IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marcelo de Almeida Filho

February 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data used collected from SpaceX API and “SpaceX & Falcon Heavy Launches” Wikipedia’s page
  - Data was analyzed using python, pandas and graphs
  - Predicted if a launcher will land successfully using some machine learning algorithm(KNN, Linear Regression, SVM and Decision Trees)
- Summary of results
  - All 3 of 4 algorithm show the probabilities of 83.33% of success in future missions the Decision Tree show probabilities of 72.22%, but the amount of data is low so the accuracy may vary

Introduction

- Background and context
    - Commercial space travel is growing and companies like Virgin Galactic, Rocket Lab, Blue Origin, & SpaceX are investing in the field
    - SpaceX has developed the technology to reuse the first stage of launching, therefore making the costs lower than the others
    - Space Y want to enter in this field and compete with Space X.
- Problems
    - Given various launch parameters, will a launch be successful? Let’s answer the question using machine learning models to make predictions.

Section 1

Methodology

Methodology

Executive Summary

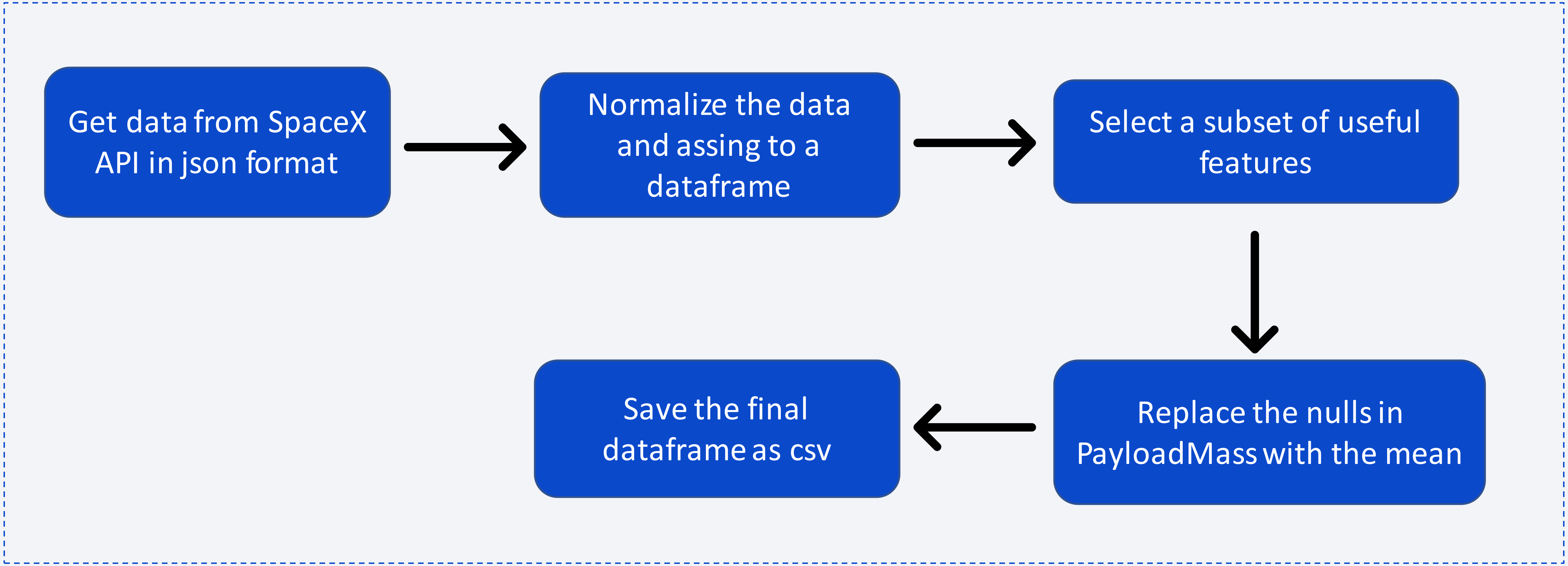
- Data collection methodology:
  - SpaceX API
  - Web scraping
- Perform data wrangling
  - Dataset sanitize(i.e. removing null values)
  - Calculated values to analyze launching
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create various models using GridSearch to find best hyperparameter

Data Collection

- SpaceX API
  - Launch site, payload and core launch was collected via SpaceX API
  - After that another call to other APIs are made to get these fields:
    - Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Finds, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude
- Web Scraping
  - Data was taken from "SpaceX & Falcon Heavy Launches" Wikipedia's page inside HTML tables.
  - In the tables has de following columns:
    - Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time

Data Collection – SpaceX API

https://github.com/MarceloHimura/ibm_data_science_capstone/blob/main/capstone%20week1%20lab%201.ipynb



Get data from SpaceX API in json format

Normalize the data and assing to a dataframe

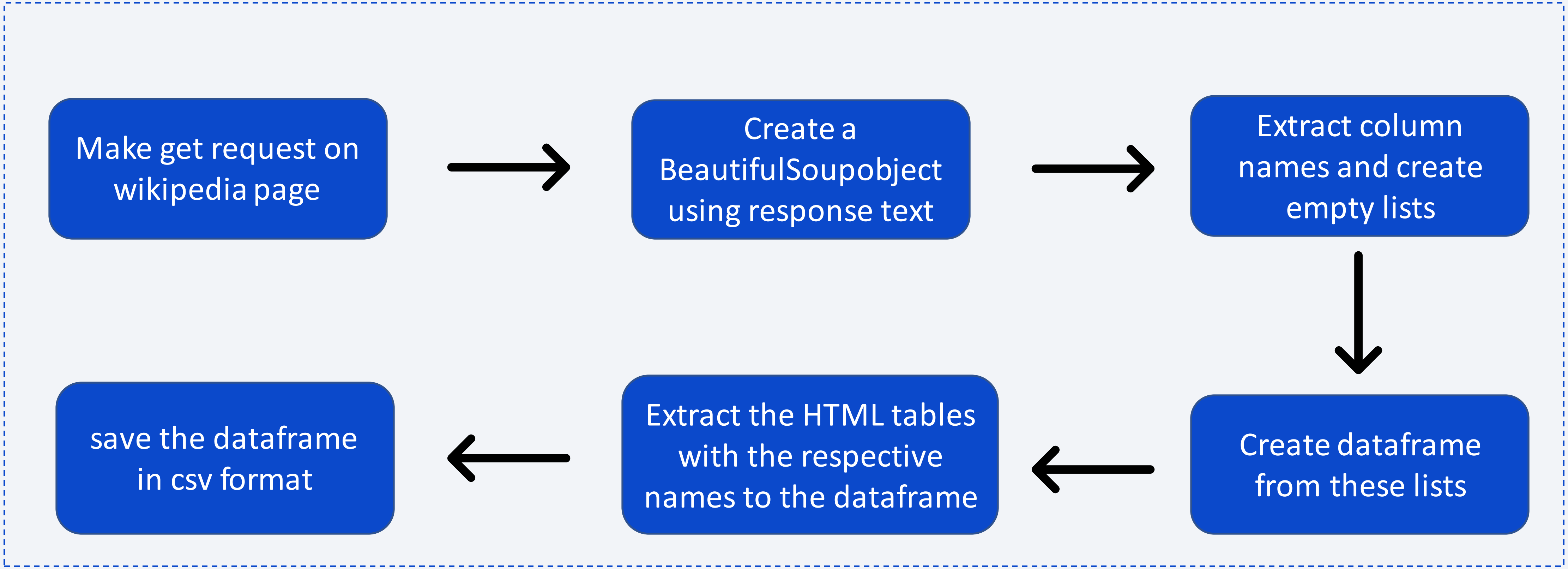
Select a subset of useful features

Replace the nulls in PayloadMass with the mean

Save the final dataframe as csv

Data Collection - Scraping

https://github.com/MarceloHimura/ibm_data_science_capstone/blob/main/capstone%20week1%20lab%202.ipynb



Make get request on wikipedia page

Create a BeautifulSoupobject using response text

Extract column names and create empty lists

Create dataframe from these lists

Extract the HTML tables with the respective names to the dataframe

save the dataframe in csv format

Data Wrangling

https://github.com/MarceloHimura/ibm_data_science_capstone/blob/main/capstone%20week1%20lab%203.ipynb

- Data was analyzed by calculating:
  - The number of launches at each site
  - The number and occurrence of each orbit
  - The number and occurrence of mission outcomes per orbit type
- The Outcome column was created to one-hot encode the outcome of each launch (0 for failed, 1 for successful)

EDA with Data Visualization - 1

https://github.com/MarceloHimura/ibm_data_science_capstone/blob/main/capstone%20week%202%20lab%202 02.ipynb

- Flight Number vs. Payload Mass (scatter plot)
  - As the number of flight increases, the first stage is more likely to succeed.
- Flight Number vs. Launch Site (scatter plot)
  - The result on the VAFB SLC 4E it seems to be associated with flight number
- Payload vs. Launch Site (scatter plot)
  - There were no heavy payloads (more than 10,000 kg) launched from VAFB SLC 4E.
  - At all sites, most payloads were less than 8,000 kg
- Success Rate by Orbit Type (bar chart)
  - On average, success rates were higher for orbit types ES-L!, GEO, HEO, & SSO. Success rates were lowest for orbit types GTO & SO

EDA with Data Visualization - 2

- Flight Number vs. Orbit Type (scatter plot)
  - Higher flight numbers appear to be successful in LEO and MEO orbits.
  - There is no relationship between flight number and orbit type for ISS and GTO orbits.
- Payload Mass vs. Orbit Type (scatter plot)
  - Heavier payloads appear to be successful for LEO orbits.
  - There appears to be no relationship between payload mass and success for ISS & GTO orbits.
- Launch Success Annual Trends (line chart)
  - The average success rate increase for the majority of years (2013-2017 and 2019).

EDA with SQL

https://github.com/MarceloHimura/ibm_data_science_capstone/blob/main/capstone%20week%202%20lab%201.ipynb

- Queried a list of unique launch sites.
- Considered 'CCA' launch sites.
- Calculated the total payload mass for boosters launched by NASA (CRS)
- Calculated the average payload mass (in kg) carried by launches with booster version F9 v1.1.
- Determined the first successful landing in which ground pad was achieved.
- Listed the boosters that have successful drone ship outcomes with payload mass between than 4,000 and 6,000 kg.
- Calculated the total success and failure outcomes for each type of mission outcome.
- Listed the booster versions with the greatest payload mass.
- Listed landing outcome, booster version and launch site for failed drone ship landings.
- Ranked the outcomes of landings that occurred between 06/04/2010 and 03/20/2017 in descending order

Build an Interactive Map with Folium

https://github.com/MarceloHimura/ibm_data_science_capstone/blob/main/capstone%20week%203%20lab%201.ipynb

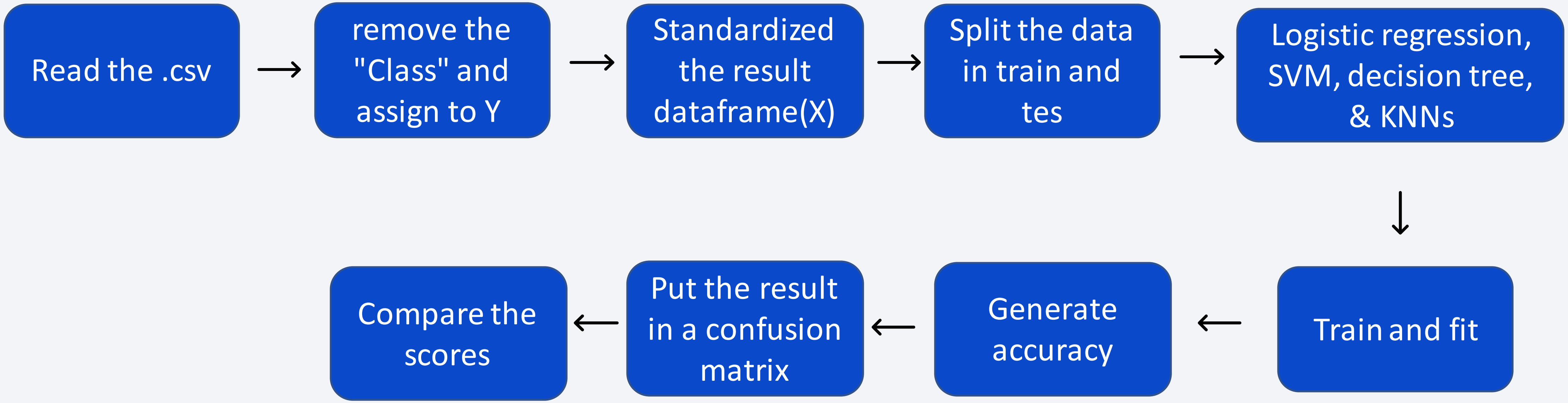
- Created a map of all the launch sites.
- Marked launches at each site, using green markers for successful launches and red markers for failed launches. When zoomed out, map displayed total number grouped by the site. when zoomed in, map displays launches sites
- Calculated the distance between launch sites and coastlines, railways, highways and cities. Visualized these distances with blue lines connected these locations.

Build a Dashboard with Plotly Dash

https://github.com/MarceloHimura/ibm_data_science_capstone/blob/main/week_3_lab_2_spacex_dash_app.py

- Pie Charts
  - A pie chart displays the percentage of launch successes and failures. Selecting different launch sites from the drop-down change the pie chart values.
- Scatter Charts
  - A scatter chart visualizes the correlation between payload and launch success. Selecting different launch sites from the drop-down menu change the scatter chart values.

Predictive Analysis (Classification)


Read the .csv

remove the "Class" and assign to Y

Standardized the result dataframe(X)

Split the data in train and tes

Logistic regression, SVM, decision tree, & KNNs

Train and fit

Generate accuracy

Put the result in a confusion matrix

Compare the scores

https://github.com/MarceloHimura/ibm_data_science_capstone/blob/main/capstone%20week%204%20lab%201.ipynb

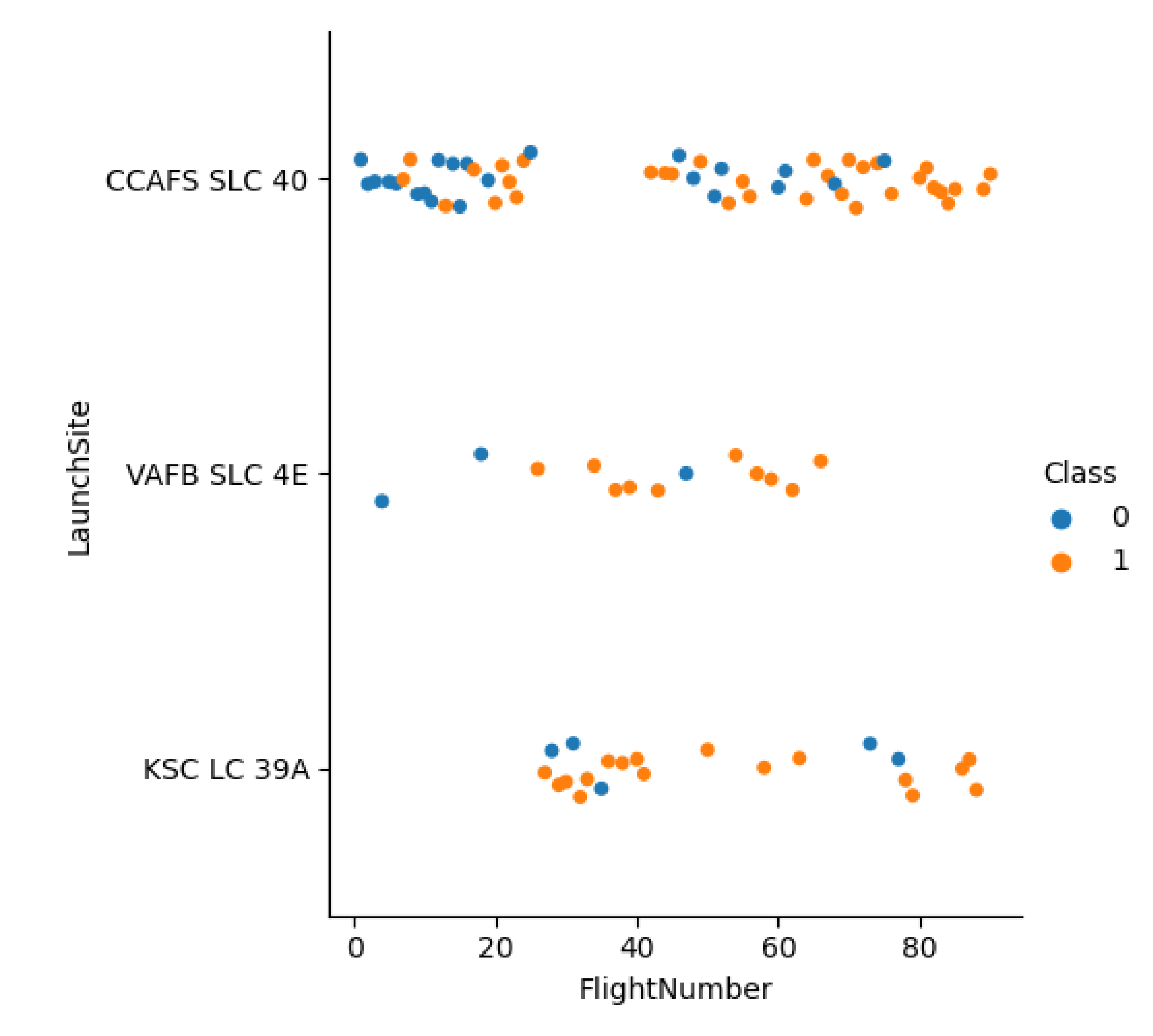
Results

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

Section 2

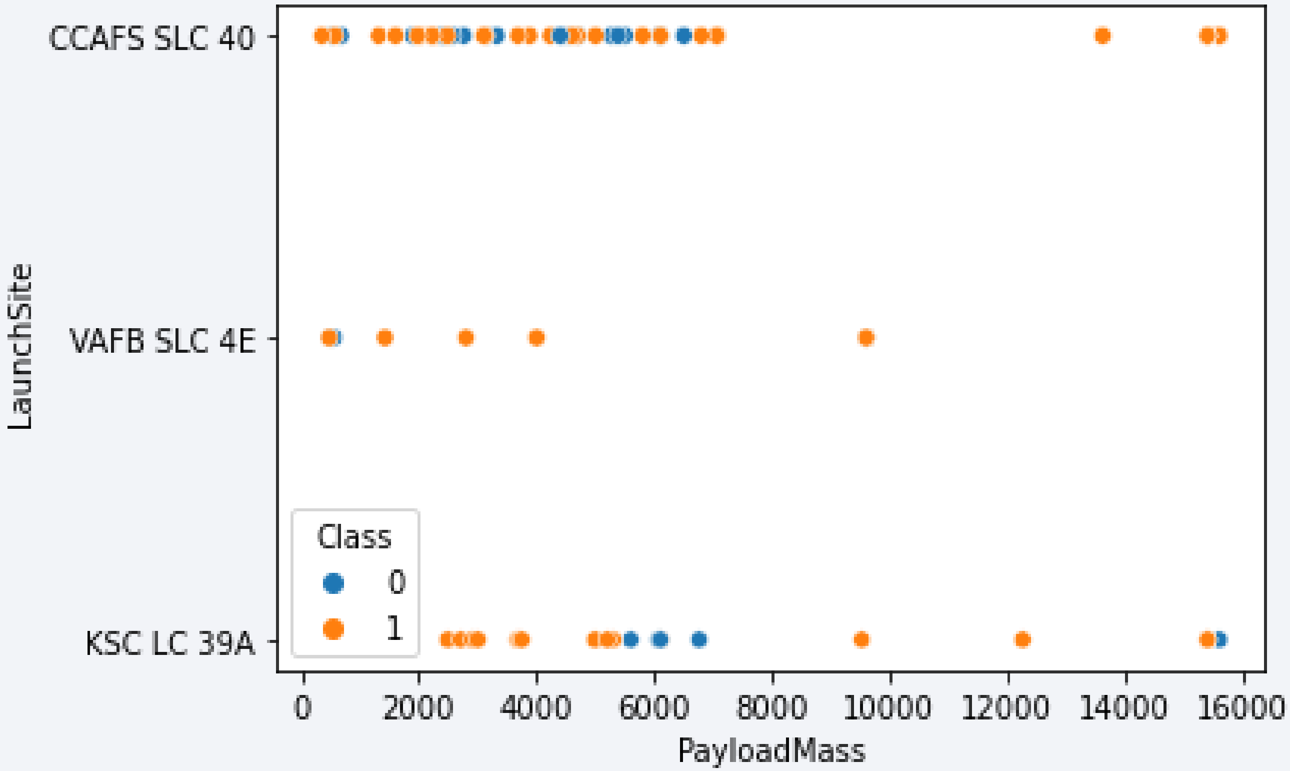
Insights drawn from EDA

Flight Number vs. Launch Site



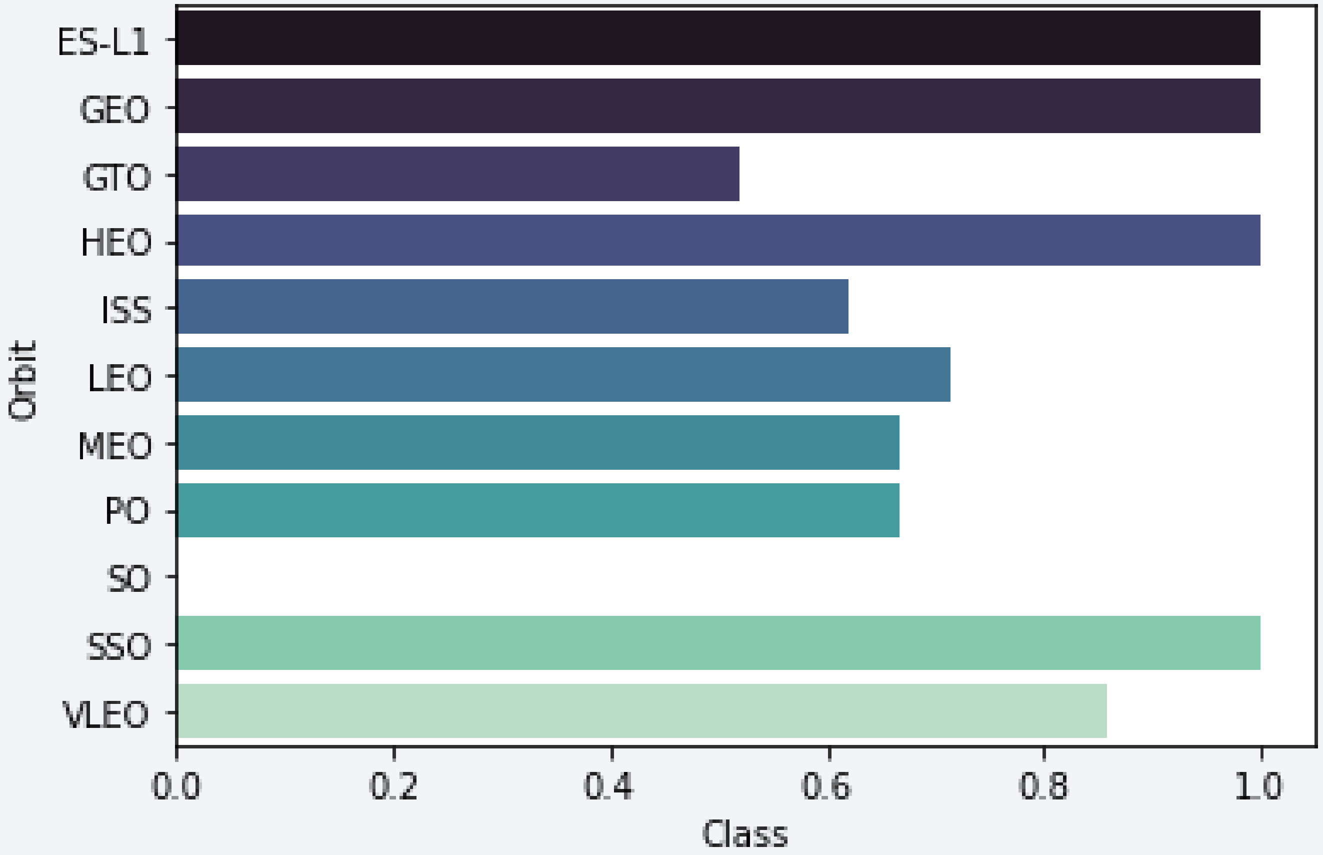
- Orange indicates a successful and blue indicates a failed.
- On VAFB SLC 4E site it seems that higher flight numbers correspond to more successful launches.
- The other launch sites, seems that there is no relationship between flight number and launch site.

Payload vs. Launch Site



- Orange indicates a successful and blue indicates a failed.
- Only payloads with less than 10,000 kg were launched from site VAFB SLC 4E
- For launch site CCAFS SCL 4G, payloads above 12,000 kg were successful.

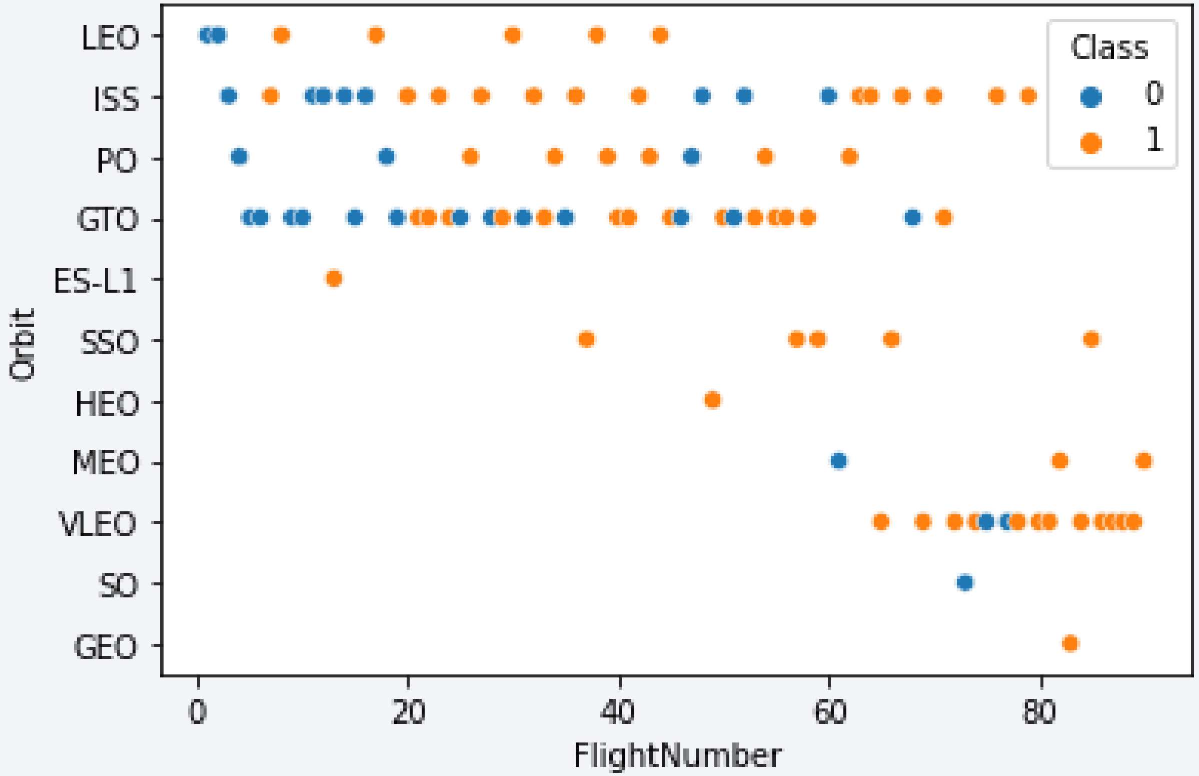
Success Rate vs. Orbit Type



- Orbit types ES-L1, GEO, HEO and SSO had the highest average success rates.

- Orbit types GTO and SO were the least successful.

Flight Number vs. Orbit Type



- Orange indicates a successful and blue indicates a failed.

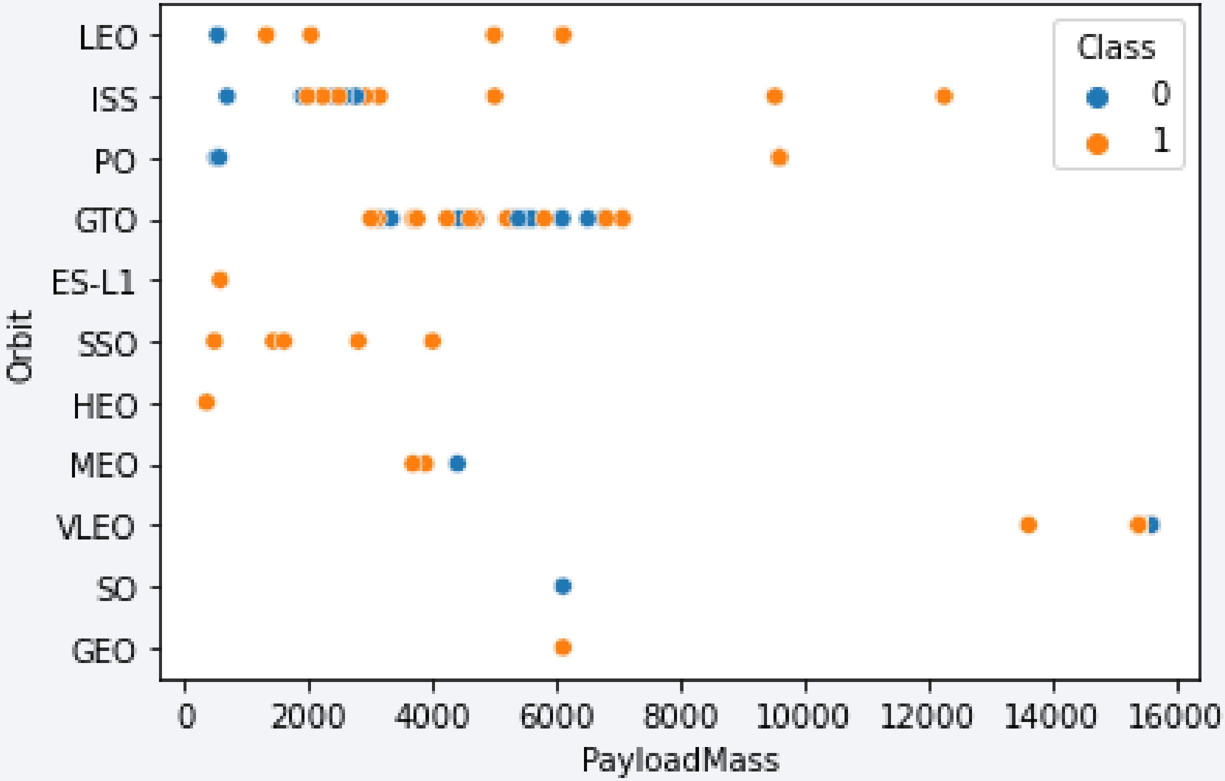
- For LEO and MEO orbits, higher flight numbers were more successful.

- For ES-L1, SSO, HEO and GEO orbit launches, all flight numbers were successful.

- For ISS, PO, GTO and VLEO orbit launches, there was no relationship between flight number and orbit type

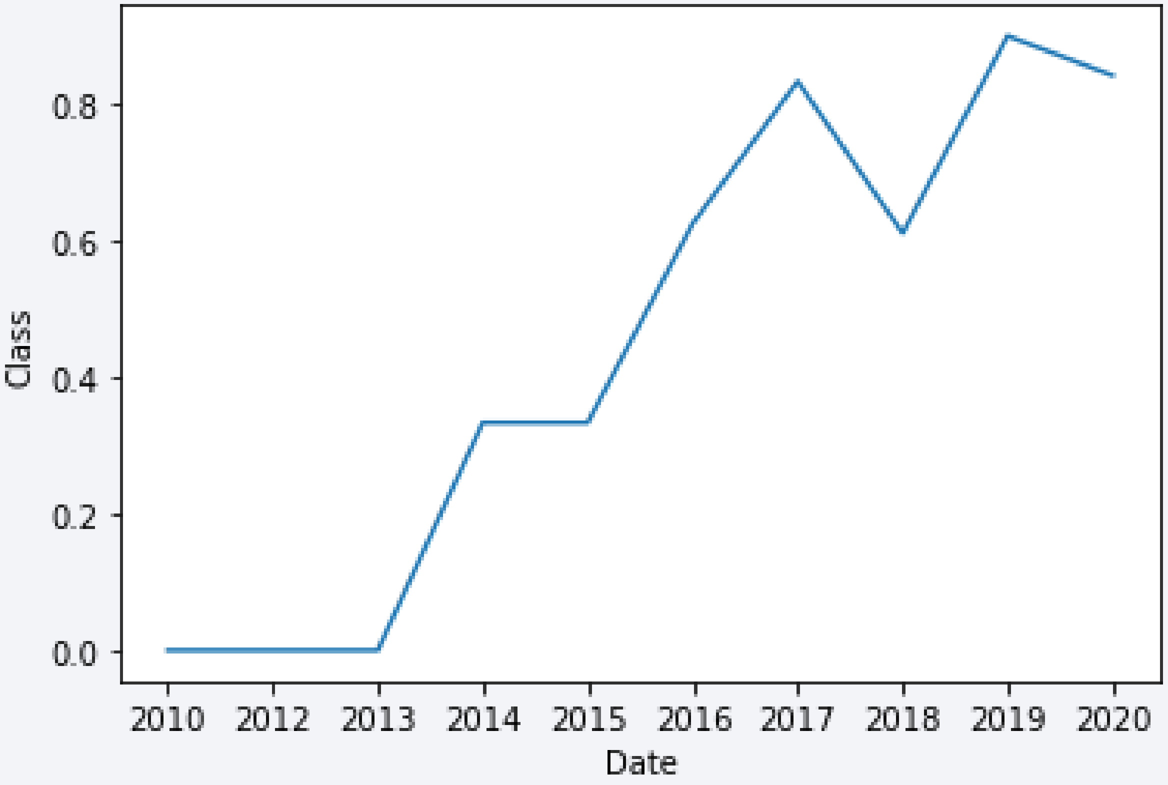
- All SO orbit launches were failure.

Payload vs. Orbit Type



- Orange indicates a successful and blue indicates a failed.
- For LEO, ISS orbit launches, heavier payloads tend to be more successful.
- The ES-L1, SSO, HEO and GEO orbit launches, launches were successful independent of the payload mass. However, all payloads were less than 8,000 kg
- For ISS and GTO orbit launches, there is no relationship between payload mass and orbit type.

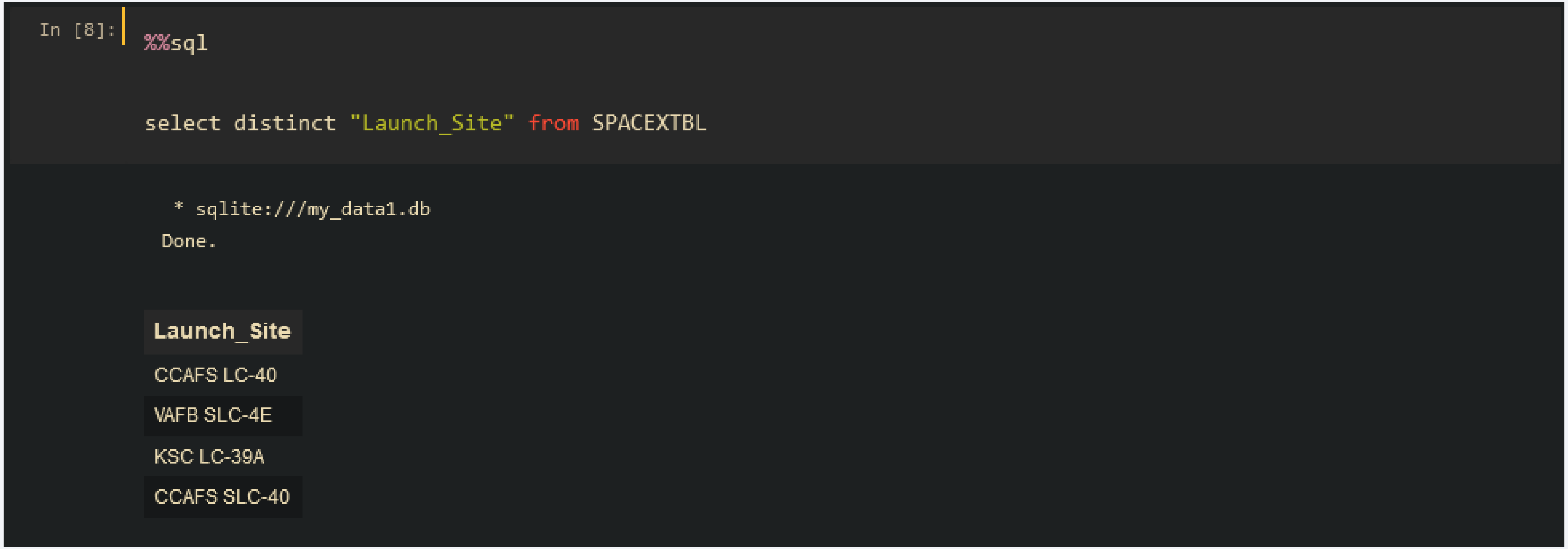
Launch Success Yearly Trend



- Overall, the average success rate increased from 2013 to 2019.
- The success rate was static from 2010 to 2013.

All Launch Site Names

- There are 4 distinct launch sites.
- Used the distinct to return a list of unique launch site names.



In [8]:

```
%%sql

select distinct "Launch_Site" from SPACEXTBL
```

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

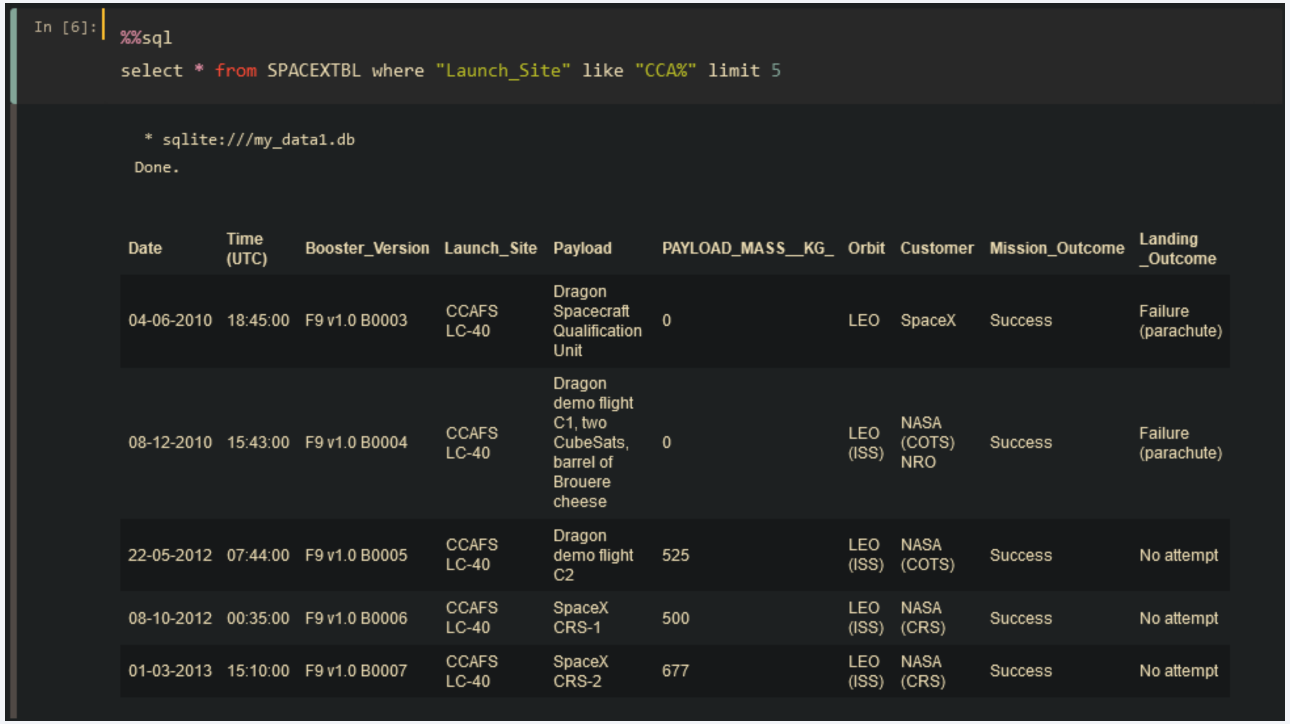
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- The first 5 occurred at CCAFS LC-40.

- Used a like to return only 'CCA' launch sites.

- Used the limit to return only the first 5 rows



In [6]:

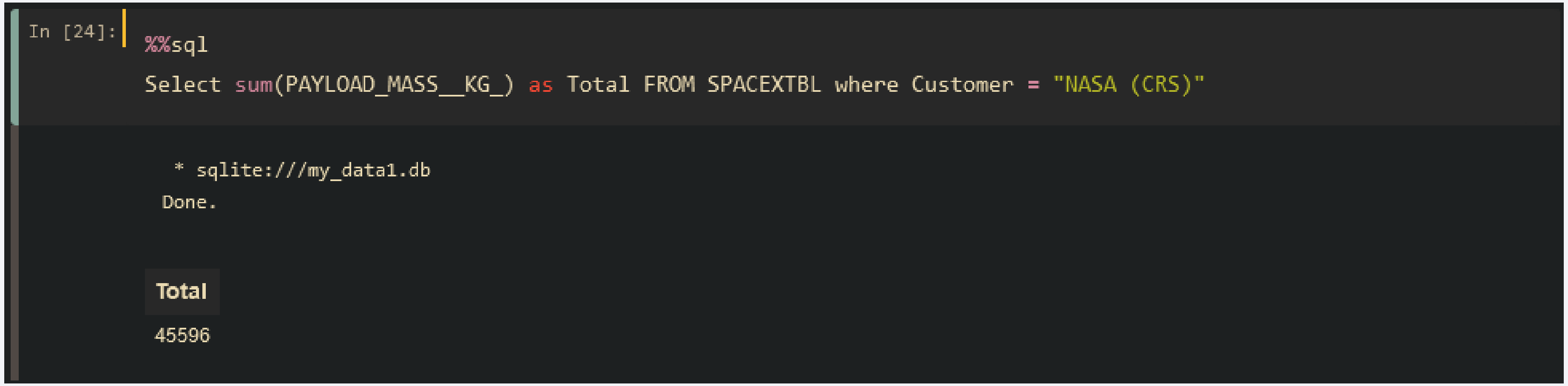
```
%%sql
select * from SPACEXTBL where "Launch_Site" like "CCA%" limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload mass for customer 'NASA (CRS)' was 45,596 kg.
- Use the where to limit results to customer 'NASA (CRS)'.
- Use the sum function to calculate the total payload mass in kg.



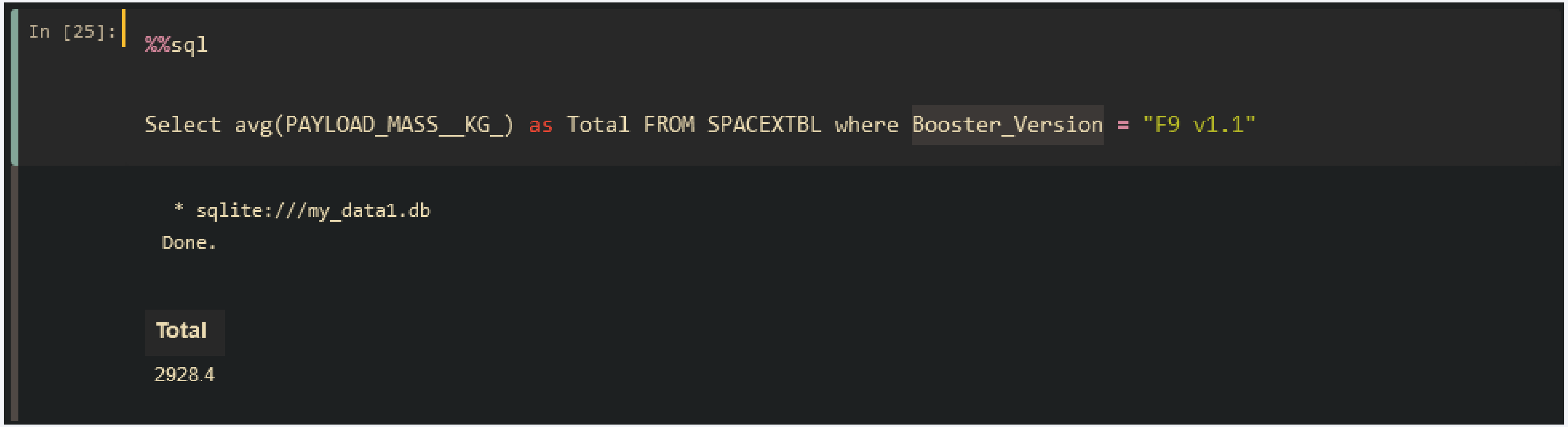
```
%%sql
Select sum(PAYLOAD_MASS__KG_) as Total FROM SPACEXTBL where Customer = "NASA (CRS)"
```

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

Total
45596

Average Payload Mass by F9 v1.1

- The average payload mass for launches with booster version 'F9 v1.1' was 2928.4 kg
- Use the where to limit results to booster version 'F9 v1.1'.
- Use the avg function to calculate the average payload mass.



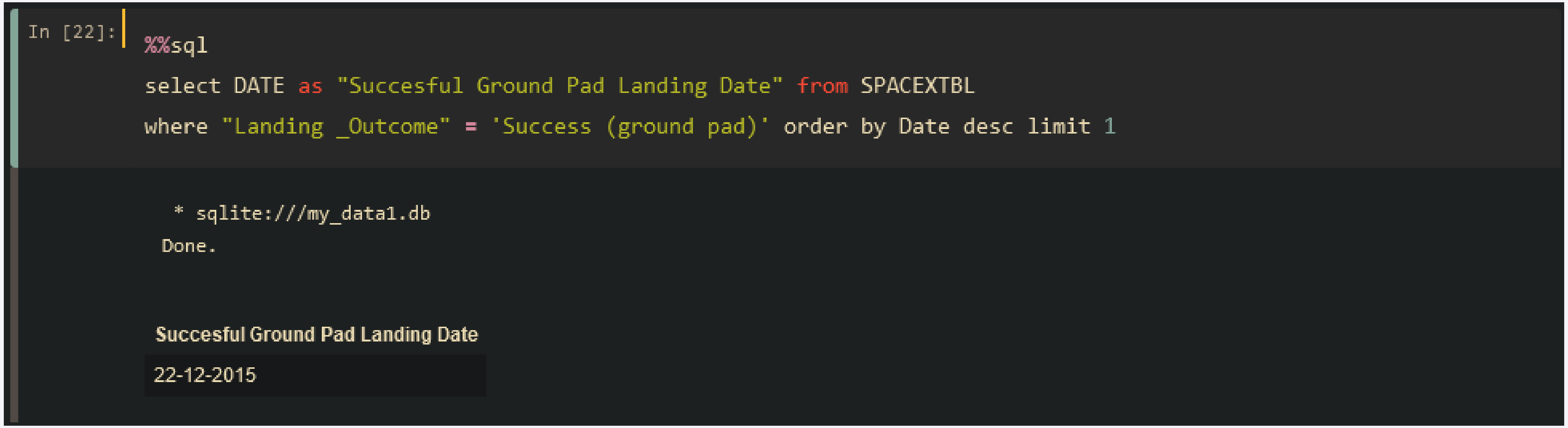
```
%%sql

Select avg(PAYLOAD_MASS__KG_) as Total FROM SPACEXTBL where Booster_Version = "F9 v1.1"
```

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

Total
2928.4

First Successful Ground Landing Date

- The first successful ground pad landing occurred on 12/22/2015.
- The where was used to limit results to successful ground rate outcomes.
- The order by and limit was used to order by date and put the order from smallest to largest and then show only the first row



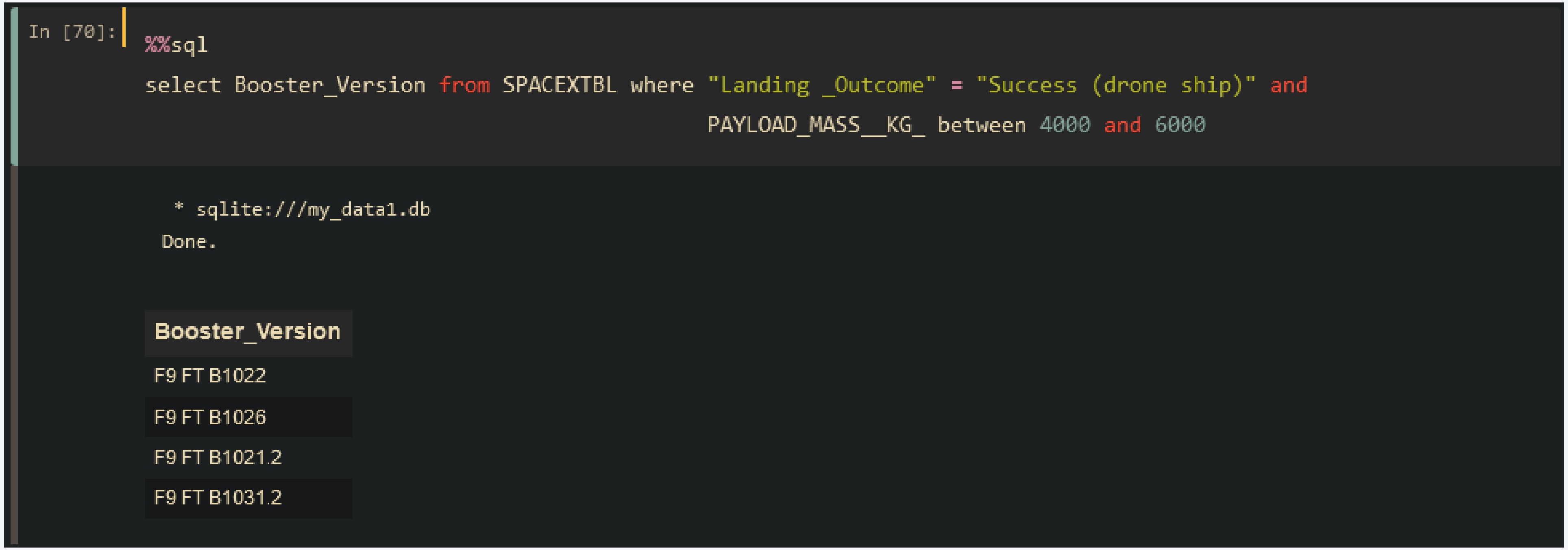
```
%%sql
select DATE as "Succesful Ground Pad Landing Date" from SPACEXTBL
where "Landing _Outcome" = 'Success (ground pad)' order by Date desc limit 1
```

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

Succesful Ground Pad Landing Date
22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 4 distinct booster version for successful drone ship landings with payload masses between 4,000 and 6,000 kg.
  • The where was utilized to limit results to successful drone ship landings with payload masses between 4,000 and 6,000 kg.
  • The distinct was utilized to return a set of distinct booster versions.



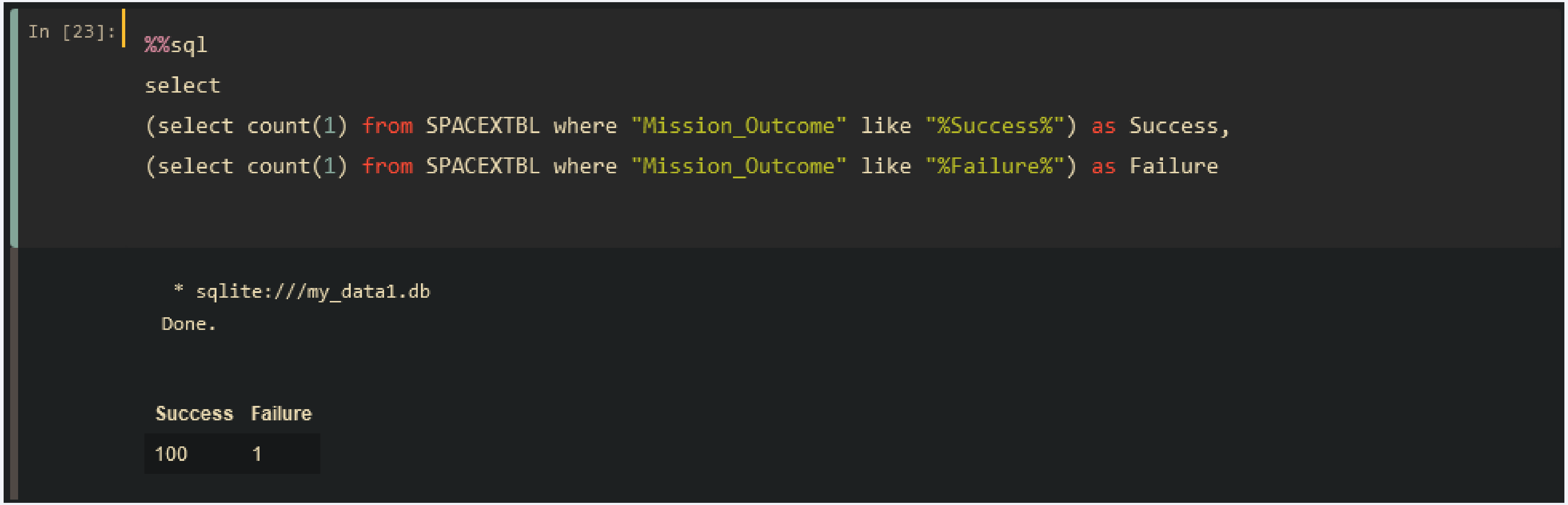
```
In [70]: %%sql
         select Booster_Version from SPACEXTBL where "Landing _Outcome" = "Success (drone ship)" and
                                                   PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * 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

- There were 100 successful missions and 1 failed mission.
- 2 subquery were used one to get the number of 'Success' and other to 'Failure' and then returned as column from the main query



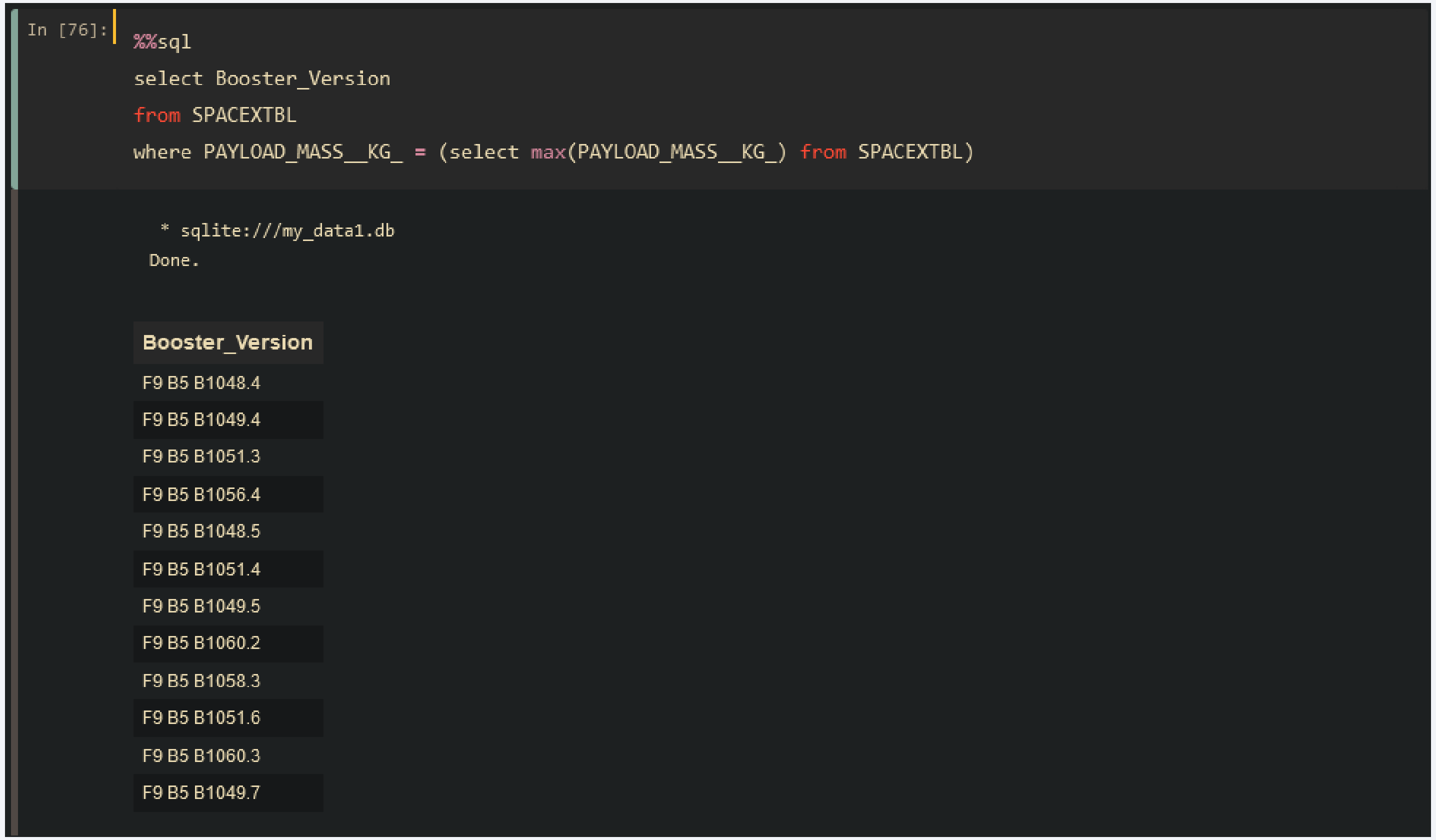
In [23]:

```
%%sql
select
(select count(1) from SPACEXTBL where "Mission_Outcome" like "%Success%") as Success,
(select count(1) from SPACEXTBL where "Mission_Outcome" like "%Failure%") as Failure
```

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

Success	Failure
100	1

Boosters Carried Maximum Payload



```
In [76]: %%sql
         select Booster_Version
         from SPACEXTBL
         where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

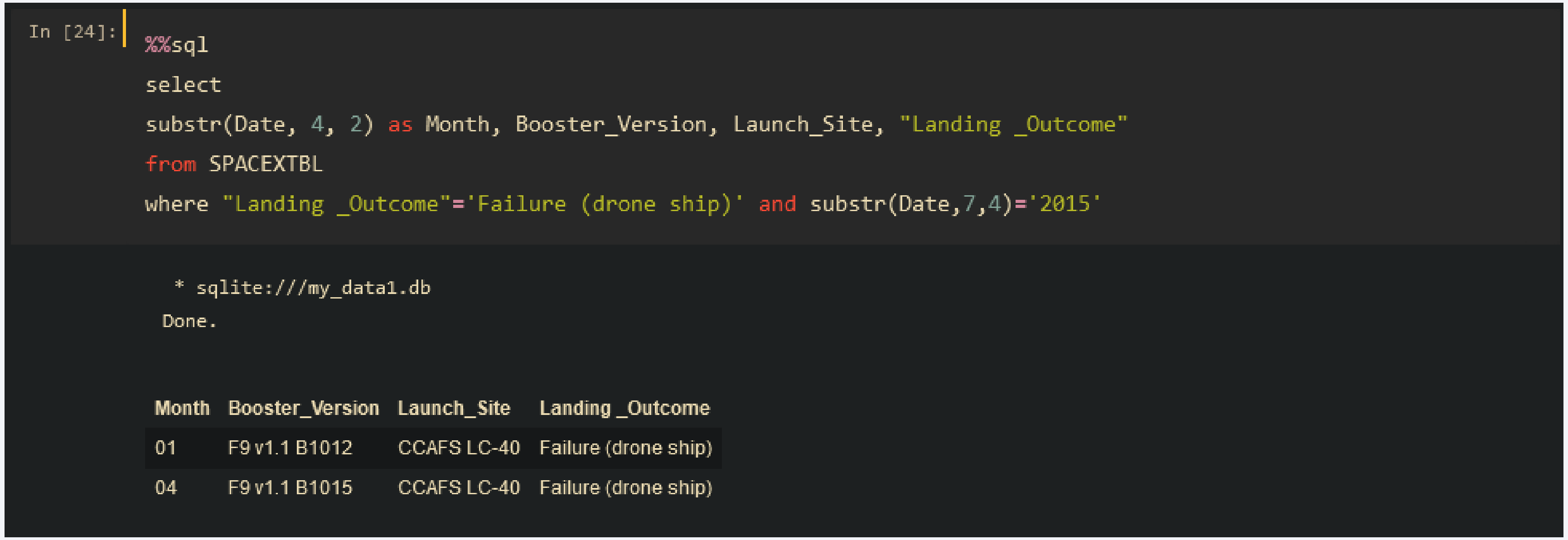
```
 * sqlite:///my_data1.db
Done.
```

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

- There were 12 different versions that had the maximum payload mass.
  - A subquery calculated the maximum payload mass. It was applied to the where to limit results to launches with the maximum payload mass.

2015 Launch Records

- There were 2 different launches with failed drone ship landings in the year 2015.
- Month was returned in number and substr function was used because SQLLite does not support monthnames/year functions
- The where limited the results to failed drone ship landings in the year 2015.



In [24]:

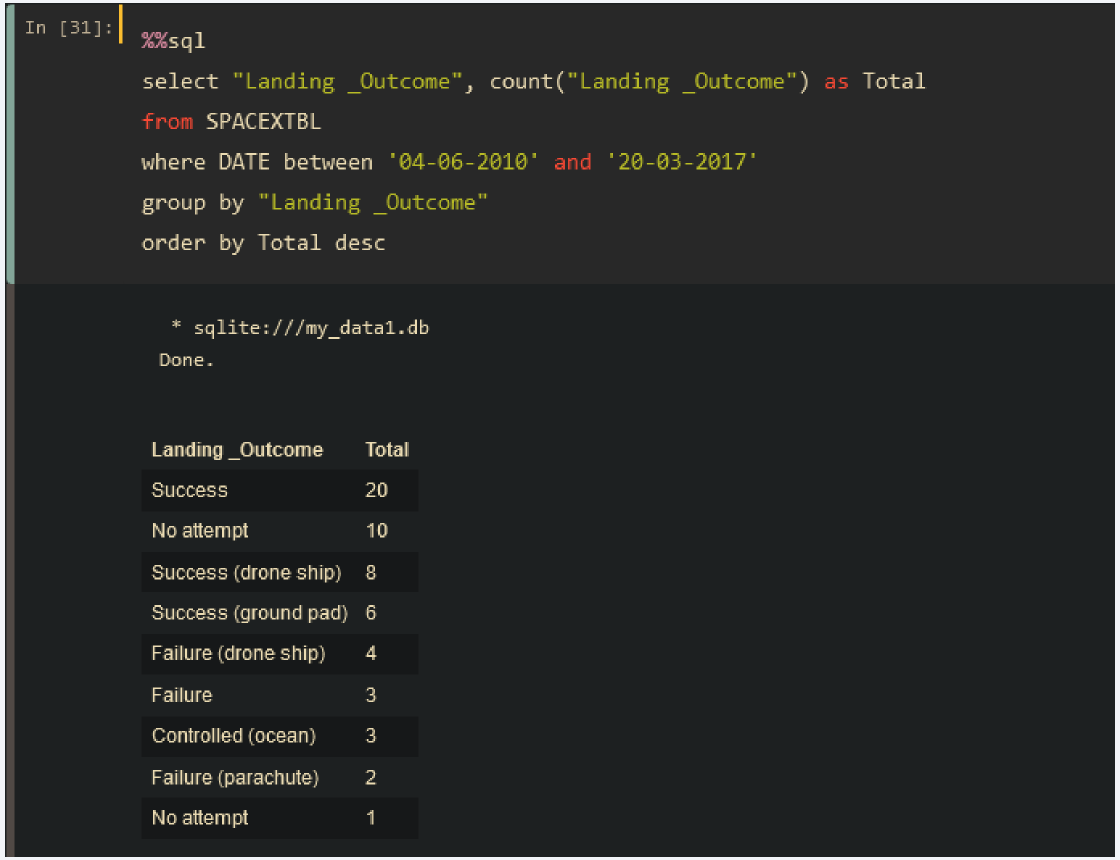
```
%%sql
select
substr(Date, 4, 2) as Month, Booster_Version, Launch_Site, "Landing _Outcome"
from SPACEXTBL
where "Landing _Outcome"='Failure (drone ship)' and substr(Date,7,4)='2015'
```

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

Month	Booster_Version	Launch_Site	Landing _Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- There are 57 landing outcomes between 06/04/2010 and 03/20/2017. When considered in descending outcomes order, 'Success' is ranked first.
- The where limited results to launches occurring between 06/04/2010 and 03/20/2017.
- 59.65% of the launches were successful in that period



```
In [31]: %%sql
         select "Landing _Outcome", count("Landing _Outcome") as Total
         from SPACEXTBL
         where DATE between '04-06-2010' and '20-03-2017'
         group by "Landing _Outcome"
         order by Total desc
```

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

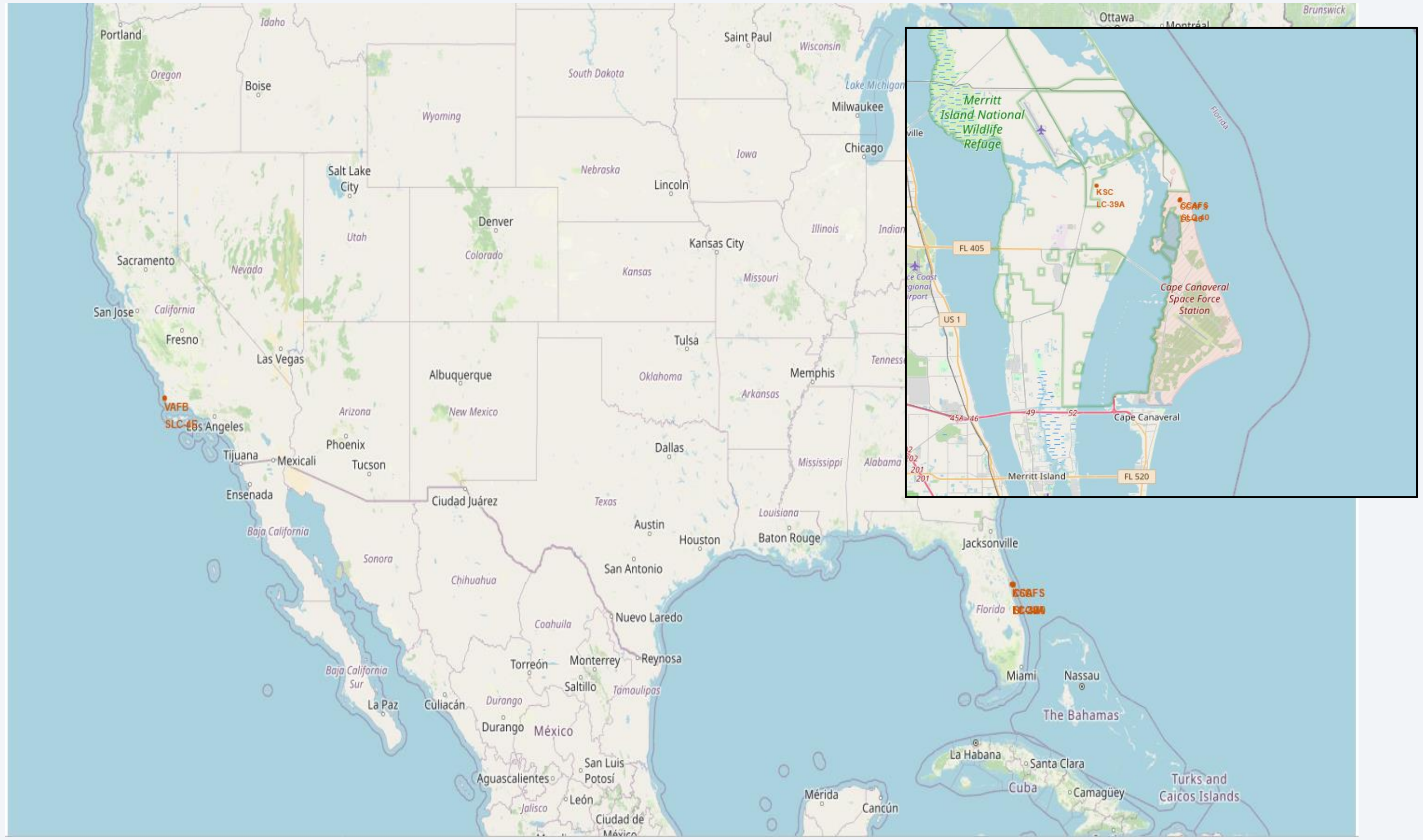
Landing _Outcome	Total
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Launch Site Map

- Using a data frame of unique launch site locations (with the latitude and longitude values) was used to create a map of launch site .
- Each launch site was marked using a red circle icon and labeled with the name of the launch site.

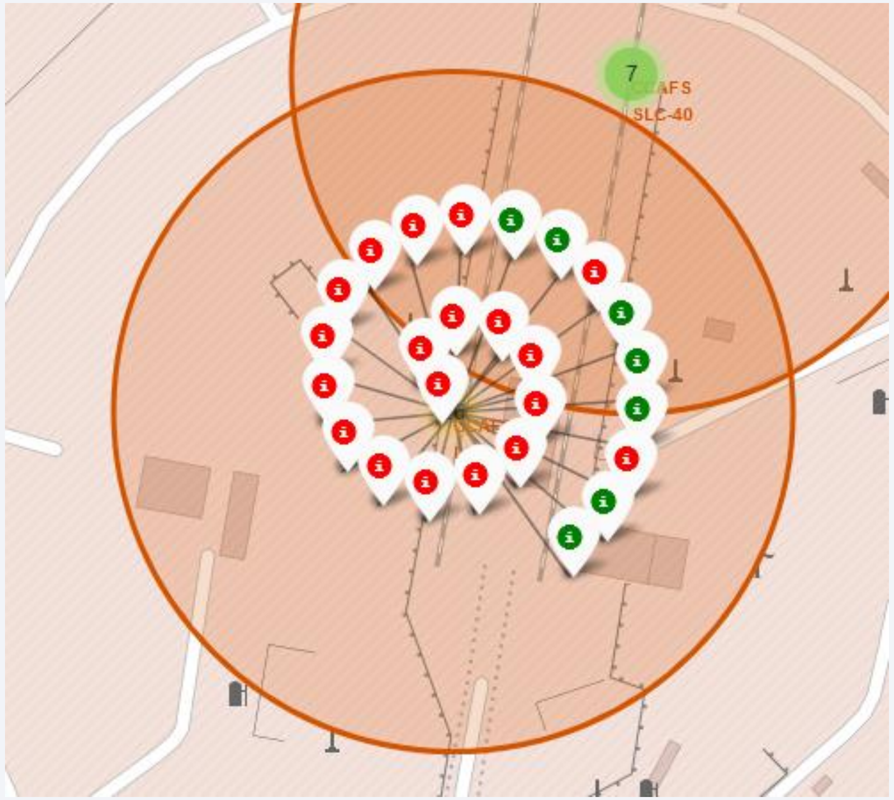

VAFB
SLC-4E
CCAFS
LC-40
KSC
LC-39A
CCAFS
SLC-40
Merritt Island National Wildlife Refuge
Cape Canaveral Space Force Station
Cape Canaveral
Merritt Island

Successful & Failed Launches per Site

- When zoomed into a specific launch site , the specific launches that have occurred at that site become visible

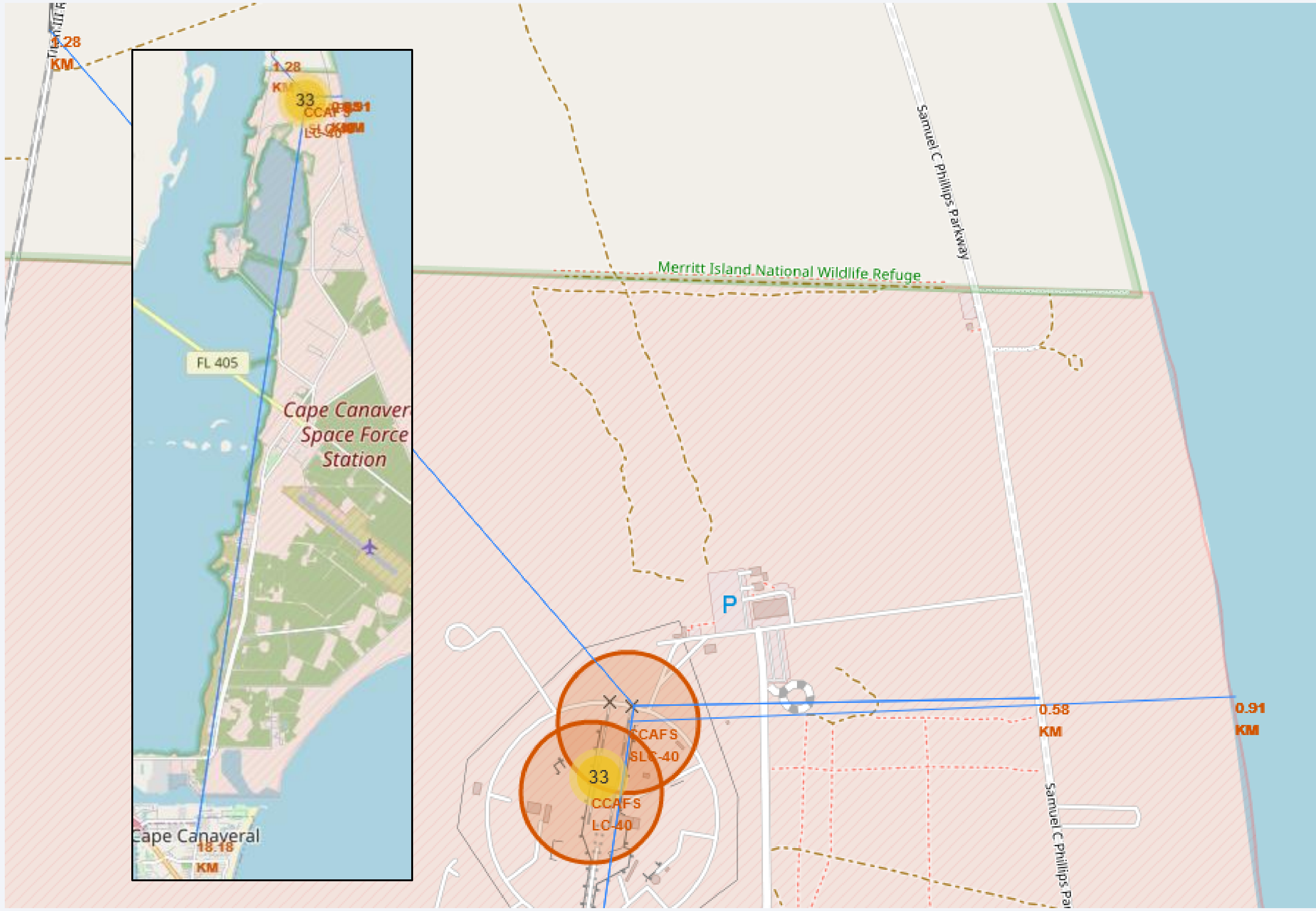
- Each launch is marked with an indicator that indicate if the launch was a success or failure.

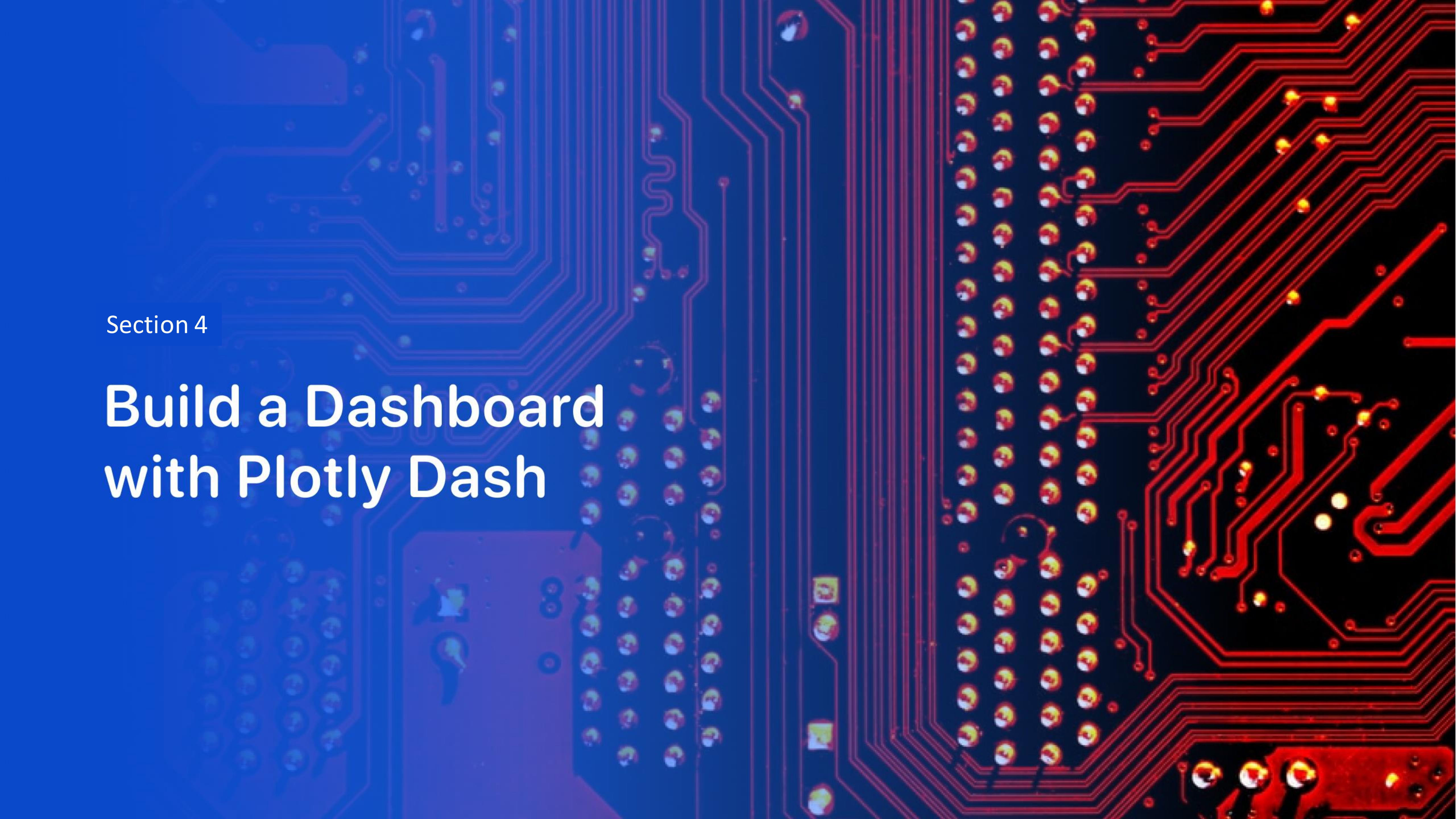
- The launch site itself is labeled with the name of the location, as well as a number indicating the total number of successful launches.



Launch Site Proximity to Railway, Highway, Coastline

- The proximity of various map features to the CCAFS SLC-40 launch site were marked on the map
- The Samuel C. Phillips Parkway highway is 0.58 km from the location.
- The nearest railway is 1.28 km from the launch location
- The nearest city (Cape Canaveral) is 18.40 km from the site.

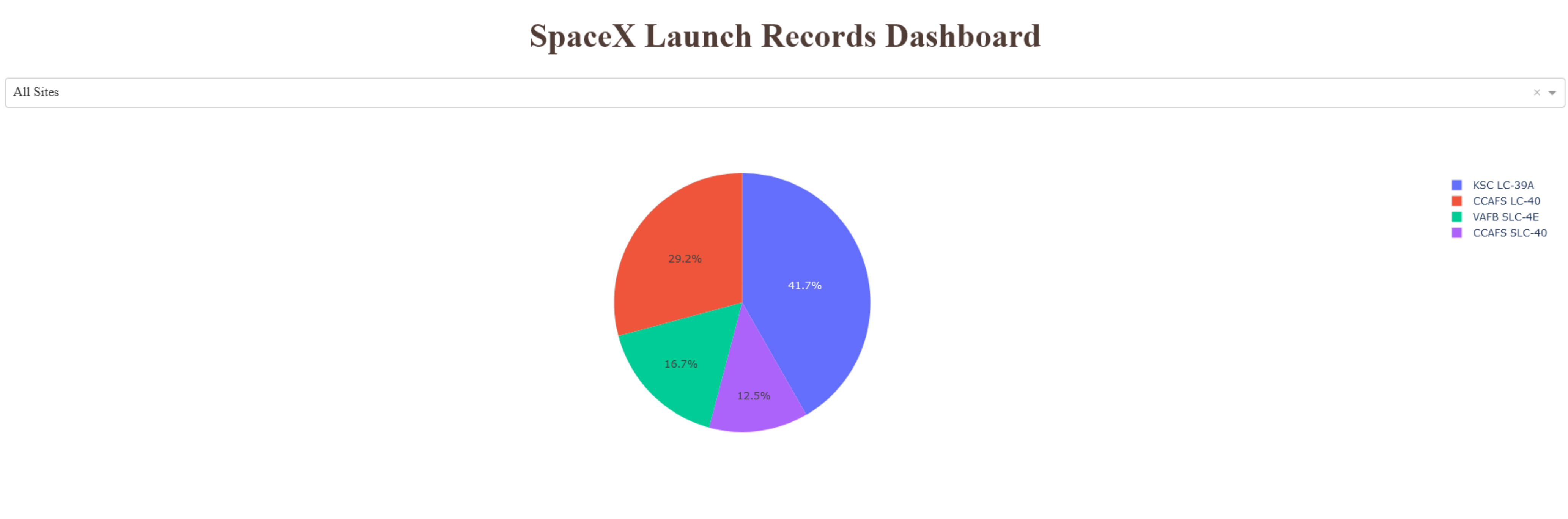




Section 4

Build a Dashboard with Plotly Dash

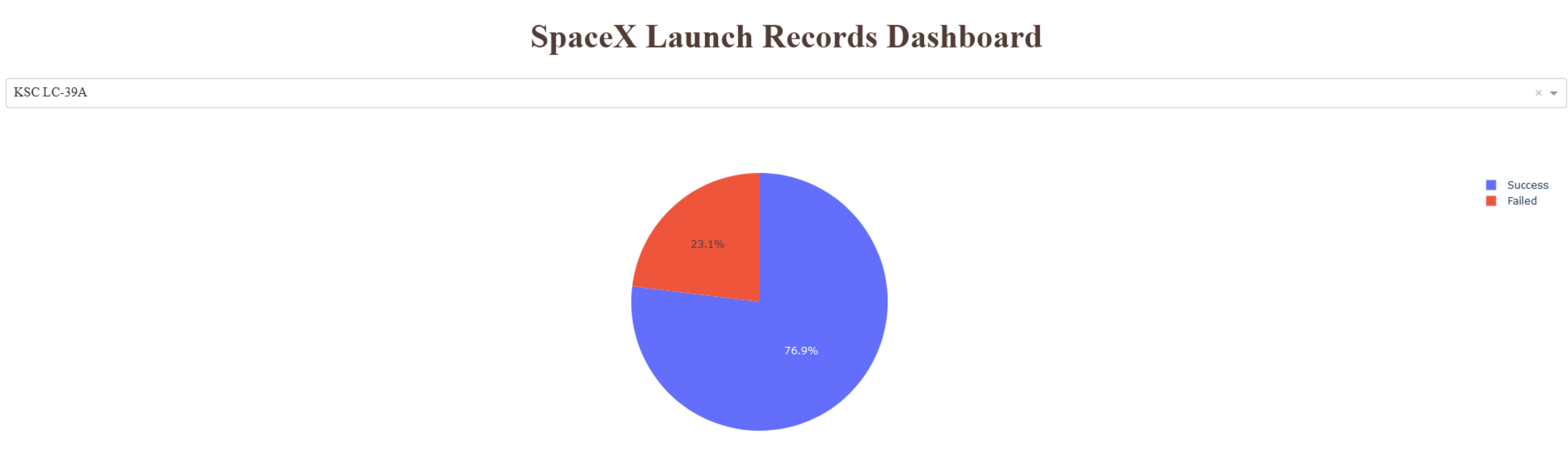
Successful Launches by Launch Site


SpaceX Launch Records Dashboard

All Sites

- Using a data frame of launch successes and failures at each launch site, a pie chart of launch success rates at each site is displayed.
- Site KSC LC-39A has the highest success rate, while site CCAFS SLC-40 has the lowest success rate.

Launch Site Most Successful



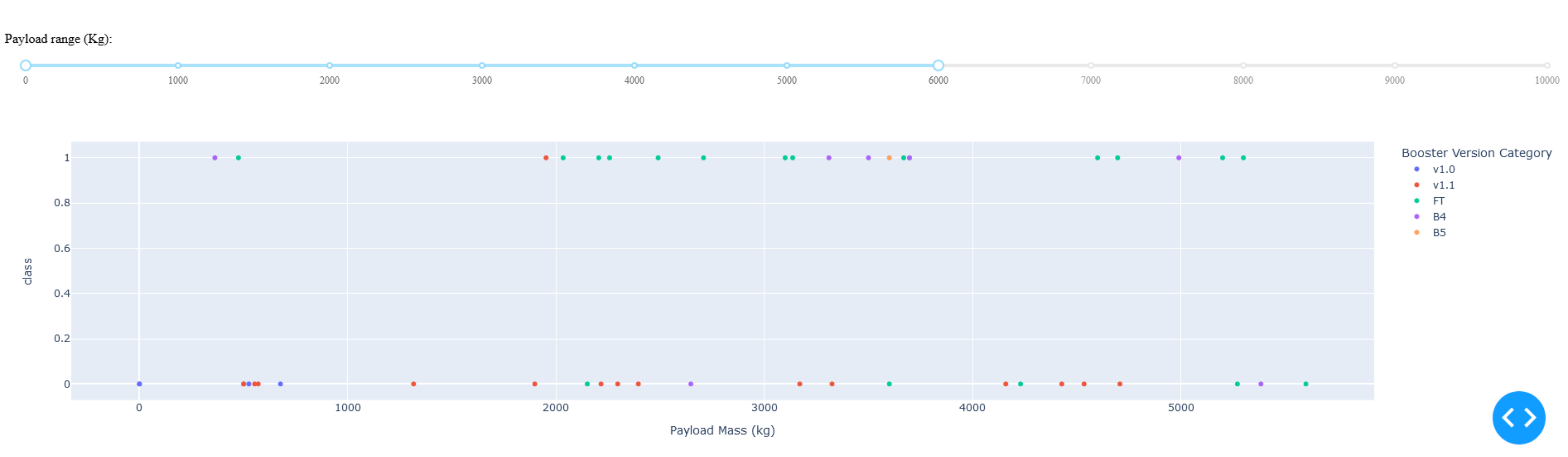
SpaceX Launch Records Dashboard

KSC LC-39A

- After filtering a data frame of launch data to include only launches from site KSC LC-39A, values were grouped by Launch Site and class (success or failure). Total successes and failures were calculated and displayed in a pie chart.

- At launch site KSC LC 39-A, 76.9% of launches were successful (displayed in purple) , while 23.1% of launches failed (displayed orange).

Payload Mass vs Launch Outcome



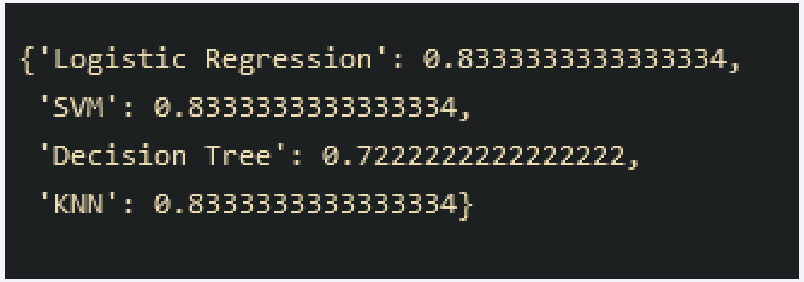
Payload range (Kg):

- Using a data frame of launch information, a scatter plot of Payload Mass (kg) is plotted against class (success or failure).
- A slider bar allows the user to define the limits of the payload masses displayed in the scatter plot. By default, the slider bar is set to a range of 0 kg to 10,000 kg (10,000 kg is the maximum value for payload masses in the data frame).

Section 5

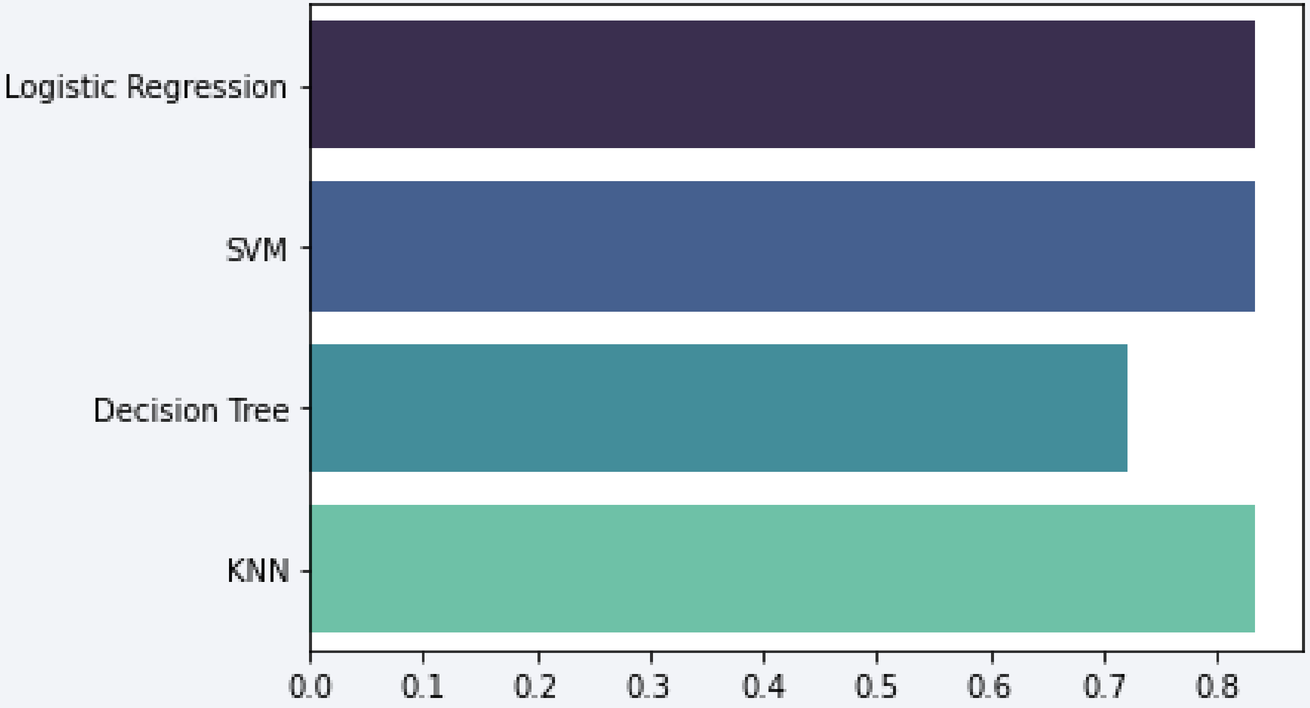
Predictive Analysis (Classification)

Classification Accuracy

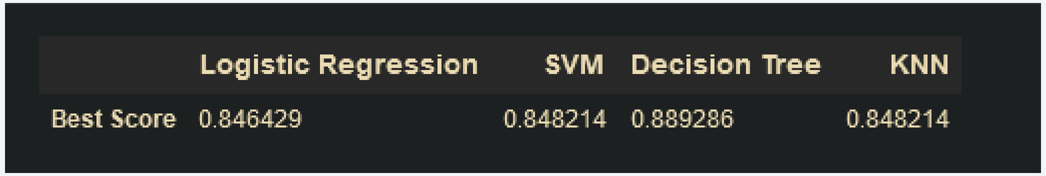


```
{'Logistic Regression': 0.8333333333333334,
 'SVM': 0.8333333333333334,
 'Decision Tree': 0.7222222222222222,
 'KNN': 0.8333333333333334}
```

- All had almost the same result of 83.33% except for Decision Tree with 72.22%.
- Since the testing set size was only 18 records, large variances in accuracy may exist. A larger data set would likely produce more precise scores.

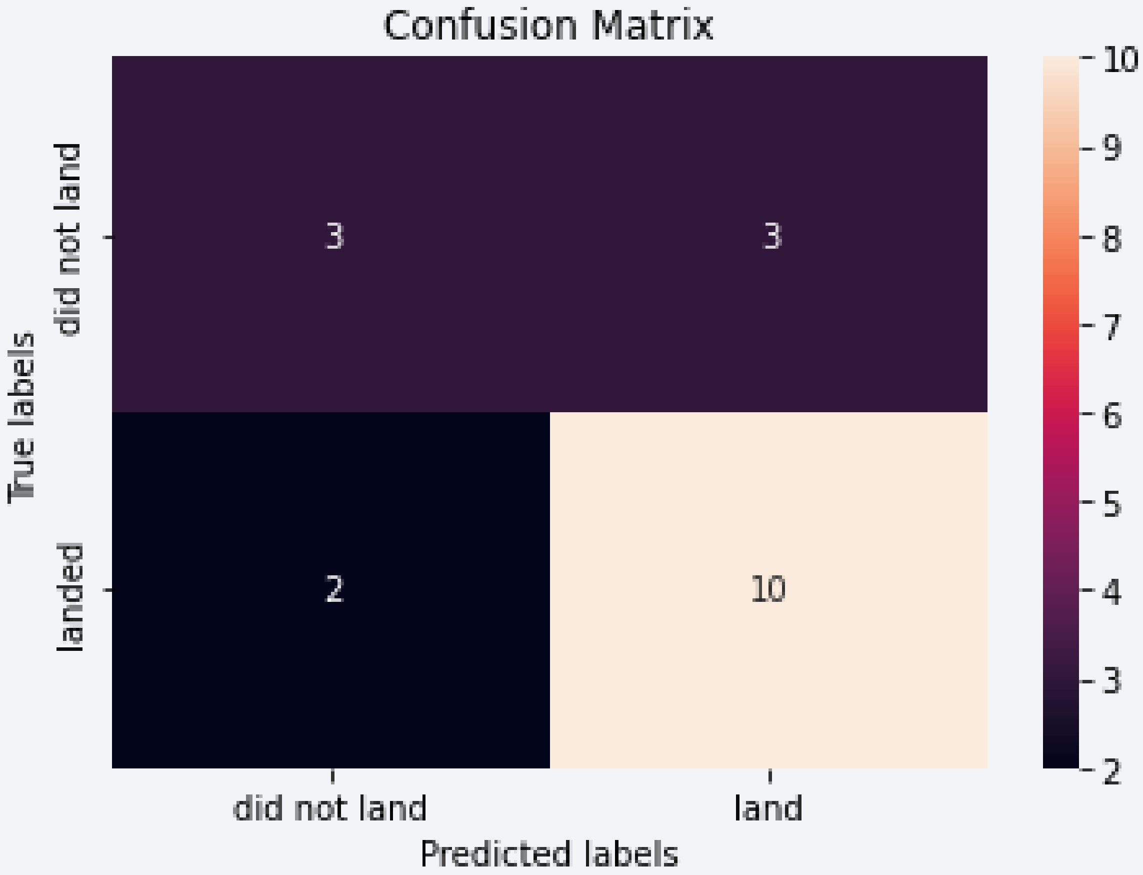


Confusion Matrix



	Logistic Regression	SVM	Decision Tree	KNN
Best Score	0.846429	0.848214	0.889286	0.848214

- despite having the lowest accuracy, the decision tree model performed the best.
- It's confusion matrix indicates that is predicted:
  - 2 true positives
  - 10 true positives
  - 3 false negatives
  - 3 false positives



Conclusions

- Analysis of launch data indicates that the average success rate of all launches is 66.67%.
- Launch sites may be in closer proximity to coastlines, highways, and railroads, but they tend to be located farther away from cities.
- Four machine learning models were used to predict the success or failure or future launches. All three of four models had the same score (83.3%) the Decision tree has 72.22% of score. This this similarity is due to the small sample size of data. More data would likely produce better predictions.
- By visualizing the data, it may be observed that:
  - Certain types of orbits (ES-L1, GEO, HEO, and SSO) had a higher success rate.
  - The relationship between payload mass and launch site has an effect of the success of a launch. Specifically, launches with payload masses greater than 12,000 kg that occurred at sites CCAFS SLC-40 and KSC LC-39A were much more likely to succeed.

Appendix

- Wikipedia’s List of Falcon 9 and Falcon Heavy Launches:
    - https://en.wikipedia.org/wiki/List_of_Falcon%5C_9%5C_and_Falcon_Heavy_launches?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2021-01-01
- Thanks to Instructors:
  Yan Luo, Ph.D., Data Scientist and Developer, IBM
  Joseph Santarcangelo, Ph.D., Data Scientist, IBM

Thank you!