IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andre Victor
2022-06-10

Outline

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

Executive Summary

- Data science methodology applied:
  - Data ingestion from 2 public data sources related to Space X data
  - Data wrangling to summarize and label data
  - Exploratory data analysis (EDA) to obtain insights about success landing
  - Predictive analysis to preview the success or failure of landing
- Summary of results
  - Insights obtained from EDA allowed us to identify the best features to predict success of landing
  - A classification model was generated with good accuracy (above 80%)

Introduction

- Main goal
  - To evaluate if the new company, Space Y, can compete with Space X in a sense to successfully to landing first stage rockets and to preview correctly their launch costs.

- Questions to be answered:
  - Which features have more influence on successful landing
  - Which features are determinant to the launch cost

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - 2 public data sources were used
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - SpaceX launching data were summarized and featured to label a new feature Outcome which represents the success or failure of Falcon 9 landing.

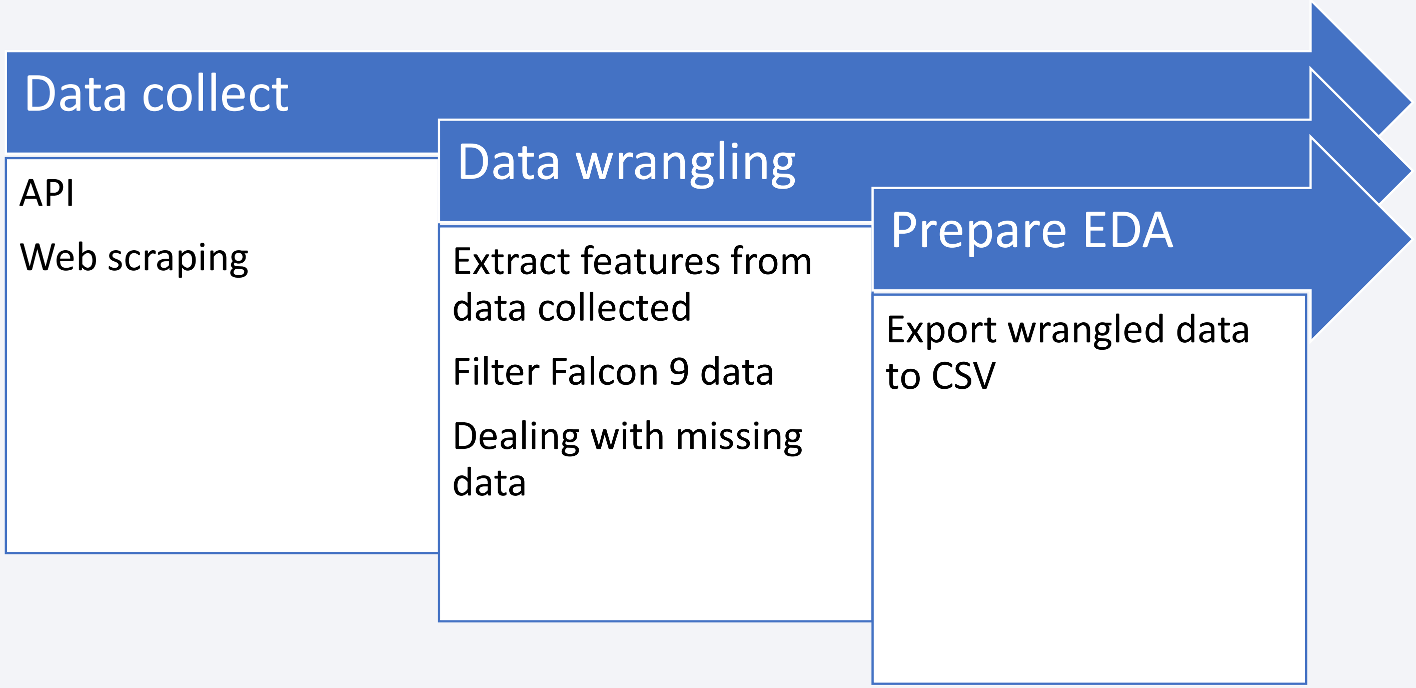
Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different classification models were tested to preview the success or failure of rocket landing, based on SpaceX data features
    - KNN, Tree Decision, Logistic regression and SVM were tested
    - They presented good performance in both training and tests datasets (accuracy above 80%)

Data Collection

- Datasets were collected using Space X API (https://api.spacexdata.com/v4/rockets/rockets/) and web scraping techniques (BeautifulSoup library).



Data collect

API

Web scraping

Data wrangling

Extract features from data collected

Filter Falcon 9 data

Dealing with missing data

Prepare EDA

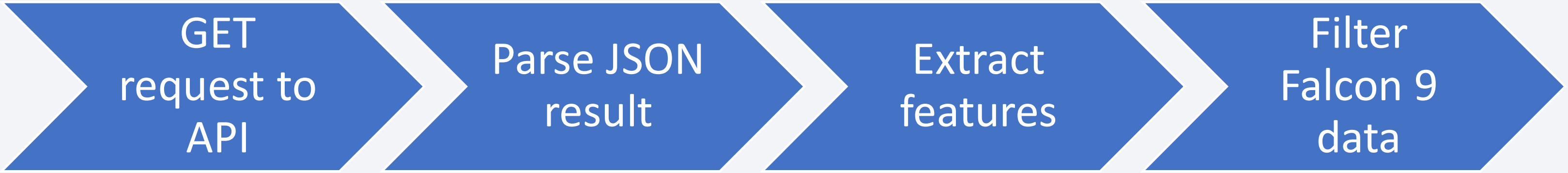
Export wrangled data to CSV

Data Collection – SpaceX API

- Using **request** library, we call a GET request to API to collect data.
- **JSON result** is parsed and features are extracted and converted to a dataframe
- Falcon 9 data is filtered

- More details:

https://github.com/ormastroni/IBM-Course/blob/master/Data%20Collection%20API%20(C10W1).ipynb



GET request to API

Parse JSON result

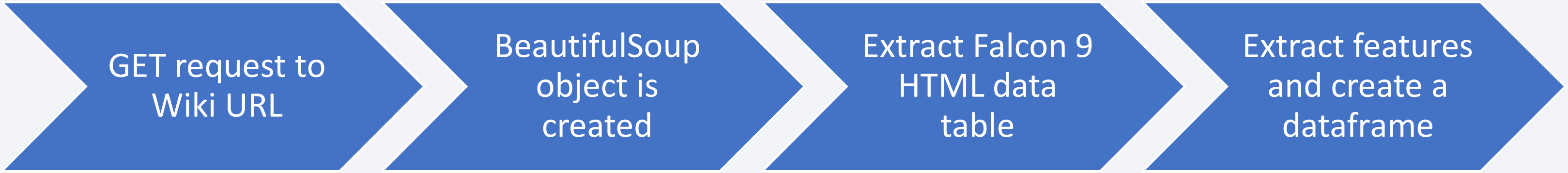
Extract features

Filter Falcon 9 data

Data Collection - Scraping

- Using **request** library, we call a GET request to Space X Wiki site
- Using **BeautifulSoup** library, we convert response text into a BeautifulSoup object with all **HTML structure**
- We **extract Falcon 9 features** and convert them into a **dataframe**
- More details:

https://github.com/ormastroni/IBM-Course/blob/master/Data%20Collection%20with%20Web%20Scrapping%20(C10W1).ipynb



GET request to Wiki URL

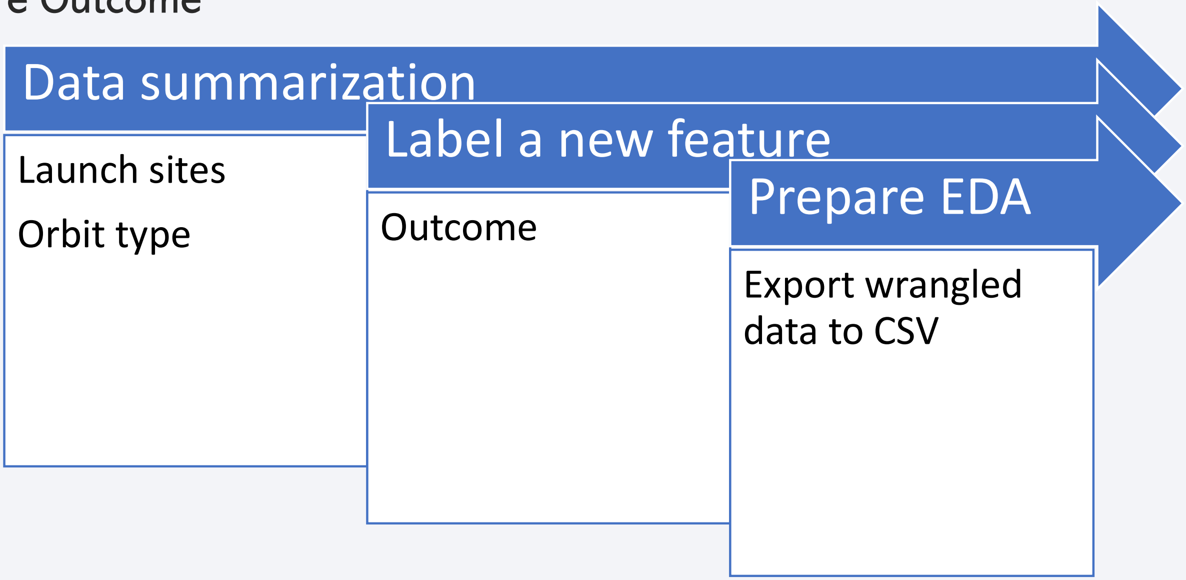
BeautifulSoup object is created

Extract Falcon 9 HTML data table

Extract features and create a dataframe

Data Wrangling

- Dealing with missing data
- Exploratory Data Analysis: summarization
  - Number of launches per site and orbit type
  - Label a new feature Outcome



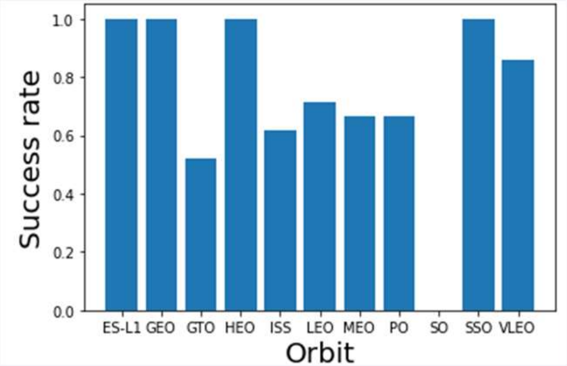
More details in

https://github.com/ormastroni/IBM-Course/blob/master/Data%20wrangling%20(C10W1).ipynb

EDA with Data Visualization

- Scatterplots were used in order to understand relationship between each pair of numerical features, showing their respective classes (success or failure)
    - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

- Barplots were used to evaluate the influence of Orbit feature in the success rate



More details in:

https://github.com/ormastroni/IBM-Course/blob/master/EDA%20with%20Data%20Visualization(C10W2).ipynb

EDA with SQL

- SQL queries executed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

More details in https://github.com/ormastroni/IBM-Course/blob/master/EDA%20with%20SQL%20(C10W2).ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers show launch sites;
  - Circles show highlighted areas around specific coordinates (e.g. NASA Johnson Space Center)
  - Marker clusters show different launch positions in a same launch site
  - Lines show distances between two coordinates.

More details in:

https://github.com/ormastroni/IBM-Course/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20(C10W3).ipynb

Build a Dashboard with Plotly Dash

- A dashboard was developed to visualize:
  - Percentage of launches by site
  - Payload range

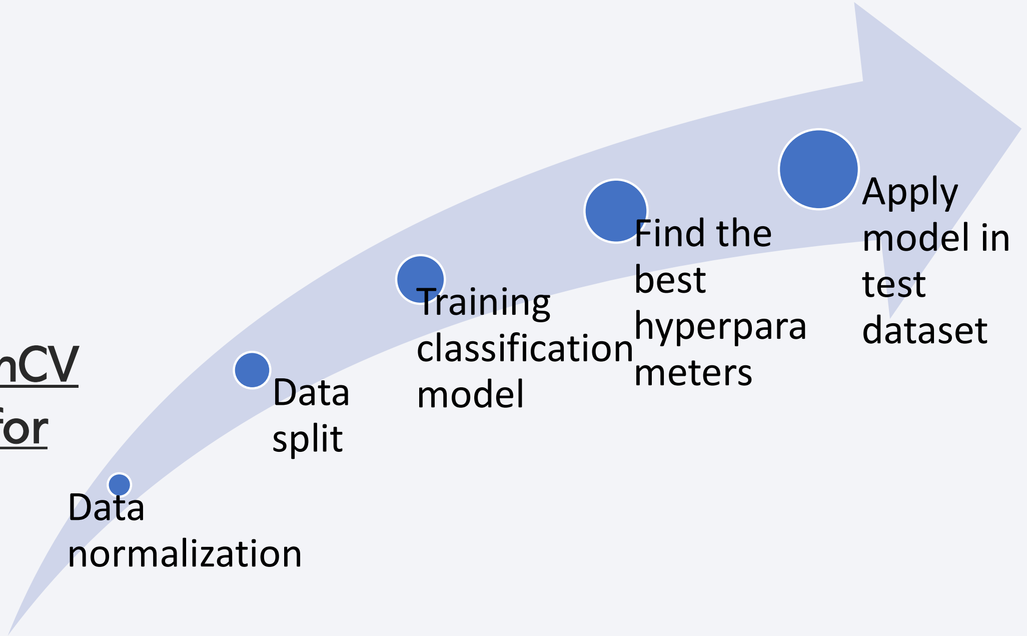
- Both graphs are suitable to show the relation between payloads and launch sites
  - It is possible to detect what are the best places to launch according to payload range

More details in:

https://github.com/ormastroni/IBM-Course/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Data are normalized and divided in 2 datasets: training and test sets
- 4 different classifiers were tested:
  - Logistic regression, Tree Decision, SVM and KNN
- Using sklearn.model_selection library, GridSearchCV was used to find out the best hyperparameters for each classifier in the training dataset
- Performance results for each classifier were compared in the test dataset



Data normalization

Data split

Training classification model

Find the best hyperparameters

Apply model in test dataset

More details in

https://github.com/ormastroni/IBM-Course/blob/master/Machine%20Learning%20Prediction%20(C10W4).ipynb

Results

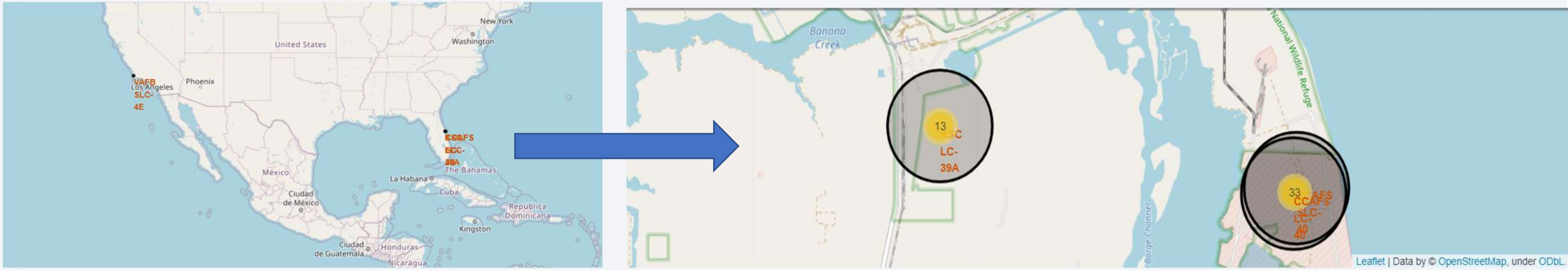
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - Falcon 9 booster versions in drone ships tend to have success at landing
  - High successful rates of mission outcomes (almost 100%)
  - Increasing rates of landing outcomes over time

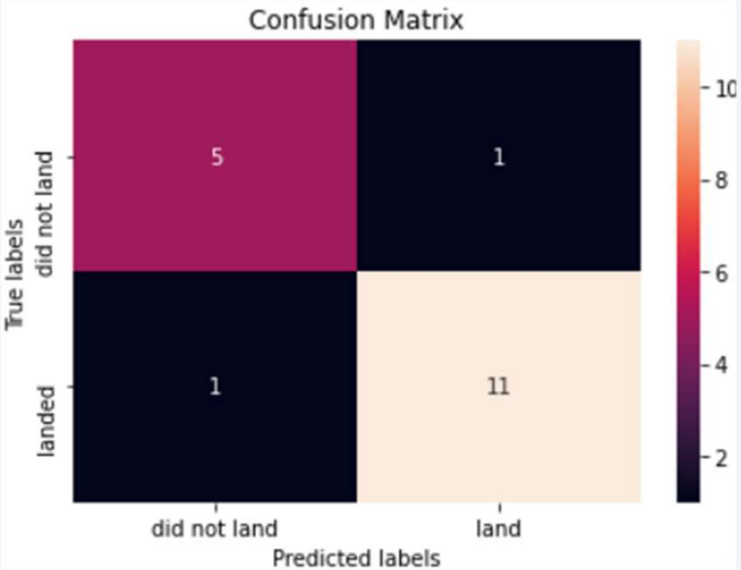
Results

- Exploratory data analysis using geographic data
  - Launch sites are kept in safe places and near coast
  - Most launches happens at east sites (46 launches)



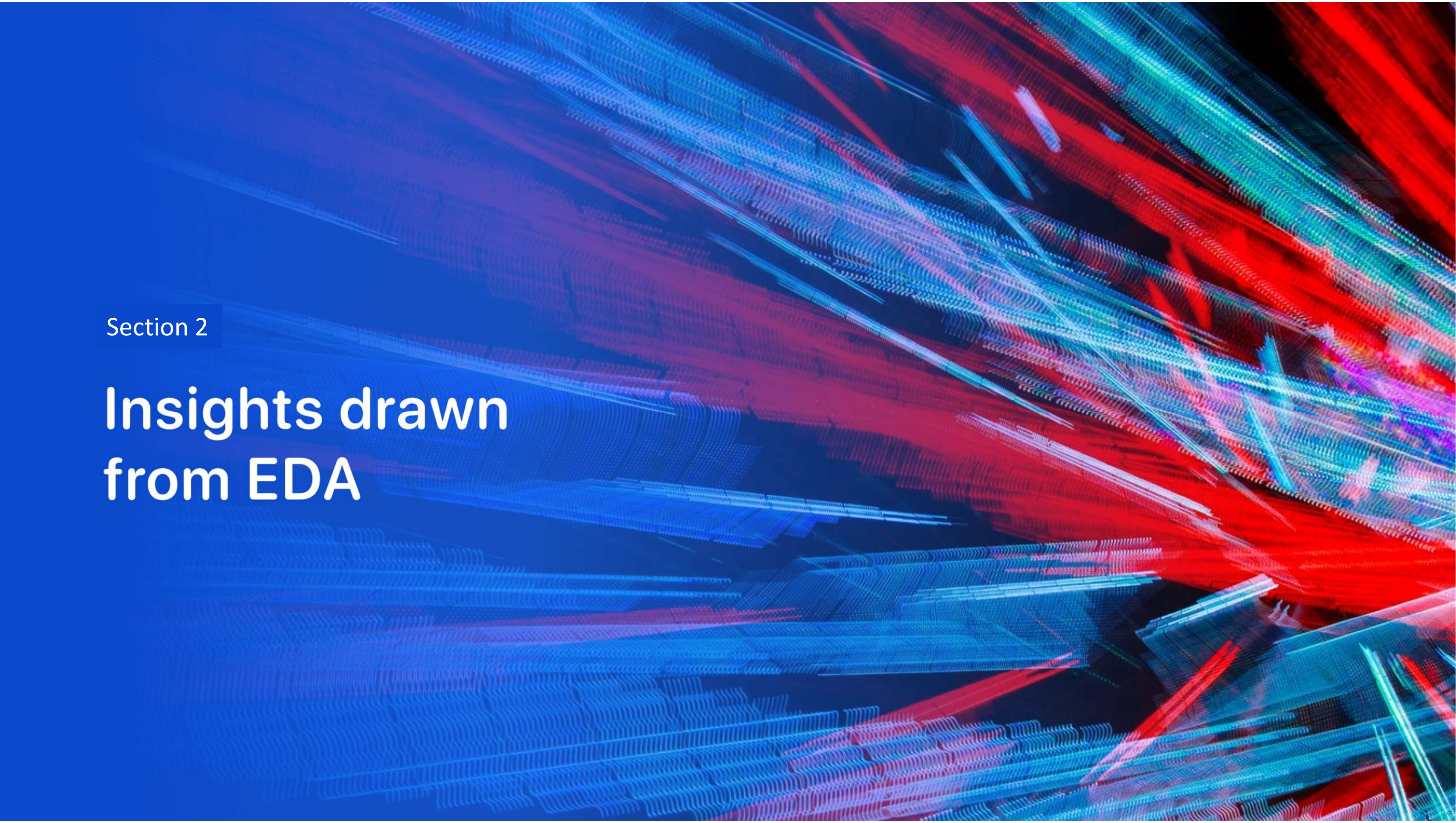
Results

- Predictive Analysis
  - Decision Tree was the classifier with the best accuracy (over 85%) on testing dataset
    - Only 1 false positive and 1 false negative

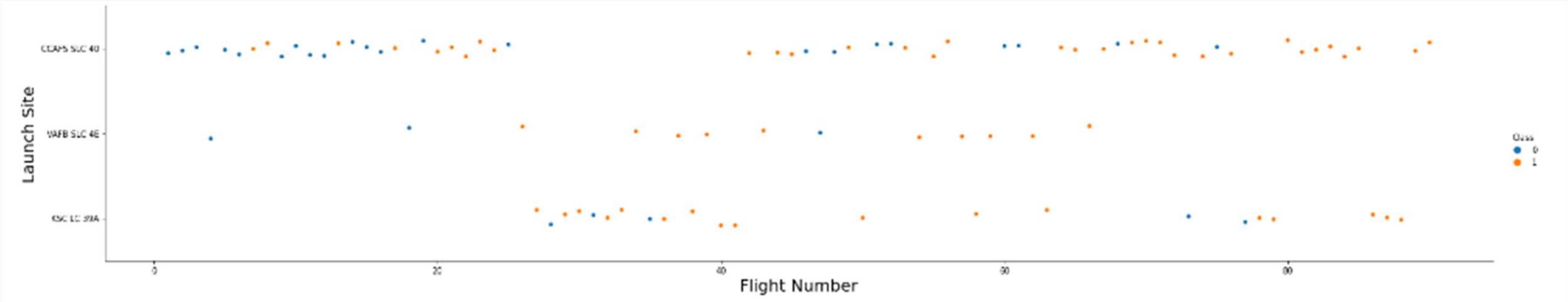


Section 2

Insights drawn from EDA



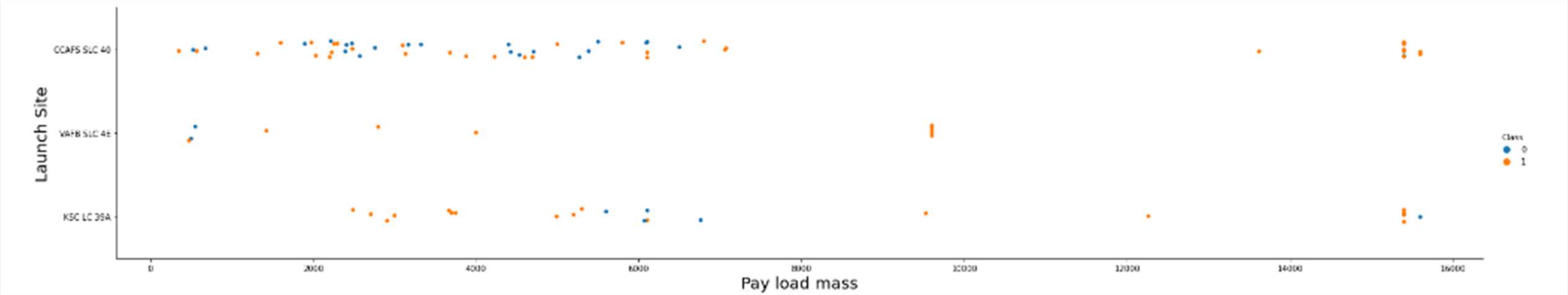
Flight Number vs. Launch Site



- The best launch site is CCAF5 SLC 40, where most of recent launches were successful;

- Success rate improves over time for all launch sites

Payload vs. Launch Site

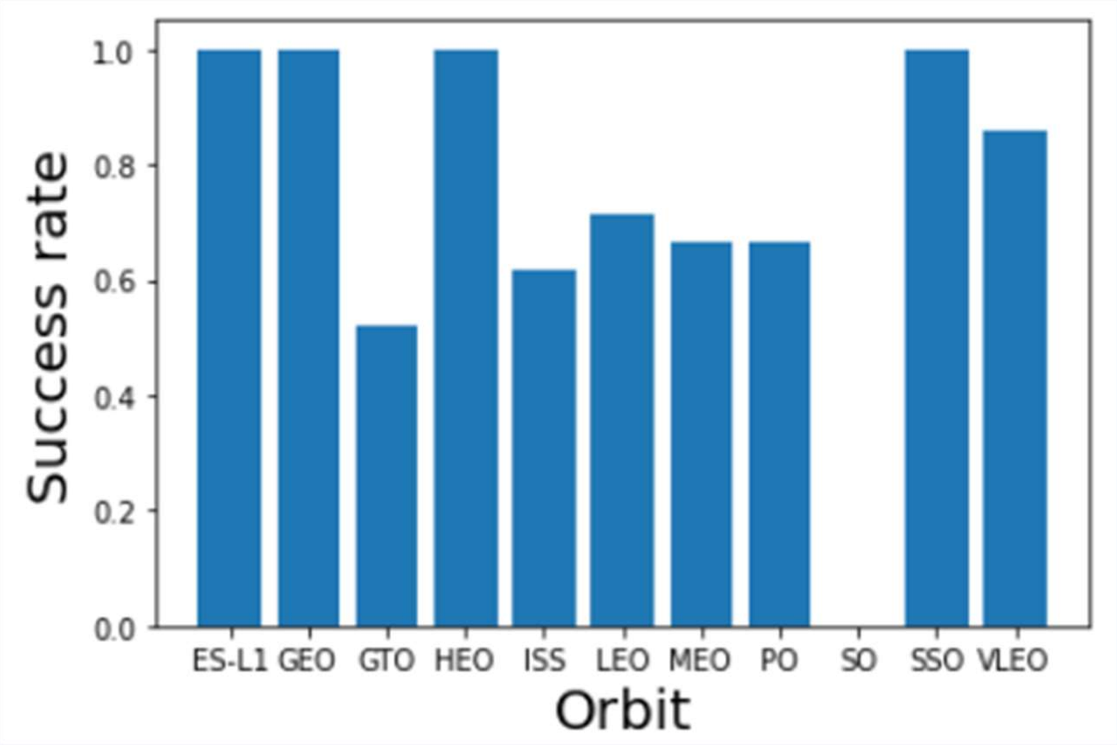


- Payloads over 9,000kg have high success rate;

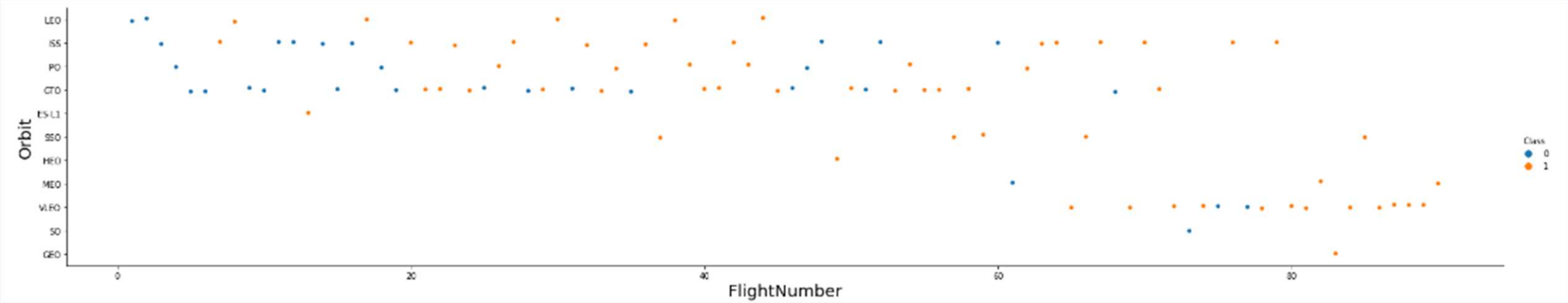
- There is no attempt on VAFB SLC 4E launch site for payload over 9,000kg

Success Rate vs. Orbit Type

- 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Other orbits with high success rate:
  - VLEO (over 80%)
  - LEO (over 70%)



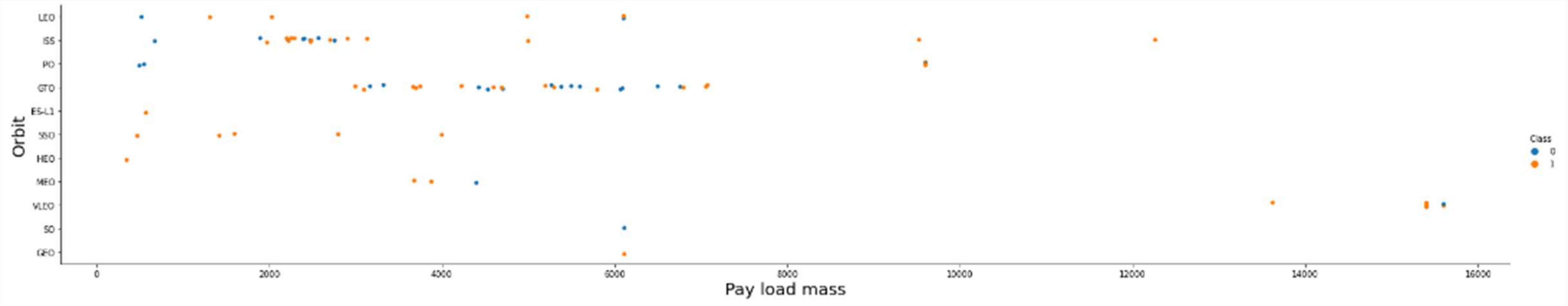
Flight Number vs. Orbit Type



- Success rate improved over time for all orbits

- VLEO orbit has been used more recently.

Payload vs. Orbit Type

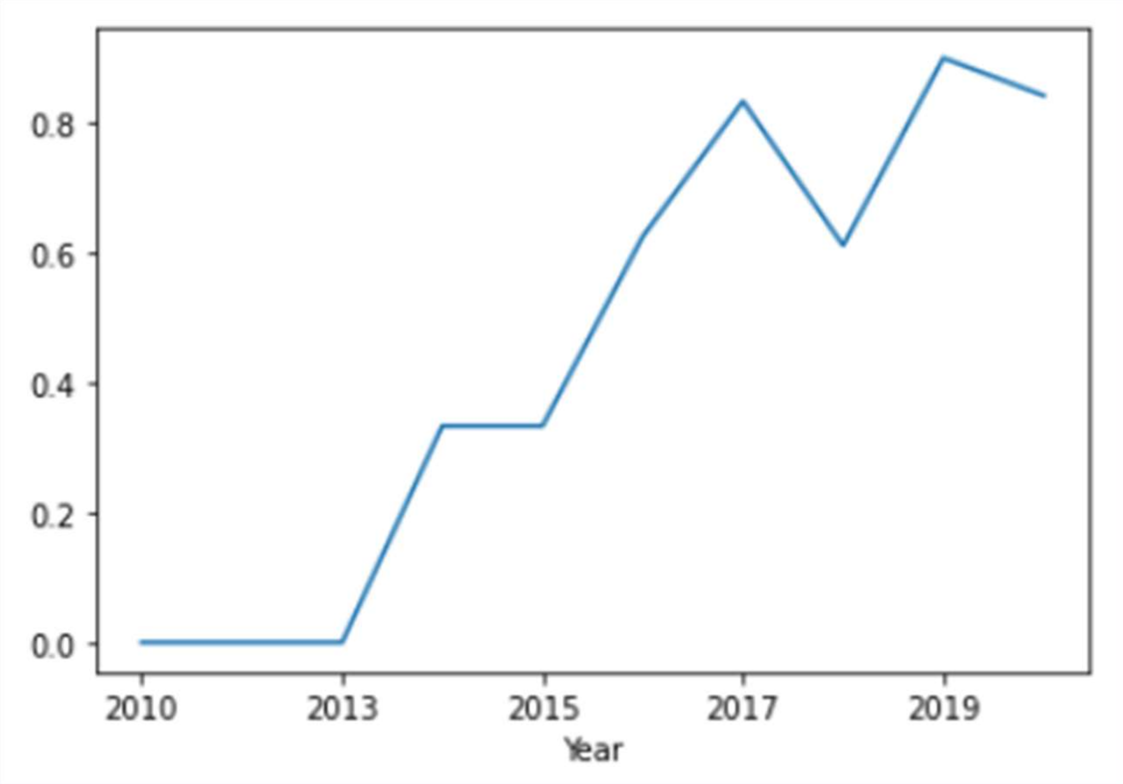


- There seems to be relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few attempts to the orbits SO and GEO.

Launch Success Yearly Trend

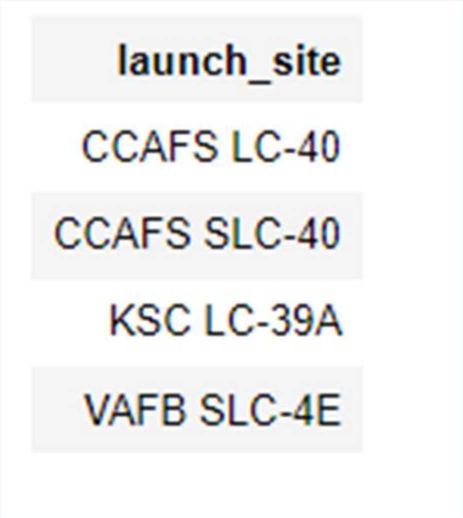
- Success rate since 2013 kept increasing till 2020

- Apparently, the period between 2010 and 2013 was used to develop and improve the technology, since there is no success.



All Launch Site Names

- There are four distinct launch sites

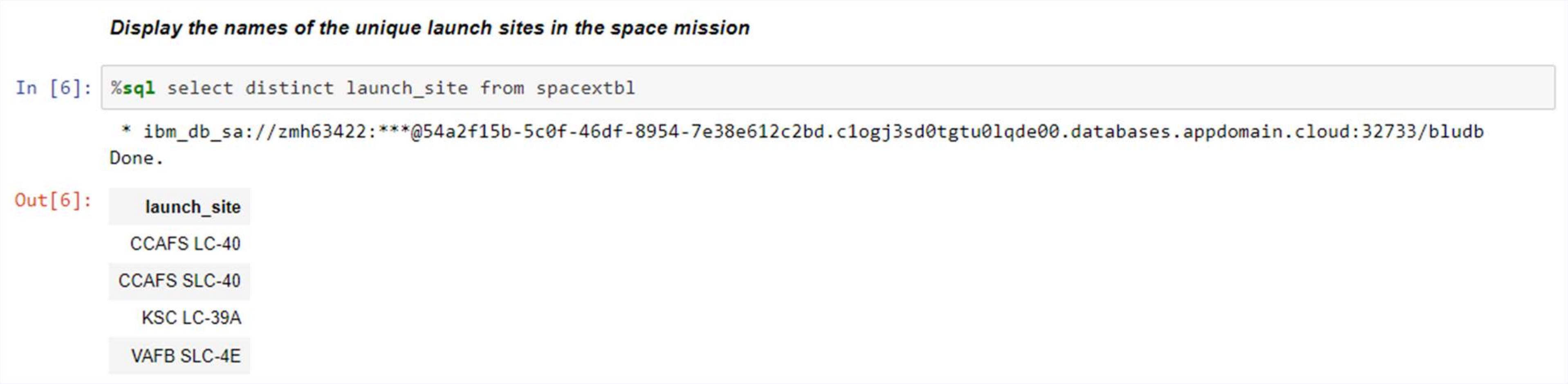


launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- They were obtained from selecting distinct record values from relational data

All Launch Site Names

- There are four distinct launch sites



Display the names of the unique launch sites in the space mission

```
In [6]: %sql select distinct launch_site from spacextbl
```

```
 * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

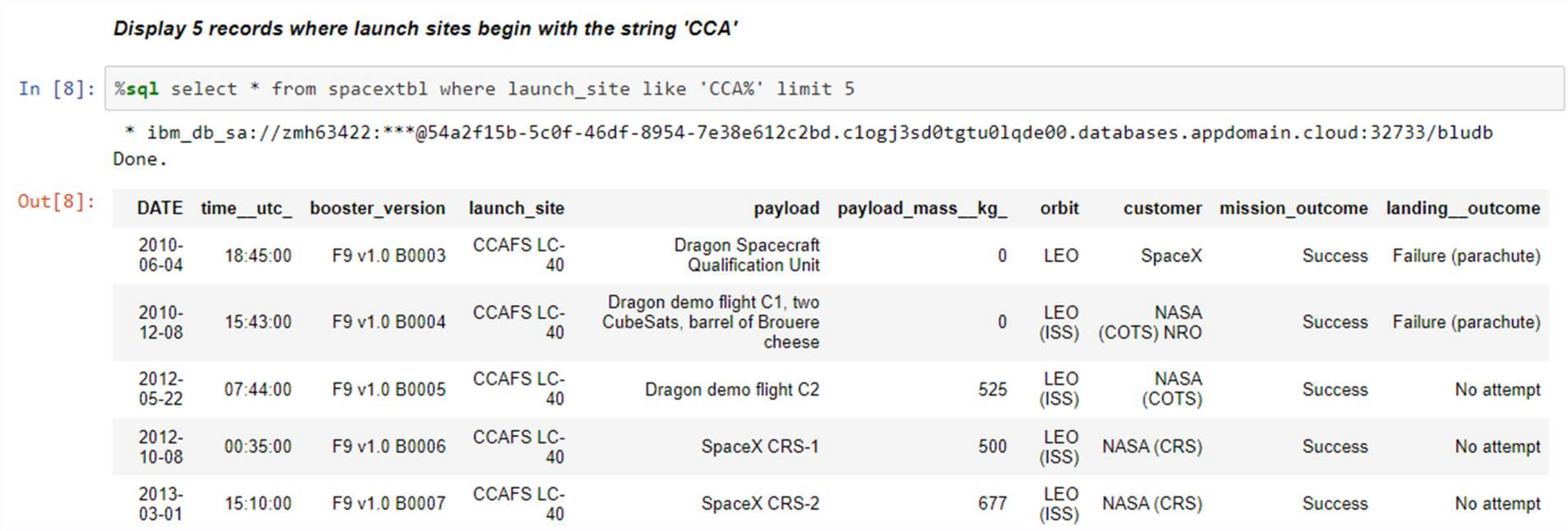
Out[6]:

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

- They were obtained from selecting distinct record values from relational data

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:



Display 5 records where launch sites begin with the string 'CCA'

In [8]:

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

```
 * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

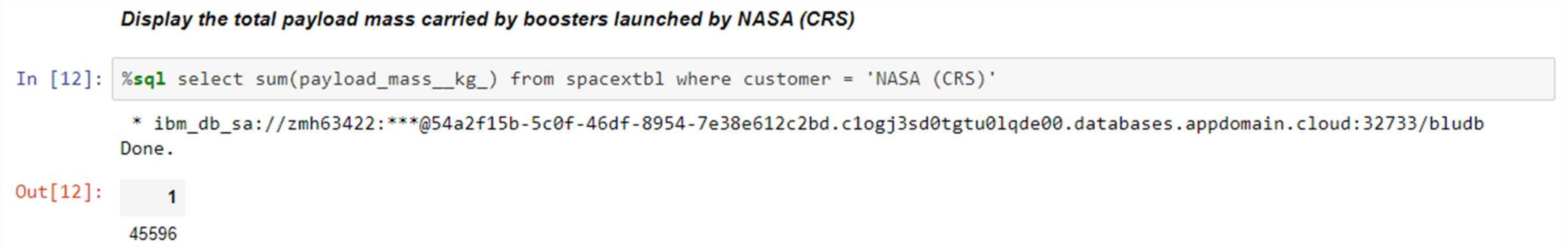
Out[8]:

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

- The samples shown are from Cape Canaveral launches

Total Payload Mass

- Total payload carried by boosters from NASA



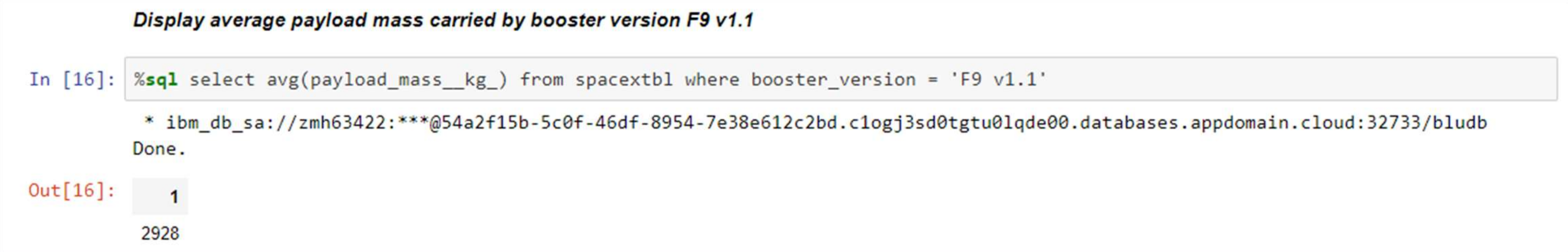
Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [12]: %sql select sum(payload_mass__kg_) from spacextbl where customer = 'NASA (CRS)'

          * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.

Out[12]:     1
         45596
```

- We used a SQL agregation function (sum) to sum all payloads whose customer is NASA (CRS)

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1



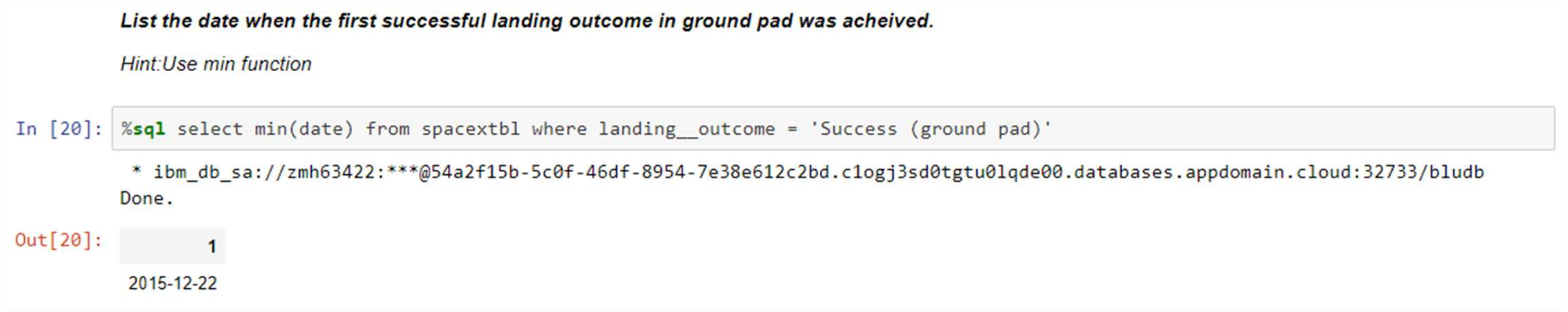
Display average payload mass carried by booster version F9 v1.1

```
In [16]: %sql select avg(payload_mass__kg_) from spacextbl where booster_version = 'F9 v1.1'

          * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.

Out[16]:    1
          2928
```

- We used a SQL agregation function (avg) to calculate the payload average whose booster version is F9 v1.1

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad



List the date when the first successful landing outcome in ground pad was acheived.

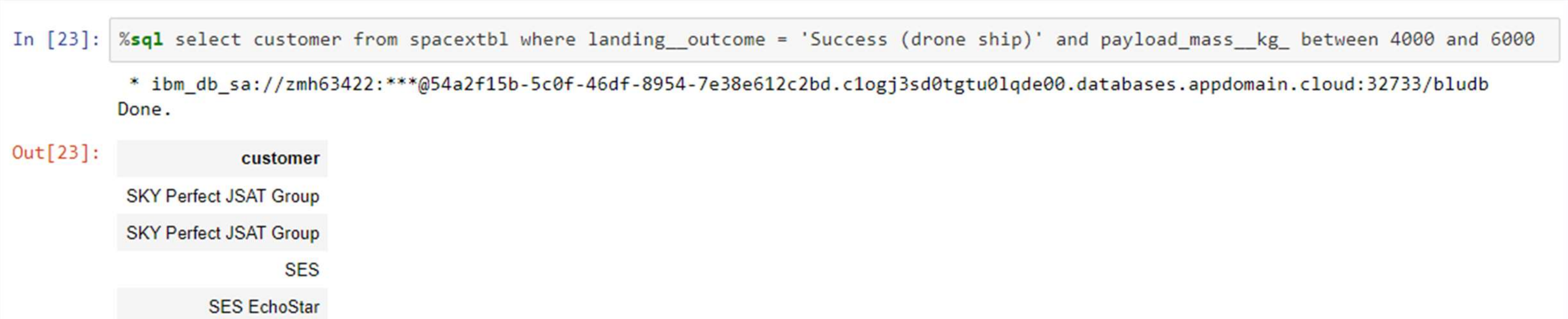
Hint:Use min function

```
In [20]: %sql select min(date) from spacextbl where landing__outcome = 'Success (ground pad)'

          * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.

Out[20]:          1
          2015-12-22
```

- As date is a column whose type is datetime, we can use the agregation function min() to get the minimum date, i.e., the lowest date where the outcome value is ‘Success (ground pad)’

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



```
In [23]: %sql select customer from spacextbl where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000

 * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

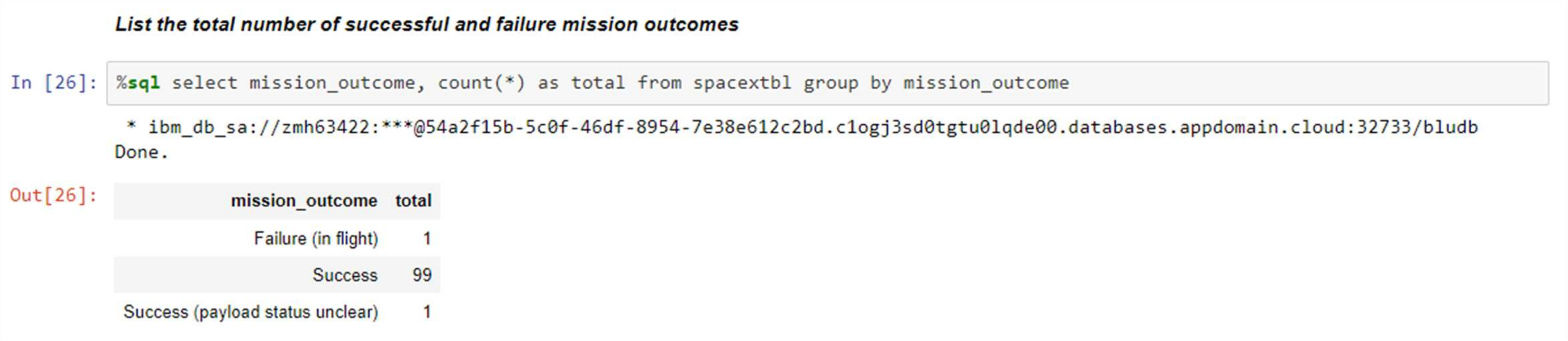
Out[23]:

customer
SKY Perfect JSAT Group
SKY Perfect JSAT Group
SES
SES EchoStar

- We applied two filters on data: Successful (drone ship) and data range. For this reason, the query has two expressions with AND clause
- In the date clause, we use the operator BETWEEN to get results in a range

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes



List the total number of successful and failure mission outcomes

```
In [26]: %sql select mission_outcome, count(*) as total from spacextbl group by mission_outcome
```

```
 * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

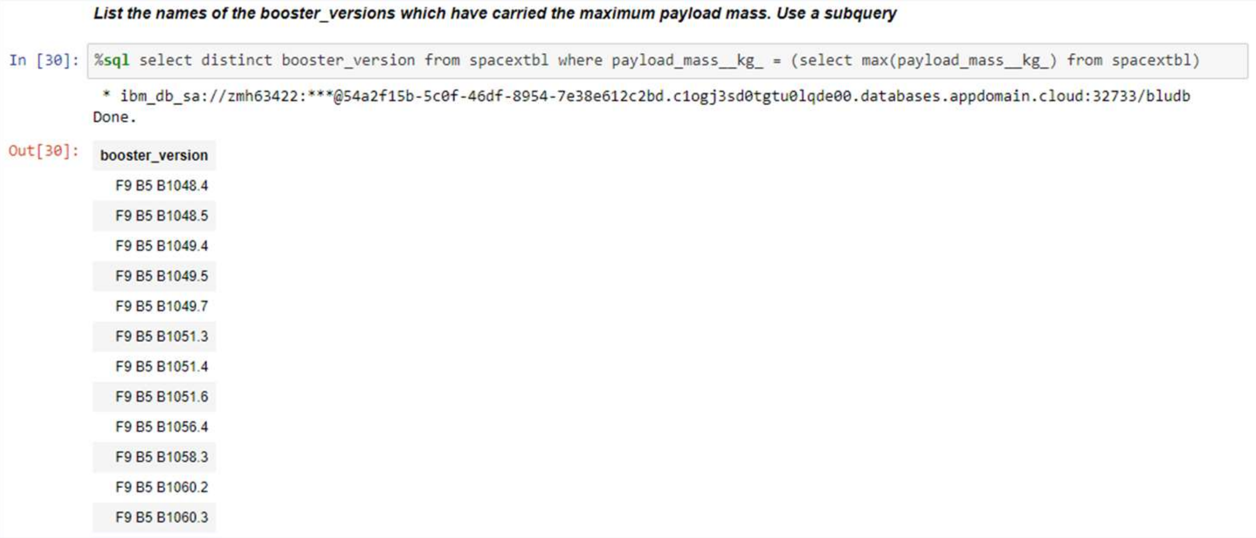
Out[26]:

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

- We group all data by mission outcomes and count all occurrences of success and failures for each group

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass



List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
In [30]: %sql select distinct booster_version from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)

 * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

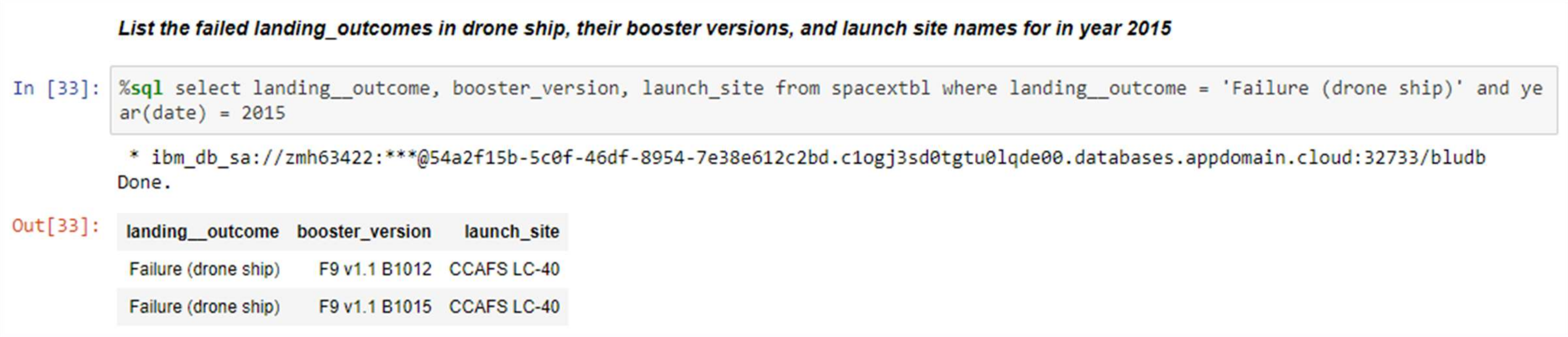
Out[30]:

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- We used a subquery to get the maximum payload mass. So, we filter all data which matches with this value

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



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

```
In [33]: %sql select landing__outcome, booster_version, launch_site from spacextbl where landing__outcome = 'Failure (drone ship)' and ye
         ar(date) = 2015
```

```
 * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

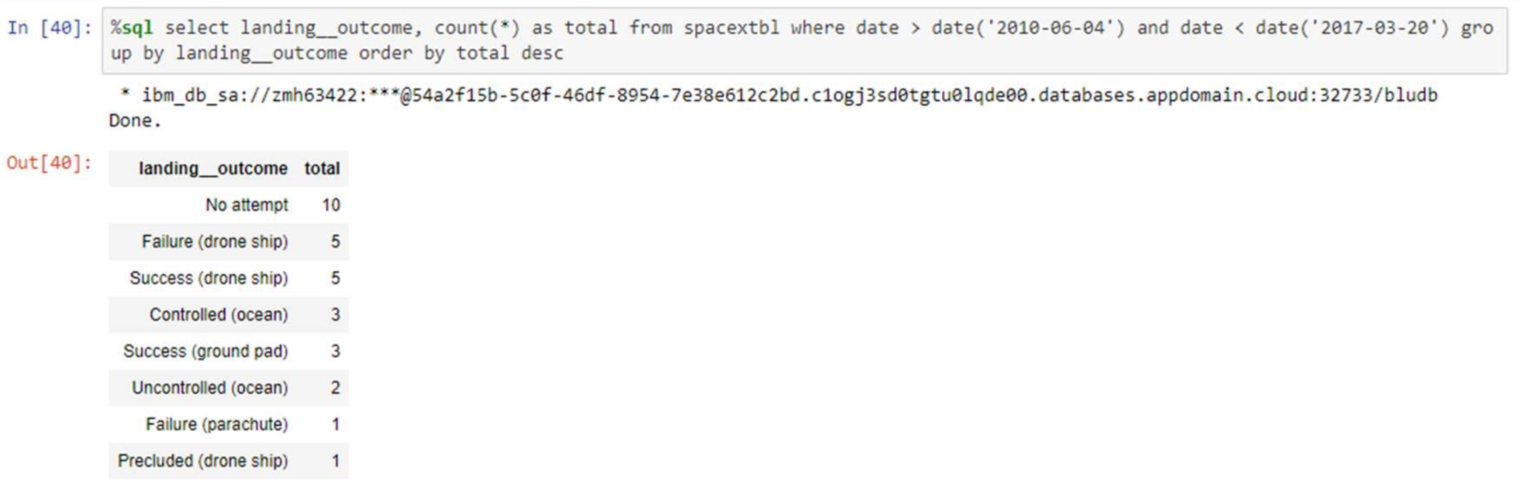
Out[33]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- We applied two filters on data: Failure (drone ship) and year 2015. For this reason, the query has two expressions with AND clause
- We used a function year to get the year part of complete datetime column

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

- Rank 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



```
In [40]: %sql select landing__outcome, count(*) as total from spacextbl where date > date('2010-06-04') and date < date('2017-03-20') gro
         up by landing__outcome order by total desc

          * ibm_db_sa://zmh63422:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.
```

Out[40]:

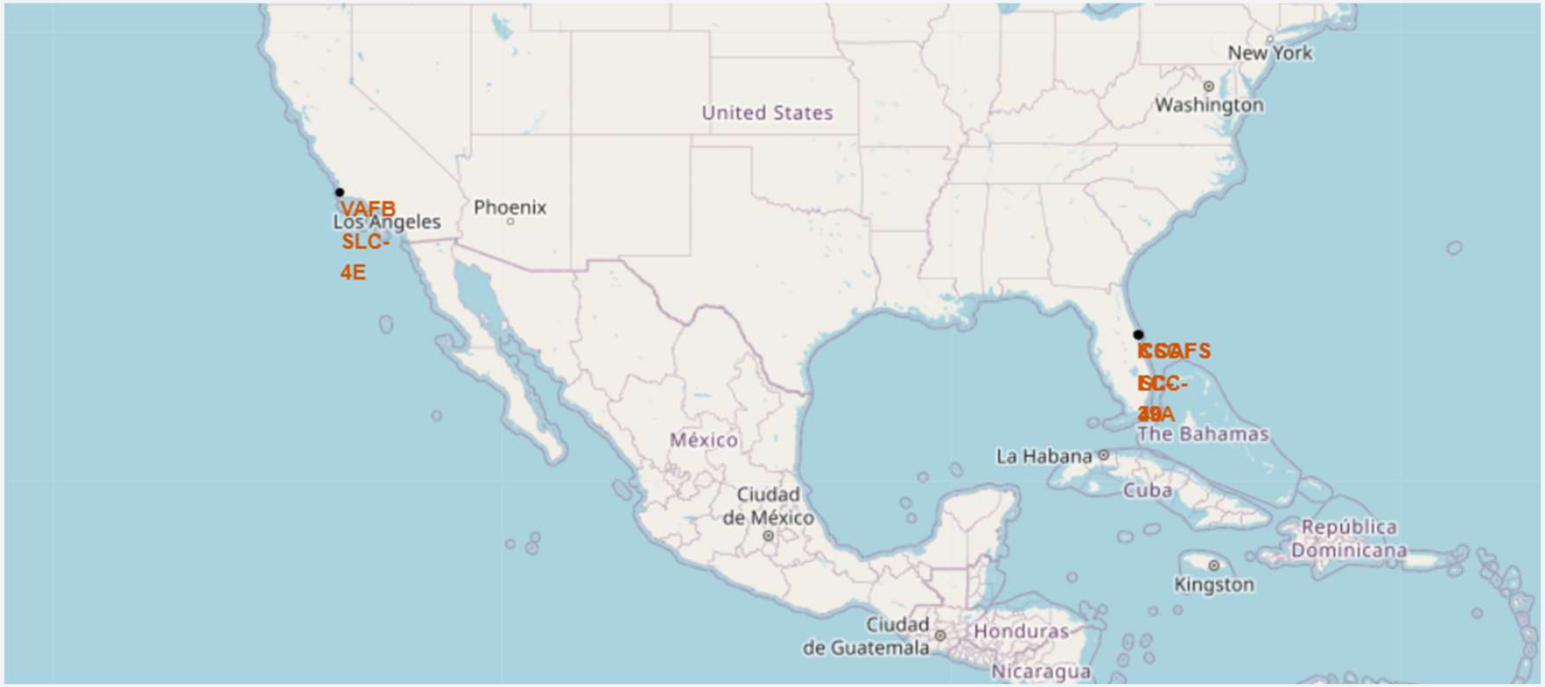
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

- We use date functions to specify date range, group the results by landing outcomes, count all occurrences for each group values, showing the results in a descending order

Section 3

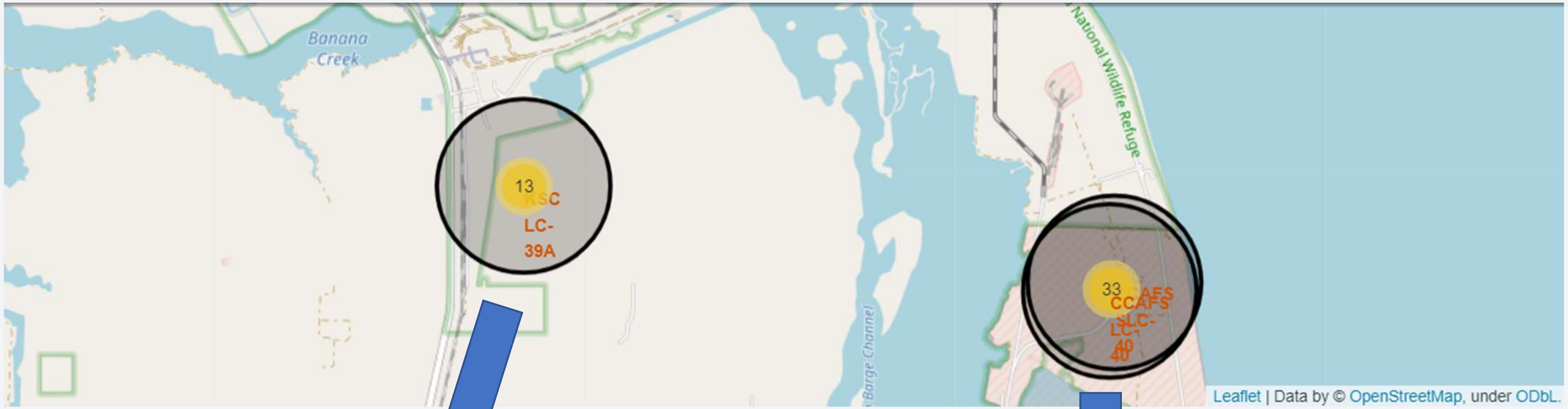
Launch Sites Proximities Analysis

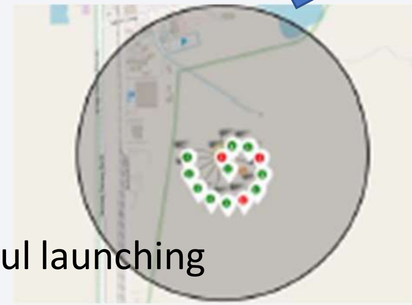
All launch sites

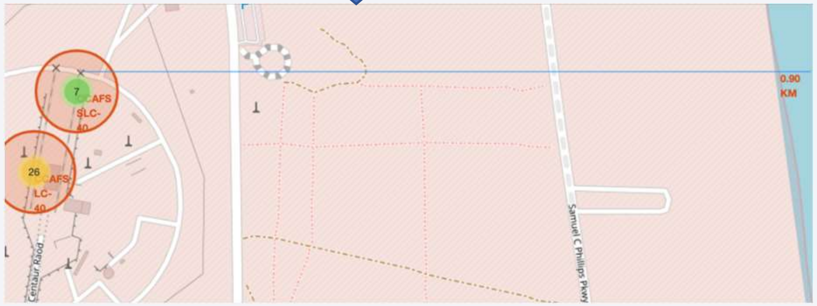


By safety, launch sites are near sea, as shown in the map

Lauch Outcomes by Site

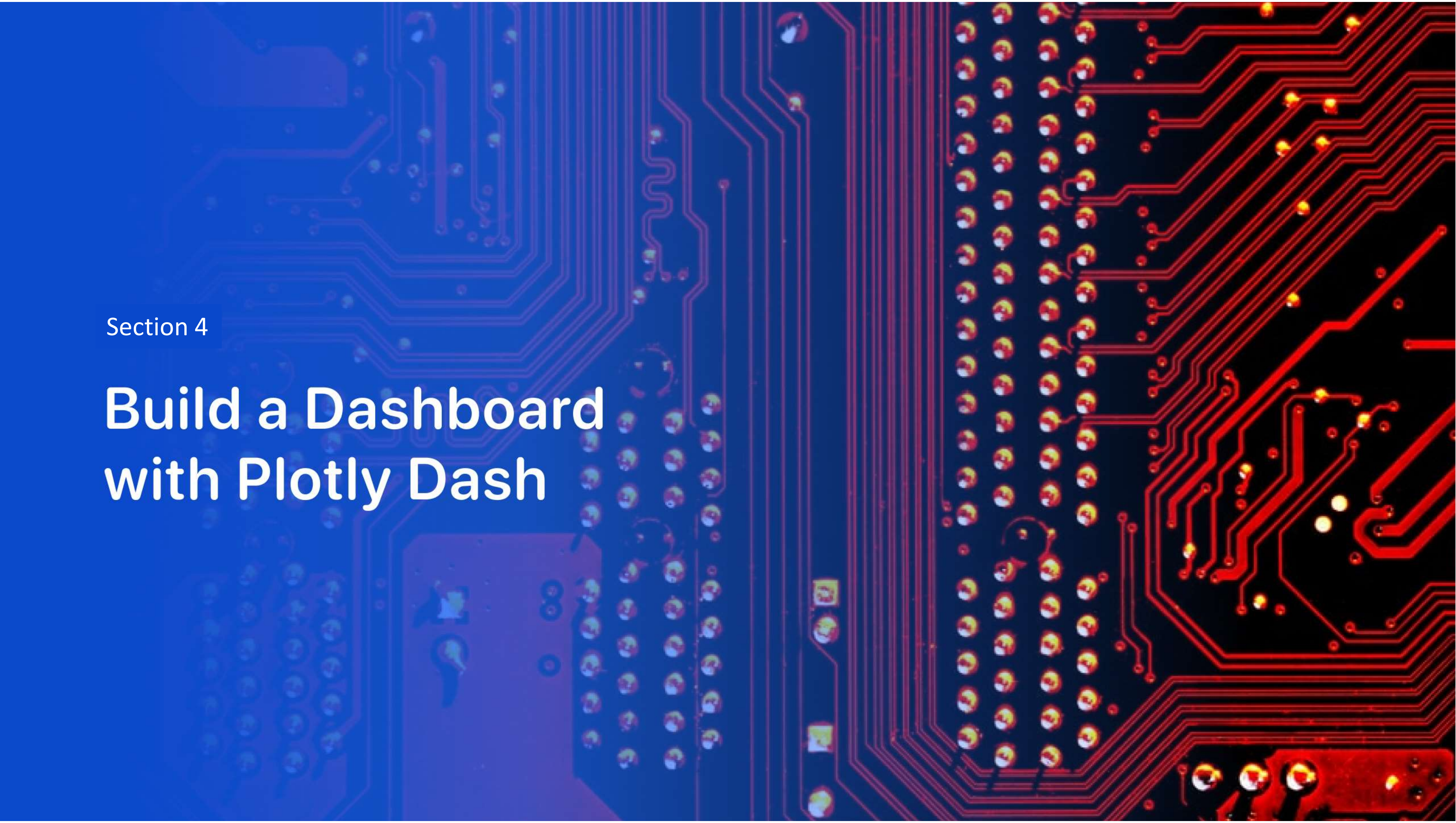






Green marks: successful launching
Red marks: failures

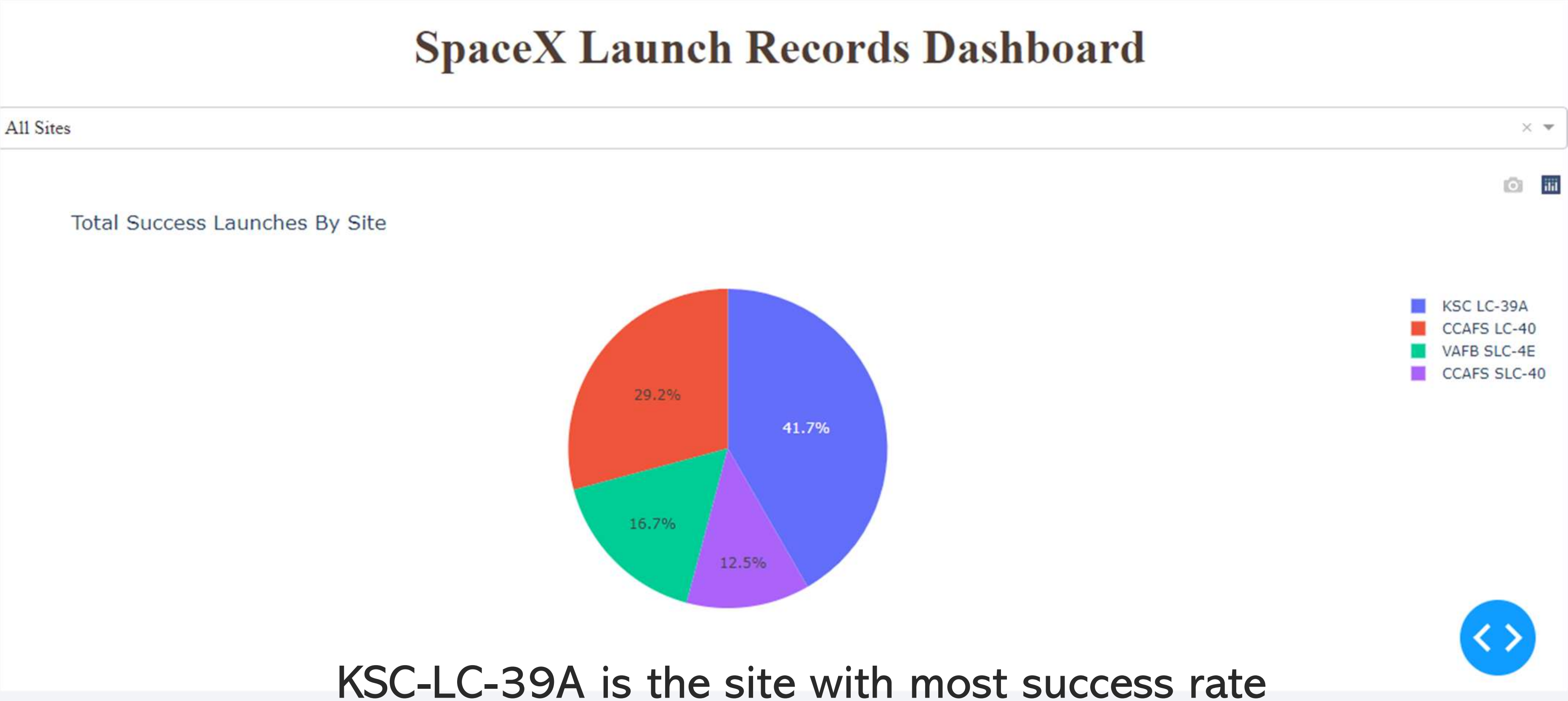
Suitable place: near roads and no habitants



Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site



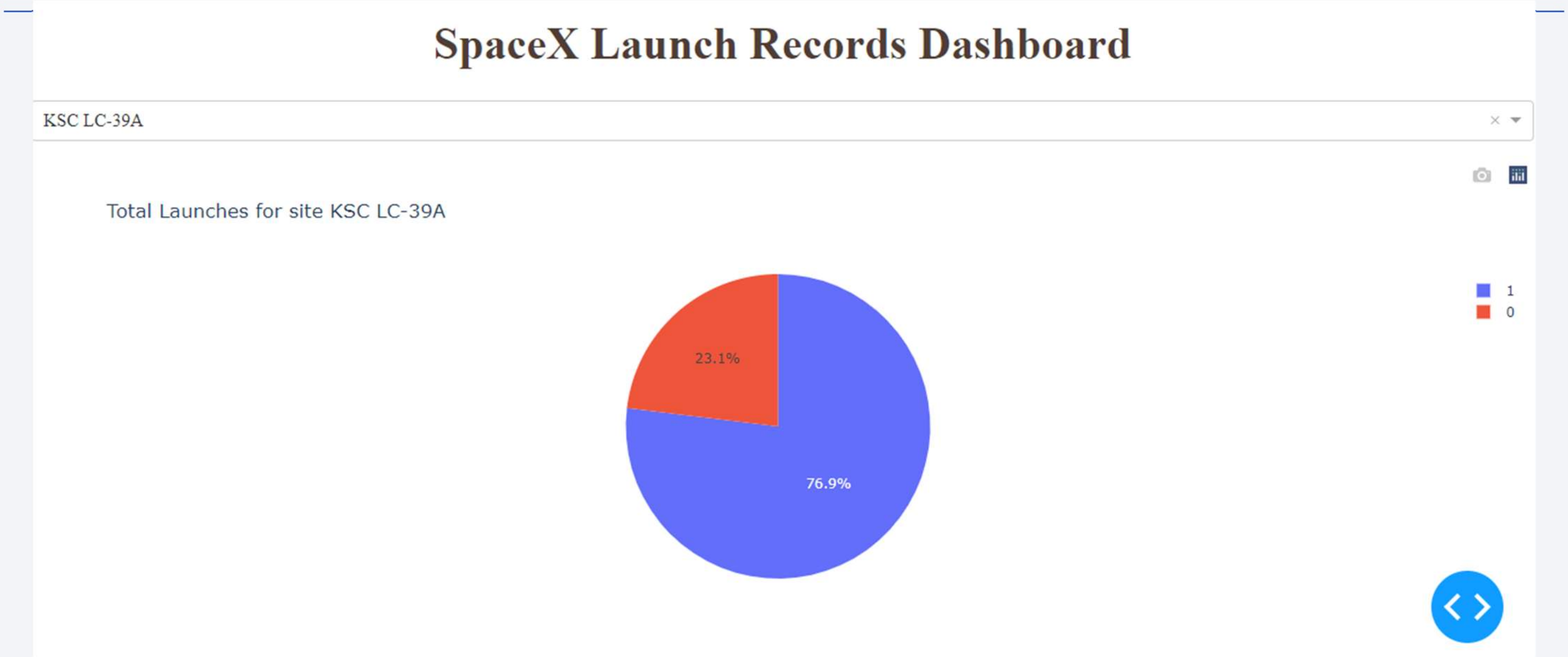
SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

KSC-LC-39A is the site with most success rate

Launch Success Ratio for KSC LC-39A



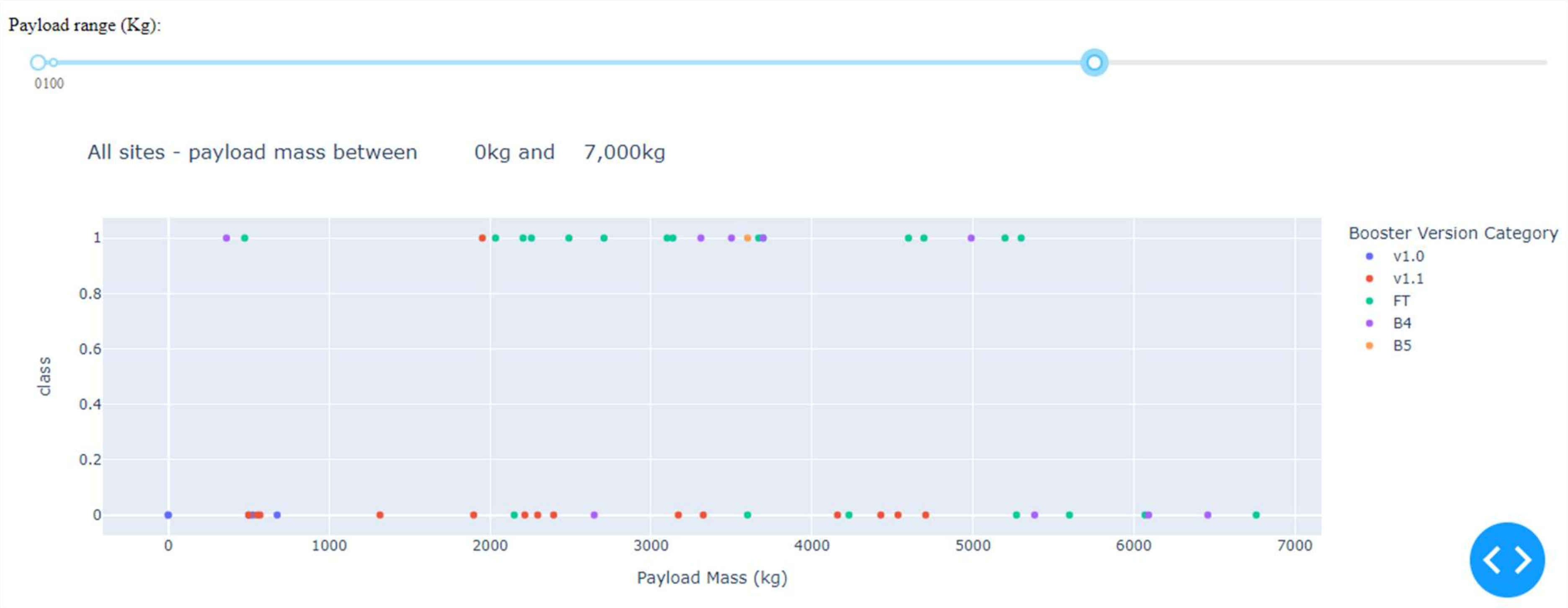
SpaceX Launch Records Dashboard

KSC LC-39A

Total Launches for site KSC LC-39A

76.9% of success rate

Payload vs. Launch Outcome



Payload range (Kg):

0100

All sites - payload mass between 0kg and 7,000kg

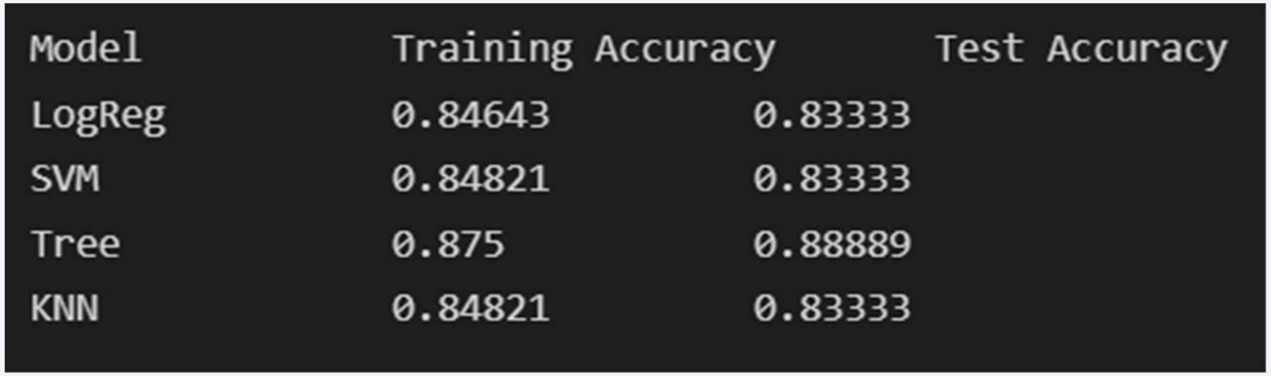
Until 7,000 kg, FT is the most effective Booster Version Type, while v1.1 is the lesser of

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- 4 classifier models were tested
  - Logistic Regression
  - SVM
  - Tree Decision
  - KNN



```
Model            Training Accuracy       Test Accuracy
LogReg           0.84643          0.83333
SVM              0.84821          0.83333
Tree             0.875            0.88889
KNN              0.84821          0.83333
```

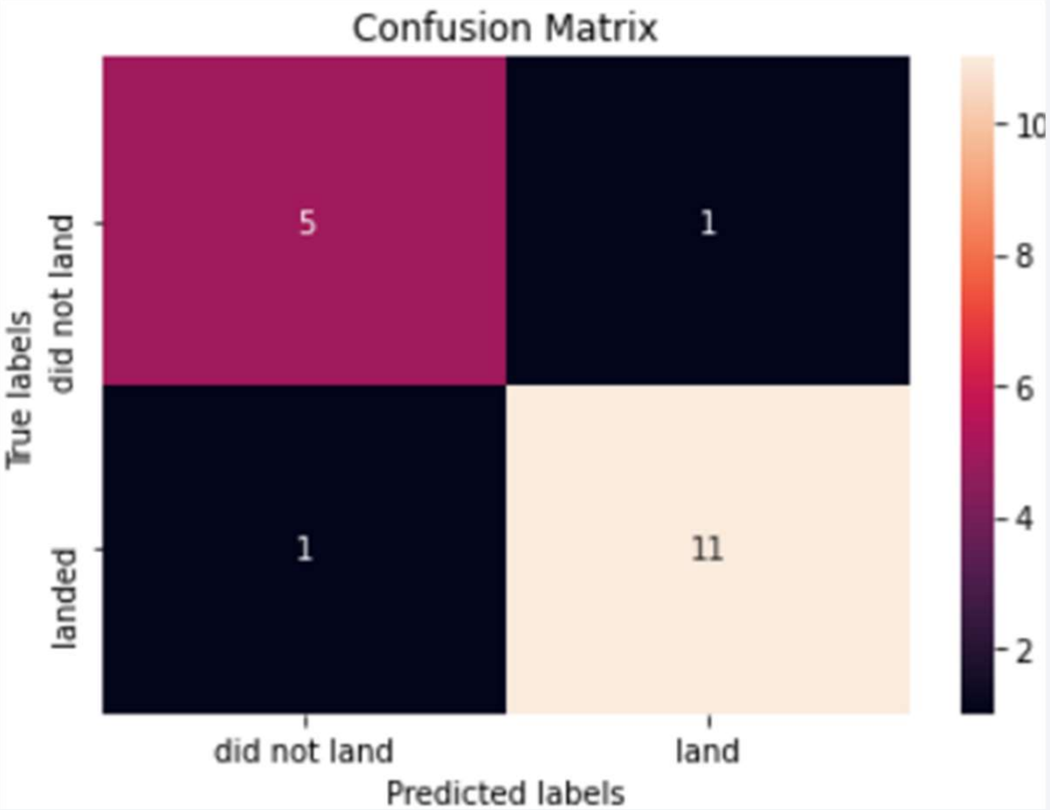
- According to the figure, Tree Decision was the model with the best accuracy in both training set and testing set

- All models were trained to find out the best hyperparameters in the training set. The same model optimized was applied in the test set to evaluate accuracy

Confusion Matrix of Decision Tree Classifier

- In 12 successful landings, our classifier previewed 11 correctly
  - 1 false positive

- In 6 failures, our classifier previewed 5 correctly
  - 1 false negative



Conclusions

- Public data sources were used to collect data
- EDA (graphical and SQL) was crucial to evaluate the influence of features on successful landing (e.g. Launch site, orbit and pay load mass)
  - Success rate have improved over time
  - The best launch site is KSC LC-39A
  - Insights obtained from dashboards and geographic visualization
- Decision Tree classifier was the supervised learning algorithm with the best accuracy to preview the success of a rocket landing
  - Over 85% of accuracy on test dataset
  - High accuracy means cost predictability

Appendix

- It would be interesting to test a different approach:
  - Data splitting first and normalization after. Hight accuracy for both training and testing datasets may indicate data leakage
- Decision tree was not the best classifier for all executions
  - All classifiers have performed similarly: high accuracy for both training and test datasets
  - Other classifiers have also achieved the best accuracy for different executions

Thank you!