IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Deeksha Pant
28 September 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Collecting data by:
    - API,
    - Web Scraping
  - Data wrangling for improving the data quality
  - EDA (Exploratory Data Analysis) of the processed data by:
    - SQL
    - statistical analysis and data visualization, to see directly how variables might be related to each other
  - Interactive Visual Analytics with Folium
  - Predictive modelling for discovering more insights
- Summary of all results
  - EDA Findings
  - Interactive Analytics Findings
  - Predictive Modelling Insights

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers
  - Correlation between various features towards success rate in landing
  - Predict if the Falcon 9 first stage will land successfully
  - Predict if SpaceX will reuse the first stage using Machine Learning Models

Section 1

Methodology

Methodology

Executive Summary

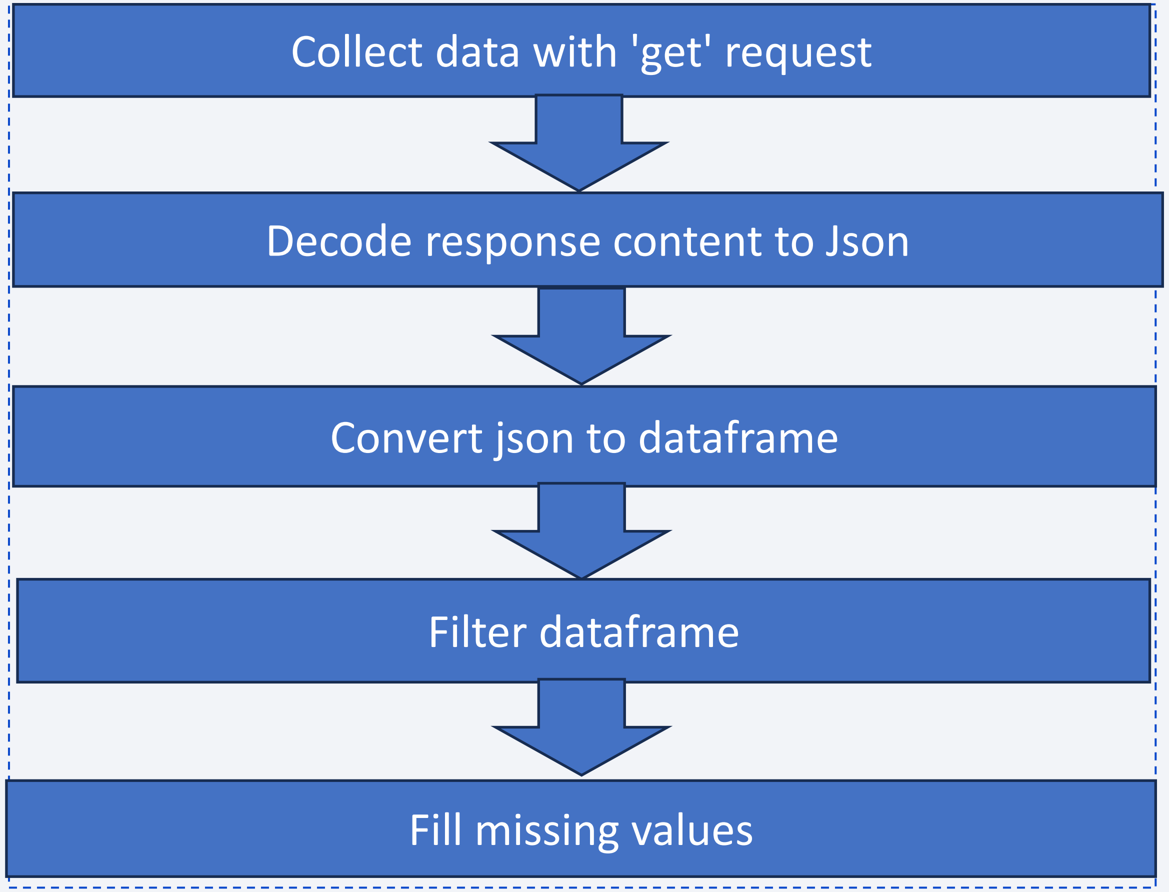
- Data collection methodology:
  - SpaceX launch Data was collected using SpaceX REST API (https://api.spacexdata.com/v4/rockets/)
  - Performing web scraping from a Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Creating a landing outcome label from Outcome label using One hot encoding technique
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardized data split into training and testing data. Trained the model & performed Grid Search to find the best hyperparameter values which determine the model with the best accuracy
  - Tested Logistic Regression, SVM, Decision Tree Classifier, and KNN M/L models

Data Collection

- Describe how data sets were collected.
  - Using API
    - Data collection was done using 'get' request to the SpaceX API
    - Decoded the response content as a Json using .json()
    - Used json_normalize method to convert the json result into a dataframe
    - Filtered the dataframe to only include 'Falcon 9' launches
    - Cleaned the data, checked for missing values and fill in missing values where necessary
  - Web Scraping
    - Performed Web-scraping using BeautifulSoup from a snapshot of the ‘List of Falcon 9 and Falcon Heavy launches’ Wiki page
    - Extracted all column/variable names from the HTML table header
    - Created a dataframe by parsing the launch HTML tables

Data Collection – SpaceX API

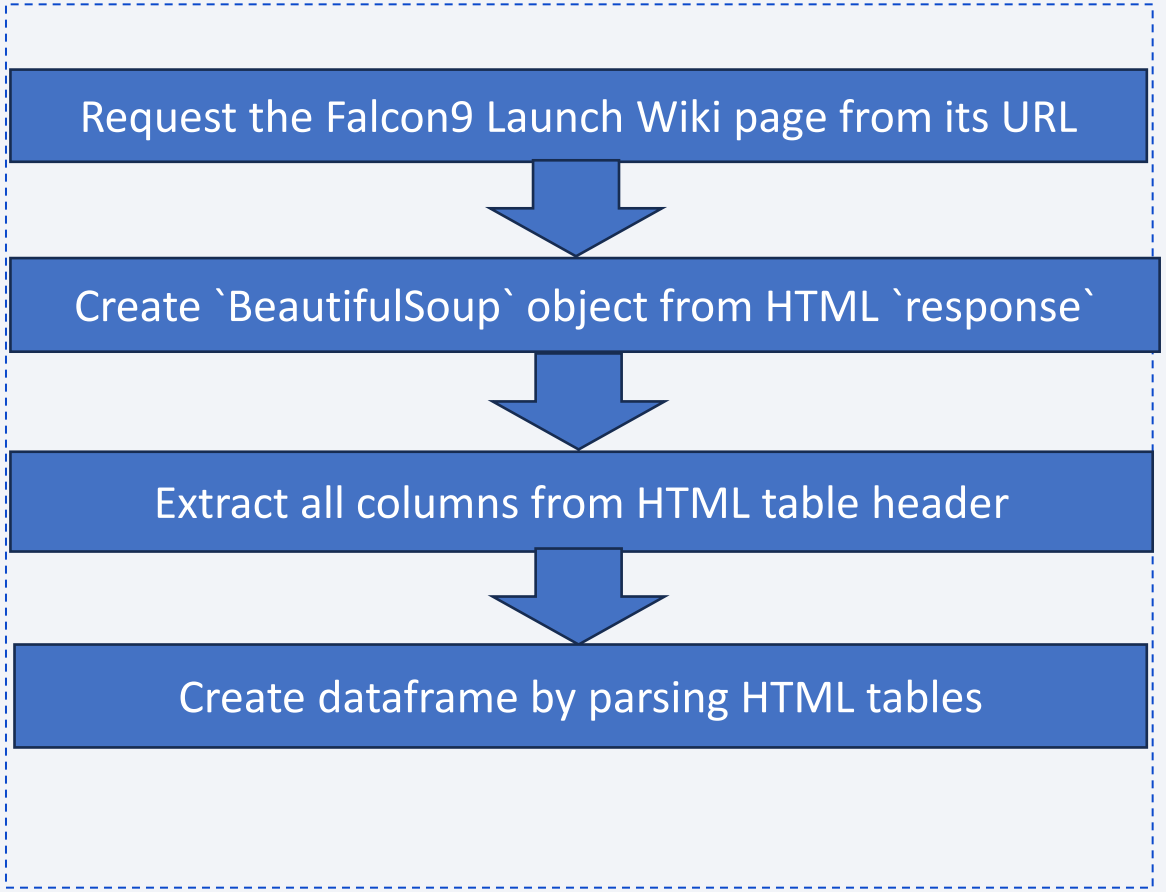
- Data collected using 'get' request to the SpaceX API was filtered and cleaned using the flowchart mentioned beside
- GitHub URL of the completed SpaceX API calls:
  - https://github.com/Deeksha-pant/IBM-Applied-Data-Science-Capstone/blob/main/spacex_data_collection_api.ipynb


Collect data with 'get' request
Decode response content to Json
Convert json to dataframe
Filter dataframe
Fill missing values

Data Collection - Scraping

- Performed web scraping from Wikipedia with BeautifulSoup using the flowchart mentioned beside

- GitHub URL of the completed web scraping notebook :
  - https://github.com/Deeksha-pant/IBM-Applied-Data-Science-Capstone/blob/main/spacex_webscraping.ipynb



Request the Falcon9 Launch Wiki page from its URL

Create `BeautifulSoup` object from HTML `response`

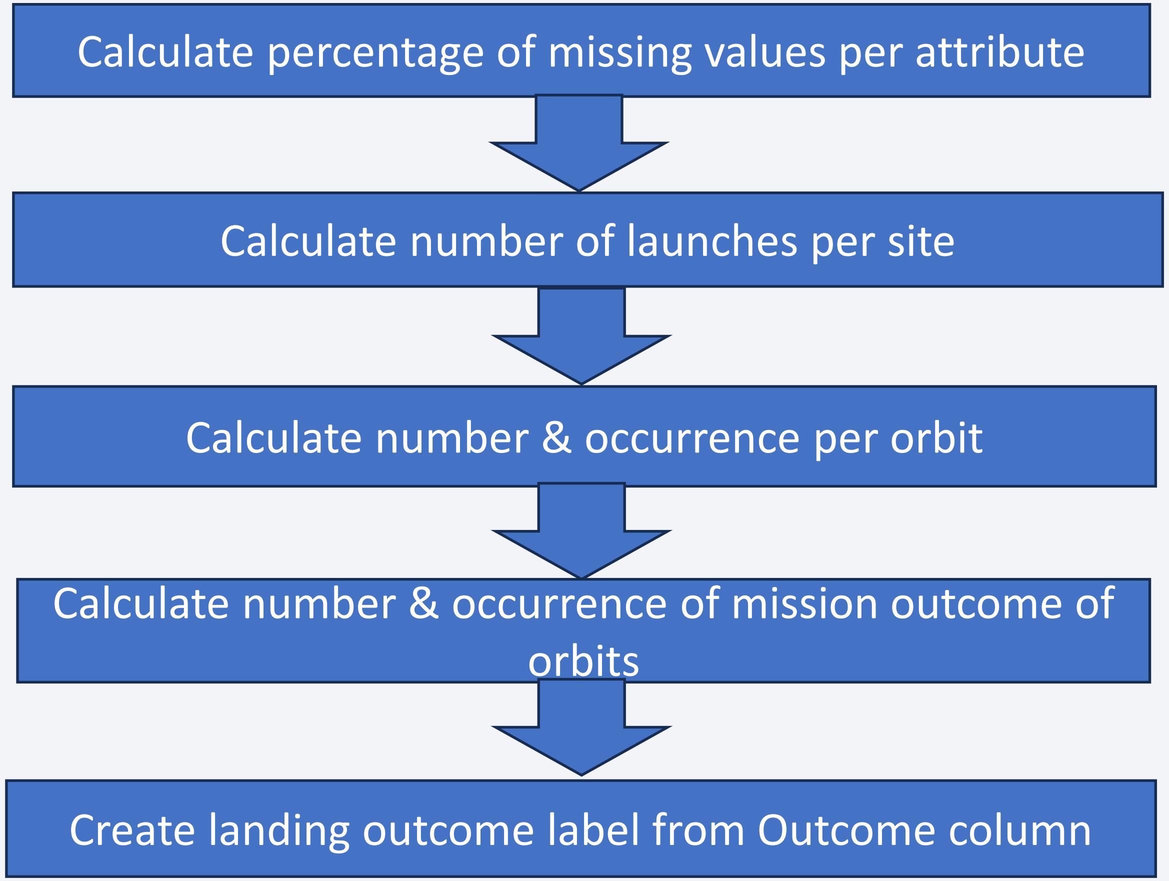
Extract all columns from HTML table header

Create dataframe by parsing HTML tables

Data Wrangling

- Performed Data wrangling by EDA (Exploratory Data Analysis) using the flowchart mentioned beside

- GitHub URL of the completed data wrangling related notebook:
  - https://github.com/Deeksha-pant/IBM-Applied-Data-Science-Capstone/blob/main/spacex_Data_wrangling.ipynb



Calculate percentage of missing values per attribute

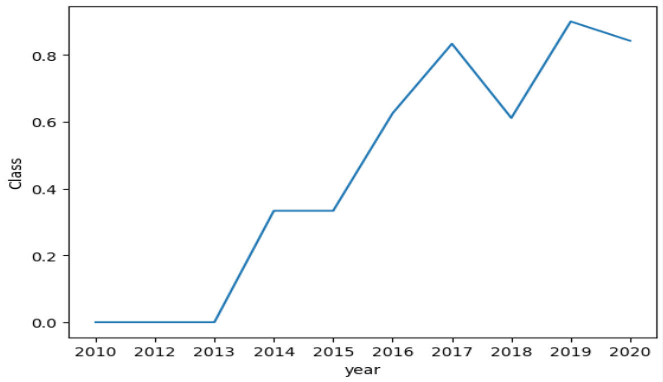
Calculate number of launches per site

Calculate number & occurrence per orbit

Calculate number & occurrence of mission outcome of orbits

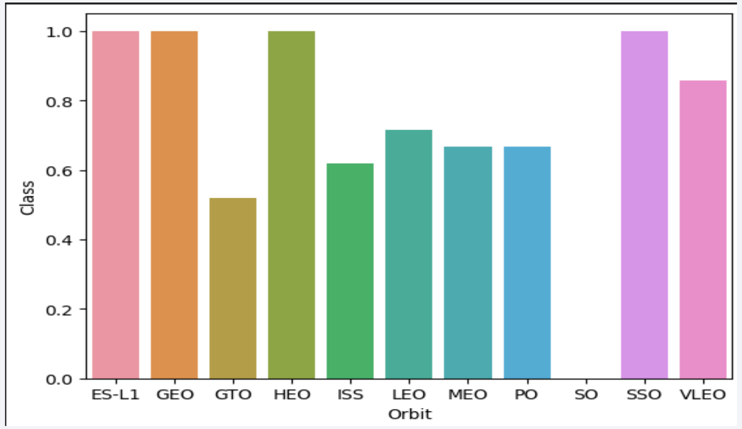
Create landing outcome label from Outcome column

EDA with Data Visualization



Visualization of the launch success yearly trend

- Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing
- 2019 has the highest success rate



Visualizing success rate of each orbit type

- ES-L1, GEO, HEO, SSO has the highest success rate
- SO, STO has the lowest success rates

GitHub URL of the completed data EDA with data visualization notebook:

- https://github.com/Deeksha-pant/IBM-Applied-Data-Science-Capstone/blob/main/spacex_EDA_Data_Visualization.ipynb

EDA with SQL

- EDA done using following SQL Queries:
  - Names of the unique launch sites in the space mission
  - Top 5 records where launch sites begin with the string 'CCA'
  - 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
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - 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
- GitHub URL of the completed EDA with SQL notebook:
  - https://github.com/Deeksha-pant/IBM-Applied-Data-Science-Capstone/blob/main/spacex_EDA_SQL_sqllite.ipynb

Build an Interactive Map with Folium

- Circle & Marker :
  - Added a highlighted circle area with a text label to provide more intuitive insights about launch sites
  - Marked the success/failed launch rates for each site on the map
    - Successful launch(class=1) -> green marker
    - Failed launch (class=0) -> red marker
- Marker clusters:
  - To simplify a map containing many markers having the same coordinate.
  - From the colour labelled markers in marker clusters, it's easy to identify which launch sites have relatively high success rates
- PolyLine :
  - To explore and analyse the proximities of launch sites, calculated the distances between a launch site to any railway, highway, coastline, etc

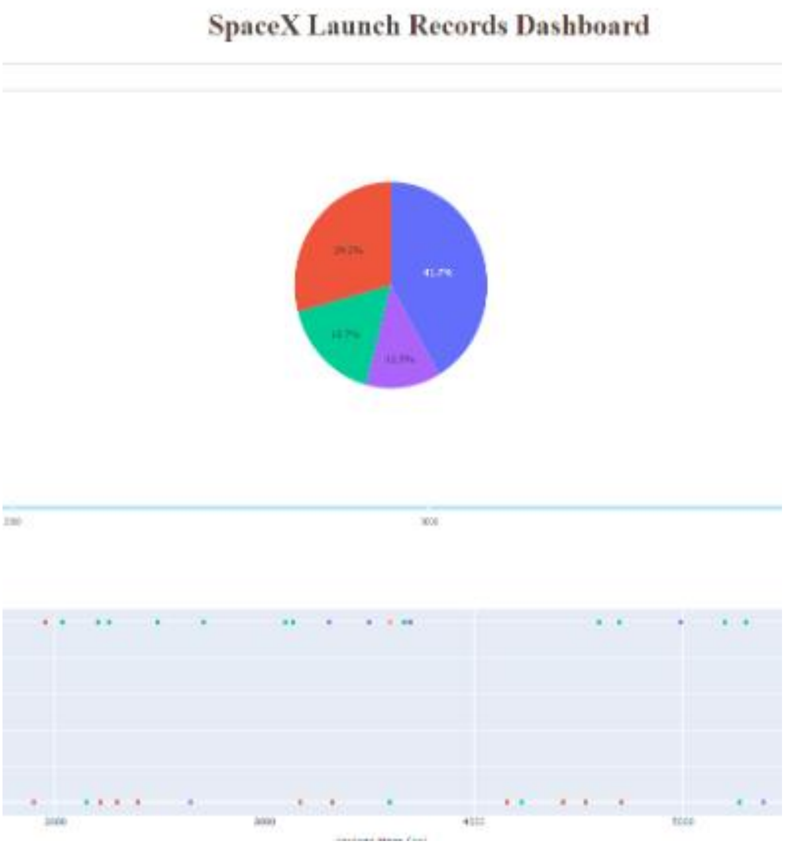
GitHub URL of the completed interactive map with Folium map:

- https://github.com/Deeksha-pant/IBM-Applied-Data-Science-Capstone/blob/main/spacex_Analysis_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- Built a Plotly Dash application (to perform interactive visual analytics on SpaceX launch data in real-time)
  - **Drop-down:** for Launch Site Options
  - **Pie chart:** to visualize launch success counts (based on the selected launch site from site-dropdown)
  - **Range Slider:** to Select Payload
  - **Scatter plot:** to visually observe correlation of payload with mission outcomes for selected site(s)
    - colour-label the Booster version on each scatter point to observe mission outcomes with different boosters

- GitHub URL of completed Plotly Dash lab:
  - https://github.com/Deeksha-pant/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

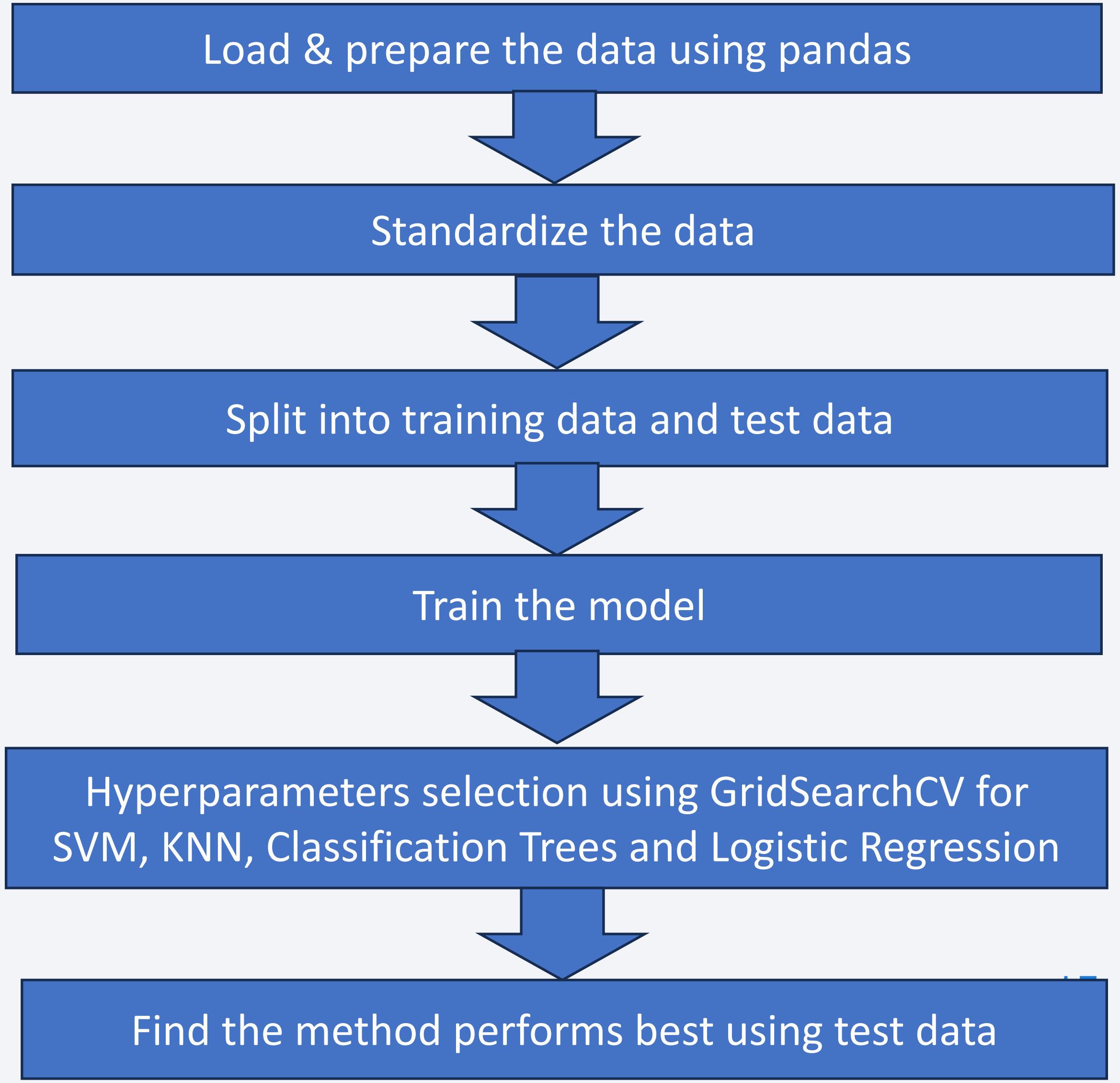


SpaceX Launch Records Dashboard

Predictive Analysis (Classification)

- Steps followed for Predictive Analysis using the flowchart mentioned beside:

- GitHub URL of completed predictive analysis lab:
  - https://github.com/Deeksha-pant/IBM-Applied-Data-Science-Capstone/blob/main/spacex_Machine_Learning_Prediction.ipynb



Load & prepare the data using pandas

Standardize the data

Split into training data and test data

Train the model

Hyperparameters selection using GridSearchCV for SVM, KNN, Classification Trees and Logistic Regression

Find the method performs best using test data

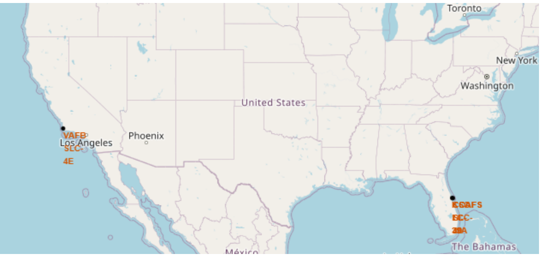
Results

Exploratory data analysis results

- **Names of the unique launch sites in the space mission**
  - ❑ CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
- **Total payload mass carried by boosters launched by NASA (CRS)**
  - ❑ 45596 KG
- **Average payload mass carried by booster version F9 v1.1**
  - ❑ 2928 KG
- **Date when the first successful landing outcome in ground pad was achieved**
  - ❑ 2015-12-22
- **Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
  - ❑ F9 FT B1022 , F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
- **Total number of successful and failure mission outcomes**
  - ❑ Success:100, Failure :1

Results

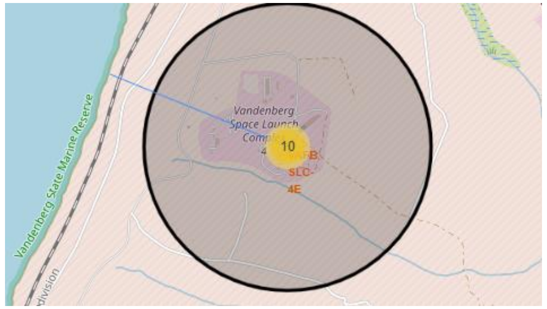
- Interactive analytics demo



Circle & Markers

CCAFS SLC-40
26

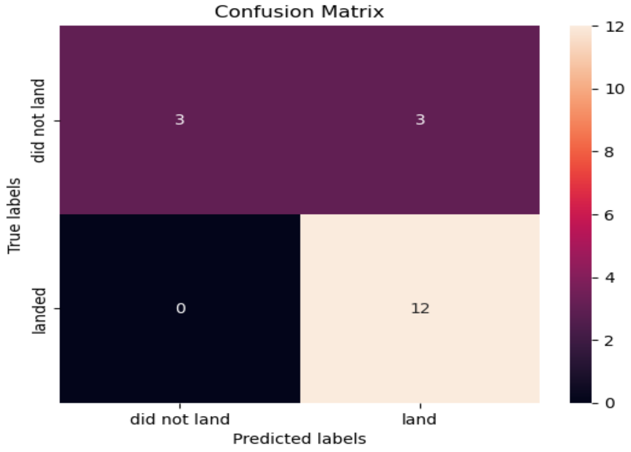
MarkerCluster



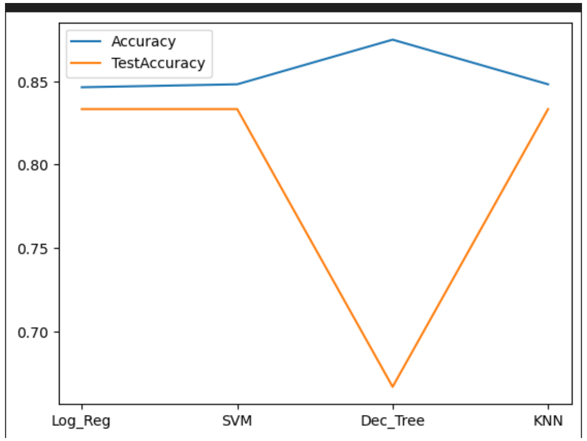
PolyLine

- Predictive analysis results

Confusion Matrix



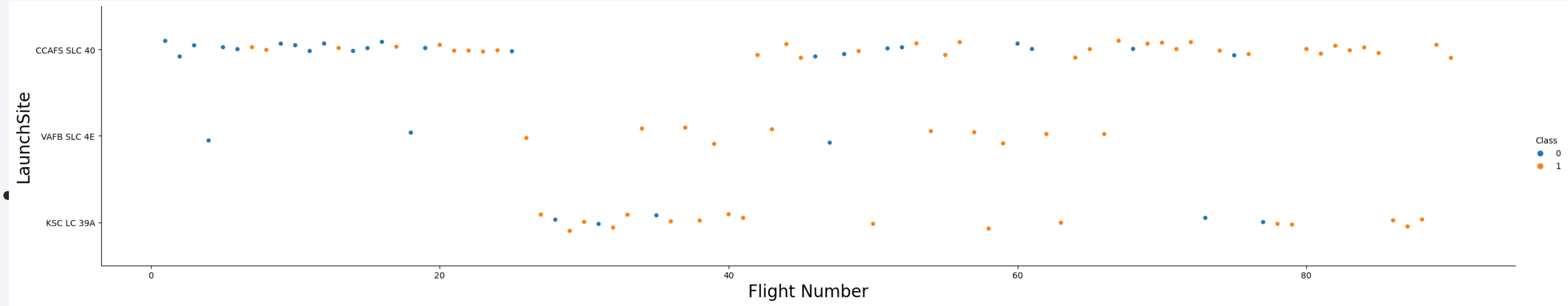
Decision Tree: Highest Accuracy
Lowest Test Accuracy



Section 2

Insights drawn from EDA

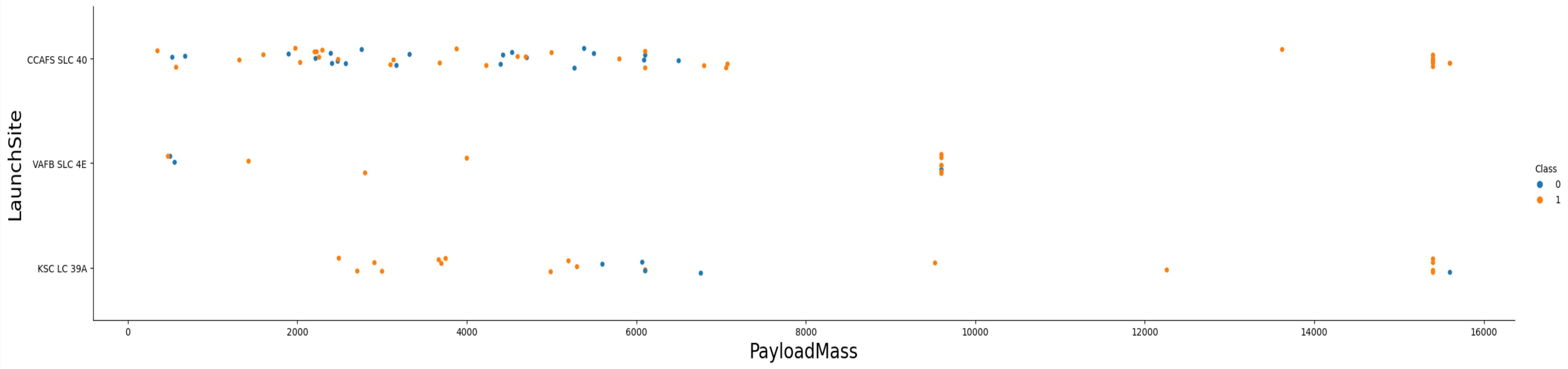
Flight Number vs. Launch Site



- Success Rate was more for higher flight Numbers
  - CCAF5 SLC 40 & KSC LC 39A: flight Numbers >80
  - VAFB SLC 4E: flight Numbers>50

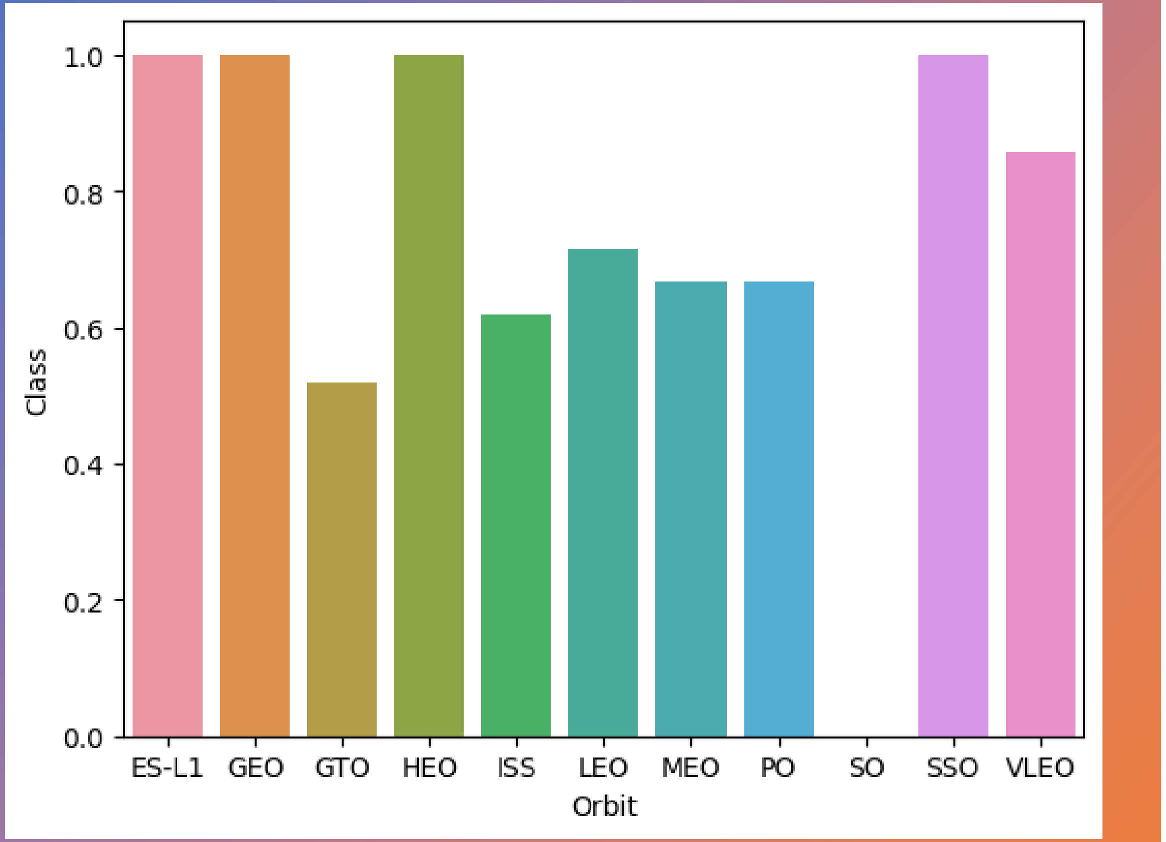
- CCAF5 SLC 40 has maximum range & number of flight Numbers & KSC LC 39A

Payload vs. Launch Site



- There are no rockets launched for heavypayload mass(greater than 10000) for VAFB-SLC launchsite
- Almost 100% success rate for all launch sites for PayLoadMass >70000

Success Rate vs. Orbit Type



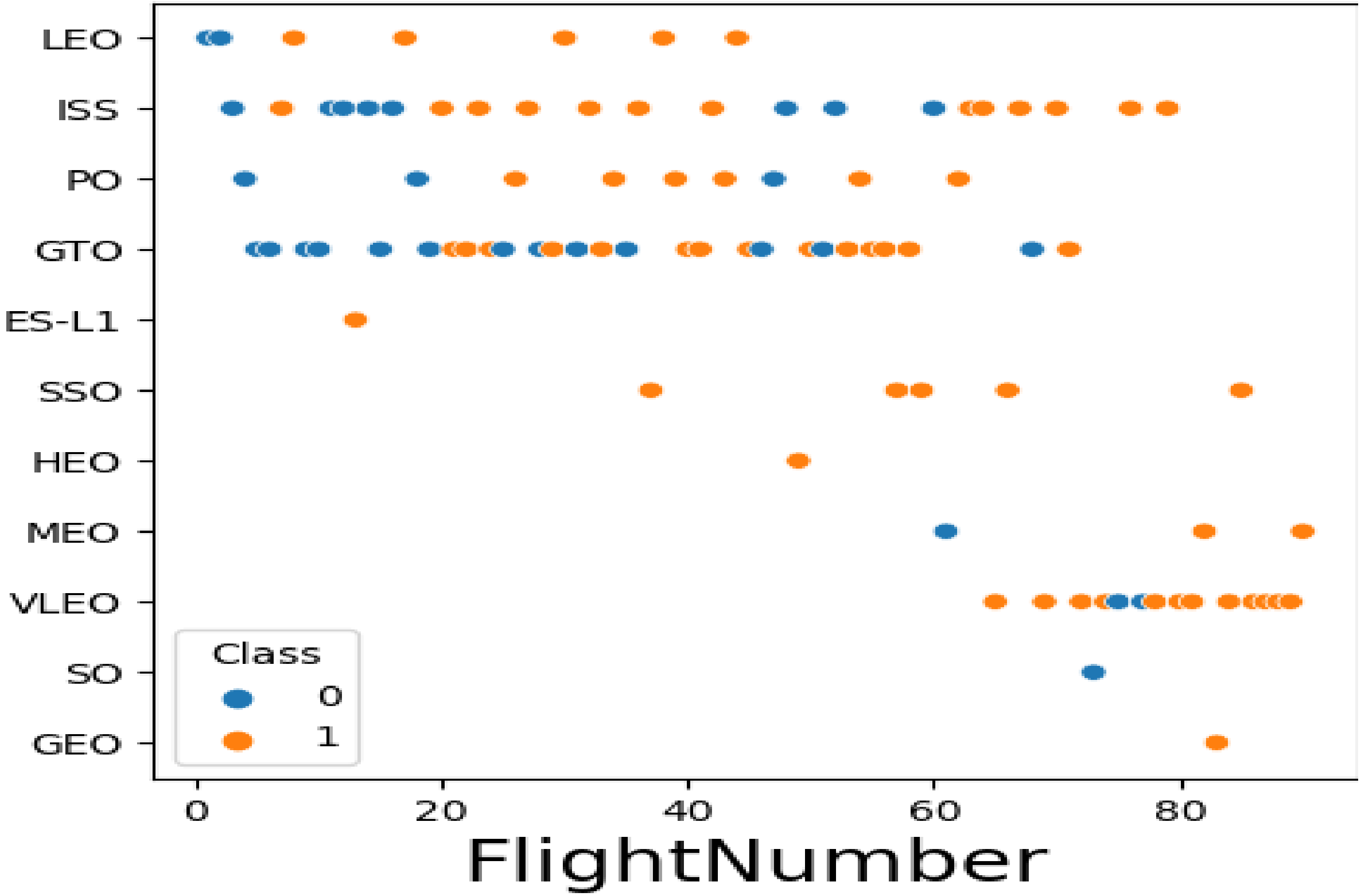
- ES-L1, GEO, HEO, SSO has the highest success rate

- SO, STO has the lowest success rates

Flight Number vs. Orbit Type

- LEO orbit the Success appears related to the number of flights

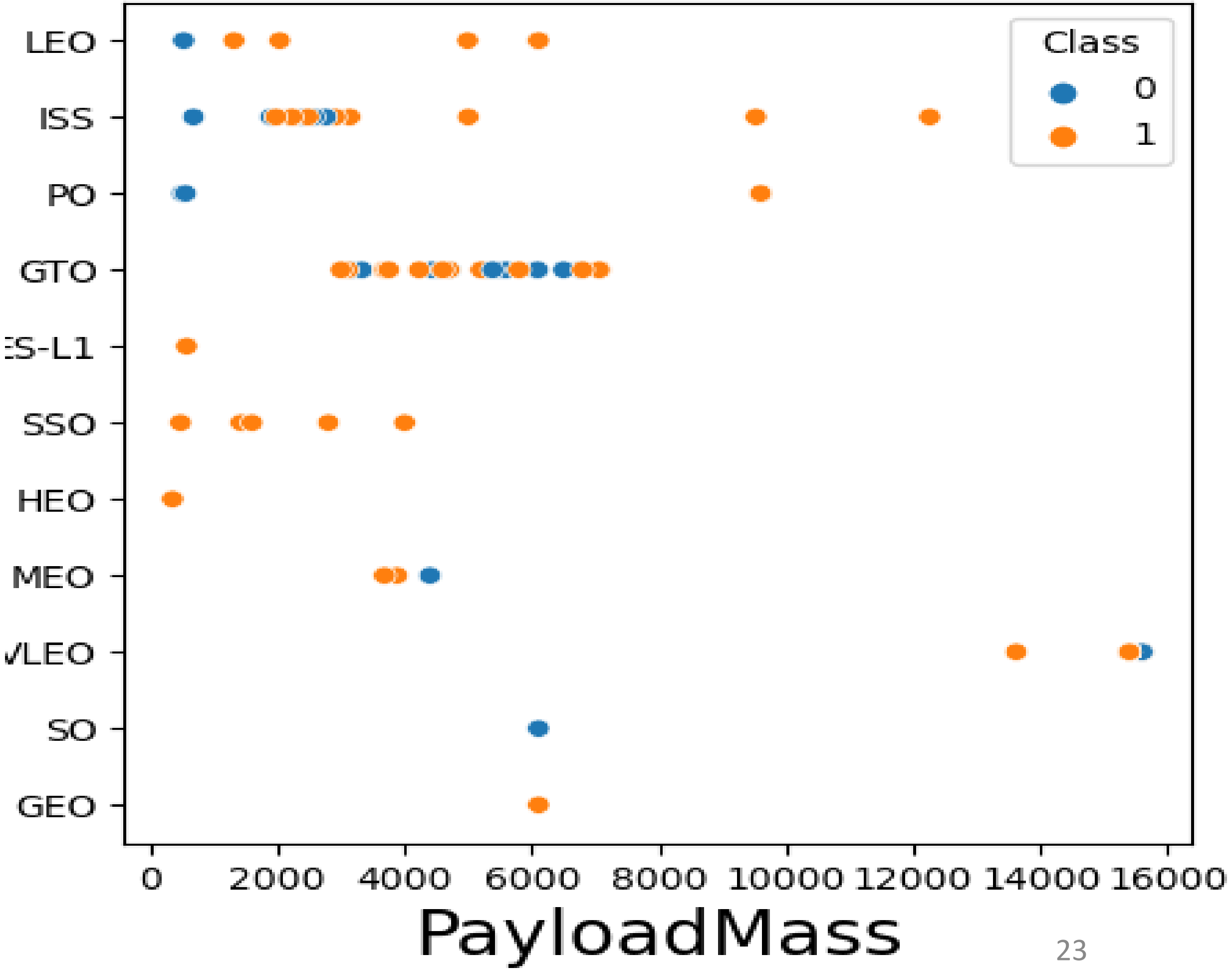
- There seems to be no relationship between flight number when in GTO orbit



Payload vs. Orbit Type

- With heavy payloads the successful landing rate are more for Polar, LEO and ISS.

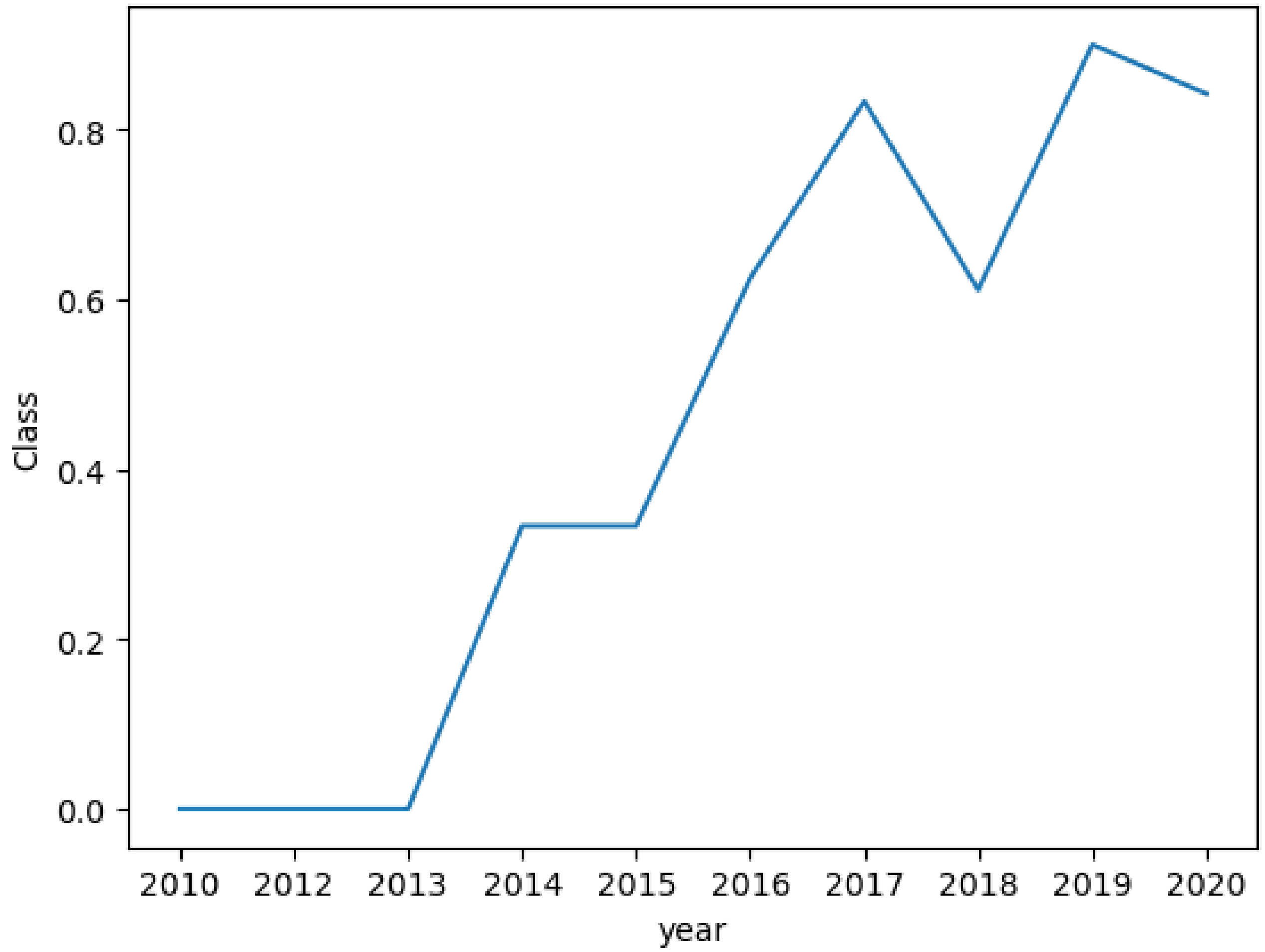
- In case of GTO, both positive landing rate and negative landing(unsuccessful mission) are there



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing

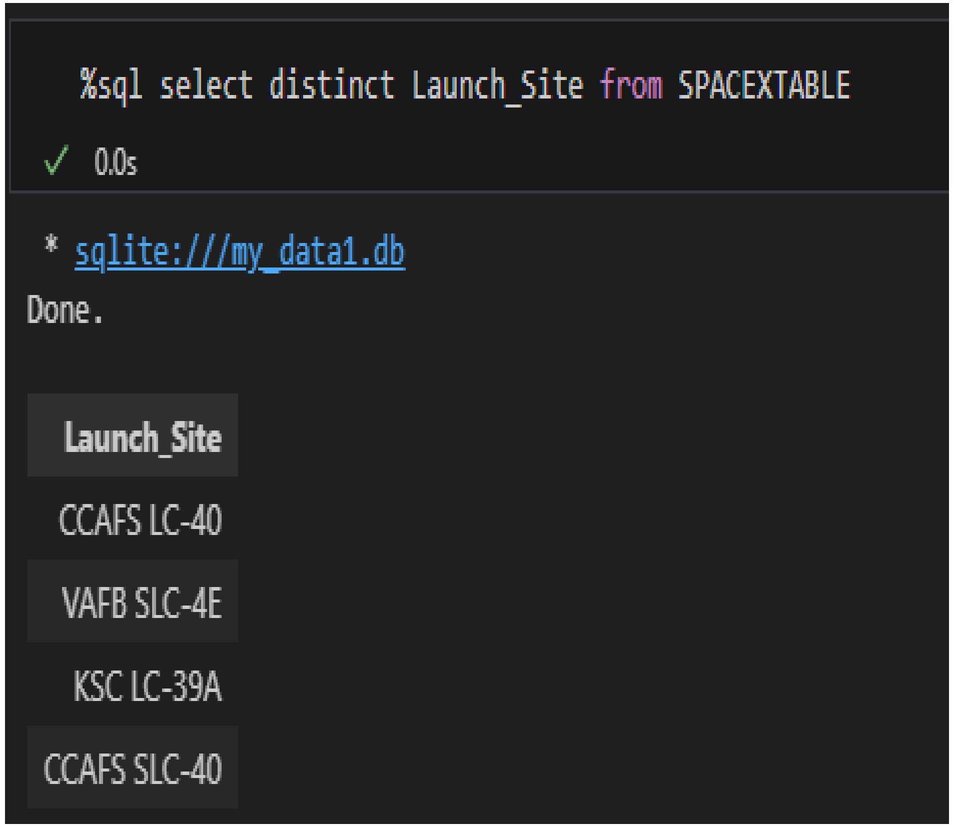
- 2019 has the highest success rate



All Launch Site Names

- There are total 4 unique launch sites

- Names of the unique launch sites are mentioned as besides using “Distinct” keyword in SQL query

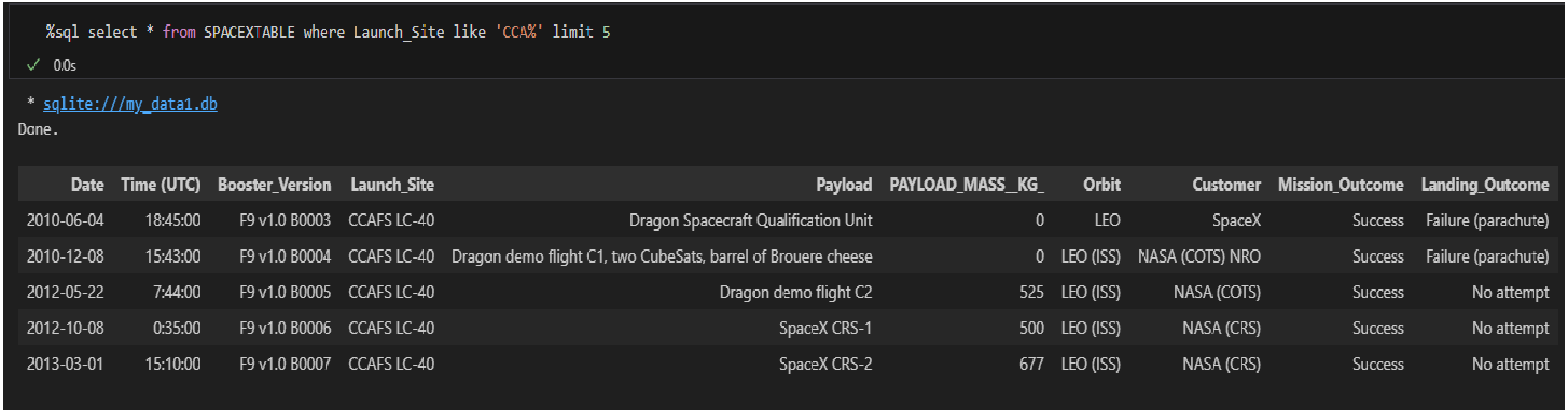


```
%sql select distinct Launch_Site from SPACEXTABLE
✓ 0.0s

* 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'



```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

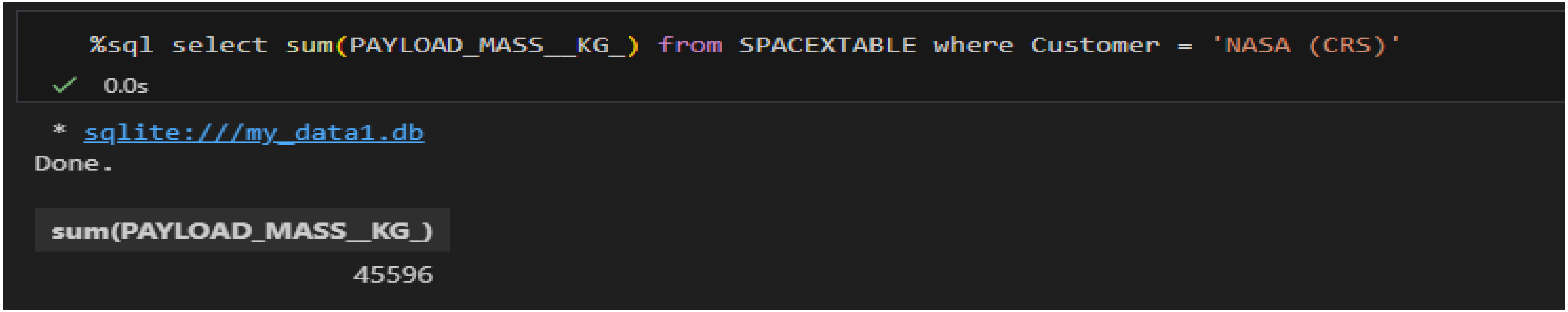
✓ 0.0s

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Sample 5 records whose launch sites begin with `CCA` are mentioned above using the “LIKE” keyword in the SQL query

Total Payload Mass



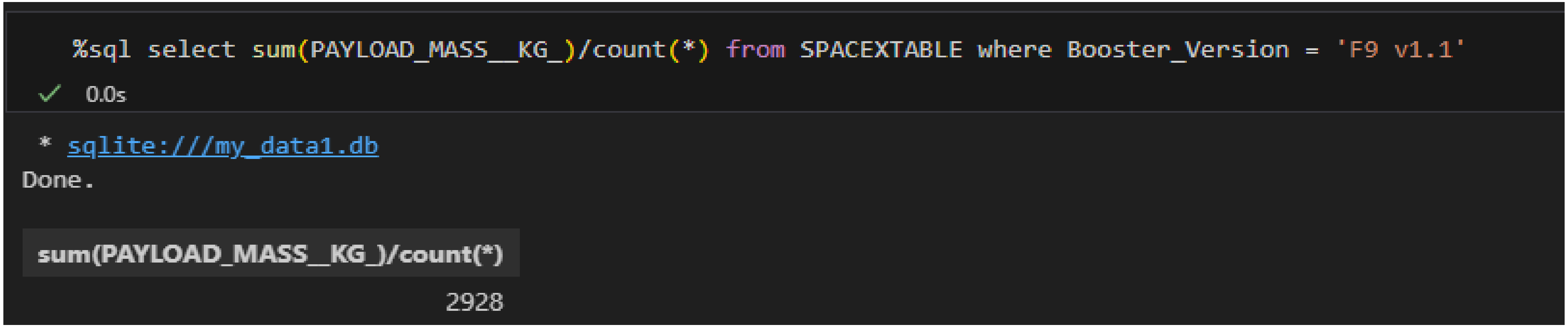
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
✓ 0.0s

* sqlite:///my_data1.db
Done.

sum(PAYLOAD_MASS__KG_)
                 45596
```

- Total payload carried by boosters from NASA(CRS) is 45596 Kg
- Query aggregated the PAYLOAD_MASS_KG column for only those records carried by boosters from NASA(CRS) using “WHERE” clause as filter

Average Payload Mass by F9 v1.1



```
%sql select sum(PAYLOAD_MASS__KG_)/count(*) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

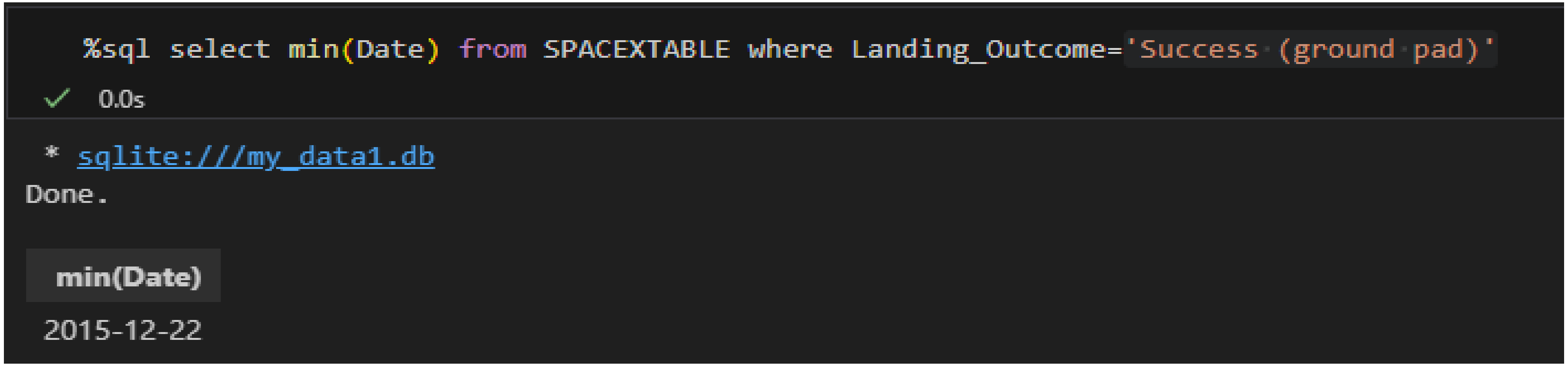
✓ 0.0s

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

sum(PAYLOAD_MASS__KG_)/count(*)
2928

- Average payload mass carried by booster version F9 v1.1 is 2928 Kgs
- Query average the total PAYLOAD_MASS_KG column per number of records using “WHERE” clause as filter to select records having Booster version F9 v1.1

First Successful Ground Landing Date



```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome='Success (ground pad)'
```

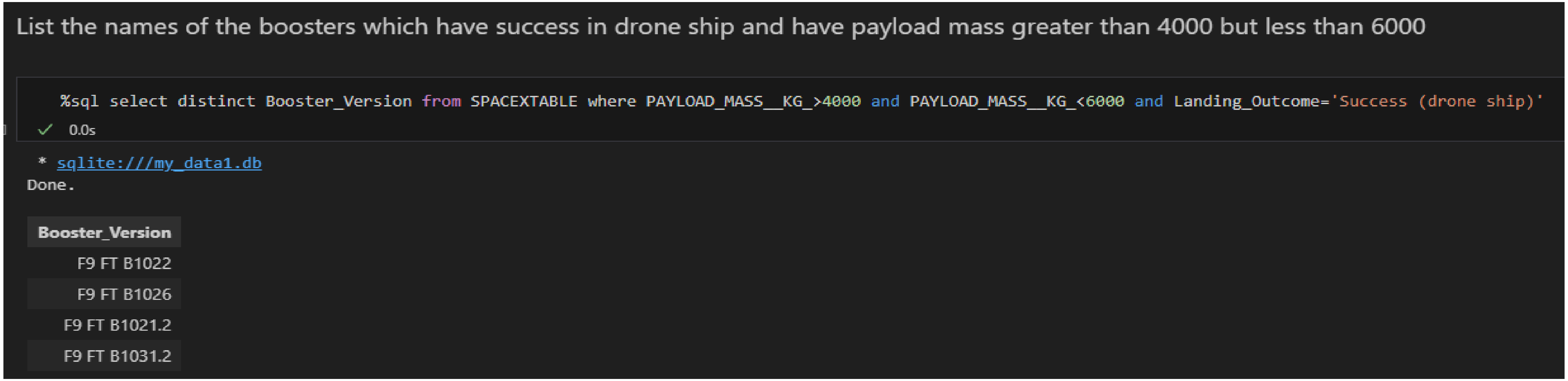
✓ 0.0s

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

min(Date)
2015-12-22

- Date of the first successful landing outcome on ground pad is 2015-12-22
- Query selected the minimum date from dataset, filtering records where Landing outcome was '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 distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000 and Landing_Outcome='Success (drone ship)'
```

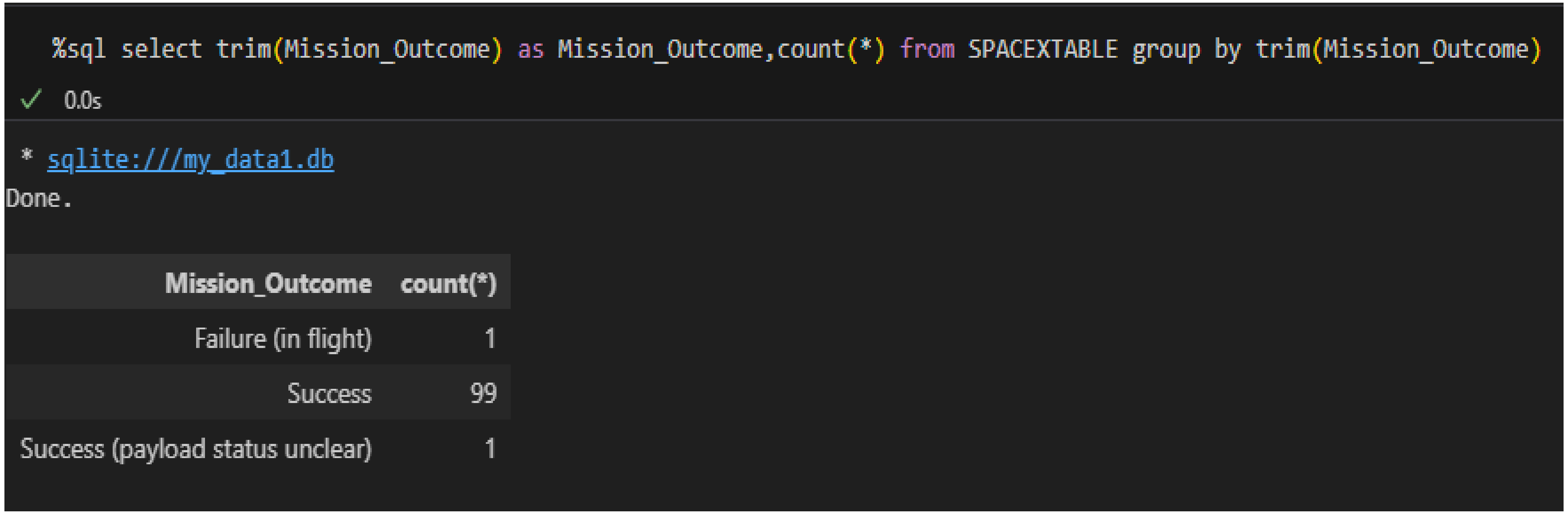
✓ 0.0s

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

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

- Query selected unique Booster versions with following filters in ‘WHERE’ clause:
    - payload mass greater than 4000 but less than 6000
    - Landing outcome was 'Success (drone ship)'

Total Number of Successful and Failure Mission Outcomes



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

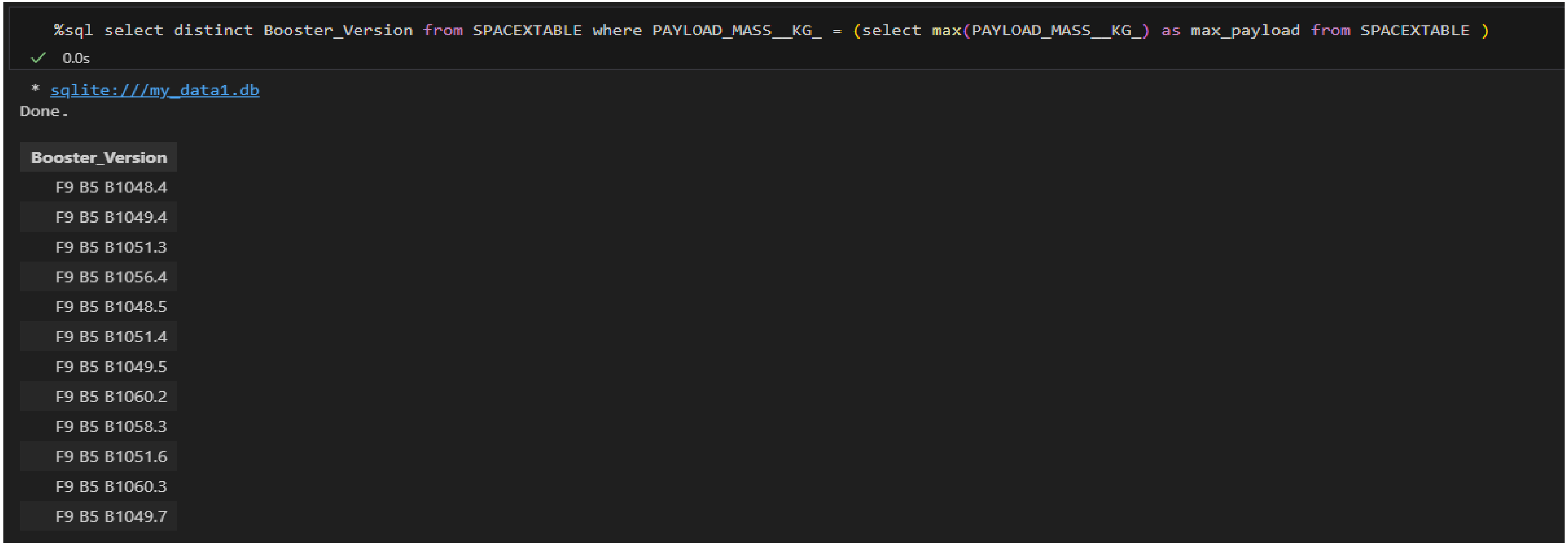
✓ 0.0s

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

Mission_Outcome	count(*)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Query grouped the record count based on the Mission Outcome column from the dataset

Boosters Carried Maximum Payload



```
%sql select distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) as max_payload from SPACEXTABLE )
```

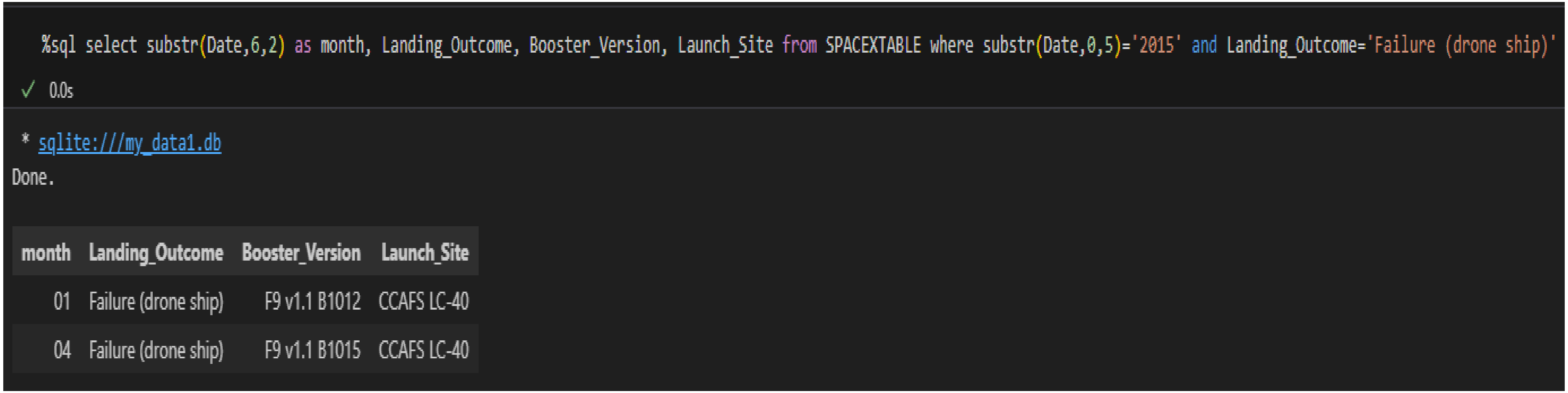
✓ 0.0s

```
 * 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
```

- Query selected the unique Booster Version values from the dataset based on filter criteria where PayLoadMass matches maximum PayLoadMass fetched by a subquery

2015 Launch Records



```
%sql select substr(Date,6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome='Failure (drone ship)'
```

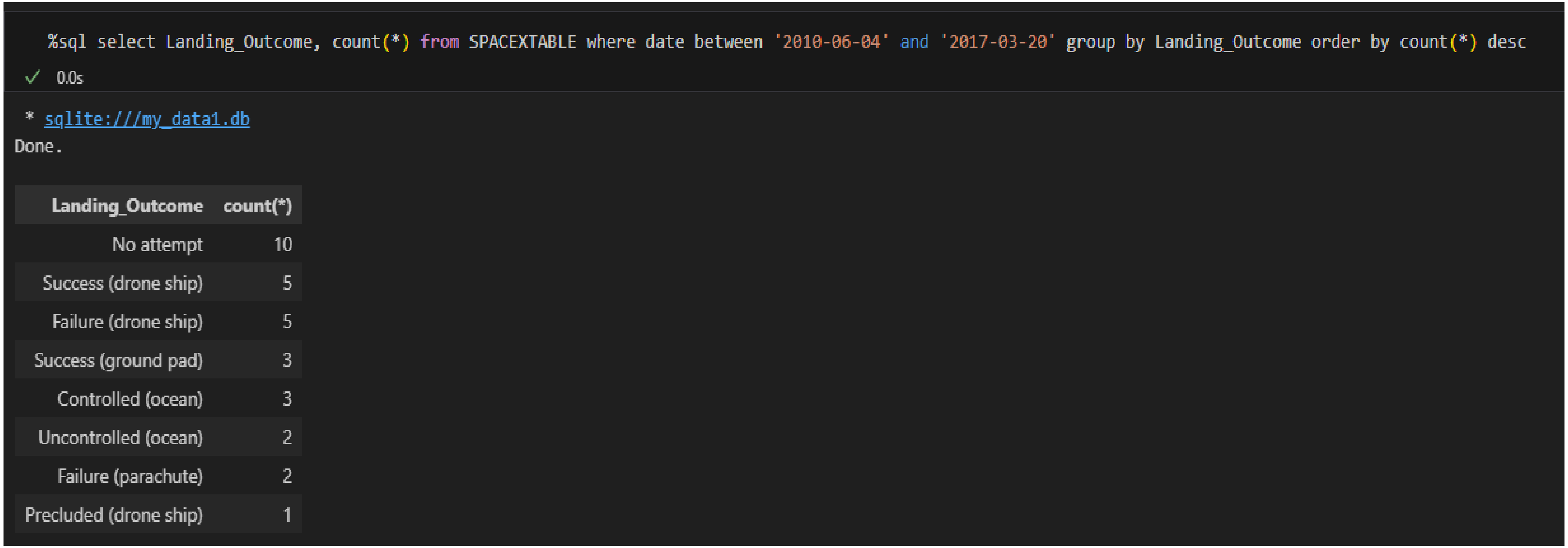
✓ 0.0s

```
 * 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

- Query selected the required columns from dataset based on following filter criteria:
  - Year was 2015
  - Landing outcome was ‘Failure (drone ship)'

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



```
%sql select Landing_Outcome, count(*) from SPACEXTABLE where date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(*) desc
✓ 0.0s
```

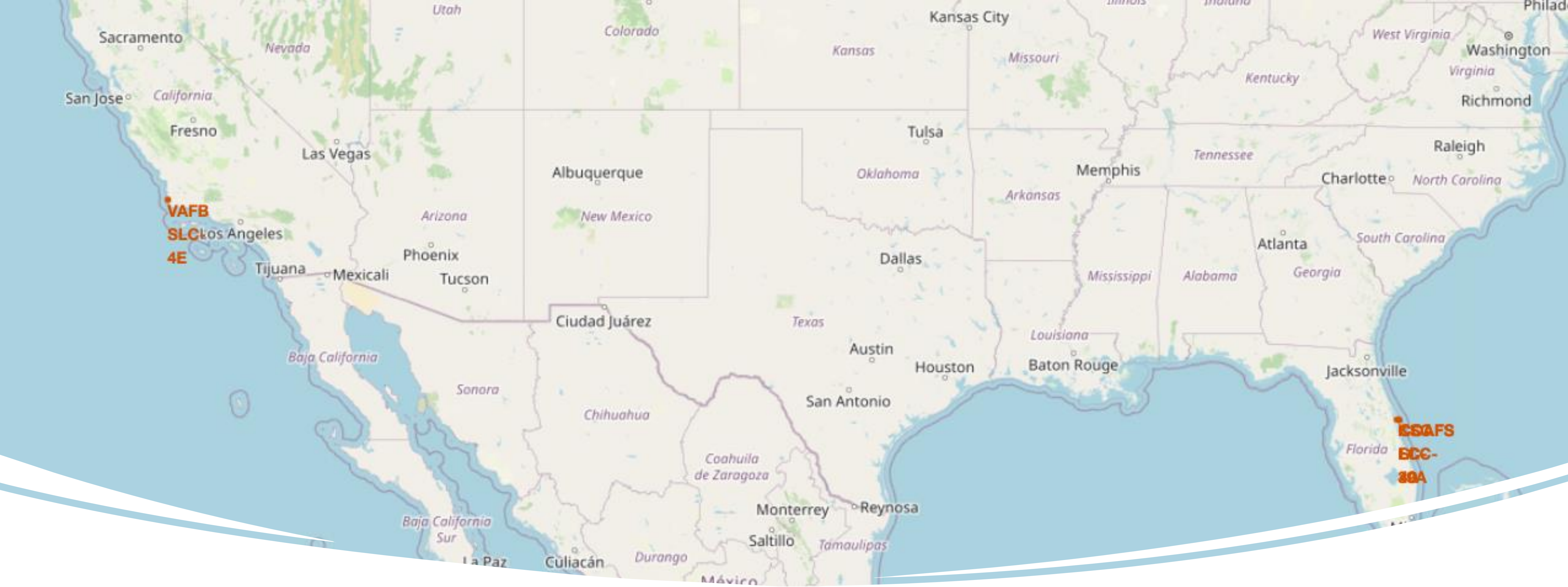
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count(*)
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

- Query grouped the record count based & ordered (in descending order) on Landing Outcome for only those records whose date ranges between the date 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

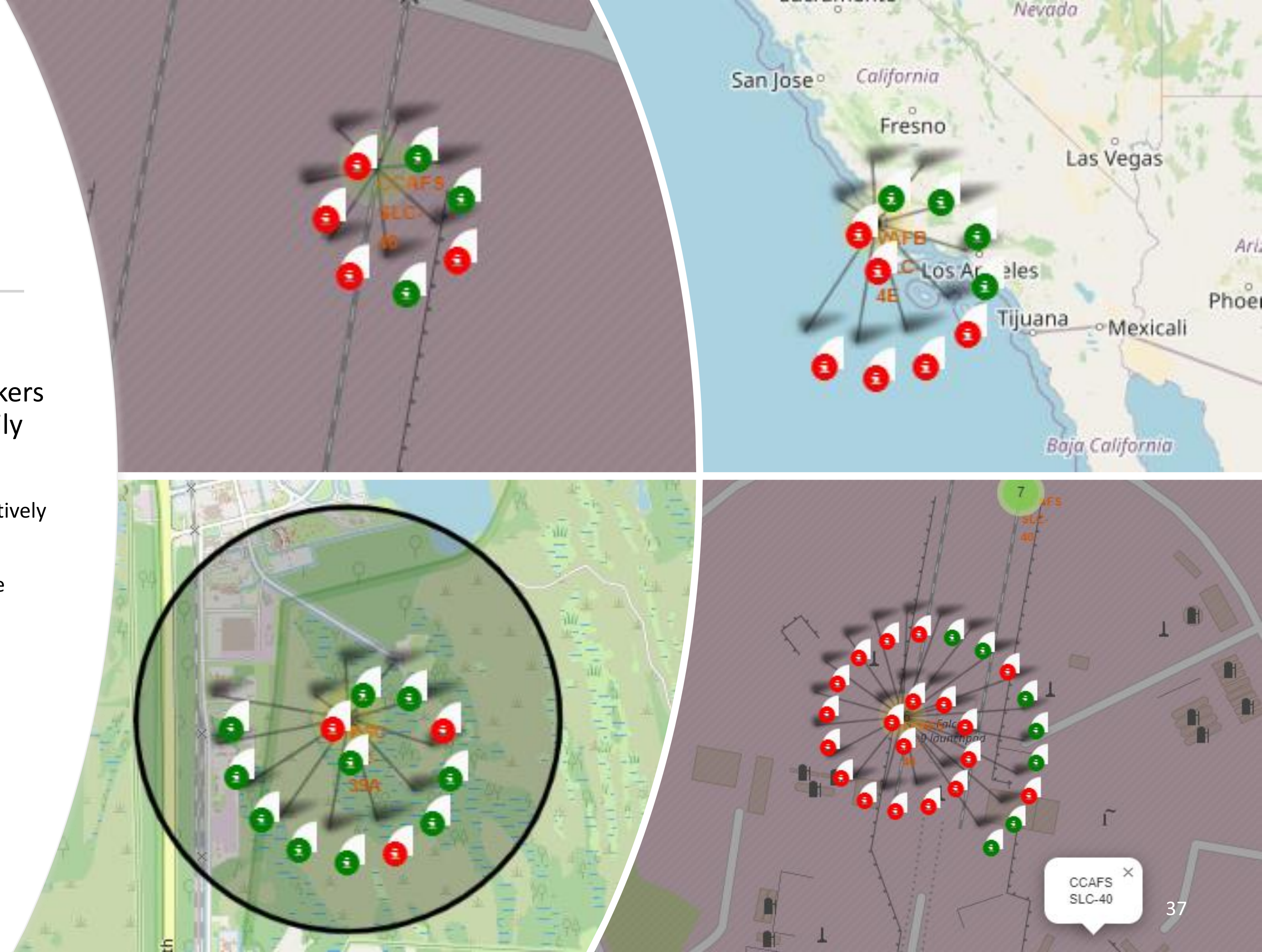


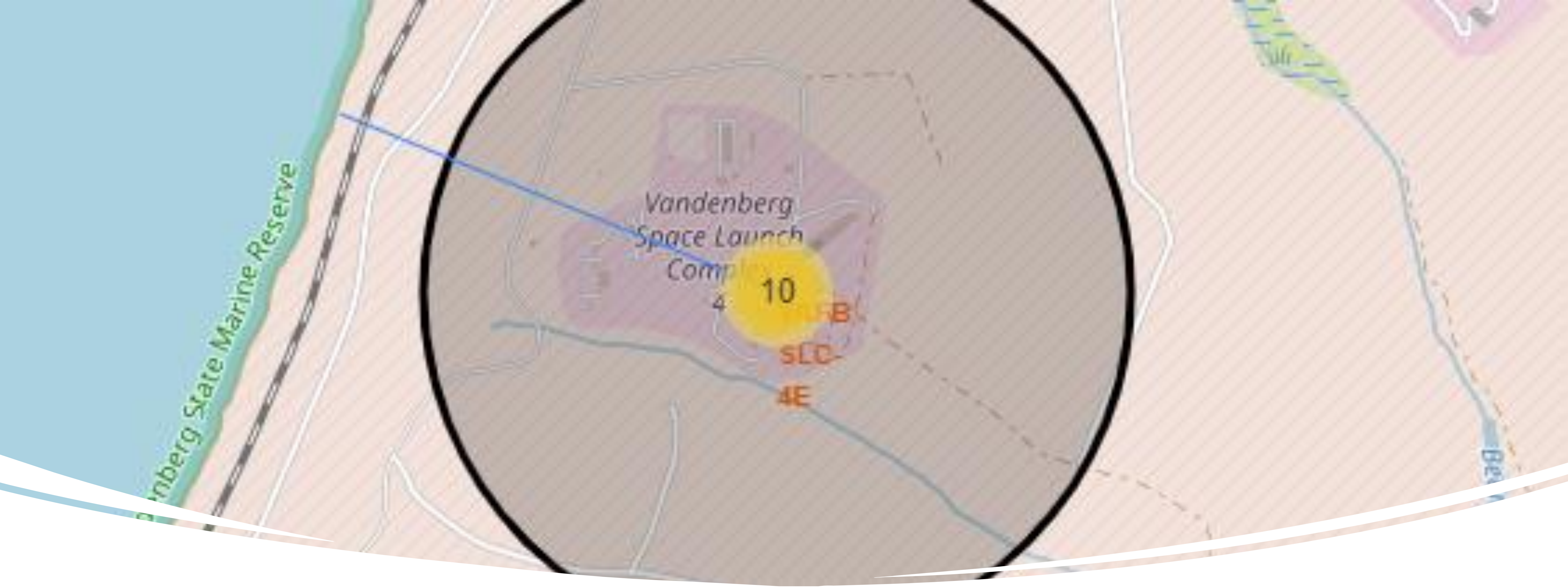
Launch Sites

- SpaceX's launch sites are near Florida & LA cities
- All launch sites' location are near costal region for safety regions

Launch Sites' Success Rates

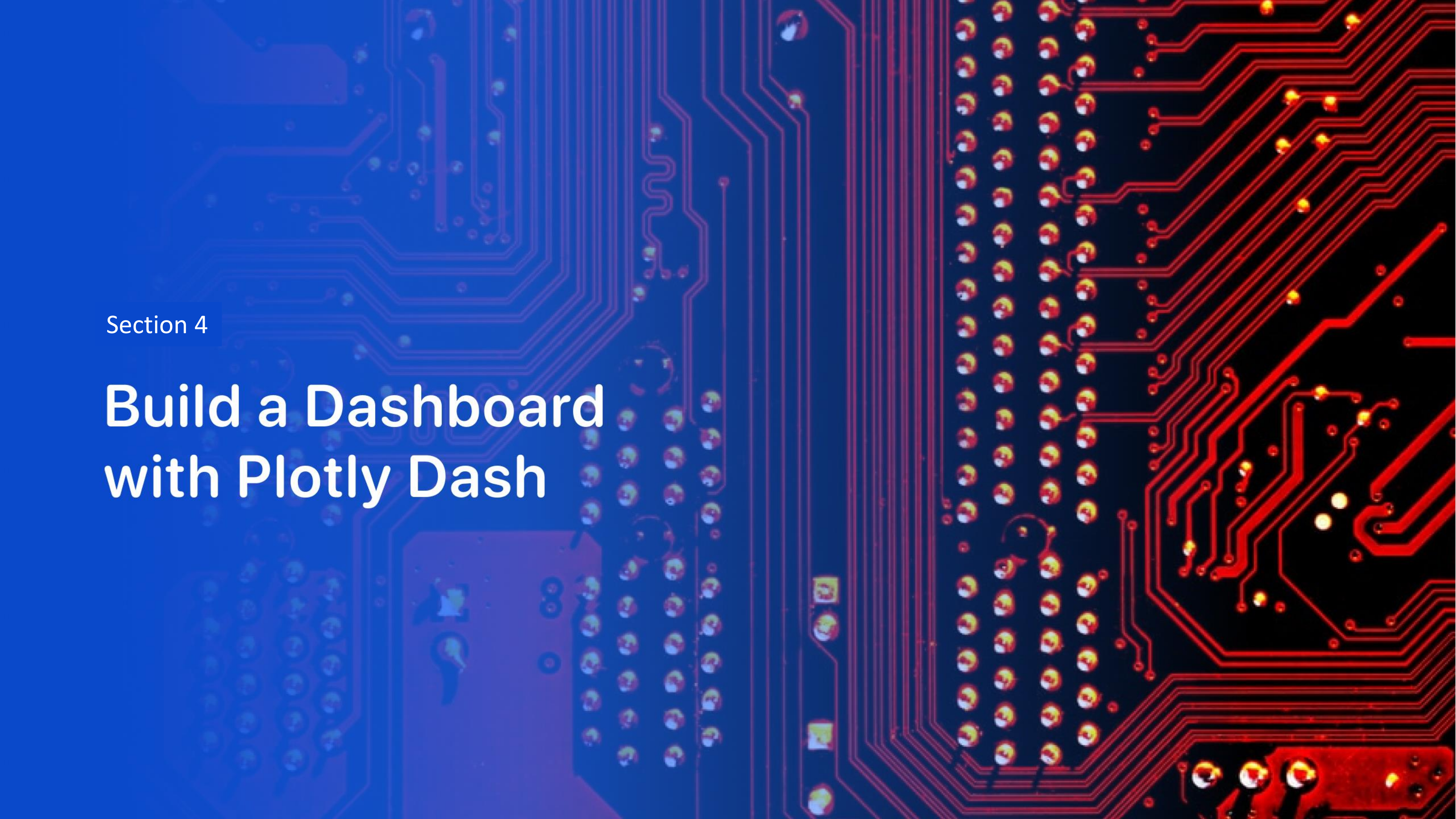
- From the colour- labelled markers in marker clusters, we can easily identify
  - KSC LC-39A launch site have relatively high success rates
  - CCAFS LC-40 has low success rate





Launch Site Proximities

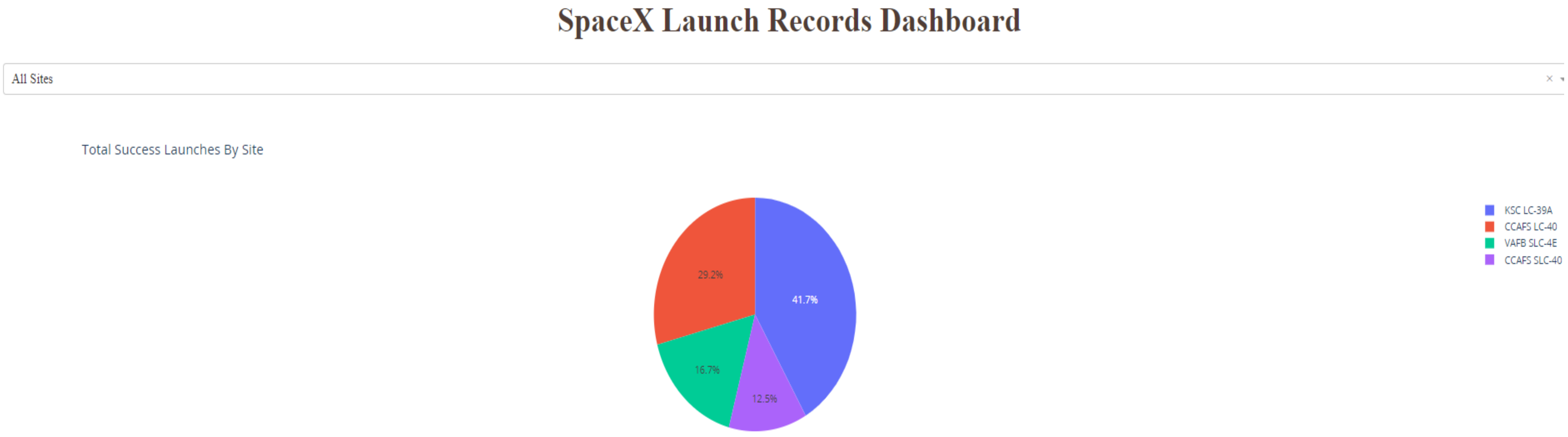
- Screenshot shows the proximities of launch site VAFB SLC- 4E to railway & coastal regions and far from city



Section 4

Build a Dashboard with Plotly Dash

Success Ratio by Launch Sites



SpaceX Launch Records Dashboard

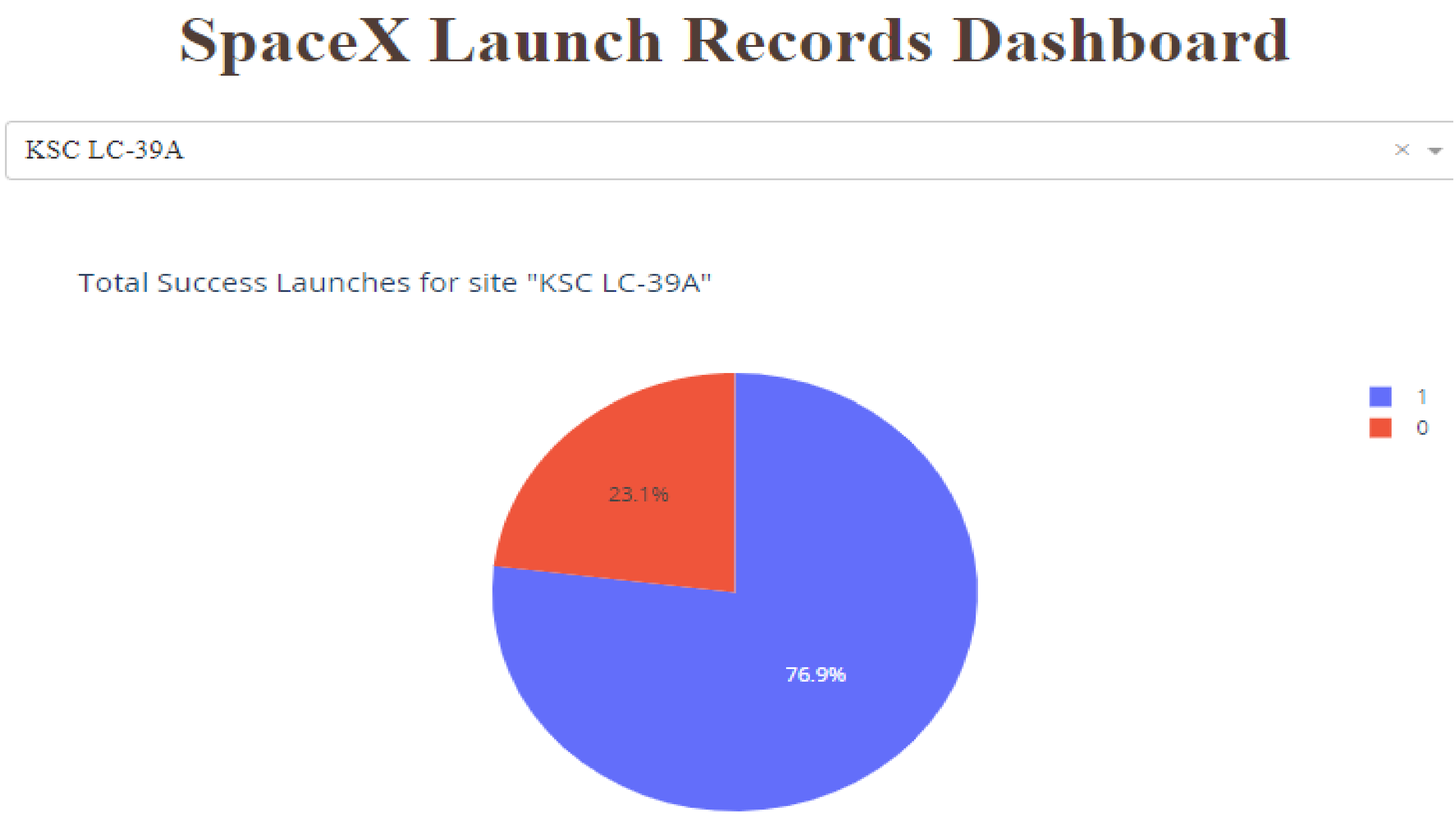
All Sites ×

Total Success Launches By Site

- KSC LC-39A has the highest success ratio of 41.7% while CCAFS SLC-40 has the lowest success rate of 12.5%

Highest Success Rate

- KSC LC-39A has the highest success rate of **76.9%**

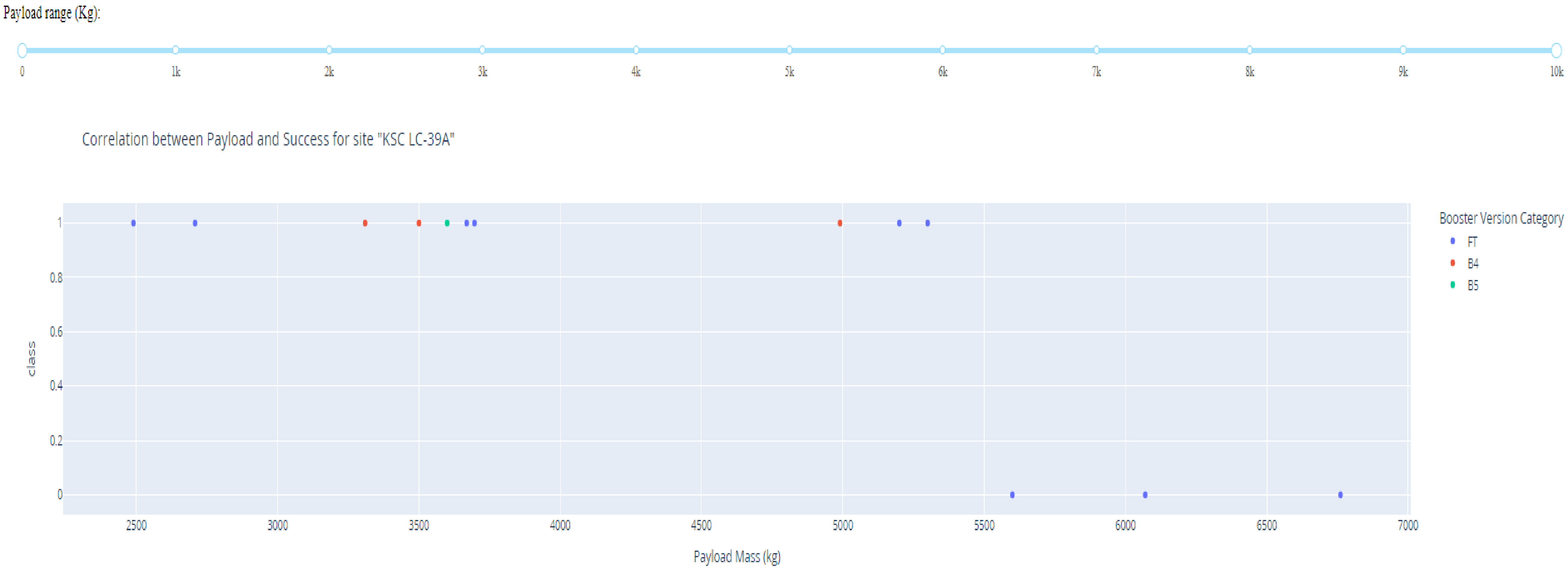


SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site "KSC LC-39A"

Payload Vs Success Rate



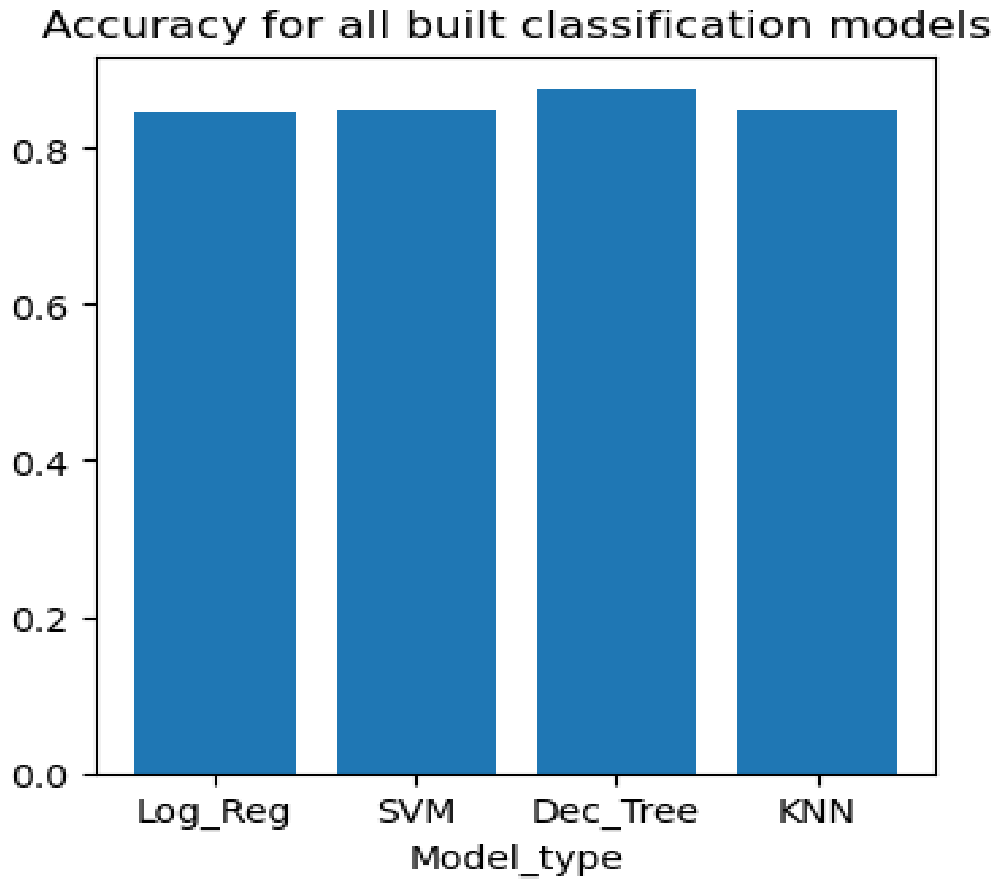
- Payload under 5500Kgs have high success Rate across all Booster Version Categories

Section 5

Predictive Analysis (Classification)

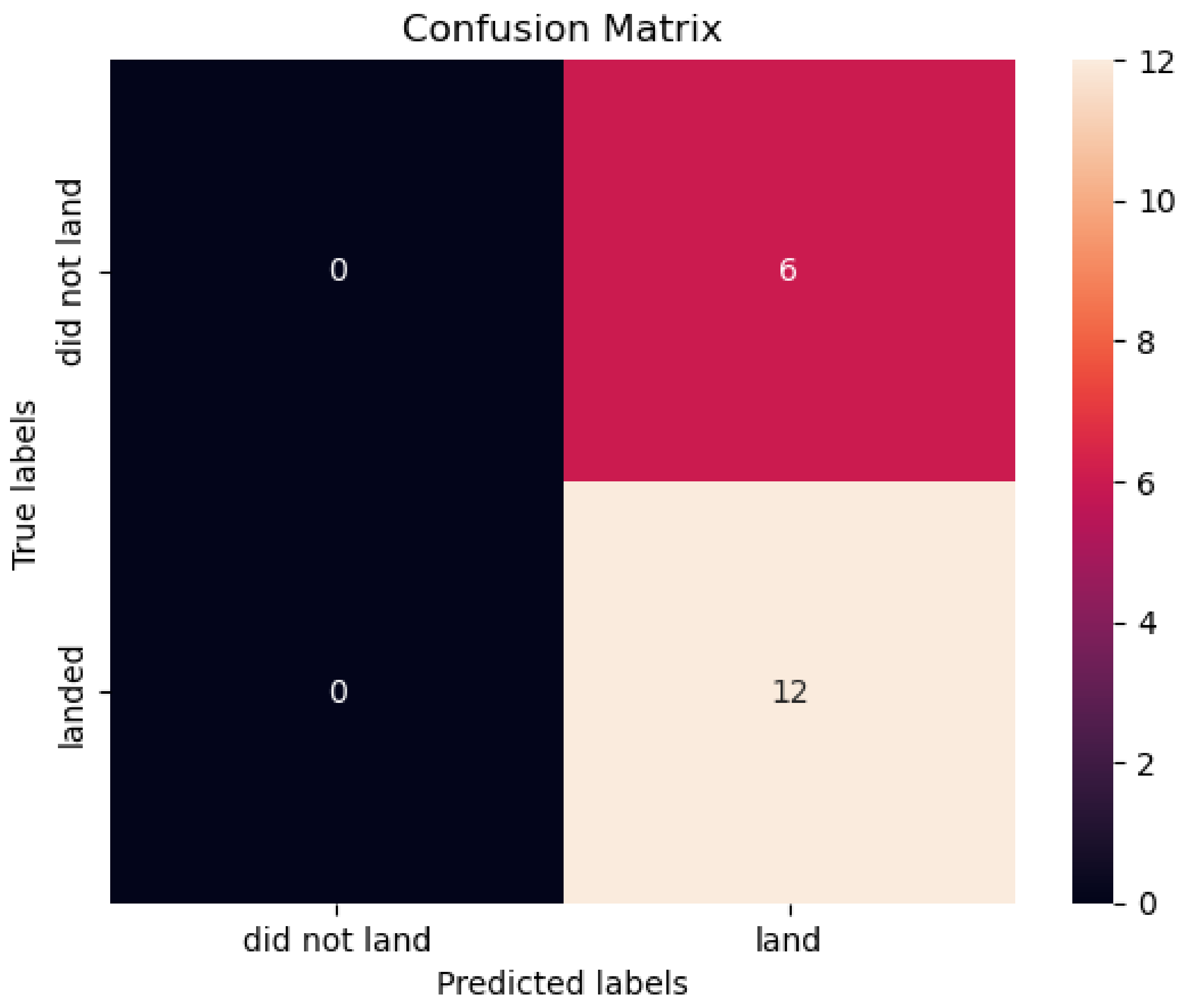
Classification Accuracy

- Decision Tree Model has the highest classification accuracy of 87.5%



Confusion Matrix

- Confusion matrix of the Decision Tree is very good at predicting success landings
- But bad at predicting the failure landings



Conclusions

- Success Rate was more for higher flight Numbers & PayLoadMass >70000
- ES-L1, GEO, HEO, SSO has the highest success rate, while STO has the lowest success rates
- For LEO orbit, the Success appears related to the number of flights while there seems to be no relationship between flight number when in GTO orbit
- With heavy payloads the successful landing rate are more for Polar, LEO and ISS
- In case of GTO, both positive landing rate and negative landing(unsuccessful mission) are there
- Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing & 2019 has the highest success rate
- KSC LC-39A launch site have relatively high success rates
- Payload under 5500Kgs have high success Rate across all Booster Version Categories
- The best model for predictions of success landings is Decision Tree Model

Thank you!