IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammed Cagri Aslantas

24.01.2025

Outline

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

Executive Summary

Methodologies Summary:

The projects followed a structured approach starting with **data collection**, utilizing **APIs** (e.g., Space API) and **web scraping** to gather data. This was followed by **data wrangling**, where we cleaned, transformed, and engineered features to improve model performance. **SQL** was used to manage and query large datasets effectively.
During **exploratory data analysis (EDA)**, we visualized relationships and identified trends. Predictive models, including **KNN**, **SVM**, **Decision Trees**, and **Logistic Regression**, were applied. **GridSearchCV** helped fine-tune hyperparameters, and models were evaluated using **accuracy**, **precision**, **recall**, and **F1-score**.

Summary of Results:

The models showed strong performance, with **KNN**, **SVM**, and **Logistic Regression** achieving around **83.33%** accuracy. The **data wrangling** process and use of **SQL** for data management were essential for quality data preparation. **APIs** and **web scraping** provided rich datasets, enabling effective analysis. Overall, the combination of these methods resulted in accurate, reliable predictive models.

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. Beyond predicting landings, an extended cost analysis would include refurbishment costs, fuel expenses, and turnaround times, allowing SpaceY to assess whether reusability leads to a true economic advantage.

Problems to Find Answers:

- **How to clean and process large-scale space data** for analysis.
- **Which machine learning models best fit space data** to predict rocket launches and mission outcomes.
- **How to leverage SQL** for managing and analyzing space-related records.
- **Evaluating model performance** to ensure accurate predictions.
- **The ultimate goal** is to develop a predictive system for rocket launches and space missions, offering valuable insights for the aerospace industry.

Section 1

Methodology

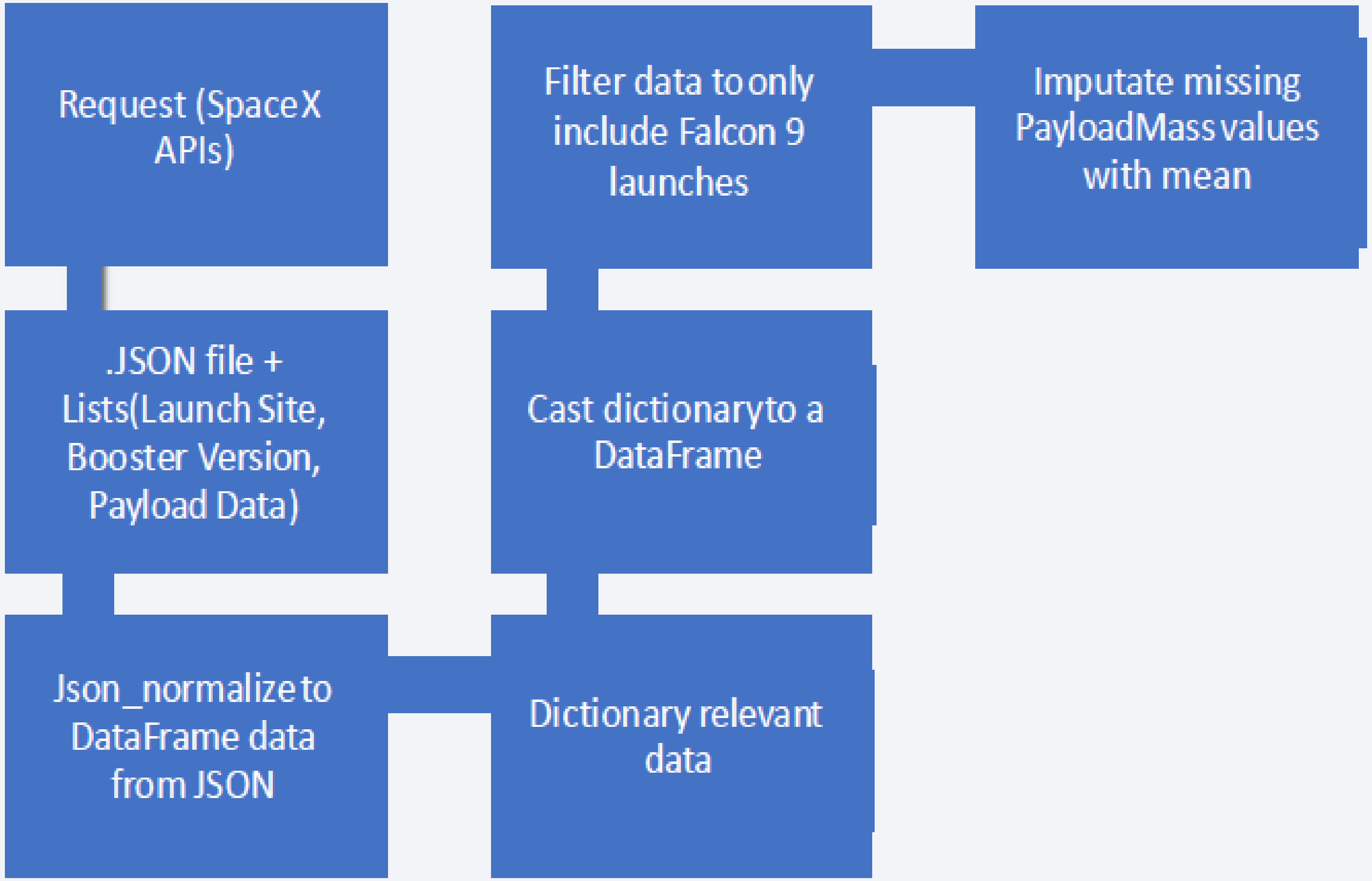
Methodology

Executive Summary

- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models

Data Collection – SpaceX API

FLOWCHART:



Request (SpaceX APIs)

.JSON file + Lists(Launch Site, Booster Version, Payload Data)

Json_normalize to DataFrame data from JSON

Dictionary relevant data

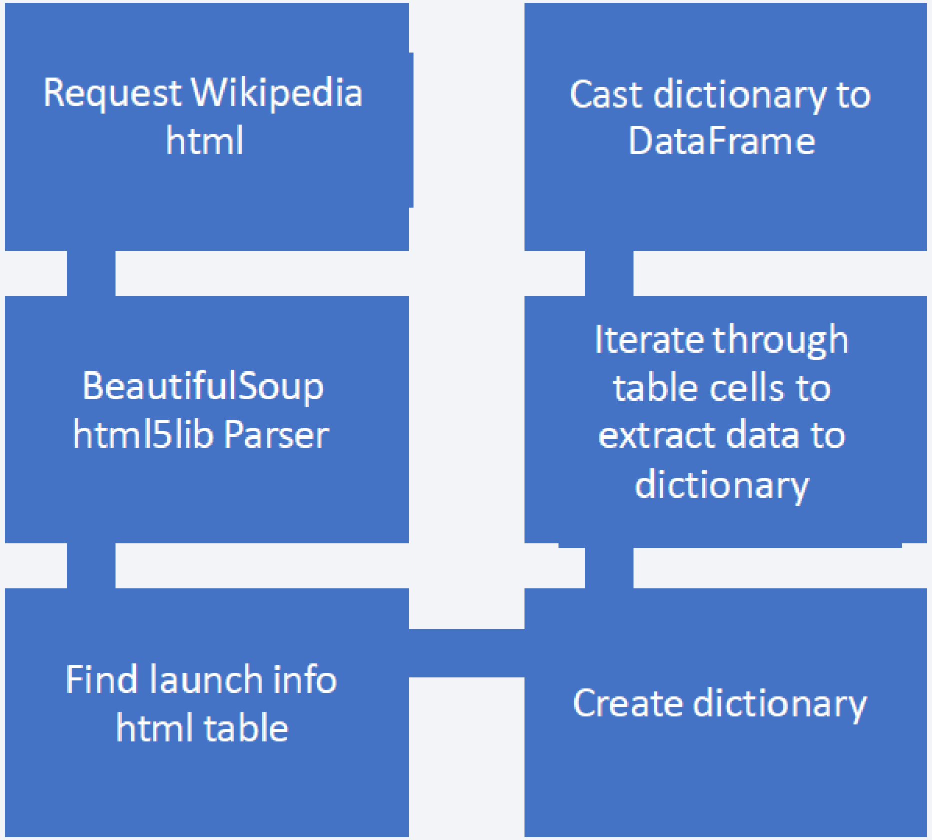
Cast dictionaryto a DataFrame

Filter data to only include Falcon 9 launches

Imputate missing PayloadMass values with mean

The full analysis and code can be reviewed on GitHub: https://github.com/maslantas46/Projekte/blob/main/SpaceX%20Api.ipynb

Data Collection - Scraping



Request Wikipedia html

BeautifulSoup html5lib Parser

Find launch info html table

Create dictionary

Iterate through table cells to extract data to dictionary

Cast dictionary to DataFrame

The full analysis and code can be reviewed on GitHub: https://github.com/maslantas46/Projekte/blob/main/Scraping.ipynb

Data Wrangling

Key Phrases:

- **Data Wrangling:** The process of cleaning and transforming raw data into a structured format suitable for analysis.
- **Handling Missing Data:** Identifying and handling missing or incomplete data.
- **Data Types Conversion:** Converting data types for consistency.
- **Data Filtering:** Removing or handling outliers and irrelevant data.
- **Feature Engineering:** Creating new features or transforming existing ones to enhance model performance.
- **Data Normalization:** Scaling data to bring features to a common scale.
- **Data Export:** Saving processed data for further analysis.

The full analysis and code can be reviewed on GitHub: https://github.com/maslantas46/Projekte/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

What Charts Were Plotted and Why Those Charts Were Used

- EDA (Exploratory Data Analysis): The process of analyzing and visualizing the data to understand its structure and relationships.
- Data Visualization: Graphical representation of data to help identify trends, patterns, and outliers.
- Charts Used:
- Line Plot: Used to track the success rate over time (year-wise) to observe trends.
- Bar Plot: Helps in visualizing categorical data, like the number of successful launches by Launch Site.
- Scatter Plot: Used to analyze relationships between numeric features, such as PayloadMass vs. Success Rate.
- Box Plot: Used to visualize the distribution of numerical data and detect outliers.
- Heatmap: Correlation matrix heatmap to show how different features correlate with each other.

The full analysis and code can be reviewed on GitHub: https://github.com/maslantas46/Projekte/blob/main/edadataviz.ipynb

The full analysis and code can be reviewed on GitHub: https://github.com/maslantas46/Projekte/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL

- **Task 1**: Displayed the unique launch sites used in SpaceX missions.
- **Task 2**: Filtered launch sites that begin with 'CCA' and displayed 5 records.
- **Task 3**: Calculated the total payload mass carried by boosters launched by NASA (CRS).
- **Task 4**: Calculated the average payload mass carried by boosters of version F9 v1.1.
- **Task 5**: Retrieved the date when the first successful landing outcome on a ground pad was achieved.
- **Task 6**: Displayed boosters that had successful landings on a drone ship with a payload mass greater than 4000 but less than 6000 kg.
- **Task 7**: Displayed the total number of successful and failed mission outcomes.
- **Task 8**: Listed the names of the booster versions that carried the maximum payload mass.
- **Task 9**: Retrieved records of missions with failure landing outcomes in drone ships, including booster versions and launch sites, for the months in the year 2015.
- **Task 10**: Ranked the count of landing outcomes (e.g., "Failure (drone ship)" or "Success (ground pad)") between the dates 2010-06-04 and 2017-03-20, in descending order.

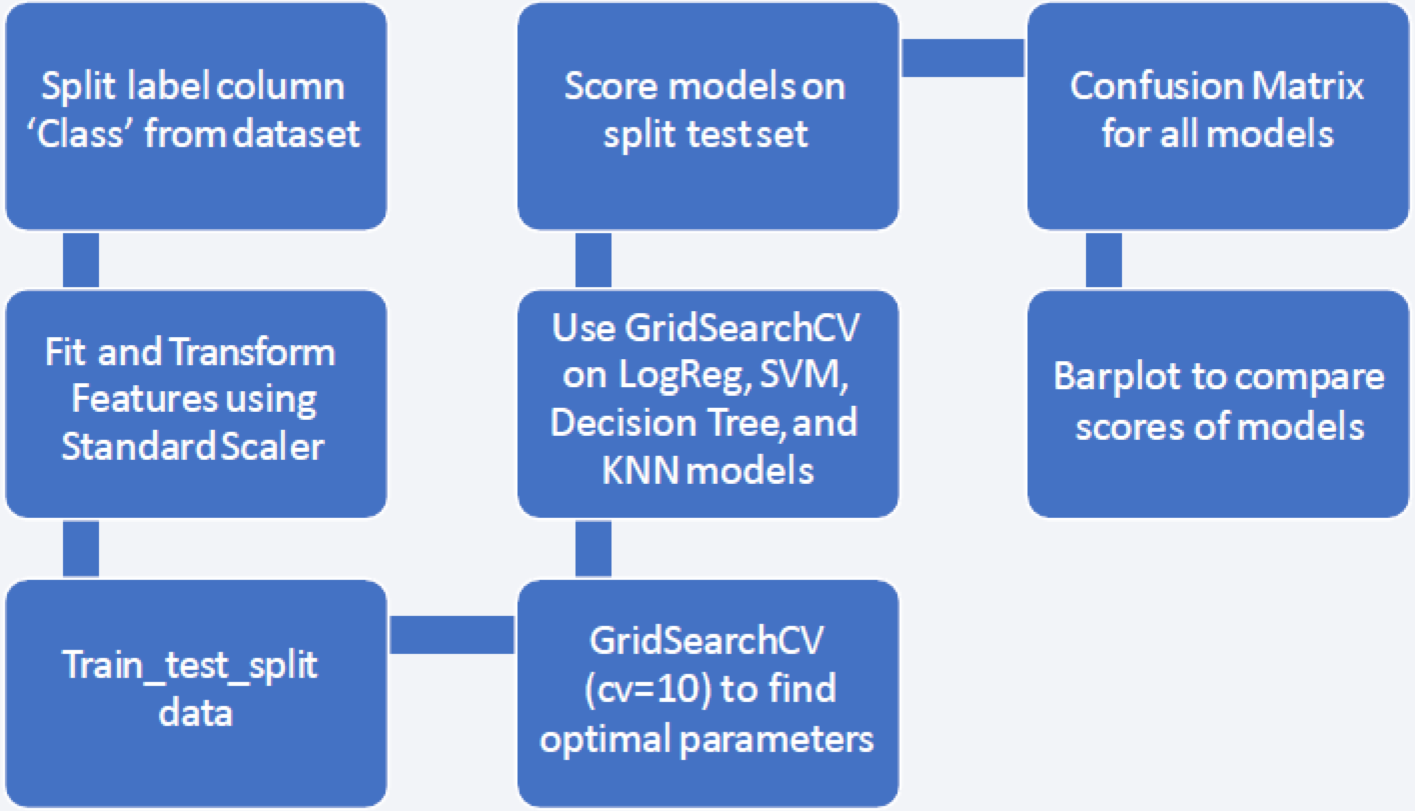
Build an Interactive Map with Folium

- Launch Sites, successful and unsuccessful landings, and a proximity example to important locations—such as the city, coast, highway, and railroad—are marked on folium maps.

- This enables us to comprehend the reasons for the possible locations of launch sites. Additionally, it shows successful landings in relation to their location.

The full analysis and code can be reviewed on GitHub: https://github.com/maslantas46/Projekte/blob/main/Predictive%20Analysis.ipynb

Predictive Analysis (Classification)


Split label column ‘Class’ from dataset

Fit and Transform Features using Standard Scaler

Train_test_split data

GridSearchCV (cv=10) to find optimal parameters

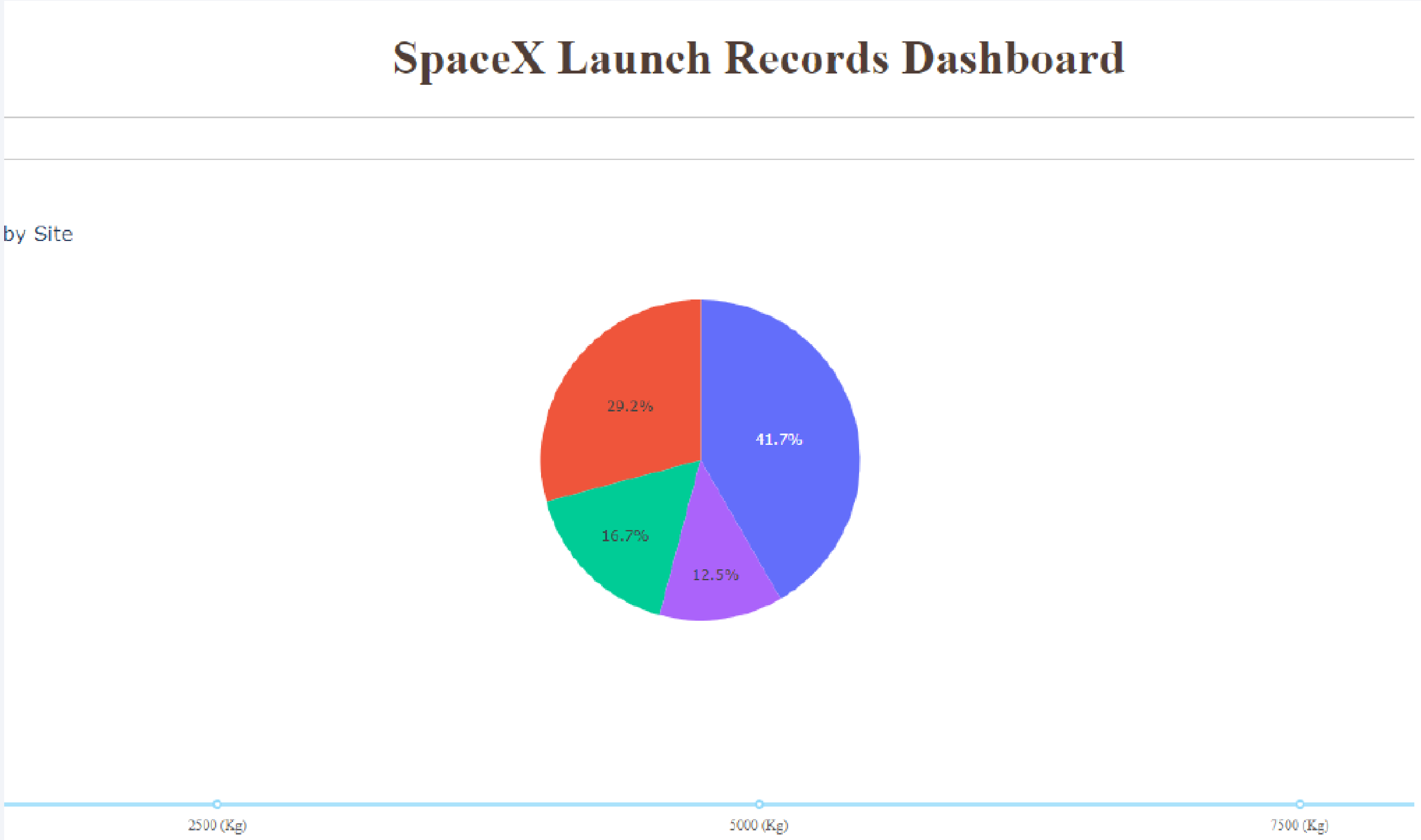
Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Score models on split test set

Confusion Matrix for all models

Barplot to compare scores of models

Results



SpaceX Launch Records Dashboard

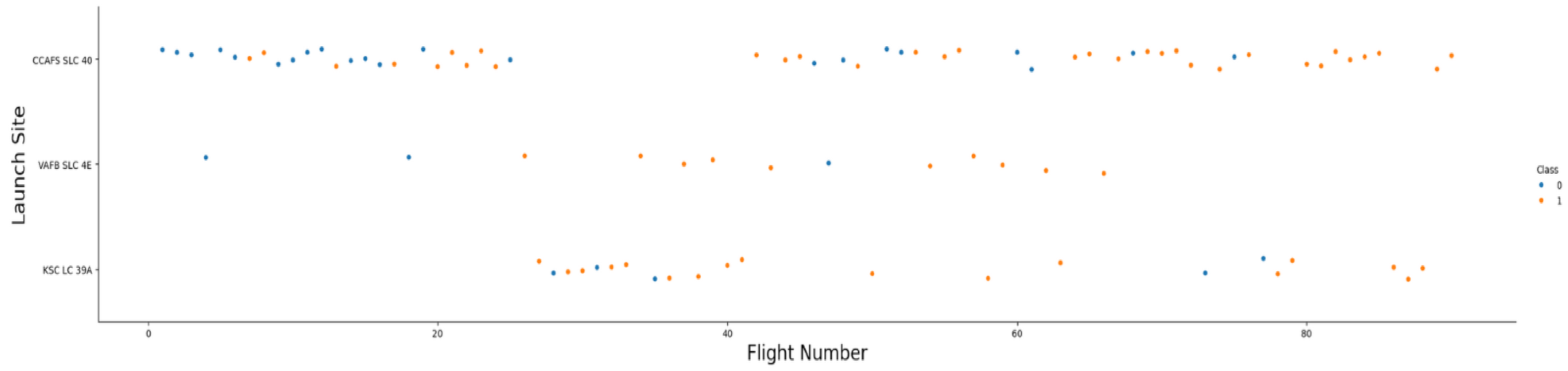
Results

A sneak peek at the Plotly dashboard is seen here. The outcomes of EDA with visualisation, EDA with SQL, Interactive Map with Folium, and ultimately our model's outputs with an accuracy of roughly 83% are displayed on the following slides.

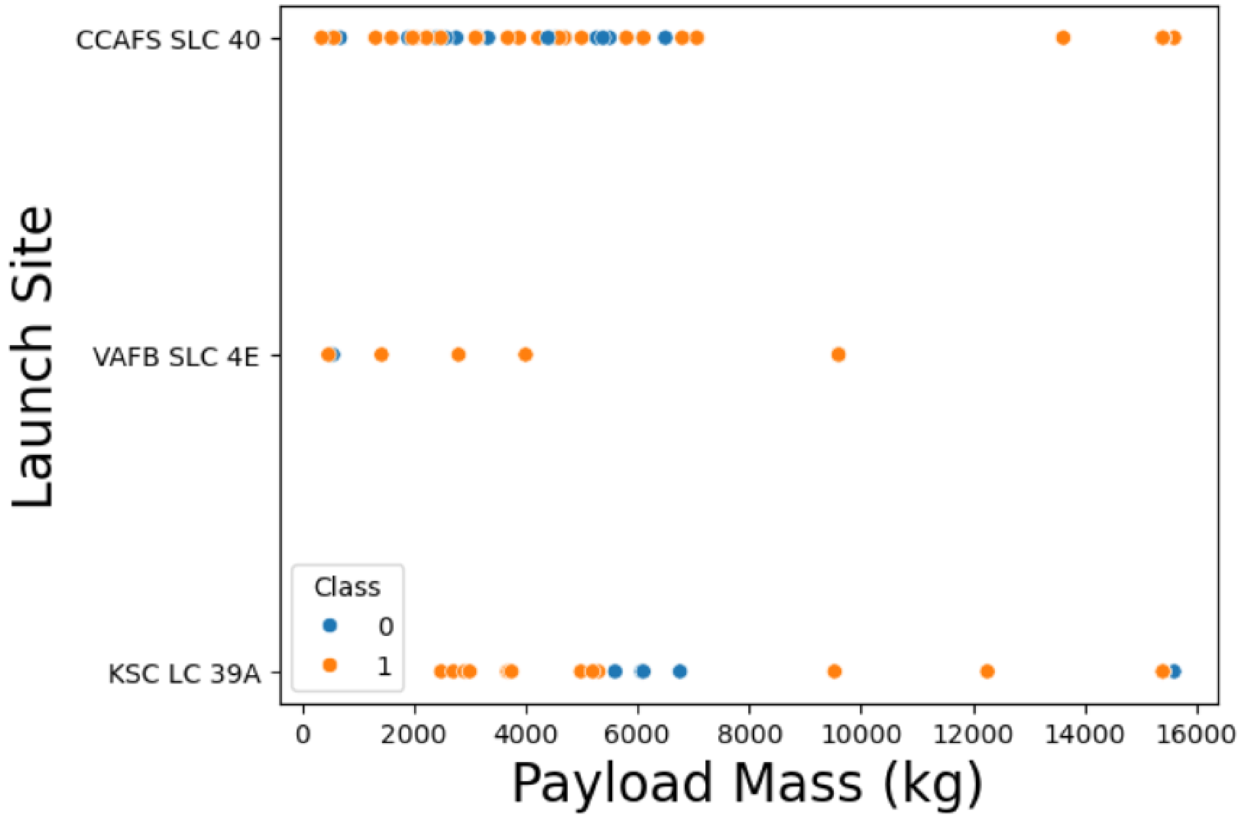
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

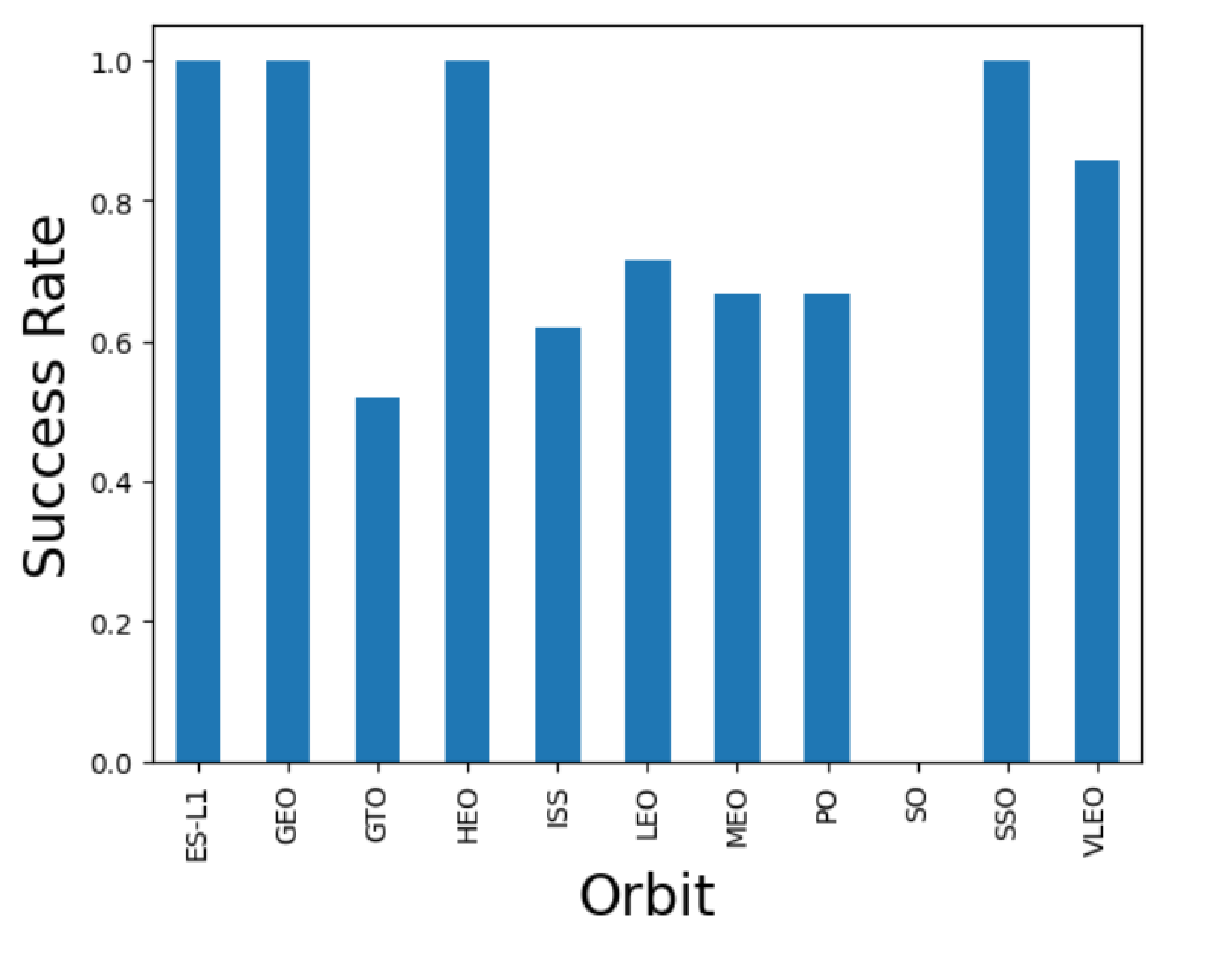


Payload vs. Launch Site

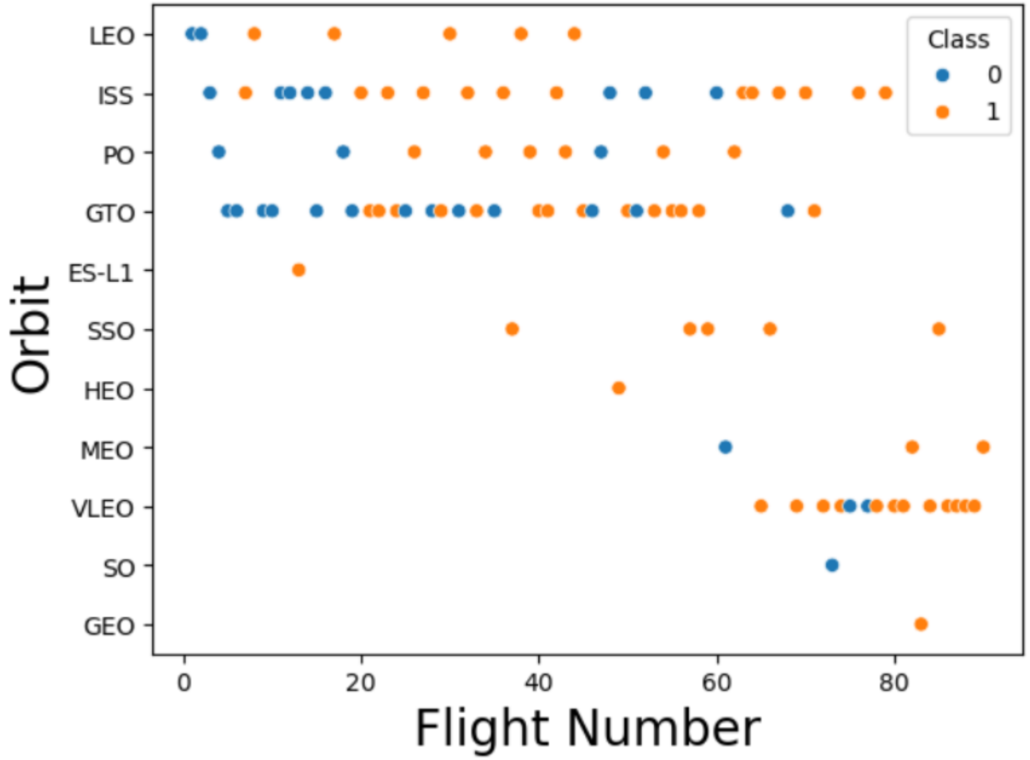


Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

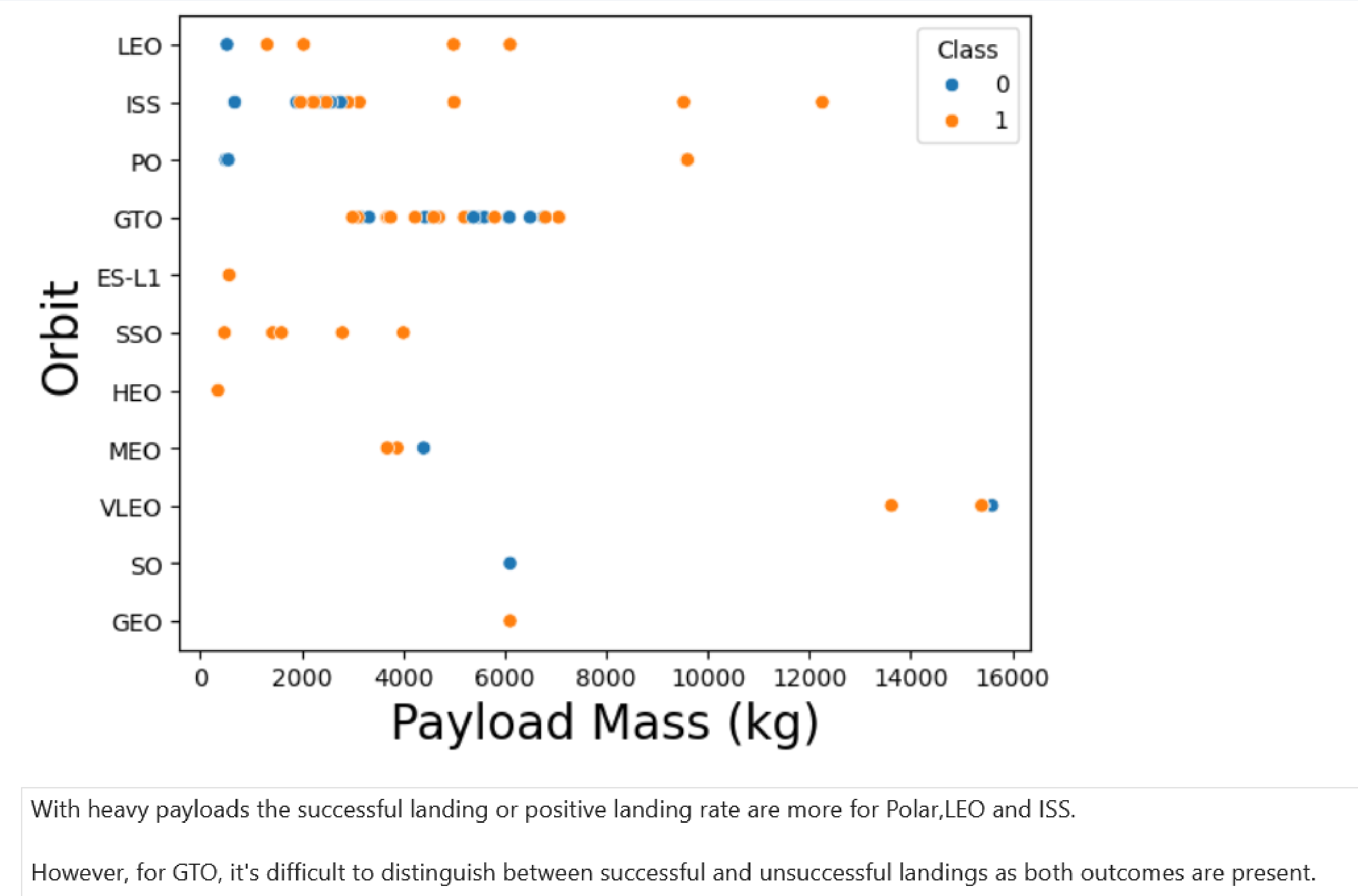


Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

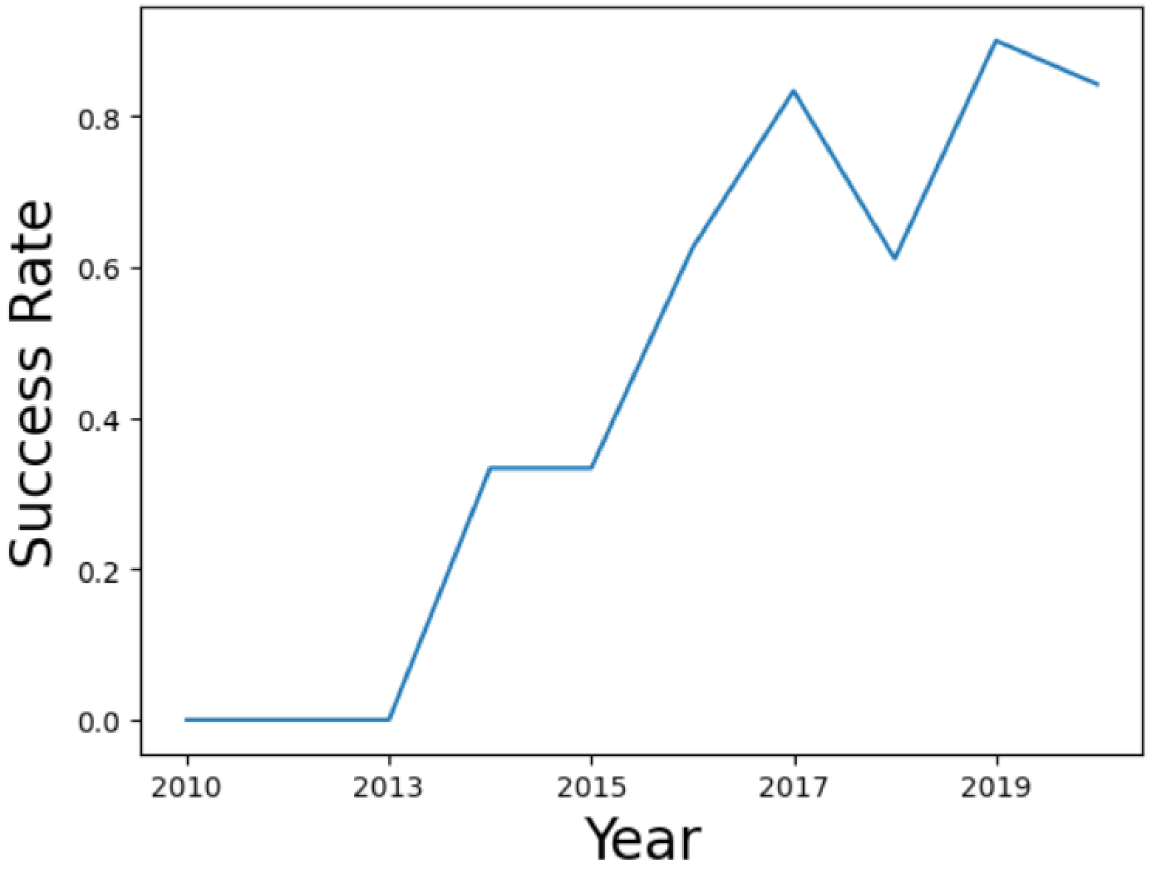
Payload vs. Orbit Type



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

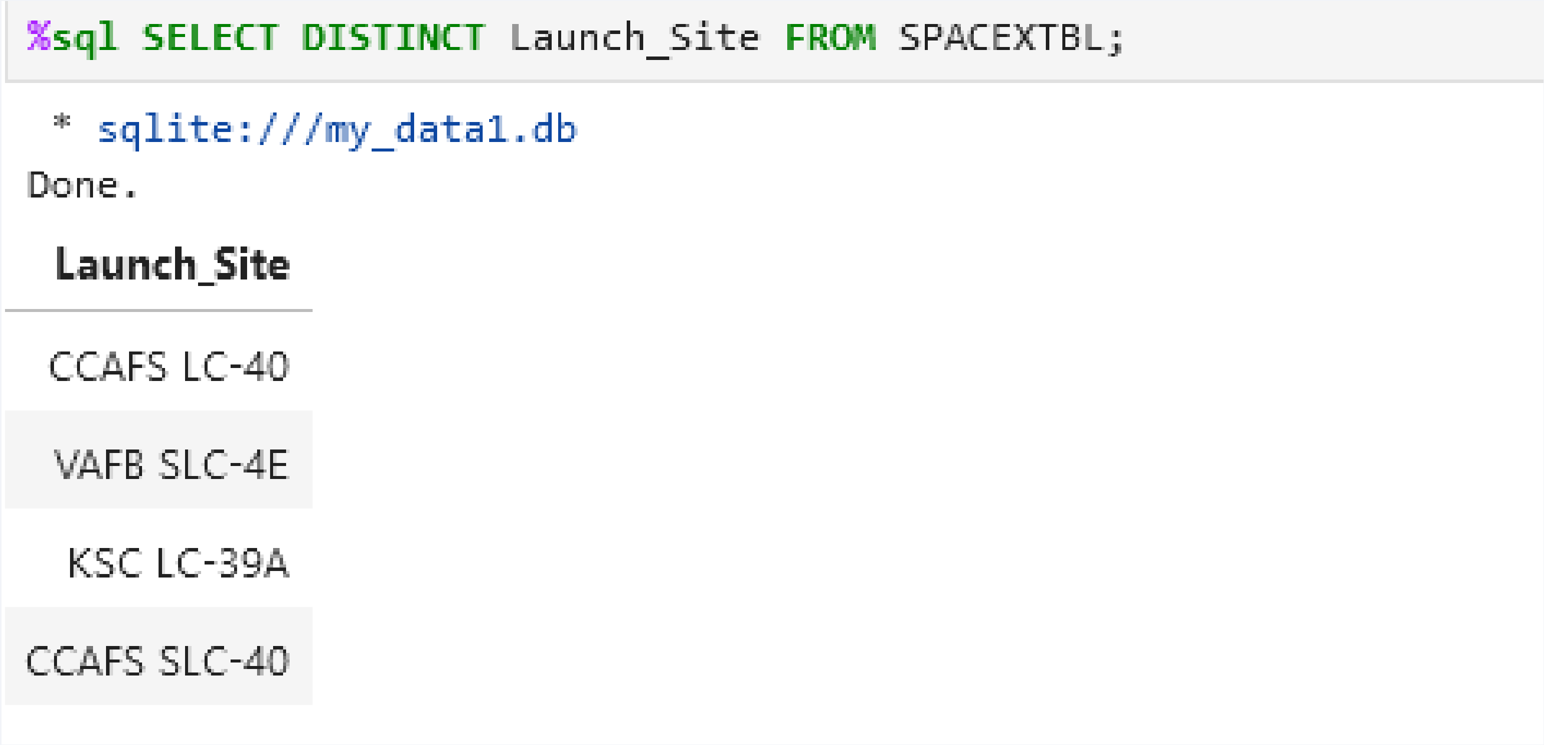
However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

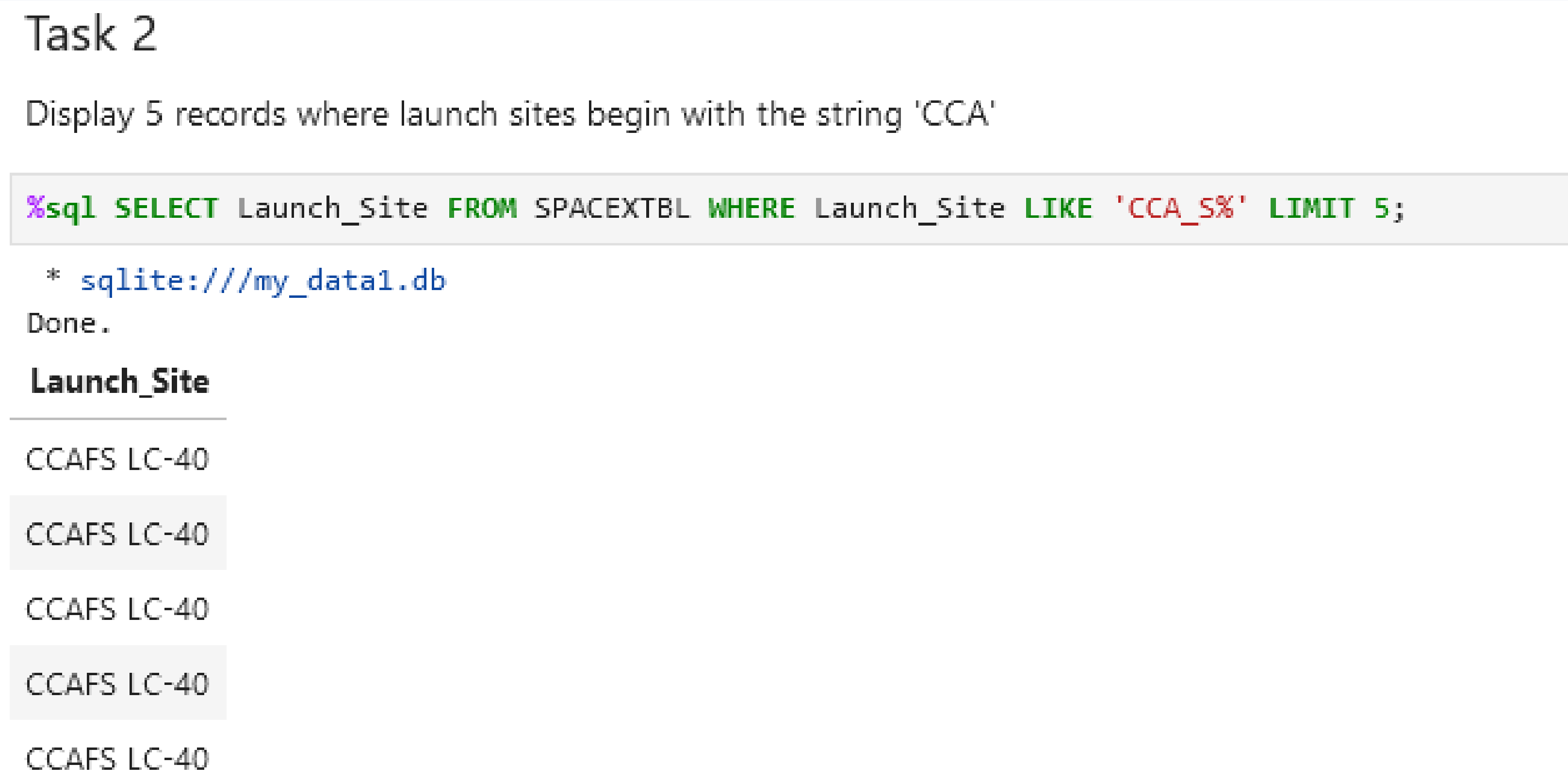


```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'



Task 2

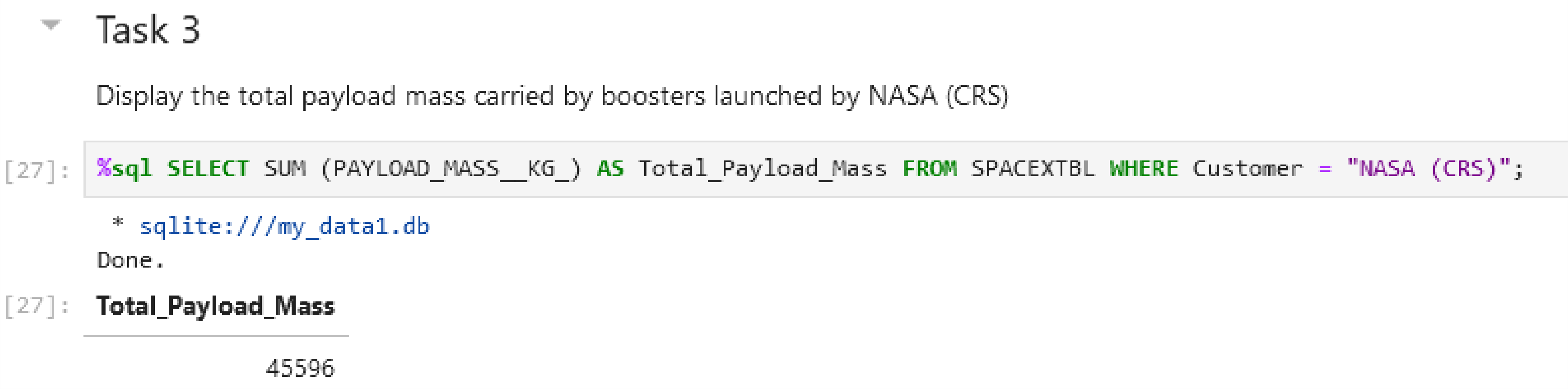
Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT Launch_Site FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA_S%' LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass



Task 3

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

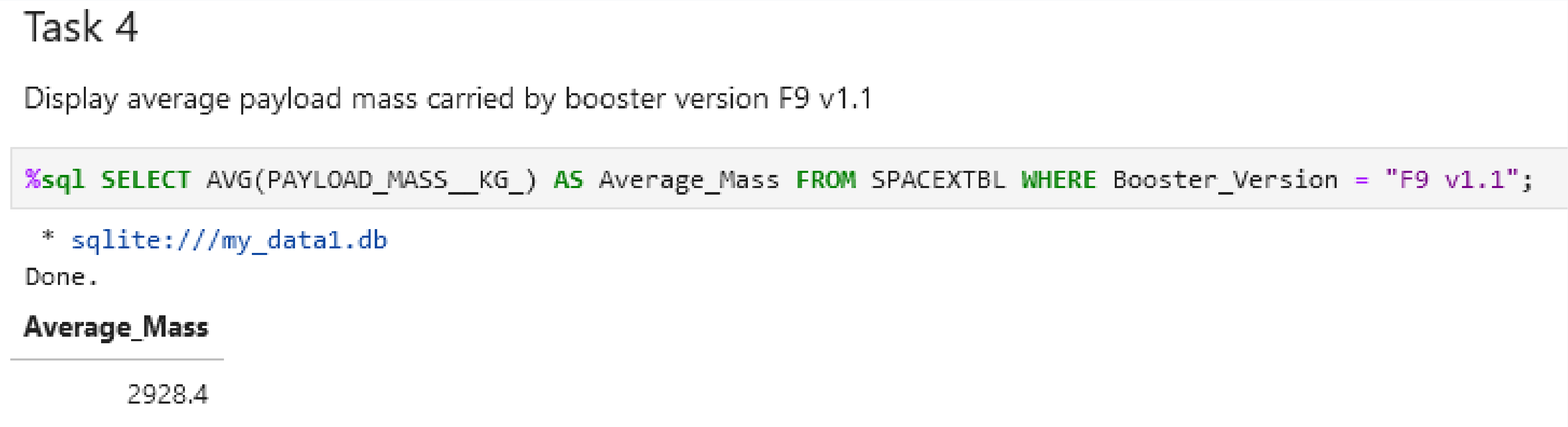
```
[27]: %sql SELECT SUM (PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTBL WHERE Customer = "NASA (CRS)";
```

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

[27]:

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1



Task 4

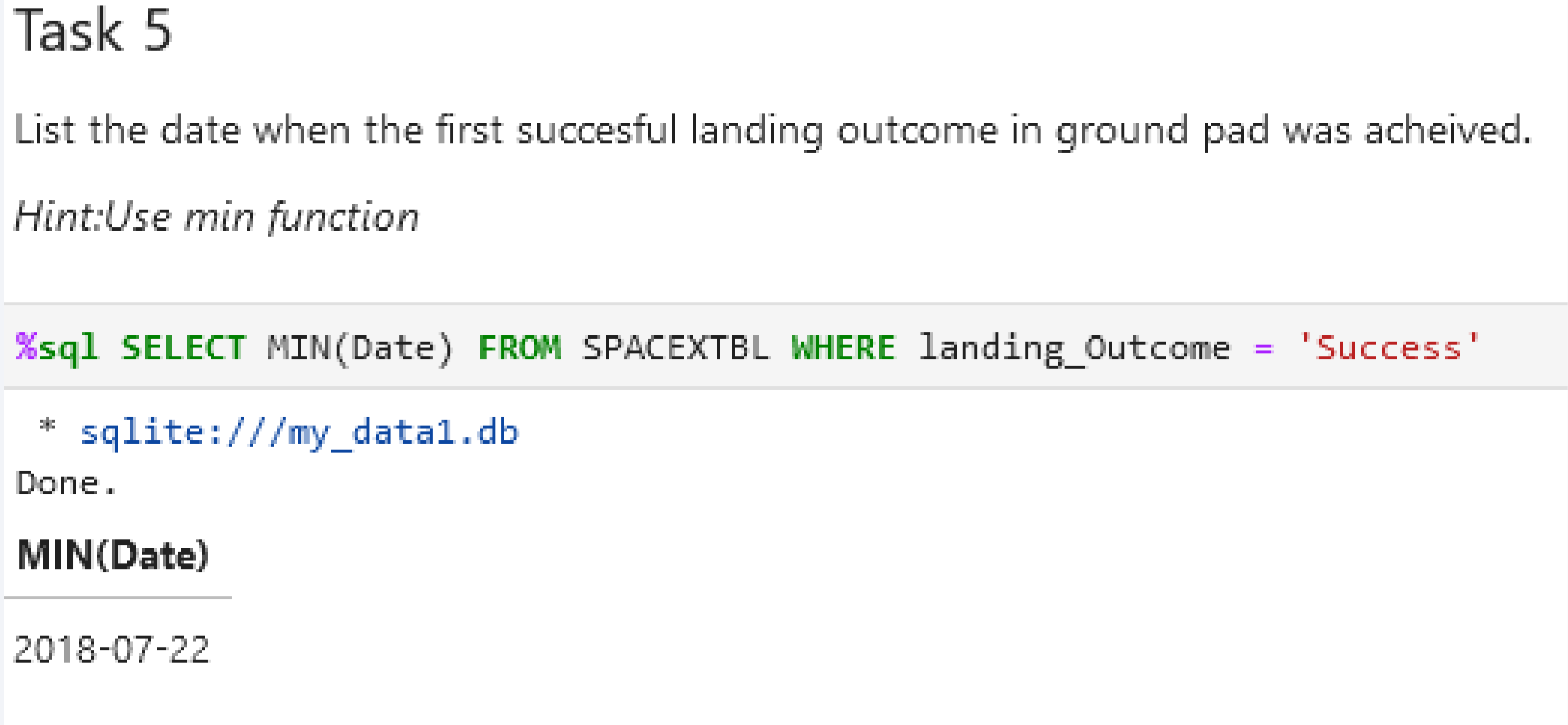
Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Mass FROM SPACEXTBL WHERE Booster_Version = "F9 v1.1";
```

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

Average_Mass
2928.4

First Successful Ground Landing Date



Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

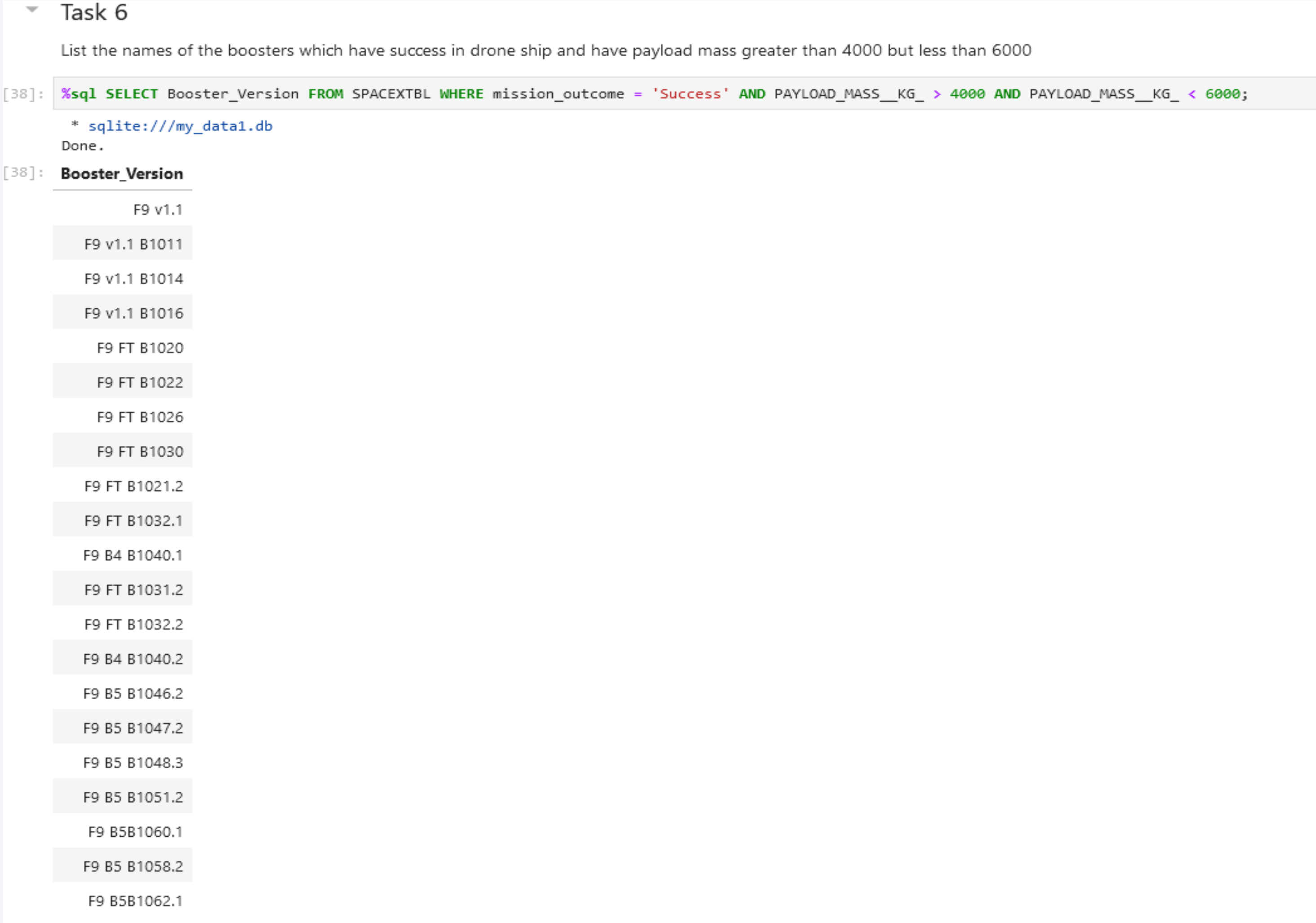
Hint:Use min function

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE landing_Outcome = 'Success'
```

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

MIN(Date)
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000



Task 6

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

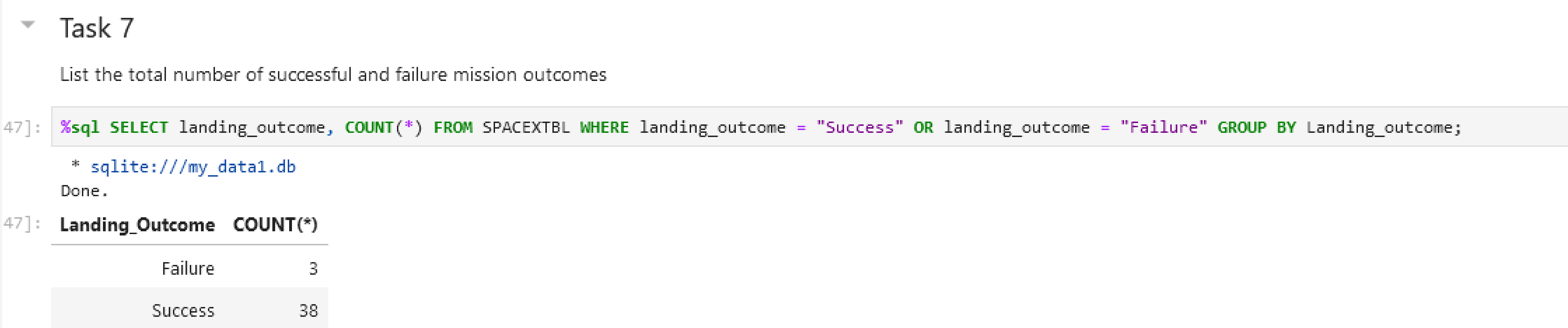
```
[38]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE mission_outcome = 'Success' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

[38]:

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes



Task 7

List the total number of successful and failure mission outcomes

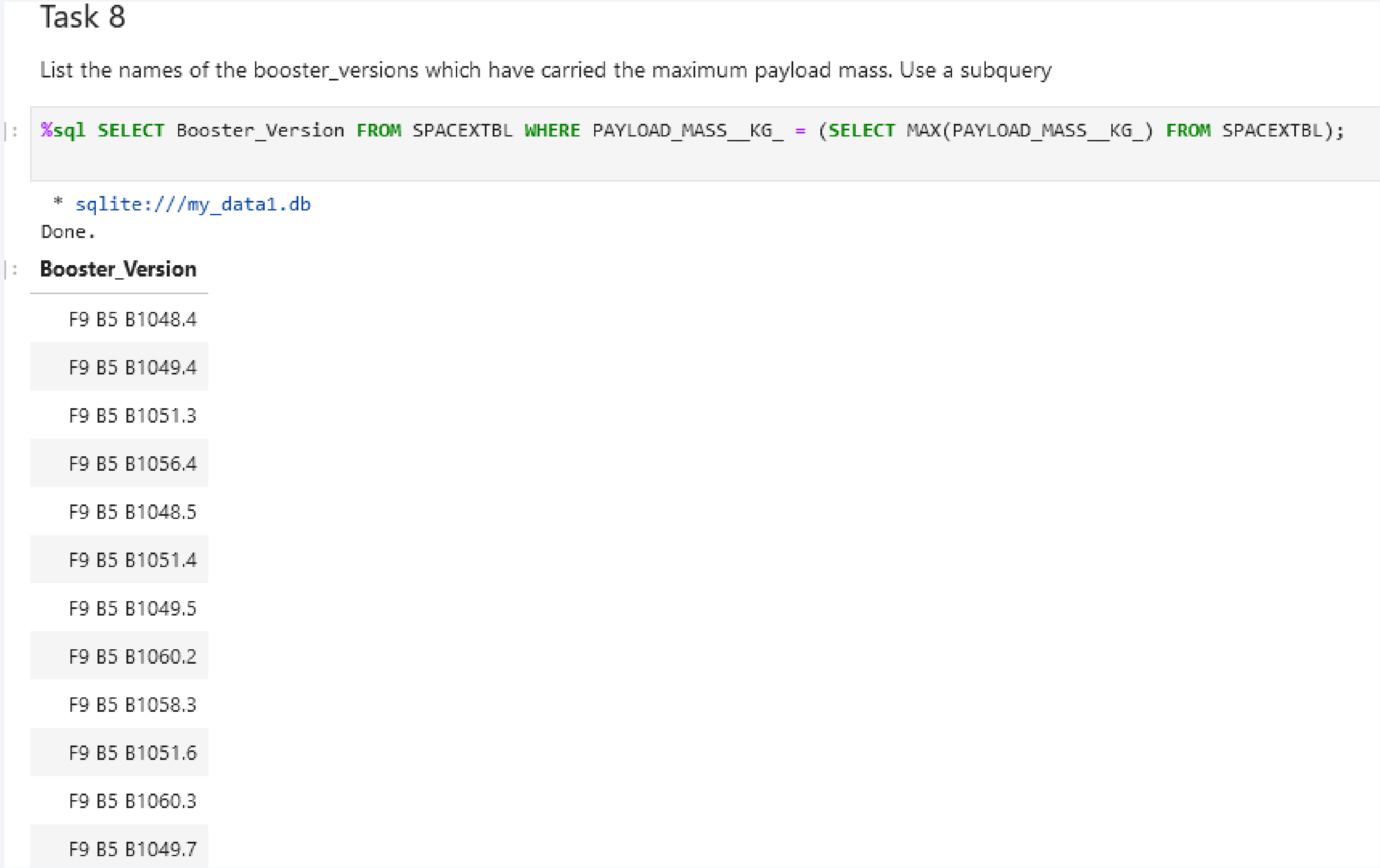
```
47]: %sql SELECT landing_outcome, COUNT(*) FROM SPACEXTBL WHERE landing_outcome = "Success" OR landing_outcome = "Failure" GROUP BY Landing_outcome;
```

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

47]:

Landing_Outcome	COUNT(*)
Failure	3
Success	38

Boosters Carried Maximum Payload



Task 8

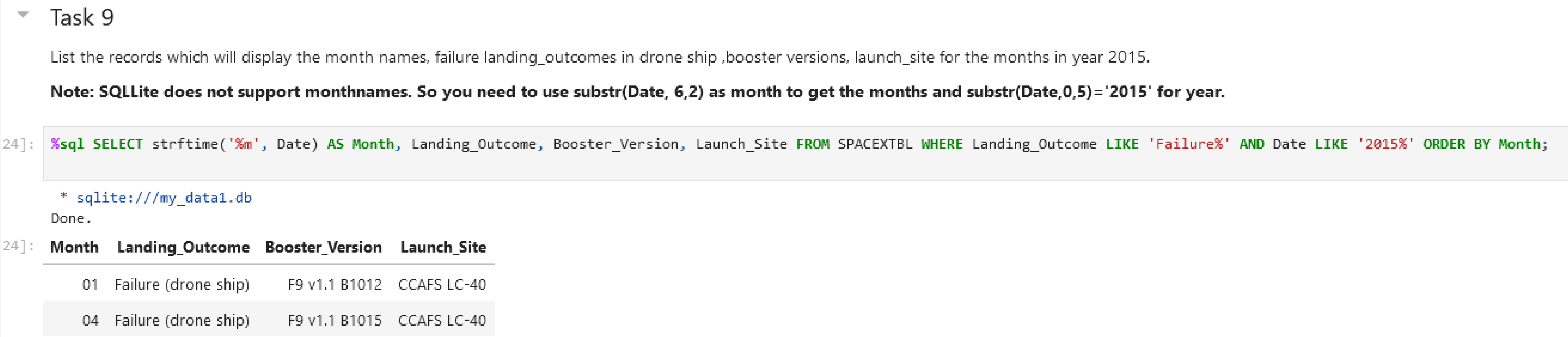
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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



Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

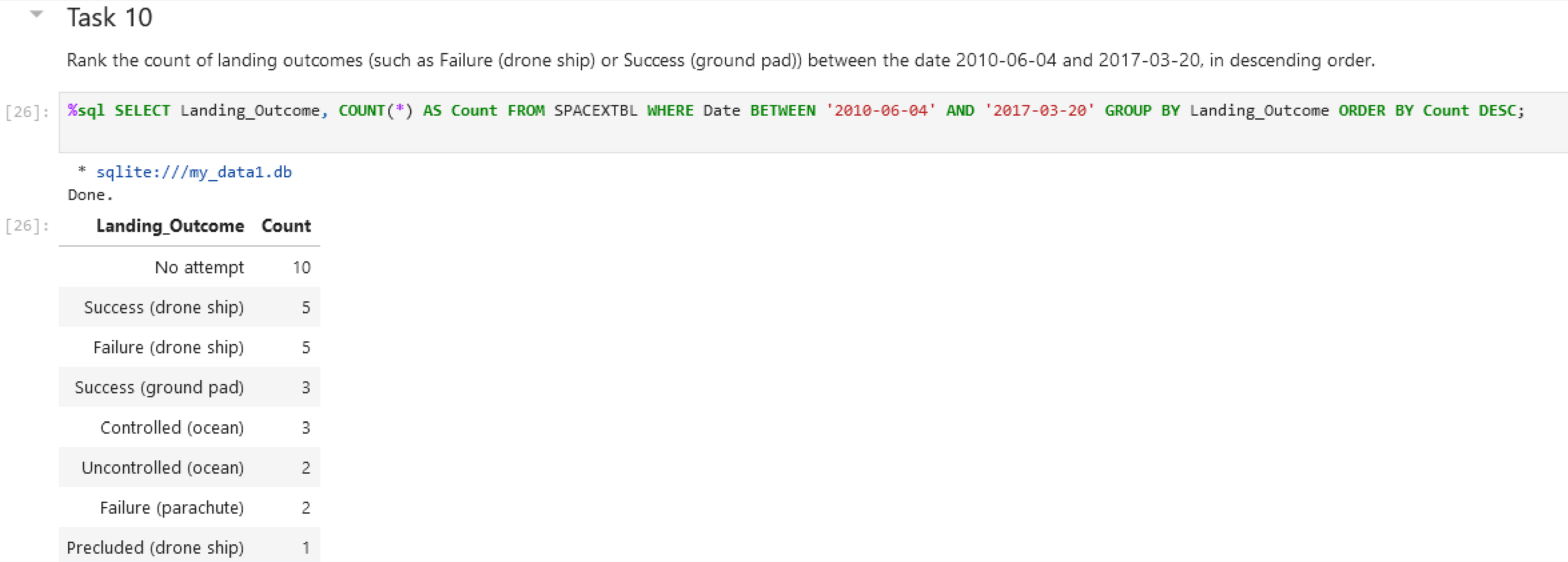
```
24]: %sql SELECT strftime('%m', Date) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%' AND Date LIKE '2015%' ORDER BY Month;
```

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

24]:

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



Task 10

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.

```
[26]: %sql SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Count DESC;
```

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

[26]:

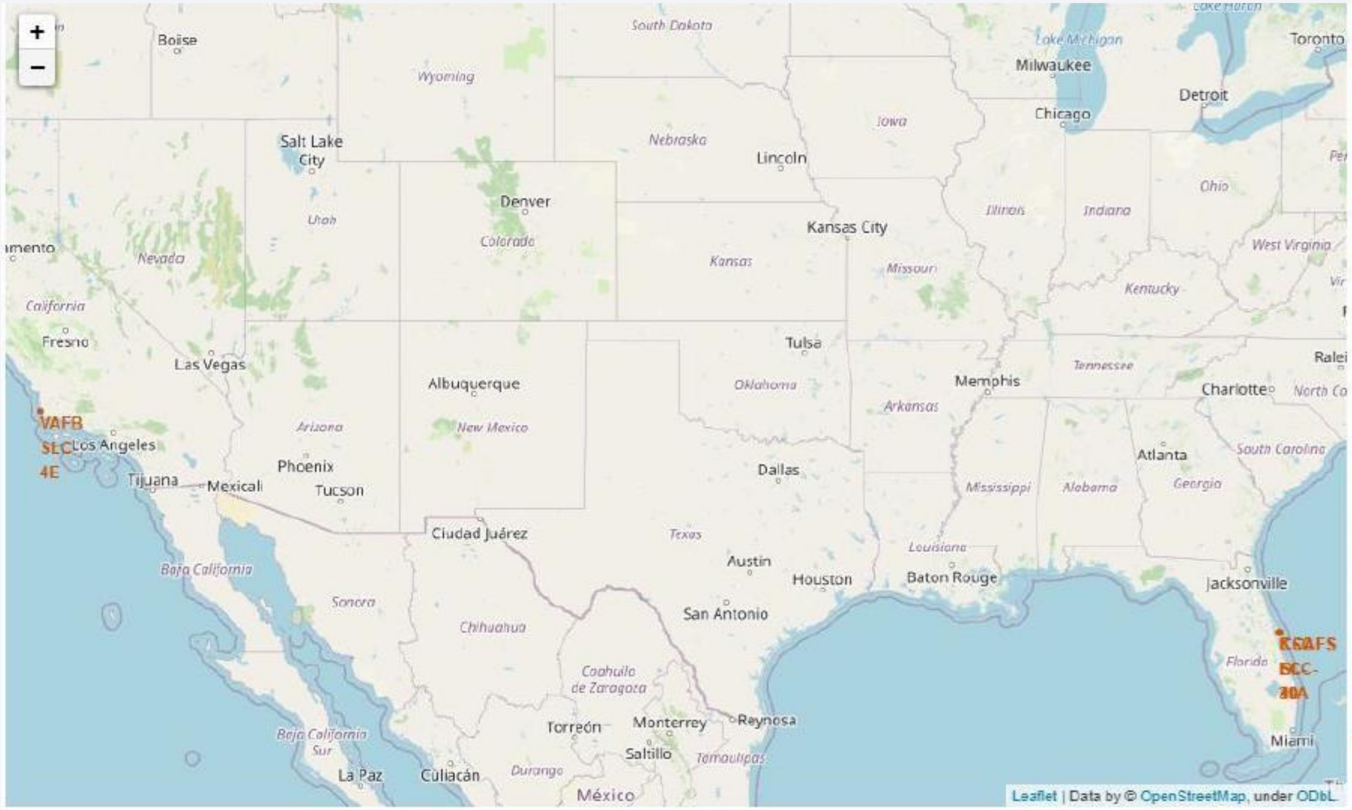
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

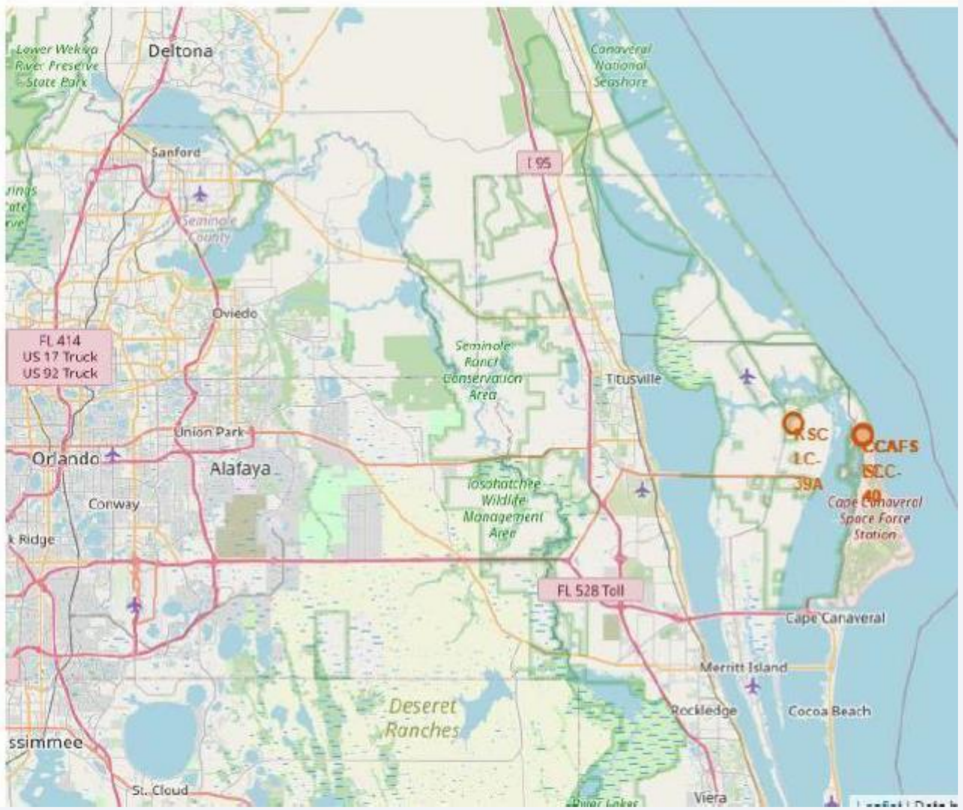
Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

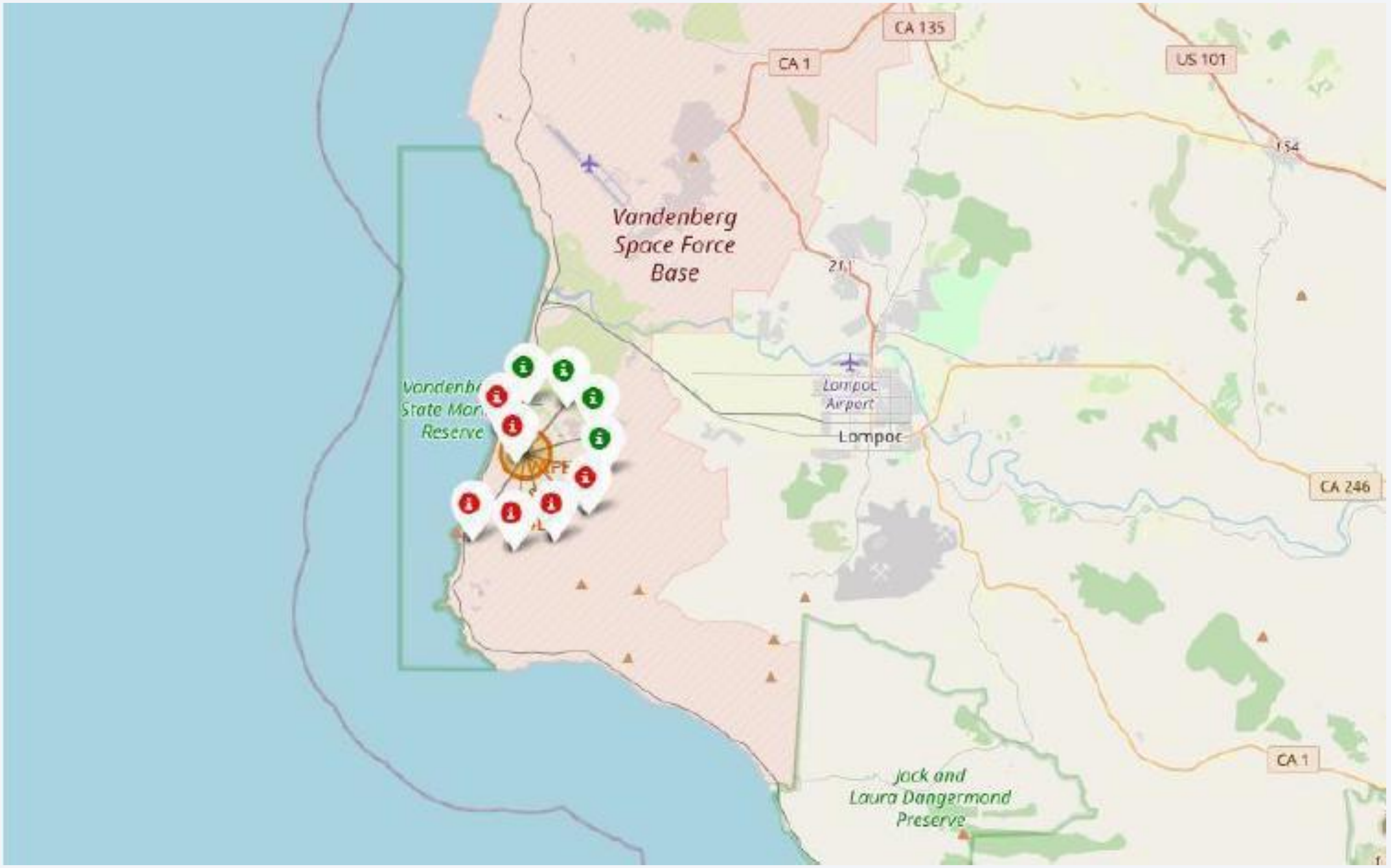
All launch sites are displayed in relation to the US map on the left. Due to their near proximity, the two launch sites in Florida are depicted on the right map. Every launch location is close to the sea.





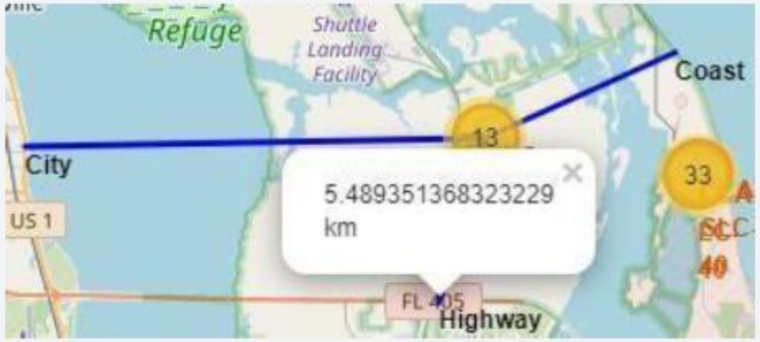
<Folium Map Screenshot 2>

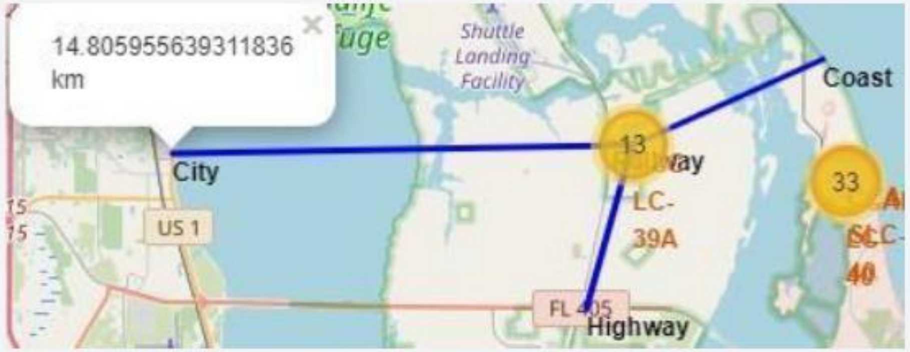
Clicking on clusters on the Folium map will show each successful landing (green icon) and unsuccessful landing (red icon). VAFB SLC-4E displays four successful landings and six unsuccessful landings in this example.

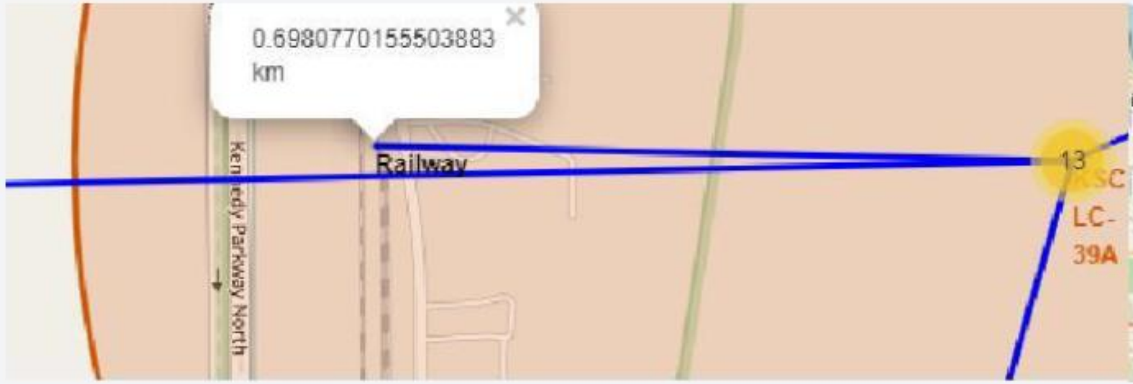

CA 135
CA 1
US 101
154
Vandenberg Space Force Base
Vandenberg State Marine Reserve
Lompoc Airport
Lompoc
CA 246
CA 1
Jack and Laura Dangermond Preserve

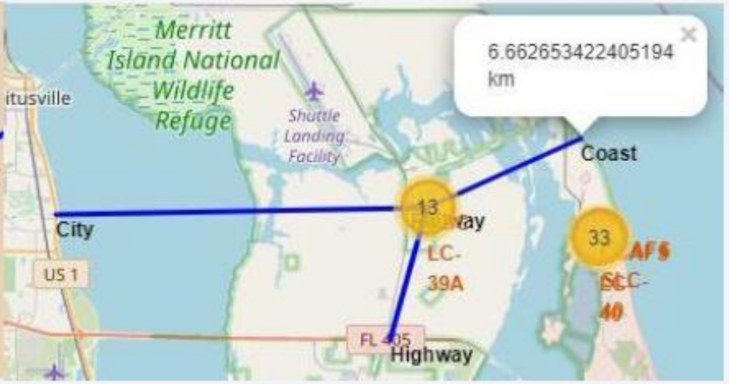
<Folium Map Screenshot 3>

Using the KSC LC-39A as an example, launch locations are generally located near railroads that provide transportation. Highways for the transportation of people and supplies are near launch locations. In order to prevent rockets from falling on heavily populated areas, launch sites are also situated near coasts and somewhat away from towns. This allows for the possibility of a launch failure landing in the sea.









Section 5

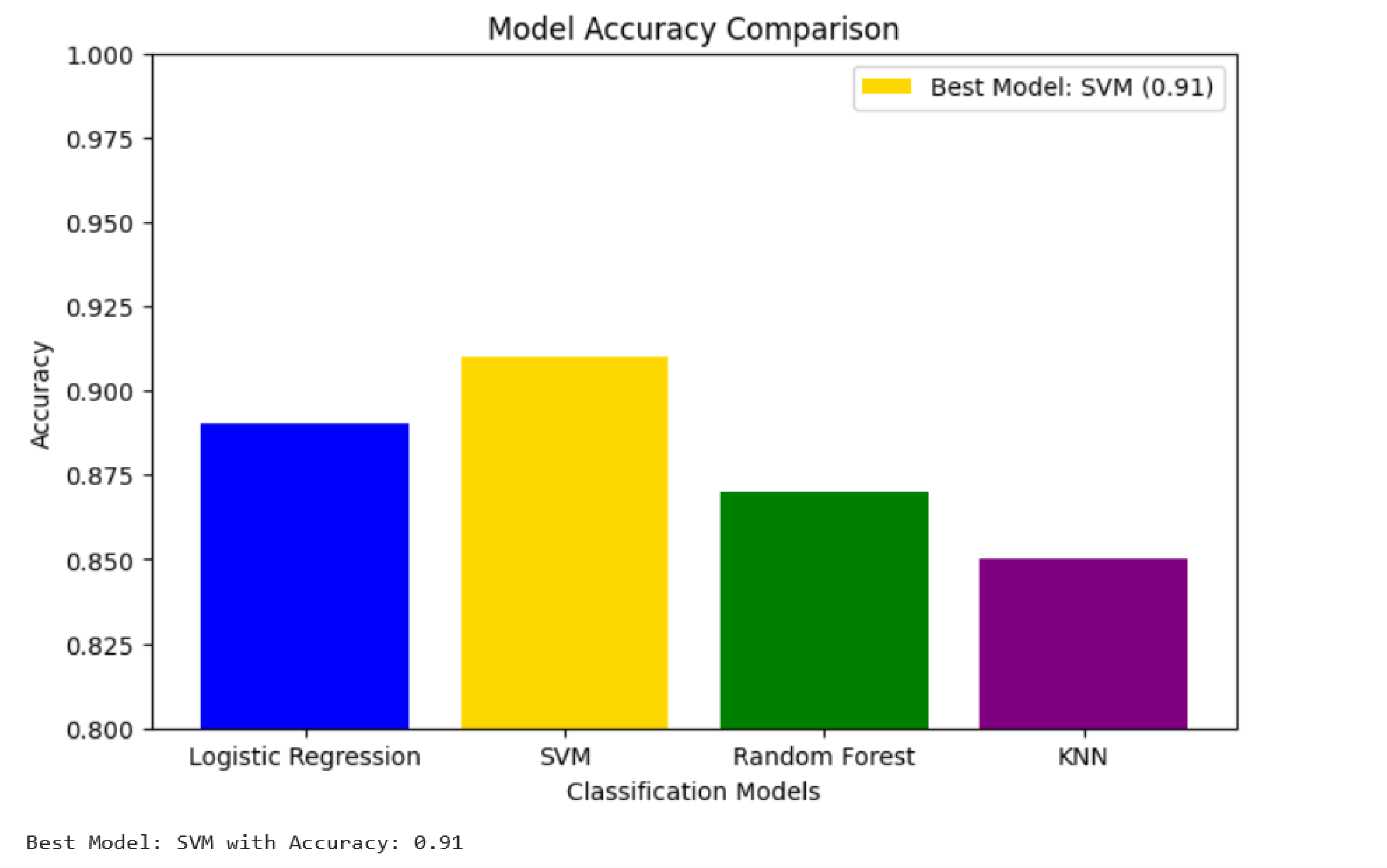
Predictive Analysis (Classification)

Classification Accuracy

- On the test set, the accuracy of each model was 83.33%, which was almost identical. It should be mentioned that the test size is tiny, with only 18 samples.

- This can lead to a wide range of accuracy outcomes, as those obtained from repeated runs of the Decision Tree Classifier model.

- In order to identify the best model, we certainly need more data.



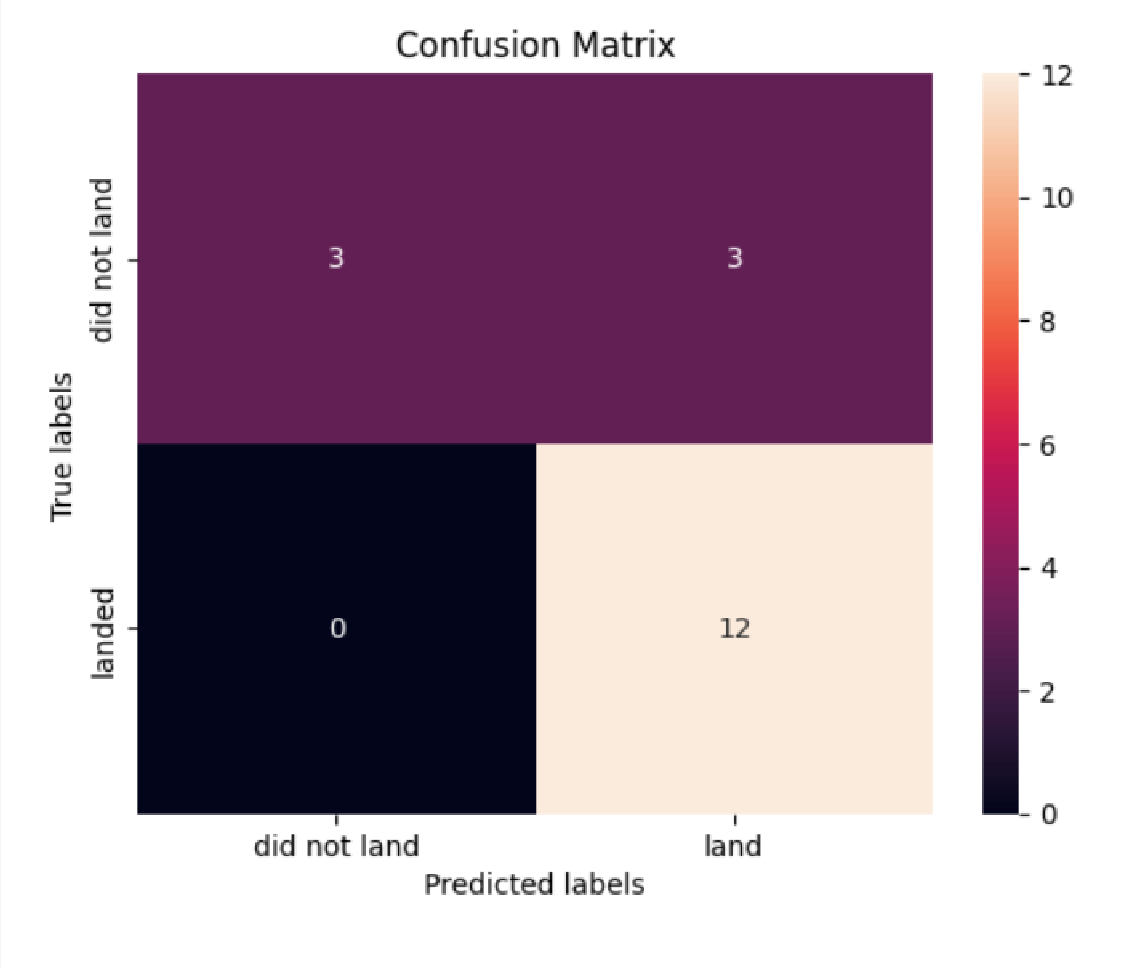
```
Best Model: SVM with Accuracy: 0.91
```

Confusion Matrix

- The confusion matrix is the same for all models since they all performed identically on the test set. Twelve successful landings were predicted by the models when the true label was successful landing.

- When the actual label was failure landing, the models projected three unsuccessful landings.

- When unsuccessful landings (false positives) were the true label, the models projected three successful landings. Successful landings are overpredicted by our models.



Conclusions

- Our goal was to develop a machine learning model that would enable SpaceY to bid against SpaceX.
- The goal of the model is to forecast that Stage 1 will land successfully, potentially saving about $100 million USD.
- However, landing success alone does not determine the overall cost efficiency. Additional factors such as refurbishment costs, launch delays, and operational expenses should be considered for a full cost analysis.
- Data was collected via scraping the SpaceX Wikipedia page and using a public SpaceX API.
- A DB2 SQL database was used to hold the data, and data labels were made.
- To create a comprehensive decision-making model, additional cost factors should be integrated. Future work could include a financial analysis component to compare the economic impact of successful and unsuccessful landings.
- The accuracy of the machine learning models were 83%.
- In order to determine whether to move forward with the launch, Elon Mask of SpaceY can use this model to forecast, with high accuracy, the likelihood of a successful Stage 1 landing.

Thank you!