Winning Space Race with Data Science

A.G.

06/26/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building interactive map visualization with Folium
  - Building a dashboard with Plotly Dash
  - Predictive classification analysis
- Summary of all results
  - Exploratory Data Analysis results
  - Predictive classification analysis results

Introduction

- Project background and context
  - This project 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. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used by other companies interested in bidding against SpaceX for a rocket launch.
- Problems you want to find answers
  - What influences if the rocket will land successfully?
  - The impact of the many variables involving a rocket launch will have on success rate
  - determining the success rate of a successful landing.
  - What conditions does SpaceX have to achieve to ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API, web scraping from Wikipedia
- Perform data wrangling
  - Irrelevant Column dropped
  - One hot encoding to prepare data for machine learning
- 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

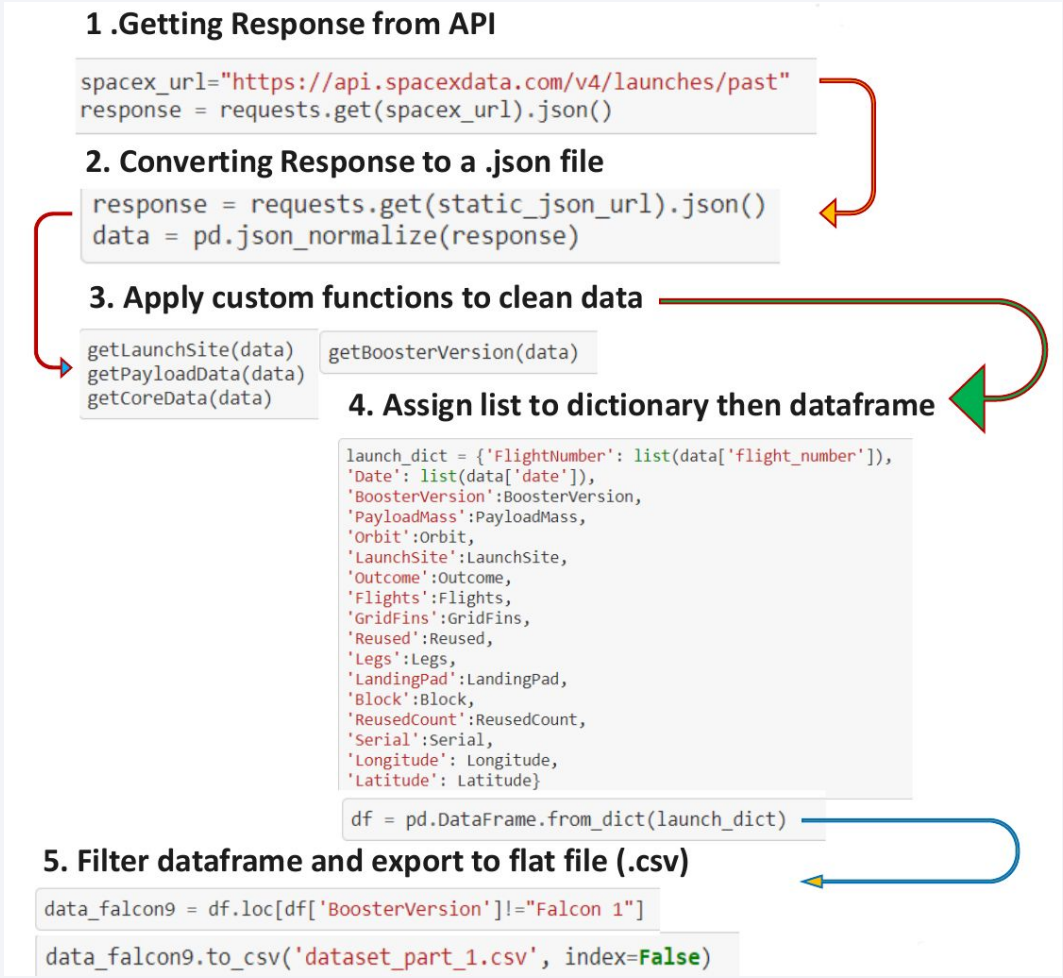
Data Collection

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

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

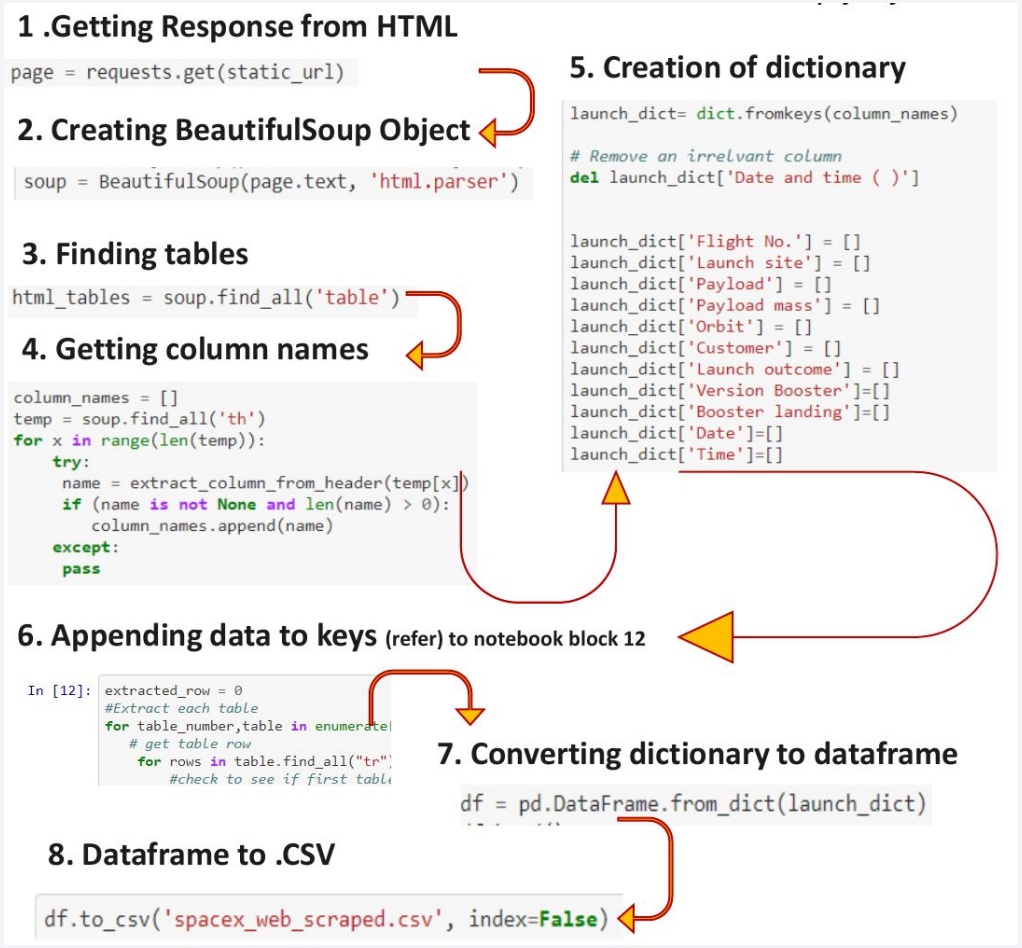
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose



Data Collection - Scraping

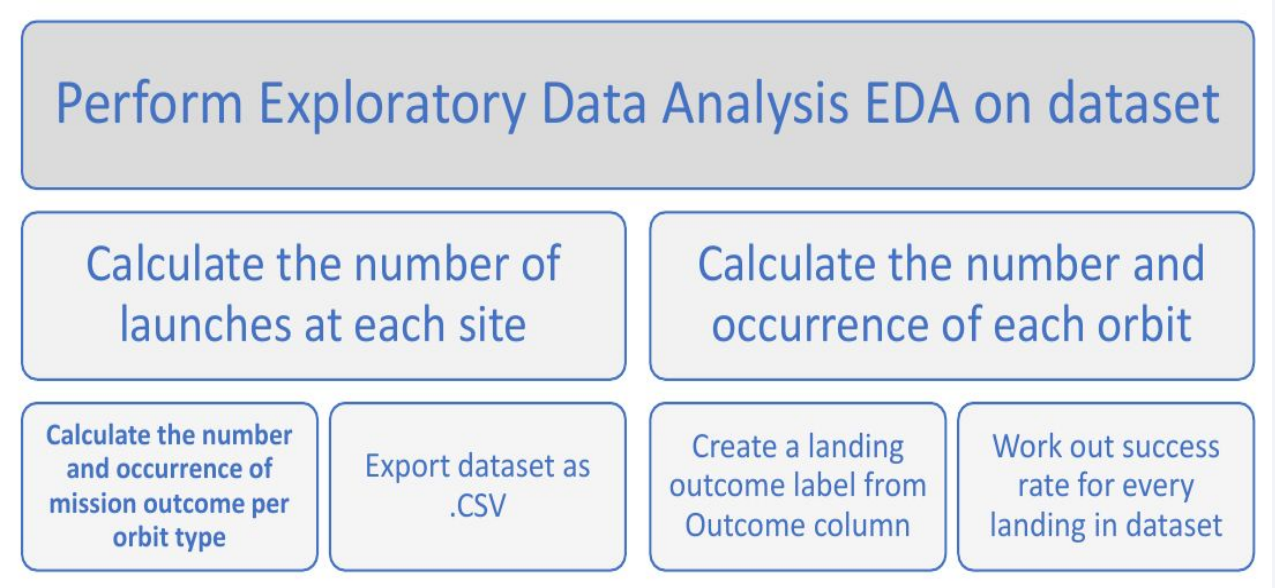
- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



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, in the dataset the following labels were used:
    - **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 use one-hot encoding to convert those outcomes into boolean labels **0** and **1** representing success or failure respectively.
- **Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose**



EDA with Data Visualization

- Scatter Graphs being drawn:
  - Flight Number VS. Payload Mass
  - Flight Number VS. Launch Site
  - Payload VS. Launch Site
  - Orbit VS. Flight Number
  - Payload VS. Orbit Type
  - Orbit VS. Payload Mass
    - Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation . Scatter plots usually consist of a large body of data.
- Bar Graph being drawn:
  - Mean VS. Orbit
    - A bar diagram makes it easy to compare sets of data between different groups at a glance.The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.
- Line Graph being drawn:
  - Success Rate VS. Year
    - Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded

- **Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose**

EDA with SQL

- Performed SQL queries to gather information about the dataset.
    - Displaying the names of the unique launch sites in the space mission
    - Displaying 5 records where launch sites begin with the string 'KSC’
    - 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 where the successful landing outcome in drone ship was achieved.
    - Listing the names of the boosters which have success in ground pad 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 records which will display the month names, successful landing_outcomes in ground pad ,booster
    - versions, launch_site for the months in year 2017
    - Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending
    - order.

- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes
    - 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks.
- These interactive maps allowed us to answer the following questions:
    - Are launch sites in close proximity to railways? **No**
    - Are launch sites in close proximity to highways? **No**
    - Are launch sites in close proximity to coastline? **Yes**
    - Do launch sites keep certain distance away from cities? **Yes**
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- The dashboard is built with Flask and Dash web framework.
- Graphs used:
  - **Pie Chart:** Shows the total launches by a certain site/all sites
    - display relative proportions of multiple classes of data. - size of the circle can be made proportional to the total quantity it represents.
  - **Scatter Graph:** Shows the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions.
    - It allows us to see if there are positive or negative correlations between variables.

Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

- **Building the Model**
  - Import dataset into NumPy and Pandas
  - Transform Data: Split data into training and test sets
  - Decide which type of machine learning algorithms we want to use
  - Set our parameters and algorithms to GridSearchCV
  - Fit our datasets into the GridSearchCV objects and train our dataset.
- **Model Evaluation**
  - Check accuracy for each model
  - Get tuned hyperparameters for each type of algorithms
  - Plot Confusion Matrix
- **Model Tuning**
  - Feature Engineering Algorithm Tuning

- The model with the best accuracy score wins the best performing model

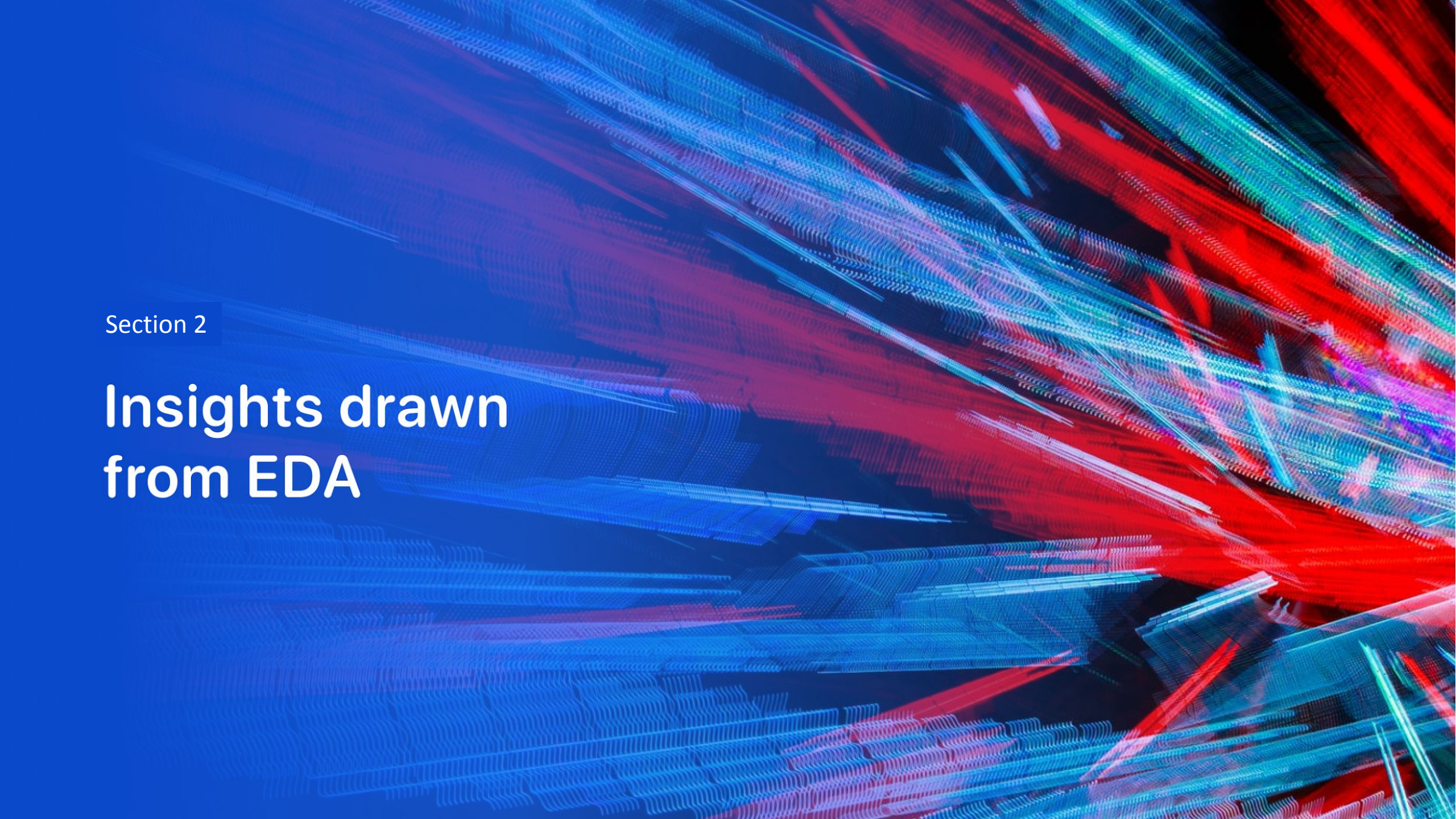
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results

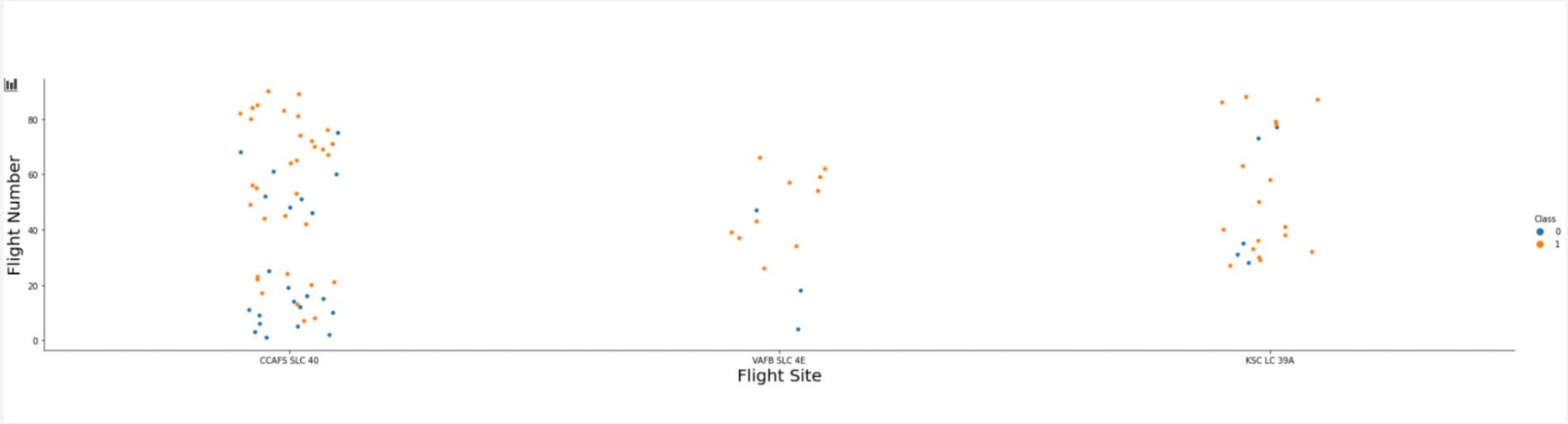
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

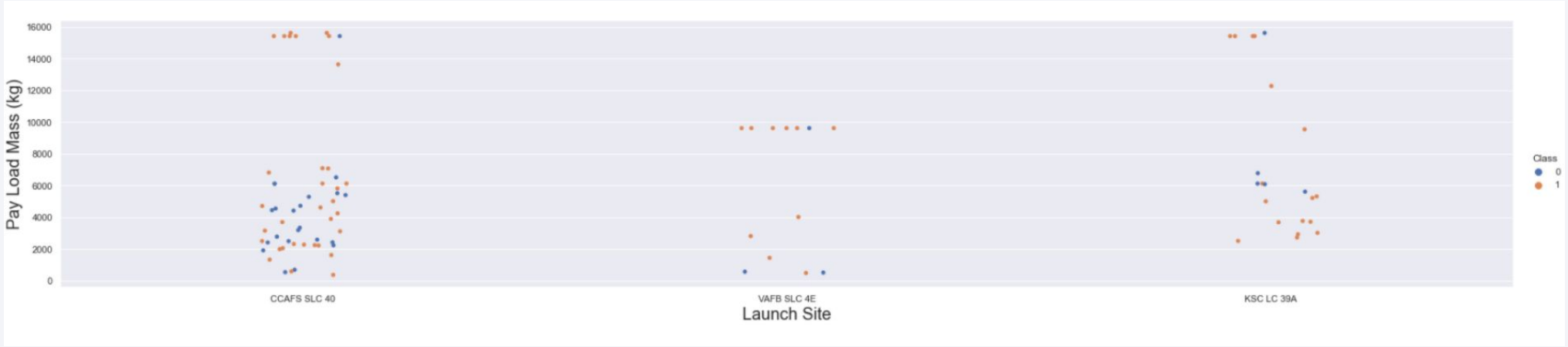


Flight Number vs. Launch Site



The more amount of flights at a launch site the greater the success rate at a launch site.

Payload vs. Launch Site

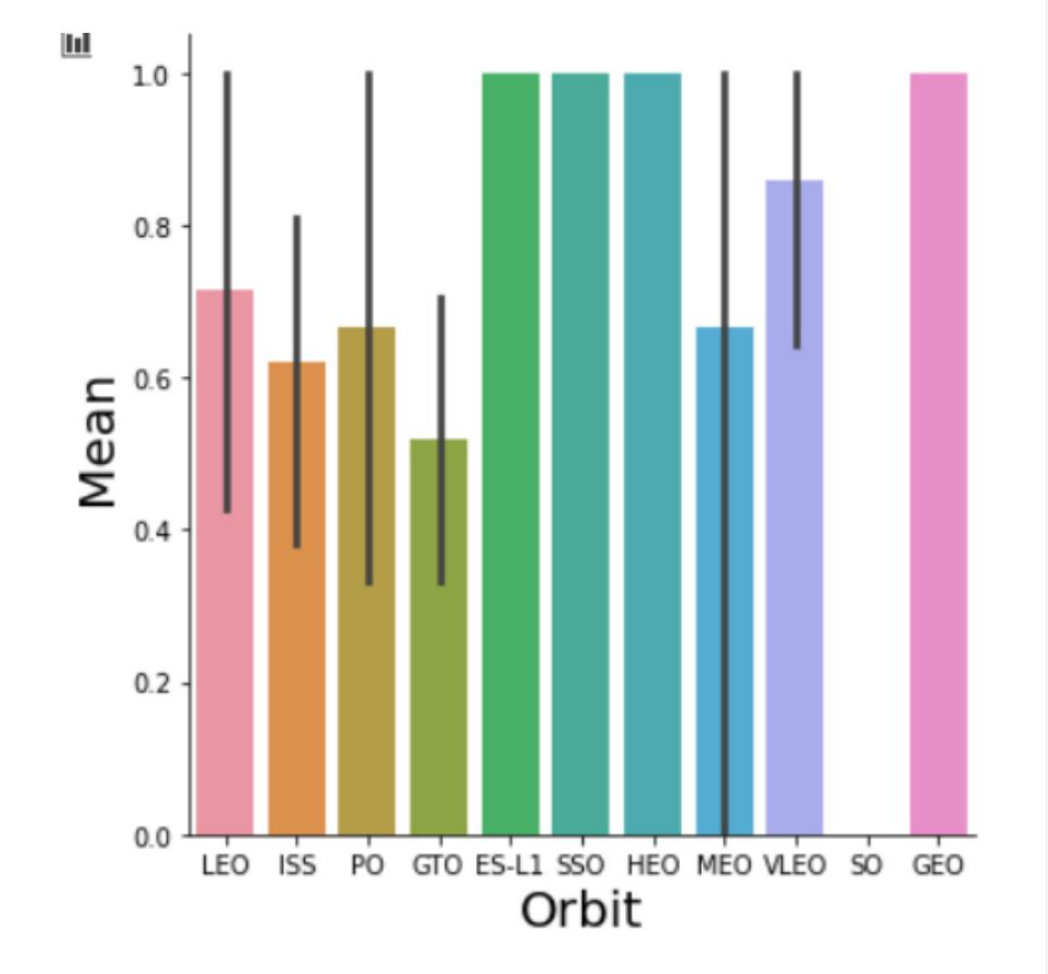


The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket.

There is no clear pattern to be found using this visualization to make a decision if correlation exists between launch site and payload mass for a successful launch.

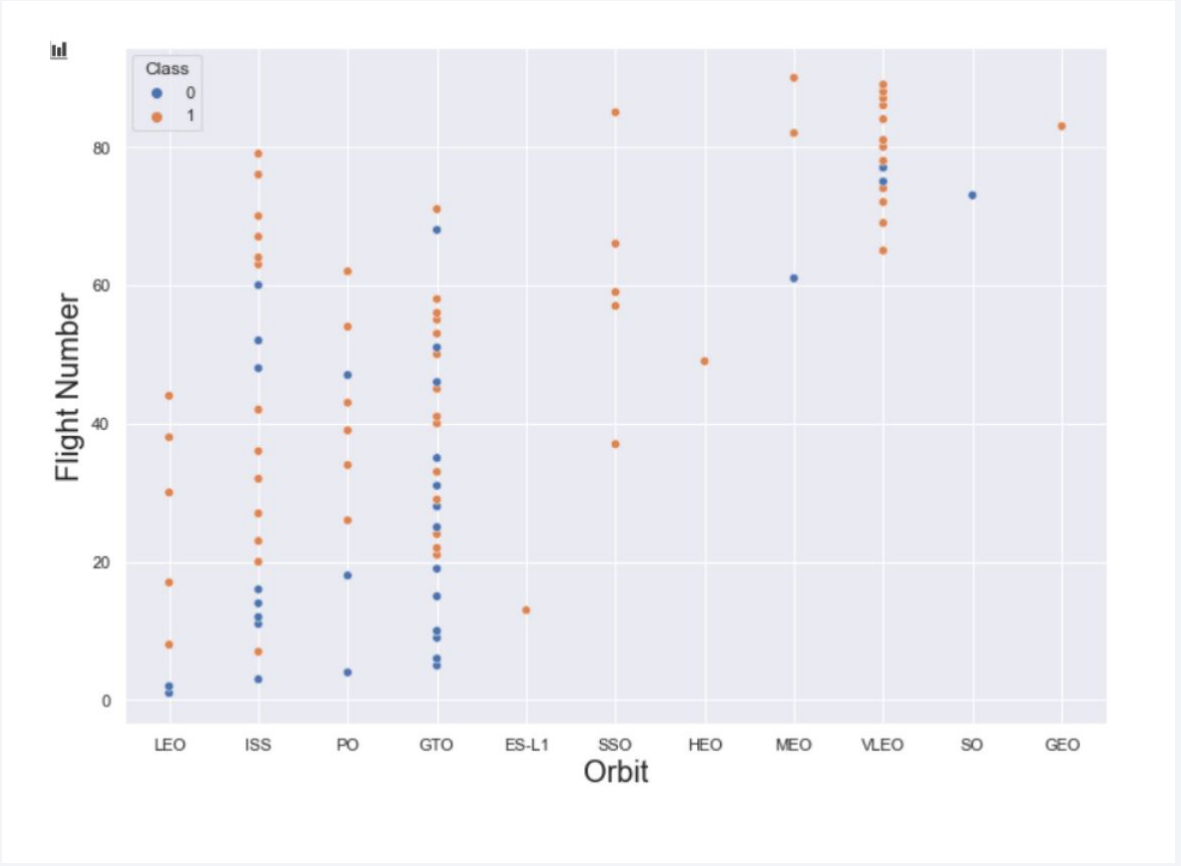
Success Rate vs. Orbit Type

Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate



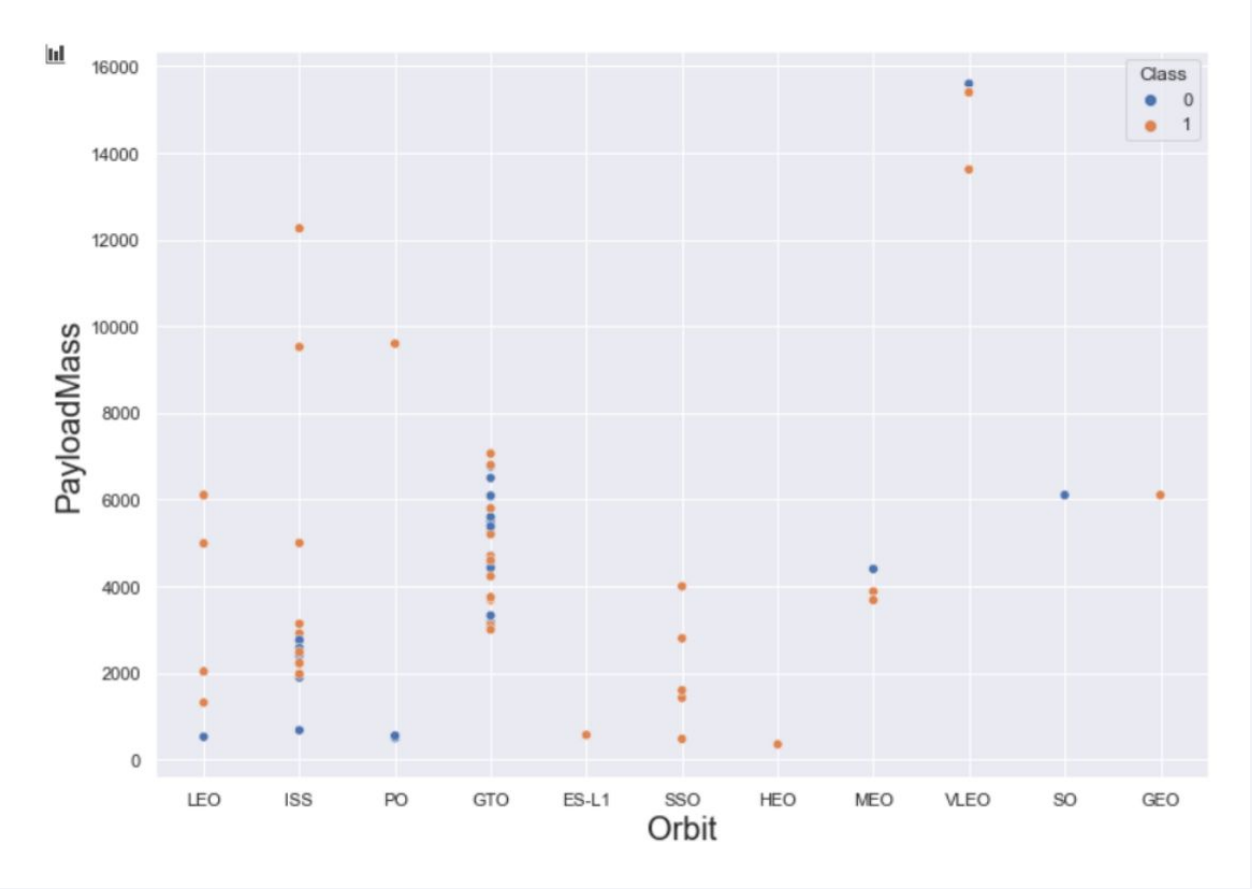
Flight Number vs. Orbit Type

- On the LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit.



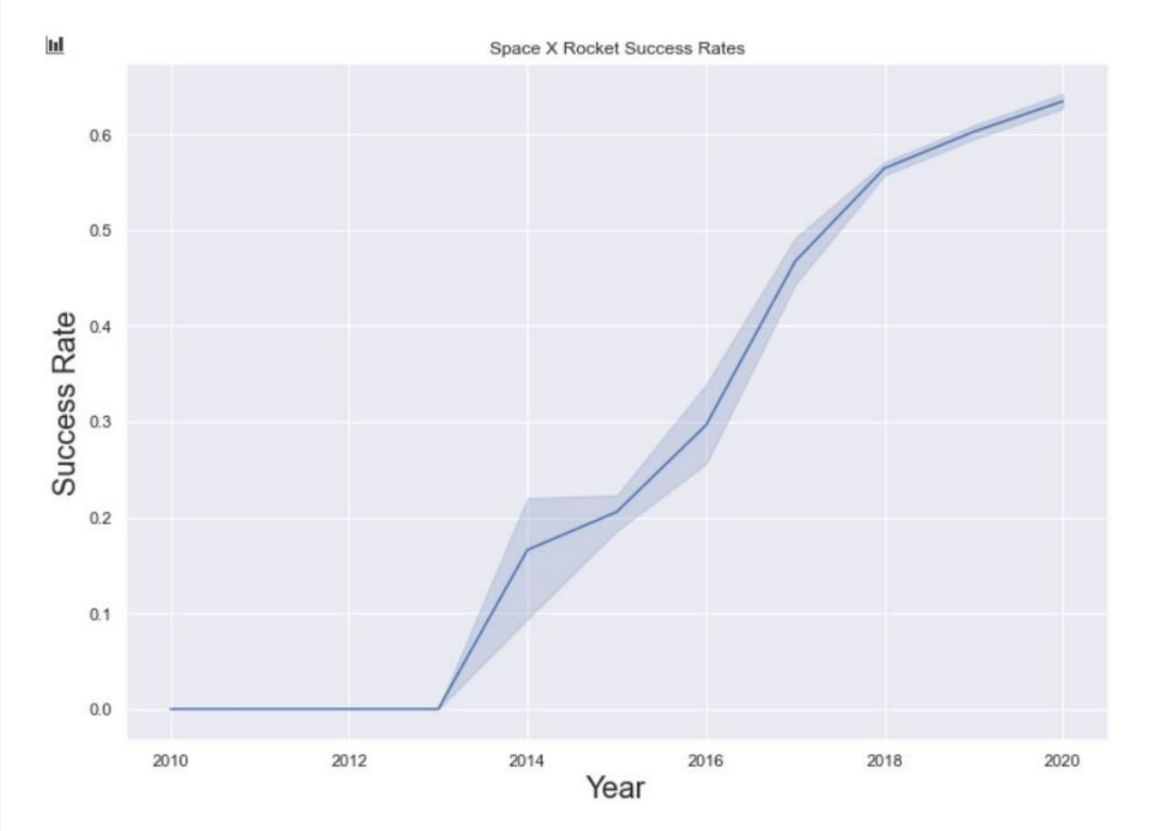
Payload vs. Orbit Type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- Success rate is increasing year over year



All Launch Site Names

SQL QUERY

```
select DISTINCT Launch_Site
from tblSpaceX
```

Unique Launch Sites
CCAFS LC-40
CCAFS SLC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- Using the word DISTINCT in the query will only return Unique values in the Launch_Site column from tblSpaceX

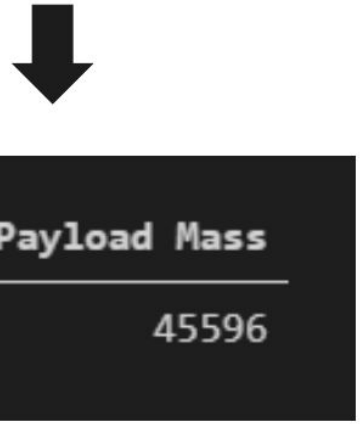
Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Total Payload Mass

SQL QUERY

```
select SUM(PAYLOAD_MASS_KG_) TotalPayloadMass from tblSpaceX
where Customer = 'NASA (CRS)'",'TotalPayloadMass
```



	Total Payload Mass
0	45596

Using the function SUM sums the total in the column PAYLOAD_MASS_KG_

The WHERE clause acts like a filter and specifies the dataset to only perform calculations on Customer NASA (CRS)

Average Payload Mass by F9 v1.1

SQL QUERY

```
select AVG(PAYLOAD_MASS_KG_) AveragePayloadMass from tblSpaceX
where Booster_Version = 'F9 v1.1'
```

Average Payload Mass
2928

- AVG function calculates the average in the column PAYLOAD_MASS_KG_
- The WHERE clause filters the dataset to only perform calculations on Booster_version F9 v1.1

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL QUERY

```
select Booster_Version from tblSpaceX where Landing_Outcome = 'Success (ground pad)'
AND Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000
```

	Date which first Successful landing outcome in drone ship was acheived.
0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

- Selecting only Booster_Version column
- The WHERE clause filters the dataset to Landing_Outcome = Success (drone ship)
- The AND clause specifies additional filter conditions Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

Total Number of Successful and Failure Mission Outcomes

SQL QUERY

```
SELECT(SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome
LIKE '%Success%') as Successful_Mission_Outcomes,
(SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome
LIKE '%Failure%') as Failure_Mission_Coutcomes
```

	Successful_Mission_Outcomes	Failure_Mission_Outcomes
0	100	1

- The LIKE wildcard shows that in the record the specified phrase is in any part of the string in the records.

Boosters Carried Maximum Payload

SQL QUERY

```
SELECT DISTINCT Booster_Version, MAX(PAYLOAD_MASS
_KG_) AS [Maximum Payload Mass]
FROM tblSpaceX GROUP BY Booster_Version
ORDER BY [Maximum Payload Mass] DESC
```

- Using the word DISTINCT in the query means that it will only show Unique values in the Booster_Version column from tblSpaceX
- GROUP BY groups the results by unique values in Booster_Version.
- DESC arranges the results into descending order

	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
...	...	...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0

97 rows × 2 columns

2015 Launch Records

SQL QUERY

```
SELECT DATENAME(month, DATEADD(month,
MONTH(CONVERT(date, Date, 105)), 0) - 1) AS Month,
Booster_Version, Launch_Site, Landing_Outcome
FROM    tblSpaceX
WHERE  (Landing_Outcome LIKE N'%Success%') AND
(YEAR(CONVERT(date, Date, 105)) = '2017')
```

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
March	F9 FT B1021.2	KSC LC-39A	Success (drone ship)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1029.2	KSC LC-39A	Success (drone ship)
June	F9 FT B1036.1	VAFB SLC-4E	Success (drone ship)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
August	F9 FT B1038.1	VAFB SLC-4E	Success (drone ship)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
October	F9 B4 B1041.1	VAFB SLC-4E	Success (drone ship)
October	F9 FT B1031.2	KSC LC-39A	Success (drone ship)
October	F9 B4 B1042.1	KSC LC-39A	Success (drone ship)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

- The function CONVERT converts NVARCHAR to Date.
- WHERE clause filters Year to be 2015

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

SQL QUERY

```
SELECT COUNT(Landing_Outcome)
FROM    tblSpaceX
WHERE  (Landing_Outcome LIKE '%Success%')
AND (Date > '04-06-2010')
AND (Date < '20-03-2017')
```

Successful Landing Outcomes Between 2010-06-04 and 2017-03-20
34

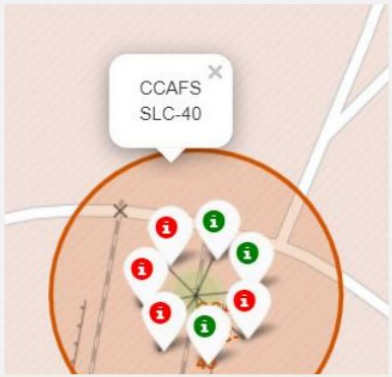
- Function COUNT counts records in column
- WHERE filters data
- LIKE AND AND(wildcard) (conditions) (conditions)

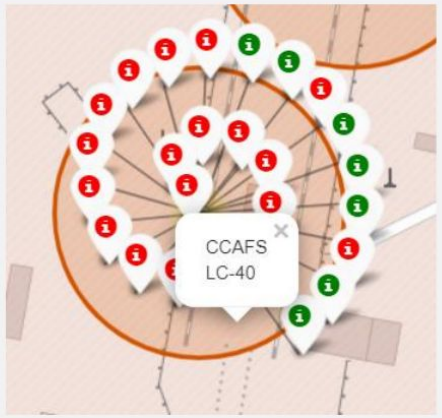

Section 3
Launch Sites
Proximities Analysis

All Launch Sites: Global Map Markers

We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

Labelled Markers





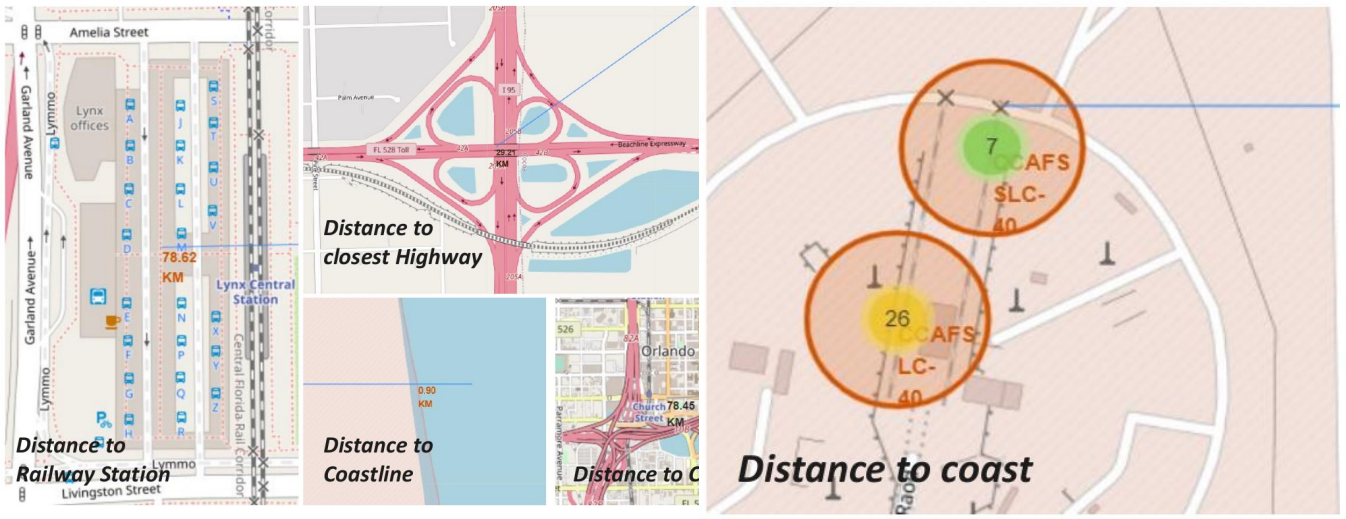
Florida Launch Sites

Green Marker shows successful Launches and Red Marker shows Failures

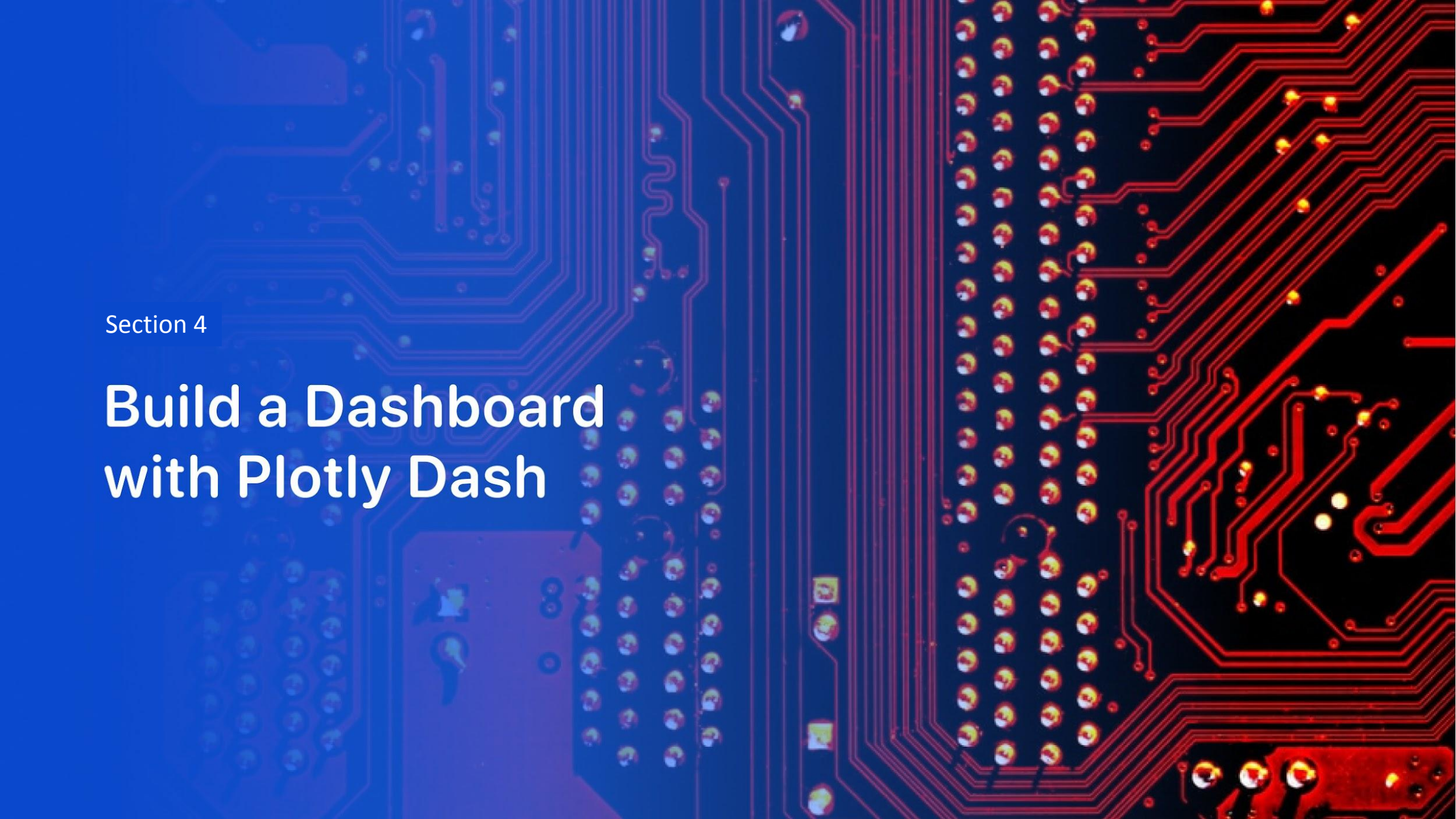
California Launch Site

37

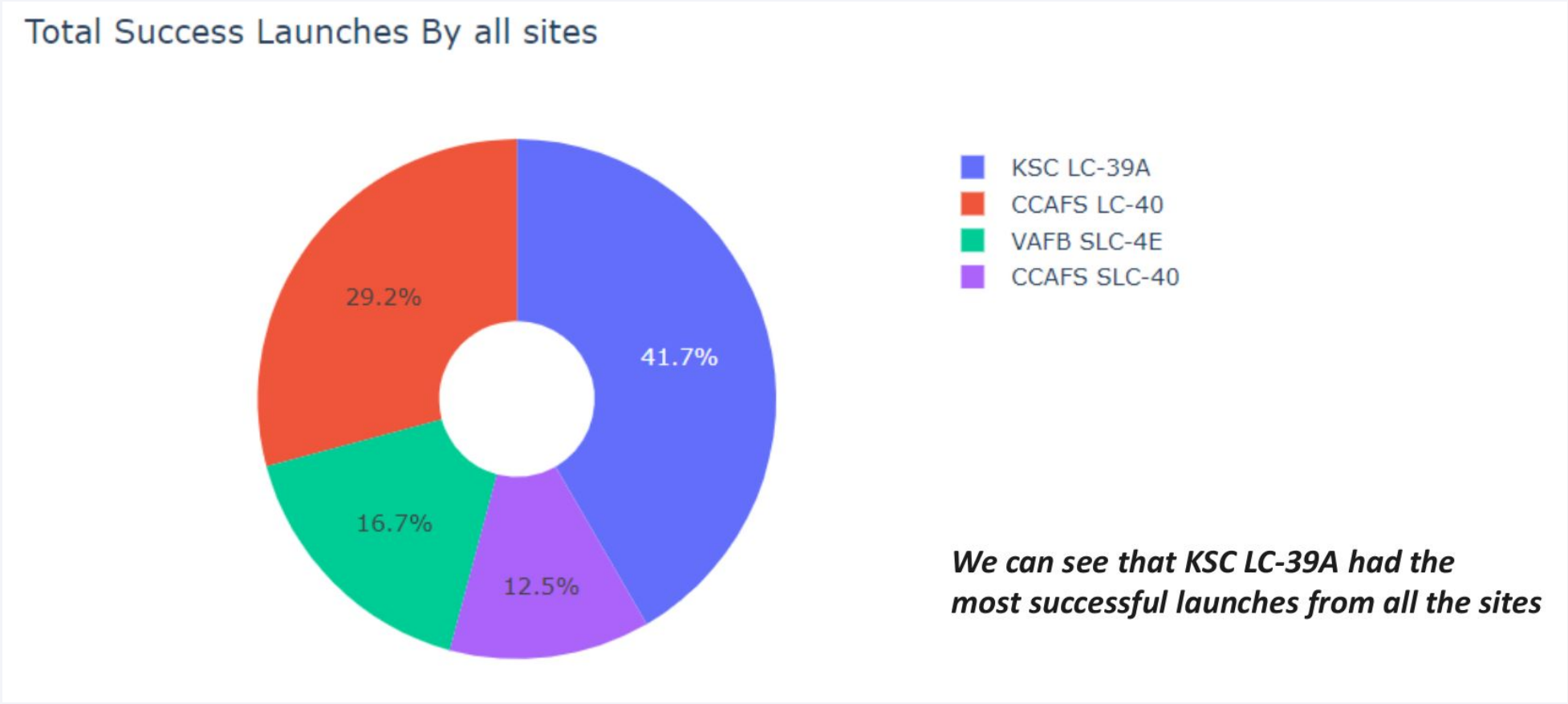
Distance to Nearby Landmarks



- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

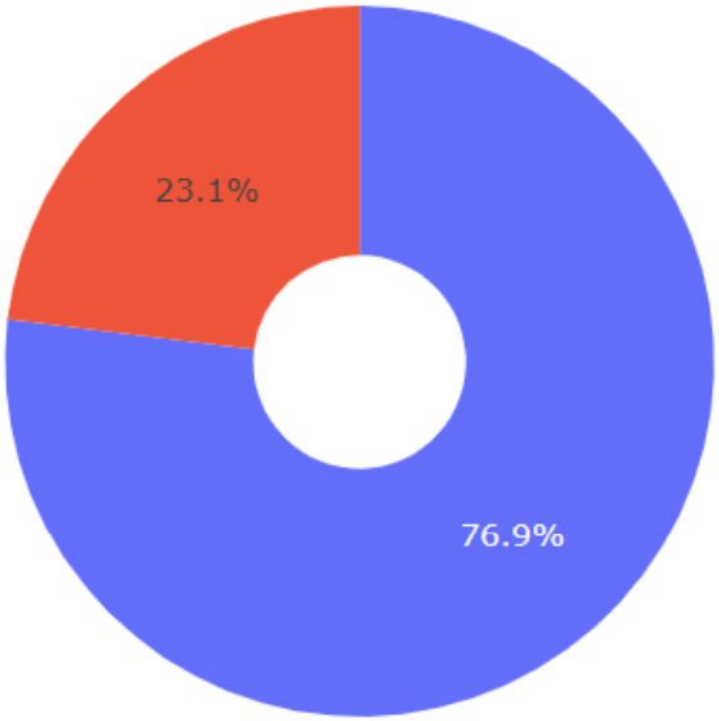

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



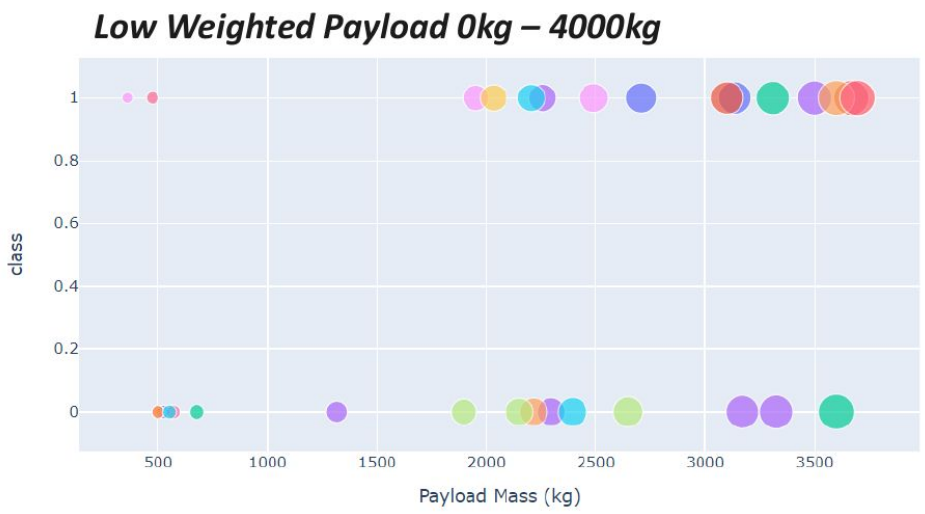
We can see that KSC LC-39A had the most successful launches from all the sites

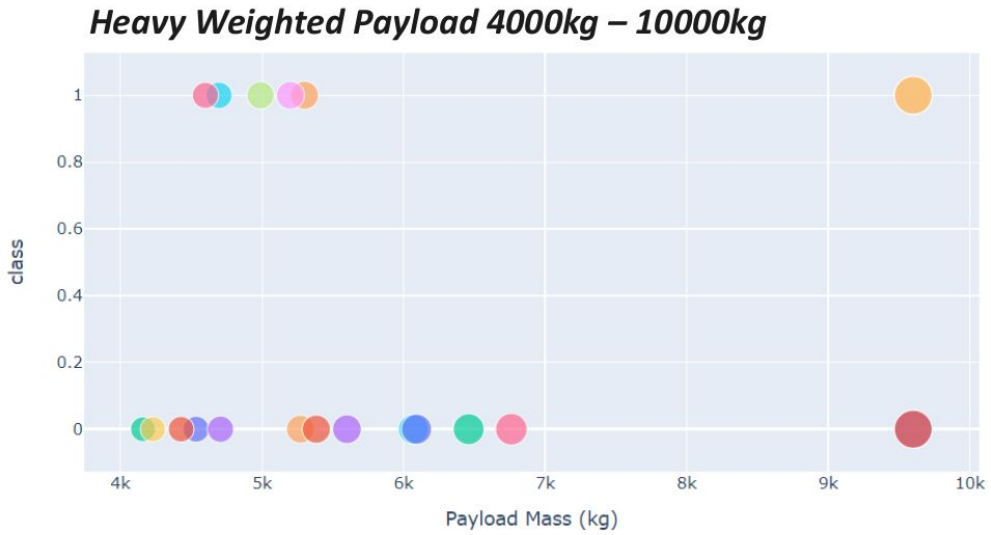
Most Successful Launch Site



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Class vs Payload Mass





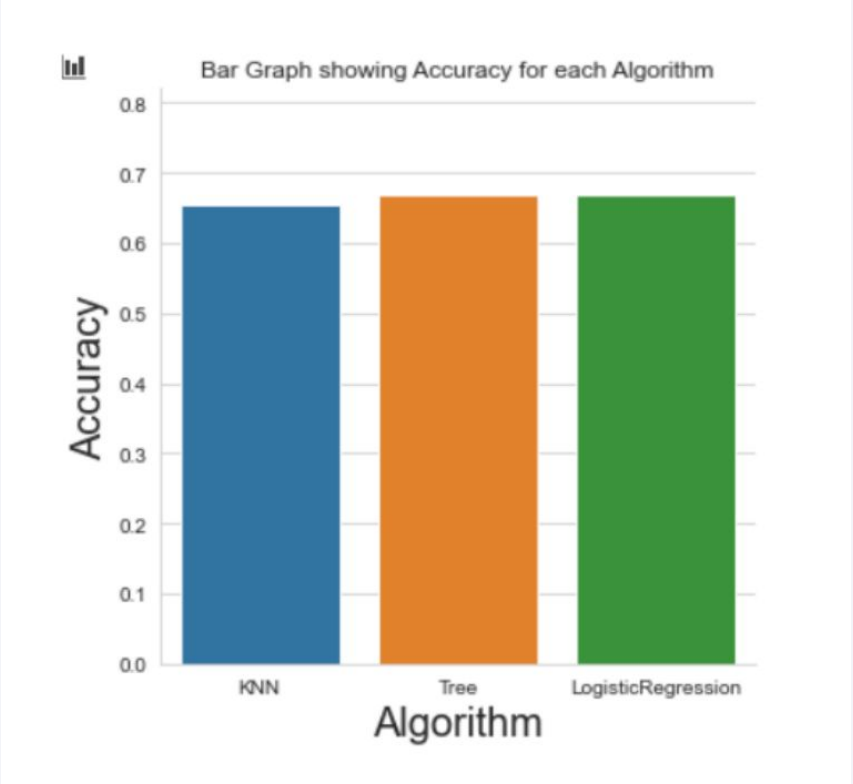
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

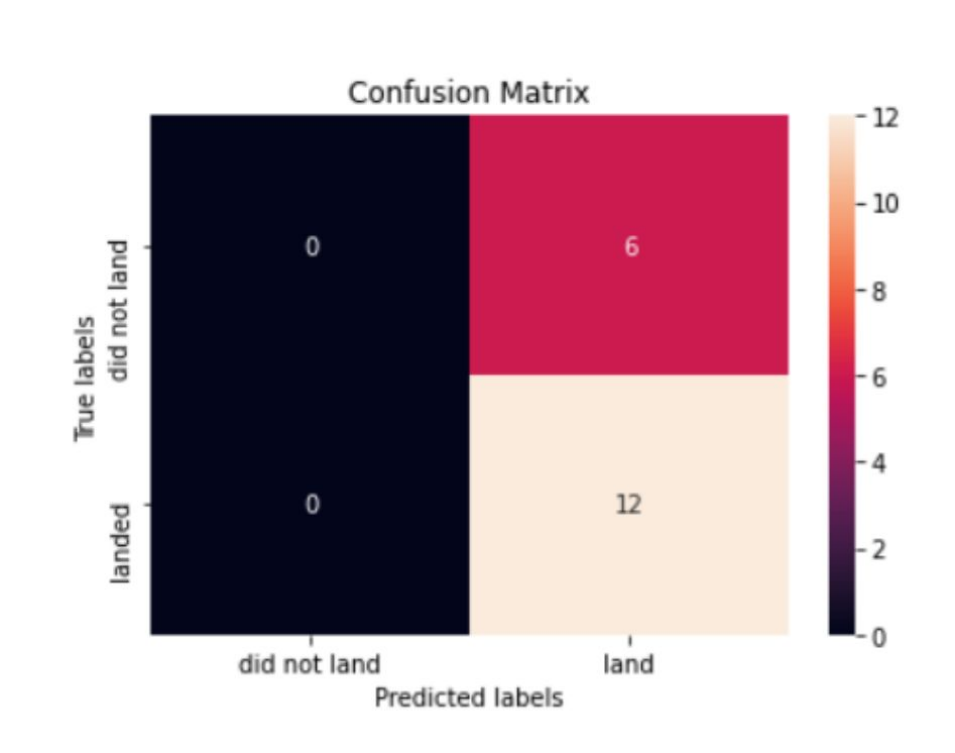
Predictive Analysis (Classification)

Classification Accuracy

- The results were close but the decision tree algorithm scored the highest for this use case.



Confusion Matrix



- Examining the confusion matrix, we see that Decision Tree Algorithm can distinguish between the different classes.
- The major hurdle is false positives.

Conclusions

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is increasing significantly year over year
- KSC LC-39A had the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

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!