IBM Developer SKILLS NETWORK

Winning Space Race with Data Science

Surya D
14-09-2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Summary of Methodologies:

Data Preprocessing: Standardized features and split data into training (80%) and testing (20%) sets.
Model Training: Used **GridSearchCV** for hyperparameter tuning and cross-validation for four models: **Logistic Regression**, **SVM**, **Decision Tree**, and **KNN**.
Evaluation: Tested models for accuracy and analyzed confusion matrices.

Summary of Results:

Logistic Regression: Best accuracy – **83.33%**.
SVM (linear kernel): Best accuracy – **83.33%**.
Decision Tree: Lower accuracy – **73.33%**.
KNN: Accuracy – **77%**.

Introduction

- Project background and context
- Problems you want to find answers

Project Background and Context:
SpaceX's Falcon 9 rocket reusability reduces launch costs significantly. This project aims to predict whether the first stage will land successfully, using historical SpaceX launch data, including factors like launch site, payload mass, and orbit type.

Problems You Want to Find Answers To:
Can we predict the Falcon 9 first stage landing success?
Which factors influence landing success (e.g., payload mass, launch site)?
Which model (Logistic Regression, SVM, etc.) performs best?
What are the best hyperparameters for each model?

Section 1

Methodology

Methodology

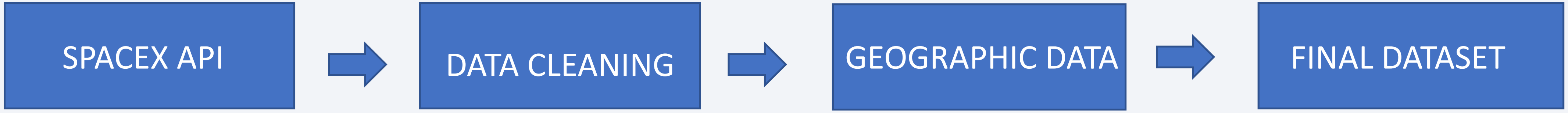
Executive Summary

- Data collection methodology:
  - Data was collected using web scarping.
- Perform data wrangling
  - Exploratory Data Analysis to find some patterns in the data and determine what would be the label for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Exploratory Data Analysis using python library (Matplotlib, Seaborn) and using SQL. Feature Engineering
- Perform interactive visual analytics using Folium and Plotly Dash
  - Use Folium to view previously observed correlations.
- Perform predictive analysis using classification models
  - Creating machine learning models to predict if the firtst stage will land.

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection Process (in Short):
SpaceX API: Data on Falcon 9 launches (e.g., launch site, payload mass, orbit, landing success) was fetched from the SpaceX API.
Geographic Data: Latitude and longitude of launch sites were added to the dataset for location analysis.
Data Cleaning: The raw data was cleaned and formatted for consistency.
Combined Dataset: Data from multiple sources (API + geographic) was merged to form a complete dataset.



Data Collection - SpaceX API

Github link:
https://github.com/CmeDinfinity/Capstone_spaxeX/blob/main/Data%20Collection%20API.ipynb

SpaceX REST API

GET /launches (Launch Details)
Provides launch site, date, success

GET /rockets (Rocket Info)
Provides details about the rocket

GET /payloads (Payload Info)
Provides mass and orbit type

GET /launchpads (Geographic Info)
Provides latitude and longitude

Aggregated Dataset
Final cleaned and combined data

Data Collection - Scraping

Github link :

https://github.com/CmeDinfinity/Capstone_spaxeX/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

1. Identify Target Website
2. Send HTTP Request (using requests)
3. Parse HTML (using BeautifulSoup)
4. Extract Data (from HTML elements)
5. Clean and Structure Data
6. Store Data (in DataFrame or CSV)

Data Wrangling

- **Load Data**: Load data from APIs and web scraping into a DataFrame.
- **Handle Missing Data**: Identify and fill or remove missing values.
- **Clean Data**: Remove unnecessary columns and correct data types.
- **Feature Engineering**: Create new features, such as one-hot encoding for categorical variables.
- **Standardization**: Scale numerical data using **StandardScaler**.
- **Data Integration**: Merge data from multiple sources into a cohesive dataset.

GitHub : https://github.com/CmeDinfinity/Capstone_spaxeX/blob/main/Data%20Wrangling.ipynb

Load Data (API/Web Scraping)

Identify and Handle Missing Data

Data Cleaning (Remove Noise)

Feature Engineering (One-Hot)

Standardize Data (Numerical)

Data Integration (Merge Sources)

Final Dataset Ready

EDA with Data Visualization

•**Scatter Plot (Flight Number vs. Launch Site)**: Show landing success based on flight number and launch site.

•**Scatter Plot (Payload vs. Launch Site)**: Analyze how payload mass affects landing success at each site.

•**Bar Plot (Success Rate per Orbit Type)**: Visualize success rate for each orbit type.

•**Line Chart (Launch Success Trend by Year)**: Track yearly trends in landing success.

•**Confusion Matrix**: Evaluate model performance by showing true/false positives and negatives.

Github:
https://github.com/CmeDinfinity/Capstone_spaxeX/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

•Retrieve Unique Launch Sites:

sql
```
SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

•Records Where Launch Site Starts with 'CCA':

sql
```
SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%';
```

•Total Payload Mass by NASA (CRS) Launches:

sql
```
SELECT SUM(Payload_Mass__kg_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

•Average Payload Mass for Falcon 9 v1.1:

sql
```
SELECT AVG(Payload_Mass__kg_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';
```

•Date of First Successful Ground Pad Landing:

sql
```
SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

•Boosters with Drone Ship Success and Payload Mass 4000-6000 kg:

sql
```
SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND Payload_Mass__kg_ BETWEEN 4000 AND 6000;
```

•Total Successful and Failed Missions:

sql
```
SELECT Landing_Outcome, COUNT(*) FROM SPACEXTBL GROUP BY Landing_Outcome;
```

•Booster Versions with Maximum Payload Mass:

sql
```
SELECT Booster_Version FROM SPACEXTBL WHERE Payload_Mass__kg_ = (SELECT MAX(Payload_Mass__kg_) FROM SPACEXTBL);
```

•Failed Drone Ship Landings in 2015 by Month:

sql
```
SELECT SUBSTR(Date, 6, 2) AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date, 1, 4) = '2015';
```

•Rank Landing Outcomes (2010-06-04 to 2017-03-20):

sql
```
SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;
```

GITHUB:
https://github.com/CmeDinfinity/Capstone_spaxeX/blob/main/EDA_sql.ipynb

Build an Interactive Map with Folium

Map Objects Added to Folium Map:

- **Markers**: Added to each launch site to pinpoint locations.
- **Circles**: Highlighted the area around each launch site.
- **Marker Clusters**: Grouped multiple launches at the same site to reduce clutter.
- **Color-Coded Markers**: Green for successful landings, red for failed landings.
- **Polylines**: Showed distances from launch sites to nearby landmarks (e.g., coastlines, cities).

Why:

- To clearly visualize launch sites, landing outcomes, and their proximity to important geographic features.

GITHUB:
https://github.com/CmeDinfinity/Capstone_spaxeX/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Plots/Graphs and Interactions Added to the Dashboard:

- **Line Plot**: Shows launch success over time.
- **Bar Chart**: Displays success rate by orbit type.
- **Scatter Plot**: Visualizes flight number vs. payload mass vs. success.
- **Map**: Interactive map of launch sites and outcomes.
- **Dropdown Filter**: Allows users to filter data by launch site.
- **Tooltips/Hover**: Provides quick data insights on hover.

Why:

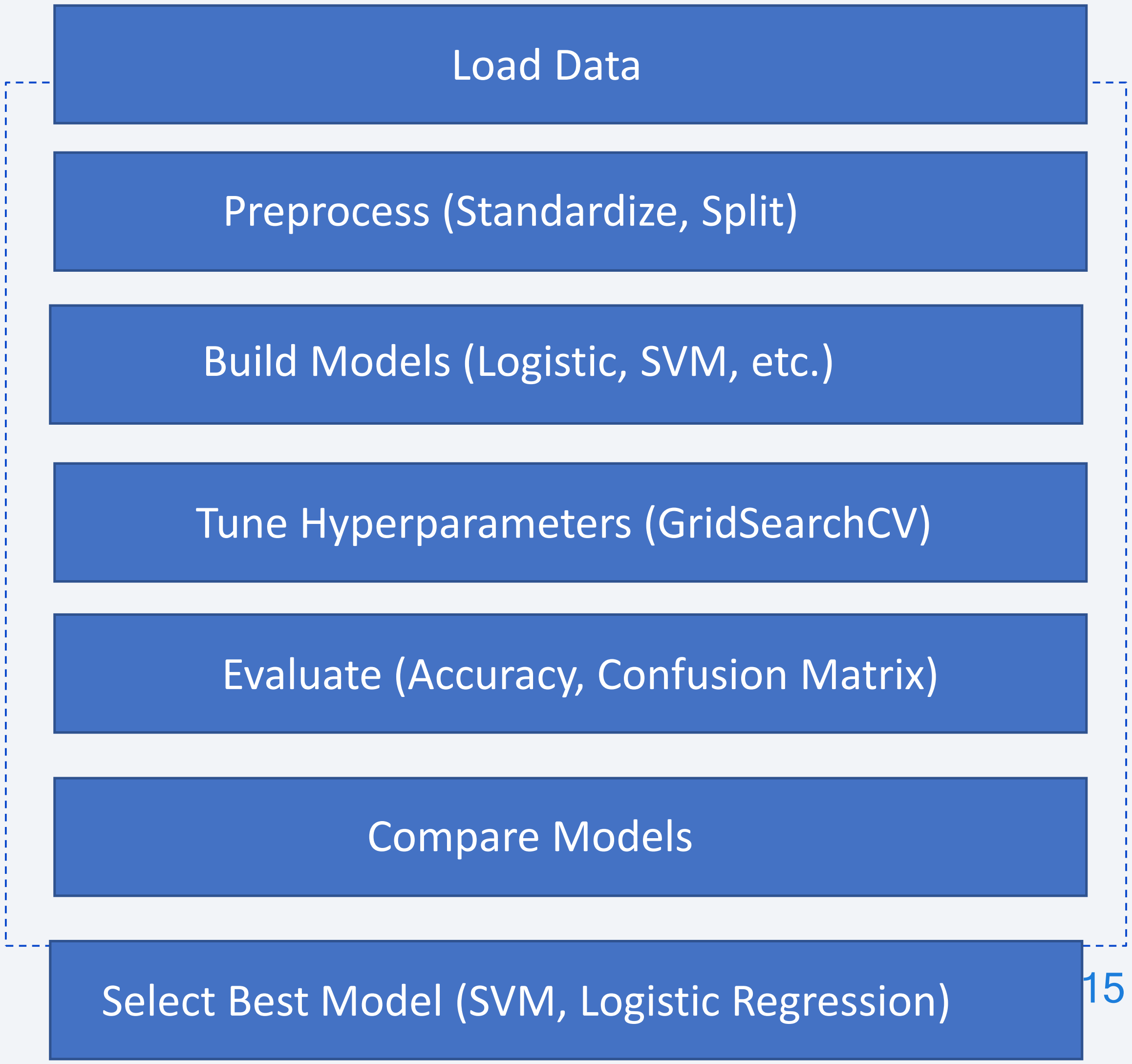
- To visualize trends, compare success rates, and enable interactive data exploration for deeper insights.

GITHUB:
https://github.com/CmeDinfinity/Capstone_spaxeX/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- **Built Models**: Created Logistic Regression, SVM, Decision Tree, and KNN models.
- **Evaluated Models**: Used accuracy and confusion matrices for performance evaluation.
- **Improved Models**: Tuned hyperparameters using **GridSearchCV** and standardized data.
- **Best Model**: **SVM (linear)** and **Logistic Regression** performed best with **83.33% accuracy**.

GITHUB:
https://github.com/CmeDinfinity/Capstone_spaxeX/blob/main/Machine%20Learning%20Prediction.ipynb



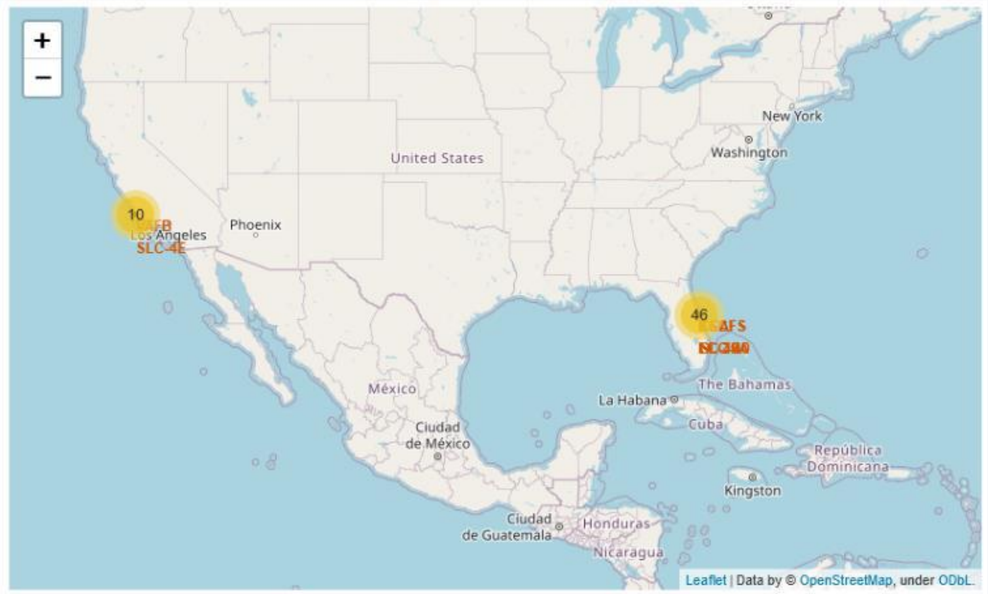
Results

Exploratory data analysis results:

•Space X uses 4 different launch sites;
•The first launches were done to Space X itself and NASA;
•The average payload of F9 v1.1 booster is 2,928 kg;
•The first success landing outcome happened in 2015 fiver year after the first launch;
•Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
•Almost 100% of mission outcomes were successful;
•Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
•The number of landing outcomes became as better as years passed.

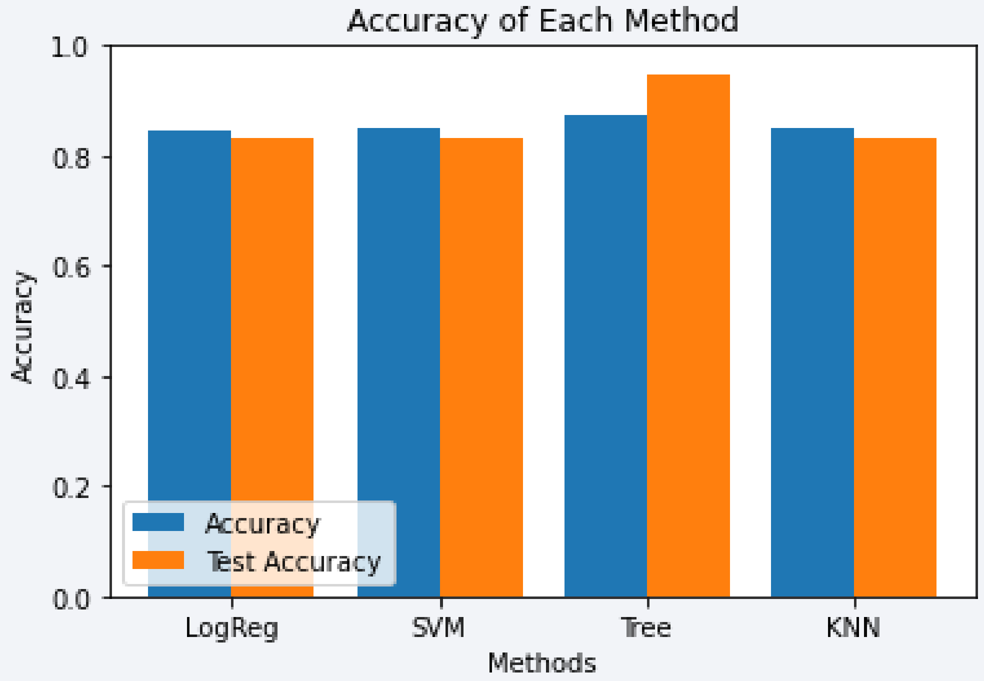
Results

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.



Results

•Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.



Section 2

Insights drawn from EDA

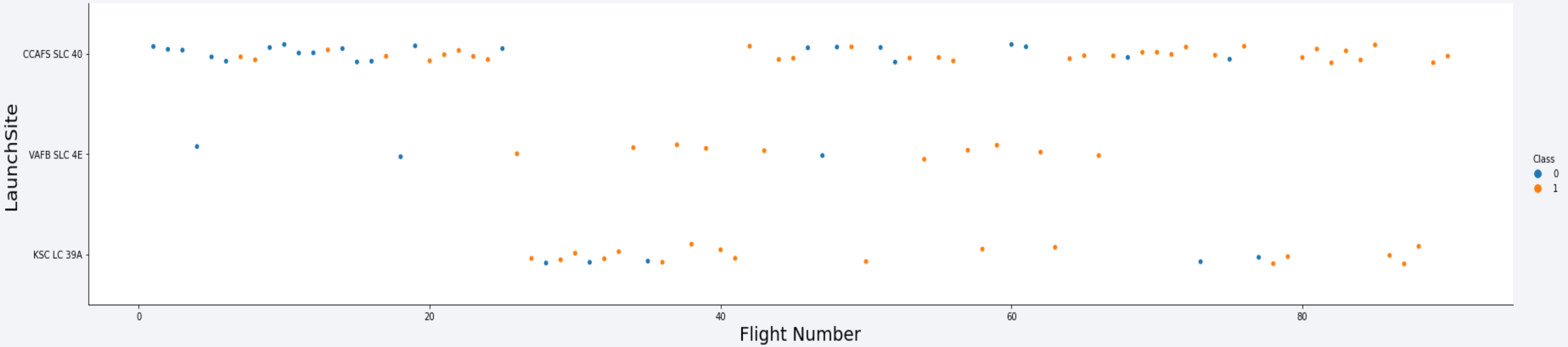
Flight Number vs. Launch Site

•**CCAFS SLC 40** is the best-performing launch site, with most recent launches being successful.

•**VAFB SLC 4E** ranks second in launch success.

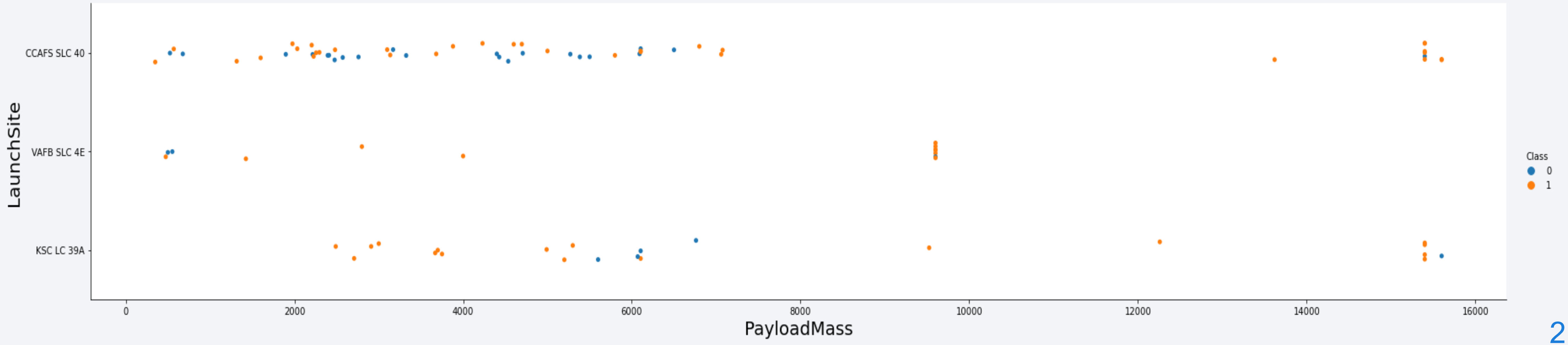
•**KSC LC 39A** holds the third position in launch success.

The overall success rate has steadily improved over time.

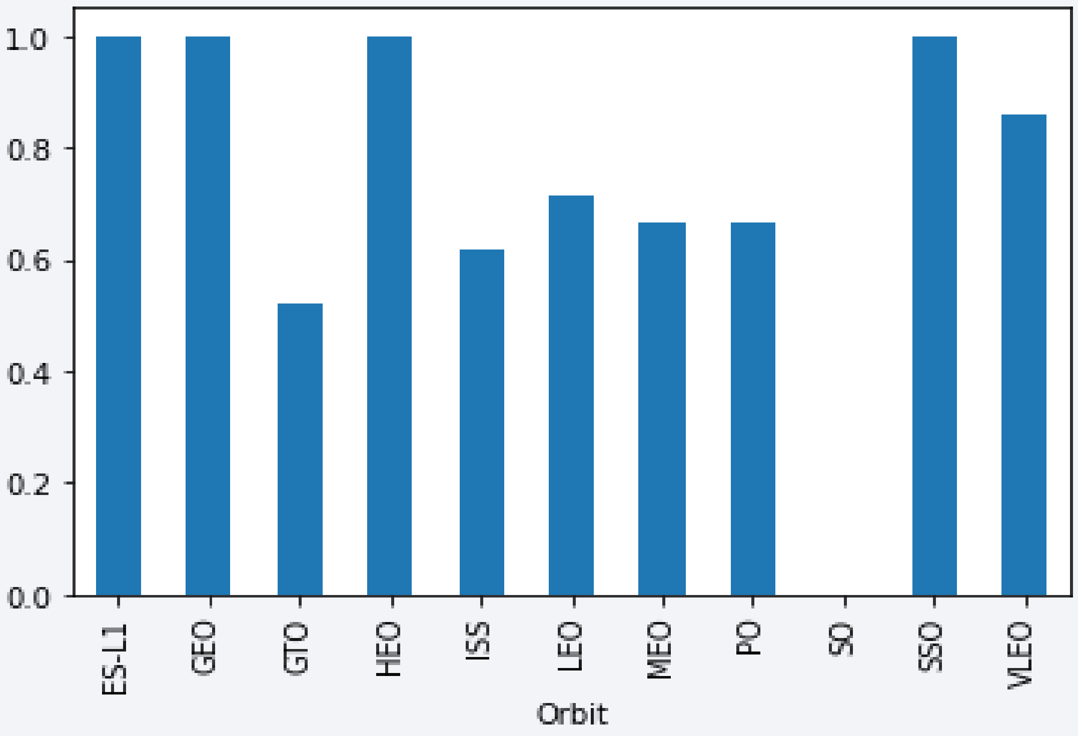


Payload vs. Launch Site

•Payloads exceeding **9,000kg** (approximately the weight of a school bus) have a high success rate.
•Payloads over **12,000kg** appear to be launched exclusively from the **CCAFS SLC 40** and **KSC LC 39A** sites.



Success Rate vs. Orbit Type



The highest success rates are observed in the following orbits:

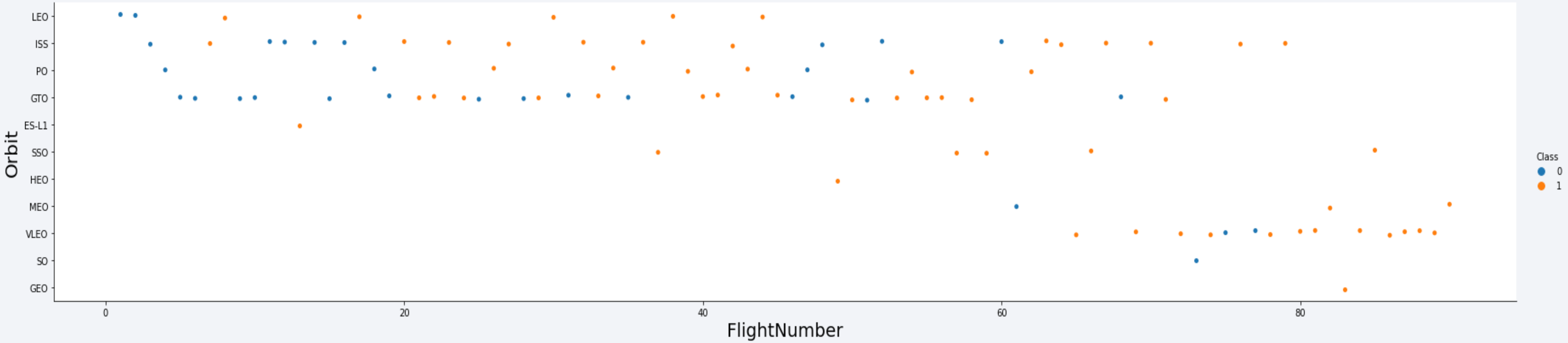
- **ES-L1**
- **GEO**
- **HEO**
- **SSO**

This is followed by:

- **VLEO** with a success rate of over 80%
- **LFO** with a success rate of over 70%

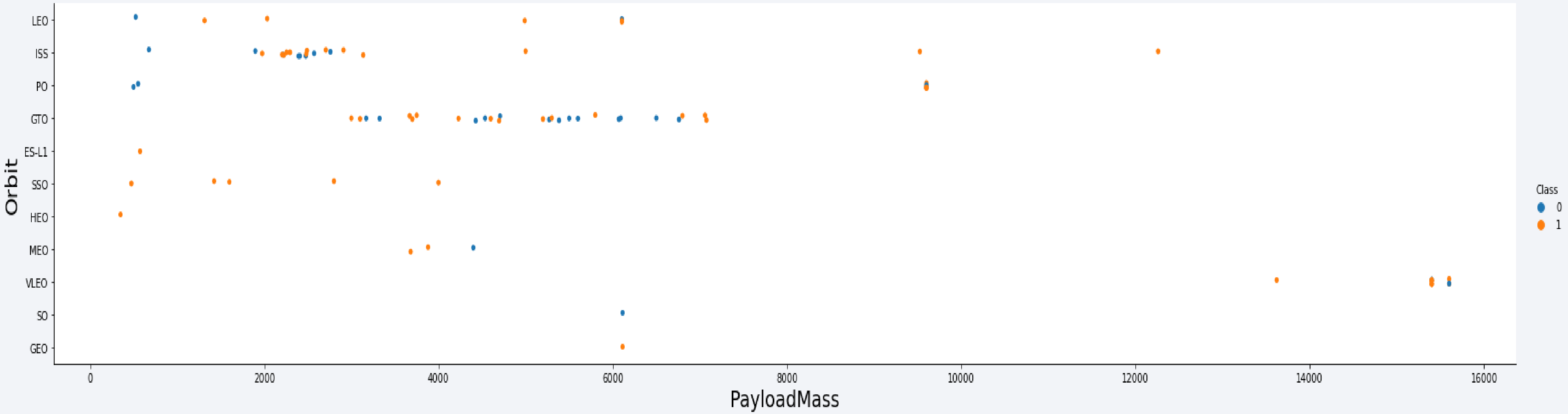
Flight Number vs. Orbit Type

•The success rate has improved over time across all orbits.
•The **VLEO** orbit presents a new business opportunity, as its frequency has recently increased.



Payload vs. Orbit Type

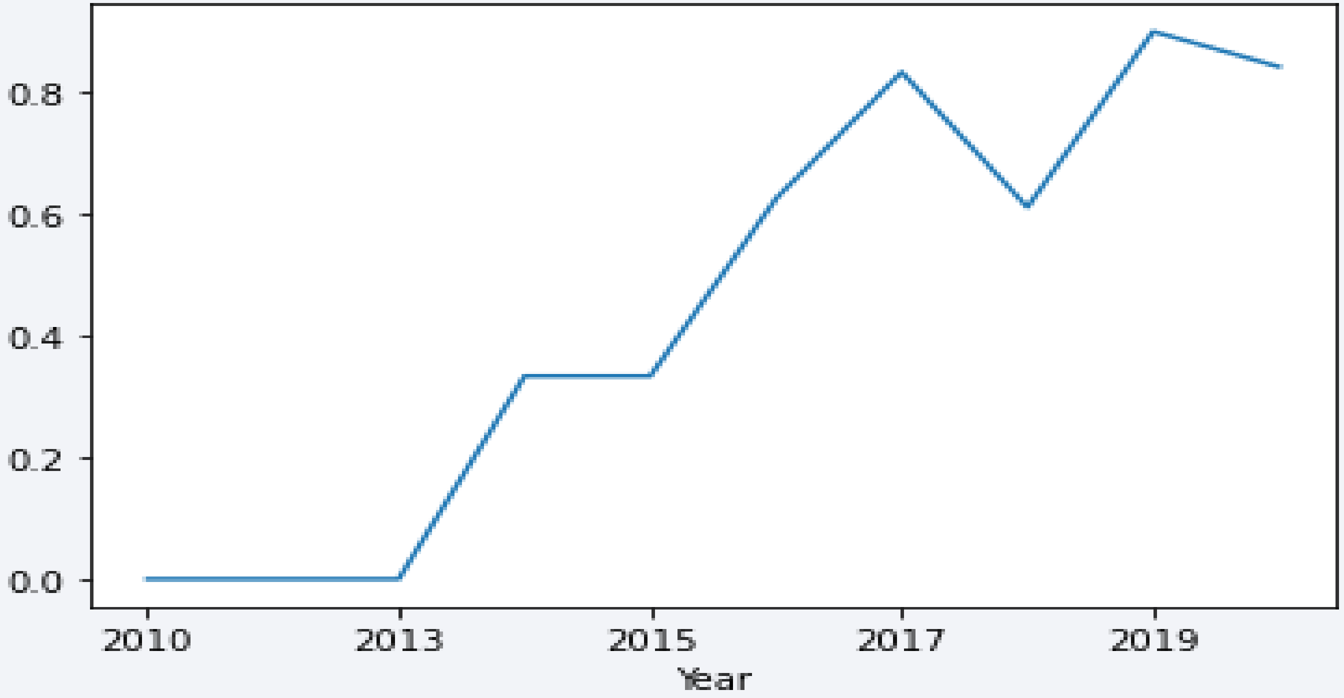
- There appears to be no correlation between payload size and success rate for the **GTO** orbit.
- The **ISS** orbit handles the widest range of payloads and maintains a good success rate.
- There are relatively few launches to the **SO** and **GEO** orbits.



Launch Success Yearly Trend

•The success rate began increasing in **2013** and continued to improve until **2020**.

•The first three years appear to have been a period of adjustments and technological improvements.



All Launch Site Names

Launch Site Names:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

Total Payload in KG : 111.268

SELECT SUM(Payload_Mass__kg_) AS Total_Payload_Mass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';

Explanation:

•**Purpose**: This query calculates the total payload mass carried by SpaceX boosters for NASA (Customer = 'NASA (CRS)').
•**SUM()**: Aggregates the payload mass (Payload_Mass__kg_) for all launches where NASA was the customer.
•**WHERE** clause filters the records to only include those for NASA's Cargo Resupply Services (CRS) missions.

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928 KG

First Successful Ground Landing Date

- First successful landing outcome on ground pad is 2015 -12 - 22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

Landing Outcome	Occurrences
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

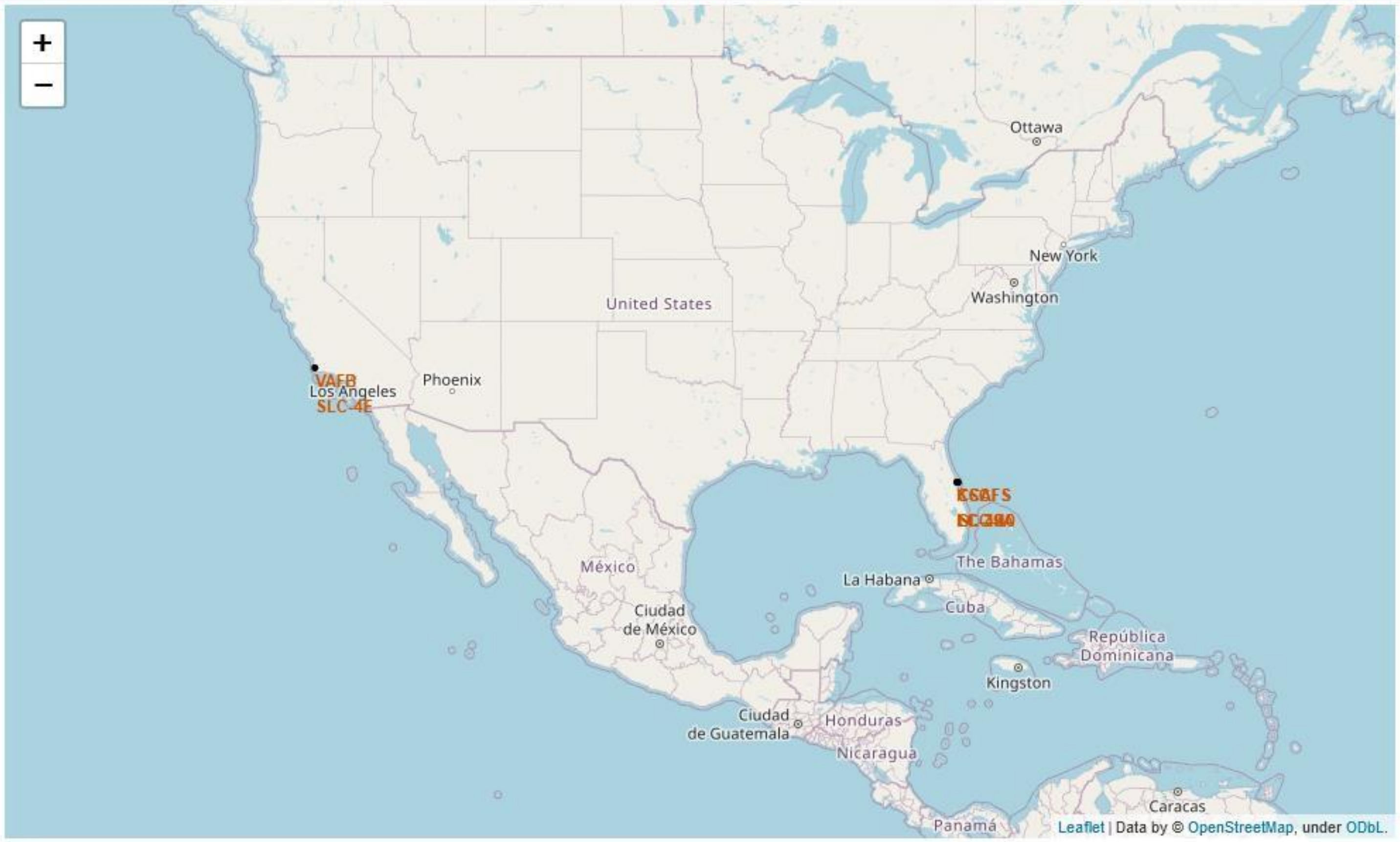
- This view of data alerts us that “No attempt” must be taken in account.

Section 3

Launch Sites Proximities Analysis

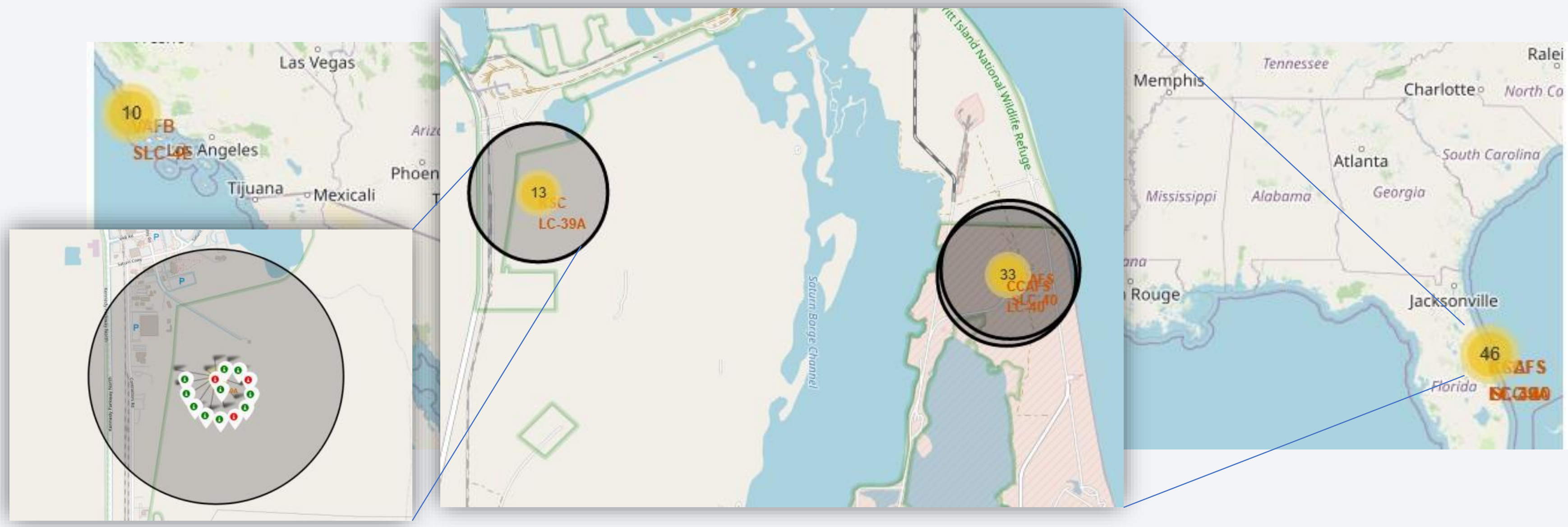
Launch Sites

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

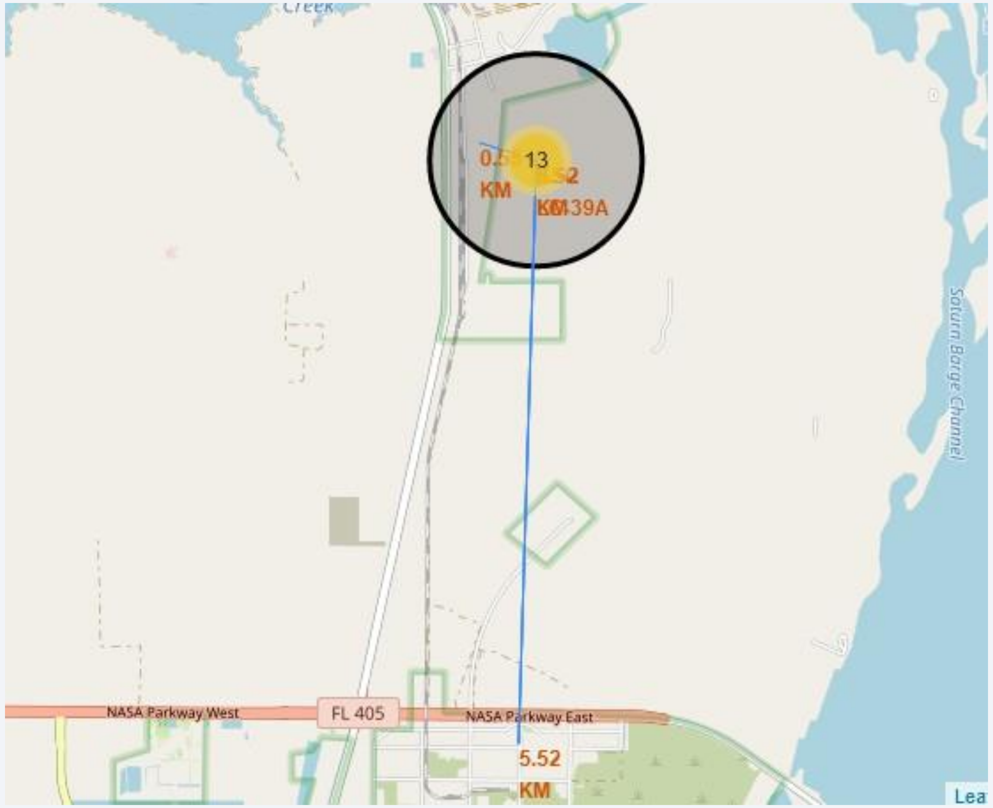


Launch Outcomes

- Green markers indicate successful and red ones indicate failure.



Logistics



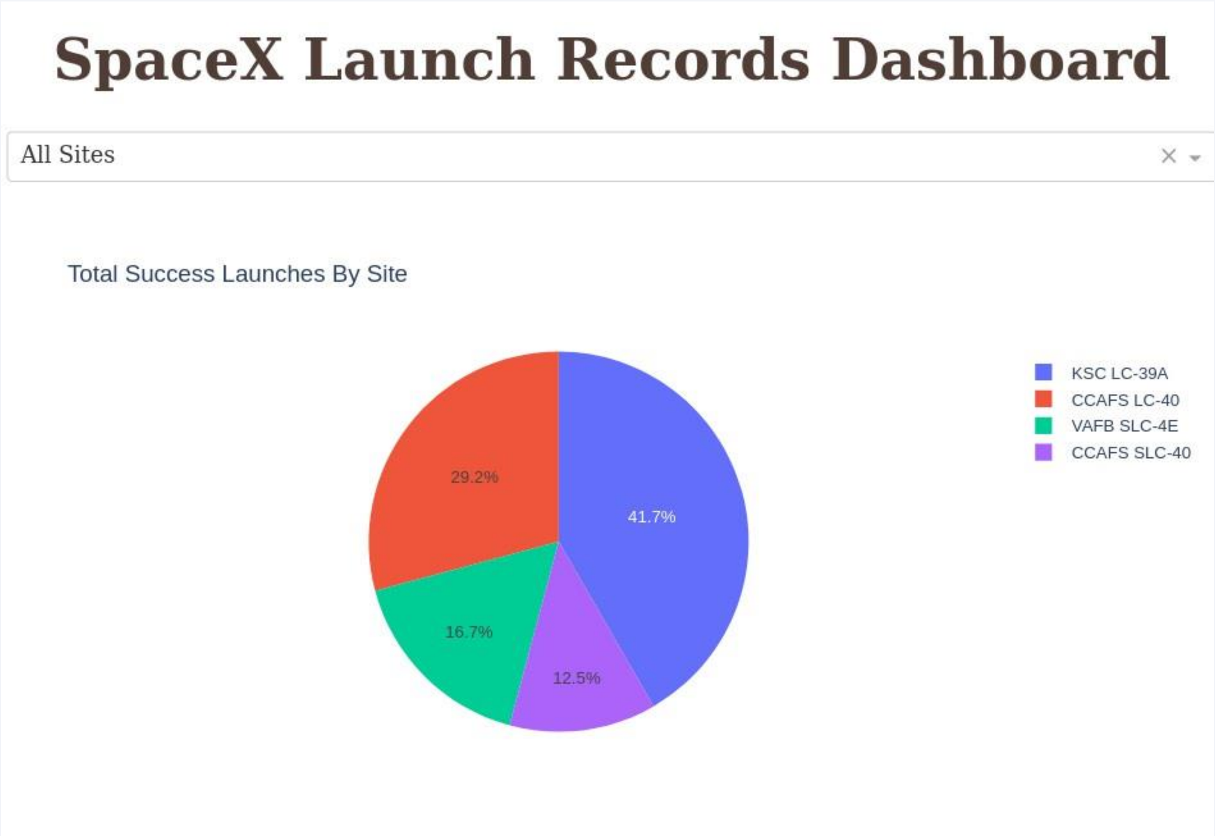
- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 4

Build a Dashboard with Plotly Dash

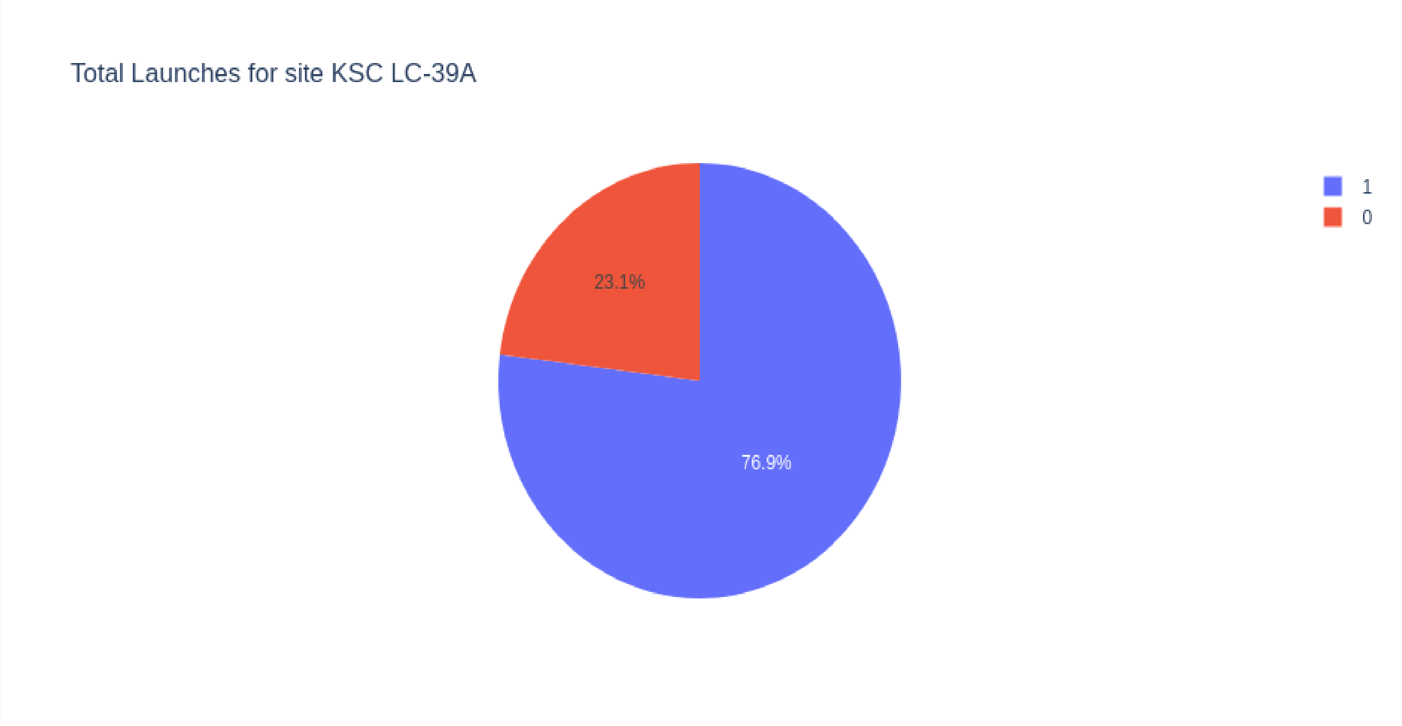
Successful Launches

- The place from where launches are done seems to be a very important factor of success of missions.



Launch Success for site KSC LC - 39A

- 76.9% of launches are successful in this site.



Payload vs Launch Outcome

- Payloads under 6,000kg and FT boosters are the most successful combination.

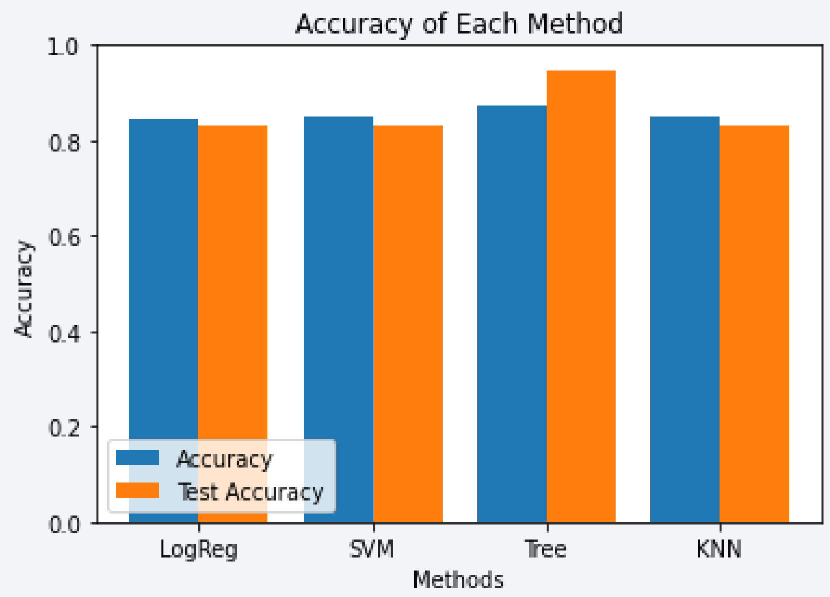
Section 5

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

Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

•Various data sources were analyzed, leading to refined conclusions throughout the process.
•**KSC LC-39A** is identified as the top launch site.
•Launches with payloads over **7,000kg** are associated with lower risk.
•While most mission outcomes are successful, the success rate of landings appears to improve over time due to advancements in processes and rockets.
•A Decision Tree Classifier can be utilized to predict successful landings and enhance profitability.

Appendix

All codes and notebooks are found in my github repo:

https://github.com/CmeDinfinity/Capstone_spaxeX

Thank you!