IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rohan Kumar Das
24/02/2023

Outline

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

Executive Summary

- Summary of methodologies

  Data Collection using web scraping and SpaceX API

  Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics

  Machine Learning Prediction

- Summary of all results

  It was possible to collected valuable data from public sources.

  EDA allowed to identify which features are the best to predict success of launchings.

  Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

The objective is to evaluate the viability of the new company Space Y to compete with Space X.

Desirable answers:

- The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
- Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features.
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

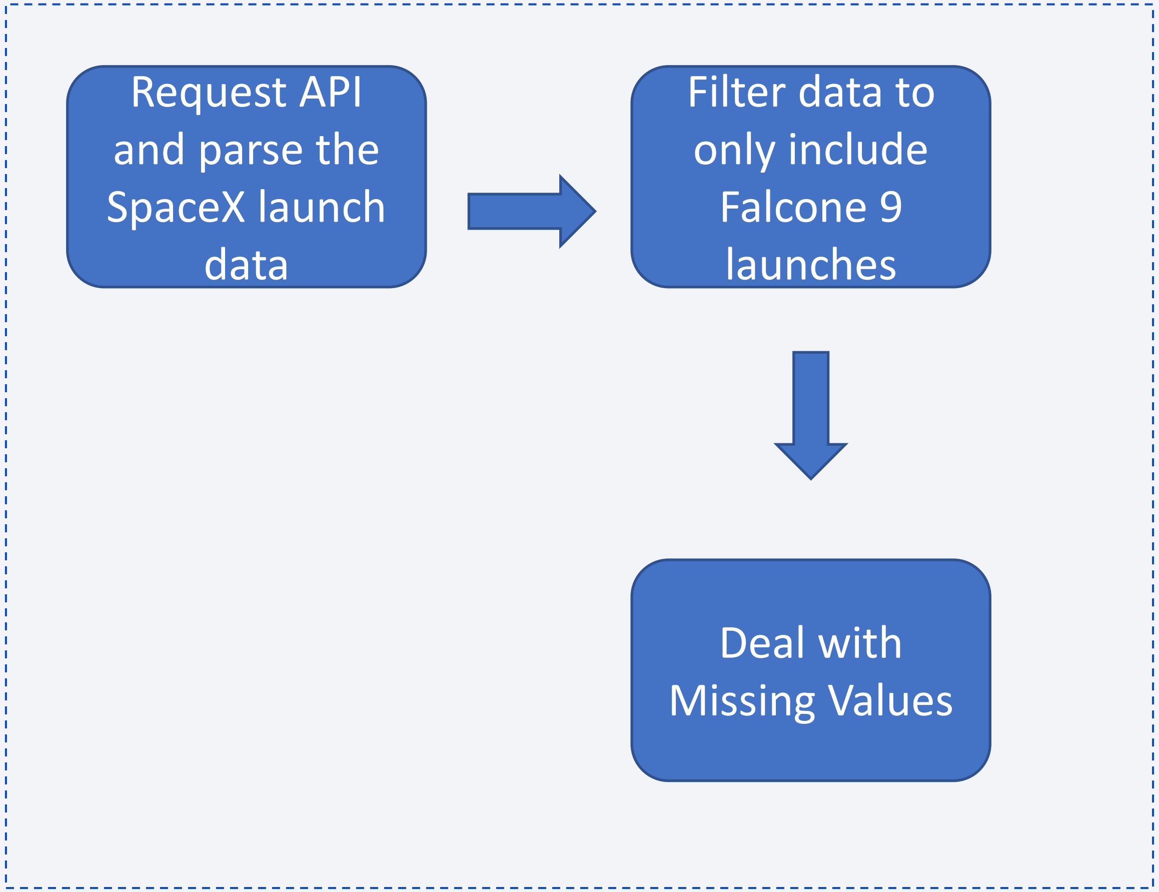
- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/)

and from Wikipedia

(https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

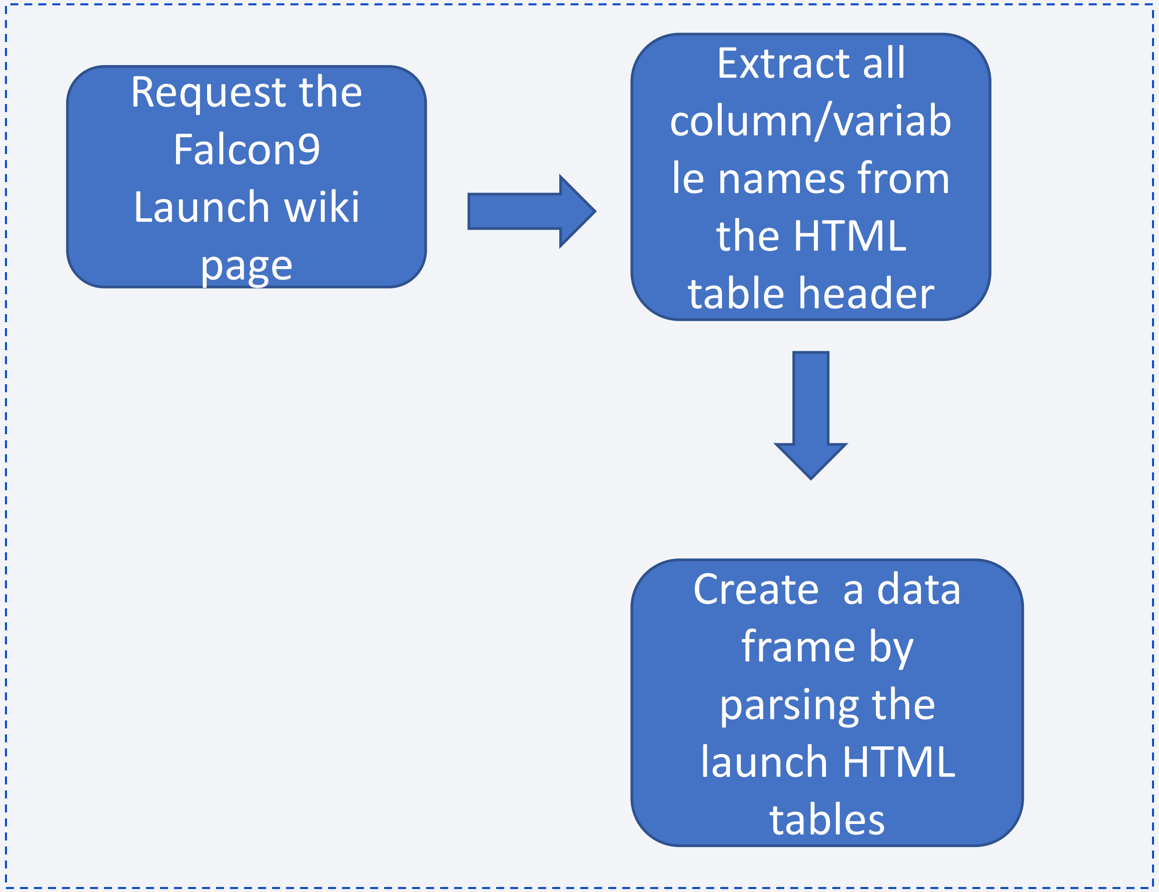
- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
- Source Code: [Coursera_capstone_project/jupyter-labs-spacex-data-collection-api.ipynb at main · Rohankumardas/Coursera_capstone_project (github.com)]



Request API and parse the SpaceX launch data
Filter data to only include Falcone 9 launches
Deal with Missing Values

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flow chart and then persisted.
- Source code : [Coursera_capstone_project/jupyter-labs-webscraping.ipynb at main · Rohankumardas/Coursera_capstone_project (github.com)]

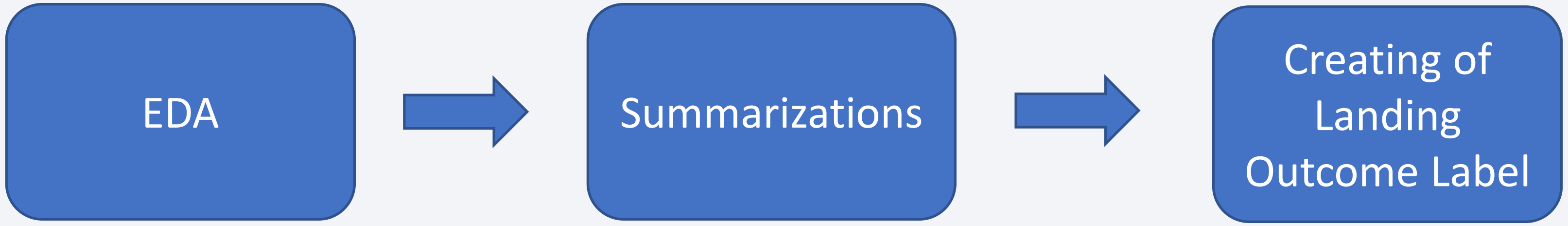

Request the Falcon9 Launch wiki page

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site , occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.
- Source code: [Coursera_capstone_project/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb at main · Rohankumardas/Coursera_capstone_project (github.com)]()

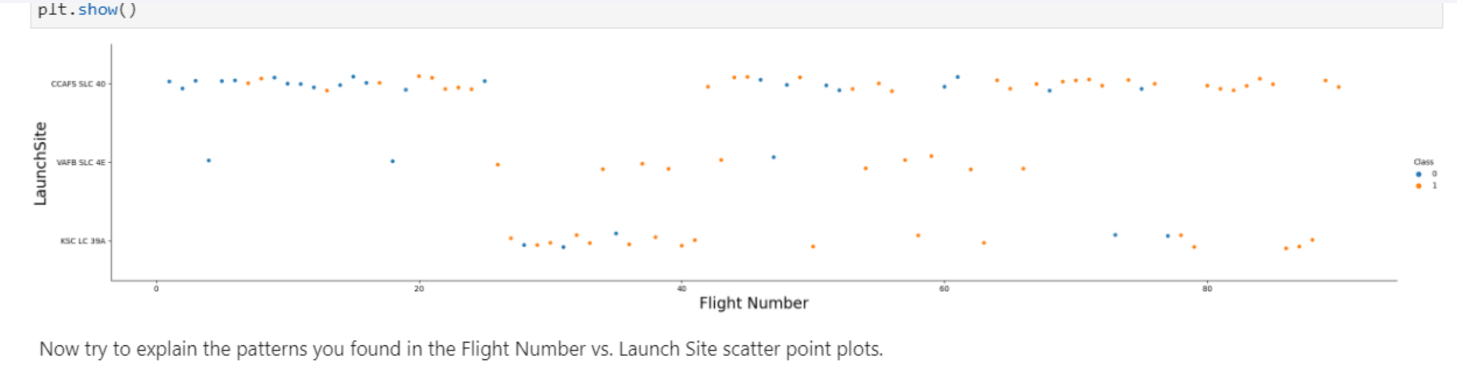


EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:

- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site Payload Mass, Orbit and Flight Number, Payload and Orbit

Source Code:Coursera_capstone_project/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb at main · Rohankumardas/Coursera_capstone_project (github.com)



```
plt.show()
```

Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

EDA with SQL

The following SQL queries were performed:

- 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.
- Source code: Coursera_capstone_project/jupyter-labs-eda-sql-coursera_sqllite (1).ipynb at main · Rohankumardas/Coursera_capstone_project (github.com)

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson SpaceCenter;
  - Marker clusters indicates groups of events in each coordinate, like l aunches in a launch site;and
  - Lines are used to indicate distances between two coordinates.

Source Code: Coursera_capstone_project/lab_jupyter_launch_site_location.jupyterlite.ipynb at main · Rohankumardas/Coursera_capstone_project (github.com)

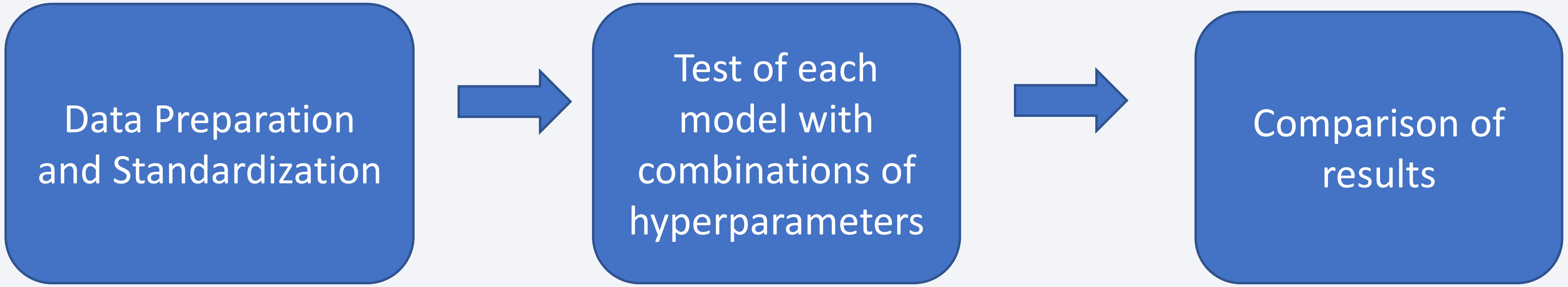
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
    - Percentage of launches by site
    - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Coursera_capstone_project/spacex_dash_app.py at main · Rohankumardas/Coursera_capstone_project (github.com)

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.
- Coursera_capstone_project/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · Rohankumardas/Coursera_capstone_project (github.com)



Data Preparation and Standardization

Test of each model with combinations of hyperparameters

Comparison of results

Results

- Exploratory data analysis results

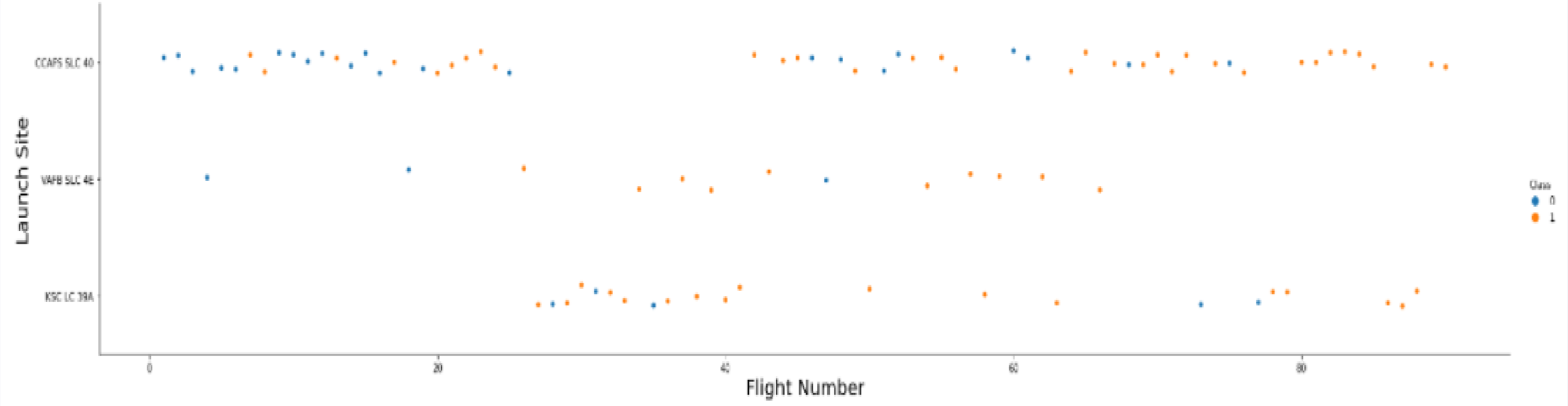
  - Space X uses 4 different launchsites;
  - The first launches were done to Space Xitself and NASA;
  - The average payload of F9 v1.1 booster is 2,928kg;
  - The first success landing outcome happened in 2015 fiver year after the firstlaunch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above theaverage;
  - Almost 100% of mission outcomes weresuccessful;
  - Twoboosterversionsfailedatlandingindroneshipsin2015:F9v1.1B1012andF9v1. 1B1015;
  - The number of landing outcomes became as better as yearspassed.

Section 2

Insights drawn from EDA

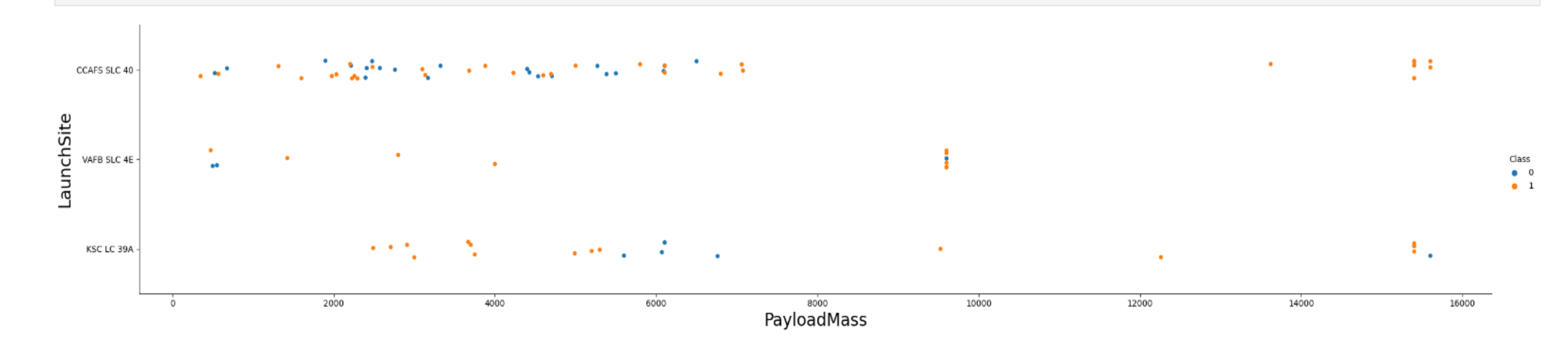
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



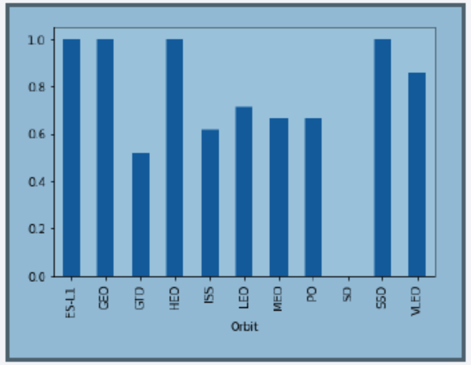
Payload vs. Launch Site

- Payloadsover9,000kg(about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSCLC 39A launch sites.



Success Rate vs. Orbit Type

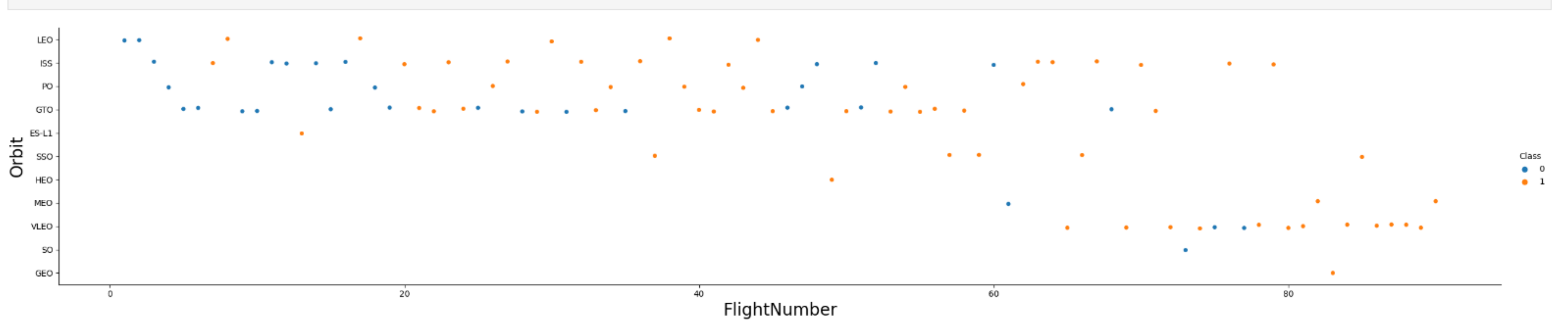
- The biggest success rates happens toorbits:
  - ES-L1;
  - GEO;
  - HEO;and
  - SSO.
- Followedby:
  - VLEO (above 80%);and
  - LFO (above70%).



Flight Number vs. Orbit Type

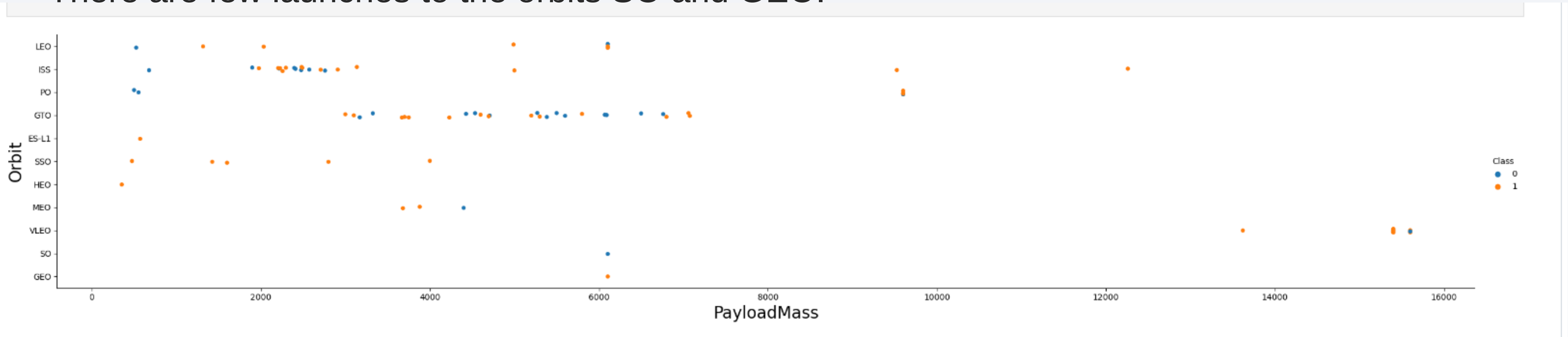
- Apparently, success rate improved over time to allorbits;

- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.



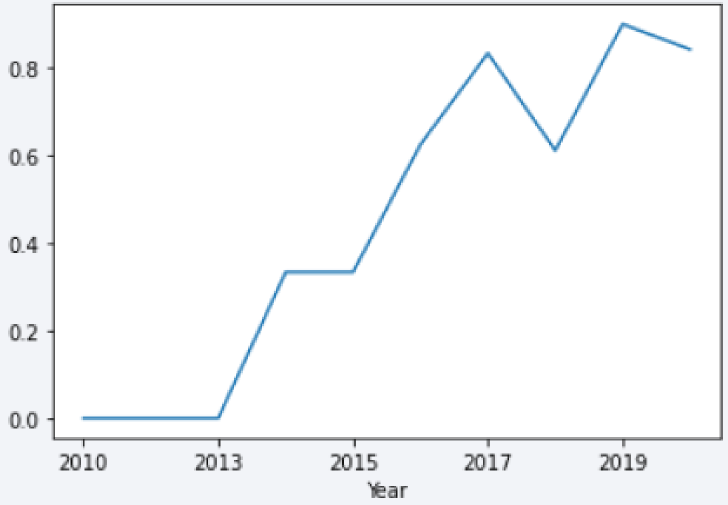
Payload vs. Orbit Type

- Apparently, there is no 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 launches to the orbits SO and GEO.



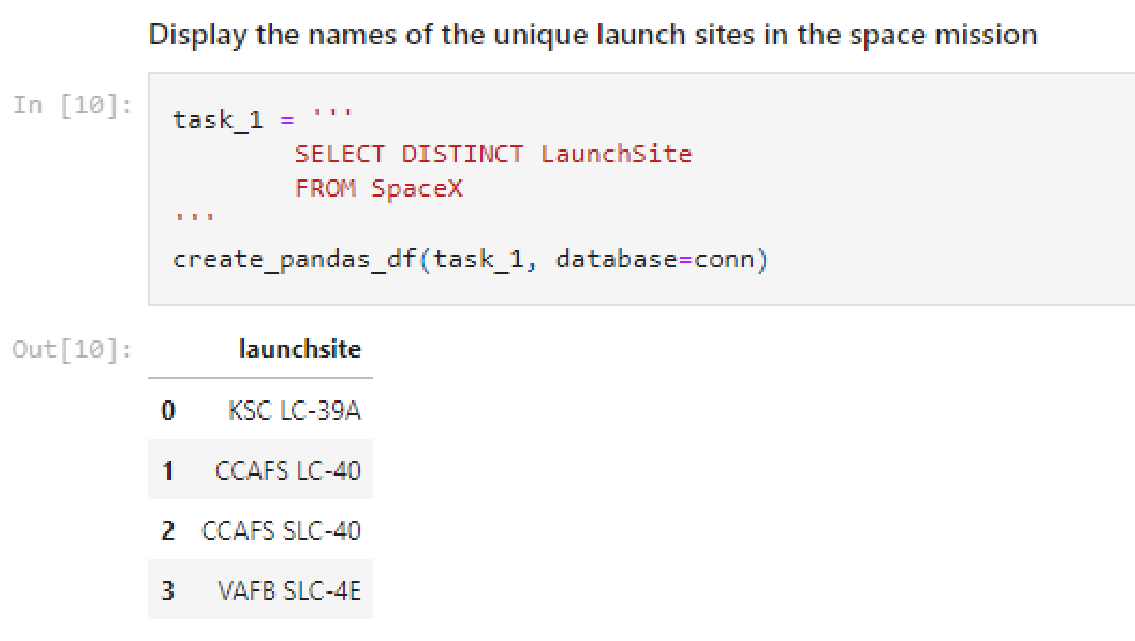
Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are four launch sites:
- They are obtained by selecting unique occurrences of “launch site” values from the dataset.



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

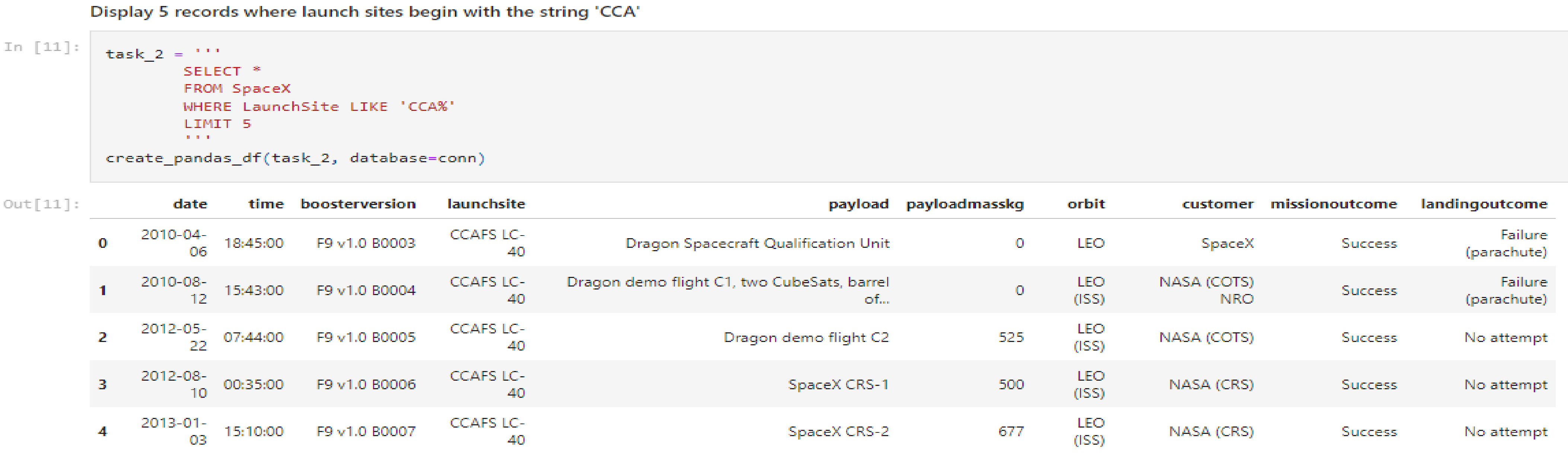
```
In [10]:
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

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



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

In [11]:

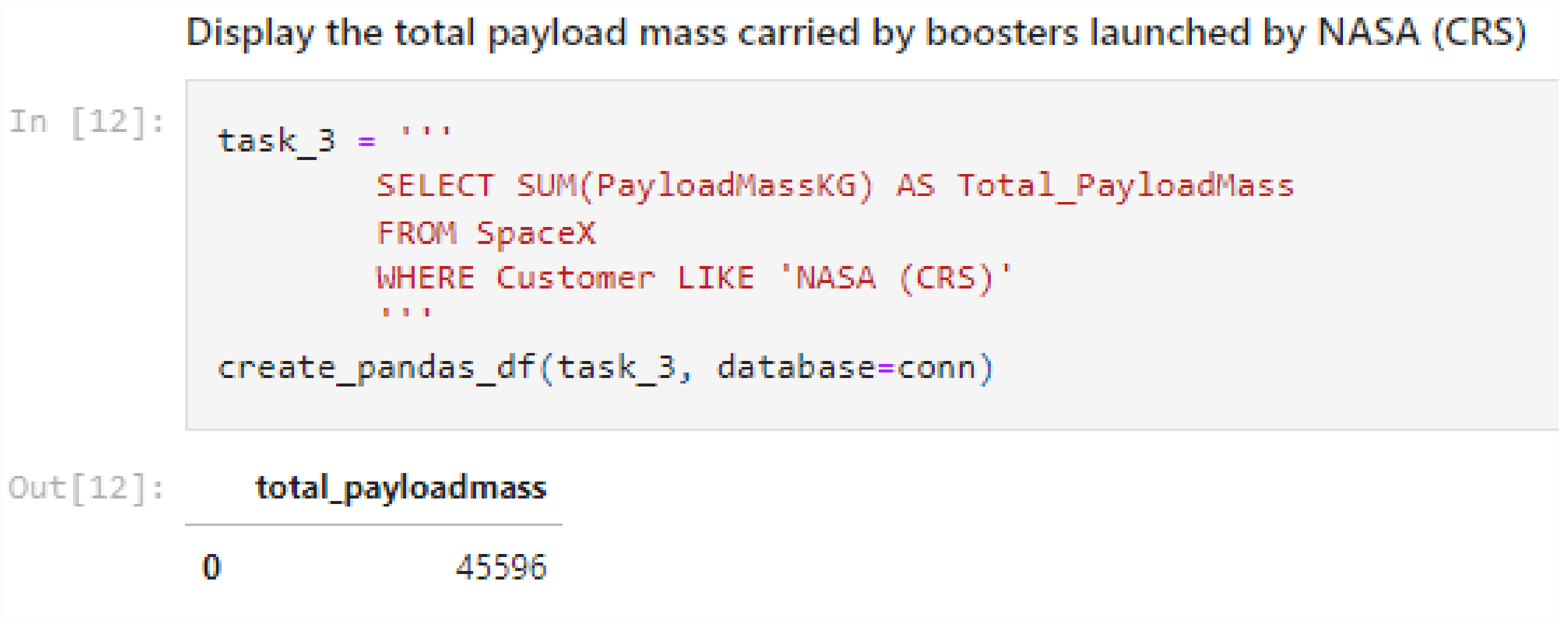
```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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 calculated the total payload carried by boosters from NASA as 45596 using the query below



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

In [12]:

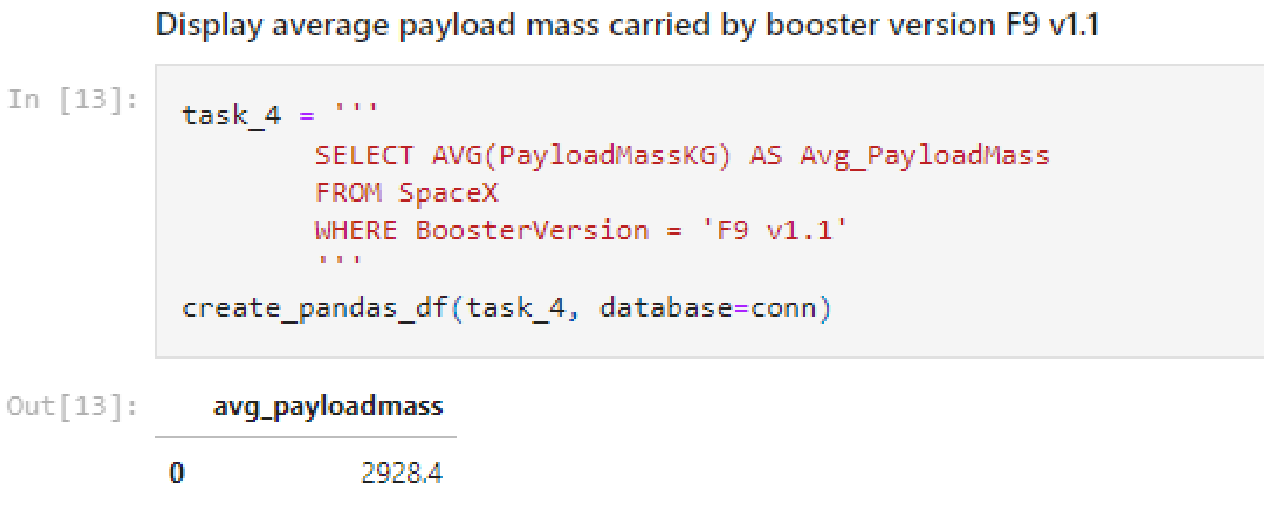
```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

Out[12]:

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4



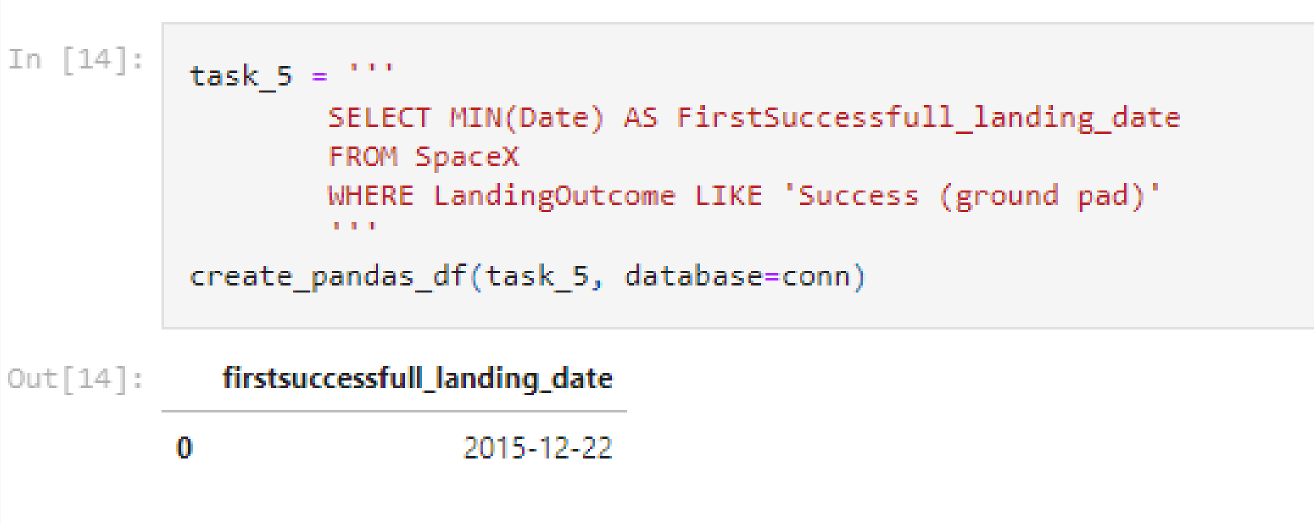
Display average payload mass carried by booster version F9 v1.1

```
In [13]:  task_4 = '''
                  SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
                  FROM SpaceX
                  WHERE BoosterVersion = 'F9 v1.1'
                  '''
          create_pandas_df(task_4, database=conn)
```

```
Out[13]:    avg_payloadmass
          0          2928.4
```

First Successful Ground Landing Date

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on12/22/2015.

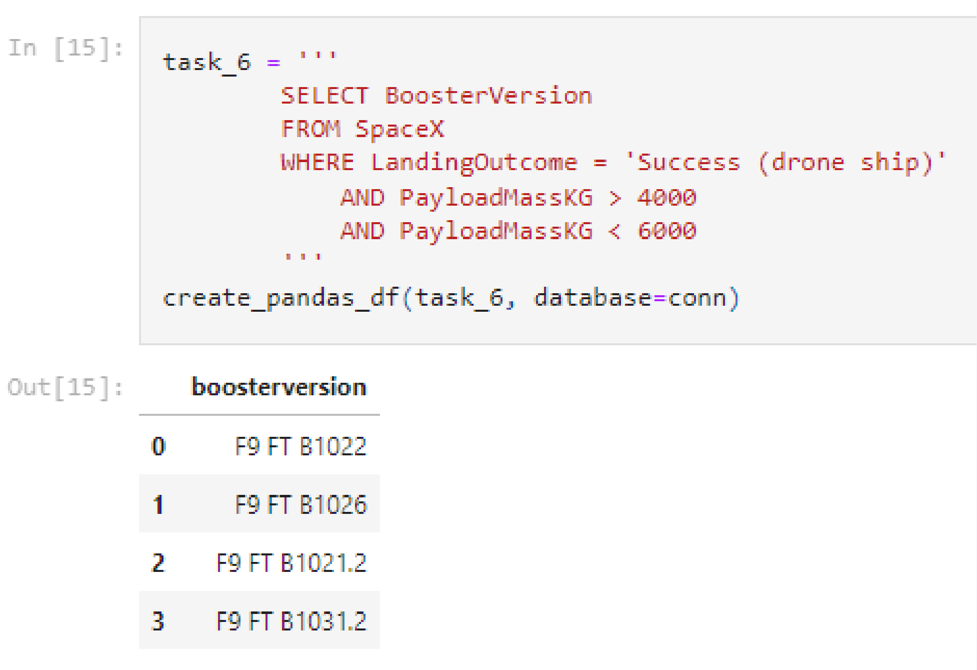


```
In [14]:  task_5 = '''
                  SELECT MIN(Date) AS FirstSuccessfull_landing_date
                  FROM SpaceX
                  WHERE LandingOutcome LIKE 'Success (ground pad)'
                  '''
          create_pandas_df(task_5, database=conn)
```

```
Out[14]:     firstsuccessfull_landing_date
          0                     2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

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



In [15]:

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

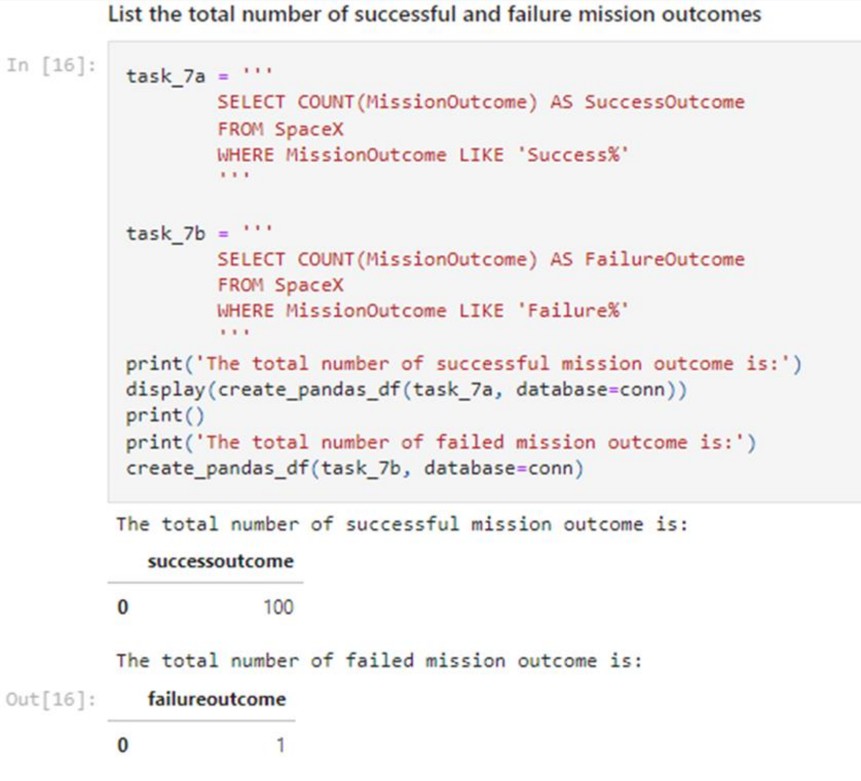
Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

- We used wildcard like ‘%’ to filter for **WHERE** MissionOutcome was a success or a failure.



List the total number of successful and failure mission outcomes

In [16]:

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''

print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

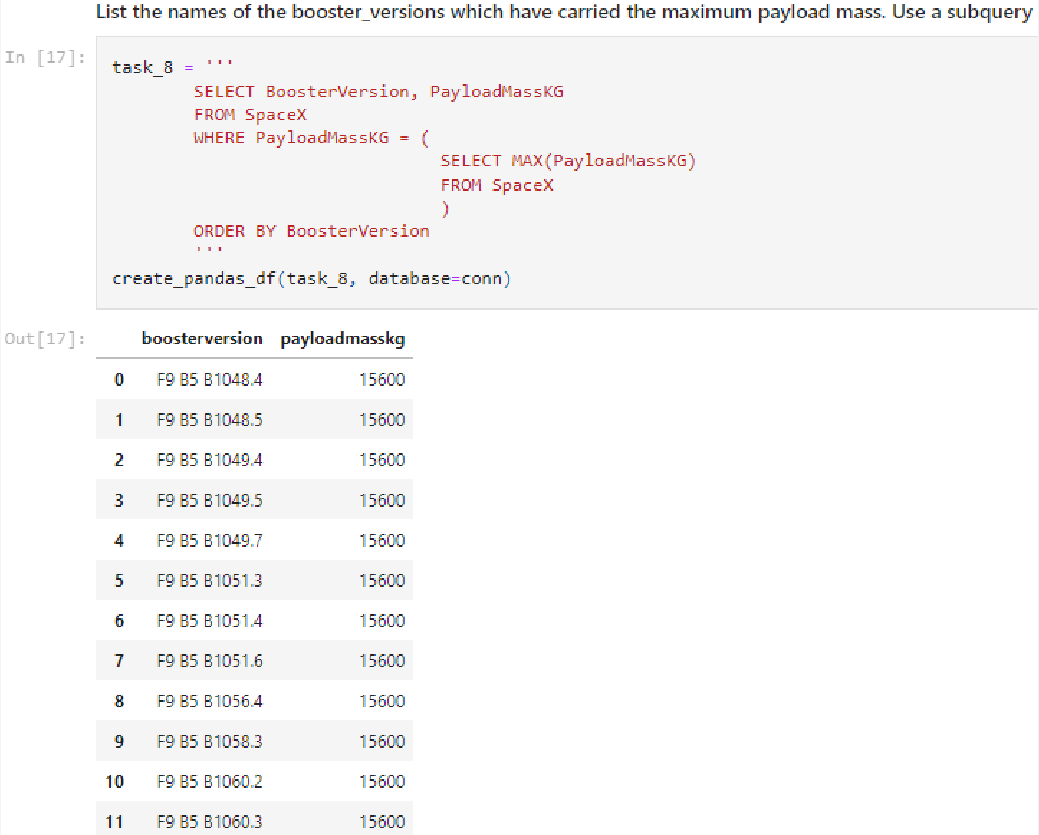
```
The total number of failed mission outcome is:
```

Out[16]:

	failureoutcome
0	1

Boosters Carried Maximum Payload

- 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 [17]:

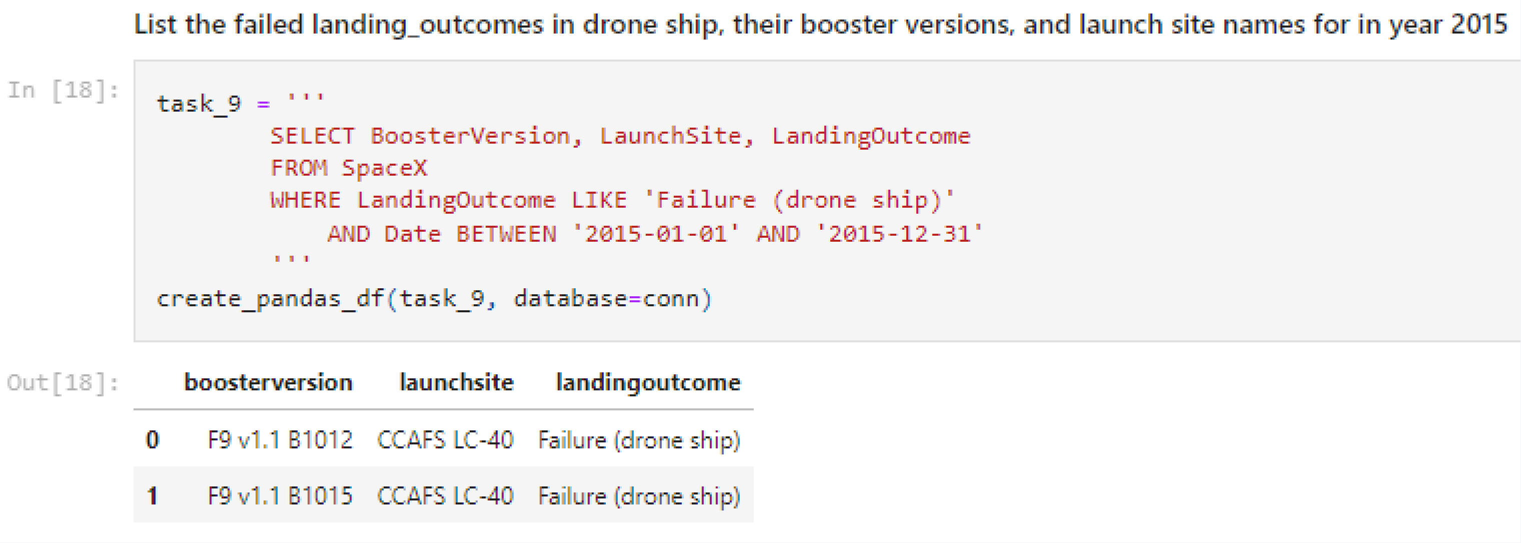
```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- We used a combinations of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015



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

In [18]:

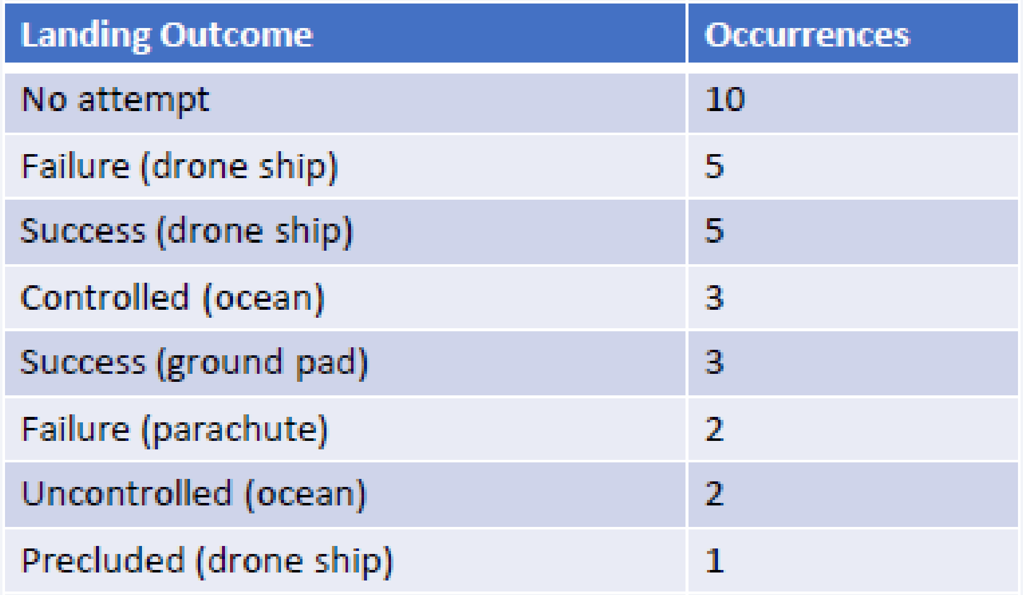
```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (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

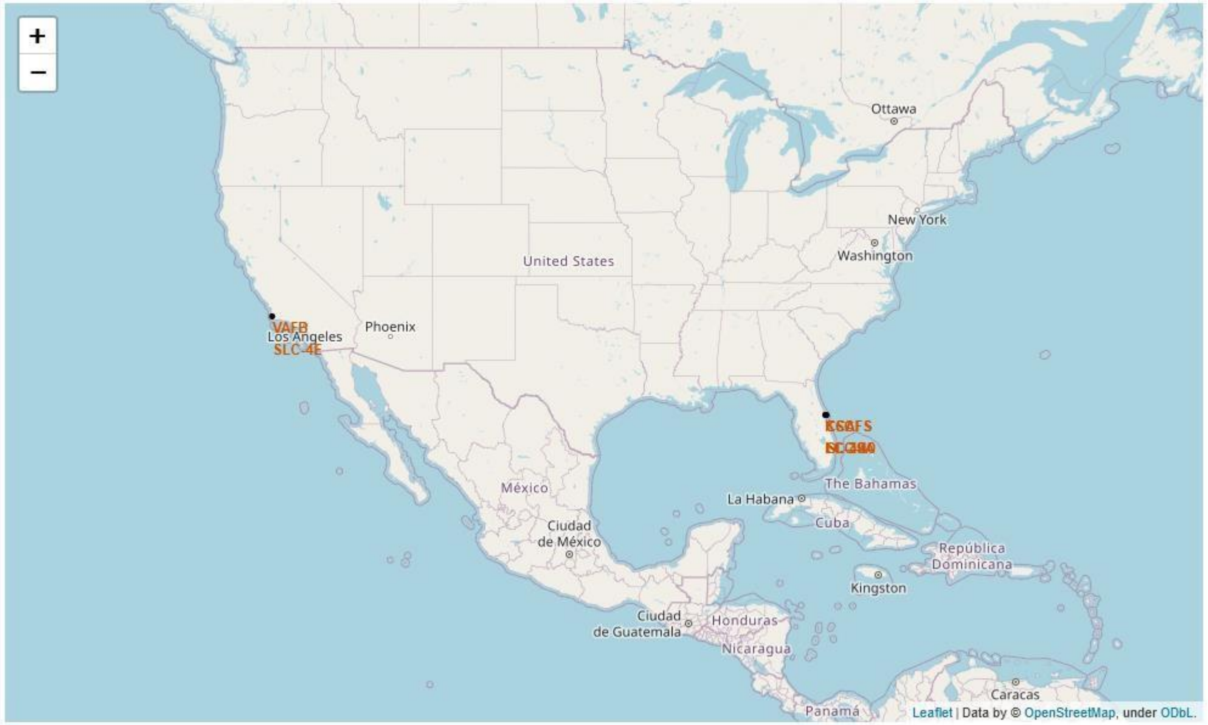


Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All launch sites global map markers



Launch sites are near sea, probably by safety, but not too far from road and railroads.

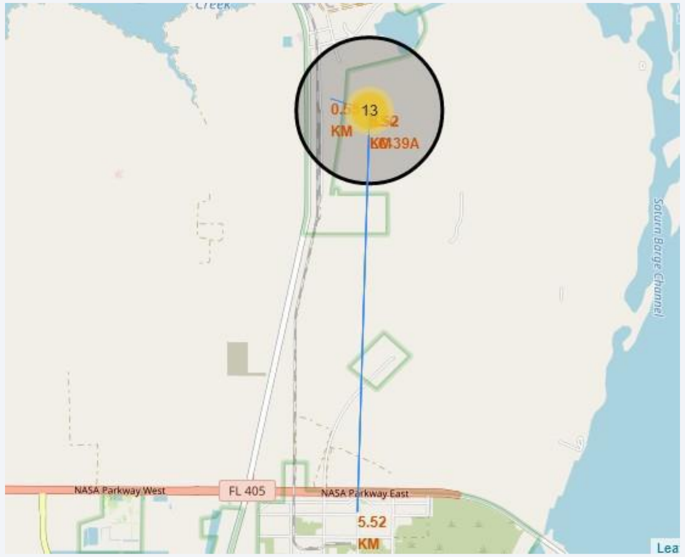
Launch Outcomes by Site

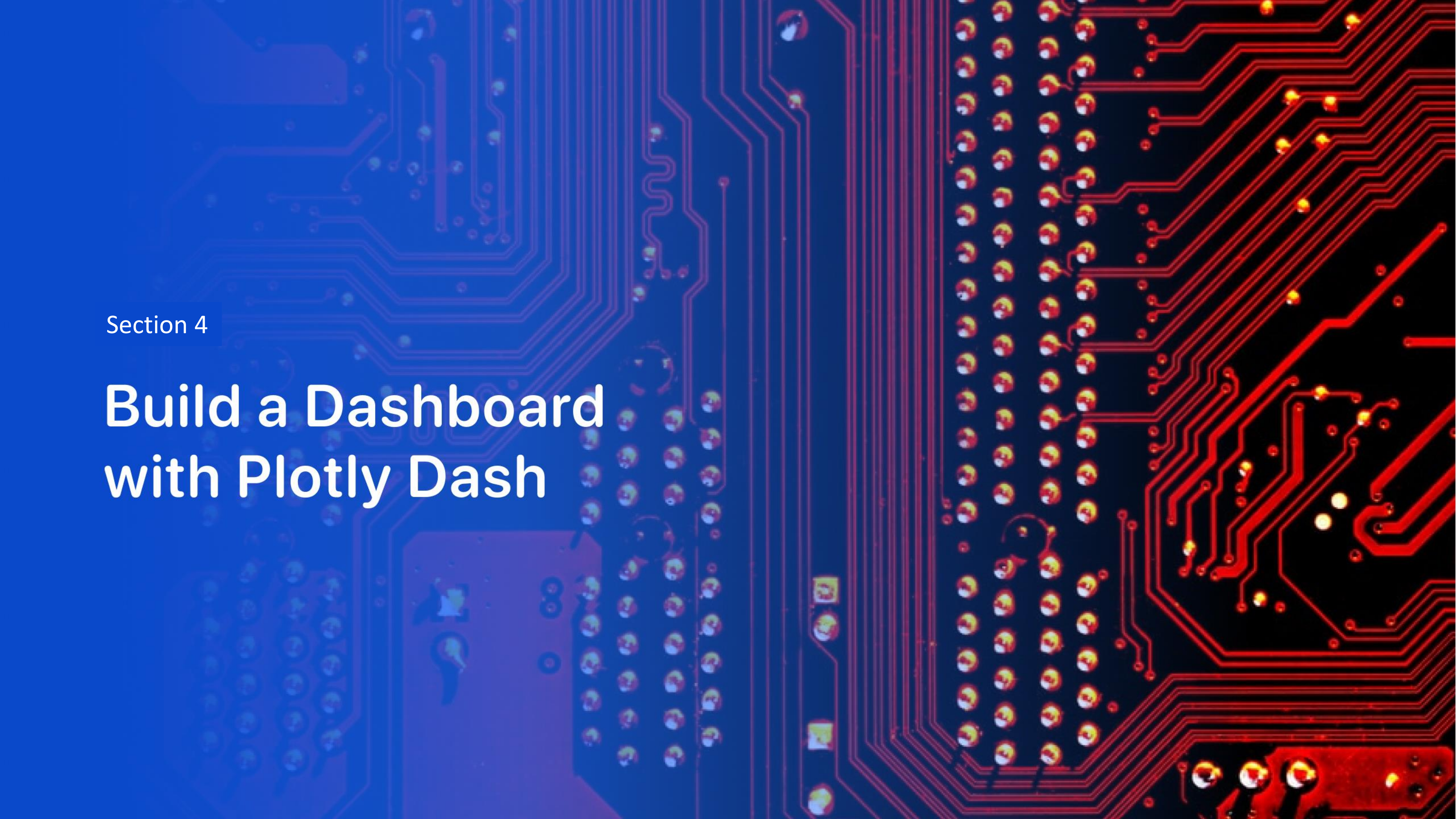
- Example of KSC LC-39A launch site launch outcomes
- Green markers indicate successful and red ones indicate failure.

Las Vegas
10
SLC-4E
Angeles
Tijuana
Mexicali
13
LC-39A
77
Memphis
Charlotte
Atlanta
Georgia
Rouge
Jacksonville
46

Logistics and Safety

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.



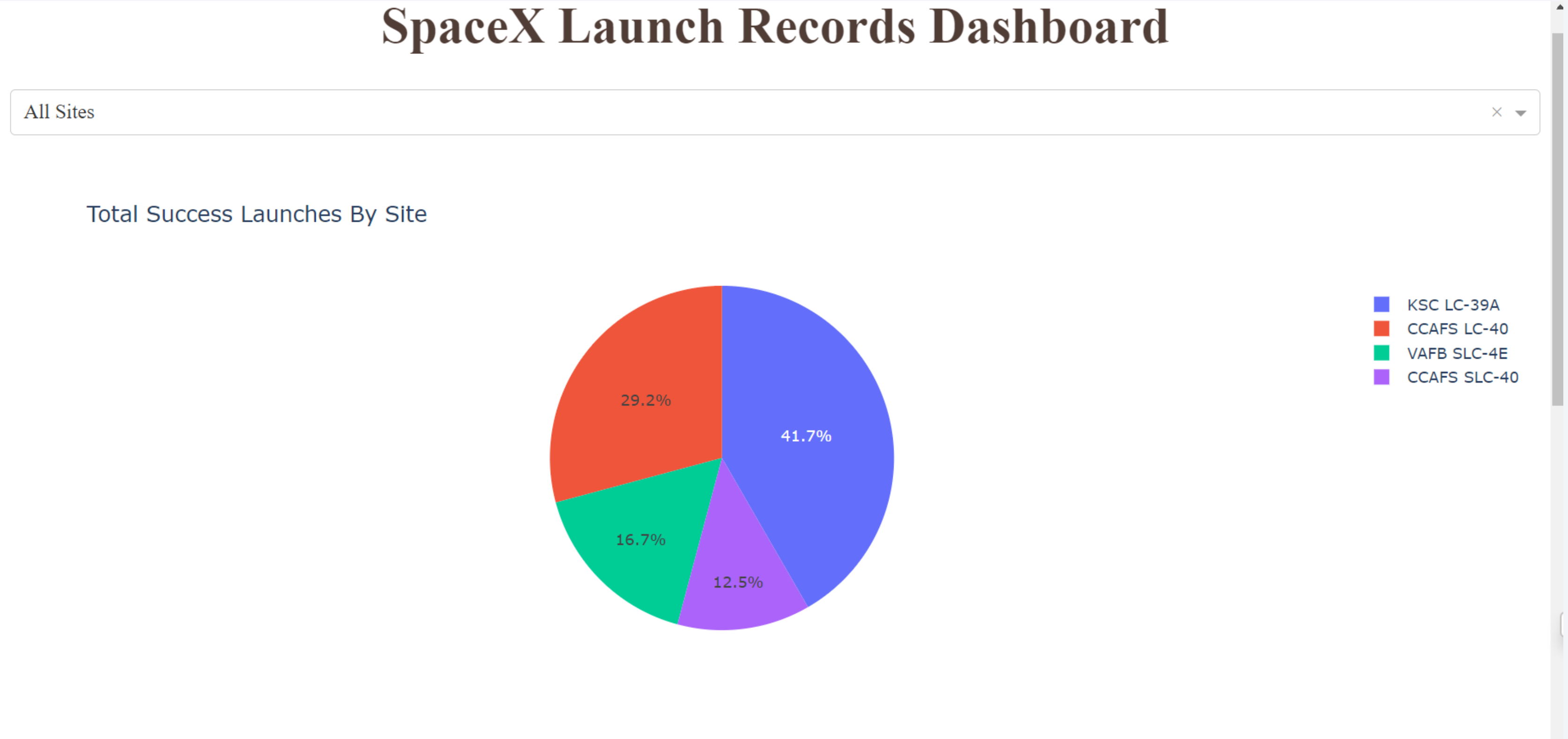


Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

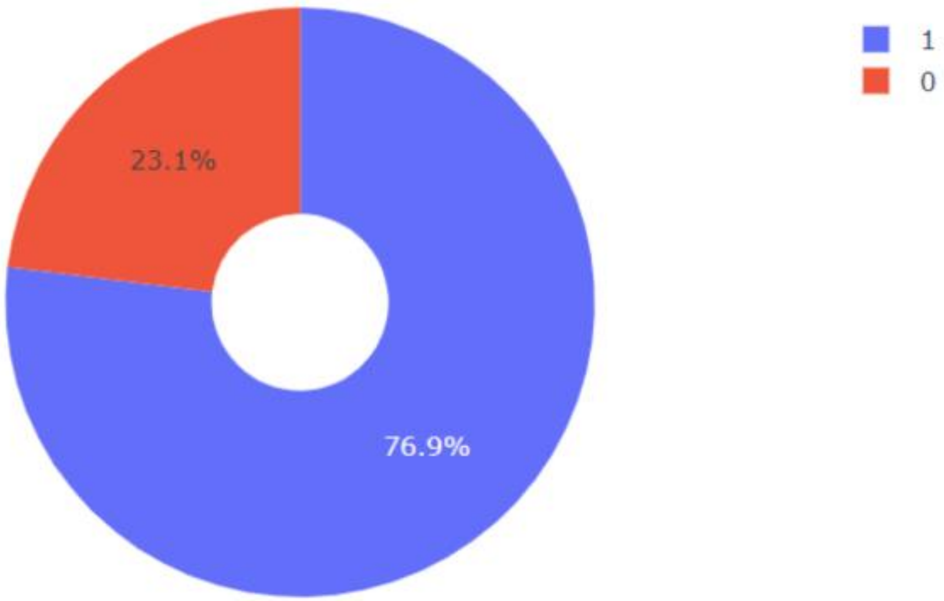
- The place from where launches are done seems to be a very important factor of success of missions.



SpaceX Launch Records Dashboard

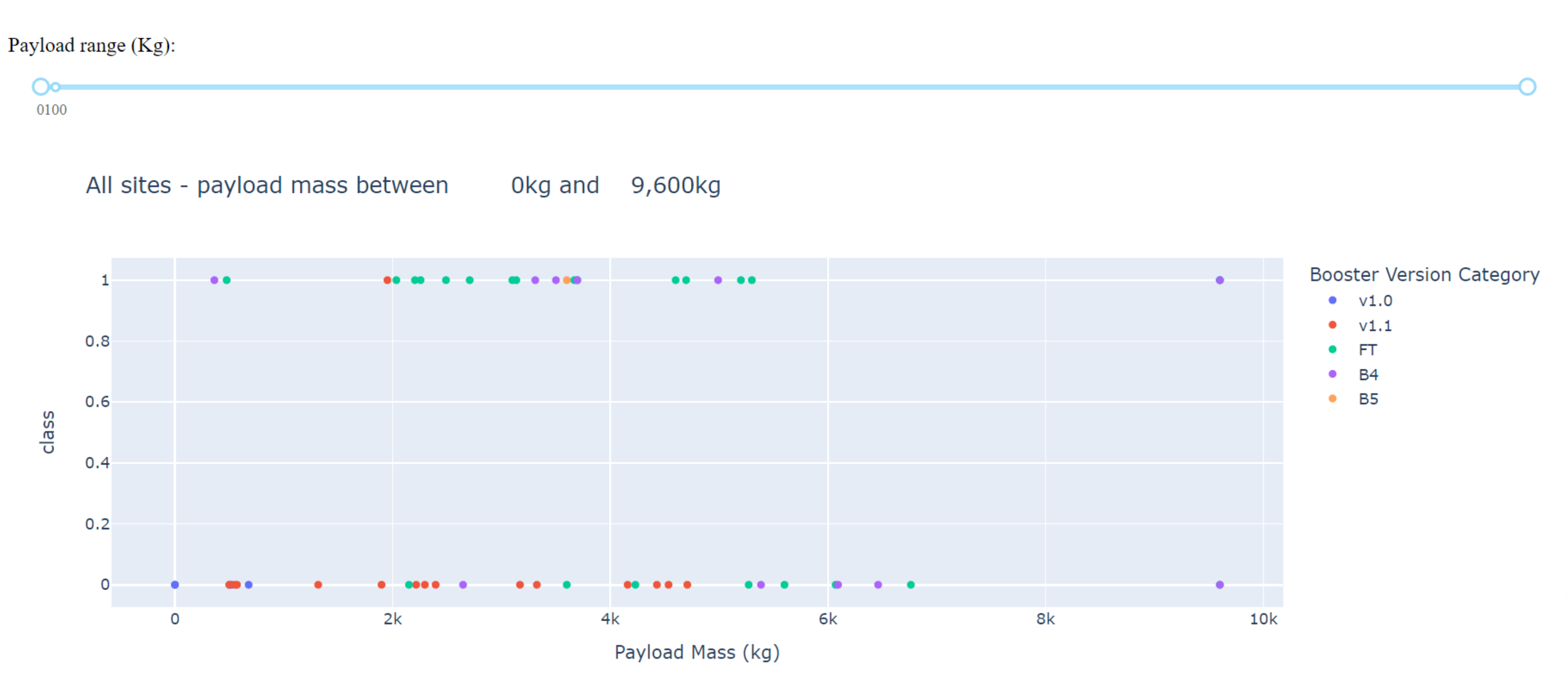
Launch Success Ratio for KSCLC-39A

- 76.9% of launches are successful in this site.



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs. Launch Outcome



Payload range (Kg):

0100

All sites - payload mass between 0kg and 9,600kg

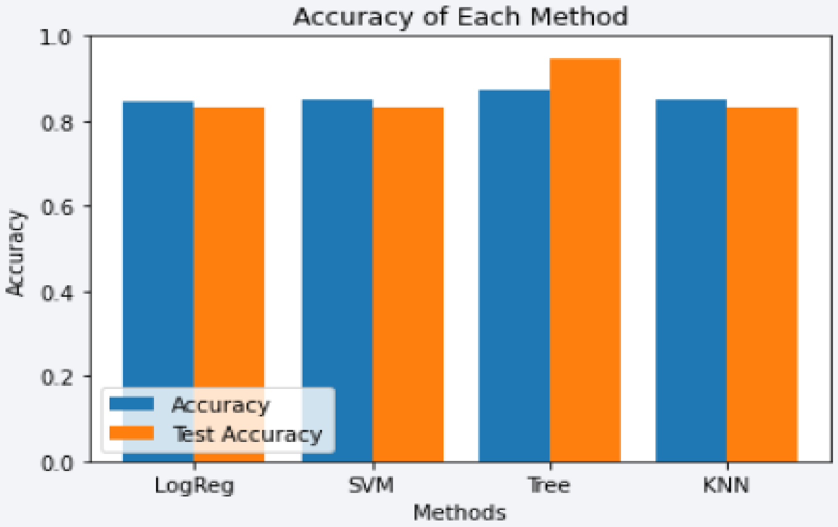
Section 5

Predictive Analysis (Classification)

Classification Accuracy

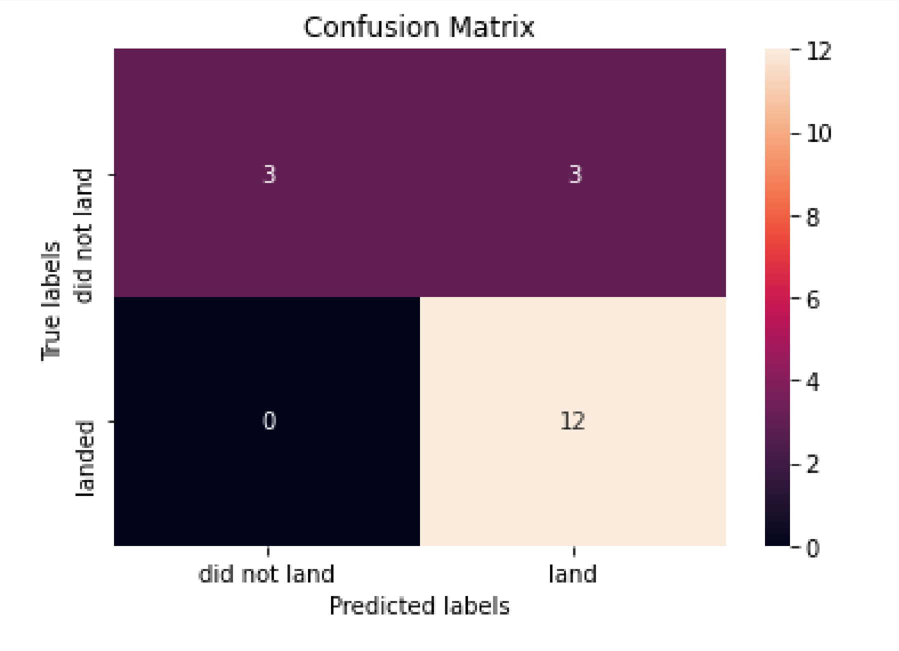
- Four classification models were tested, and their accuracies are plottedbeside;

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than87%.



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- Different data sources were analyzed,refining conclusions along the process;
- The best launch site is KSCLC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and Increase profits.

Thank you!