IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Radu-Constantin Pîrva
20.03.2022

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection
  - Data wrangling
  - Exploratory data analysis with data visualization
  - Exploratory data analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)

Executive Summary

- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics in screenshots
  - Predictive analytics results

Introduction

- Project background and context
  - Space X 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 Space X 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 space X for a rocket launch. The goal of this project is to use machine learning techniques in order to be able to predict if the first stage will land succesfuly.

- Problems we want to find answers for:
  - Which factors correlate with the rocket landing successfully?
  - The impact different variables have on the landing outcome.
  - What conditions are needed in order to achieve the best results?

Section 1

Methodology

Methodology – Executive Summary

Data collection methodology:

- SpaceX Rest API
- Web Scrapping

Perform data wrangling:

- One Hot Encoding data fields
- Dropping irrelevant columns
- Dealing with missing values

Methodology – Executive Summary

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

Using scatter and bar graphs to show relationship between variables and to show patterns of data.

Performing interactive visual analytics using Folium and Plotly Dash.

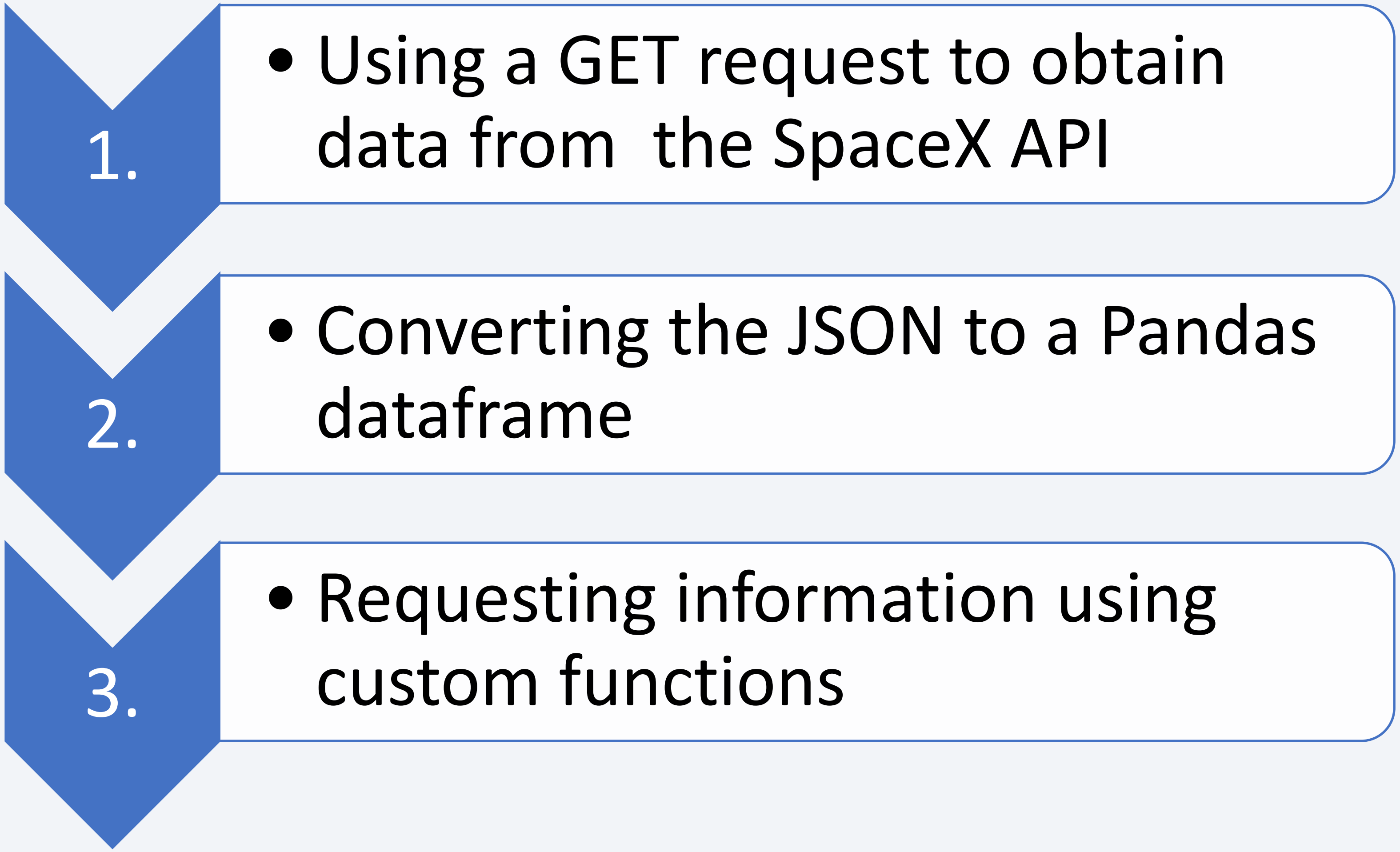
Performing predictive analysis using classification models:

- Logistic Regression
- SVM
- KNN
- Decision tree classifier

Data Collection

Data collection included obtaining data from the SpaceX REST API and Web Scraping from a Wikipedia table. Both of these sources were used in order to get a more complete view.

Data Collection – SpaceX API



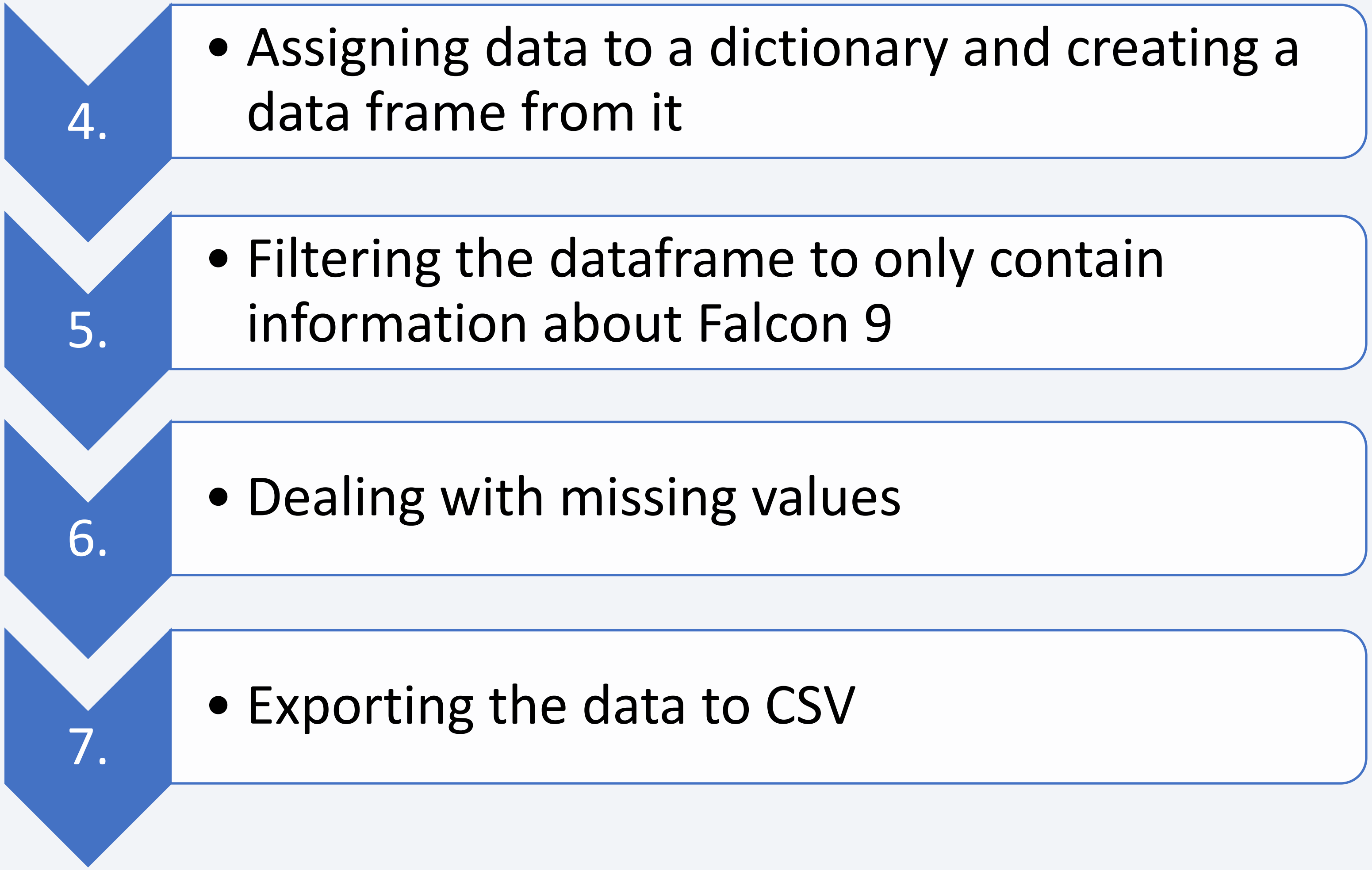
1.
- Using a GET request to obtain data from the SpaceX API

2.
- Converting the JSON to a Pandas dataframe

3.
- Requesting information using custom functions

https://github.com/radu177/Data-Science-Project/blob/master/Data%20Collection%20API.ipynb

Data Collection – SpaceX API



4. Assigning data to a dictionary and creating a data frame from it

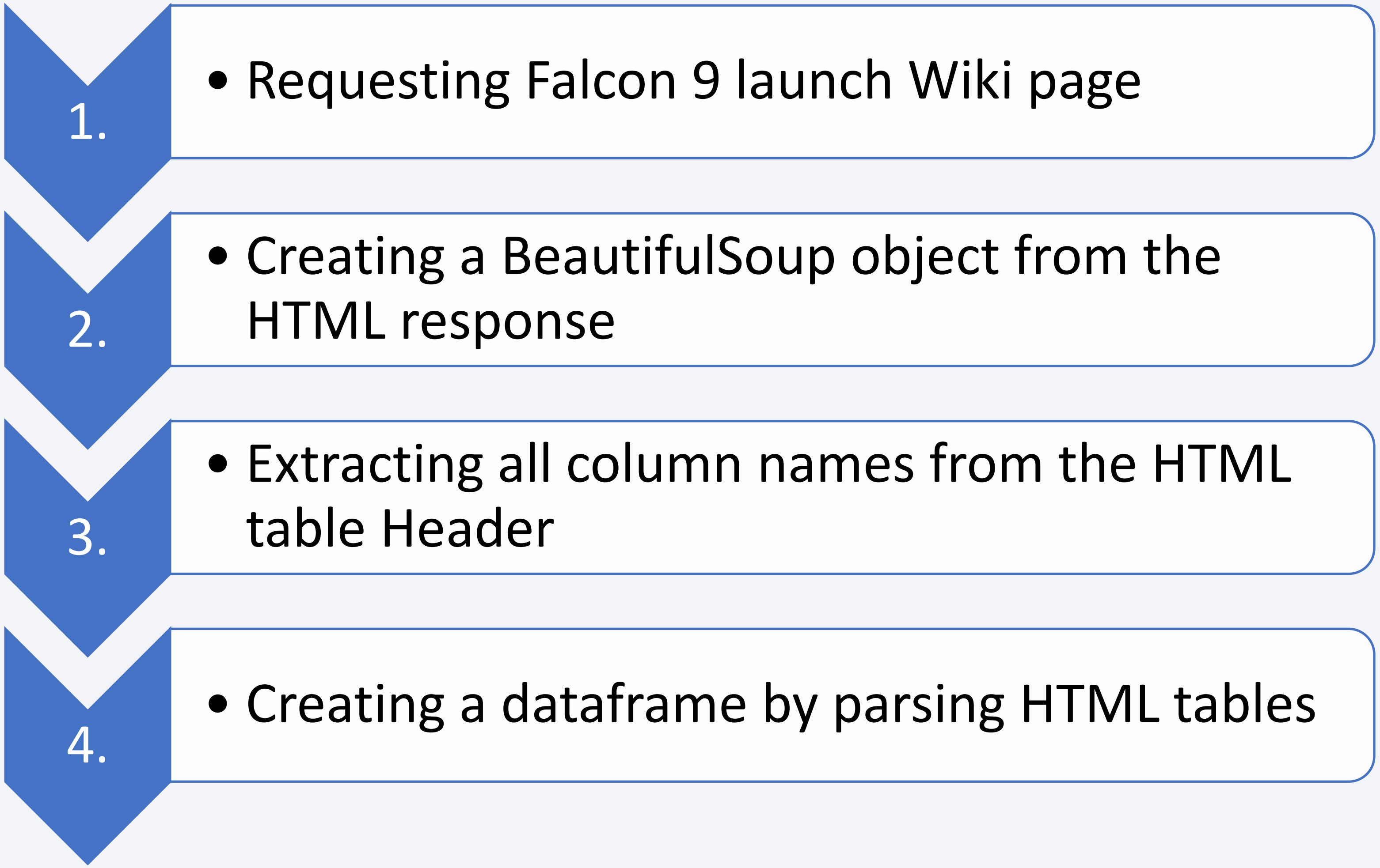
5. Filtering the dataframe to only contain information about Falcon 9

6. Dealing with missing values

7. Exporting the data to CSV

https://github.com/radu177/Data-Science-Project/blob/master/Data%20Collection%20API.ipynb

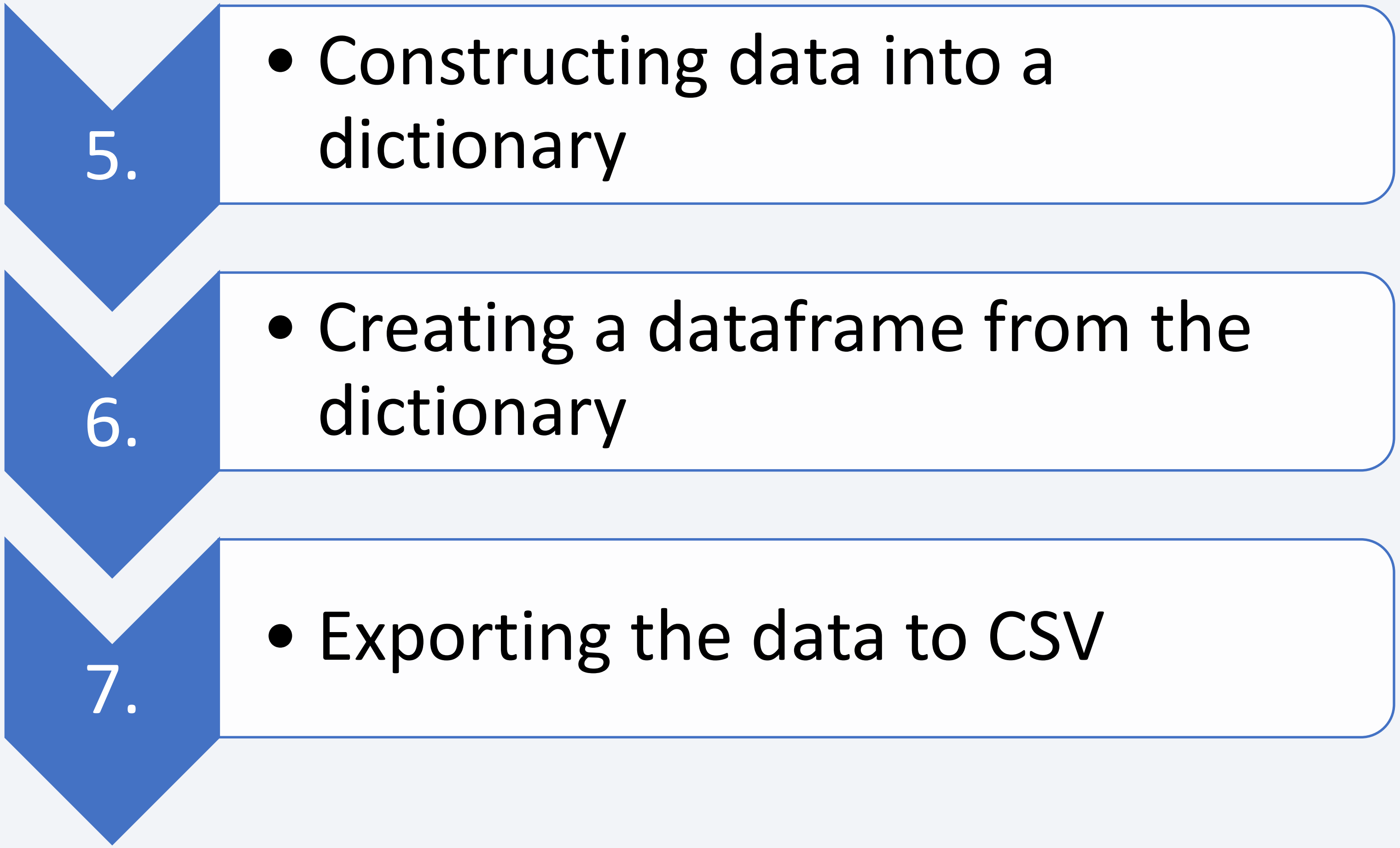
Data Collection - Scraping



1.
   - Requesting Falcon 9 launch Wiki page
2.
   - Creating a BeautifulSoup object from the HTML response
3.
   - Extracting all column names from the HTML table Header
4.
   - Creating a dataframe by parsing HTML tables

https://github.com/radu177/Data-Science-Project/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Collection - Scraping



5.

- Constructing data into a dictionary

6.

- Creating a dataframe from the dictionary

7.

- Exporting the data to CSV

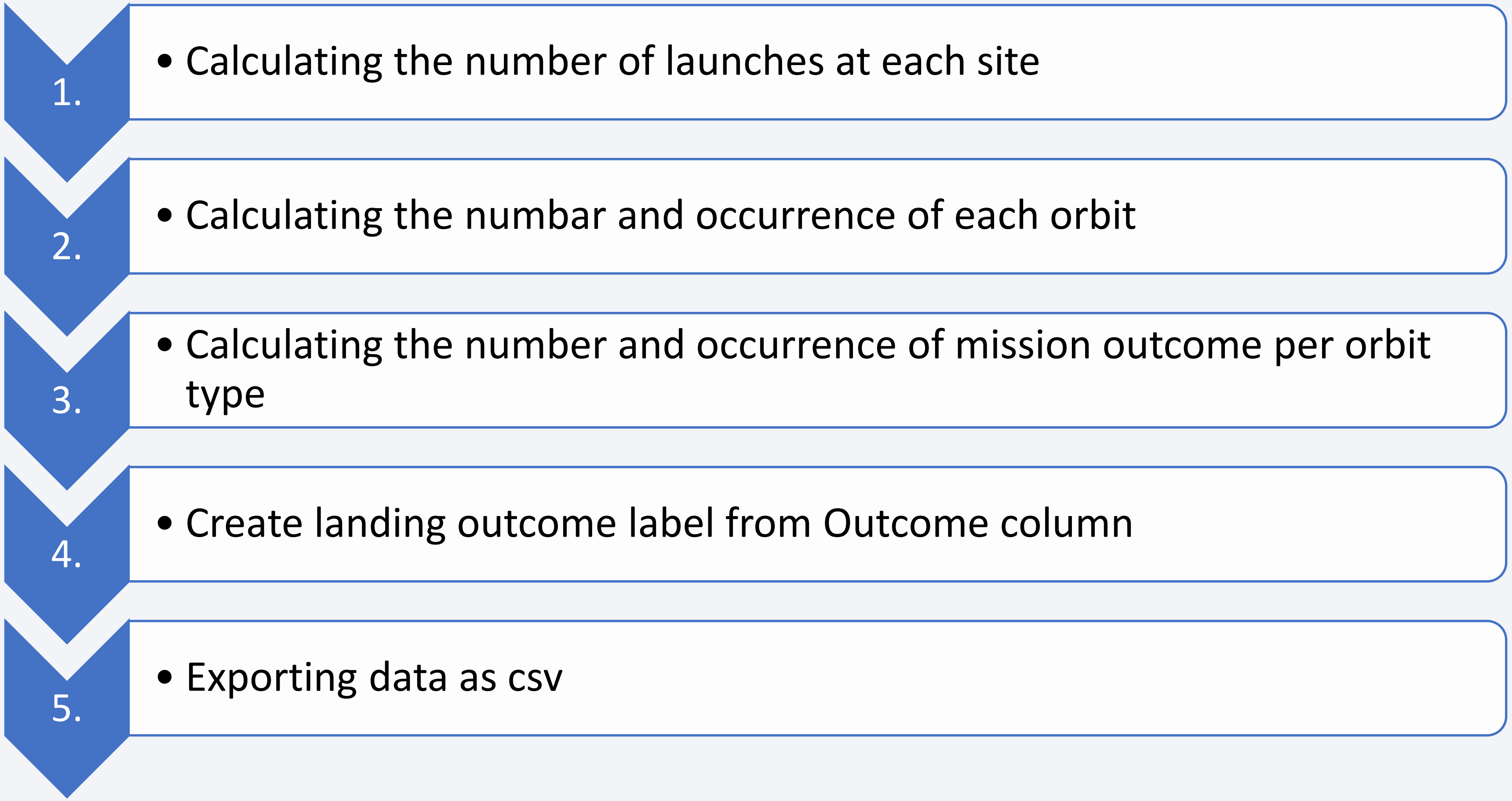
https://github.com/radu177/Data-Science-Project/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Exploratory Data Analysis (EDA) was performed in order to find some patterns in the data and determine what would be the label for training supervised models.
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.
- The outcomes were converted into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

https://github.com/radu177/Data-Science-Project/blob/master/Data%20Wrangling.ipynb

Data Wrangling



1.
   - Calculating the number of launches at each site
2.
   - Calculating the numbar and occurrence of each orbit
3.
   - Calculating the number and occurrence of mission outcome per orbit type
4.
   - Create landing outcome label from Outcome column
5.
   - Exporting data as csv

https://github.com/radu177/Data-Science-Project/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

Multiple plot types were used:

A **scatterplot** shows the relationship between two quantitative variables.

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Flight Number vs. Orbit Type
- Payload vs. Orbit type

A **bar chart** shows the relationship between a numeric and a categoric variable. Each entity of the categoric variable is represented as a bar. The size of the bar represents its numeric value.

- Orbit Type and Success Rate

A **line chart** shows the trend of a variable over time.

- Success Rate VS. Year

https://github.com/radu177/Data-Science-Project/blob/master/EDA%20Visualization.ipynb

EDA with SQL

SQL queries that were performed:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was acheived.

https://github.com/radu177/Data-Science-Project/blob/master/EDA.ipynb

EDA with SQL

SQL queries that were performed:

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass
- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranking 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

https://github.com/radu177/Data-Science-Project/blob/master/EDA.ipynb

Build an Interactive Map with Folium

Using Folium, the launch data was visualised in an interactive map.

Using the coordinates of each launch site a **Circle Marker** was added around them and labeled.

The launch outcomes were represented by green or red markers on a map using a **MarkerCluster**.

By using Haversine’s formula the distances from the launch site to various landmarks were calculated in order to find trends about the surrounding area. Lines were drawn on the map using **PolyLine** in order to represent the distances.

https://github.com/radu177/Data-Science-Project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

The dashboard app contains a pie chart and a scatter chart.

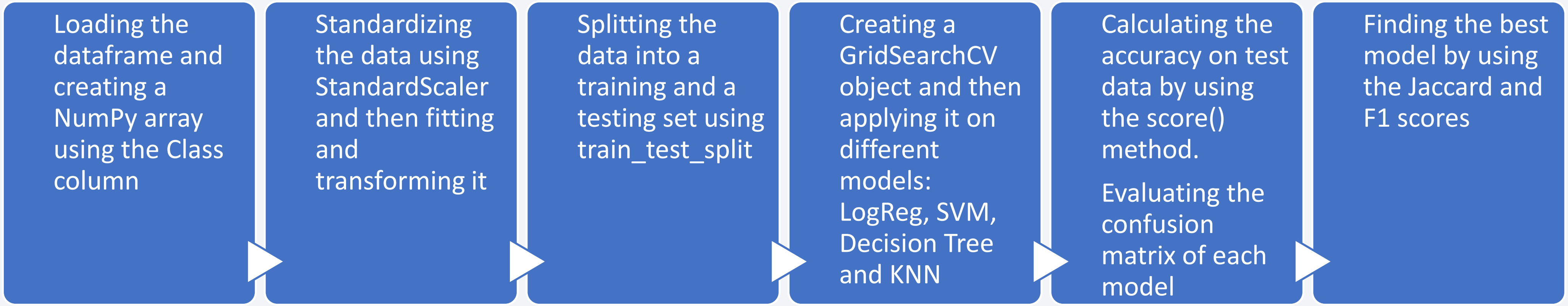
Pie chart: by selecting from a dropdown menu it can show the total success percentage for all sites or for a specific launch site.

Scatter chart: shows the relationship between outcomes and payload mass. It has slider between 0 and 10000 which allows you to select a range of values for the payload mass to be displayed. The chart also shows the booster version.

https://github.com/radu177/Data-Science-Project/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

The main steps are building, evaluating, improving the model and then finding the best performing one.



Loading the dataframe and creating a NumPy array using the Class column

Standardizing the data using StandardScaler and then fitting and transforming it

Splitting the data into a training and a testing set using train_test_split

Creating a GridSearchCV object and then applying it on different models: LogReg, SVM, Decision Tree and KNN

Calculating the accuracy on test data by using the score() method.

Evaluating the confusion matrix of each model

Finding the best model by using the Jaccard and F1 scores

https://github.com/radu177/Data-Science-Project/blob/master/Machine%20Learning%20Prediction.ipynb

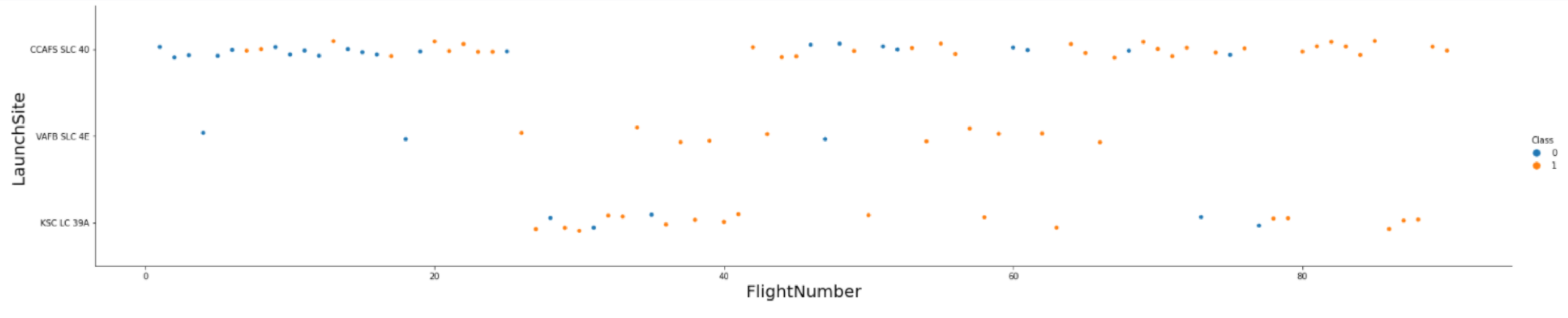
Results

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

Section 2

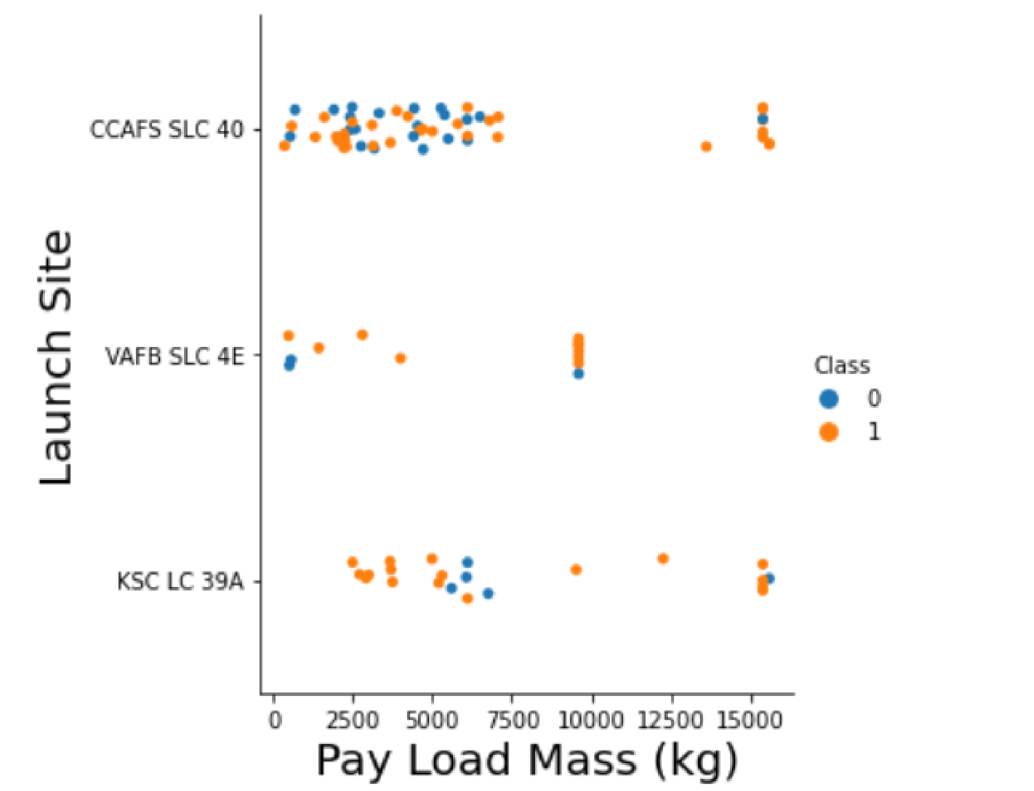
Insights drawn from EDA

Flight Number vs. Launch Site



- This scatter plot suggests that as the flight number increases the success rate also increases.
- Launch site CCAFS SLC-40 has the highest number of launches.

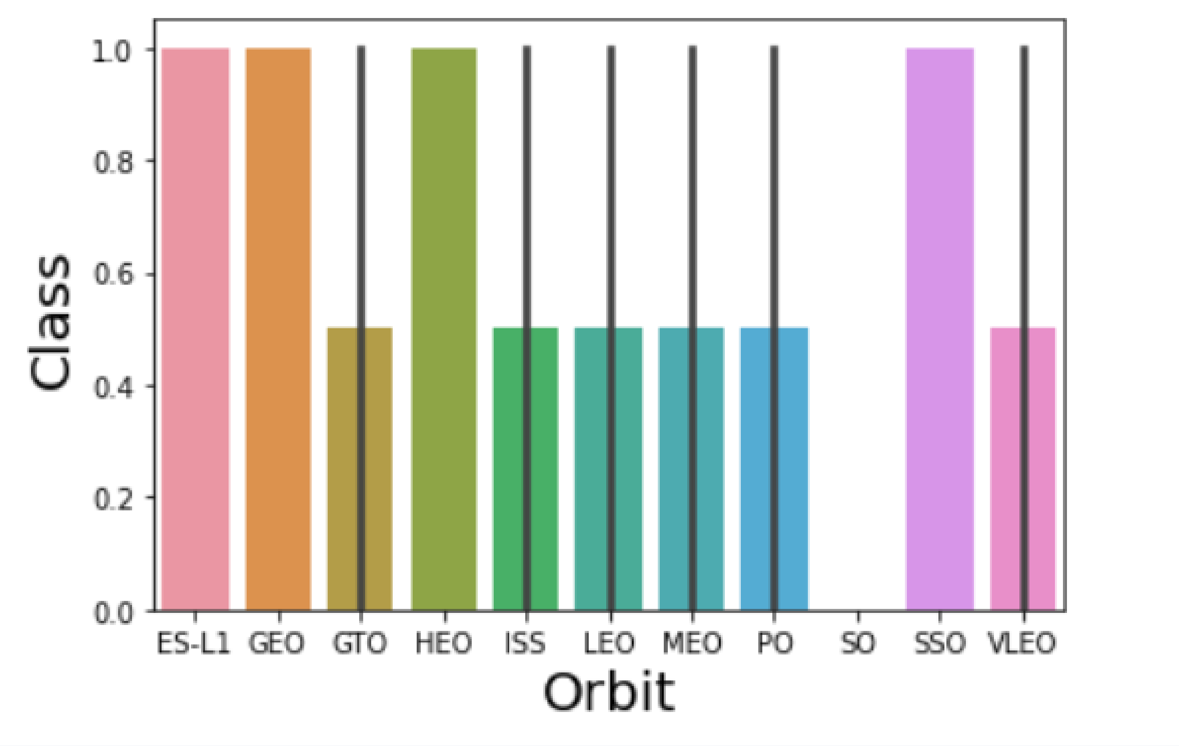
Payload vs. Launch Site



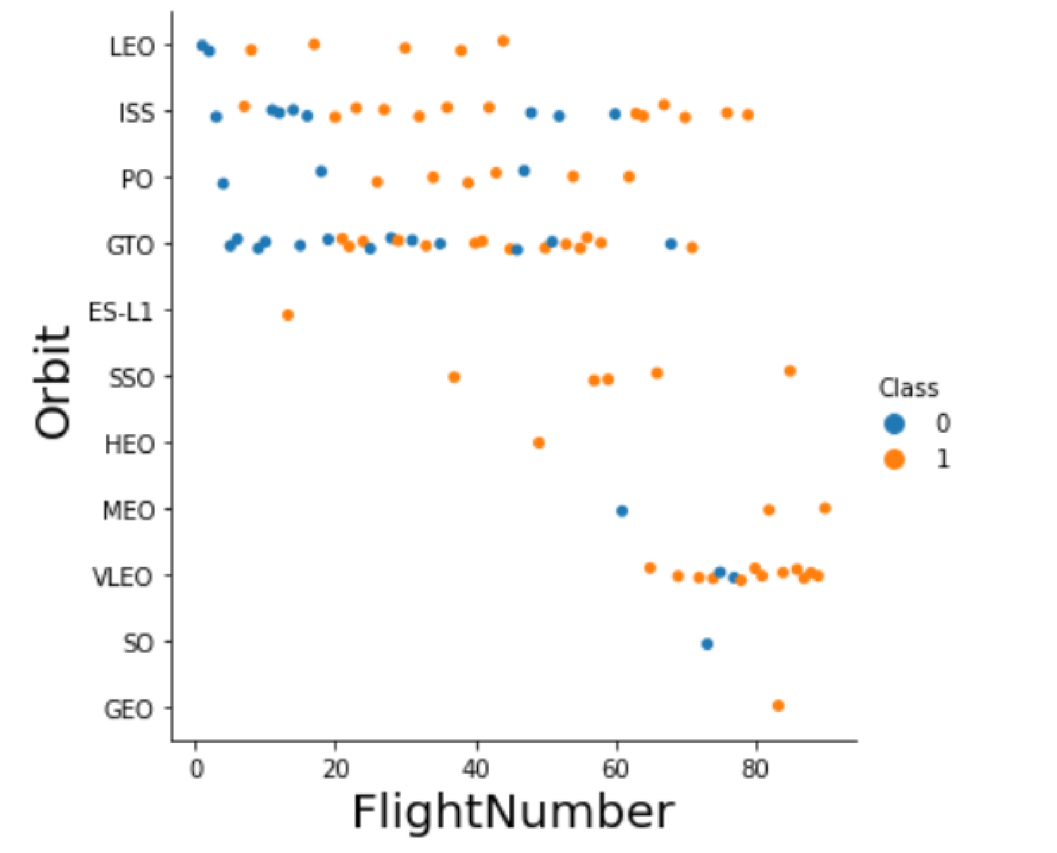
- In this graph we can observe that with higher payload mass there seems to be an increase in successful landings, especially at launch site CCAFS SLC 40.
- Most launches with a payload of over 7500 kg were successful.
- At launch site KSC LC 39A all launches with a payload of less than 5000 kg were successful.

Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, SSO have a 100% success rate.
- Orbit SO has a 0% success rate.
- Other orbits have a success rate of around 50%.

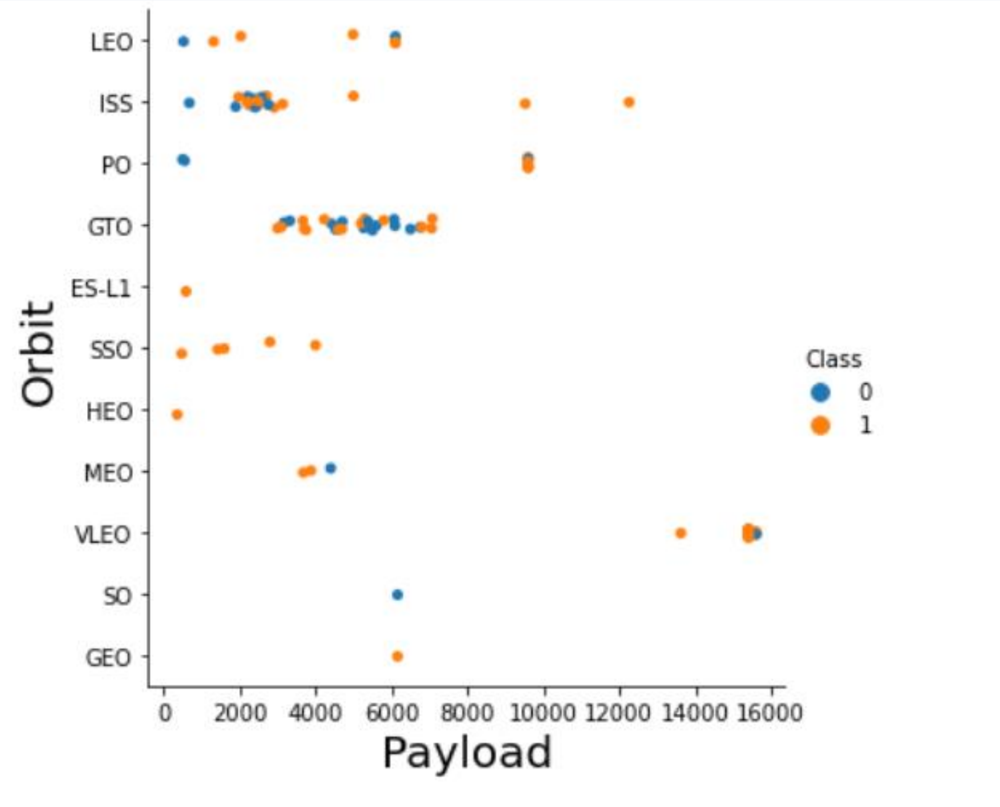


Flight Number vs. Orbit Type



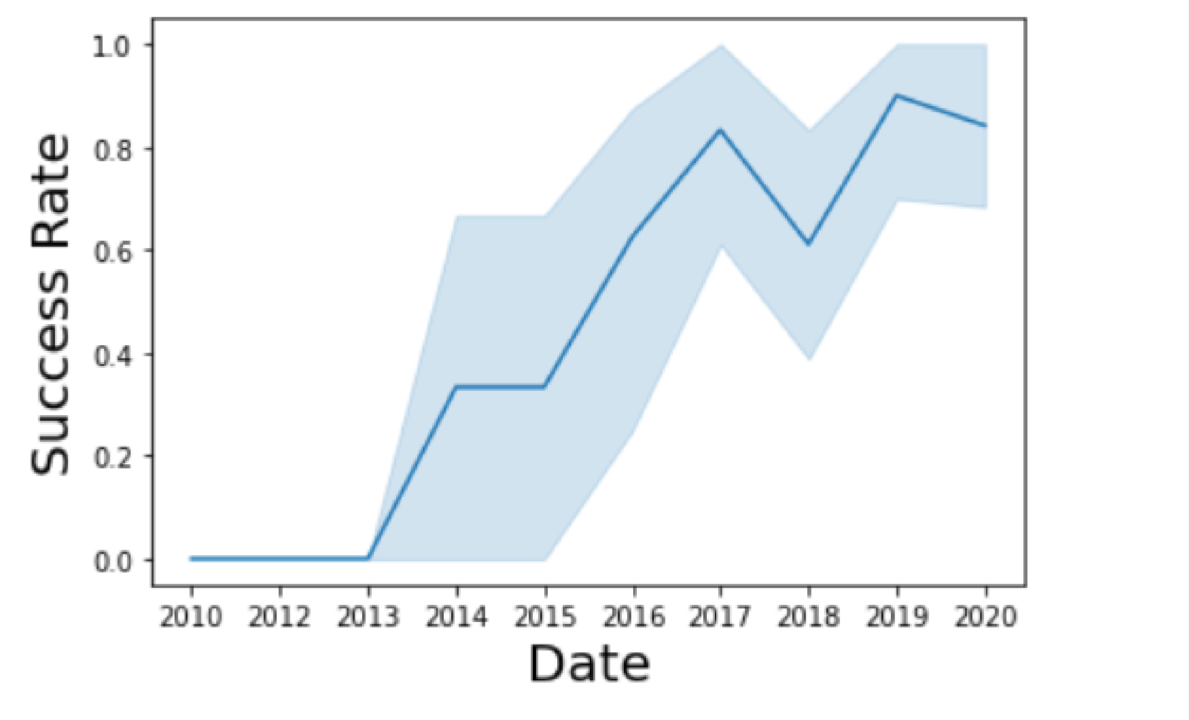
- A trend towards higher success rates can be seen in multiple orbits with increasing flight numbers, especially for the LEO orbit.
- All launches in the SSO orbit were successful.

Payload vs. Orbit Type



- Heavy payloads seem to have had a negative influence for VLEO and MEO orbits and a positive influence on the succes rate of the ISS and PO.

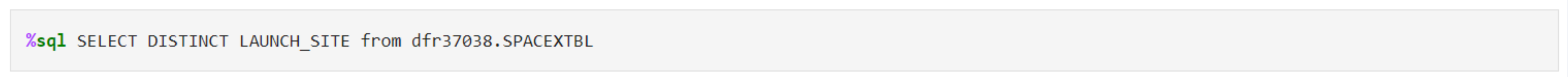
Launch Success Yearly Trend



The success rate has increased from 2013 to 2017 after which it had a drop to about 0.6 in 2018, but which later increased again for 2019. A small drop in success rate is seen in 2020, but the overall trend of the graph suggests an increase in success rate over time.

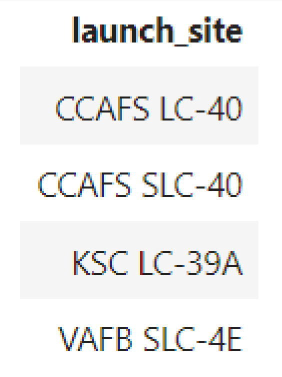
All Launch Site Names

- Finding the names of the unique launch sites. The sql query used:



```
%sql SELECT DISTINCT LAUNCH_SITE from dfr37038.SPACEXTBL
```

- Query results:

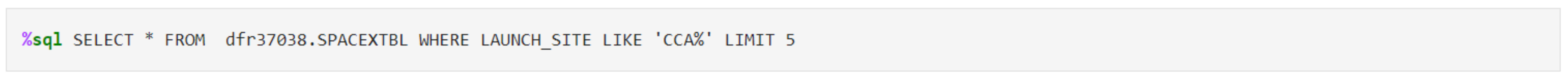


launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- Using the word **DISTINCT** in the query will return unique values from the LAUNCH_SITE column.

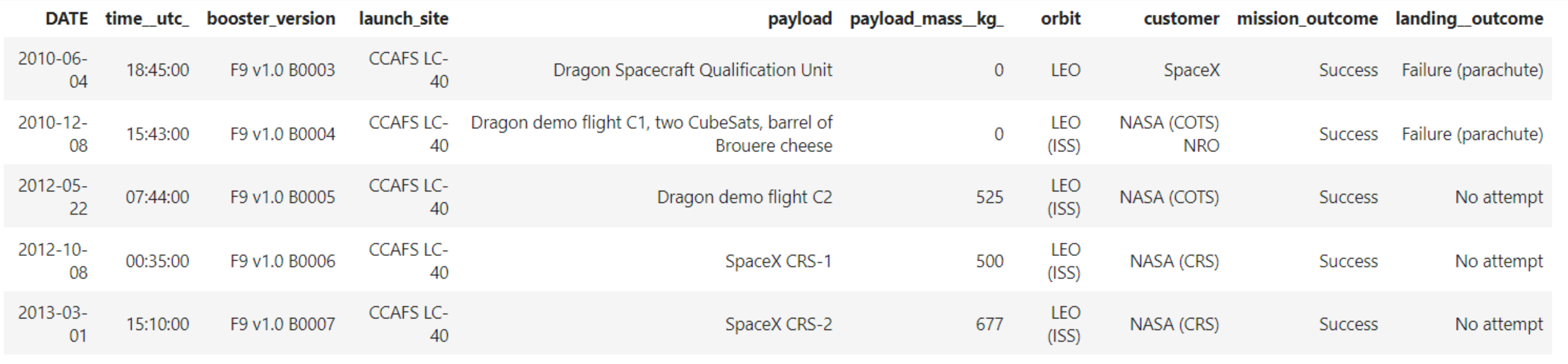
Launch Site Names Begin with 'CCA'

- Finding 5 records where launch sites begin with `CCA`. The query used:



```
%sql SELECT * FROM  dfr37038.SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

- Query result:

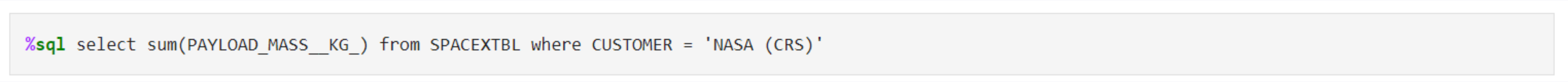


DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Using the keywords **LIMIT 5** in the query will return 5 records from table. Then, the condition **LIKE** was used with a wild card “CCA%” – this will filter the results to show only names that begin with ‘CCA’.

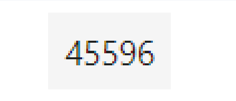
Total Payload Mass

- Calculating the total payload carried by boosters from NASA. The query used:



```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

- Result:

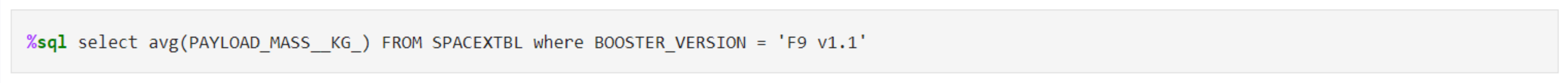


```
45596
```

- The **SUM** keyword adds up the values in the selected column. **WHERE** filters the data(here it is set to return only the rows in which the customer is ‘NASA (CRS)’).

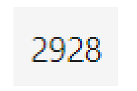
Average Payload Mass by F9 v1.1

- Calculating the average payload mass carried by booster version F9 v1.1. The query used:



```
%sql select avg(PAYLOAD_MASS__KG_) FROM SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

- Result:

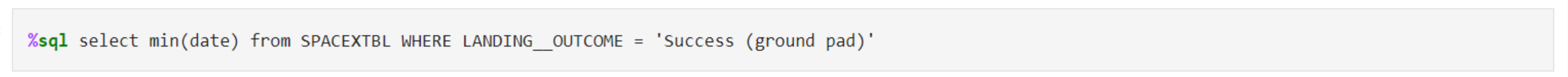


```
2928
```

- The **AVG** function calculates the average of the values from the selected column. In this case the values used from the column were filtered by the **WHERE** clause to include only the rows with a booster version F9 v1.1.

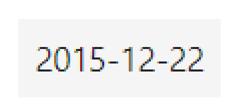
First Successful Ground Landing Date

- Finding the dates of the first successful landing outcome on ground pad. The query used:



```
%sql select min(date) from SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

- The result:

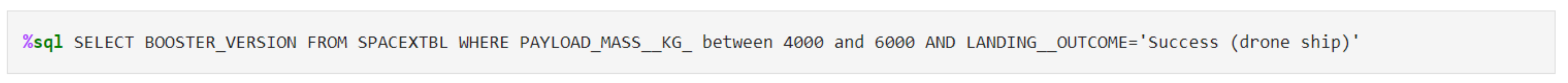


```
2015-12-22
```

- The **MIN** function finds the minimum value from the selected column and the **WHERE** clause filters the data to only include the rows with a successful landing outcome.

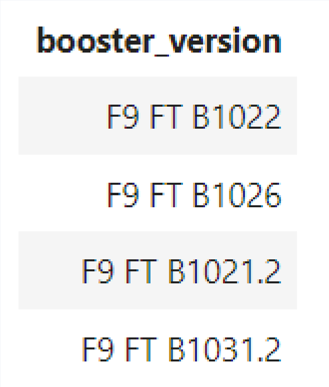
Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. The query used:



```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING__OUTCOME='Success (drone ship)'
```

- The result:

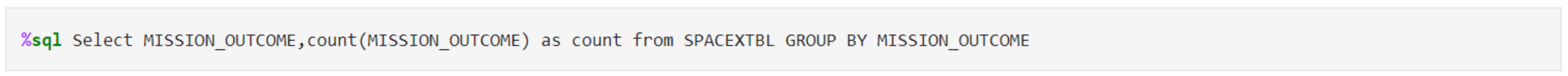


booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

In this case the **WHERE** clause was used with multiple filters linked together by the **AND** keyword.
The **BETWEEN** 4000 and 6000 condition was used to specify an interval of values.

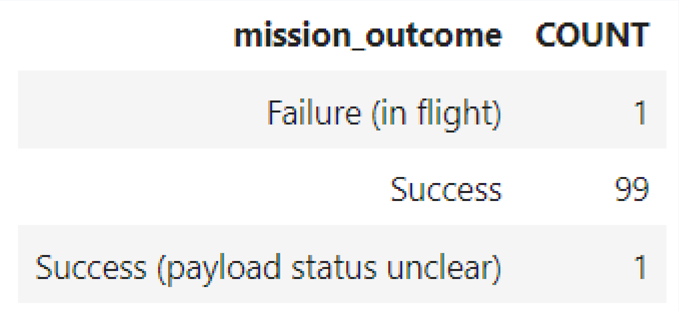
Total Number of Successful and Failure Mission Outcomes

- Calculating the total number of successful and failure mission outcomes. The query used:



```
%sql Select MISSION_OUTCOME,count(MISSION_OUTCOME) as count from SPACEXTBL GROUP BY MISSION_OUTCOME
```

- The result:

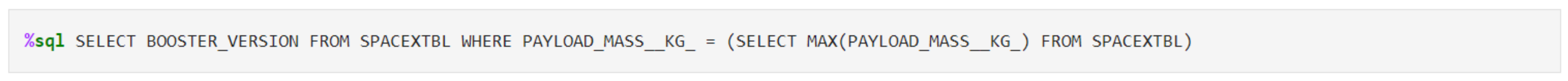


mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

The data was grouped by the outcome using **GROUP BY** and then each group was counted by using the **COUNT** function.

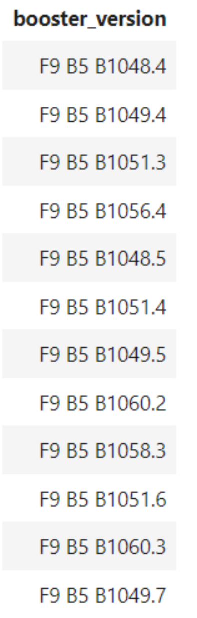
Boosters Carried Maximum Payload

- Listing the names of the boosters which have carried the maximum payload mass. The query used:



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

- Result:

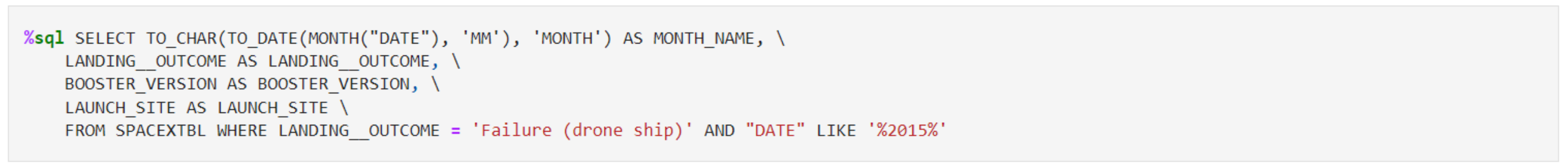


booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

The function **MAX** finds the maximum payload in the column PAYLOAD_MASS_KG. Using these values, the clause **WHERE** filters the results to only include the rows with the maximum payload mass.

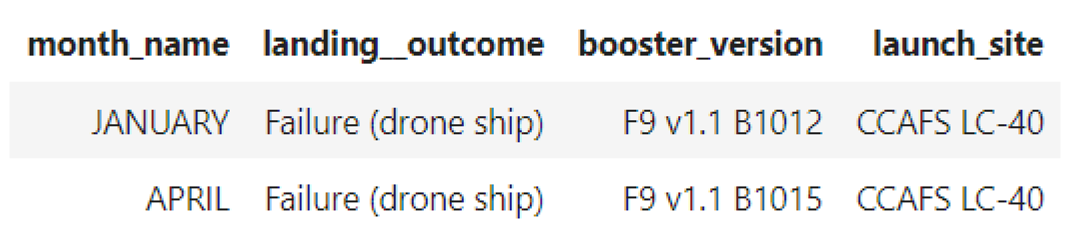
2015 Launch Records

- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015. The query used:



```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, \
    LANDING__OUTCOME AS LANDING__OUTCOME, \
    BOOSTER_VERSION AS BOOSTER_VERSION, \
    LAUNCH_SITE AS LAUNCH_SITE \
    FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND "DATE" LIKE '%2015%'
```

- Results:

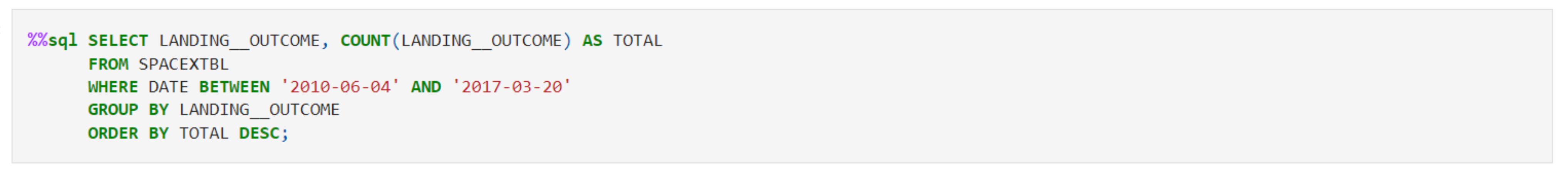


month_name	landing__outcome	booster_version	launch_site
JANUARY	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APRIL	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Two conditions were used for the **WHERE** clause: a date **LIKE** ‘%2015%’ and an outcome of ‘Failure (drone ship)’.

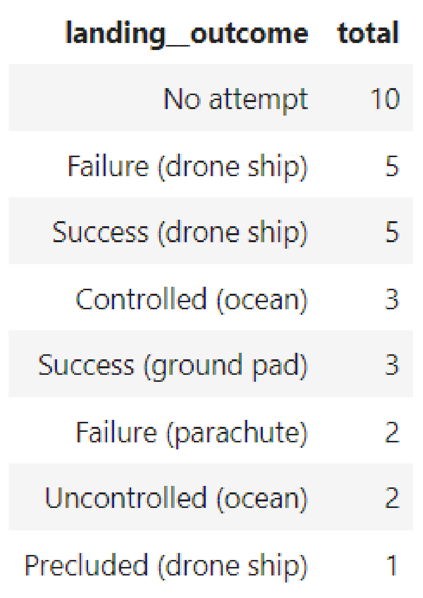
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:



```
%%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL
      FROM SPACEXTBL
      WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
      GROUP BY LANDING__OUTCOME
      ORDER BY TOTAL DESC;
```

- Result:



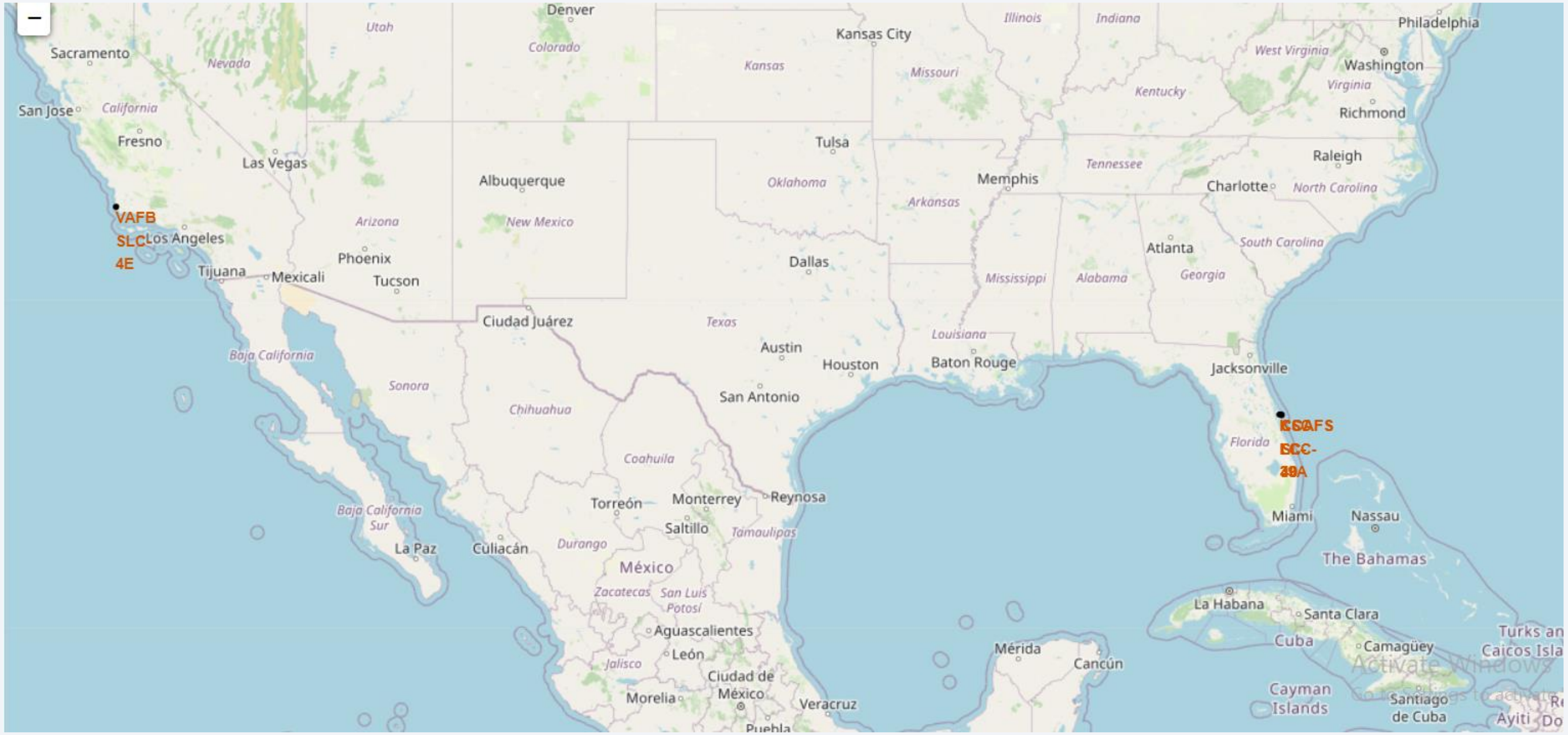
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

This query uses the function **COUNT** in order to count the records in the column LANDING__OUTCOME filtered from a certain time interval by using the **WHERE** clause and that are grouped by the LANDING_OUTCOME.
The **ORDER BY** clause and then the **DESC** keyword used at the end places the results in descending order.

Section 3

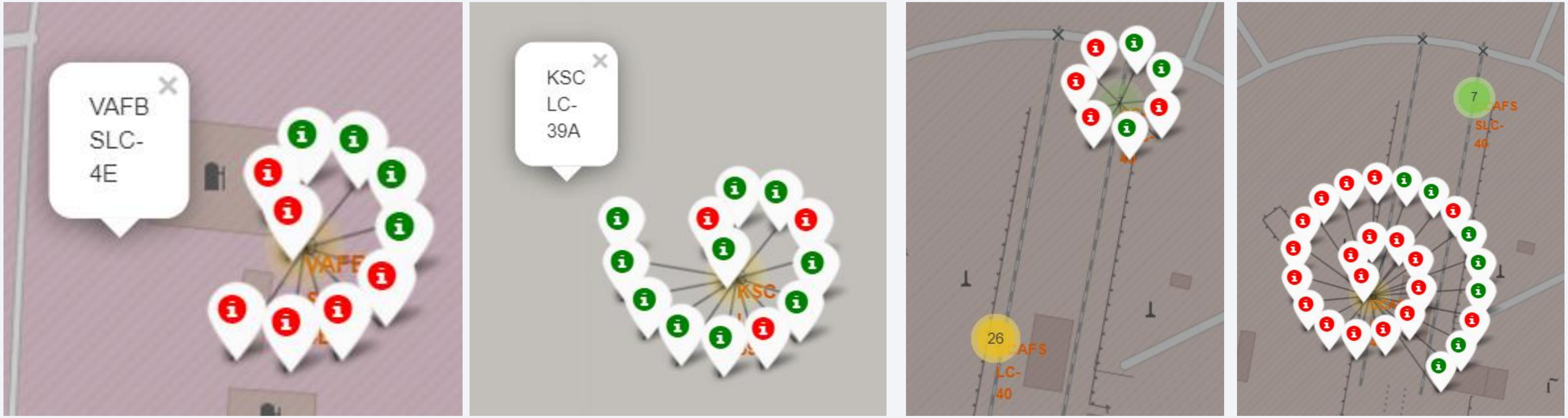
Launch Sites Proximities Analysis

Launch sites marked on a Folium map



This map shows the locations of the launch sites which are in California and Florida. They are located near the coastline and relatively far away from densely populated areas.

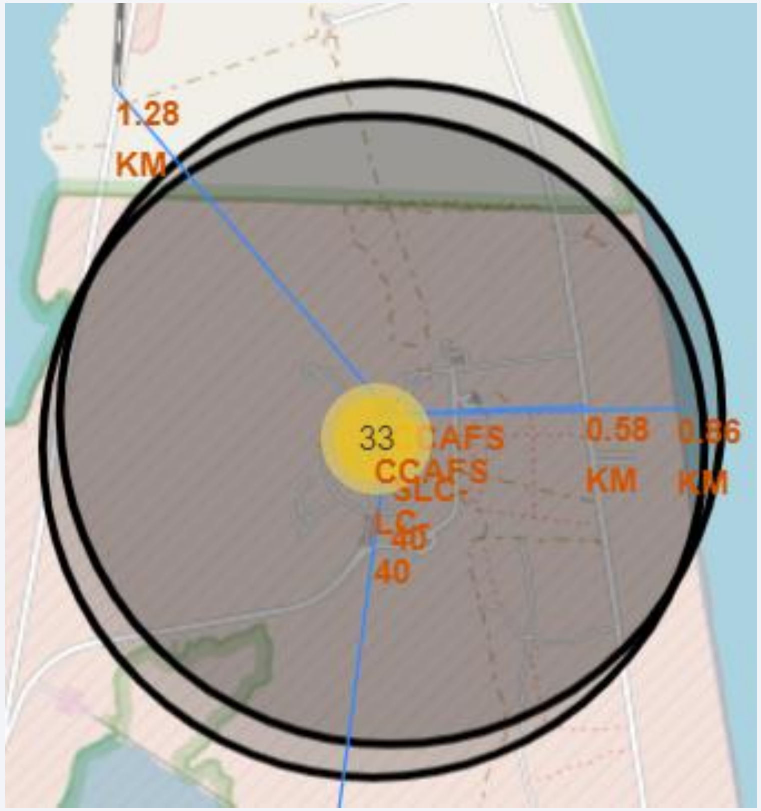
Color success markers

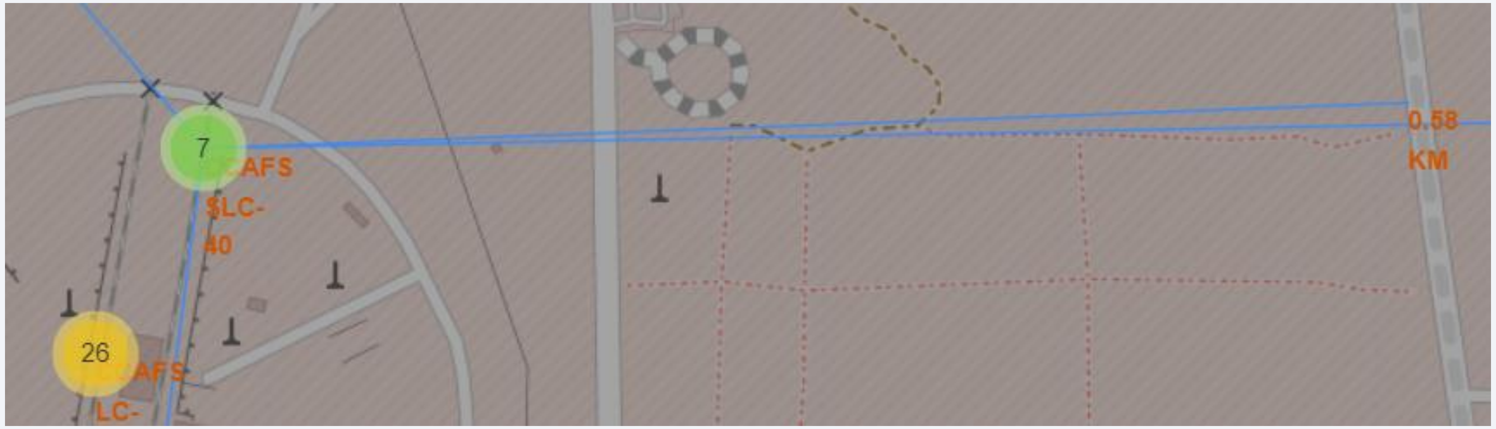


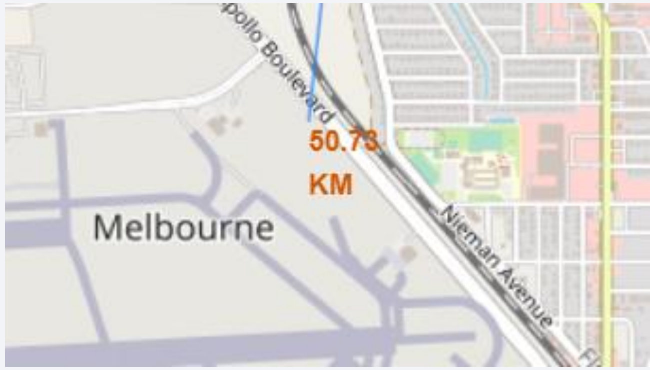
- In these marker clusters successful landings are displayed in green and failed landings in red.

Launch site vicinities

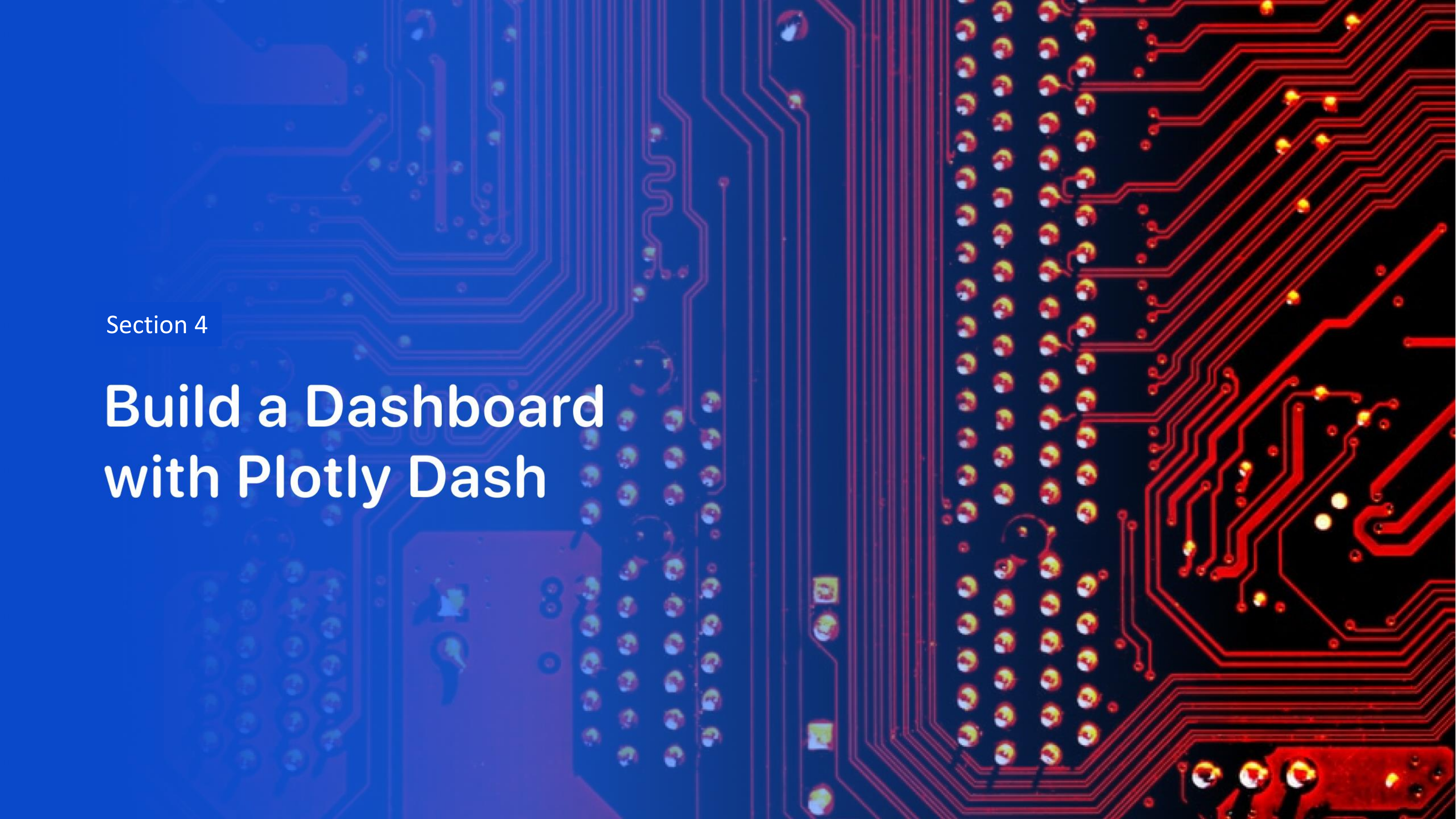
Calculating distances between the CCAFS-SLC-40 site and different landmarks.







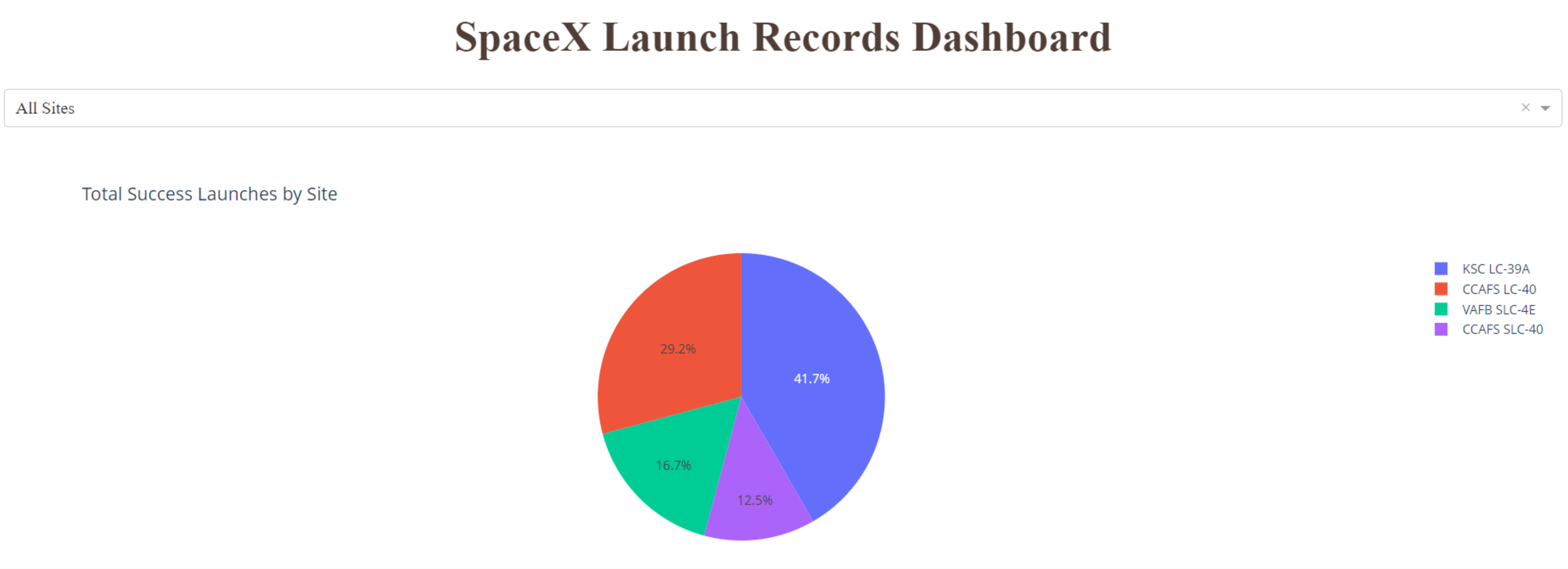
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities?Yes



Section 4

Build a Dashboard with Plotly Dash

Success rate across all sites

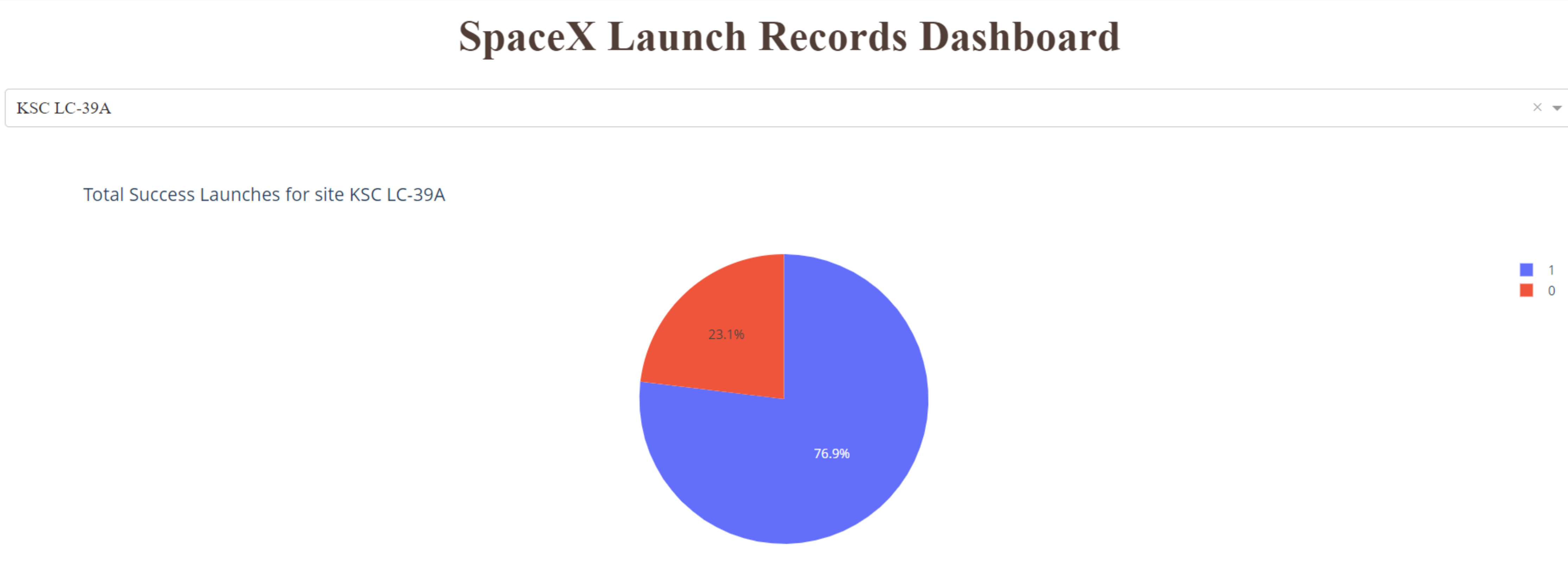


SpaceX Launch Records Dashboard

All Sites

The pie chart shows KSC LC-39A as being the most successful launch site.

Most successful launch site



SpaceX Launch Records Dashboard

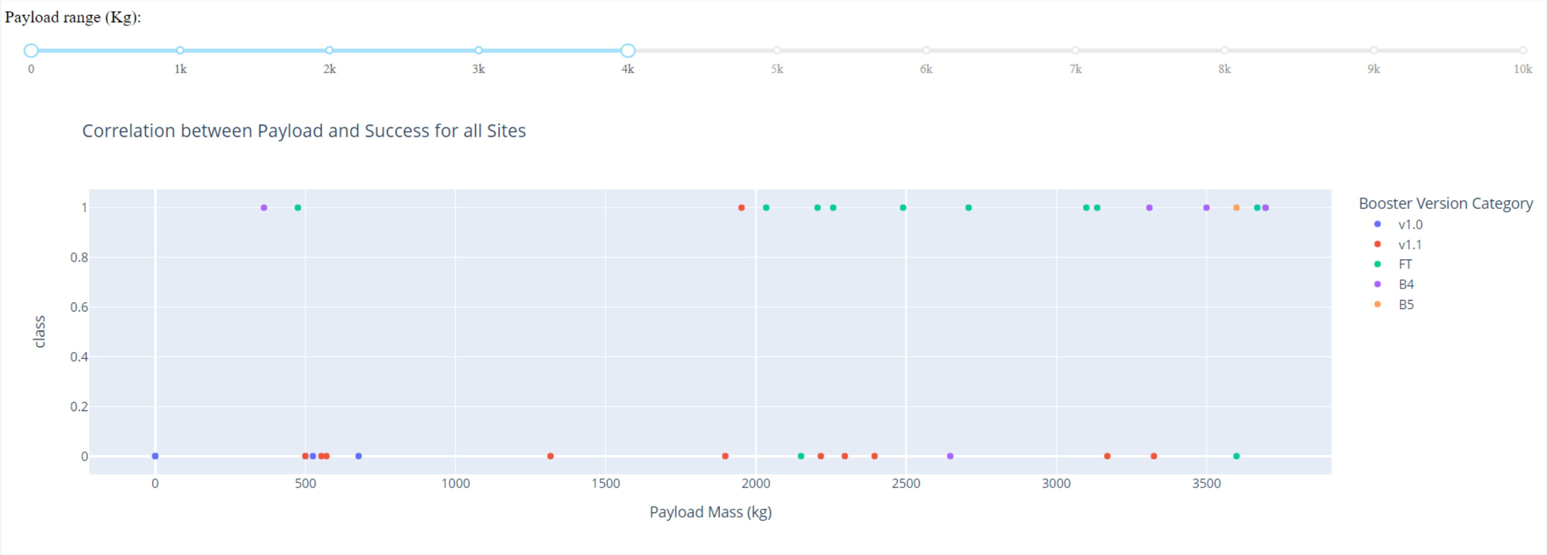
KSC LC-39A

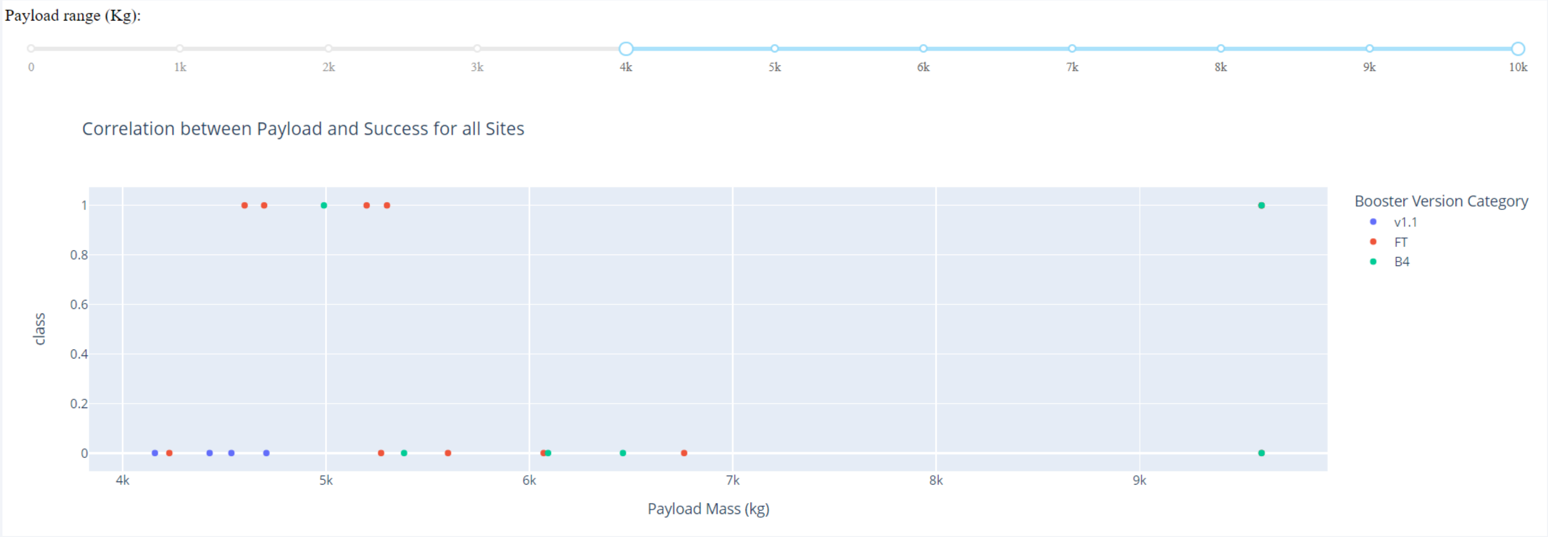
Total Success Launches for site KSC LC-39A

KSC LC-39A achieved a 76.9% success rate.

Payload vs. Launch Outcome scatter plot

This scatter plot suggests that the success rates for lighter payloads are higher than for the heavy ones.

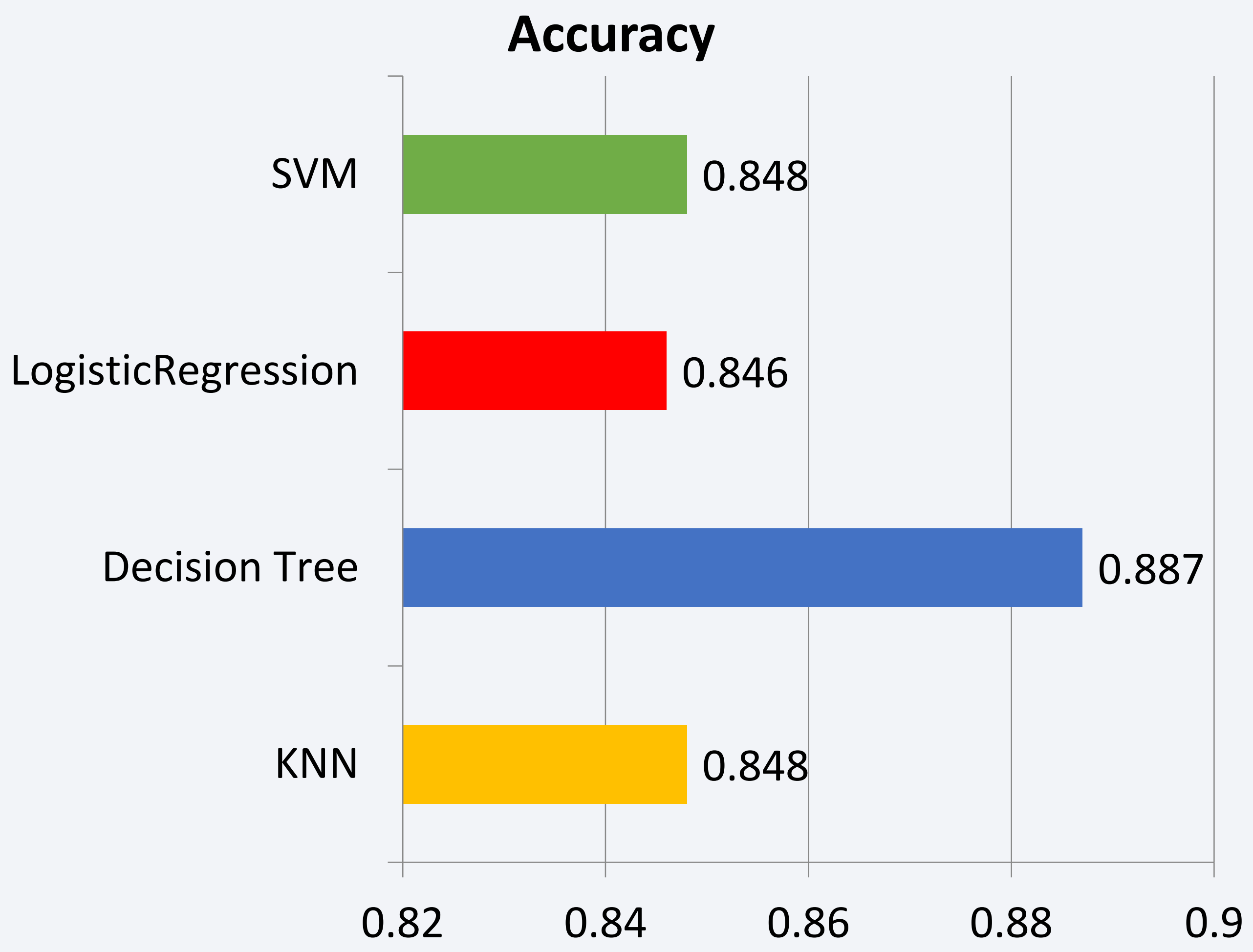

Payload range (Kg):


Payload range (Kg):

Section 5

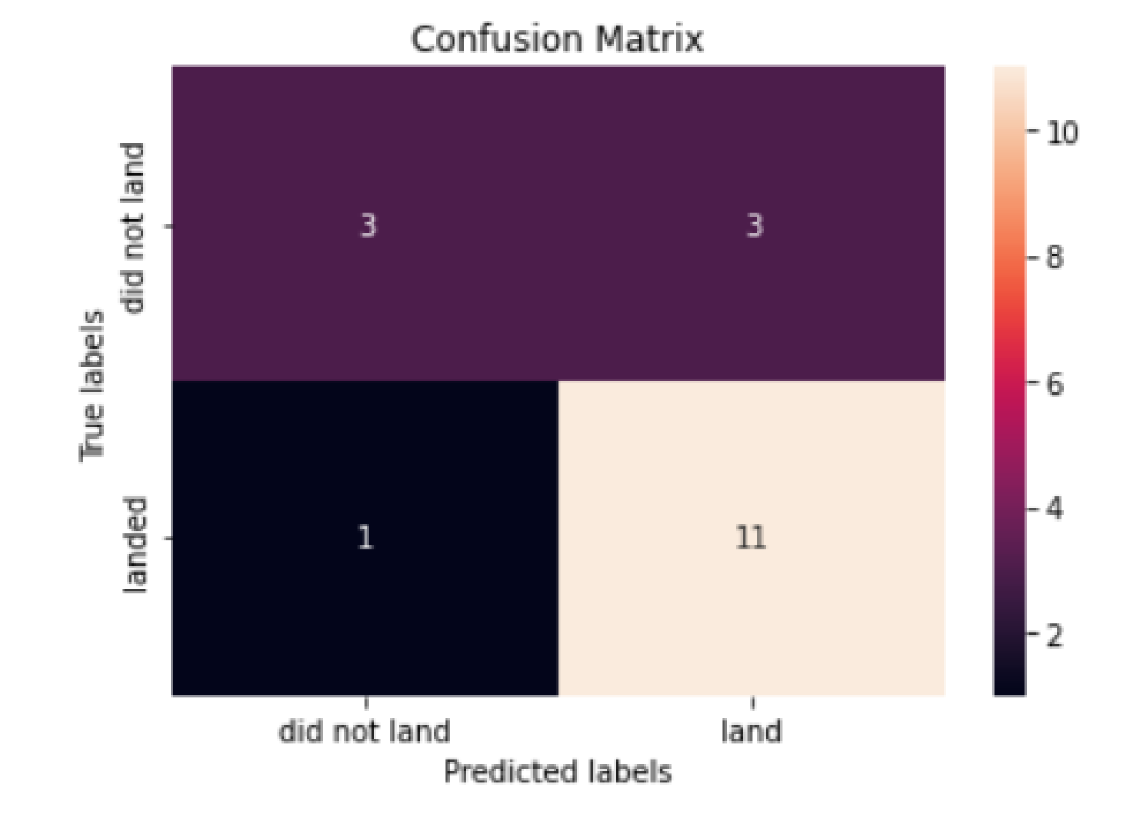
Predictive Analysis (Classification)

Classification Accuracy



Best performing algorithm is Decision Tree, with a calculated accuracy of approximately 0.887.

Decision tree confusion matrix



Confusion matrix metrics:

True Positive: 11
False Positive: 3
True Negative Rate: 3
False Negative: 1

Accuracy: (TP+TN)/total = 0.77
Misclassification Rate: (FP+FN)/total = 0.22
Precision: TP/TP + FP = 0.78

Conclusions

- The success rate for SpaceX launches has generally been increasing with time.
- The decision tree algorithm is the best machine learning model for this datatset, having an accuracy of approximately 0.88.
- A positive correlation between number of flights and success rate has been noted.
- Launch site KSC LC-39A had the highest success rate (76.9%).
- Orbits ES-L1, GEO, HEO, SSO have a 100% success rate.
- Lighter payloads seem to be correlated with higher success rates.

Appendix

GitHub repo link:

https://github.com/radu177/Data-Science-Project/tree/master

Thank you!