IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cejana Maciel
12-14-2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an Interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars each. Much of the savings are because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- Problems you want to find answers
    - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
    - Does the rate of successful landings increase over the years?
    - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data for 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

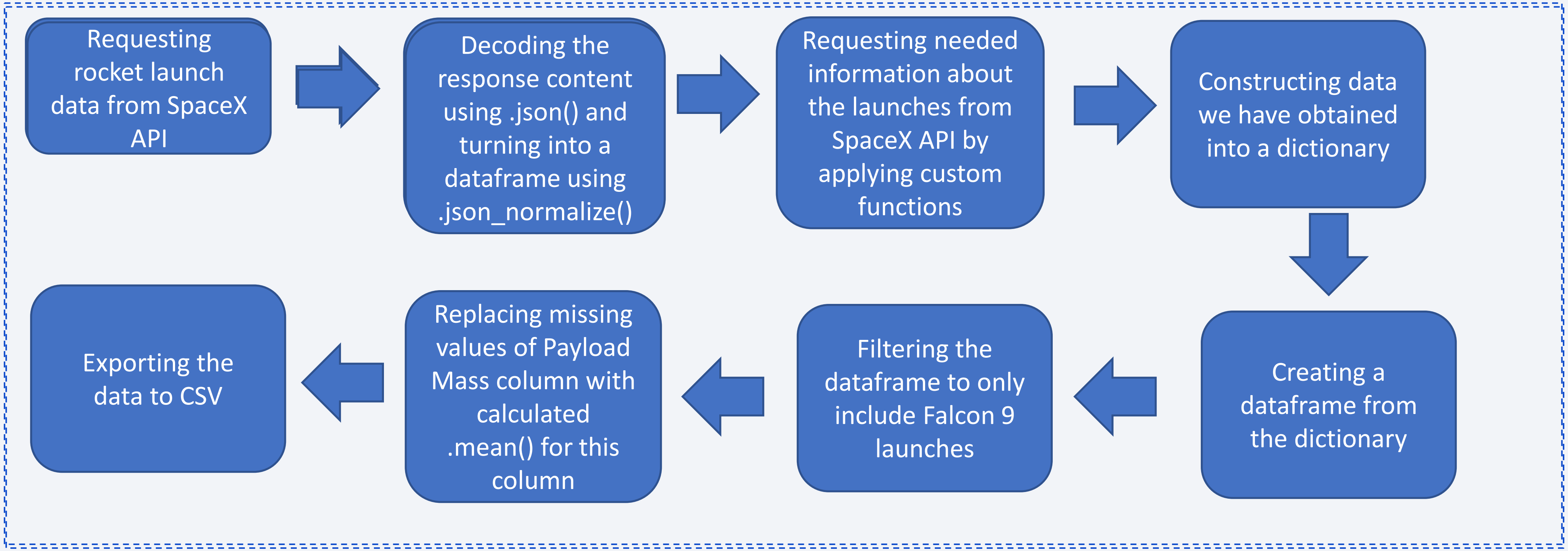
Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX`s Wikipedia entry.

We had to use both data collection methods in order to get complete information about the launches for a more detailed analysis.

- Data Columns are obtained by using SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Collection – SpaceX API



Requesting rocket launch data from SpaceX API

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

Requesting needed information about the launches from SpaceX API by applying custom functions

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

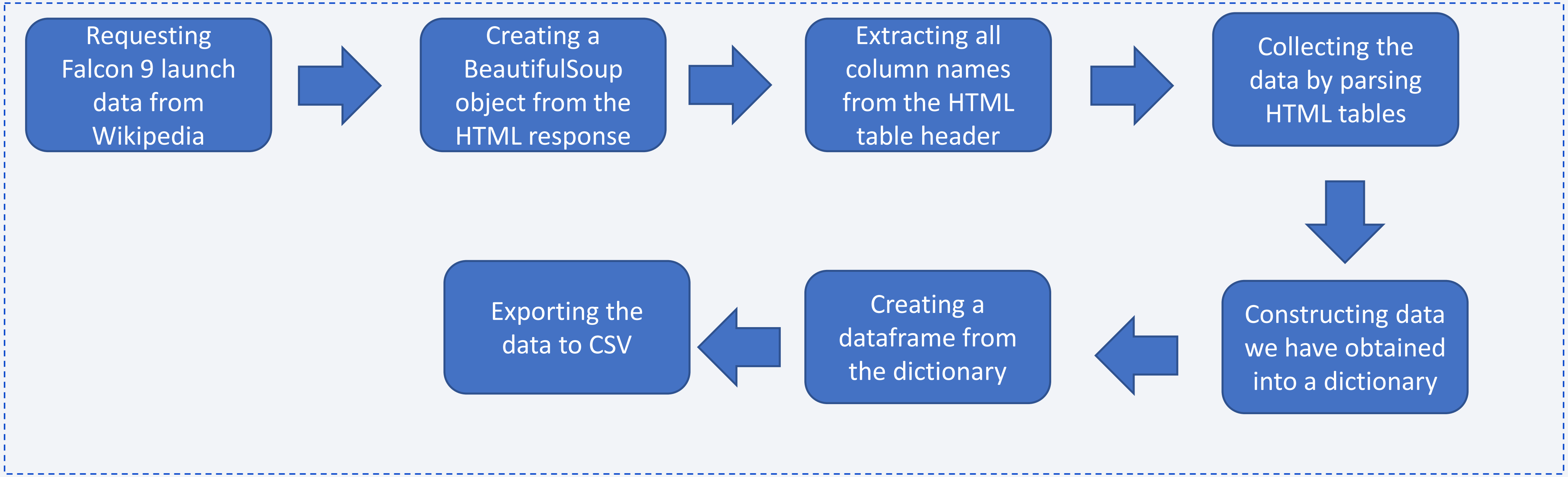
Filtering the dataframe to only include Falcon 9 launches

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

Exporting the data to CSV

Data Collection API

Data Collection - Scraping



Requesting Falcon 9 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 we have obtained into a dictionary

Creating a dataframe from the dictionary

Exporting the data to CSV

Data Collection with Web Scraping

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed in a specific region of the ocean, while False Ocean means the mission outcome was unsuccessfully landed in a specific region of the ocean. True RTLS means the mission outcome was successfully landed on a ground pad, and False RTLS means the mission outcome was unsuccessfully landed on a ground pad. True ASDS means the mission outcome was unsuccessful, landing on a drone ship.
- We mainly convert those outcomes into Training Labels with "1", which means the booster successfully landed, and "0" means it was unsuccessful

- Perform exploratory Data Analysis and determine Training Labels
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcomes per orbit type
  - Create a landing outcome label from the Outcome column
  - Exporting the data to CSV

Data Wrangling

EDA with Data Visualization

- Charts were plotted:
  - 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 show the relationship between variables. If a relationship exists, they could be used in the machine learning model.
  - Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
  - Line charts show trends in data over time (time series)

[EDA with Data Visualization]

EDA with SQL

- Performed SQL Queries:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string “CCA”
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying the average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in the ground pad was achieved
  - Listing the names of the boosters which have success in drone ships and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failed mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ships, their booster versions, and launch site names for the months in the year 2015
  - Ranking 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

EDA with SQL

Build an Interactive Map with Folium

- Markers of all Launch Sites:
  - Added marker with Circle, Popup Label, and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circles, Popup labels, and Text labels of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator coasts.

- Colored Markers of the launch outcomes for each Launch Site:
  - Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

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

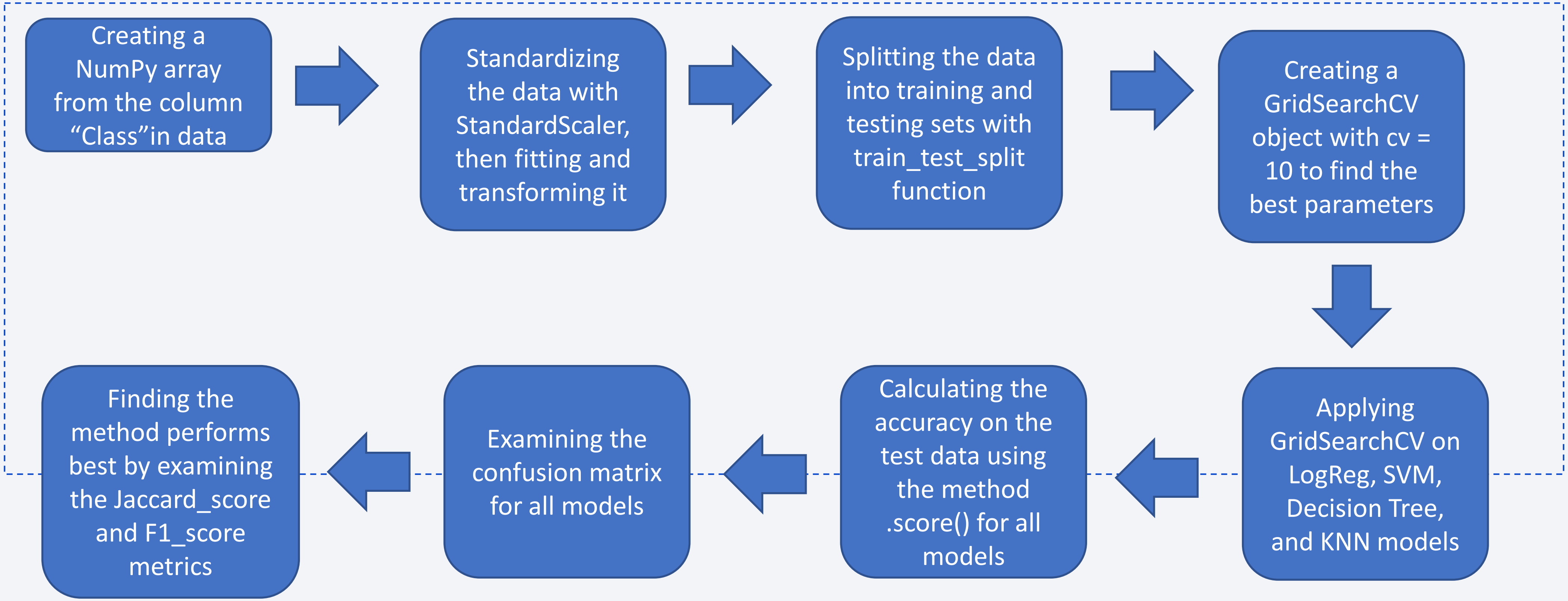
Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

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

SpaceX Dash App

Predictive Analysis (Classification)



Creating a NumPy array from the column “Class”in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

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

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

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

Machine Learning Prediction

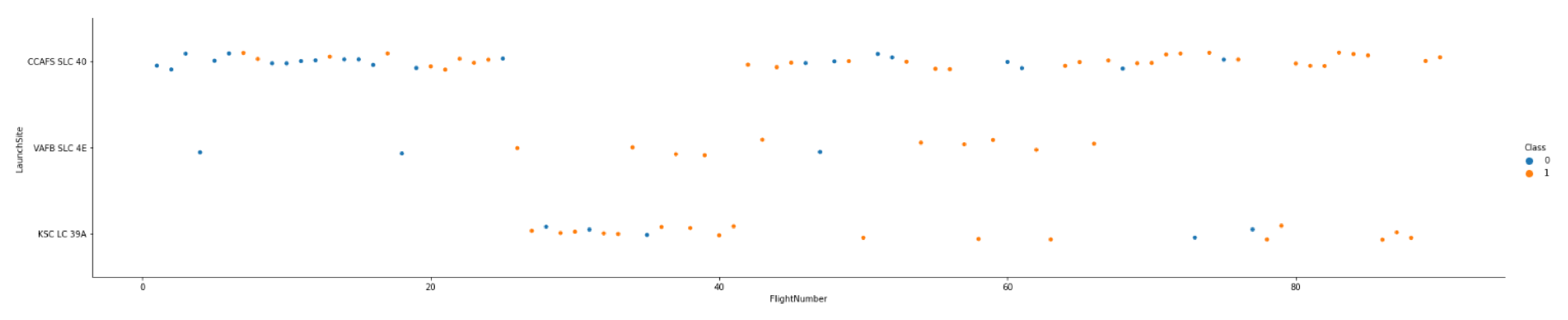
Results

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

Section 2

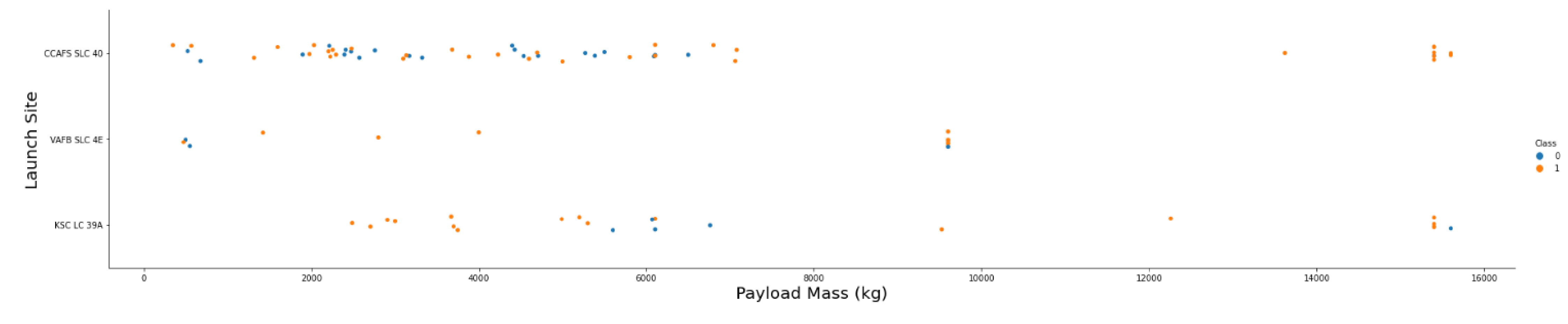
Insights drawn from EDA

Flight Number vs. Launch Site



- Explanation:
  - The earliest flights all failed, while the latest flights all succeeded
  - The CCAFS SLC 40 launch site has about half of all launches
  - VAFB SLC 4E and KSC LC 39A have higher success rates
  - It can be assumed that each new launch has a higher rate of success

Payload vs. Launch Site

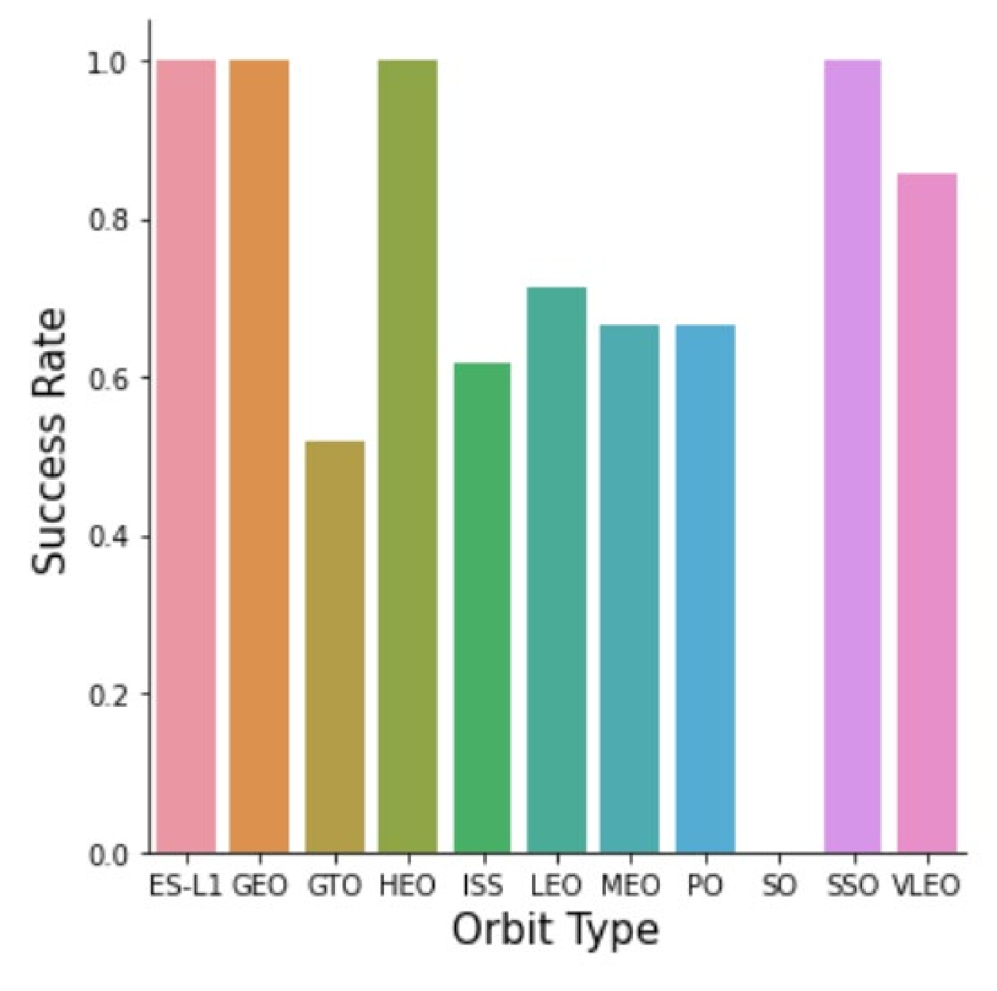


- Explanation:
  - For every launch site, the higher the payload mass, the higher the success rate
  - Most of the launches with payload mass over 7000 kg were successful
  - KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

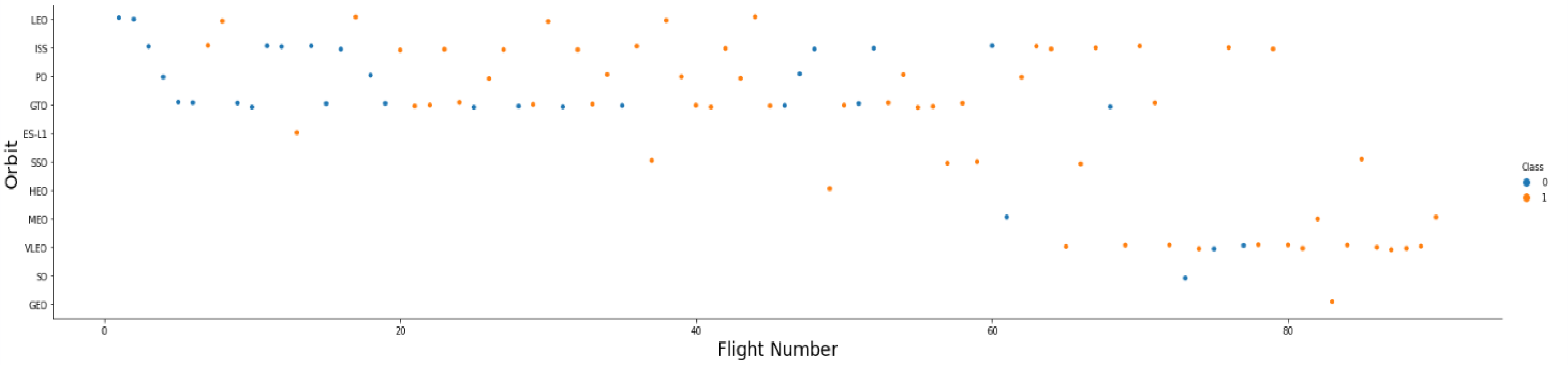
Success Rate vs. Orbit Type

Explanation :

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



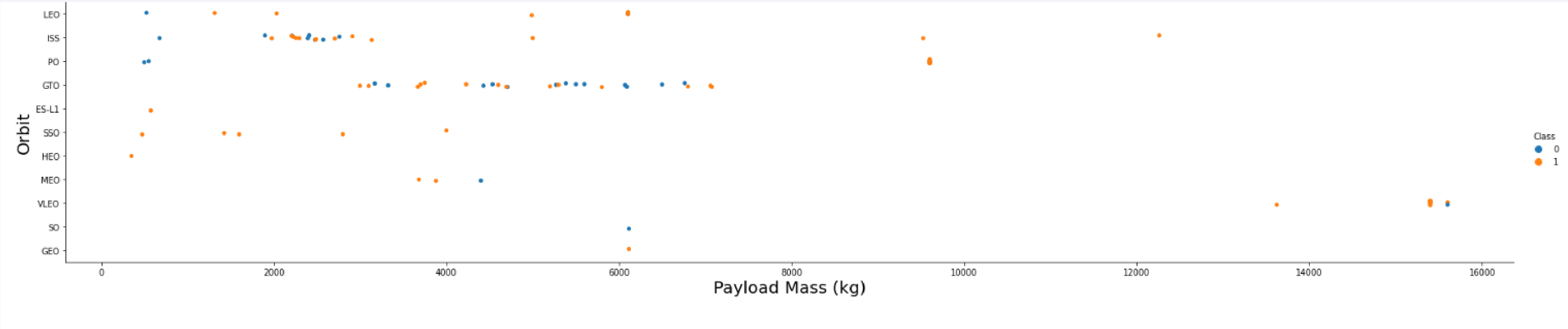
Flight Number vs. Orbit Type



Explanation:

- In the LEO orbit 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



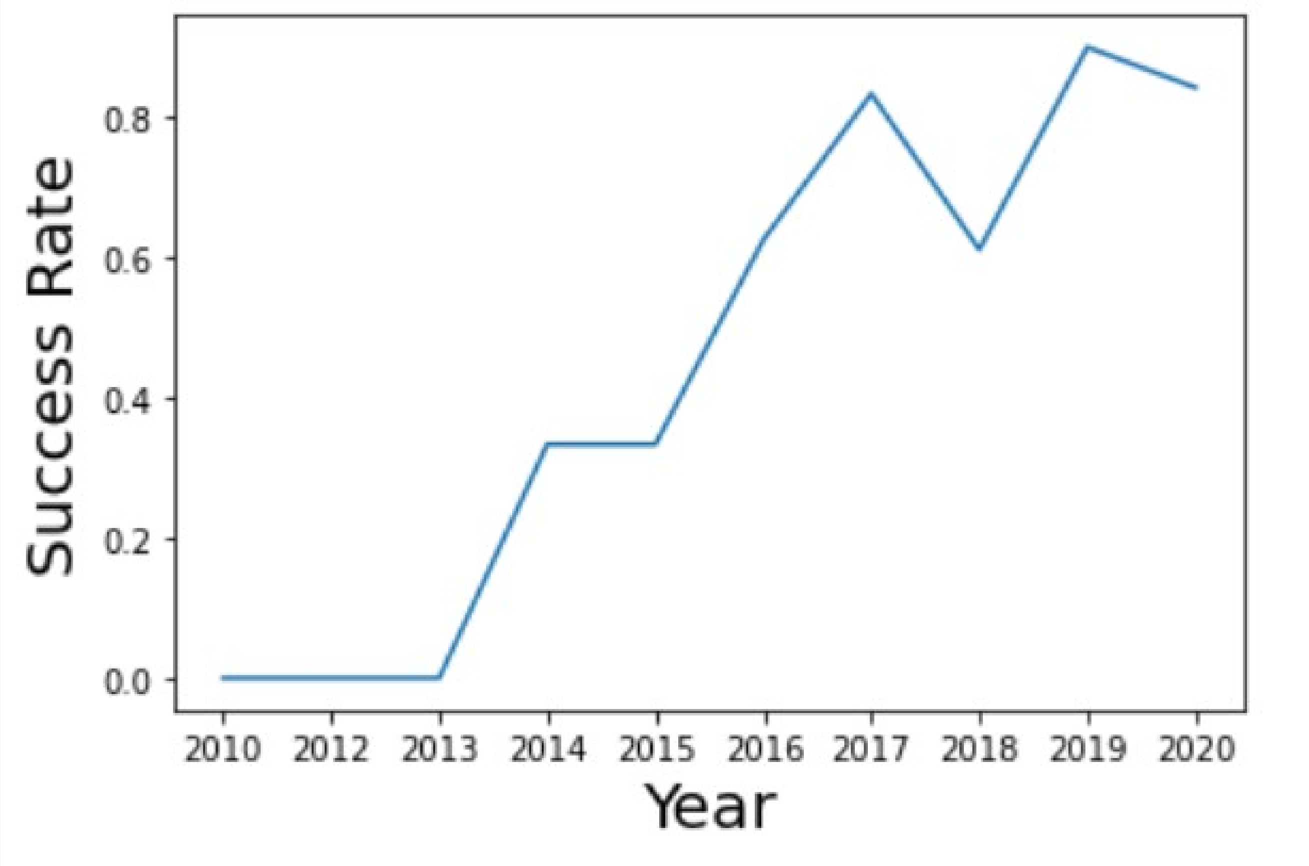
Explanation:

- Heavy payloads have a negative influence on GTO orbits and a positive on GTO and Polar LEO (ISS) orbits

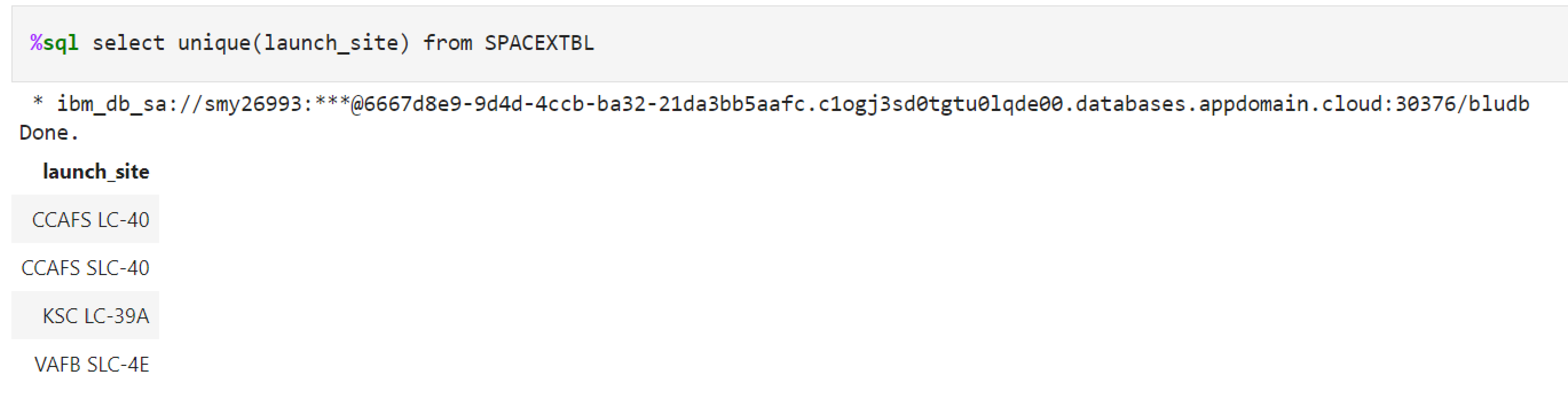
Launch Success Yearly Trend

Explanation:

- The success rate since 2013 kept increasing until 2020



All Launch Site Names



```
%sql select unique(launch_site) from SPACEXTBL
```

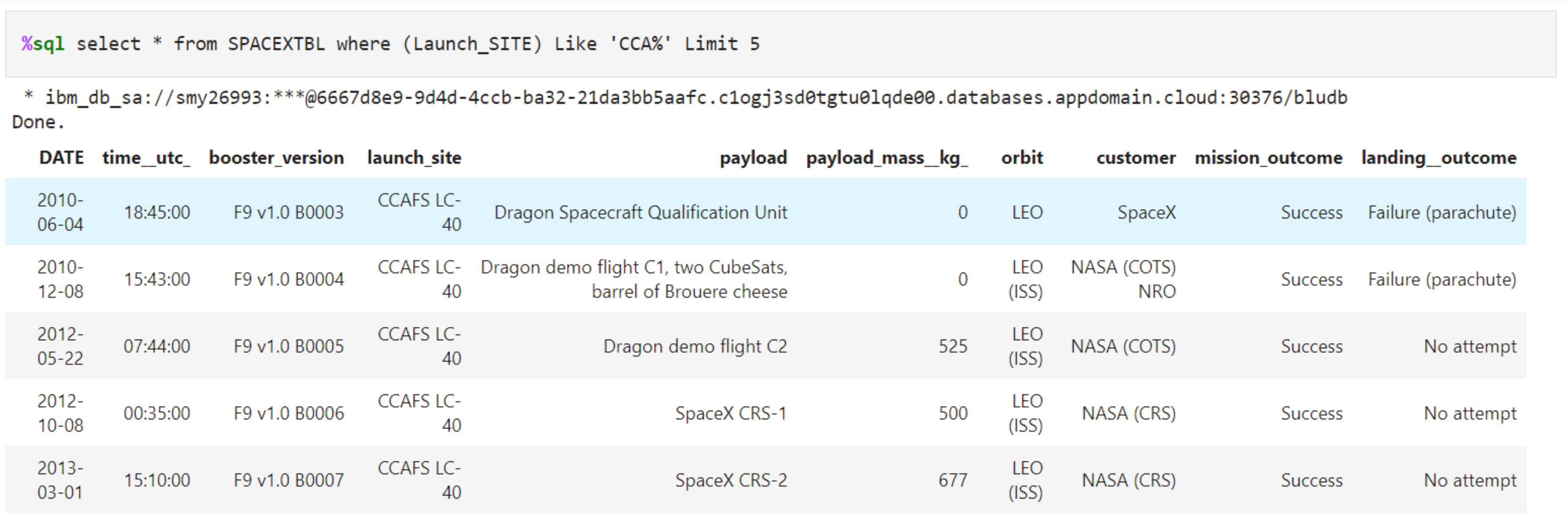
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

Explanation:

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

Launch Site Names Begin with 'CCA'



```
%sql select * from SPACEXTBL where (Launch_SITE) Like 'CCA%' Limit 5
```

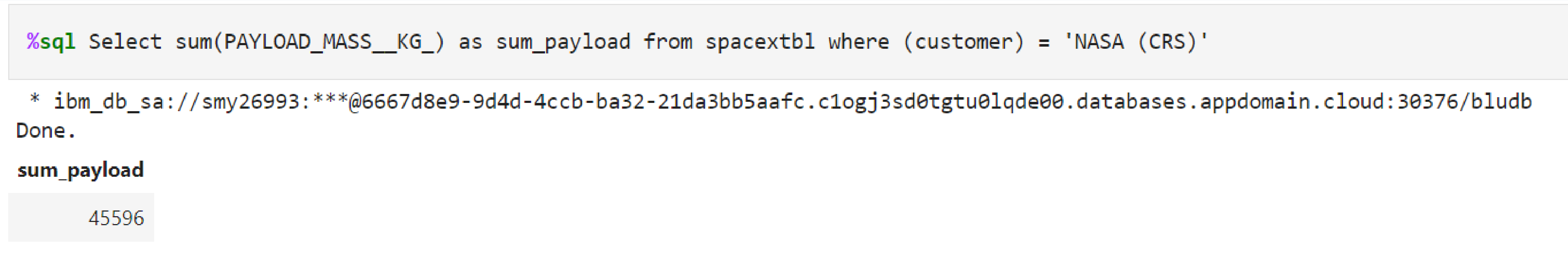
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation:

- Displaying 5 records where launch sites begin with the string “CCA”

Total Payload Mass



```
%sql Select sum(PAYLOAD_MASS__KG_) as sum_payload from spacextbl where (customer) = 'NASA (CRS)'
```

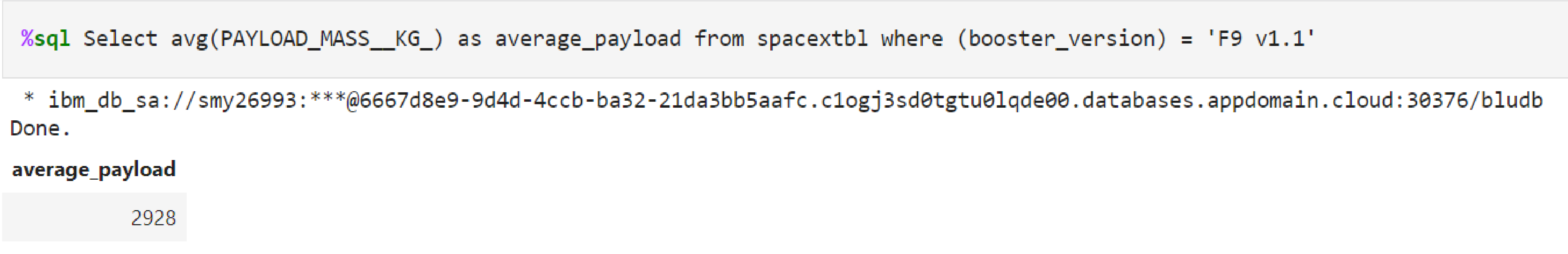
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

sum_payload
45596

Explanation:

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1



```
%sql Select avg(PAYLOAD_MASS__KG_) as average_payload from spacextbl where (booster_version) = 'F9 v1.1'
```

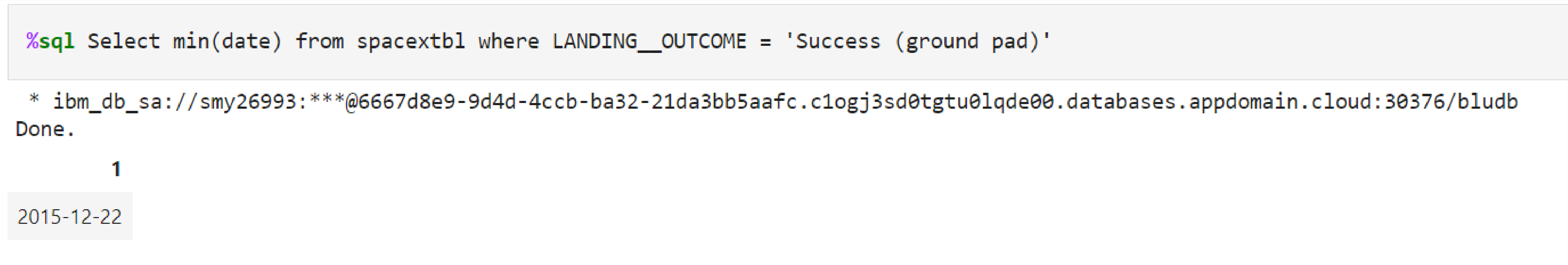
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

average_payload
2928

Explanation:

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

First Successful Ground Landing Date



```
%sql Select min(date) from spacextbl where LANDING__OUTCOME = 'Success (ground pad)'
```

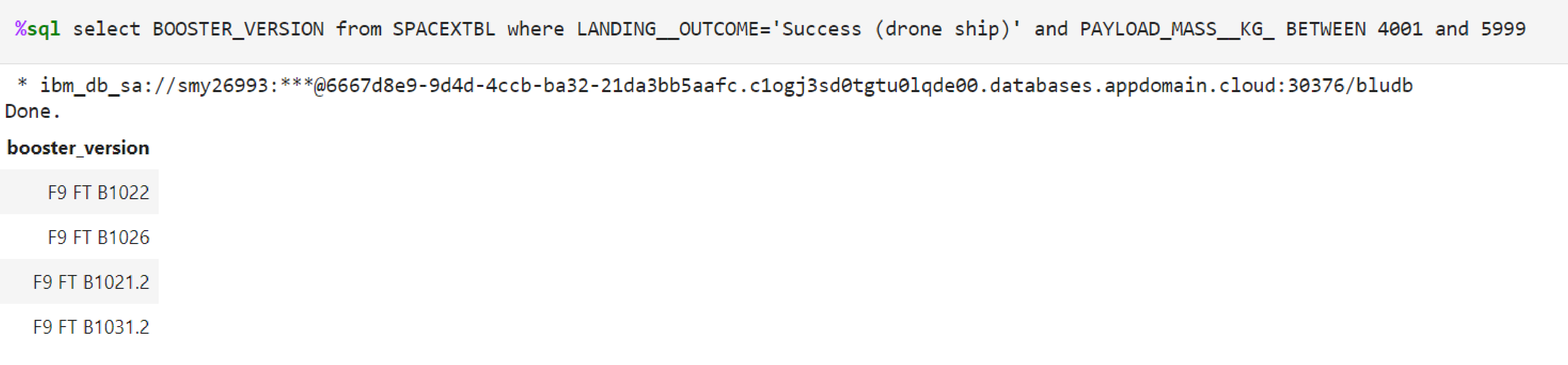
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

1
2015-12-22

Explanation:

- Listing the date when the first successful landing outcome in the round pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000



```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999
```

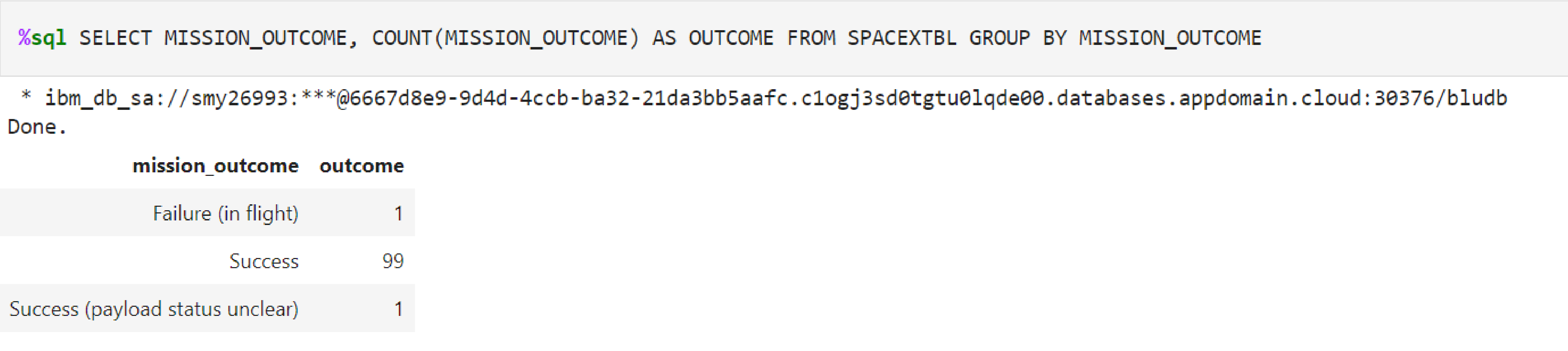
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

Explanation:

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes



```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

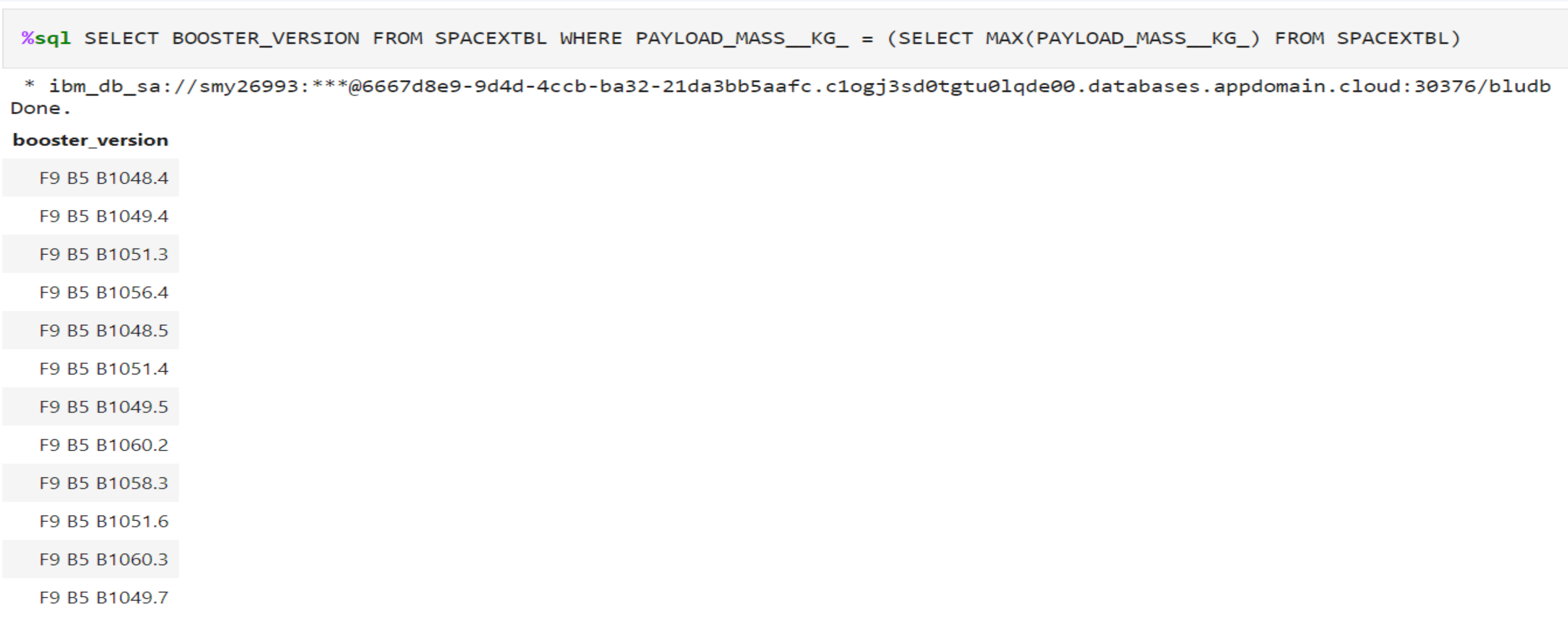
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

mission_outcome	outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

- Listing the total number of successful and failed mission outcomes

Boosters Carried Maximum Payload



```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

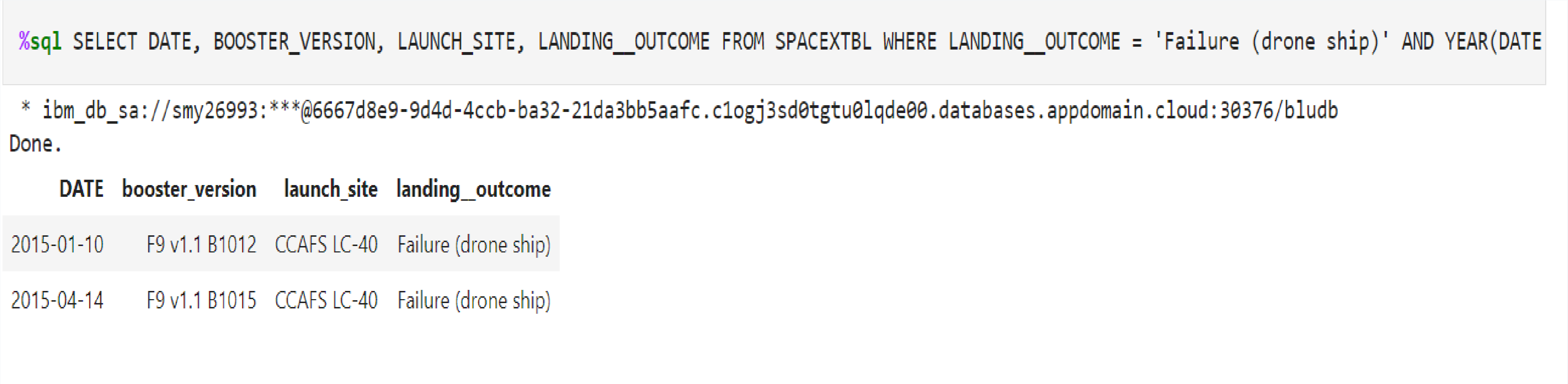
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

Explanation:

- Listing the names of the booster versions which have carried the maximum payload mass

2015 Launch Records



```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE
```

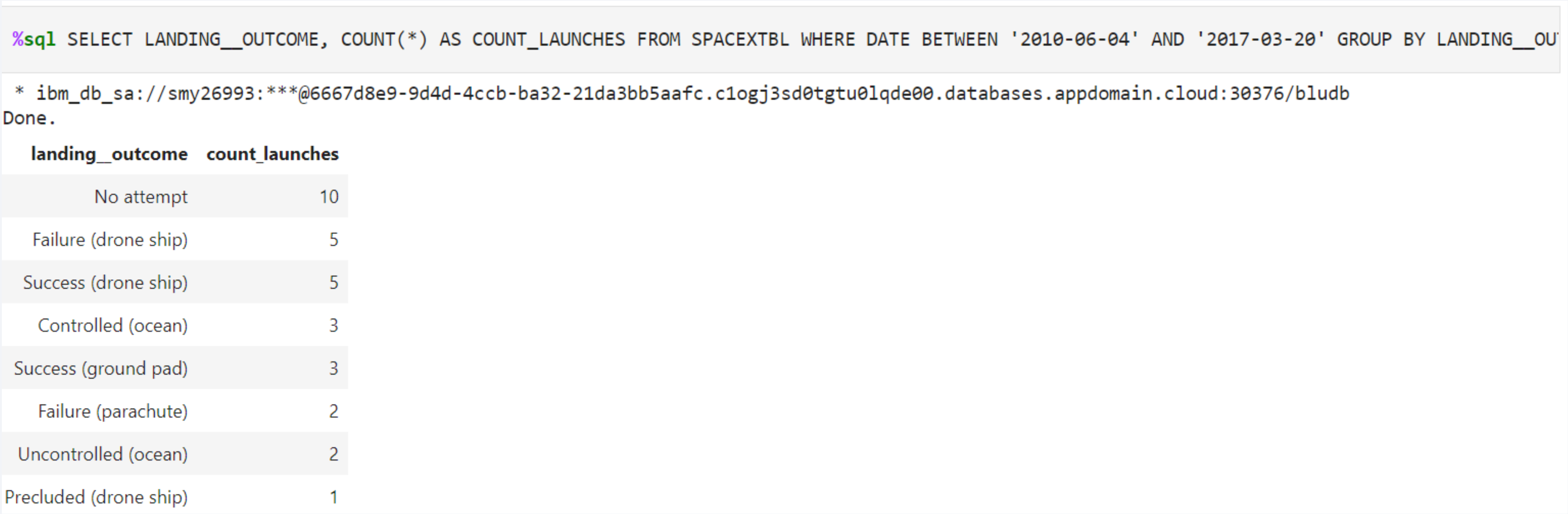
```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:

- Listing the failed landing outcomes in drone ships, their booster versions, and launch site names for the months in the year 2015

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



```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OU
```

```
 * ibm_db_sa://smy26993:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

Explanation:

- 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.

Section 3

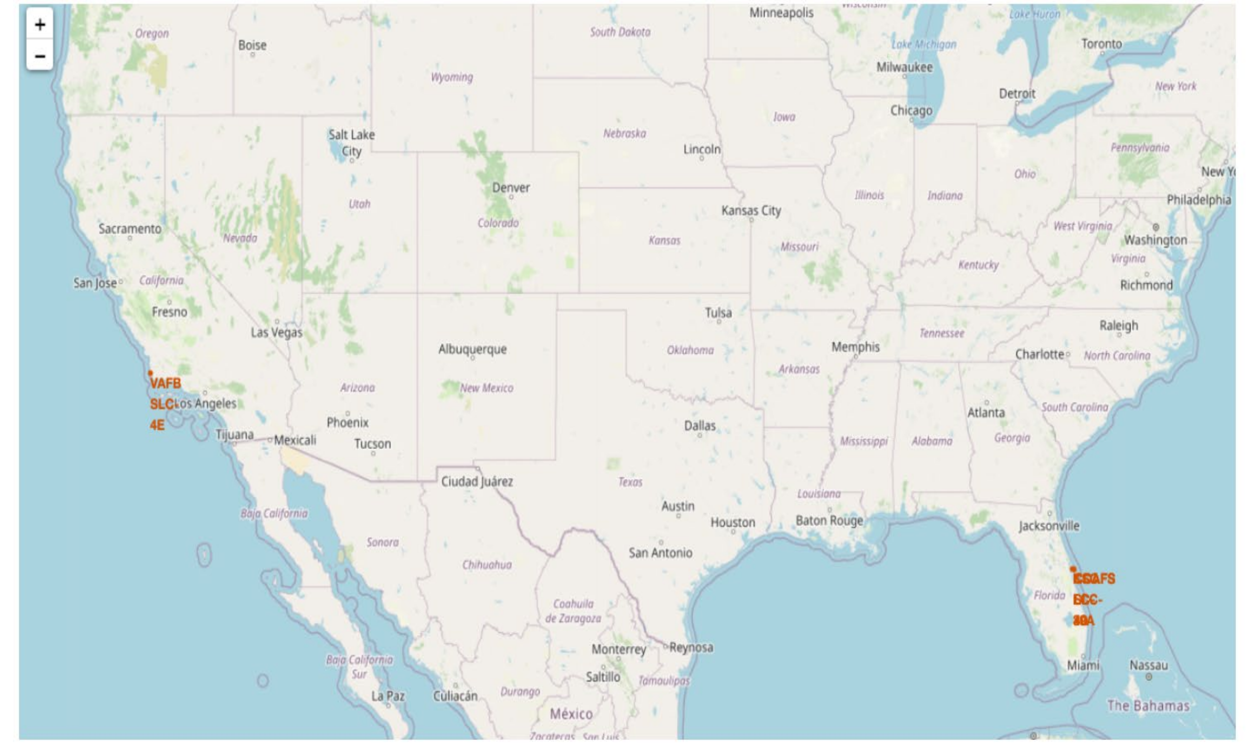
Launch Sites Proximities Analysis

All launch sites location markers on a global map

Explanation:

- Most of the launch sites are in proximity to the Equator line. The land is moving faster at the Equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the Equator is already moving at 1670 km/hour. If a ship is launched from the Equator, it goes up into space and is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit

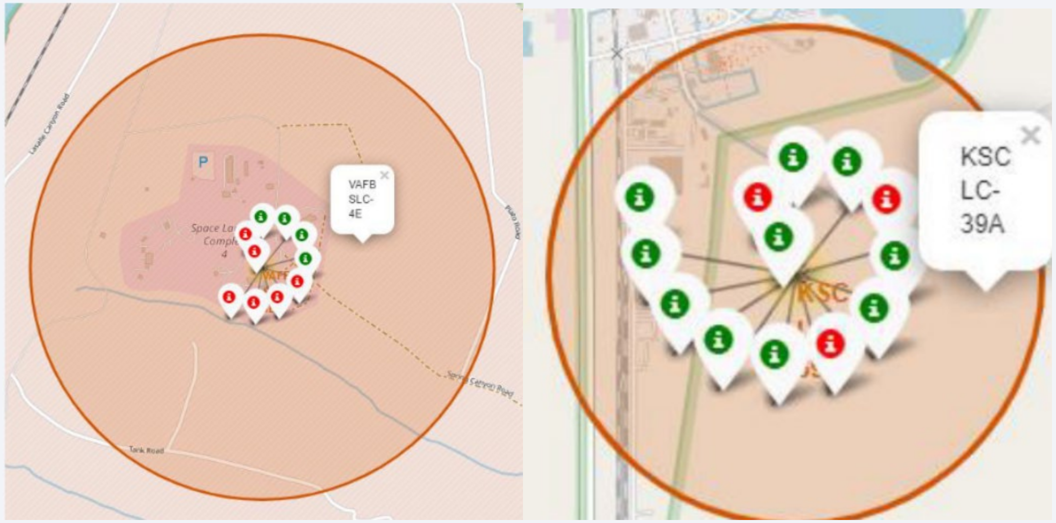
- All launch sites are in very close proximity to the coast. While launching rockets toward the ocean minimizes the risk of having any debris dropping or exploding near people.



Color-labeled launch records on the map

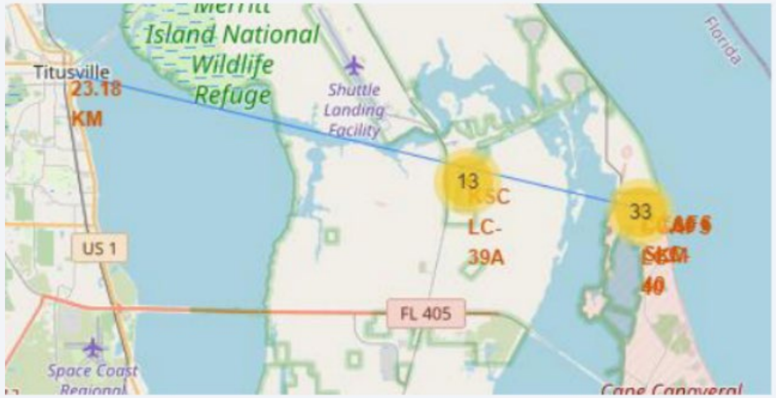
Explanation:

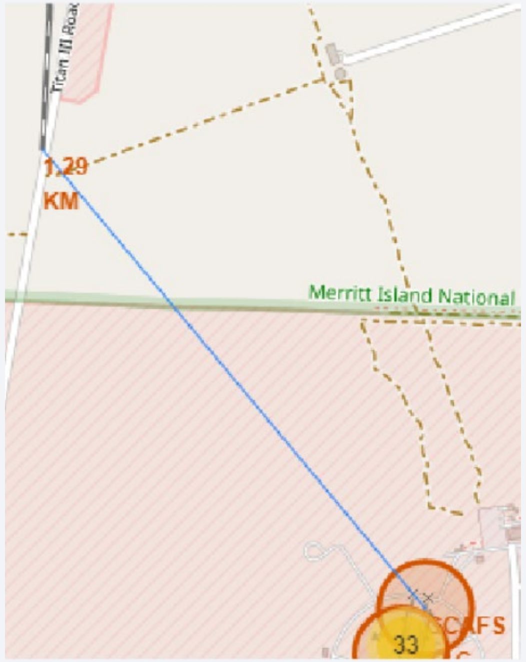
- From the color-labeled markers, we should easily identify which launch sites have relatively high success rates.
    - Green Marker = Successful launch
    - Red Marker = Failed launch
- Launch site KSC LC-39A has a very high Success Rate

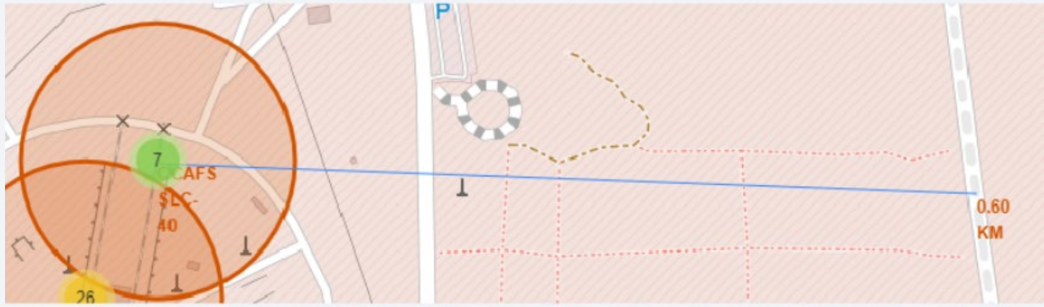


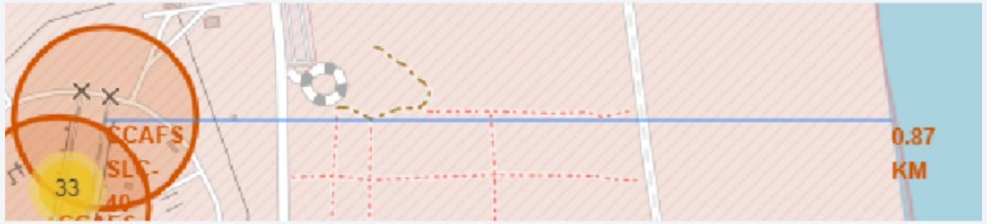
Distances between CCAFS SLC-40 and its proximities

- Explanation:
  - CCAFS SLC-40 is close to:
    - Railways
    - Highways
    - Coastline
    - Cities



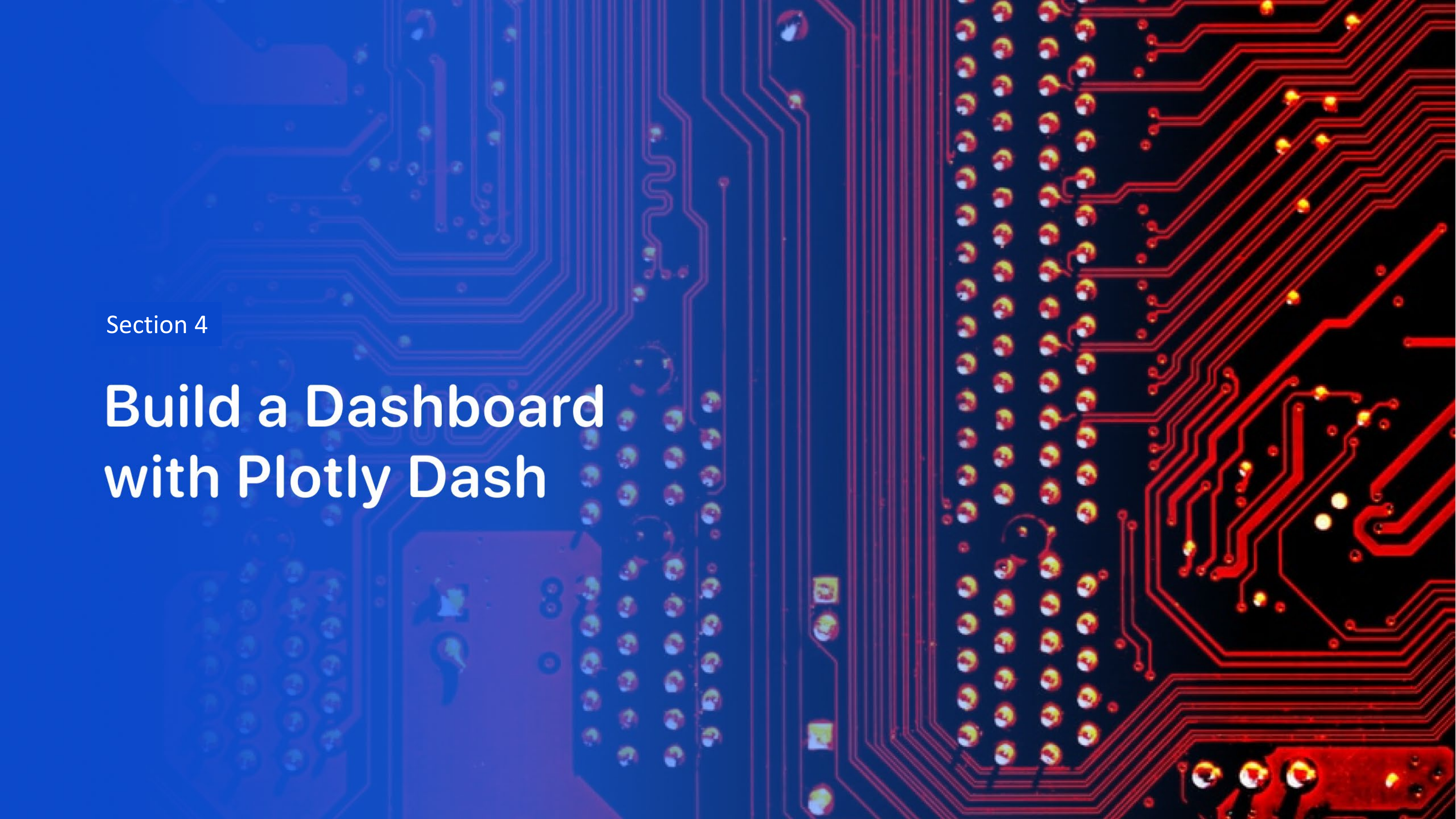




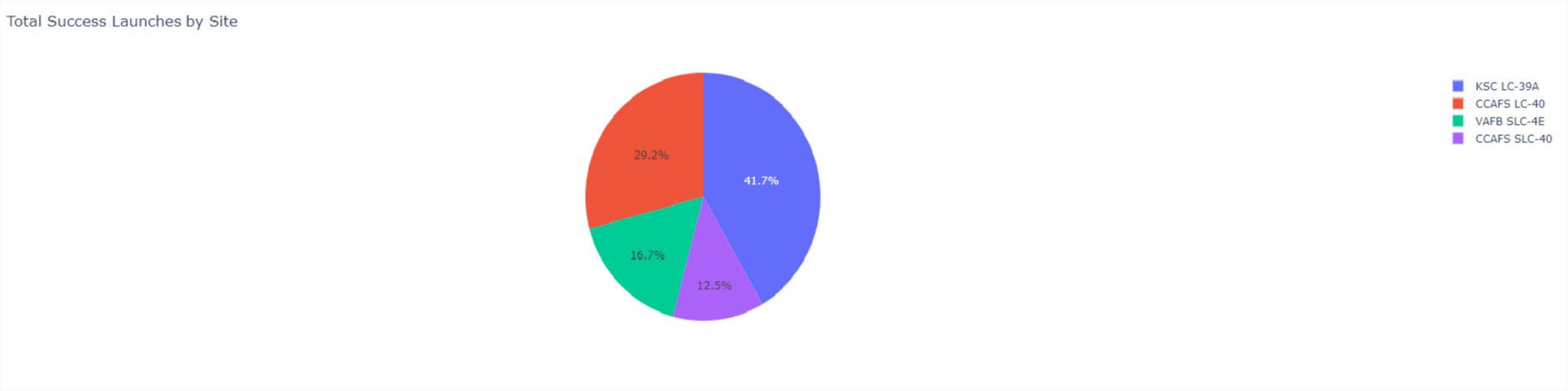


Section 4

Build a Dashboard with Plotly Dash



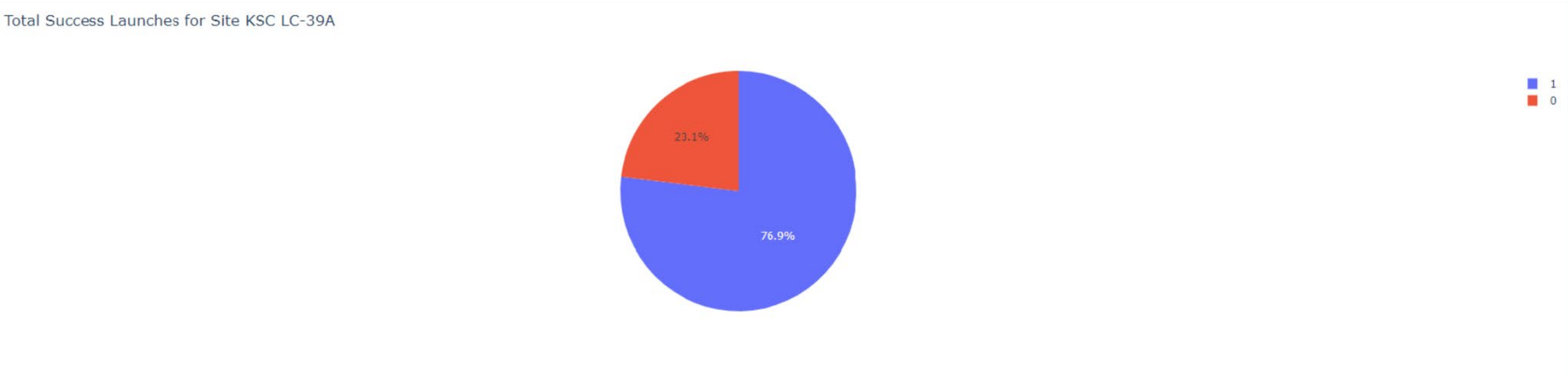
Total Success Launches by Site



Explanation:

- KSC LC39A has the best success rate of launches

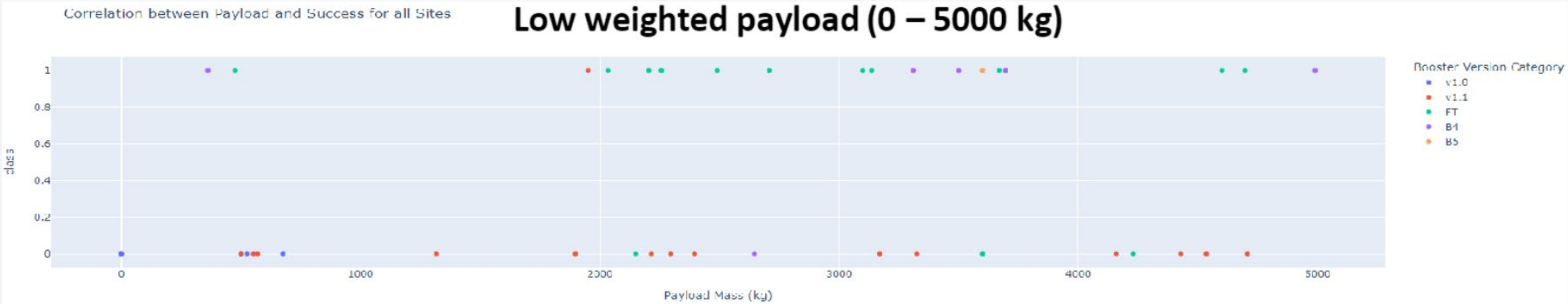
Total Success Launches for Site KSC LC-39A

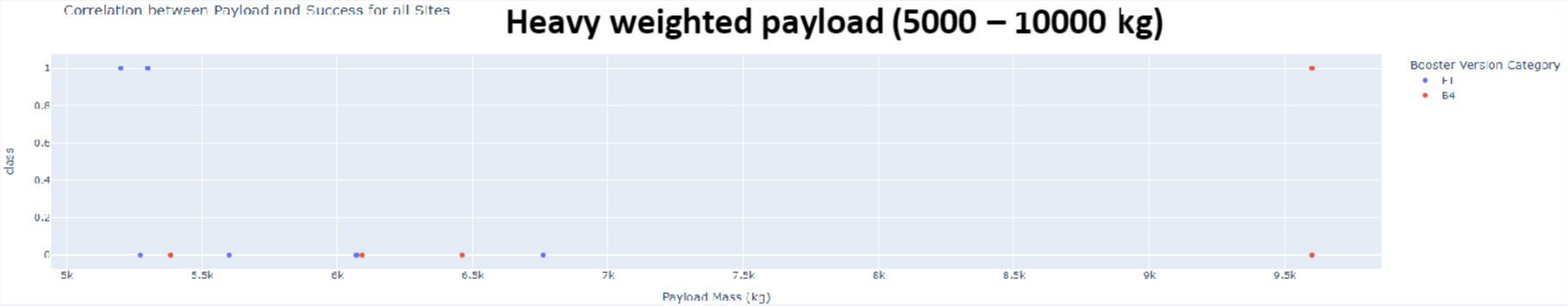


Explanation:

- KSC LC-39A has achieved a 76.9% success rate and a 23.1% failure rate

Dashboard – Payload mass vs. Outcome for all sites with different payload mass selected





Explanation:

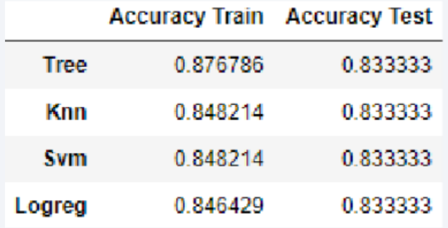
- Low weighted payloads have a better success rate than the heavily weighted payloads

Section 5

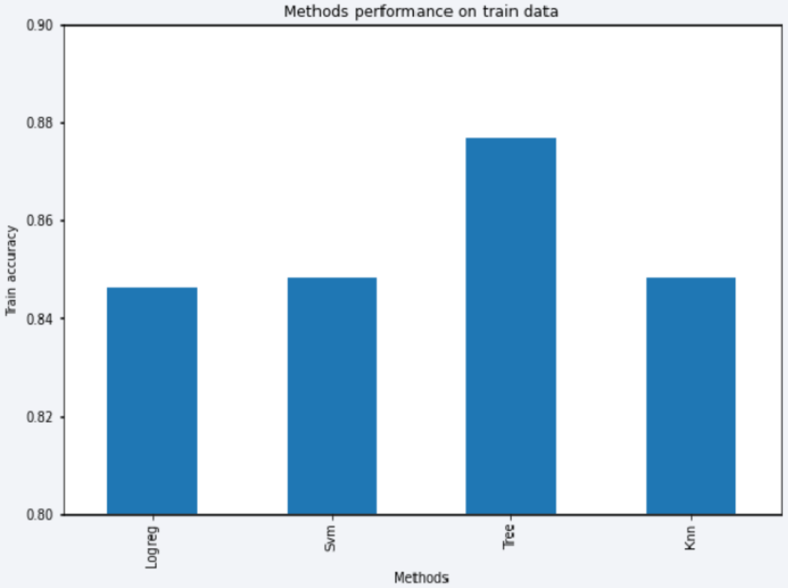
Predictive Analysis (Classification)

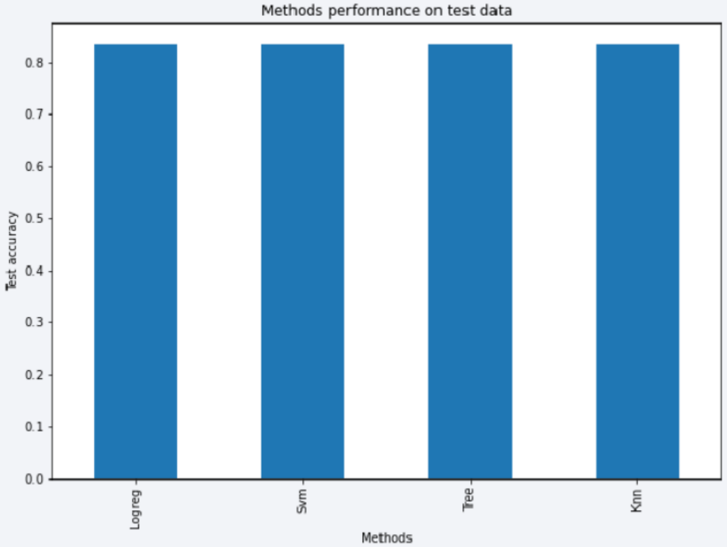
Classification Accuracy

- Explanation:
  - For the accuracy test, all methods performed similarly. We could get more test data to decide between them.
  - If we need to choose one, would be the decision tree.

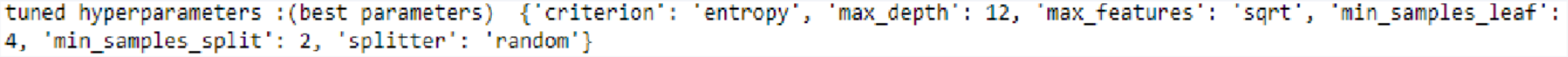


	Accuracy Train	Accuracy Test
Tree	0.876786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333





Decision tree best parameters

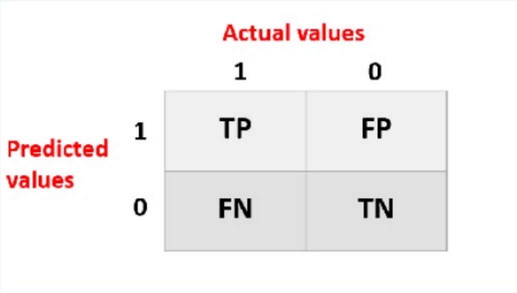


```
tuned hyperparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf':
4, 'min_samples_split': 2, 'splitter': 'random'}
```

Confusion Matrix

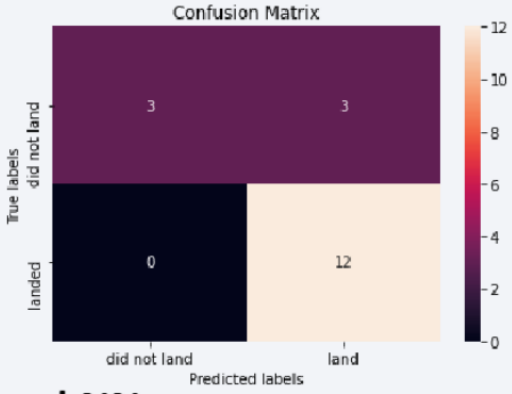
Explanation

- As the test accuracy is equal, the confusion matrices are also identical. The main problem of these models are the false positives

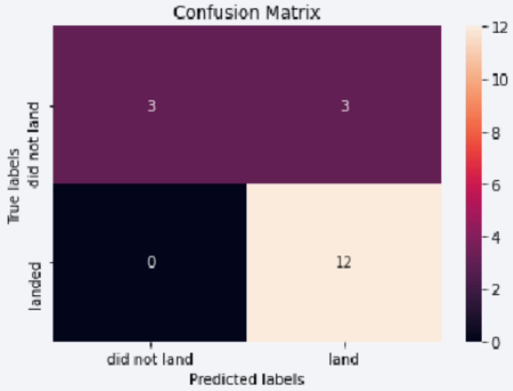


		Actual values	
		1	0
Predicted values	1	TP	FP
	0	FN	TN

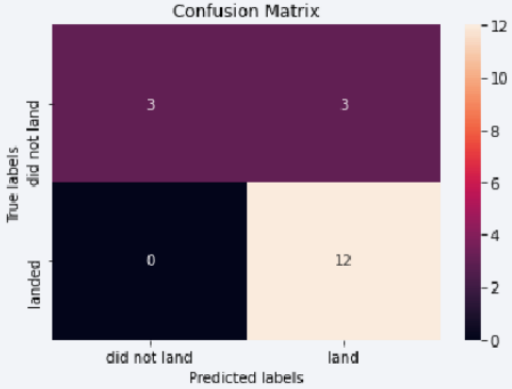
Logistic regression



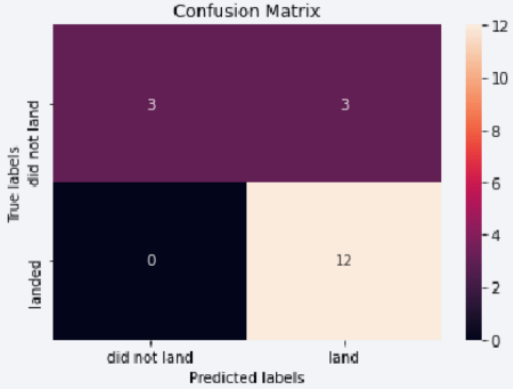
Decision Tree



kNN



SVM



Conclusions

- The success of a mission can be explained by several factors, such as the launch site, the orbit, and the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed them to go from failure to success
- The orbits with the best success rate are GEO, HEO, SSO, and ES-L1
- Depending on the orbits, the payload mass can be a criterion to consider for a mission's success. Some orbits require a light or heavy payload mass. But generally, low-weighted payloads perform better than heavily-weighted payloads
- With the current data, it is impossible to explain why some launch sites are better than others (KSC LC-39A) is the best launch site). To get an answer to this question, it should be necessary to obtain atmospheric or other relevant data
- For this dataset, we chose the Decision Tree Algorithm as the best model even if the test accuracy between the models were identical. The choice considered the better train accuracy.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!