IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Franco Messetzian
October 2021

OUTLINE

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

EXECUTIVE SUMMARY

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive Analytics shown in screenshots
  - Predictive Analytics result

INTRODUCTION

- SpaceX is a company that designs, manufactures and launches advanced rockets and spacecraft. One of their main competitive advantages in the rocket launching market is the cost at which they offer their launch, which is nearly a third of the price of their competitors. They achieve this by re using the first stage.
- In this project we are going to analyze the available information in order to determine (and therefore be able to confidently predict) the best indicators for their Falcon9 successful landings. In term, this will accurately determine the cost of each launch, and valuable information for a SpaceX competitor.

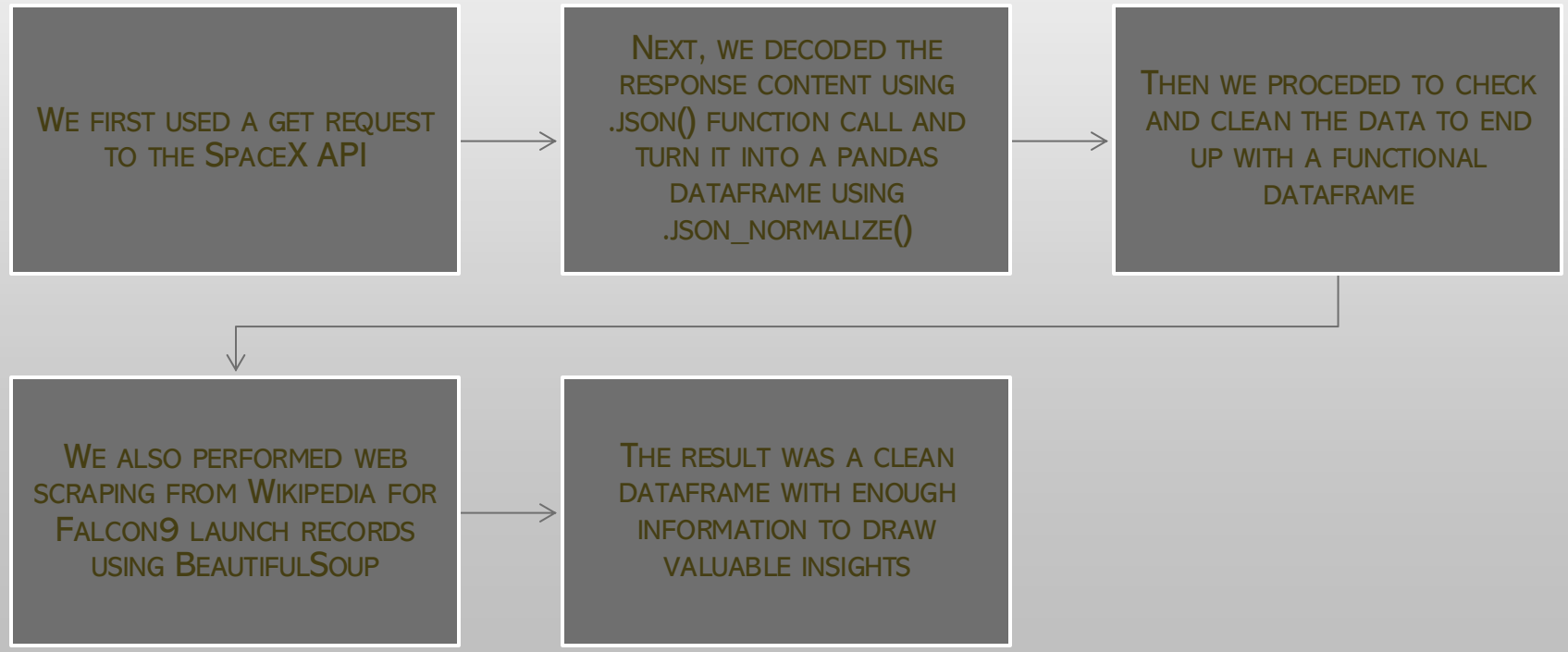
Section 1

Methodology

METHODOLOGY

- Executive Summary
- Data collection methodology:
  - Data was collected by web scraping SpaceX API's from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- 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

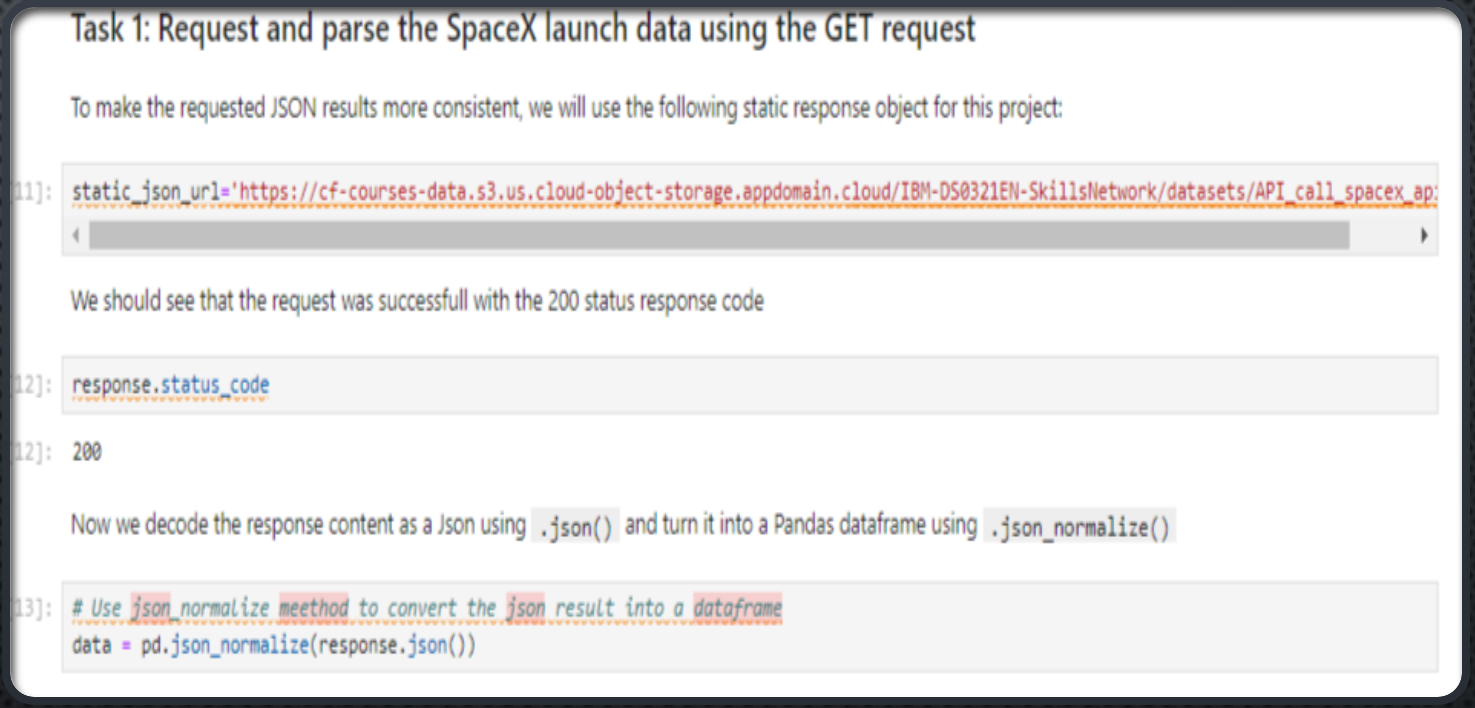
DATA COLLECTION



DATA COLLECTION – SPACEX API

- We used the GET request command on SpaceX API to collect and clean the data. Then did some basic data wrangling and formatting.

- Click on the Data Collection notebook for details



Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
[11]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_ap
```

We should see that the request was successfull with the 200 status response code

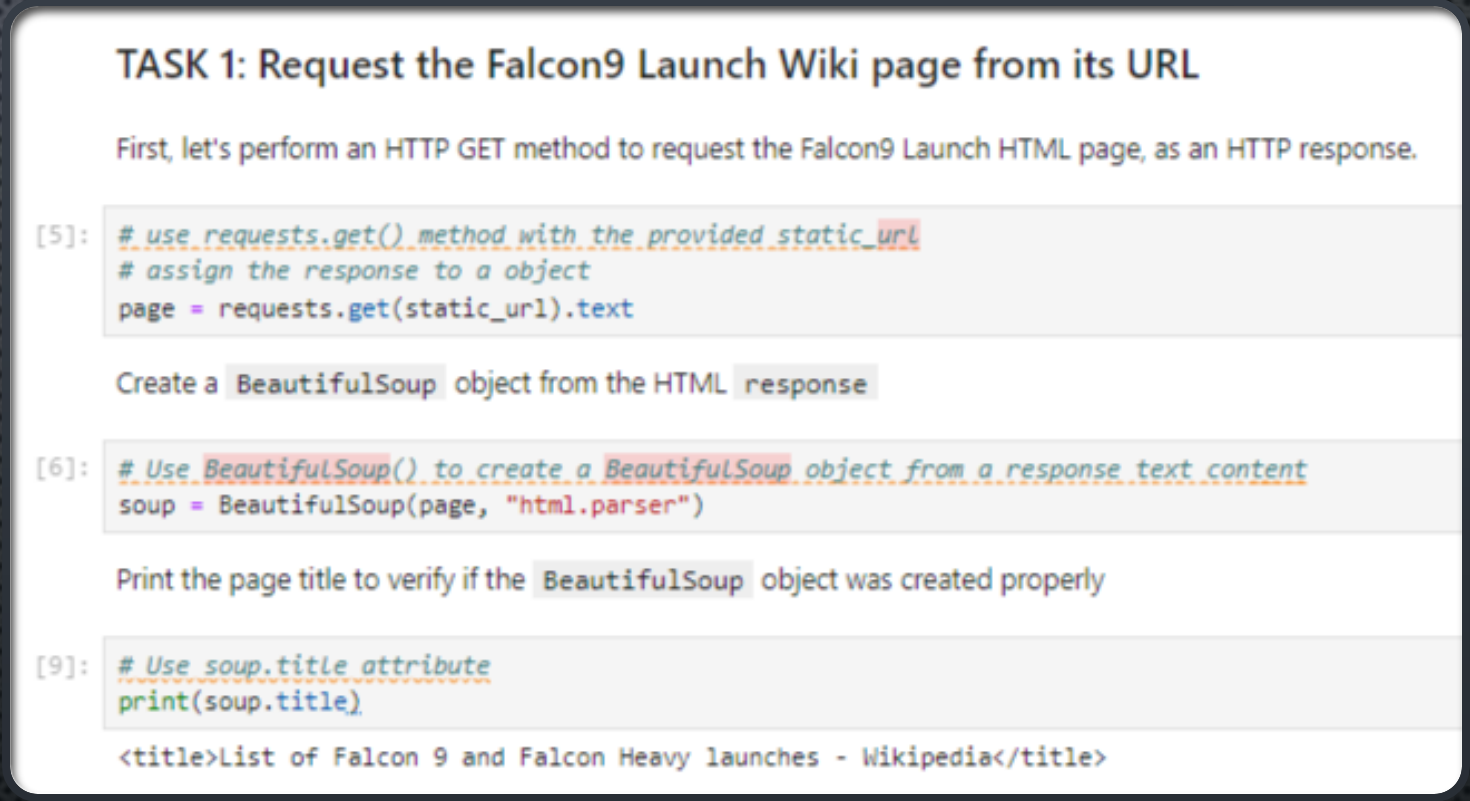
```
[12]: response.status_code

[12]: 200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

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

DATA COLLECTION - SCRAPING

- We did web scraping for Flacon9 launch records using BeautifulSoup
- We parsed the table and converted it into a pandas dataframe
- Click on the Web Scraping notebook for details



TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

[5]:
```
# use requests.get() method with the provided static_url
# assign the response to a object
page = requests.get(static_url).text
```

Create a `BeautifulSoup` object from the HTML `response`

[6]:
```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(page, "html.parser")
```

Print the page title to verify if the `BeautifulSoup` object was created properly

[9]:
```
# Use soup.title attribute
print(soup.title)
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

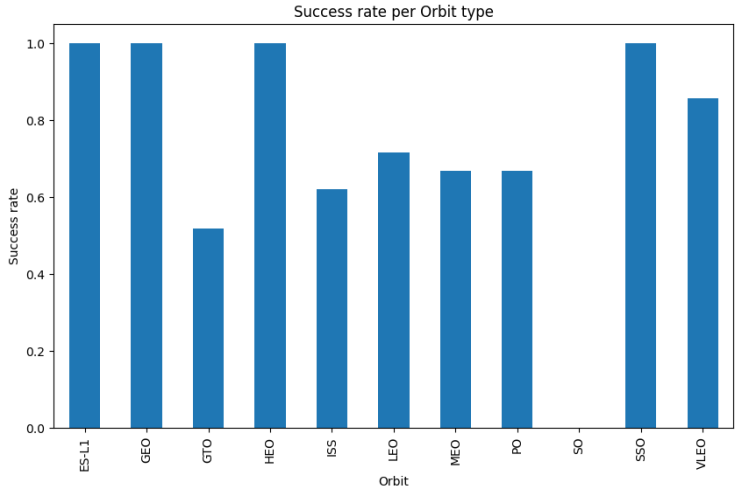
DATA WRANGLING

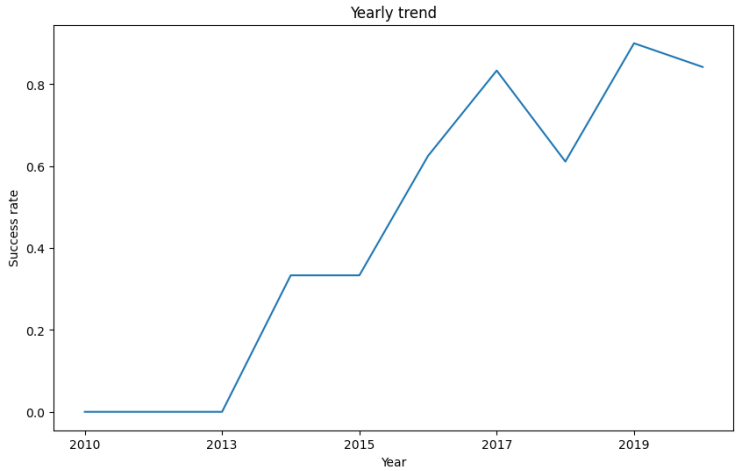
- We performed exploratory data analysis and determined training levels
- We calculated the number of launches at each site and the number of occurance of each orbit
- We then created a landing outcome label from the outcome column of our dataframe

- Refer to the Data Wrangling notebook for details

BUILD A DASHBOARD WITH PLOTLY DASH

- We explored the data by visualizing the relationship between different variables in order to detect dependencies for further analysis.
- The graphs on the right show how certain orbits have better success rate (top), and the general positive relationship between success rate and years since first launch (bottom)
- See EDA data visualization notebook for details





EDA WITH SQL

- We loaded the SpaceX dataset into a PostgreSQL database
- Using SQL commands, we explored the dataset for insights such us:
  - Names of unique launch sites
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Failed landing outcomes in drone ship with their booster version and launch site names

•See EDA with SQL for further details

BUILD AN INTERACTIVE MAP WITH FOLIUM

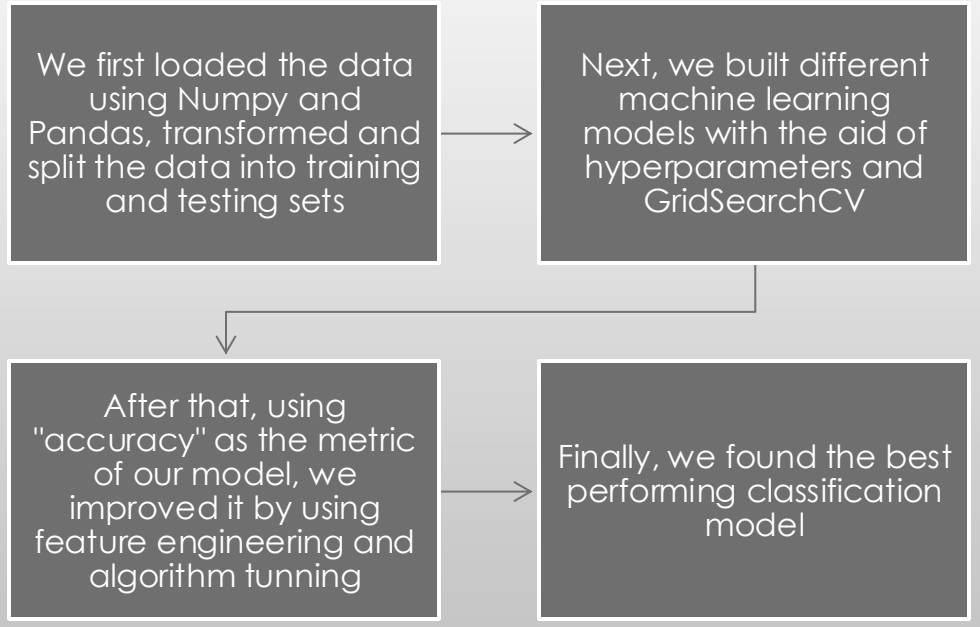
- We created a Folium map, then marked all launch sites and added map objects such as markers, circles and lines.
- These markers helped us visualize the success/failure launches for each site on the map

- Refer to Folium notebook for full details

BUILD A DASHBOARD WITH PLOTLY DASH

- We built an interactive dashboard with Plotly Dash
- We plotted pie charts showing the total launches given a site
- We also plotted a scatter graph for different booster versions, where we can see Payload and Outcome for each
- See Dashboard notebook for details

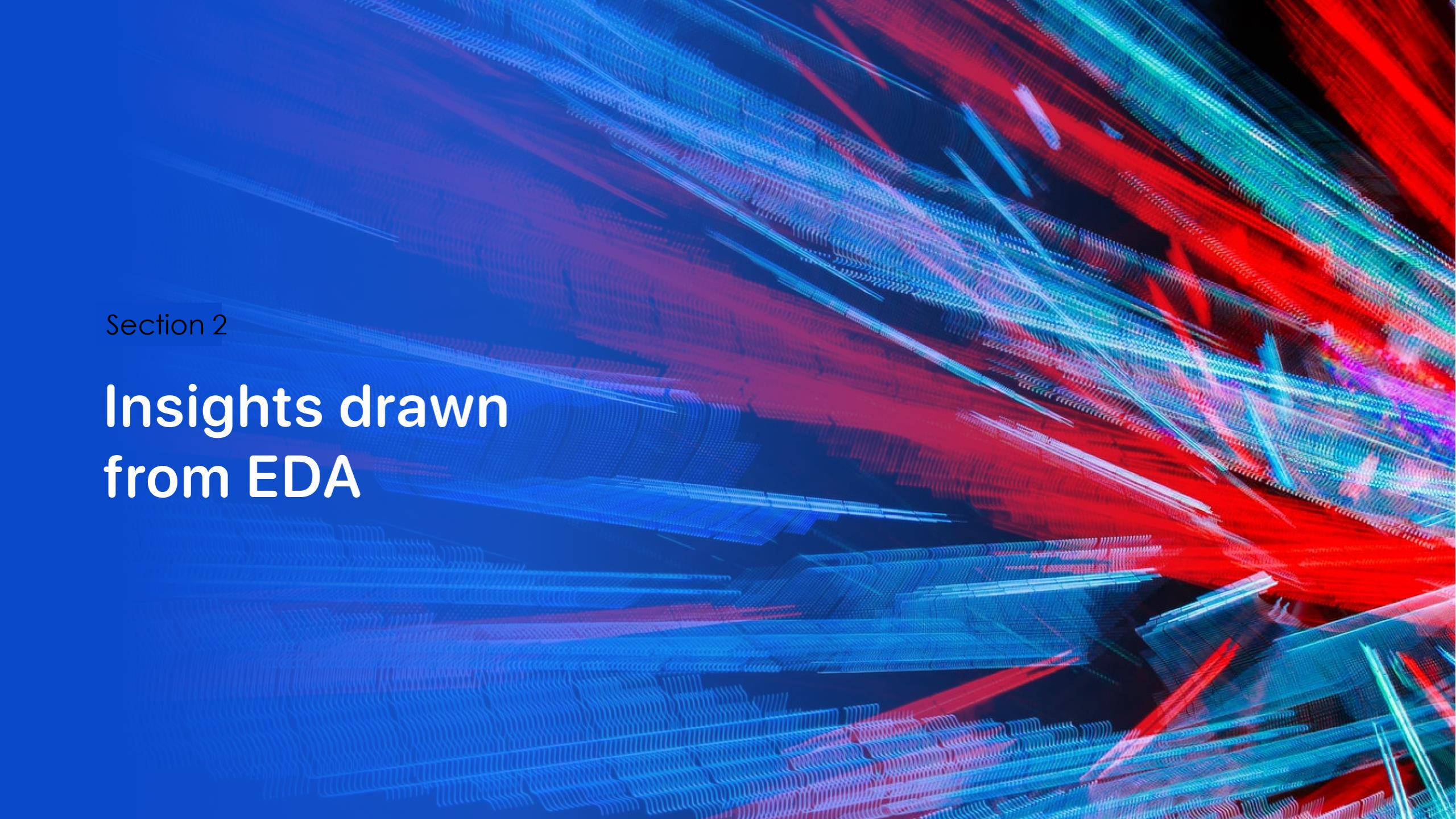
PREDICTIVE ANALYSIS (CLASSIFICATION)



- FIND THE FULL DETAILS AT PREDICTIVE ANALYSIS NOTEBOOK

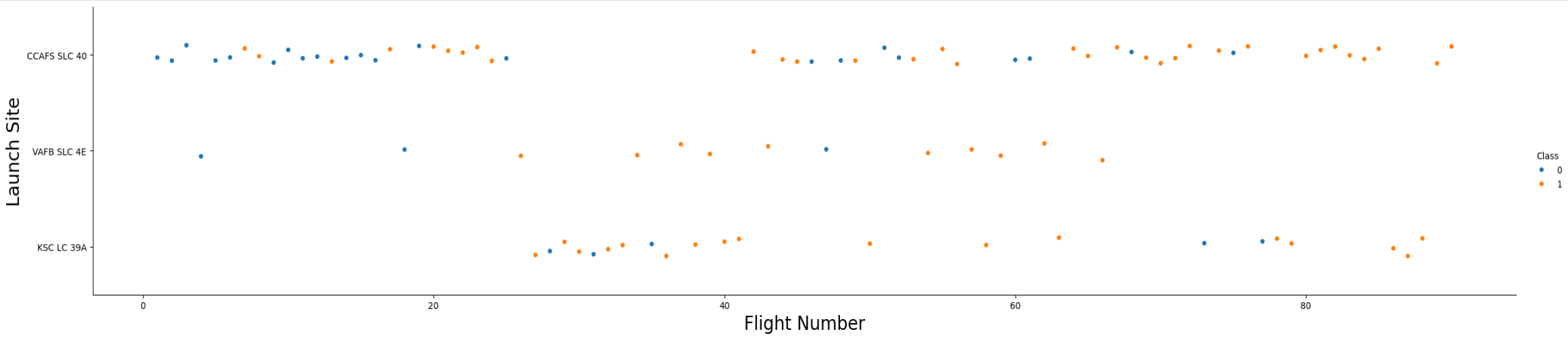
RESULTS

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


Section 2
Insights drawn from EDA

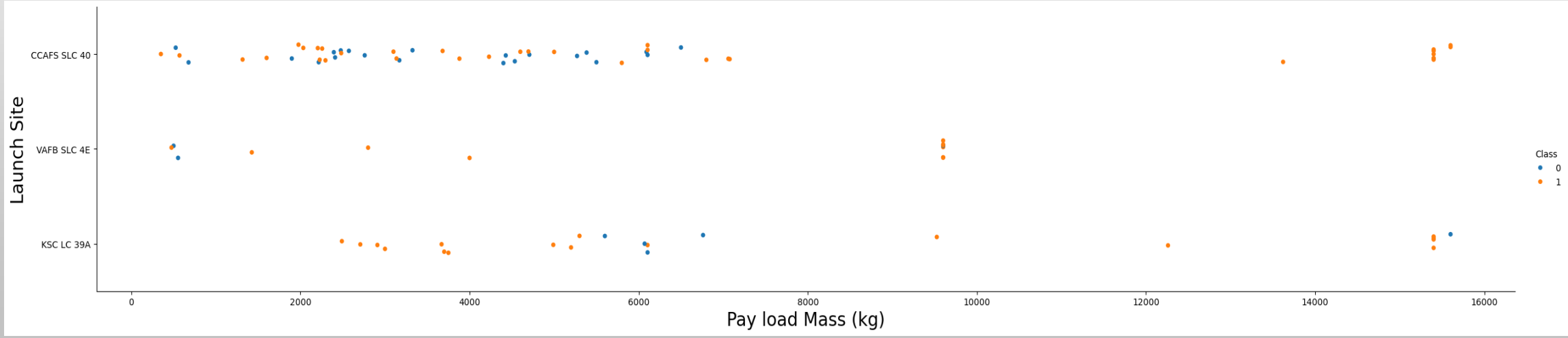
FLIGHT NUMBER VS. LAUNCH SITE

- WE CAN SEE ON THE PLOT THAT AS FLIGHT NUMBERS INCREASED, SO DID THE SUCCESS RATE FOR EACH SITE



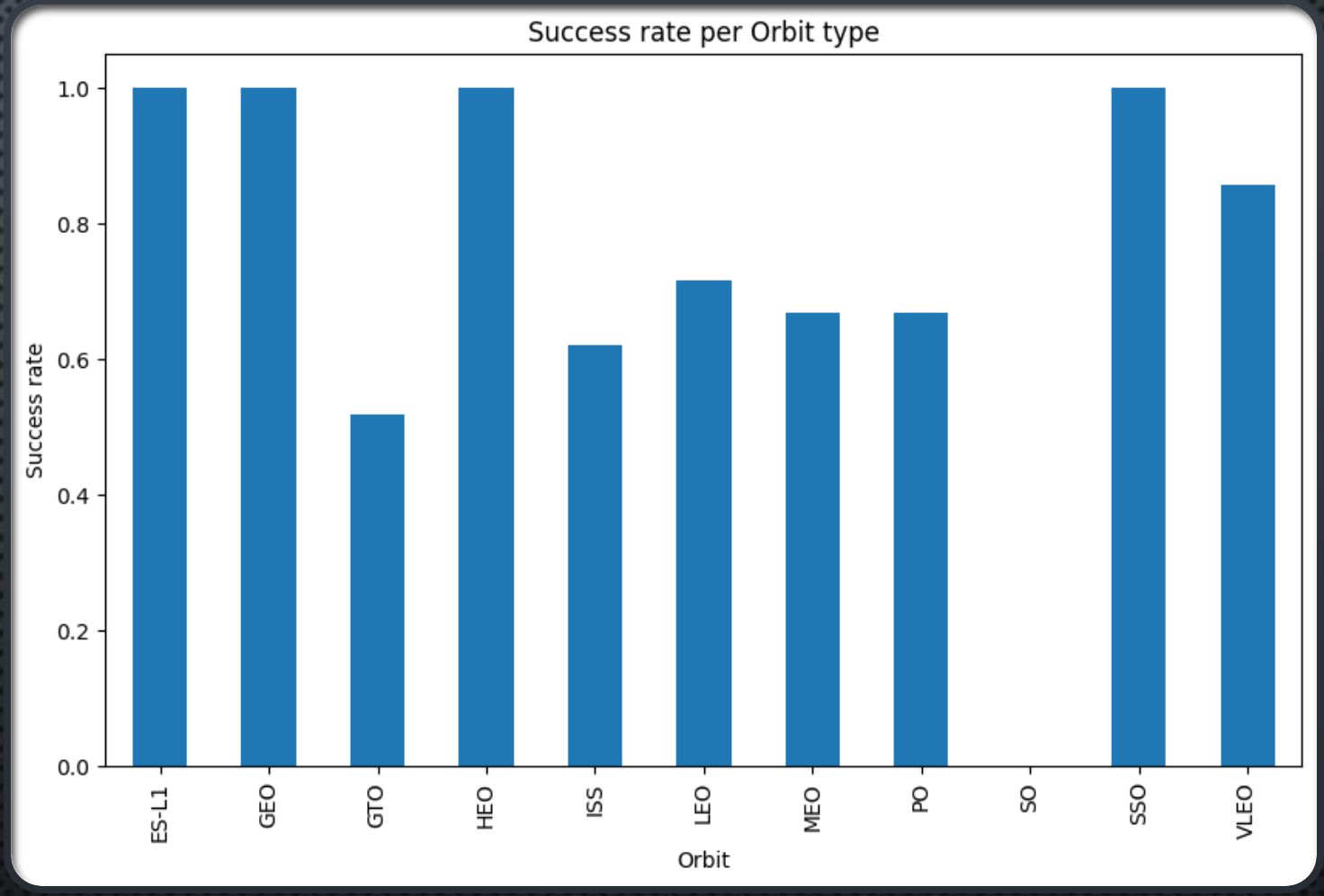
PAYLOAD VS. LAUNCH SITE

- We can observe that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)



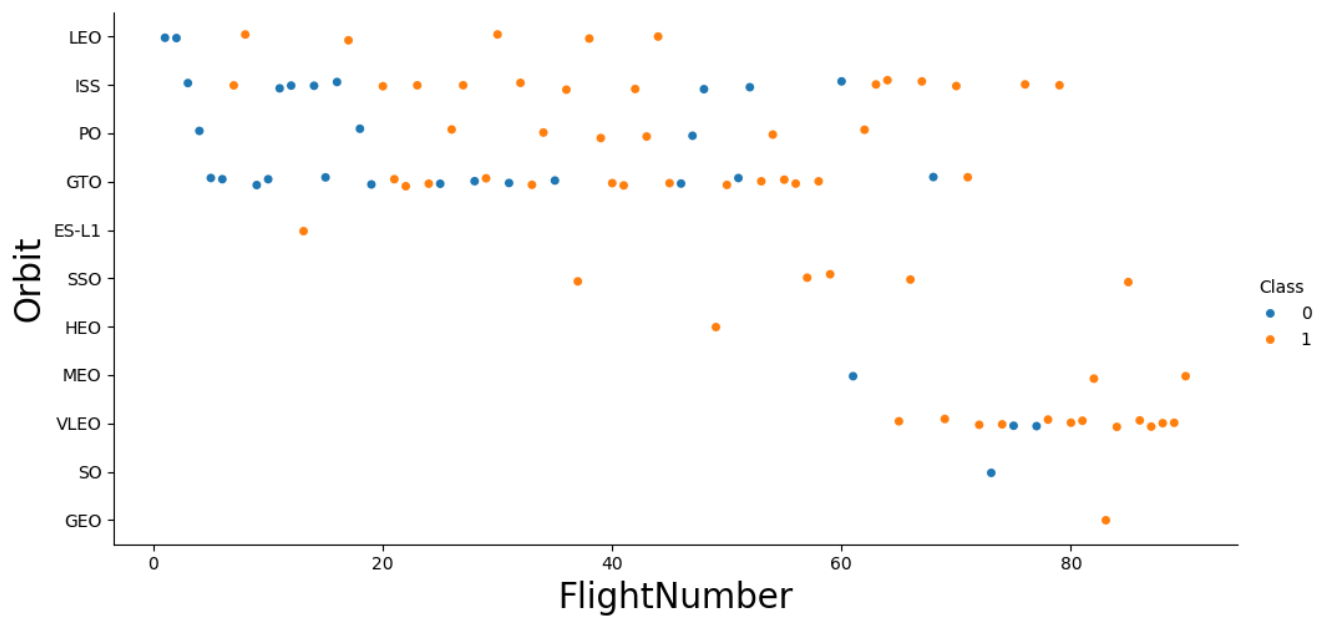
SUCCESS RATE VS. ORBIT TYPE

- We can clearly establish that ES-L1, GEO, HEO and SSO lead the Orbit type success rate



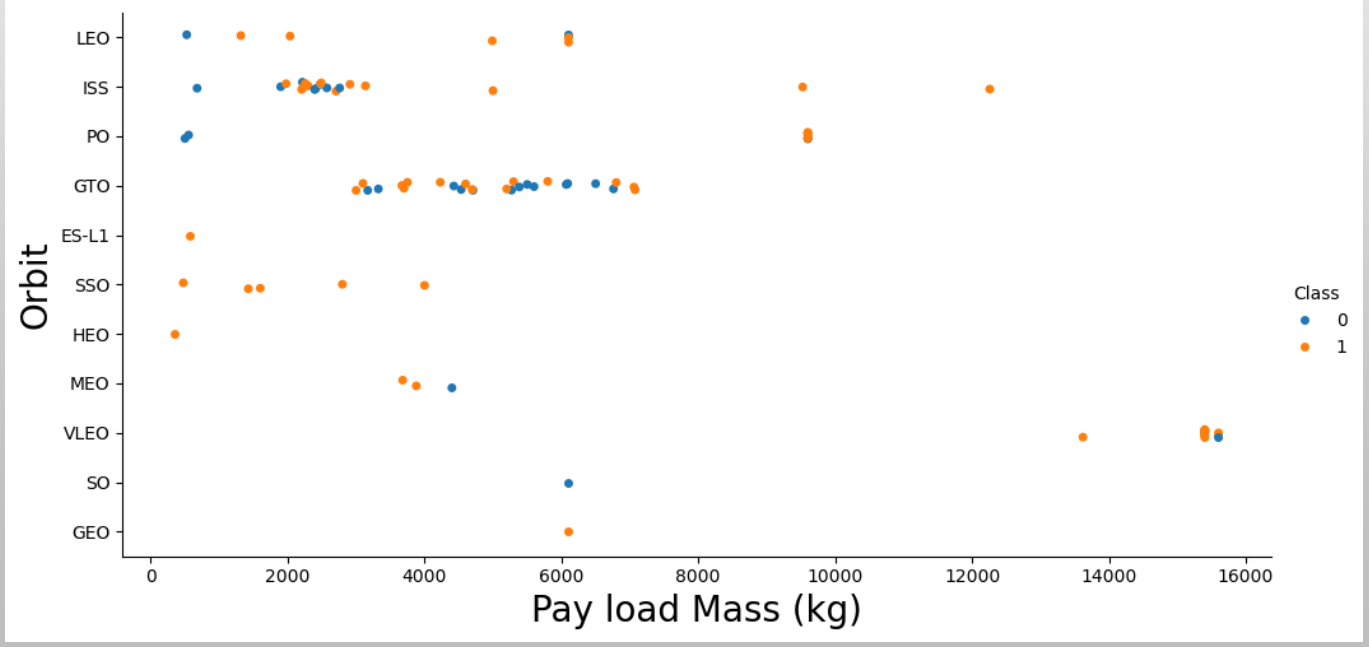
FLIGHT VS. ORBIT TYPE

- IT'S INTERESTING TO NOTE THAT FOR SOME ORBITS, SUCH AS LEO, THERE IS A DIRECT RELATIONSHIP BETWEEN THE NUMBER OF FLIGHTS AND SUCCESS RATE.ON THE CONTRARY, ORBITS LIKE GTO OR ISS, THAT DOESN'T SEEM THE CASE



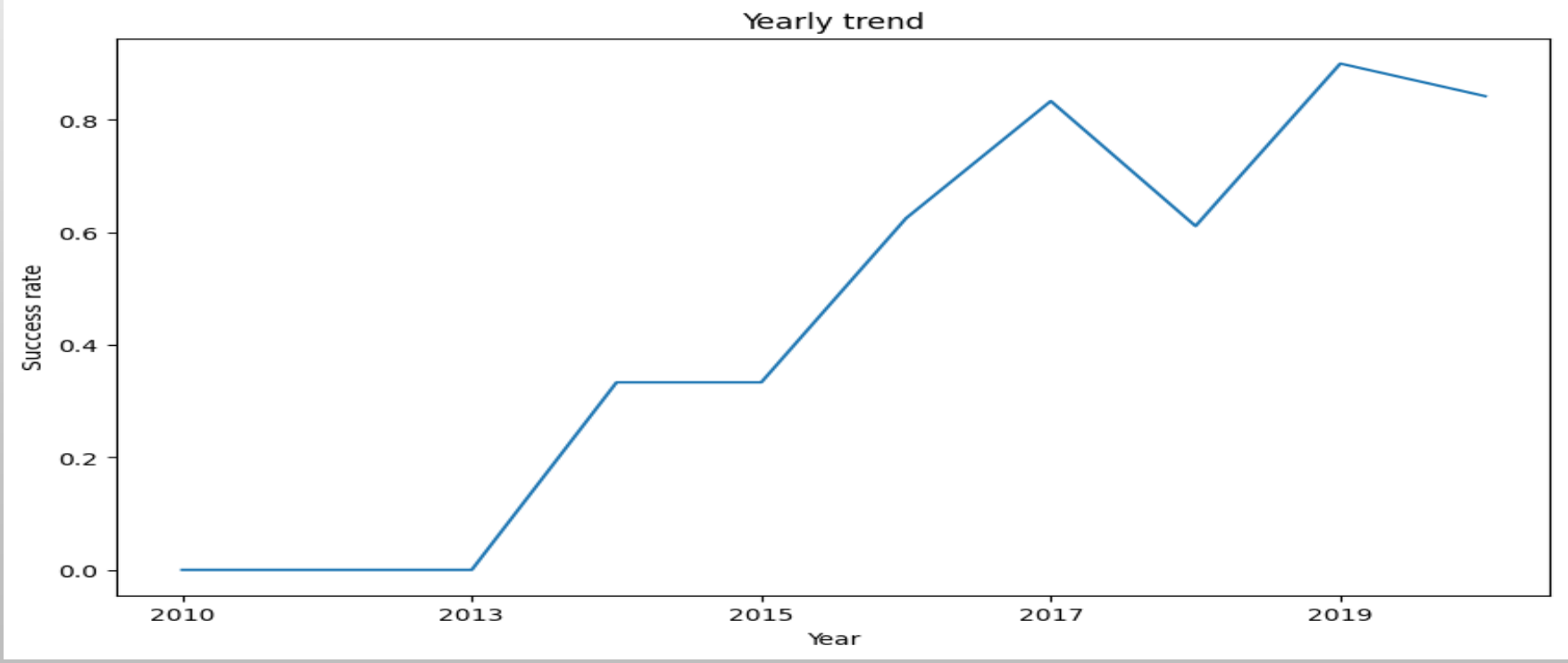
PAYLOAD VS. ORBIT TYPE

- We can see that for heavy payloads, success rate correlates more with PO, LEO and ISS orbits



LAUNCH SUCCESS YEARLY TREND

- FROM THE GRAPH WE CAN SEE HOW SUCCESS INCREASED STEADILY FROM 2013 ONWARDS



ALL LAUNCH SITE NAMES

- USING THE COMMAND DISTINCT IN SQL WE CAN GET THE SITES LIST

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

```
[7]: %sql select distinct "Launch_Site" from SPACEXTABLE
```

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

[7]:

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

LAUNCH SITE NAMES BEGIN WITH 'KSC'

- USING THE CONDITIONAL 'WHERE' AND 'LIKE' IN SQL WE GET THE DESIRED OUTCOME

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

[9]:
```
%sql select * from SPACEXTABLE where "Launch_Site" like 'KSC%' limit 5
```

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

[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

TOTAL PAYLOAD MASS

- USING SQL'S FUNCTION 'SUM' AND CONDITIONAL '

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

```
[15]: %sql select sum("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer" = 'NASA (CRS)'
```

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

[15]:

sum("PAYLOAD_MASS__KG_")
45596

AVERAGE PAYLOAD MASS BY F9 V1.1

- We calculated the average payload mass carried by F9 v1.1 as 2928.4

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

```
16]: %sql select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" = 'F9 v1.1'
```

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

16]:

avg("PAYLOAD_MASS__KG_")
2928.4

FIRST SUCCESSFUL (DRONE SHIP) GROUND LANDING DATE

- We can see that the first successful landing was achieved on the 8th April 2016

List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

```
[36]: %sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)'
```

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

[36]:

min("Date")
2016-04-08

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

- As shown below, the following booster versions have successfully landed payloads between 4000 and 6000: F9 FT B1032.1, F9 B4 B1040.1 and F9 B4 B1043.1

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

```
39]: %sql select distinct("Booster_Version") from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)' and "PAYLOAD_MASS__KG_" >4000 an
```

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

39]:

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

- AS SHOWN BELOW, THERE HAVE BEEN 100 SUCCESSES AND 1 FAILURE

List the total number of successful and failure mission outcomes

```
%sql select count("Mission_Outcome") as "Failure_count" from SPACEXTABLE where "Mission_Outcome" like 'Failure%'
```

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

Failure_count
1

```
%sql select count("Mission_Outcome") as "Success_count" from SPACEXTABLE where "Mission_Outcome" like 'Success%'
```

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

Success_count
100

BOOSTERS CARRIED MAXIMUM PAYLOAD

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

```
%sql select distinct("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

- SELECTING THE DISTINCT BOOSTER VERSION AND APPLYING THE CONDITION FOR MAX PAYLOAD, WE GET THE RESULTS AS SHOWN ON THE LEFT

2017 LAUNCH RECORDS

List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

Note: SQLLite does not support monthnames. So you need to use substr(Date,6,2) for month, substr(Date,9,2) for date, substr(Date,0,5),='2017' for year.

```
[60]: %sql select substr("Date",6,2), "Landing_Outcome", "Booster_Version", "launch_Site" from SPACEXTABLE where "Landing_Outcome" = 'Success (g
```

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

[60]:

substr("Date",6,2)	Landing_Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- SELECTING THE DESIRED COLUMNS, AND APPLYING THE CONDITIONS FOR SUCCESSFUL LANDING AND SUBSTR("DATE",0,5) = '2017' WE CAN SEE THE LAUNCH RECORDS FOR EVERY ACTIVE MONTH IN 2017

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 "Date", "Landing_Outcome", count("Landing_Outome") as "Count of Landing Outcomes" from SPACEXTABLE where "Date" between "2010-
```

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

Date	Landing_Outcome	Count of Landing Outcomes
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

- We selected "Date" and COUNT of landing outcomes. Then used the WHERE clause to filter the desired range.
- We applied the GROUPBY clause on Landing Outcomes and ORDERBY the newly created column 3 in descending order

Section 3

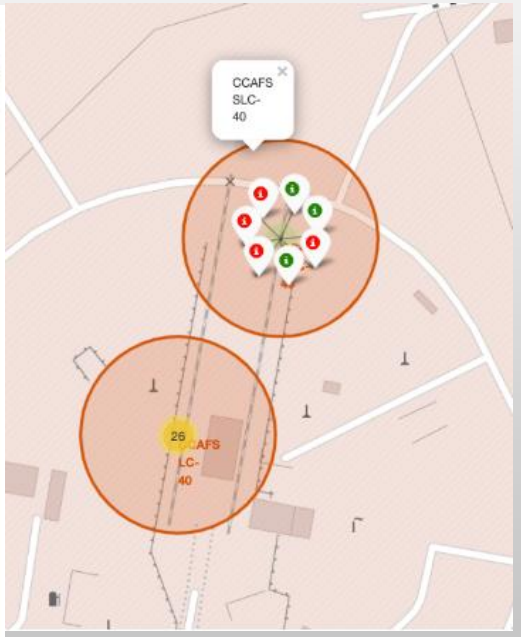
Launch Sites Proximities Analysis

ALL LAUNCH SITES WORLD MAP

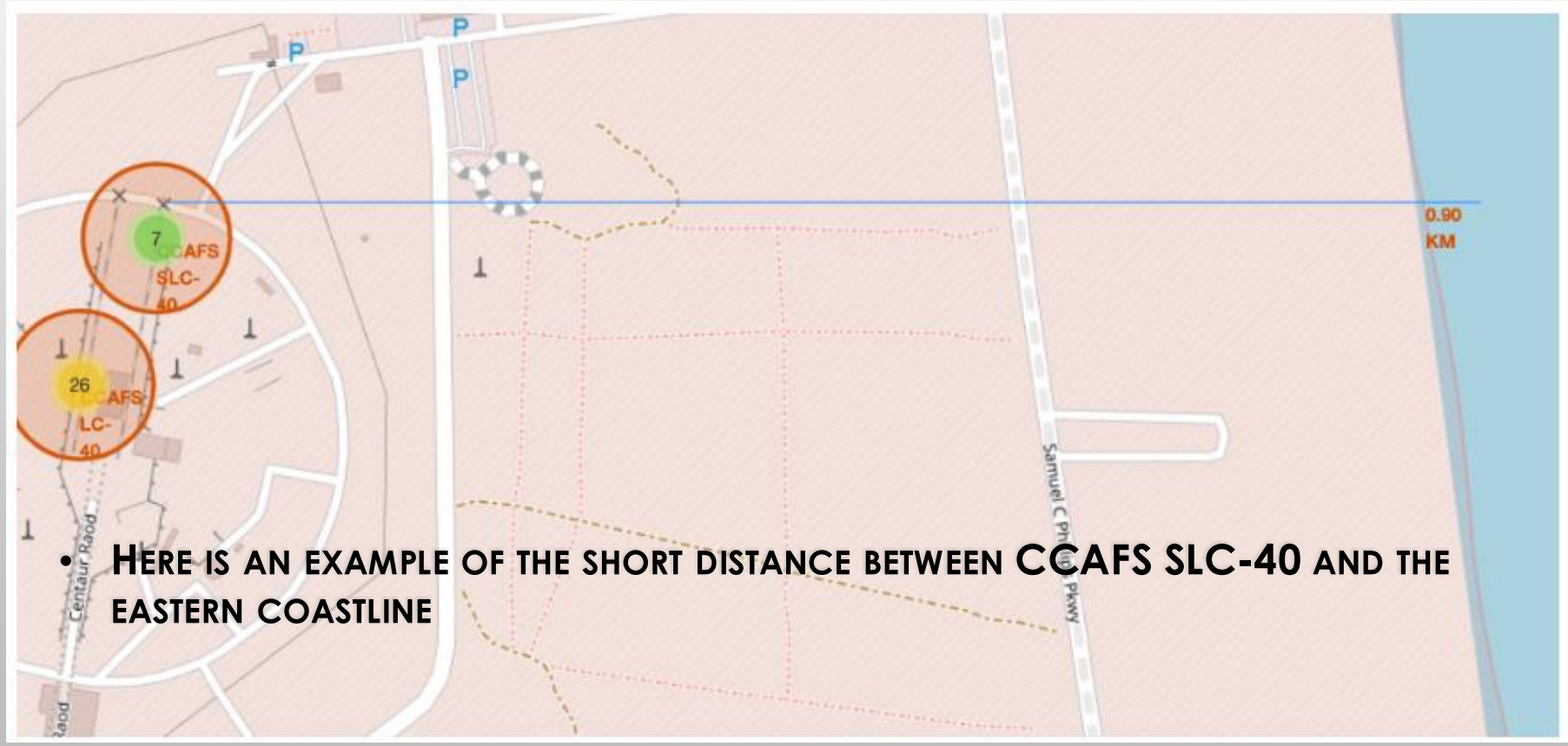
- We can see that all of SpaceX's launch sites are located on coastal American zones, more specifically in California and Florida

MARKERS FOR COLOR LABELED LANDING OUTCOMES

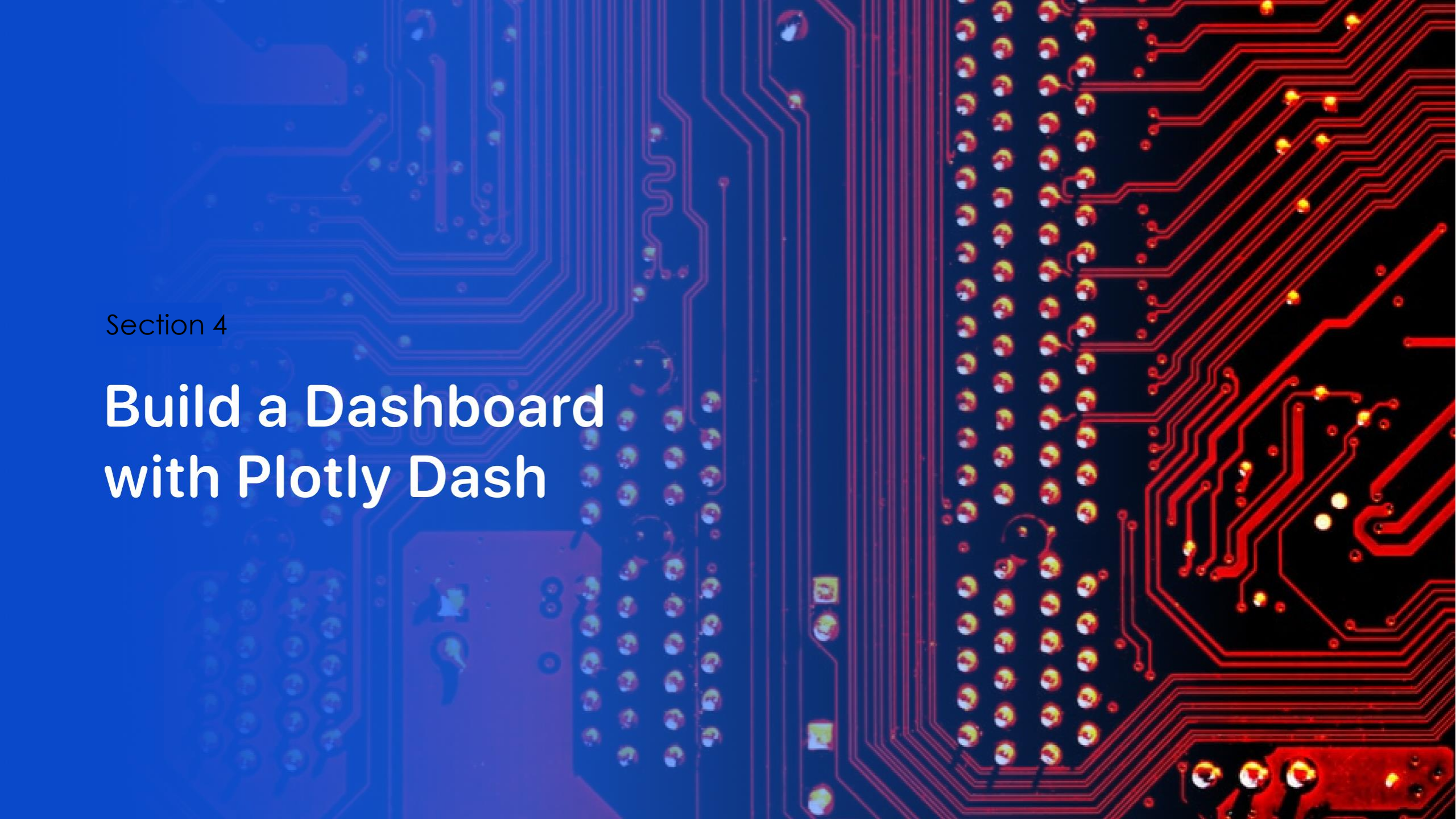
- If we zoom into the sites in Florida, we can appreciate the color coded labels, where green represent successful labels and red unsuccessful



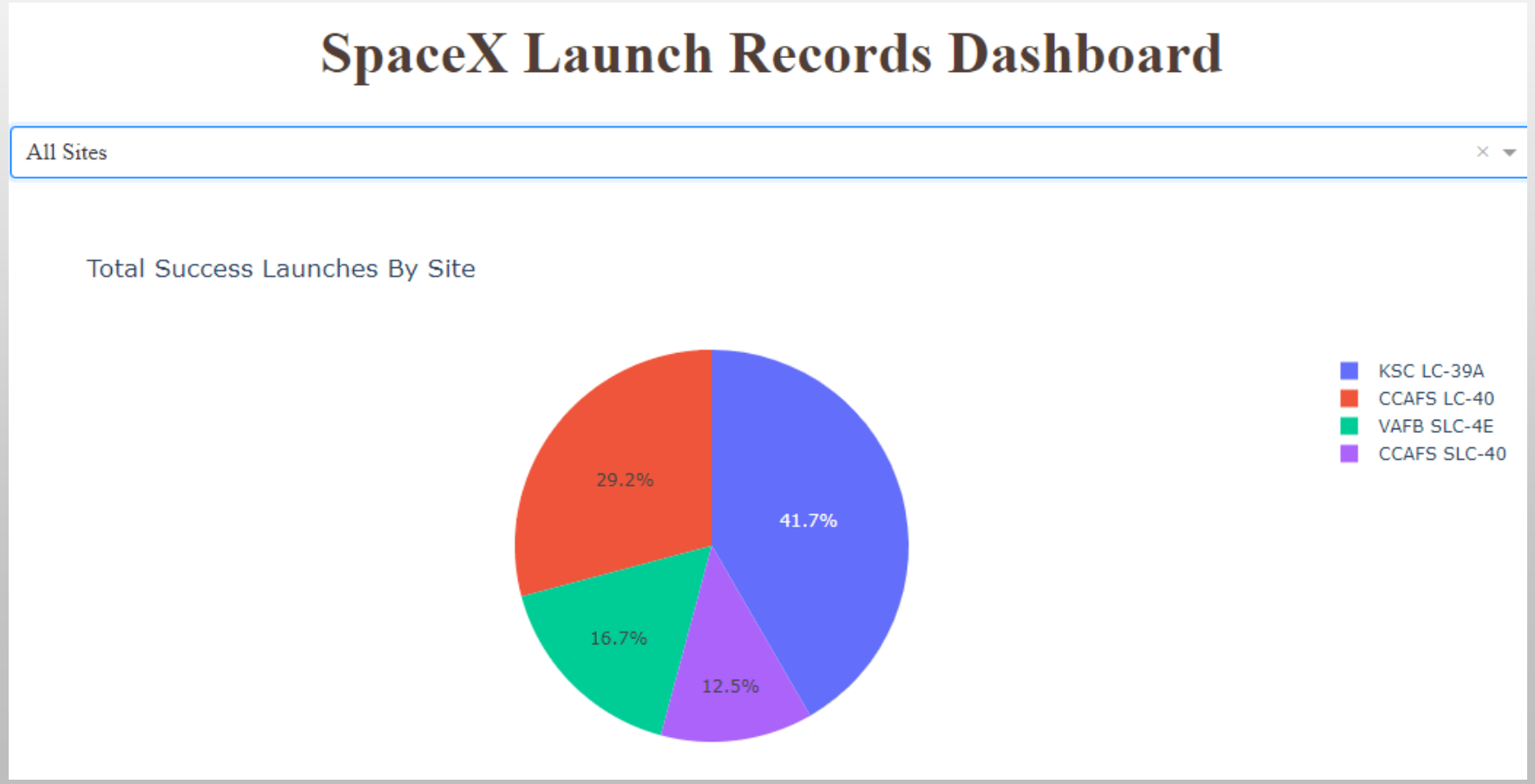
SITES PROXIMITY TO LANDMARKS



- HERE IS AN EXAMPLE OF THE SHORT DISTANCE BETWEEN CCAFS SLC-40 AND THE EASTERN COASTLINE

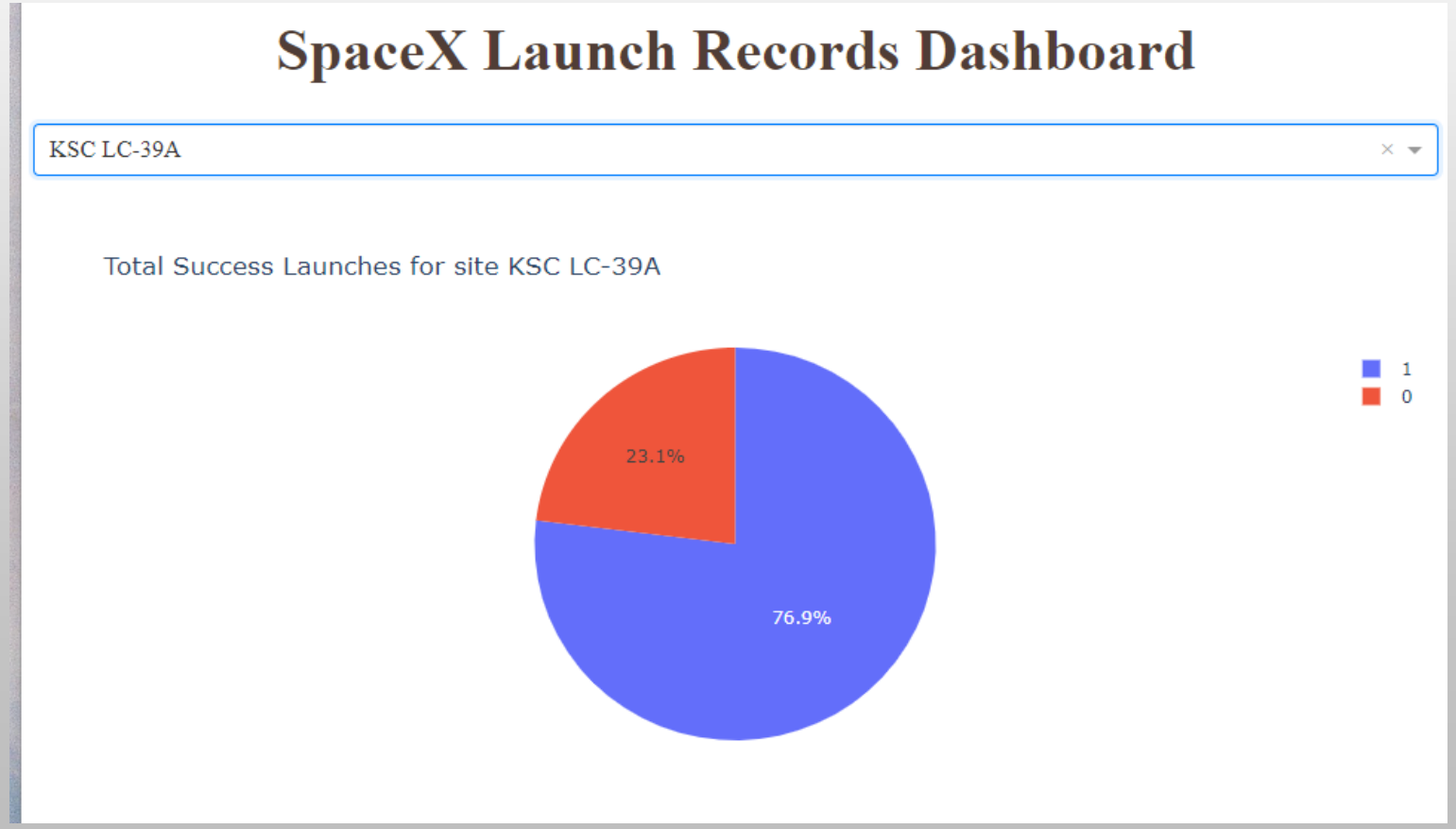

Section 4
Build a Dashboard with Plotly Dash

DASHBOARD – SUCCESS LAUNCHES FOR ALL SITES



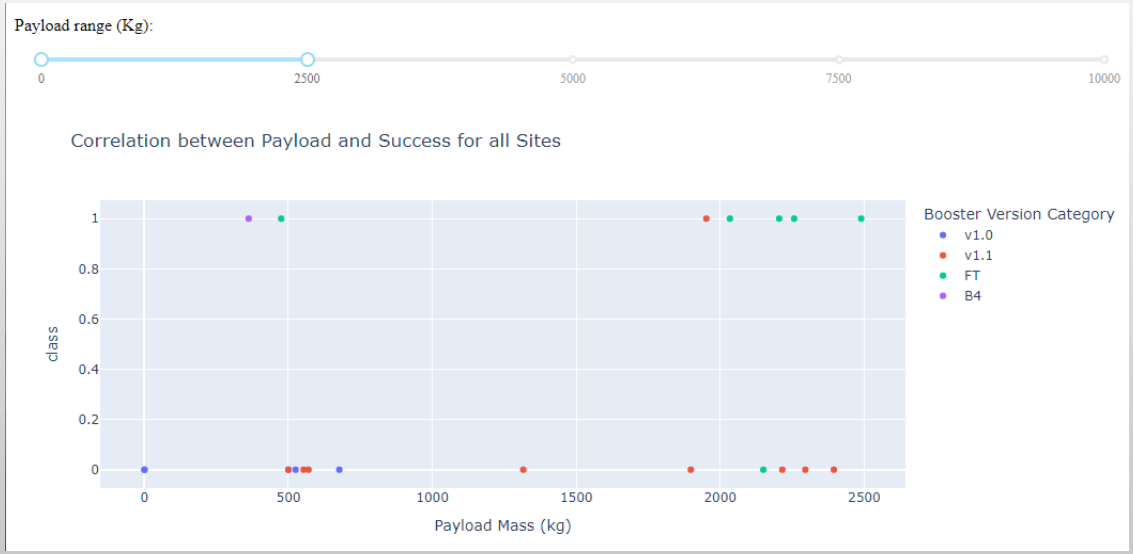
- THE PIE CHART SHOWS THE SUCCESS RATE FOR THE 4 DIFFERENT SITES
- WE CAN SEE THAT KSC LC-39A HAS THE HIGHEST SUCCESS RATE

DASHBOARD – 'KSC LC-39A' SUCCESS RATIO



- THE SITE HAS ACHIEVED A 76.9% SUCCESS RATE

DASHBOARD – SCATTER PLOT OF PAYLOAD VS LAUNCH OUTCOME FOR ALL SITES

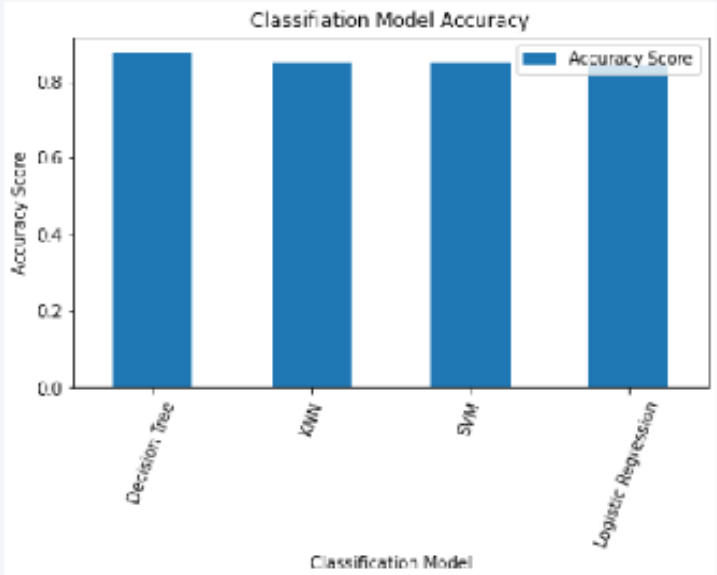


- **LIGHT PAYLOAD FOR ALL SITES**
- **HEAVY PAYLOAD FOR ALL SITES**

Section 5

Predictive Analysis (Classification)

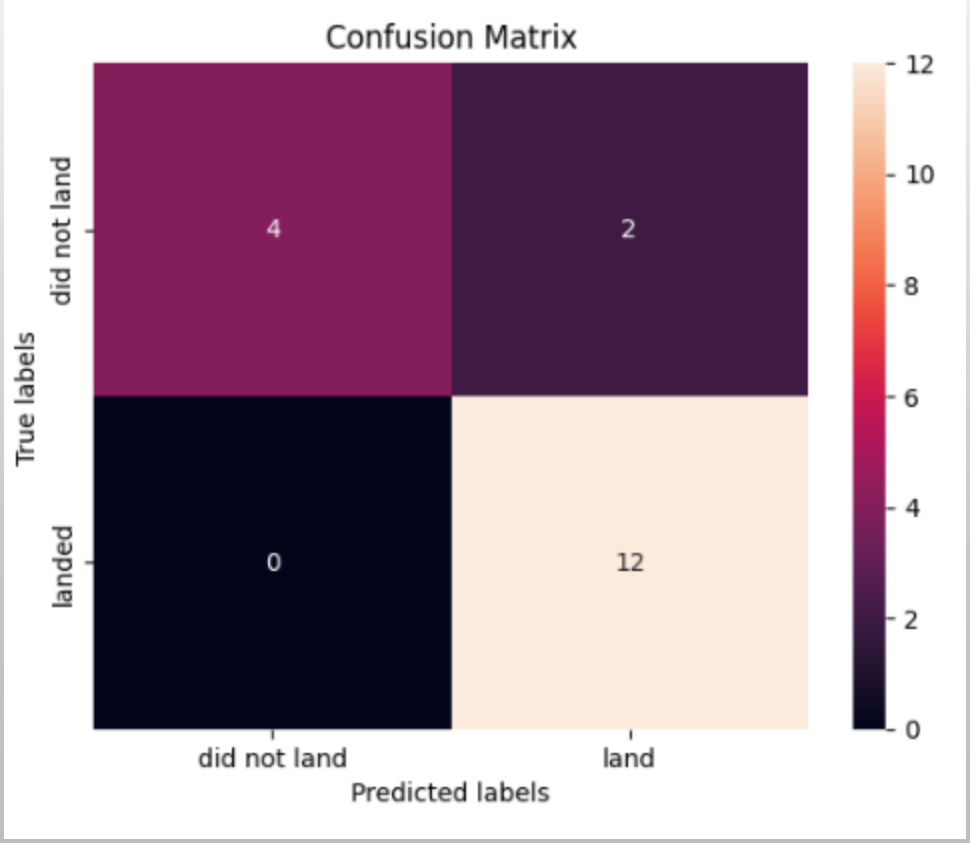
CLASSIFICATION ACCURACY



	Algo Type	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.875000	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

- BASED ON THE ACCURACY SCORE, AND AS SEEN ON THE BAR CHART, WE KNOW THE DECISION TREE IS THE MOST ACCURATE MODEL WITH A SCORE OF **0.8750**

CONFUSION MATRIX



- THE CONFUSION MATRIX FOR THE DECISION TREE MODEL MADE A TOTAL OF 18 PREDICTIONS
- WE HAD 12 TRUE POSITIVES AND 4 TRUE NEGATIVES
- WE ALSO HAD 0 FALSE NEGATIVES AND 2 FALSE POSITIVES
- OVERALL, THE CLASSIFIER WAS CORRECT ABOUT 89% OF THE TIME ((TP + TN)/TOTAL) AND HAD A MISCLASSIFICATION ((FN + FP)/TOTAL) OF 11%

CONCLUSIONS

- As the number of flights increase, so does the likelihood of a successful first stage landing
- Success rate seems to go up as Payload increases, but there is no conclusive correlation between Payload mass and success rate
- Launch success rate increased by about 80% from 2013 to 2020
- Launch site 'KSC LC-39A' has the highest launch success rate
- ES-L1, GEO, HEO and SSO lead the Orbit type success rate
- Launch sites are strategically located away from cities and close to coastlines, railroads and highways, which obviously minimizes the risk of casualties due to a landing failure
- The best performing ML Classification model is the Decision Tree with an accuracy of about 87.5%. Despite the relatively high score, more data may be needed in order to better fine tune predictions

APENDIX

Thank you!