IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michele Chersich
07.07.2025

Outline

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

Executive Summary

- Preventing unsuccessful rocket launches can save aerospace companies millions of dollars for each launch. To this end, we identify the key features of a successful launch and devise a machine learning model to predict launch success or failure. Following the data science methodology, we do data collection, data wrangling, exploratory data analysis, data visualisation, and machine learning prediction.
- Including features like payload mass, number of flights, orbit type, launch site location, the use of specific components (e.g. gridfins and legs), the reuse of rocket components, we tested several machine learning models to predict successful vs. unsuccessful launches.

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each. Much of the savings is because Space X can reuse the first stage.
- If we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

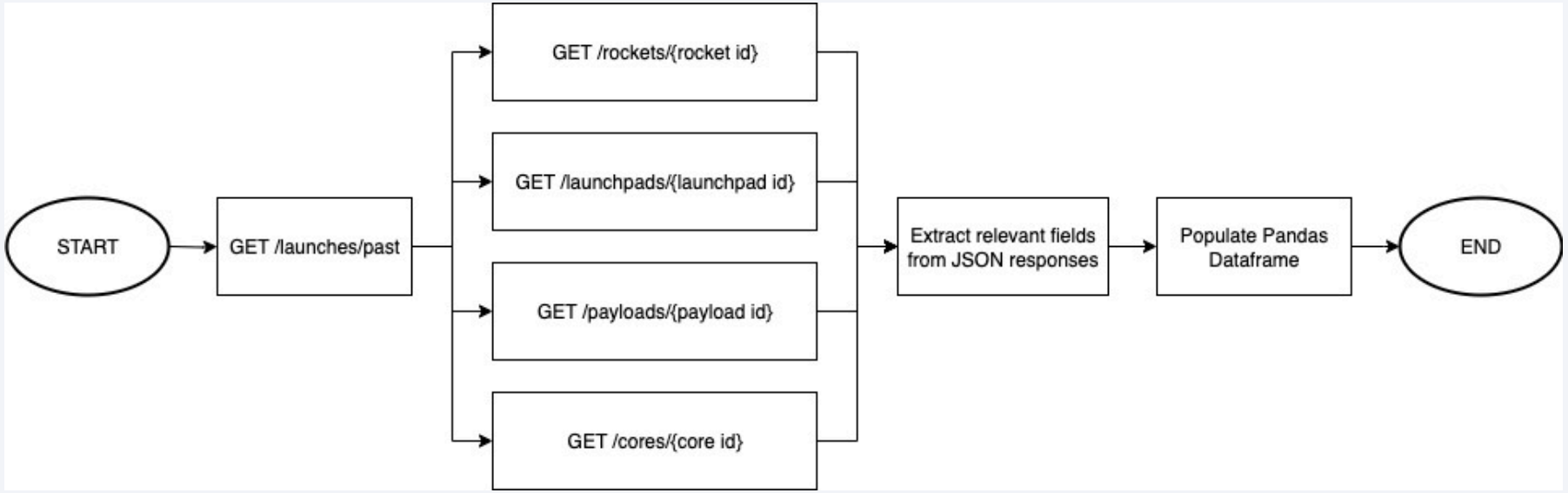
- Data collection methodology:
  - HTTP request to SpaceX API
- Perform data wrangling
  - Convert launch outcomes from string to binary representation (1 for success, 0 for failure)
  - Impute missing payload mass values with the mean
- Perform exploratory data analysis (EDA) using visualisation and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using the Scikit-Learn library, we standardise the data; split the dataset into training and test sets; perform GridSearch to find the best set of hyper-parameters, then train the classifier; evaluate the model's performance as training and test accuracy; repeat for different classification models and compare performance

Data Collection

- The datasets are collected via web requests to the SpaceX API.
- Data collection process:
  - Define business problem: predicting launch outcome
  - Identify data requirements: booster name, launch site name and location, payload mass and orbit type, core related data, and launch outcome
  - Choose data collection method: SpaceX API
  - Collect data from SpaceX API
  - Validate data: check completeness and accuracy

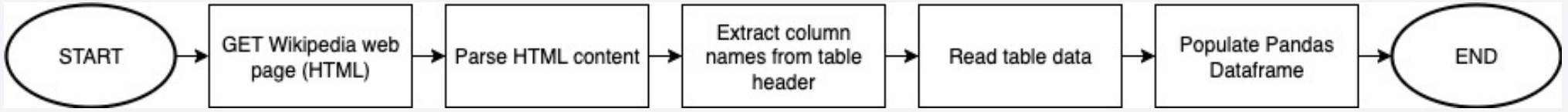
Data Collection – SpaceX API

- Call the SpaceX API: retrieve launches, then for each launch retrieve rocket, launchpad, payload, and core data; extract relevant data from JSON responses and create a Pandas Dataframe
- https://github.com/k33rs/data-science-capstone/blob/main/1-data/spacex-data-collection-api-v2.ipynb



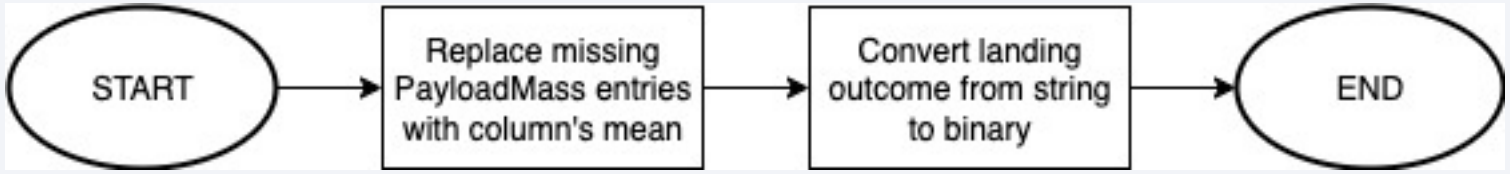
Data Collection - Scraping

- Fetch the HTML list of Falcon 9 launches (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922); parse HTML content with BeautifulSoup; extract column names from table header; read table data and populate Pandas Dataframe
- https://github.com/k33rs/data-science-capstone/blob/main/1-data/spacex-webscraping.ipynb



Data Wrangling

- Replace missing PayloadMass entries with column's mean; convert landing Outcome from string ("True …" for success, else failure) to binary (1 for success, 0 for failure)
- https://github.com/k33rs/data-science-capstone/blob/main/1-data/spacex-data-wrangling-v2.ipynb



EDA with Data Visualization

- Several scatter plots show relationship between variables (e.g. FlightNumber vs PayloadMass, FlightNumber vs LaunchSite, PayloadMass vs LaunchSite, etc.), highlighting launch success or failure for each data point; a bar plot is used to display the success rate for each Orbit type; a line chart displays the yearly success rate from 2010 to 2020
- https://github.com/k33rs/data-science-capstone/blob/main/2-exploratory-data-analysis/spacex-eda-dataviz-v2.ipynb

EDA with SQL

- The following SQL queries are performed:
  1. Display the names of unique launch sites in the space mission
  2. Display 5 records where launch sites begin with the string ‘CCA'
  3. Display the total payload mass carried by boosters launched by NASA (CRS)
  4. Display the average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome in ground pad was achieved
  6. List the name of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  7. List the total number of success and failure mission outcomes
  8. List all the booster versions that have carried the maximum payload mass
  9. List the records which will display month names, failure landing outcomes in drone ship, booster versions, and launch site for the months in year 2015
  10. Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- https://github.com/k33rs/data-science-capstone/blob/main/2-exploratory-data-analysis/spacex-eda-sqlite.ipynb

Build an Interactive Map with Folium

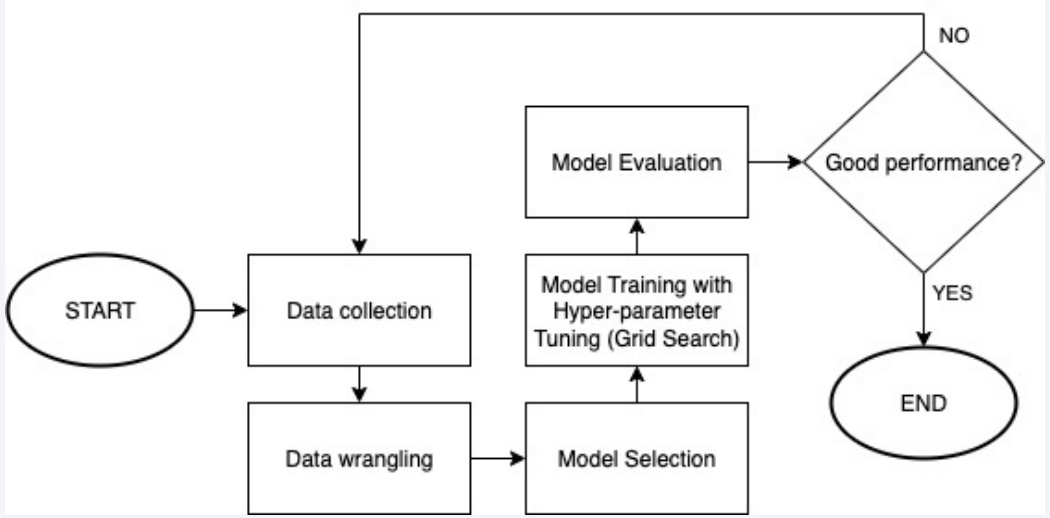
- We mark launch sites on the map with circles, markers, marker clusters, and poly lines
- Circles highlight launch site locations, markers label locations with launch site names; marker clusters are used to display each launch site's missions with green and red icons (green for success, red for failure); poly lines are drawn to highlight the distance between launch sites and nearest points of interest (e.g. coastline, city, railway, highway, etc.)
- https://github.com/k33rs/data-science-capstone/blob/main/3-visual/spacex-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- The upper part of the dashboard features a dropdown menu and a pie chart: the options are either all launch sites or individual launch sites: the former displays each site's success rate out of all successful launches, the latter shows each launch site's success vs failure rate (1 for success, 0 for failure)
- The bottom part features a slider and a scatter plot, which plots payload mass vs launch outcome for the current dropdown selection, with booster version as hue; the slider allows to select the range of the payload mass (i.e. x axis)
- The pie chart allows to rank successful launch sites, while the scatter plot investigates a correlation between payload mass and launch outcome
- https://github.com/k33rs/data-science-capstone/blob/main/3-visual/spacex_dash_app.py

Predictive Analysis (Classification)

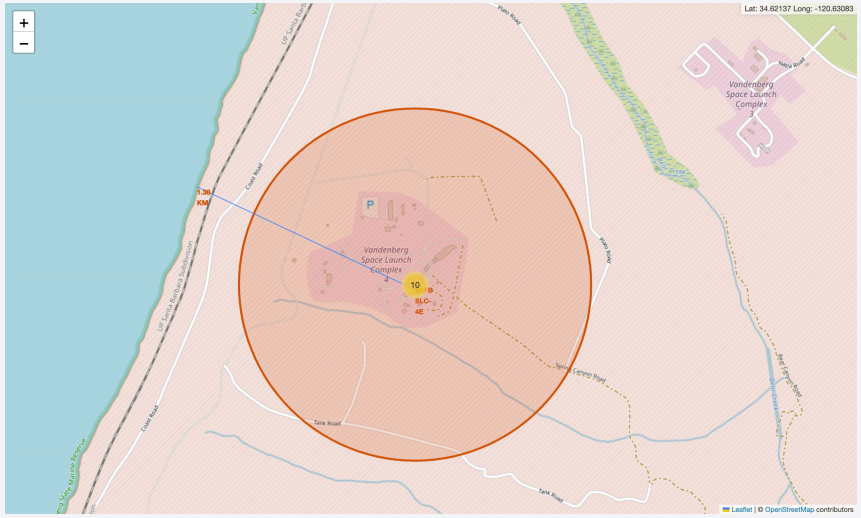
- Tested four classification models: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors; found the best hyper-parameters for each model via grid search, i.e. trained each model multiple times and retained the instance with the best parameters; evaluated each model's performance based on prediction accuracy on test data
- N.B. In addition to what previously done in the data wrangling phase, we standardise the input, and split it into training and test sets
- https://github.com/k33rs/data-science-capstone/blob/main/4-machine-learning/spacex-machine-learning-prediction-v1.ipynb



Results: Exploratory data analysis

- Success rate grows with the number of flights
- Payload mass needs to be within some limit to control risks
- ES-L1, GEO, HEO, and SSO orbit types have best success rate
- No correlation between flight number and orbit type with respect to success rate
- The yearly success rate surged from 2013 to 2017

Results: Interactive analytics demo



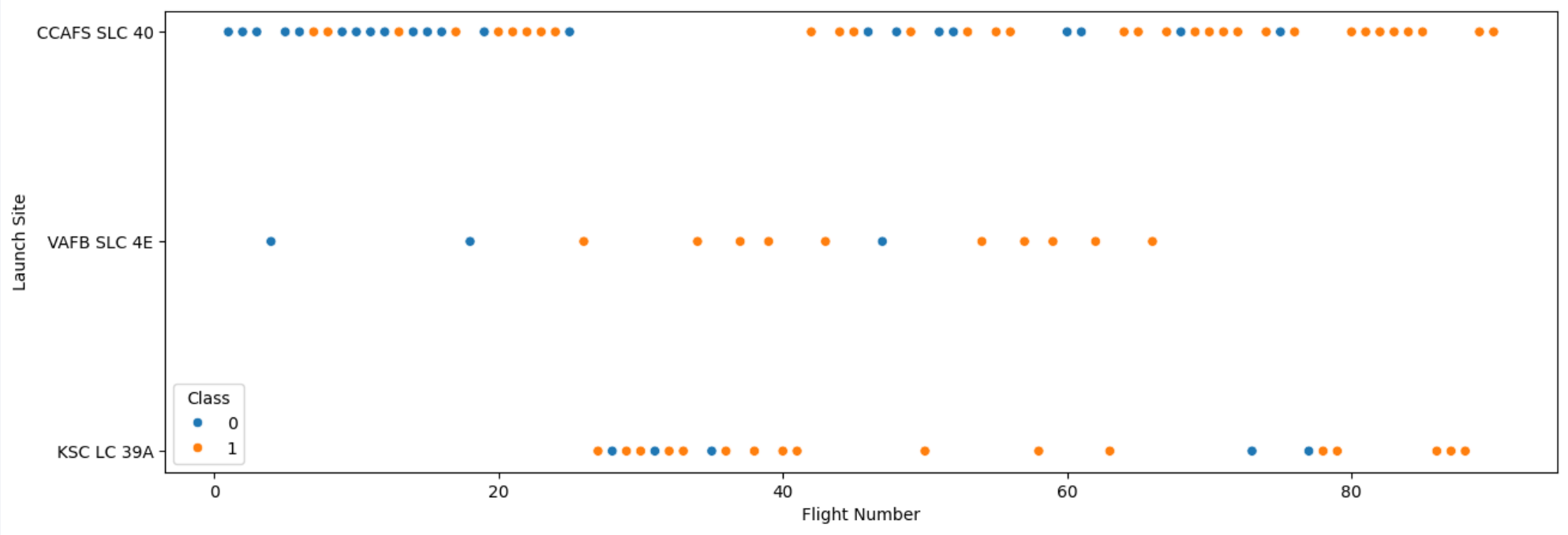
Results: Predictive Analysis

- All tested models (Logistic Regression, Decision Trees, Support Vector Machine, and K-Nearest Neighbors) achieve an accuracy score of 83.33% on test data
- The performance is good for all models, i.e. any of these is a good choice

Section 2

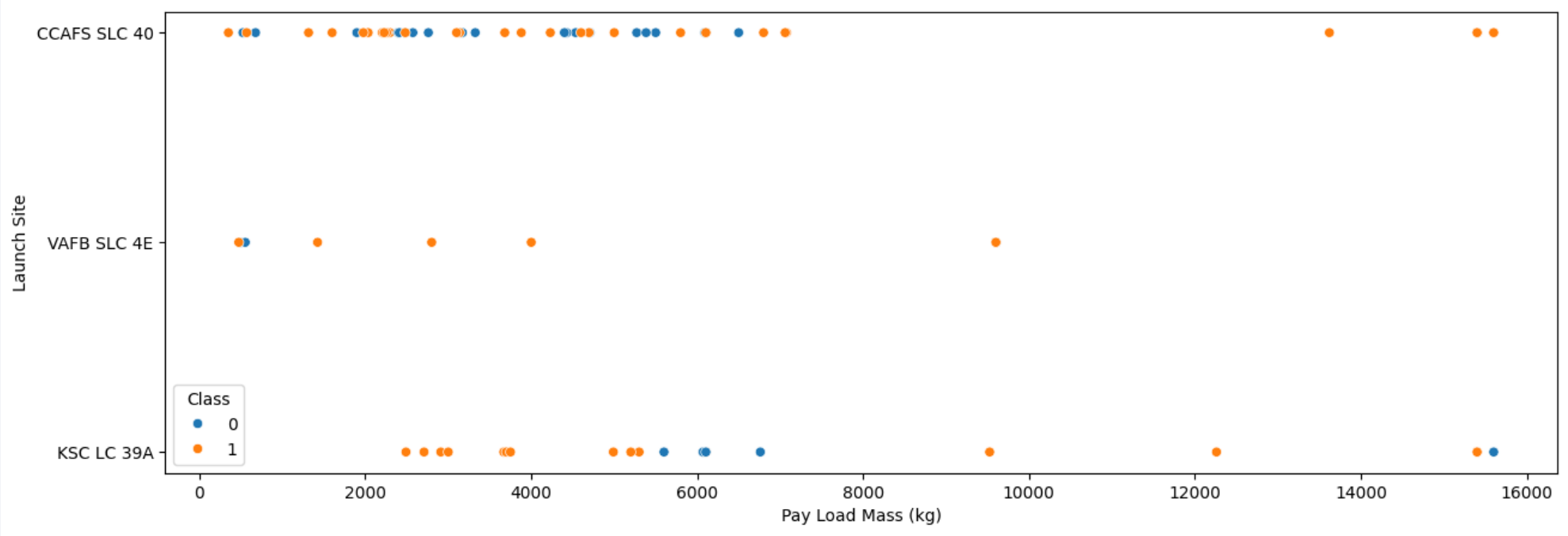
Insights drawn from EDA

Flight Number vs. Launch Site



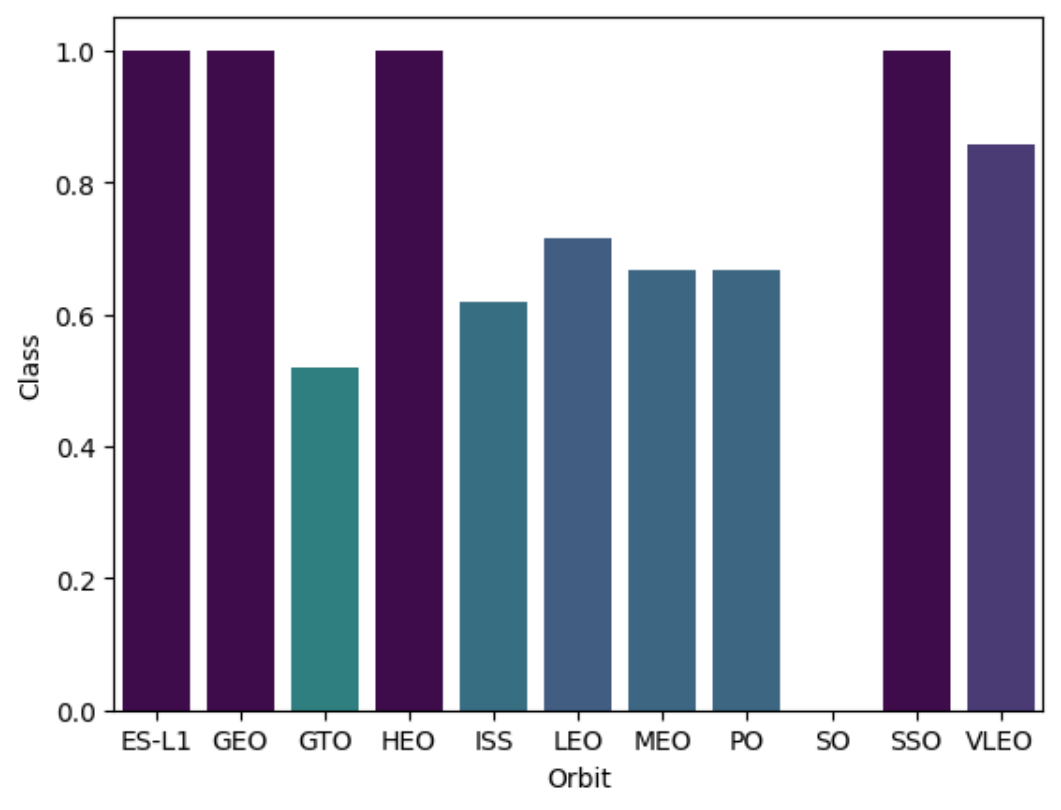
- Mission outcome (1 for success, 0 for failure) as a function of number of flights and launch sites
- We observe a positive correlation between number of flights and success rate

Payload vs. Launch Site



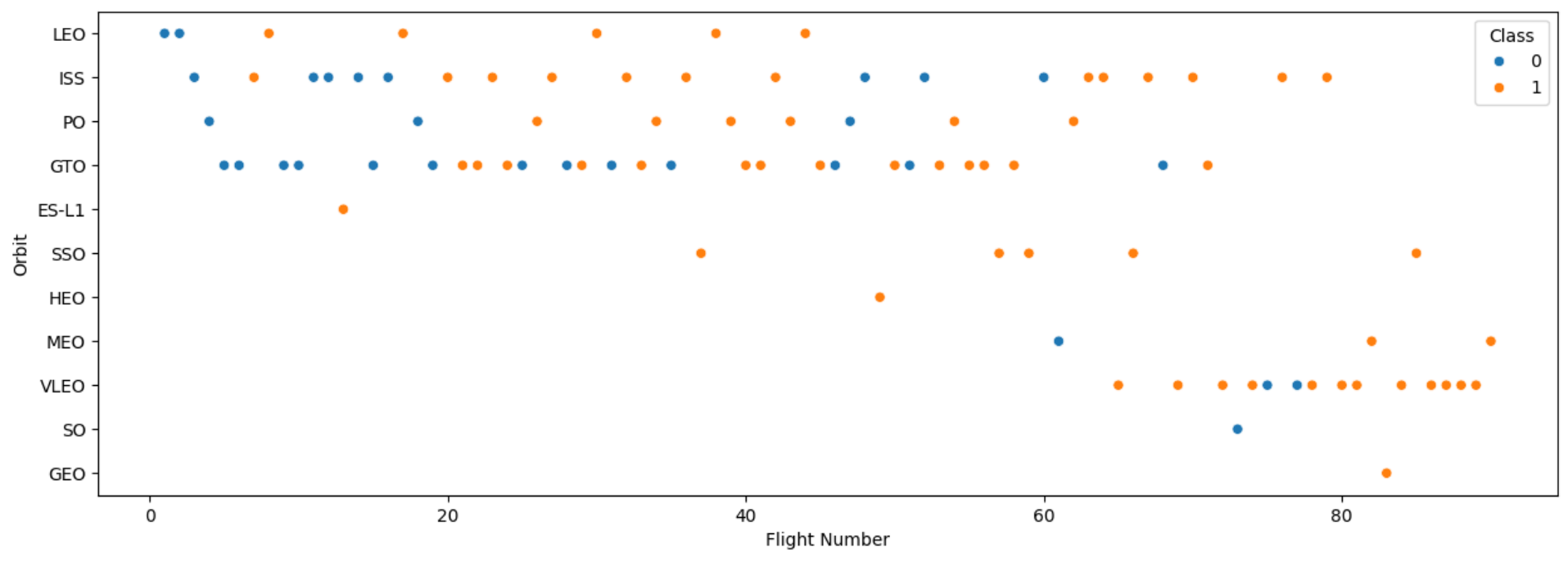
- Mission outcome (1 for success, 0 for failure) as a function of payload mass and launch sites
- We observe higher success rate when payload does not exceed some threshold (e.g. KSC LC 39A below 6000)
- Very few launches have high payload (i.e. above 8000)

Orbit Type vs. Success Rate



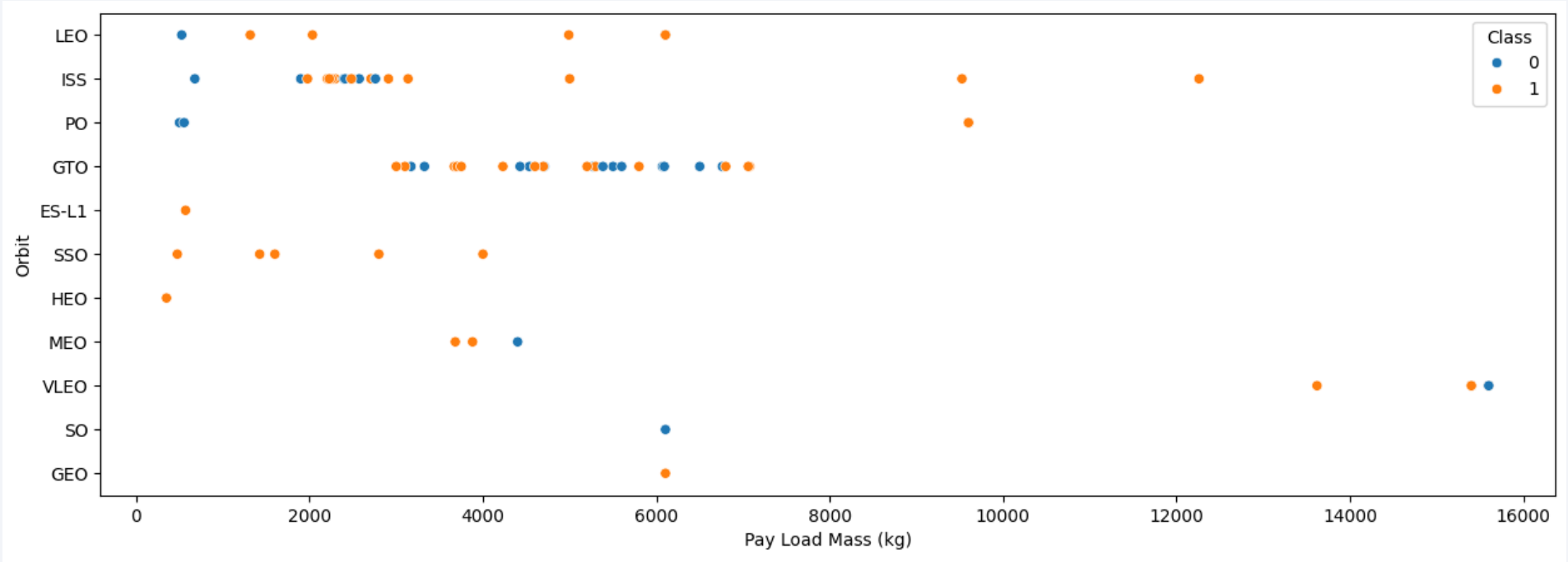
- A bar chart to quantify success rate for each orbit type
- ES-L1, GEO, HEO, and SSO have maximum success rate, followed by VLEO, LEO, etc.

Flight Number vs. Orbit Type



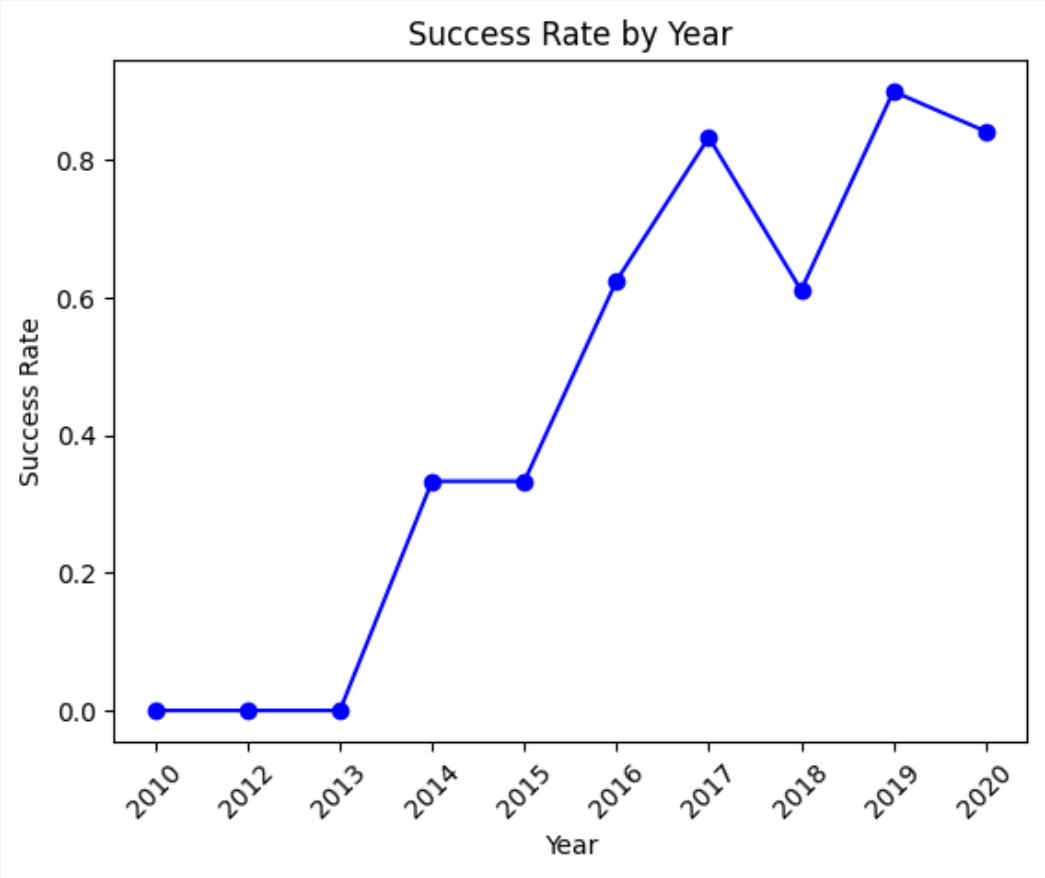
- In the LEO orbit, the success appears to be correlated to the number of flights
- No relationship in the GTO orbit
- Therefore, the correlation between number of flights and success is not affected by the orbit type

Payload vs. Orbit Type



- With heavy payloads, the success rate is higher for LEO, ISS, and PO
- For GTO, no correlation; for other orbits, not enough data points
- Therefore, correlation is possible but not certain

Launch Success Yearly Trend



- Success rate from 2010 to 2013 was stable
- Surged from 2013 to 2017, with a plateau from 2014 to 2015

All Launch Site Names

- Find the names of the unique launch sites
- There are many records in the SQL table, so the “Launch_Site” column contains repetitions
- We use the “distinct” keyword to remove duplicates (see Appendix A)

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Use a regular expression to filter launch sites, then select the first 5 records (see Appendix B)

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

- Calculate the total payload carried by boosters from NASA
- Filter the customer column for ‘NASA (CRS)’, then sum payload mass entries (see Appendix C)

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Filter the booster version column for ‘F9 v1.1’, then average payload mass entries (see Appendix D)

average_payload_mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Filter landing outcome for ‘Success (ground pad)’, then select the minimum date (see Appendix E)

earliest_success_ground_pad
2015-12-22

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
- Filter landing outcome for ‘Success (drone ship)’ and payload mass between 4000 and 6000; booster version contains duplicates, so use ‘distinct' keyword for selection (see Appendix F)

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- We must perform two queries, so we nest them; we use regular expressions to filter mission outcome for success and failure, then count the number of entries for each (see Appendix G)

successful_missions	unsuccessful_missions
100	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Filter payload for maximum value by using a nested query, then select booster version (see Appendix H)

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 the failed landing_outcomes in drone ship, their booster versions, and launch site names in year 2015
- Filter date for year 2015 using *substr* function, and filter landing outcome for 'Failure (drone ship)', then select columns (see Appendix I)

launch_year	launch_month	Booster_Version	Launch_Site	Landing_Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

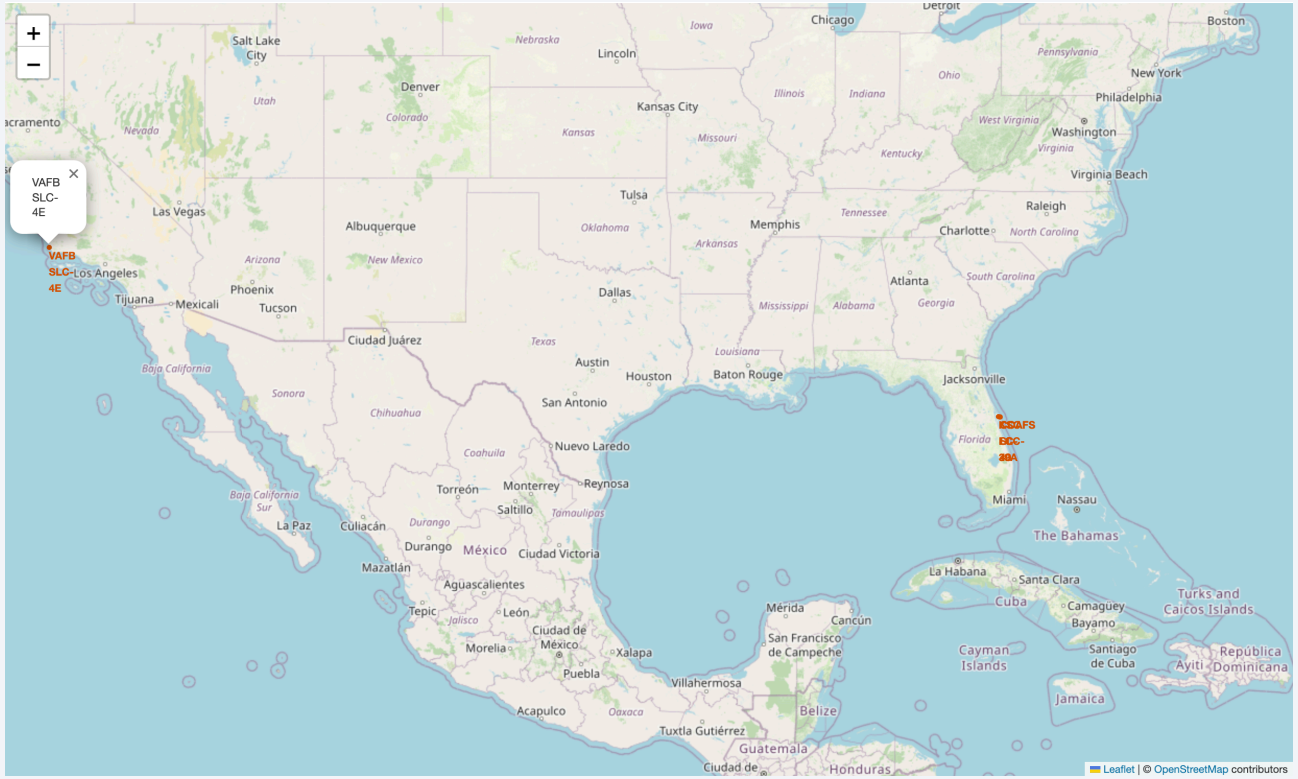
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Filter date between 2010-06-04 and 2017-03-20, group by landing outcome, then select landing outcome and count the number of entries in each group (see Appendix J)

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

Section 3

Launch Sites Proximities Analysis

Folium: launch site locations

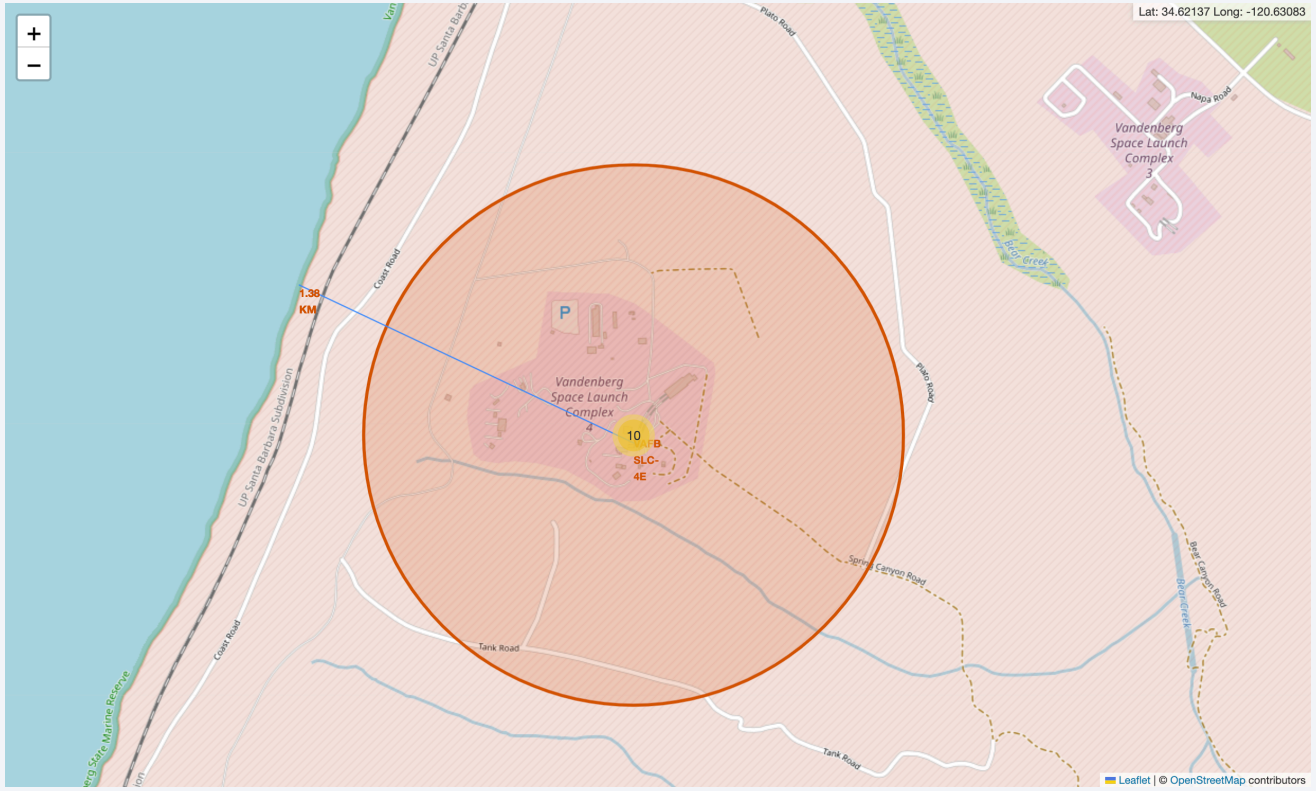


- Locations are marked using circles and markers (launch site names)
- Clicking on the circles shows popups with launch site names

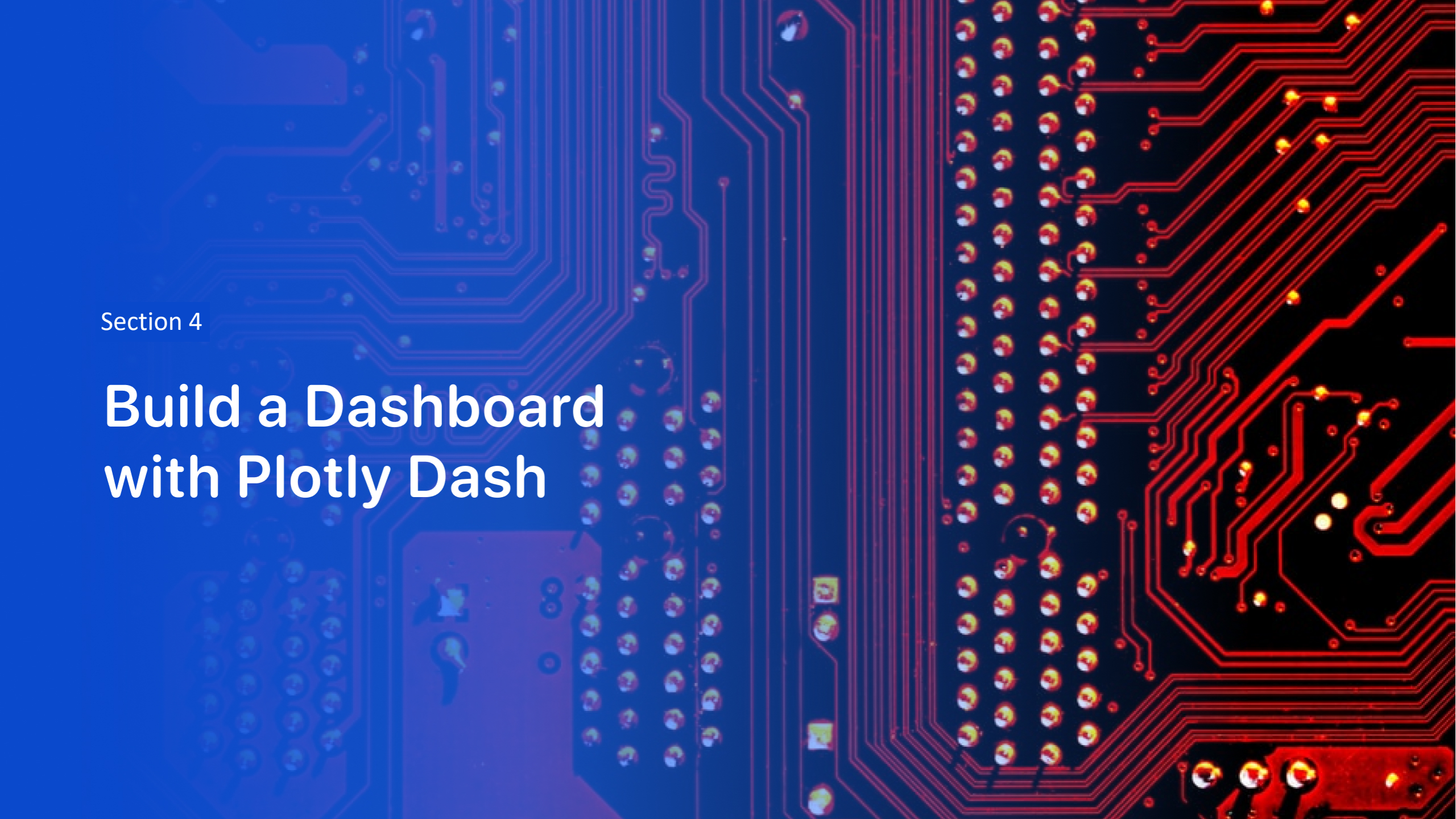
Folium: launch sites' mission outcomes

- We add multiple icons at once using marker clusters, choosing the color depending on mission outcome
- Popup items contain launch site name and mission outcome

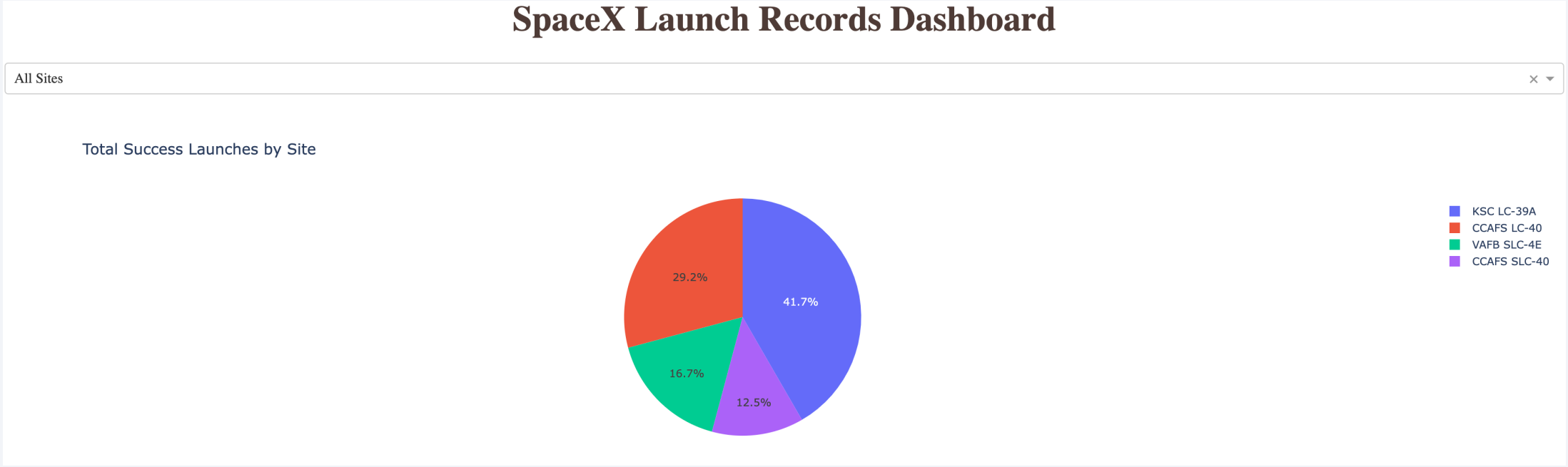
Folium: launch site and proximities



- We draw a poly line from the launch site location to a nearby point of interest (e.g. coastline)
- The locations are obtained by adding a mouse position tracker to the map to read latitude and longitude on the top right
- We then add a marker at the coastline to display the distance from the launch site

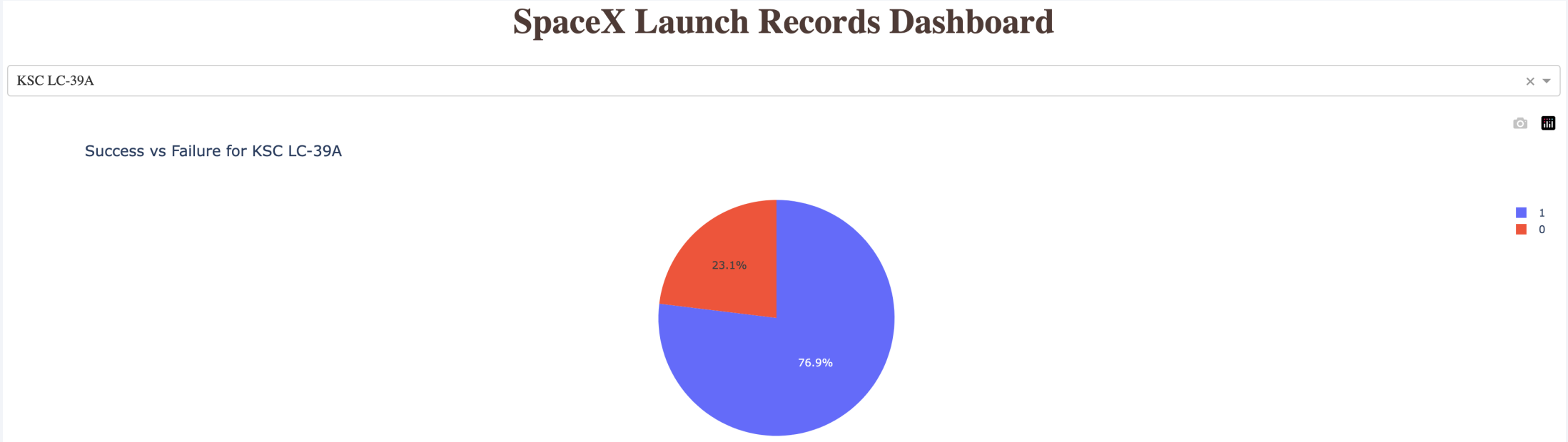

Section 4
Build a Dashboard with Plotly Dash

Launch success: all sites



- Selection is done through the dropdown menu on top
- KSC LC-39A is the most successful launch site (with 41.7% of all successful launches), followed by CCAFS LC-40 (with 29.2%)

Launch success: best site



- Success rate of launch site KSC LC-39A is 76.9%

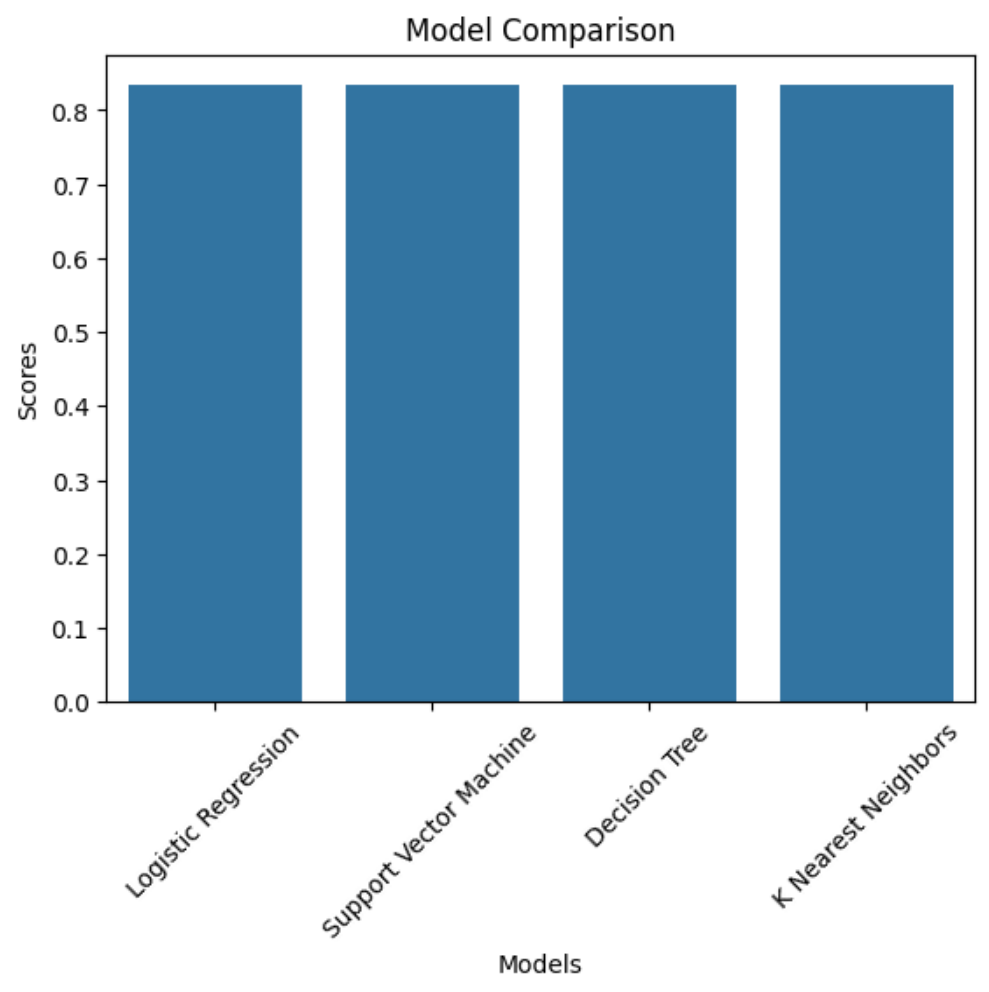
Payload vs. Launch outcome for all sites

- Above a payload mass of 6000 kg, the success rate is very low, so we omit it from the chart
- The most successful range is 2000-4000 kg
- The most successful booster version is FT

Section 5

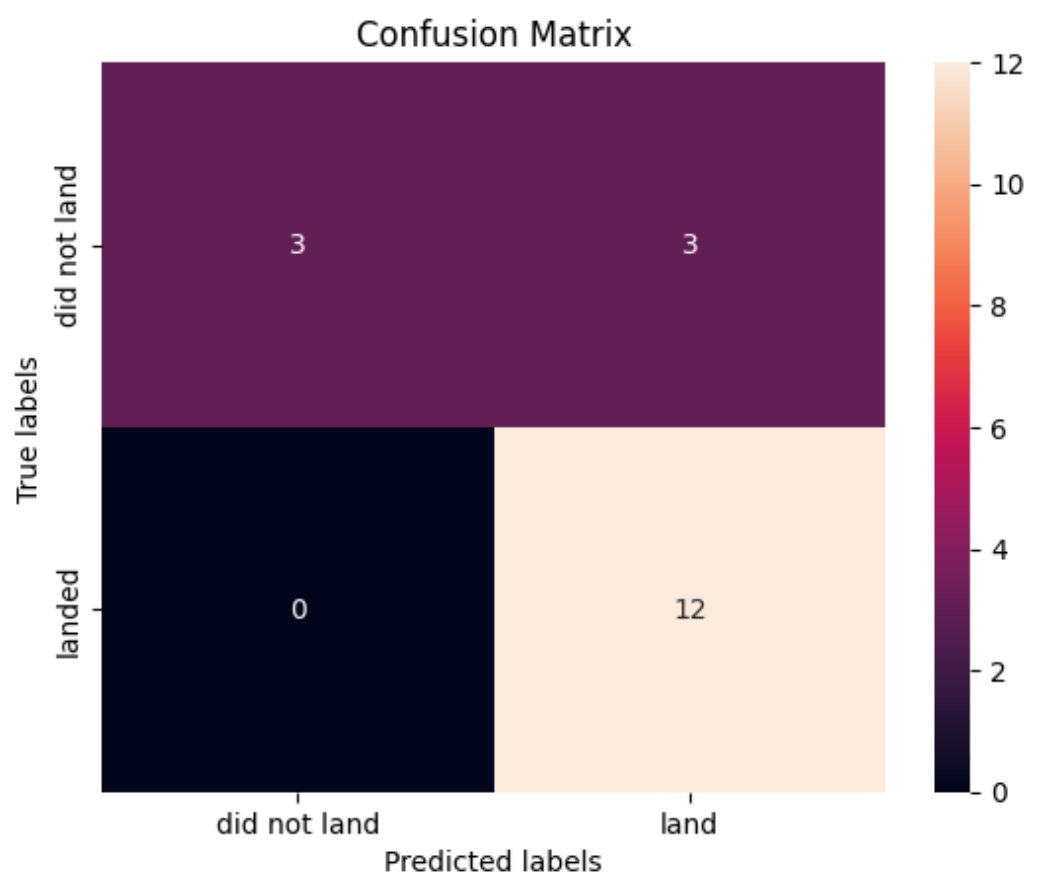
Predictive Analysis (Classification)

Classification Accuracy



- All models have the same classification accuracy of 83.33%, so any of these is a good choice

Confusion Matrix



- The confusion matrix is the same for all models
- The only problem is false positives (3 out of 15, i.e. a false positive rate of 20%)

Conclusions

- Identifying best data sources is crucial to the entire workflow
- Structuring data in proper format, including feature engineering, enables effective analysis
- Several channels of analysis are useful: data visualisation, database queries, interactive visualisation (maps and dashboards), and predictive analytics with machine learning
- If the problem is not complex, several machine learning models may be a viable choice
- The data science workflow is iterative, with subsequent improvements based on feedback collected at different stages

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

A. select distinct “Launch_Site” from SPACEXTABLE;

B. select * from SPACEXTABLE where “Launch_Site” like 'CCA%' limit 5;

C. select sum("PAYLOAD_MASS__KG_") as total_payload_mass from SPACEXTABLE where "Customer" = 'NASA (CRS)’;

D. select avg("PAYLOAD_MASS__KG_") as average_payload_mass from SPACEXTABLE where "Booster_Version" = 'F9 v1.1’;

E. select min(Date) as earliest_success_ground_pad from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)’;

F. select distinct “Booster_Version" from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)’ and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;

G. select
(select count("Mission_Outcome") from SPACEXTABLE where "Mission_Outcome" like 'Success%') as successful_missions,
(select count("Mission_Outcome") from SPACEXTABLE where "Mission_Outcome" like 'Failure%') as unsuccessful_missions;

H. select “Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);

I. select substr(Date, 0, 5) as launch_year, substr(Date, 6, 2) as launch_month, “Booster_Version", “Launch_Site", “Landing_Outcome" from SPACEXTABLE where substr(Date, 0, 5) = ‘2015' and "Landing_Outcome" = 'Failure (drone ship)’;

J. select "Landing_Outcome", count(*) as outcome_count from SPACEXTABLE where Date between '2010-06-04' and ‘2017-03-20' group by “Landing_Outcome" order by outcome_count desc;

Thank you!