Data Science Capstone Project

Prepared By:

Name : Lee Wie Jie

Date : 12/5/2025

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context

SpaceX has emerged as a leading force in commercial spaceflight, revolutionizing the industry by making space travel more cost-effective. Their Falcon 9 rockets are advertised at around $62 million per launch—significantly cheaper than the $165 million charged by other providers. This reduced cost is largely due to their ability to reuse the rocket's first stage. If we can accurately predict whether the first stage will land successfully, we can also better estimate the overall launch cost. Using publicly available data and machine learning techniques, this project aims to forecast whether SpaceX will successfully reuse the first stage of its rockets.

Key Questions Explored

- How do factors like payload weight, launch location, flight count, and orbit type influence first stage landing success?
- Has the frequency of successful landings improved over time?
- Which machine learning model is most effective for making a binary classification in this scenario?

Methodology

Data collection methodology:

- Using SpaceX REST API
- Using Web Scraping from Wikipedia

Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data for binary classification

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

- Building, tuning, and evaluation of classification models to ensure the best results

Data Collection - SpaceX API

To gather the necessary information, we used a combination of data sources: the SpaceX REST API and web scraping from a Wikipedia table related to SpaceX launches. Leveraging both methods allowed us to compile a more comprehensive dataset for in-depth analysis of rocket launches.

Columns extracted from SpaceX REST API include:

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

Columns retrieved through Wikipedia web scraping include:

- Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome
- Booster Version, Booster Landing, Date, Time

Data Collection - Scraping

The following steps were followed to extract and structure data from Wikipedia:

1. **Sent a request** to retrieve Falcon 9 launch information from the Wikipedia page
2. **Parsed the HTML content** using **BeautifulSoup to handle the** web page structure
3. **Extracted table headers** to identify all relevant column names
4. **Parsed the HTML tables** to collect launch data row by row
5. **Organized the data** into a structured dictionary format
6. **Converted the dictionary** into a pandas DataFrame for easier handling
7. **Exported the final dataset** to a CSV file for storage and further analysis

Data Wrangling

In the dataset, various outcomes describe whether the booster landed successfully or not. For instance, a **"True"** label (e.g., *True Ocean, True RTLS, True ASDS*) indicates a successful landing—whether on the ocean, a ground pad, or a drone ship. On the other hand, **"False"** labels (e.g., *False Ocean, False RTLS, False ASDS*) represent failed landing attempts at the respective locations.

To simplify model training, these detailed outcomes are converted into binary labels:

- **1** for a successful landing
- **0** for an unsuccessful landing

EDA With Data Visualisation

Visualizations Created

The following relationships were visualized through various plots:

- 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 Trends Over Time (Yearly)

Purpose of Each Chart Type

- **Scatter Plots** help identify correlations between numerical variables, which can inform feature selection for machine learning models.
- **Bar Charts** are used to compare values across different categories, highlighting how specific categories influence outcomes.
- **Line Charts** track how values change over time, revealing trends in time-series data.

EDA With SQL

SQL Queries Performed:

- Retrieved a list of all distinct launch site names used in space missions
- Selected 5 entries where the launch site name starts with **'CCA'**
- Calculated the total payload mass from launches conducted by NASA (CRS missions)
- Computed the average payload mass for booster version **F9 v1.1**
- Identified the date of the first successful landing on a ground pad
- Extracted booster names that successfully landed on drone ships and carried a payload mass between **4000 and 6000**
- Counted the number of successful and failed mission outcomes
- Found the booster version that carried the **heaviest payload**
- Queried failed drone ship landings in 2015, including booster version and launch site details
- Ranked landing outcomes (e.g., drone ship failure, ground pad success) occurring between **June 4, 2010** , and **March 20, 2017** , in descending order

Build an Interactive Map with Folium

Launch Site Mapping and Visualization

Launch Site Markers

- A marker with a circle, popup, and label was placed at NASA Johnson Space Center using its latitude and longitude as the initial reference point.
- Additional markers were added for all launch sites, using their coordinates to illustrate their geographical locations and proximity to features like the Equator and coastlines.

Colored Markers for Launch Outcomes

- Markers were color-coded to indicate launch outcomes: green for success and red for failure.
- Marker clustering was used to visualize which sites have higher concentrations of successful launches.

Distance Visualizations from Launch Sites

- Colored lines were drawn to represent the distances between Launch Site KSC LC-39A (as an example) and nearby locations such as the nearest railway, highway, coastline, and city.

Build a Dashboard with Plotly Dash

Launch Site Dropdown Menu

- Implemented a dropdown component to allow users to choose and view data for specific launch sites.

Success Launches Pie Chart (All Sites or Specific Site)

- Designed a pie chart to visualize the number of successful launches across all sites or for a selected launch site. It also displays the success-to-failure ratio when a specific site is selected.

Payload Mass Range Slider

- Added an interactive slider to allow users to filter and analyze data within a selected payload mass range.

Scatter Plot: Payload Mass vs. Success Rate by Booster Version

- Included a scatter plot to explore the relationship between payload mass and launch success, categorized by different booster versions.

Predictive Analysis (Classification)

Model Training and Evaluation Workflow

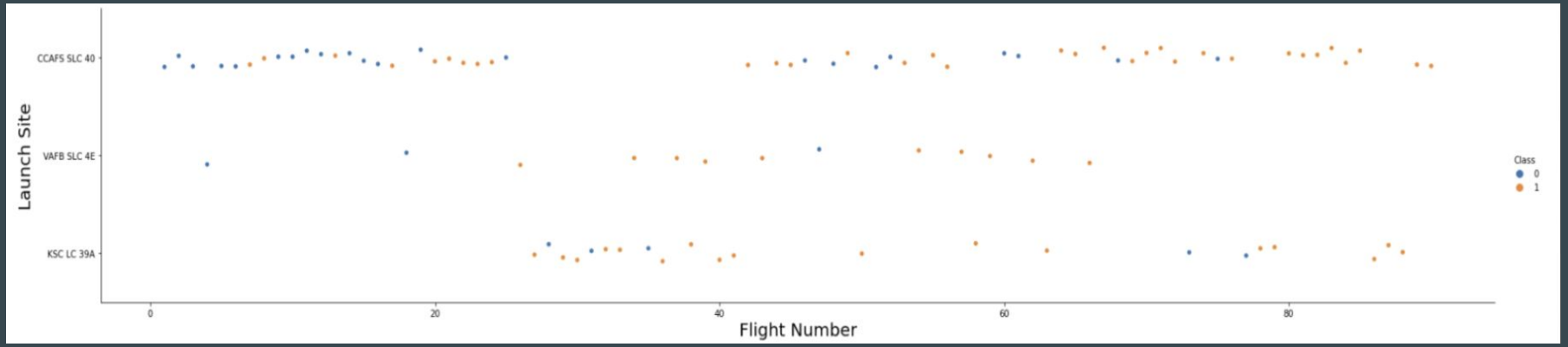
1. **Creating a NumPy array** from the column **“Class”** in the dataset
2. **Standardizing the data** using `StandardScaler`, followed by fitting and transforming
3. **Splitting the data** into training and testing sets using the `train_test_split` function
4. **Creating a `GridSearchCV` object** with `cv=10` to determine the optimal hyperparameters
5. **Applying `GridSearchCV`** to Logistic Regression, SVM, Decision Tree, and KNN models
6. **Calculating the model accuracy** on test data using the `.score()` method
7. **Examining the confusion matrix** for all models
8. **Identifying the best-performing model** based on `Jaccard_score` and `F1_score` metrics

Results

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

EDA with Visualization

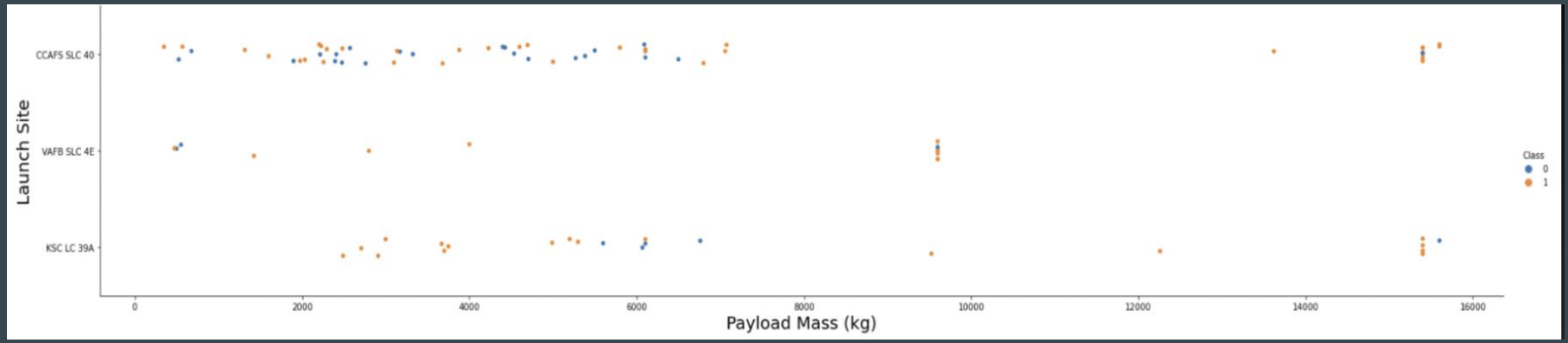
Flight Number vs Launch Site



Interpretation of Findings:

- Initial Falcon 9 missions experienced a high failure rate, whereas more recent launches have consistently succeeded.
- Approximately 50% of all launches were conducted from the **CCAFS SLC 40** site.
- Launch sites **VAFB SLC 4E** and **KSC LC 39A** demonstrated comparatively higher success rates.
- The data suggests a positive trend in launch reliability, with newer missions showing an increased likelihood of success.

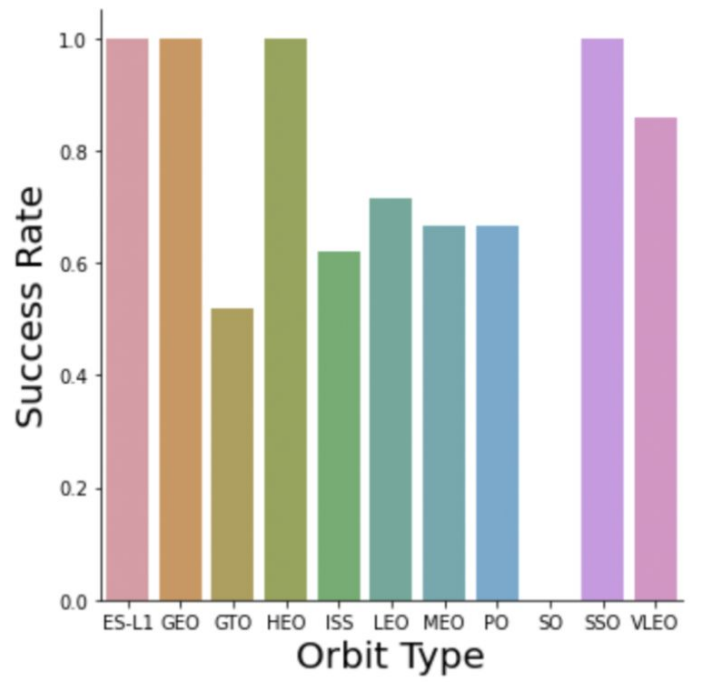
Payload vs Launch Site



Insights Based on Payload Mass and Launch Success

- Across all launch sites, a positive correlation was observed: higher payload mass tends to be associated with higher success rates.
- Launches carrying payloads exceeding **7000 kg** were predominantly successful.
- The **KSC LC 39A** site demonstrated a **100% success rate** for launches with payload masses below **5500 kg**.

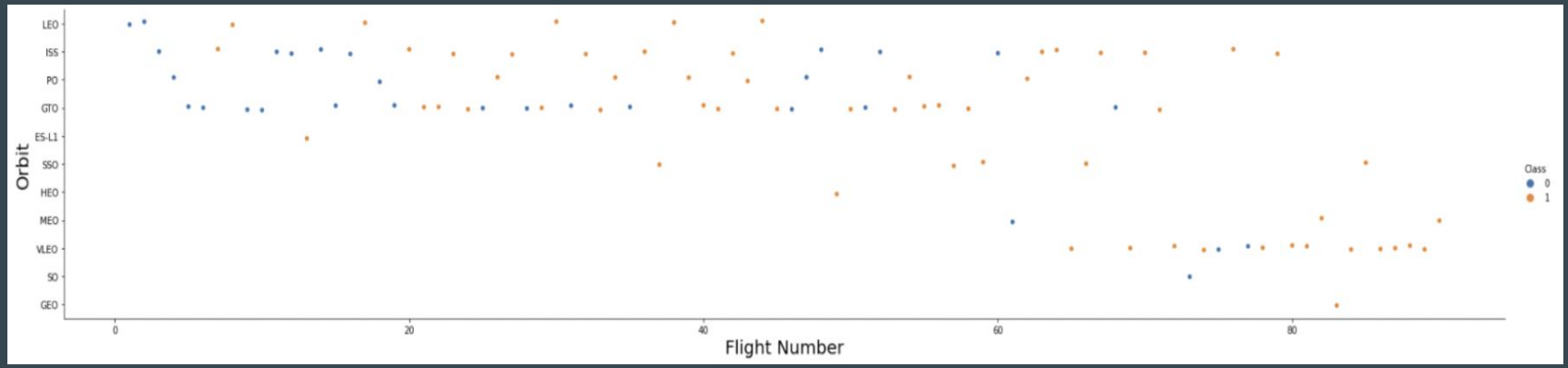
Success Rate vs Orbit Type



Success Rates by Orbit Type

- **Orbits with 100% Launch Success:**
  ES-L1, GEO, HEO, and SSO missions achieved complete success across all attempts.
- **Orbits with 0% Launch Success:**
  The SO orbit experienced no successful launches.
- **Orbits with Moderate Success Rates (50%–85%):**
  GTO, ISS, LEO, MEO, and PO demonstrated partial success within this range.

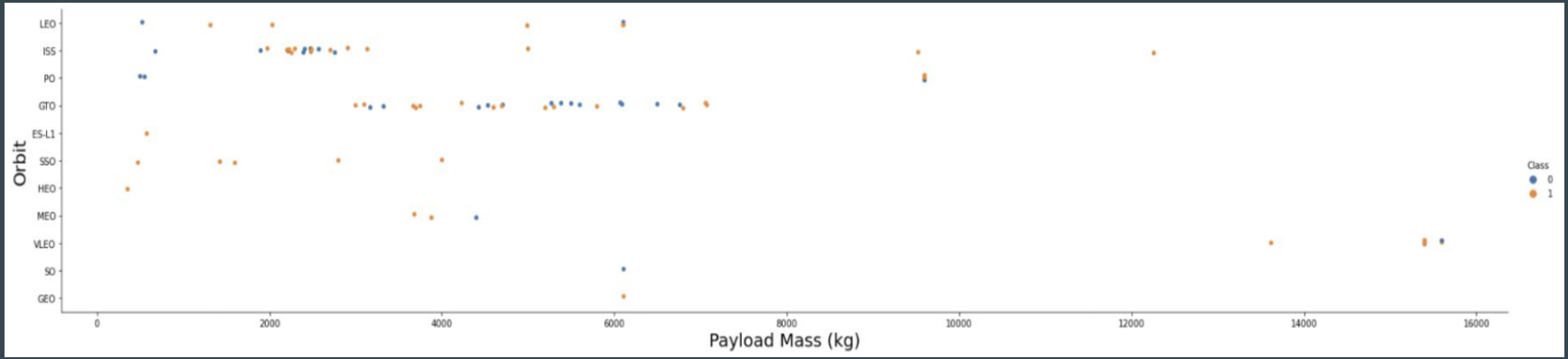
Flight Number vs Orbit Type



Orbit-Specific Observation

- In **LEO (Low Earth Orbit)** , there is a noticeable trend where launch success improves with the number of flights, indicating a positive correlation.
- However, in the case of **GTO (Geostationary Transfer Orbit)** , no such correlation between flight number and success rate is observed.

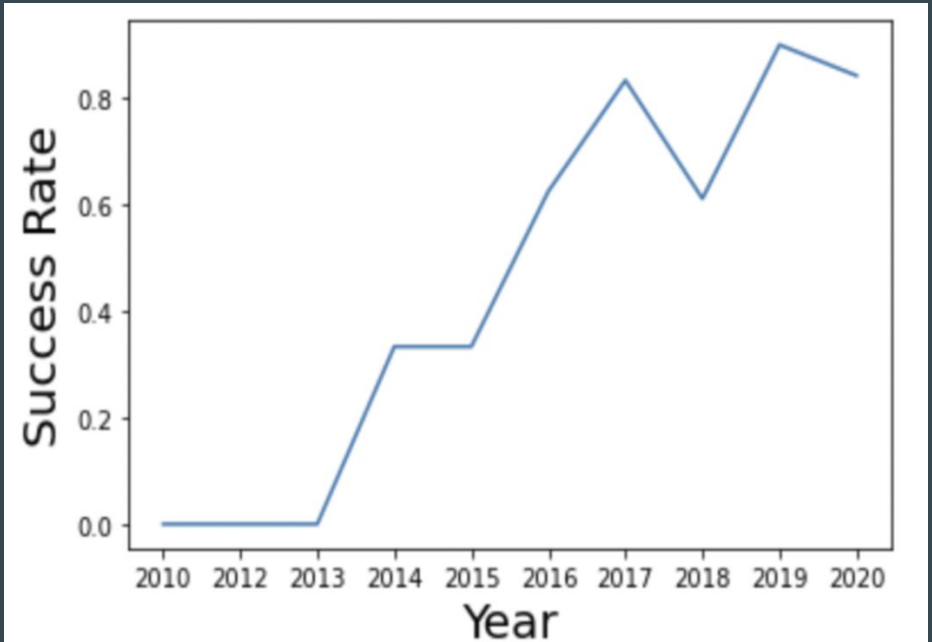
Payload Mass vs Orbit Type



Impact of Payload Mass on Orbital Success

- Heavier payloads tend to negatively affect launch success in **GTO orbits** .
- In contrast, heavy payloads appear to have a **positive impact** on success rates in **Polar LEO orbits** , including missions to the **International Space Station (ISS)** .

Launch Success Yearly Trend



Trend in Launch Success Over Time

- The launch success rate showed a consistent upward trend from **2013 through 2020** , indicating continuous improvement in mission outcomes during this period.

EDA with SQL

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

- Extracted and presented the list of all distinct launch site names involved in the space missions.

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

- Listed all unique launch site names recorded in the dataset for the space missions.

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 carried by boosters from NASA

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.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

Explanation

- Displaying the average payload mass by 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

- Displaying the date when the fiest successful landing outcome in the ground pad was achieved

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 4
        000 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:

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

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:

- Summarized the total count of mission outcomes, including successes, failures, and partially successful missions with unclear payload status.

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:

- Identified and displayed the booster versions that successfully delivered the highest recorded payload mass in the dataset.

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

- Retrieved details of failed drone ship landings during 2015, including the month, date, booster version, and corresponding launch site.

Rank success count 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]:

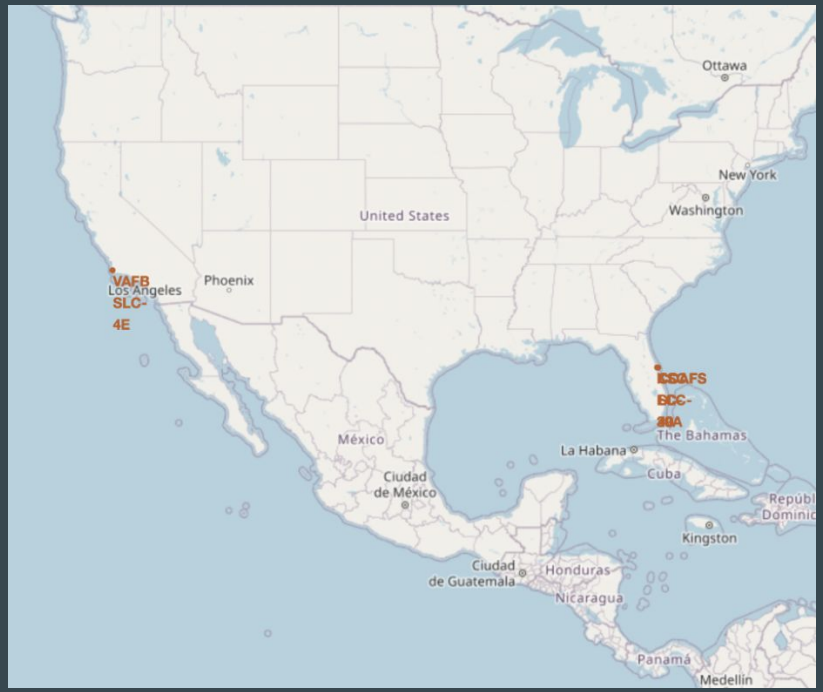
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Explanation:

- Ranked the frequency of various landing outcomes between **June 4, 2010** , and **March 20, 2017** , in descending order to identify the most common types of results during this period.

Interactive Map with Folium

All launch sites' location markers on a global map

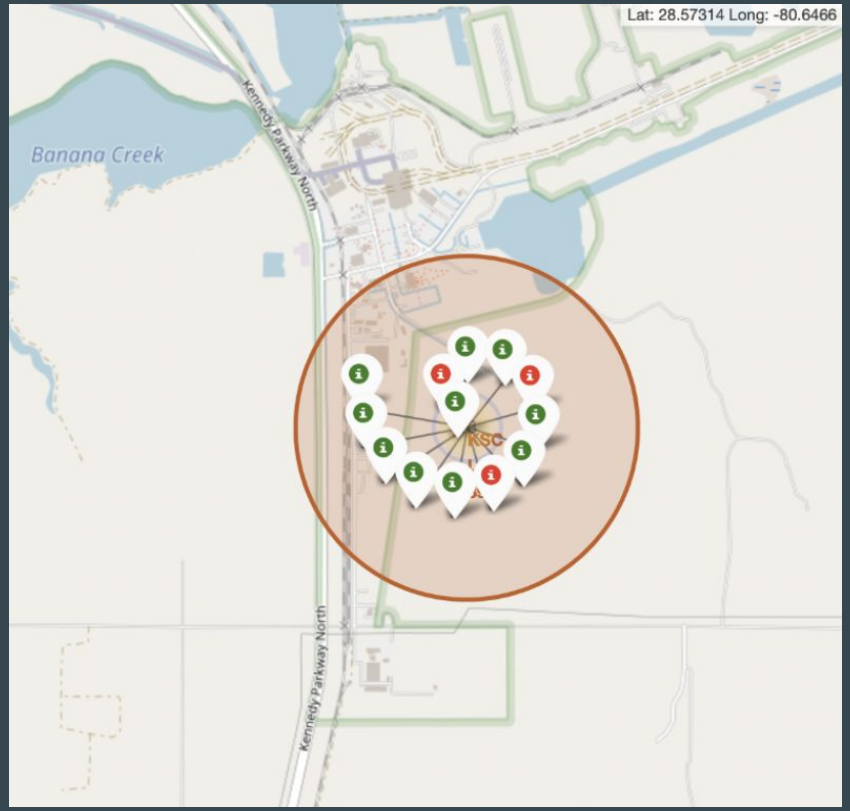


Explanation:

- Most launch sites are near the Equator. Land at the Equator rotates faster than at other latitudes—about 1670 km/hour. When a spacecraft launches from this region, it already carries that velocity, aiding its journey into orbit. This effect, due to inertia, helps rockets maintain enough speed to remain in space.

- Additionally, launch sites are typically near coastlines. Launching over oceans reduces risk, minimizing the chance of debris falling or explosions occurring in populated areas.

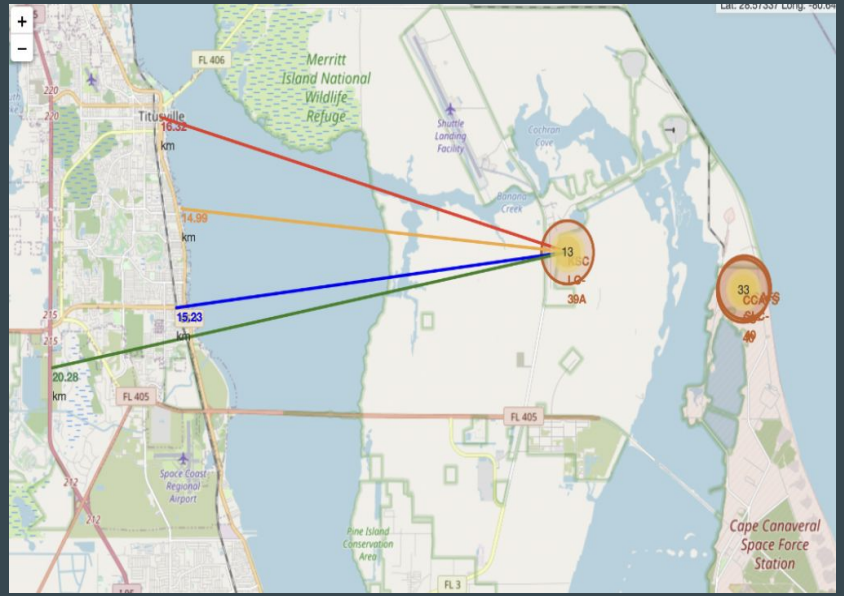
Colour-labeled launch records on the map



Explanation:

- Colour-coded markers help quickly identify launch sites with higher or lower success rates.
- **Green Marker** = Successful Launch
- **Red Marker** = Failed Launch
- Launch site **KSC LC-39A** demonstrates a very high success rate.

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

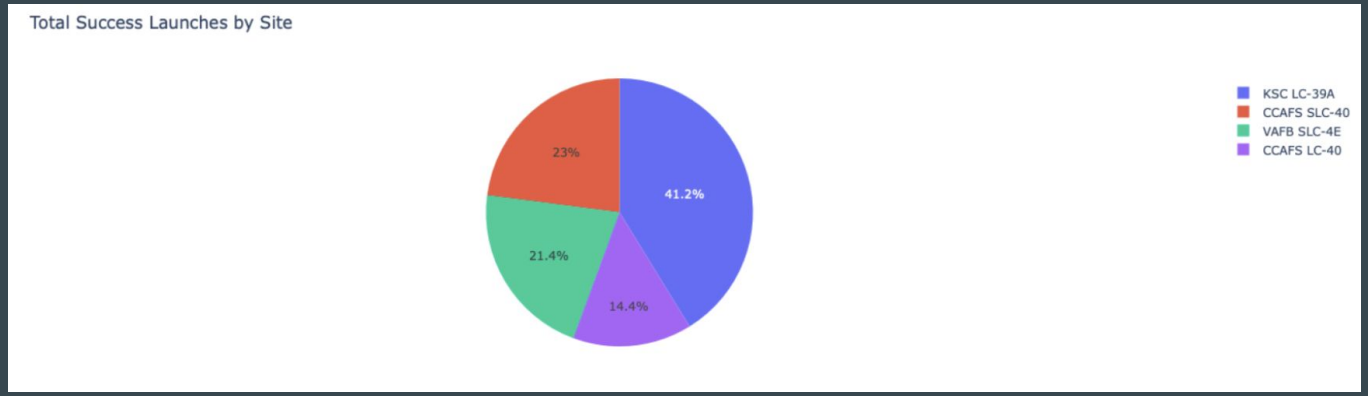


Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
    - relative close to railway (15.23 km)
    - relative close to highway (20.28 km)
    - relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative 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.

Build a Dashboard with Plotly Dash

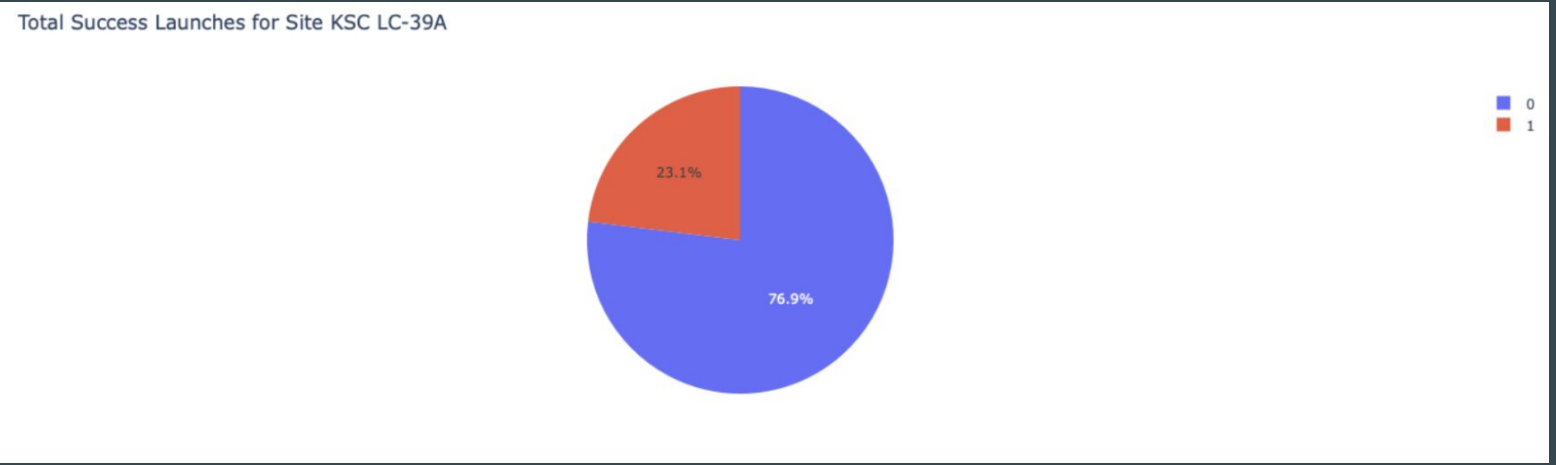
Launch success count for all sites



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.

Payload Mass vs. Launch Outcome for all sites

Explanation:

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

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- Based on the scores of the Test Set, we cannot 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.

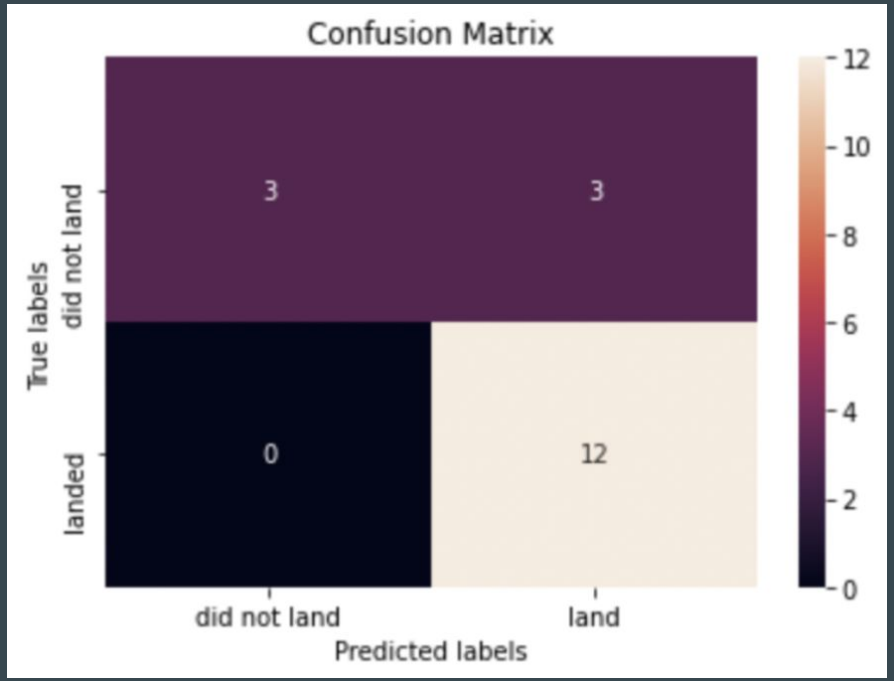
Score & Accuracy of the Test Set

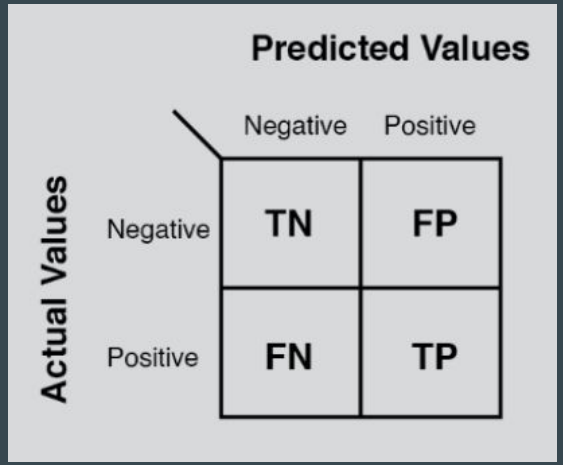
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Score & 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:

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

Conclusion

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

Thank You Very Much