IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jared Kay
23 August 2024

Outline

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

Executive Summary

- Methodologies
  - Data Collection via API
  - Data Collection via Webscraping
  - Data Wrangling
  - Data Analysis with SQL
  - Data Analysis with Visualization
  - Interactive Data Visualization with Folium and Plotly
  - Data Predictions using Machine Learning
- Summary of Results
  - Data Analysis Results and Predictions
  - Data Visualizations
  - Predictive Analysis

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. This will be accomplished by utilizing data science methodologies to garner insights related to the real-world problem.
- We look to explore how launch site, payload mass, orbit type, and time affect the rates of success and failure of launches.
- Predictive models will also be evaluated for how well they can predict a successful launch based on these variables.

Section 1

Methodology

Methodology

Executive Summary

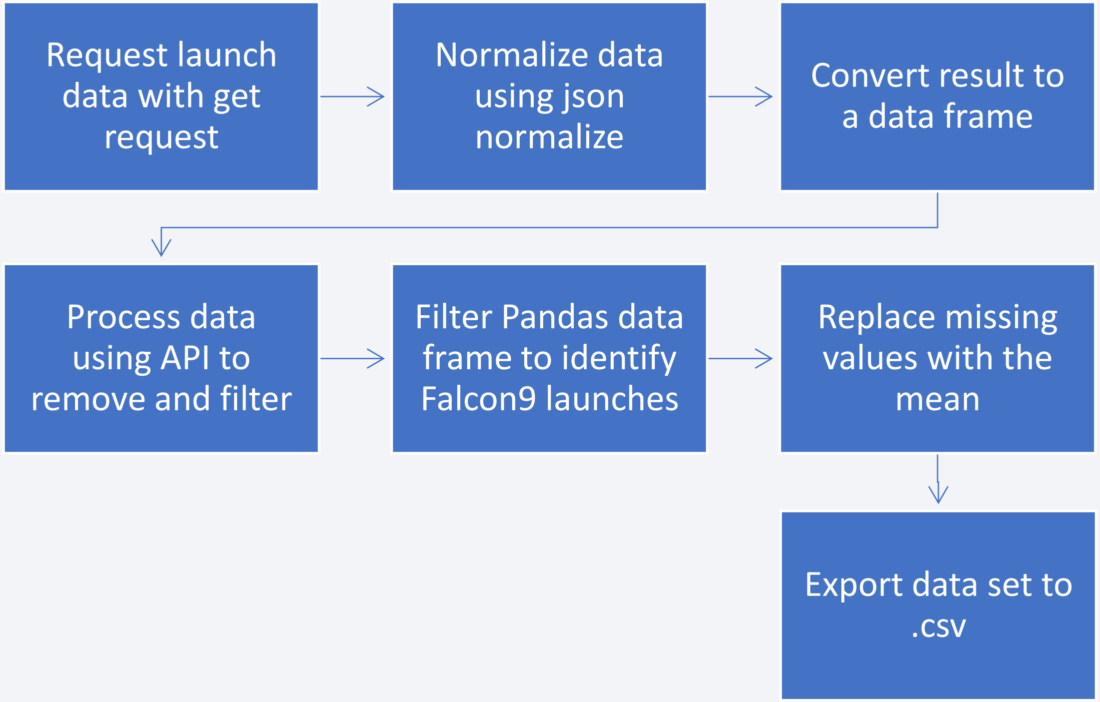
- Data collection methodology:
  - Data about launches was collected using the SpaceX REST API and web scraping using Beautiful Soup on relevant Wiki pages.
- Perform data wrangling
  - Data was processed and transformed into a cleaned data set using an API and sampling data while dealing with null values in the data set. One hot encoding was used when applicable to further make the data more manageable.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build and train predictive models before testing them to determine their accuracy

Data Collection

- Data was initially collected using the SpaceX REST API by performing a get request
- Results were viewed via the .json() method prior to using json_normalize transforming the data into a flat table
- Additional data was collected using web scraping of relevant Wiki sites using Beautiful Soup and converted to a Pandas data frame
- Once the data was collected it was transformed and cleaned by using APIs, data sampling, data filtering, and addressing null values
- One hot encoding was used where applicable to further make the data set more manageable for future analysis

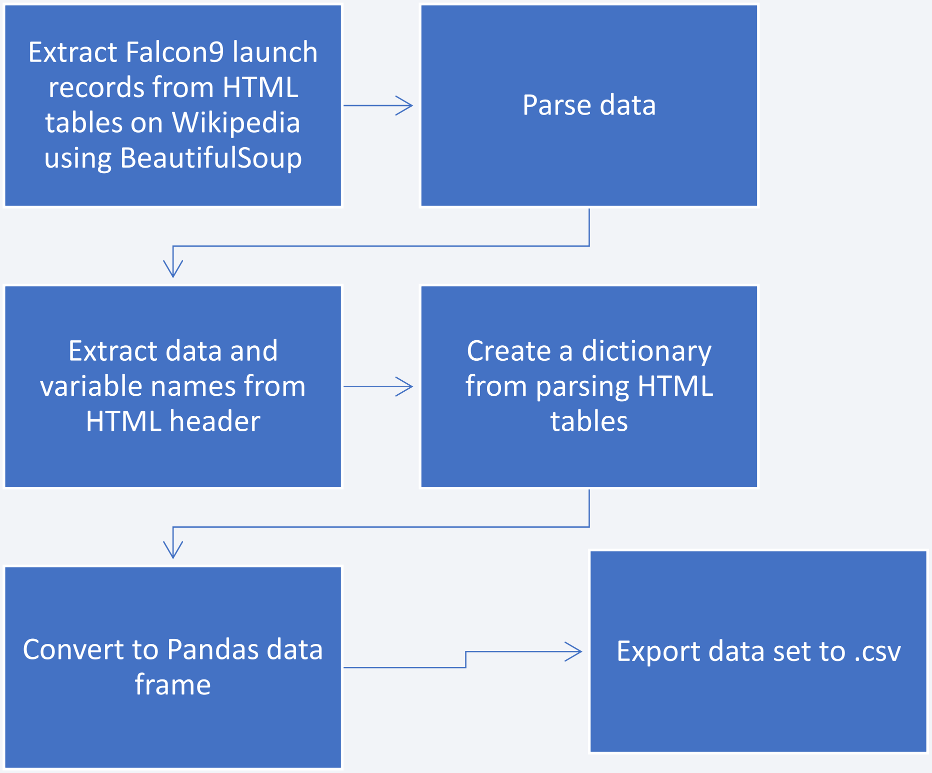
Data Collection – SpaceX API

- Request launch data from the SpaceX API using a get request
- Used json_normalize method to convert the result into a data frame
- Processed data using the API to remove irrelevant data and obtain information within a specific timeframe
- The resulting Pandas data frame was filtered to remove anything unrelated to Falcon9 launches
- Missing values were identified and addressed by replacing (.replace) them using the mean (.mean)
- Data was exported to a .cvs file for further analysis
- https://github.com/unknown-user-error-py/IBM-DataScience-Capstone/blob/main/1%20-%20spacex-data-collection.ipynb


Request launch data with get request
Normalize data using json normalize
Convert result to a data frame
Process data using API to remove and filter
Filter Pandas data frame to identify Falcon9 launches
Replace missing values with the mean
Export data set to .csv

Data Collection - Scraping

- Falcon 9 launch records were extracted from HTML tables on Wikipedia using Beautiful Soup
- Data was parsed prior to extracting column and variable names from the HTML header
- Created a dictionary by parsing the HTML tables prior to converting it to a Pandas dataframe
- Data was exported to a .csv file for future analysis
- https://github.com/unknown-user-error-py/IBM-DataScience-Capstone/blob/main/2%20-%20spacex-webscraping.ipynb



Extract Falcon9 launch records from HTML tables on Wikipedia using BeautifulSoup

Parse data

Extract data and variable names from HTML header

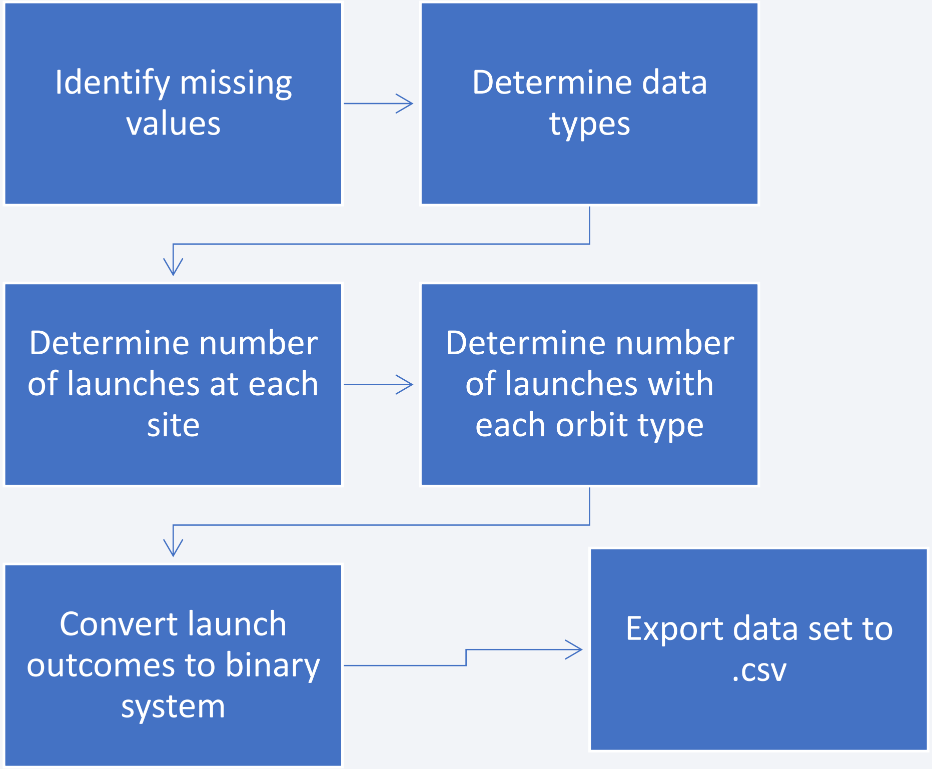
Create a dictionary from parsing HTML tables

Convert to Pandas data frame

Export data set to .csv

Data Wrangling

- Data was first evaluated to identify missing values and determine the data types
- The number of launches at each site was determined in addition to the number and occurrence of each orbit
- They outcome for these launches was also changed to a binary system using 0 bad outcomes and 1 for good outcomes
- https://github.com/unknown-user-error-py/IBM-DataScience-Capstone/blob/main/3%20-%20spacex-data-wrangling.ipynb



Identify missing values

Determine data types

Determine number of launches at each site

Determine number of launches with each orbit type

Convert launch outcomes to binary system

Export data set to .csv

EDA with Data Visualization

- The following charts were plotted:
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Success Rate vs Orbit Type
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
  - Success Rate vs Year
- These plots allow for visualization and identification of relationships between the compared data inputs

- https://github.com/unknown-user-error-py/IBM-DataScience-Capstone/blob/main/5%20-%20spacex-eda-data-visulization.ipynb

EDA with SQL

- SQL Queries Performed:
  - Names of unique launch sites
  - 5 records of launch sites beginning with ‘CCA’
  - Total payload mass in kg for boosters launched by NASA
  - Average payload mass in kg for booster version F9 v1.1
  - The date of the first successful ground pad landing
  - Names of the boosters than have had a successful drone ship landing with a payload between 4,000kg and 6,000kg
  - Total number of successful and failure mission outcomes
  - Names of boosters that have carried a maximum payload mass in kg
  - Records, including months and booster version, of failed landing outcomes on drone ships in 2015
  - Count of landing outcomes listed in descending order between 2010-06-04 and 2017-03-20

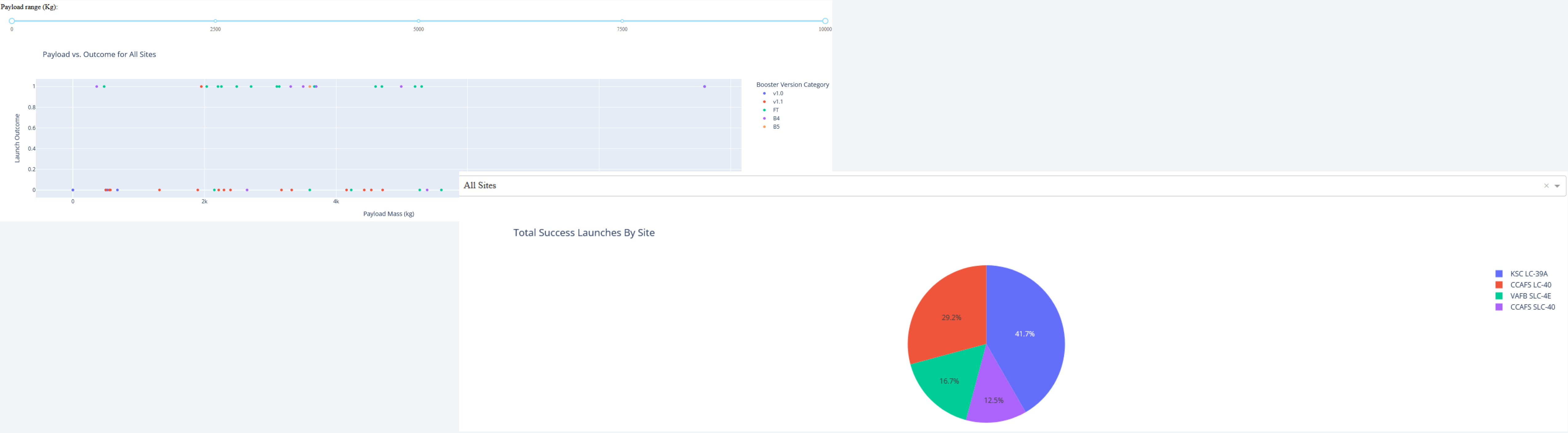
- https://github.com/unknown-user-error-py/IBM-DataScience-Capstone/blob/main/4%20-%20spacex-eda-sql.ipynb

Build an Interactive Map with Folium

- Created a blue circle at coordinates of NASA Johnson Space Center displaying both name and coordinates
- Created red circles for all launch sites displaying both name and coordinates
- Added green and red markers to indicate successful or failed launches at each site
- Added lines from launch site ‘CCAFS SLC-40’ to the nearest coastline, railroad, highway, and city displaying the distance to each in km
- These additions to the map allow for visualization of location and distance to their proximities to identify trends based on launch site location and each sites success rates
- https://github.com/unknown-user-error-py/IBM-DataScience-Capstone/blob/main/6%20-%20sapcex_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Plots to determine success rate of launches for each site in addition to all launches were included to visually analyze success rates
- Payload mass options were included to refine the weight and identify trends of successful launches when looking at booster version vs payload mass



- https://github.com/unknown-user-error-py/IBM-DataScience-Capstone/blob/main/7%20-%20spacex_dashboard.py

Predictive Analysis (Classification)

- Created a NumPy array using to_numpy() while assigning it to variable ‘Y’
- Data attached to variable ‘X’ was standardized using StandardScaler
- Split the data into test and training sets using train_test_split
- Created logistic regression and GridSearchCV with cv=10 to determine the best parameters
- Determined the best parameters using GridSearchCV for the following algorithms:
- Logistic regression, support vector machine, decision tree classifier, and k nearest neighbor
- Calculated accuracy for each using the test method – returning an accuracy score
- Plotted each on a confusion matrix to compare predicted vs true outcomes
- Determined the best performing model
- https://github.com/unknown-user-error-py/IBM-DataScience-Capstone/blob/main/8%20-%20spacex_machine_learning_prediction.ipynb

Results

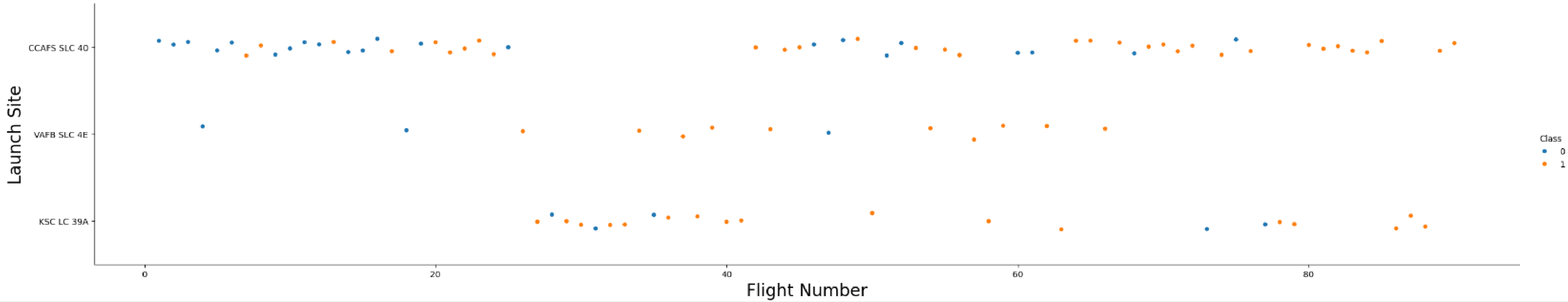
- Exploratory data analysis results
  - Launch success rate improved with time
  - Flights with heavier payloads are typically more successful
  - Orbit type ES-L1, GEO, HEO, SSO have 100% success rates
- Interactive analytics demo in screenshots
  - Launch sites are located near the equator and coastlines away from population centers
  - They maintain reasonable distance to highways and railroads presumably to transport items to launch sites
- Predictive analysis results
  - The decision tree performed best with a test accuracy of .944

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

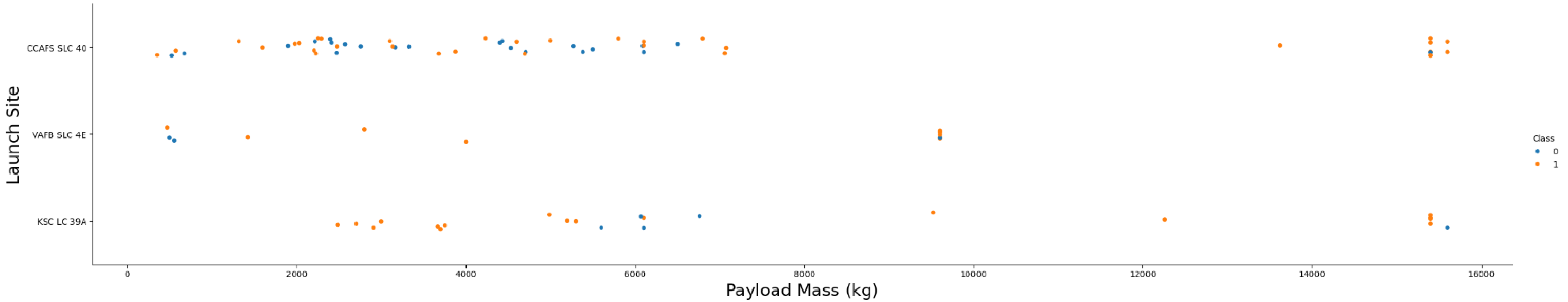
- Scatter Plot of Flight Number vs. Launch Site



- Earlier flights were more likely to be unsuccessful relative to later flights
  - Failed flights indicated by blue dots; successful flights indicated by orange dots
- Flights from the launch sites KSC LC-39A and VAFB SLC 4E were more likely to be successful than those from CCAFS LC-40

Payload vs. Launch Site

- Scatter Plot of Payload Mass (kg) vs Launch Site

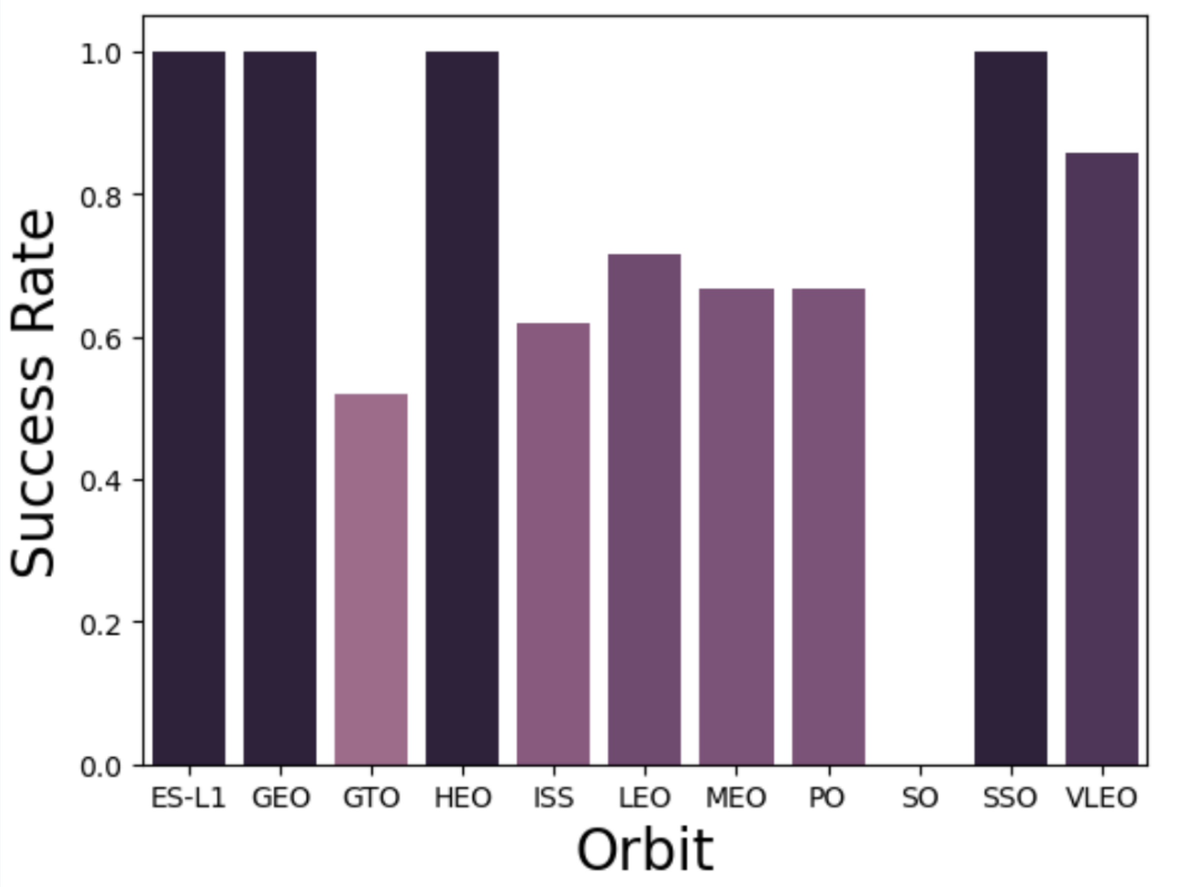


- Launches with heavier payloads were more likely to be successful relative to lighter payloads
  - Failed flights indicated by blue dots; successful flights indicated by orange dots
- Launch site VAFB SLC 4E has never had a launch with a payload greater than 10,000 kg
- Launch site KSC LC 39A has never had a failed launch with a payload less than ~5,000 kg

Success Rate vs. Orbit Type

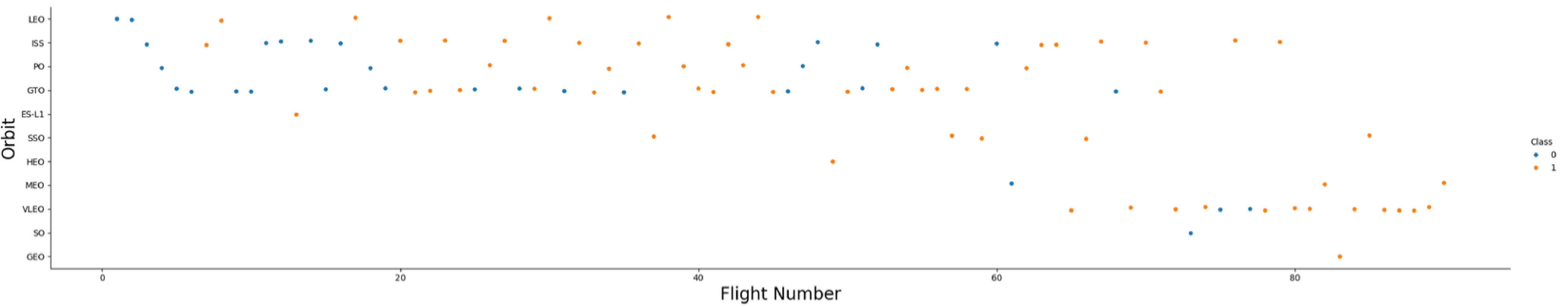
- Bar Plot of Success Rate vs Orbit Type

- 100% Success Rate
  - ES-L1, GEO, HEO, SSO
- 50%-85% Success Rate
  - GTO, ISS, LEO, MEO, PO, VLEO
- 0% Success Rate
  - SO



Flight Number vs. Orbit Type

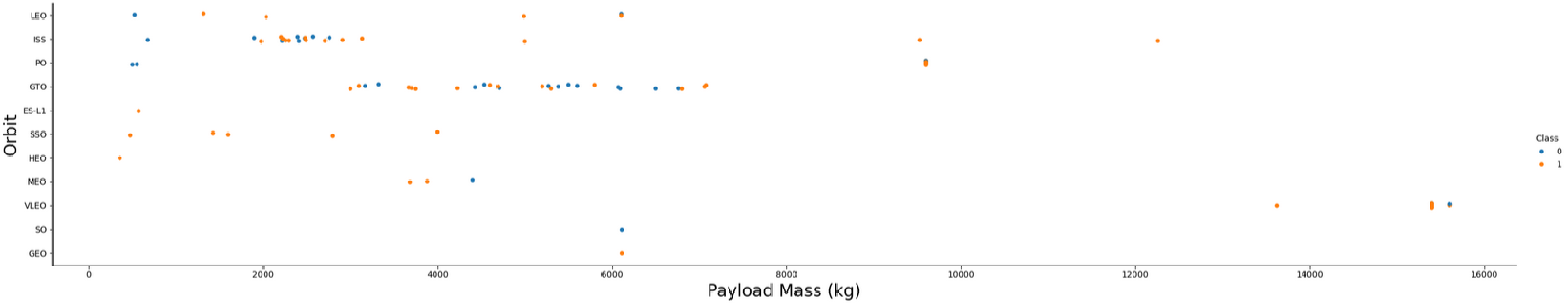
- Scatter Plot of Flight Number vs Orbit Type



- Earlier launches were more frequently conducted at launch sites: LEO, ISS, PO, GTO
- Later launches were more frequently conducted at launch sites: VLEO, SO, GEO
- Success rate typically increase with additional launches at a site, GTO appears to be the exception
  - Failed flights indicated by blue dots; successful flights indicated by orange dots

Payload vs. Orbit Type

- Scatter Plot of Payload Mass (kg) vs Orbit Type

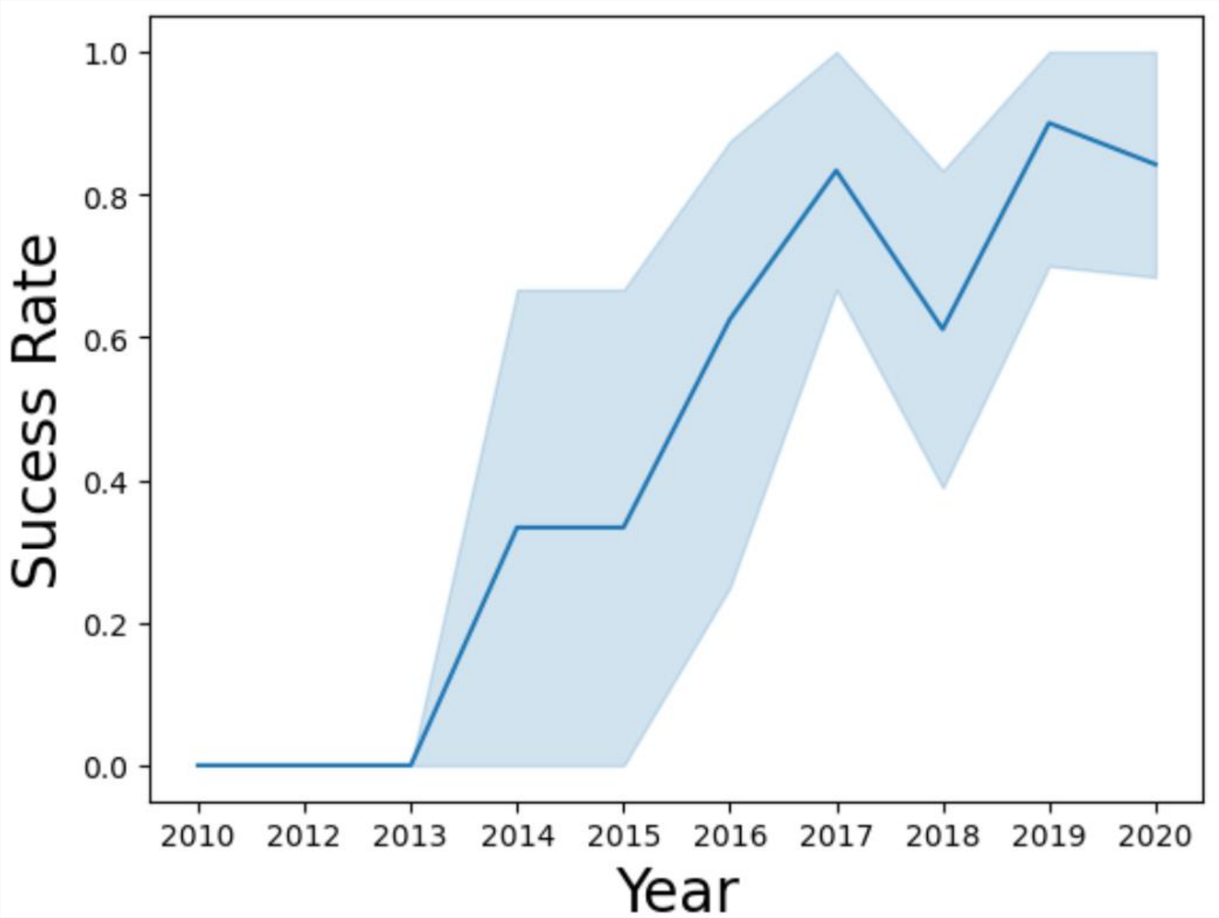


- Heavier payloads are more successful when orbit type is: LEO, ISS, PO
- Payload mass appears to have no impact on success rate with orbit type GTO
  - Failed flights indicated by blue dots; successful flights indicated by orange dots

Launch Success Yearly Trend

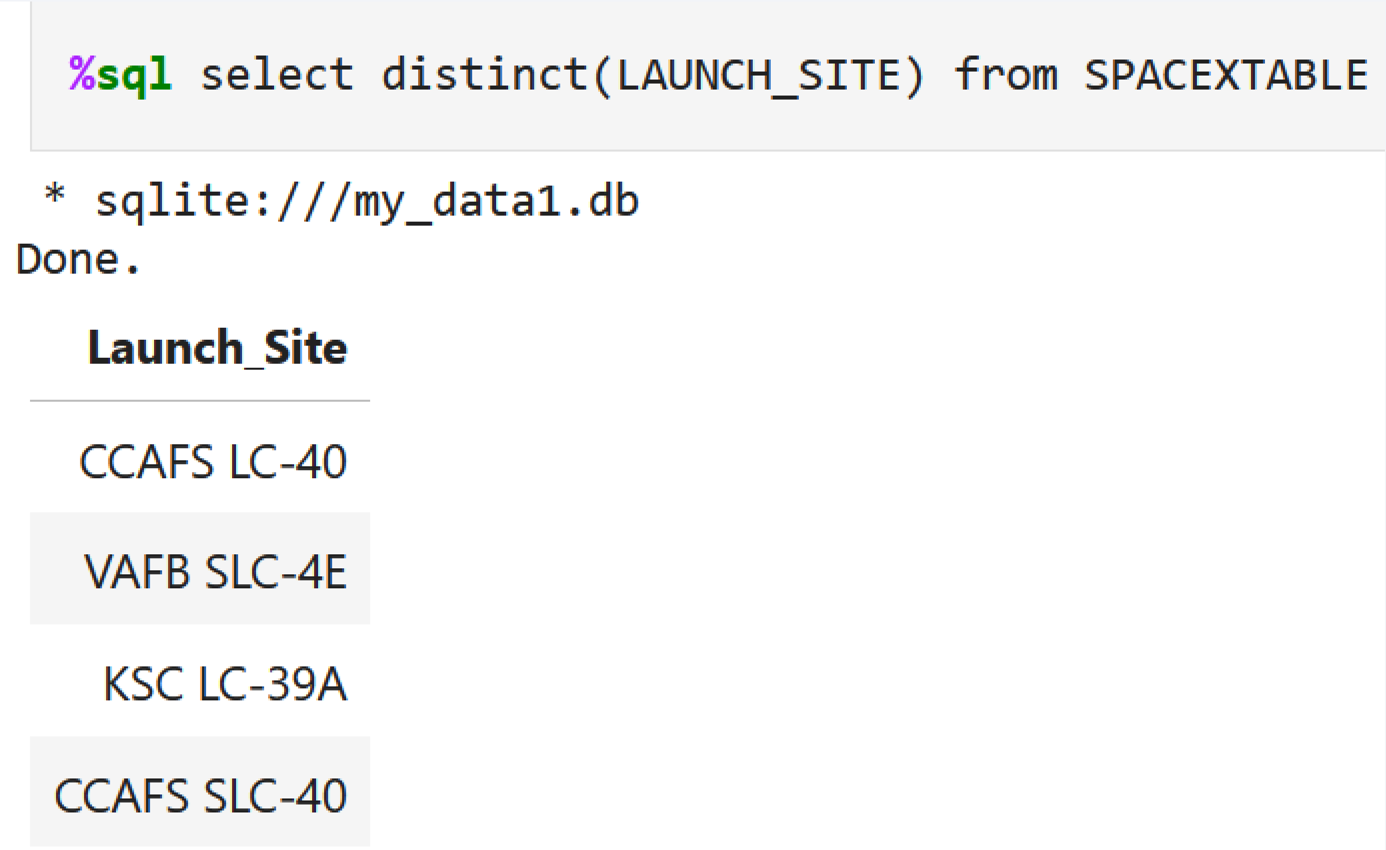
- Line Plot of Success Rate vs Year

- Success rate shows a positive correlation with later years
- There was a minor downturn in 2018 but the overall trend shows launches to be more successful as time goes on



All Launch Site Names

- The SPACEXTABLE was queried to request the names of distinct LAUNCH_SITE which returned the following four sites:
    - CCAFS LC-40
    - VAFB SLC-4E
    - KSC LC-39A
    - CCAFS SLC-40



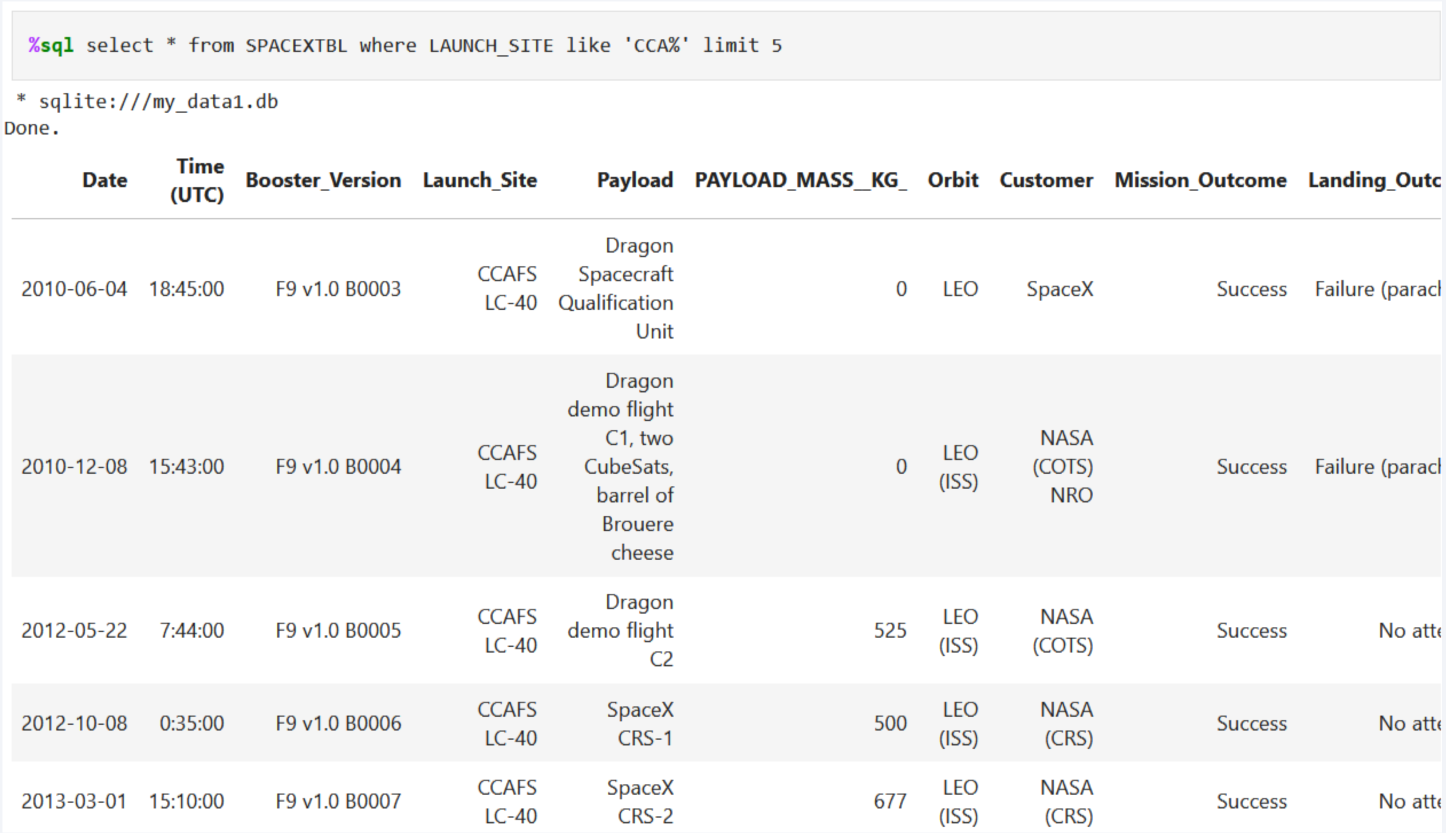
```
%sql select distinct(LAUNCH_SITE) from SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

- The SPACEXTBL was queried to request all information where the launch site began with the letters ‘CCA’ while limiting the number of results to 5



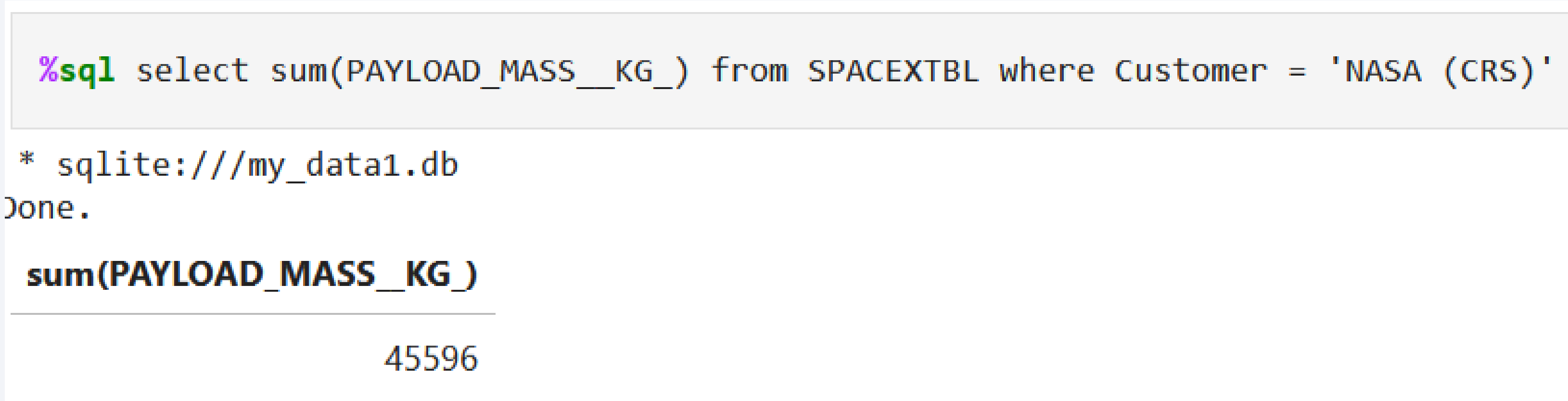
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (paracl
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 (paracl
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

- The SPACEXTBL was queried to select the sum (total) PAYLOAD_MASS_KG while only selecting those where the customer was NASA



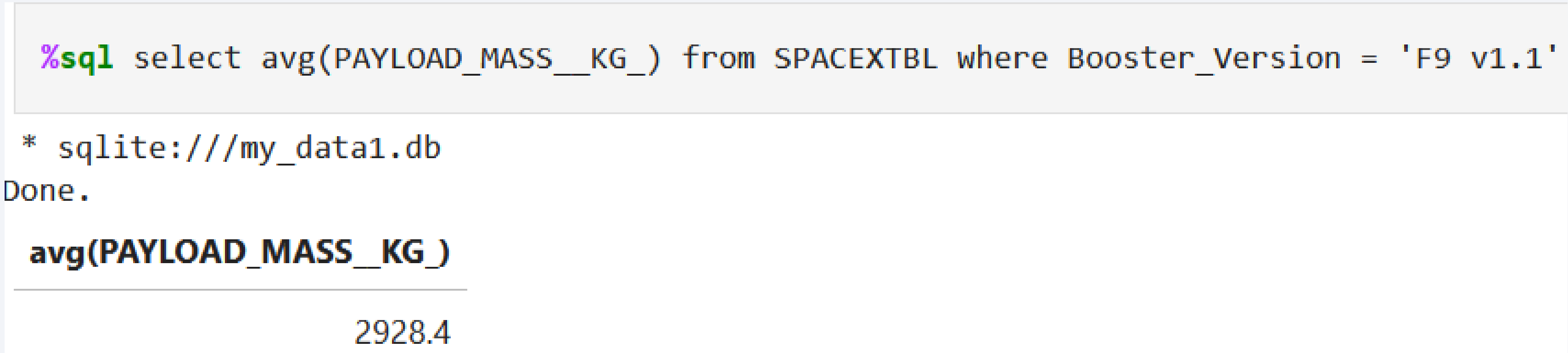
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The SPACEXTBL was queried to determine the average (avg) PAYLOAD_MASS_KG while only selecting those where the booster version was F9 v1.1



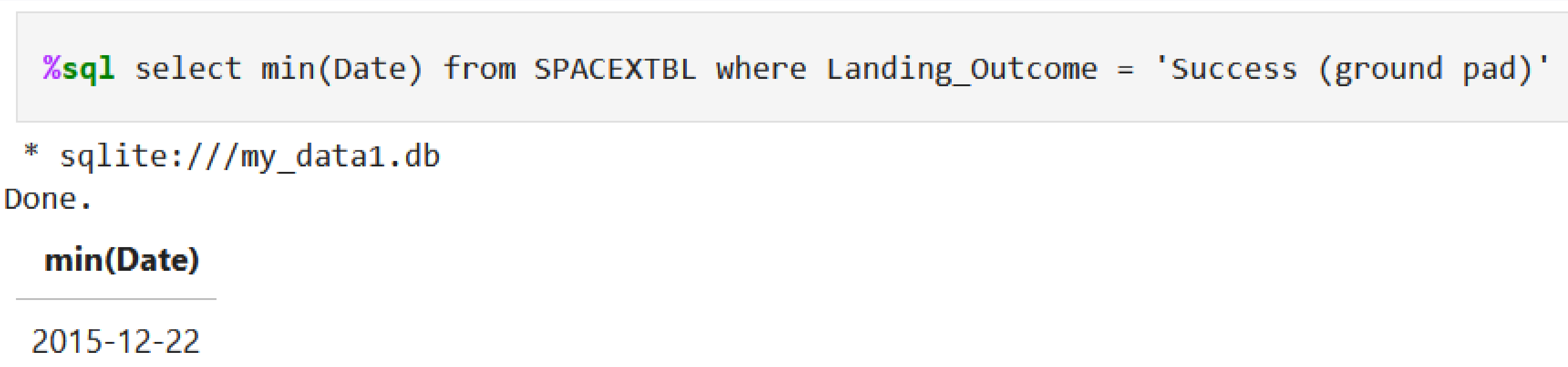
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The SPACEXTBL was queried to determine the min(Date), or earliest date, while only selecting those where the launch had a successful LANDING_OUTCOME on a ground pad



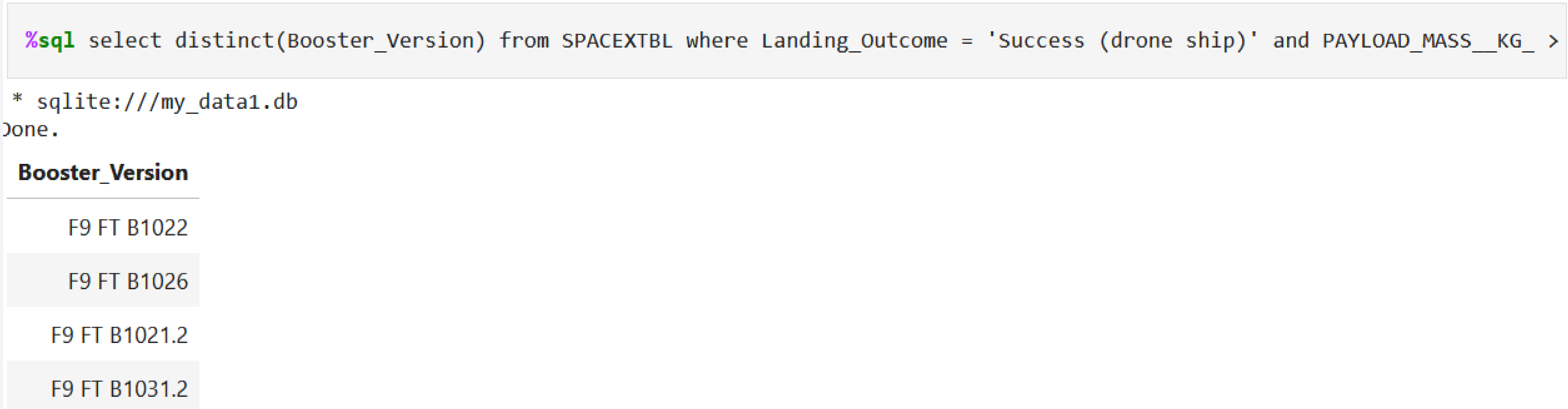
```
%sql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The SPACEXTBL was queried to determine which Booster_Version had a successful LANDING_OUTCOME on a drone ship while constraining the PAYLOAD_MASS_KG to be greater than 4,000 but less than 6,000



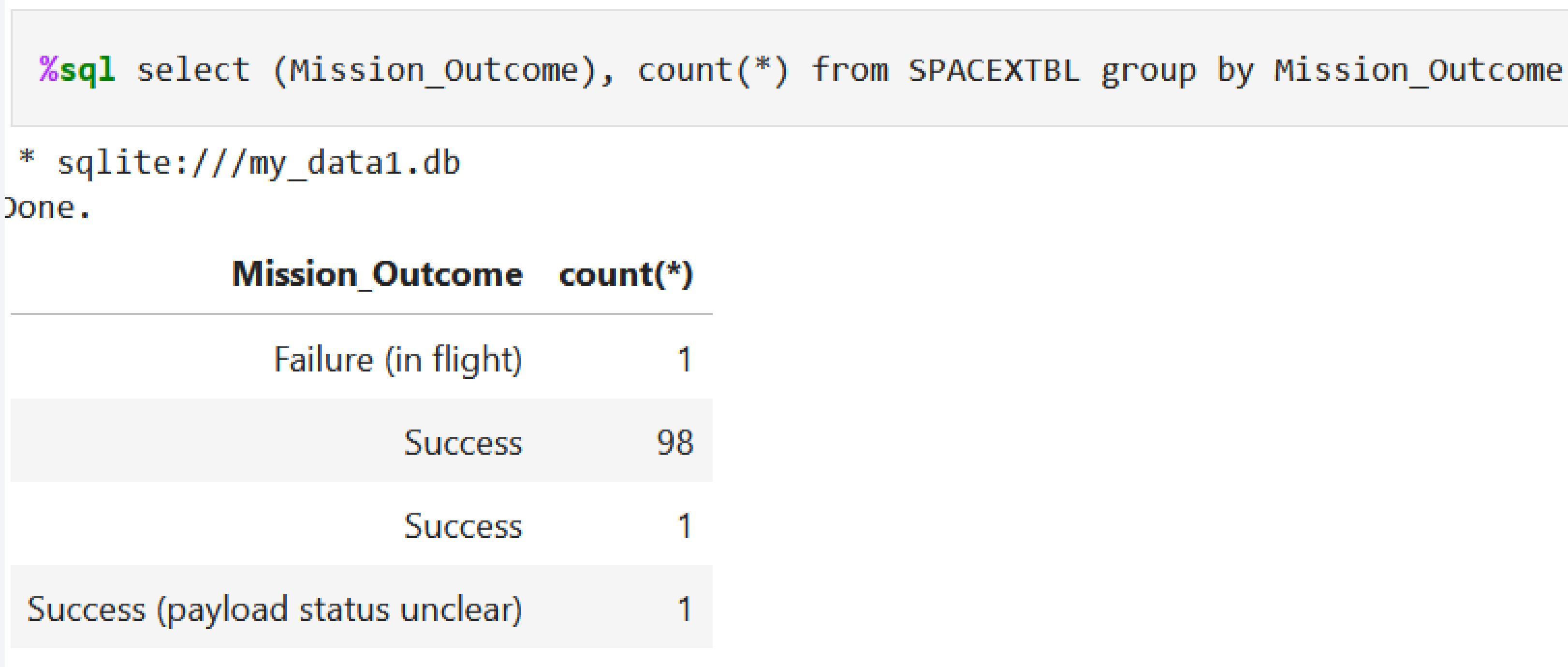
```
%sql select distinct(Booster_Version) from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ >
```

```
 * 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

- The SPACEXTBL was queried for a count of the Mission_Outcome – there is an issue in the result of counting one of the successful missions as a different outcome when the total should be 99



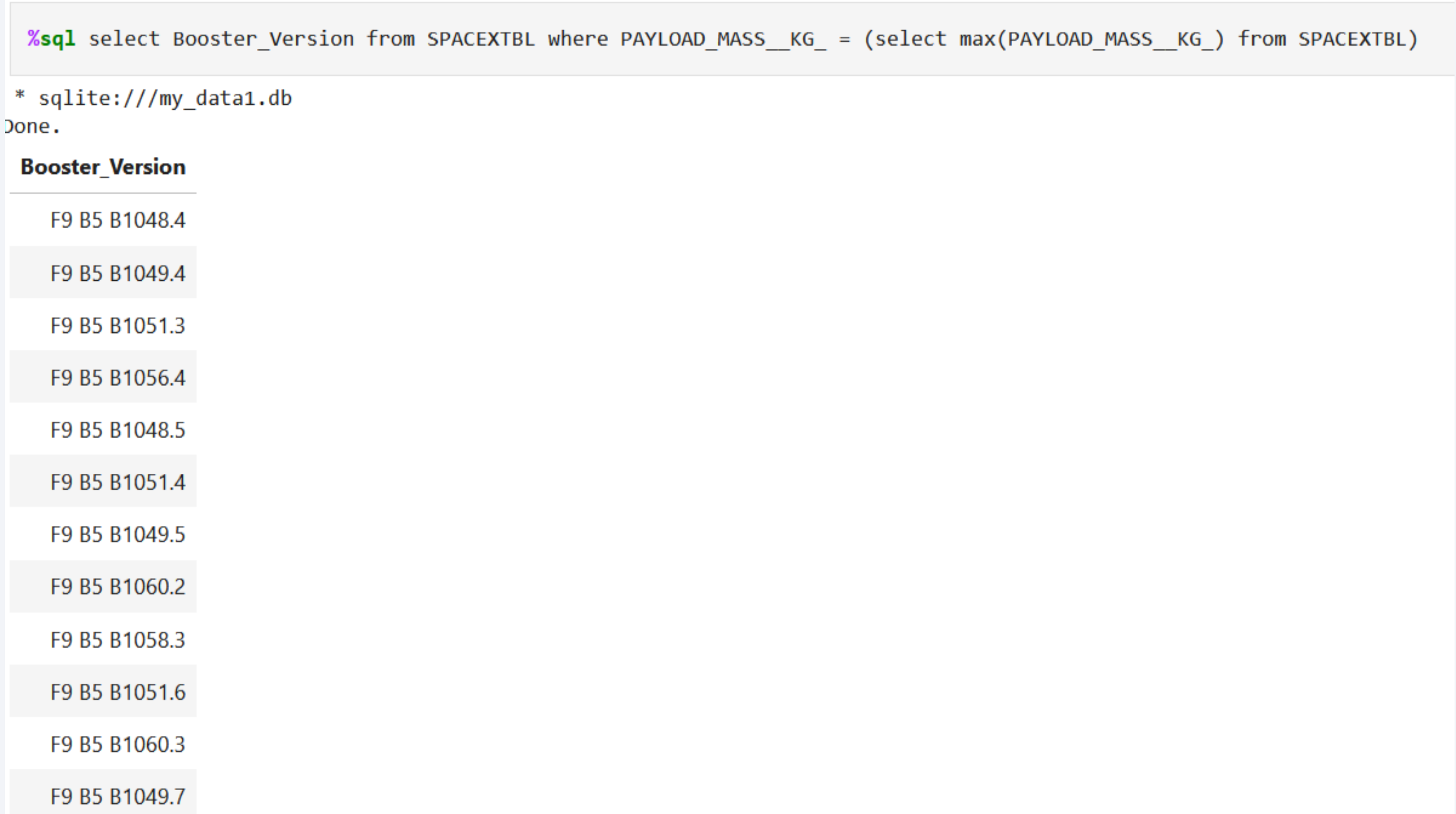
```
%sql select (Mission_Outcome), count(*) from SPACEXTBL group by Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- The SPACEXTBL was queried to find the Booster_Version that has carried a PAYLOAD_MASS_KG equal to the max(PAYLOAD_MASS_KG) found in the table



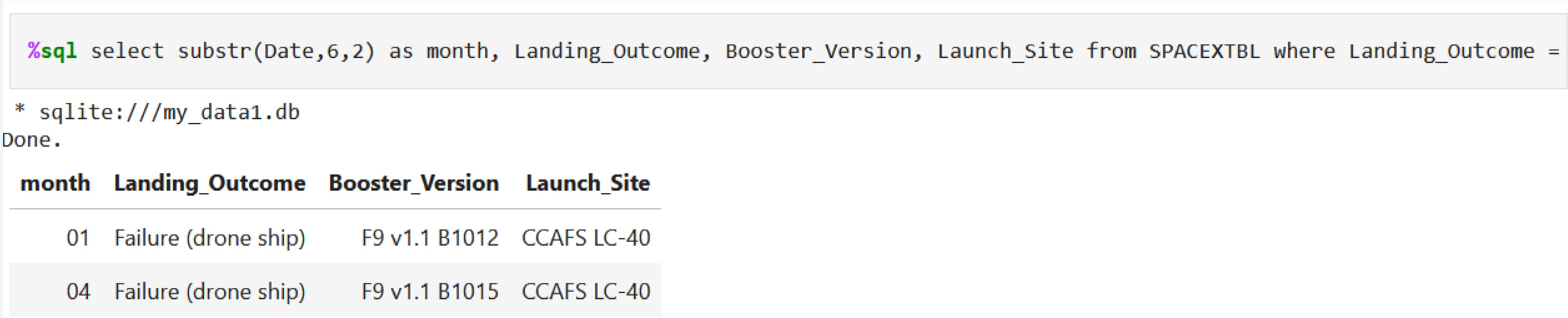
```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * 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

- The SPACEXTBL was queried to find the records pertaining to failed launches during the year 2015



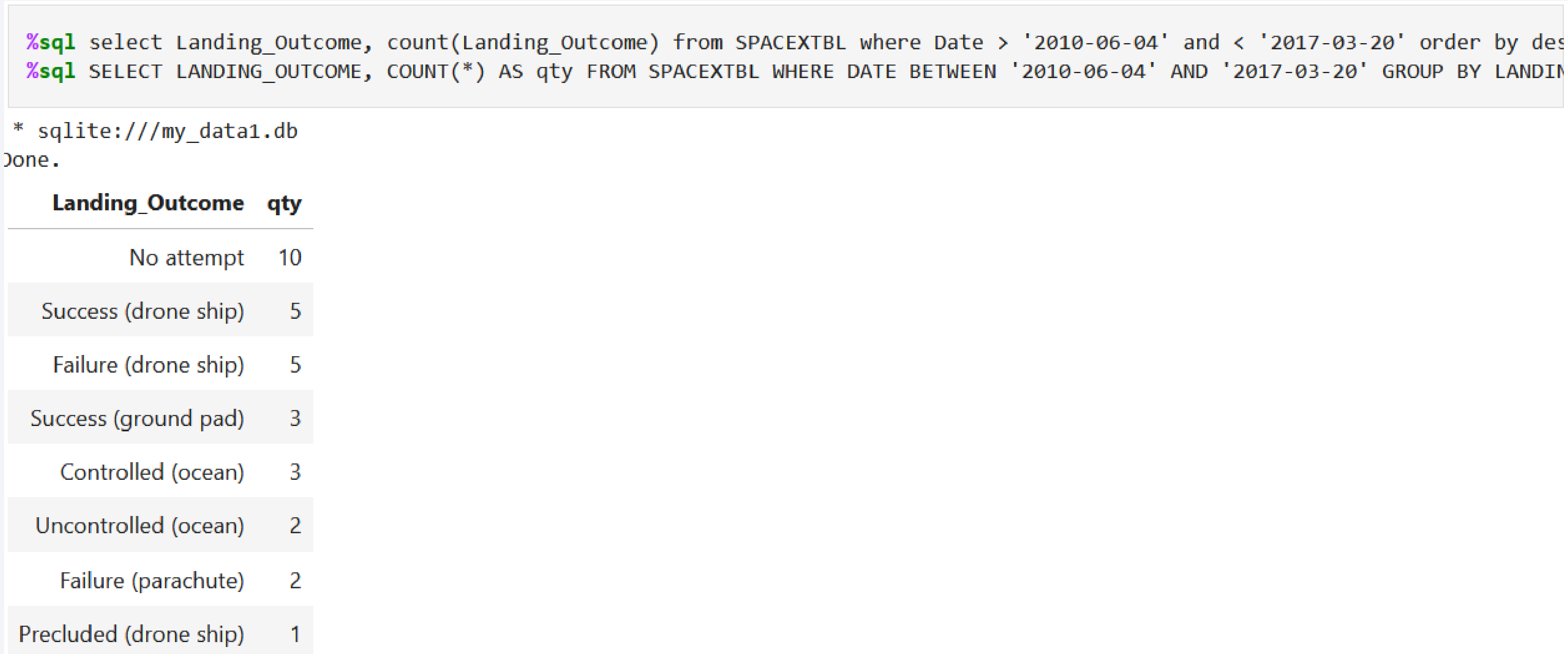
```
%sql select substr(Date,6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome =
```

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

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

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

- The SPACEXTBL was queried to count the Landing_Outcome between the dates of 2010-06-04 and 2017-03-20 and ranking them in descending order



```
%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTBL where Date > '2010-06-04' and < '2017-03-20' order by des
%sql SELECT LANDING_OUTCOME, COUNT(*) AS qty FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDIN
```

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

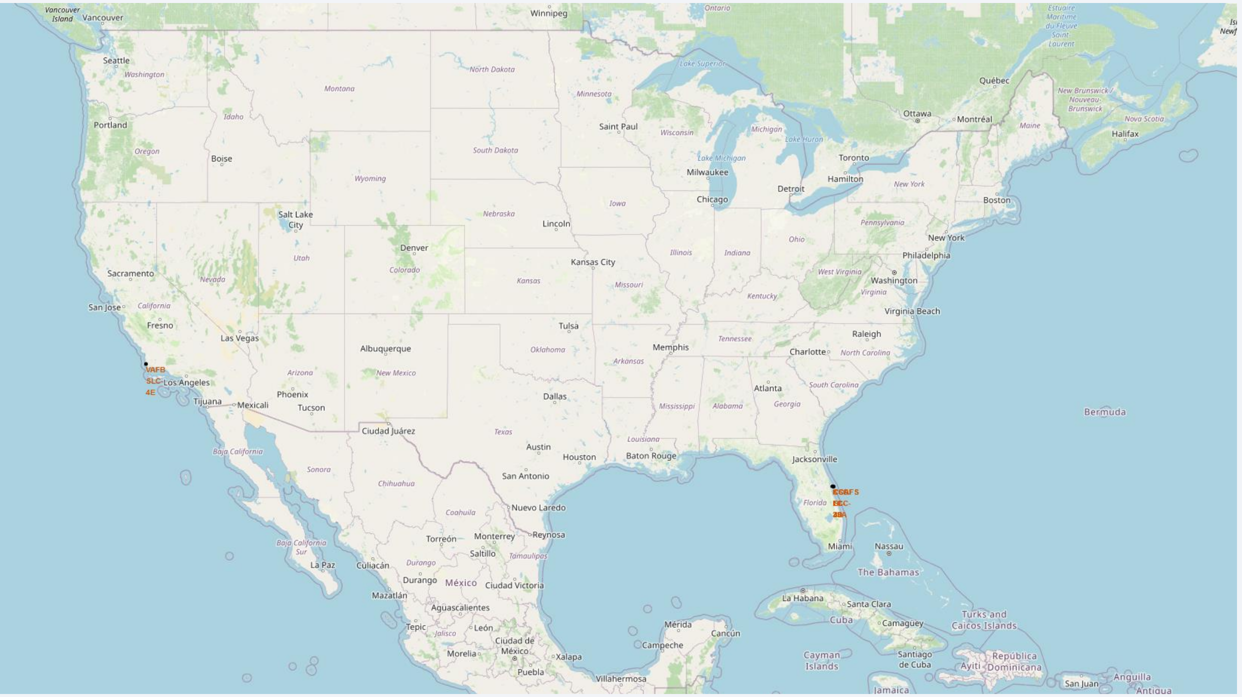
Landing_Outcome	qty
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

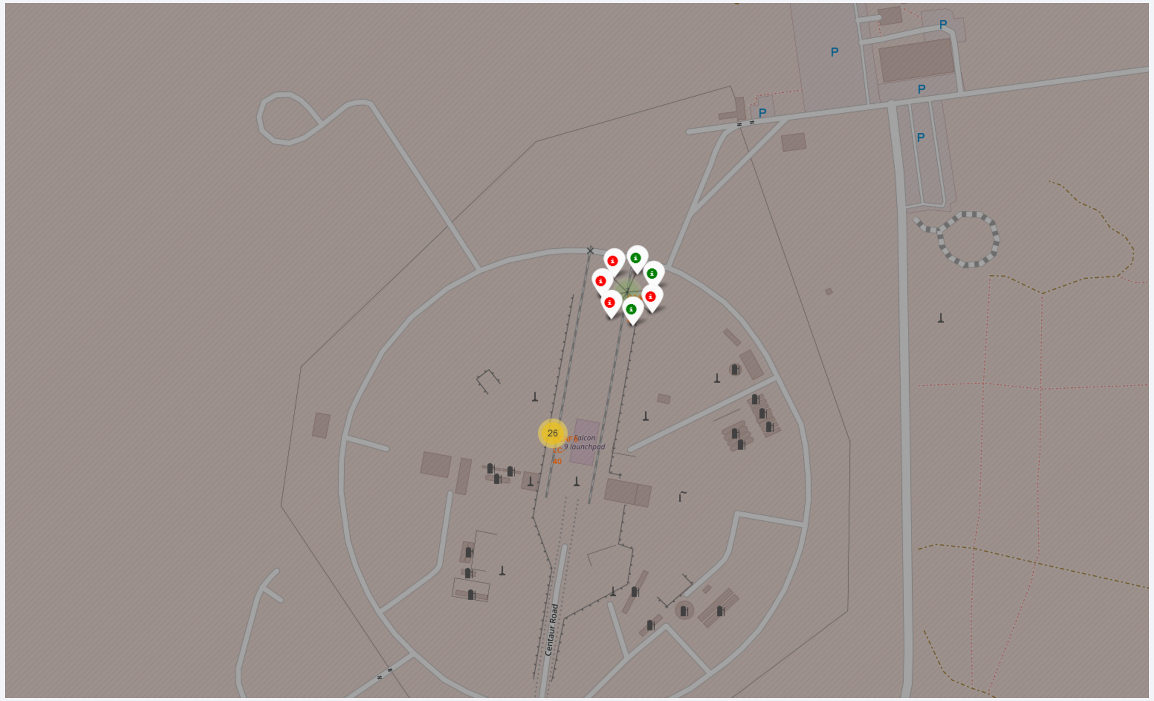
Launch Site Locations

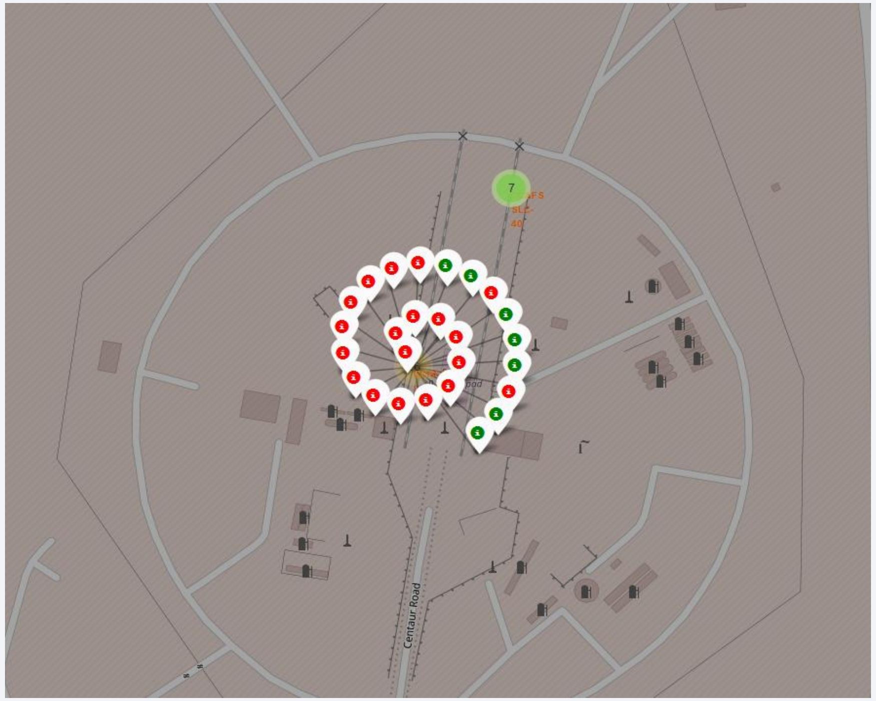
- The launch sites are located close to the equator to take advantage of Earth's rotational speed. The Earth rotates at a speed >1650km/hr and this can be applied to the speed required to orbit the Earth. This cuts down on the amount of propellant needed to launch a spacecraft.
- The launch sites are also located near the coasts far from population centers to reduce the impact to people accounting for the chance it may experience a catastrophic event.



Launch Outcomes

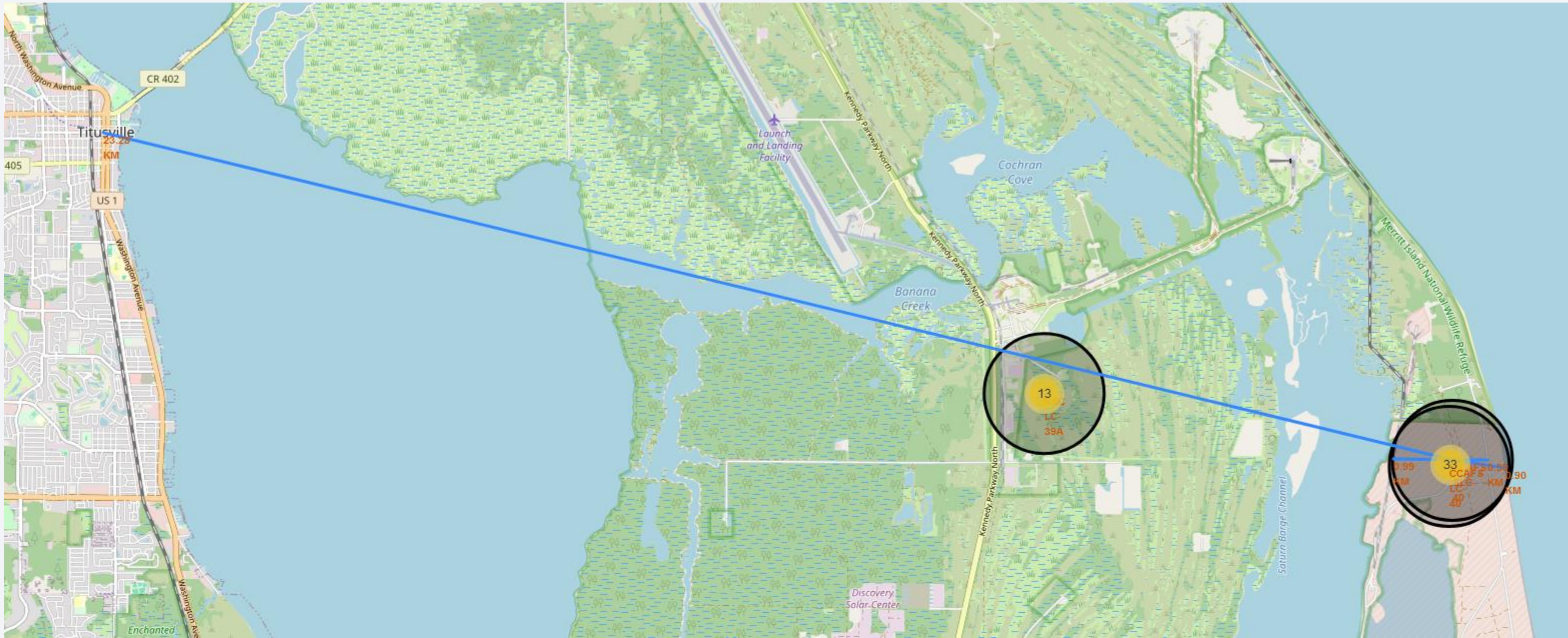
- The map displays launches at each site color coded to display the outcome
  - Colors indicate Successful or Unsuccessful
- Using the colors it is easier to identify sites with high or low success rates

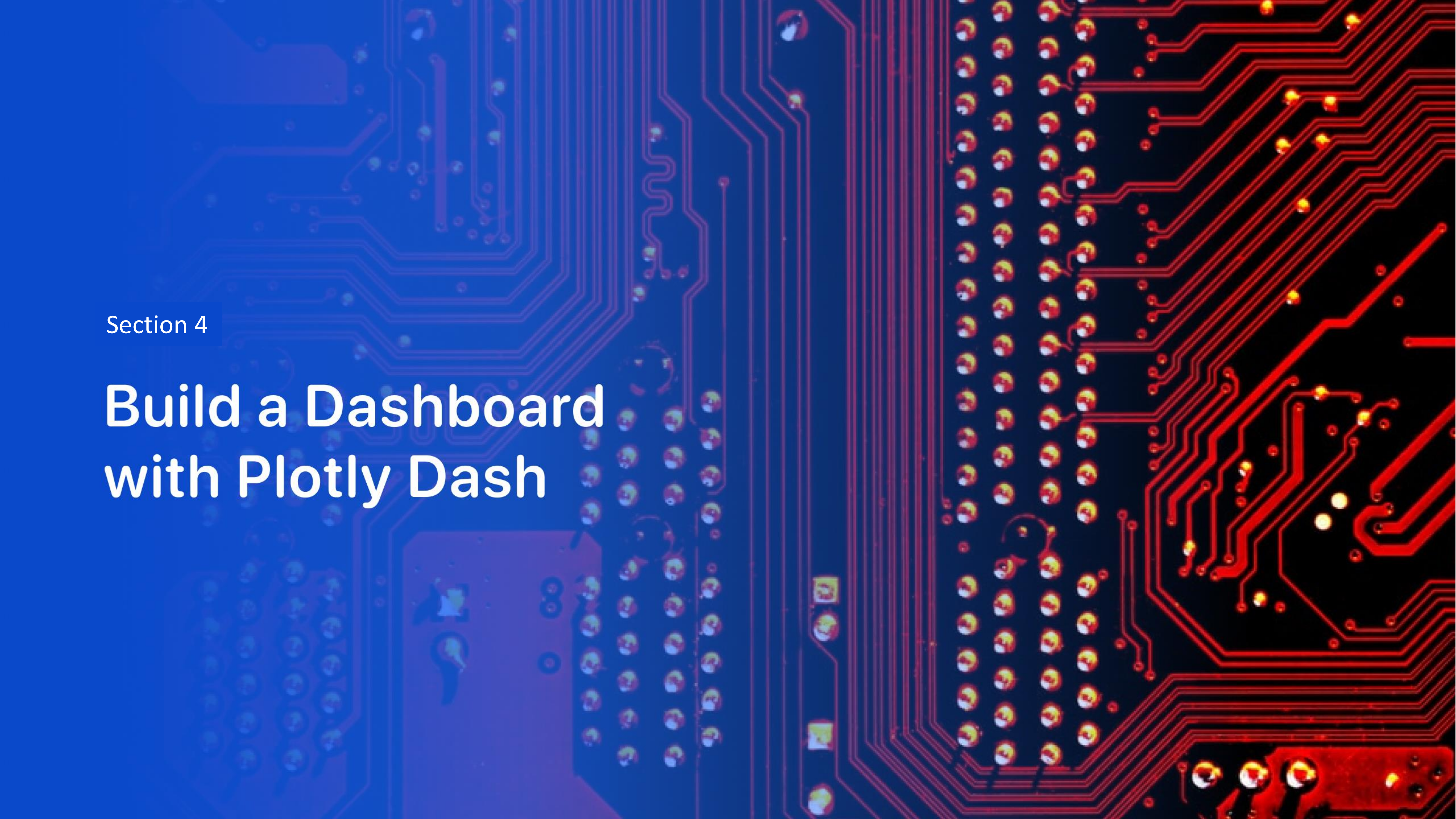




Distance of Launch Site to Proximities

- Launch site CCAFS SLC-40 is located:
- 0.90km to the coastline, 0.59km to the nearest highway, 0.99km to the nearest railroad, and 23.28km to the city of Titusville
- The launch site is located away from a population center and near the coast while maintain adequate distance to highways and railroads to move materials and people to the site as needed



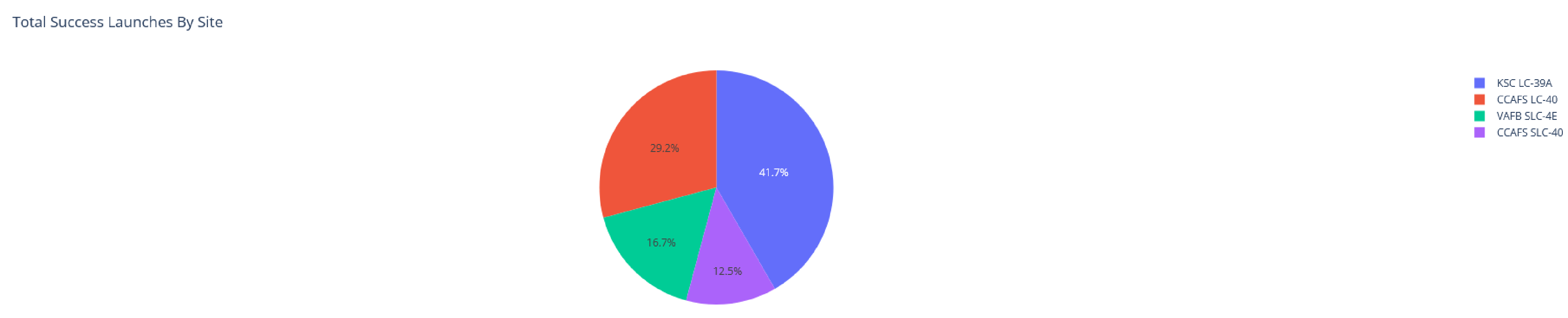


Section 4

Build a Dashboard with Plotly Dash

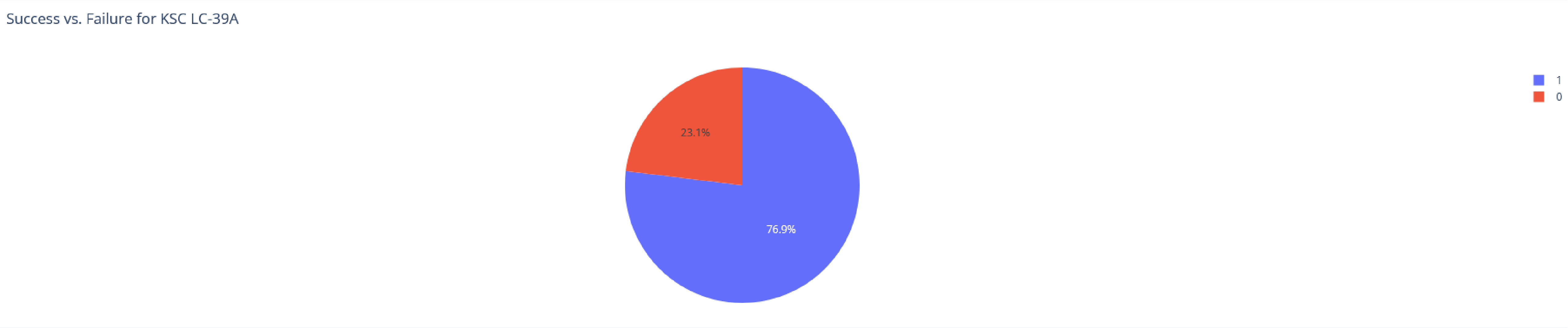
Total Successful Launches – All Sites

- The pie chart shows the percentage of successful launches for each site
- The most successful site is KSC LC-39A with 41.7% of all successful launches coming from this site
- The least successful site is CCAFS SLC-40 with 12.5% all successful launches coming from this site



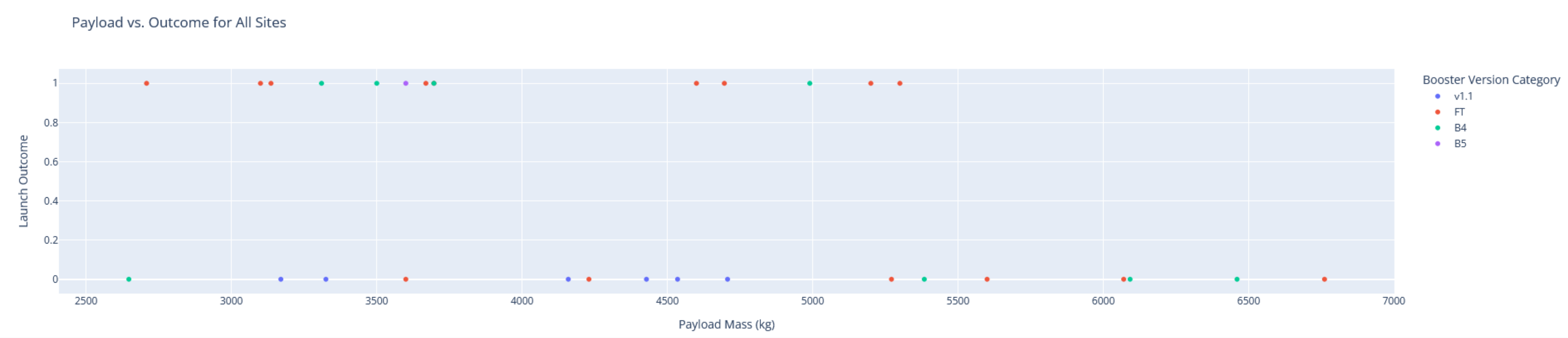
Success & Failure Percentage for Site KSC LC-39A

- KSC LC-39A has a 76.9% successful launch rate while 23.1% of launches fail



Launch Success w. Payload >2,500kg and <,7,500kg

- Booster version FT is the most frequently used for payload mass between 2,500kg and 7,500kg
- Booster version v1.1 always fails for payload mass between 2,500kg and 7,500kg
- Booster version v1.1 and B5 are more frequently used <5,000kg

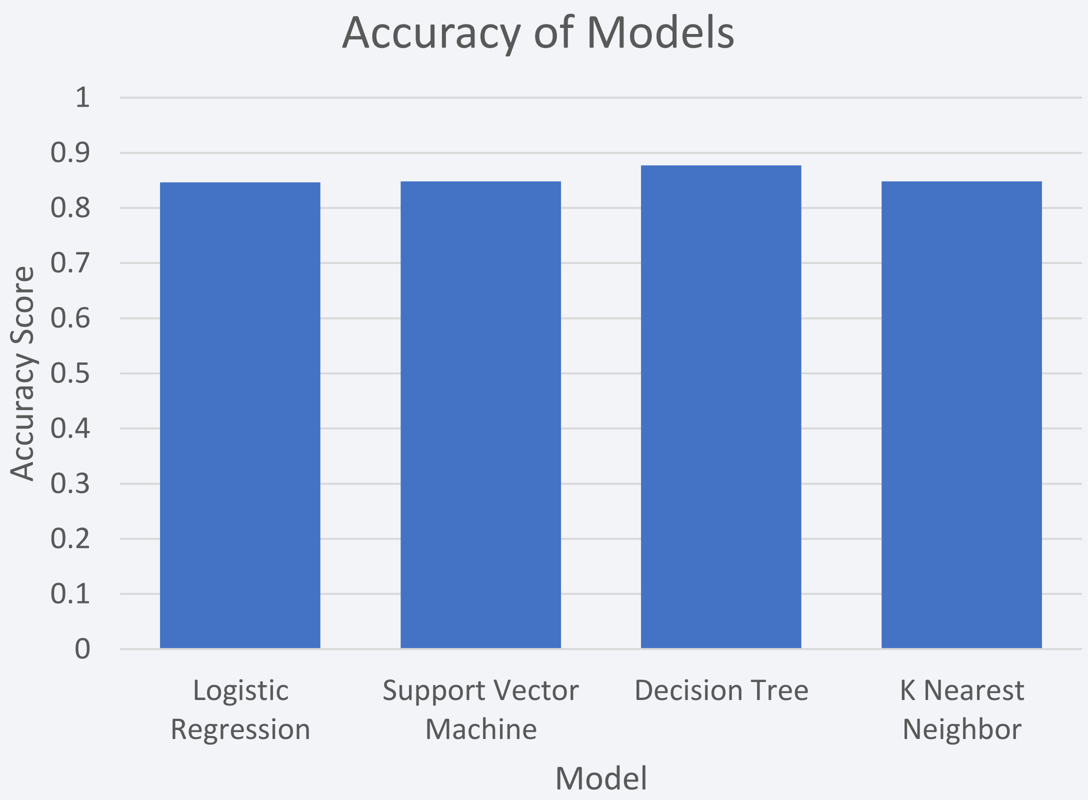


Section 5

Predictive Analysis (Classification)

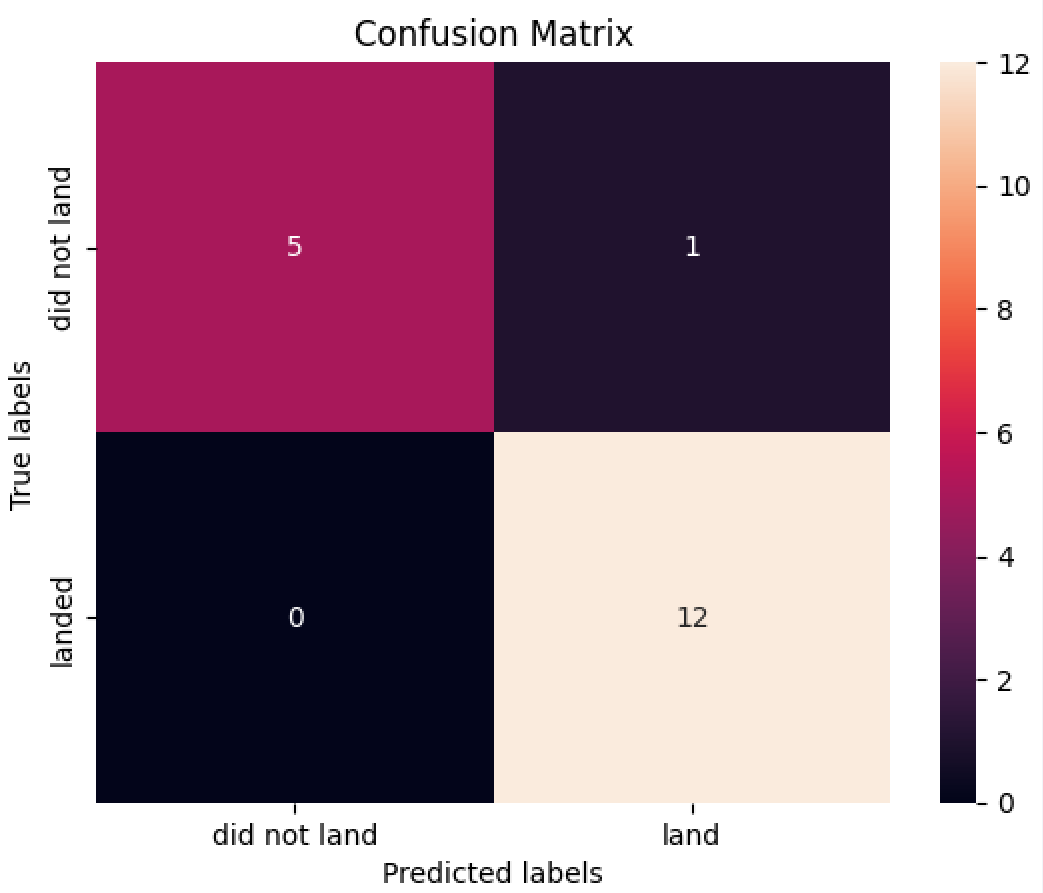
Classification Accuracy

- Decision Tree Classification returned the highest accuracy score for both training and test data sets



Confusion Matrix

- The Decision Tree Classification performed best compared to other models tested
- The confusion matrix showed that this model predicted
  - 12 successful landings of actual successful landings, true positive
  - 5 failed landings that were actual failures, true negative
  - 1 successful landing that was actually a failed landing, false positive



Conclusions

- Earlier flights were more likely to be unsuccessful relative to later flights
- Flights from the launch sites KSC LC-39A and VAFB SLC 4E were more likely to be successful than those from CCAFS LC-40
- Launches with heavier payloads were more likely to be successful relative to lighter payloads
- ES-L1, GEO, HEO, SSO orbit types are the most successful
- Success rate typically increase with additional launches at a site, GTO appears to be the exception
- The launch site is located away from a population center and near the coast while maintaining adequate distance to highways and railroads
- The Decision Tree Classification performed best compared to other when predicting a successful launch

Thank you!