Winning Space Race with Data Science

Tan Shan Mei
18 Apr 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building interactive map with Folium
  - Building Dashboard with Plotly Dash
  - Predictive Modelling (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics screenshots
  - Predictive modelling results

Introduction

- Project background and context
  - SpaceX has revolutionized the space industry with significant reductions in launch costs, primarily through the innovation of reusable rocket technology.
  - Our project revolves around analyzing data from SpaceX's past launches to gain insights into their cost-saving strategies and predict future launch successes.
- Problems you want to find answers
  - Identify patterns in launch outcomes based on variables like payload mass, destination orbit, and landing site conditions.
  - Predictive Modeling: Develop a predictive model to assess whether the first stage of a Falcon 9 rocket will land successfully.
  - Cost estimation bidding: Determine the cost of a launch based on success of rocket landings.

Section 1

Methodology

Methodology

Executive Summary

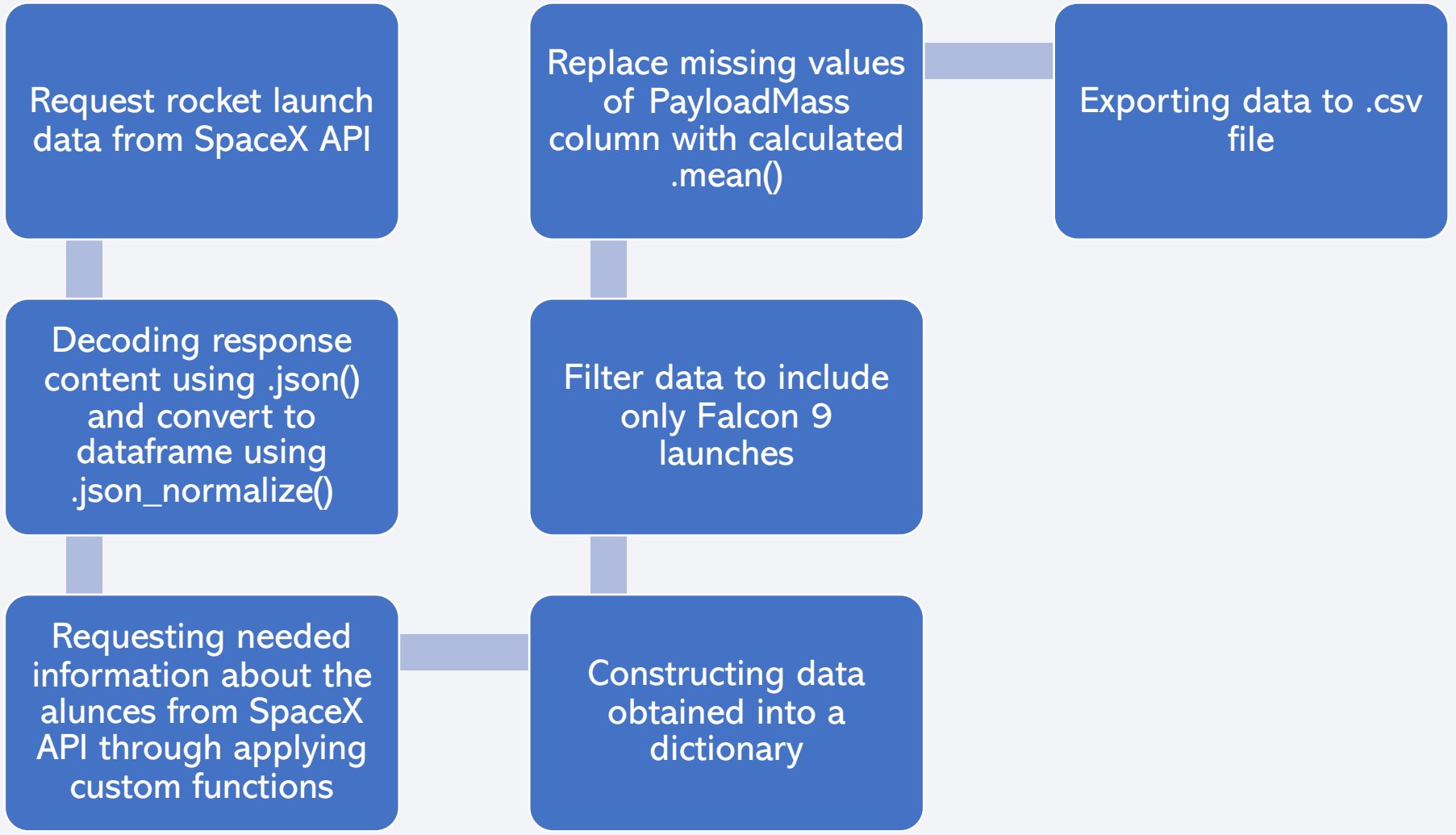
- Data collection methodology:
  - Utilized the SpaceX REST API to fetch historical launch data in JSON format.
  - Web Scraped relevant Wikipedia pages for Falcon 9 launch records using BeautifulSoup.
- Perform data wrangling
  - Cleaned and structured raw JSON and HTML data into usable Pandas DataFrames.
  - Filtered data to focus on Falcon 9 and handled missing or null values in the dataset and One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Applied classification algorithms (e.g., Logistic Regression, Decision Trees) to predict rocket landing successes.
  - Tuned models' hyperparameters to improve predictive performance using evaluation metrics like accuracy, precision, recall, and F1-score.

Data Collection

- Data collection process involve combining data from both SpaceX's API and web scraping SpaceX's Wikipedia tables in order to get a detailed analysis of the rocket launches.

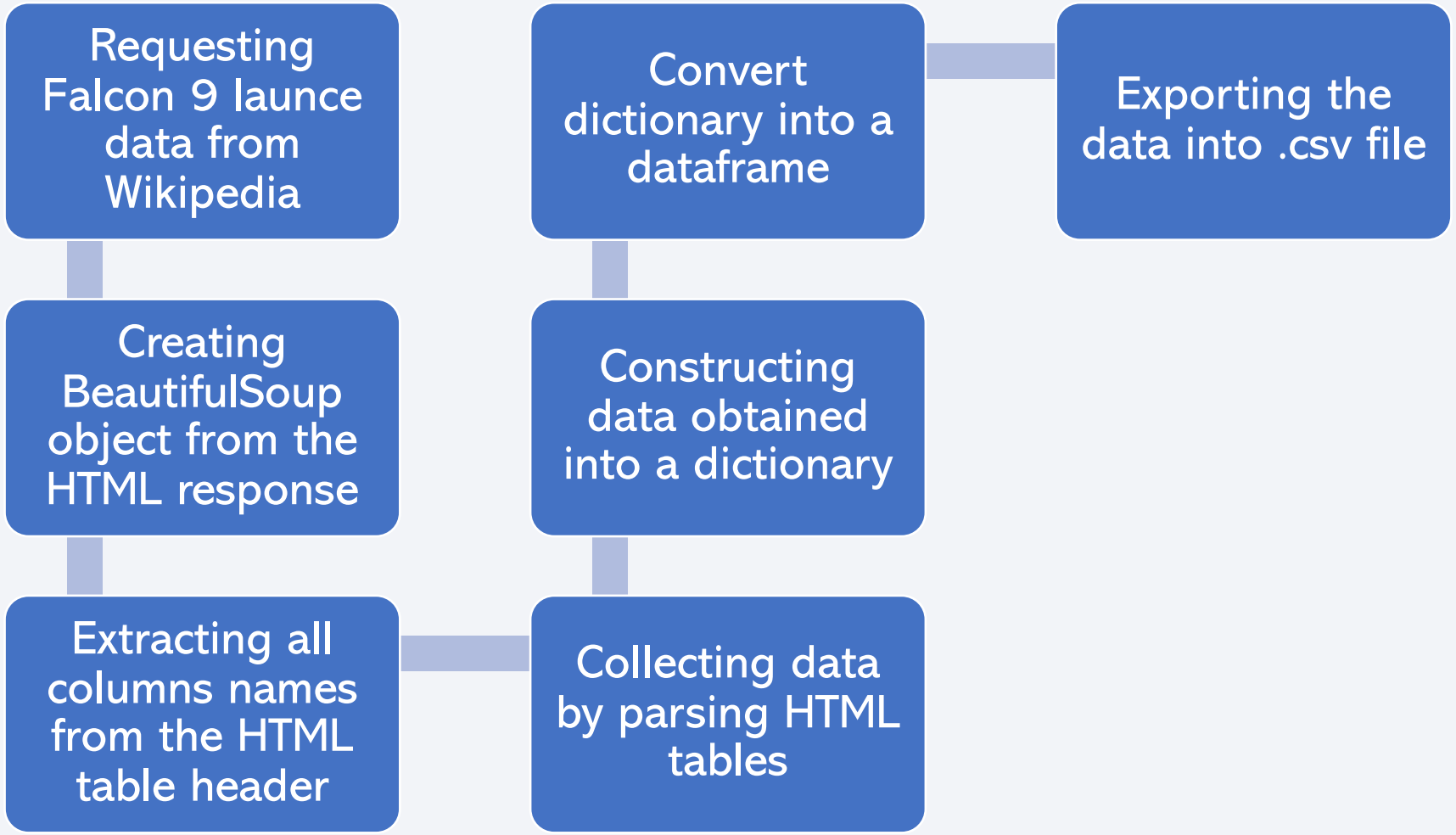
Features from SpaceX Rest API	FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
Features from Wikipedia Web Scraping	Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL: Data Collection - SpaceX API

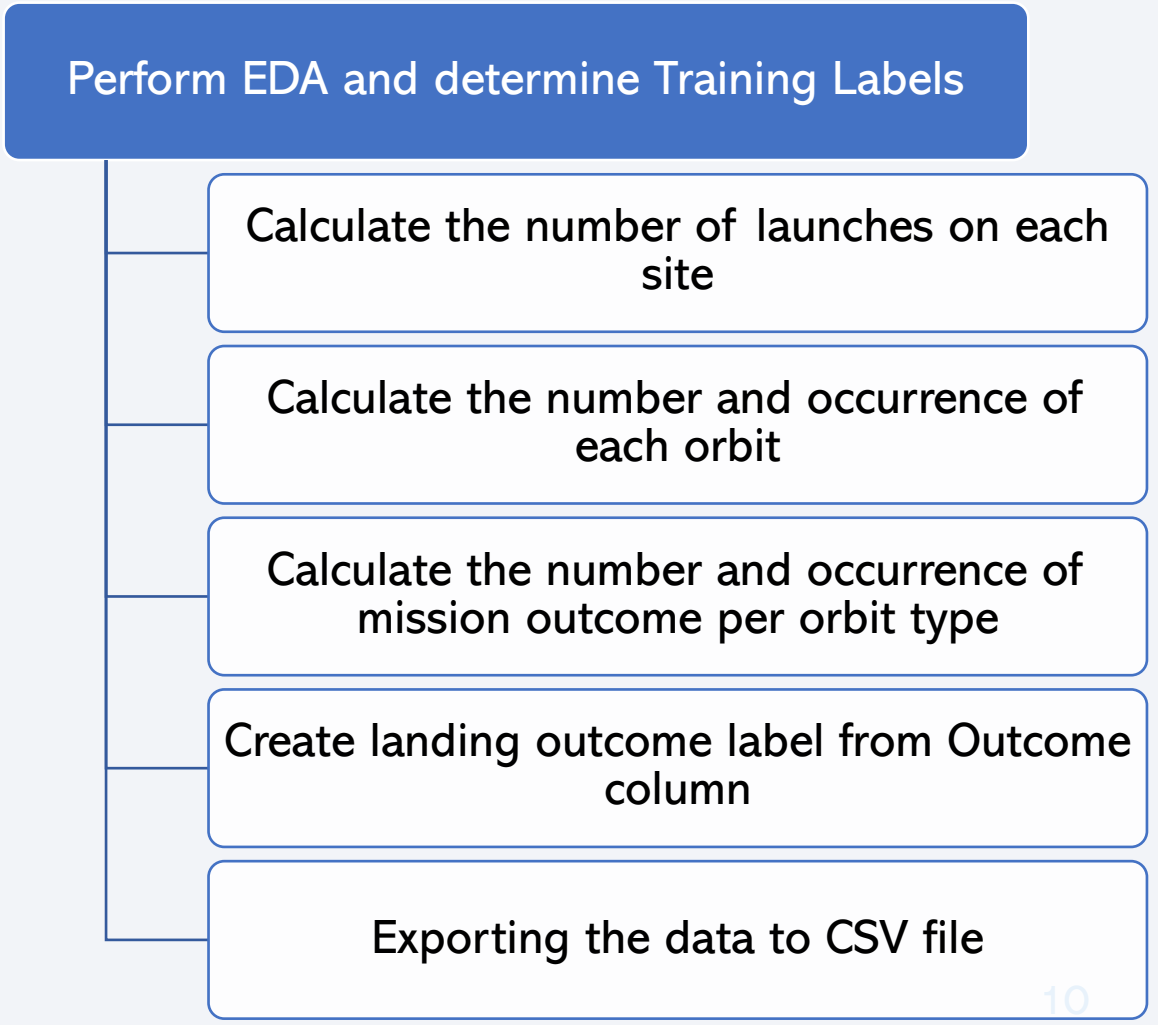
Data Collection - Scraping



GitHub URL: Data Collection - Scraping

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, **True Ocean** means the mission outcome was successfully landed to a specific region of the ocean while **False Ocean** means the mission outcome was unsuccessfully landed to a specific region of the ocean. **True RTLS** means the mission outcome was successfully landed to a ground pad **False RTLS** means the mission outcome was unsuccessfully landed to a ground pad. **True ASDS** means the mission outcome was successfully landed on a drone ship **False ASDS** means the mission outcome was unsuccessfully landed on a drone ship.
- We will mainly convert those outcomes into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful.



GitHub URL: Data Wrangling

EDA with Data Visualization

- Charts were plotted: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
  - Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
  - Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
  - Line charts show trends in data over time (time series).
- Prepare Data Feature Engineering

GitHub URL: EDA with Data Visualization

EDA with SQL

- Exploratory Data Analysis using SQL queries are executed to understand more about the SpaceX Dataset through connection to the database
- SQL queries:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS) • Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking the 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

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- Markers of all Launch Sites:
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Coloured Markers of the launch outcomes for each Launch Site:
  - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
  - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

GitHub URL: Build an Interactive Map with Folium

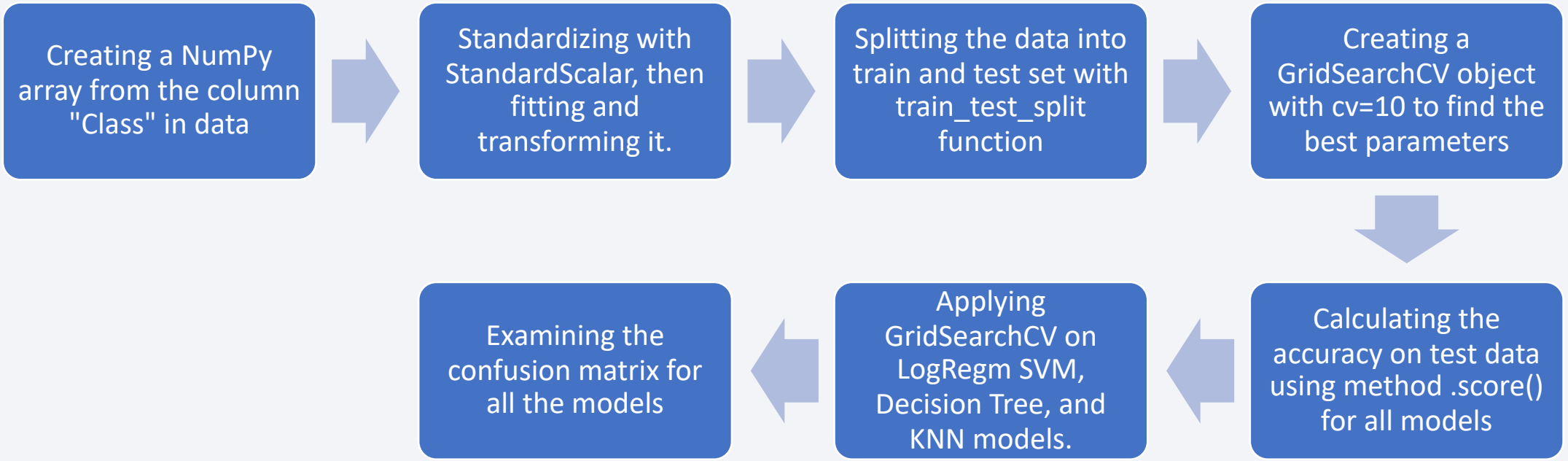
Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection. Pie Chart showing
- Success Launches (All Sites/Certain Site):
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range. Scatter Chart of Payload Mass vs. Success.
- Rate for the different Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub URL: Build a Dashboard with Plotly Dash

Predictive Analysis (Classification)

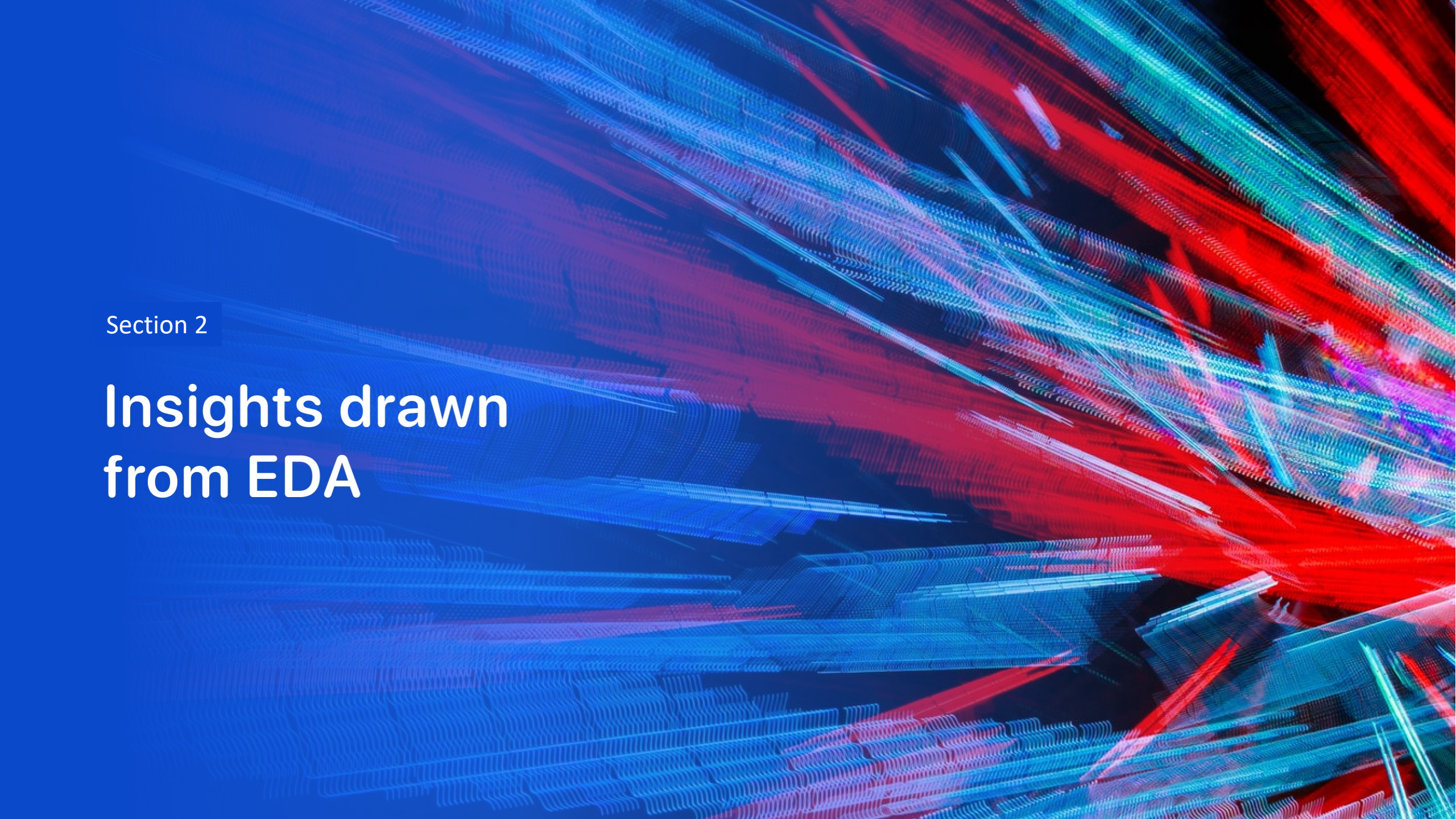
- Creation of a machine learning pipeline to predict if the first stage will land given the data collected and feature engineered.



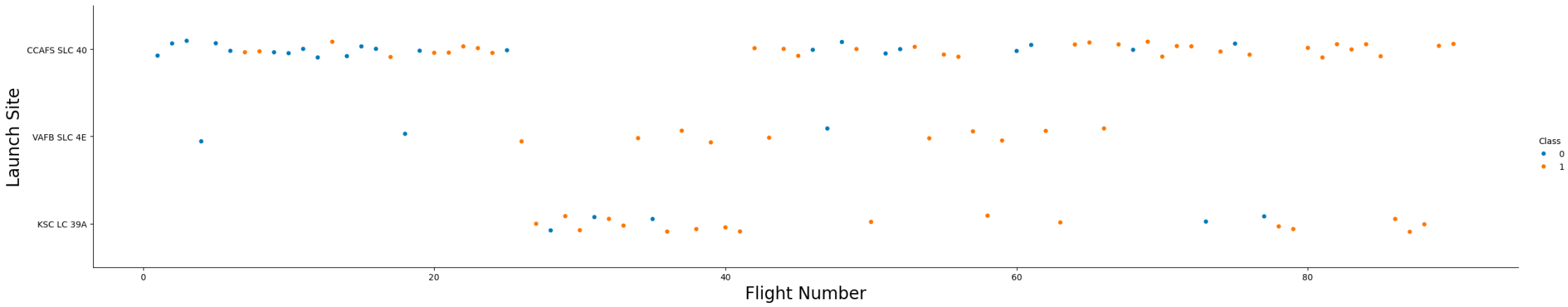
GitHub URL: [Predictive Analysis (Classification)]

Results

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

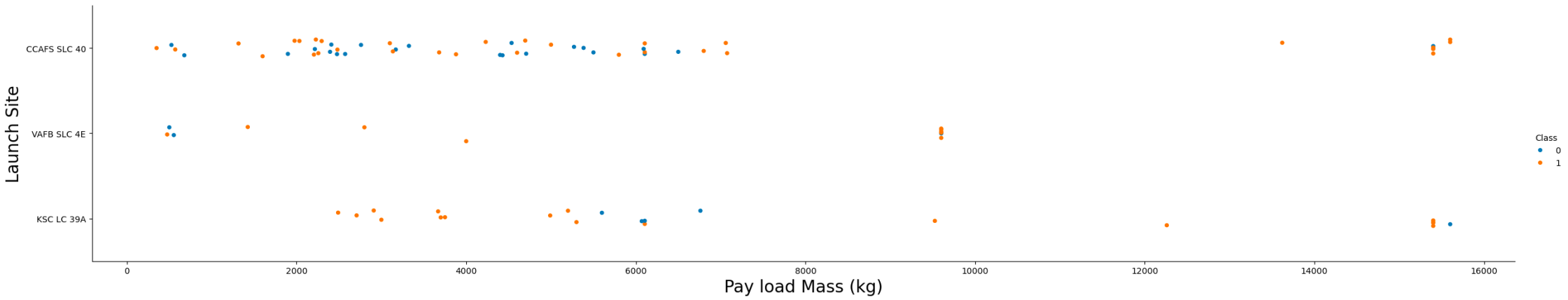

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Explanations:
  - VAFB SLC 4E and KSC LC 39A have higher success rates
  - There were higher proportions of launces from CCAFS SLC 40

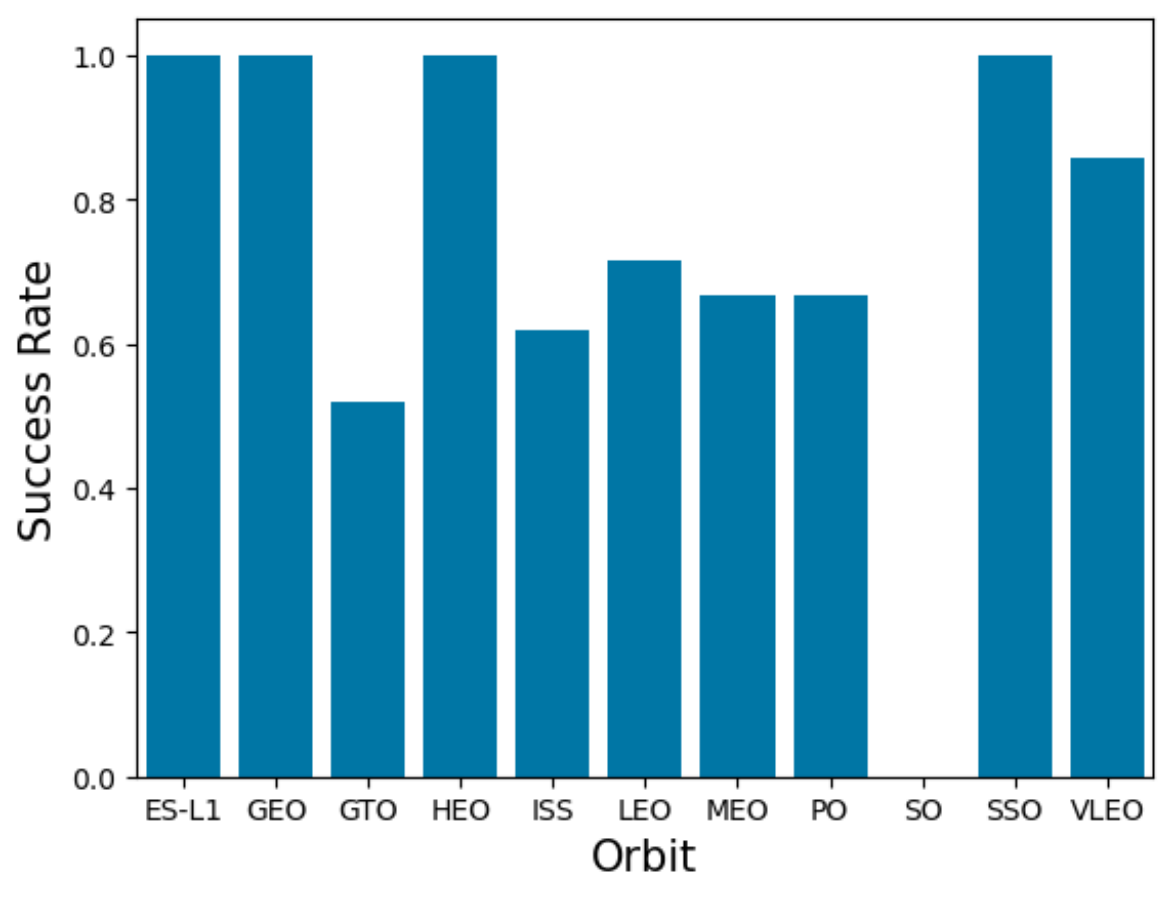
Payload vs. Launch Site



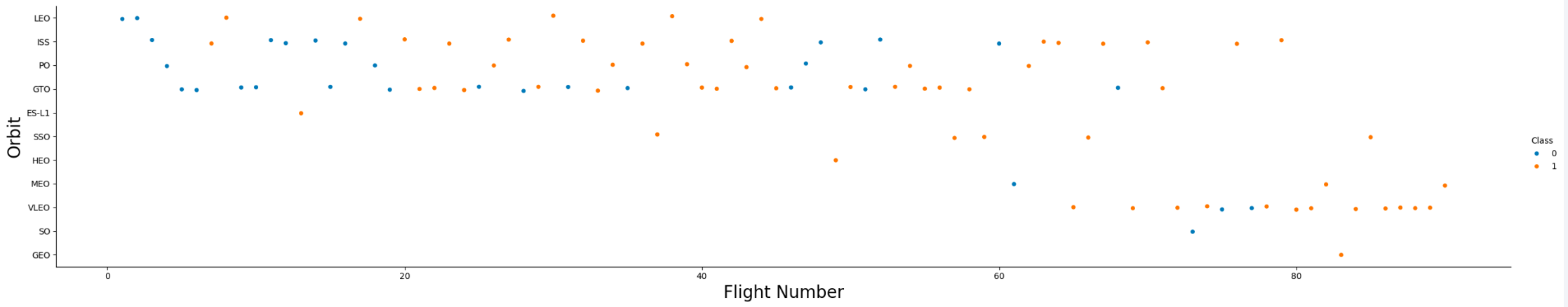
- Explanations:
  - Majority of the launches are of PayloadMass < 8000kg.
  - For launches with PayloadMass > 9000, there is a higher success rate.

Success Rate vs. Orbit Type

- Explanations:
  - Orbits with 100% Success Rate are ES-L1, GEO, HEO, and SSO
  - Oribits with 0% Success Rate is SO.
  - Orbits with Success Rate between 50% - 85% are GTO, ISS, LEO, MEO, and PO.

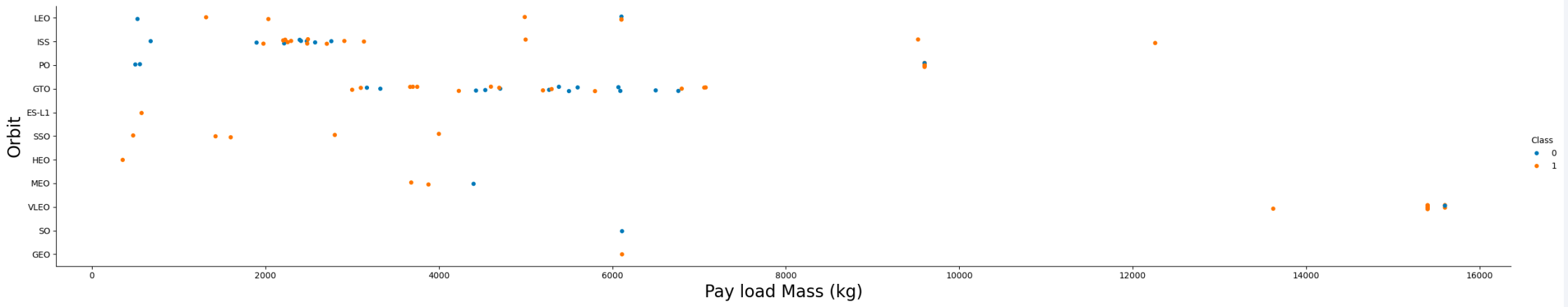


Flight Number vs. Orbit Type



- Explanations:
  - No visible relationship can be observed between Flight Number and Orbit Type

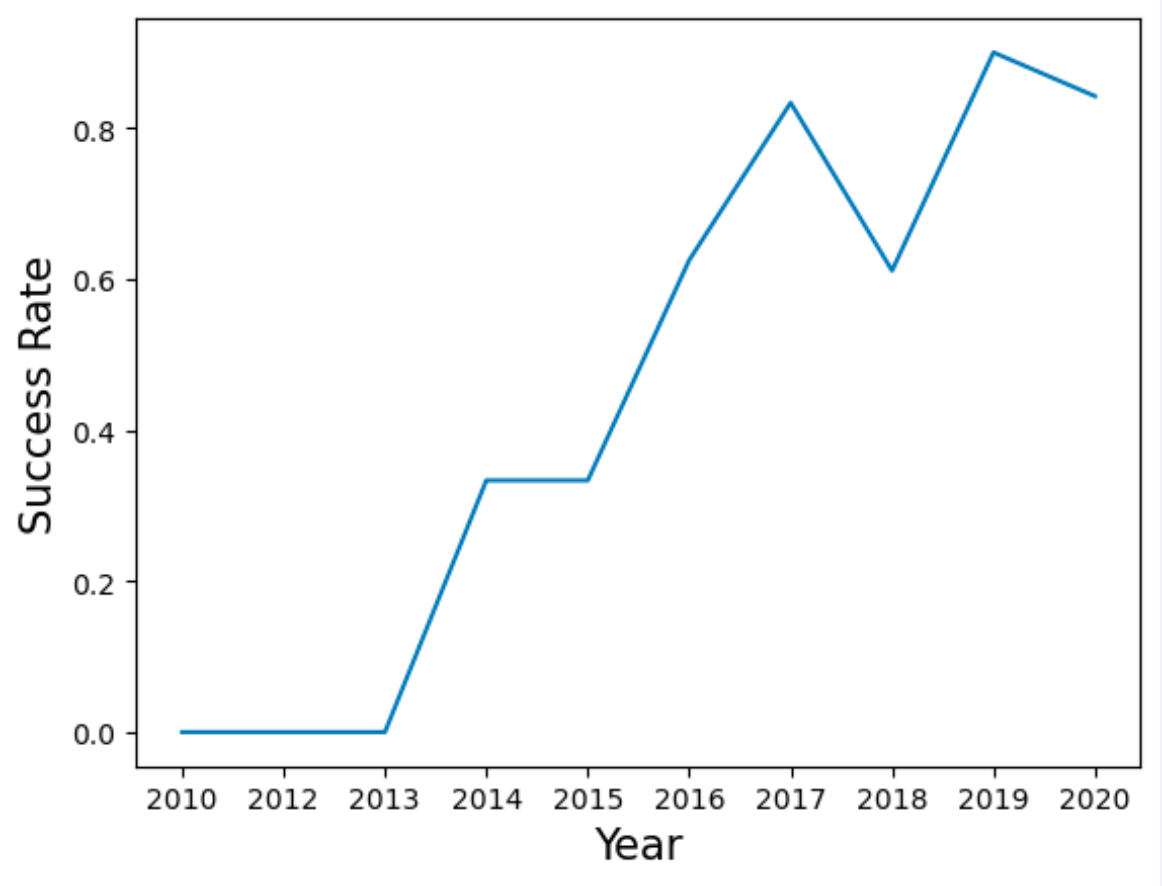
Payload vs. Orbit Type



- Explanations:
  - Heavy payloads have negative influence on GTO orbits and positive influence on GTO and ISS orbits.

Launch Success Yearly Trend

- Explanations:
  - The success rate is increasing on an increasing rate from 2013 to 2020.
  - The success rate from 2010 – 2013 is 0%



All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

Python

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

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

- Explanations:
  - Displaying the unique launch site names from SpaceX mission dataset.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

Python

```
 * 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

- Explanations:
  - Displaying 5 records where Launch_Site begins with the string “CCA”.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

Python

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

SUM(PAYLOAD_MASS__KG_)
45596

- Explanations:
  - Displaying the total payload mass carried by boosters launched by NASA (CRS) which equals to 45596.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'
```

Python

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Explanations:
  - Displaying average payload mass carried by booster version F9 v1.1 which is equal to 2928.4.

First Successful Ground Landing Date

Hint:Use min function

+ Code + Markdown

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
```

Python

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

MIN(Date)
2015-12-22

- Explanations:
  - The date when the 1st successful landing outcome (ground pad) was achieved which is 22 December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

Python

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

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

- Explanations:
  - Displaying the booster version names which have success in drone ship and have a 4000 < payload mass < 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTBL GROUP BY Mission_Outcome
```

Python

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

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

- Explanations:
  - Listing the total number of successful and failure SpaceX mission outcomes from the dataset.

Boosters Carried Maximum Payload

- Explanations:
  - Displaying the names of the booster version names which have carried out the maximum payload mass.

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * 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

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.

```
%sql SELECT substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,0,5)='2015' AND Landing_Outcome = 'Failure (drone ship)'
```

Python

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

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Explanations:
  - Listing the failed landing outcomes in drone ship, their booster version, and launch site names for the months in year 2015.

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

```
%sql SELECT Landing_Outcome, COUNT(*) FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC
```

Python

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

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

- Explanations:
  - Ranking the count of landing outcomes from the period between 2010-06-04 and 2017-03-20 in descending order.

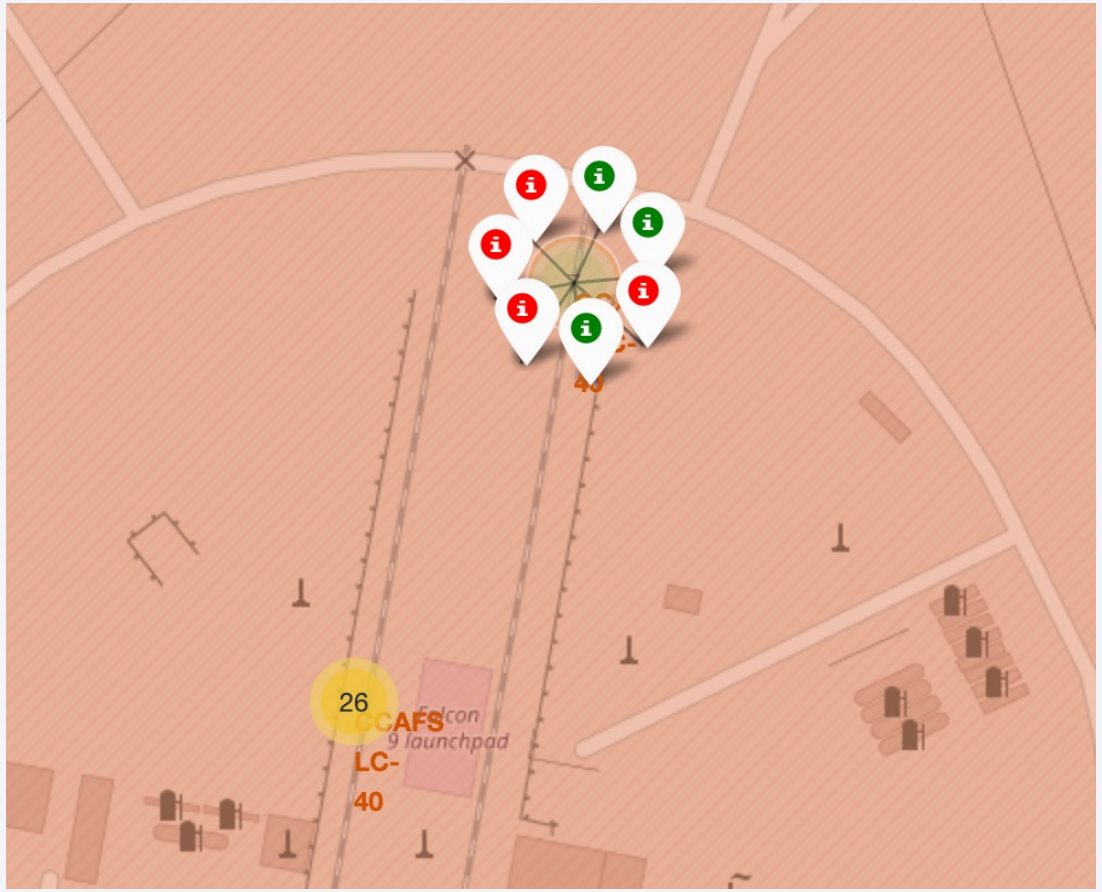

Section 3
Launch Sites
Proximities Analysis

<Folium Map Screenshot 1>

- All launch sites' location markers on a global map
- All launch sites are in close proximity to the coast.

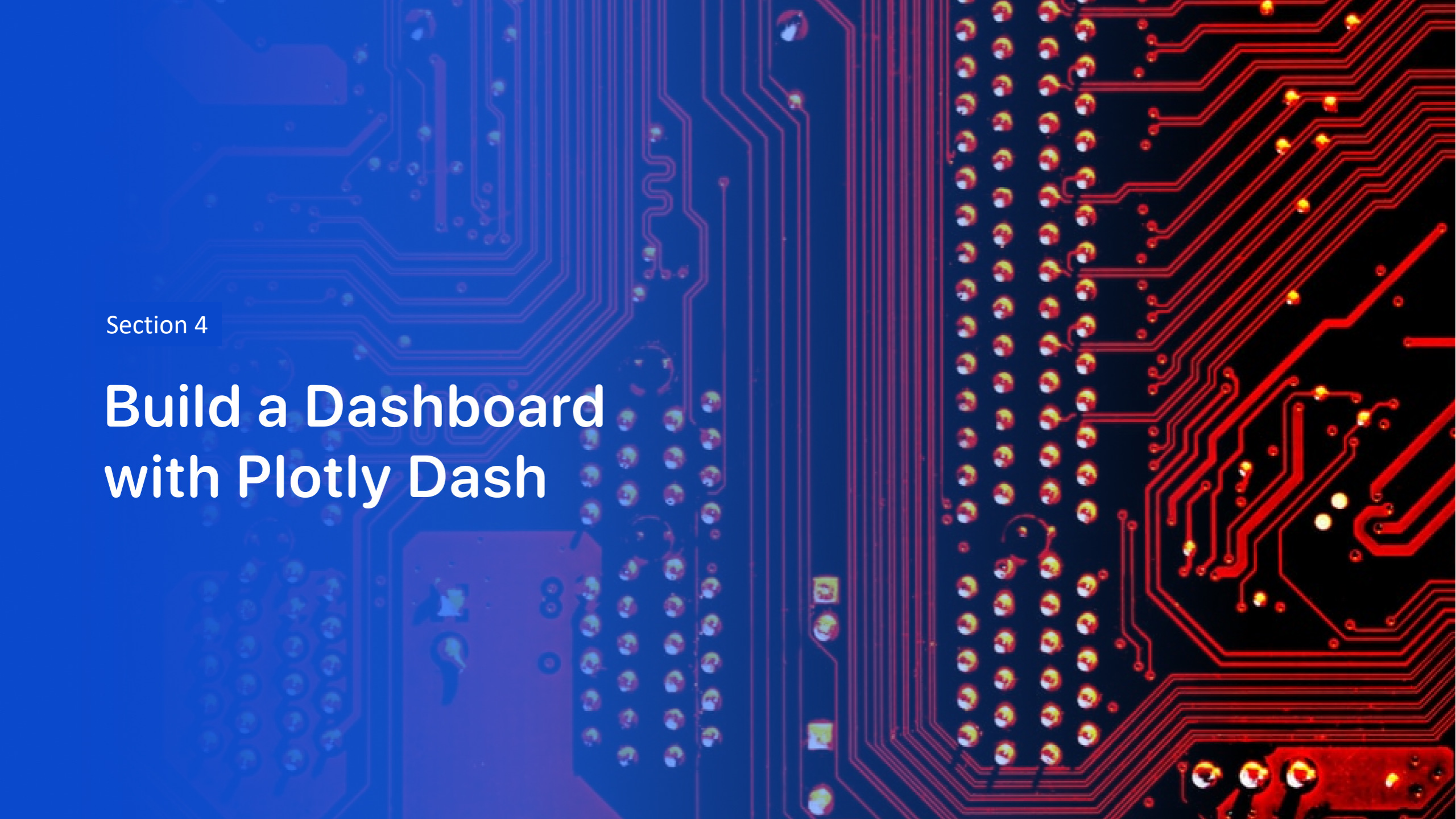
<Folium Map Screenshot 2>

- Mark the success/failed launches for each site on the map
  - Success -> green marker
  - Failed -> red marker
- Launch sites with relatively high success rates can be easily identified



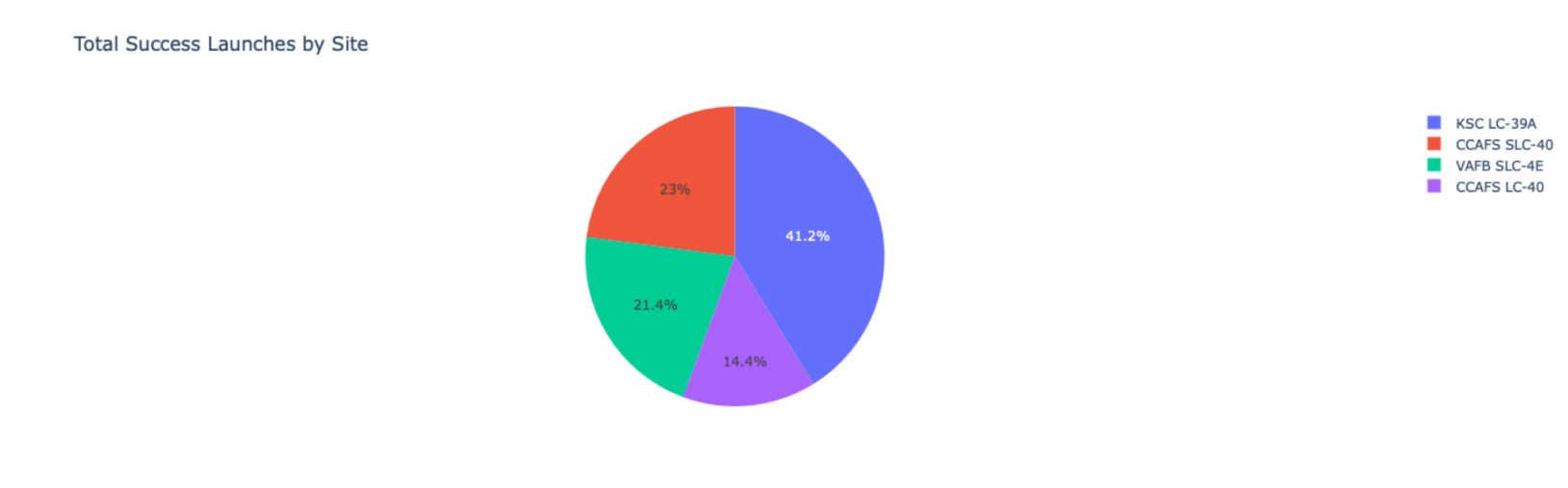
<Folium Map Screenshot 3>

- Calculate distances between launch site to its proximities
- For example: launch site KSC LC-39A is 15.23km away from the railway.
- Have a gauge on the impact of failed rocket launches to its proximities.


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

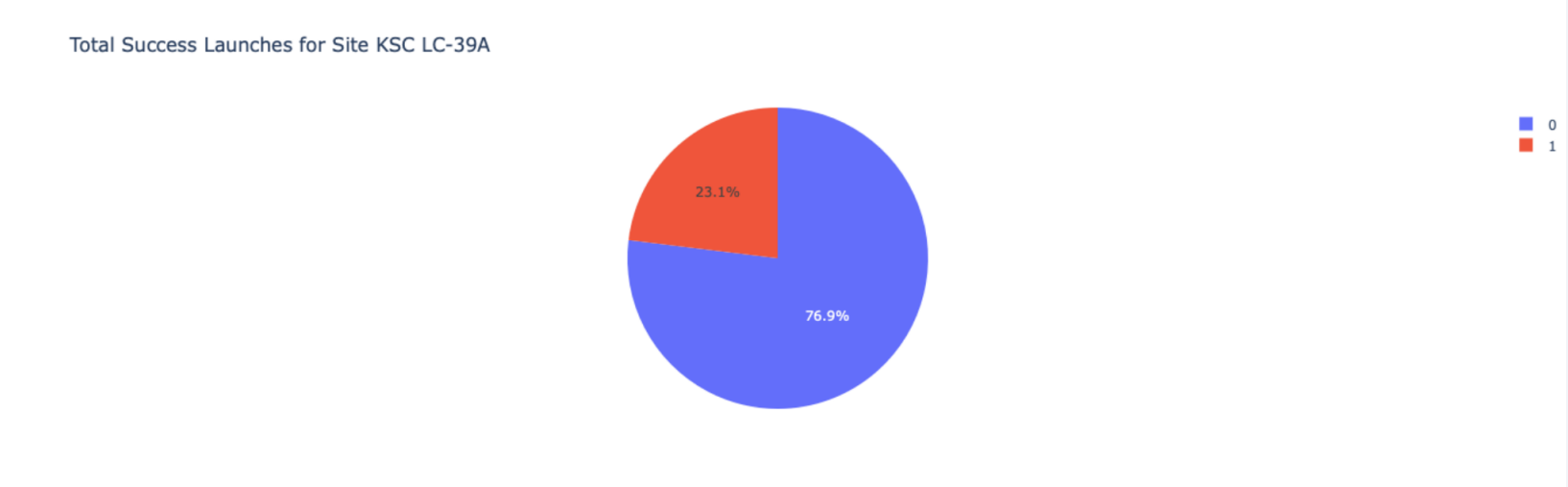
- Launch success count for all sites



- Explanations:
  - The chart shows that KSC LC-29A has a higher proportion of successful launches compared to the rest.

<Dashboard Screenshot 2>

- Launch site with the highest launch success ratio



- Explanations:
  - The pie chart shoes that the launch site KSC LC-29A has a high success rate of (76.9%) with 10 successful launches and only 3 failed launches.

<Dashboard Screenshot 3>

- Payload Mass vs. Launch Outcomes for all sites

- Explanations:
  - The chart shows that payloads between 2000 and 5500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Model is evaluated using Jaccard Score, F1 Score and Accuracy of the Test set.
- All the models have the same Test Set scores.
- Due to limited sample size, the models are tested on whole Dataset, in order to find the best performing model.
- Upon evaluation, **Decision Tree model** has the highest score and accuracy making it the best model.

Evaluation on Test Set

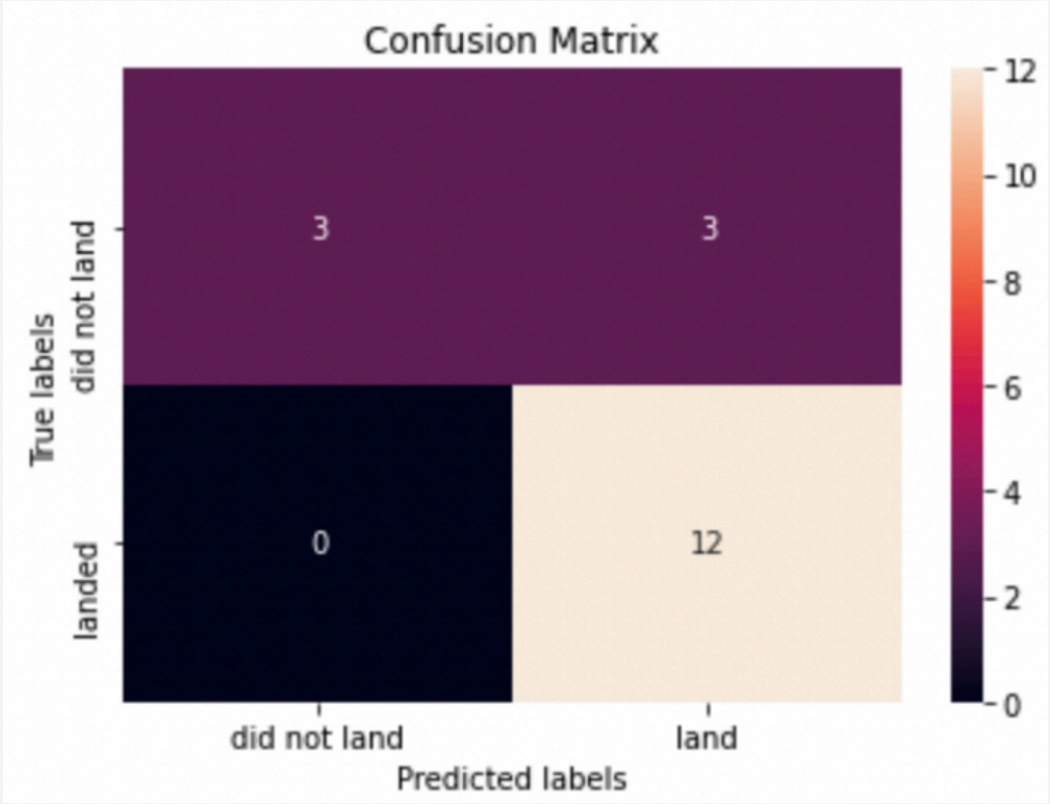
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Evaluation on Whole Dataset

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

- Evaluating the confusion matrix, the decision tree model can distinguish between the different classes.
- The model correctly predicted 12 instances where the rocket's first stage did land.
- Accuracy: (TP + TN) / (TP + TN + FP + FN)
- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)



Conclusions

- Out of all the models, Decision Tree model is the best suited algorithm for the spaceX dataset.
- Most of the sites are in close proximity to the coast.
- The success rates of the launches increases over the years.
- KSC LC-39A has the highest success rate of launches compared to other launch sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Thank you!