IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sam
14 December 2024

Outline

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

Executive Summary

- Summary of methodologies
  1. Data Collection:
     - SpaceX API launch data and web scrapping
  2. Data Wrangling:
     - Clean and format data
     - Store data to database and do SQL queries
     - Standardization
  3. Visualizations:
     - Interactive visualization with Folium
     - Dashboard with Plotly Dash
  4. Model Building & Evaluation:
     - SVM, Decision Trees, K-Nearest Neighbors, GridSearchCV
     - Evaluate models by accuracy

- Summary of Results
  1. Data analysis:
     - Find factors contributing to landings successful rates
     - Visualizations correlations between factors
  2. Model evaluation:
     - Test accuracy of different models
  3. Key findings:
     - Launch location and payload mass influence success rates
     - Decision Tree model perform the best (94% accuracy)

Introduction

Background

SpaceX revolutionized the space industry by offering rocket launches at significantly lower costs, primarily due to their ability to reuse the first stage of the rocket. This project aims to predict the successful landing of the Falcon 9 first stage, helping estimate launch costs and providing insights for companies competing with SpaceX.

Problems to be included

1. Factors that determine successful landings
2. Relationship between factors and its importance
3. Best machine learning model that performs best prediction

Section 1

Methodology

Methodology

Executive Summary:

Data collection methodology: SpaceX API and web scrapping

Perform data wrangling: Handle missing values, standardizing data, extract and create new features.

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models:

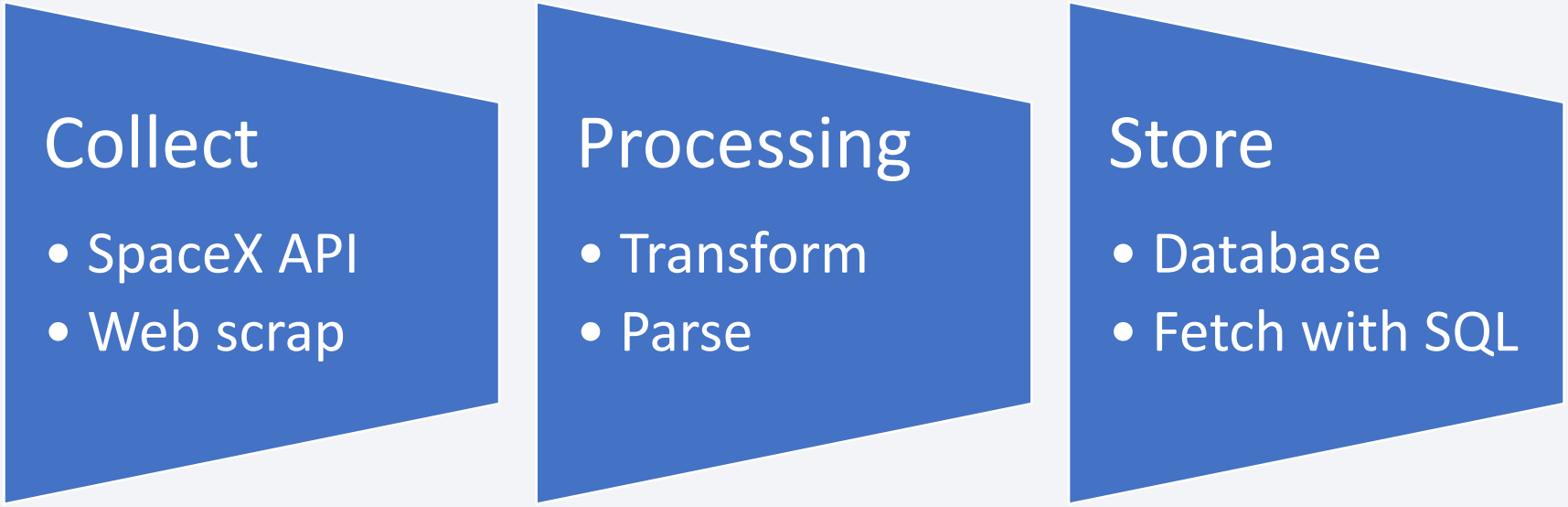
Built models like Logistic Regression, SVM, KNN and Decision Trees.

Use GridSearchCV for hyperparameter tuning

Data Collection

Data is collected by SpaceX REST API and Web Scrapping from Wikipedia.

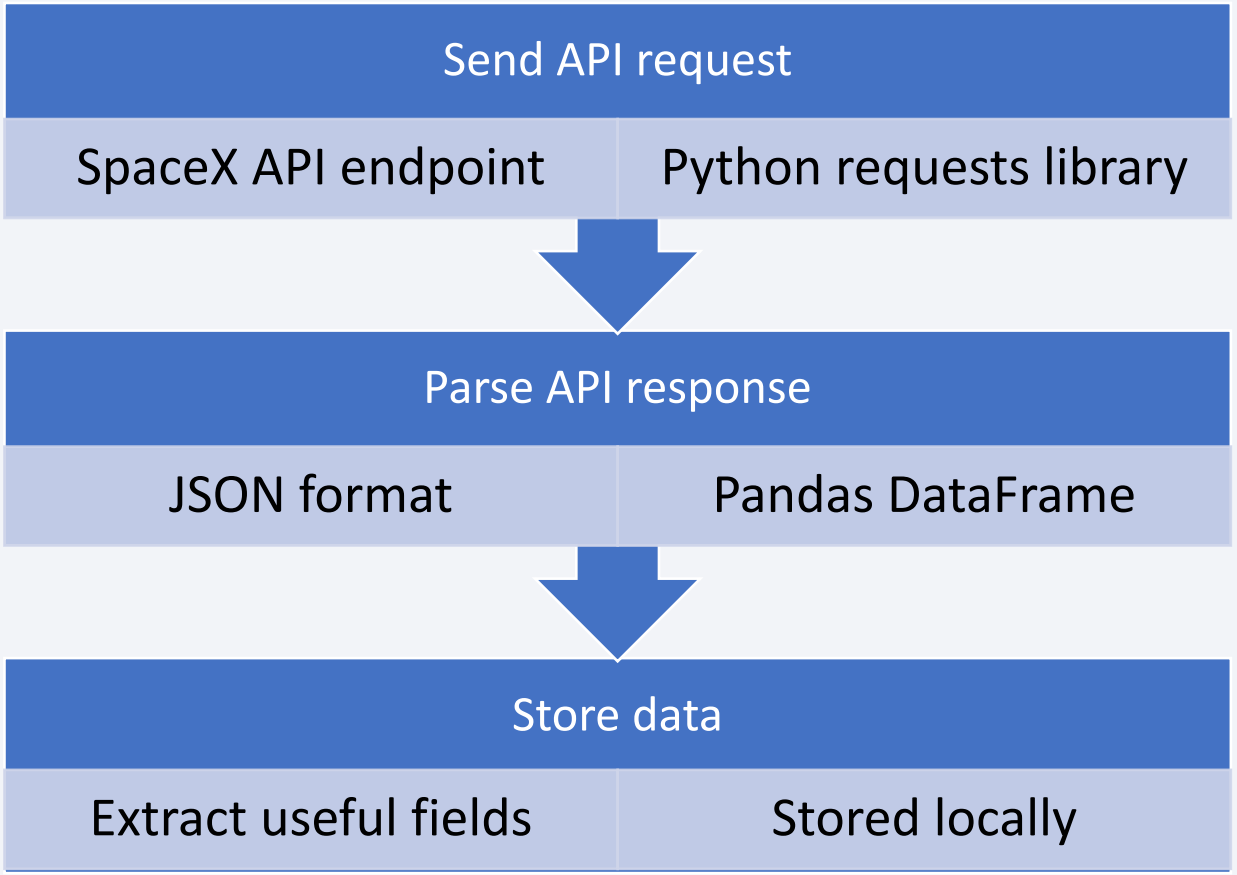
Request is made follow by data fetching and parsing. Those collected data is then streamlined to perform cleaning and transformation. In collection process, it is important to keep data structured well organized to avoid revisiting collection process during data analysis steps.



Data Collection – SpaceX API

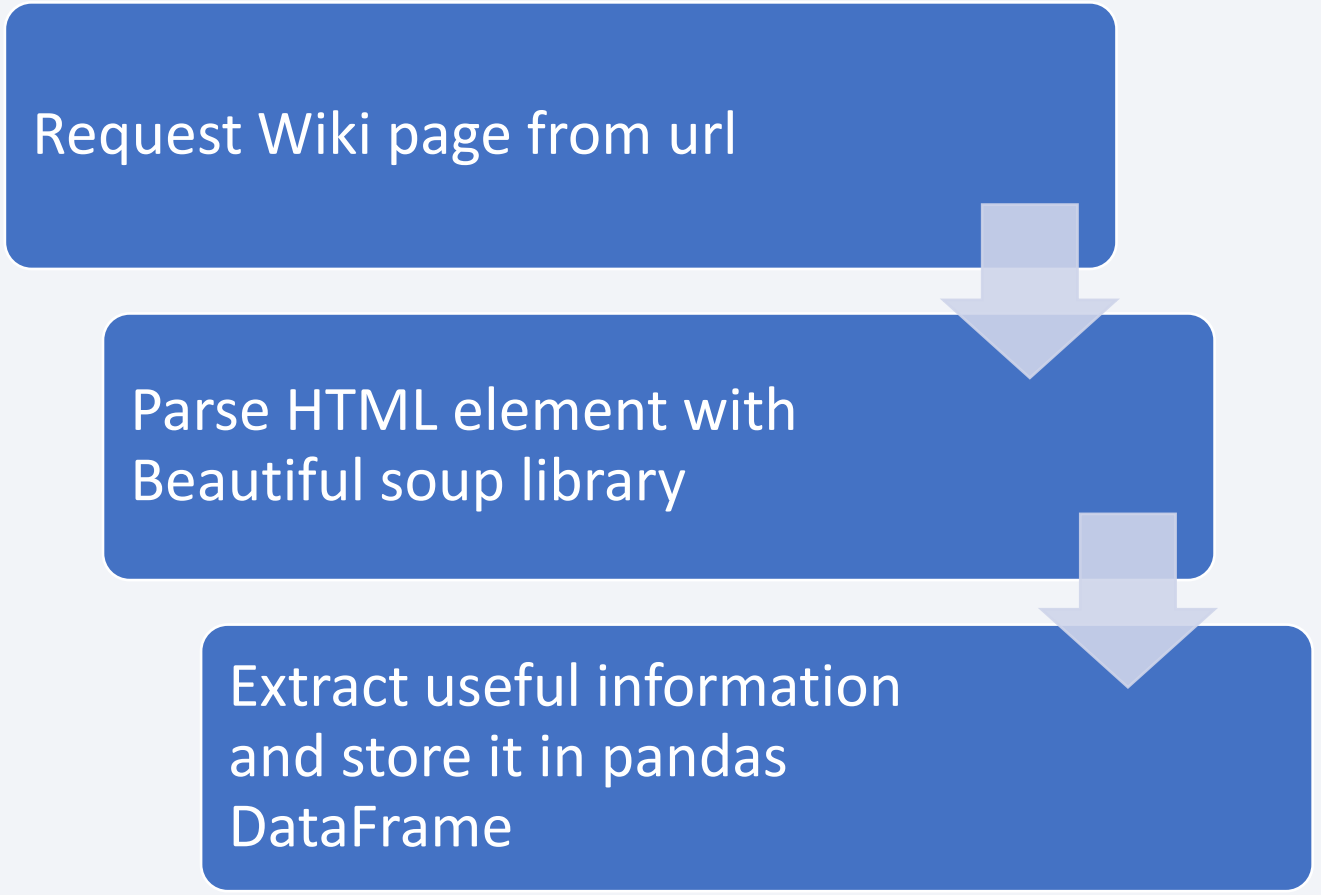
1. Send API request
   - Python requests library allow user to send get request to end point which is connected to SpaceX API
2. Parse response
   - Received response will be parsed to a pandas dataframe in JSON format
3. Store data:
   - Extract useful fields and saved it locally

Github: https://github.com/SamLo322/course-project-AppliedDataScienceCapstone/blob/be484cdbcc510814d5deffea5e1b46fde9dcea87/1.%20jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

1. Send API request
   - Python requests library can fetch target page HTML content
2. Parse response
   - Python Beautiful soup library can parse HTML content to capture useful information
3. Store data
   - Extract useful data and store in Pandas DataFrame

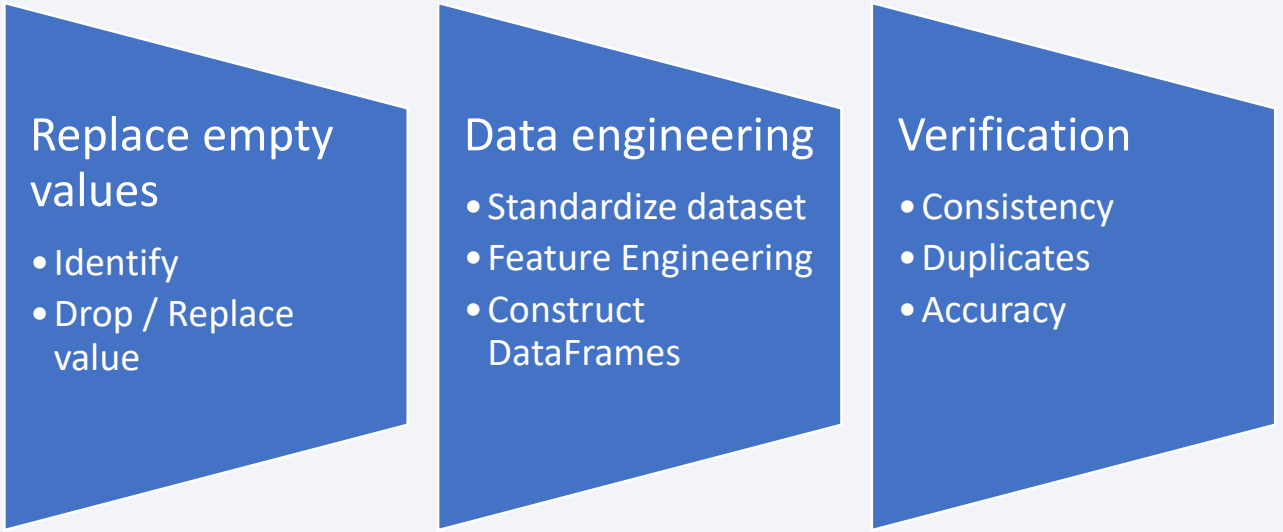


Github: https://github.com/SamLo322/course-project-AppliedDataScienceCapstone/blob/be484cdbcc510814d5deffea5e1b46fde9dcea87/2.%20jupyter-labs-webscraping.ipynb

Data Wrangling

Data Wrangling process is to transform and clean messy and unstructured data into useful form that can be carried forward for exploratory data analysis.

1. Data Cleaning
   - Replace nan values with mean value of dataset
   - Extract useful information by dropping rows and columns
2. Data integration and transformation
   - Standardize column format into DataFrames object
   - Integrate DataFrames object from different data source
3. Data Validation
   - Remove duplicate recores and perform consistency check



Github: https://github.com/SamLo322/course-project-AppliedDataScienceCapstone/blob/be484cdbcc510814d5deffea5e1b46fde9dcea87/3.%20labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Histograms – Visualize statistic such as launch success rates and payload mass, it helps generate data insights such as tendency and inter-category comparisons
  - Bar Charts – Visualize launch outcomes across launch sites or rocket types to compare success rates and find best type
  - Line Charts – Visualize success rate over time, time series plot over period helps identify trends
  - Scatter Plots – Show relationship across two dimension to perform comparison and find dependencies and correlations
  - Heatmaps – show matrices across variables that help identify correlation trends

Github: https://github.com/SamLo322/course-project-AppliedDataScienceCapstone/blob/be484cdbcc510814d5deffea5e1b46fde9dcea87/5.%20edadataviz.ipynb

EDA with SQL

SQL Queries:

- Grab success and fail launches, get average or total data grouped by launch site and rocket type
- Filter data by launch outcomes, dates, payload masses and rocket types
- Sort and order data by wanted means such as date or type
- Nested queries to find max or average statistic such as average payload mass grouped by launch site

Github: https://github.com/SamLo322/course-project-AppliedDataScienceCapstone/blob/be484cdbcc510814d5deffea5e1b46fde9dcea87/4.%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

In order to visualize rocket launch location, we capture the launch latitude and longitude coordinates and mark them on the interactive map using Folium

Map objects plotted:

Markers:

- Outline the exact launch locations of rocket launch in order to give clear image of where launch location is distributed over the past period

Circles:

- Outlines the areas that is critical to ensure controllability over rocket launch such as safety and operational needs

Lines:

- Highlights the distance between launch locations and connected nearby areas to enhance understanding from an overview

Github: https://github.com/SamLo322/course-project-AppliedDataScienceCapstone/blob/be484cdbcc510814d5deffea5e1b46fde9dcea87/6.%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Interactive dashboard of Pie Chart and Scatter plots

- Pie Chart
  - Shows success and fail launches
- Scatter Plots
  - Shows correlations of payload mass and success rates to give insight where threshold shall be set

Github: https://github.com/SamLo322/course-project-AppliedDataScienceCapstone/blob/47184b02f92bb4f0192dfb308b0f9f3c34bde3c0/spacex_dash_app.py

Predictive Analysis (Classification)

Construct model

- Input dataset to DataFrame
- Transform data into trainable format
- Split dataset into training and testing dataset
- Pick Machine Learning model to use with

Model Evaluation

- Test the accuracy scores with the report and confusion matrix
- Rank the accuracy across each models
- Adjust hyperparameters of each model

Pick best model

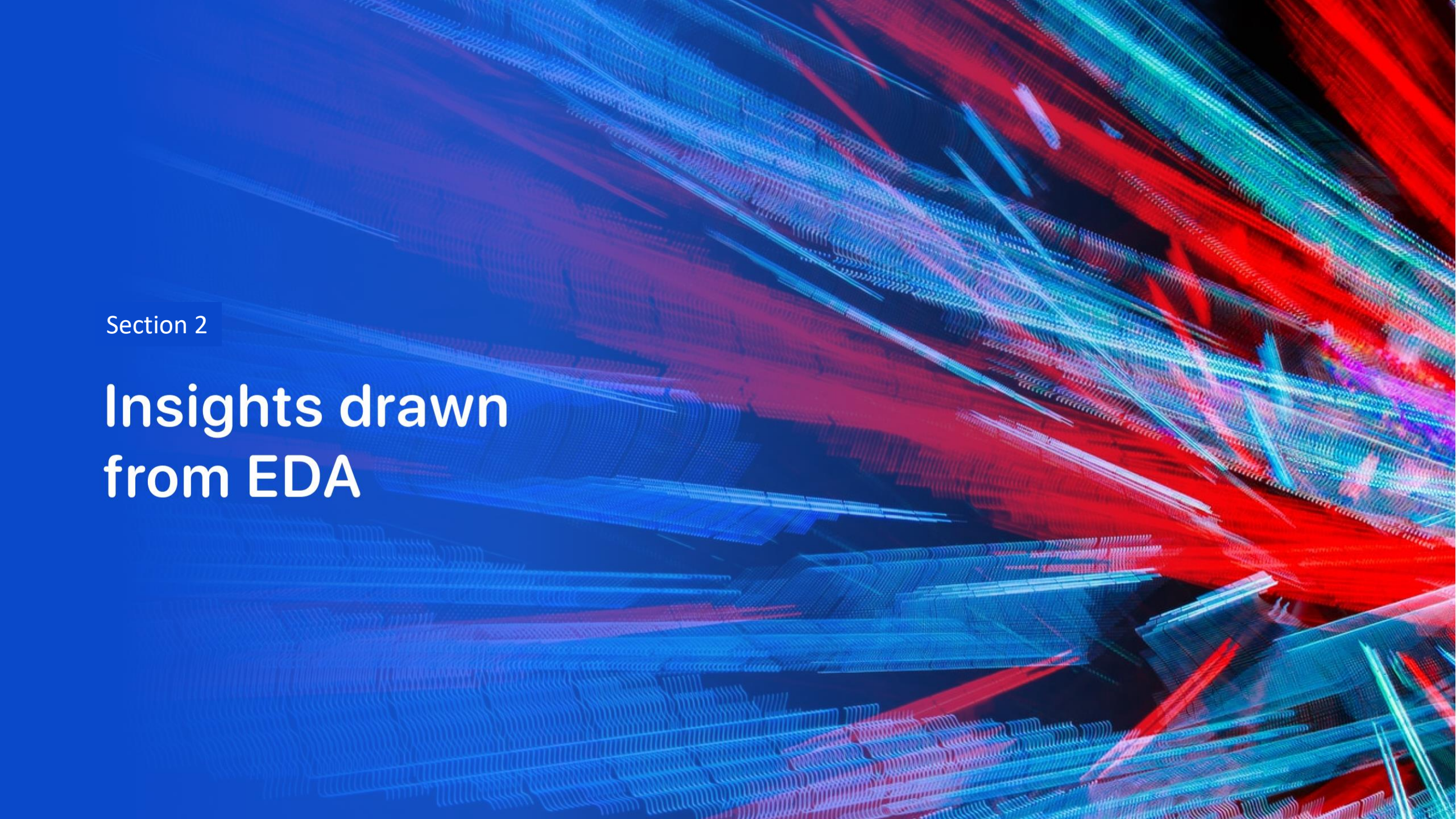
- Feature engineering and parameters tuning of each model
- Select the model with best accuracy for this dataset

Github: https://github.com/SamLo322/course-project-AppliedDataScienceCapstone/blob/47184b02f92bb4f0192dfb308b0f9f3c34bde3c0/7.%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

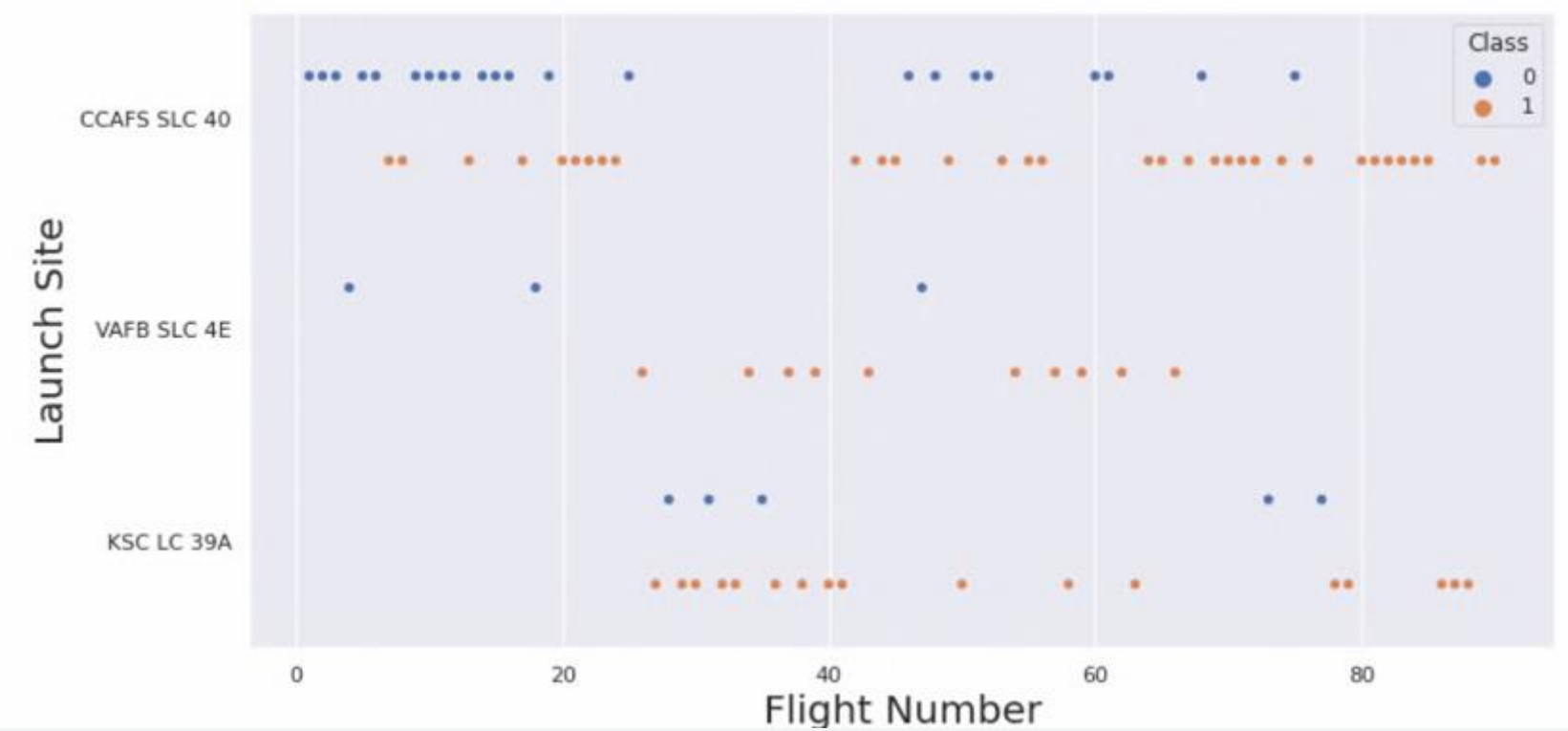
Results

Results is then displayed as

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Blue dot indicate unsuccessful landings while orange ones indicate successful landings.

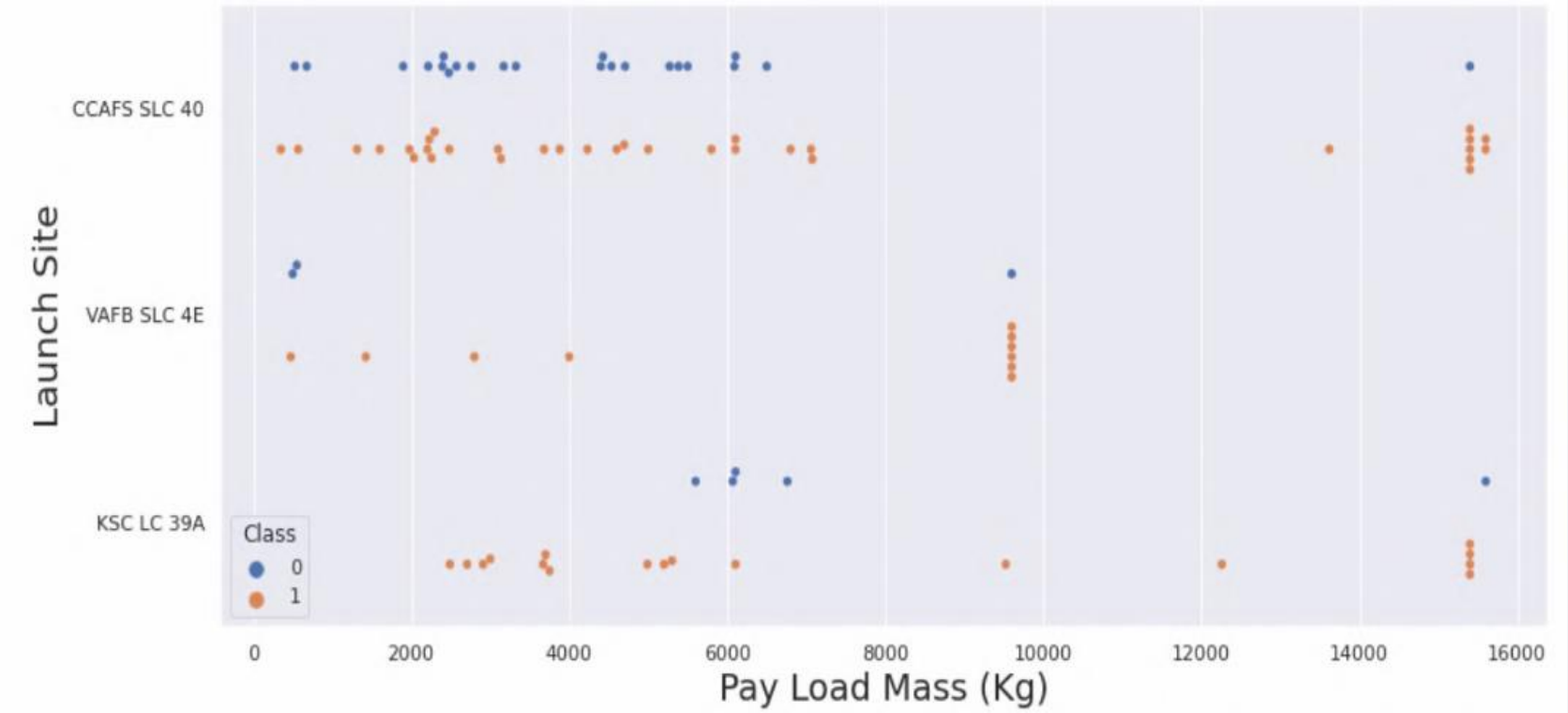
On the Y-axis it shows the distribution of successful rates over different launch sites.

Combine both shows that successful rates increase as number of launches is sent in each site

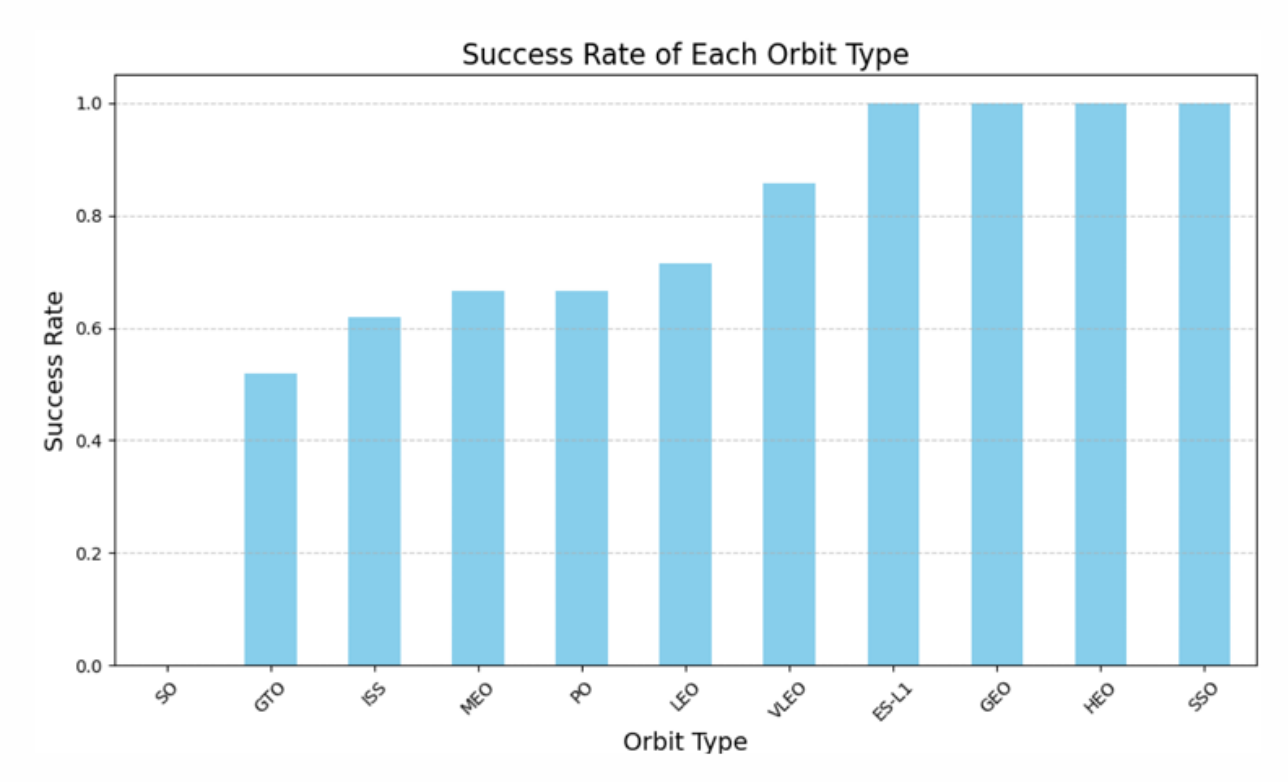
Payload vs. Launch Site

This plot show that the successful rate of launches increased with payload over 9000 kg.

Besides payload below 4000 kg on KSC LC 39A site has a 100% successful rate



Success Rate vs. Orbit Type



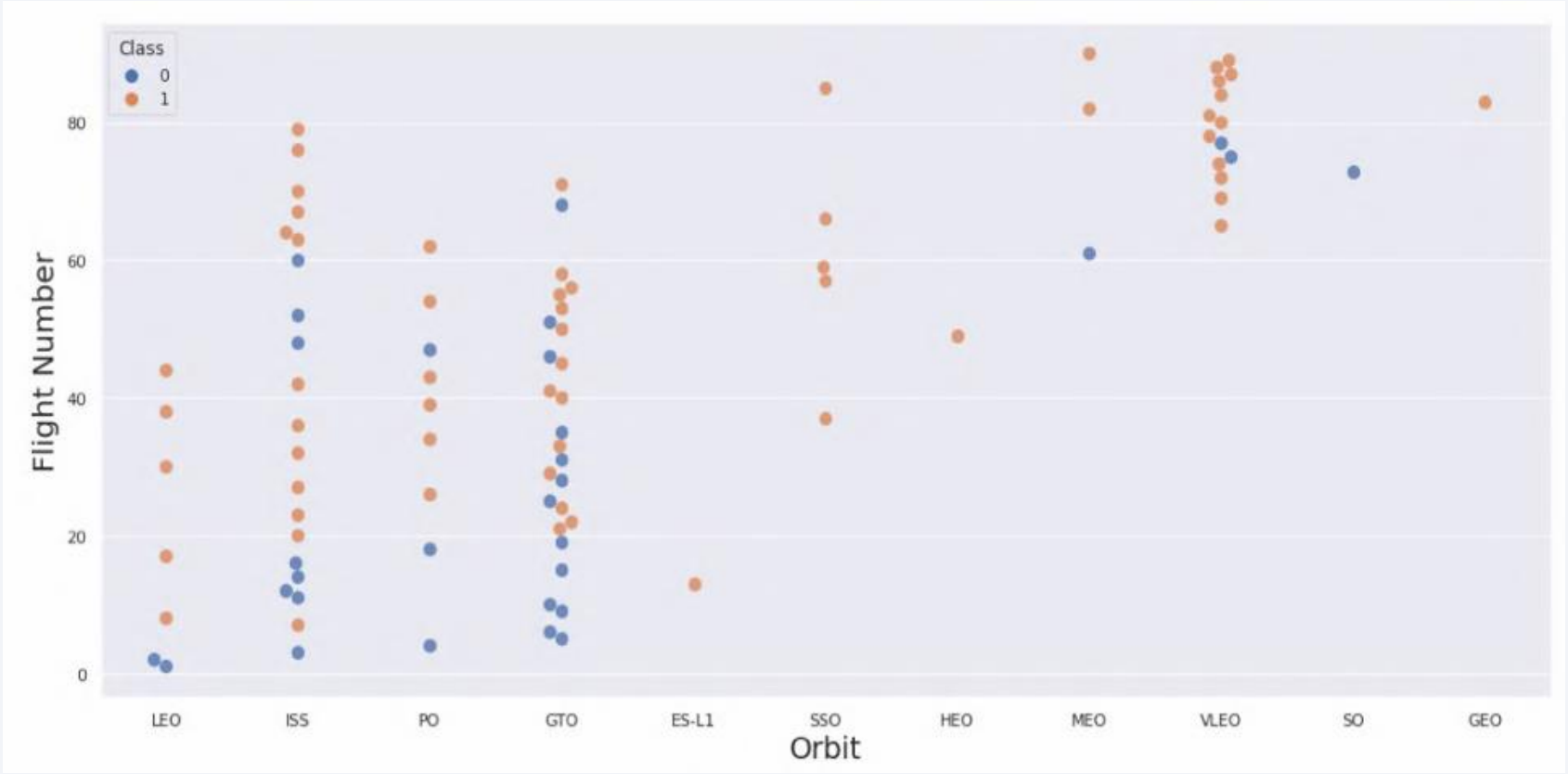
This plot shows that orbit type highly influence success rate, and achieve 100% successful rate for SSO.

However should bare in mind spots like GEO, SO, and SEO only has 1 occurrence indicating that the figure is biased at the moment

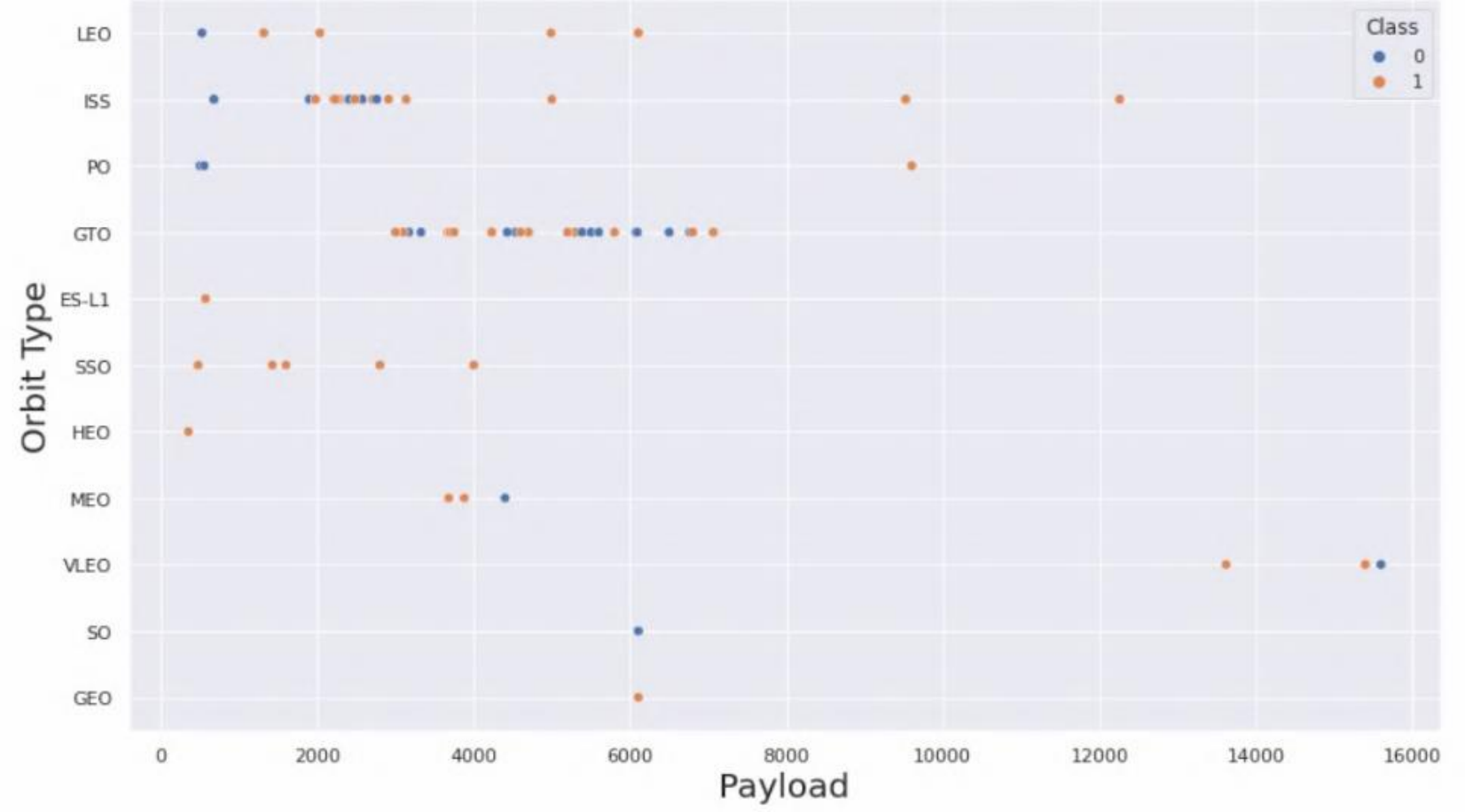
Flight Number vs. Orbit Type

The plot show that flight is shifted from LEO ISS and GTO to VLEO, which has a higher successful rate.

ISS and GTO has the most occurrence with a trend of increasing successful rates.



Payload vs. Orbit Type



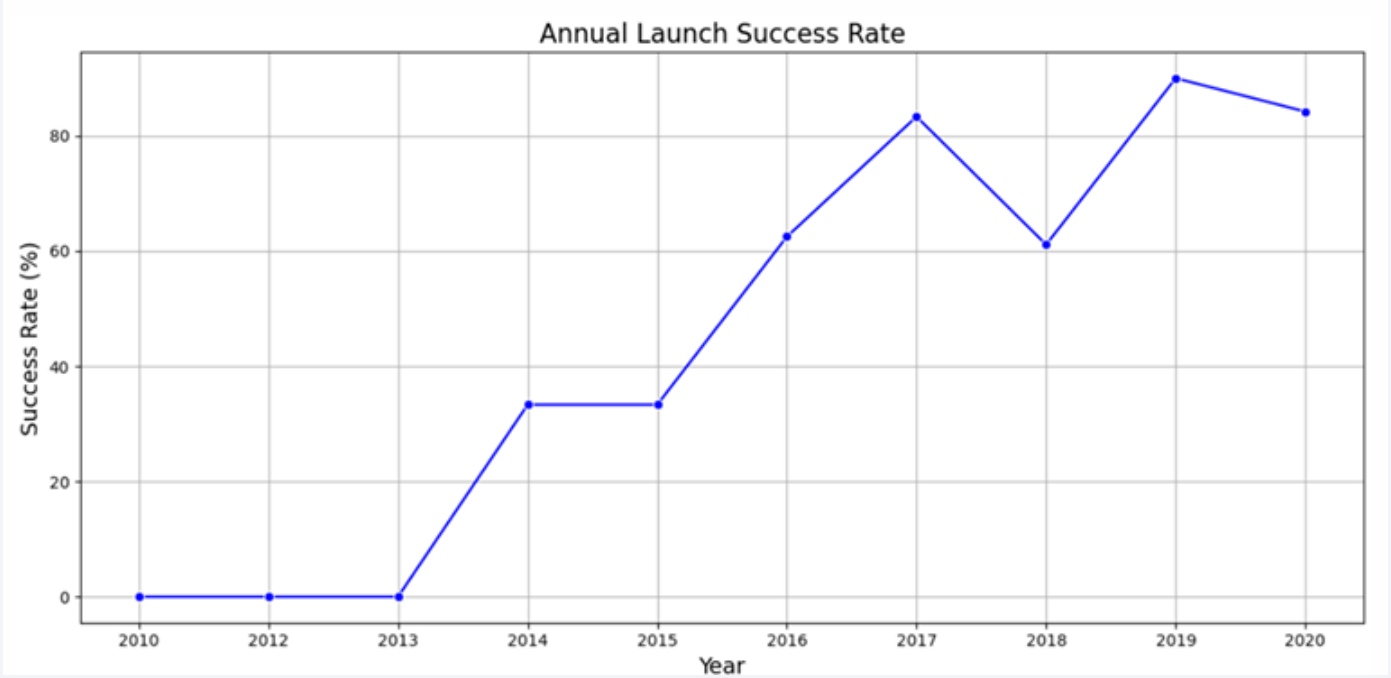
This plot show relationship between Payload and Orbit type.

Figure shows an unequal distribution of orbit types and launches, as well as the Payload.

For instance SSO is relatively more good at launching low Payload with GTO and ISS is good at 4000~6000 and 2000~3000 Payload respectively

Launch Success Yearly Trend

This time series plot shows the average launch success rate over each year, this shows that improvement and optimization from the engineering team has did a good job in optimizing launch steps and aim for a higher successful ate.



All Launch Site Names

The SQL query used to show unique launch sites:

```
%sql select distinct Launch_Site from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

SQL to find first 5 records with launch sites begin with `CCA`:

The SQL use limit use “like” keyword to match similar wordings and “limit” to specify 5 records

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
[11] 1 %sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total payload carried by boosters from NASA

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[13] 1 %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1 by filtering booster version and use avg() keyword on payload mass column

Task 4

Display average payload mass carried by booster version F9 v1.1

```
[18] 1 %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Dates of the first successful landing outcome on ground pad:

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[21] 1 %sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success%'
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Successfully landed booster with payload mass between 4000 and 6000

Use keyword between, “and” with specified landing outcome.

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

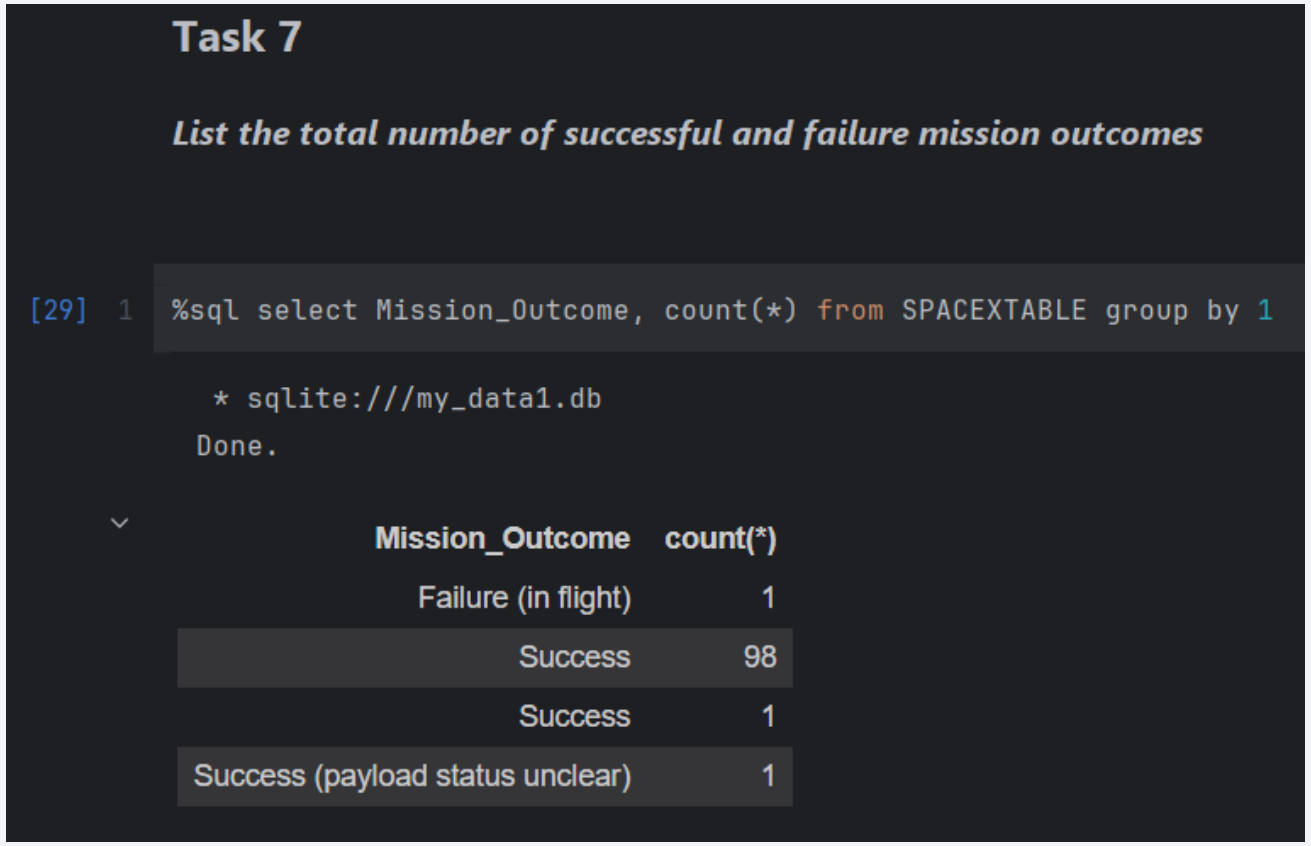
```
[22] 1 %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ between 4000 and 6000 and Landing_Outcome = "Success (drone ship)"
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes. Using count() keyword and group keyword to calculate numbers of individual outcomes



Task 7

List the total number of successful and failure mission outcomes

```
[29] 1 %sql select Mission_Outcome, count(*) from SPACEXTABLE group by 1
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

Names of the booster with maximum payload mass using subquery that gets the max payload for filtering

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[30] 1 %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

Failed landing outcomes in year 2015 with specified columns and date requirements

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
[31] 1 %sql select substr(Date, 6,2), Landing_Outcome from SPACEXTABLE where Landing_Outcome = "Failure (drone ship)" and substr(Date,0,5)='2015'
```

```
 * sqlite:///my_data1.db
Done.
```

substr(Date, 6,2)	Landing_Outcome
01	Failure (drone ship)
04	Failure (drone ship)

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

Rank number of landing outcomes between 2010-06-04 and 2017-03-20.

Use Count, where, group by and order by keywords to formulate this result

```
%sql SELECT LANDING__OUTCOME as "Landing Outcome", COUNT(LANDING__OUTCOME) AS "Total Count" FROM SPACEX \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC ;
```

Landing Outcome	Total 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

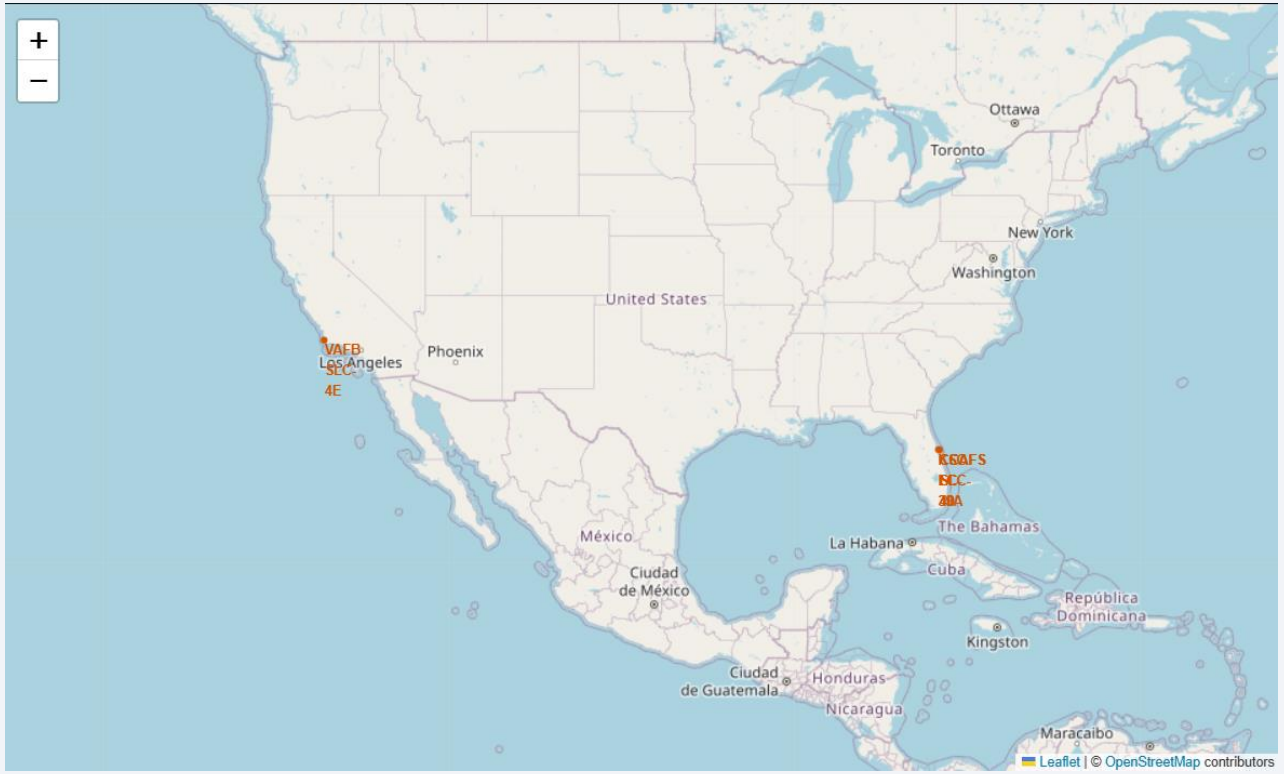
Section 3

Launch Sites Proximities Analysis

Location of launch sites in USA

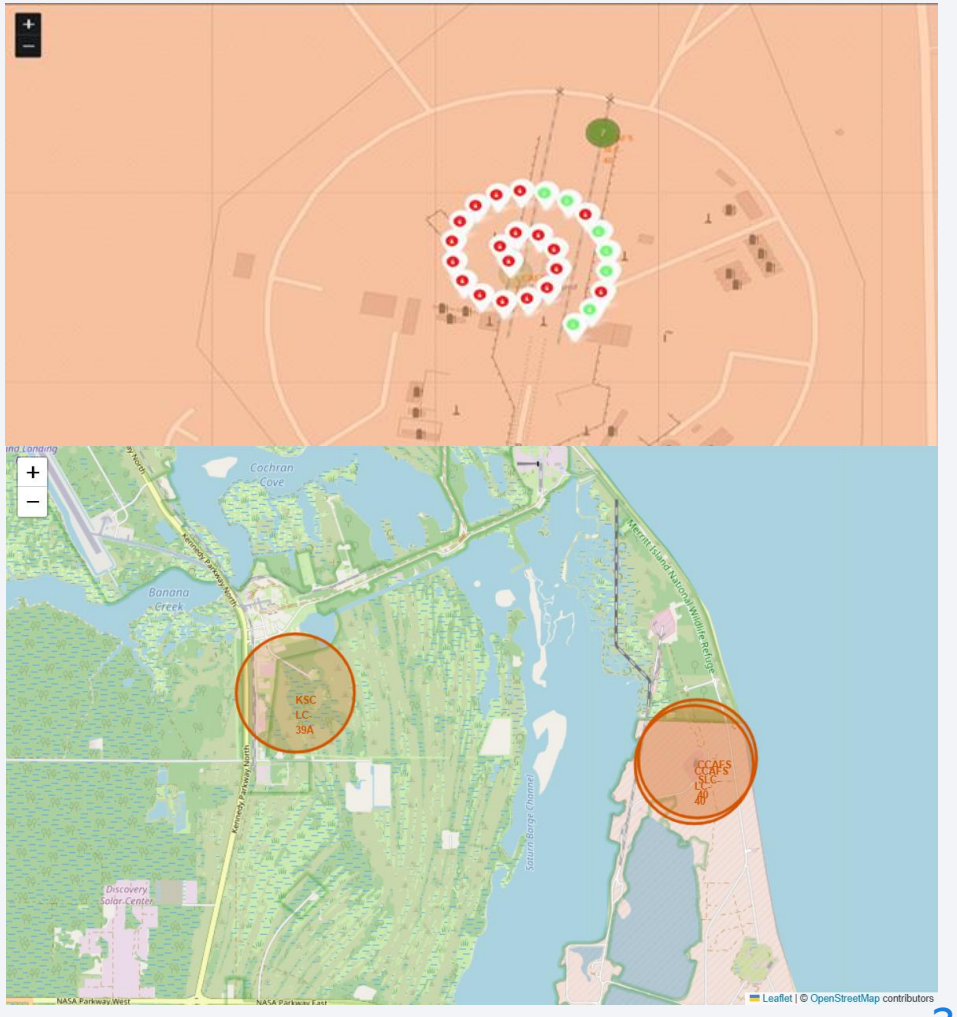
The map includes all launch sites locations.

All launch site is located near the coastal area and with developed communication infrastructure



Launch outcomes for each site

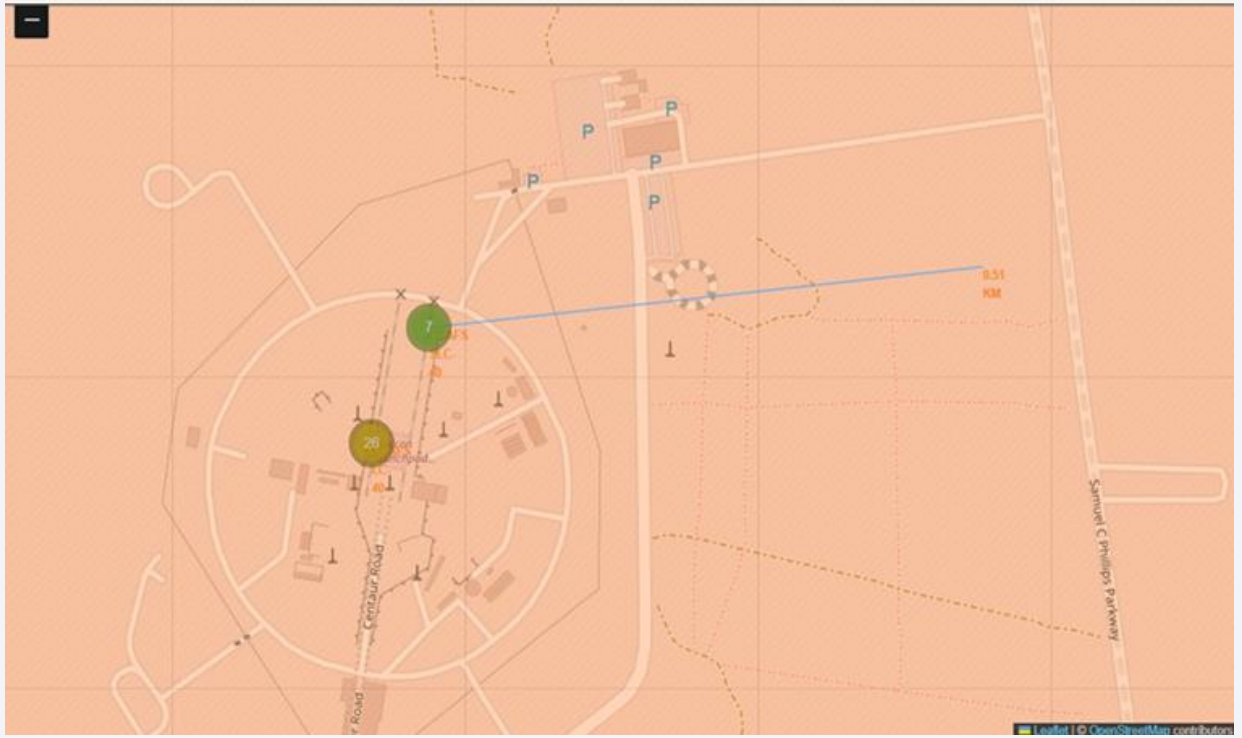
This plot highlight the success and fail launch outcomes on map that is labelled with green and red label for identification.

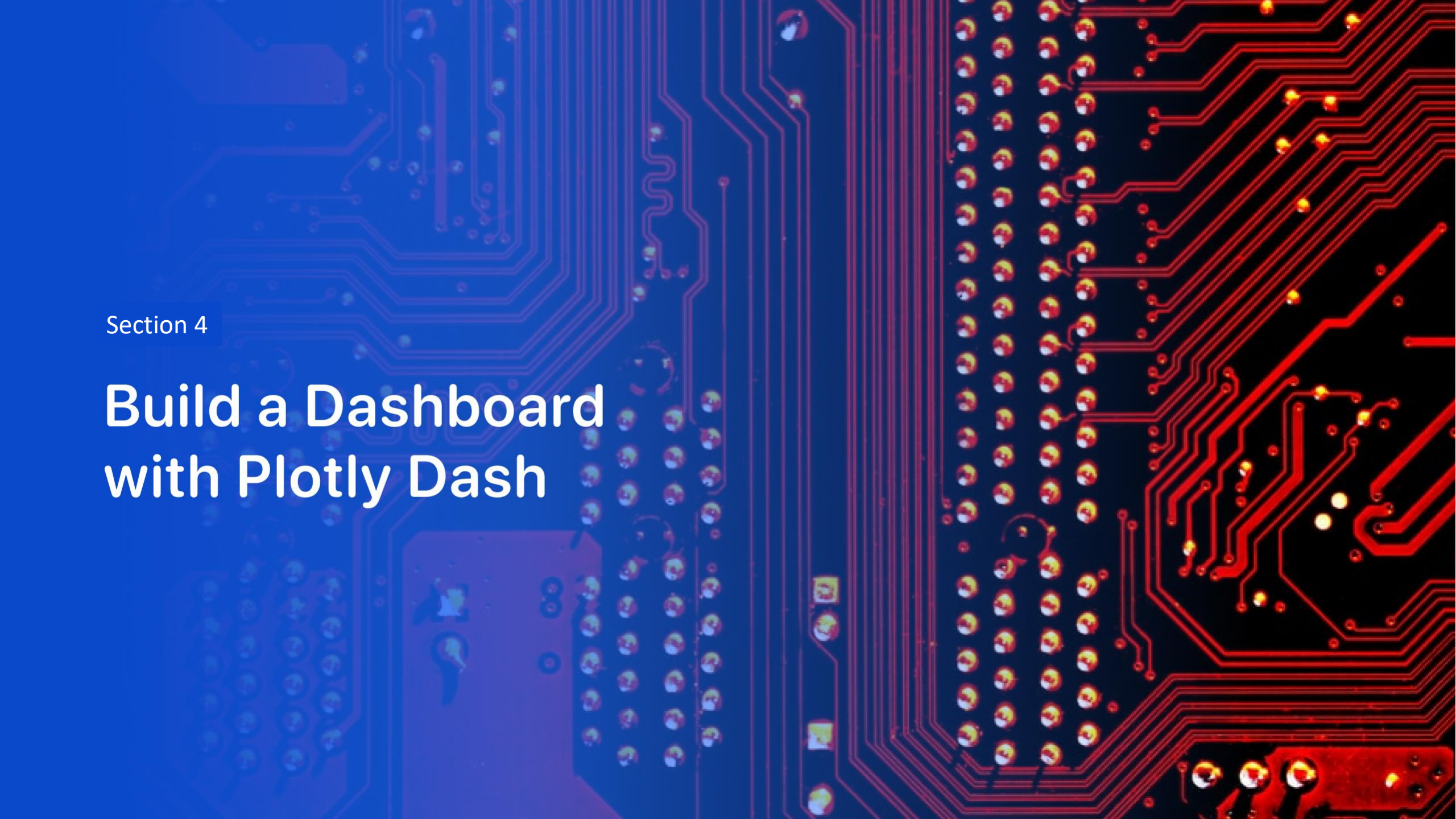


Distance from launch site to important locations

The launch site is close to the coastline but away from city.

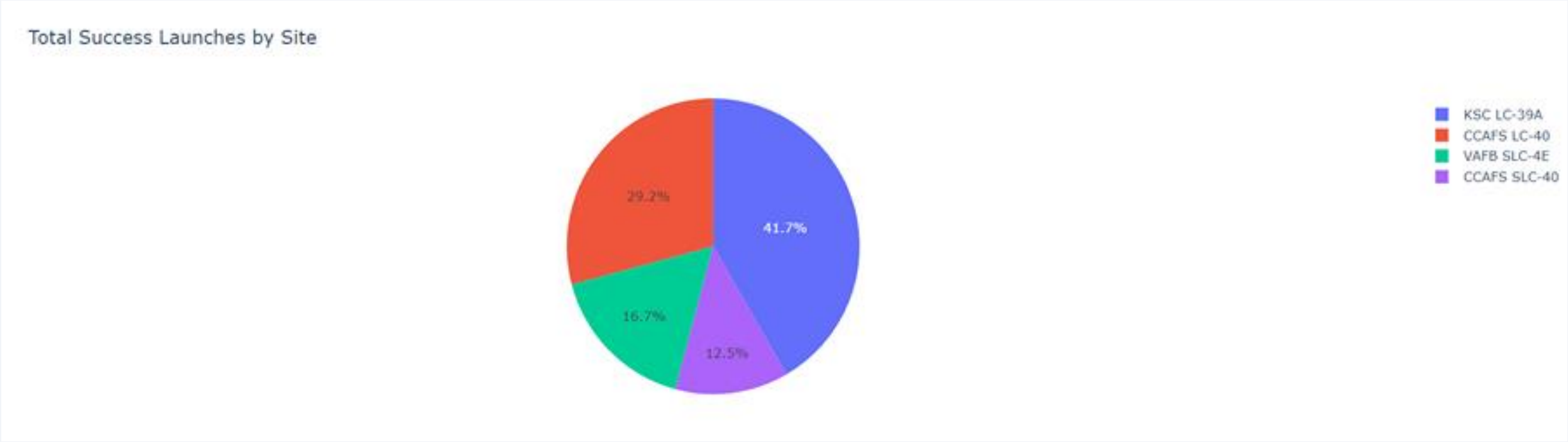
- 0.9 km to coastline
- 29 km to highway
- 79 km to railway station
- 79 km to city




Section 4
Build a Dashboard with Plotly Dash

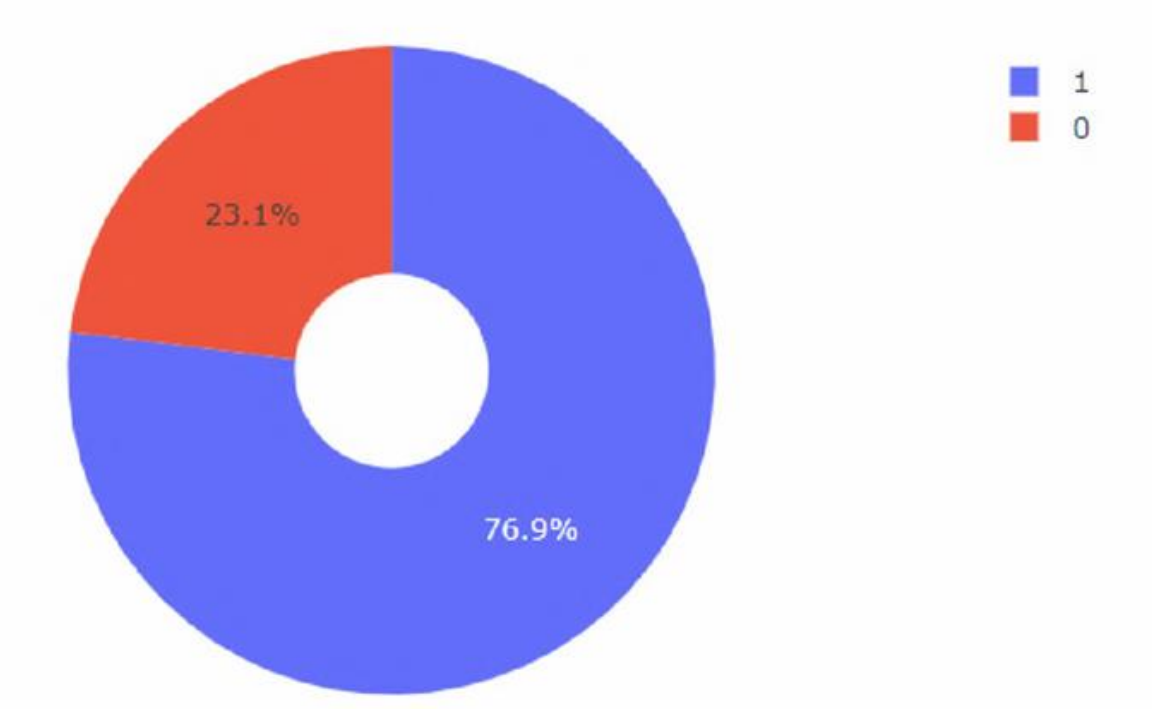
Distribution of successful launches by sites

From the Pie Chart we sees that KSC LC-39A have the highest number of successful launches compare to other sites



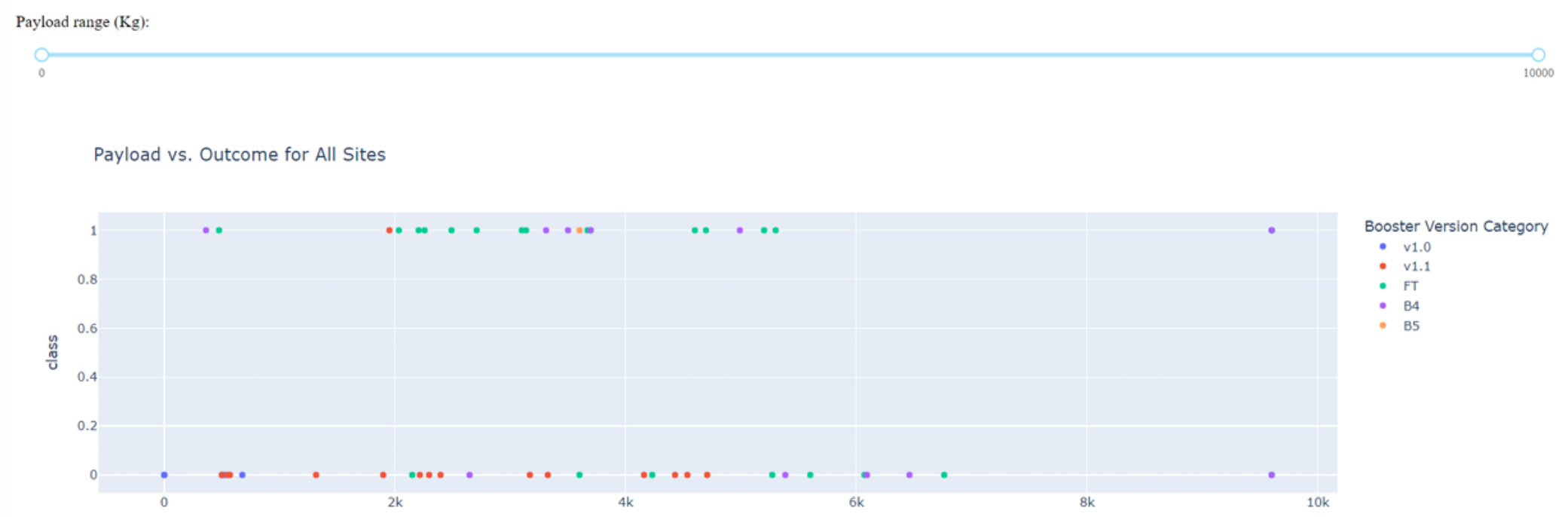
The launch statistic of KSC LC-39A site

The pie chart shows that KSC LC-39A site have a 76.9% success rate and a 23.1% unsuccessful rate.



Distribution of launches across booster version over launch outcomes

The Scatter Plot shows that smaller payload have a relatively higher success rate compare to those with a larger payload



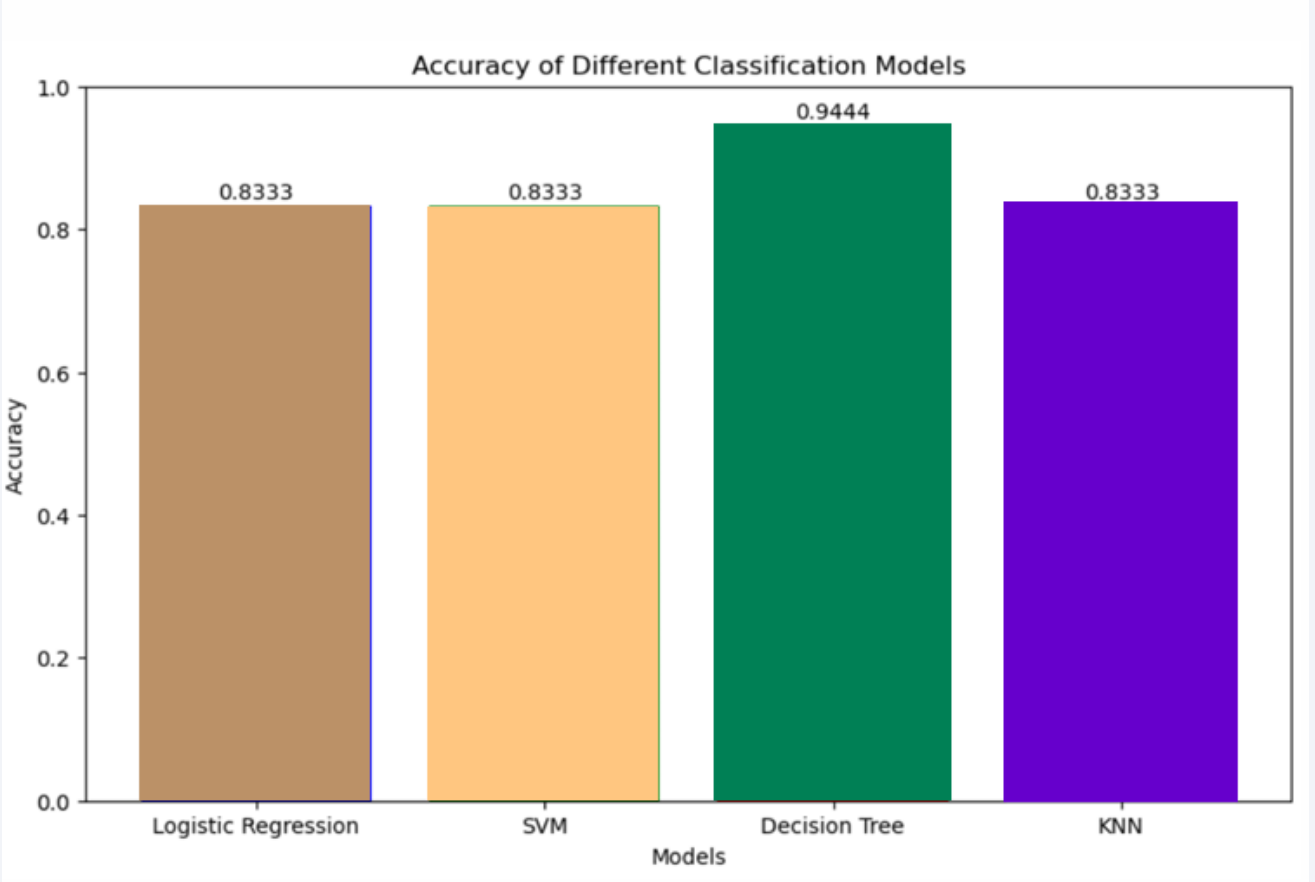
Section 5

Predictive Analysis (Classification)

Classification Accuracy

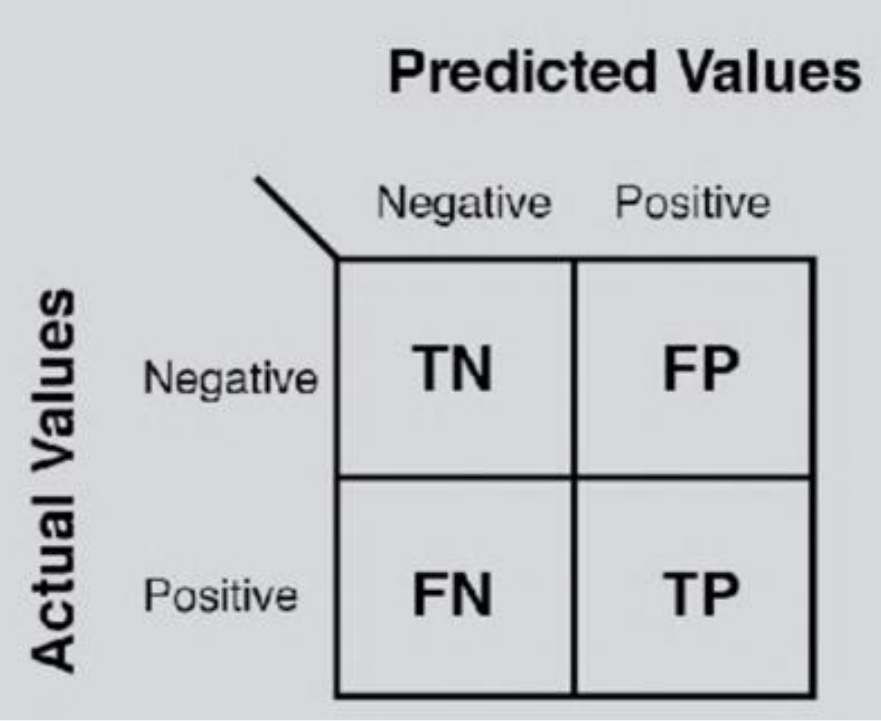
Among the four classifications model, the decision tree classification models has the highest 94% accuracy.

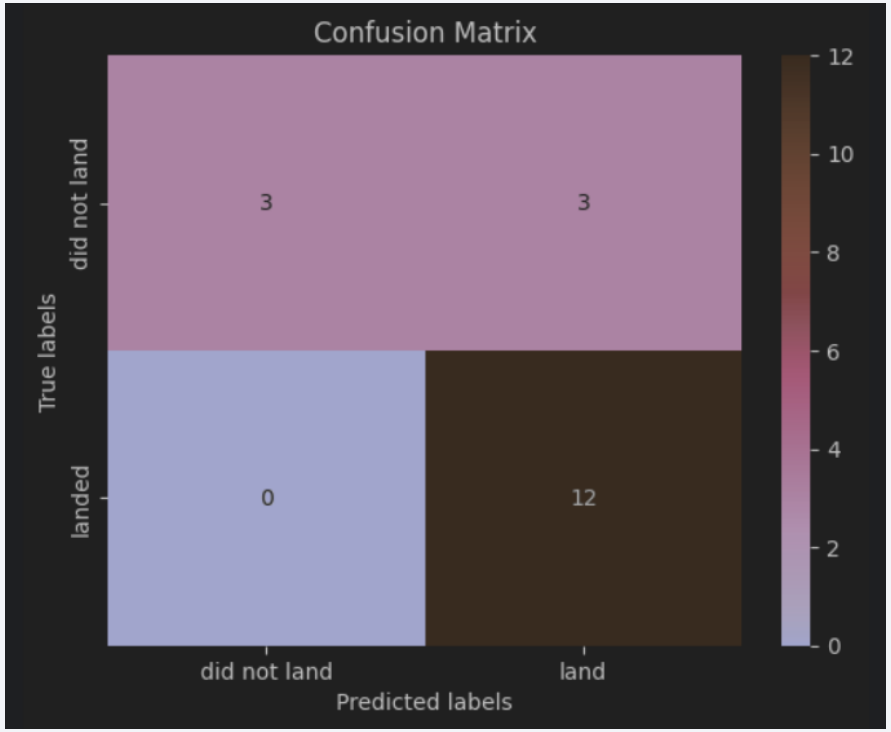
The implies that the decision tree model best fit the dataset compare to other classification models



Confusion Matrix

Confusion matrix from the decision tree classifier shows that there're several False Positive records made by the classifier, and the classifier has a high accuracy in predicting negative value.





Conclusions

- SpaceX rocket launch techniques has been improved over the years
- More large payload rockets is tested and launched, although with a slightly lower successful rate than small payload rocket
- The launch site KSC LC-39A is the most ideal launch site with a highest successful rate
- In terms of prediction models, the decision tree classifier algorithm perform the best in predicting the successful landing outcomes by absorbing the dataset provided by SpaceX API.

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!