IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Justin Stewart

10/24/2022

Outline

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION

Executive Summary

Summary of methodologies

Data Collection

Data Collection via Web scraping

Data Wrangling

EDA w/ data visualization

EDA w/ SQL

Building an interactive map w/ Folium

Building a Dashboard w/ Plotly Dash

Predictive analysis w/ Classification Models

Summary of all results

EDA results

Interactive analytics

Predictive analysis via classification models

Introduction

- The modern commercial space age is currently geared toward making space travel affordable for everyone. One of, if not the most successful company accomplishing this task is SpaceX. SpaceX advertises launches of their Falcon 9 rocket at a cost of 62 million dollars. Typically, this sort of launch may cost upwards of 165 million dollars, but due to SpaceX's ability to recover and reuse the first stage of their rockets, they are able to minimize costs. Through successful replication of this process, the price of space travel is allowed to decline even further. As a Data Scientist of a startup rivaling SpaceX, the goal of this project is to predict future landing outcomes in order to accurately predict the cost of rocket launches.

- Problems To Explore:

  - What factors contribute the most to a successful rocket landing?

  - What is the relationship between each variable, and how is it affecting the landing outcome?

  - What combination of variables produces the highest success rate?

Section 1

Methodology

Methodology

Executive Summary:

Data Collection Methodology

- SpaceX API
- Web scraping tables

Perform Data Wrangling

- Determine labels to train the classification model
- One Hot Encoding categorical variables
- Data cleaning of null values and non-essential columns

Perform Exploratory Data Analysis (EDA) Using Visualization & SQL

- Exploratory analysis performed utilizing various forms of visualization (Bar, Scatter, Line) to aid in the feature selection/engineering process

Perform Interactive Visual Analytics Using Folium & Plotly Dash

- Create interactive map to determine the correlation between launch sites and successful landing outcomes

Perform Predictive Analysis Using Classification Models

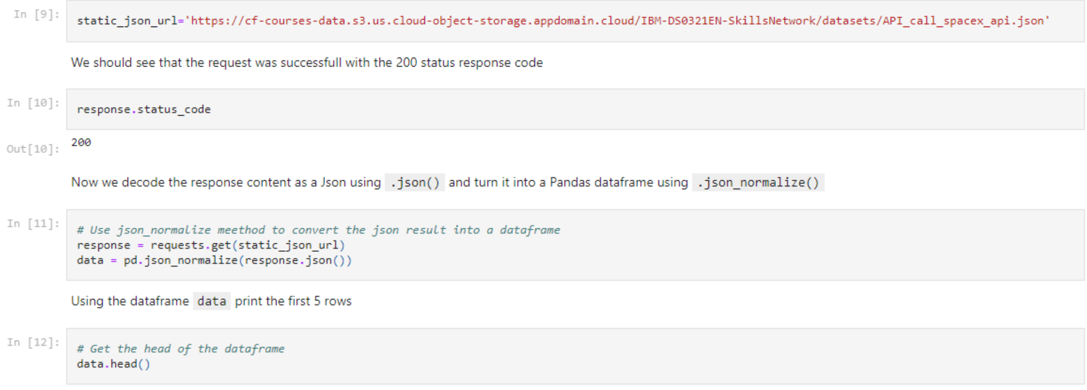
- LR, KNN, SVM, DT models built and evaluated to determine the best classification model

Data Collection

- Methods of Data Collection:
    - SpaceX API requests (Requests, Pandas & SQL)
        - Converted response content to .json using .json()
        - Converted .json to pandas dataframe with .json_normalize()
    - Web scraping data from SpaceX’s Wikipedia (Requests, Pandas & BeautifulSoup4)
        - Launch records HTML table located on the SpaceX Wikipedia scraped via Pandas, BeautifulSoup4 and converted into pandas dataframe for analysis

Data Collection – SpaceX API

- Utilized the requests library to collect data from the SpaceX API
- Data Wrangling / Preparation performed on the requested data.
- Link to Notebook: https://github.com/NotBlasto/Applied-Data-Science-Capstone-Project/blob/main/Data%20Collection.ipynb



```
In [9]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

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

```
In [10]: response.status_code

Out[10]: 200
```

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

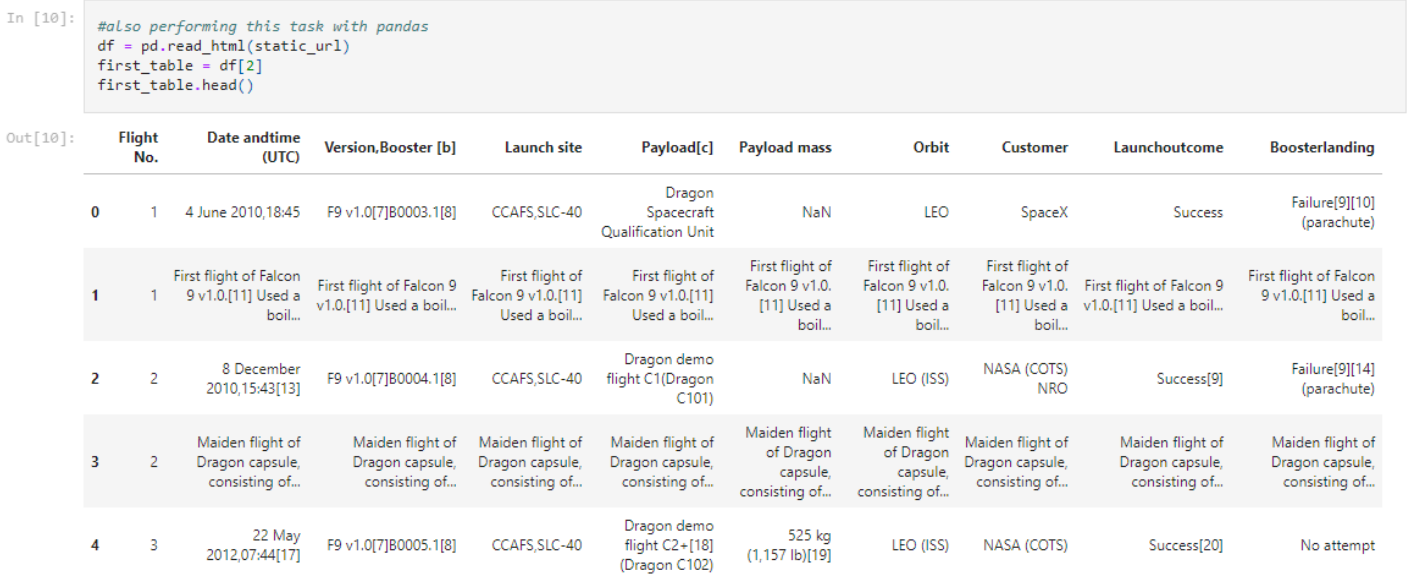
```
In [11]: # Use json_normalize meethod to convert the json result into a dataframe
         response = requests.get(static_json_url)
         data = pd.json_normalize(response.json())
```

Using the dataframe `data` print the first 5 rows

```
In [12]: # Get the head of the dataframe
         data.head()
```

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- Link to Github: https://github.com/NotBlasto/Applied-Data-Science-Capstone-Project/blob/main/Data%20Collection%20W_%20Web%20Scraping.ipynb



In [10]:

```
#also performing this task with pandas
df = pd.read_html(static_url)
first_table = df[2]
first_table.head()
```

Out[10]:

	Flight No.	Date andtime (UTC)	Version,Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launchoutcome	Boosterlanding
0	1	4 June 2010,18:45	F9 v1.0[7]B0003.1[8]	CCAFS,SLC-40	Dragon Spacecraft Qualification Unit	NaN	LEO	SpaceX	Success	Failure[9][10] (parachute)
1	1	First flight of Falcon 9 v1.0.[11] Used a boil...	First flight of Falcon 9 v1.0.[11] Used a boil...	First flight of Falcon 9 v1.0.[11] Used a boil...	First flight of Falcon 9 v1.0.[11] Used a boil...	First flight of Falcon 9 v1.0.[11] Used a boil...	First flight of Falcon 9 v1.0.[11] Used a boil...	First flight of Falcon 9 v1.0.[11] Used a boil...	First flight of Falcon 9 v1.0.[11] Used a boil...	First flight of Falcon 9 v1.0.[11] Used a boil...
2	2	8 December 2010,15:43[13]	F9 v1.0[7]B0004.1[8]	CCAFS,SLC-40	Dragon demo flight C1(Dragon C101)	NaN	LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
3	2	Maiden flight of Dragon capsule, consisting of...	Maiden flight of Dragon capsule, consisting of...	Maiden flight of Dragon capsule, consisting of...	Maiden flight of Dragon capsule, consisting of...	Maiden flight of Dragon capsule, consisting of...	Maiden flight of Dragon capsule, consisting of...	Maiden flight of Dragon capsule, consisting of...	Maiden flight of Dragon capsule, consisting of...	Maiden flight of Dragon capsule, consisting of...
4	3	22 May 2012,07:44[17]	F9 v1.0[7]B0005.1[8]	CCAFS,SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt

Data Wrangling

- Exploratory data analysis performed
  - Determine labels to train the classification model
  - One Hot Encoding categorical variables
  - Data cleaning of null values and non-essential columns

- Link to Notebook:
- https://github.com/NotBlasto/Applied-Data-Science-Capstone-Project/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- Various forms of data visualization were utilized during the exploratory data analysis process:
- Scatter Plots were used to identify the correlation between various independent and dependent variables such as the success of a flight based on its payload mass and orbit
- Bar charts were used to identify the success rate of launches based on each type of orbit
- Link To Notebook: https://github.com/NotBlasto/Applied-Data-Science-Capstone-Project/blob/main/EDA%20W_%20DATA%20VIZ.ipynb

EDA with SQL

- Connected and read .csv file to Sqlite database utilizing SQL magic commands
- Performed exploratory data analysis with SQL utilizing queries to find:
    - The names of unique launch sites
    - Total payload mass carried by boosters launched by NASA (CRS)
    - Average payload mass carried by F9 v1.1 boosters
    - Total amount of successful and unsuccessful launches
    - Identified the booster version and launch site name of failed landing outcomes on drone ships.
- Link To Notebook: https://github.com/NotBlasto/Applied-Data-Science-Capstone-Project/blob/main/Spacex%20SQL%20Magic%20EDA.ipynb

Build an Interactive Map with Folium

- Utilizing Folium to build an interactive map we:
  - Created a map and marked all launch sites
  - Added map objects such as markers, circles, and lines to provide additional information such as site distance from geographic landmarks
  - Utilized color-labeled marker clusters to identify launch sites with high success rates
- Link To Notebook: https://github.com/NotBlasto/Applied-Data-Science-Capstone-Project/blob/main/Spacex%20Launch%20Site%20Locations%20Map%20W_%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Constructed an interactive dashboard with Plotly Dash that allows the user to view the data from their desired perspective
- Plotted pie charts showing the total successful launches by all sites, as well as the ability to filter to specific sites only.
- Plotted scatter plot to present the user with detailed information about the relationship between landing outcomes and payload mass for different booster versions.
- Link To Notebook: https://github.com/NotBlasto/Applied-Data-Science-Capstone-Project/blob/main/spacex_dash_app_final.py

Predictive Analysis (Classification)

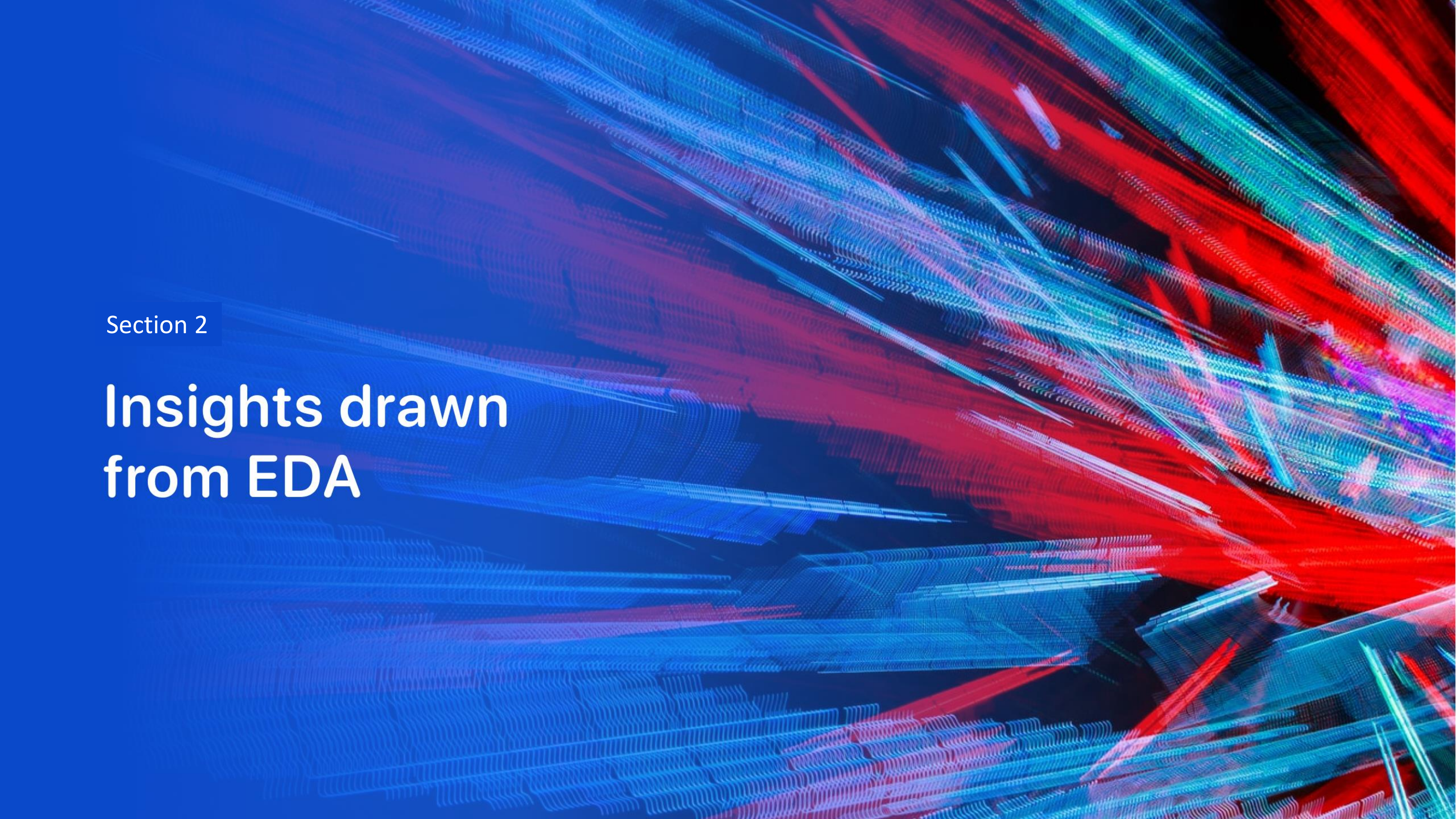
- In order to accurately predict launch outcomes we:
  - Loaded data utilizing numpy and pandas into dataframes
  - Split our data into training and testing sets (80/20 split)
  - Built four separate classification machine learning models tuned with various hyperparameters through grid search cross-validation
  - Evaluated each model by its accuracy, best accuracy, and area under the curve to determine the best model.
- Link To Notebook: https://github.com/NotBlasto/Applied-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

Results

EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS

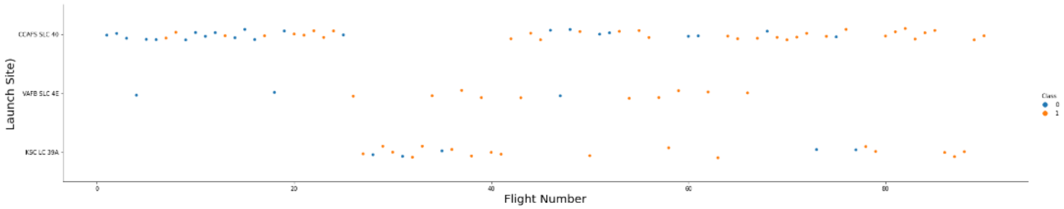


Section 2

Insights drawn from EDA

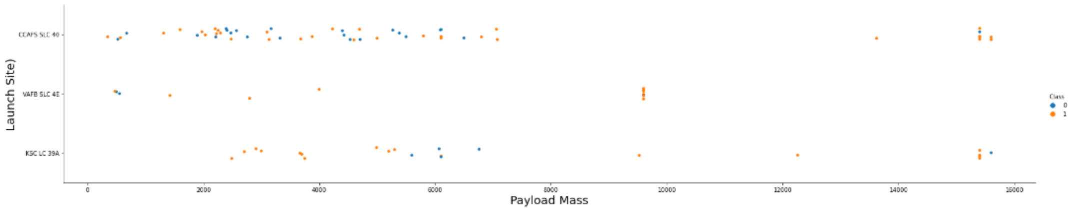
Flight Number vs. Launch Site

- From the above scatter plot, we can see that the amount of flights, or Flight Number, is positively correlated with a successful launch at each launch site. I.E. the more flights at a launch site, the greater the success rate at that launch site.



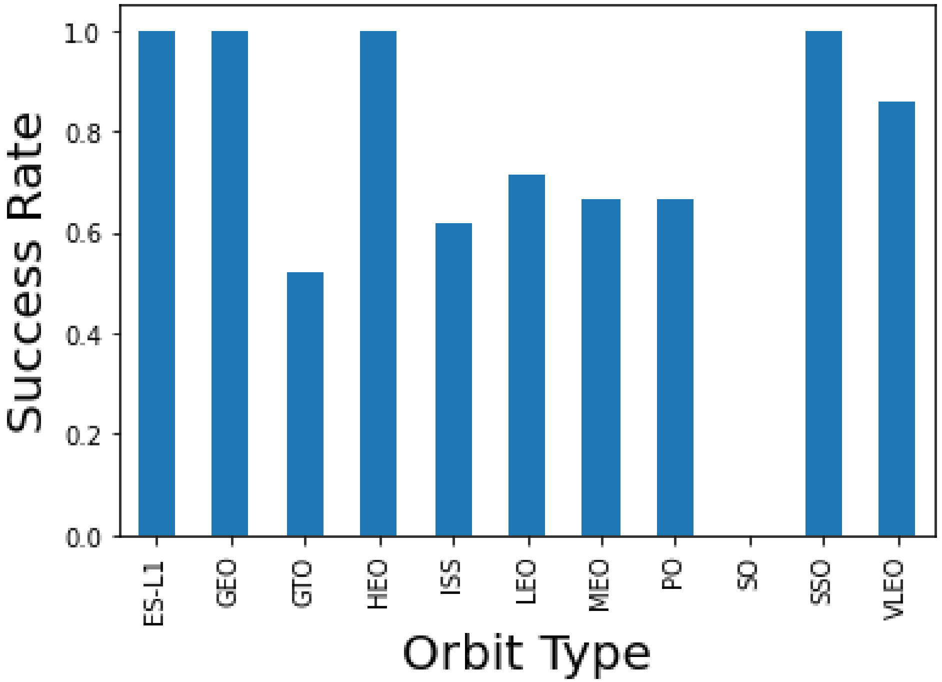
Payload vs. Launch Site

- From the scatter plot above, we gain a few insights:
    - The greater the payload mass the less likely it is that the first stage will return successfully.
    - Launch site VAFB-SLC has no launches with a payload mass greater than 10000 kg.



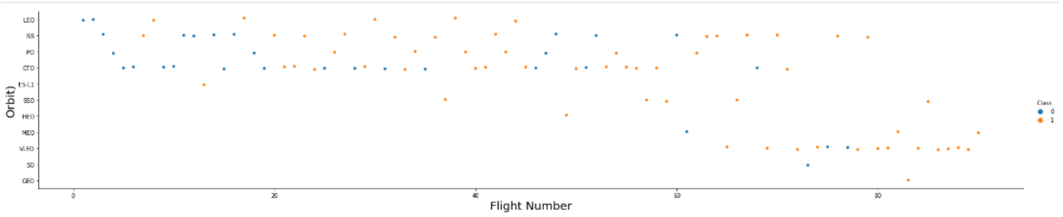
Success Rate vs. Orbit Type

- From the provided bar plot, we can see that launches dedicated to the orbits ES-L1, GEO, HEO, and SSO have the highest success rate.



Flight Number vs. Orbit Type

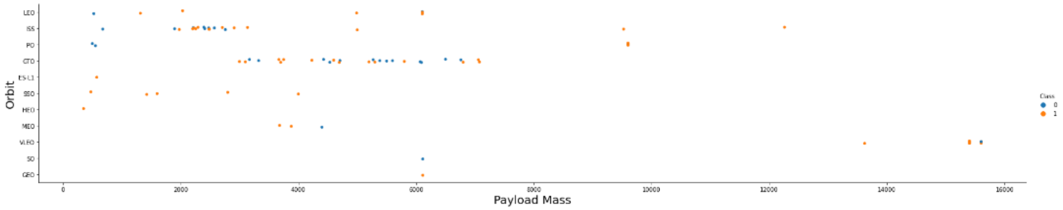
- By creating a scatterplot between the number of flights (Flight Number) and Orbit type two interesting relationship insights emerge:
    - In the LEO orbit, successful landing outcomes appear to be strongly correlated to the number of flights.
    - In the GTO orbit, the number of flights seem to have no correlation to success landing outcomes.



Payload vs. Orbit Type

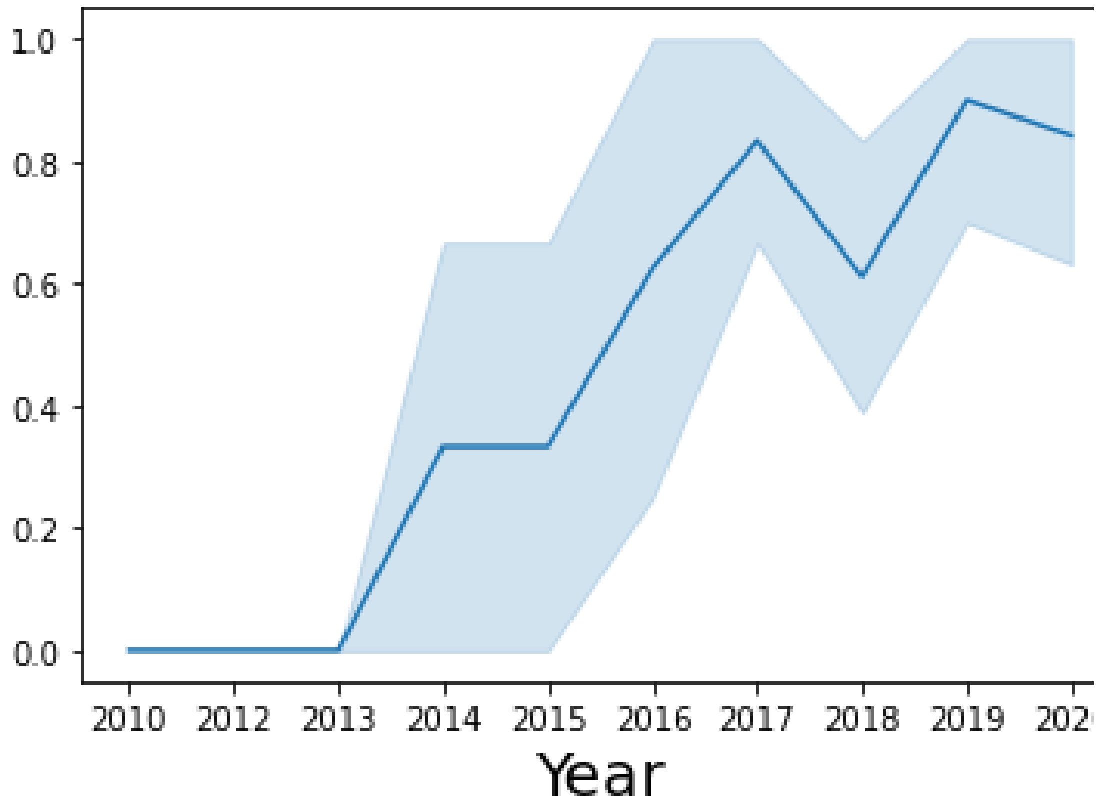
- From the above scatter plot we can observe that heavy payloads have a much higher rate of success in the PO, LEO, and ISS orbits.

- For the GTO orbit the data is less conclusive, as there are many failures and successes in the same payload ranges.



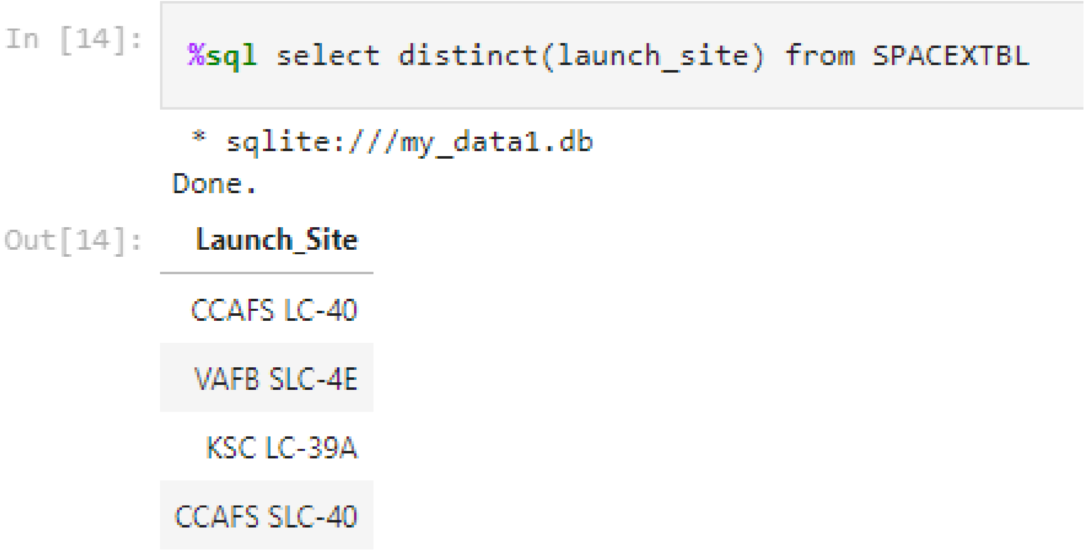
Launch Success Yearly Trend

- The line plot provided shows us that there is a strong positive correlation between years and success rate, meaning as time increases, so does our rate of success.
- We also see that success rate did not begin increasing until 2013.



All Launch Site Names

- Utilizing SQL’s **DISTINCT** keyword, we are able to show each unique launch site available in the SpaceX dataset.



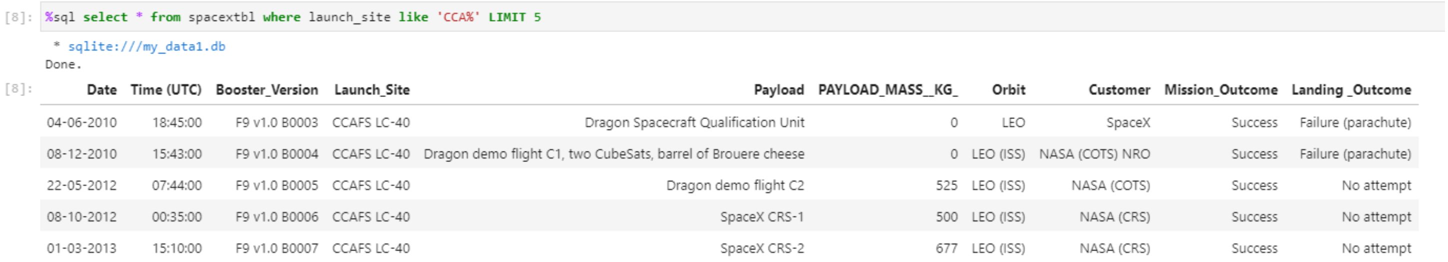
```
In [14]:  %sql select distinct(launch_site) from SPACEXTBL

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

Out[14]:

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

Launch Site Names Begin with 'CCA'

- We used the following query to display 5 records in which launch sites began with ‘CCA’



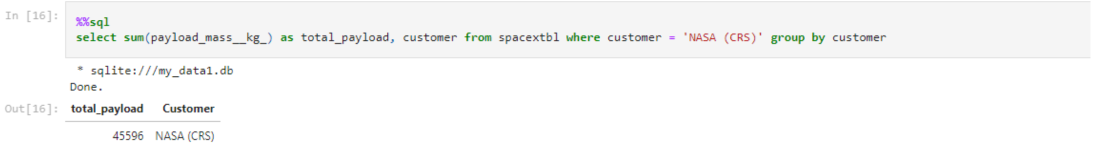
```
[8]: %sql select * from spacextbl where launch_site like 'CCA%' LIMIT 5
 * sqlite:///my_data1.db
Done.
```

[8]:

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

- The query above shows us that the total payload mass carried by boosters from NASA (CRS) is 45,596 kg through the usage of the **SUM()** SQL function.



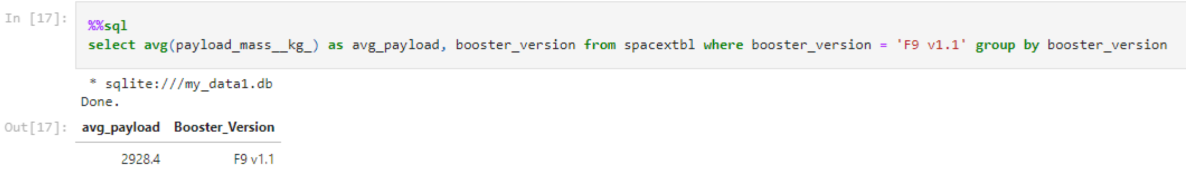
```
In [16]:  %%sql
          select sum(payload_mass__kg_) as total_payload, customer from spacextbl where customer = 'NASA (CRS)' group by customer

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

Out[16]:

total_payload	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- The query below utilizes the **AVG()** SQL function to display the average payload weight carried by F9 v1.1 boosters in kilograms.
- The average payload carried by F9 v1.1 boosters is 2928.4 kg.



```
In [17]: %%sql
select avg(payload_mass__kg_) as avg_payload, booster_version from spacextbl where booster_version = 'F9 v1.1' group by booster_version
```

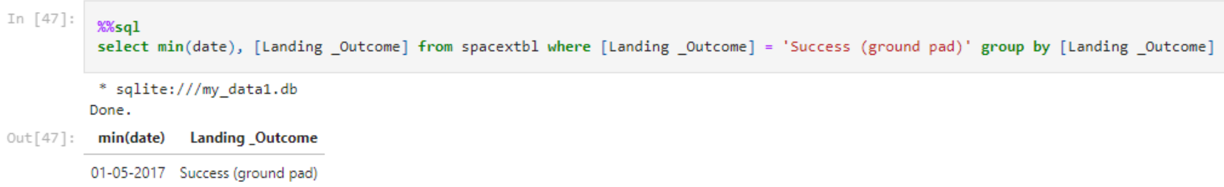
```
 * sqlite:///my_data1.db
Done.
```

Out[17]:

avg_payload	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

- Utilizing the **MIN()** function we observe that the first successful ground landing was on May 1st, 2017.



In [47]:

```
%%sql
select min(date), [Landing _Outcome] from spacextbl where [Landing _Outcome] = 'Success (ground pad)' group by [Landing _Outcome]
```

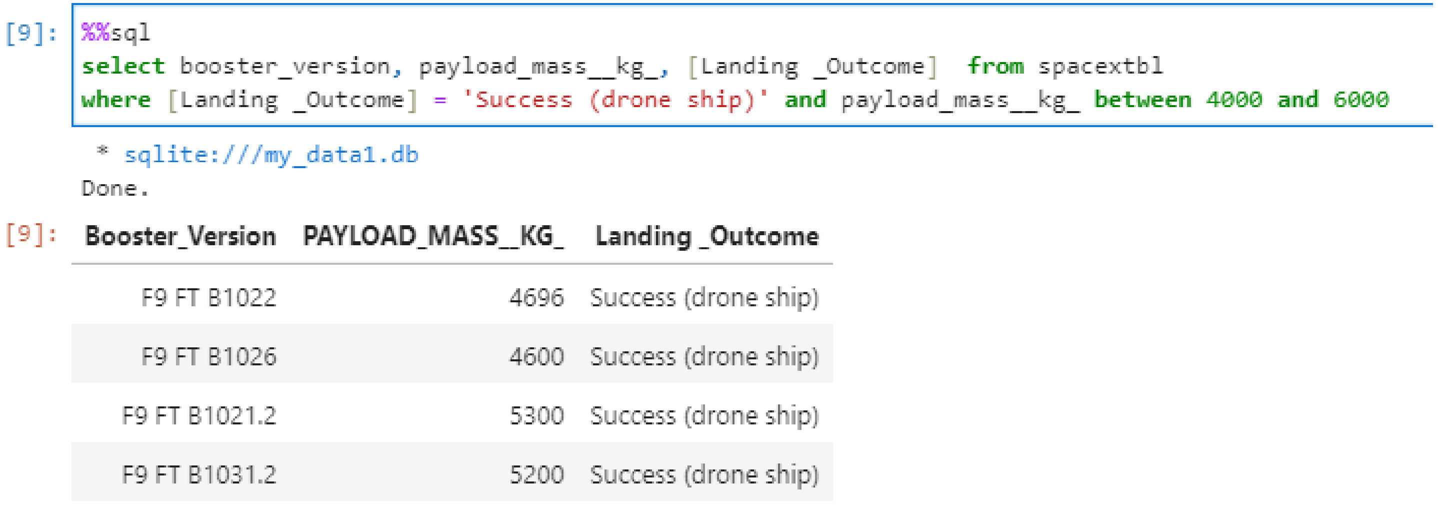
```
 * sqlite:///my_data1.db
Done.
```

Out[47]:

min(date)	Landing _Outcome
01-05-2017	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The **WHERE**, **AND**, and **BETWEEN** clauses were utilized to filter the data for successful drone ship landings with a payload mass between 4000 and 6000 kg.



[9]:

```
%%sql
select booster_version, payload_mass__kg_, [Landing _Outcome]  from spacextbl
where [Landing _Outcome] = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

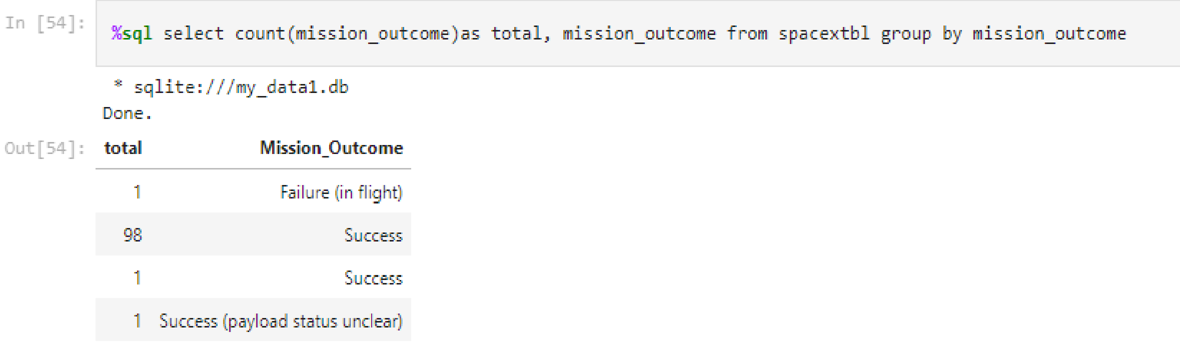
```
 * sqlite:///my_data1.db
Done.
```

[9]:

Booster_Version	PAYLOAD_MASS__KG_	Landing _Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Through the usage of the COUNT() SQL function, we observe that there have been a total of 100 successful mission outcomes, and 1 failure that occurred in flight.
- We are, however, unsure of the payload status in one of these successes.



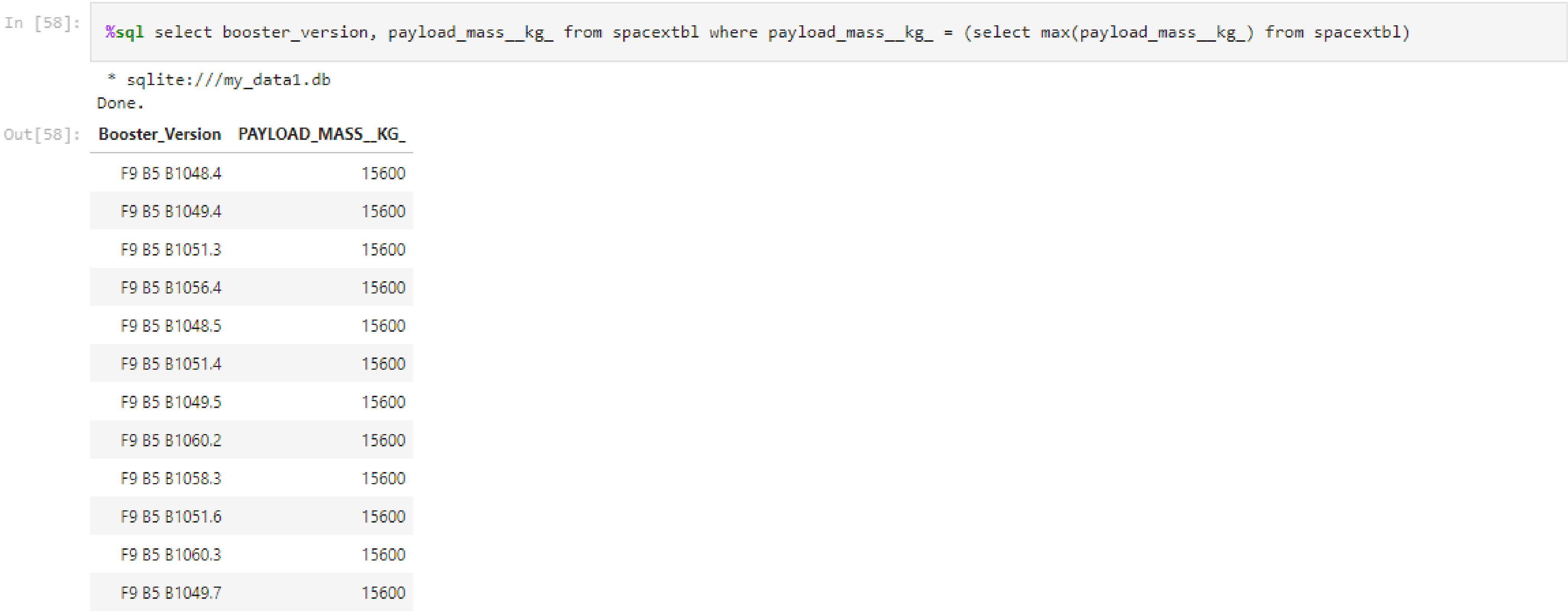
```
In [54]: %sql select count(mission_outcome)as total, mission_outcome from spacextbl group by mission_outcome

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

Out[54]:

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

Boosters Carried Maximum Payload

- By utilizing a subquery to select the maximum payload value for each booster version, we observe that all boosters carrying the maximum payload are F9 B5 boosters.



```
In [58]: %sql select booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
```

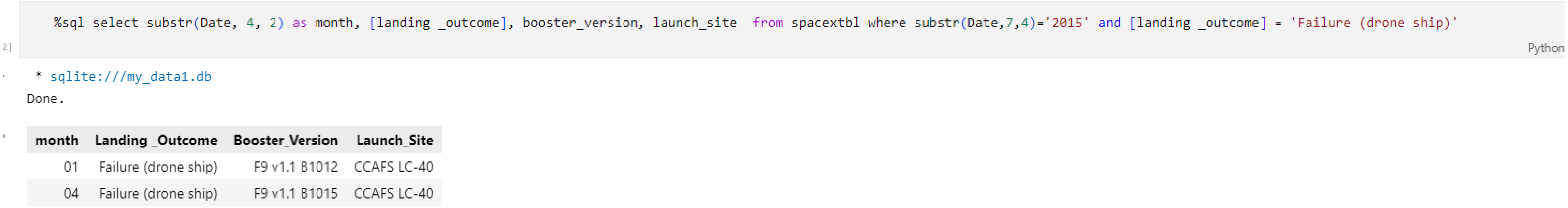
```
 * sqlite:///my_data1.db
Done.
```

Out[58]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- By using the substr() SQL function in the query below, we are able to retrieve the months in which drone ship landing failures occurred in 2015.
- From this query we observe that both of these failures occurred at the same launch site.



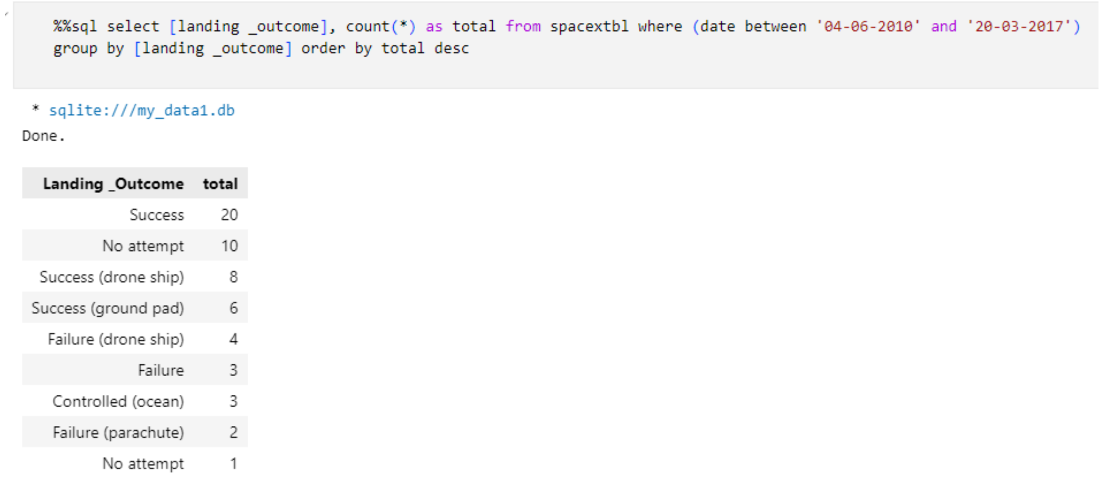
```
%sql select substr(Date, 4, 2) as month, [landing _outcome], booster_version, launch_site  from spacextbl where substr(Date,7,4)='2015' and [landing _outcome] = 'Failure (drone ship)'
```

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

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

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

- Through the combined usage of SQL’s COUNT and BETWEEN keywords, the query produces a result that ranks landing outcomes between June 4th, 2010, and March 20th, 2017.



```
%%sql select [landing _outcome], count(*) as total from spacextbl where (date between '04-06-2010' and '20-03-2017')
group by [landing _outcome] order by total desc
```

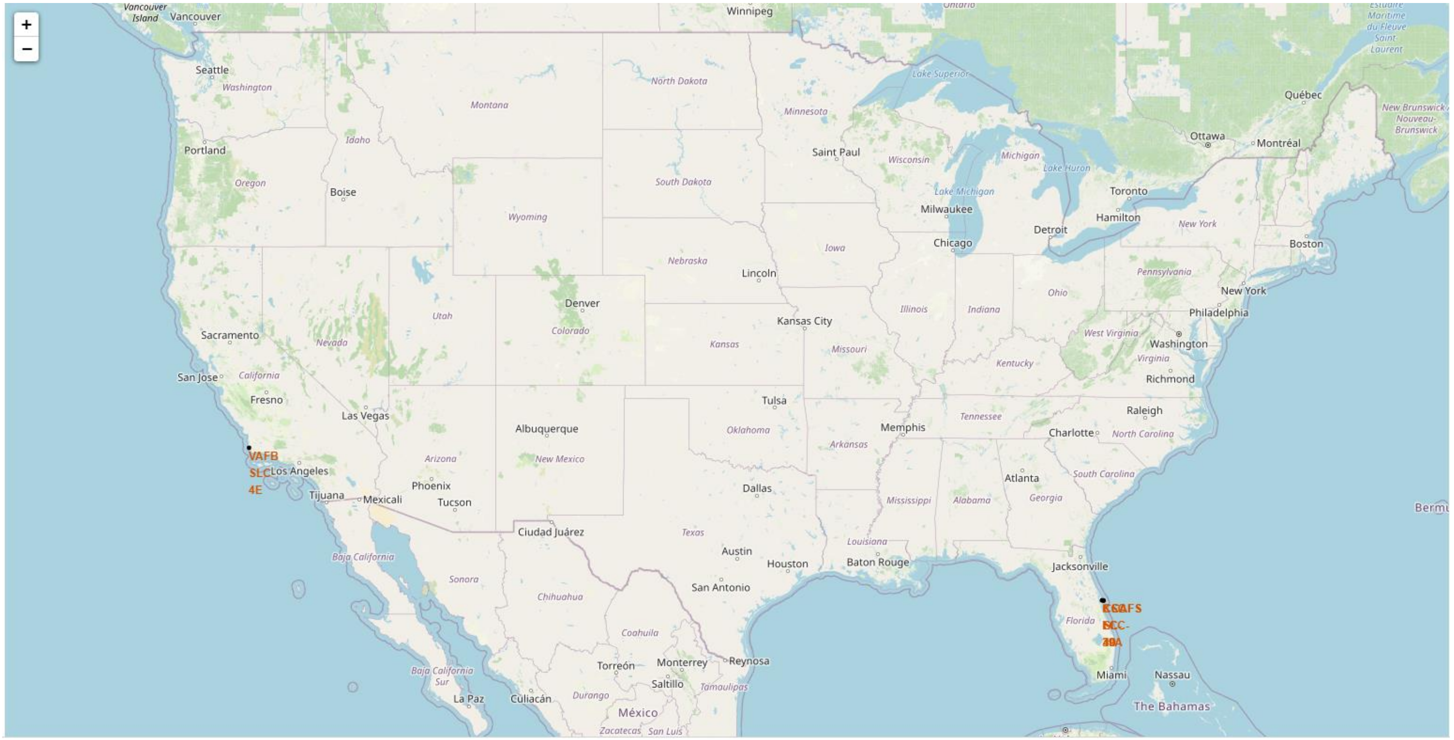
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	total
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

All Launch Sites With Interactive Global Map Markers

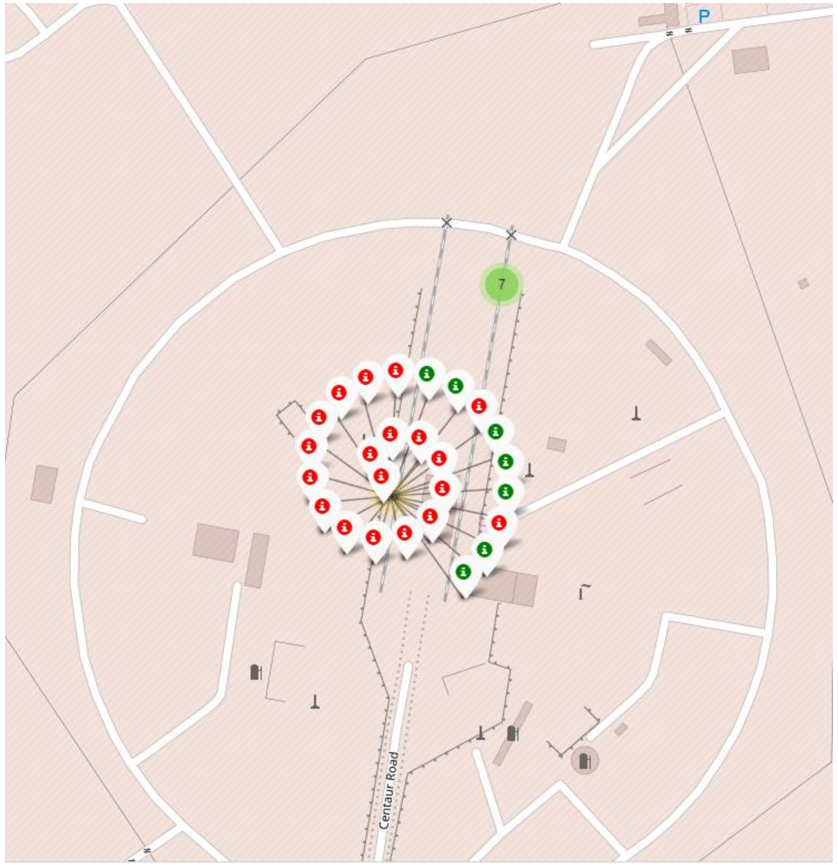


- In the above screenshot we see that each launch site (many are overlapping due to proximity) are located on the southern coastal regions of the United States.

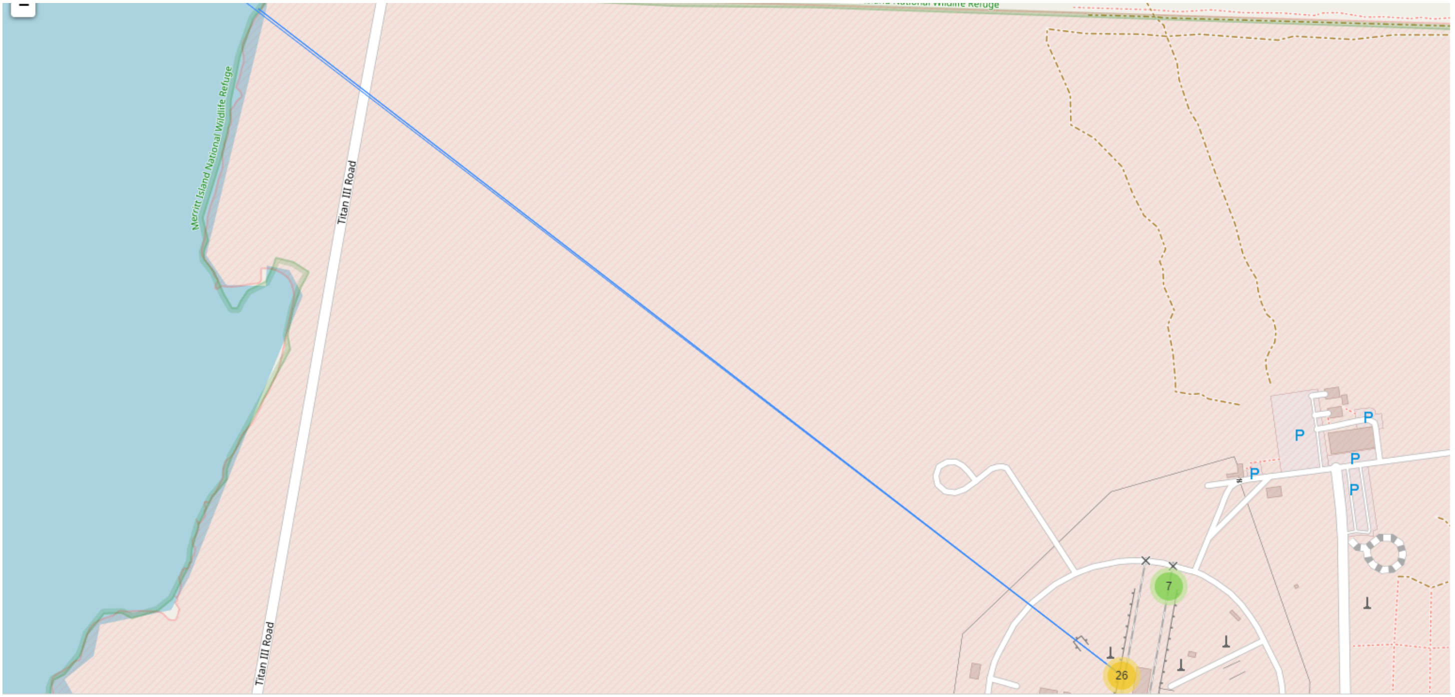
Launch Sites In Close Proximity

By utilizing marker clusters, we are able to identify the exact location of launch sites that otherwise would be too close in proximity to identify.

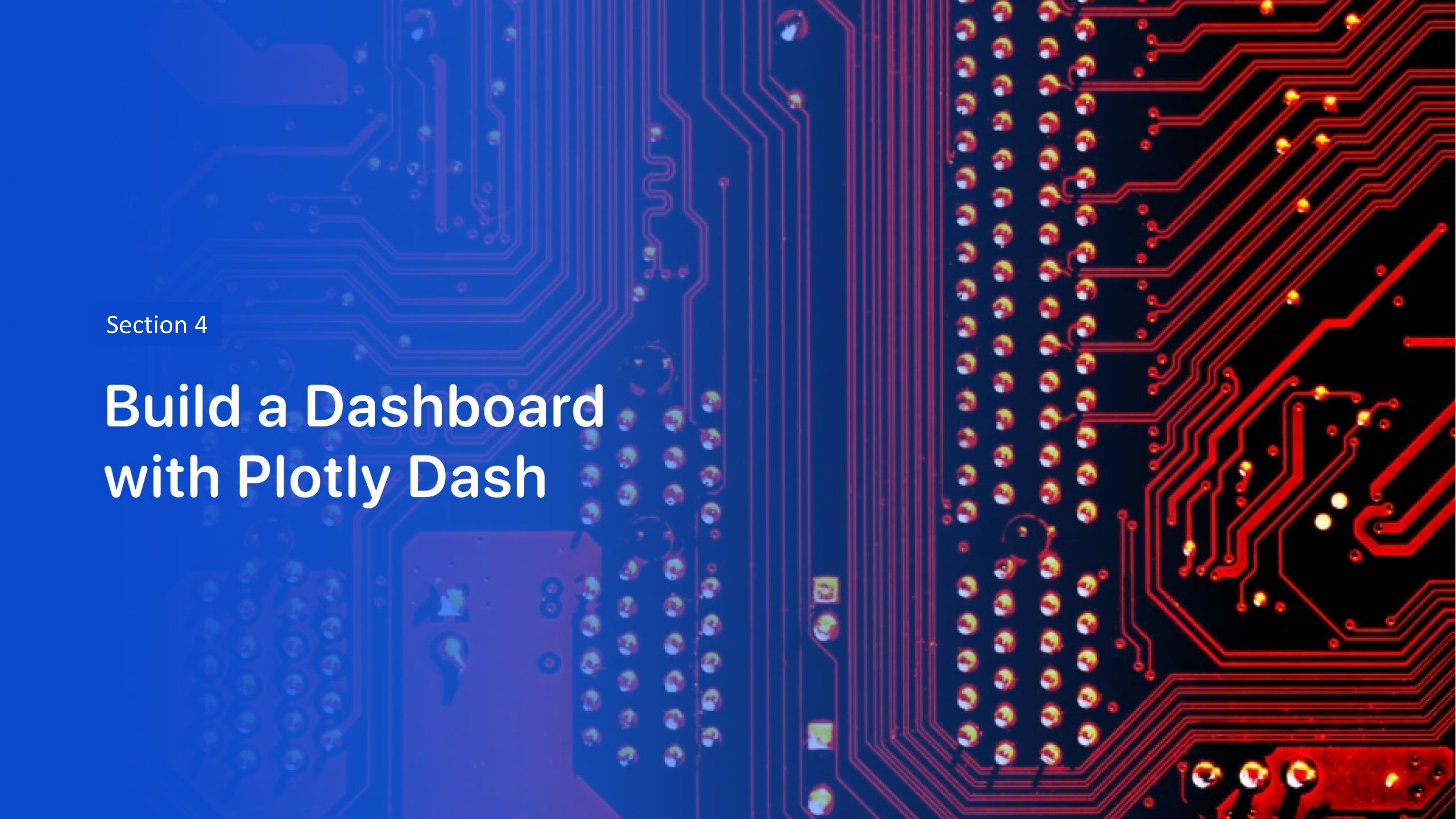
Through the combined usage of labels and colors we are able to identify which launches were successful as well as the site in which launches occurred.



Geographic Significance of Launch Sites



By adding lines to our map, we can see the distance of launch sites to geographic features such as water, railroads, and highways. It appears launch sites have a strong preference to keep away from cities, and near water sources.

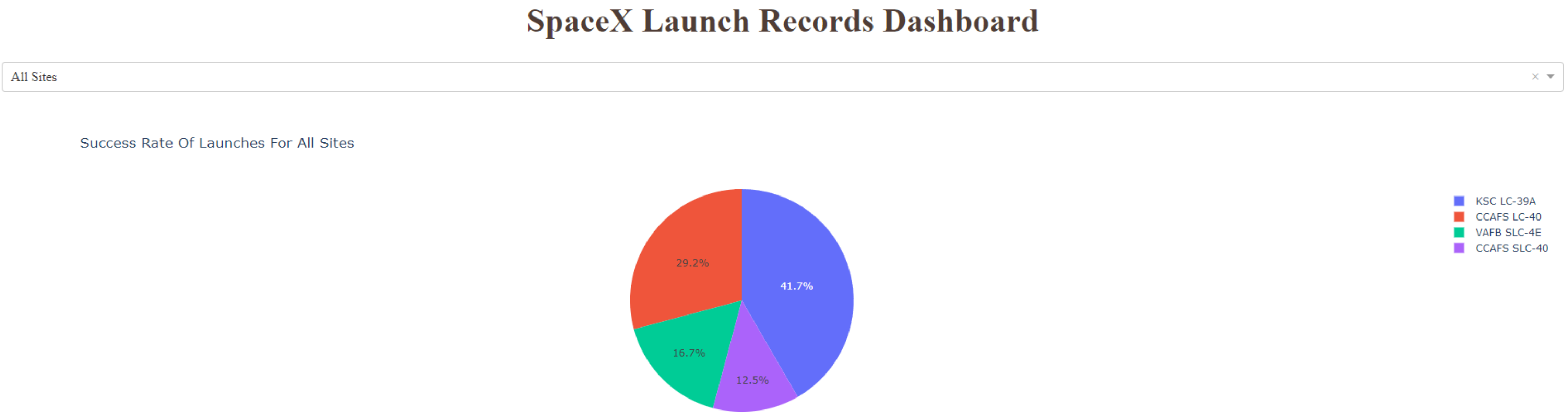


Section 4

Build a Dashboard with Plotly Dash

Pie Chart: Success Rate Of All Launch Sites

- We observe that the KSC LC-39A launch site had the most successful launch rate, making up 41.7% of all successful launches.
- We also observe that the overall success rate of launches is below 50%, meaning not even half of the total launches across all sites successfully land to be recovered.



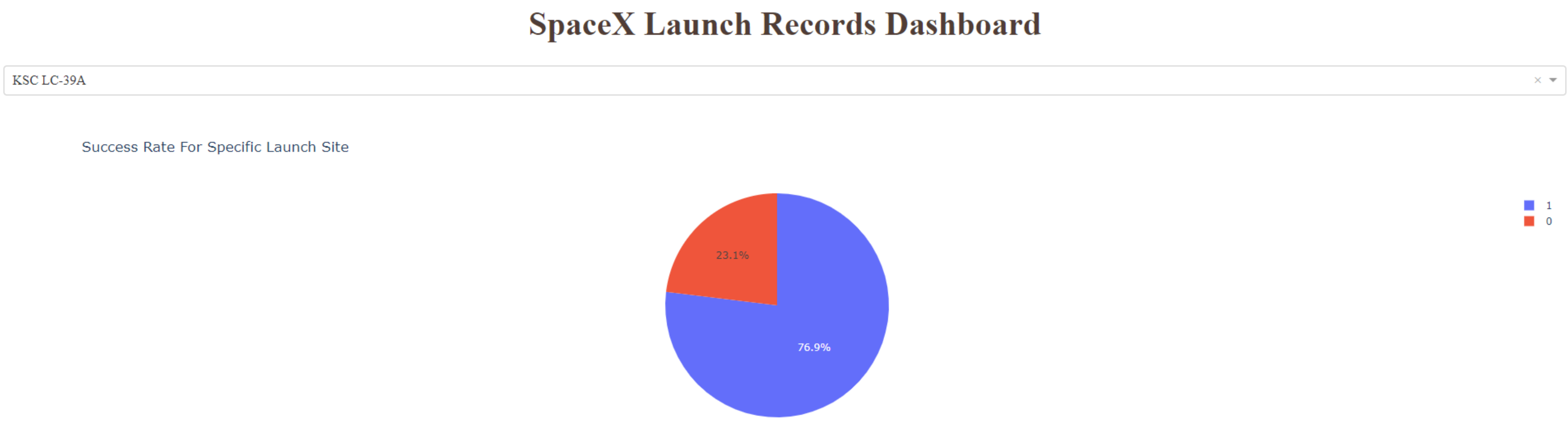
SpaceX Launch Records Dashboard

All Sites

Success Rate Of Launches For All Sites

Pie Chart: Most Successful Launch Site

The most successful launch site, KSC LC-39A has a success rate of 76.9% and a failure rate of 23.1%.



SpaceX Launch Records Dashboard

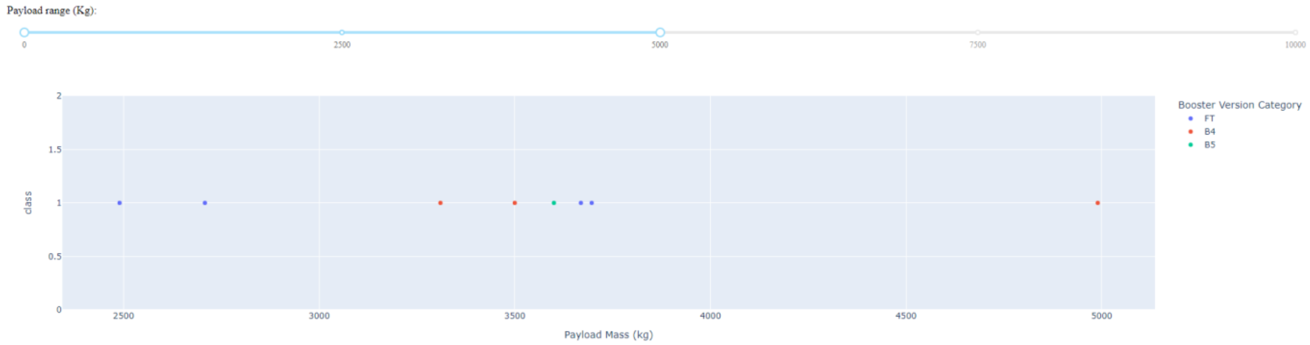
KSC LC-39A

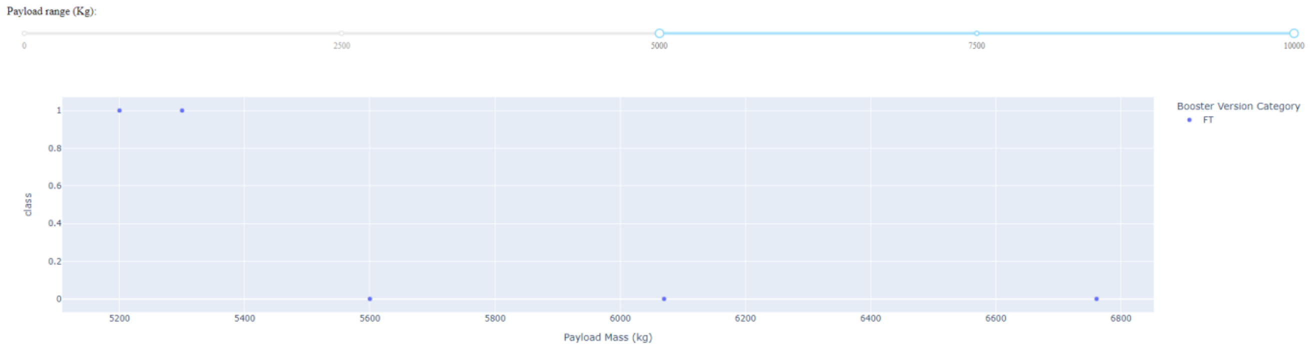
Success Rate For Specific Launch Site

Scatter Plot: Success By Payload Mass

From the interactive scatter plots presented we observe:

1. FT boosters are the only booster type used for payloads with a mass greater than 5000 kg.
2. It is not until payload mass reaches 5600kg that we see our first failure.
3. Launches across all booster types are most successful with payloads ranging from ~2400 – 3800 kg.





Section 5

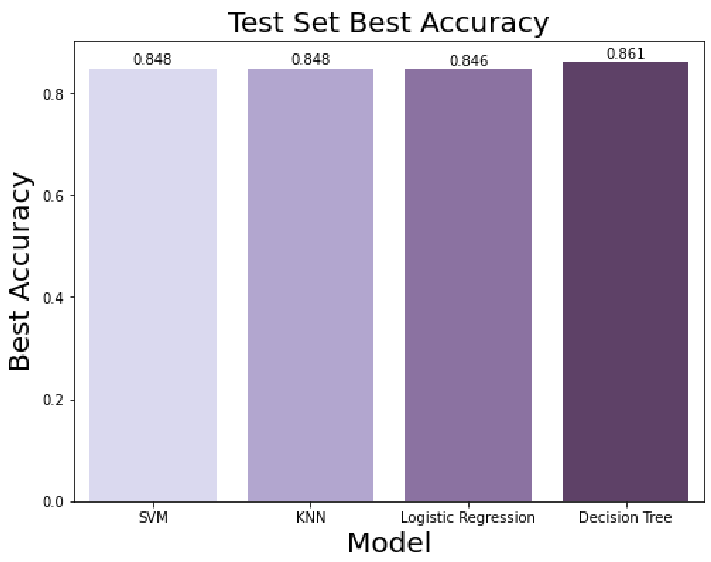
Predictive Analysis (Classification)

Classification Accuracy

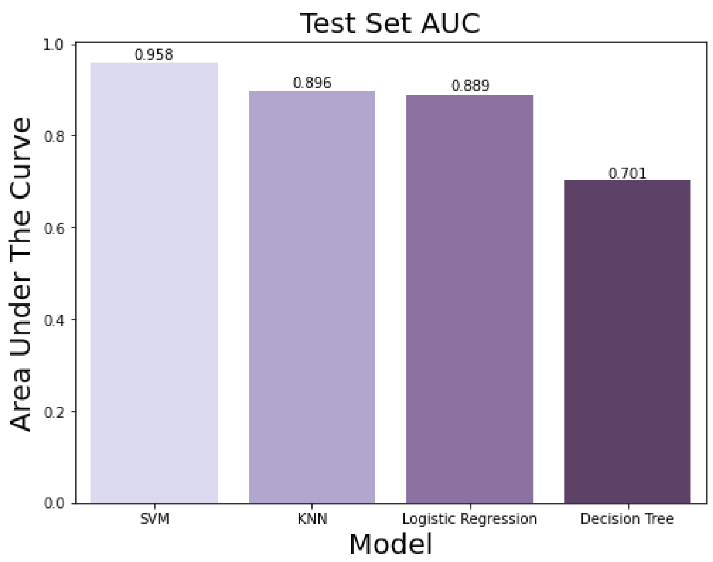
All four models have been evaluated by three metrics:

- Best Accuracy
- Accuracy (actual prediction)
- Area Under The Curve (AUC)

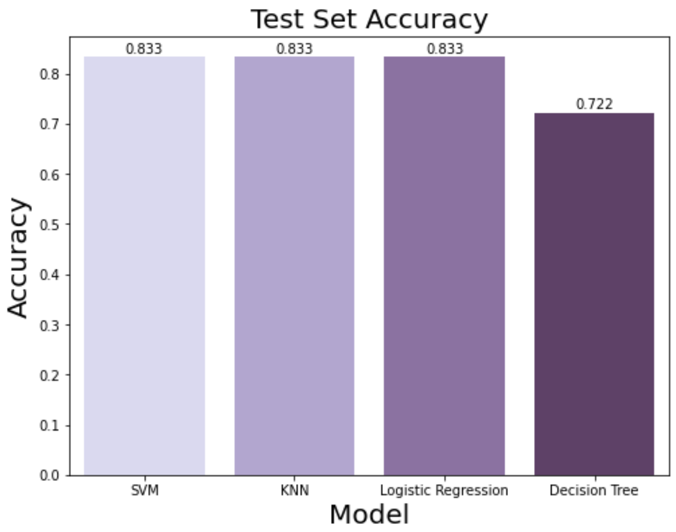
The Support Vector Machine, or SVM, performs the best as it has the highest values among all three metrics.



Test Set Best Accuracy



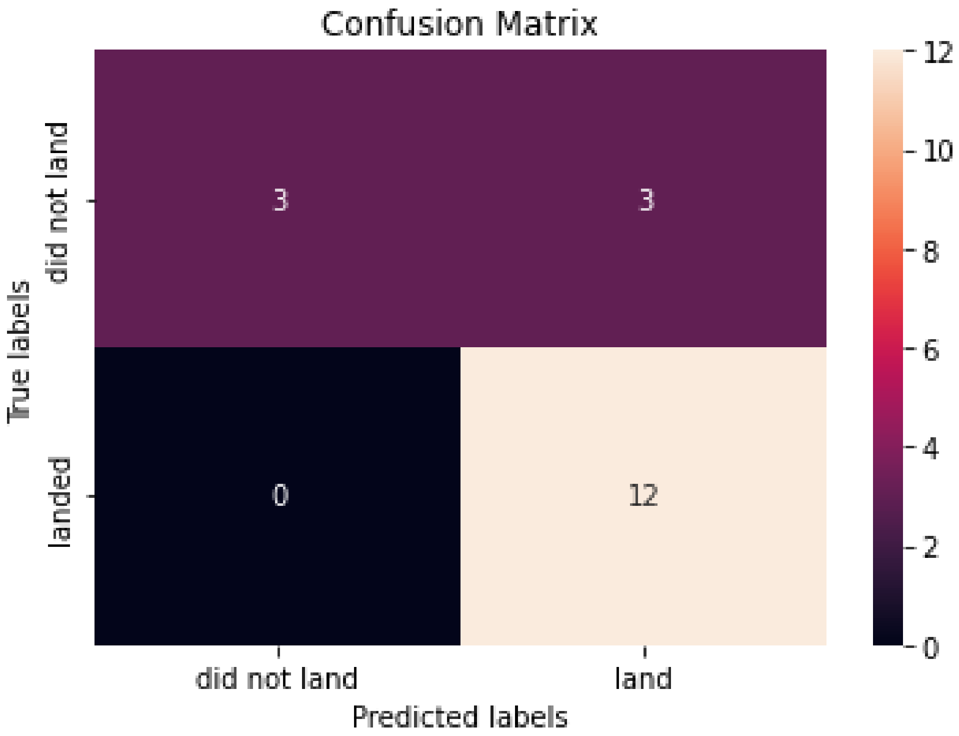
Test Set AUC



Test Set Accuracy

Confusion Matrix

- The confusion matrix for the Support Vector Machine (SVM) classification model displays how many values were accurately predicted by the model.
- In essence, the SVM correctly predicted fifteen out of eighteen outcomes.



Conclusions

- Launch sites have a strong preference to keep away from cities, and near water sources.
- The most successful launch site is KSC LC-39A at a success rate of 76.9% and a failure rate of 23.1%.
- Launches dedicated to the orbits ES-L1, GEO, HEO, and SSO have the highest success rate.
- The more launches, and the more time that passes, the higher the success rate.
- Launch success rates increased from 2013 onwards, but not from 2010 – 2013.
- The greater the payload mass the less likely it is that the first stage will return successfully.
- Launches across all booster types are most successful with payloads ranging from ~2400 – 3800 kg.
- FT boosters are the only booster type used for payloads with a mass greater than 5000 kg
- The Support Vector Machine, or SVM, classification model is the best model for predicting landing outcomes.

Thank you!