IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mahesh Wadhvana
May 2025

Outline

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

Executive Summary

This project focuses on predicting whether the Falcon 9 rocket’s first stage will successfully land. Accurate predictions help estimate SpaceX's lower launch costs due to reusable rockets.

SpaceX Falcon 9 launch data was sourced from a public API and web scraping (e.g., Wikipedia), cleaned by excluding non-Falcon 9 launches, addressing missing values, and formatting for analysis. A machine learning pipeline employed Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, and K-Nearest Neighbors (KNN) on a 90-sample dataset, with features standardized and split into 80% training and 20% test sets. Hyperparameter tuning via GridSearchCV optimized model performance. Decision Tree and Logistic Regression achieved the highest test accuracy (83%), SVM (78%) showed overfitting, and KNN underperformed (72%). To enhance accuracy, we recommend incorporating additional features, expanding the dataset, and exploring ensemble methods.

Introduction

Project background and context:

This project focuses on predicting whether the Falcon 9 rocket's first stage will successfully land. Accurate predictions help estimate SpaceX's lower launch costs due to reusable rockets.

Problems you want to find answers to:

Can we build a model that accurately predicts Falcon 9 first stage landings? This helps determine launch cost efficiency and supports competitive pricing for other companies.

Section 1

Methodology

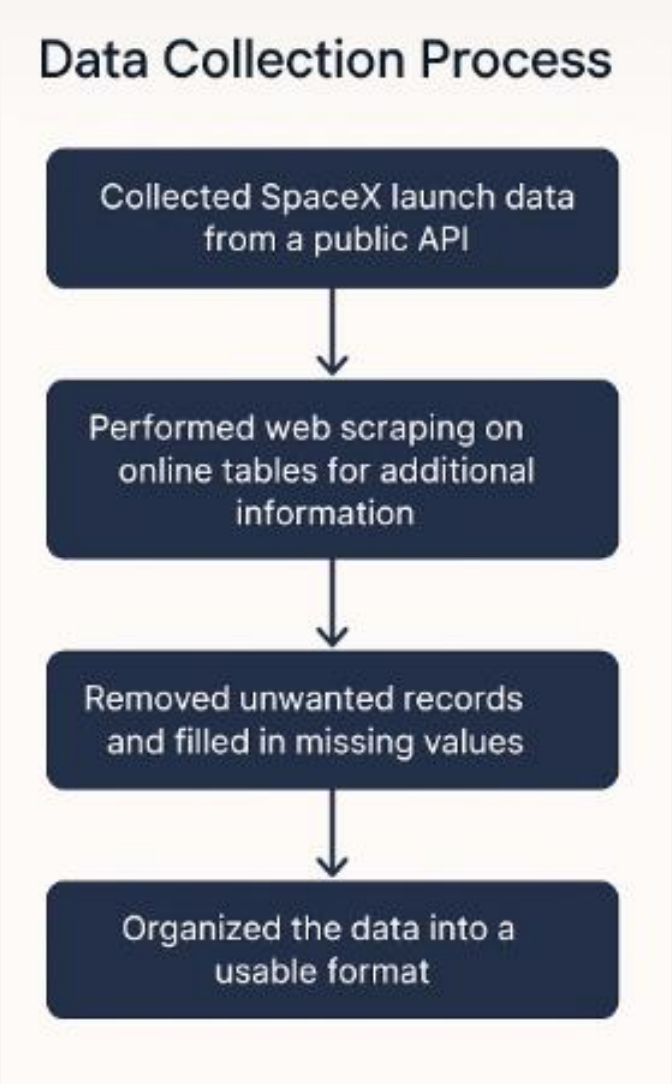
Methodology

Executive Summary

SpaceX Falcon 9 launch data was sourced from a public API and web scraping (e.g., Wikipedia), cleaned by excluding non-Falcon 9 launches, addressing missing values, and formatting for analysis. A machine learning pipeline employed Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, and K-Nearest Neighbors (KNN) on a 90-sample dataset, with features standardized and split into 80% training and 20% test sets. Hyperparameter tuning via GridSearchCV optimized model performance. Decision Tree and Logistic Regression achieved the highest test accuracy (83%), SVM (78%) showed overfitting, and KNN underperformed (72%). To enhance accuracy, we recommend incorporating additional features, expanding the dataset, and exploring ensemble methods.

Data Collection

- Describe how data sets were collected.
  - SpaceX launch data was collected from a public online source (an API) that provides detailed records about past rocket launches.
  - Additional information was gathered by pulling details from online tables, such as those found on Wikipedia, using a technique called web scraping.
  - The raw data was cleaned by removing unwanted records (like non-Falcon 9 launches), filling in missing values, and organizing it into a usable format for analysis.
- You need to present your data collection process use key phrases and flowcharts

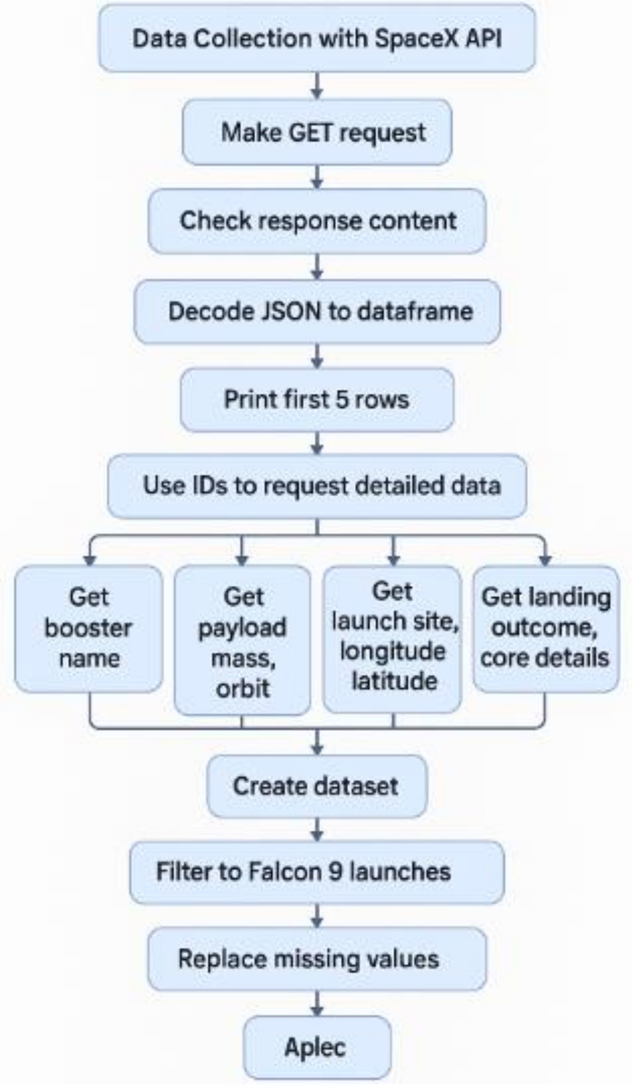


Data Collection – SpaceX API

- The flow chart illustrates how an API was used to acquire SpaceX data and create a working data set
- As an overview, the main data is acquired in and process from a Json file, additional data is acquired and combined to create a dataset then cleaned.

GitHub Notebook URL:

https://github.com/mwadhvan/IBM-SpaceX-Applied-Data-Science-Capstone-Project/blob/main/1.%20jupyter-labs-spacex-data-collection-api.ipynb

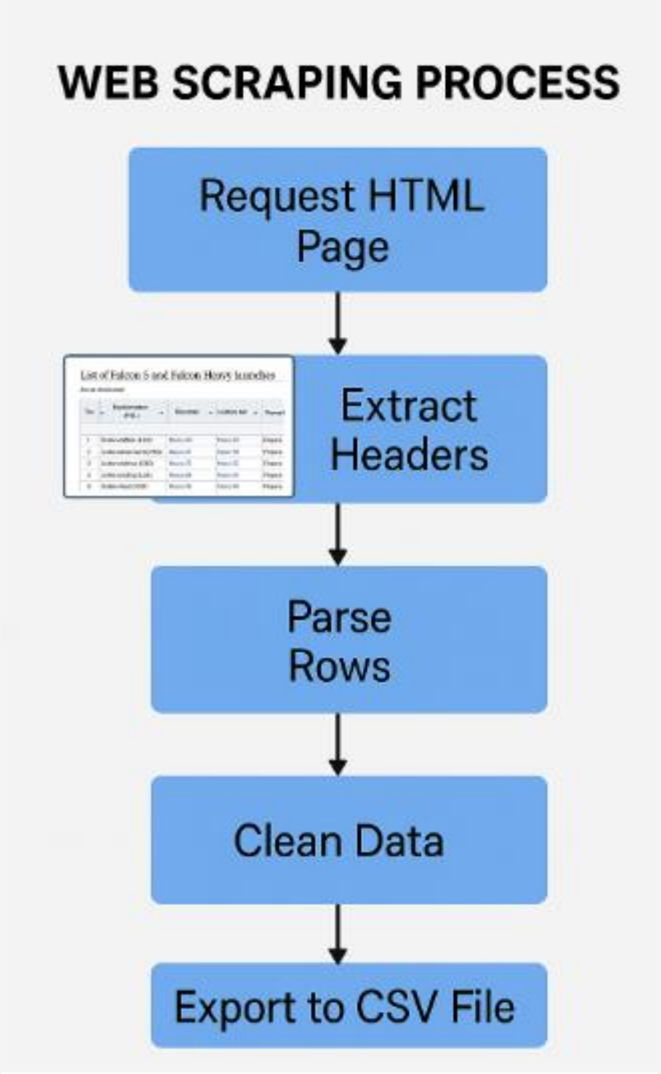


Data Collection - Scraping

- The flowchart illustrates how webscaping was utilized to gather data and create a dataset for the project

GitHub notebook URL:

https://github.com/mwadhvan/IBM-SpaceX-Applied-Data-Science-Capstone-Project/blob/main/2.%20jupyter-labs-webscraping.ipynb



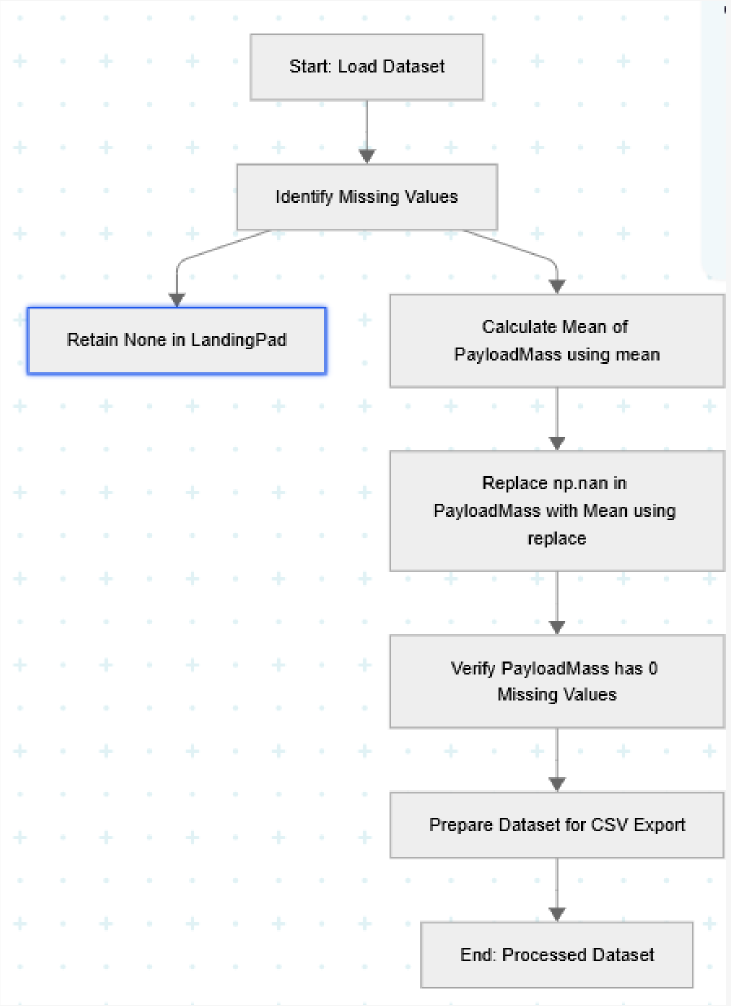
Data Wrangling

The data was cleaned and prepared for use by addressing missing information in a straightforward way.

- Kept blank entries in the LandingPad column to indicate when landing pads were not used.
- Calculated the average value for the PayloadMass column.
- Filled in missing PayloadMass entries with the calculated average.
- Ensured no gaps remained in the PayloadMass data.
- Prepared the cleaned data for saving, with a plan to use a specific date range for consistency in the next step.

GitHub notebook URL:

https://github.com/mwadhvan/IBM-SpaceX-Applied-Data-Science-Capstone-Project/blob/main/3.%20labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

To understand how flight number and payload variables influence launch outcomes, we created various following charts to visualize relationships and trends in the data.

1) Scatter Plot to Show Flight Number vs. Launch Site Success Patterns
   Plotted flight number vs. launch sites, colored by success class, revealing uneven flight distribution and more failures at CCAFS SLC 40 for lower flight numbers.
2) Scatter Plot to Identify Payload Mass Trends by Launch Site
   Plotted payload mass vs. launch sites, highlighting that VAFB-SLC had no launches with heavy payloads (>10,000 kg).
3) Bar Chart to Compare Success Rates by Orbit Type
   Bar chart of success rates per orbit type, identifying higher success in orbits like LEO and ISS.
4) 4: Scatter Plot to Explore Flight Number and Orbit Type Success Correlation
   Plotted flight number vs. orbit types, colored by success, showing strong success correlation in LEO but none in GTO.
5) Scatter Plot to Examine Payload Mass Impact by Orbit Type
   Plotted payload mass vs. orbit types, colored by success, noting higher success for heavy payloads in Polar, LEO, and ISS, but mixed results in GTO.
6) Bar Chart to Track Yearly Success Trends
   Bar chart of yearly average success rate, showing a steady increase from 2013 to 2020.

GitHub notebook URL:
https://github.com/mwadhvan/IBM-SpaceX-Applied-Data-Science-Capstone-Project/blob/main/5.%20edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

- The data was analyzed to extract key insights about space missions.

    - Listed all unique launch site names.
    - Summed total payload weight for NASA commercial resupply missions.
    - Calculated average payload weight for F9 v1.1 rocket model.
    - Identified the date of the first successful ground pad landing.
    - Listed rockets with successful drone ship landings and payloads between 4000-6000 kg.
    - Counted total successful and failed mission outcomes.
    - Found rockets that carried the heaviest payloads.
    - Detailed 2015 drone ship landing failures, including month, rocket, and launch site.
    - Ranked mission outcomes from June 4, 2010, to March 20, 2017.

GitHub URL:
https://github.com/mwadhvan/IBM-SpaceX-Applied-Data-Science-Capstone-Project/blob/main/4.%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Map Object Summary and Purpose

The following summary outlines the key map objects used in the visualization, along with their purposes for enhancing spatial understanding, site identification, and launch analysis.

1. Circles: Black circles (1000m radius) highlight each launch site (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E) for visibility and identification, with site name popups. An orange circle marks NASA JSC for differentiation.

2. Markers: Orange text markers identify launch sites. Rocket-icon markers—green for successful and red for failed launches—visualize launch outcomes. Distance markers (1.34 KM to coastline, 26.67 KM to city) aid in proximity analysis.

3. Polylines: Lines from CCAFS LC-40 to the coastline and city provide spatial context and support safety assessments. These elements collectively enhance clarity and support effective spatial analysis.

GitHub notebook URL:
https://github.com/mwadhvan/IBM-SpaceX-Applied-Data-Science-Capstone-Project/blob/main/6.%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Visuals & Interactions

1. Pie Chart: Shows total successful launches per site, or success vs. failure for a selected site

2. Scatter Plot: Displays correlation between payload mass and launch outcome, colored by booster version

3. Dropdown Menu: Filters both charts by selected launch site

4. Payload Slider: Refines scatter plot by payload mass range

Purpose of these is to provide a clear picture of launch performance by site, payload, and booster type - enabling fast, interactive analysis of trends and outcomes.

GitHub notebook URL:

https://github.com/mwadhvan/IBM-SpaceX-Applied-Data-Science-Capstone-Project/blob/main/7.%20spacex-dash-app.py

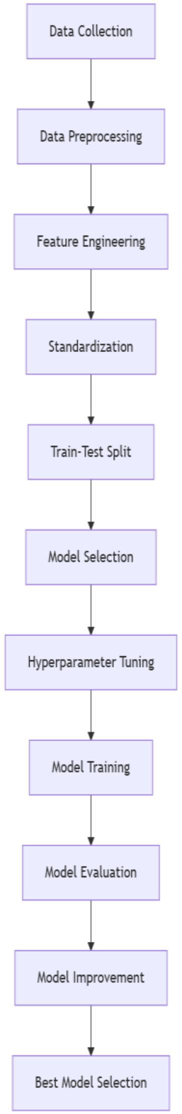
Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found th best performing classification model

- Standardized features, split data into train/test sets, used GridSearchCV to tune and compare classifiers (SVM, Logistic Regression, Decision Tree, KNN), and selected the model with the highest test accuracy.

- You need present your model development process using key phrases and flowchart

GitHub notebook URL:
https://github.com/mwadhvan/IBM-SpaceX-Applied-Data-Science-Capstor Project/blob/main/8.%20SpaceX_Machine%20Learning%20Prediction_Pa .ipynb



Results

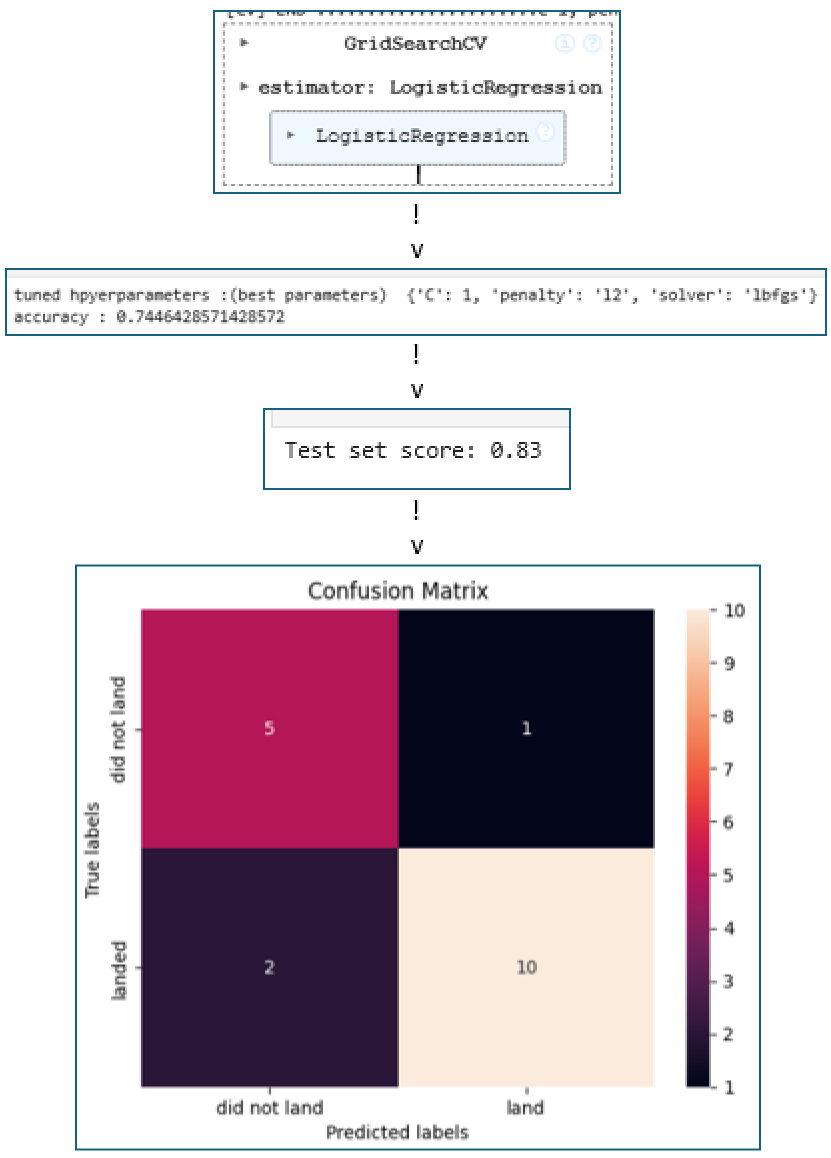
1. Exploratory Data Analysis Results
    - Loaded Falcon 9 launch data; identified key features and target (Class).
    - Data showed both numerical and categorical variables; some missing values.

2. Interactive Analytics Demo in Screenshots *(see diag. on right)*
    - Displayed sample data and feature set using .head().
    - Visualized model performance with confusion matrix plots.

3. Predictive Analysis Results
    - Standardized features, split data (80/20), trained/tested multiple classifiers.
    - SVM achieved highest test accuracy (~94%); selected as best model.

Model	Best Parameters	Accuracy (%)
SVM (RBF Kernel)	C=10, gamma=1	**94.44**
Logistic Regression	C=1	88.89
KNN	n_neighbors=7	88.89
Decision Tree	max_depth=3	83.33

Sample of Linear Regression Analysis

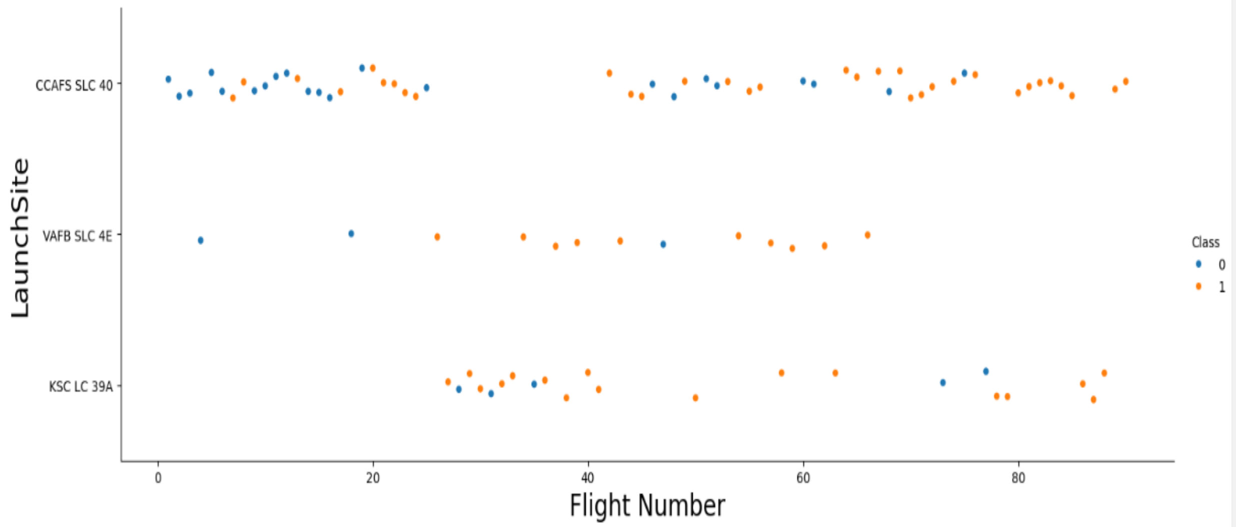


Section 2

Insights drawn from EDA

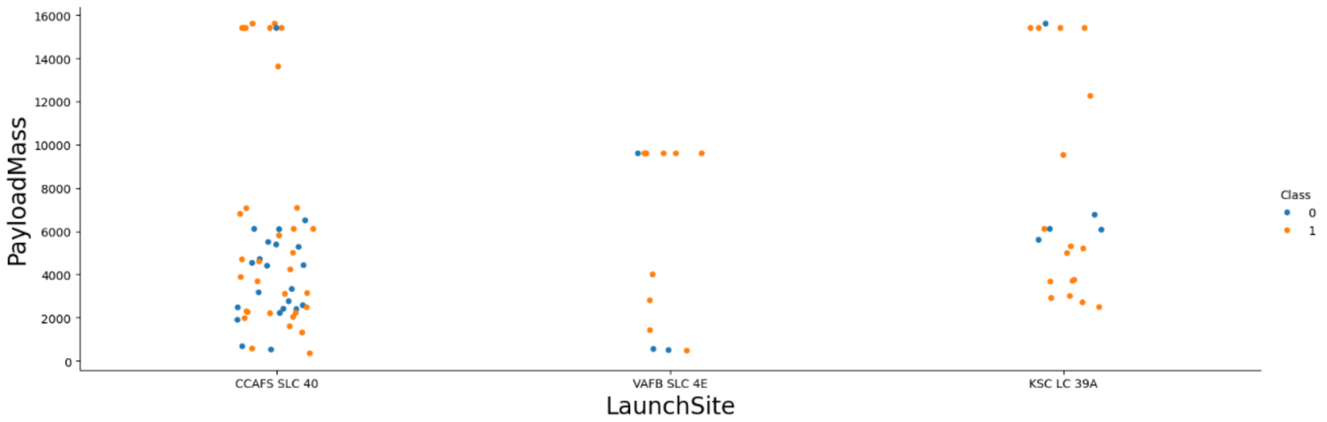
Flight Number vs. Launch Site

- This is a plot of Flight Number vs. Launch Site.
- CCAFS SLC 40 seem to have most flight numbers and its success rate increase with the flight number.
- VAFB SLC 4E have greater success rate in the 20 to 70 flight number range.
- KSC LC 39A has success rate from 30 to 40 and 75 to 85 flight number range.



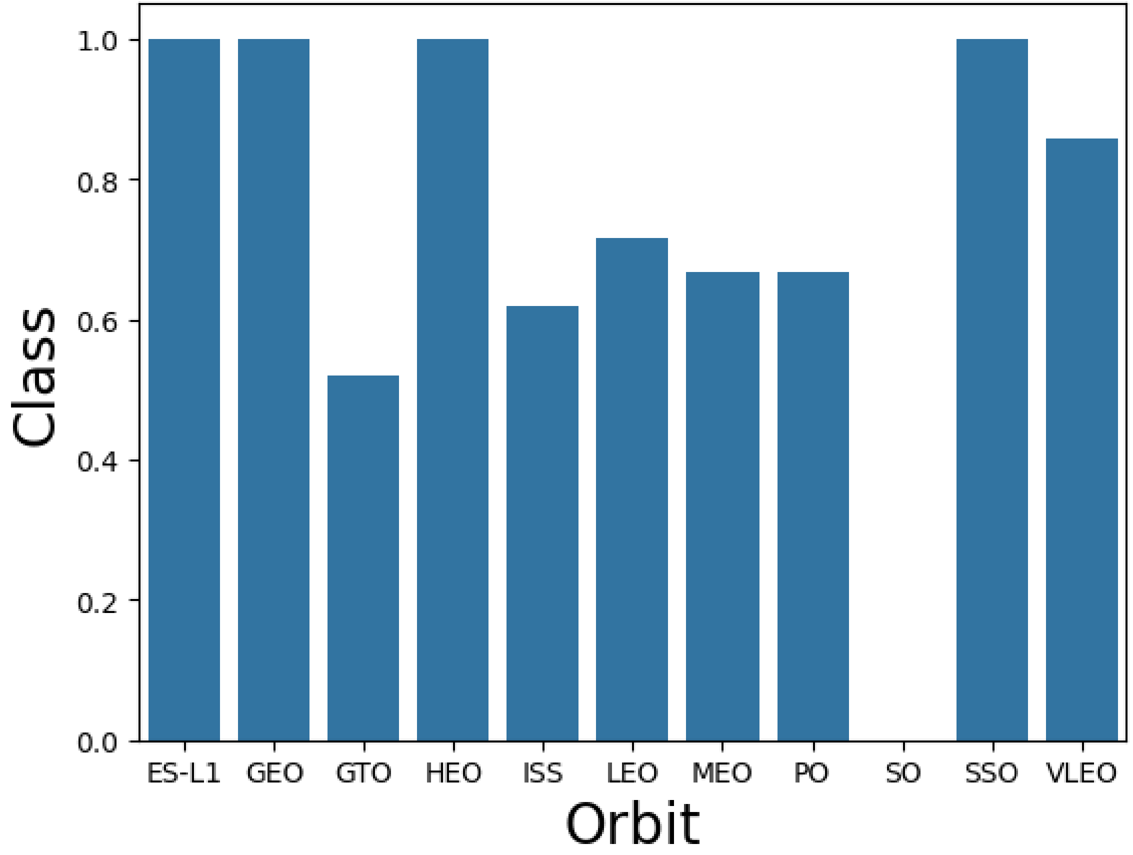
Payload vs. Launch Site

- This is a plot of Payload vs. Launch Site
- Site CCAFS SLC 40 predominately launches payloads masses in the range 0 to 8000Kg with a mixture of success.
- Site VAFB SLC 4E launches payloads masses ranging from 0 to 4000Kg and 10000Kg, a majority of which are a success.
- Site KSC LC 39A launches payloads masses ranging from 0 to 3000Kg and 16000Kg, a majority of which are a success.



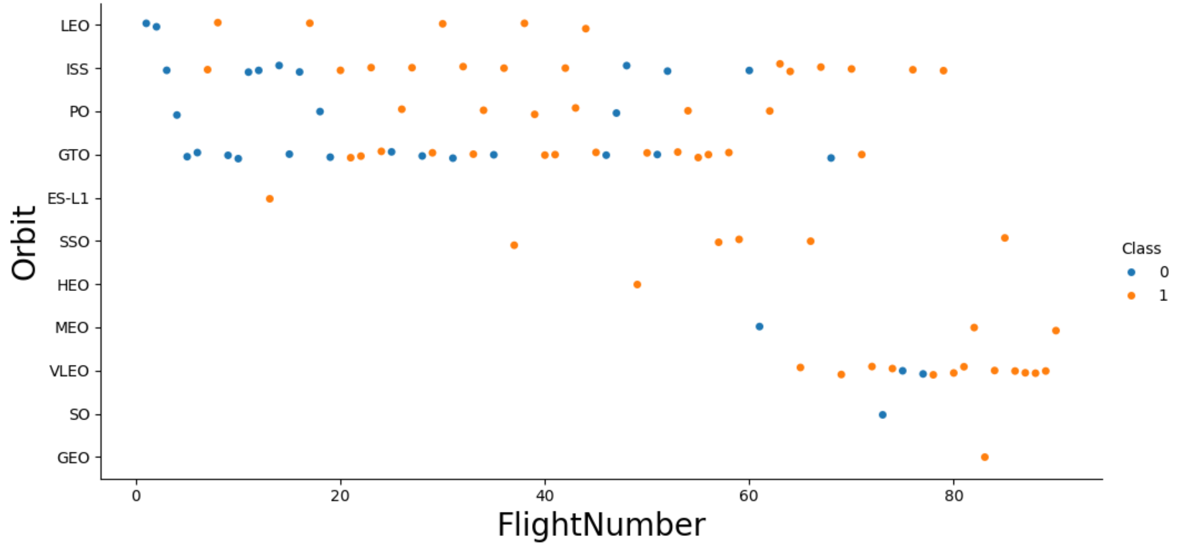
Success Rate vs. Orbit Type

- The a bar chart Illustrates the success rate of each orbit type
- The most successful orbit types are ES-L1, GEO, and SSO orbit types are the most successful.
- The least successful orbit types are GTO, ISS, MEO and PO.



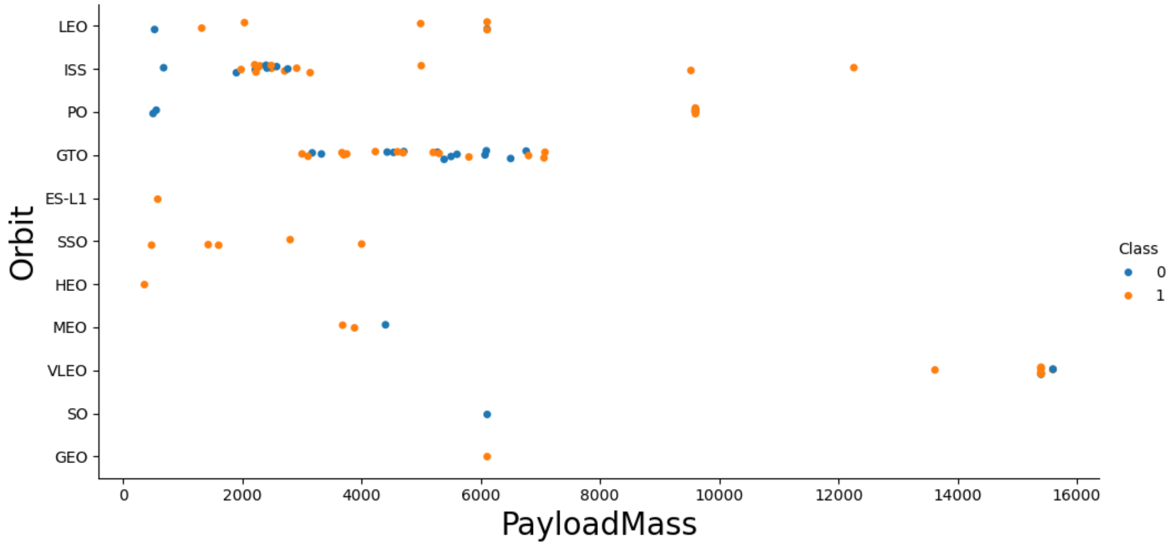
Flight Number vs. Orbit Type

- This is a plot of the Flight number vs. Orbit type
- Flight number ranging from 0 to 60 are usually utilized for Orbit Types LEO, ISS, PO and GTO. The higher Flight Numbers in the range are more successful.
- Flight number ranging from 60 to 85 are usually utilized for Orbit Types SSO, MEO and VLEO. The higher Flight Numbers in the range are more successful.



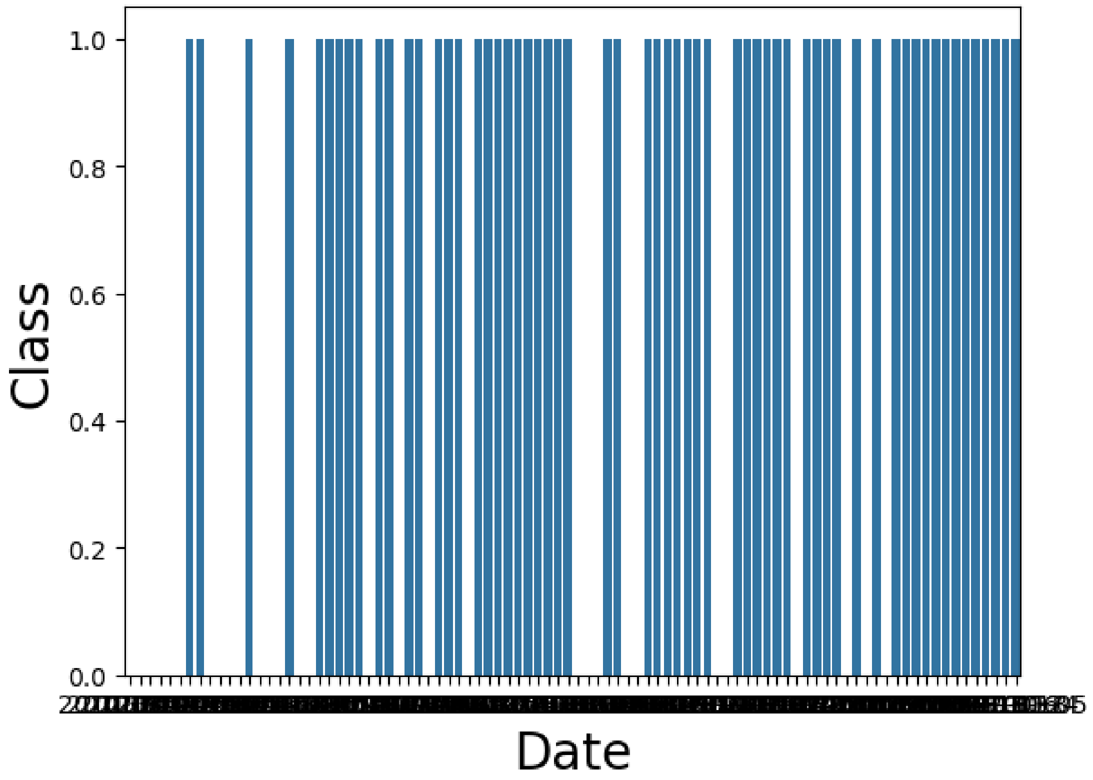
Payload vs. Orbit Type

- This is a plot of payload vs. orbit type
- Orbit type GTO has payload masses ranging from 2500Kg to 8000Kg. There is a mixture of success of launching these masses.
- Orbit type ISS has payload masses ranging from 2000Kg to 4000Kg. There seems to be greater success of launching these masses.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations



All Launch Site Names

- The screen shot to the right is a SQL query to find a list of unique launch sites.

- The query successfully returns four (4) sites, as illustrated.

Display the names of the unique launch sites in the space mission

```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

- The screen shot to the right is a SQL query that returns five (5) sites which begin with ‘CCA’.
- The query successfully returns five (5) sites, as illustrated.

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select Launch_Site from SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- The screen shot to the right is a SQL query that returns the total payload carried by boosters from NASA

- The query successfully returns a total mass of 45596 Kg, as illustrated.

Display the total payload mass carried by boosters launched by NASA (CRS) ¶

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer ="NASA (CRS)";
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The screen shot to the right is a SQL query calculates the average payload mass carried by booster version F9 v1.1

- The query successfully returns an average payload mass of 2928.4 Kg, as illustrated.

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The screen shot to the right is a SQL query returns the date of the first successful landing outcome on ground pad
- The query successfully returns the first ground pad landing on “2015-12-22”

```
%sql SELECT Landing_Outcome, Date FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' order by Date Limit 1;
```

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

Landing_Outcome	Date
Success (ground pad)	2015-12-22

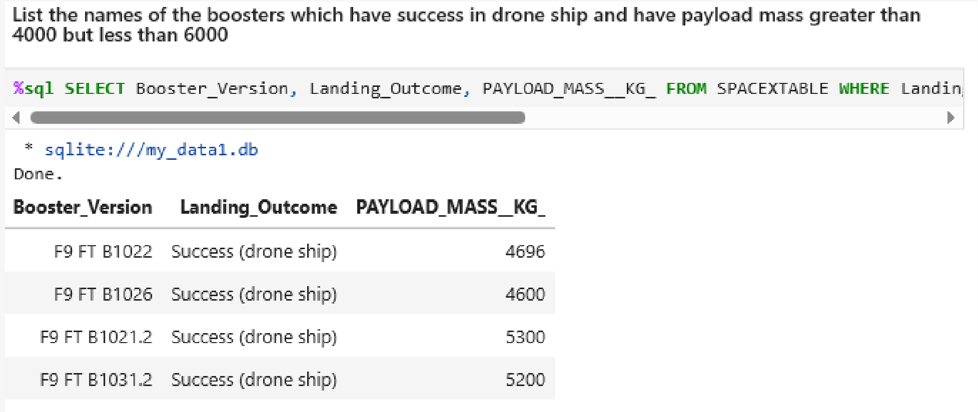
%sql SELECT Landing_Outcome, Date FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' order by Date Limit 1;

Successful Drone Ship Landing with Payload between 4000 and 6000

- The screen shot to the right is a SQL query that returns names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- The query successfully returns the following four (4) records: *F9 FT B1022 (4696 Kg), F9 FT B1026 (4600 Kg), F9 FT B1021.2 (5300 Kg) and F9 FT B1031.2 (5200 Kg)*

%sql SELECT Landing_Outcome, count(Landing_Outcome) FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success%' or Landing_Outcome LIKE 'Failure%' group by Landing_Outcome ;



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

```
%sql SELECT Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Landin
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- The screen shot to the right is a SQL query calculates the total number of successful and failure mission outcomes
- There are 38 successes and 14 successful outcomes on drone ships. There were 3 failures and 5 failures on drone ships.

%sql SELECT Landing_Outcome, count(Landing_Outcome) FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success%' or Landing_Outcome LIKE 'Failure%' group by Landing_Outcome ;

List the total number of successful and failure mission outcomes

```
Landing_Outcome LIKE 'Success%' or Landing_Outcome LIKE 'Failure%' group by Landing_Outcome ;
```

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

Landing_Outcome	count(Landing_Outcome)
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
Success	38
Success (drone ship)	14
Success (ground pad)	9

Boosters Carried Maximum Payload

- The screen shot to the right is a SQL query that returns the names of the booster versions that have carried the maximum payload mass
- The SQL query returned 12 boosters that carried the maximum payload of 15,600Kg. Four (4) of these boosters include F9 B5 B1048.4, F9 B5 B1049.4, F9 B5 B1051.3, and F9 B5 B1056.4

%sql SELECT distinct Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE where PAYLOAD_MASS__KG_= (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

```
%sql SELECT distinct Booster_Version, PAYLOAD_MASS__KG_  FROM SPACEXTABLE where PAYLOAD_MASS_
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- The screen shot to the right is a SQL query that return the failed landing outcomes on drone ship, their booster versions, and launch site names for in year 2015
- The SQL query 2 records of drone ship failures of booster versions of "F9 v1.1 B1012" and "F9 v1.1 B1015" from launch sites "CCAFS LC-40" in January and April of 2015, respectively.

%sql SELECT Landing_Outcome, Booster_Version, Launch_Site, substr(Date,0,5) as 'Year', substr(Date, 6,2) as 'Month' FROM SPACEXTABLE where Landing_Outcome ='Failure (drone ship)' and substr(Date,0,5)='2015';

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT  Landing_Outcome, Booster_Version, Launch_Site, substr(Date,0,5) as 'Year',  subs
```

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

Landing_Outcome	Booster_Version	Launch_Site	Year	Month
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015	01
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015	04

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

- The screen shot to the right is a SQL query that landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

```
%sql SELECT Mission_Outcome, Date FROM SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' order by Date Desc;
```

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

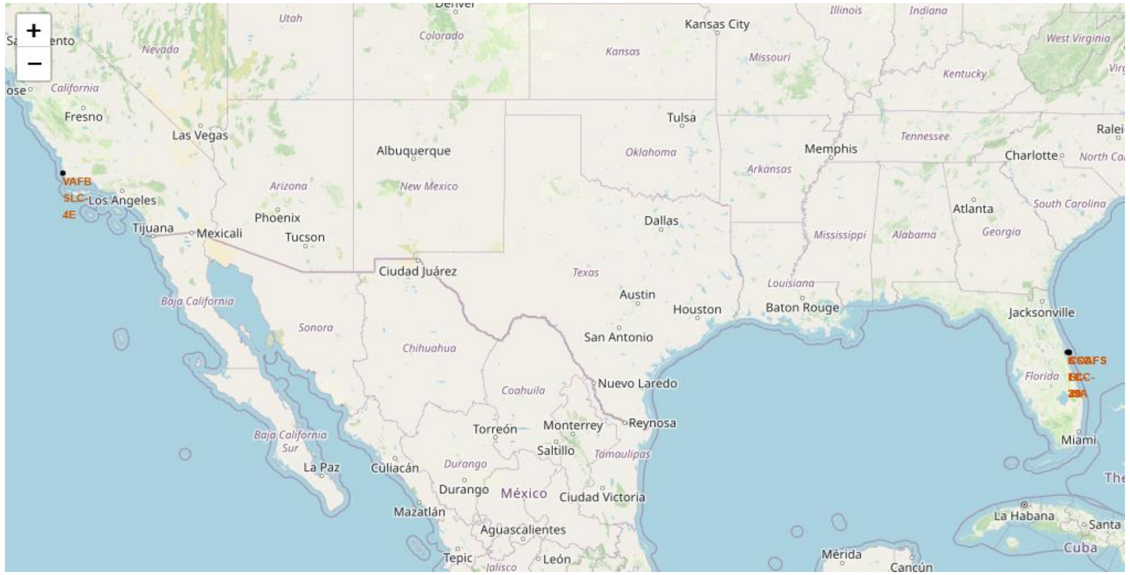
Mission_Outcome	Date
Success	2017-03-16
Success	2017-02-19
Success	2017-01-14
Success	2016-08-14
Success	2016-07-18
Success	2016-06-15
Success	2016-05-27
Success	2016-05-06
Success	2016-04-08
Success	2016-03-04
Success	2016-01-17
Success	2015-12-22
Failure (in flight)	2015-06-28
Success	2015-04-27
Success	2015-04-14
Success	2015-03-02
Success	2015-02-11
Success	2015-01-10
Success	2014-09-21
Success	2014-09-07
Success	2014-08-05
Success	2014-07-14
Success	2014-04-18
Success	2014-01-06
Success	2013-12-03
Success	2013-09-29
Success	2013-03-01
Success	2012-10-08
Success	2012-05-22
Success	2010-12-08

Section 3

Launch Sites Proximities Analysis

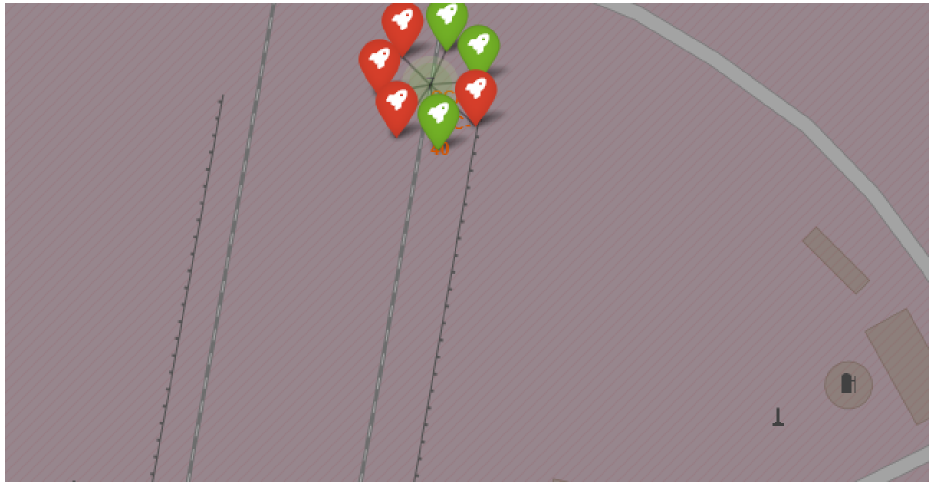
SpaceX Launch Sites

- The Map on the right displays SpaceX launch site in the continental United States.
- There are three sites located in the southern eastern and western coastlines.
- Due to the hazards of rocket launches all launch sites are located away from populated areas, but accessible by road.
- Location near coast permit the termination of rockets over water.



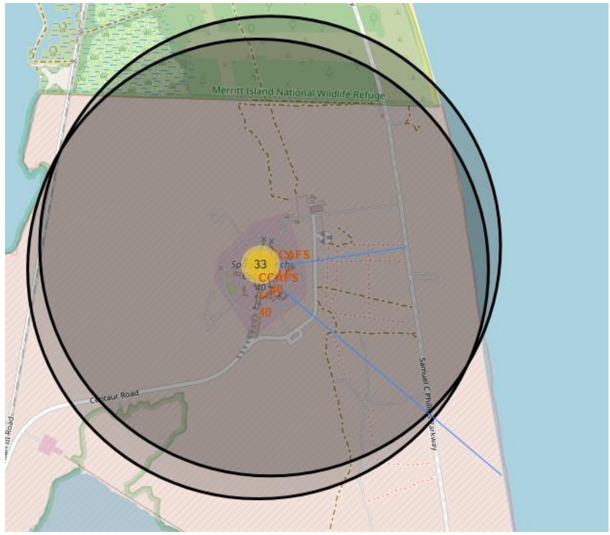
Status of Launches from West Coast Launch Sites

- The maps to the right illustrates the status of launches from the SpaceX west coast launch sites. Green markers are successful launches, and red markers are unsuccessful launches.
- The top map shows a lower number of launches with a mix rate of success.
- The bottom map shows a higher number of launches with a greater number of failures.
- The launch dates are a factor in success rates, SpaceX may still have been experimenting and designing their rockets during this period.



SpaceX Launch Site Area Geography.

- SpaceX are located on the southern east and west coasts away from populated areas and near the ocean.
- The sites are located near railways and roads to facilitate transport of materials to SpaceX factories and launch sites. Road access is also vital for SpaceX workers access factory and launch sites.
- Being near a large body of water and away from city provides ability to land and abort launches with out endangering people and other infrastructure.



Section 4

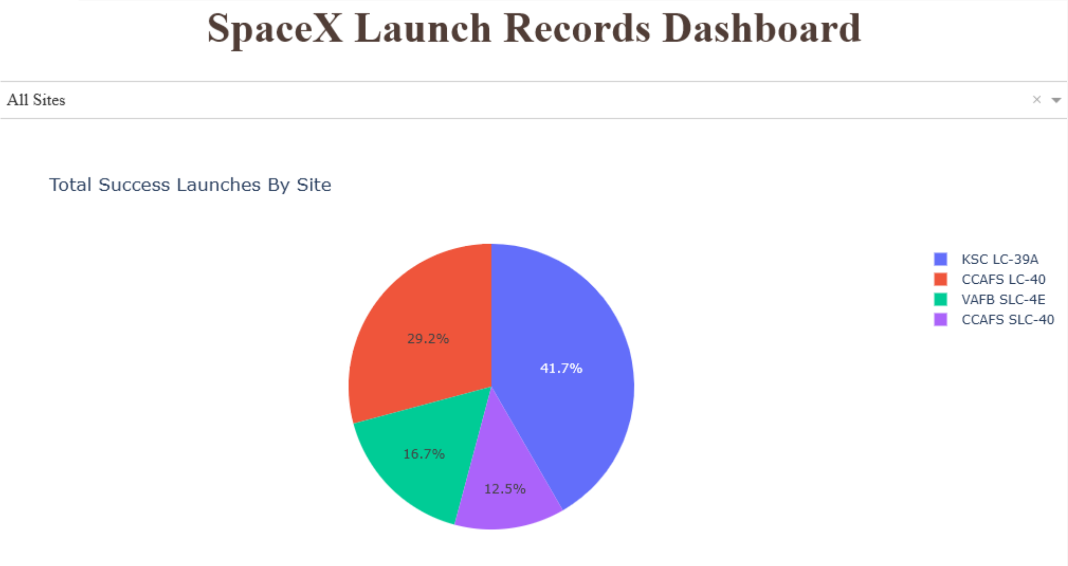
Build a Dashboard with Plotly Dash

Successful Launches From All Sites

- The Pie chart displays successful launches from all rocket site.

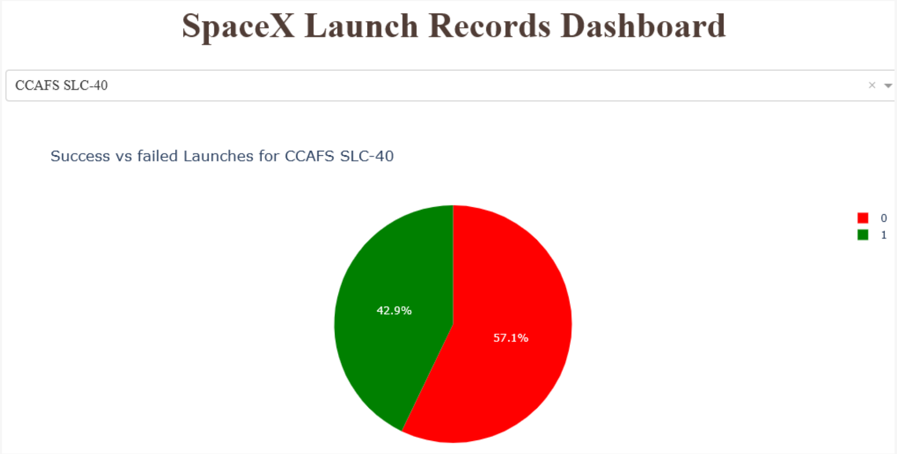
- The KSC LC-39A site has the most success launch rate of 41.7%.

- CCAFS SLC-40 has the least success launch rate of 12.5%.



Launch Site With Highest Launch Ratio

- The Pie chart shows the launch ratio of CCAFS SLC-40.
- Of all the launches from this site, 42.9.% were successful and 57.1% were unsuccessful.
- This makes it the site with highest launch ratio
- VAFB SLC-4E as the second highest launch ratio with 40% success and 60% unsuccessful launches.



Payload vs. Launch Outcome scatter plot for all sites.

- The top chart displays payload range of 2500Kg to 5000 Kg. FT and B4 rockets seems to have the most successful launches, while v1.1 seem to have failures.

- The bottom chart displays payload range of 5000Kg to 7500 Kg. FT rocket seems to have only 2 successful launches while the remaining Ft and B4 rocket launches were failures.

Section 5

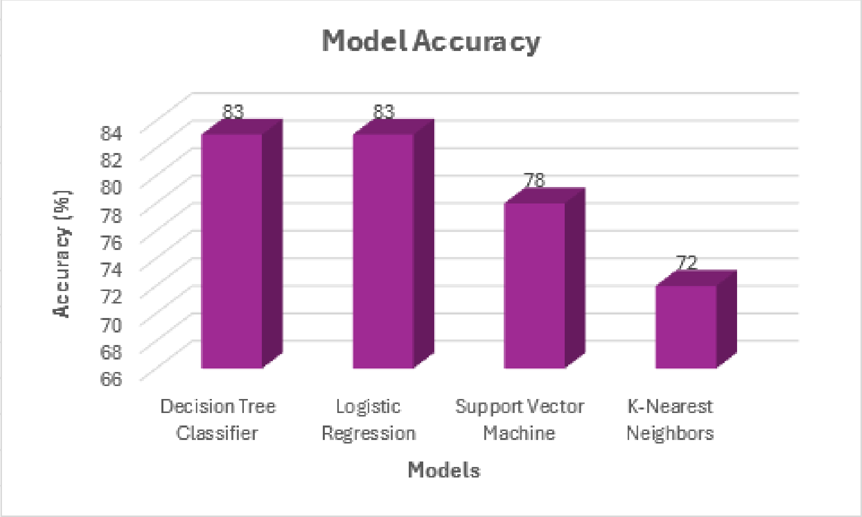
Predictive Analysis (Classification)

Classification Accuracy

- The chart on the right is list of all the models attempted in this project, the Best (hyper) Parameters determined by the process and the Model Accuracy in descending order. The graph below illustrates the project model accuracy.

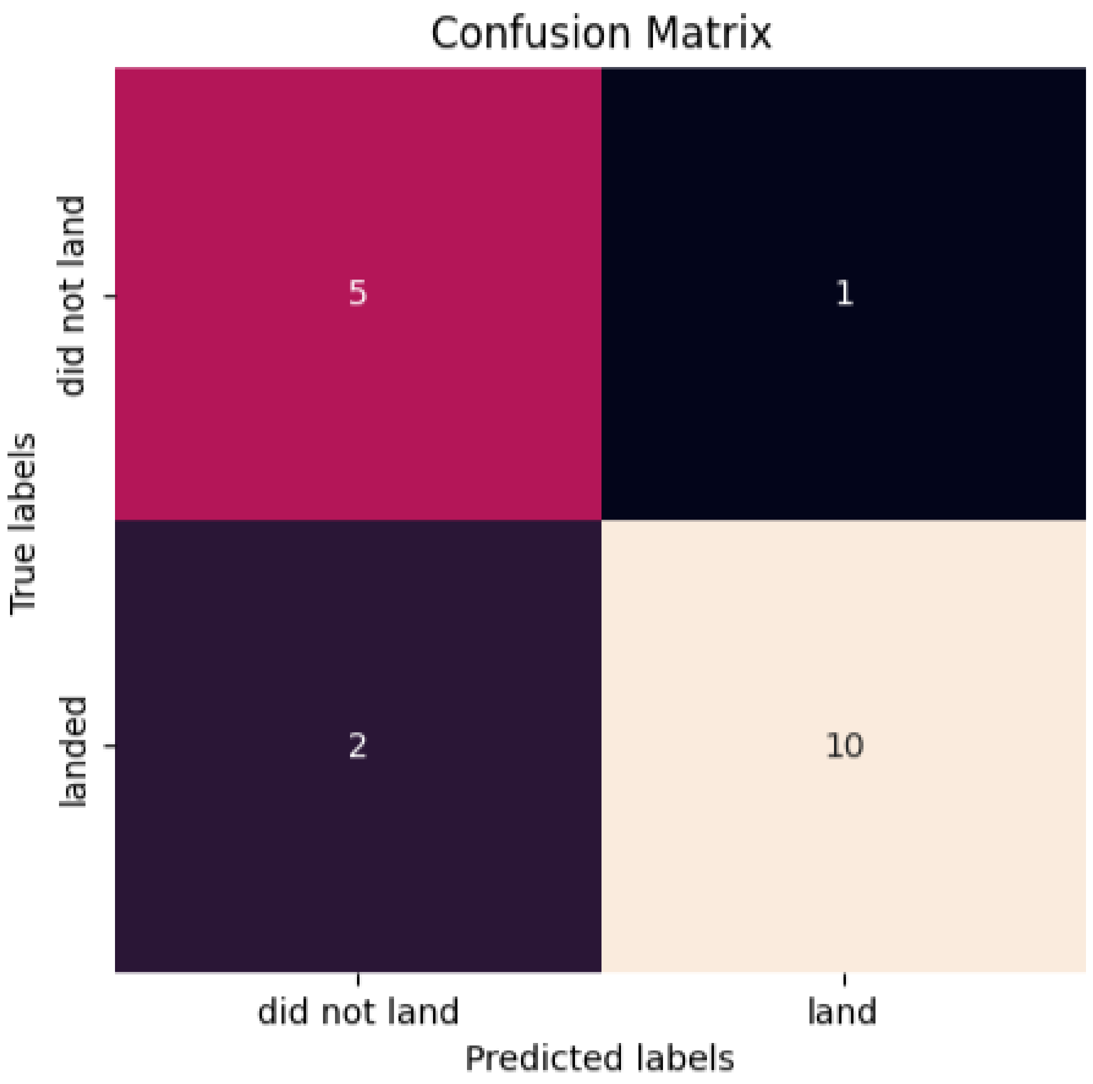
- Decision Tree Classifier and Logistic Regression models both had the best accuracy rate of 83%. K-Nearest Neighbors model has the worst accuracy rate of 72%

Model	Best Parameters	Accuracy (%)
Decision Tree Classifier	{'criterion': 'gini', 'max_depth': 4, 'min_samples_leaf': 2, 'min_samples_split': 10}	83
Logistic Regression	{'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}	83
Support Vector Machine	{'C': 1, 'kernel': 'rbf', 'gamma': 0.001}	78
K-Nearest Neighbors	{'algorithm': 'auto', 'n_neighbors': 3, 'p': 2}	72



Confusion Matrix

- The Logistic Regression and Decision Tree Classifier models both were the best performing models.
- The right is the **Logistic Regression** Confusion Matrix.
- The model correctly predicted 5 cases where the rocket did not land and 10 cases where it did land. Conversely, the model incorrectly predicted 1 rockets as landed when they didn't, and missed 2 actual landing by predicting it as did not land.



Conclusions

- **Decision Tree and Logistic Regression Perform Best**: Both achieved the highest test set accuracy (83%), indicating strong predictive performance for Falcon 9 landing success.

- **SVM Shows Overfitting**: Despite the highest cross-validation accuracy (83.2%), SVM's test accuracy dropped to 78%, suggesting overfitting.

- **KNN Underperforms:** KNN had the lowest test accuracy (72%), making it the least effective model for this task.

- **Data Standardization is Critical**: Standardizing the feature set (X) was necessary to ensure consistent model performance across algorithms.

- **Hyperparameter Tuning Improves Results**: GridSearchCV identified optimal parameters, enhancing model accuracy for all algorithms.

Recommendation to potentially improve model perform:

- Incorporate Additional Features: Include more relevant features (e.g., weather conditions, rocket configuration details) to capture factors influencing landing success.

- Increase Dataset Size: Collect more data to expand the training set, reducing overfitting and improving model generalization. ance

Thank you!