IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Veaceslav Zagaevschi

23.03.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  The website of Space X showcases Falcon 9 rocket launches at a price of 62 million dollars, whereas other providers charge over 165 million dollars per launch. Space X achieves substantial cost savings by reusing the first stage of their rockets. Hence, by assessing the probability of a successful first-stage landing, one can determine the launch cost. Such information is invaluable for companies interested in competing with Space X for rocket launches. The project's objective is to develop a machine learning pipeline that can accurately predict the likelihood of a successful first-stage landing.

- Problems you want to find answers

  - What are the factors that influence the successful landing of a rocket?
  - The successful landing rate of a rocket depends on the interplay of multiple factors.
  - What are the necessary operating conditions to ensure a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from Wikipedia using SpaceX API and web scraping.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built, tuned, and evaluated classification models.

Data Collection

- Different techniques were utilized to gather the information.
  - We utilized the SpaceX API to retrieve the data through a get request.
  - Afterwards, we converted the response content, which was in Json format, to a pandas dataframe with the help of the .json() and .json_normalize() functions.
  - The data was then subjected to a cleaning process where we checked for missing values and filled them in when necessary.
  - We conducted web scraping from Wikipedia for Falcon 9 launch records using BeautifulSoup.
  - Our aim was to extract the launch records in the form of an HTML table, parse it, and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

- GitHub URL https://github.com/vzagaevschi/Data-Analysis-Courses/blob/main/IBM%20Data%20Science%20Professional%20Certificate/Applied%20Data%20Science%20Capstone/Week1/jupyter-labs-spacex-data-collection-api.ipynb

Task 1: Request and parse the SpaceX launch data using the GET request

Task 2: Filter the dataframe to only include Falcon 9 launches

Task 3: Dealing with Missing Values

Data Collection - Scraping

- Applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup
- Parsed the table and converted it into a pandas dataframe.
- GitHub URL https://github.com/vzagaevschi/Data-Analysis-Courses/blob/main/IBM%20Data%20Science%20Professional%20Certificate/Applied%20Data%20Science%20Capstone/Week1/jupyter-labs-webscraping.ipynb

TASK 1: Request the Falcon9 Launch Wiki page from its URL

TASK 2: Extract all column/variable names from the HTML table header

TASK 3: Create a data frame by parsing the launch HTML tables

Data Wrangling

- Exploratory data analysis was conducted, and the training labels were determined.
- The analysis involved calculating the number of launches at each site, as well as the number and frequency of each orbit.
- Created a landing outcome label from the outcome column and exported the results to a csv file.
- GitHub URL https://github.com/vzagaevschi/Data-Analysis-Courses/blob/main/IBM%20Data%20Science%20Professional%20Certificate/Applied%20Data%20Science%20Capstone/Week1/labs-jupyter-spacex-Data%20wrangling.ipynb

TASK 1: Calculate the number of launches on each site

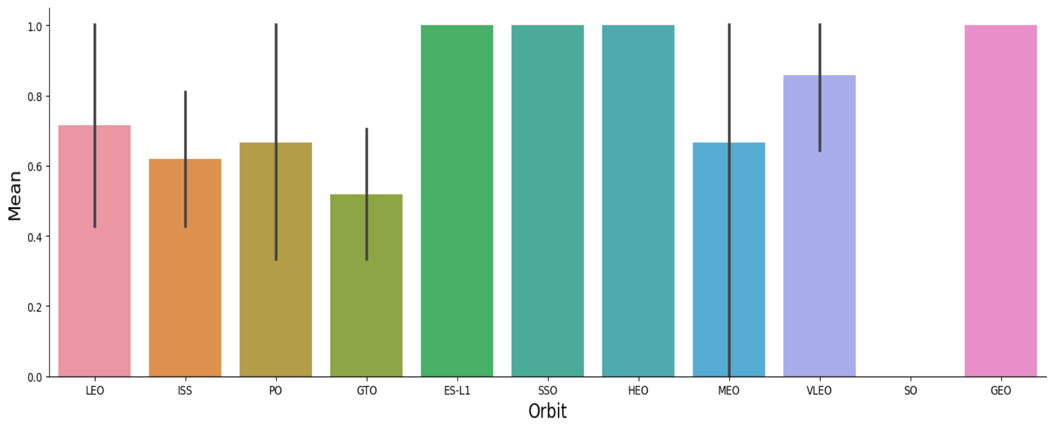
TASK 2: Calculate the number and occurrence of each orbit

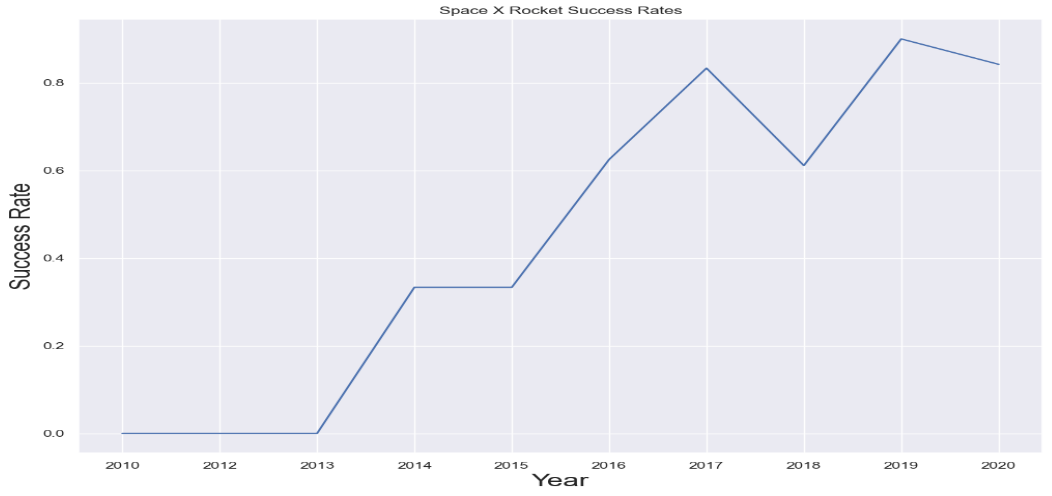
TASK 3: Calculate the number and occurrence of mission outcome per orbit type

TASK 4: Create a landing outcome label from Outcome column

EDA with Data Visualization

- We conducted data exploration by creating visualizations that highlighted the relationships between various data points.
- Specifically, we examined the relationship between flight number and launch site, as well as the connection between payload and launch site.
- We also visualized the success rate of each orbit type, investigated the connection between flight number and orbit type, and analyzed the yearly trend in launch success rates.
- GitHub URL
  https://github.com/vzagaevschi/Data-Analysis-Courses/blob/main/IBM%20Data%20Science%20Professional%20Certificate/Applied%20Data%20Science%20Capstone/Week2/jupyter-labs-eda-dataviz.ipynb





EDA with SQL

- Loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.
- Applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.
- GitHub URL https://github.com/vzagaevschi/Data-Analysis-Courses/blob/main/IBM%20Data%20Science%20Professional%20Certificate/Applied%20Data%20Science%20Capstone/Week2/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Task 1: Display the names of the unique launch sites in the space mission

Task 2: Display 5 records where launch sites begin with the string 'CCA'

Task 3: Display the total payload mass carried by boosters launched by NASA (CRS)

Task 4: Display average payload mass carried by booster version F9 v1.1

Task 5: List the date when the first succesful landing outcome in ground pad was acheived.

Task 6: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Task 7: List the total number of successful and failure mission outcomes

Task 8: List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

Task 9: 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.

Task 10: Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

- We incorporated several map objects, including markers, circles, and lines, into the folium map to denote the success or failure of launches at each launch site.
- We classified the launch outcomes as either 0 for failure or 1 for success, and then used marker clusters labeled with colors to identify which launch sites had higher success rates.
- We also computed the distances between each launch site and its surrounding areas, answering questions such as whether launch sites were located near railways, highways, or coastlines, and whether they maintained a certain distance from nearby cities.
- GitHub URL https://github.com/vzagaevschi/Data-Analysis-Courses/blob/main/IBM%20Data%20Science%20Professional%20Certificate/Applied%20Data%20Science%20Capstone/Week3/lab_jupyter_launch_site_location.ipynb

TASK 1: Mark all launch sites on a map

TASK 2: Mark the success/failed launches for each site on the map

TASK 3: Calculate the distances between a launch site to its proximities

Build a Dashboard with Plotly Dash

- We developed an interactive dashboard using Plotly Dash.
- To provide a visual representation of the total number of launches at specific sites, we created pie charts.
- We also used scatter graphs to demonstrate the correlation between outcome and payload mass (Kg) for various booster versions.GitHub URL https://github.com/vzagaevschi/Data-Analysis-Courses/blob/main/IBM%20Data%20Science%20Professional%20Certificate/Applied%20Data%20Science%20Capstone/Week3/plotly.py

TASK 1: Add a Launch Site Drop-down Input Component

TASK 2: Add a callback function to render success-pie-chart based on selected site dropdown

TASK 3: Add a Range Slider to Select Payload

TASK 4: Add a callback function to render the success-payload-scatter-chart scatter plot

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model
- GitHub URL https://github.com/vzagaevschi/Data-Analysis-Courses/blob/main/IBM%20Data%20Science%20Professional%20Certificate/Applied%20Data%20Science%20Capstone/Week4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

TASK 1: Create a NumPy array from the column Class

TASK 2: Standardize the data in X then reassign it to the variable X

TASK 3: Use the function train_test_split to split the data X and Y into training and test data.

TASK 4: Create a logistic regression object

TASK 5: Calculate the accuracy

TASK 6: Create a support vector machine object

TASK 7: Calculate the accuracy on the test data

TASK 8: Create a decision tree classifier object

TASK 9: Calculate the accuracy of tree_cv on the test data

TASK 10: Create a k nearest neighbors object

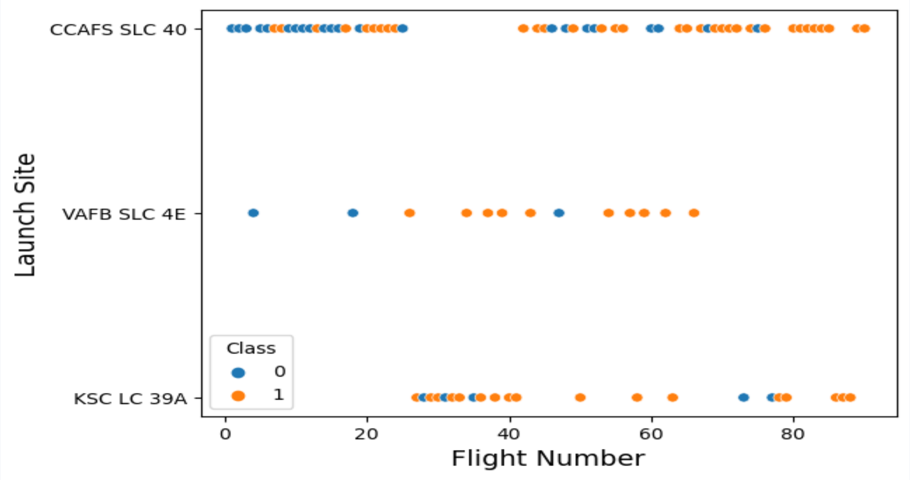
TASK 11: Calculate the accuracy of tree_cv on the test data

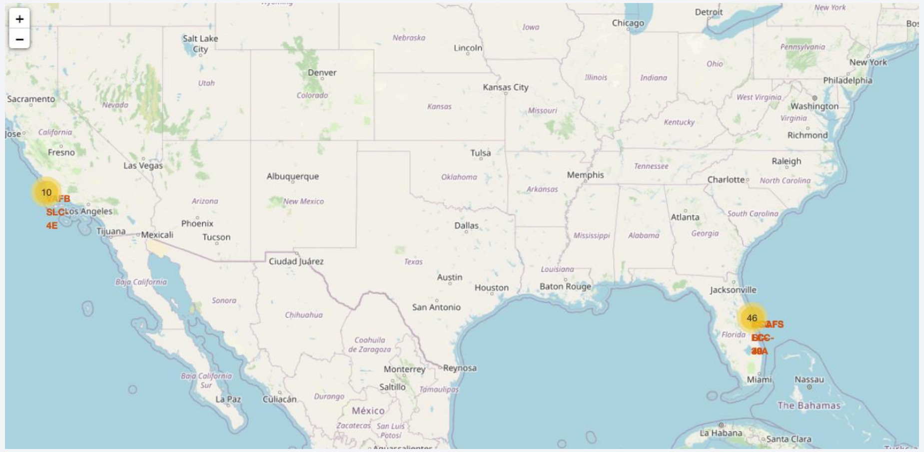
TASK 12: Find the method performs best

Results

- Exploratory data analysis results
    - After making more than 20 flights (aproximatively) there is a increase of succesfull flights. After 80 flights there was no failures.
    - There are more succsesful flights with increased payload. Success could not be related to payload mass but to increased experience and number of flights. Perhaps payload increased gradualy with flights numbers.
    - Launch sites are strategically located near the equator and coastline due to practical reasons. The proximity to the equator allows for efficient use of fuel during space launches, taking advantage of Earth's rotation. Additionally, launch sites situated near the coast are a reasonable safety precaution.
- Interactive analytics demo in screenshots
- Predictive analysis results
    - The best performing method is Decision Tree and has a score of 0.9142857142857143

	Accuracy
KNN	0.848214
Decision Tree	0.914286
Logistic Regression	0.846429
SVM	0.848214



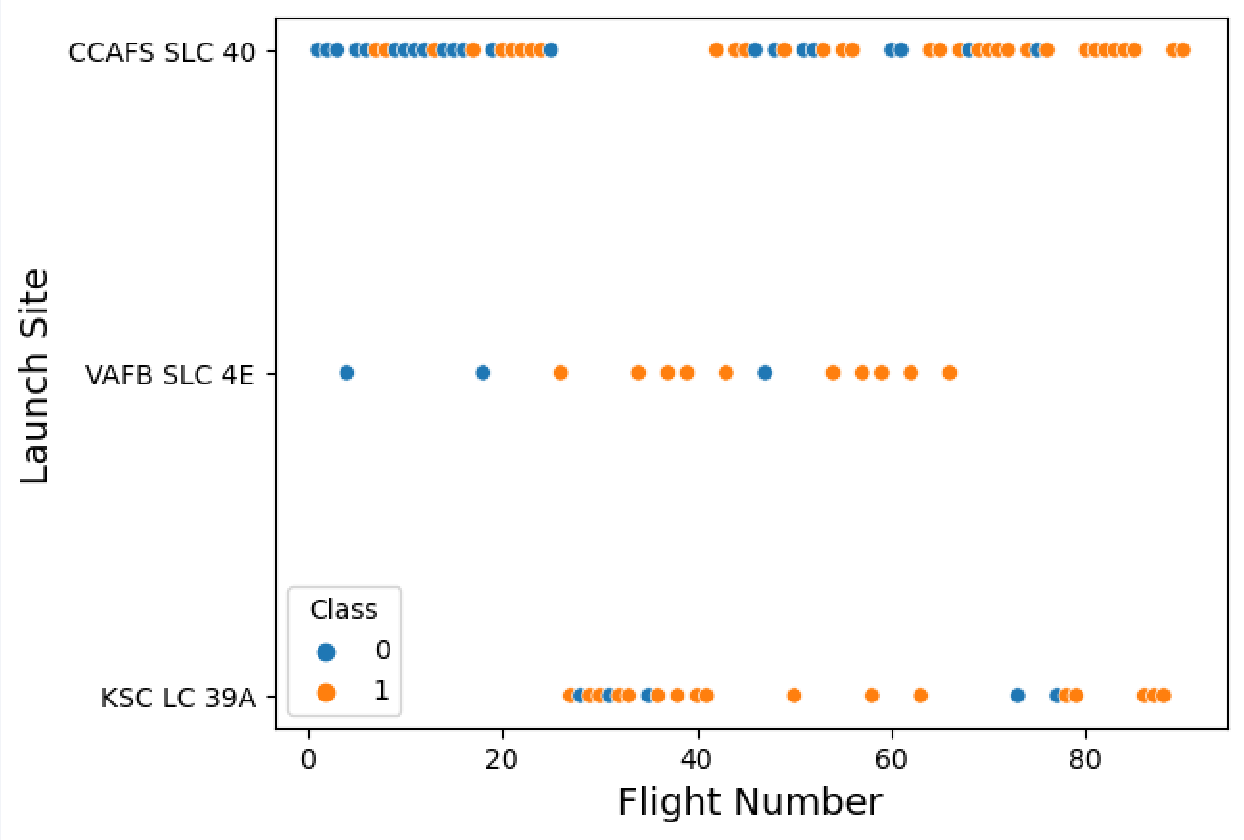


Section 2

Insights drawn from EDA

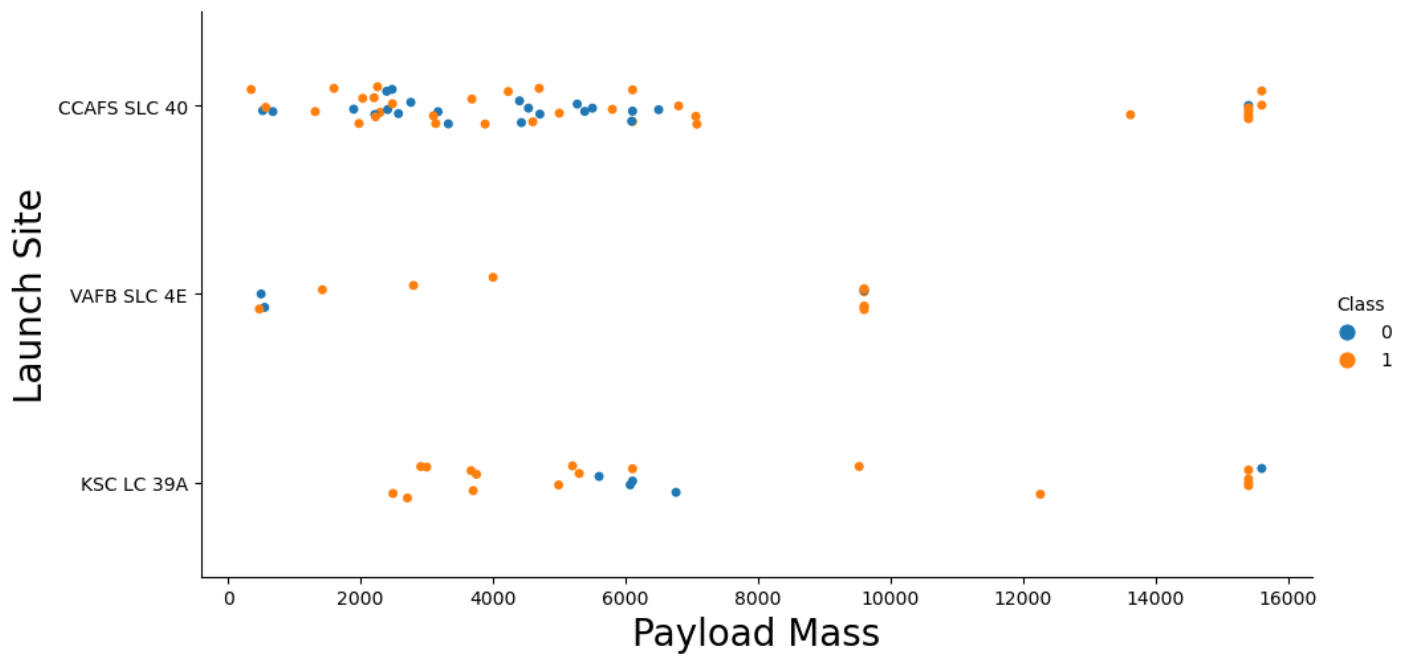
Flight Number vs. Launch Site

- After making more than 20 flights (approximatively) there is a increase of successful flights.
- After 80 flights there was no failures.



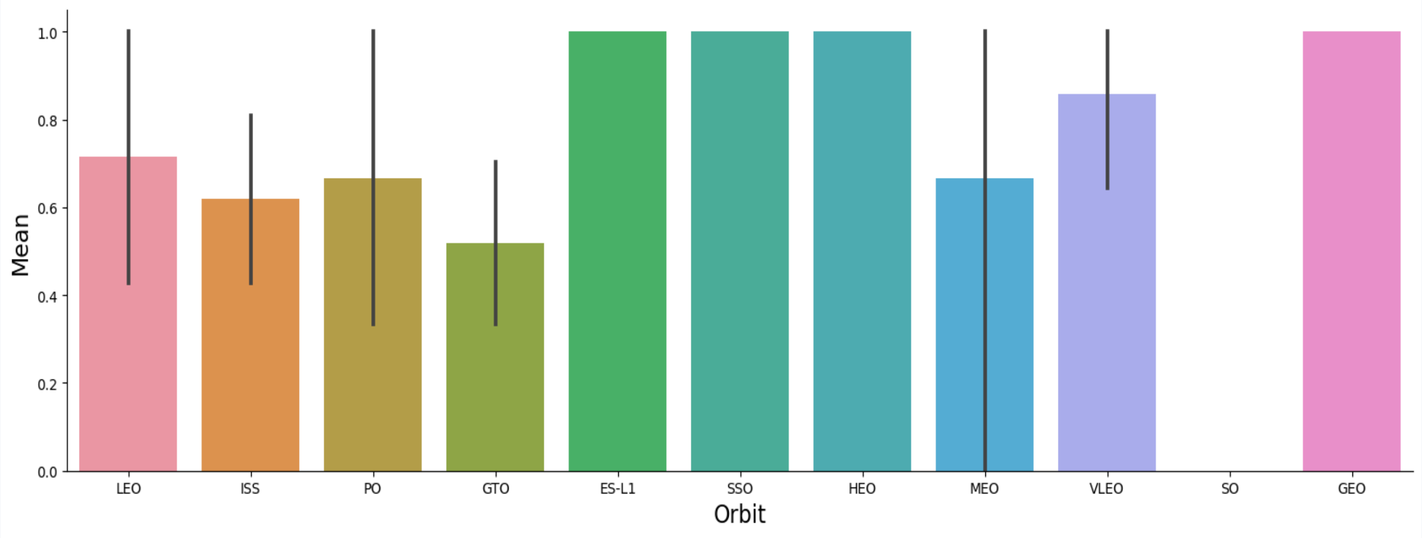
Payload vs. Launch Site

- There are more successful flights with increased payload. Success could not be related to payload mass but to increased experience and number of flights.
- Perhaps payload increased gradually with flights numbers.



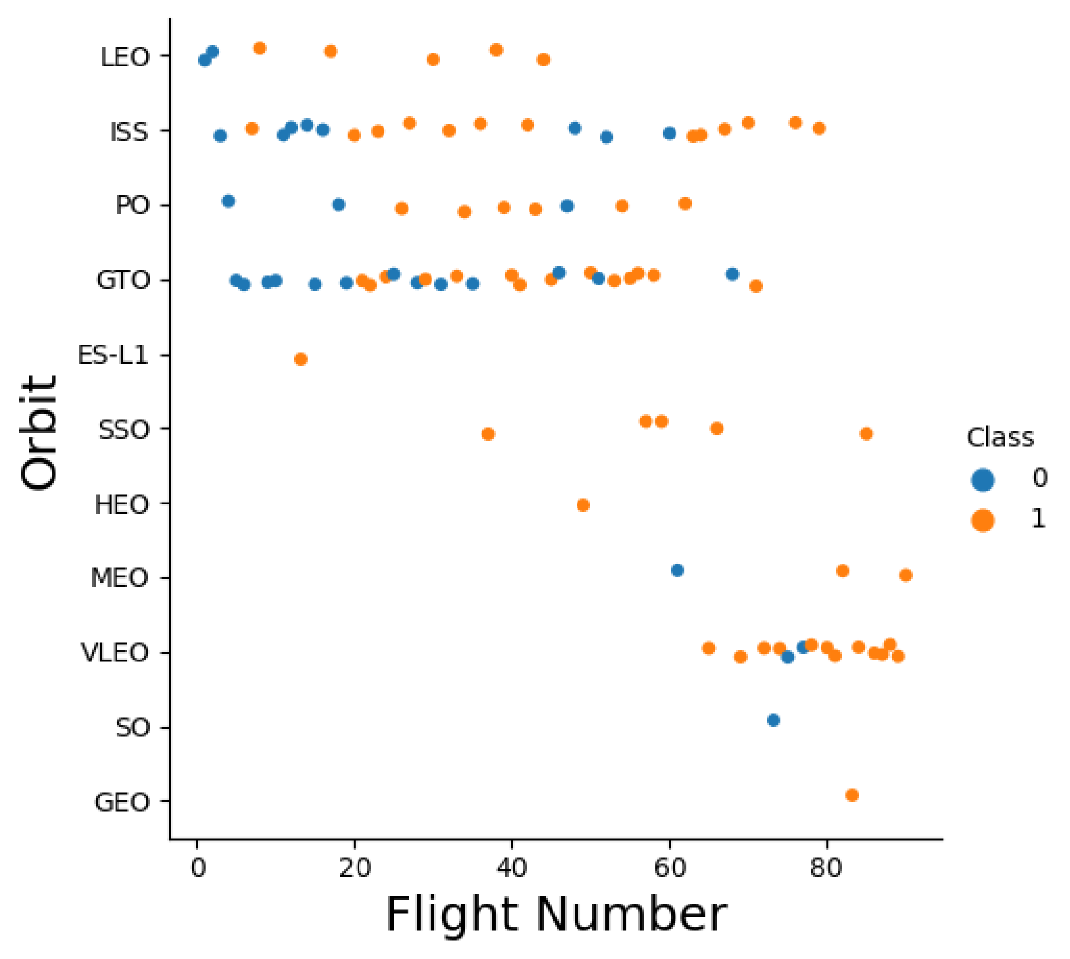
Success Rate vs. Orbit Type

- Highest Success rates: ES-L1, GEO, HEO, SSO. Lowest Sucess rates: SO.



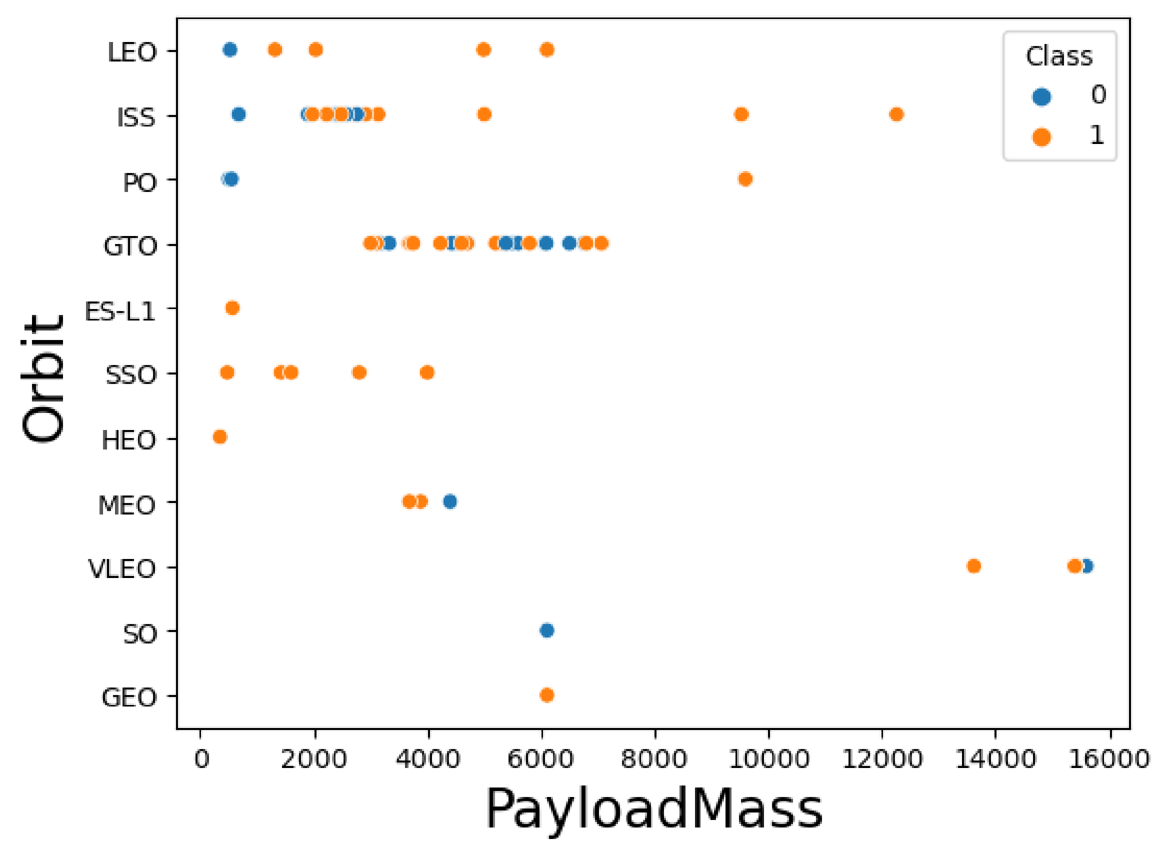
Flight Number vs. Orbit Type

- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



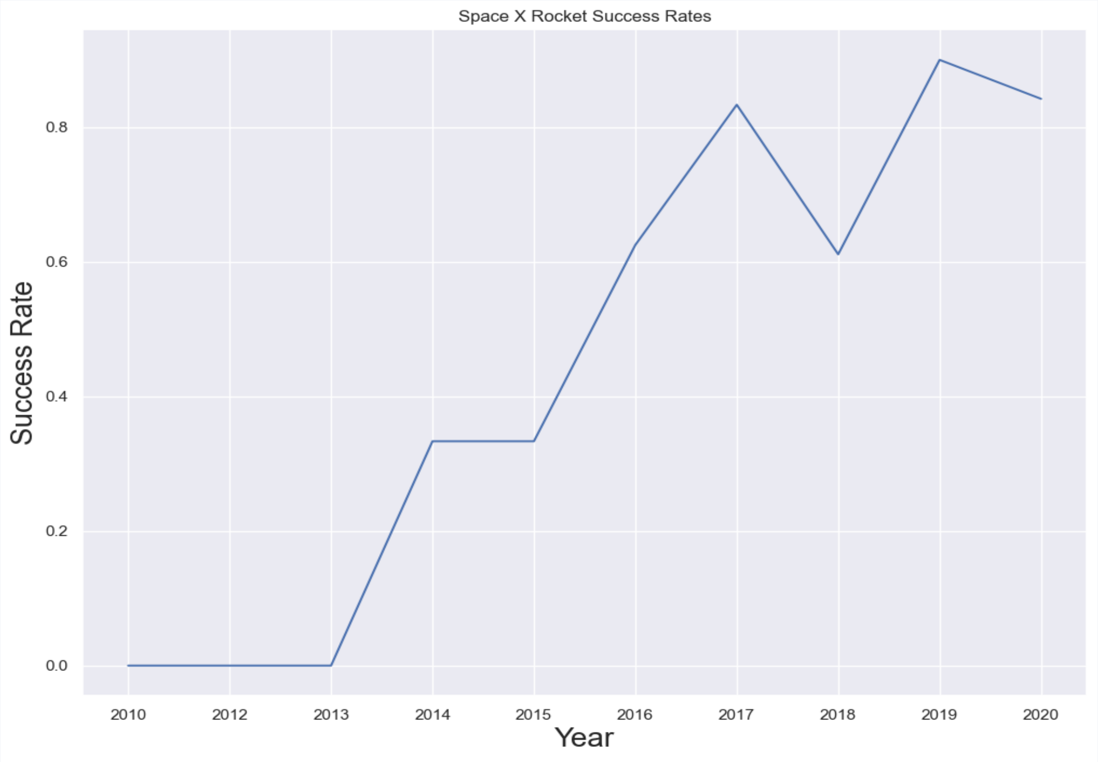
Payload vs. Orbit Type

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



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2020



All Launch Site Names

- Used the key word **DISTINCT** to show only unique launch sites from the SpaceX data.

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

Launch Site Names Begin with 'CCA'

- Used the query to display 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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculated the total payload carried by boosters from NASA as 45596

- SELECT SUM(PAYLOAD_MASS__KG_)
  FROM SPACEXTBL
  WHERE Customer="NASA (CRS)";

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

- SELECT AVG(PAYLOAD_MASS__KG_)
  FROM SPACEXTBL
  WHERE Booster_Version="F9 v1.1";

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad was 01.05. 2015
- SELECT MIN(Date)
  FROM SPACEXTBL
  WHERE "Landing _Outcome" = 'Success (ground pad)';

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000
- SELECT DISTINCT Booster_Version
  FROM SPACEXTBL
  WHERE "Landing _Outcome" = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);

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

Total Number of Successful and Failure Mission Outcomes

- Used wildcard like ‘%’ to filter for **WHERE** MissionOutcome was a success or a failure and UNION.

- SELECT 'Success' as Result, COUNT(*)
  FROM SPACEXTBL
  WHERE Mission_Outcome LIKE "Success%“
  UNION
  SELECT 'Failure', COUNT(*)
  FROM SPACEXTBL
  WHERE Mission_Outcome = "Failure (in flight)";

Result	COUNT(*)
Failure	1
Success	100

Boosters Carried Maximum Payload

- Determined the booster that have carried the maximum payload using a subquery in the **WHERE** clause and the **MAX()** function.
- SELECT DISTINCT Booster_Version
  FROM SPACEXTBL
  WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

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

2015 Launch Records

- Used a **WHERE** clause, **SUBSTR func** to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015
- SELECT SUBSTR(Date, 4, 2), Booster_Version, Launch_Site
  FROM SPACEXTBL
  WHERE SUBSTR(Date, 7, 4) = '2015'
  AND "Landing _Outcome" = 'Failure (drone ship)';

SUBSTR(Date, 4, 2)	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

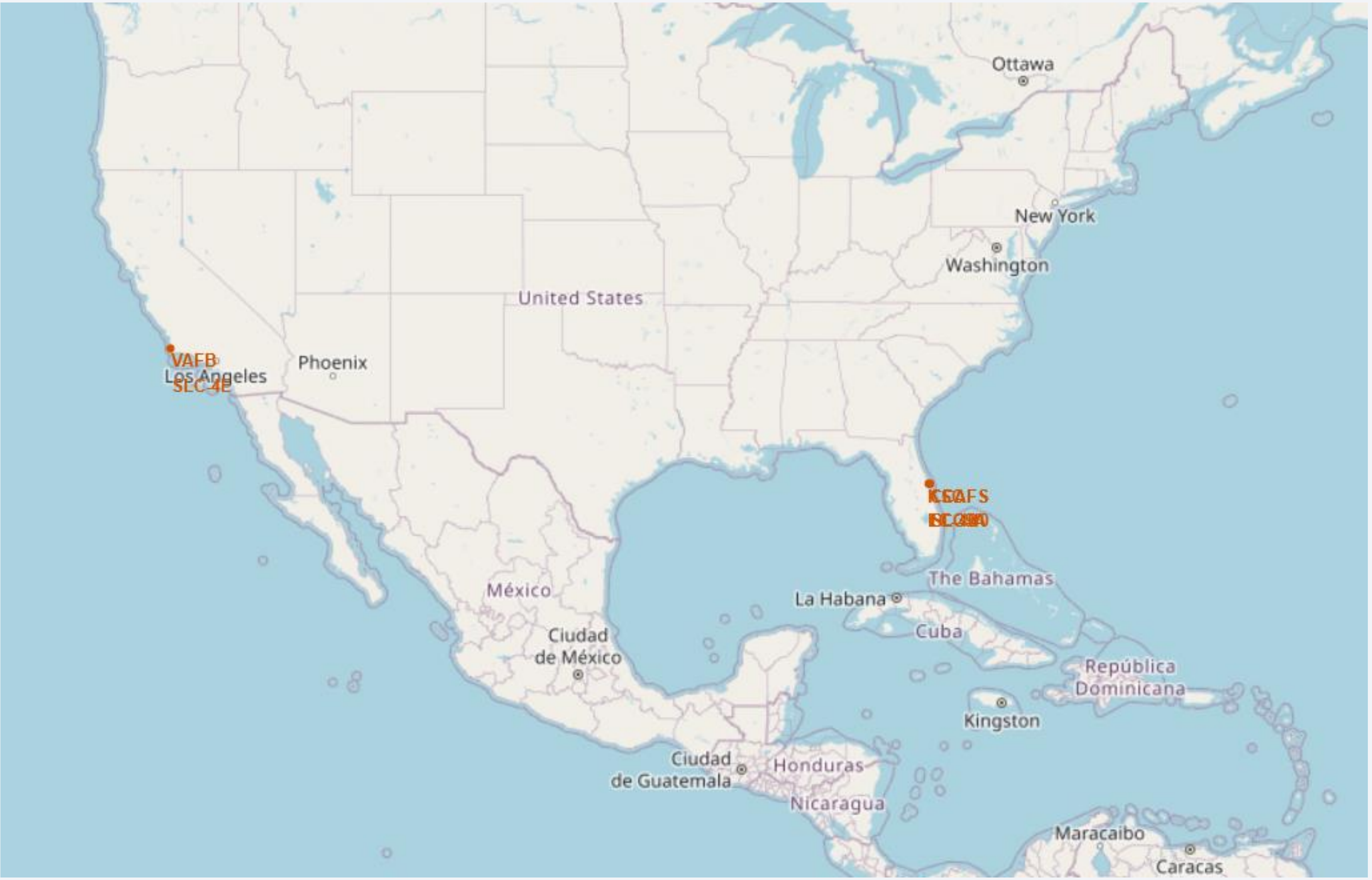
- Selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 2010-06-04 to 2010-03-20.
- Applied the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.
- SELECT "Landing _Outcome", COUNT("Landing _Outcome")
  FROM SPACEXTBL
  WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017'
  GROUP BY "Landing _Outcome"
  ORDER BY COUNT("Landing _Outcome") DESC;

Landing _Outcome	COUNT("Landing _Outcome")
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

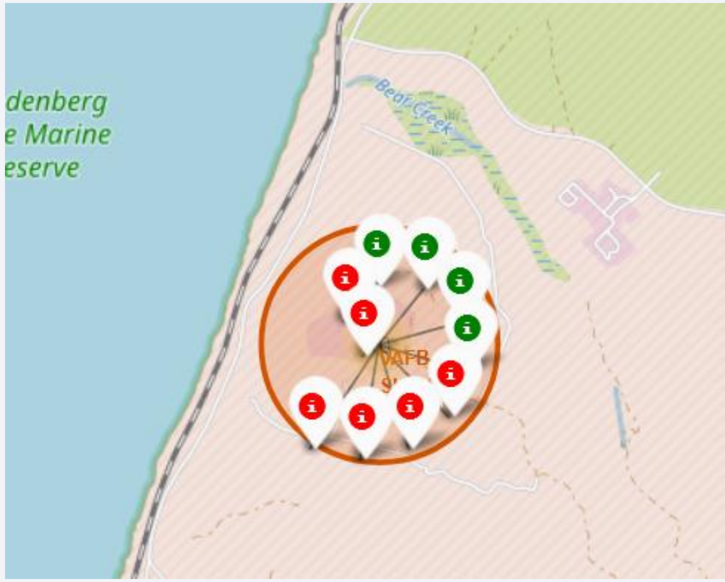
Section 3

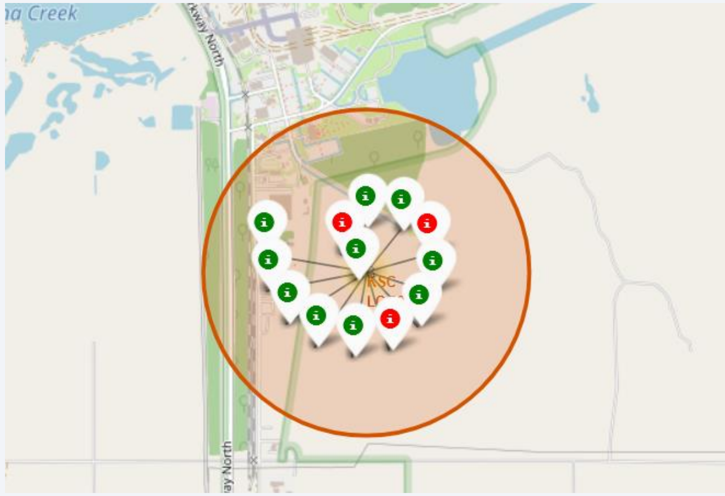
Launch Sites Proximities Analysis

Mark all launch sites on a map

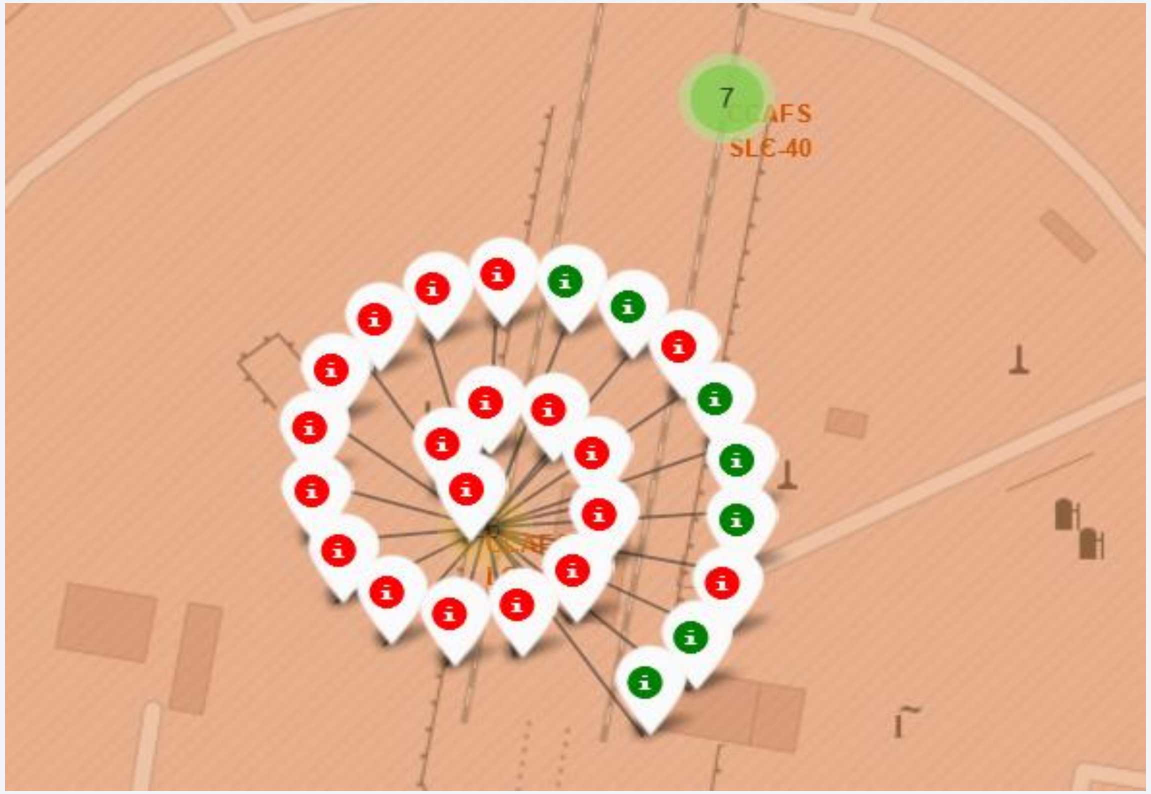


Success/failed launches for each site on the map



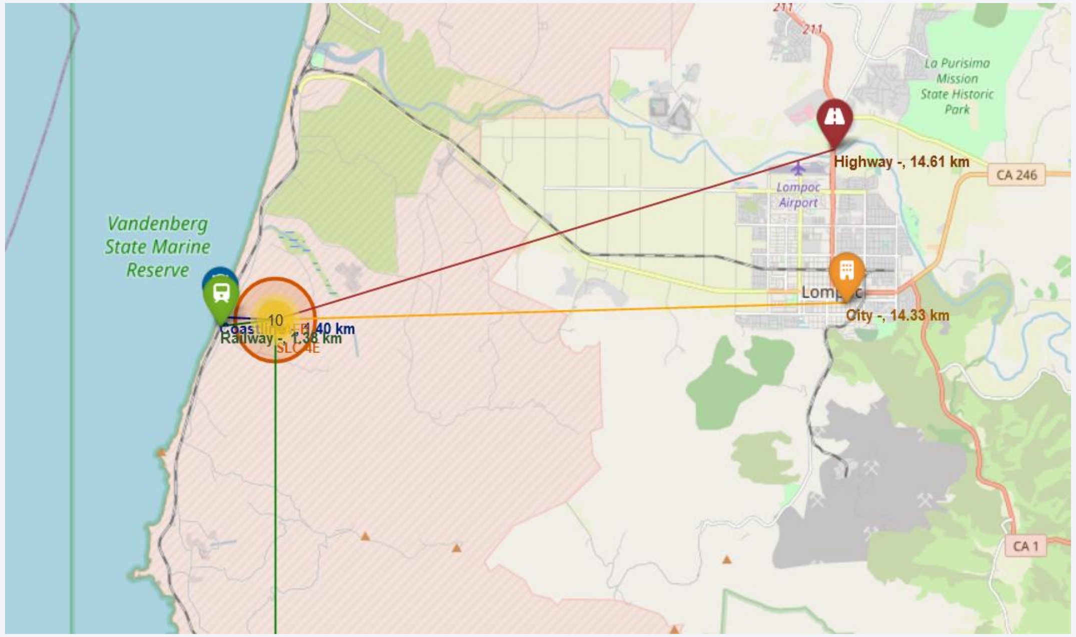


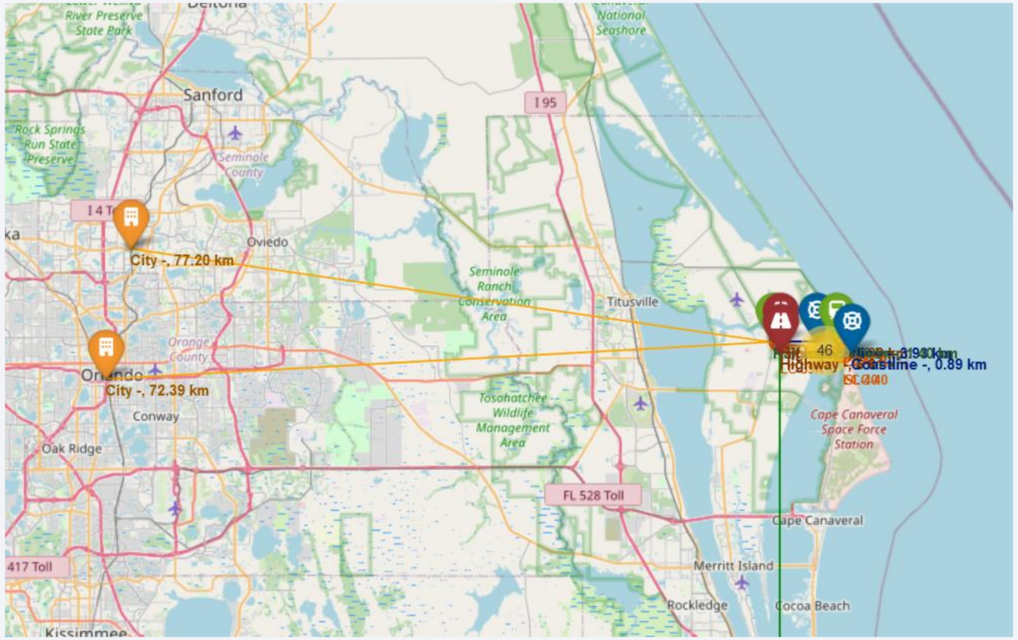
From the color-labeled markers in marker clusters, we are able to easily identify which launch sites have relatively high success rates.



Distances between a launch site to its proximities

- Are launch sites in close proximity to railways? - No
- Are launch sites in close proximity to highways? - No
- Are launch sites in close proximity to coastline? - Yes
- Do launch sites keep certain distance away from cities? - Yes

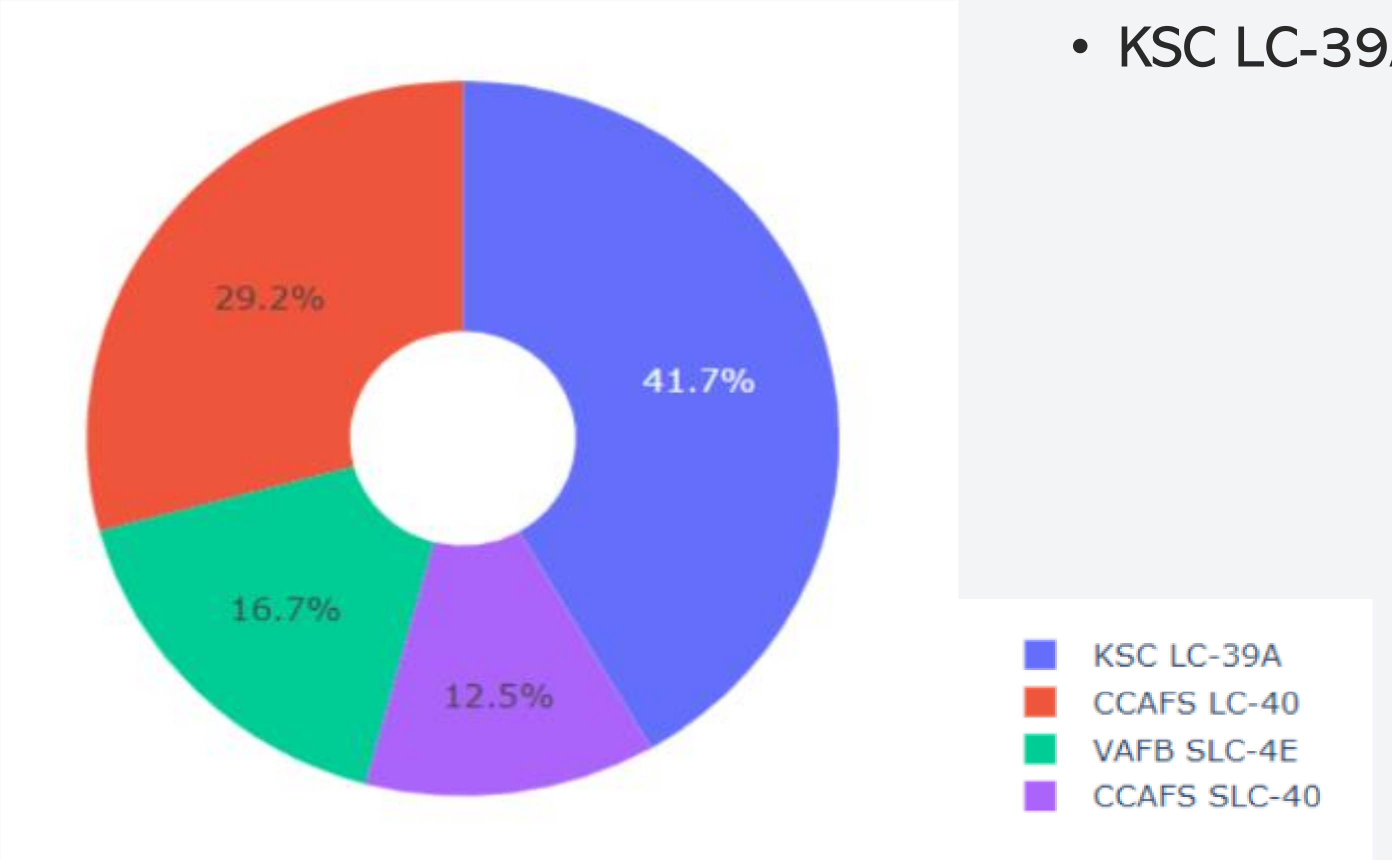




Section 4

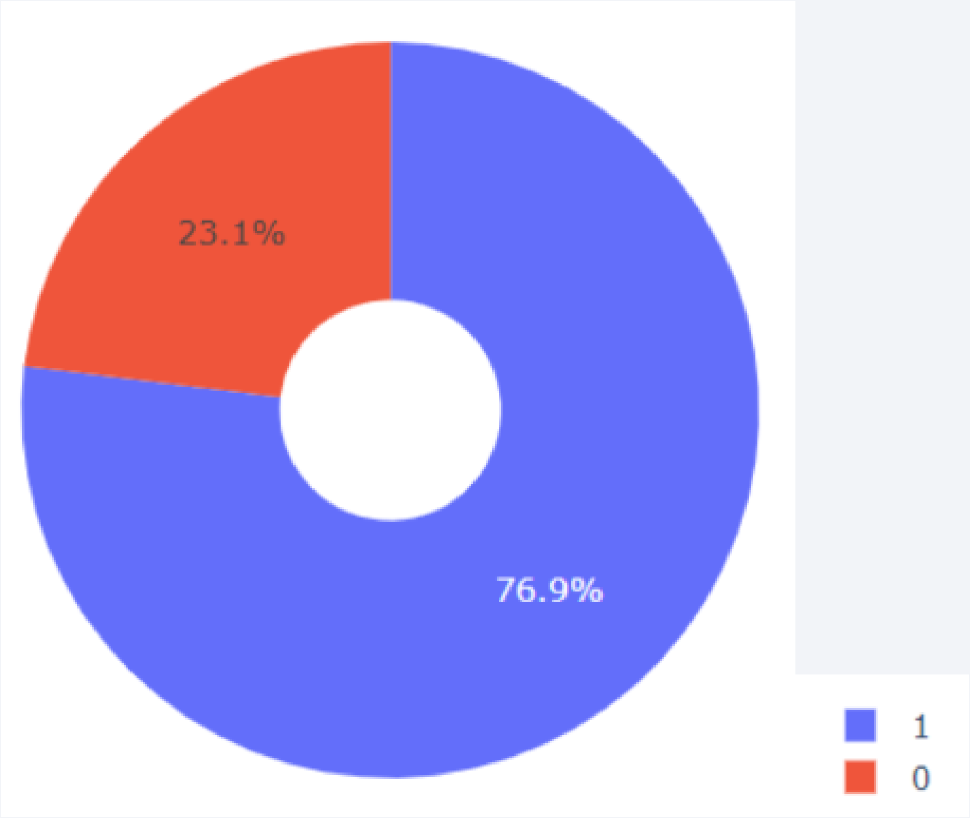
Build a Dashboard with Plotly Dash

Success percentage achieved by each launch site



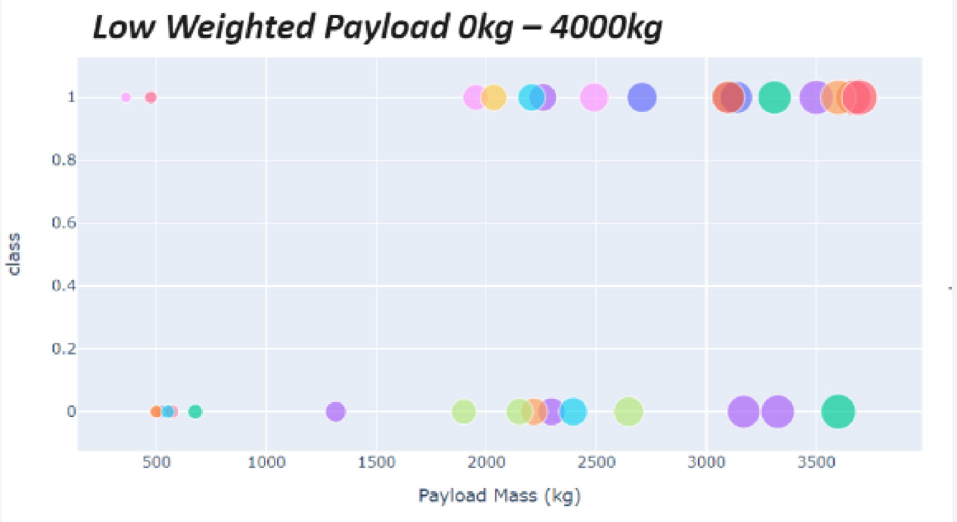
- KSC LC-39A is most successful.

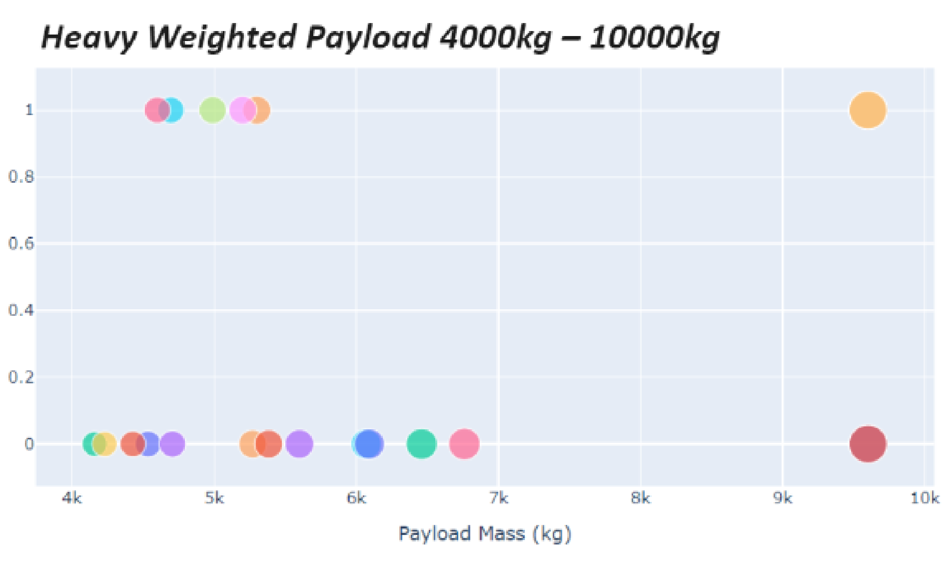
Launch site with the highest launch success ratio



For LC-39A there are 76.9% success and 23.1% failure.

Payload vs Launch Outcome for all sites





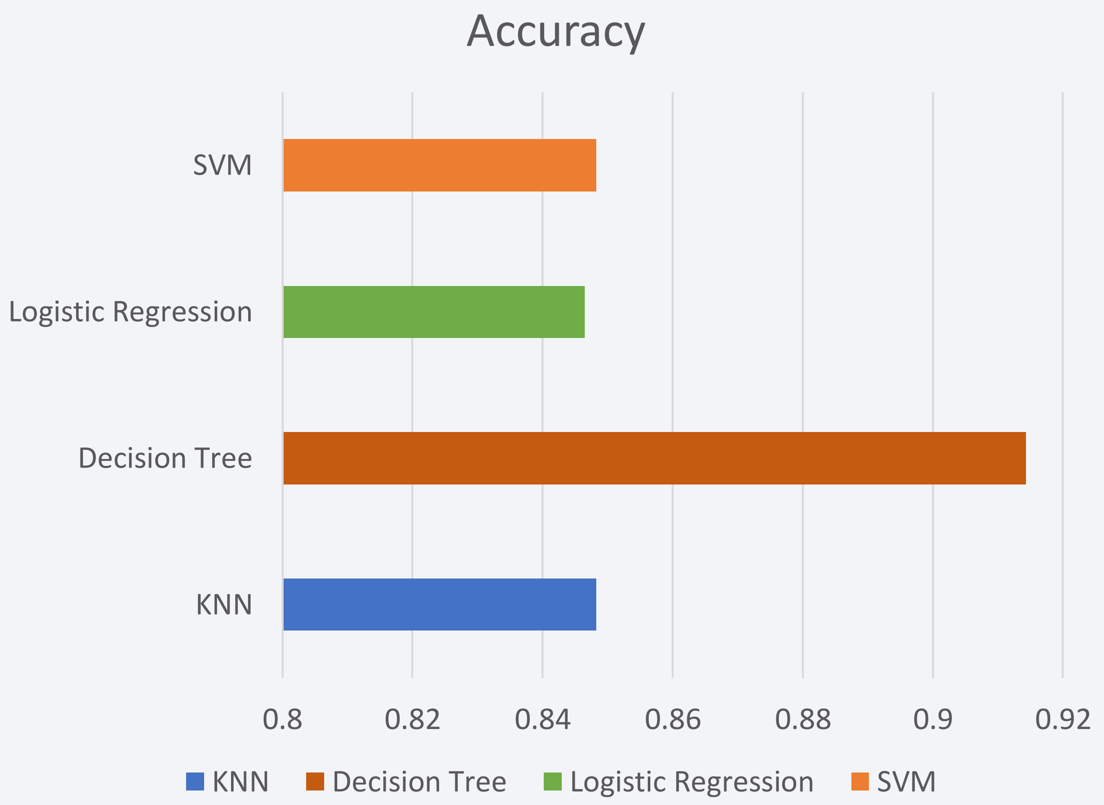
Success rates for low weighted payloads is higher than for heavy weighted payloads.

Section 5

Predictive Analysis (Classification)

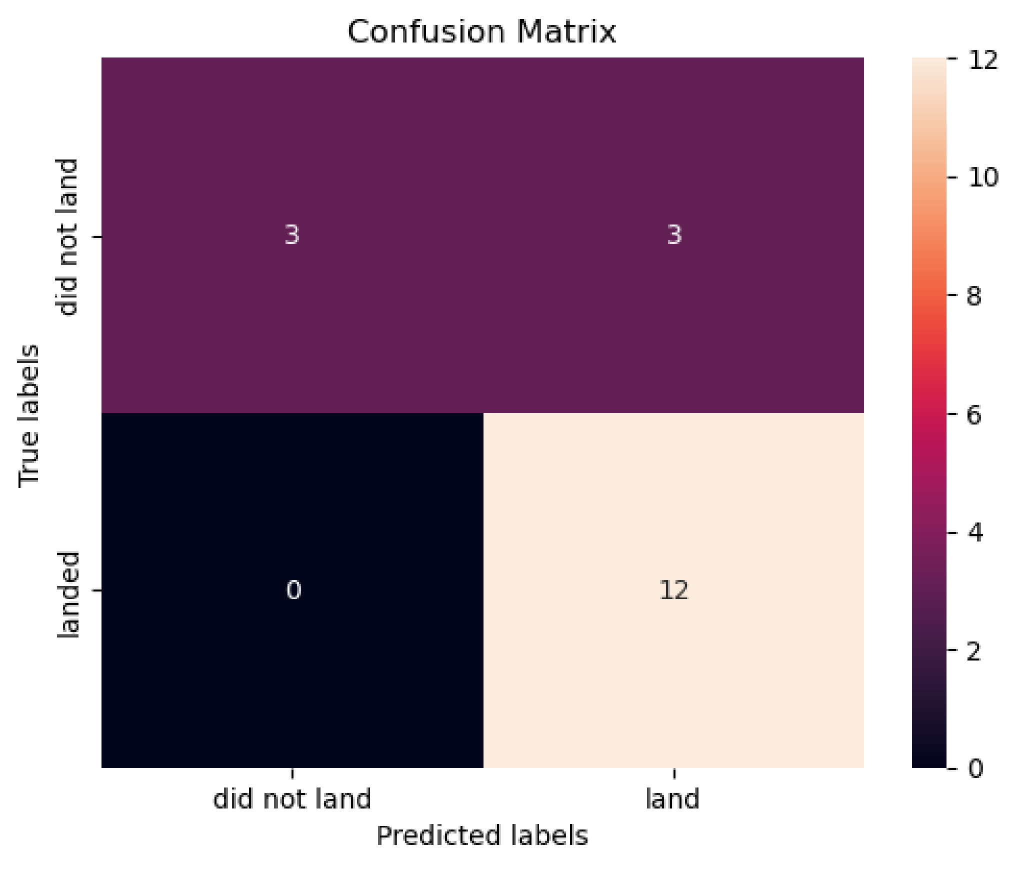
Classification Accuracy

The best performing method is Decision Tree and has a score of 0.91



Confusion Matrix

Classifier can distinguish between the different classes. The major problem is the false positives - unsuccessful landing marked as successful landing by the classifier.



Conclusions

- We observed that there is a positive correlation between the number of flights launched from a site and the success rate at that site.
- We found that the launch success rate has been steadily increasing since 2013 and continued to do so until 2020.
- Among all the orbital types, ES-L1, GEO, HEO, SSO, and VLEO had the highest success rates.
- Moreover, we identified that KSC LC-39A had the most successful launches out of all the launch sites.
- Launch sites keep certain distance away from cities, railways and highways and are in close proximity to coastline

Thank you!