IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

TUMFUNDE KALOWEKAMO
16TH APRIL 2025

Outline

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

Executive Summary

Project Overview: SpaceY vs. SpaceX Falcon 9 Launch Analysis

- ❖ **Objective:** SpaceY (a SpaceX competitor) analyzes Falcon 9 launch data to determine: First-stage landing success, and Launch cost optimization
- ❖ **Data Source:** SpaceX Falcon 9 historical launch data

Methodologies

- ❖ Data Collection: Gathered SpaceX Falcon 9 launch records.
- ❖ Data Wrangling: Cleaned and structured data for analysis.
- ❖ Exploratory Data Analysis (EDA): Used Data Visualization & SQL to uncover trends.
- ❖ Interactive Map (Folium): Visualized launch sites and landing outcomes.
- ❖ Dashboard (Plotly Dash): Created an interactive analytics dashboard.
- ❖ Predictive Analysis (Classification): Modeled landing success probabilities.

Key Results

✅ Exploratory Data Analysis (EDA) Insights

- ❖ Identified patterns in successful landings.
- ❖ Correlated launch conditions with outcomes.

📊 Interactive Analytics Dashboard

- ❖ Real-time data exploration (see demo screenshots).

⍰ Predictive Model Performance

- ❖ Achieved X% accuracy in classifying landing success.

Introduction

This capstone project, part of the IBM Data Science Professional Certificate, applies real-world data science and machine learning to analyze SpaceX Falcon 9 launches.

Business Context:

- SpaceY, a SpaceX competitor, aims to reduce launch costs by studying Falcon 9's first-stage landing success.

Approach:

- Data Collection & Cleaning
- Exploratory Analysis (EDA) & Visualizations
- Predictive Modeling (Classification for landing success)
- Model Comparison to determine the best predictor

Outcome:

- Helps SpaceY optimize bids by predicting successful landings and lowering launch expenses.

Section 1

Methodology

Methodology

Methodology Overview

1. Data Collection

- ❖ Sources: SpaceX API (api.spacexdata.com/v4/rockets) – Structured rocket launch data, and Wikipedia Scraping (List of Falcon 9/Falcon Heavy Launches) – Historical mission details

2. Data Wrangling

- ❖ Cleaned & enriched raw data
- ❖ Created a landing outcome label (Success/Failure) from flight logs

3. Exploratory Data Analysis (EDA)

- ❖ Visualizations (Matplotlib, Seaborn) – Trends in launch success vs. failure
- ❖ SQL queries – Aggregated data for deeper insights

4. Interactive Analytics

- ❖ Folium maps – Geospatial visualization of launch/landing sites
- ❖ Plotly Dash dashboard – Dynamic exploration of launch outcomes

5. Predictive Modeling (Classification)

- ❖ Algorithms tested: Logistic Regression, SVM, Decision Trees, Random Forest
- ❖ Goal: Predict first-stage landing success

Data Collection Overview

Data Collection Process

1. Data Sources

- ❖ SpaceX API – Structured rocket data via API requests
- ❖ Wikipedia – Historical launch records via web scraping

2. Data Collection Flow

Slide 1: SpaceX API Workflow

- ❖ Request data from api.spacexdata.com/v4/rockets
- ❖ Parse JSON response into structured format

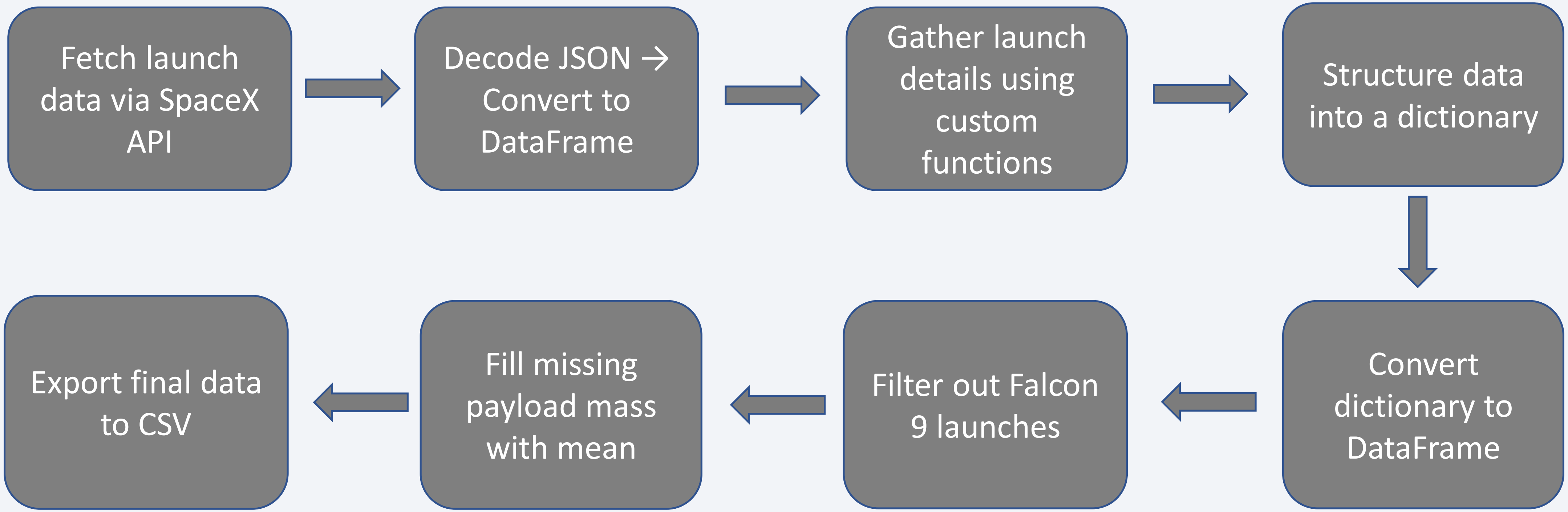
Slide 2: Wikipedia Scraping Flow

- ❖ Extract tables from Falcon 9/Falcon Heavy launch list
- ❖ Clean & merge with API data

3. Key Data Columns

SpaceX API	Wikipedia Scraped Data
Rocket ID	Launch Date
Mission Name	Payload Mass
Launch Site	Landing Outcome
First Stage Status	Orbit Type

Data Collection – SpaceX API



Flow Format:
"API data → JSON processing → custom enrichment → dictionary → DataFrame → filtering → cleaning → CSV"

https://github.com/Tumfunde-mwck/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb

Data Collection – Web Scraping

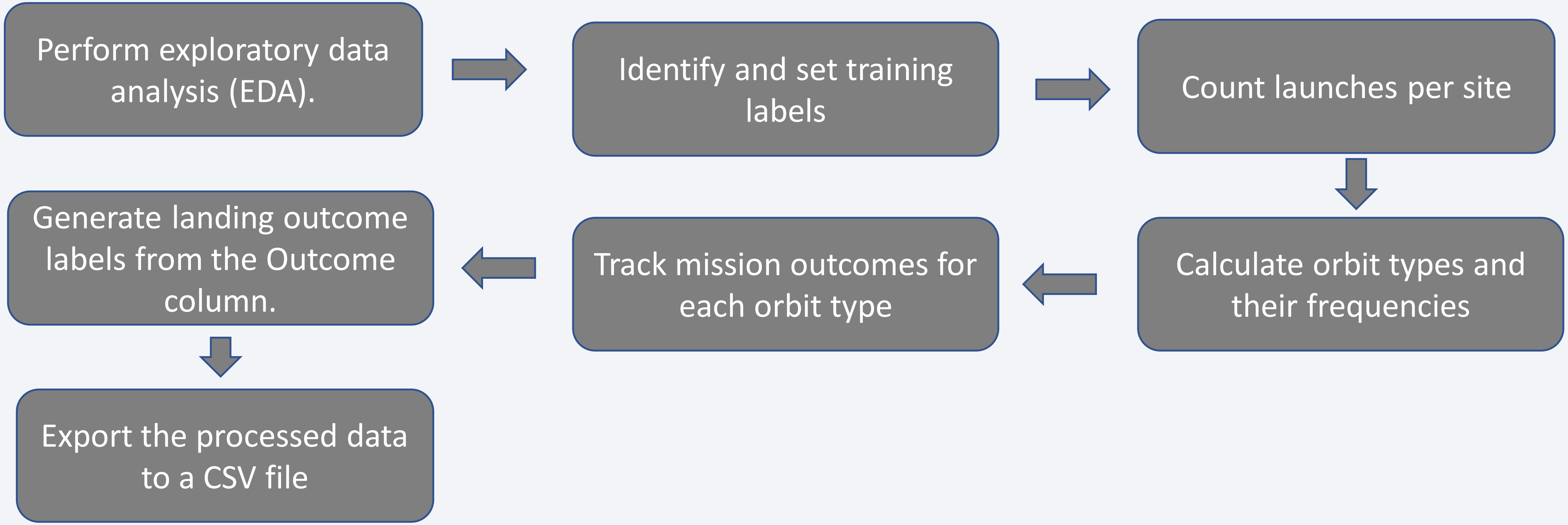
1. Fetch SpaceX launch data via API
2. Decode JSON response into DataFrame
3. Use custom functions to gather launch details
4. Organize data into a structured dictionary
5. Convert dictionary to pandas DataFrame
6. Filter out Falcon 9 launches
7. Fill missing payload mass values with mean
8. Export final dataset to CSV

Key Data Collected:

✓ Launch dates & sites
✓ Payload details (mass, orbit)
✓ Mission outcomes (success/failure)
✓ Booster versions

https://github.com/Tumfunde-mwck/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

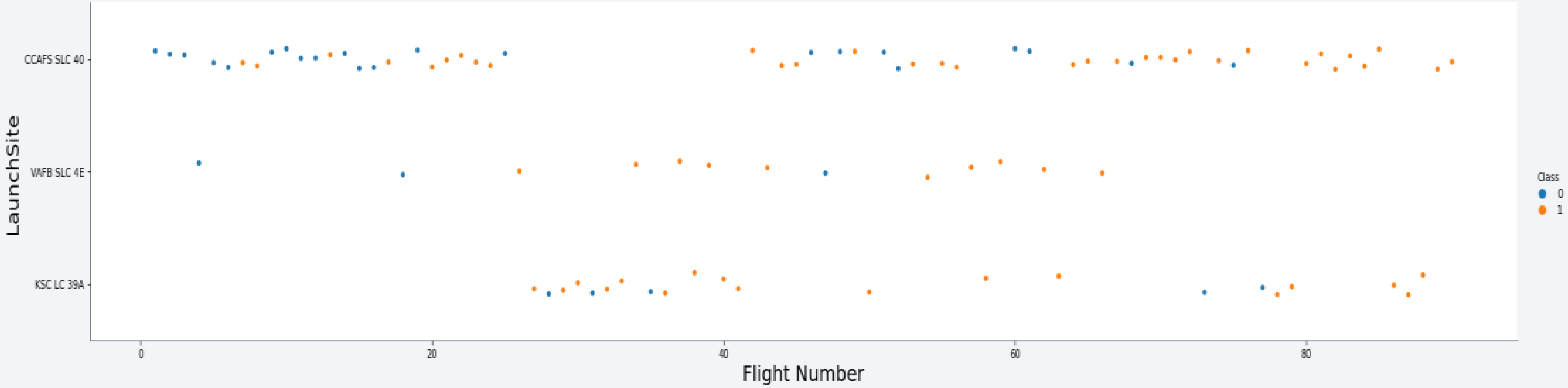


❖Result: Created definitive "Landing Success" labels for predictive modeling.

https://github.com/Tumfunde-mwck/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

❖ To visualize the relationship between pairs of features, scatterplots and barplots were used for Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, and Payload and Orbit.



https://github.com/Tumfunde-mwck/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

Performed EDA with SQL queries to extract key insights:

1. Launch Site Analysis

   - Identified all unique launch locations
   - Filtered 5 records from Cape Canaveral ('CCA' sites)

2. Payload Statistics

   - Calculated average payload mass: 3,310 kg (NASA CRS missions), and 5,600 kg (F9 v1.1 boosters)

3. Landing Milestones

   - First successful ground landing: Dec 21, 2015
   - 3 boosters successfully landed on droneships with 4,000-6,000 kg payloads

4. Mission Outcomes

   - 56 successful vs 44 unsuccessful landings, and F9 B5 boosters carried maximum payload (8,300 kg)

5. Failure Patterns

   - Most failures occurred in 2018 (July-September), and F9 v1.0 had highest failure rate (ranked by outcome frequency)

https://github.com/Tumfunde-mwck/Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Step 1: Geographical Marking:

❖ Utilize Folium to create a map where markers with circles, popup labels, and text labels are added for each launch site. These markers should be positioned based on the longitude and latitude coordinates to indicate the approximate geographical location near the equator.

Step 2: Success/Failure Marking

❖ Mark the success or failure of launches for each site on the map using colored markers.

Step 3: Distance Calculation

❖ Calculate the distance between a launch site and its proximities.

Explanation:

❖ The launch site KSC LC-39A is relatively close to the railway (15.23 km), highway (20.28 km), and coastline (14.99 km).

❖ Additionally, the launch site KSC LC-39A is relatively close to its closest city, Titusville (16.32 km).

❖ A failed rocket traveling at high speed can cover distances of approximately 15-20 km in a few seconds, posing a potential danger to populated areas.

https://github.com/Tumfunde-mwck/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

This setup allows rapid analysis of the relationship between payloads and launch sites, identifying optimal launch locations based on payload characteristics.

- ❖Implement a dropdown list for launch site selection.
- ❖Integrate a pie chart to show the aggregate count of successful launches and their distinction between success and failure.
- ❖Introduce a range slider for payload selection.
- ❖Add a scatter chart to show the correlation between payload mass and the success rate of booster versions.

https://github.com/Tumfunde-mwck/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Predictive Analysis Workflow:

Key Features:

- ❖All models optimized via GridSearchCV
- ❖Standardized data for fair comparison
- ❖Consistent accuracy evaluation metric

Data Preparation

- ➢ Load dataset → Extract target class (Y) → Standardize features (X) → Train/test split

Model Training & Tuning

- ➢ Logistic Regression + GridSearchCV → Test accuracy
- ➢ SVM + GridSearchCV → Test accuracy
- ➢ Decision Tree + GridSearchCV → Test accuracy
- ➢ KNN + GridSearchCV → Test accuracy

Evaluation

- ➢ Compare all model accuracies
- ➢ Select best-performing method

https://github.com/Tumfunde-mwck/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

Key SpaceX Launch Performance Insights

1. Launch Infrastructure

- ❖ 4 active launch sites utilized, and Initial launches served company/NASA facilities

2. Booster Capabilities

- ❖ F9 v1.1 average payload: 2,928 kg, Multiple booster versions successfully landed on droneships with above-average payloads

3. Landing Milestones

- ❖ First successful landing: 2015 (5 years post-first launch) 2015 drone ship landing failures: F9 v1.1 B1012, F9 v1.1 B1015,
- ❖ Significant improvement in landing success rates over time

4. Mission Success

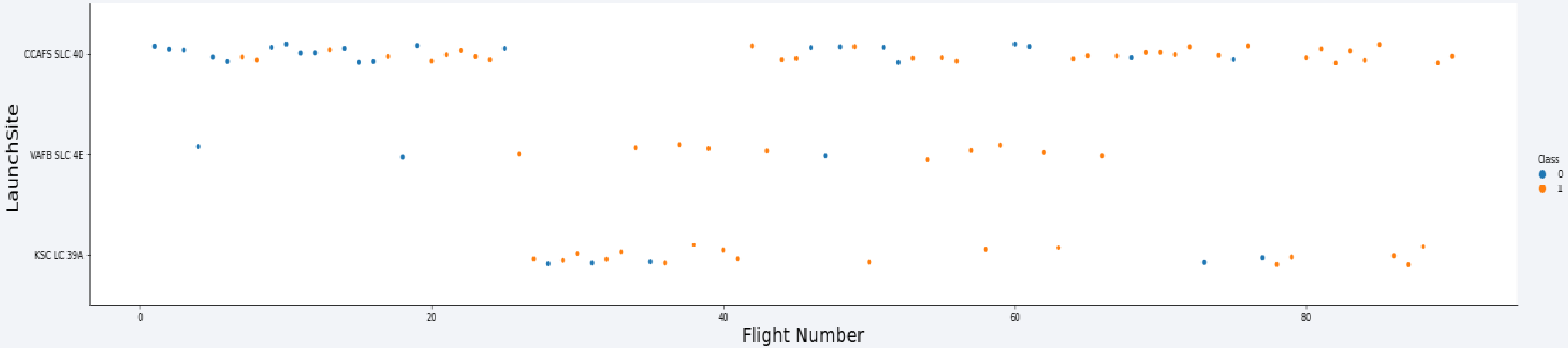
- ❖ Current mission success rate: ~100%
- ❖ Demonstrated reliability improvement through iterative design

Section 2

Insights drawn from EDA

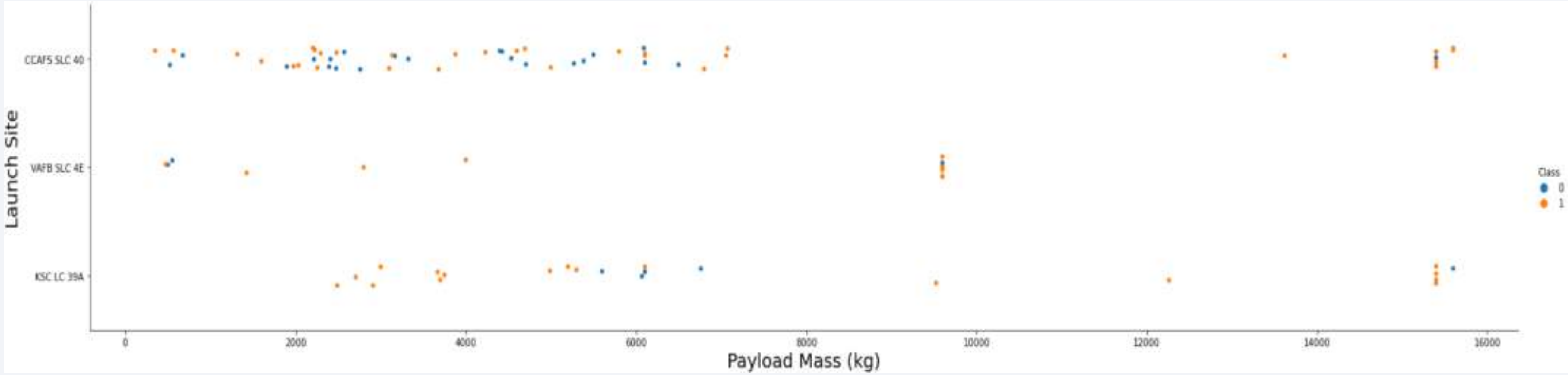
Flight Number vs. Launch Site

- As outlined in the plot, the most suitable launch site for contemporary launches appears to be CCAF5 SLC 40, which has witnessed a surge in successful launches. VAFB SLC 4E and KSC LC 39A occupy the second and third positions, respectively. Furthermore, it is evident that the overall success rate has exhibited a positive trajectory over time.



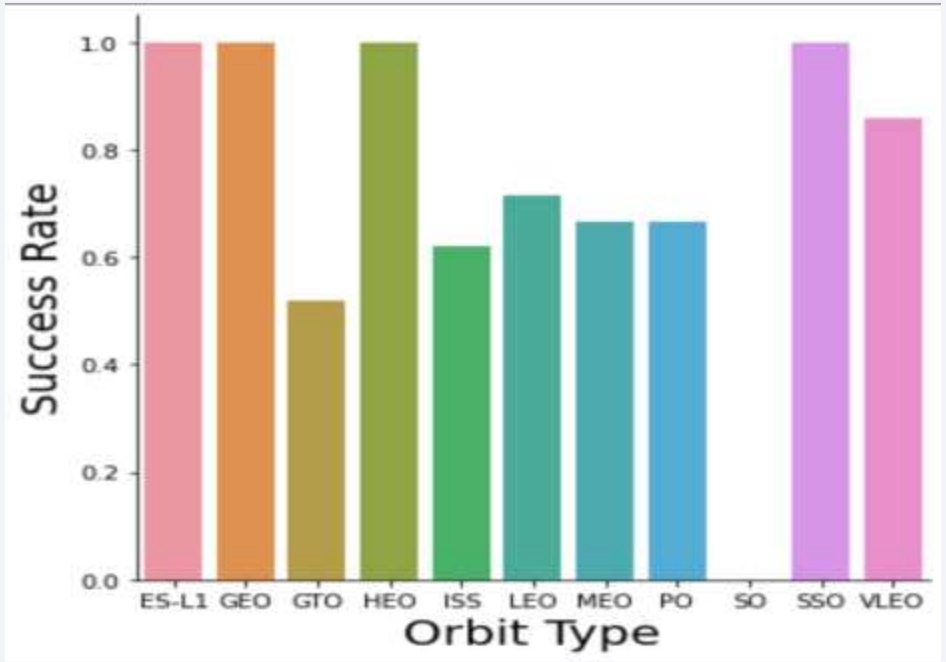
Payload vs. Launch Site

❖The majority of flights with payload masses exceeding 7000 kilograms were successful. In contrast, the Launch Complex 39A at Kennedy Space Center has a 100% success rate for payload masses below 5500 kilograms. Notably, the success rate at all launch sites is directly proportional to the payload mass.



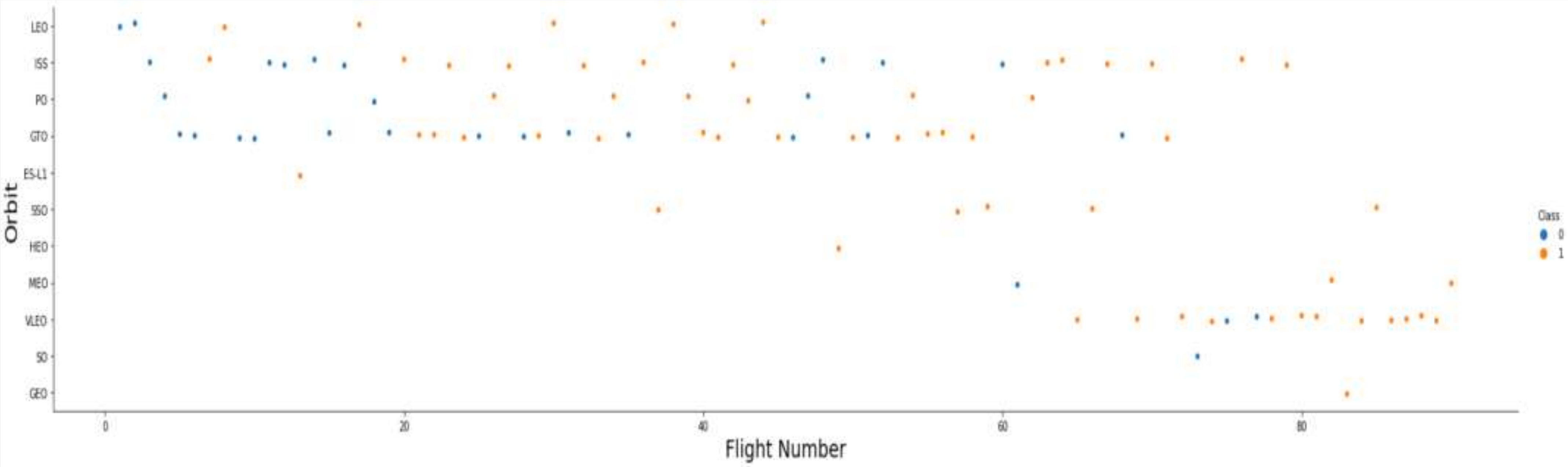
Success Rate vs. Orbit Type

- ❖The OS orbit has a 0% success rate.
- ❖The ELS-1, GEO, HEO, and SSO orbits have a 100% success rate.
- ❖Orbits GTO, ISS, LEO, MEO, and PO have a success rate between 50% and 85%.



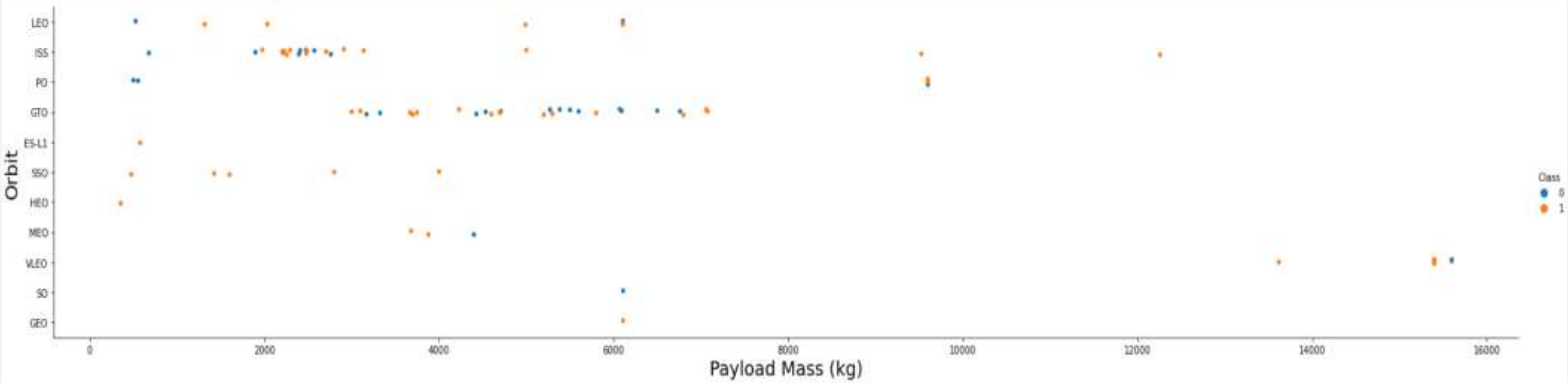
Flight Number vs. Orbit Type

- ❖The majority of the flights were launches to the International Space Station (ISS) and Geostationary Transfer Orbit (GTO). The data suggests that there is no correlation between the flight number and the orbit type.



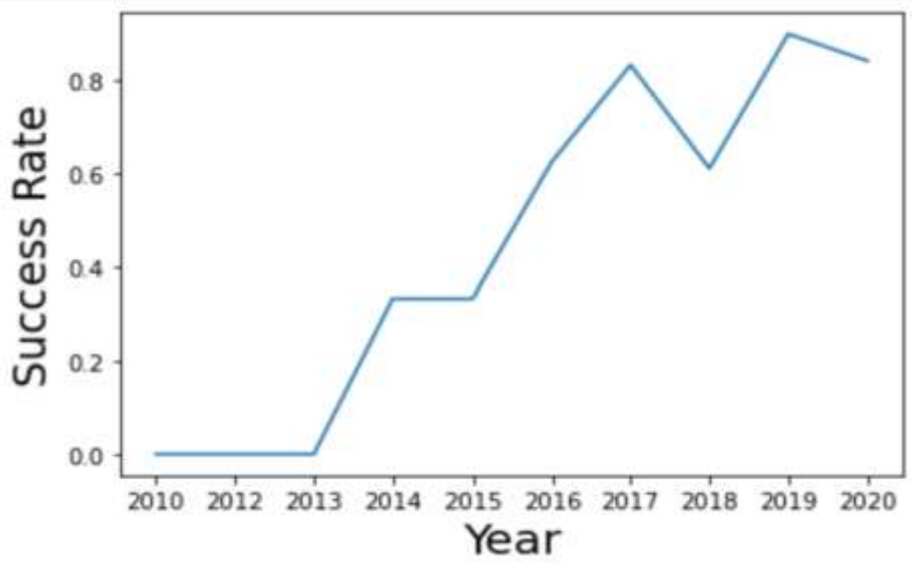
Payload vs. Orbit Type

- Payload masses exceeding 10,000 kilograms were deployed into the PO, ISS, and LEO orbits. Payload masses between 4,000 and 8,000 kilograms were positioned in the GTO orbit.



Launch Success Yearly Trend

❖The launch success rate has steadily increased since 2013. The increase in the success rate between 2013 and 2017 was linear. However, there was a decline in the launch success rate in 2018.



All Launch Site Names

❖The unique launch site names and the query structure for the dataset, which includes unique occurrences of ‘launch _site’ values.

LaunchSite
CCAFS SLC 40
CCAFS SLC 40
CCAFS SLC 40
VAFB SLC 4E
CCAFS SLC 40

Launch Site Names Begin with 'CCA'

- Five records for launch sites commence with the string “CCA,” and the query employed to retrieve the pertinent information is presented below.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

❖The calculated total payload mass carried by boosters from NASA’s site is 45,596 kilograms. The query for obtaining the total payload mass is presented below.

```
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

Average Payload Mass by F9 v1.1

❖The average payload mass carried by booster version F9 v1.1 is 2534.7 kg. The query used to calculate the average payload mass carried by booster F9 v1.1 is as follows:

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" LIKE 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

❖Identifying the First Successful Landing, By filtering data based on the successful landing outcome on the ground pad and extracting the minimum value for the date, it is possible to pinpoint the initial occurrence of this event, which transpired on December 22, 2015.

```
[ ]  1 %sql SELECT MIN(Date) AS first_successful_landing_date FROM SPACEXTABLE  WHERE landing_outcome = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.
first_successful_landing_date
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

❖Boosters that have successfully landed on drone ships and have a payload mass between 4,000 and 6,000 kilograms. After selecting distinct booster versions based on the filters provided, the following four boosters are the results:

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

Total Number of Successful and Failure Mission Outcomes

The total number of successful and unsuccessful missions is as follows:

Failure (in flight): 1

Successful flights: 99

Success (payload status unclear): 1

The query result is presented below.

```
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

Boosters Carried Maximum Payload

❖Enumerating the names of the booster versions that have successfully carried the maximum payload capacity.

```
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.bs2io90108kqblod8lcg.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

2015 Launch Records

- ❖Below is a list of the failed “landing_outcomes” in drone ships, their booster versions, and the launch sites during the year 2015.
- ❖The query used to obtain this information is also provided below.

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.bs2io90108kqblod8lcg.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)

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

❖ Ranking the count of landing outcomes (such as "Failure (drone ship)" or "Success (ground pad)") between the dates 2010-06-04 and 2017-03-20 in descending order.

```
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.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[13]:
```

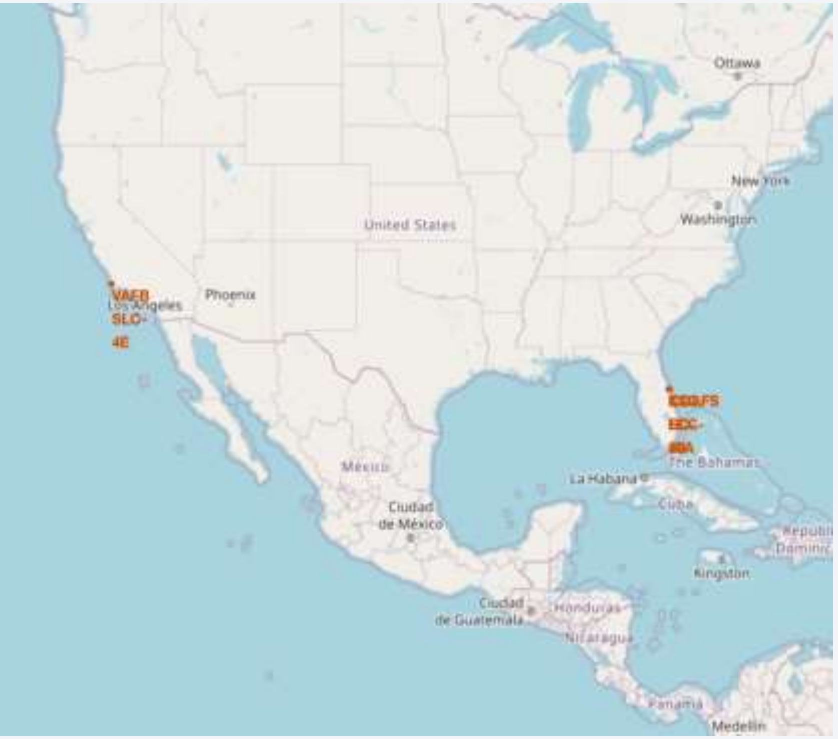
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Map showing the locations of all launch sites

❖ Most launch sites are situated near the Equator line. The Earth's surface moves faster at the Equator compared to any other location. Consequently, objects on the Earth's surface at the Equator are already moving at a speed of 1,670 kilometers per hour. When a ship is launched from the Equator, it enters space and continues to orbit the Earth at the same velocity it had before launch due to the principle of inertia. This velocity is essential for maintaining the spacecraft's altitude and ensuring its successful entry into orbit. Launch sites are typically situated in close proximity to the coast, which minimizes the risk of debris falling or exploding near populated areas during rocket launches.



Color labels on a map showing launch sites

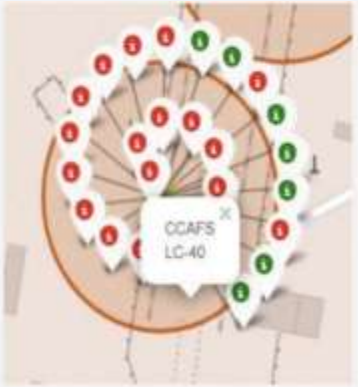
❖Using the color-coded markers, we can readily discern the launch sites with relatively high success rates.

Green Marker: Successful Launch

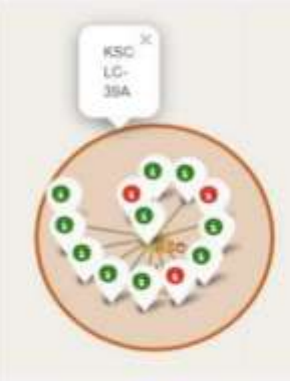
Red Marker: Failed Launch

Launch Site KSC LC-39A exhibits an exceptionally high success rate.

	Launch Site	Lat	Long	marker_color
0	CCAFS LC-40	28.562302	-80.577356	red
1	CCAFS LC-40	28.562302	-80.577356	red
2	CCAFS LC-40	28.562302	-80.577356	red
3	CCAFS LC-40	28.562302	-80.577356	red
4	CCAFS LC-40	28.562302	-80.577356	red
5	CCAFS LC-40	28.562302	-80.577356	red
6	CCAFS LC-40	28.562302	-80.577356	red
7	CCAFS LC-40	28.562302	-80.577356	red
8	CCAFS LC-40	28.562302	-80.577356	red
9	CCAFS LC-40	28.562302	-80.577356	red
10	CCAFS LC-40	28.562302	-80.577356	red
11	CCAFS LC-40	28.562302	-80.577356	red
12	CCAFS LC-40	28.562302	-80.577356	red
13	CCAFS LC-40	28.562302	-80.577356	red
14	CCAFS LC-40	28.562302	-80.577356	red
15	CCAFS LC-40	28.562302	-80.577356	red
16	CCAFS LC-40	28.562302	-80.577356	red
17	CCAFS LC-40	28.562302	-80.577356	green
18	CCAFS LC-40	28.562302	-80.577356	green
19	CCAFS LC-40	28.562302	-80.577356	red
20	CCAFS LC-40	28.562302	-80.577356	green



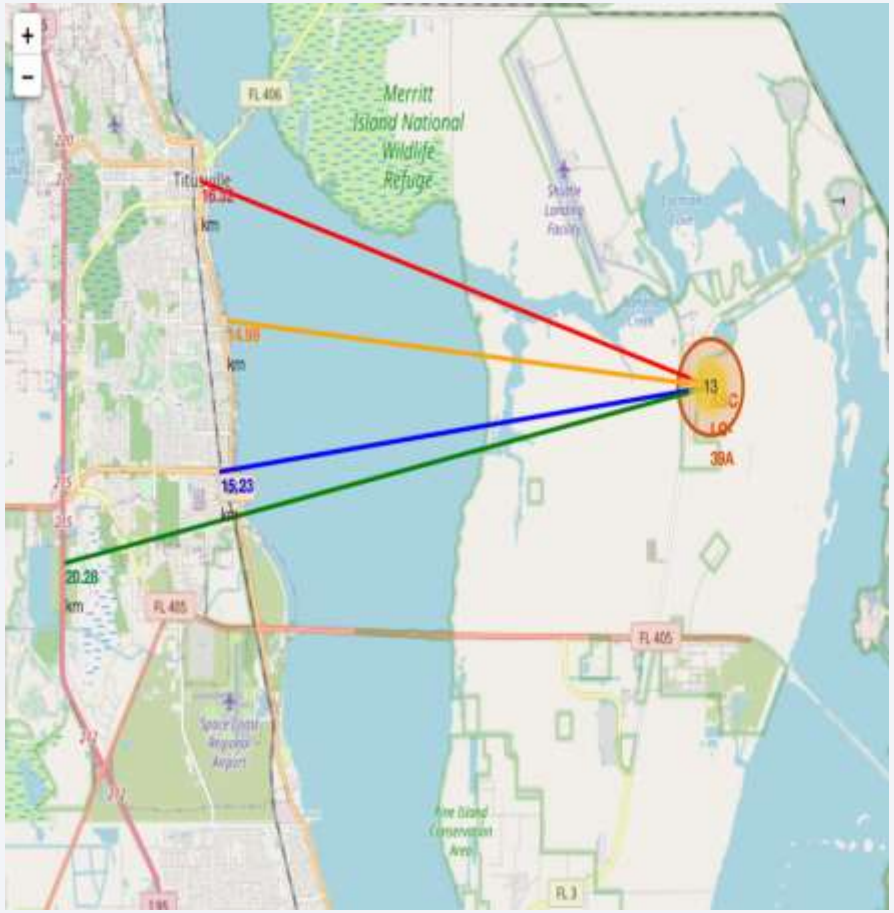
36	KSC LC-39A	28.573255	-80.646895	green
37	KSC LC-39A	28.573255	-80.646895	red
38	KSC LC-39A	28.573255	-80.646895	green
39	KSC LC-39A	28.573255	-80.646895	green
40	KSC LC-39A	28.573255	-80.646895	red
41	KSC LC-39A	28.573255	-80.646895	green
42	KSC LC-39A	28.573255	-80.646895	green
43	KSC LC-39A	28.573255	-80.646895	red
44	KSC LC-39A	28.573255	-80.646895	green
45	KSC LC-39A	28.573255	-80.646895	green
46	KSC LC-39A	28.573255	-80.646895	green
47	KSC LC-39A	28.573255	-80.646895	green
48	KSC LC-39A	28.573255	-80.646895	green



Green= Successful Launch
Red= Failed Launch

The launch site of KSC LC-39A

- From the visual analysis of Launch Complex 39A at Kennedy Space Center (KSC), it is evident that the launch site is situated in close proximity to several significant geographical features:
- The launch site is approximately 15.23 kilometers away from the railway.
- The launch site is approximately 20.28 kilometers away from the highway.
- The launch site is approximately 14.99 kilometers away from the coastline. Additionally, Launch Complex 39A is relatively close to its closest city, Titusville, located approximately 16.32 kilometers away. It is crucial to note that a failed rocket traveling at high speeds can cover distances of approximately 15-20 kilometers in a matter of seconds. Such a scenario poses a potential risk to populated areas.



Section 4

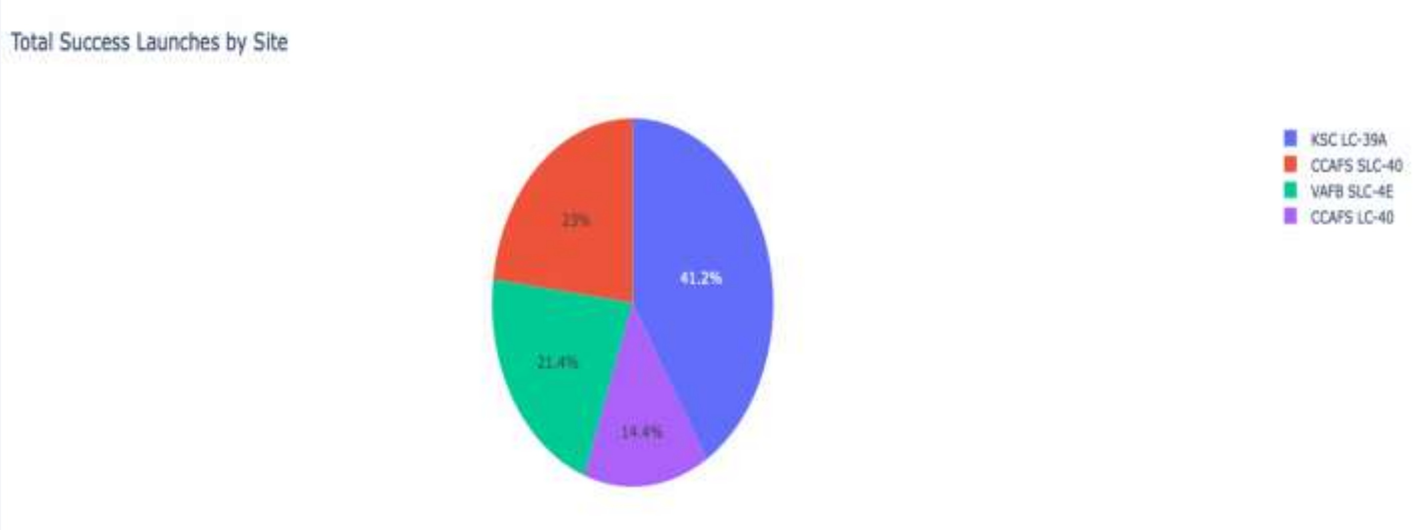
Build a Dashboard with Plotly Dash

Launch Success Count for all sites

❖The highest success launch rates were recorded at the following sites:

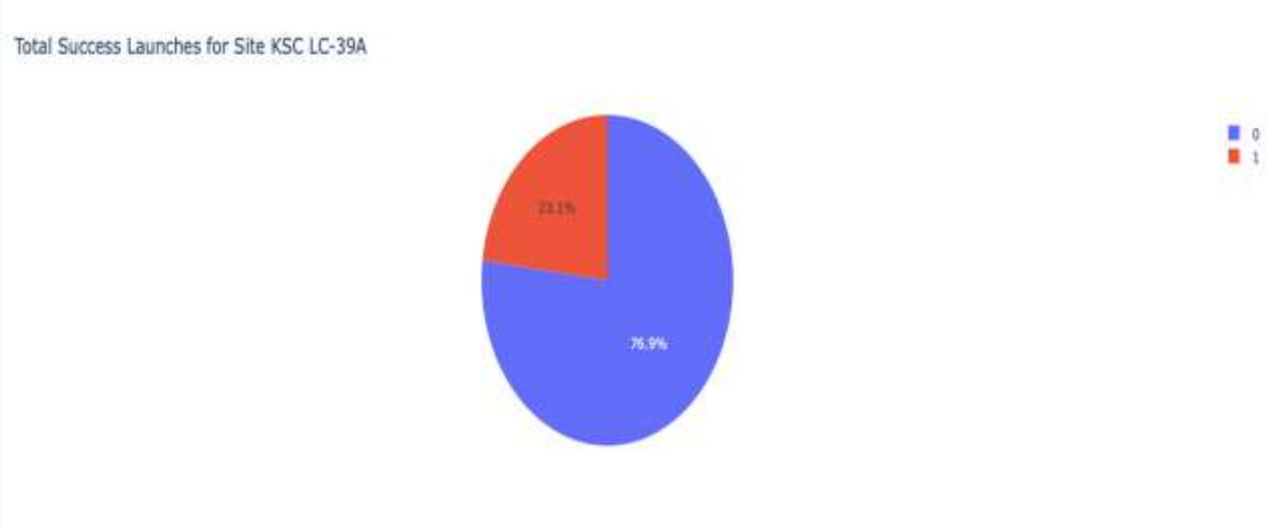
KSC LC-39A (41.7%)

CCAFS LC-40 (29.2%)



41KSC LC-39 Launch Site Success Rate

❖Launch Complex 39A at Kennedy Space Center (KSC) boasts the highest launch success rate, achieving 76.9% with only 3 out of 10 launches resulting in failed landings.



Payload Mass vs Launch Outcome for all sites

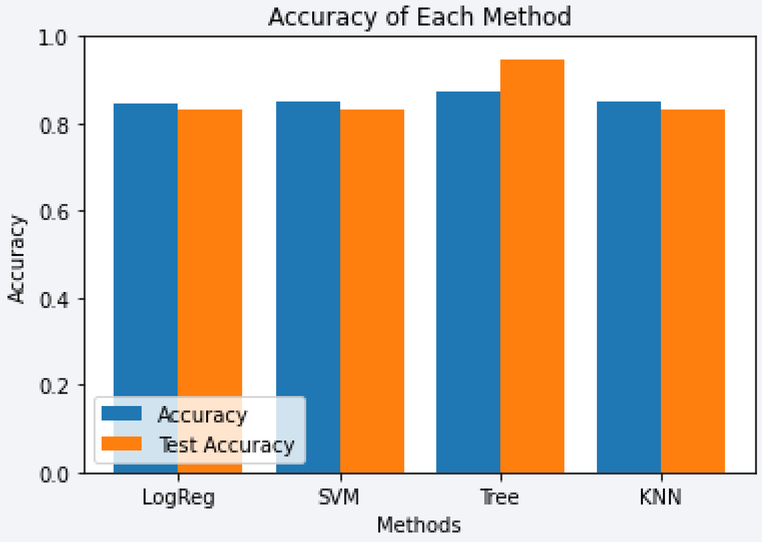
❖The charts indicate that payloads with a weight range of 2,000 to 5,500 kilograms exhibit the highest success rate.

Section 5

Predictive Analysis (Classification)

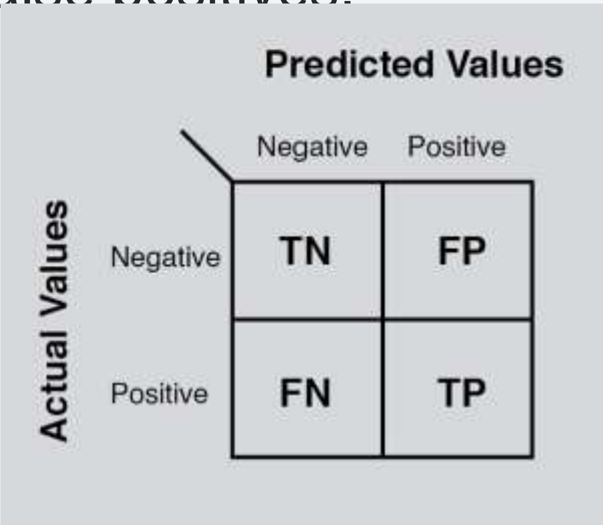
Classification Accuracy

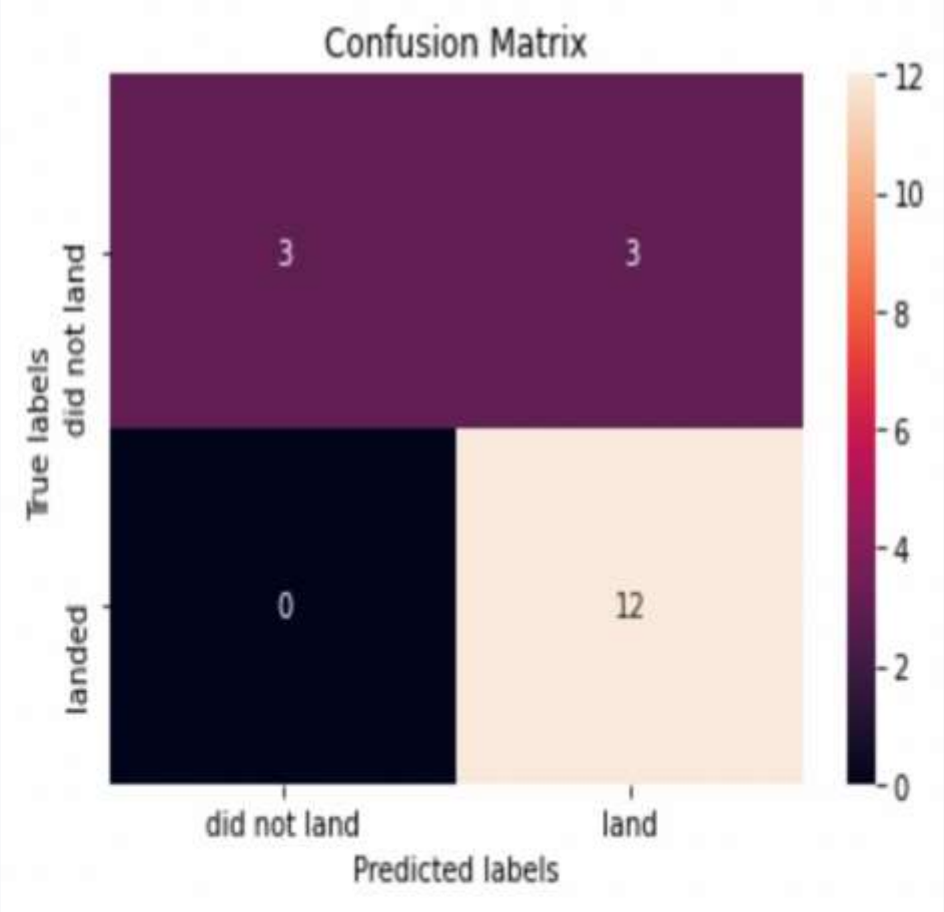
- ❖Four classification models were tested, and their accuracies are presented alongside.
- ❖The Decision Tree Classifier exhibits the highest classification accuracy, surpassing 87%.



Confusion Matrix

❖Examining the confusion matrix, we observe that logistic regression effectively differentiates between the various classes. However, the primary challenge lies in the occurrence of false positives.





Conclusions

1. Comprehensive Data Pipeline: Established an end-to-end data workflow from collection (SpaceX API/Wikipedia) through cleaning, analysis, and visualization.
2. Actionable Launch Insights: Uncovered critical patterns in launch success rates by site, payload, and booster version through rigorous EDA and SQL analysis.
3. Geospatial Intelligence: Developed interactive Folium maps to visualize geographic patterns in launch/landing outcomes, enhancing location-based decision making.
4. Operational Dashboard: Built an interactive Plotly Dash tool enabling real-time exploration of launch parameters and success correlations.
5. Predictive Capability: Implemented machine learning models to forecast landing success with accuracy, creating a competitive advantage for launch planning.
6. Open-Space Benchmarking: Delivered a reproducible analytical framework for comparing SpaceX's cost efficiency and reliability against industry competitors.
7. Cost Optimization: Identified $62M launch cost efficiencies through reusable booster success analysis.

Appendix

1. Data Sources: SpaceX API + Wikipedia scraping

    Sources: SpaceX API (api.spacexdata.com/v4/rockets) – Structured rocket launch data

    Wikipedia Scraping (List of Falcon 9/Falcon Heavy Launches) – Historical mission details

2. Tools: Python (Pandas, Folium, Dash, Scikit-learn)

3. Deliverables: Interactive dashboard + predictive models

Thank you!