IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wilson Carvalho
October 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API and Web Scaping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium and Dashboard with Ploty Dash
  - Machine Learning Prediction

- Summary of all results
  - Optimal modeling
  - Visualization for decision making

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
    - How do variables determine if the rocket will land successfully?
    - The interaction amongst various features that determine the success rate of a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using Rest API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

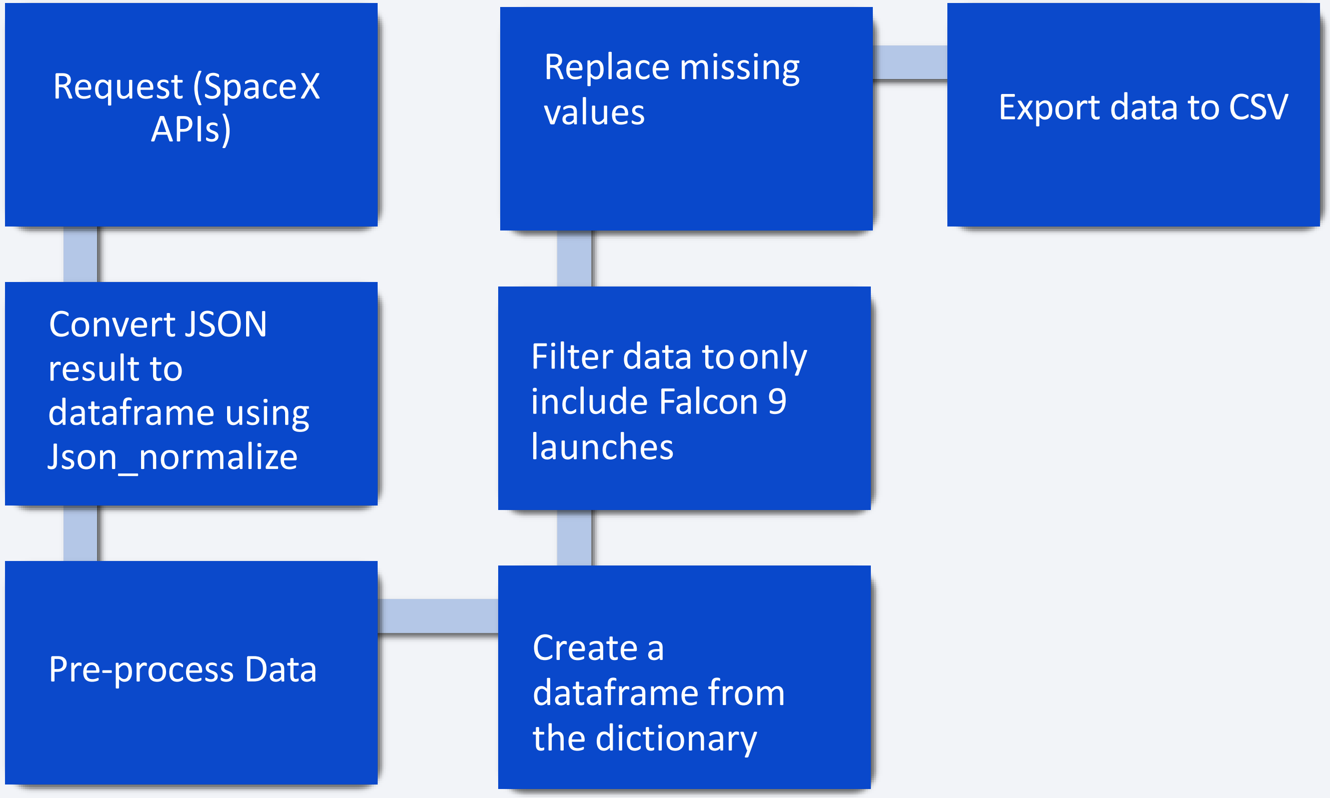
Data Collection

- Data was collected using various methods and steps:
  - Data collection using get request to the SpaceX API.
  - Response content decoded as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
  - Data cleaned, checked for missing values and fill in missing.
  - Web scraping perfromed from Wikipedia for Falcon 9 launch records with BeautifulSoup.
  - Launch records extracted as HTML table, parse the table and convert to a pandas dataframe for future analysis.

Data Collection – SpaceX API

- Data collection with SpaceX REST calls using get request. Data wrangling and formatting was performed.

- GitHub link to the notebook: https://github.com/Cawi27/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Request (SpaceX APIs)

Convert JSON result to dataframe using Json_normalize

Pre-process Data

Create a dataframe from the dictionary

Filter data to only include Falcon 9 launches

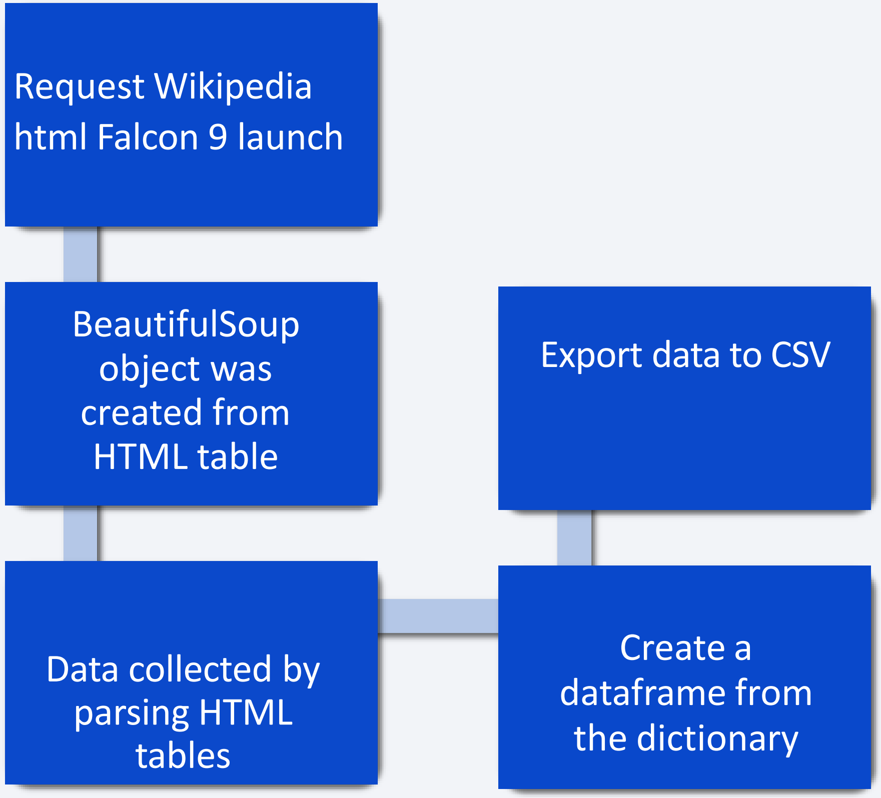
Replace missing values

Export data to CSV

Data Collection - Scraping

- Web scrapping was applied to Falcon 9 launch records with BeautifulSoup. The table was converted into a pandas dataframe.

- GitHub link to the notebook: https://github.com/Cawi27/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Request Wikipedia html Falcon 9 launch

BeautifulSoup object was created from HTML table

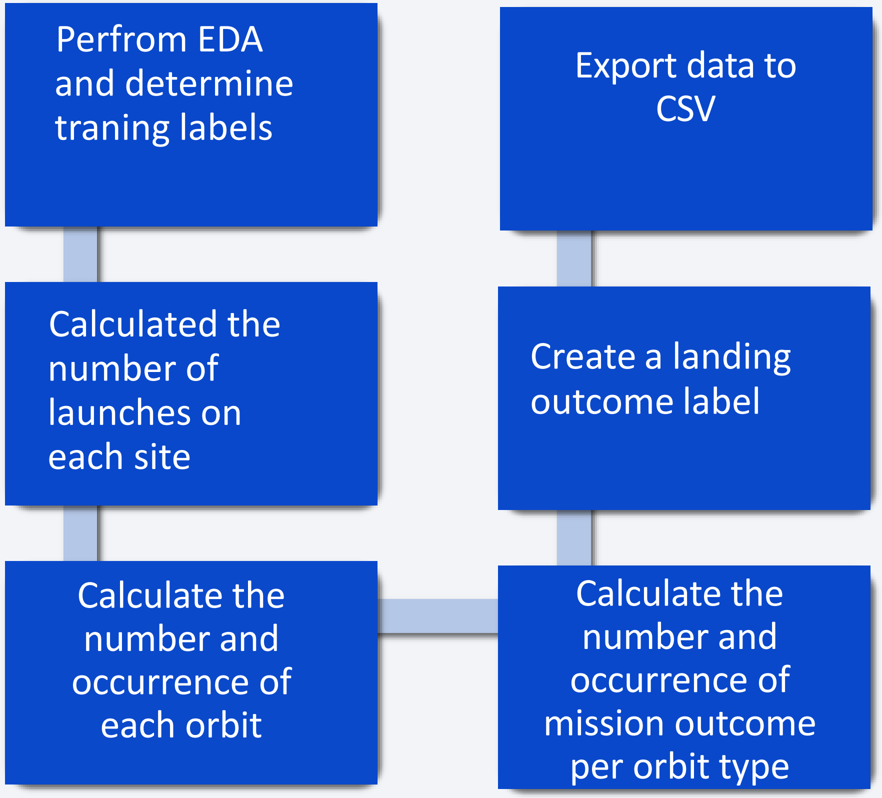
Data collected by parsing HTML tables

Create a dataframe from the dictionary

Export data to CSV

Data Wrangling

- Performed exploratory data analysis and determined the training labels.
- Calculated the number of launches at each site, and the number and occurrence of each orbits
- Created landing outcome label from outcome column and exported the results to csv.
- GitHub link to the notebook: https://github.com/Cawi27/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



Perfrom EDA and determine traning labels

Calculated the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

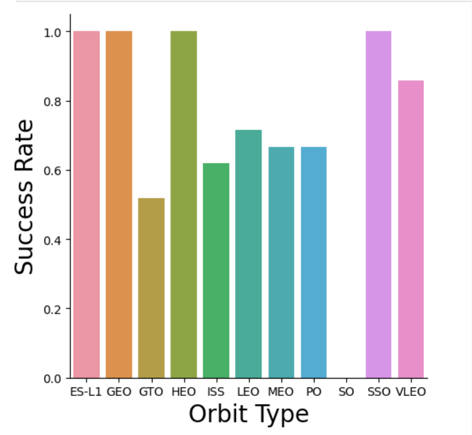
Create a landing outcome label

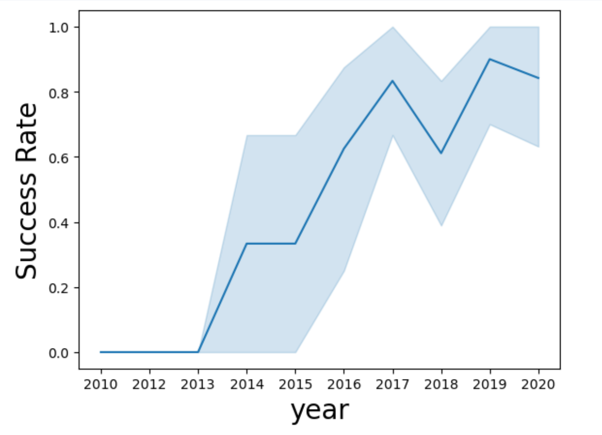
Export data to CSV

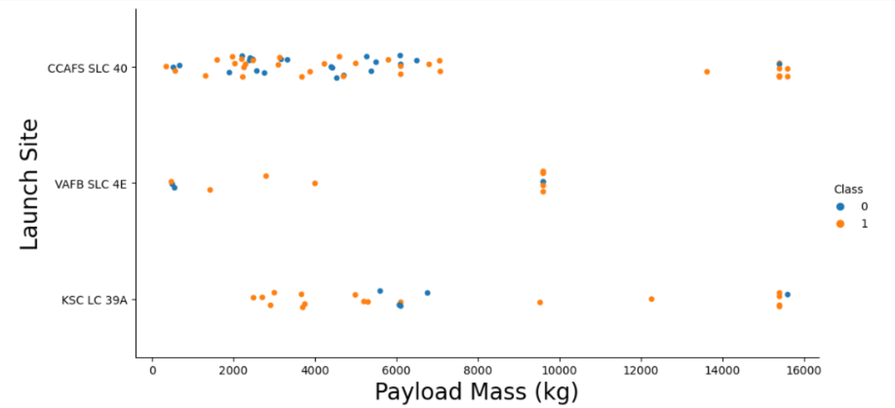
EDA with Data Visualization

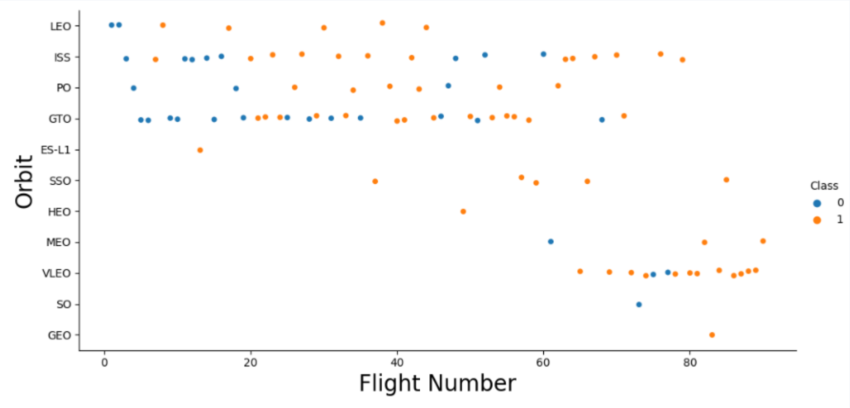
- Charts were used to visualize the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.

- All charts are available on this link: https://github.com/Cawi27/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb









EDA with SQL

- SQL queries performed:
  - Names of unique launch sites in the space mission.
  - 5 records where launch sites begin with the string 'CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Listed the date when the first successful landing outcome in ground pad was achieved.
  - Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - Listed the total number of successful and failure mission outcomes
  - Listed the names of the booster_versions which have carried the maximum payload mass.
  - Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Ranked 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.
- The link to the notebook is https://github.com/Cawi27/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- All launch sites were marked, and map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map were added.
- Feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success were assigned.
- Using the color-labeled marker clusters, launch sites have relatively high success rate were identified.
- The distances between a launch site to its proximities (like railways, highways and coastlines) were calculated. We answered some question for instance:

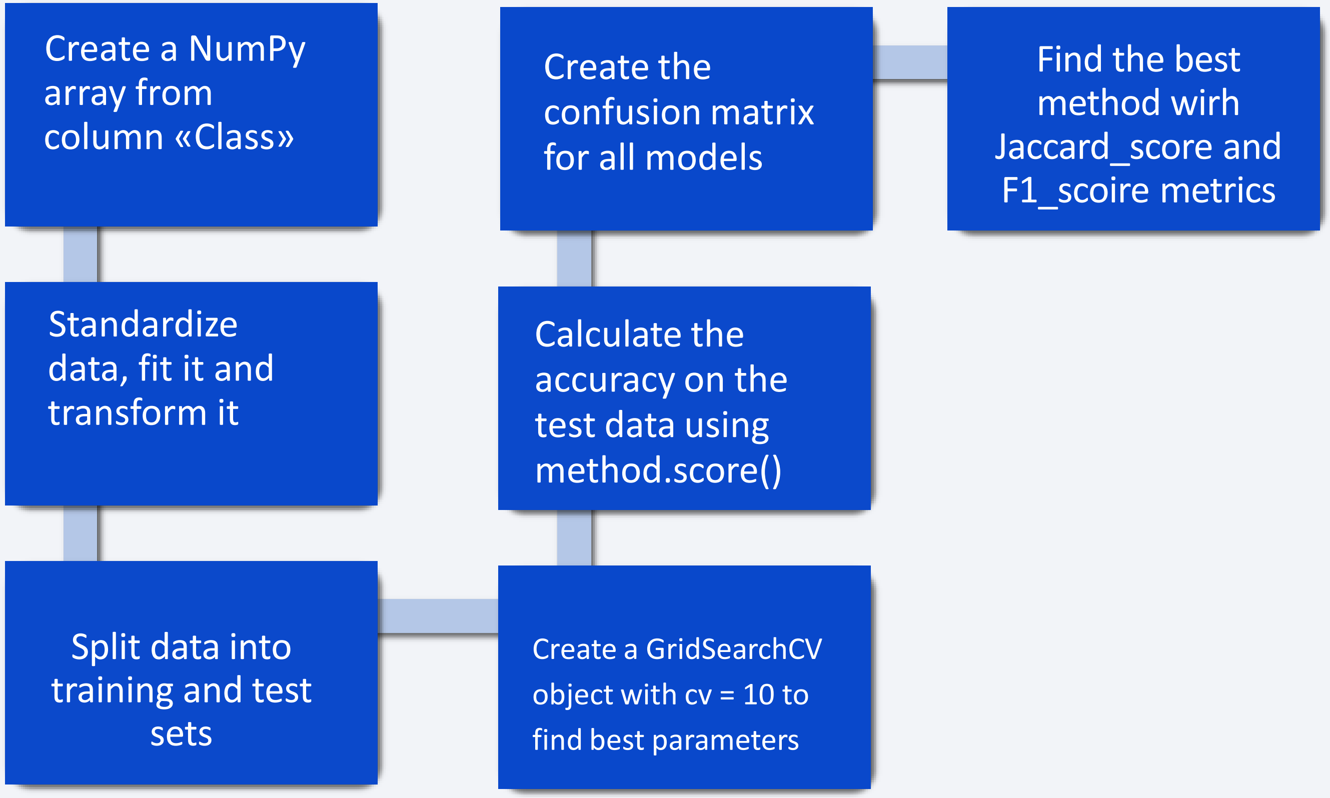
- GitHub link to the notebook: https://github.com/Cawi27/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard with Plotly dash was built with the following charts:
  - Pie charts showing the total launches by a certain sites
  - Scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- GitHub link to the notebook: https://github.com/Cawi27/Applied-Data-Science-Capstone/blob/main/TASK%201%20Add%20a%20Launch%20Site%20Drop-down%20Input%20Component.png

Predictive Analysis (Classification)

- Models were built to make predications
- GitHub link to the notebook: https://github.com/Cawi27/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Create a NumPy array from column «Class»

Standardize data, fit it and transform it

Split data into training and test sets

Create a GridSearchCV object with cv = 10 to find best parameters

Calculate the accuracy on the test data using method.score()

Create the confusion matrix for all models

Find the best method wirh Jaccard_score and F1_scoire metrics

Results

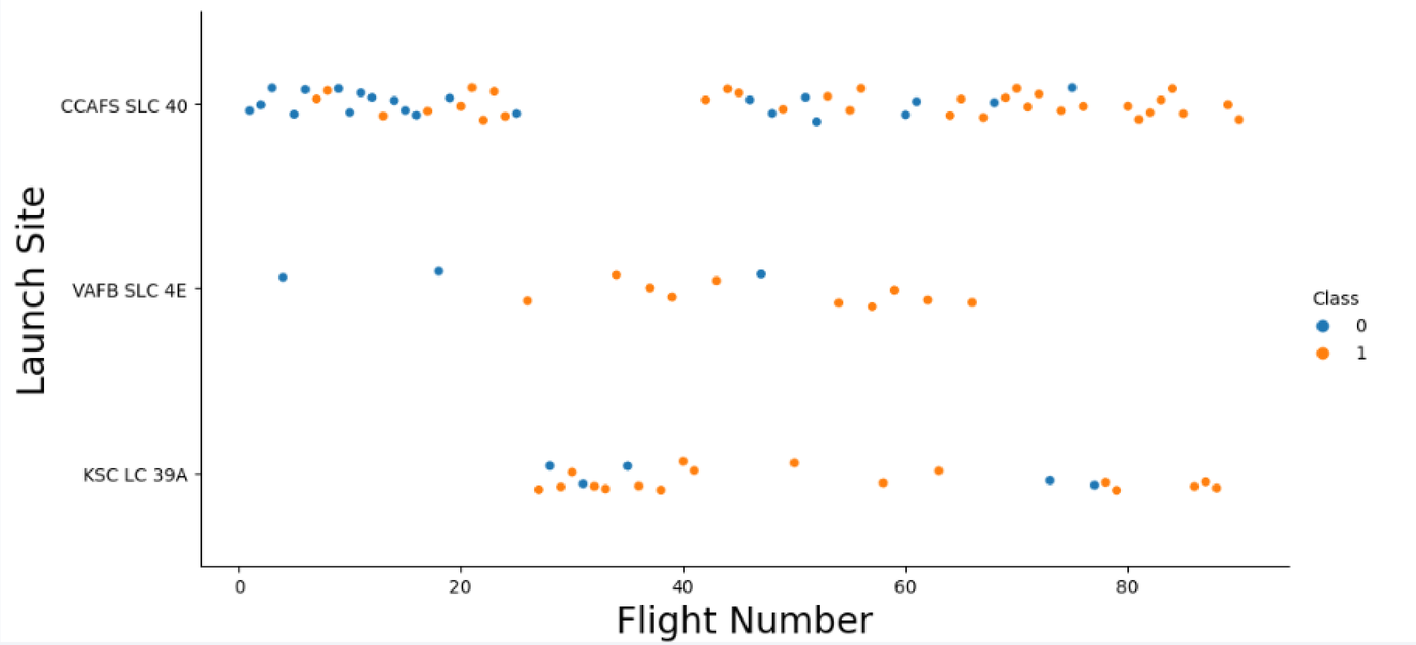
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

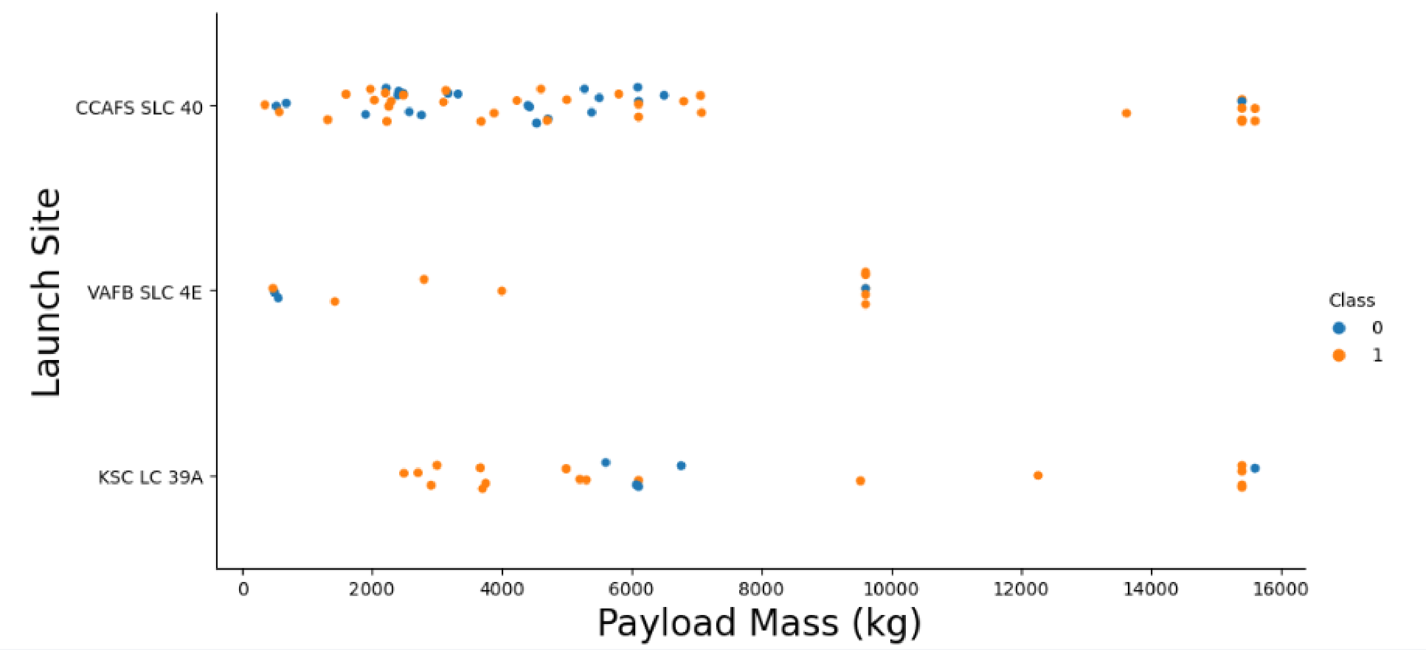
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



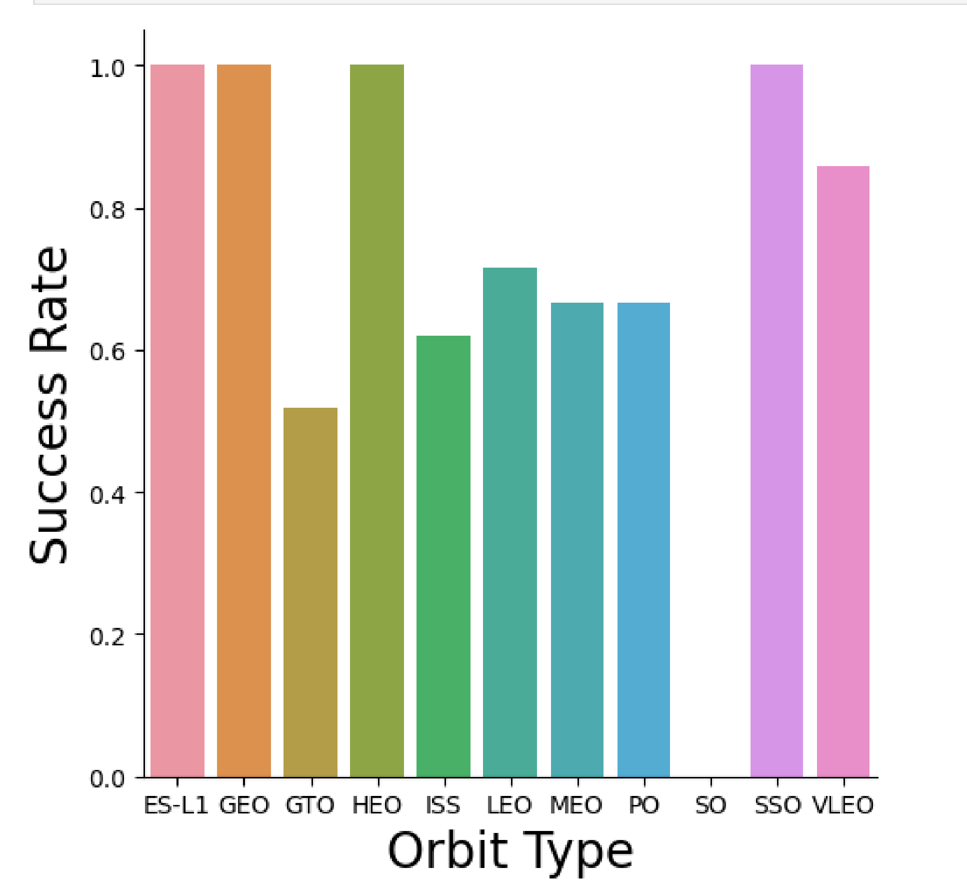
Payload vs. Launch Site

- The greater the payload mass launch site CCAFS SLC 40 the higher the success rate.



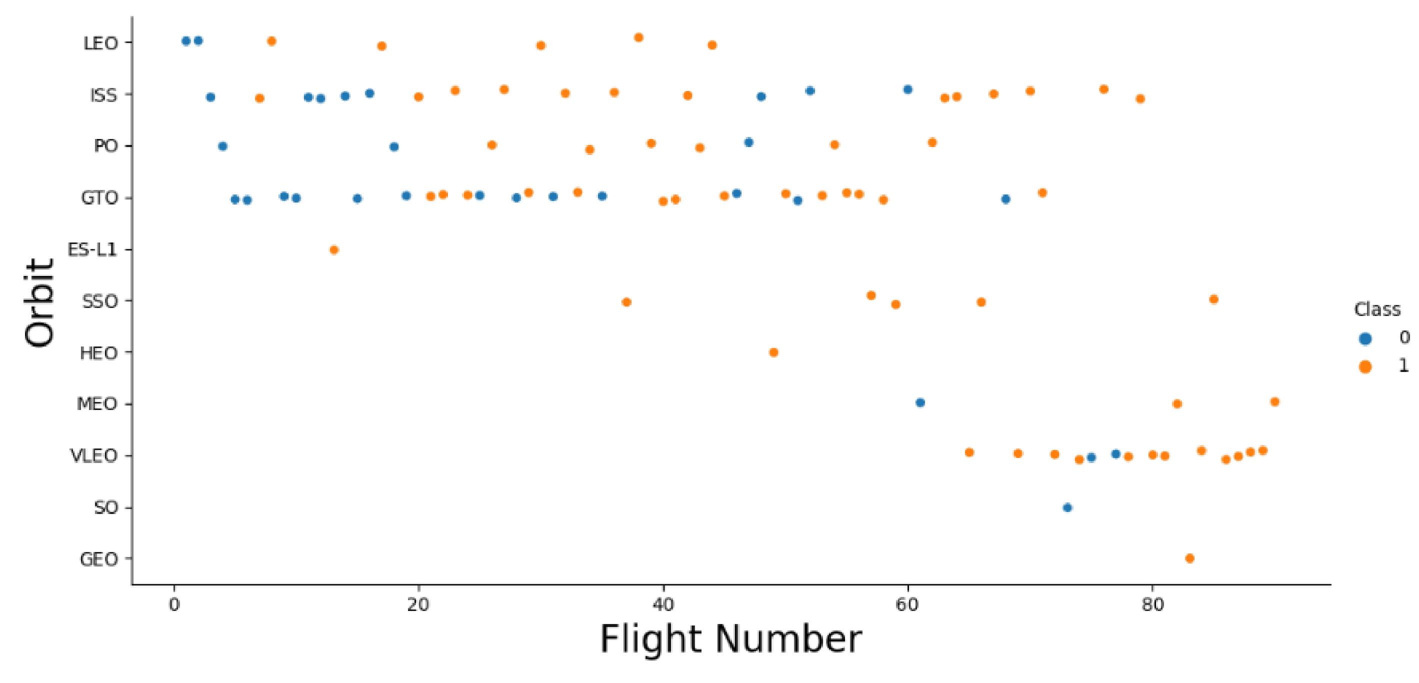
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO had the most success rate.



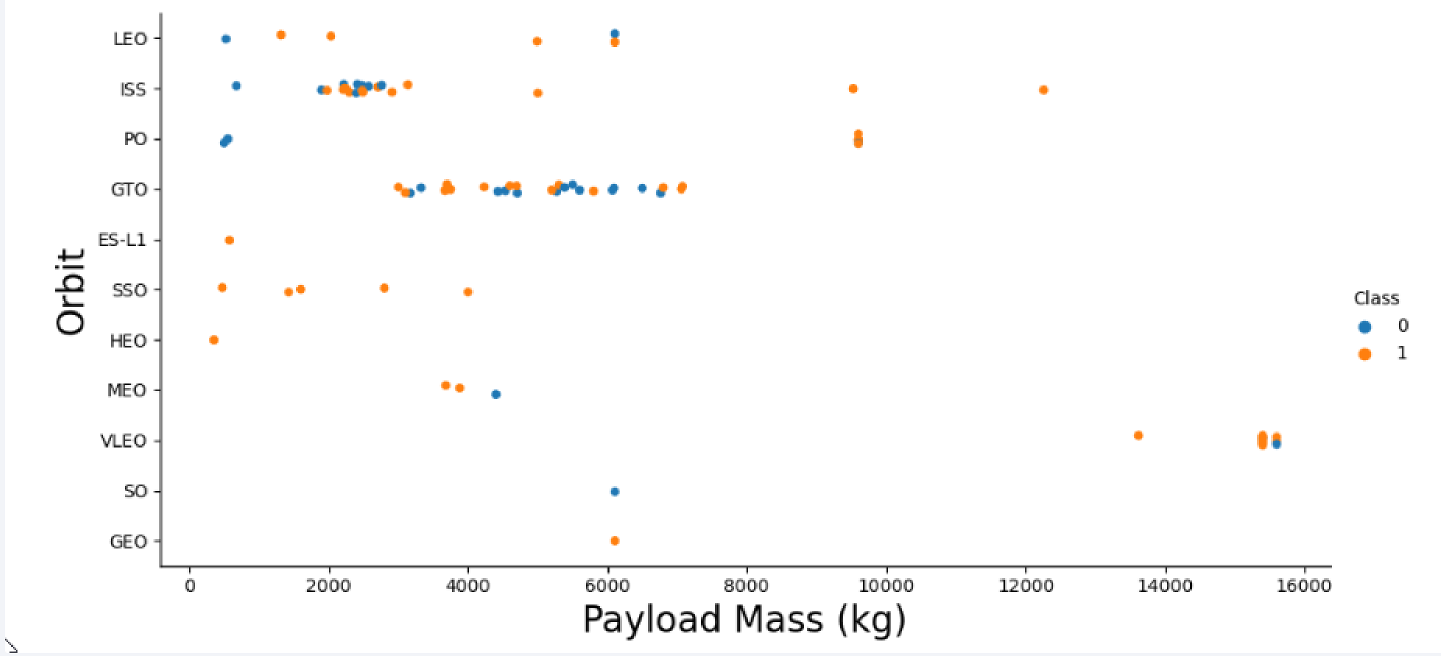
Flight Number vs. Orbit Type

- In the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



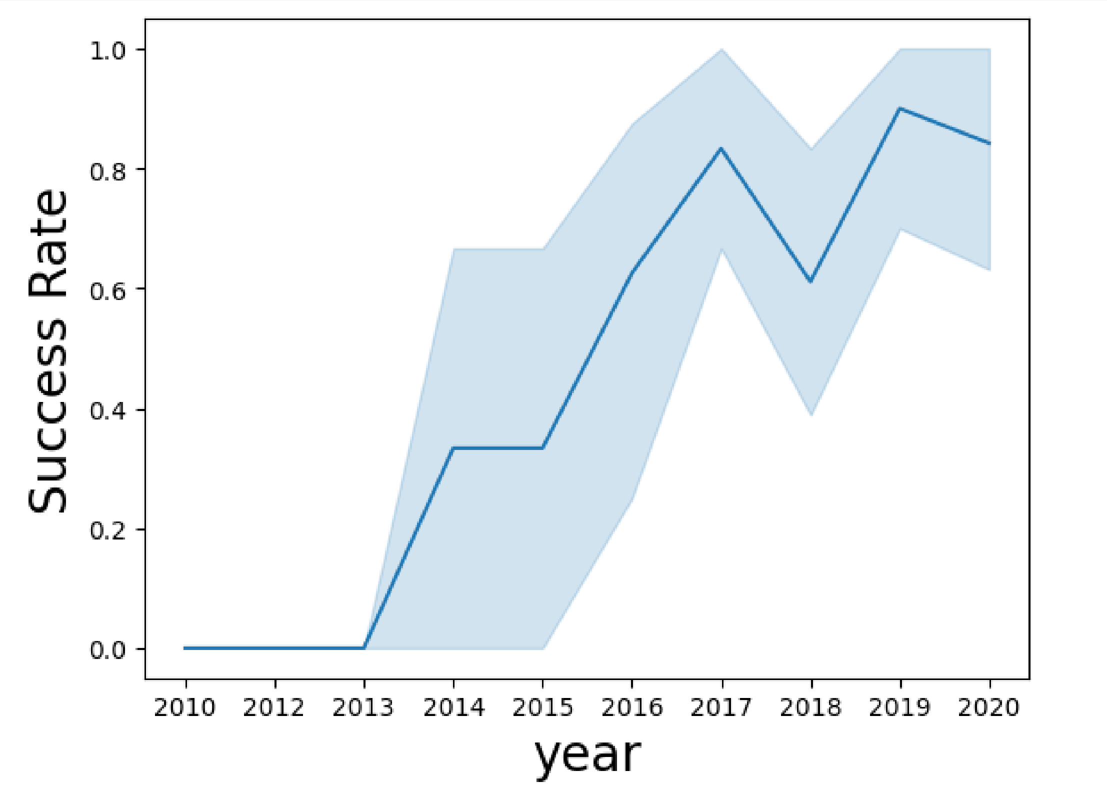
Payload vs. Orbit Type

- With heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



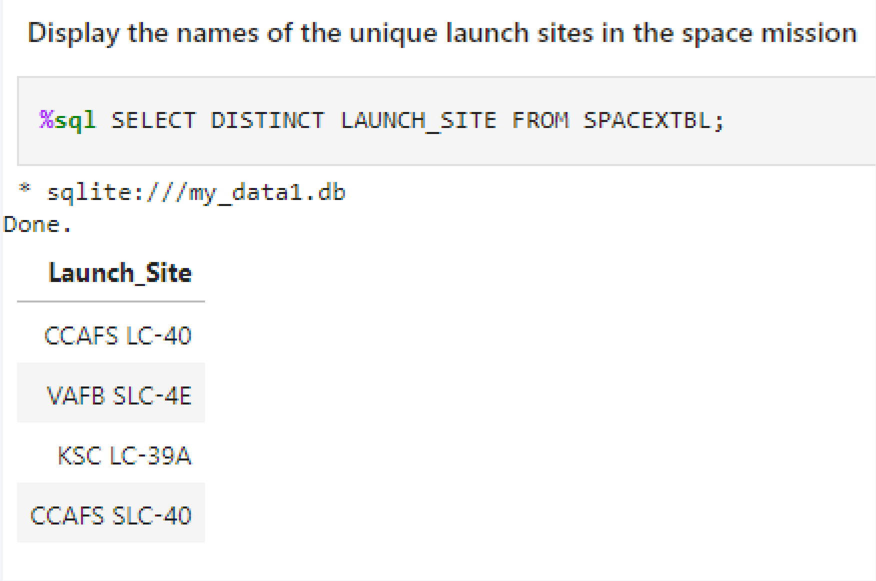
Launch Success Yearly Trend

- Success rate since 2013 kept on increasing till 2020. despite a drop in 2018.



All Launch Site Names

- The key word DISTINCT was used to show only unique launch sites from the SpaceX data.



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

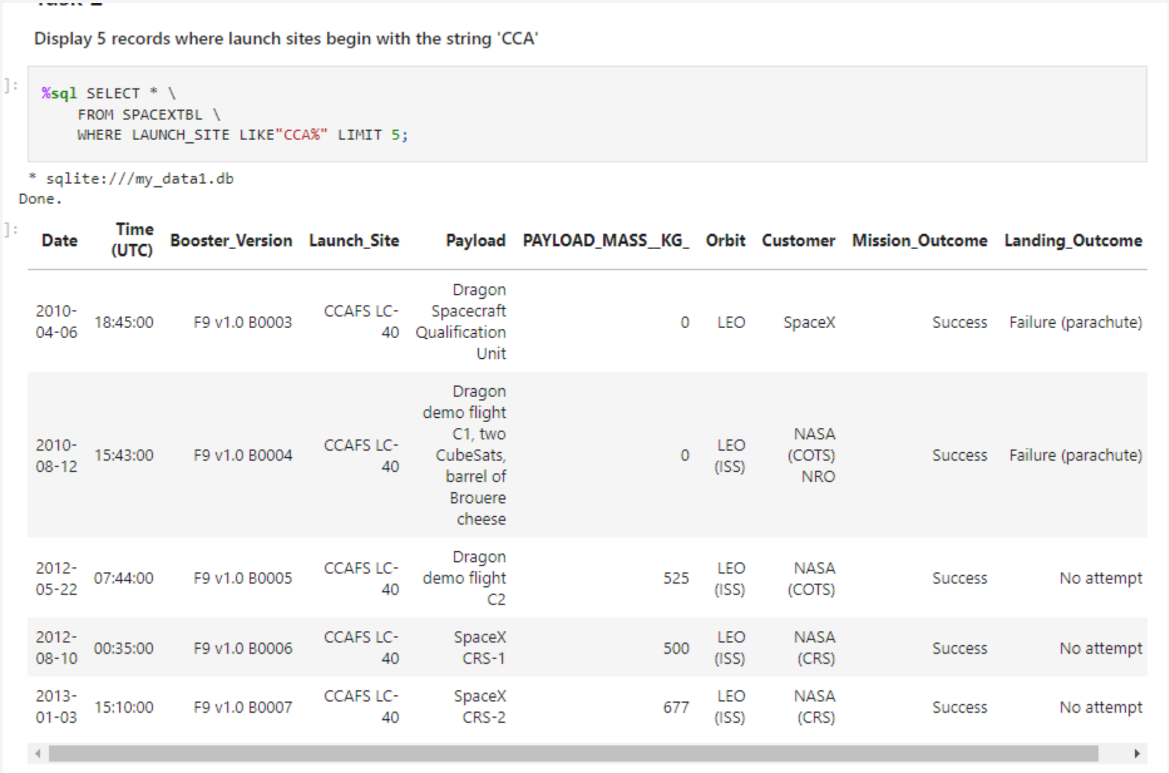
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

```
 * 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'

- The query below was used to display 5 records where launch sites begin with “CCA”.



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

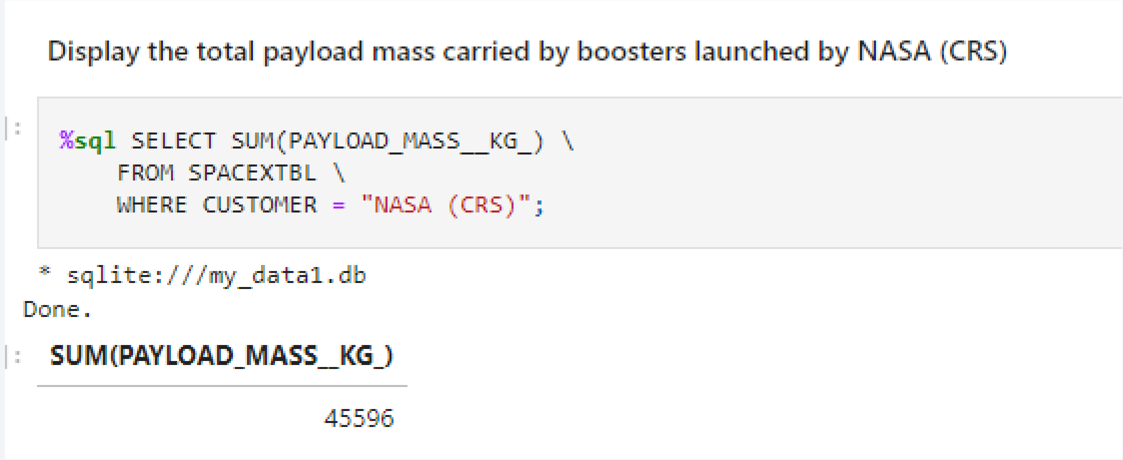
```
%sql SELECT * \
    FROM SPACEXTBL \
    WHERE LAUNCH_SITE LIKE"CCA%" LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
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

Total Payload Mass

- Total payload carried by boosters from NASA was 45596



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

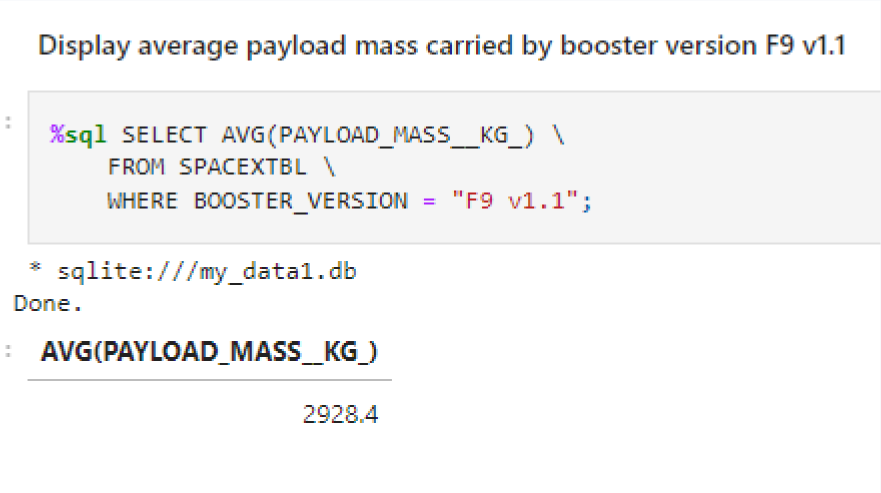
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = "NASA (CRS)";
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 was 2928.4



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

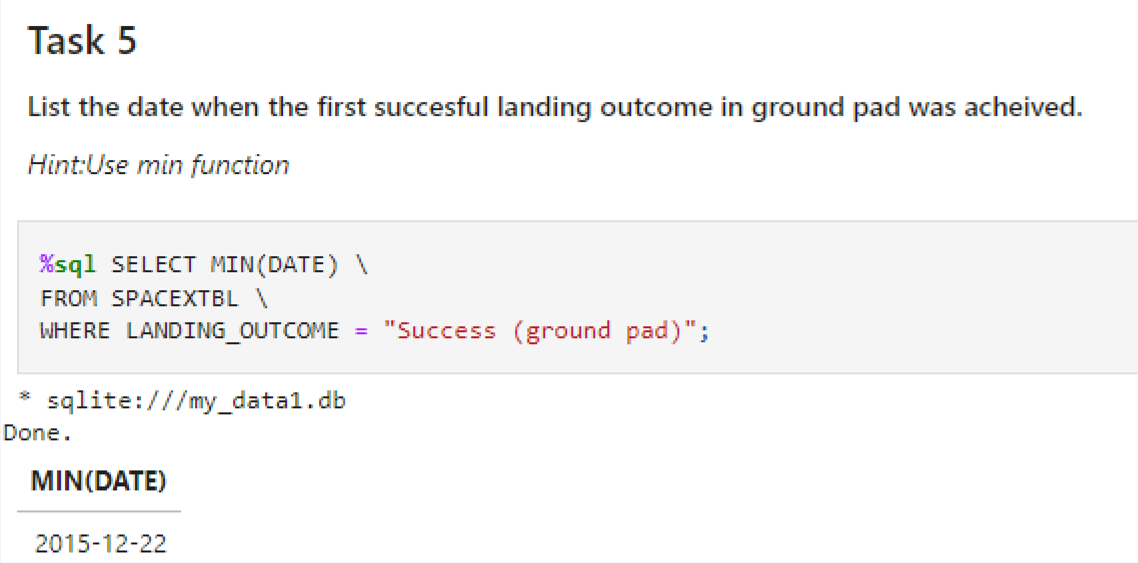
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = "F9 v1.1";
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First successful landing outcome on ground pad was 22nd December 2015



Task 5

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

Hint:Use min function

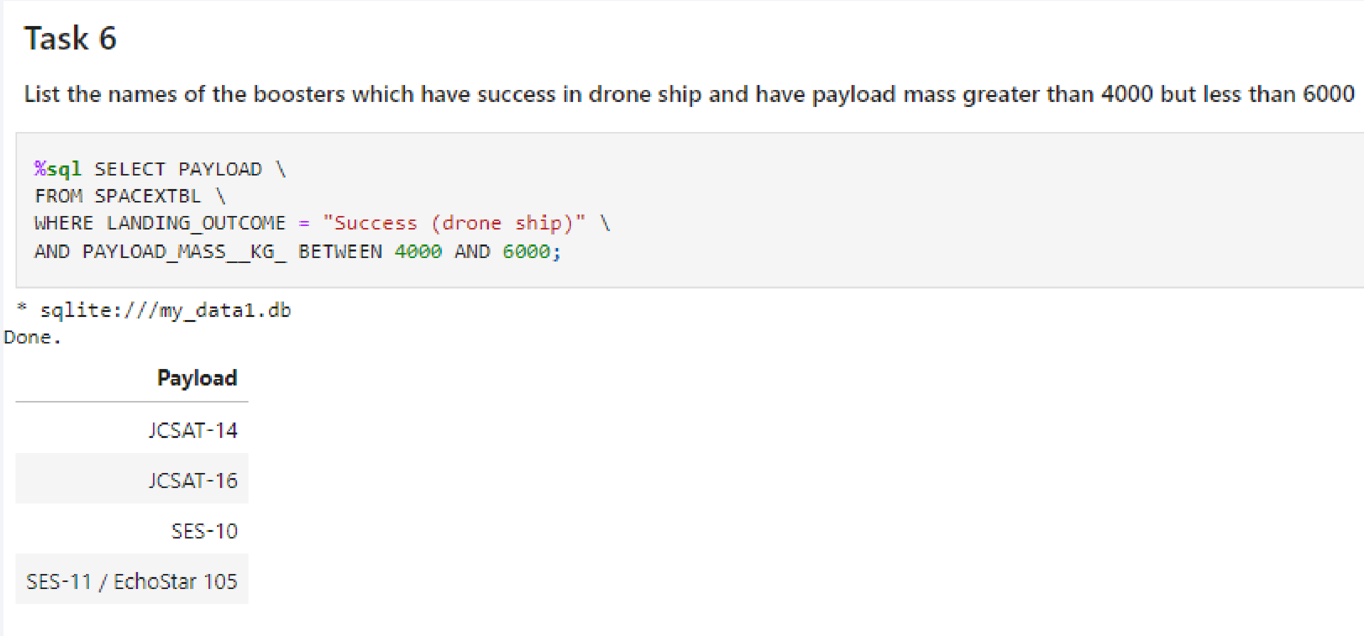
```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = "Success (ground pad)";
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- WHERE clause were used 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



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

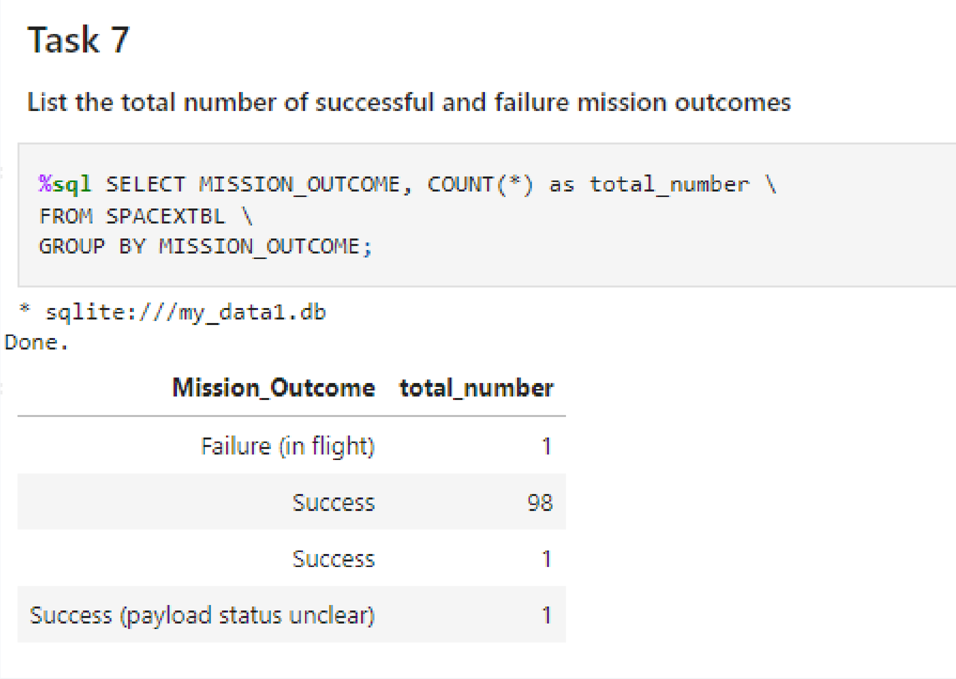
```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = "Success (drone ship)" \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Wildcard like ‘%’ to filter for WHERE MissionOutcome was a success or a failure was used.



Task 7

List the total number of successful and failure mission outcomes

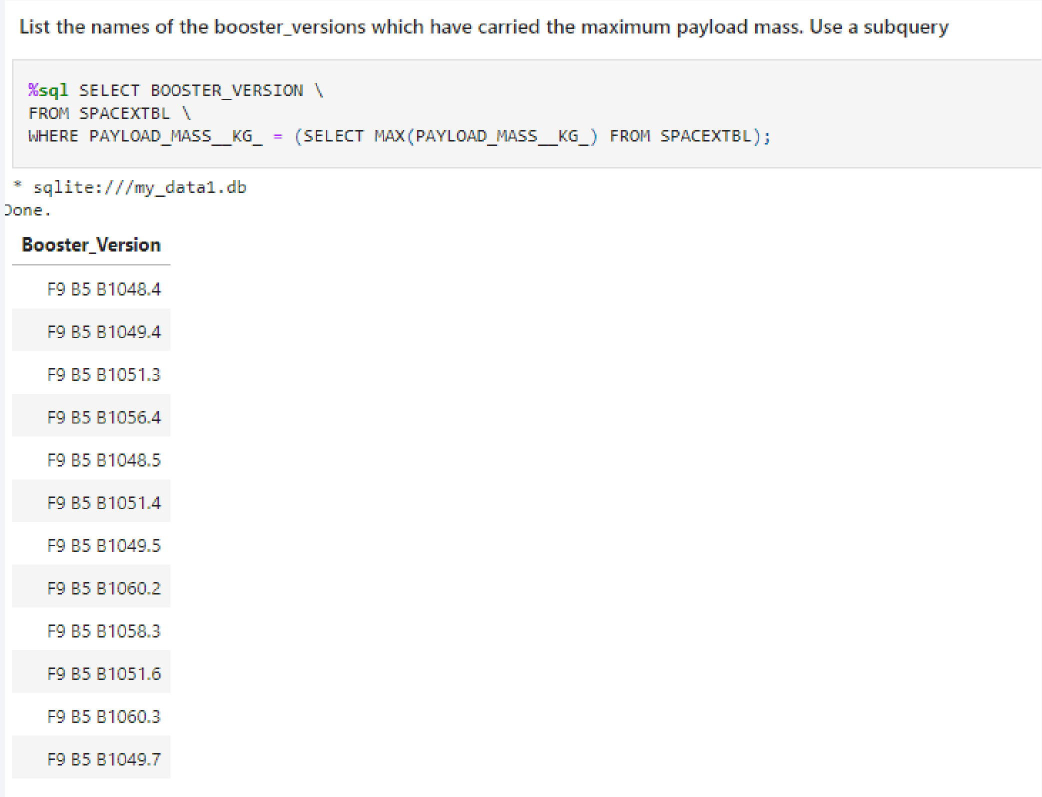
```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- Thebooster that have carried the maximum payload was determined using a subquery in the WHERE clause and the MAX() function.



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

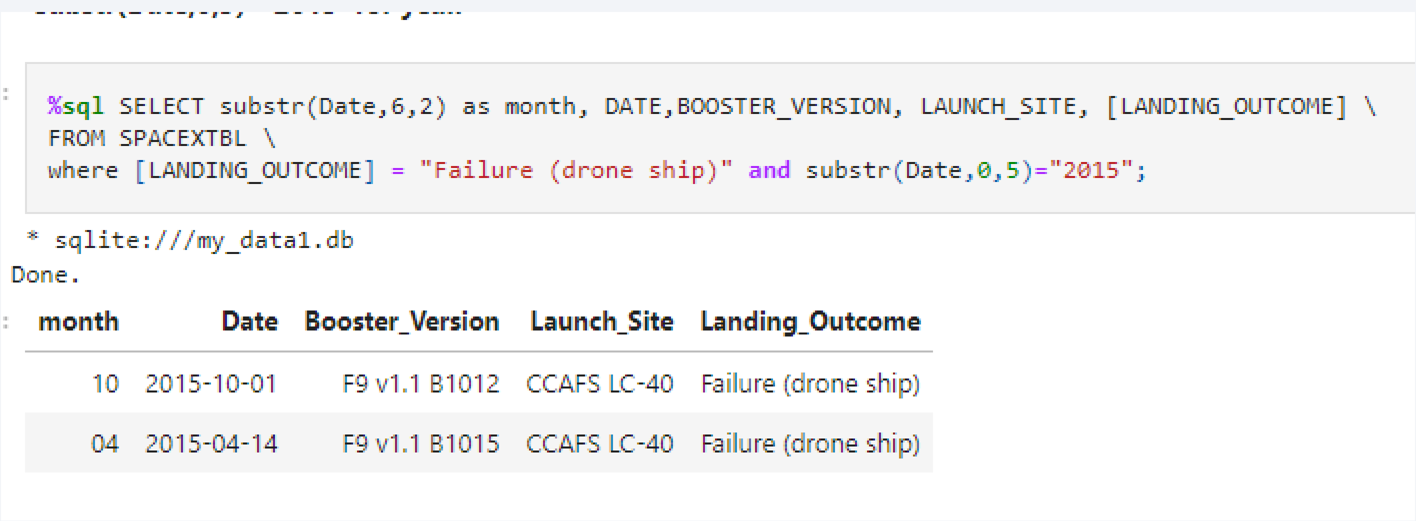
```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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

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



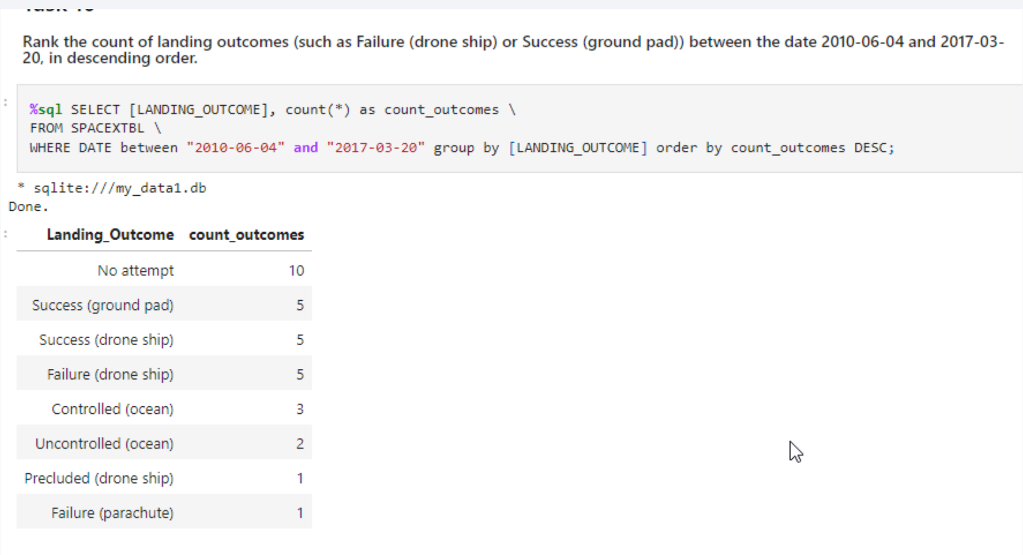
```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [LANDING_OUTCOME] \
FROM SPACEXTBL \
where [LANDING_OUTCOME] = "Failure (drone ship)" and substr(Date,0,5)="2015";
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Landing outcomes and the COUNT of landing outcomes were selected from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.
- The GROUP BY clause was apllied to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.



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.

```
%sql SELECT [LANDING_OUTCOME], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between "2010-06-04" and "2017-03-20" group by [LANDING_OUTCOME] order by count_outcomes DESC;
```

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

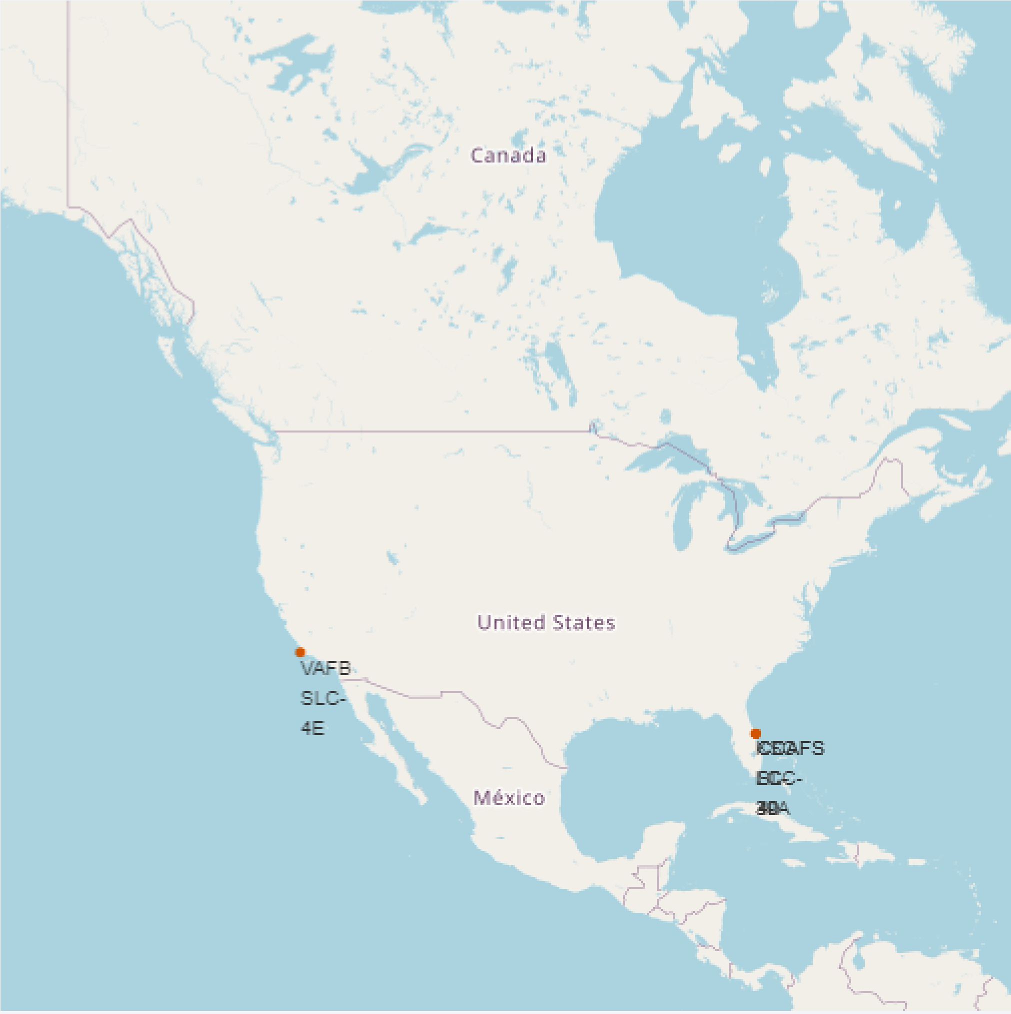
Landing_Outcome	count_outcomes
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

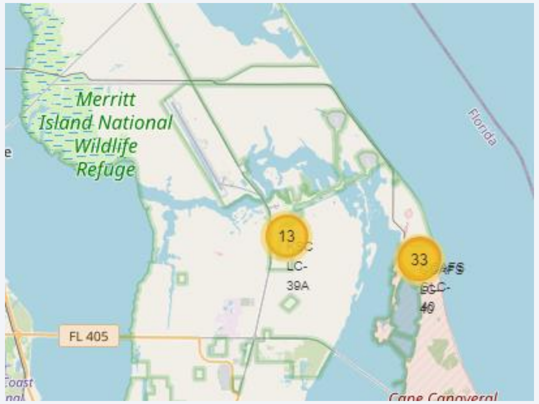
SpaceX Launch Sites

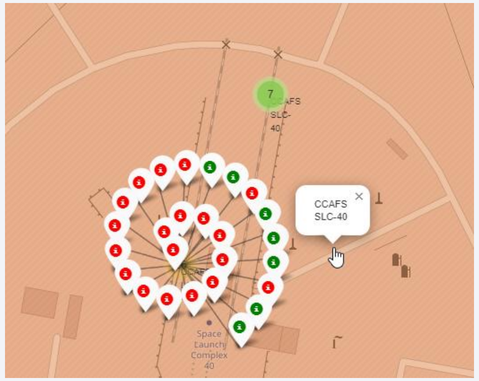
- SpaceX has only launch sites in the United Stated.

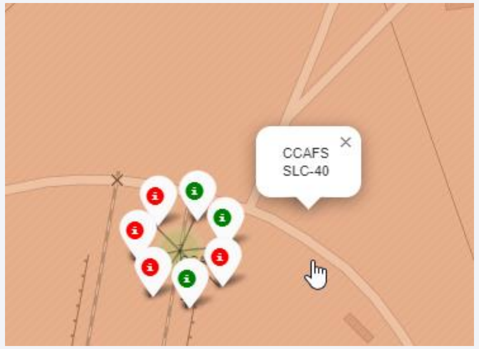


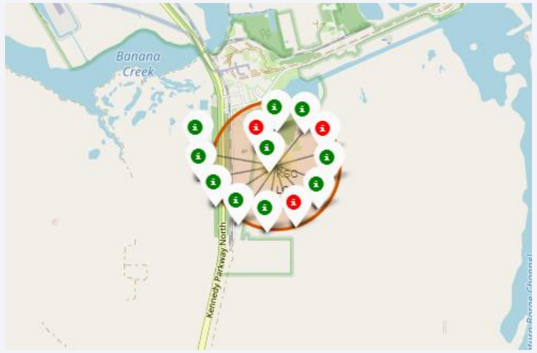
Launch Sites with Markers for success and failures

- Florida Launch Sites

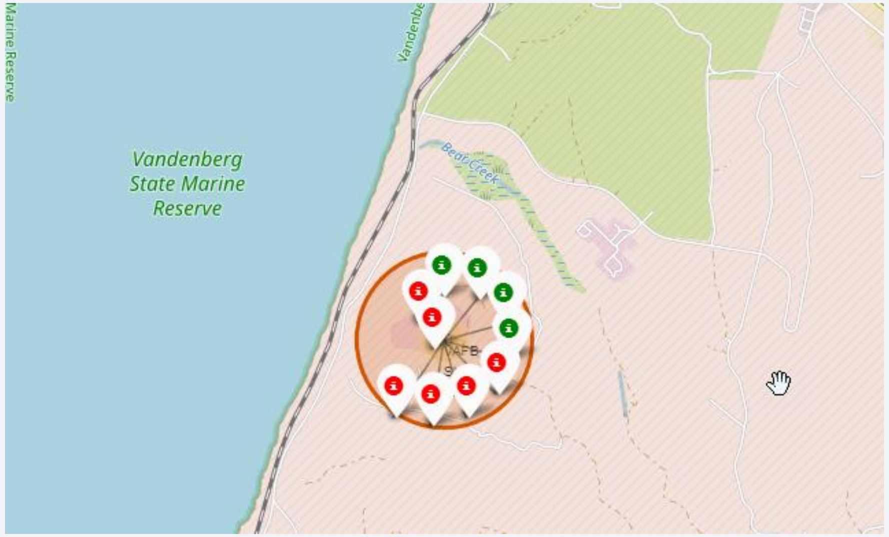








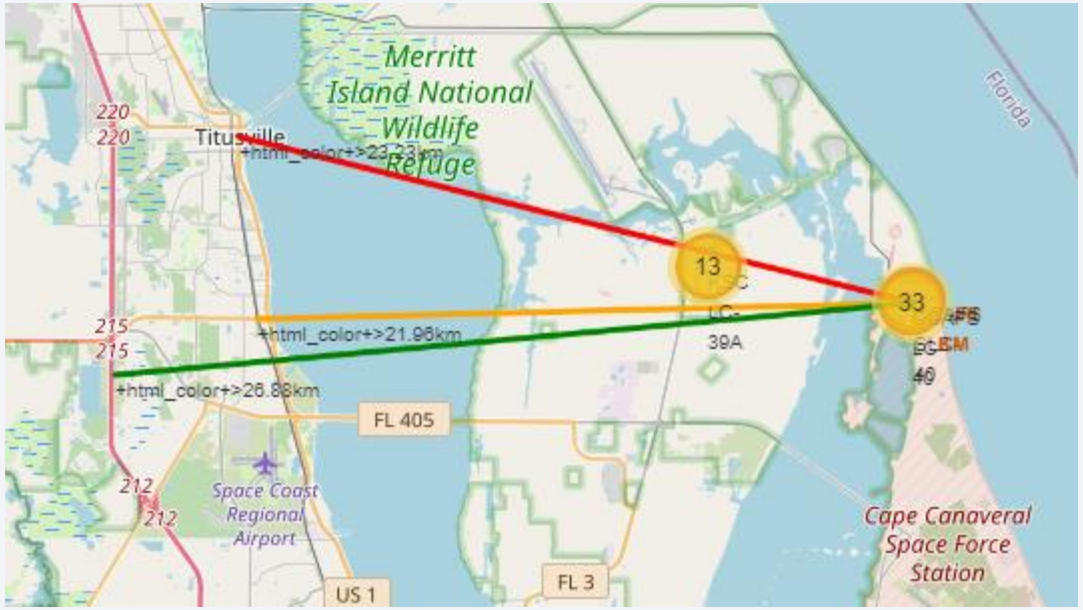
- California Launch Sites

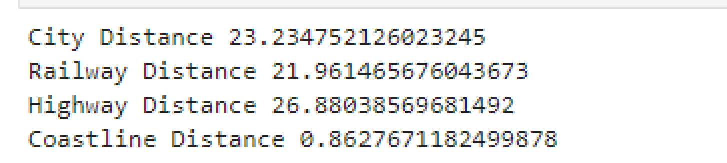


Green markers show successful launches and red markers show failures.

Launch Site distances 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? No
- Do launch sites keep certain distance away from cities? Yes

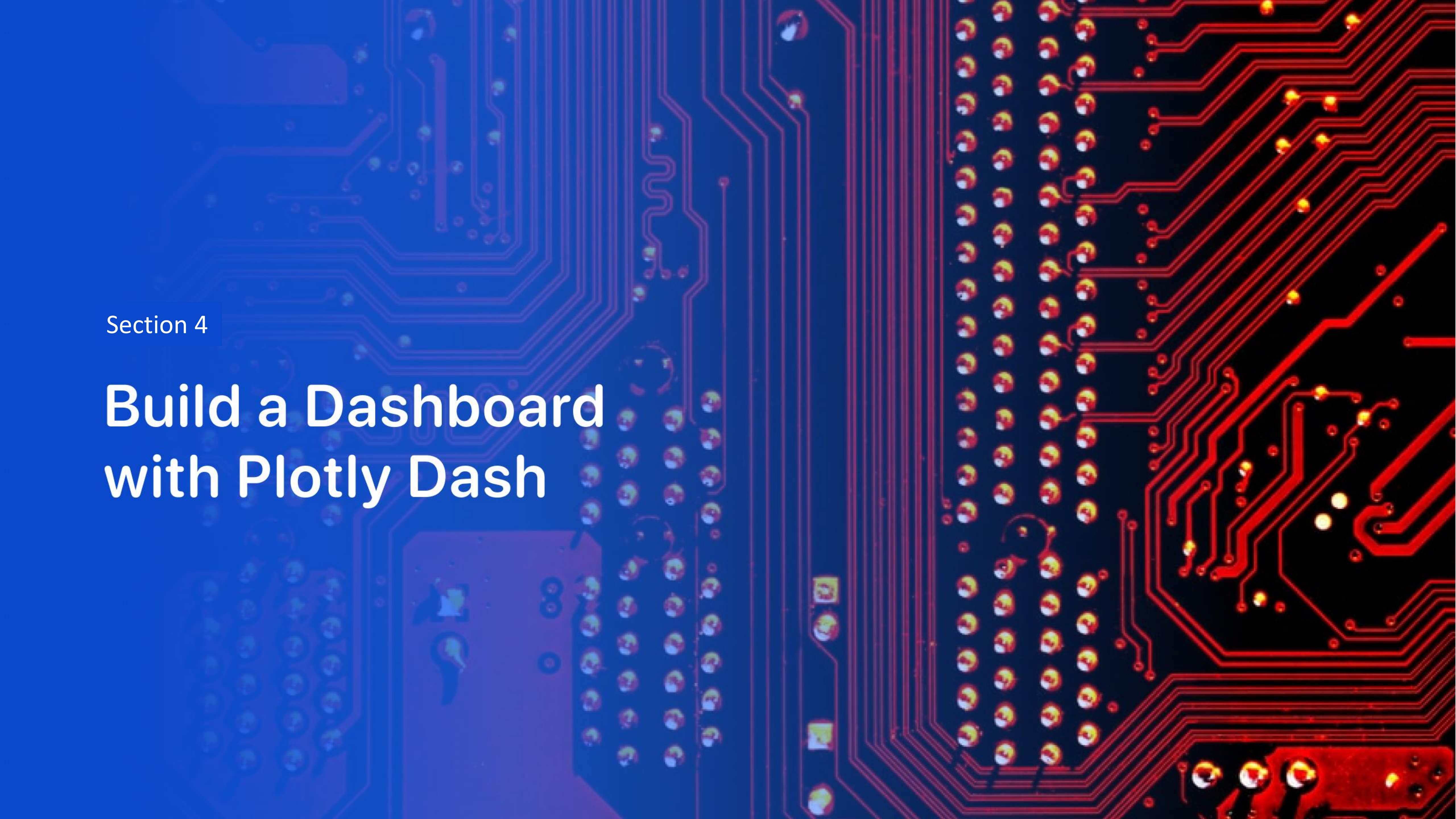




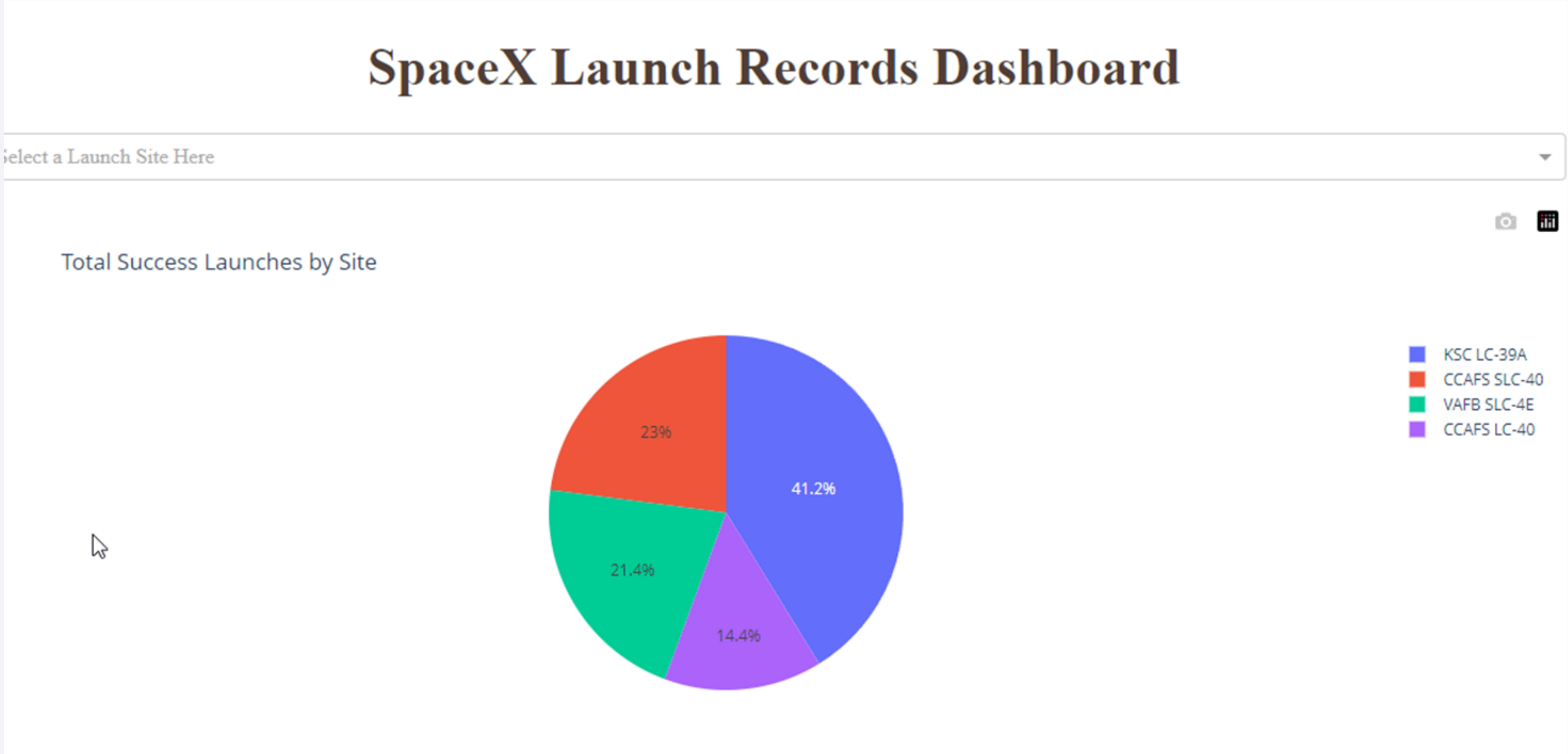
```
City Distance 23.234752126023245
Railway Distance 21.961465676043673
Highway Distance 26.88038569681492
Coastline Distance 0.8627671182499878
```

Section 4

Build a Dashboard with Plotly Dash



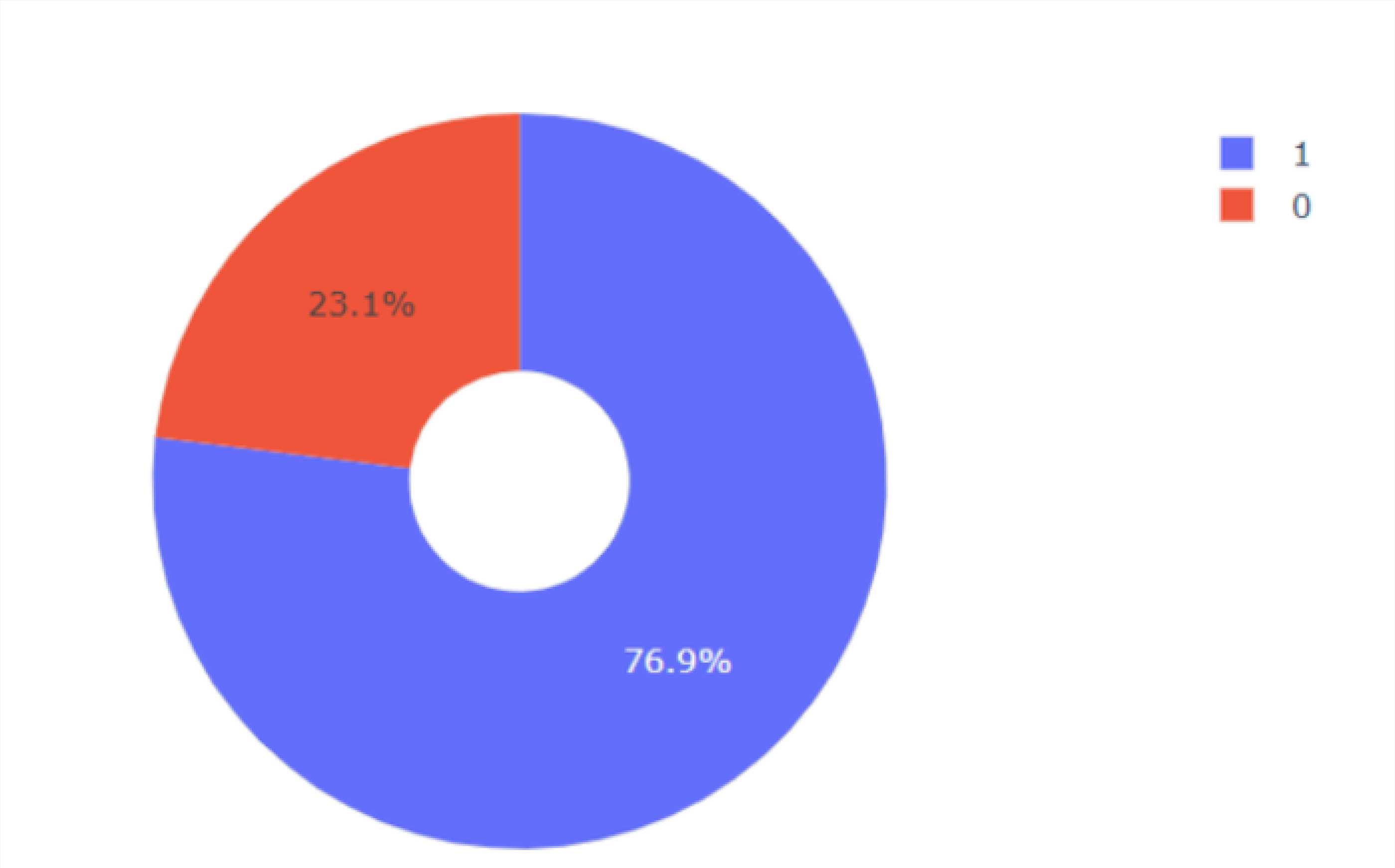
Total Success Launches by Site



SpaceX Launch Records Dashboard

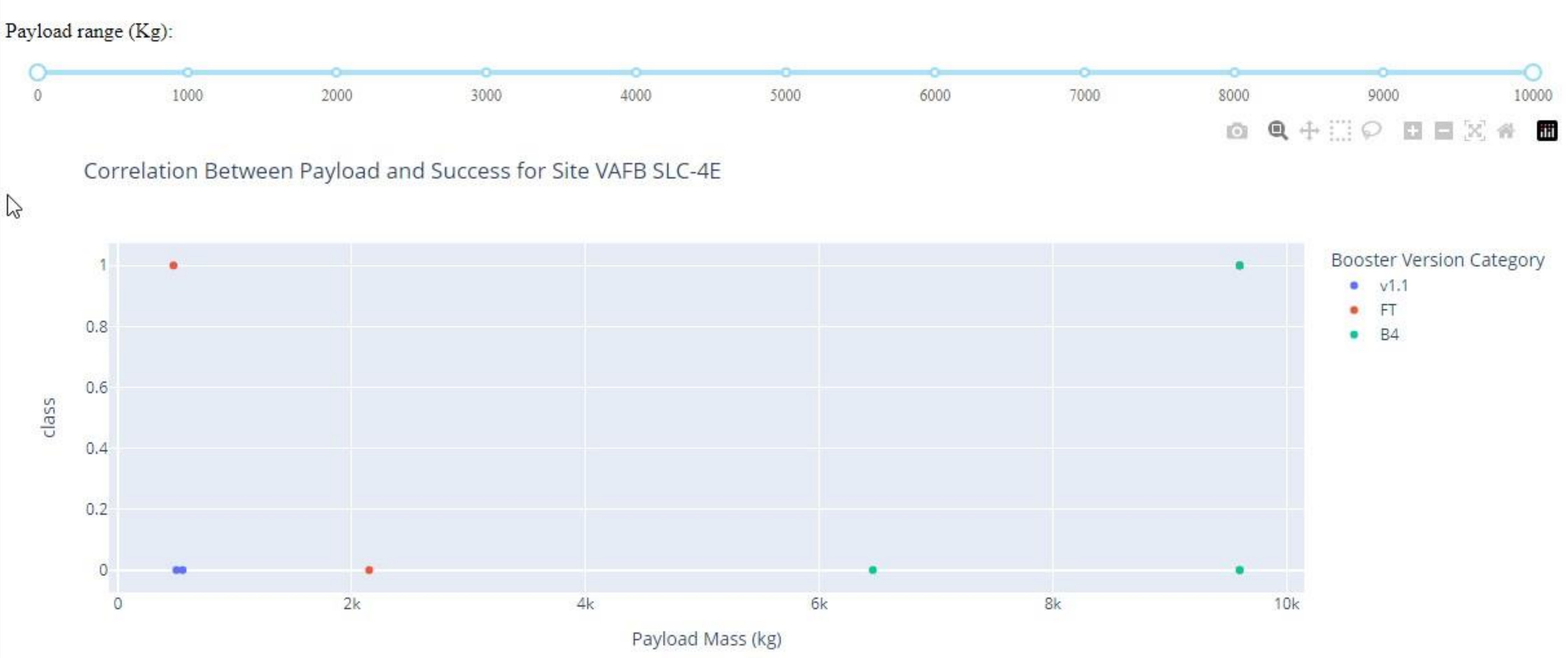
- KSC LC-39A had the highest successful launches

Launch Site with the highest launch success ratio



- KSC LC-39A had a 76,9% success rate.

Payload vs Launch Outcome for VAFB SLC-4E site

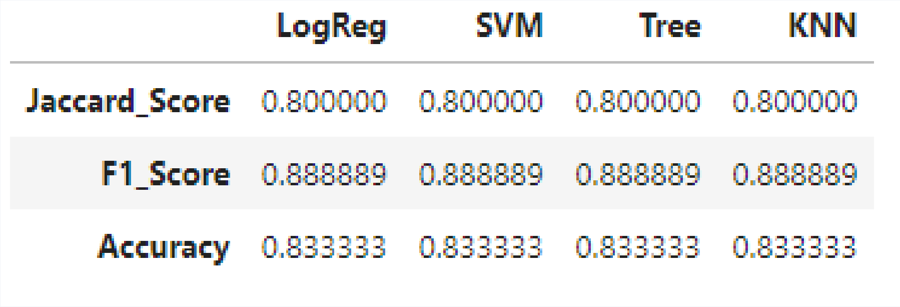


Section 5

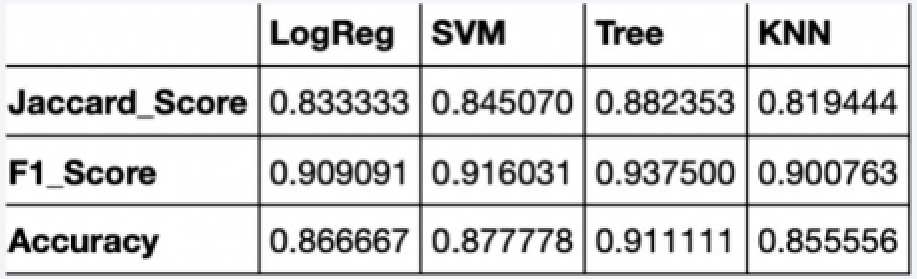
Predictive Analysis (Classification)

Classification Accuracy

- Based on the scores of the test set, all methods have the same performance.
- Same test set scores may be due to the small test sample size. Using the whole dataset, the best model is Decision Tree Model. This Model has higher scores and accuracy.



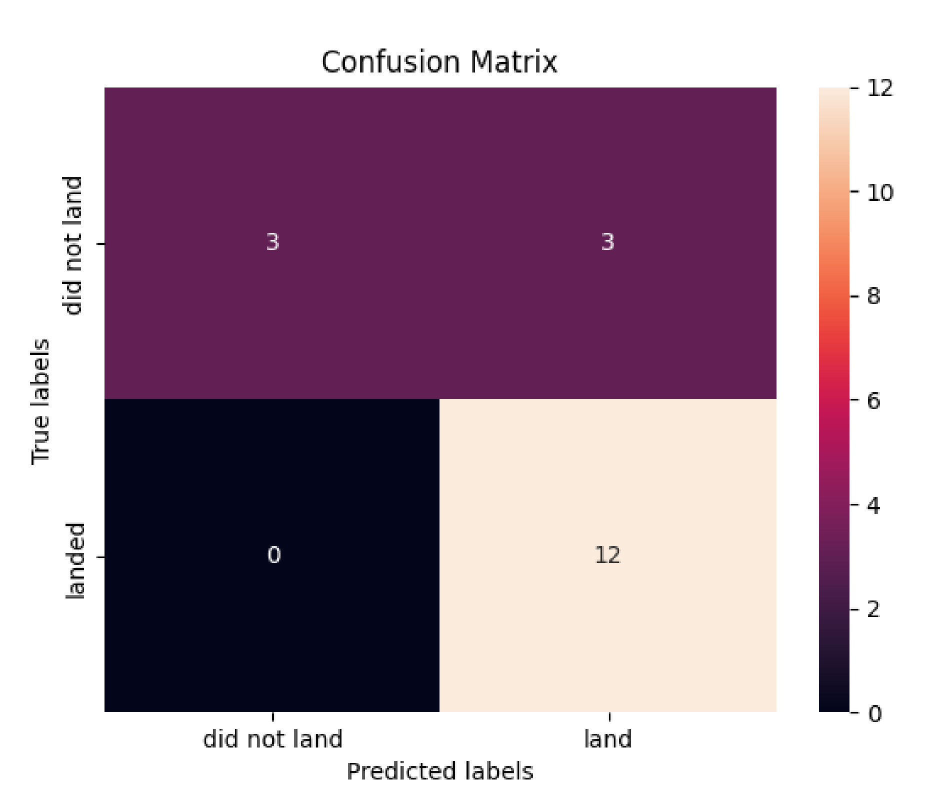
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

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, unsuccessful landing marked as successful landing by the classifier.



Conclusions

- Decision tree classifier is the best machine learning model.
- Launches with low payload mass have better results
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of all launch sites
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- The larger the flight amount at a launch site, the greater the success rate at a launch site.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!