IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ismael L. Lozano
28/03/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection from API and Web scraping
  - Data wrangling
  - Exploratory Data Analysis (EDA) with data visualization: Pandas and Matplotlib
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- **Project background and context**

SpaceX's Falcon 9 rocket family has revolutionized spaceflight with its reusable first-stage boosters, significantly reducing launch costs.

Since its first successful propulsive landing in 2015, the Falcon 9 has completed over 460 launches, with a remarkable success rate.

The Falcon Heavy, a heavy-lift derivative, enhances SpaceX's capabilities for deep-space and high-payload missions.

A major portion of Falcon 9 launches is dedicated to deploying Starlink satellites, alongside critical missions for NASA, and others. The rocket has delivered satellites to various orbits, including low Earth orbit (LEO), geostationary transfer orbit (GTO), and interplanetary destinations.

This project explores Python-based data analysis techniques to examine Falcon 9 launch success rates, cost efficiency, and mission performance. By leveraging machine learning and statistical methods, we aim to predict launch outcomes and gain insights into SpaceX's advancements in reusable rocket technology.

Introduction

- The aim of this project is to solve the following critic questions:
  - What influences if the rocket will land successfully?
  - The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
  - What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API: https://api.spacexdata.com/v4/rockets/
    - WebScraping: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analysing features. Thus, data was wrangled/cleaned in preparation for visualizations, queries, and machine learning model training.

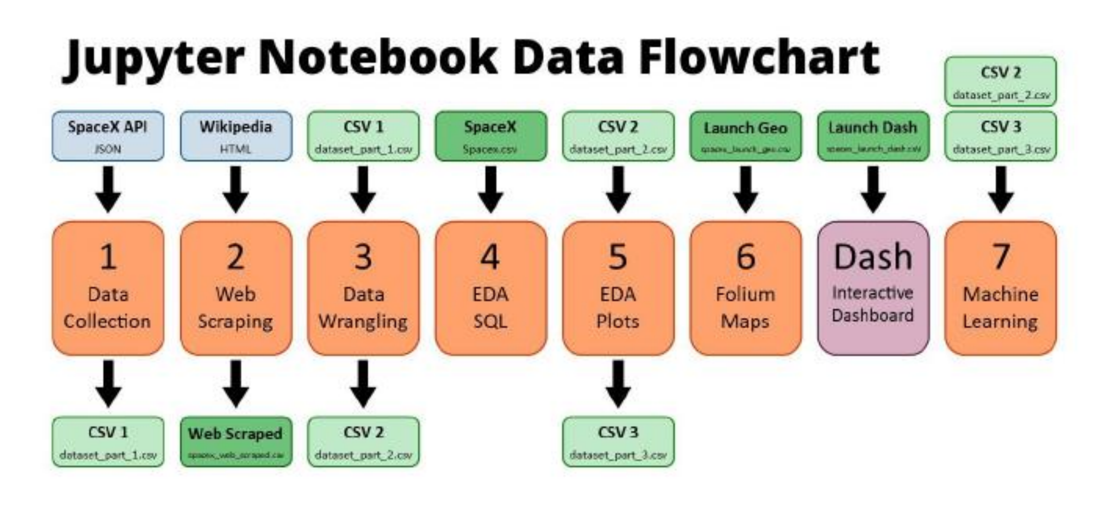
Methodology

Executive Summary

- Performed exploratory data analysis (EDA) using visualization and SQL
  - Plotting : Scatter Graphs, Bar Graphs to show relationships between variables to show patterns of data.
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Building, tuning, and evaluating classification models to determine if Falcon 9 will land successfully.

Data Collection

- Datasets were collected from Space X API and Wikipedia using web scraping technics, throw the following links:
  - https://api.spacexdata.com/v4/rockets/ rockets/) and from Wikipedia https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- From the above links, several csv were store and utilise in several steps of the course. See the below diagram for details.



Data Collection – SpaceX API

- SpaceX data was available publicly in the API endpoint (not affiliated with SpaceX): https://api.spacexdata.com/
- Data was extracted from the response from the API and loaded into a Pandas DataFrame for further analysis.
- The used code for this purpose can be followed in the following link:

https://github.com/lozais/DataScience-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Flowchart of API Data Processing

- Send GET request to API
- Extract nested data
- Convert date format
- Use defined functions to generate specific columns of data
- Combine separate columns into a DataFrame
- Filter out all launches with rockets other than the Falcon 9
- Handle missing values

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia, from HTML tables (https://en.wikipedia.org/wiki/SpaceX).
- Launch data was extracted from these tables and loaded into a Pandas DataFrame for further analysis.
- The code can be followed in the following link: https://github.com/lozais/DataScience-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Flowchart of Wikipedia Web Scraping

- Web Scrape the page to get the entire HTML text
- Create a BeautifulSoup object from the response text content
- Select the tables
- From the launch table, extract the column names from the <th></th> tags
- Create a Pandas DataFrame by parsing the launch tables

Data Wrangling

- The CSV file from the first section contained the data in need of cleaning/wrangling.
- Firstly, there was performed an initial Exploratory Data Analysis (EDA) on the dataset. Here the mission outcomes were processed and reformatted.
- Then, the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated, where 1 represented the Falcon 9 first stage landing being a success and 0 represented a failure.
- Finally, the new mission outcome classification column was added to the DataFrame. Follow the code in the following link:

https://github.com/lozais/DataScience-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Flowchart of Data Cleaning / Wrangling

- Load data from CSV file from earlier
- Find the number of launches at each site
- Find the quantity of each type of orbit
- Find the quantity of each type of mission outcome
- Create a DataFrame column from the mission outcome data
- Compile data into a single DataFrame

EDA with Data Visualization

- The following charts were created to look at Launch Site trends
  - Scatterplot to see the relationship between Flight Number and Payload Mass.
  - Scatterplot to see the relationship between Flight Number and Launch site.
  - Scatterplot to see the relationship between Payload Mass and Launch Site.
- The following charts were created to look at Orbit Type trends
  - Bar chart to visualize the relationship between success rate and Orbit Type.
  - Scatterplot to see mission outcome relationship between Flight Number and Orbit type.
  - Scatterplot to see mission outcome relationship between Payload Mass and Orbit type.
- The following chart was created to look at trends based on time
  - Line plot to see the mission launch success trend by year.
- Follow the code in the following link:

https://github.com/lozais/DataScience-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- We have performed SQL queries to gather information about the dataset.
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string ‘CCA’
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date where the successful landing outcome in ground pad was achieved.
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass.
  - Listing the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
  - 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.
- Follow the code in:

https://github.com/lozais/DataScience-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map objects were created and added to the Folium map
  - First, all launch sites were marked on the map
    - Circles were added for the launch sites
  - Second, Success/failed launches for each site have been marked on the map
    - Icon red for failure and green for successful launch were added
  - Finally, lines were added to show the distance to the nearby features:
    - To the coastline
    - To the city, railway and highway
- Follow the code in the following link:

  https://github.com/lozais/DataScience-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of success launches by site
  - Payload range
- This combination allowed to quickly analyse the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- Follow the code in the following link:

https://github.com/lozais/DataScience-Capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- The dataset was standardized and then splitted into training and testing sets.
- The following machine learning models were trained on the training data set:
  - Logistic Regression
  - SVM (Support Vector Machine)
  - Decision Tree
  - KNN (k-Nearest Neighbors)
- Hyper-parameters were evaluated using GridSearchCV() and the best was selected using the best_params method.
- Using the best hyper-parameters, each of the four models were scored on accuracy by using the testing data set. Additionally, the accuracy of each model was visually evaluated with a confusion matrix.
- Follow the code in the link:

https://github.com/lozais/DataScience-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

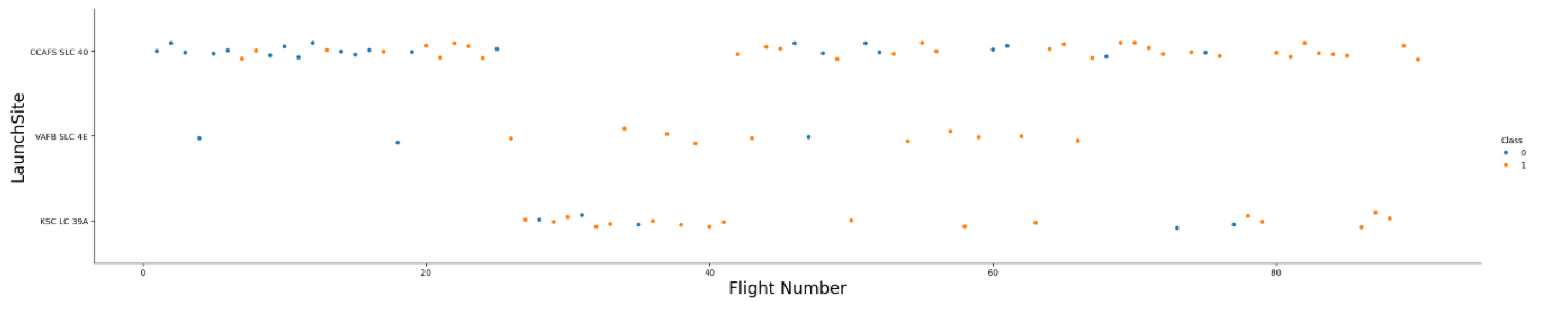
- Exploratory data analysis results
  - Space X uses 4 different launch sites.
  - The average payload of F9 v1.1 booster is 2,928 kg
  - The first success landing outcome happened in 2015 fiver year after the first launch
  - Almost 100% of mission outcomes were successful
- Interactive analytics demo in screenshots
  - 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.
- Predictive analysis results
  - After comparing accuracy of above methods, they all preformed practically# the same, except for tree which fit train data slightly better but test data worse.

Section 2

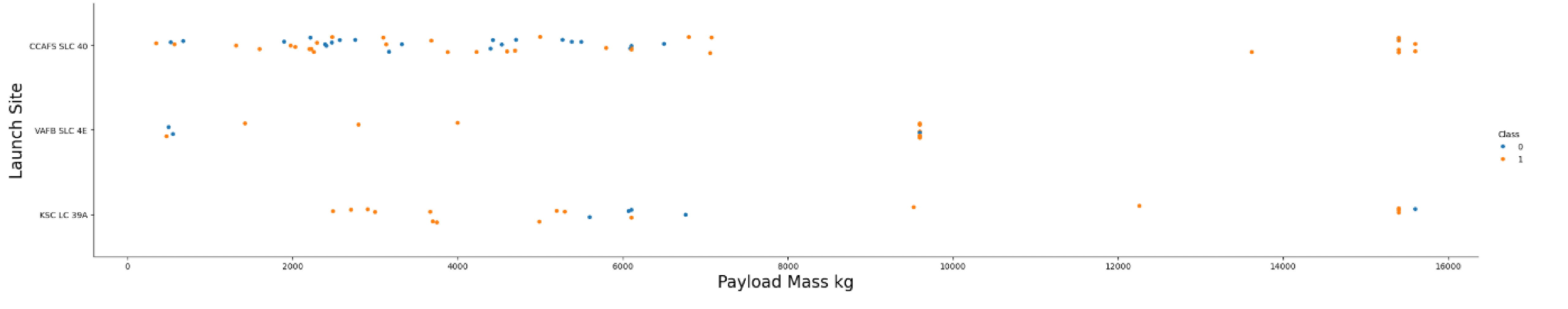
Insights drawn from EDA

Flight Number vs. Launch Site

- The best launch site nowadays is CCAF5 SLC 40 (see the figure below), where most of recent launches were successful.
- Second and third place for VAFB SLC 4E and KSC LC 39A respectively
-

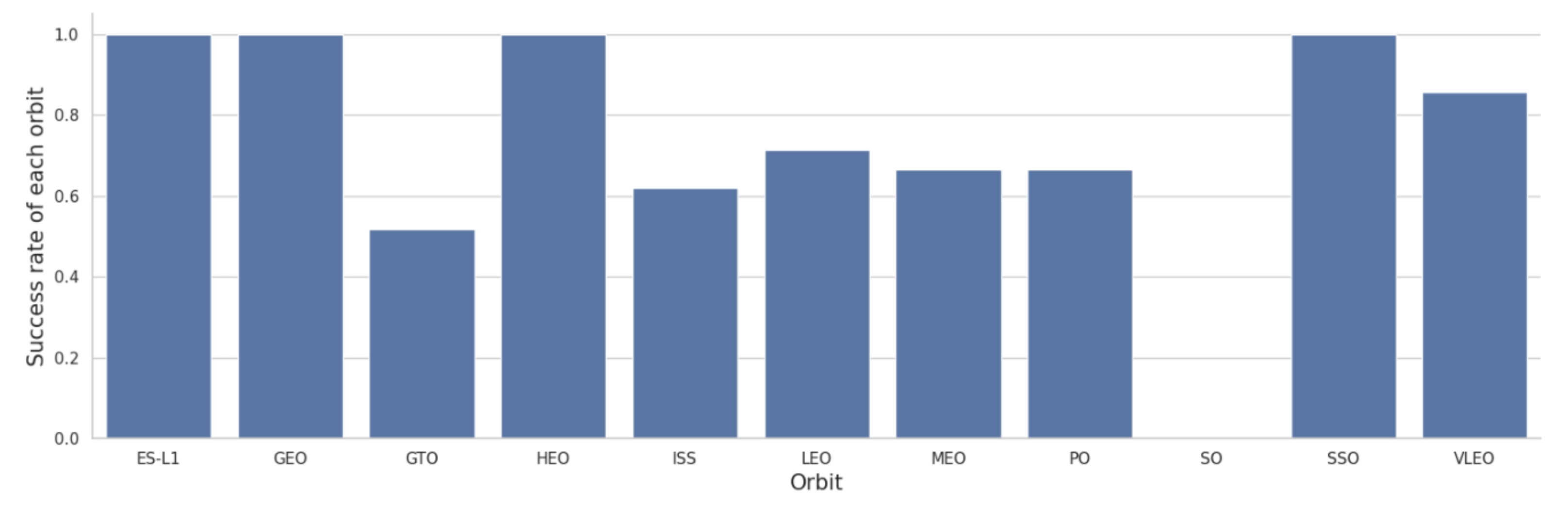


Payload vs. Launch Site



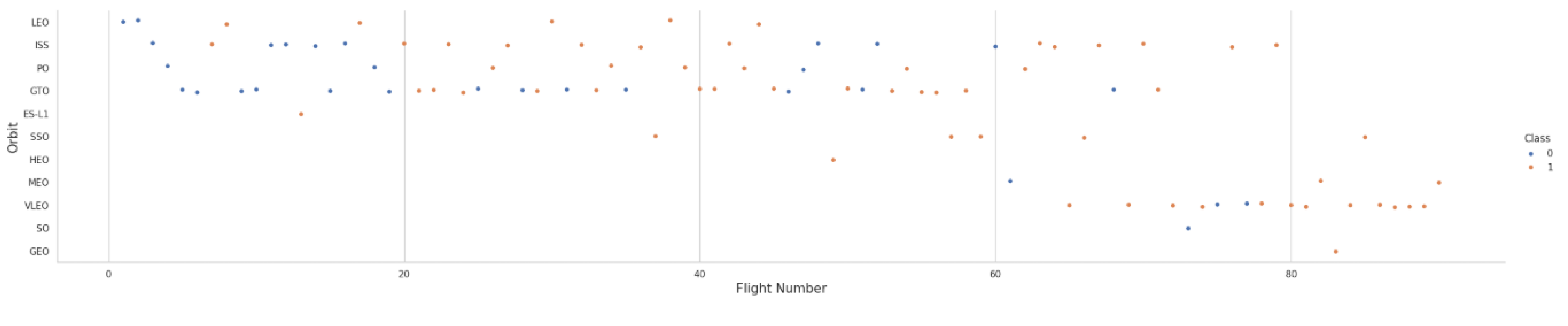
- Payloads over 9.000kg have excellent success rate
- Payloads over 12.000kg seems to be possible only at CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type



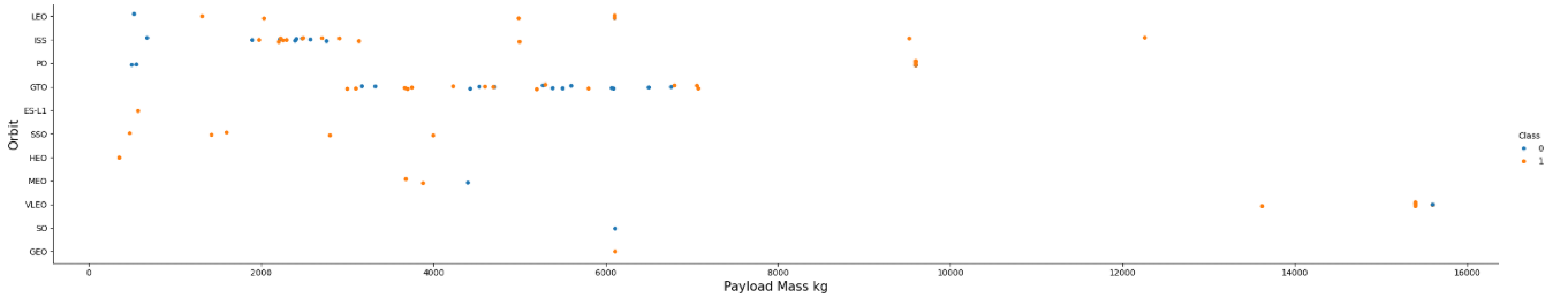
- The biggest success rates happens to orbits ESL1, GEO, HEO, and SSO.
- Followed by VLEO (>80%); and LFO (~70%).

Flight Number vs. Orbit Type



- Ssuccess rate improved over time to all orbits;

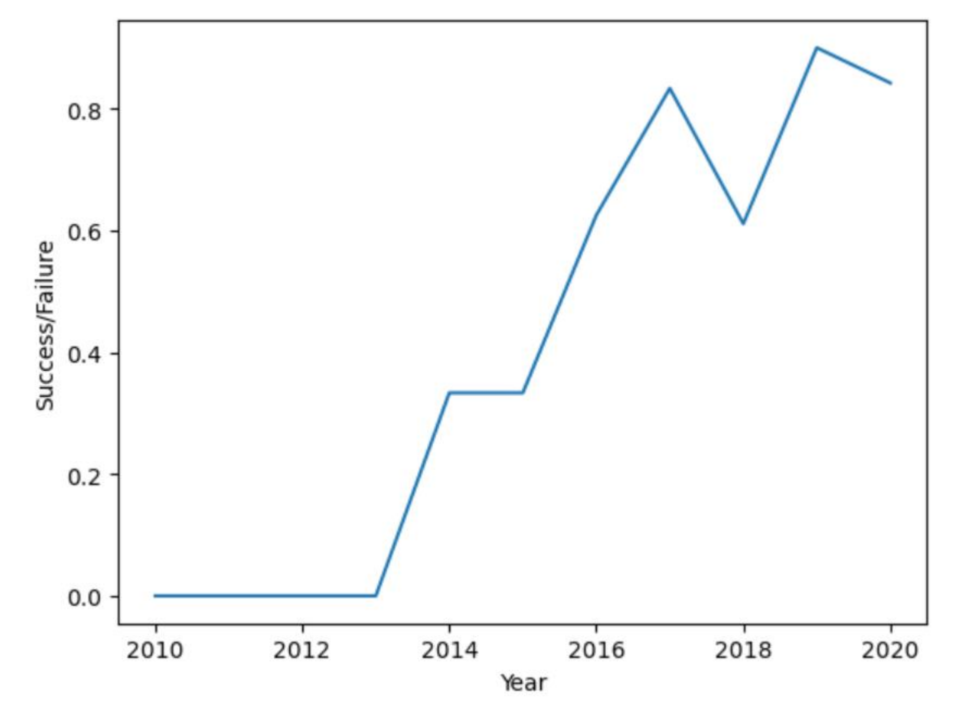
Payload vs. Orbit Type



- Over 10.000 kg payloads there is not success rate for most of the orbit

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

- Names of the unique launch sites, obtained with the query: “sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;”

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

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	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 attemp

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total Payload (kg)
111.268

- Total pay load calculated by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

Avg Payload (kg)
2.928

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928kg.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

Min Date
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Selecting distinct booster versions according to the filters these 4 are the result.

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

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:
- Grouping mission outcomes and counting records for each group led us to the summary as follows

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

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass registered in the dataset.

Booster Version (...)
F9 B5 B1048.4
[illegible] 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

2015 Launch Records

- List the 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 twooccurrences.

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

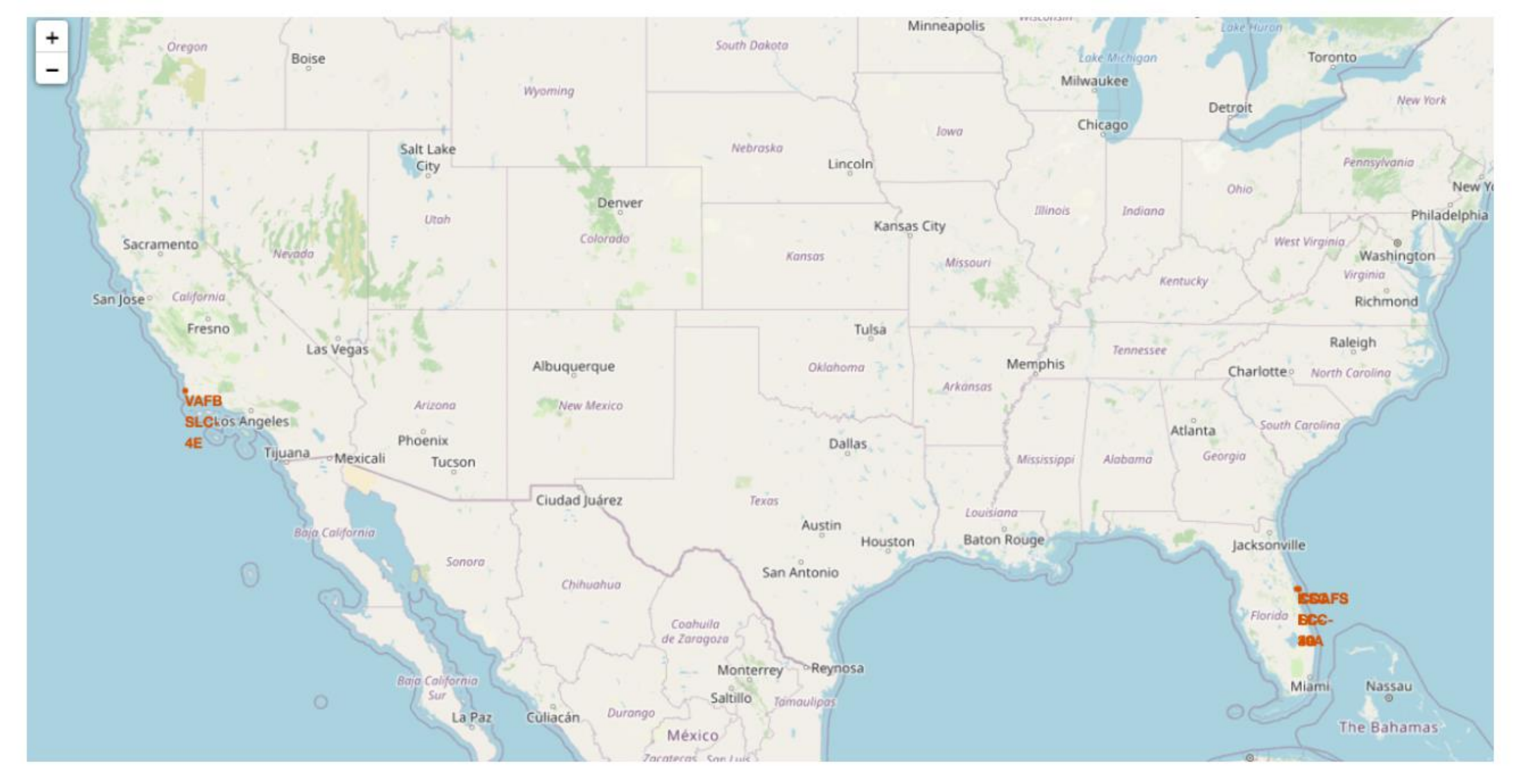
- This view of data alerts us that “No attempt” must be taken in account.

Section 3

Launch Sites Proximities Analysis

All Launch sites with Folium maps

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



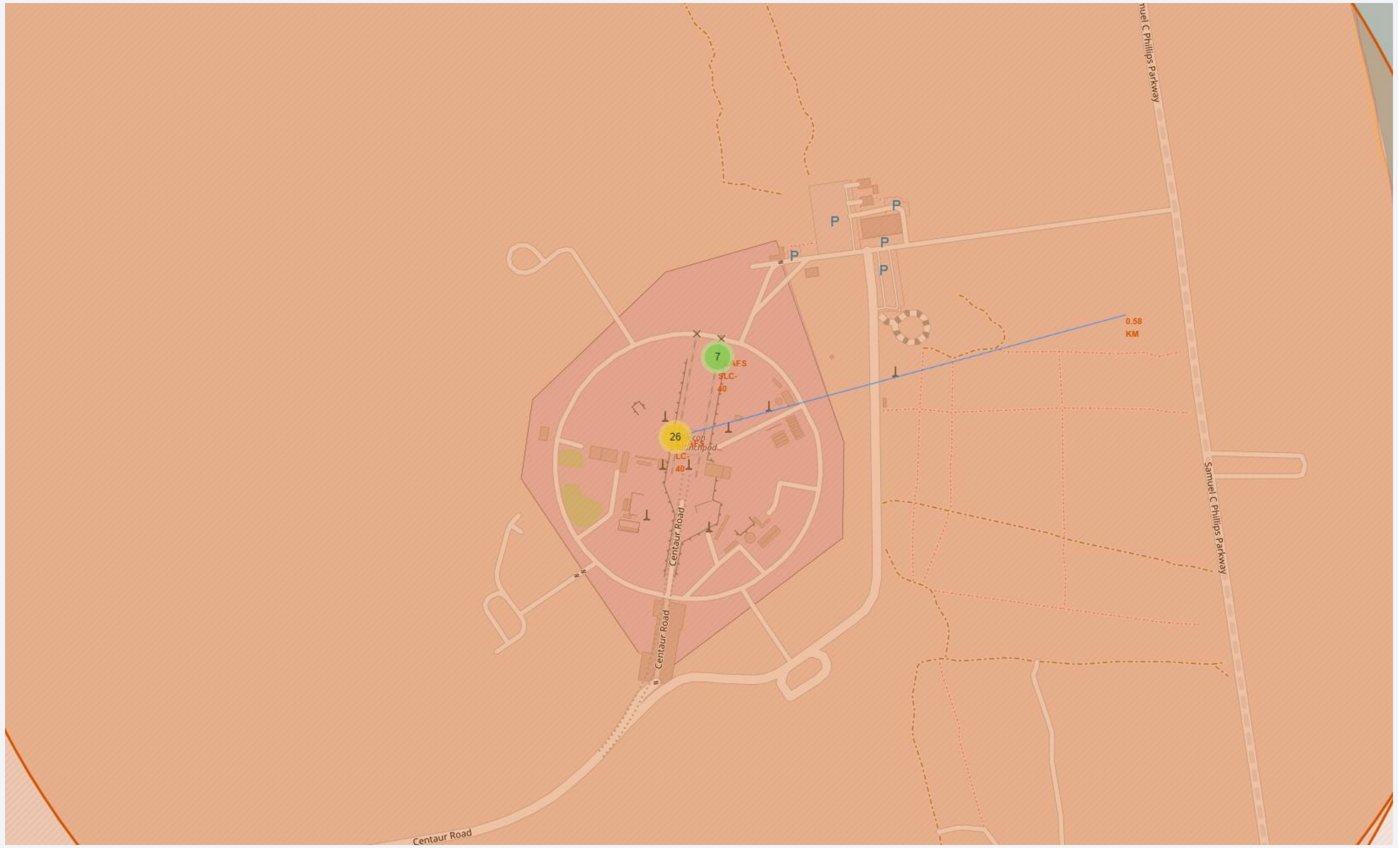
Launch outcomes by site

- Example of CCAFS SLC 40 launch site launch outcomes

- Green markers indicate successful and red ones indicate failure.

Logistics

- CCAFS SLC 40 has good logistics aspects, being near railroad and road and relatively far from inhabited areas.



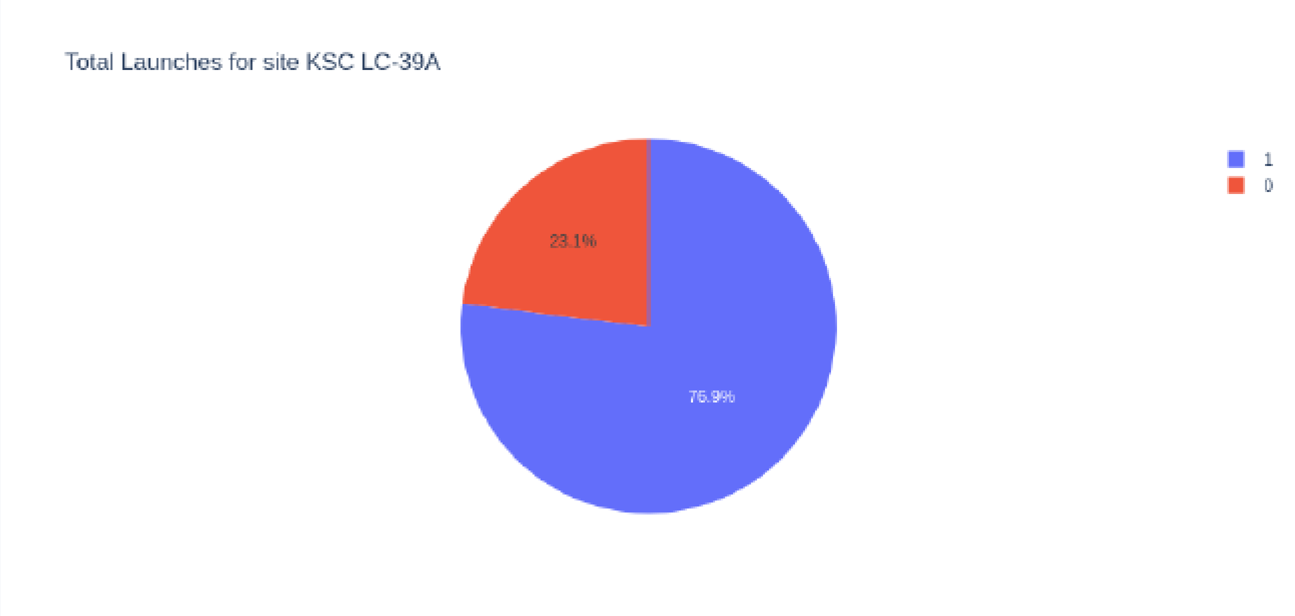
Section 4

Build a Dashboard with Plotly Dash

Successful Launches by site

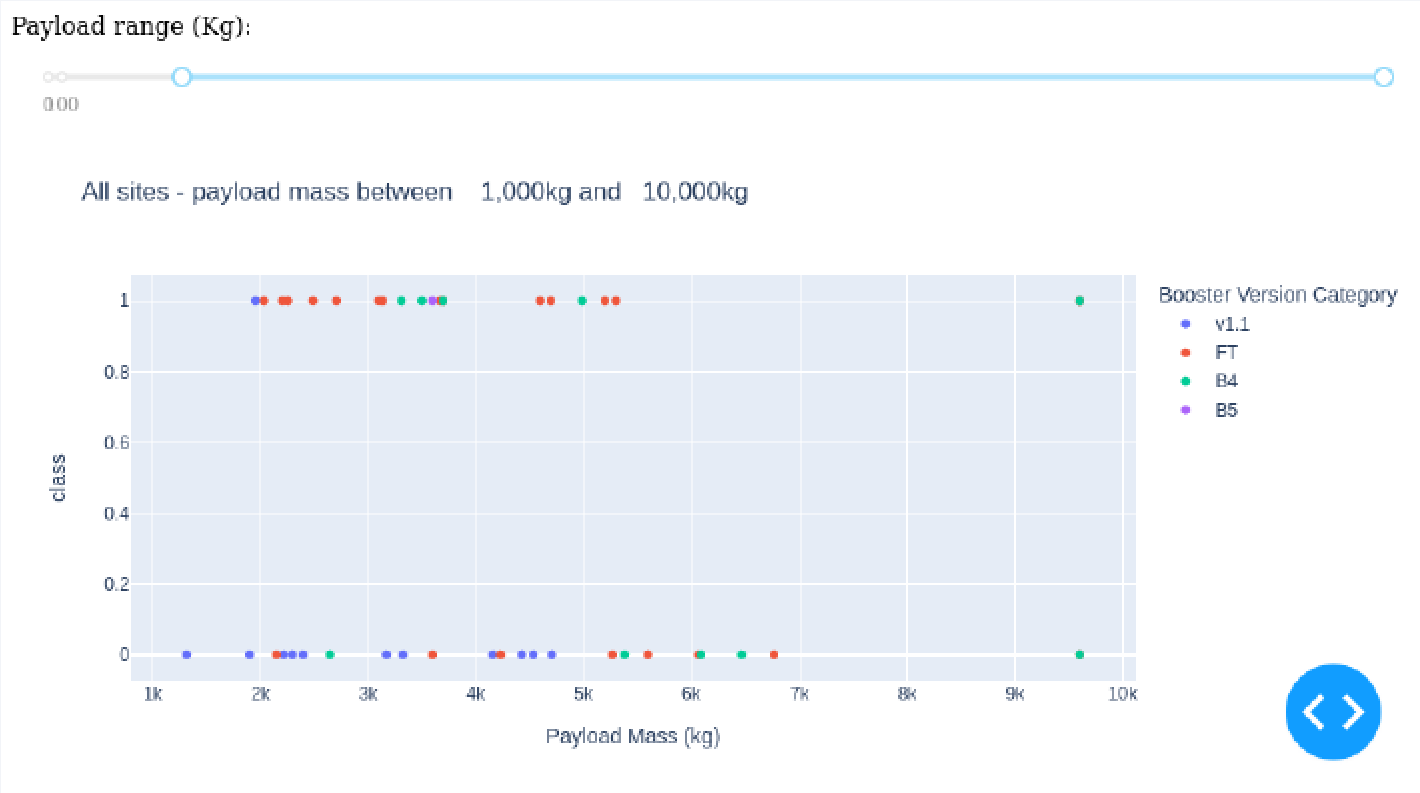
The place from where launches are done seems to be a very important factor of success of missions.

High Launch Success ratio site



76.9% of launches are successful in KSC LC 39A.

Payload vs Launch outcome



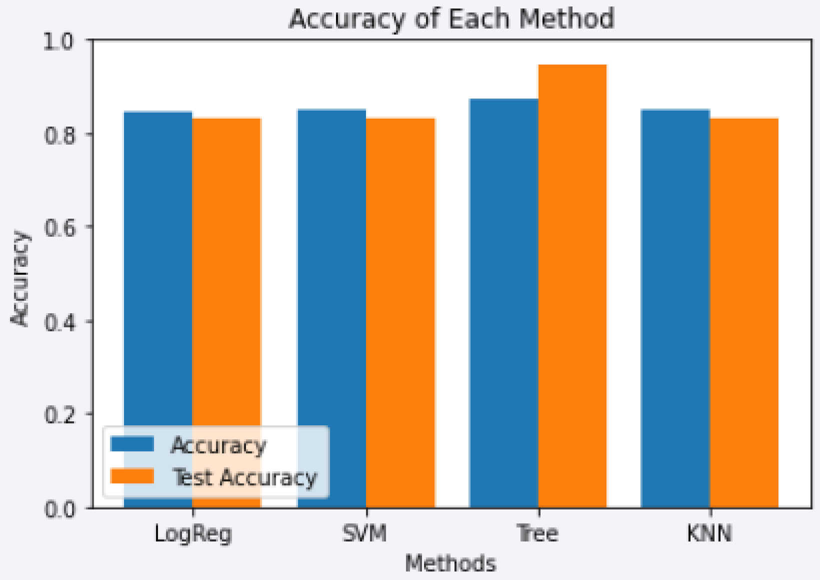
Payloads under 6.000kg and FT boosters are the most successful combination.

Section 5

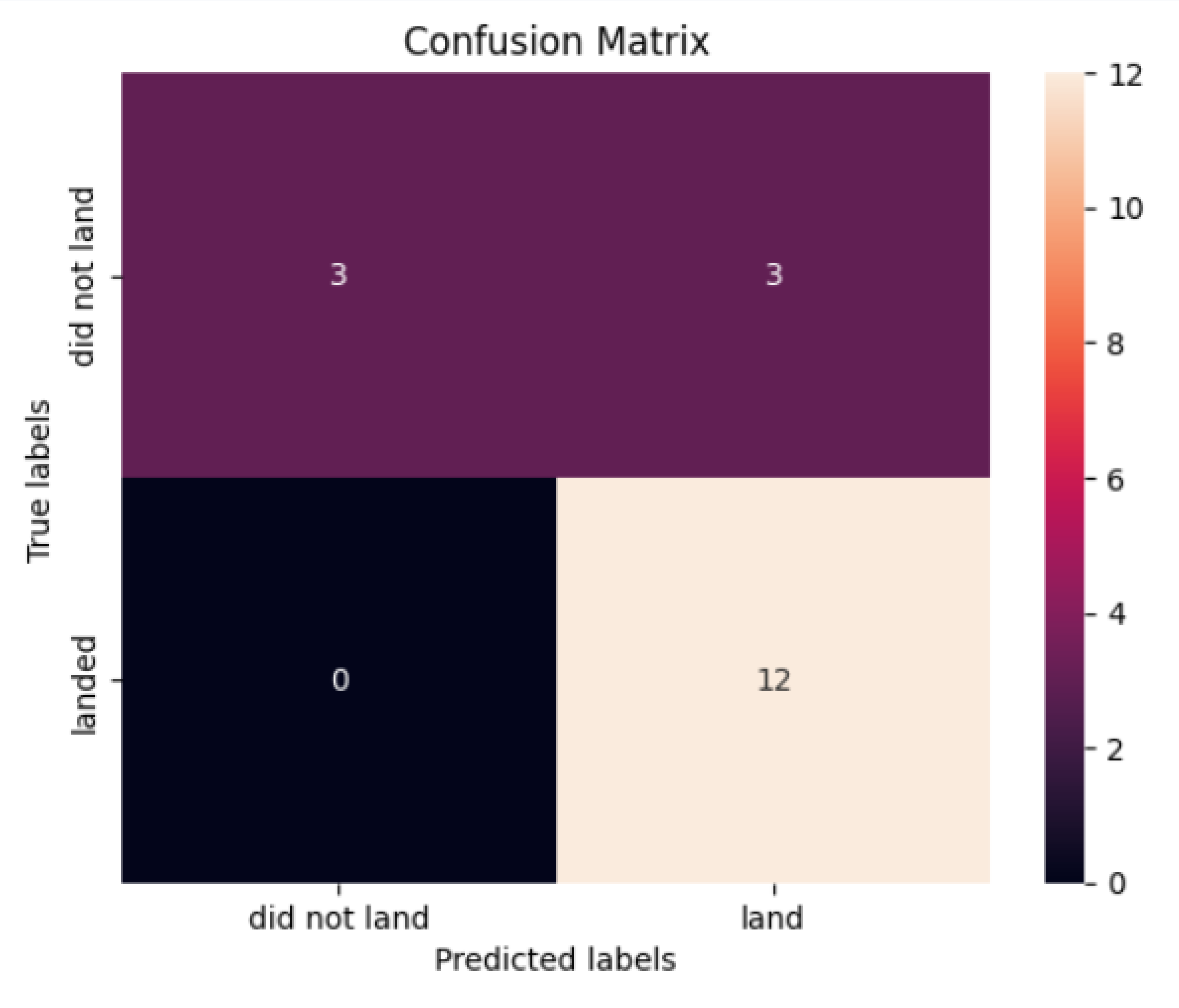
Predictive Analysis (Classification)

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

- Different data sources were analysed, refining conclusions along the process
- The best launch site is KSC LC 39A
- Launches above 7.000kg are less risky
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- All the jupyter notebooks files can be found in the following github:
- https://github.com/lozais/DataScience-Capstone/tree/main

Thank you!