SpaceX Falcon9 first stage Landing Prediction

Erkut Koral

05.12.2022

Reference:
IBM Data Science Professional Certification curriculum, hosted on Coursera.com and CognitiveClass.ai

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion

EXECUTIVE SUMMARY

- Data Collection
  - Data collection with API
  - Data collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis
  - EDA with SQL
  - EDA with Data Visualization
- Interactive Visual Analytics with Folium
- Features Engineering
- Machline Learning Predictions
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive Analytics with screenshots
  - Prediction results

INTRODUCTION

- Project Background
  - 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 because SpaceX 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 SpaceX for a rocket launch.

IBM Developer

SKILLS NETWORK

INTRODUCTION

- Questions we want to answer:

1. Will Falcon9’s first stage be land successfully?
2. What factors affect this landing?

For reaching the notebooks, check out the github icon below of the pages.

IBM Developer SKILLS NETWORK

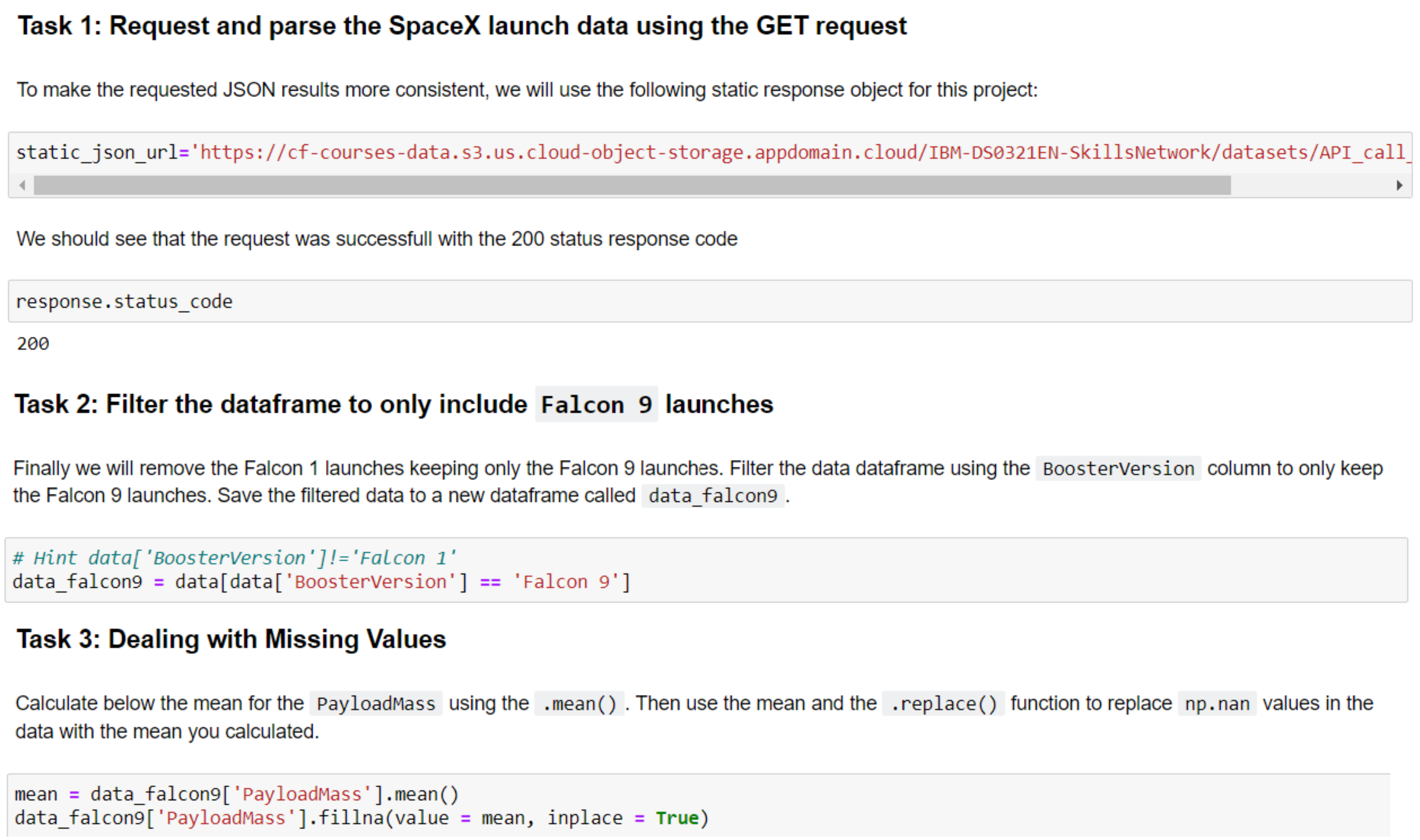
Data Collection

METHODOLOGY
1)Data Collection

- The data was collected in different methods:

1. First, we created requests.get() object and took the reponse from API.
2. Then, we decoded that response and used .json_normalize() to turning it to pandas DataFrame.
3. After that, we took our features from that DataFrame.
4. Finally, we cleaned the data, checked for missing values and filled them with mean value of that features.
5. In addiction, we collected the data with Web Scraping method, using BeautifulSoup.

METHODOLOGY
1)Data Collection with API



Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_
```

We should see that the request was successfull with the 200 status response code

```
response.status_code
```

```
200
```

Task 2: Filter the dataframe to only include `Falcon 9` launches

Finally we will remove the Falcon 1 launches keeping only the Falcon 9 launches. Filter the data dataframe using the `BoosterVersion` column to only keep the Falcon 9 launches. Save the filtered data to a new dataframe called `data_falcon9`.

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = data[data['BoosterVersion'] == 'Falcon 9']
```

Task 3: Dealing with Missing Values

Calculate below the mean for the `PayloadMass` using the `.mean()`. Then use the mean and the `.replace()` function to replace `np.nan` values in the data with the mean you calculated.

```
mean = data_falcon9['PayloadMass'].mean()
data_falcon9['PayloadMass'].fillna(value = mean, inplace = True)
```

IBM Developer

SKILLS NETWORK

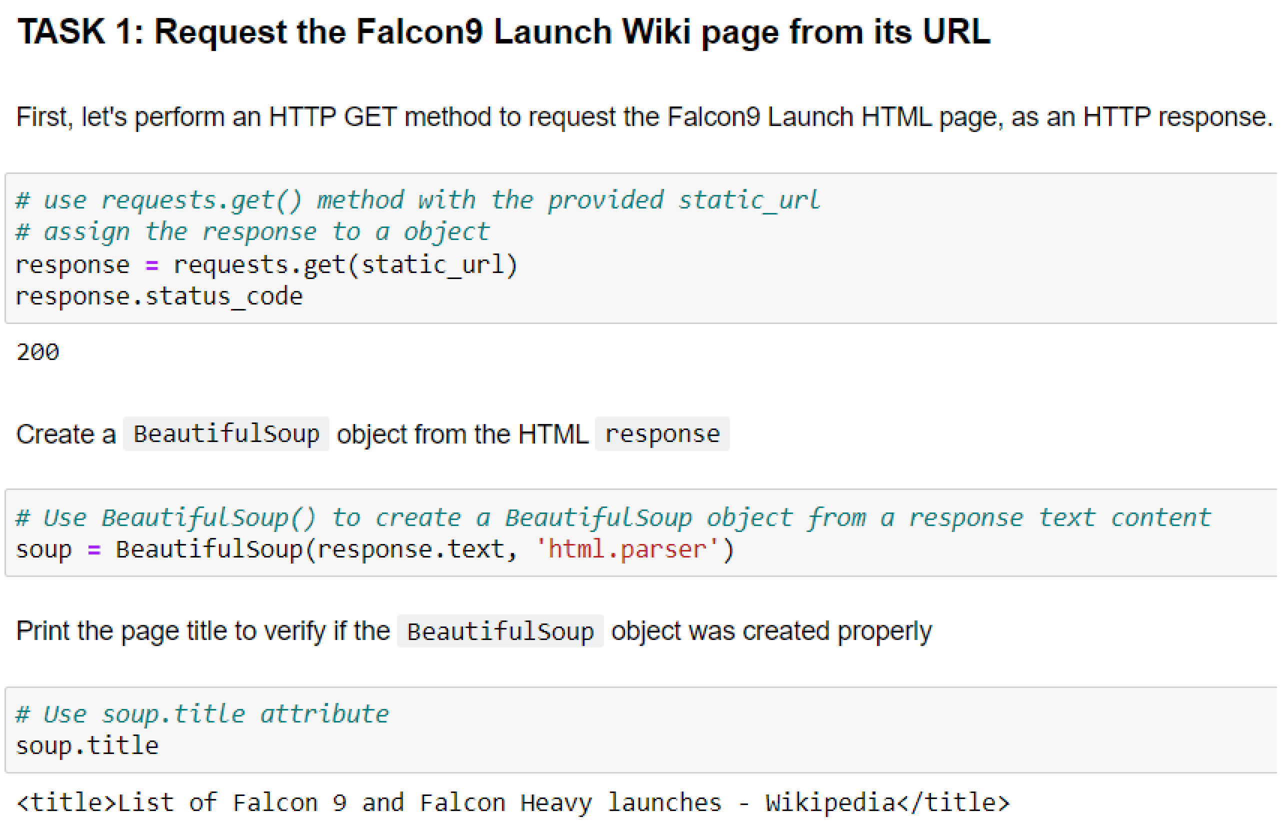
METHODOLOGY
1)Data Collection with Web Scraping

We started with Task1.

1. We created column names object with its function for taking column names from Wikipedia.

2. Created an empty dictionary for taking which features we want to take.

3. Put that features in empty dictionary and created a dataframe from it.



TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
response.status_code
```

```
200
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.text, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
# Use soup.title attribute
soup.title
```

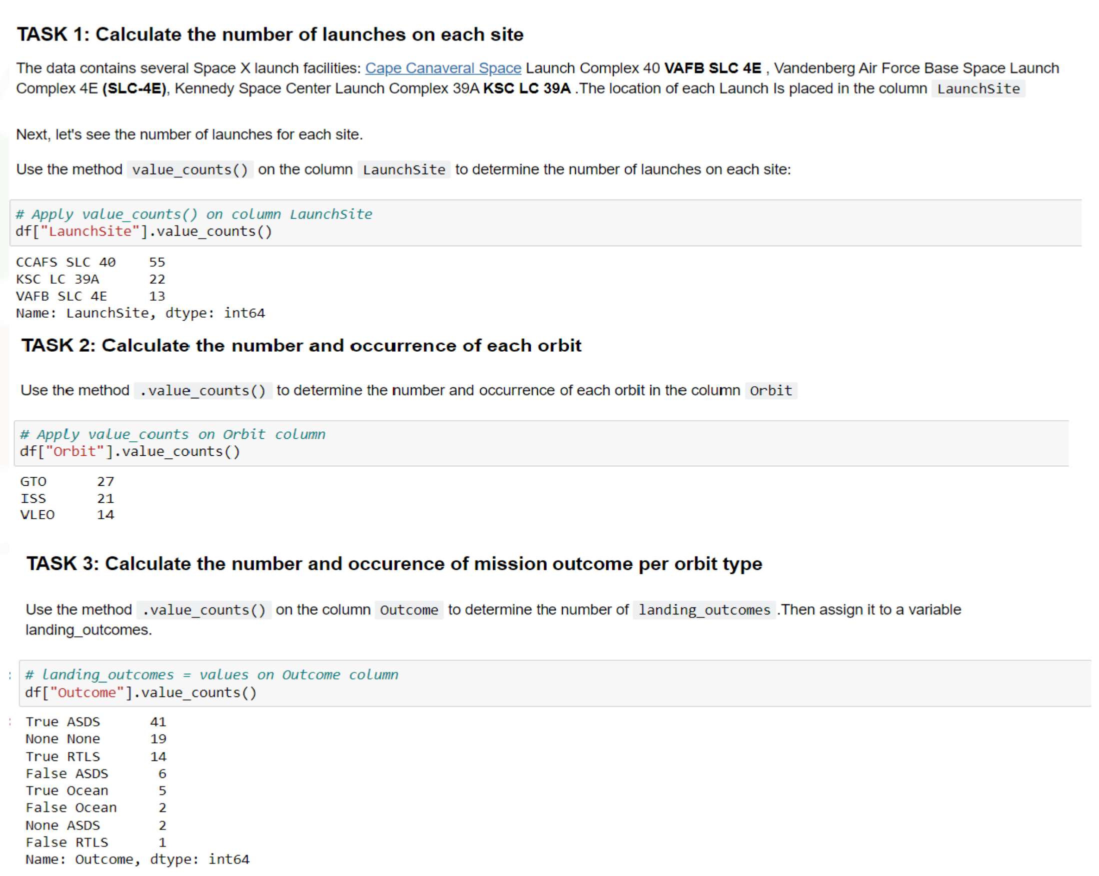
```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

Data Wrangling

METHODOLOGY
2)Data Wrangling

After, all tasks were done, for taking outcome to the class column we need to parse them like:

1. **Bad outcomes (did not land)(0)**
2. **Other outcomes (land) (1)**



TASK 1: Calculate the number of launches on each site

The data contains several Space X launch facilities: Cape Canaveral Space Launch Complex 40 **VAFB SLC 4E** , Vandenberg Air Force Base Space Launch Complex 4E **(SLC-4E)**, Kennedy Space Center Launch Complex 39A **KSC LC 39A** .The location of each Launch Is placed in the column `LaunchSite`

Next, let's see the number of launches for each site.

Use the method `value_counts()` on the column `LaunchSite` to determine the number of launches on each site:

```
# Apply value_counts() on column LaunchSite
df["LaunchSite"].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

TASK 2: Calculate the number and occurrence of each orbit

Use the method `.value_counts()` to determine the number and occurrence of each orbit in the column `Orbit`

```
# Apply value_counts on Orbit column
df["Orbit"].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
```

TASK 3: Calculate the number and occurence of mission outcome per orbit type

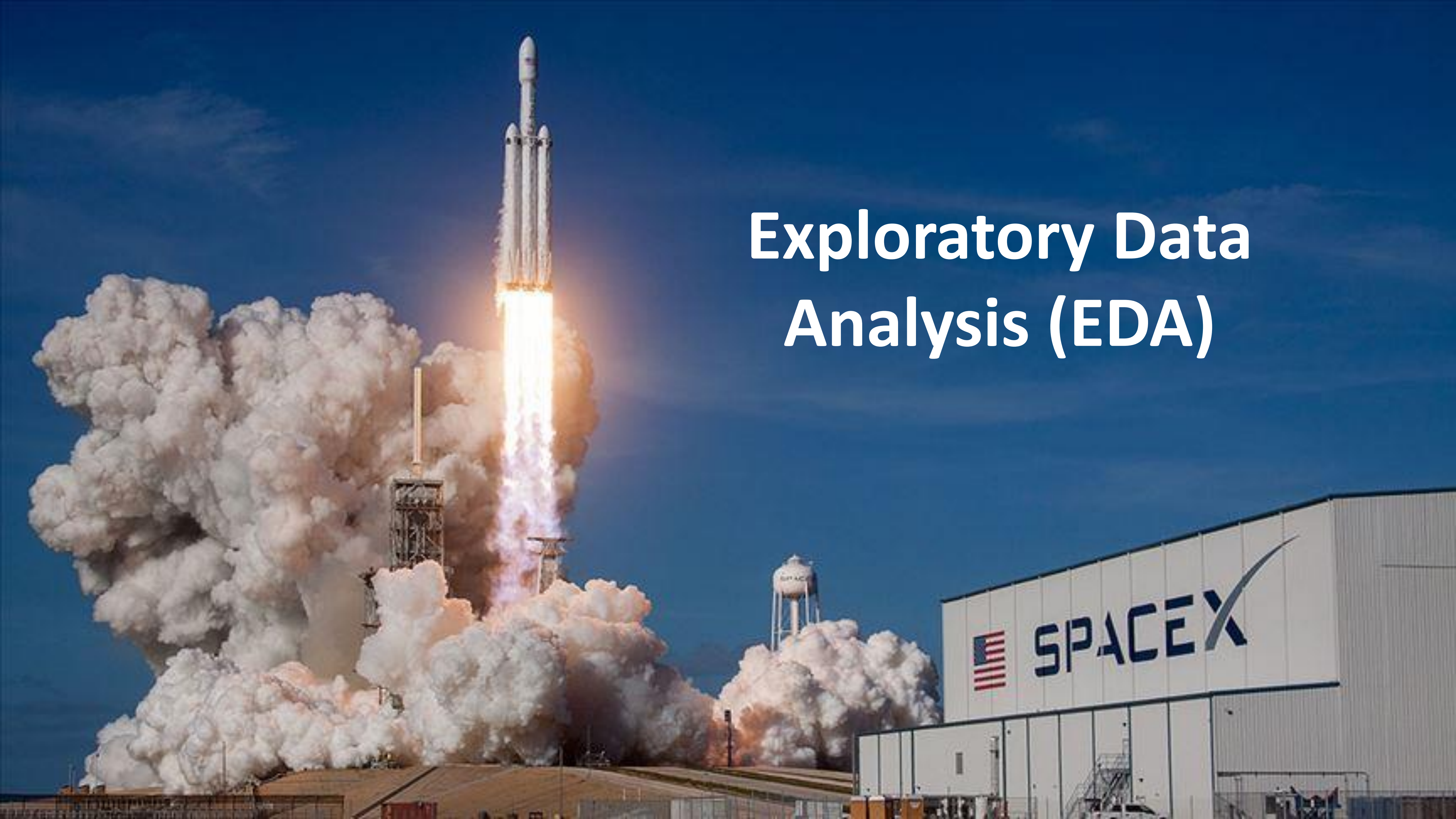
Use the method `.value_counts()` on the column `Outcome` to determine the number of `landing_outcomes` .Then assign it to a variable landing_outcomes.

```
# landing_outcomes = values on Outcome column
df["Outcome"].value_counts()
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

IBM Developer

SKILLS NETWORK



Exploratory Data Analysis (EDA)

METHODOLOGY
3)EDA with SQL

- We explored data with some questions and applying that questions to SQL.
- These questions are:

1. *Display the names of the unique launch sites in the space mission*
2. *Display 5 records where launch sites begin with the string 'CCA'*
3. *Display the total payload mass carried by boosters launched by NASA (CRS)*
4. *Display average payload mass carried by booster version F9 v1.1*
5. List the date when the first successful landing outcome in ground pad was acheived.
6. *List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*
7. *List the total number of successful and failure mission outcomes*
8. *List the names of the booster_versions which have carried the maximum payload mass. Use a subquery*
9. *List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015*
10. *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*

METHODOLOGY
3)EDA with Data Visualization

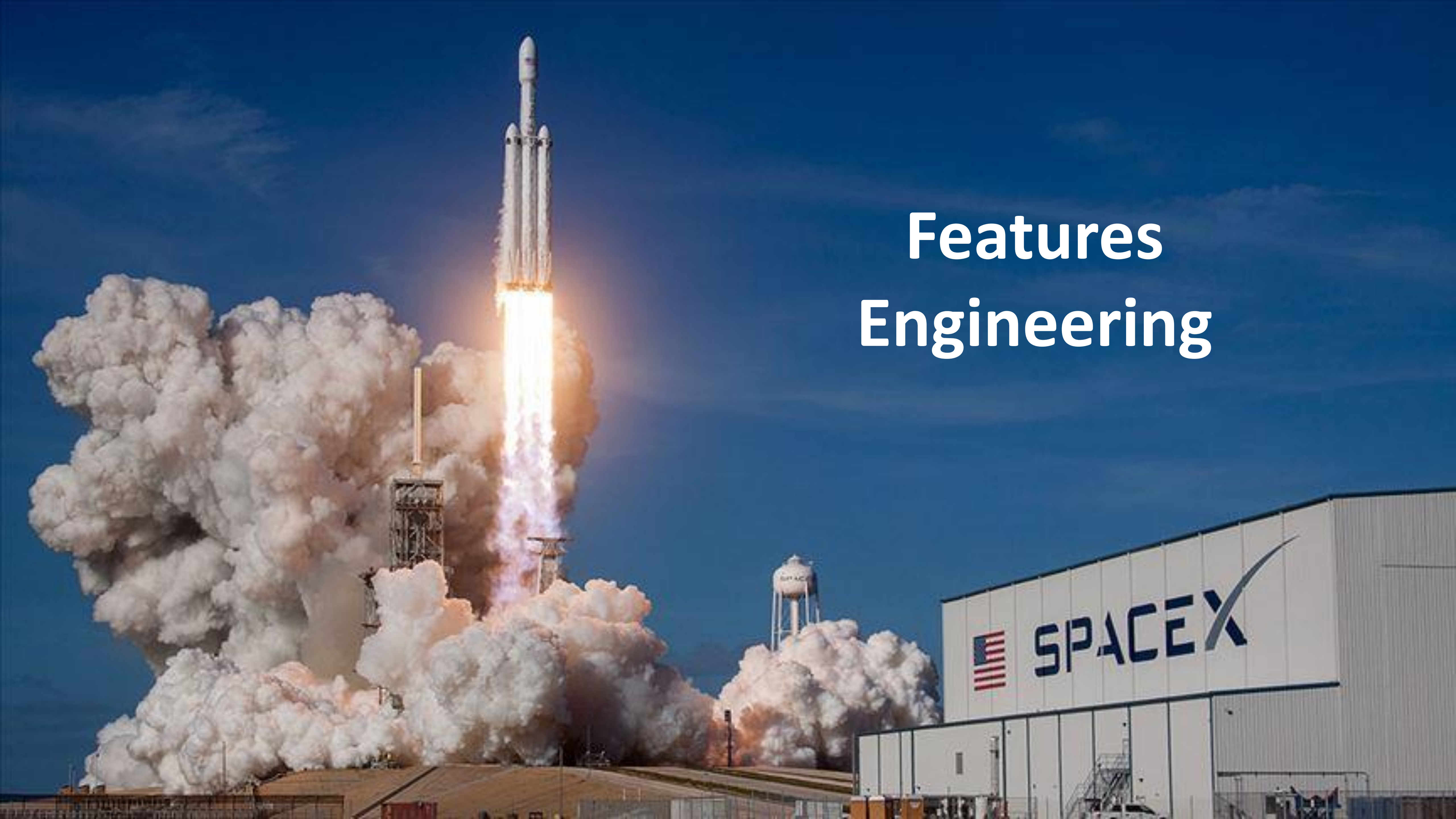
- We explored data with data visualization and tried to answer these following questions:

1. Visualize the relationship between Flight Number and Launch Site
2. Visualize the relationship between Payload and Launch Site
3. Visualize the relationship between success rate of each orbit type
4. Visualize the relationship between Flight Number and Orbit type
5. Visualize the relationship between Payload and Orbit type
6. Visualize the launch success yearly trend

METHODOLOGY
4)Interactive Visual Analytics with Folium

- The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.
- **This section contains the following tasks:**

1. Mark all launch sites on a map
2. Mark the success/failed launches for each site on the map
3. Calculate the distances between a launch site to its proximities



Features Engineering

METHODOLOGY
5)Features Engineering

- When predicting the landing outcome, we need to take the features that will serve this purpose.
- So, we dropped Date, BoosterVersion, Outcome, Longitude and Latitude features.
- Used get_dummies() function to apply One Hot Encoding to categorical columns.
- After that, we need to transform the dataframe type to float64.

Machine Learning Predictions

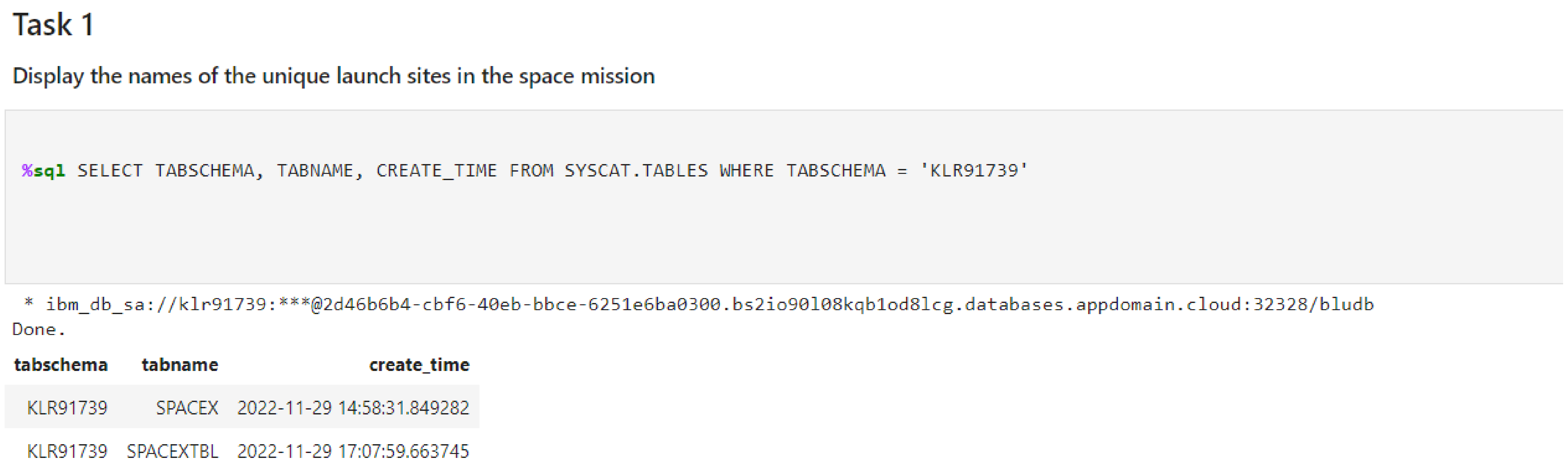
METHODOLOGY
5)Machine Learning Predictions

- In this section we performed following objectives:

1. **Standardized the data**
2. **Splitted into training data and test data**
3. **Determined Machine Learning Algorithms for predictions**
   Logistic Regression, SVM, Decision Trees, KNearestNeighbors
4. **Finded the best Hyperparameters for selected algorithms**
5. **Finded the method performs best using test data**

Insights from EDA

RESULTS for EDA with SQL



Task 1

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

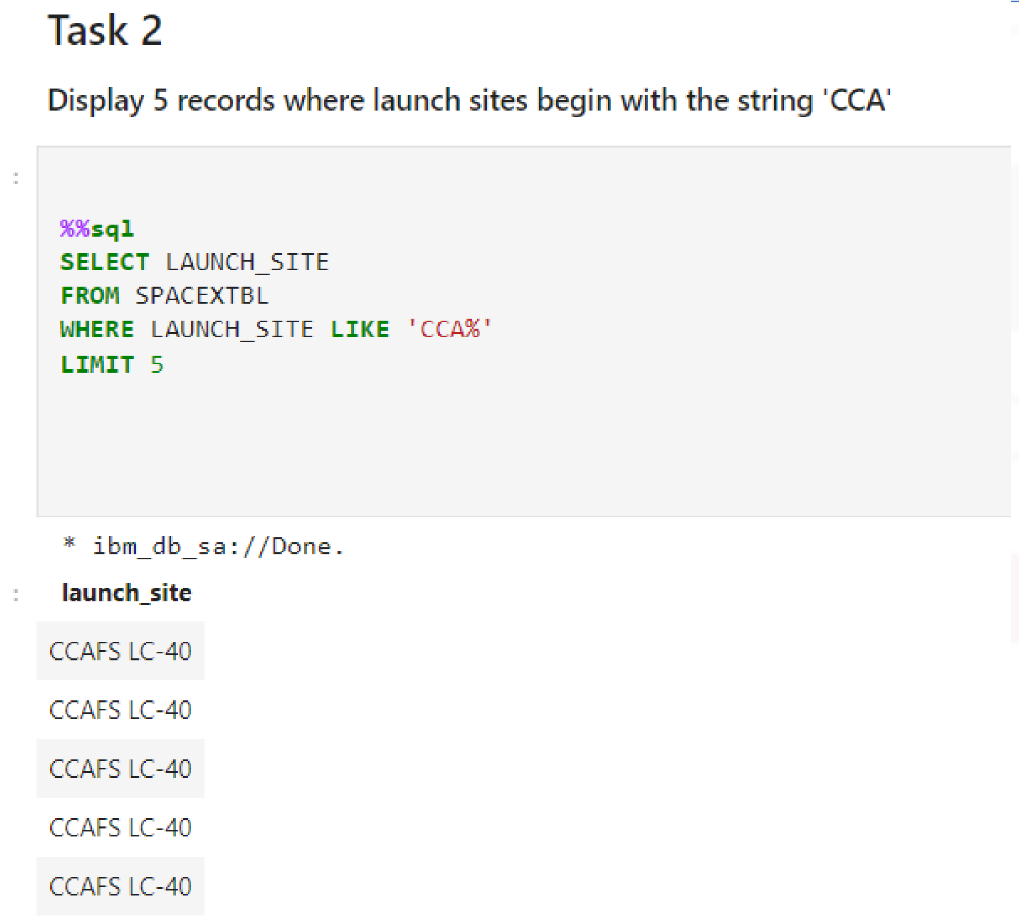
```
%sql SELECT TABSCHEMA, TABNAME, CREATE_TIME FROM SYSCAT.TABLES WHERE TABSCHEMA = 'KLR91739'
```

```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

tabschema	tabname	create_time
KLR91739	SPACEX	2022-11-29 14:58:31.849282
KLR91739	SPACEXTBL	2022-11-29 17:07:59.663745

IBM Developer SKILLS NETWORK

RESULTS for EDA with SQL



Task 2

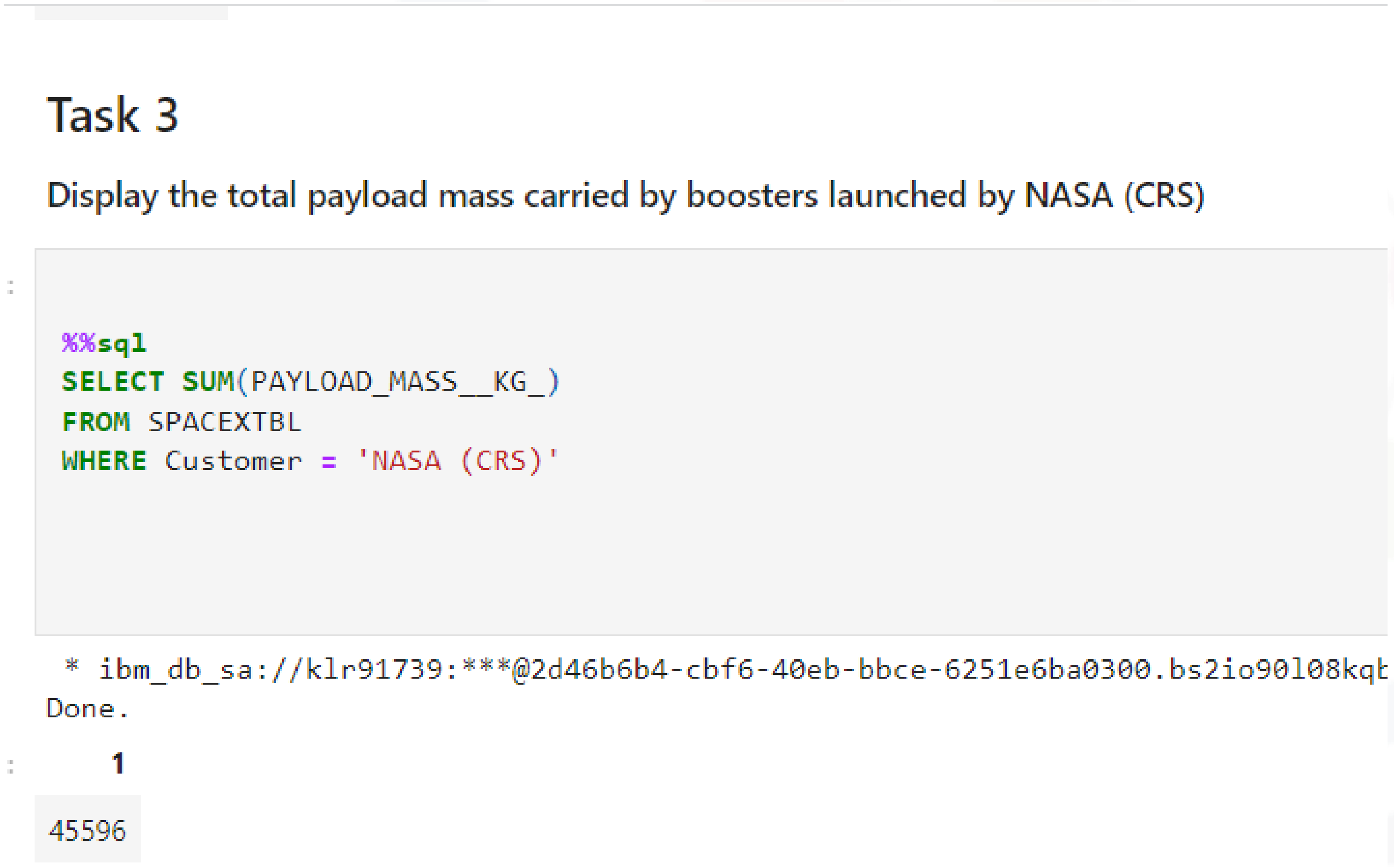
Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

```
 * ibm_db_sa://Done.
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

RESULTS for EDA with SQL



Task 3

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)'
```

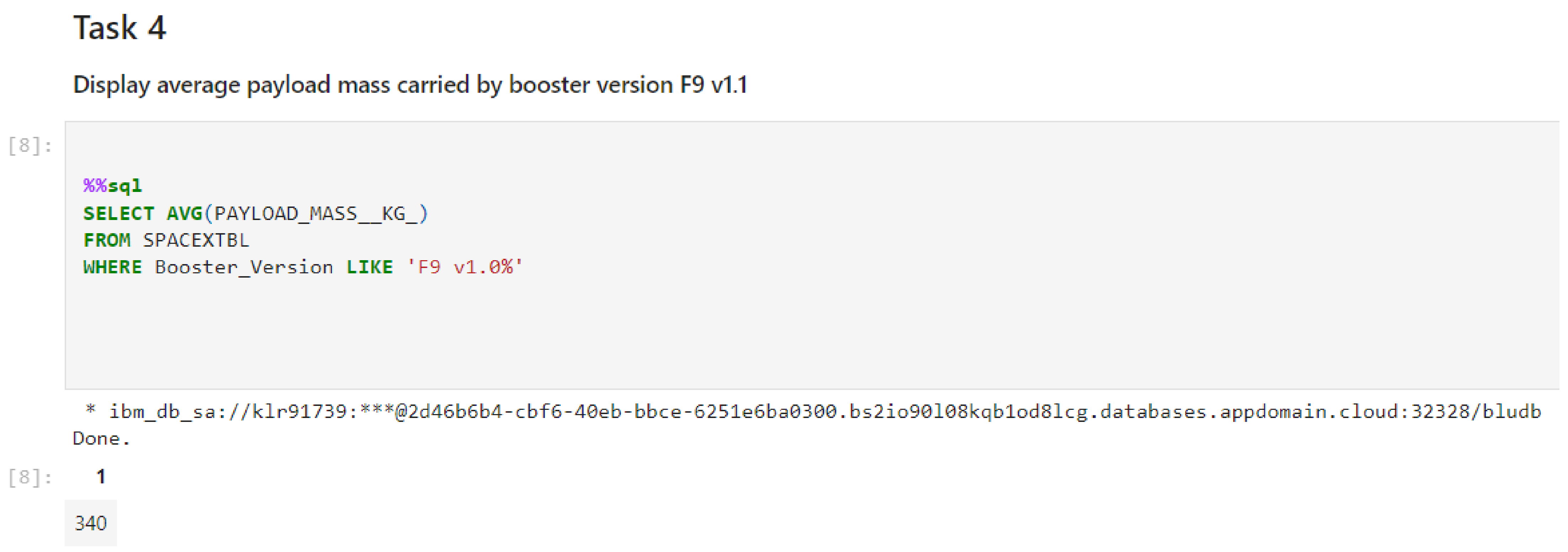
```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb
Done.
```

1
45596

IBM Developer

SKILLS NETWORK

RESULTS for EDA with SQL



Task 4

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

```
[8]:
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%'
```

```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

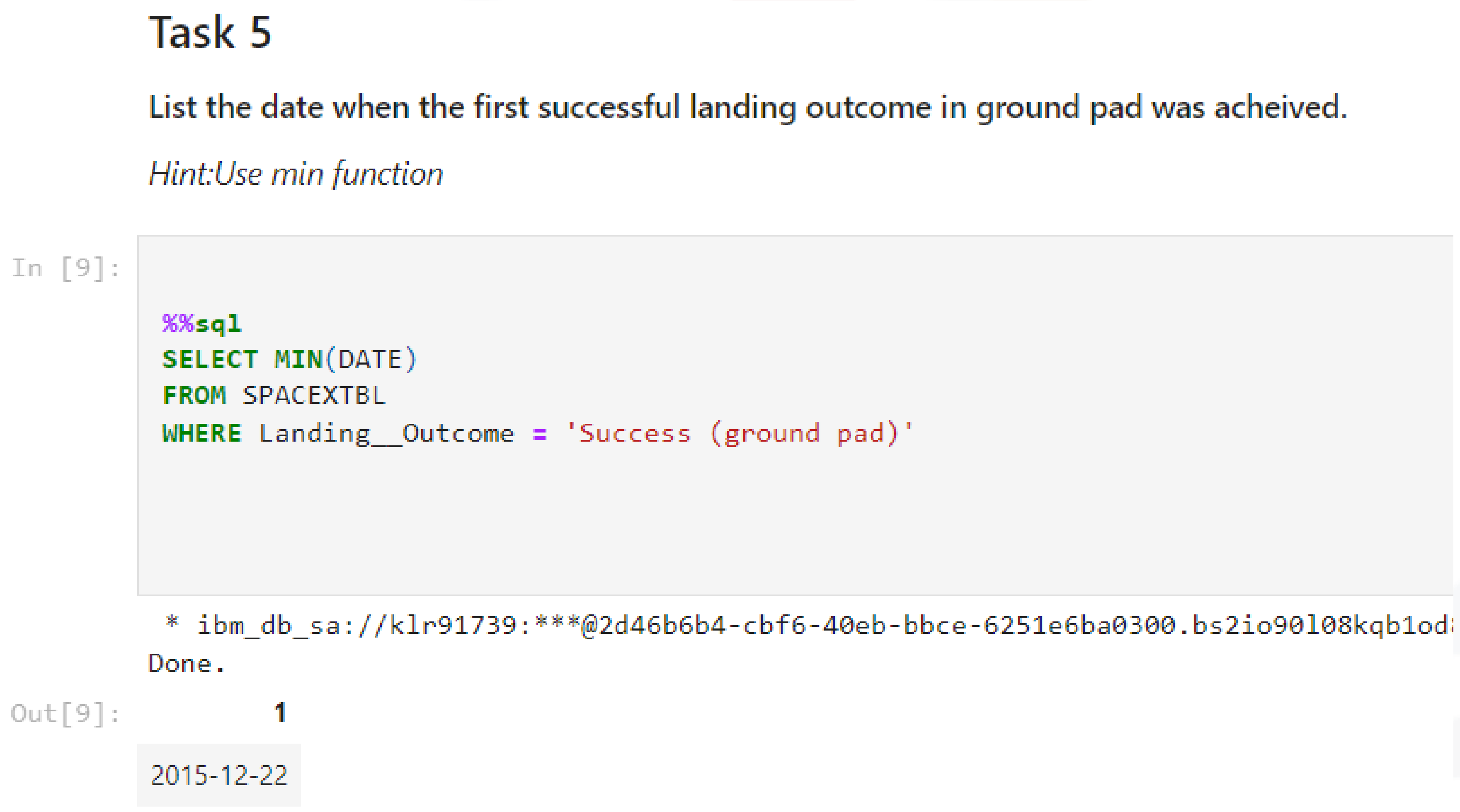
[8]:

1
340

IBM Developer

SKILLS NETWORK

RESULTS for EDA with SQL



Task 5

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

Hint:Use min function

```
In [9]:

%%sql
SELECT MIN(DATE)
FROM SPACEXTBL
WHERE Landing__Outcome = 'Success (ground pad)'
```

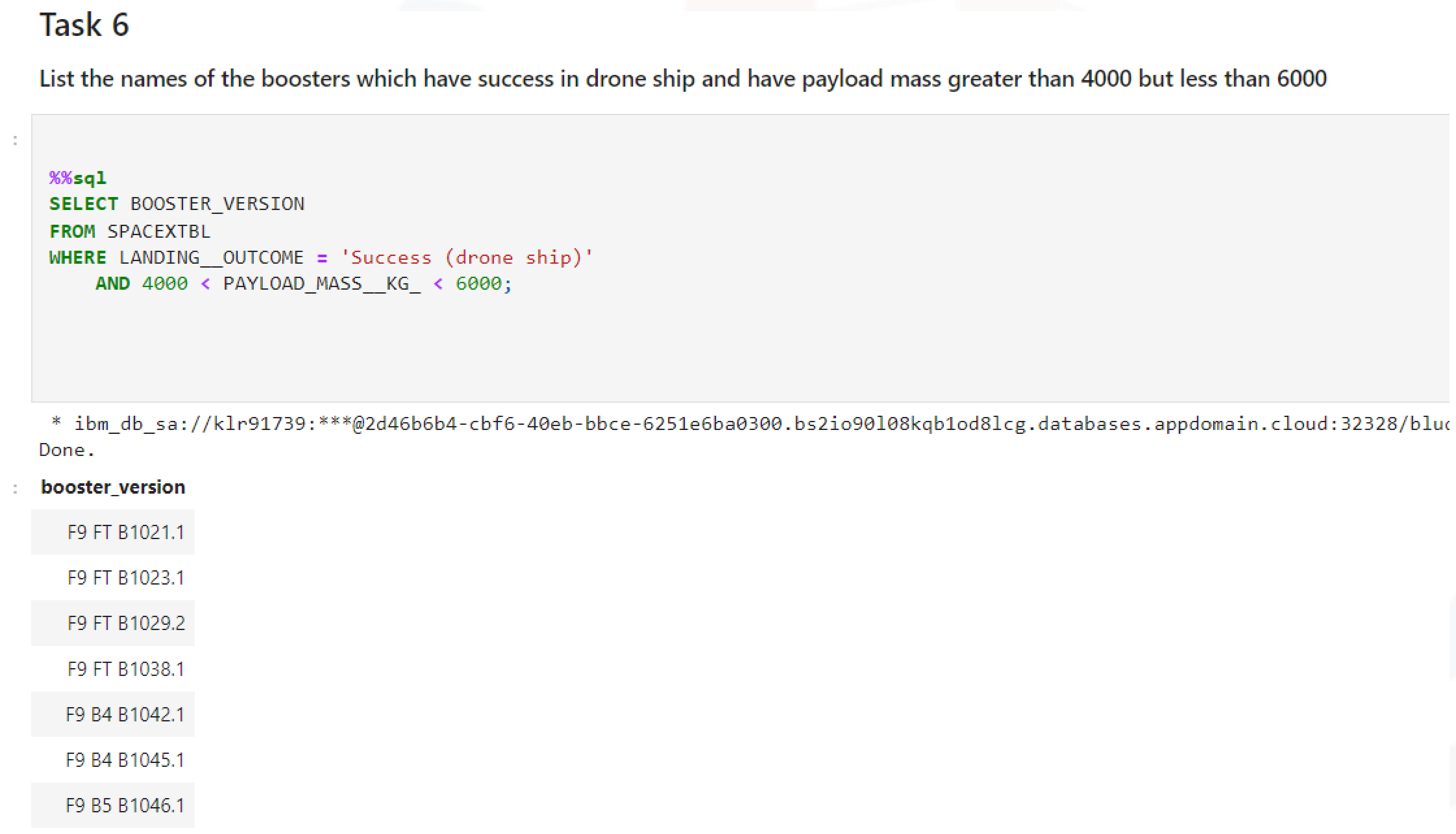
```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od
Done.
```

```
Out[9]:          1
        2015-12-22
```

IBM Developer

SKILLS NETWORK

RESULTS for EDA with SQL



Task 6

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

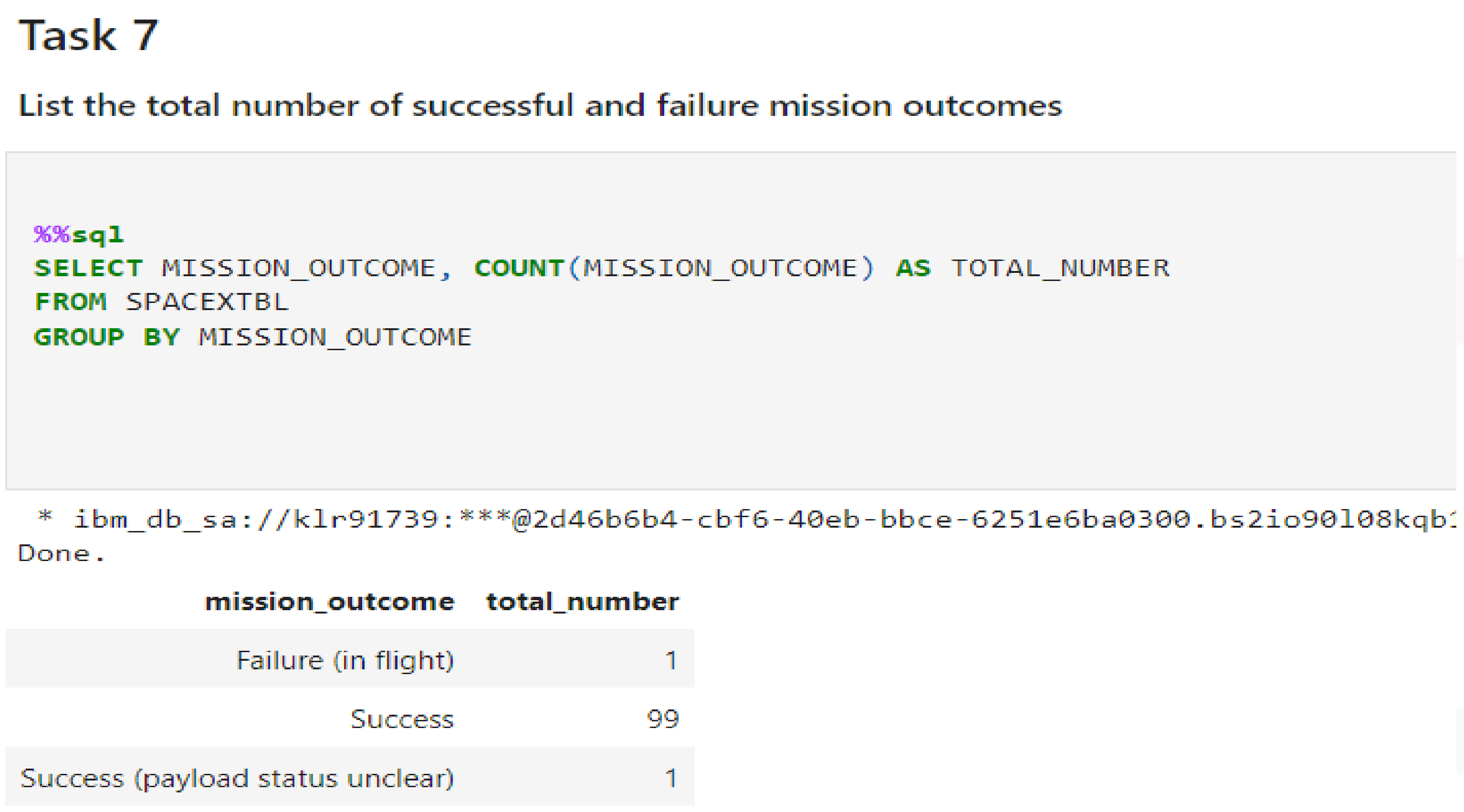
```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/blu
Done.
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

IBM Developer SKILLS NETWORK

RESULTS for EDA with SQL



Task 7

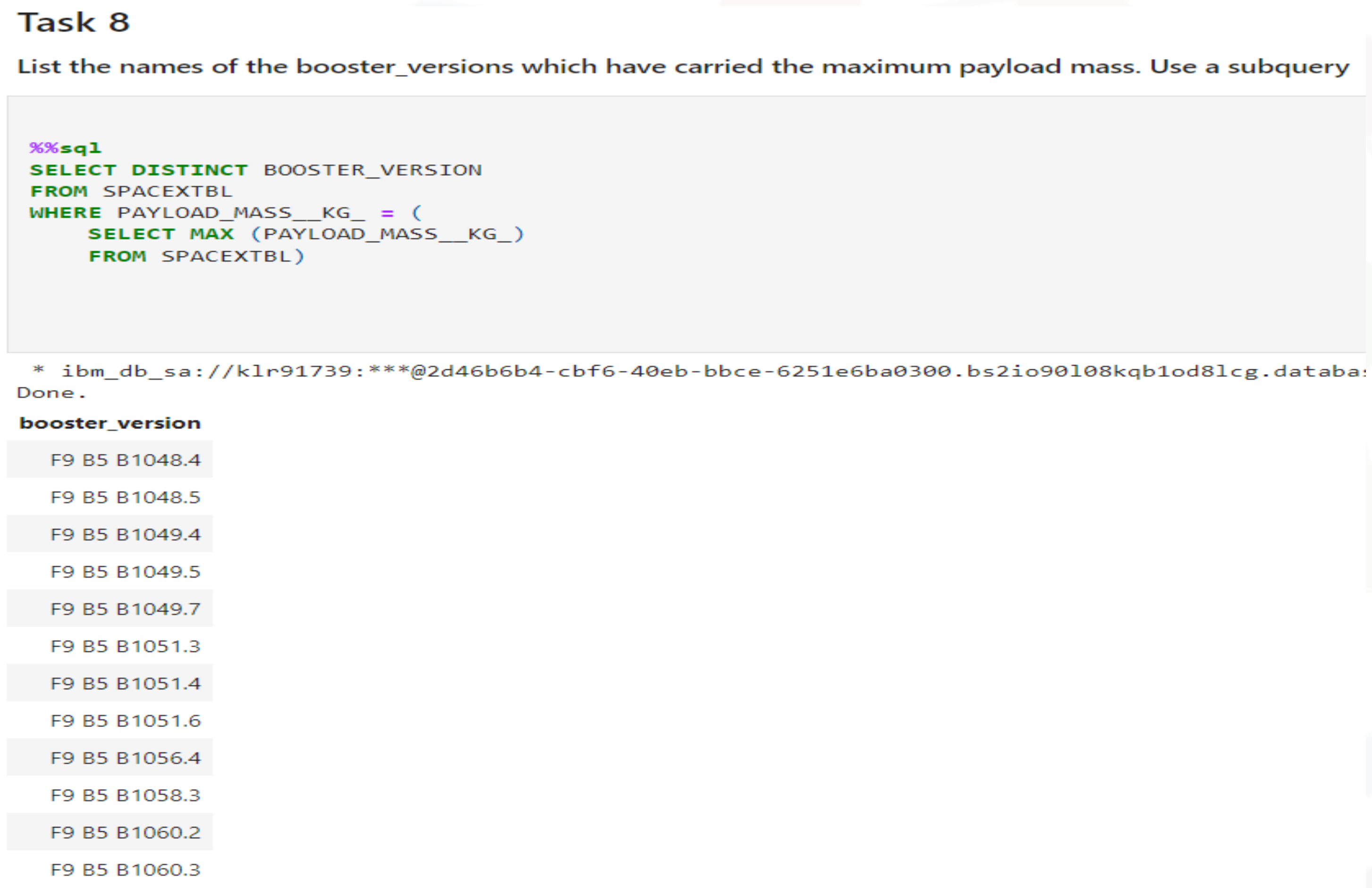
List the total number of successful and failure mission outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1
Done.
```

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

RESULTS for EDA with SQL



Task 8

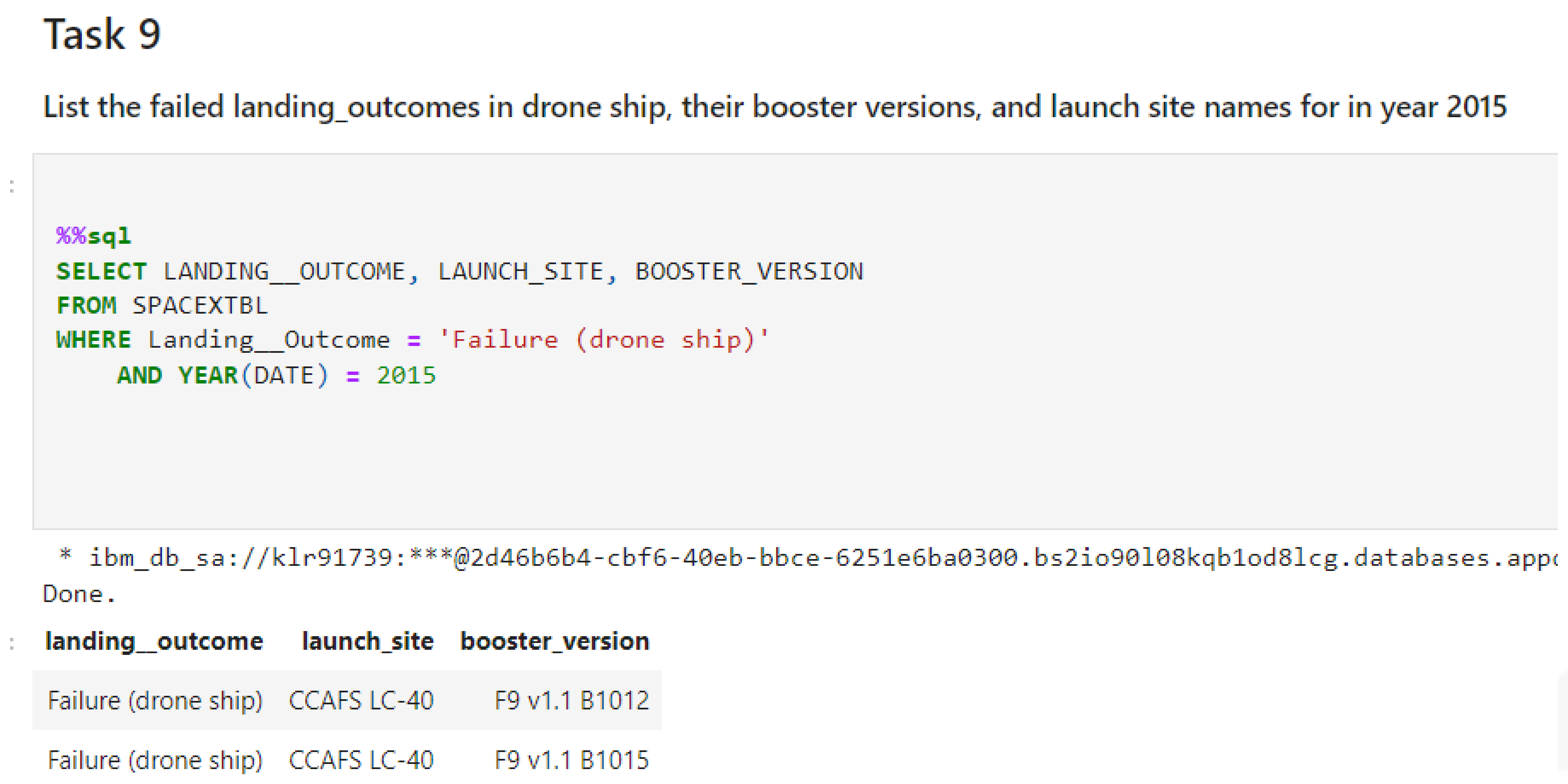
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX (PAYLOAD_MASS__KG_)
    FROM SPACEXTBL)
```

```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databa
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

RESULTS for EDA with SQL



Task 9

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

```
%%sql
SELECT LANDING__OUTCOME, LAUNCH_SITE, BOOSTER_VERSION
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015
```

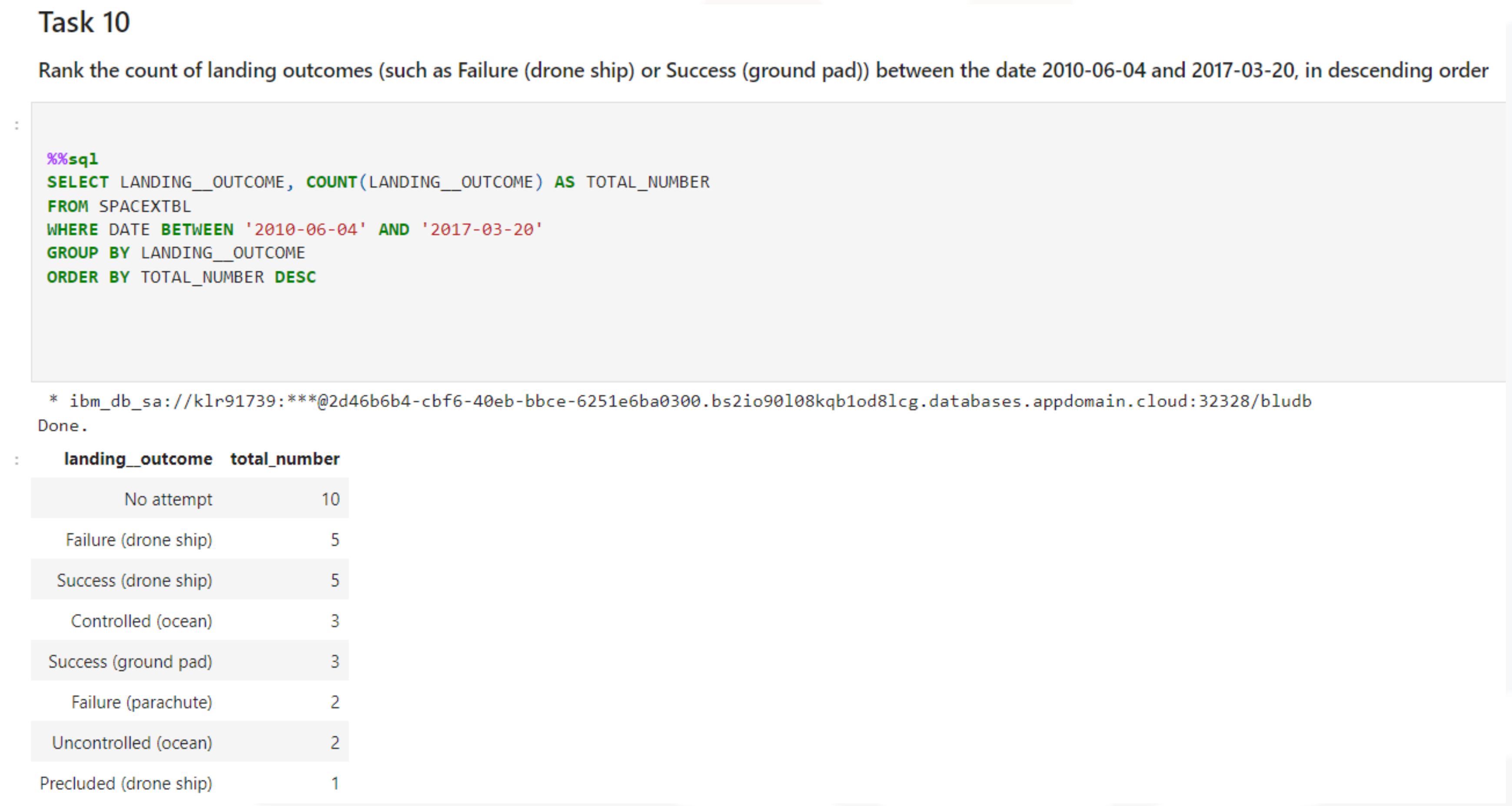
```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appd
Done.
```

landing__outcome	launch_site	booster_version
Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1012
Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1015

IBM Developer

SKILLS NETWORK

RESULTS for EDA with SQL



Task 10

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

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

```
 * ibm_db_sa://klr91739:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

landing__outcome	total_number
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

IBM Developer

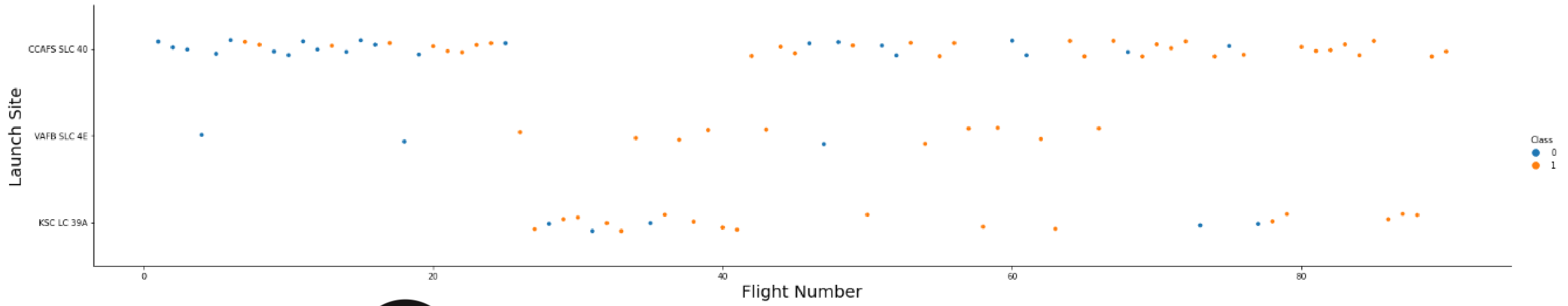
SKILLS NETWORK

RESULTS for EDA with Data Visualization

TASK 1: Visualize the relationship between Flight Number and Launch Site

Use the function `catplot` to plot `FlightNumber` vs `LaunchSite`, set the parameter `x` parameter to `FlightNumber`,set the `y` to `Launch Site` and set the parameter `hue` to `'class'`

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(x = "FlightNumber", y = "LaunchSite", hue = "Class", data = df, aspect = 5)
plt.xlabel("Flight Number", fontsize = 20)
plt.ylabel("Launch Site", fontsize = 20)
plt.show()
```

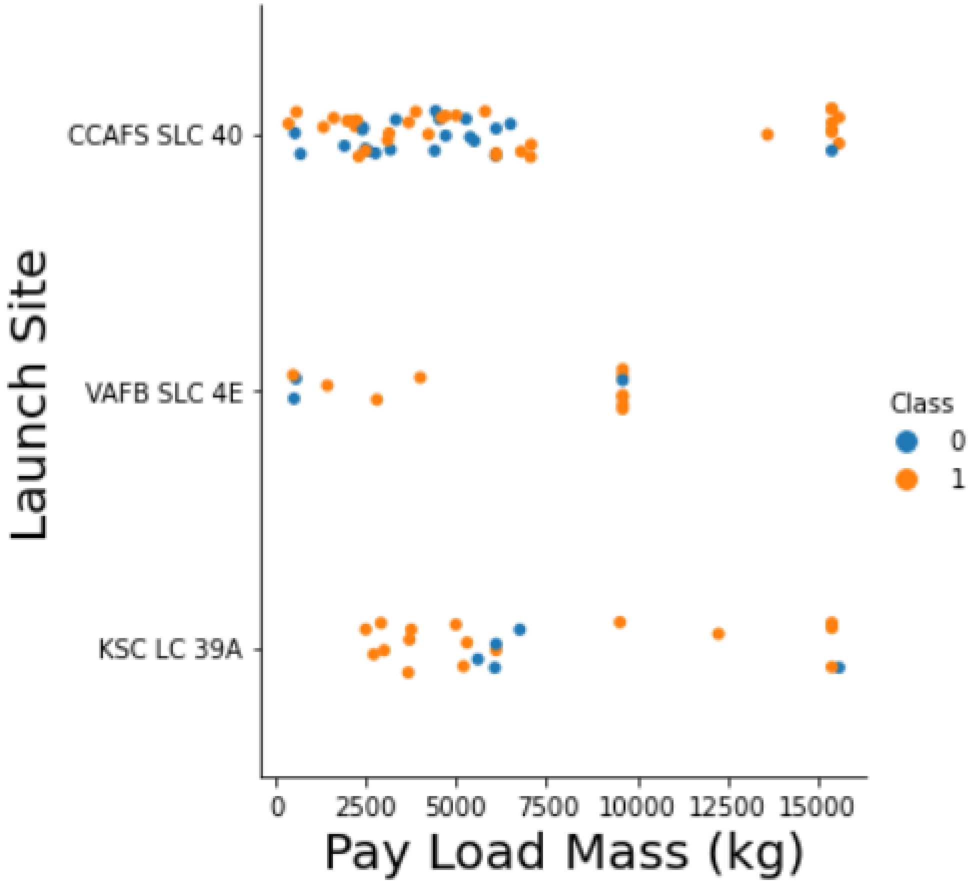


IBM Developer

SKILLS NETWORK

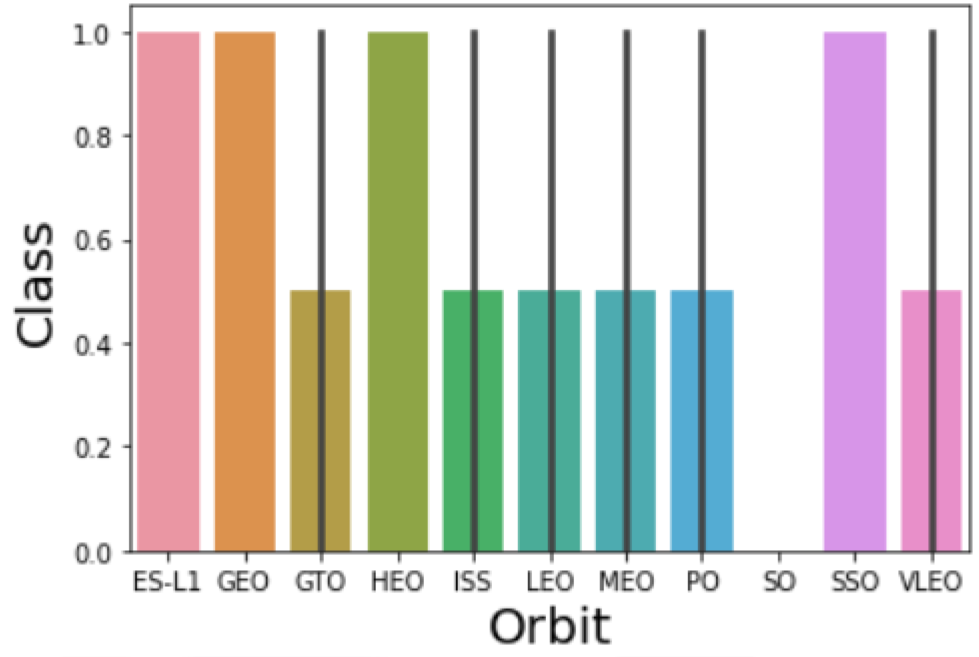
RESULTS for EDA with Data Visualization

We can observe CCAFS SLC 40 Launch Site has greater Payload Mass with success rate.



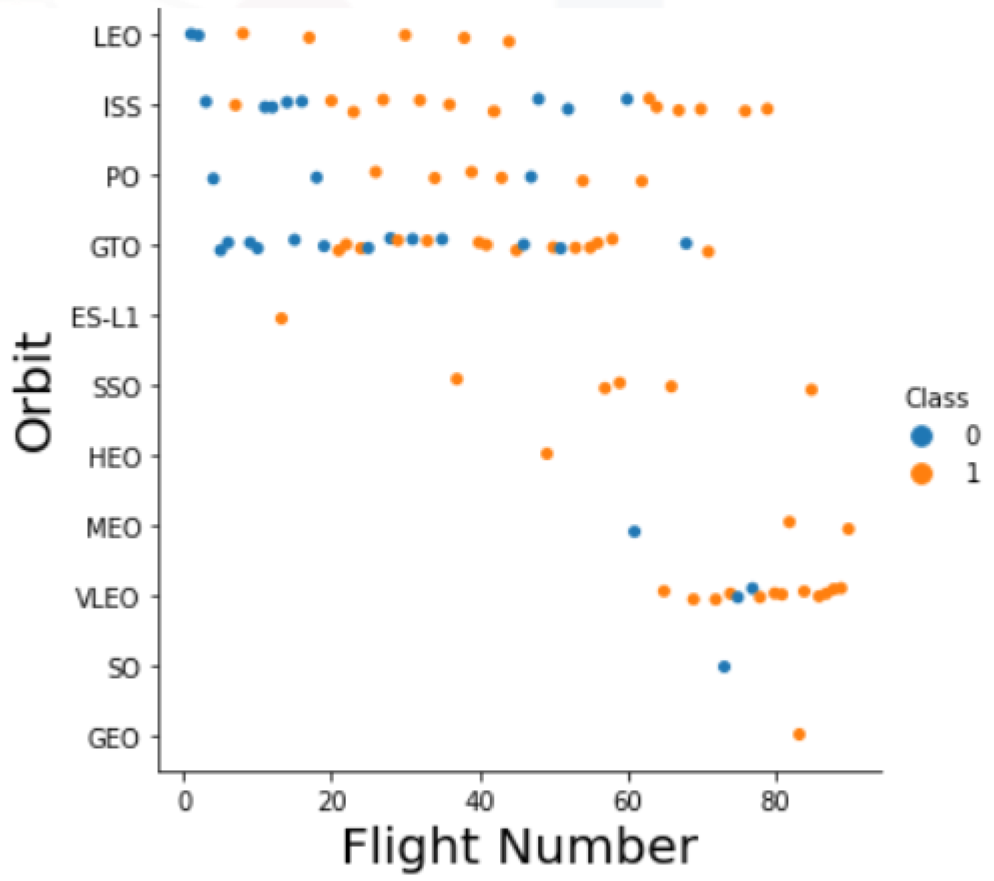
RESULTS for EDA with Data Visualization

We can observe ES-L1, GEO, HEO and SSO orbits have high success rate than the others.



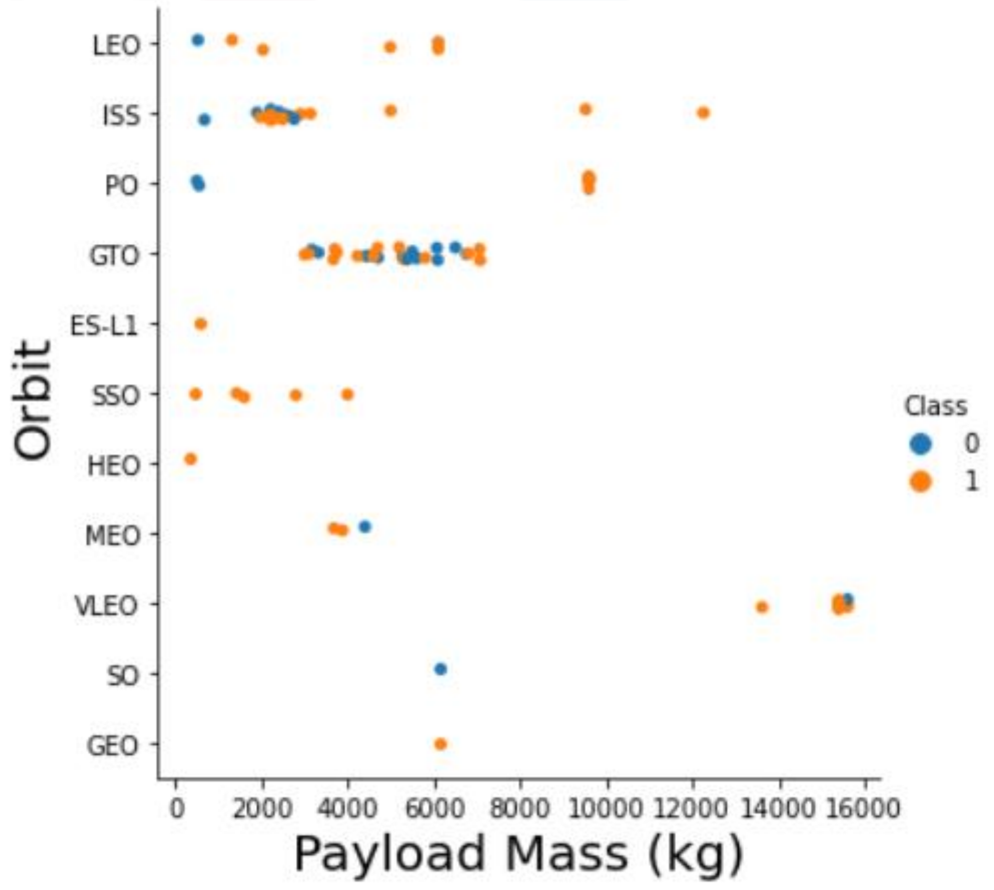
RESULTS for EDA with Data Visualization

We can observe for the LEO orbit success is related with flight number but we can't say that for GTO orbit.



RESULTS for EDA with Data Visualization

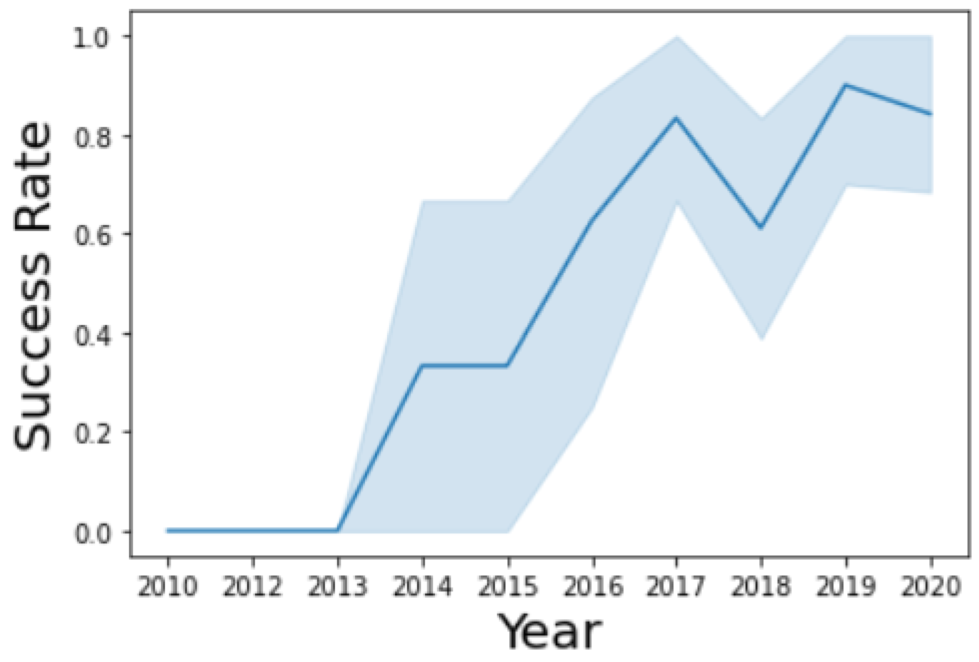
We can observe with heavy loads the successful landing or positive landing rate are more for VLEO, LEO, and ISS.



RESULTS for EDA with Data Visualization

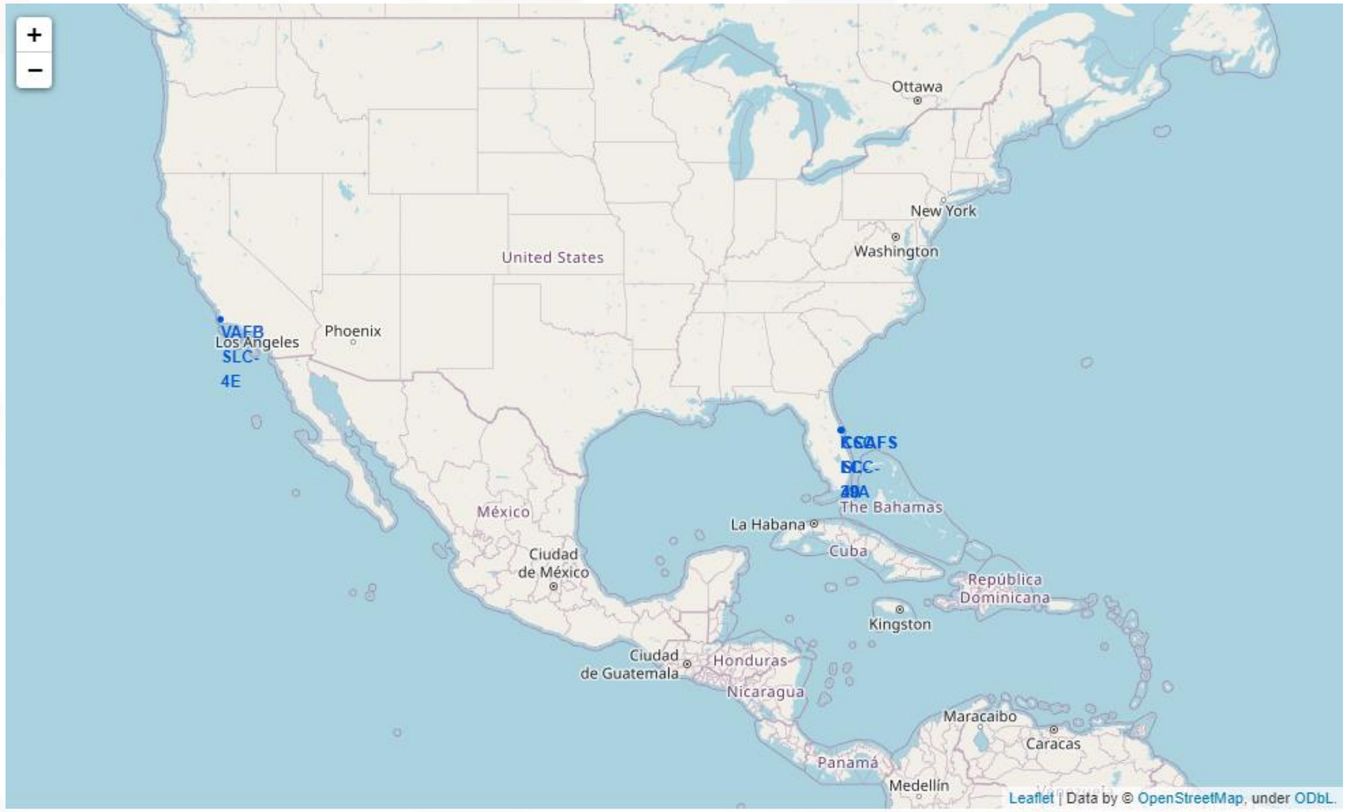
We can observe success rate since 2015 kept increasing almost linearly till 2017.

We can observe there is fluctuation between some years.



Results for Interactive Analytics with Folium

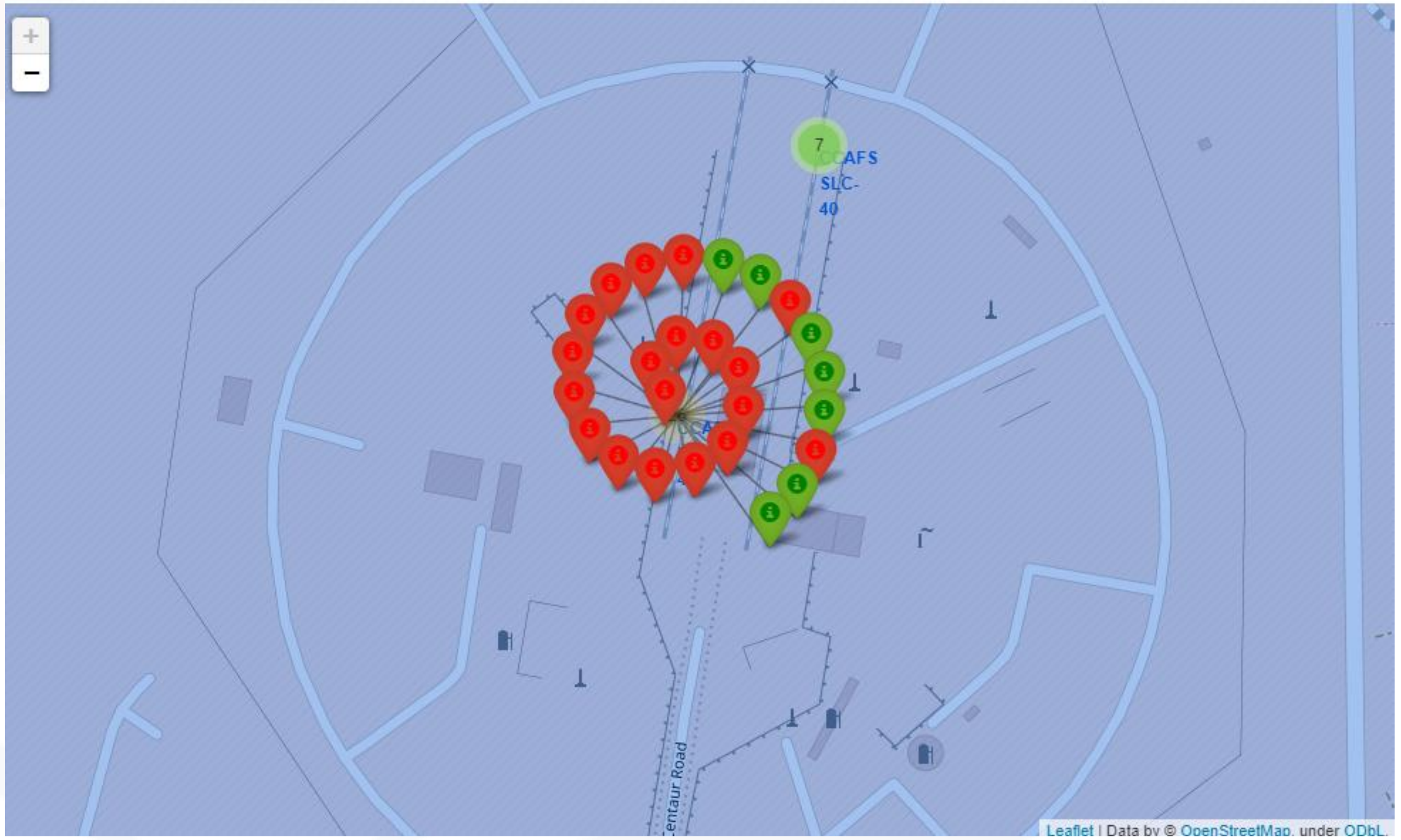
We can observe that SpaceX launch sites are in USA's Coasts. Especially Florida and California coasts.



Results for Interactive Analytics with Folium

After, we marked the success/failed launches for each site on map.

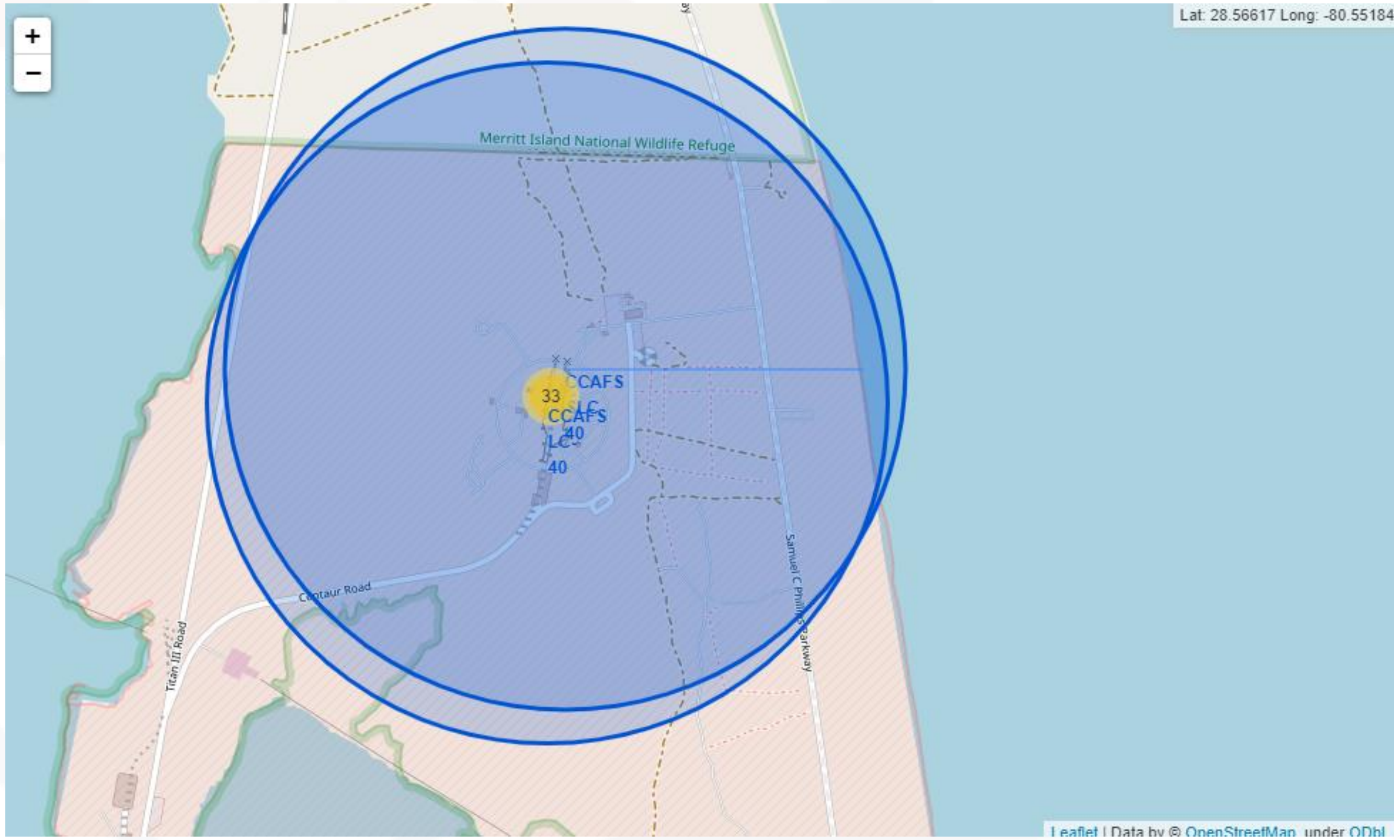
Green markers shows success launches.
Red markers shows failed launches.



Results for Interactive Analytics with Folium

We calculated the distance between launch sites to its proximities.

For example, we calculated the distance between selected launch site to coast.

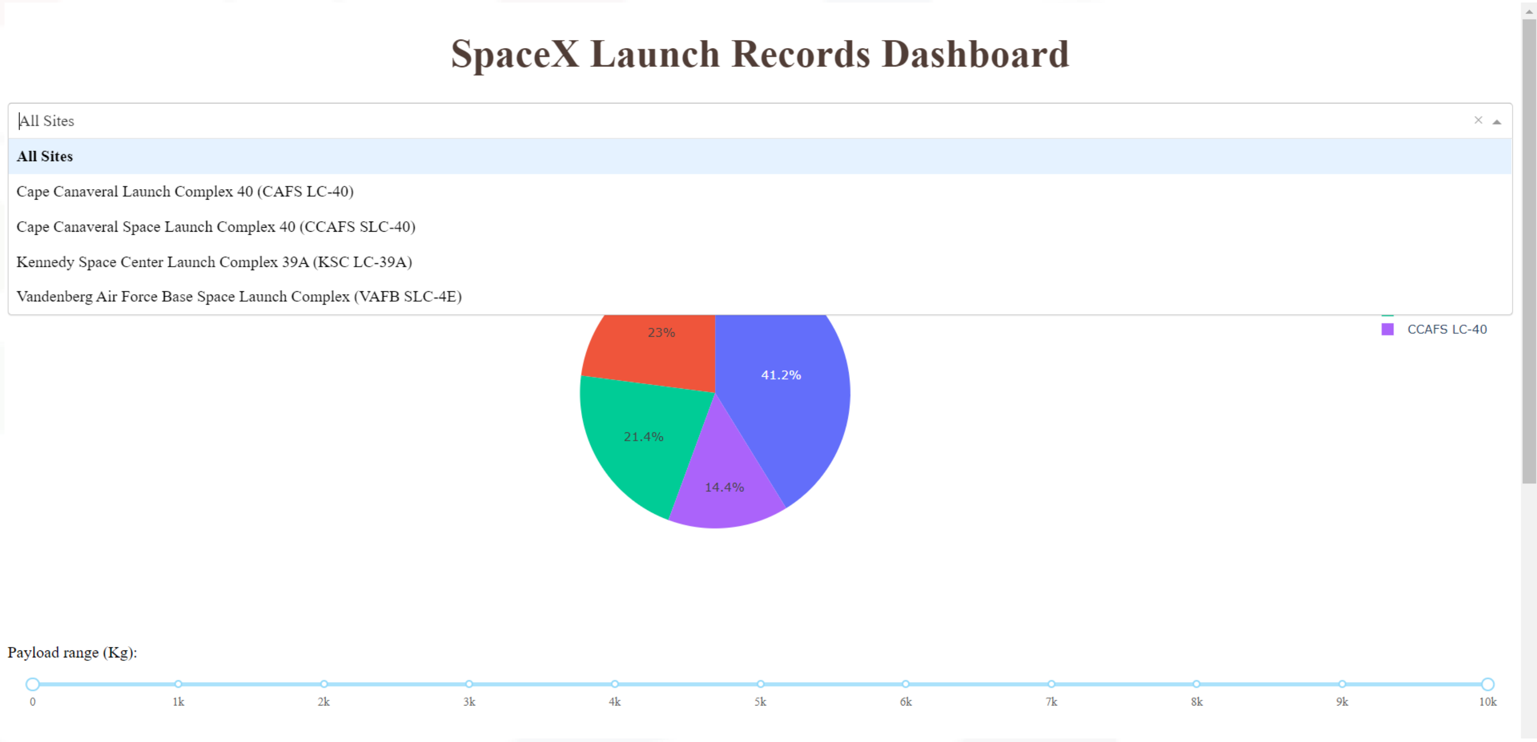

+
−
Lat: 28.56617 Long: -80.55184
Merritt Island National Wildlife Refuge
CCAFS
33
CCAFS
LC-40
40
Centaur Road
Titan III Road
Samuel C Phillips Parkway
Leaflet | Data by © OpenStreetMap, under ODbL

Results for Dashboard with Plotly Dash

We created dashboard with Plotly Dash. Following pictures shows structure of dashboard.

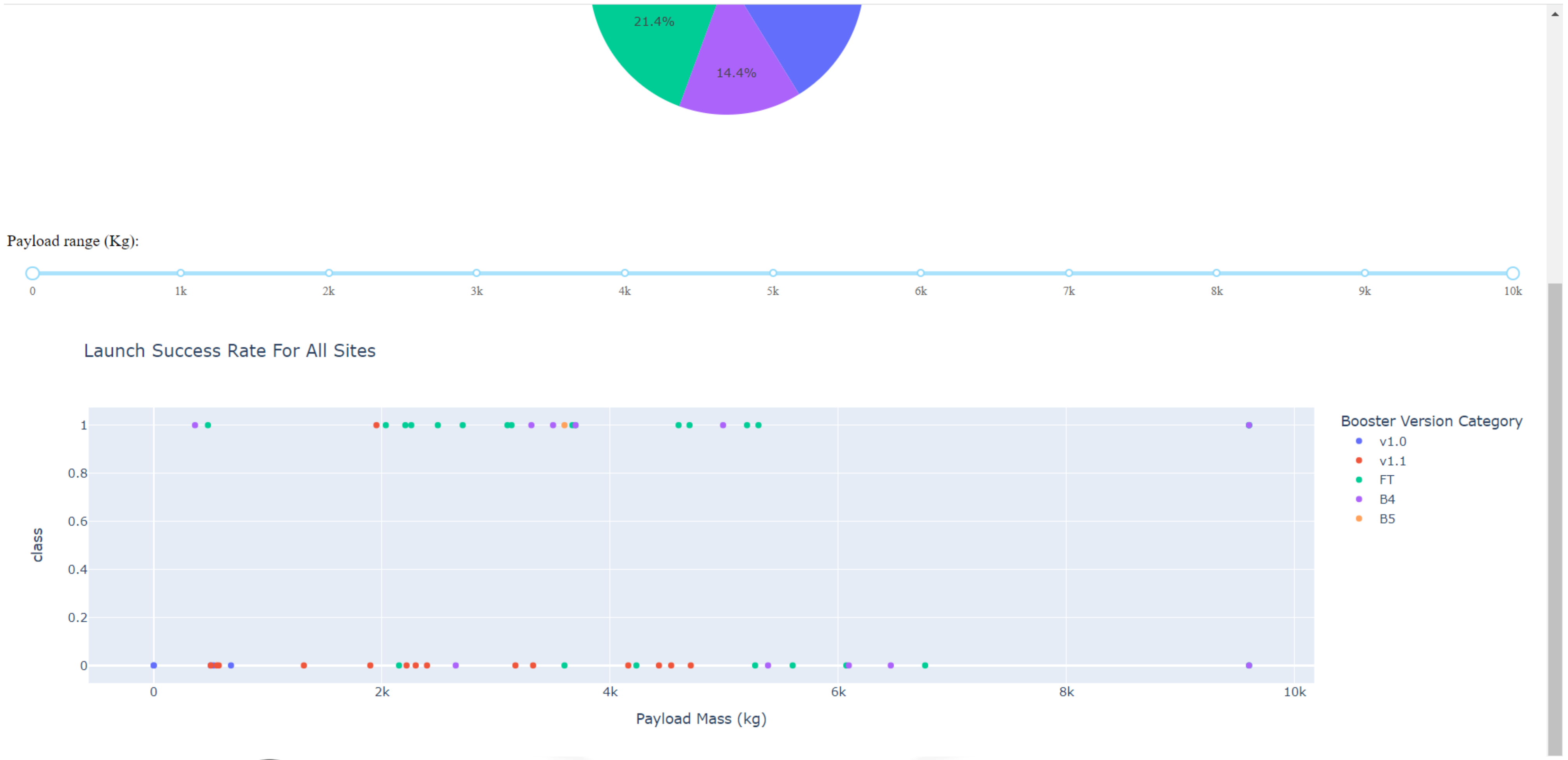
Maybe you will search answers for different questions but these are our findings:

We observed that KSC LC-39A site had most success rate compare to the others.

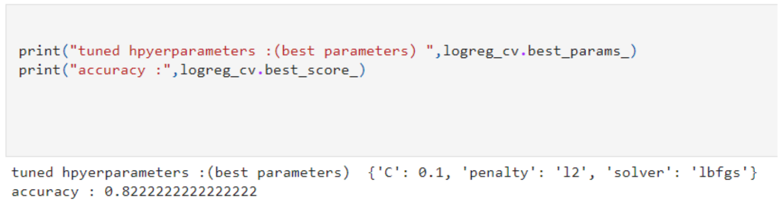

SpaceX Launch Records Dashboard

All Sites
- **All Sites**
- Cape Canaveral Launch Complex 40 (CAFS LC-40)
- Cape Canaveral Space Launch Complex 40 (CCAFS SLC-40)
- Kennedy Space Center Launch Complex 39A (KSC LC-39A)
- Vandenberg Air Force Base Space Launch Complex (VAFB SLC-4E)

Results for Dashboard with Plotly Dash

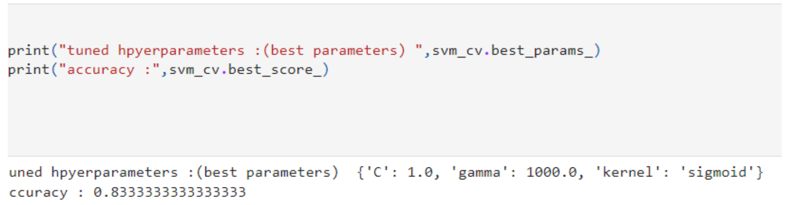


Results for ML Predictions



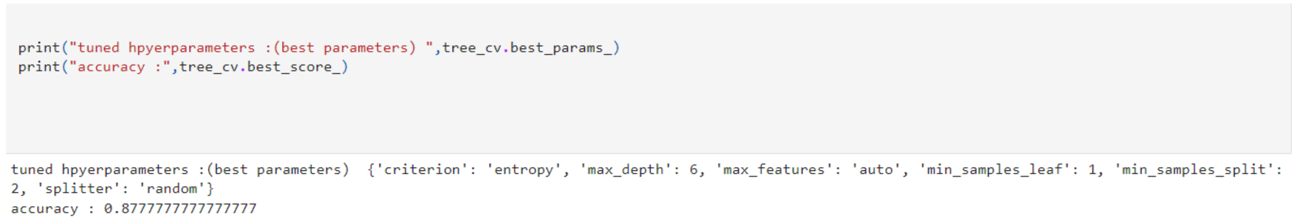
```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8222222222222222
```



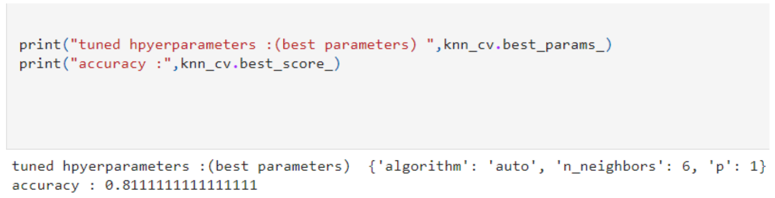
```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)
```

```
uned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 1000.0, 'kernel': 'sigmoid'}
ccuracy : 0.8333333333333333
```



```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split':
2, 'splitter': 'random'}
accuracy : 0.8777777777777777
```



```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 6, 'p': 1}
accuracy : 0.8111111111111111
```

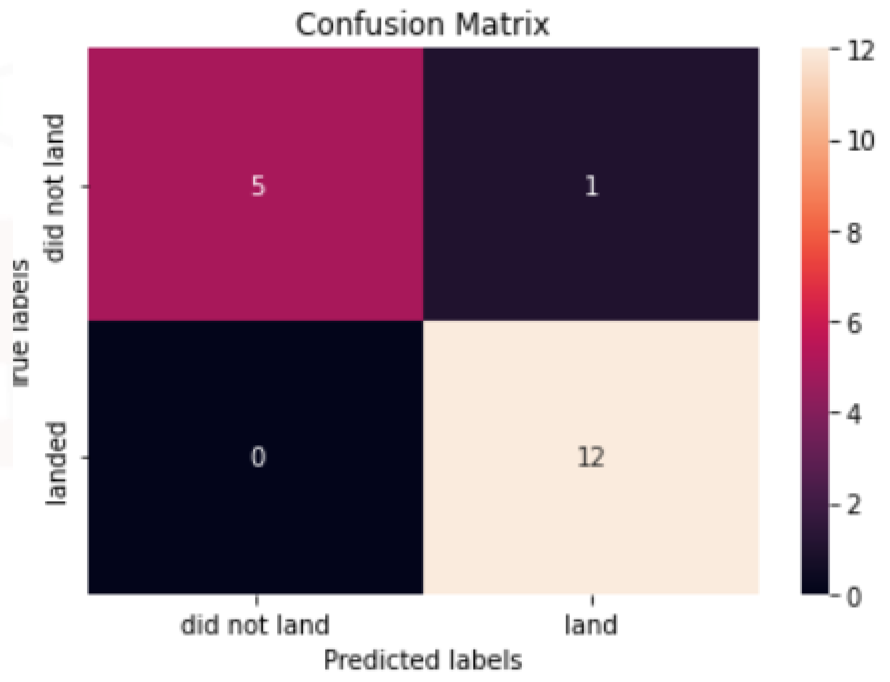
Best hyperparameters for selected ML algorithms.

Insights from Machine Learning Predictions

Results for ML Predictions

Logistic Regression confusion matrix.

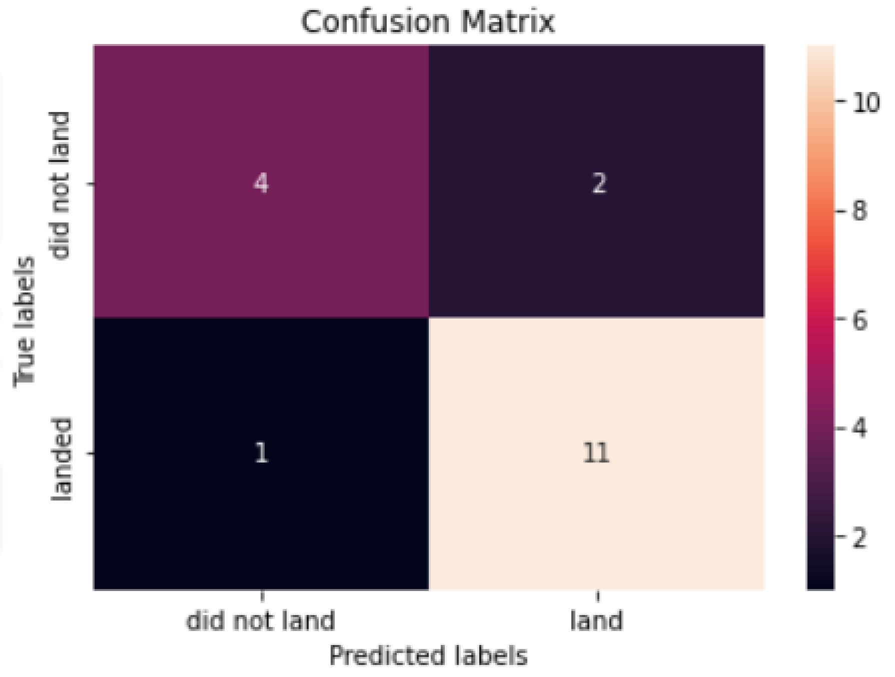
Examining the confusion matrix, we can observe that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Results for ML Predictions

SVM confusion matrix.

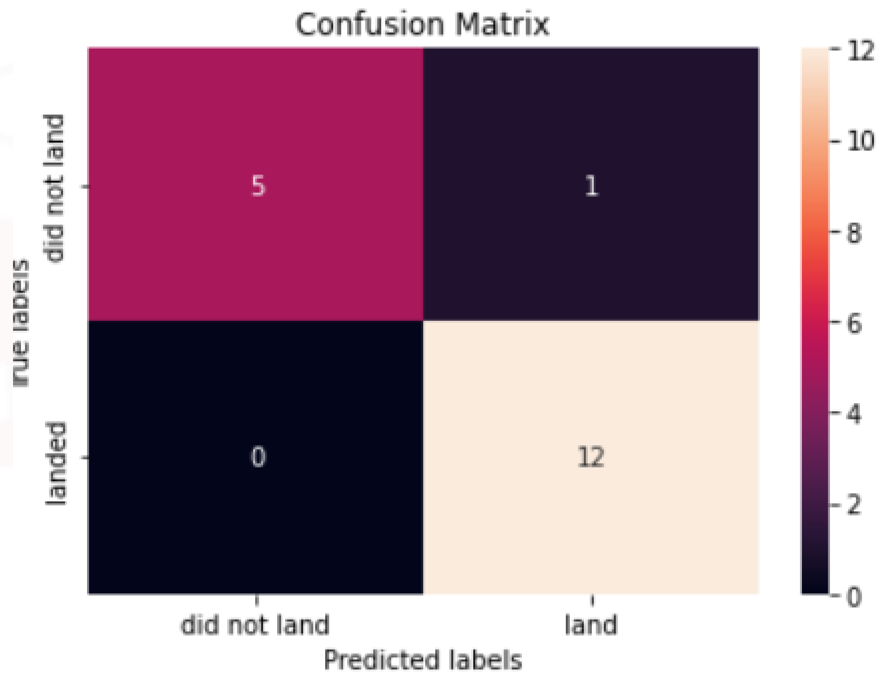
Examining the confusion matrix, we can observe that logistic regression can distinguish between the different classes. We see that the major problem is false positives. Also we see that the minor problem is false negatives



Results for ML Predictions

Decision Trees confusion matrix.

Examining the confusion matrix, we can observe that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



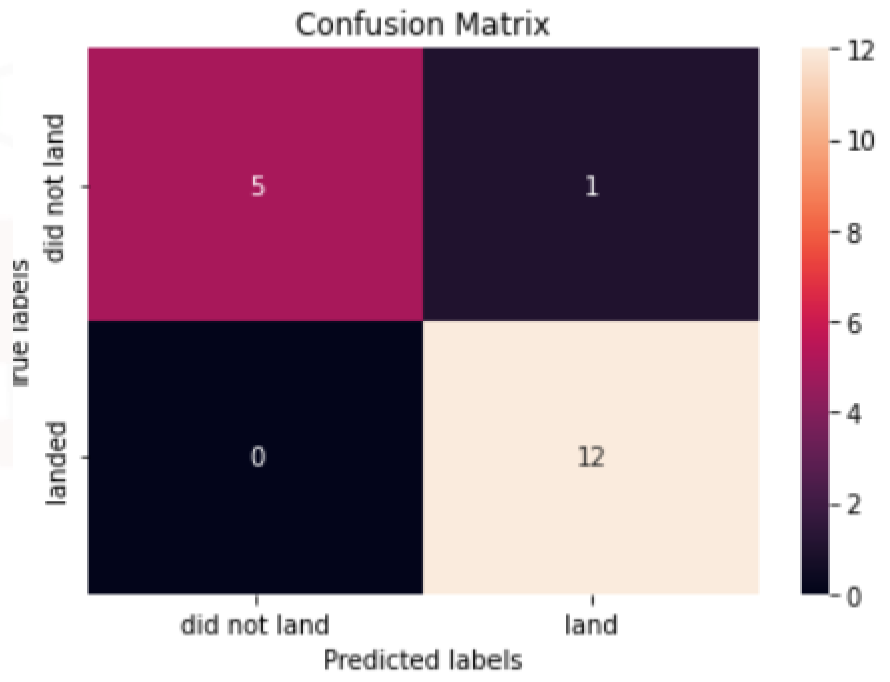
IBM Developer

SKILLS NETWORK

Results for ML Predictions

KNN confusion matrix.

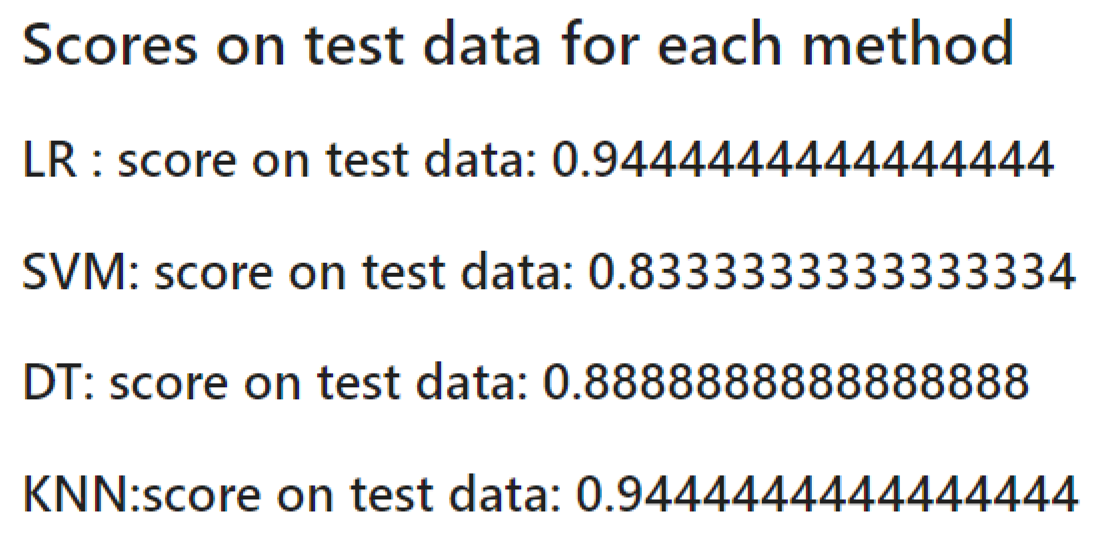
Examining the confusion matrix, we can observe that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Results for ML Predictions

We can observe that most successful algorithms are:

1. **Logistic Regression**
2. **KNN**



Scores on test data for each method

LR : score on test data: 0.9444444444444444

SVM: score on test data: 0.8333333333333334

DT: score on test data: 0.8888888888888888

KNN:score on test data: 0.9444444444444444

DISCUSSION

We performed 5 steps at this Data Science Capstone Project.

1. Data Collection with 2 different methods:
    1. Collection with API
    2. Collection with Web Scraping
2. Data Wrangling
3. Data Visualization
    1. Visualization with SQL
    2. Visualization with Graphs
    3. Visualization with Dashboard (Plotly Dash)
4. Features Engineering
5. Machine Learning Predictions
    1. Logistic Regression
    2. SVM
    3. Decision Trees
    4. KNN

For further projects we can improve that Project with different ranges of hyperparameters but it needs more processing time.

CONCLUSION

- Launch success rate since 2015 kept increasing almost linearly till 2017.
- KSC LC-39A site have most success rate compare to the others.
- ES-L1, GEO, HEO and SSO orbits have high success rate than the others
- CCAFS SLC 40 Launch Site has greater Payload Mass with success rate.
- Logistic Regression and KNN classifiers have best test accuracy.

Thank You

If you want to reach this presentation in pdf format check out the icon.

In addiction, I want to work and meet my peers. Reach out to me on Github. Check out the icon.