IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chihiro Takeda

9/11/2023

Outline

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

Executive Summary

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

- Summary of all results
  - Data analysis results
  - Data visualization, interactive dashboard
  - Predictive model analysis results

Introduction

Project background and context

The goal of this project is to predict if the Falcon 9 first stage performed by SpaceX will land successfully.

SpaceX, the most successful company of the commercial space trip, states its Falcon 9 rocket could launch significantly cheaper than its competitors by reusing the first stage.

Therefore, determining the first stage's success is the key for the cost estimation of a launch. With public information and machine learning models, this project will predict the success of the first stage.

- Problems to be answered
  - Identify variables affecting the success of the first stage landing and their impact
  - Find if time affects the success rate
  - Correlations between variables
  - Find the best algorithm for machine learning model to predict the success

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scraping Falcon 9 and Falcon Heavy launch records from Wikipedia
- Perform data wrangling
  - Filter data, deal with missing values, create binary landing outcome feature,
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - With different classification algorithm, build a machine learning model, use split data for training and test, then evaluate classification models based on the results.

Data Collection

- Data sets used in this project were collected from:
  - SpaceX REST API
  - Web Scraping from the SpaceX's Wikipedia entry

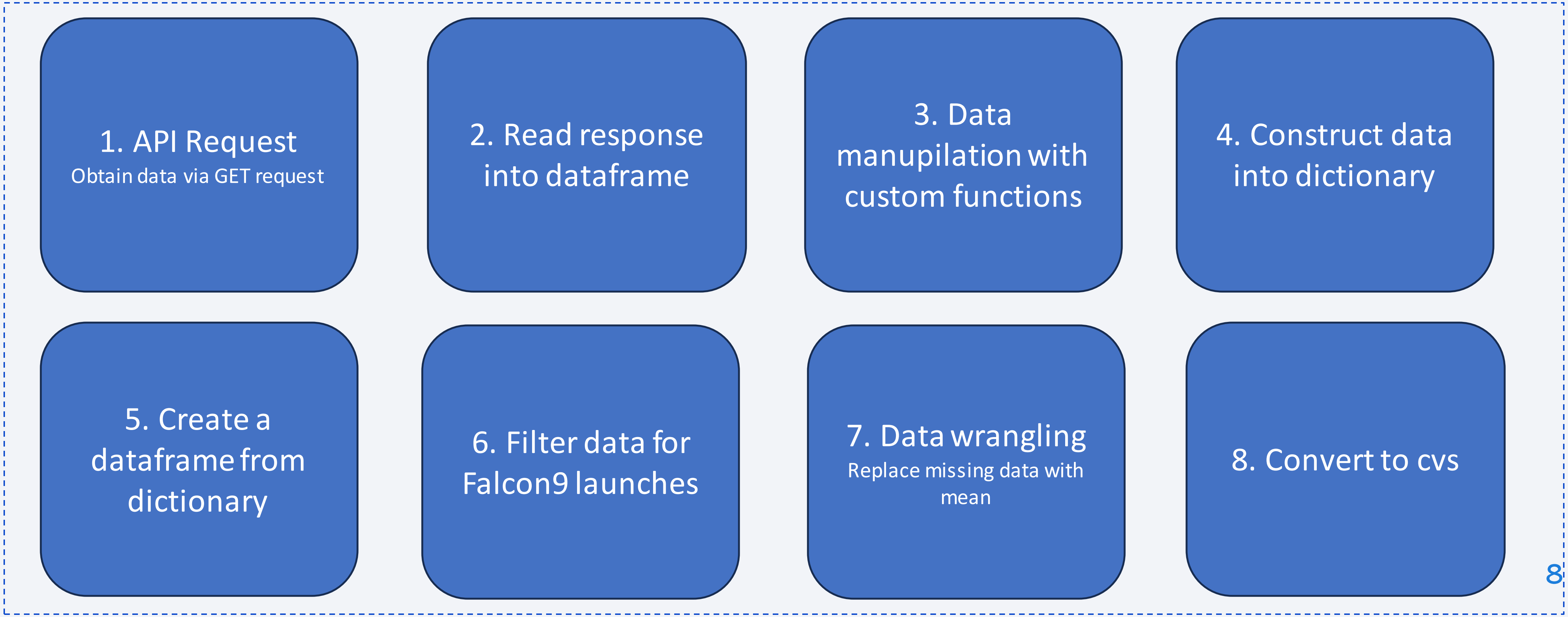
- Data from SpaceX Rest API
  - Flight number
  - Date
  - Booster version
  - Payload mass
  - Orbit
  - Launch site
  - Outcome
  - Flights
  - Grid fins etc

- Data from SpaceX Wikipedia
  - Flight number
  - Launch site
  - Payload
  - Payload mass
  - Orbit
  - Customer
  - Launch outcome
  - Version booster
  - Date etc

Data Collection – SpaceX API

- Data collection with SpaceX REST calls flowcharts with key phrases

GitHub URL: Data Collection API



1. API Request
Obtain data via GET request

2. Read response into dataframe

3. Data manupilation with custom functions

4. Construct data into dictionary

5. Create a dataframe from dictionary

6. Filter data for Falcon9 launches

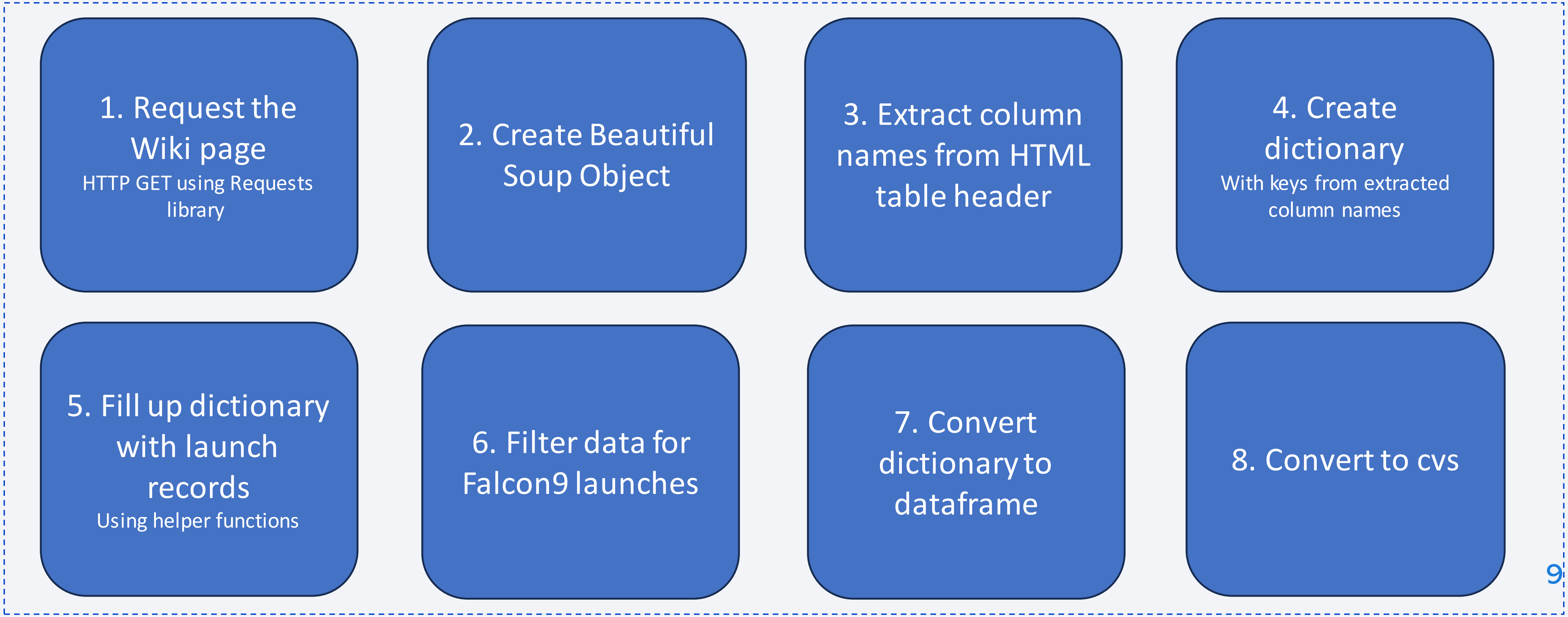
7. Data wrangling
Replace missing data with mean

8. Convert to cvs

Data Collection – Scraping

- Data collection via web scraping flowcharts with key phrases

GitHub URL: Data Collection Web Scraping



1. Request the Wiki page
HTTP GET using Requests library

2. Create Beautiful Soup Object

3. Extract column names from HTML table header

4. Create dictionary
With keys from extracted column names

5. Fill up dictionary with launch records
Using helper functions

6. Filter data for Falcon9 launches

7. Convert dictionary to dataframe

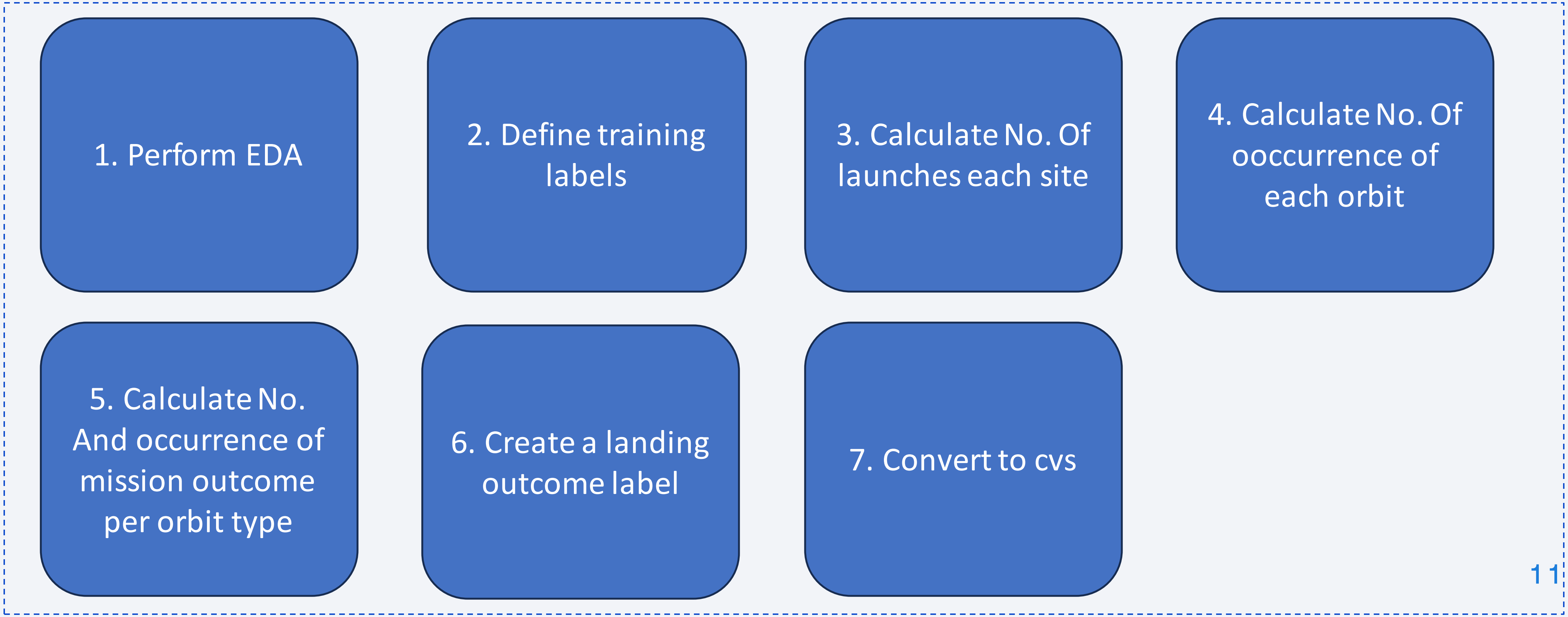
8. Convert to cvs

Data Wrangling

- Goal:
  - Conduct Exploratory Data Analysis (EDA) to find patterns in data
  - Define labels for training supervised models
  - Convert landing outcomes into binary score (0: failure, 1: success)

Data Wrangling

- Data wrangling flowcharts with key phrases GitHub URL: Data wrangling



1. Perform EDA

2. Define training labels

3. Calculate No. Of launches each site

4. Calculate No. Of ooccurrence of each orbit

5. Calculate No. And occurrence of mission outcome per orbit type

6. Create a landing outcome label

7. Convert to cvs

EDA with Data Visualization

GitHub URL: EDA with Data Visualization

- Charts plotted to gain further insights into the data
  - Scatter plots: shows correlation between two variables making patterns easy to observe
    - Flight number and Launch site
    - Payload and Launch site
    - Flight number and orbit type
    - Payload and orbit type
  - Bar charts: shows proportion of variables
    - Success rate of each orbit type
  - Line charts: tracks changes over a period of time
    - Average launch success yearly trend

EDA with SQL

- Summary of the SQL queries performed
  - Names of unique launch sites in the space mission
  - 5 records of launch sites beginning with 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date of the first successful landing outcome in the ground pad
  - Names of boosters with success in drone ship and payload mass between 4000 and 6000
  - Total number of successful and failed mission outcomes
  - Names of booster versions that carried the maximum payload mass
  - Month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
  - Ranking the count of landing outcomes or success between 2010/6/4 and 2017/3/20 in descending order

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

Folium interactive map was used for interactive visual analysis of geospatial data. It helped better understand factors such as location and proximity of launch sites that impacted launch success rate.

- Added items to the map:
  - Markers of all Launch Sites:
    - Marker with circle, popup label, text label of all launch sites and NASA Johnson Space Center to show their geographical locations to proximity to Equator and coasts
  - Colored Markers of the launch outcomes for each Launch Site
    - Colored markers of success and failed launches using Marker Cluster to identify launch sites with success rate
  - Distance between a launch site to its proximities
    - Colored lines to show distance between selected launch sites and its proximities like railway, highway, coastline and closest city

GitHub URL: interactive map with Folium map

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - A dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - the total successful launches count for all sites
  - Success vs. Failed counts for the site, if a specific Launch Site was selected
- Slider of Payload Mass Range:
  - A slider to select Payload range
- Scatter plot of Payload Mass vs. Success Rate for the different Booster Versions:
  - A scatter plot to show the correlation between Payload and Launch Success.

GitHub URL: Plotly Dash

Predictive Analysis (Classification)

- Summary of how to build, evaluate, improve, and find the best performing classification model

GitHub URL: predictive analysis

Build

- Transform data to scale the columns
- Split data into testing and training sets
- Select machine learning algorithms
- Find best tuning paramters via grid search

Evaluate

- Check accuracy of each model
- Plot confusion matrix

Improve

- Feature engineering
- Algorithm tuning

Select

- Select the model with the best accuracy score

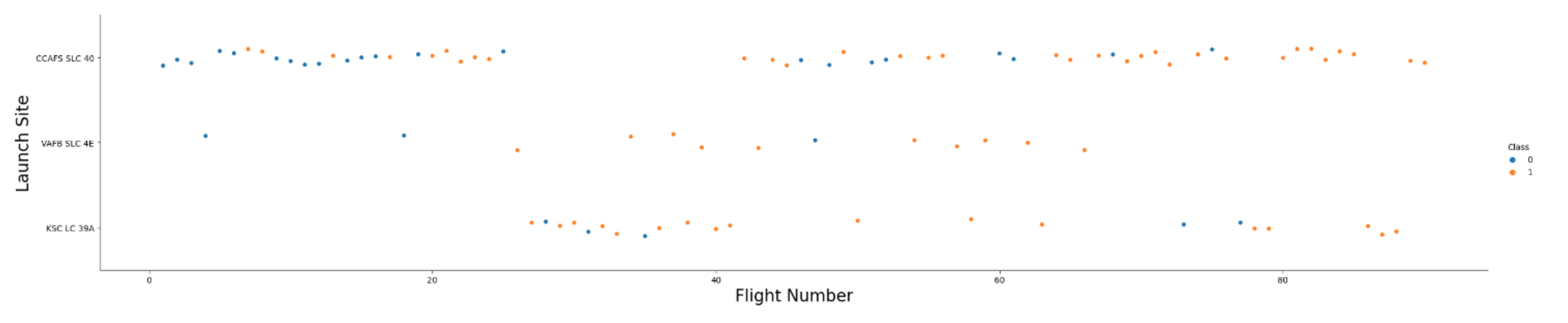
Results

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

Section 2

Insights drawn from EDA

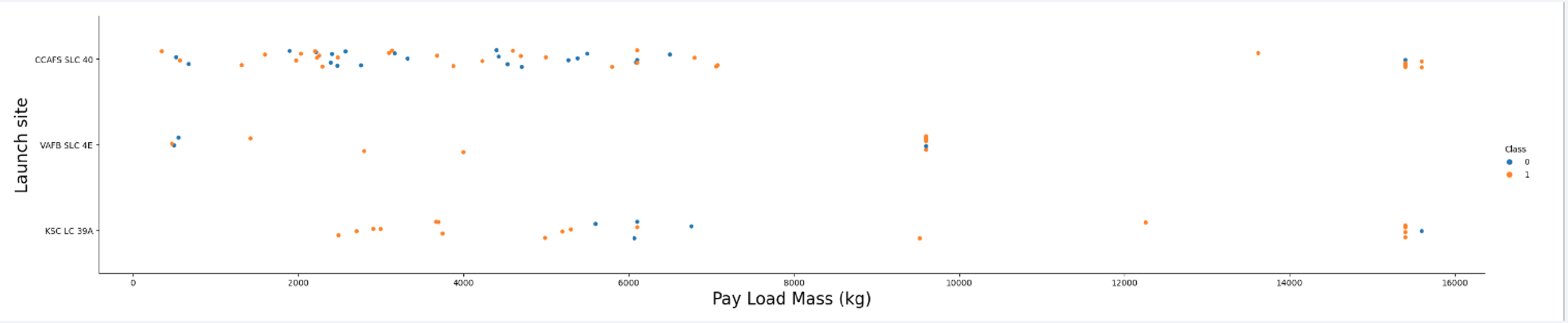
Flight Number vs. Launch Site



Explanations:

- The earliest flights tend to fail while the latest flights all succeeded
- Success rate goes up overtime
- About a half of all launches came from launch site CCAFS SLC 40
- Launch sites VAFB SLC 4E and KSC LC 39A have higher success rates

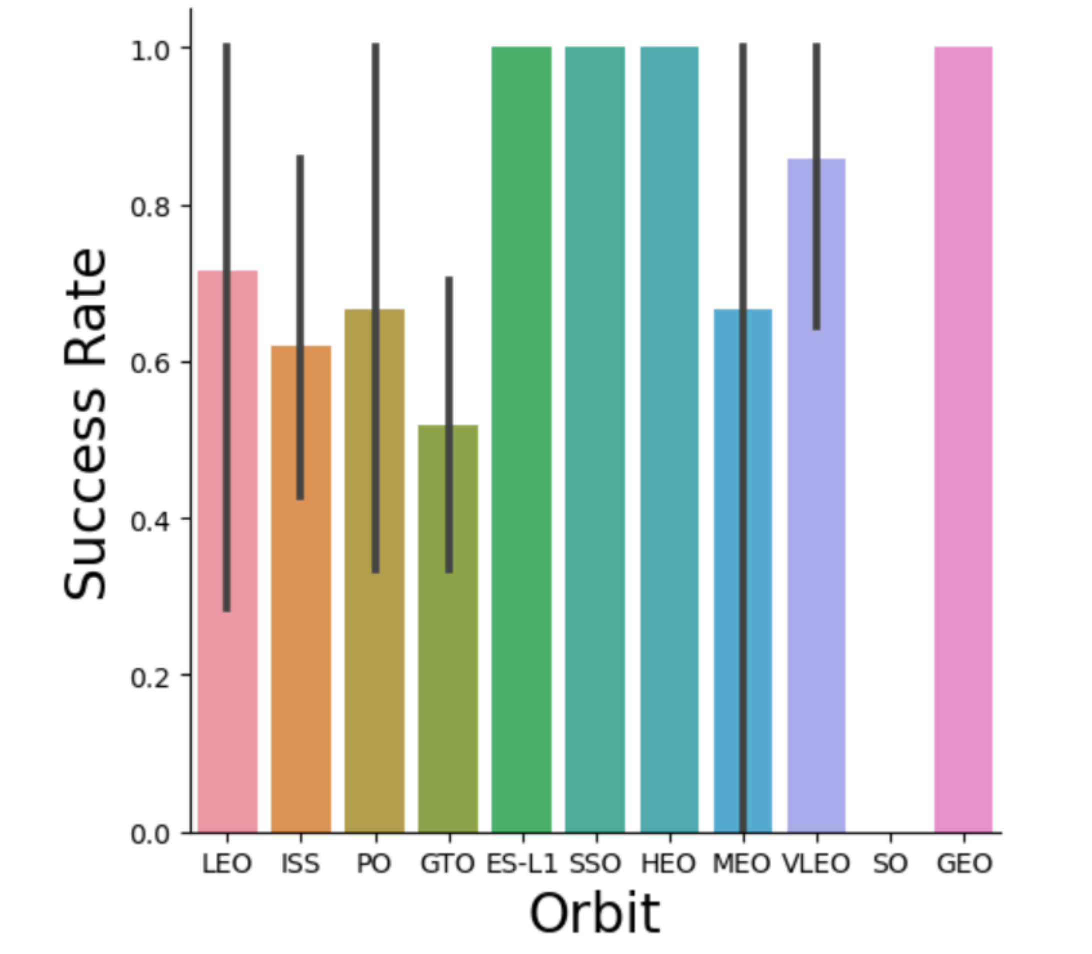
Payload vs. Launch Site



- Launches with higher payload mass have higher success rate in every launch site
- Most of the launches with payload mass over 7000 kg were successful.
- In KSC LC 39A, launches with payload mass under 5500 kg has a 100% success rate

Success Rate vs. Orbit Type

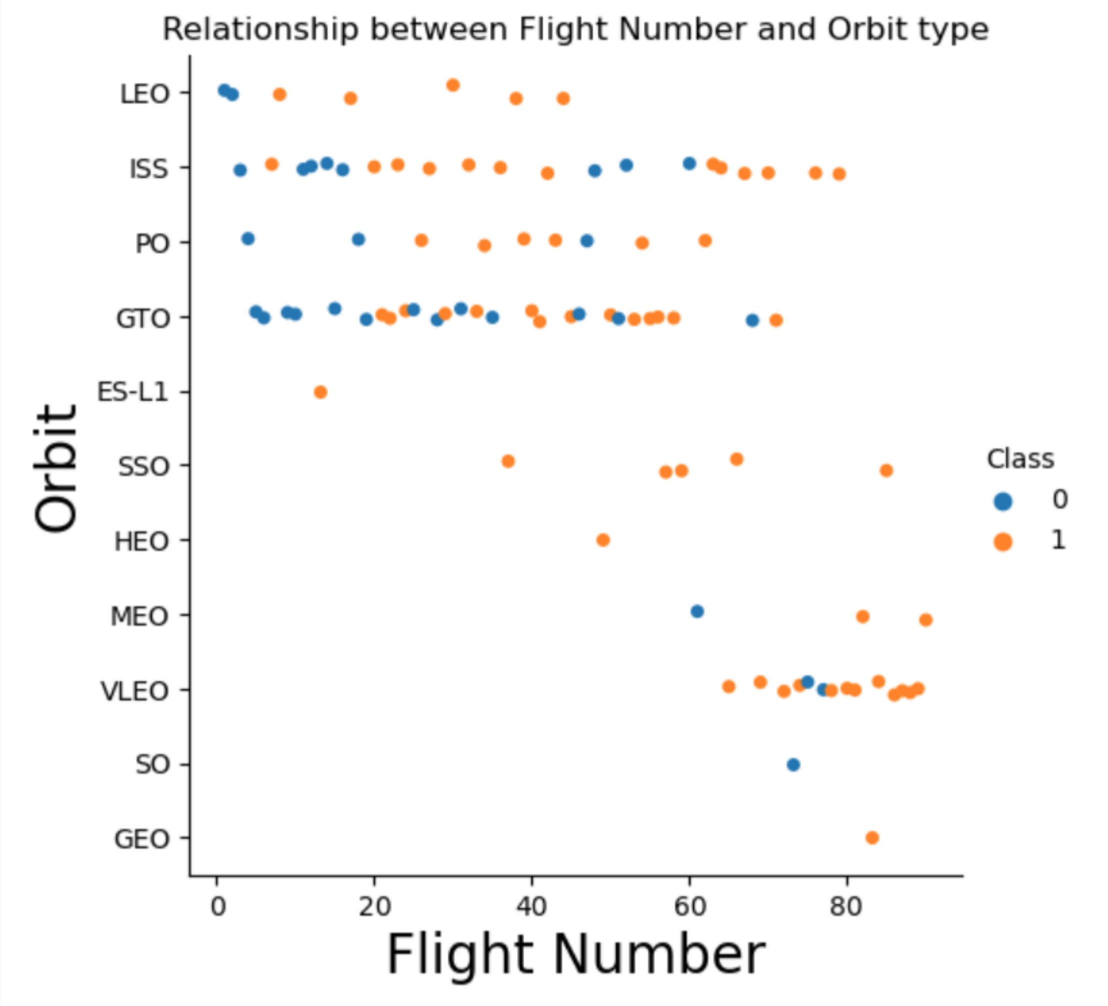
- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO



Flight Number vs. Orbit Type

Explanation:

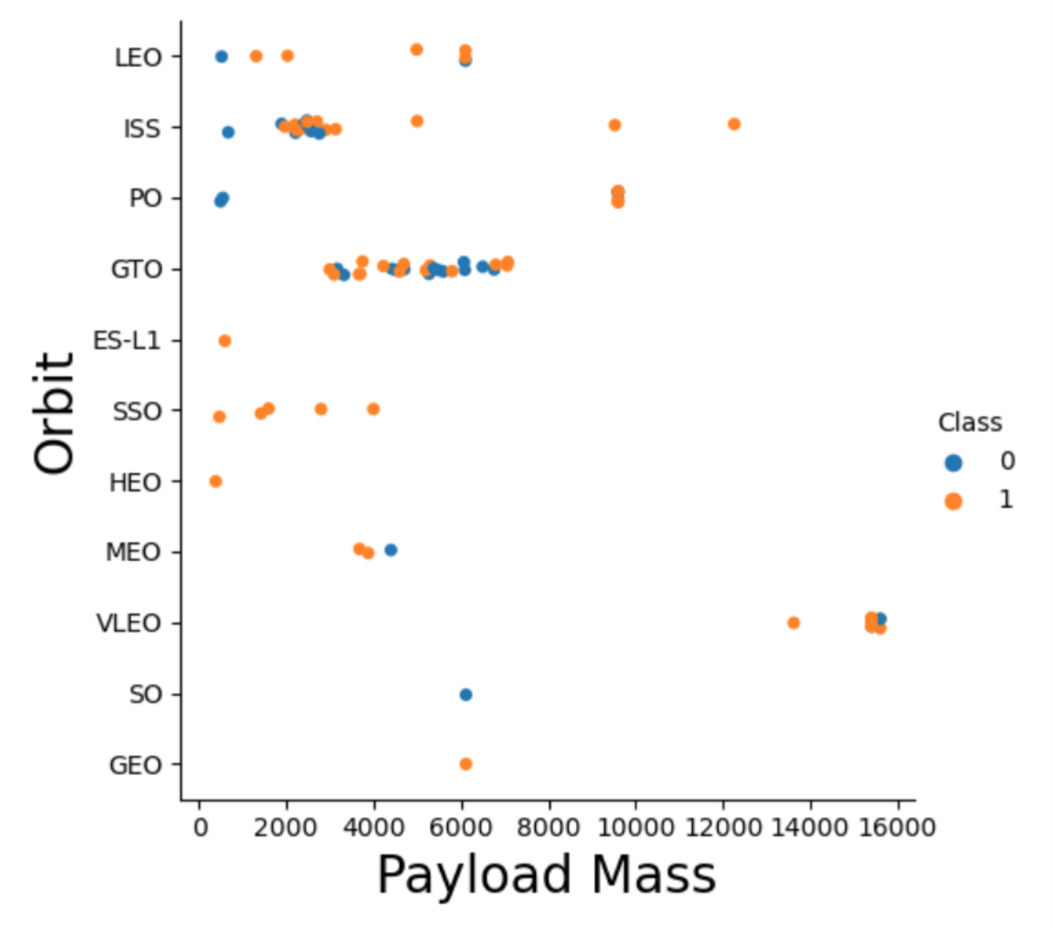
- In LEO, PO, MEO, the success rate goes up as the number of flights increases
- No relationship between flight number and success rate in GTO
- ES-L1, SSO, HEO, GEO have 100% success rate



Payload vs. Orbit Type

Explanation:

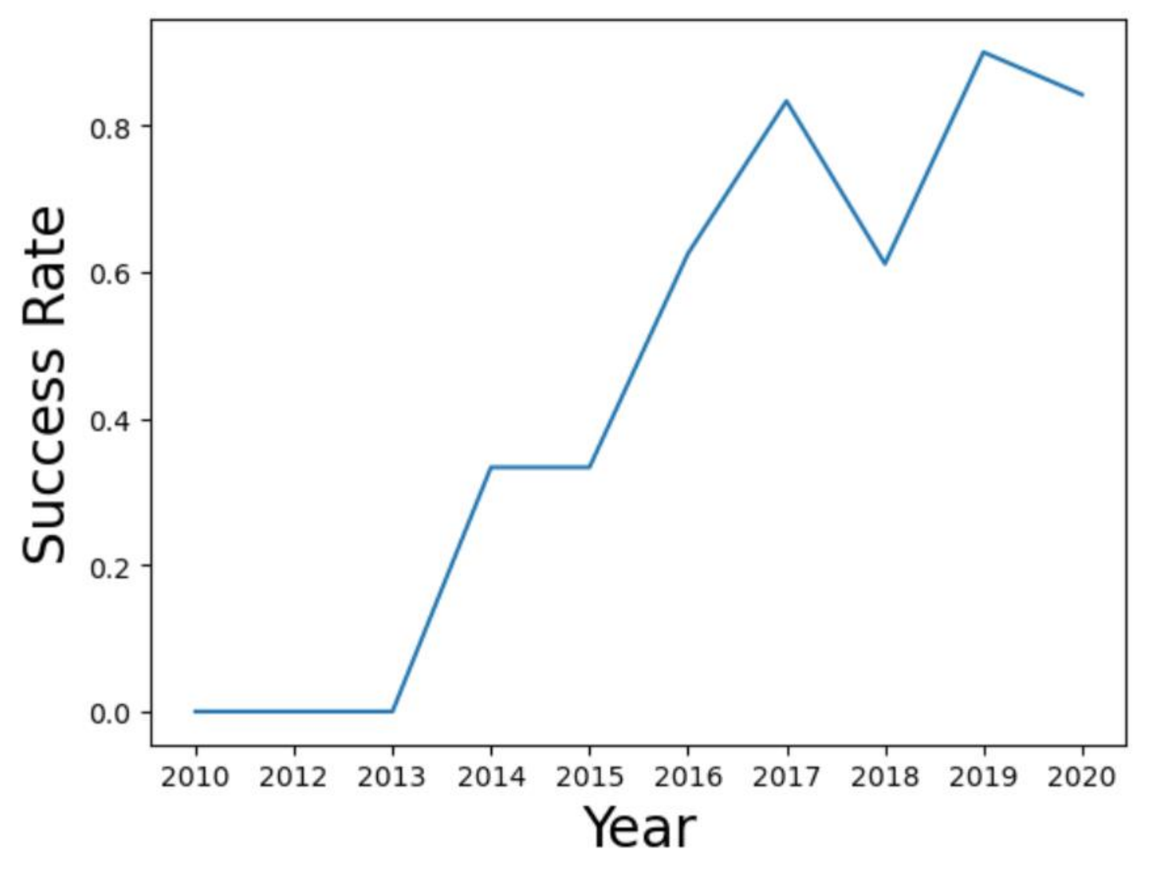
- In ISS, PO, LEO, success rate goes up as payload mass increases
- GTO does not have correlation between payload mass and orbit
- Not enough data to show relationship between payload mass and orbit type in other orbits



Launch Success Yearly Trend

Explanation:

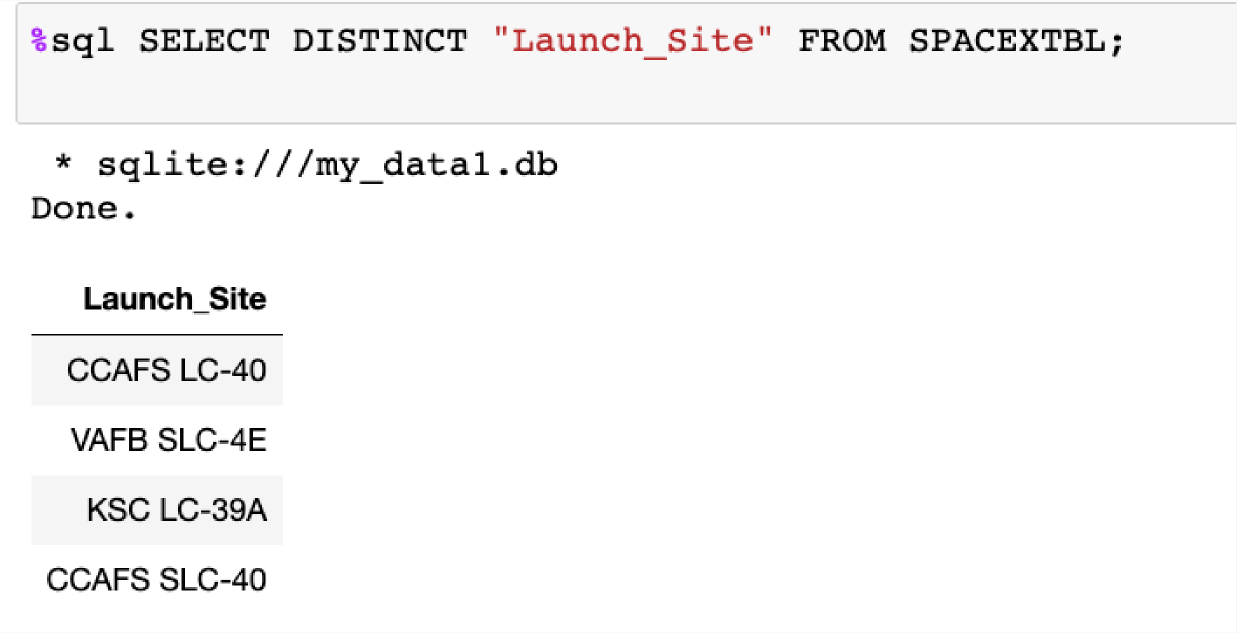
- Yearly average success rate was 0 prior to 2013
- Success rate shows upward trend after 2013 until 2019
- Success rate went down by 25% in 2018



All Launch Site Names

Explanation:

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



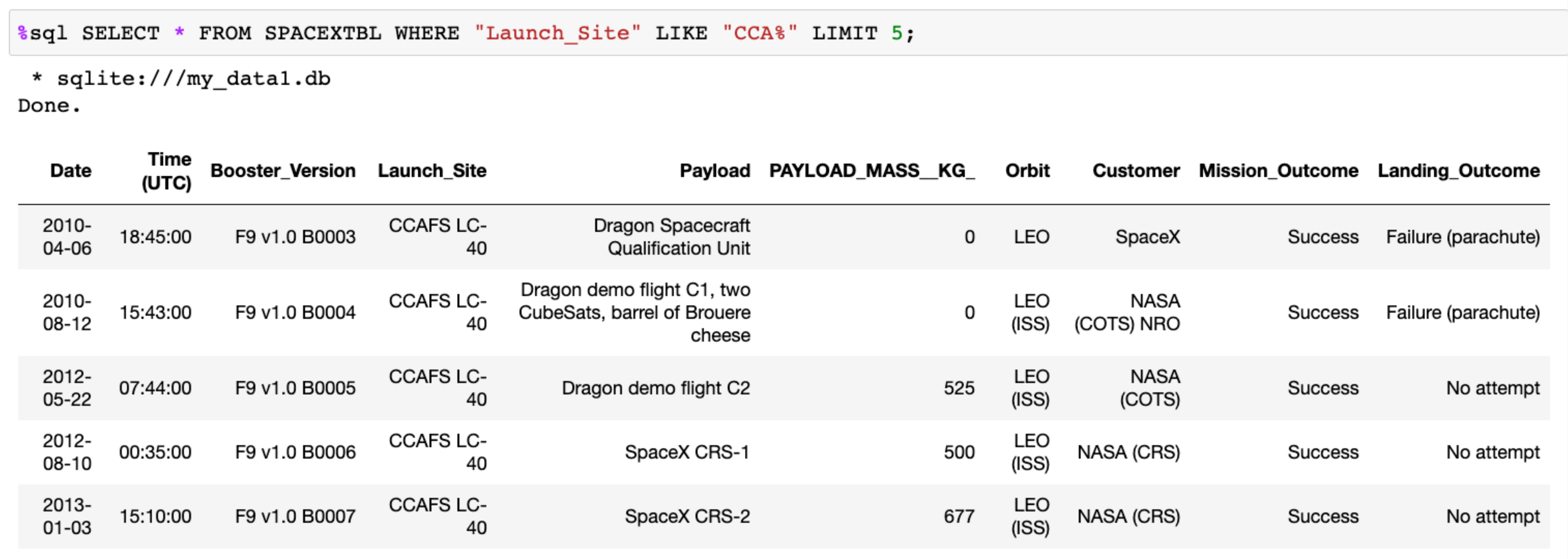
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

```
 * 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'

- Displaying 5 records where launch sites begin with `CCA`



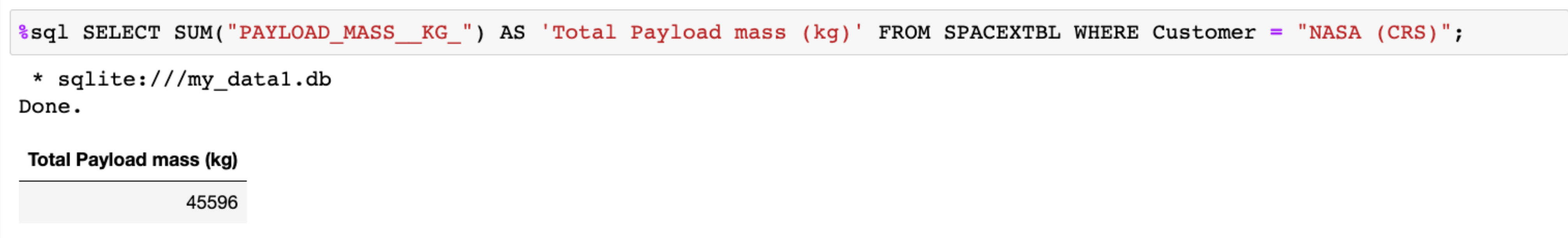
```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE "CCA%" LIMIT 5;
```

```
 * 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

- Displaying total payload carried by boosters from NASA (CRS) calculated



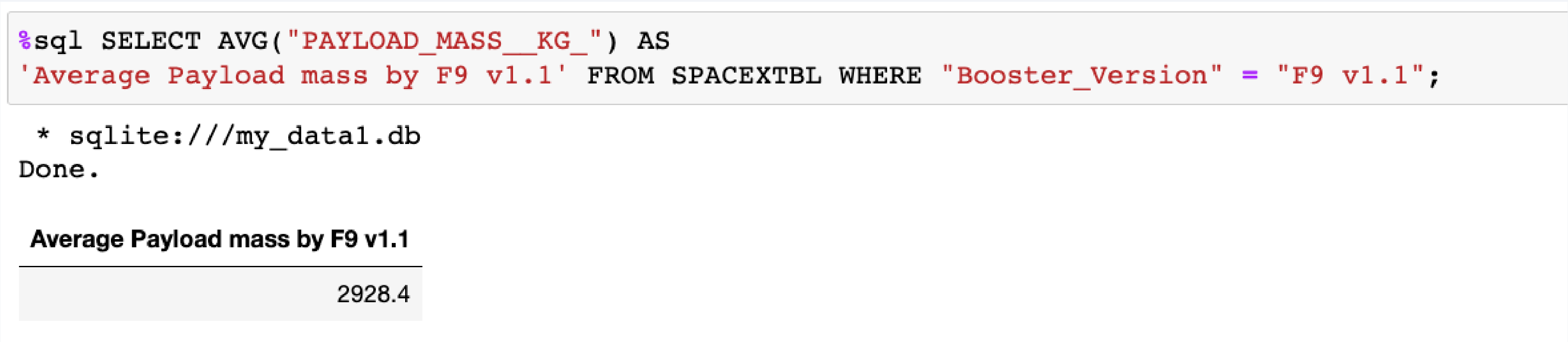
```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS 'Total Payload mass (kg)' FROM SPACEXTBL WHERE Customer = "NASA (CRS)";
```

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

Total Payload mass (kg)
45596

Average Payload Mass by F9 v1.1

- Displaying calculate the average payload mass carried by booster version F9 v1.1



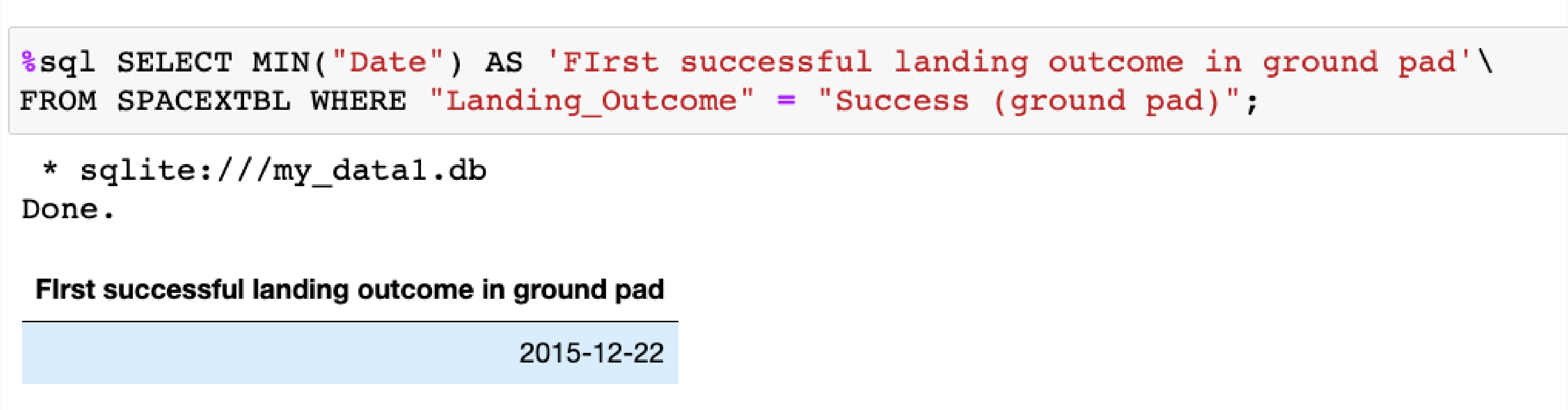
```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS
'Average Payload mass by F9 v1.1' FROM SPACEXTBL WHERE "Booster_Version" = "F9 v1.1";
```

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

Average Payload mass by F9 v1.1
2928.4

First Successful Ground Landing Date

- Displaying the date of the first successful landing outcome on ground pad



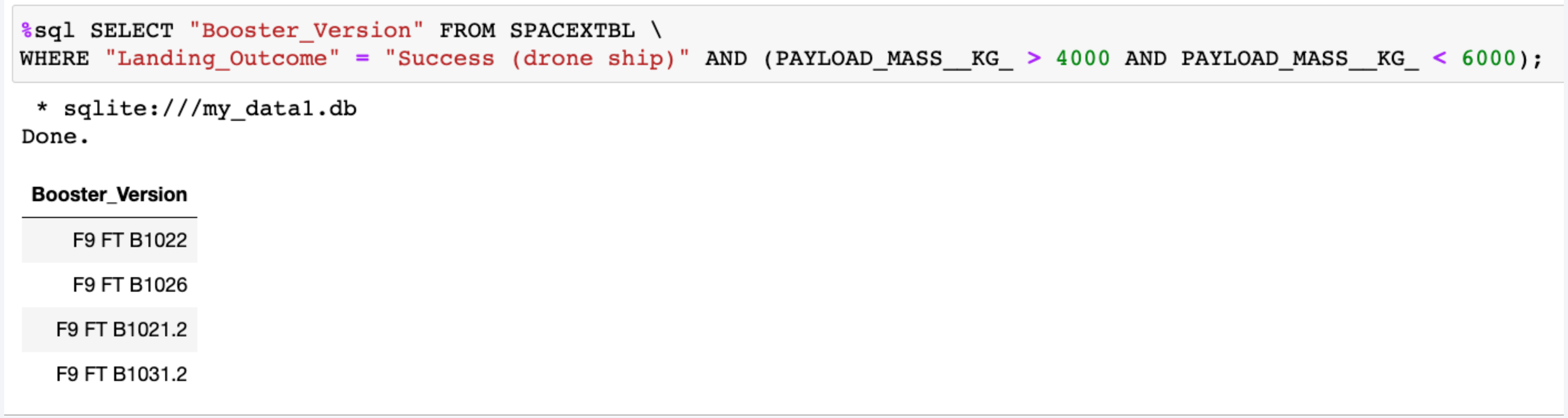
```
%sql SELECT MIN("Date") AS 'FIrst successful landing outcome in ground pad'\
FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)";
```

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

FIrst successful landing outcome in ground pad
2015-12-22

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



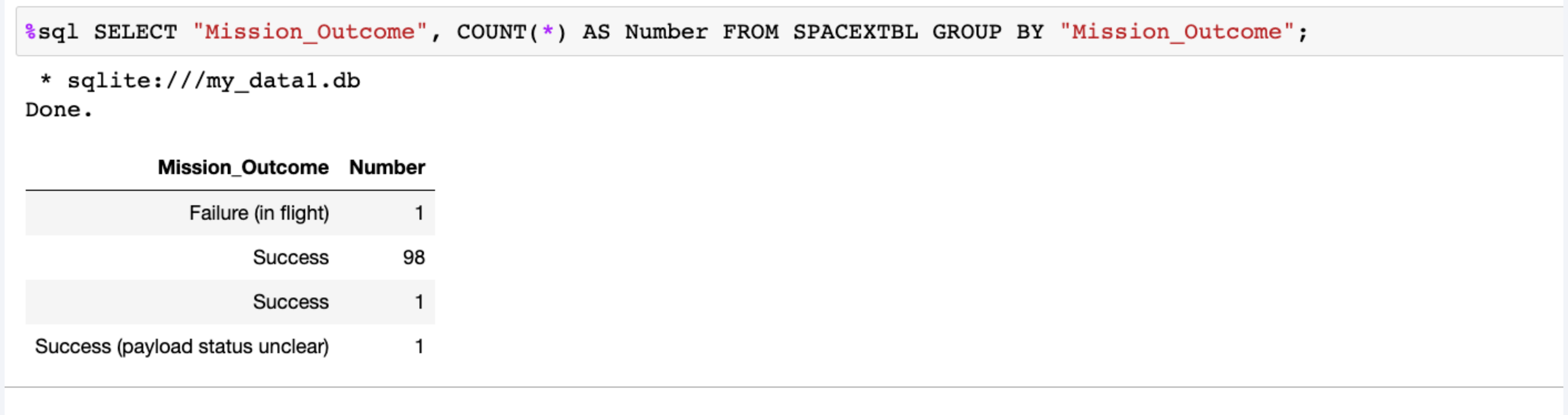
```
%sql SELECT "Booster_Version" FROM SPACEXTBL \
WHERE "Landing_Outcome" = "Success (drone ship)" AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000);
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



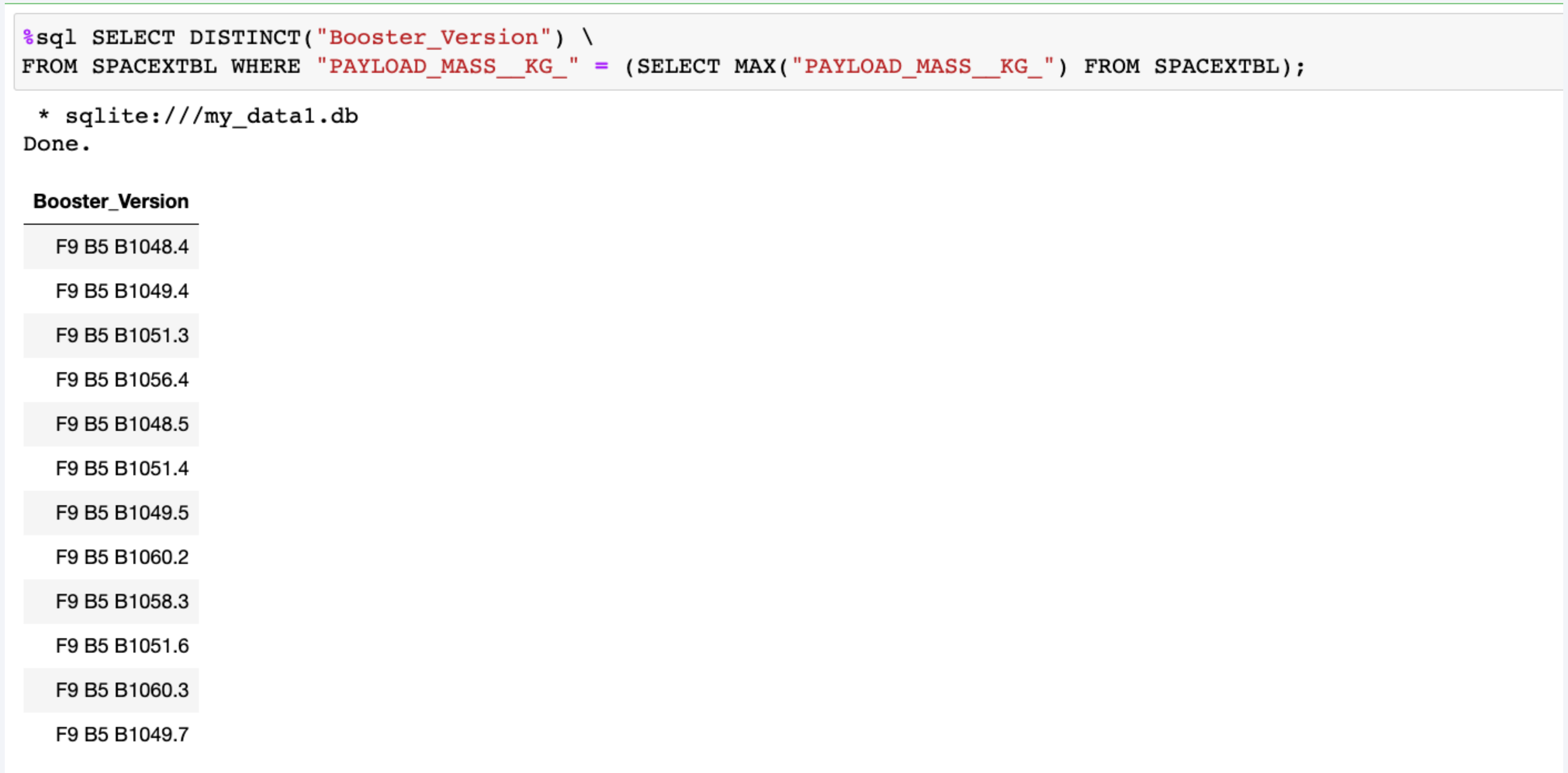
```
%sql SELECT "Mission_Outcome", COUNT(*) AS Number FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass



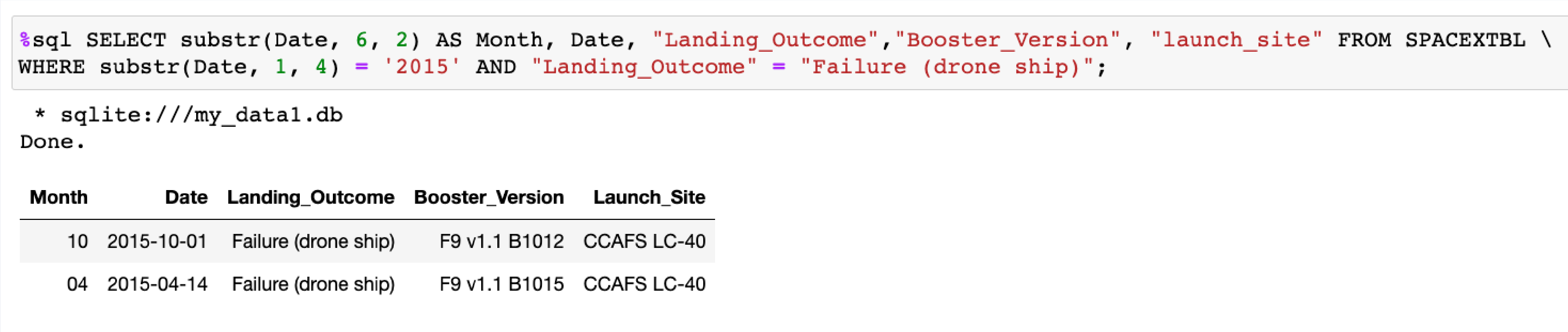
```
%sql SELECT DISTINCT("Booster_Version") \
FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

```
 * 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

- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



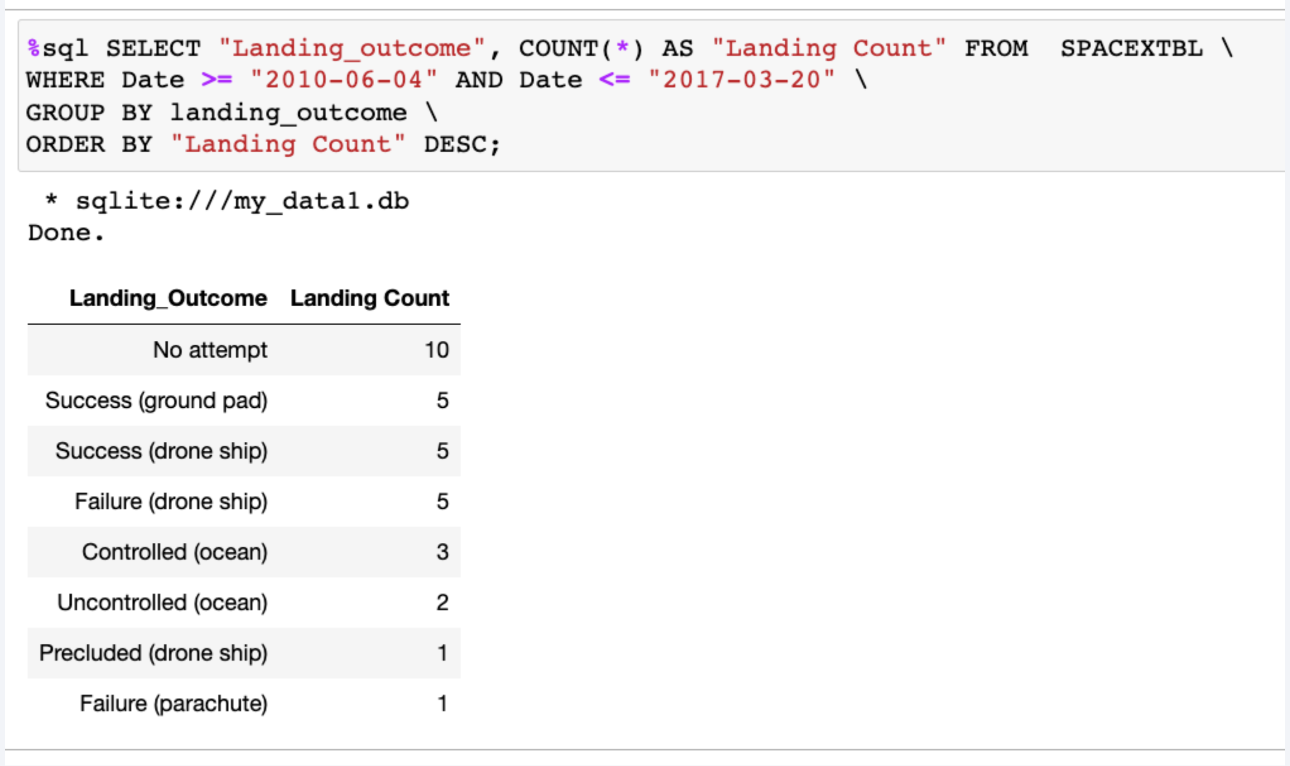
```
%sql SELECT substr(Date, 6, 2) AS Month, Date, "Landing_Outcome","Booster_Version", "launch_site" FROM SPACEXTBL \
WHERE substr(Date, 1, 4) = '2015' AND "Landing_Outcome" = "Failure (drone ship)";
```

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

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

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

- 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 "Landing Count" FROM  SPACEXTBL \
WHERE Date >= "2010-06-04" AND Date <= "2017-03-20" \
GROUP BY landing_outcome \
ORDER BY "Landing Count" DESC;
```

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

Landing_Outcome	Landing Count
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

All launch sites' location markers on a google map

Explanation:

- Fig 1 is the Global map with Falcon 9 launch sites located in the United States (in California and Florida).
- Each launch site contains a circle, label, and a popup to highlight the location and the name of the launch site.
- All launch sites are near the coast and in proximity to the Equator.

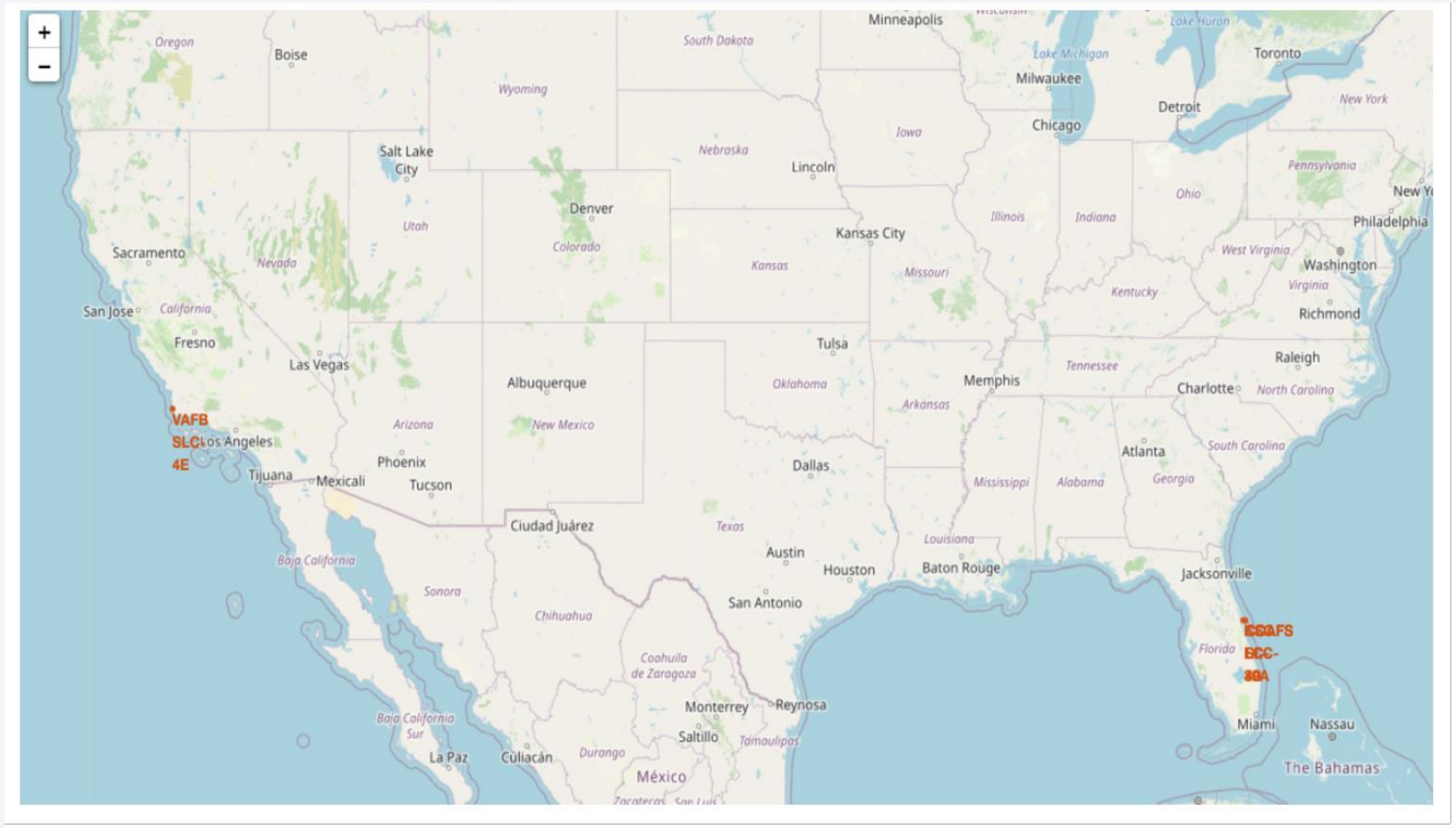


Fig 1

All launch sites' location markers on a google map

Explanation:

- Fig 2 zooms in to VAFB SLC-4E (CA)
- Fig 3 zooms in to the following launch sites to display:
  - CCAFS LC-40 (FL)
  - KSC LC-39A (FL)
  - CCAFS SLC-40 (FL)

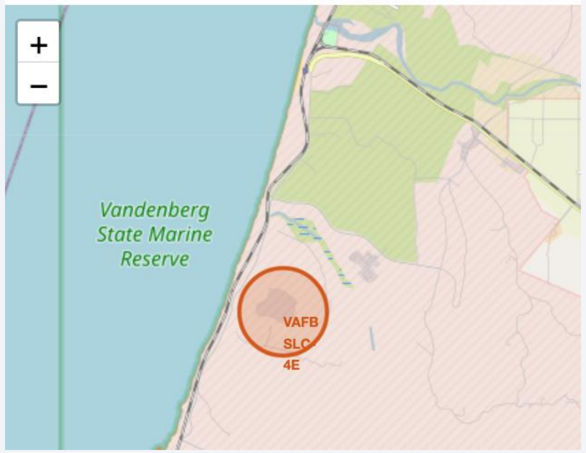


Fig 2

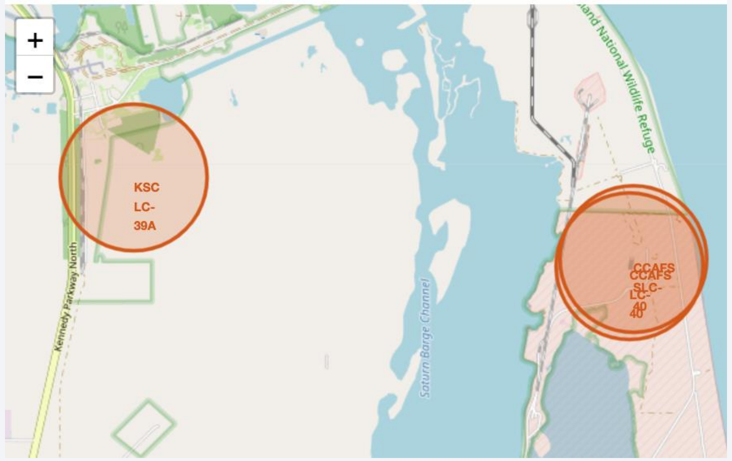
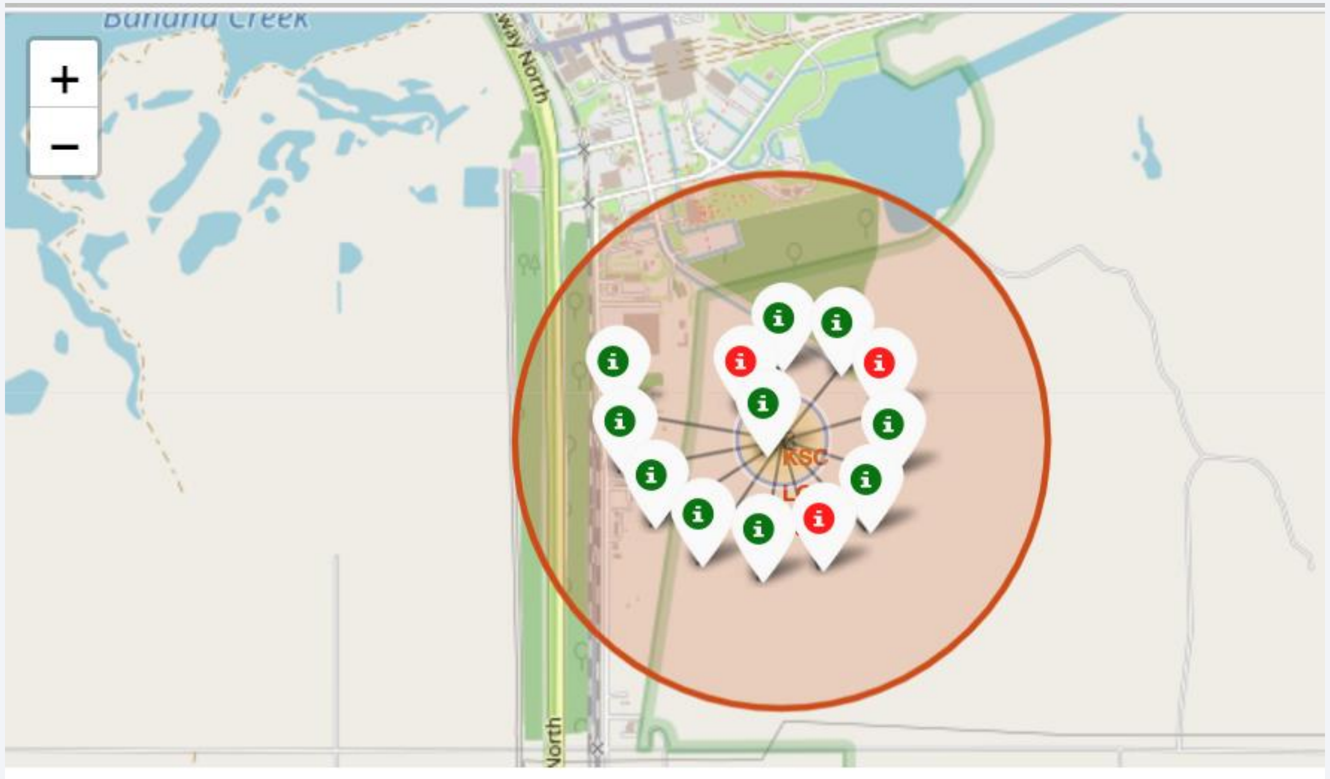


Fig 3

Color-labeled launch records on the map

Explanation:

- color-labeled markers tell success (green) / failure (red) launches

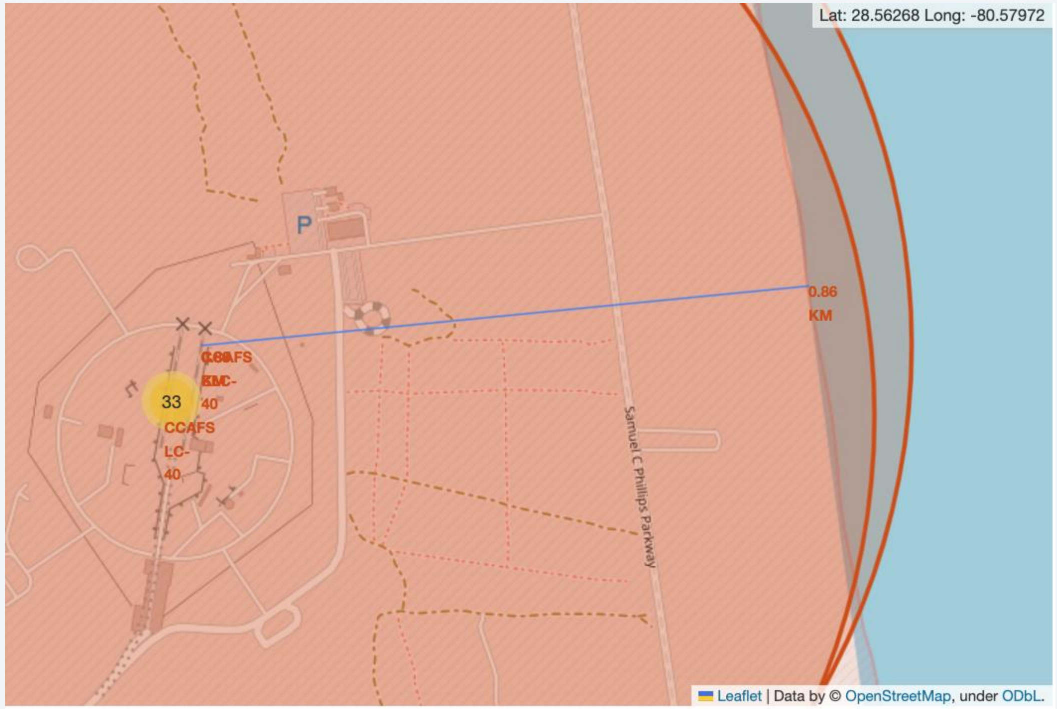


Distance from the launch site to its proximities

Explanation:

The visual analysis of CCAFS SLC-40

- close to coastal line (0.86 km)

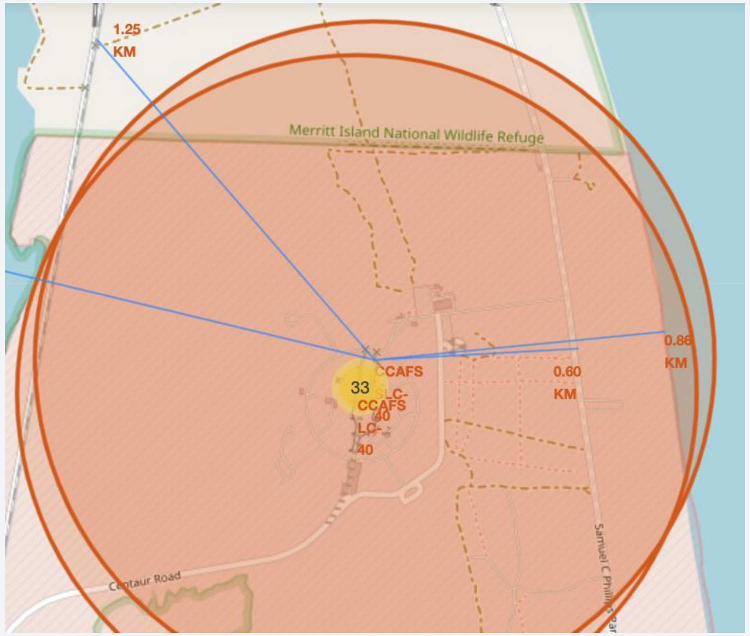

Lat: 28.56268 Long: -80.57972
P
0.86 KM
CCAFS SLC-40
33
CCAFS LC-40
Samuel C Phillips Parkway
Leaflet | Data by © OpenStreetMap, under ODbL.

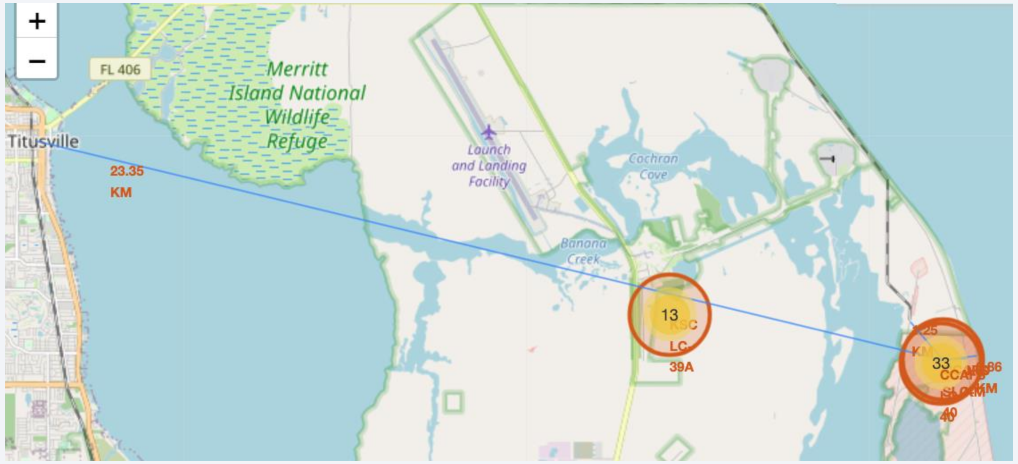
Distance from the launch site to its proximities

Explanation:

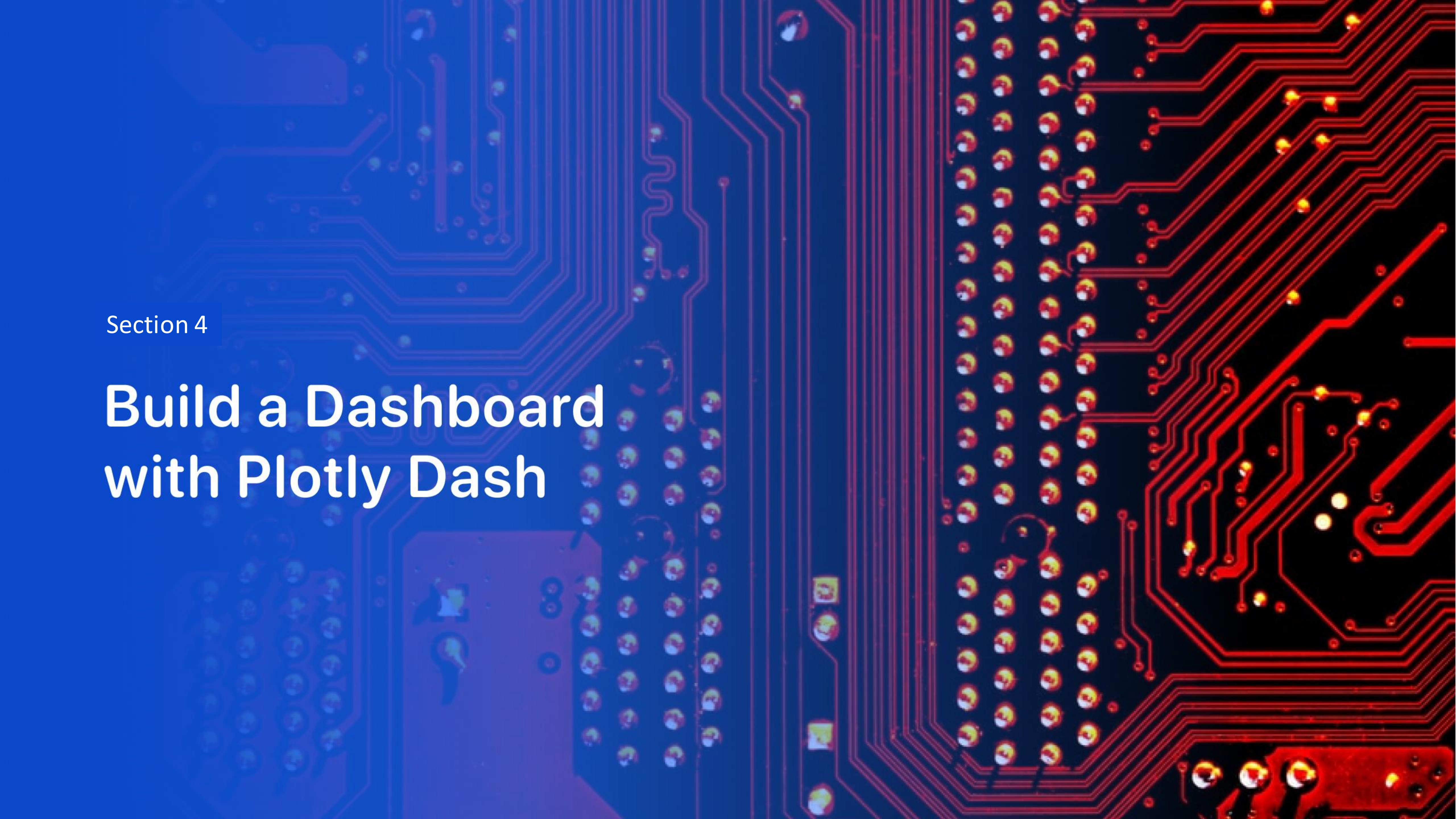
The visual analysis of CCAFS SLC-40

- close to railway (1.25 km)
- close to highway (0.6 km)
- Keeps certain distance away from its closest city Titusville (23.35 km).





+
−
FL 406
Merritt Island National Wildlife Refuge
Titusville
23.35 KM
Launch and Landing Facility
Cochran Cove
Banana Creek
13
KSC LC 39A
33

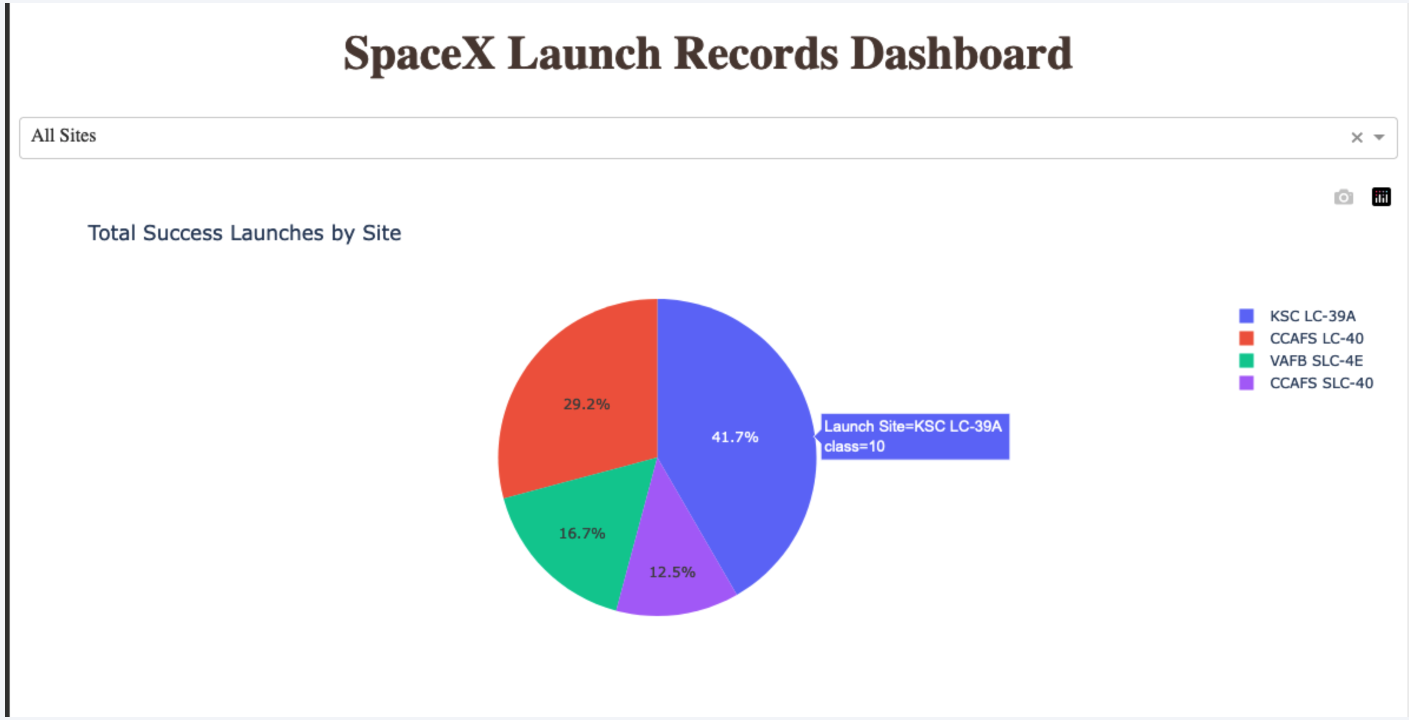


Section 4

Build a Dashboard with Plotly Dash

Launch success count for all sites

Explanation: the pie chart shows that KSC LC-39A has the most successful launches, which accounts for over 40% of all success launches.



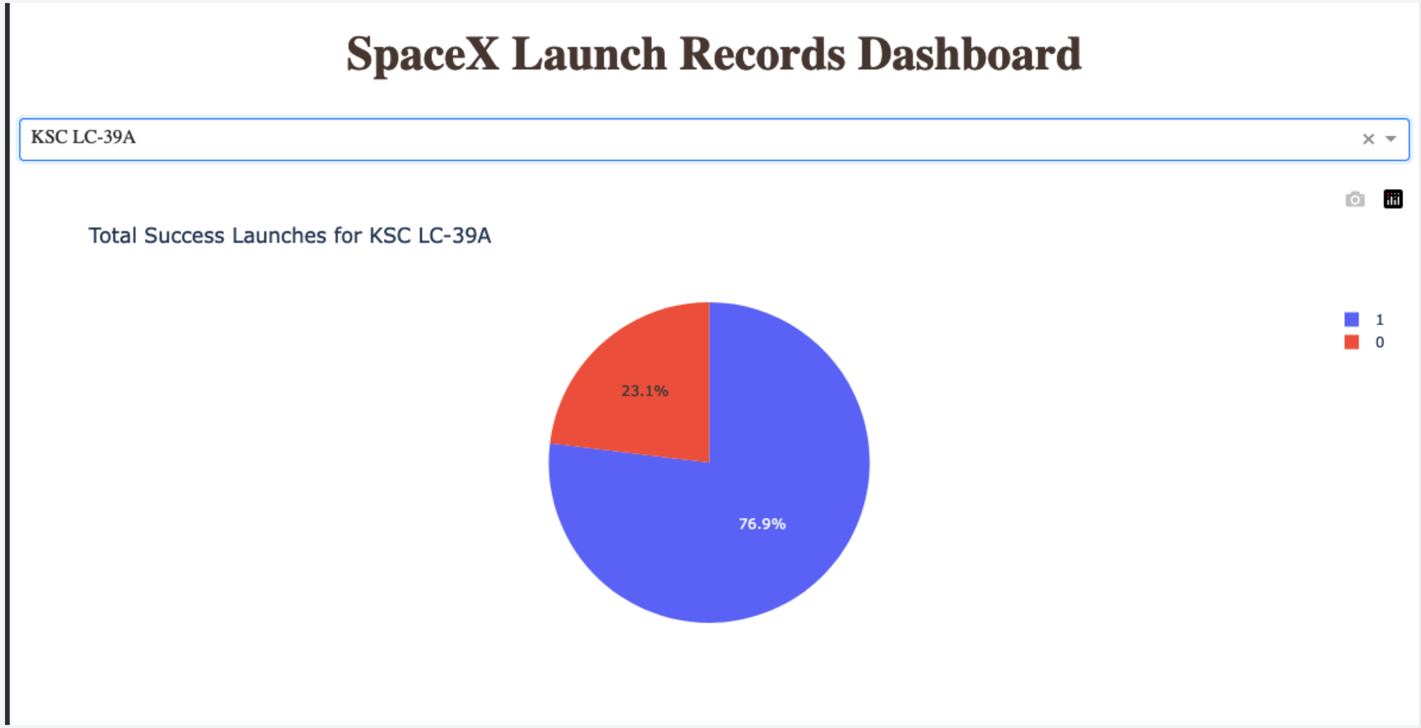
SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site

Launch site with highest launch success ratio

Explanation: KSC LC-39A has the highest launch success rate (76.9%)



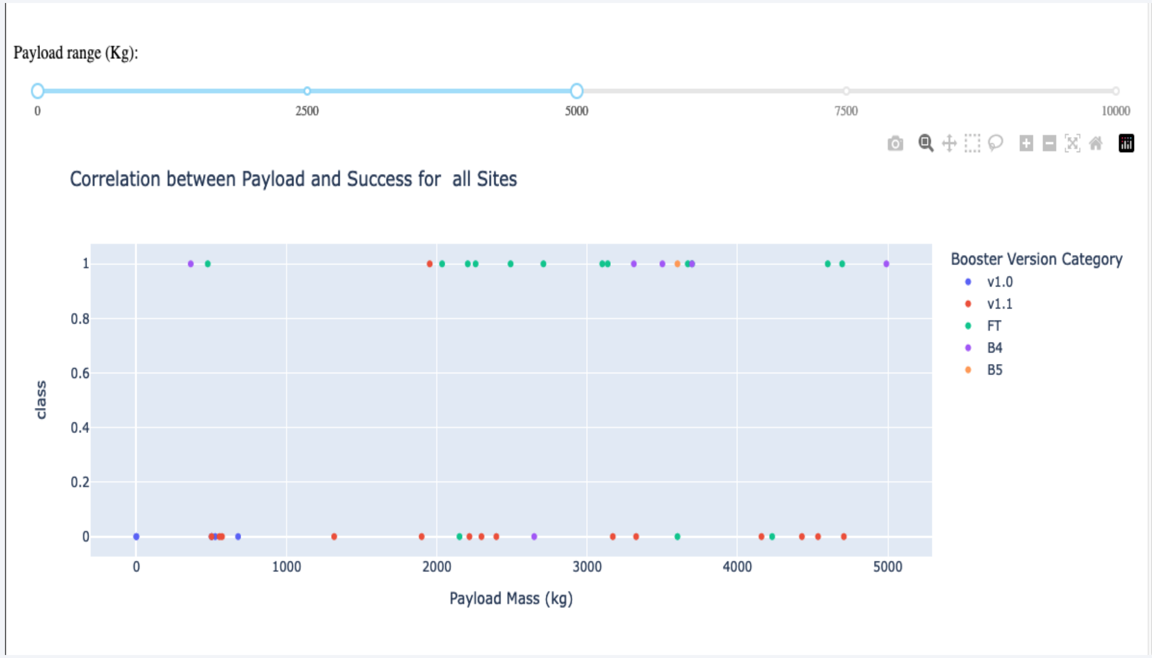
SpaceX Launch Records Dashboard

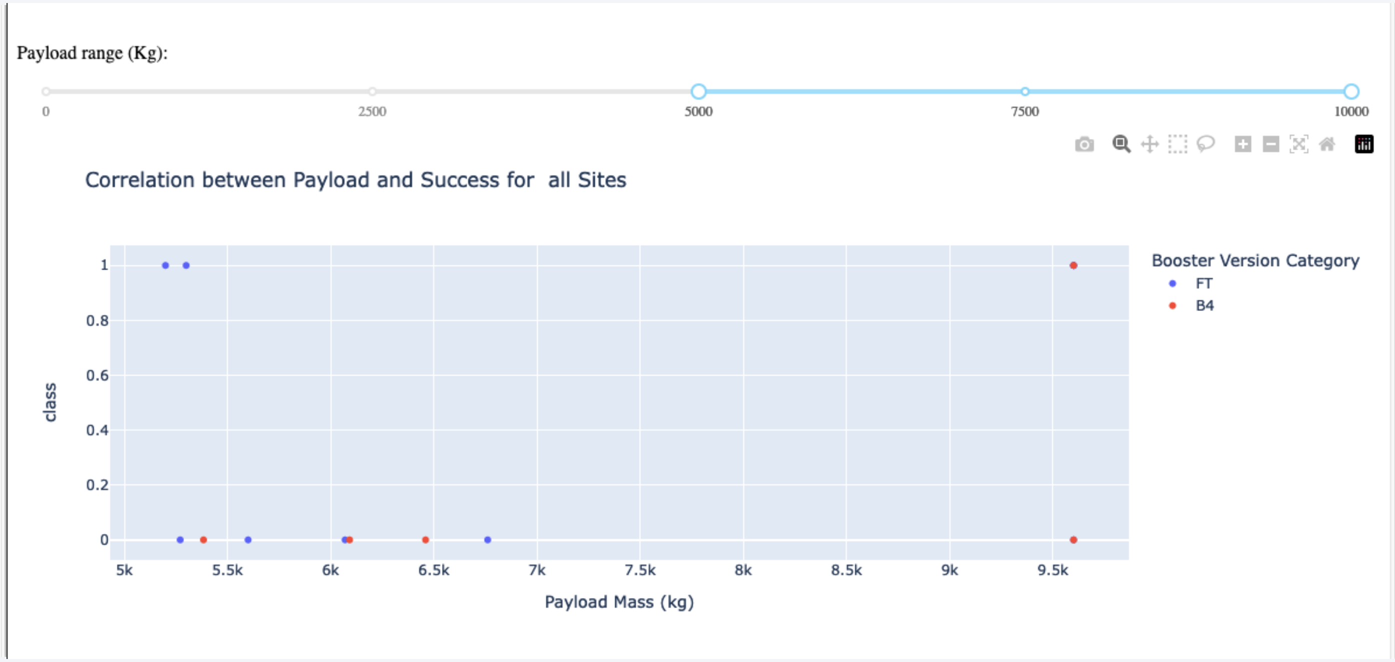
KSC LC-39A

Total Success Launches for KSC LC-39A

Payload vs. Launch Outcome for all sites

Explanation: the charts show that payloads between 2000 and 5500 kg have the highest success rate



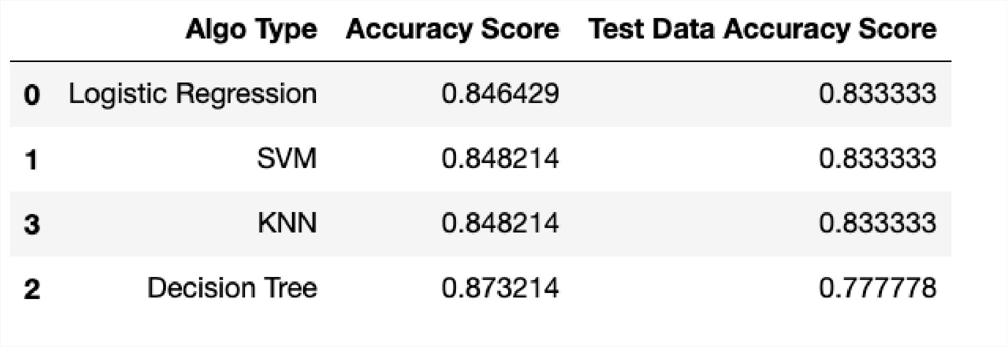


Section 5

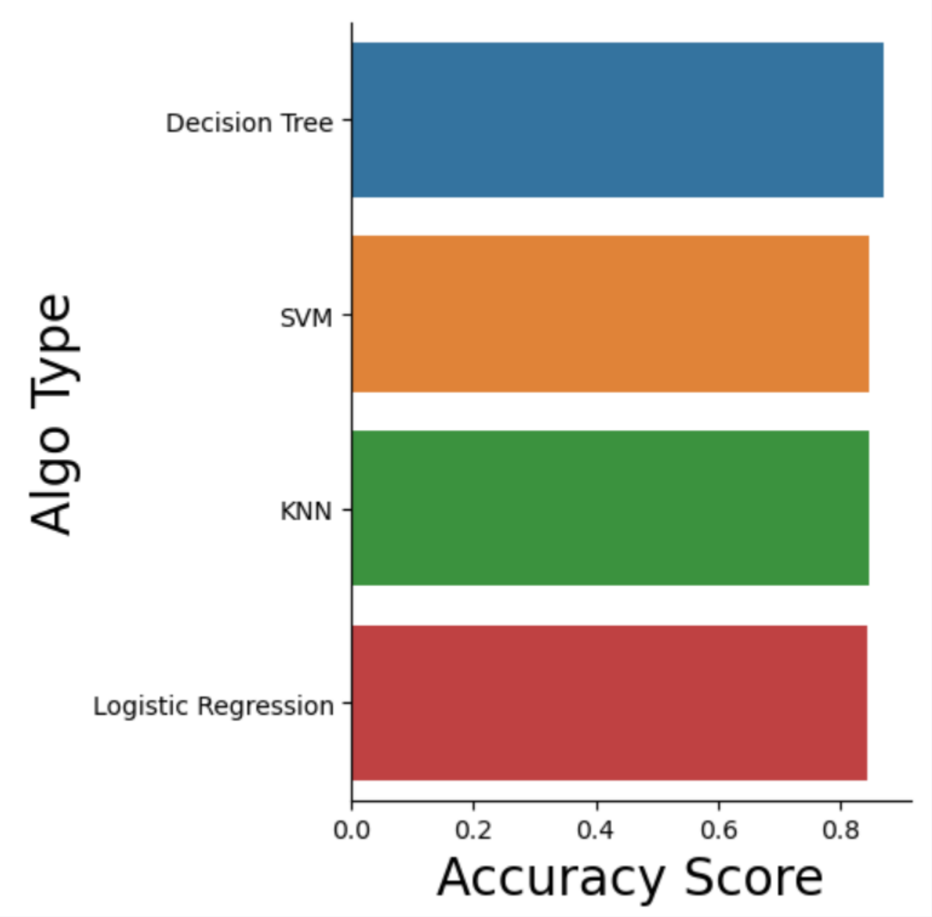
Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree model has the highest classification accuracy (87.32%) based on Accuracy Score
- However, Test Data Accuracy Score of Decision Tree model is the lowest among other models
- This could be due to size of data set. Bigger data set would be needed for model tuning



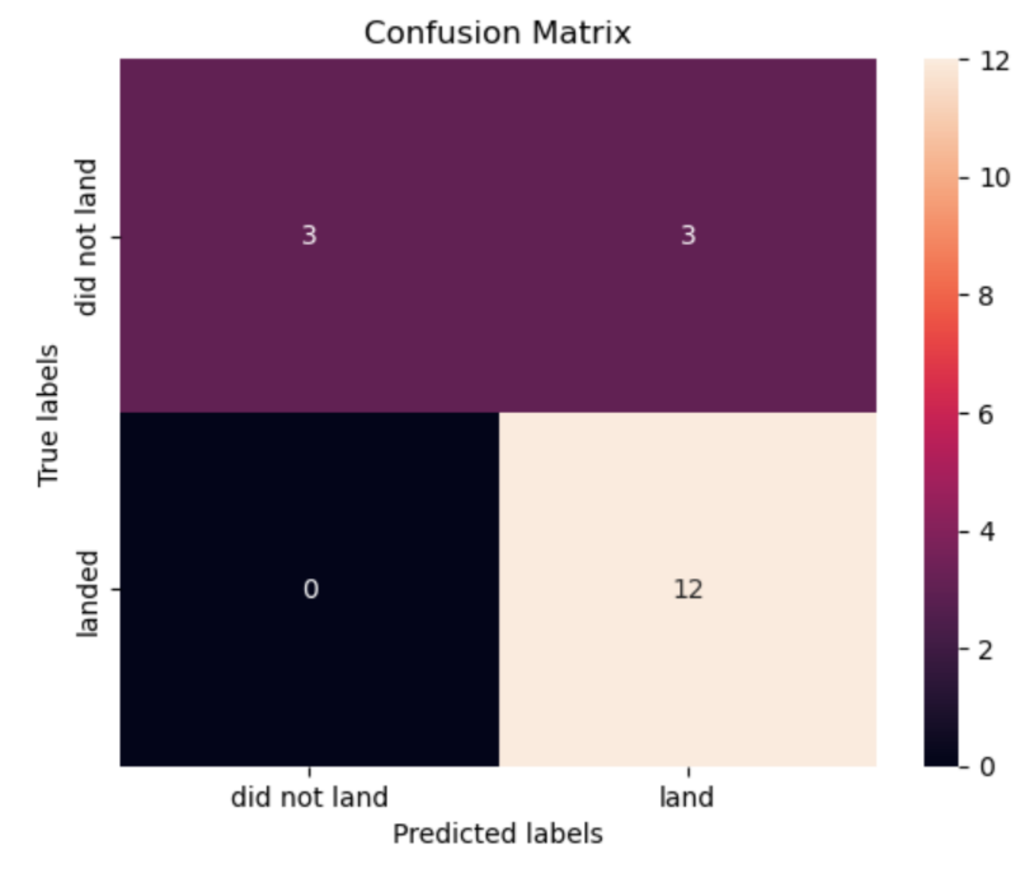
	Algo Type	Accuracy Score	Test Data Accuracy Score
0	Logistic Regression	0.846429	0.833333
1	SVM	0.848214	0.833333
3	KNN	0.848214	0.833333
2	Decision Tree	0.873214	0.777778



Confusion Matrix

Confusion matrix

- 12 true positive: predicted landing success, landed successfully
- 3 true negative (top left): predicted landing failure, landing failed
- 3 false negative (top right): predicted landing success, but landing failed
- 83 % accuracy of classifier overall ((TP + TN) / Total)
- 16.5% of misclassification or error rate ((FP + FN) / Total)



Conclusions

- Decision Tree is the best algorithm with an accuracy of about 87.5%
- Launches with higher payload mass have higher success rate in every launch site
- Launch success rate increased by about 80% from 2013 to 2020
- Site ‘KSC LC-39A’ has the highest launch success rate and ‘CCAFS SLC-40’ the lowest
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates (100%) and GTO the lowest
- Lunch sites are located strategically away from the cities and closer to coastline, railroads, and highways

Appendix

Thank you!