Winning Space Race with Data Science

Dario

06.03.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via API, SQL. And Web Scraping
  - Data Wrangling and analysis
  - Interactive Maps with Folium
  - Predictive Analysis for each classification model
- Summary of all results
  - Data Analysis along with Interactive Visualization
  - Best model for Predictive Analysis

Introduction

- Project background and context

Here we will predict if the Falcon 9 first stage will land succefully. SpaceX advertise Falcon 9 rocket launches on its website, with a cost of 62 millions dollars; other providers cost upwards of 165 millions dollars each, much of the saving is because SpaceX can reuse the first stage. Therefore, it crucial to determine if the first stage will land successfully. This information can be used if an alternative company wants to bid against SpaceX for a rocket launch.

- Problems we want to find answers
  - With what factors, the rocket will land successfully?
  - The effect of each relationship of rocket variables on outcome
  - Conditions which will aid SpaceX have to achieve the best result

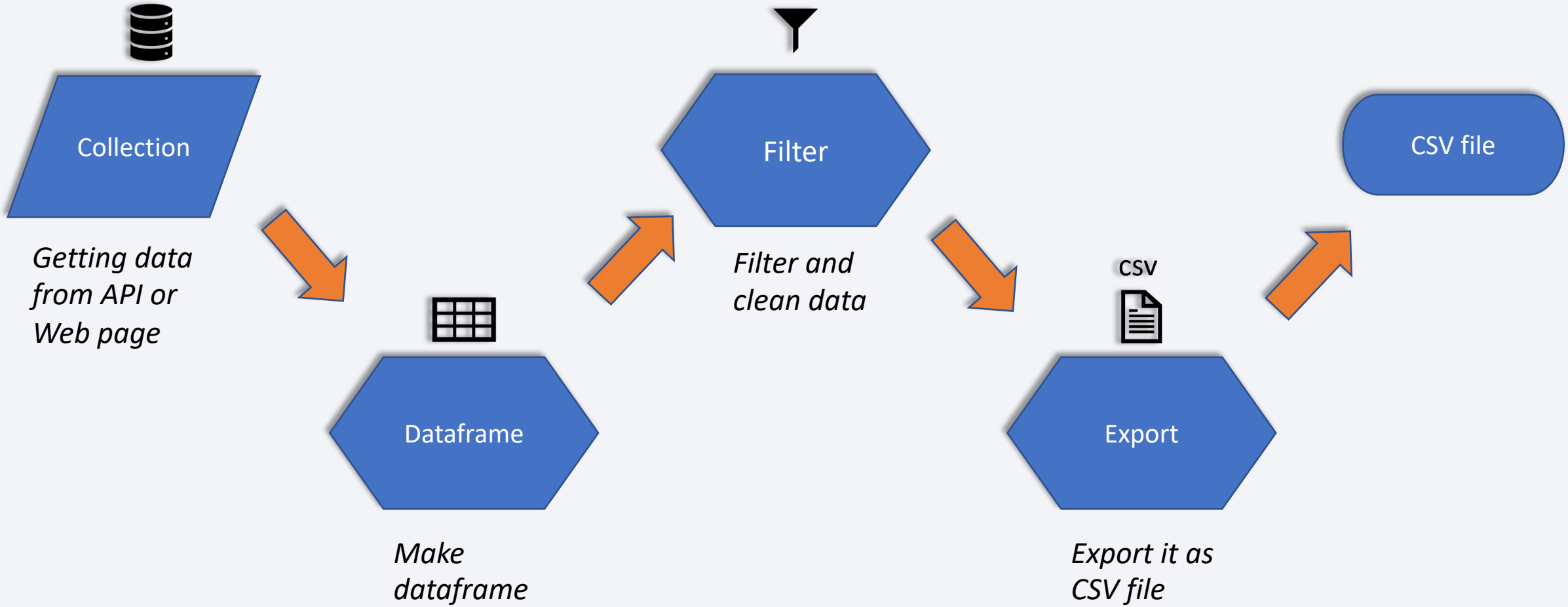
Section 1

Methodology

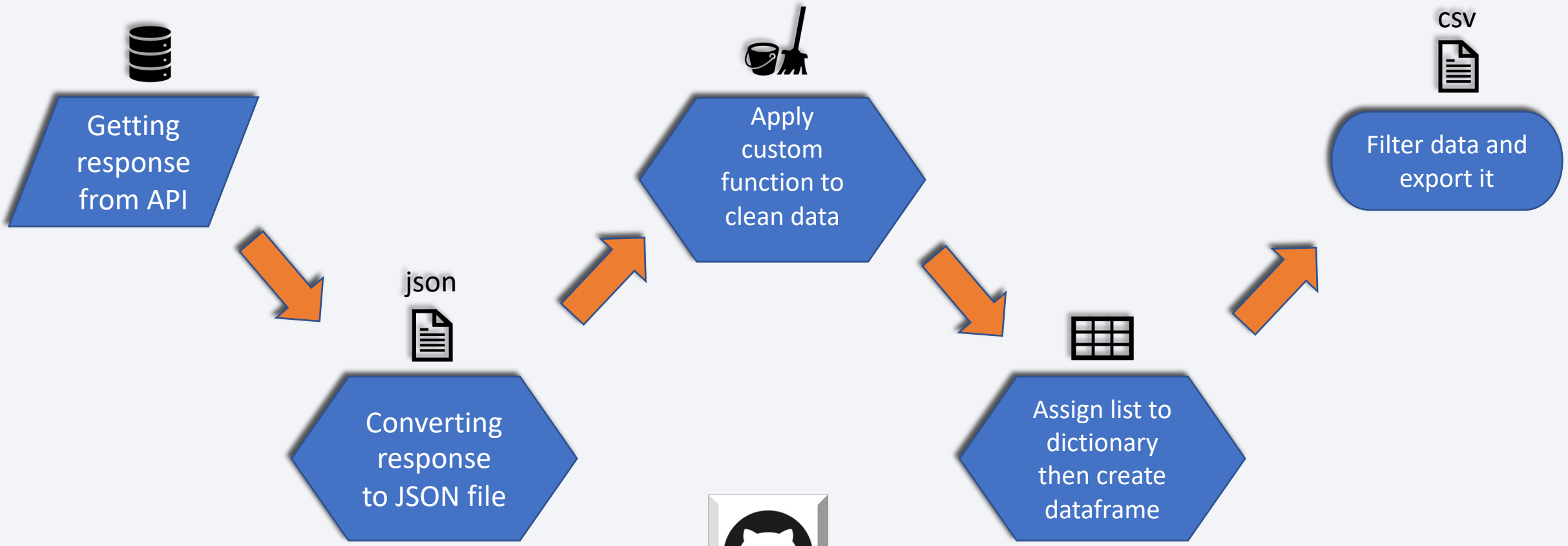
Methodology

- Data collection methodology:
  - Via SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Replacing null values with the mean, one hot encoding data filed for machine learning and dropping irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Scatter and bar graphs to show pattern between data
- Perform interactive visual analytics
  - Using Folium, Plotly and Dash
- Perform predictive analysis using classification models
  - Build and evaluate classification models

Data Collection

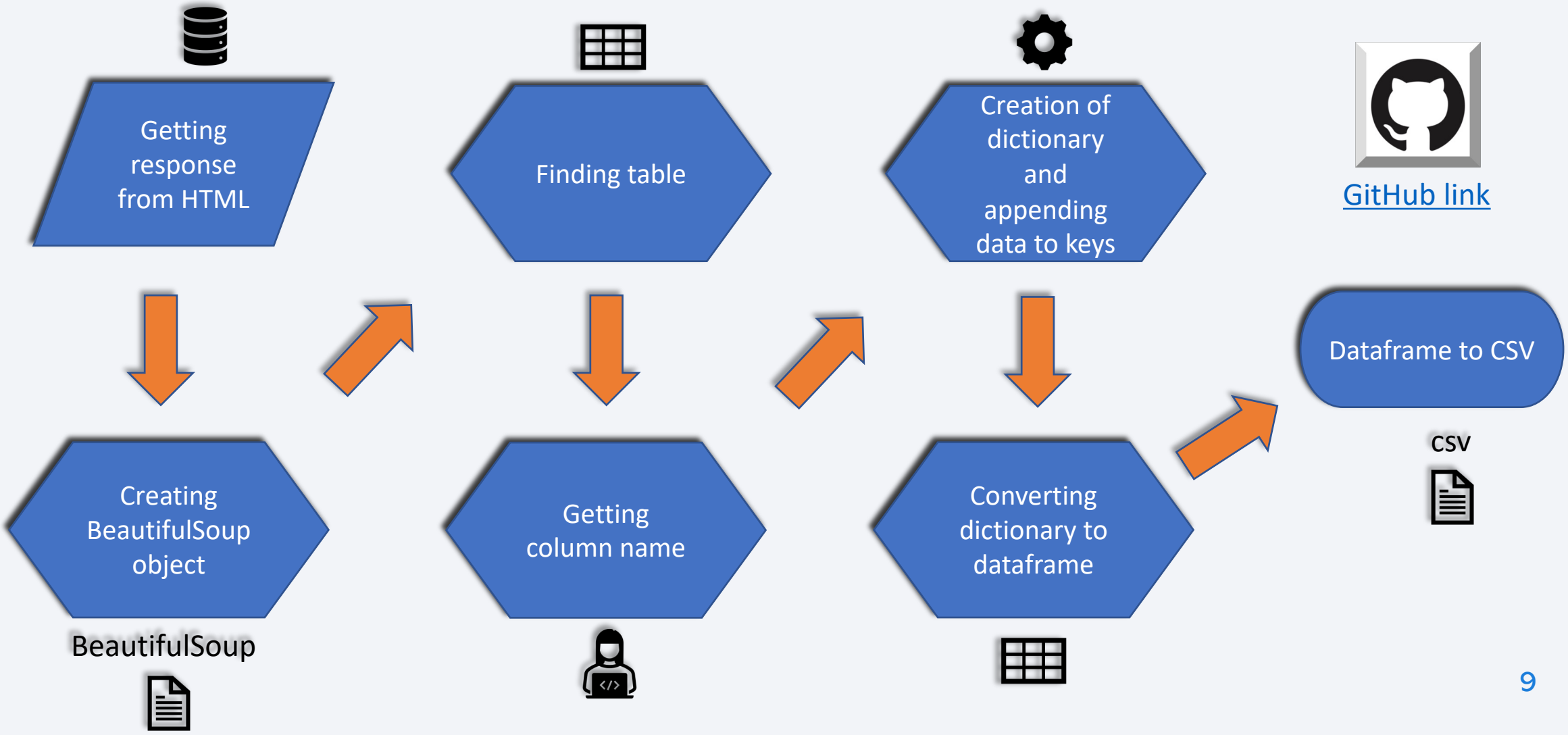


Data Collection – SpaceX API

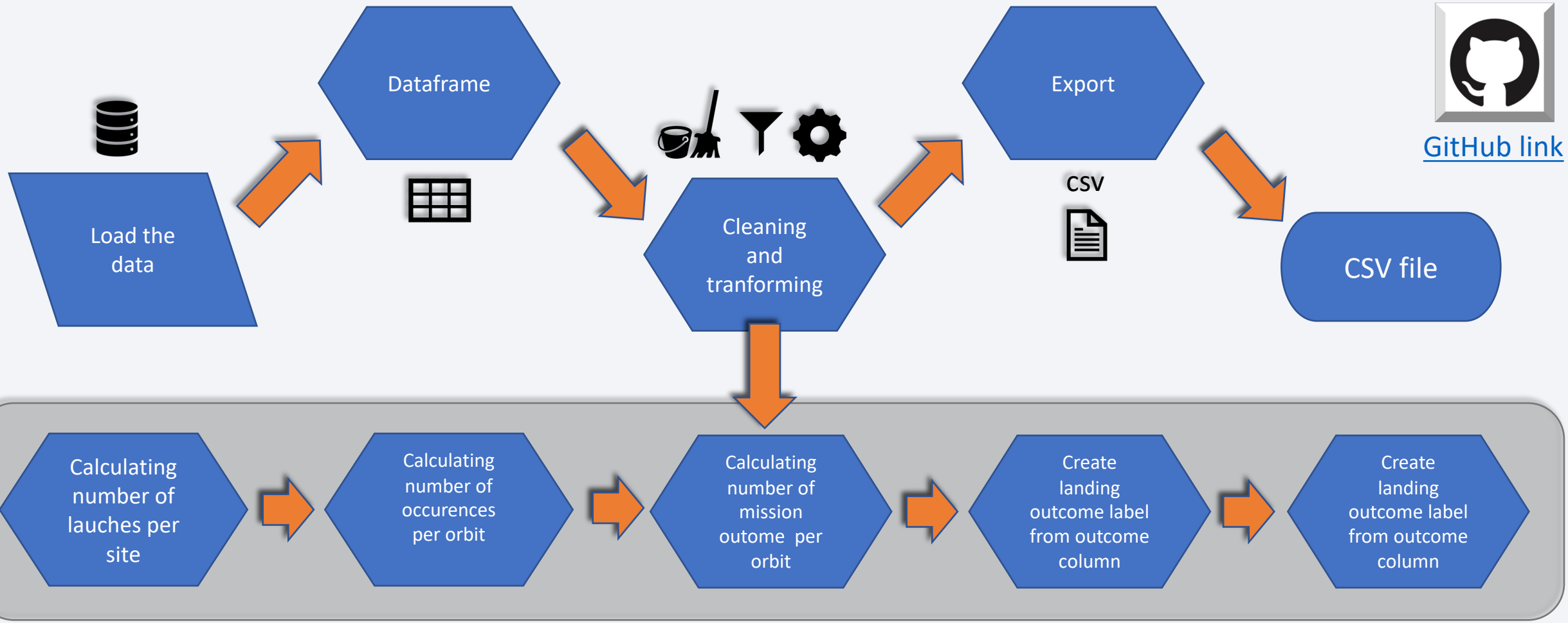


GitHub link

Data Collection - Scraping



Data Wrangling



EDA with Data Visualization – part 1

- Scatter plot

Scatter plots show dependency of attributes on each other. Once a pattem is determined from the grans its very easy to predict which factors will lead to maximum probability of success in both outcome and landing

- Payload and Flight Number
- Flight Number and Launch Site
- Pavload and Launch Site
- Flight Number and Orbit Type
- Payload and Orbit Type

GitHub link

EDA with Data Visualization – part 2

- Bar plot

Bar graphs are easy to interpret a relationship between attributes. Via this bar graph we can easily determine which orbits have the highest probability of success

  - Success Rate VS. Orbit Type

GitHub link

- Line plot

Line graphs are useful in that they show trends clearly and can aid in predictions for the future.

  - Launch Success Yearly Trend

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Display 5 records where lauch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters lauchned by NASA (CRS)
- Displaying average payload mass carried by booster version F9v1.1
- 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 bosster_version which have carried the maximum payload mass
- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for the 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 link

Build an Interactive Map with Folium

- Map Marker
  - Map object to make a mark on map
- Icon Marker
  - Create an icon on map
- Circle Marker
  - Create a circle where Marker is being placed
- PolyLine
  - Create a line between points
- Marker Cluster Object
  - This is a good way to simplify a map containing many markers having the same coordinate

GitHub link

Build a Dashboard with Plotly Dash

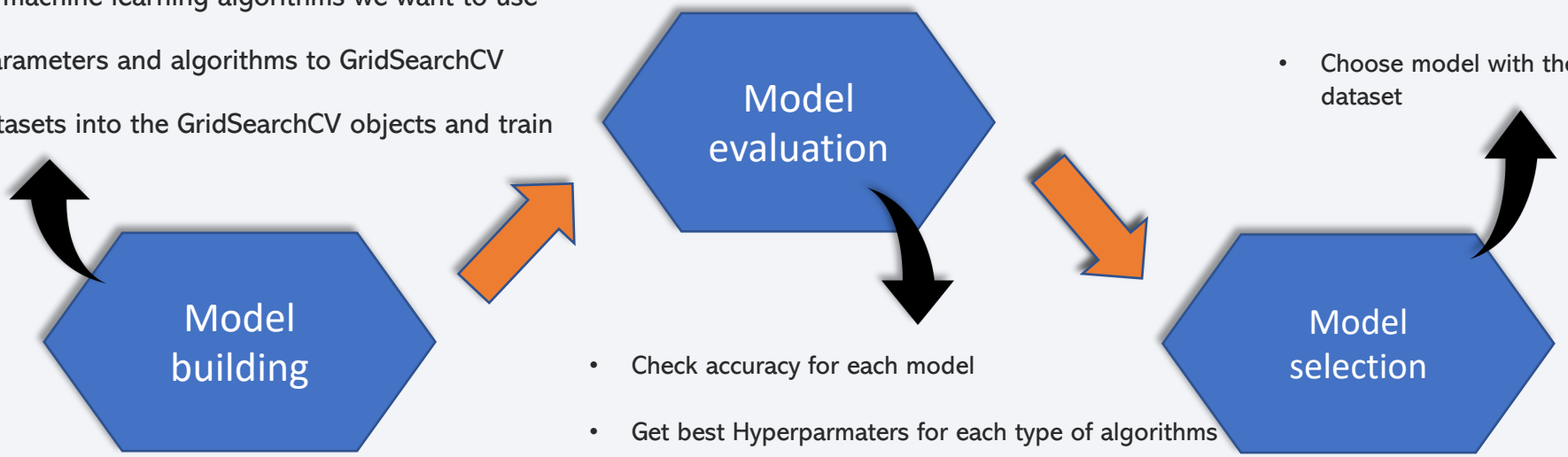
- Pie Chart showing the total success for all sites or by certain launch site
  - Percentage of success in relation to launch site
- Scatter Graph showing the correlation between Payload and Success for all sites or by certain launch site
  - It shows the relationship between Success rate and Booster Version Category

GitHub link

Predictive Analysis (Classification)

- Load our features engineered data into dataframe
- Transform it into Numpy array
- Standardize and transtorm data
- split data into training and test data sets
- Check how many test samples has been created
- List down machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our model

GitHub link

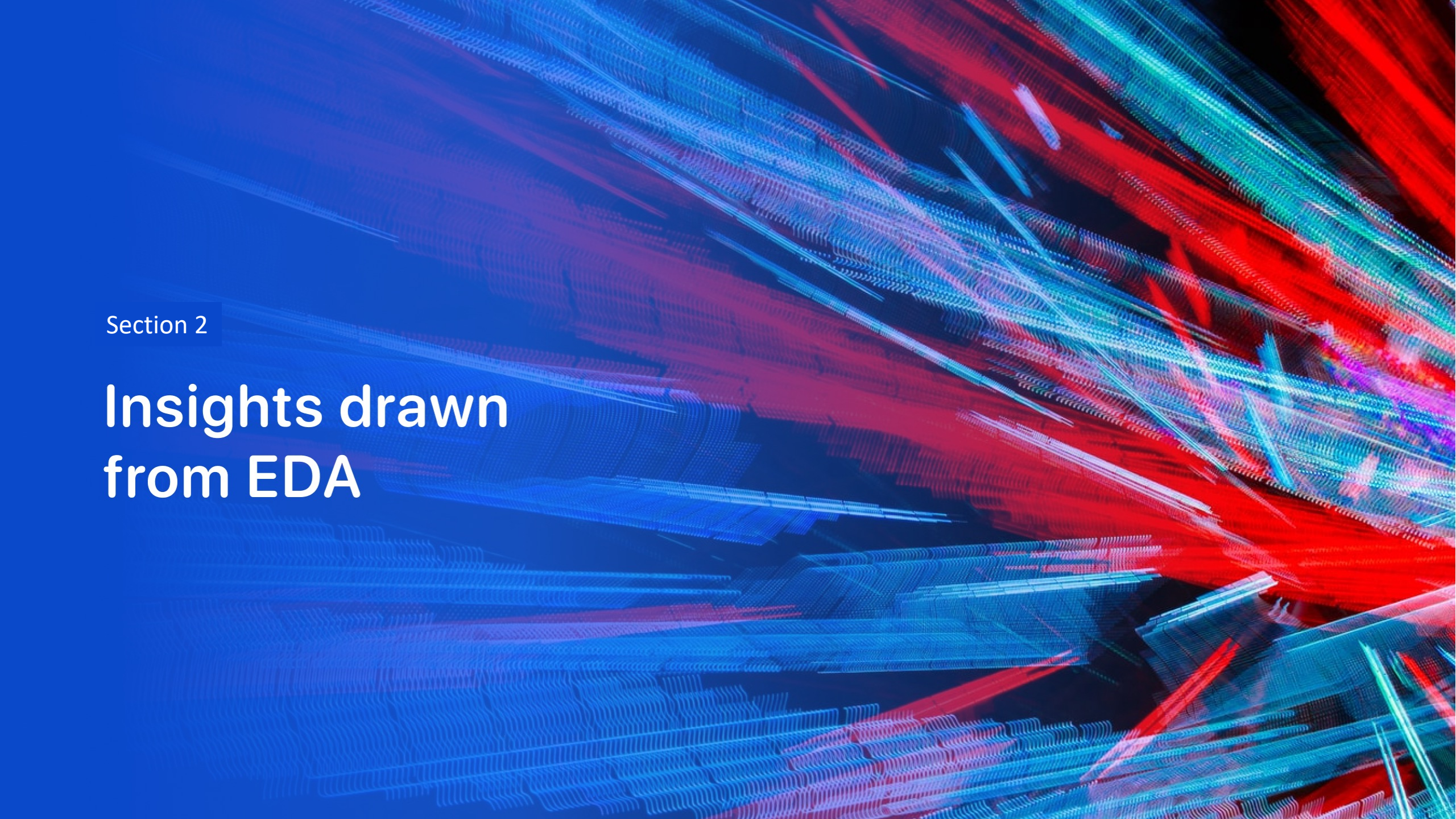


- Check accuracy for each model
- Get best Hyperparmaters for each type of algorithms
- Plot Confusion Matrix

- Choose model with the highest performance in test dataset

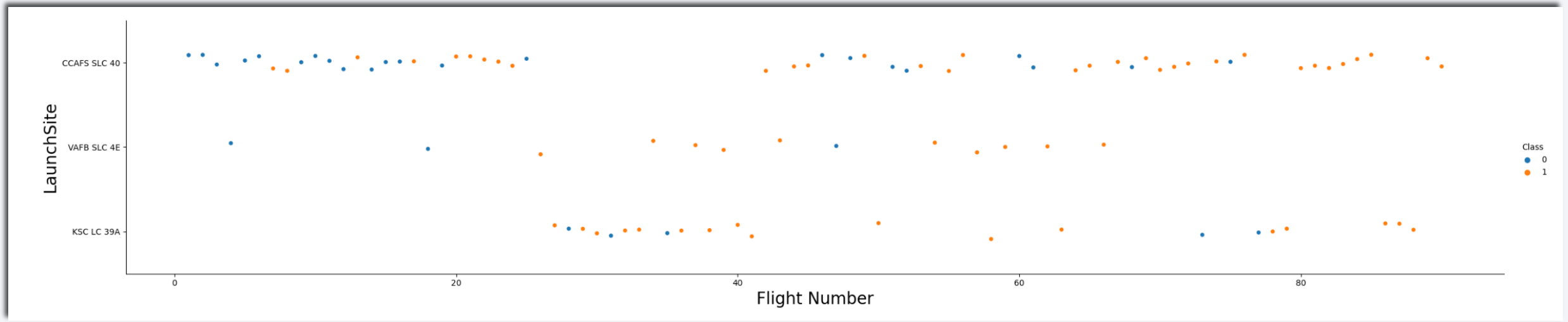
Results

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


Section 2
Insights drawn from EDA

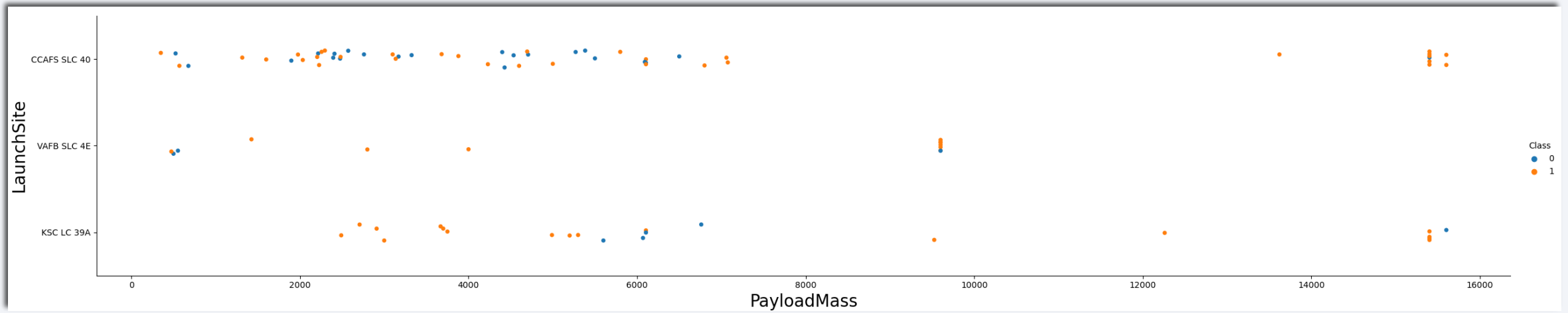
Flight Number vs. Launch Site

- With higher flight numbers (greater than 30) the success rate for the Rocket is increasing.



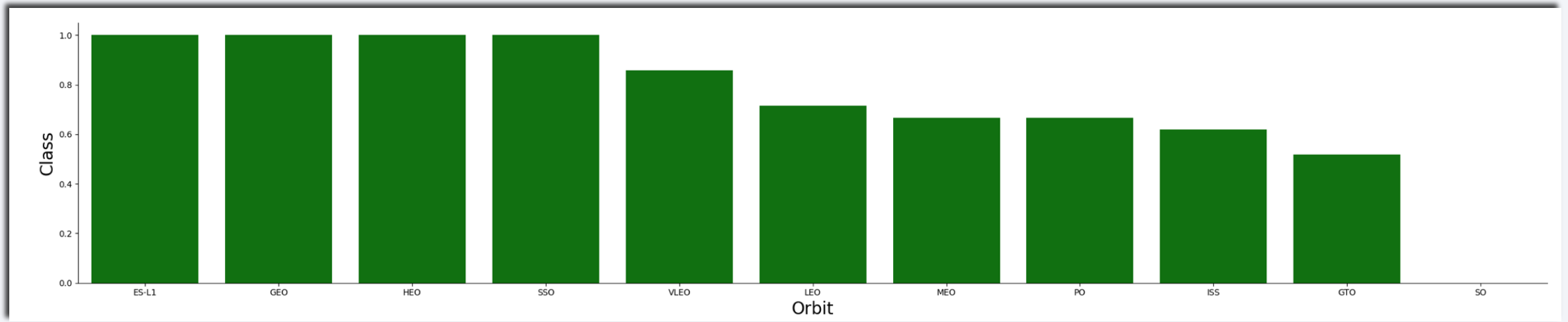
Payload vs. Launch Site

- The greater the payload mass (greater than 7000 Kg) higher the success rate for the Rocket. But there's no clear pattern to take a decision, if the launch site is dependent on Pay Load Mass for a success launch.



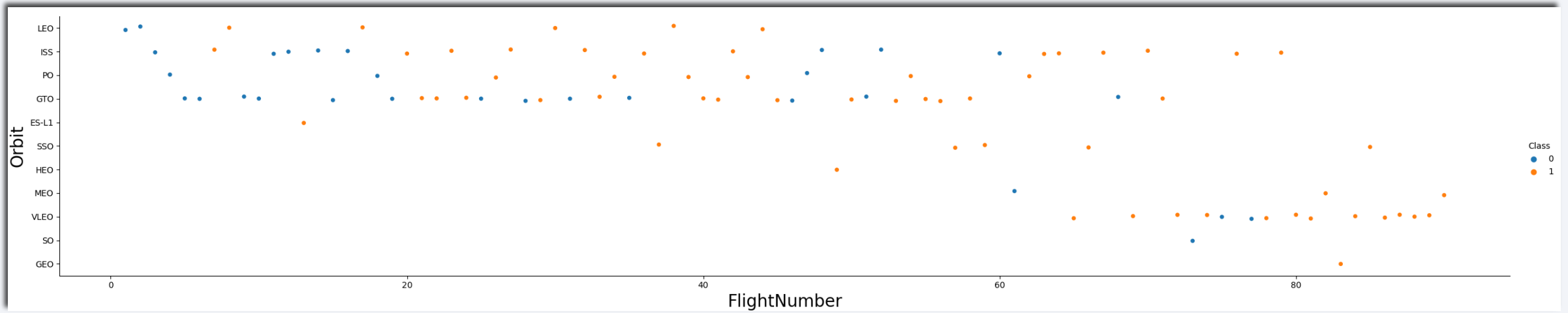
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO has highest Sucess rates.



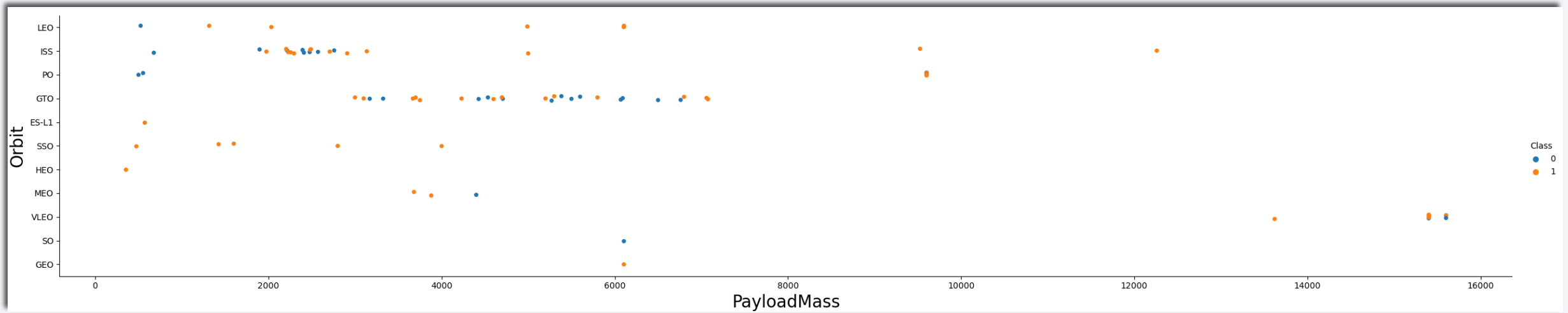
Flight Number vs. Orbit Type

- We see that for LEO orbit the success increases with the number of flights
- On the other hand, there seems to be no relationship between flight number and the GTO orbit.



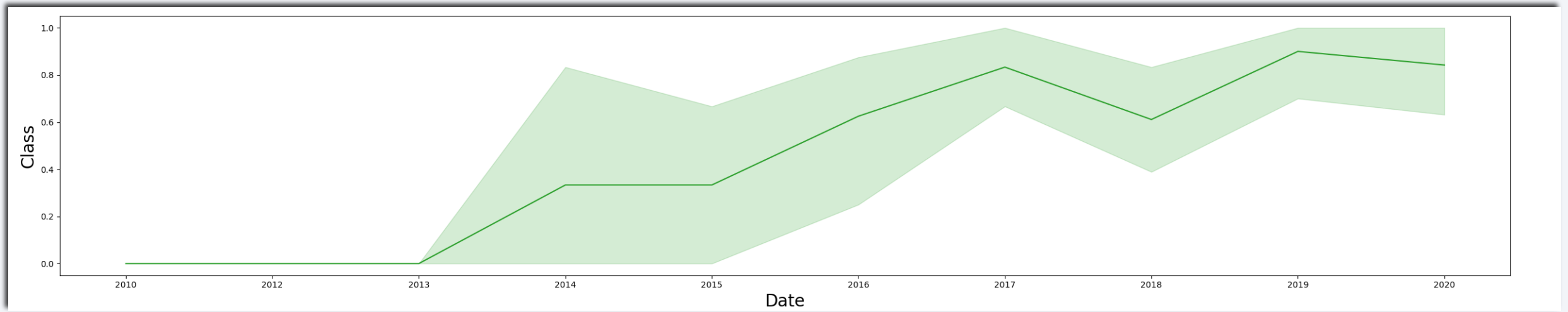
Payload vs. Orbit Type

- We observe that heavy payloads have a negative influence on MEO, GTO, VLEO orbits
- Positive on LEO, ISS orbits



Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing relatively though there is slight dip after 2019.



All Launch Site Names

- Using DISTINCT in the query allow us to retrieve unique launch sites

```
%%sql

select DISTINCT(Launch_Site) from SPACEXTBL
```

```
 * 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'

- Using keyword 'LIMIT 5' in the query we fetch 5 records from table space and with condition LIKE keyword with wild card - 'CCA%' . The percentage in the end suggests that the Launch_Site name must start with CCA.

```
%%sql

select * from SPACEXTBL 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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Using the function SUM calculates the total in the column PAYLOAD_MASS_KG_ fetch Customer's name containing ”(CRS)".

```
%%sql

select SUM(PAYLOAD_MASS__KG_) AS "Total payload mass" from SPACEXTBL WHERE Customer LIKE "%NASA (CRS)%"
```

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

Total payload mass
48213

Average Payload Mass by F9 v1.1

- Using the function AVG works out the average ni the column PAYLOAD_MASS_KG_
- The WHERE clause filters the dataset to only perform calculations on Booster_version "F9 v1.1'

```
%%sql

select AVG(PAYLOAD_MASS__KG_) as "Average payload mass" from SPACEXTBL WHERE Booster_Version LIKE "%F9 v1.1%"
```

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

Average payload mass
2534.6666666666665

First Successful Ground Landing Date

- Using the function MIN works out the minimum date in the column Date and WHERE clause filters the data to only perform calculations on LandingOutcome with values "Success (ground pad)"

```
%%sql

select MIN(Date) AS "First successful landing" from SPACEXTBL WHERE [Landing _Outcome] = 'Success (ground pad)'
```

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

First successful landing
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%%sql

SELECT Booster_Version
FROM SPACEXTBL
WHERE [Landing _Outcome] = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000)
```

```
 * 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

- Not complete query but you can see that successful mission outcome sum up to 100 and failed mission outcome to 1.

```
%%sql
SELECT Outcome as Mission_Outcome, SUM(Quantity)  as Quantity
FROM (SELECT Mission_Outcome, CASE WHEN Mission_Outcome LIKE "%Success%" THEN 'Success'  ELSE 'Failure' END AS Outcome, COUNt(*) AS Quant:
FROM SPACEXTBL
GROUP BY Mission_Outcome)
GROUP BY Outcome
```

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

Mission_Outcome	Quantity
Failure	1
Success	100

Boosters Carried Maximum Payload

- Using the function MAX works out the maximum payload ni the column PAYLOAD_MASS__KG_ in sub query
- WHERE clause filters Booster Version which had that maximum pavload

```
%%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

- We need to 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
- Via year function we extract the year and future where clause 'Failure (drone ship)' fetches our required values

```
%%sql

SELECT substr(Date, 4, 2) AS Month, Booster_Version, Launch_Site, [Landing _Outcome]
FROM SPACEXTBL
WHERE [Landing _Outcome] = "Failure (drone ship)" and substr(Date,7,4)='2015'
```

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

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

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

- Selecting only LANDINGOUTCOME,
- WHERE clause filters the data with DATE BETWEEN '2010-06-04' AND '2017-03-20' and LANDING_OUTCOME contains “Success”
- Grouping by LANDING_OUTCOME
- Order by COUNT(*) in Descending Order.

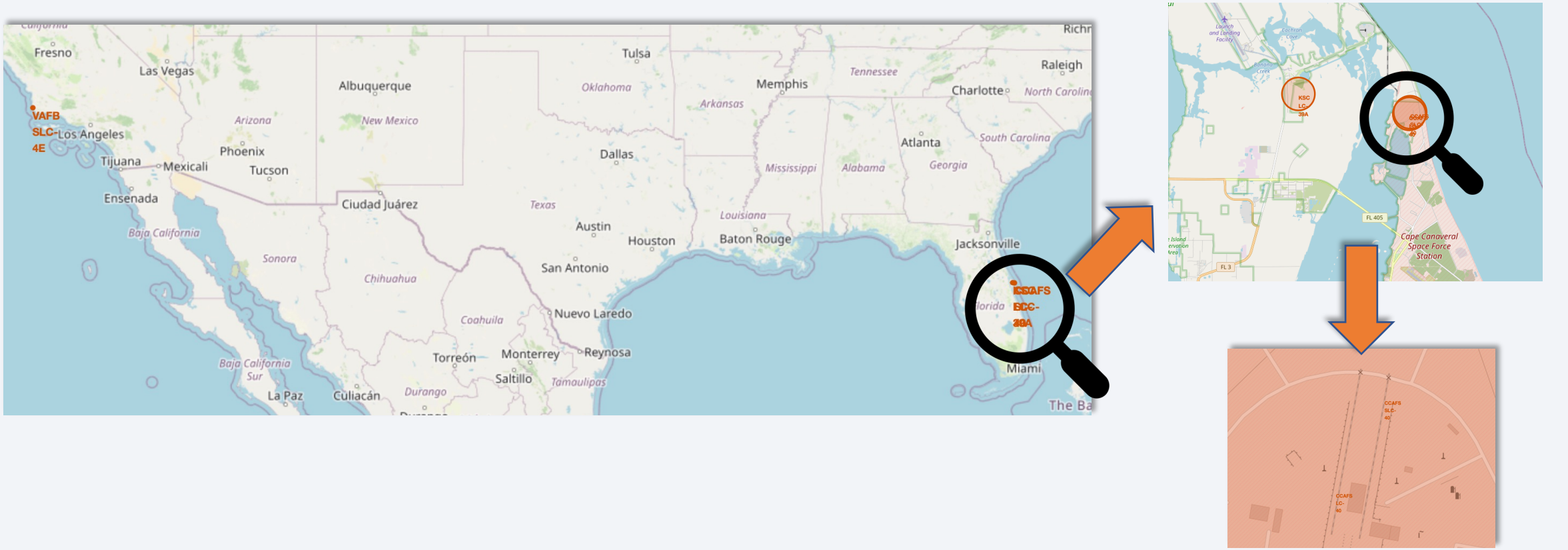
```
%%sql

SELECT [Landing _Outcome], COUNT(*) as Quantity
FROM SPACEXTBL
WHERE (Date > '04-06-2010' AND Date < '20-03-2017') AND [Landing _Outcome] LIKE '%Success%'
GROUP BY [Landing _Outcome]
ORDER BY Quantity DESC
```

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

Landing _Outcome	Quantity
Success	20
Success (drone ship)	8
Success (ground pad)	6

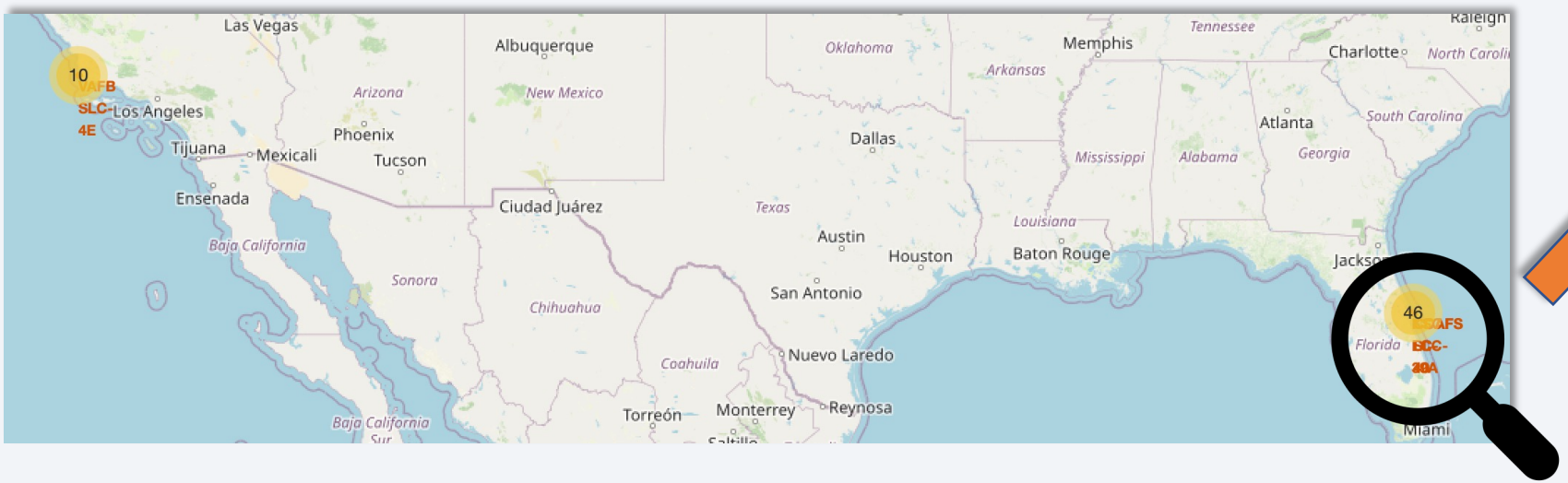

Section 3
Launch Sites
Proximities Analysis

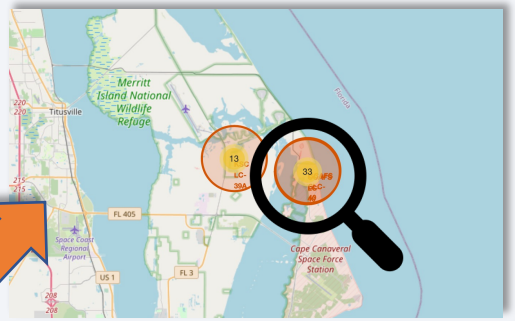
AI Launch Sites on Folium Map

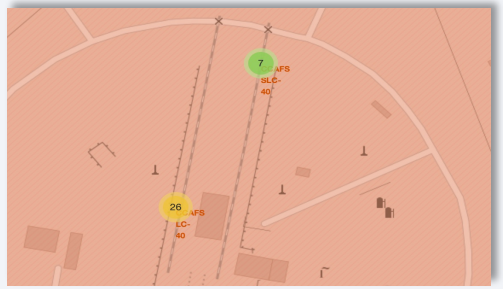


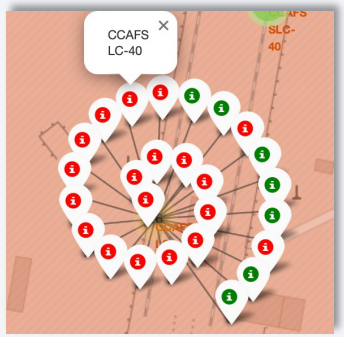
- We can see that the Space launch sites are near to the United States of America coasts i.e., Florida and California Regions.

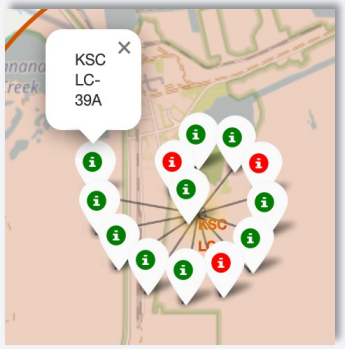
Color Labeled Launch Records

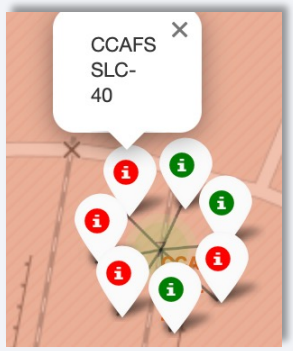






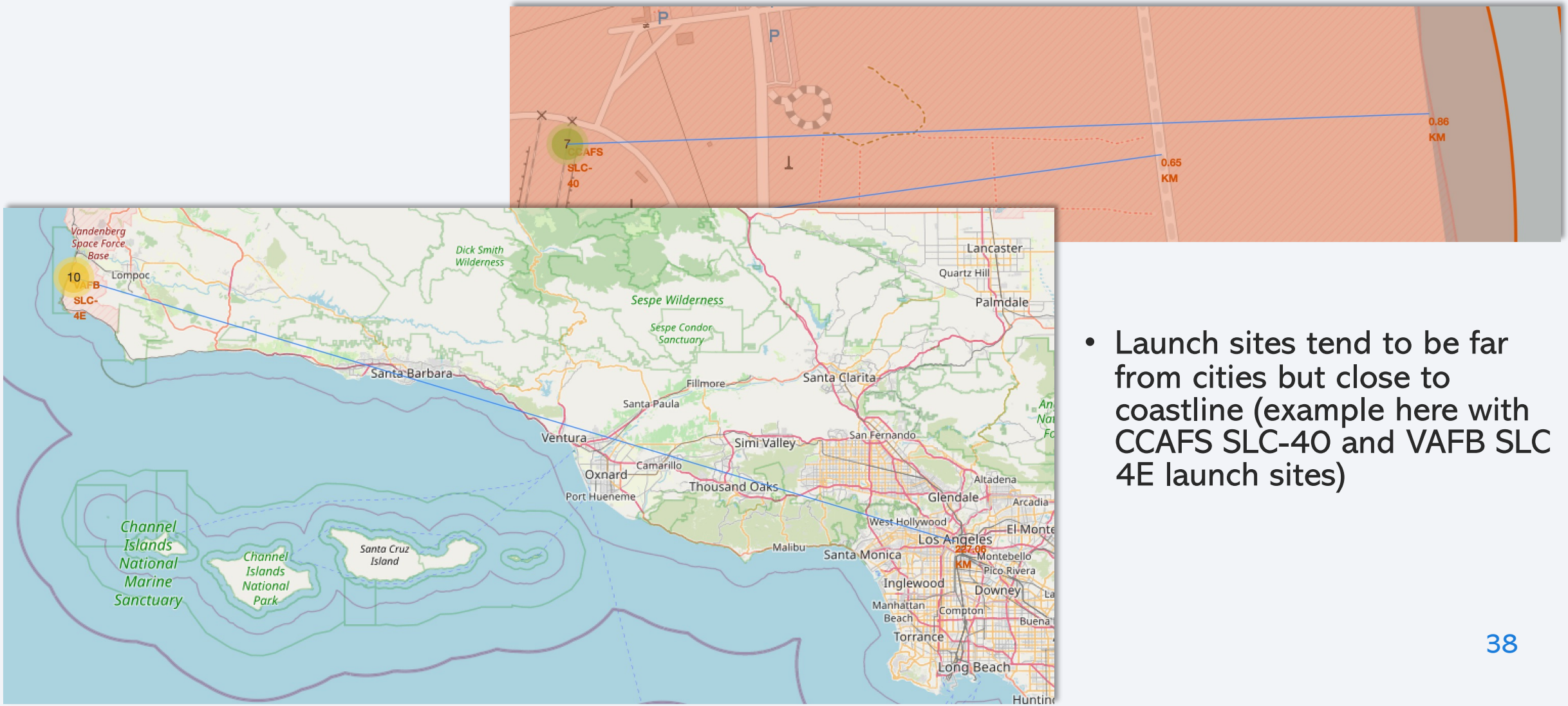






- From these screenshots its easily understandable that KSC LC- 39A has the maximum probability of success.

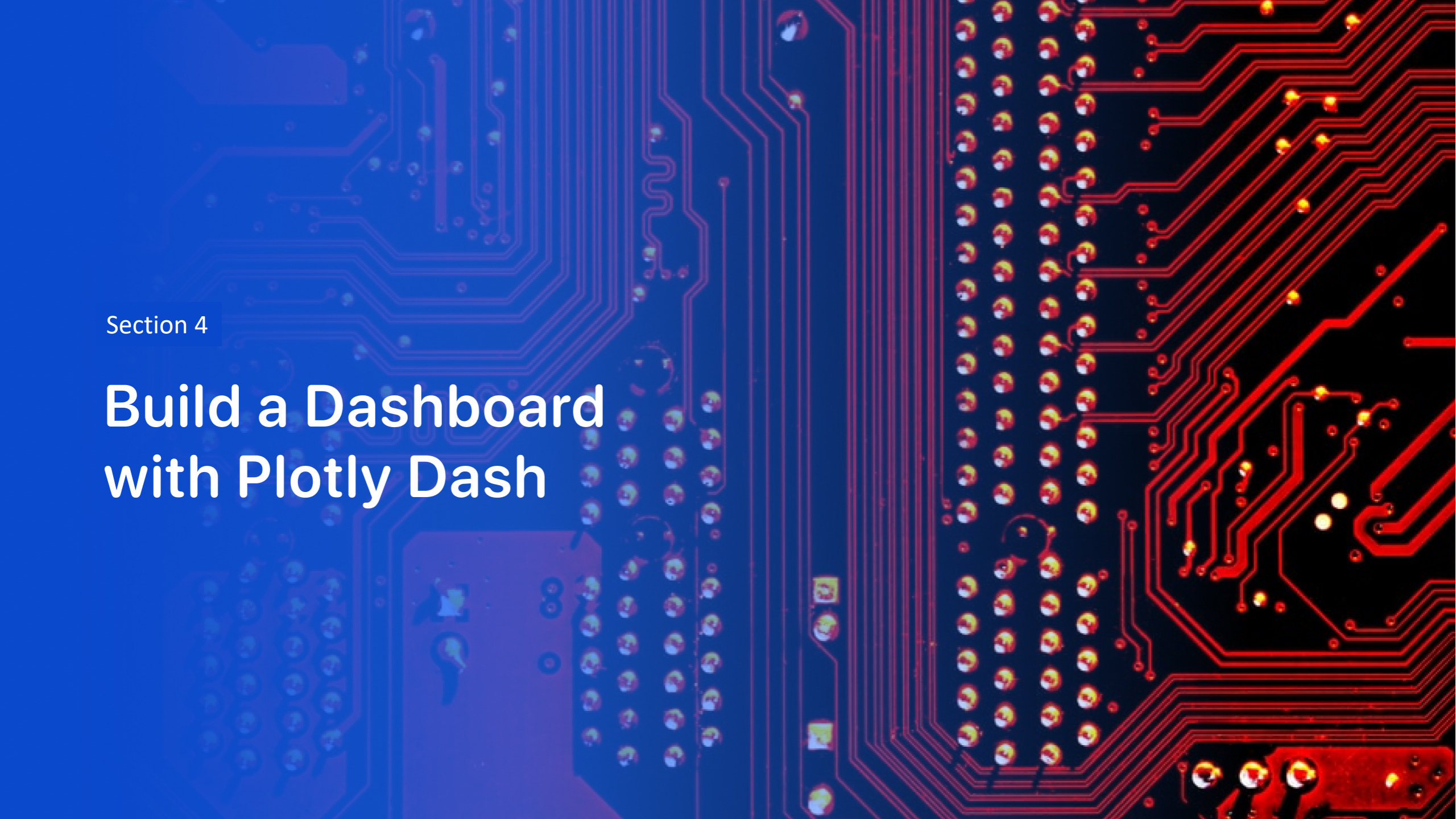
Launch Site Distances from cities and Coastlines



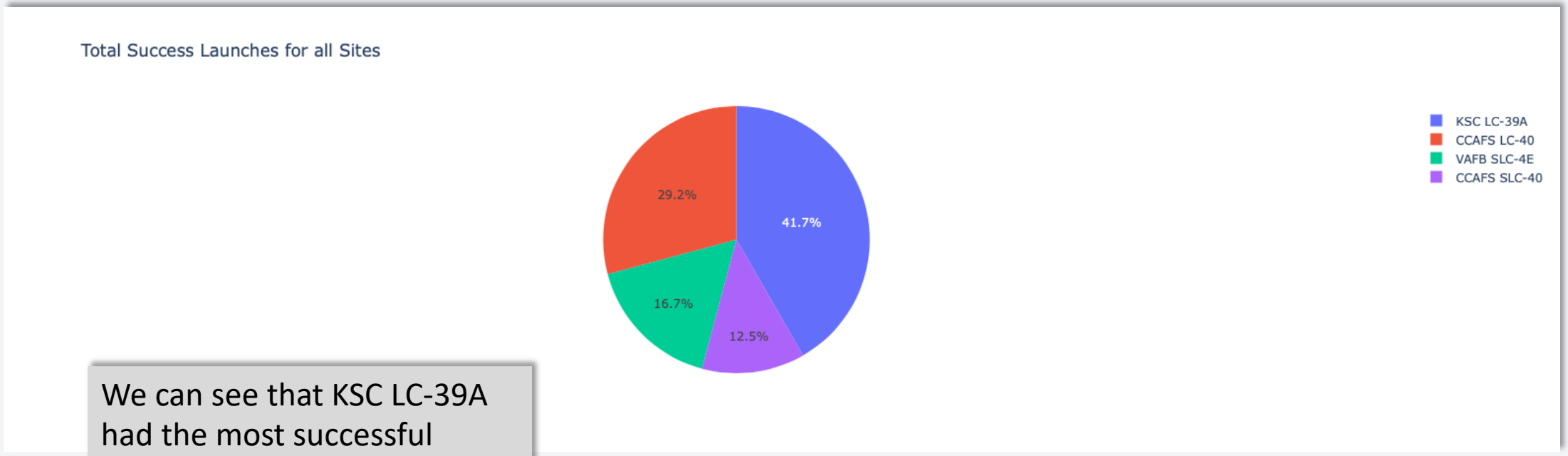
- Launch sites tend to be far from cities but close to coastline (example here with CCAFS SLC-40 and VAFB SLC 4E launch sites)

Launch Site Distances from highway and railway

- Launch sites tend to be close to highway and railway (example here with CCAFS LC-40 and KSC LC 39A launch sites)

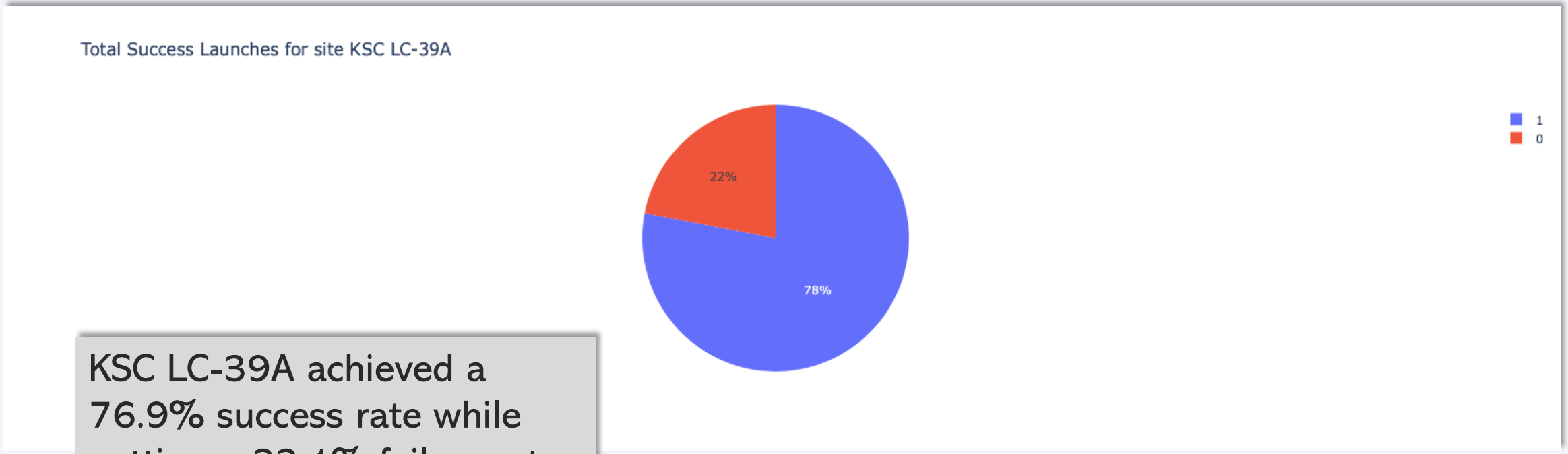

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



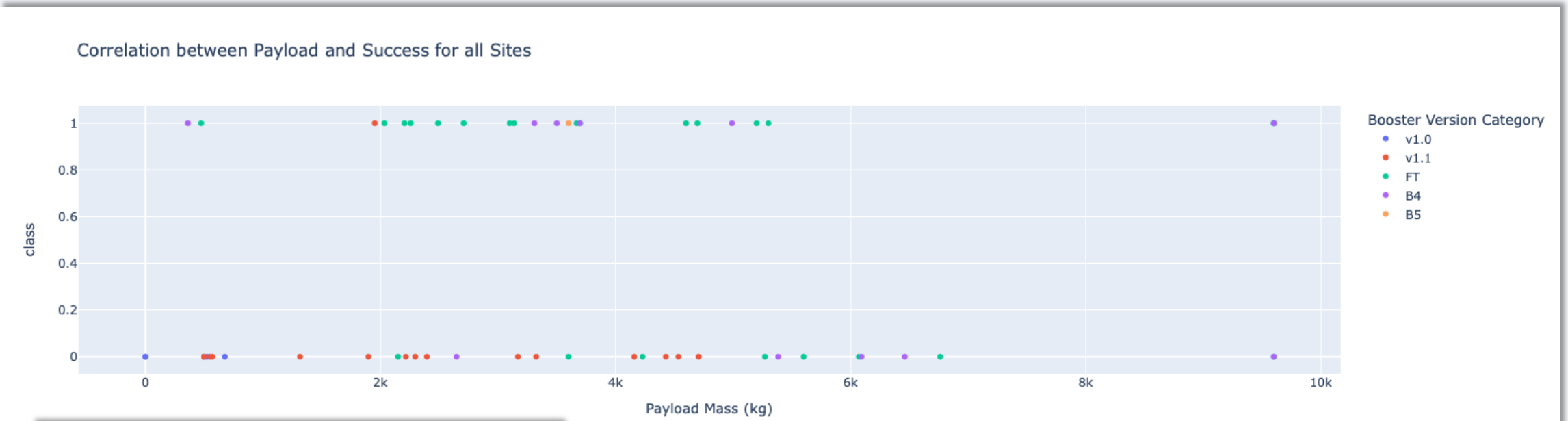
We can see that KSC LC-39A had the most successful launches from all the sites.

Launch Site with Highest Launch Success Ratio



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

Payload vs. Launch Outcome Scatter Plot for All Sites



- We can see the success rates for low weighted payloads (below 4000 kg) is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

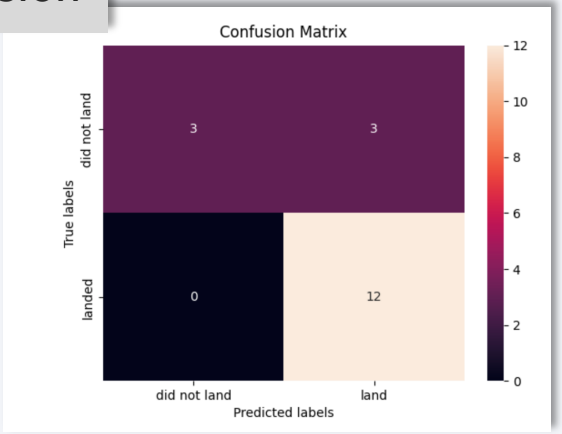
Classification Accuracy

- Our cross-validation method suggests us to choose Decision tree algorithm with an accuracy score around 88%

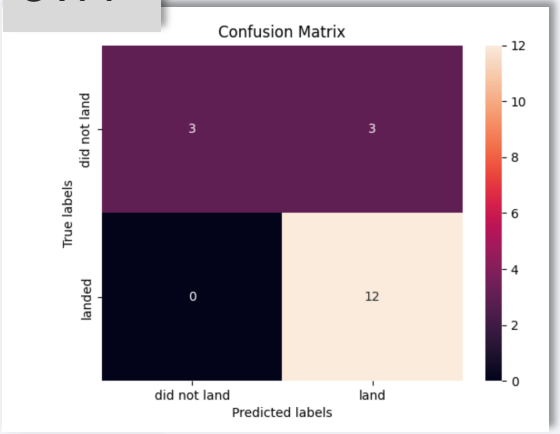
Confusion Matrix

- The confusion matrix tells us that the model need to be improved for predicting negative outcome. Indeed, only half of the negative outcome have been correctly assigned.
- Furthermore, we can see here that, on the test set, our tree model is not that good (overfitting possibility)

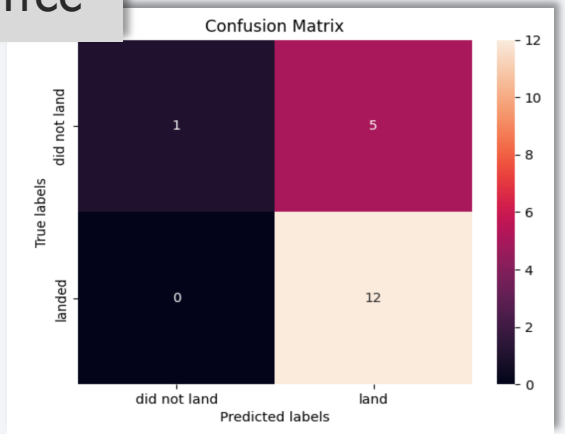
Logistic regression



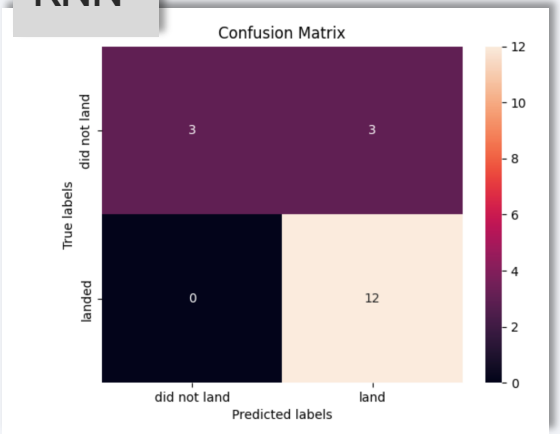
SVM



Tree



KNN



Conclusions

- Orbits ES-L1, GEO, HEO, SO has highest Sucsess rates
- Success rates for Space launches has been increasing relatively with time and it looks like soon they will reach the required target
- SCLC-39A had the most successful launches but increasing payload mass seems to have negative impact on success
- Decision Tree Classifier Algorithm is the best for Machine Learning Model for provided dataset regarding Cross-Validation error but is not performing well on the test set

Thank you!