
IBM Developer
SKILLS NETWORK
Dimiter Georgiev
19 January 2024

Outline

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

Executive Summary

Methodologies:

- Data Collection with API, Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) with SQL, Pandas, Matplotlib
- Interactive Visual Analytics with Folium
- Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Results:

- EDA Results
- Predictions

Introduction

Background and context:

- SpaceY will target the commercial space travel sector with inexpensive rocket launches by building on SpaceX successes. In order to do this, SpaceY must be ableto predict if the Falcon 9 first stage will land successfully. SpaceX operating costs are much smaller than its competitors’ as a result of it reusing its rocket first stages. We investigate this technology.

Answers to problems:

- SpaceY must determine whether rockets can reuse their first stage, by determining the parameters for successful landings. Under which conditions do SpaceX rockets have the best landing success rate?

Section 1

METHODOLOGY

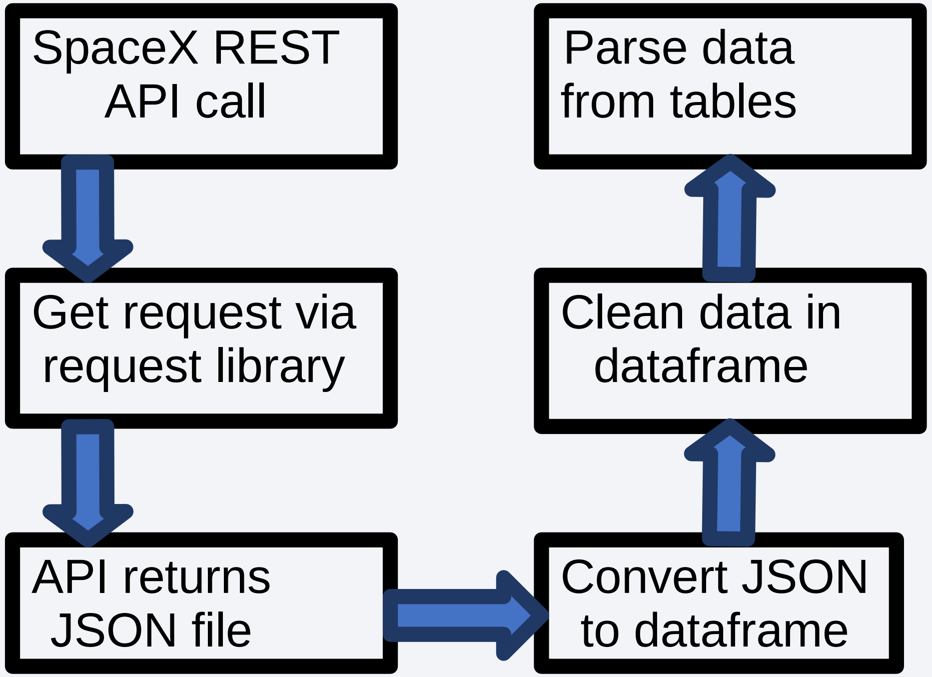
Methodology

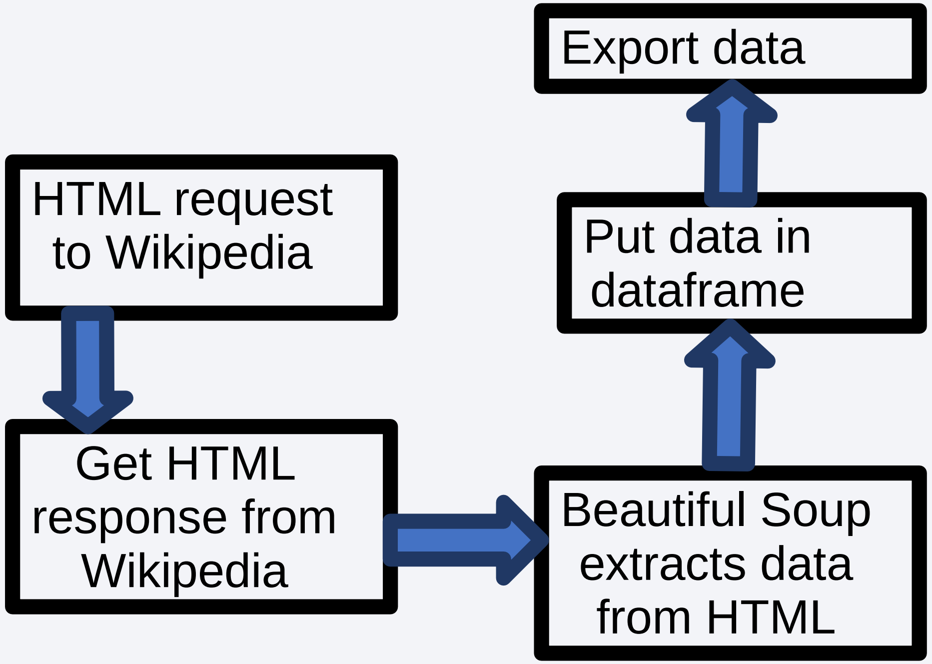
Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web scraping Wikipedia
- Perform data wrangling
  - Remove unnecessary data from JSON files
  - One Hot encoding classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Machine Learning to determine the best first stage parameters

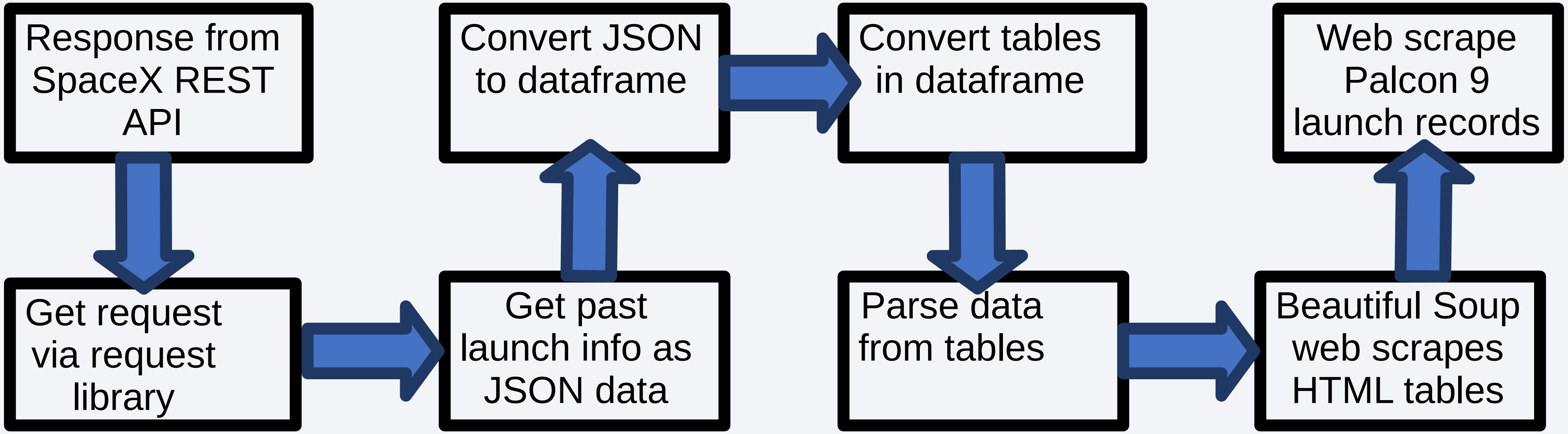
Data Collection

Data gathered from SpaceX REST API and web scraped from Wikipedia



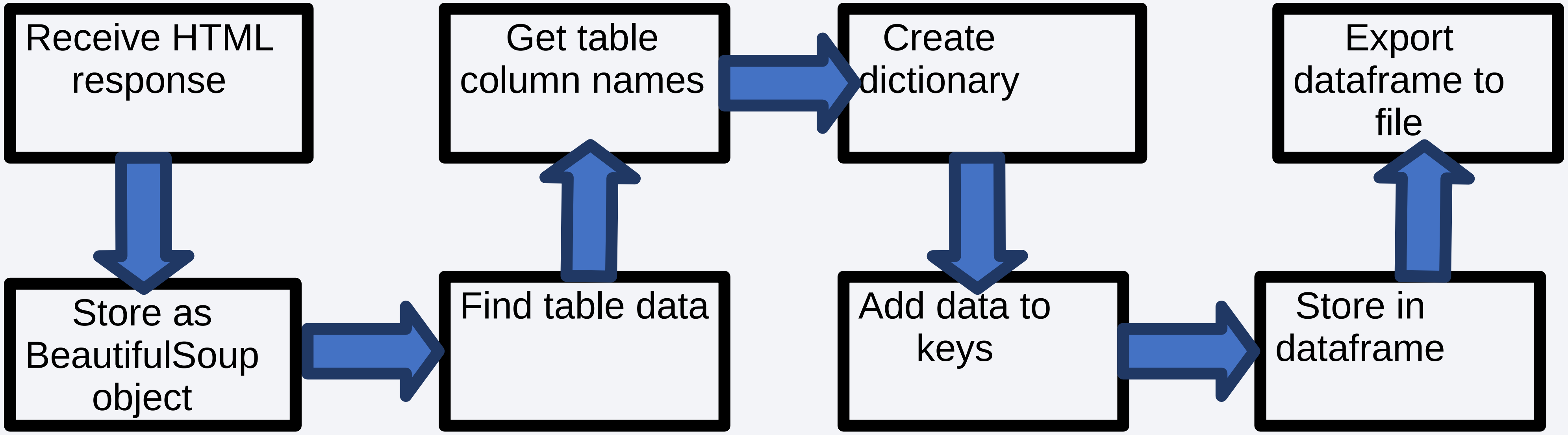


Data Collection – SpaceX API



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

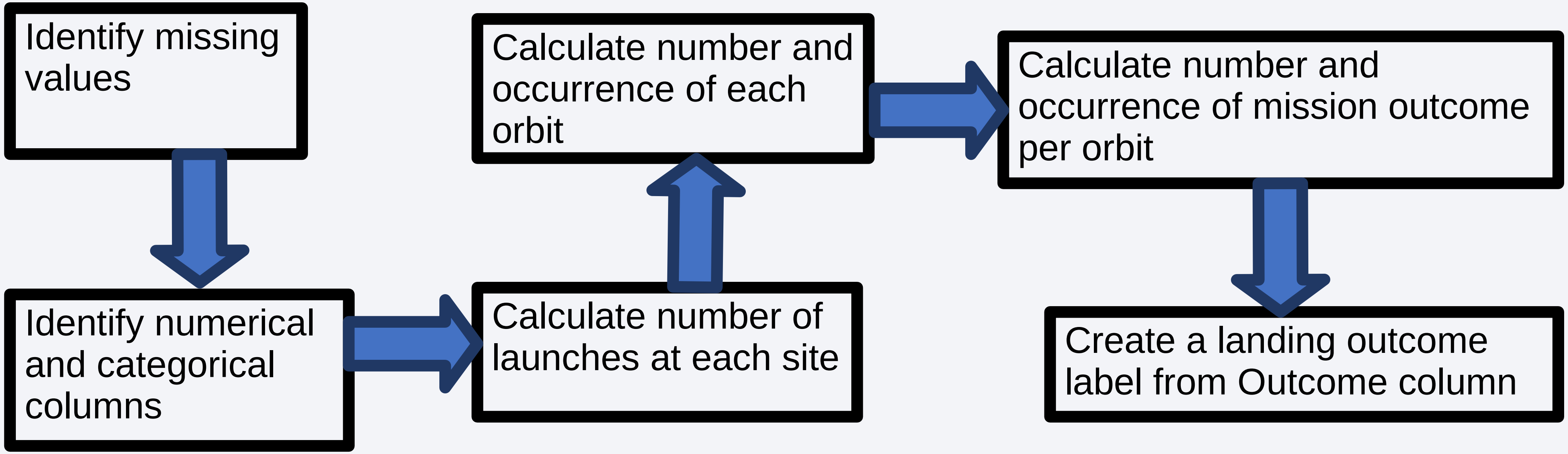
Data Collection - Scraping



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

Data Wrangling

Exploratory Data Analysis (EDA) was used to find patterns in the data and determine how to label trained supervised models



https://github.com/thisisthedgz/testrepo/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Summary of plotted charts:

- Catplot visualizes Flight Number and Payload relationship
- Catplot visualizes Flight Number and Launch Site relationship
- Catplot visualizes Payload and Launch Site relationship
- Bar Chart visualizes success rate of each Orbit type
- Catplot visualizes Flight Number and Orbit type relationship
- Catplot visualizes Payload and Orbit type relationship
- Line Chart visualize annual Launch Success trends

https://github.com/thisisthedgz/testrepo/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL queries performed to gather and understand data:

- Display names of unique launch sites in space mission
- Display 5 records where launch sites begin with string
- Display total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster F9 v1.1
- List date when first successful landing outcome in ground pad achieved
- List names of boosters which have success in drone ship and payload mass between 4000 and 6000
- List total number of successful and failed outcomes
- List names of booster_versions which have carried maximum payload mass
- List failed landing_outcomes in drone ship, their booster versions, and launch_site names for year 2015
- Rank 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

https://github.com/thisisthedgz/testrepo/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The map object is centered on NASA Johnson Space Center on a map, and populated with objects created to help understand the data:

- folium.Circle and folium.Marker to create highlighted circles with text label at specific coordinate for each launch site
- MarkerCluster object simplifies map containing many markers with same coordinate.
- MousePosition on the map gets coordinate when mousing over a point on the map.
- folium.PolyLine object shows distances between key locations by drawing a line.

https://github.com/thisisthedgz/testrepo/blob/main/lab_jupyter_launch_site_location.jupyterlite(1).ipynb

Build a Dashboard with Plotly Dash

Dashboard summarizes plots/graphs and interactivity is added to provide visual analytics on SpaceX launch data in real-time.

The dropdown list and range slider are input components to interact with a pie chart and scatter plot.

Callback function renders success-pie-chart based on a site selected in the dropdown menu. This retrieves the selected launch site from site-dropdown and renders the pie chart with launch successes.

A range slider selects payload mass. The Slider select different payload ranges so we can try to identify visual patterns.

The scatter chart shows the relationship between two variables, Success and Payload Mass.

https://github.com/thisisthedgz/testrepo/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Data preparation

- Load dataset
- Normalize data
- Split data into training and test sets.

Model preparation

- Select machine learning algorithms
- Set parameters for each algorithm to GridSearchCV
- Train GridSearchModel models with training dataset

Model evaluation

- Find best hyperparameters for each model
- Compute model accuracy with test dataset
- Plot Confusion Matrix

Model comparison

- Compare model accuracy
- Choose most accurate model

https://github.com/thisisthedgz/testrepo/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

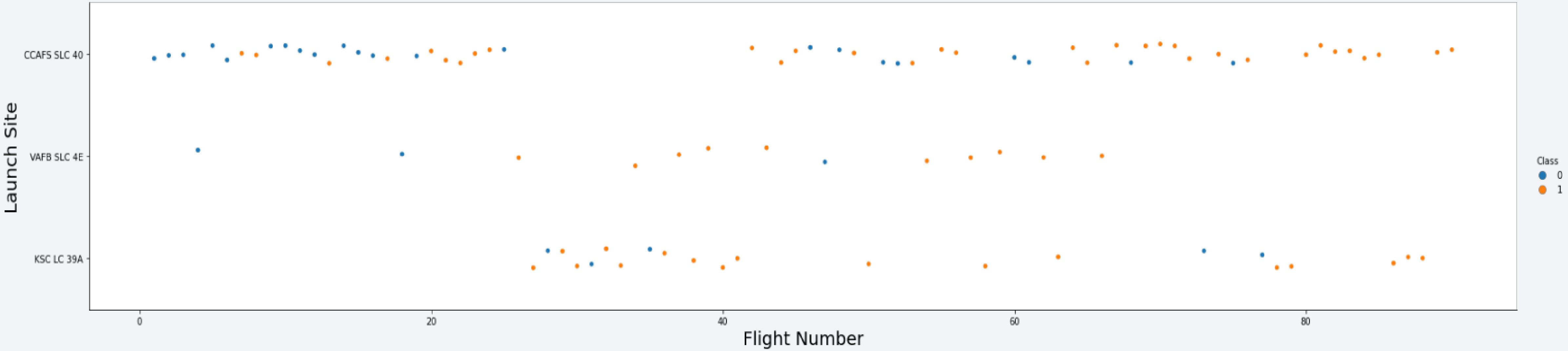
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

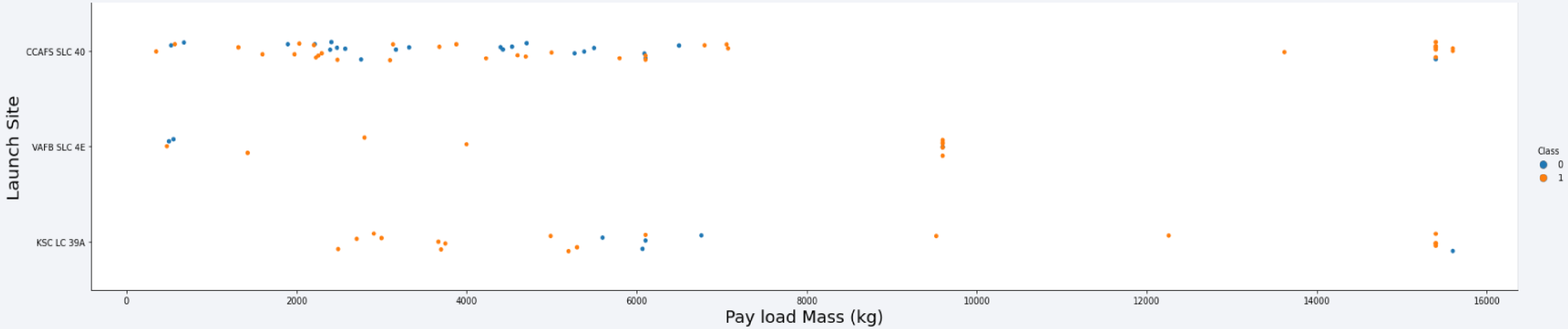
INSIGHTS DRAWN FROM EDA

Flight Number vs. Launch Site



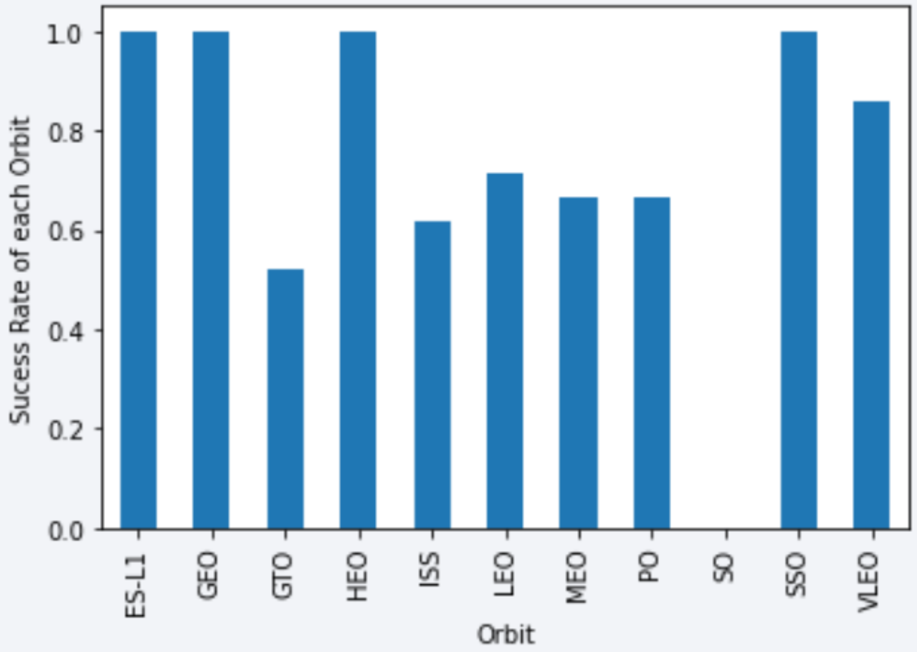
- The success rate for every launch site is increasing with time, especially for CCAFS SLC 40
- VAFB SLC 4E and KSC LC 39A have a higher success rate but represent one third of launches

Payload vs. Launch Site



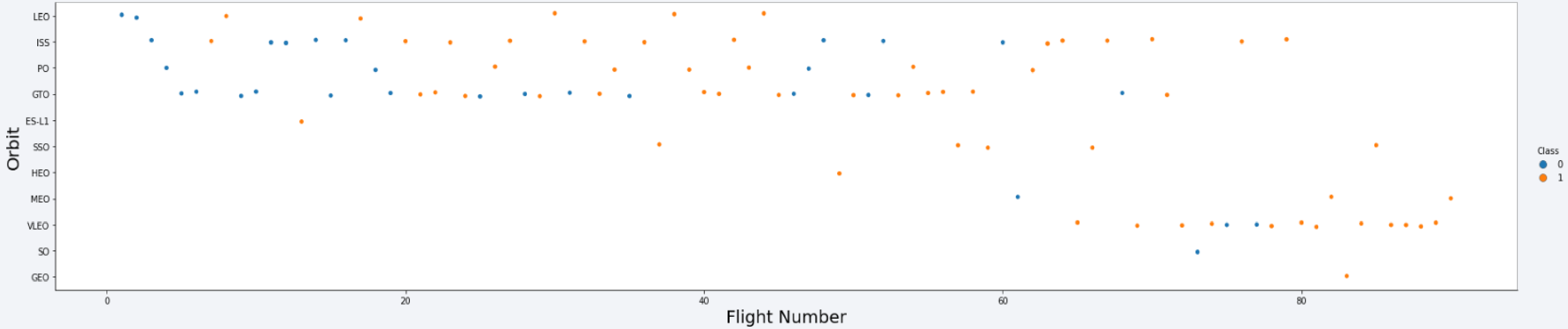
- In VAFB-SLC launch site there are no rockets launched for heavy payload mass (> 10 000kg)
- In KSC LC launch site there are no rockets launched for lower paylod mass (< 2500 kg)
- CCAFS SLC has only launched rockets with a payload mass < 7500 kg and > 13 000 kg

Success Rate vs. Orbit Type



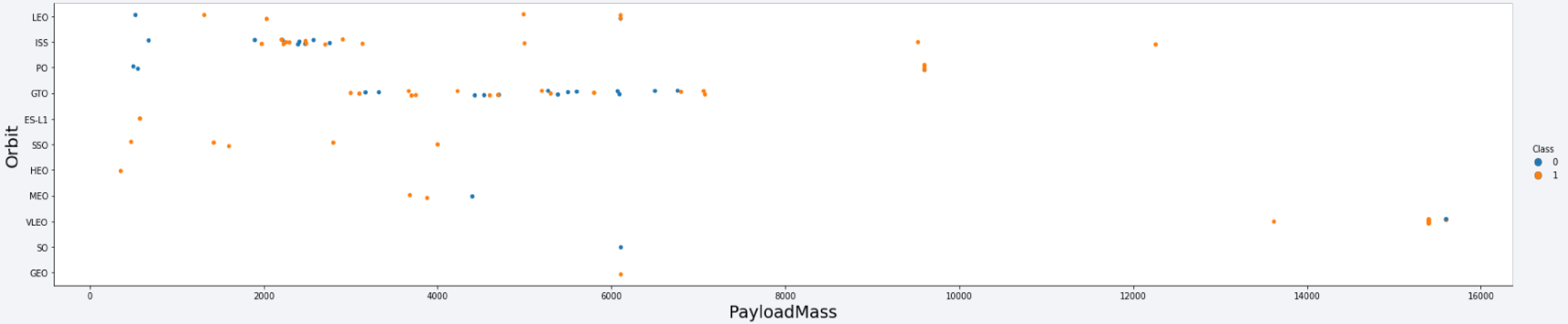
- The first 4 Orbit types have the best success rate. How many attempts are per Orbit type?
- Interpret the bar chart with the number of launches per orbit type

Flight Number vs. Orbit Type



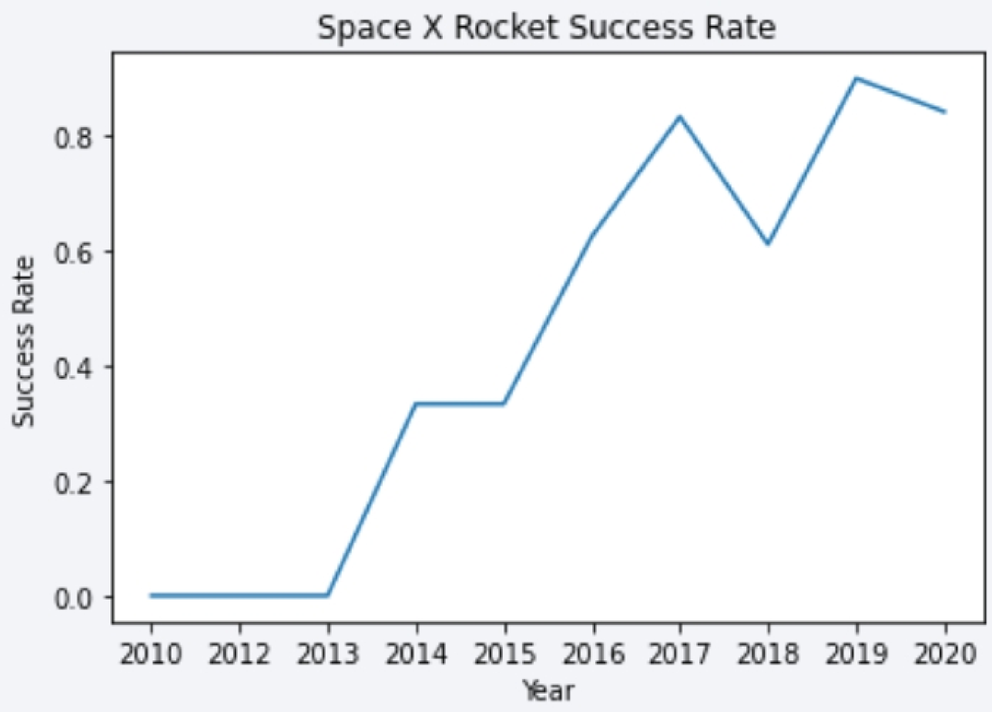
- There are more failures at the beginning of the series of launches but after the first 40 launches, the ratio improves by reducing the 50 percent of unsuccessful landings
- GTO and ISS orbits has the higher concentration of launches with the lowest ratio of successful landings
- The orbits with higher successful rate, has one or just a few number of launches

Payload vs. Orbit Type



- Exists a visible limit of payload around 7600 kg. Less than 10 launches exceed that limit.
- With heavy payloads, the successful landing rate is more for Polar, LEO and ISS
- But for GTO we cannot distinguish this well since both positive and negative landing are both there here

Launch Success Yearly Trend



- Show a line chart of yearly average success rate
- Explanation: the success rate improved throughout the 2010s, starting at 0.0% and reaching a high of 0.9% in 2019

All Launch Site Names

The four unique launch sites in the space mission are

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

SQL query: SELECT DISTINCT “LAUNCH_SITE” FROM SPACEXTBL

The keyword DISTINCT removes duplicate launch sites

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

SQL query: SELECT * FROM SPACEXTBL WHERE “LAUNCH SITE” LIKE ‘%CCA’ LIMIT 5

The WHERE and LIKE clauses together filter launch sites containing “CCA”. LIMIT 5 shows 5 records.

Total Payload Mass

SUM(“PAYLOAD_MASS__KG_”)

45596

SQL query: SELECT SUM(“PAYLOAD_MASS__KG_”) FROM SPACEXTBL WHERE “CUSTOMER” = ‘NASA (CRS)’

This query returns the sum of all payload masses of the customer NASA (CRS).

Average Payload Mass by F9 v1.1

AVG(“PAYLOAD_MASS__KG”)

2534.6666666666665

SQL query: SELECT AVG(“PAYLOAD_MASS__KG_”) FROM SPACEXTBL WHERE “BOOSTER_VERSION” LIKE ‘%F9 V1.1%’

Query returns average of all payload masses where booster version contains substring F9 v1.1.

First Successful Ground Landing Date

MIN(“DATE”)

01-05-2017

SQL query: SELECT MIN(“DATE”) FROM SPACEXTBL WHERE “Landing_Outcome” LIKE ‘%Success%’

WHERE clause limits dataset to records where landing succeeded, and MIN function selects the oldest recorded date.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

SQL query: %sql SELECT “BOOSTER_VERSION” FROM SPACEXTBL WHERE “LANDING_OUTCOME ” = ‘Success (drone ship)’ AND “PAYLOAD_MASS__KG_” > 4000 AND “PAYLOAD_MASS__KG_” < 6000;

WHERE and AND clauses filter the dataset to return booster version where landing succeeded and payload mass is between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

Success: 100

Failure: 1

SQL query: %sql SELECT (SELECT COUNT(“MISSION_OUTCOME”) FROM SPACEXTBL WHERE “MISSION_OUTCOME” LIKE ‘%Success%’) AS SUCCESS, (SELECT COUNT(“MISSION_OUTCOME”) FROM SPACEXTBL WHERE “MISSION_OUTCOME” LIKE ‘%Failure%’) AS FAILURE

The first SELECT displays the subqueries that return results. The first subquery counts successful missions, and the second subquery counts the failed missions. The WHERE and LIKE clauses filter the mission outcome, and COUNT counts the records filtered.

Boosters Carried Maximum Payload

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

SQL query: %sql SELECT DISTINCT “BOOSTER_VERSION” FROM SPACEXTBL WHERE “PAYLOAD_MASS__KG_” = (SELECT max(“PAYLOAD_MASS__KG_”) FROM SPACEXTBL)

MAX functions filters data to return only heaviest payload mass with. Main query uses subquery results and returns unique booster version (SELECT DISTINCT) with heaviest payload mass.

2015 Launch Records

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

SQL query: %sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL WHERE "LANDING_OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'

Query returns 2015 failed landings with month, booster version, and launch site. Substr function processes date to take month or year. Substr(DATE, 4, 2) shows month and Substr(DATE,7, 4) shows year.

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

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

SQL query: %sql “”LANDING_OUTCOME”, COUNT(“LANDING_OUTCOME”) FROM SPACEXTBL WHERE “DATE” >= ‘04-06-2010’ and “DATE”<= ‘20-03-2017’ and “LANDING_OUTCOME” LIKE ‘%Success%’ GROUP BY “LANDING_OUTCOME” ORDER BY COUNT(“LANDING_OUTCOME”) DESC;

Query returns landing outcomes and their count where mission succeeded and date is from 2010/06/04 and 2017/03/20. GROUP BY clause groups results by landing outcome and ORDER BY COUNT DESC shows results in decending order.

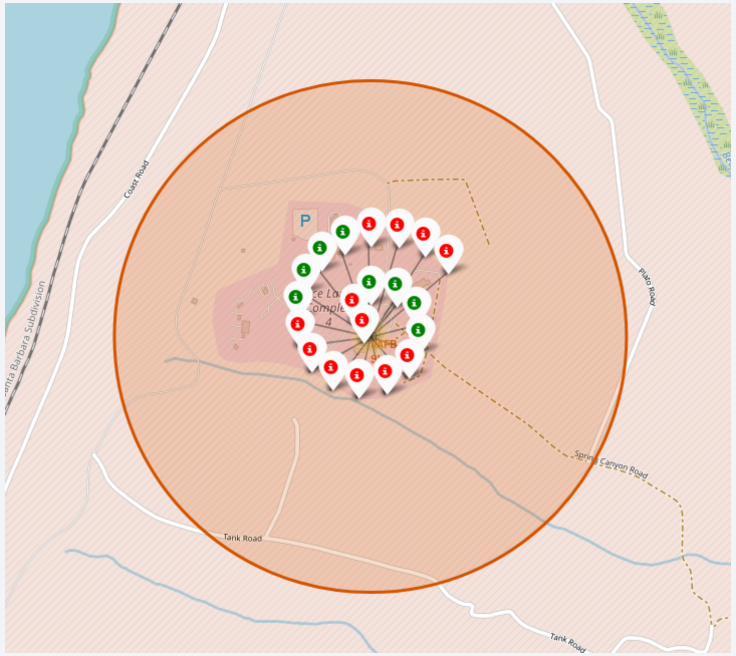
Section 3

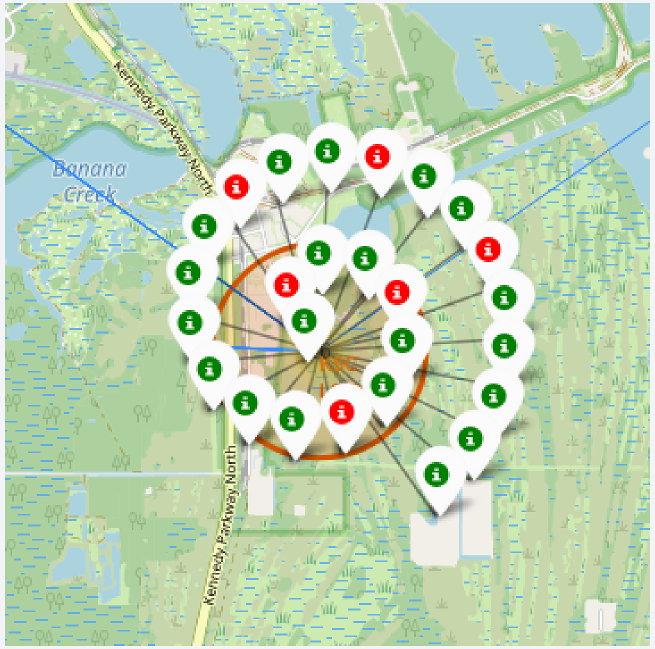
LAUNCH SITES PROXIMITIES ANALYSIS

All Launch Sites

All launch sites are in very close proximity to the coast and intro restricted areas.

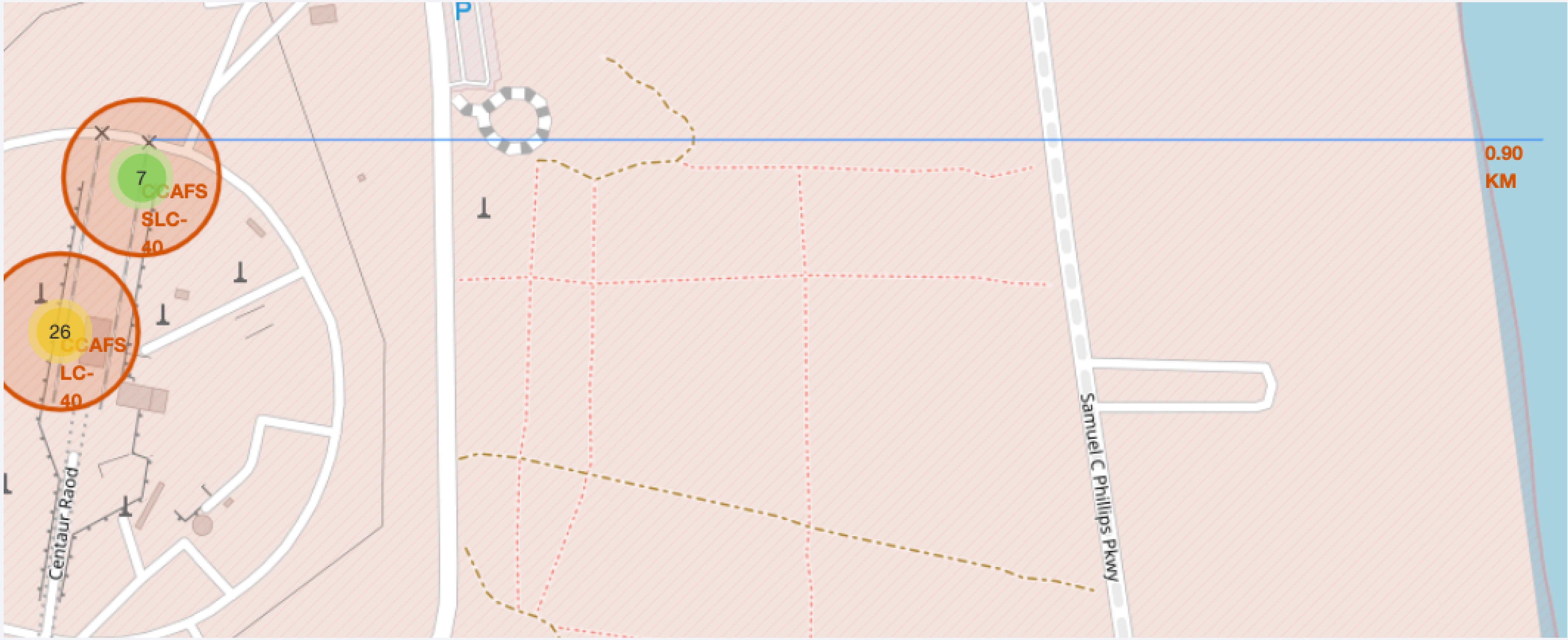
Successful/Failed Launches for Each Site





The green markers show successful launches, and the red markers show failed launches.

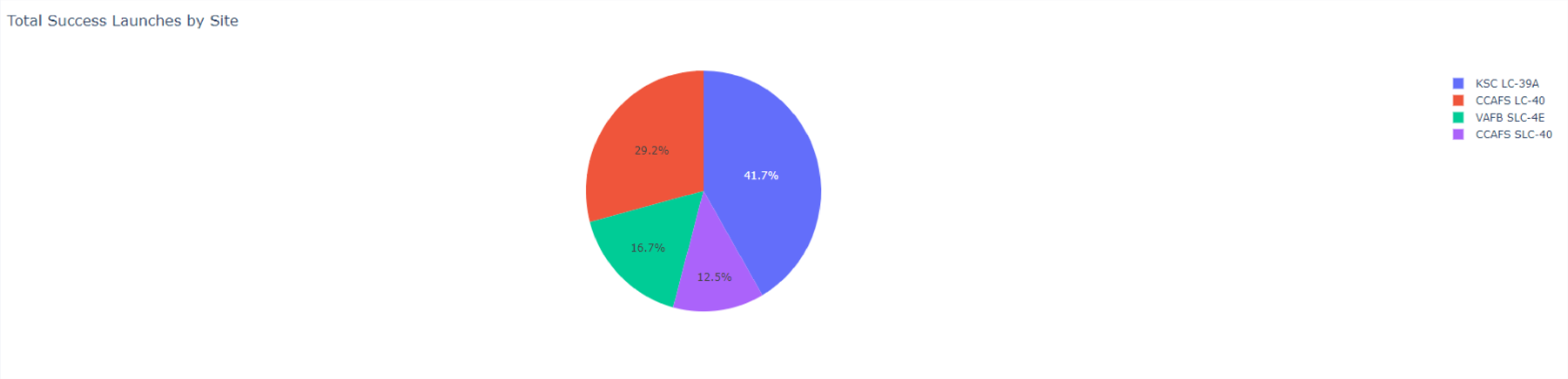
A Launch Site and Its Proximities



Section 4

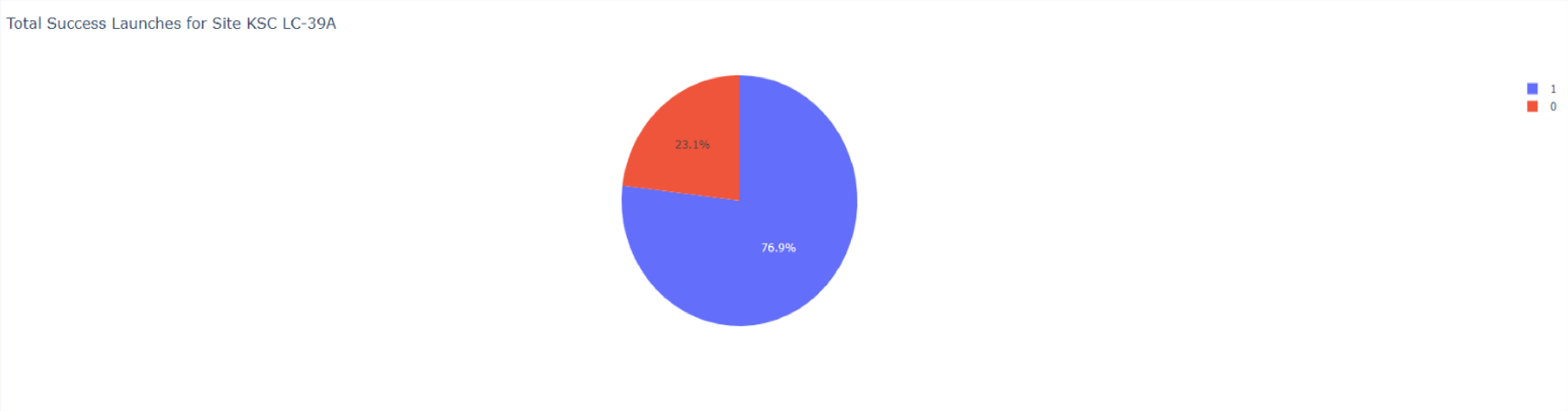
BUILD A DASHBOAD WITH PLOTLY DASH

Total Success Launches By Site



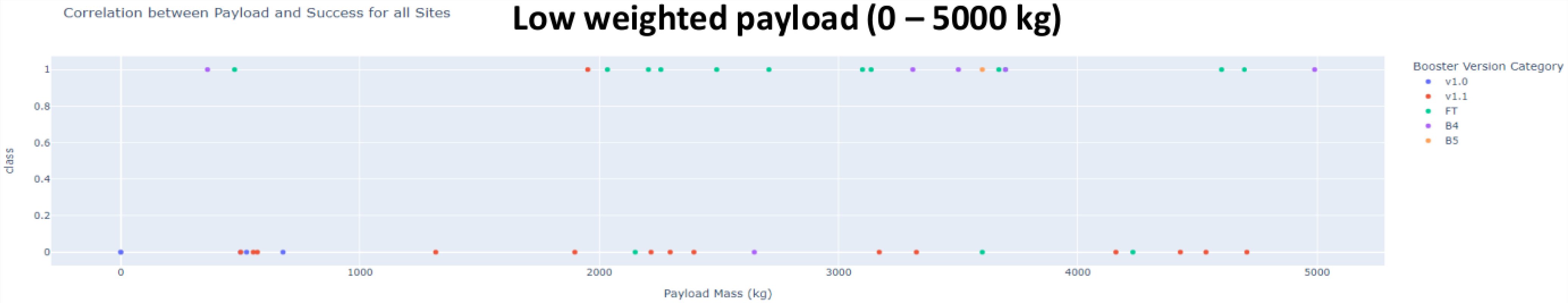
KSC LC-39A is the site with the most successes, and CCAFS LC-40 has the second-most successes

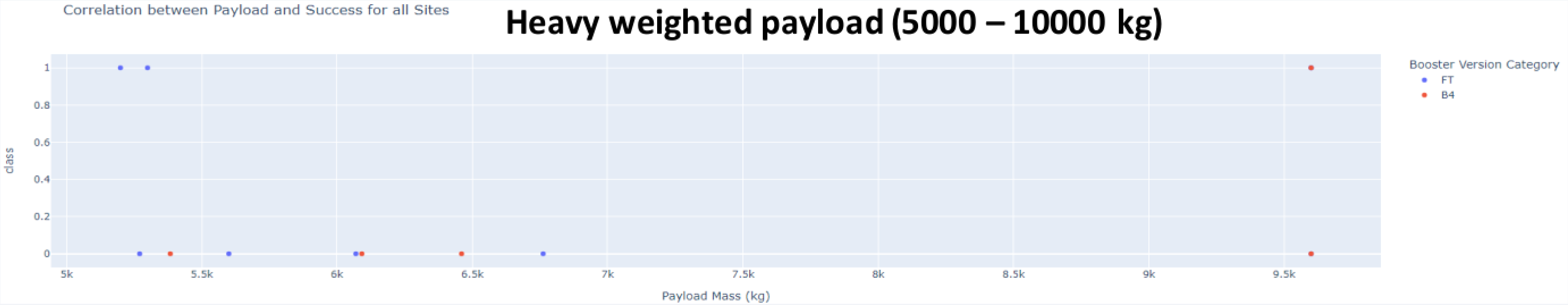
KSC LC-39A



KSC LC-39A’s pie chart shows it is the site with the highest launch success rate.

Payload vs. Launch Outcome





Scatter plots for all sites with up to 5000 kg, and between 5000 and 10 000 kg payload ranges

The majority of successes are at lower payloads.

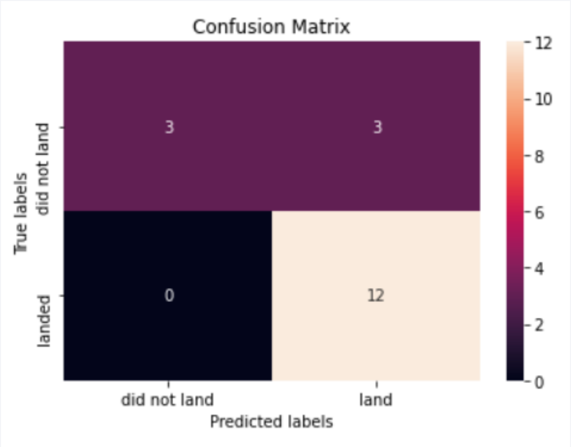
Section 5

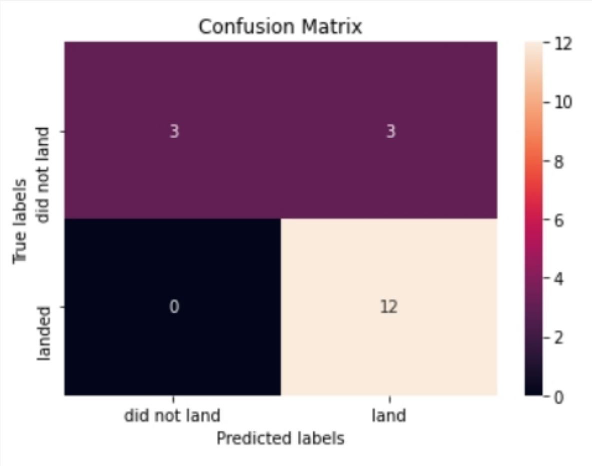
PREDICTIVE ANALYSIS (CLASSIFICATION)

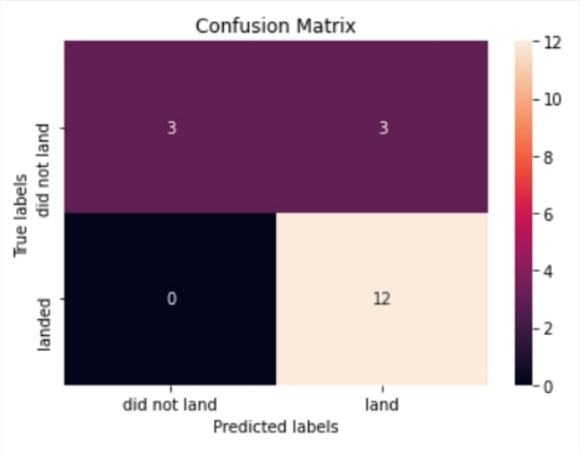
Classification Accuracy

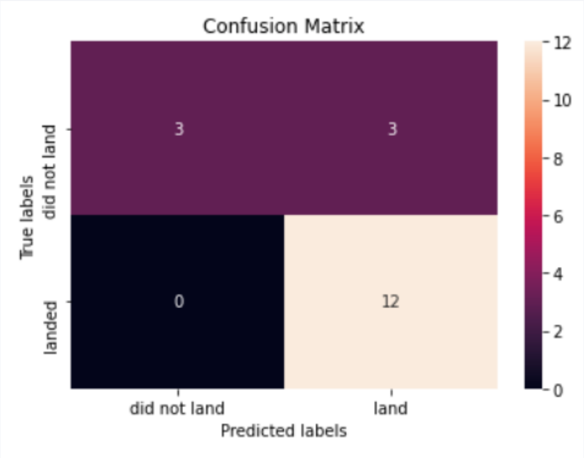
All of the models have similar accuracy levels.

Confusion Matrix









Top left: Logistic regression; Top Right: Decision tree; Bottom left: kNN; Bottom right: SVM.

As seen, the confusion matrix is the same for all models.

Conclusions

Mission successes or failures are affected by a multitude of factors such as the launch site, the orbit and the number of previous launches, indicating an increase in experience and knowledge for SpaceX launches.

The orbits that have the best success are GEO, HEO, SSO, ES-L1., and lower payloads also increases the chances of success.

Since all of the algorithms have similar accuracy, it doesn’t matter which we use presently. The machine learning model can predict if the fist state of the rocket will land successfully or not.

SpaceY can use this data to become competitive.

Appendix

All Capstone material, including the code for all of the projects, contained here:

https://github.com/thisisthedgz/testrepo/tree/main


SEE YOU LATER,
SPACE COWBOY