IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Ana Marques
30–Dec-2022

# Outline

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

# Executive Summary

- The following methodologies were used to analyze data:

**Summary of methodologies**

- Data Collection using web scrapping and SpaceX API;
- Exploratory Data Analysis (EDA). Including:
- data wrangling,
- data visualization and
- Interactive visual analytics
- Dashboard
- Predictive analysis (Classification)

**Summary of all results**

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

# Introduction

## Project background and context

- We predicted if the Falcon 9 first stage will land successfully.
- If we can determine if the first stage will land we can determine the cost of a launch, information essential other companies who wants to bid against SpaceX for a rocket launch.

## Common problems that needed solving

- What influences if the rocket will land successfully?
- The effect each relationship with certain variables will impact in determining the success rate of successful landing
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate

Section 1

# Methodology

# Methodology

## Executive Summary

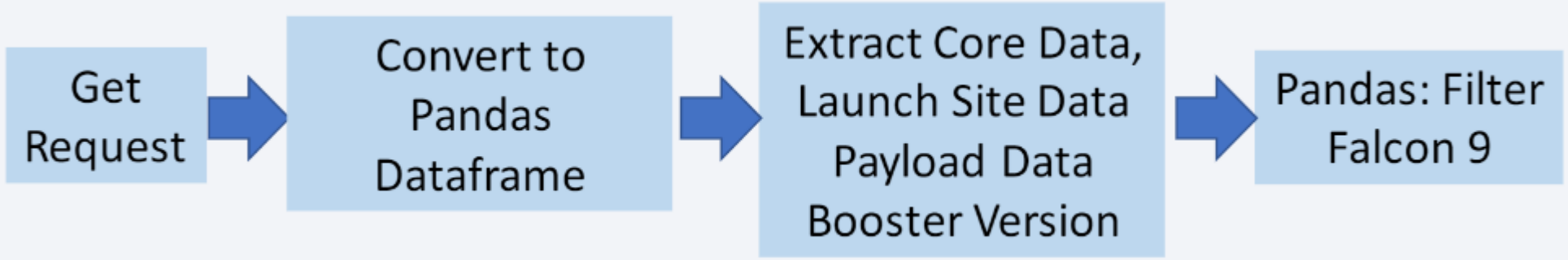
- Data collection methodology:
  - Requesting data from SpaceX Rest API , and with web scrapping from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and dropping irrelevant colums
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Using visualization tools suchs as Python´s matplotlib and seaborn libraries, as well as answering questions using SQL queries
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The models that were used are logistic regression, k-nearest neighbour and decision tree. Each model was trained, tuned and evaluated to find the best one

# Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

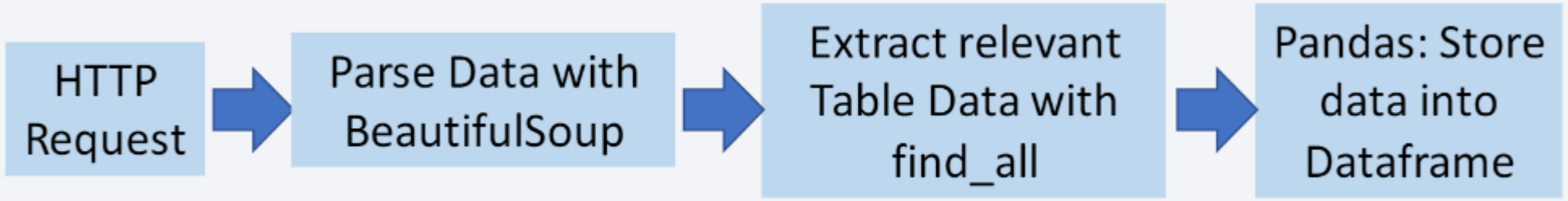
# Data Collection – SpaceX API

- Request and parse the SpaceX launch data using the GET request
- Normalize JSON response into a dataframe and extract only useful columns using auxiliary functions
- Create new pandas dataframe from dictionary an filter dataframe  ti only include Falcon9 launches
- Handle missing values
- Export csv file
- GitHub URL: Lab 1: Collecting the data



# Data Collection - Scraping

- Request rocket launch data from Wikipedia page and extract column/variable names from the HTML table header.
- Create a data frame by parsing launch HTML tables
- Export to CSV file
- GitHub URL [jupyter-labs-webscraping.]



# Data Wrangling

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome by orbit type
- Design a landing outcome label from Outcome column using one-hot encoding
- Export to cv

GitHub URL: Lab 2 Data wrangling

# EDA with Data Visualization

Charts:

- Playload mass vs Flight number vs Success Rate: Development of the playload mass and the success rate over ime
- Launch site vs Flight number vs Success rate: Success rate of each launch site over time
- Launch vs Payload mass vs Success rate: Wich Playload is best to have success at a specific launch site
- Orbit type vs Success rate: Which orbit types have the highest success rates
- Orbit type vs Flight number vs Success rate: Development of orbit types over time
- Orbit type vs Payload mass vs Success rates: success rate for orbit type/playload mass
- Success rate vs Year: Show the success development over time



GitHub URL:Lab - EDA with Visualization

# EDA with SQL

- names of the unique launch sites in the space mission
- records where launch sites begin with the string 'KSC'
- total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- date where the first successful landing outcome in drone ship was acheived.
- names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- total number of successful and failure mission outcomes
- names of the booster_versions which have carried the maximum payload mass. Use a subquery
- records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

GitHub URL:Lab - EDA with SQL lab

# Build an Interactive Map with Folium

- Objects were created and added to Follium map. Marker objects used to wshow all launch sites on a map as the successful/failed launches for each site on map.

- Map Objects:
  - Edge Circles (radius 1000m): Space launch sites
  - Markers: for labeling all objects
  - MarkerCluster: for creating a bunch of markers around space lauch sites to indicate success or failure of the landing of the rocket's first stage
  - Lines: Measure the distance between the launch site and the next coat or next city

GitHub URL: Lab - Interactive Visual Analytics with Folium

# Predictive Analysis (Classification)

Perform exploratory Data Analysis and determine Training Labels

- create a column for the class
- Standardize the data
- Split into training data and test data

Find best Hyperparameter for SVM, Classification Trees and Logistic Regression (Model Building for each method (logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor)

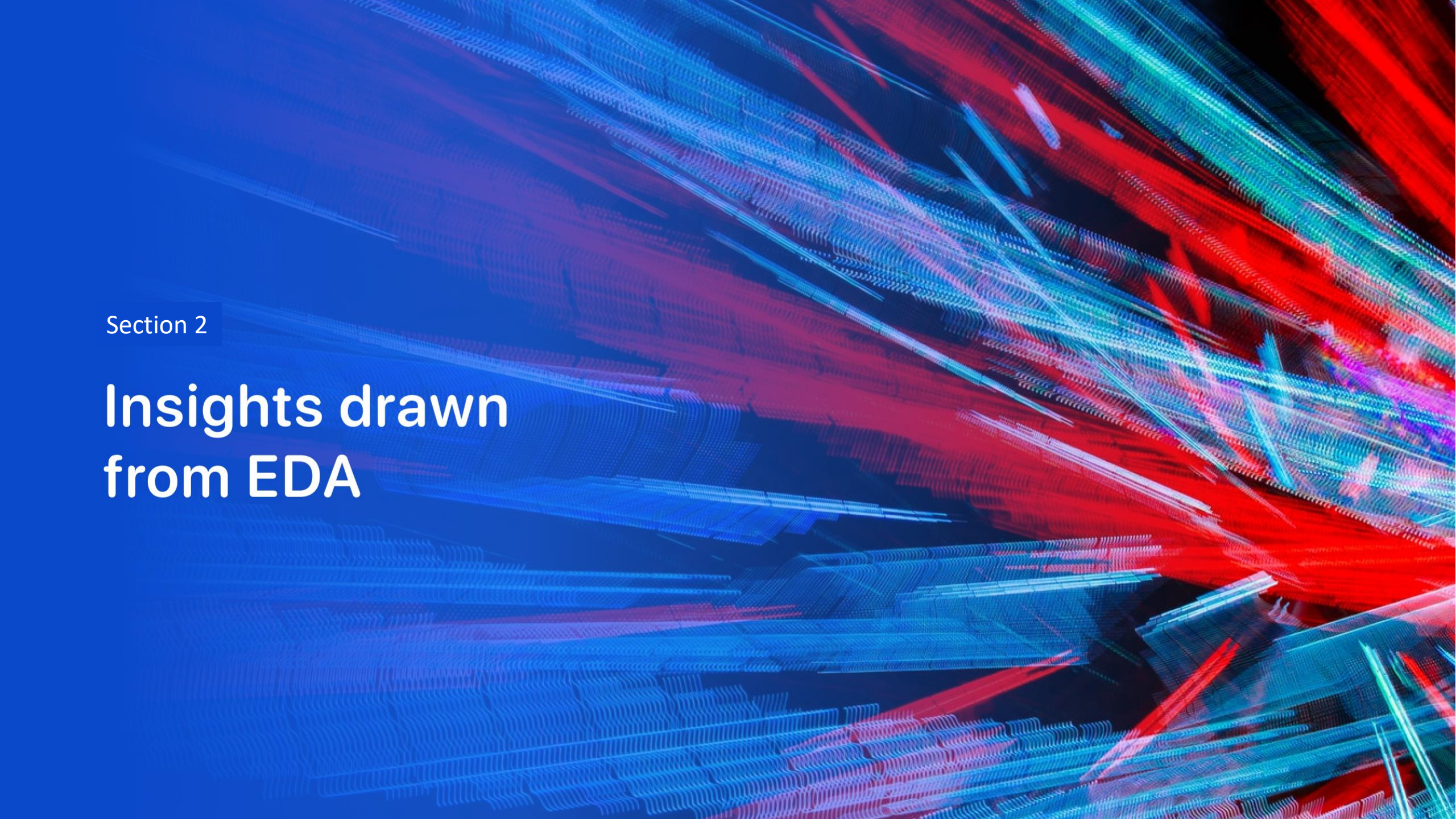
Optimization

Evaluation

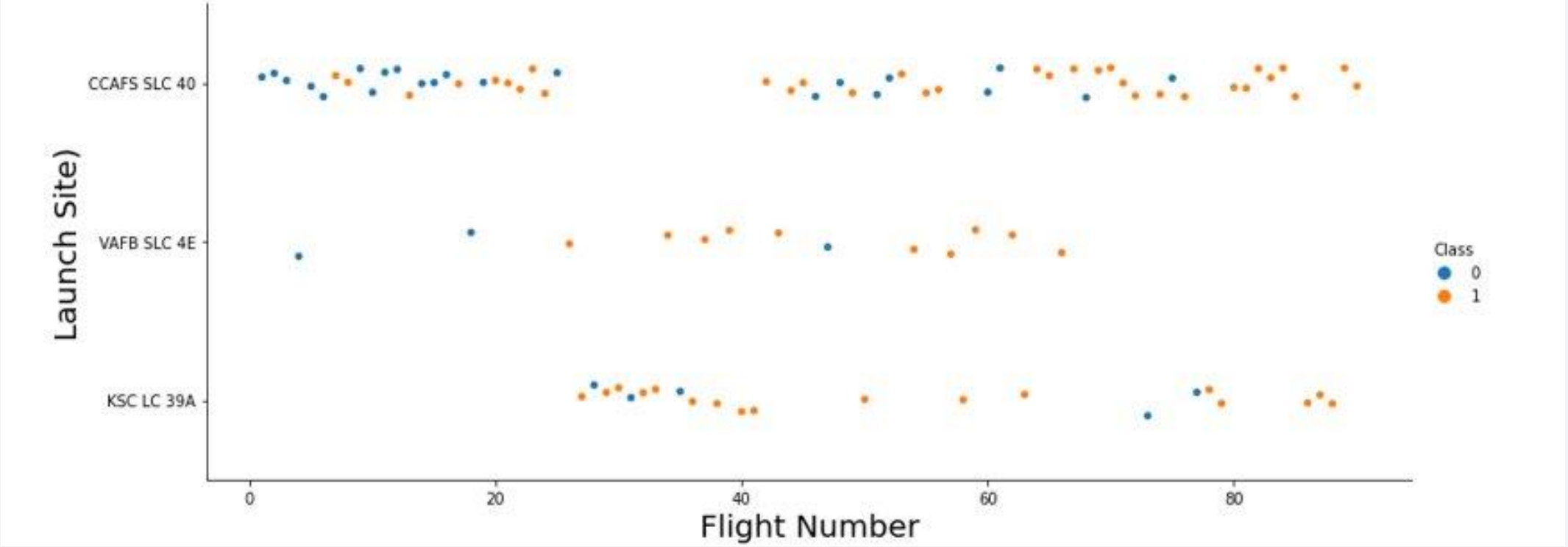
GitHub URL: Lab Machine Learning Prediction lab

# Results

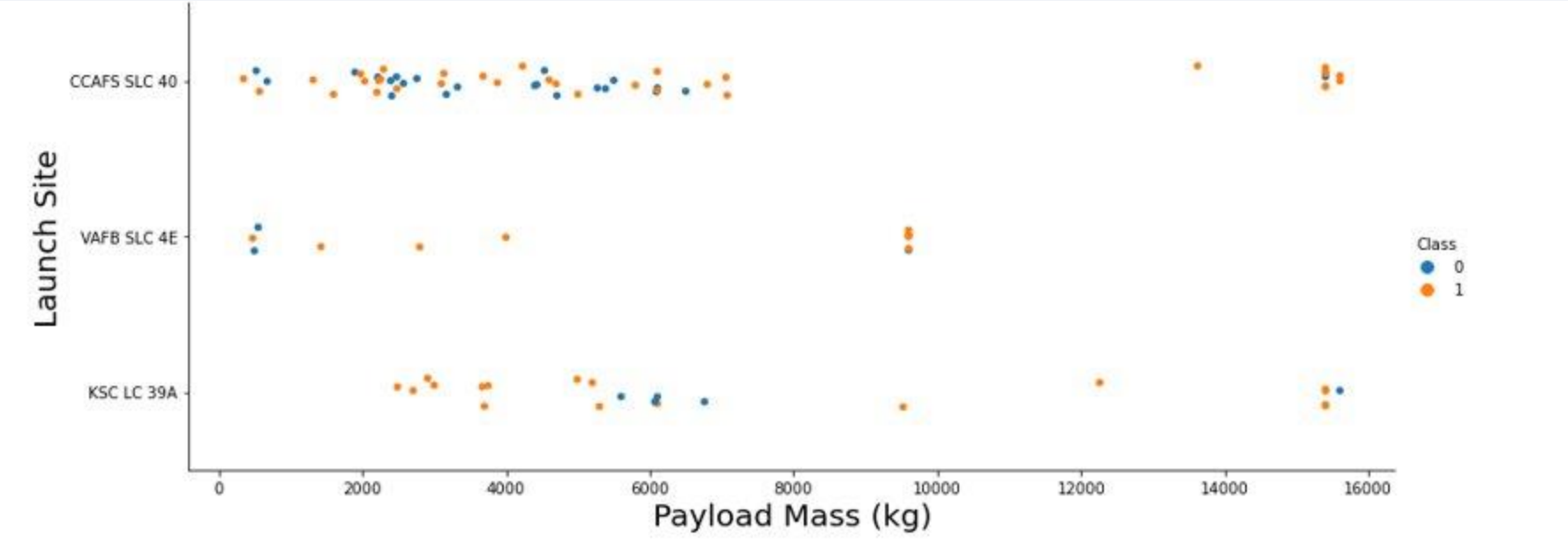
- Exploratory data analysis results
  - Launch success rate increases over time
  - Higher success rate for higher orbits
- Interactive analytics demo in screenshots
  - Higher success rate for higher playload mass
- Predictive analysis results
  - Best prediction results with Logistic Regression and Support Vector Machine


Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site

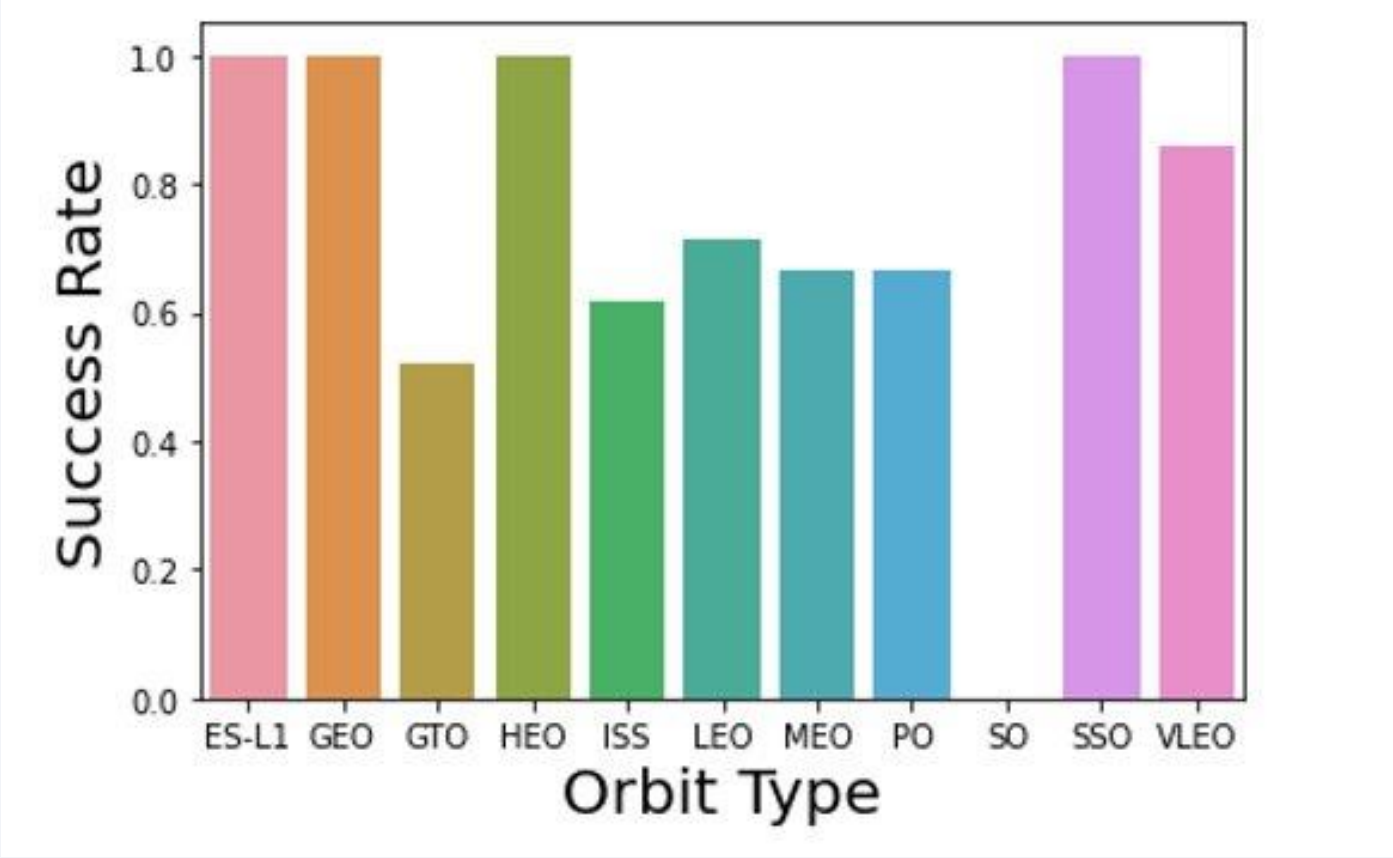


# Payload vs. Launch Site




Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

# Success Rate vs. Orbit Type

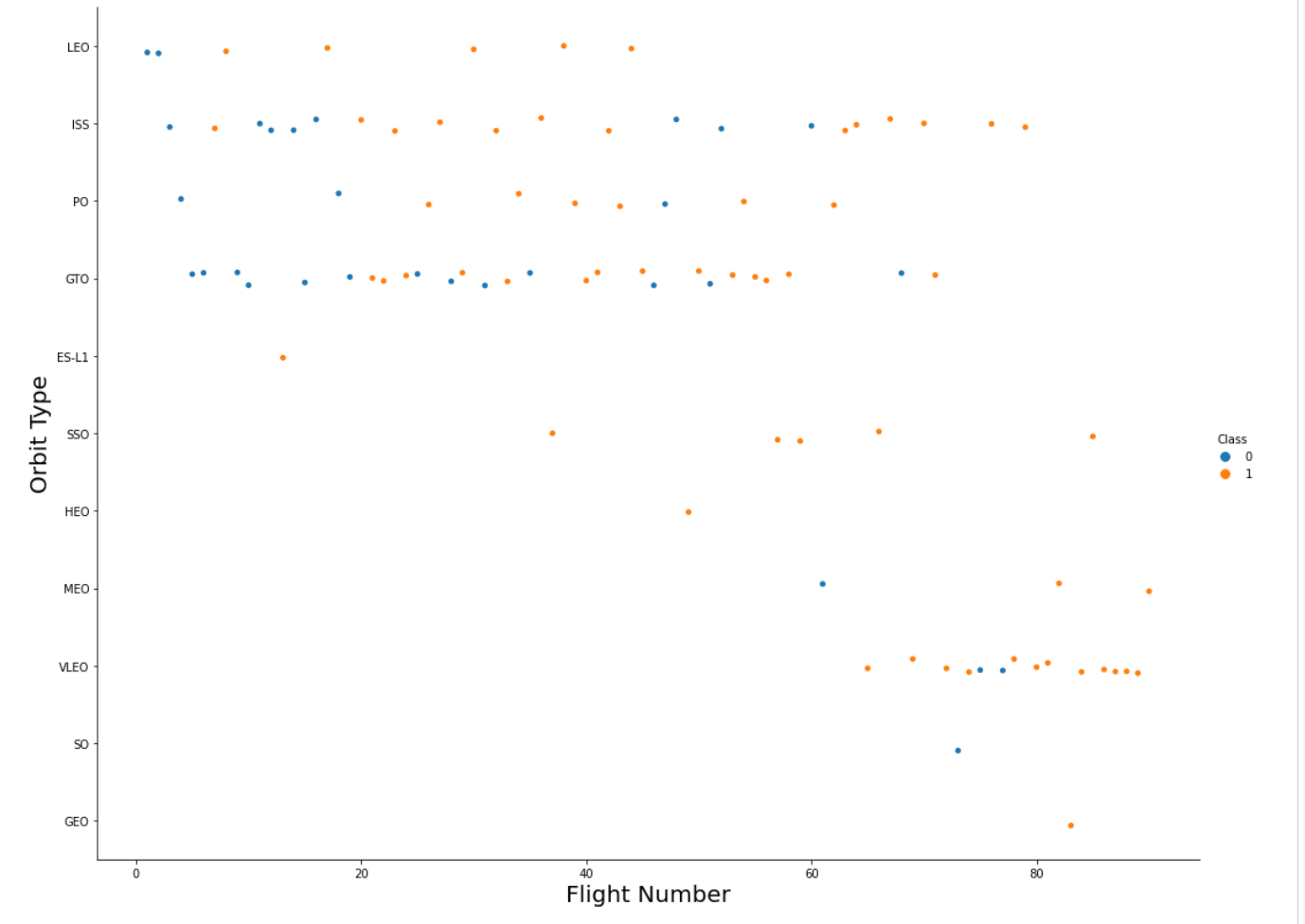



Low Earth Orbits
SSO, HEAO, GEO, ES-LO

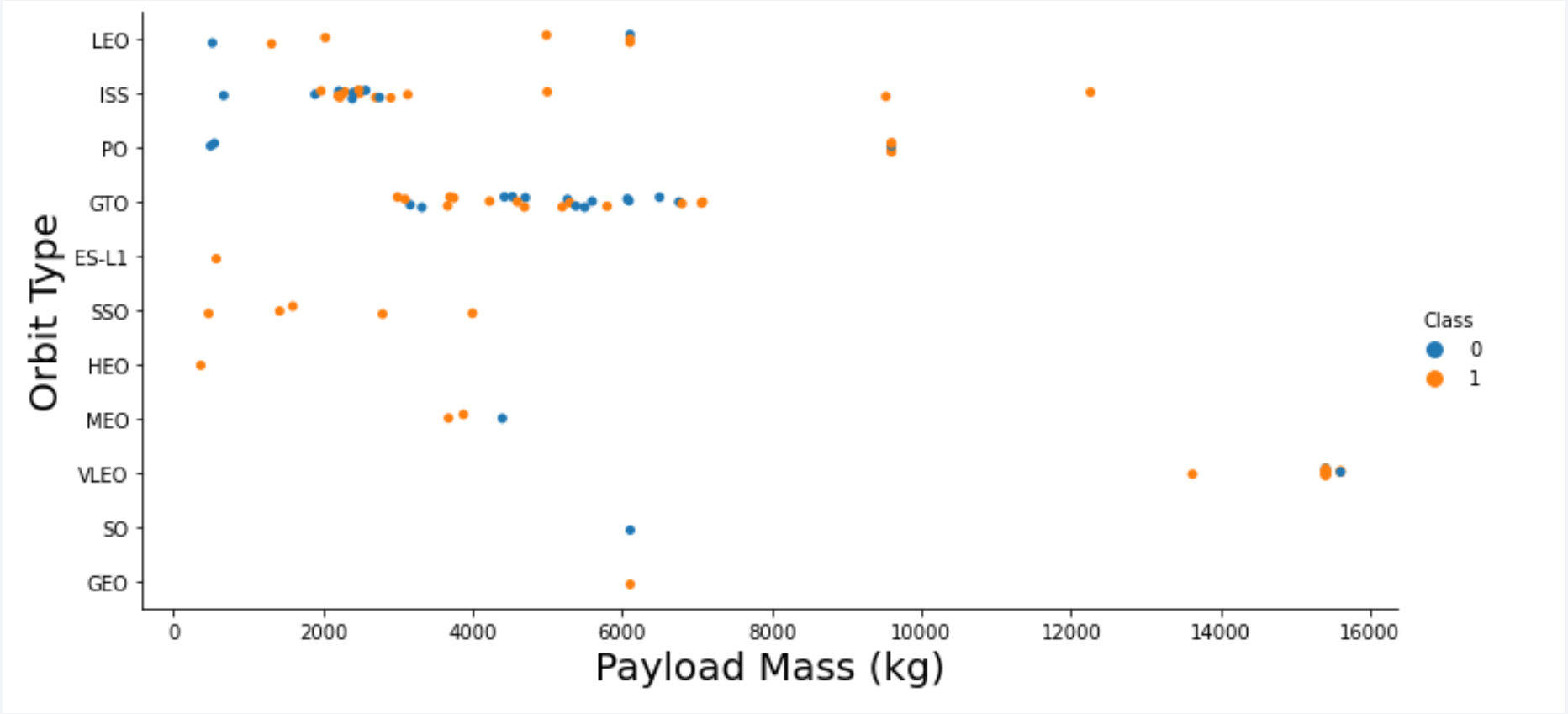
Low Earth Orbits
GTO, ISS, LEO, MEO, PO, VLEO

# Flight Number vs. Orbit Type

In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

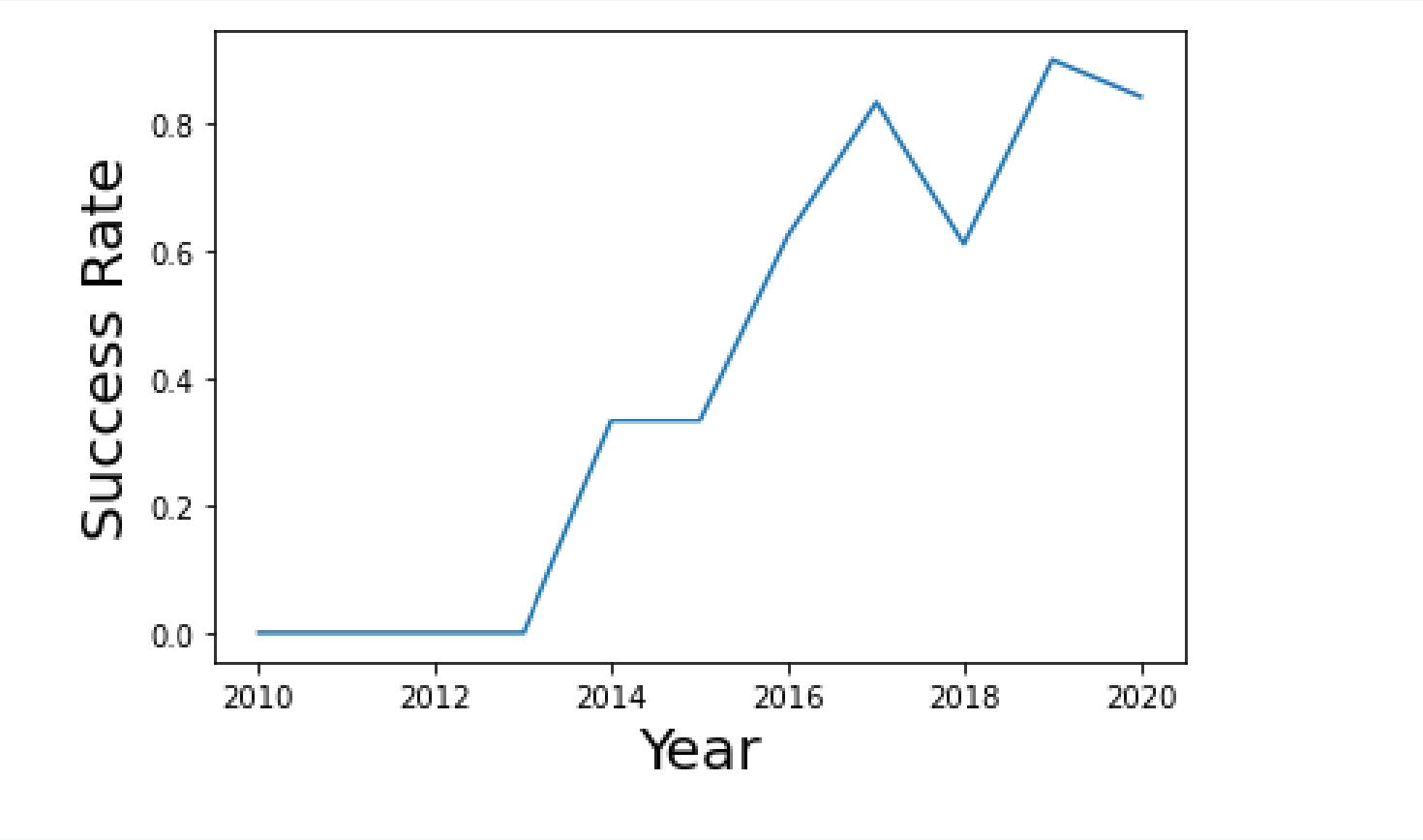


# Payload vs. Orbit Type



# Launch Success Yearly Trend

Launch success has increased over the years



# All Launch Site Names

- the names of the unique launch sites:

```
SELECT DISTINCT launch_site FROM SPACEXTBL;
```

| launch_site |
| --- |
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |

# Launch Site Names Begin with 'KSC'

- 5 records where launch sites' names start with `KSC`

| DATE | time_utc_ | booster_version | launch_site | payload | payload_mass_kg_ | orbit | customer | mission_outcome | landing_outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2017-02-19 | 14:39:00 | F9 FT B1031.1 | KSC LC-39A | SpaceX CRS-10 | 2490 | LEO (ISS) | NASA (CRS) | Success | Success (ground pad) |
| 2017-03-16 | 06:00:00 | F9 FT B1030 | KSC LC-39A | EchoStar 23 | 5600 | GTO | EchoStar | Success | No attempt |
| 2017-03-30 | 22:27:00 | F9 FT B1021.2 | KSC LC-39A | SES-10 | 5300 | GTO | SES | Success | Success (drone ship) |
| 2017-05-01 | 11:15:00 | F9 FT B1032.1 | KSC LC-39A | NROL-76 | 5300 | LEO | NRO | Success | Success (ground pad) |
| 2017-05-15 | 23:21:00 | F9 FT B1034 | KSC LC-39A | Inmarsat-5 F4 | 6070 | GTO | Inmarsat | Success | No attempt |

# Total Payload Mass

**Total payload mass (NASA (CRS)):** 45.596

How we have caculated this:

```
SELECT sum(payload_mass__kg_) AS "Total payload mass (NASA (CRS))"
FROM SPACEXTBL WHERE customer = 'NASA (CRS)';
```

# Average Payload Mass by F9 v1.1

How we have caculated this:

```
SELECT AVG(payload_mass__kg_) AS "Average payload mass (booster version F9 v1.1)" FROM
SPACEXTBL WHERE booster_version LIKE 'F9 v1.1%';
```

**Average payload mass (booster version F9 v1.1): 2.534**

# First Successful Ground Landing Date

**First successful landing outcome in drone ship** 2016-04-08

How we have caculated this:

```
SELECT min(DATE) AS "First successful landing outcome in drone ship" FROM SPACEXTBL
WHERE landing__outcome = 'Success (drone ship)';
```

# 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

| booster_version |
| --- |
| F9 FT B1032.1 |
| F9 B4 B1040.1 |
| F9 B4 B1043.1 |

- Query:

```
SELECT booster_version FROM SPACEXTBL WHERE landing__outcome =
'Success (ground pad)' AND payload_mass__kg_ BETWEEN 4000 AND 6000;
```

# Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

| Outcome | Count |
|---|---|
| Success | 61 |
| Failure | 40 |
| (All) | 101 |

# Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass:

| booster_version |
| --- |
| F9 B5 B1048.4 |
| F9 B5 B1048.5 |
| F9 B5 B1049.4 |
| F9 B5 B1049.5 |
| F9 B5 B1049.7 |
| F9 B5 B1051.3 |
| F9 B5 B1051.4 |
| F9 B5 B1051.6 |
| F9 B5 B1056.4 |
| F9 B5 B1058.3 |
| F9 B5 B1060.2 |
| F9 B5 B1060.3 |

# 2015 Launch Records

Records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

| booster_version | launch_site | landing__outcome |
|---|---|---|
| F9 FT B1031.1 | KSC LC-39A | Success (ground pad) |
| F9 FT B1032.1 | KSC LC-39A | Success (ground pad) |
| F9 FT B1035.1 | KSC LC-39A | Success (ground pad) |
| F9 B4 B1039.1 | KSC LC-39A | Success (ground pad) |
| F9 B4 B1040.1 | KSC LC-39A | Success (ground pad) |
| F9 FT B1035.2 | CCAFS SLC-40 | Success (ground pad) |

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

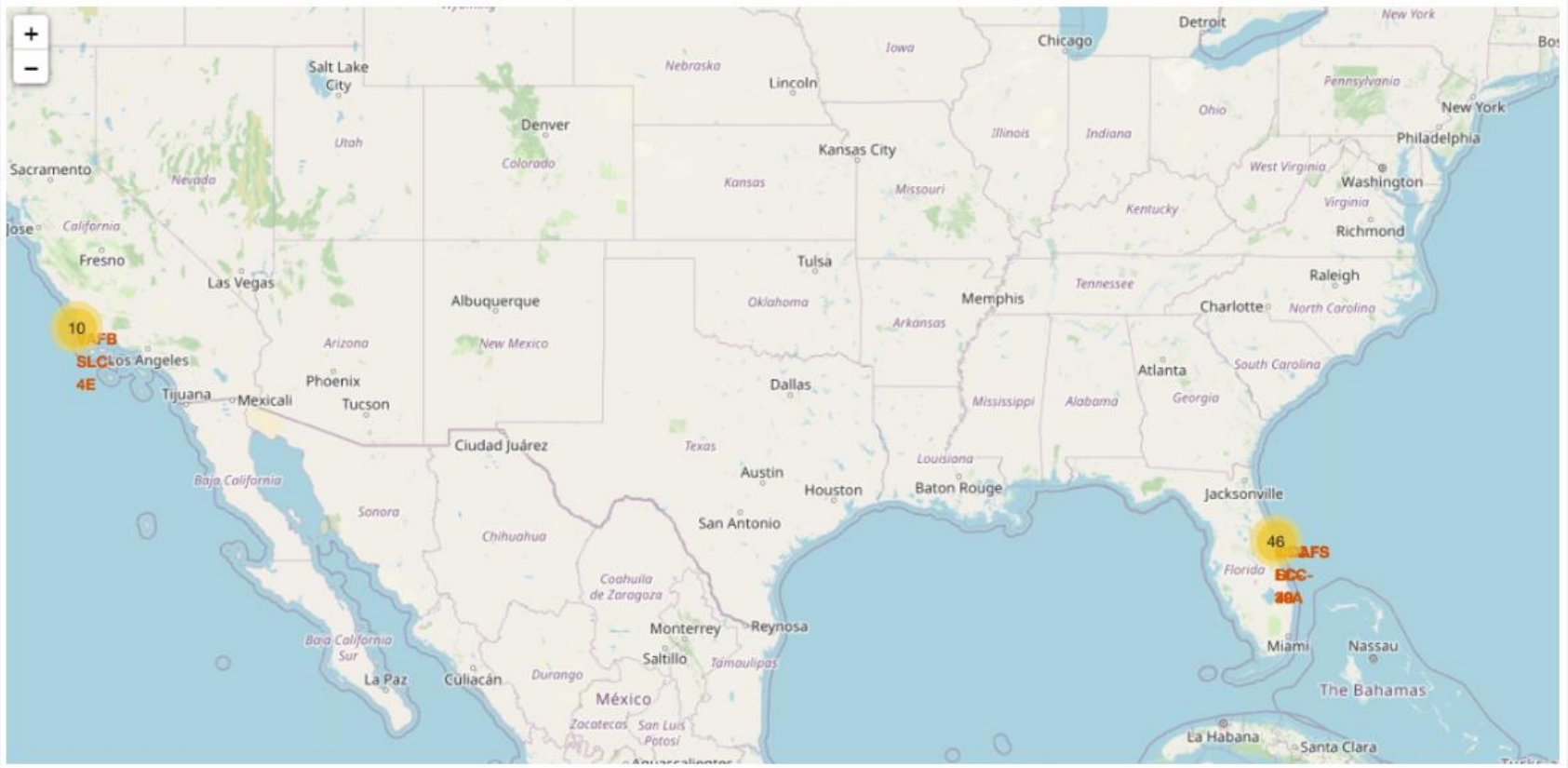
- Rank of the successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

| landing_outcome | Count |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

Section 3

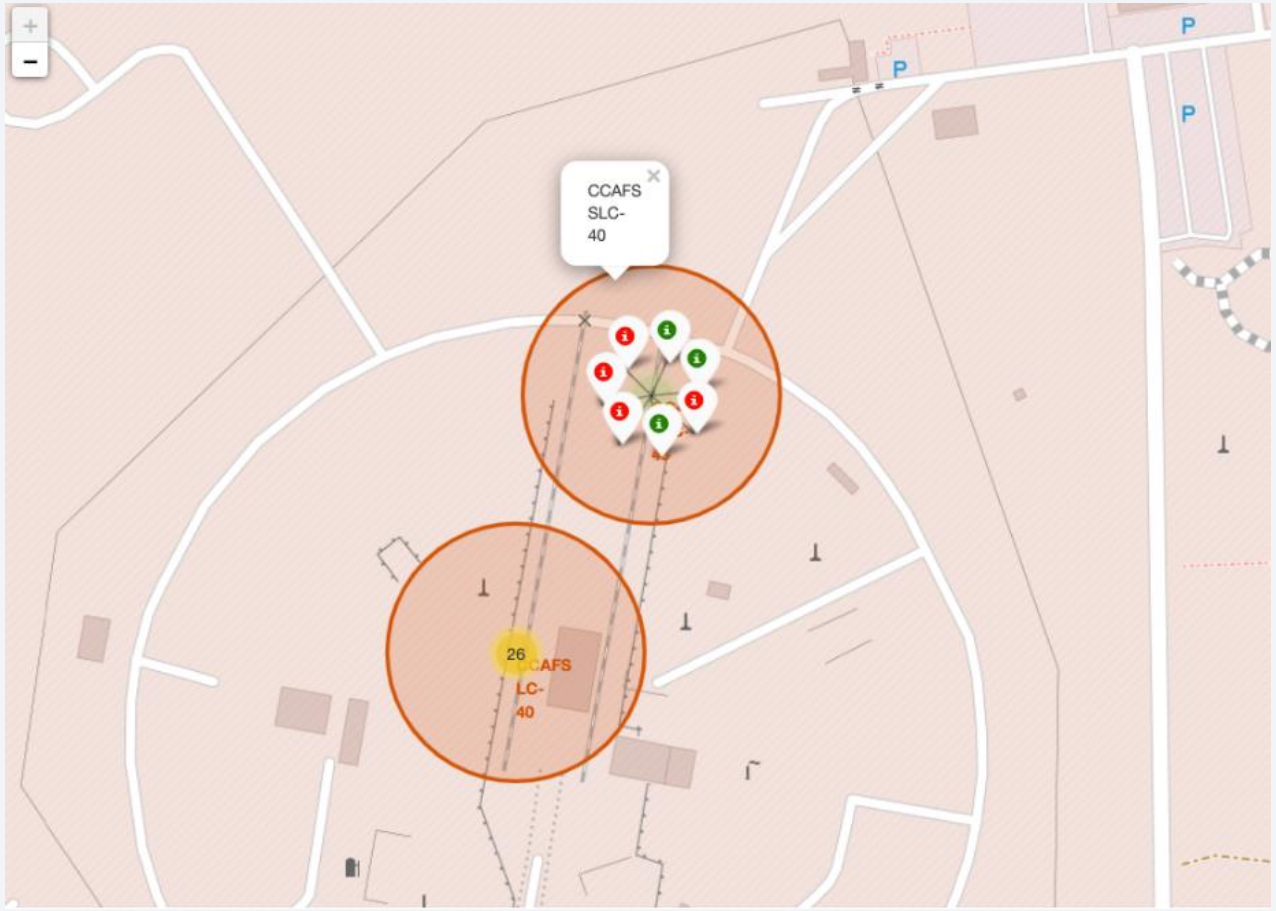
# Launch Sites Proximities Analysis

# Space X Launch Site Locations



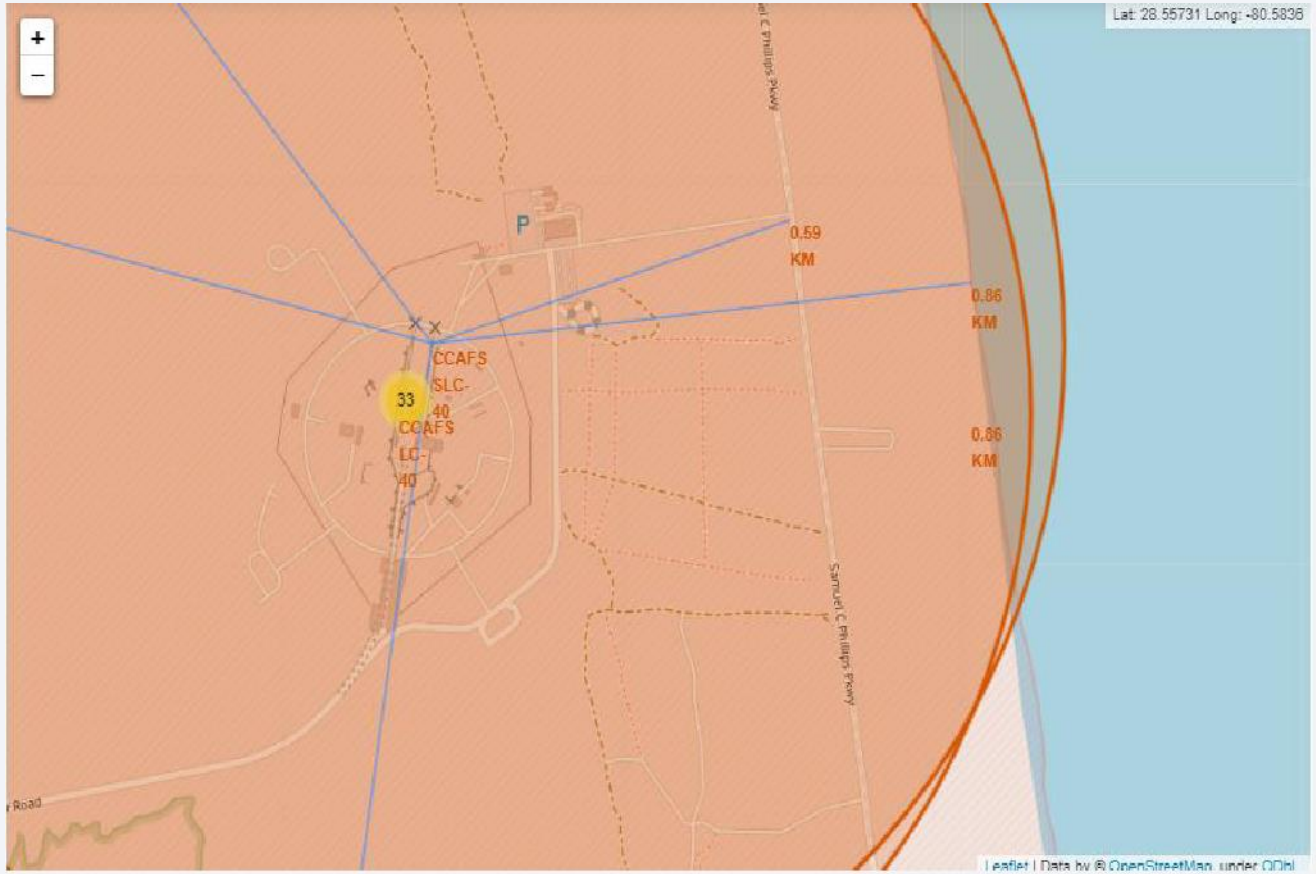

- The Yellow markers are indicators of where the locations of all the SpaceX launch sites are situated, in the US.
- All of them were planned near the coast

# Success or Failure



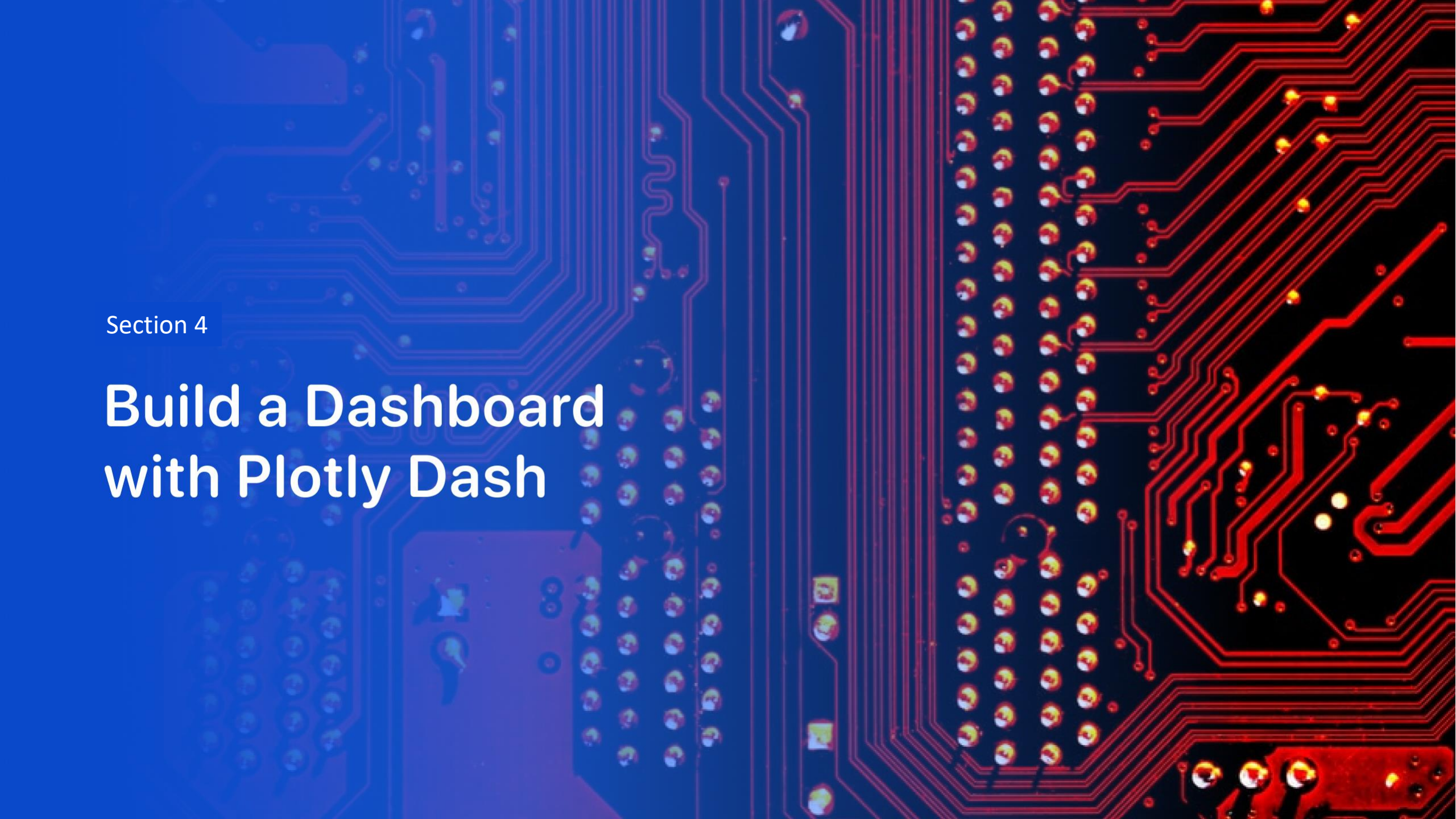

- From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.
- When we zoom in on a launch site, we can click on the launch site which will display marker clusters of:
- successful landings (green),
- or failed landing (red)

# Launch Site Proximities




- The generated map shows that the selected launch site is close to a highway for transportation of personnel and equipment

- The launch site also maintain a certain distance from the cities (can be viewed in notebook)

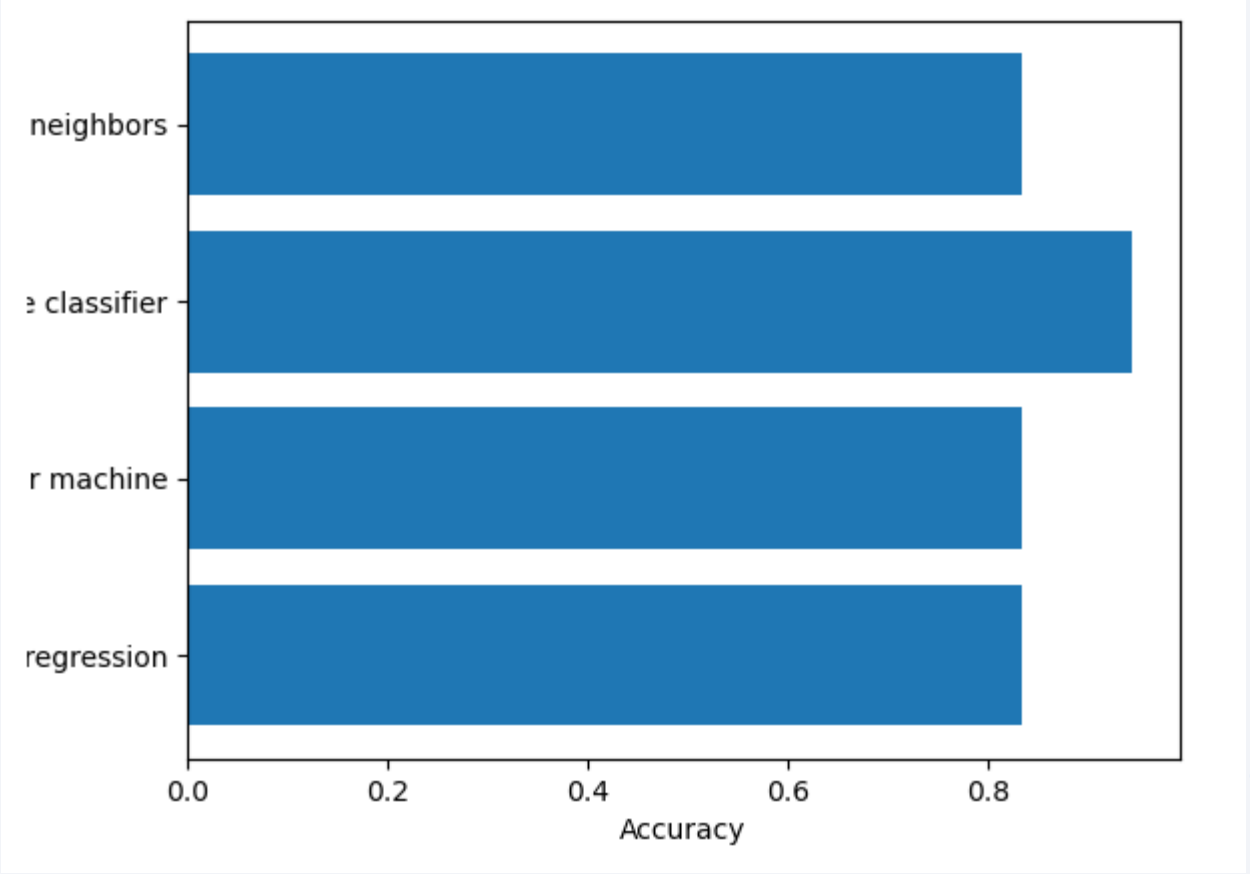

Section 4
Build a Dashboard with Plotly Dash

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- According to the models fit and tested with the split dataset,
- all models performed similarly with 83.33% test set accuracy, except for decision tree
- at 72.22%
- • The logistic regression
- model will be
- considered for
- future slides



# Confusion Matrix

All models except the decision tree performed similarly well

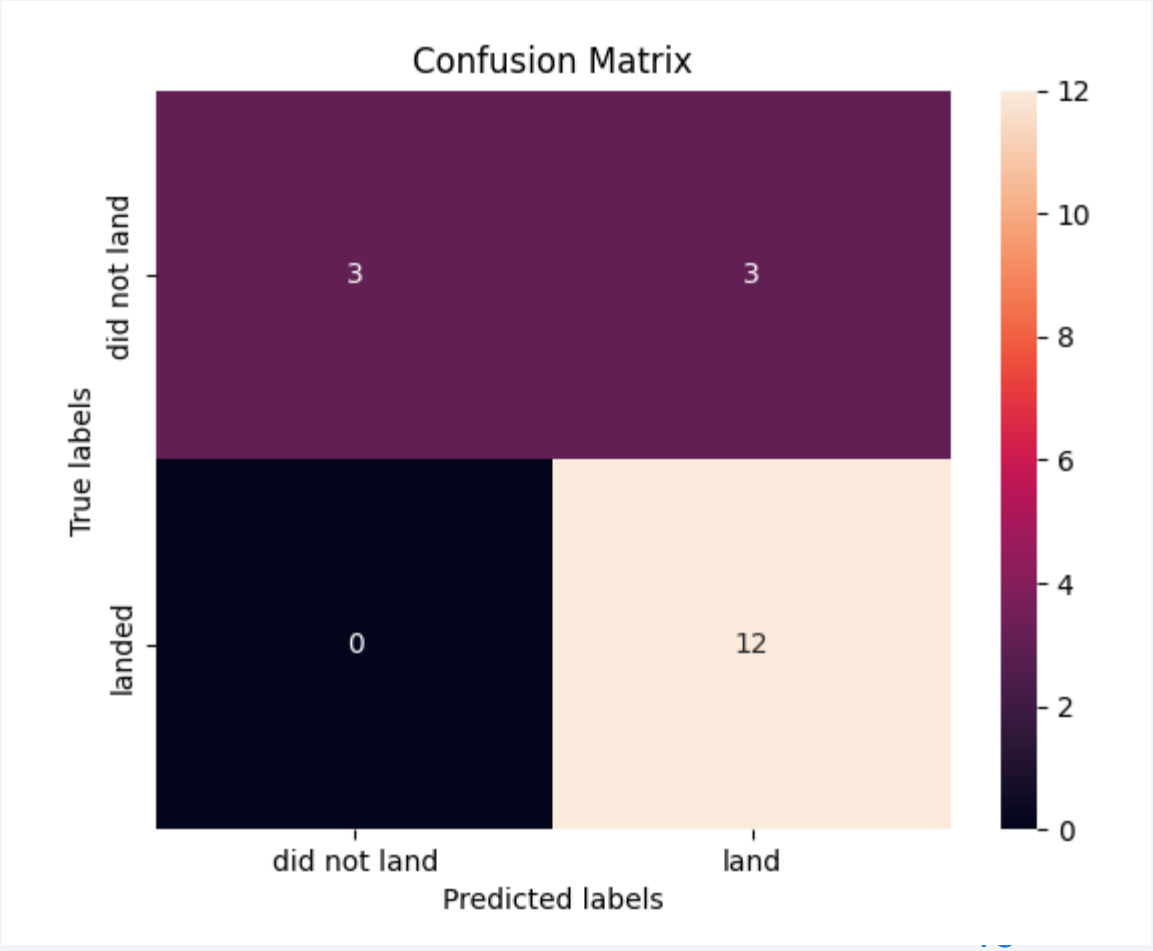
Provided is the confusion matrix for logistic regression

- This matrix was the same SVM and KNN

Generally good performance, but the

false positive rate is unacceptably high

- False positive rate is 50%
- Sensitivity is 100%
- Specificity is 50%



# Conclusions

- Many factors influence rocket science and space mission success
    - Visualizations may assist with determining optimal launch sites
    - Arbitrary SQL queries may not provide much insight
    - Some patterns exist, but are more simple associations to site variables
- rather than actual progress in engineering capabilities and accuracies
    - Factors outside the scope of the available data appear to have greater influence
    - ML methods fit to available data mostly performed similarly well
    - This is resultant of a limited dataset → More data needed

# 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

# Thank you!