IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tejas K

1 July 2024

Outline

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

Executive Summary

- This project leverages historical data to predict the success of Falcon 9 first stage landings, providing critical insights into cost efficiency for space launches. Utilizing techniques such as data wrangling, exploratory data analysis with SQL, Pandas, and Matplotlib, interactive visual analytics using Folium and Plotly Dash, and predictive modeling with Scikit-Learn, we derived several key findings:
- **Evolution of Landing Success**: Analysis from 2013 to 2020 indicates a significant increase in successful landings, showcasing SpaceX's advancements in reusable rocket technology.

Executive Summary

- **Orbit Type and Payload Mass**: Meaningful relationships between orbit type, payload mass, and landing success rates were identified. Specific orbit types and payload masses were found to correlate with higher success probabilities.
- **Predictive Modeling**: Our predictive models demonstrated robust accuracy in forecasting landing success, providing a valuable tool for assessing potential cost savings and competitive bidding strategies for alternate companies.
- The project's methodology included data collection via JSON files and web scraping, followed by data wrangling to identify and process missing values, and exploratory data analysis using visualization tools. We then employed interactive visual analytics and predictive analysis using classification models to test and select the most accurate model. Our findings offer strategic insights for improving space mission outcomes and cost efficiency.

Introduction

- The landscape of space travel has experienced a significant transformation with the advent of reusable rocket technology. SpaceX, a pioneer in this field, offers Falcon 9 rocket launches at a cost of 62 million dollars per launch, in contrast to other providers whose costs exceed 165 million dollars. A major factor contributing to this cost efficiency is the reusability of the Falcon 9's first stage.
- This project aims to predict the likelihood of a successful landing of the Falcon 9 first stage, a critical determinant in evaluating launch costs. The ability to accurately forecast landing success provides valuable insights for alternate companies bidding against SpaceX for rocket launches. The project encompasses data wrangling, data collection, exploratory data analysis, and predictive modeling to understand and predict the factors influencing landing success.
- Key questions addressed are:

  How has the success rate of Falcon 9 first stage landings evolved over the years?

  What is the relationship between orbit type, payload mass, and landing success rate?

  Can a predictive model accurately forecast the success of future landings based on historical data?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  The data was collected using JSON files and web scraping, respective python libraries were used to extract data from their respective sources.

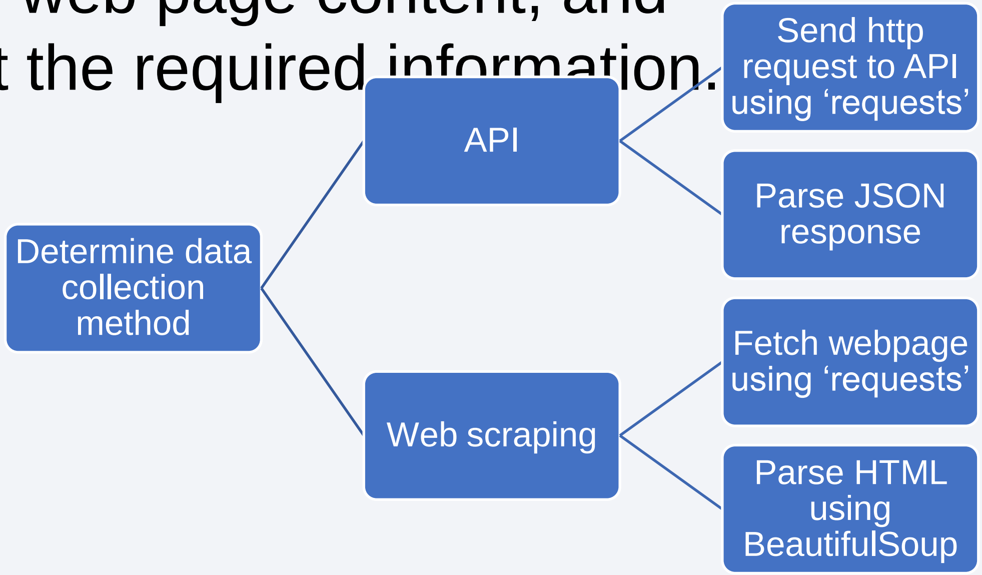
- Perform data wrangling

  First the percentage of missing values were identified, then the columns were identified as either numerical or categorical. Then basic tasks were performed on the data. It was then converted into a CSV file for further processing.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We test the data with each of the classification models and choose the one which gives us the most accurate data. We do this using scores and confusion matrices

Data Collection

- Data was collected using APIs with the requests library. By sending HTTP requests to the API endpoints, structured data (in JSON format) was retrieved.
- The libraries requests and BeautifulSoup were used to scrape data from websites. The requests library fetches the web page content, and BeautifulSoup parses the HTML to extract the required information.



Determine data collection method

API

Send http request to API using ‘requests’

Parse JSON response

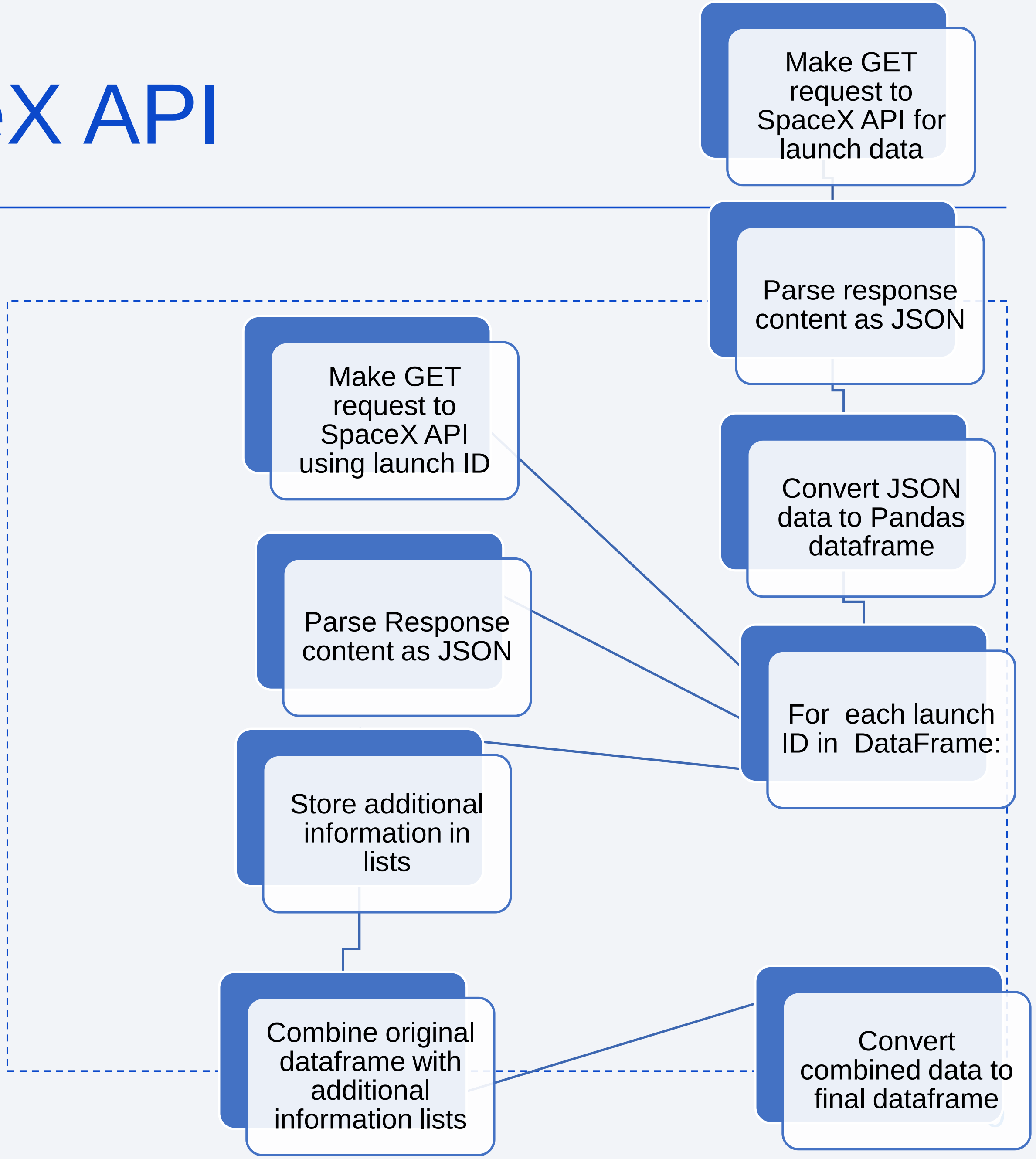
Web scraping

Fetch webpage using ‘requests’

Parse HTML using BeautifulSoup

Data Collection – SpaceX API

- First we request and parse the SpaceX launch data with the GET request.
- Then we decode the response content as a JSON and turn it into a pandas dataframe.
- Then we again use the API to get additional information related to launches using the IDs given for each launch.
- The information is then stored in lists and then ultimately converted to a dataframe.
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Make GET request to SpaceX API for launch data

Parse response content as JSON

Convert JSON data to Pandas dataframe

For each launch ID in DataFrame:

Make GET request to SpaceX API using launch ID

Parse Response content as JSON

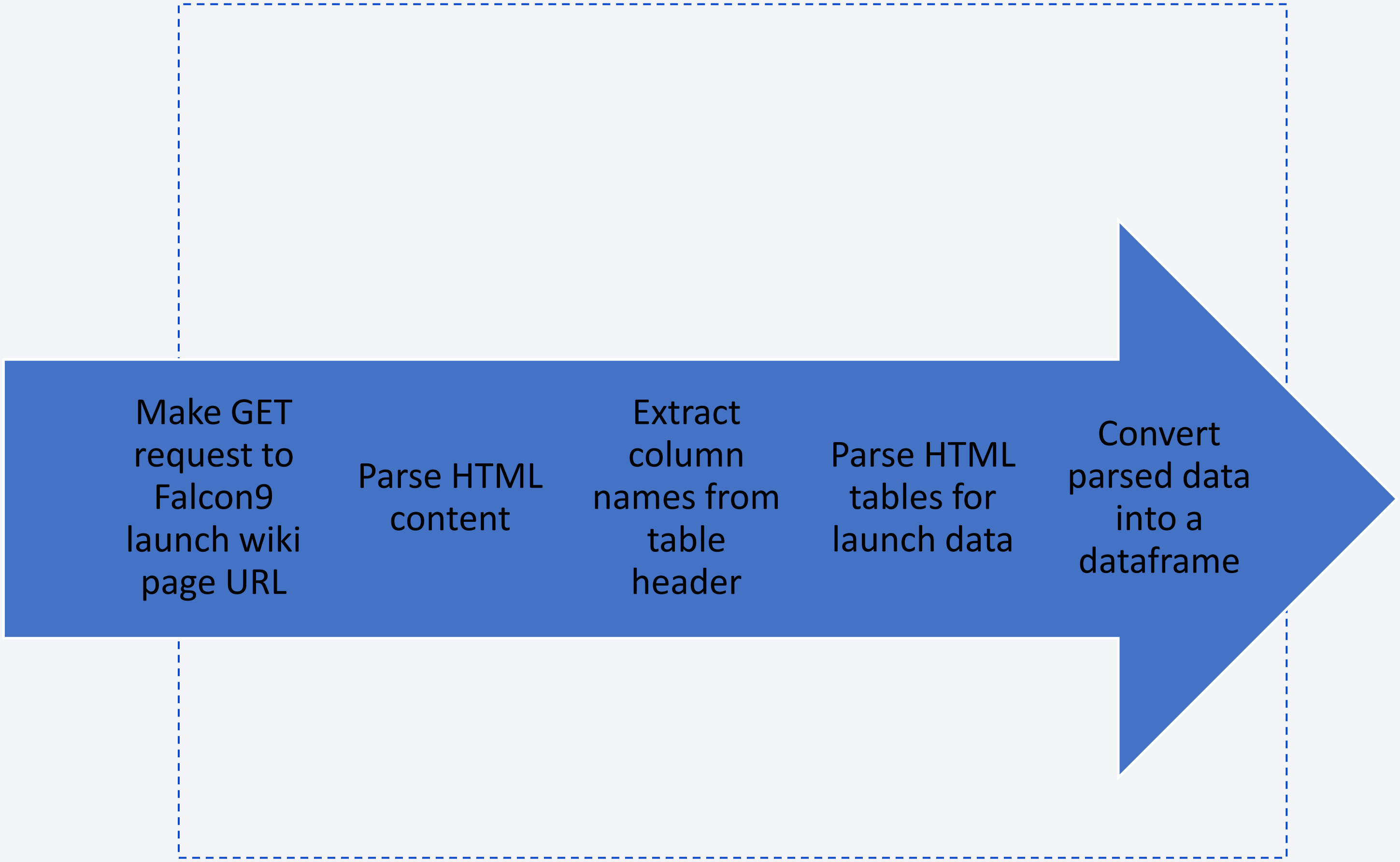
Store additional information in lists

Combine original dataframe with additional information lists

Convert combined data to final dataframe

Data Collection - Scraping

- First we request the Falcon9 launch wiki page from it’s URL.
- Then we extract all column names from the HTML table header.
- Then we create a dataframe by parsing the launch HTML tables.
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/jupyter-labs-webscraping%20(1).ipynb



Make GET request to Falcon9 launch wiki page URL

Parse HTML content

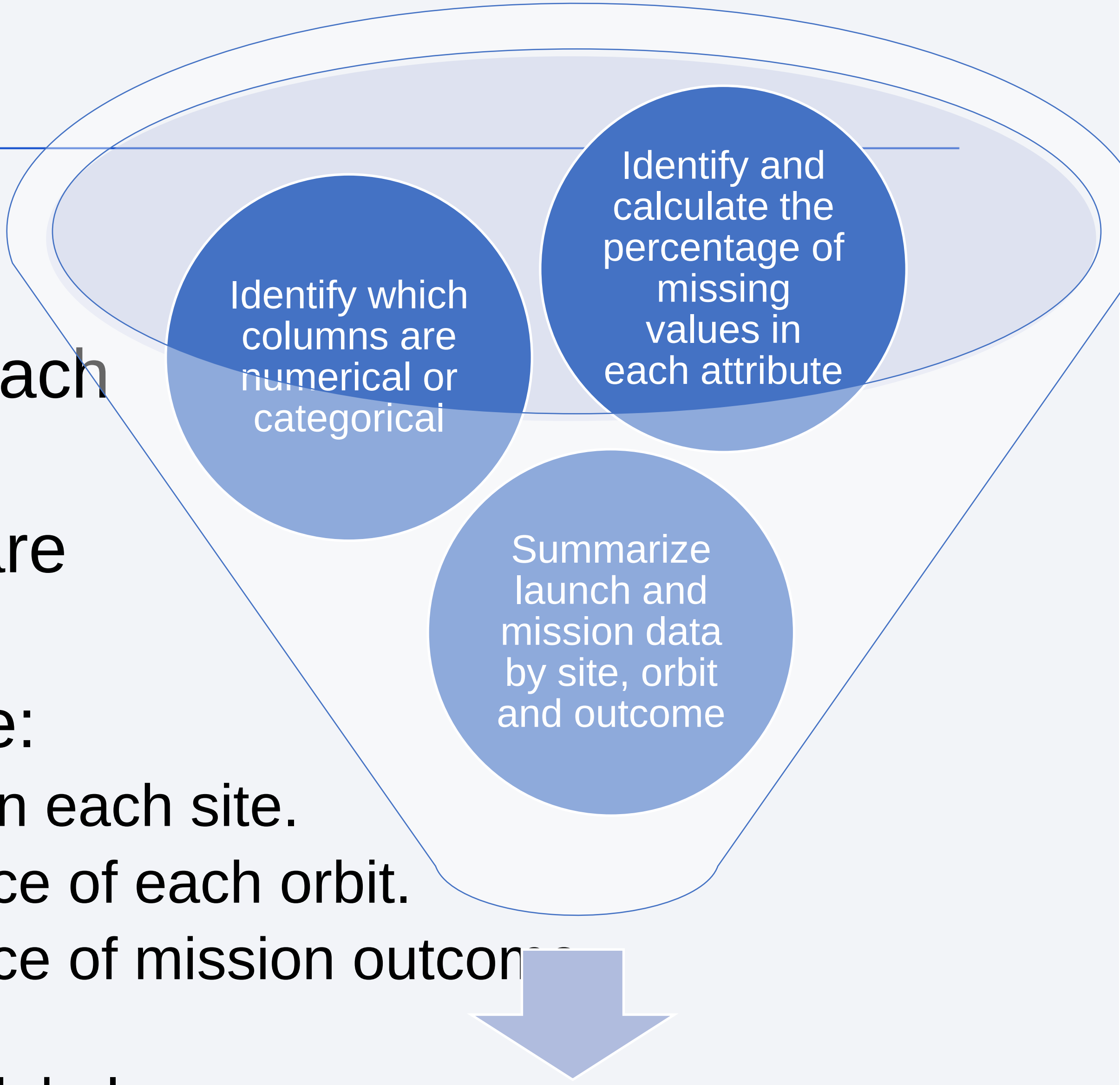
Extract column names from table header

Parse HTML tables for launch data

Convert parsed data into a dataframe

Data Wrangling

- First we identify and calculate the percentage of missing values in each attribute.
- Then we identify which columns are numerical or categorical.
- Then to determine launch data we:
  - ➢Calculate the number of launches on each site.
  - ➢Calculate the number and occurrence of each orbit.
  - ➢Calculate the number and occurrence of mission outcome of the orbits
  - ➢Finally, create the landing outcome label


Identify and calculate the percentage of missing values in each attribute
Identify which columns are numerical or categorical
Summarize launch and mission data by site, orbit and outcome

Successful creation of classification variable

https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

EDA with Data Visualization

- We performed correlations between Flight Numbers and Launch Sites, Payloads and Launch Sites, success rates of orbit types, Flight Numbers and Orbit types, Payloads and Orbit types, and performed a correlation which shows the yearly trend in launch success.
- These charts are essential for identifying key patterns and relationships in space mission data, which can help optimize future launches, improve mission success rates, and enhance our understanding of how different variables influence spaceflight outcomes.
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/edadataviz.ipynb

EDA with SQL

- In this intricate analysis of space mission data, we enumerated unique launch sites, extracted five records with launch sites starting with 'CCA,' calculated the total payload mass of NASA's CRS missions, and determined the average payload mass for booster version F9 v1.1.
- Additionally, we identified the date of the first successful ground pad landing, listed boosters with successful drone ship landings carrying 4000-6000 units, and tallied the total number of successful and failed mission outcomes.
- Furthermore, we identified booster versions with the maximum payload mass using a subquery, displayed 2015 records with specific failure outcomes on drone ships, and ranked landing outcomes between 2010 and 2017 in descending order.
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- We added markers for VAFB SLC 4E, KSC LC 39A, CCAFS SLC 40, and CCAFS LC 40, and then added distance lines from space stations to the nearest coastline, railways, city, and highway.
- The markers for VAFB SLC 4E, KSC LC 39A, CCAFS SLC 40, and CCAFS LC 40 were added to provide clear and distinct identifiers for specific launch sites within the space mission data.
- These markers serve to facilitate quick reference and analysis of launch activities associated with each site, aiding in spatial and logistical planning, operational coordination, and historical tracking of space missions.

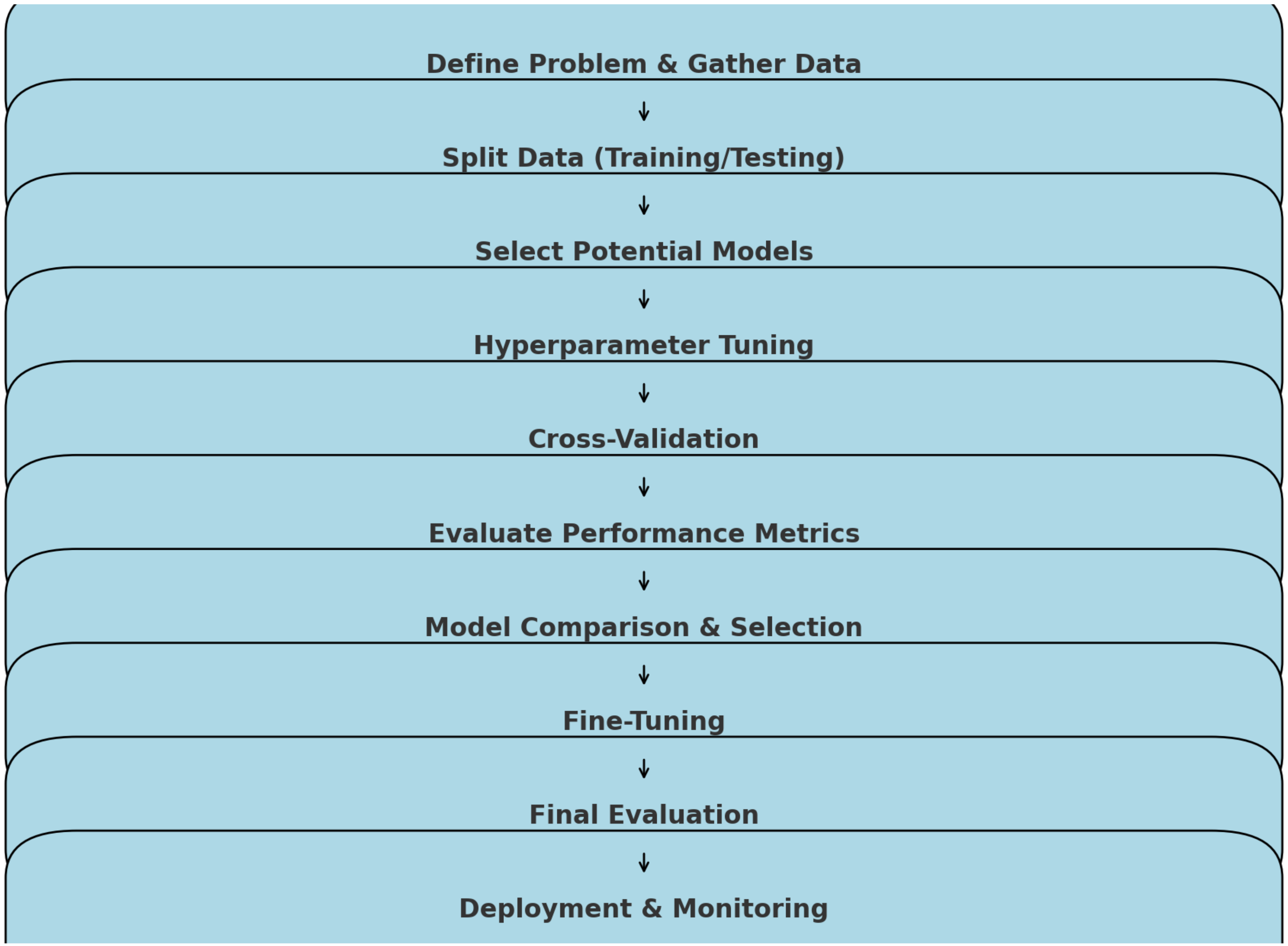
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- We incorporated pie charts to present comprehensive data on all launch sites collectively, as well as for each launch site individually. Additionally, we introduced a category plot to illustrate the success rate of various booster versions, both in aggregate and separately.
- We added pie charts to provide a clear and intuitive visualization of the distribution of launches across all sites and to highlight the specific contributions of each individual site.
- This approach ensures a comprehensive understanding of site-specific performance and trends.
- Additionally, we introduced a category plot to illustrate the success rates of different booster versions, both collectively and individually, enabling stakeholders to identify which versions excel and which may require further refinement.
- This enhances the clarity and actionable insights of our data presentation.
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/Dashboard%20(2).docx

Predictive Analysis (Classification)

- To find the optimum classification model, we start by defining the problem and gathering data. (Flowchart in next slide)
- The data is then split into training and testing sets. We select potential models and perform hyperparameter tuning.
- Cross-validation is used to ensure consistency, followed by evaluating performance metrics.
- Based on these evaluations, we compare and select the best model, fine-tune it, and conduct a final evaluation.
- The chosen model is then deployed and monitored to ensure it performs well in real-world scenarios.
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Define Problem & Gather Data

↓

Split Data (Training/Testing)

↓

Select Potential Models

↓

Hyperparameter Tuning

↓

Cross-Validation

↓

Evaluate Performance Metrics

↓

Model Comparison & Selection

↓

Fine-Tuning

↓

Final Evaluation

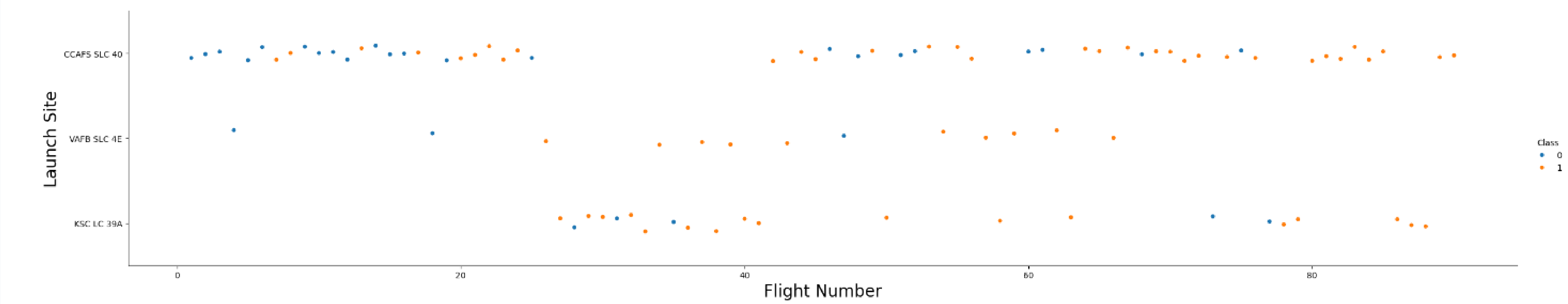
↓

Deployment & Monitoring

Section 2

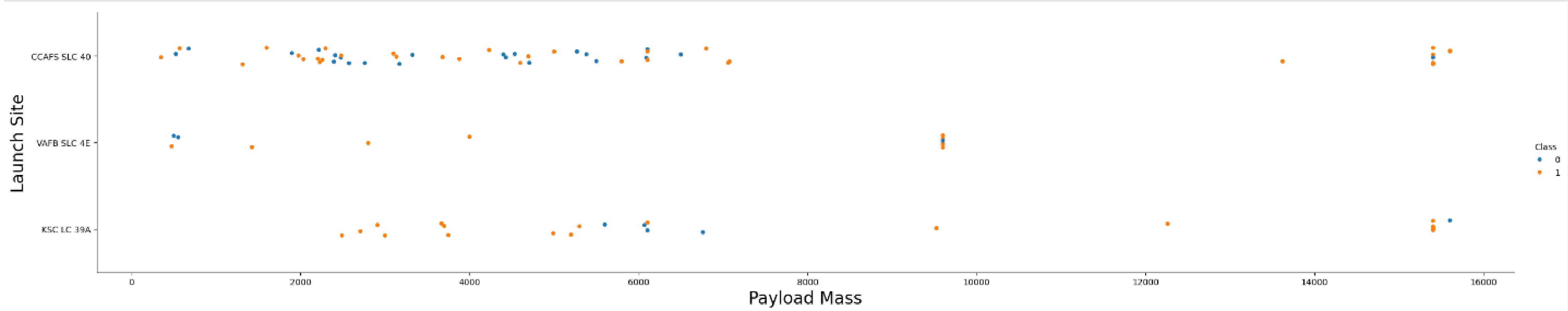
Insights drawn from EDA

Flight Number vs. Launch Site



- We observe in this scatter plot that in all cases, as the number of launches increase, the success rate also increases.
- We can also note that at the launch site VAFB SLC 4E has had a high success rate in general.
- An interesting thing to note is that at the launch site CCAFS SLC 40 there seems to be a long gap between launches between flight numbers 20 and 40 suggesting some deformities.
- We also can notice that at the launch site KSC LC 39A, the launches have started after some time suggesting that it is a newer launch site.

Payload vs. Launch Site



- In this scatter plot we notice that there is a better landing outcome as the payload mass increases.

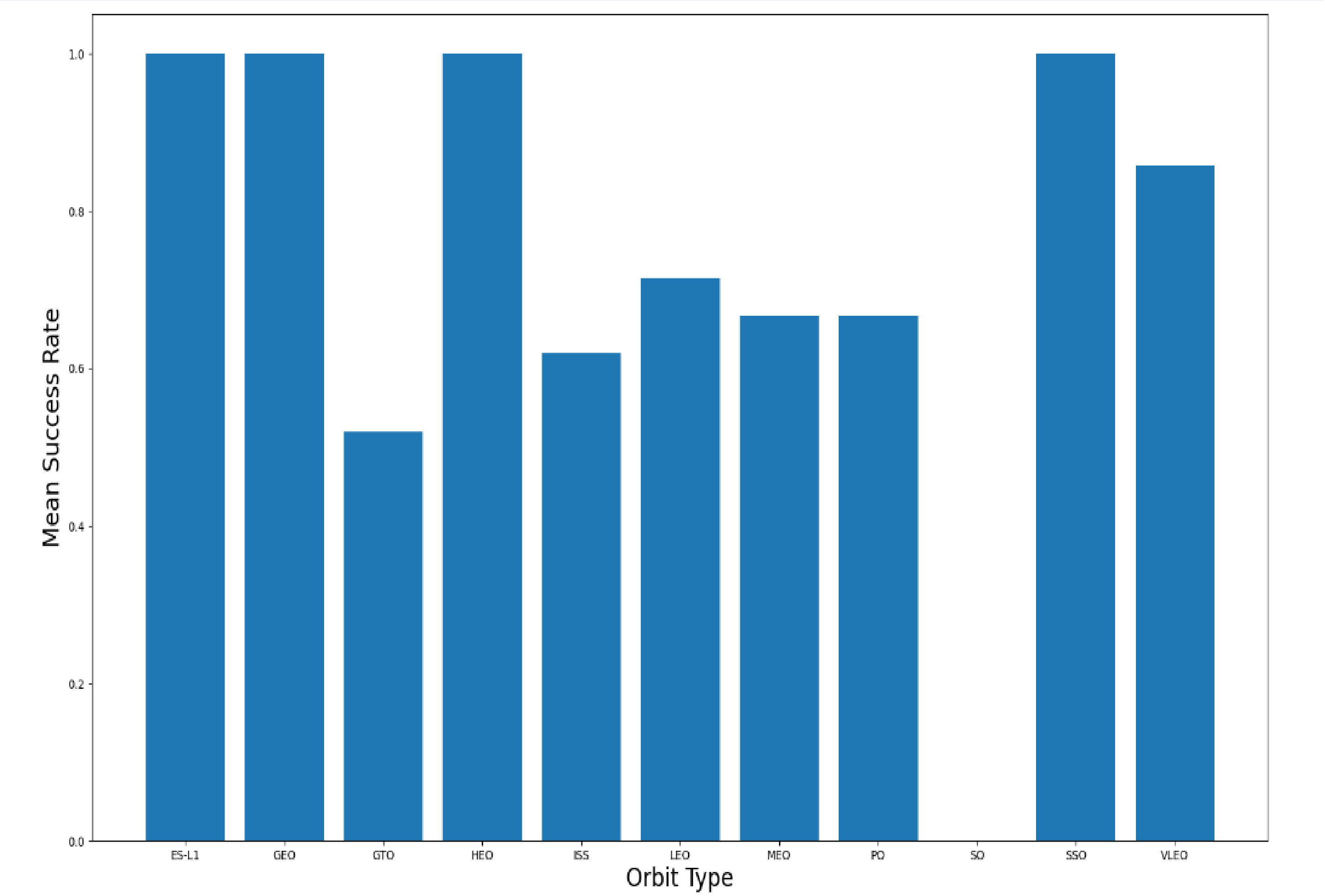
- We also notice that there are many launches where the payload mass is lesser than 10000 kg, which suggests that the average load is below that.

- We can also notice very few launches at the VAFB SLC 4E center, suggesting that it is not an active hotspot for launches.

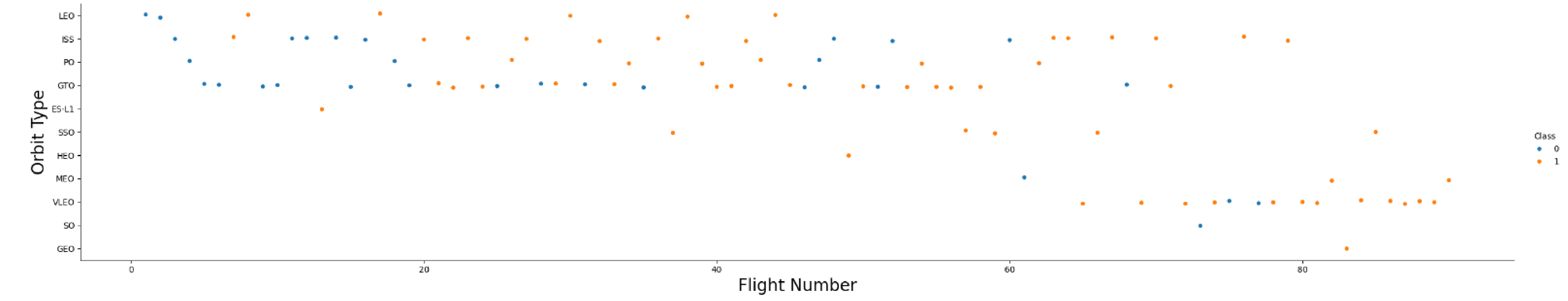
- We can see that the KSC LC 39A space center has a better landing outcome with lower payload masses(<6000 kg), suggesting that it is a better place to launch low payload rockets.

Success Rate vs. Orbit Type

- In this bar chart we can observe that the average success rate is quite high.
- We also notice that the orbit types which are at a higher height or a very low height have a noticeably higher success rate compared to medium altitude orbits.
- This gives us an idea of what type of orbits usually work more effectively, thus helping us in organizing their production and deployment accordingly

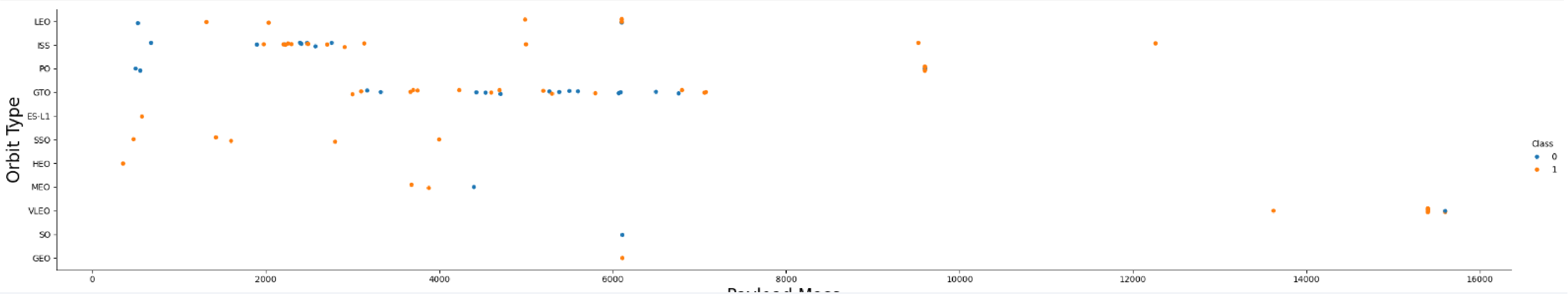


Flight Number vs. Orbit Type



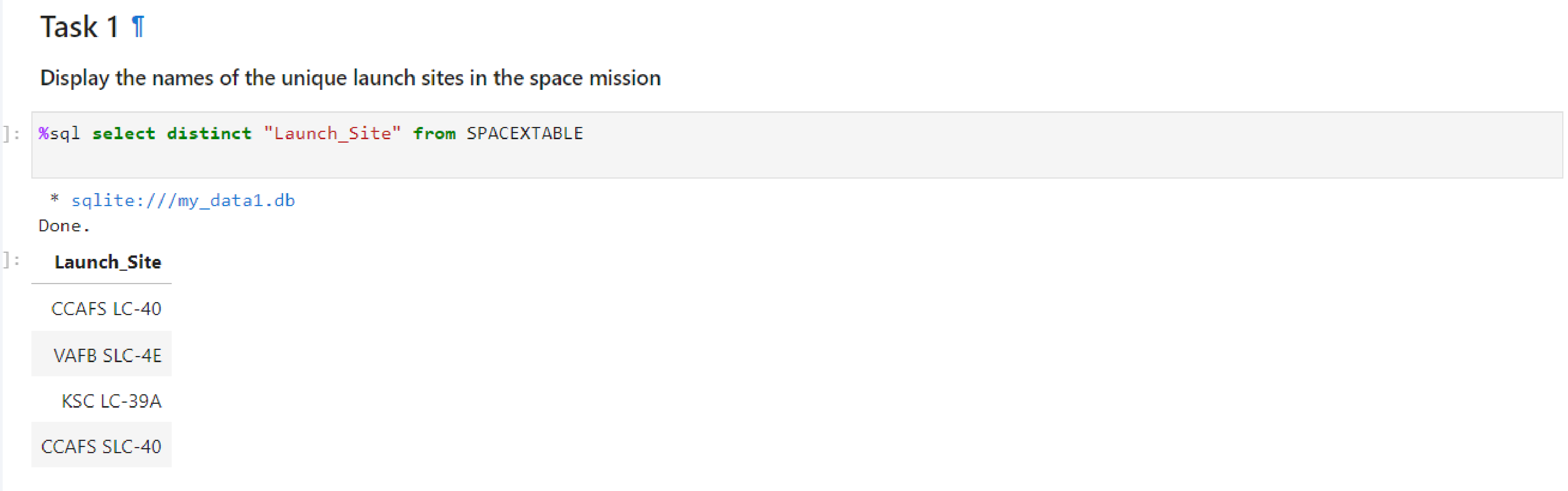
- We notice in this scatter plot that there as the launches increase the probability of landing increases, with an exception being the GTO with a haphazard pattern.
- The launches of LEO and VLEO have a better landing outcome as the flight number increases.
- We notice that there is a better chance of landing to those orbit types which have to remain synchronous with respect to another body(Earth, Sun).
- This plot can help us understand the nature of the more successful landing orbit type satellites.
- We also notice that there is a 100% landing rate of SSO orbit satellites

Payload vs. Orbit Type



- In this scatter plot we notice that there are a lot of launches with payload mass < 4000 kg in the ISS orbit and <8000 kg in GTO orbit type satellites.
- The average payload mass for all orbit types is <10000kg.
- We notice that as the payload mass increases, the landing rate increases in LEO satellites.
- Another point to note is that VLEO satellites usually carry very high payload masses.
- All these points suggest the nature of the payload and the orbit types can be planned according to these success metrics and the payload masses

All Launch Site Names



Task 1 ¶

Display the names of the unique launch sites in the space mission

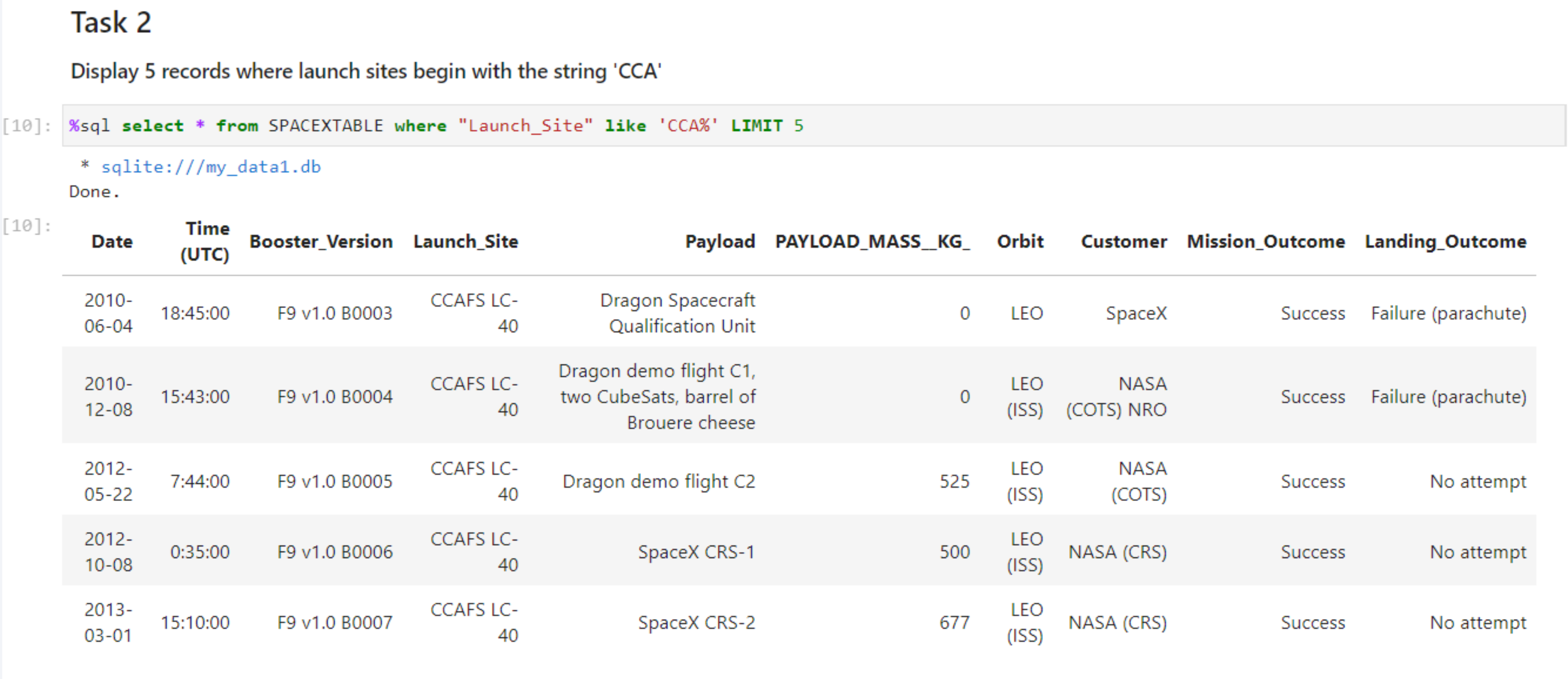
```
]: %sql select distinct "Launch_Site" from SPACEXTABLE
```

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

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

Here we can see that the distinct launch sites listed are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'



Task 2

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

```
[10]: %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' LIMIT 5
```

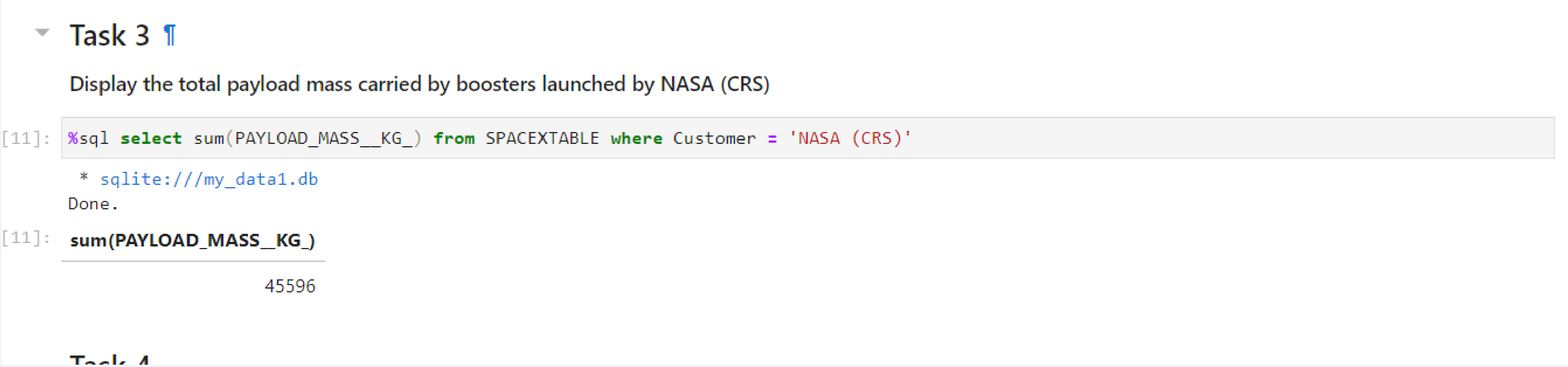
```
 * sqlite:///my_data1.db
Done.
```

[10]:

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

Here we can see that that 5 records from the launch sites located at Cape Canaveral.

Total Payload Mass



Task 3 ¶

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

```
[11]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

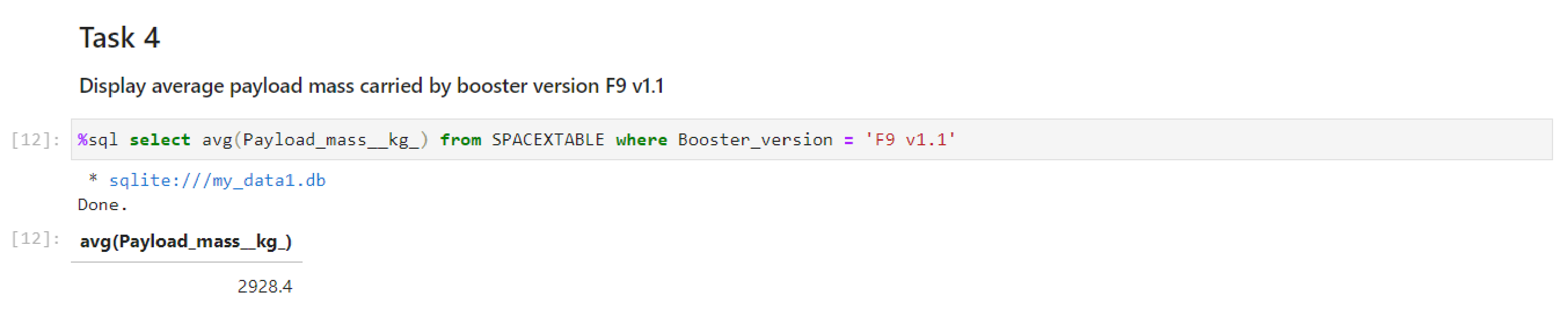
[11]:

sum(PAYLOAD_MASS__KG_)
45596

Task 4

Here we notice that the total payload mass from NASA(CRS) is 45596 kg.

Average Payload Mass by F9 v1.1



Task 4

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

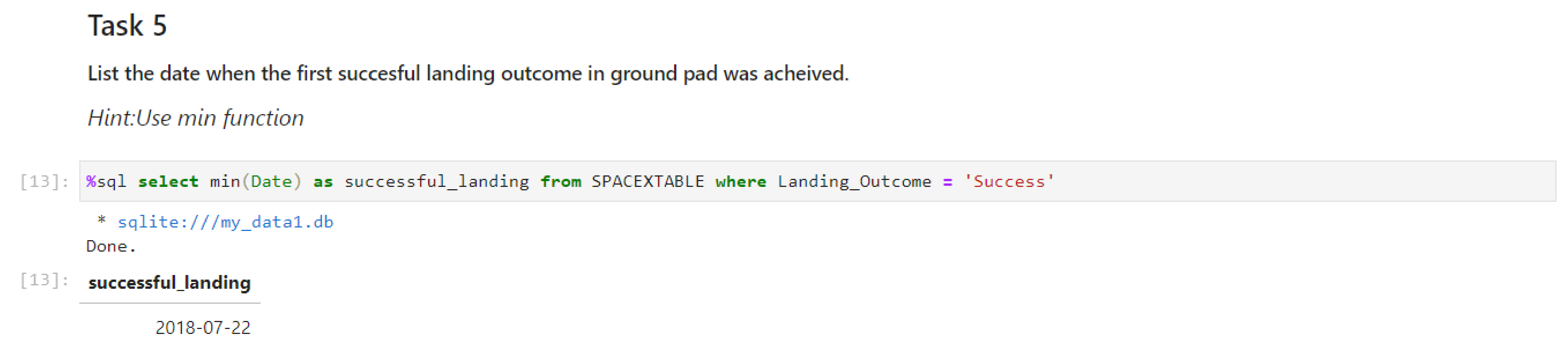
```
[12]: %sql select avg(Payload_mass__kg_) from SPACEXTABLE where Booster_version = 'F9 v1.1'

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

[12]:

avg(Payload_mass__kg_)
2928.4

Here we have calculated the average payload mass carried by booster version F9 v1.1 which turns out to be 2928.4 kg.

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

```
[13]: %sql select min(Date) as successful_landing from SPACEXTABLE where Landing_Outcome = 'Success'
```

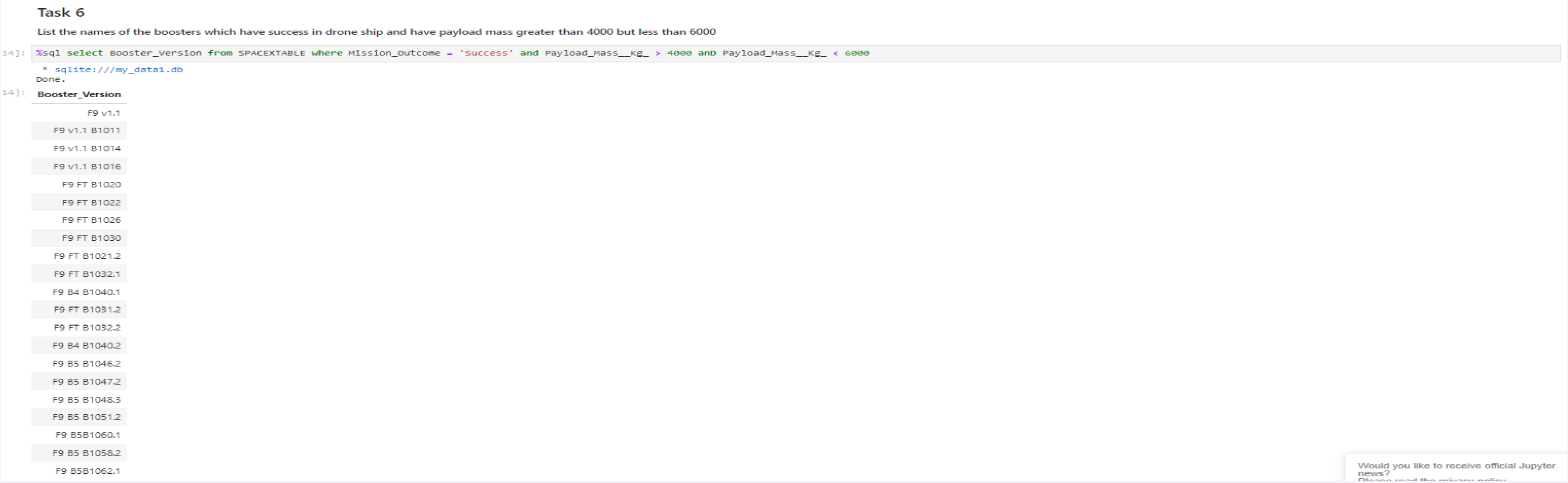
```
 * sqlite:///my_data1.db
Done.
```

[13]:

successful_landing
2018-07-22

Here we found out the first successful landing outcome and it turned out to be on the 22nd of July in the year 2018.

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

```
%sql select Booster_Version from SPACEXTABLE where Mission_Outcome = 'Success' and Payload_Mass__Kg_ > 4000 anD Payload_Mass__Kg_ < 6000
```

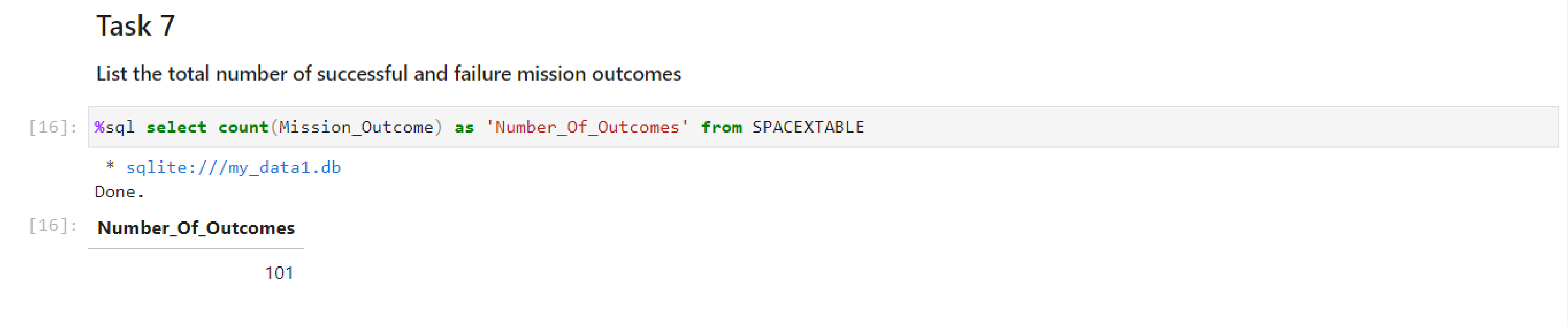
```
 * sqlite:///my_data1.db
Done.
```

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

Would you like to receive official Jupyter news?
Please read the privacy policy

Here we listed out all the successful drone ship missions with a payload mass between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes



Task 7

List the total number of successful and failure mission outcomes

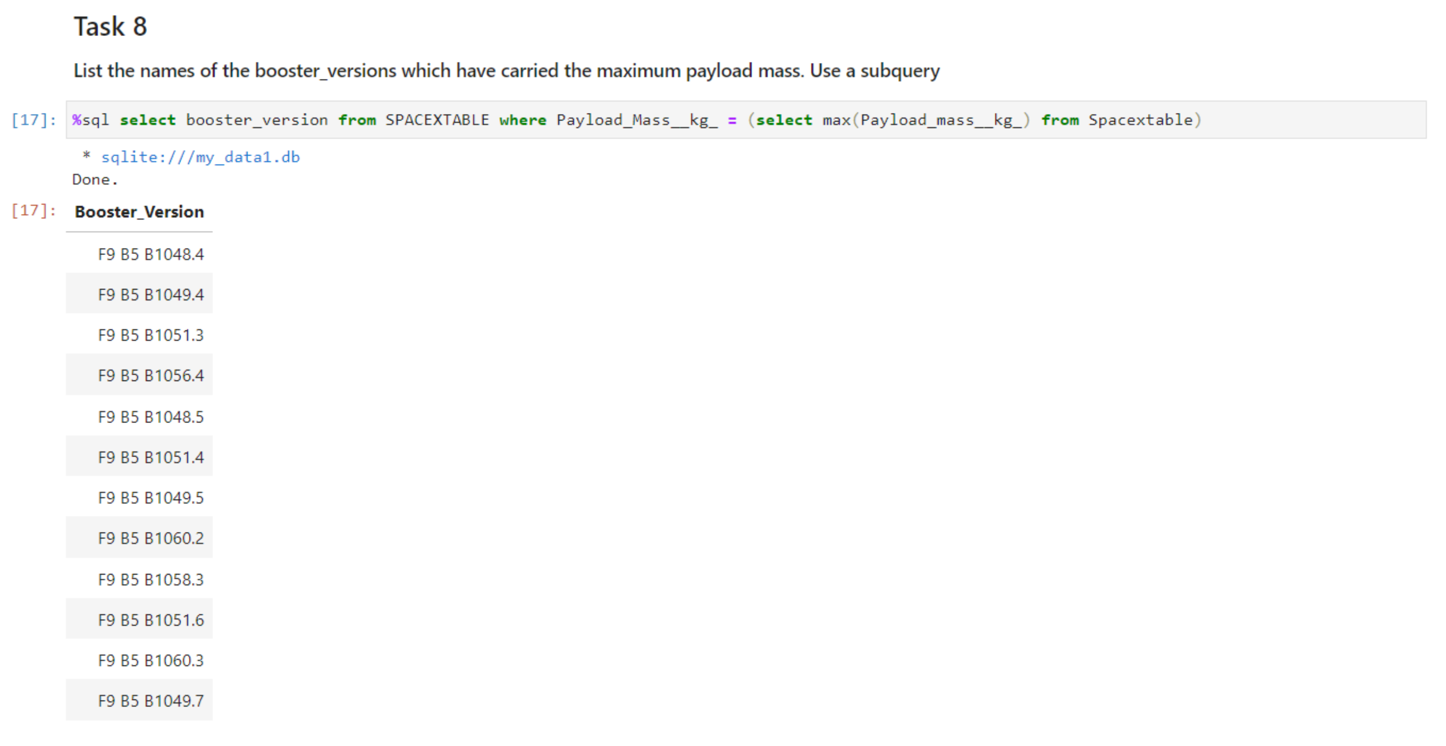
```
[16]: %sql select count(Mission_Outcome) as 'Number_Of_Outcomes' from SPACEXTABLE
```

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

```
[16]: Number_Of_Outcomes
                     101
```

Here we have displayed the total number of mission outcomes.

Boosters Carried Maximum Payload



Task 8

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

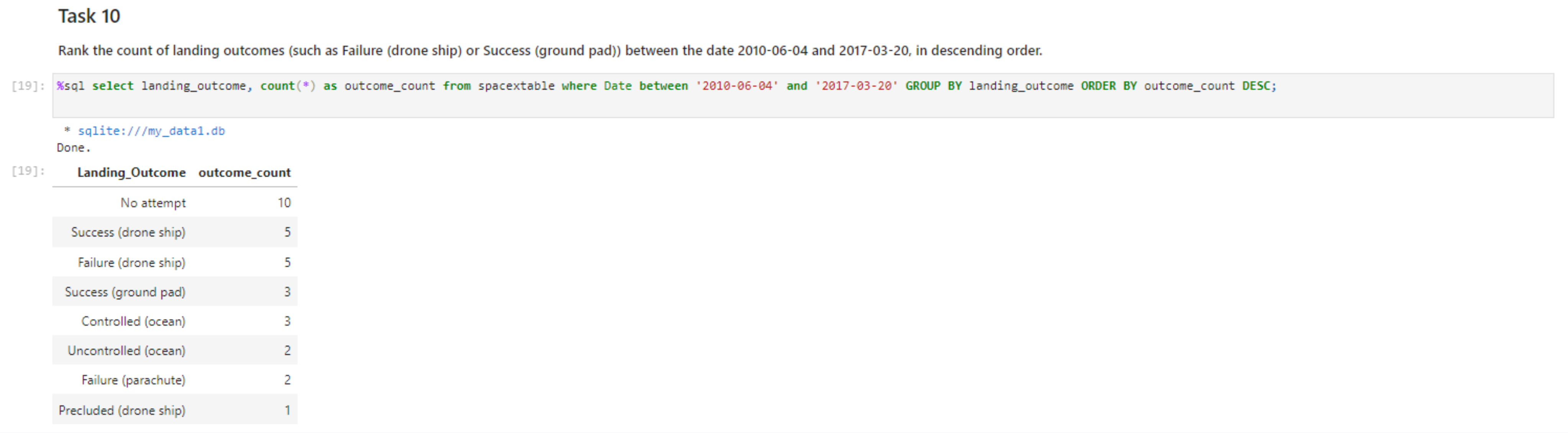
```
[17]: %sql select booster_version from SPACEXTABLE where Payload_Mass__kg_ = (select max(Payload_mass__kg_) from Spacextable)

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

[17]:

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

Here we listed the names of the booster versions which have carried the highest payload mass.

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.

```
[19]: %sql select landing_outcome, count(*) as outcome_count from spacextable where Date between '2010-06-04' and '2017-03-20' GROUP BY landing_outcome ORDER BY outcome_count DESC;
```

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

[19]:

Landing_Outcome	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

Here we made a list of the number of different types of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

Section 3

Launch Sites Proximities Analysis

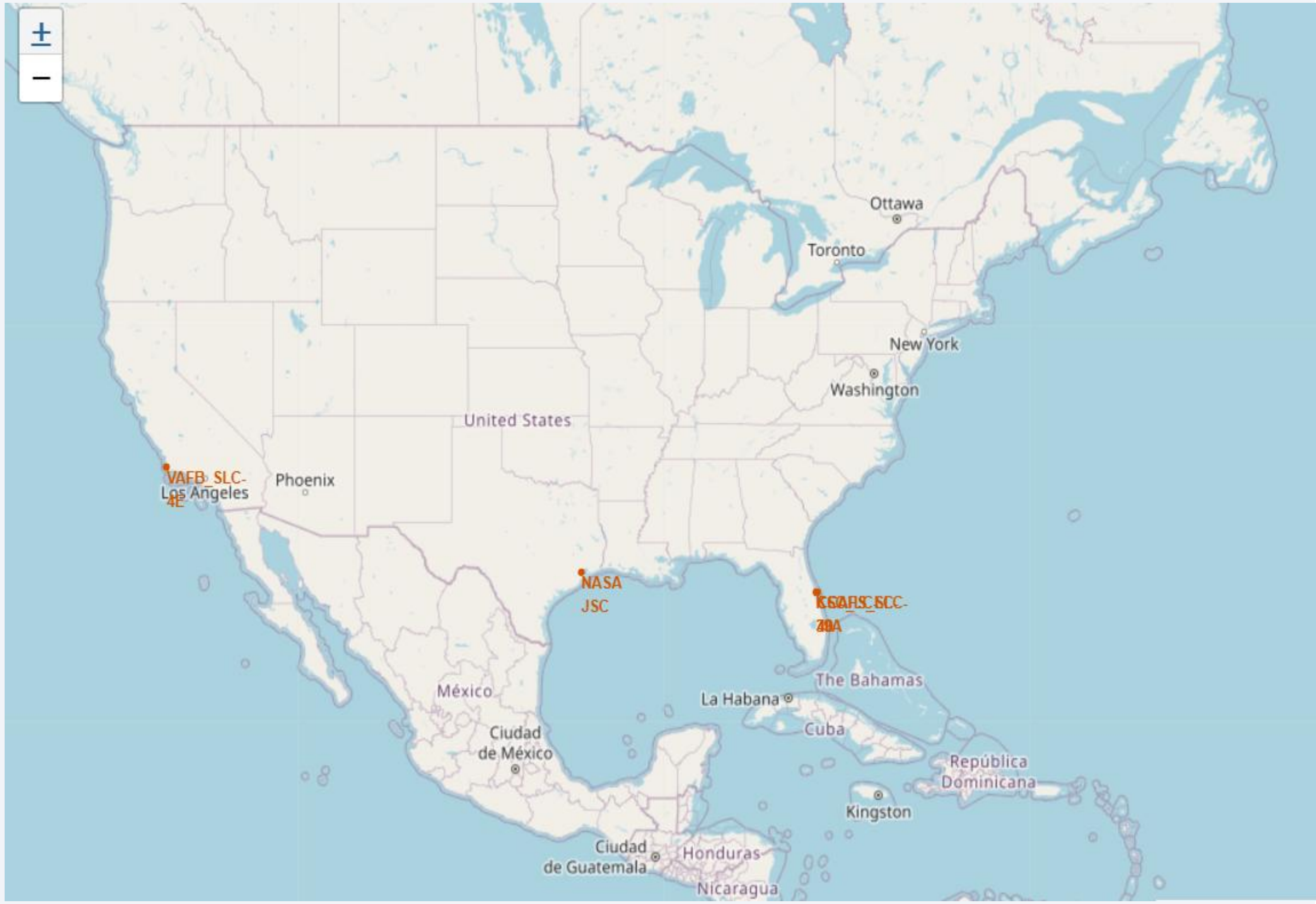
Key Locations and Insights

Some key insights are as follows:

1. Geographic Distribution: The launch sites are strategically positioned along the coasts and near key space infrastructure hubs. This allows for diverse launch trajectories and access to different orbits, including equatorial, polar, and geostationary.

2. Proximity to Water: Most sites are near large bodies of water, which is crucial for safety reasons, ensuring that any potential mishaps during launch do not endanger populated areas.

3. Coverage and Accessibility: The sites are spread across the country, allowing for logistical flexibility and redundancy. This distribution also facilitates a variety of mission types, from commercial satellite launches to deep space exploration missions.

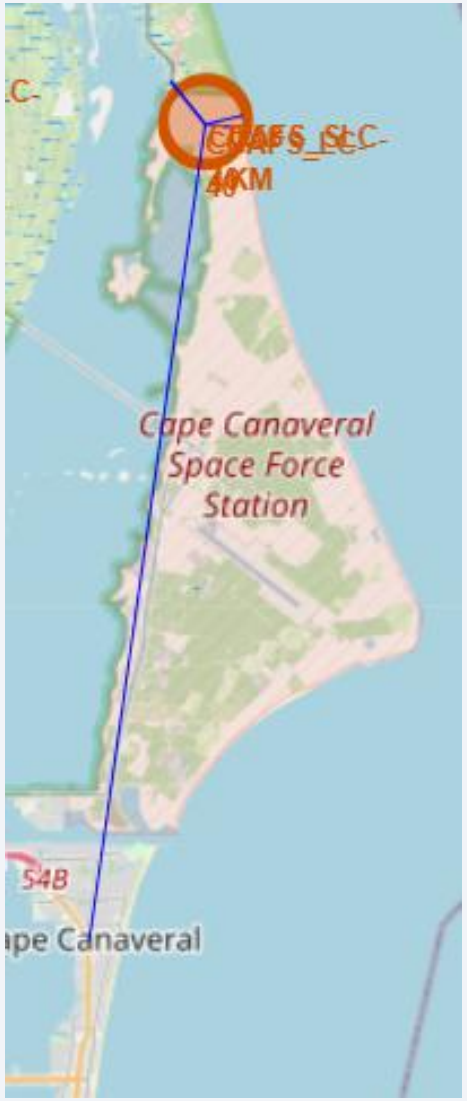


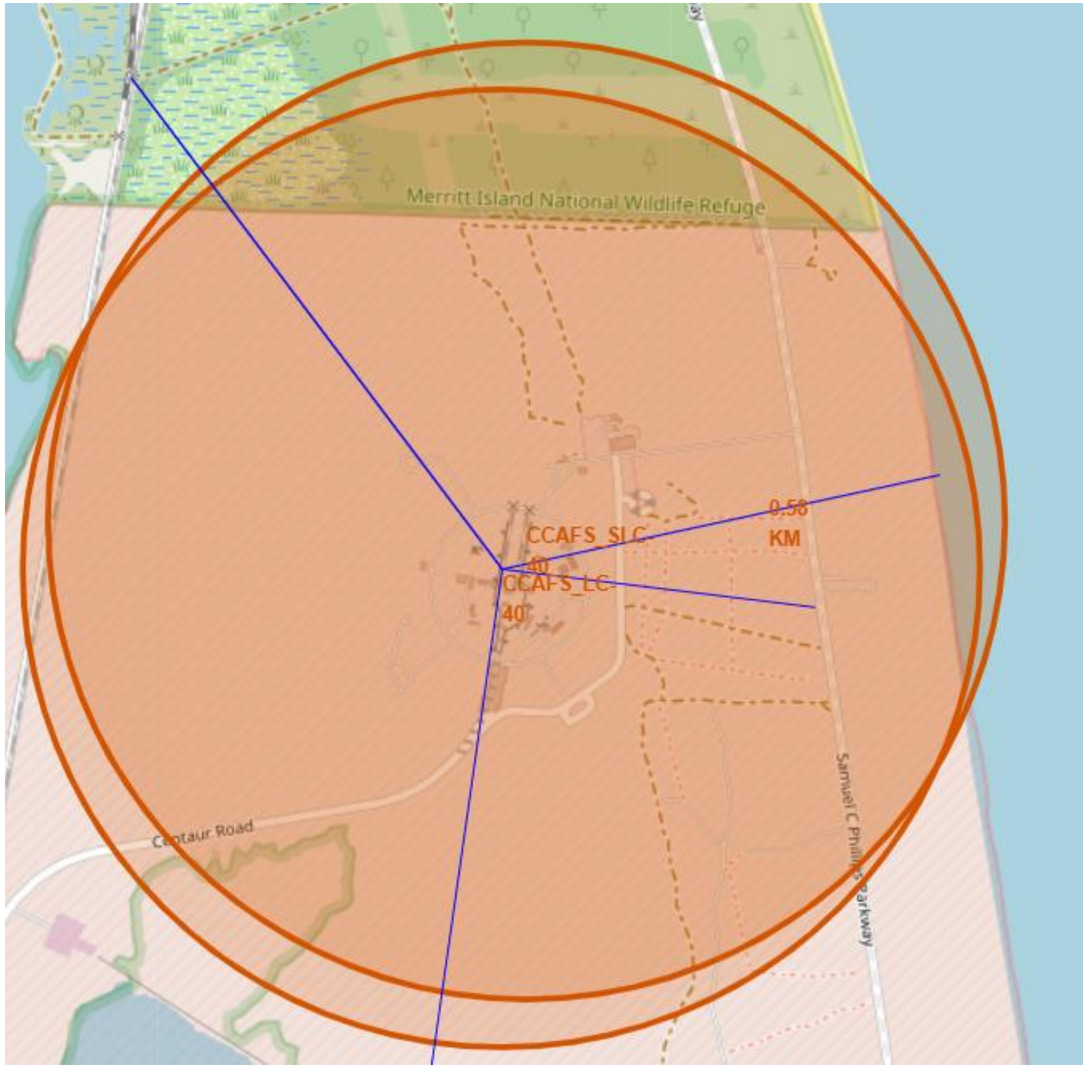
Strategic Proximity of Space Launch Complexes to Key Transportation Hubs

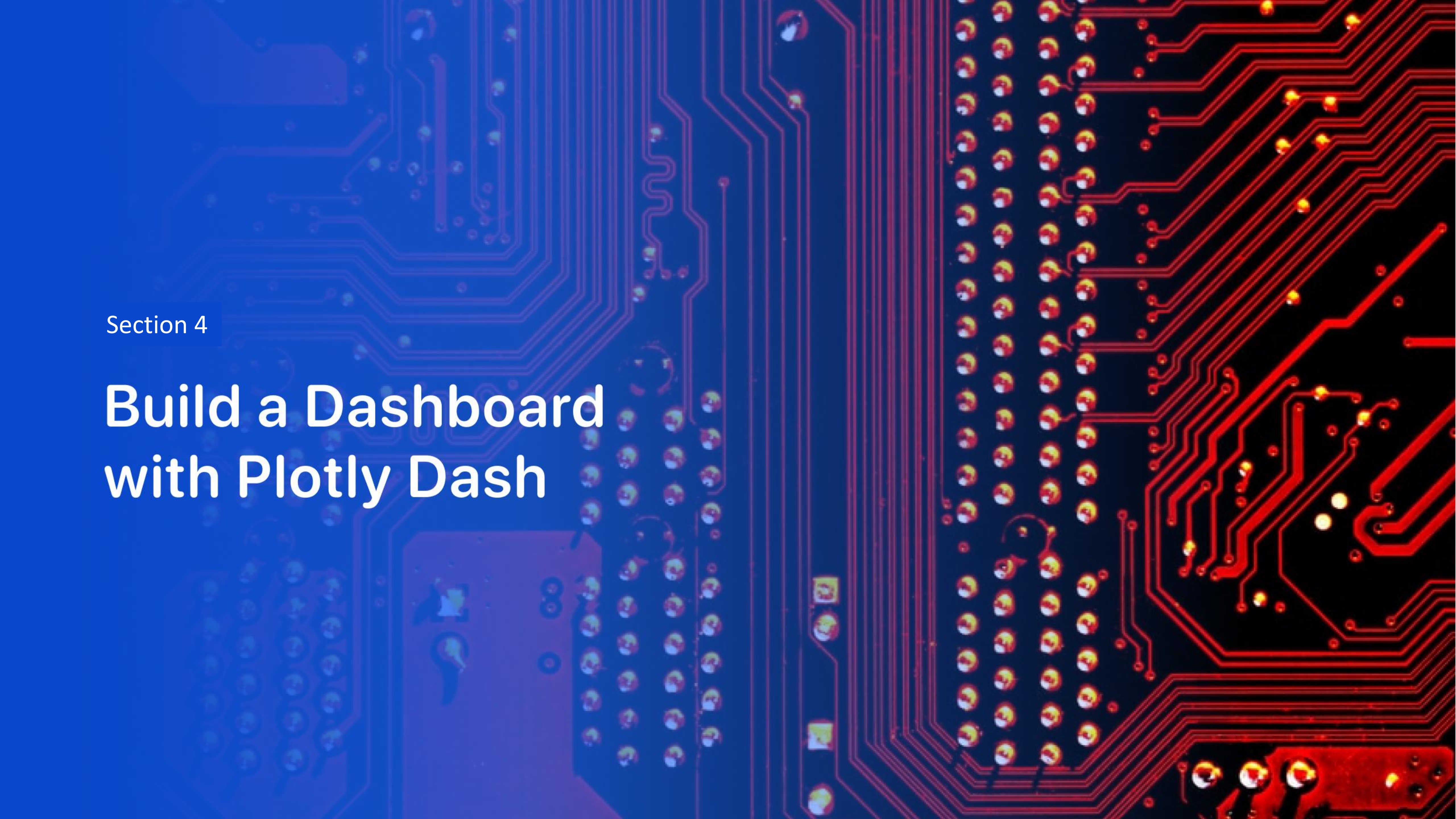
We can see in this screenshot the proximities of transportation hubs with respect to the space stations near Cape Canaveral.(Detailed image in next slide)

We see that the nearest city is Cape Canaveral and the nearest railway line is is the NASA railroad.

The nearest coastline is the Atlantic Ocean coastline of Florida and the nearest highway is the




Merritt Island National Wildlife Refuge
CCAFS_SLC-40
CCAFS_LC-40
0.58 KM
Centaur Road
Samuel C Phillips Parkway



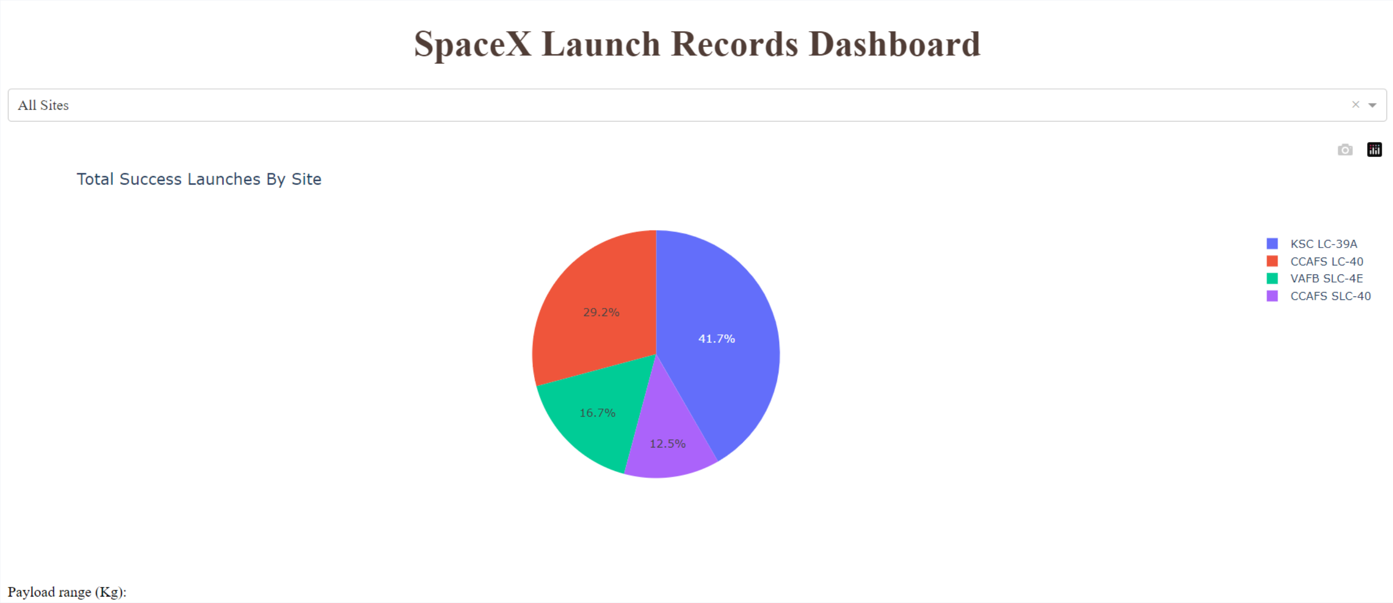
Section 4

Build a Dashboard with Plotly Dash

Success rates of each launch site

In this pie chart we can see that the largest number of successful launches have been from the KSC LC-39A center, followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40.

This pie chart gives us insights on choosing the best center to launch our rockets for the best success rate.



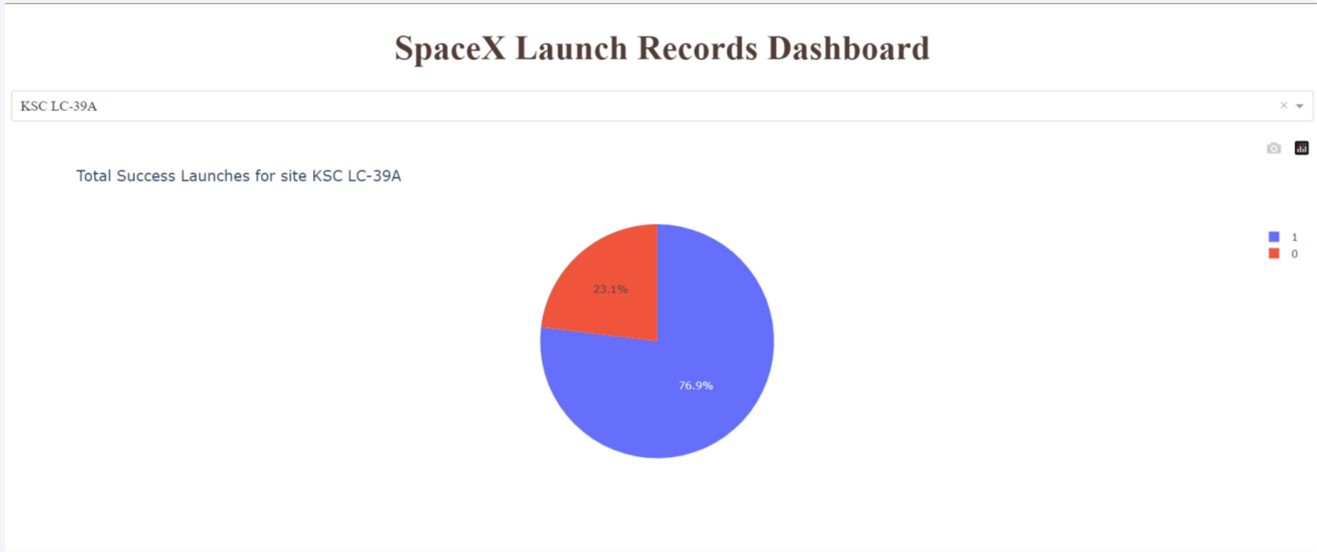
SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Success rate of launches at KSC LC-39A

- We can see that at KSC LC-39A, we have a high success rate of 76.9% which makes it the best launch site for successful launches.
- The chance of failing the launch sits at a relatively low 23.1%.
- To componsenate for this, we can launch single use rockets or satellites with higher success ratios



SpaceX Launch Records Dashboard

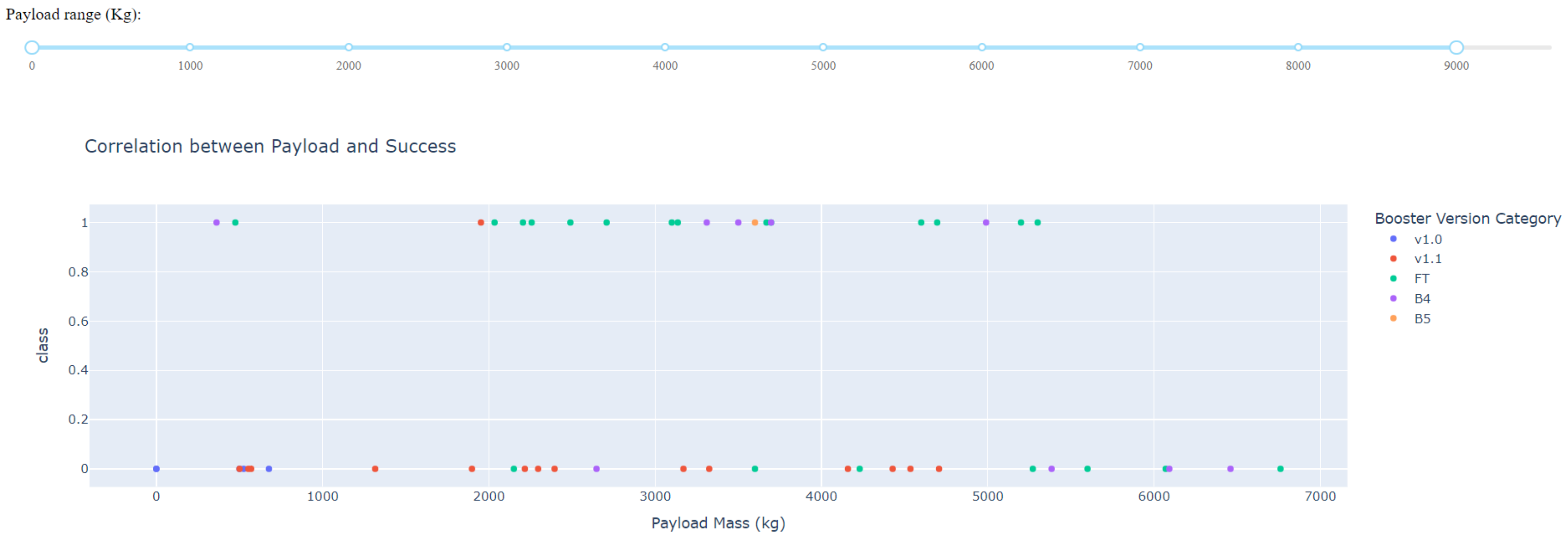
KSC LC-39A

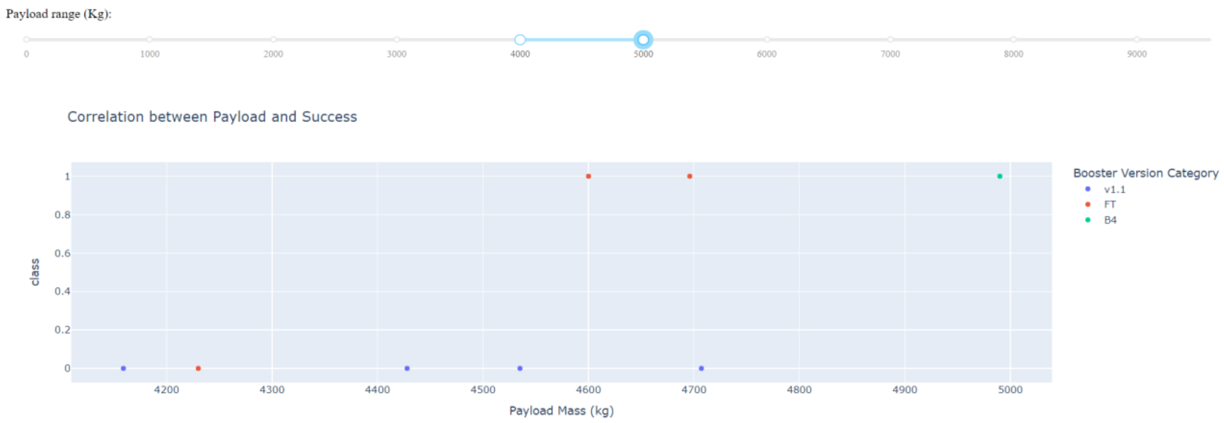
Total Success Launches for site KSC LC-39A

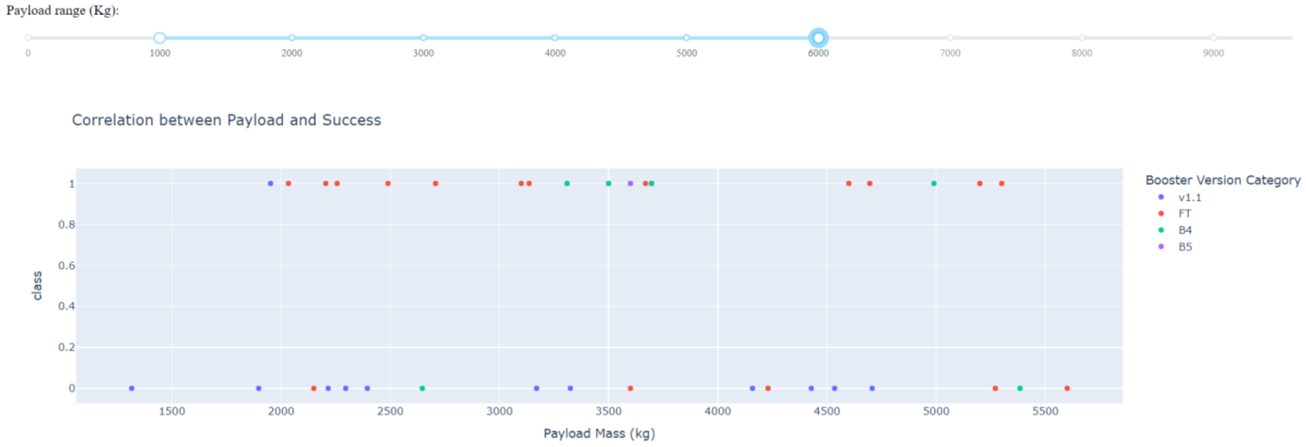
Payload vs launch outcome

- We can see in this scatter plot that there is a relatively high success rate for rockets carrying payloads between 2000 kg and 4000 kg.
- There is also a higher failure chance when the payload is between 0 kg and 1000 kg, there is also a high failure chance for most launches post 5000 kg of payload.
- In all these booster versions, the booster VT has a higher success chance(it has a lower success rate with heavy payloads) and the booster v1.1 has the least success chance. Only one launch has been done with the booster version B5 which turned out to be successful, so it can be experimented with.
- This data helps us make an informed choice about the payload mass and booster version to get the best success rate.

<Screenshots in next slide>


Payload range (Kg):
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Correlation between Payload and Success
class
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
Payload Mass (kg)
Booster Version Category
● v1.0
● v1.1
● FT
● B4
● B5


Payload range (Kg):
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Correlation between Payload and Success
class
1
0.8
0.6
0.4
0.2
0
4200
4300
4400
4500
4600
4700
4800
4900
5000
Payload Mass (kg)
Booster Version Category
● v1.1
● FT
● B4

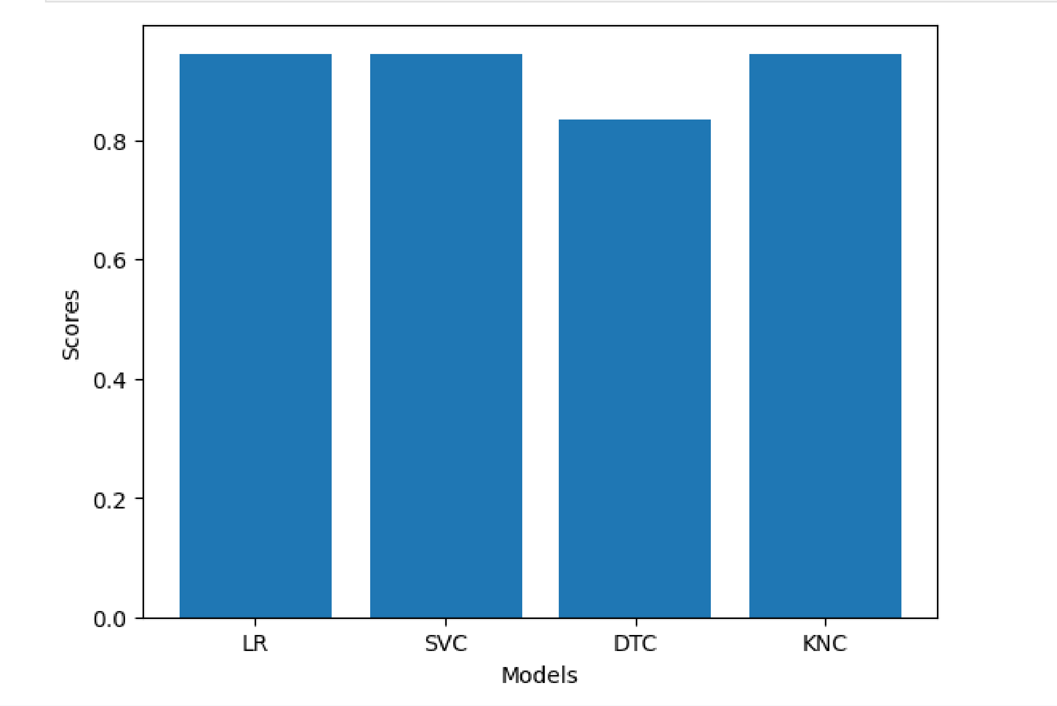

Payload range (Kg):
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Correlation between Payload and Success
class
1
0.8
0.6
0.4
0.2
0
1500
2000
2500
3000
3500
4000
4500
5000
5500
Payload Mass (kg)
Booster Version Category
● v1.1
● FT
● B4
● B5

Section 5

Predictive Analysis (Classification)

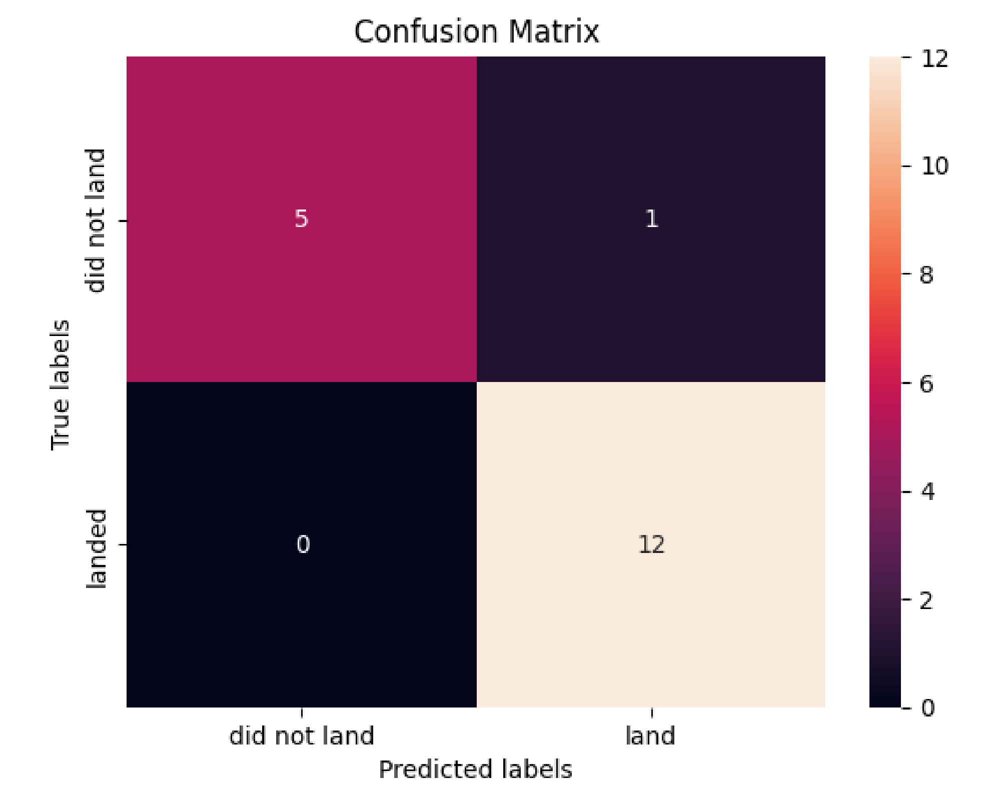
Classification Accuracy

- Here we see that logistic regression, support vector machines and K nearest classification models have the highest accuracy.
- Whereas the Decision tree classification model has the least accuracy of 0.8.



Confusion Matrix

- This is the confusion matrix of the logistic regression model.
- It shows 5 true positives, 1 false positive, 0 false negatives and 12 true negatives.
- It is the most accurate model along with SVC and K nearest classification models



Conclusions

- Our comprehensive analysis has led to several important conclusions:
- **Increased Landing Success**: The success rate of Falcon 9 first stage landings has significantly improved over the years, reflecting SpaceX's continuous advancements in reusable rocket technology.
- **Impact of Orbit Type and Payload Mass**: We discovered that certain orbit types and payload masses have higher success rates, suggesting that mission planning should consider these factors to maximize success probabilities.
- **Predictive Modeling Accuracy**: The predictive models we developed are highly accurate in forecasting landing success. This capability can help companies optimize their bidding strategies and enhance cost efficiency by predicting the likelihood of successful landings.

Conclusions

- **Strategic Launch Site Utilization**: Analysis of launch sites indicates that certain locations, such as KSC LC-39A, have higher success rates, making them preferable for future launches. Understanding the geographical distribution and logistical advantages of these sites can further improve mission success rates.
- **Optimization of Booster Versions**: Different booster versions exhibit varying success rates based on payload mass, highlighting the importance of selecting the appropriate booster for specific mission requirements.
- In conclusion, this project has provided valuable insights into the factors influencing the success of Falcon 9 landings, offering practical recommendations for optimizing space launch strategies and enhancing cost efficiency in the competitive landscape of space exploration.

Appendix

- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/jupyter-labs-webscraping%20(1).ipynb
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb
- https://github.com/Tejasvi-K/IBM-Applied-Data-science-capstone/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Thank you!