IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sanket Kamble
25-Feb-24

Outline

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

Executive Summary

- **Summary of methodologies**

  The following methodologies were used to analyze data:

  - Data collection using Web scraping and SpaceX API
  - Exploratory Data Analysis including data wrangling, data visualization, SQL
  - Building an interactive map with Folium
  - Machine learning

- **Summary of all results**
  - Collected data from the public sources
  - EDA helped to predict success of launchings
  - Machine learning model suggested which features should be consider for a launching mission

Introduction

- **Project background and context**
  - SpaceX is private company of space transportation and aerospace manufacturer
  - It is established with the goal of reducing space transportation costs and ultimately developing a sustainable colony on the Mars.
  - The company advertises Falcon 9 rocket launches which reuse the first stage.
  - We will determine the success of launchings and cost of launch.

- **Problems you want to find answers**
  - Which is the best place to make successful launches
  - use of machine learning models to determine successful landings
  - impact of payload mass, launch site etc on the success of first stage landing

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - Using SpaceX REST API - (https://api.spacexdata.com/v4/rockets/)
  - Web scraping from Wikipedia

(https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- **Perform data wrangling**
  - Filtered the data and dealing with missing values
  - After summarizing and analyzing features data collected was enriched by creating a landing outcome label based on outcome data
- **Perform exploratory data analysis (EDA) using visualization and SQL**

Methodology continued

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Collected data divided in training and test data tests and evaluated by four different classification models
  - the accuracy of model evaluated using different features and combinations

Data Collection

- Data sets were collected from SpaceX API - Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/ rockets/)
- Below data columns obtained by using SpaceX API:
  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Below Data columns obtained by using Web scraping:

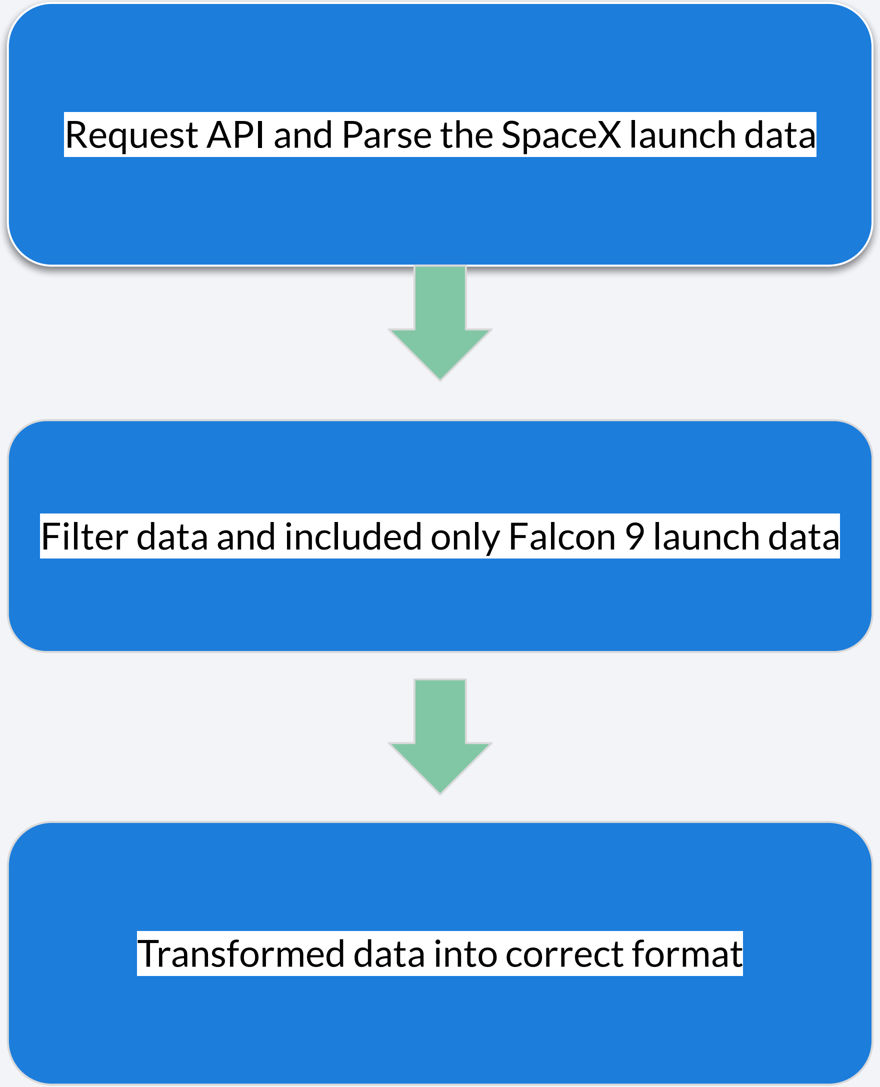
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

- Data collected from the SpaceX public API

- Please find below Github link for reference:

  https://github.com/sanketrk92/Sanket_Coursera_Applied_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



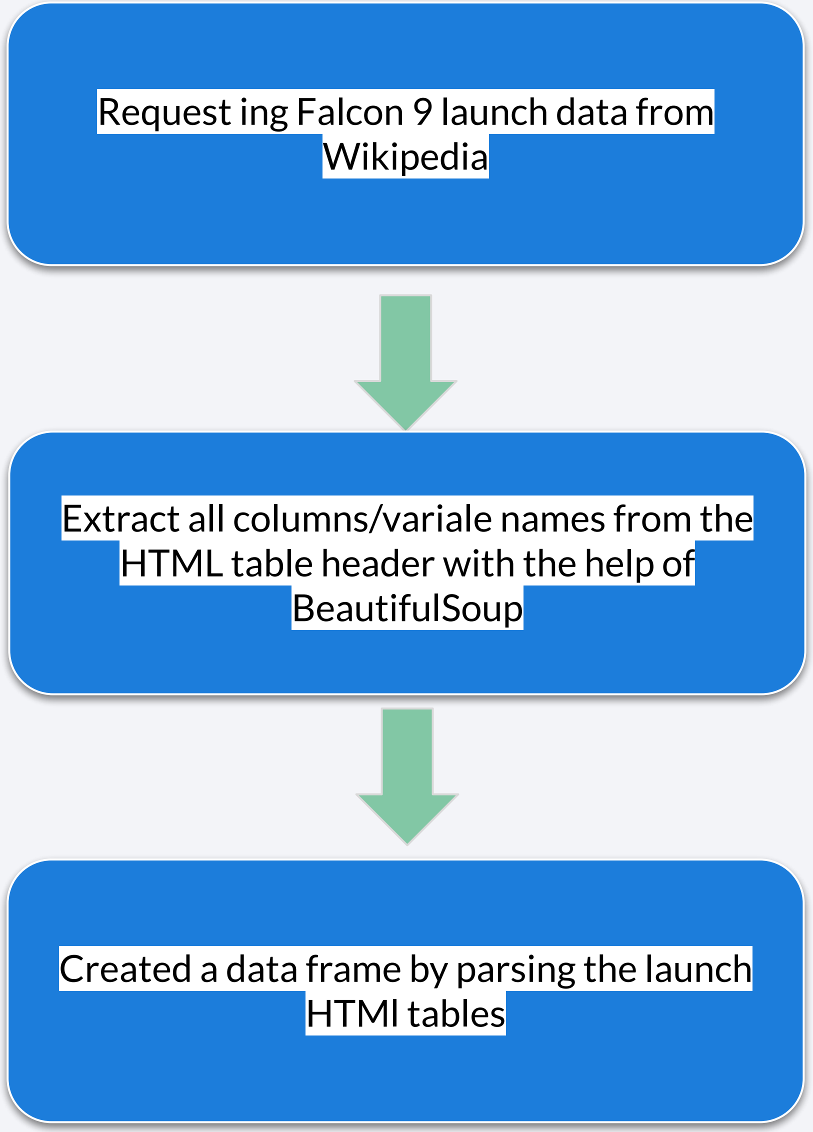
Data Collection - Scraping

- Obtained data for SpaceX launches from Wikipedia

- Downloaded data transformed to usable data and then moved to the next steps

- GitHub URL

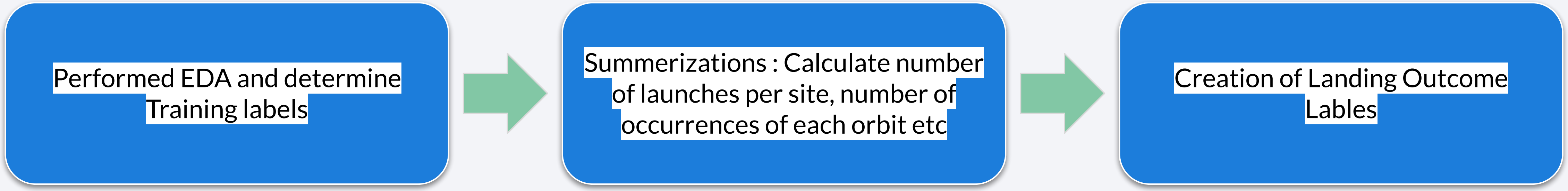
  https://github.com/sanketrk92/Sanket_Coursera_Applied_Data_Science_Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Using exploratory data analysis, multiple instances of unsuccessful booster landings were discovered in the dataset.
- Launches summarized by site, orbital occurrences, and mission outcome occurrences per orbit
- If the mission outcome is True, it indicates that the landing was successful, and if it is Falses, it indicates that the landing was unsuccessful.

- URL - https://github.com/sanketrk92/Sanket_Coursera_Applied_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:

  Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass,
  Orbit and Flight Number, Payload and Orbit

- GIthub URL

  https://github.com/sanketrk92/Sanket_Coursera_Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name 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;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

URL - https://github.com/sanketrk92/Sanket_Coursera_Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers of all Launch Sites:
- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:
- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:
- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

URL - https://github.com/sanketrk92/Sanket_Coursera_Applied_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:
- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):
- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:
- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
- Added a scatter chart to show the correlation between Payload and Launch Success

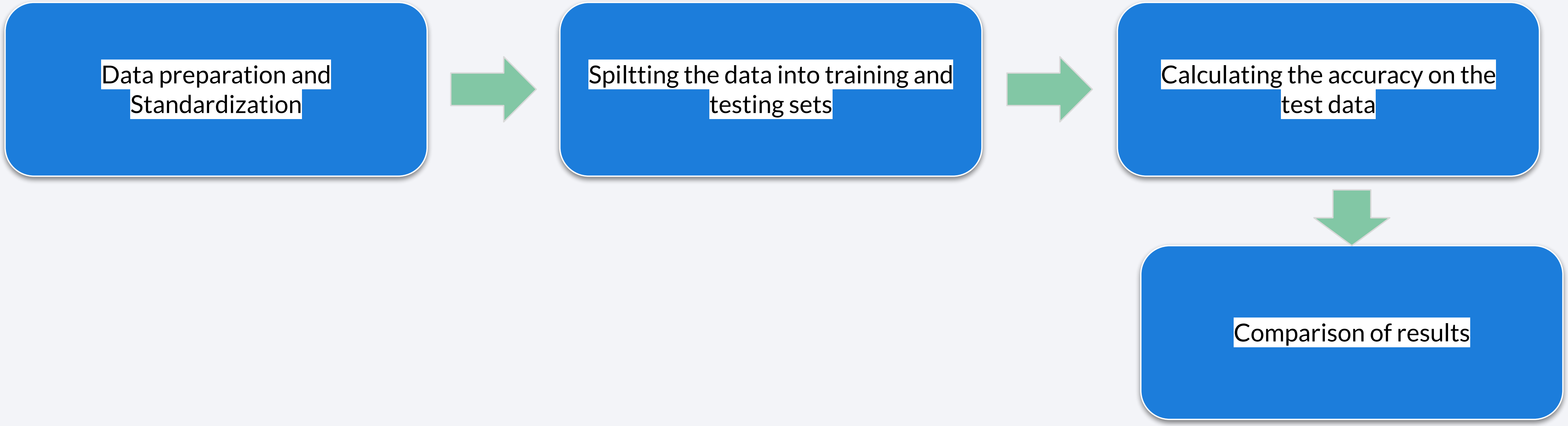
This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

URL -
https://github.com/sanketrk92/Sanket_Coursera_Applied_Data_Science_Capstone/blob/main/Coursera_spacex_dash_app.py

Predictive Analysis (Classification)

Logistic regression, support vector machine, decision tree and k nearest neighbors models have been compared.

URL - https://github.com/sanketrk92/Sanket_Coursera_Applied_Data_Science_Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



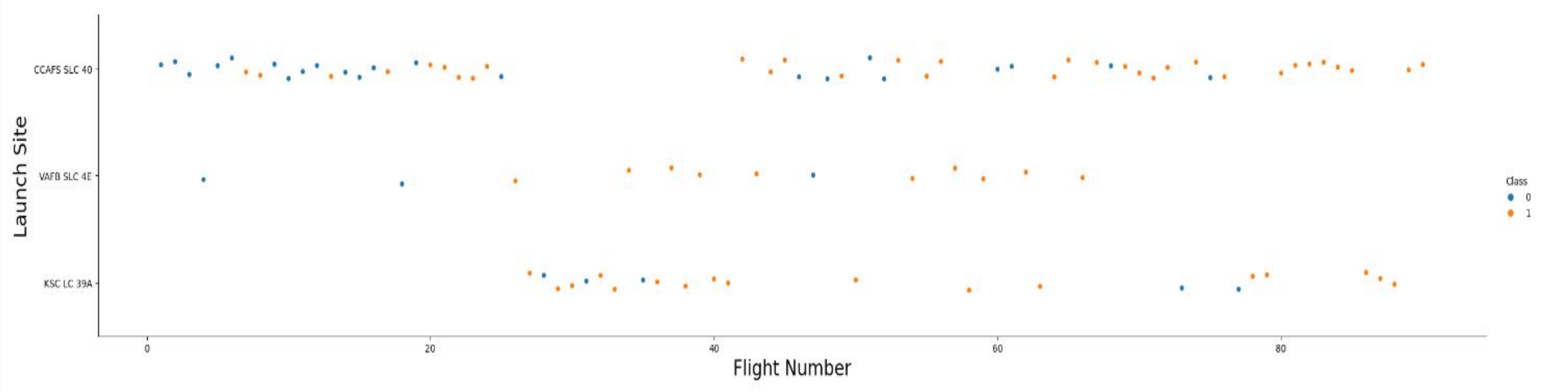
Results

- **Exploratory data analysis results**
  - SpaceX uses 4 different launch sites
  - The average payload of F9 v1.1 booster is 2,928 kg
  - Many Falcon 9 booster versions were successful at landing
  - Two booster versions failed at landing in drone ships
  - Interactive analytics demo in screenshots

Section 2

Insights drawn from EDA

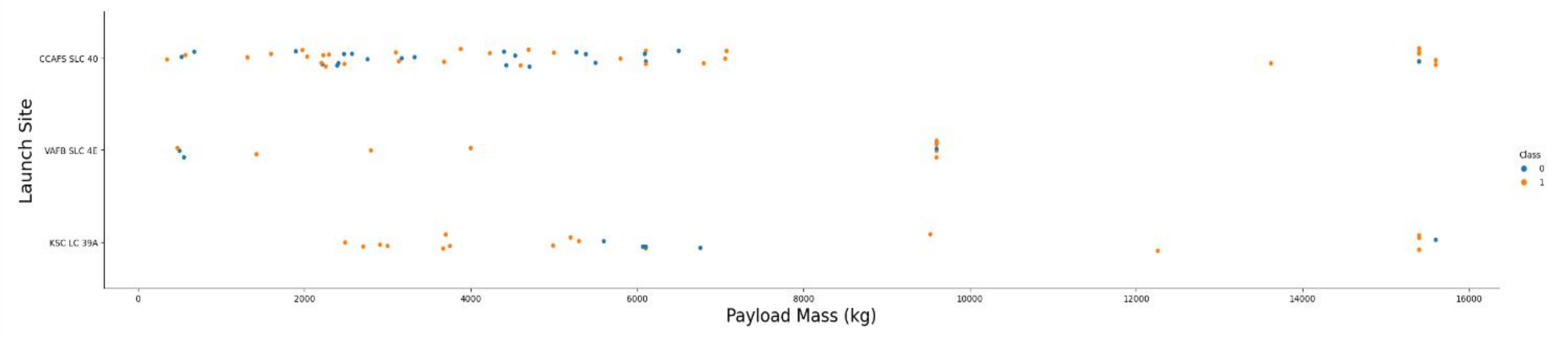
Flight Number vs. Launch Site



As per the above plot, below are the findings

- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

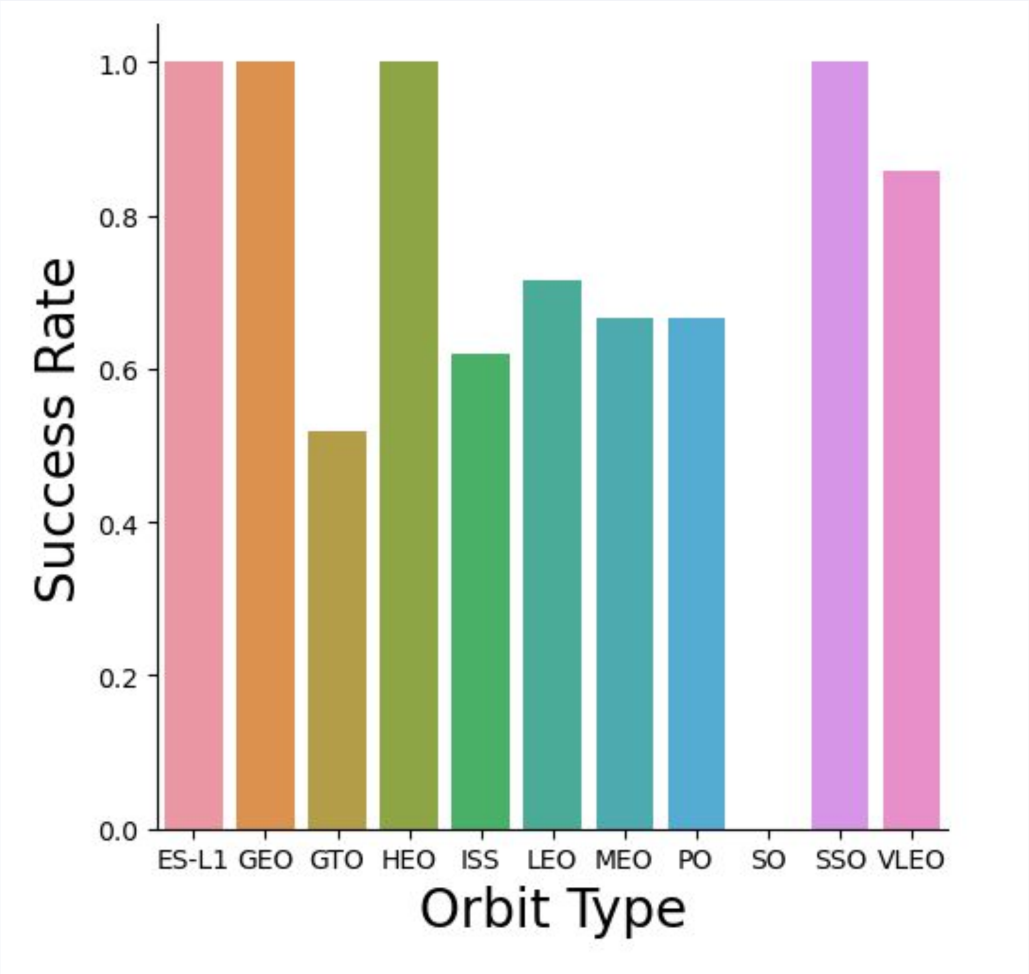
Payload vs. Launch Site



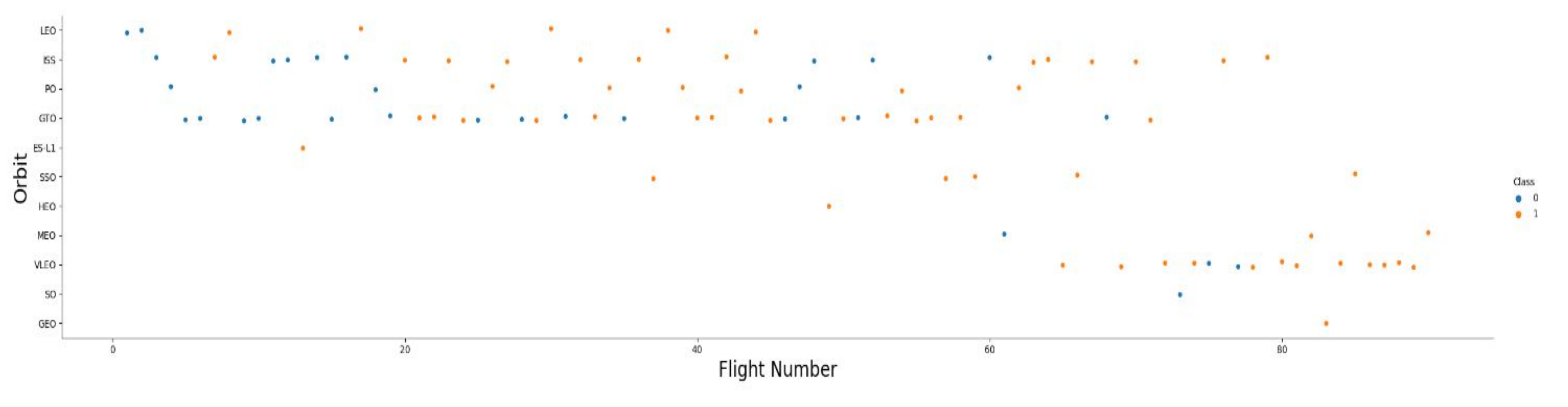
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- **The biggest success rates happens to orbits:**
  - ES L1
  - GEO
  - HEO
  - SSO
- **Orbits with a success rate between 50% and 85%:**
  - GTO, ISS, LEO, MEO, PO

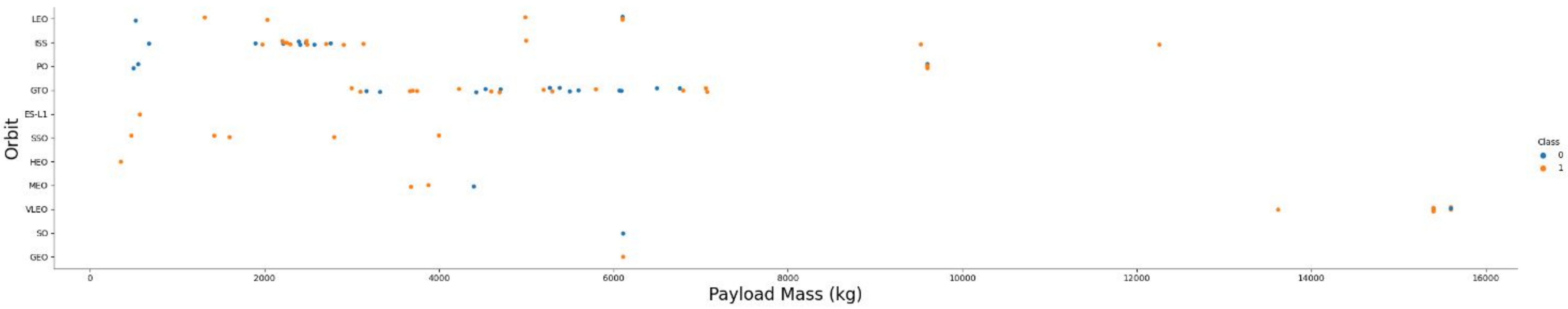


Flight Number vs. Orbit Type



While in GTO orbit, there doesn't seem to be any correlation between the number of flights and success; however, in LEO orbit, it does appear to be related to the number of flights.

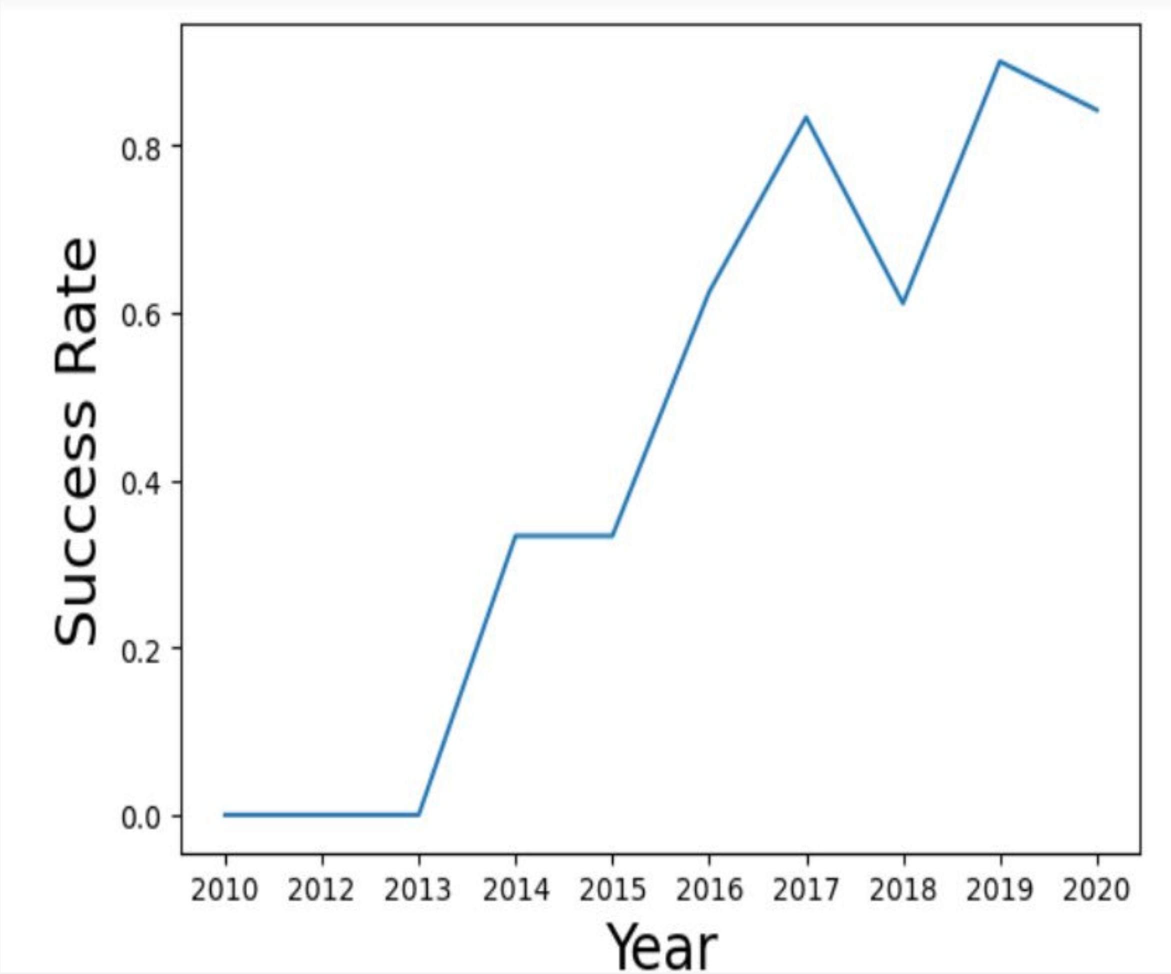
Payload vs. Orbit Type



- On GTO orbits, heavy payloads have a negative impact; on GTO and Polar LEO (ISS) orbits, they have a favorable impact.

Launch Success Yearly Trend

- The success rate appears to have increased during the first three years due to technological adjustments and advancements, which persisted until 2020.



All Launch Site Names

- There are 4 launch sites, as per the data.

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

Launch Site Names 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 attempt

- There are 5 Cape Canaveral launches

Total Payload Mass

- Showing the total mass of payload carried by NASA-launched rockets (CRS).

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Displaying the booster version F9 v1.1's average payload mass.

average_payload_mass
2534

First Successful Ground Landing Date

- Finding the minimal number for the date and applying a filter to the data based on the ground pad's successful landing outcome allows one to pinpoint the first instance, which took place on December 22, 2015.

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of the rockets' names that have been successful in drone ship applications and whose payload mass is more than 4,000 but less than 6,000.

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

Total Number of Successful and Failure Mission Outcomes

- List of all mission outcomes, both successful and unsuccessful.

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass.

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

- List of the drone ship's unsuccessful landings, along with the names of the launch sites and booster versions, for each month in 2015.

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

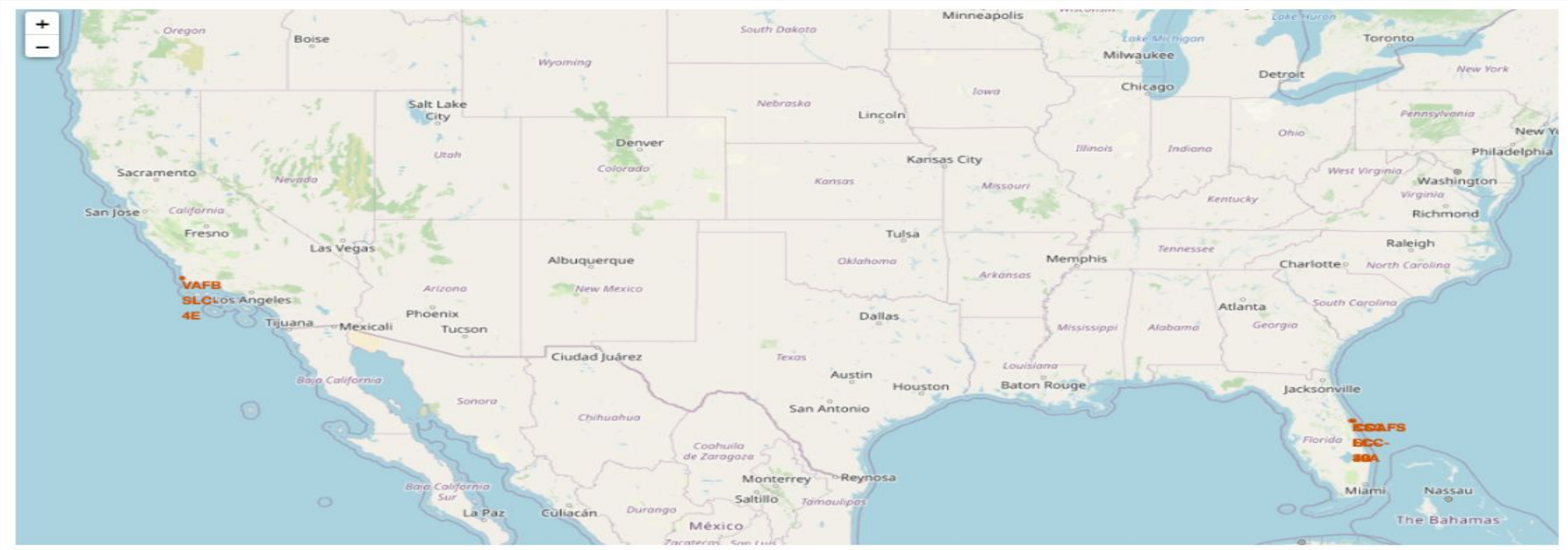
- Sorting the number of landing results (e.g., ground pad success or drone ship failure) between 2010-06-04 and 2017-03-20 in descending order.

landing__outcome	count_outcomes
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

Section 3

Launch Sites Proximities Analysis

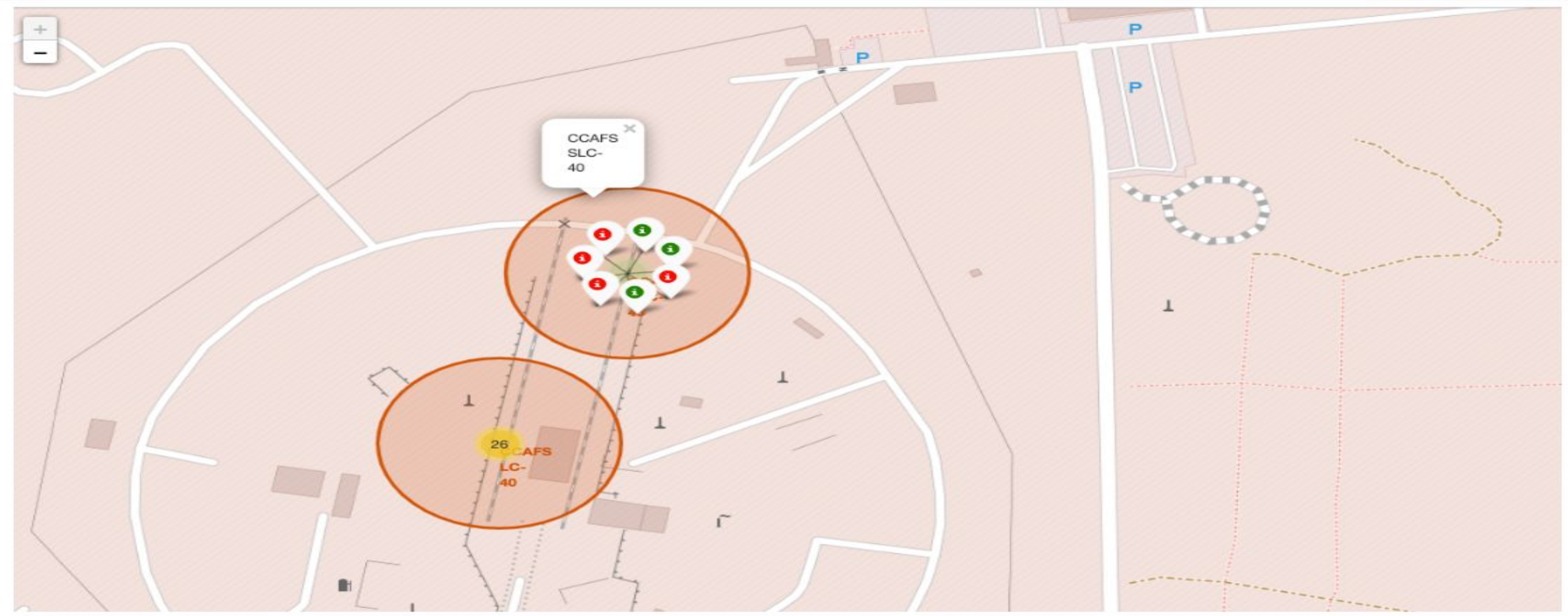
All launch sites



Since all launch sites are located fairly close to the coast, there is little chance that any debris will fall or explode in the vicinity of people when rockets are launched toward the ocean.

Colour-labeled launch records on the map

Which launch sites with reasonably high success rates should be easy to determine from the color-labeled markers in marker clusters.

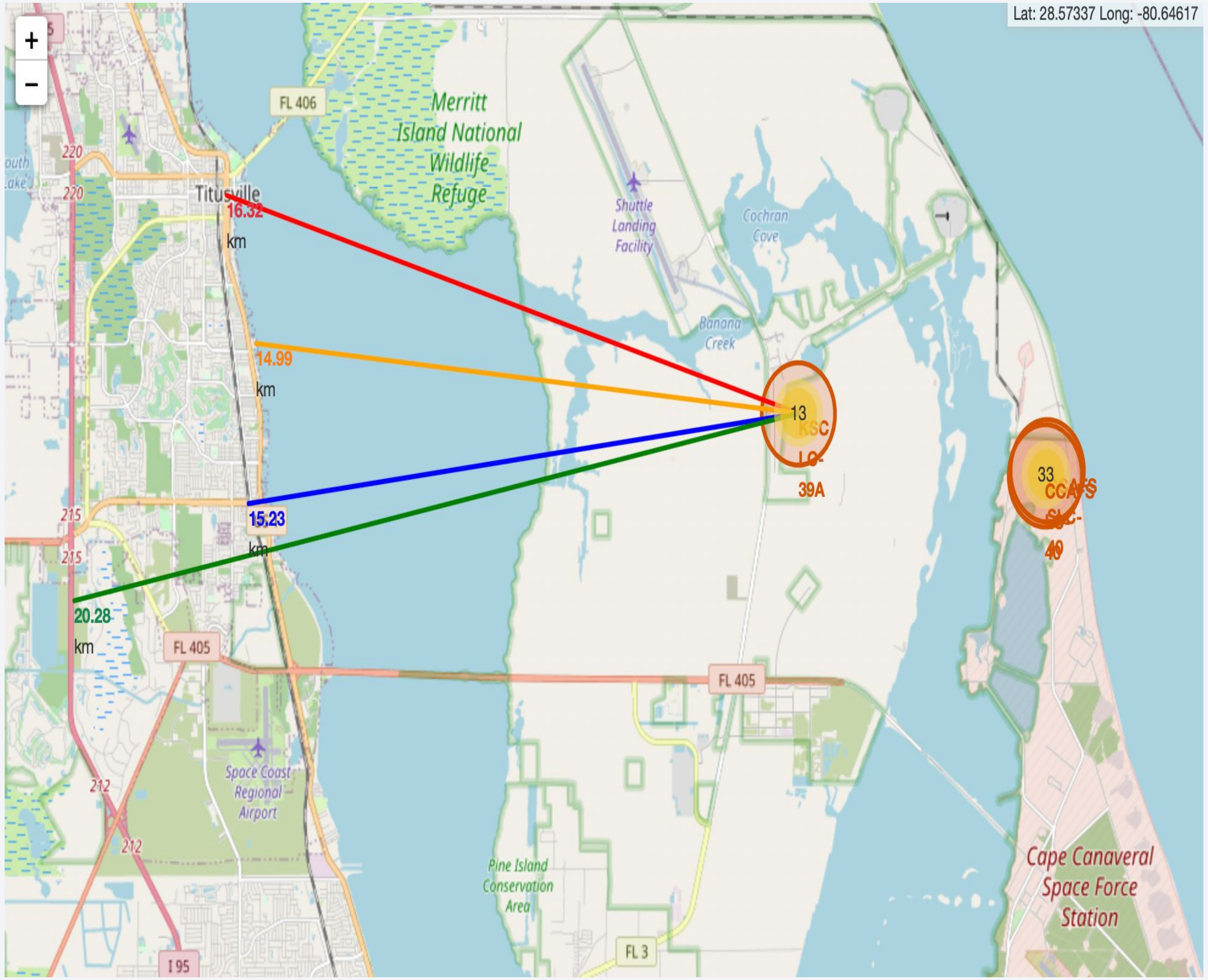


Distance from the launch site KSC LC-39A to its proximities

The launch site KSC LC-39A is plainly visible from the visual analysis, showing that it is: - relatively close to railway (15.23 km)
- in proximity to a highway (20.28 km)
- quite near the coast (14.99 kilometers)

• Additionally, Titusville, the closest city to the launch site KSC LC-39A, is only 16.32 kilometers away.

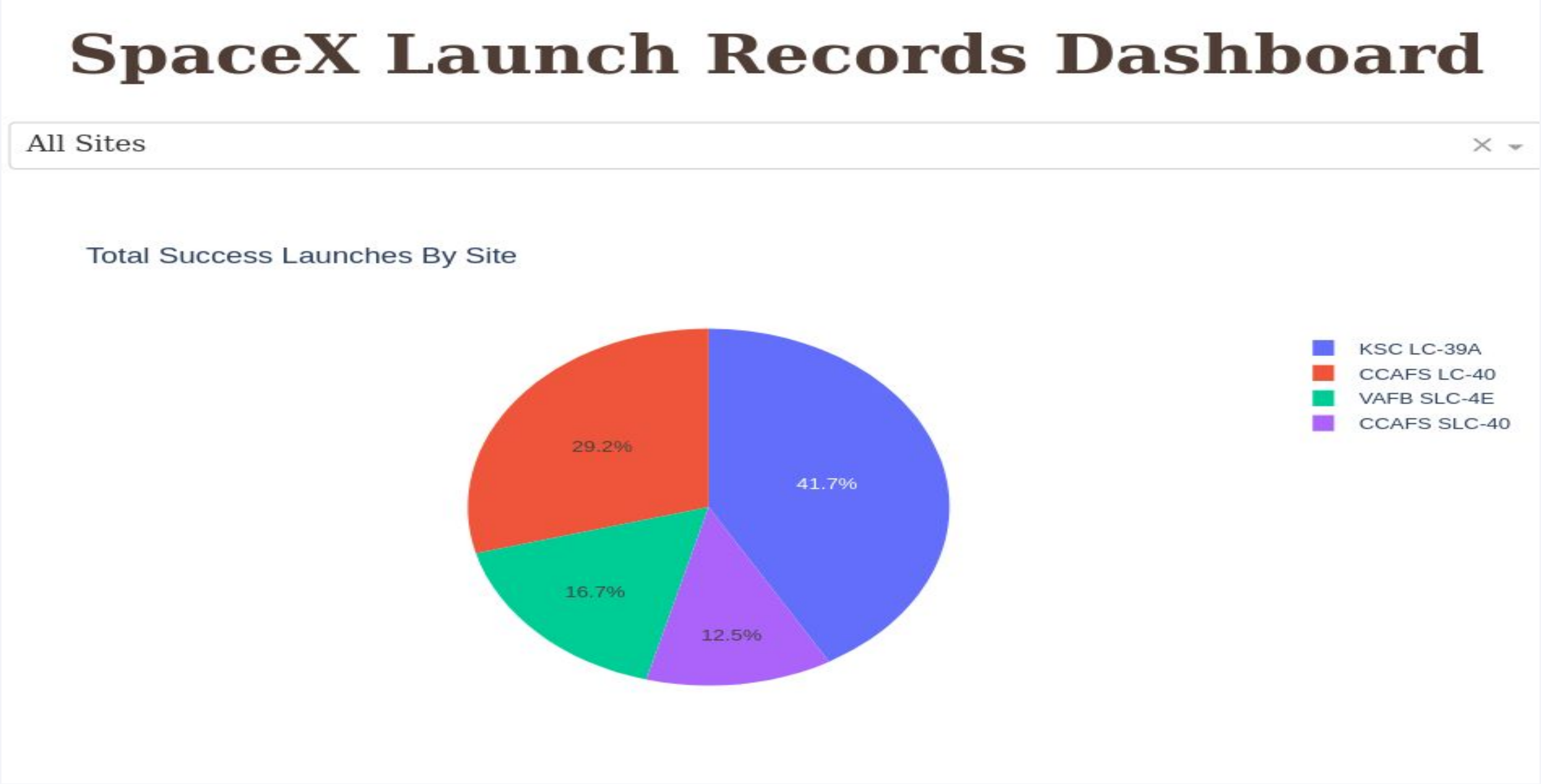
• A failed rocket can go up to 15-20 kilometers in a matter of seconds due to its high speed. It might pose a threat to densely inhabited areas.



Section 4

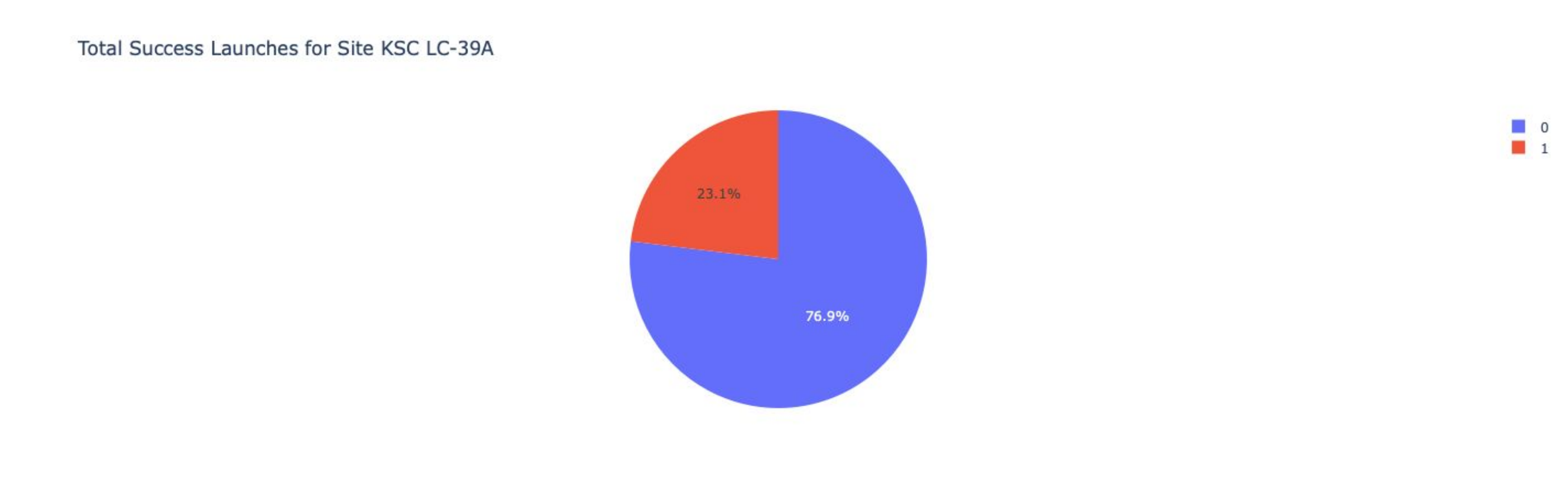
Build a Dashboard with Plotly Dash

Launch success count for all sites



The location of launches appears to be a critical component in the mission's success.

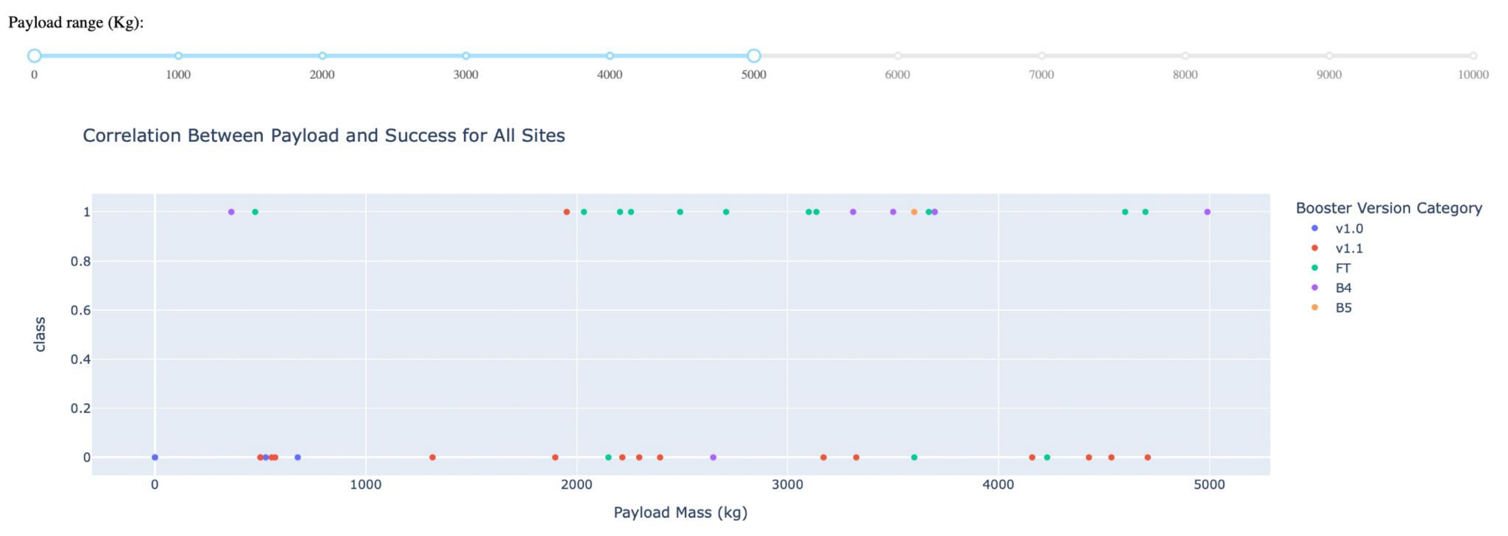
Launch site with highest launch success ratio



With 76.9% launch success, KSC LC-39A has the highest landing success rate, with 10 successful landings and only 3 failures.

Payload Mass vs. Launch Outcome for all sites

Payloads weighing between 2000 and 5500 kg had the best success rate, according to the data.

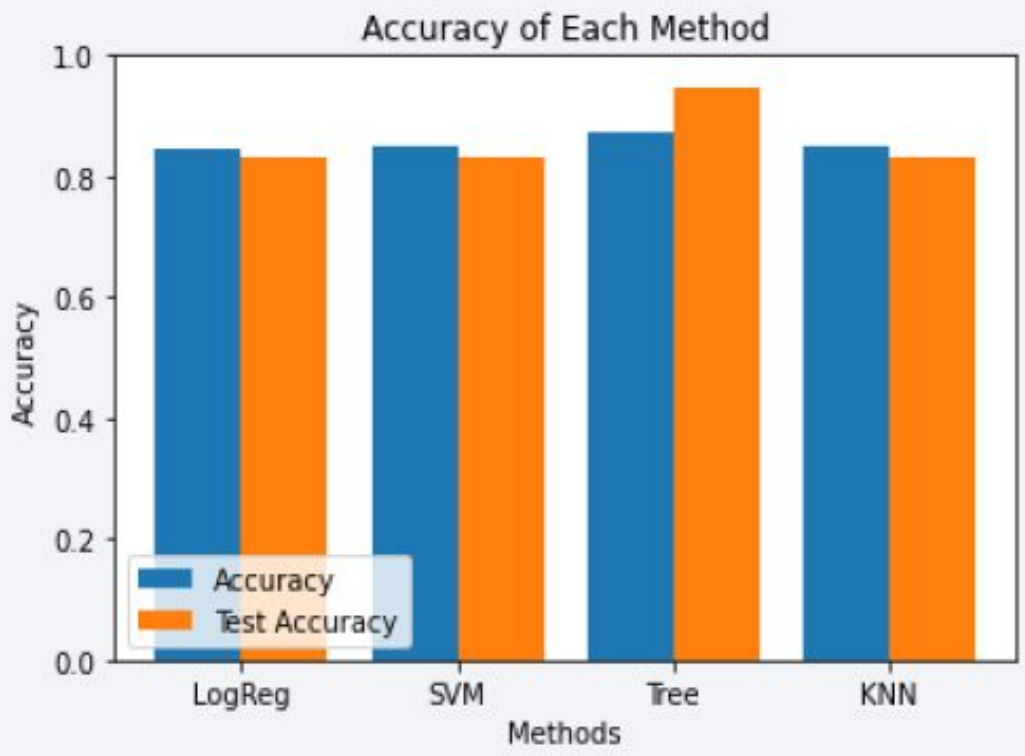


Section 5

Predictive Analysis (Classification)

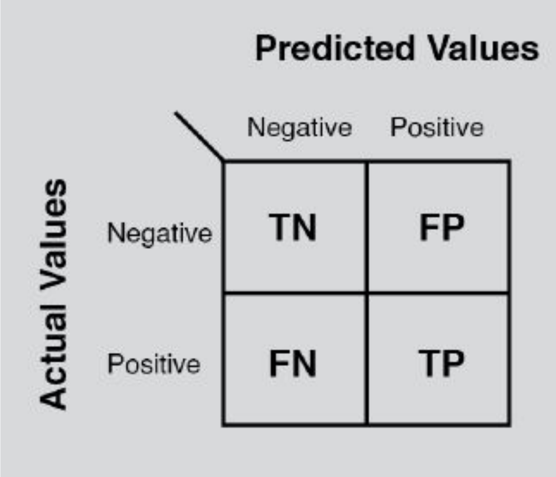
Classification Accuracy

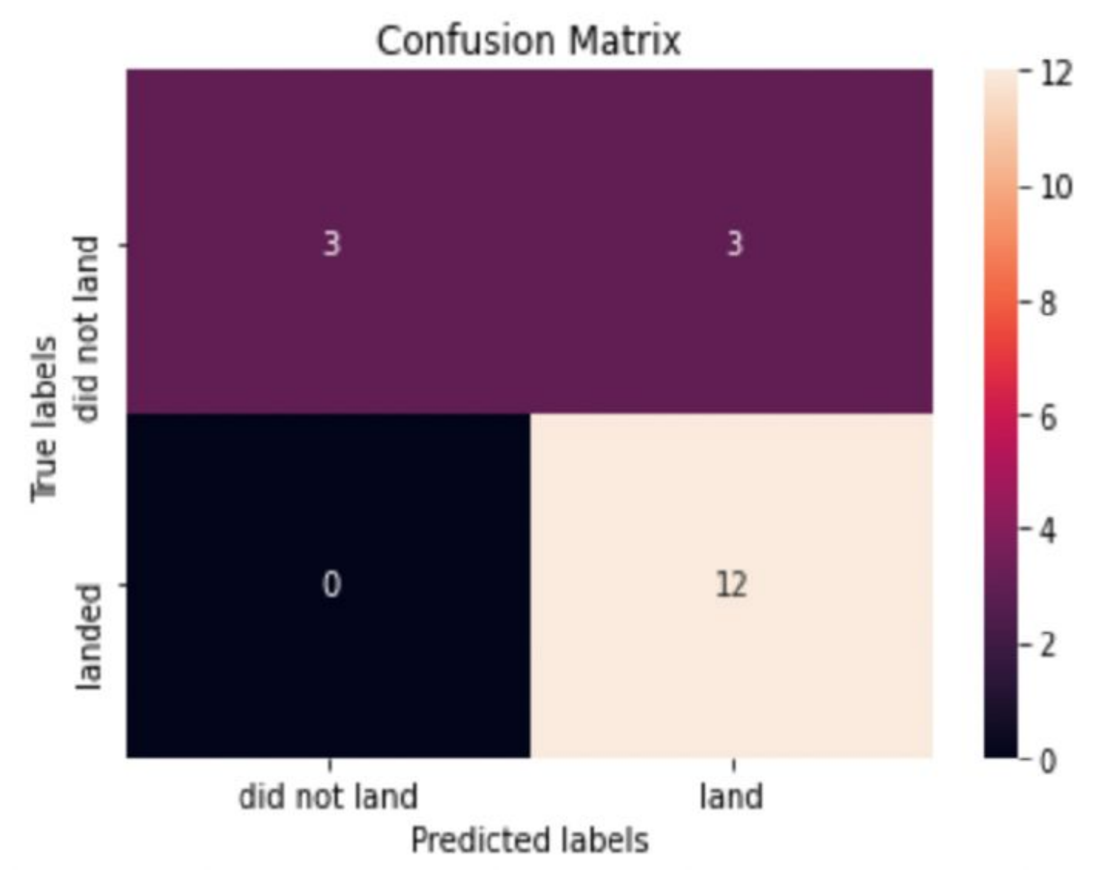
Plotting the accuracies of four evaluated classification models side by side reveals which model has the highest classification accuracy: the Decision Tree Classifier, with accuracies greater than 87%.



Confusion Matrix

It is evident by looking at the confusion matrix that logistic regression is capable of differentiating between the various classes. It is evident that false positives are the main issue.





Conclusions

1. For this dataset, the optimal algorithm is the Decision Tree Model.
2. Results from launches with a smaller payload mass are superior to those from launches with a greater payload mass.
3. All of the launch sites are extremely near to the coast, and the majority are close to the Equator line.
4. Over time, the success rate of launches rises.
5. Out of all the locations, KSC LC-39A has the best success percentage for launches.
6. The success rate for orbits ES-L1, GEO, HEO, and SSO is 100%.

Appendix

- Coursera
- IBM
- Instructors
- Wikipedia


Thank you!