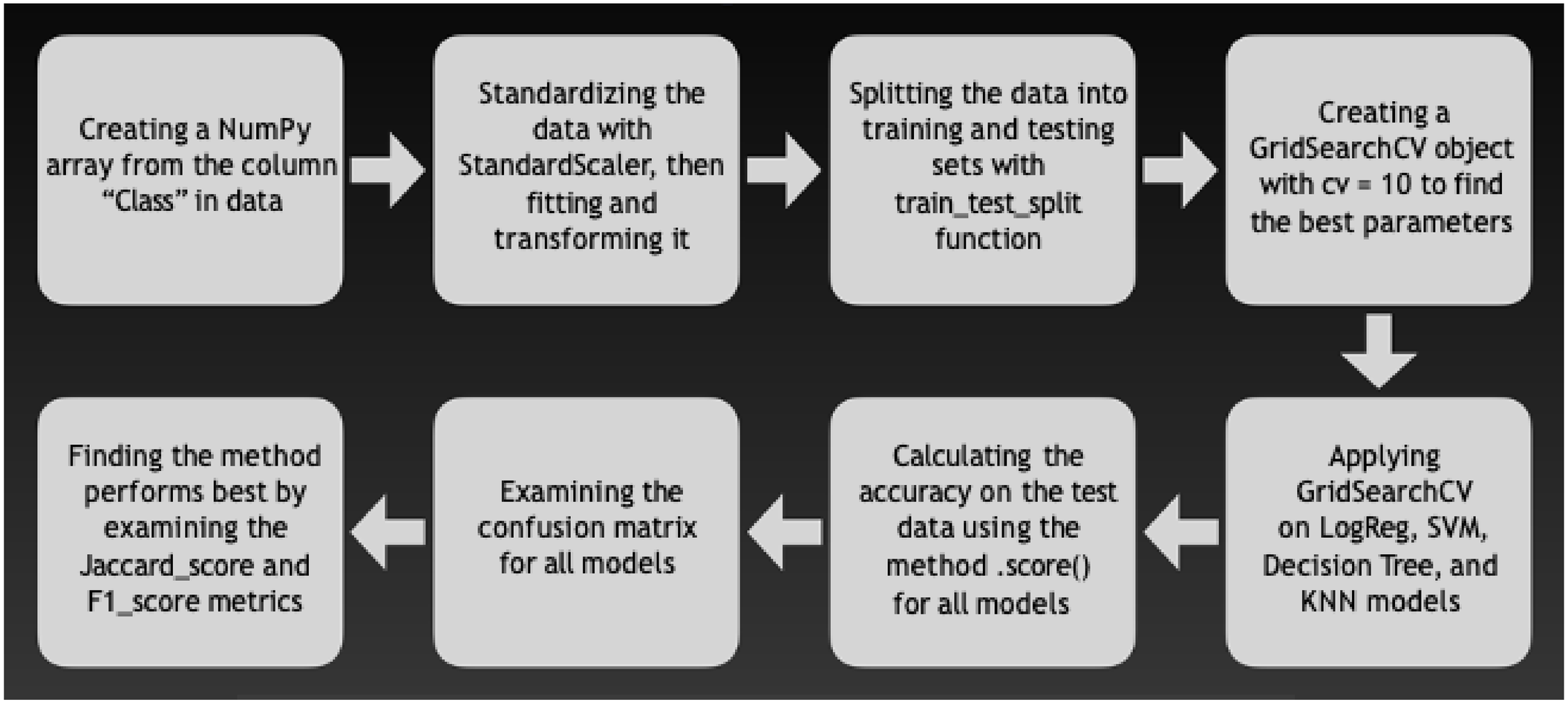
GitHub URL

Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List:**
- Implemented a dropdown to select a specific Launch Site.
- **Pie Chart for Success Launches (All Sites/Specific Site):**
- Added a pie chart to display total successful launches for all sites and Success vs. Failure counts for a selected site.
- **Payload Mass Range Slider:**
- Introduced a slider to filter data based on payload mass range.
- **Scatter Chart of Payload Mass vs. Success Rate (Booster Versions):**
- Created a scatter chart to visualize the relationship between payload mass and launch success for different booster versions.

GitHub URL

Predictive Analysis (Classification)



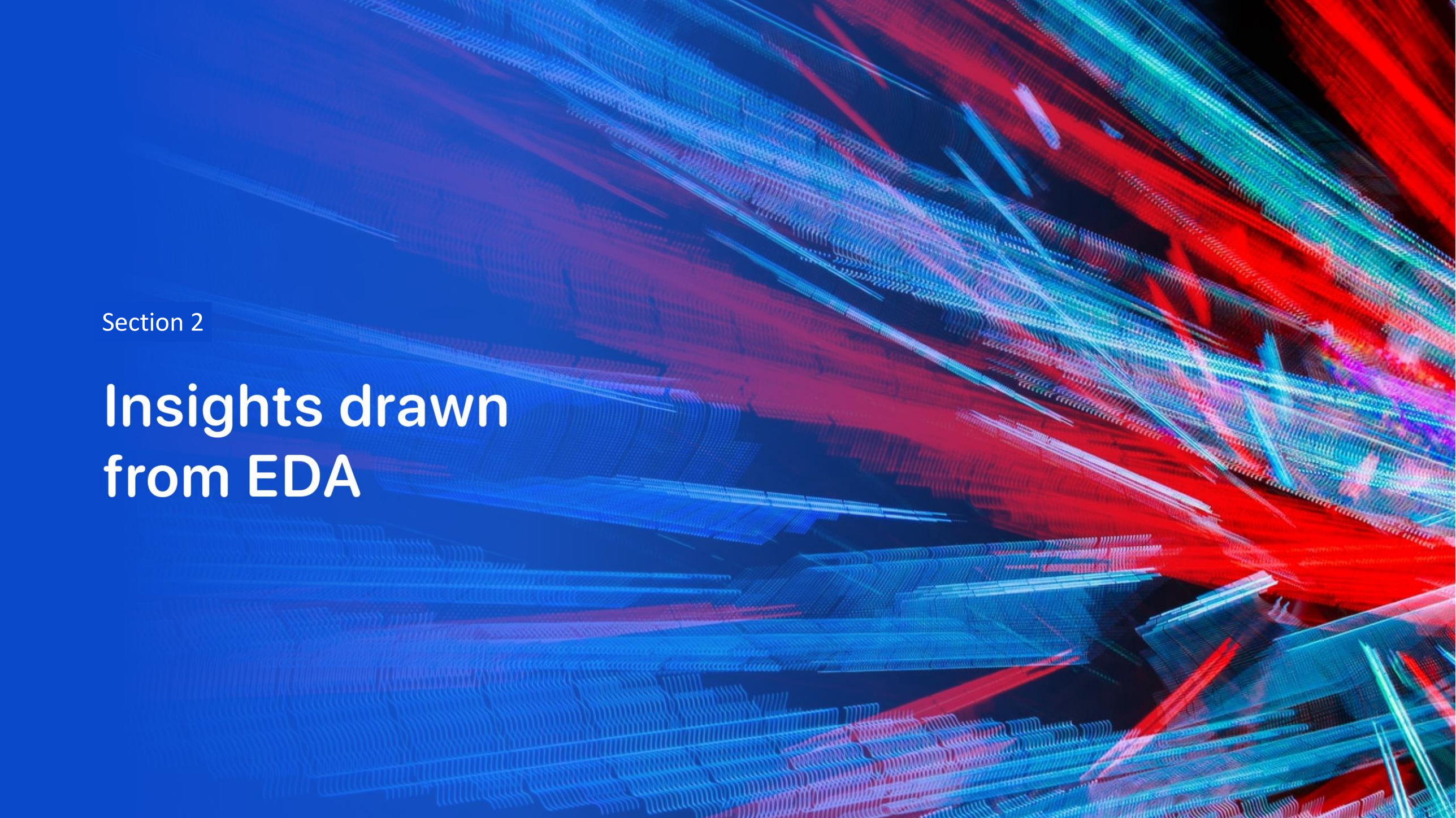
GitHub URL

Results

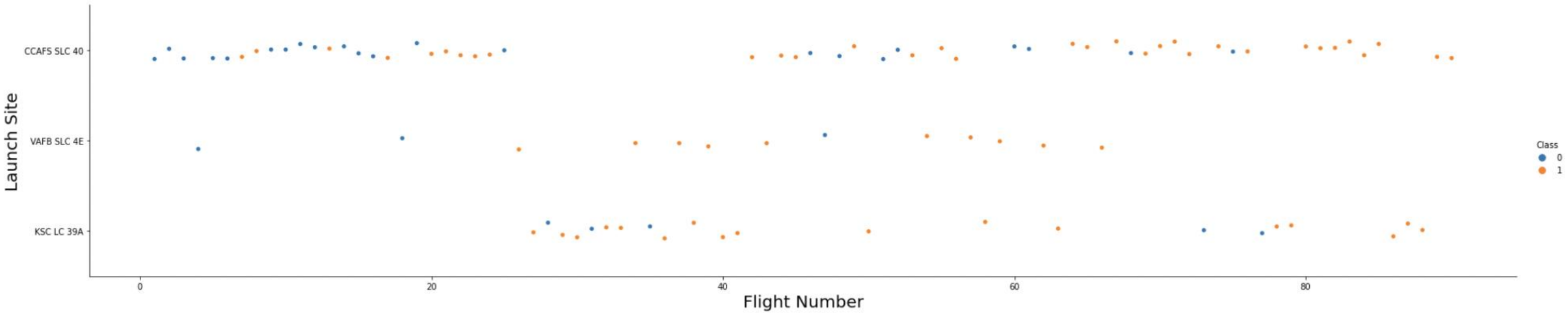
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



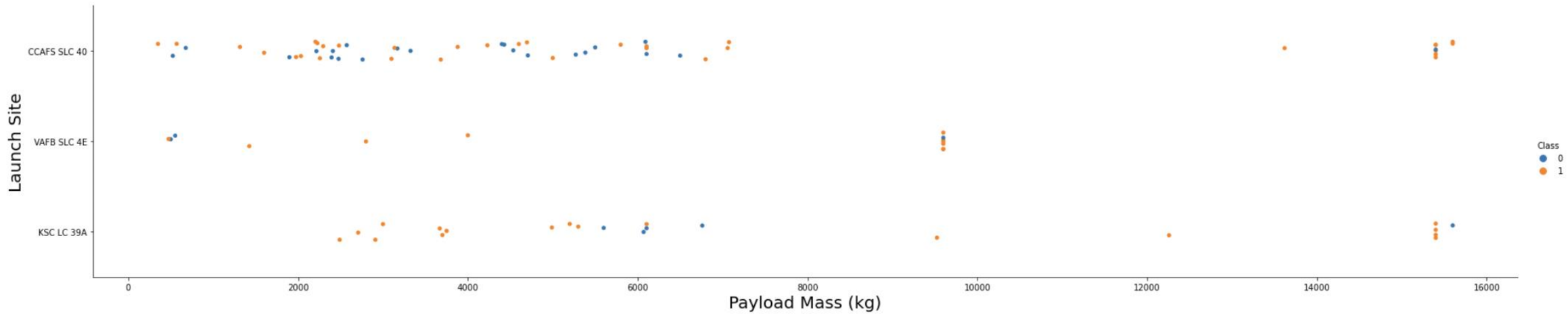
Flight Number vs. Launch Site



Explanation:

- Early flights experienced failures, while recent flights have all been successful.
- The CCAFS SLC 40 launch site accounts for nearly half of all launches.
- Higher success rates are observed at VAFB SLC 4E and KSC LC 39A.
- It can be inferred that newer launches tend to achieve greater success rates.

Payload vs. Launch Site



Explanation:

- Across all launch sites, higher payload mass correlates with higher success rates.
- Most launches carrying payloads over 7000 kg were successful.
- KSC LC 39A achieved a 100% success rate for payloads under 5500 kg.

Success Rate vs. Orbit Type

Explanation:

Orbits with 100% success rate:

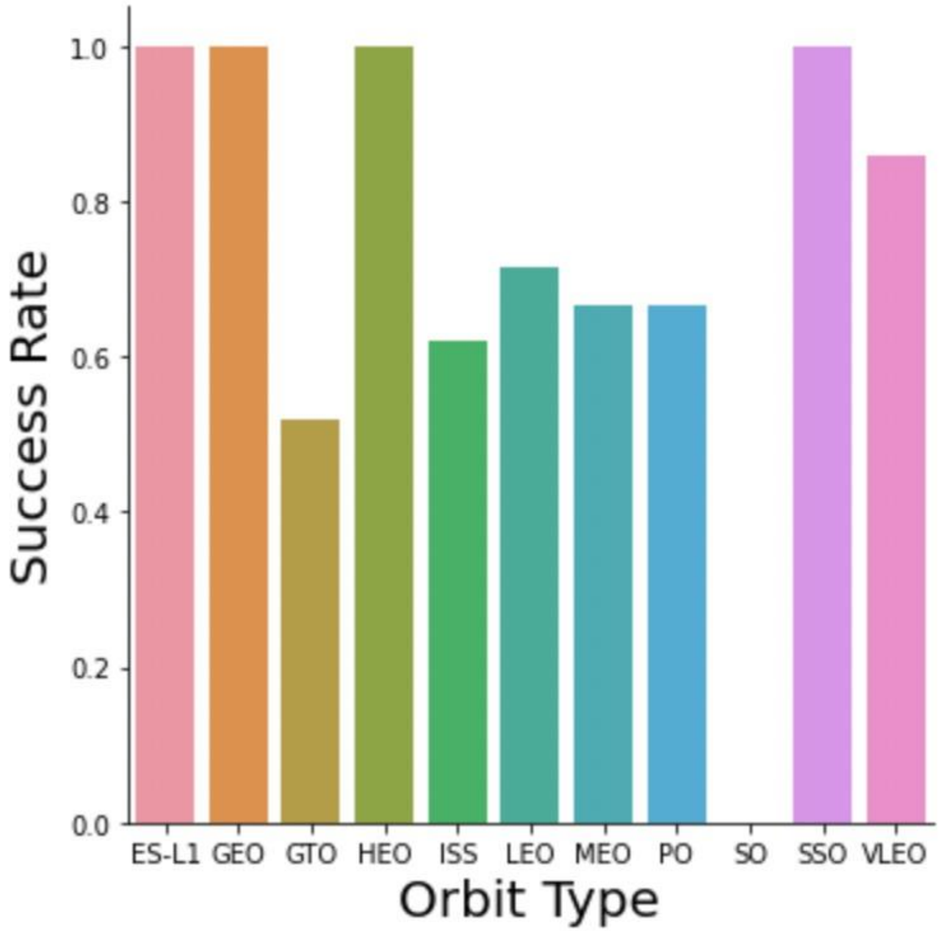
- ES-L1, GEO, HEO, SSO

Orbits with 0% success rate:

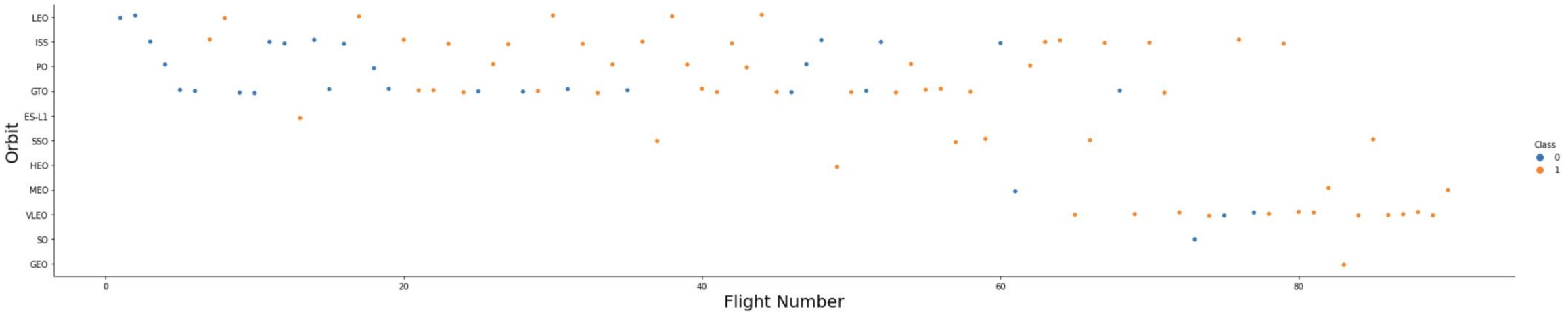
- SO

Orbits with success rate between 50% and 85%:

- GTO, ISS, LEO, MEO, PO



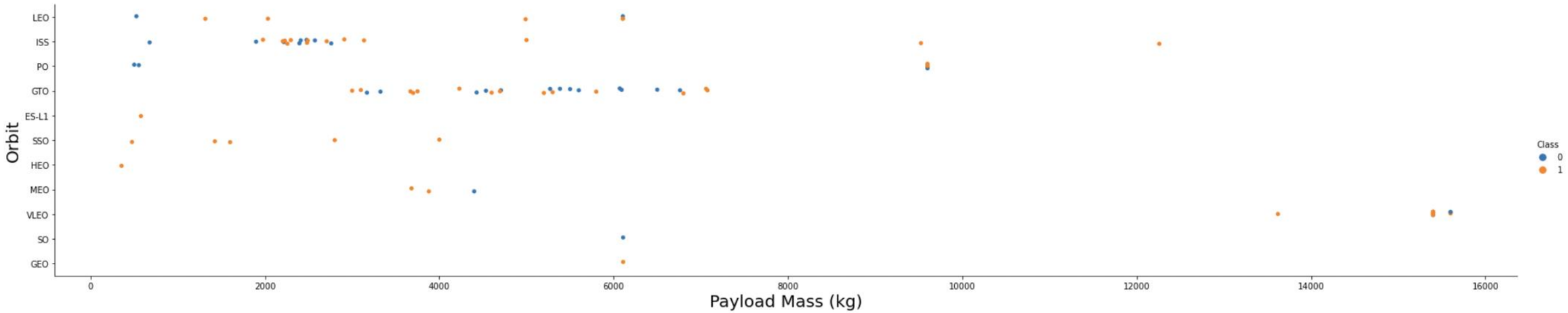
Flight Number vs. Orbit Type



Explanation:

- In the LEO orbit, success seems linked to the number of flights.
- Conversely, no clear relationship exists between flight number and success in the GTO orbit.

Payload vs. Orbit Type



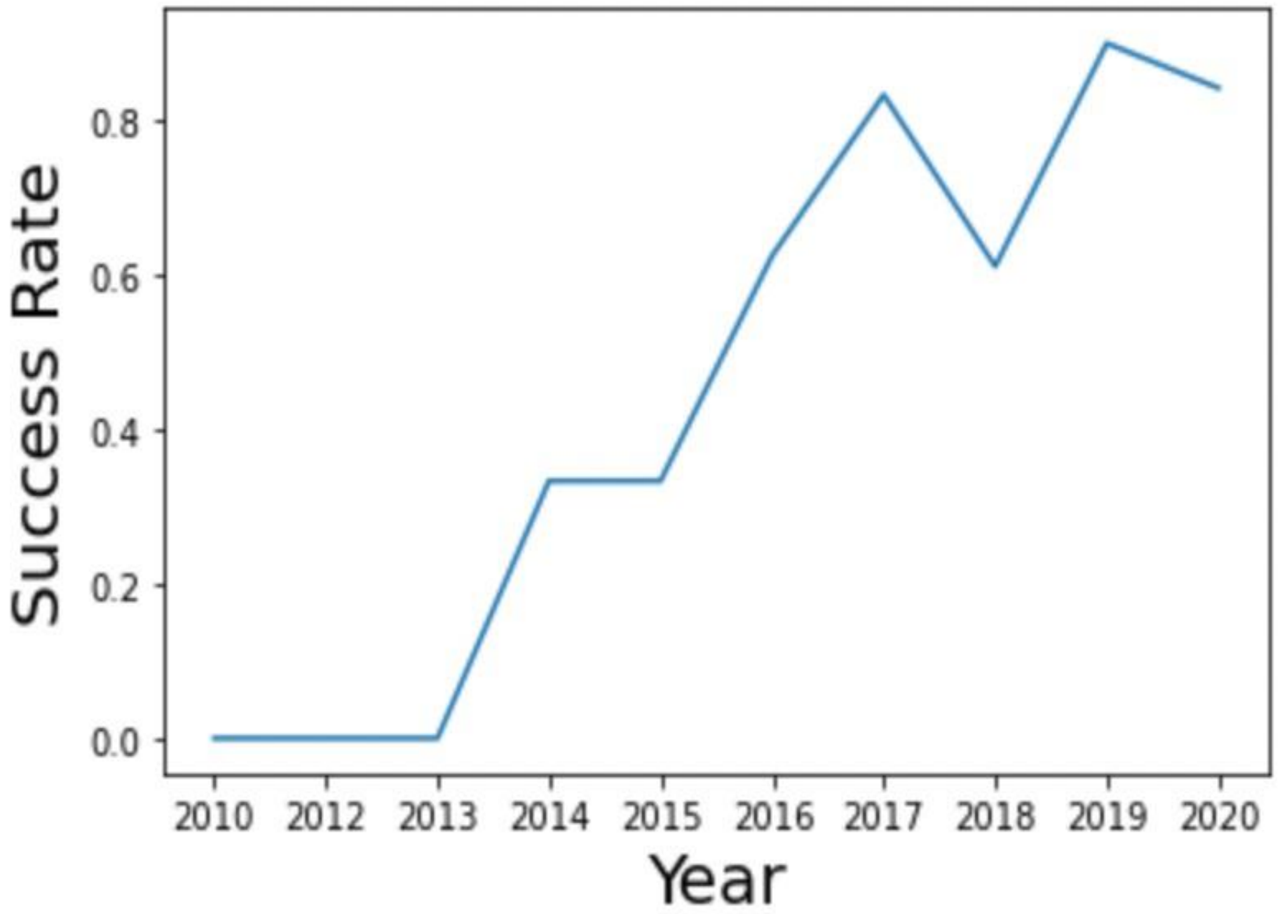
Explanation:

•Heavy payloads negatively impact GTO orbits but positively influence GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

Explanation:

•The success rate steadily increased from 2013 to 2020.



All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Explanation:

- Displaying the unique launch site names in the space mission.

Launch Site Names Begin with '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

Explanation:

- Displaying five records of launch sites starting with the string 'CCA'.

Total Payload Mass

```
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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

Explanation:

- Displaying the total payload mass transported by boosters launched by NASA (CRS).

Average Payload Mass by 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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

Explanation:

- Displaying the average payload mass transported by booster version F9 v1.1.

First Successful Ground Landing Date

```
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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Explanation:

- Listing the date of the first successful ground pad landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [9]:
```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

Out[9]:

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

Explanation:

•Identifying the boosters that successfully landed on a drone ship and carried payloads with a mass between 4000 and 6000 kg.

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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

Explanation:

- Listing the total count of mission outcomes, including both successes and failures.

Boosters Carried Maximum Payload

```
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.bs2io90l08kqb1od8lcg.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

Explanation:

- Identifying the booster versions that have transported the highest payload mass.

2015 Launch Records

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)

Explanation:

- Detailing the failed landing outcomes on drone ships, including their booster versions and launch site names, for the months of the year 2015.

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

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.bs2io90l08kqb1od8lcg.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

Explanation:

- Ranking the landing outcomes, including failures on drone ships and successes on ground pads, within the date range of June 4, 2010, to March 20, 2017, in descending order.

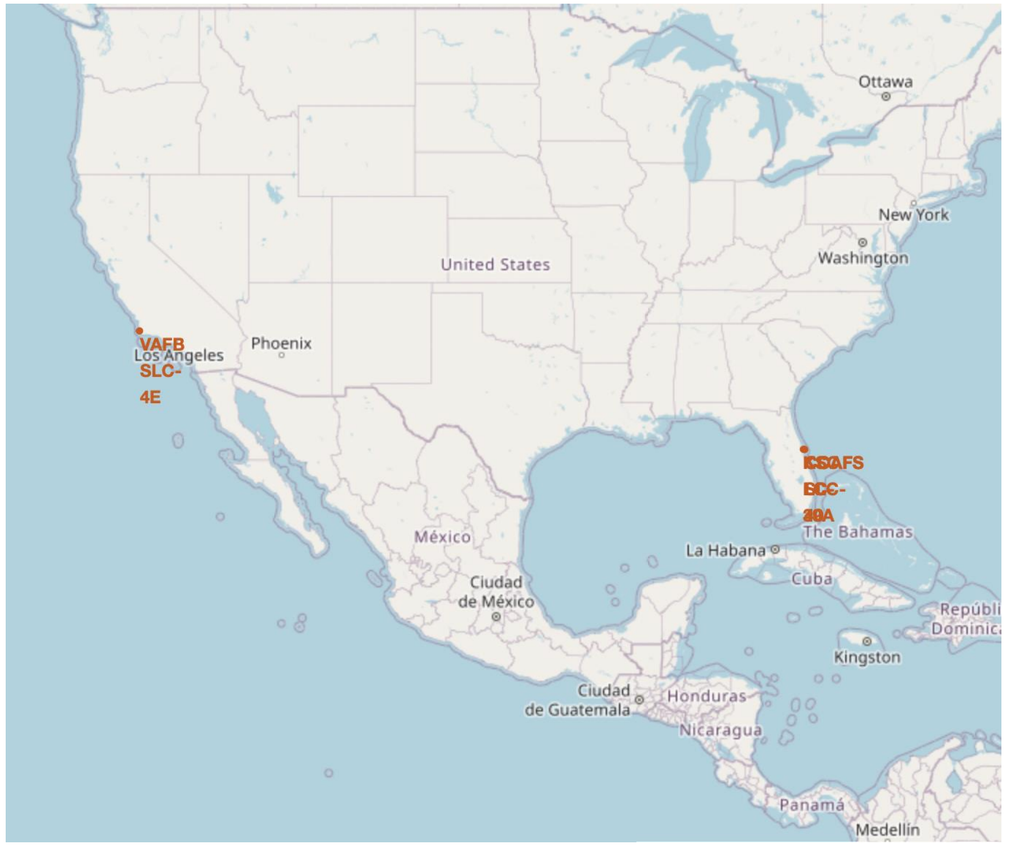
Section 3

Launch Sites Proximities Analysis

All launch sites’ location markers on the global map

Explanation:

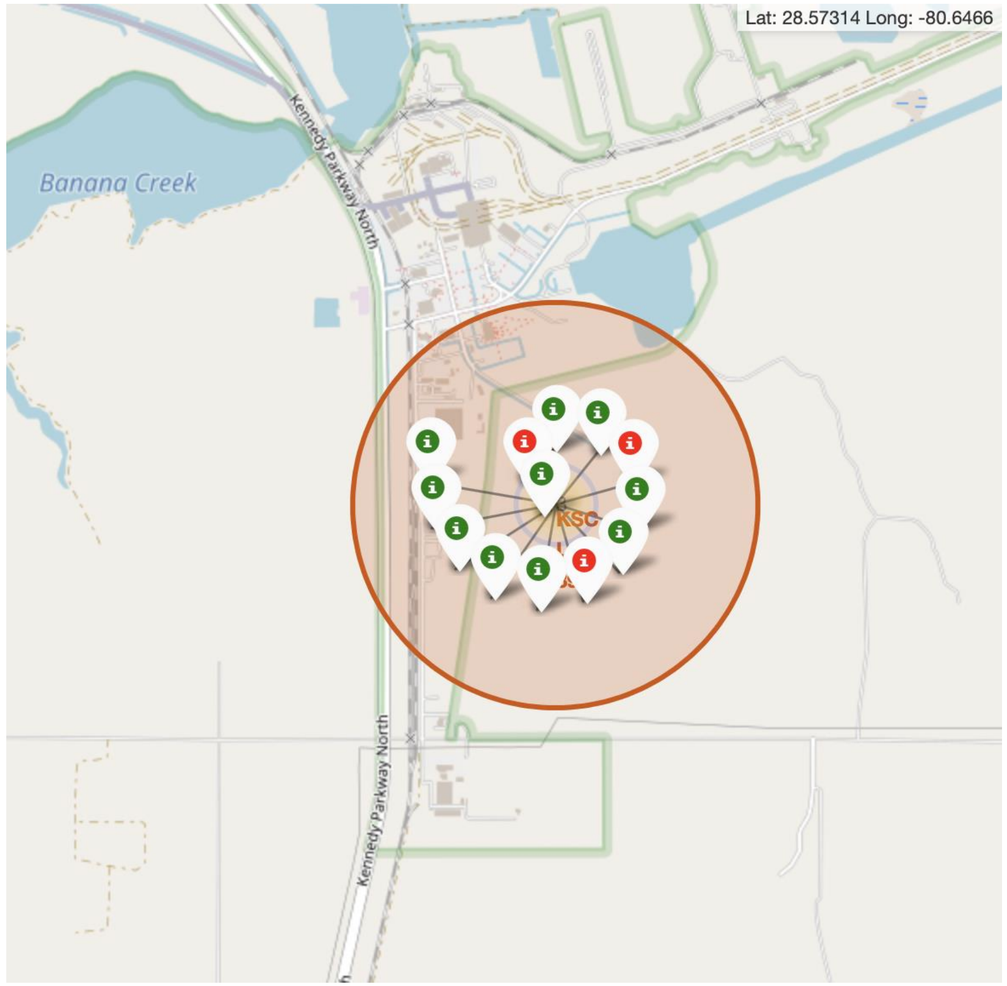
- Most launch sites are near the Equator, where the land moves faster due to Earth's rotation (1670 km/hour).
- This speed, combined with inertia, helps spacecraft maintain orbit after launch.
- Launch sites are also near coasts to minimize risks of debris or explosions affecting populated areas.



Color labeled launch records

Map Explanation:

- The color-coded markers allow for easy identification of launch sites with higher success rates.
  - Green Marker: Successful Launch
  - Red Marker: Failed Launch
- Launch Site KSC LC-39A stands out with a notably high success rate.



Distance from the launch site KSC LC-39A to its proximities

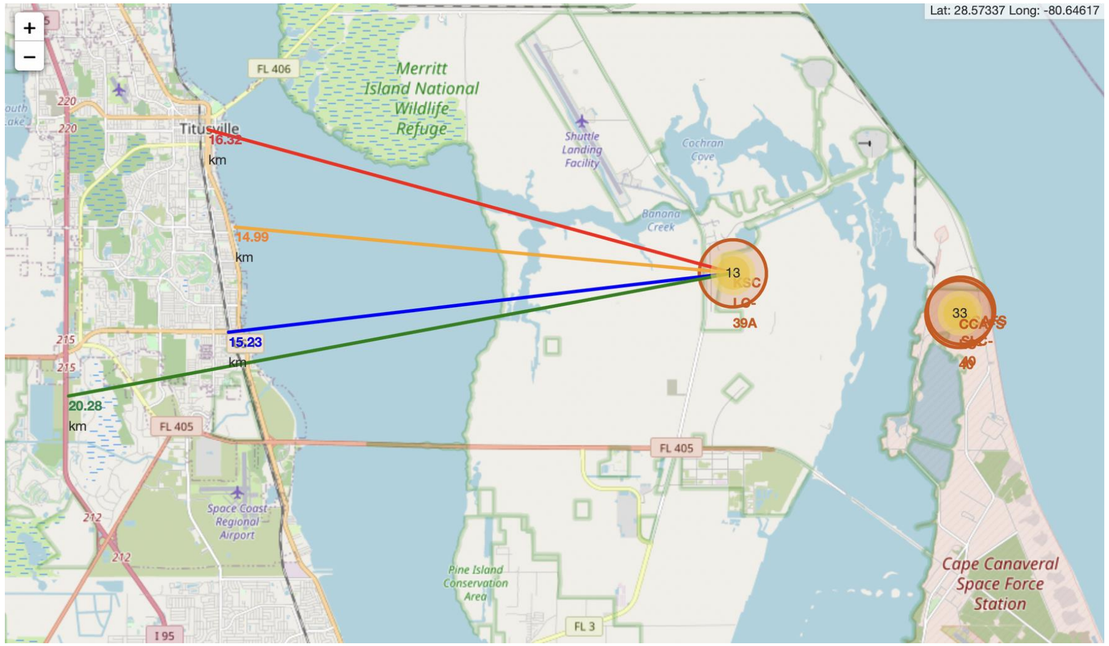
Explanation:

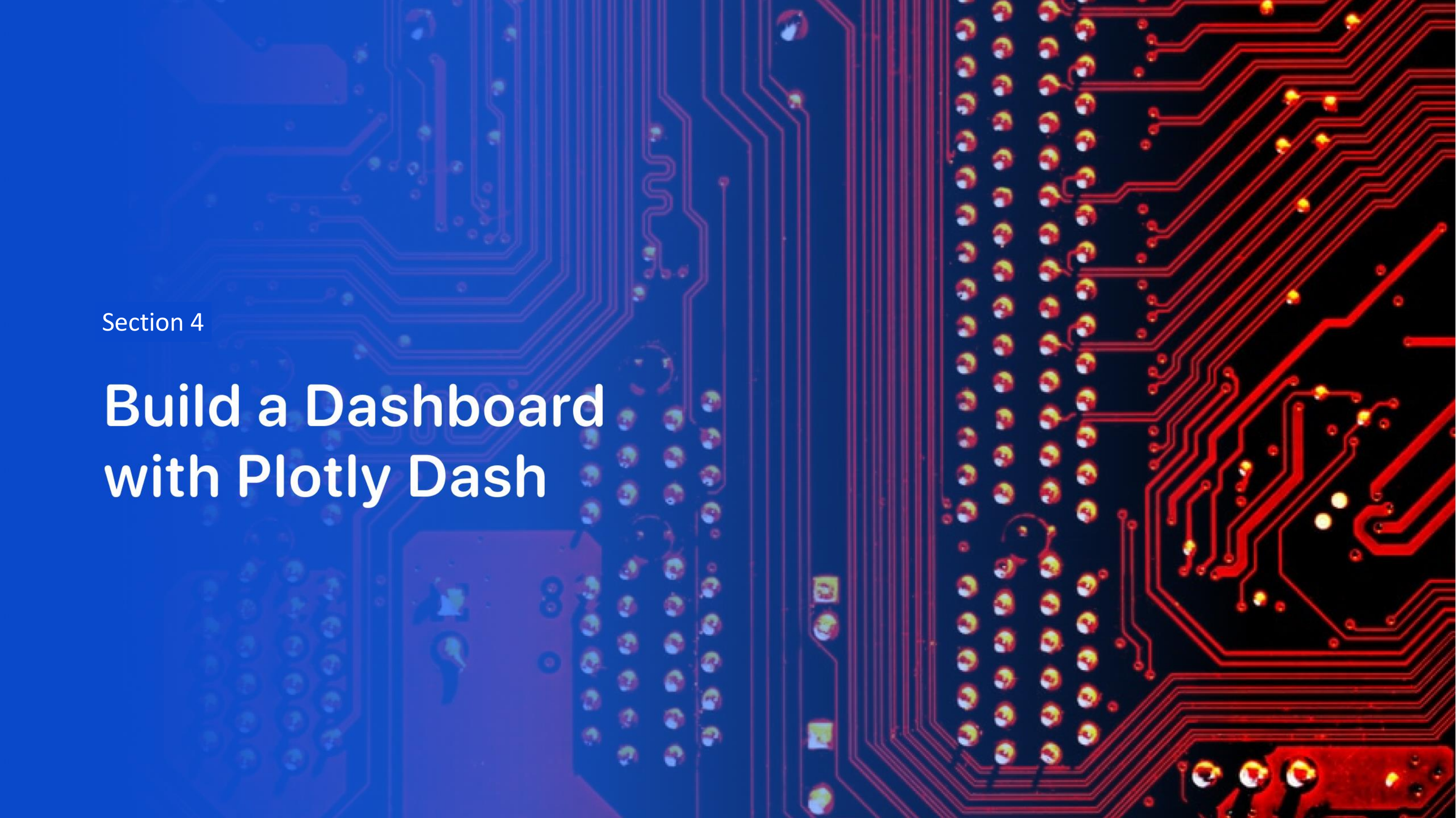
•Visual analysis of the KSC LC-39A launch site reveals its proximity to key locations:

- Railway: 15.23 km
- Highway: 20.28 km
- Coastline: 14.99 km

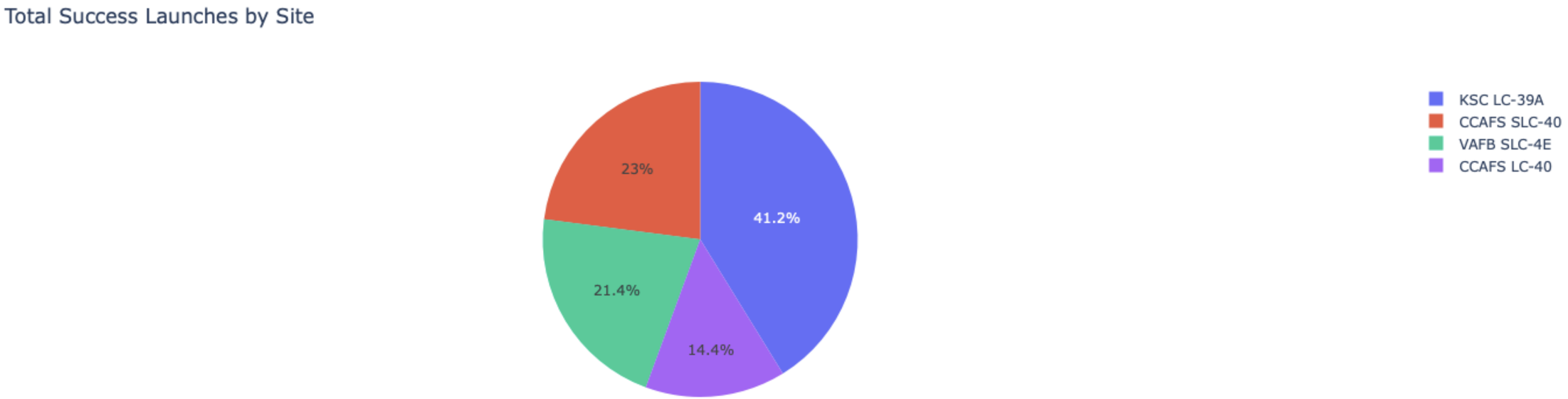
•Additionally, the launch site is relatively close to the nearest city, Titusville, at 16.32 km.

•A failed rocket traveling at high speeds can cover distances of 15-20 km in mere seconds, posing potential risks to populated areas.



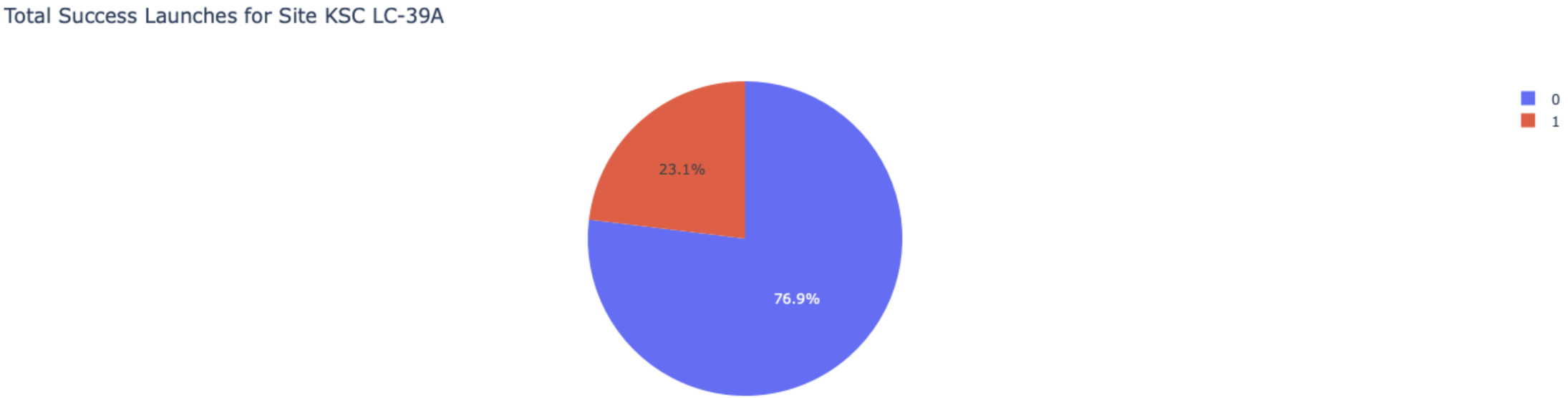

Section 4
Build a Dashboard with Plotly Dash

Total success launches by site



Explanation:
The chart highlights that KSC LC-39A stands out as the site with the highest number of successful launches among all locations.

Launch site with the highest launch success ratio



Explanation:
KSC LC-39A boasts the highest launch success rate of 76.9%, with 10 successful landings and only 3 failures.

Payload Mass vs Launch outcome for all the sites

Explanation:
The charts indicate that payloads ranging from 2000 to 5500 kg achieve the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

•Test Set scores alone cannot confirm the best-performing method due to the small sample size (18 samples).

•To address this, all methods were evaluated on the entire dataset.

•Results show that the Decision Tree Model outperforms others, achieving the highest scores and accuracy.

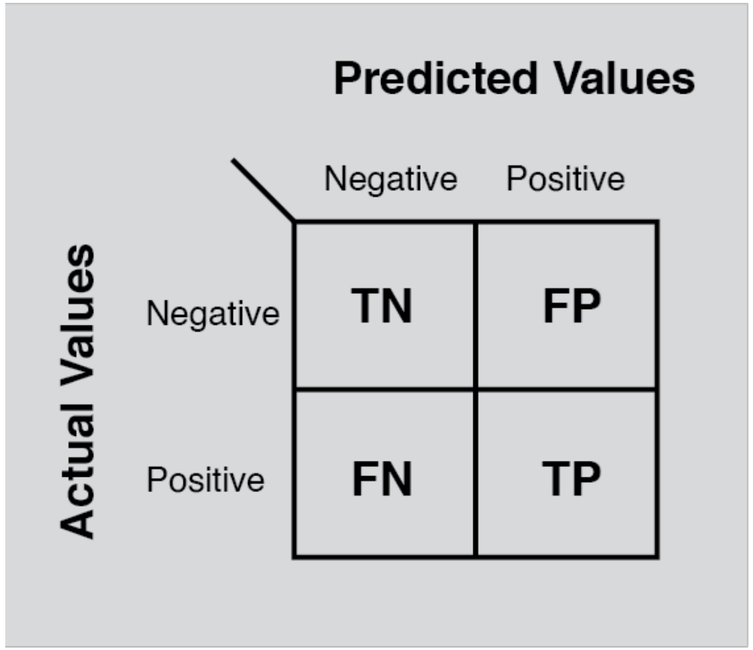
Scores and Accuracy of the Test Set

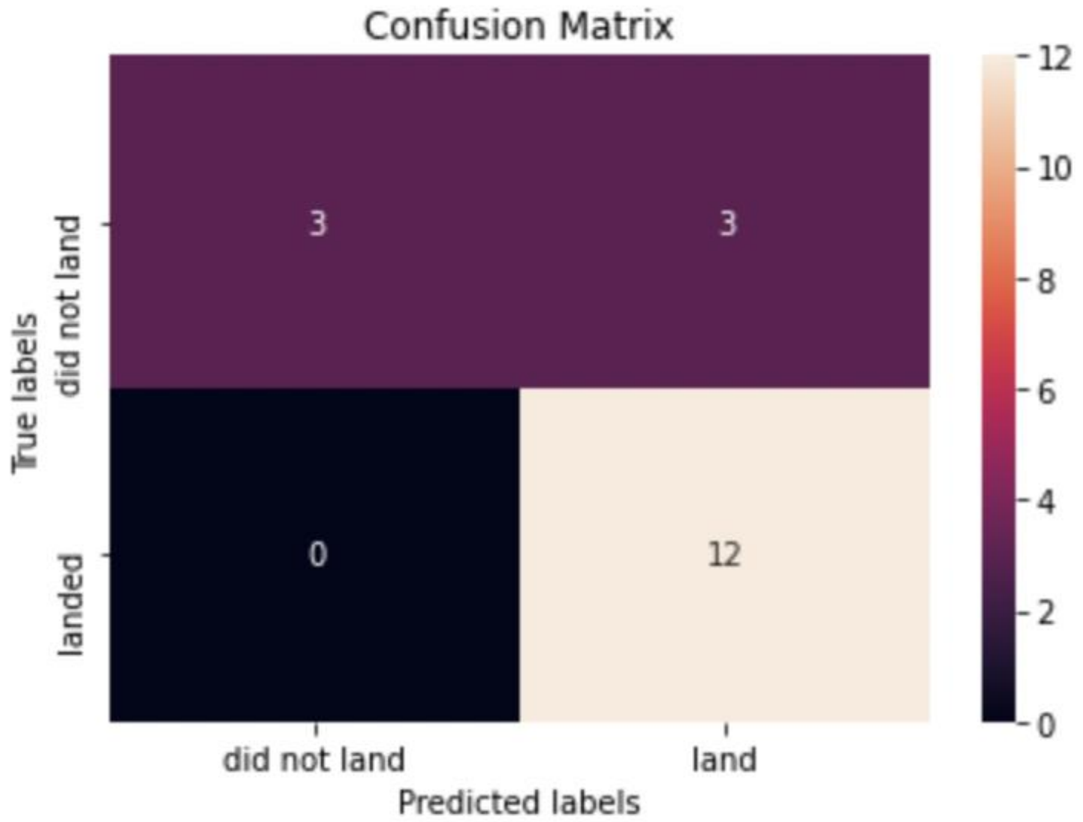
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

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

Confusion Matrix





Explanation: The confusion matrix reveals that logistic regression effectively differentiates between classes, though it struggles with false positives.

Conclusions

- Decision Tree Model is the most effective algorithm for this dataset.
- Launches with smaller payload masses yield better outcomes compared to those with larger payloads.
- Most launch sites are close to the Equator and in proximity to coastal areas.
- The success rate of launches has consistently improved over the years.
- KSC LC-39A has the highest success rate among all launch sites.
- Orbits such as ES-L1, GEO, HEO, and SSO achieve a 100% success rate.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!