IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thomas Solomon

9th Jan 2023

Outline / Table of contents

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

Executive Summary

- The aim of this project was to predict the success of first-stage landing during SpaceX launches.
- Data was collected using SpaceX Rest API and web scraping from Wikipedia, and then wrangled. Exploratory data analysis (using data viz and SQL) was used to help understand the rate of landing success based on historical data. Interactive maps were build using Folium and a dashboard was created using Plotly Dash. Finally, machine learning was used for predictive analysis (classification). Several machine learning models were applied and model that best predicted landing success was determined.
- Analysis found that KSC LC-39A has the highest success rate, launches with a low payload mass are more successful than launches with a larger payload mass, and a decision tree model is the best performing algorithm for predicting landing success in this data set.

Introduction

- SpaceX advertises Falcon 9 rocket launches, at a cost of $62 million. Other companies cost in excess of $165 million dollars. SpaceX is cheaper because it can reuse the first stage. Therefore, by predicting if the first stage will land, it is possible to determine the cost of a launch. Using publicly-available information and applying machine learning models, it is possible to predict if SpaceX will reuse the first stage during a launch.
- This project will answer the following questions:
  - ✓ *How do variables including payload mass, launch site, number of flights, orbits, etc., affect the success of first-stage landing?*
  - ✓ *Which machine learning classification algorithm is best for predicting success?*

Section 1

Methodology

Methodology

Executive Summary

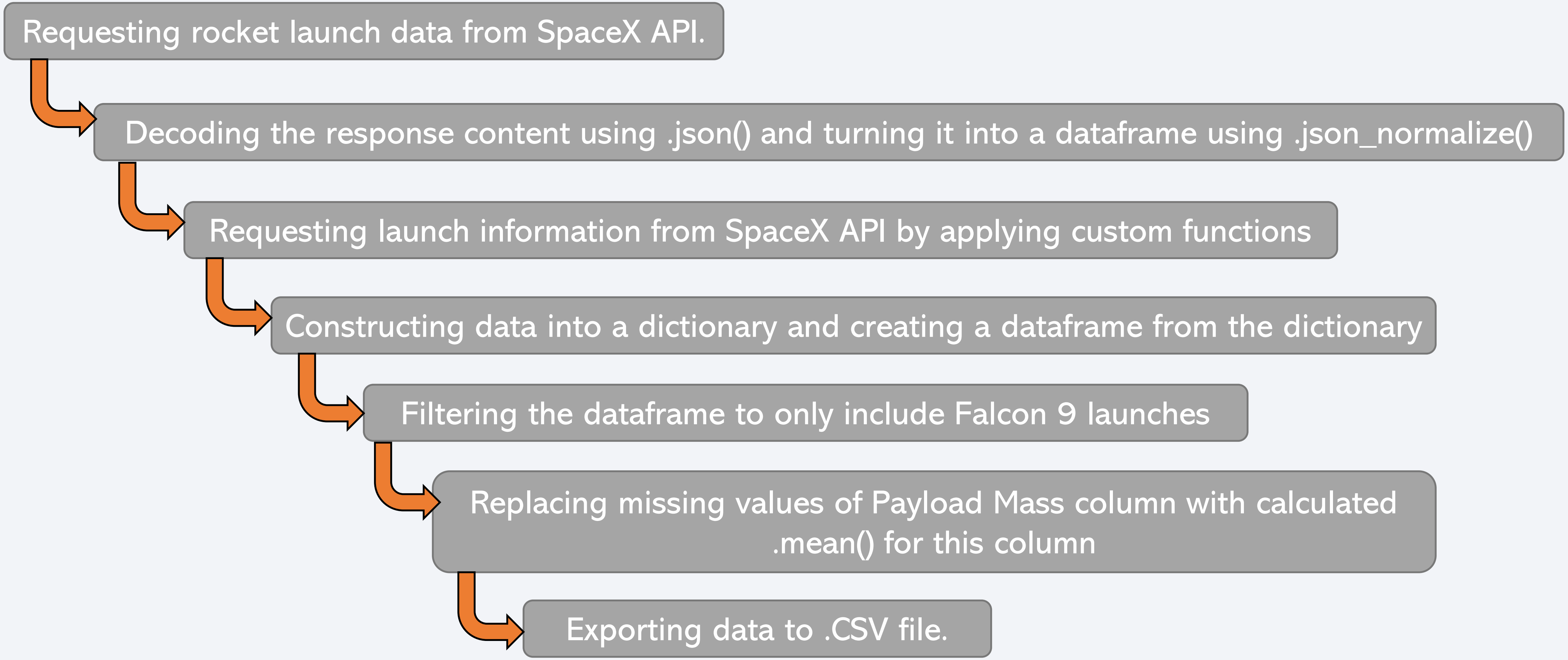
- Data collection methodology:
    - Using SpaceX Rest API and Web Scrapping from Wikipedia
- Perform data wrangling
    - Filtering the data, dealing with missing values, using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Building, tuning and evaluating classification models to ensure the best results.

Data Collection

- To obtain complete data about rocket launches, data collection involved API requests from SpaceX REST API and Web Scraping data from SpaceX's Wikipedia entry.
- SpaceX REST API provided the following: *FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude*.
- Web scraping provided: *Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time*.

Data Collection – SpaceX API

- GitHub URL: https://github.com/tpjsolomon/IBM-data-science-capstone/blob/main/1-Data_collection.ipynb



Requesting rocket launch data from SpaceX API.

Decoding the response content using .json() and turning it into a dataframe using .json_normalize()

Requesting launch information from SpaceX API by applying custom functions

Constructing data into a dictionary and creating a dataframe from the dictionary

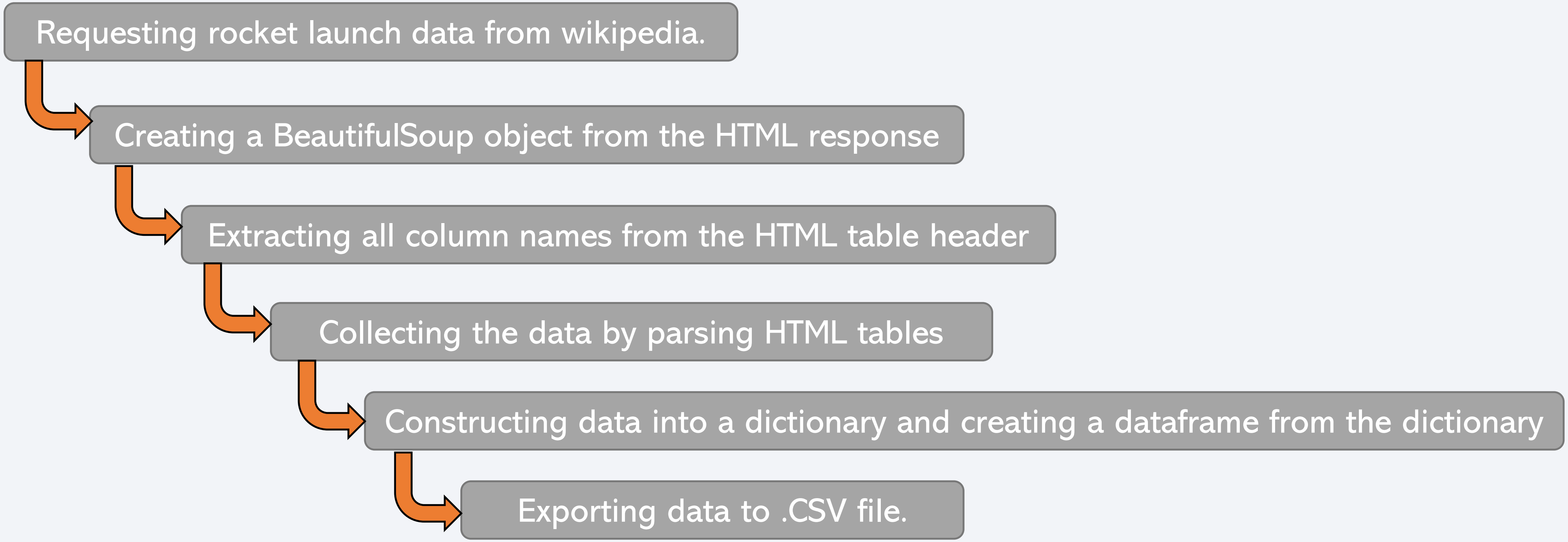
Filtering the dataframe to only include Falcon 9 launches

Replacing missing values of Payload Mass column with calculated .mean() for this column

Exporting data to .CSV file.

Data Collection – Web Scraping

- GitHub URL: https://github.com/tpjsolomon/IBM-data-science-capstone/blob/main/2-Web_scraping.ipynb



Requesting rocket launch data from wikipedia.

Creating a BeautifulSoup object from the HTML response

Extracting all column names from the HTML table header

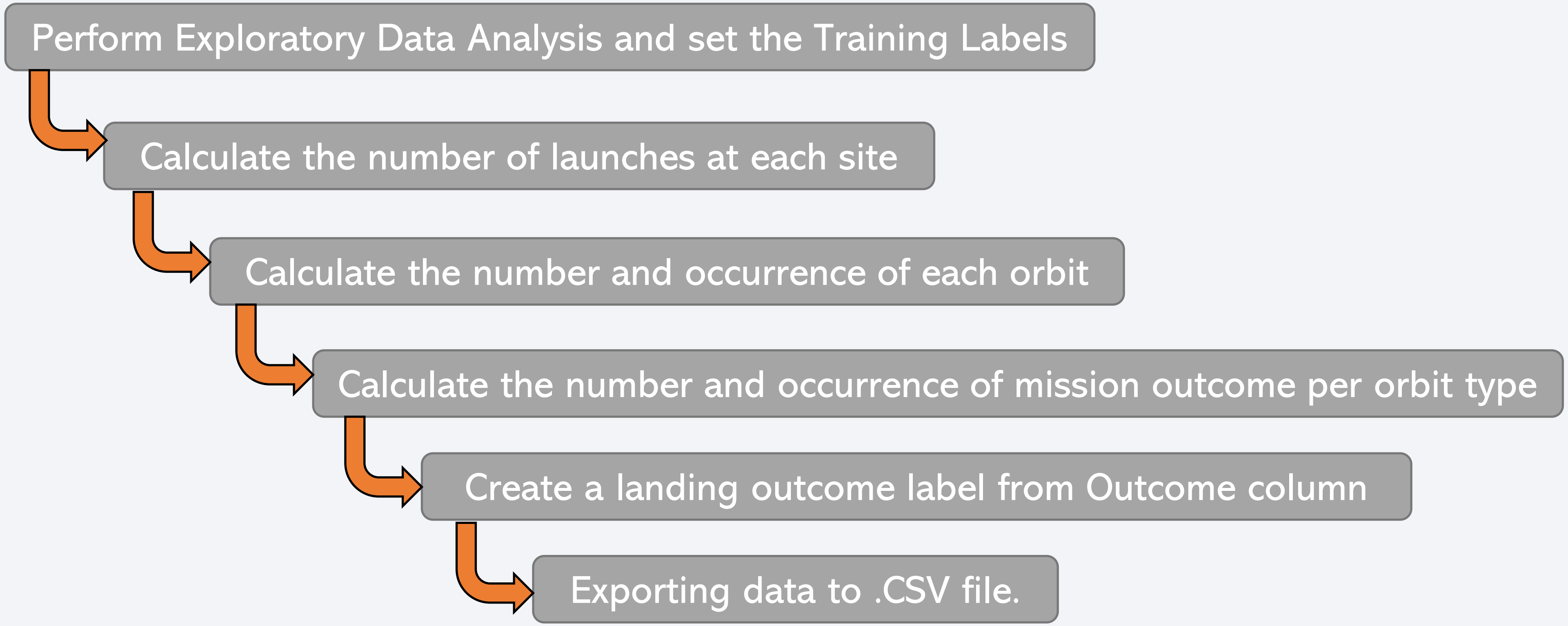
Collecting the data by parsing HTML tables

Constructing data into a dictionary and creating a dataframe from the dictionary

Exporting data to .CSV file.

Data Wrangling

- GitHub URL: https://github.com/tpjsolomon/IBM-data-science-capstone/blob/main/3-Data_wrangling.ipynb
- Because the booster either does or does not land successfully, the outcomes were converted into *Training Labels* with “1” meaning successful landing and “0” meaning unsuccessful.



Perform Exploratory Data Analysis and set the Training Labels

Calculate the number of launches at each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting data to .CSV file.

EDA with Data Visualization

- GitHub URL: https://github.com/tpjsolomon/IBM-data-science-capstone/blob/main/5-EDA_dataviz.ipynb
- The following plots were built: *Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type*, and *Success Rate Yearly Trend*.
- Scatter plots help show the relationship between variables. Bar charts help compare discrete categories. Line charts show trends in data over time.

EDA with SQL

- GitHub URL: https://github.com/tpjsolomon/IBM-data-science-capstone/blob/main/4-EDA_SQL.ipynb

The following SQL queries were made:

- Displaying the names of the unique launch sites in the space mission; the 5 records where launch sites begin with the string ‘CCA’; the total payload mass carried by boosters launched by NASA (CRS); and the average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved; the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000; the total number of successful and failure mission outcomes; the names of the booster versions which have carried the maximum payload mass; the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium

- GitHub URL: https://github.com/tpjsolomon/IBM-data-science-capstone/blob/main/6-Dataviz_with_Folium.ipynb

Added Markers to all Launch Sites:

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured markers (success = green; failed = red) using *Marker Cluster* to identify high success rates.

Distances between a Launch Site to its proximities:

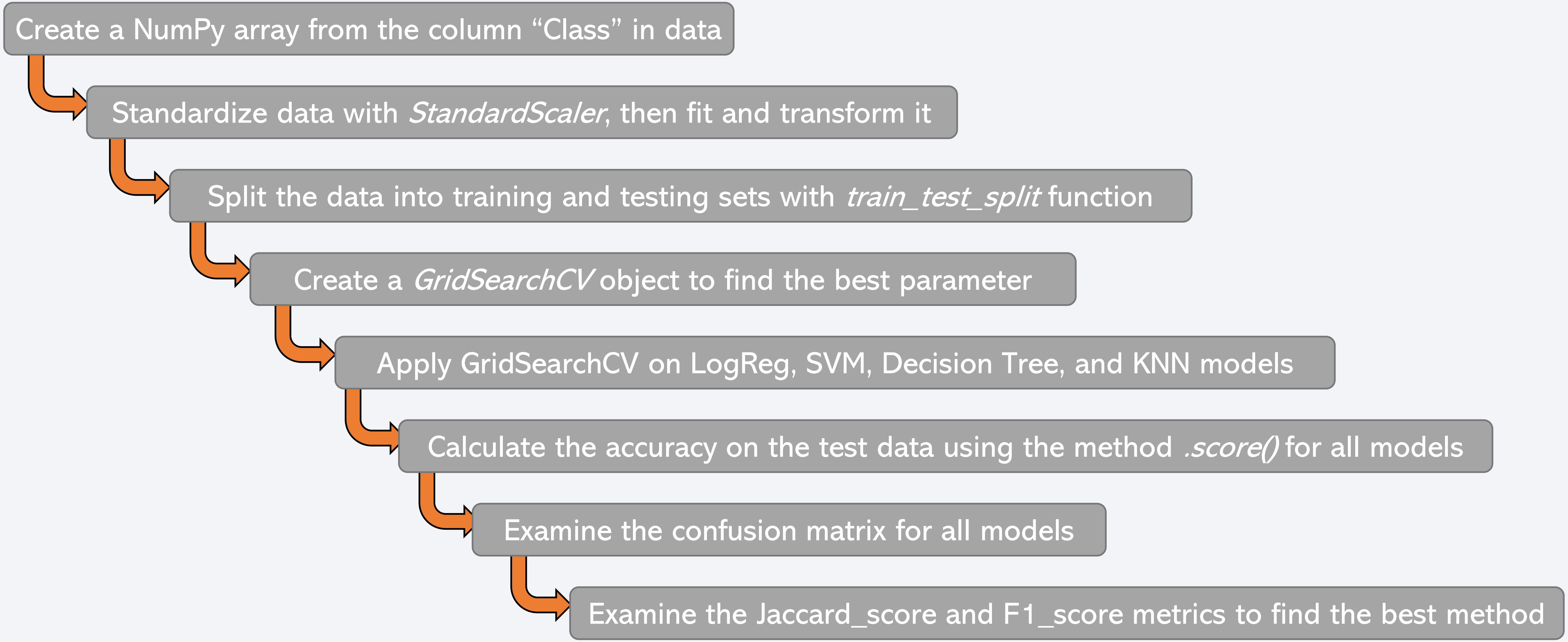
- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

Build a Dashboard with Plotly Dash

- GitHub URL: https://github.com/tpjsolomon/IBM-data-science-capstone/blob/main/6-Dataviz_with_Plotly
- Added a dropdown list to enable Launch Site selection.
- Added a pie chart to show the total successful launches count for all sites and the success/failed counts for the site.
- Added a slider to select Payload range.
- Added a scatter chart to show the correlation between Payload and Launch Success.

Predictive Analysis (Classification)

- GitHub URL: https://github.com/tpjsolomon/IBM-data-science-capstone/blob/main/7-Machine_learning_prediction.ipynb



Create a NumPy array from the column "Class" in data

Standardize data with *StandardScaler*, then fit and transform it

Split the data into training and testing sets with *train_test_split* function

Create a *GridSearchCV* object to find the best parameter

Apply GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculate the accuracy on the test data using the method *.score()* for all models

Examine the confusion matrix for all models

Examine the Jaccard_score and F1_score metrics to find the best method

Results

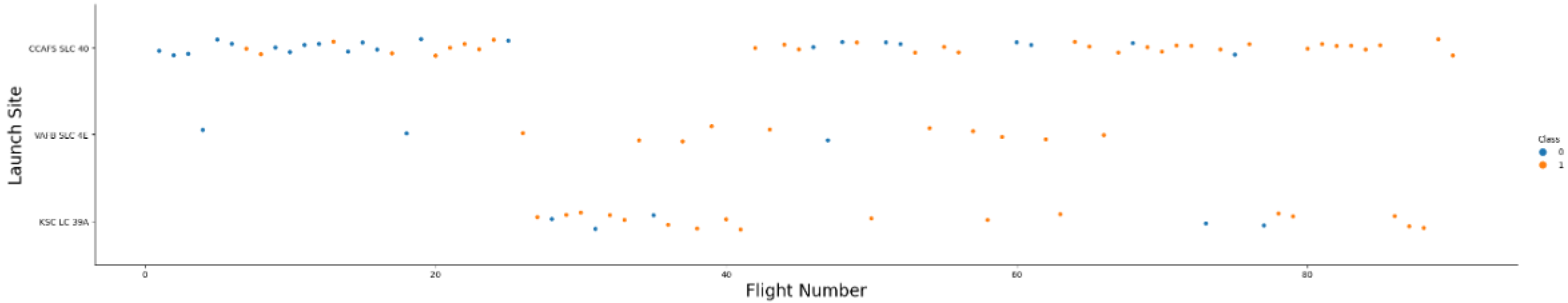
The following pages will present the following:

- Exploratory data analysis (EDA) results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

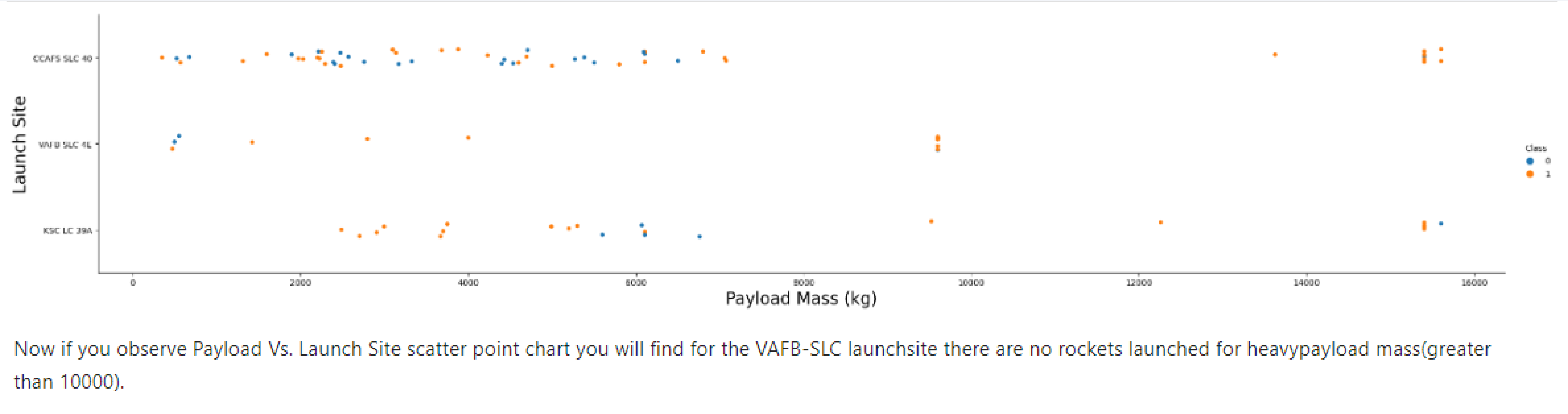
Flight Number vs. Launch Site



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots:

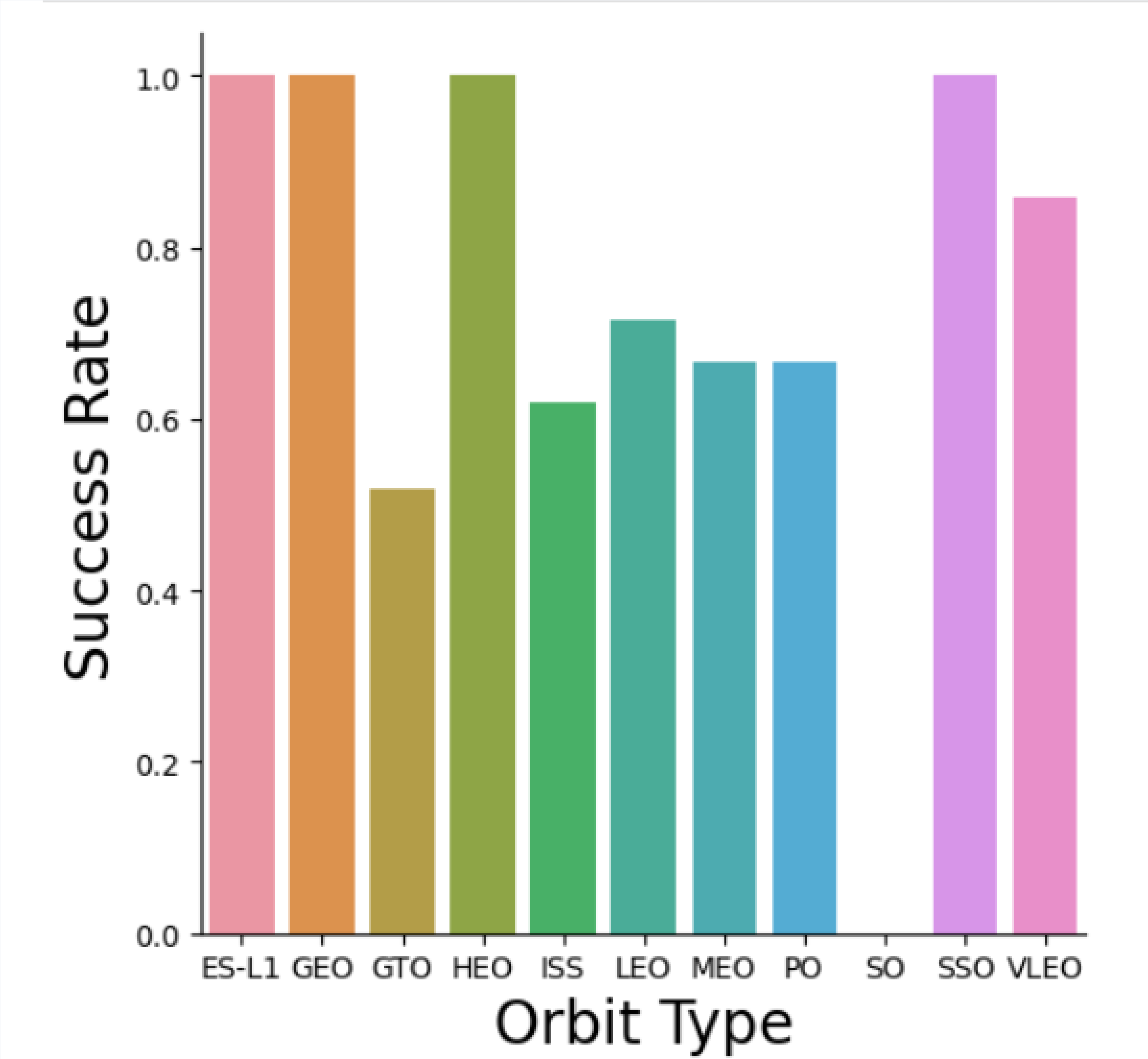
1. Most of the early flights (less than flight number approx. 25) were from launch site CCAFS SLC 40
2. Most of these early flights failed.
3. Lauches from KSC LC-39A were mostly successful.
4. Later flights from site CCAFS SLC 4 were more successful than early flights from this site.

Payload vs. Launch Site



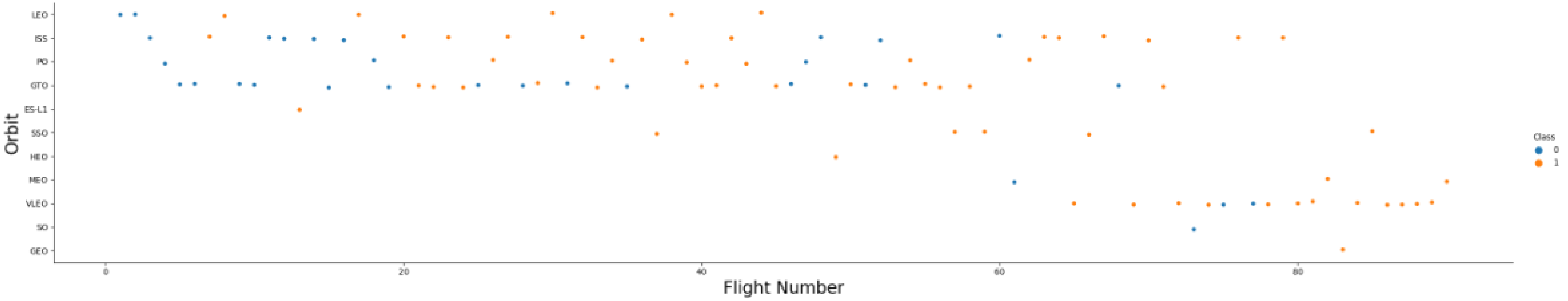
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



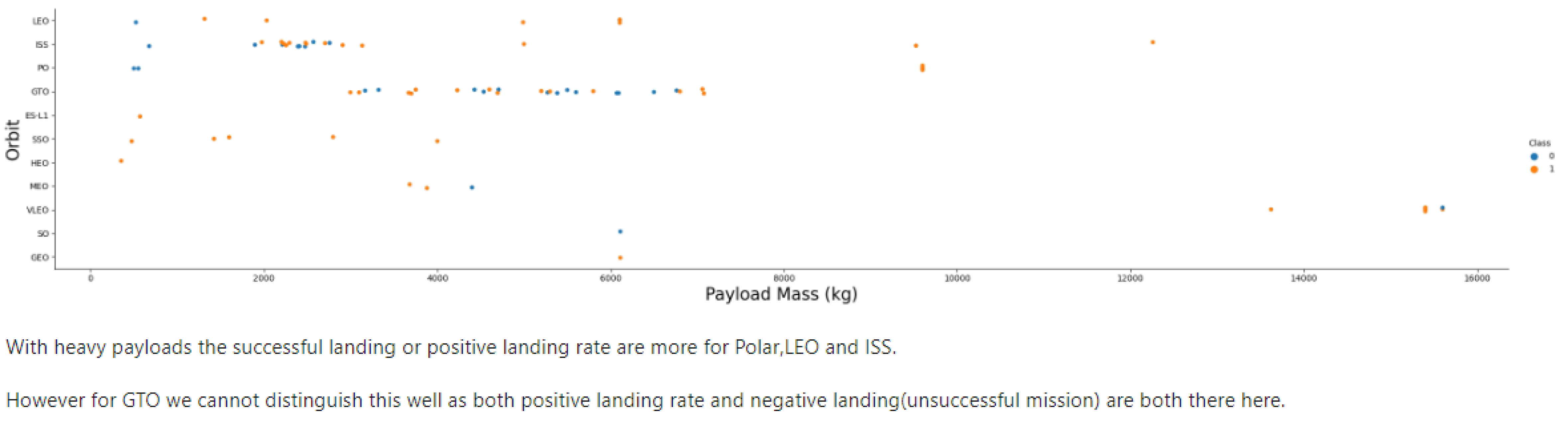
There is 100% success rate during ES-L1, GEO, HEO, and SSO orbits.

Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

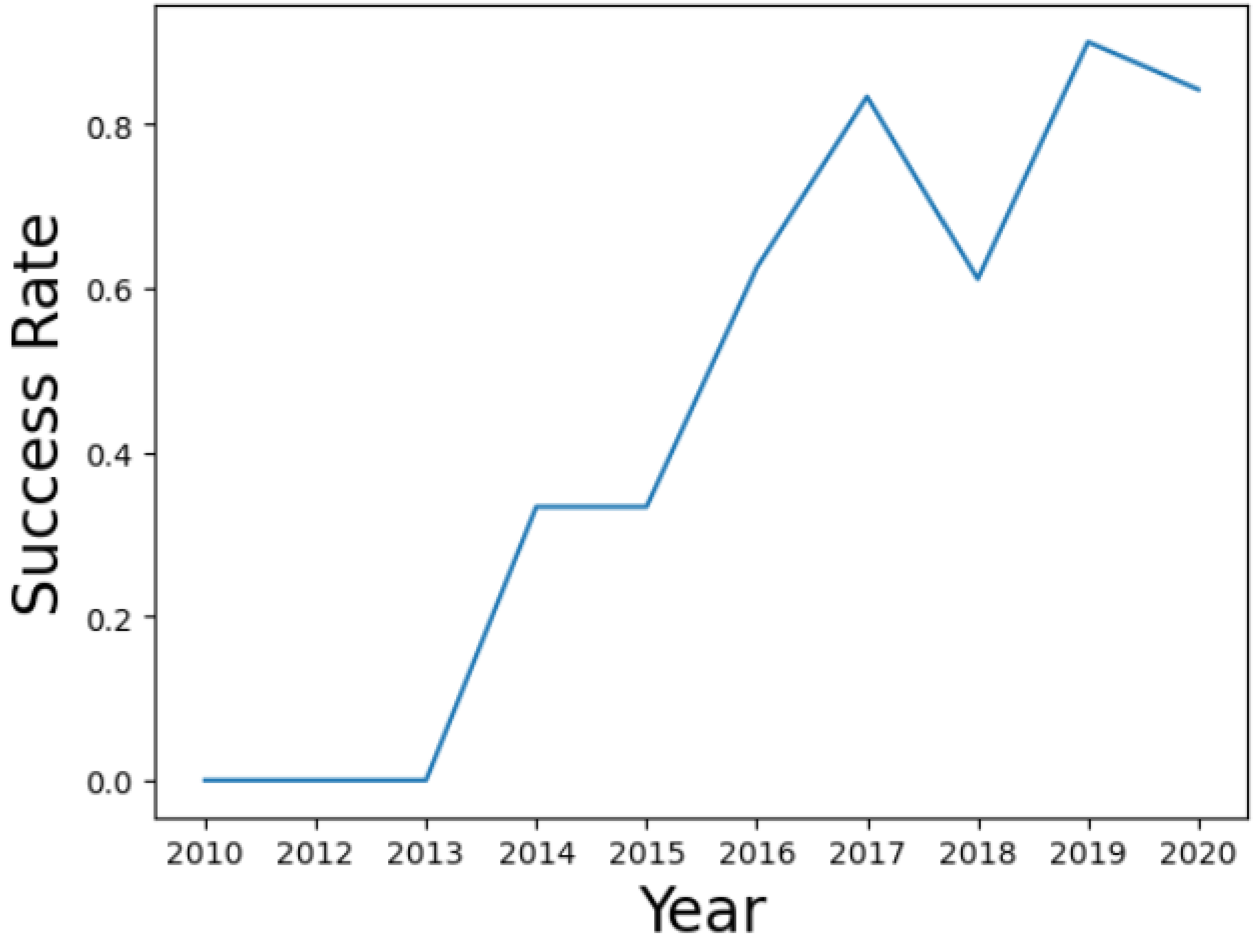
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

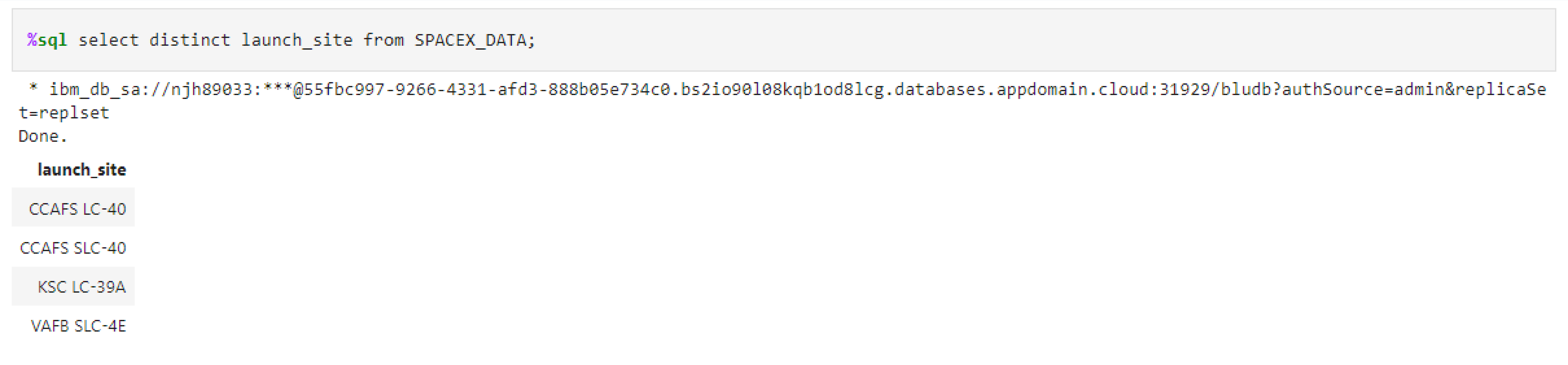
Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

The names of the unique launch sites in the space mission:



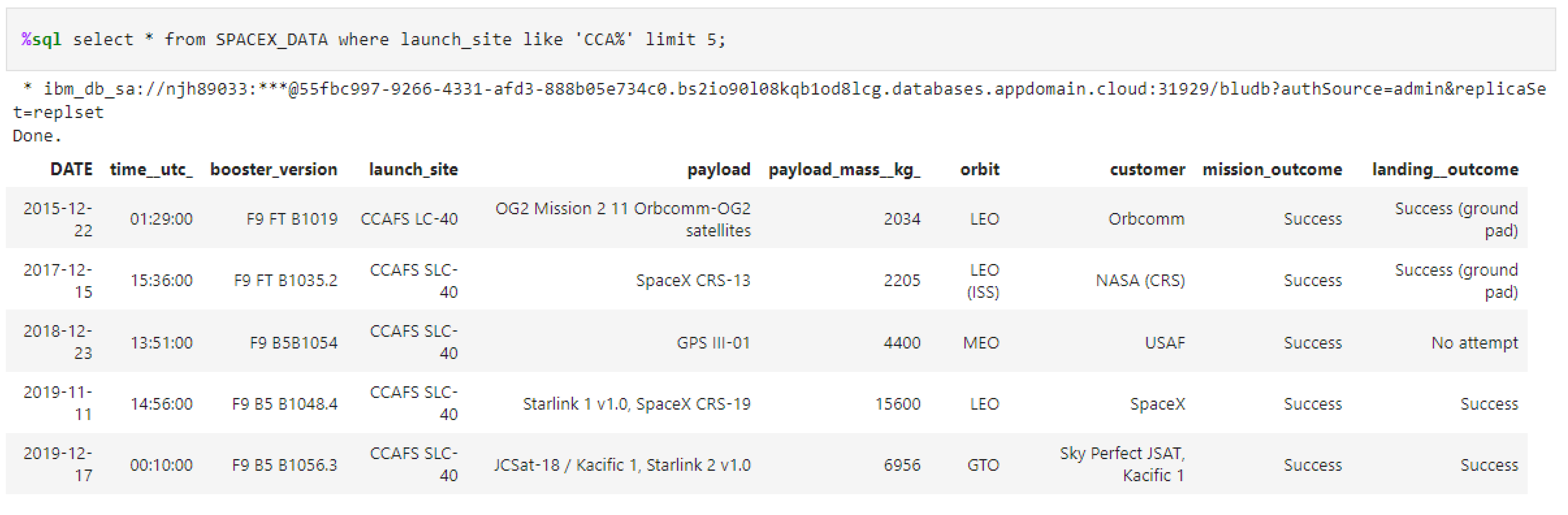
```
%sql select distinct launch_site from SPACEX_DATA;
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

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

Launch Site Names Begin with 'CCA'

Five records where launch sites begin with `CCA`:



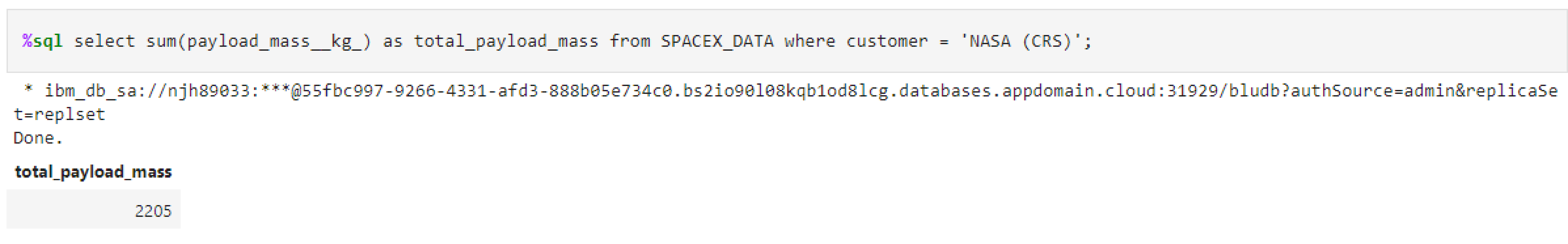
```
%sql select * from SPACEX_DATA where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSet=replset
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
2017-12-15	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2018-12-23	13:51:00	F9 B5B1054	CCAFS SLC-40	GPS III-01	4400	MEO	USAF	Success	No attempt
2019-11-11	14:56:00	F9 B5 B1048.4	CCAFS SLC-40	Starlink 1 v1.0, SpaceX CRS-19	15600	LEO	SpaceX	Success	Success
2019-12-17	00:10:00	F9 B5 B1056.3	CCAFS SLC-40	JCSat-18 / Kacific 1, Starlink 2 v1.0	6956	GTO	Sky Perfect JSAT, Kacific 1	Success	Success

Total Payload Mass

The total payload carried by boosters from NASA:



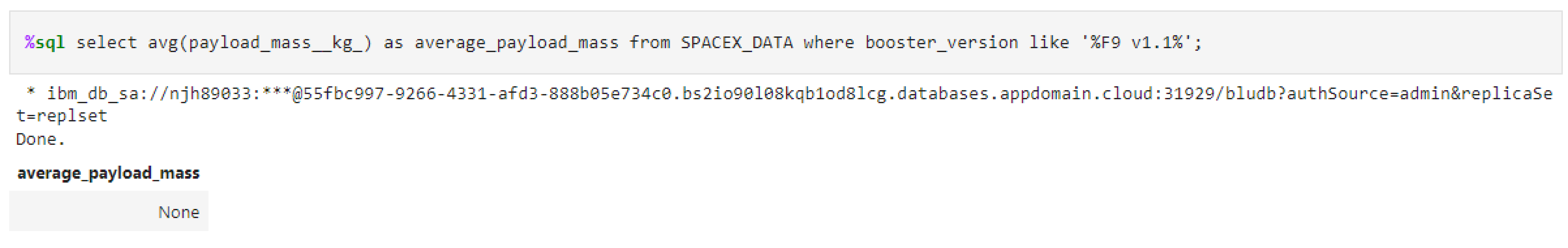
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEX_DATA where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

total_payload_mass
2205

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1:



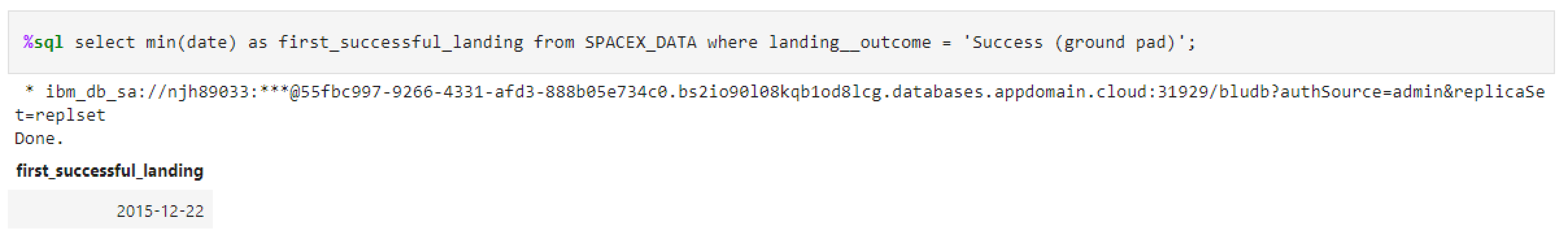
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEX_DATA where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

average_payload_mass
None

First Successful Ground Landing Date

The first successful landing outcome on ground pad:



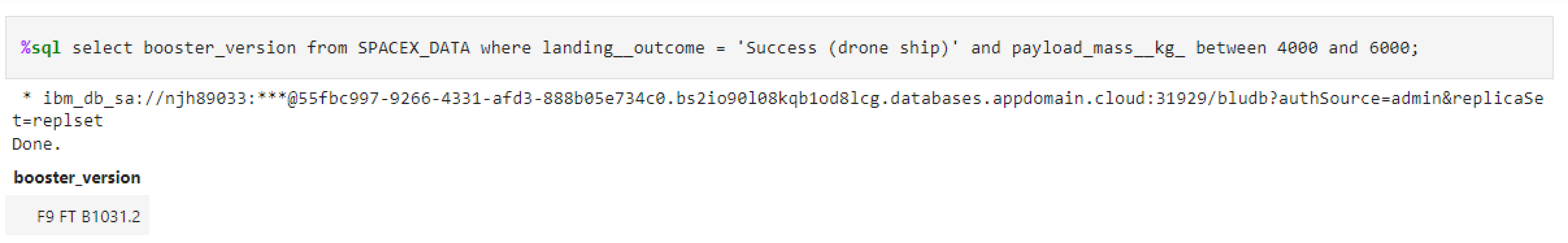
```
%sql select min(date) as first_successful_landing from SPACEX_DATA where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:



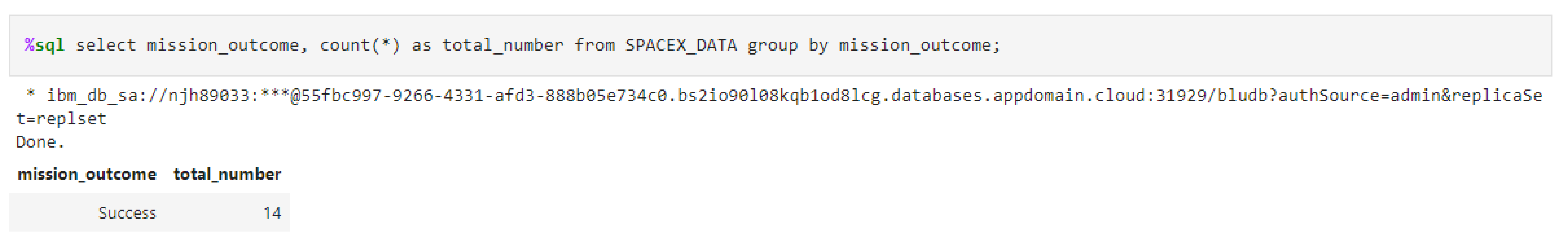
```
%sql select booster_version from SPACEX_DATA where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

booster_version
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes:



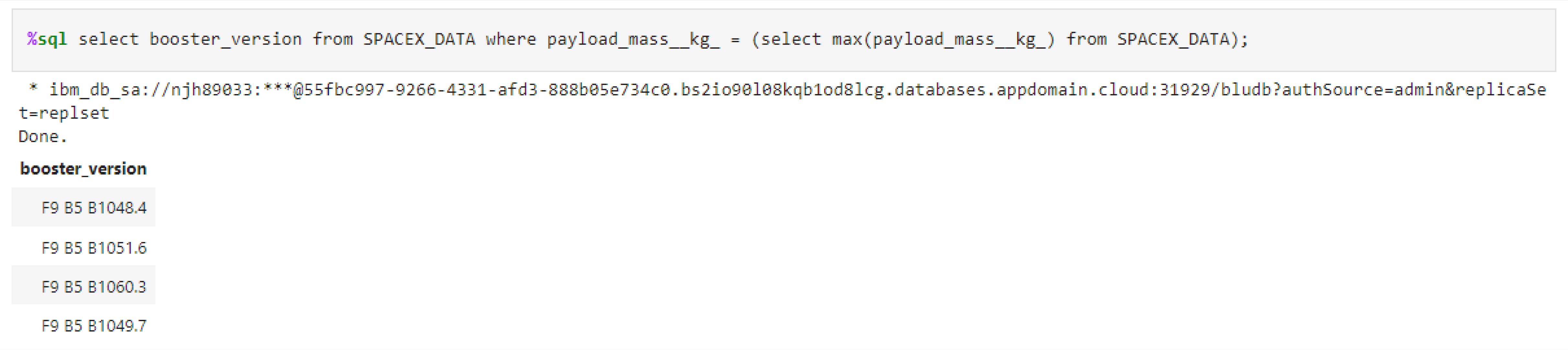
```
%sql select mission_outcome, count(*) as total_number from SPACEX_DATA group by mission_outcome;
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

mission_outcome	total_number
Success	14

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass:



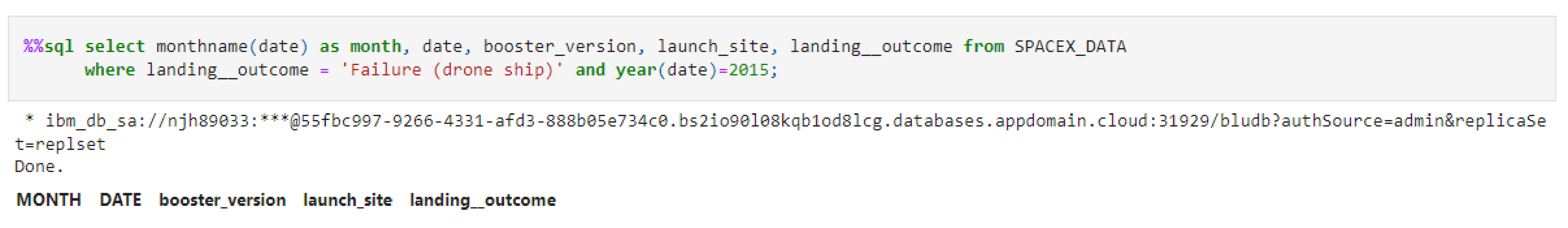
```
%sql select booster_version from SPACEX_DATA where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX_DATA);
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:



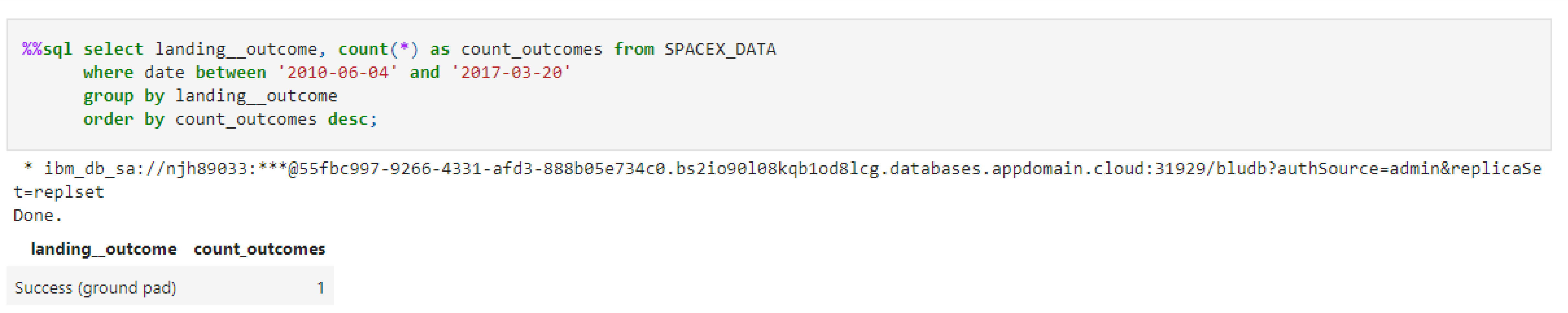
```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEX_DATA
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome

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

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 count_outcomes from SPACEX_DATA
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://njh89033:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb?authSource=admin&replicaSe
t=replset
Done.
```

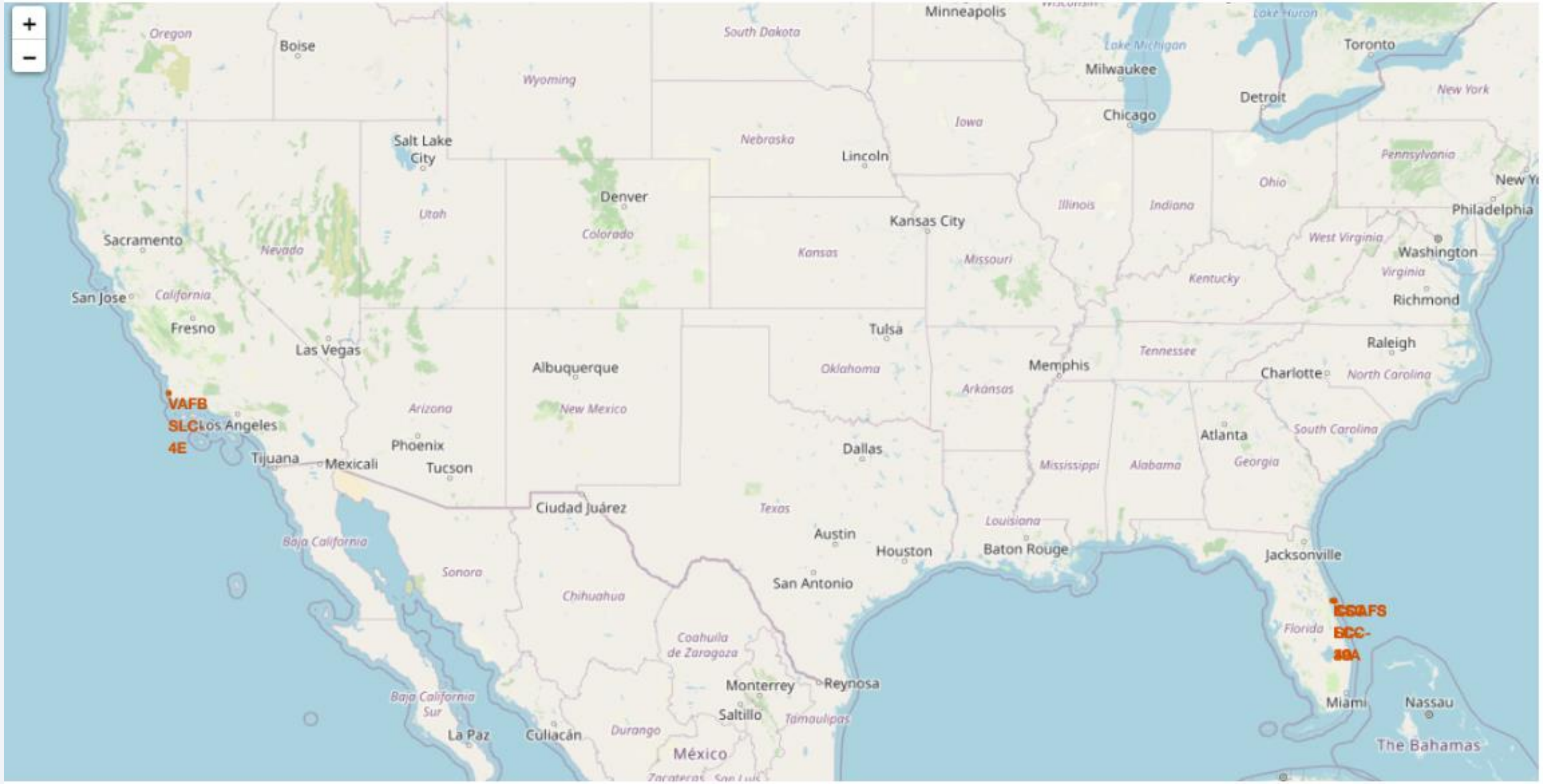
landing__outcome	count_outcomes
Success (ground pad)	1

Section 3

Launch Sites Proximities Analysis

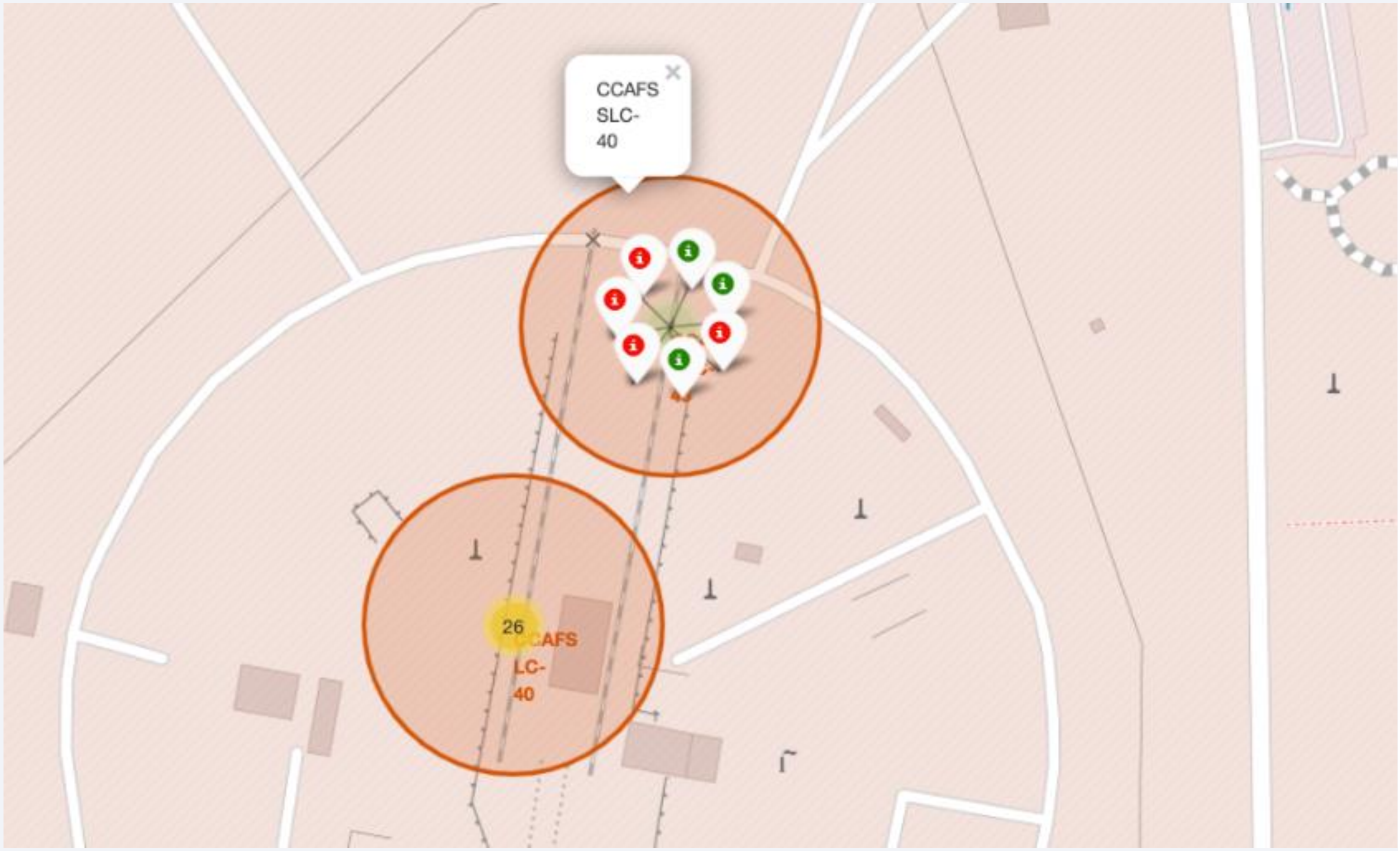
Folium Map 1 - All launch sites' location markers on a global map.

- All launch sites are in proximity to the Equator line and all launch sites are in very close proximity to the coast.



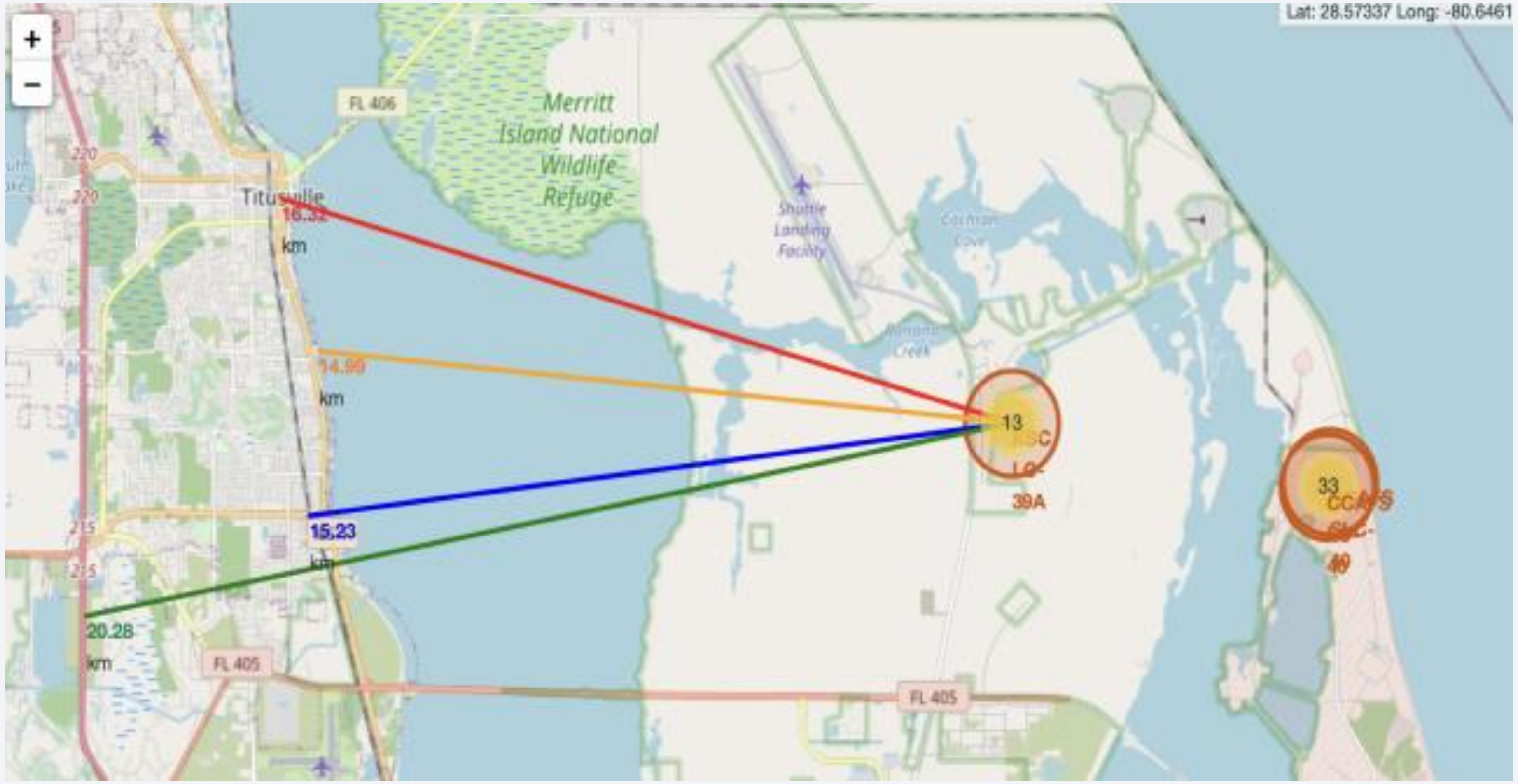
Folium Map 2 - Colour-labeled launch records

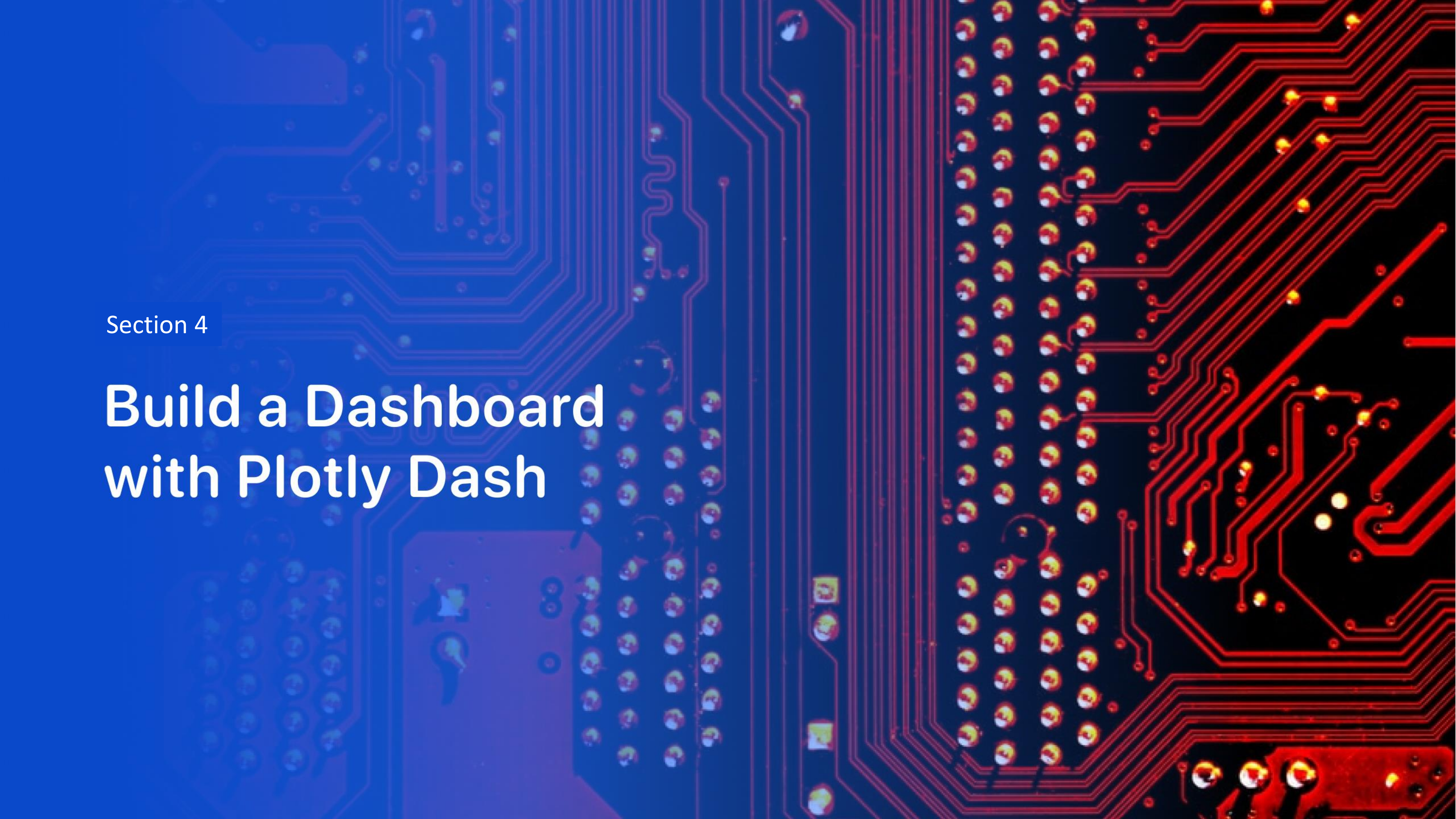
- The colour-labeled markers identify which launch sites have relatively high success rates. Green Marker = Successful Launch. Red Marker = Failed Launch



Folium Map 3 - Distance of launch site KSC LC-39A to its proximities.

The visual analysis shows that launch site KSC LC-39A is relatively close to a railway, highway, and coastline, and the closest city is ~16 km away.



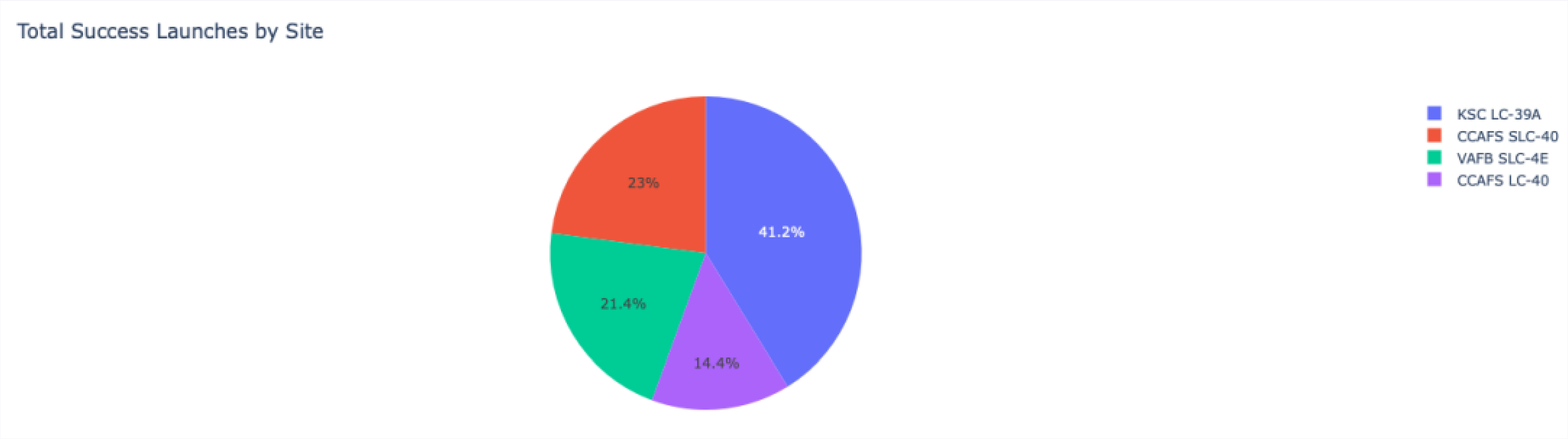


Section 4

Build a Dashboard with Plotly Dash

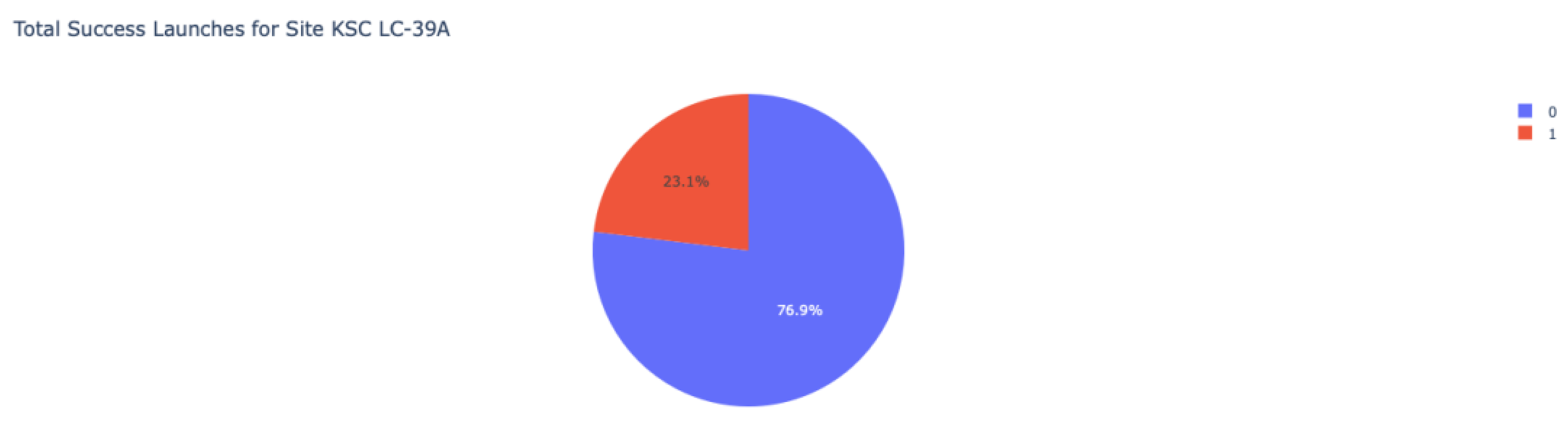
Dashboard 1 - Launch success count for all sites

The chart shows that launch site KSC LC-39A has the most successful launches:



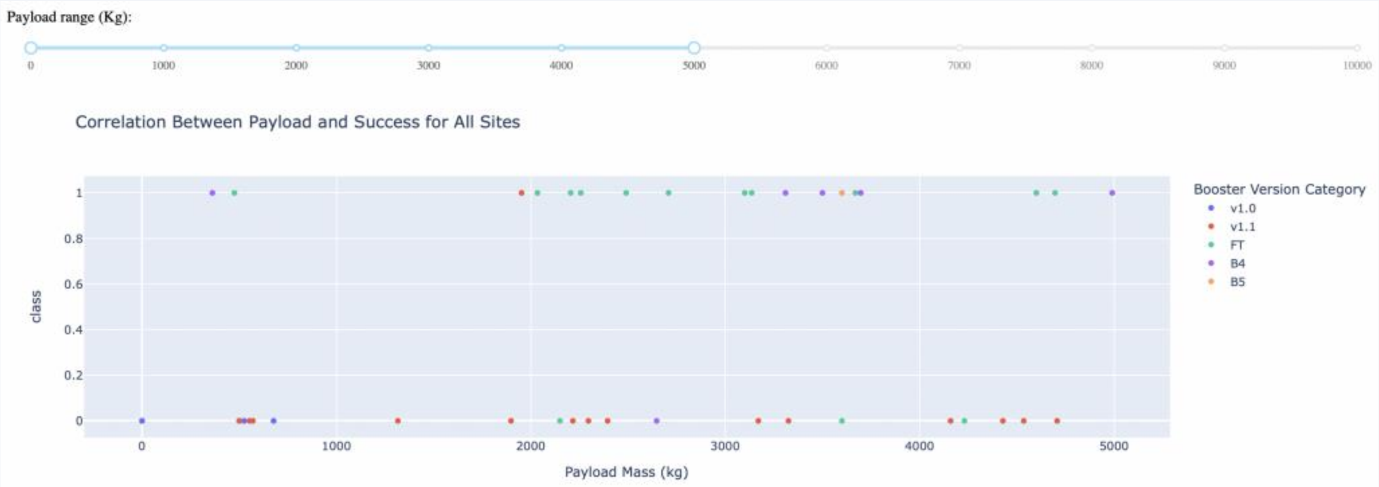
Dashboard 2 - Launch site with highest success

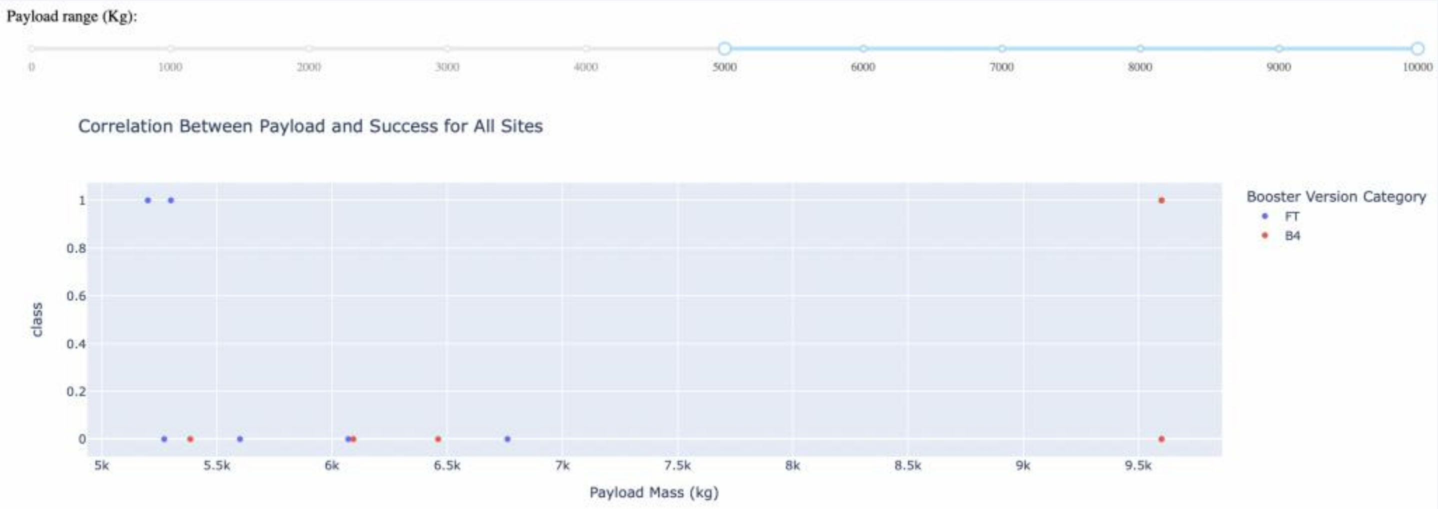
Lunch site KSC LC-39A has the highest success rate (blue, 76.9%):



Dashboard 3 - Payload Mass vs. Launch Outcome

Payloads between 2000 and 5500 kg have the highest success rate:



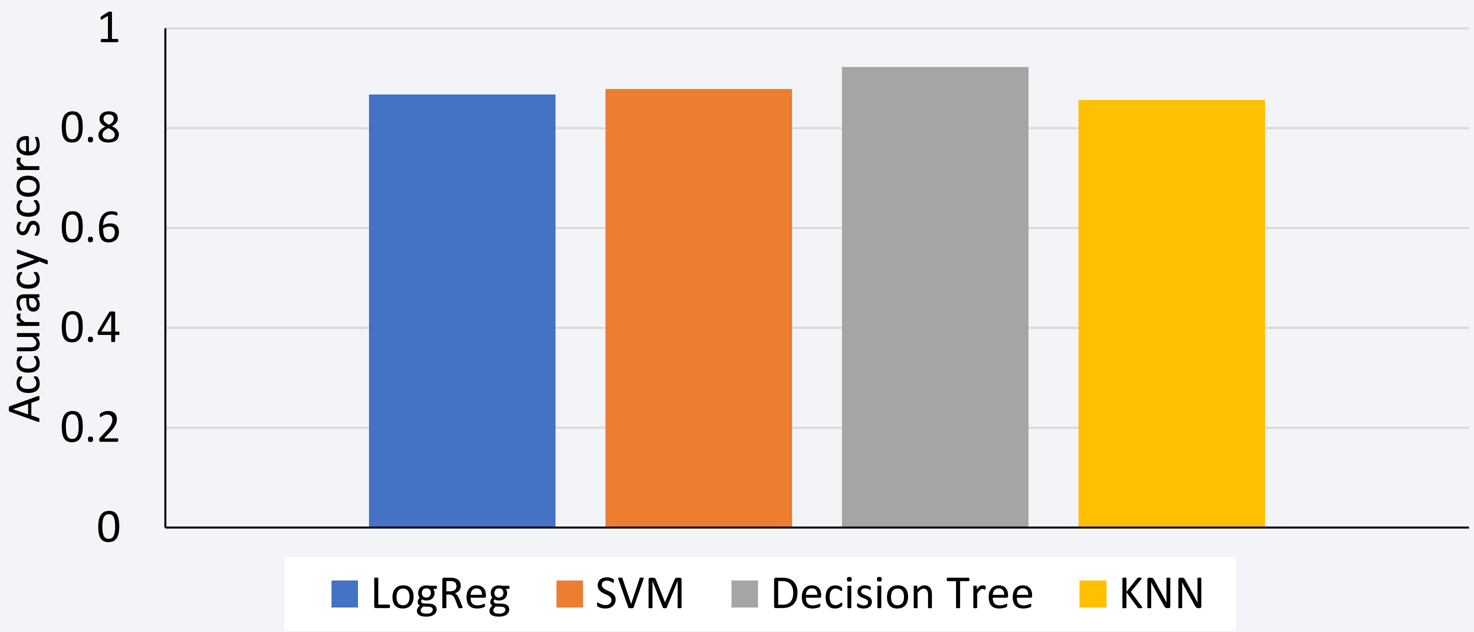


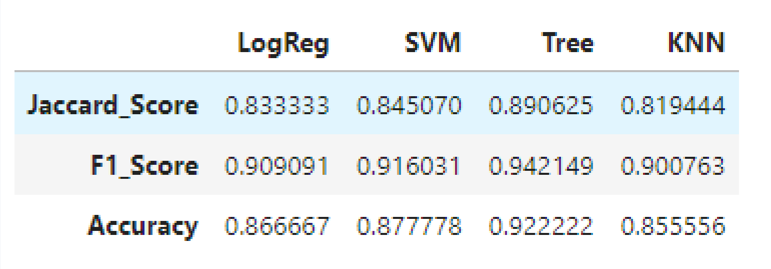
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- A bar chart showing the model accuracy for all classification models:



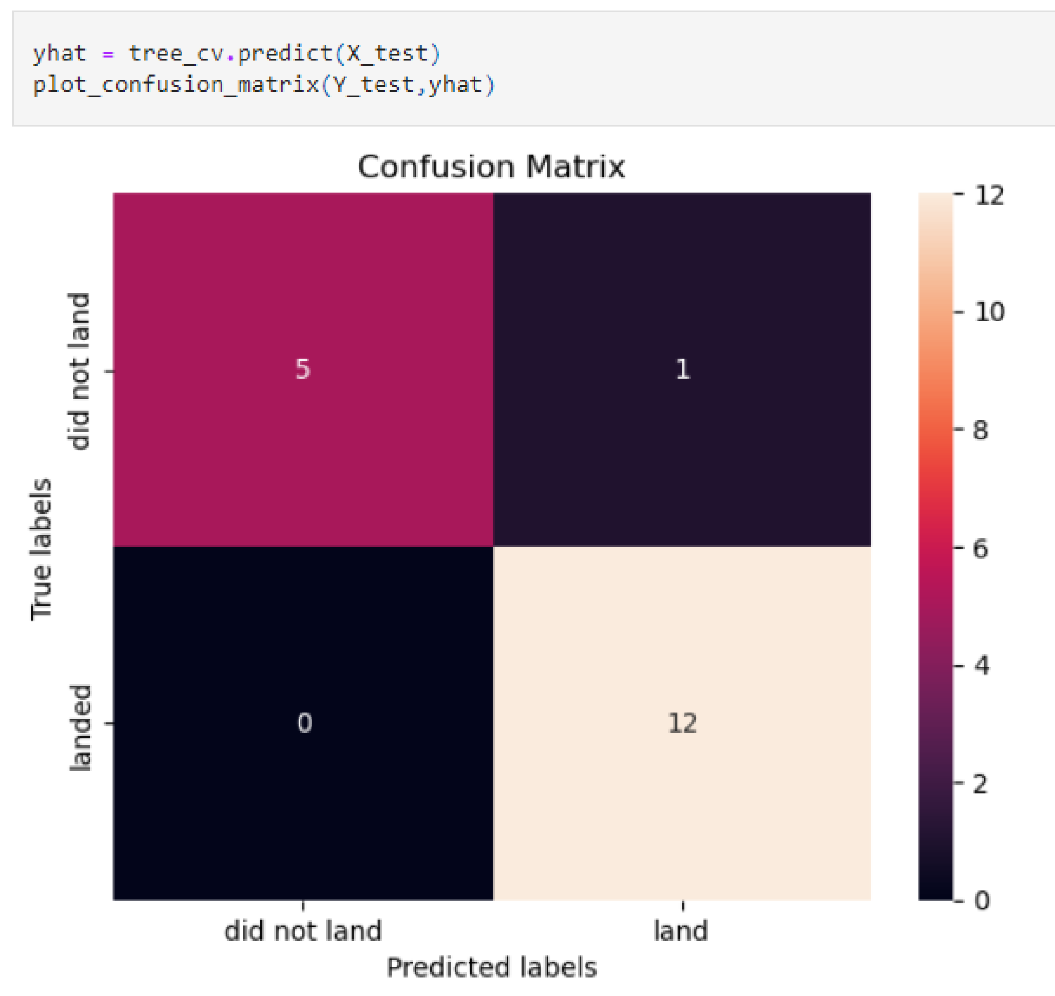


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.890625	0.819444
F1_Score	0.909091	0.916031	0.942149	0.900763
Accuracy	0.866667	0.877778	0.922222	0.855556

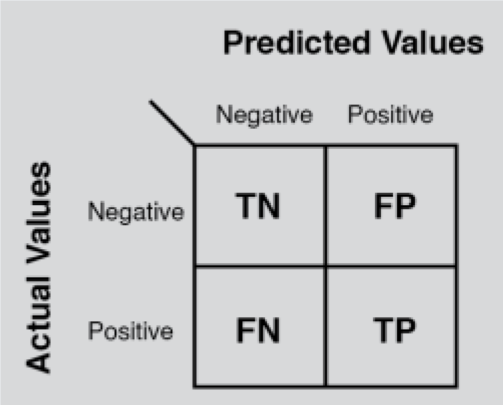
The Decision Tree model has the highest classification accuracy (0.92).

Confusion Matrix

The confusion matrix of the Decision Tree model, the best performing model:



```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP

The confusion matrix shows that the Decision Tree model has a 100% True Positive rate (12/12) and a 0% False Negative rate but a minor problem with False Positives (1/12).

Conclusions

- Launch site KSC LC-39A has the highest success rate.
- Launches with a low payload mass have a greater success rate than launches with a larger payload mass.
- The Decision Tree Model is the best machine learning algorithm for predicting landing success rate for this data set.

Appendix

- All Jupiter Notebooks for this capstone project can be found at: https://github.com/tpjsolomon/IBM-data-science-capstone

Acknowledgements

- I'd like to thank the course instructors, IBM, and Coursera for this professional certificate course.

Thank you!