IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cong Thanh Vu
Sep 30th 2021

Outline

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

Executive Summary

- Summary of methodologies

  I mined data on SpaceX rocket launches for analysis, then I applied machine learning models to predict the success of the Falcon 9 first stage landing.

- Summary of all results

  The mined data gave me a lot of information about SpaceX's rocket launches, through which I also analyzed and extracted the characteristics of the launches to choose the right model to predict the success of the launches.

Introduction

- Project background and context

As working for SpaceY which is a competitor of SpaceX. I want to mine the data on SpaceX's Falcon 9 rocket launches for analysis to make strategies for the development of other SpaceY rocket technologies.

- Problems I want to find answers

I want to analyze the data to answer the hypothesis of whether there is a relationship between the known factors and the success of the Falcon 9 first stage landing. From there I will be able to plan and improve SpaceY's products.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data is collected by 2 methods: from SpaceX API and scraping data from Wikipedia
- Perform data wrangling
    - I convert the "Outcomes" column in the data from categorical data to numeric data and put it in the new column "Class"
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - I use GridSearchCV in the sklearn library to survey models with different params and then choose the model with the highest score on the test dataset.

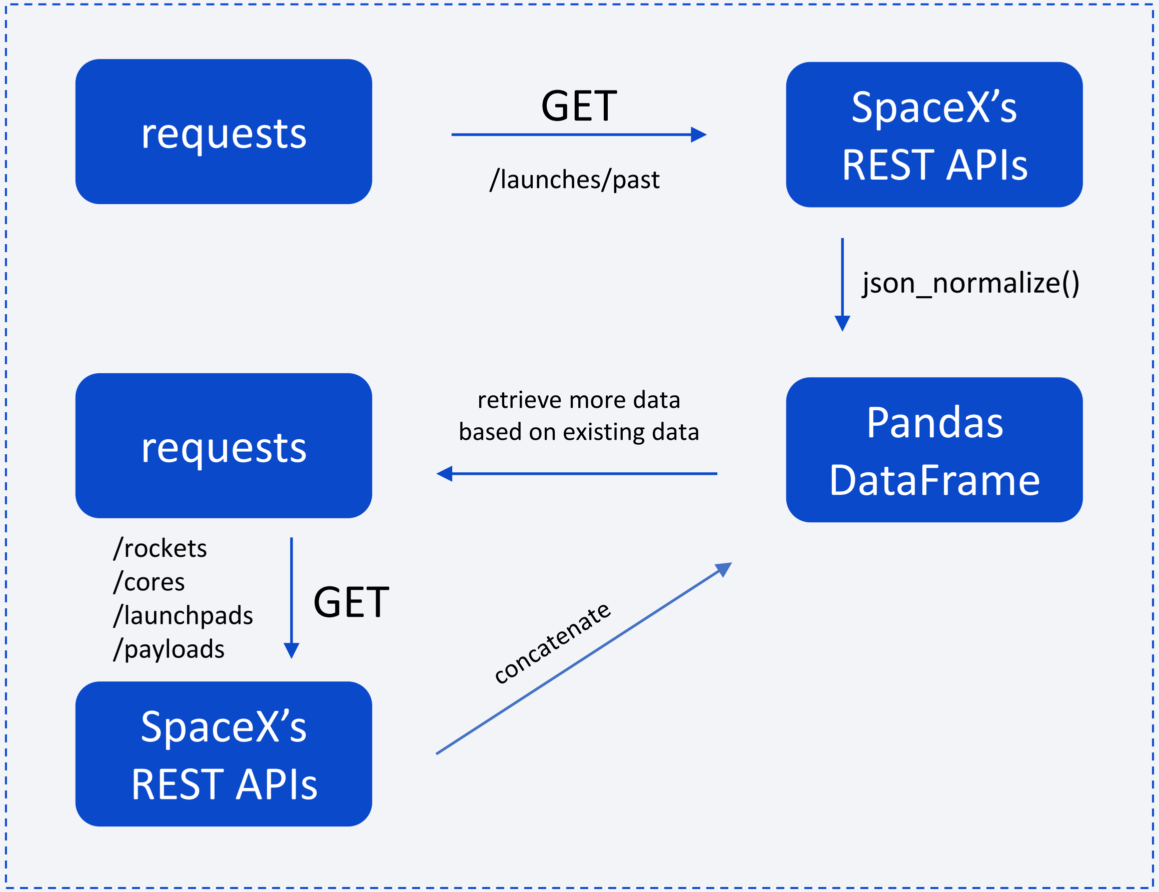
Data Collection

- Dataset are collected by 2 methods: through SpaceX REST API and web scraping on Wikipedia

Data Collection – SpaceX API

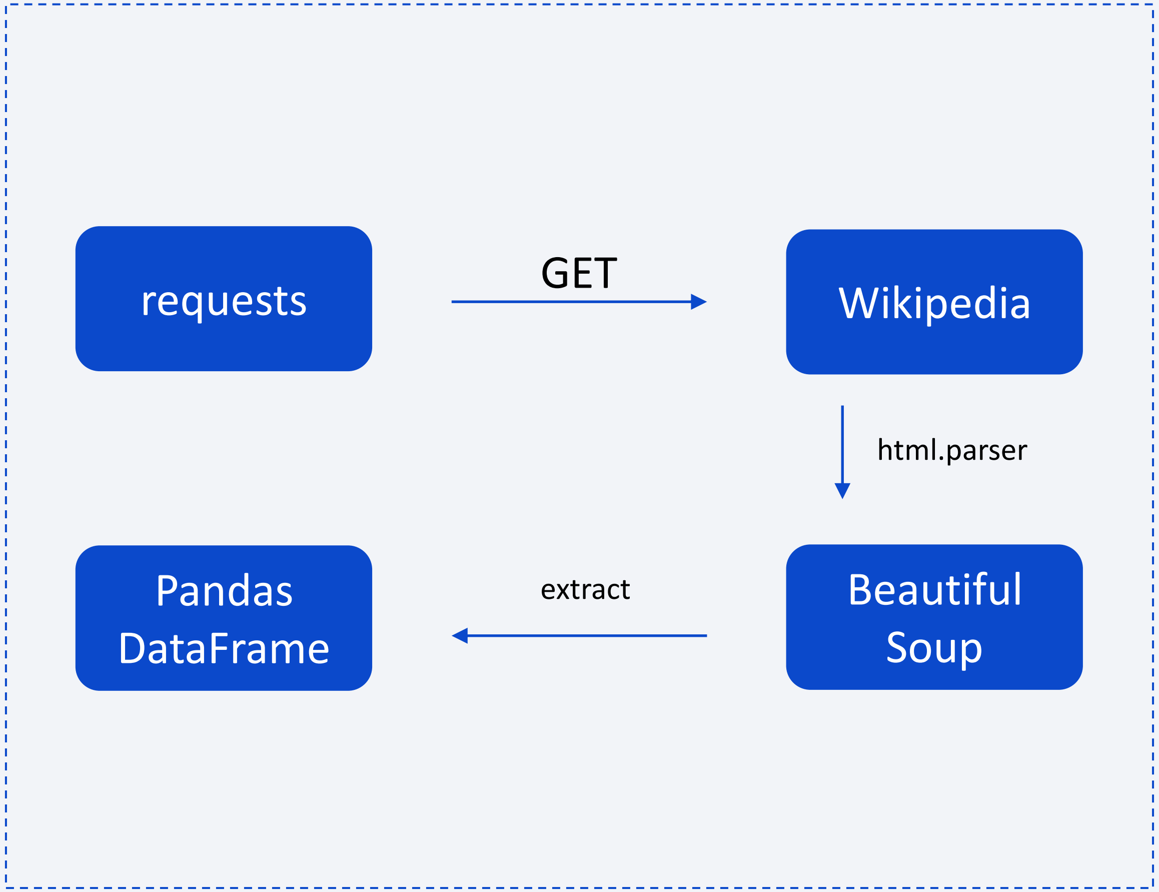
- Use requests library to call to REST API endpoints start with https://api.spacexdata.com/v4/

- GitHub URL of the notebook: Data Collection API Lab



Data Collection - Scraping

- Use requests library to get the Wikipedia page, then use Beautiful Soup to parse HTML and then extract data and put them into a dataframe
- GitHub URL of the notebook: [Data Collection with Web Scraping]()



Data Wrangling

- Filter the data so it should contain only Falcon 9 launches.
- Turn categorical data from “Outcome” column into numerical data with 1 means the booster successfully landed and 0 means it was not successful.
- GitHub URL: [Data Wrangling]()

EDA with Data Visualization

- Scatter charts and bar charts were plotted
- The scatter charts to show the relationship between features
- The bar charts are easy to understand and quickly figure out which orbit had the highest success rate
- GitHub URL: EDA with Data Visualization

EDA with SQL

- SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXDATASET
- SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE CUSTOMER='NASA (CRS)’
- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE BOOSTER_VERSION LIKE 'F9 v1.1%’
- SELECT MIN(DATE) FROM SPACEXDATASET WHERE LANDING__OUTCOME='Success (ground pad)’
- SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE LANDING__OUTCOME='Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000
- SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXDATASET GROUP BY MISSION_OUTCOME
- SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET)
- SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXDATASET WHERE LANDING__OUTCOME='Failure (drone ship)' AND DATE>='2015-01-01' AND DATE<='2015-12-31’
- SELECT LANDING__OUTCOME, COUNT(*) FROM SPACEXDATASET WHERE DATE>='2010-06-04' AND DATE<='2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT(*) DESC

- GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- Objects used: markers, circles, lines, MarkerCluster, MousePosition
- Markers: check the location on the map
- Circles: add a circle to the map
- Lines: represents the distance in crow's path between 2 points
- MarkerCluster: mark multiple point in the same location on the map
- MousePosition: track the position of the cursor

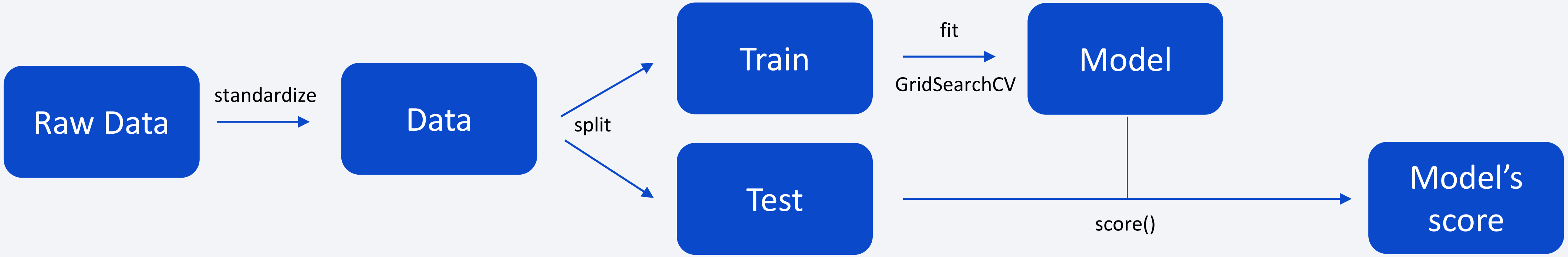
- GitHub URL: Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Items added: scatter chart, pie chart, dropdown, slider
- I added a scatter chart to plot the relationship between Payload Mass and Success Rate (class), a pie chart to plot the success/failed launch count, a dropdown to pick the site, and a slider to pick the payload range.
- GitHub URL: Build a Dashboard with Plotly Dash

Predictive Analysis (Classification)

- At first, I standardized the data, then use the train_split_test() to split the dataset into training set and test set, and then I use the GridSearchCV module from sklearn.model_selection to apply on different models and parameters to choose the best option. After that, I use the score() method to calculate the accuracy of the model on the test set.



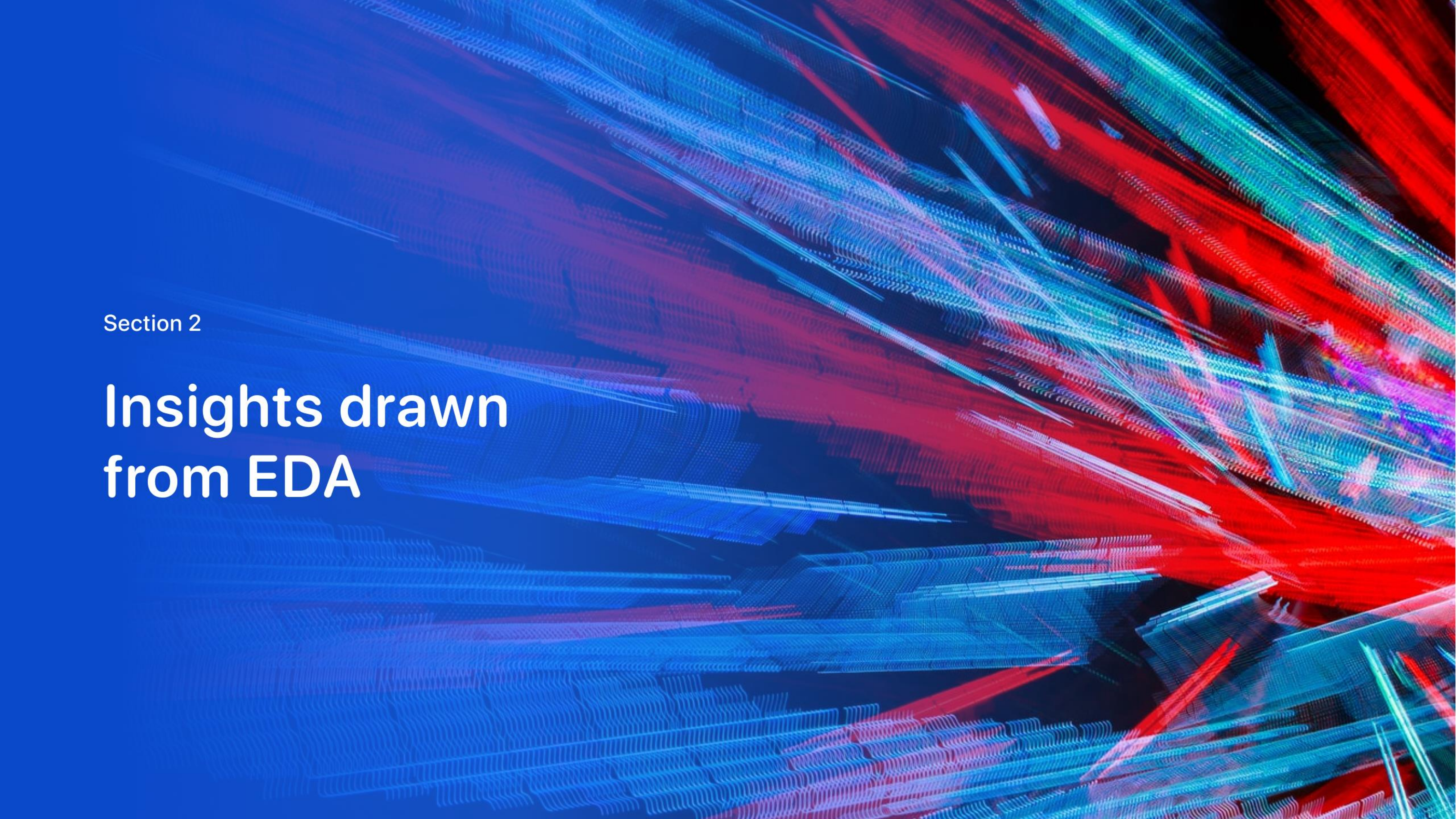
- GitHub URL: Predictive Analysis

Results

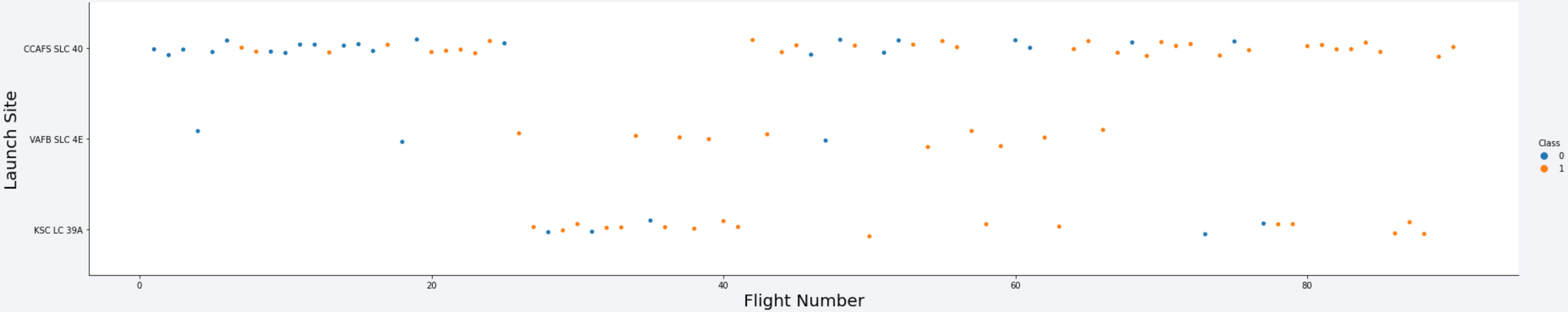
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

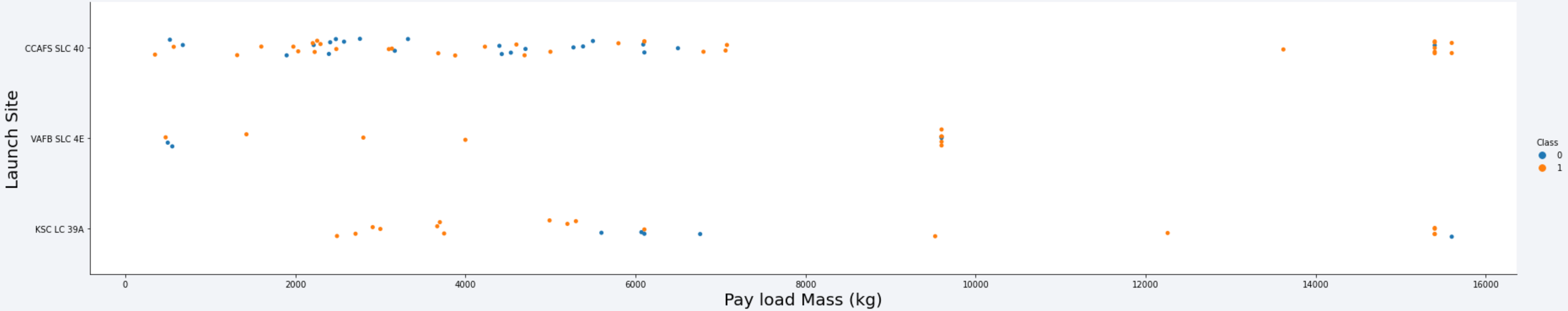


Flight Number vs. Launch Site



As the Flight Number increases, the first stage is more likely to be landed successfully. This fact is true for all of those launch sites

Payload vs. Launch Site

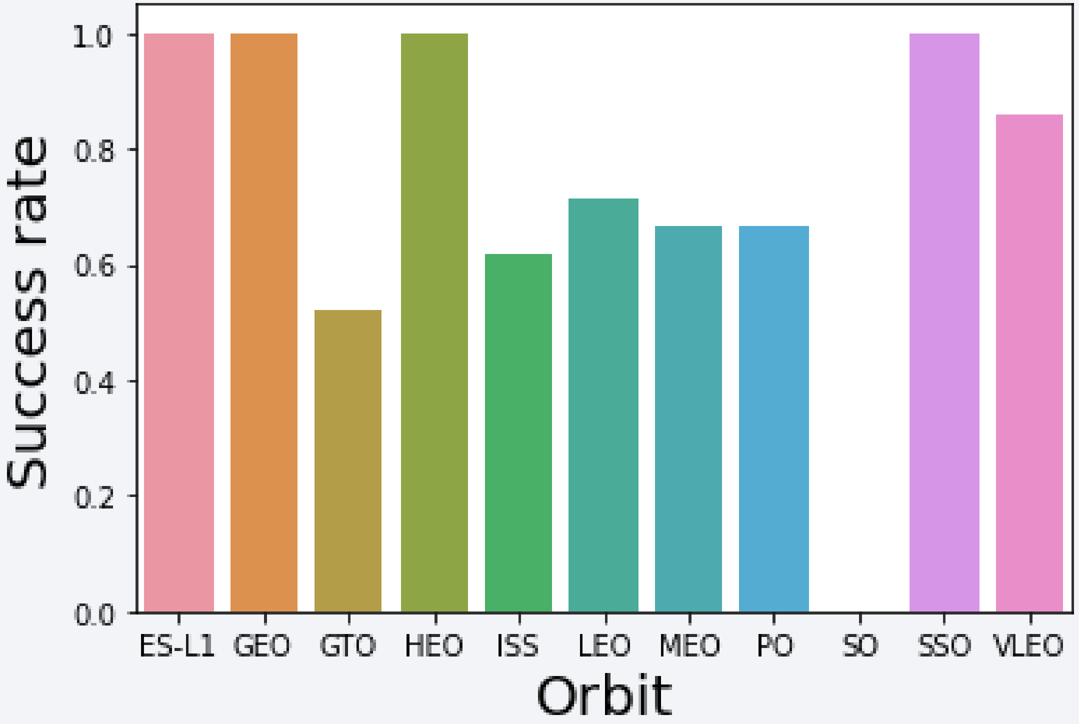


Launches with payload mass bigger than 8000kg are likely to have higher success rate.

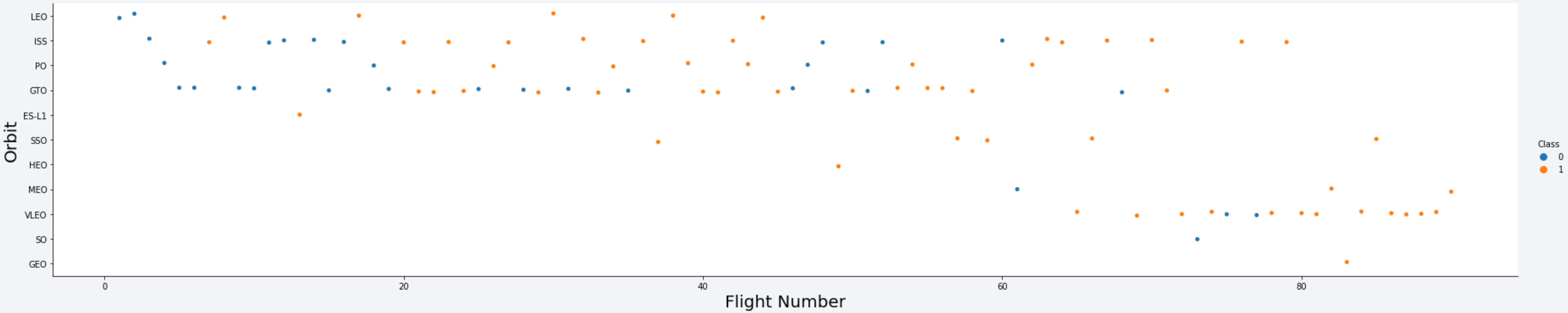
Launches in KSC LC 39A with low payload mass (lower than 5000kg) also work well.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO are the orbits which have the highest success rate.
- GTO has the worst success rate

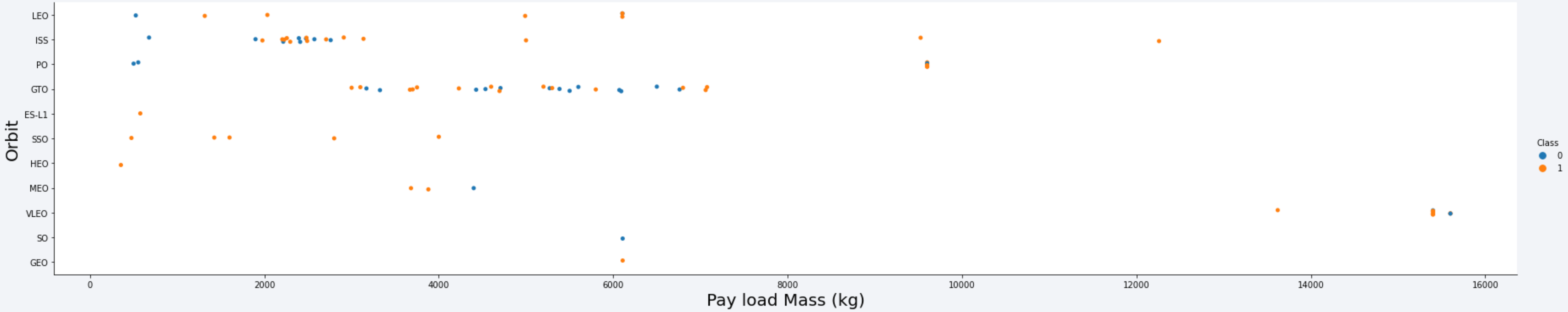


Flight Number vs. Orbit Type



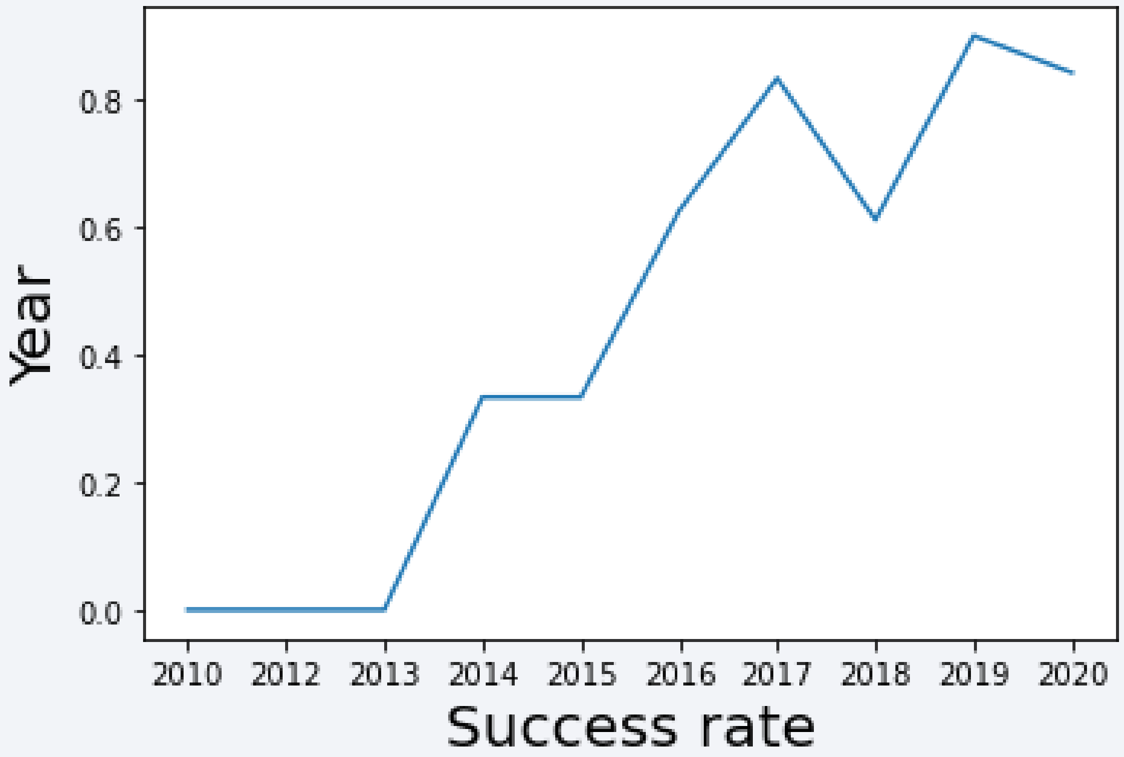
- LEO, VLEO: The higher flight number, the higher success rate
- ES-L1, GEO, HEO and SSO have the success rate of 100% but the number of launches is small

Payload vs. Orbit Type



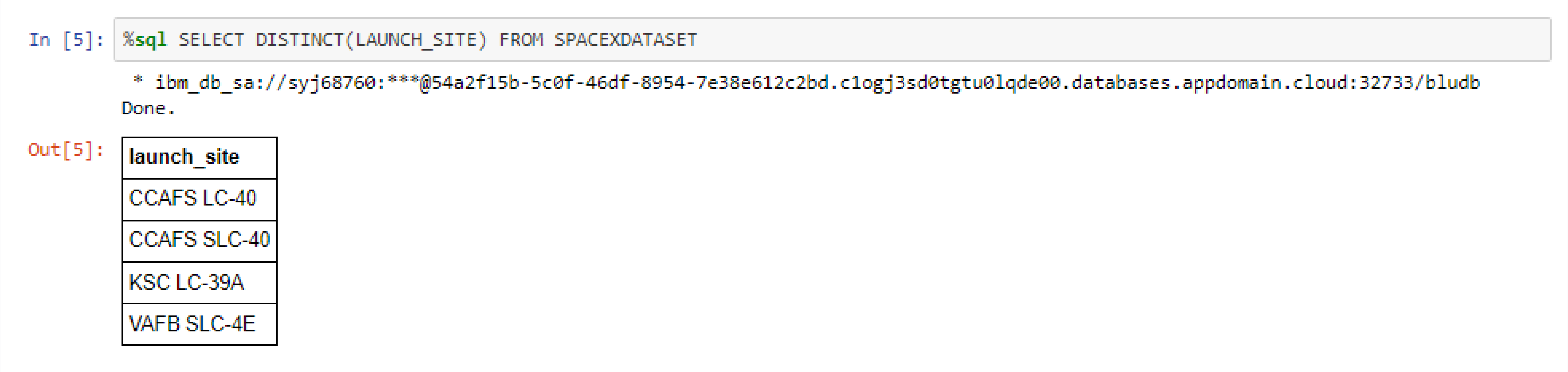
Higher payload mass are good for LEO, ISS (which have higher success rate on more heavy payloads). But it is bad for GTO.

Launch Success Yearly Trend



The success rate is increasing year by year thanks to the development of technology

All Launch Site Names



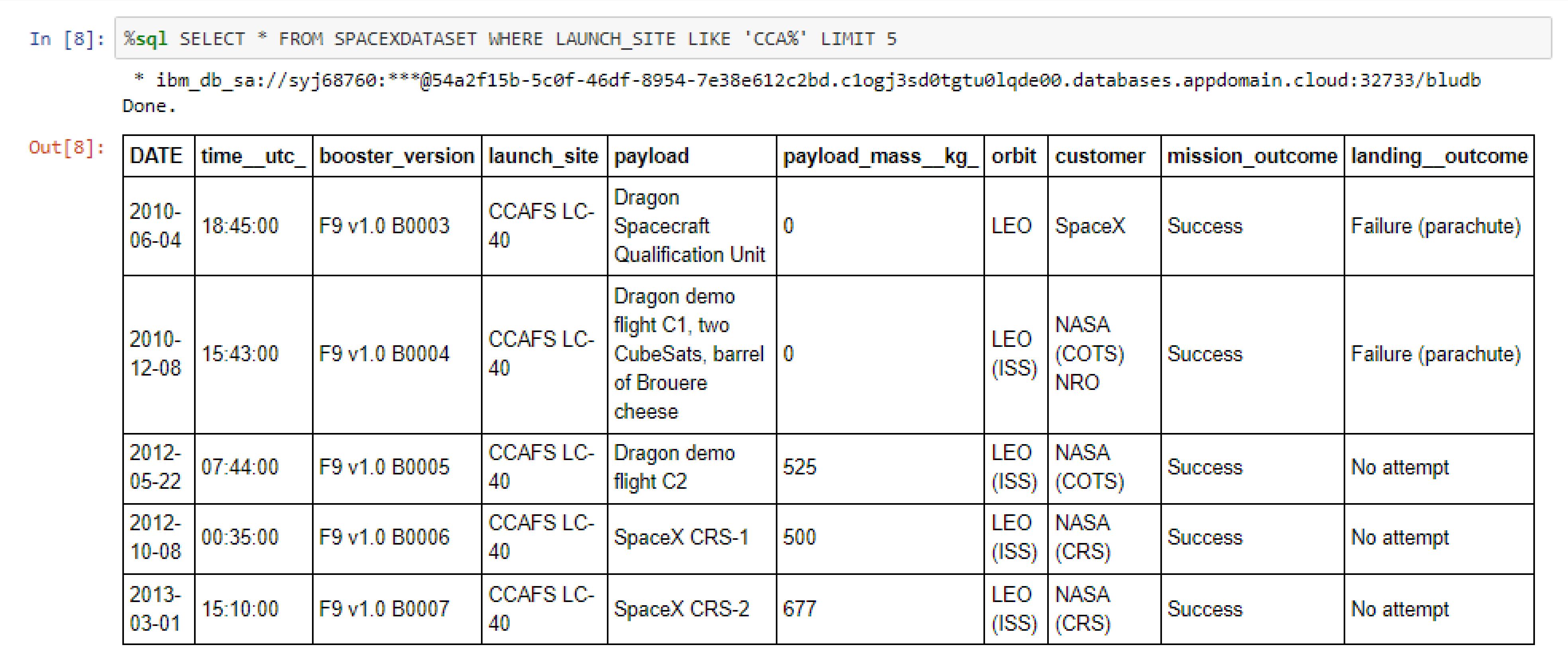
```
In [5]: %sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXDATASET

 * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[5]:

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

SELECT the LAUNCH_SITE column because the name of launch sites are record in that column. I used DISTINCT because, there are many launches that can be launched in the same launch site. So DISTINCT gives me the list of unique launch site names

Launch Site Names Begin with 'CCA'



```
In [8]: %sql SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

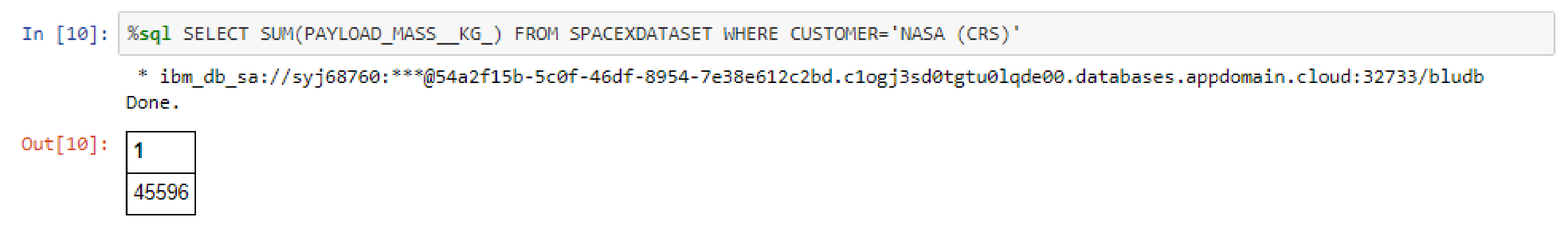
Out[8]:

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

Condition: LAUNCH_SITE LIKE ‘CCA%’ means the launch site names need to start with CCA

LIMIT 5: just print the first 5 results

Total Payload Mass



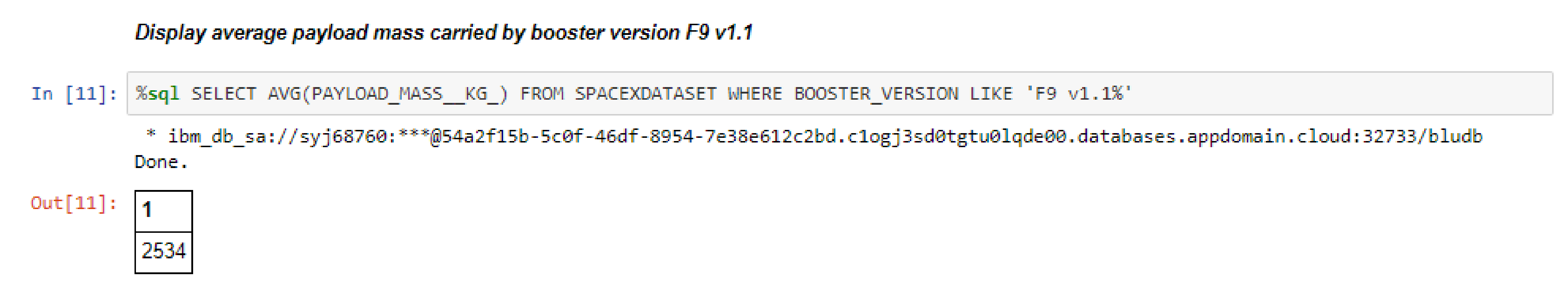
```
In [10]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE CUSTOMER='NASA (CRS)'

 * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

Out[10]:
```

1
45596

SUM: calculate the total payload mass

CUSTOMER=‘NASA (CRS)’: all booster launched by NASA

Average Payload Mass by F9 v1.1



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

```
In [11]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'

          * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.
```

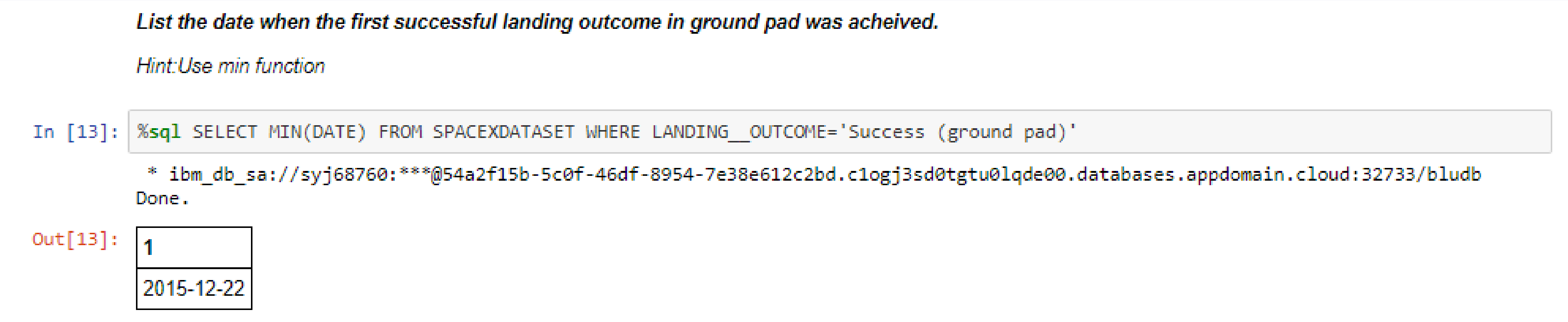
Out[11]:

1
2534

AVG(PAYLOAD_MASS__KG_): calculate the average payload mass

Because the BOOSTER_VERSION column have the form like ‘F9 v1.1 XXXXX’ so that I use BOOSTER_VERSION LIKE ‘F9 v1.1%’ to find all booster version of F9 v1.1

First Successful Ground Landing Date



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

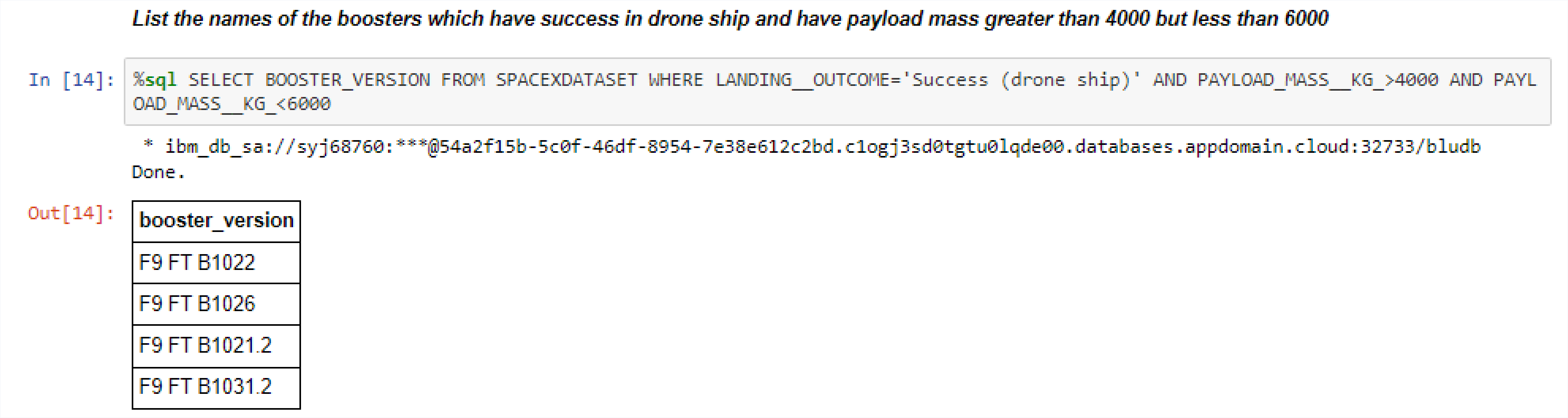
Hint:Use min function

```
In [13]: %sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE LANDING__OUTCOME='Success (ground pad)'

 * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

Out[13]:
```

1
2015-12-22

The first successful landing outcome has the minimum date, so that I use MIN(DATE) to print the date.

Successful Drone Ship Landing with Payload between 4000 and 6000



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

```
In [14]: %sql SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE LANDING__OUTCOME='Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYL
         OAD_MASS__KG_<6000

          * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.
```

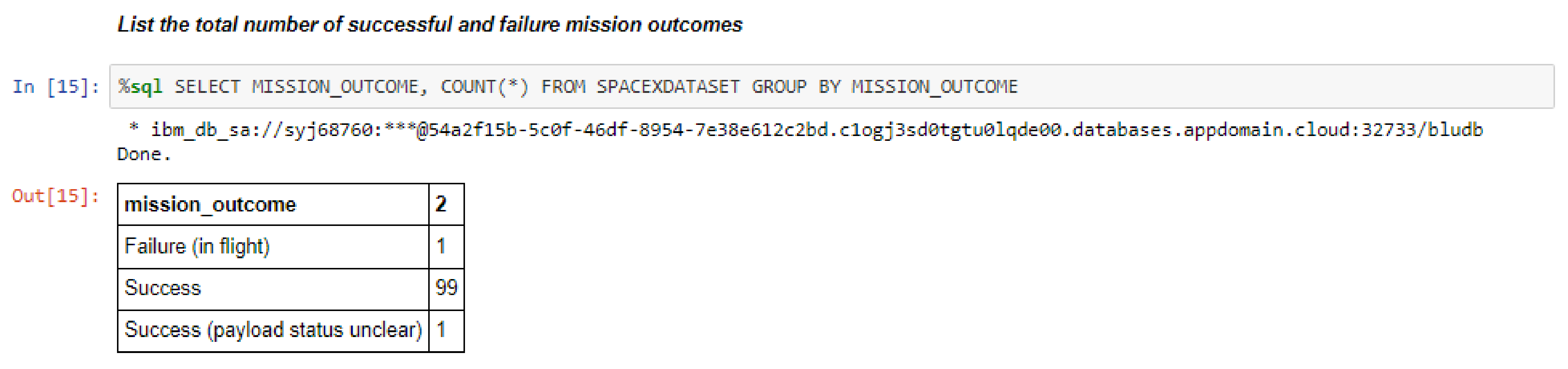
Out[14]:

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

Add condition: PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

It means that the result's payload should be in the range [4000, 6000]

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

```
In [15]: %sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXDATASET GROUP BY MISSION_OUTCOME
```

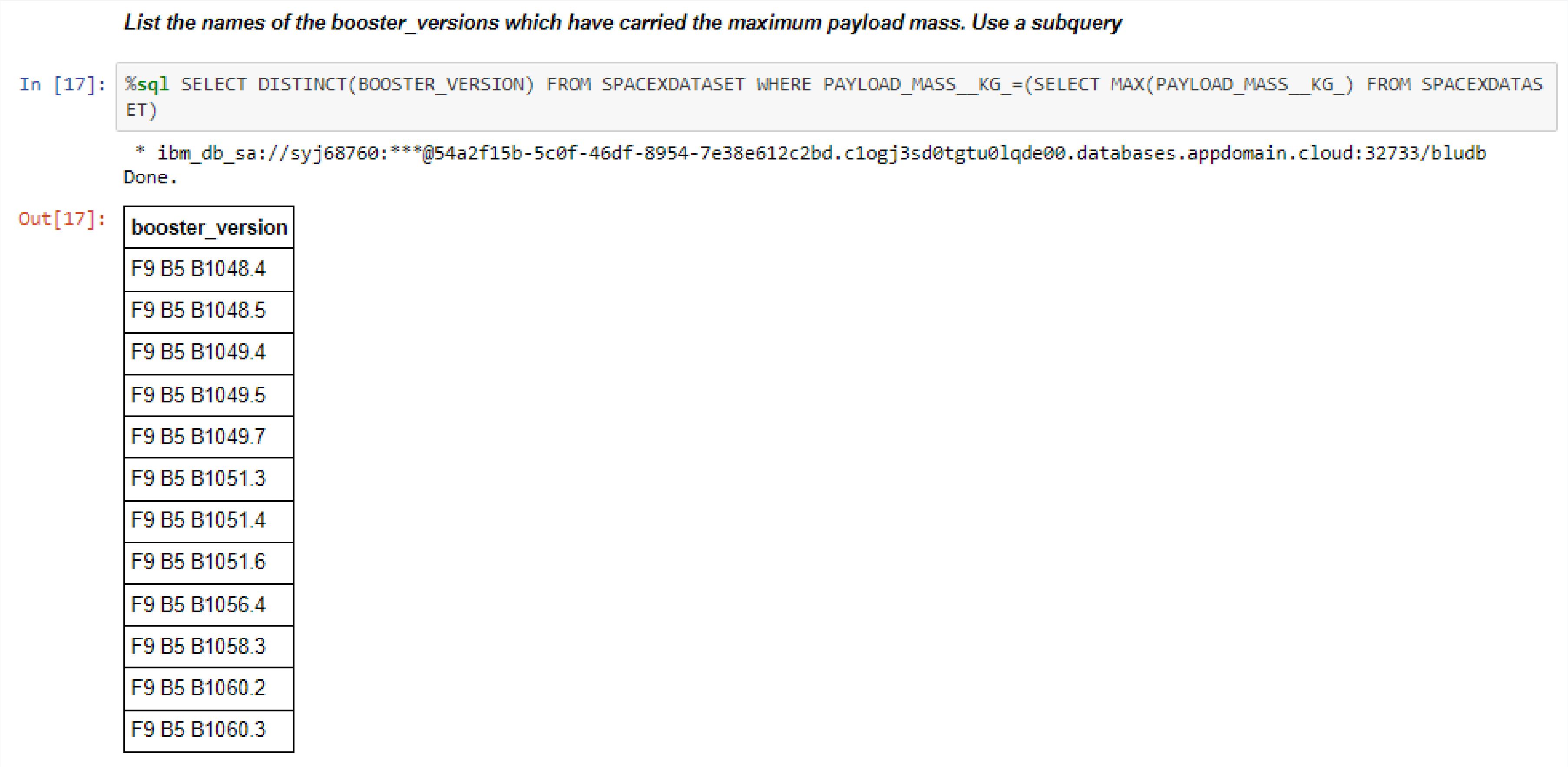
```
 * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[15]:

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

Group the table by MISSION_OUTCOME and count the records in each group.

Boosters Carried Maximum Payload



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

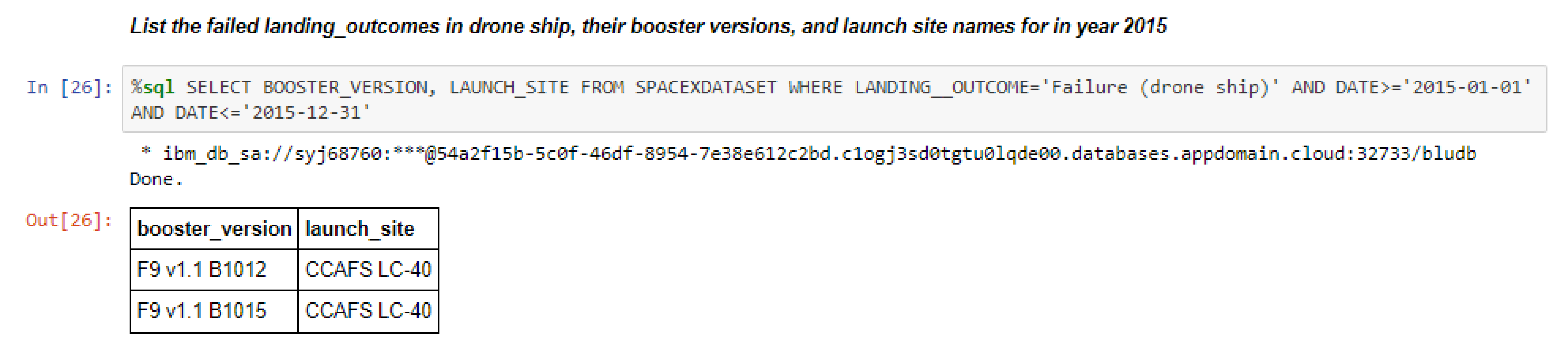
```
In [17]: %sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATAS
         ET)

          * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.
```

Out[17]:

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

Use a subquery to query the max payload and then add it to the condition of the parent query

2015 Launch Records



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

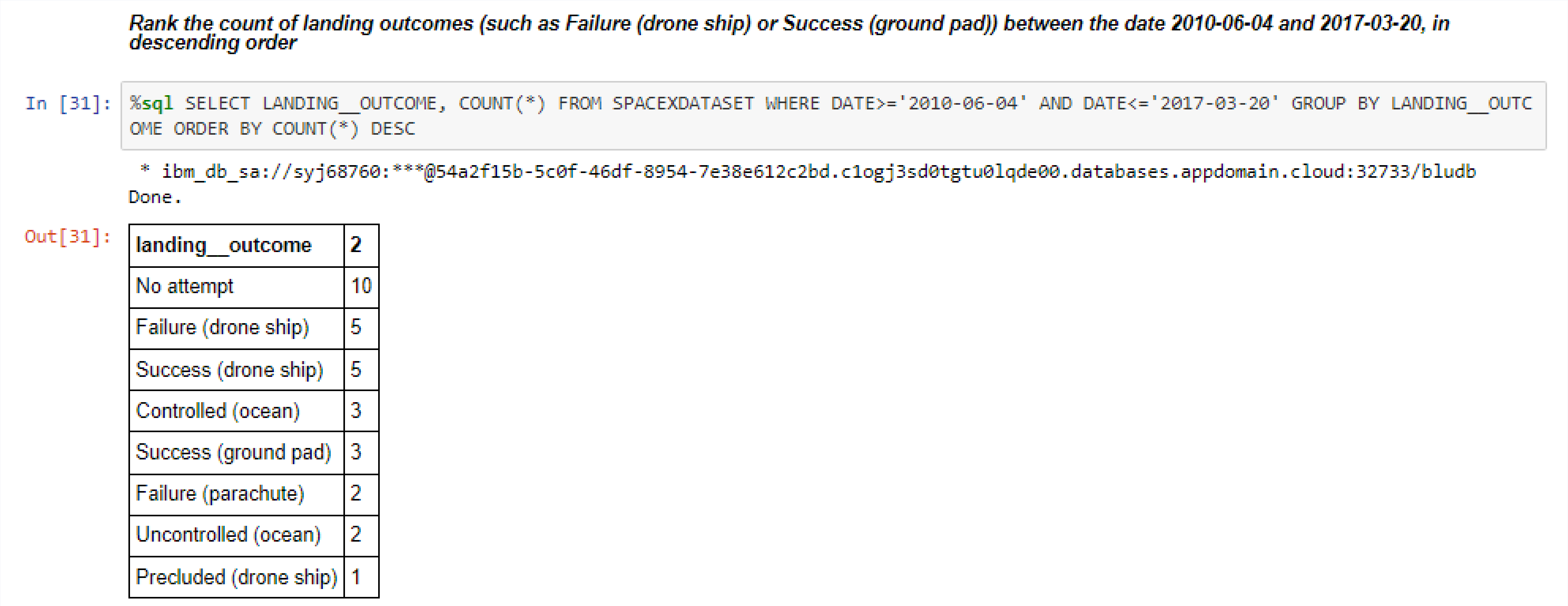
```
In [26]: %sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXDATASET WHERE LANDING__OUTCOME='Failure (drone ship)' AND DATE>='2015-01-01'
         AND DATE<='2015-12-31'

          * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.
Out[26]:
```

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

In year 2015 means the DATE should be between Jan 1st 2015 and Dec 31st 2015

So I added it to the condition. The LANDING_OUTCOME='Failure (drone ship)' to query launches failed in drone ship

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 [31]: %sql SELECT LANDING__OUTCOME, COUNT(*) FROM SPACEXDATASET WHERE DATE>='2010-06-04' AND DATE<='2017-03-20' GROUP BY LANDING__OUTC
         OME ORDER BY COUNT(*) DESC

          * ibm_db_sa://syj68760:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.
```

Out[31]:

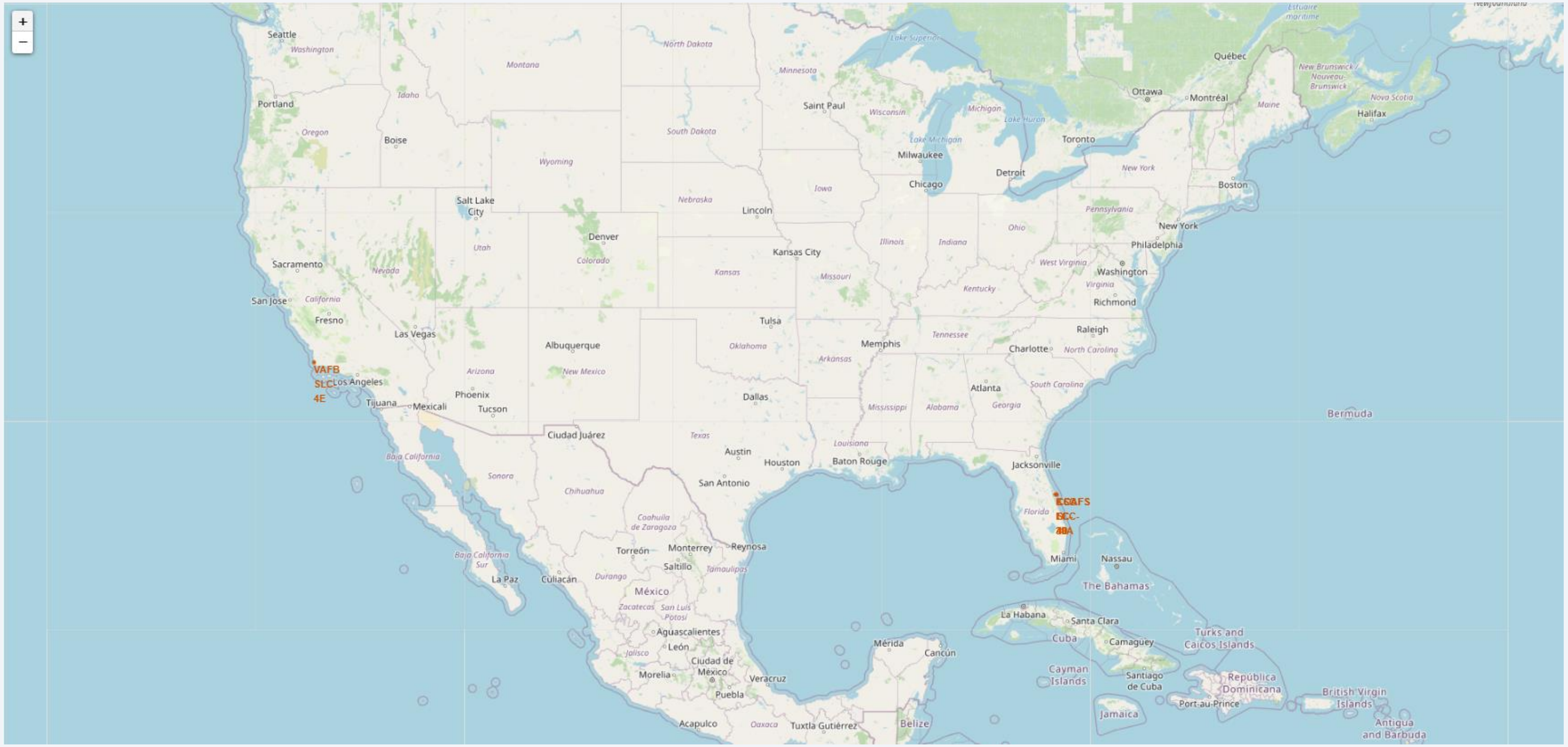
landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Filter the DATE between 2010-06-04 and 2017-03-20, then group the table by LANDING_OUTCOME, then sort by the count of landing outcomes

Section 4

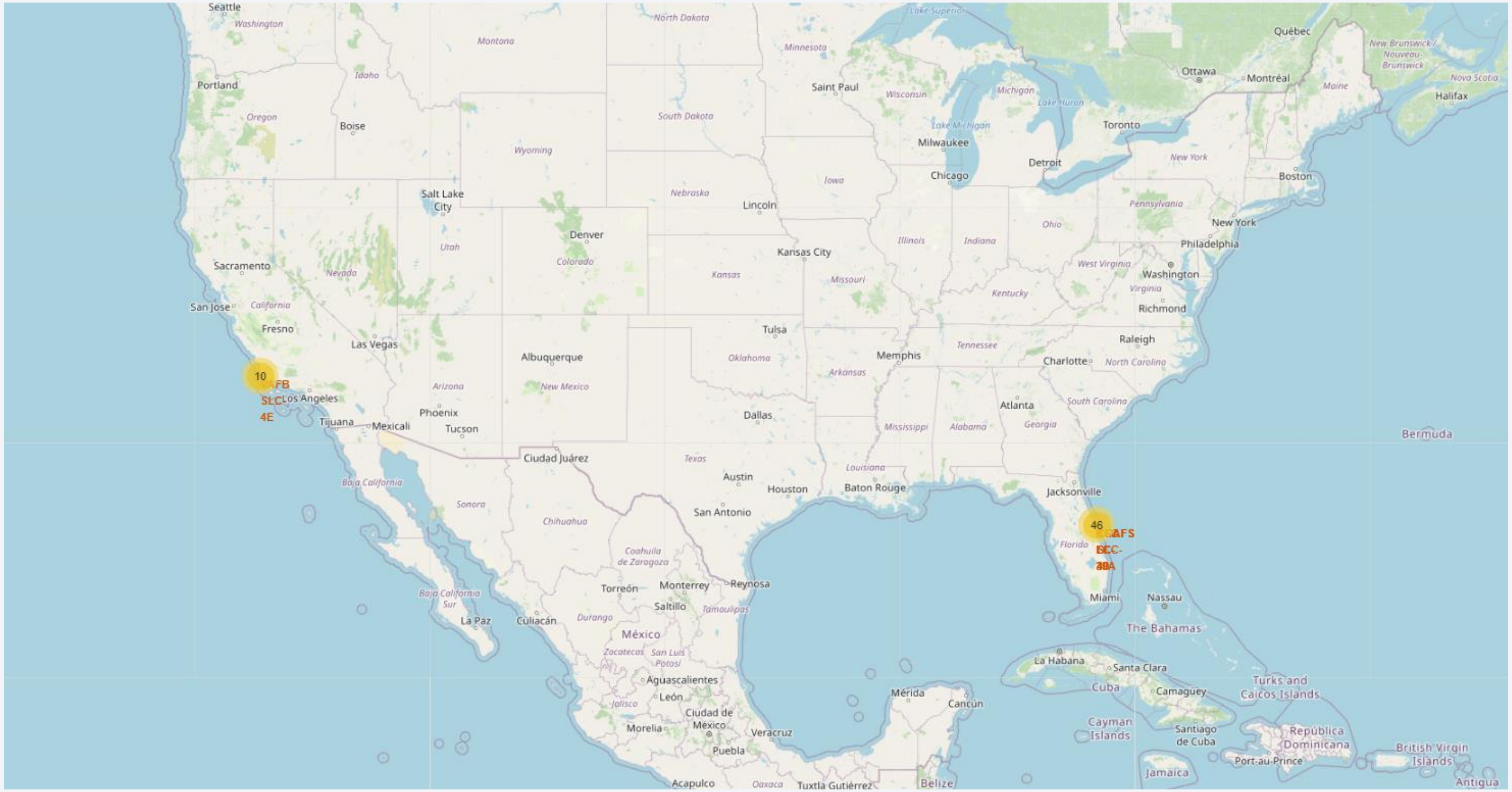
Launch Sites Proximities Analysis

Launch sites map



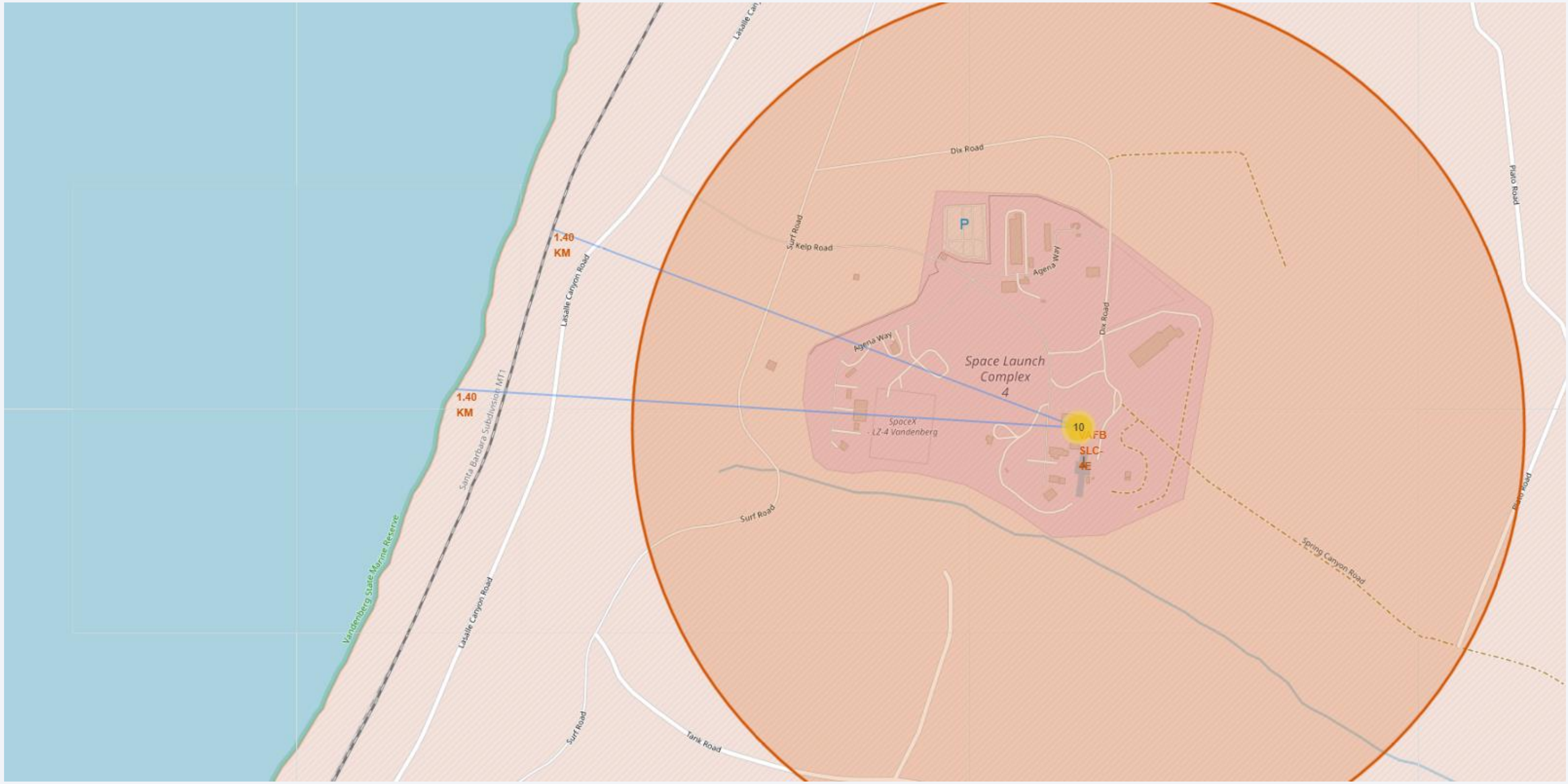
Launch sites are located near the coast

Color-labeled launch outcomes

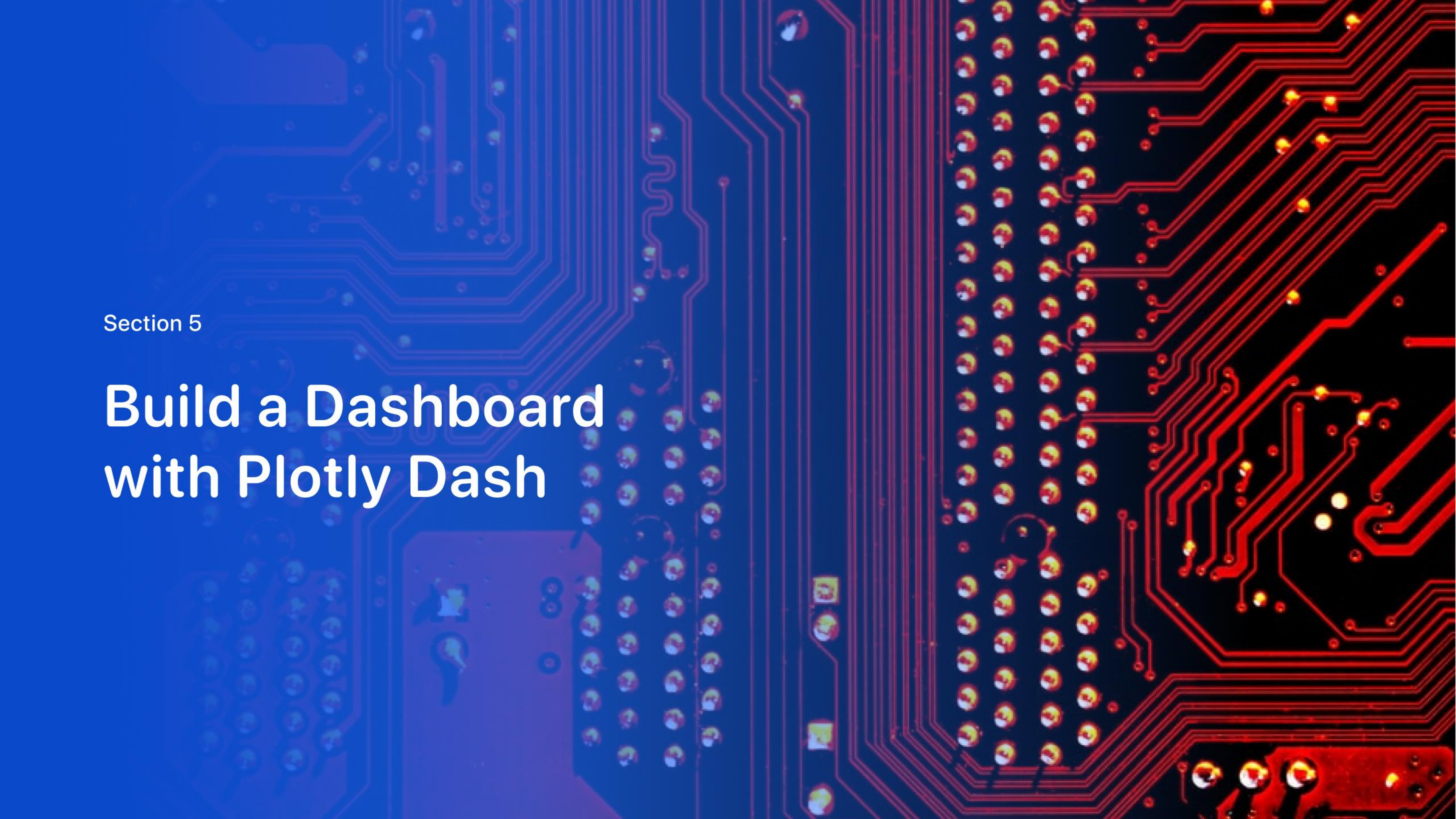


The number of launch outcomes in the launch sites on the right is more than on the left side. Both have roughly equal success rate (based on color)

Proximities distance



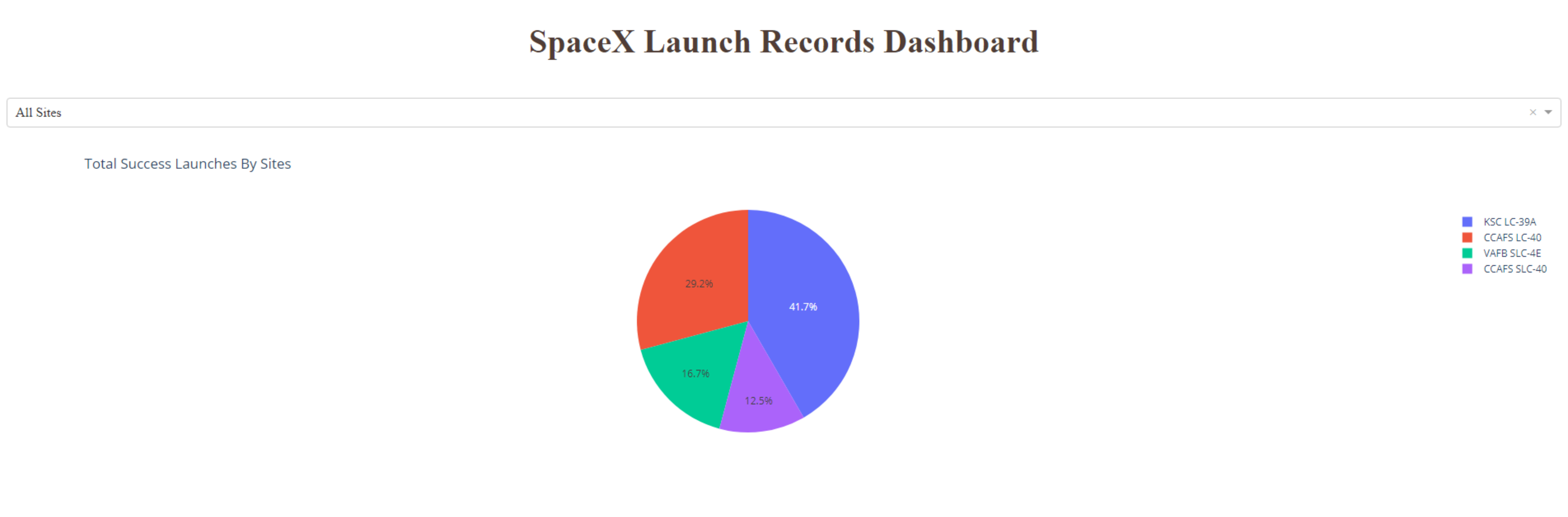
The distance from the launch sites to the railway and the coastline is not too far, just over 1km



Section 5

Build a Dashboard with Plotly Dash

Success count for all sites



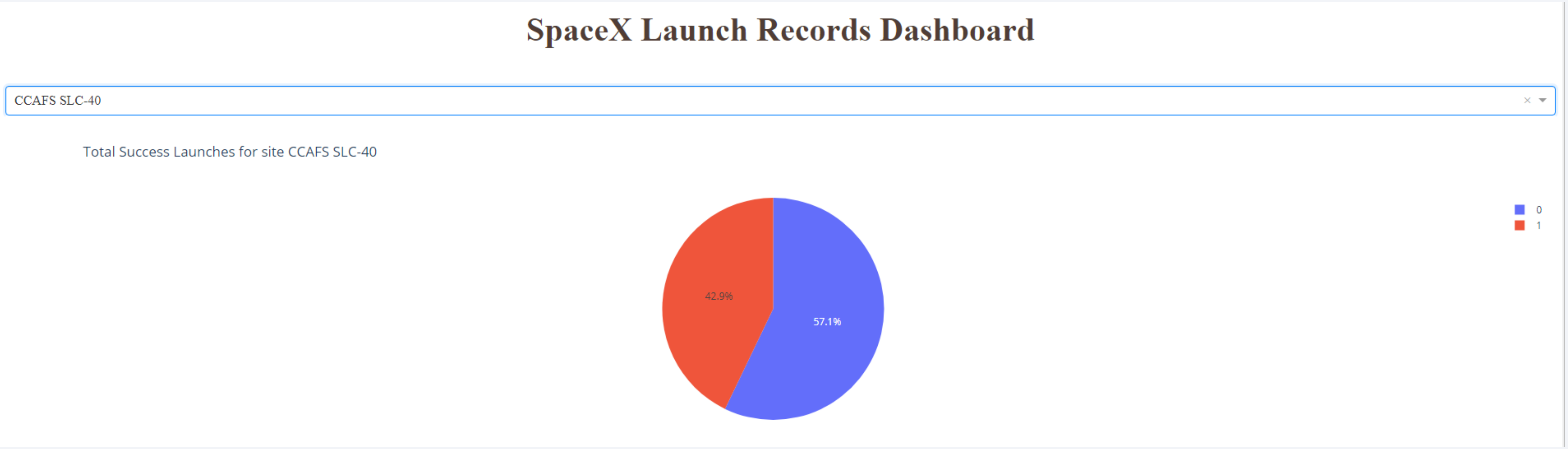
SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Sites

The KSC LC-39A (blue) has the largest number of successful launches (41.7% of all successful launches). The CCAFS SLC-40 (purple) has the smallest number of successful launches (12.5%)

Highest success rate launch site



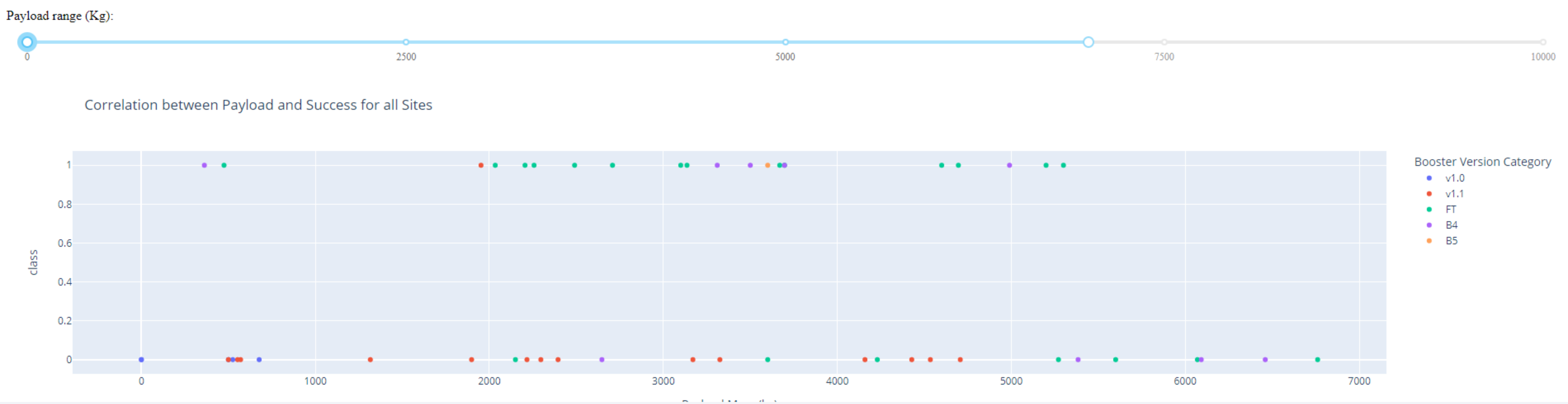
SpaceX Launch Records Dashboard

CCAFS SLC-40

Total Success Launches for site CCAFS SLC-40

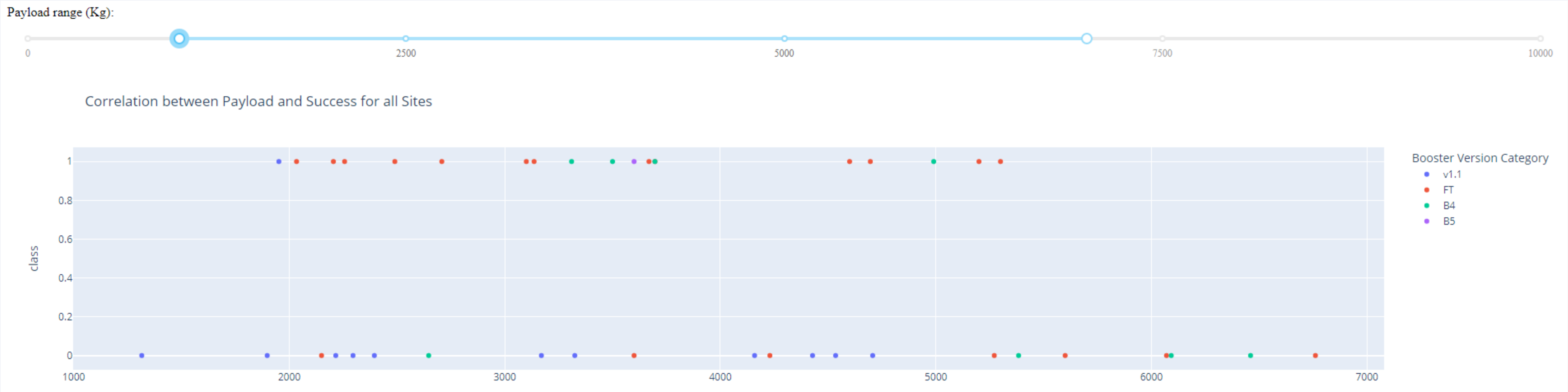
The launch site has the highest success rate is CCAFS SLC-40, 42.9% launches are successfully landed.

Payload vs. Launch Outcome



There are 5 categories of Booter Version: v1.0, v1.1, FT, B4, B5. The FT Booster Version have highest success rate, as you can see there are so many green dots in the class 1

Payload vs. Launch Outcome



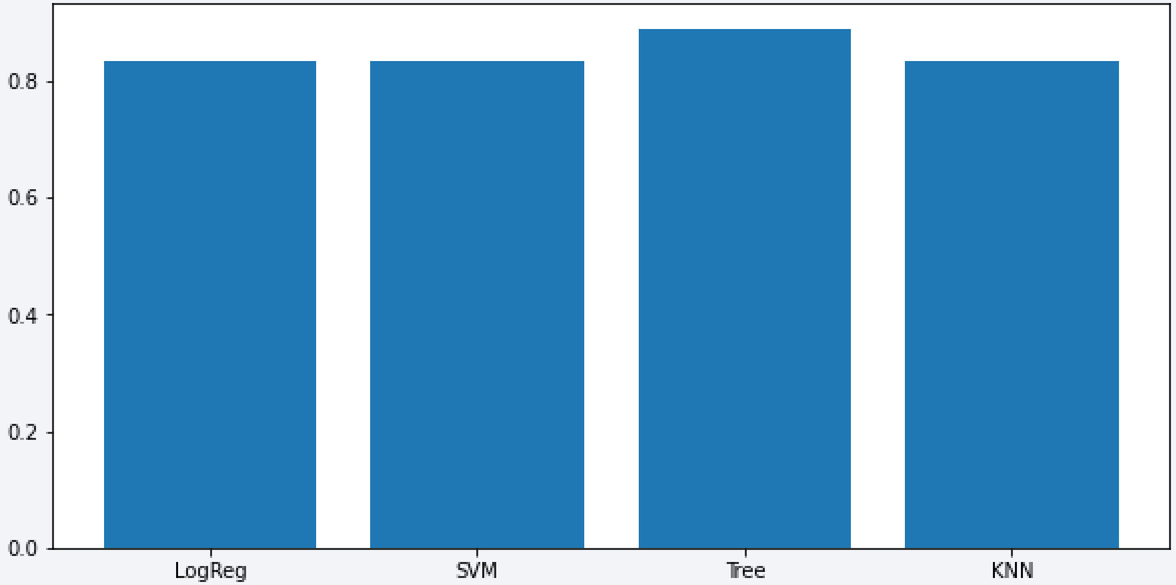
When change the slider to the range of 1000-7000, there is no launch with the booster version of v1.0, so now there are 4 categories of booster version

Section 6

Predictive Analysis (Classification)

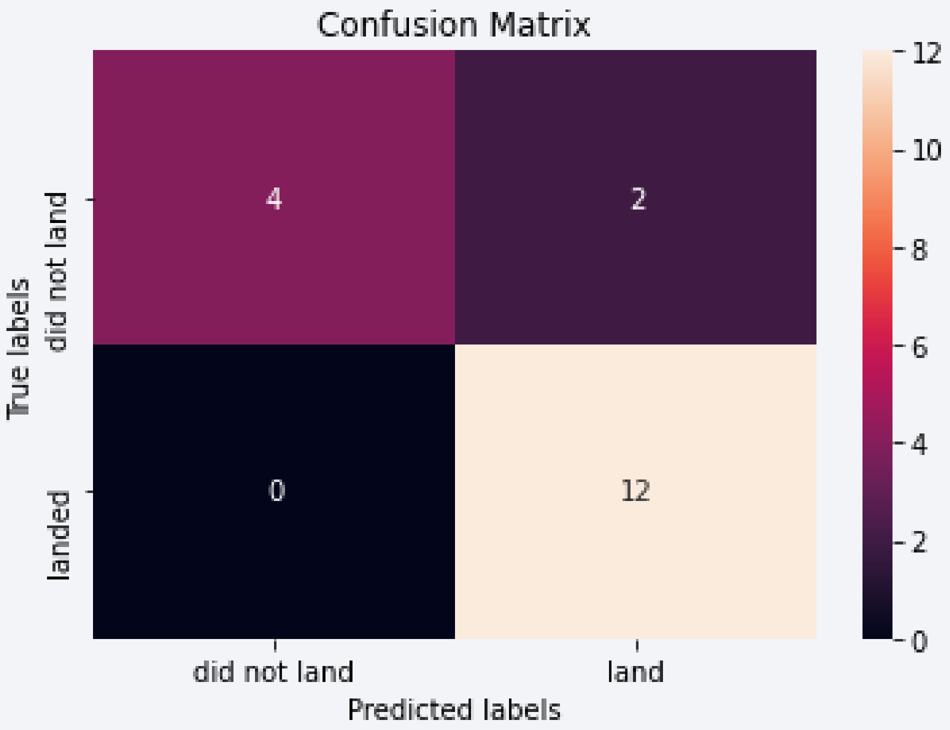
Classification Accuracy

- Decision tree model has accuracy on test set of 88.88% while other models has accuracy of 83.33%
- So the best model based on its accuracy on test set is decision tree model



Confusion Matrix

- The best model (Decision Tree) prediction contains 2 samples of false positive (the upper right tile), 4 samples of true negative, and the remaining 12 samples are true positive
- All landed launches have been predicted correctly by the model



Conclusions

- ES-L1, GEO, HEO and SSO are the orbits which have the highest success rate.
- Success rate of launches increases year by year
- 99/101 mission outcomes are successful
- The KSC LC-39A has the largest number of successful launches
- The launch site has the highest success rate is CCAFS SLC-40 with 42.9% of success
- The best model is the decision tree model with 88% accuracy on test set.

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!