IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Data Science Capstone Project

Lorita Rahimi, 2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection via API, Web Scraping
  - Exploratory Data Analysis (EDA) with Data Visualization
  - EDA with SQL
  - Interactive Map with Folium
  - Dashboards with Plotly Dash
  - Predictive Analysis
- **Summary of all results**
  - Exploratory Data Analysis results
  - Intereactive maps and dashboard
  - Predictive results

Introduction

- Project background and context
  - The objective of this project is to forecast the successful landing of the Falcon 9 first stage. According to SpaceX's website, the cost of launching a Falcon 9 rocket is $62 million, whereas other providers charge over $165 million per launch. This price disparity is due to SpaceX's ability to reuse the first stage. By accurately determining the landing outcome, we can ascertain the total launch cost. This information holds significance for any company intending to rival SpaceX in the rocket launch market.
- Problems you want to find answers
  - What are the primary attributes distinguishing a successful or unsuccessful landing?
  - How do various rocket variables affect the outcome of a landing, and what are their respective impacts?
  - What conditions need to be met for SpaceX to attain the highest rate of landing success?

Section 1

Methodology

Methodology

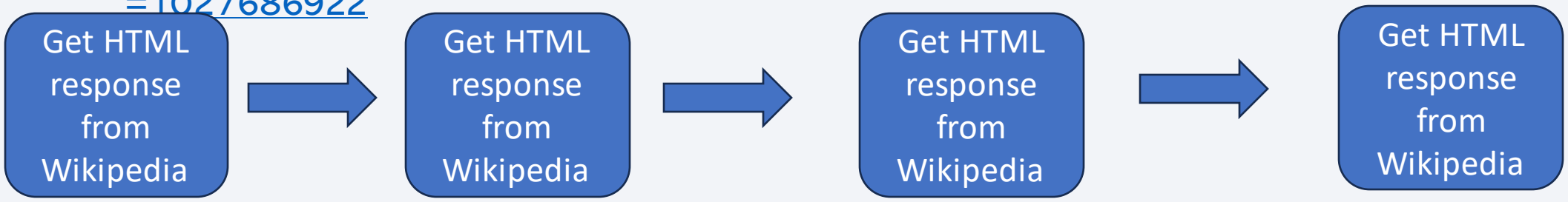
Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- 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

- Datasets are collected from SpaceX API and webscrapping Wikipedia.
  - The SpaceX API URL is api.spacexdata.com/v4/

- You need to present your data collection process use key phrases and flowcharts
  - The Wikipedia URL is https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



Data Collection – SpaceX API

- 1.Requesting rocket launch data from SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)
```

- 2.Converting Response to a JSON file

```
# Use json_normalize meethod to convert the json result into a dataframe
data=pd.json_normalize(response.json())
```

- 3.Using custom functions to clean data

```
# Call getLaunchSite
getLaunchSite(data)

# Call getPayloadData
getPayloadData(data)
```

```
# Call getCoreData
getCoreData(data)
```

- 4.Combining the columns into a dictionary to create data frame

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

- 5.Filtering dataframe and exporting to a CSV

```
data_falcon9 = data.loc[data['BoosterVersion']!='Falcon 1']
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

GitHub URL

Data Collection - Scraping

- 1.Getting response from HTML

```
html_page = requests.get(static_url)
```

- 2.Creating a BeautifulSoup object

```
soup = BeautifulSoup(html_page.text, 'html.parser')
```

- 3.Finding all tables and assigning the result to a list

```
html_tables = soup.find_all('table')
```

- 4.Extracting column name one by one

```
column_names = []

for i in first_launch_table.find_all('th'):
    if extract_column_from_header(i) != None and len(extract_column_from_header(i)) > 0:
        column_names.append(extract_column_from_header(i))
```

- 5.Creating an empty dictionary with keys

```
launch_dict= dict.fromkeys(column_names)
# Remove an irrelvant column
del launch_dict['Date and time ( )']
# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

- 6.Filling up the launch_dict with launch records (please refer to the notebook)

- 7.Creating a Dataframe and exporting it to a CSV

```
df=pd.DataFrame(launch_dict)
```

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

GitHub URL

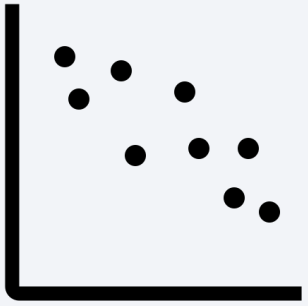
Data Wrangling

- There have been instances where the booster was unable to successfully land on the dataset, and in certain cases, it made landing attempts but failed due to accidents.
  - True Ocean: the mission result has successfully landed in a specific area of the ocean
  - False Ocean: the mission result has not successfully landed in a specific area of the ocean
  - True RTLS: the mission result successfully landed on the ground pad
  - False RTLS: the mission result has not successfully landed on the ground pad
  - True ASDS: the mission result has successfully landed on the drone ship
  - False ASDS: the mission result has not landed on the drone ship
- Coverting these results into training labels:
  - 1 = successful / 0 = failure

GitHub URL

EDA with Data Visualization

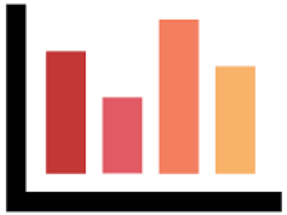
- Scatter Graphs
  - ❖ Flight Number vs. Payload Mass
  - ❖ Flight Number vs. Launch Site
  - ❖ Payload vs. Launch Site
  - ❖ Orbit vs. Flight Number
  - ❖ Payload vs. Orbit Type
  - ❖ Orbit vs. Payload Mass



The pattern of dots on the graph can provide valuable insights into the correlation or lack thereof between the two variables

- Bar Graph
  - ❖ Success rate vs Orbit

Bar graphs show the relationship between numeric and categoric variables.



- Line Graph
  - ❖ Success rate vs. Year

Line graphs show data variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data

GitHub URL

EDA with SQL

- Executing SQL queries to answer these questions:
  - 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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site 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

Build an Interactive Map with Folium

- Objects incorporated and included in a folium map:
  - Markers illustrating all launch sites on the map.
  - Markers denoting the success or failure of launches for each respective site on the map.
  - Lines indicating the distances between a launch site and its nearby locations.

- The following questions are easily answered:
  - Are launch sites in close proximity to railways? YES
  - Are launch sites in close proximity to highways? YES
  - Are launch sites in close proximity to coastline? YES
  - Do launch sites keep certain distance away from cities? YES

GitHub URL

Build a Dashboard with Plotly Dash

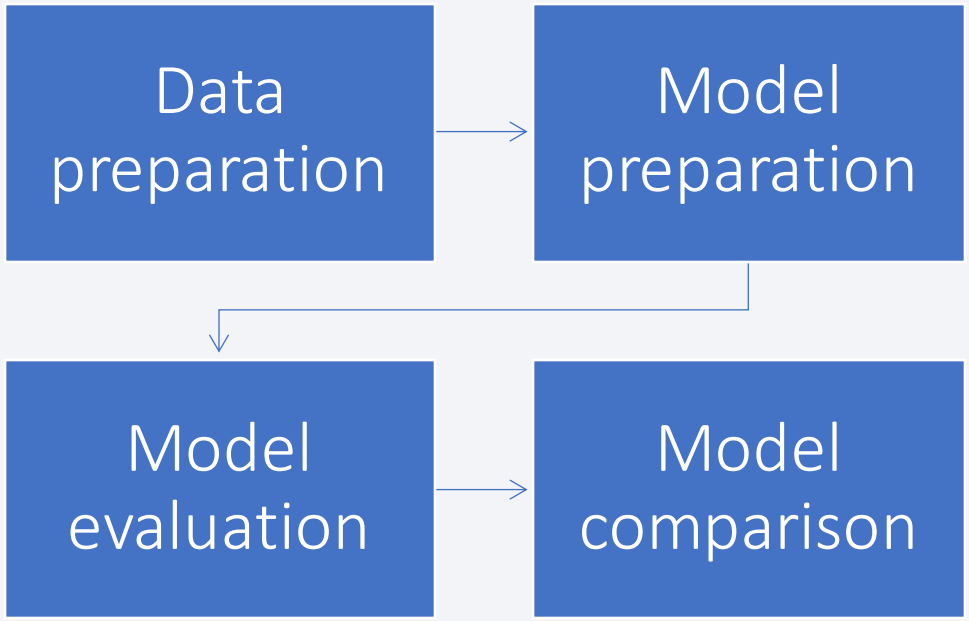
- The dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
  - ❑ Dropdown allows a user to chose the launch site or all launch sites.
  - ❑ Pie chart shows the total success and the total failure for the launch site chosen.
  - ❑ Range slider allows a user to select a payload mass in a fixed range
  - ❑ Scatter chart shows relationship between two variables (Success vs Payload Mass)

GitHub URL

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
  - ❖ Find the method performs best using test data

GitHub URL

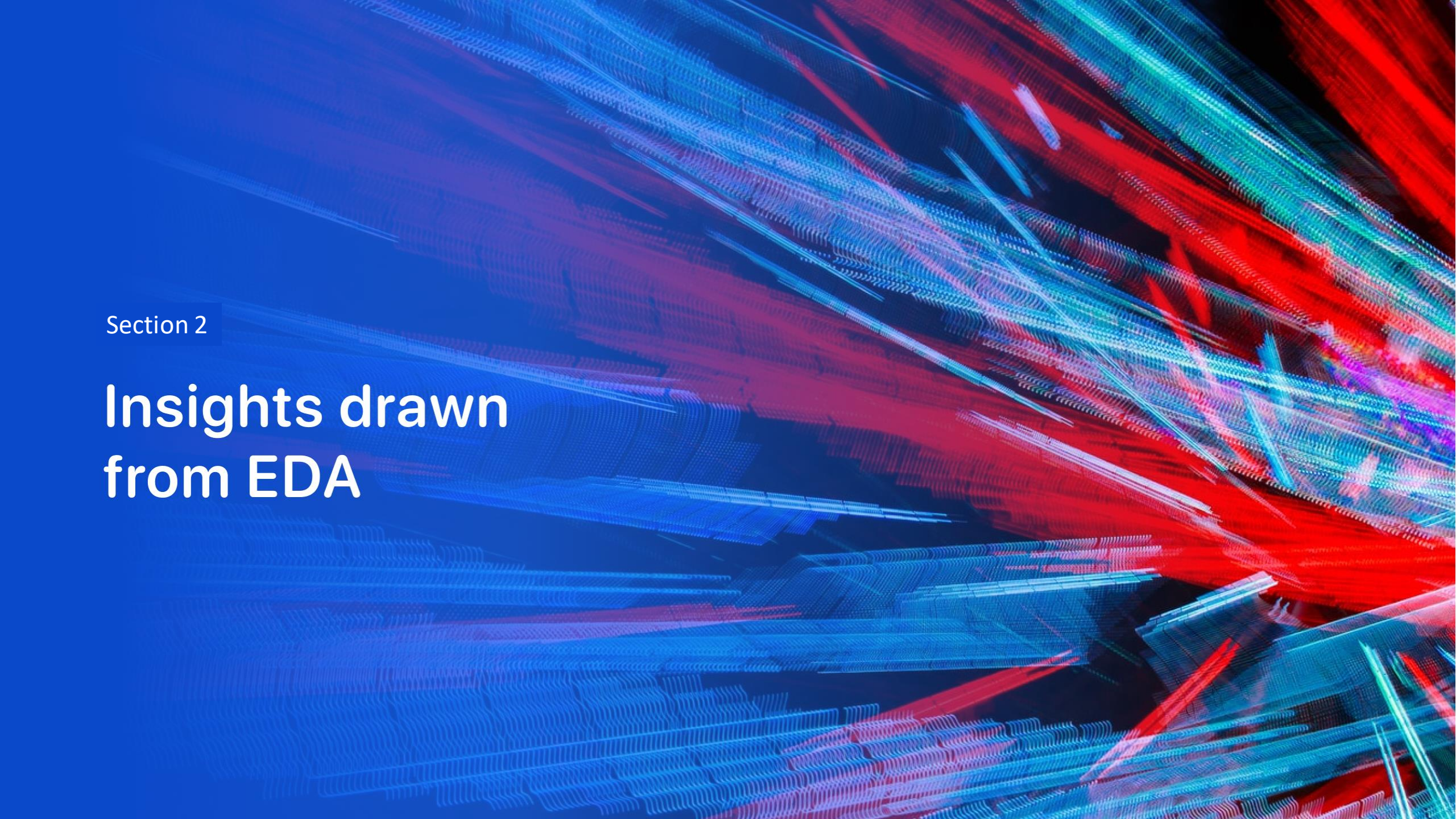


Results

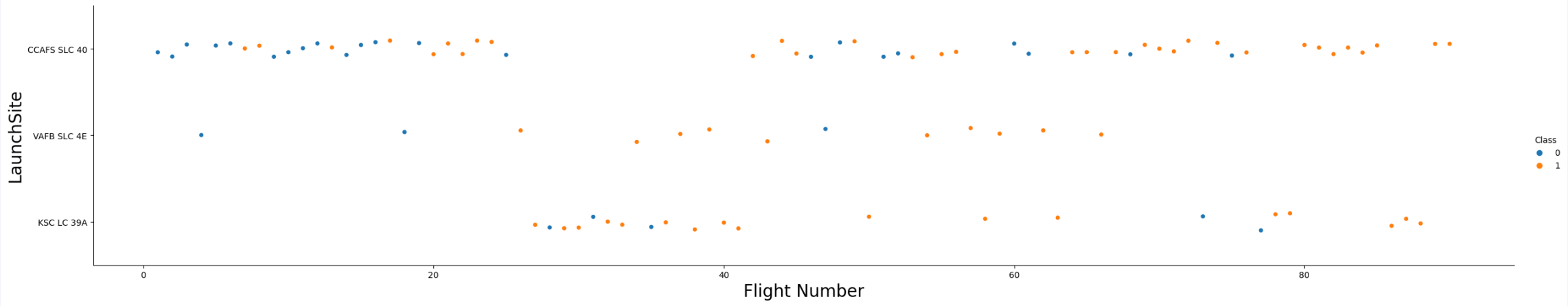
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

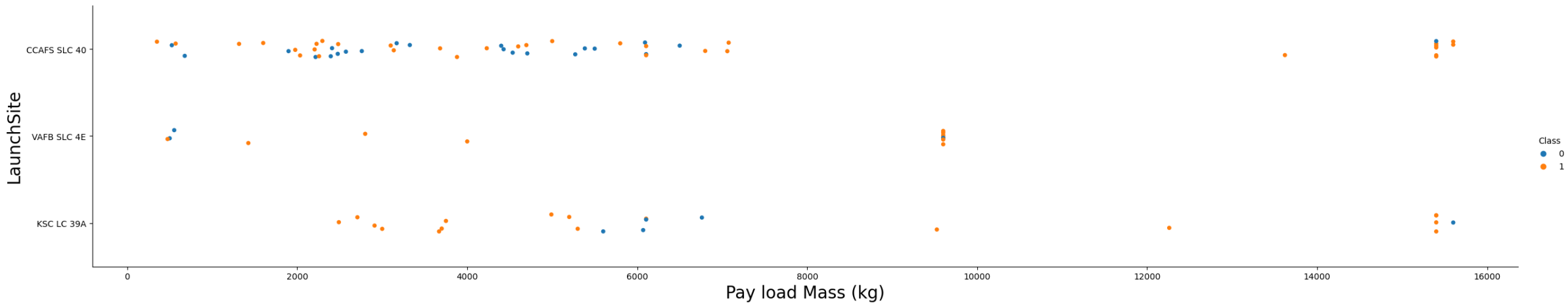


Flight Number vs. Launch Site



- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- This figure shows that the success rate increased as the number of flights **increased**.

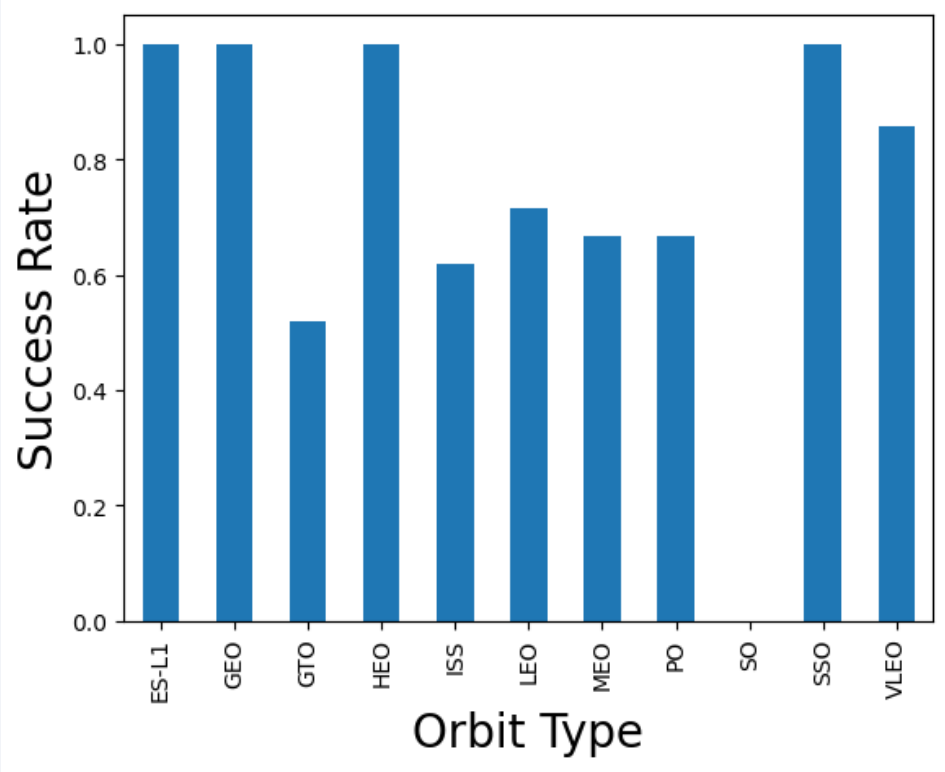
Payload vs. Launch Site



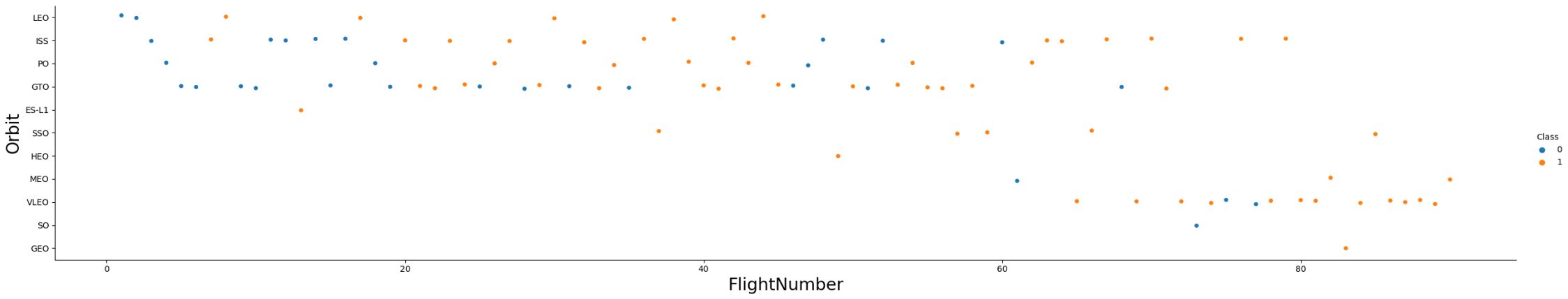
- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- Upon initial observation, there appears to be a correlation between a rocket's success rate and its larger payload mass. However, making decisions solely based on this correlation may be challenging as there is no clear and consistent pattern between successful launches and payload mass.

Success Rate vs. Orbit Type

- By examining this plot, we can observe the success rates for various orbit types. It becomes evident that **ES-L1**, **GEO**, **HEO**, and **SSO** orbits exhibit the highest success rates among the data.

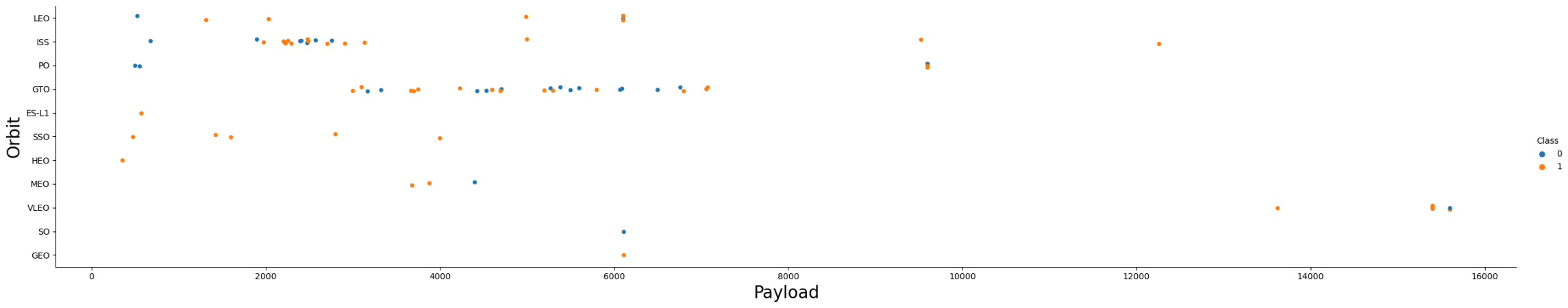


Flight Number vs. Orbit Type



- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- In most cases, the launch outcome seems to be correlated with the flight number. But for some orbits like GTO, there is no relation between the success rate and the number of flights.

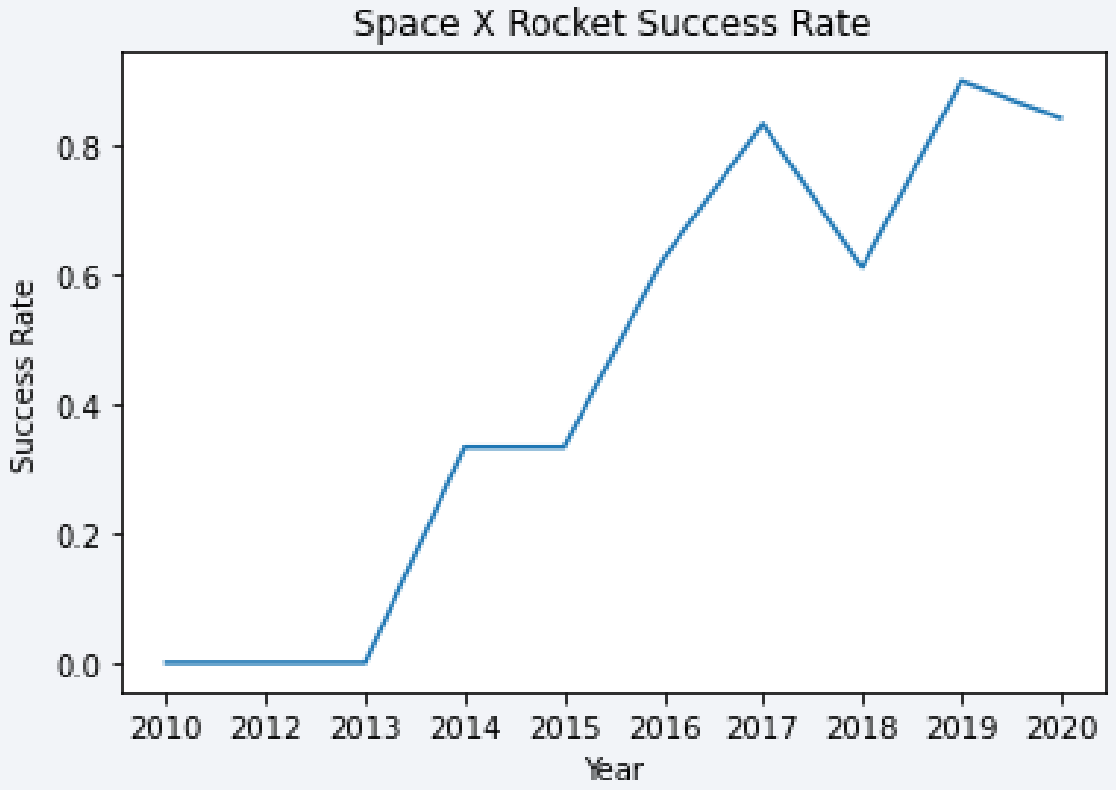
Payload vs. Orbit Type



- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- The weight of payloads can significantly impact the success rate of launches in specific orbits. For instance, heavier payloads tend to enhance the success rate for the LEO. Conversely, reducing the payload weight for a GTO has been found to improve the launch success.

Launch Success Yearly Trend

- Starting from 2013, there has been a noticeable rise in the success rate of SpaceX rockets.



All Launch Site Names

❖ Query

```
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

❖ Result

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

❖ When the SQL DISTINCT clause is used in the query, only unique values are displayed in the Launch_Site column from the SpaceX table.

Launch Site Names Begin with 'CCA'

❖ Query

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

❖ The WHERE clause followed by LIKE clause filters launch sites that contain the substring CCA. LIMIT 5 shows 5 records from filtering.

❖ Result

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

❖ Query

```
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM FROM SPACEXTBL WHERE
CUSTOMER LIKE 'NASA (CRS)'
```

❖ Result

sum
45596.0

❖ Using the SUM() function to calculate the sum of column PAYLOAD_MASS__KG_.

Average Payload Mass by F9 v1.1

❖ Query

```
SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE
FROM SPACEXTBL
WHERE BOOSTER_VERSION LIKE 'F9 V1.1%'
```

❖ Result

average
2534.6666666666665

❖ Using the AVG() function to calculate the average value of column PAYLOAD_MASS__KG_.

First Successful Ground Landing Date

❖ Query

```
SELECT MIN(DATE) AS DATE
FROM SPACEXTBL
WHERE MISSION_OUTCOME LIKE 'SUCCESS'
```

❖ Result

date
01/06/2014

❖ Using the MIN() function to find out the earliest date in the column DATE.

Successful Drone Ship Landing with Payload between 4000 and 6000

❖ Query

```
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (drone ship)'
    AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

❖ Result

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

❖ This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg. The WHERE and AND clauses filter the dataset.

Total Number of Successful and Failure Mission Outcomes

❖ Query

```
SELECT MISSION_OUTCOME, COUNT(*) AS COUNT
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
ORDER BY MISSION_OUTCOME
```

❖ Result

Mission_Outcome	count
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

❖ This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg. The WHERE and AND clauses filter the dataset.

Total Number of Successful and Failure Mission Outcomes

❖ Query

```
SELECT MISSION_OUTCOME, COUNT(*) AS COUNT
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
ORDER BY MISSION_OUTCOME
```

❖ Result

Mission_Outcome	count
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

❖ Using the COUNT() function to calculate the total number of columns. Using the GROUP BY statement, groups rows that have the same values into summary rows to find the total number in each Mission_outcome.

Boosters Carried Maximum Payload

❖ Query

```
SELECT BOOSTER_VERSION
FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM
SPACEXTBL)
```

❖ Using a subquery, first, find the maximum value of the payload by using MAX() function, and second, filter the dataset to perform a search if PAYLOAD_MASS__KG_ is the maximum value of the payload.

❖ Result

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

❖ Query

```
select substr(Date, 4, 2) as month,
landing_outcome,
booster_version,
launch_site
from SPACEXTBL
where substr(Date,7,4)='2015'
```

❖ Result

month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
11	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
02	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

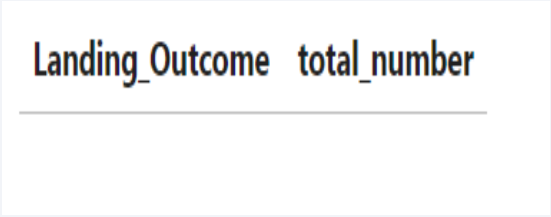
❖ This query returns month, booster version, launch site where landing was unsuccessful and landing date took place in 2015. Substr function process date in order to take month or year. Substr(DATE, 4, 2) shows month. Substr(DATE,7, 4) shows year.

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

❖ Query

```
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME)
AS total_number
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY total_number DESC
```

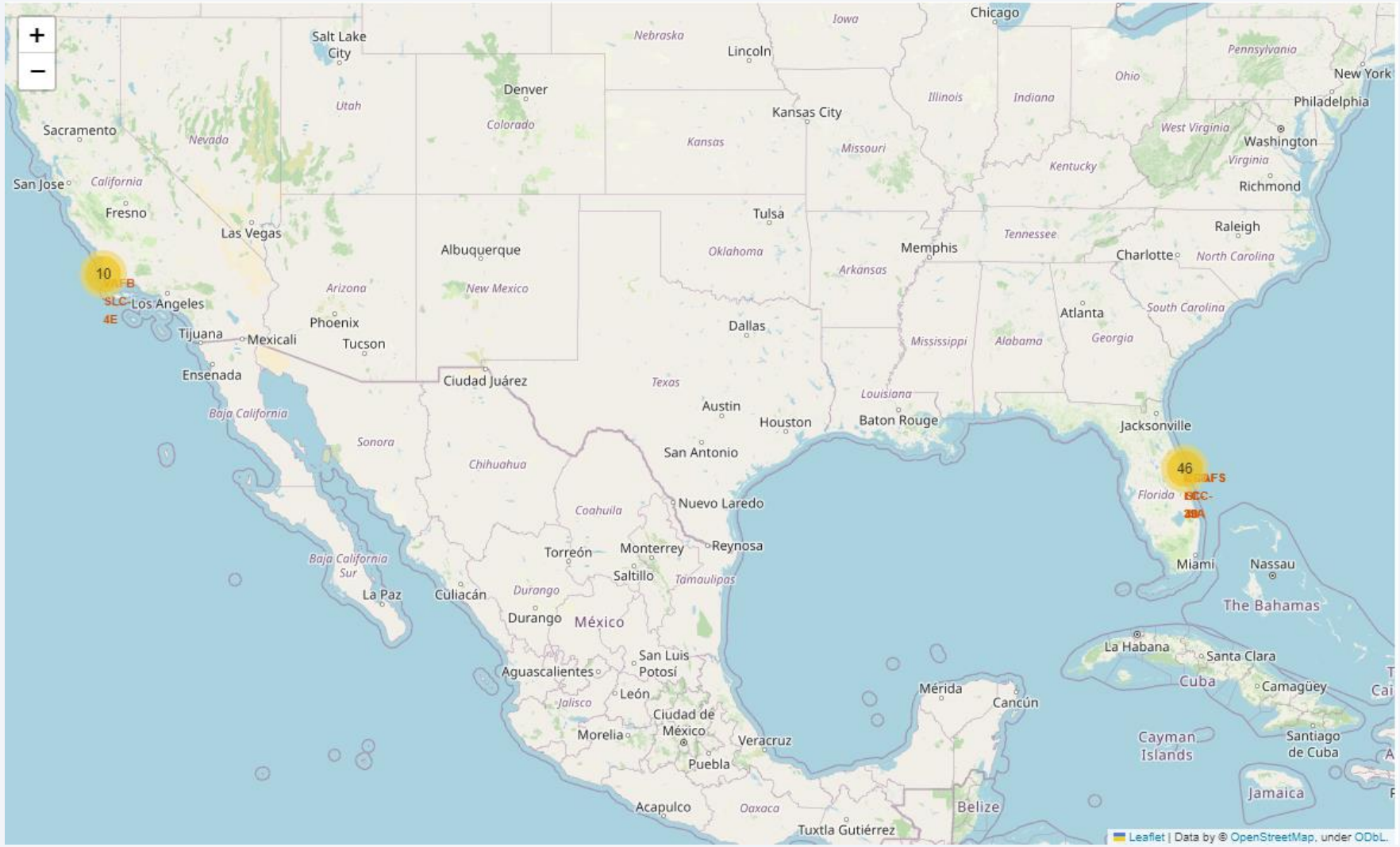
❖ Result



❖ In the WHERE clause, filter the dataset to perform a search if the date is between 2010-06-04 and 2017-03-20.

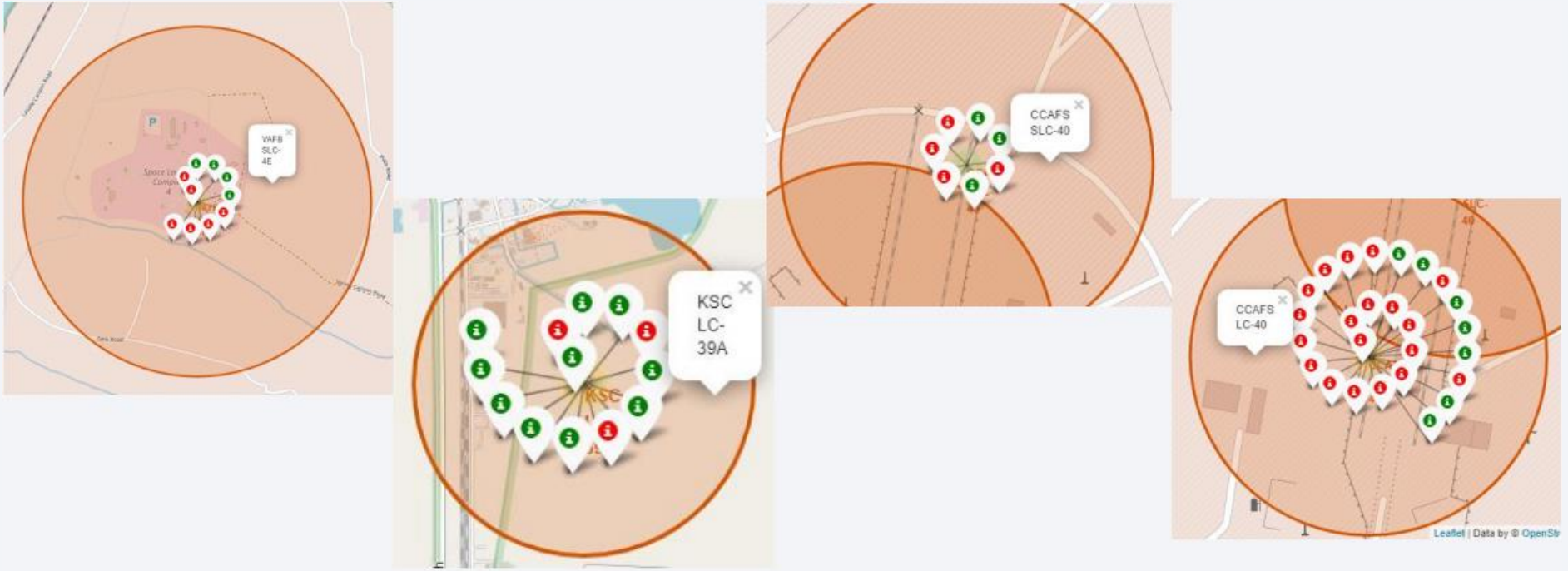

Section 3
Launch Sites
Proximities Analysis

Folium Map – Ground stations



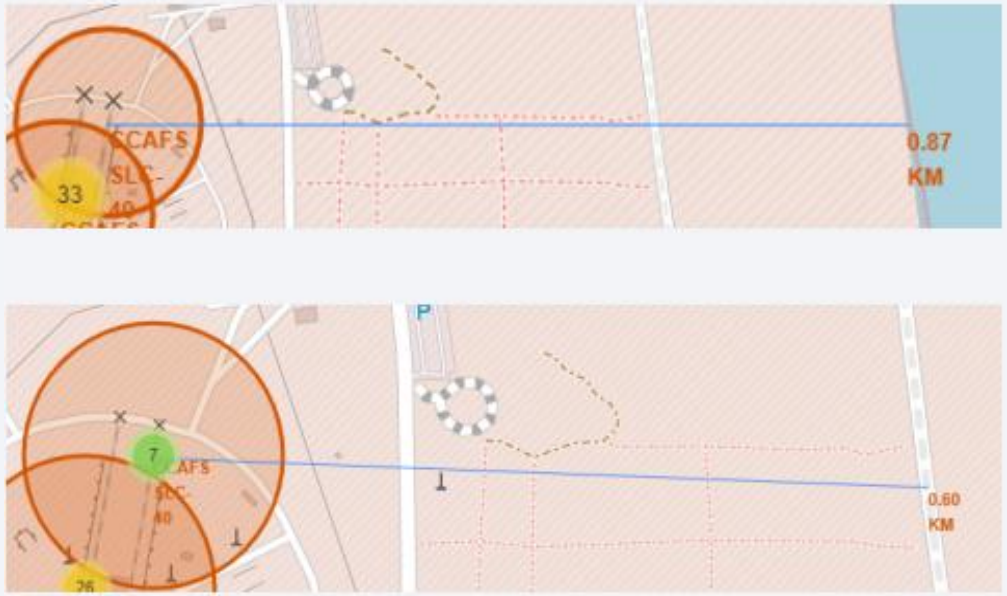
- We see that Space X launch sites are located on the coast of the United States

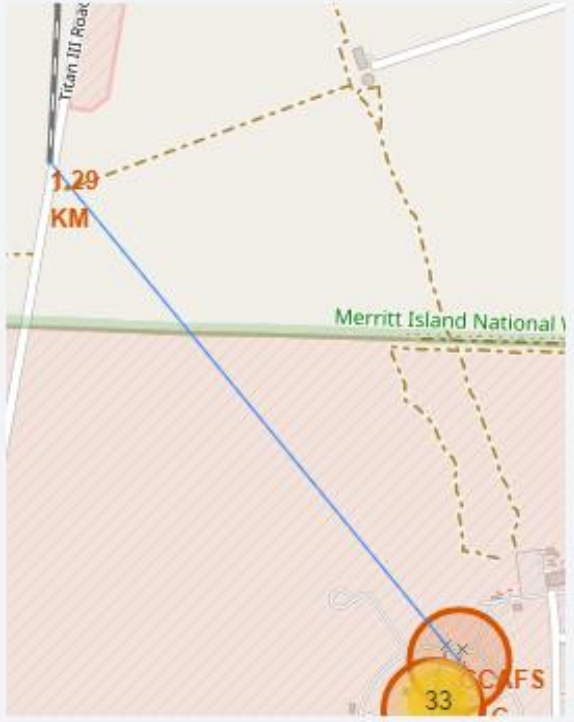
Folium Map – Color Labeled Markers



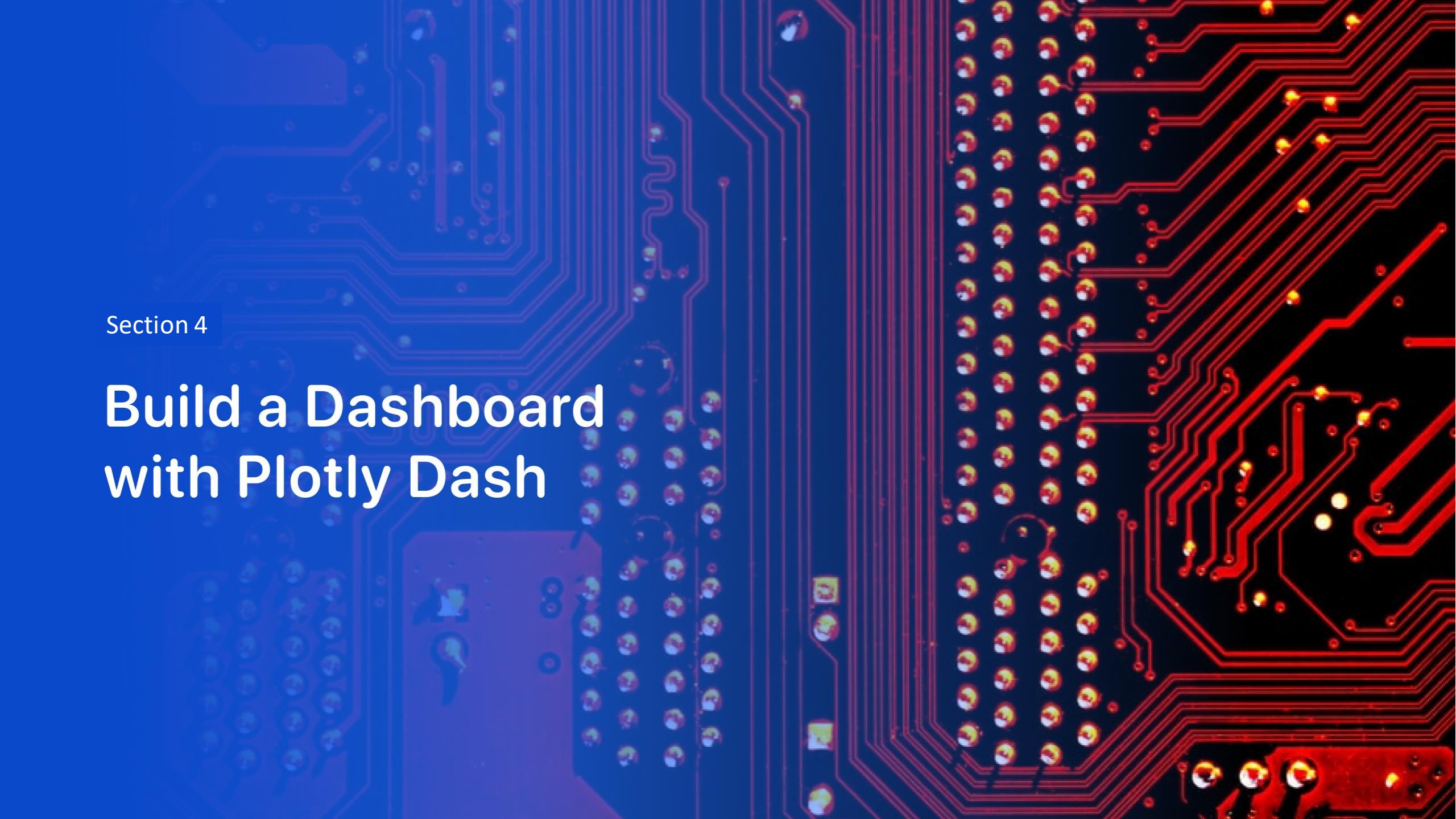
- Green marker represents successful launches. Red marker represents unsuccessful launches. We note that KSC LC-39A has a higher launch success rate.

Folium Map – Proximities of Launch Sites

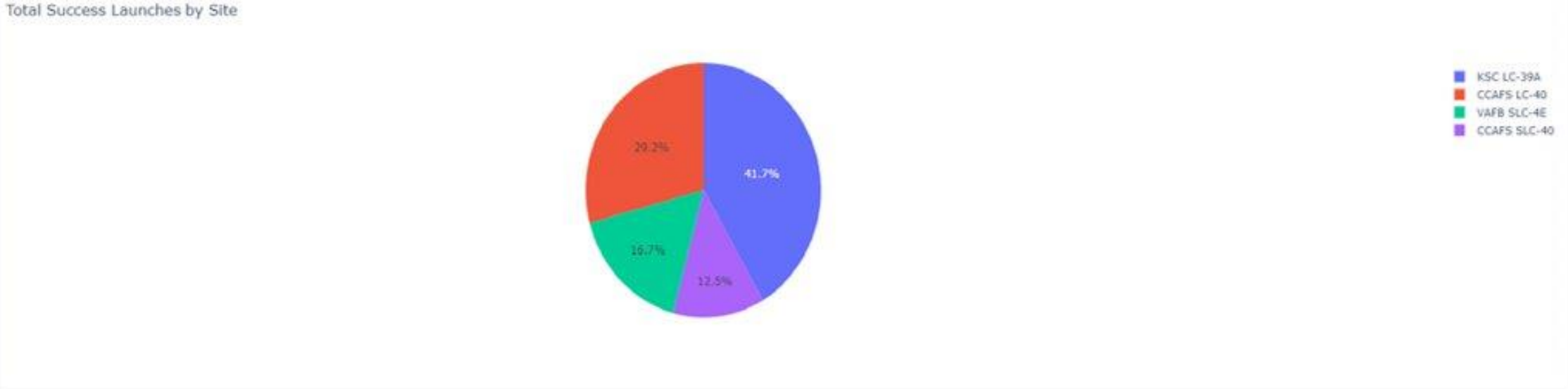




Is CCAFS SLC-40 in close proximity to railways ? Yes
Is CCAFS SLC-40 in close proximity to highways ? Yes
Is CCAFS SLC-40 in close proximity to coastline ? Yes
Do CCAFS SLC-40 keeps certain distance away from cities ? No

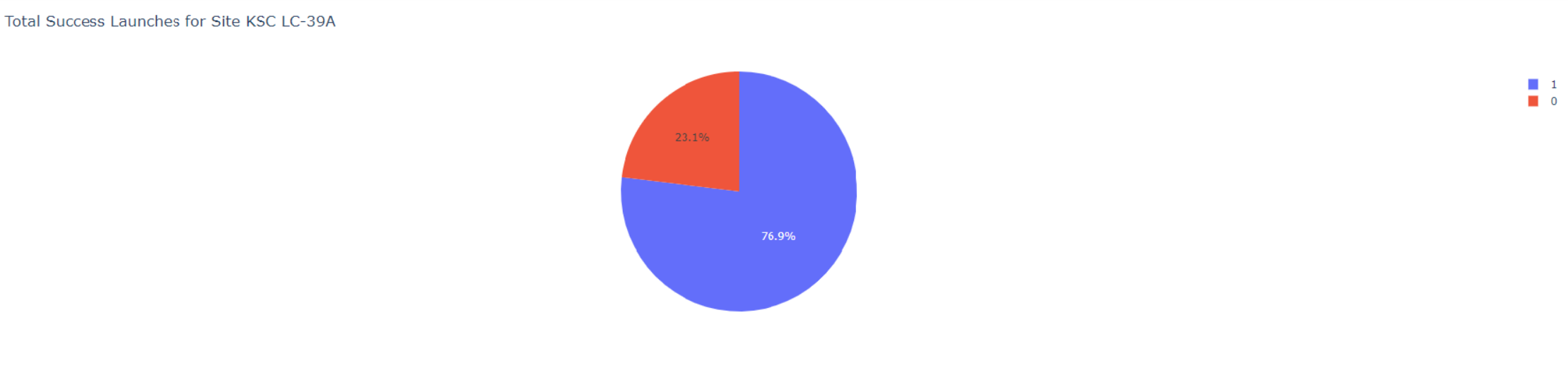

Section 4
Build a Dashboard with Plotly Dash

Dashboard – Total Success Launches by Site



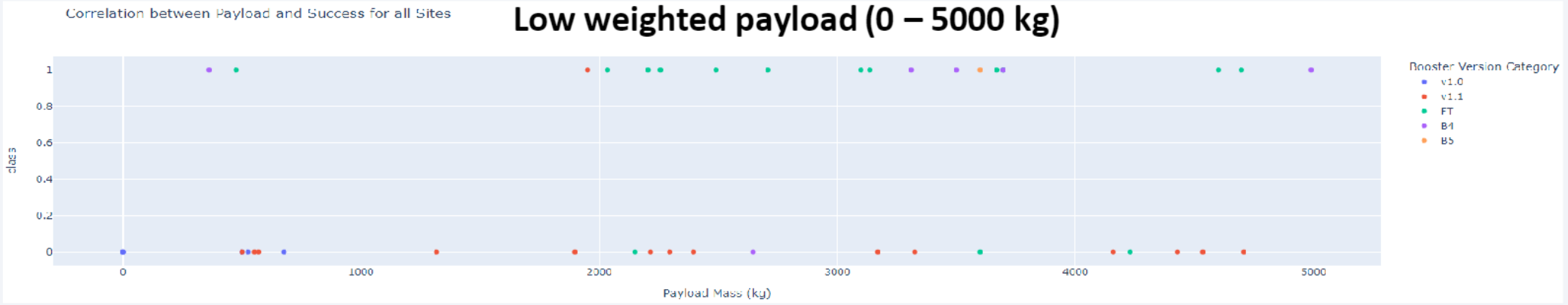
- KSC LC-39A has the best success rate of launches.

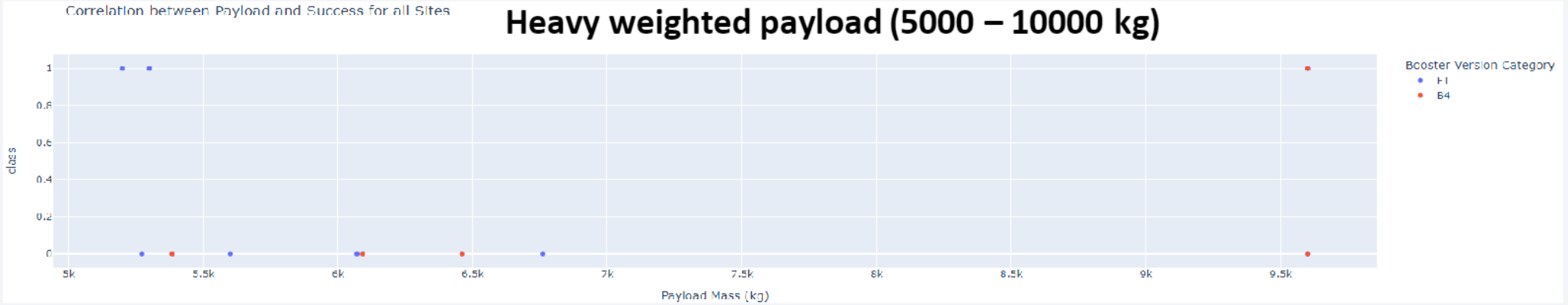
Dashboard – Total success launches for Site KSC LC-39A



- The launch site KSC LC-39A has attained an impressive success rate of 76.9% with a failure rate of 23.1%.

Dashboard - Payload vs. Launch Outcome for all sites with different payload selected in the range slider



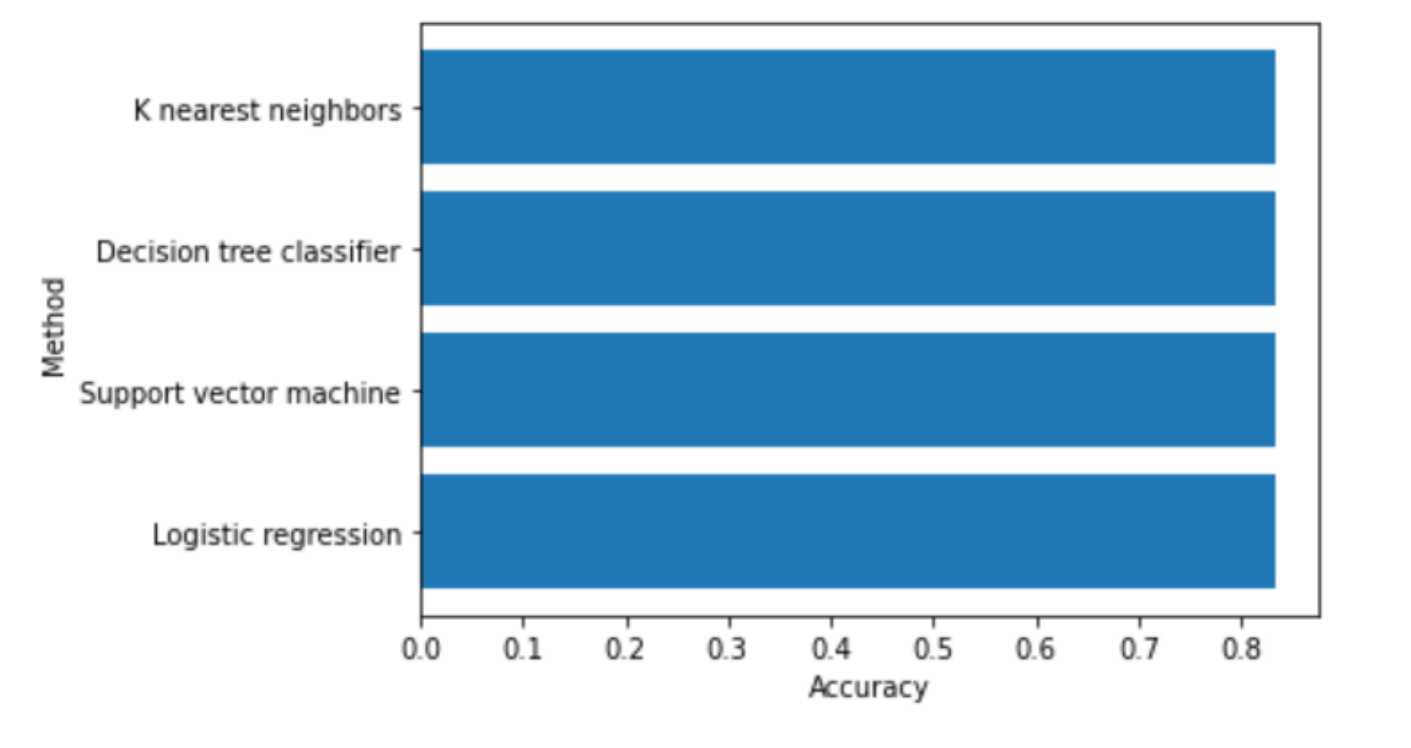


- The data suggests that low-weighted payloads exhibit a higher success rate compared to heavy-weighted payloads.

Section 5

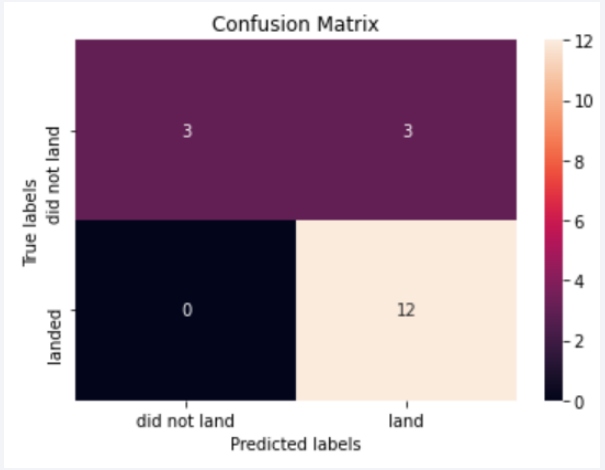
Predictive Analysis (Classification)

Classification Accuracy

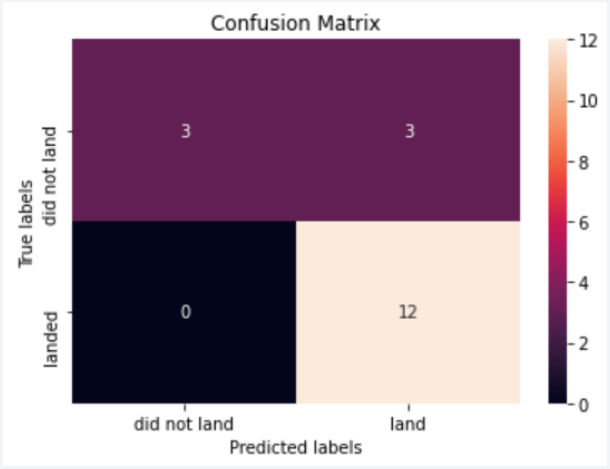


- For accuracy test, all methods performed similar. We could get more test data to decide between them. But if we really need to choose one right now, we would take the decision tree.

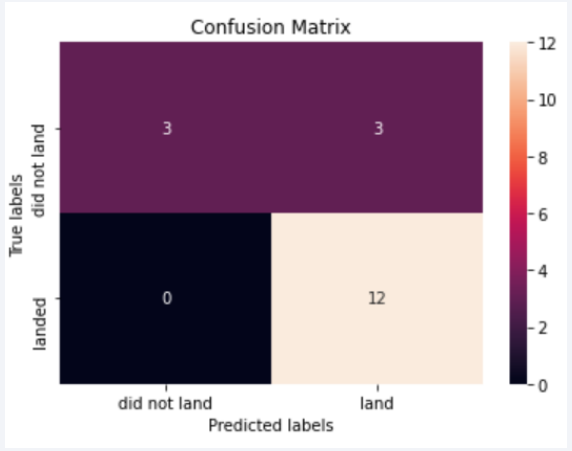
Confusion Matrix



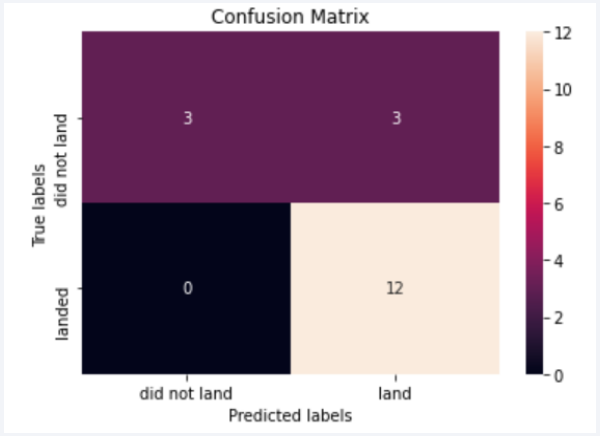
Logistic regression



SVM



Decision Tree



kNN

- The models correctly predicted 12 successful landings when the true label was indeed successful and accurately identified 3 failed landings when the true label was failure. However, the models also had 3 false positive predictions, indicating successful landings when the true label was failure.
- Overall, these models predict successful landings.

Conclusions

- The success of a mission can be attributed to various factors, including the launch site, the orbit, and notably the number of previous launches. It is reasonable to assume that there has been a significant knowledge gain over time, enabling the progression from launch failures to successful missions. This accumulated experience likely plays a crucial role in achieving successful outcomes.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- KSLC-39A has the highest number of launch successes and the highest success rate among all sites.
- In general, it appears that low-weighted payloads tend to outperform heavy-weighted payloads.
- In this dataset, all models have the same accuracy (83.33%), but it seems that more data is needed to determine the optimal model due to the small data size.

Appendix

- GitHub URL
- Applied Data Science Capstone Course URL

Thank you!