Winning Space Race with Data Science

Jair Camargo del Águila
04/12/2024

Outline

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

Note: all code links are in the Git Hub and can be found in the appendices.

Executive Summary

Summary of methodologies:

- Data collection.
- Data manipulation and cleaning.
- Exploratory data analysis with graphical visualization.
- Exploratory data analysis using SQL.
- Creation of an interactive map with Folium.
- Development of an interactive dashboard with Plotly Dash.
- Predictive analysis (classification).

Summary of all results:

- Results of exploratory data analysis.
- Demonstration of interactive analysis using screenshots.
- Results of the predictive analysis.

Introduction

Project Background and Context

SpaceX stands out as the most successful company in the era of commercial space exploration, revolutionizing affordability in space travel. According to its website, SpaceX offers Falcon 9 rocket launches for $62 million, a fraction of the over $165 million charged by other providers. This significant cost reduction is largely attributed to SpaceX's ability to reuse the rocket's first stage. Consequently, predicting whether the first stage can be reused allows for estimating the launch cost. Using publicly available data and machine learning models, we aim to predict SpaceX's ability to reuse the first stage.

Questions to answer

- How do variables such as payload mass, launch site, number of flights, and orbits affect first stage landing success?
- Has the rate of successful landings increased over time?
- What is the best algorithm to use for binary classification in this case?

Section 1

Methodology

Methodology

Data Collection Methodology

- Leveraged the SpaceX REST API
- Utilized web scraping techniques from Wikipedia

Data Wrangling

- Applied data filtering
- Addressed missing values
- Employed One-Hot Encoding to prepare the dataset for binary classification

Exploratory Data Analysis (EDA)

- Conducted analysis using visualizations and SQL

Interactive Visual Analytics

- Created interactive visualizations with Folium and Plotly Dash

Predictive Analysis

- Developed, fine-tuned, and evaluated classification models to achieve optimal performance

Data Collection

The data collection process combined API requests from the SpaceX REST API with web scraping from a table in SpaceX's Wikipedia entry. Both methods were necessary to gather comprehensive information about the launches, enabling a more detailed analysis.

Data Columns from the SpaceX REST API:

- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Columns from Wikipedia Web Scraping:

- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

1. Retrieving rocket launch data via SpaceX API requests
2. Parsing the API response with .json() and converting it into a DataFrame using .json_normalize()
3. Extracting relevant launch details from the SpaceX API by implementing custom functions
4. Organizing the retrieved data into a structured dictionary
5. Generating a DataFrame from the constructed dictionary
6. Refining the DataFrame to focus exclusively on Falcon 9 launches
7. Filling missing values in the Payload Mass column with the calculated mean
8. Saving the processed data as a CSV file

Data Collection - Scraping

1. Fetching Falcon 9 launch information from Wikipedia
2. Initializing a BeautifulSoup object to process the HTML response
3. Retrieving all column headers from the HTML table
4. Parsing the HTML tables to extract the required data
5. Organizing the extracted data into a structured dictionary
6. Saving the processed data as a CSV file
7. Saving the resulting data to a CSV file

Data Wrangling

The dataset includes various scenarios where the booster landing was either successful or unsuccessful. For example, a successful ocean landing is labeled as "True Ocean," while a failed attempt is labeled as "False Ocean." Similarly, "True RTLS" and "True ASDS" represent successful landings on a ground pad or drone ship, respectively, while their "False" counterparts indicate failures. These outcomes are simplified into training labels: a successful booster landing is marked as "1," while an unsuccessful attempt is marked as “0.“

1. Conduct exploratory data analysis and define training labels.
2. Determine the total number of launches at each site.
3. Analyze the frequency and distribution of each orbit type.
4. Examine the count and frequency of mission outcomes for each orbit type.
5. Generate a landing outcome label based on the Outcome column.
6. Save the processed data as a CSV file.

EDA with Data Visualization

Various charts were created:

- Scatter plots to visualize relationships between variables like Flight Number and Payload Mass, Flight Number and Launch Site, and others. These can help identify potential correlations for machine learning models.
- Bar charts to compare discrete categories and highlight their relationships with a measured value.
- Line charts to display trends over time (time series), illustrating yearly success rates and other patterns.

EDA with SQL

SQL queries were executed to:

- Retrieve unique launch site names involved in space missions.
- Display 5 records for launch sites starting with "CCA."
- Calculate the total payload mass carried by NASA (CRS) boosters.
- Determine the average payload mass for the F9 v1.1 booster version.
- Identify the date of the first successful ground pad landing.
- List boosters with successful drone ship landings and a payload mass between 4000 and 6000.
- Additional queries included:
- Counting the total number of successful and failed mission outcomes.
- Identifying booster versions with the highest payload capacity.
- Listing failed drone ship landings in 2015, including booster versions and launch site names.
- Ranking landing outcomes (e.g., Failure on drone ship, Success on ground pad) between June 4, 2010, and March 20, 2017, in descending order.

Build an Interactive Map with Folium

Launch site markers were created as follows:

- Placed a marker with a circle, popup label, and text label for NASA Johnson Space Center using its coordinates as the initial location.
- Added markers with circles, popup labels, and text labels for all launch sites, using their coordinates to display their geographic positions and proximity to the Equator and coastlines.

Colored markers represented launch outcomes at each site:

- Used green markers for successful launches and red markers for failed ones, with a marker cluster to highlight sites with higher success rates.
- Distances from a launch site to nearby areas:
- Added color-coded lines to indicate the distances from KSC LC-39A (as an example) to nearby features like railways, highways, coastlines, and the nearest city.

Build a Dashboard with Plotly Dash

Dropdown for selecting launch sites:

- Implemented a dropdown menu to choose a launch site.

Pie chart for launch success (overall or by site):

- Included a pie chart displaying the total count of successful launches across all sites, or the success and failure counts for a selected site.

Payload mass range slider:

- Added a slider to set the payload mass range.

Scatter plot of payload mass vs. success rate for booster versions:

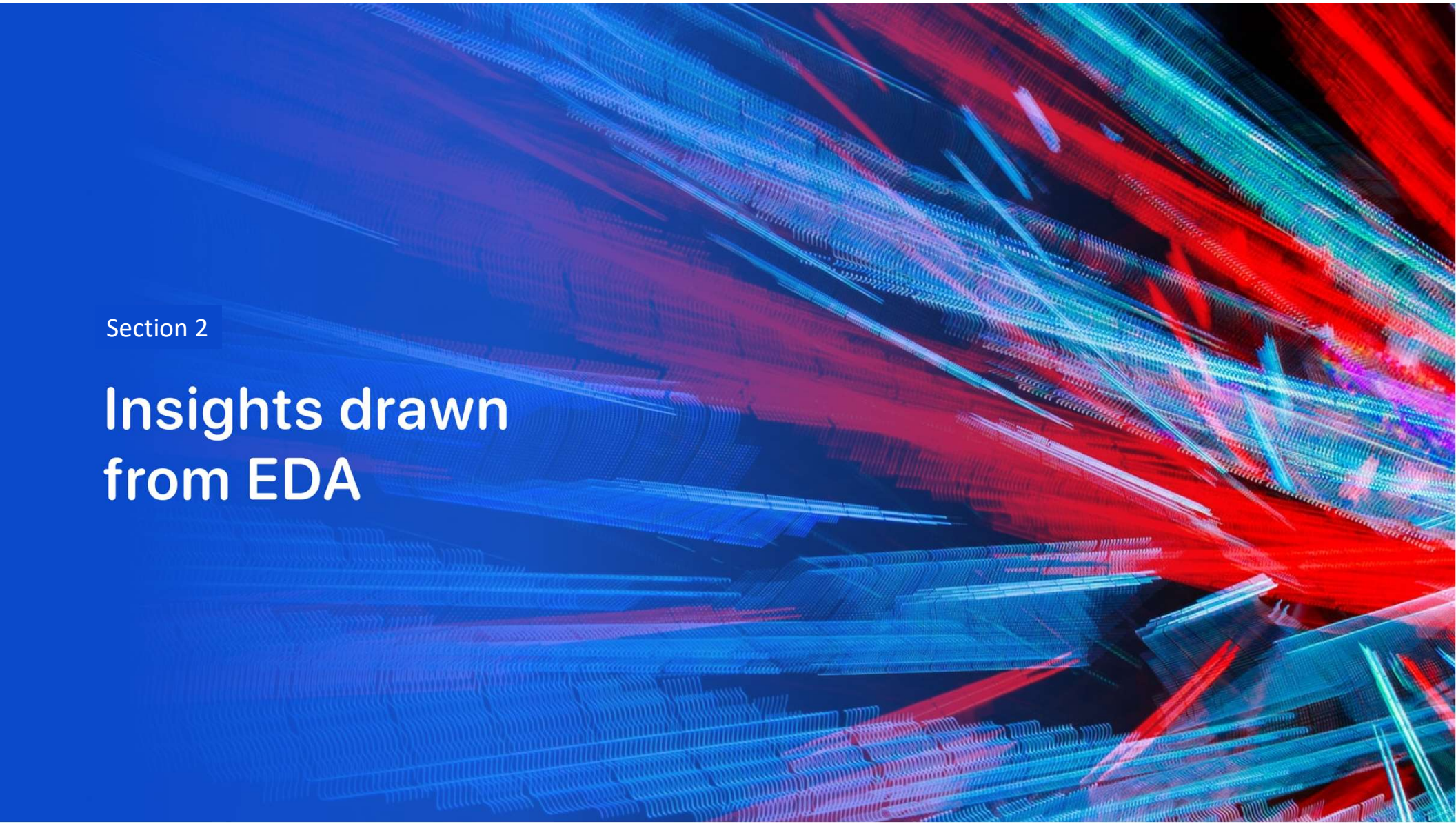
- Created a scatter plot to illustrate the relationship between payload mass and launch success rate for different booster versions.

Predictive Analysis (Classification)

1. Generating a NumPy array from the "Class" column in the dataset.
2. Standardizing the data using StandardScaler and applying the fit and transform methods.
3. Dividing the data into training and testing sets using the train_test_split function.
4. Creating a GridSearchCV instance with 10-fold cross-validation to determine optimal parameters.
5. Running GridSearchCV for LogReg, SVM, Decision Tree, and KNN models.
6. Evaluating model accuracy on the test set with the .score() method for all models.
7. Reviewing the confusion matrix for each model.
8. Identifying the best-performing method by analyzing Jaccard_score and F1_score metrics.

Results

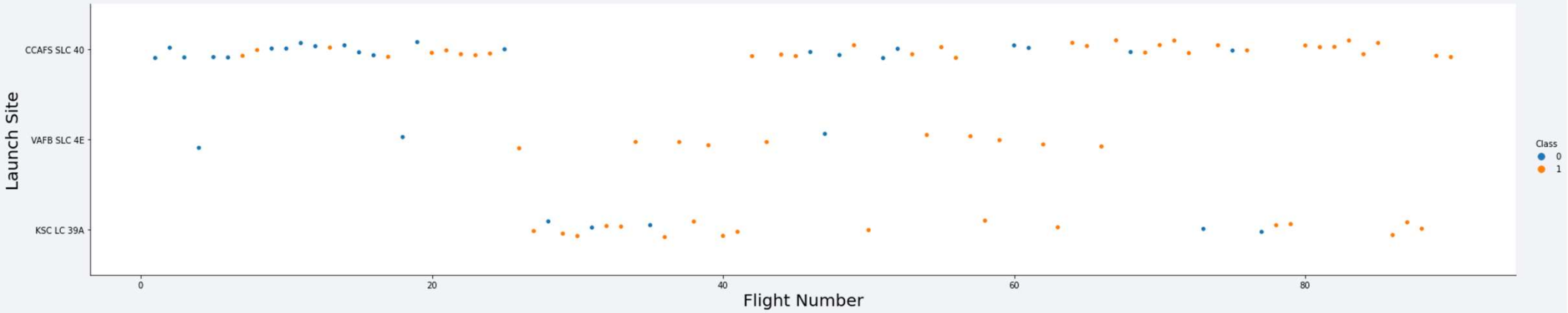
- Exploratory Data Analysis Findings
- Demonstrated interactive analytics through screenshots.
- Presented the results of the predictive analysis.



Section 2

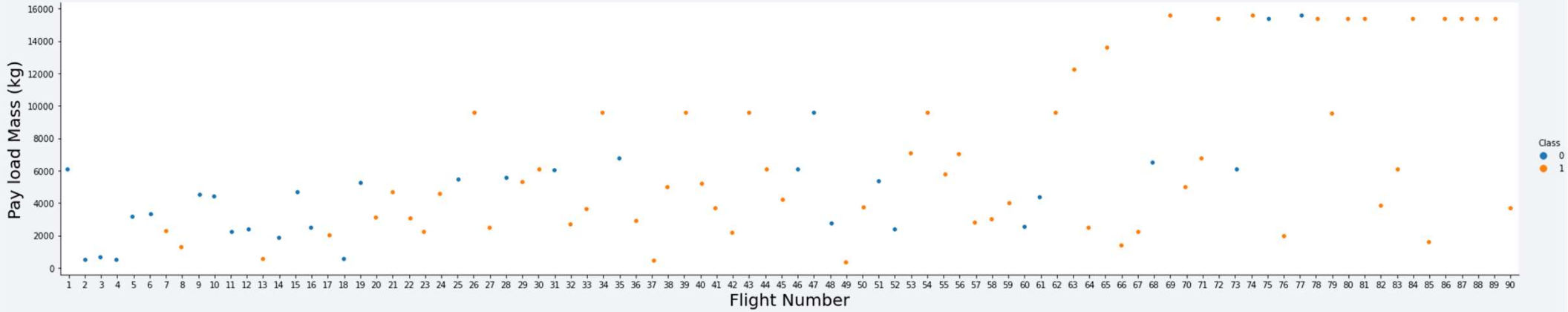
Insights drawn from EDA

Flight Number vs. Launch Site



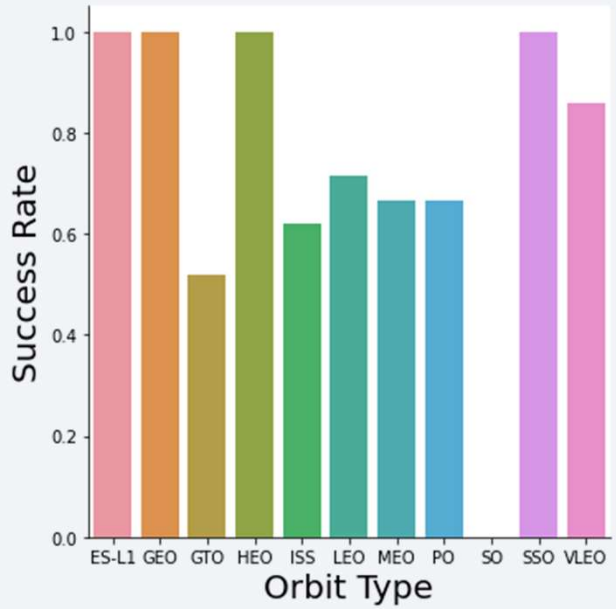
- The initial flights were all unsuccessful, whereas the most recent ones have all been successful.
- Approximately half of all launches took place at the CCAFS SLC 40 site.
- The VAFB SLC 4E and KSC LC 39A launch sites show higher success rates.
- It can be inferred that newer launches tend to have a greater success rate.

Payload vs. Launch Site



- At each launch site, a higher payload mass correlates with a higher success rate.
- The majority of launches with a payload mass exceeding 7000 kg were successful.
- KSC LC 39A also demonstrates a 100% success rate for payload masses below 5500 kg.

Success Rate vs. Orbit Type



Orbits that achieved a 100% success rate include:

- ES-L1, GEO, HEO, SSO

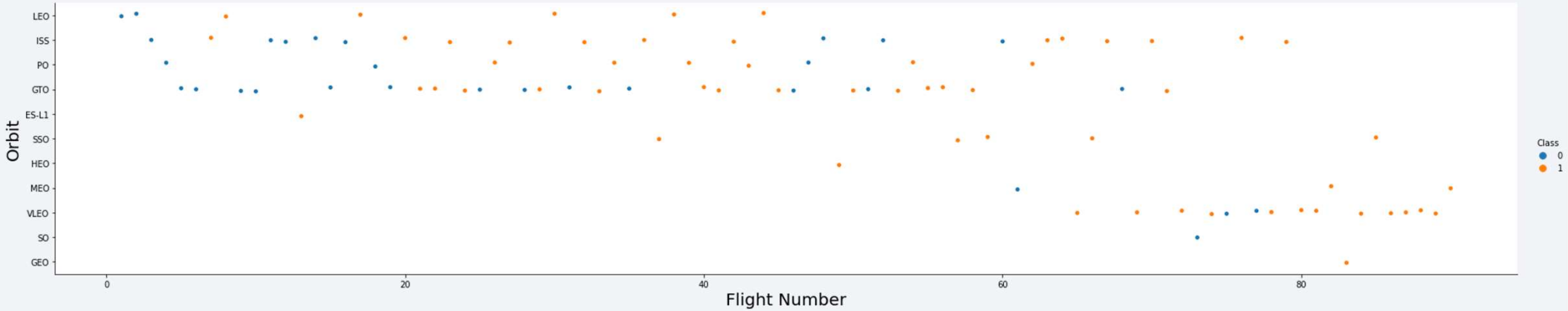
Orbits that had a 0% success rate:

- SO

Orbits with success rates ranging from 50% to 85%:

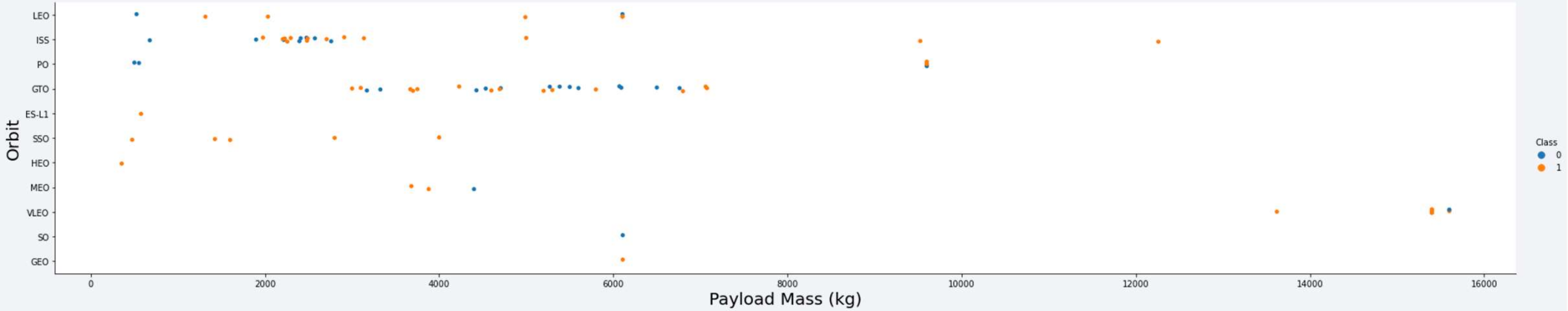
- GTO, ISS, LEO, MEO, PO

Flight Number vs. Orbit Type



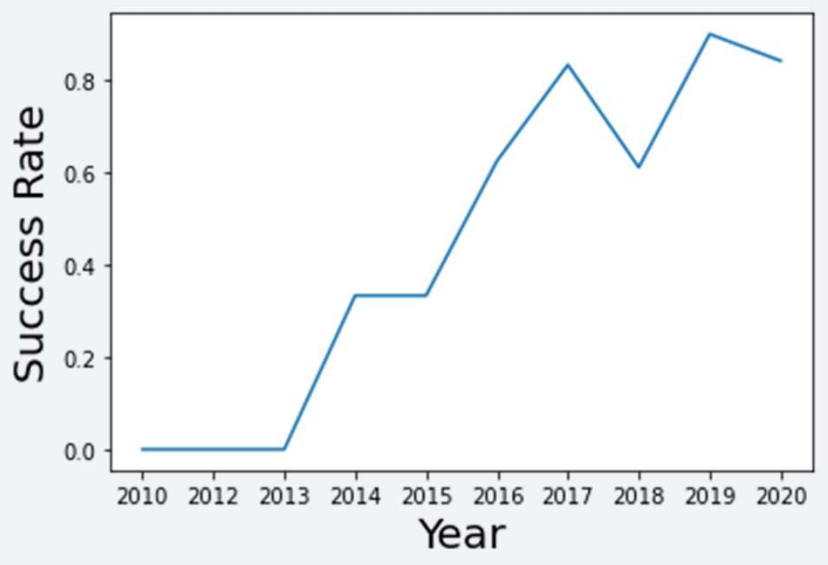
- In the LEO orbit, the success rate appears to be associated with the number of flights. Conversely, there doesn't seem to be any correlation between the number of flights and success rate in the GTO orbit.

Payload vs. Orbit Type



- Heavy payloads have a negative impact on GTO orbits but a positive effect on both GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



The success rate has steadily increased from 2013 up to 2020.

All Launch Site Names

```
%sql select distinct launch_site from SPACEXDATASET;
```

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

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

Listing the distinct names of the launch sites used in the space mission.

Showing 5 records where the launch site names start with the string 'CCA'.

Launch Site Names Begin with 'CCA'

```
%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.
```

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

- Showing 5 records where the launch site names start with the string 'CCA'.

Total Payload Mass

```
%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.
total_payload_mass
45596
```

Displaying the total payload mass delivered by boosters launched under NASA's CRS program.

Average Payload Mass by F9 v1.1

```
%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.
average_payload_mass
2534
```

Showing the average payload mass delivered by the F9 v1.1 booster version.

First Successful Ground Landing Date

```
%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.
first_successful_landing
2015-12-22
```

Displaying the date of the first successful landing on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Displaying the names of boosters that successfully landed on a drone ship with a payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

```
%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.
```

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

Displaying the total count of successful and failed mission outcomes.

Boosters Carried Maximum Payload

```
%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.
```

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

Displaying the booster versions that have transported the highest payload mass.

2015 Launch Records

```
%%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.
```

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)

Displaying the failed landing outcomes on drone ships, including the corresponding booster versions and launch site names for the months of 2015.

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

```
%%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.
```

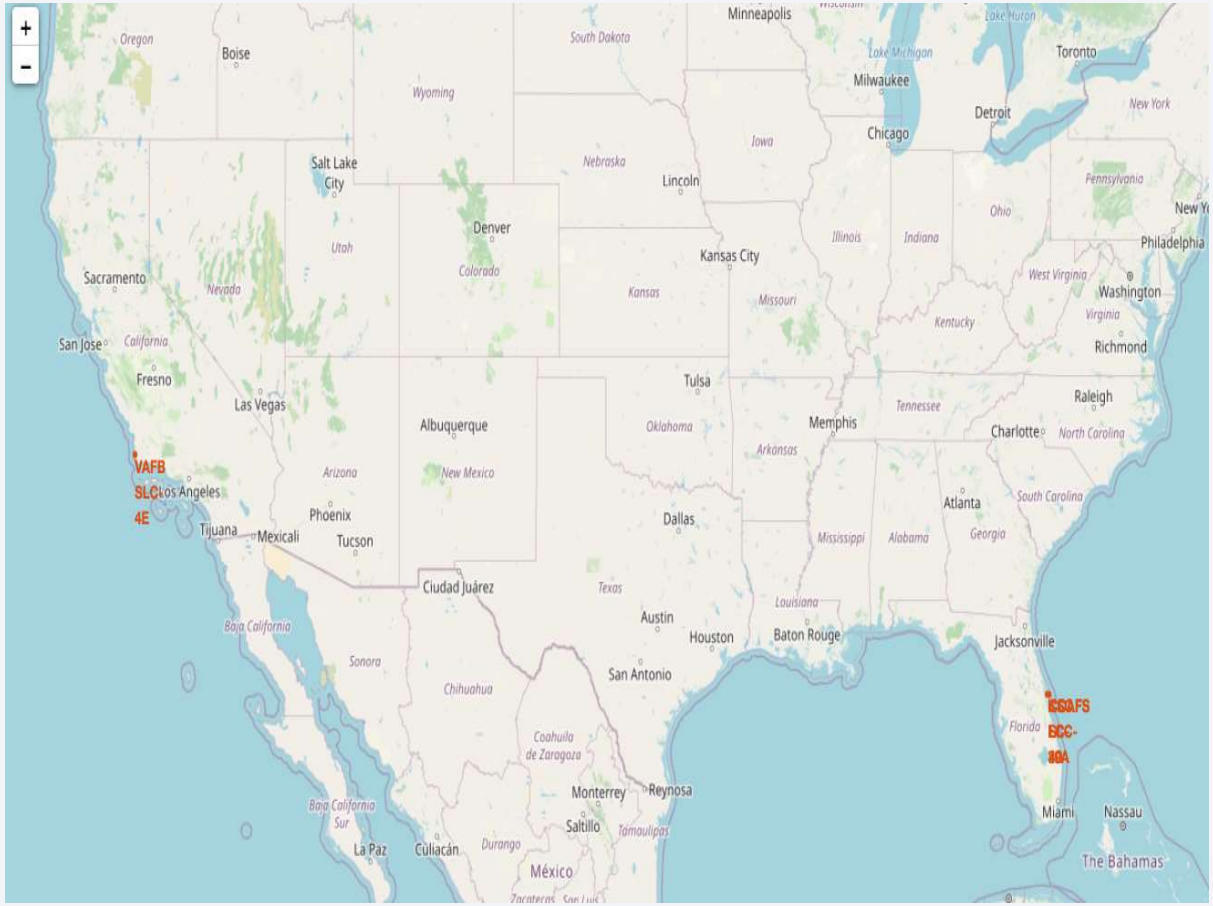
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Ranking the landing outcomes (e.g., Failure on drone ship or Success on ground pad) between June 4, 2010, and March 20, 2017, in descending order based on their count.

Section 3

Launch Sites Proximities Analysis

Displaying the location markers for all launch sites on a global map.



Most launch sites are located near the Equator, where the Earth's surface moves fastest at approximately 1670 km/h. This high rotational speed provides an initial boost to spacecraft launched from the equator, aiding them in achieving the necessary velocity to maintain orbit due to inertia. Additionally, all launch sites are situated close to coastlines, which reduces the risk of debris falling or exploding near populated areas when rockets are launched over the ocean.

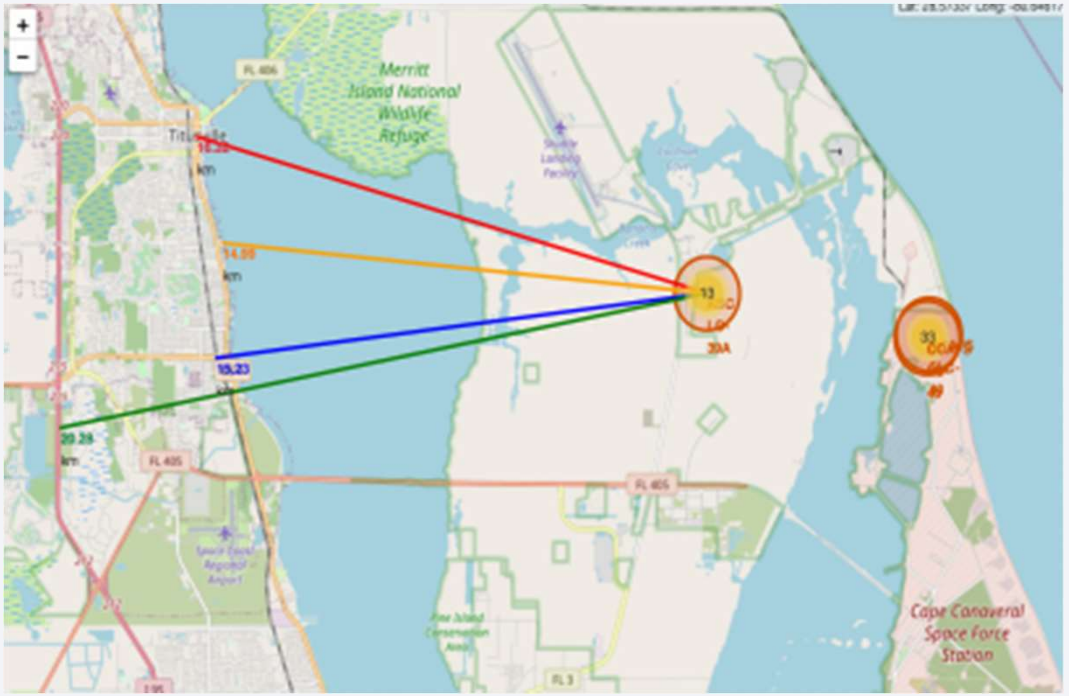
Launch records on the map are marked with different colors for easy identification.

The color-coded markers make it easy to identify launch sites with higher success rates:

- Green Marker = Successful Launch
- Red Marker = Failed Launch

The KSC LC-39A launch site shows a very high success rate.

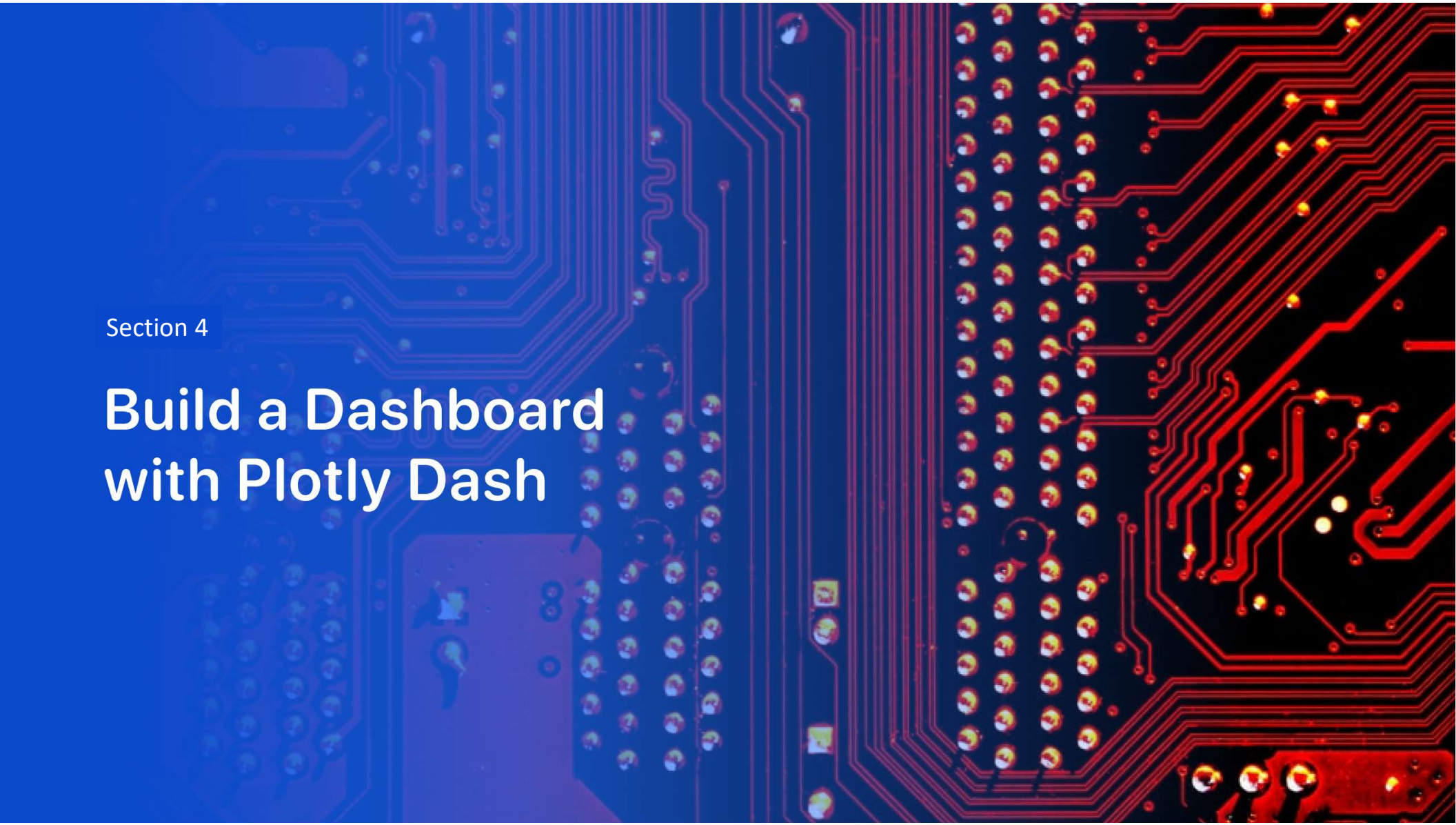
Proximity distances from the launch site KSC LC-39A to nearby locations.



The visual analysis of the KSC LC-39A launch site shows that it is:

- Approximately 15.23 km from the railway
- About 20.28 km from the highway
- Roughly 14.99 km from the coastline
- Close to the city of Titusville, at a distance of 16.32 km

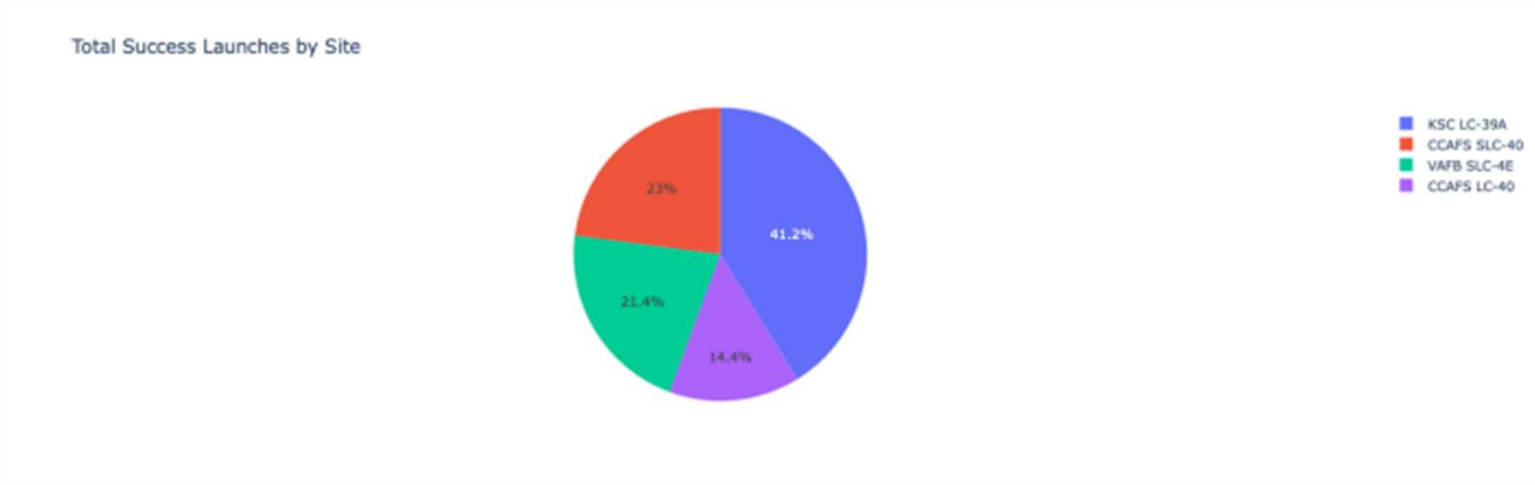
Given that failed rockets can travel distances of 15-20 km in a matter of seconds, this proximity poses a potential risk to nearby populated areas.



Section 4

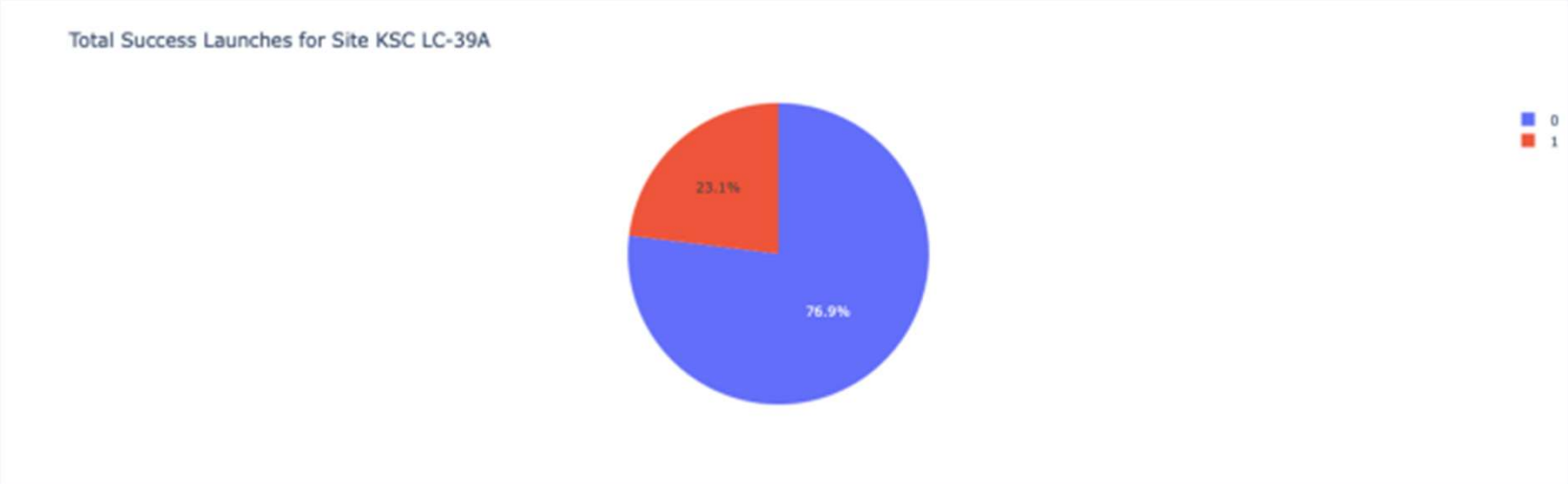
Build a Dashboard with Plotly Dash

Total count of successful launches for all sites.



The chart clearly indicates that KSC LC-39A has the highest number of successful launches among all the sites.

Launch site with highest launch success ratio



KSC LC-39A boasts the highest launch success rate at 76.9%, with 10 successful landings and just 3 failures.

Payload Mass vs. Launch Outcome for all sites

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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and Accuracy of the Test Set

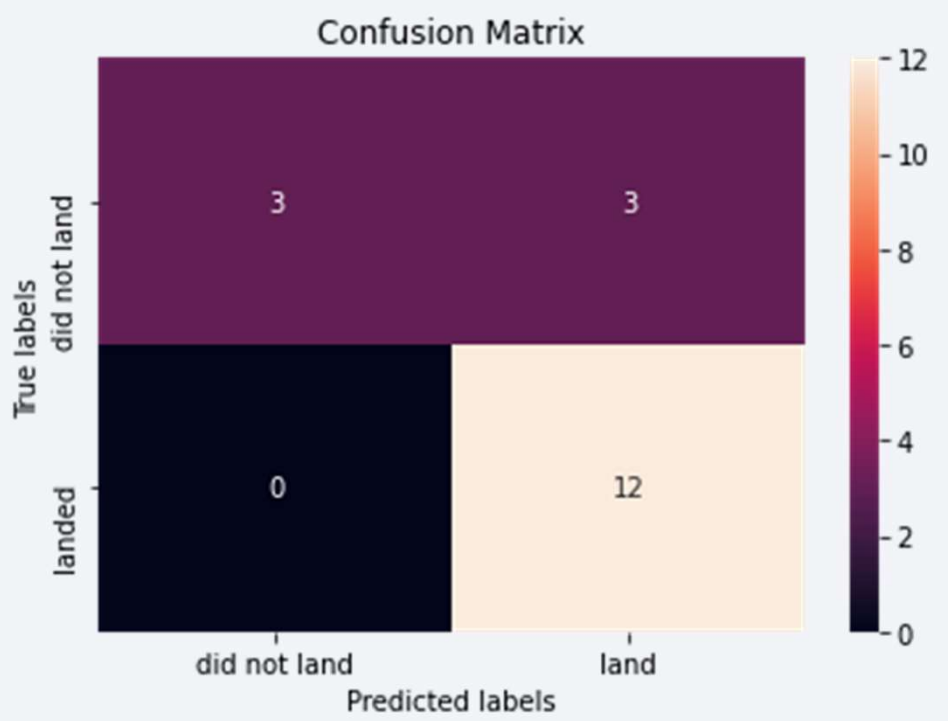
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

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

- The Test Set scores do not provide a clear indication of the best-performing method.
- The similarity in Test Set scores might be attributed to the small sample size (18 samples). For a more comprehensive evaluation, we tested all methods using the entire dataset.
- Analysis of the full dataset confirms that the Decision Tree Model is the most effective. This model not only achieved higher scores but also demonstrated the highest accuracy.

Confusion Matrix



An analysis of the confusion matrix indicates that logistic regression is capable of distinguishing between the different classes. However, the main issue is the presence of false positives.

Conclusions

- The Decision Tree Model is the most effective algorithm for this dataset.
- Launches with smaller payload masses tend to show better performance compared to those with larger payloads.
- Most launch sites are located near the Equator, and all are positioned close to the coast.
- The success rate of launches has shown an upward trend over the years.
- KSC LC-39A has the highest launch success rate among all sites.
- The orbits ES-L1, GEO, HEO, and SSO all have a 100% success rate.

Appendix

- Git Hub Link: https://github.com/JairCamargo02/Data-Science-Capstone-Project-Final-Paper/issues

Thank you!