IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dana Ghioca-Robrecht
May 31, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection from SpaceX REST API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) with SQL, Pandas, and Matlibplot
  - Interactive Visual Analytics and Dashboard with Folium and Plotly Dash
  - Machine Learning Predictive models using logistic regression, support vector machine, decision tree and K-nearest neighbor
- Summary of all results
  - Launching sites are in coastal areas and close to the equator
  - KSC LC-39A is the landing site with the highest success rate
  - The four predictive model had similar accuracies

Introduction

- Space X is a leader in the space industry that aims at making space travel affordable.
- They accomplish this with rocket launches that are relatively inexpensive by reusing the first stage of the Falcon 9 rocket. Therefore, Space X advertises space trips at only $62 millions for each launch compared to $165 million trips offered by the competitors, who are not reusing the first stage.
- To estimate total cost of Space X launches, we will use Data Science and Machine Learning models to predict if the first stage will land successfully and which location is best for launches

Section 1

Methodology

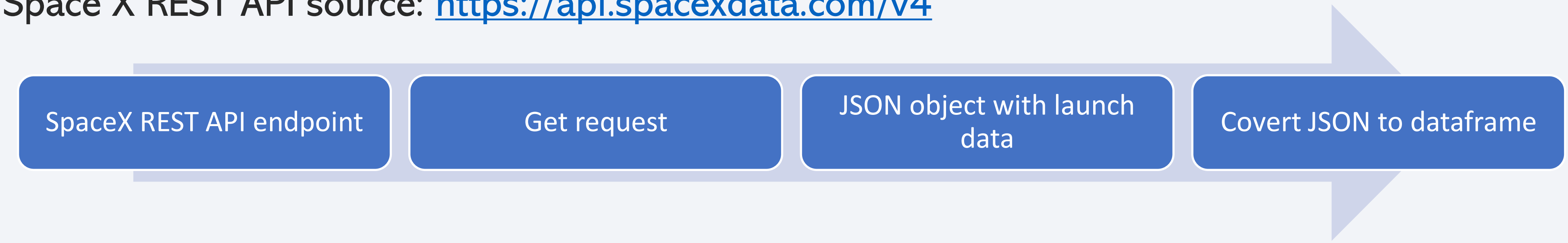
Methodology

Executive Summary

- Data collection methodology:
  - Data was gathered from Space X REST API and wiki web scraping
- Perform data wrangling
  - Methods includes data filtering, missing data handling, and one hot encoding
  - New variable for landing outcome was created
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data were normalized, training and test sets were created, and four classification models were applied, tuned, and evaluated for their accuracy in predicting the successful landing of the first stage of Falcon 9.

Data Collection

- Data sets were collected from two sources:
  - Space X REST API source: https://api.spacexdata.com/v4



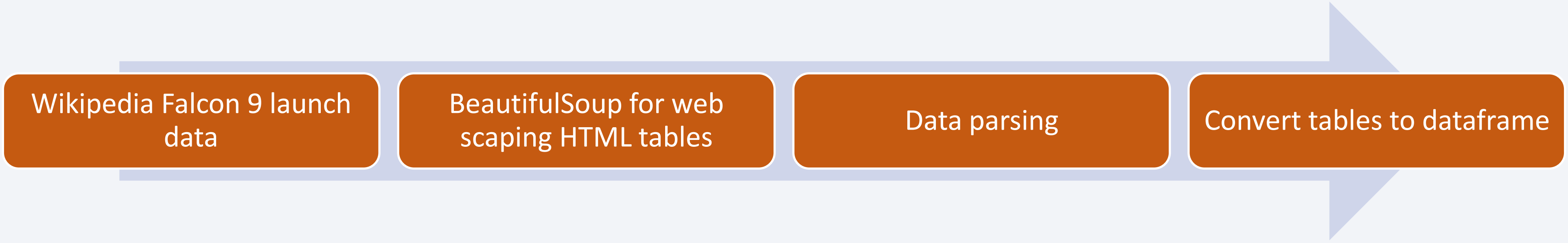
SpaceX REST API endpoint

Get request

JSON object with launch data

Covert JSON to dataframe

  - Web scraping source: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



Wikipedia Falcon 9 launch data

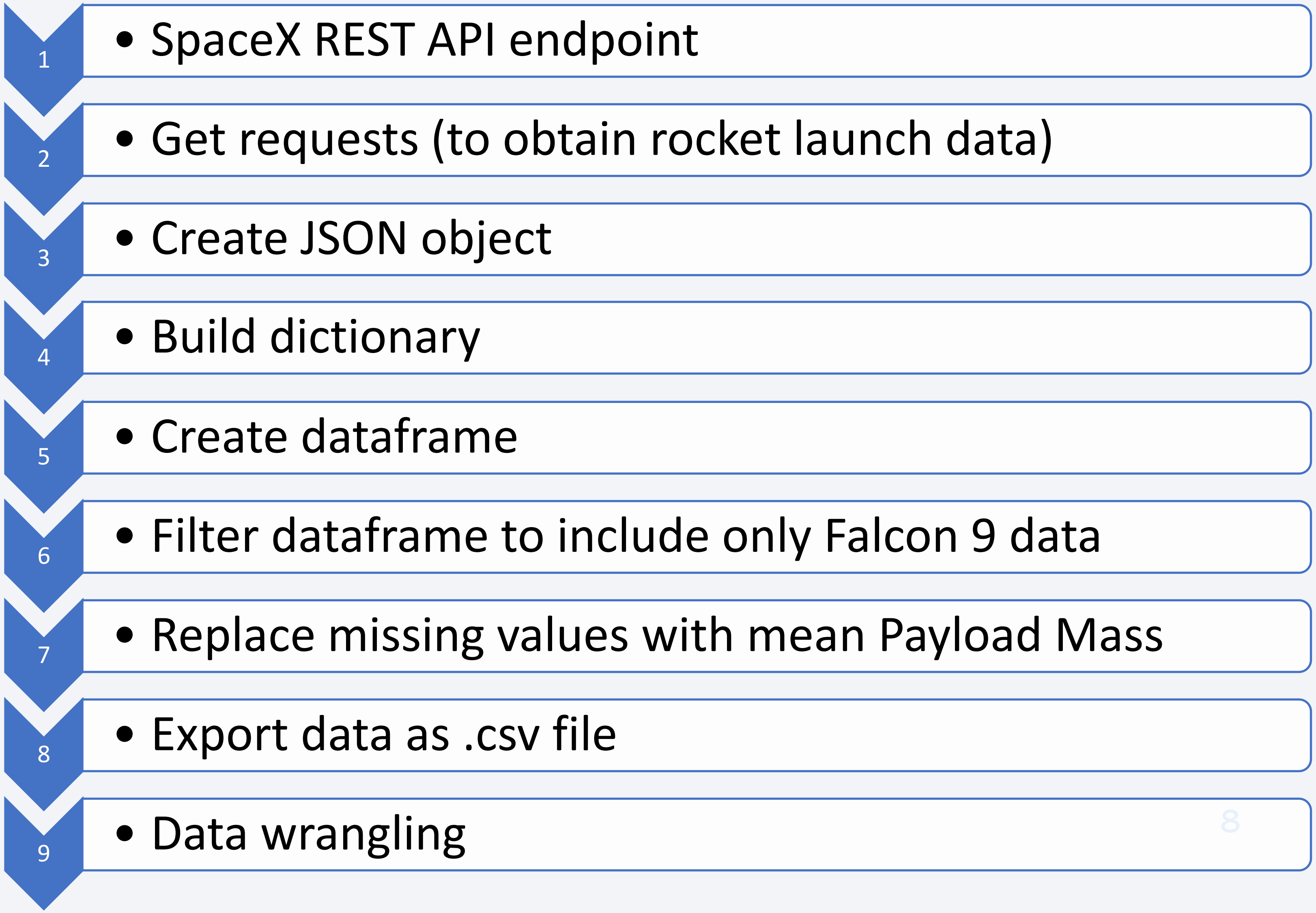
BeautifulSoup for web scaping HTML tables

Data parsing

Convert tables to dataframe

Data Collection – SpaceX API

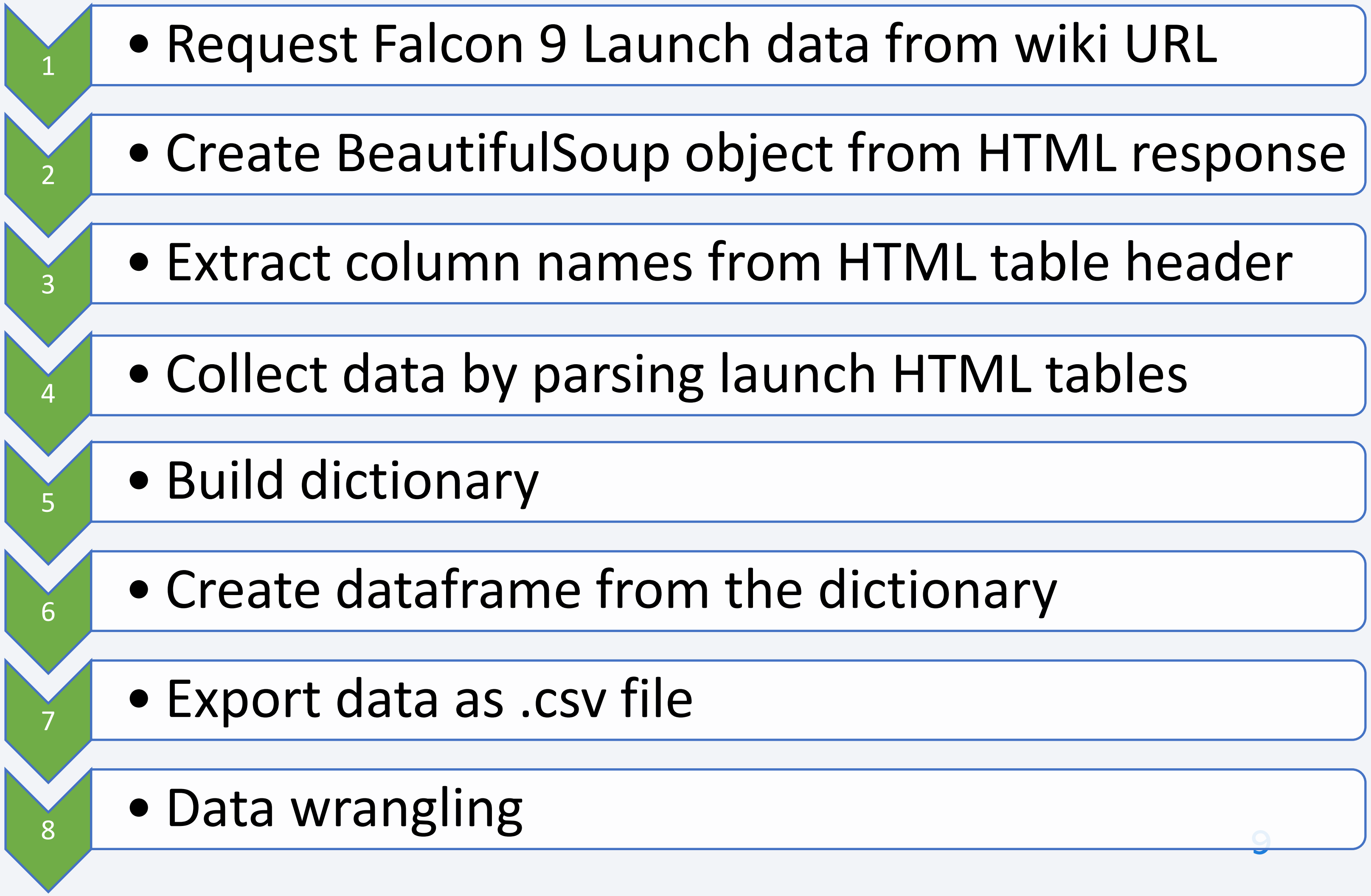
- Jupyter notebook #1 - data collection from API



1. SpaceX REST API endpoint
2. Get requests (to obtain rocket launch data)
3. Create JSON object
4. Build dictionary
5. Create dataframe
6. Filter dataframe to include only Falcon 9 data
7. Replace missing values with mean Payload Mass
8. Export data as .csv file
9. Data wrangling

Data Collection - Scraping

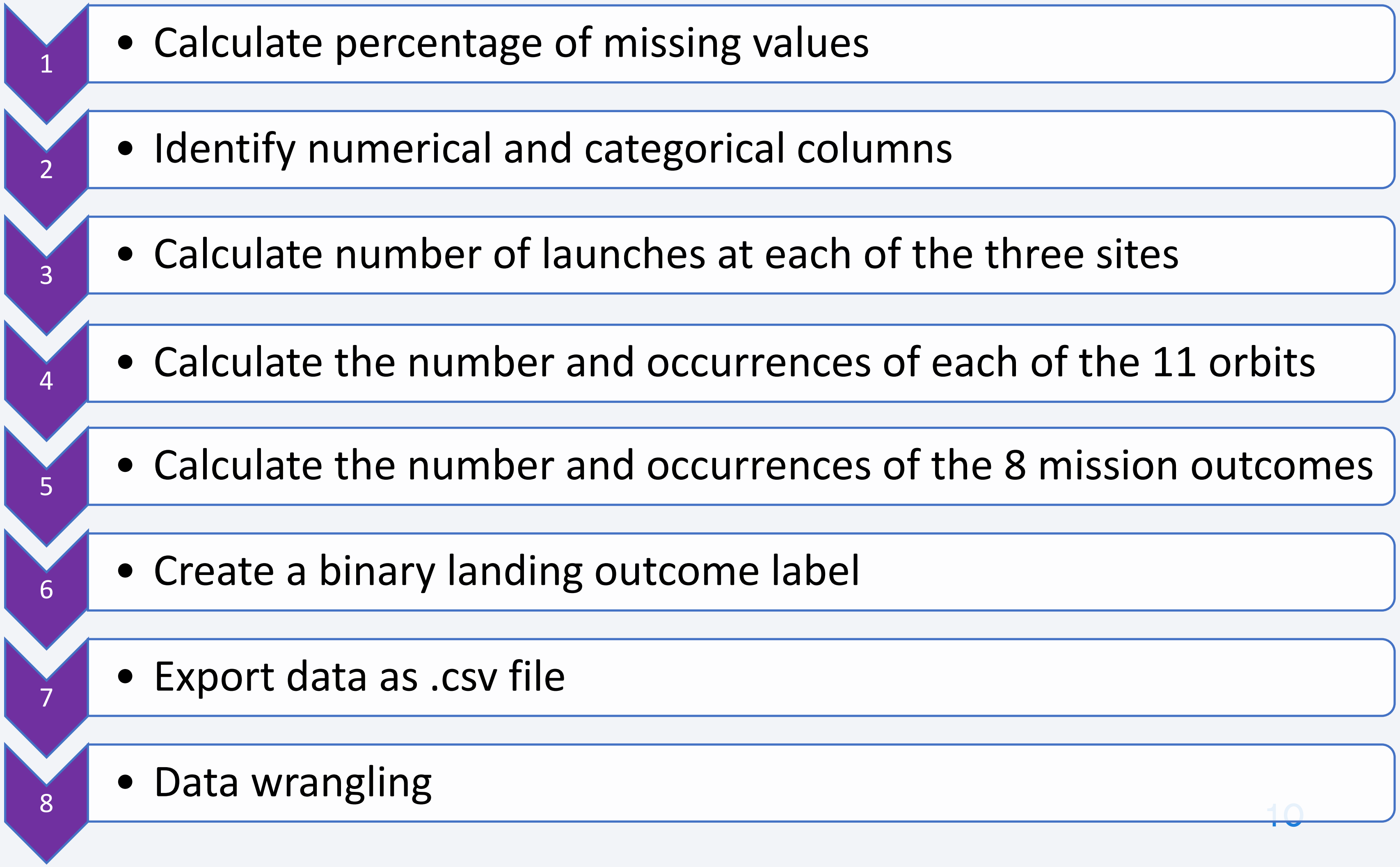
- Jupyter notebook #2 - data collection using web scraping



1. Request Falcon 9 Launch data from wiki URL
2. Create BeautifulSoup object from HTML response
3. Extract column names from HTML table header
4. Collect data by parsing launch HTML tables
5. Build dictionary
6. Create dataframe from the dictionary
7. Export data as .csv file
8. Data wrangling

Data Wrangling

- This step includes Exploratory Data Analysis (EDA) followed by summaries and creation of the Landing Outcome variable which will be the label for supervised models
- Landing Outcome is a binary variable with 1 representing a successful landing and 0 an unsuccessful landing
- Jupyter notebook #3 – data wrangling



1
- Calculate percentage of missing values

2
- Identify numerical and categorical columns

3
- Calculate number of launches at each of the three sites

4
- Calculate the number and occurrences of each of the 11 orbits

5
- Calculate the number and occurrences of the 8 mission outcomes

6
- Create a binary landing outcome label

7
- Export data as .csv file

8
- Data wrangling

EDA with Data Visualization

- Several scatterplots were built to be explore relationships between continuous variables and how these affected the launch outcome (that is, whether the launch was successful or not):
  - FlightNumber vs PayloadMass
  - FlightNumber vs LaunchSite
  - PayloadMass vs. LaunchSite
  - FlightNumber vs. Orbit
  - Payload vs. Orbit
- A bar chart was created to compare the launch success rate by orbit type
- A line chart was created to see how launch success varied over time

- Jupyter notebook #4 – EDA with Visualization

EDA with SQL

- SQL queries you performed
  - Display the names of the unique launch sites
  - Display the 5 launch sites with names that began with CCA
  - Display the total payload mass carried by booster launched by NASA (CRS)
  - Display the average payload mass carried by booster version F9 v1.1
  - List the date where the successful landing outcome in drone ship was achieved
  - List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the records which will display the month names, successful landing outcomes in ground pad , booster versions, launch site for the months in year 2017
  - Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- Jupyter notebook #5 – EDA with SQL

Build an Interactive Map with Folium

- Map objects created on Folium map:
  - Folium.circle and folium.marker were created to highlight and add text to circle areas around the specific coordinates corresponding to each launch site and to code successful and unsuccessful launch sites
  - MarkerCluster was added to the map to simplify markers with the same coordinates
  - MousePosition was added to obtain coordinates when the mouse is passes over a point on the map to further be able to calculate distances
  - Folium.polyline was created to draw lines and calculate distances between a launch site (VAFB SLC-4E) and individual locations of significance, such as the closest coastline, city, highway, or railway.

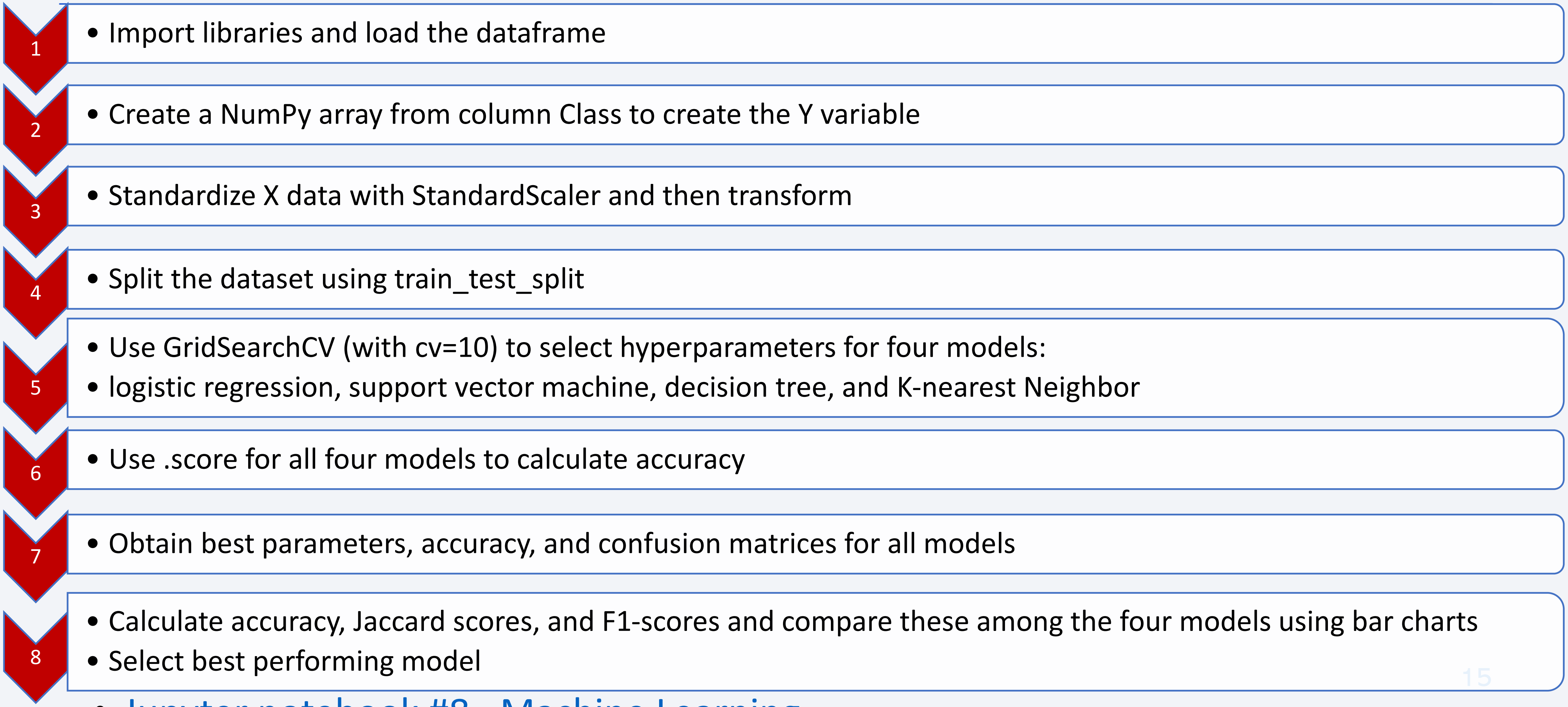
- [Jupyter notebook #6 – Folium map]()

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard:
    - **Dropdown list** with the four launch sites was created to allow user selection of individual sites
    - **Success pie chart** was created using a callback function to allow a visual depiction of successful lunch counts per site or per all sites
    - **Slider on payload mass range** was added to allow an easy observation of how the variation in payload affects the success count per launch site
    - **Scatterplot of payload mass vs success count** color coded for the five different booster versions

- Jupyter notebook #7 – Plotly Dash code

Predictive Analysis (Classification)



1
- Import libraries and load the dataframe

2
- Create a NumPy array from column Class to create the Y variable

3
- Standardize X data with StandardScaler and then transform

4
- Split the dataset using train_test_split

5
- Use GridSearchCV (with cv=10) to select hyperparameters for four models:
- logistic regression, support vector machine, decision tree, and K-nearest Neighbor

6
- Use .score for all four models to calculate accuracy

7
- Obtain best parameters, accuracy, and confusion matrices for all models

8
- Calculate accuracy, Jaccard scores, and F1-scores and compare these among the four models using bar charts
- Select best performing model

- Jupyter notebook #8 –Machine Learning

Results

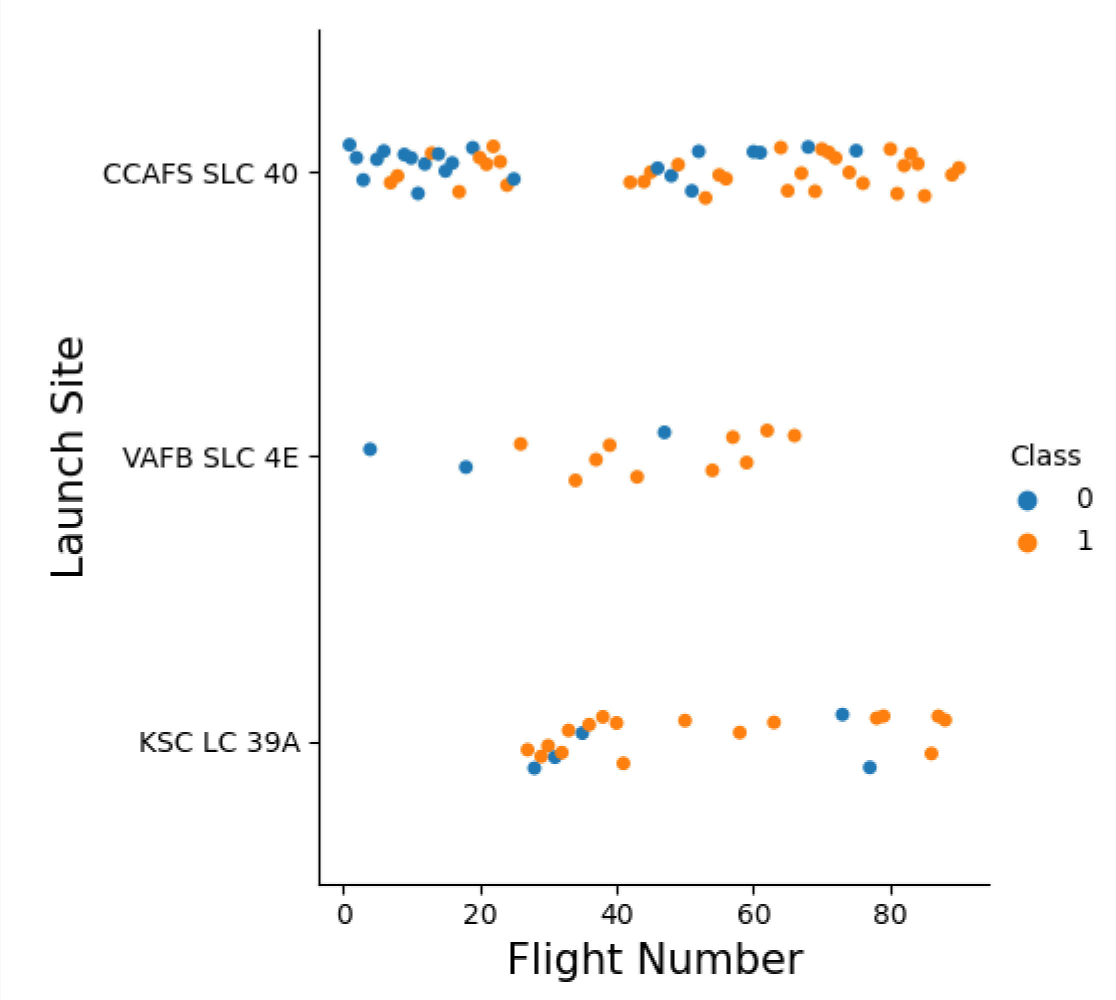
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

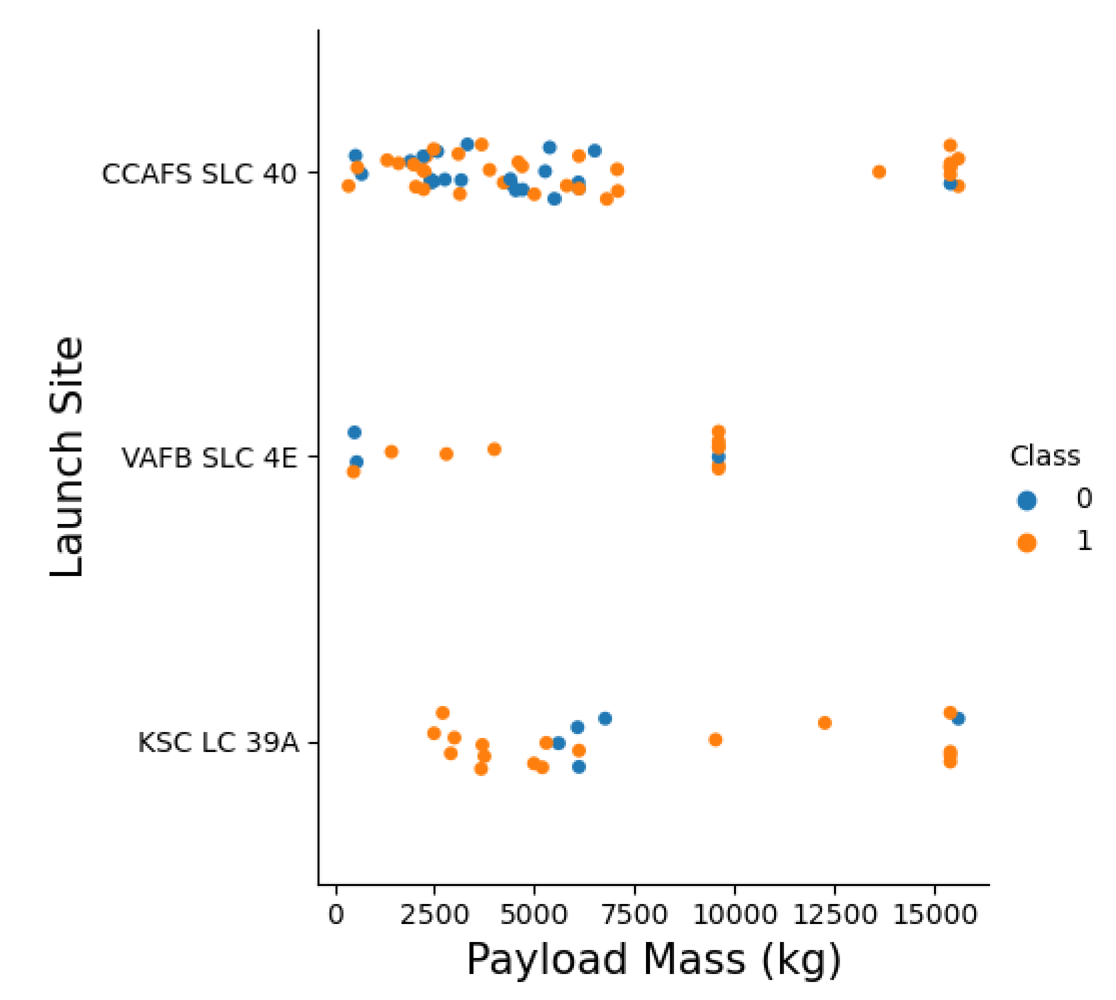
Flight Number vs. Launch Site

- Regardless of the launch site, more recent flights (that is, larger numbers on the x-axis) had greater success rates than earlier flights (more orange dots on the right side of the chart)
- Most launches were from SLC 40, but SLC 4E and LC 39A had higher success rates
- SCL 40 had very low success rates during the early flights which were mostly launched from this site



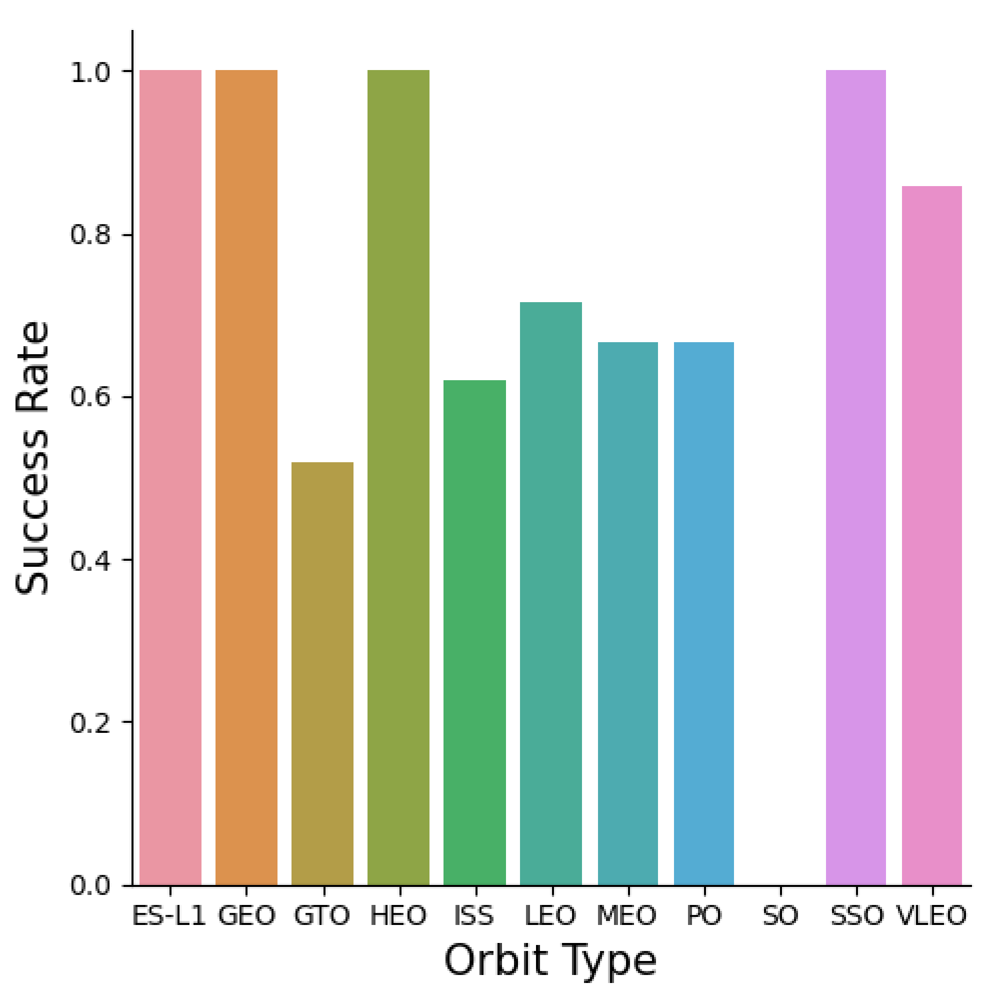
Payload vs. Launch Site

- Regardless of launch site, the larger the payload mass (especially above 7,500 kg), the higher the launch success rate (more orange dots on the right side of the chart)
- SLC 4E site did not launch payload above 10,000 kg
- LC 39A had great success rate below 5,000 kg payload



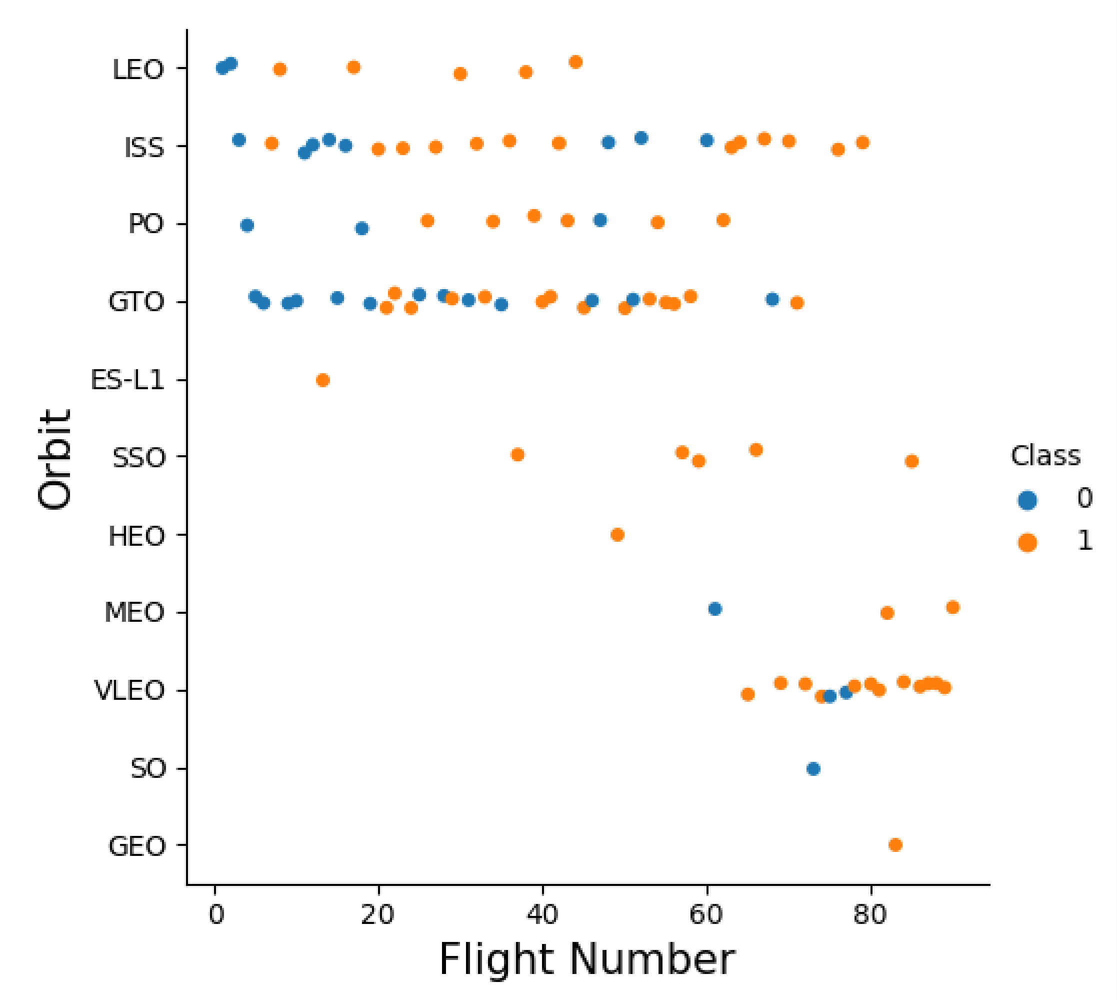
Success Rate vs. Orbit Type

- Success rate ranged from 0 to 100% and depended on the orbit type
  - 0% success rate for SO
  - 50-80% success rate for GTO, ISS, LEO, MEO, PO
  - >85% success rate for ES-L1, GEO, HEO, SSO, VLEO



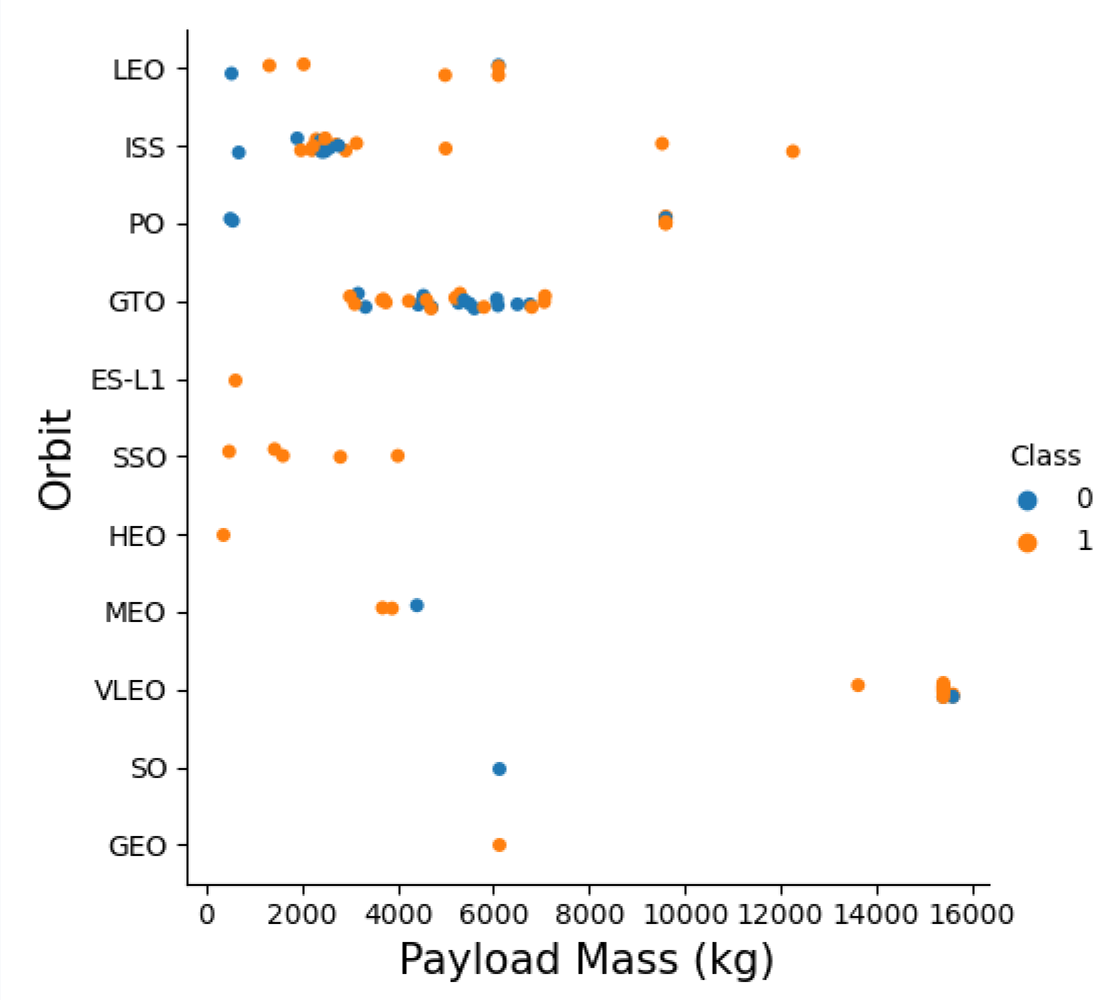
Flight Number vs. Orbit Type

- Regardless of the orbit type, more recent flights had greater success rates than earlier flights (more orange dots on the right side of the chart)
- Most used orbits were LEO, ISS, GTO from the beginning of the program
  - **LEO** had higher success rate than ISS and GTO, but all between 50-80%
- Orbits used after about the 40th flight with several launches (**SSO and VLEO**), had greater success rates (>85%) than earlier used orbits
- ES-L1, HEO, and GEO had 100% success rates but only one launch each

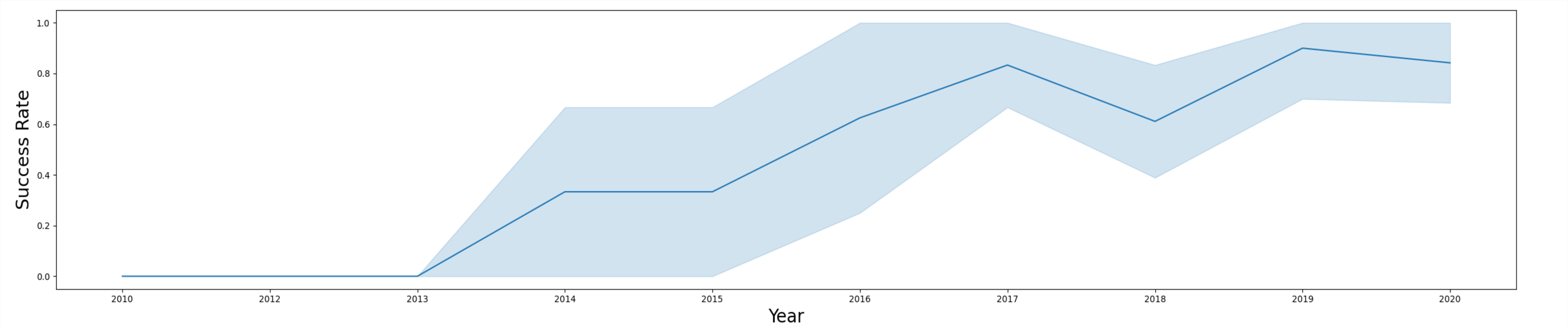


Payload vs. Orbit Type

- Only a few orbits had high payloads (>9,500 kg), including ISS, PO, and VLEO
- The two more recently used orbits that had high success rates from several launches (SSO and VLEO) had payloads at the opposite end of the range
  - SSO (100% success rate) had payloads <5,000 kg
  - VLEO (~85% success rate) had payloads >13,000 kg



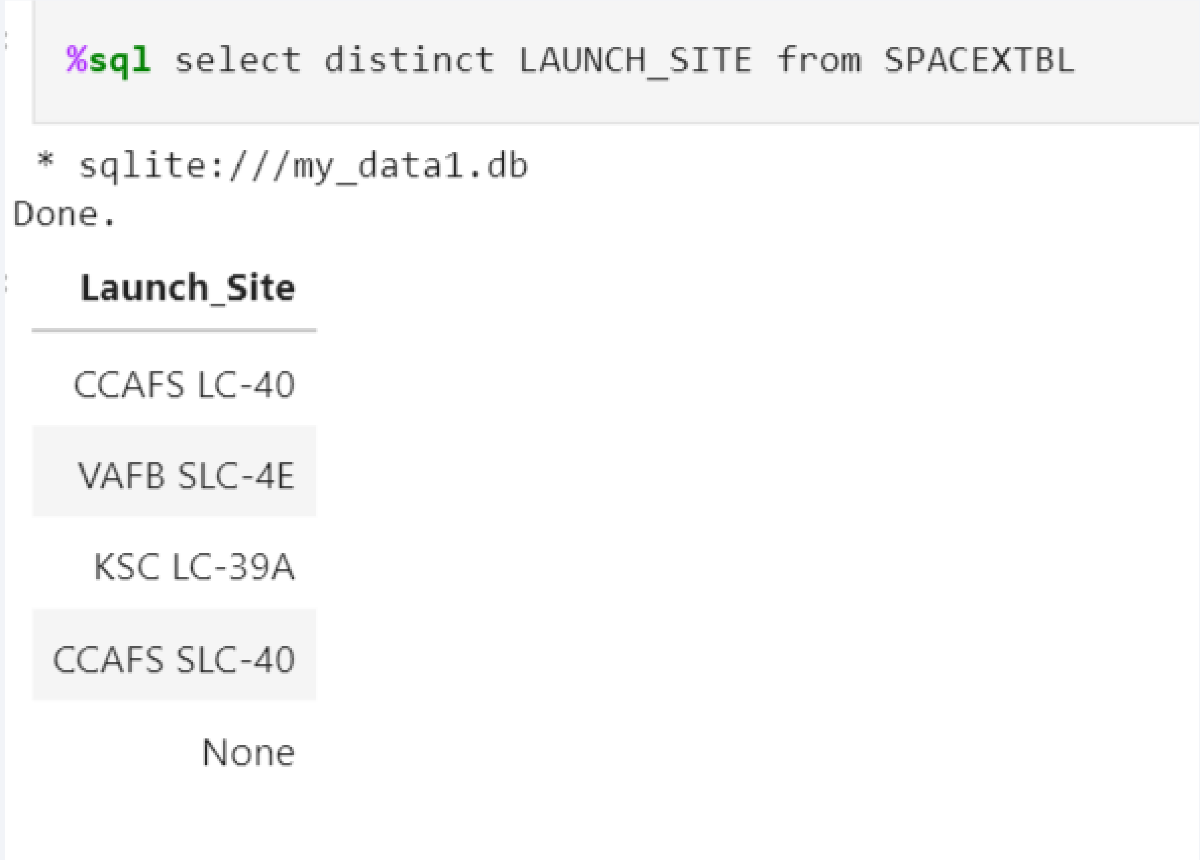
Launch Success Yearly Trend



- Clear trend of increase rate over time after 2013, except for year 2018 which saw a drop in the success rate before increasing again
- Great hike in success rate from 2015 to 2017
- Maximum success rate so far was reached in 2019 (~ 85% success rate)

All Launch Site Names

- There were four unique launch sites
- DISTINCT statement in the query provided the unique launch sites



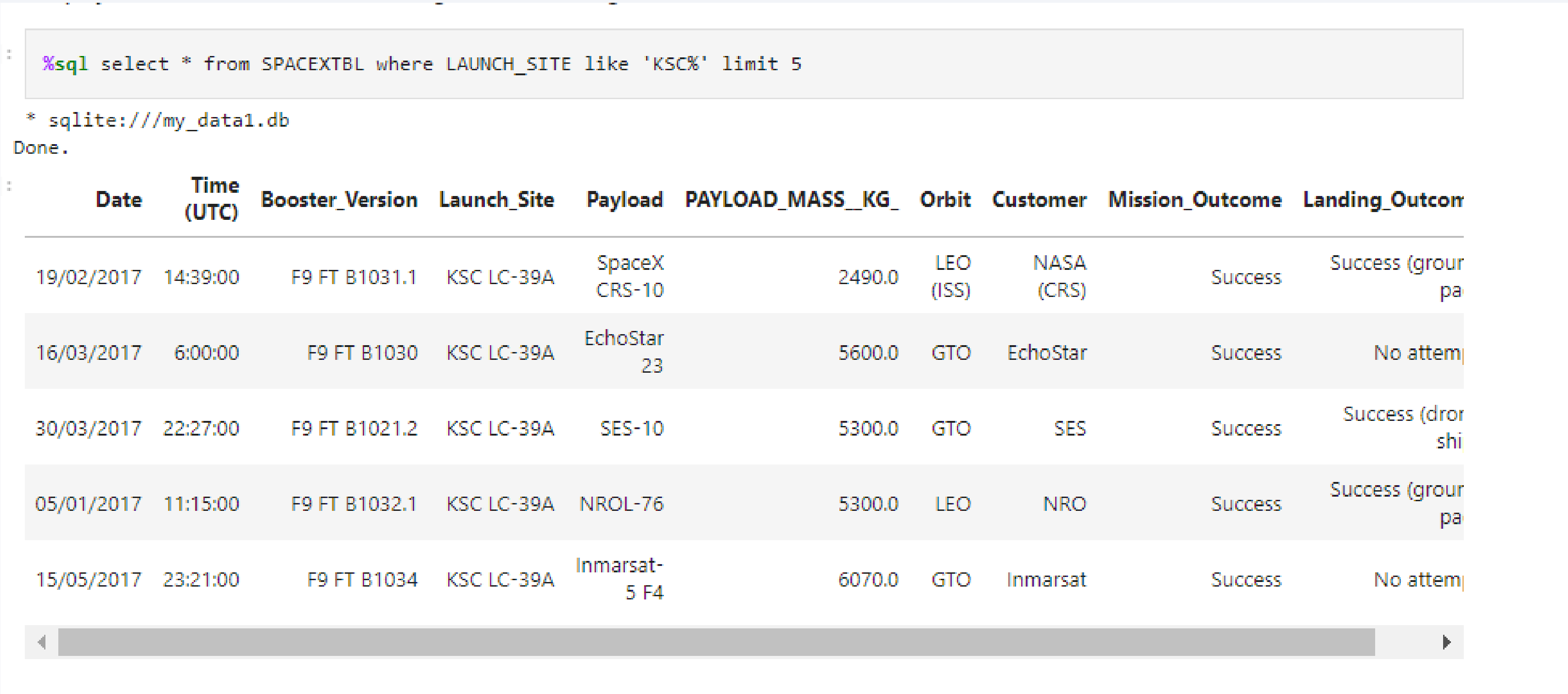
```
%sql select distinct LAUNCH_SITE from SPACEXTBL
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'KSC'

- The table presents 5 records where launch sites' names start with `KSC`
- WHERE LIKE LIMIT statements were used in the query

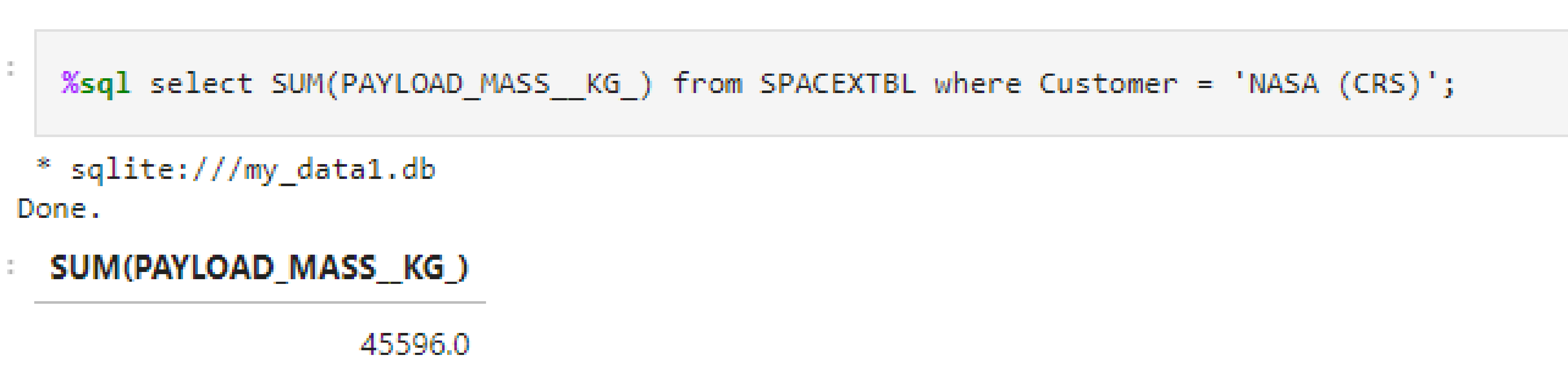


```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'KSC%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcon
19/02/2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490.0	LEO (ISS)	NASA (CRS)	Success	Success (grour pa
16/03/2017	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600.0	GTO	EchoStar	Success	No attem
30/03/2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300.0	GTO	SES	Success	Success (dror shi
05/01/2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300.0	LEO	NRO	Success	Success (grour pa
15/05/2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070.0	GTO	Inmarsat	Success	No attem

Total Payload Mass



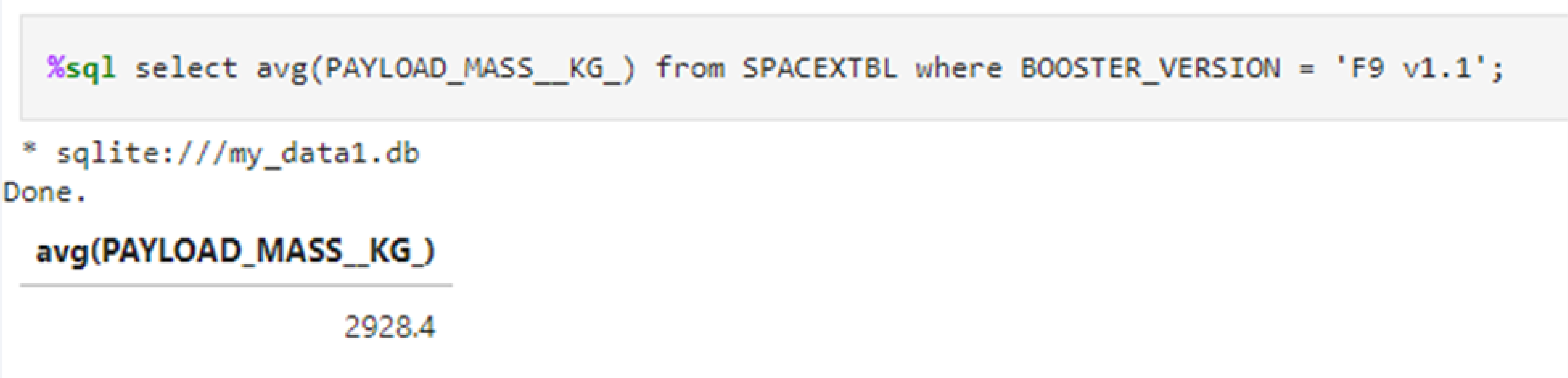
```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596.0

- Total payload carried by boosters from NASA was 45,596 kg
- SUM FROM WHERE statements were used in the query

Average Payload Mass by F9 v1.1



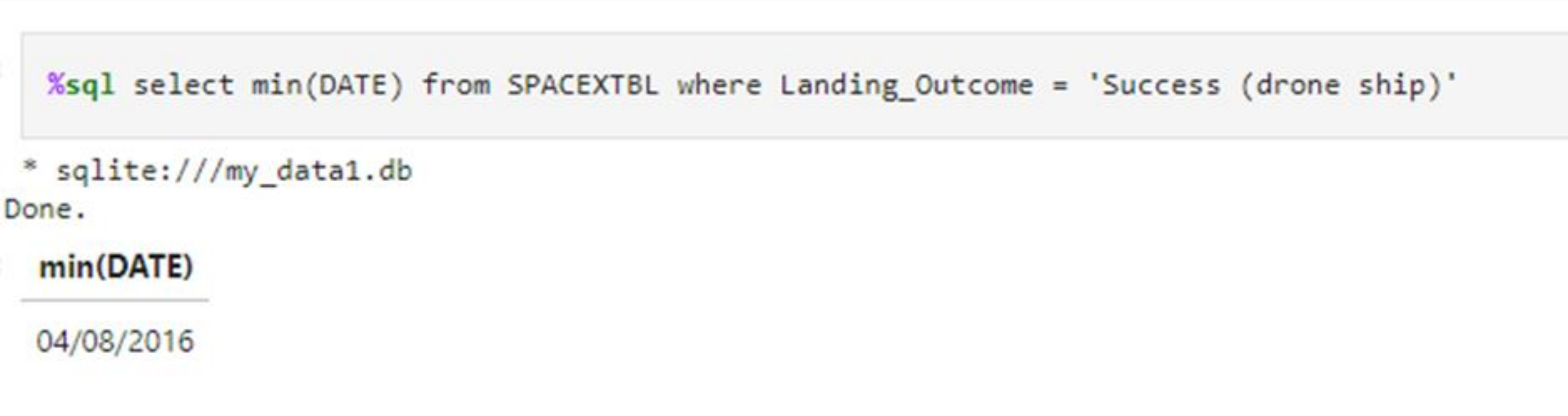
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1';
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

- Average payload carried by booster version F9 v1.1 was 2,928.4 kg
- SUM FROM WHERE statements were used in the query

First Successful Drone Ship Landing Date



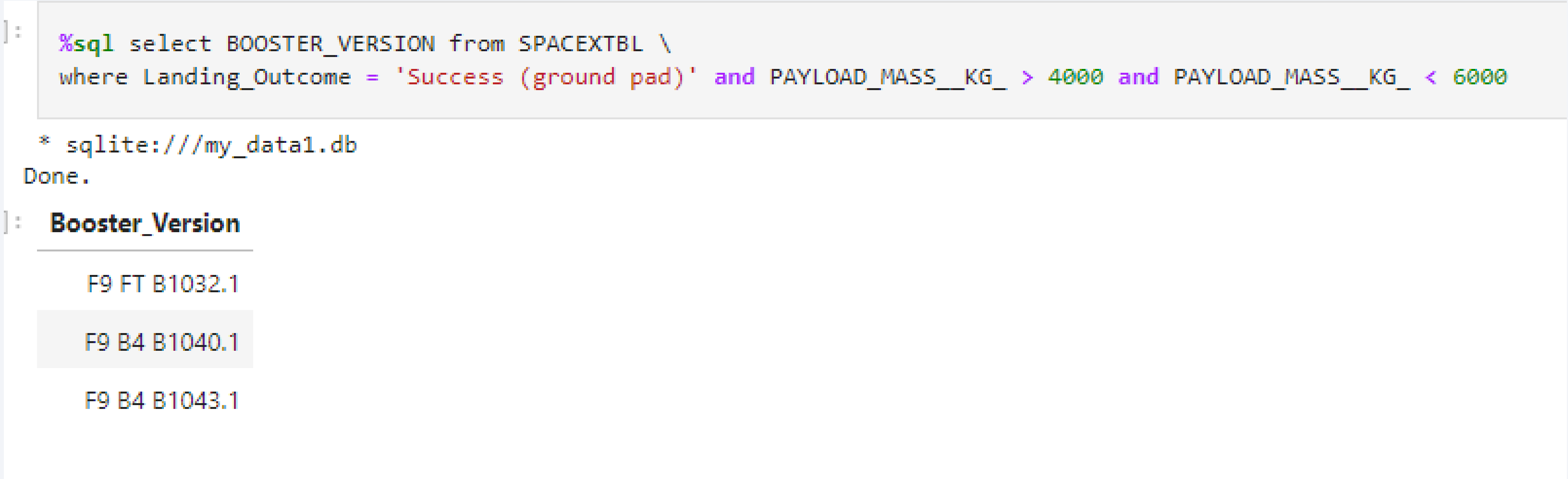
```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (drone ship)'
```

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

min(DATE)
04/08/2016

- First successful landing on a drone ship was on 04/08/2016
- MIN FROM WHERE statements were used in the query

Successful Ground Pad Landing with Payload between 4000 and 6000



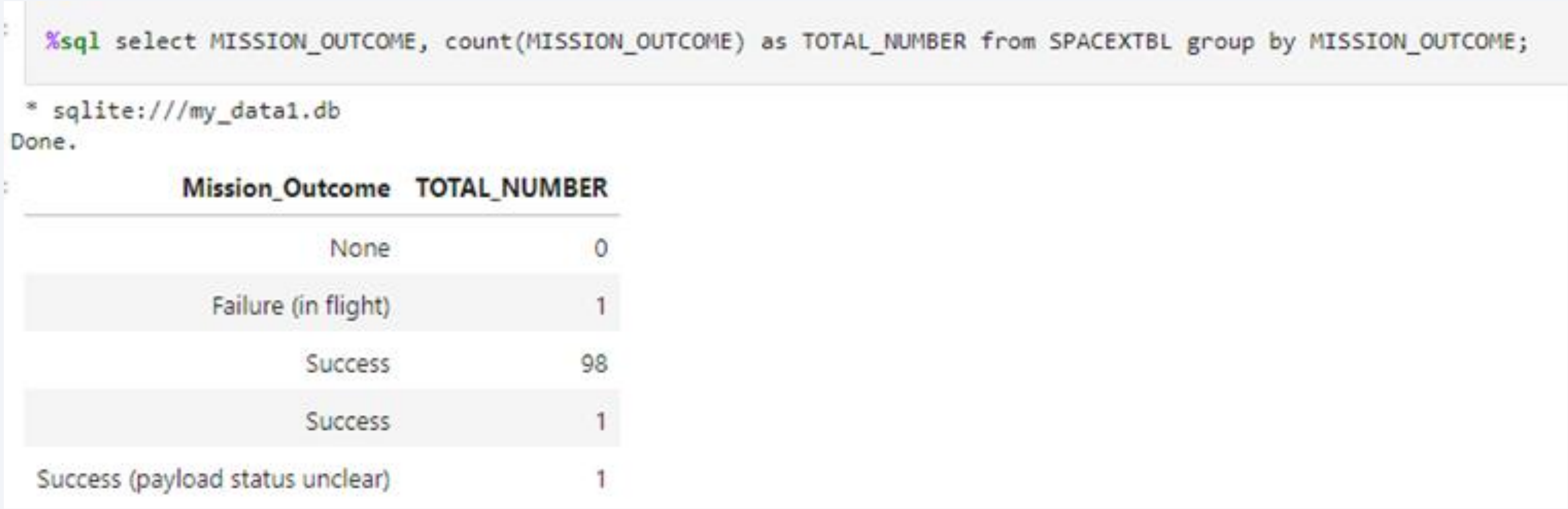
```
%sql select BOOSTER_VERSION from SPACEXTBL \
where Landing_Outcome = 'Success (ground pad)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

- Three booster versions had successfully landed on ground pads carrying loads between 4,000 and 6,000 kg
- FROM WHERE AND statements were used in the query

Total Number of Successful and Failure Mission Outcomes



```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as TOTAL_NUMBER from SPACEXTBL group by MISSION_OUTCOME;
```

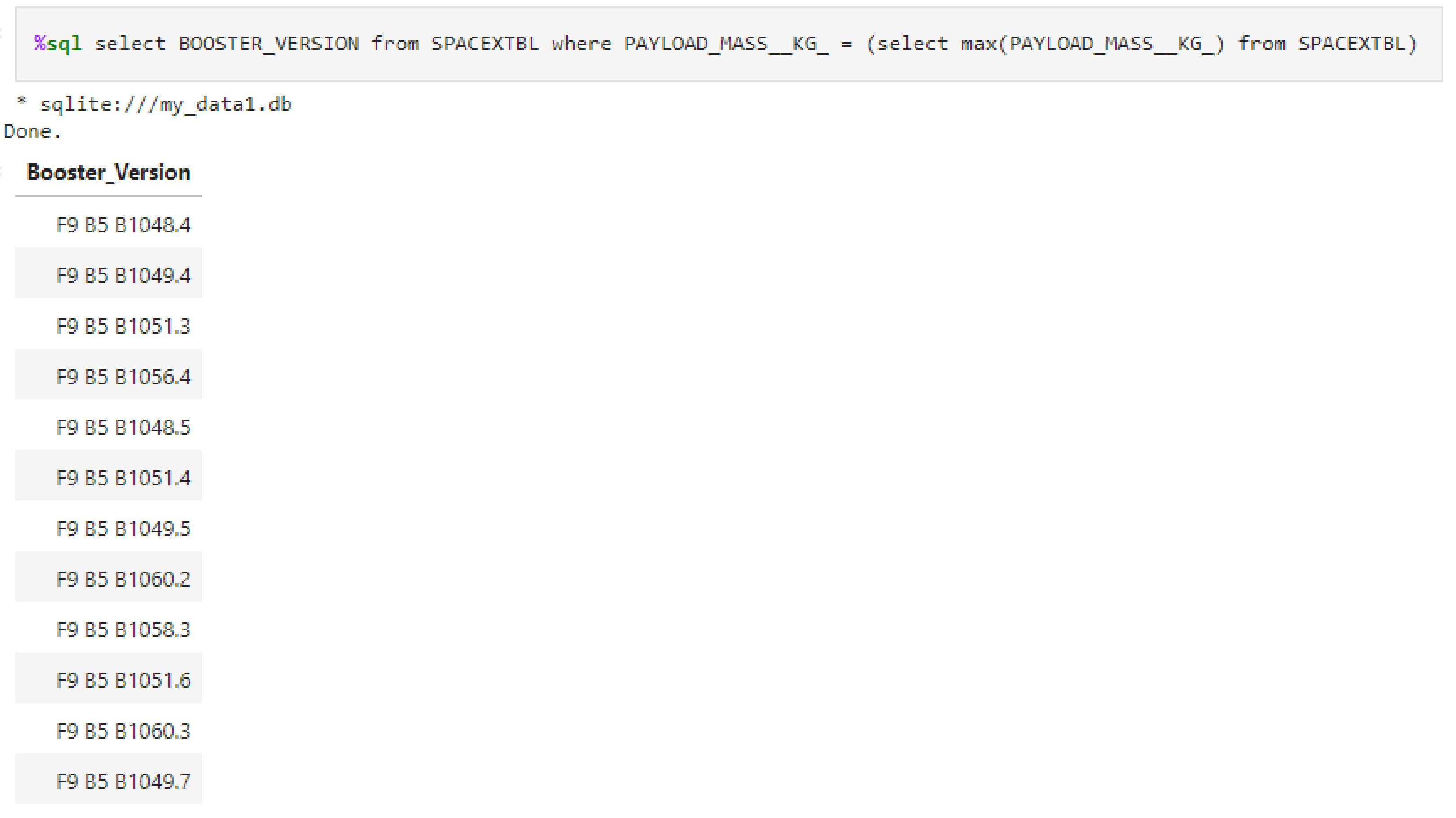
```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	TOTAL_NUMBER
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The number of successful mission outcomes was 100 and there was one failure (in flight)
- COUNT AS GROUP BY statements were used in the query

Boosters Carried Maximum Payload

- There were 12 different boosters that carried maximum payload successfully in space
- FROM WHERE were used in the query
- Subquery with MAX statement was used



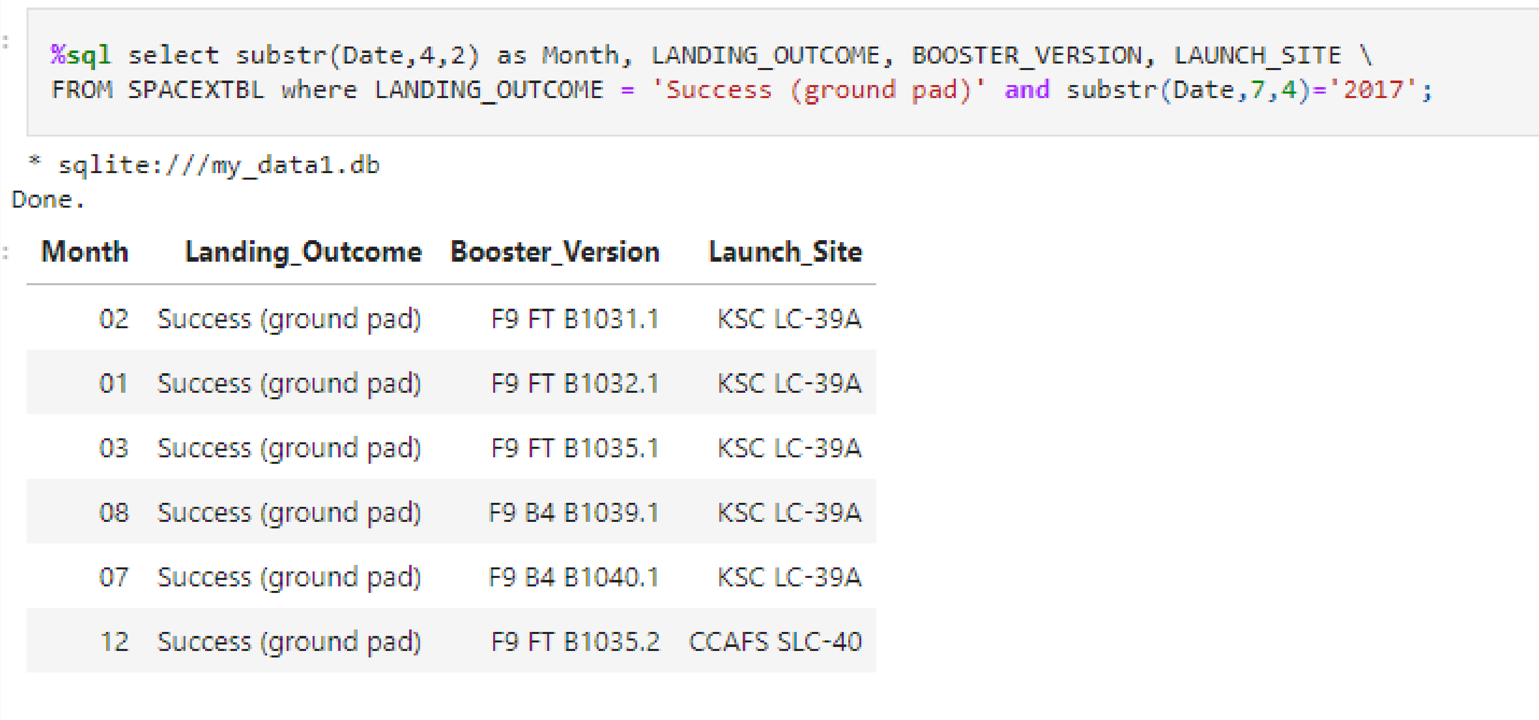
```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * 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

2017 Launch Records

- There were 6 successful landing on ground pad in 2017



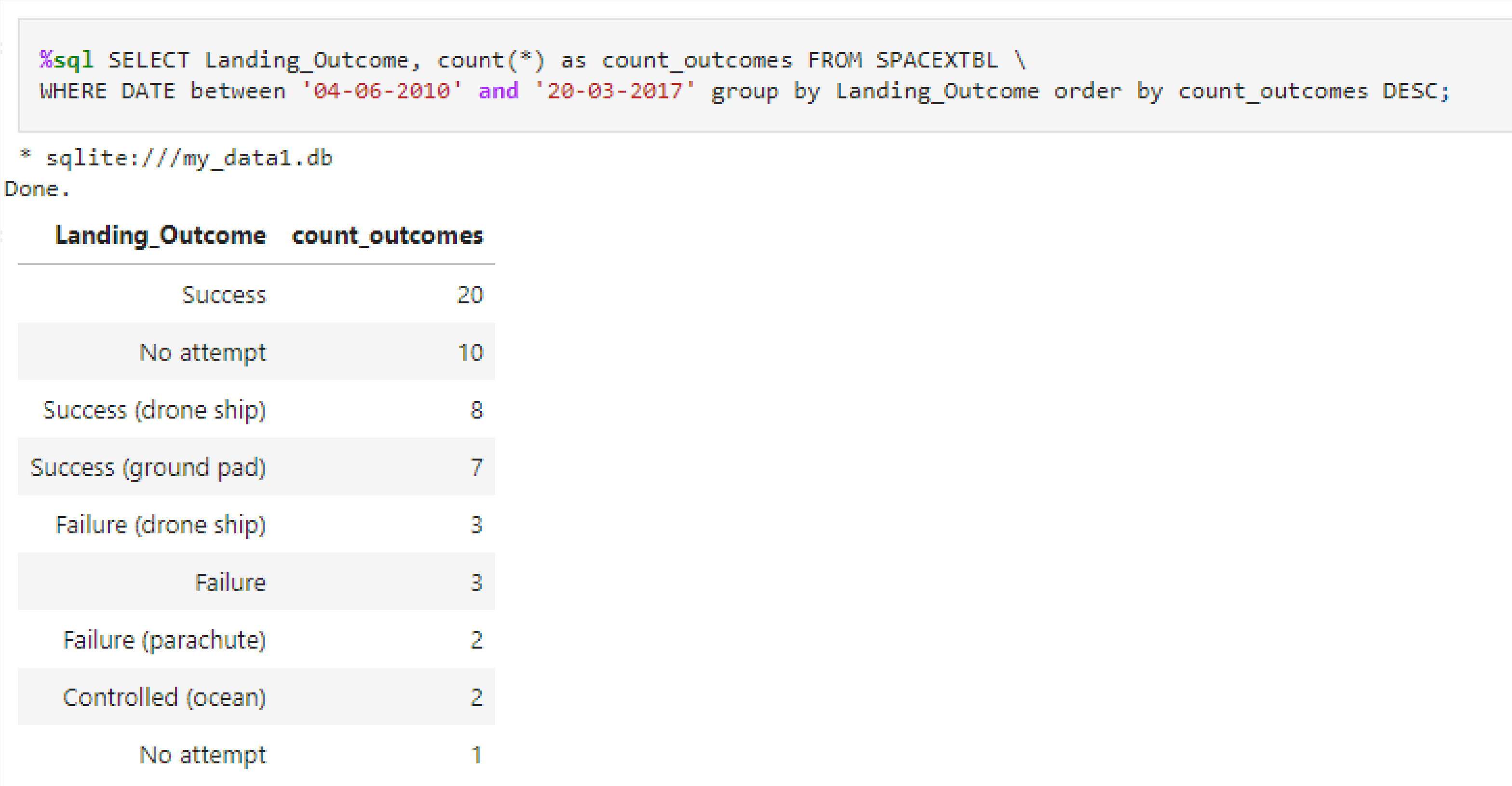
```
%sql select substr(Date,4,2) as Month, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE \
FROM SPACEXTBL where LANDING_OUTCOME = 'Success (ground pad)' and substr(Date,7,4)='2017';
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
01	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
03	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
07	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

- This table shows the rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order
- COUNT WHERE BETWEEN GROUP BY statements used



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

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

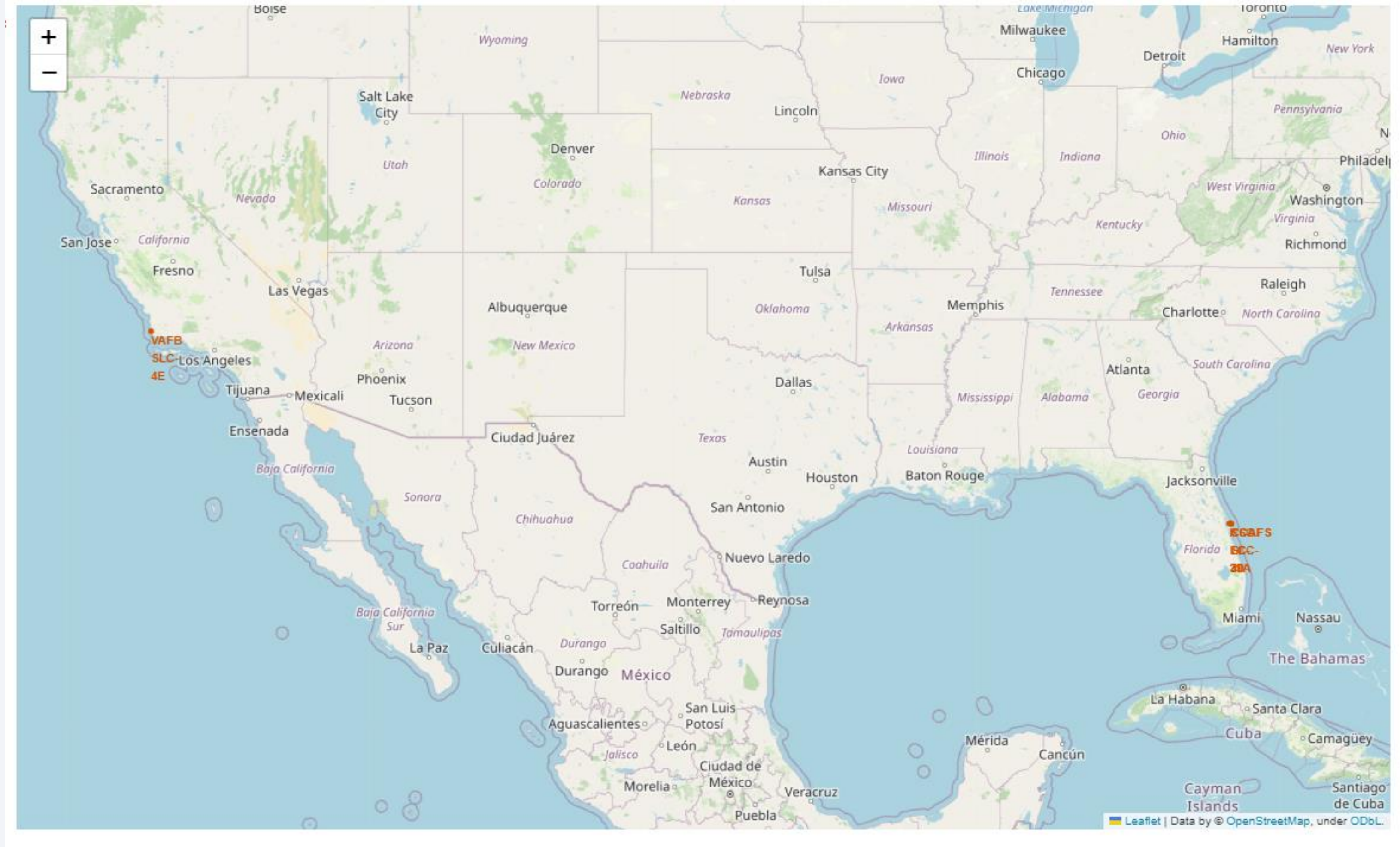
Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

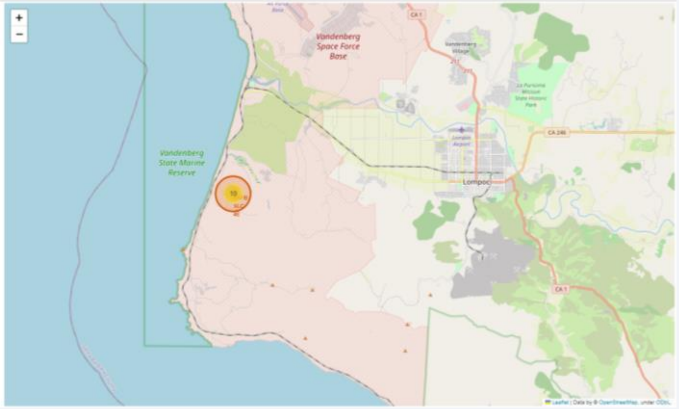
Launch Sites Proximities Analysis

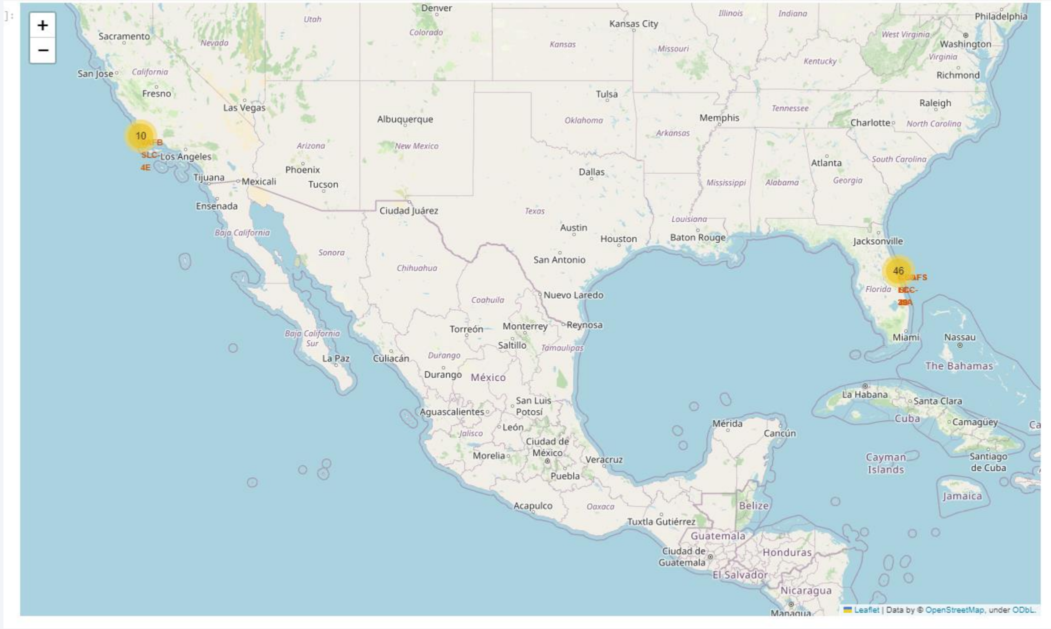
Launch Sites

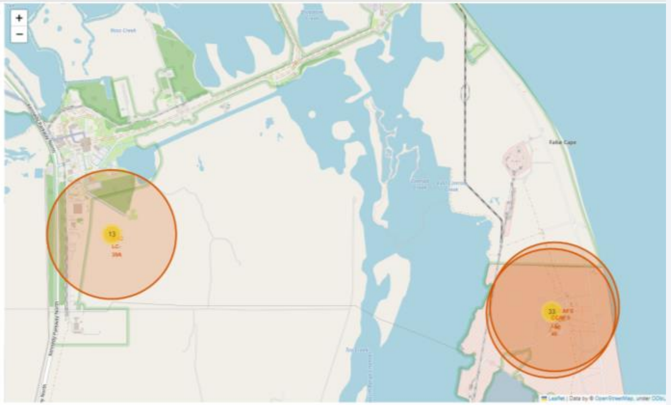
- All launch sites are located on the coast and closer to the equator
    - easier to launch on an equatorial orbit while the rockets also get a boost from the Earth’s rotation speed



Launch Success Rates by Sites



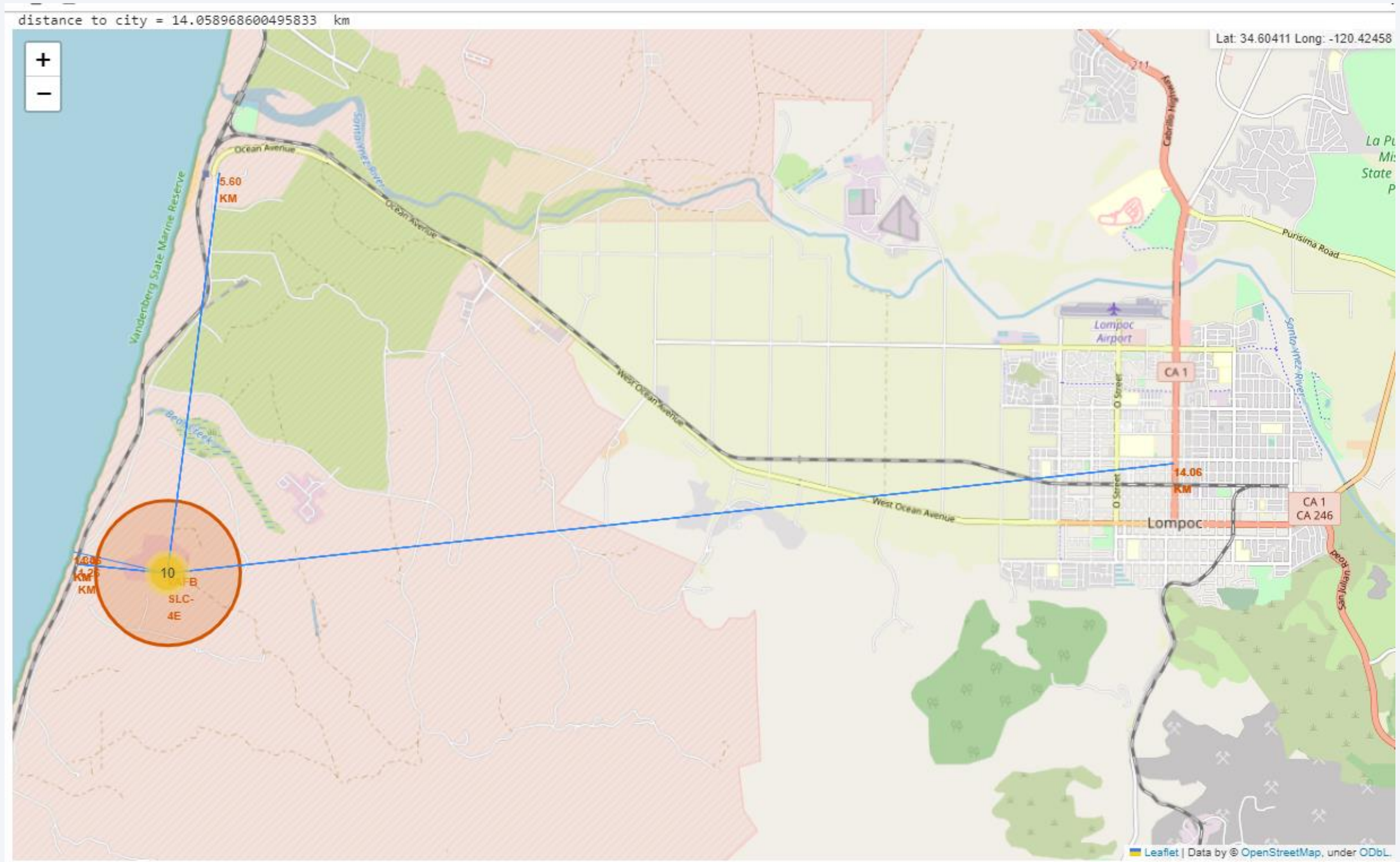




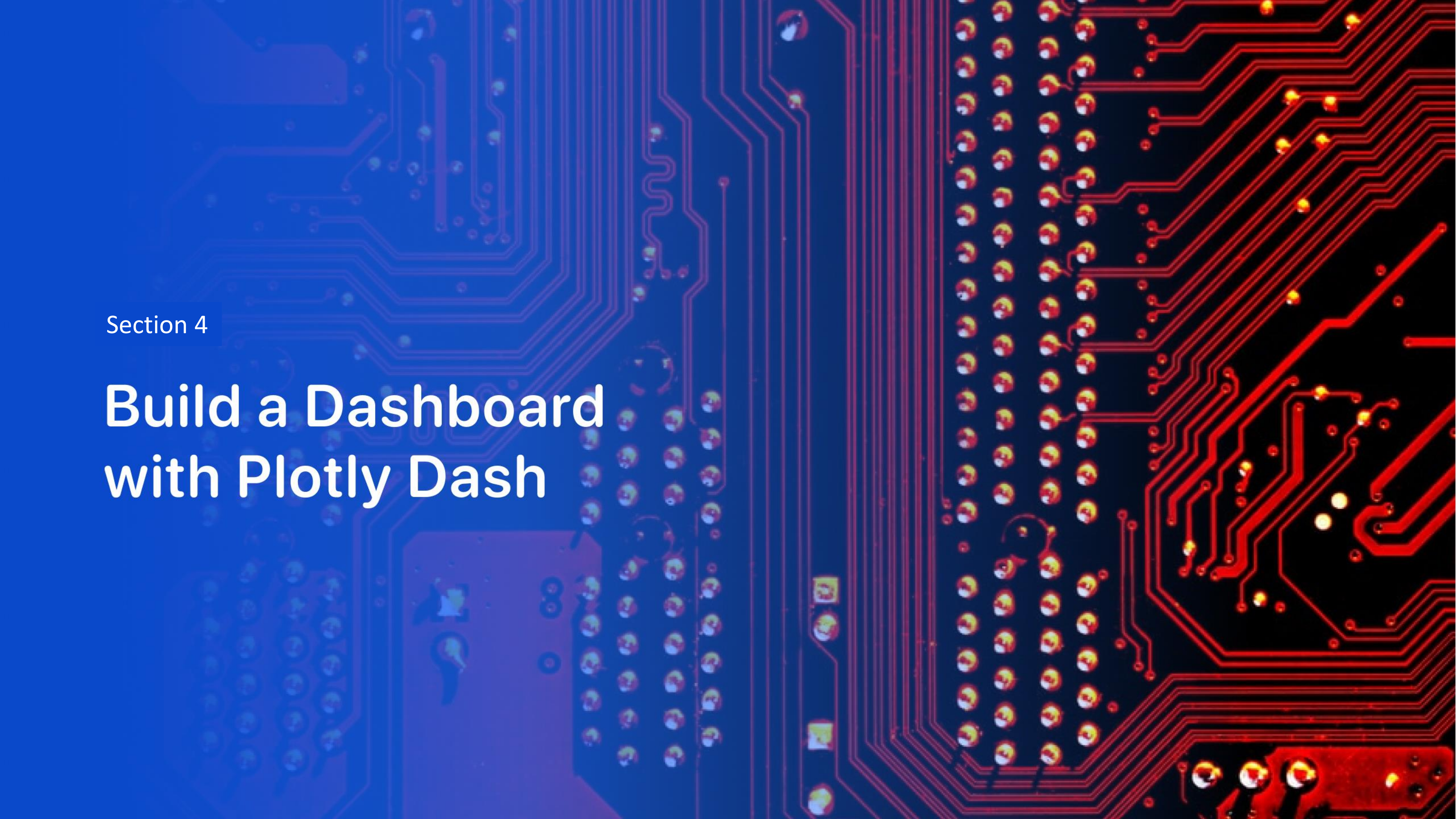
- These maps show the number of successful launches for each site. There were more successful launches from the East Coast sites than from the west Coast site.

VAFB SLC-4E launch site and its proximities

- This site, like the other sites, is very close to the coast (about 1 km) ensuring safety of the population and property in case parts of the rocket fall
- This site is also close to the railroad (about 1 km), to the highway (5.6 km) and to a city (14 km) allowing for transportation of materials and workers to the site while still keeping a safe distance from centers of population (i.e., the city)



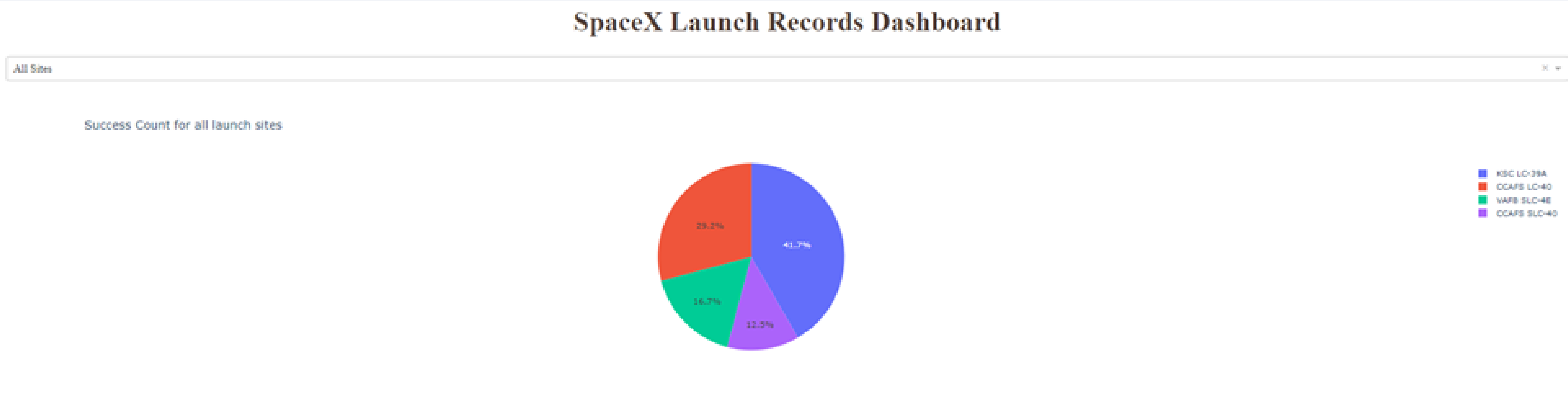
```
distance to city = 14.058968600495833   km
```



Section 4

Build a Dashboard with Plotly Dash

SpaceX Launch Success by Site

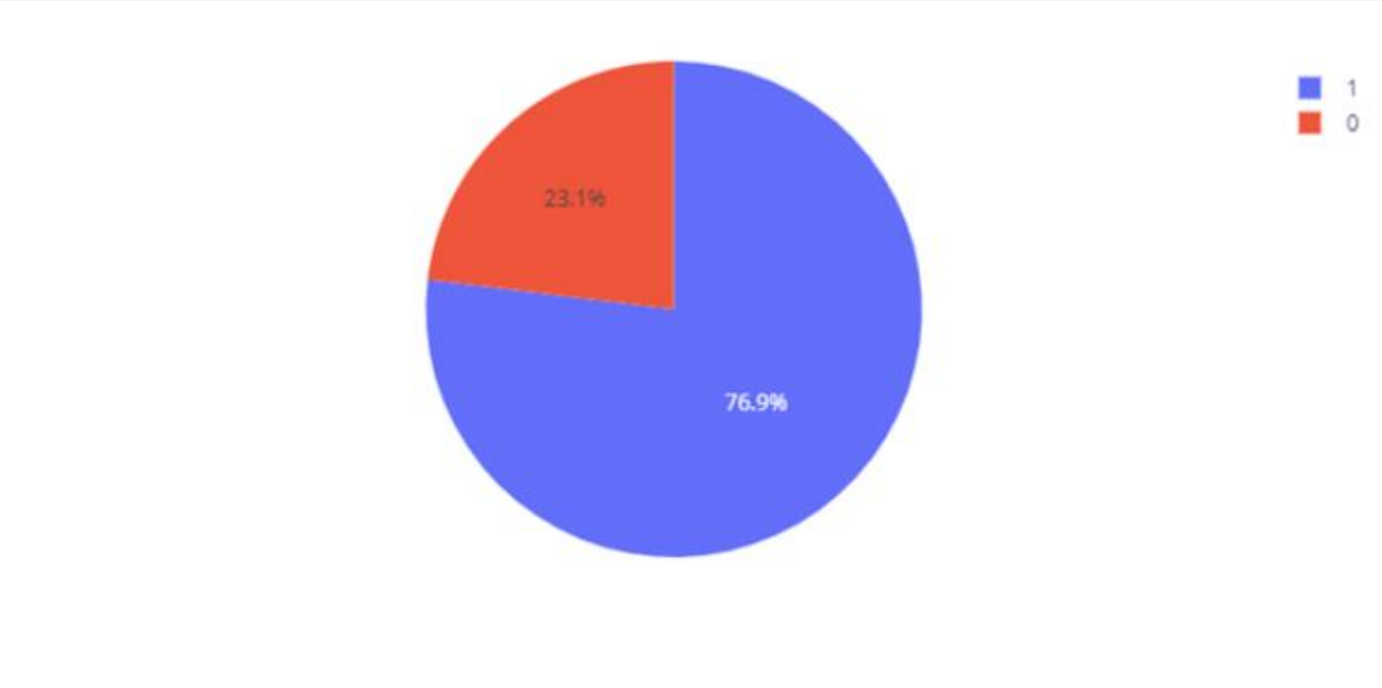


SpaceX Launch Records Dashboard

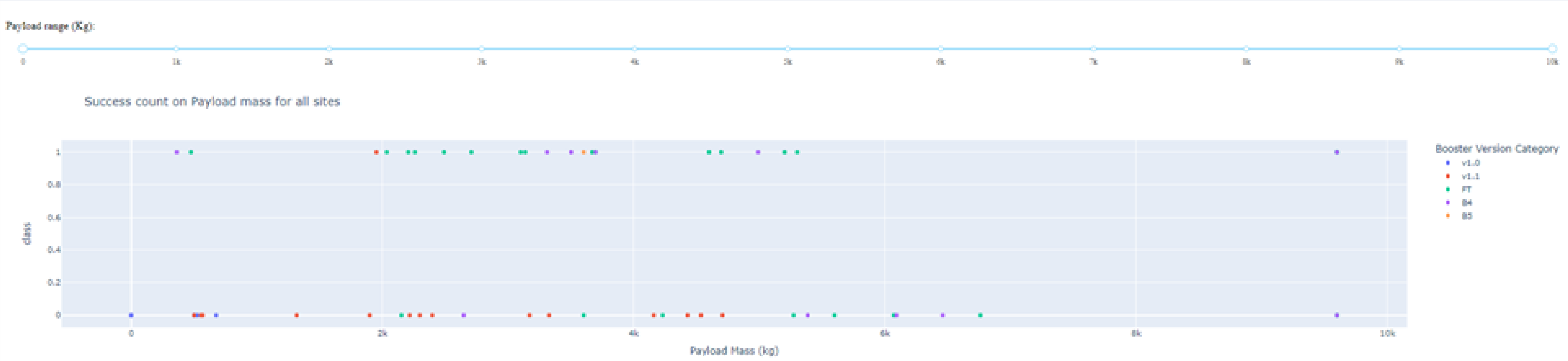
- KSC LC-39A has the highest success rate for launches among the four sites (41.2%)
- CCAFS SLC-40 has the lowest success rate compared to the other sites (12.5%

KSC LC-39A success vs failure rate

- 77% of the launches were successful versus 23% failed at KSC LC-39A, the site with the highest success rate among the four sites used by SpaceX



Payload Mass and Success

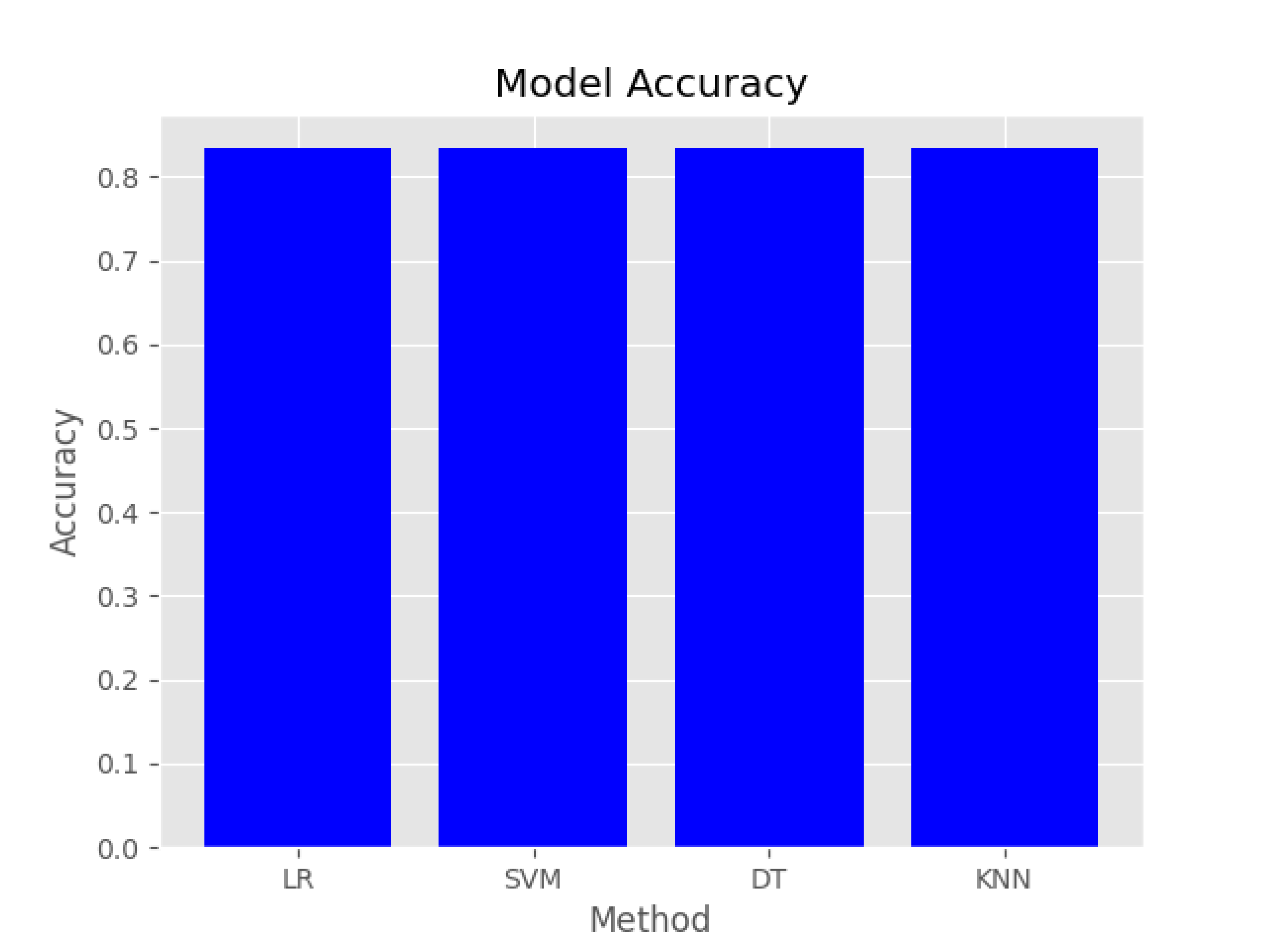


Section 5

Predictive Analysis (Classification)

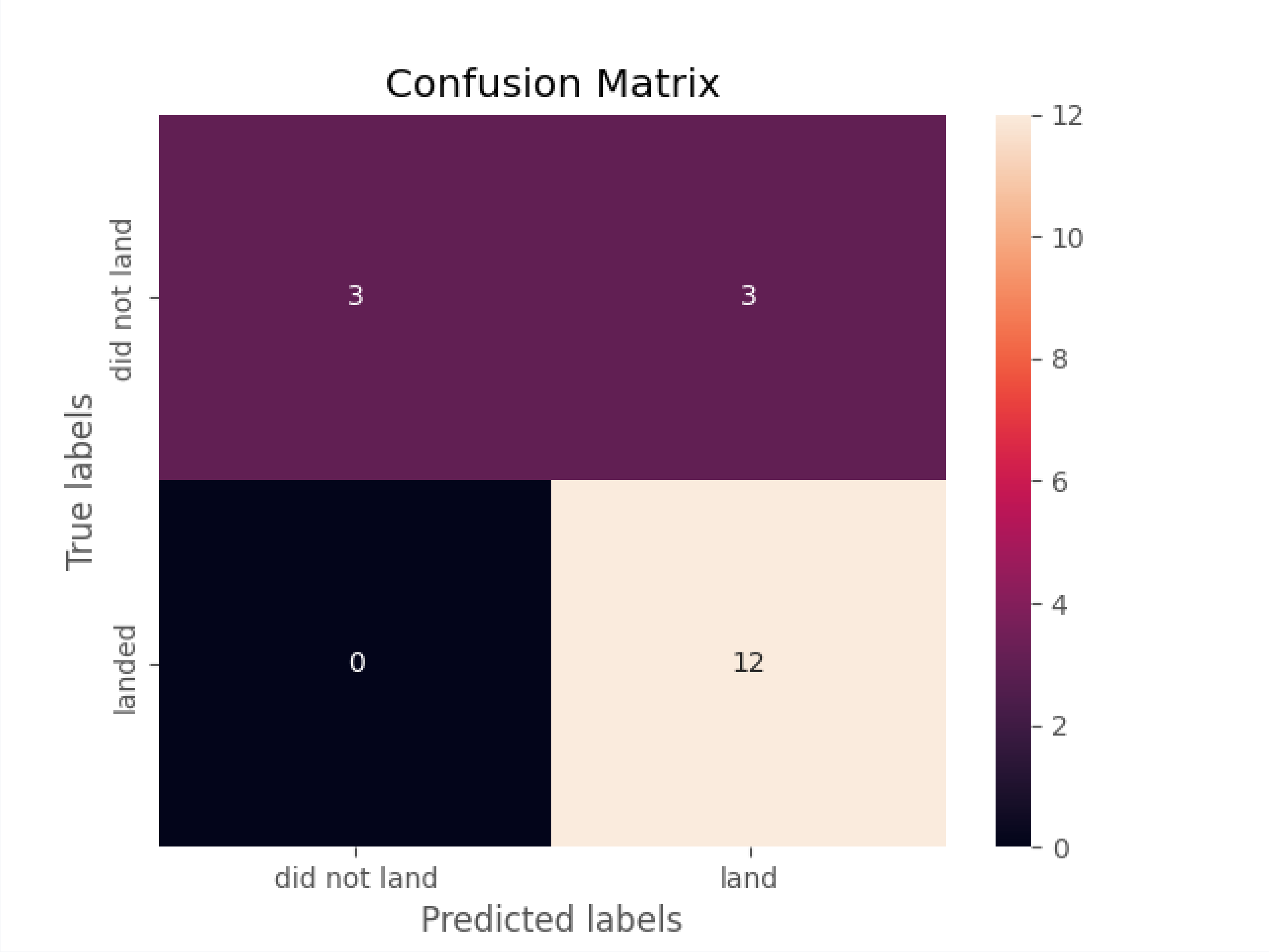
Classification Accuracy

- All four models had the same accuracy, F1-scores, and Jaccard scores



Confusion Matrix

- All four confusion matrices were identical, looking just like the confusion matrix included on this slide



Conclusions

- Launching sites are in coastal areas and close to the equator
- Over time, launch success rate has increased
- The greater the payload mass, the higher the success rate
- KSC LC 39-A is the launch site with the highest success rate, especially for lower payload mass
- SSO, VLEO, and LEO orbits had high success rates based on more than one launch
- Any of the four machine algorithms developed can be used to make launch success predictions with accuracy greater than 0.8
- Future research should consider increasing the dataset and continue to test other algorithms in search for better accuracy

Thank you!