IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Santiago Marr
23/10/2023

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection performed using API, webscraping
  - Data wrangling performed with Python
  - Exploratory data analysis performed using SQL
  - Folium & Plotly to create interactive maps and dashboard
  - Predictive analysis using logistic regression, service vector machine, decision tree classifier & K nearest neighbor
- Summary of all results

Introduction

- In this report, we will predict if the Falcon 9 first stage will land successfully.
- SpaceX 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 SpaceX 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

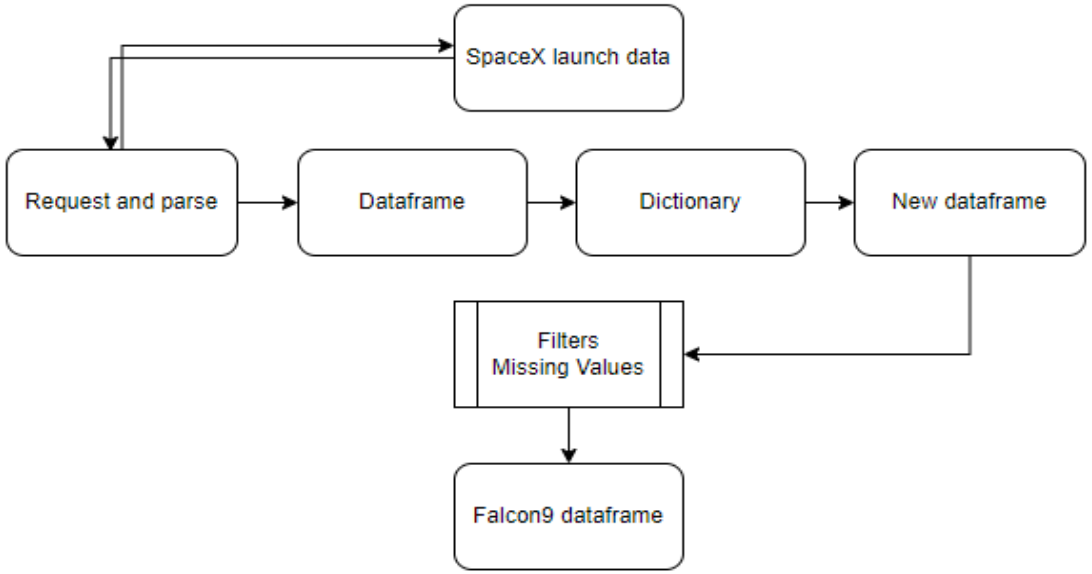
Methodology

Executive Summary

- Data collection methodology:
  - REST calls to SpaceX API & webcraping from Wikipedia
- Perform data wrangling
  - Missing values, data type identification, key metrics calculations using Python
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

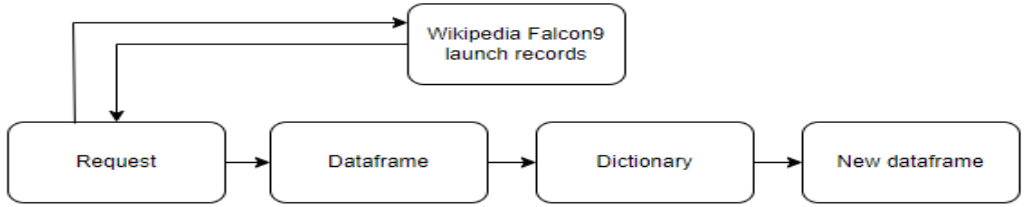
Data Collection

API



- API get request to call SpaceX launch data
- Dataframe created from dataset
- Relevant data extracted from dataframe and dictionary created and saved to new dataframe
- New dataframe filtered for Falcon9 and missing values handled
- Final Falcon9 dataframe created

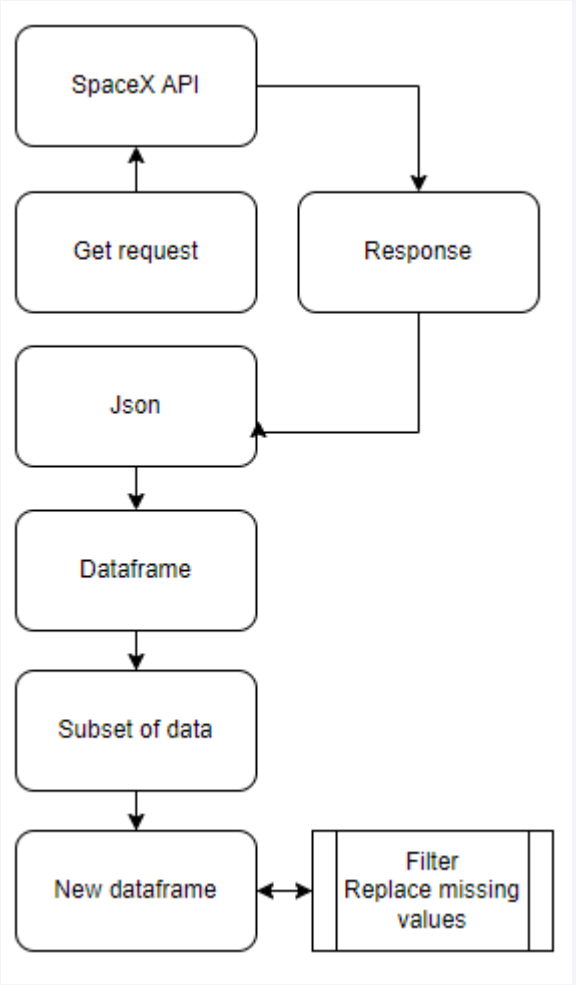
Webscraping



- Get request to Wikipedia page “List of Falcon 9 and Falcon Heavy Launches”
- Launch HTML tables created, parsed and passed to dataframe
- Extract tables from data to fill dictionary
- New dataframe created from parsed data into dictionary

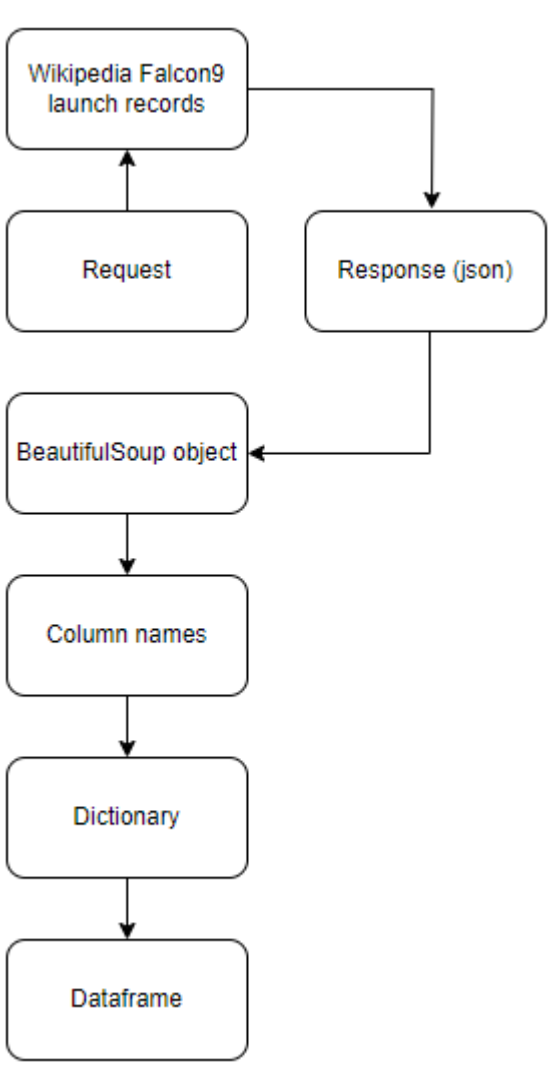
Data Collection – SpaceX API

- Source: SpaceX API
- Get request to API to return response
- Decode response as json and return as Pandas dataframe
- Extract subset of data from dataframe and store in lists
- Construct dataset using data extracted and combine to dictionary
- Create new dataframe from dictionary
- Filter new dataframe for Falcon9 launches and create another dataframe
- Replace missing values from Payload mass column with mean
- Link To Data Collection API notebook



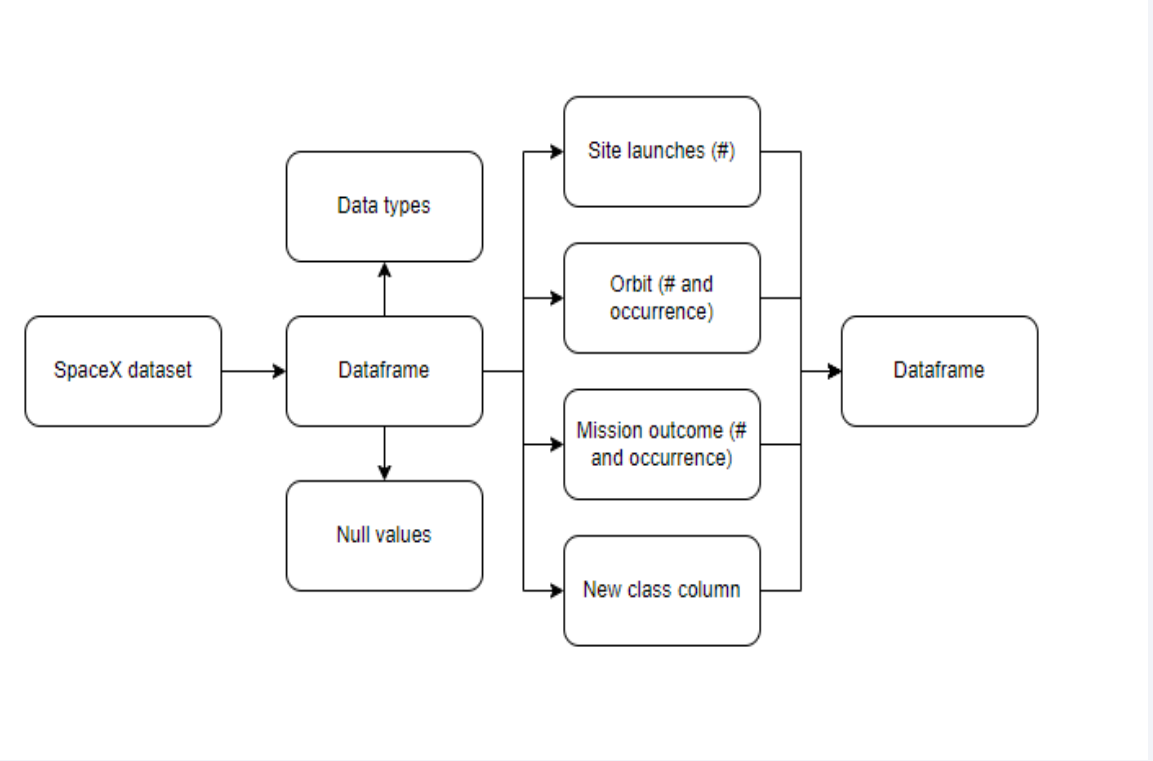
Data Collection - Scraping

- Source: List of Falcon 9 and Falcon Heavy launches Wikipedia page updated on June 9, 2021
- Get request to web page to return a response in json
- Create BeautifulSoup object and store the response in soup object
- Extract column names from HTML table header
- Parse HTML tables and create dataframe by passing data into dictionary
- Link to Webscraping Notebook via Github



Data Wrangling

- Using SpaceX dataset and Python, missing values were identified in each column, data types were identified for each attribute, number of launches at each site, number and occurrence of each orbit & mission outcome were calculated. A new column for Class was created from the Outcome column to identify successful and unsuccessful outcomes for each launch.
- Link to Data Wrangling notebook



EDA with Data Visualization

- Scatter plots
  - Flight number vs payload mass to determine flight outcome (success or fail) related to payload mass
  - Flight number vs launch site determine success or failure by launch site
  - Payload mass vs launch site to determine the explore the success or failure of payload masses at the launch sites
  - Flight number vs orbit to determine the success or failure of each flight by orbit
  - Payload mass vs orbit to determine the success or failure of payload masses at each orbit
- Bar chart
  - Orbit success rate to determine the average success rate of each orbit
- Line chart
  - Average success rate by year to view the trend of success or failure of all launches over time

- Link to EDA Data Viz Notebook

EDA with SQL

- Query to select the names of the unique launch sites in the space mission
- Query to display the first five records where launch site begin with the letters “CCA”
- Query to display the total payload mass carried by boosters launched by NASA (CRS)
- Query to display the average payload mass carried by booster version F9 v1.1
- Query to list the date of the first successful landing outcome in ground pad
- Query to list the names of the boosters that have success in drone ship and have payload mass between 4000 and 6000 kg
- Query to list the total number of successful and failure mission outcomes
- Query to list the names of the booster versions that have carried the maximum payload mass
- Query to list the records that display the month names, failure landing outcomes in drone ships, booster version, launch sites for the months in 2015
- Query to rank the count of landing outcomes between June 4, 2010 and March 20, 2017

- Link to SQL notebook

Build an Interactive Map with Folium

- Geo world maps with zoom in/out functionality with marker and circle features to map the launch sites, marker clusters to highlight the successful (green) or failed (red) launches at each site, lines pointing to the nearest coastline, railroad, highway & nearest city of one selected site to show proximity of launch site to large bodies of a water, major transportation & population centers.

- Link to Interactive Map with Folium notebook

Build a Dashboard with Plotly Dash

- Pie charts
  - Successful launches by launch site
  - Drop down feature to filter charts by launch site
- Scatter chart
  - Correlation between payload and launch success
  - Slider to select payload range

- Link to Plotly Dash notebook

Predictive Analysis (Classification)

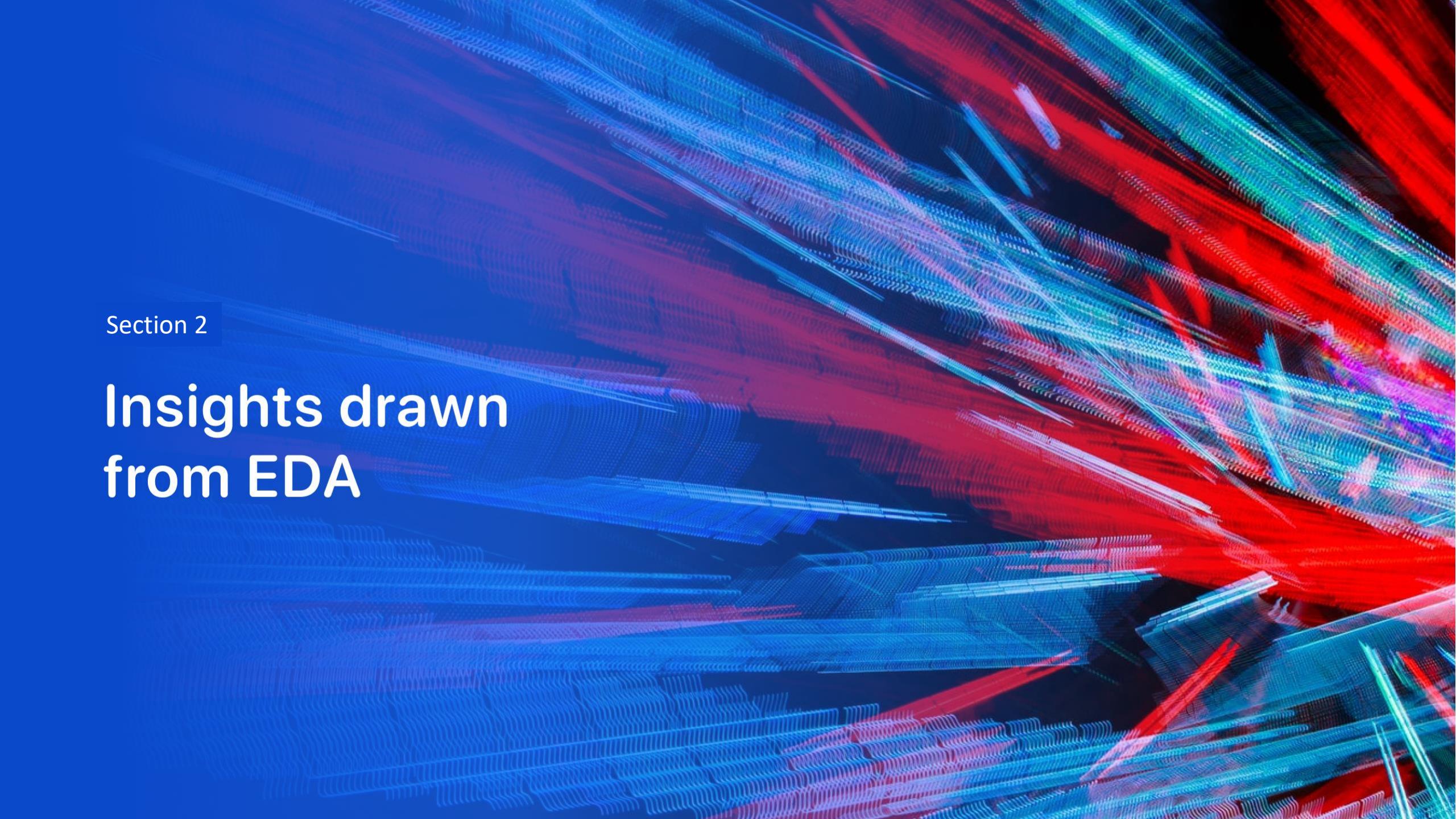
- Source: Falcon9 dataframe
- Create numpy array from Class column, assign to Y variable (Pandas output)
- Standardize data using StandardScaler then reassign to variable X
- Split data into X_train, X_test, Y_train, Y_test training and test data
- Create logistic regression, support vector machine, decision tree classifier and k nearest neighbor models

- You need present your model development process using key phrases and flowchart

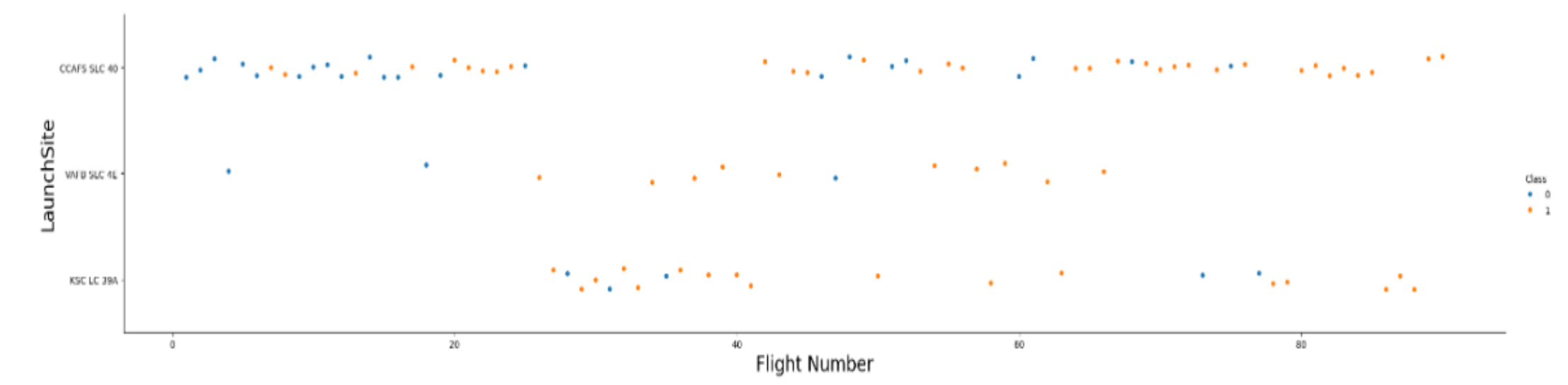
- Link to Predictive Analysis notebook

Results

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

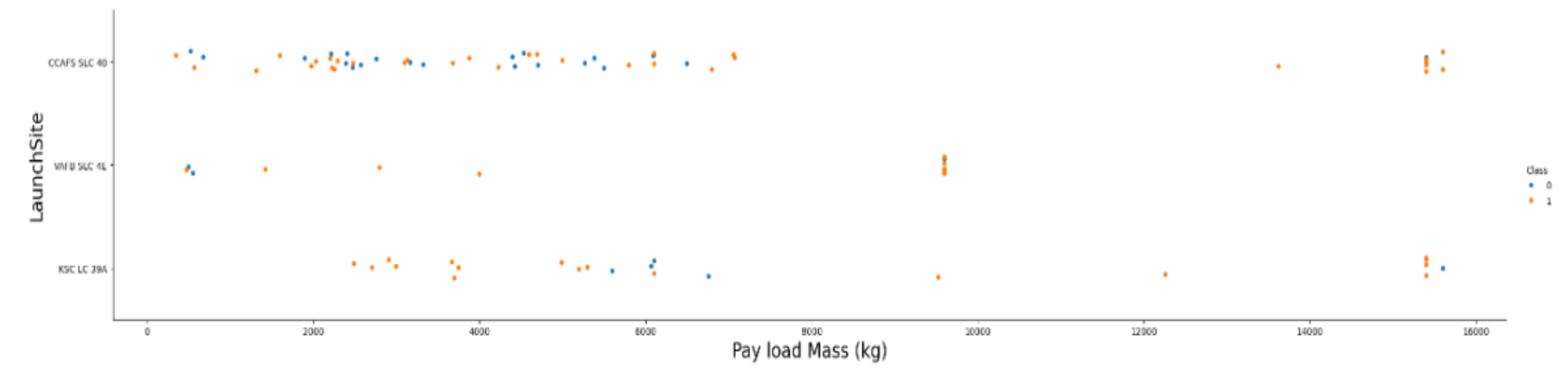

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Launch site KSC LC 39A has highest launch success rate at 77.27%
- VAFB SLC 4E has the second-highest launch success rate at 76.92%, but the lowest number of total launches at 13 launches.
- CCAFS SLC 40 has the highest number of total launches at 55, with the lowest launch success rate at 60%.

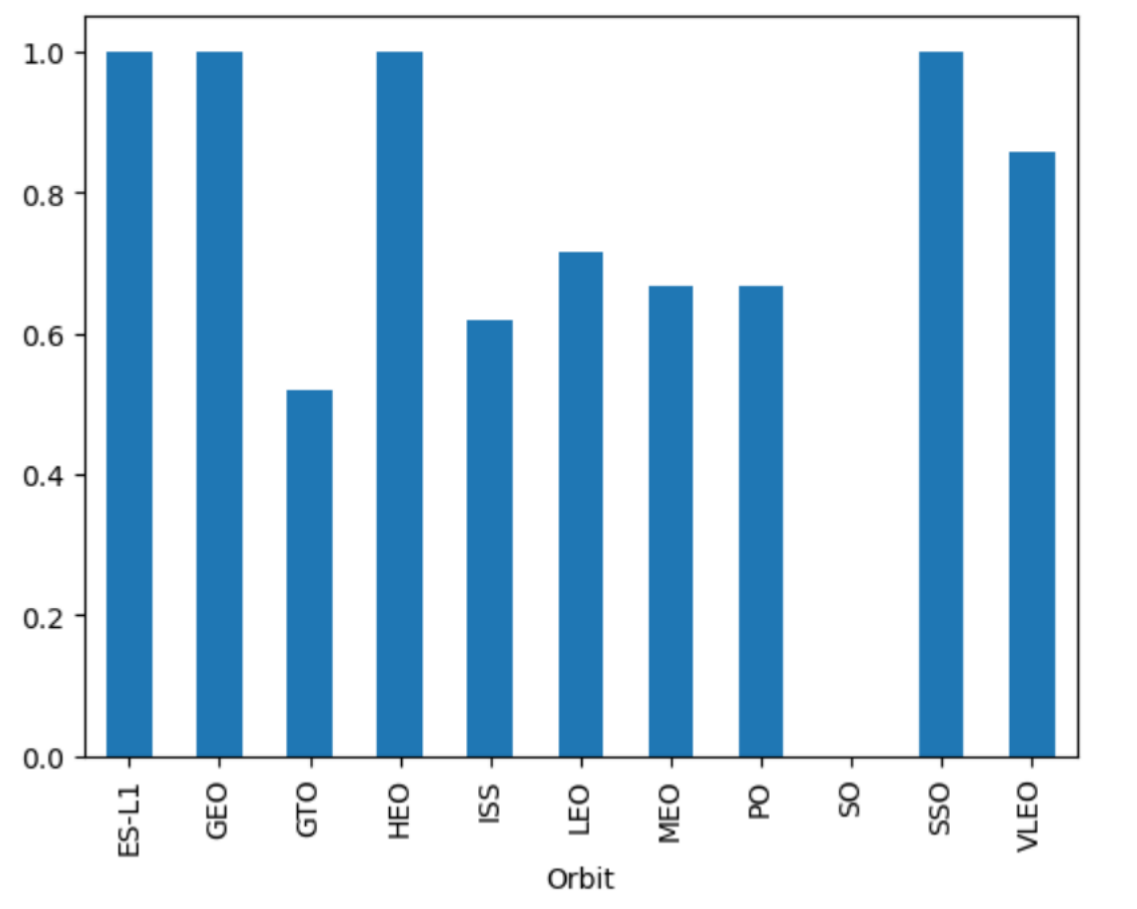
Payload vs. Launch Site



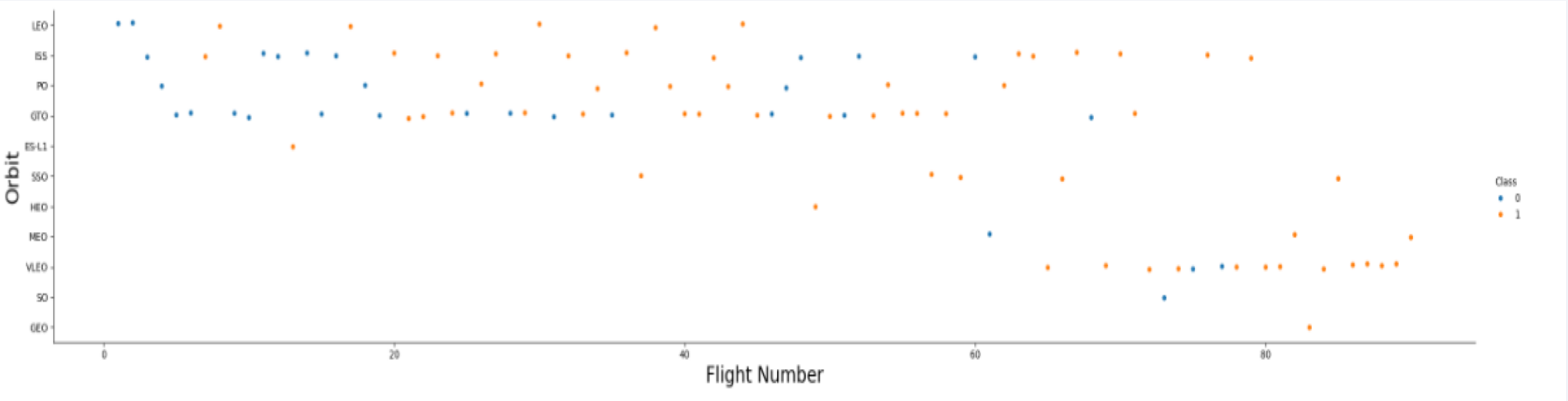
- KSC LC 39A has the best launch success rate overall with high, medium and light payloads.
- VAFB SLC 4E was not tested with high payloads and success with light and medium payloads in not convincing.
- CCAFS SLC 40 completed the highest number of successful launches with the heaviest payloads, but has a high number of failed launches overall, especially with lighter payloads.

Success Rate vs. Orbit Type

- 100% success rates with orbits ES-L1, GEO, HEO and SSO.
- 80% success rate with VLEO.
- Poor success rates with GTO, ISS, LEO, MEO, PO.
- 0% success rate with orbit SO.

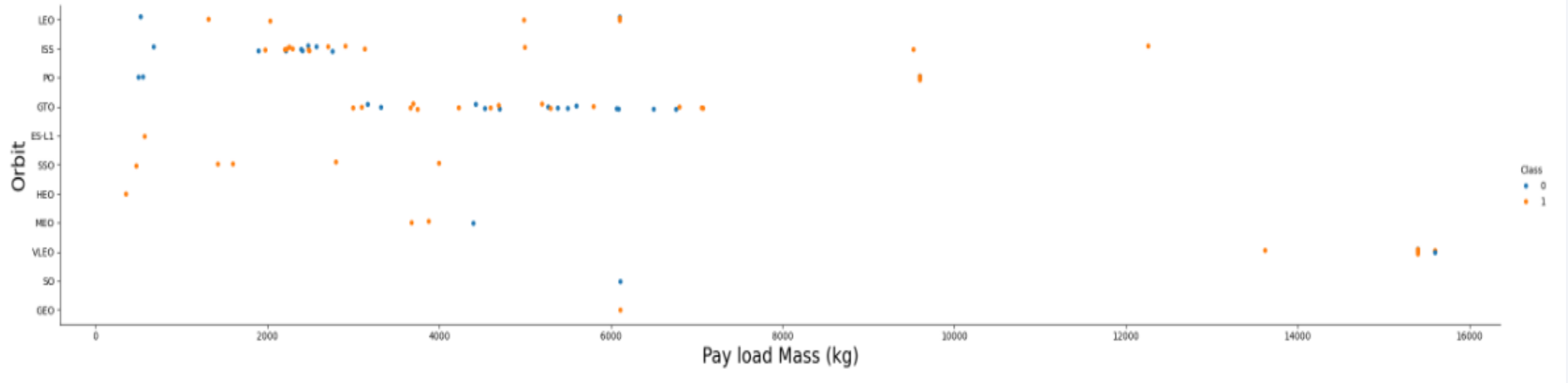


Flight Number vs. Orbit Type



- Orbits with the highest success rates were tested the least amount of times.
- Among the orbits with the highest success rates, SSO was tested the most.
- Orbits with mid-range success rates between 40% and 70% were tested most.
- Orbit SO was only tested once and failed, resulting in a 0% success rate.
- The success rate of the LEO orbit appears related to the total number of flights, yet no relationship between success and total number of flights in the GTO orbit.

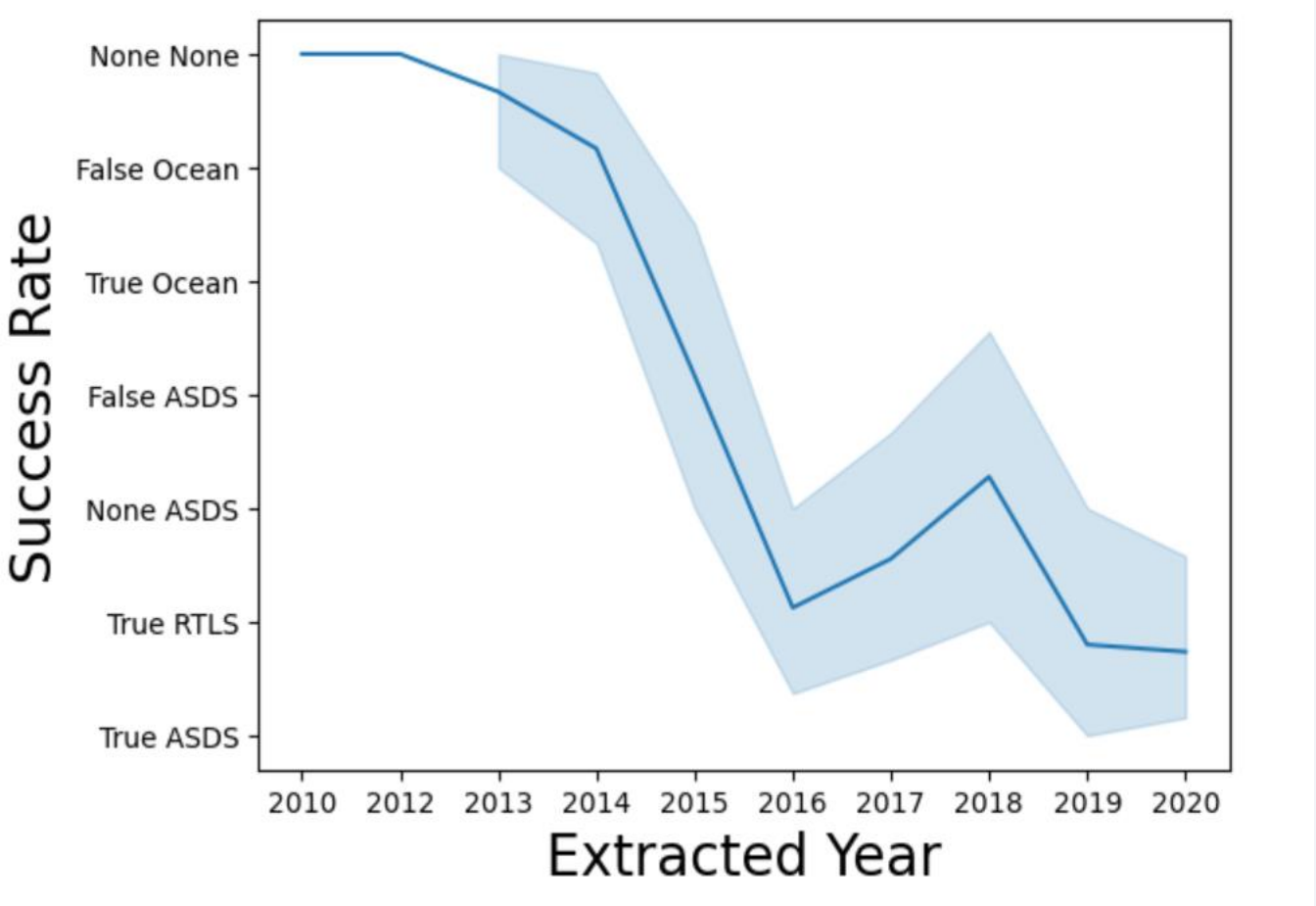
Payload vs. Orbit Type



- With heavy payloads the successful landing rates are greater for Polar, LEO and ISS.
- GTO is indistinguishable as both positive and negative landing results are mixed.

Launch Success Yearly Trend

- The success rate of all launches increased overall between the years 2013 and 2020.



All Launch Site Names

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

- SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
  - SQL query to extract the unique launch site names from the column “Launch_Site” from the “SPACEXTBL” table

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The first five rows of data in the SpaceX dataset where the launch site name begins with the letters “CCA” are from launch site CCAFS LC-40 dated April 6, 2010, August 12, 2010, May 22, 2012, August 10, 2012 and January 3, 2013
  - SELECT * FROM SPACEXTBL WHERE Launch_Site Like "CCA%" LIMIT 5
  - SQL query returns all columns from five rows of the SpaceX dataset where the launch site names begin with the letters “CCA”
  - “Like” and “%” used in the WHERE clause to return launch site names beginning with the letter “CCA” *(“%” placement after the letters)*; LIMIT clause used to return only the first five rows meeting those conditions.

Total Payload Mass

Total payload carried by boosters from NASA CRS

45,596 kg

- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer Like "NASA (CRS)"
  - SQL query returns the sum of values in the “PAYLOAD_MASS_KG_” column for customer NASA (CRS) from the SpaceX dataset
  - Sum total of the payload mass carried by all boosters launched by NASA (CRS) from all launch sites

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

2928.40 kg

- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version == "F9 v1.1“
  - SQL query that returns the average from the PAYLOAD_MASS_KG_ column where the values in the Booster_Version column are F9 v1.1

First Successful Ground Landing Date

The first successful landing outcome on a ground pad

December 22, 2015

- SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome == "Success (ground pad)“
  - SQL query selecting the first date from the SpaceX dataset where the landing outcome for a ground pad launch was successful returns the first successful launch date for that type

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 AND Landing_Outcome = "Success (drone ship)"

Total Number of Successful and Failure Mission Outcomes

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

- SELECT Mission_Outcome, COUNT(*) AS TOTAL FROM SPACEXTBL GROUP BY Mission_Outcome ORDER BY Mission_Outcome

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

- SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)

2015 Launch Records

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

- SELECT "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr(Date,1,4) = '2015'

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

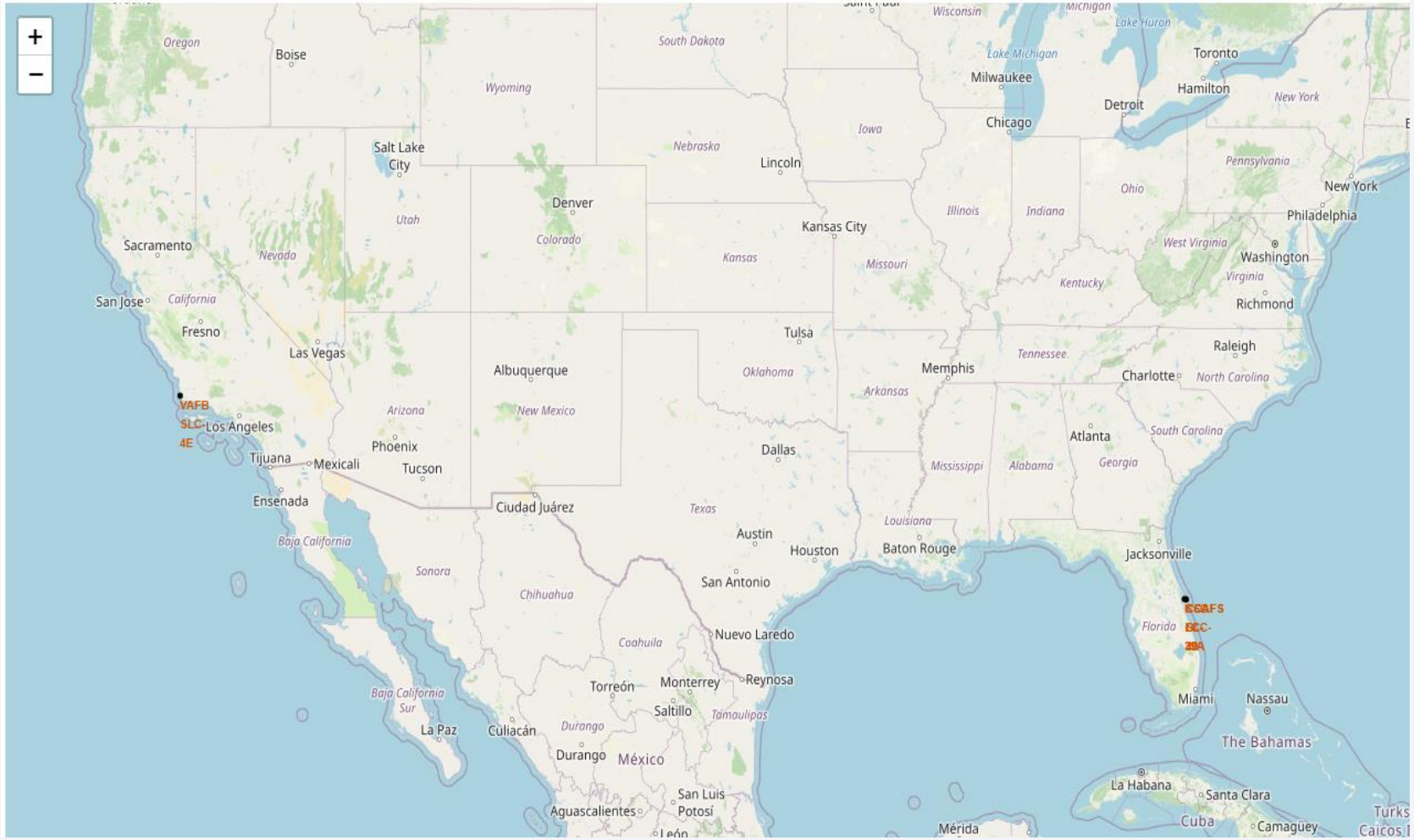
Landing_Outcome	COUNT
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- SELECT "LANDING_OUTCOME", COUNT(*) as 'COUNT' FROM SPACEXTBL WHERE substr(Date,1,4) || substr(Date,6,2) || substr(Date,9,2) between '20100604' and '20170320' GROUP BY "Landing_Outcome" ORDER BY "COUNT" DESC

Section 3

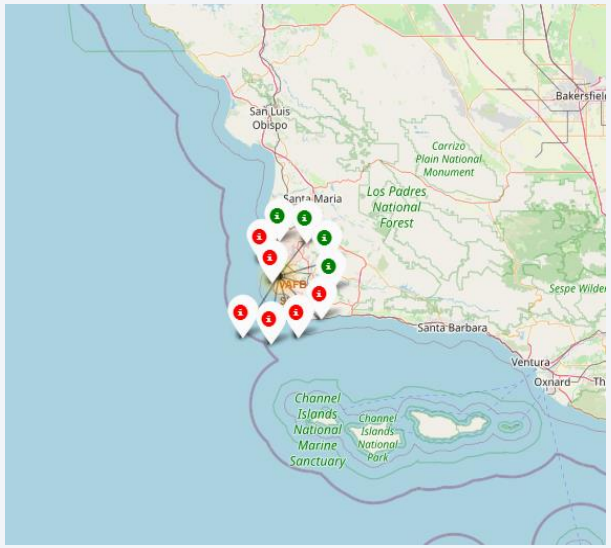
Launch Sites Proximities Analysis

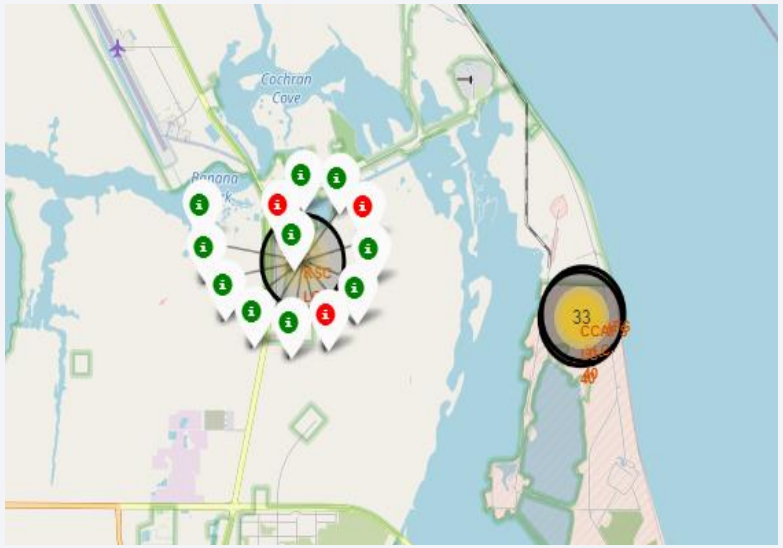
Location Map of All Launch Sites

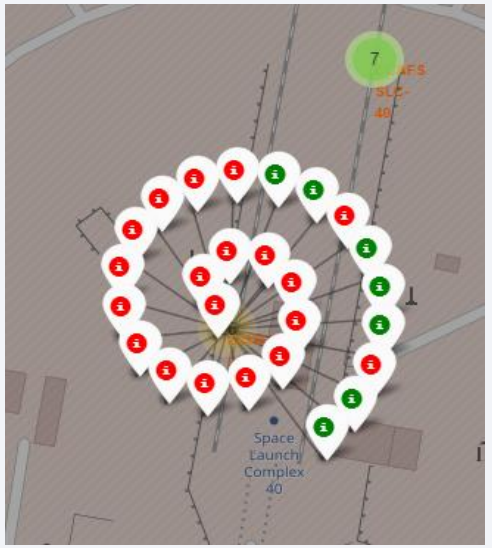


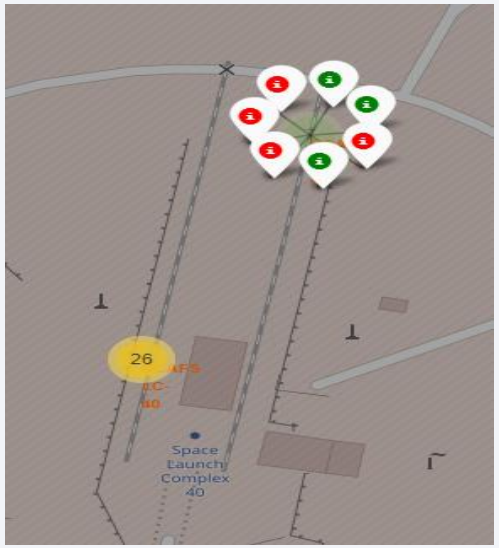
- Geo world map marking launch sites.
- Zoom in/out functionality.
- Launch sites located near coastline / large body of water
- Launch sites located in southernly areas of continental United States.

Launch Outcomes Map



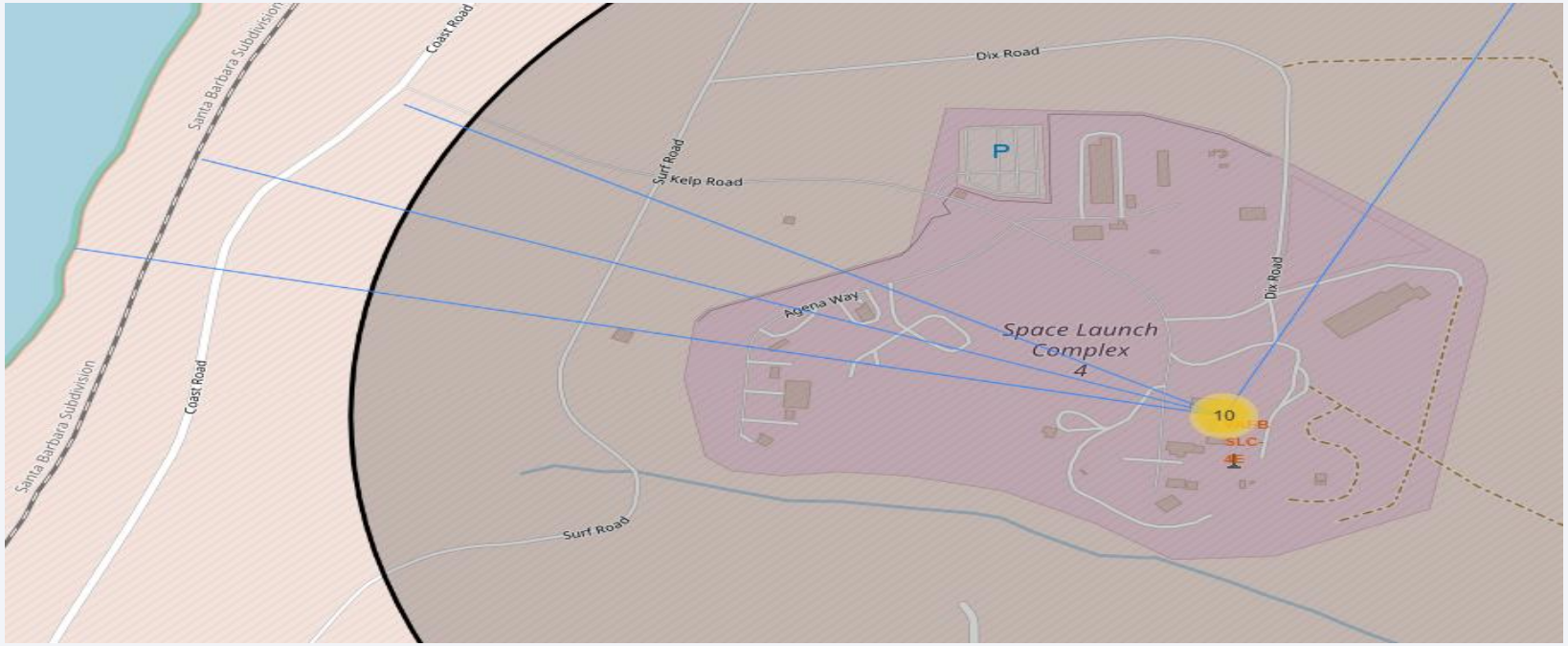




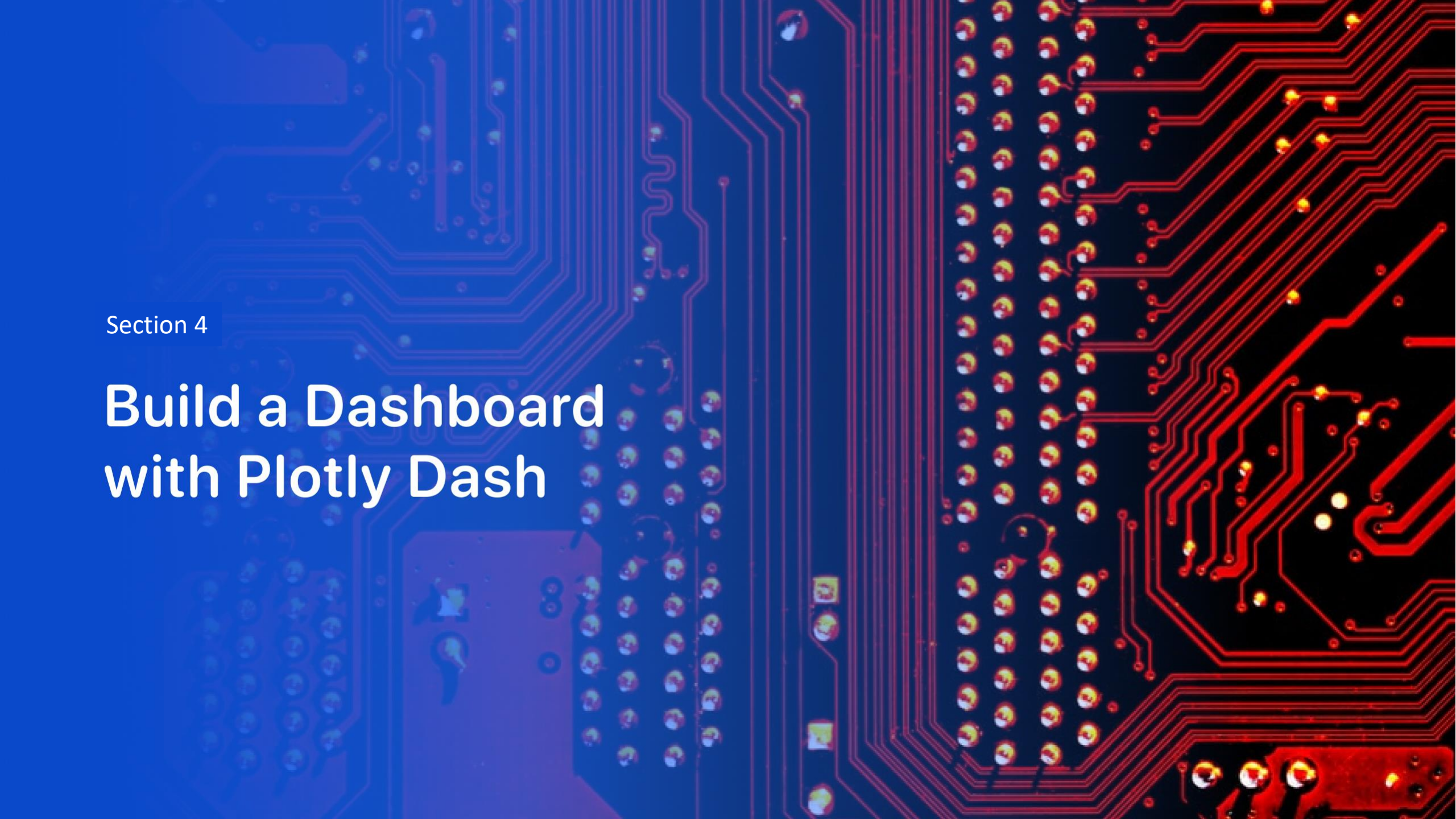


- Geo map of launch sites with zoom in/out functionality with successful and unsuccessful launches marked in green and red marker clusters, respectively.

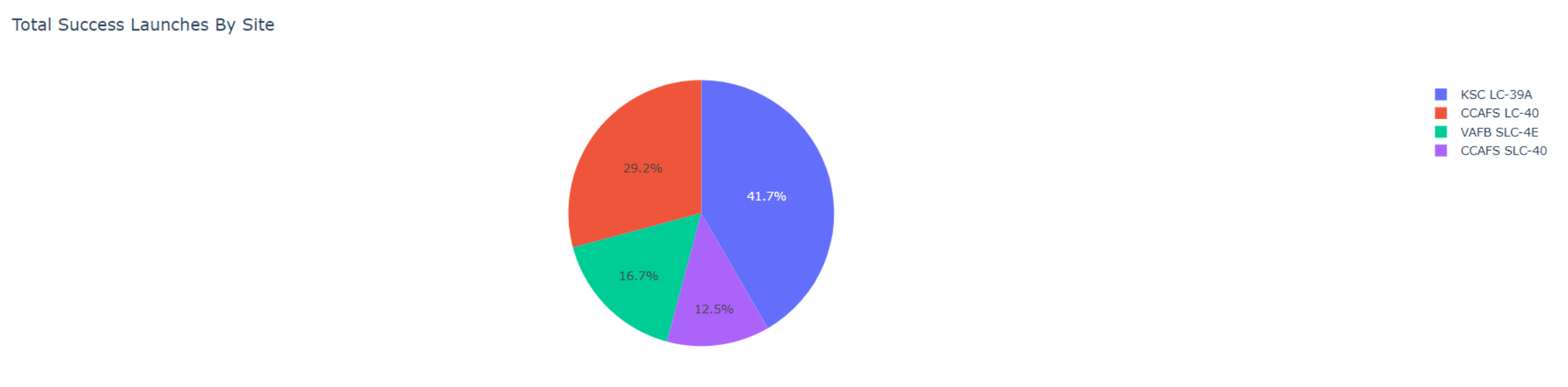
VAFB SLC-4E Proximity Map



- Site VAFB SLC-4E is located near US1, the Pacific coastline, the Santa Barbara Subdivision train line.
- Launch site in close proximity to coastline, large body of water, railroad and highway, yet distant from population centers.

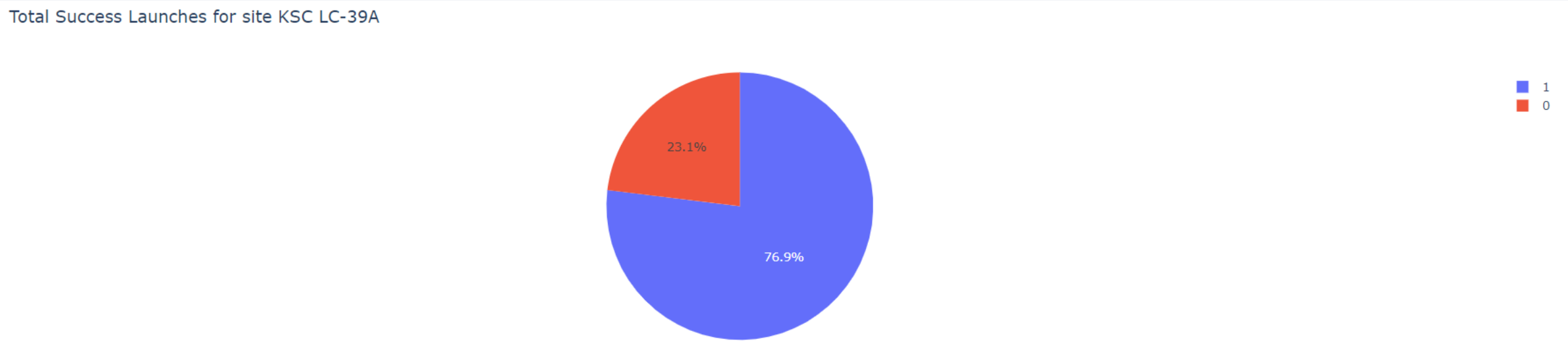

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches By Site



- Launch site KSC LC-39A leads launch sites with the most successful launches at 41.7%.
- Site CCAFS LC-40 ranks second at 29.2% - 12.5 points lower than KSC LC-39A.
- VAFB SLC-4E and CCAFS SLC-40 are not successful launch sites by comparison.

Highest Launch Success Ratio



- Site KSC LC-39A ranks the most successful among all launch sites.
- 76.9% of all total launches at KSC LC-39A were successful.
- Only 23.1% of all total launches at KSC LC-39A were unsuccessful.

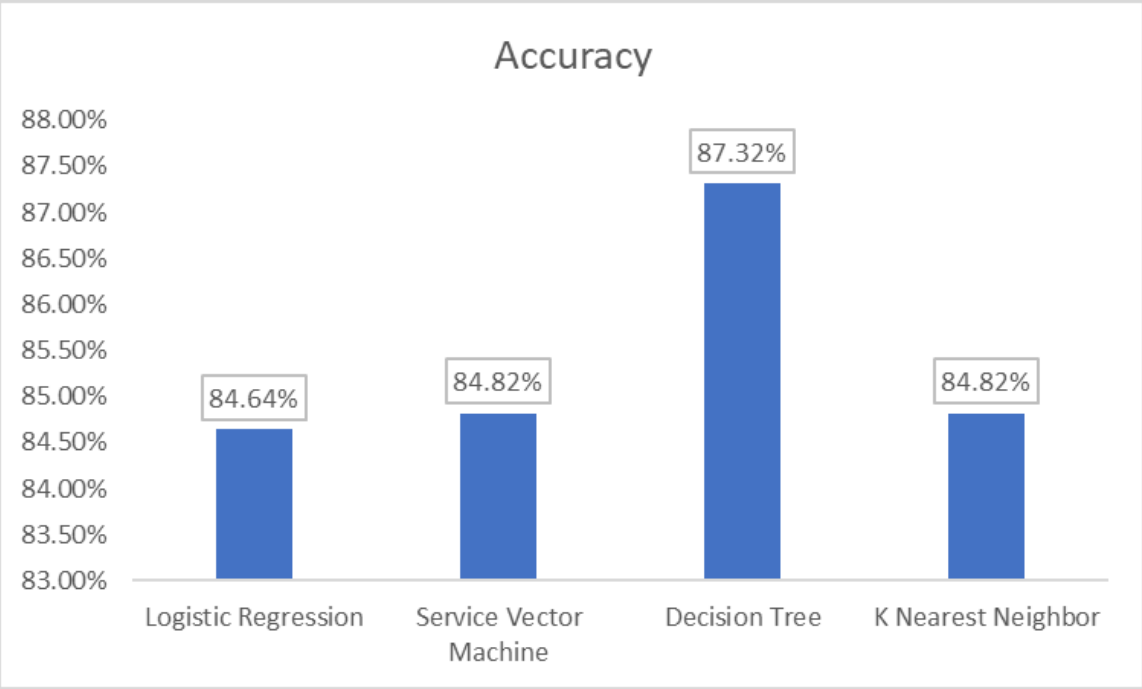
Payload vs. Launch Outcome

- At the heaviest payloads, booster B4 resulted with the highest number of successful launch outcomes.
- Booster FT performed notably as well, ranking in second behind booster B4 with the highest number of successful launch outcomes at the heaviest payload, while performing successfully overall at varying light, medium and heavy payloads.

Section 5

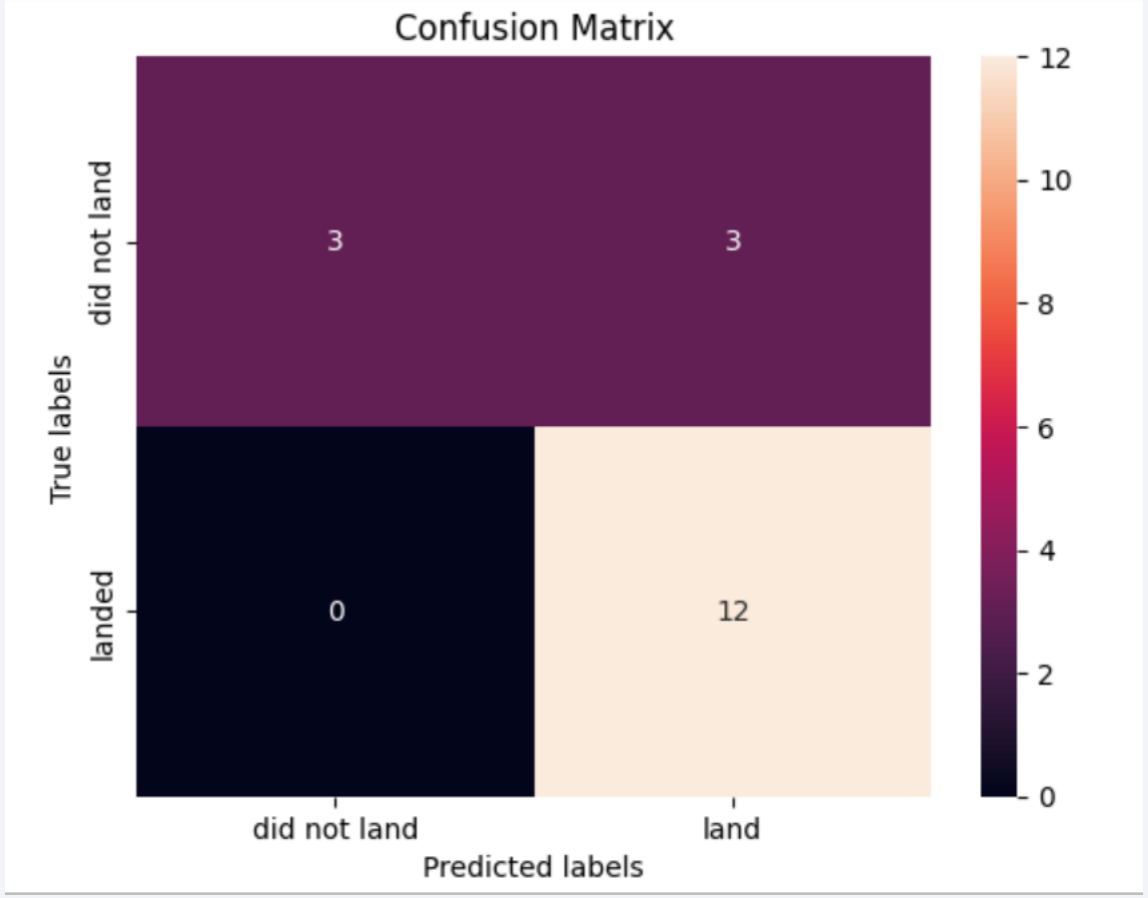
Predictive Analysis (Classification)

Classification Accuracy



- When tuned to best parameters, the decision tree classifier model returns the highest accuracy of the four models with 87.32% accuracy.

Confusion Matrix



- Decision tree classifier: 87% accuracy
- High True Positive (12) and high True Negative (3) values
- Low False Positive (3) and low False Negative (0) values

Conclusions

- KSC LC-39A ranks the most successful among all launch sites
- KSC LC 39A has the best launch success rate overall with high, medium and light payloads
- Booster B4 & FT are overall best performing boosters across varying payloads
- 100% success rates with orbits ES-L1, GEO, HEO and SSO

Appendix

- Github repository of Python, SQL, charting and dashboard notebooks

Thank you!