
IBM Developer
SKILLS NETWORK
DIVING INTO SPACE WORLD
WITH DATA SCIENCE
Aziz Sayadi
9/16/2023

OUTLINE

EXECUTIVE SUMMARY

Methodologies

- Data Collection and Data Wrangling
- EDA and Interactive Visual Analytics
- Predictive Analysis

Results

- EDA with visualization
- EDA with SQL
- Interactive Map with Folium
- Plotly Dash dashboard
- Predictive Analysis

INTRODUCTION

The SpaceX Falcon 9 rocket has transformed the modern space industry since its first launch in 2010.
In this data science project, we will analyze SpaceX's publicly available Falcon 9 mission data and gain insights.

Main Goal:

- Predicting The Successful landing of Falcon 9 First Stage

Key Problems:

- What factors (orbit, payload mass, etc..) most influence whether a mission achieves a landed/recovered first stage?
- What are the possible dangers in the case of failed landing?

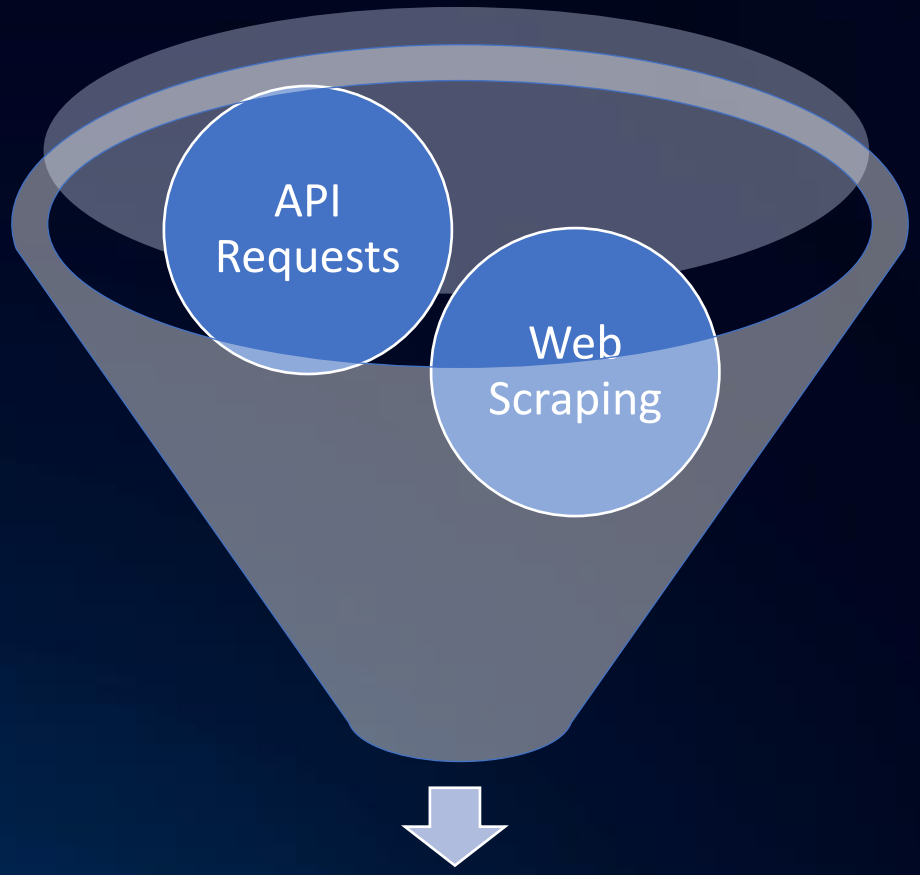
SECTION 1
METHODOLOGY

METHODOLOGY

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

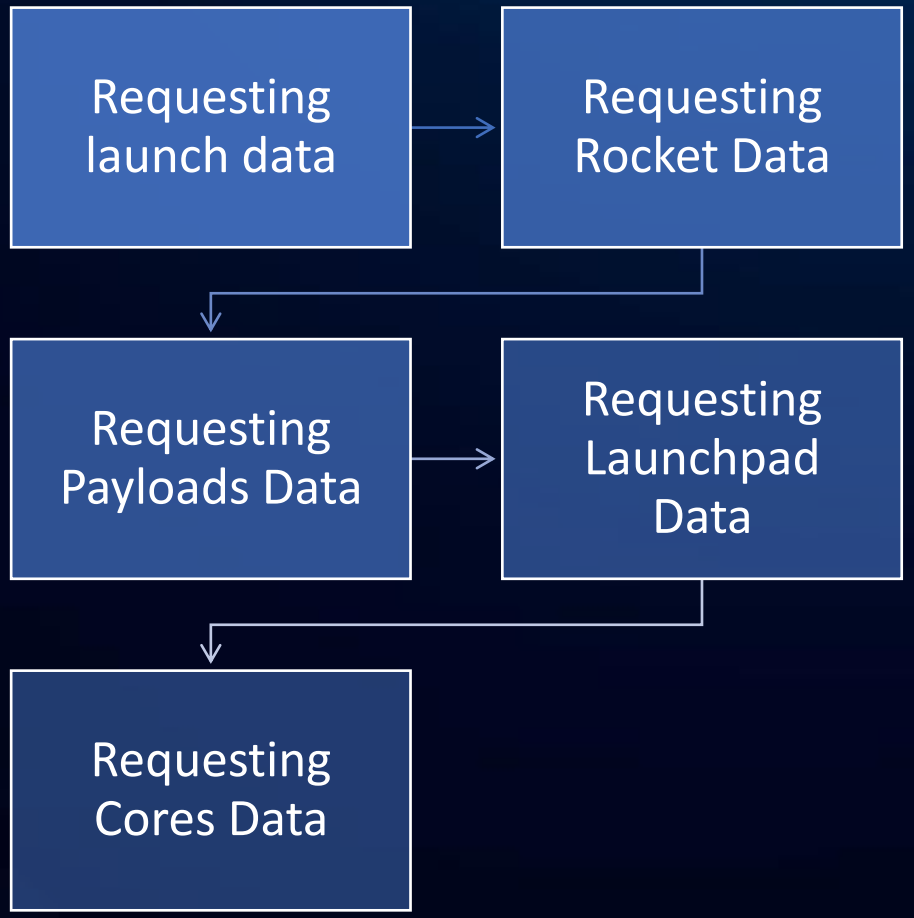


- Using Get Requests:
  - Collecting Data From SpaceX API
- Using Web Scraping with Beautiful Soup:
  - Collecting Data From Wikipedia Falcon 9 launches records

DATA COLLECTION – SPACEX API

After Requesting Data from SpaceX API, We've noticed that a lot of the data are IDs.
So, we used the API again to get informations about rocket, payloads, launchpad and cores.

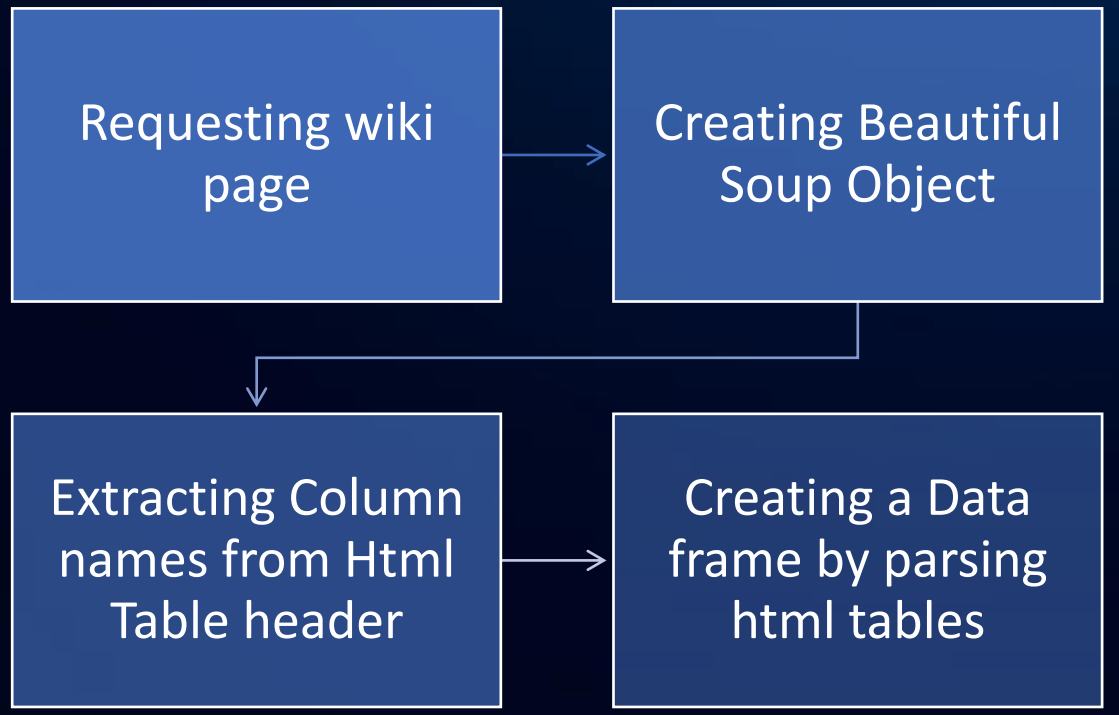
https://github.com/azizsayadi1155/applied-datascience-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



DATA COLLECTION – SCRAPING

After getting the html page content, we've collected the relevant column names from the HTML table header, Then we've created a dictionary to save the headers and to append each element to the appropriate list, after that we've created the data frame.

https://github.com/azizsayadi1155/applied-datascience-capstone/blob/main/jupyter-labs-webscraping.ipynb

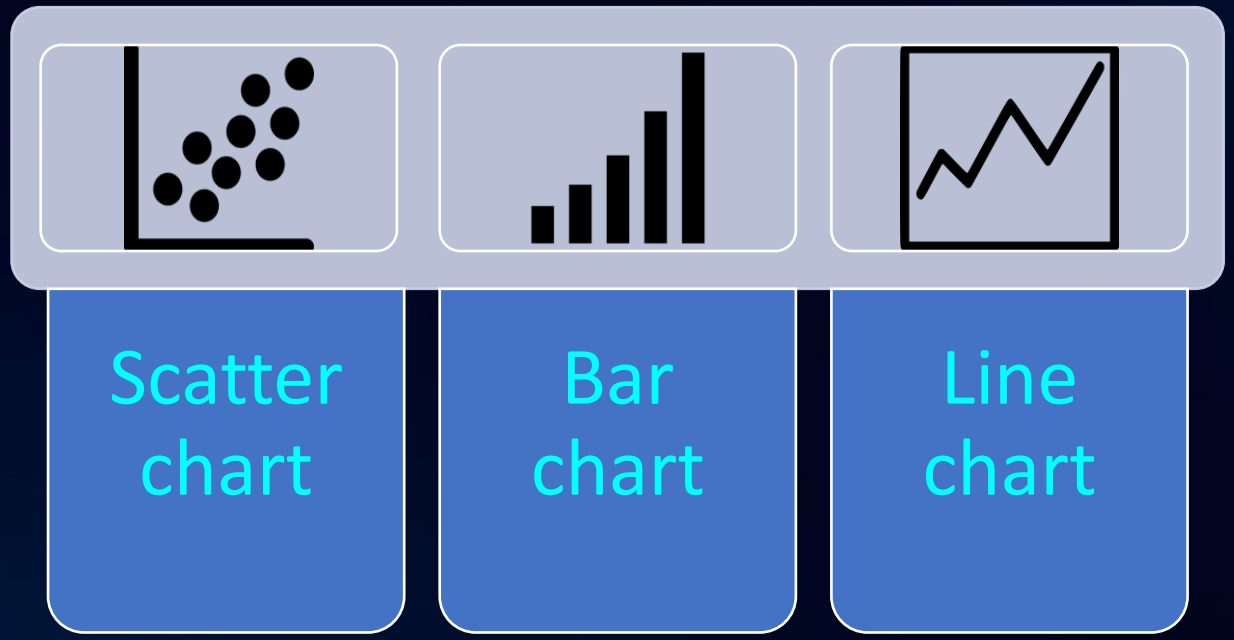


DATA WRANGLING

Missing Values	Launches	Orbits	Outcomes	Outcome Label
1	2	3	4	5
Replace Missing values with the mean	Determine the number of launches on each site	Determine the number and the occurrence of each orbit	Determine the number of mission outcomes per orbit type	Create a landing outcome label

https://github.com/azizsayadi1155/applied-datascience-capstone/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization



https://github.com/azizsayadi1155/applied-datascience-capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

- Scatter chart:
  - Revealing relationship between features (e.g., FlightNumber and Orbit)
- Bar chart:
  - Finding which orbit have high success rate
- Line chart:
  - Visualizing Success rate trend over years

EDA with SQL

- Display the names of unique launch sites
- Display 5 records where launch sites begin with ‘CCA’
- Display the total payload mass carried by NASA(CRS) Boosters
- Display average payload mass by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- 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
- Rank 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

https://github.com/azizsayadi1155/applied-datascience-capstone/blob/main/jupyter-labs-eda-sql.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

Markers

- Mark Launch Sites Locations

Circles

- Identify which launch sites have high success rate

Lines

- Determine the distance between a launch site and its proximities

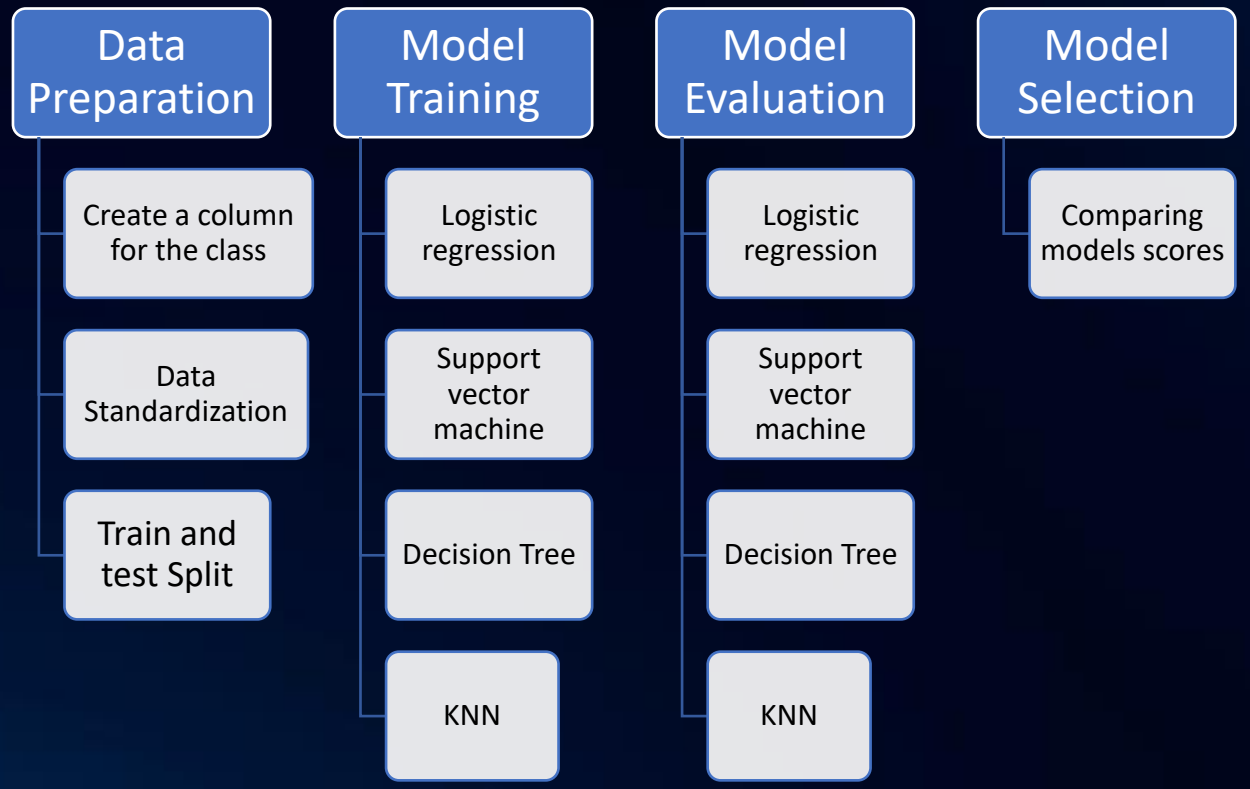
https://github.com/azizsayadi1155/applied-datascience-capstone/blob/main/jupyter_launch_site_location_folium.jupyterlite.ipynb

BUILD A DASHBOARD WITH PLOTLY DASH

Pie Chart	Scatter Chart	Payload Slider	Site dropdown
• Show Total Success launches by site • Show Total Success launches for a selected site	• Show Correlation between payload and class for all sites • Show Correlation between payload and class for a selected site	• Precise the range of payload mass	• Easily selecting a site

https://github.com/azizsayadi1155/applied-datascience-capstone/blob/main/spacex_dash_app.py

PREDICTIVE ANALYSIS (CLASSIFICATION)



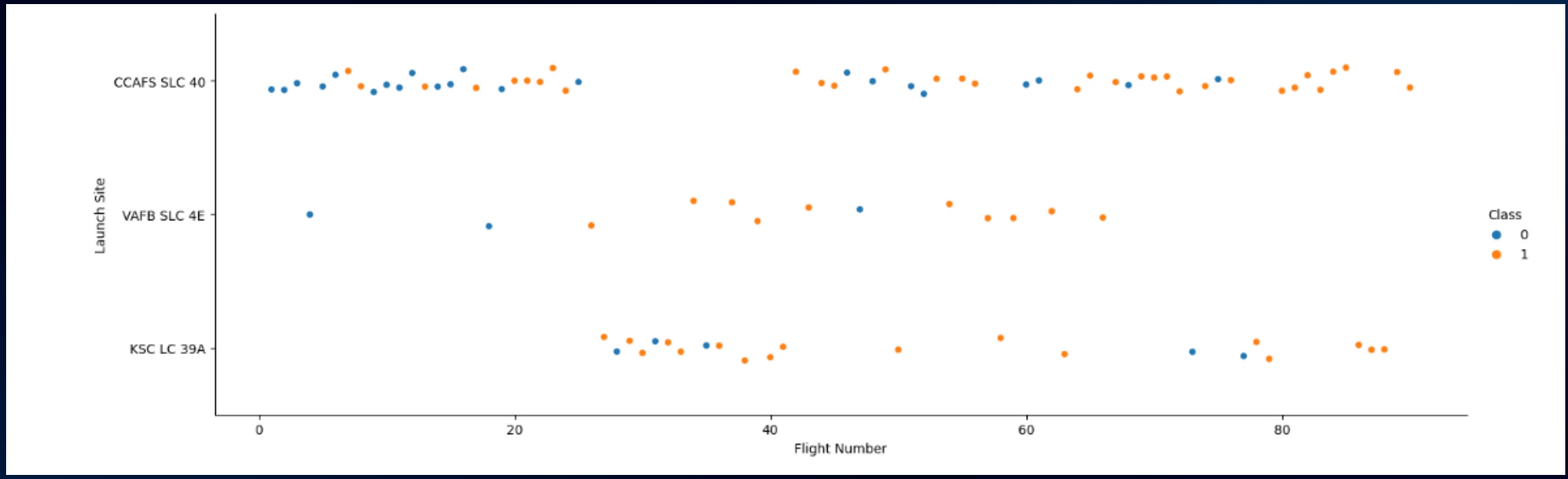
- Data standardization transforms data into a standard format and easy to understand for models
- We used GridSearchCV to find the optimal parameter values for a given model

https://github.com/azizsayadi1155/applied-datascience-capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

SECTION 2

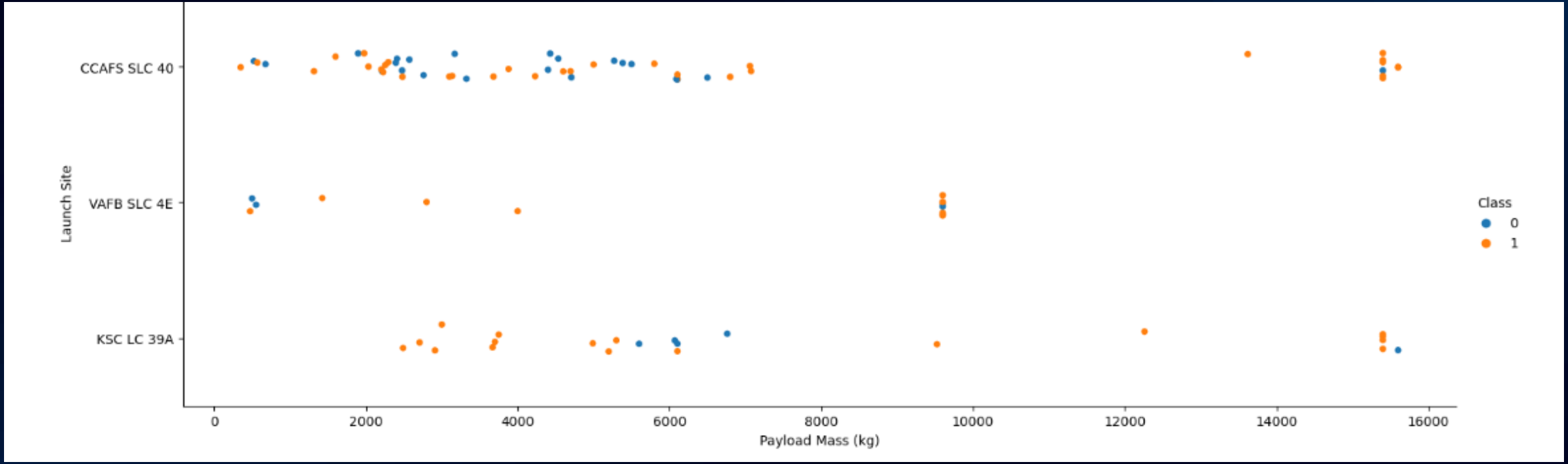
INSIGHTS DRAWN FROM EDA

FLIGHT NUMBER VS. LAUNCH SITE



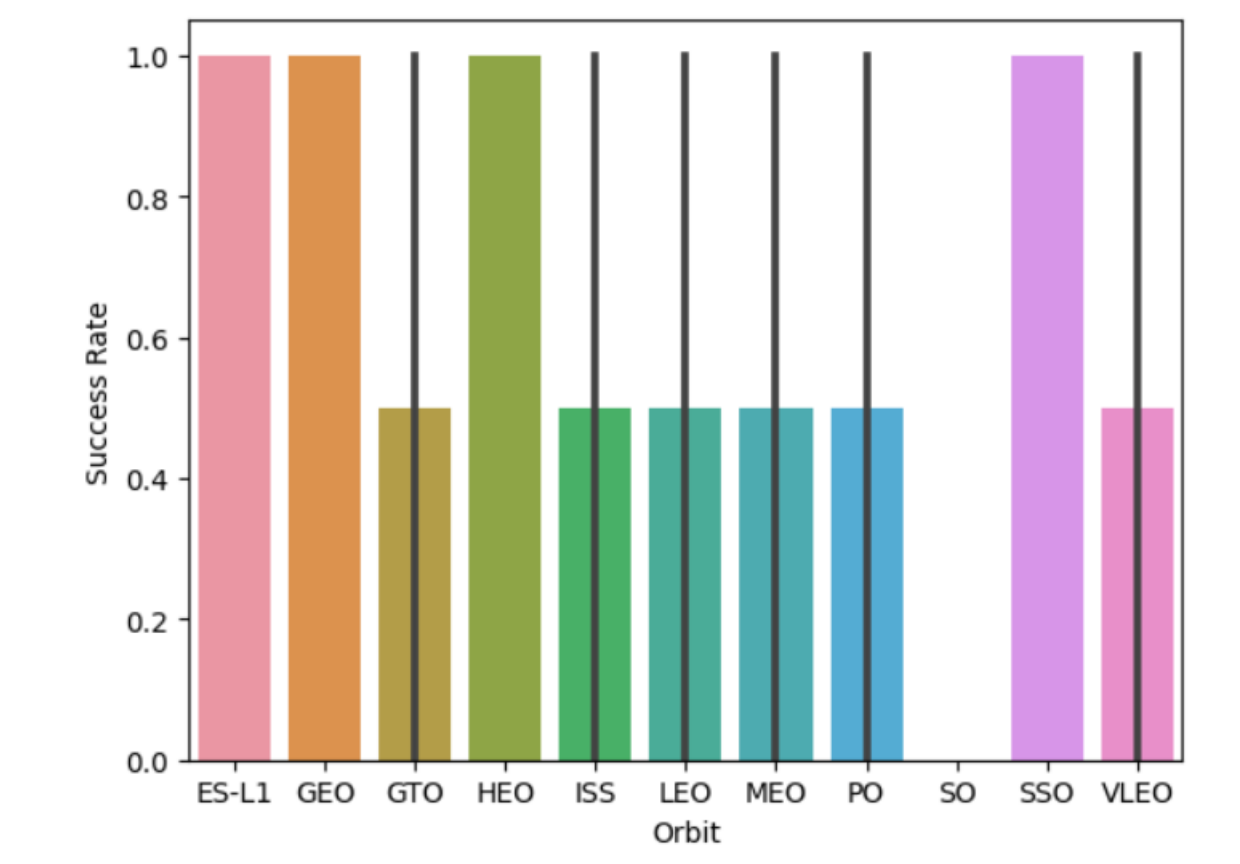
- CCAFS SLC 40 has more flights but no clear success rate
- VAFB SLC 4E has less flights with high success rate
- KSC LC 39A high success rate with a significant flights

PAYLOAD VS. LAUNCH SITE



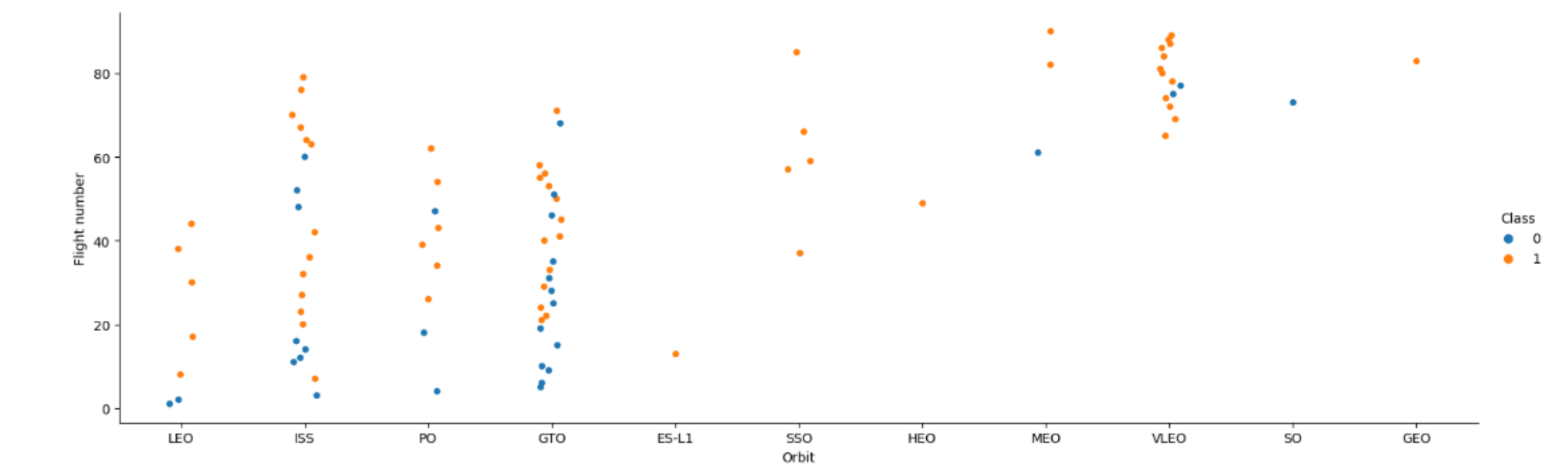
- CCAFS SLC 40 has more success rate above 15k payload mass
- VAFB SLC 4E has no more than 10k payload mass but high success rate just before 10k
- KSC LC 39A high success rate between 2000 and 5000, above 15k also

SUCCESS RATE VS. ORBIT TYPE



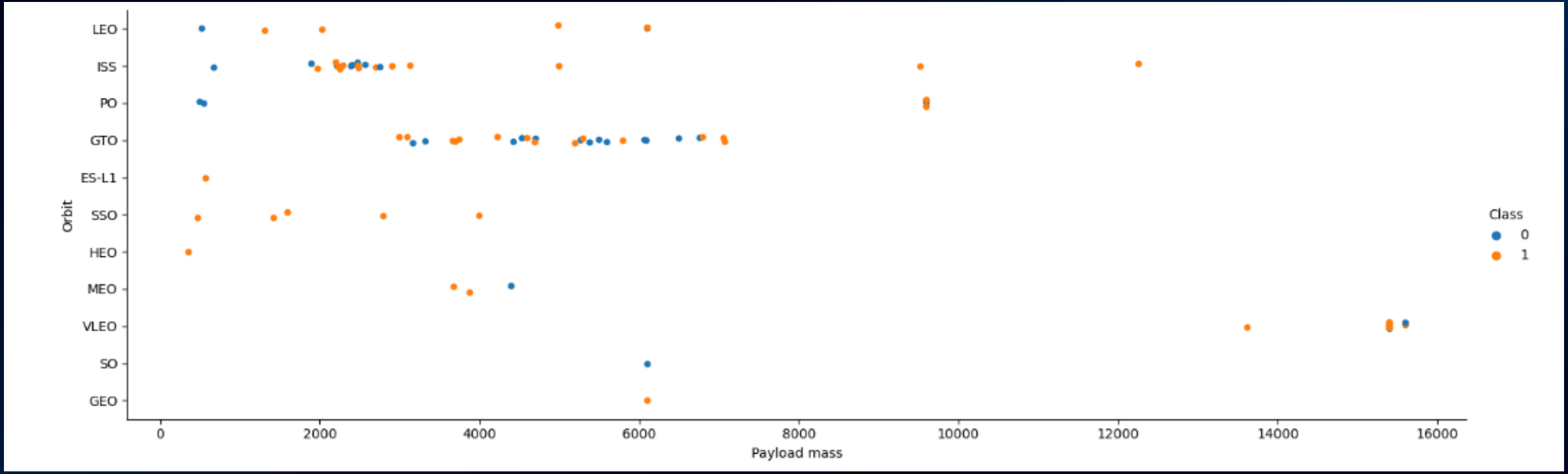
- ES-L1, GEO, HEO, SSO have the highest success rate

FLIGHT NUMBER VS. ORBIT TYPE



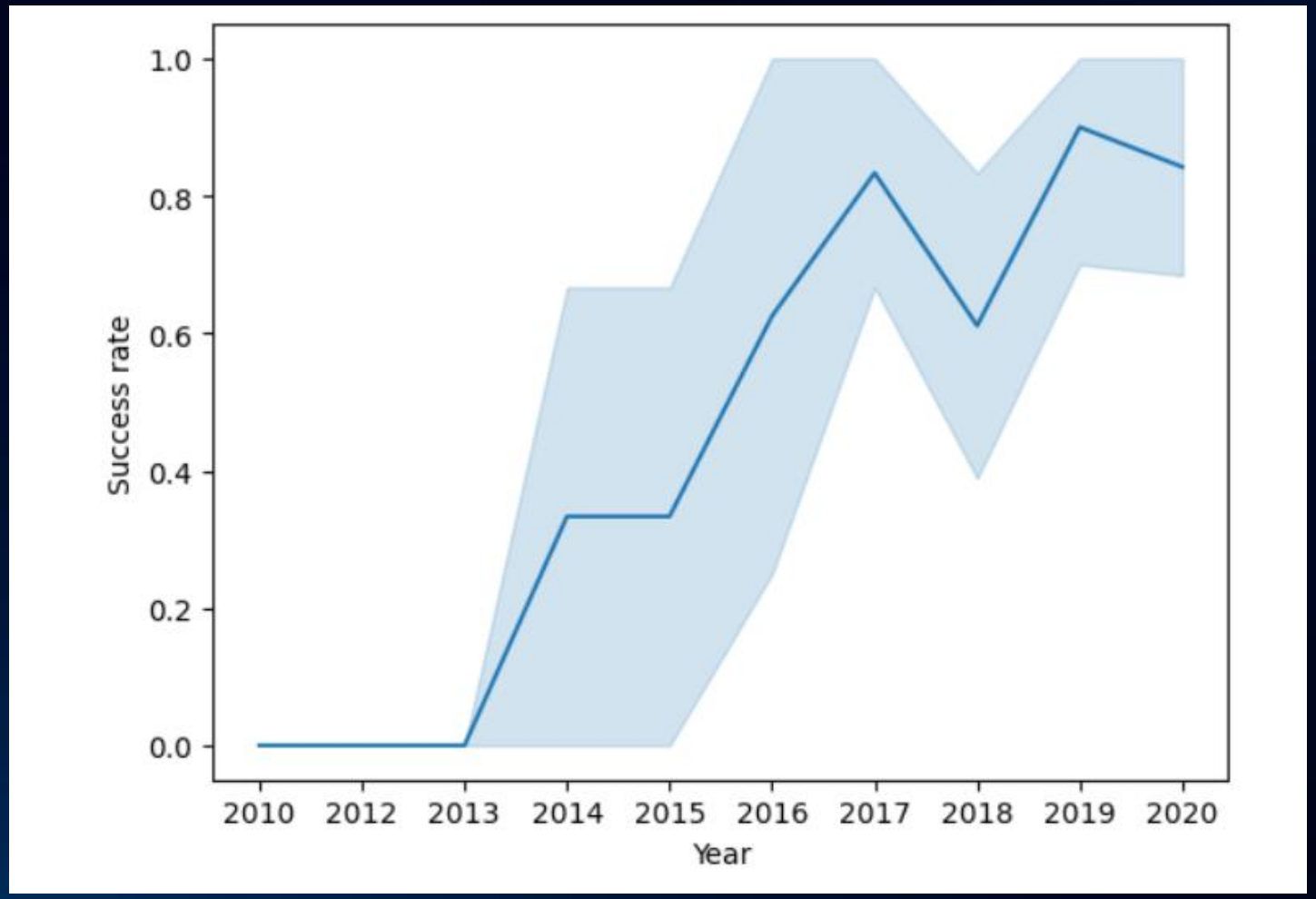
- ISS with more flights and success rate
- SSO less flights but all succeed

PAYLOAD VS. ORBIT TYPE



- SSO high success rate but below 5000 payload mass

LAUNCH SUCCESS YEARLY TREND



- Success rate highly increased after 2013

ALL LAUNCH SITES NAMES

Display the names of the unique launch sites in the space mission

In [9]:
```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

Out[9]:

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

- Distinct used when there are duplicates and we want to see only unique values

LAUNCH SITE BEGIN WITH 'CCA'

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

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

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

[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- We use like for filtering with % means any other characters, we also set the limit 5 to show only 5 records

TOTAL PAYLOAD MASS NASA

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

```
[12]: %sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"
```

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

Out[12]:

SUM(PAYLOAD_MASS__KG_)
45596

- We use sum to get the total of values with setting condition on customer must be “NASA(CRS)”

AVERAGE PAYLOAD MASS F9 V1.1

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

In [13]:
```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like "F9 v1.1"
```

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

Out[13]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- We use avg to get the average of values with setting condition on booster_version must be “F9 v1.1”

FIRST SUCCESSFUL GROUND LANDING DATE

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

Hint:Use min function

In [15]:
```
%sql select MIN(Date) from SPACEXTABLE where Landing_Outcome = "Success (ground pad)"
```

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

Out[15]:

MIN(Date)
2015-12-22

- We use min to get the minimum value in that case the first date with setting condition on landing_outcome must be “Success (ground pad)”

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

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

```
%sql select Customer from SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and
```

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

Customer
SKY Perfect JSAT Group
SKY Perfect JSAT Group
SES
SES EchoStar

- We use between to specify a range (min,max)

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
```

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

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

- We use group by to cluster the same values together and then apply count to each one

BOOSTERS CARRIED MAXIMUM PAYLOAD

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

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- In this case we used a subquery to determine the max mass because we cannot apply it directly in the where clause

2015 LAUNCH RECORDS

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

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql select substr(Date,6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE \
where substr(Date,1,4)='2015' and Landing_Outcome = "Failure (drone ship)"
```

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

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

- In this case we used substr to extract the month and the year from the date

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

Rank 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.

```
%sql select Landing_Outcome,count(*) as count from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Lar
```

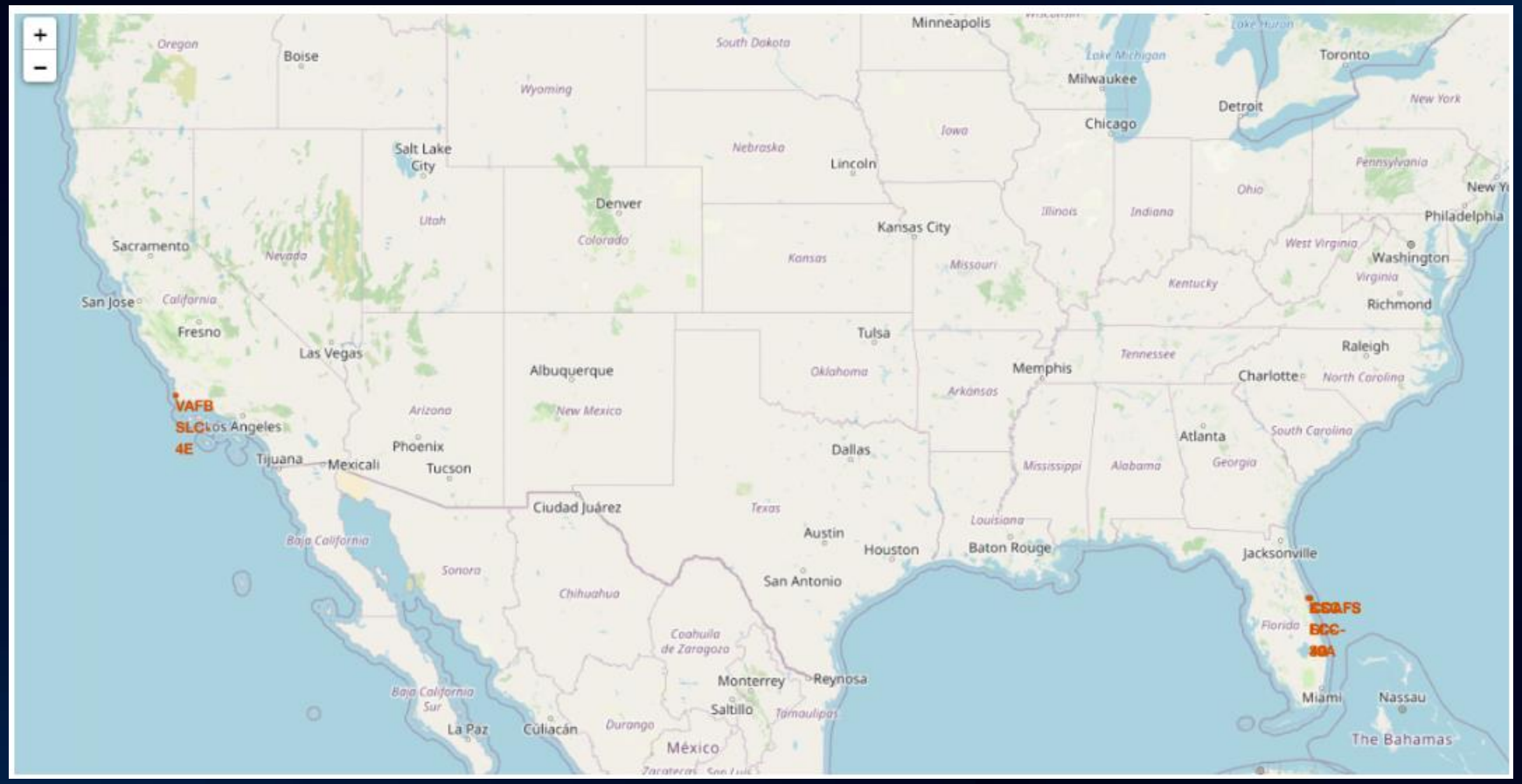
```
 * sqlite:///my_data1.db
Done.
```

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

- Here we grouped by landing_outcome and apply count to each group then order by count column then we added desc to let them in descending order

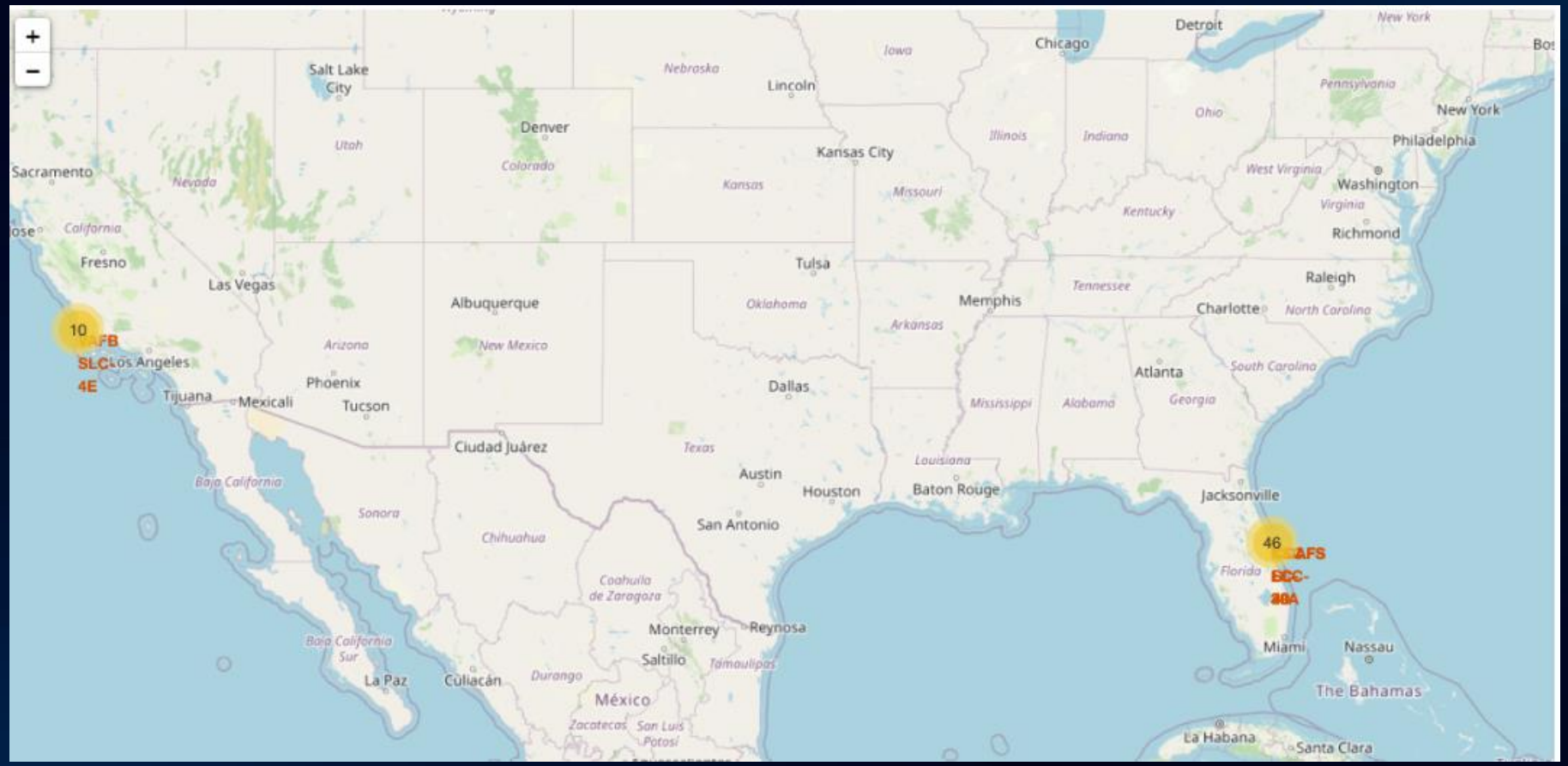
SECTION 3
LAUNCH SITES PROXIMITIES ANALYSIS

MARKED LAUNCH SITES



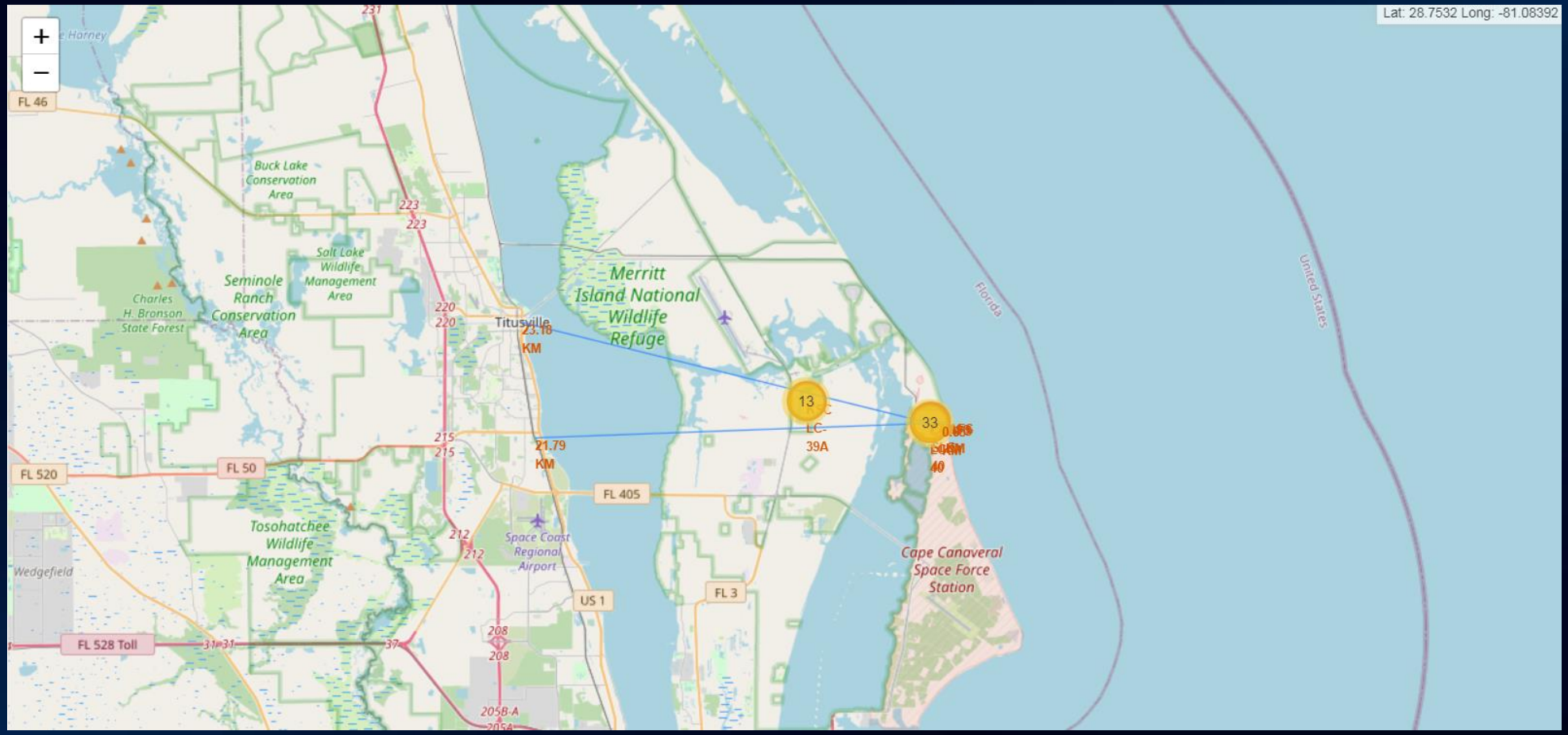
- All the launch sites are so close to coast and near the equator line

LAUNCH CLASS CLUSTERS



- CCAFS sites have relatively higher success rate

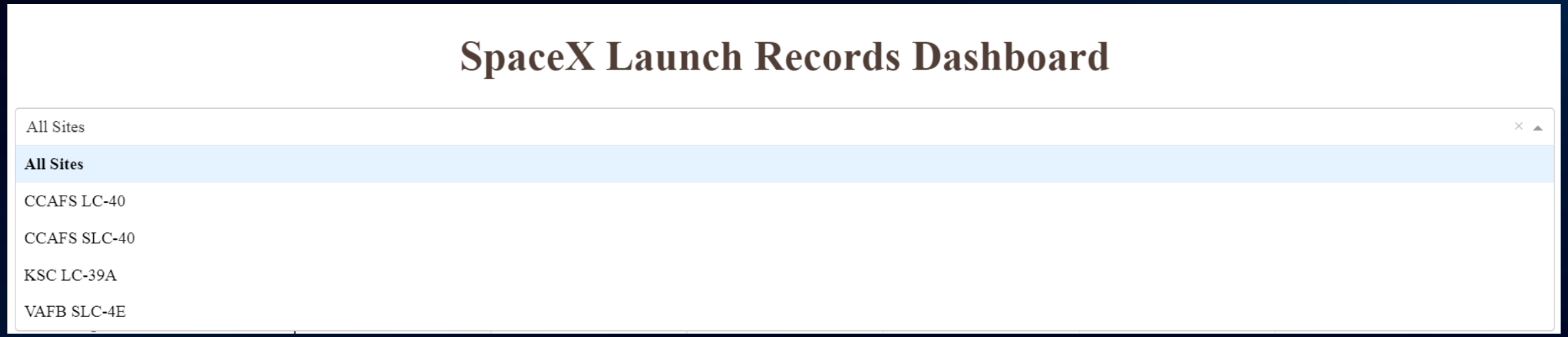
LAUNCH SITES PROXIMITIES



- CCAFS launch sites are close to parkway highway but still not close to any railway or city

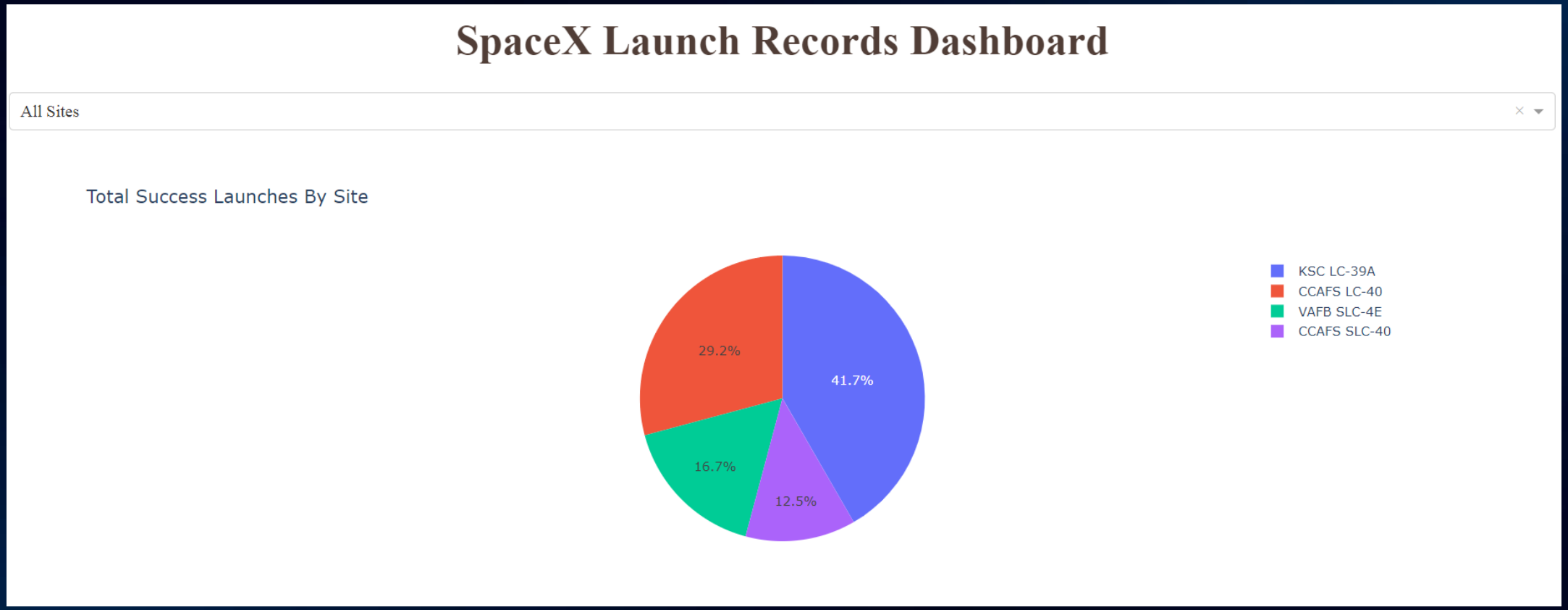
SECTION 4
BUILD A DASHBOARD WITH PLOTLY DASH

LAUNCH SITES Dropdown list



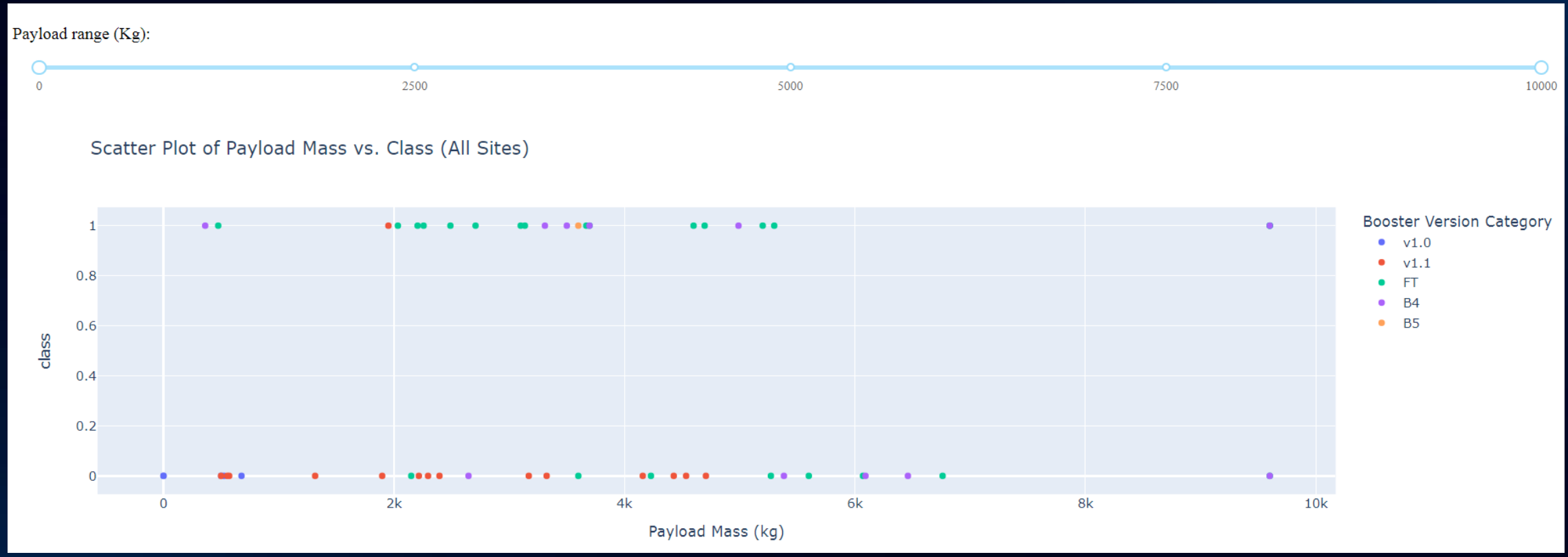
- Once you select a site dash will automatically render the appropriate figure for that site

TOTAL SUCCESS LAUNCHES BY SITE

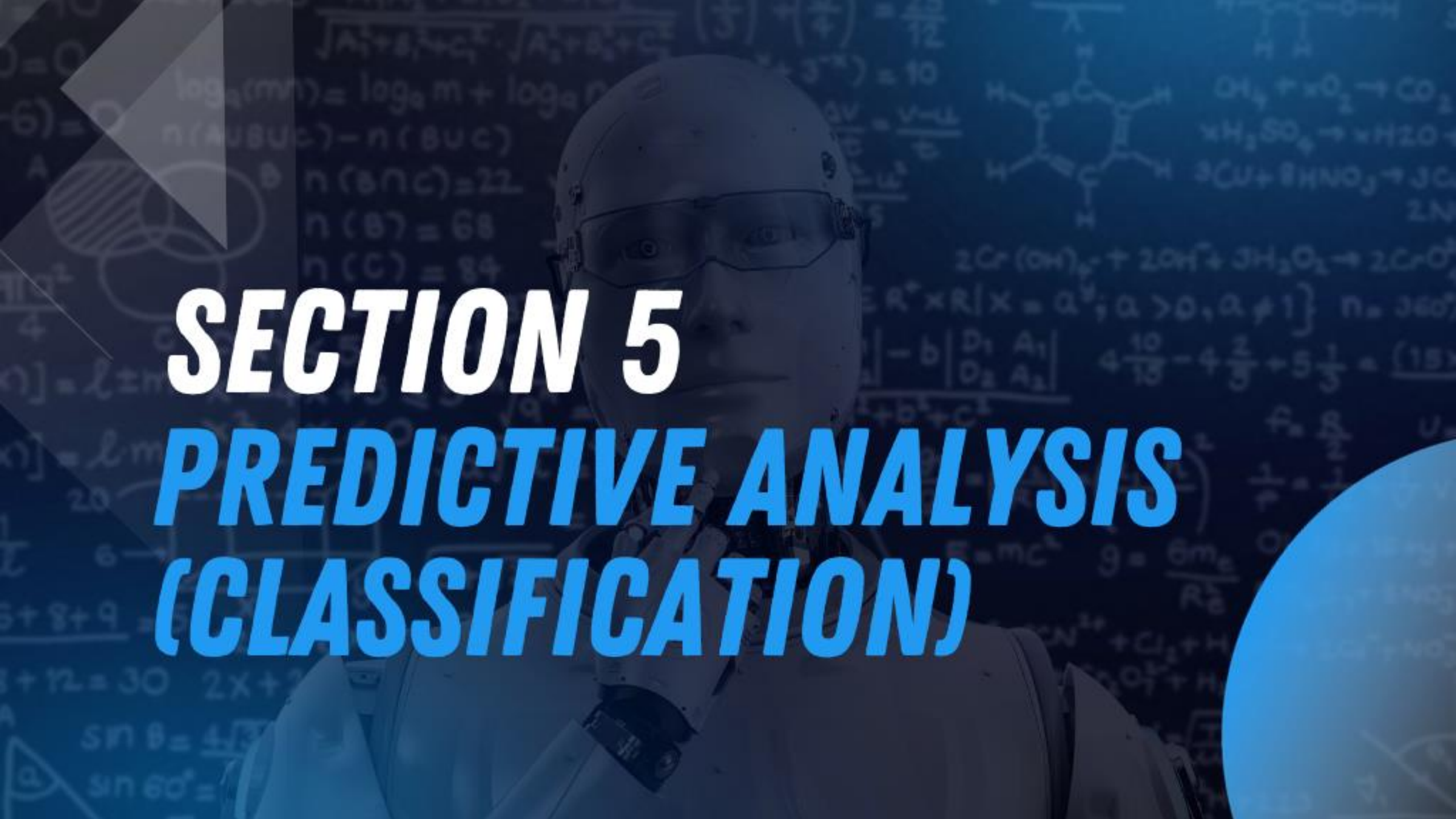


- KSC LC 39A has the most success launches

CORRELATION PAYLOAD VS. LAUNCH

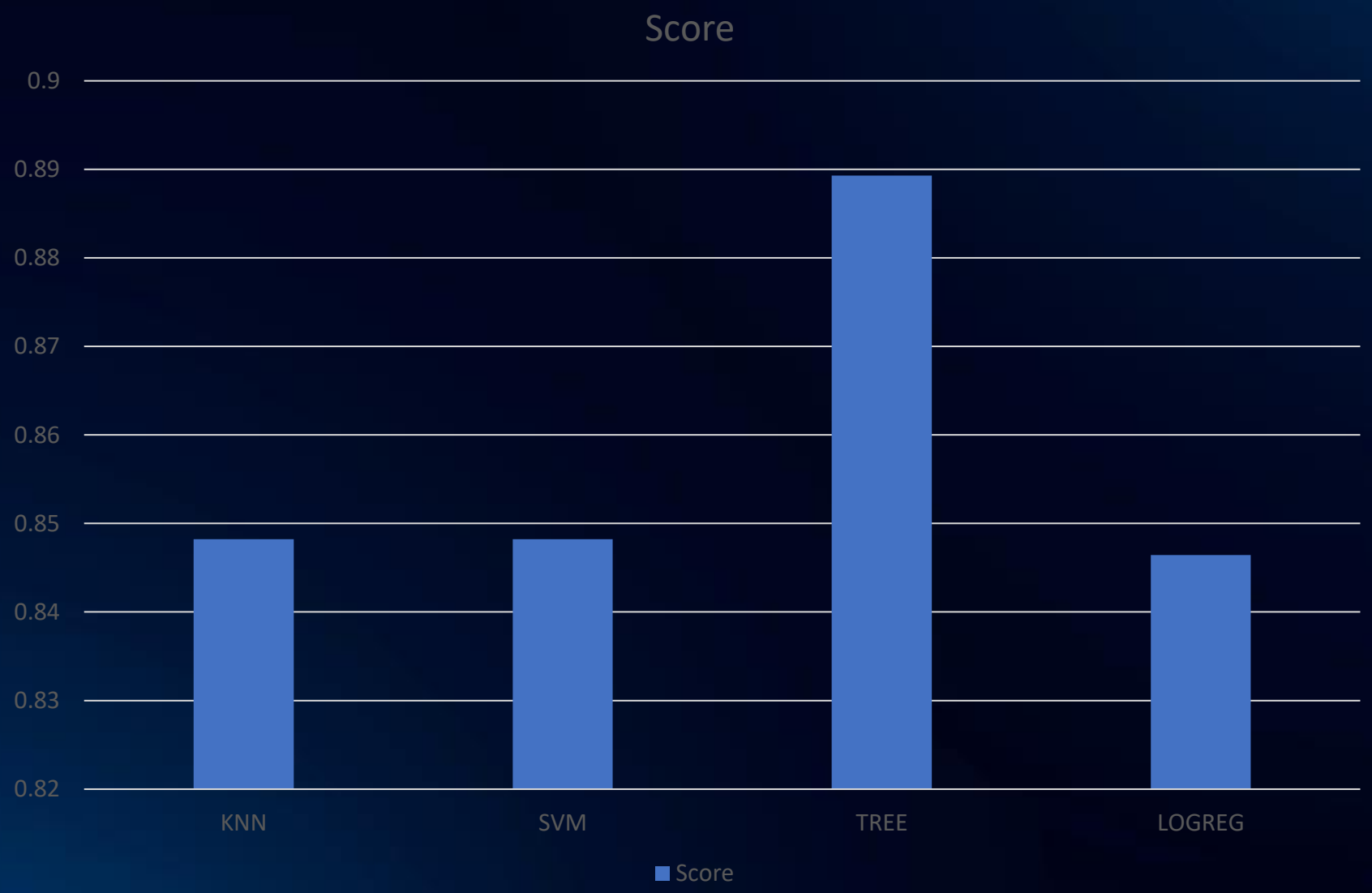


- FT has the most success launches
- Payload mass between 4k and 7k contributed to more failed launches

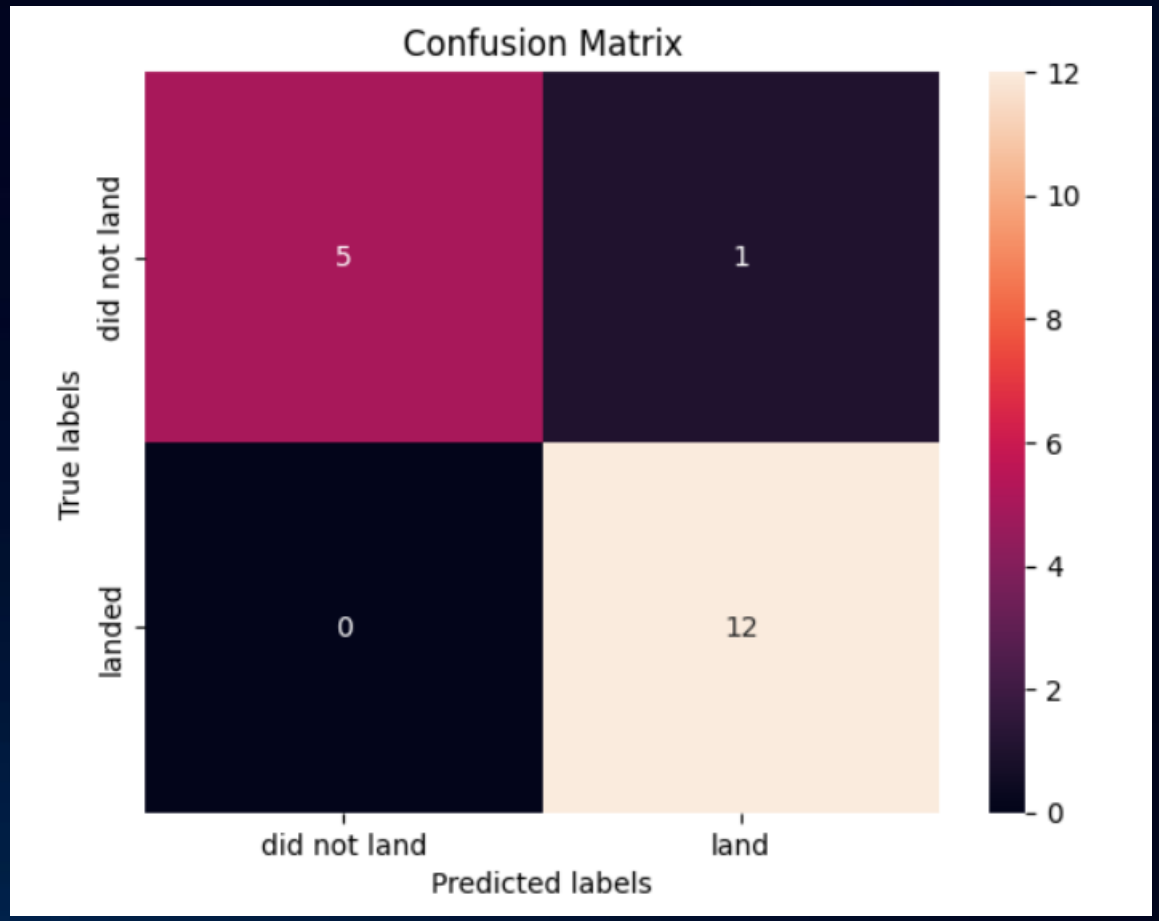


SECTION 5
PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY



CONFUSION MATRIX



- For the not landing case, the model predicts 5 correctly out of 6 which is good
- For the landing case, the model successfully predicts 12 out of 12 correctly which is so good

CONCLUSION

- Most of the launch sites are locating near the equator line and so close to the coast
- ES-L1, GEO, HEO, and SSO have a 100% success rate
- KSC LC-39A has the highest success rate
- The decision tree model scored higher than others and has better predictions according to the confusion matrix

APPENDIX

Dash Documentation : https://dash.plotly.com
SpaceX Falcon9 Datasheet : https://sma.nasa.gov/LaunchVehicle/assets/spacex-falcon-9-data-sheet.pdf
Falcon 9 Wiki : https://en.wikipedia.org/wiki/Falcon_9
SpaceX API Docs : https://docs.spacexdata.com

THANK YOU