IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Griffon
01/02/2022

Outline

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

Executive Summary

In this project, data on SpaceX rocket first stage landing are gathered from SpaceX API and Wiki webpage, in an aim to predict whether the landing will be successful such that the cost can be saved. Explorative data exploration techniques are applied to evaluate the relationship of different variables and to select important features. These features are used to establish a classification model to predict success/failure of the landing. Various classification models are attempted, and all show promising results in terms of model accuracy.

Introduction

- Rocket launch is resource-consuming; it can save a lot of cost and resource if the launch first stage is successful, and the rocket can be recycled.
- Therefore, it's helpful to know whether a landing can be successful based on information such as, payload mass, launch site and etc.
- With a lot of data available online, they can be scraped from the internet and utilized to establish machine learning models for the purpose of prediction.

Section 1

Methodology

Methodology

Executive Summary

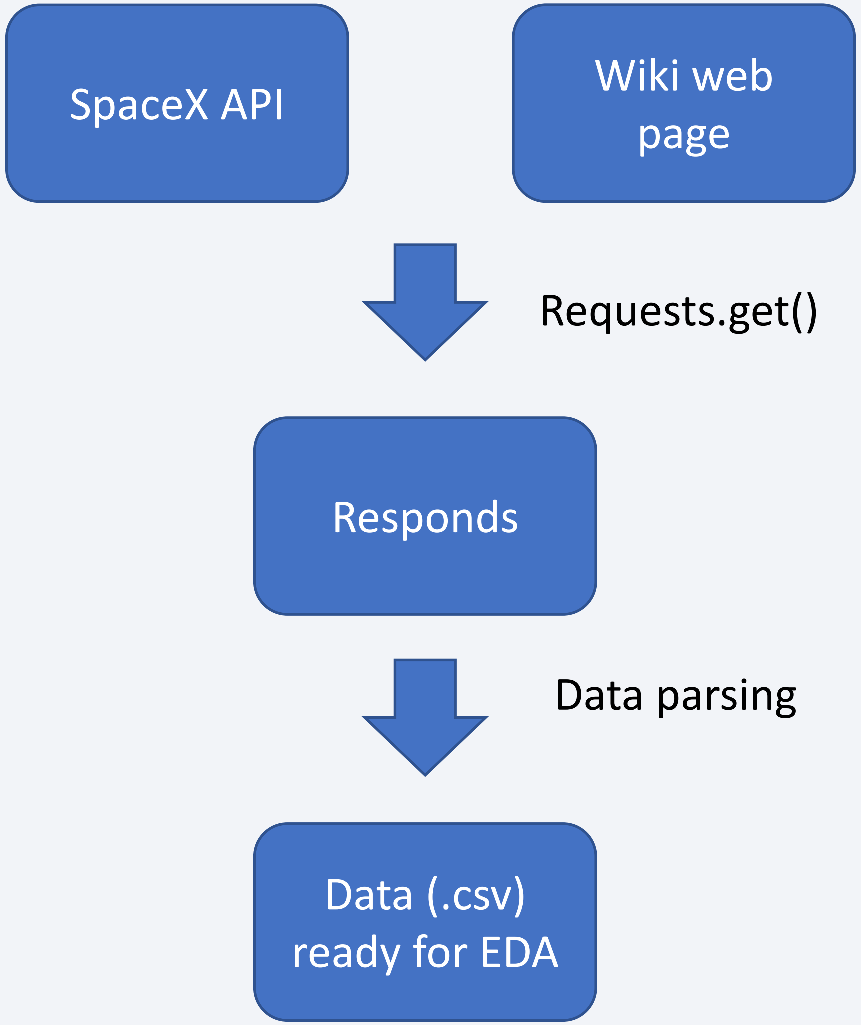
- Data collection methodology:
    - SpaceX API and web scraping
- Perform data wrangling
    - Data filtering, replacing missing data and determining training labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Using Scikit-learn to import model object
    - Using Grid search to find the optimal hyperparameters
    - Use accuracy, confusion matrix and so on for evaluation

Data Collection

- Data are collected in two manners:
  - Request from SpaceX API
  - Scrape from a WIKI web page

- Both involve using requests module to obtain web information via HTTP connection

- Both involve data processing of raw content from the web


SpaceX API

Wiki web page

Requests.get()

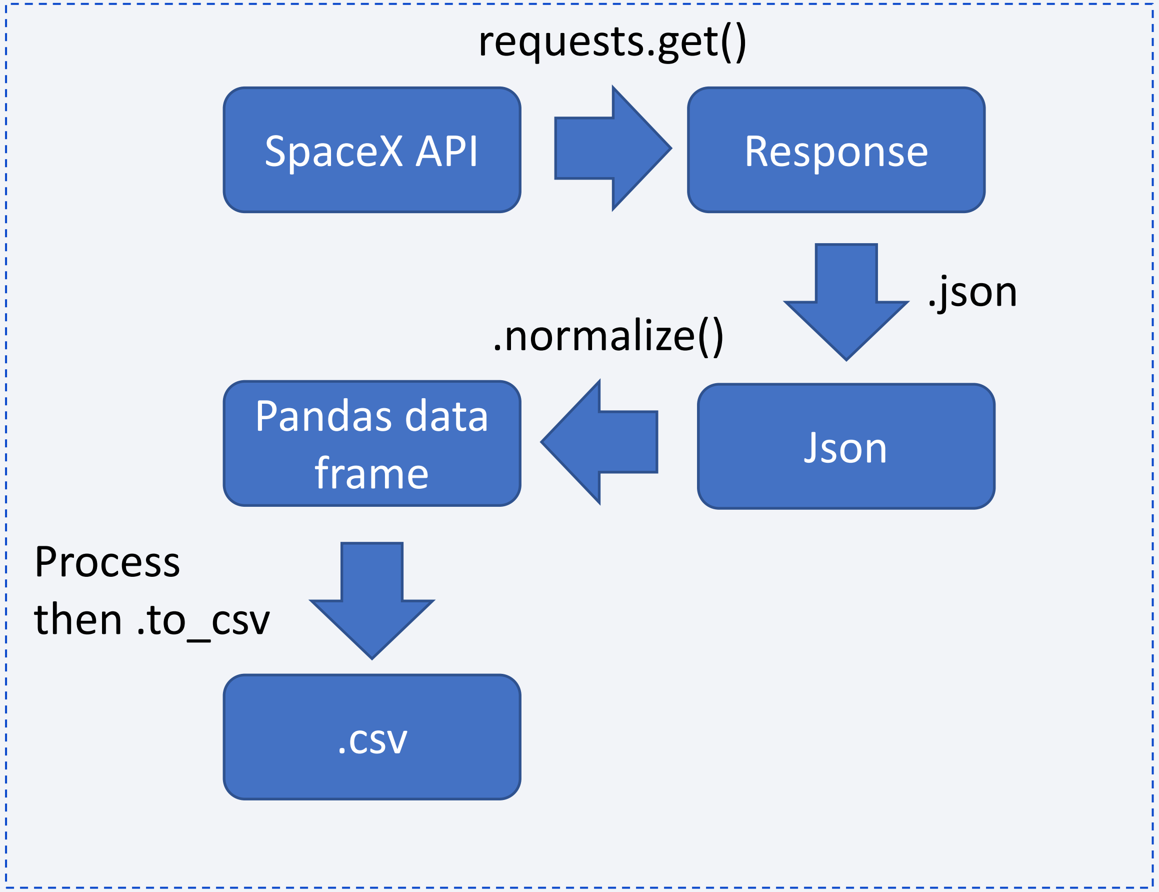
Responds

Data parsing

Data (.csv) ready for EDA

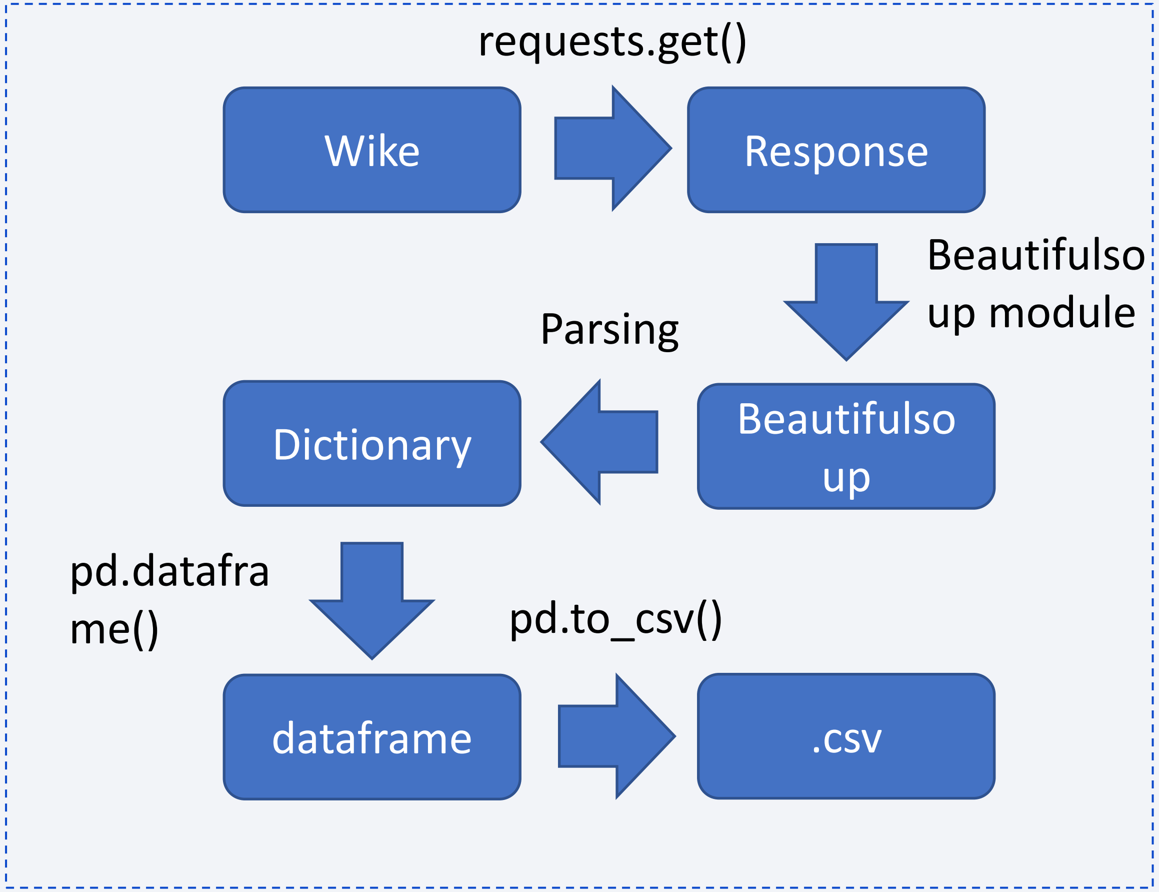
Data Collection – SpaceX API

- Use get function in requests module to request rocket launch data from SpaceX API.
- Decode the response content as a json and “normalize” the result into a pandas dataframe.
- Use the IDs in the original dataframe to request the actual information that is needed, e.g., booster name and payload.
- Turn the data obtained into a new pandas data frame
- Filter the data to keep only Falcon 9 launches
- Replace missing payload data with its mean value
- GitHub URL of the SpaceX API calls notebook: SpaceX data collection API (Ctrl + click to follow link)



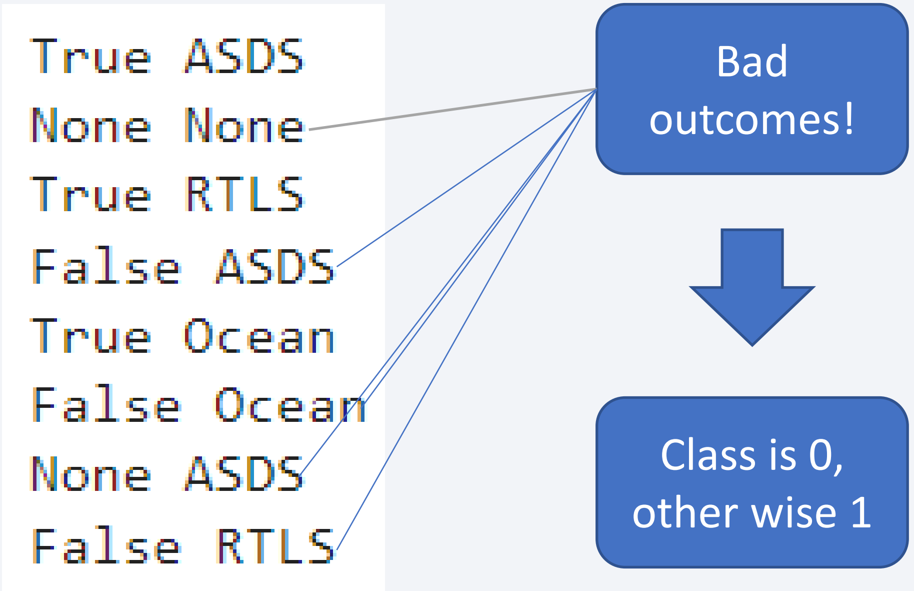
Data Collection - Scraping

- Use requests.get() method to request the Falcon 9 launch wiki web page.
- Create a BeautifulSoup object from the response.
- Extract column names from the HTML table header
- Parse the HTML table to create a launch dictionary which is then converted to a pandas dataframe.
- GitHub URL of the web scraping notebook: [Data Collection - Scraping]() (Ctrl + click to follow link)



Data Wrangling

- Read data from CSV to pandas dataframe with pd.read_csv() method
- Calculate the number of launches on 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 and success rate
- GitHub URL of the data wrangling notebook: Data wrangling (Ctrl + click to follow link)



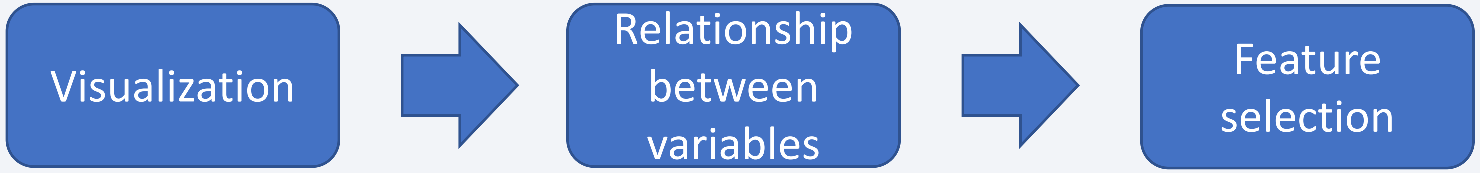
```
True ASDS
None None
True RTLS
False ASDS
True Ocean
False Ocean
None ASDS
False RTLS
```

Bad outcomes!

Class is 0, other wise 1

EDA with Data Visualization

- Scatter plot
  - Pay load mass vs. flight number
  - Launch site vs. flight number
  - Launch site vs. pay load mass
  - Orbit vs. fight number
  - Orbit vs. pay load mass
- Bar plot – class vs. orbit
- Line plot – class vs. year
- Feature engineering
- One hot encoder and convert type to float
- GitHub URL of the EDA notebook: [EDA with data visualization]() (Ctrl + click to follow link)



EDA with SQL

- The names of the unique launch sites in the space mission
- 5 records of launch sites beginning with 'CCA’
- The total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- 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 failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranking of landing outcomes count between the date 2010-06-04 and 2017-03-20
- GitHub URL of the web scraping notebook: EDA with SQL (Ctrl + click to follow link)

Build an Interactive Map with Folium

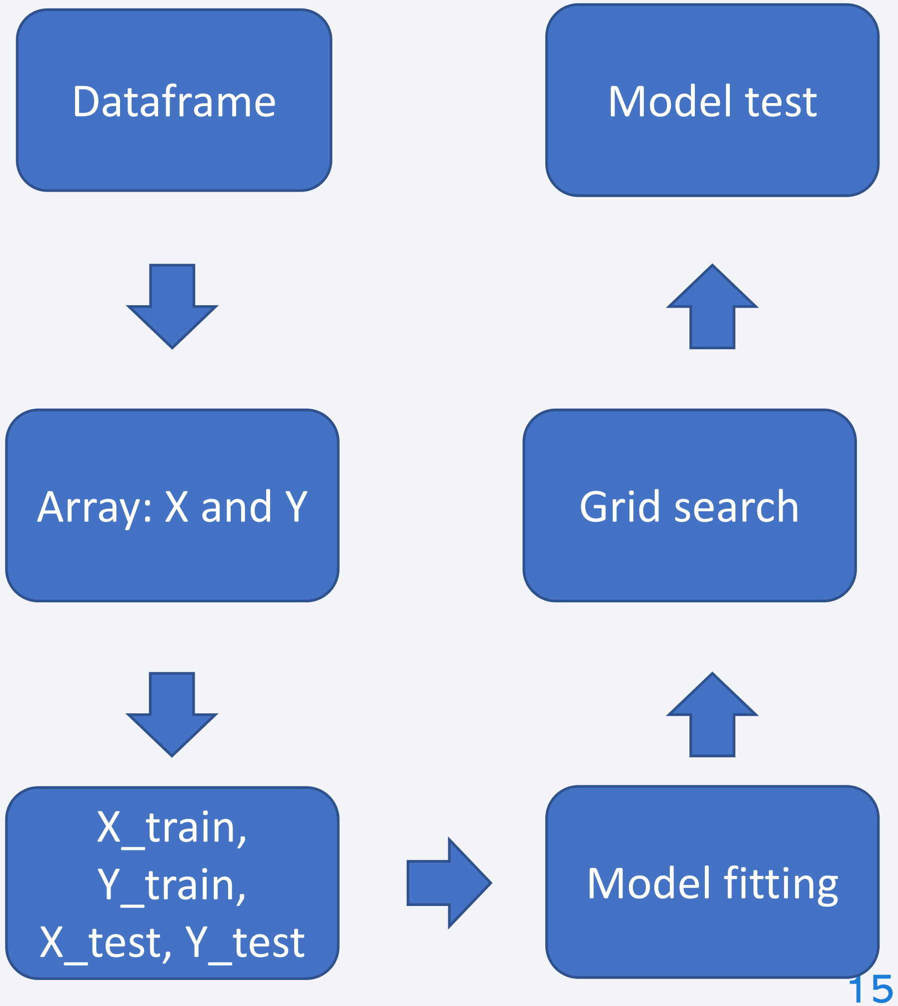
- Map objects that were created and added to a folium map
  - Circle and marker --- NASA Johnson Space Centre's coordinate
  - Circle and marker --- All launch sites locations
  - Mark the success/failed launches for each site on the map with marker cluster
  - Line denoting the distance between launch site and the coastline
- To evaluate potential dependency of launch outcome and the location/proximities of a launch site.
- GitHub URL of the interactive map with Folium notebook: Interactive Map with Folium (Ctrl + click to follow link)

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions that are added to a dashboard
  - a Launch Site Drop-down Input Component to select launch site
  - success-pie-chart generated based on selected site dropdown
  - a range slider to Select Payload
  - success-payload-scatter-chart scatter plot based on selected launch site and payload range
- For easy and dynamic demonstration of data exploration results.
- GitHub URL of the Dashboard with Plotly Dash notebook: Dashboard with Plotly Dash (Ctrl + click to follow link)

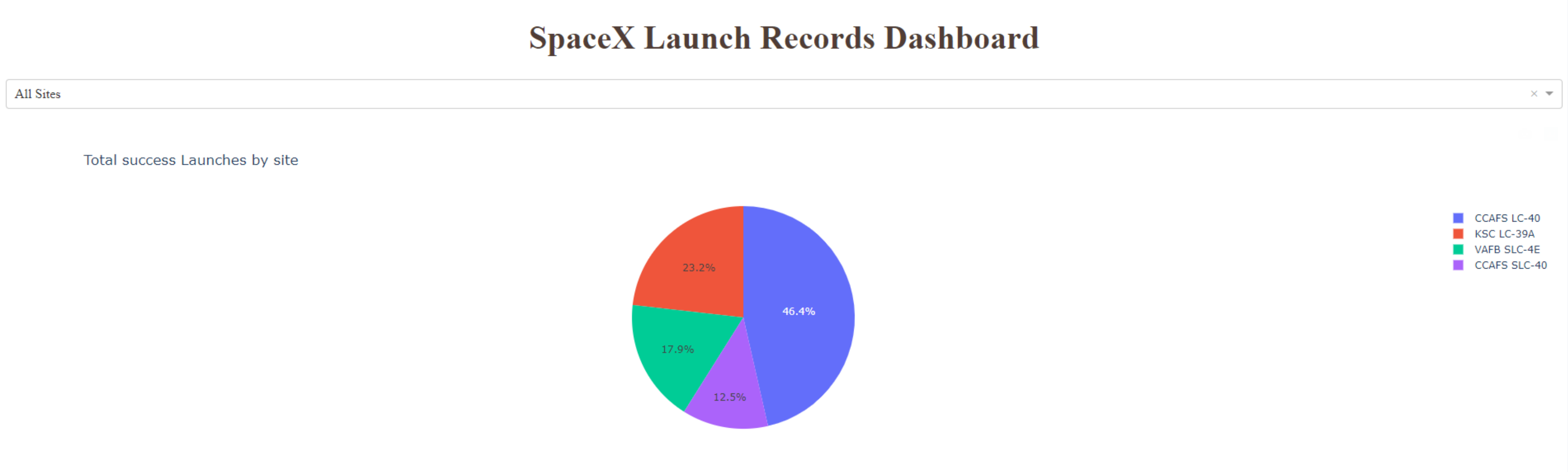
Predictive Analysis (Classification)

- Process of classification model setup
  - Read data frame, turn dataframe to Numpy array
  - Preprocessor to standardize
  - Train, test, split
  - GridsearchCV
  - Logistic regression, SVM, Decision Tree, KNN
- GitHub URL of the Predictive analysis notebook: Predictive analysis (Ctrl + click to follow link)



Results

- Launch results are relevant to factors such as orbits and launch site.



SpaceX Launch Records Dashboard

All Sites

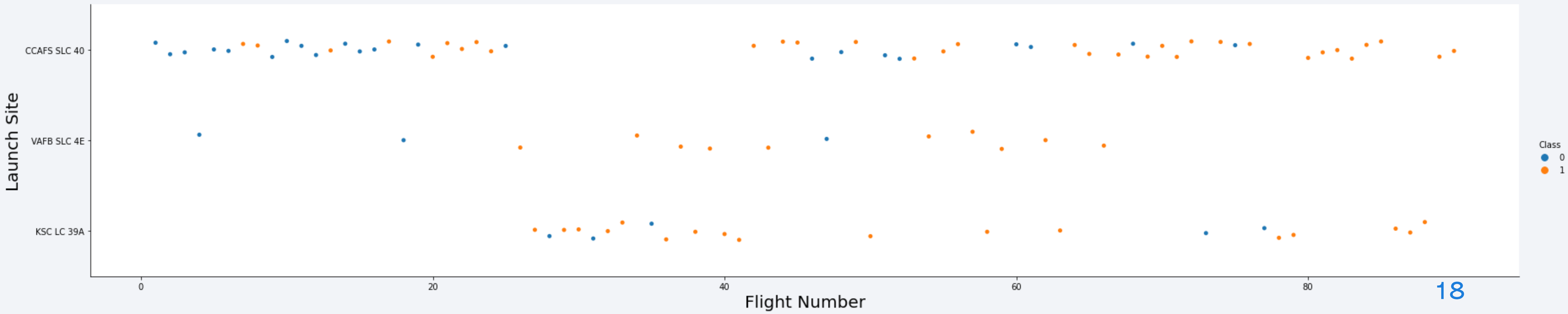
- All predictive models show sound accuracy (up to 0.833), except the decision tree model.

Section 2

Insights drawn from EDA

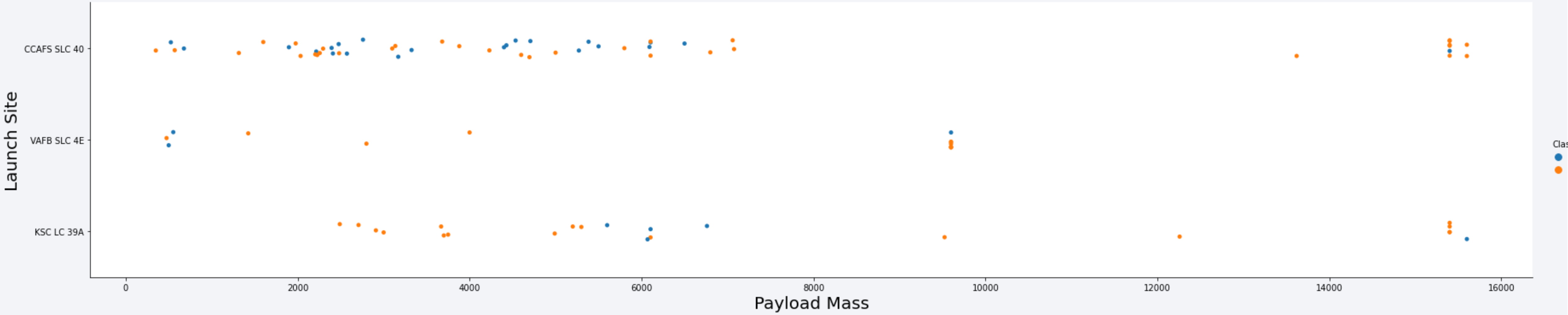
Flight Number vs. Launch Site

- Most launches took place in CCAFS SLC 40.
- VAFB SLC 4E yields the most successes.
- All launches after number 80 were successful.



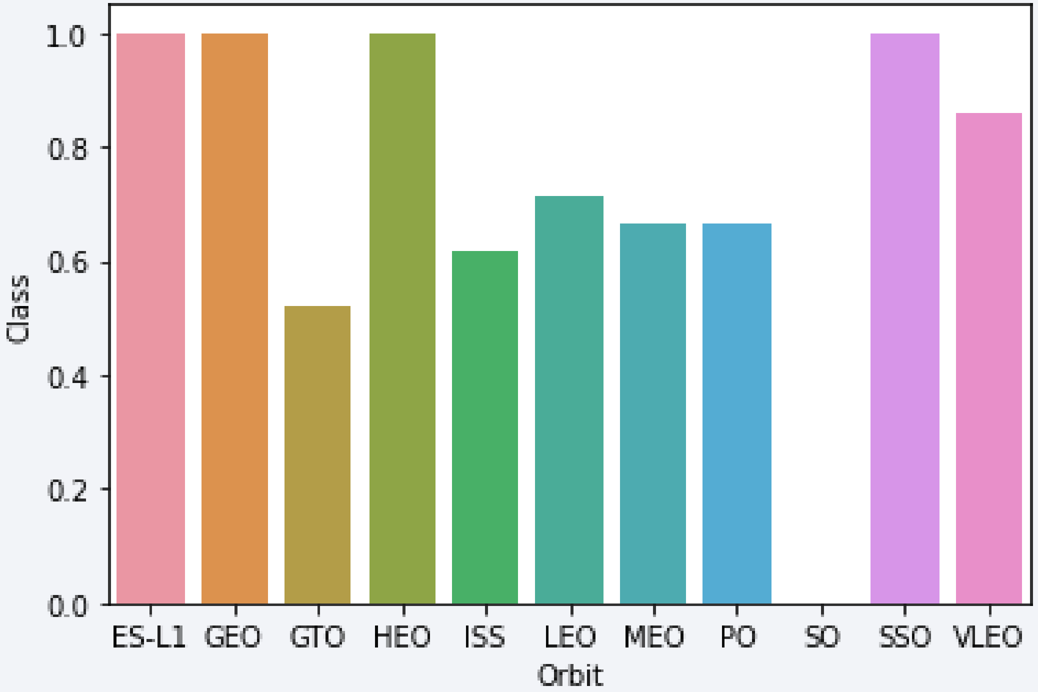
Payload vs. Launch Site

- In CCAFS SLC 40, flights with only either small or very large payload mass were launched.
- Larger payload mass is more likely to lead to successful launches.



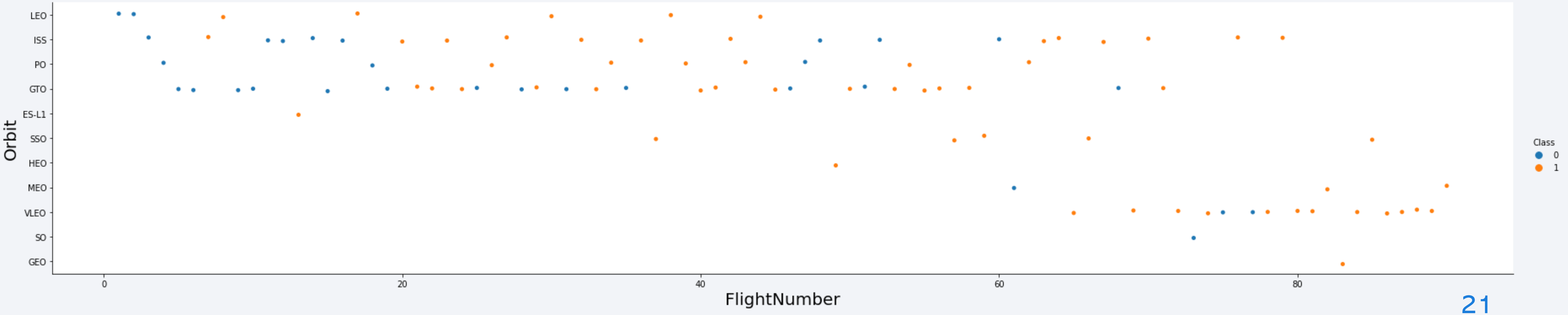
Success Rate vs. Orbit Type

- Success rates are high for orbit ES-L1, GEO, HEO and SSO, i.e., 100%.
- Orbit SO has 0 success rate



Flight Number vs. Orbit Type

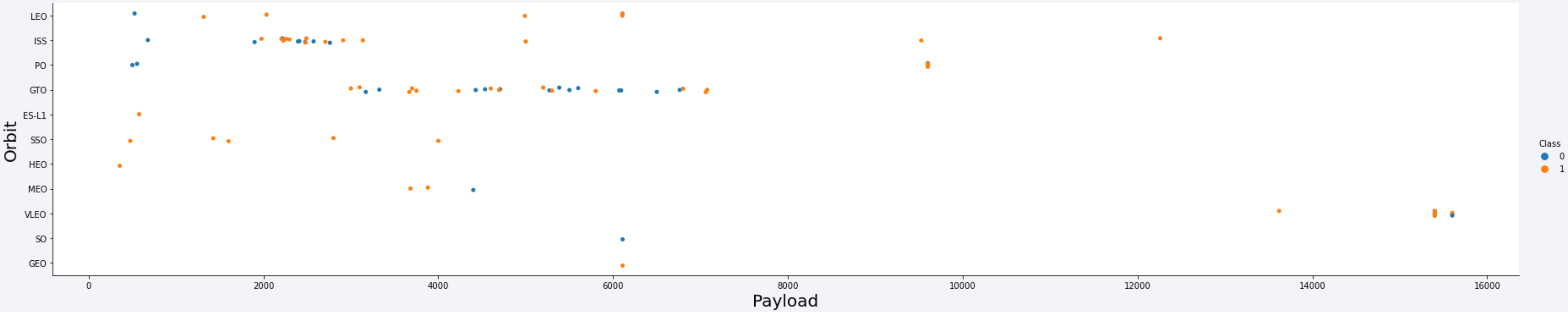
- Most launches have the orbit type of LEO, ISS, PO, GTO.
- The other orbits are used less, thus explaining the high success rate in the previous slide.



Payload vs. Orbit Type

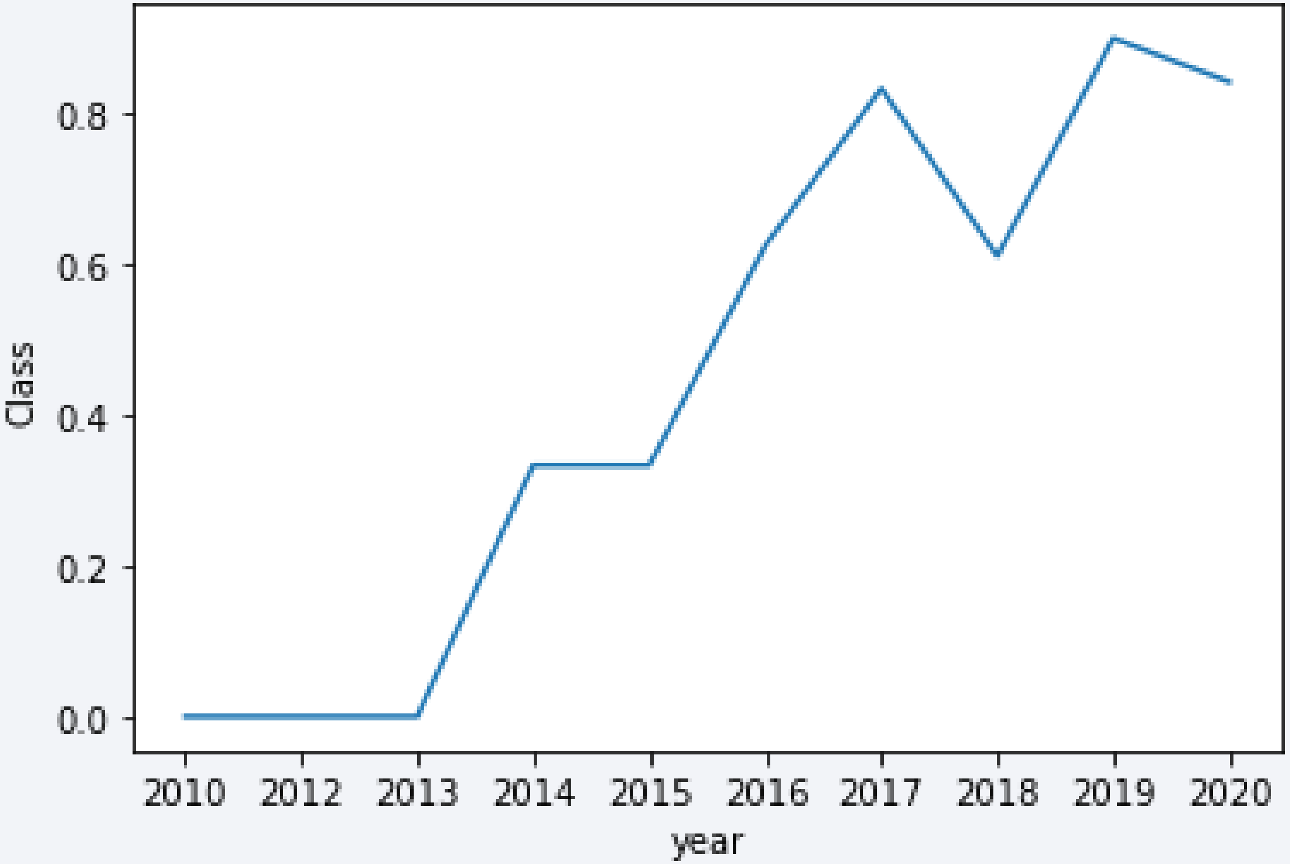
- Very large payload mass is only launched with VLEO orbit.

- Most common payload mass and orbit combo are 3000 – 7000 kg with GTO orbit.



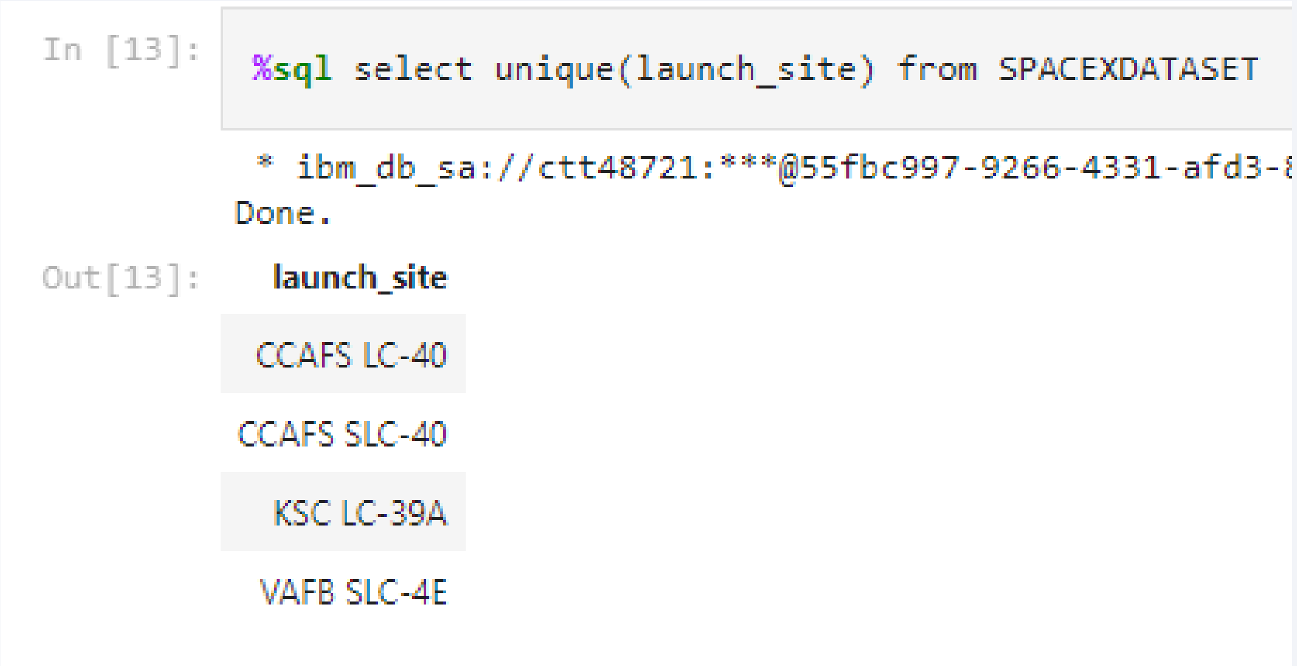
Launch Success Yearly Trend

- Success rate generally increases with time.
- Since 2013, the launches began to succeed for the first time.
- After 2018, the success rate is close to 90%.



All Launch Site Names

- Unique() function is used to retrieve distinct launch site names.



In [13]:

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

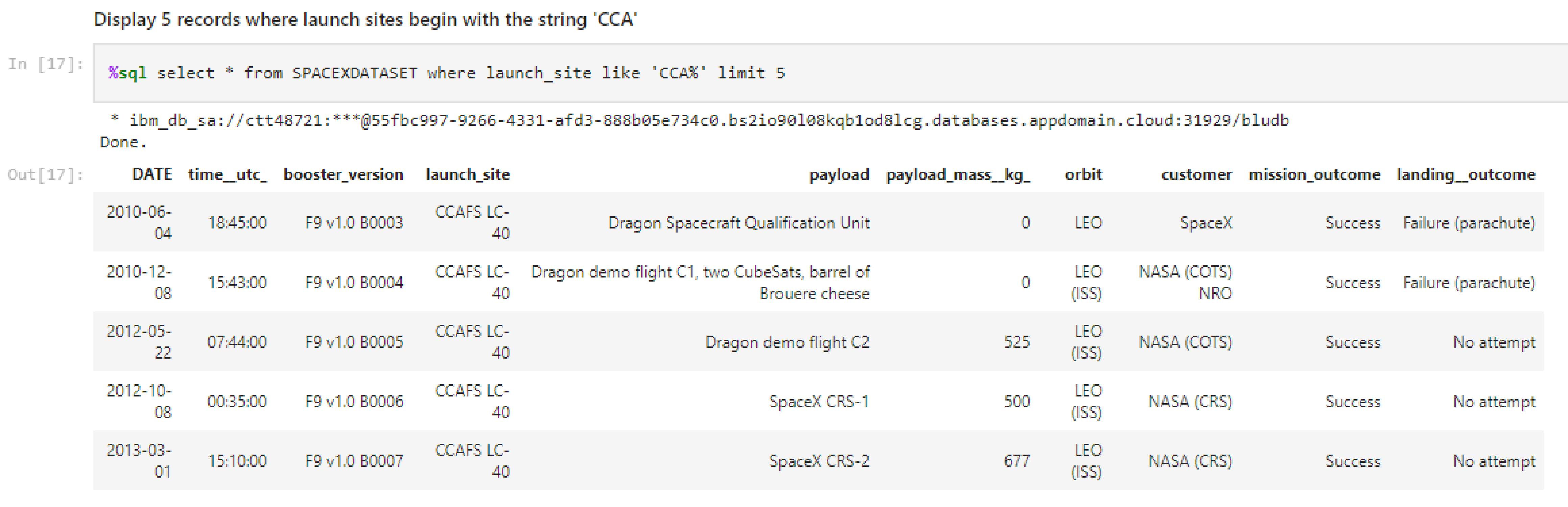
```
 * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-
Done.
```

Out[13]:

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

Launch Site Names Begin with 'CCA'

- Wildcard here is used to retrieve those records with the string ‘CCA’.



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

```
In [17]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

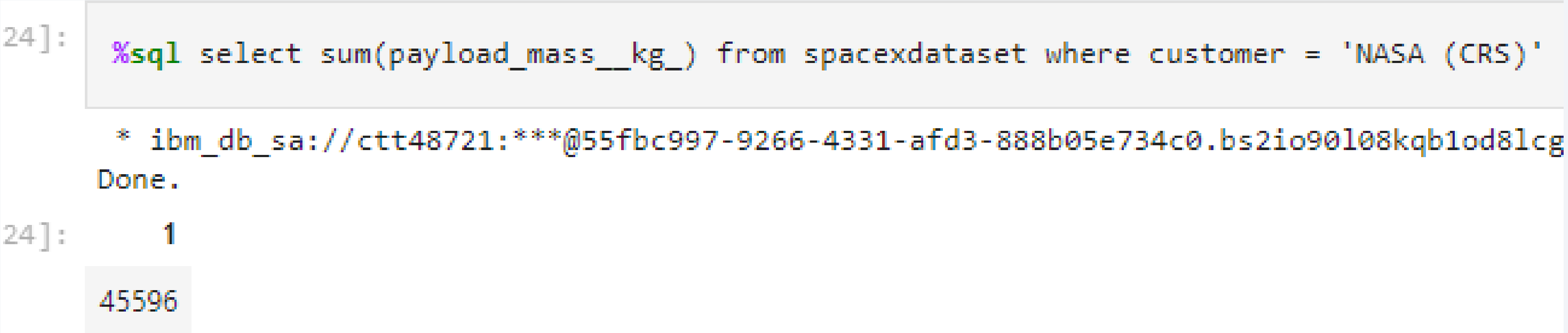
```
 * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[17]:

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 by boosters from NASA is calculated to be 25596 kg.
- Here the sum() function is used in conjunction with a conditional clause.

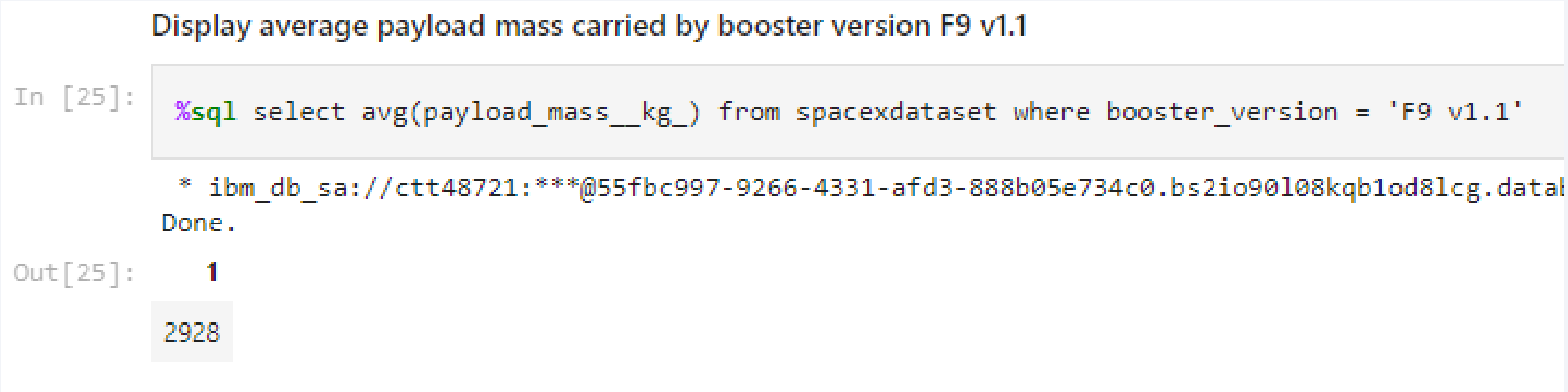


```
24]:  %sql select sum(payload_mass__kg_) from spacexdataset where customer = 'NASA (CRS)'

       * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg
      Done.
24]:      1
      45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is calculated to be 2928 kg.
- Here the avg() function is used.



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

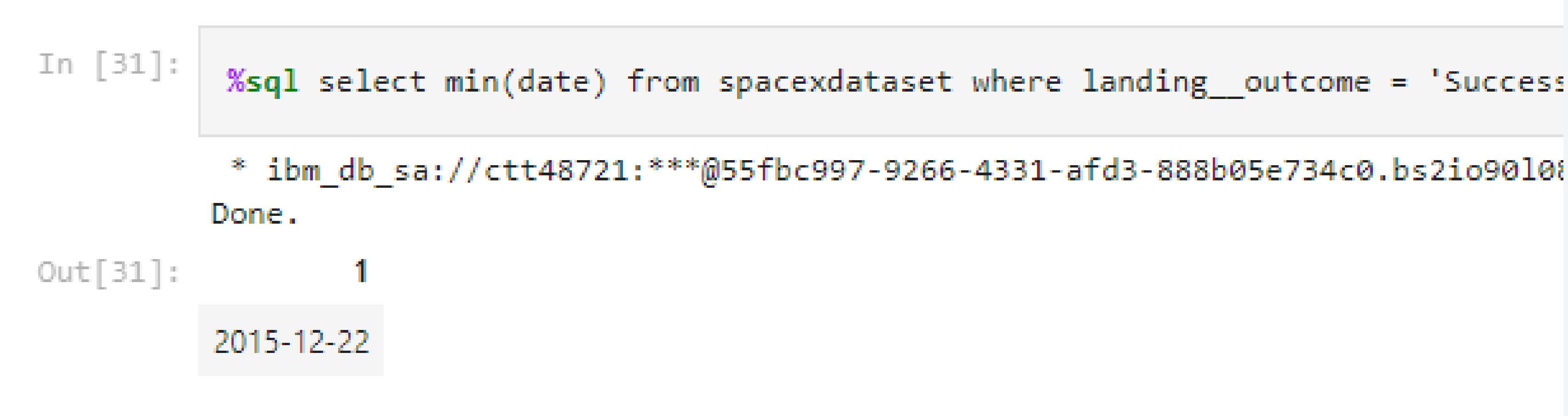
```
In [25]:  %sql select avg(payload_mass__kg_) from spacexdataset where booster_version = 'F9 v1.1'

 * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.datal
Done.

Out[25]:    1
         2928
```

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 22nd Dec 2015.
- This is identified using the min() function

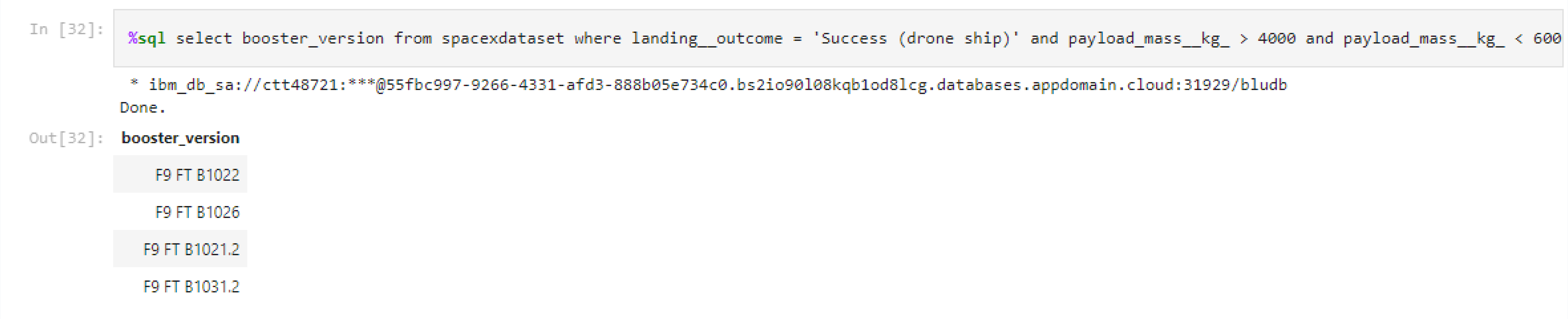


```
In [31]: %sql select min(date) from spacexdataset where landing__outcome = 'Success

          * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08
         Done.
Out[31]:            1
         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 are listed.



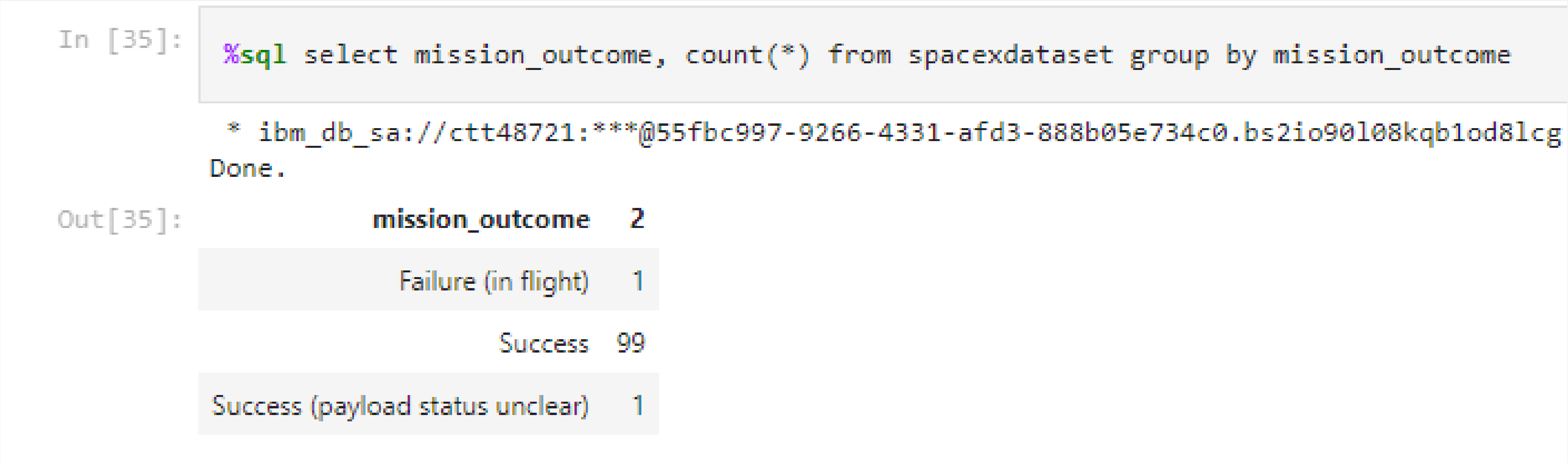
```
%sql select booster_version from spacexdataset where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 600
```

```
 * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are calculated.
- The count() function is used along with “group by”.



```
In [35]: %sql select mission_outcome, count(*) from spacexdataset group by mission_outcome
```

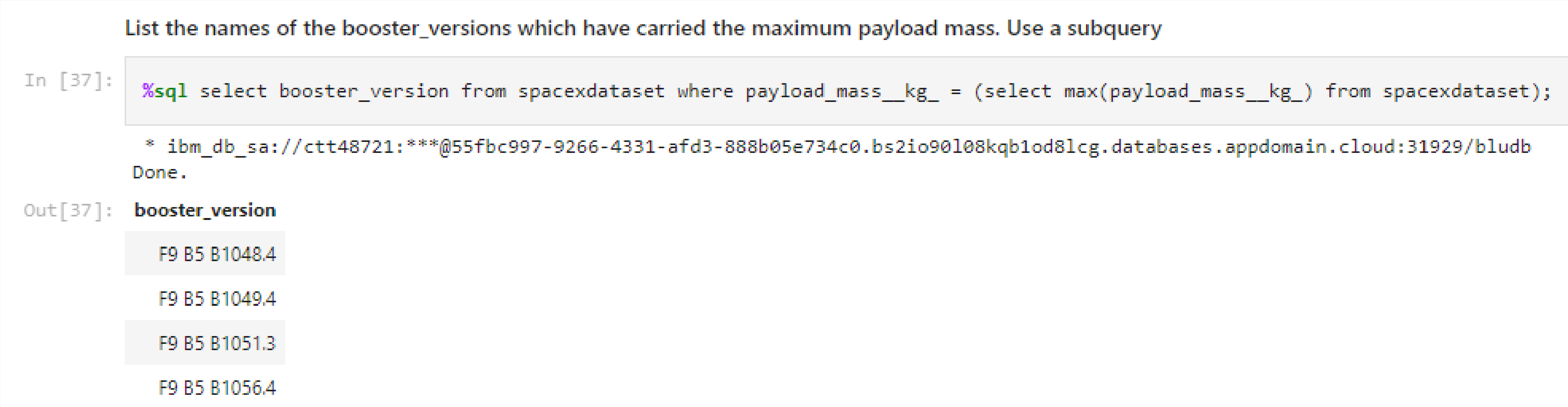
```
 * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg
Done.
```

Out[35]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are listed
- A subquery is used in the where clause.



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

```
In [37]: %sql select booster_version from spacexdataset where payload_mass__kg_ = (select max(payload_mass__kg_) from spacexdataset);
```

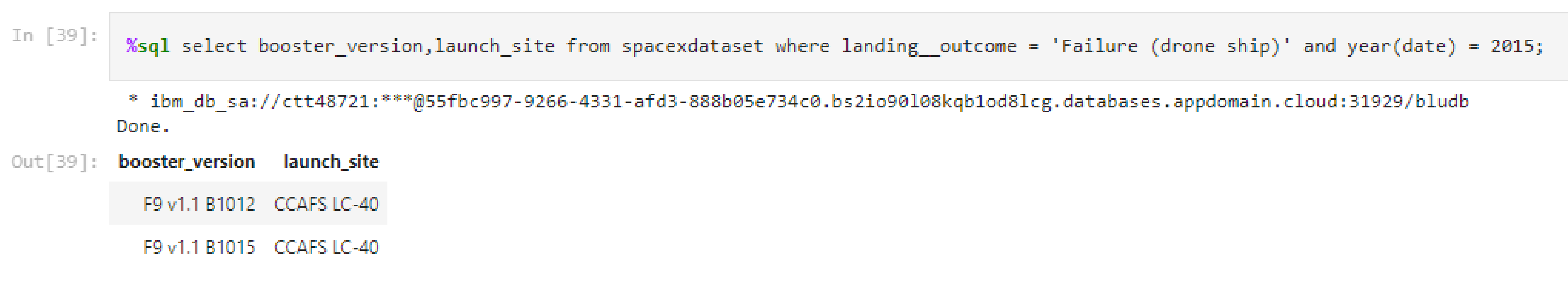
```
 * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[37]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed using where conditional clause.



In [39]:

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

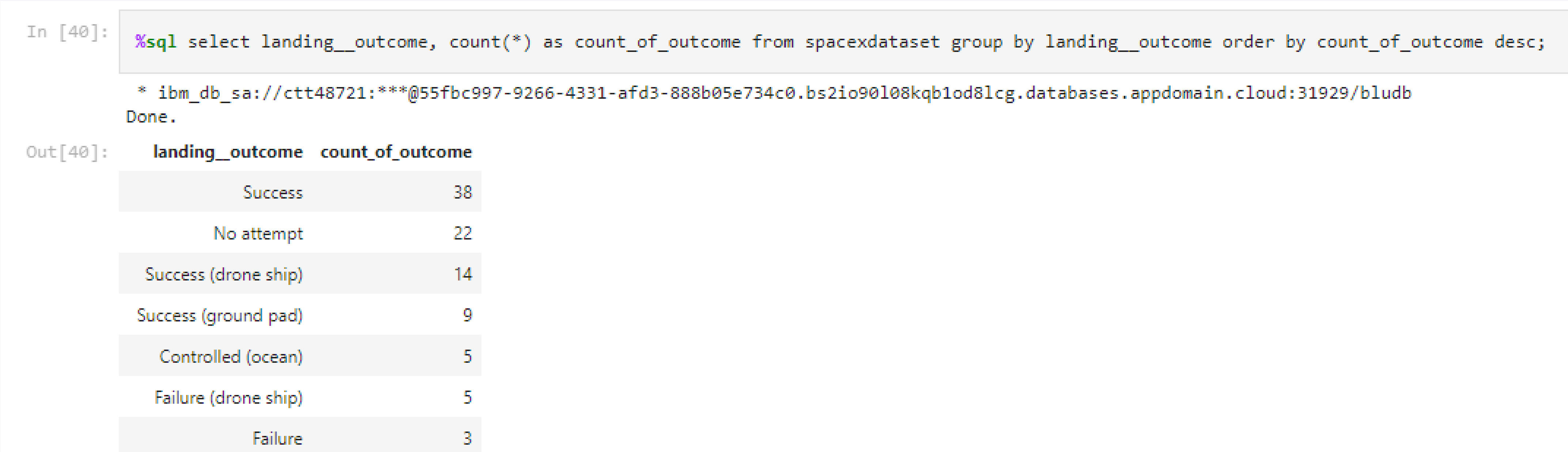
```
 * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[39]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- The ranking of the number of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 is shown here.
- Count() function, and “group by” and “order by” clause are used.



In [40]:

```
%sql select landing__outcome, count(*) as count_of_outcome from spacexdataset group by landing__outcome order by count_of_outcome desc;
```

```
 * ibm_db_sa://ctt48721:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[40]:

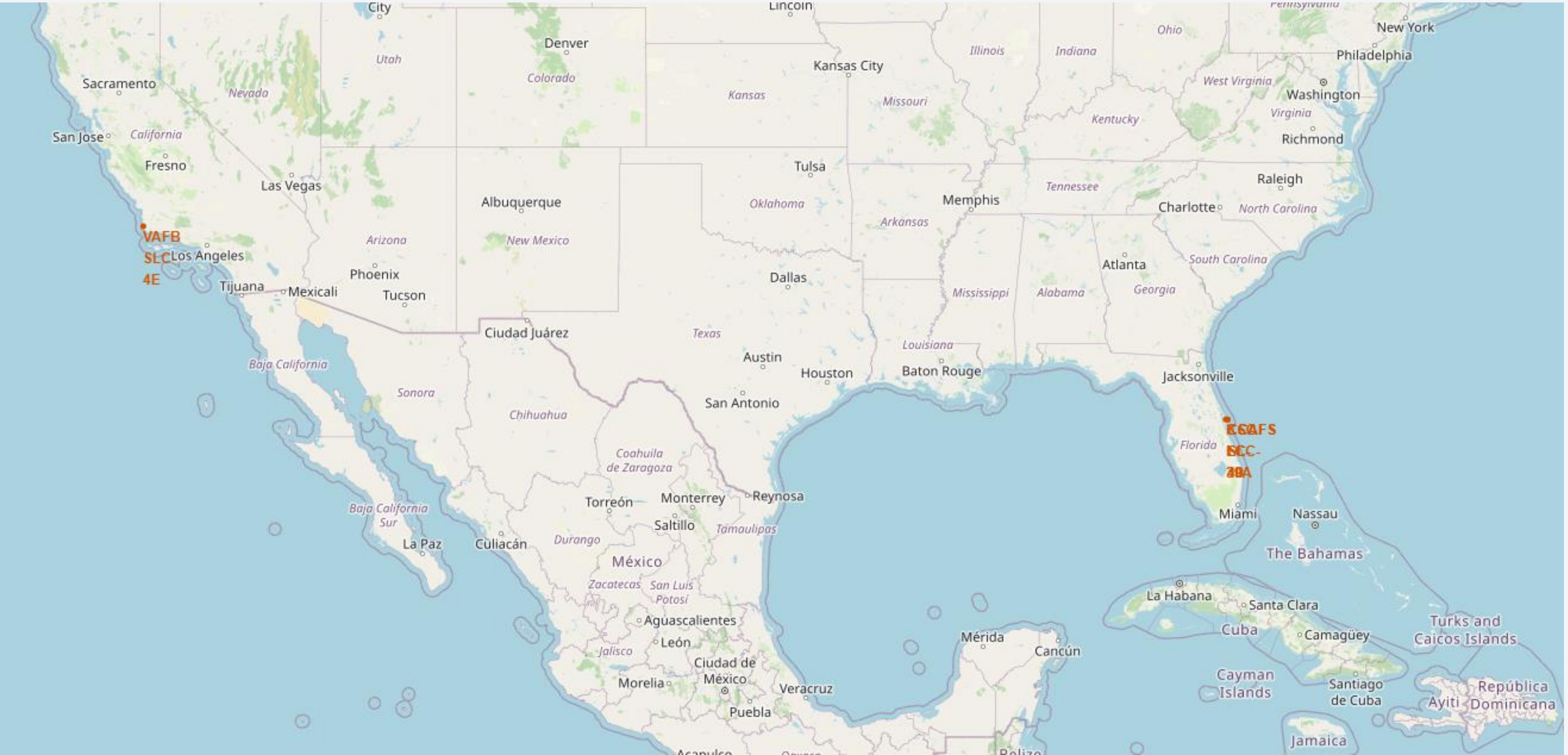
landing__outcome	count_of_outcome
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3

Section 3

Launch Sites Proximities Analysis

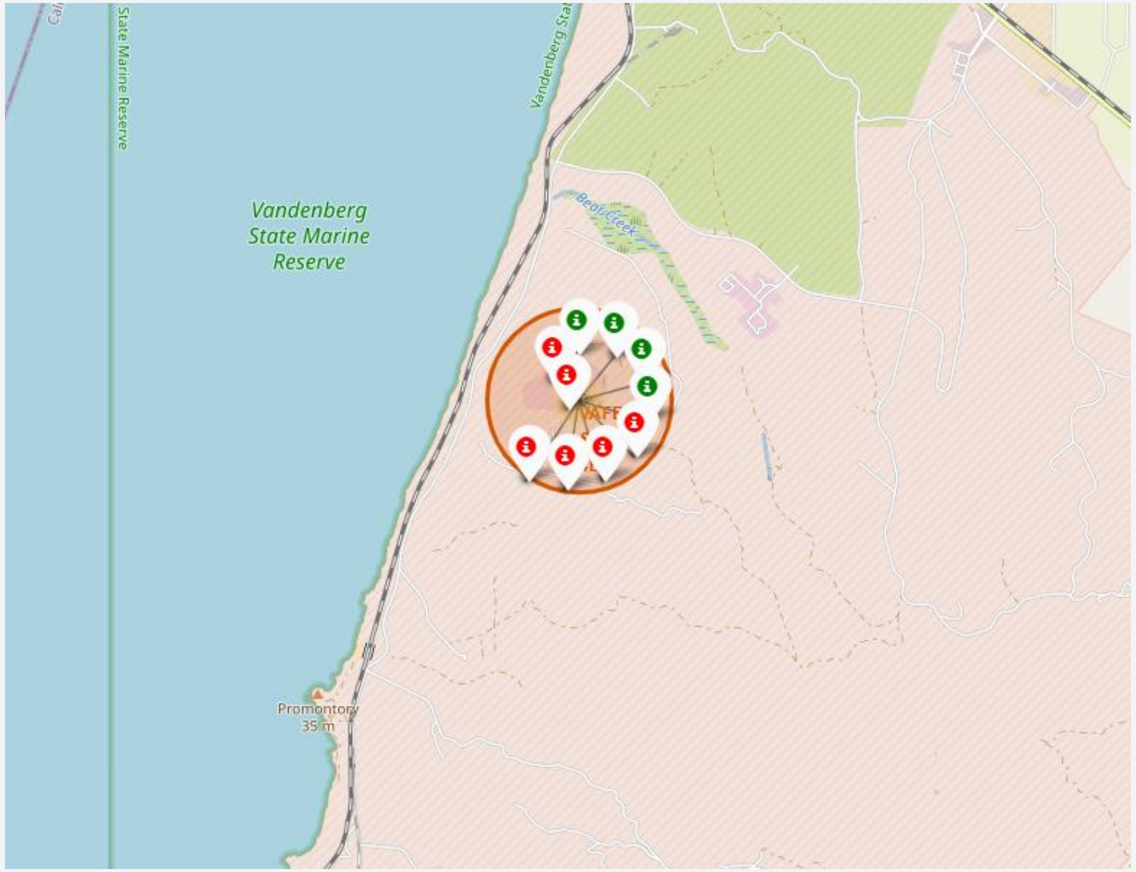
All launch sites’ location markers on a global map

- All launch sites' locations are marked with circles and labels showing their names.



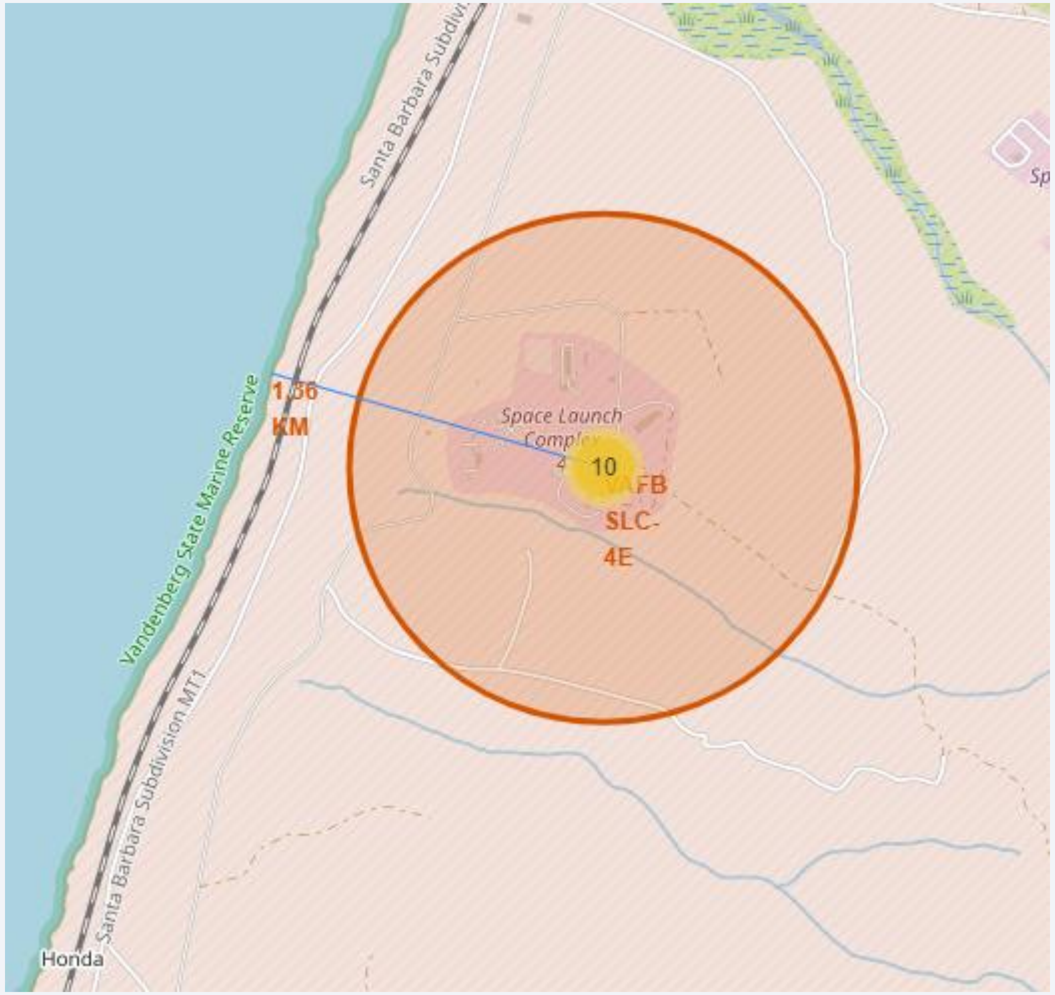
Color-labeled launch outcomes

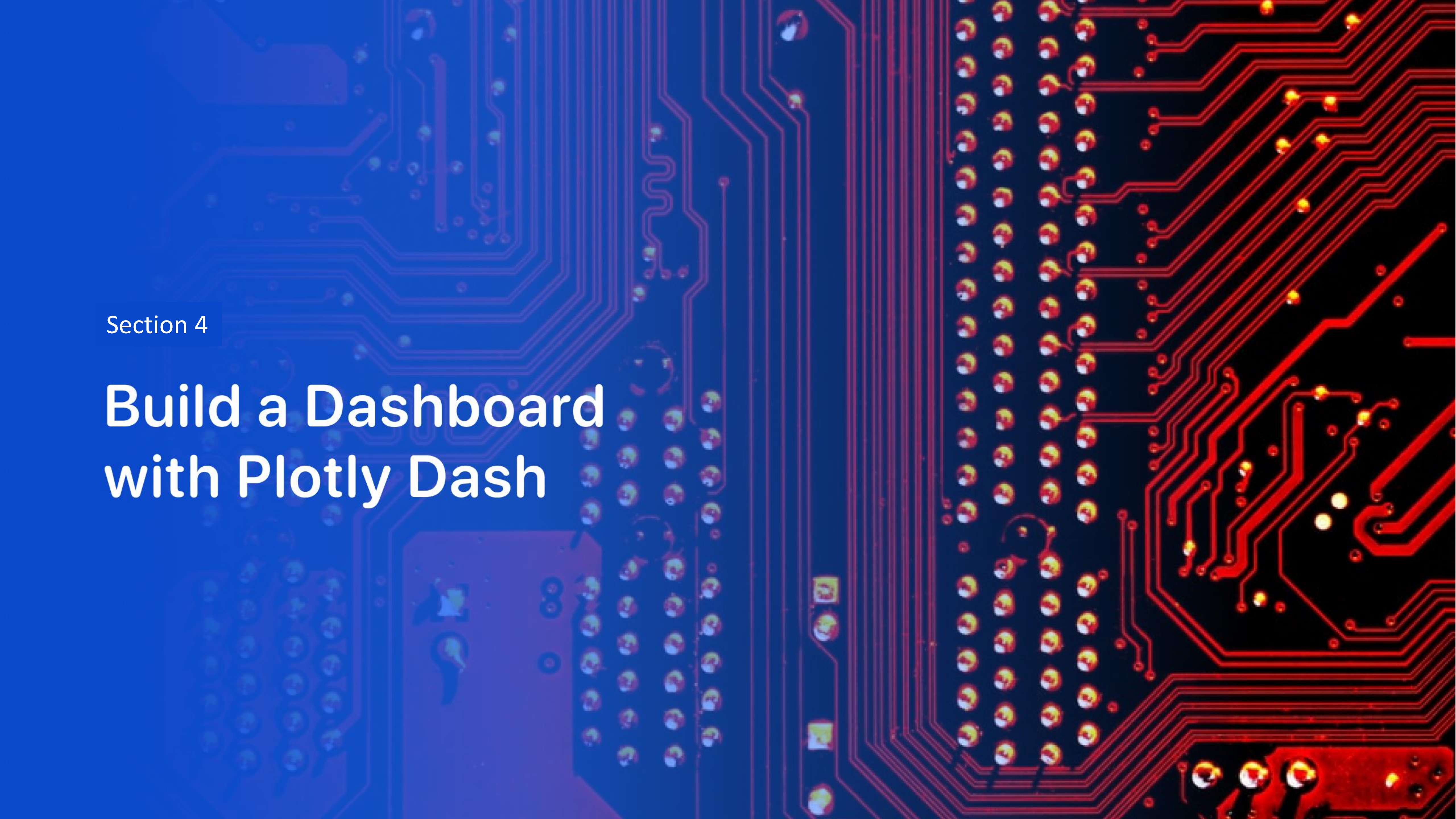
- The success or failure attempts of launches are marked at the launch site location with the green and red sign.



The distances between a launch site to its proximities

- The VAFB SLC-4E launch site is shown here. It is very close to the coastline with a distance of only 1.36kg.


1.36 KM
Space Launch Complex
10
VAFB SLC-4E
Santa Barbara Subdivision
Vandenberg State Marine Reserve
Santa Barbara Subdivision MT1
Honda
Sp

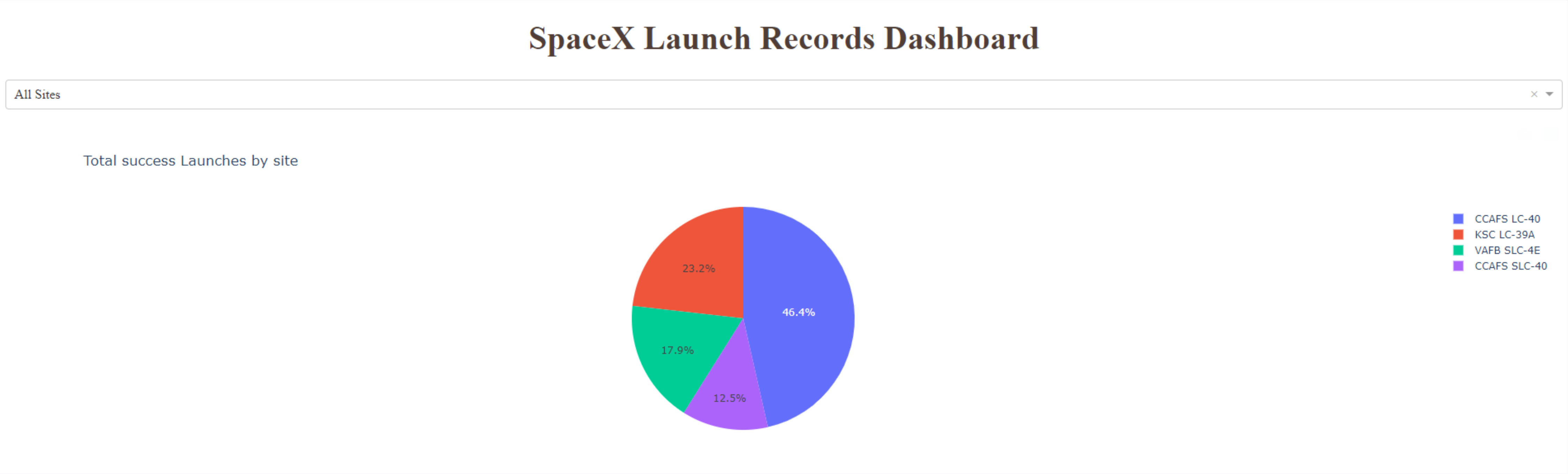


Section 4

Build a Dashboard with Plotly Dash

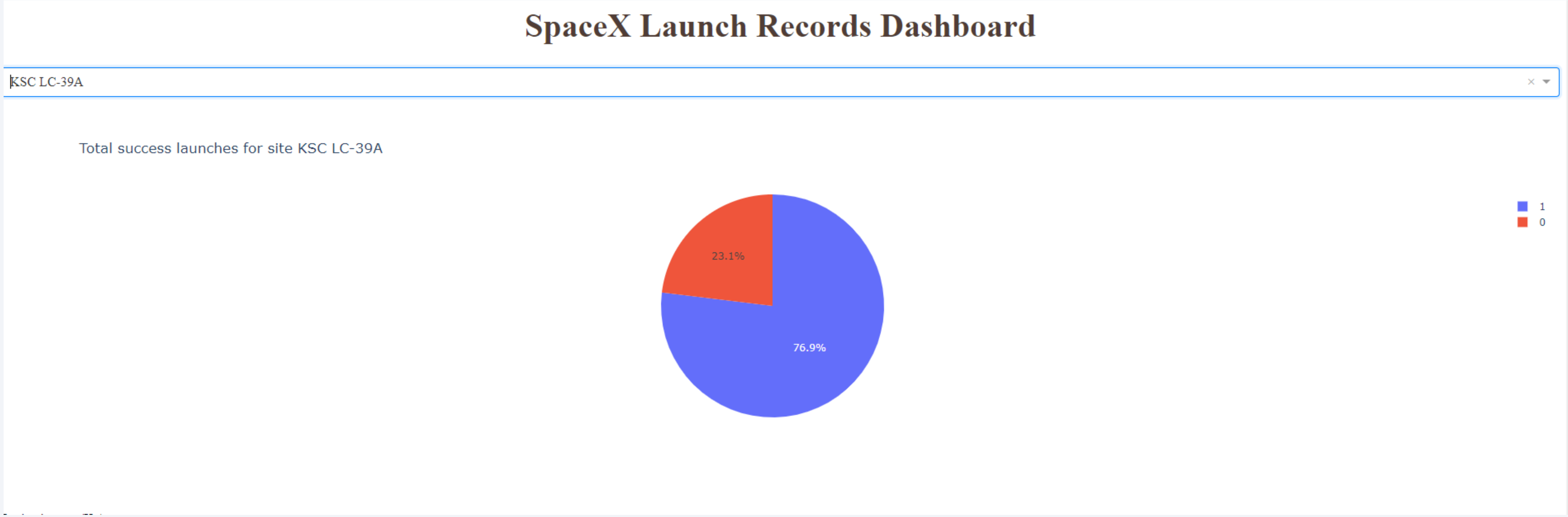
Launch success count for all sites

- CCAFS LC-40 has the most success launches of 46.4%



Pie chart for launch site KSC LC-39A

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



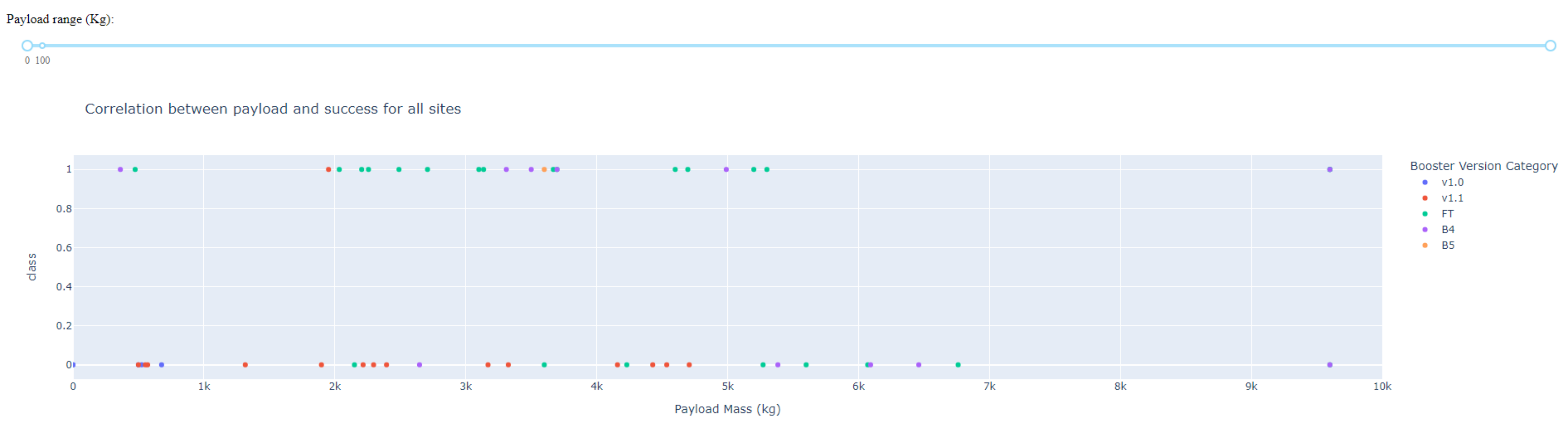
SpaceX Launch Records Dashboard

KSC LC-39A

Total success launches for site KSC LC-39A

Payload vs. Launch Outcome scatter plot for all sites

- 3k to 4k payload range or booster version have the largest success rate.
- FT booster also has many success launches.

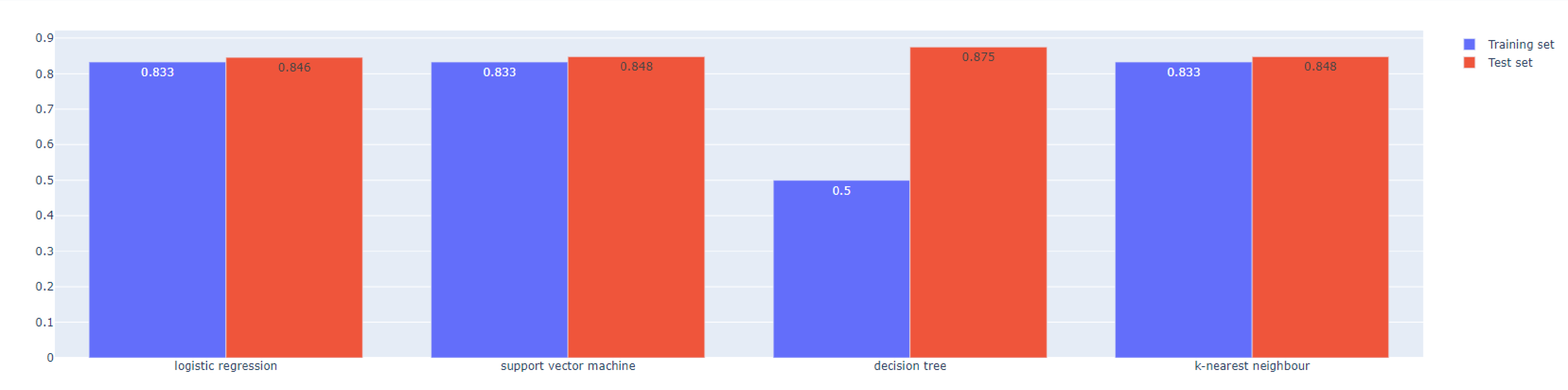


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Support vector machine and k-nearest neighbor have the same performance in terms accuracy of training set and test set.

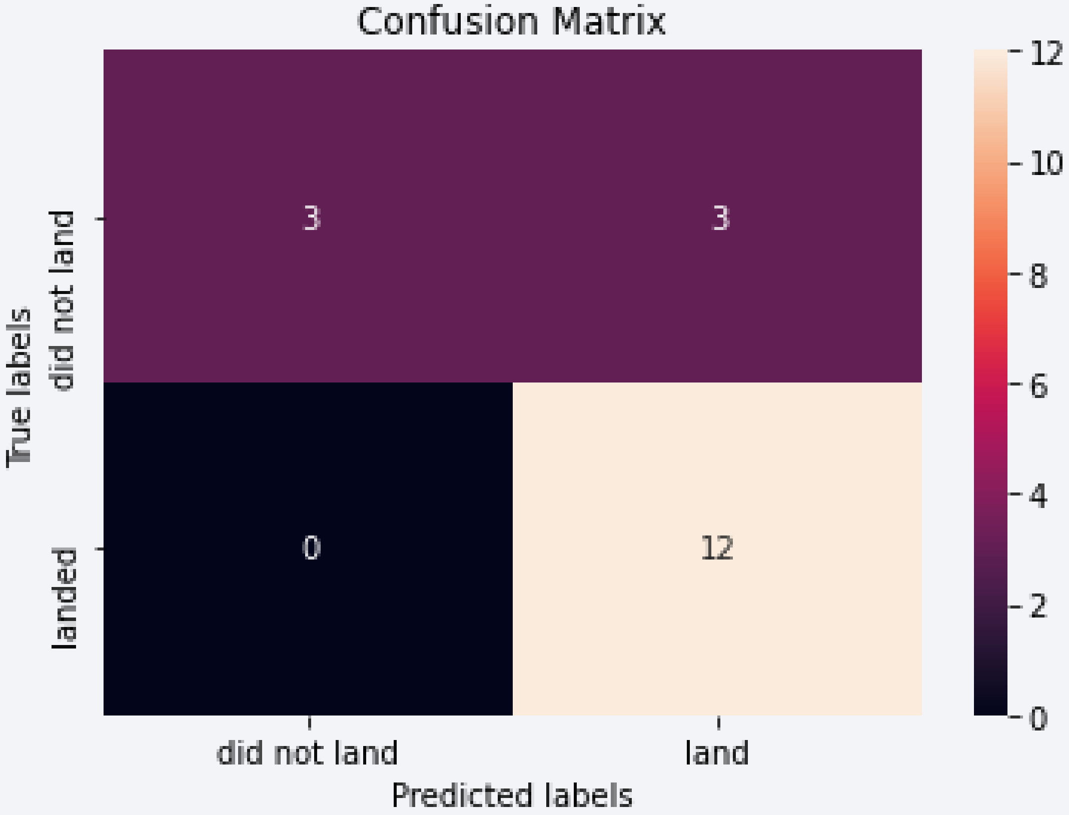


Confusion Matrix

- It has very high precision as 12 out of 15 positive prediction are true.

- The accuracy is 15 out of 18 which is very high too.

- The recall is 1.



Conclusions

- Features such as Launch site and pay load mass are important for the launch success.
- Machine learning model can be built to predict the success/failure of a launch.
- The best performing models are the support vector machine and k-nearest neighbor model with both good in-sample accuracy and out-of-sample accuracy.

Appendix

- Please find all relevant information using the GitHub link in the previous slides.

Thank you!