IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Houssam Eddine Ariche

Nov 22, 2025

Outline

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

Executive Summary

In this capstone project, we aim to predict whether the SpaceX Falcon 9 first stage will successfully land using various machine learning classification algorithms.

The project involves several key steps:

- Data collection, cleaning, and preparation
- Exploratory data analysis (EDA)
- Interactive data visualization
- Machine learning model development and prediction

Our analyses indicate that certain features of rocket launches are correlated with the launch outcomes—success or failure. Based on the results, the Decision Tree algorithm appears to be the most effective model for predicting the successful landing of the Falcon 9 first stage.

Introduction

- Since SpaceX is able to reuse the first stage of its Falcon 9 rocket, the launch cost is approximately **$62 million**, whereas launches from other providers can cost **upwards of $165 million** each.

- So, if we can determine whether the first stage will land, we can determinethe cost of a launch. This information can be used if an alternate companywants to bid against SpaceX for a rocket launch.

- Can we predict a new launch with success of the first stage landingaccording to the historical launch data?
- Can we tell what is the best choice for a successful launch?

Section 1

Methodology

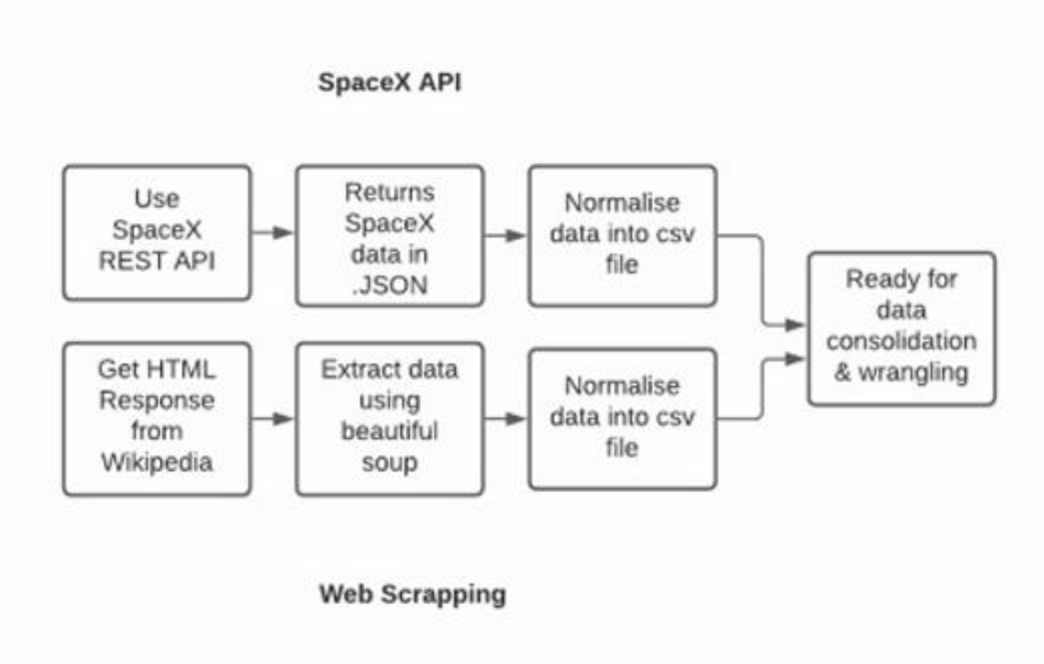
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping (Wikipedia)
- Perform data wrangling
  - Generate landing Class from Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearchCV to find best fit model

Data Collection

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. As mentioned, the dataset was collected by REST API and WebScrapping from Wikipedia


SpaceX API
Use SpaceX REST API
Returns SpaceX data in .JSON
Normalise data into csv file
Ready for data consolidation & wrangling
Get HTML Response from Wikipedia
Extract data using beautiful soup
Normalise data into csv file
Web Scrapping

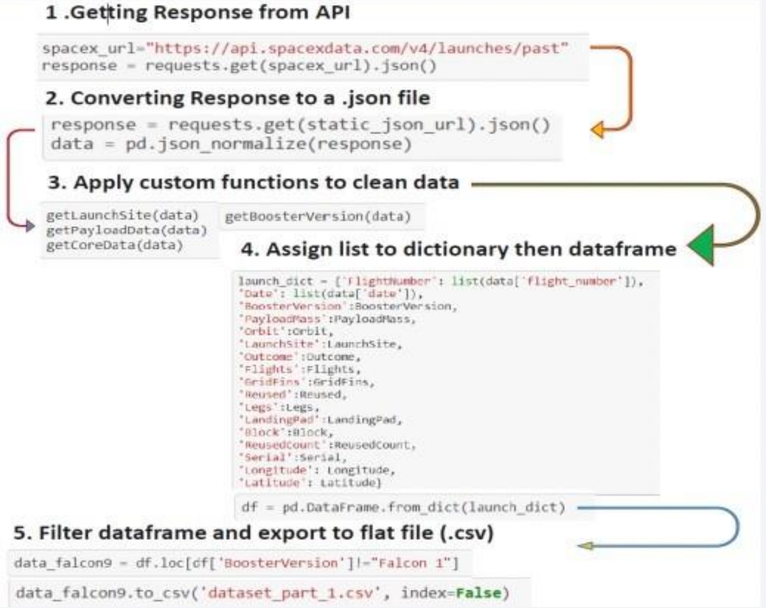
For REST API, its started by using the get request. Then, we decoded the response content as Json and turn it into a pandas dataframe using json_normalize(). We then cleaned the data, checked for missing values and fill with whatever needed.

For web scrapping, we will use the BeautifulSoup to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for further analysis

Data Collection – SpaceX API

- Data collection is the process of gathering data from available sources. This data can be structured, unstructured, or semi-structured. For this project, data was collected via SpaceX API and Web scrapping Wiki pages for relevant launch data.

- Get request for rocket launchdata using API
- Use json_normalize method toconvert json result to dataframe
- Performed data cleaning andfilling the missing value



1 .Getting Response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url).json()
```

2. Converting Response to a .json file

```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

3. Apply custom functions to clean data

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

```
getBoosterVersion(data)
```

4. Assign list to dictionary then dataframe

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

```
df = pd.DataFrame.from_dict(launch_dict)
```

5. Filter dataframe and export to flat file (.csv)

```
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
```

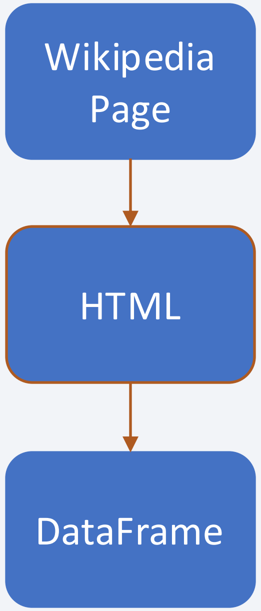
```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

The link to the notebook is https://github.com/HoussamEdar/Applied-Data-Science-Capstone/blob/main/1_spacex_data_collection_api.ipynb

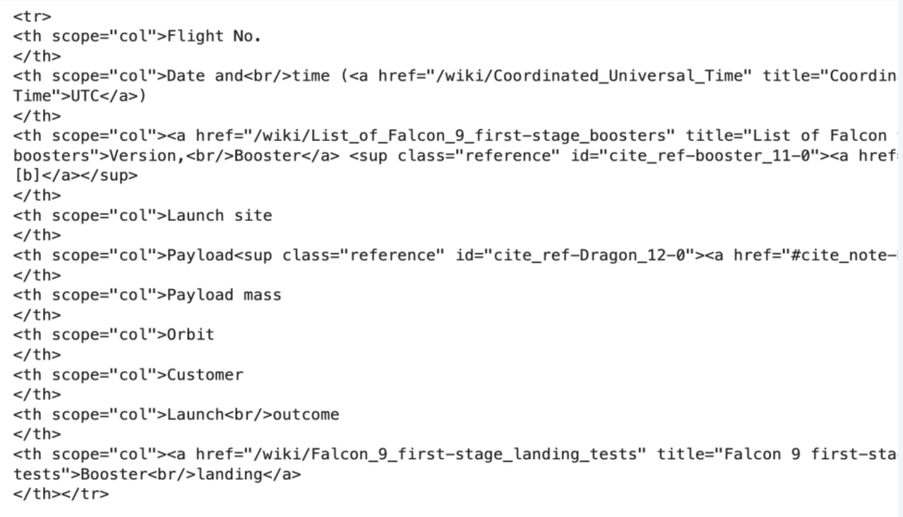
Data Collection - Scraping

- Wikipedia Falcon Page
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- My Notebook
  https://github.com/HoussamEdar/Applied-Data-Science-Capstone/blob/main/2_webscraping_spacex_launches_from_wikipedia.ipynb



```
import requests
from bs4 import BeautifulSoup
url =
'https://en.wikipedia.org/wiki/List_of_Falcon
_9_and_Falcon_Heavy_launches'
response = requests.get(url)
html_data = response.text
soup = BeautifulSoup(html_data)
```

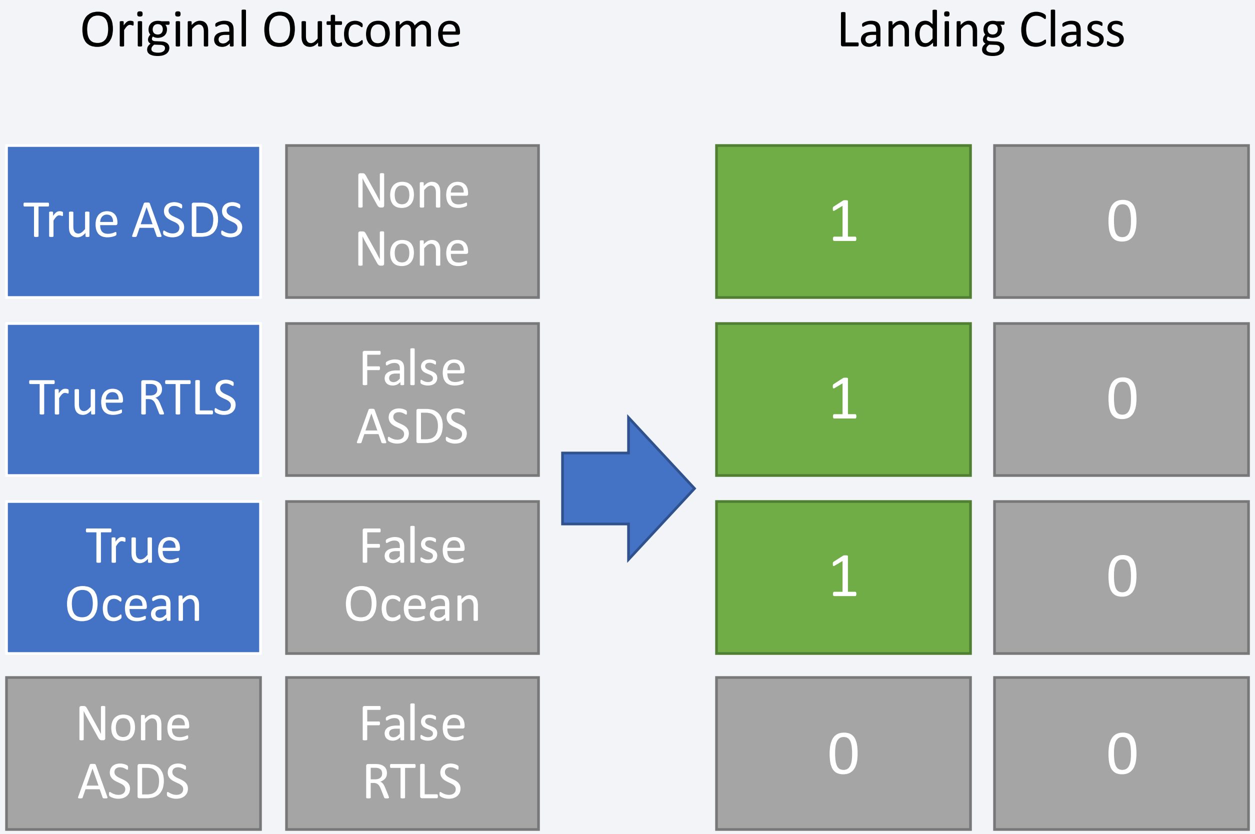


```
<tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordin
Time">UTC</a>)
</th>
<th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Falcon
boosters">Version,<br/>Booster</a> <sup class="reference" id="cite_ref-booster_11-0"><a href
[b]</a></sup>
</th>
<th scope="col">Launch site
</th>
<th scope="col">Payload<sup class="reference" id="cite_ref-Dragon_12-0"><a href="#cite_note-
</th>
<th scope="col">Payload mass
</th>
<th scope="col">Orbit
</th>
<th scope="col">Customer
</th>
<th scope="col">Launch<br/>outcome
</th>
<th scope="col"><a href="/wiki/Falcon_9_first-stage_landing_tests" title="Falcon 9 first-sta
tests">Booster<br/>landing</a>
</th></tr>
```

Data Wrangling

- **My Notebook**
  https://github.com/HoussamEdar/Applied-Data-Science-Capstone/blob/main/3_spacex_data_wrangling.ipynb

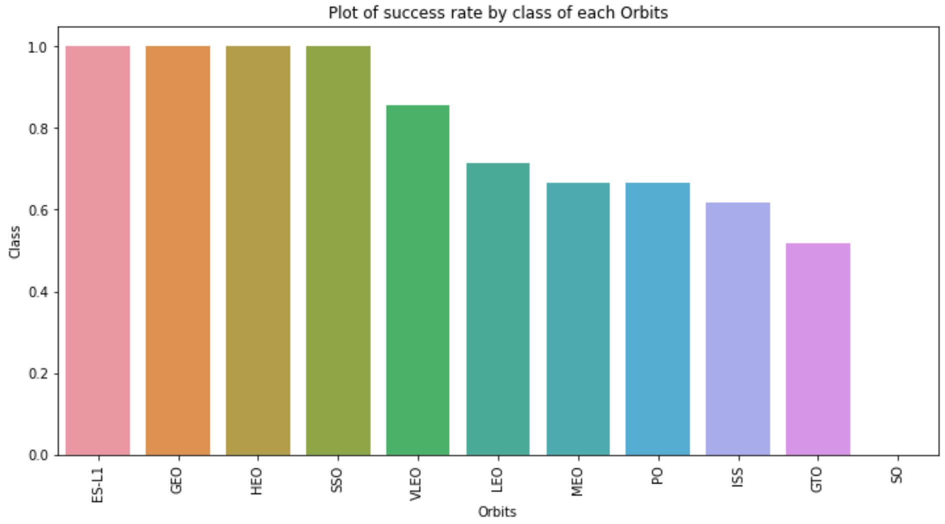
- Transform raw data to **useful** data. For example, convert original outcome labels into landing class that represent landing classification which will be our new landing prediction target.
  - 1 for success
  - 0 for failure

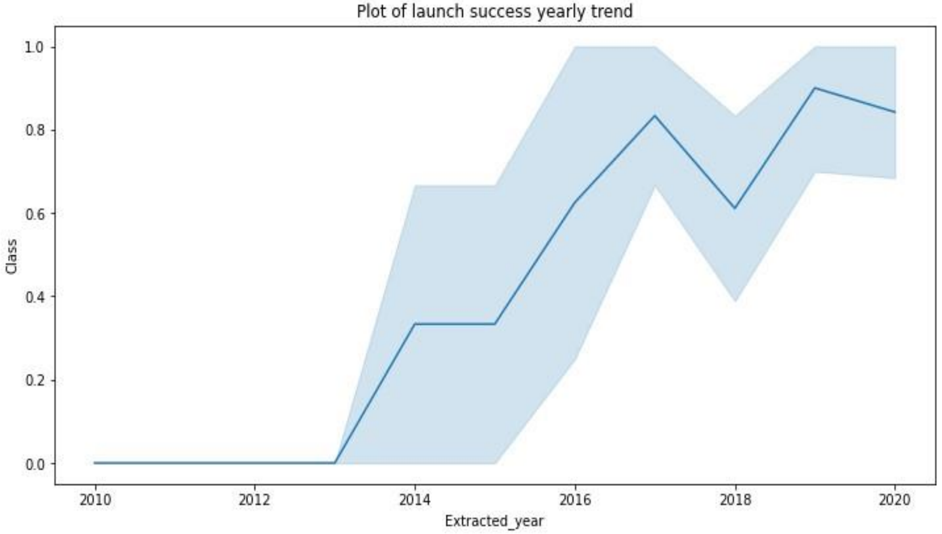


Original Outcome		Landing Class	
True ASDS	None None	1	0
True RTLS	False ASDS	1	0
True Ocean	False Ocean	1	0
None ASDS	False RTLS	0	0

EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.
- The link to the notebook is https://github.com/HoussamEdar/Applied-Data-Science-Capstone/blob/main/4_eda_data_visualization.ipynb





EDA with SQL

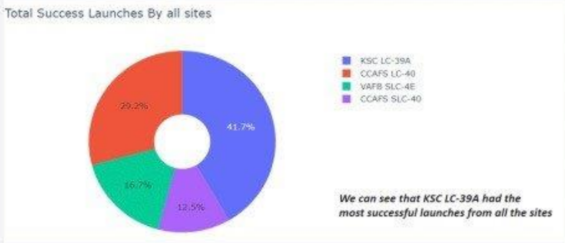
- We loaded the SpaceX dataset into a PostgreSQL database without leaving thejupyter notebook.
- We applied EDA with SQL to get insight from the data. We wrote queries tofind out for instance:
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.
- The link to the notebook is https://github.com/HoussamEdar/Applied-Data-Science-Capstone/blob/main/5_eda_using_sql.ipynb

Build an Interactive Map with Folium

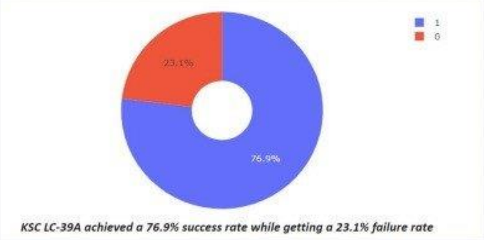
- We marked all launch sites, and added map objects such asmarkers, circles, lines to mar k the success or failure oflaunches for each site on the folium map.
- We assigned the feature launch outcomes (failure orsuccess) to class 0 and 1.i.e., 0 for failure,and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to itsproximities. We answered so me question for instance:
  - Are launch sites near railways, highways and coastlines.
  - Do launch sites keep certain distance away from cities.
  - The link to the notebook is https://github.com/HoussamEdar/Applied-Data-Science-Capstone/blob/main/6_launch_site_location_visualization_with_folium.ipynb

Build a Dashboard with Plotly Dash

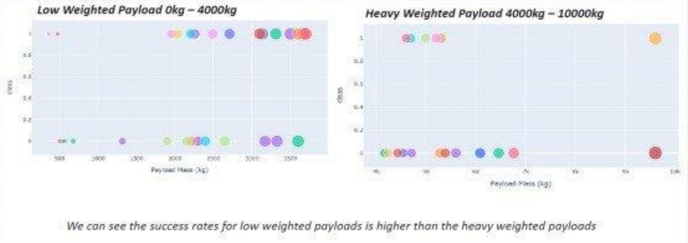
- We built an interactive dashboardwith Plotly dash
- We plotted pie charts showing thetotal launches by a certain sites
- We plotted scatter graph showingthe relationship with Outcome andPayload Mass (Kg) for the differentbooster version.
- The link to the notebook https://github.com/HoussamEdar/Applied-Data-Science-Capstone/blob/main/7_spacex_data_intereactive_dashboard_app.py



We can see that KSC LC-39A had the most successful launches from all the sites



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate



We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our datainto training and testing.
- We built different machine learning models and tune different hyperparametersusing GridSearchCV.
- We used accuracy as the metric for our model, improved the model using featureengineering and algorithm tuning.
- We found the best performing classification model.
- The link to the notebook is https://github.com/HoussamEdar/Applied-Data-Science-Capstone/blob/main/8_spacex_machine_learning_prediction.ipynb

Results

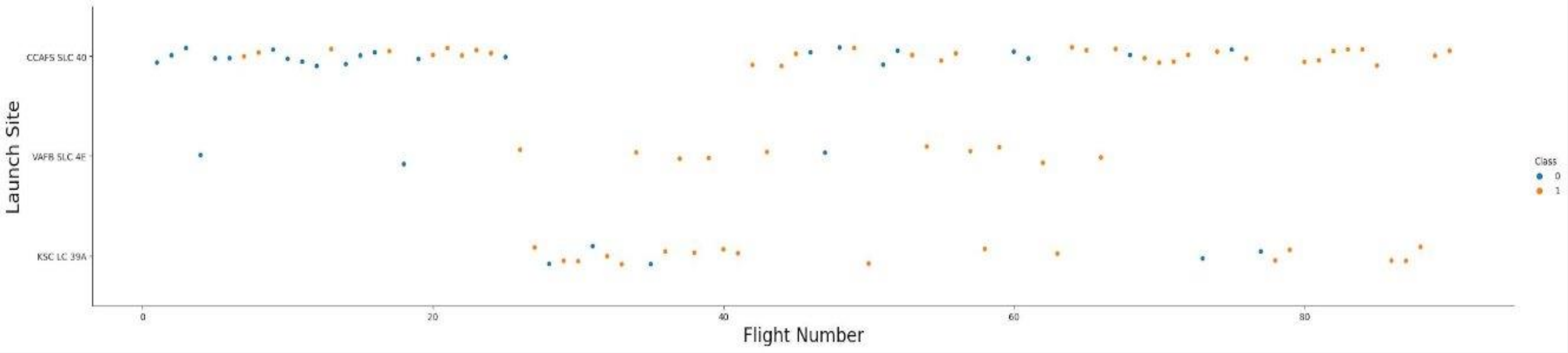
- The larger the flight amount at a launch site, the greater the success rate ata la unch site
- Low weighted payloads perform better than the heavier payloads.
- Launch success rate started to increase in 2013 till 2020.
- KSC LC 39A had the most successful launches from all the sites.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- The Decision tree classifier is the best machine learning algorithm for thistas k.

Section 2

Insights drawn from EDA

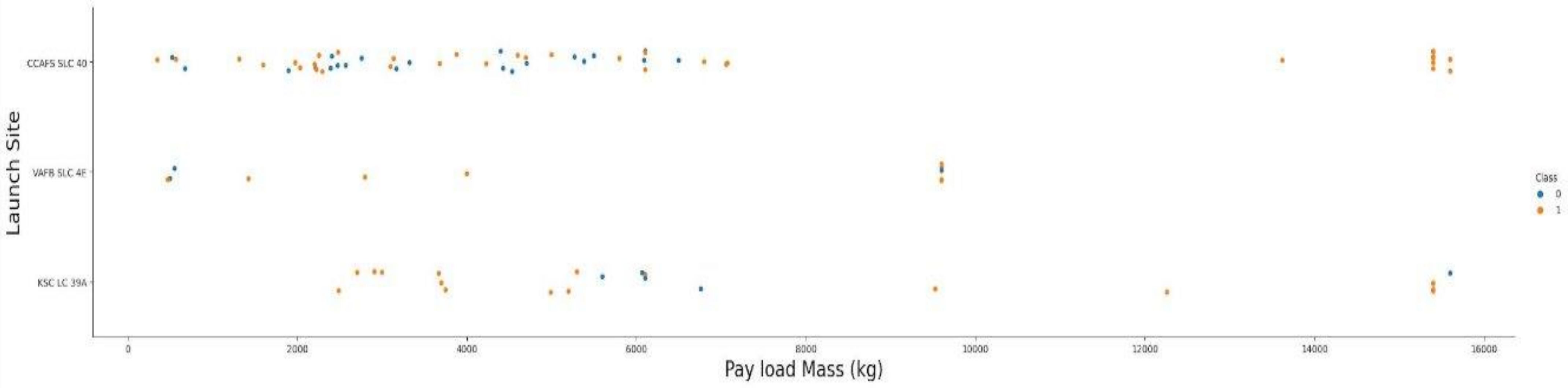
Flight Number vs. Launch Site

- From the figure below, it can be noted that the launch site CCAFS SLC 40 has launched the highest number of rockets compared to the other sites.
- Also, it is shown that the later flights from launch sites VAFB SLC 4E and KSC LC 39A showed a higher success rate compared to the earlier flights.



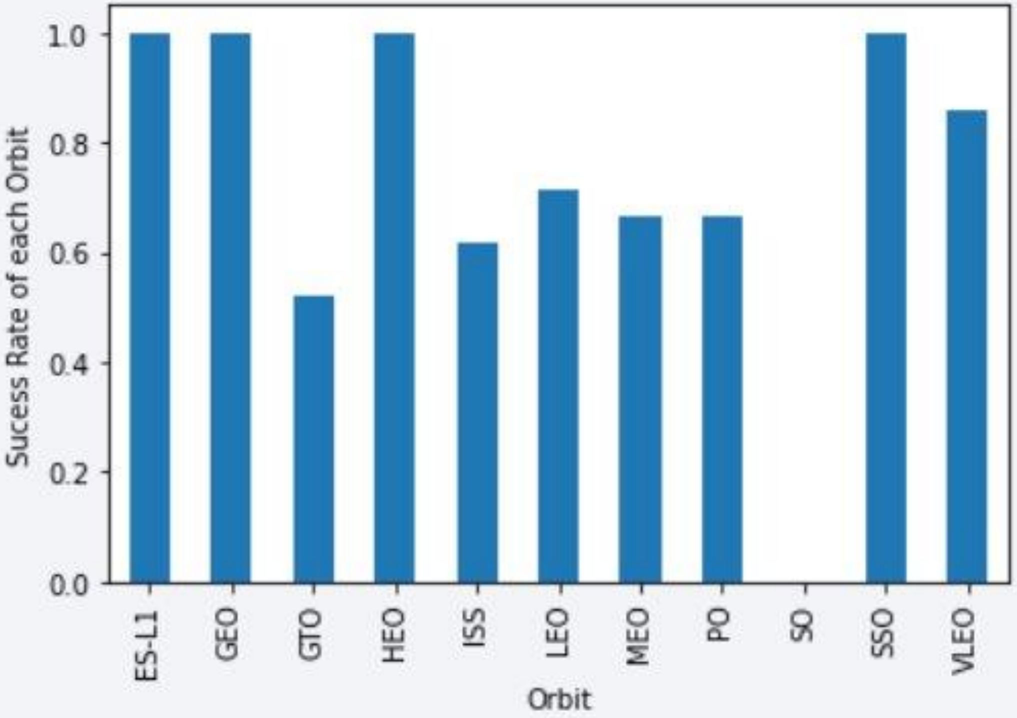
Payload vs. Launch Site

- From the figure below, it is observed that VAFB-SLC 4E launch site has no rockets launched for heavy payload mass greater than 10000kg.
- It is also observed that most of the rockets launched in all launch sites have a payload mass of less than 9000kg.
- Compared to VAFB-SLC 4E and KSC LC 39A, CCAFS SLC 40 has a higher success rate for rockets launched with a heavy payload mass of 14000kg and 16000kg.



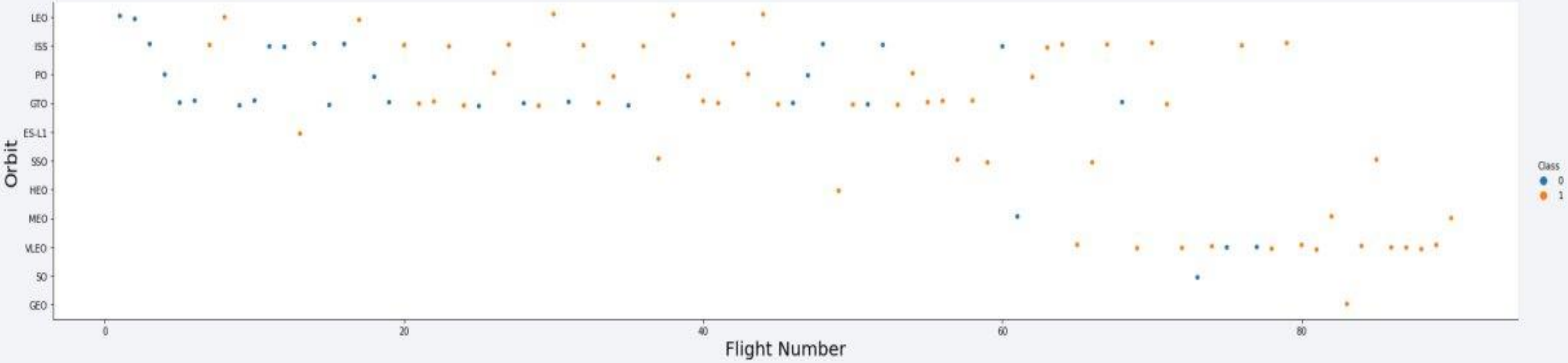
Success Rate vs. Orbit Type

- The bar chart shows the success rate for each orbit. It can be seen that four orbits — ESL1, GEO, HEO, and SSO — achieved a perfect success rate of 1, while SO had a success rate of 0. The remaining orbits had success rates ranging between 0.5 and 0.8.



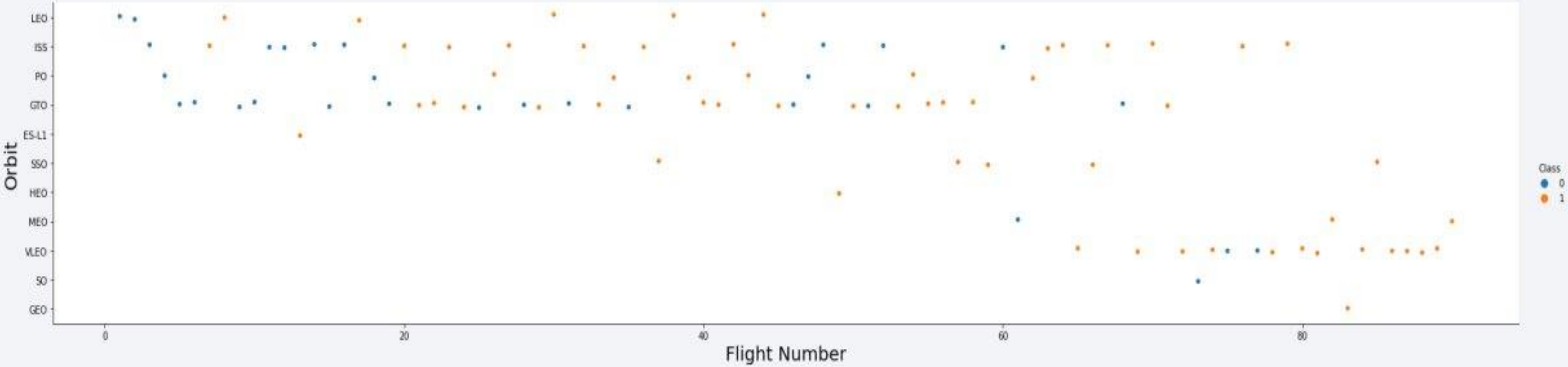
Flight Number vs. Orbit Type

- From the scatter plot, it can be observed that as the number of flightsincreases the success rate also increases. Hence, we can say thatthe past launches helped in improving the success rate



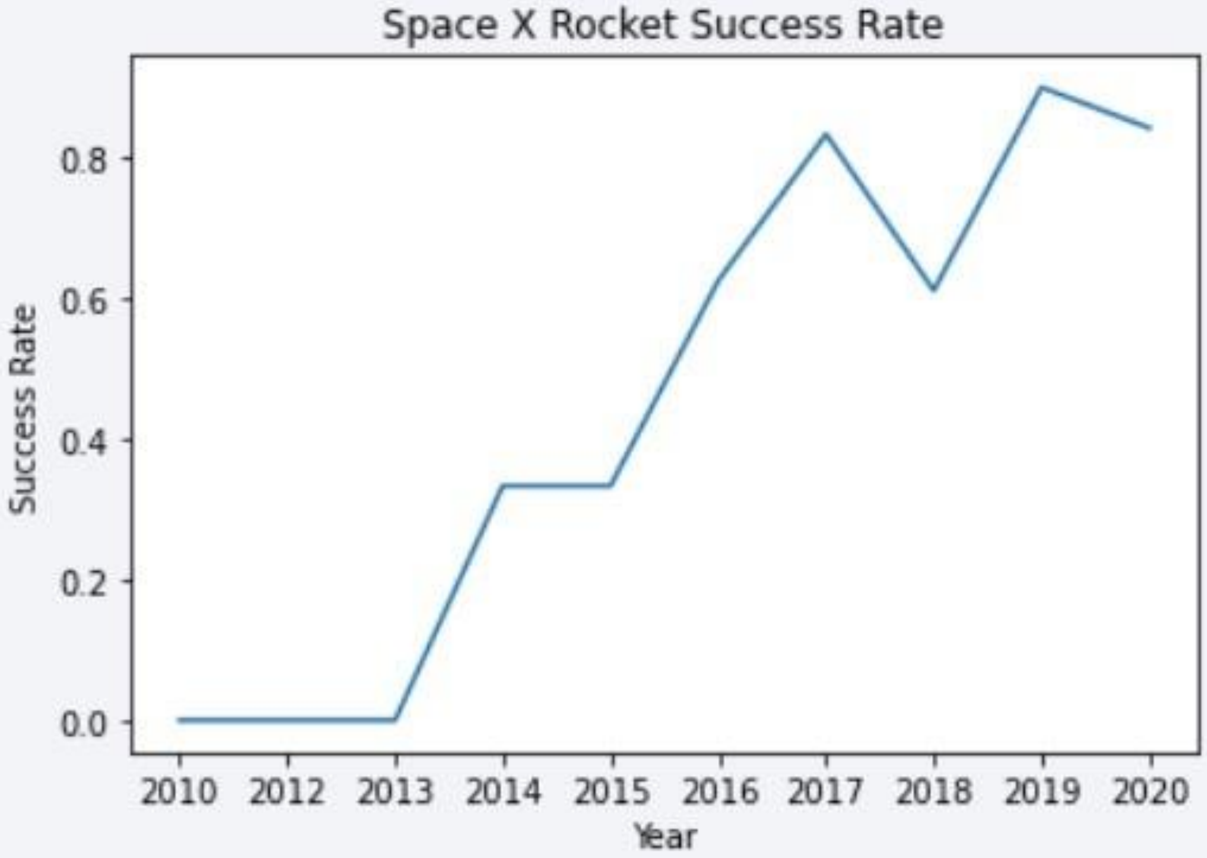
Payload vs. Orbit Type

- It can be seen that for some orbit, a heavier pay load mass helps t oimprove the success rate, for example the LEO orbit. However, f or someorbit such as GTO, increasing the pay load mass does not seem toimprove the success rate.



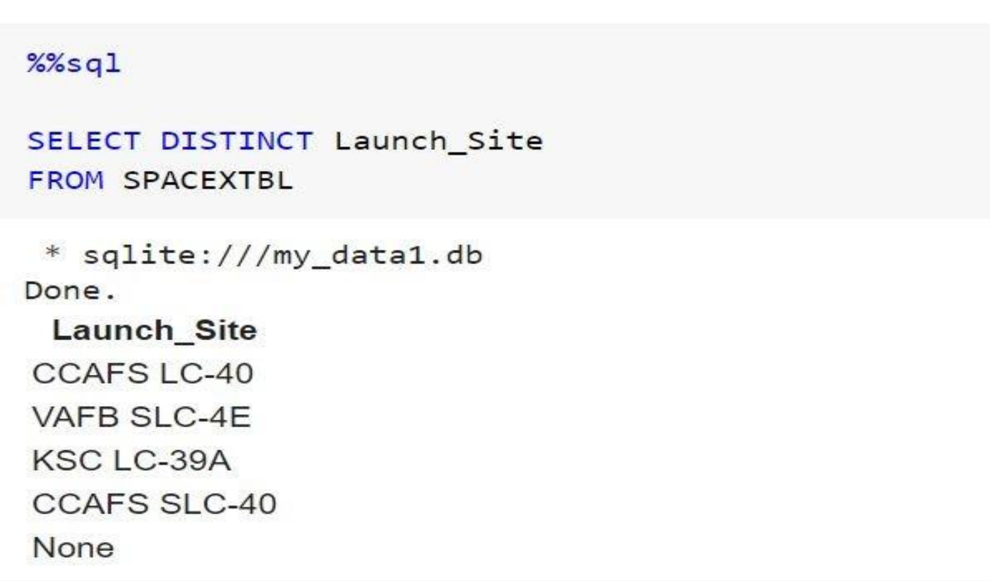
Launch Success Yearly Trend

- Launch success rate has increased significantly since 2013 and has stabilized since 2019, potentially due to advance in technology and lessons learned



All Launch Site Names

- An SQL table called SPACEXTBL using the existing data frame.
- To find the Unique Launch Sites, the keyword DISTINCT was used on the column.

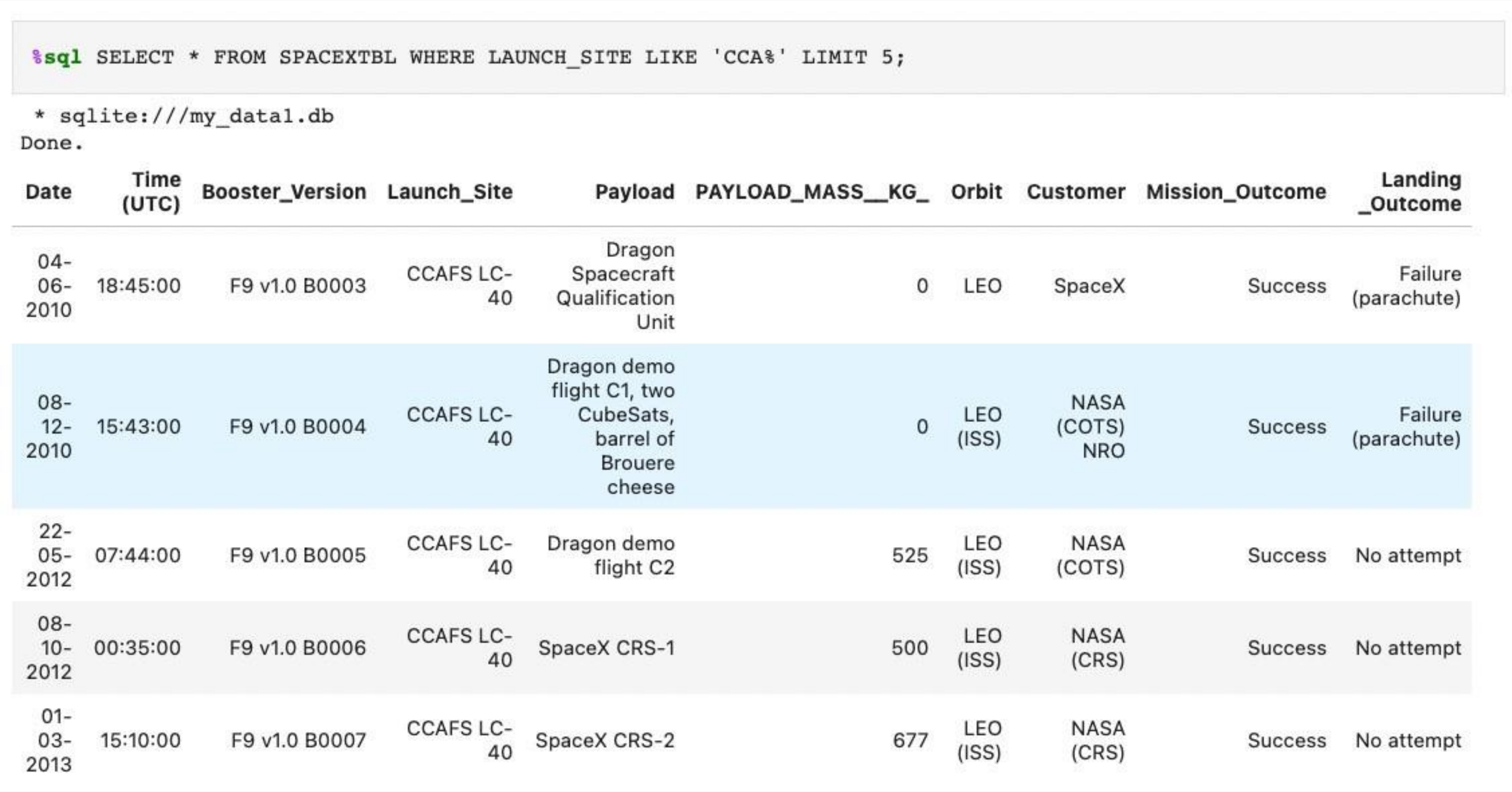


```
%%sql

SELECT DISTINCT Launch_Site
FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
 Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None
```

Launch Site Names Begin with 'CCA'

- We used the query bellow to display 5 records where launch sites begin with `CCA`



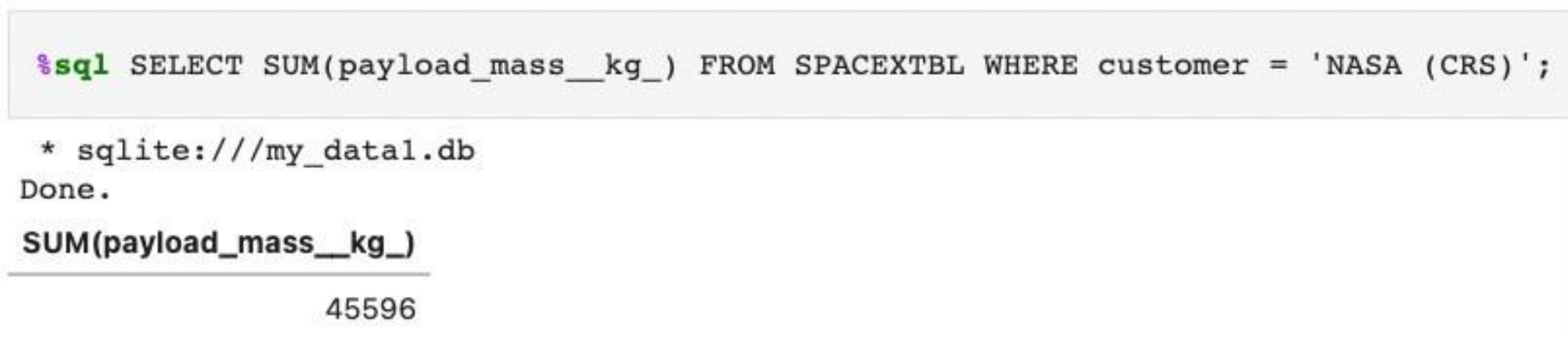
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

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

- We calculated the total payload carried by boosters from NASA as 45596 using the query below



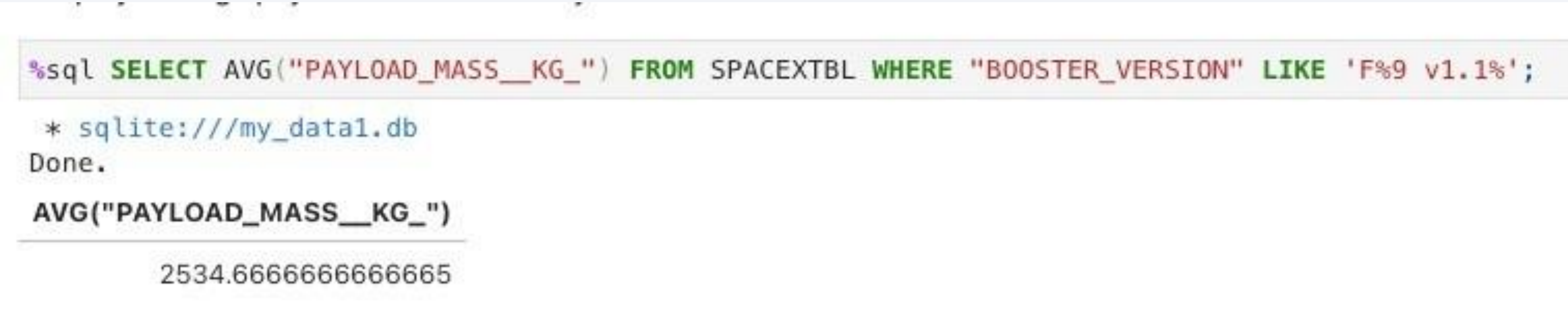
```
%sql SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE customer = 'NASA (CRS)';
```

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

SUM(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2534.66

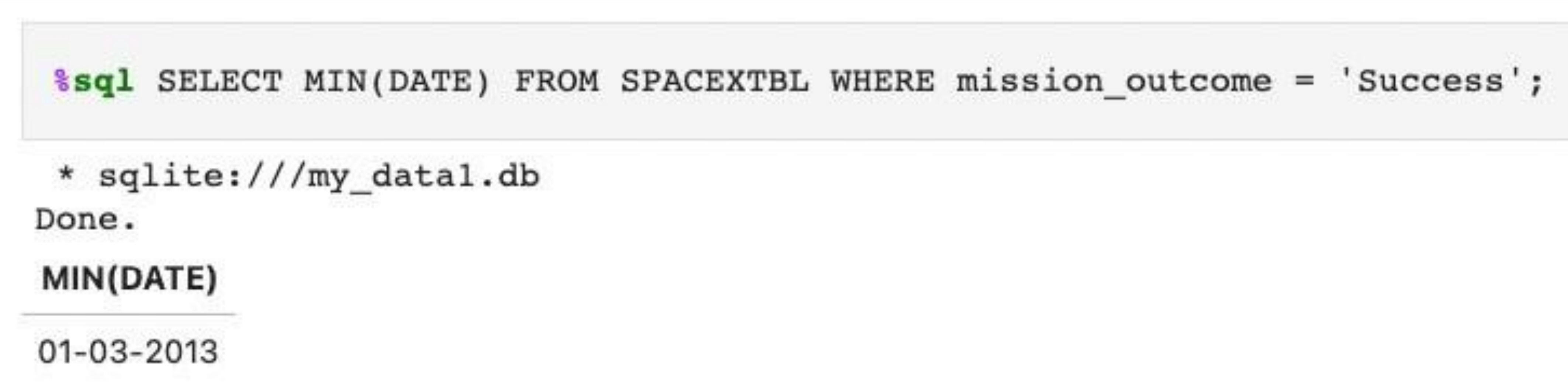


```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE 'F%9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
AVG("PAYLOAD_MASS__KG_")
       2534.666666666665
```

First Successful Ground Landing Date

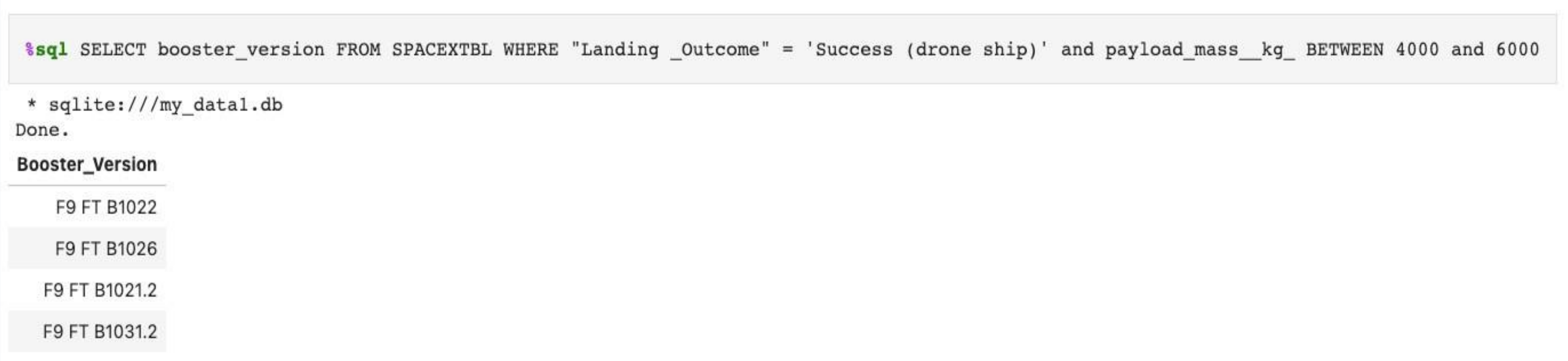
- We observed that the dates of the first successful landing outcome on ground pad was January 3rd, 2013.



```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE mission_outcome = 'Success';

 * sqlite:///my_data1.db
Done.

MIN(DATE)
01-03-2013
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000.



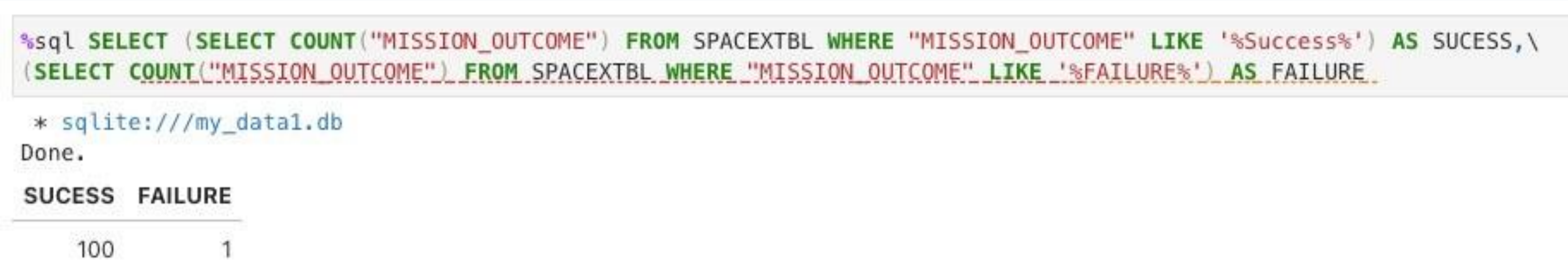
```
%sql SELECT booster_version FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' and payload_mass__kg_ BETWEEN 4000 and 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- We used wildcard like ‘%’ to filter for **WHERE** Mission Outcome was a success or a failure.



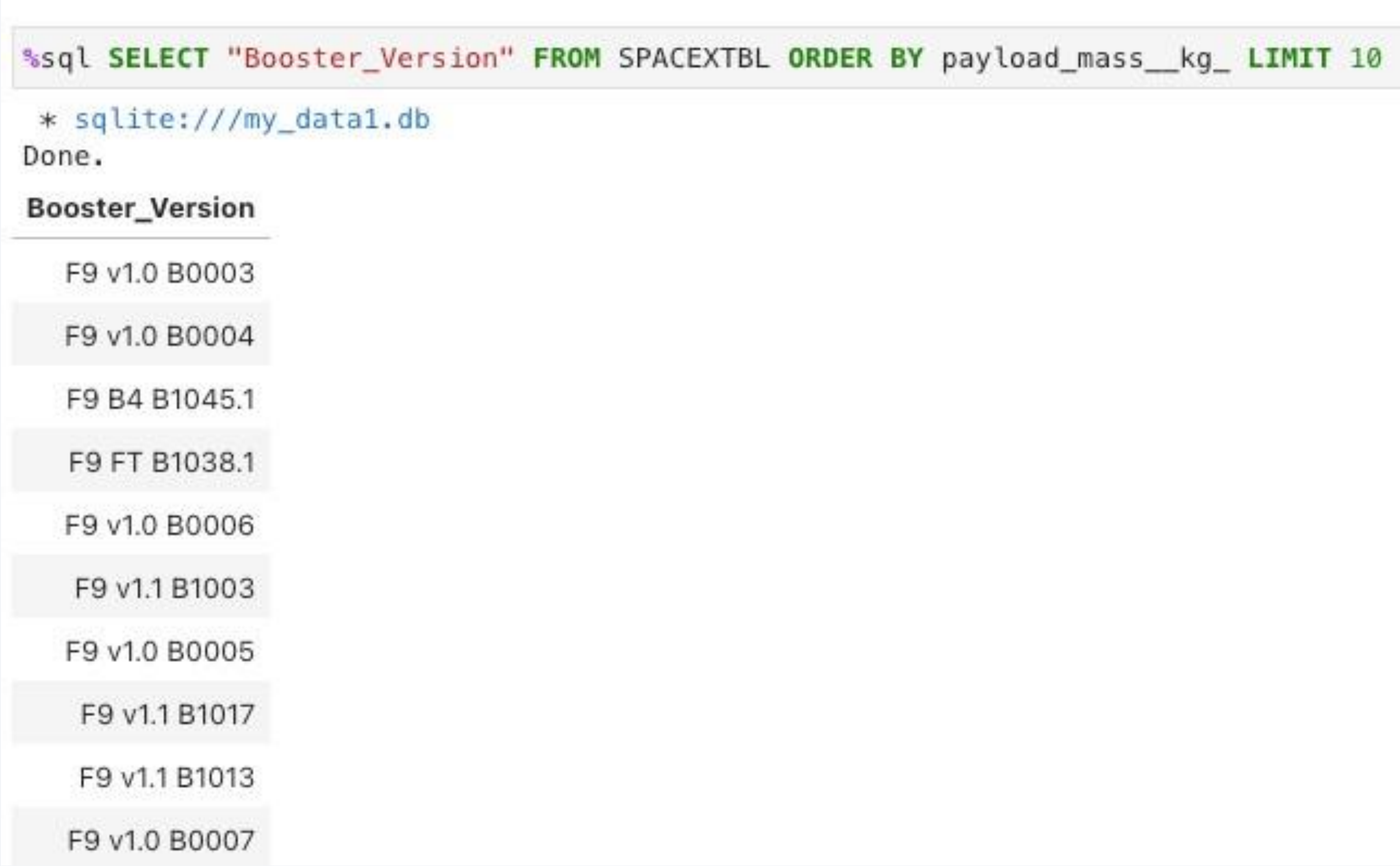
```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCESS,\
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%FAILURE%') AS FAILURE
```

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

SUCESS	FAILURE
100	1

Boosters Carried Maximum Payload

- These are the names of the booster which have carried the maximum payload mass



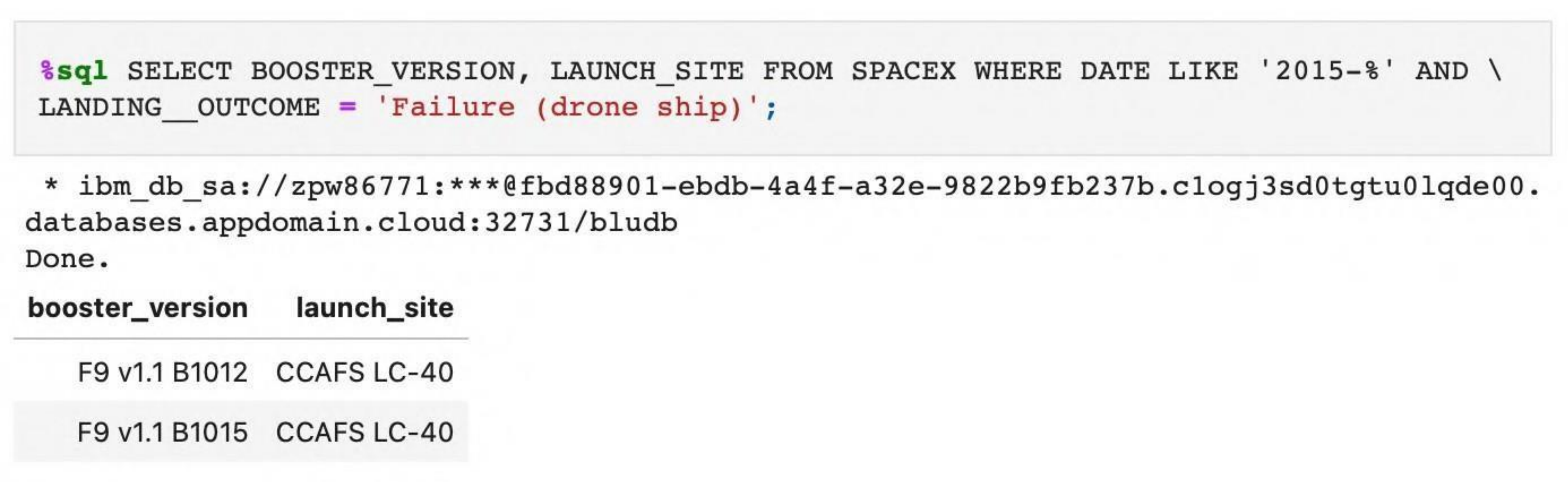
```
%sql SELECT "Booster_Version" FROM SPACEXTBL ORDER BY payload_mass__kg_ LIMIT 10
```

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

Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004
F9 B4 B1045.1
F9 FT B1038.1
F9 v1.0 B0006
F9 v1.1 B1003
F9 v1.0 B0005
F9 v1.1 B1017
F9 v1.1 B1013
F9 v1.0 B0007

2015 Launch Records

We used a combinations of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015



```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE DATE LIKE '2015-%' AND \
LANDING__OUTCOME = 'Failure (drone ship)';
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

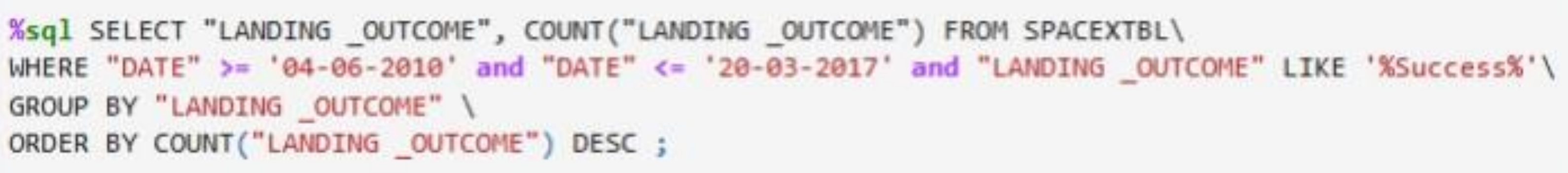
booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- We selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 04-06-2010 to 20-03-2017

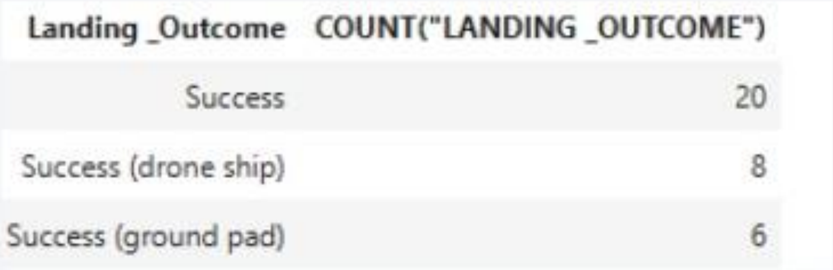
- We applied the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.

SQL Query:



```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

Output:

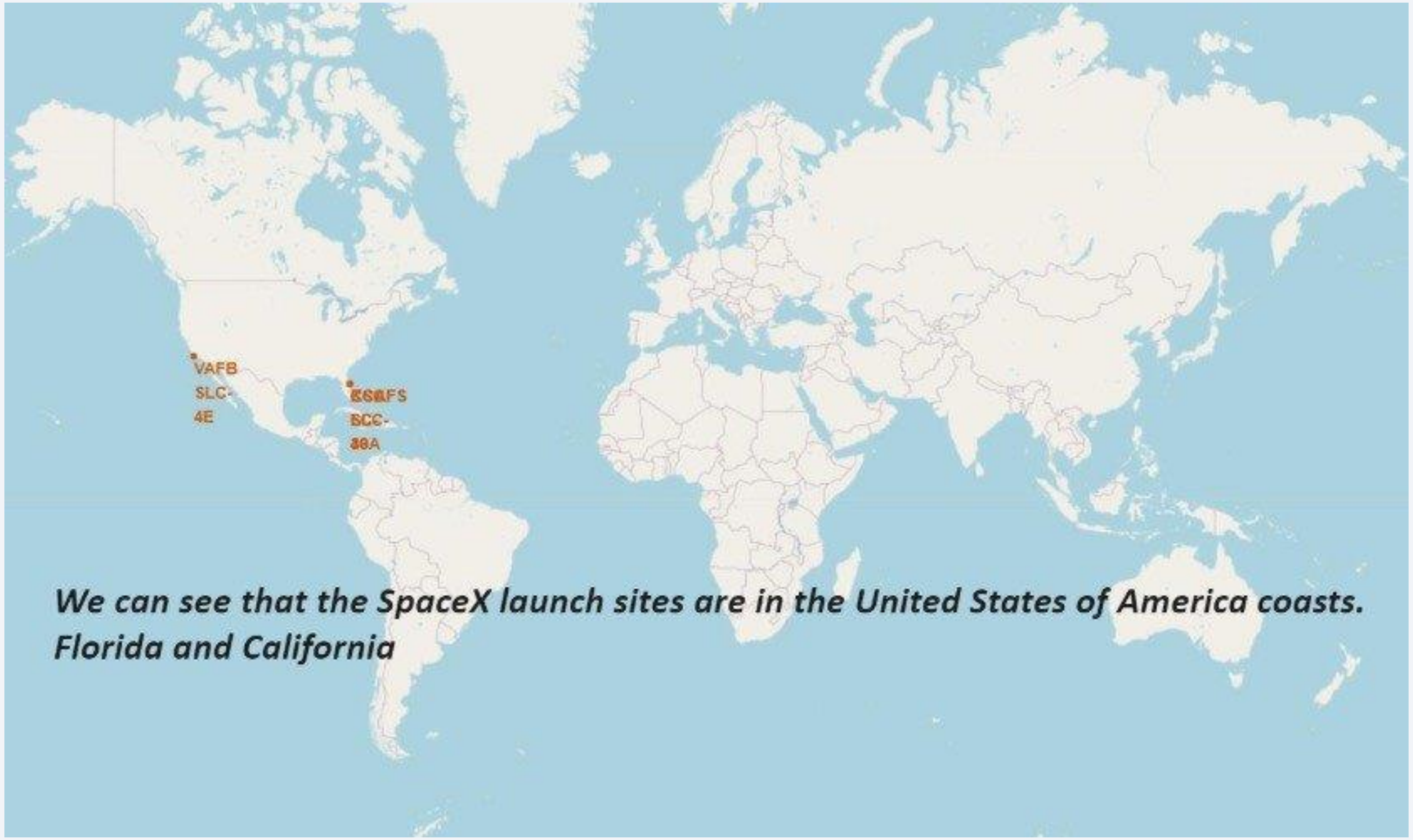


Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

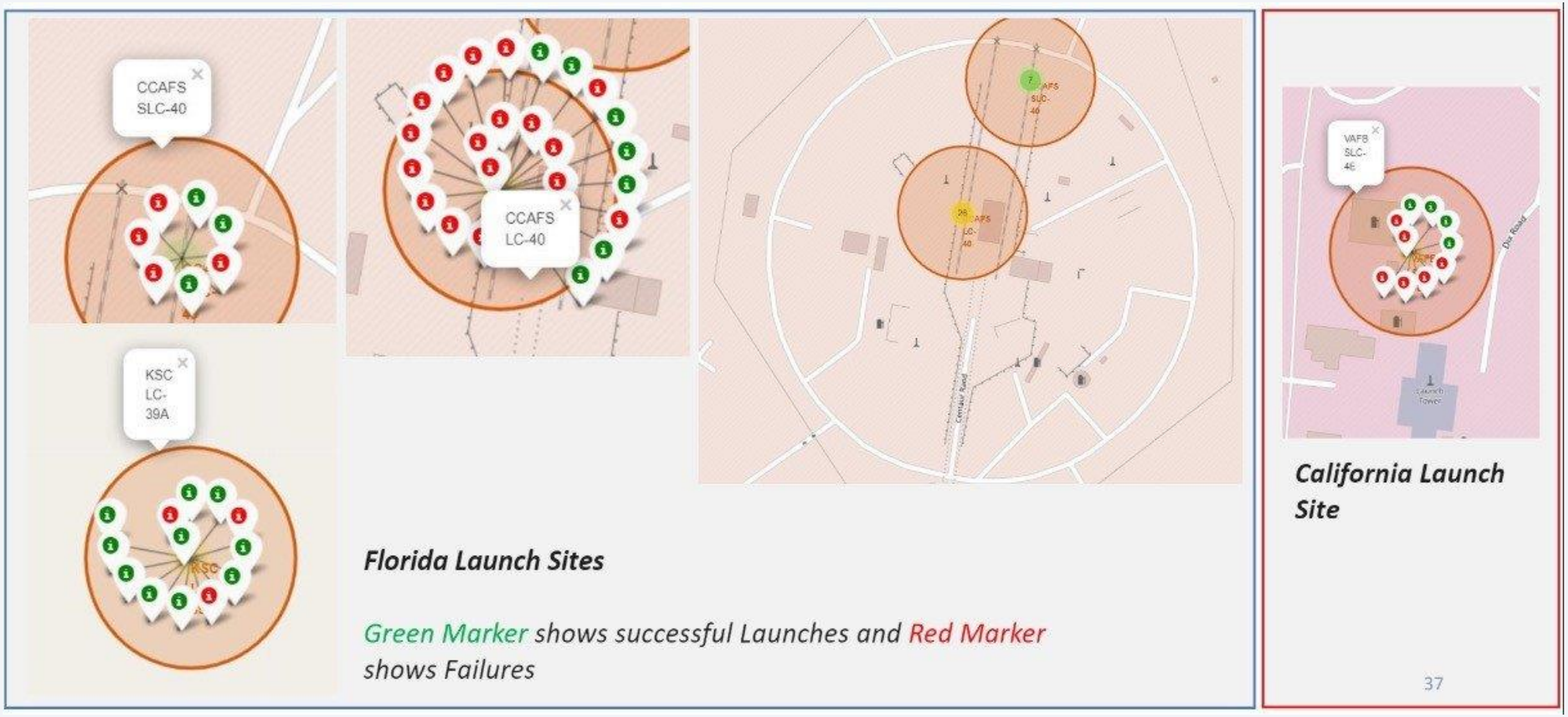
Launch Sites Proximities Analysis

All launch sites global map markers



We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

Launch Outcomes for Space X Falcon 9



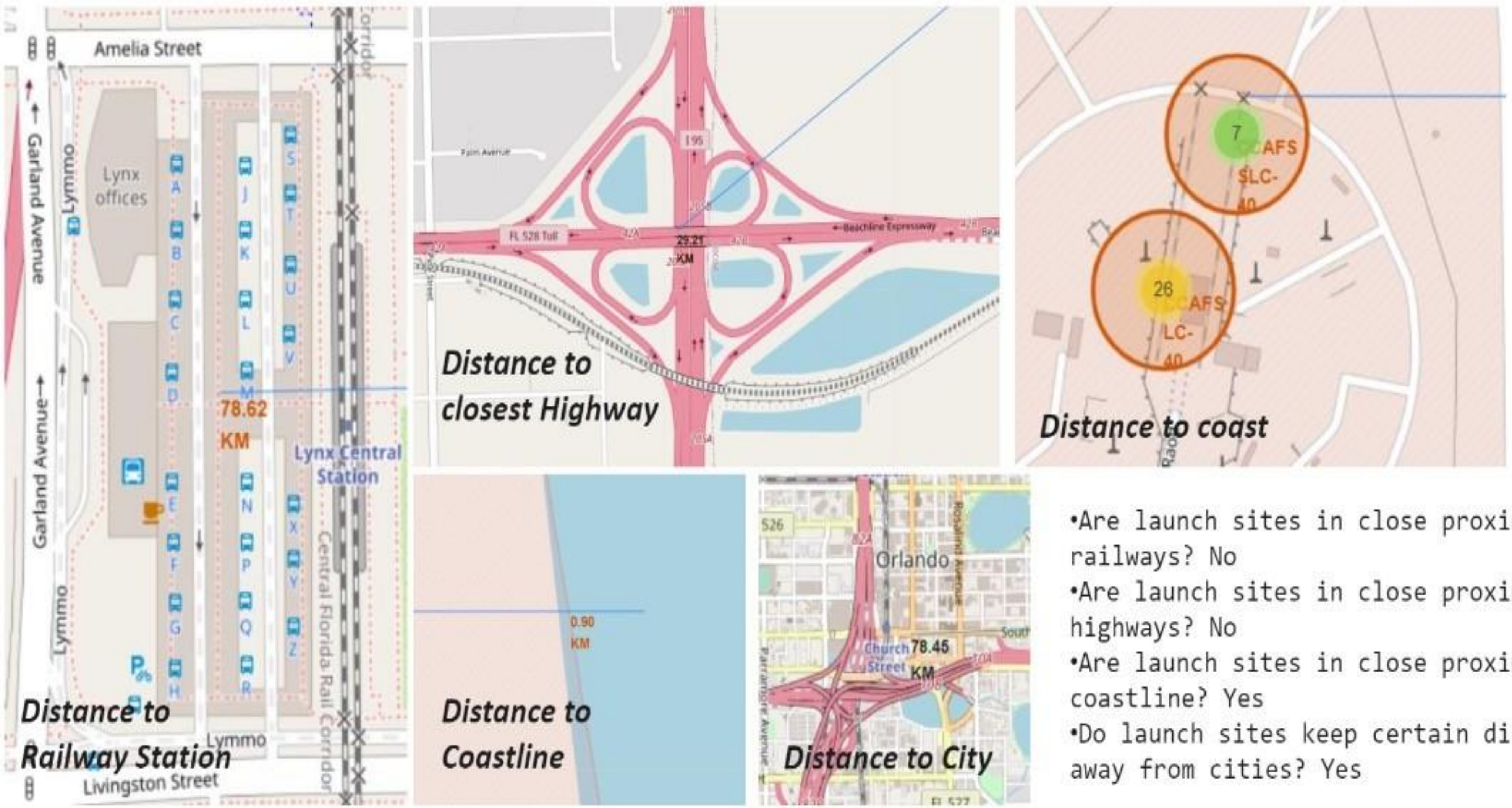
Florida Launch Sites

Green Marker shows successful Launches and Red Marker shows Failures

California Launch Site

37

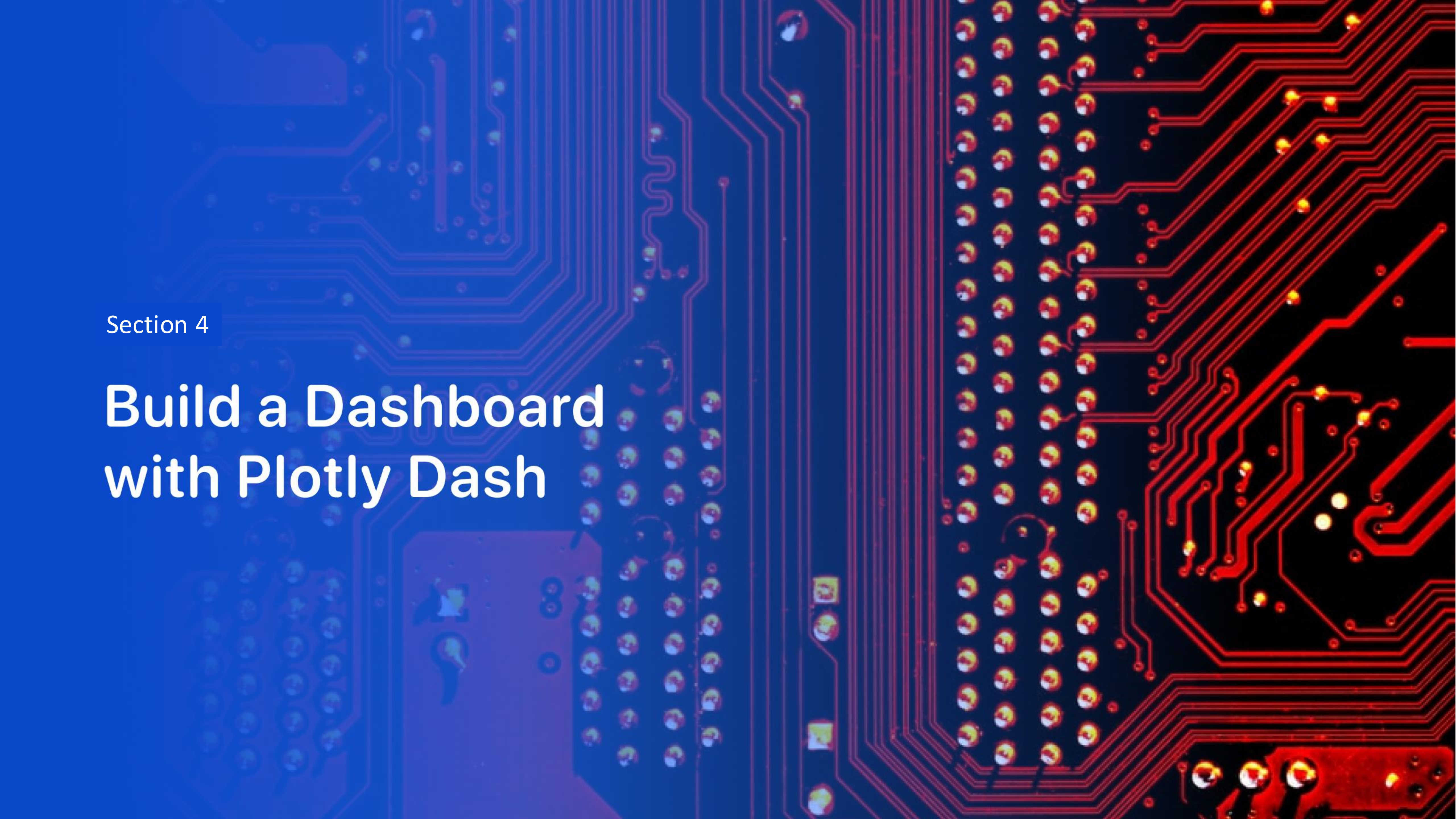
Launch Site distance to landmarks



- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

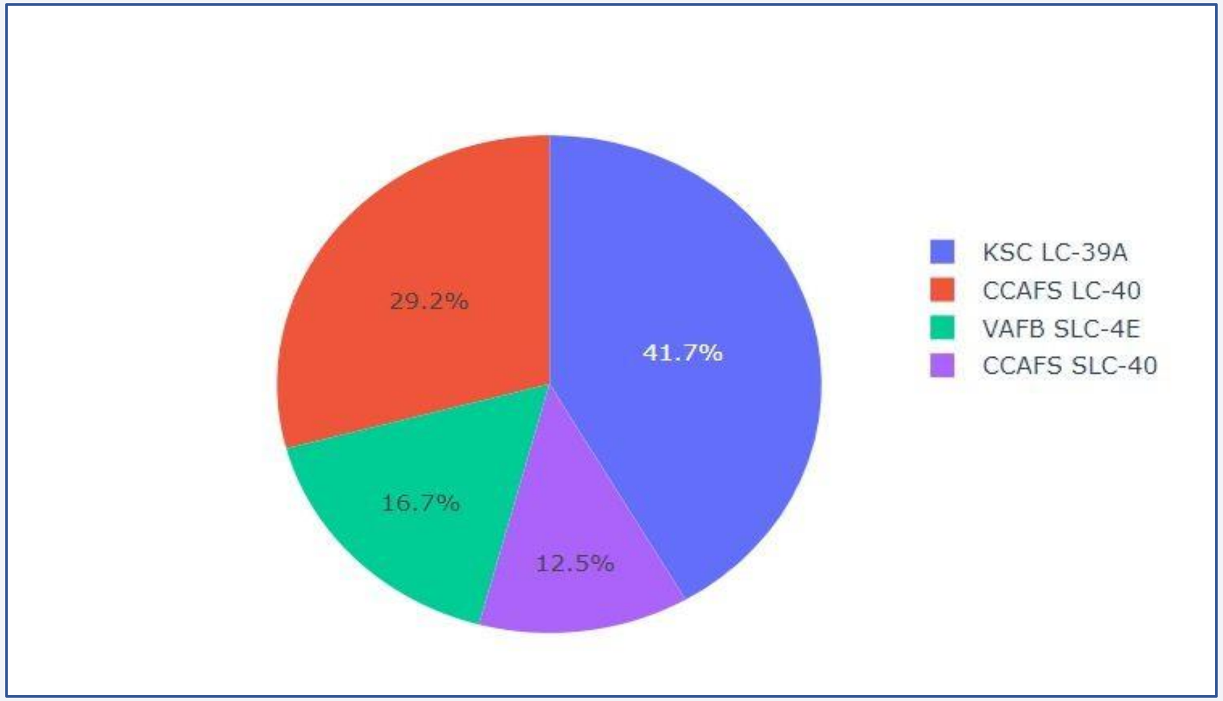
Section 4

Build a Dashboard with Plotly Dash



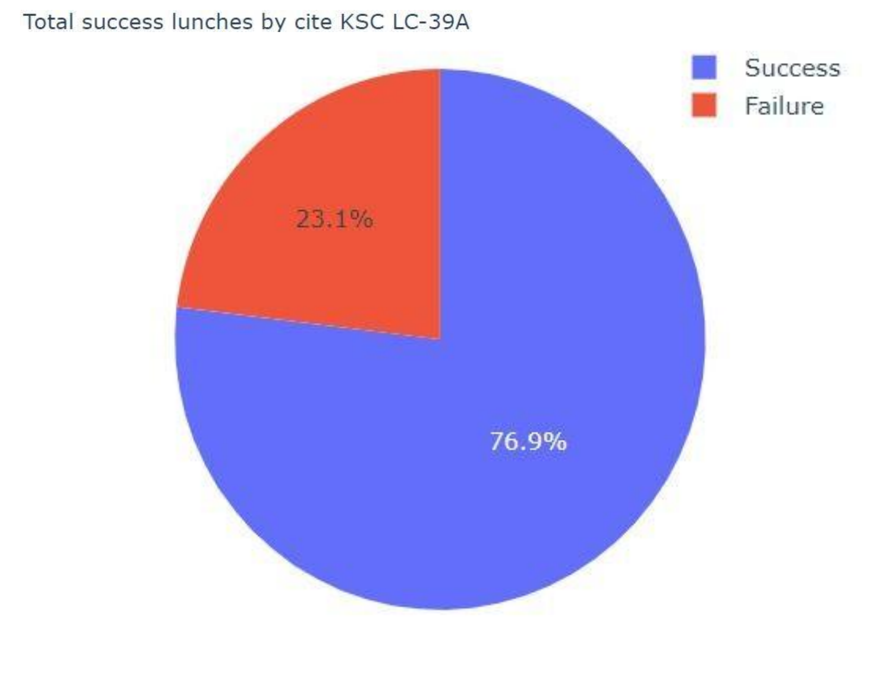
Pie Chart of Launch Success for all Sites

- it is shown that KSC LC-39A has the largest success rate with about 41.7% of the total success ratio with other sites.



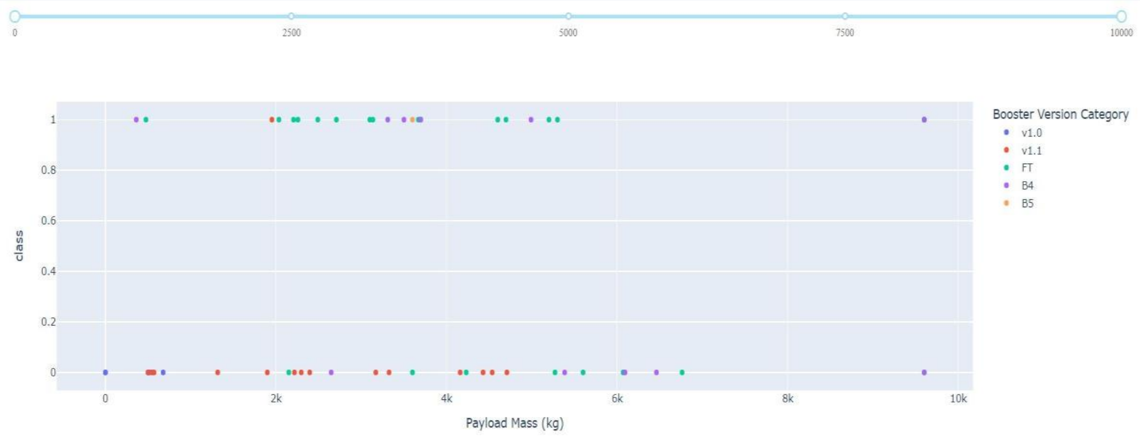
Pie chart of Launch site with highest success ratio

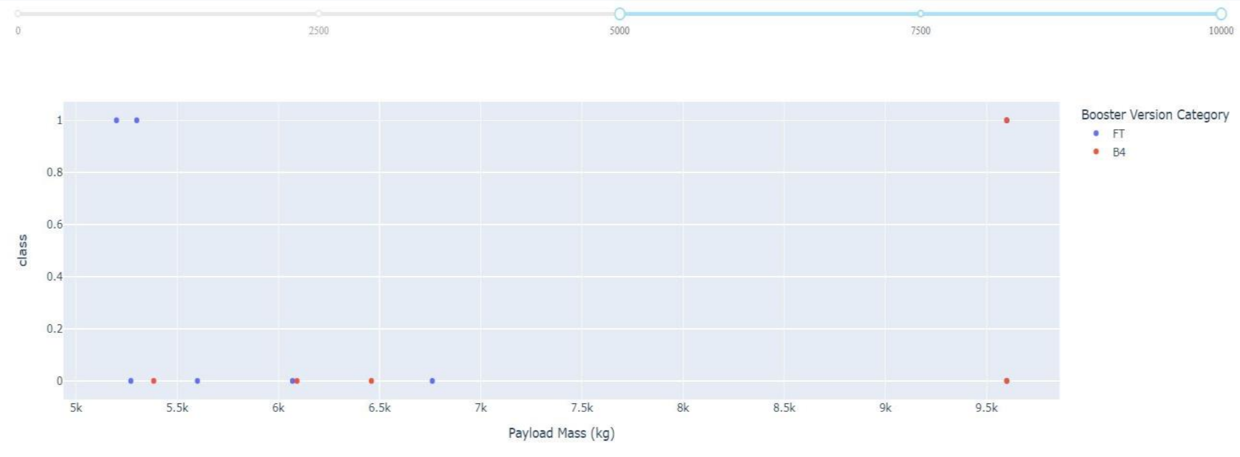
- It is also evident from Figure 25 that KSC LC-39A has the highest success ratio with about 76.9%, compared to the other sites;
    - 73.1% for CCAFS LC-40
    - 60% for VAFB SLC-4E
    - 57.1% for CCAFS SLC-40



Payload vs Launch outcome for all sites

- From the figures below, Booster version FT has the highest success rate with its payload mass of about between 700kg to 5,500kg.
- It is also shown that rockets with payload mass above 5,500kg have a lower success rate, which means the heavier the payload, the slimmer the chance of a successful outcome.



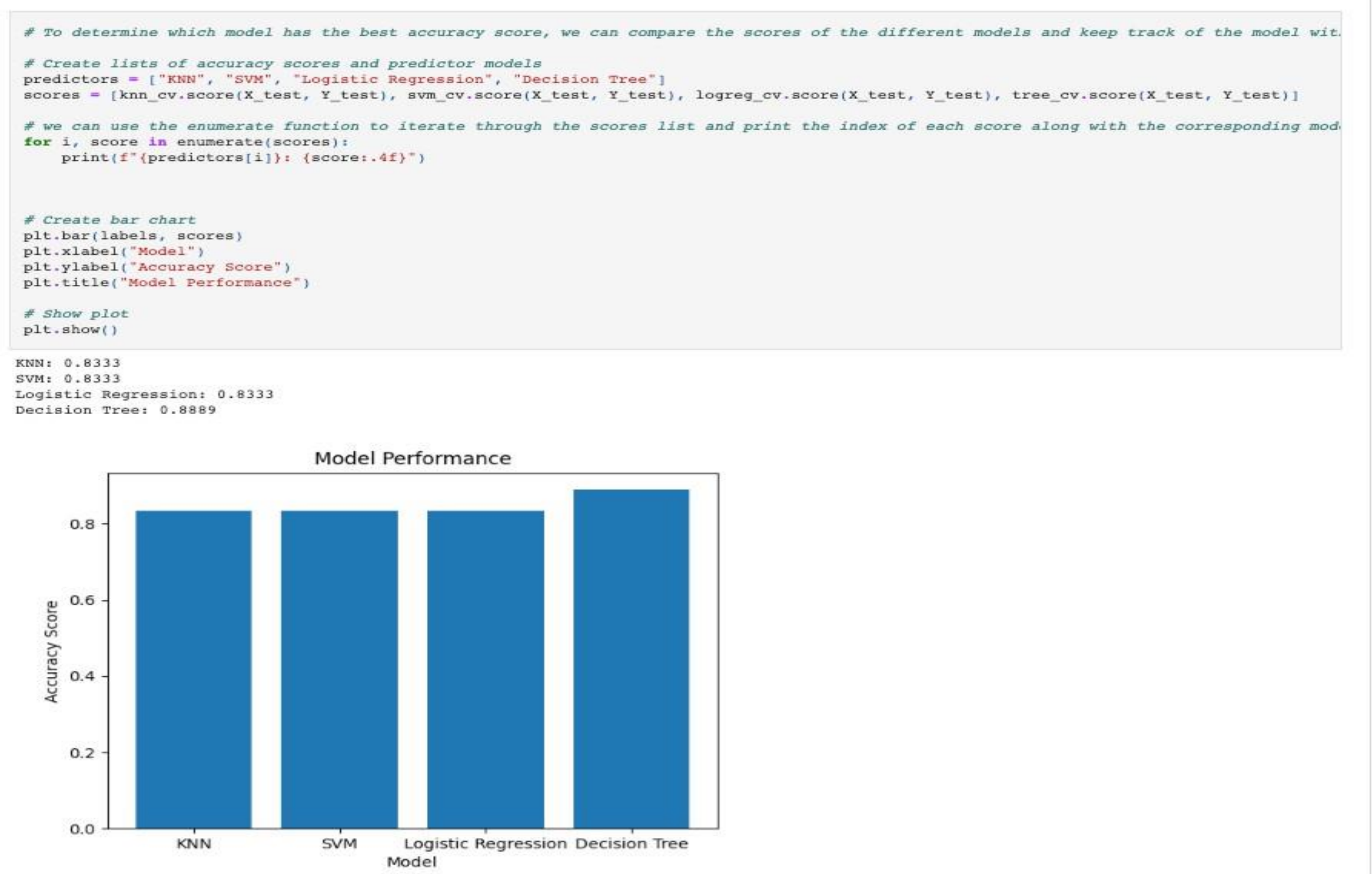


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classifier is the model with the highest classification accuracy with a 88.8%

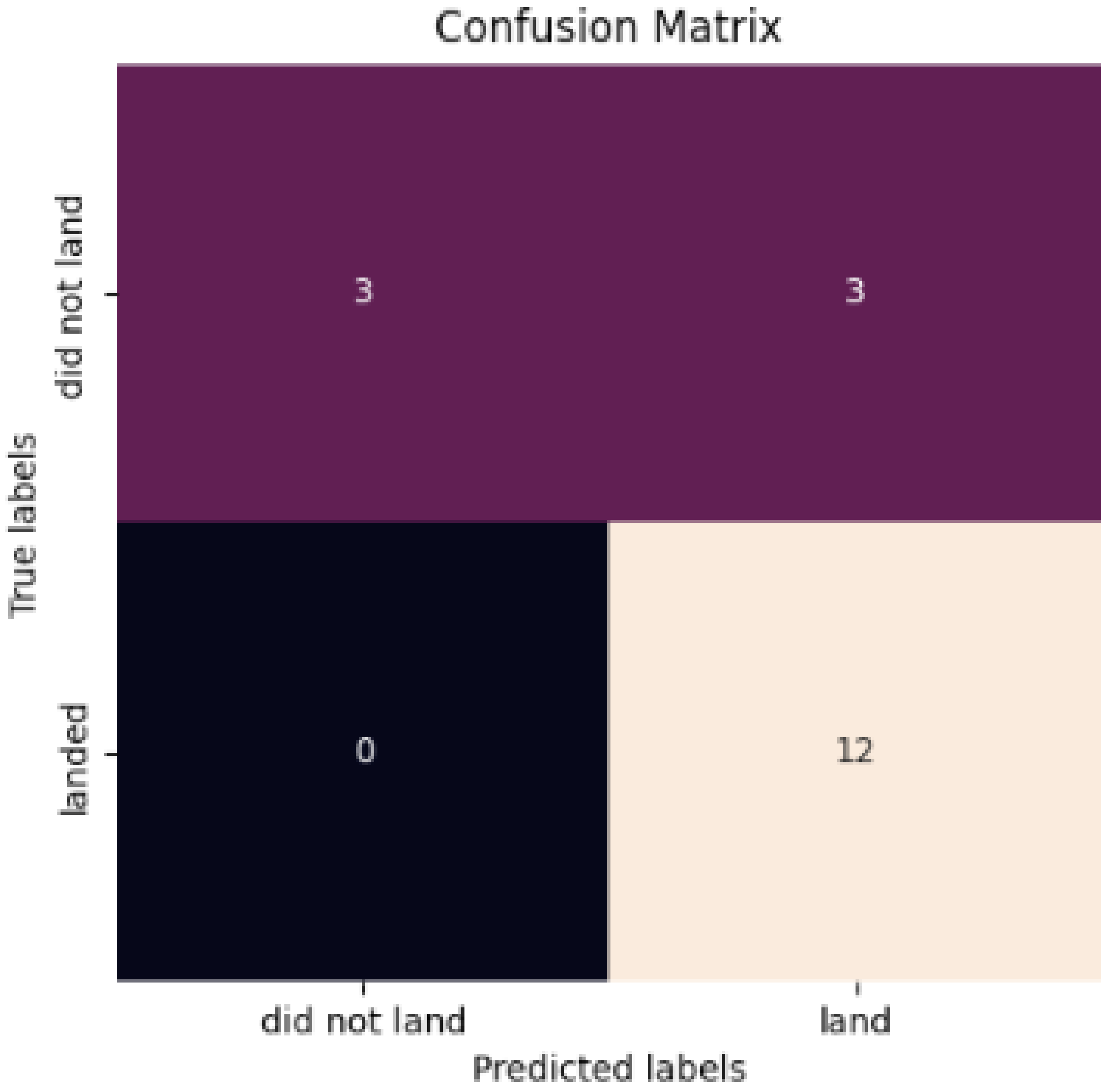


```
# To determine which model has the best accuracy score, we can compare the scores of the different models and keep track of the model wit

# Create lists of accuracy scores and predictor models
predictors = ["KNN", "SVM", "Logistic Regression", "Decision Tree"]
scores = [knn_cv.score(X_test, Y_test), svm_cv.score(X_test, Y_test), logreg_cv.score(X_test, Y_test), tree_cv.score(X_test, Y_test)]

# we can use the enumerate function to iterate through the scores list and print the index of each score along with the corresponding mod
for i, score in enumerate(scores):
    print(f"{predictors[i]}: {score:.4f}")


# Create bar chart
plt.bar(labels, scores)
plt.xlabel("Model")
plt.ylabel("Accuracy Score")
plt.title("Model Performance")

# Show plot
plt.show()
```

```
KNN: 0.8333
SVM: 0.8333
Logistic Regression: 0.8333
Decision Tree: 0.8889
```

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

- For a successful mission, the mass of the payload should be considered as rockets with smaller payload had a higher success rate.
- Orbit type should also be considered because rockets launched to certain orbits (VLEO, ES-L1, GEO, HEO, and SSO) had higher success rates compared to others.
- Launch sites are located in coastal cities for easy retrieval/recovery and far from busy areas like major highways and cities to minimize casualties in the event of a failure.
- In recent years the outcome has been more successful as later flight launches had a higher success rate.

Appendix

- https://github.com/HoussamEdar/Applied-Data-Science-Capstone/tree/main

Thank you!