IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tony Nguyen
12/01/23

Outline

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

Executive Summary

- Collected data using both the SpaceX API and web-scraping off Wikipedia and cleaned it
- Performed EDA using SQL and charts
- Performed IDA using Folium and Plotly Dash
- Fed data into multiple classification models

- Results show:
  - SpaceX landing success rate improved between 2010 to 2020.
  - Most missions were successful regardless of landing
  - Most successful launches were held at Kennedy Space Center Launch Complex

Introduction

The cost of space travel is going down, and commercial space flight is becoming more and more viable for the general public.

SpaceX is one example of a company that has used innovation to reduce the cost of launching a rocket by landing and then reusing the initial stage of the rocket.

Thus, we wish to analyze SpaceX’s past rocket launches and determine if there’s any pattern and characteristics which would lead to a successful landing of its 1st stage rocket.

Section 1

Methodology

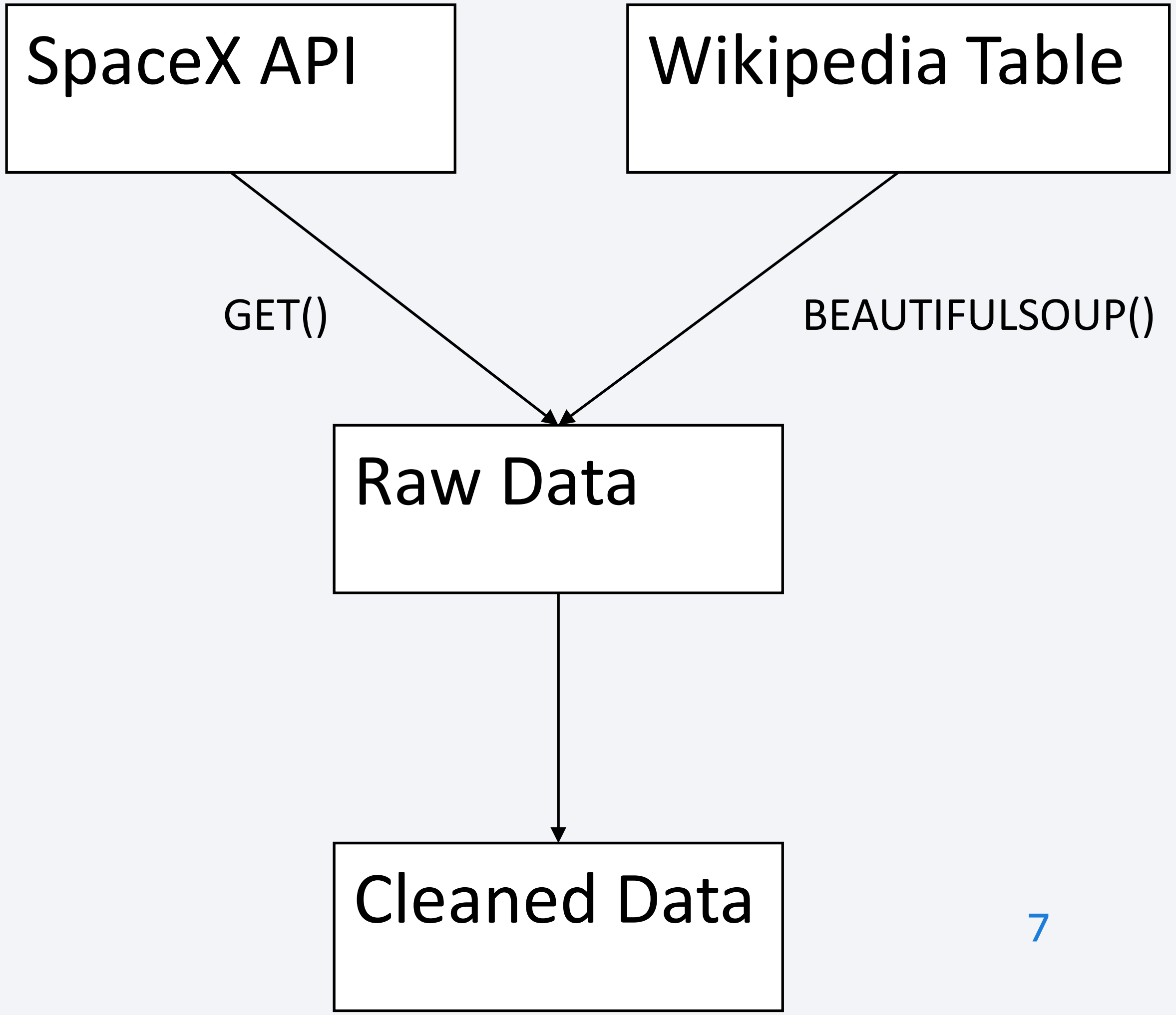
Methodology

Executive Summary

- Data collection methodology:
  - Data collected using SpaceX API with a GET request
  - Web-scrapped launch data using BeautifulSoup
- Perform data wrangling
  - Resulting table filtered by booster version; null values replace with their column mean
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardized the data, split the data into training and testing sets before feeding it into multiple predictive machine learning models and comparing accuracies.

Data Collection

- Collected data using a GET() request from the SpaceX API and a web-scrapped table from Wikipedia
- Cleaned data by filtering out irrelevant columns and replaced missing values with their variable's mean as an approximation.



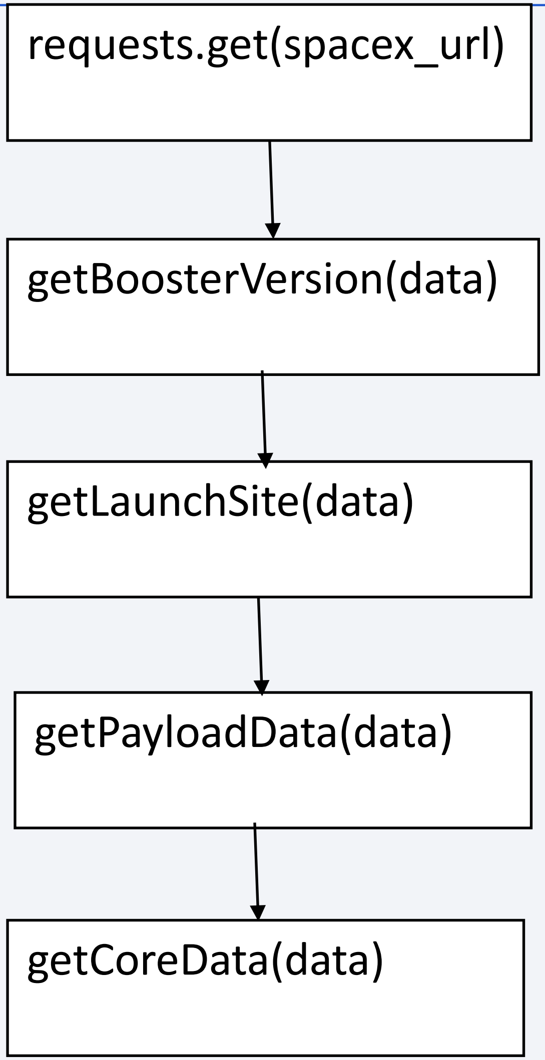
Data Collection – SpaceX API

- Collected raw data from API using the GET request

requests.get(spacex_url)

- Used API again to replace ID values in table with specified data

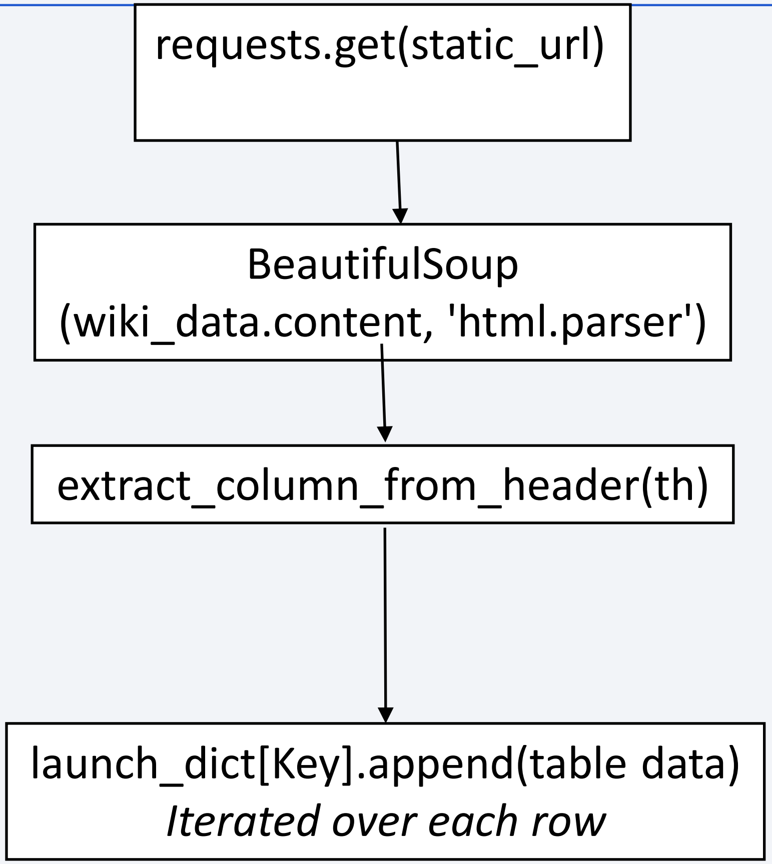
e.g. getBoosterVersion(data)



https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/Complete%20the%20Data%20Collection%20API%20Lab_v2.ipynb

Data Collection - Scraping

- HTTP GET method to request the Falcon9 Launch HTML page
- Extracted column names from the HTML table header using BeautifulSoup
- Created dataframe from extracted row data



requests.get(static_url)

BeautifulSoup (wiki_data.content, 'html.parser')

extract_column_from_header(th)

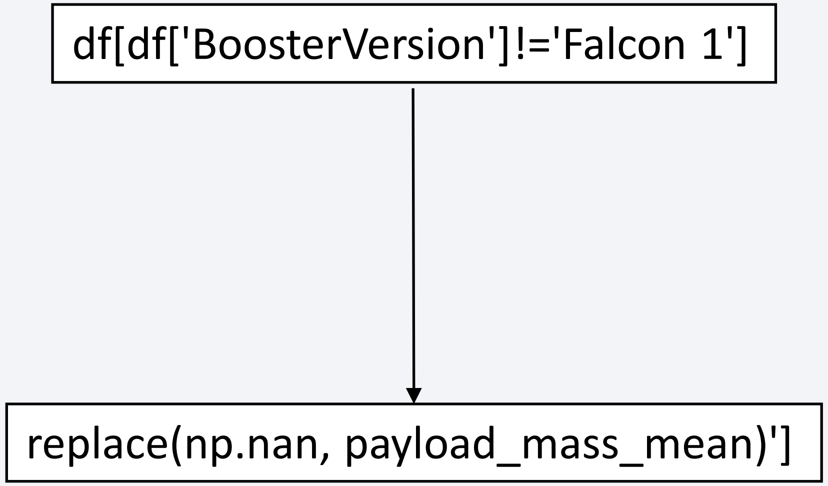
launch_dict[Key].append(table data)
Iterated over each row

https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/Data%20Collection%20with%20Web%20Scraping%20Lab.ipynb

Data Wrangling

- Filtered database to include only Falcon 9 launches

- Used replace() to replace Null values with the PayloadMass mean



df[df['BoosterVersion']!='Falcon 1']

replace(np.nan, payload_mass_mean)']

https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/Complete%20the%20Data%20Collection%20API%20Lab_v2.ipynb

https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/Data%20Collection%20with%20Web%20Scraping%20Lab.ipynb

EDA with Data Visualization

- Charted used include:
  - Scatterplot - FlightNumber vs. PayloadMass
  - Scatterplot - FlightNumber vs LaunchSite
  - Scatterplot - FlightNumber vs Orbit
  - Scatterplot - Orbit vs Payload Mass
  - Line chart – Year vs Success Rate
  - Bar chart – Orbit vs Success Rate

- Doing so allows us to compare how rocket launches change over time, especially their success rate

https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/EDA%20with%20Data%20Viz%20Lab.ipynb

EDA with SQL

- Listed unique launch sites used
- Calculated total and average payload mass sent
- Found boosters with successful drone ship landings
- Listed booster versions that carried maximum payloads
- Counted successful and failed landings

https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/EDA%20SQL%20Lab.ipynb

Build an Interactive Map with Folium

- Marked all launch sites and all successful/failed launches for each one on a world map.

- Doing so we can compare which launch site was used most often and which provided the most successful launches

https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/Data%20Viz%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- A pie chart to compare the proportion of successful to failed launches
- A scatterplot to compare successful launches at different payload masses within a specific range
- Both can information for all launches or a specific launch site

https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Standardized table data
- Split both target variable and predictor variables into training and testing sets
- Used GridSearchCV on:
  - Logistic Regression
  - SVM
  - Decision Tree Classifier
  - K-nearest neighbours
- Determined best classification model

https://github.com/Magician960/IBM-DS-Capstone-Project/blob/main/Machine%20Learning%20Prediction%20Lab.ipynb

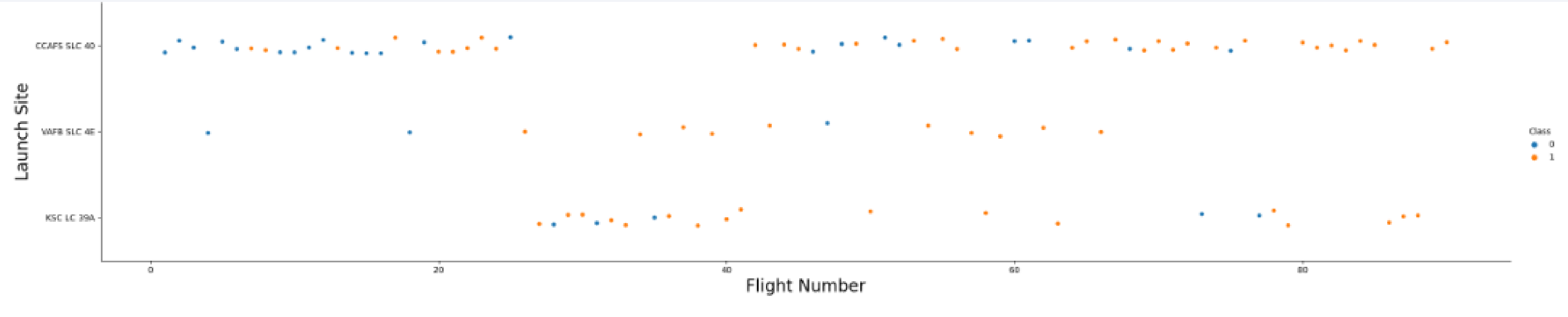
Results

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

Section 2

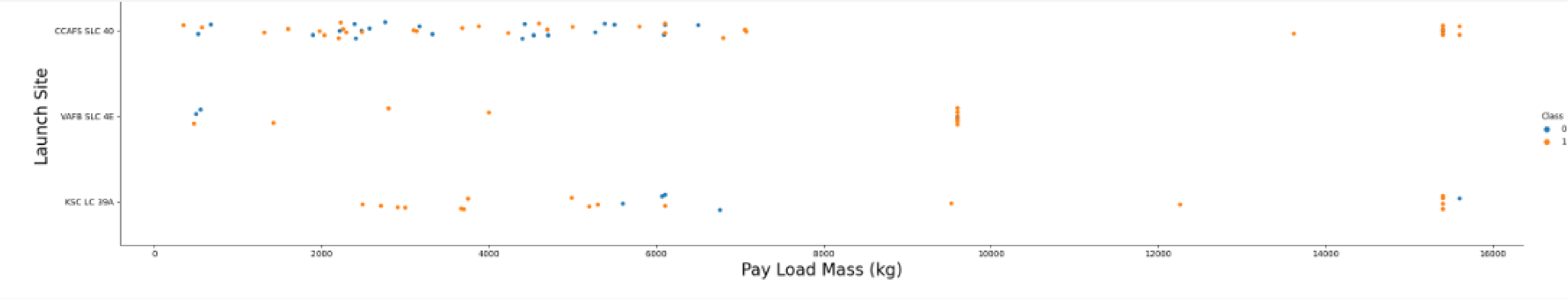
Insights drawn from EDA

Flight Number vs. Launch Site



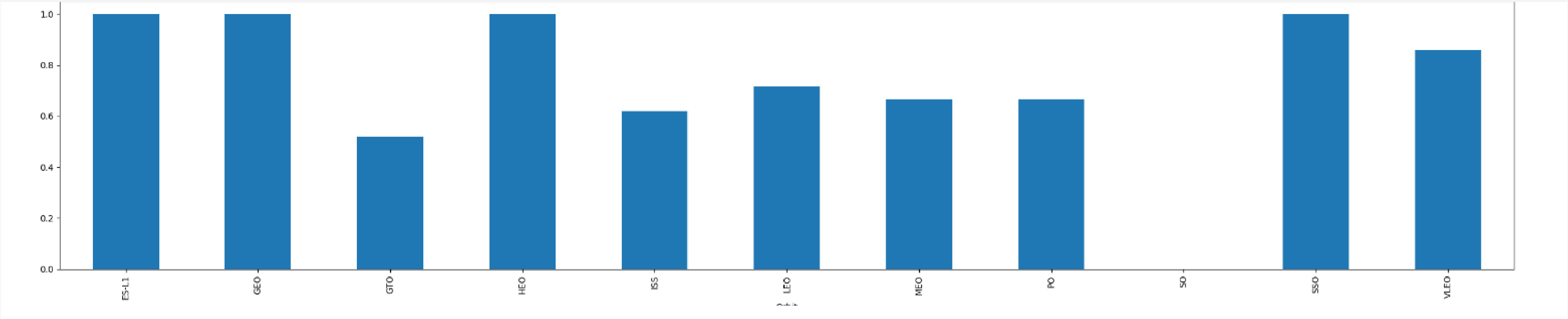
- Flight Number on the x-axis plotted against Launch Site on the y-axis.
- Successful landings are in orange (Class = 1), unsuccessful in green (Class = 0)

Payload vs. Launch Site



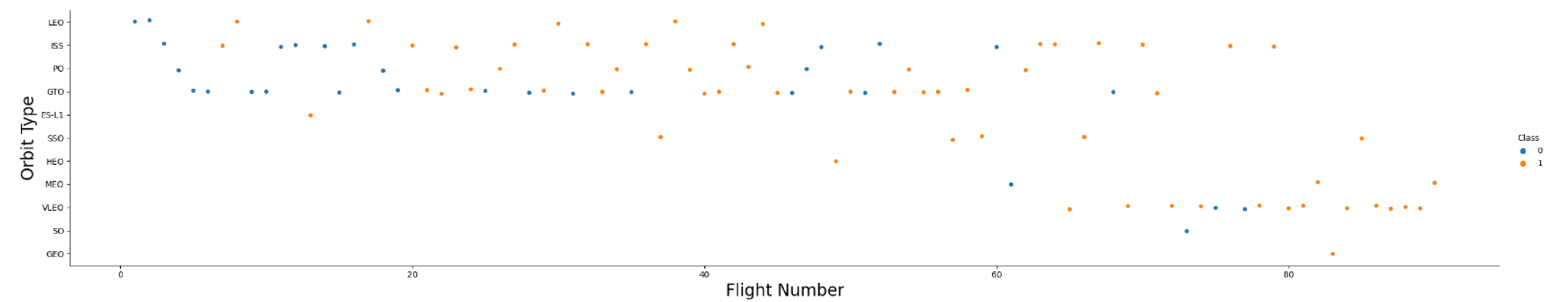
- Payload mass (kg) on the x-axis plotted against Launch Site on the y-axis.
- Successful landings are in orange (Class = 1), unsuccessful in green (Class = 0)

Success Rate vs. Orbit Type



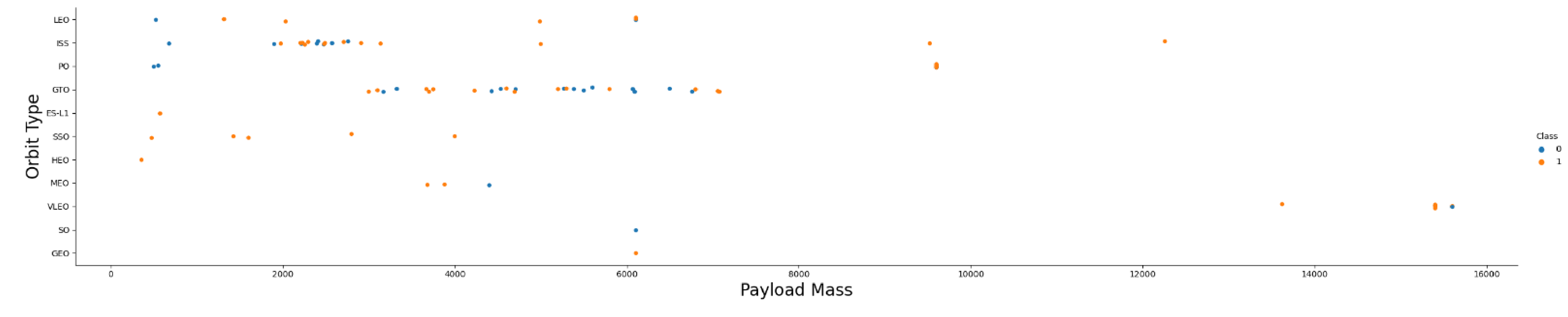
- Type of orbit of the rocket on the x-axis with the average success rate of landing on the y-axis.

Flight Number vs. Orbit Type



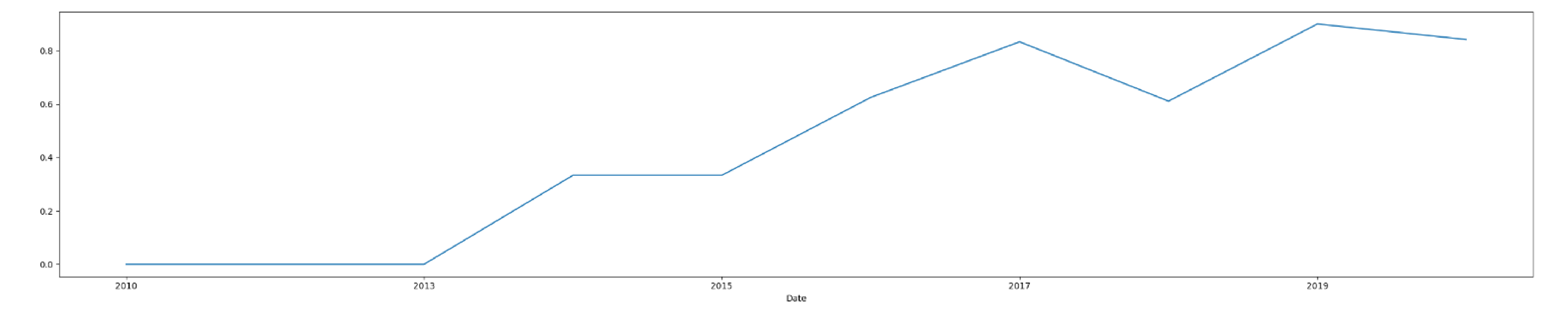
- Flight Number on the x-axis plotted against orbit type on the y-axis.
- Successful landings are in orange (Class = 1), unsuccessful in green (Class = 0)

Payload vs. Orbit Type



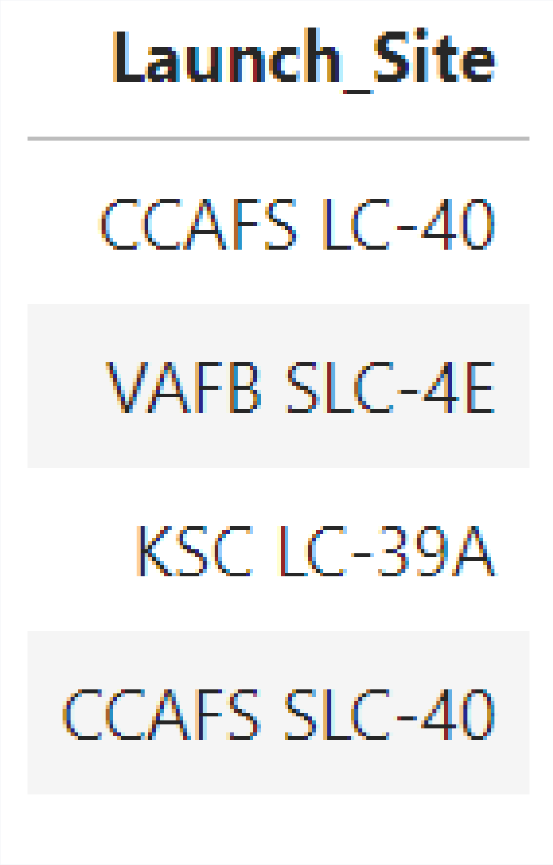
- Payload mass on the x-axis plotted against orbit type on the y-axis.
- Successful landings are in orange (Class = 1), unsuccessful in green (Class = 0)

Launch Success Yearly Trend



- Line plot charting success rate of landing from 2013 to 2020.

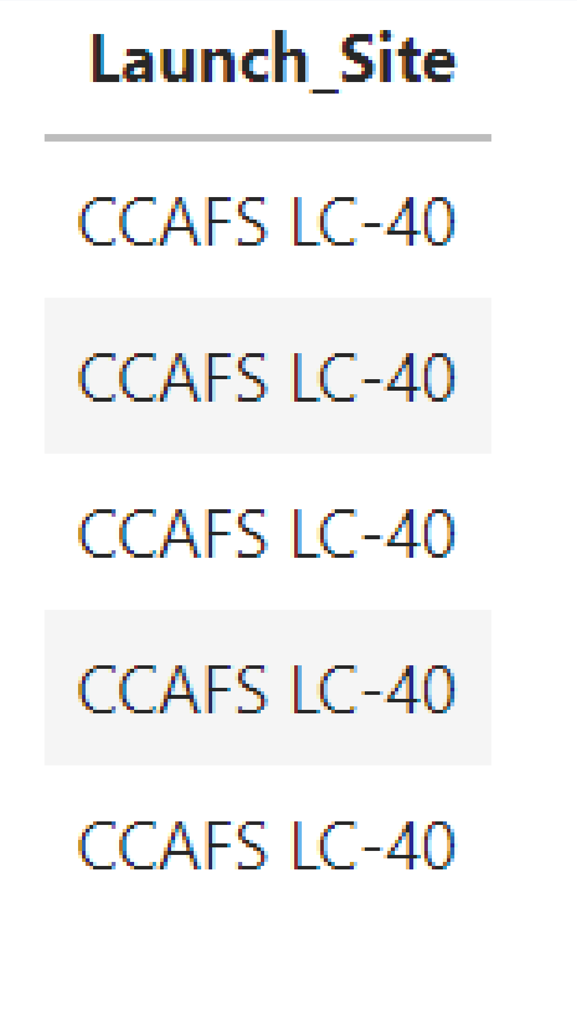
All Launch Site Names



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

- SpaceX uses 4 launch sites which are:
- CCAFS LC-40 - Cape Canaveral Launch Complex 40
- VAFB SLC-4E - Vandenberg Space Launch Complex 4
- KSC LC-39A - Kennedy Space Center Launch Complex 39
- CCAFS SLC-40 - Cape Canaveral Space Launch Complex 40

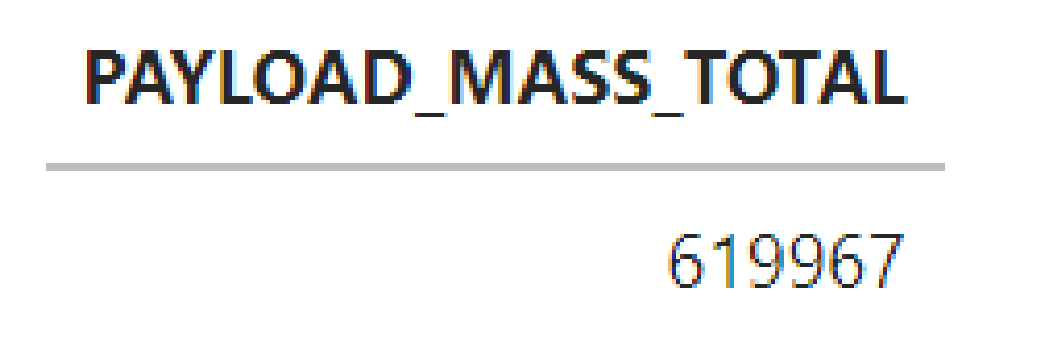
Launch Site Names Begin with 'CCA'



Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- Used “like 'CCA%’” in an SQL query to filter records with launch site names beginning with ‘CCA’.

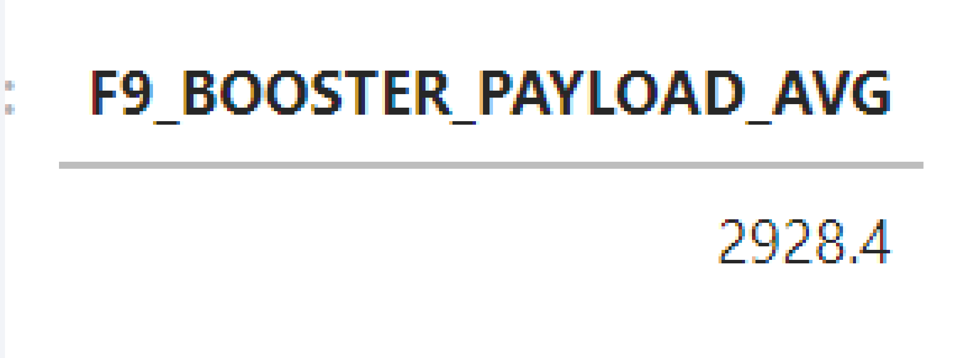
Total Payload Mass



PAYLOAD_MASS_TOTAL
619967

- Used SUM() in an SQL query to calculate the total amount of payload taken into orbit by SpaceX (in kg).

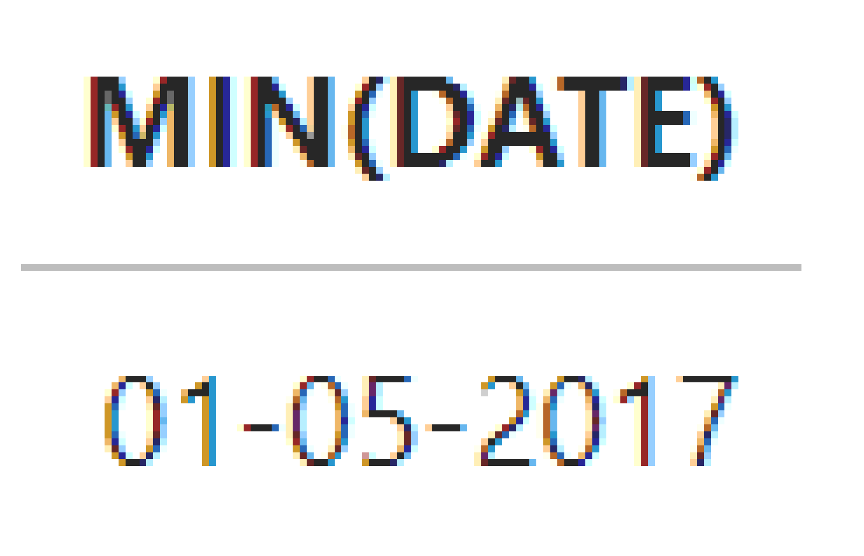
Average Payload Mass by F9 v1.1



F9_BOOSTER_PAYLOAD_AVG
2928.4

- Used AVG() function in an SQL query to calculate the average payload mass launched by the F9 v1.1

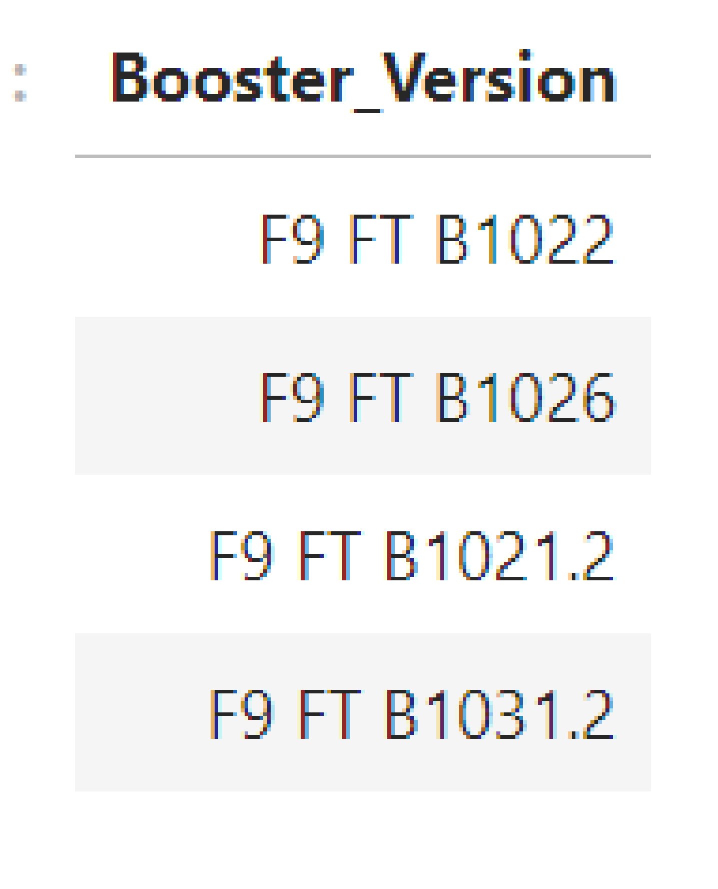
First Successful Ground Landing Date



MIN(DATE)
01-05-2017

- Used MIN() function in an SQL query to select the first successful ground landing and its date

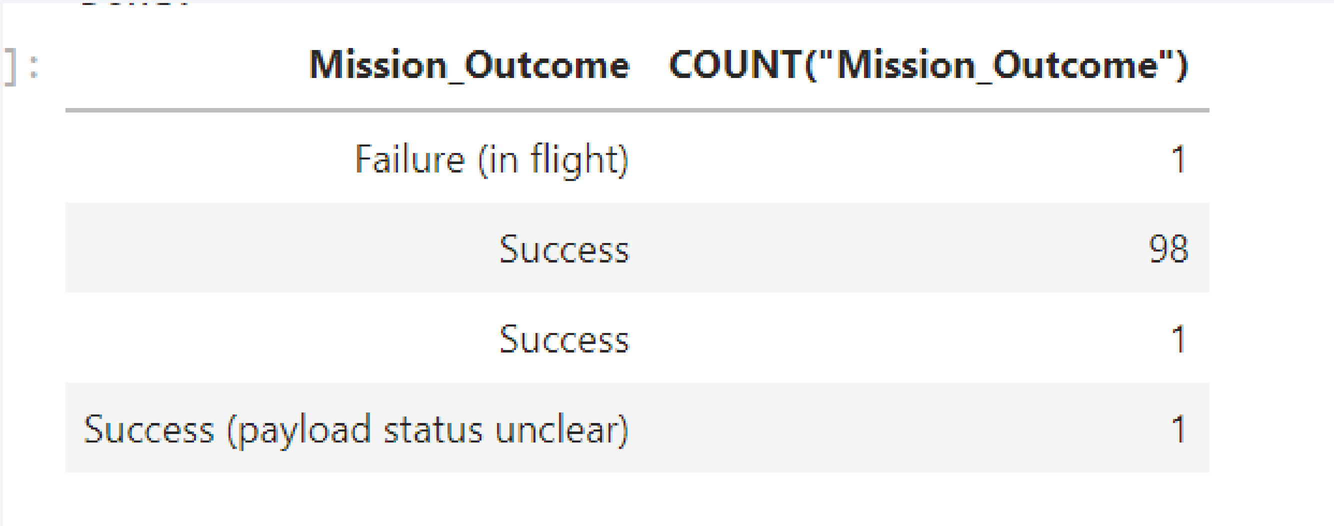
Successful Drone Ship Landing with Payload between 4000 and 6000



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

- Used an SQL query to select which boosters have successfully landed on a drone ship with a payload mass between 4000-6000 kg

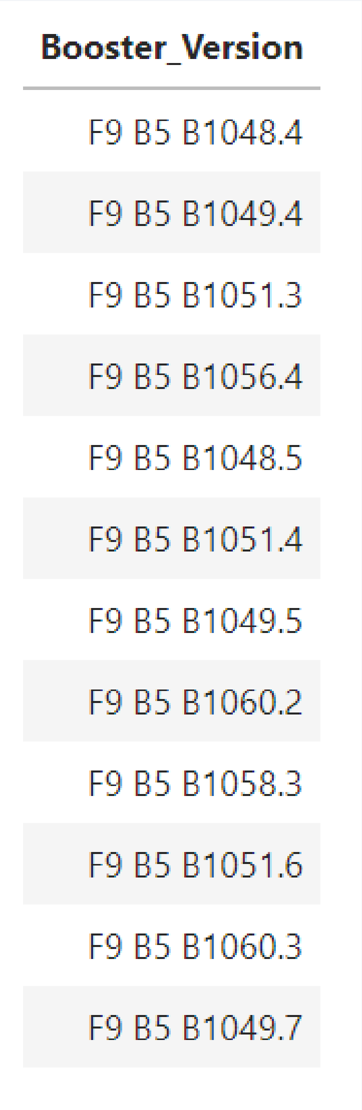
Total Number of Successful and Failure Mission Outcomes



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

- Used COUNT() function in an SQL query to sum total mission based on mission outcomes

Boosters Carried Maximum Payload

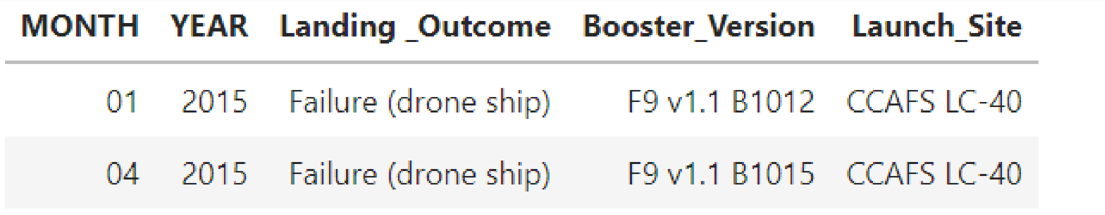


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- Used MAX() within a subquery of an SQL query to select all booster versions that carried the maximum possible payload mass

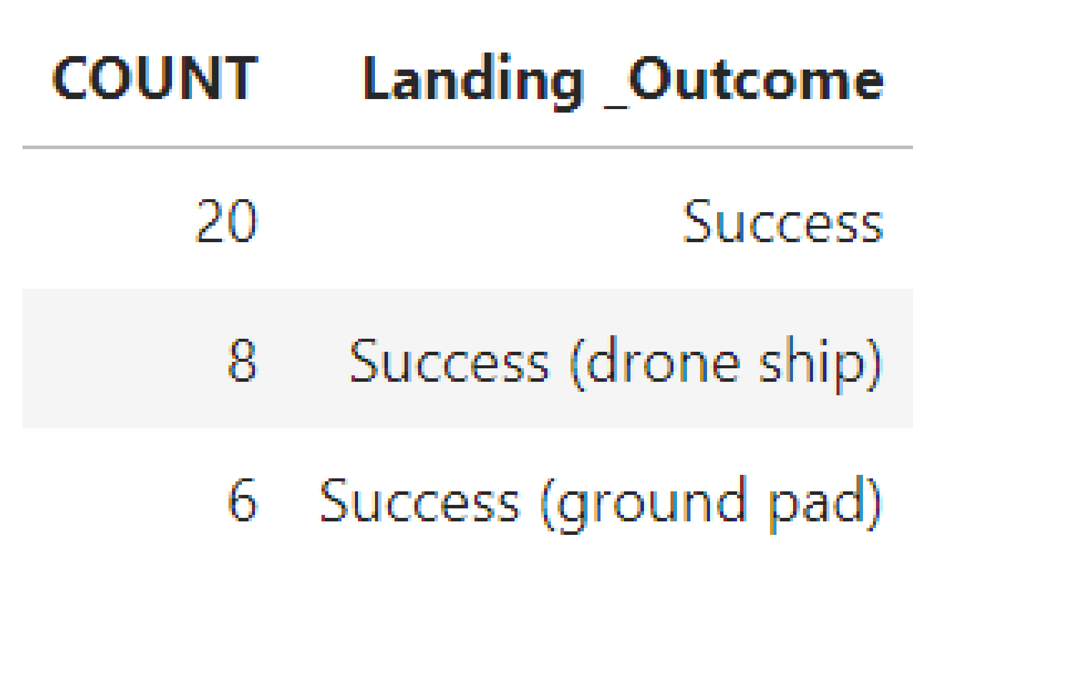
2015 Launch Records

- Used an SQL query to select all missions in 2015 that had a failed landing outcome



MONTH	YEAR	Landing _Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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



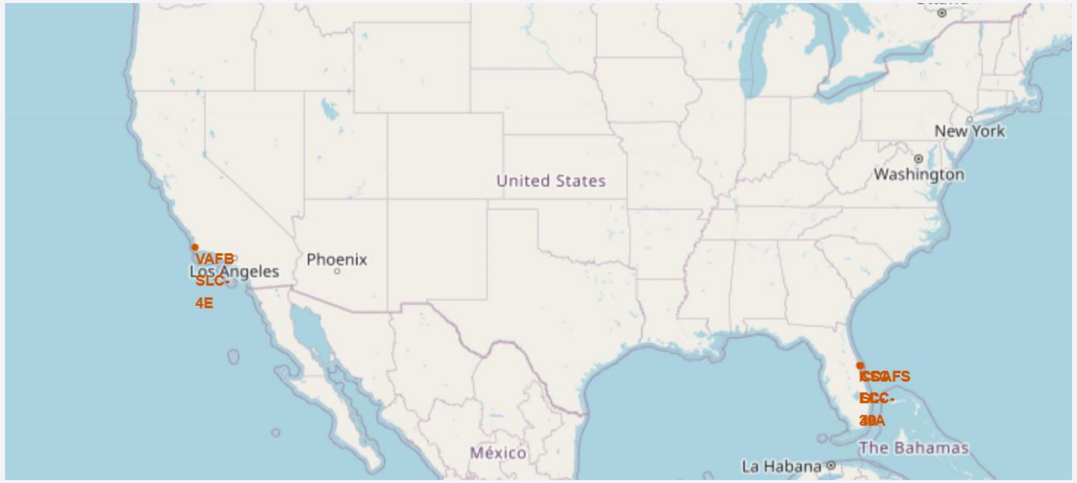
COUNT	Landing _Outcome
20	Success
8	Success (drone ship)
6	Success (ground pad)

- Used an SQL query to count all missions with successful landing outcomes between the dates of 2010-06-04 and 2017-03-20

Section 3

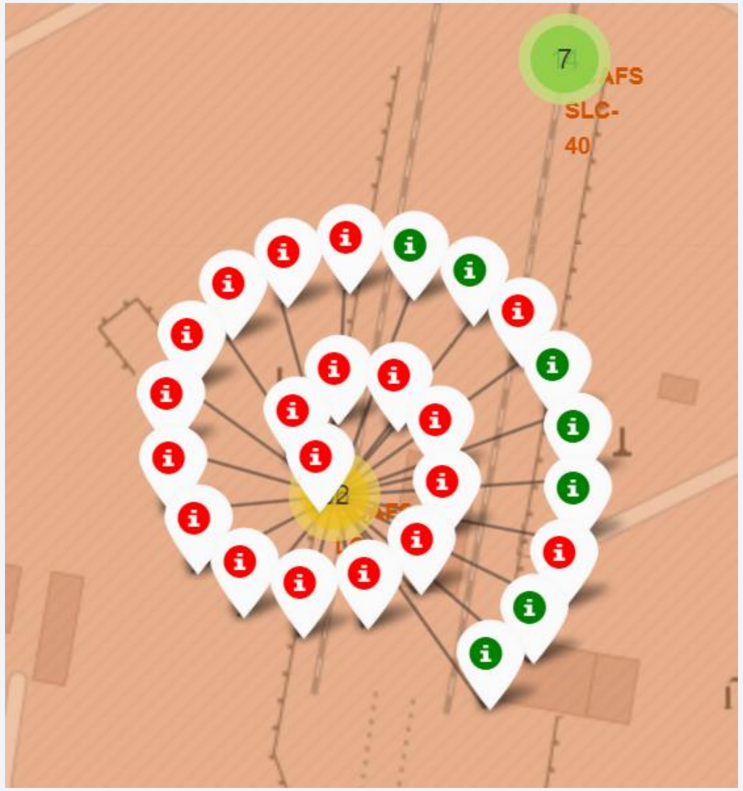
Launch Sites Proximities Analysis

Launch Sites used by SpaceX



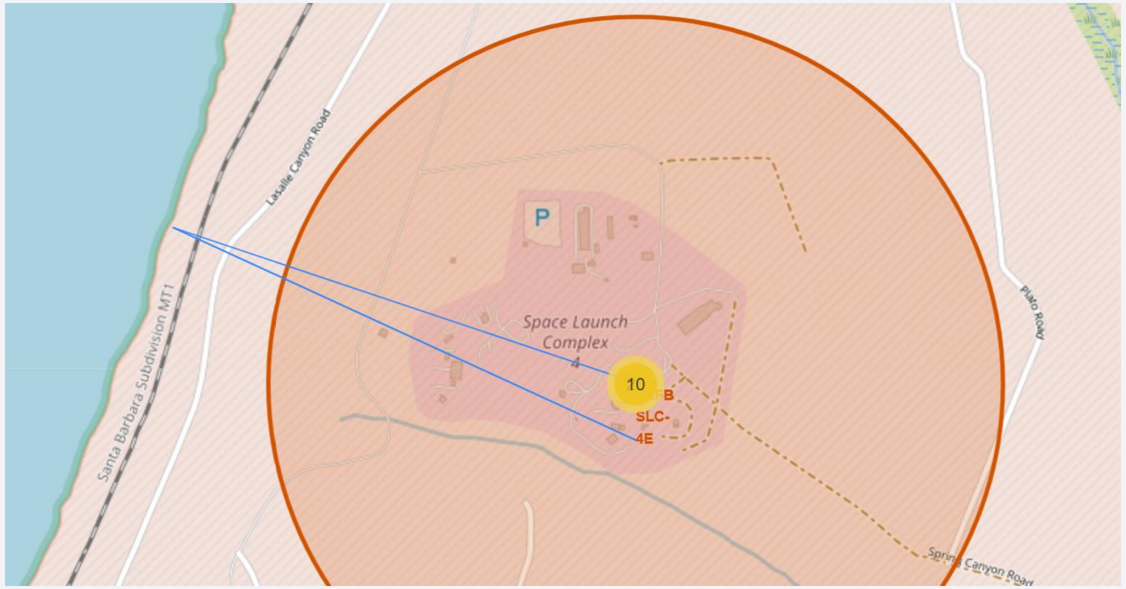
- Launch sites marked in red
- Launch sites located in the US and as south as possible

Launch attempts held at CCAFS LC-40

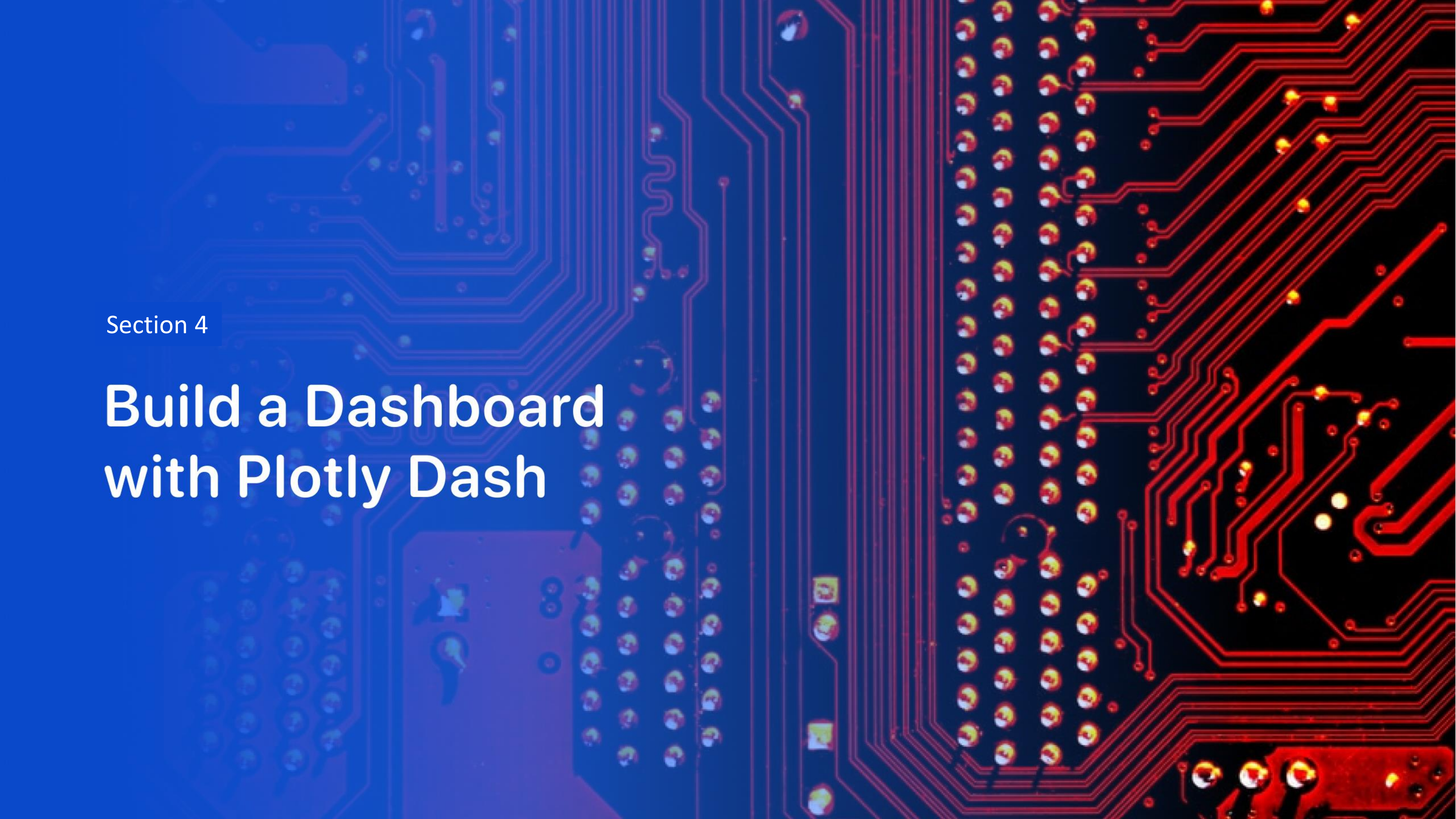


- Each marker represents a launch attempt; a red marker indicates a failed launch, a green marker indicates a successful launch, ordered by launch date starting from centre of spiral.
- Shows initial launch attempts were all failures but most recent ones were mostly successful.

VAFB SLC-4E proximity to ocean



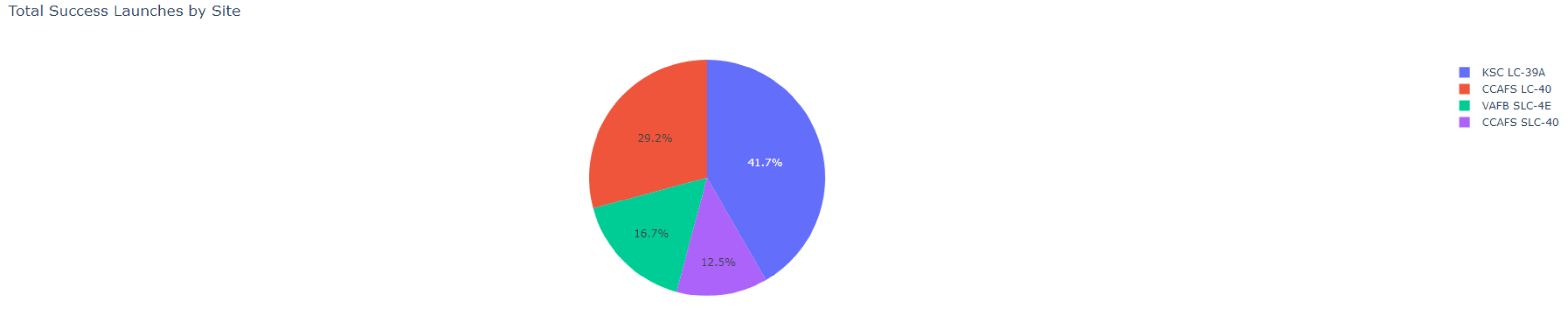
- Shortest distance to ocean from Vandenberg Space Launch Complex 4 indicated with (upper) blue line
- Close proximity minimizes chance of collateral damage during a failed mission



Section 4

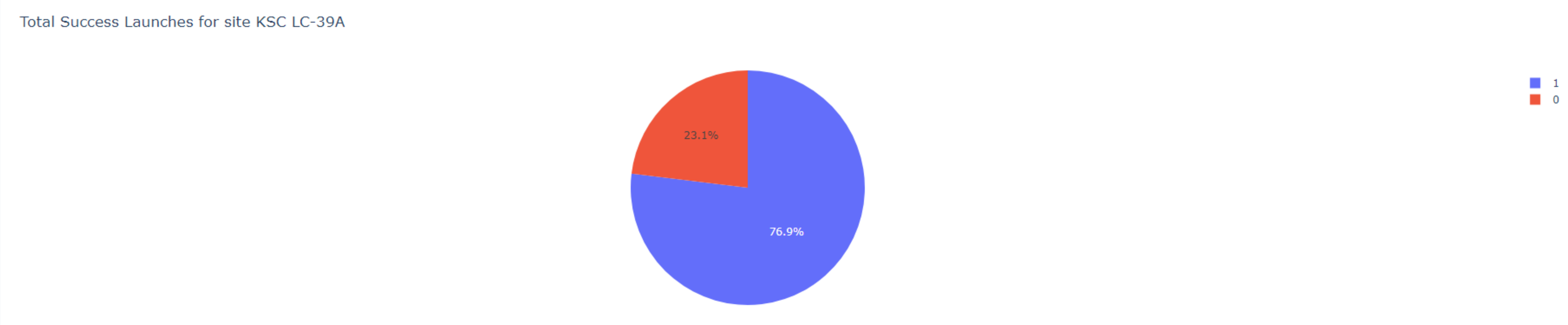
Build a Dashboard with Plotly Dash

Launch Success from all Sites



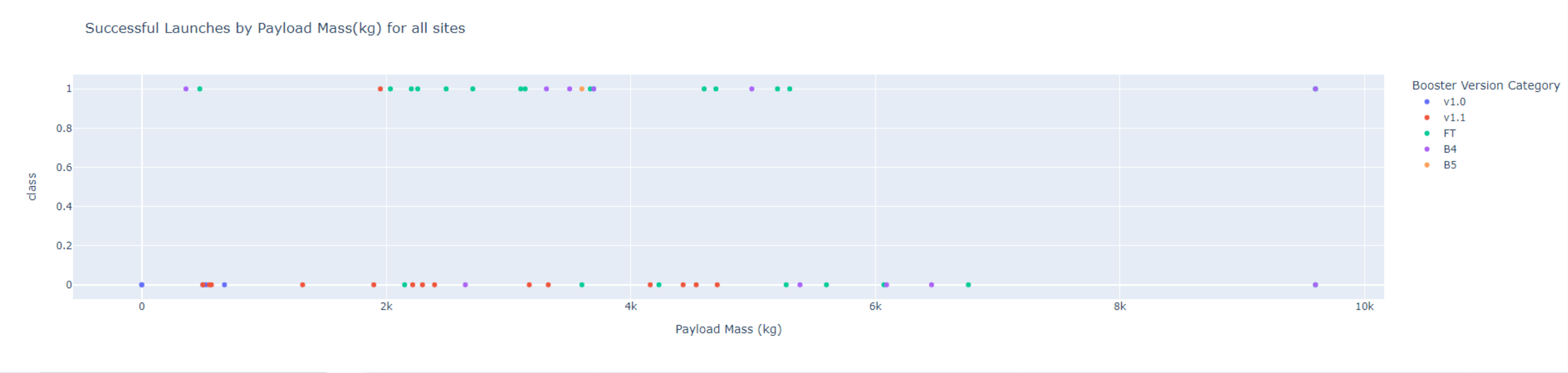
- Successful launches plotted in a pie chart, colour coded by launch site
- The launch site with the most successful launches is Kennedy Space Center Launch Complex 39 (KSC LC-39A) with 41.7%

KSC LC-39A launch success rate



- Successful launches are in blue, failed launches are in red.
- KSC LC-39A also held the highest ratio of successful launches to failed launches.

Launches by Payload Mass

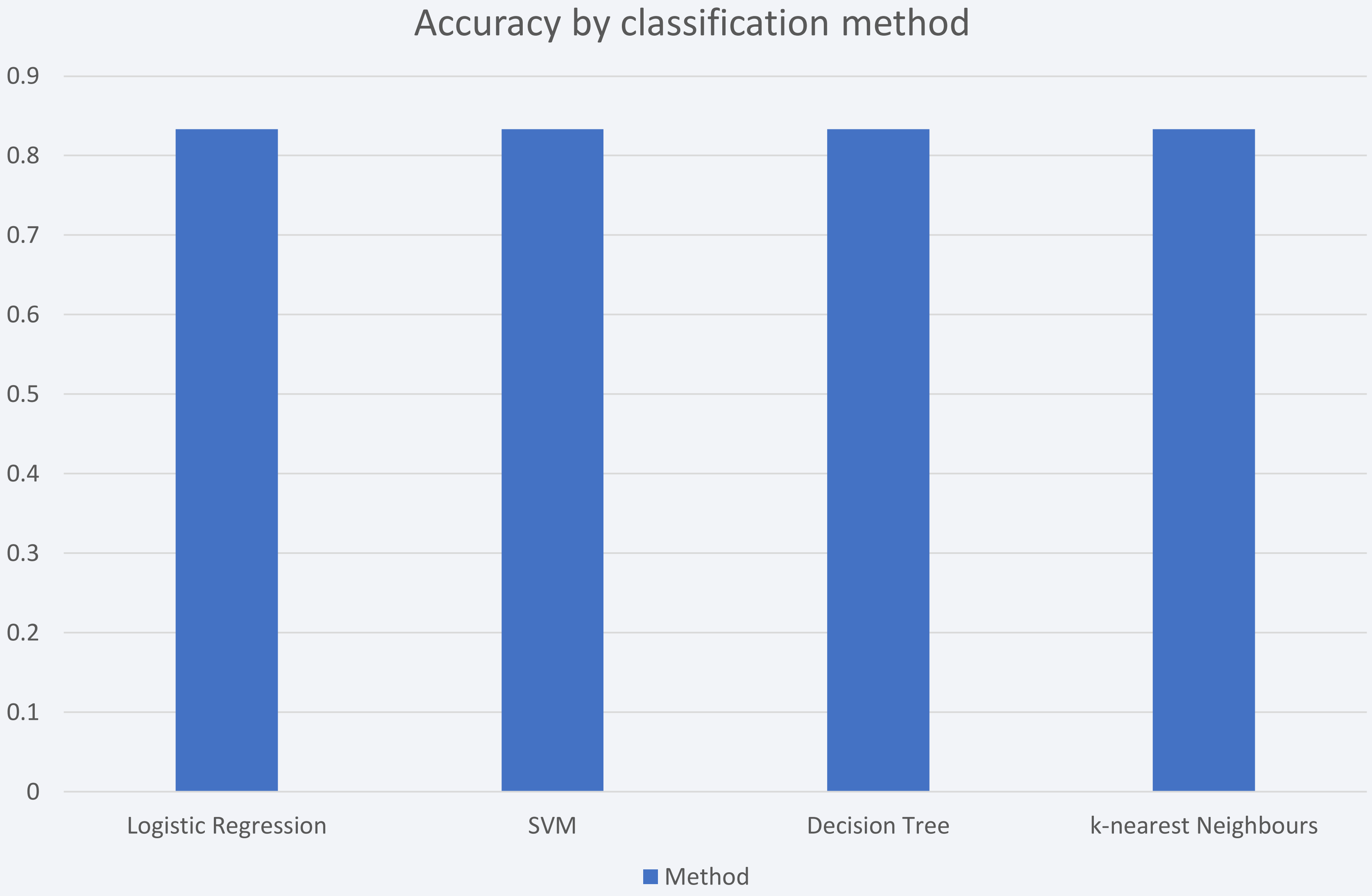


- Payload mass on the x-axis, launch outcome on the y-axis where 0 is fail, 1 is success.
- Colour coded by booster version.
- Most launches (and therefore most successful launches) were with a payload mass between 0 and 5000kg

Section 5

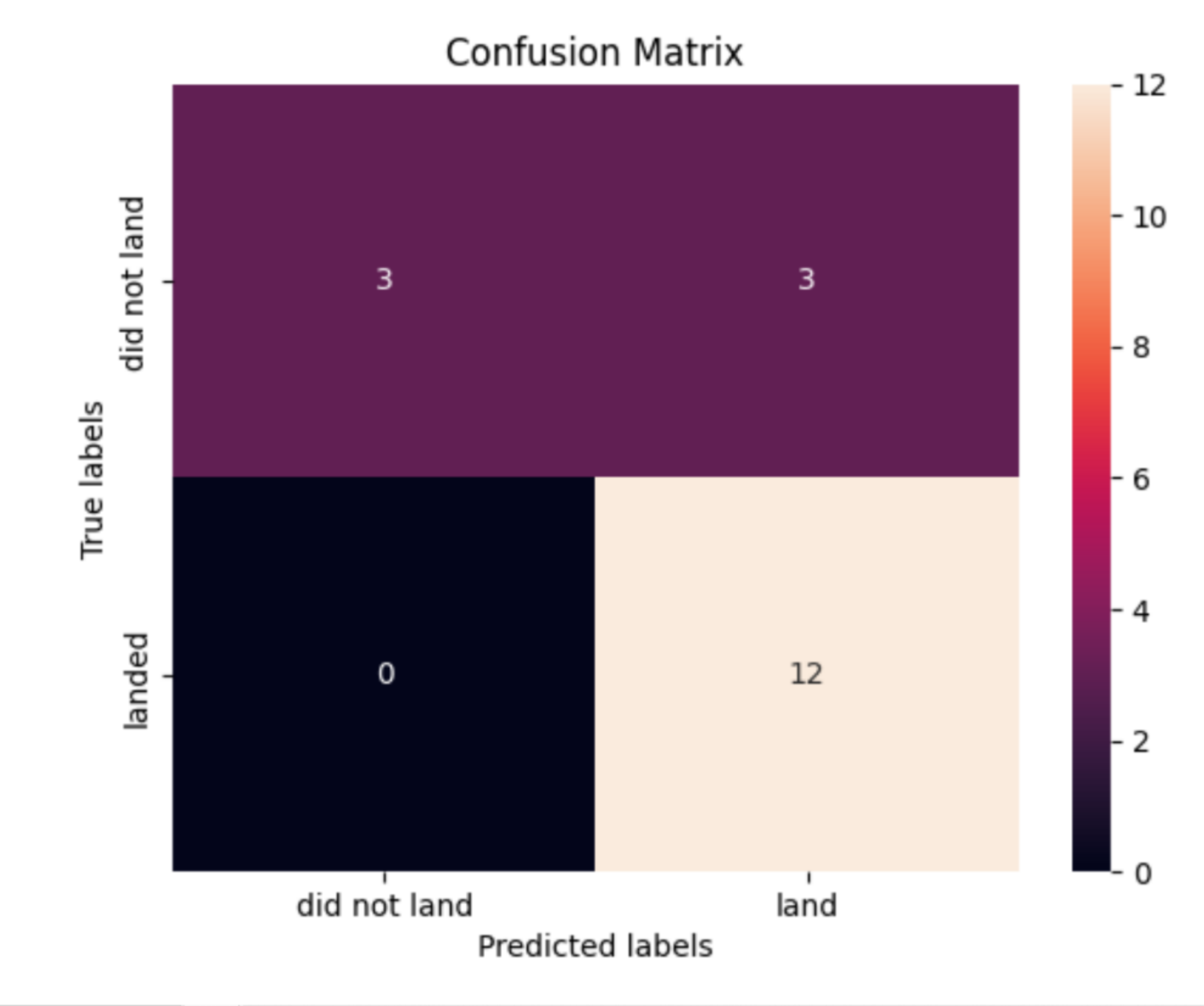
Predictive Analysis (Classification)

Classification Accuracy



- All 4 models showed the same level of accuracy at 83.34%

Confusion Matrix



- Same confusion matrix created from all 4 methods.
- All methods can successfully predict a successful landing but failed at predicting a failed landing (false positive).

Conclusions

- The success rate of SpaceX has improved due to learning for their past experiences and iterating on their process.
- It however took them until 2017 to have their first successful ground landing.
- Kennedy Space Center was their main launch site, due to both the highest number of missions held there as well as their best-performing launch site.

Appendix

- Link to Github repository containing all notebook and .py files:
- https://github.com/Magician960/IBM-DS-Capstone-Project

Thank you!