IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yash Kumar

February 8th, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

The aim of this project was to make models for predicting whether the first stage of the Falcon 9 will land. The Data was collected using the SpaceX API as well as web scraping techniques. Data Wrangling was done to clean the data and assign outcome labels for the machine learning models.

SQL queries were used to investigate the data. The data was visualized by plotting several charts of different variables. Folium maps were plotted to visualize Launch Site locations and their proximities. An interactive Plotly Dash application was created to visualize the success rate for different payload ranges and booster versions. Different types of classification models were built and GridSearchCV was used to optimize the models.

The yearly success rate has continued to increase since 2013 and is high. The classification models built had a high accuracy in predicting and had an identical test accuracy of 83.3 %. However, the models need improvement to reduce false positives. Further insight was gained through data visualization such as the Site KSC LC-39A has the highest success rate. The data also showed that launches with certain booster versions such as B5 and FT and orbit types such as ES-L1, GEO, HEO and SSO had very high landing probability.

Introduction

- SpaceX is the most successful private company providing space travel. Its rocket launches cost only a fraction compared to other companies. SpaceX is pioneering because of its ability to reuse the first stage by re-landing it. The first stage is the largest and most expensive part of the spacecraft and does most of the work of propelling.
- The purpose of this project is to be able to predict whether the first stage of Falcon 9 will land successfully and subsequently predict the cost of the mission. Machine learning models were built to predict the landing outcome of the first stage. This gives insight of what specific factors affect the success rate and to what extent. By predicting the cost of a mission, there is necessary information and understanding to bid against SpaceX for launch missions.

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:

  - Data Collection using SpaceX API and Web scraping

- Perform data wrangling

  - Data Cleaning Techniques

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

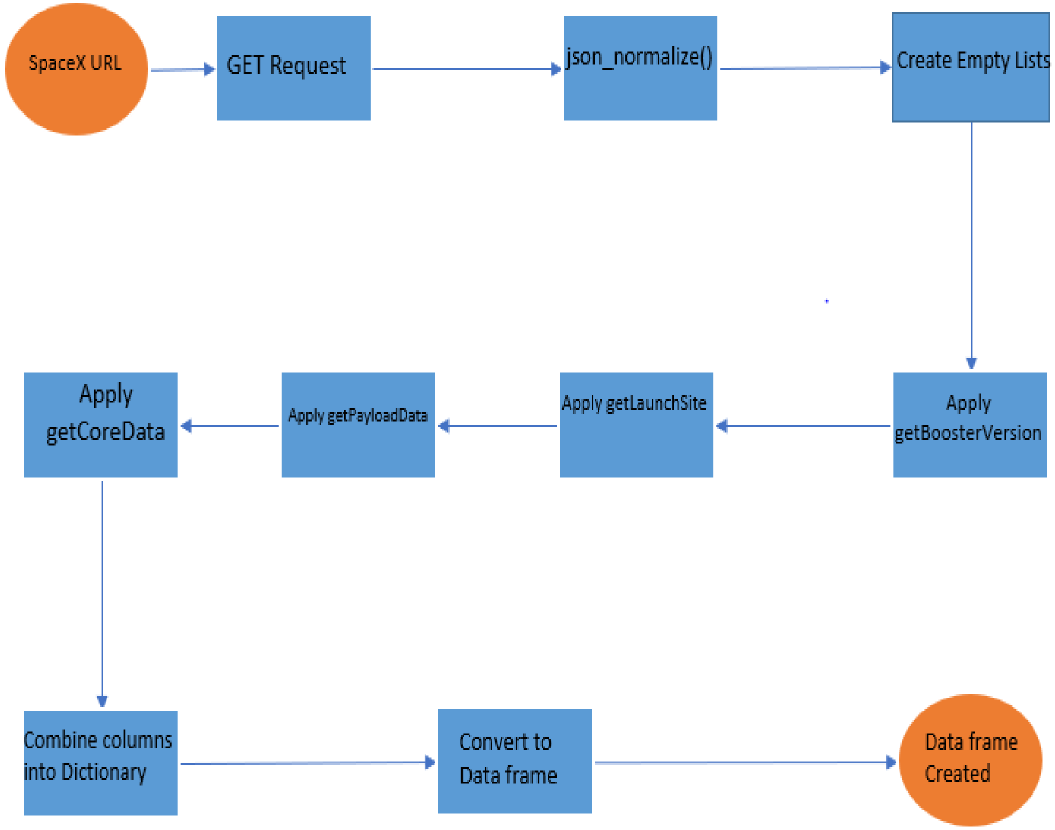
  - Model building, optimization and testing processes

Data Collection

- The datasets were collected using the SpaceX API and helper functions.
- Web scraping was used to collect data from a specific Launch records table on a Wikipedia page.

Data Collection- SpaceX API

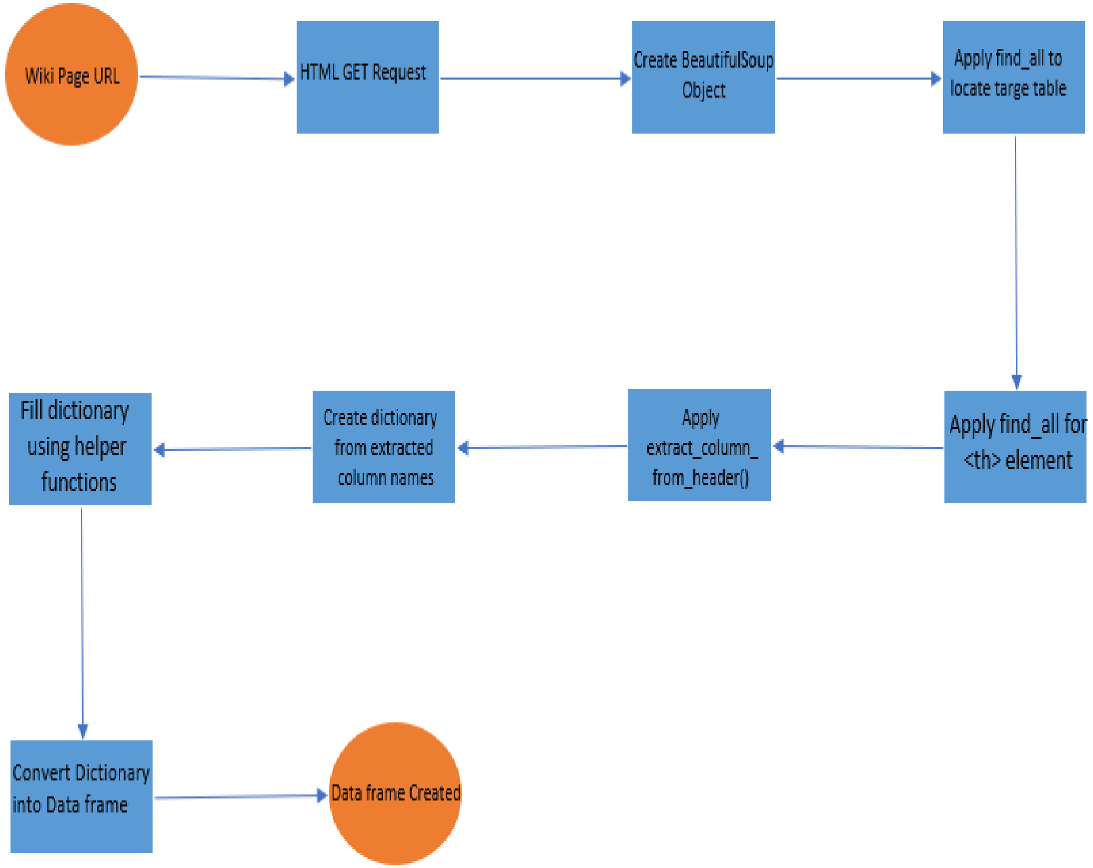
- The GET request method was used to connect to the API and get a static response object from a URL.
- The response object was decoded as JSON file and converted into a data frame using .json_normalize().
- Empty lists were created, and the auxiliary functions were applied to append them by connecting to the API.
- The lists were combined into a dictionary, which was converted into a data frame.



Notebook Link: https://github.com/yash500/Capstone-Project/blob/master/Data%20Collection%20API.ipynb

Data Collection – Web Scraping

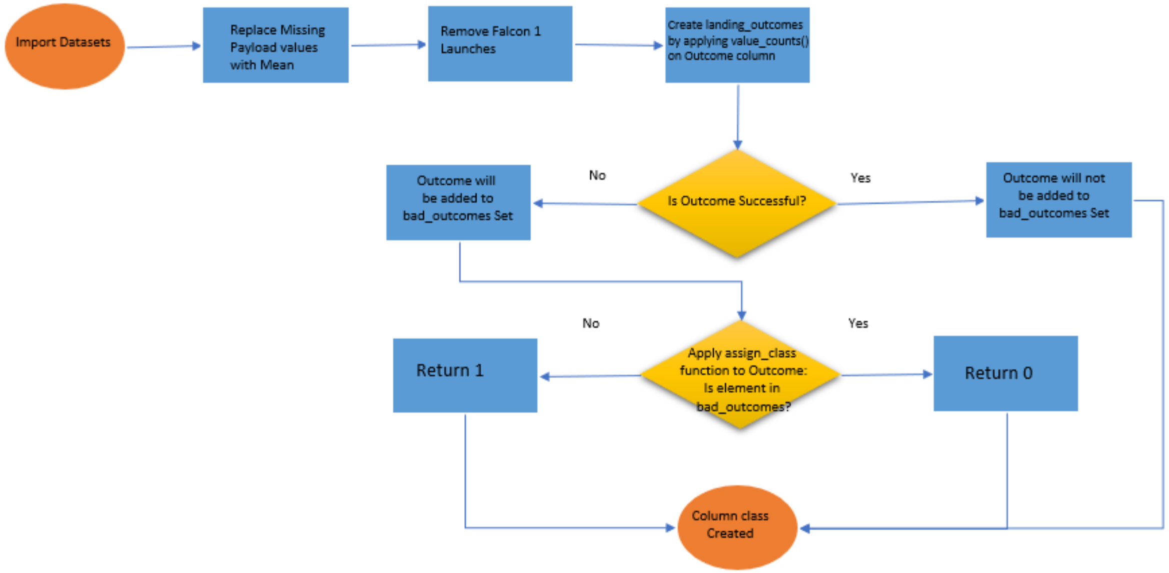
- The GET Request method was used on a Wikipedia URL. A BeautifulSoup object was created from the response.
- The find_all function was used to locate target table. The function extract_column_from_header was applied to the table.
- A dictionary was created using extracted column names and appended with extracted records using provided helper functions. Then, it was converted into a data frame.


Wiki Page URL
HTML GET Request
Create BeautifulSoup Object
Apply find_all to locate targe table
Apply find_all for <th> element
Apply extract_column_from_header()
Create dictionary from extracted column names
Fill dictionary using helper functions
Convert Dictionary into Data frame
Data frame Created

Notebook Link: https://github.com/yash500/Capstone-Project/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Missing Payload values were replaced by mean Payload values.
- The data was filtered by removing Falcon 1 launches.
- Using .value_counts() on column Outcome the variable landing_outcomes was created that showed the number of each unique landing outcome.
- The set bad_outcomes was created that included all the unsuccessful outcomes.
- A function was defined that returns 0 if element is in bad_outcomes, otherwise returns 1. Then the function was applied on Outcome to make a new column Class.



Import Datasets

Replace Missing Payload values with Mean

Remove Falcon 1 Launches

Create landing_outcomes by applying value_counts() on Outcome column

Is Outcome Successful?

No

Outcome will be added to bad_outcomes Set

Yes

Outcome will not be added to bad_outcomes Set

Apply assign_class function to Outcome: Is element in bad_outcomes?

No

Return 1

Yes

Return 0

Column class Created

Notebook Link: https://github.com/yash500/Capstone-Project/blob/master/EDA.ipynb

EDA with Data Visualization

- Scatter plots were plotted to visualize patterns of how launch outcomes were affected by different variables. The following scatter charts were plotted:

  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit Type
  - Payload vs Orbit Type

- A bar chart was plotted to visualize the success rate of each orbit type.
- A line chart was plotted to visualize the yearly success rate of launches.

Notebook Link: https://github.com/yash500/Capstone-Project/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- The SELECT DISTINCT statement was used to find unique values of a column such as unique launch sites and total number of successful and failed missions.
- The SUM and AVG statements were used in performing arithmetic functions on payload column e.g., displaying the average payload carried by a booster version.
- The MIN statement was useful in finding the first date a certain landing outcome was achieved.

Notebook Link: https://github.com/yash500/Capstone-Project/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers were created to mark and label the position of the launch sites, proximities, distances and the failed/successful launches.
- Circles were added to make it easier to visualize the launch site locations.
- A MarkerCluster was created to add the failed/successful launches.
- MousePosition was used to determine the coordinates of the launch sites and proximities.
- The distance between the launch sites and their proximities were calculated using the provided distance function.
- PolyLine objects were created to draw lines between the launch sites and their proximities.

Notebook Link: https://github.com/yash500/Capstone-Project/blob/master/Data%20Visuzlization%20with%20Folium.ipynb

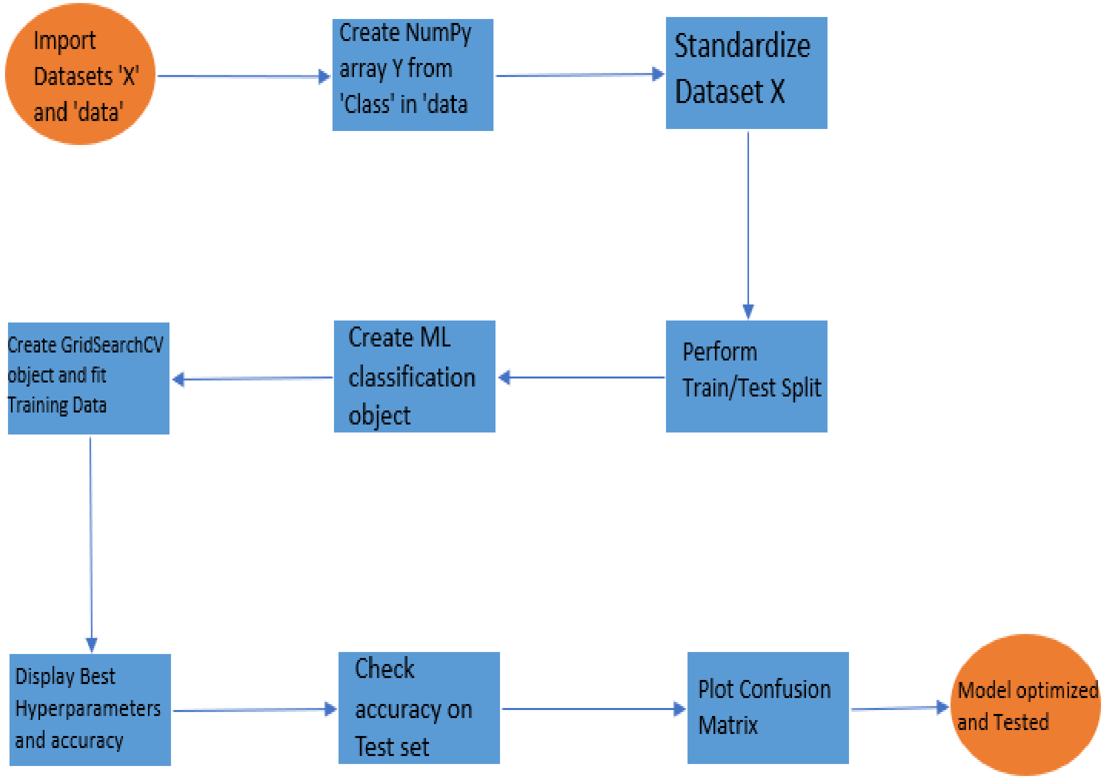
Build a Dashboard with Plotly Dash

- A total success launches pie chart for each launch site and a scatter plot of outcome against Payload mass were added.
- A Launch Site Drop-down input component was added to filter the graphs to show data for either all the sites or a specific site.
- A range slider was added to filter the payload mass ranges for the scatter plot.

Notebook Linkhttps://github.com/yash500/Capstone-Project/blob/master/Interactive%20Dashboard%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

- The datasets X and data were loaded. A numpy array Y was created from ‘class’ in data.
- The dataset X was standardized using the preprocessing.StandardScaler().fit(X).transform(X) function.
- The train_test_split function was used to split the data into training and testing data.
- Logistic Regression, SVM, Decision Tree and KNN models were created to predict the outcome.
- The GridSearchCV objects were used to find the best hyperparameters and optimize the models.
- Confusion matrices were plotted for each model to analyze the accuracy and check whether there were false positives of false negatives.


Import Datasets 'X' and 'data'
Create NumPy array Y from 'Class' in 'data
Standardize Dataset X
Perform Train/Test Split
Create ML classification object
Create GridSearchCV object and fit Training Data
Display Best Hyperparameters and accuracy
Check accuracy on Test set
Plot Confusion Matrix
Model optimized and Tested

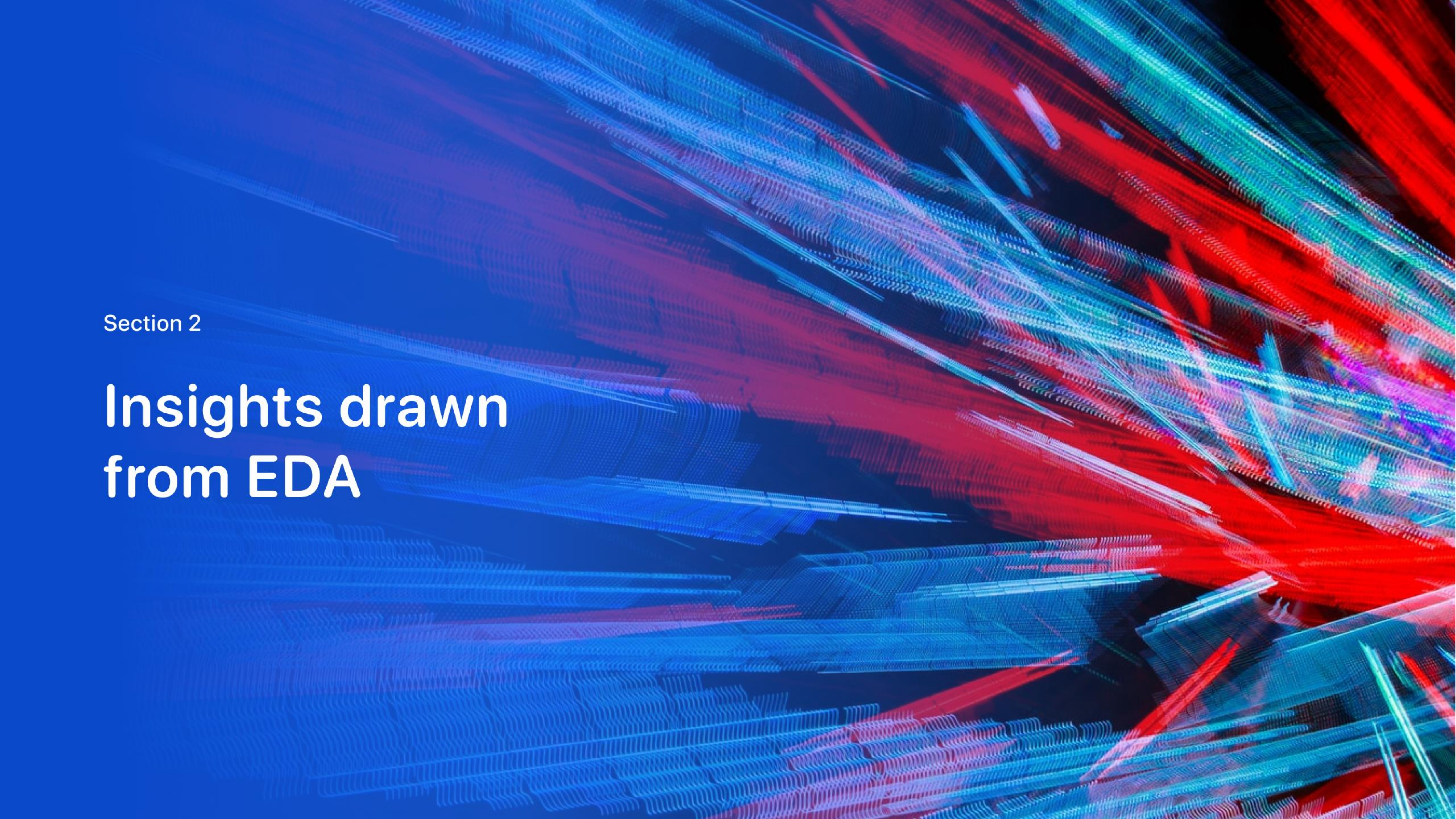
Notebook Link: https://github.com/yash500/Capstone-Project/blob/master/Predictive%20Analysis

Results

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

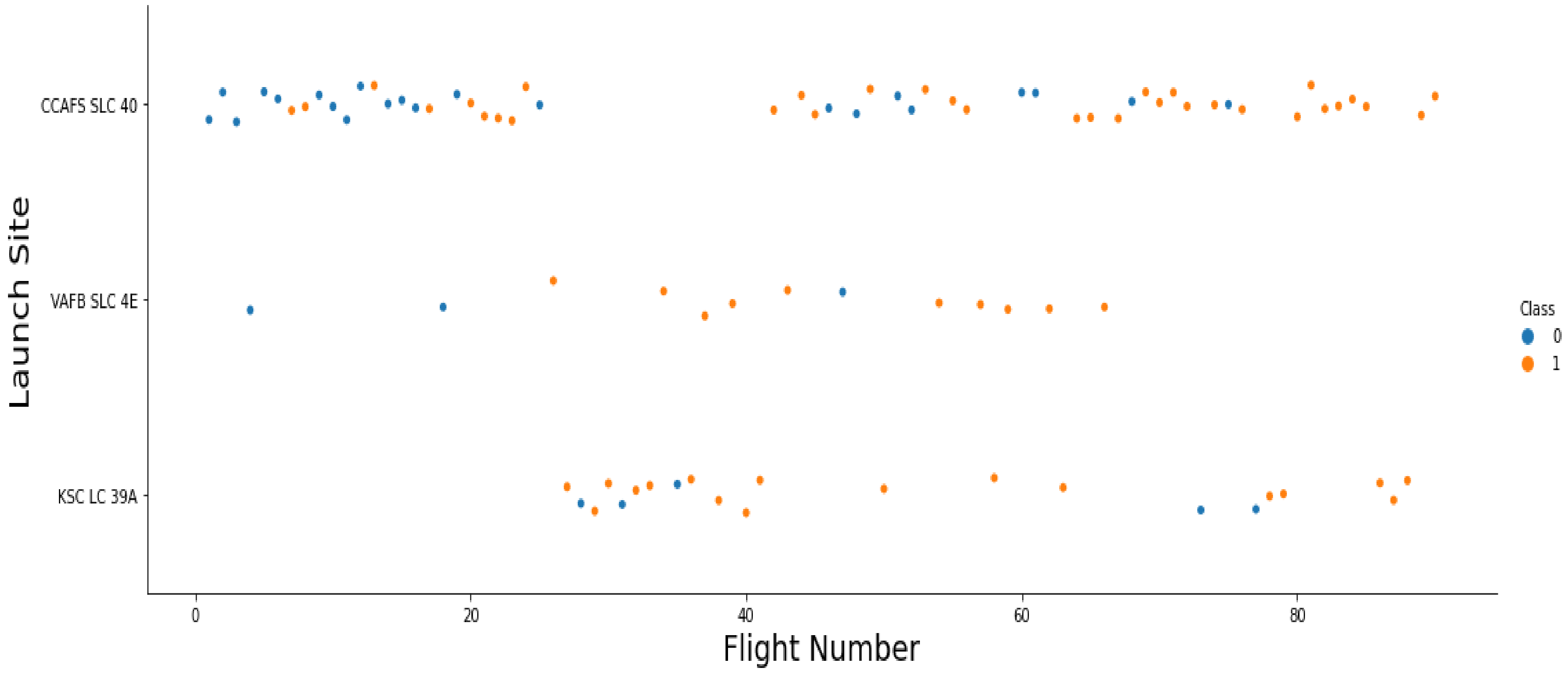
Section 2

Insights drawn from EDA



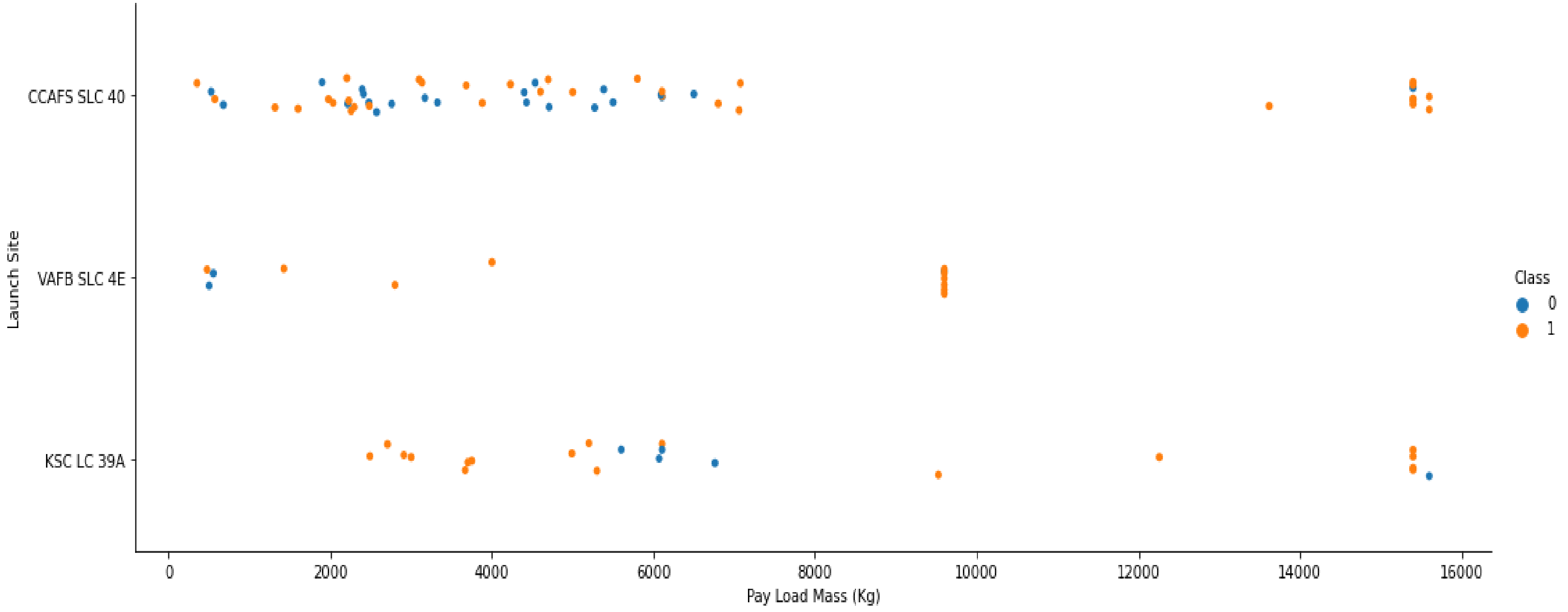
Flight Number vs. Launch Site

- The success rate has increased with flight number for VAFB SLC-4E.



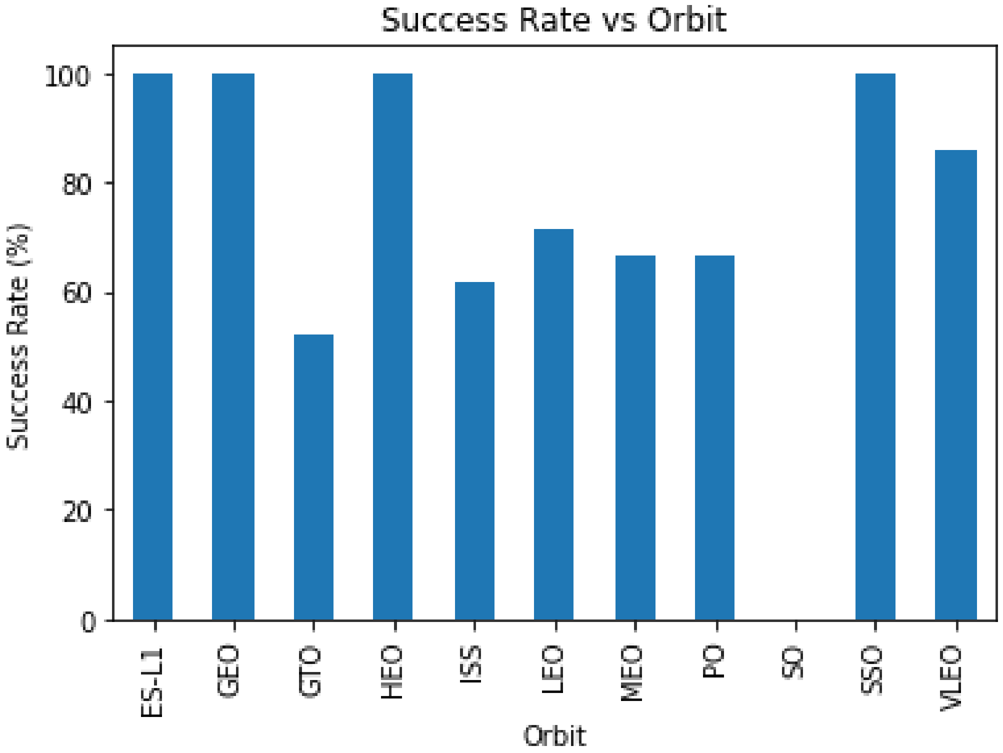
Payload vs. Launch Site

- For VAFB SLC-4E the success rate increases as Payload Mass increases.
- The site VAFB SLC-4E does not have launches for Payload mass greater than 10000 Kg



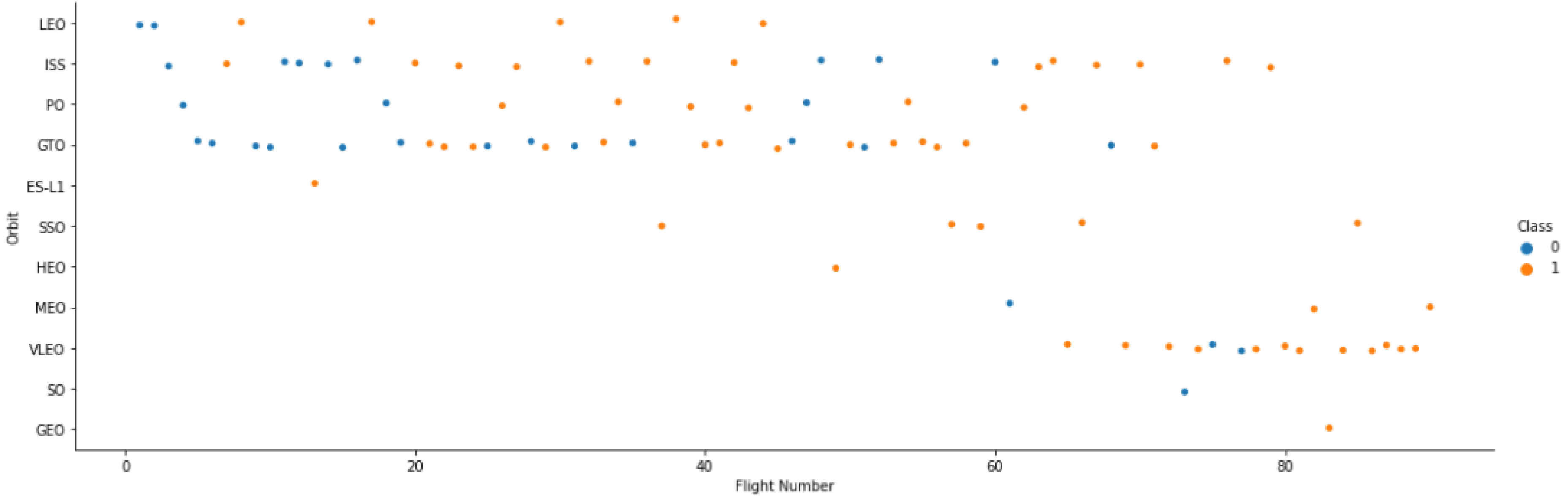
Success Rate vs. Orbit Type

- The graph shows success rates for different orbit types.
- The orbits ES-L1, GEO, HEO and SSO have very high success rates of 100%
- The orbit SO has a success rate of 0%.



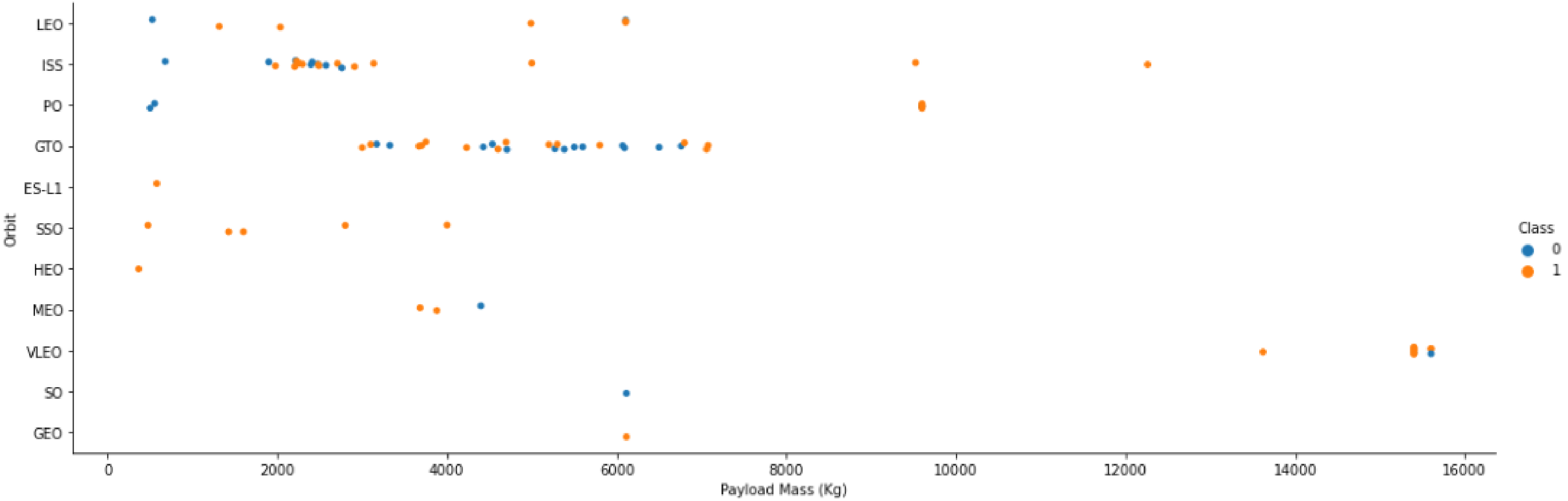
Flight Number vs. Orbit Type

- This is a scatter plot of flight number against orbit type shows that for the orbit type LEO, the success rate increases as flight number increases.
- For the orbit type GTO there is not relation between success rate and flight number



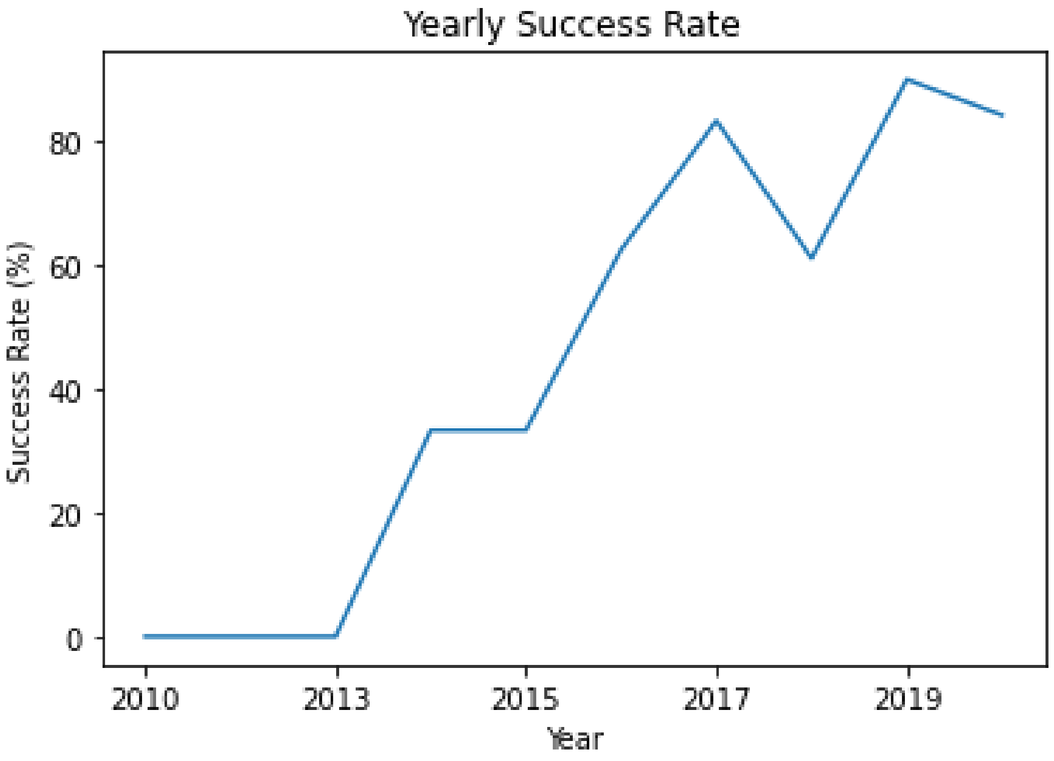
Payload vs. Orbit Type

The scatter plot shows that for orbit types LEO, ISS and Polar the success rate increases as Payload mass increases



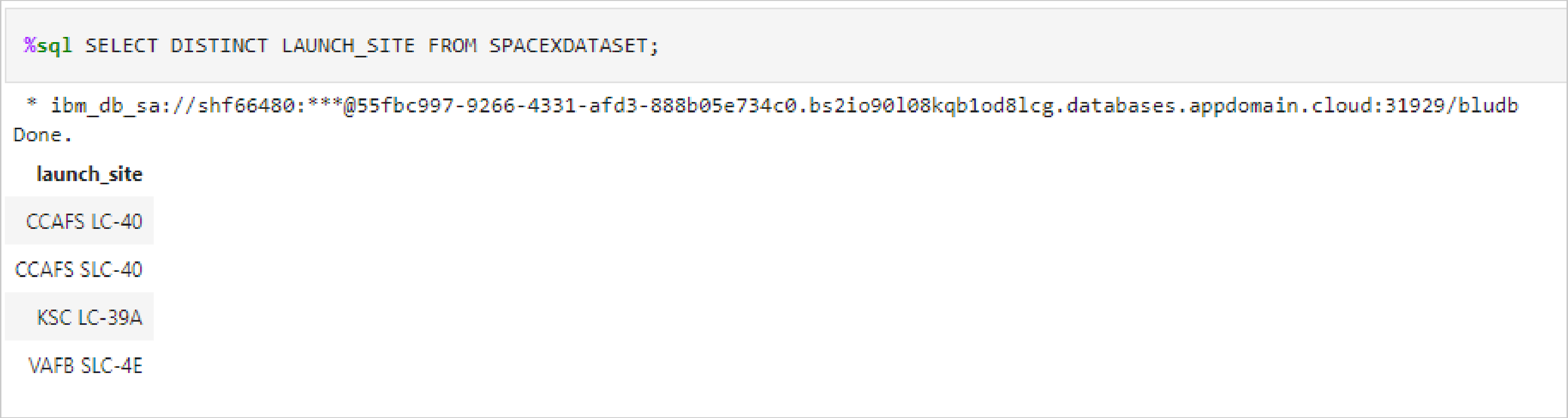
Launch Success Yearly Trend

From 2013 onwards the yearly success rate has continued to increase going as high as 90%.



All Launch Site Names

There were 4 distinct Launch Sites found in the Data Set using the SELECT DISTINCT statement.



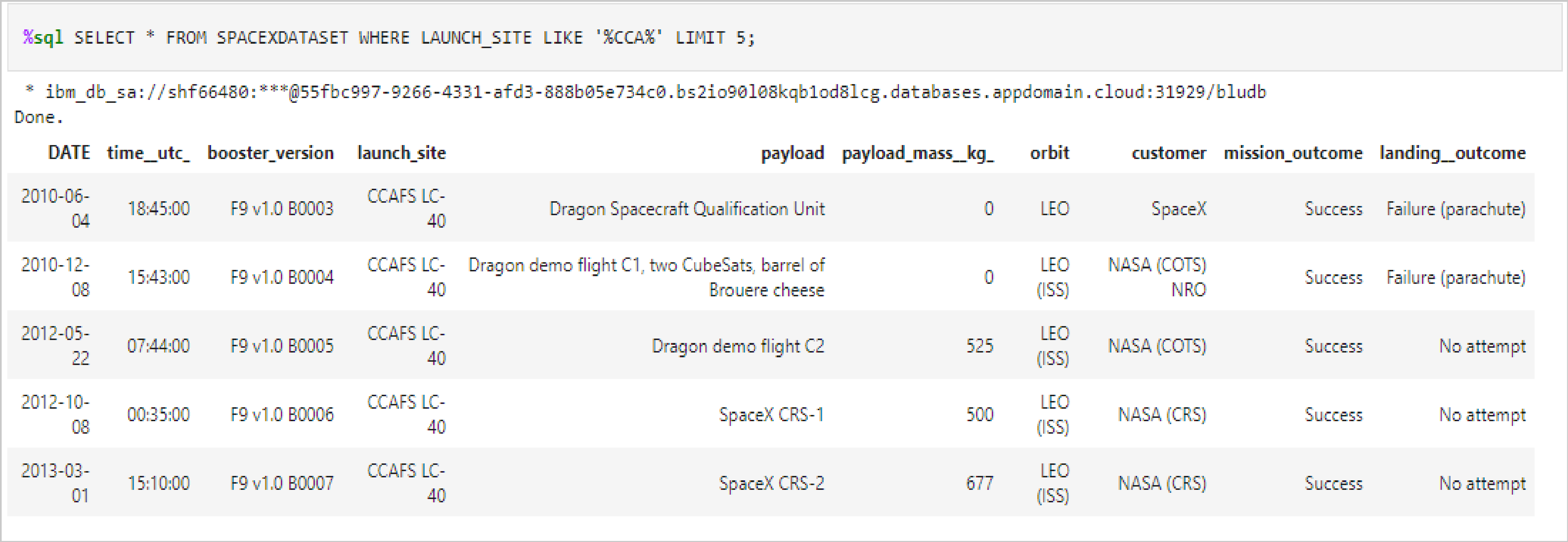
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXDATASET;
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

The SELECT * statement was used to return 5 records where launch sites begin with ‘CCA’



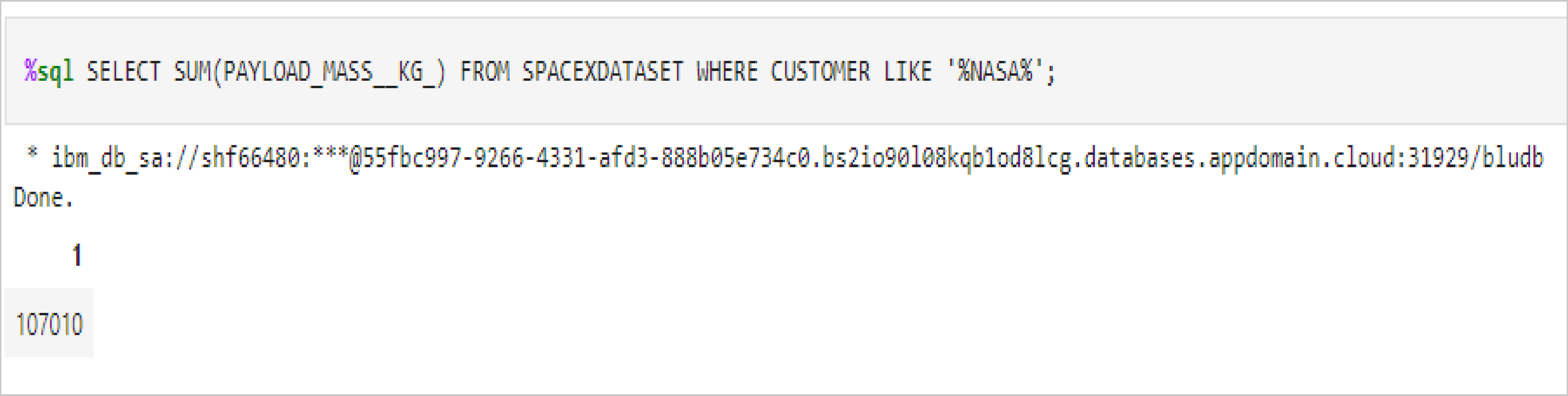
```
%sql SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE '%CCA%' LIMIT 5;
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/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

Total Payload Mass

The total Payload carried for NASA was found to be 107010 Kg using the SUM statement.

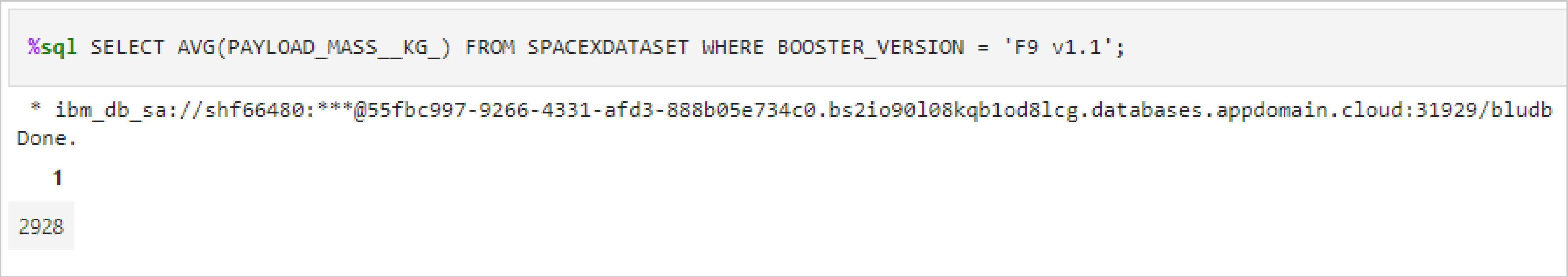


```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE CUSTOMER LIKE '%NASA%';
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
     1
107010
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was found to be 2928 KG using the SELECT AVG statement.
- This shows that the booster version F9 v1.1 was used for lower payload masses.



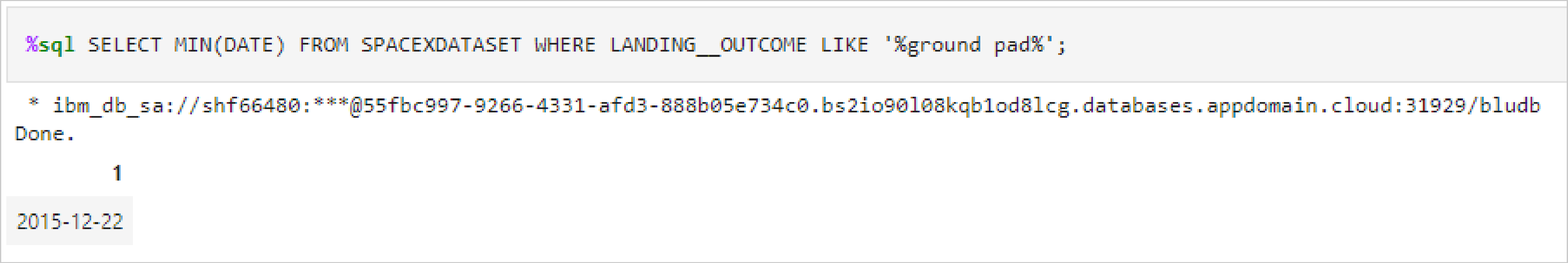
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

1
2928

First Successful Ground Landing Date

- The first successful landing on ground pad was achieved on December 22, 2015.
- This result was found using the SELECT MIN statement.



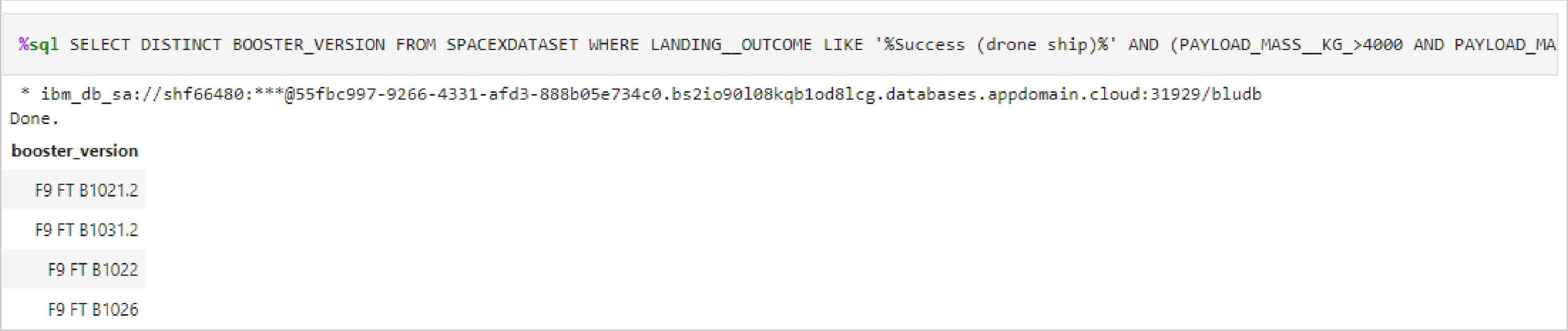
```
%sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE LANDING__OUTCOME LIKE '%ground pad%';
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The booster versions that have successfully landed on a drone ship and had a payload mass between 4000 and 6000 KG are F9 FT B1021.2, F9 FT B1031.2, F9 FT 1022 and F9 FT B1026.
- All these boosters were of type FT
- These were found using the SELECT DISTINCT statement.



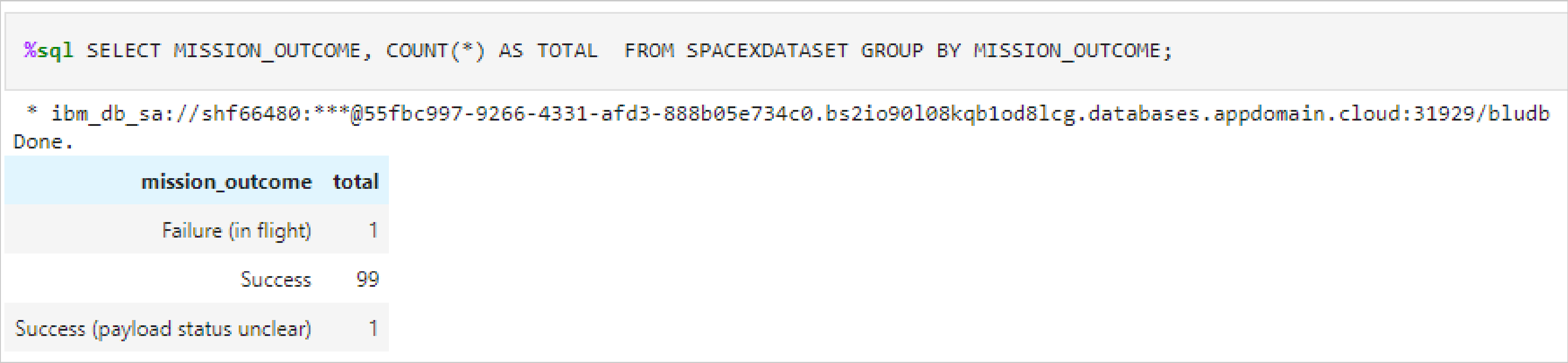
```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXDATASET WHERE LANDING__OUTCOME LIKE '%Success (drone ship)%' AND (PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MA
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes was 100 and the total number of failed mission outcomes was just 1. One of the successful mission outcomes had payload status unclear.
- This result was found using the SELECT , COUNT (*) statements.



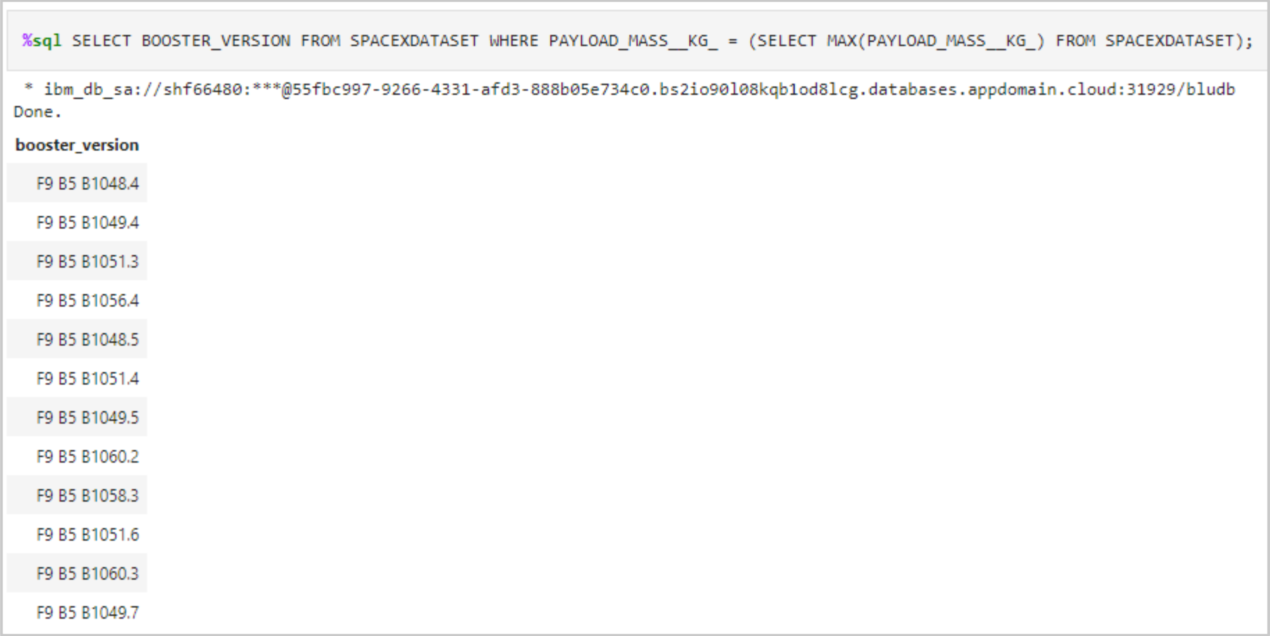
```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL  FROM SPACEXDATASET GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

Boosters Carried Maximum Payload

- Using the SELECT and SELECT MAX statement the booster versions which carried the maximum payload mass were found.
- All the booster versions were of the type B5.



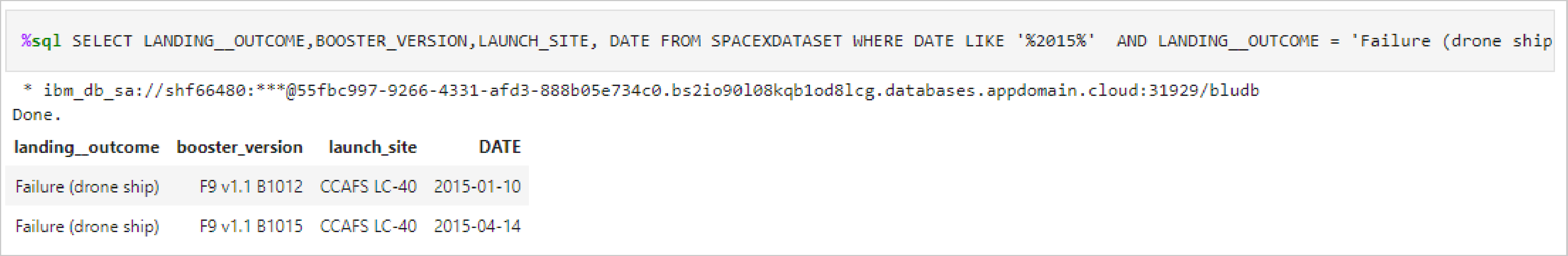
```
%sql SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/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

2015 Launch Records

- The following is a list of the failed landing outcomes in drone ship in 2015.
- The result was found using the SELECT statement.
- Both the records were for booster version F9 v1.1 launched from CCAFS LC-40.



```
%sql SELECT LANDING__OUTCOME,BOOSTER_VERSION,LAUNCH_SITE, DATE FROM SPACEXDATASET WHERE DATE LIKE '%2015%'  AND LANDING__OUTCOME = 'Failure (drone ship
```

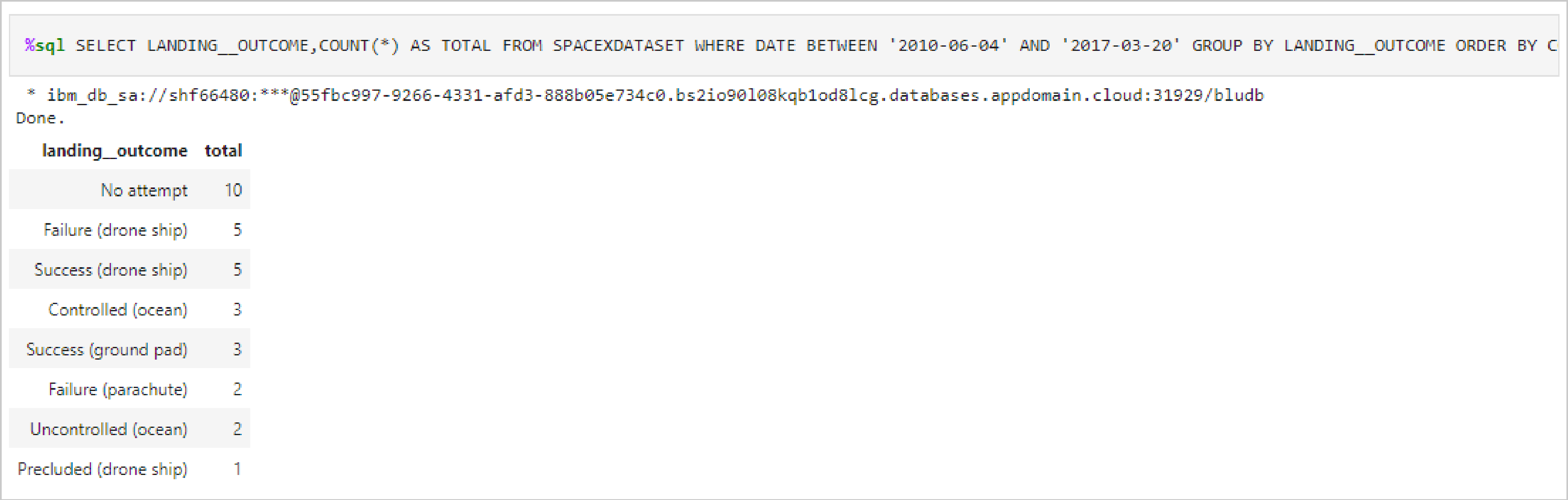
```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

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

- The following is a rank of the count of landing outcomes between June 4, 2010, and 20 March, 2017.

- This shows that the largest landing outcome type was no attempt in this time period.



```
%sql SELECT LANDING__OUTCOME,COUNT(*) AS TOTAL FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY C
```

```
 * ibm_db_sa://shf66480:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

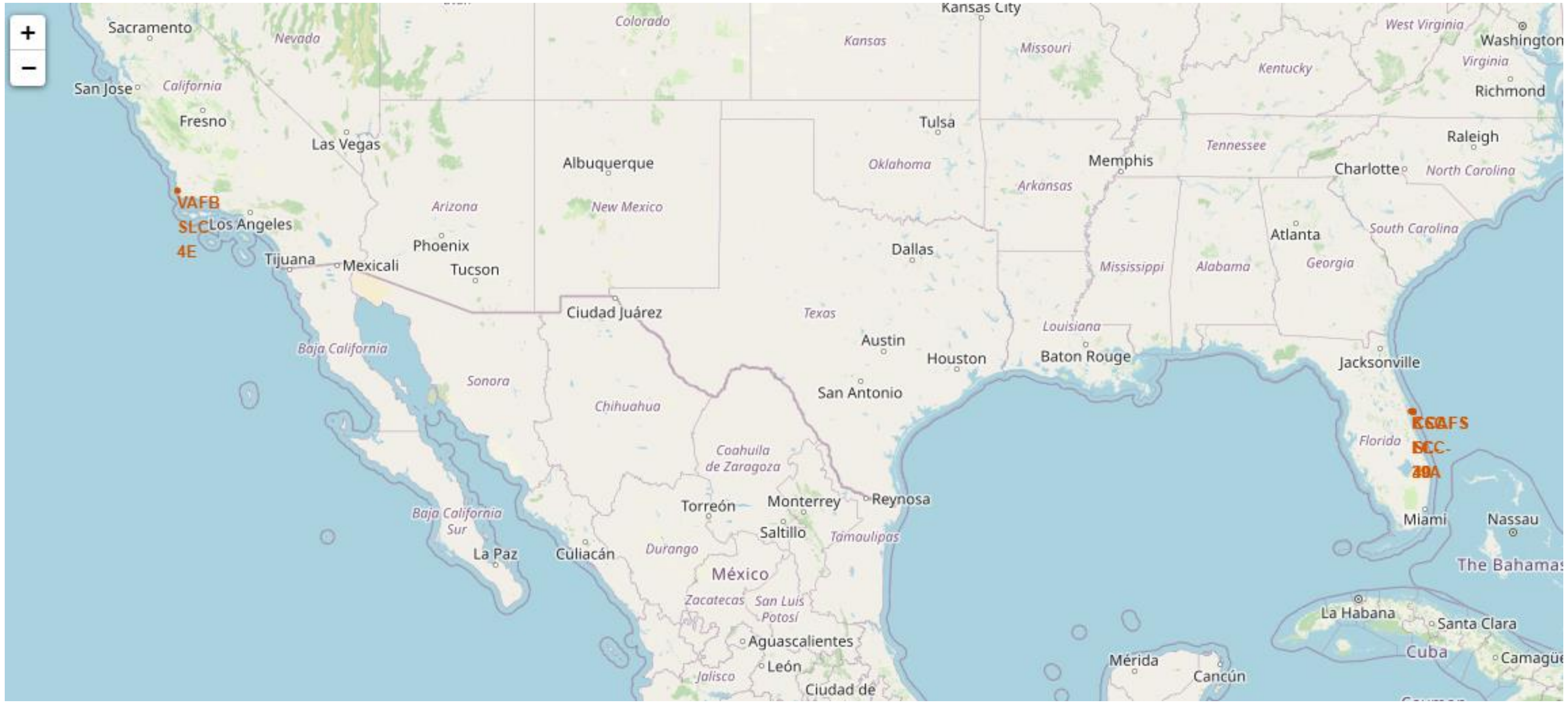
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

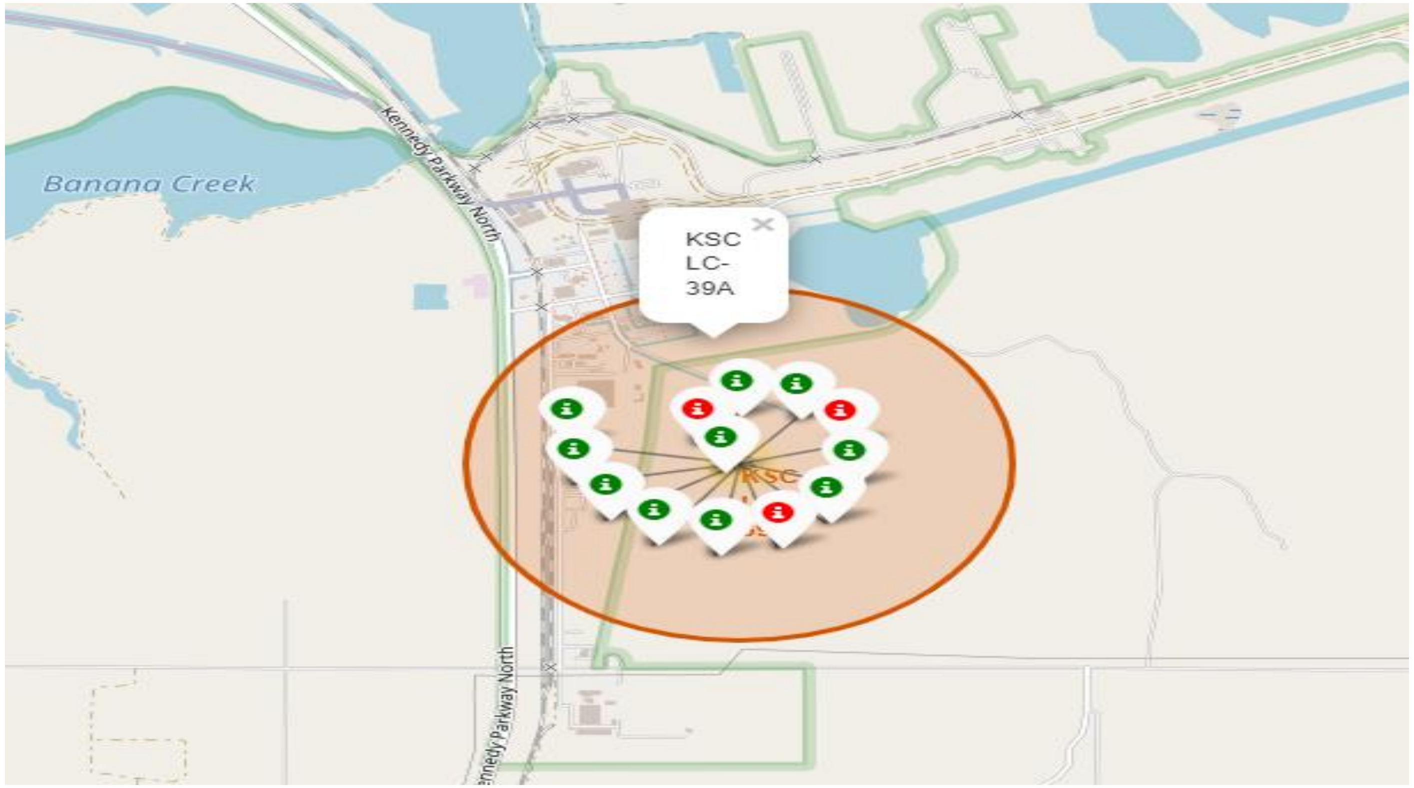
Launch Site Location Markers on Folium Maps

This graph shows that all the launch site locations are close to the coastline and close to the equator.



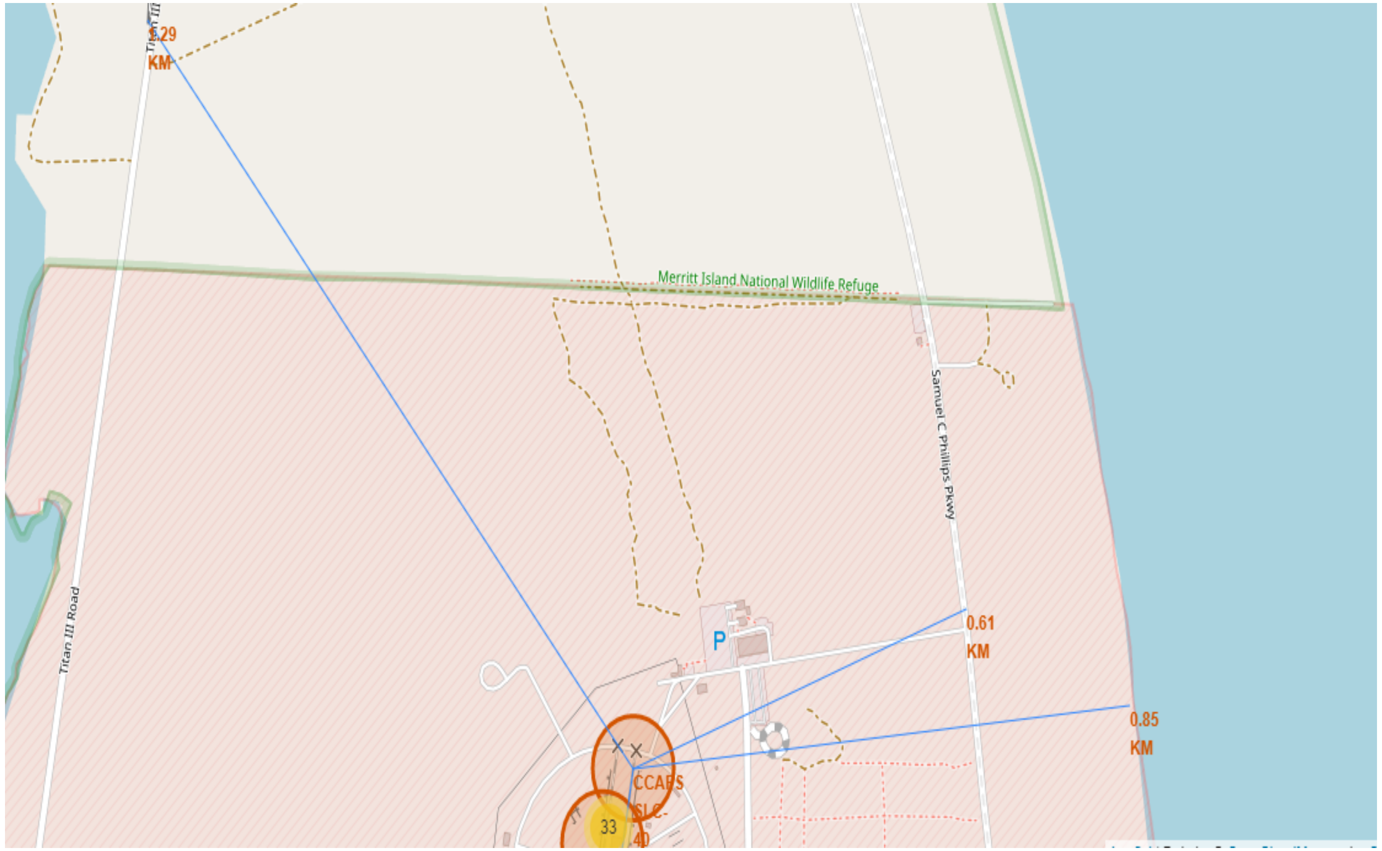
Launch Outcome Markers on Folium Maps

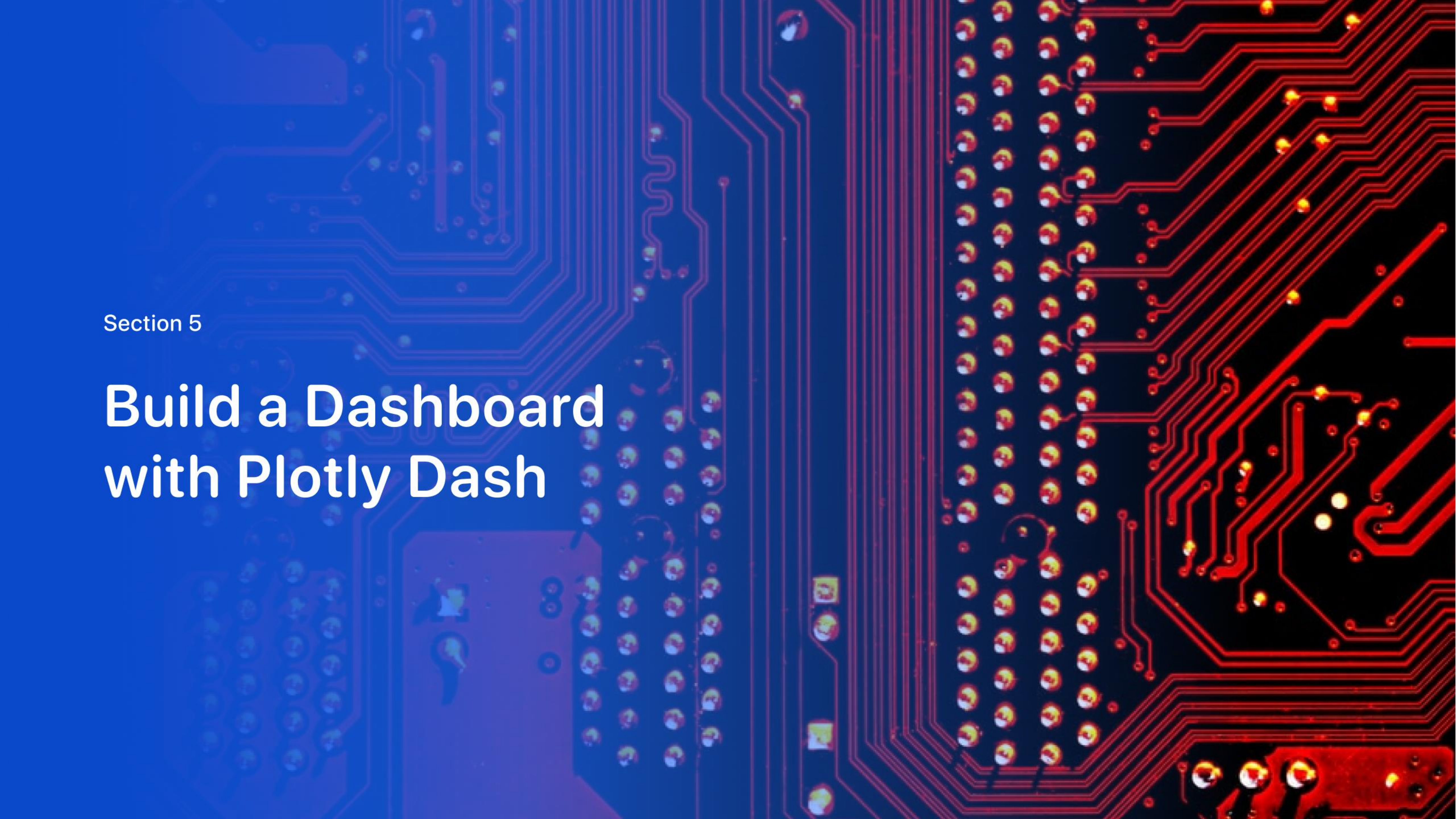
- The color labeled markers clearly show the successful(green) and unsuccessful(red) launches.
- This is a clear indicator of the high success rate of KSC LC-39A



PolyLine Distance Markers on Folium Maps

- The distance markers show the distance between the launch site and its proximities.
- The launch site is close to the coastline, which is a safe location for launching and landing.
- The launch site is close to the highways and railways which is convenient for transporting materials.



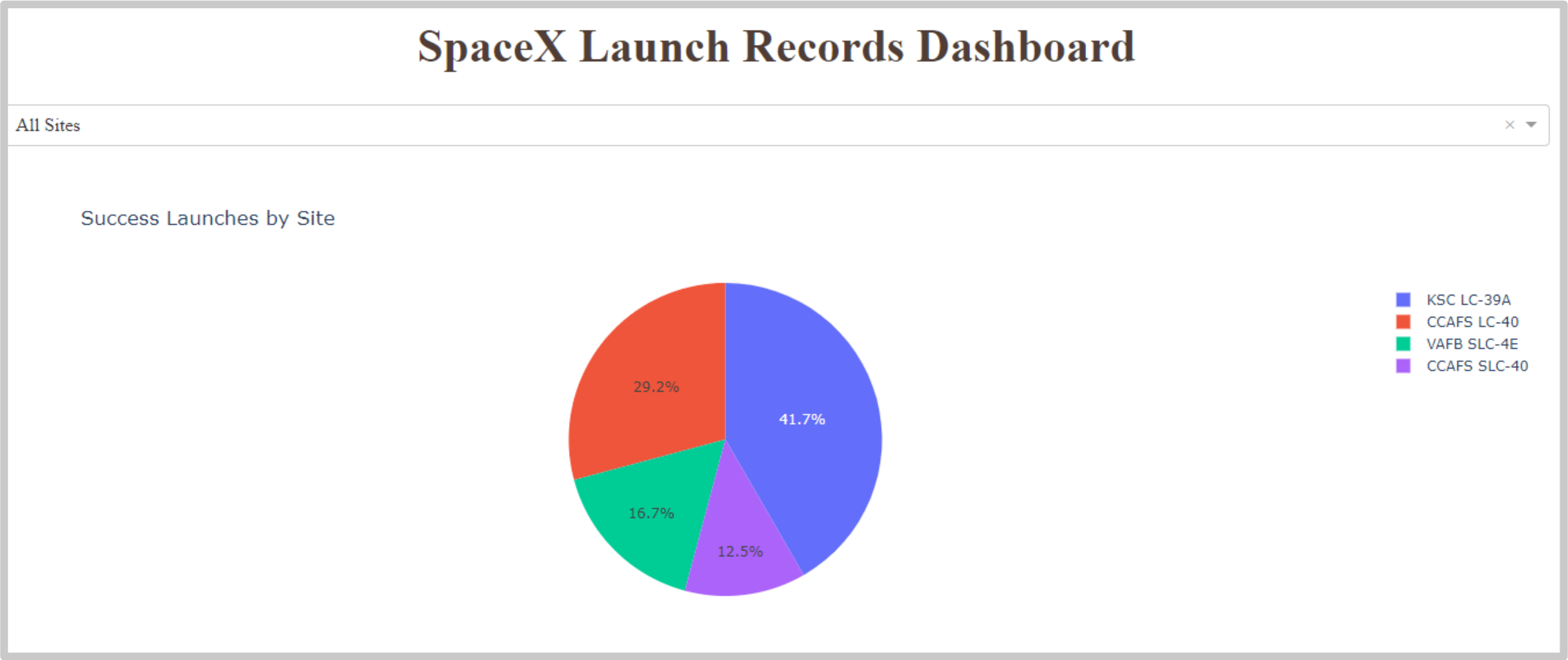


Section 5

Build a Dashboard with Plotly Dash

Proportion of Successful Launches by Site Dashboard

The highest number of successful launches was for KSC LC-39A followed by CCAFS LC-40.

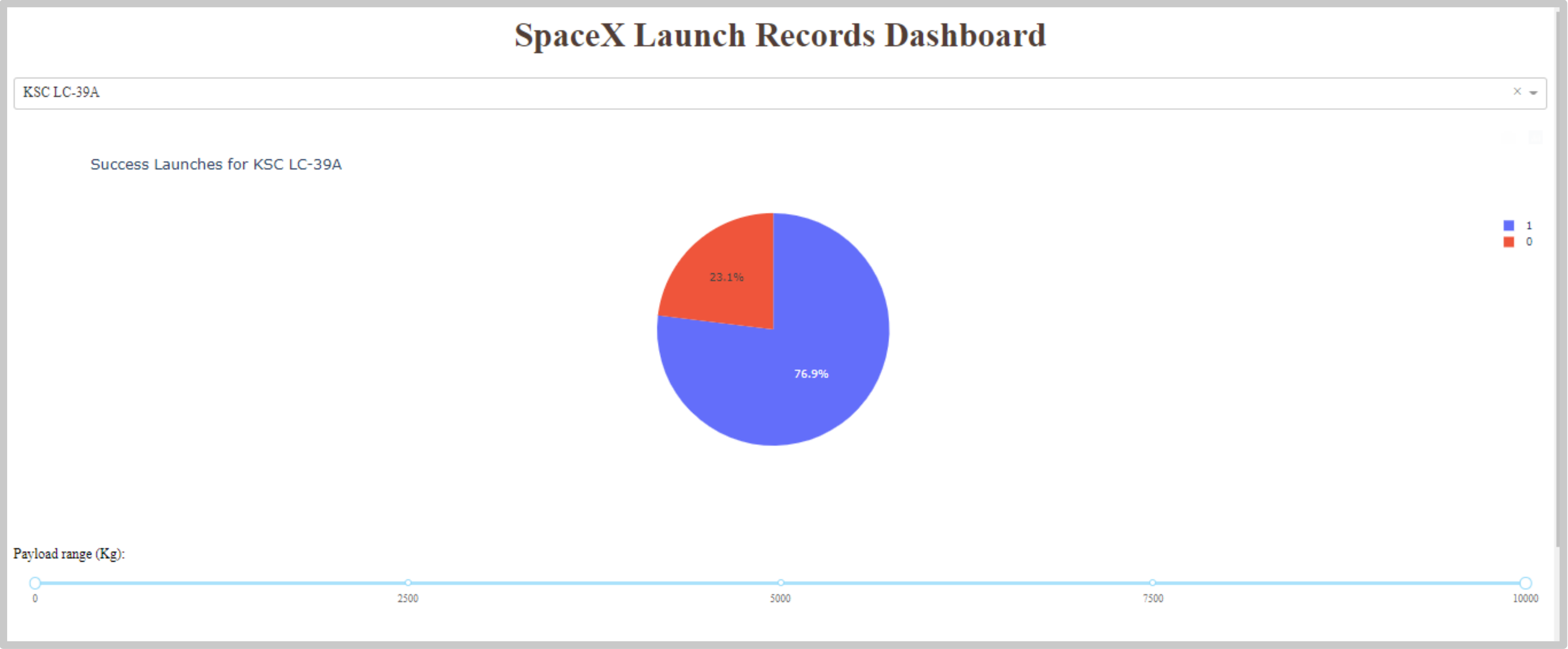


SpaceX Launch Records Dashboard

All Sites

Launch Site with Highest Success Rate

The launch site KSC LC-39A has the highest success rate of 76.9%



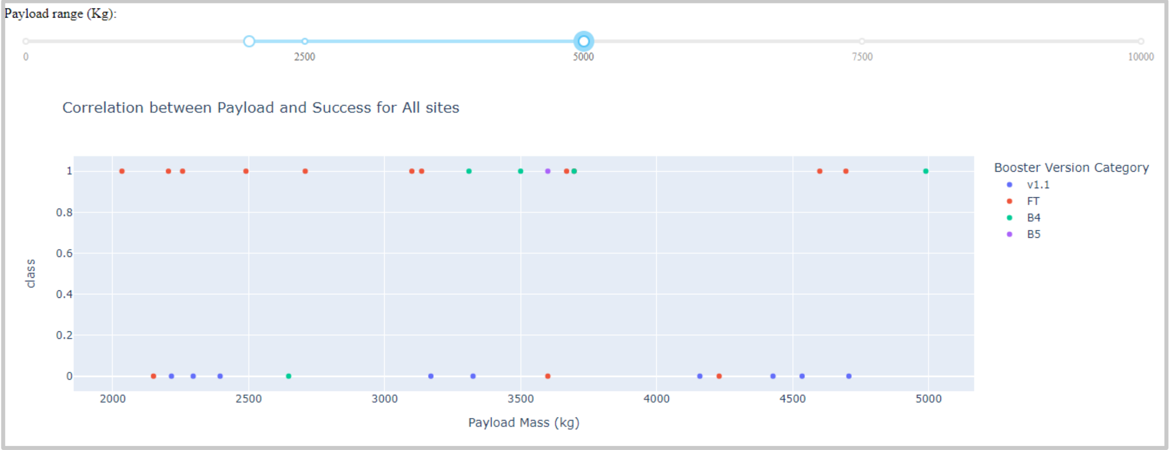
SpaceX Launch Records Dashboard

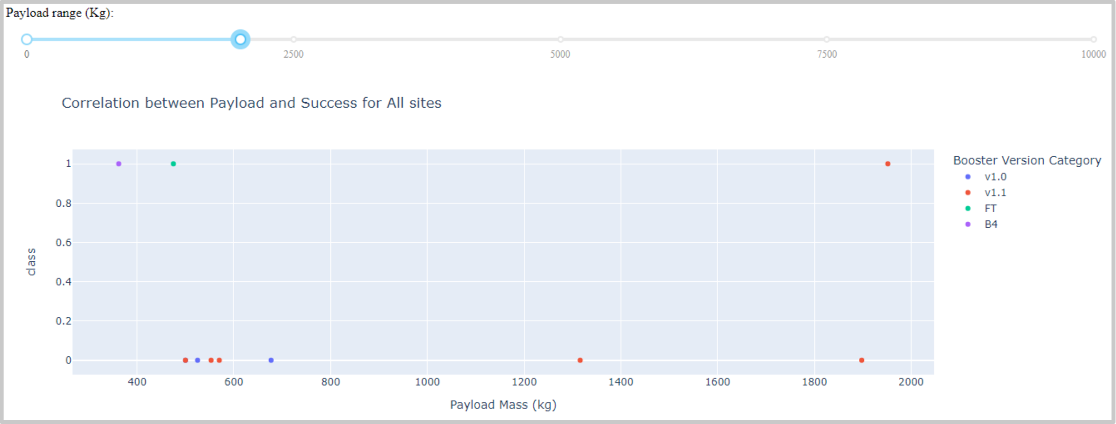
KSC LC-39A

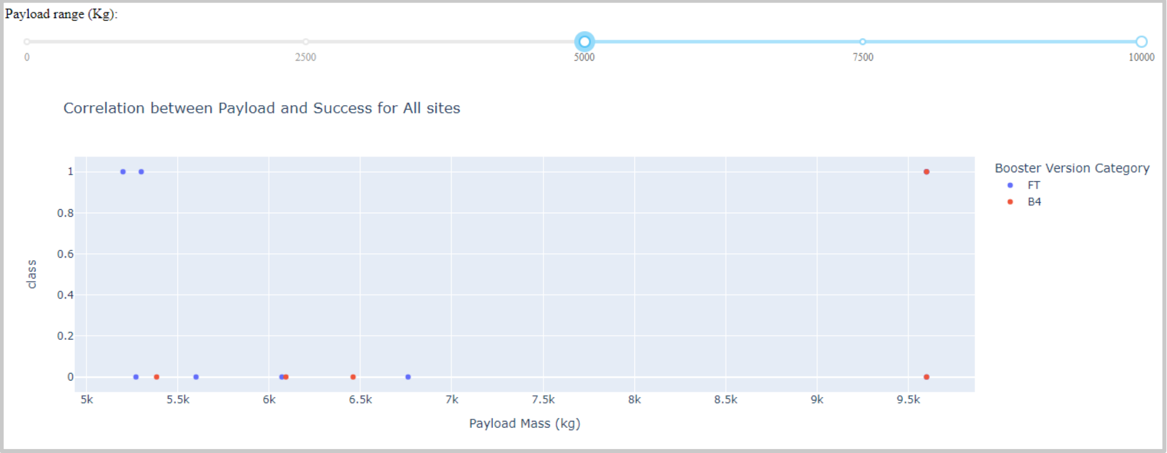
Payload range (Kg):

Payload vs Launch Outcome Dashboard

- The following scatter plots show the success rate for different payload ranges.
- The highest success rate was for the range of 2000-4000 kg
- The payload range of 6000 kg and above has the lowest success rate.
- The booster versions B5 and FT have the highest success rate.





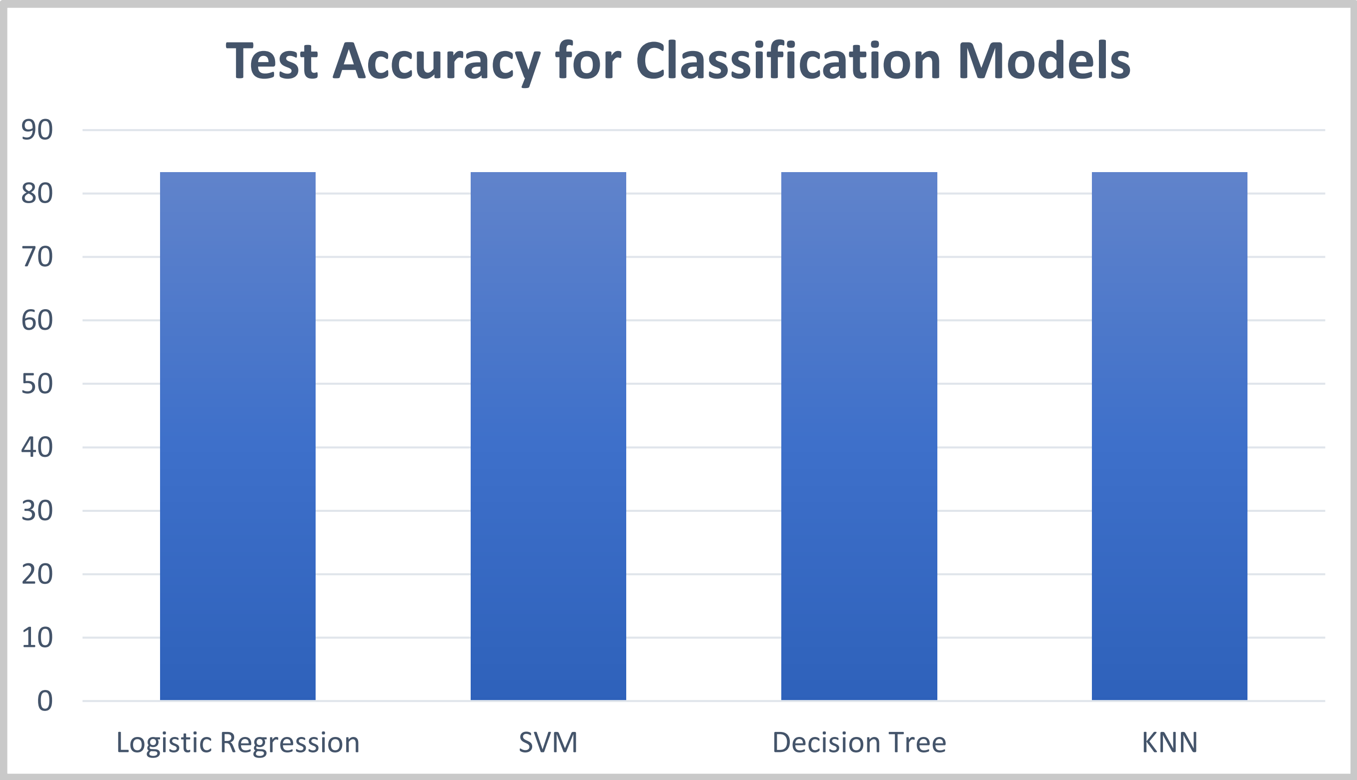


Section 6

Predictive Analysis (Classification)

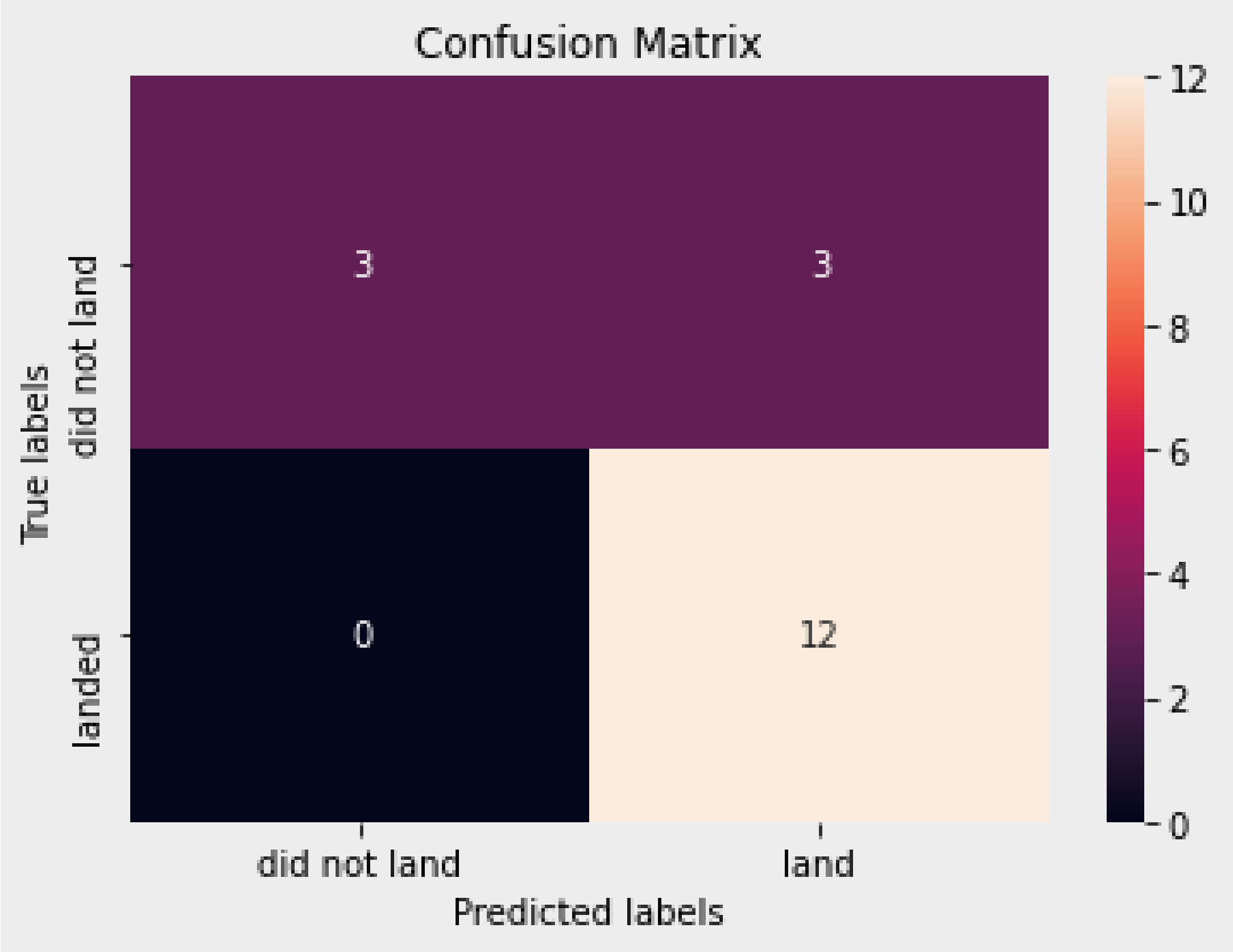
Classification Accuracy

All the models had the same accuracy of 83.33% using the .score() method on the test data.



Confusion Matrix

- Each model’s results created an identical confusion matrix.
- There was a very high accuracy for True label – landed.
- However, the accuracy for True label – did no land was not accurate as there were false positives.



Conclusions

- The aim was to be able to predict the landing outcome of a Falcon 9 launch. We built different types of classification models to predict the outcome.
- All the models built (KNN, SVM, Decision Tree and Logistic Regression) had an accuracy of 83.3 %. They were excellent in predicting for True label – landed as there were no false negatives. However, the accuracy could be improved for True label – not landed as there were some false positives.
- The success rate of Falcon 9 has continuously increased since 2013 and has a high probability of landing.
- Through visualizing the data, we found the following insight:
  - The Site KSC LC-39A has the highest success rate.
  - The Payload Mass Range 2000-4000 kg has the highest success rate compared to higher and lower ranges.
  - The orbit types ES-L1, GEO, HEO and SSO have perfect success rates.
  - Booster versions B5 and FT had the highest success rates.
- The built classification models had high accuracy in predicting, though they need improvement. The probability of landing increased considerably for certain launch sites, payload mass ranges and booster versions.

Thank you!