IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Advay Singh
August 27, 2024

Outline

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

Executive Summary

- Methodologies used for Data Analysis:
  1. Data Collection was performed using:
     - Web Scraping from Html page sources
     - SpaceX Web API (https://api.spacexdata.com/v4)
  2. Data wrangling, outcome label creations, data clean up
  3. Performed Exploratory Data Analysis (EDA) using SQL and Visualizations
  5. Predictive analysis using Machine Learning pipeline
- Summary of all results
  - Data was successfully pre-processed using the public sources
  - Best Machine learning model characteristics were identified and used
  - Machine learning pipeline was able to learn and predict the landing

Introduction

- Project background and context
  - The commercial space industry is evolving rapidly with several key players.
  - **SpaceX's Falcon 9** is notable for its cost efficiency due to its reusable first stage, which reduces launch costs to $62 million, compared to over $165 million from competitors.
  - Use machine learning models and data analysis to provide insights for Space Y, a new company looking to compete with SpaceX in the space industry.
- Problems you want to find answers:
  - What factors influence whether the first stage of a Falcon 9 rocket can be reused?
  - Best locations to make the launches
  - Analyze whether SpaceX's Falcon 9 first stage will land successfully and be reused based on available data.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web Scraping from source: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches
  - Space X API was used to obtain json data: https://api.spacexdata.com/v4
- Perform data wrangling
  - Add landing outcome label
  - Data summarization for various features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - Data normalization
  - Train test split for input data
  - Utilize different machine learning models:
    - Logistic Regression
    - Support Vector Machines
    - Decision Tree Classifier
    - K-nearest neighbors.
  - Perform hyperparameter tuning using Grid Search.
  - Model Evaluation: Determine the model with the best accuracy using the training data.

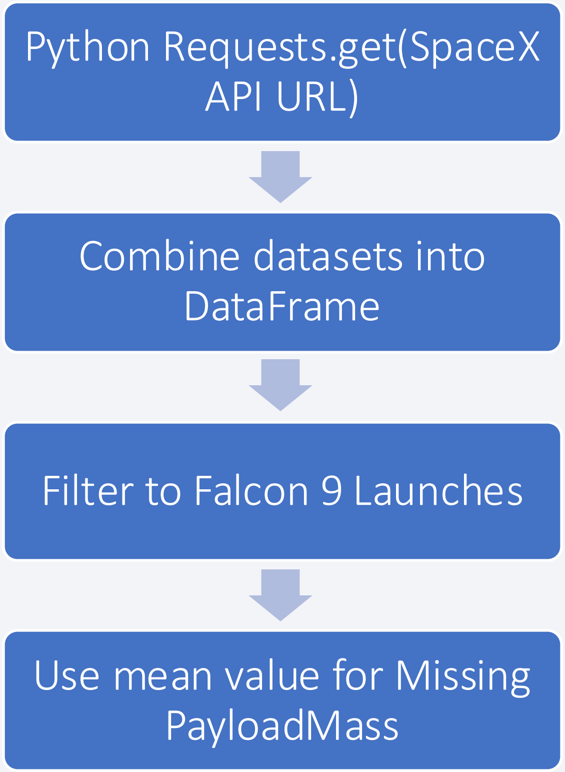
Data Collection

- Describe how data sets were collected.
  - Web Scraping from source: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches
  - Space X API was used to obtain json data: https://api.spacexdata.com/v4
- Data collection process
  - Web scraping using the Python BeautifulSoup package to extract Falcon 9 launch data from HTML tables.
  - Convert data from Html Tables and Web API json into Data Frames
  - Filter data to focus only on Falcon 9 launches

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

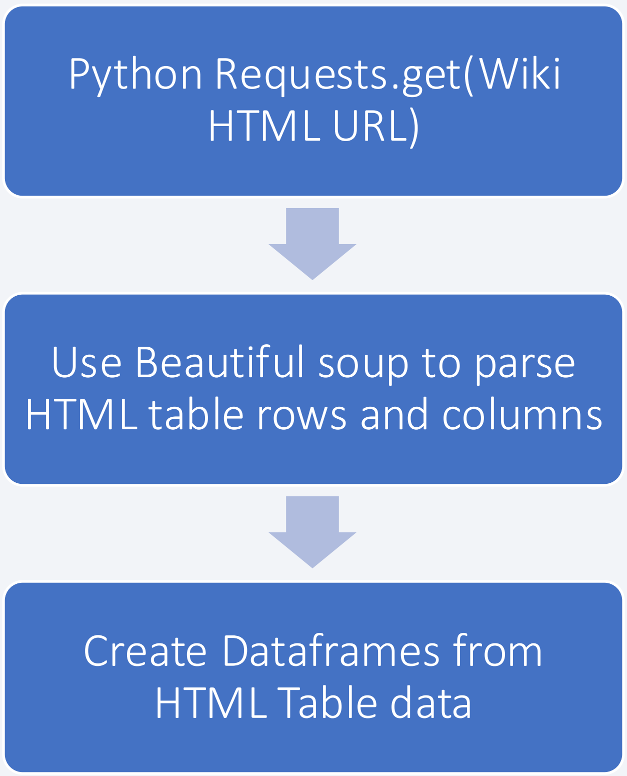
- https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scraping using the Python BeautifulSoup package to extract Falcon 9 launch data from HTML tables.

- https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Cleaning and transforming raw data into a format that is suitable for analysis
- Convert the landing outcomes into classes, where 0 represents a bad outcome (the booster did not land) and 1 represents a good outcome (the booster did land)
- https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Use value_count() to analyze data summary

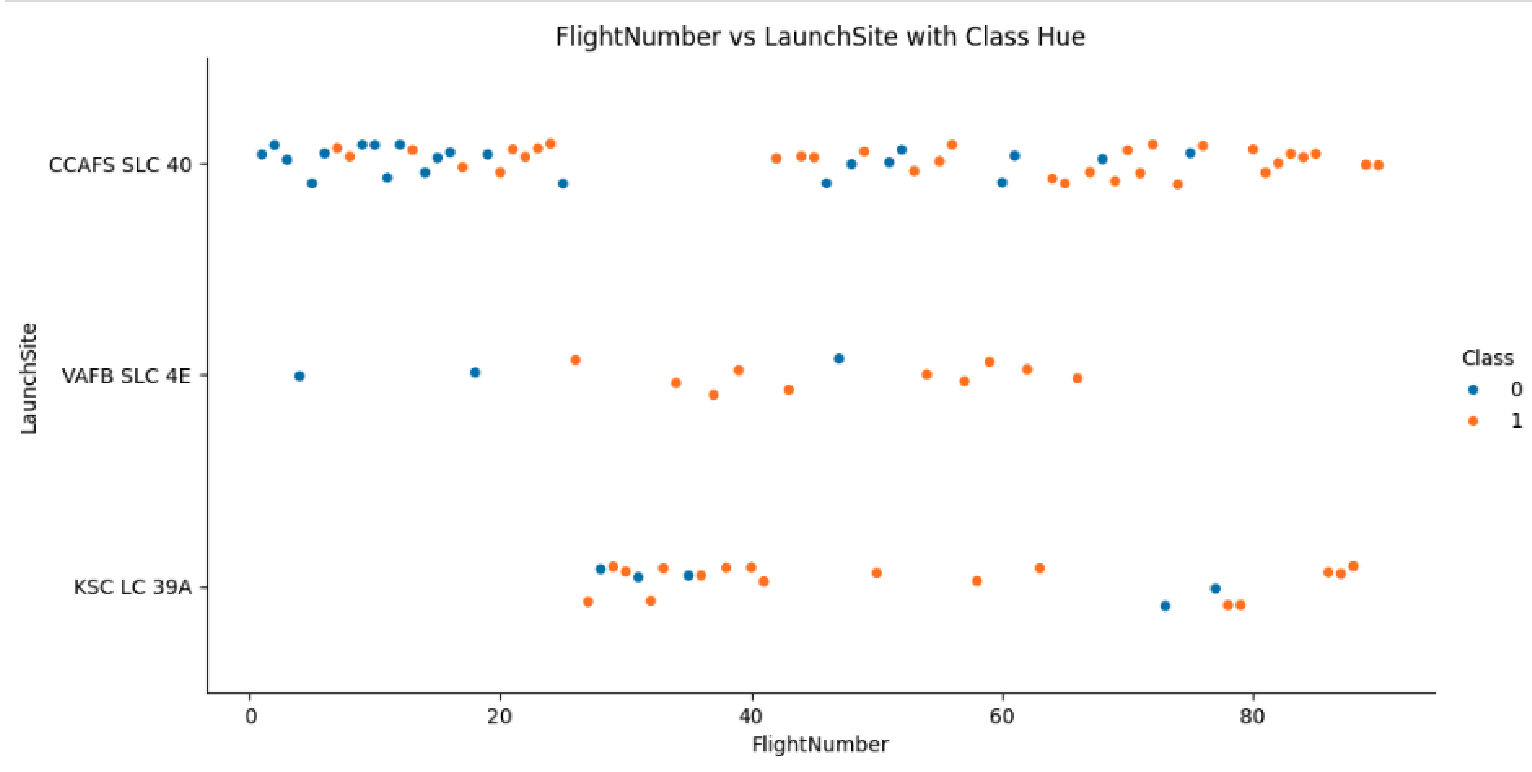
Determine the number of launches at each site

Create a landing outcome label from Outcome column

Determine success rate from Outcome

EDA with Data Visualization

- Scatterplot to analyze relationship between Flight Number and Launch site and flight outcome (class) hue
- Scatterplot to analyze relationship between Flight Number and Payload size and flight outcome hue

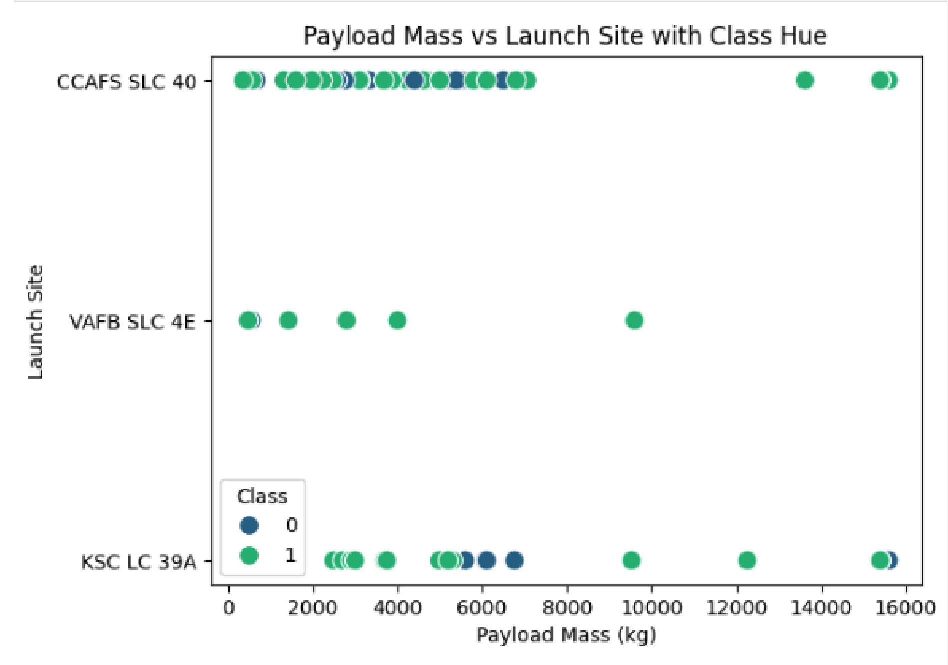


EDA with Data Visualization

- Scatterplot to analyze relationship between Launching site and Payload size and flight outcome hue

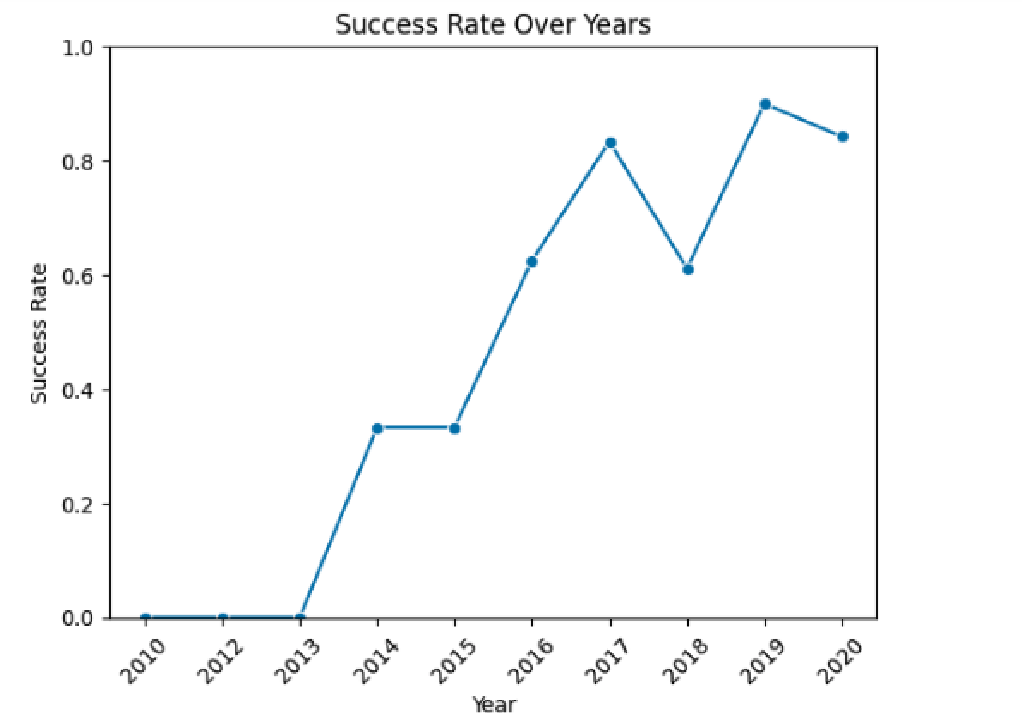
- Barchart to visualize the relationship between

  success rate of each orbit type



EDA with Data Visualization

- Line plot to visualize success rate over years



https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/eda_with_datavisualization.ipynb

EDA with SQL

Summary of the SQL queries performed

- List DISTINCT Launch Sites for the space missions
- Display 5 records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Distinct Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA with SQL

Summary of the SQL queries performed

- Total number of successful and failure mission outcomes
- Display 5 records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Distinct names of the Booster Versions that have carried the maximum payload mass by using a subquery
- Rank the count of landing outcomes such as Failure (drone ship) or Success (ground pad) in descending order.

https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, lines, marker clusters, and circles were created and added with Folium Maps

Why those objects were added:

- Markers used to represent specific points, such as launch sites
- Circles highlight areas around particular coordinates, like the Space Center
- Marker clusters group events at the same location, such as multiple launches at a launch site
- Line depicts the distances between two coordinates.
- https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/lab_jupyter_launch_site_location_folio.ipynb

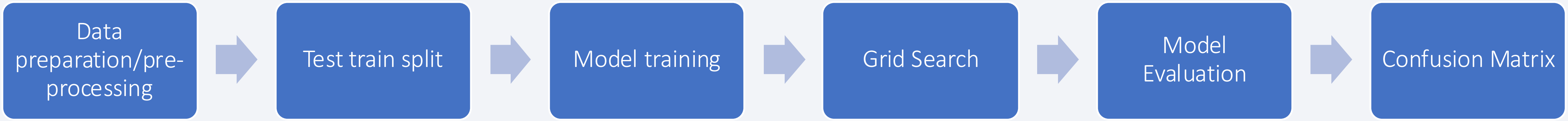
Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Pie chart to visualize the success count percentages by all sites
- Add interaction to filter and show success vs fail percentage count by a selected site
- Range slider interaction with scatter plot to select payload range
- That interaction with scatter plot allowed to study the relationship between payload size range vs. launch site vs. Booster category and success rates

- https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Used four different machine learning classification models:
  - Logistic Regression
  - Support Vector Machines
  - Decision Tree Classifier
  - K-nearest neighbors.



- https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results
  - SpaceX operates from four different launch sites.
  - The average payload for Falcon F9 is 2,928.4 kg.
  - The first successful landing occurred in December 2015
  - Several Falcon 9 booster versions successfully landed on drone ships with payloads above the average.
  - Mission success rate is almost 100%

Results

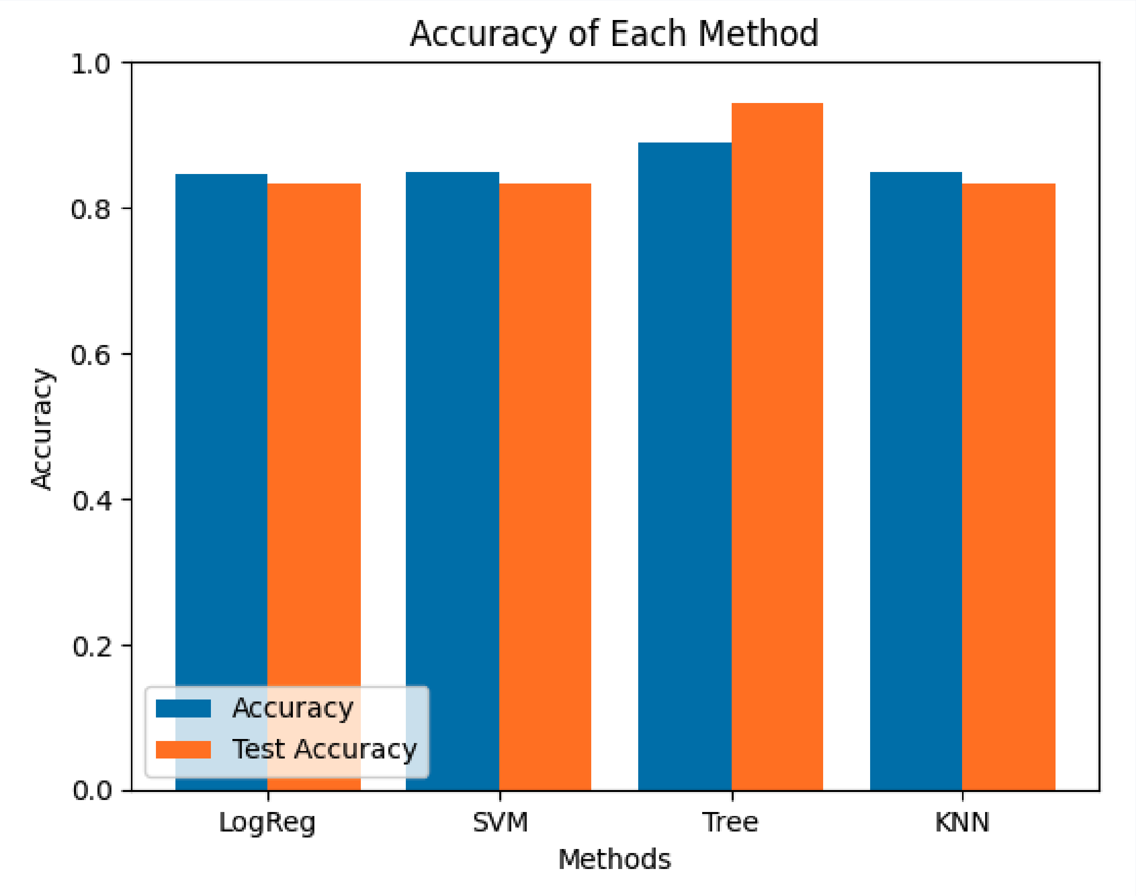
- Interactive analytics demo in screenshots
  - Most of the launch sites and launches are done from Florida sites

  - Launch sites are near to coastline

- Predictive analysis results

Results

- Machine learning predictive analysis model building and comparison showed that Decision Tree classifier model is the most accurate and the best model to predict successful landings for Space Y team.
- This model has 94% test accuracy and 89% train accuracy.

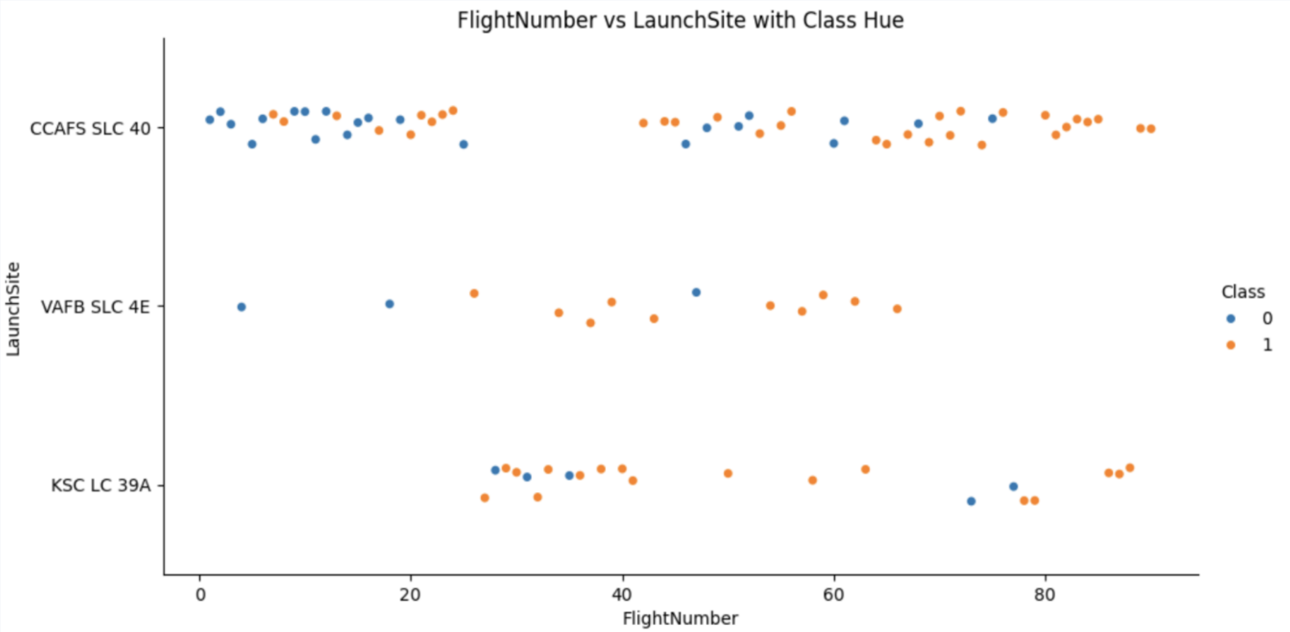


Section 2

Insights drawn from EDA

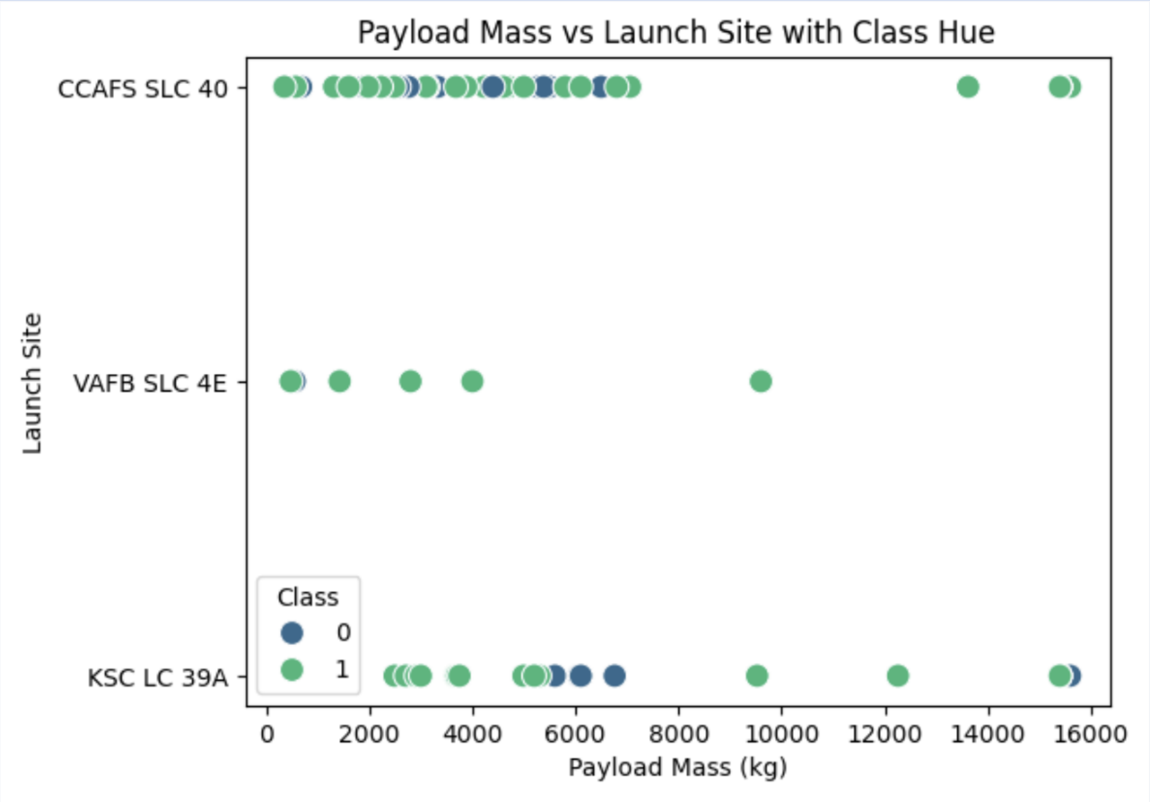
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



- CCAFS SLC 40 is top Launch site and has had more success
- Launch success has improved with time or recent flights

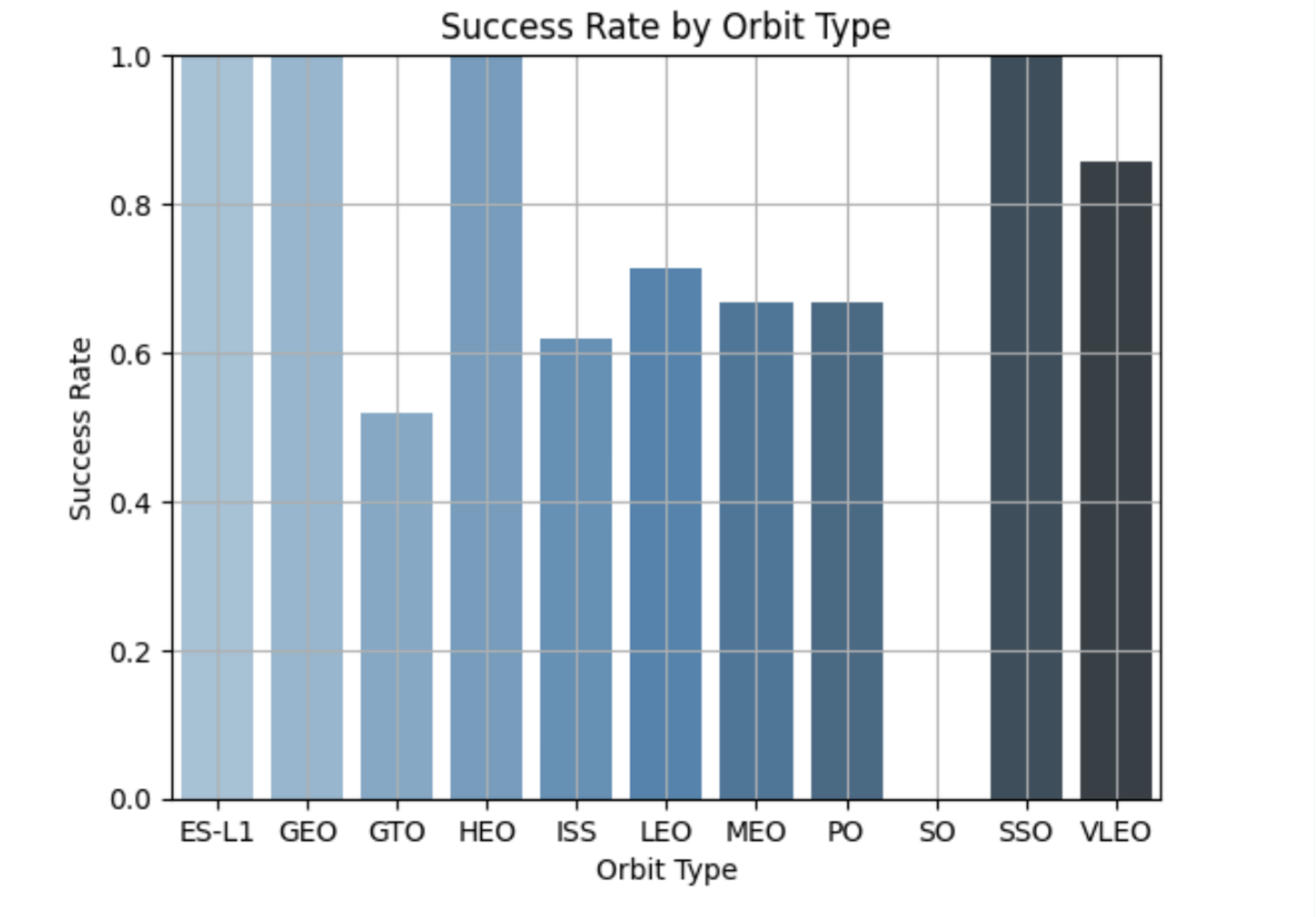
Payload vs. Launch Site



- VAFB-SLC did not launch any payload more than 10000 kg
- More payloads across all weights launched from site CCAFS SLC 40 and better success with heavier payloads of 16000 kg

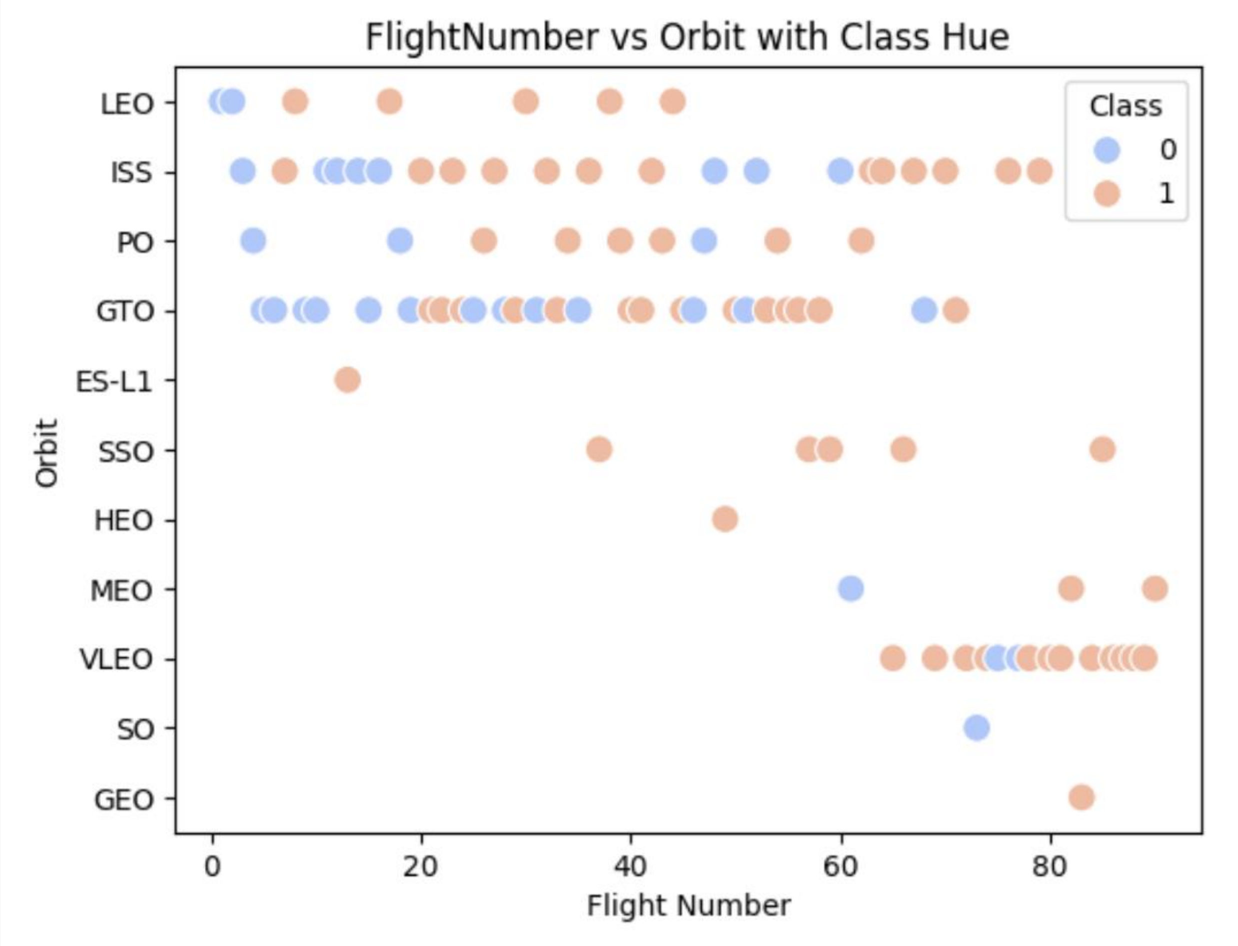
Success Rate vs. Orbit Type

- Most successful Orbits are
  - ES-L1
  - GEO
  - HEO
  - SSO
- LEO orbit success rate is close to 75%
- SO orbit has no success rate



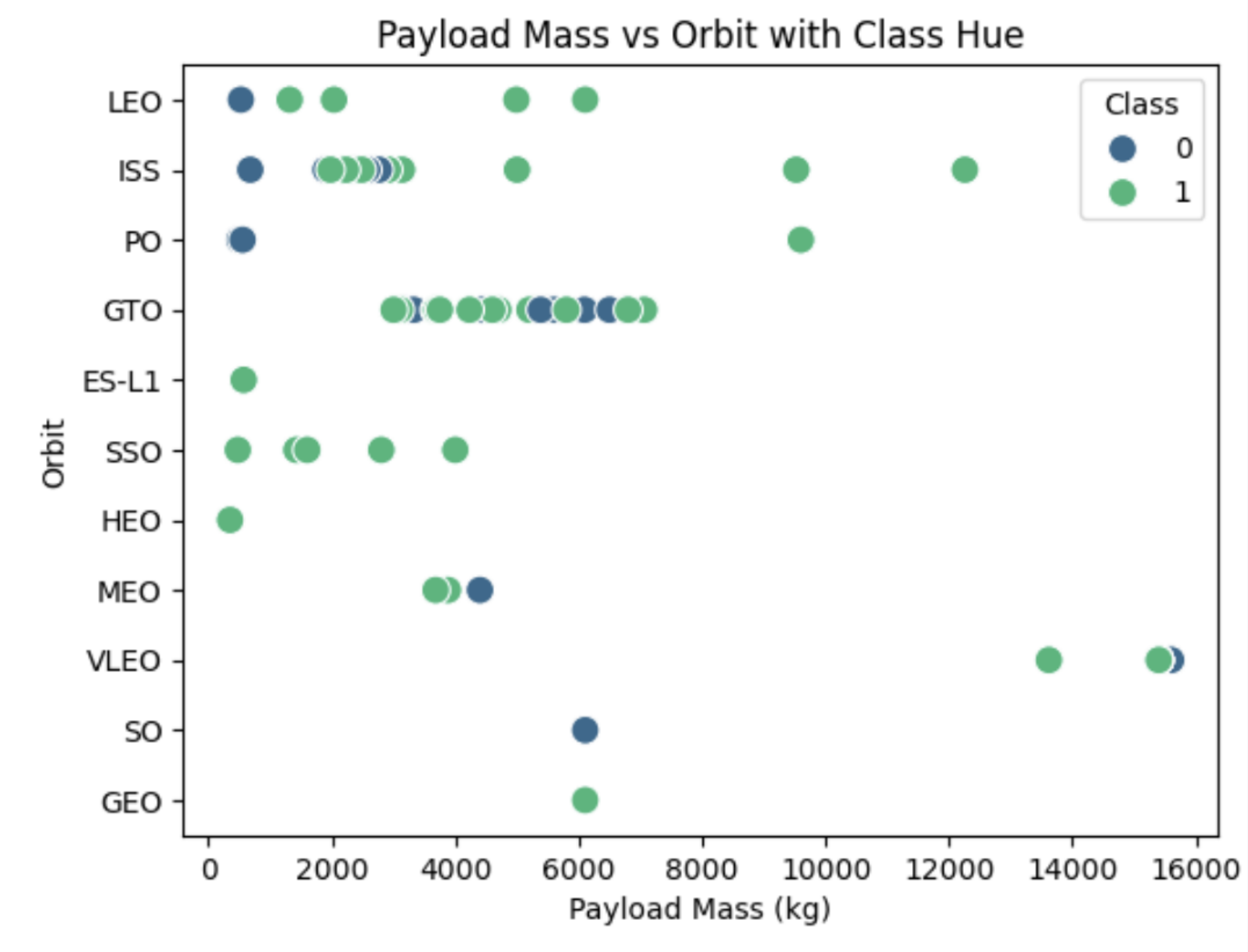
Flight Number vs. Orbit Type

- In GTO Orbit there is no relation between flight number and success
- VLEO orbit has had most recent flights and high success rate
- More success in recent flights in VLEO orbit
- One flight and that also failed in SO orbit
- Successful flight proportion increased in general with recent flights over time



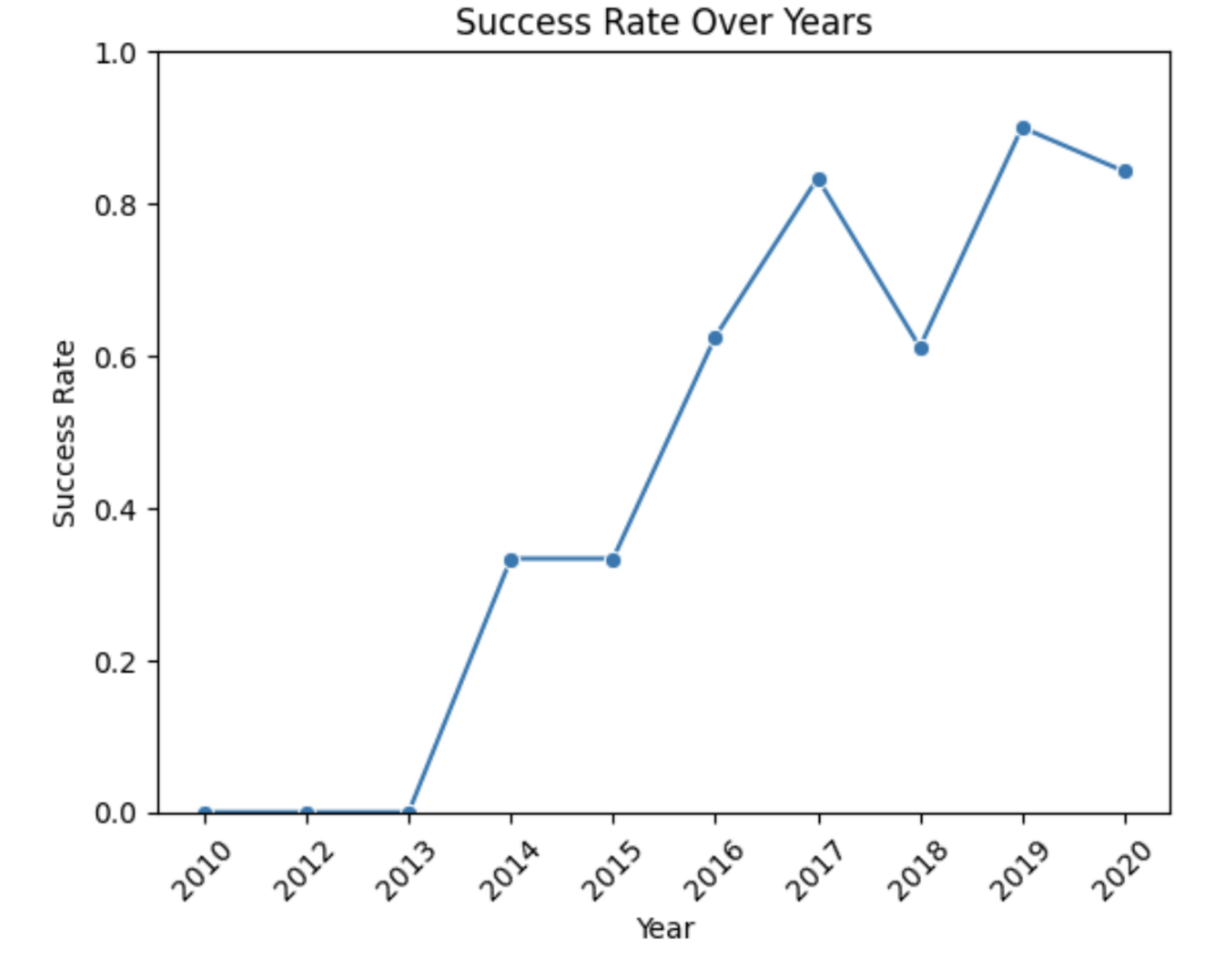
Payload vs. Orbit Type

- With heavy payloads the successful landing rate are more for Polar (PO) , VLEO and ISS.
- No pattern of relation between Payload mass and success in Orbit GTO
- Least number of payload launches from SO and GEO
- ISS has the most variety of payloads



Launch Success Yearly Trend

- No success before year 2013
- Success rate kept increasing from 2013 to 2020
- Highest success rate in year 2019



All Launch Site Names

- Names of the unique launch sites

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

- Query result with a short explanation here:
  - SELECT DISTINCT Launch_Site from SPACEXTABLE
  - Selects the unique data values from Launch site column

Launch Site Names Begin with 'CCA'

- Find 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	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The LIMIT 5 records query shows 5 different data points for "CCAFS LC-40"

Total Payload Mass

- Total payload carried by boosters from NASA is 45596 KG

- Query result with a short explanation here:
  - SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer = 'NASA (CRS)'
  - The SUM query calculates the total payload mass in the entire SpaceX flight data

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 KG

- Query: "SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'"

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad is 2015-12-22
- Query: SELECT min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
- It selects the least value of the date from the records for Landing_Outcome = 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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 B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Four booster versions have been able to land successfully with payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

Mission_Outcome	QTY
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- GROUP BY MISSION_OUTCOME gave the grouped counts by outcome

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass:

- Sub-query to find max payload results was used as an inner query to list all the unique booster versions in those records

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Failures were in the month of January and April.

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

- Ranked 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

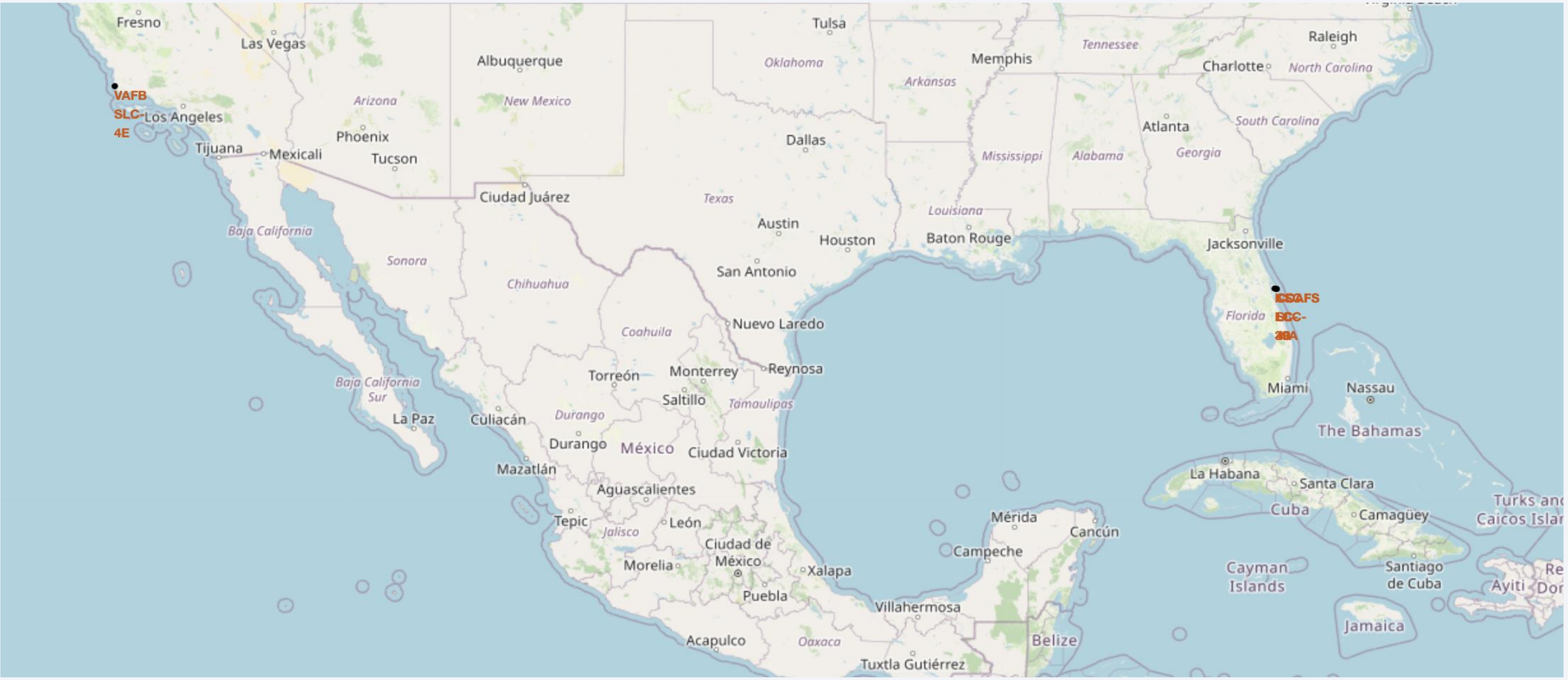
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- "No attempt" has the highest count and Precluded (drone ship) has the least outcomes

Section 3

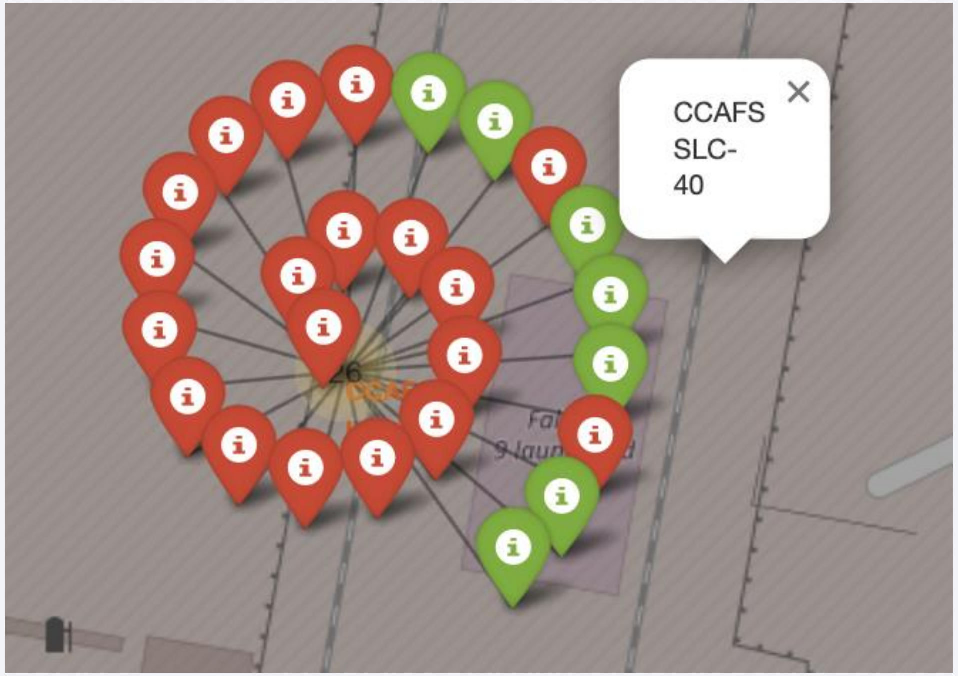
Launch Sites Proximities Analysis

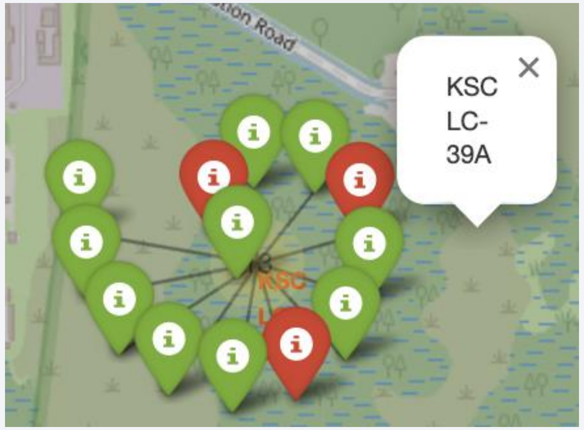
Launch site locations on global map



- Launch sites are located near coasts
- Located in California and Florida

Launch outcome markers by site

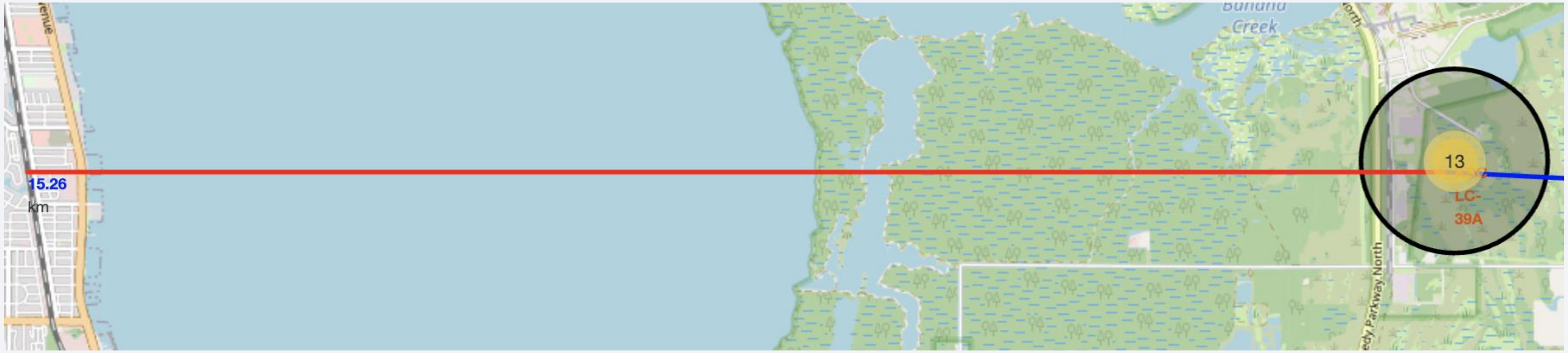




- Success and failure launch outcomes are marked by Green and Red colors respectively
- Success rates are visible by site locations/names

Proximities of Launch sites

- Distance to coastline from KSC LC- 39A: 7.73 KM

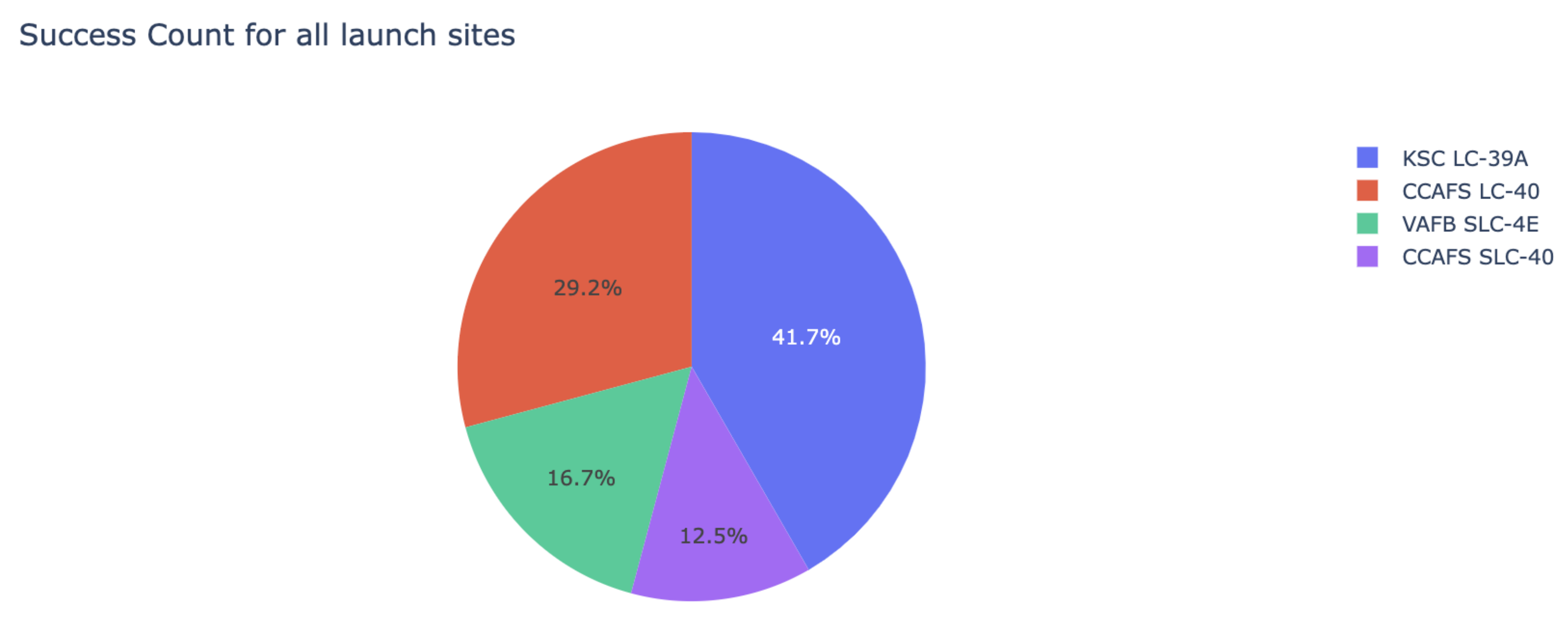


- Launch site is located near coastline and not very near to the cities and railway

Section 4

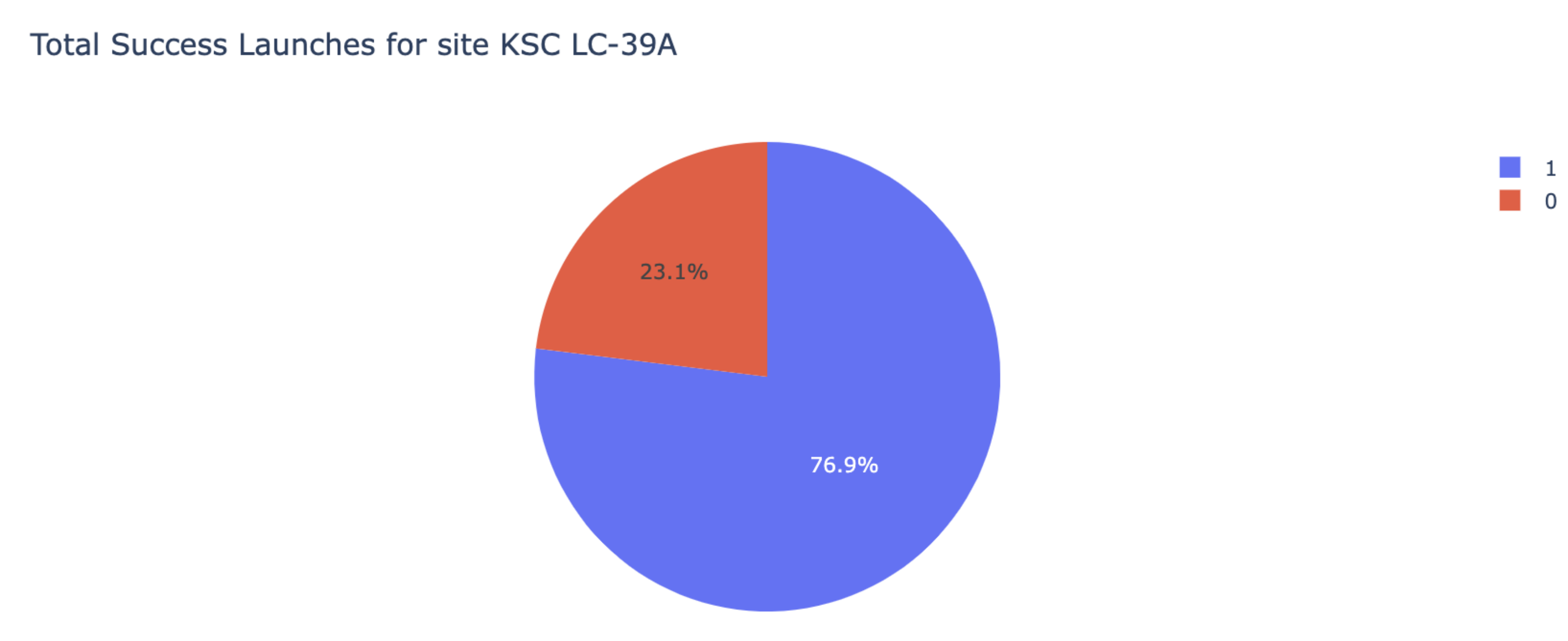
Build a Dashboard with Plotly Dash

Launch success count for all sites



- Success rate differs by sites
- KSC LC-39A has the highest success rate

Piechart for the launch site with highest launch success ratio



- Highest success ratio of a site is 76.9 %

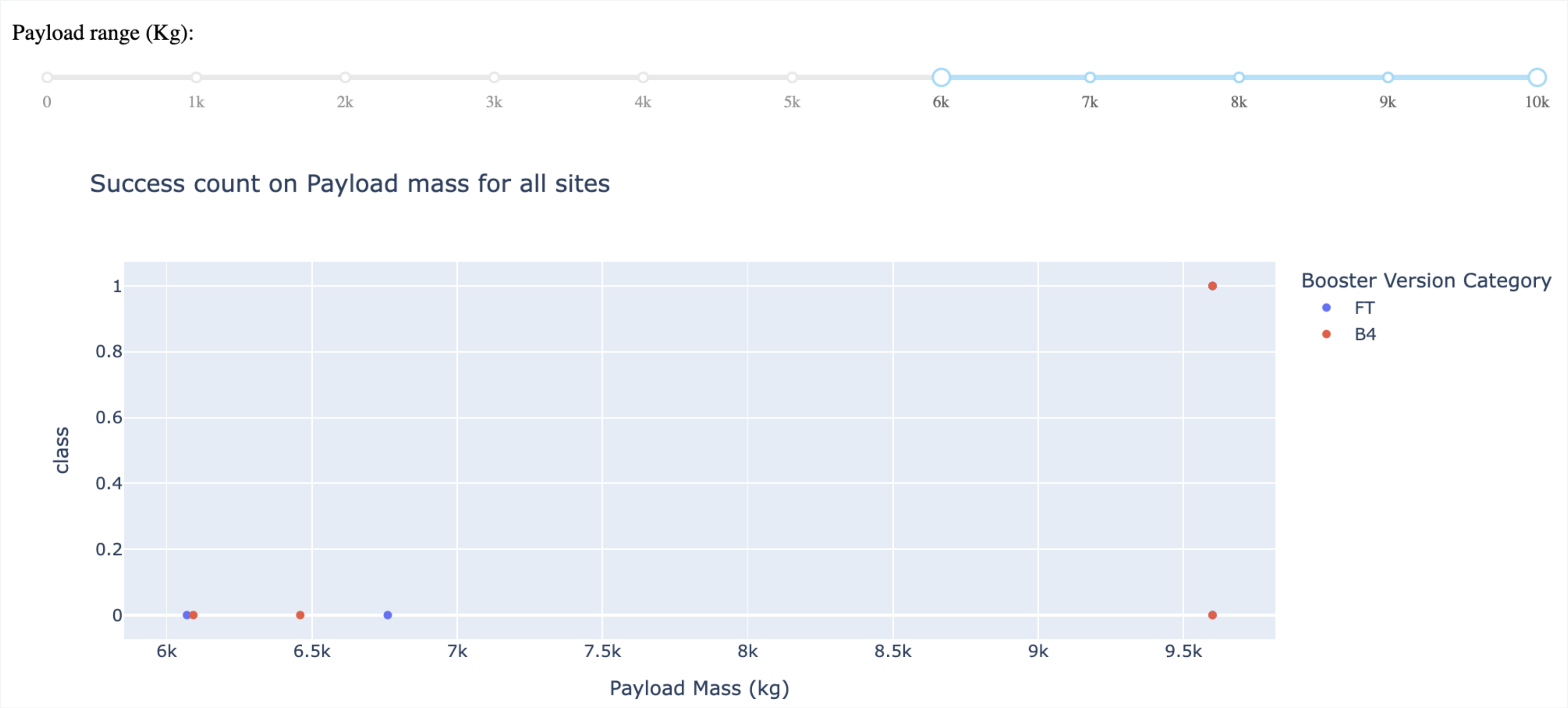
Payload vs. Launch Outcome Scatter plot

- Payload vs. Launch Outcome scatter plot for all sites, with different payloads selected in the range 2000 kg to 10,000 kg.

- Booster version "FT" in payload range 2000 kg to 5500 kg has the highest success rate.

Payload vs. Launch Outcome Scatter plot

- Payload vs. Launch Outcome scatter plot for all sites, with different payloads selected in the range 6000 kg to 10,000 kg.



- Only the Booster version "B4" in payload range 6000 kg to 10000 kg has the successful launch outcome.

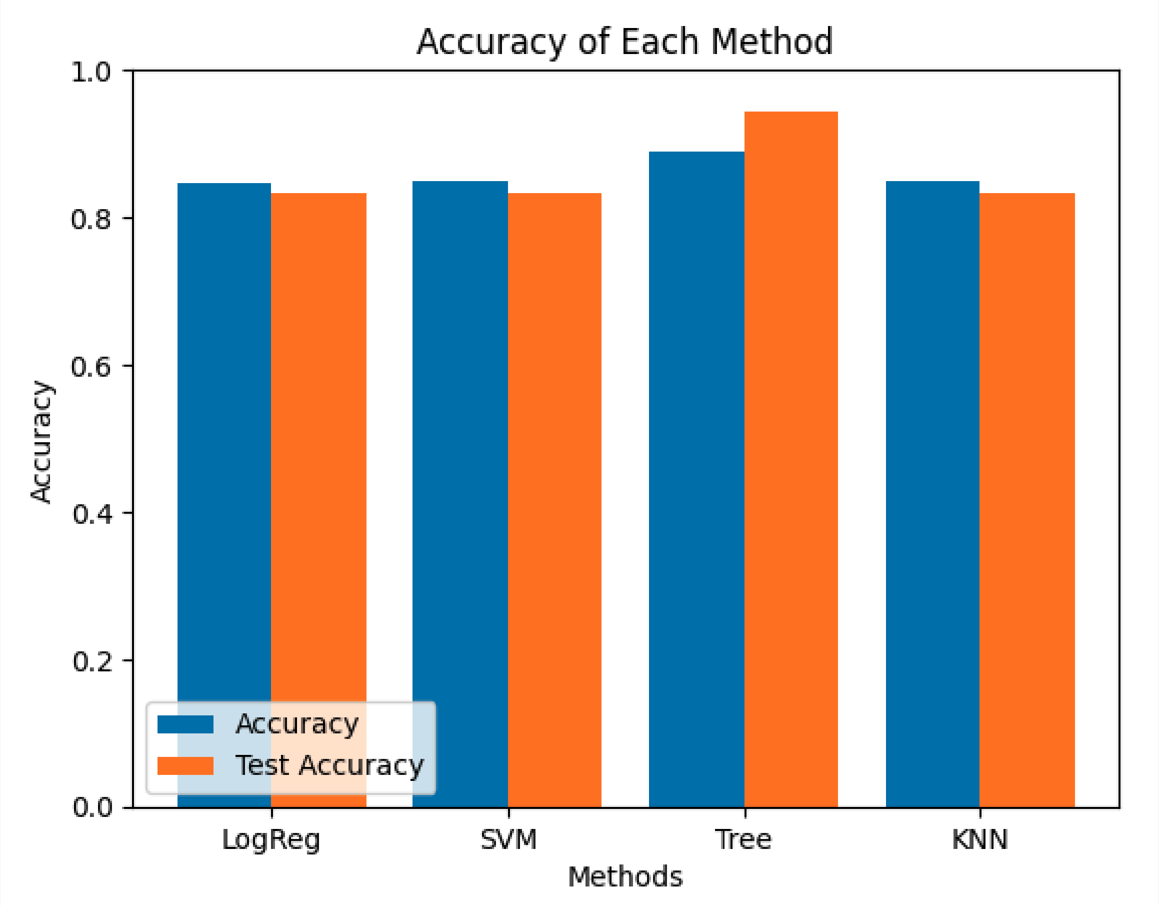
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

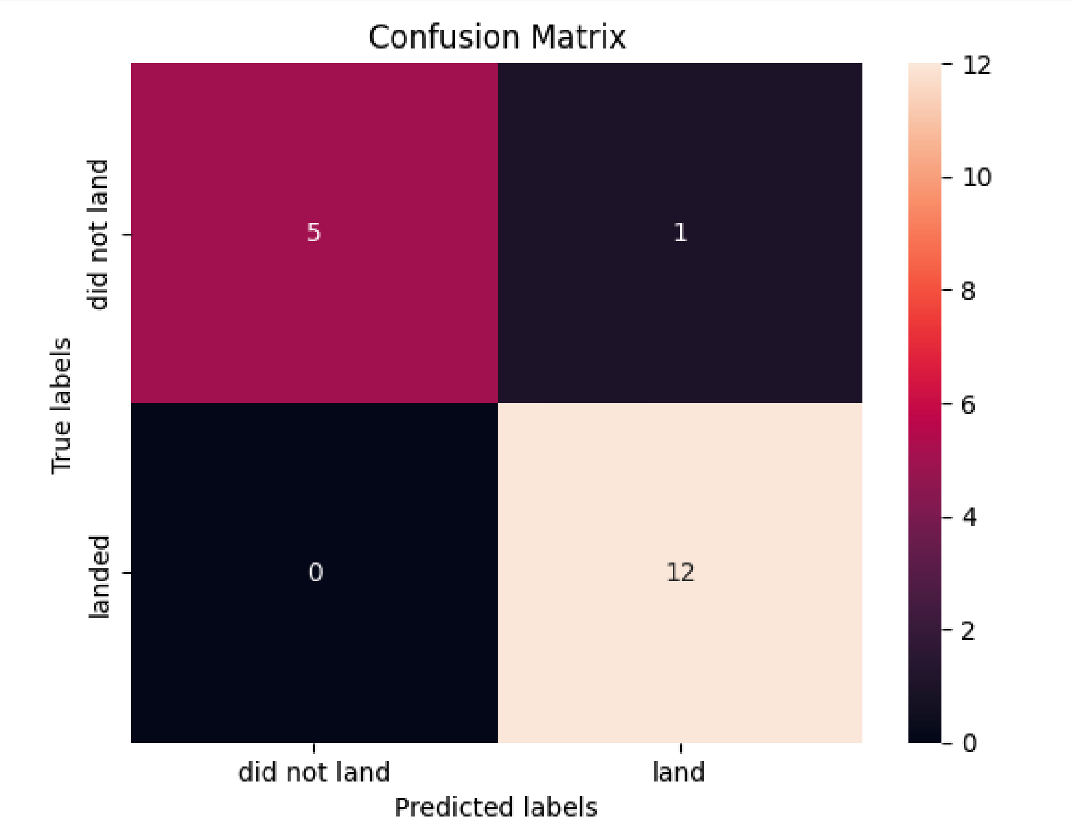
- Decision Tree model has the highest classification accuracy of 94%



https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

Confusion Matrix

- Confusion matrix of the best performing Decision Tree model
- The only error it had was a false positive while predicting the landed label



Conclusions

- The site for best launch success is KSC LC-39A
- Launch success has increased over the years
- Space Y should use machine learning analysis to predict launch success outcome. They can utilize Decision Tree classifier model to predict the outcome and plan.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
    - https://github.com/advaykumarsingh/coursera-ibm-data-science-professional-certificate/tree/main

Thank you!