IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Miguel Angel Garcia
11.03.2024

Outline

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

Executive Summary

Data collection
Data Wrangling
Exploratory Data Analysis with Data Visualization
Exploratory Data Analysis with SQL
Building an interactive map with Folium
Building a Dashboard with Plotly Dash
Predictive analysis
Visual comparison
SQL Data extraction
Folium interactive map insights
Prediction model comparison.

Introduction

- SpaceX is a private aerospace manufacturer and space transportation company aimed at reducing space transportation costs to enable the colonization of Mars. The Falcon 9 is advertised as a revolutionary instrument since its debut in 2010 achieving significant milestones for humanity.
- Questions to answer:
- What is the relationship between the variables along with each one of the launches?
- Numerous times the company has launched their missions but, how successful have they been?
- How can we predict the outcome of a launch by looking at the external features?

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Data was collected from many sources
- 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

- The data was collected using the SpaceX API from the URL https://api.spacexdata.com/v4/launches/past
- Using the *requests* to make HTTP requests which were utilized to download the data.
- The response was decoded as a JSON file and normalized along with the pandas library.
- Therefore, we got a Data frame that can be used for data transformation and analysis.

Data Collection – SpaceX API

After getting the data, we could confirm it is a set composed of 187 rows and 43 columns with raw data from 2006.

Important factors is that the raw data contains several entries with no information (considered None or NaN) being “fairings” with 187 rows of null information discarded.

The process of data pre-processing involved transforming the format of some of the columns, (such as “date_utc”) to datetime to facilitate the analysis.

The columns 'rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc' were used for analysis, the remaining columns were removed from the study since, after review, were not relevant.

Data Collection - Scraping

- The URL used was from Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The libraries used to get the data were *requests* and *BeautifulSoup*

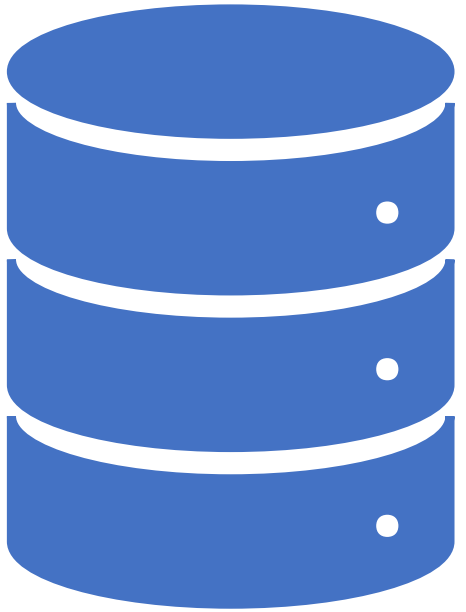
Using requests, the data was extracted and processed with assistance of a BeautifulSoup object.

Then the tables of the website were located, and the columns extracted so therefore can be stored in a dictionary.

This dictionary was transformed into a Data Frame and exported to CSV.

Data Wrangling

- This Data frame was downloaded from: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv.
- It records the flight information and categorizes it into successful and unsuccessful types of landing showing us an improvement over time in the performance of the company.
- The data was downloaded as raw data and, after verification, confirmed their data types, null values and shape.
- We were able to confirm it has 90 rows with 17 columns of information and the column “LandingPad” had a 28.88% of null values in it



EDA with Data Visualization

- The plots that were plotted were:
- Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload Mass vs Launch Site, Orbit Type vs Success Rate, Flight Number vs Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend.
- With this information it was possible to see the relations between the variables and get to make a visual comparison for a better understanding of the data.

EDA with SQL

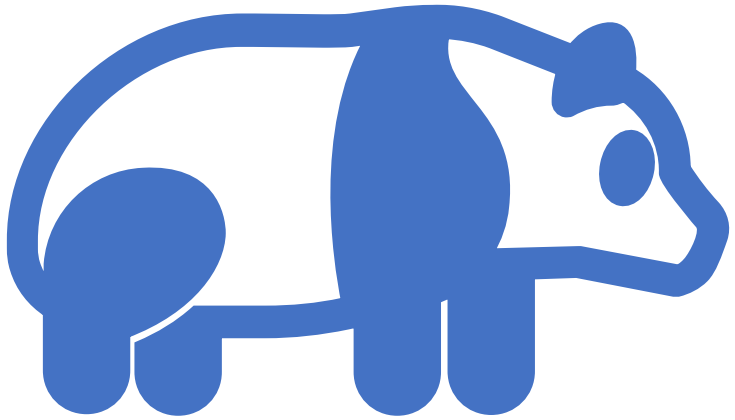
- The SQL queries used were:
- Names of the unique launch sites in the space mission
- 5 records where the launch sites began with CCA
- The total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Name of the boosters which have success in drone ship and have payload mass between 4000 and 6000
- Total number of successful and failed mission outcomes.
- Names of the booster versions which have carried the maximum payload mass.
- Failed landing outcomes in drone ship, with booster version and launch site, during 2015.
- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium

- Within the interactive map:
- Added circle markers, pop-up labels and text lables of the NASA Johnson Space Center.
- Additionally added circle markers on each launch site.
- Each marker was colored with Green for success and Red for failed launches.
- Also, pop-up labels and text labels were added.
- Since many of them were done very close, they were clustered for better comprehension.

Build a Dashboard with Plotly Dash

- Features of the Dashboard
- “Launch Sites” Dropdown List.
- Pie chart showing successful and failed launches by Launch site selected by “Launch Sites” Dropdown.
- “Payload Range” slider for KG.
- Scatter chart representing the relation between Payload Mass and success rate selected by “Payload Range” slider

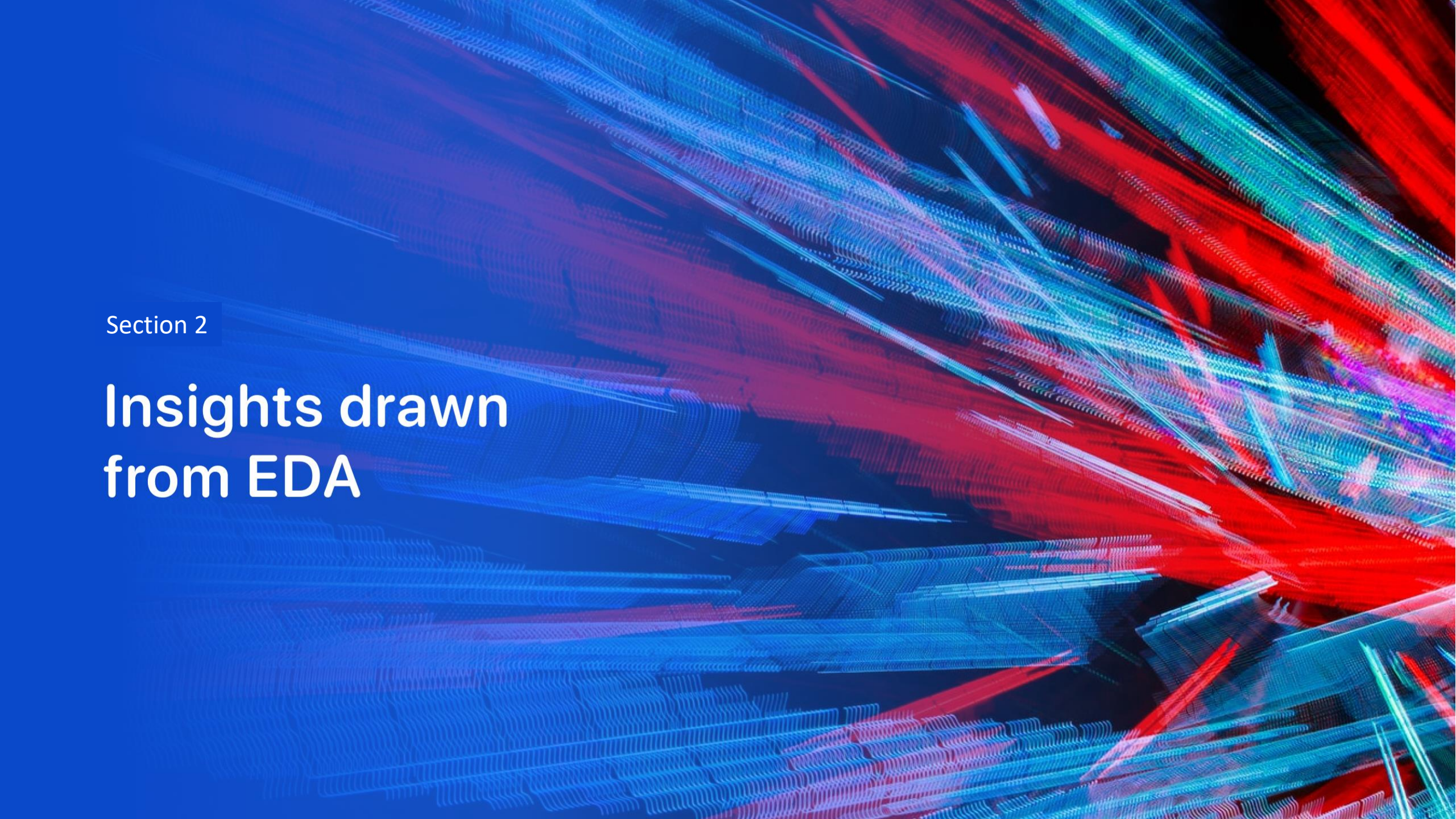


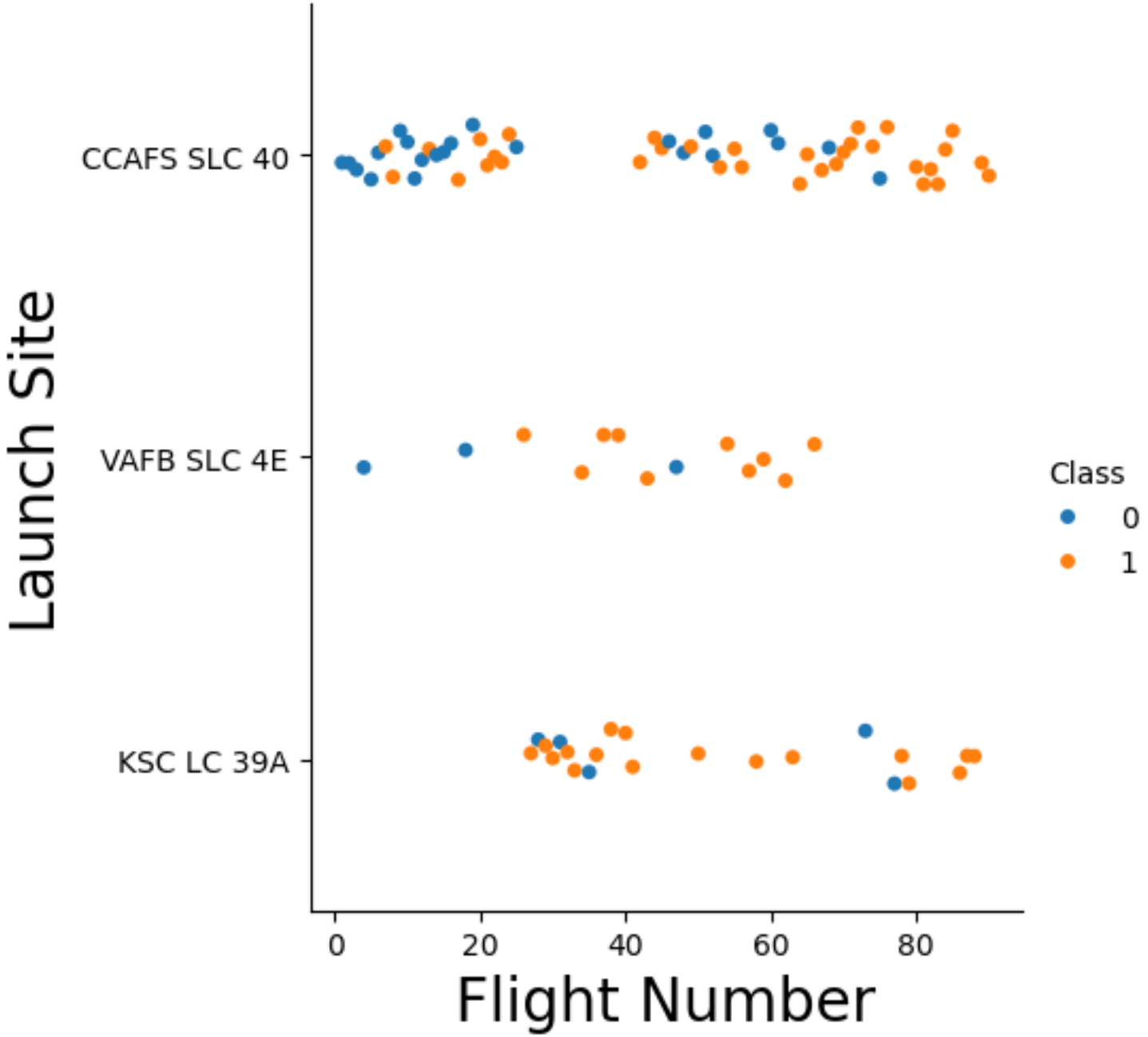
Predictive Analysis (Classification)

- After data extraction using Pandas, the Standardization process was completed using ".StandardScaler()" for better understanding on the scaling during the visualization.
- The data was separated using "train_test_split" with 20% of the data used for testing and 80% for training for verification.
- Then used 4 techniques: Logistic Regression, SVM, Decision tree, KNN with the training and testing data to verify their accuracy comparing each technique.

Results

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


Section 2
Insights drawn from EDA

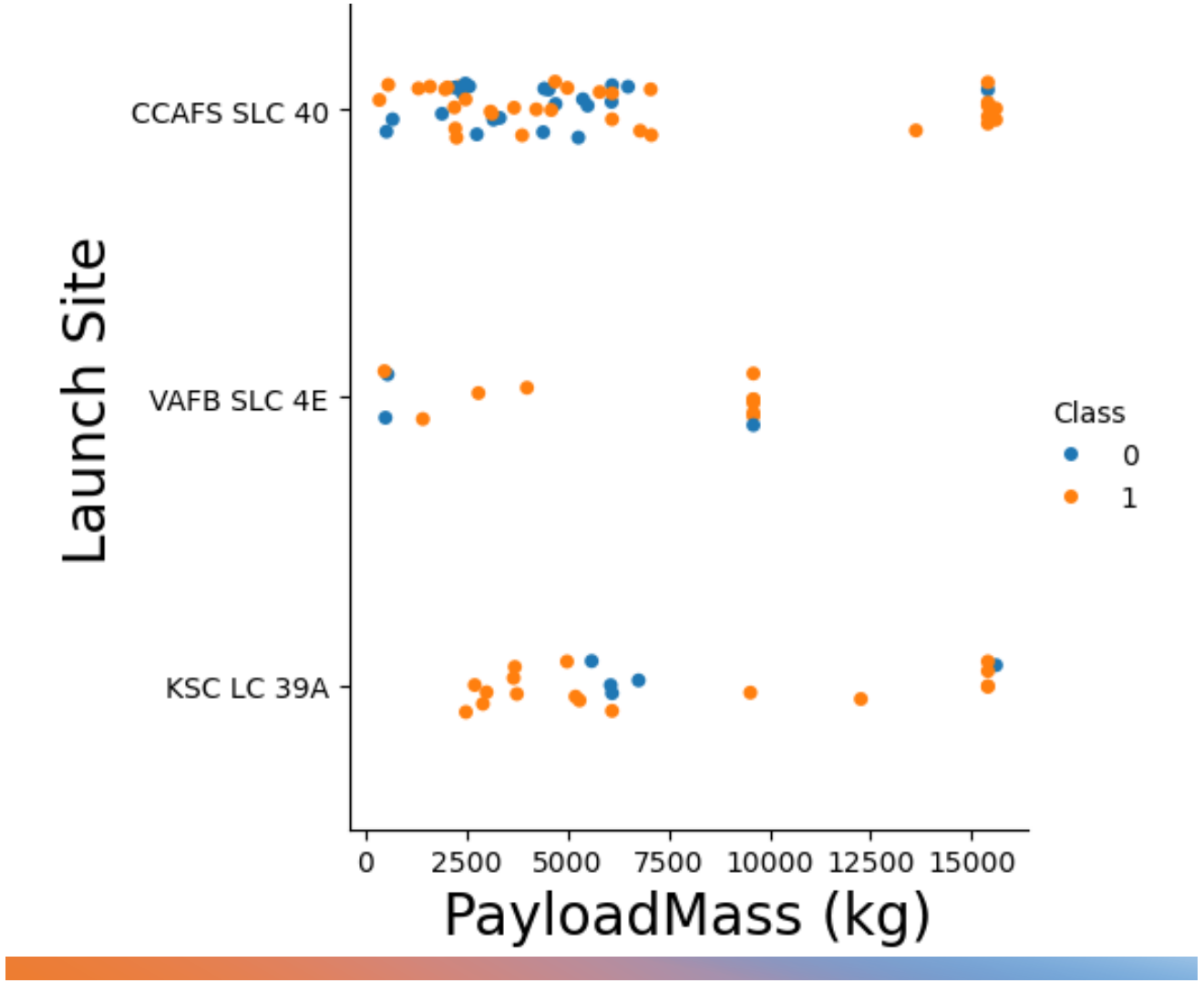


Flight Number vs. Launch Site

- We can see a trend of improvement in the success of the flights as we move from left to right with more orange (success) dots.
- CCAFS SLC 40 is by far the most used launch site.
- There was a period between flight ~25 to 40 where CCAFS SLC 40 was not used.

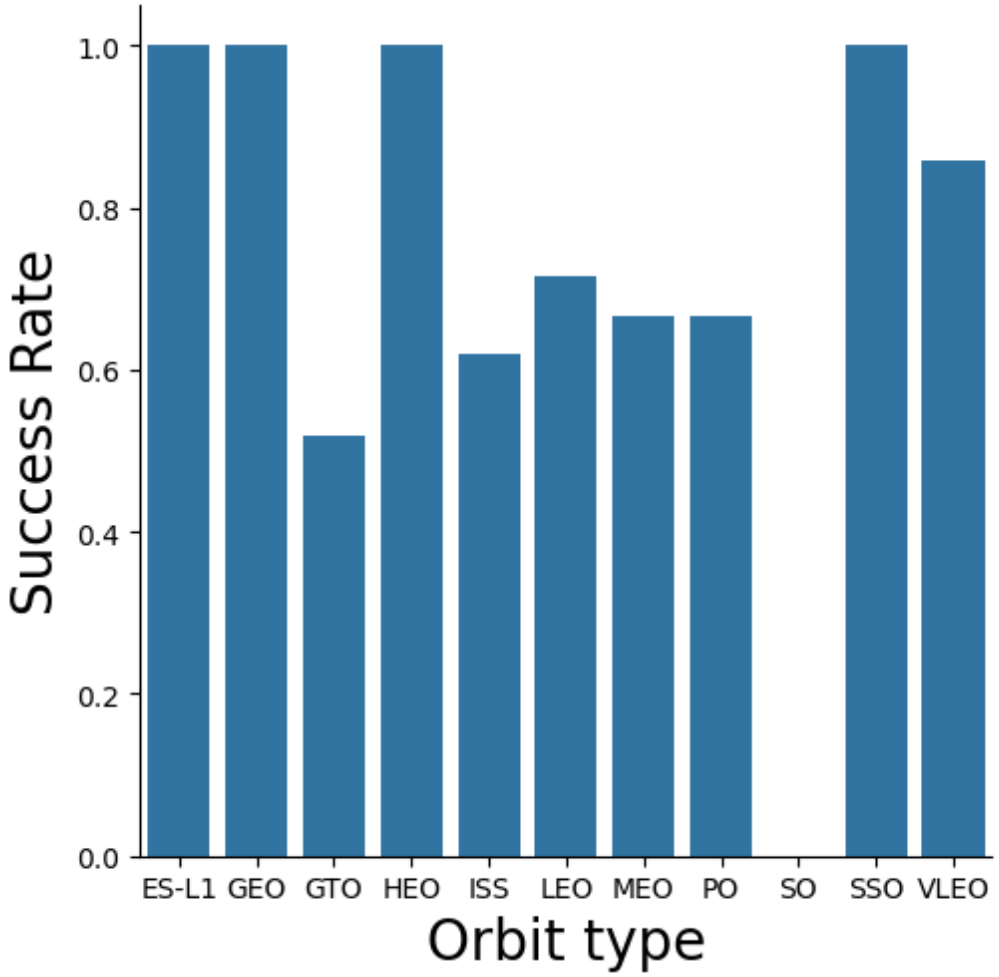
Payload vs. Launch Site

- Most of the launches with payload mass over 7500 KG were successful.
- The maximum payload launched from VAFB SLC 4E was close to 10000 KG
- There was 100% success rate in payload mass from 2500 KG to 5000 KG launched from KSC LC 39A.



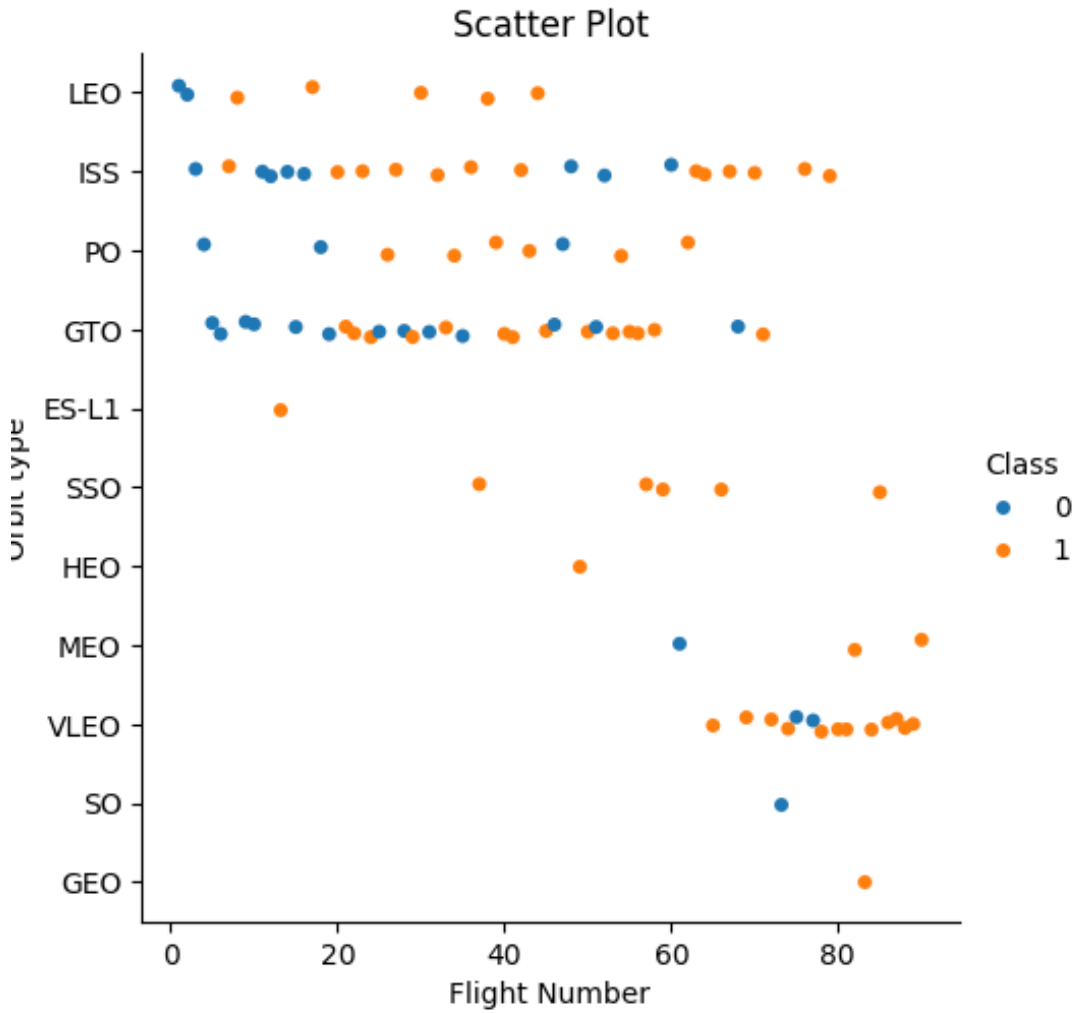
Success Rate vs. Orbit Type

- The orbits ES-L1, GEO, HEO, and SSO had 100% success rate.
- The lowest success rate was SO with 0%
- ISS, LEO, MEO, and PO ha have a similar success rate around the 60%



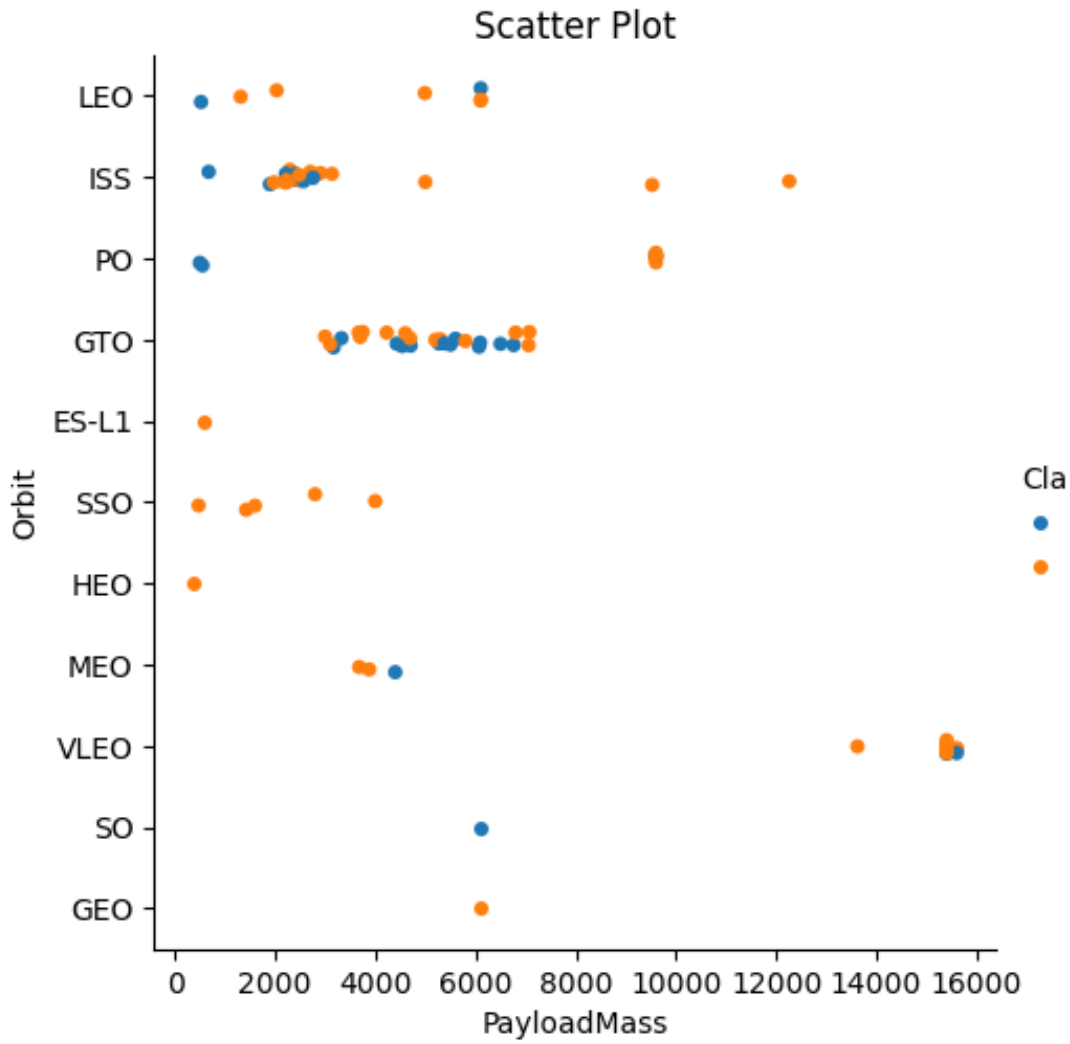
Flight Number vs. Orbit Type

- There seemed to be a trend In the earliest flight numbers on the orbits LEO, ISS, PO, and GTO
- From the flight number 60, there is a trend on the orbit VLEO as well as a positive tendency on success rate.



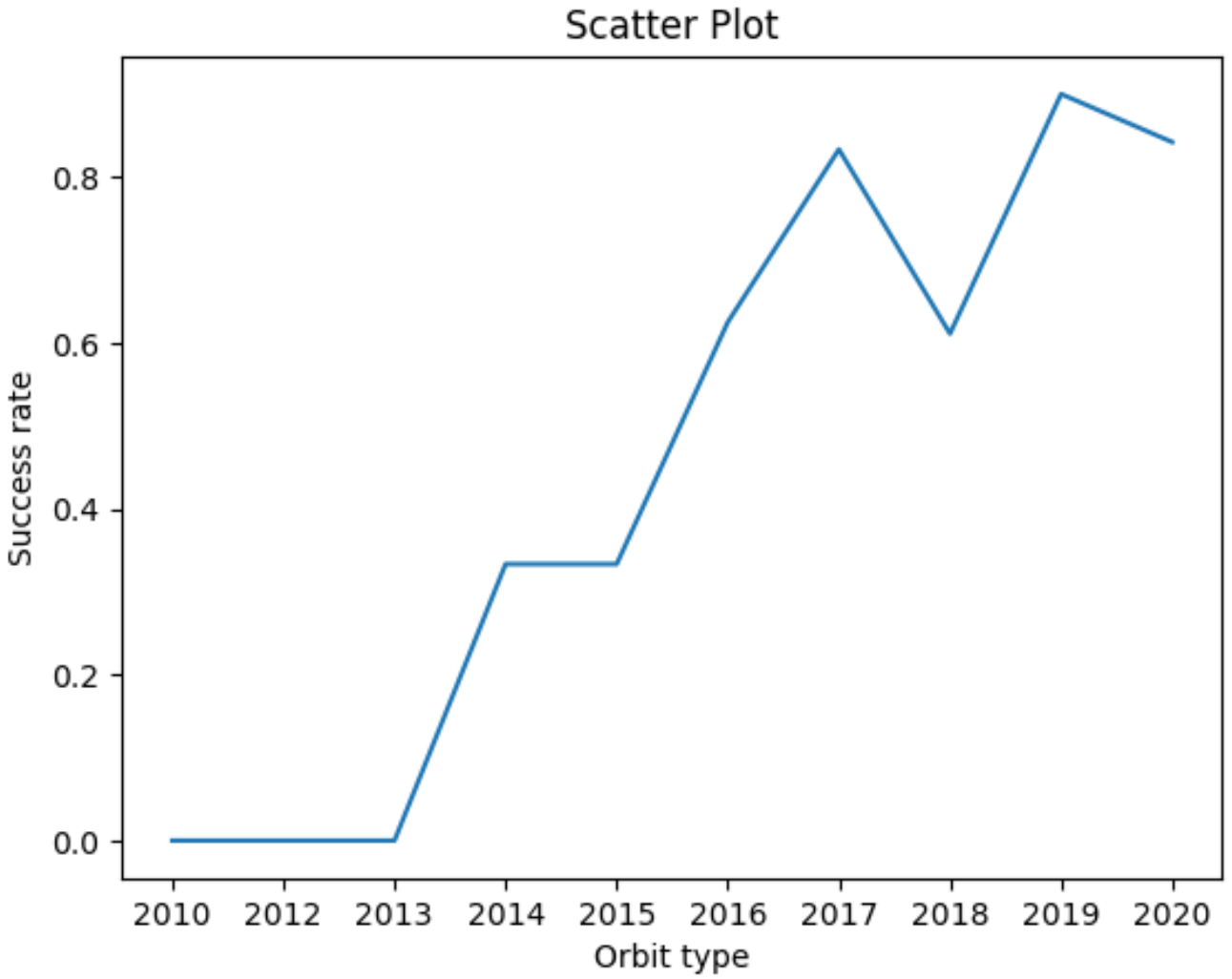
Payload vs. Orbit Type

- Lower Payload mass is related to LEO, ISS, PO, GTO and SSO orbits.
- High Payload Mass is seen for VLEO orbit.
- VLEO payload mass is in general double the mass that GTO has.



Launch Success Yearly Trend

- There is a very clear positive success rate over the years
- There was decrease between 2017 and 2018 and between 2019 and 2020.



All Launch Site Names

- This query displays the unique values in the “Launch_Site” column in the table
- This information is interpreted as the list of Launch sites present in the table.

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

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

```
%sql select * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Cust
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	Sp
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA

Launch Site Names Begin with 'CCA'

- This query displays the table information on the first 5 rows for entries where, within the Launch_Site column, the name starts with CCA.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS sum_of_payload FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

sum_of_payload
45596

Total Payload Mass

- This query displays the sum of all values within the “Payload Mass KG” column, for the rows where the “Customer” column is “NASA (CRS”.
- The result is renamed as sum_of_payload.

Average Payload Mass by F9 v1.1

- This query displays the average of all values within the “Payload Mass KG” column, for the rows where the “Booster Version” column is “F9 v1.1”.
- The result is renamed as avg_of_payload.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_of_payload FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

avg_of_payload
2928.4

```
%sql SELECT * FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' ORDER BY Date LIMIT 1;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

First Successful Ground Landing Date

This query displays the first row where the Landing Outcome is “Success (ground pad)”, ordered by date.

Which shows us the first success in landing outcome on the table

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query shows the Booster_Version name on the table where the Landing_Outcome column was Success in drone ship, and their Payload Mass was between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

- This query calculates the total number of flights by outcome

```
%sql SELECT Mission_Outcome, COUNT(*) AS Total_number FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- This query displays the Booster_Version that carried the maximum Payload.
- This is done by subquery, selecting the max payload mass and using it as the parameter for the main query

```
%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

2015 Launch Records

- This query gets the Landing outcome, booster version and launch site where, in the date column, the year portion is 2015 and the landing outcome column is failure (drone ship)

```
%sql SELECT strftime('%M', date) AS month_name,Landing_Outcome,Booster_Version,Launch_Site FROM SPACEXTABLE
WHERE strftime('%Y', date) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

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

month_name	Landing_Outcome	Booster_Version	Launch_Site
00	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
00	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This query Ranks 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_outcomes from SPACEXTBL
where date between '2010-06-04' and '2017-03-20' group by landing_outcome order by count_outcomes desc;
```

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

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

Section 3

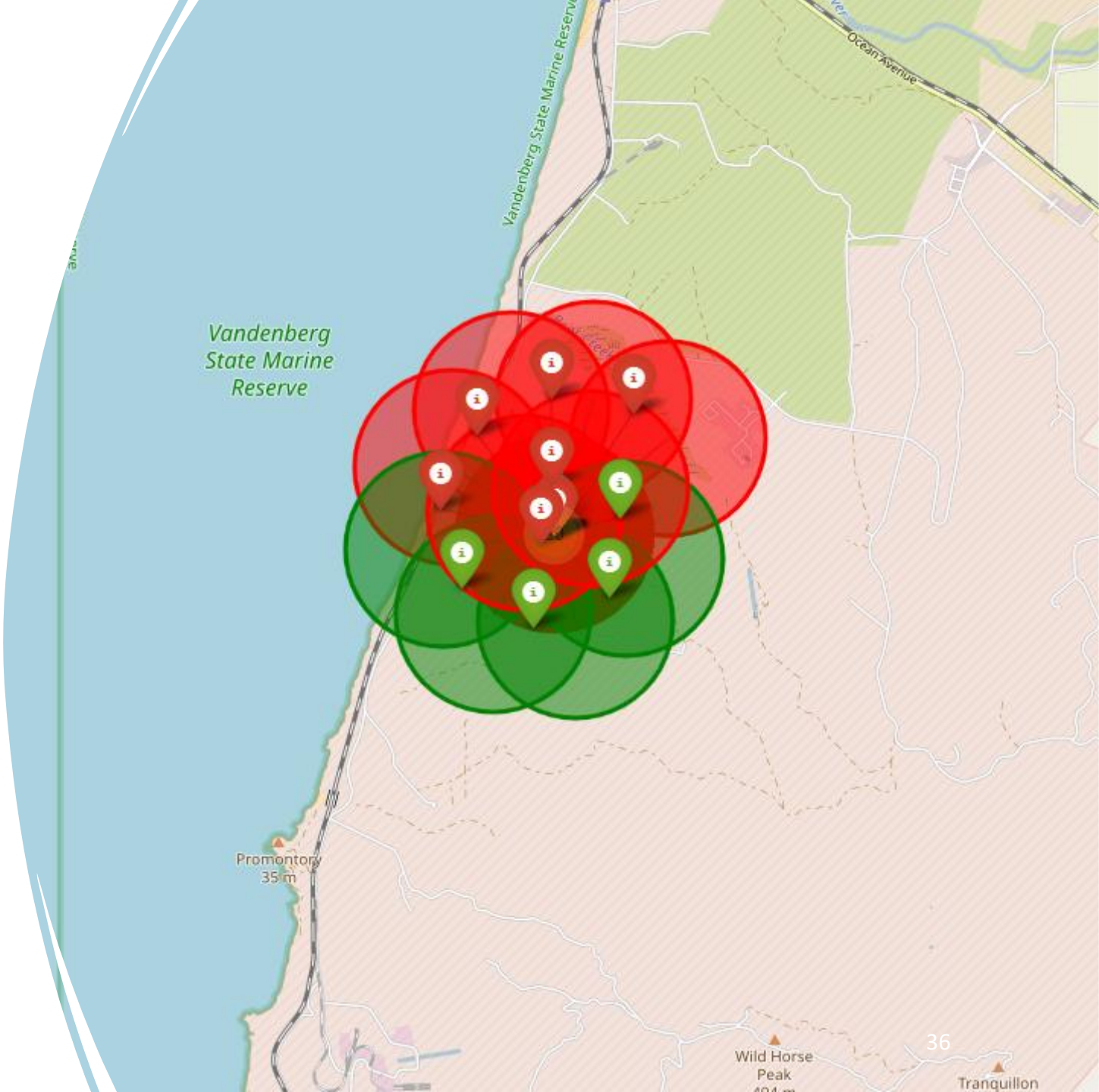
Launch Sites Proximities Analysis

Launch sites on the map

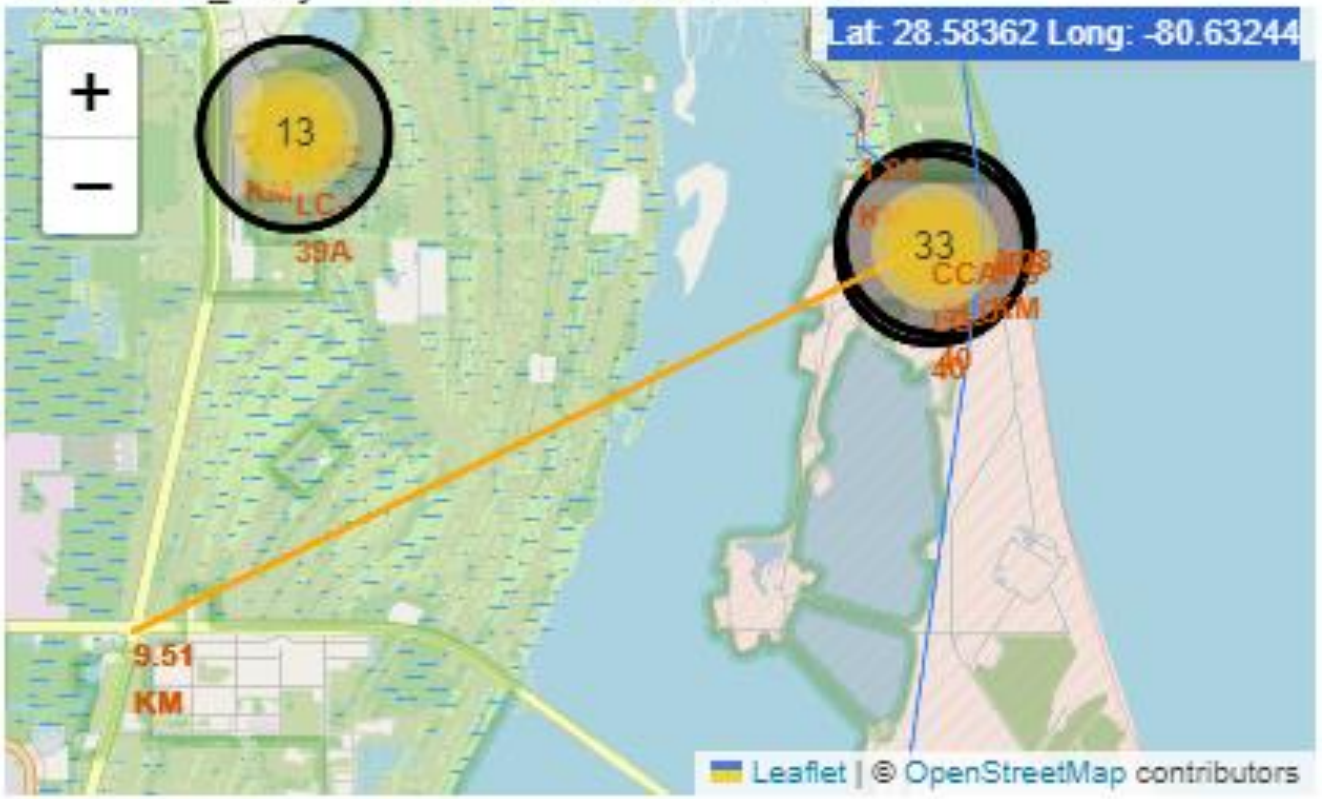
- This is the map of the US with 3 marks: The three: 2 of the launch sites (California and Florida) and the NASA Johnson Space Center (Texas)
- The map show the strategic position where this events happen close to the equator and to costal areas.

Color-coded markers

- Each launch is labeled as green for successful or red for failed, making it easy to compare.
- Here you can see the VAFB SLC 4E in California clustered, once we zoom in we can view with more detail each lunch entry.
- Same happens to the cluster in Florida

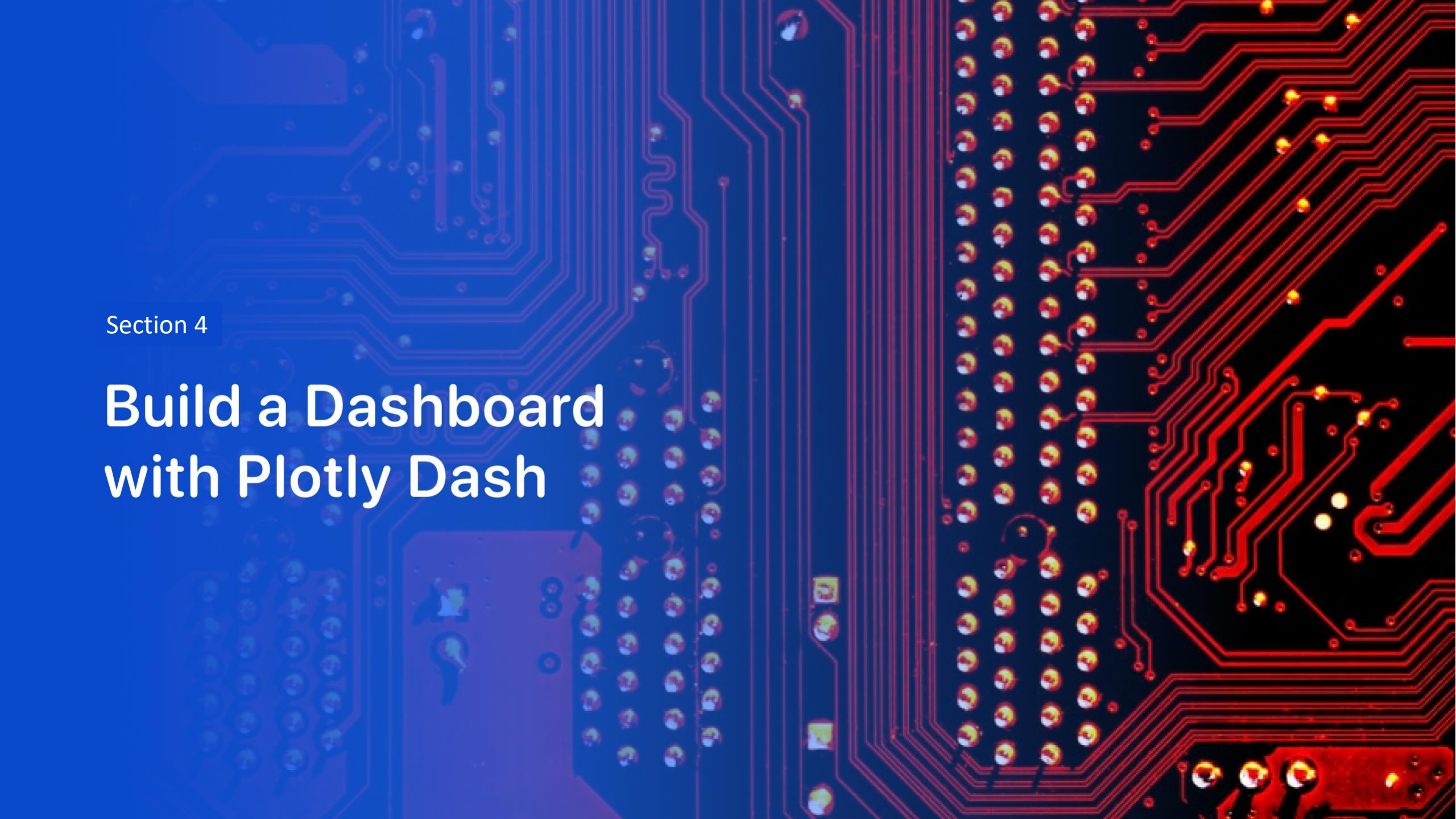


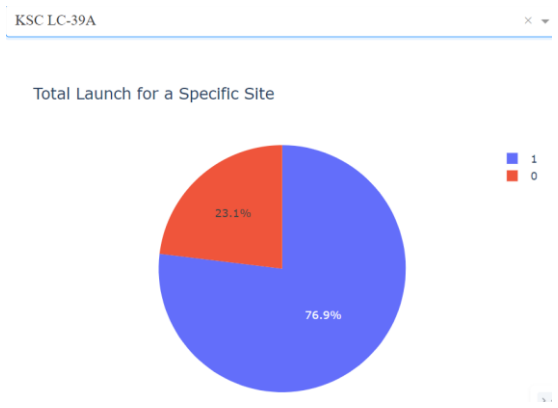
```
distance_highway = 0.08410278857048983  km
distance_railroad = 1.6251601735058 42  km
distance_city = 51.55496834371294  km
```

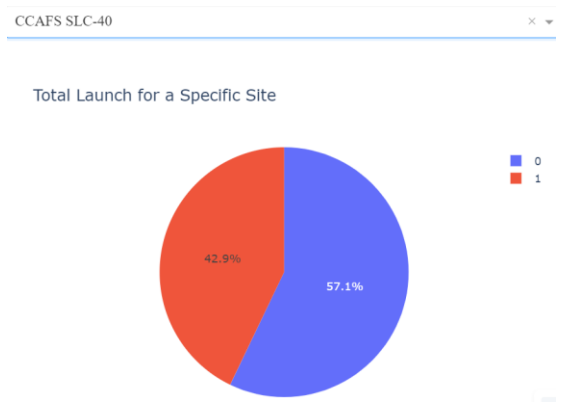


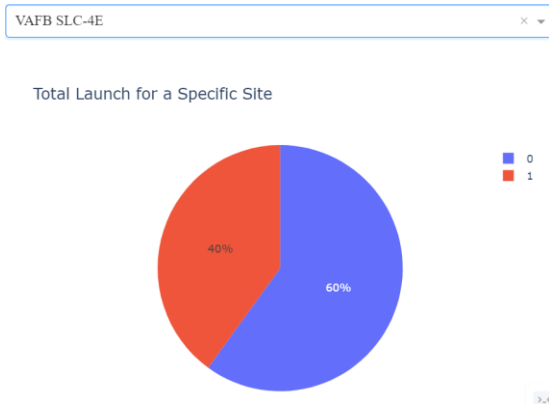
Distance markers

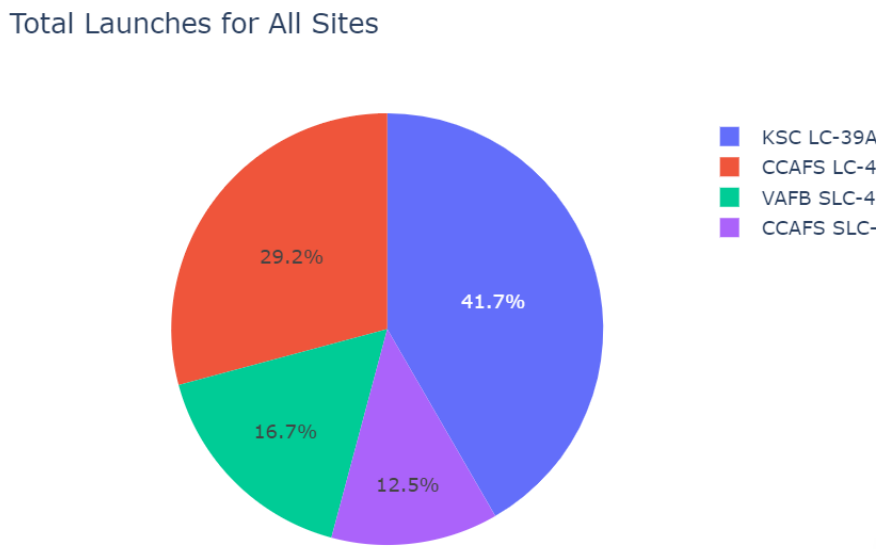
- This screenshot displays in the map how it is marked for the distance to the railroad from Florida's marker


Section 4
Build a Dashboard with Plotly Dash







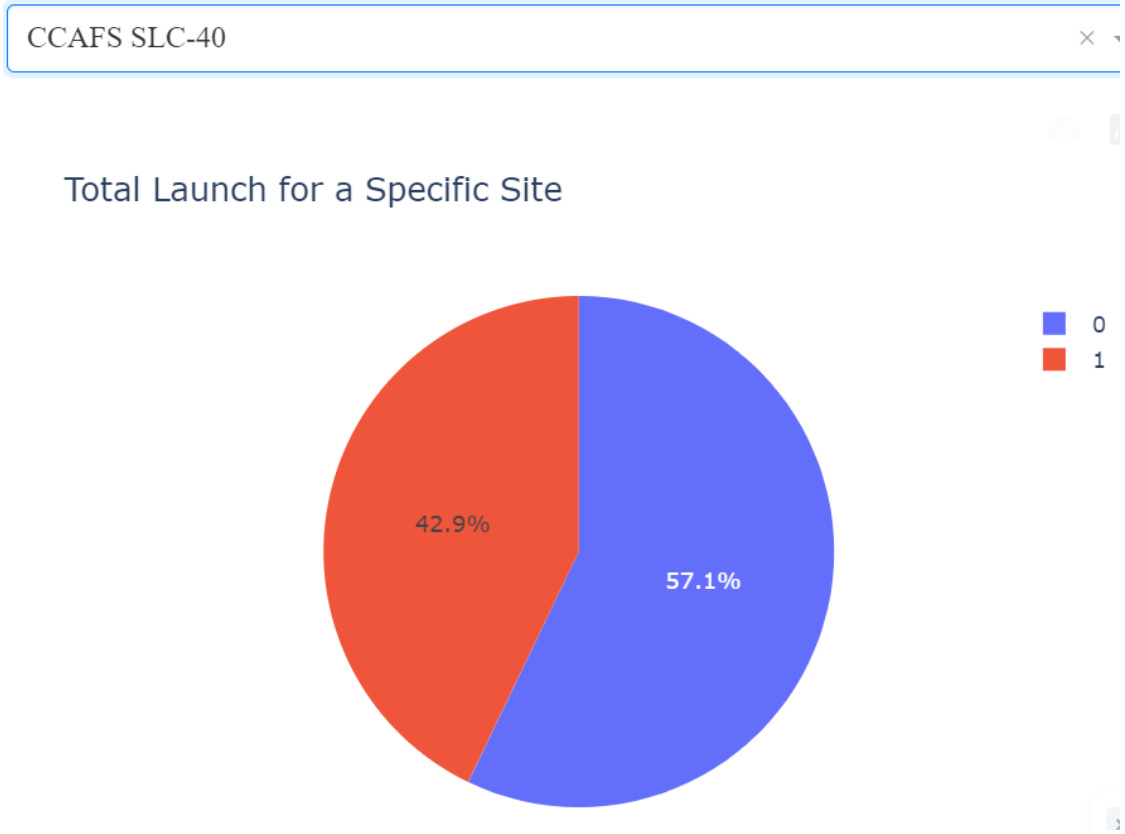


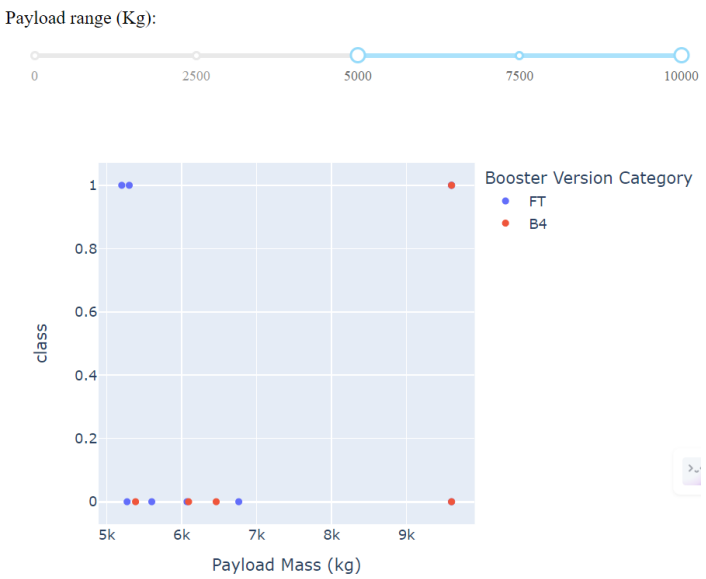
Dash Pie charts

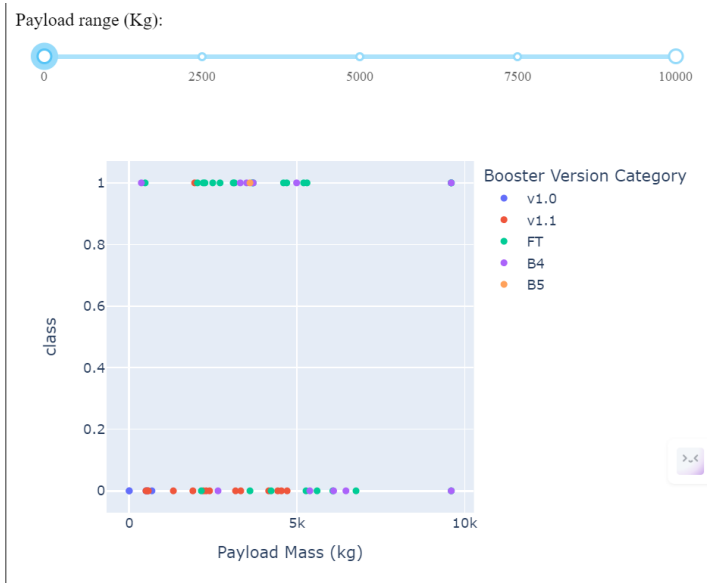
- We can see first a Pie chart with the total launches for all sites, and then the success rate for each one of the sites.
- Important to mention that in the 4 sites, the blue color represents the failures (class= 0) and the red success (class= 1)

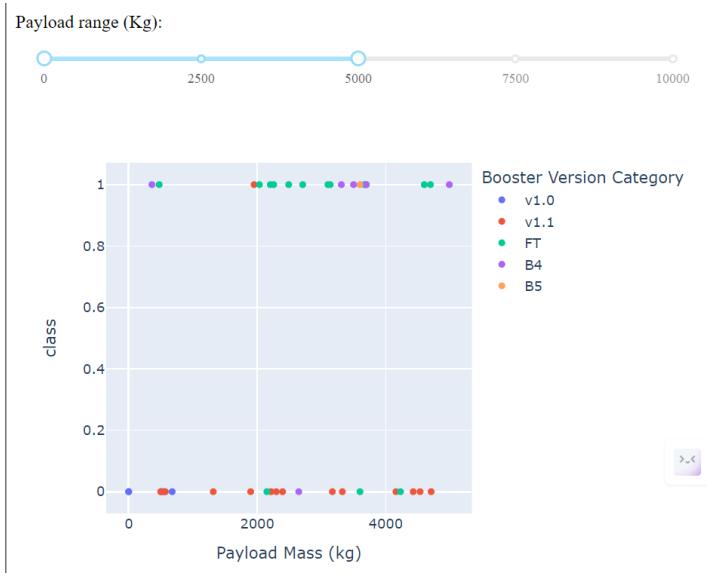
Highest success rate

- Here we can see CCAFS SLC-40, with a 42.9% success rate is the highest success rate among the launch sites.







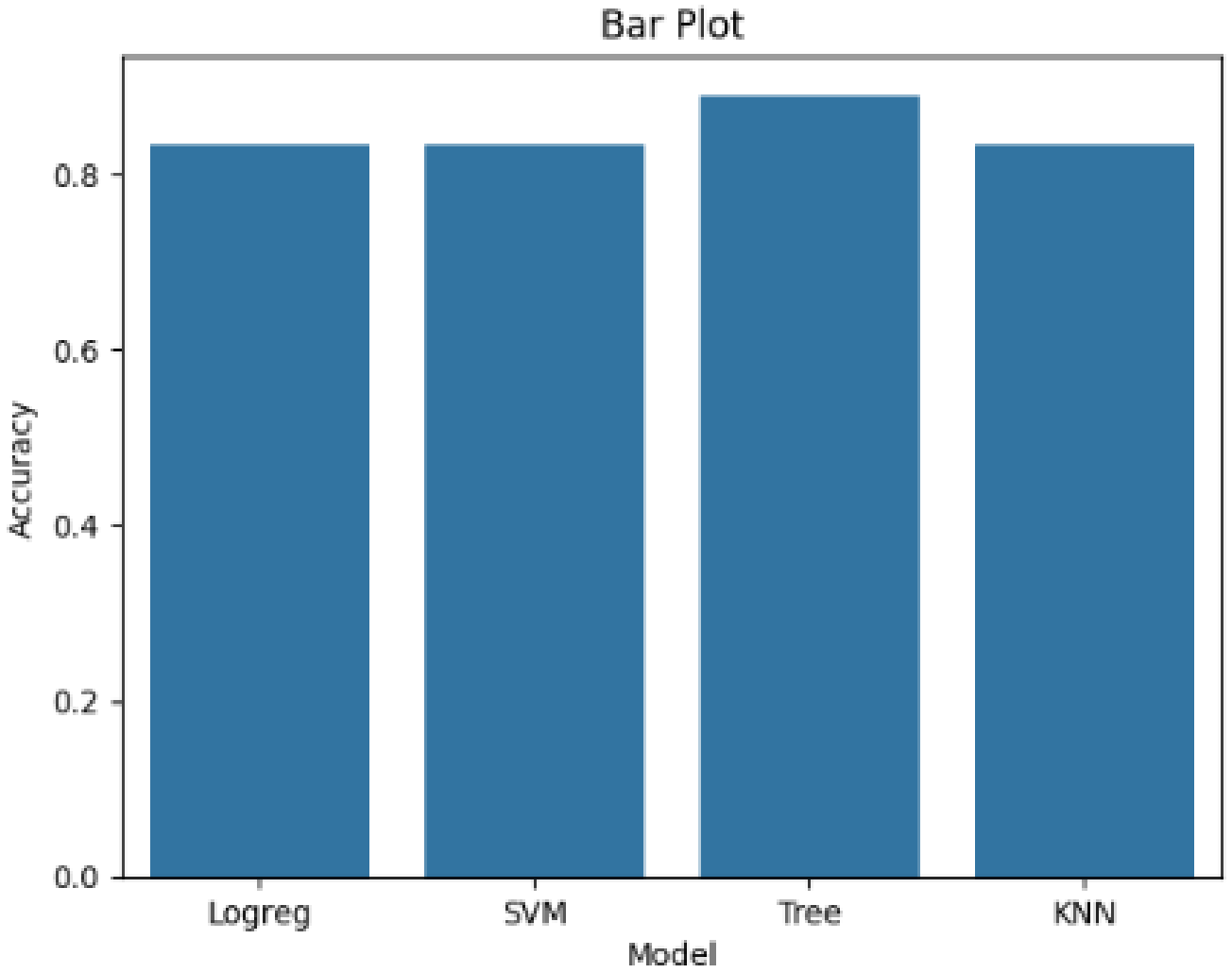


Class vs Payload interactive slider

- Here we can see plotted the payload mass vs class comparison (being 1 Successful and 0 Failure)
- The first screenshot goes from 5000 KG to 10000 KG
- The second goes from 0 KG to 10000 KG
- The last one goes from 0 KG to 5000 KG

Section 5

Predictive Analysis (Classification)



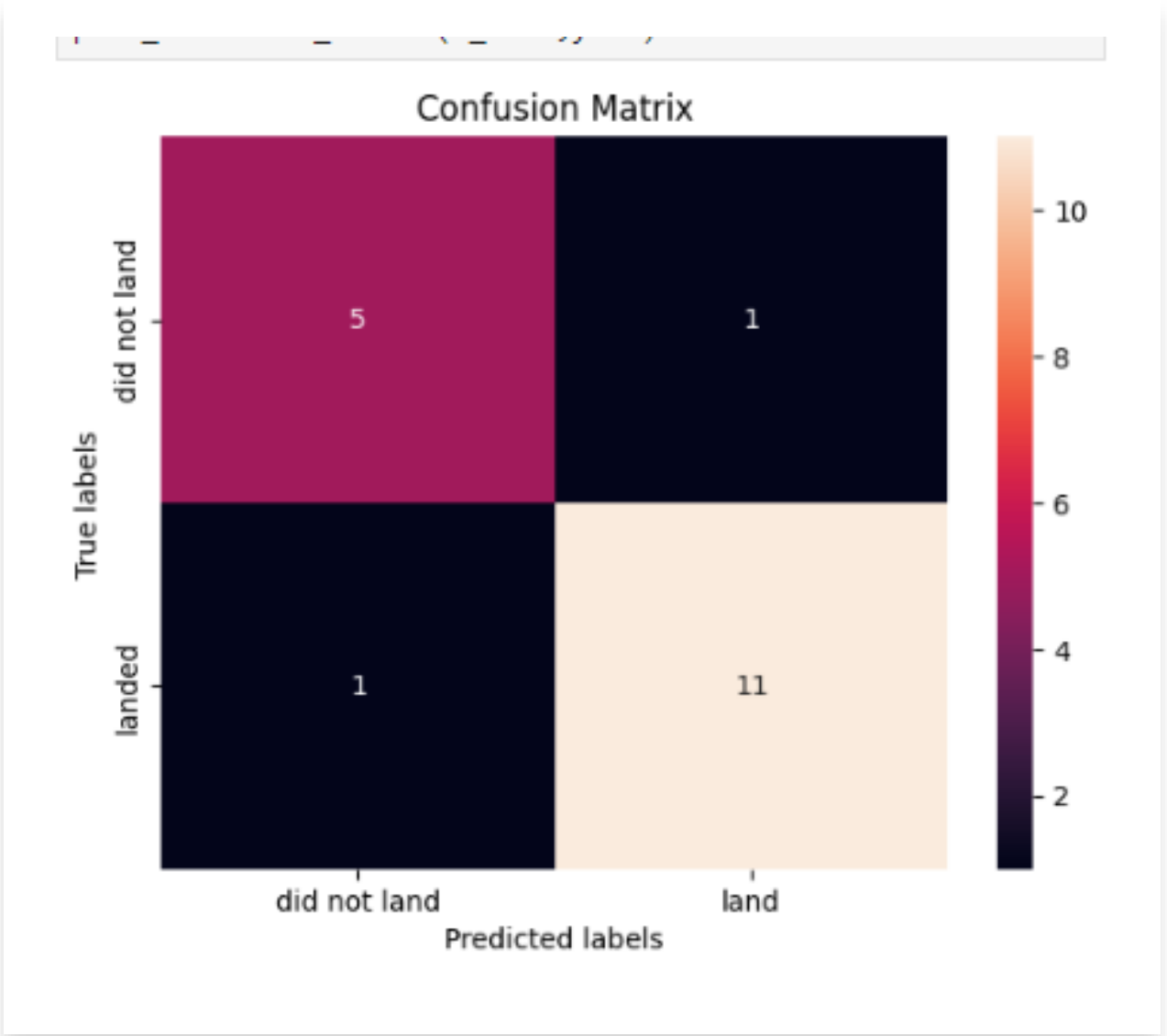
Classification Accuracy

- This plot compares the accuracy of the models when it comes to predict values from a training-test set.
- We can see that Decision Tree is the one with the highest accuracy.

This is the confusion matrix represent the accuracy on the predictions.

We can see 16 accurate predictions and 2 false predictions

Confusion Matrix: Decision Tree



Conclusions

- Over time, the improvement into the success rate of SpaceX has shown proven success.
- KSC LC-39A has the highest success rate of the launches from all sites
- Decision Tree model had the highest accuracy to predict results
- CCAFS SLC 40 in Florida was the most popular launch site

Thank you!