IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matías Cruz Ortega

December 30, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Visualization
  - Interactive map with Folium
  - Interactive Dashboards with Dash
  - Model prediction with Machine Learning
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive Analytics visuals
  - Predictive modeling results

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

- What factors determine if launch was successful?
- The interaction amongst various features that determine the success
- rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful
- landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- SpaceX rest API
- Web scraping

- Perform data wrangling
- One hot encoding data fields for machine learning and dropping irrelevant columns

- Perform exploratory data analysis (EDA) using visualization and SQL

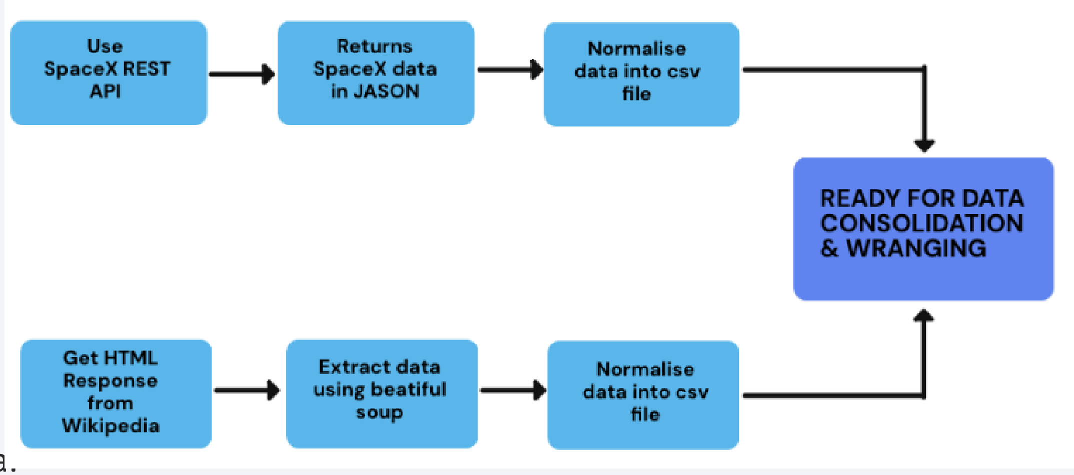
- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models

  - LR, KNN, SVM, DT models have been built and evaluated for the best classifier.

Data Collection

- **The following datasets was collected:**
  - Primary Data Source: SpaceX REST API (api.spacexdata.com/v4/)
    - Purpose: Programmatic retrieval of comprehensive and granular
    - Data Points Collected:
      - Specific rocket used (e.g., Falcon 9, Falcon Heavy)
      - Payload characteristics and nature
      - Precise launch site coordinates
      - Launch timing and date
      - Landing attempt outcome (success or failure)
- **Secondary Data Source: Wikipedia's Falcon 9 Launch Logs**
  - Purpose: Extraction of supplementary information to augment API data.
  - Method: Web scraping using the Beautiful Soup library.


Use SpaceX REST API
Returns SpaceX data in JASON
Normalise data into csv file
READY FOR DATA CONSOLIDATION & WRANGING
Get HTML Response from Wikipedia
Extract data using beatiful soup
Normalise data into csv file

- Overall Approach: Dual-source strategy to ensure a robust and comprehensive dataset

  for in-depth analysis of SpaceX launch history.

- Benefit: Leveraging both structured API data and publicly available information provides a multi-faceted understanding of each launch event.

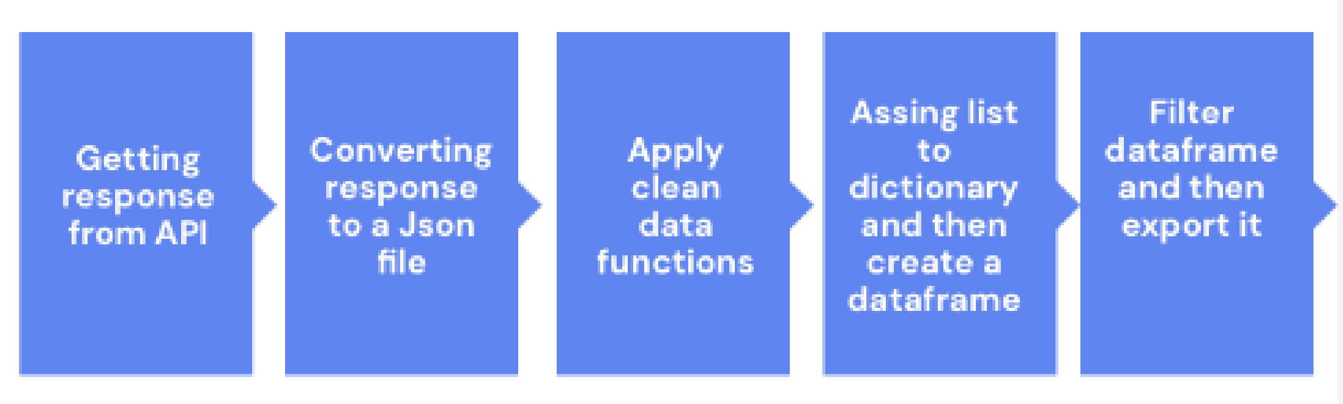
Data Collection – SpaceX API

- import request import pandas as pd spacex_url= "https://api.spacexdata.com/v4/launches/past"

response=request.get(spacex_url) data=pd.jason_normalize(response -json())

- GitHub URL of the completed SpaceX API calls notebook:

https://github.com/M02men311/data-science-capstone



Getting response from API

Converting response to a Json file

Apply clean data functions

Assing list to dictionary and then create a dataframe

Filter dataframe and then export it

Data Collection - Scraping

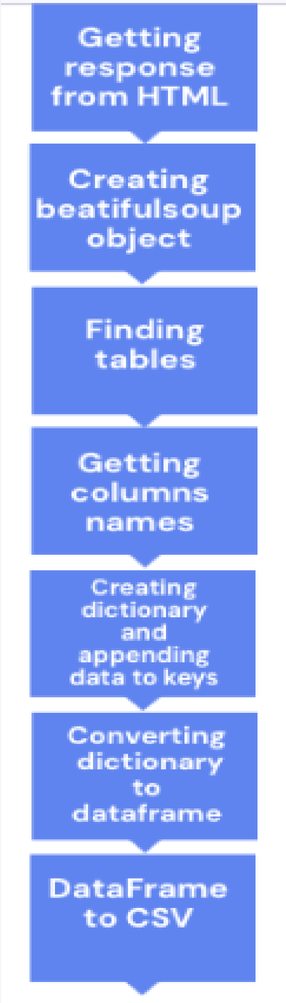
- import request from bs4 import BeatifulSoup url=

"https://en.wikipedia.org/wiki/list_of_Falcon_9_and_Falcon_Heavy_launches"

response=request.get(url)html_data=reponse.texts

Soup=BeatifulSoup(html_data)

- http://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



Getting response from HTML

Creating beatifulsoup object

Finding tables

Getting columns names

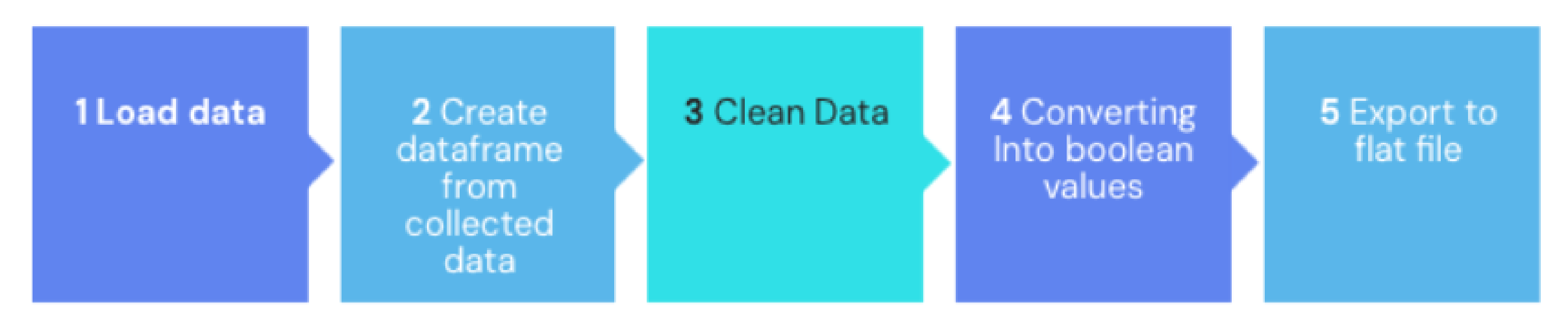
Creating dictionary and appending data to keys

Converting dictionary to dataframe

DataFrame to CSV

Data Wrangling

- Our data preprocessing pipeline involves a rigorous data quality check subsequent to data acquisition. This includes identification and handling of missing data, as well as verification of data type accuracy. To prepare the data for analysis, we perform the following cleansing steps:
  - **Missing Data Imputation:** Employing techniques such as mean imputation, or other appropriate methods based on the distribution of the variable, to address missing values and maintain data integrity.
  - **Data Type Harmonization:** Ensuring data types are consistent with the nature of the variable and the requirements of our analytical models.
  - **Numerical Encoding of Categorical Data:** Utilizing techniques like one-hot encoding to represent categorical variables numerically, enabling their inclusion in quantitative analyses.



1 Load data

2 Create dataframe from collected data

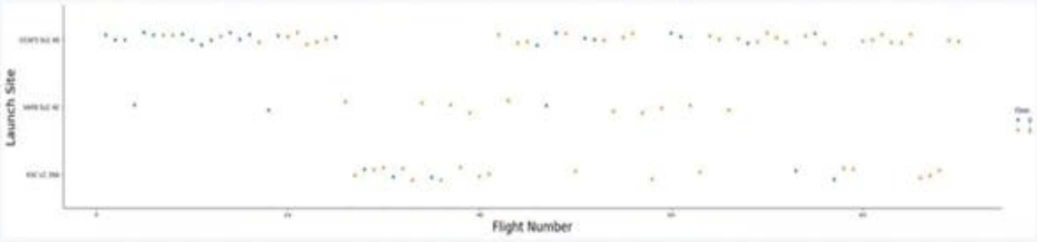
3 Clean Data

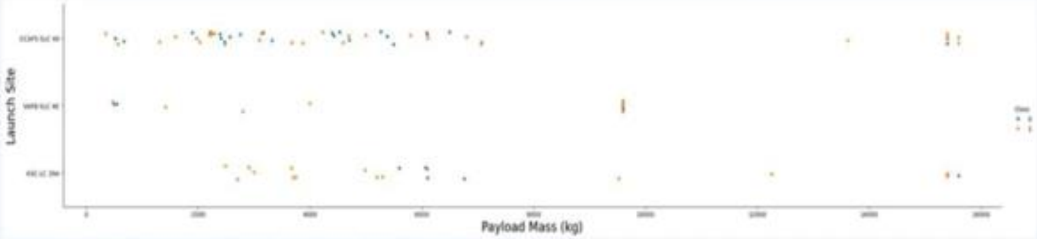
4 Converting Into boolean values

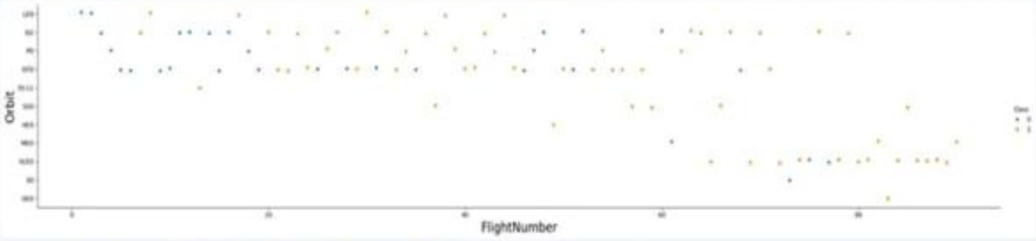
5 Export to flat file

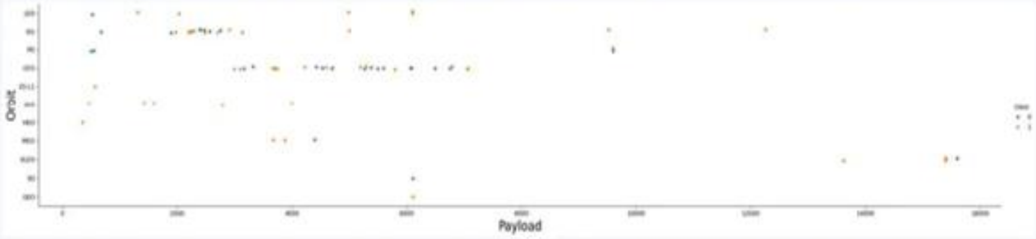
EDA with Data Visualization

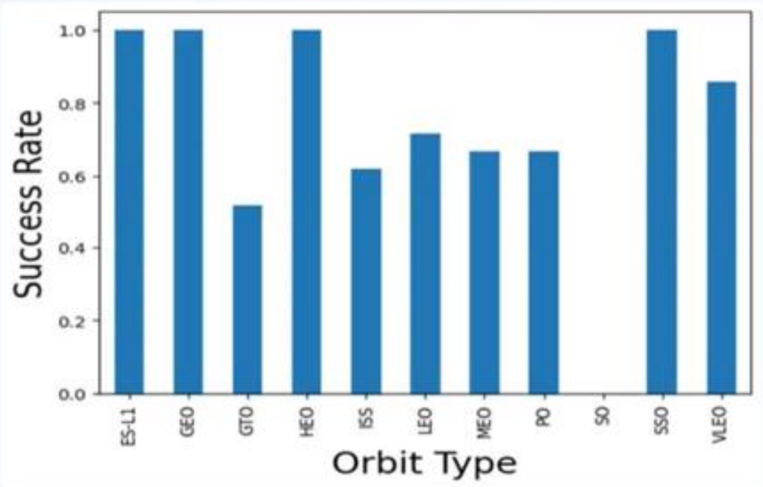
- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch.

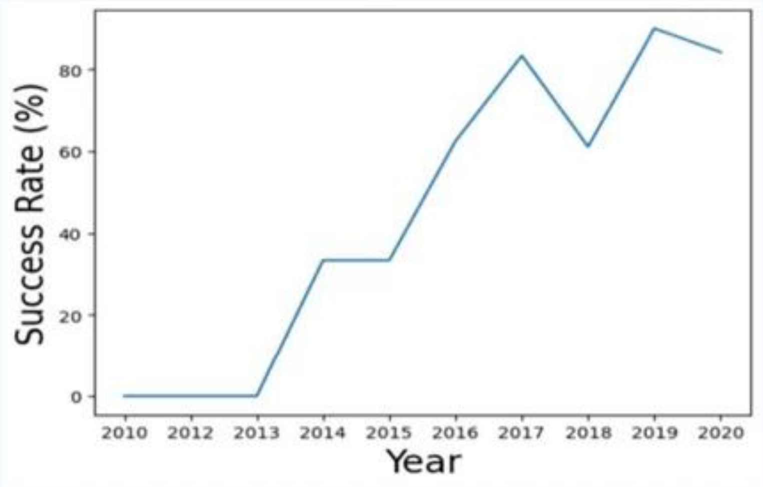




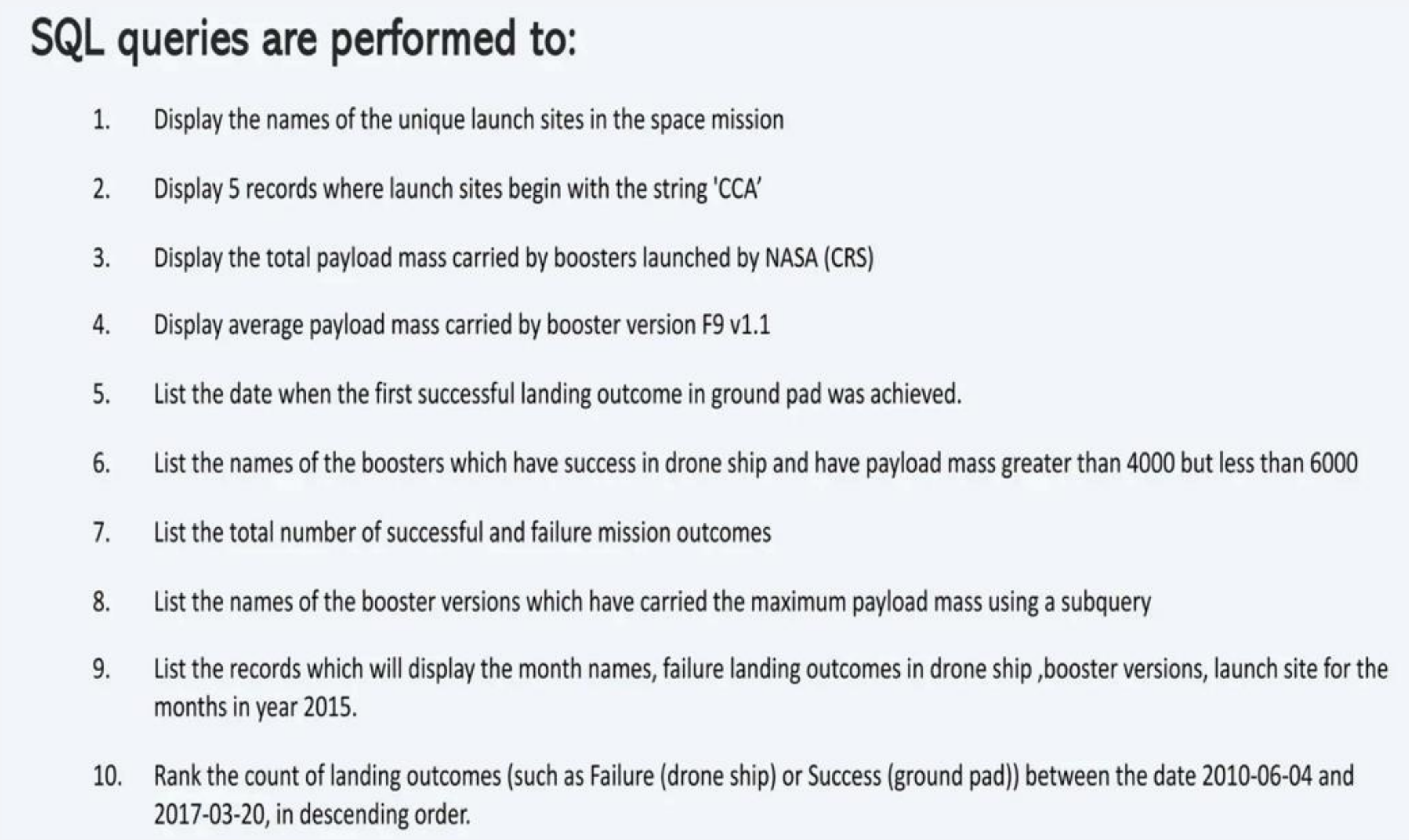








EDA with SQL

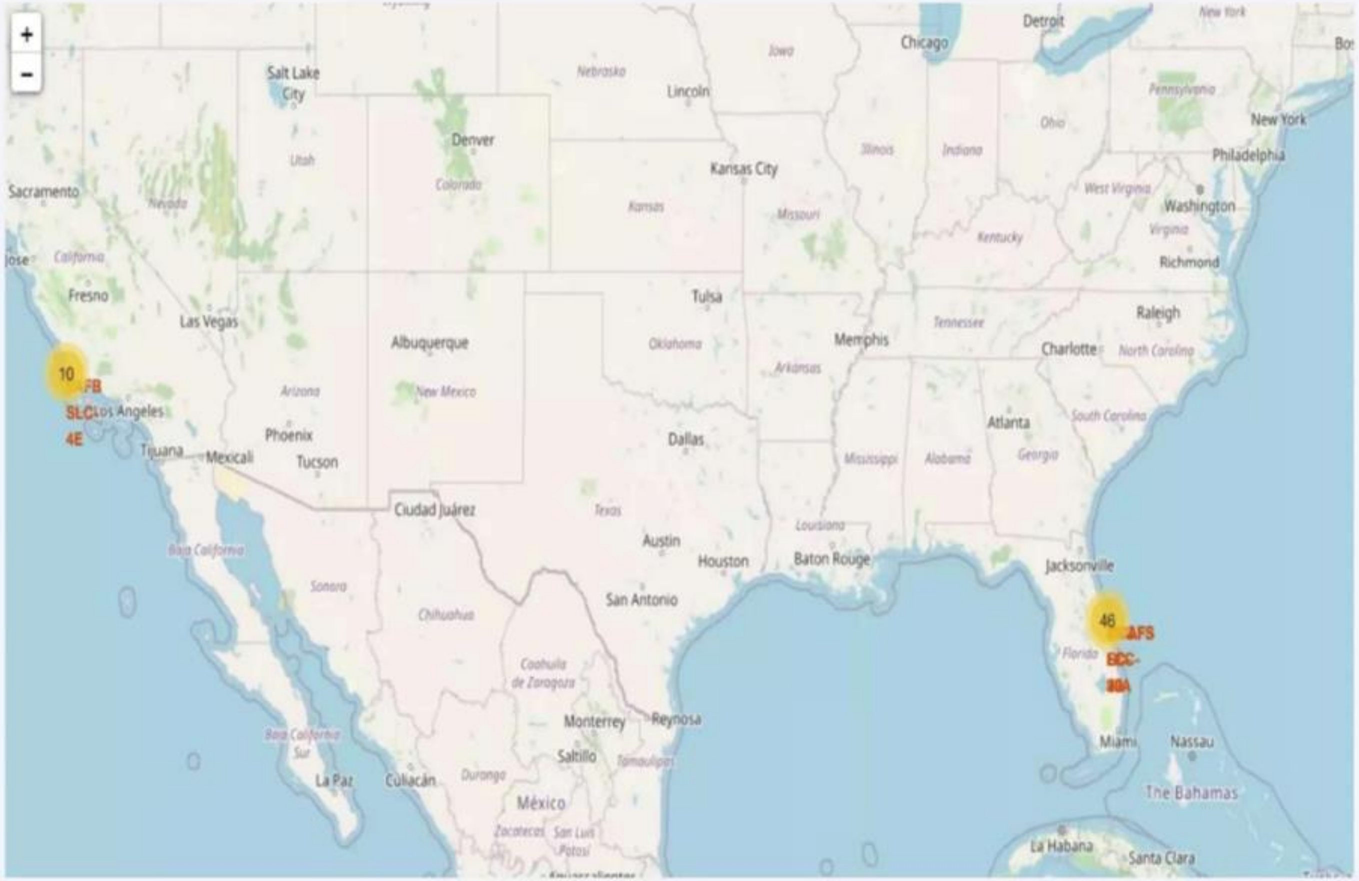


SQL queries are performed to:

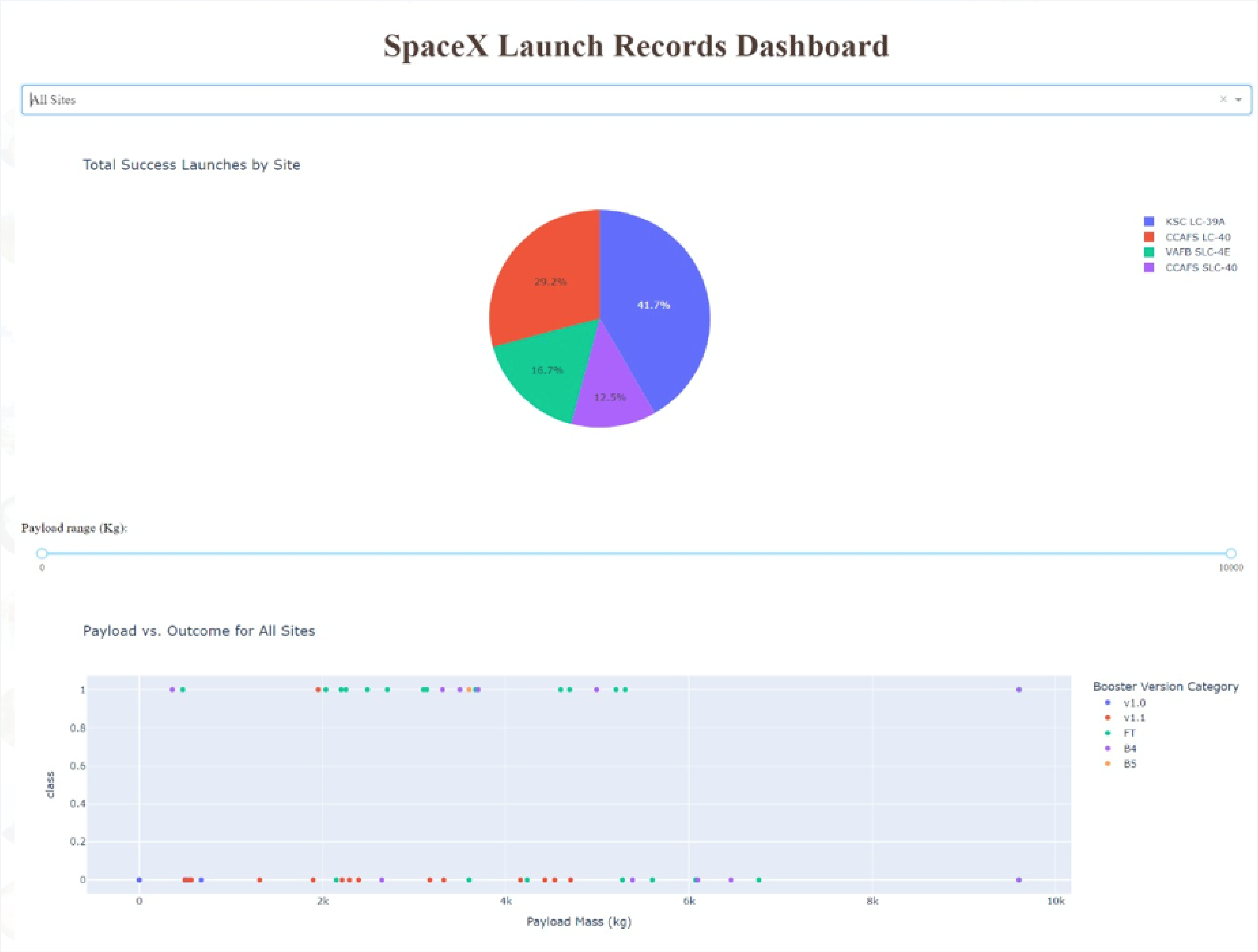
1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass using a subquery
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.
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.

Build an Interactive Map with Folium

- I marked potential sites to determine the best location for launch facility construction.



Build a Dashboard with Plotly Dash



SpaceX Launch Records Dashboard

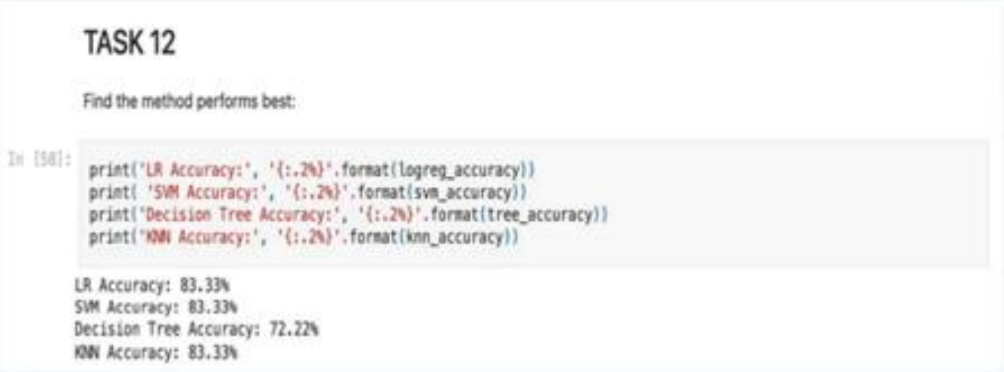
All Sites

Payload range (Kg):

- https://github.com/matiasinfgit/appliedDS_capstone/blob/main/spacex_dash_app_screenshot.png

Predictive Analysis (Classification)

- LR, SVM, Decision Tree and KNN objects are created and fit with GridSearchCV object to find the best parameters, then the models are trained on the training set.
- The accuracy of test data are calculated for each machine learning model. It is found that the methods performed best are LR, SVM, KNN where all 3 achieved the highest accuracy of 83.33%.

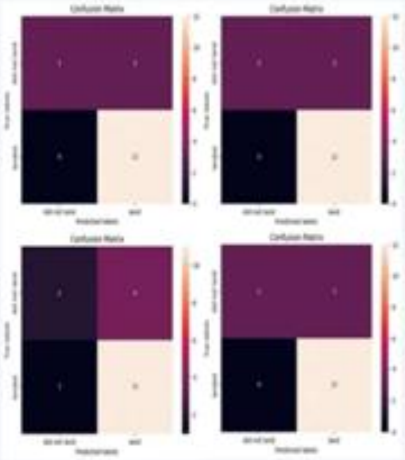


TASK 12

Find the method performs best:

```
In [58]:
print('LR Accuracy:', '{:.2%}'.format(logreg_accuracy))
print( 'SVM Accuracy:', '{:.2%}'.format(svm_accuracy))
print('Decision Tree Accuracy:', '{:.2%}'.format(tree_accuracy))
print('KNN Accuracy:', '{:.2%}'.format(knn_accuracy))
```

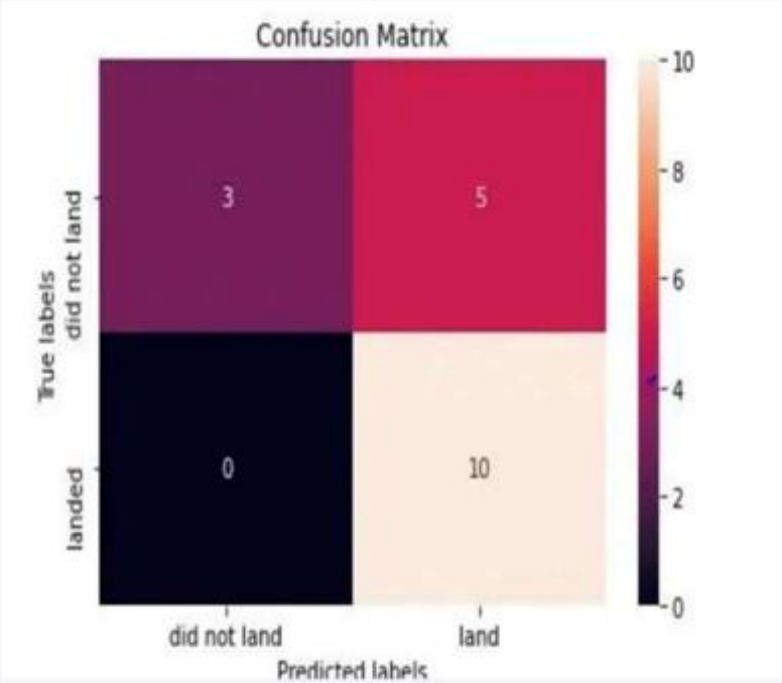
```
LR Accuracy: 83.33%
SVM Accuracy: 83.33%
Decision Tree Accuracy: 72.22%
KNN Accuracy: 83.33%
```



https://github.com/matiasinfgit/appliedDS_capstone/blob/main/7.SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

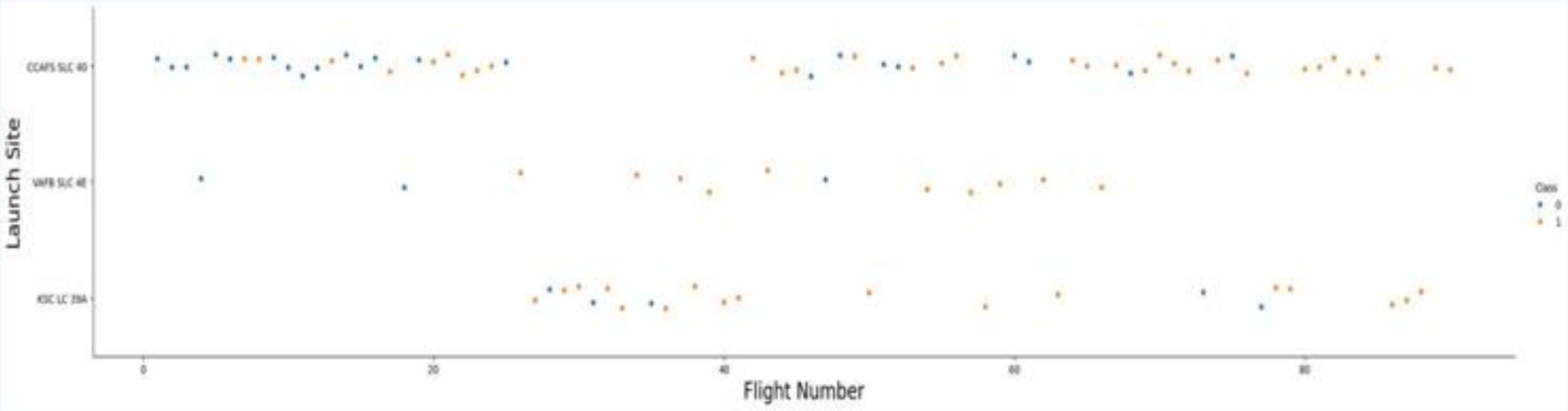
- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.



Section 2

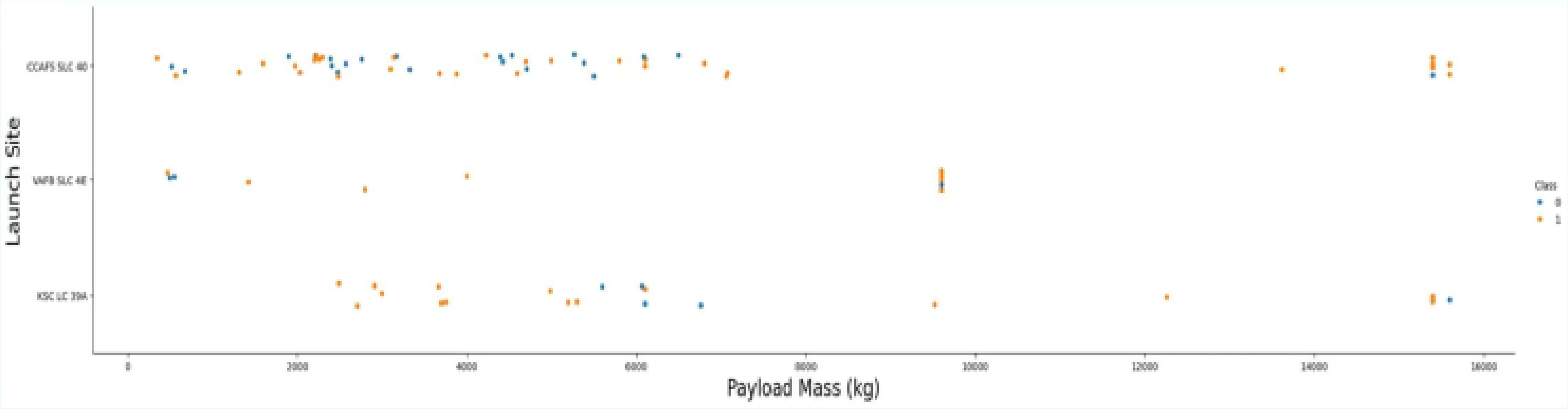
Insights drawn from EDA

Flight Number vs. Launch Site



- Total number of launches from launch site CCAFS SCL 40 are significantly higher than the other launch sites.

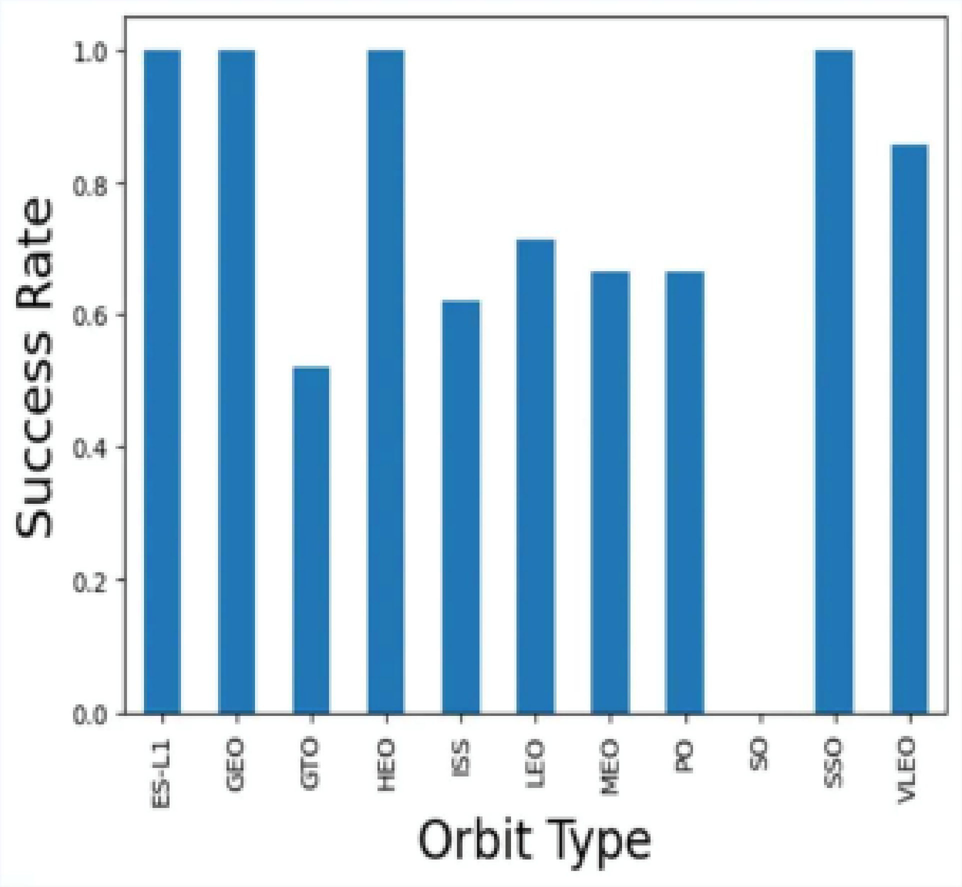
Payload vs. Launch Site



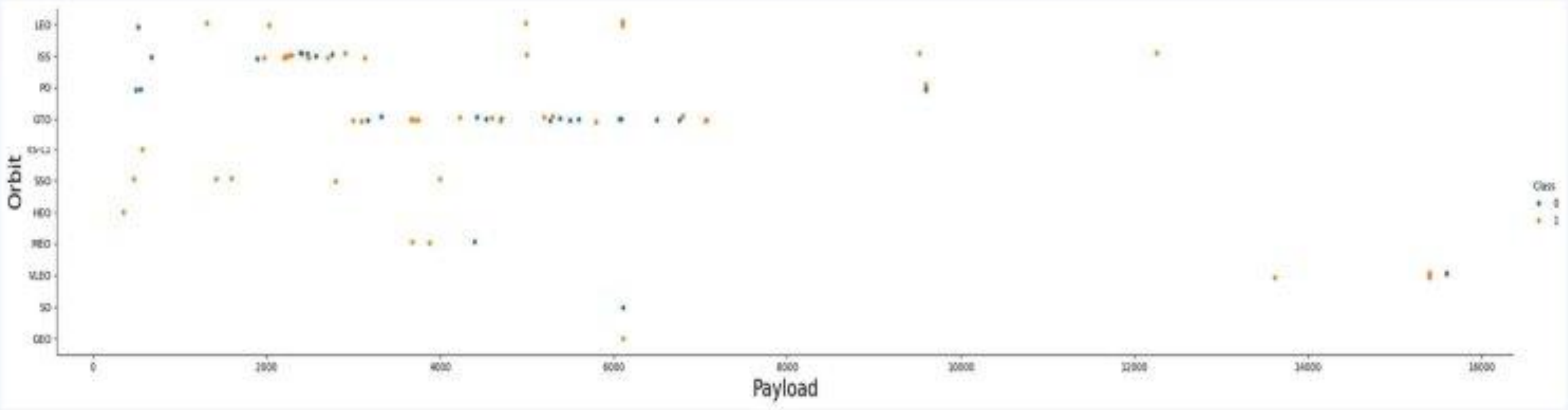
- Payloads with lower mass are have more launches compared to those with higher mass across all three launch sites.

Success Rate vs. Orbit Type

- Orbit types ES-L1, GEO, HEO, SSO have the highest success rate among all.

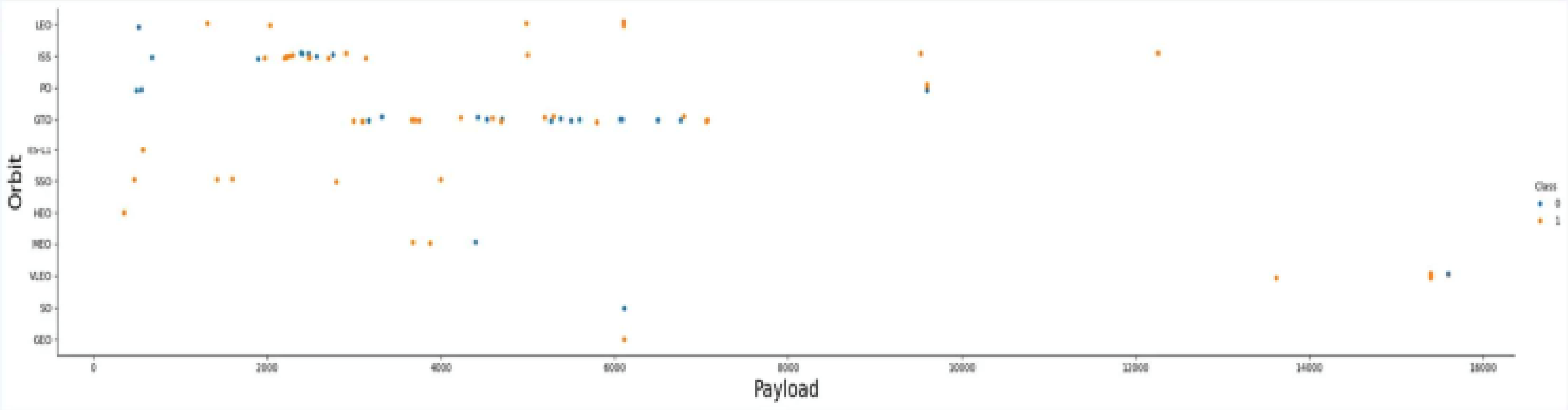


Flight Number vs. Orbit Type



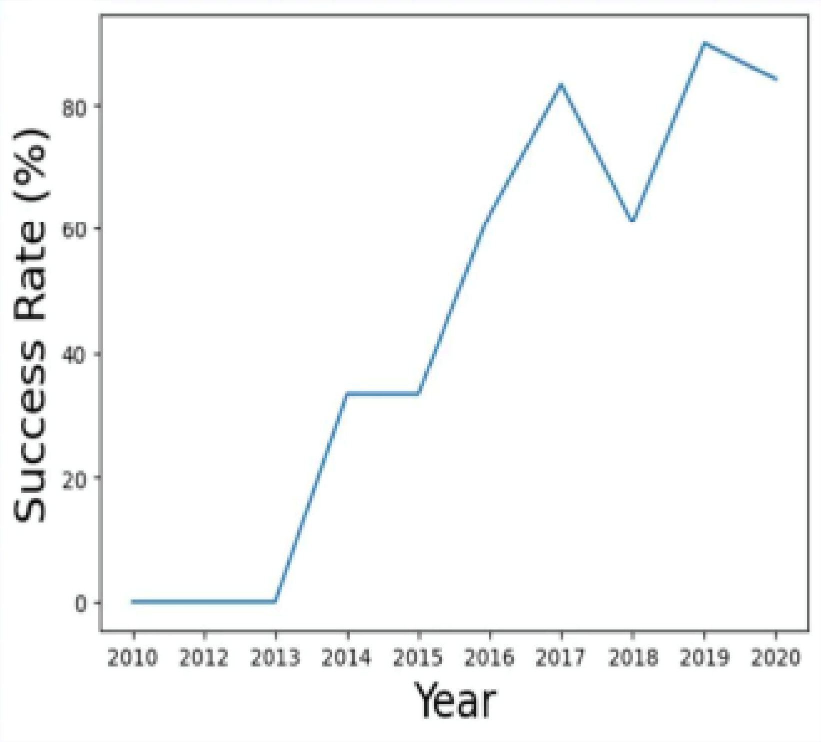
LEO, ISS, PO, GTO orbits have the most launches in the earlier years, but it slowly shifted to VLEO orbits in the later years.

Payload vs. Orbit Type



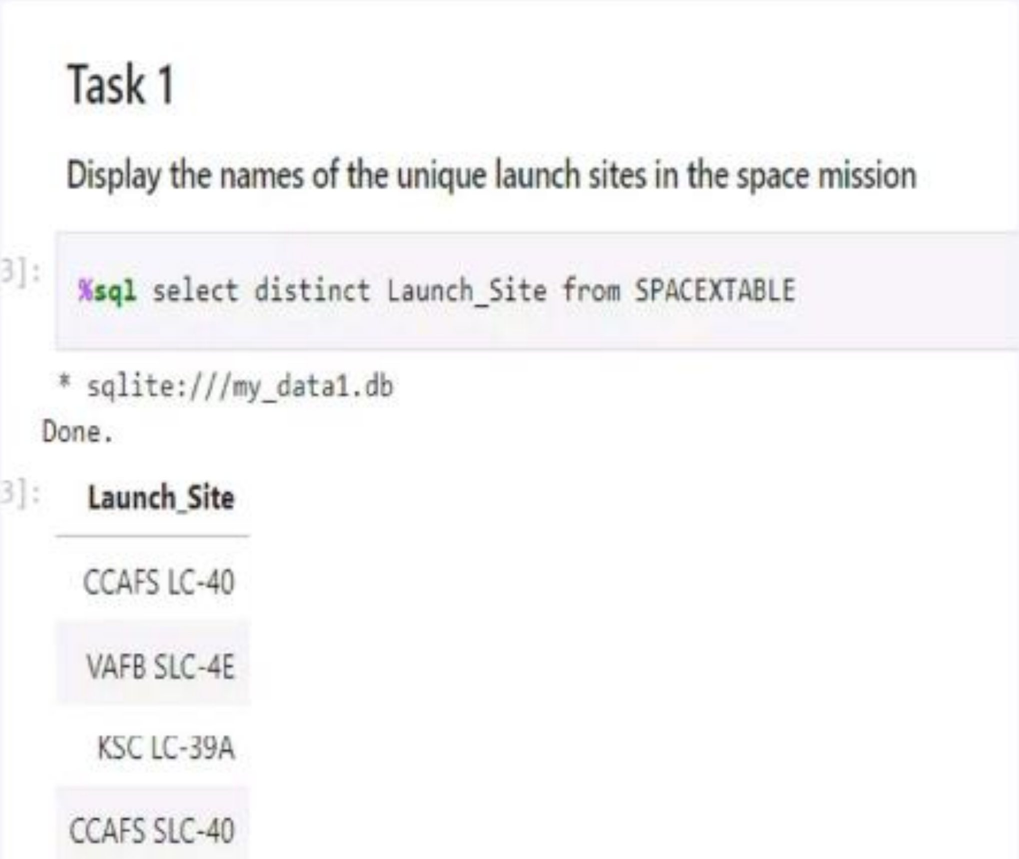
- Heavy payloads tend to have a higher successful landing rates for PO, LEO, and ISS orbits, but for GTO orbit, success is less predictable with an almost equal mix of success and failures.

Launch Success Yearly Trend



- The success rate of launches have been increasing since 2013 till 2020, possibly due to technology advancement and experience.

All Launch Site Names



Task 1

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

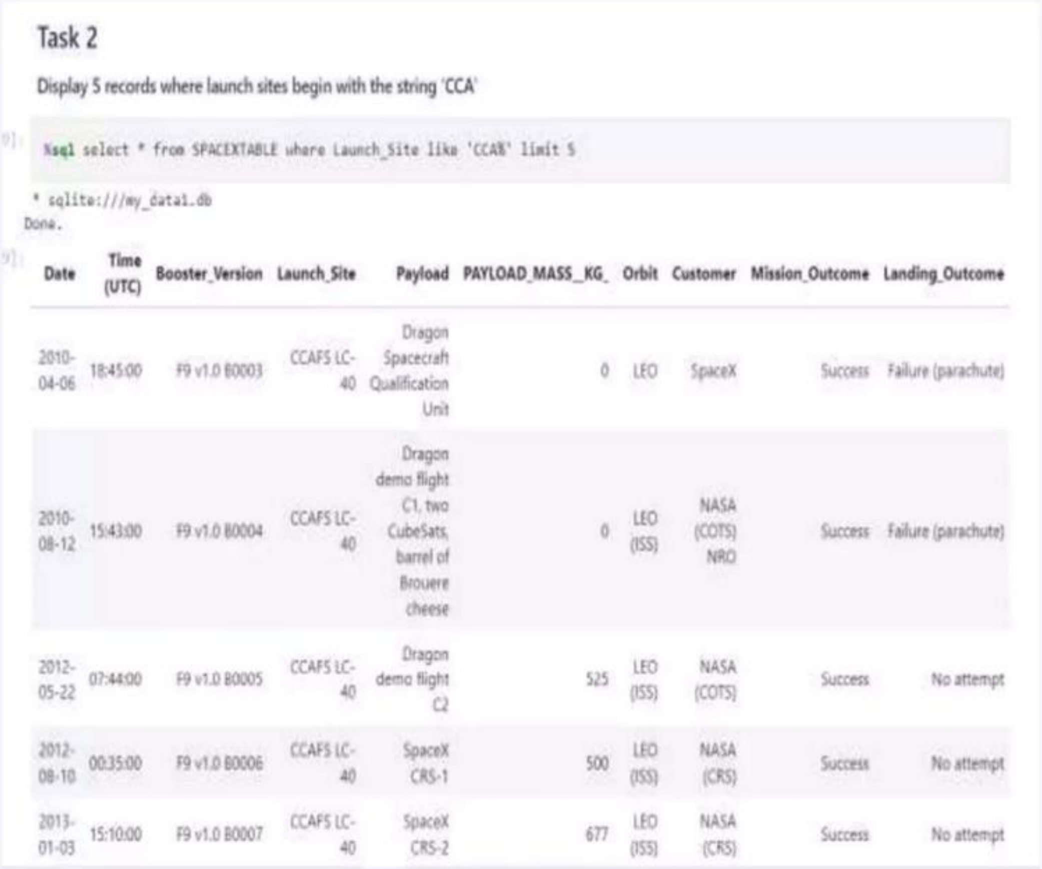
```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

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

- I used the keyword distinct to show only unique launch sites from the SpaceX data.

Launch Site Names Begin with 'CCA'



Task 2

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

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

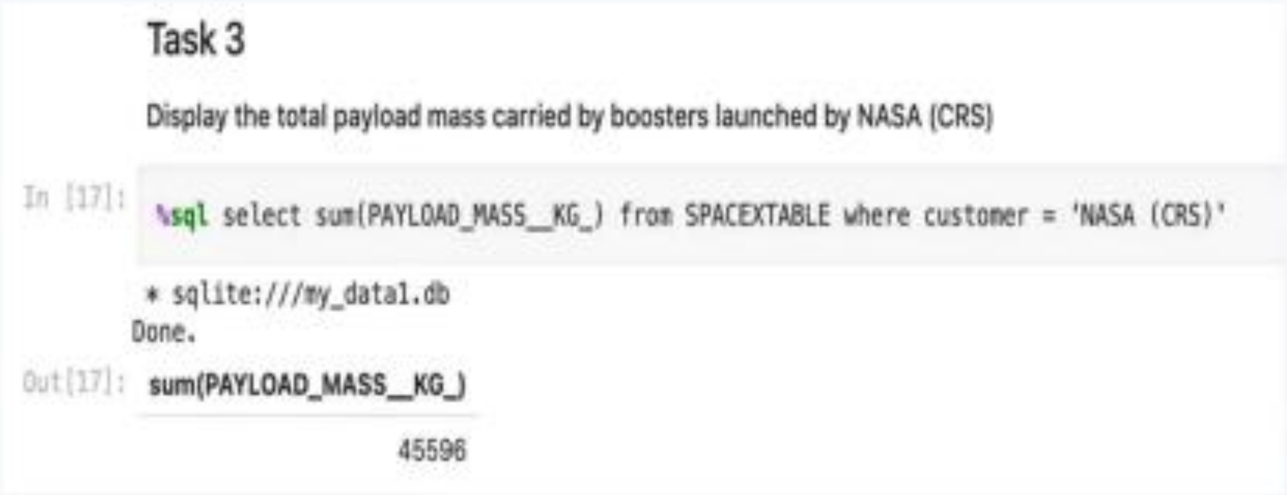
```
* sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- This query display 5 records where the launch site begin with "CCA"

Total Payload Mass

- Performed an SQL query to obtain the total payload mass carried by boosters launched by NASA (CRS)



Task 3

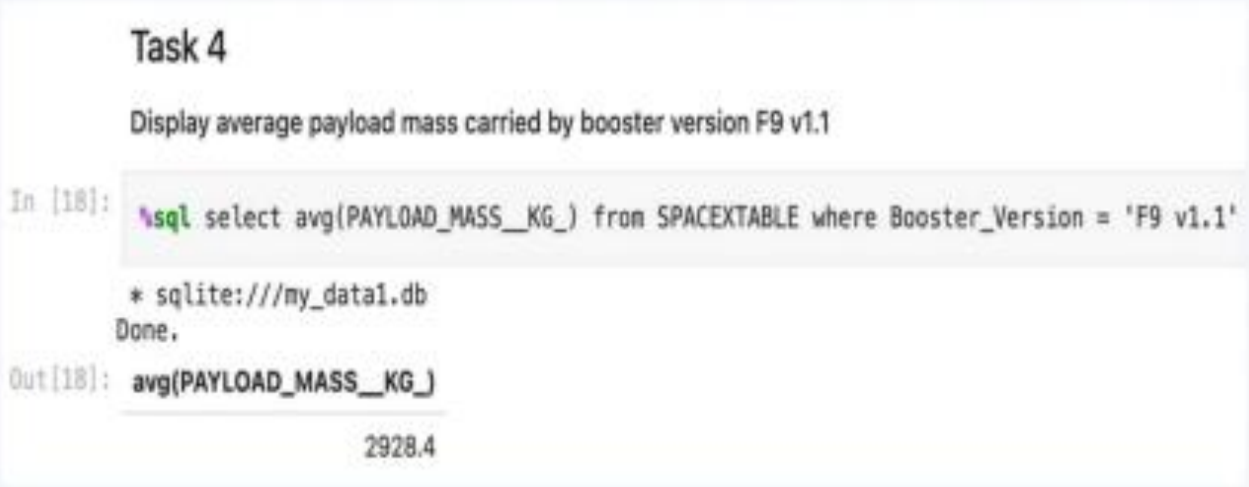
Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [17]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where customer = 'NASA (CRS)'
```

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

```
Out[17]: sum(PAYLOAD_MASS__KG_)
                          45596
```

Average Payload Mass by F9 v1.1



Task 4

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

```
In [18]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

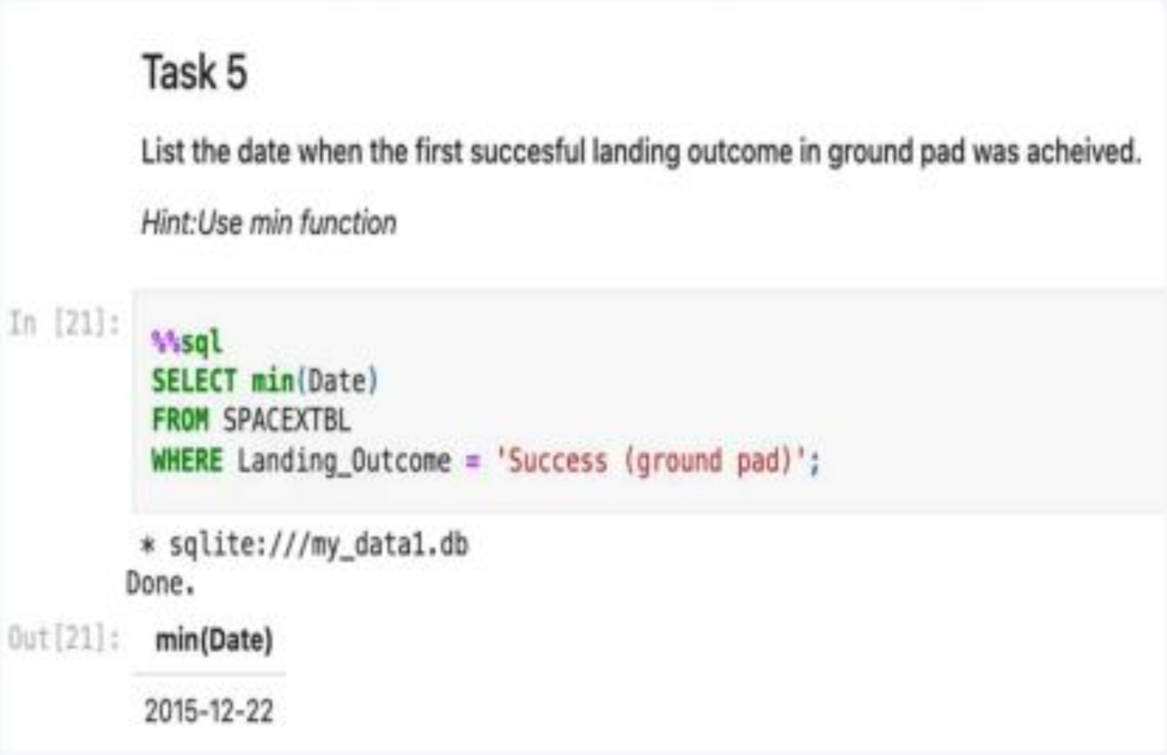
```
 * sqlite:///my_data1.db
Done.
```

Out[18]:

avg(PAYLOAD_MASS__KG_)
2928.4

- Performed an SQL query to calculate the average payload mass carried by booster version F9 v1.1

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

```
In [21]:  %%sql
          SELECT min(Date)
          FROM SPACEXTBL
          WHERE Landing_Outcome = 'Success (ground pad)';
```

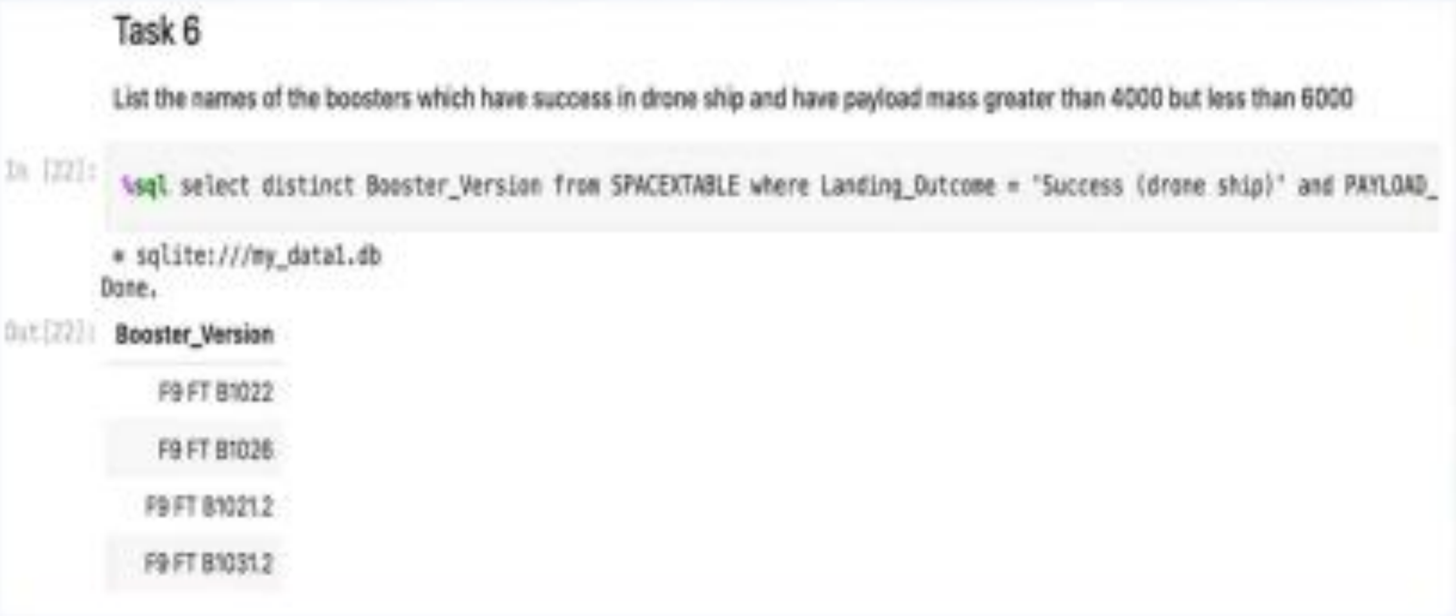
```
 * sqlite:///my_data1.db
Done.
```

```
Out[21]:  min(Date)
          2015-12-22
```

- Performed an SQL query to find the dates of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

- Performed an SQL query to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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

In [22]:

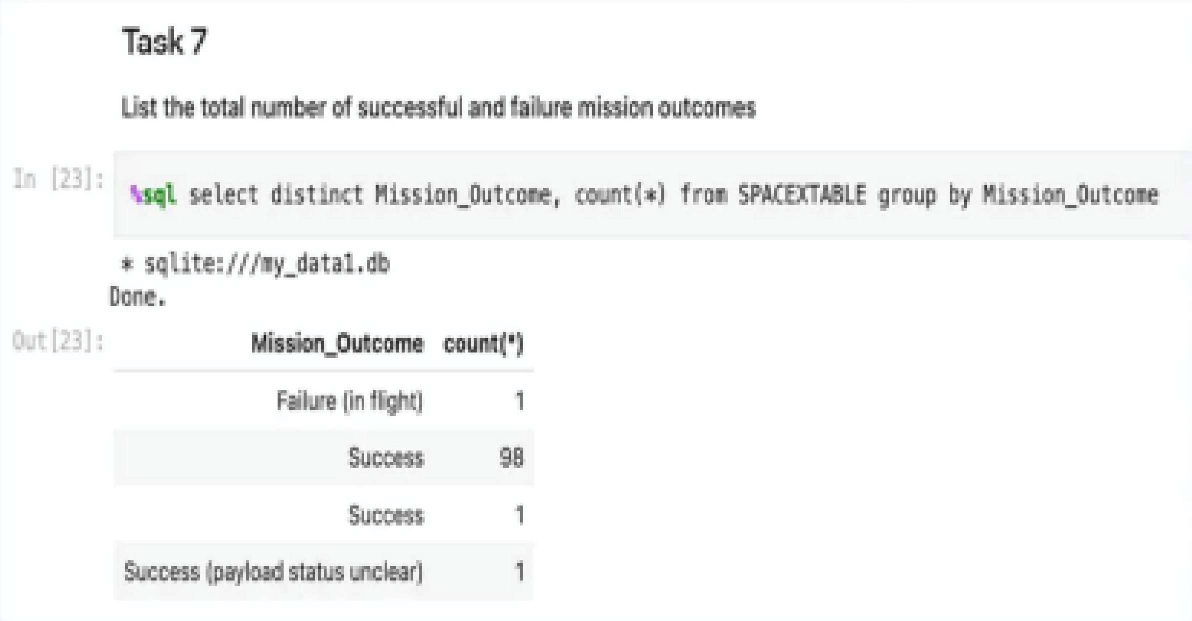
```
%sql select distinct Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_
```

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

Out[22]:

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

Total Number of Successful and Failure Mission Outcomes



Task 7

List the total number of successful and failure mission outcomes

In [23]:
```
%sql select distinct Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
```

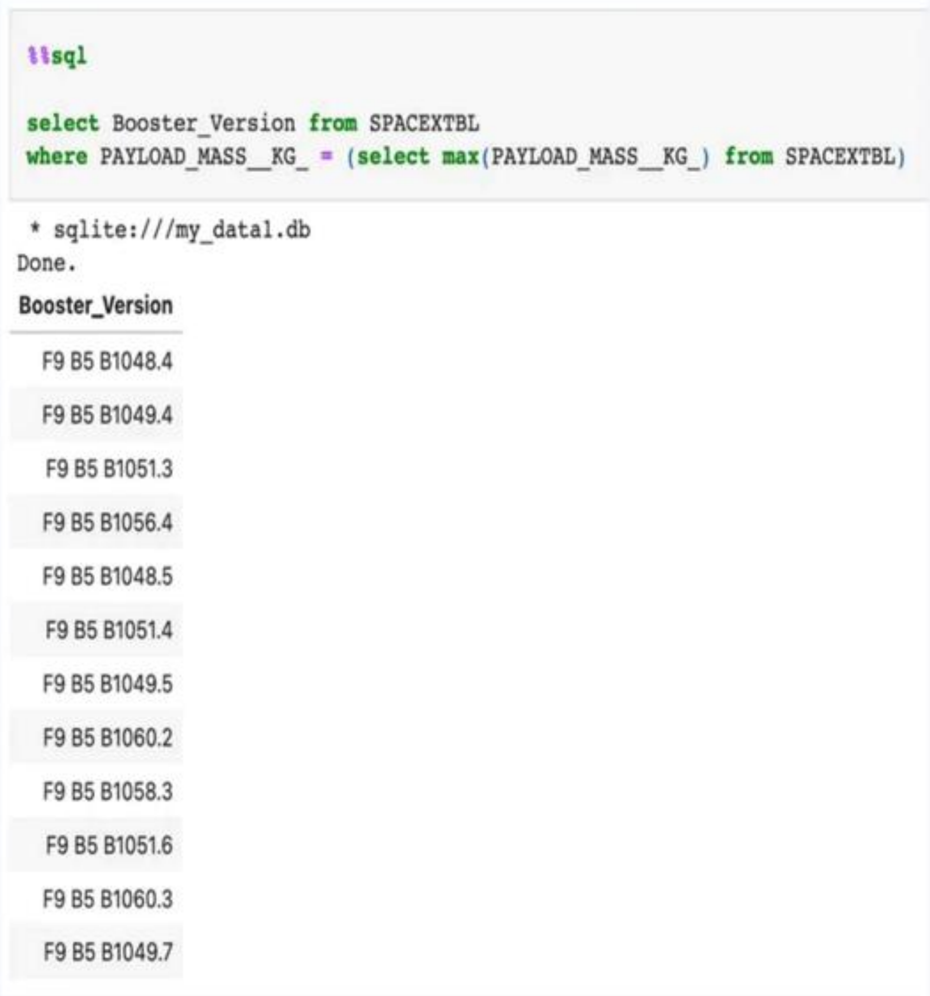
```
 * sqlite:///my_data1.db
Done.
```

Out[23]:

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

- Performed an SQL query to calculate the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload



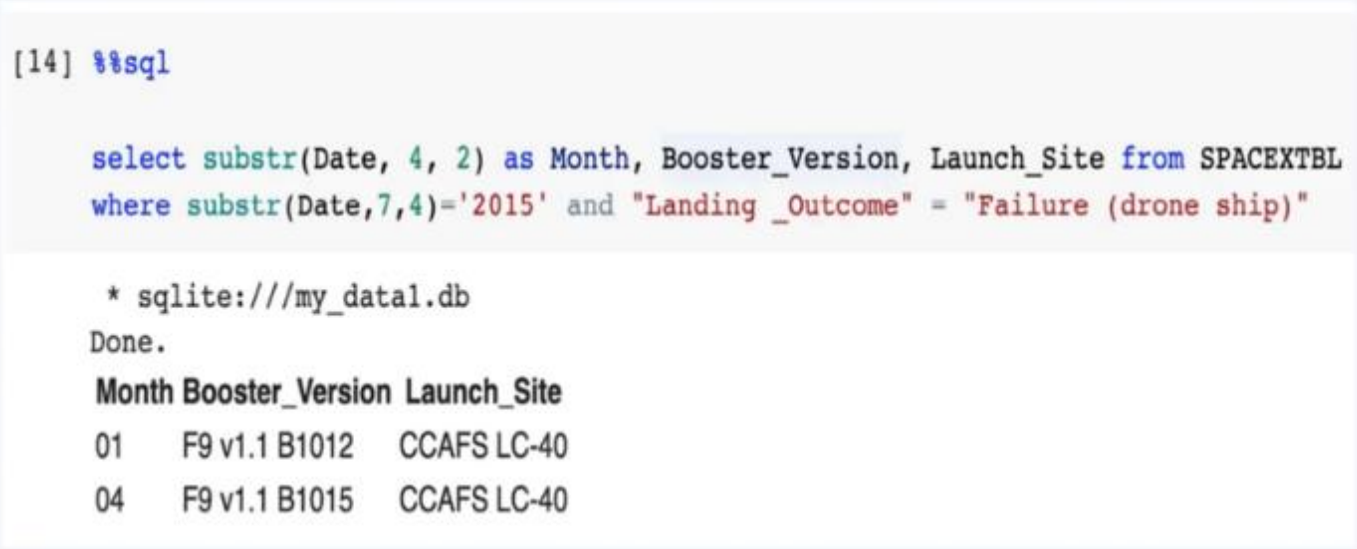
```
%%sql

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

- The names of the booster which have carried the maximum payload mass.

2015 Launch Records



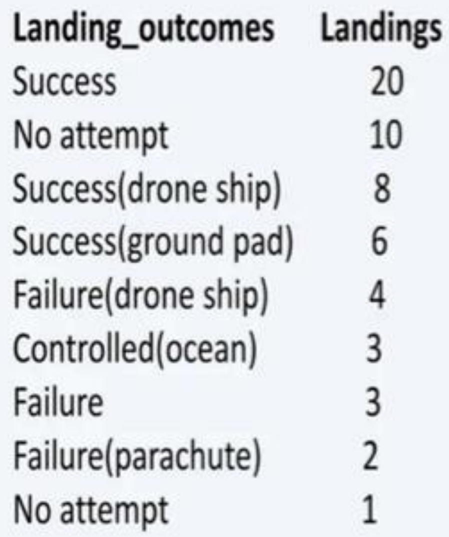
```
[14] %%sql

     select substr(Date, 4, 2) as Month, Booster_Version, Launch_Site from SPACEXTBL
     where substr(Date,7,4)='2015' and "Landing _Outcome" = "Failure (drone ship)"
```

```
 * sqlite:///my_data1.db
Done.
Month Booster_Version  Launch_Site
01    F9 v1.1 B1012    CCAFS LC-40
04    F9 v1.1 B1015    CCAFS LC-40
```

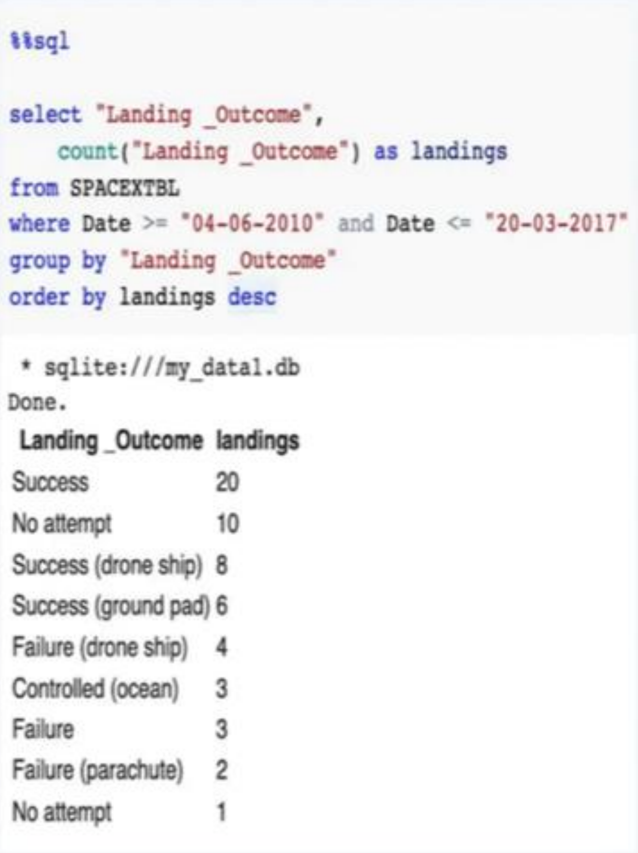
Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

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



Landing_outcomes	Landings
Success	20
No attempt	10
Success(drone ship)	8
Success(ground pad)	6
Failure(drone ship)	4
Controlled(ocean)	3
Failure	3
Failure(parachute)	2
No attempt	1



```
%%sql

select "Landing _Outcome",
    count("Landing _Outcome") as landings
from SPACEXTBL
where Date >= "04-06-2010" and Date <= "20-03-2017"
group by "Landing _Outcome"
order by landings desc
```

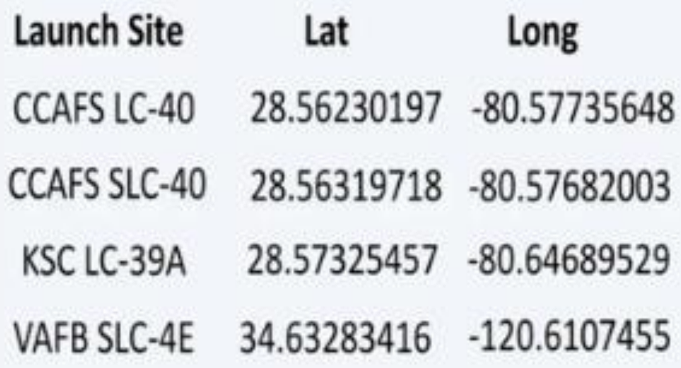
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	landings
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Controlled (ocean)	3
Failure	3
Failure (parachute)	2
No attempt	1

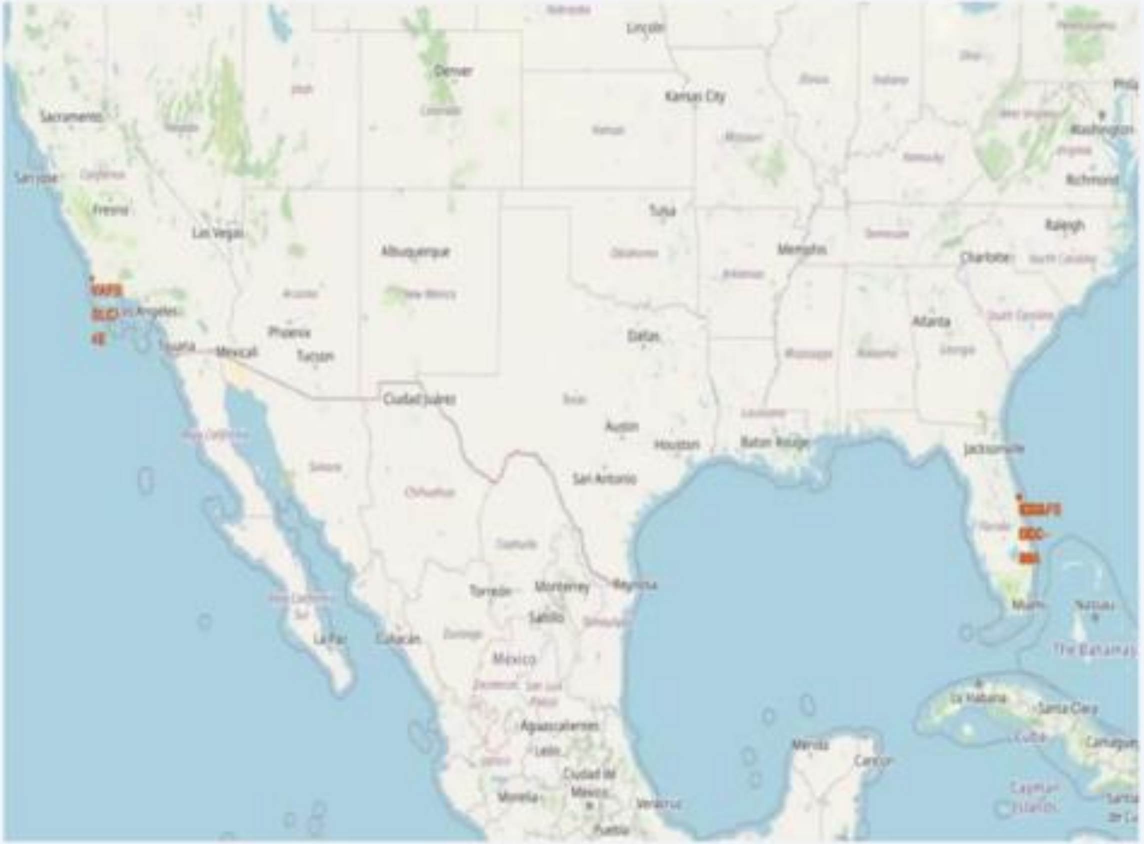
Section 3

Launch Sites Proximities Analysis

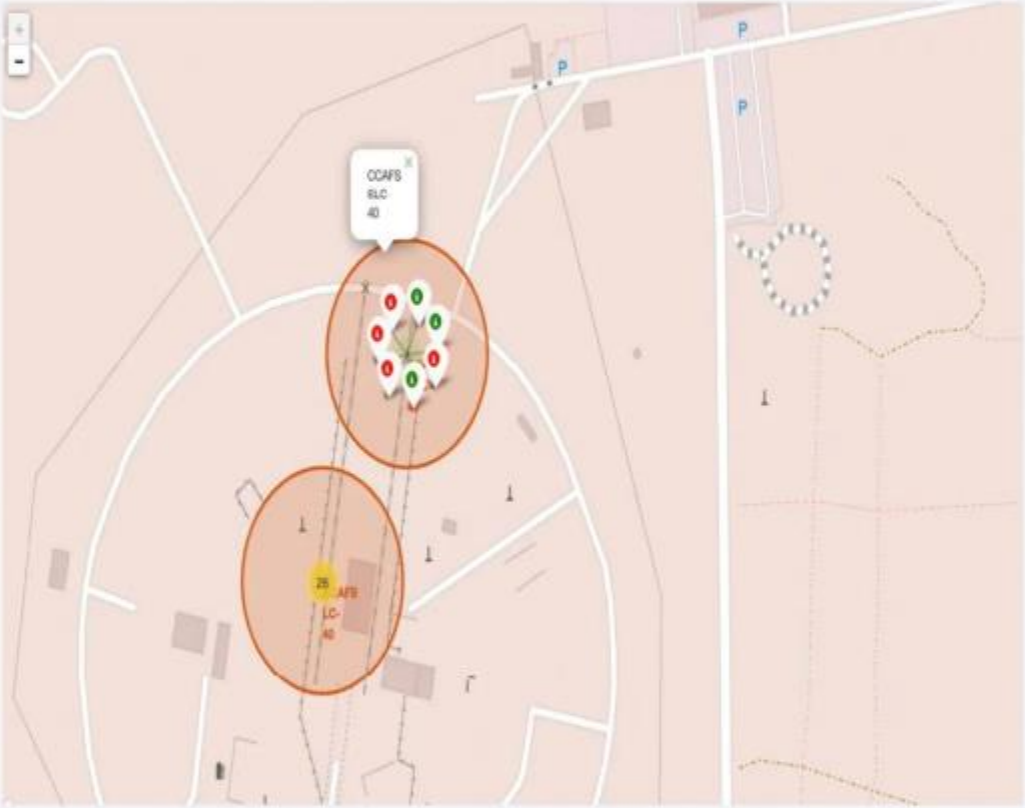
All launch sites on a map



Launch Site	Lat	Long
CCAFS LC-40	28.56230197	-80.57735648
CCAFS SLC-40	28.56319718	-80.57682003
KSC LC-39A	28.57325457	-80.64689529
VAFB SLC-4E	34.63283416	-120.6107455

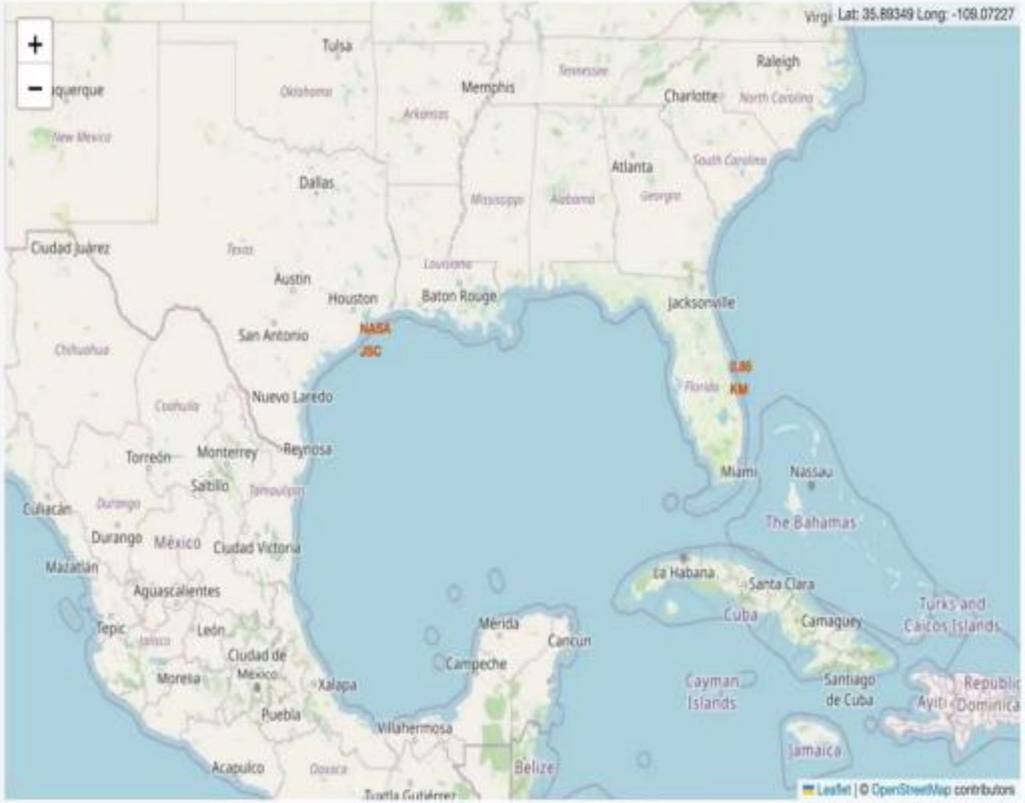


All success/failed launches for each site on the map

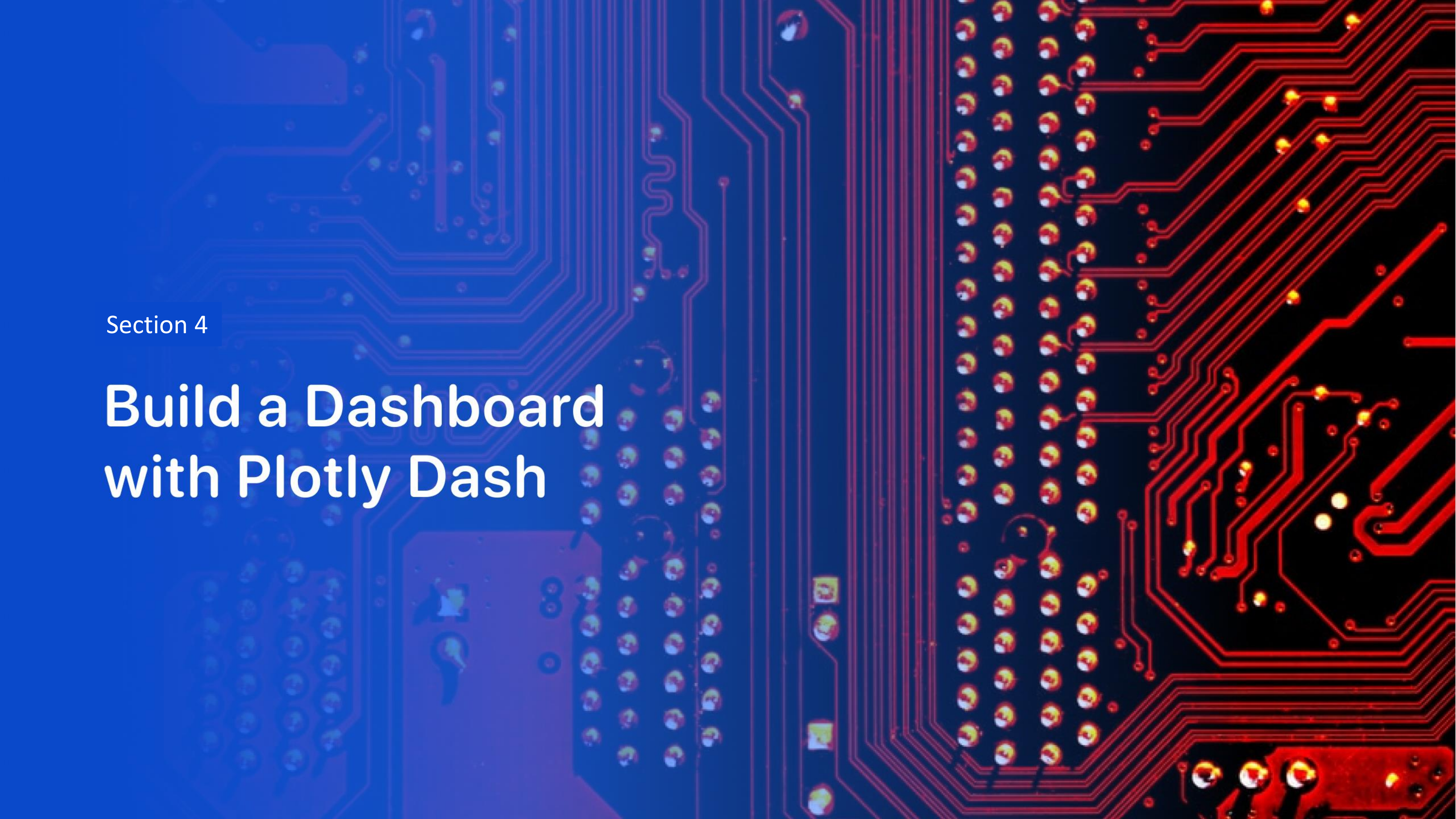


- The launch records are grouped in clusters on the map, then labelled by green markers for successful launches, and red markers for unsuccessful ones.

Distances between a launch site to its proximities



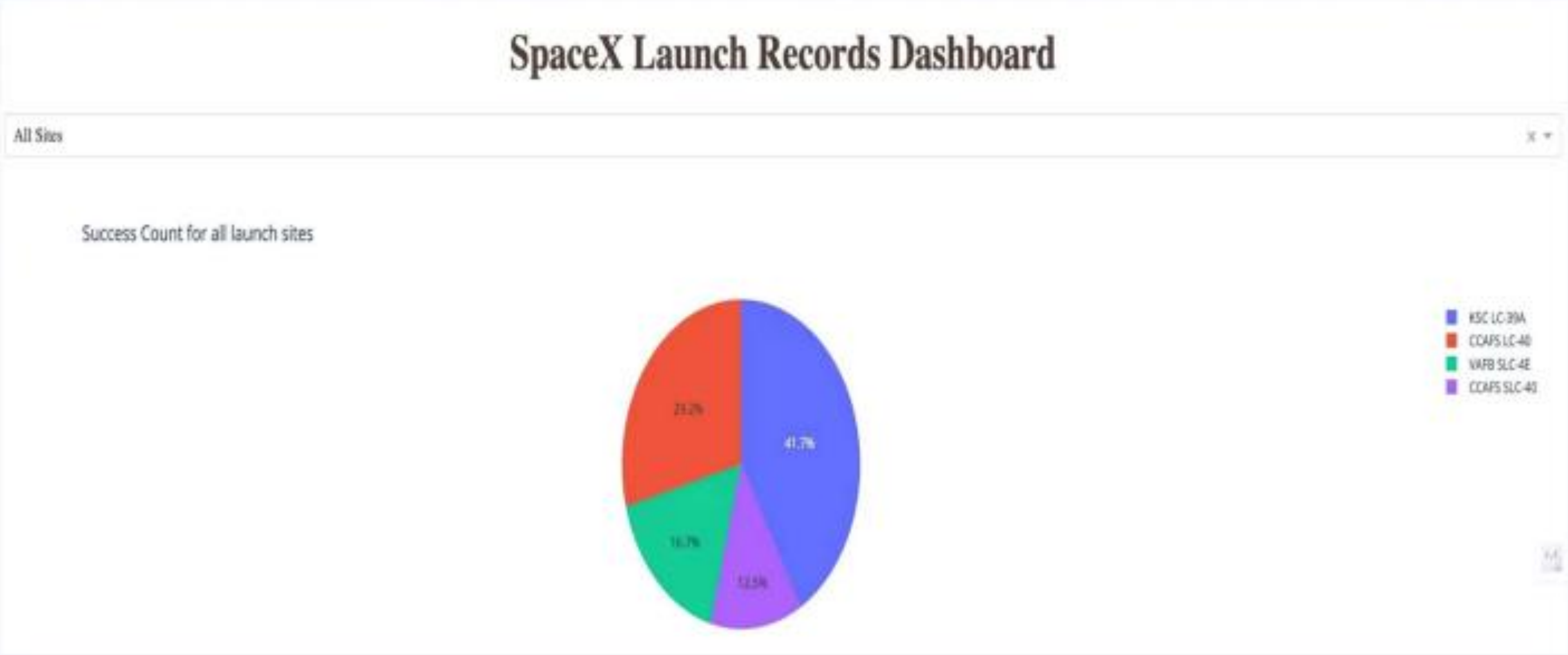
- The closest coastline from NASA JSC is marked as a point using Mouse Positions and the distance between the coastline point and the launch site, which is approximately 0.86 km.



Section 4

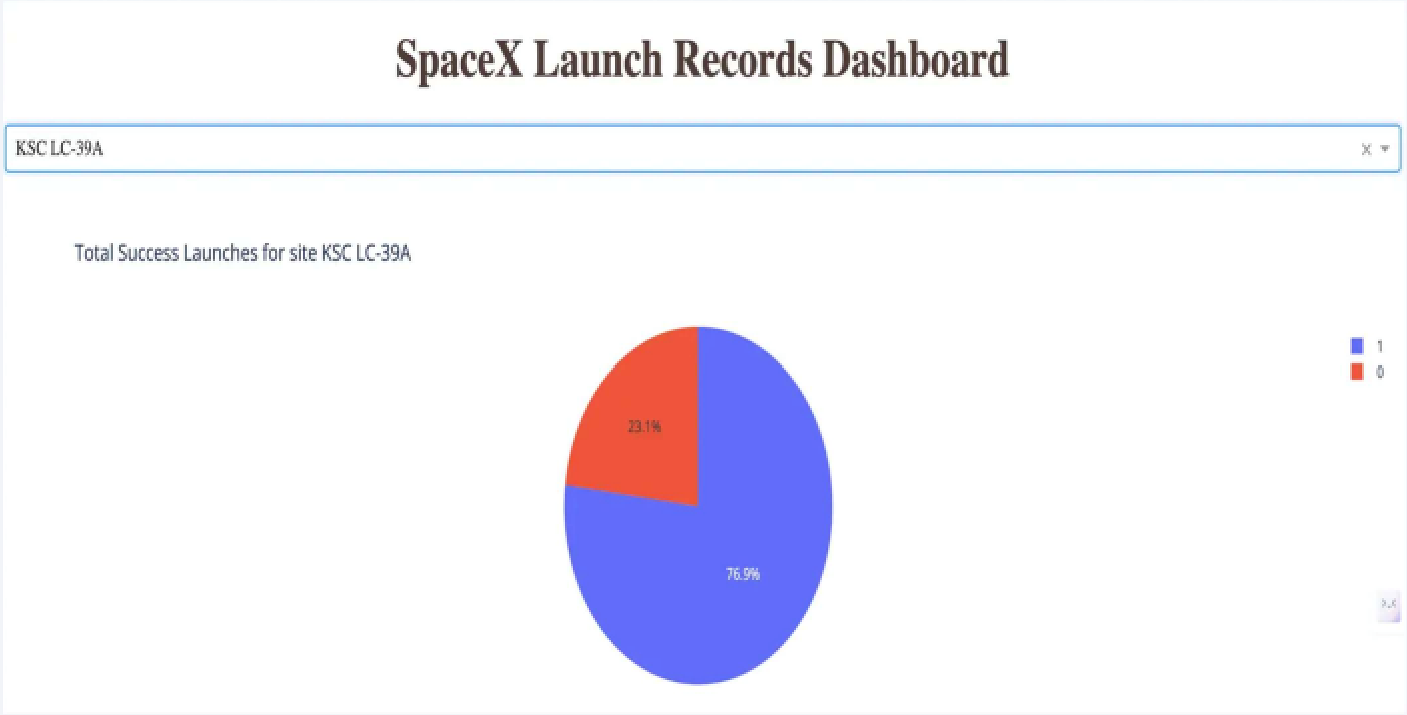
Build a Dashboard with Plotly Dash

Total success launches for all sites



- KSC LC-39A has the highest amount of success launches with a 41.7% from the entire record, whereas CCAFS SLC-40 has the lowest amount of success launches with only 12.5%

Success ratio of the launch site with the highest success launches



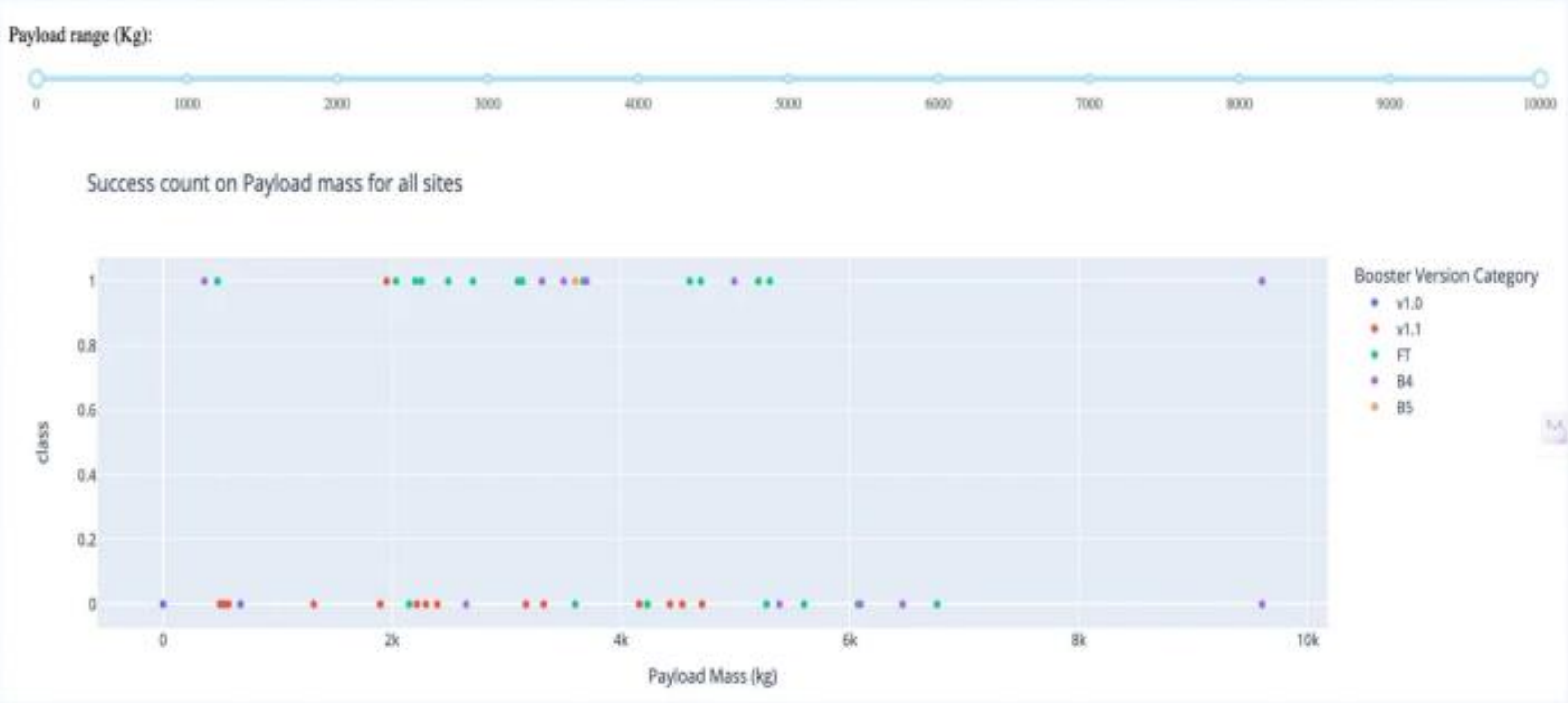
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

- KSC LC-39A which is the launch site with highest amount of success, has 76.9% success rate for the launches from its site, and 23.1% failure rate.

Payload vs. Launch outcome



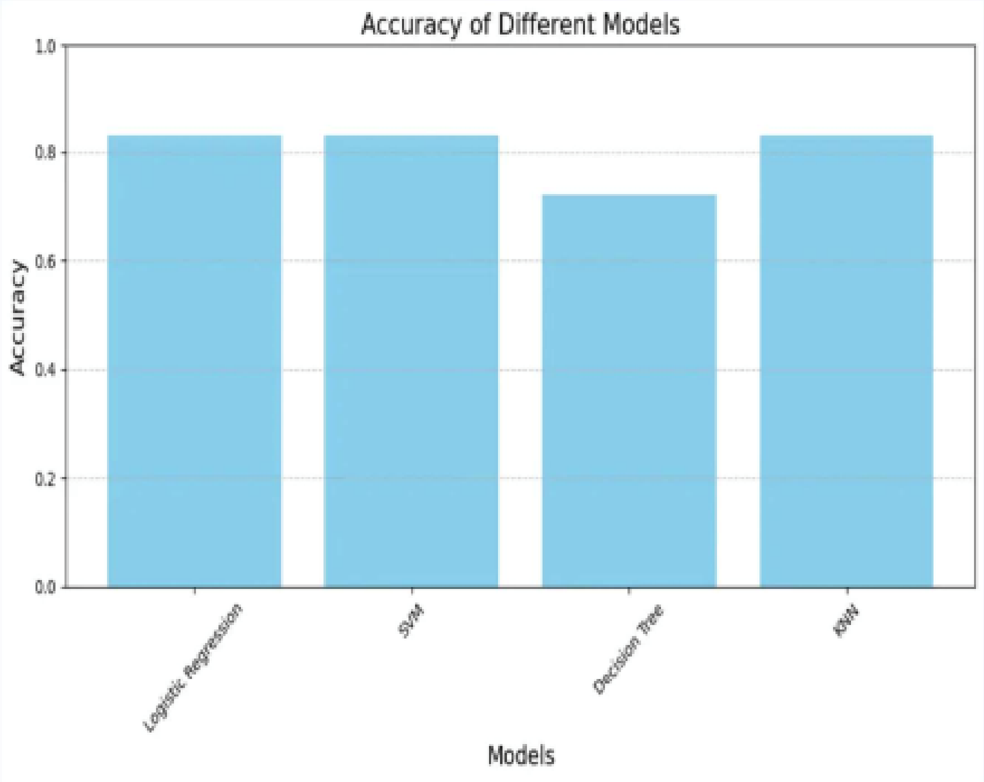
- The payload range that has the highest success launches is between 2,000 to 4,000 kg, which can be seen by the greatest number of plots in that range, followed by the payload range of 4,000 to 6,000 kg, with the second the greatest number of plot.

- Booster version FT (green spots) has the highest success launches, among all boster versions.

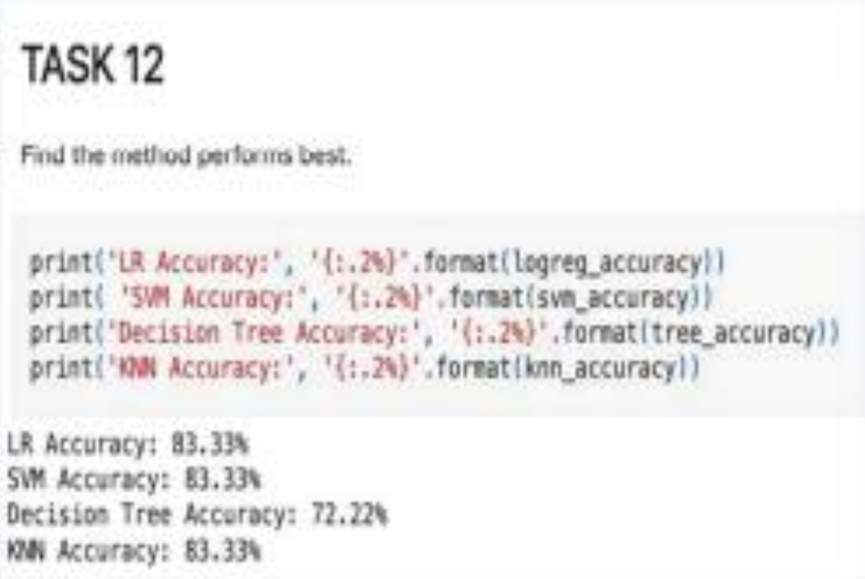
Section 5

Predictive Analysis (Classification)

Classification Accuracy



- The model that performed best are LR, SVM, KNN where all 3 achivied the highest accuracy of 83.33%



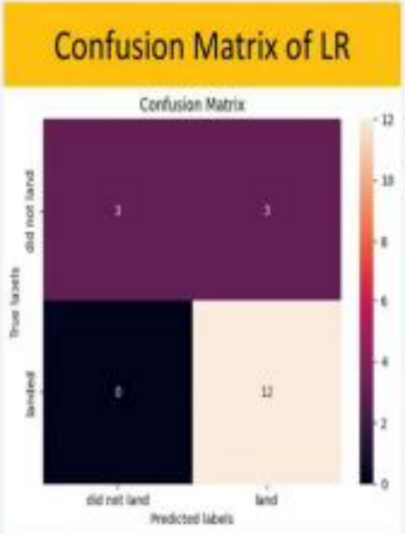
TASK 12

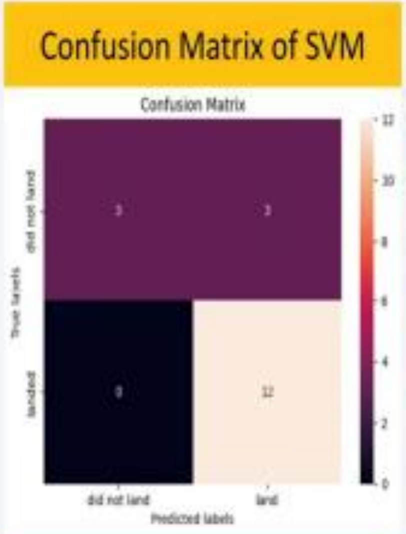
Find the method performs best.

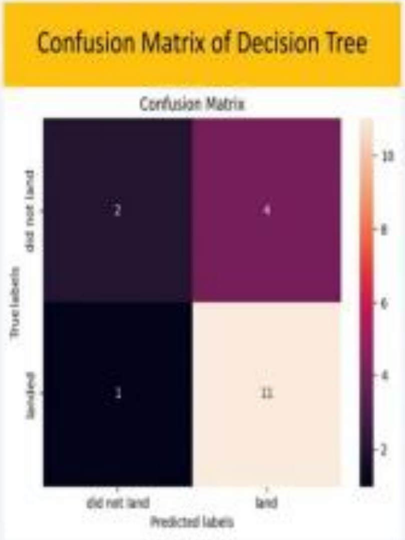
```
print('LR Accuracy:', '{:.2%}'.format(logreg_accuracy))
print( 'SVM Accuracy:', '{:.2%}'.format(svm_accuracy))
print('Decision Tree Accuracy:', '{:.2%}'.format(tree_accuracy))
print('KNN Accuracy:', '{:.2%}'.format(knn_accuracy))
```

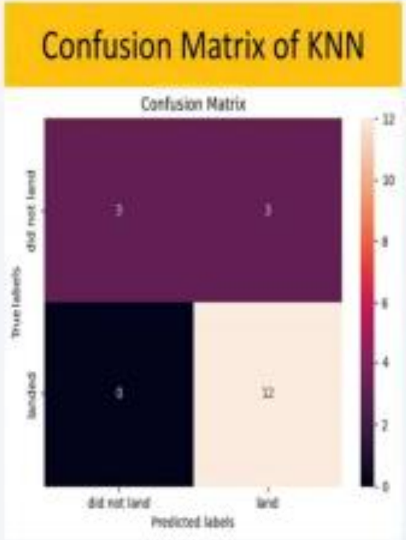
```
LR Accuracy: 83.33%
SVM Accuracy: 83.33%
Decision Tree Accuracy: 72.22%
KNN Accuracy: 83.33%
```

Confusion Matrix


Confusion Matrix of LR

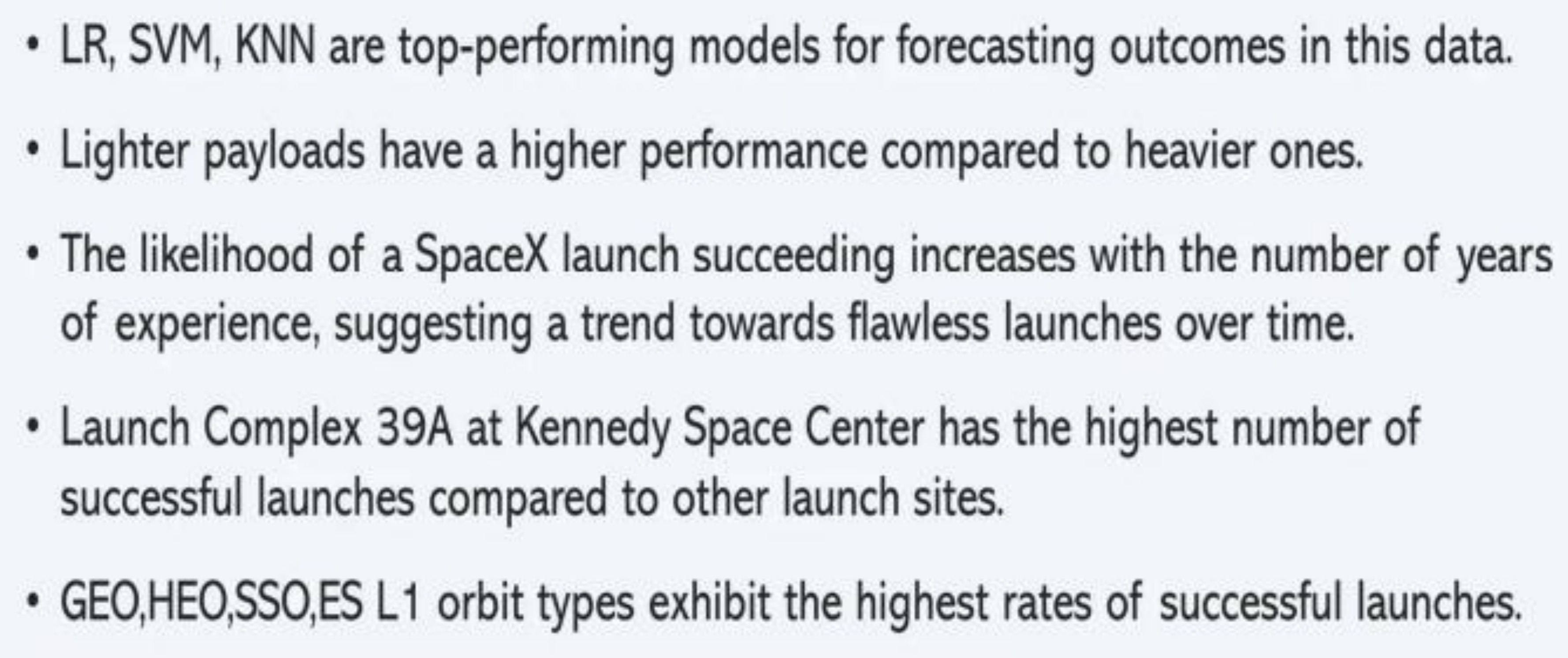

Confusion Matrix of SVM


Confusion Matrix of Decision Tree


Confusion Matrix of KNN

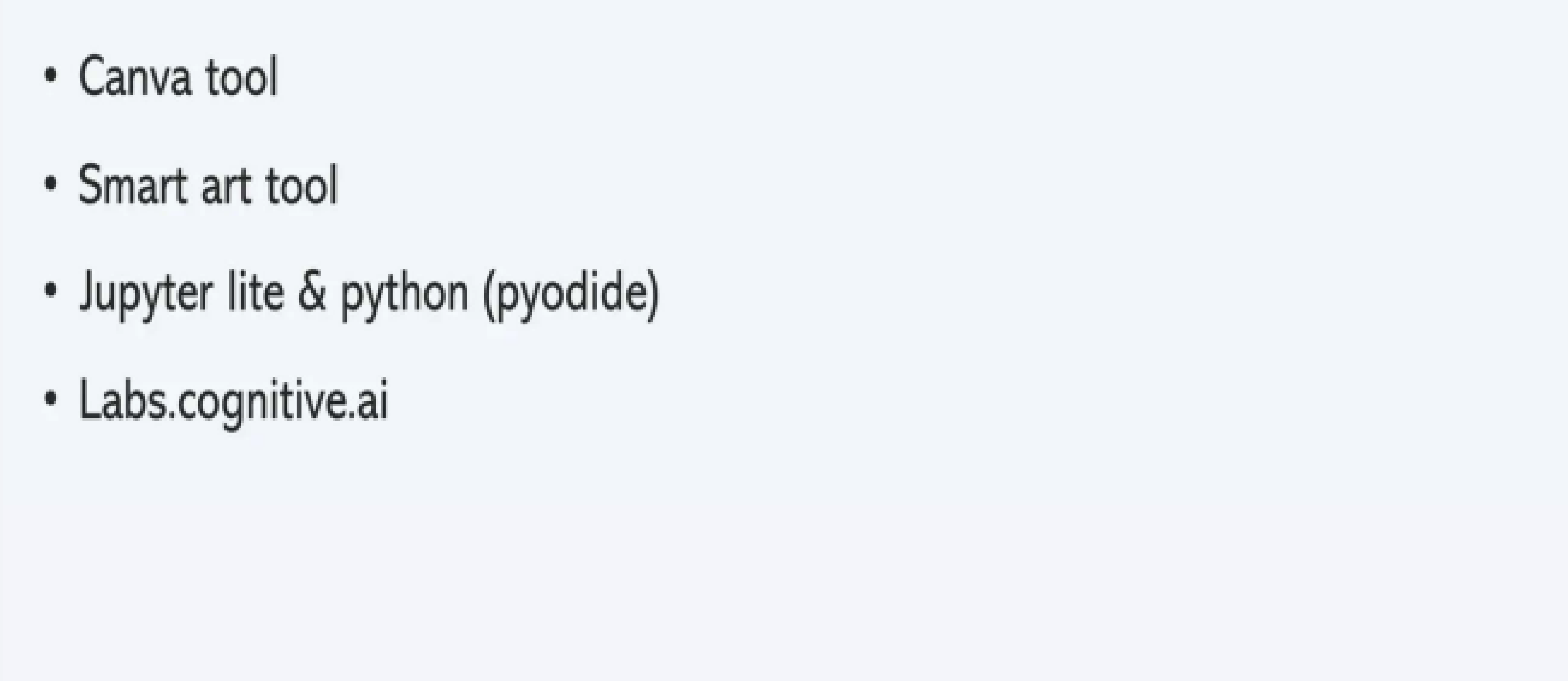
- LR, SVM, KNN models are good as their confusion matrix show that they predicted all 12 successful landing correctly, with 0 error.
- However, the Decision Tree model only predicted 11 successful landing correctly, with one of them wrongly predicted as a failed / did not land.
- LR, SVM, KNN models have the same accuracy of 83.33% as displayed earlier, hence the same confusion matrix.

Conclusions



- LR, SVM, KNN are top-performing models for forecasting outcomes in this data.
- Lighter payloads have a higher performance compared to heavier ones.
- The likelihood of a SpaceX launch succeeding increases with the number of years of experience, suggesting a trend towards flawless launches over time.
- Launch Complex 39A at Kennedy Space Center has the highest number of successful launches compared to other launch sites.
- GEO,HEO,SSO,ES L1 orbit types exhibit the highest rates of successful launches.

Appendix



- Canva tool
- Smart art tool
- Jupyter lite & python (pyodide)
- Labs.cognitive.ai

Thank you!