IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Henry Wong
Oct 18, 2023

Outline

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

Executive Summary

- We collected historical launch data of Space X from its open API, explored the data using different techniques, and properly transformed it to fit the data to different prediction model
- To predict whether future launches of SpaceX will have a successful first stage landing, we determined it is best to make use a decision tree prediction model, which makes use of a number of features for the prediction, including booster version, payload mass, destination orbit, and launch site used.

Introduction

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

We would like to use Machine Learning approach to predict if a first stage rocket will land given its operating parameters.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using an open API provided by SpaceX
- Perform data wrangling
  - To facilitate subsequent steps, we converted outcomes in the data with `1` means the booster successfully landed, and `0` means it was unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression; and then find the method performs best using test data

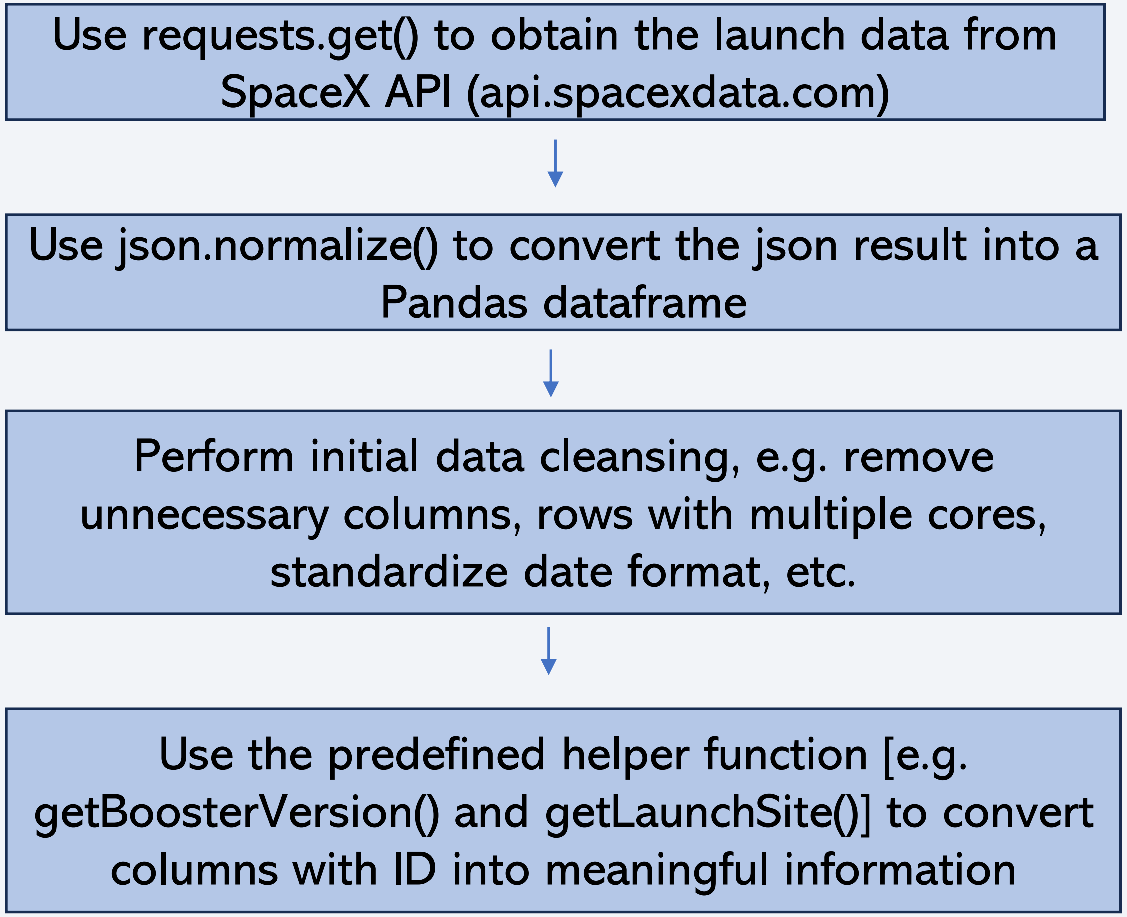
Data Collection

1. Data collection was done using get request to the SpaceX API
2. We then decode the response contents as a JSON using .json(), and turn it into a Pandas data frame using json_normalize()
3. As some fields are in IDs, we use the API again to get more specific launch information, including rocket, payloads, launch pad, cores, and booster version
4. We then cleaned the data, checked for missing values and fill in missing values where

Data Collection – SpaceX API

- The data from SpaceX API is collected the cleaned using the steps illustrated by the flow chart on the right.

https://github.com/iamhenrywong/Coursera-data-science-capstone/blob/8d8bb1715180190e70bce52897bc075628fe8898/jupyter-labs-spacex-data-collection-api.ipynb



Use requests.get() to obtain the launch data from SpaceX API (api.spacexdata.com)

Use json.normalize() to convert the json result into a Pandas dataframe

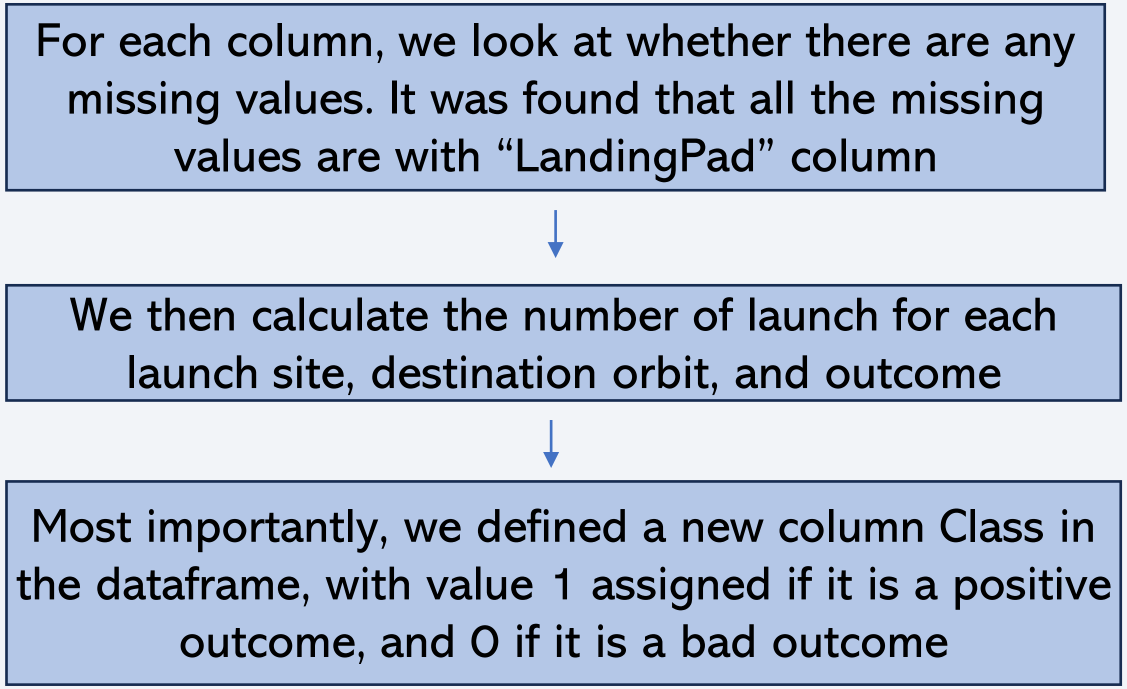
Perform initial data cleansing, e.g. remove unnecessary columns, rows with multiple cores, standardize date format, etc.

Use the predefined helper function [e.g. getBoosterVersion() and getLaunchSite()] to convert columns with ID into meaningful information

Data Wrangling

- We performed Data Wangling with the process described in the flow chart on the right

https://github.com/iamhenrywong/Coursera-data-science-capstone/blob/8d8bb1715180190e70bce52897bc075628fe8898/labs-jupyter-spacex-Data%20wrangling.ipynb



For each column, we look at whether there are any missing values. It was found that all the missing values are with "LandingPad" column

We then calculate the number of launch for each launch site, destination orbit, and outcome

Most importantly, we defined a new column Class in the dataframe, with value 1 assigned if it is a positive outcome, and 0 if it is a bad outcome

EDA with Data Visualization

- We used scattered plot to see, initially, whether there is any relationship between these factors: flight number, payload mass, launch site, and destination orbit type.
- In general, the success rate of landing for first stage has been increasing from 2013 to 2020

https://github.com/iamhenrywong/Coursera-data-science-capstone/blob/8d8bb1715180190e70bce52897bc075628fe8898/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The data was loaded onto IBM Db2 on cloud, to enable exploratory analysis using SQL
- The below SQL queries were performed:
    - display names of the unique launch sites, and selected records from one of the sites
    - display total payload mass carried for launches commissioned by NASA
    - display average payload mass carried by a certain booster
    - list the date when the first successful landing outcome in ground pad was achieved
    - list the names of the boosters which have success in drone ship and have a specific payload mass range
    - list the total number of successful and failure mission outcomes
    - list the names of the booster versions which have carried the maximum payload mass, using a subquery
    - list the failed landing in drone ship, their booster versions, and launch site names for in year 2015
    - rank the count of landing outcomes in a specific period

https://github.com/iamhenrywong/Coursera-data-science-capstone/blob/8d8bb1715180190e70bce52897bc075628fe8898/jupyter-labs-eda-sql-coursera%20(completed).ipynb

Build an Interactive Map with Folium

- We added below objects to the relevant map:
  - circle and marker: to mark the launch sites
  - marker cluster: to mark the successful/fail launches at each site
  - calculate the mark the distance from the launch sites to proximities, such as coastline, railway, highway, and city

- The reasons we added these objects are:
  - visualize the launches and their success/failure more easily
  - understand whether proximity to other transport infrastructure is a necessary condition for a launch site

https://github.com/iamhenrywong/Coursera-data-science-capstone/blob/8d8bb1715180190e70bce52897bc075628fe8898/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Two interactive charts were created:
  - a pie chart that shows the success rate of all or one specific launch site, at user's selection.
  - a scattered plot for success/failure vs. the payload mass, for all or one specific launch site. In addition, user could zoom in a specific payload mass range.
- These charts would allow users to zoom into specific launch site and payload range, get the success rate visuals instantly

https://github.com/iamhenrywong/Coursera-data-science-capstone/blob/8d8bb1715180190e70bce52897bc075628fe8898/Spacex-dash.py

Predictive Analysis (Classification)

- The properly-transformed SpaceX launch data was split into training and testing data set.
- We then used applied below predictive models use Grid Search approach to tune the parameters
- Finally we compared the accuracy across different models to identify the best one

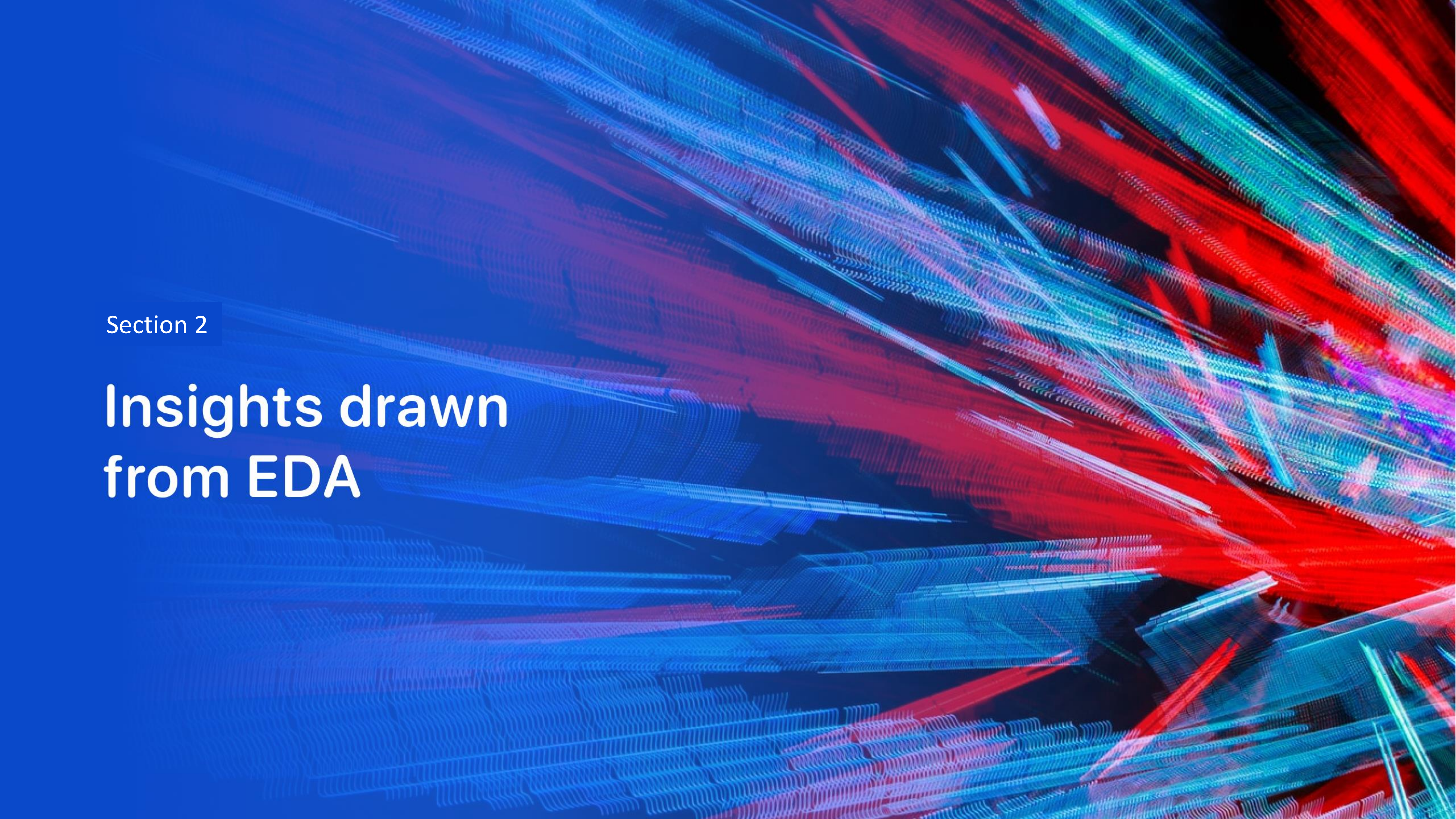
https://github.com/iamhenrywong/Coursera-data-science-capstone/blob/8d8bb1715180190e70bce52897bc075628fe8898/SpaceX_Machine_Learning_Prediction%20(completed).ipynb

Results

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

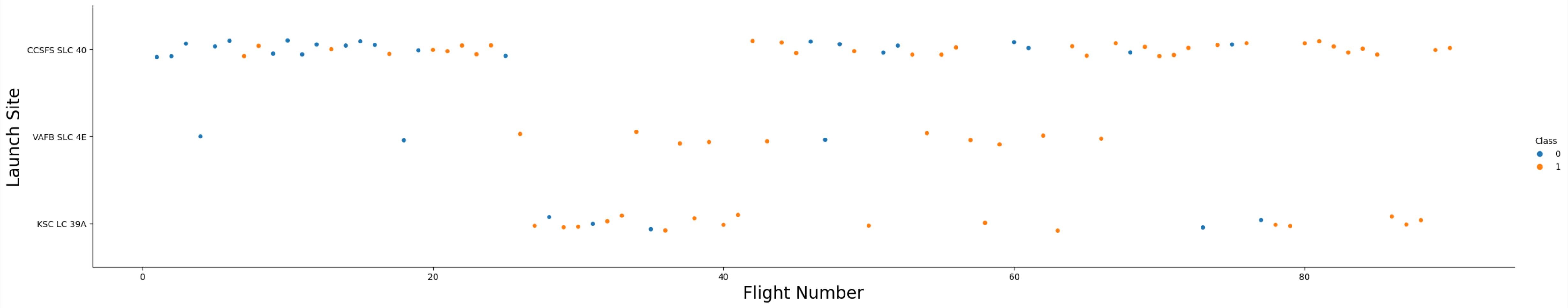
Section 2

Insights drawn from EDA



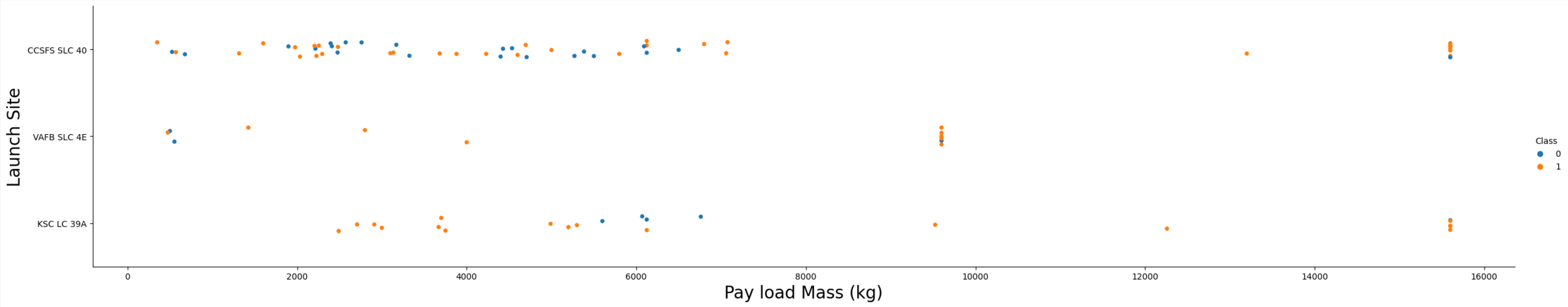
Flight Number vs. Launch Site

- It seems that for all 3 launch sites, the success rate has been increasing with flight number (i.e. later flights)



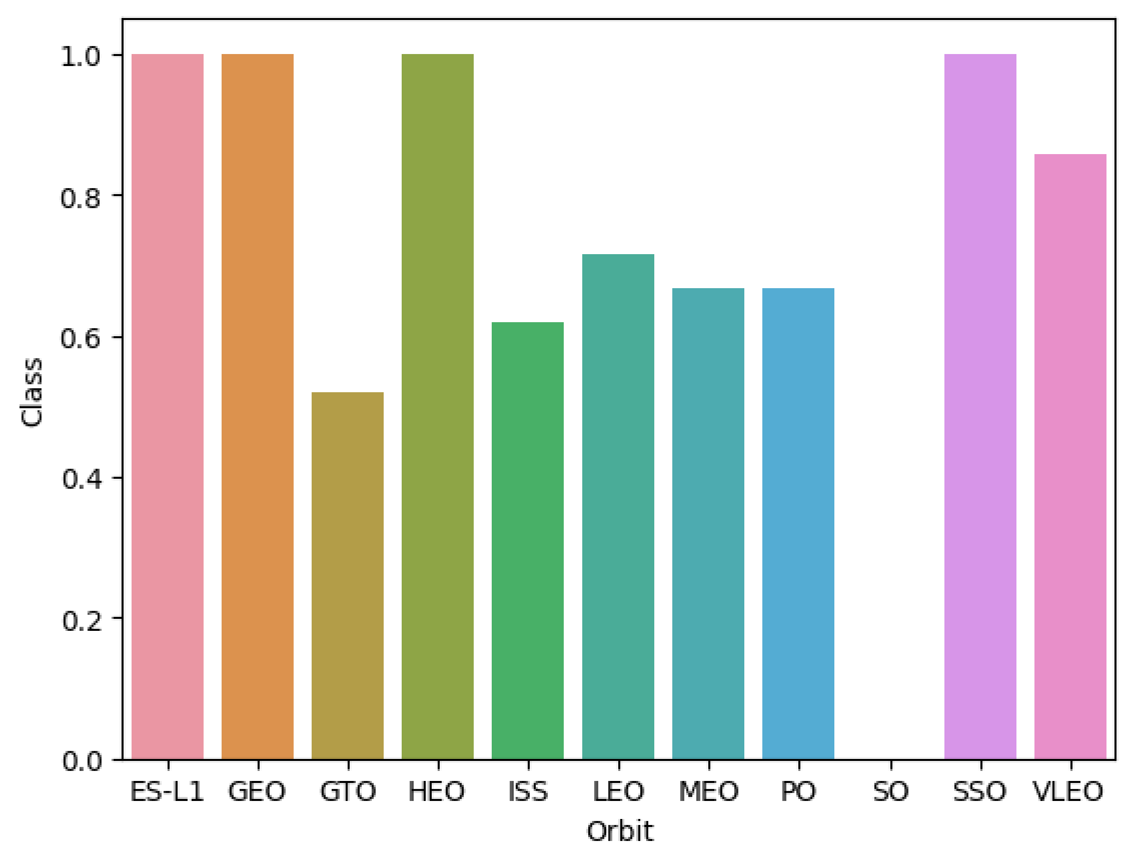
Payload vs. Launch Site

- It seems that the higher the payload, the greater the success rate
- Also for the VAFB-SLC launch site, there are no rockets launched for heavy payload mass or greater than 10,000kg



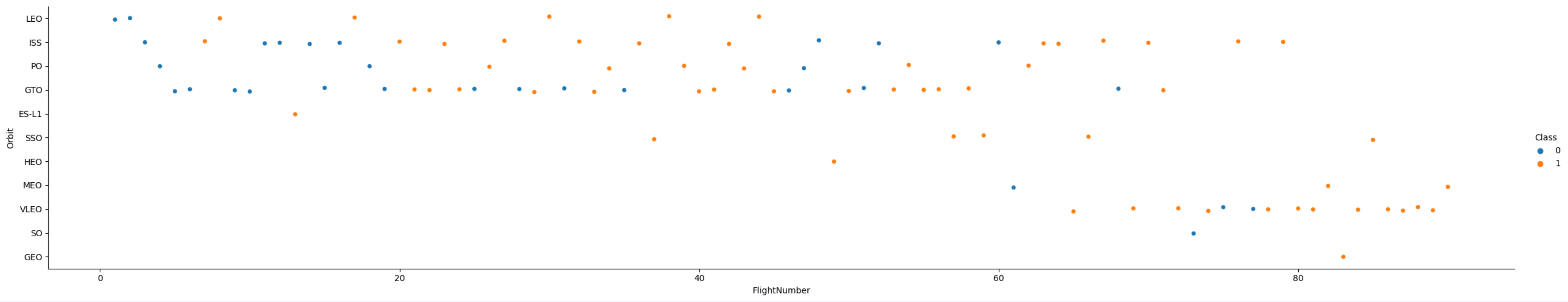
Success Rate vs. Orbit Type

- The success rate has been 100% for launches with destination orbit as ES-L1, GEO, HEO, and SSO.



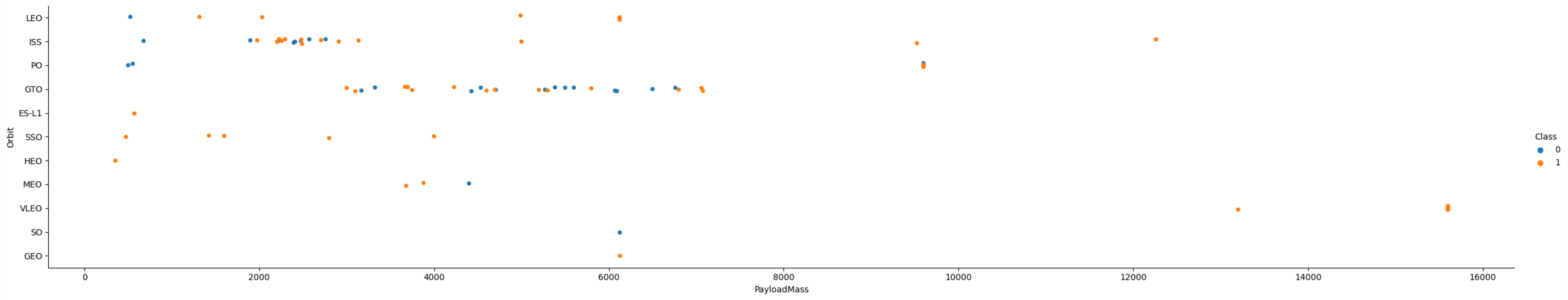
Flight Number vs. Orbit Type

- For most destination orbits, especially for some of them such as VLEO and ISS, the higher the flight number (i.e. later flights), the higher the success rate.



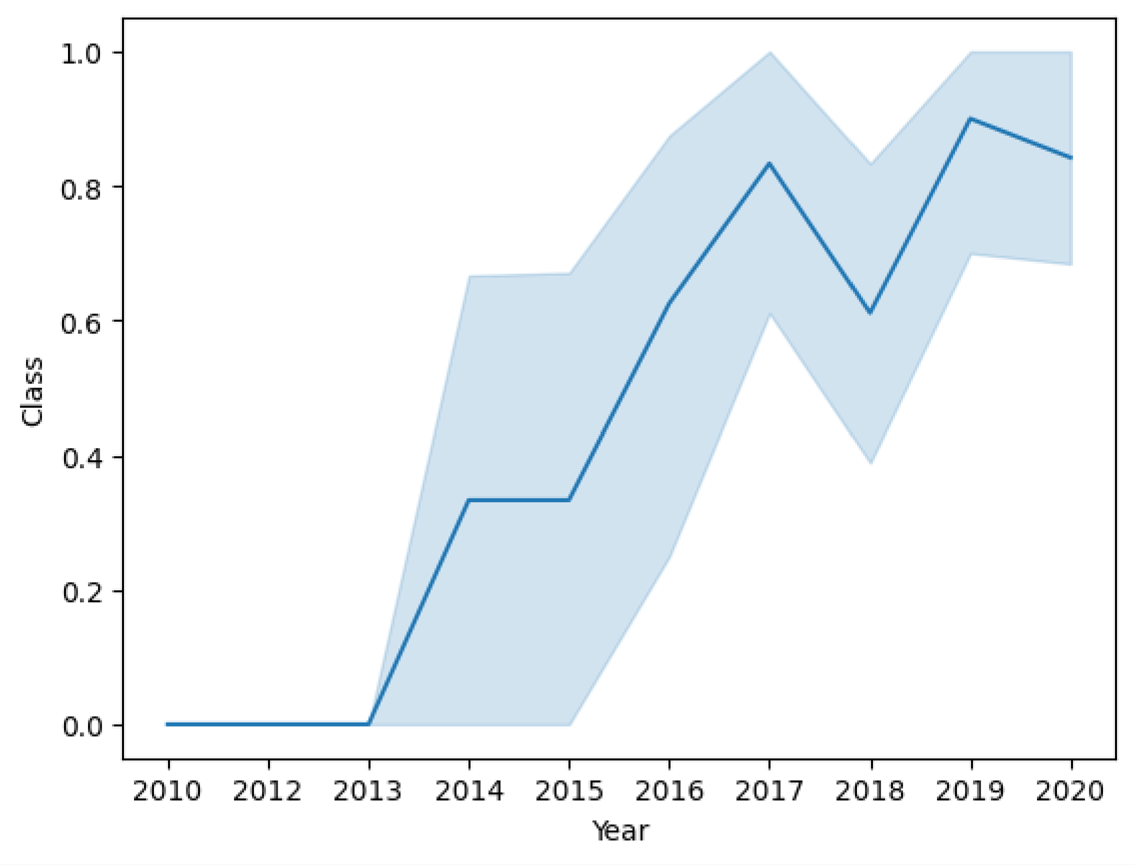
Payload vs. Orbit Type

- In general, for higher payload mass, no matter which destination orbit, and success rate is higher.



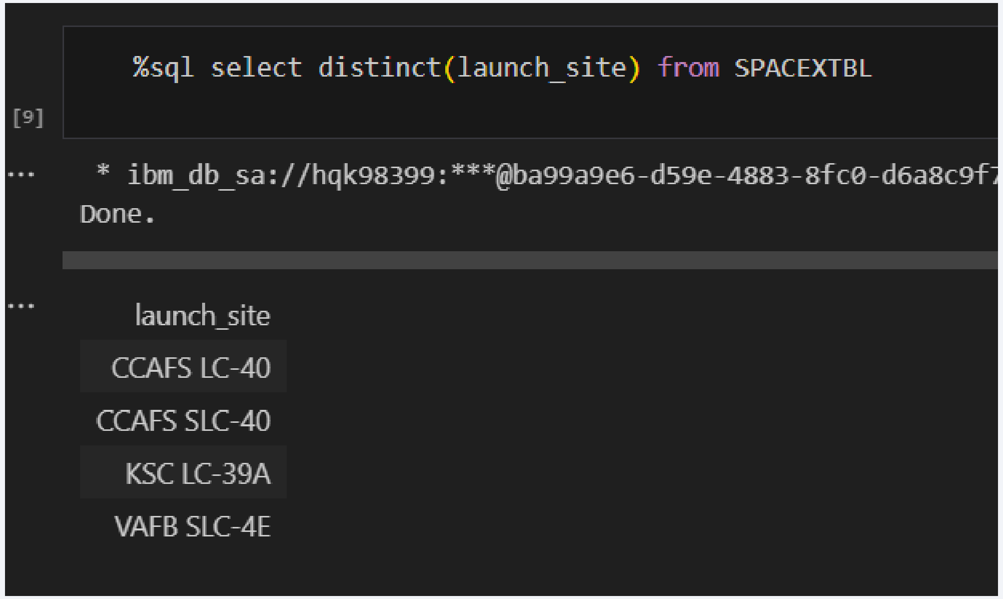
Launch Success Yearly Trend

- In general, the success rate increased with year



All Launch Site Names

- We used SELECT DISTINCT(launch_site) to find the 4 unique Launch Site values



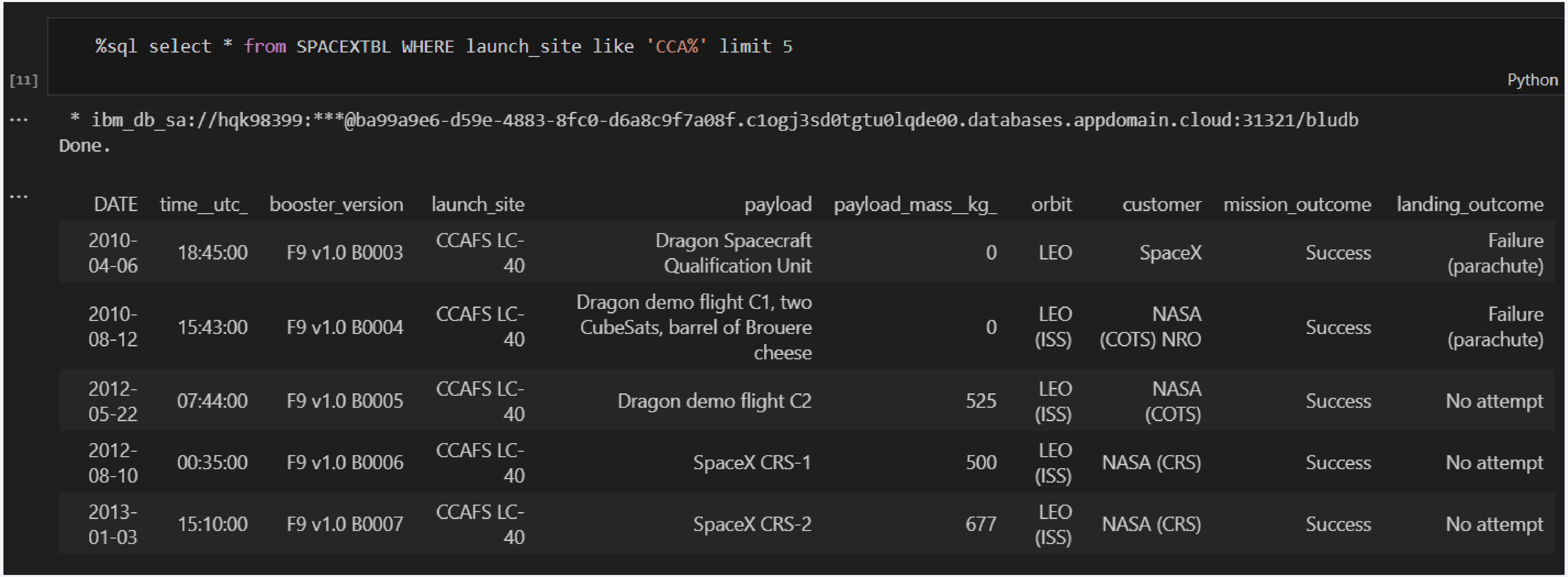
```
%sql select distinct(launch_site) from SPACEXTBL
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7
Done.
```

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

Launch Site Names Begin with 'CCA'

- We used “WHERE launch_site like ‘CCA%’” in SELECT to filter records with Launch Sites begin with ‘CCA’, plus used “Limit 5” to just retrieve the first 5 records that fit the criteria



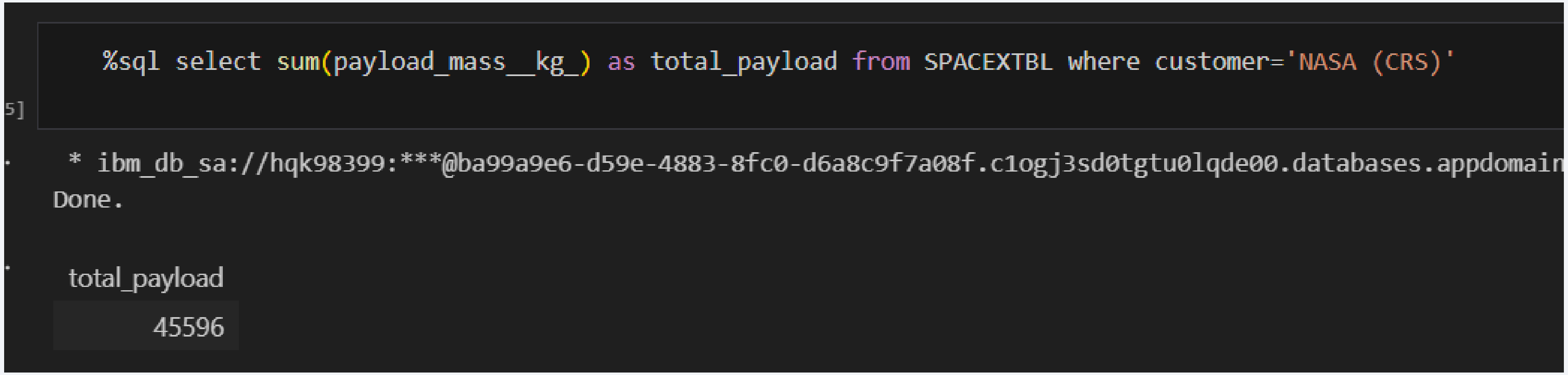
```
%sql select * from SPACEXTBL WHERE launch_site like 'CCA%' limit 5
```

```
* ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- We used “SELECT SUM(payload_mass__kg_)” to calculate the respective total, plus added a “WHERE customer=‘NASA (CRS)’ to filter the correct rows



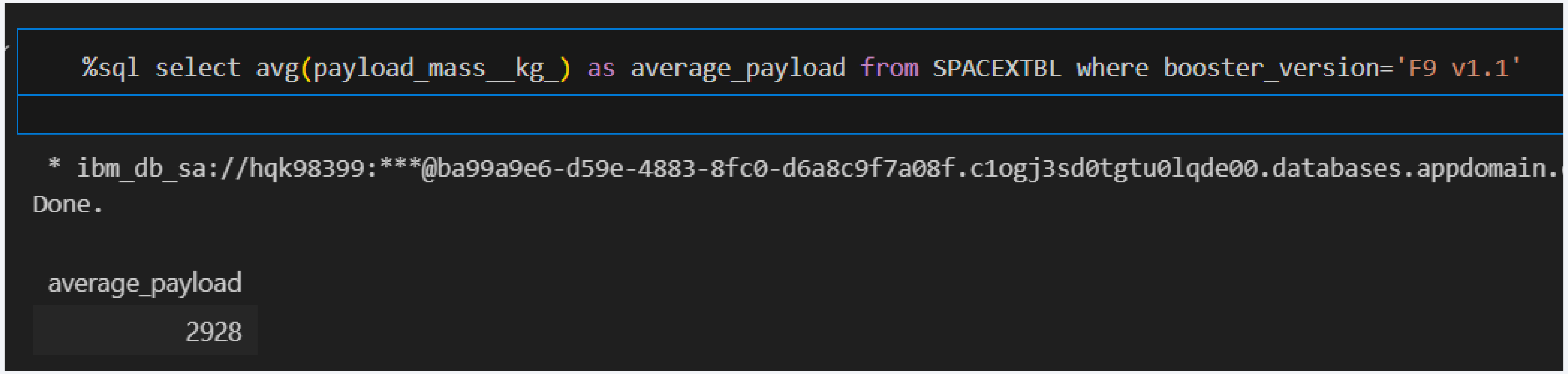
```
%sql select sum(payload_mass__kg_) as total_payload from SPACEXTBL where customer='NASA (CRS)'
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain
Done.
```

total_payload
45596

Average Payload Mass by F9 v1.1

- We used “SELECT AVG(payload_mass__kg_)” to calculate the average payload, and added “WHERE booster_version=‘F9 v1.1’” to filter the correct rows

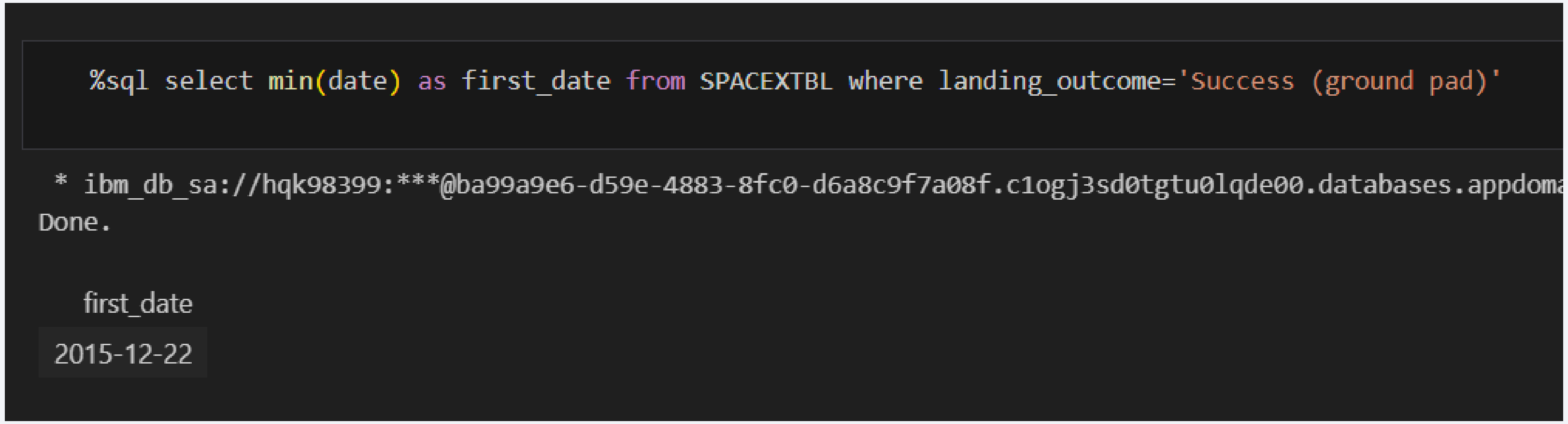


```
%sql select avg(payload_mass__kg_) as average_payload from SPACEXTBL where booster_version='F9 v1.1'
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.
Done.

 average_payload
            2928
```

First Successful Ground Landing Date

- We used “SELECT MIN(date)” to find the smallest date, and “WHERE landing_outcome=‘Success (ground pad)” to filter the correct rows



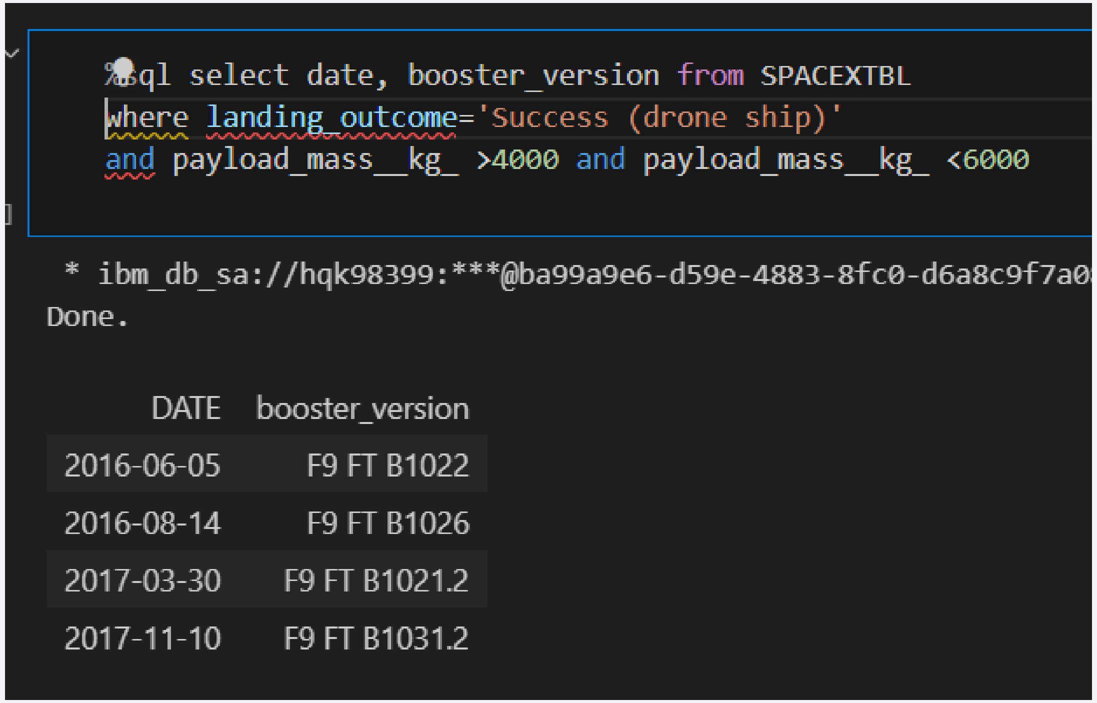
```
%sql select min(date) as first_date from SPACEXTBL where landing_outcome='Success (ground pad)'
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdoma
Done.
```

first_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used “WHERE landing_outcome=‘Success (drone ship)’, plus two payload_mass__kg conditions to filter the correct rows



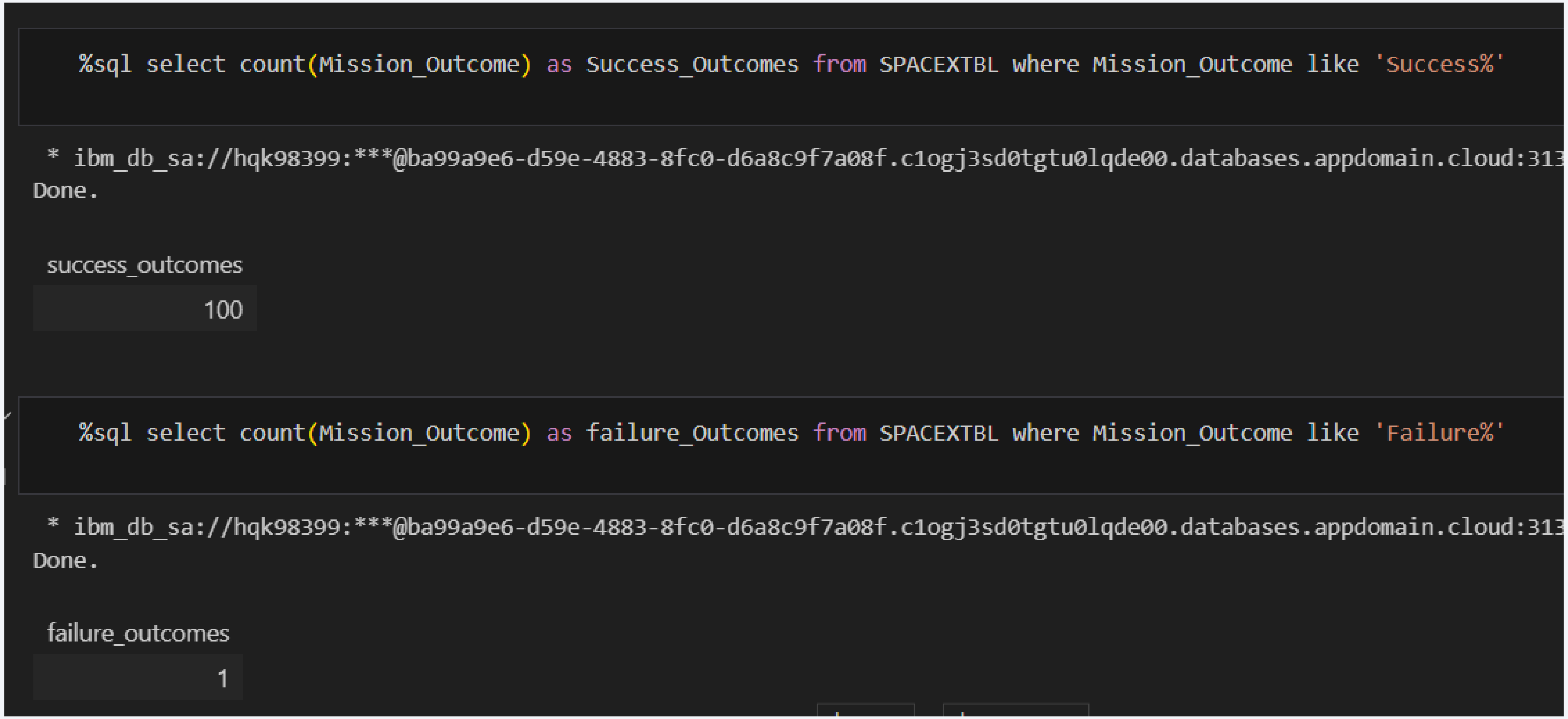
```
%sql select date, booster_version from SPACEXTBL
where landing_outcome='Success (drone ship)'
and payload_mass__kg_ >4000 and payload_mass__kg_ <6000
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a0
Done.
```

DATE	booster_version
2016-06-05	F9 FT B1022
2016-08-14	F9 FT B1026
2017-03-30	F9 FT B1021.2
2017-11-10	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- We used “WHERE Mission_Outcome like” to filter the correct rows for each case, and “SELECT COUNT(Mission_Outcome)” to give the count for each case



```
%sql select count(Mission_Outcome) as Success_Outcomes from SPACEXTBL where Mission_Outcome like 'Success%'
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:313
Done.
```

success_outcomes
100

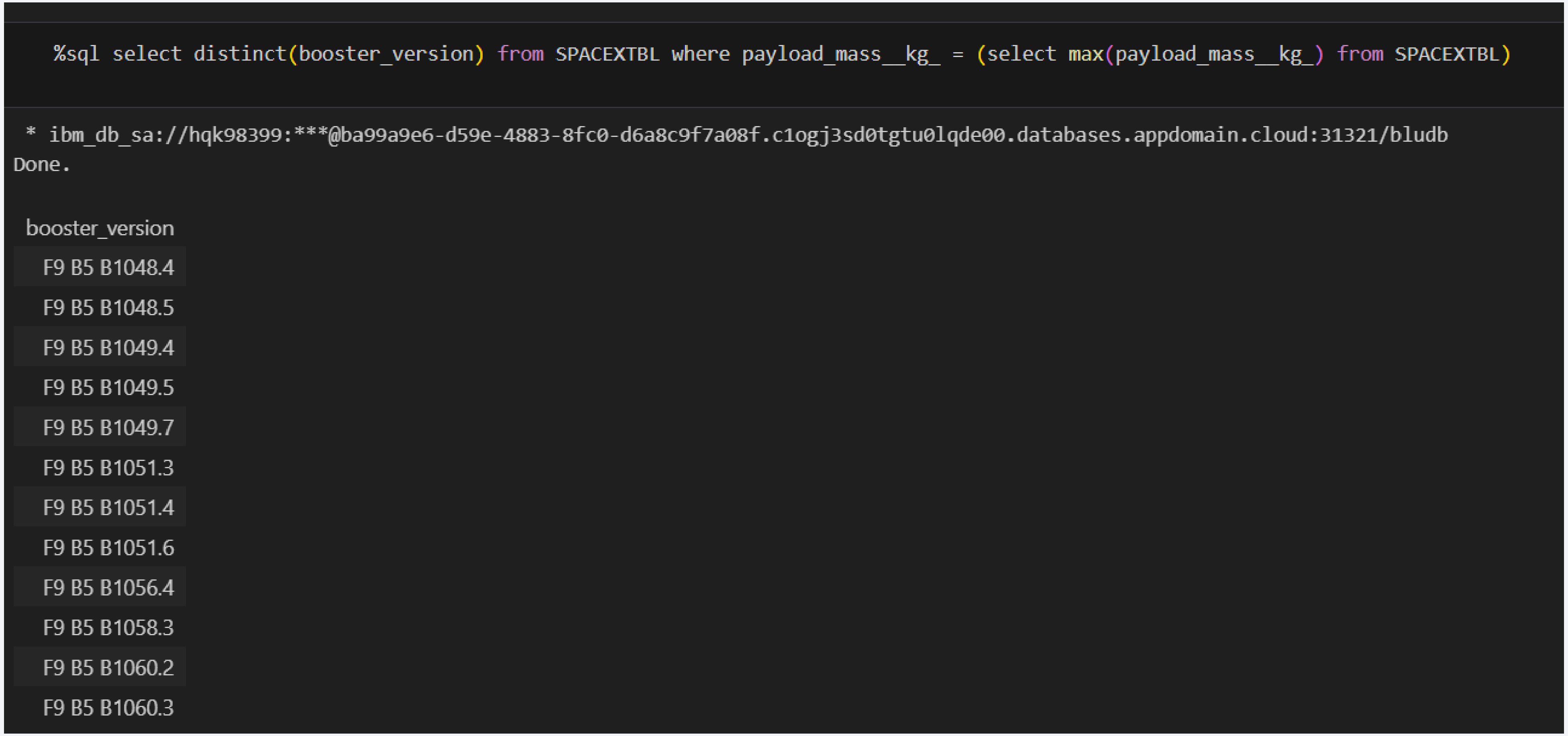
```
%sql select count(Mission_Outcome) as failure_Outcomes from SPACEXTBL where Mission_Outcome like 'Failure%'
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:313
Done.
```

failure_outcomes
1

Boosters Carried Maximum Payload

- We used a subquery “WHERE payload_mass__kg_ = (SELECT MAX (payload_mass__kg_))” to filter the rows with maxmimum payload mass from the table



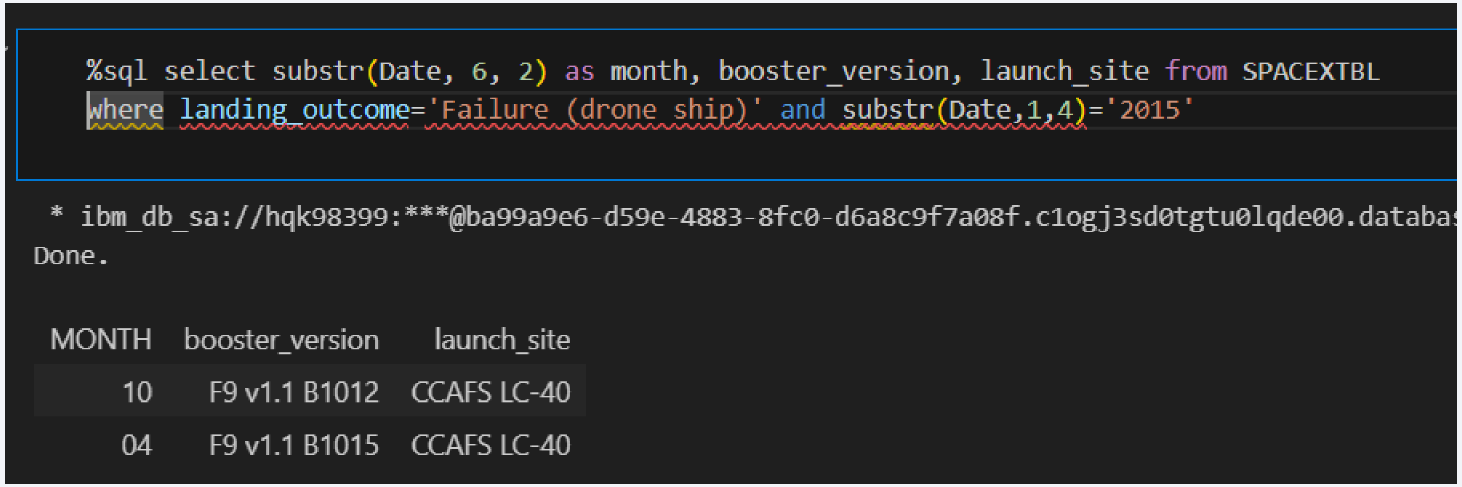
```
%sql select distinct(booster_version) from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

2015 Launch Records

- We used the suggested SUBSTR(Date, 1, 4) to zoom into the year of the Date volume, and found the related 2015 launched; plus used “WHERE landing_outcome=‘Failure (drone ship)’” to identify the correct rows



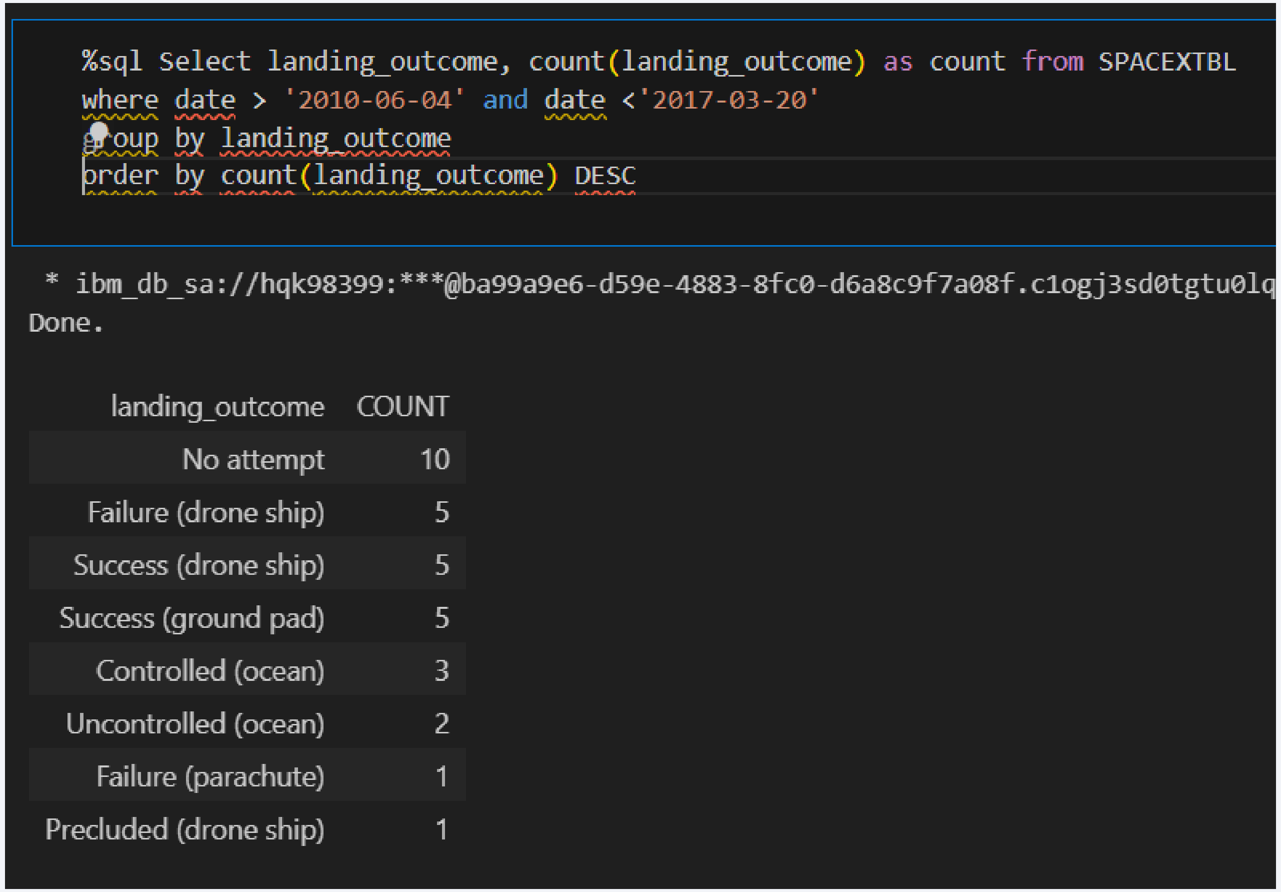
```
%sql select substr(Date, 6, 2) as month, booster_version, launch_site from SPACEXTBL
where landing_outcome='Failure (drone ship)' and substr(Date,1,4)='2015'
```

```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databa
Done.
```

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

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

- We used two “WHERE date” conditions to filter the rows with the required date range, and then used GROUP BY to group the landing outcomes, plus ORDER BY to do the required sorting



```
%sql Select landing_outcome, count(landing_outcome) as count from SPACEXTBL
where date > '2010-06-04' and date <'2017-03-20'
group by landing_outcome
order by count(landing_outcome) DESC
```

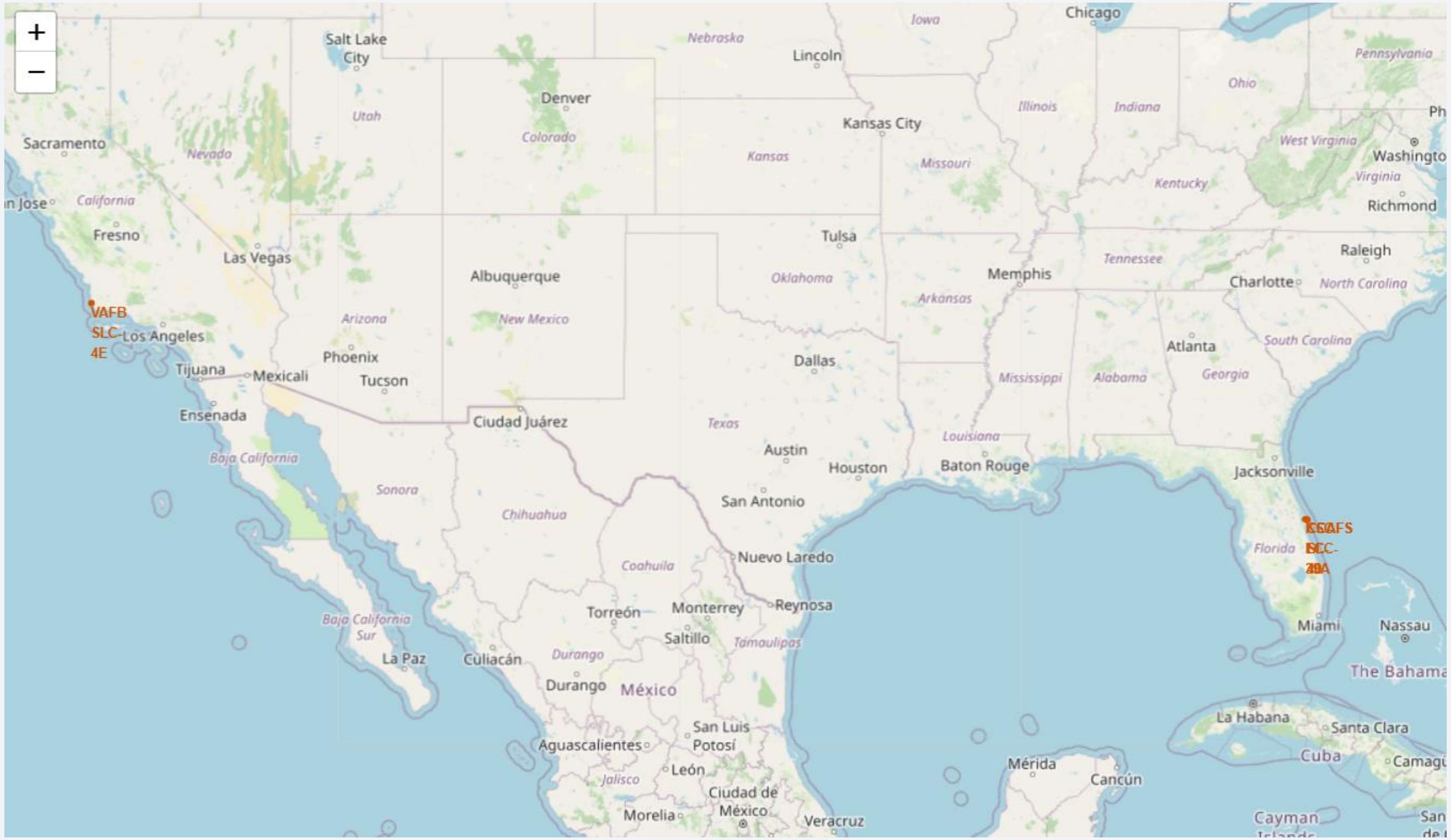
```
 * ibm_db_sa://hqk98399:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lq
Done.

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

Section 3

Launch Sites Proximities Analysis

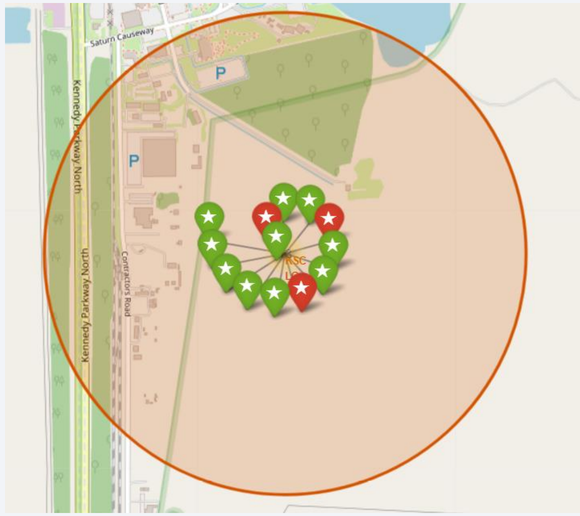
Global map with all Launch Sites

- All Launch Sites used by SpaceX were located in either the State of Floria or California



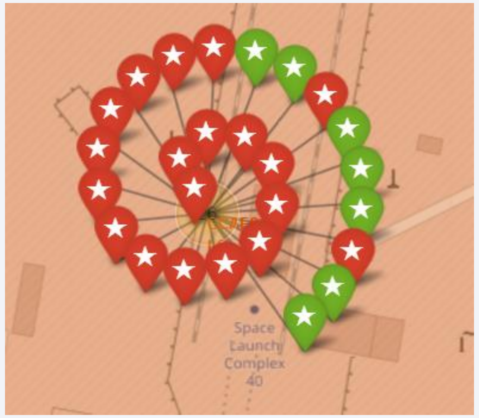
Successful/Failed Launches at each Launch Site

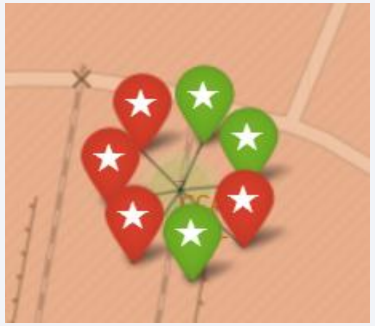
- Marker Clusters were inserted to each launch site, to indicate successful (green) vs. failed (red) launch at each location



KSC LC-39A

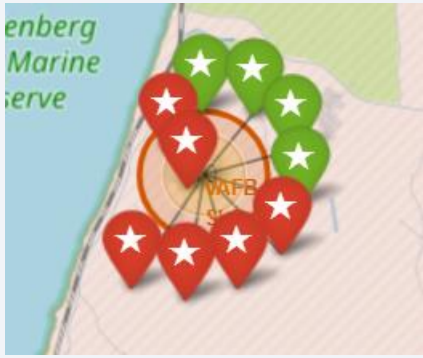
CCAFS LC-40





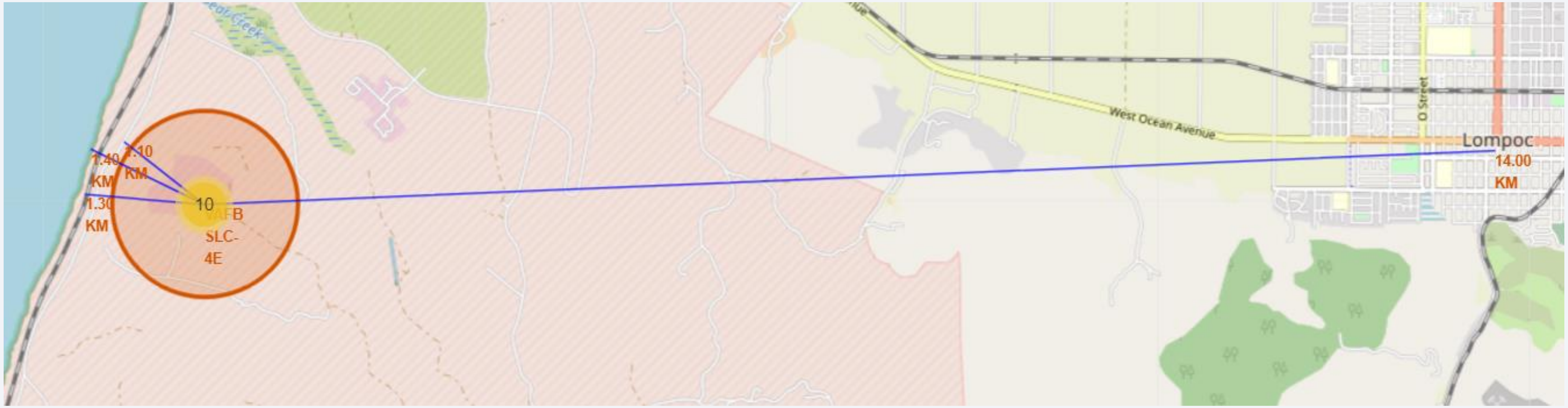
CCAFS SLC-40

VAFB SLC-4E



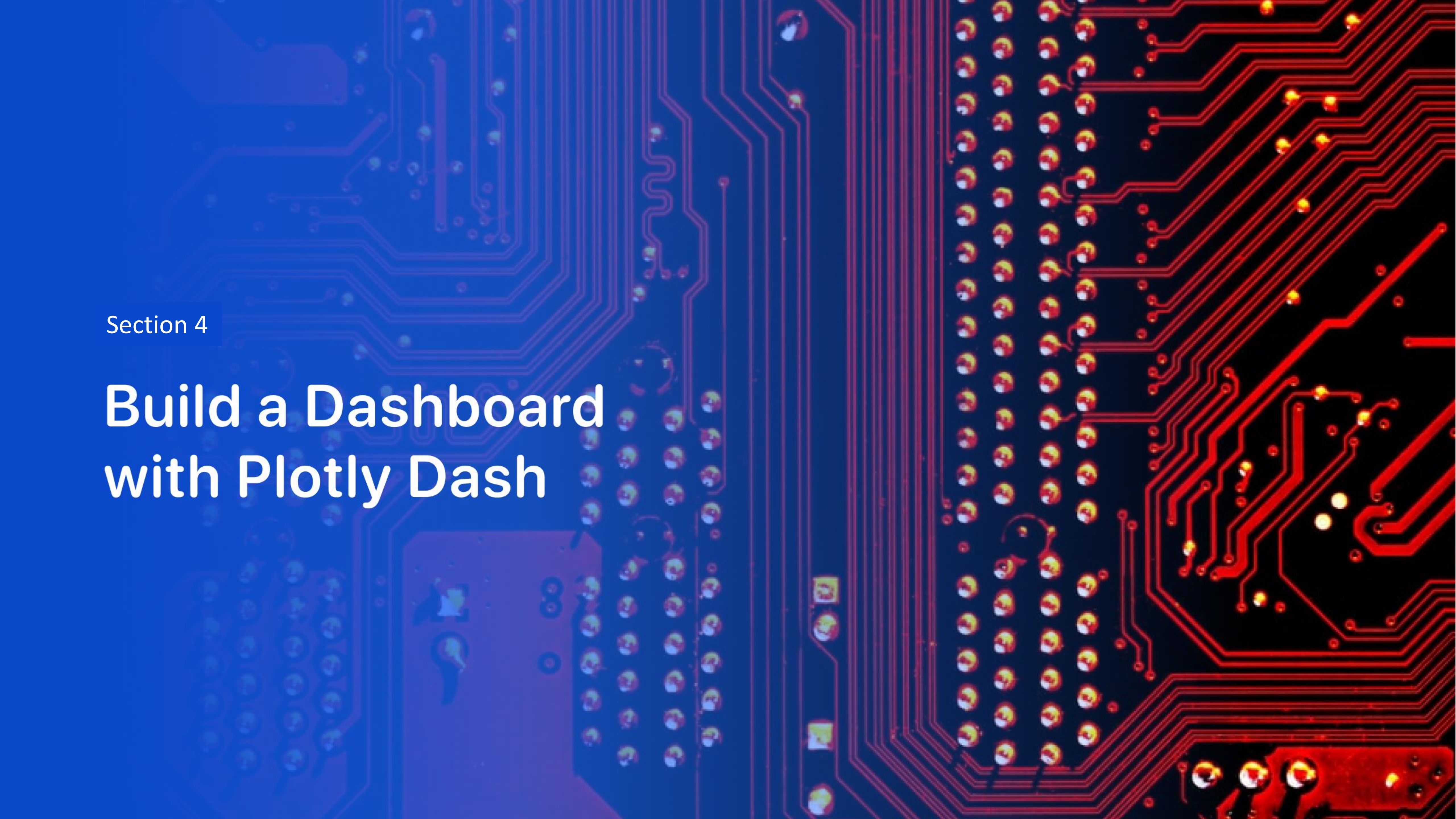
Importance of proximities to supporting infrastructure

- All launch sites are close in distance to railway, highway, coastline, and also cities. It shows that these supporting infrastructure are very important to rocket launches



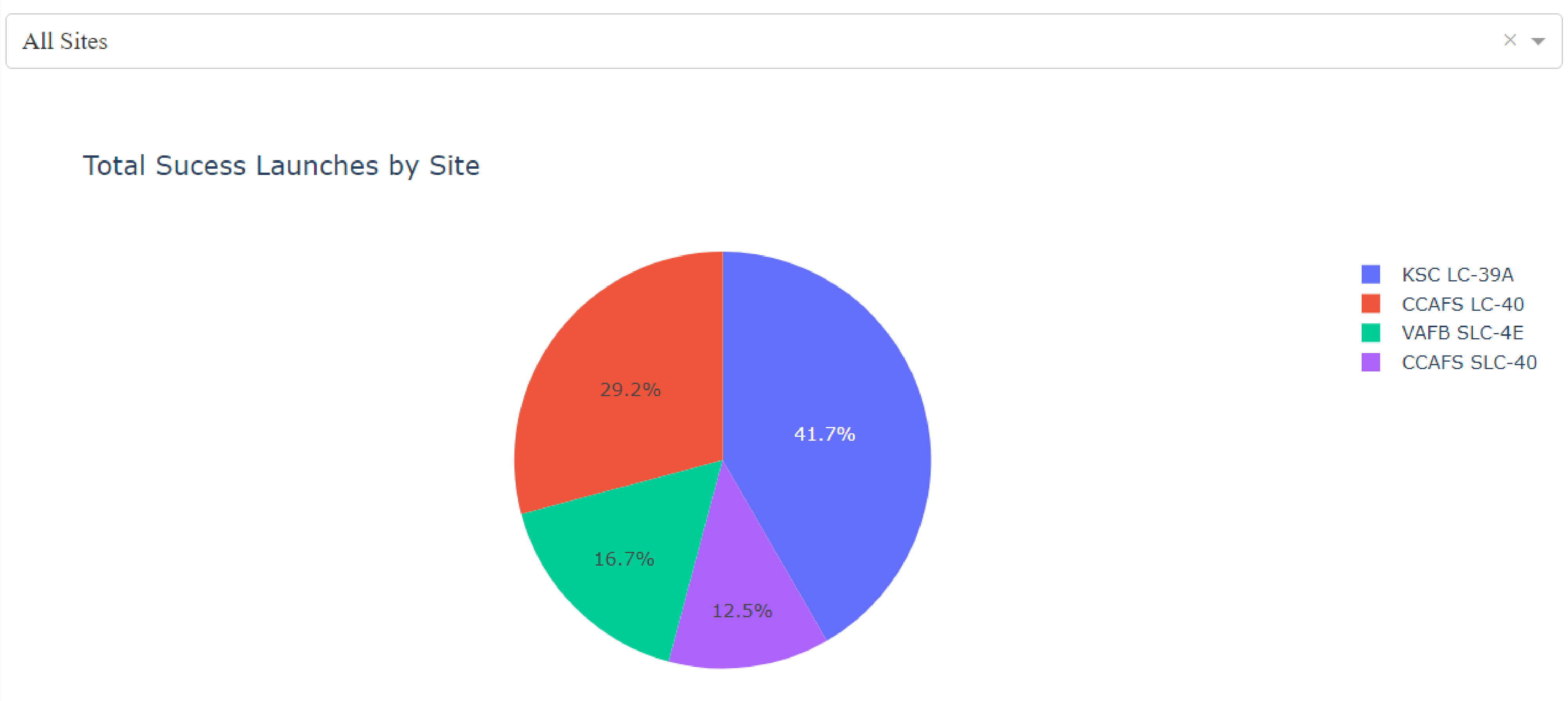
Section 4

Build a Dashboard with Plotly Dash



Distribution of successful launch by launch site

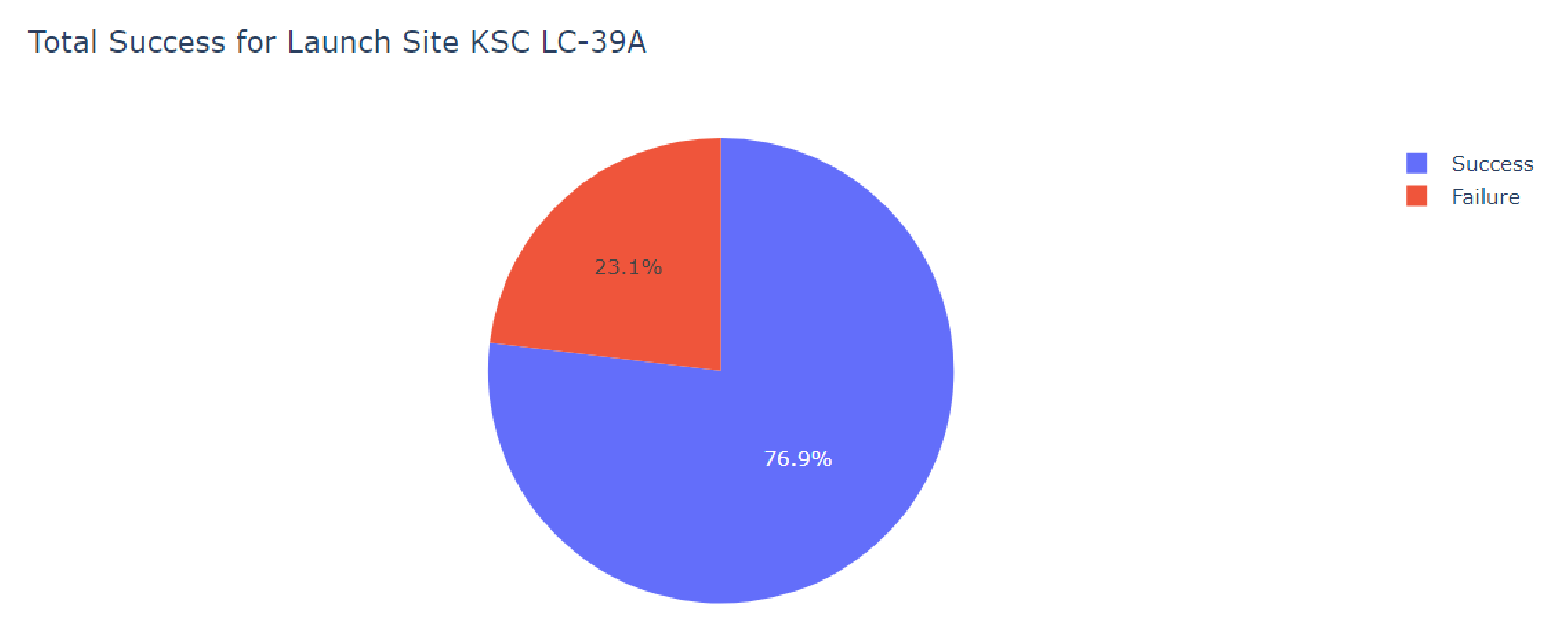
- The launch site KSC LC-39A has the highest number of successful launch among the 4 launch sites



All Sites

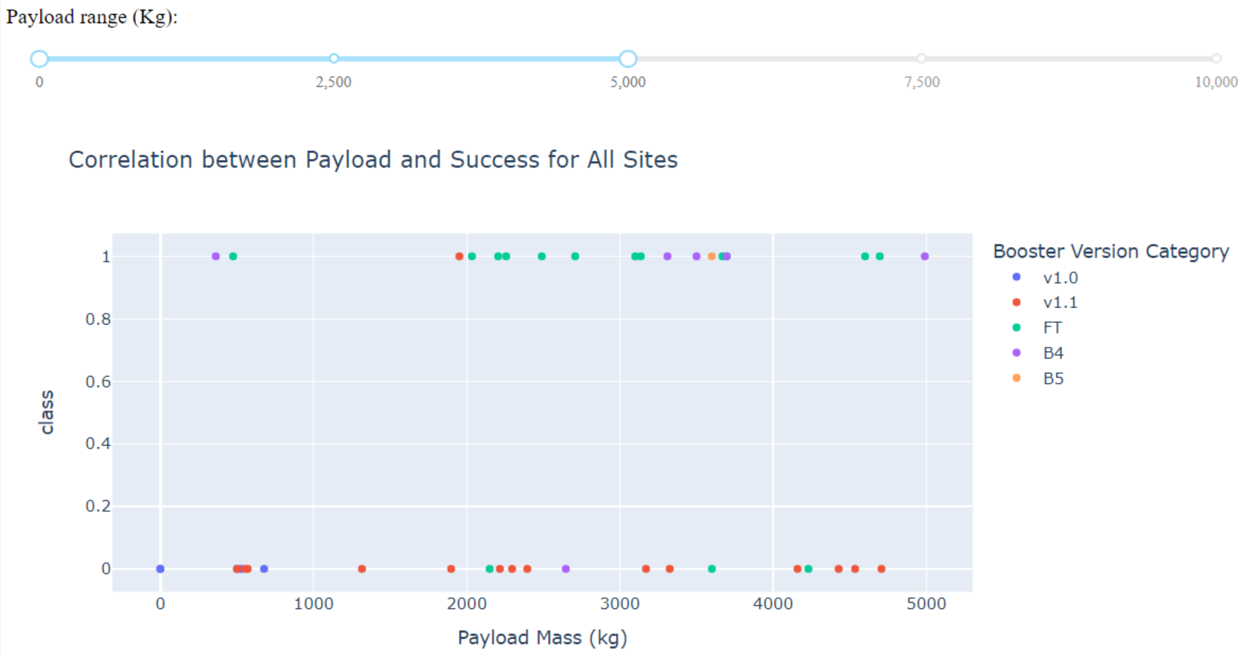
KSC LC-39A has a success rate of 76.9%

- For all launches from KSC LC-39A, the success rate was 76.9%

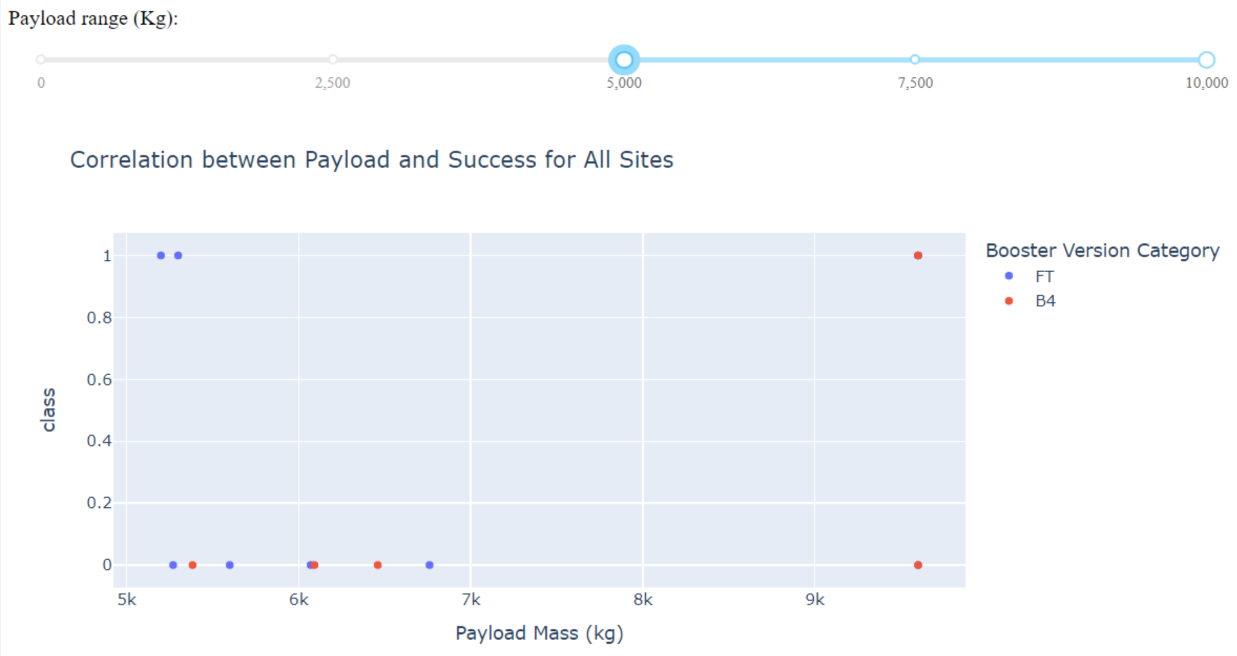


Relationship between Payload Mass, Booster Version and Success Rate

- The success rate seems to be higher for payload mass between 2,000kg and 4,000kg, and with booster version FT and B4



Scatter plot for Payload mass <=5,000kg



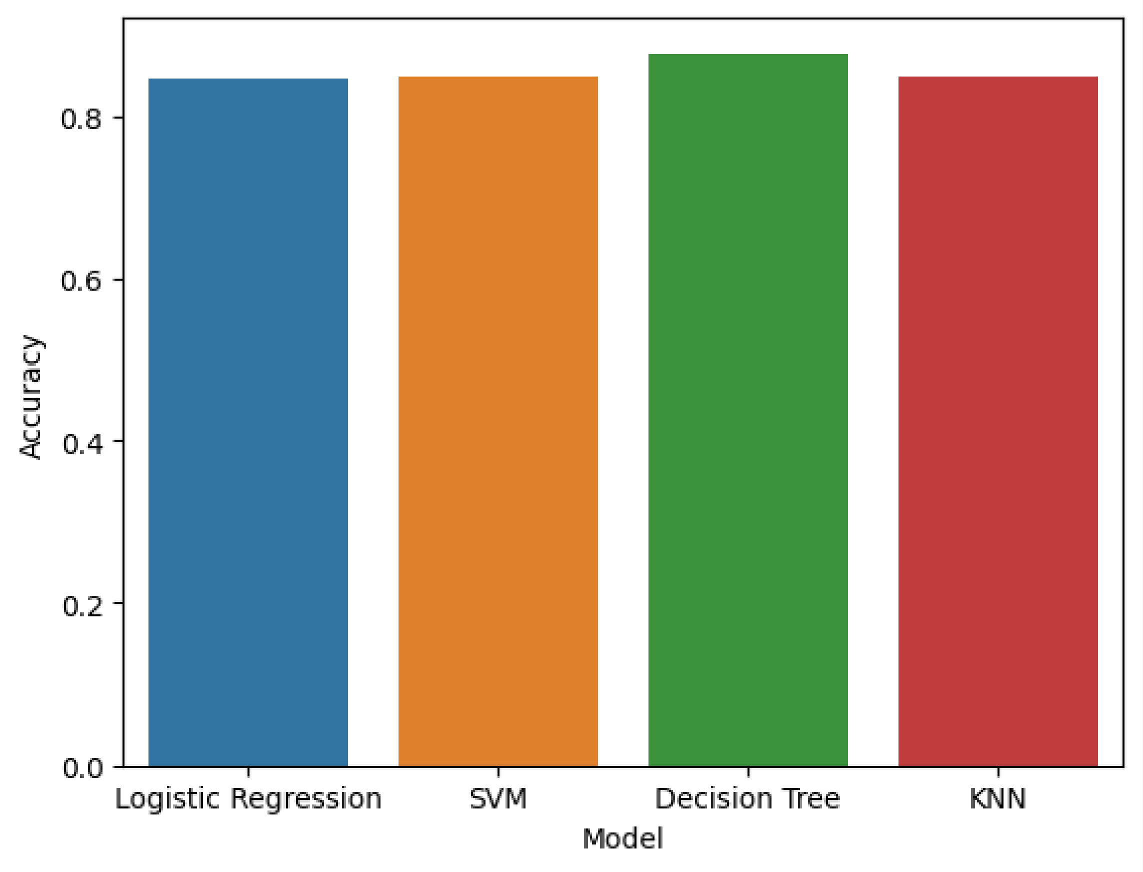
Scatter plot for Payload mass >=5,000kg

Section 5

Predictive Analysis (Classification)

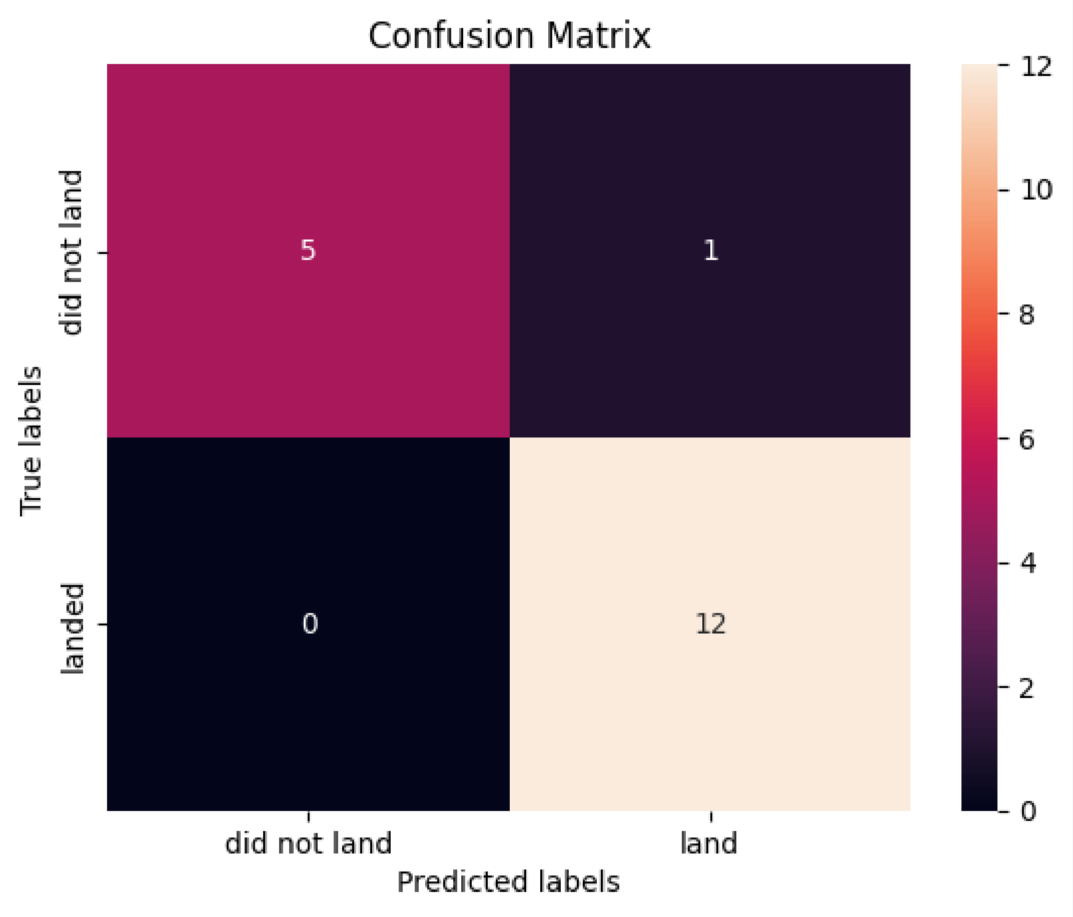
Classification Accuracy

- Among the 4 models we considered and tested, the Decision Tree model has the highest level of accuracy, at around 88%



Confusion Matrix

- Below is the confusion matrix of the decision tree model; it shows that it has high accuracy in terms of predicting the successful and failed landing



Conclusions

- The success rate of first stage landing has been generally increasing for SpaceX during the year 2013 to 2020
- Certain factors proved to be contributing to higher success rate, including the use of KSC LC-39A launch site, having the payload mass between 2,000kg and 4,000kg, etc.
- It would be best to use a Decision Tree to predict whether the first stage rocket will successfully land for future launches

Thank you!