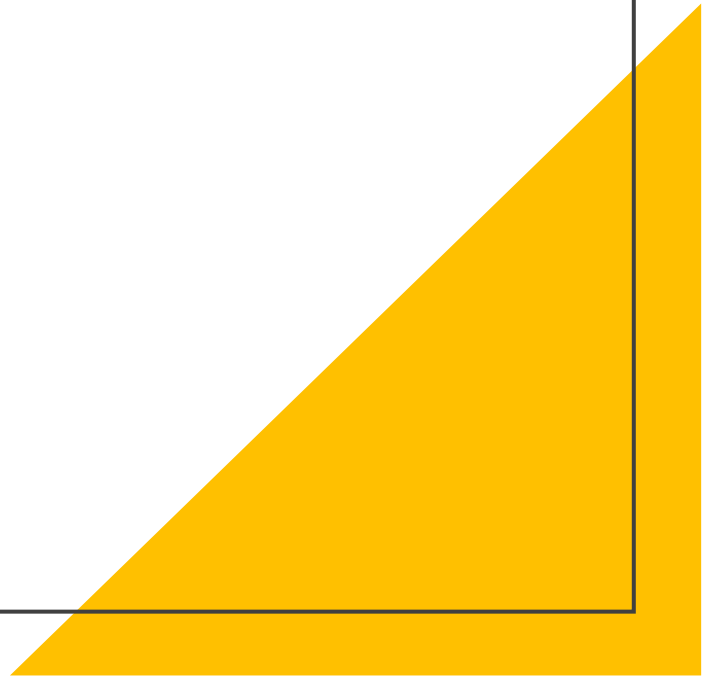


IBM Data Science Capstone

Stephen Mukuze

07/06/2024



Contents

- Overview
- Introduction
- Methods
- Results
- Conclusion

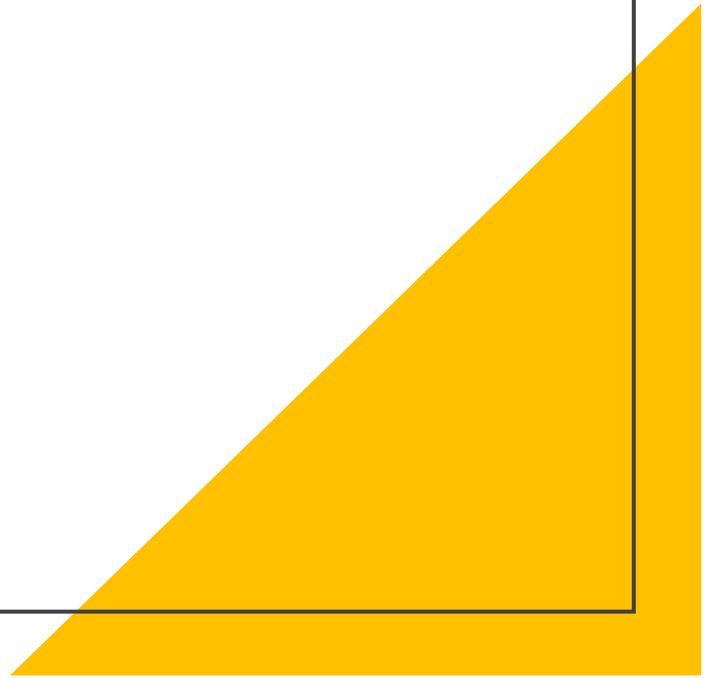
Overview

Methods

- 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

Results

- Results: exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Introduction

Project Focus:

SpaceX has revolutionized the space industry by drastically reducing launch costs, largely attributed to their reusable Falcon 9 first stage. This project aims to develop a predictive model that determines the likelihood of a successful first stage landing based on various mission parameters. This information can offer valuable insights into launch cost estimations.

Key Research Questions:

- Impact of Mission Variables: How do factors like payload mass, launch site selection, the rocket's flight history, and target orbit influence the success rate of first stage landings?
- Landing Success Trend Analysis: Has SpaceX improved its first stage recovery rate over time? Does the data reveal a trend of increasing landing success?
- Optimal Algorithm Selection: Which machine learning algorithm is best suited for predicting first stage landing success, given the binary nature of the outcome (success or failure)?

Project Significance:

- Cost Estimation Refinement: Provide more precise cost estimates for future SpaceX launches based on the likelihood of first stage reusability.
- Understanding Landing Dynamics: Enhance our understanding of the factors influencing successful Falcon 9 first stage landings.
- Benchmarking SpaceX's Progress: Track SpaceX's advancements in reusable rocket technology over time.

Methods: Data collection

- **To construct a comprehensive dataset for analyzing SpaceX launches**, we employed a two-pronged data collection strategy. We utilized API requests to directly access SpaceX's launch records via their REST API. This provided structured data on key mission parameters, including FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- **To supplement the API data and ensure completeness**, we implemented web scraping techniques on SpaceX's Wikipedia entry. This allowed us to capture additional information presented in tabular format, such as Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time. By combining data from both the SpaceX API and Wikipedia, we aimed to overcome potential limitations of using a single source, ensuring a more robust and comprehensive dataset for our analysis.
- This approach allowed us to capture a wider range of variables and mitigate potential biases or gaps in individual data sources.

[GitHub link to Data Collection API](#)

A large yellow triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

Methods: Web scrapping

For web scrapping, the following procedures were followed:

- **Data Retrieval:** We began by fetching the HTML content of the relevant Wikipedia page containing the Falcon 9 launch table.
- **HTML Parsing:** Using BeautifulSoup, we transformed the raw HTML into a structured, parsable format.
- **Column Header Extraction:** We extracted the table headers to define the columns (data fields) for our dataset.
- **Data Collection:** We systematically collected data from each row and cell within the HTML table, associating it with its corresponding column header.
- **Dictionary Construction:** The extracted data was organized into a dictionary, using column headers as keys and their corresponding data values.
- **Data Frame Creation:** This dictionary was then converted into a structured data frame to facilitate data manipulation and analysis.
- **CSV Export:** Finally, we exported the assembled data frame as a CSV file for convenient storage, sharing, and analysis in data-oriented tools.

[GitHub link to Web Scrapping](#)

A large yellow triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

Methods: Data wrangling

For data wrangling, the following procedures were followed:

- **Exploratory Data Analysis (EDA) and Target Label Definition:** We initiated our process with EDA to understand the dataset's characteristics and identify potential predictors for our target variable—landing outcome. Based on these initial findings, we defined the training labels for our predictive model.
- **Launch Site Frequency Analysis:** We calculated the frequency of launches from each unique launch site, revealing potential relationships between launch location and mission success.
- **Orbit Type Analysis:** Similarly, we analyzed the distribution of missions across different orbit types, identifying the most frequently targeted orbits.
- **Orbit-Specific Outcome Analysis:** To uncover potential correlations, we examined mission outcomes in relation to specific orbit types, looking for patterns that might influence landing success.
- **Landing Outcome Label Engineering:** We transformed the existing "Outcome" column into a dedicated "Landing Outcome" label, providing a clear and concise binary target variable for our predictive models.
- **Data Preservation:** Finally, we saved the processed dataset as a CSV file, preserving the results of our data wrangling efforts for subsequent analysis.

[GitHub link to Data Wrangling](#)

Methods: Data visualisation

For the EDA with data visualization, the following were used:

- Scatter Plots: Used to explore the relationship between two numerical variables. A visible pattern in the scatter plot might suggest a relationship suitable for machine learning modeling.
- Bar Charts: Effectively compare a measured value across different categories. For example, a bar chart could compare the success rate of launches from various launch sites.
- Line Charts: Ideal for showing how a variable changes over time. This could include tracking the yearly trend of successful launches.

[GitHub link to Data Visualisation](#)

A large yellow triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

Methods: SQL queries

Space Mission Data Analysis with SQL:

- Launch Site Identification:
- Identified and listed all unique launch sites used for space missions.
- Launch Site Filtering:
- Extracted records of five launches specifically from launch sites starting with "CCA".
- Payload Capacity Analysis:
- Calculated the total payload mass transported by boosters launched by NASA (CRS).
- Determined the average payload mass carried by boosters with the version designation "F9 v1.1".
- Mission Success Tracking:
- Found the date of the first successful landing on a ground pad.
- Identified boosters that successfully landed on drone ships and carried a payload mass between 4000 and 6000.
- Determined the total count of successful and failed missions.
- Booster Performance Evaluation:
- Identified booster versions that have carried the maximum payload mass.
- Failure Analysis:
- Extracted details of failed drone ship landings in 2015, including booster versions and launch site names.
- Landing Outcome Trends:
- Ranked the frequency of different landing outcomes.

[GitHub link to SQL Queries](#)

A large yellow triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

Methods: Folium Interactive Map

Visualizing Space Mission Launch Data on a Map:

Launch Site Mapping:

- A map was created with markers indicating the locations of all launch sites.
- NASA Johnson Space Center was marked as the starting location.
- Markers included labels with launch site names and were styled to highlight their proximity to the equator and coastlines.

Launch Outcome Visualization:

- Successful and failed launches were represented by green and red markers, respectively, using marker clusters. This visualization helped identify launch sites with higher success rates.

Proximity Analysis:

- Taking KSC LC-39A as an example, colored lines on the map were used to show the distances between the launch site and key infrastructure points like railways, highways, coastlines, and the nearest city.

[GitHub link to Folium Map](#)

Methods: Plottly dash dashboard

Interactive Dashboard for Space Mission Analysis:

Launch Site Selection: A dropdown menu allows users to focus on a specific launch site for more detailed exploration.

Success Rate Visualization: A pie chart dynamically displays:

- The overall success rate of launches from all sites.
- A breakdown of successful and failed launches for the selected launch site.

Payload Mass Filtering: A slider enables users to define a specific range of payload mass, narrowing down the data for analysis.

Payload Mass vs. Success Rate: A scatter chart reveals the relationship between payload mass and launch success rates across different booster versions. This visualization helps identify potential correlations between payload weight and mission success.

[GitHub link to Dashboard App](#)

Methods: Classification predictive analysis

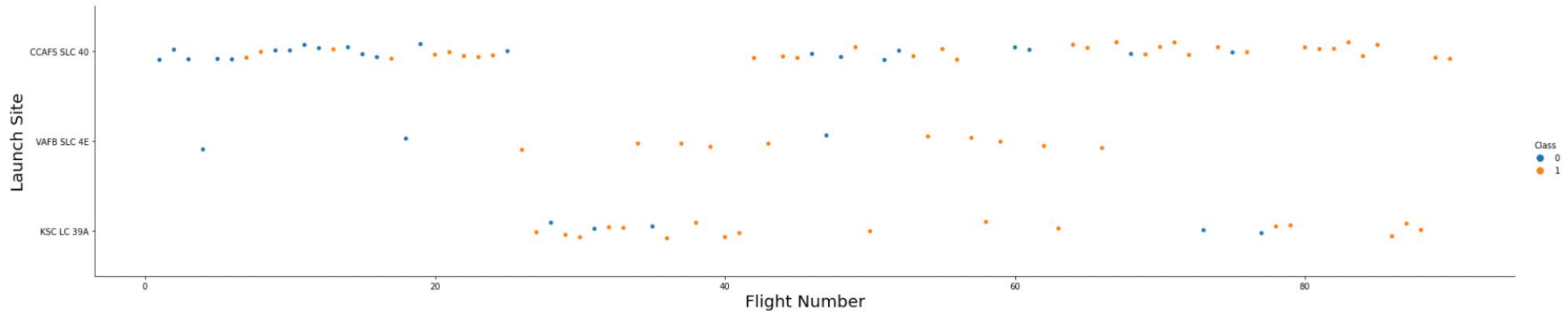
For the classification predictive analysis, the following steps were taken:

- Generating a NumPy array using the "Class" column from the dataset
- Normalizing the dataset using the StandardScaler and applying both fit and transform operations
- Dividing the dataset into subsets for training and evaluation using the train_test_split utility
- Initializing a GridSearchCV instance with 10-fold cross-validation to optimize model parameters
- Implementing GridSearchCV to evaluate the performance of Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors algorithms
- Assessing the prediction accuracy of each model on the testing subset by utilizing the .score() function
- Reviewing the confusion matrices to understand the performance of each model
- Determining the most effective algorithm by analyzing the Jaccard score and F1 score indicators

[GitHub link to Machine Learning Prediction](#)

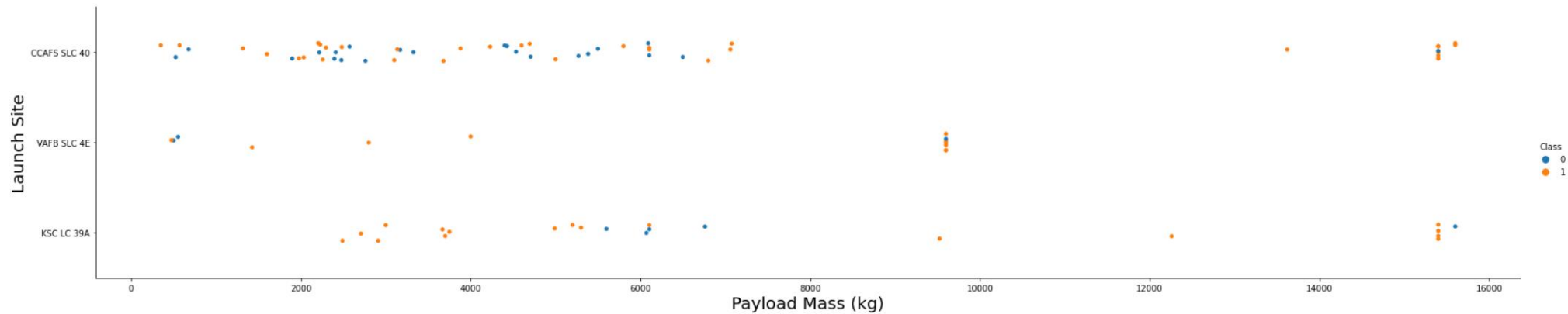
A large yellow triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

Results: Launch site v flight number



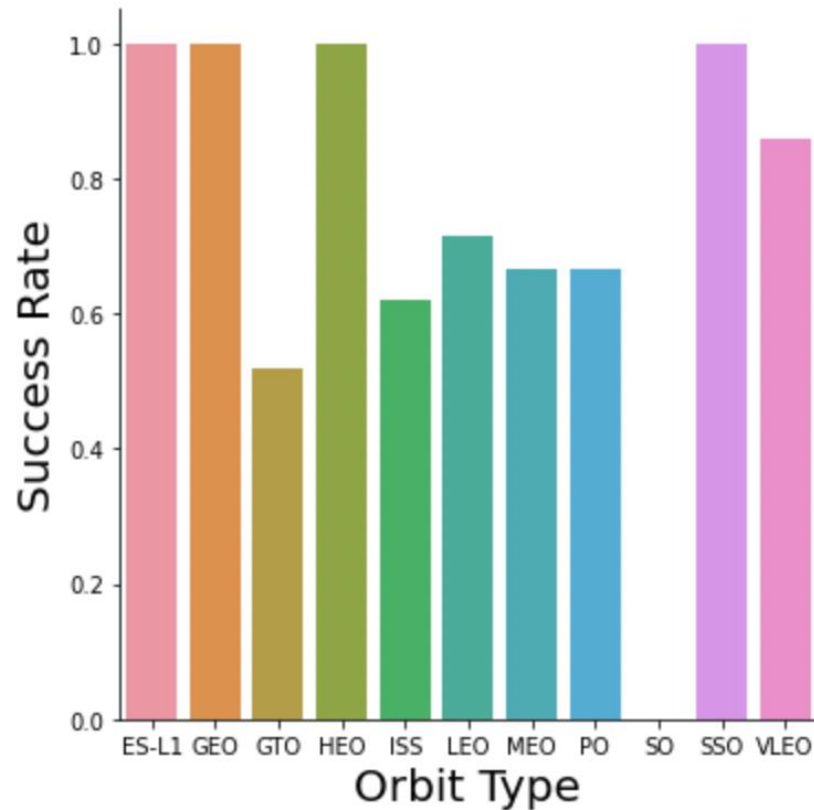
- Initial attempts at flight were unsuccessful, whereas more recent endeavors have been consistently successful.
- Approximately 50% of all launches take place at the CCAFS SLC 40 launch site.
- The launch sites VAFB SLC 4E and KSC LC 39A exhibit higher rates of successful launches.
- It is reasonable to infer that the likelihood of success increases with each subsequent launch.

Results: Launch site v pay load



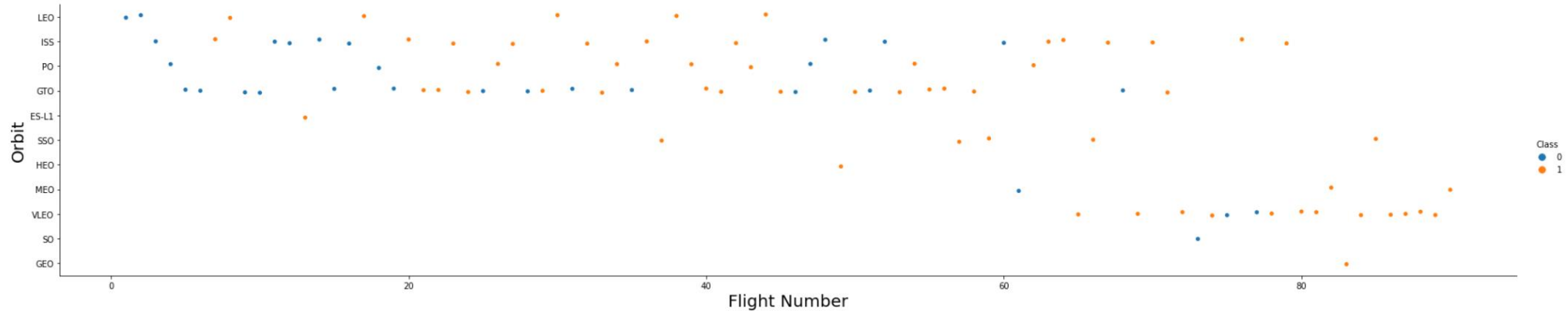
- A trend is observed at each launch site where an increase in payload mass correlates with an improved success rate.
- Launches carrying a payload exceeding 7000 kg have predominantly resulted in success.
- The KSC LC 39A launch site boasts a perfect success record for launches with payloads weighing less than 5500 kg.

Results: Launch site v payload



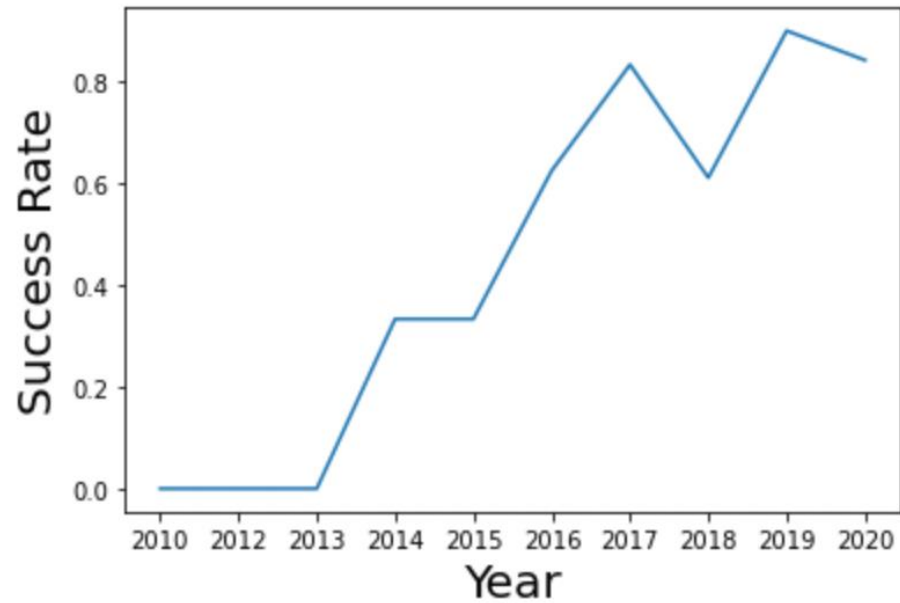
- Orbital paths achieving a complete success rate include ES-L1, GEO, HEO, SSO
- Orbital paths with no recorded successes: SO
- Orbital paths with a success rate ranging from 50% to 85%: GTO, ISS, LEO, MEO, PO

Results: flight number v orbit type



- In the Low Earth Orbit (LEO), there appears to be a correlation between the number of flights conducted and the success rate. Conversely, in the Geostationary Transfer Orbit (GTO), the number of flights does not seem to influence the success outcome.

Results: success rate



- The success rate of launches has been on an upward trajectory from the year 2013 through to 2020.

SQL EDA: launch site names

In [4]: %sql select distinct launch_site from SPACEXDATASET;

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[4]:

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

SQL EDA: CCA names

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.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

SQL EDA: total pay load 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.bs2io90108kqblod81cg.databases.appdomain.cloud:31198/bludb  
Done.
```

```
Out[6]:
```

total_payload_mass
45596

SQL EDA: ave pay load

```
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.bs2io90108kqblod81cg.databases.appdomain.cloud:31198/bludb  
Done.
```

```
Out[7]:
```

average_payload_mass
2534

SQL EDA: first successful landing

```
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.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bludb  
Done.
```

```
Out[8]:
```

first_successful_landing
2015-12-22

SQL EDA: successful drone ship landing

```
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.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

SQL EDA: total success and failure

```
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.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bludb  
Done.
```

Out[10]:

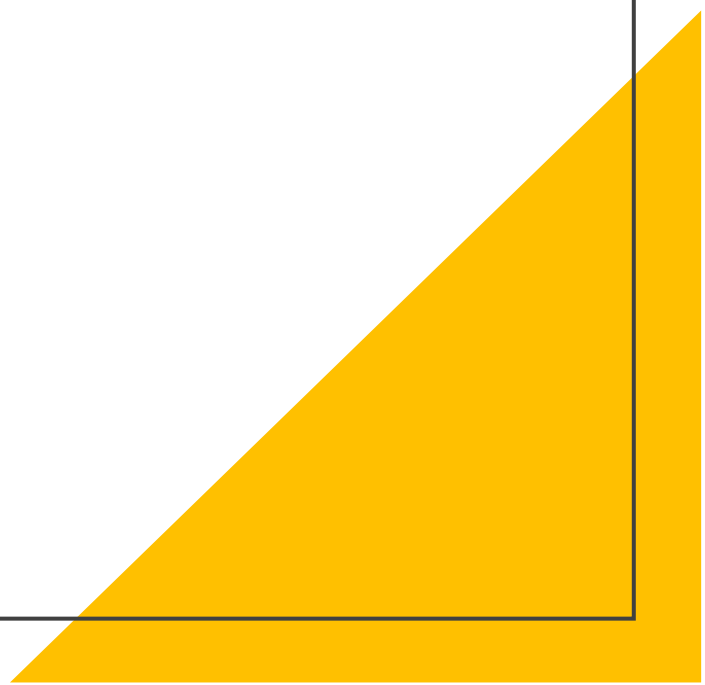
mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

SQL EDA: booster max pay load

```
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.bs2io90108kqb1od8l1cg.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



SQL EDA: 2015 launches

```
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.bs2io90108kqblod8l1cg.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)

SQL EDA: rank success count

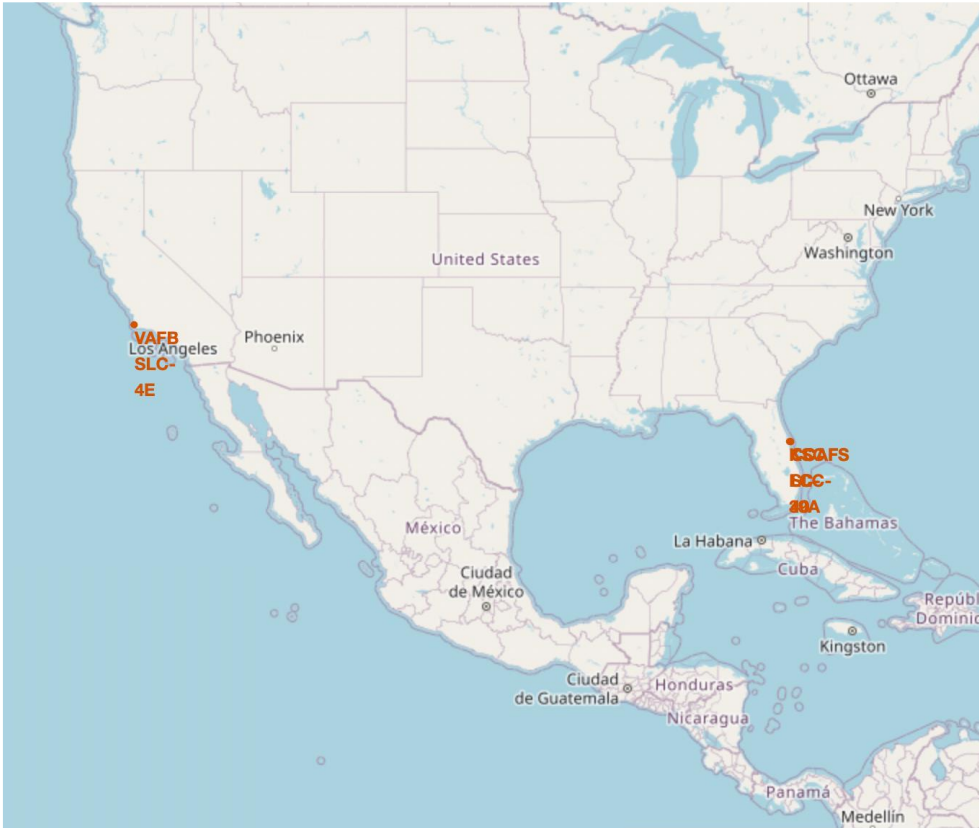
```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
        where date between '2010-06-04' and '2017-03-20'
        group by landing__outcome
        order by count_outcomes desc;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.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

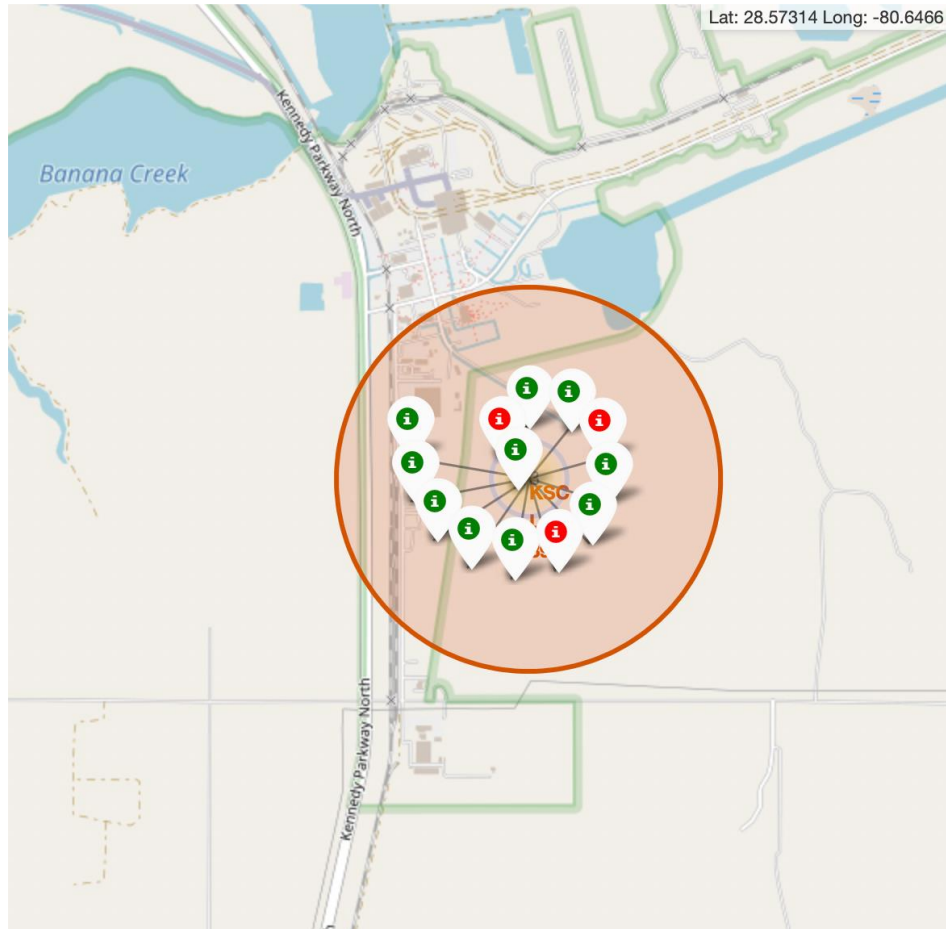
Folium map: all launch sites



Launch facilities are typically situated near the equator, where the rotational speed of the Earth is at its maximum. At the equator, the Earth's surface moves at approximately 1670 kilometers per hour. This rotational velocity is imparted to a spacecraft upon launch, aiding it in achieving the necessary speed for orbital insertion due to the principle of inertia.

Additionally, all launch sites are strategically located near coastlines. By directing rocket launches over the ocean, the risk of debris falling on populated areas or causing harm to individuals is significantly reduced.

Folium map: colour labeled launches

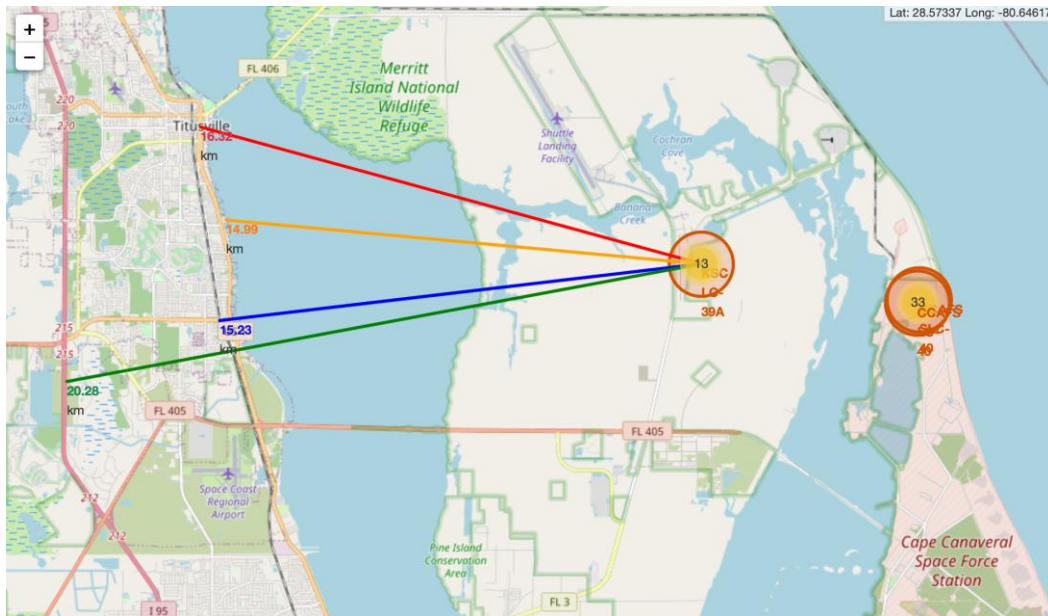


The use of color-coded markers enables quick visual identification of launch sites with varying success rates:

- A Green Marker indicates a Successful Launch.
- A Red Marker signifies a Failed Launch.

The Launch Site KSC LC-39A is noted for having a particularly high success rate.

Folium map: launch site to proximity distance



The spatial analysis of the launch site KSC LC-39A reveals its proximity to key infrastructure and geographic features:

- It is situated approximately 15.23 kilometers from the nearest railway.
- It is around 20.28 kilometers away from the closest highway.
- The site is also relatively close to the coastline, at a distance of about 14.99 kilometers.

Additionally, the launch site KSC LC-39A is located near the city of Titusville, which is 16.32 kilometers away.

Given the high velocities attained by rockets, a distance of 15-20 kilometers could be traversed in a matter of seconds in the event of a failure. This proximity poses a potential risk to nearby populated areas.

Plotly Dashboard: launch count success

Total Success Launches by Site



KSC LC 39A most successful

Plotly Dashboard: highest launch success ratio

Total Success Launches for Site KSC LC-39A



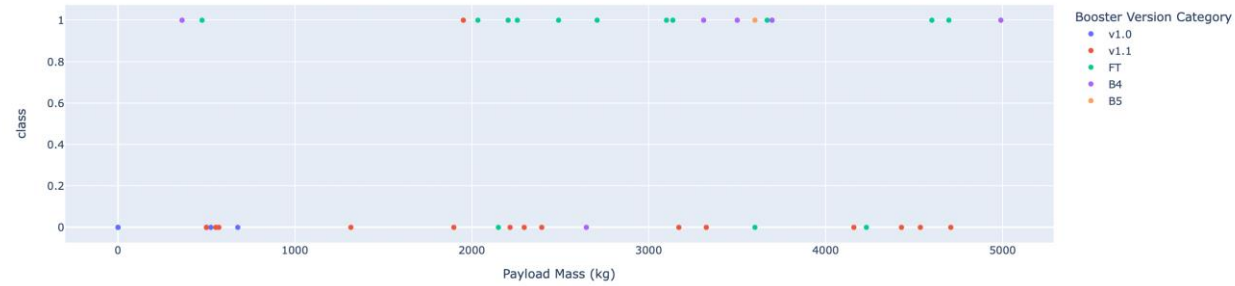
KSC LC 39A most successful at 76.9%

Plotly Dashboard: launch outcome v pay load mass

Payload range (Kg):



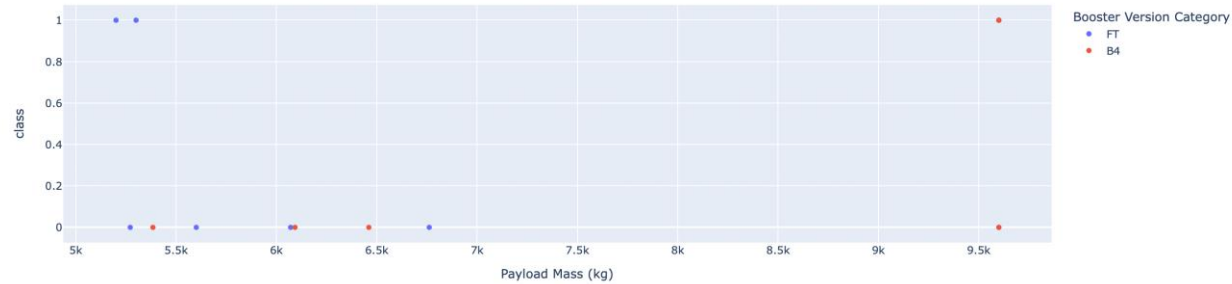
Correlation Between Payload and Success for All Sites



Payload range (Kg):



Correlation Between Payload and Success for All Sites



Classification: accuracy

	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

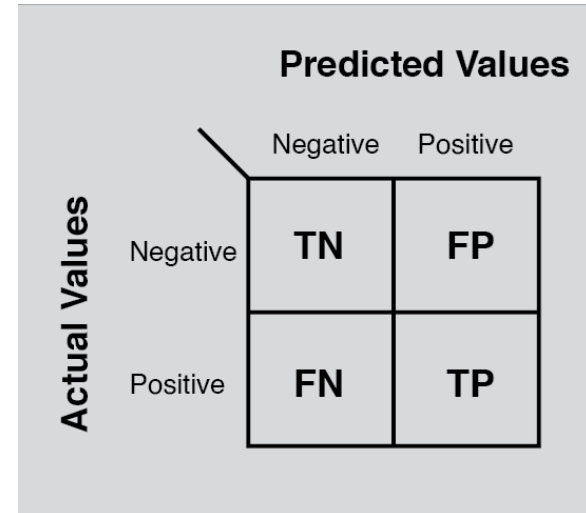
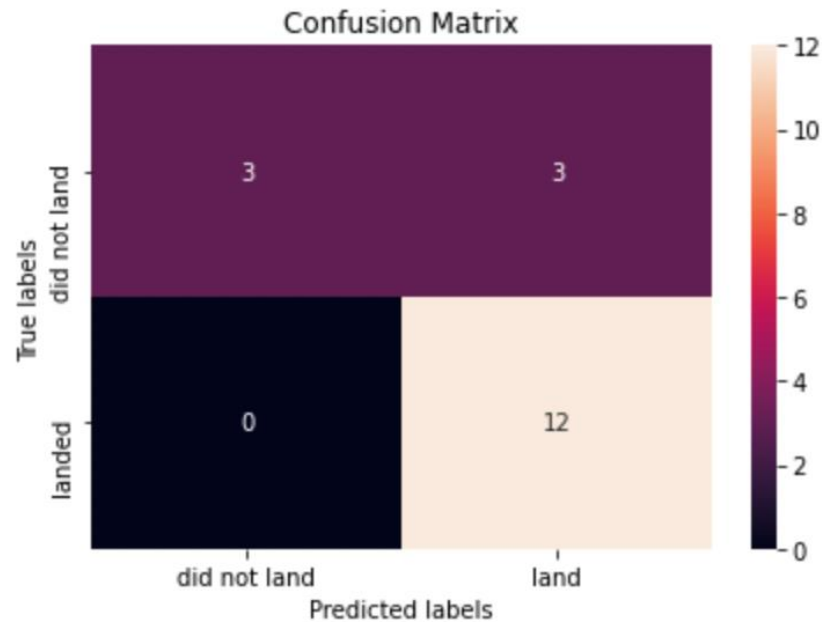
Scores and Accuracy of the Test Set

	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

Scores and Accuracy of the Entire Data Set

- The performance of various methods cannot be conclusively determined from the test set scores alone.
- The similarity in test set scores might be attributed to the limited size of the test sample, which consists of only 18 samples. Consequently, to obtain a more reliable assessment, all methods were evaluated using the entire dataset.
- Upon analyzing the scores from the full dataset, it becomes evident that the Decision Tree Model outperforms the others. This model not only achieved higher scores but also demonstrated the greatest accuracy.

Classification: confusion matrix



- The analysis of the confusion matrix indicates that the logistic regression model is capable of differentiating between the various classes. However, the primary issue identified within this model is the occurrence of false positives, where the model incorrectly predicts the positive class.

Conclusion

- The Decision Tree Model emerges as the most effective algorithm for this particular dataset.
- Launches that involve a lower payload mass tend to yield more favorable outcomes compared to those with heavier payloads.
- The majority of launch sites are strategically located near the equator, and all sites are positioned very close to coastal areas.
- There has been a noticeable improvement in the success rate of launches over time.
- Among all the launch sites, KSC LC-39A stands out with the highest success rate for its launches.
- Orbits such as ES-L1, GEO, HEO, and SSO have achieved a success rate of 100%.