IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Izal Jibrilly Winnar
26th February 2022

Outline

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

Executive Summary

- First methodology is collecting data from public SpaceX API and SpaceX Wikipedia web page through API and Web Scraping. Data Wrangling from SpaceX dataset. Exploratory data analysis with SQL and Data Visualization. Find a successful landing using visual analytics with Folium. Predict next successful landing using Machine Learning Prediction.
- Results in this presentation are exploratory data analysis, interactive visual analytics and result from machine learning prediction

Introduction

- Project background and context

SpaceX 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 due to the fact that SpaceX can reuse the first stage. We can predict cost of the launch and factor that affect success rate of landing. The purpose of this project is to predict a successful rocket landing with machine learning.

- Problems you want to find answers

1. What factors that affect rocket will land successfully ?
2. Where is the optimal location if want to landing successfully ?
3. Is there any operating condition to ensure successful landing ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected from SpaceX API and Web Scraping from Wikipedia
- Perform data wrangling
  - Using one hot encoding to categorize landing 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
  - Using SVM, Decision Trees, k-Nearest Neighbors and Logistic Regression

Data Collection

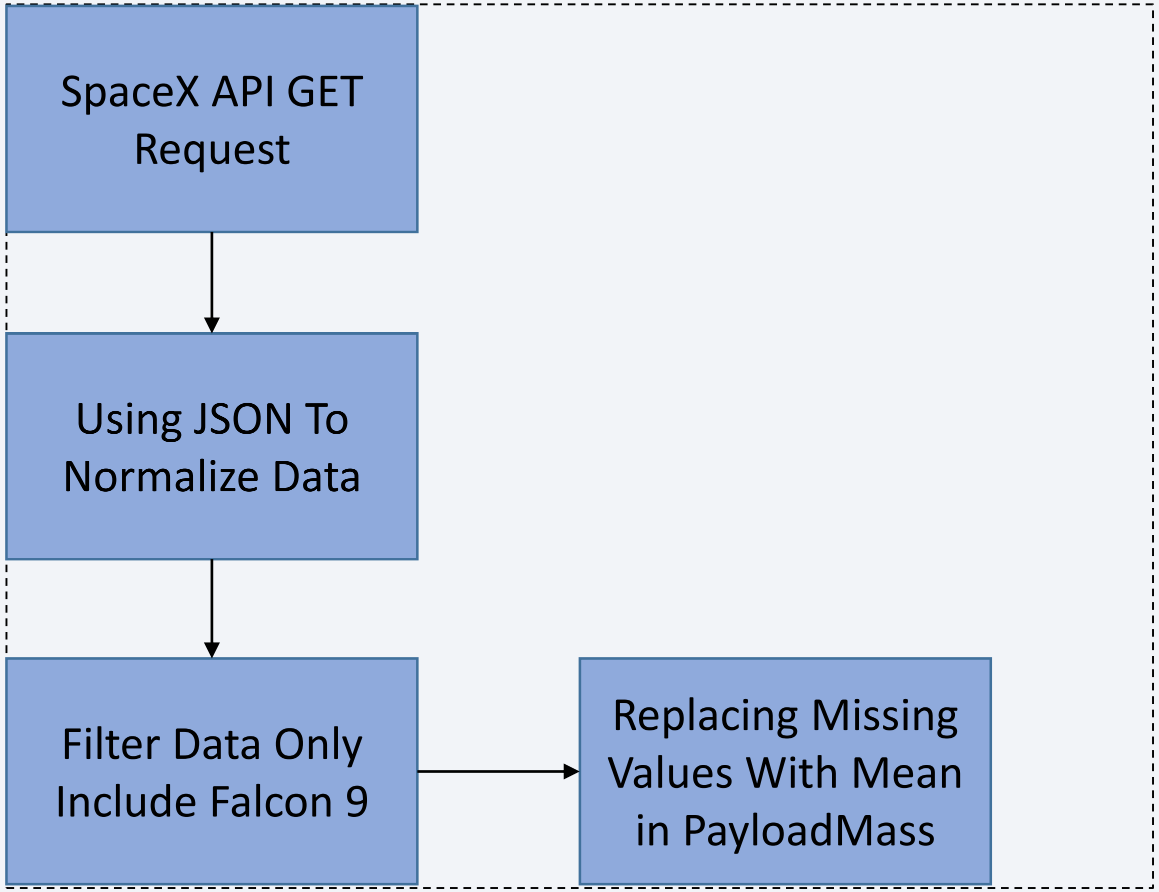
- Describe how data sets were collected.

1. First collect data with requesting to SpaceX API using GET request
2. Filtering dataframe that only include Falcon 9 launches
3. Dealing with missing values
4. Calculate mean for the PayloadMass to replace missing value in PayloadMass column

Data Collection – SpaceX API

- Using GET request to collect data, clean data and did data formatting and data wrangling

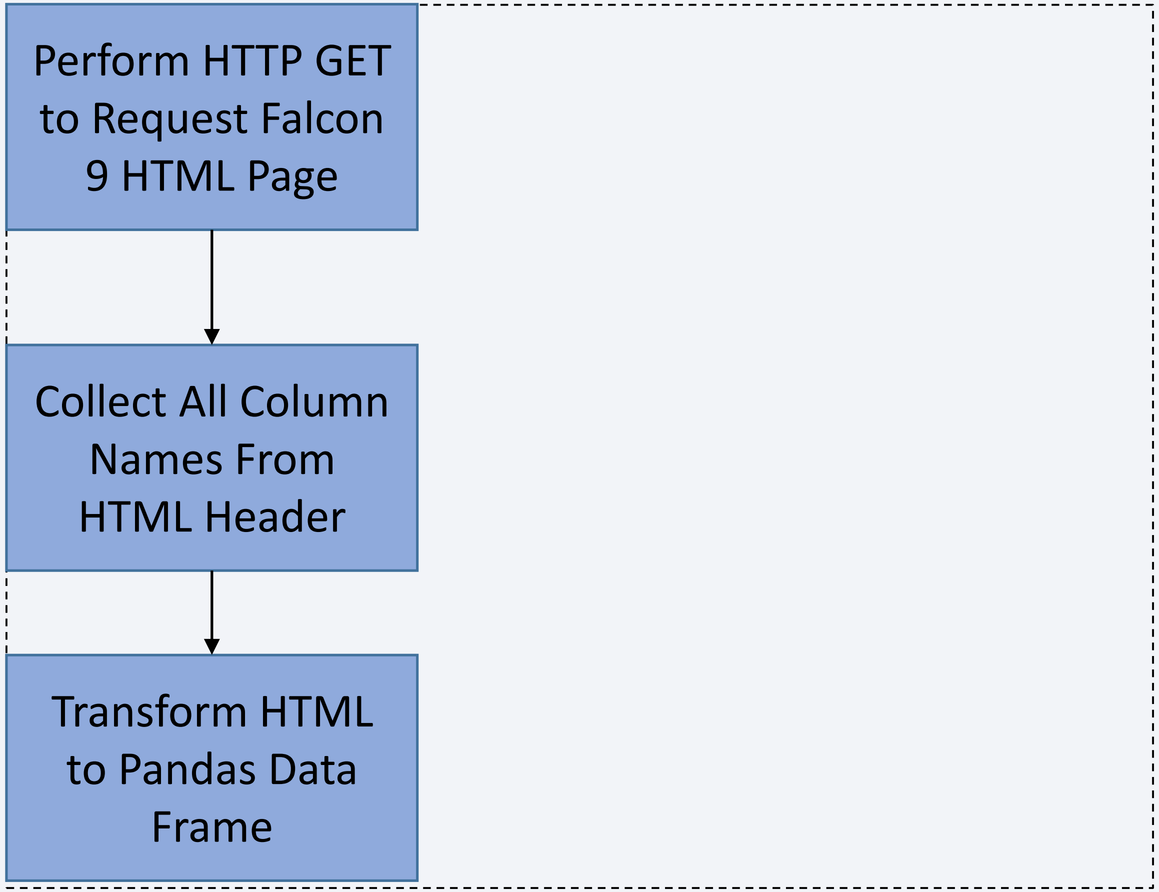
- The link to the notebook is https://github.com/izaljibrilly/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb


SpaceX API GET Request
Using JSON To Normalize Data
Filter Data Only Include Falcon 9
Replacing Missing Values With Mean in PayloadMass

Data Collection - Scraping

- Request data from Wikipedia using BeautifulSoup, Extract Falcon 9 Launch, Transform to Data Frame

- The link to the notebook is https://github.com/izaljibrilly/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20Web%20Scraping.ipynb



Perform HTTP GET to Request Falcon 9 HTML Page

Collect All Column Names From HTML Header

Transform HTML to Pandas Data Frame

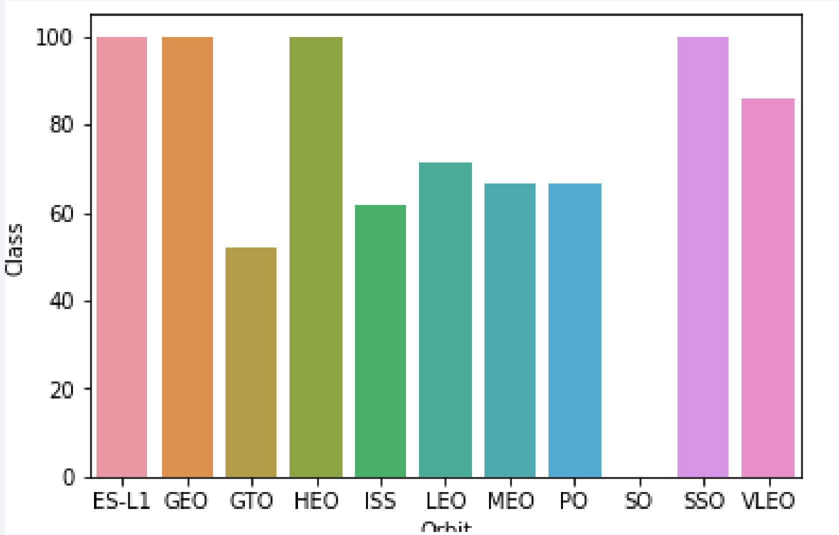
Data Wrangling

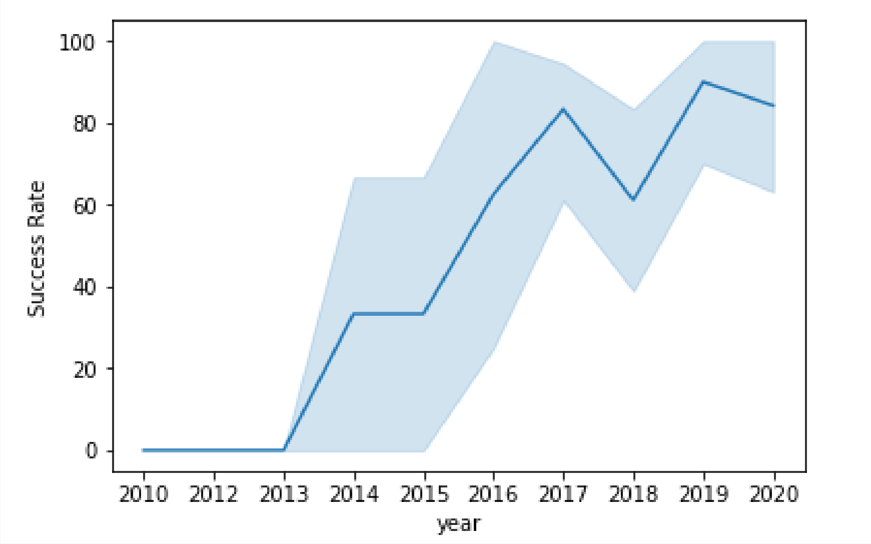
- Exploratory Data Analysis is to find some patterns in the data and determine what would be the label for training supervised models. In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident.
- First calculate number of launches each sites, calclate number and occurence of each orbit, calculate number and occurence of mission outcome per orbit type, and create landing outcome label.
- The link to the notebook is https://github.com/izaljibrilly/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis performed by visualizing relationship between flight number and payload mass, flight number and launch site, launch site and payload mass, and success rate of each orbit type.

- Link to the notebook is https://github.com/izaljibrilly/Applied-Data-Science-Capstone/blob/main/EDA%20Data%20Visualization.ipynb





EDA with SQL

- We performed SQL queries to obtain some data like :

1. Launch sites in space mission
2. Total payload mass carried by booster
3. Failed landing outcome and their booster versions
4. Booster which carried maximum payload mass
5. Count of landing outcome success or failure

- Link to the notebook is https://github.com/izaljibrilly/Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We created map that marked all launch sites, and using marker we can see where place has a successful landing and failure landing
- Marker, circles and lines to identify success or failed launches for each site on the map. Then do some calculations the distances between a launch site to its proximities such as railway and highway
- The link to the notebook is https://github.com/izaljibrilly/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20Folium.ipynb

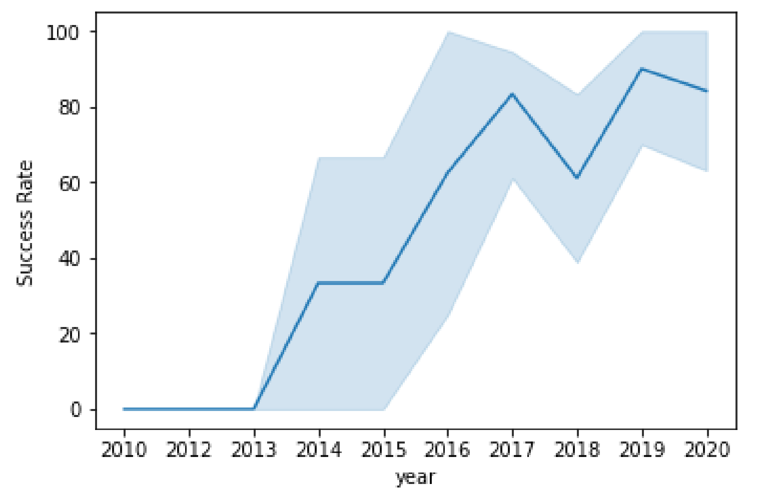
Build a Dashboard with Plotly Dash

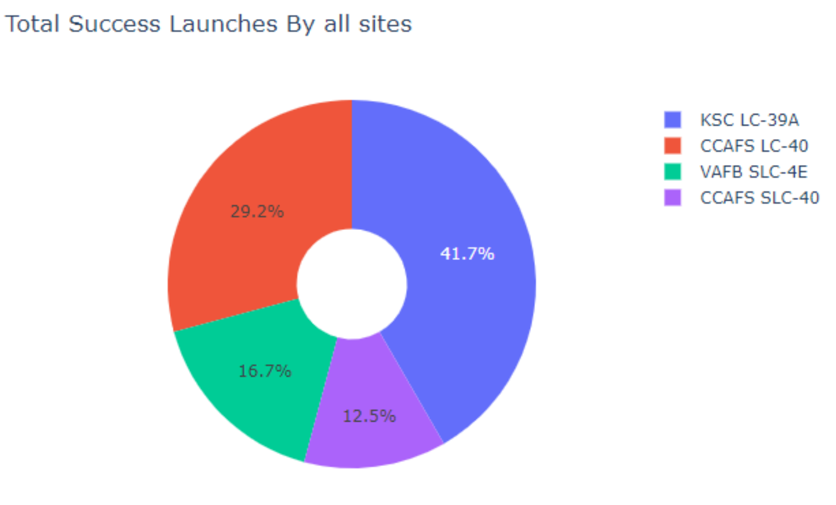
- We built a dashboard using Plotly Dash
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- The link to the notebook is https://github.com/izaljibrilly/Applied-Data-Science-Capstone/blob/main/app.py

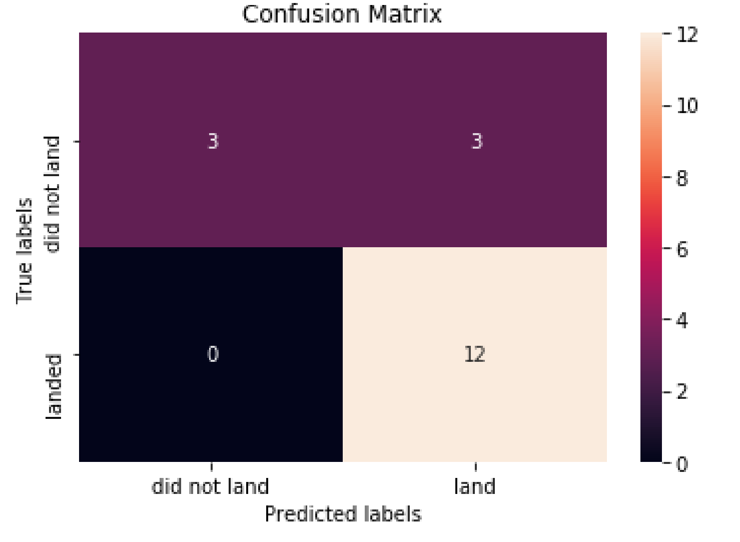
Predictive Analysis (Classification)

- We performed data standardize and take the data to train and test split
- Algorithms that used to predict rocket landing are Logistic Regression, Support Vector Machine, Decision Tree Classifier and K Nearest Neighbors Classifier
- The link to the notebook is https://github.com/izaljibrilly/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results




Total Success Launches By all sites

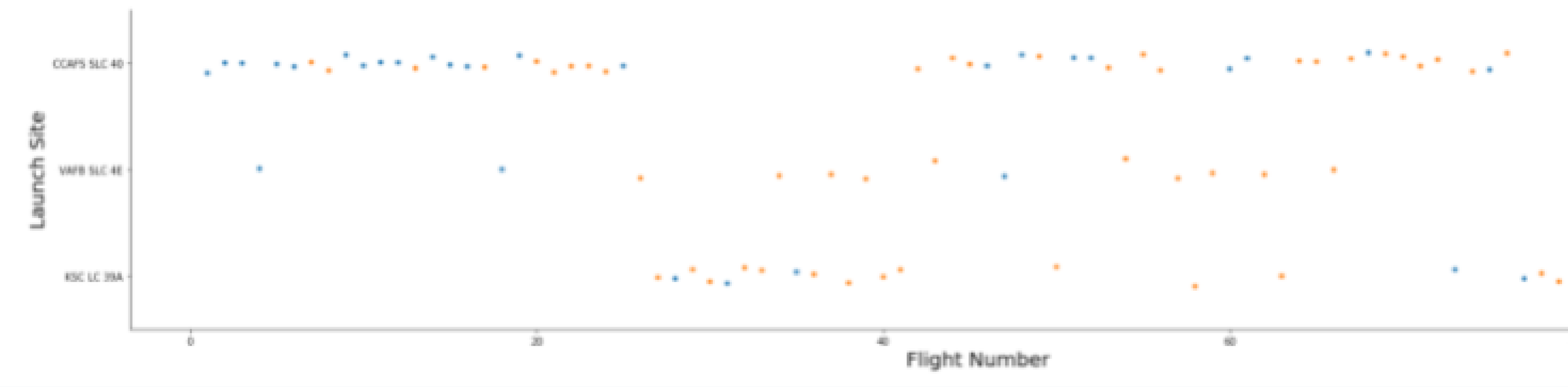

Confusion Matrix

Section 2

Insights drawn from EDA

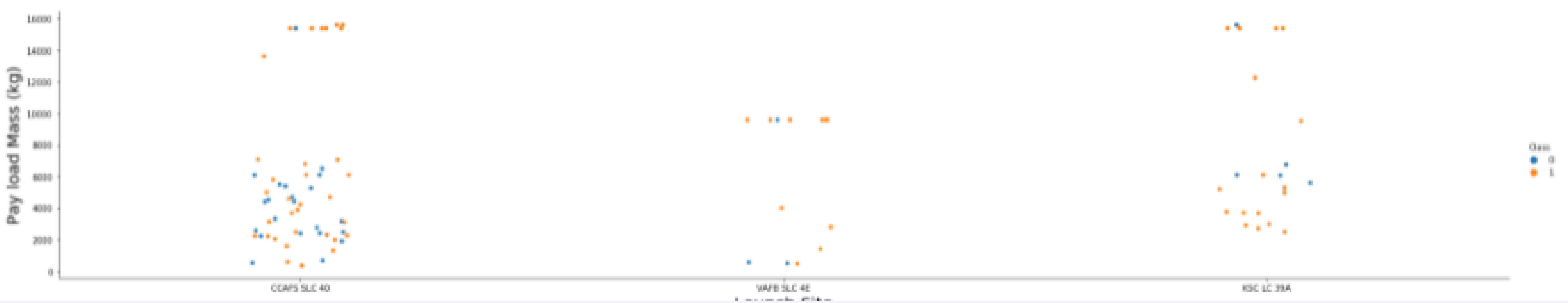
Flight Number vs. Launch Site

- From the plot, we can know relationship between flight number and launch site. Launch site CCAFS LFC 40 have a more flight number than other launch sites



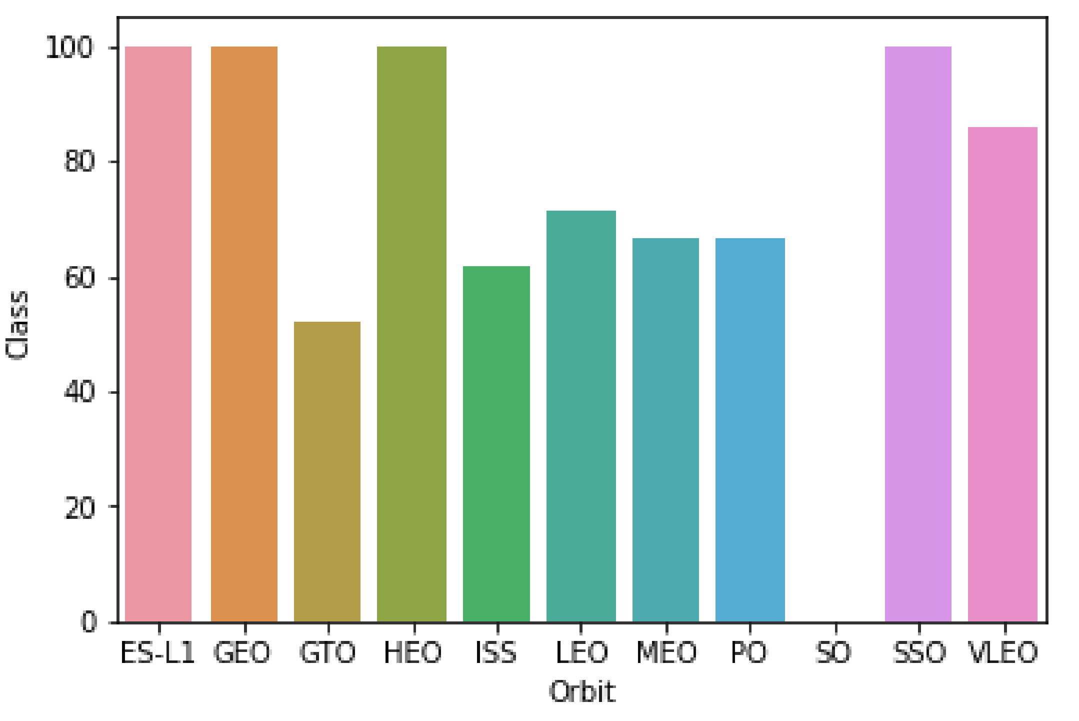
Payload vs. Launch Site

- From the plot, we can know relationship between payload and launch site. There is more launch site in flight number CCAFS LFC 40 that carry payload less than 8000 Kg



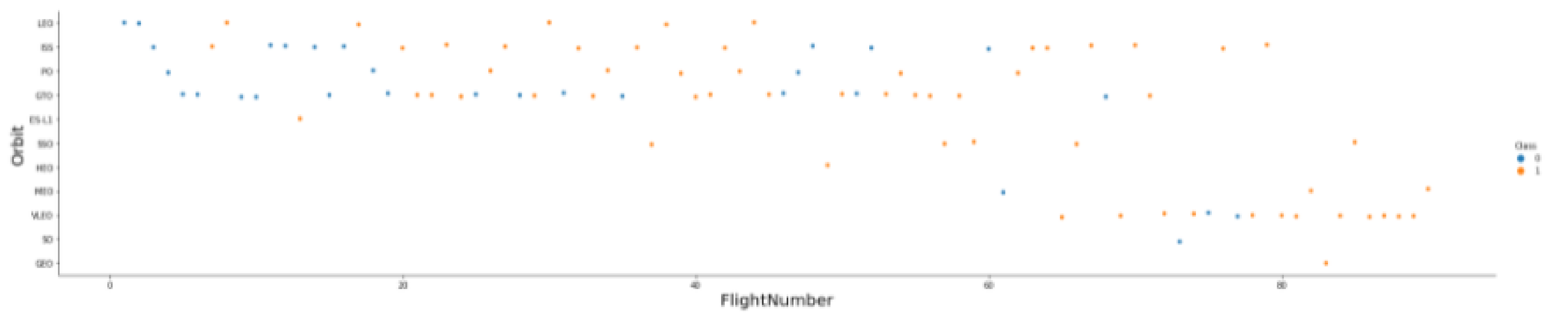
Success Rate vs. Orbit Type

- From the plot, we can know relationship between success rate and orbit type. There are 100% successful landing on ES-L1 orbit, Geosynchronous Equatorial Orbit, Highly Elliptical Orbit and Sun Synchronous Orbit.



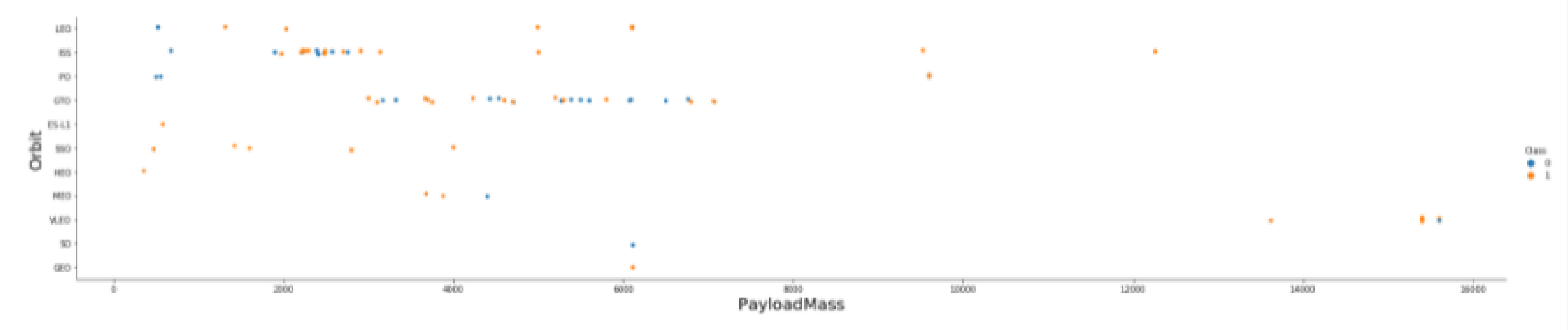
Flight Number vs. Orbit Type

- From the plot, we can know relationship between flight number and orbit type. There is a high success rate in Low Eart Orbit than other type of orbit.



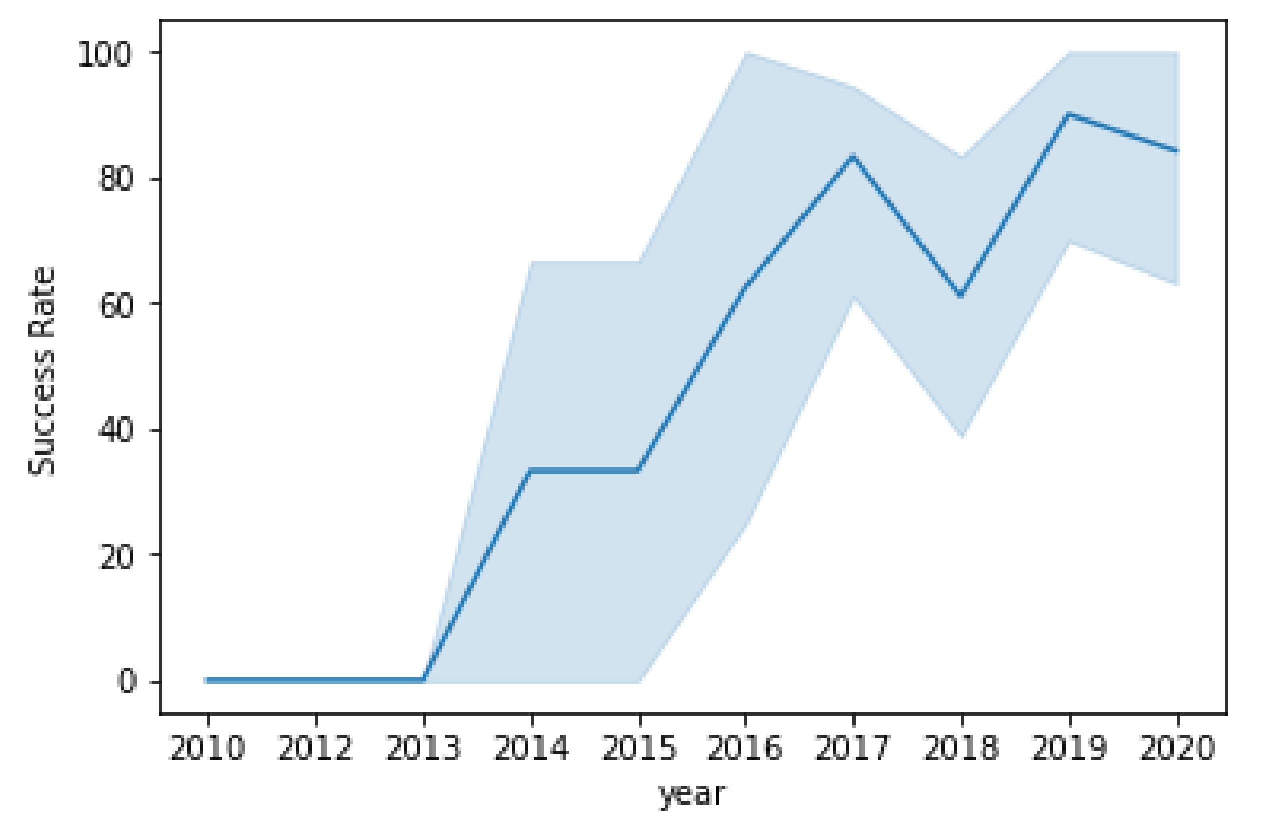
Payload vs. Orbit Type

- From the plot, we can know relationship between payload and orbit type. There is high successful landing on heavy payload in Low Earth Orbit and International Space Station Orbit.



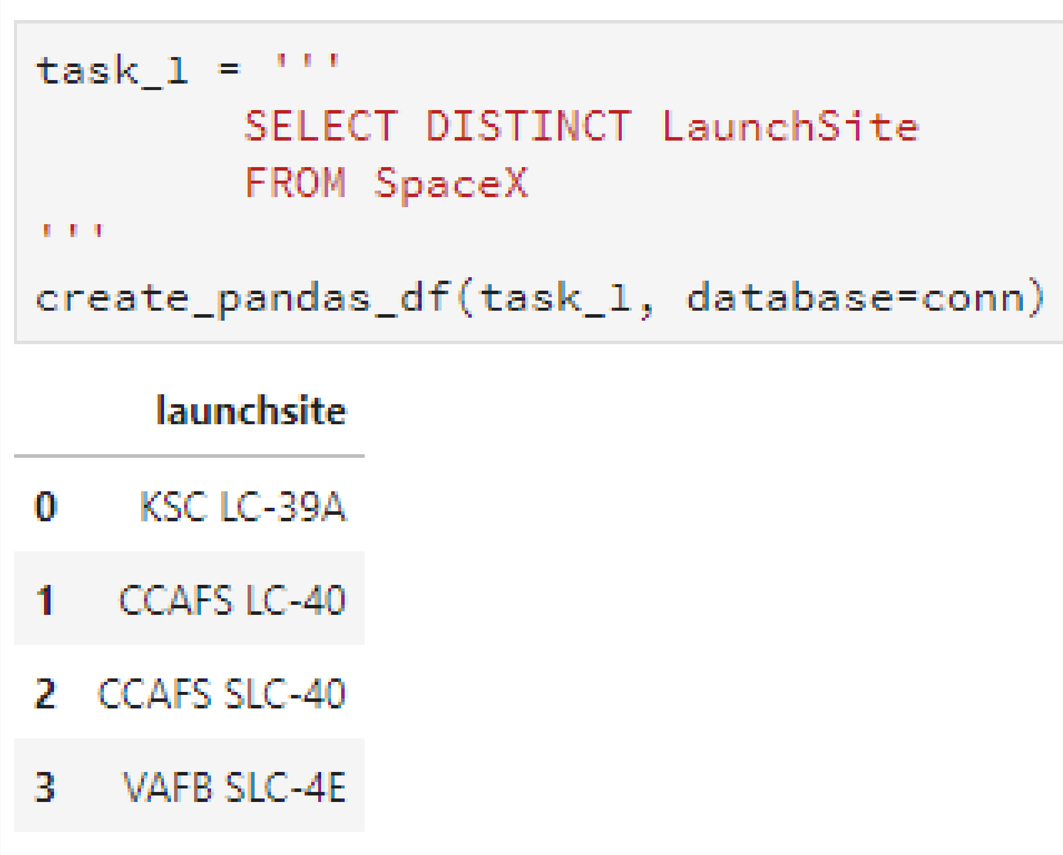
Launch Success Yearly Trend

- From the trend we can know that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

- We can get all of launch site names with DISTINCT query, and the result there are 4 launch site

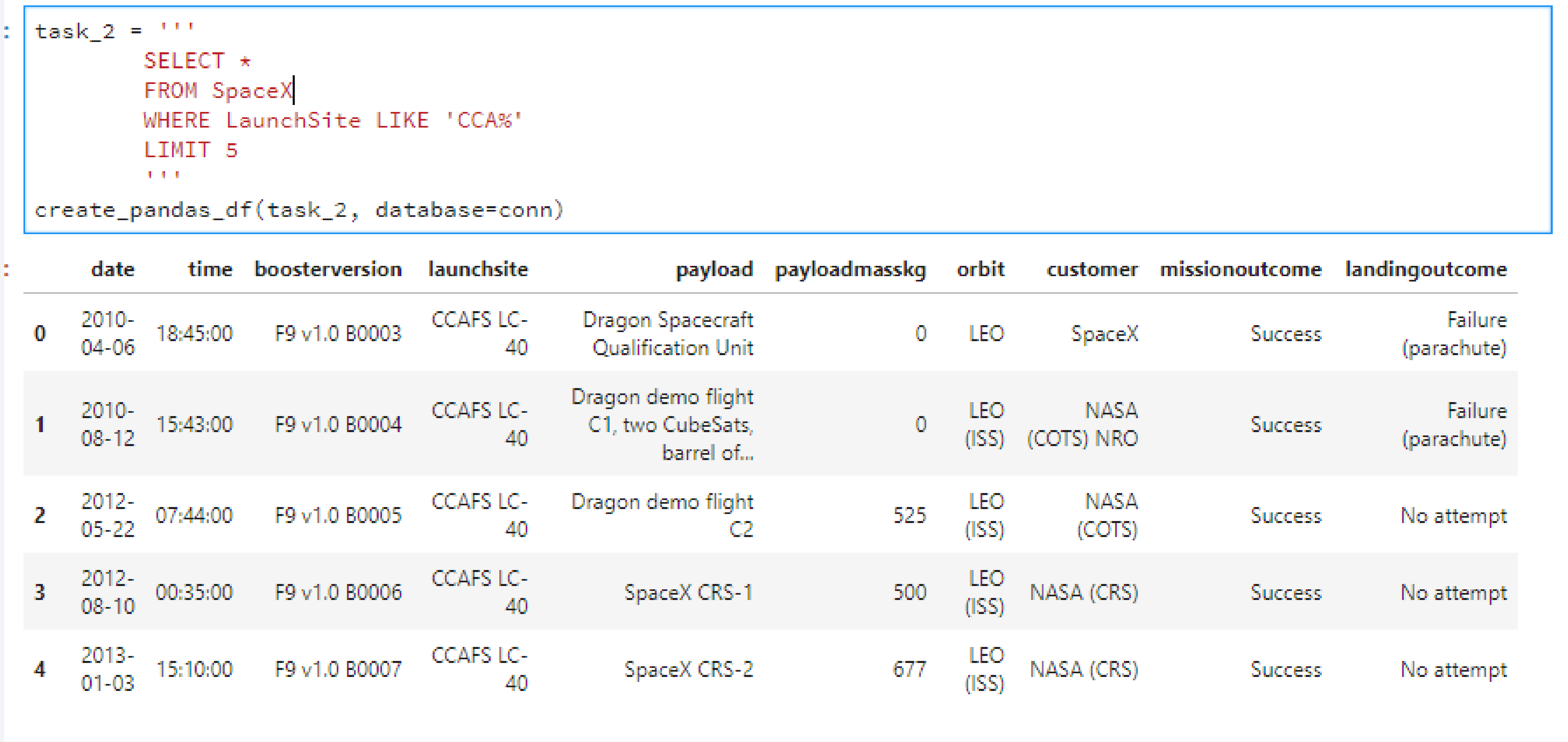


```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

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

Launch Site Names Begin with 'CCA'

- We can get launch site names begin with CCA using WHERE and LIKE syntax. There are 5 results and the name is CCAFS LC-40

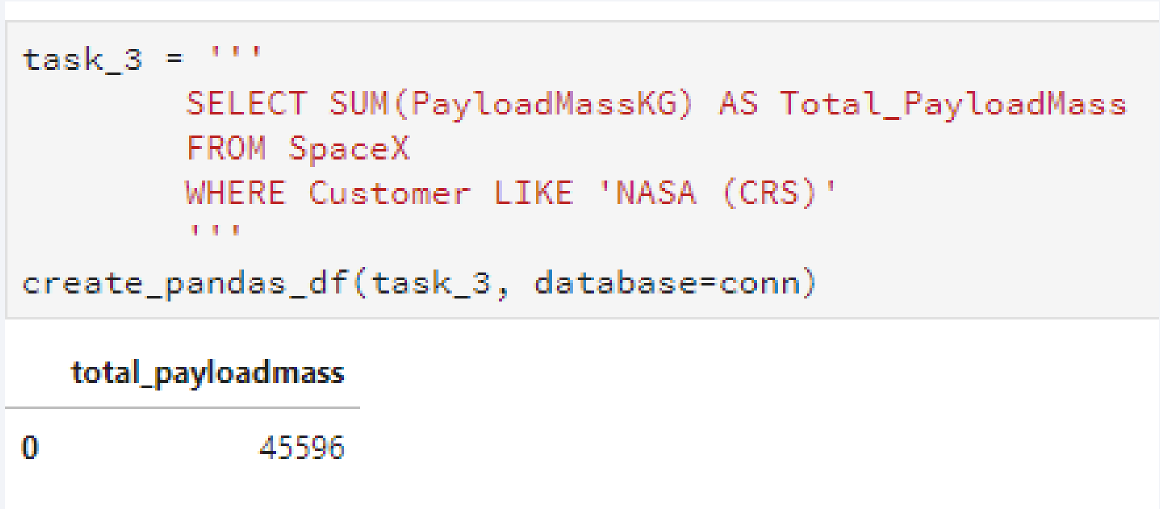


```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

	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 can calculate total payload mass using SUM syntax to sum PayloadMassKG column then using LIKE syntax to find customer like NASA CRS

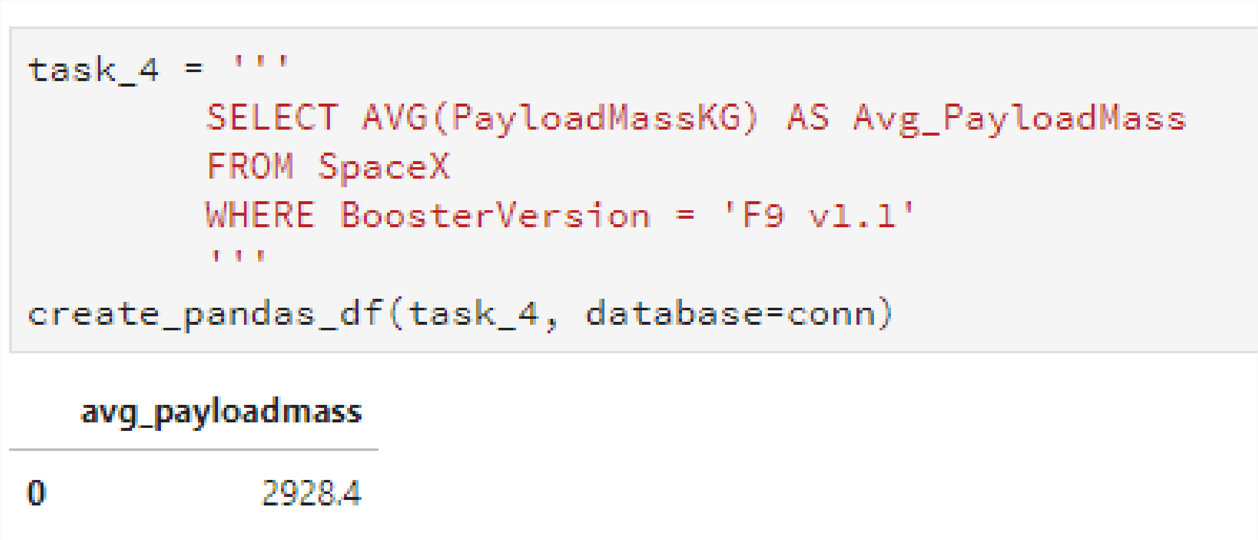


```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- We can calculate average payload mass by booster version F9 v1.1 using AVG syntax and WHERE clause to find F9 v1.1

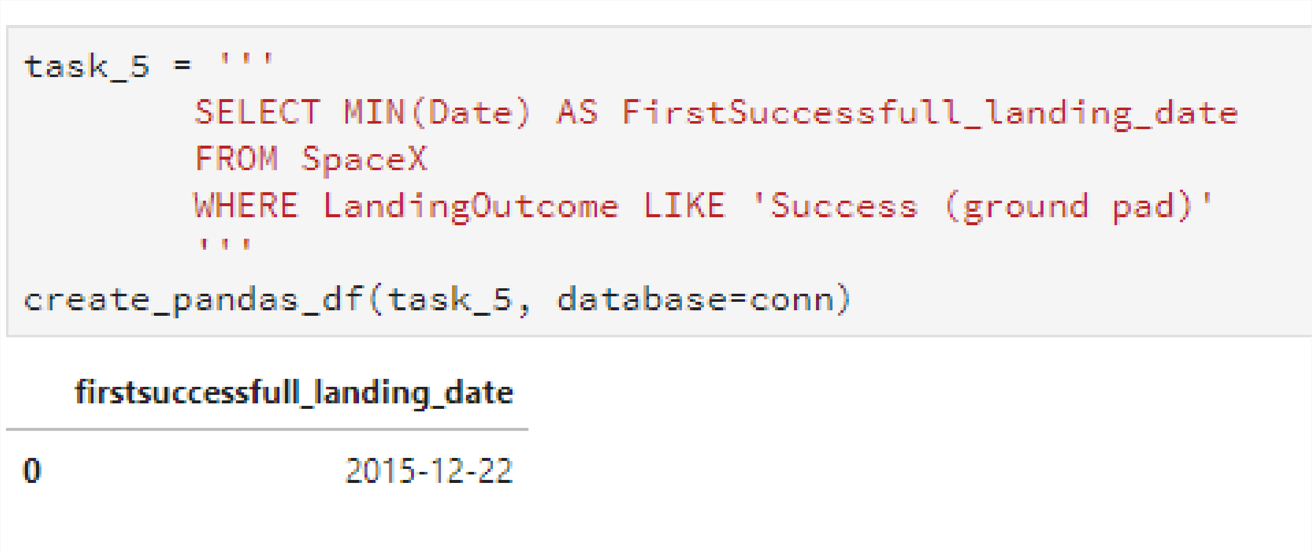


```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

- We can find first successful ground landing by finding minimum date and using LIKE clause to find 'Success (ground pad)'

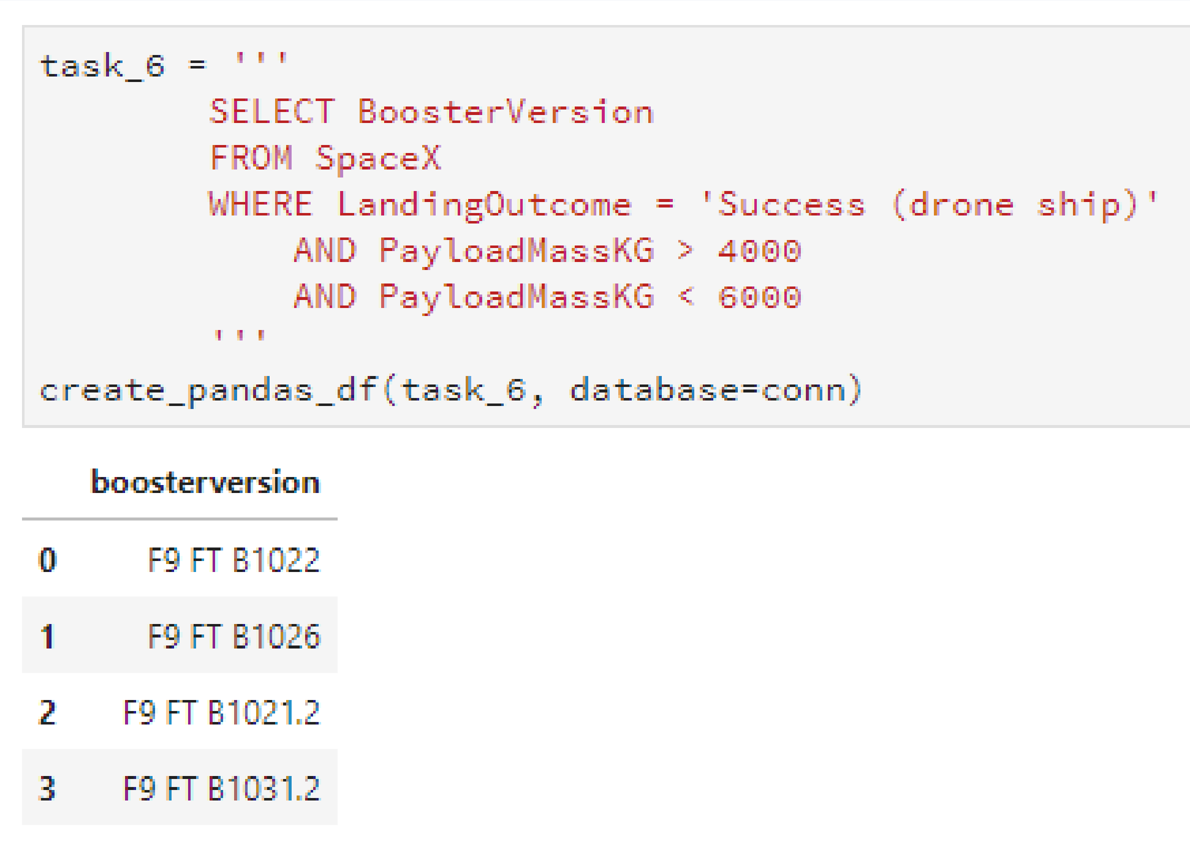


```
task_5 = '''
       SELECT MIN(Date) AS FirstSuccessfull_landing_date
       FROM SpaceX
       WHERE LandingOutcome LIKE 'Success (ground pad)'
       '''
create_pandas_df(task_5, database=conn)
```

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We can find successful drone ship landing using WHERE clause to find ‘success (drone ship)’ and limit the payload value using AND syntax greater than 4000 and lower than 6000

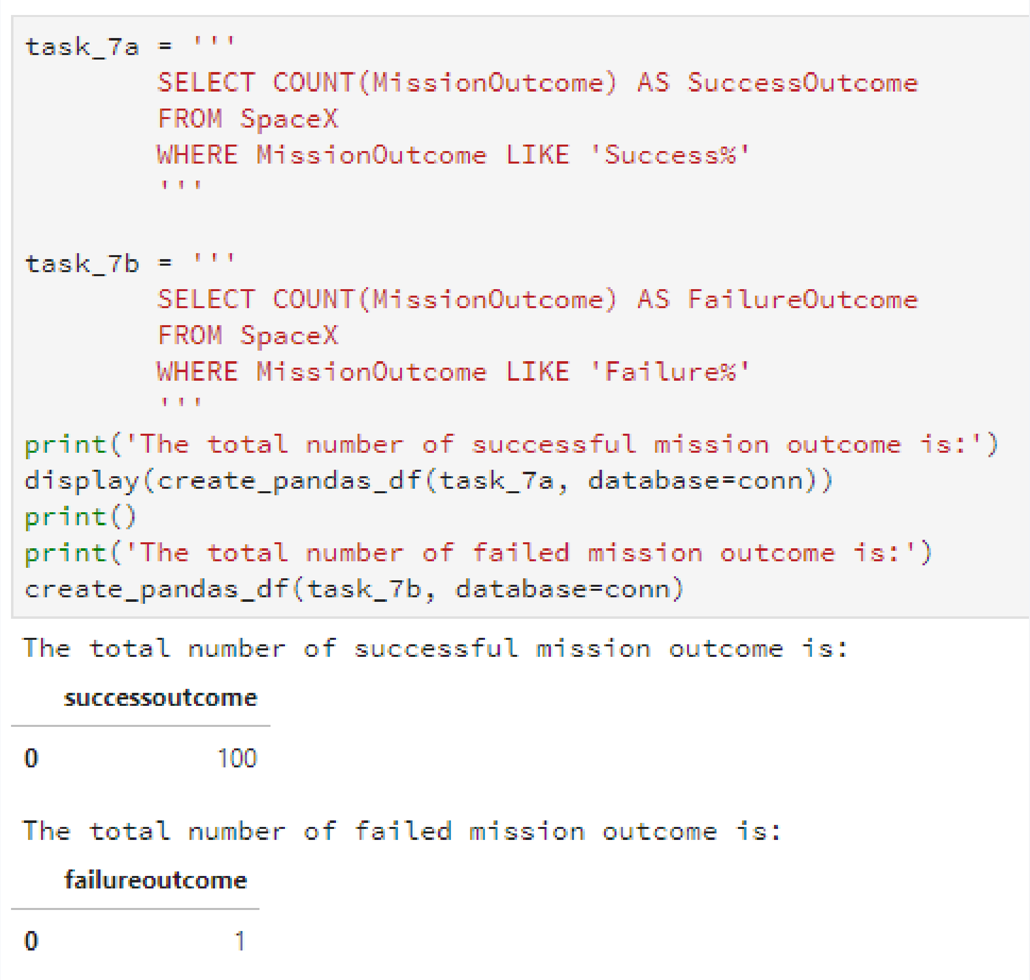


```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	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

- We can find total number of successful and failure using COUNT syntax and LIKE clause to find ‘Success%’ and ‘Failure%’ in Mission Outcome column



```
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:
```

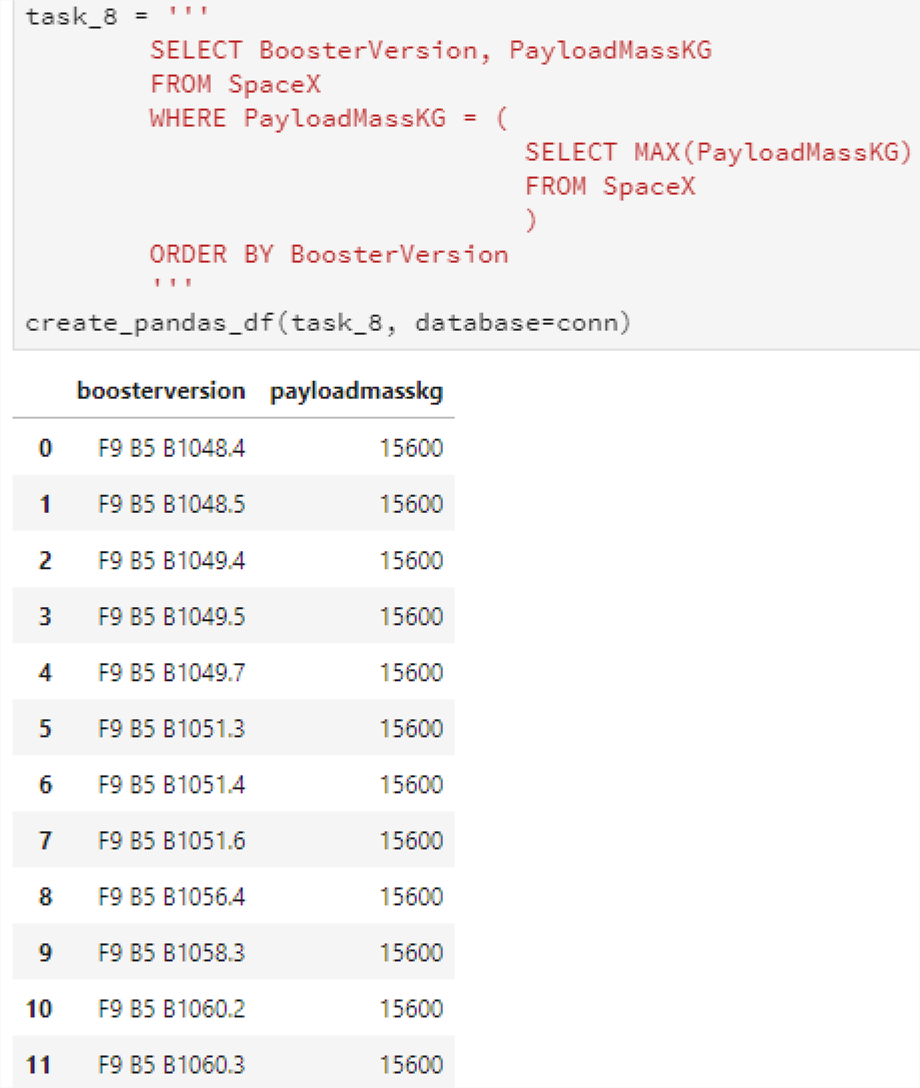
	successoutcome
0	100

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

	failureoutcome
0	1

Boosters Carried Maximum Payload

- We can list booster that carried maximum payload using MAX syntax in subquery and order it by booster version. There are 12 result of booster version.

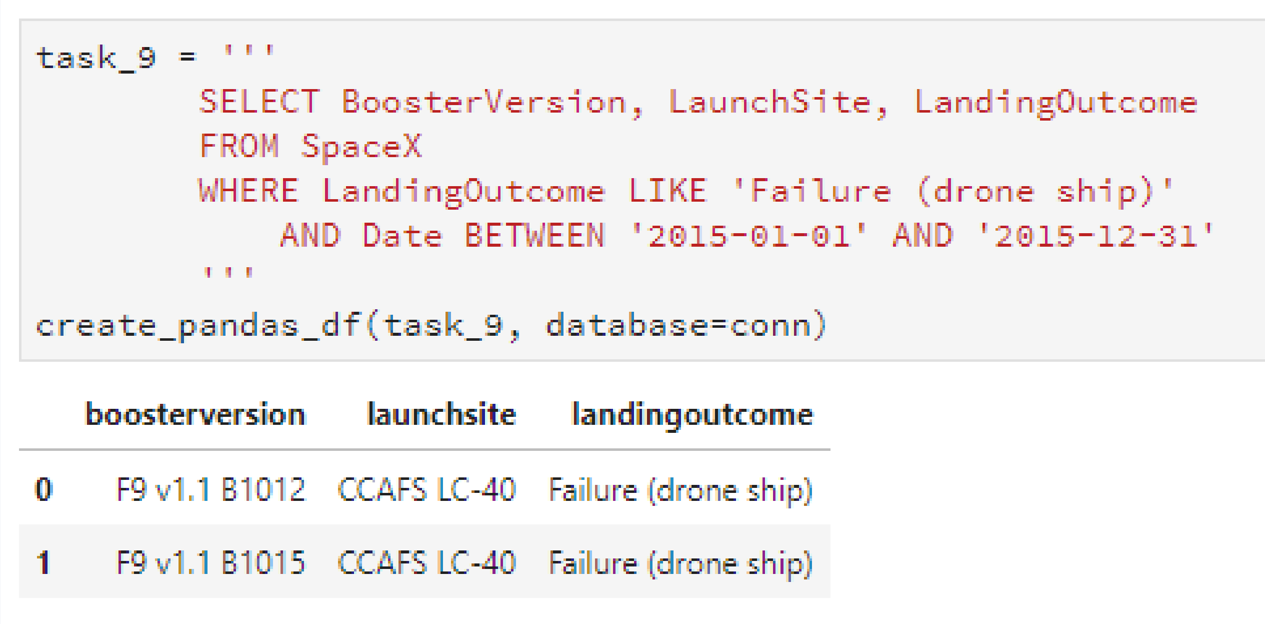


```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8,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

	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

- We can list failed landing outcome, booster version and launch site using LIKE clause to find ‘Failure (drone ship)’ and BETWEEN clause to find date between 1st january in 2015 and 31th december 2015.

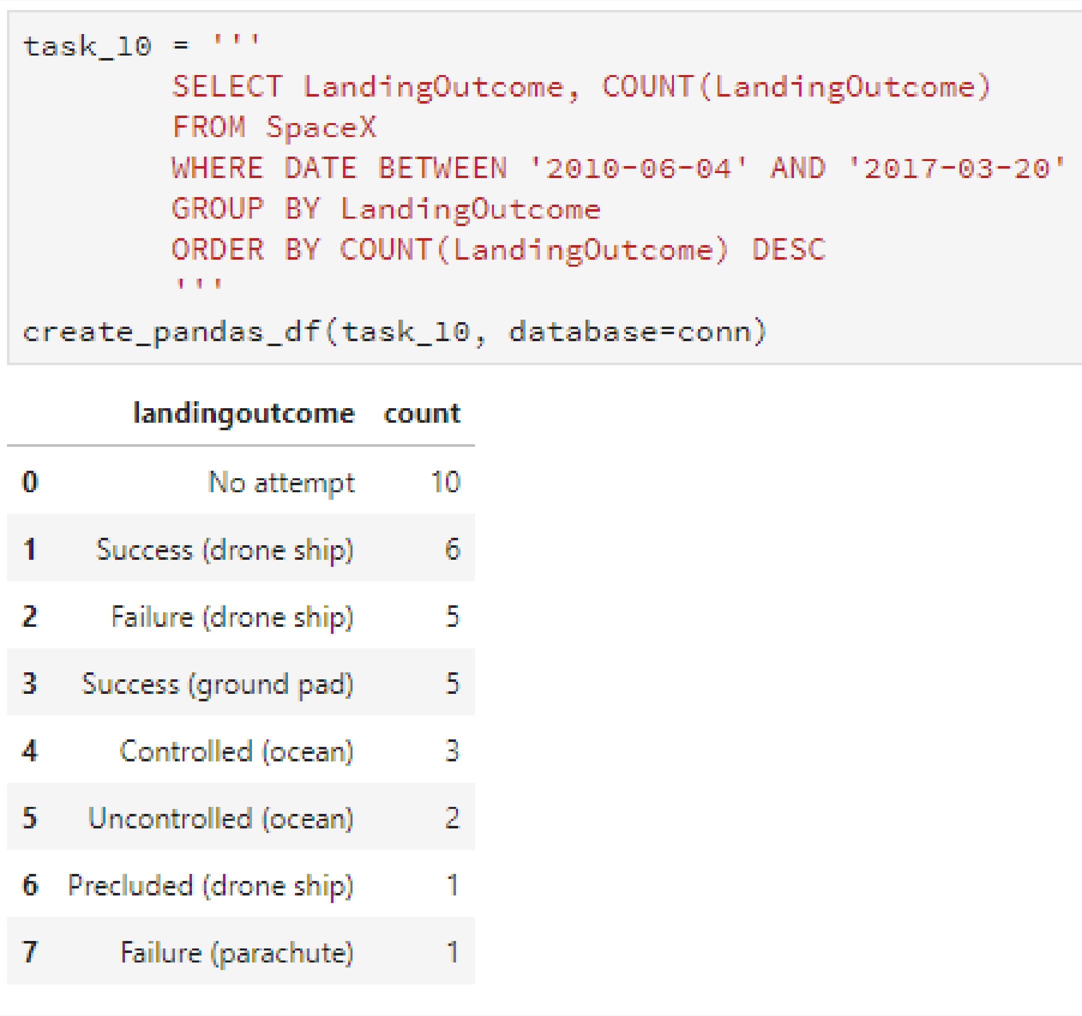


```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	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

- Finally we can rank the landing outcomes between 2010 and 2017 using COUNT syntax to count LandingOutcome column, BETWEEN clause to bound the date and year, GROUP BY syntax to group LandingOutcome column and ORDER BY syntax to order LandingOutcome descending.



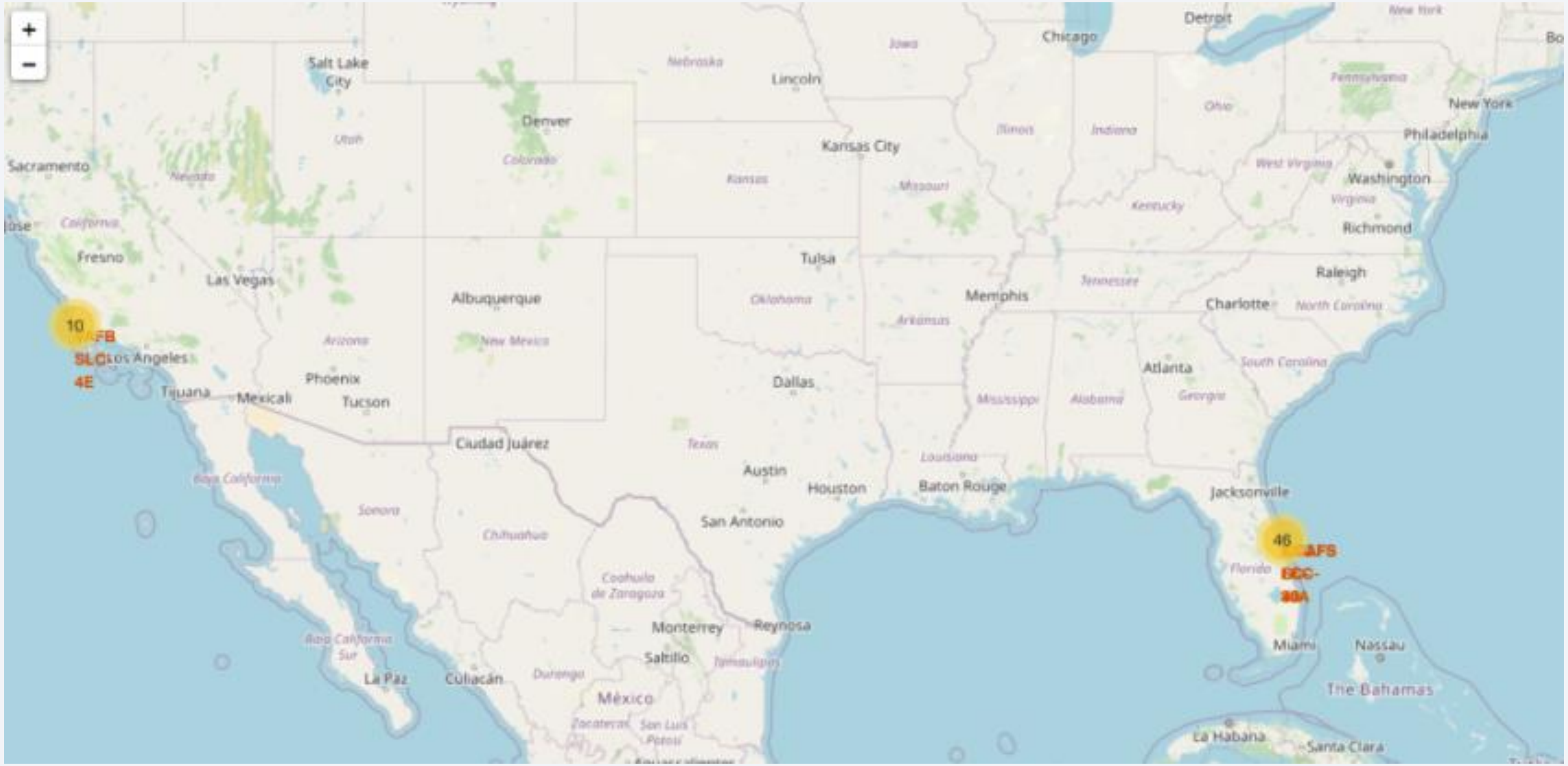
```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

Section 3

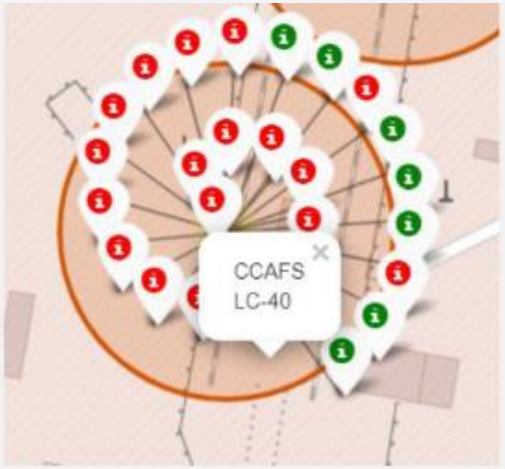
Launch Sites Proximities Analysis

Launch Sites Maps From Florida To California

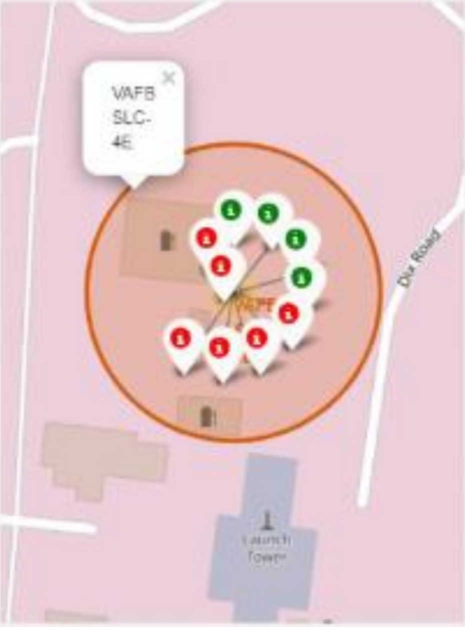


Marker Cluster To Show Launch Sites

- Florida launch sites and California launch sites are marked with green and red markers. Green markers show successful landing and red markers show failure landing.



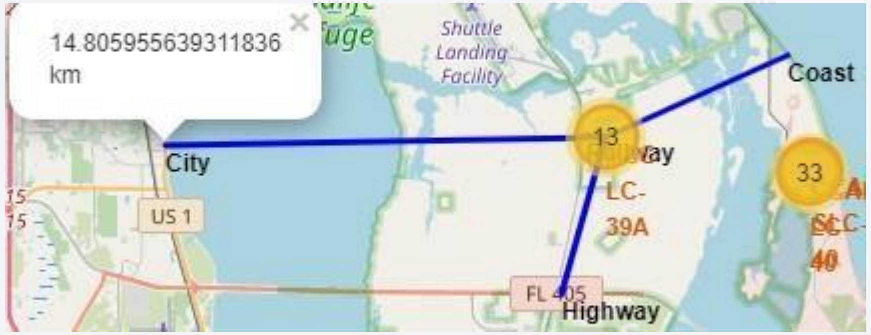
Florida launch sites

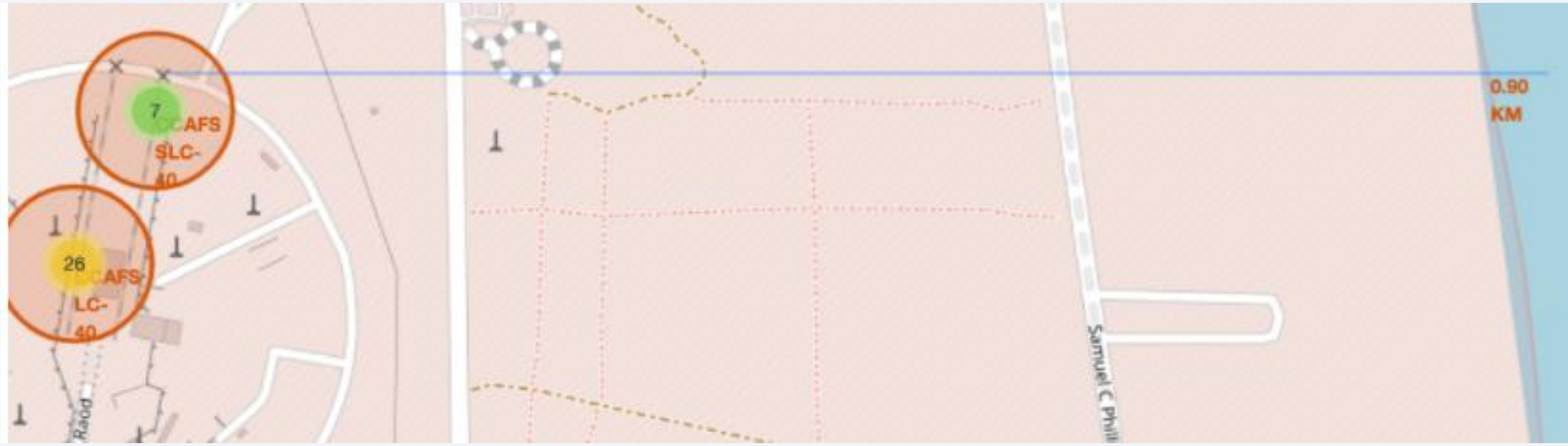


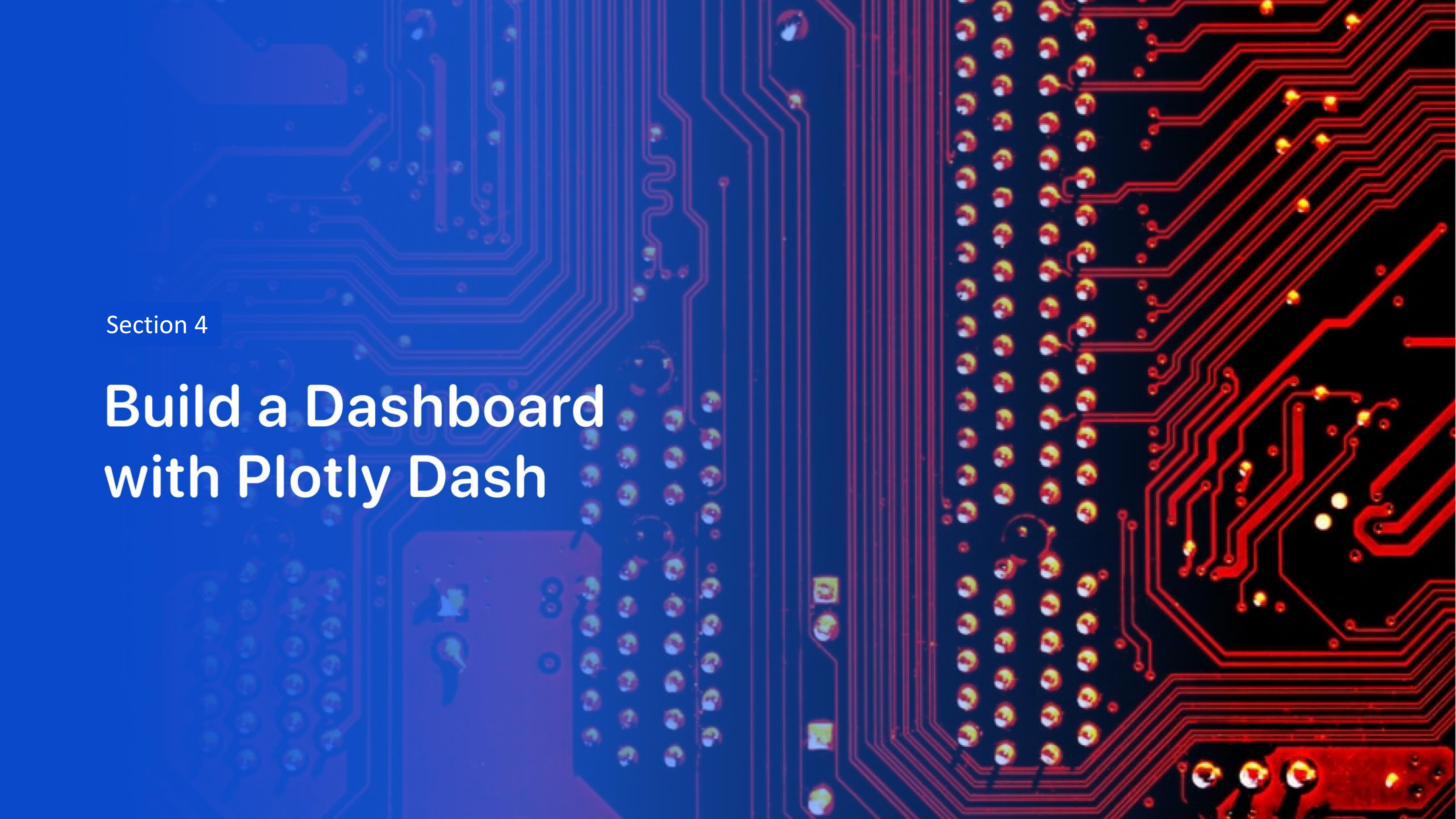
California launch sites

Distance To City Landmarks

We can measure launch distance from any city landmarks like railway, highway or coast line. Launch site distance are far from railways and highways but close to the coast line.





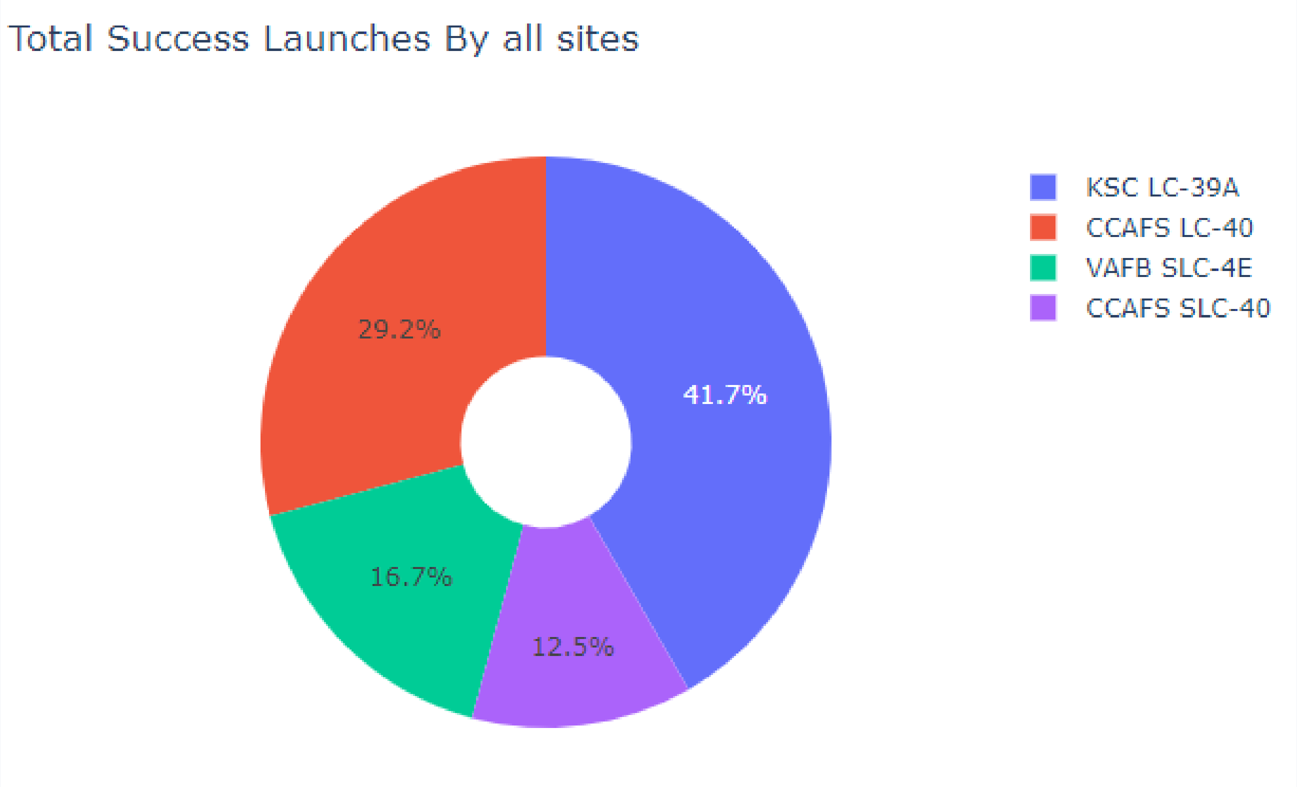


Section 4

Build a Dashboard with Plotly Dash

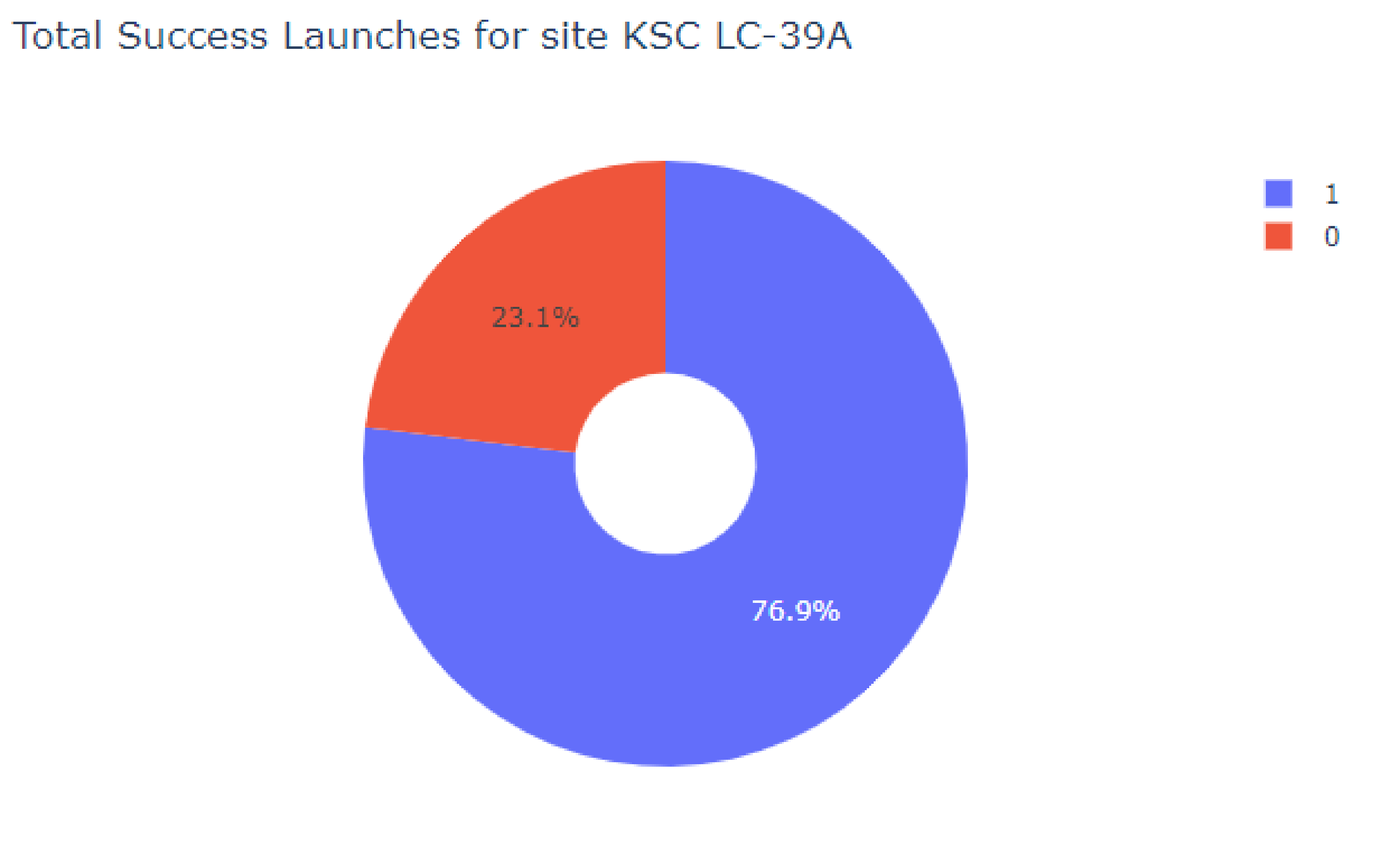
Success Launch By All Sites

- From pie chart below we can conclude that KSC LC-39A have a greater successful landing rate rather than other sites



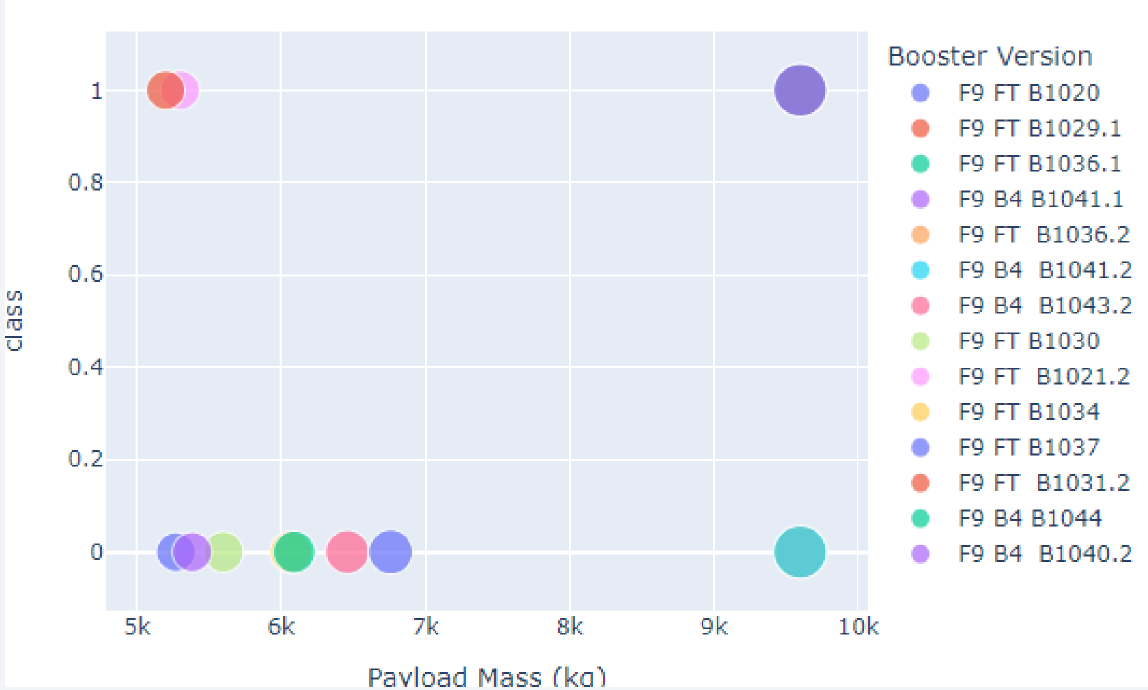
Total Success Launch For Site KSC LC-39A

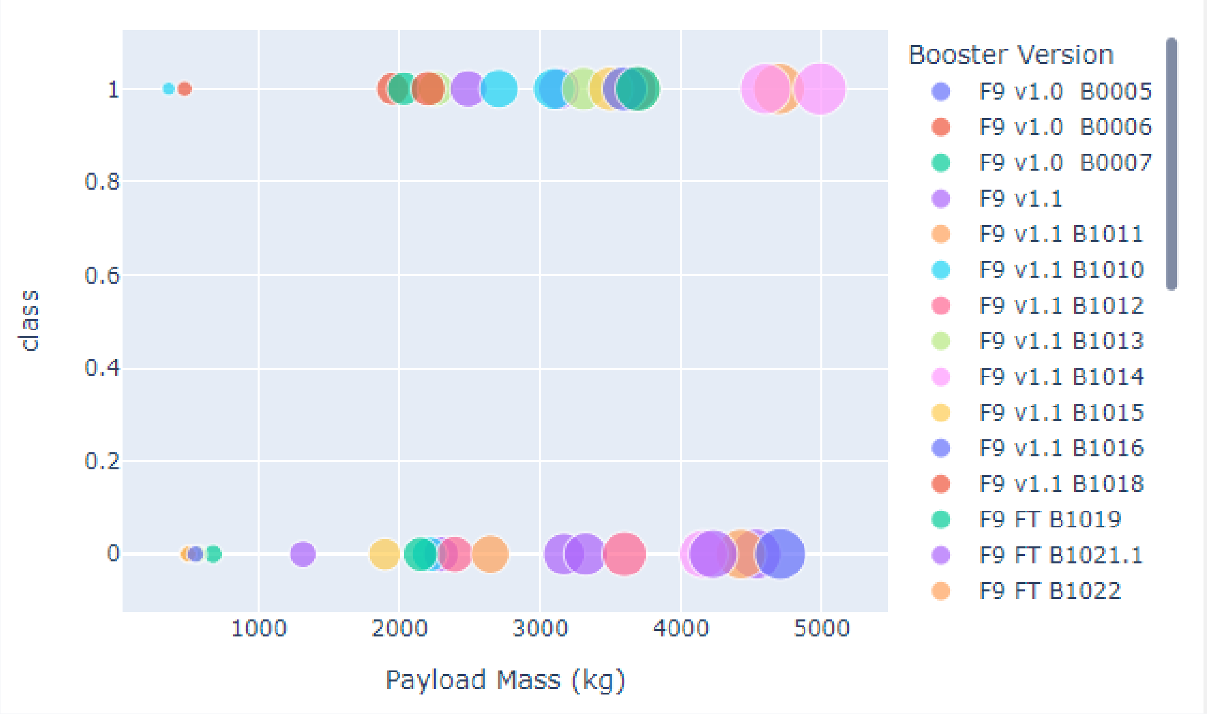
- There is 76.9 % successful rate and 23.1 % failure rate for KSC LC-39A site.



Payload Mass Vs Launch Outcome Plot

- There are higher success rate outcome in lighter payload below 5000 Kg and higher failure rate in heavier payload above 5000 Kg



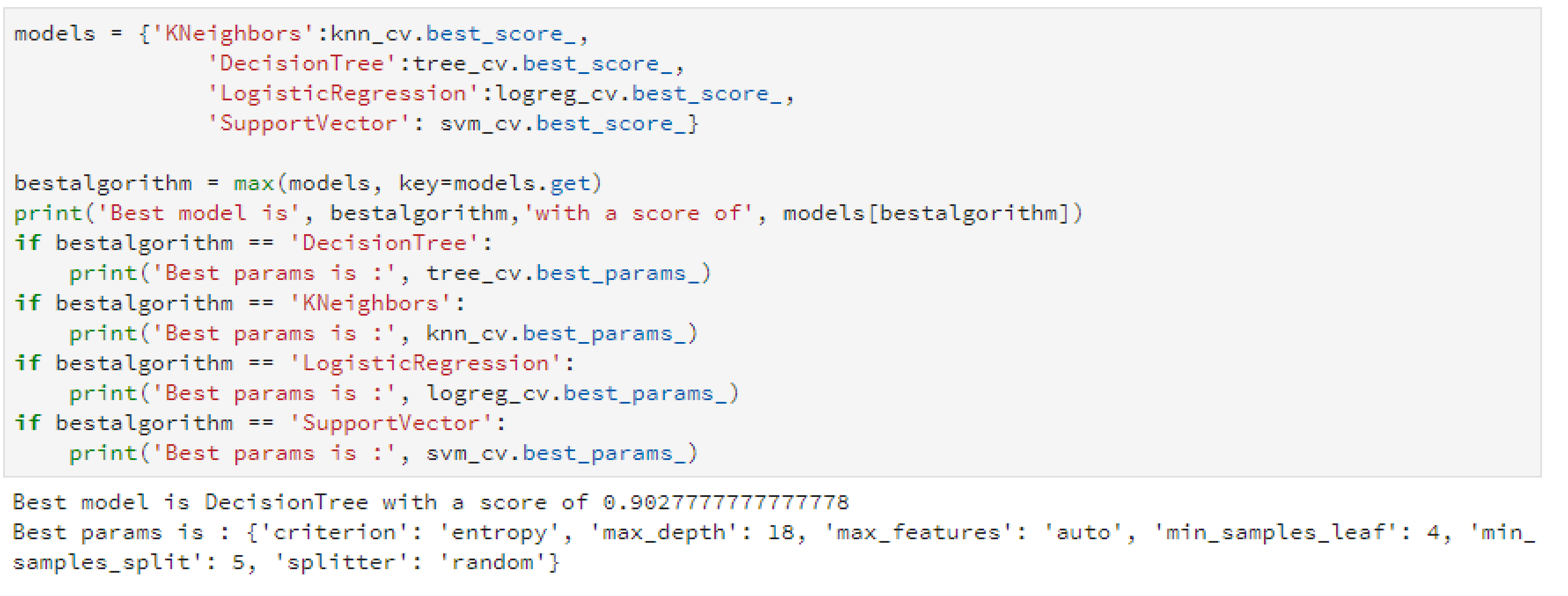


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Hgihest classification accuracy achieved by Decision Tree algorithm with score 0.903 using entropy criterion as parameter



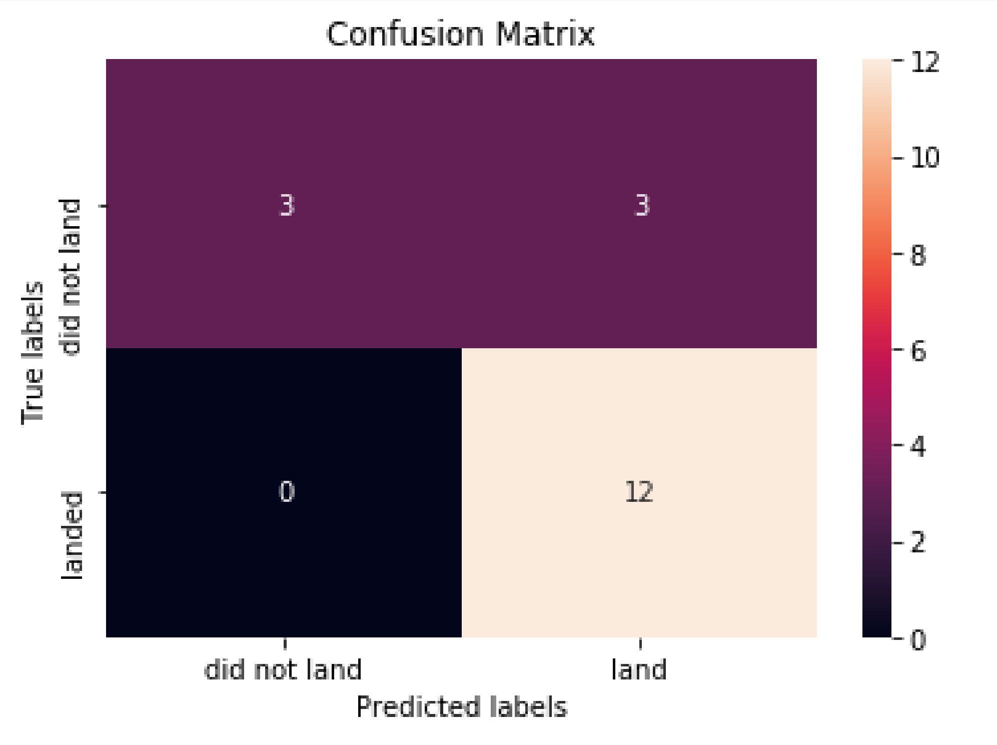
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.9027777777777778
Best params is : {'criterion': 'entropy', 'max_depth': 18, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_
samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- Confusion matrix of Decision Tree shows that classifier could distinguish different classes. Classifier can predicted all landing outcome successfully and only predict 3 didn't land.



Conclusions

- Launch success rate started to increase in 2013 till 2020.
- There is high successful landing on heavy payload in Low Earth Orbit and International Space Station Orbit.
- There are 100% successful landing on ES-L1 orbit, Geosynchronous Equatorial Orbit, Highly Elliptical Orbit and Sun Synchronous Orbit.
- Hgihest classification accuracy achieved by Decision Tree algorithm with score 0.903 using entropy criterion as parameter

Thank you!