

IBM Data Science: SpaceX Launch Data Project

Nabila Shaikh



Outline

01

Executive
Summary

04

Results

02

Introduction

05

Conclusion

03

Methodology

06

Appendix



Executive summary

Methodology:

Data collection

Data wrangling

EDA with data visualization

EDA with SQL

Building an interactive map with Folium

Building a Dashboard with Plotly Dash

Predictive analysis (Classification)

Results:

Exploratory Data Analysis

Interactive Visual Analytics

Predictive Analysis

Project Background

SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each. Much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

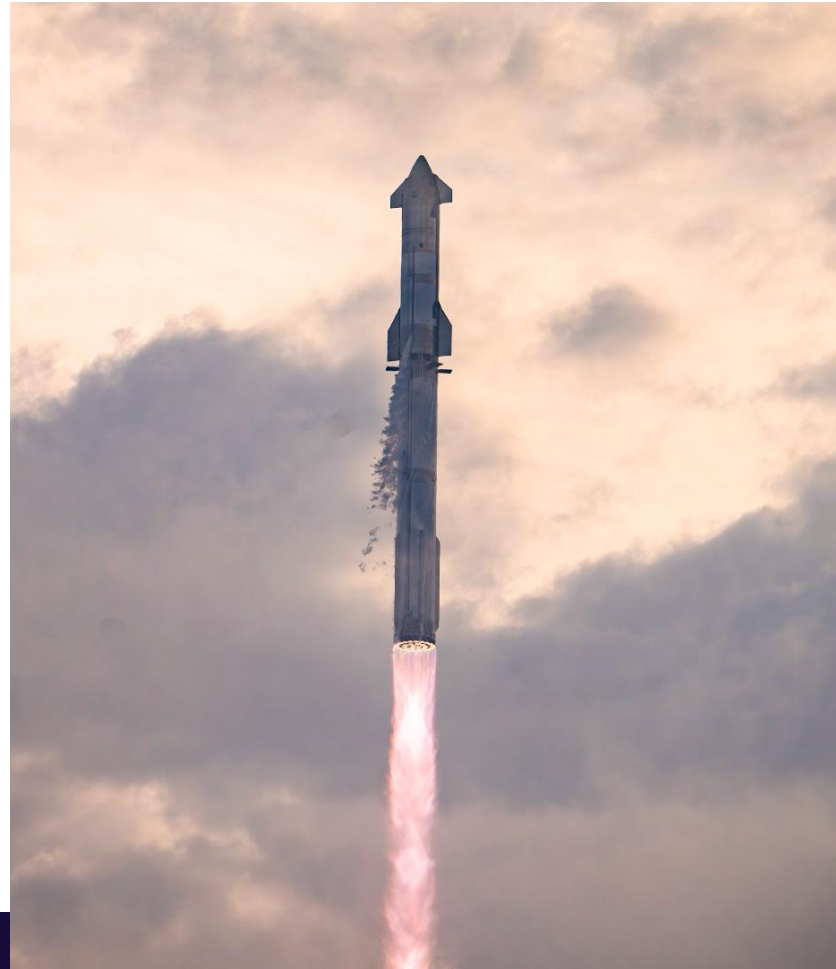
Key Insights:

- How factors like payload mass, launch site, number of flights, and orbits affect first-stage landing success
- The rate of successful landings over time
- What is the best predictive model for successful landing



Methodology

Methods used to perform the data analysis



Methodology

- Perform Data Collection by using SpaceX API and web scraping
- Perform Data Wrangling – by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- Perform Exploratory Data Analysis (EDA) using data visualization and SQL
- Perform Interactive Visual Analytics using Folium and Plotly Dash
- Perform Predictive Analysis using Classification Models

Data Collection - API

- Request and parse the SpaceX launch data
- Filter the data frame to only include Falcon 9 launches
- Dealing with Missing Values



GitHub URL: <https://github.com/NS719/Applied-Data-Science-Capstone-SpaceX-Data/blob/main/Space%20X%20Falcon%209%20-%20Data%20Collection%20%26%20Wrangling.ipynb>

Data Collection – Web Scrapping

- Request the Falcon 9 Launch Wiki page from its URL
- Create a BeautifulSoup object from the HTML response
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

GitHub URL: <https://github.com/NS719/Applied-Data-Science-Capstone-SpaceX-Data/blob/main/Space%20X%20Falcon%209%20-%20WebScrapping.ipynb>

Webscrap Launch
Records HTML
table using
BeautifulSoup

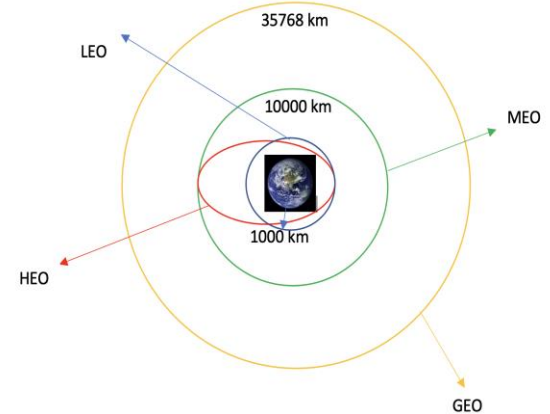


Convert to Pandas
Dataframe

Data Wrangling

The data contains several Space X launch facilities:
Cape Canaveral Space Launch Complex 40 VAFB SLC
4E , Vandenberg Air Force Base Space Launch
Complex 4E (SLC-4E), Kennedy Space Center Launch
Complex 39A KSC LC 39A

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column



Some common orbit types used for the launches

GitHub URL: <https://github.com/NS719/Applied-Data-Science-Capstone-SpaceX-Data/blob/main/Space%20X%20Falcon%209%20-%20Data%20Wrangling.ipynb>

EDA with Data Visualization

Plotting the following charts to perform Exploratory Data Analysis and prepare data for Feature Engineering

Scatter Plots

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Flight Number vs. Orbit type
- Payload Mass vs. Orbit type

Bar Plot

- Success Rate vs. Orbit Type

Line Plot

- Launch success vs. Year

Selecting the features that will be used in success prediction

Create dummy variables to categorical columns

Apply OneHotEncoder

Assign the value to the variable

Display the results

EDA with SQL

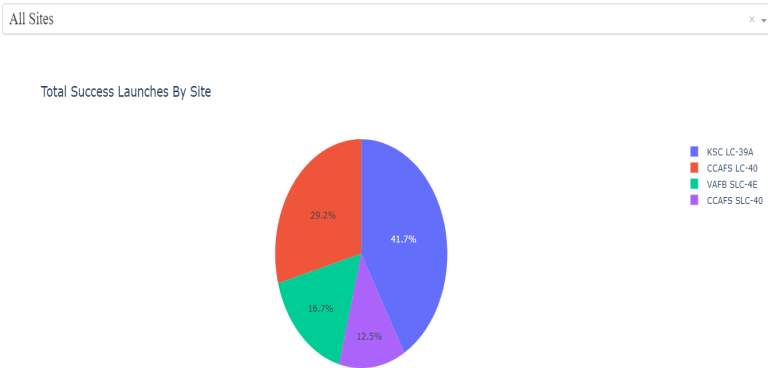
- Connect to the database – load the SQL extension and establish a connection
- Execute following SQL queries:
 - Display the names of the unique launch sites in the space mission
 - Display 5 records where launch sites begin with the string 'CCA'
 - Display the total payload mass carried by boosters launched by NASA (CRS)
 - Display average payload mass carried by booster version F9 v1.1
 - List the date when the first successful landing outcome in ground pad was achieved.
 - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
 - List the total number of successful and failure mission outcomes
 - List the names of the booster versions which have carried the maximum payload mass. Use a subquery
 - 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.
 - Rank 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.

Interactive Map with Folium

- Mark all launch sites on a map
 - Created a **blue** circle at NASA Johnson Space Center's coordinate with a popup label showing its name
 - Created **red** circle on Launch Site locations with Launch Site name as a popup label
- Mark the success/failed launches for each site on the map
 - Successful launches shown with **green** marker and unsuccessful launches shown with **red** marker at each launch site to easily identify which launch sites have relatively high success rates.
- Calculate the distances between a launch site to its proximities
 - Created markers showing distance lines from launch site CCAFS SLC-40 to closest city, railway and highway.

Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



- **Drop-down (input) Component** - Select a Launch Site (or All Sites)
- **Pie chart (output) component** - Shows pie chart displaying successful and unsuccessful launches based on selected site
- **Range slider (input) component** - Select Payload range
- **Scatter chart (output) component** - Shows scatter plot of Payload Mass vs. Success Rate by Booster Version

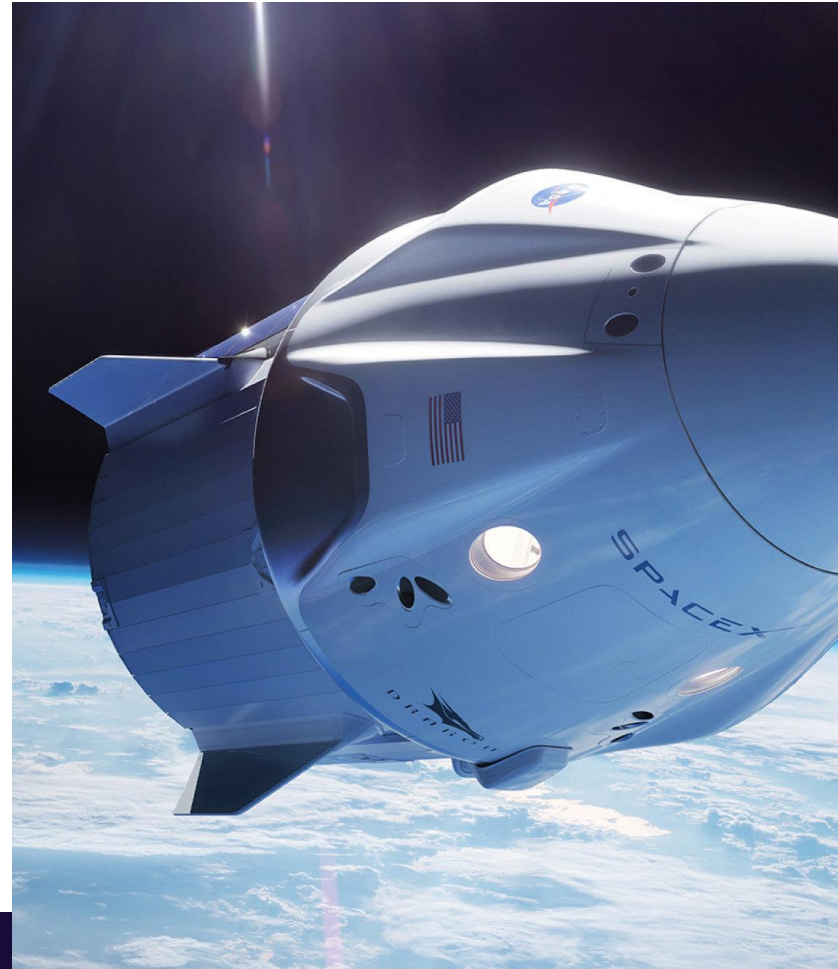
Predictive Analysis (Classification)

- Load the data.
- Create a NumPy array from the column Class in data.
- Standardize the data in X then reassign and transform it.
- Split the data X and Y into training and test data.
- Perform methods of logistic regression, support vector machine, decision tree classifier, k nearest neighbors then create a GridSearchCV object with cv = 10 for each. Fit the objects to find the best parameters from the dictionary parameters.
- Calculate the accuracy on the test data for all.
- Find the method performs best.



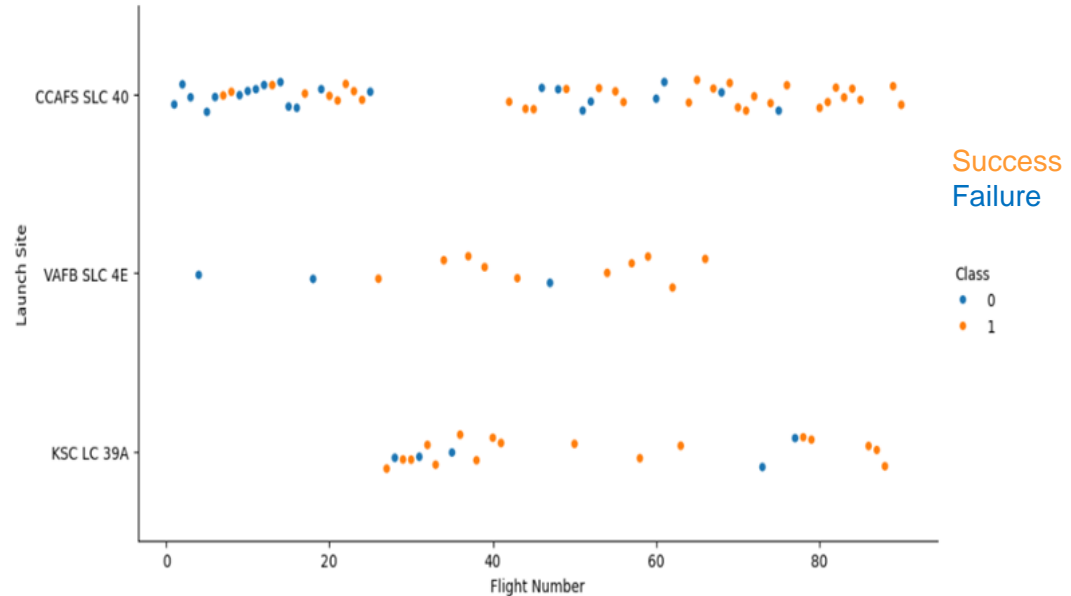
Insights from EDA

Exploratory Data Analysis with Visualization



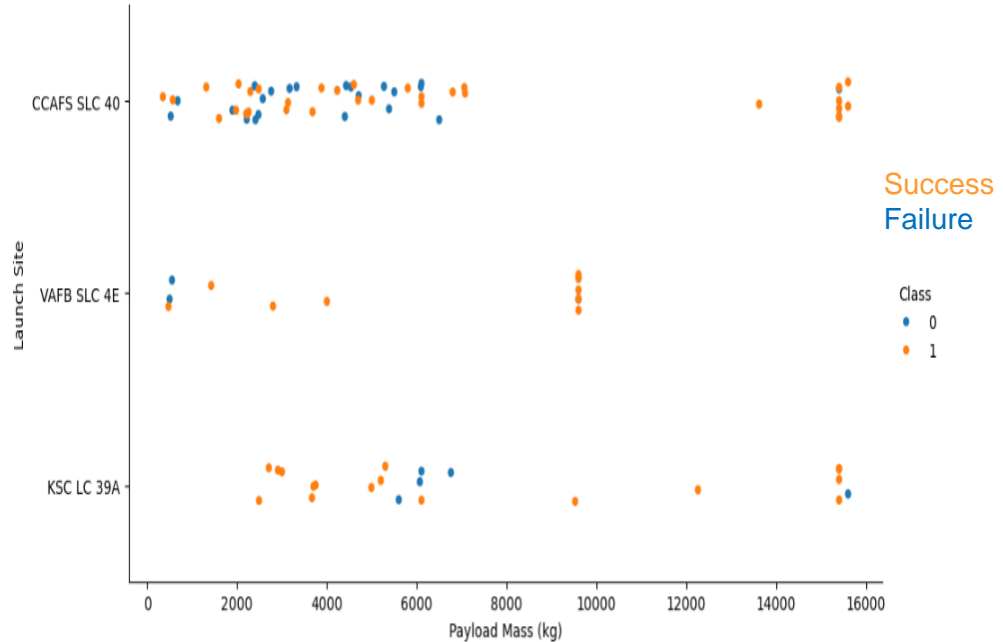
Flight Number vs. Launch Site

- Most launches occurred from CCAFS SLC 40 Launch Site
- Launch Sites VAFB SLC 4E and KSC LC 39A show high success rates per launch.

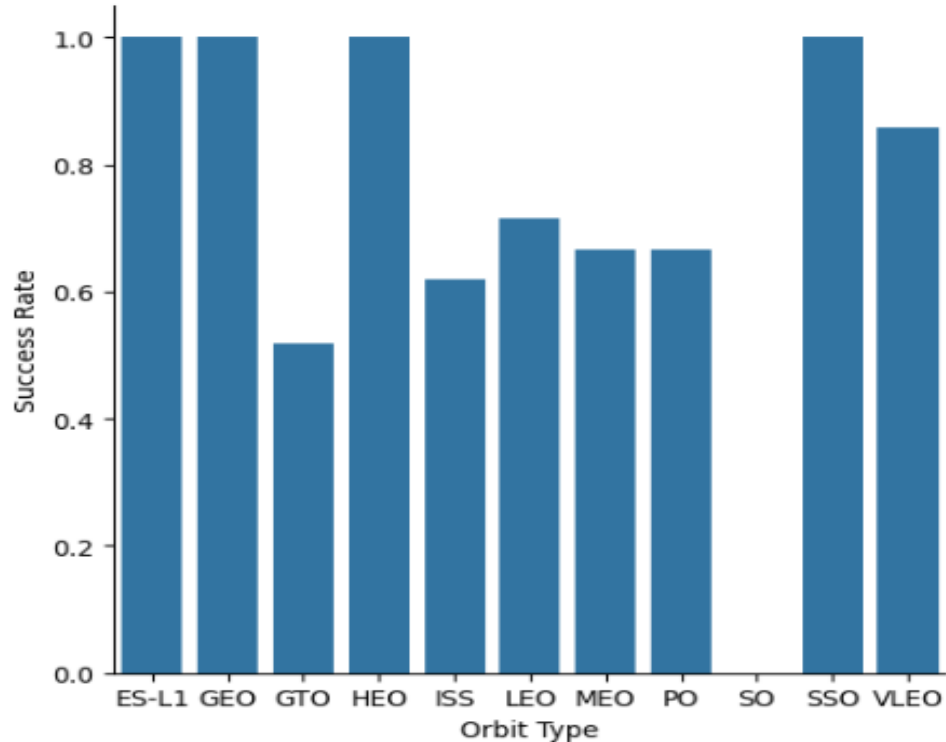


Launch Site vs. Payload

- VAFB-SLC Launch Site there are no rockets launched for Heavy Payload Mass (greater than 10000).



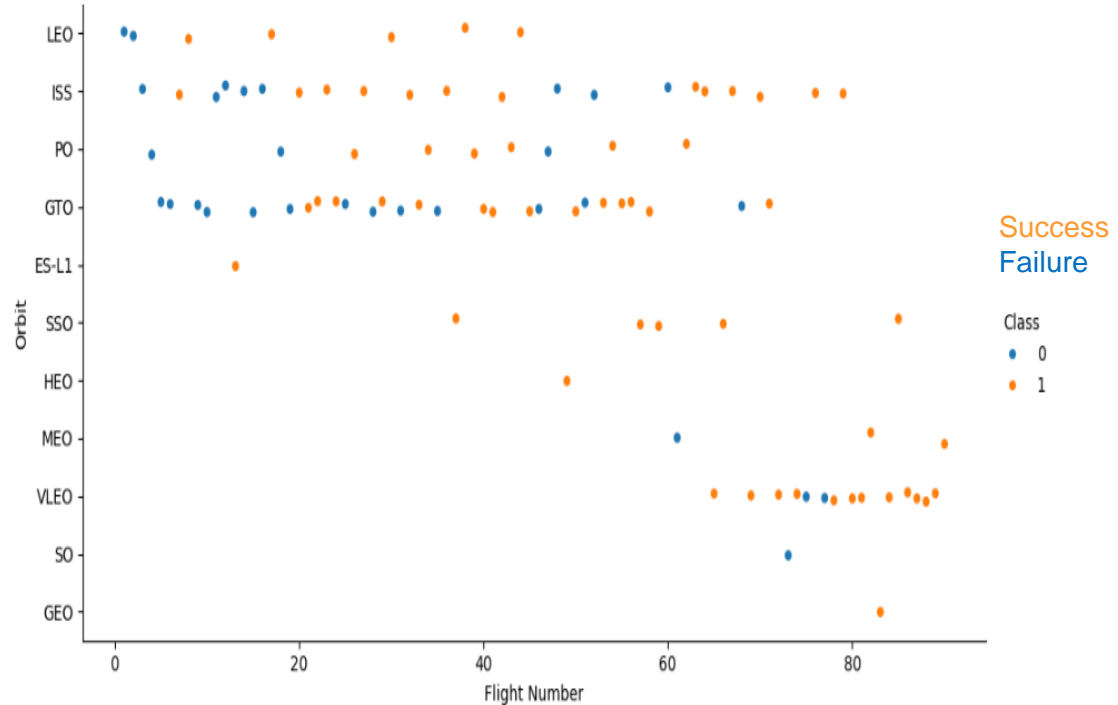
Success Rate vs. Orbit Type



- **ES-L1, GEO, HEO and SSO** orbits show maximum success rate.
- **GTO, ISS, LEO, MEO, PO** orbits have moderate rate of success.
- **SO** orbit has had zero rate of success.

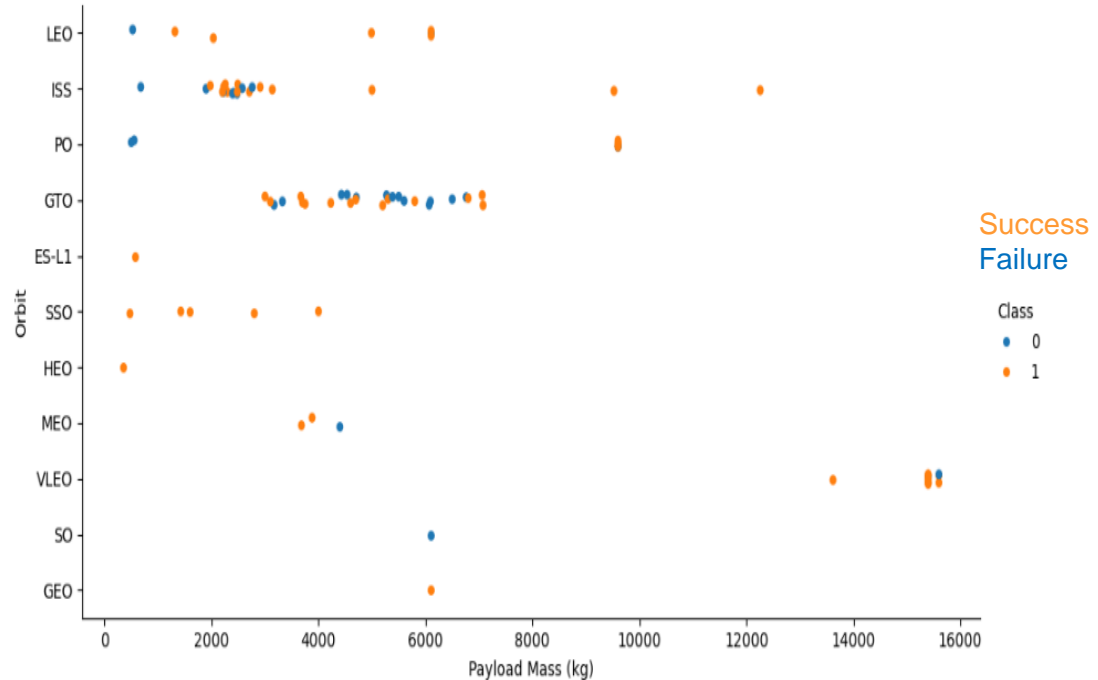
Flight Number vs. Orbit Type

- It is observed that in the LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.

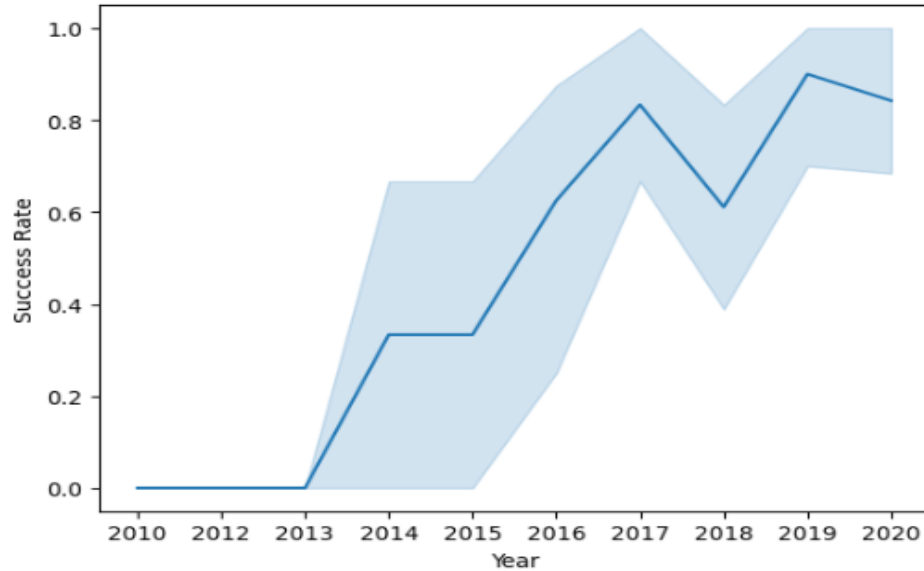


Payload vs. Orbit Type

- With **heavy payloads** the successful landing or positive landing rate are more for **Polar, LEO and ISS** orbit types.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both present.



Launch Success Yearly Trend



- It is observed that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

```
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

* sqlite:///my_data1.db

Using SQL

.....

Launch_Site

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A


CCAFS SLC-40

Launch Site Names Begin with 'CCA'

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

Using SQL



.....

Launch_Site

CCAFS LC-40

CCAFS LC-40

CCAFS LC-40

CCAFS LC-40

CCAFS LC-40

Total Payload Mass & Average Payload Mass by F9 v1.1

.....

SUM(PAYLOAD_MASS_KG_)

45596

AVG(PAYLOAD_MASS_KG_)

2928.4

First Successful
Ground Landing
Date

MIN(Date)

2015-12-22

Successful Drone Ship Landing
with Payload between 4000
and 6000

Booster_Version

F9 FT B1021.1

F9 FT B1022

F9 FT B1023.1

F9 FT B1026

F9 FT B1029.1

F9 FT B1021.2

F9 FT B1029.2

F9 FT B1036.1

F9 FT B1038.1

F9 B4 B1041.1

F9 FT B1031.2

F9 B4 B1042.1

F9 B4 B1045.1

F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	TOTAL_NUMBER
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

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

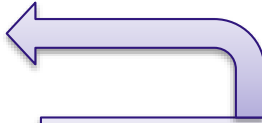
F9 B5 B1060.2

F9 B5 B1058.3

F9 B5 B1051.6

F9 B5 B1060.3

F9 B5 B1049.7



Names of Boosters that
have carried the
Maximum Payload
Mass

2015 Launch Records

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Records displaying the month, booster versions, launch site and failure landing outcomes in drone ship - in the year 2015.

Rank Landing Outcomes

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

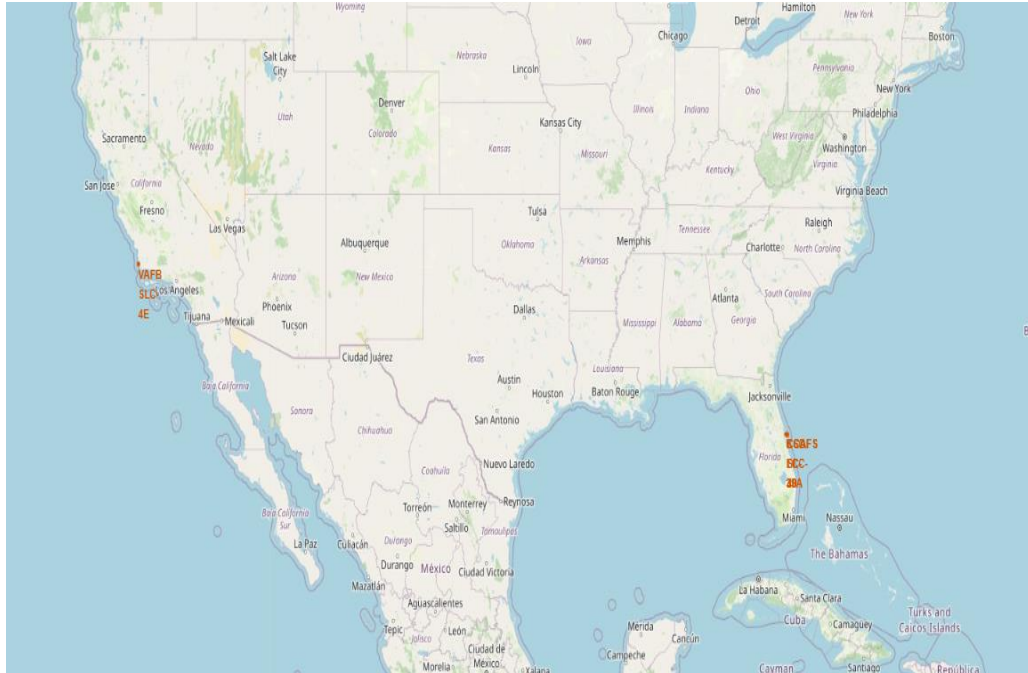
Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the dates *2010-06-04 and 2017-03-20*

Launch Sites Proximities Analysis

With Folium



All Launch Sites Location



Launch sites seem to be somewhat in proximity to the Equator line.
All launch sites are however in very close proximity to the coast

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

Launch Outcomes



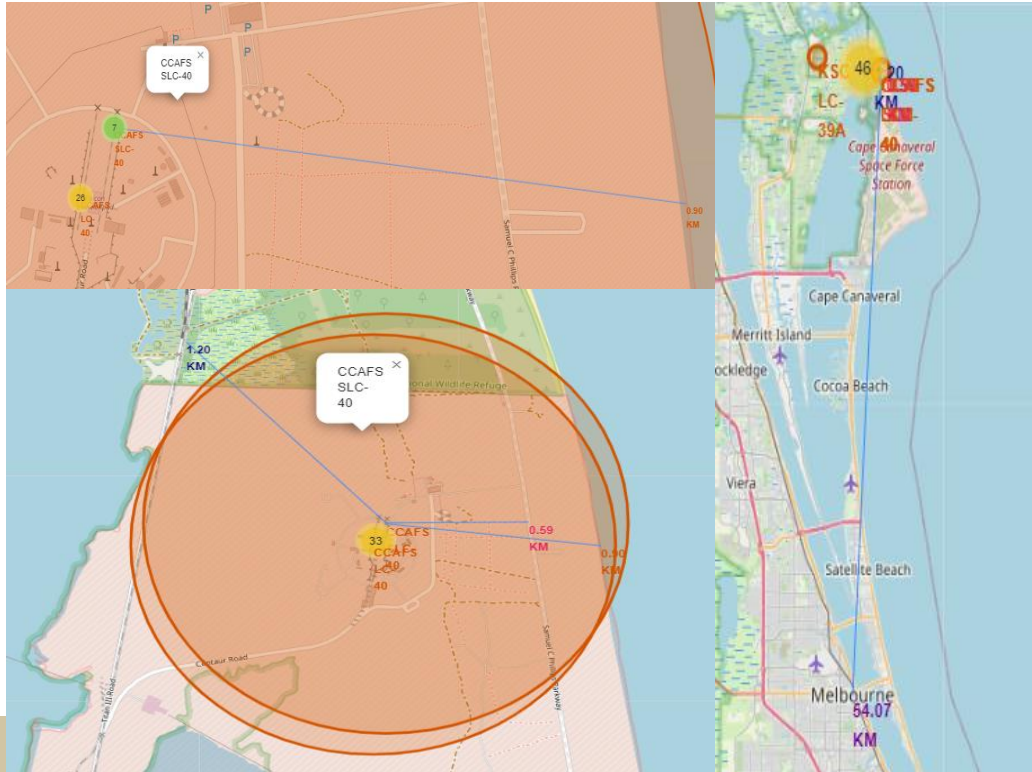
Green - Successful Launches

Red - Unsuccessful Launches

CCAFS SLC-40 Launch site has shown majority unsuccessful launches, and following it is the VAFB SLC-4E Launch Site.

KSC LC-39A Launch Site has the highest success rate of launches.

Launch Site Proximities



Proximity of CCAFS SLC-40 Launch Site to:

Closest coastline = 0.89 km

Nearest Railway = 1.20 km

Nearest Highway = 0.59 km

Nearest City = 54.07 km

Launch Sites seem to be placed near the coast, away from city, minimizing the risk of having any debris dropping or exploding near people.

Interactive Dashboard

With Plotly Dash



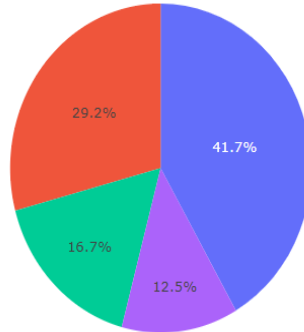
Launch Success Count For All Sites

SpaceX Launch Records Dashboard

All Sites



Total Success Launches By Site



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

KSC LC-39A has the Highest Launch Success = 41.7%

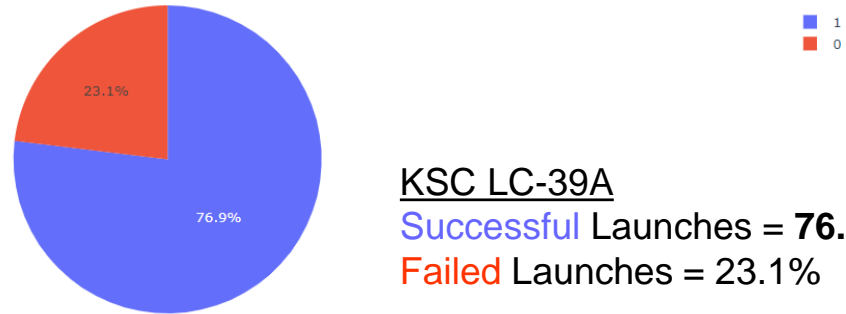
Launch Site With Highest Launch Success Ratio

SpaceX Launch Records Dashboard

KSC LC-39A

× ▾

Total Success Launches for site KSC LC-39A



KSC LC-39A

Successful Launches = **76.9%**

Failed Launches = 23.1%

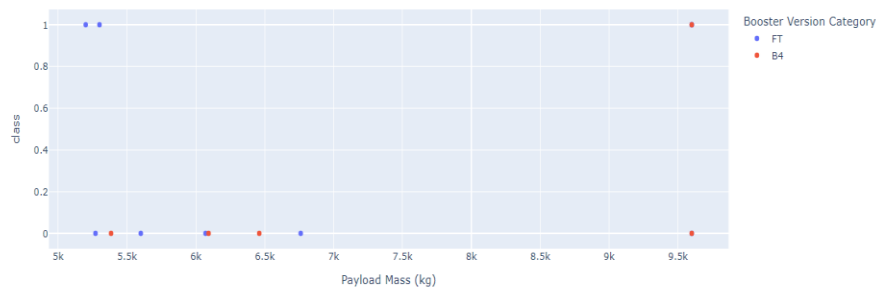
Payload vs. Launch Outcome

Payload Range
0 to 5K Kg

Payload range (Kg):



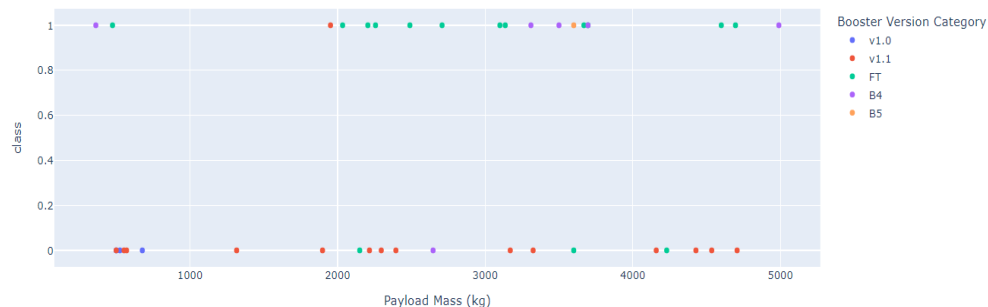
Correlation between Payload and Success for all Sites



Payload range (Kg):



Correlation between Payload and Success for all Sites



Payload Range
5K to 10K Kg

Predictive Analysis

Classification



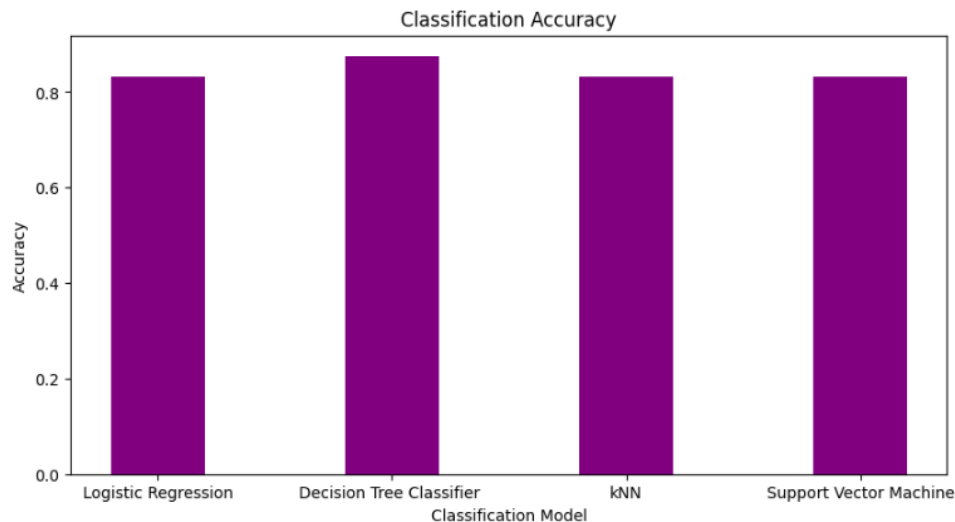
Classification Accuracy

```
methods = {'KNN':knn_cv.best_score_, 'Decision Tree Classifier':tree_cv.best_score_, 'Support Vector Machine':svm_cv.best_score_, 'Logistic Regression':logreg_cv.best_score_}
bestmethod = max(methods, key=methods.get)
print('Best Method is',bestmethod,'with accuracy',methods[bestmethod])
if bestmethod == 'Decision Tree Classifier':
    print('Best Parameter is :',tree_cv.best_params_)
if bestmethod == 'Support Vector Machine':
    print('Best Parameter is :',svm_cv.best_params_)
if bestmethod == 'KNN':
    print('Best Parameter is :',knn_cv.best_params_)
if bestmethod == 'Logistic Regression':
    print('Best Parameter is :',logreg_cv.best_params_)
```

Best Method is Decision Tree Classifier with accuracy 0.875

Best Parameter is : {'criterion': 'entropy', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}

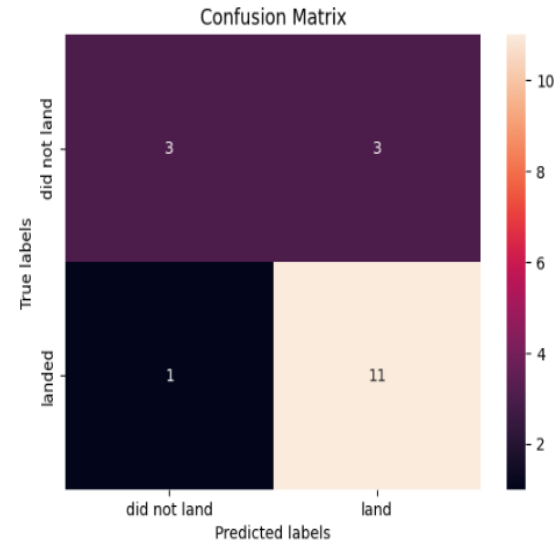
- The classification method that is calculated to work best is the Decision Tree Classifier.
- The other models perform just as well relatively, as observed from the graph.
- Perhaps a larger dataset would provide more concrete results with distinct accuracies.



Confusion Matrix

- A confusion matrix is a table that is used to define the performance of a classification algorithm.
- Precision = 0.5000 – Measure of how accurate a model's positive predictions are.
- Accuracy = 0.7778 – Measure the performance of the model.
- Sensitivity = 0.7500 – Measures the effectiveness of a classification model in identifying all relevant instances from a dataset.
- F1 Score = 0.6000 – Evaluate the overall performance of a classification model.
- Specificity = 0.7857 – Measures the ability of a model to correctly identify negative instances.

All the classification models gave the same confusion matrix.



Conclusions



Model Performance

The best performing classification model for the machine learning program with this dataset is found to be the **Decision Tree Classification Model**.



Location

Launch site near Earth's equator can take optimum advantage of the Earth's substantial rotational speed, increase fuel efficiency. Coastal site also prevents damage from debris.



Factors

Some of the factors that were observed to affect the rate of success were:
Launch Orbits
Payload Mass
Launch Site



Success Insights

- **KSC LC-39A** Launch Site showed the highest Launch Success Rate.
- **ES-L1, GEO, HEO and SSO** orbits show maximum success rate.
- Payload Mass in the **lower ranges** are preferable for success.



THANK YOU!