Winning Space Race with Data Science

Priya Dharshini D
07-04-2025

Outline

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

Executive Summary

Methodologies:

- ✓ Data Collection (Web scraping, SpaceX API)
- ✓ Data Wrangling (Cleaning and preprocessing)
- ✓ 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 with Logistic Regression, SVM, Decision Trees, KNN)

Summary of Results:

1. EDA Results (Identifying patterns and correlations)
2. Interactive Analytics Demo (Screenshots of dashboard and map)
3. Predictive Analysis Results (Model performance and evaluation metrics)

Introduction

Project Background and Context:

SpaceX has become the leading company in commercial space exploration, reducing costs through innovations like the reusable first stage of its Falcon 9 rockets. Each launch costs around $62 million, significantly lower than competitors due to this reusability. By predicting whether the first stage will successfully land, we can estimate the total cost of a launch and further optimize space missions. This project uses public data and machine learning models to analyze and predict the reusability of SpaceX's first stage, which plays a critical role in determining launch expenses.

Questions to be Answered:

1. How do variables like payload mass, launch site, number of flights, and orbits affect the success of first stage landings?
2. Does the rate of successful landings increase over the years?
3. What is the best algorithm for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Performed data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a 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

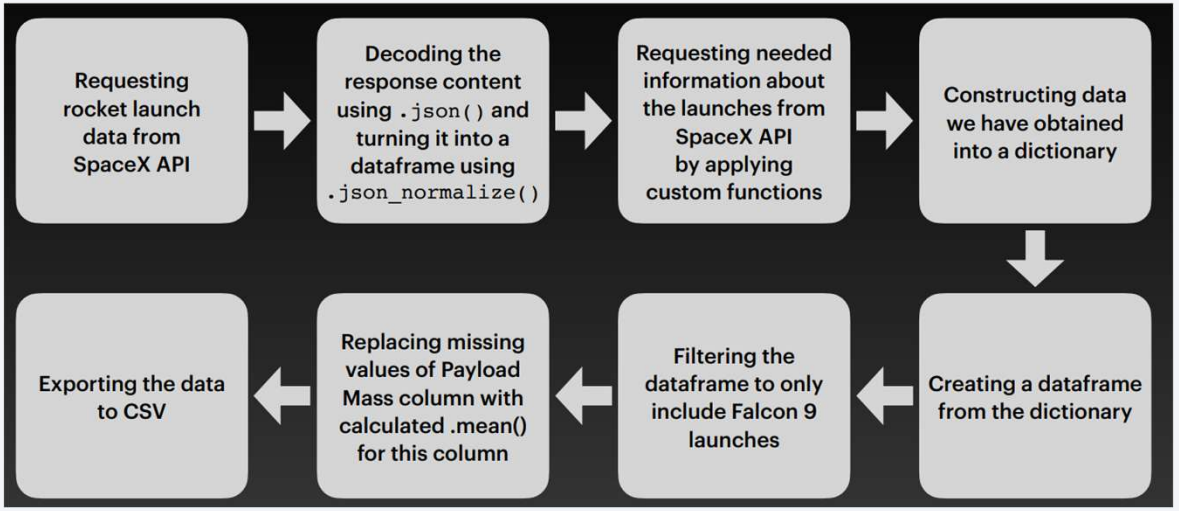
Data Collection

- ✓ The data collection for this project involved a combination of API requests and web scraping. The primary data sources were SpaceX's REST API and SpaceX's Wikipedia entry. These methods were used to gather comprehensive information on SpaceX's rocket launches, providing more detailed insights into various factors that influence launch success.

- ✓ Data was collected from the SpaceX API (https://api.spacexdata.com/v4/rockets/) and Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches), using web scraping techniques to fill gaps and ensure a thorough dataset for analysis.

Data Collection – SpaceX API

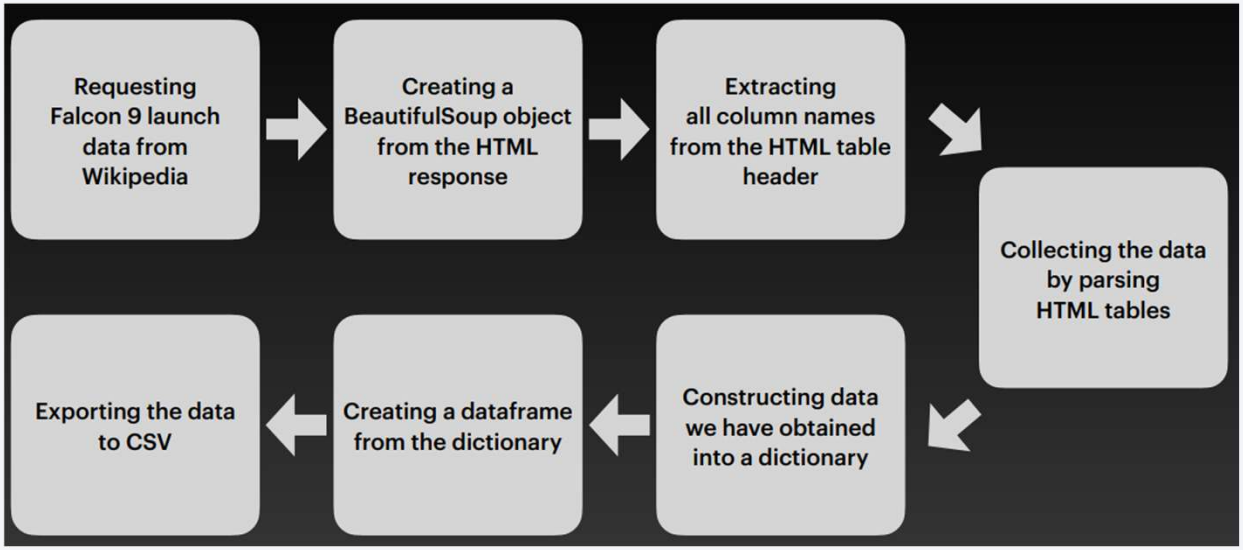
- SpaceX offers a public API from where data can be obtained and then used.
- The API was utilized following the process outlined in the flowchart, and the data is then saved.



GitHub URL – https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/Data-collection-API%20-1.ipynb

Data Collection - Scraping

- Data on SpaceX launches can also be sourced from Wikipedia.
- The data is downloaded from Wikipedia as per the flowchart and then stored.



GitHub URL - https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping%20-2.ipynb

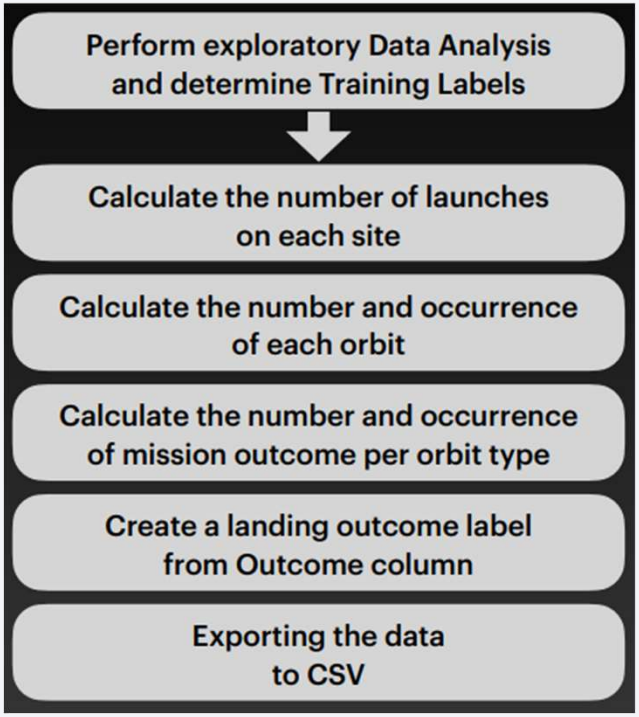
Data Wrangling

Exploratory Data Analysis (EDA) was performed to understand the dataset's structure and identify any missing or inconsistent data.

The data was summarized by calculating the number of launches per site, the frequency of each orbit type, and the occurrences of mission outcomes per orbit type.

A new column for landing outcomes was created, assigning "1" for successful landings (True Ocean, True RTLS, True ASDS) and "0" for unsuccessful landings (False Ocean, False RTLS, False ASDS).

This transformation converted categorical outcomes into binary labels for machine learning models.



GitHub URL - https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Wrangling%20-3.ipynb

EDA with Data Visualization

- Various charts were plotted to explore relationships between variables.
- Scatter plots were used to examine connections between continuous or categorical variables, such as Flight Number vs. Payload Mass, Launch Site vs. Flight Number, and Success Rate by Orbit Type.
- These plots help identify patterns useful for machine learning models.
- Bar charts compared discrete categories, like Launch Site vs. Flight Number and Orbit Type vs. Success Rate, showing differences between categories.
- Line charts were used to show trends over time, like the Success Rate Yearly Trend.
- These visualizations helped reveal insights into the relationships between features like Payload Mass, Flight Number, and Launch Site.

GitHub URL - https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization%20-5.ipynb

EDA with SQL

The following SQL queries were performed:

- Displayed the names of unique launch sites in the space mission.
- Displayed the top 5 launch sites whose names begin with the string "CCA".
- Calculated the total payload mass carried by boosters launched by NASA (CRS).
- Computed the average payload mass carried by the booster version F9 v1.1.
- Listed the date when the first successful landing outcome on a ground pad was achieved.
- Displayed the names of boosters with success on a drone ship and a payload mass between 4000 and 6000 kg.
- Counted the total number of successful and failed mission outcomes.
- Listed the names of booster versions that carried the maximum payload mass.
- Displayed failed landing outcomes on a drone ship, including booster versions and launch site names for the year 2015.
- Ranked the count of landing outcomes (Failure on drone ship, Success on ground pad) between 2010-06-04 and 2017-03-20 in descending order.

GitHub URL - https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20With%20SQL%20-4.ipynb

Build an Interactive Map with Folium

- Markers, circles, marker clusters, and lines were added to a Folium map to visualize launch sites and their characteristics.
- Markers pinpointed the locations of launch sites, including NASA Johnson Space Center, while circles highlighted key areas, such as proximity to the Equator and coasts.
- Marker clusters grouped launches at each site to identify areas with higher activity, and colored lines displayed distances between launch sites (e.g., KSC LC-39A) and nearby features like railways, highways, coastlines, and cities.
- These objects were used to provide a clear, organized view of the launch sites, their proximity to key landmarks, and launch success rates.

GitHub URL - https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab%20-6.ipynb

Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List**:

  Added a dropdown list to enable Launch Site selection.

- **Pie Chart showing Success Launches (All Sites/Certain Site)**:

  Added a pie chart to display the total successful launches for all sites and the Success vs. Failed counts for a specific Launch Site, if selected.

- **Slider of Payload Mass Range**:

  Added a slider to select the Payload range.

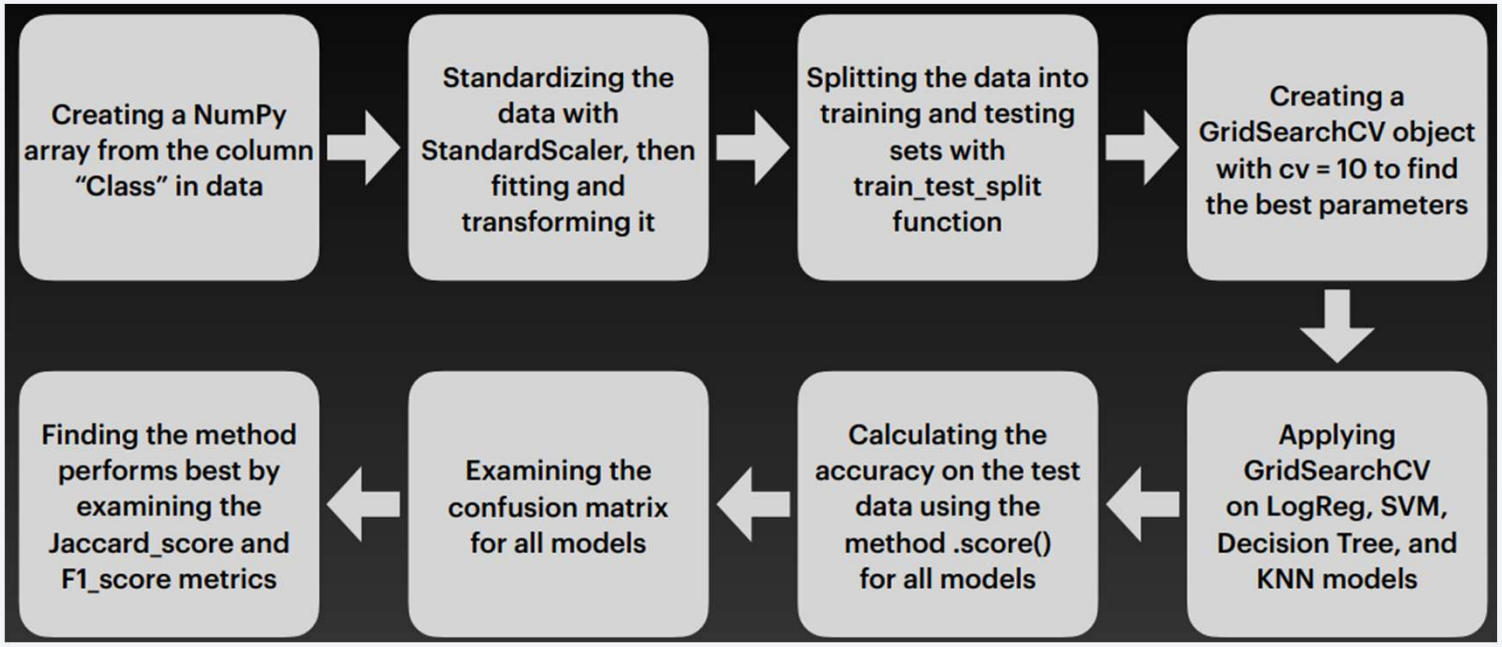
- **Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions**:

  Added a scatter chart to show the correlation between Payload Mass and Launch Success for different booster versions.

GitHub URL - https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k - nearest neighbors.



GitHub URL - https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction%20-7.ipynb

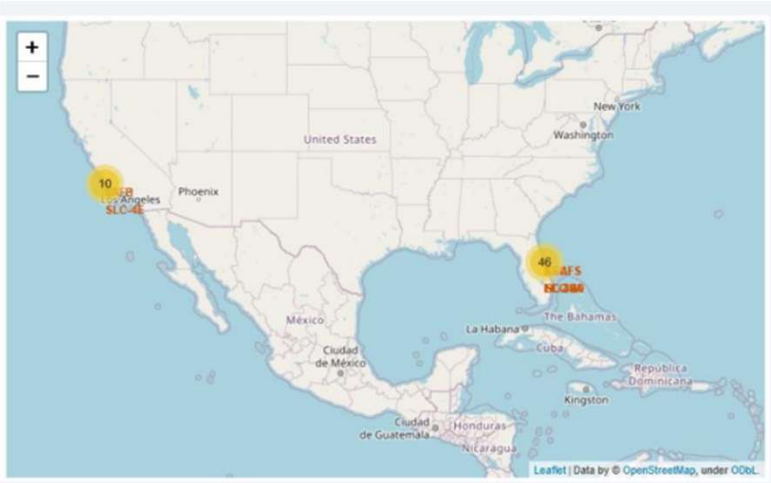
Results

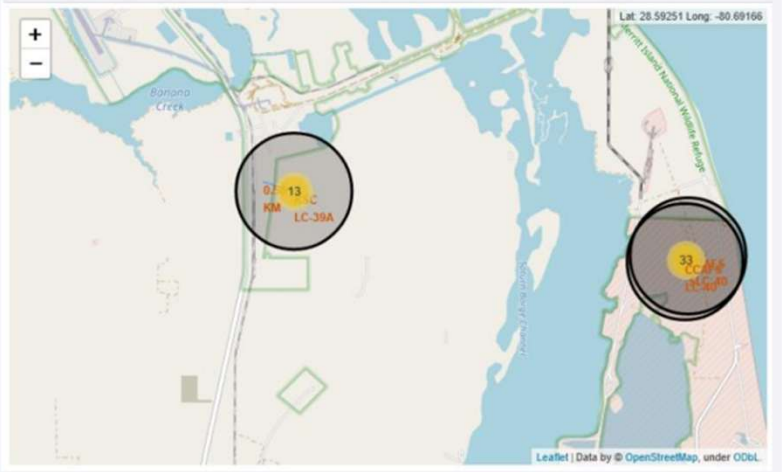
Exploratory Data Analysis Results:

- SpaceX operates 4 different launch sites.
- The first launches were conducted by SpaceX itself and NASA.
- The average payload of the F9 v1.1 booster is 2,928 kg.
- The first successful landing outcome occurred in 2015, five years after the first launch.
- Many Falcon 9 booster versions successfully landed on drone ships, with payloads above the average.
- Nearly 100% of mission outcomes were successful.
- Two booster versions (F9 v1.1 B1012 and F9 v1.1 B1015) failed to land on drone ships in 2015.
- The number of successful landing outcomes improved over the years.

Interactive analytics demo in screenshots:

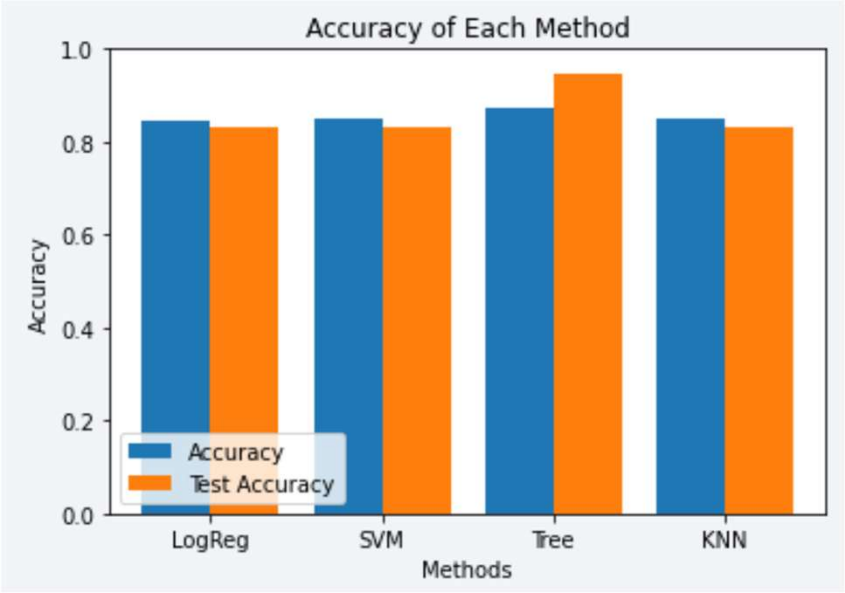
- Interactive analytics revealed that launch sites are typically located in safe areas, often near the sea, and have strong logistical infrastructure nearby.
- The majority of launches occur at launch sites on the east coast.

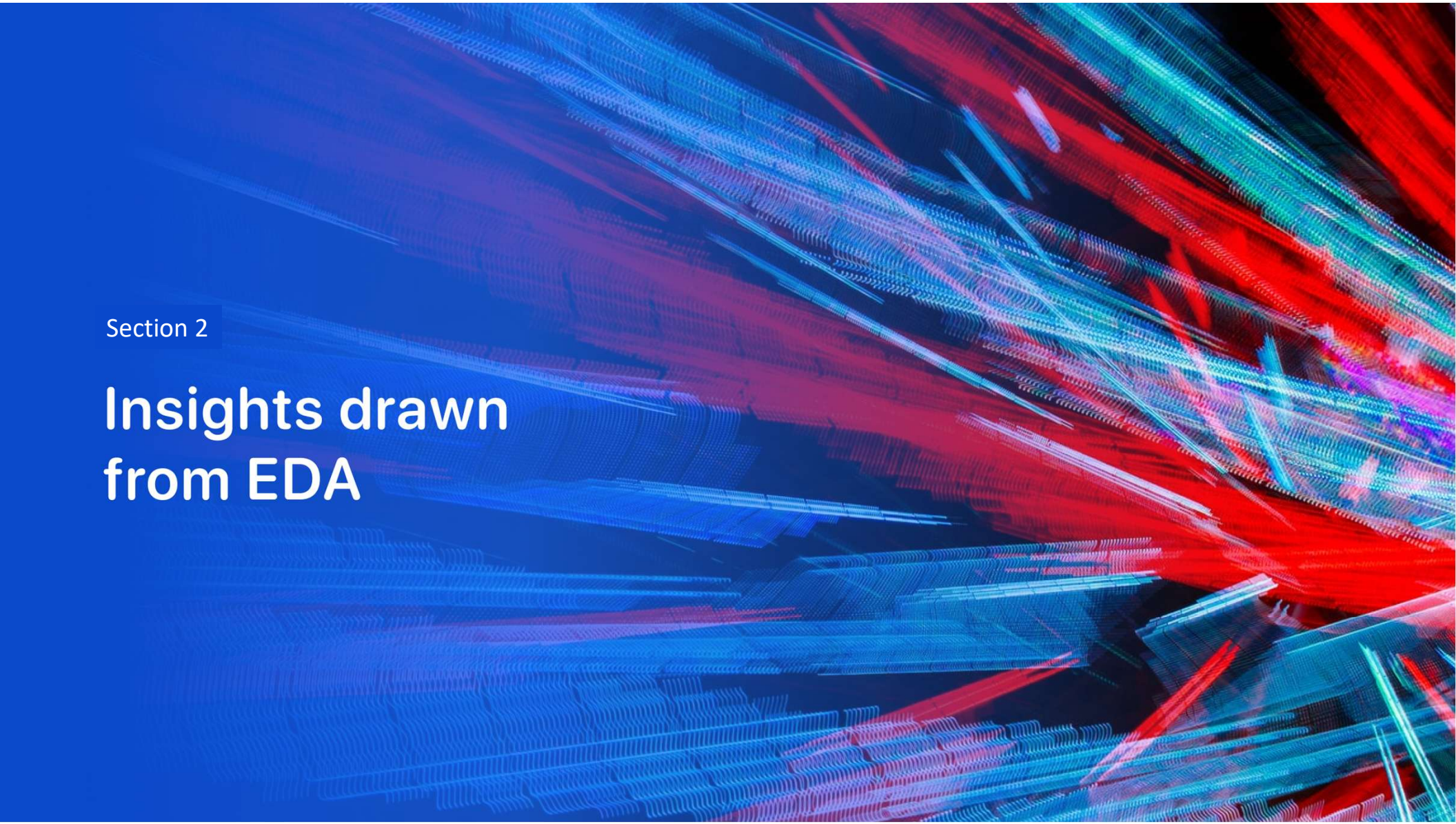




Predictive analysis results:

- Predictive analysis revealed that the Decision Tree Classifier is the most effective model for predicting successful landings, achieving an accuracy of over 87% and a test data accuracy exceeding 94%.





Section 2

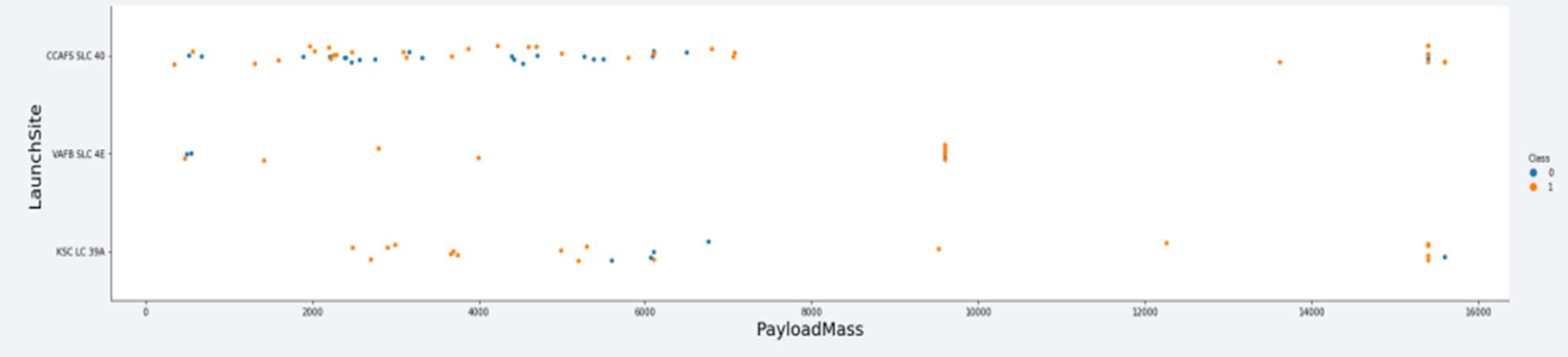
Insights drawn from EDA

Flight Number vs. Launch Site

Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



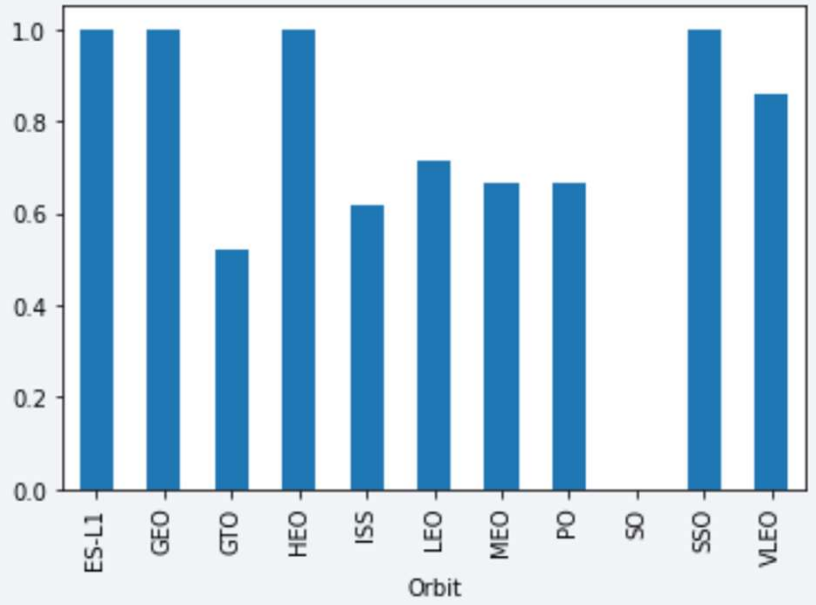
Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type

Explanation:

- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: - SO
- Orbits with success rate between 50% and 85% : - GTO, ISS, LEO, MEO, PO

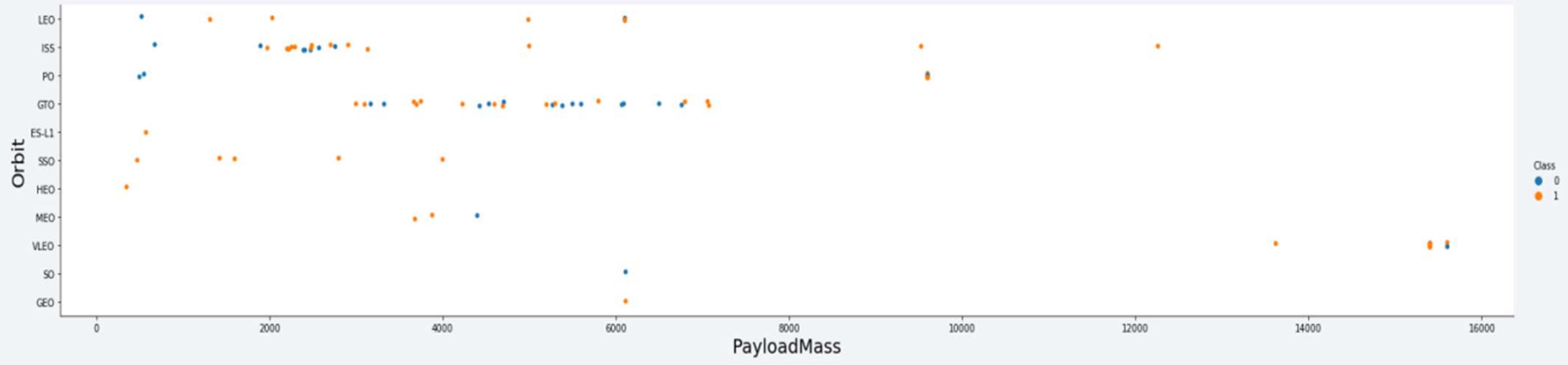


Flight Number vs. Orbit Type

Explanation:

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

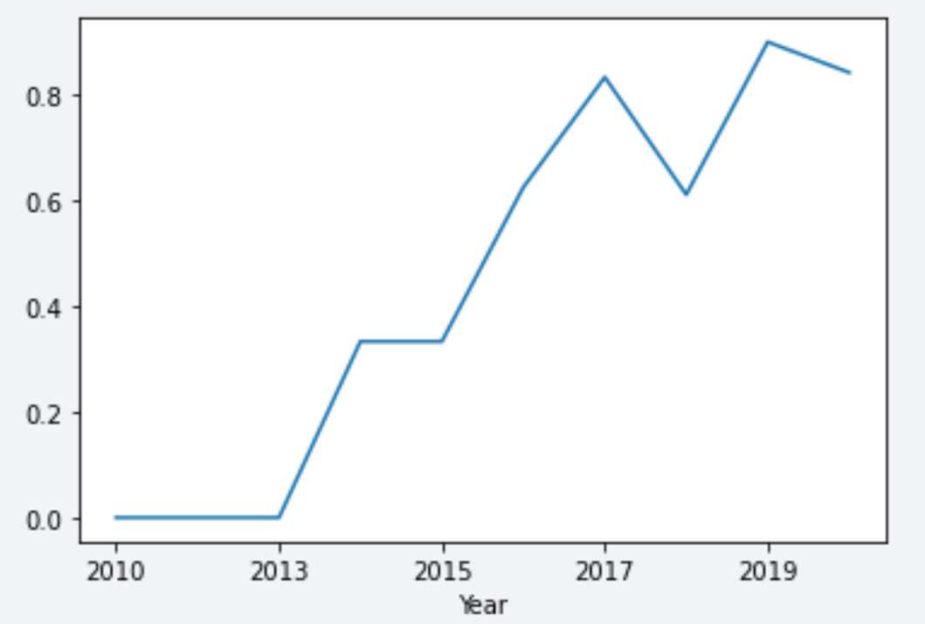


Explanation:

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020.
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

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

Out[4]:

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

Explanation:

- Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

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

Out[5]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation:

- Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

total_payload_mass
45596

Explanation:

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

Average Payload Mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[7]:
```

average_payload_mass
2534

Explanation:

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

First Successful Ground Landing Date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[8]:
```

first_successful_landing
2015-12-22

Explanation:

- Listing the date when the first successful landing outcome in 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:

- Listing the total number of successful and failure mission outcomes.

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:

- Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
             where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

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

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
             where date between '2010-06-04' and '2017-03-20'
             group by landing__outcome
             order by count_outcomes desc;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

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

Explanation:

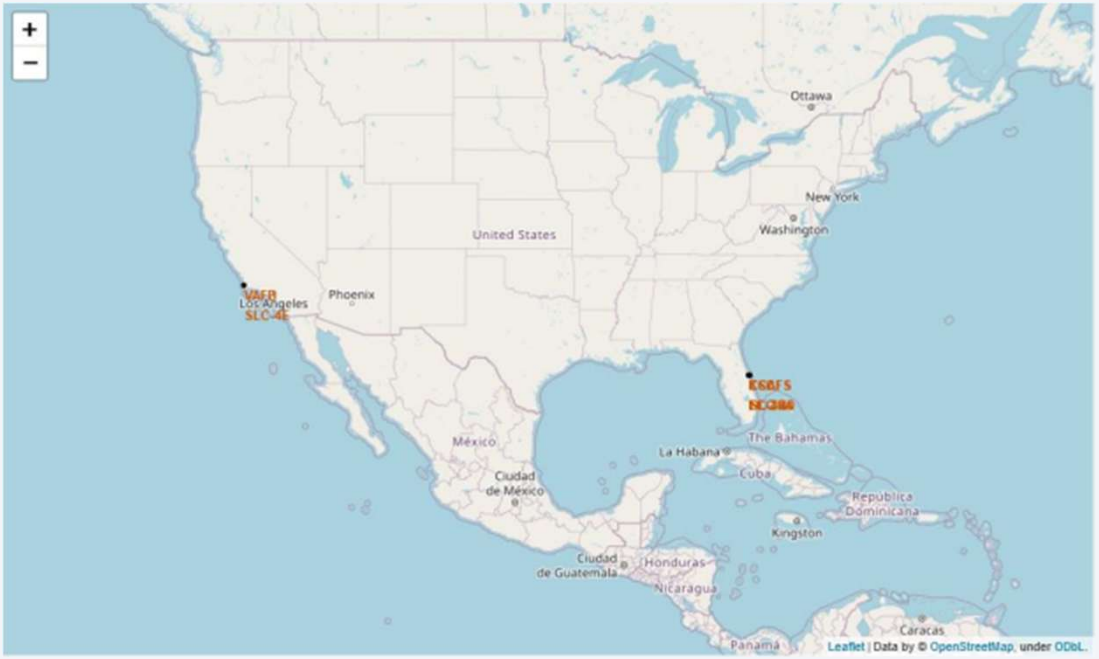
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

Section 3

Launch Sites Proximities Analysis

All launch sites

- Launch sites are near sea, probably by safety, but not too far from roads and railroads.



Launch Outcomes by Site

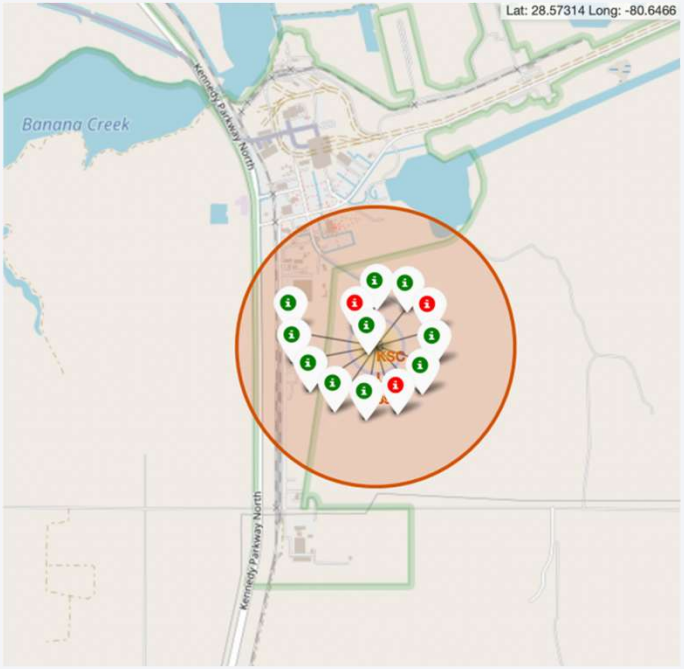
Explanation:

- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- Green Marker = Successful Launch

- Red Marker = Failed Launch

- Launch Site KSC LC-39A has a very high Success Rate.



Logistics and Safety

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)

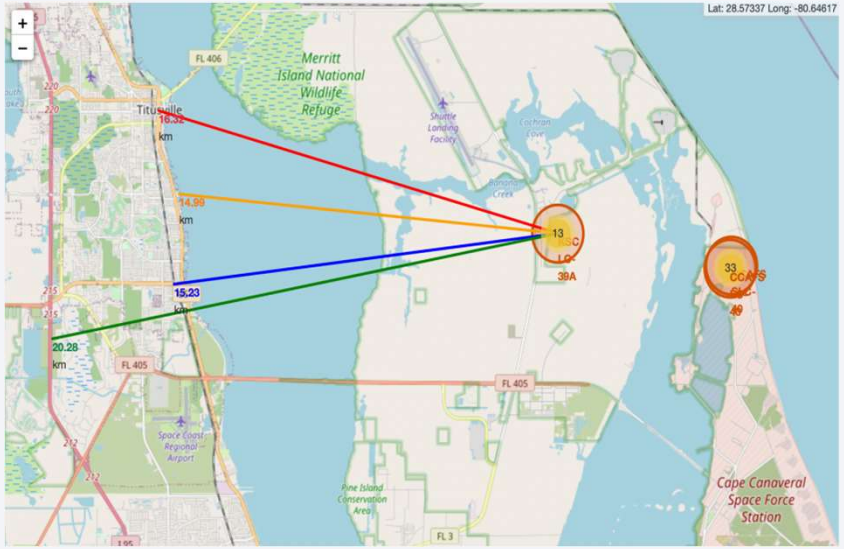
- relative close to highway (20.28 km)

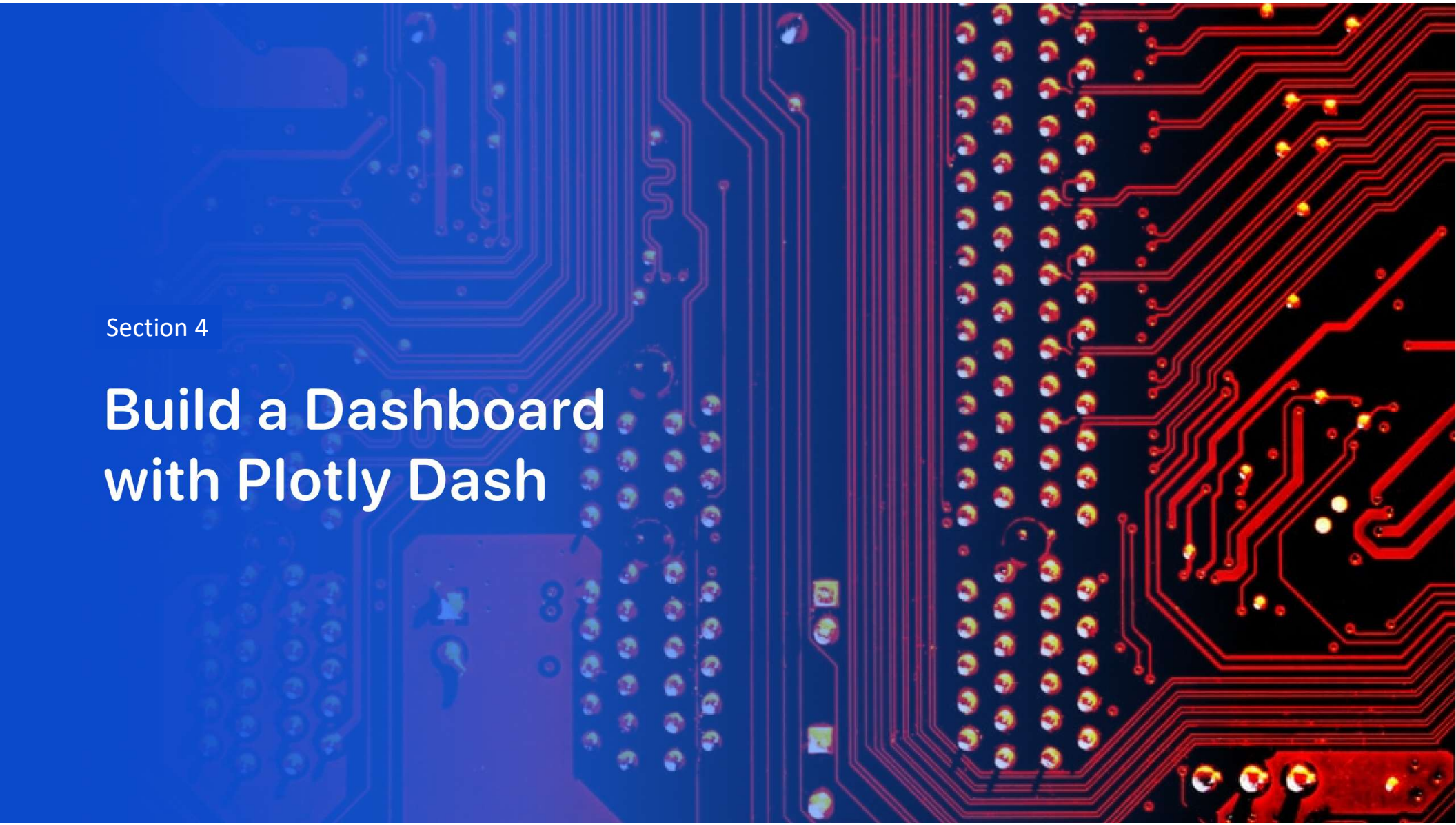
- relative close to coastline (14.99 km)

- 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.

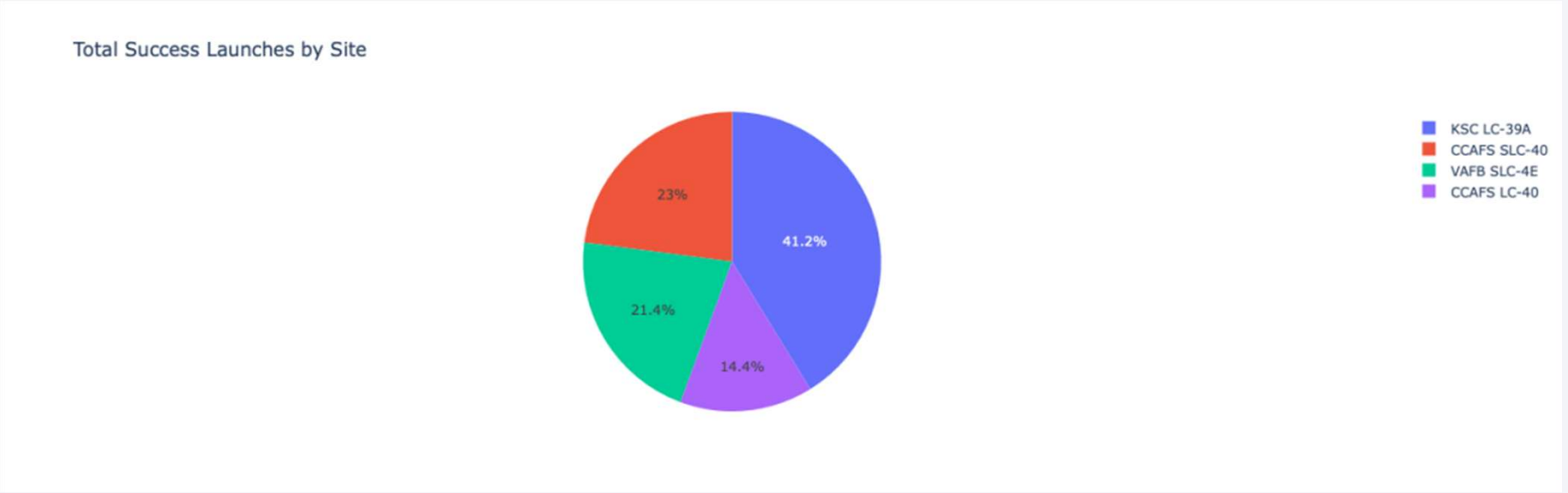




Section 4

Build a Dashboard with Plotly Dash

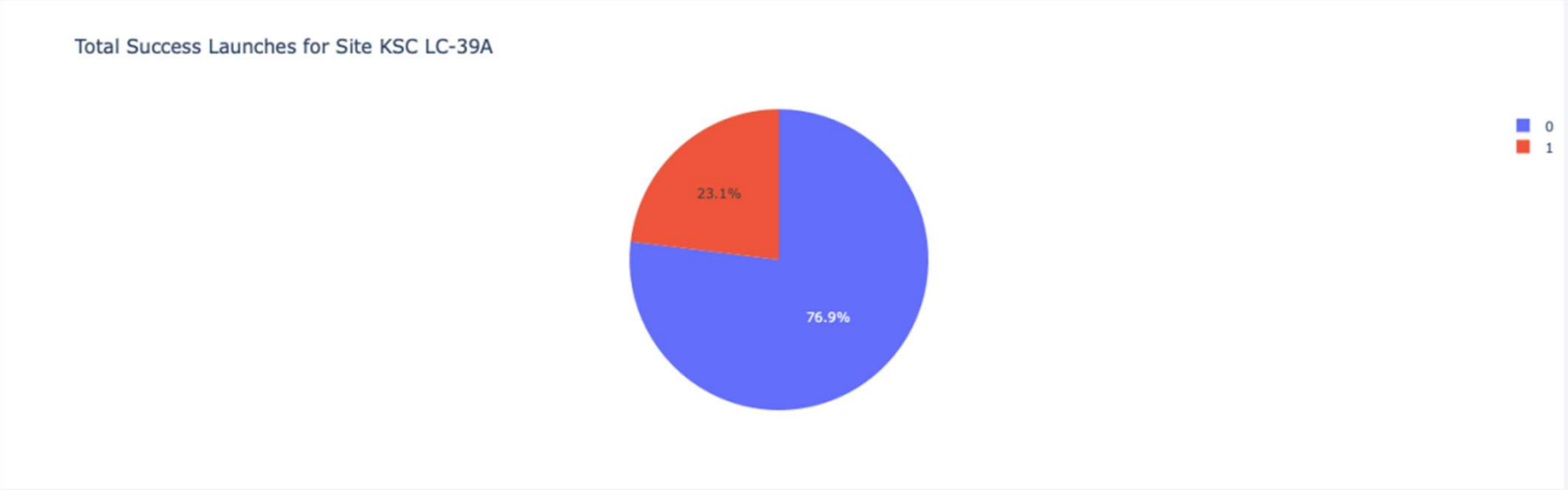
Successful Launches by Site



Explanation:

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

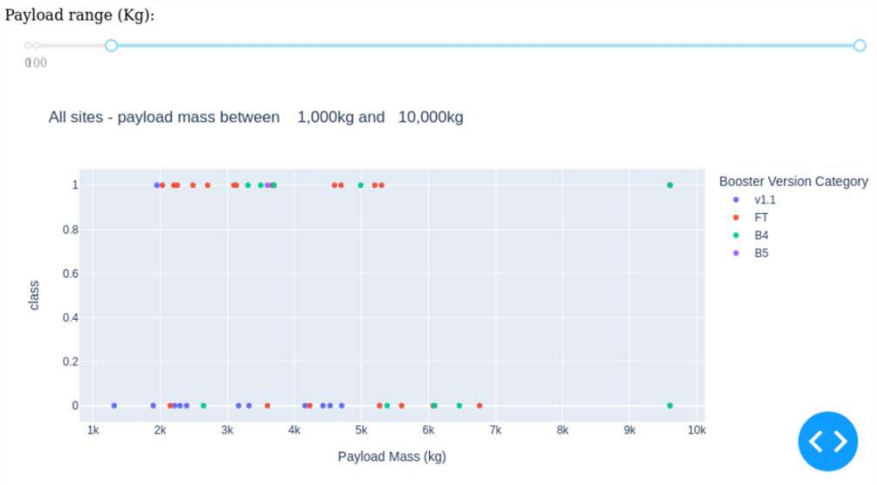
Launch site with highest launch success ratio

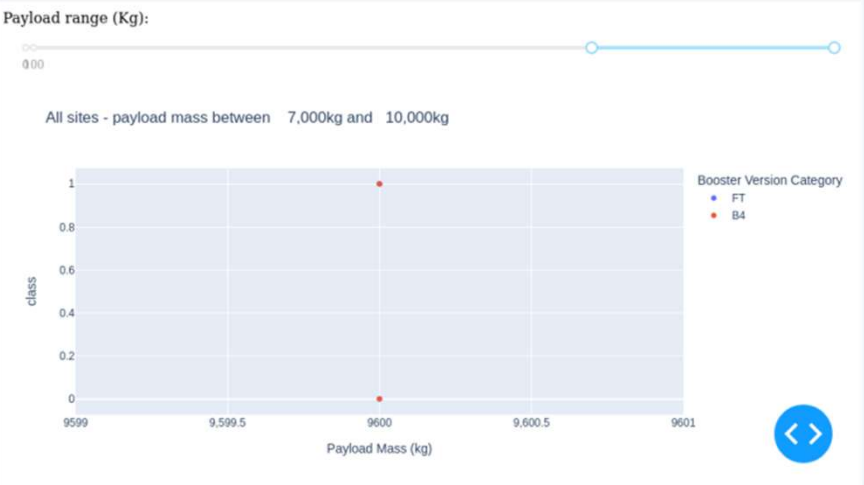


Explanation:

- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload vs. Launch Outcome





Explanation:

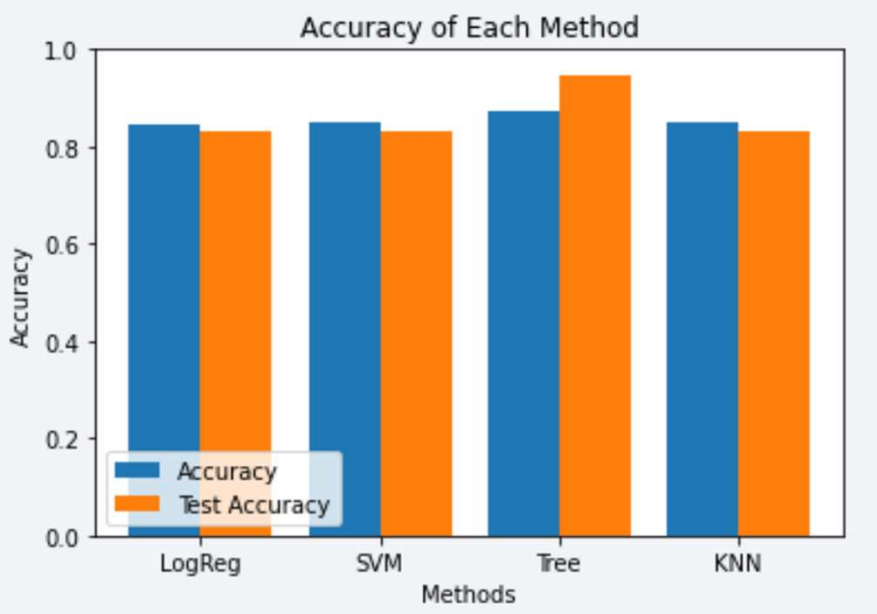
- Payloads under 6,000kg and FT boosters are the most successful combination.
- There's not enough data to estimate risk of launches over 7,000kg

Section 5

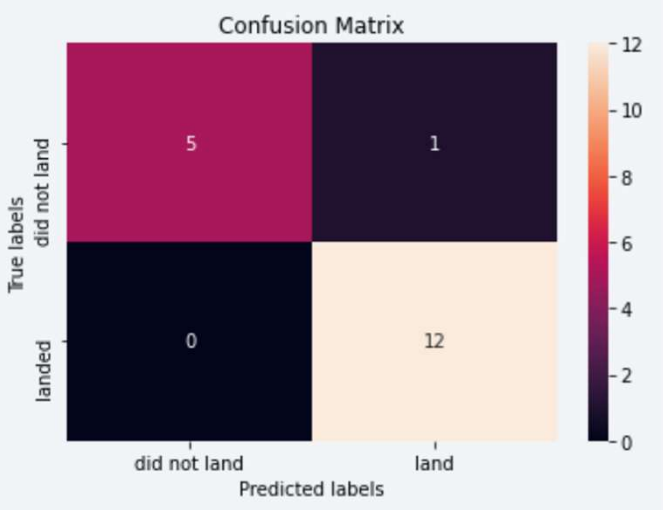
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside.
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



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.

Conclusions

- KSC LC-39A is the top launch site with the highest success rate.
- Launches with payloads over 7,000kg are less risky, indicating better success with heavier payloads.
- Most sites are near the Equator and coast, optimizing launch conditions.
- Success rates have improved over time, reflecting advances in rocket technology.
- The Decision Tree Classifier is effective for predicting successful landings and enhancing decision-making.
- Orbits like ES-L1, GEO, HEO, and SSO have a 100% success rate.
- Overall, successful landings are increasing as SpaceX improves its processes and rockets.

Appendix

Git Hub Link- (README Page)

https://github.com/priya-dharshini-d/IBM-Applied-Data-Science-Capstone/blob/main/README.md

Thank you!