IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Kwan
August 14, 2025

Outline

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

Executive Summary

- Methodologies
  - Coding were performed using Python with appropriate libraries.
  - Publicly available data (from SpaceX API and Wikipedia) were compiled into a dataset.
  - The dataset were analyzed to determine which feature impacted the landing outcome.
  - Classification models were developed to predict landing outcomes and compared with each other.
- Summary of all results
  - Individual mission parameters such as orbit, payload mass, and launch site do not show substantial correlation with the landing outcome.
  - From 2013 to 2020, the success rate of SpaceX has increased substantially (to over 80%).
  - Launch site does not correlate substantially with success rate and payload mass.
  - A hyperparameter-tuned decision tree model can predict the landing outcome of a Falcon 9 rocket with 88.9% accuracy.

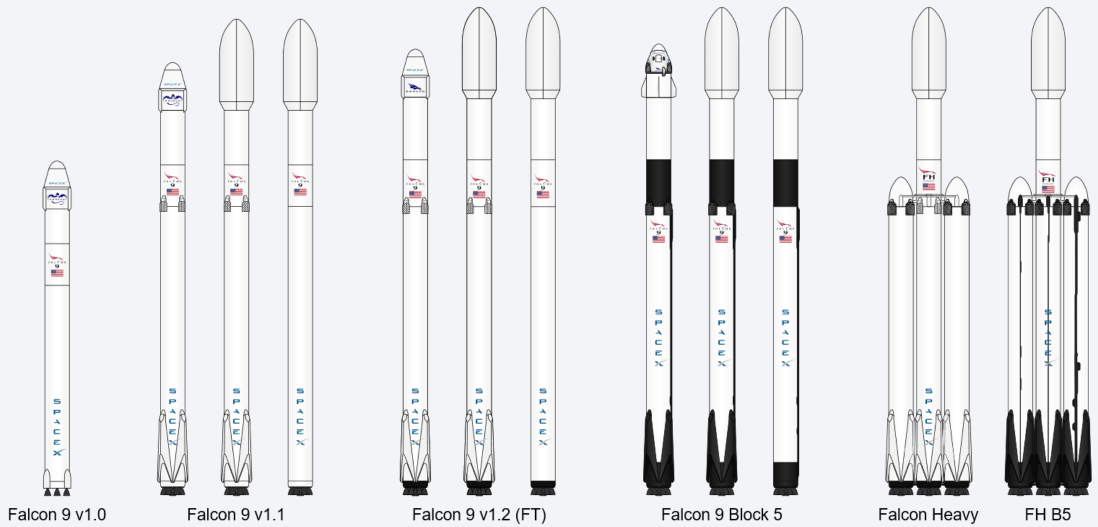
Introduction – SpaceY vs SpaceX

SpaceY: The Future of Space Flight Missions

- We at SapceY aim to be the leader of the future space flight missions.
- We will be competing with other space flight companies, and our major competitor is SpaceX.
- SpaceX's competitive advantage comes from their low operating cost for individual space mission (~ 62M USD) againsts others (~ 165M USD).
- Their low operating cost is largely due to their ability to re-use the Falcon 9 first stage rocket booster after it lands successfully.
- Besides rocket design, which is substantially different between ours and other companies, some external factors may affect the landing outcome.

Introduction – SpaceX's Mission Parameters

- Some of the SpaceX's mission parameters include:
  - Flight number, Rocket & Booster ID, Date
  - Crew, Customer & Payload Mass (kg)
  - Orbit
  - Launch Site (with latitude and Longitute)
  - Mission outcome
  - Links to othere resources
- Which of these factors is important?



Source: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Introduction – Objectives and Goal

Objectives:

1. To compile a dataset from different publicly accessible sources (e.g., APIs, Wikipedia).
2. To explore what features in the compiled dataset may affect landing outcome of a Falcon 9 rocket.
3. To determine which launch site(s) SpaceY should be operating on based on SpaceX's operation.
4. To build and evaluate the performance of different classification models in predicting the landing outcome of a Falcon 9 rocket.

Goal

- To develop a classification model to predict future landing outcome of a Falcon 9 rocket based on the selected features.

Section 1

Methodology

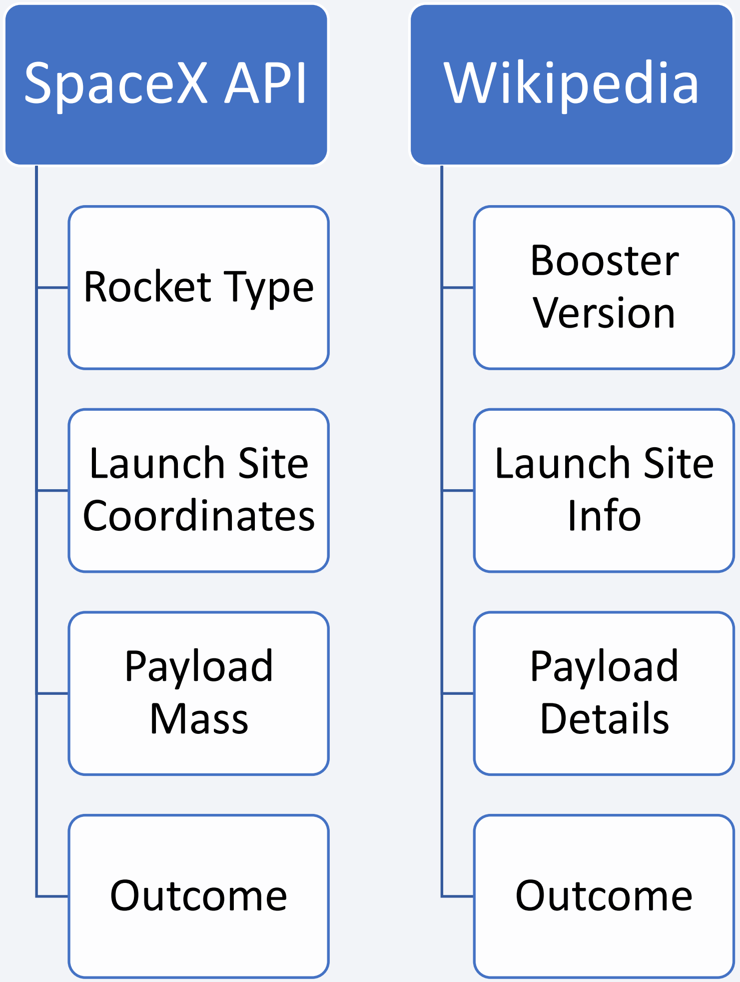
Methodology

Overview

- Data collection methodology:
  - Getting historical records from the SpaceX API
  - Web-scraping of the List of Falcon 9 and Falcon Heavy launches Wikipedia page
- Perform data wrangling
  - Replacing Missing data, Standardizing values, etc.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, and evaluate 4 classification models

Data Collection

- Python codes were used to collect historical data from:
  1. SpaceX API - https://api.spacexdata.com/v4/launches/past
  2. Wikipedia page: List of Falcon 9 and Falcon Heavy launches
- The 2 datasets were used as complements of each other.



Data Collection – SpaceX API

Call

- Call the SpaceX API (url) with a get request and store the response.

Format

- Read the response JSON file into a dataframe.
- Convert codified data into human-readable formats.

Clean-up

- Filter for only Falcon9 data
- Reset flight numbers

For details, please check out the following file:
https://github.com/danielkwan783/IBM_DS_Capstone/blob/main/Notebooks/1_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Request

- Use a get request to get the html content of the Wikipedia page List of Falcon 9 and Falcon Heavy launches

Extract

- Extract the table contents using BeautifulSoup.

Format

- Parse the extracted data into a dataframe.

For details, please check out the following file:
https://github.com/danielkwan783/IBM_DS_Capstone/blob/main/Notebooks/2_jupyter-labs-webscraping.ipynb

Data Wrangling

Missing values

- Replaced missing values with the mean (PayloadMass)
- Leaving as-is (LandingPad)

Quick summaries

- Calculate number of launches for each site
- Calculate the occurrences of each orbit
- Calculate the success rates

Defining target

- Grouped landing outcomes into success (class = 1) or failure (class = 0)

For details, please check out the following file:
https://github.com/danielkwan783/IBM_DS_Capstone/blob/main/Notebooks/3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To answer if there is any correlation between flight number, payload mass, orbit type and outcome, the following plots were created:
  1. Scatterplot of Flight Number vs Payload Mass coloured by outcome
  2. Scatterplot of Flight Number vs Launch Site coloured by outcome
  3. Scatterplot of Payload Mass vs Launch Site coloured by outcome
  4. Bar graph of the success rate of each orbit type
  5. Scatterplot of Flight Number vs Orbit coloured by outcome
  6. Scatterplot of Payload Mass vs Orbit coloured by outcome
- A line plot was also created to see the trend of changes in the success rate between 2010 and 2020

For details, please check out the following file:
https://github.com/danielkwan783/IBM_DS_Capstone/blob/main/Notebooks/5_edadataviz.ipynb

EDA with SQL

- Ten (10) SQL queries were written to gain further insights into the data:
    1. Identity of the unique launch sites
    2. Records for launch site(s) starting with “CCA”
    3. Total payload mass carried by the boosters launched by NASA (CRS)
    4. Average payload mass by booster version F9 v1.1
    5. The first successful landing outcome in ground pad date
    6. The boosters landing successfully on drone ship with payload between 4000 and 6000 kg
    7. Total number of successful and failure mission outcomes
    8. The booster versions which have carried the maximum payload mass
    9. The month for the booster failed to land on a drone ship
    10. Rank of the landing outcomes by descending order

For details, please check out the following file:
https://github.com/danielkwan783/IBM_DS_Capstone/blob/main/Notebooks/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To obtain a visual representation of the launch outcomes for the different launch sites, the following objects were added to a Folium map showing the launch sites:
  - Coloured circles for each launch site
  - Markers for each launch (red = failed; green = success) at the respective launch sites
  - Lines with labeled distance for points of interest around the launch site VAFB SLC-4E

For details, please check out the following file:
https://github.com/danielkwan783/IBM_DS_Capstone/blob/main/Notebooks/6_lab_jupyter_launch_site_location.ipynb

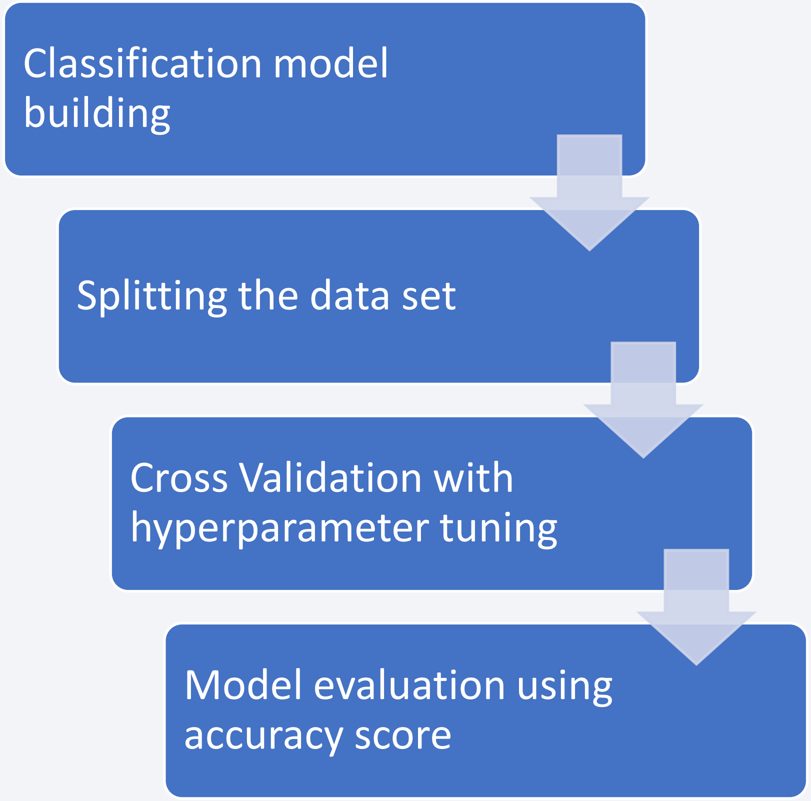
Build a Dashboard with Plotly Dash

- To better understand the outcome for each launch site, an interactive dashboard has been created with the following controls and graphs:
  - Dashboard Controls:
    1. Launch site selection menu (including options for "ALL" sites)
    2. Payload slider selection (0 to 10,000 kg)
  - Output Graphs (for "ALL" sites):
    1. A pie chart displaying the percentage of successful launches from each of the launch sites.
    2. A scatter plot of payload mass versus the launch outcome
  - Output Graphs (for each launch site):
    1. A pie chart displaying the outcome percentages (success (1) vs. failed (0)) of the selected launch site
    2. A scatter plot of the selected payload mass versus the launch outcome for the selected launch site

For details, please check out the following file:
https://github.com/danielkwan783/IBM_DS_Capstone/blob/main/Python%20Script/spacex-dash-app.py

Predictive Analysis (Classification)

- The following classification models were developed using the Sci-Kit Learn Library for Python:
  - Four (4) classification models were built: logistic Regression classifier, support vector machine (SVM) classifier, decision tree classifier, and k-Nearest Neighbors (KNN) classifier
  - Models were trained by 80% of the prepared data set, and accuracy score were calculated for evaluation.
  - A 10-fold GridSearchCV algorithm was used to search for the best hyperparameters for each model
  - The best models were selected based on the accuracy scores from the test data set



For details, please check out the following file:
https://github.com/danielkwan783/IBM_DS_Capstone/blob/main/Notebooks/7_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

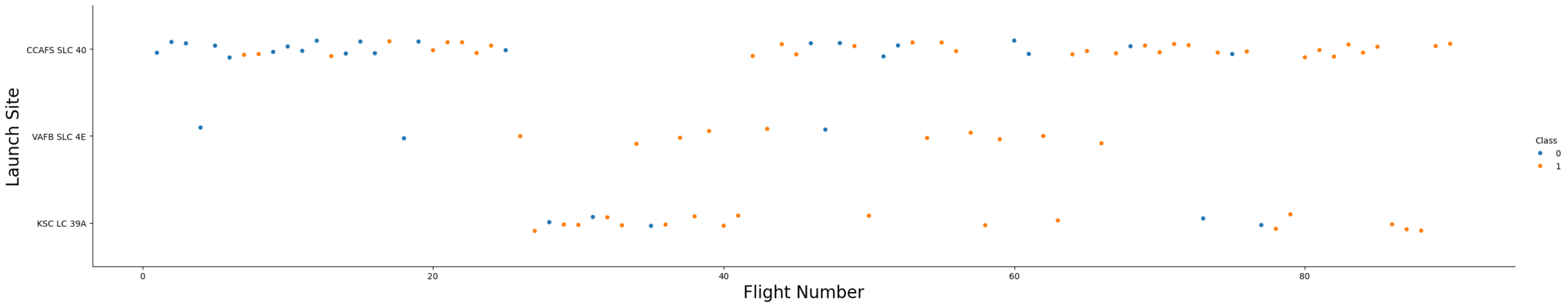
- The following results are presented in the next section:
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

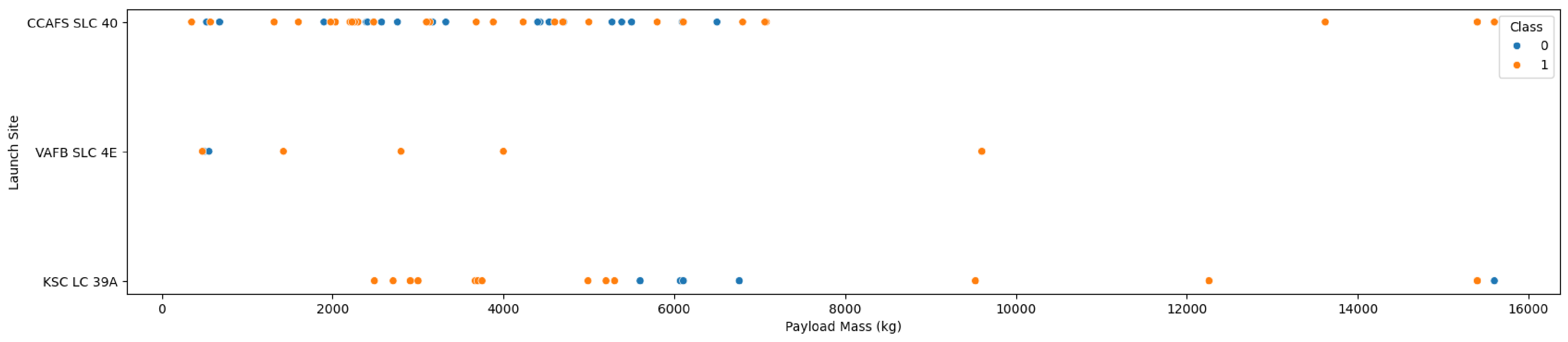
- Flight numbers are plotted against launch site, with the markers indicating successful landing (Class = 1, orange) or failure (Class = 0, blue)



- Failed missions were more frequent for the earlier flights.
- Once the launch site KSC LC 38A are in operation (Flight #27), the success rate for the different sites does not vary substantially.

Payload vs. Launch Site

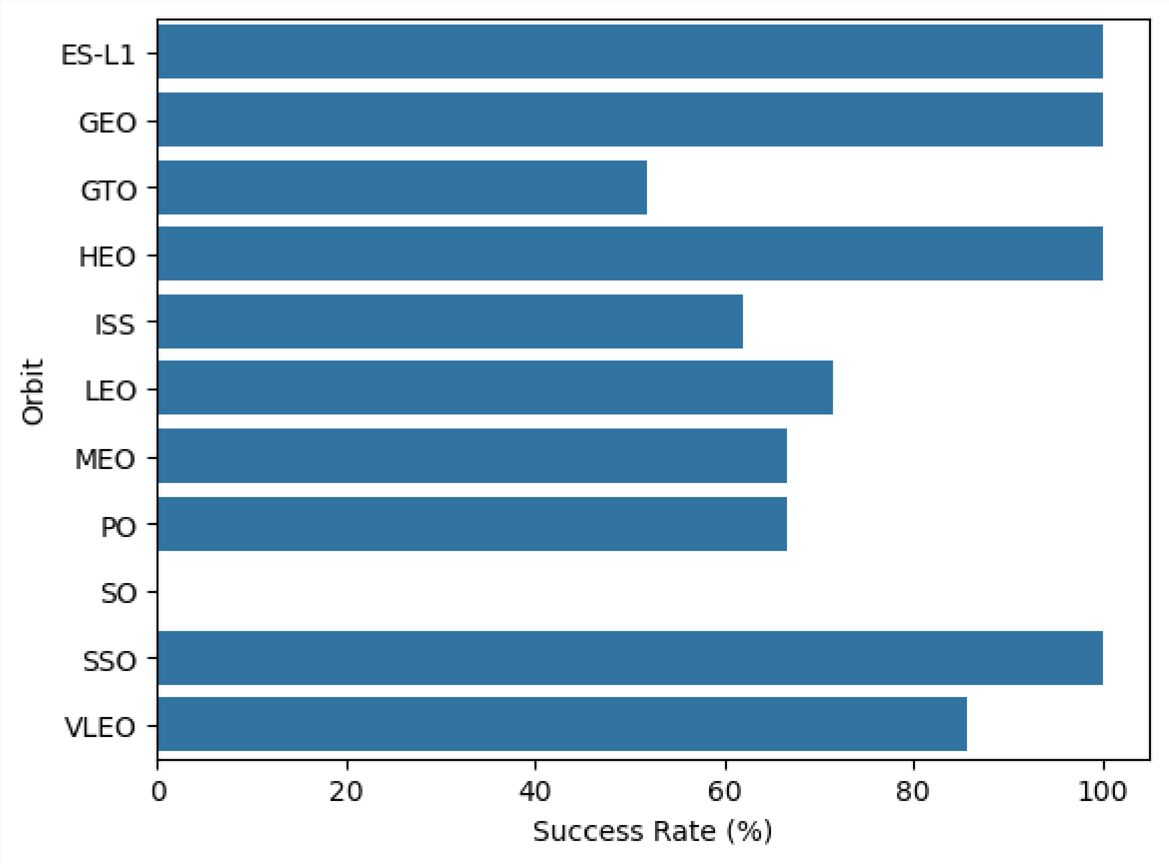
- Payload mass (kg) are plotted against launch site, with the markers indicating successful landing (Class = 1, orange) or failure (Class = 0, blue)



- The heavier payloads (> 10,000 kg) are only launched from CCAFS SLC 40 and KSC LC 39A in Florida.
- There is no clear correlation between payload mass and outcome.

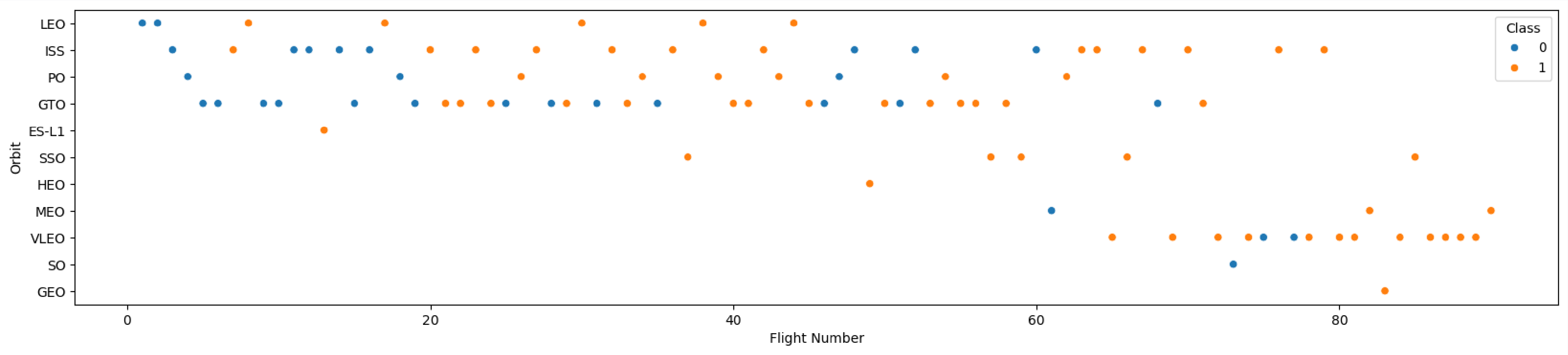
Success Rate vs. Orbit Type

- Success rate (%) for each of the mission orbit types is shown.
- Missions to ES-L1, GEO, HEO, and SSO have the highest success rate (100%)



Flight Number vs. Orbit Type

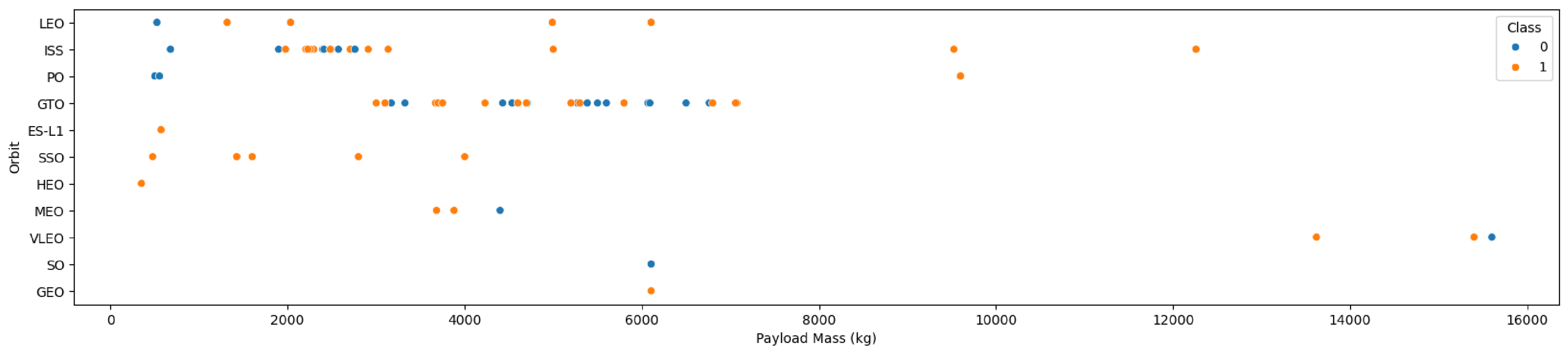
- Flight numbers are plotted against the orbit types, with the markers indicating successful landing (Class = 1, orange) or failure (Class = 0, blue)



- Success rate for missions to LEO appears to be dependent on flight number.
- For missions to the other orbit types, flight number does not appear to affect the success rate

Payload vs. Orbit Type

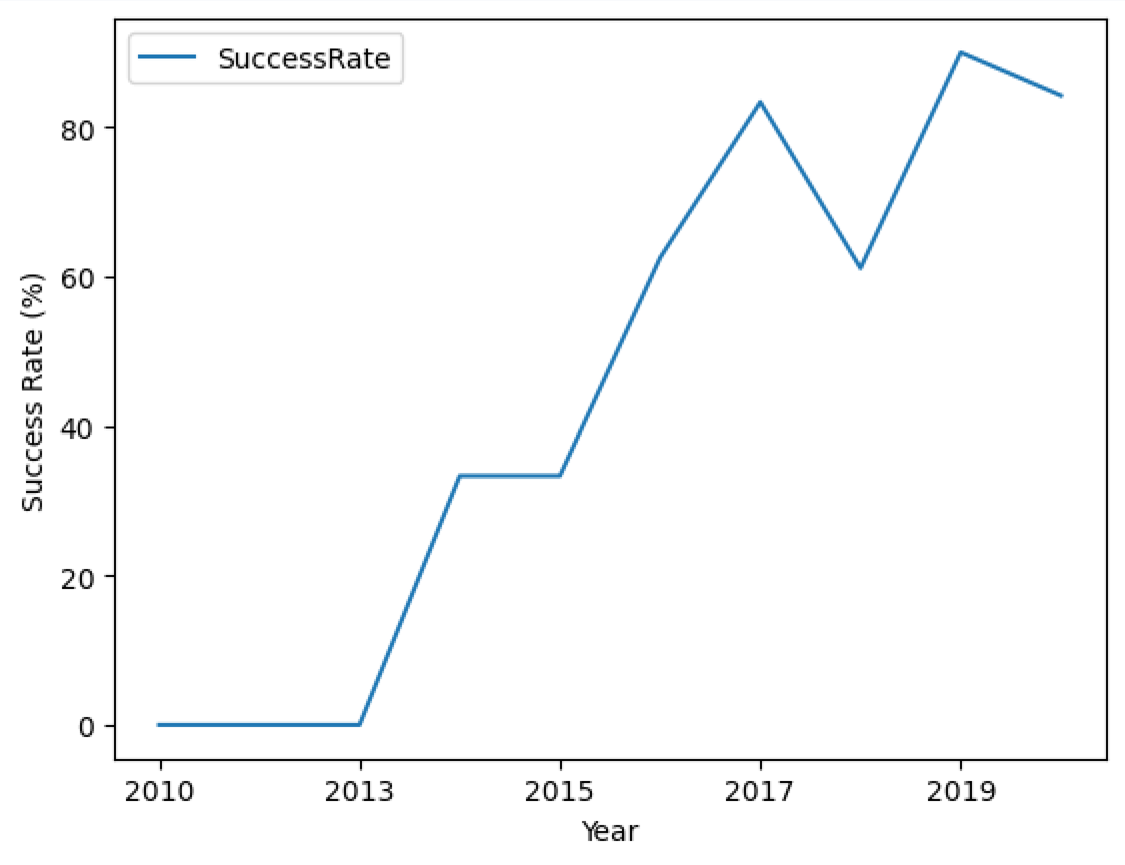
- Payload mass (kg) are plotted against the orbit types, with the markers indicating successful landing (Class = 1, orange) or failure (Class = 0, blue)



- Heavier loads (> 10,000 kg) are mainly for ISS and VLEO.
- Payload mass and orbit type do not appear to be related to the outcome.

Launch Success Yearly Trend

- Success rate (%) for each year is plotted in a line graph.
- Success rate has been increasing substantially from 2013 (0%) to 2020 (over 80%).



All Launch Site Names

- The unique launch sites are:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- CCAFS LC-40 is being renamed to CCAFS SLC-40 to align with SpaceX's launch site code convention (for details, please refer to https://en.wikipedia.org/wiki/Cape_Canaveral_Space_Launch_Complex_40)

Launch Site Names Begin with 'CCA'

- The older code name of CCAFS LC-40 is used for the earlier missions as shown below:

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

Total Payload Mass

- Among the 20 missions for NASA (CRS), the total payload mass is 45596 kg as shown below:

count(PAYLOAD_MASS__KG_)	sum(PAYLOAD_MASS__KG_)
20	45596

Average Payload Mass by F9 v1.1

- Among the 15 missions carried by the booster version F9 v1.1, the average payload mass is 2535 kg as shown below:

Count	AVG(PAYLOAD_MASS__KG_)
15	2534.666666666665

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is 2015-12-22.
- Mission details are shown below:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

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 are:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are:
  - Failed: 1
  - Success: 100
- The numbers are calculated from the query result:

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The names of the 12 boosters which have carried the maximum payload mass (15,600 kg) are:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- In 2015, the failed landing outcomes on drone ship, their booster versions, and launch site names are:

month_	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The counts of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order are shown below:

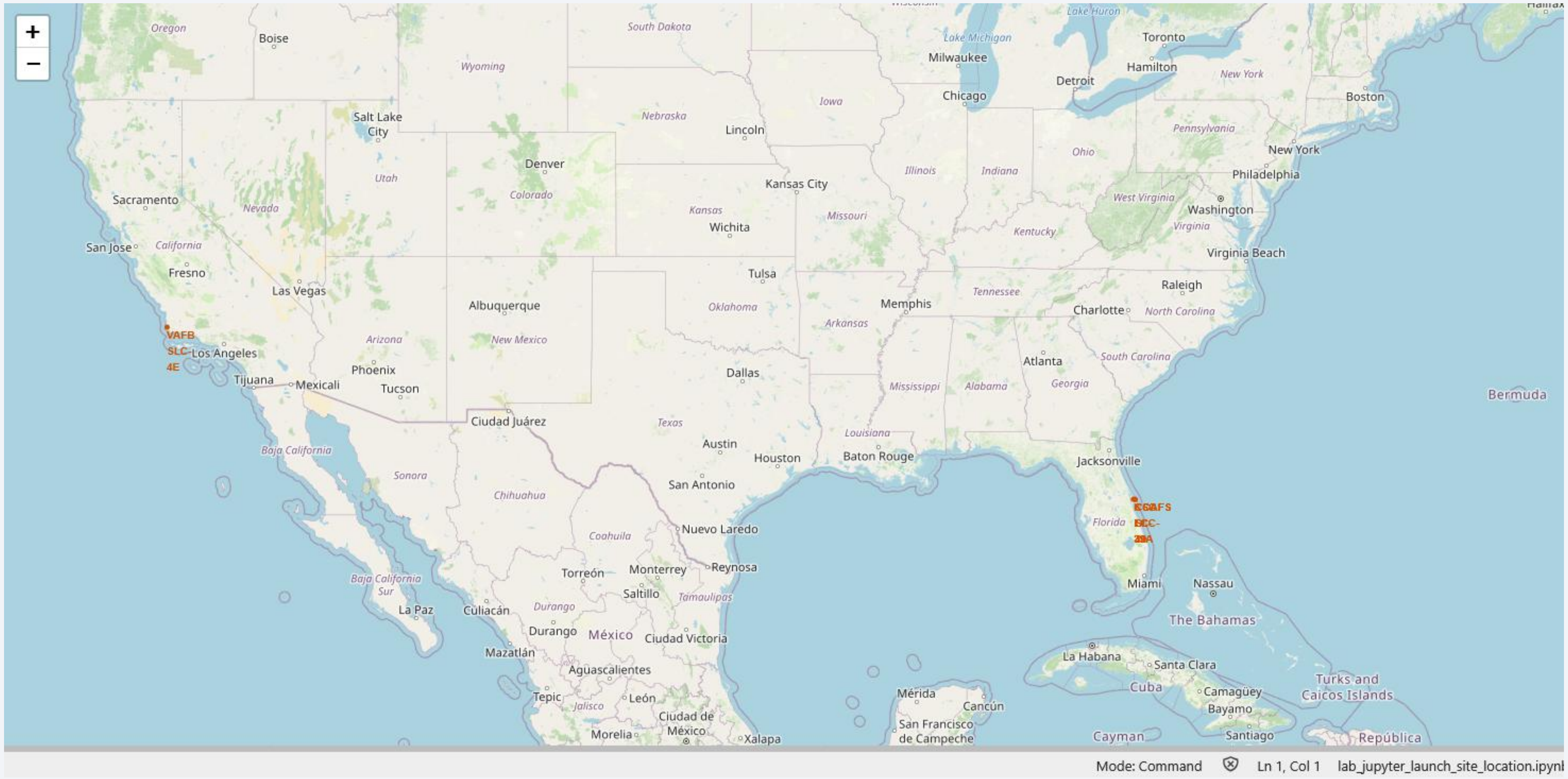
Landing_Outcome	count(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

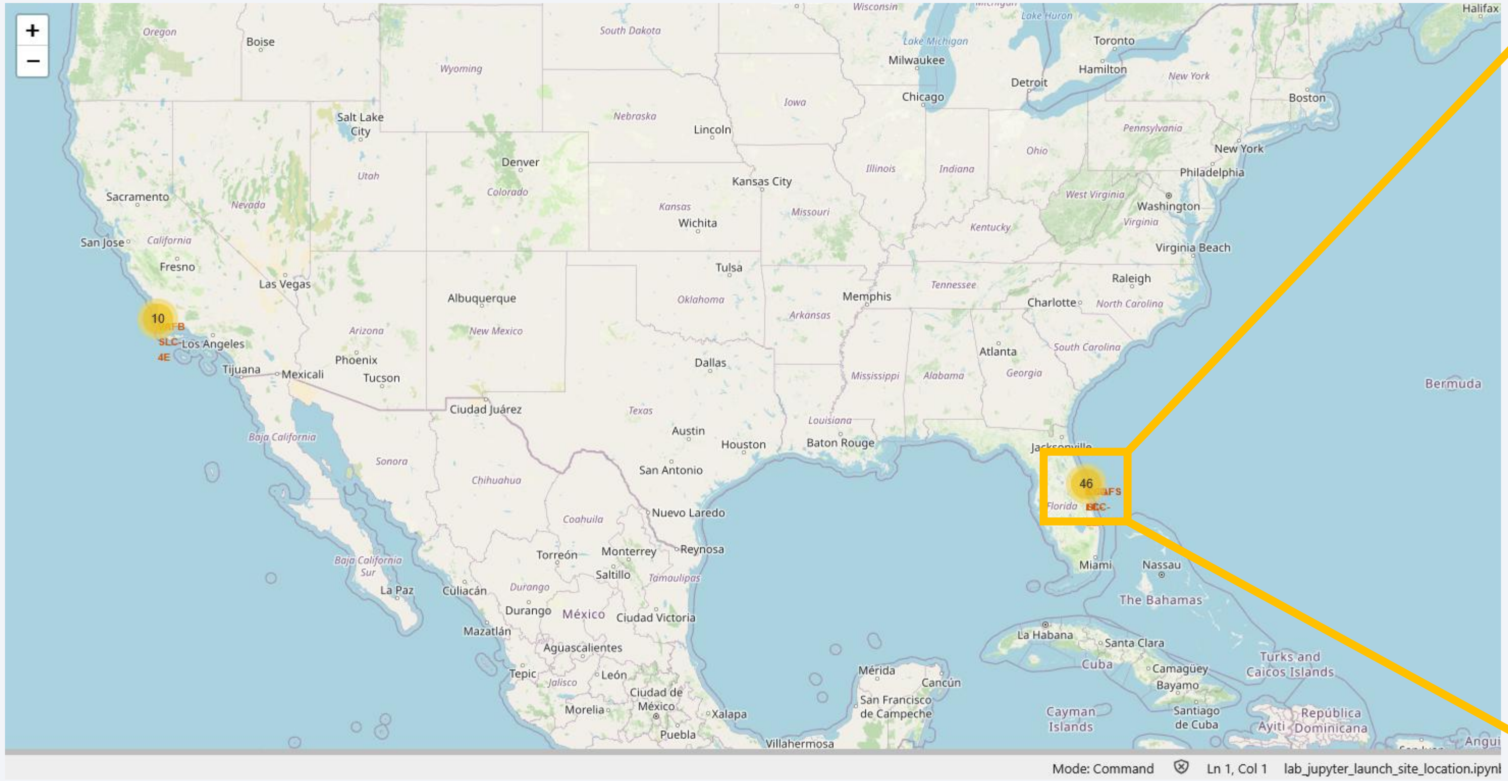
SpaceX Launch Sites

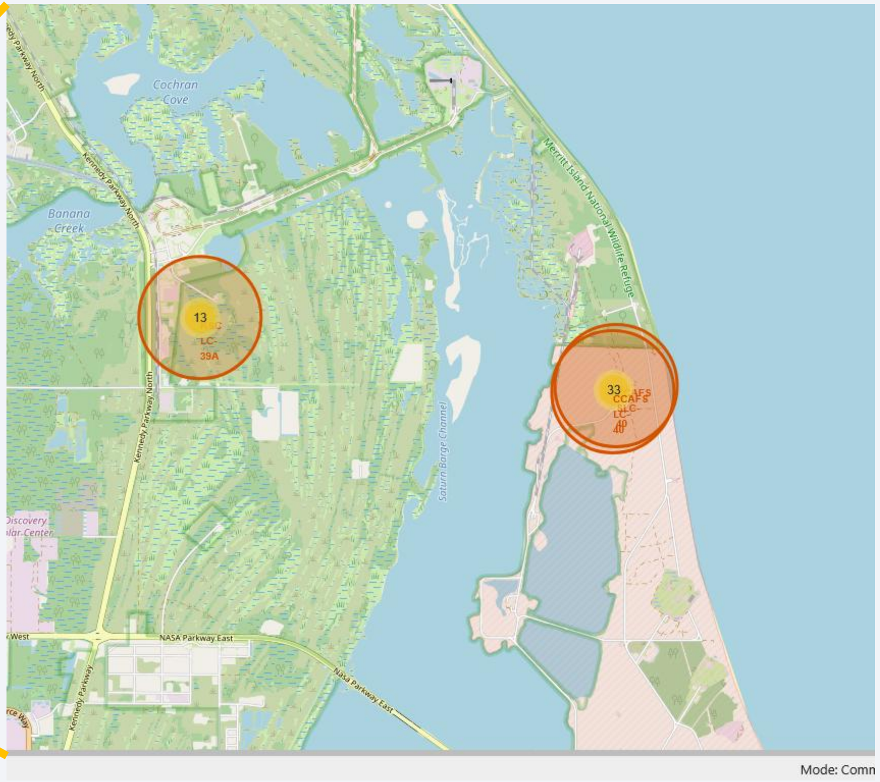
- The following map shows the SpaceX launch sites (in Red) in Southern California and Florida



Falcon 9 Rockets Successful Launch Locations

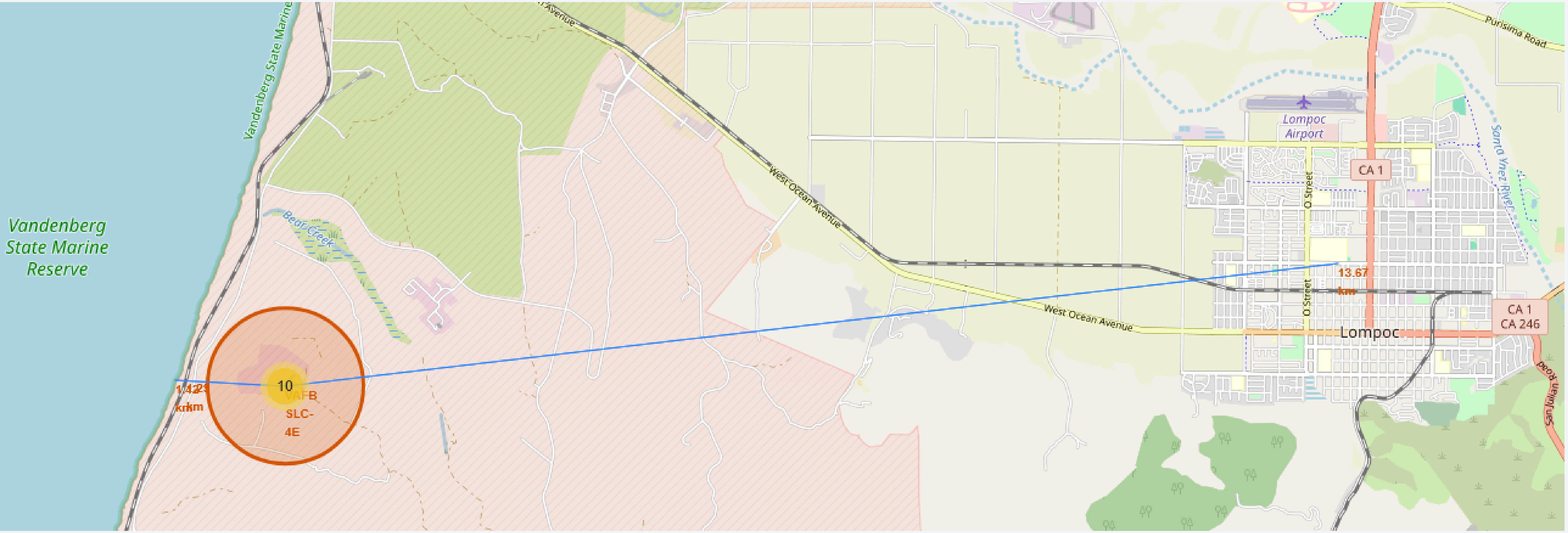
- The following maps show the number of successful landing missions launched from each of the launch sites.





Distance from VAFB SLC-4E

- The following map shows the distance from the launch site VAFB SLC-4E to the closest coast line (1.42 km) and city centre (Lompoc; 13.67 km).

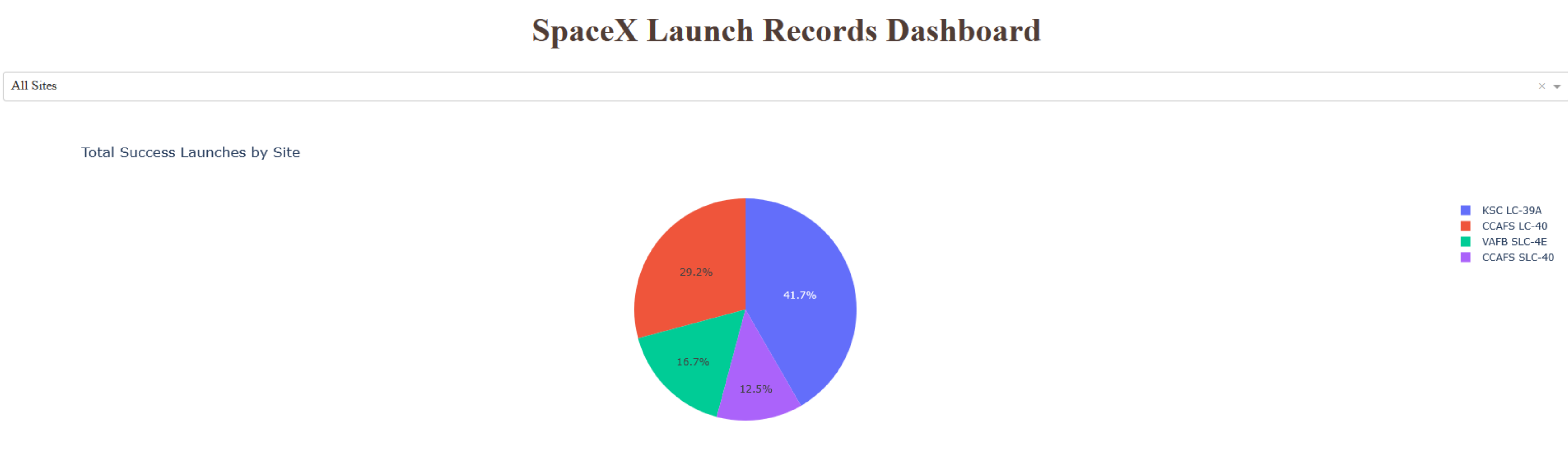


Section 4

Build a Dashboard with Plotly Dash

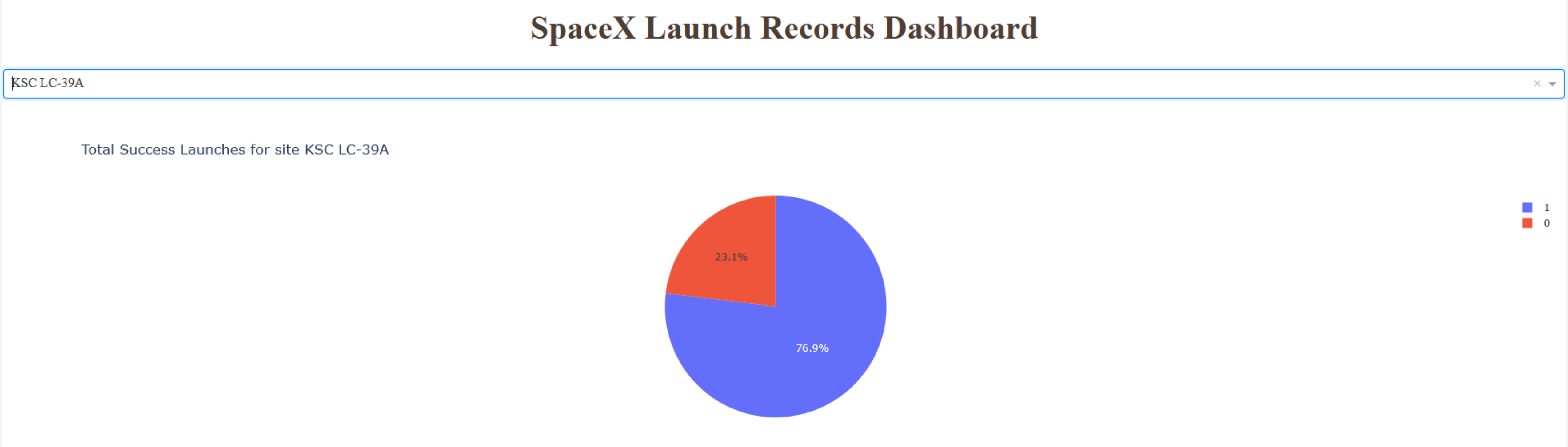
Dashboard – Successful Landing of All Sites

- The following screenshot shows the percentage of successful landing mission from each of the launch sites



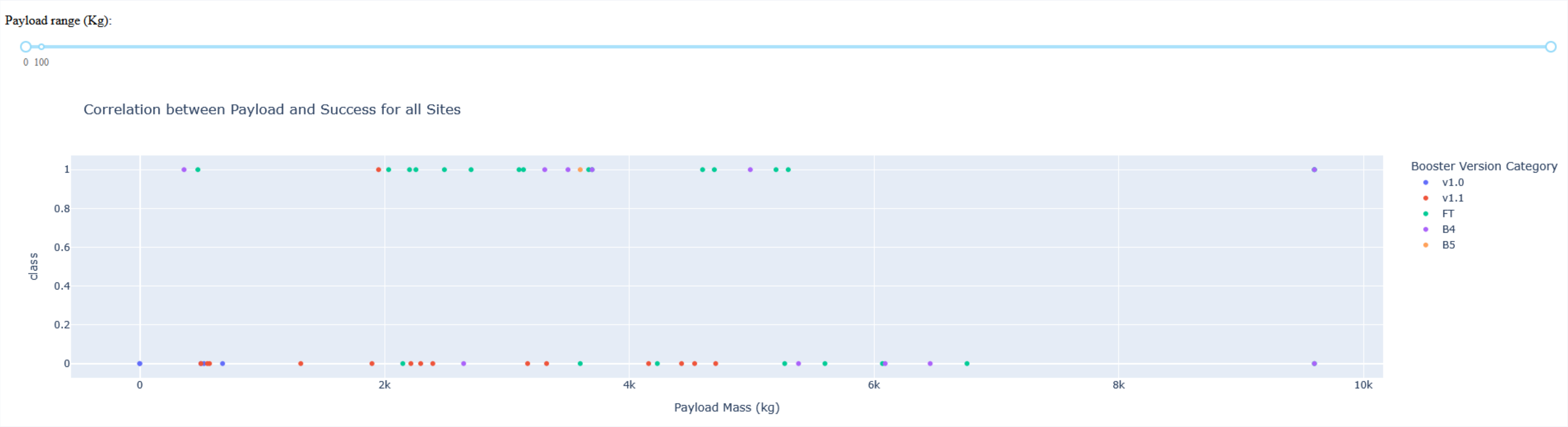
Dashboard – Launch Site with Highest Success Rate

- When a specific launch site is selected from the dropdown list, the percentage of successful landing (blue; 1) and failed landing (Red; 0) are shown as a pie chart.



Dashboard – Payload Mass Slider

- The dashboard also shows Payload Mass vs. Launch Outcome as a scatter plot for all sites with the booster versions shown as different colours.



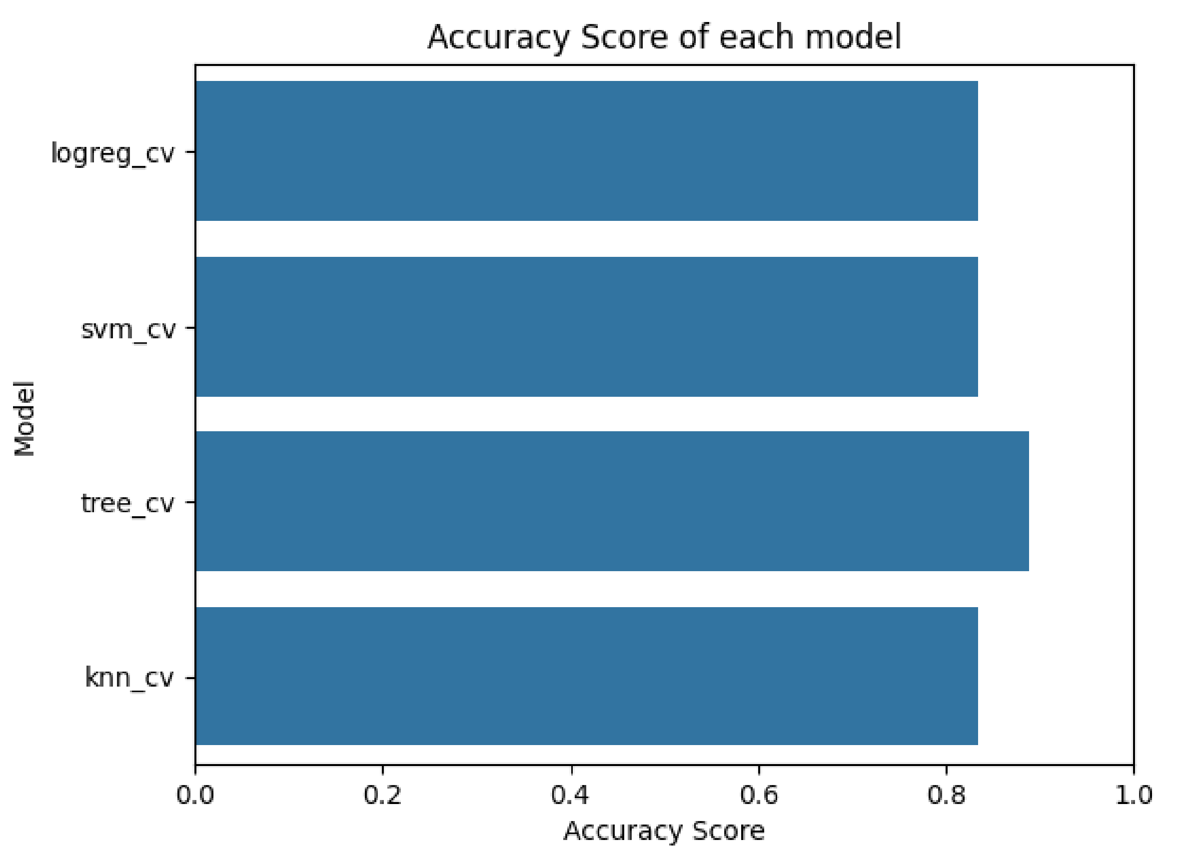
- The FT booster appears to have higher success rate than other boosters.

Section 5

Predictive Analysis (Classification)

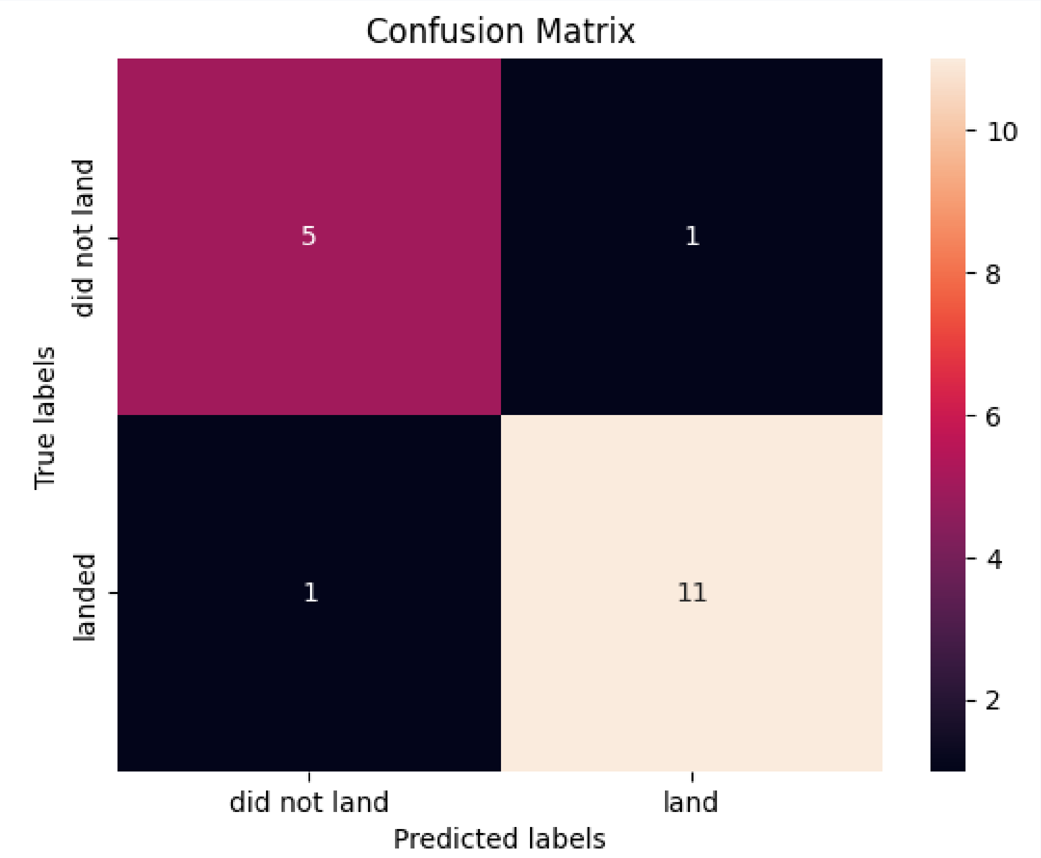
Classification Accuracy

- The bar graph on the right shows the model accuracy (score) for all built classification models.
- The Decision Tree Classification model has the highest score (0.889).



Confusion Matrix

- The Decision Tree Classification Model correctly predicts 11 out of the 12 successful missions (bottom row) and 5 out of the 6 failed missions (top row).



Conclusions

- Individual features such as launch sites, payload mass, and orbit type do not seem to correlate substantially with successful landing mission.
- There is a clear improvement of success rate for SpaceX from 2013 (0%) to 2010 (over 80%).
- Each of the launch sites is much closer to the coastline than to a city.
- The highest success rate observed for a launch site is 76.9%.
- Using features such as launch site, payload mass, orbit, booster version, grid fins collectively in a decision tree classification model can predict the correct landing outcome with 88.9% accuracy.

Appendix

- For details, please refer to the following files on GitHub:

  1. 1_jupyter-labs-spacex-data-collection-api.ipynb
  2. 2_jupyter-labs-webscraping.ipynb
  3. 3_labs-jupyter-spacex-Data%20wrangling.ipynb
  4. 4_jupyter-labs-eda-sql-coursera_sqllite.ipynb
  5. 5_edadataviz.ipynb
  6. 6_lab_jupyter_launch_site_location.ipynb
  7. 7_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb
  8. spacex-dash-app.py

Thank you!