IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fahad Muhammad
18/07/2024

Outline

- ❖ Executive Summary P3
  - ➢ Summary of methodologies
  - ➢ Summary of results
- ❖ Introduction P5
- ❖ Methodology P6
- ❖ Results P18
- ❖ Conclusion P46

Executive Summary (methodologies)

The following methodology is used to address the factors for a successful rocket landing:

- Collection of Data
  - SpaceX REST API and web scraping techniques
- Wrangling of Data
  - data wrangling to create success/fail outcome variable
- Visualization of data
  - exploration of data with the following factors are considered:
    - payload, launch site, flight number and yearly trend
- Analysis of Data
  - the following statistics are obtained with SQL: total payload, payload range for successful launches, and total no. of successful and failed outcomes

- Exploration of Data
  - launch site success rates and proximity to geographical markers
- Visualization
  - Visualize the launch sites with the most success and successful payload ranges
- Model Building
  - Build Models to predict landing outcomes using
    - logistic regression,
    - support vector machine (SVM),
    - decision tree
    - and K-nearest neighbor (KNN)

Executive Summary (results)

- Exploratory Data Analysis:
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO, and SSO have a 100% success rat
- Visualization/Analytics:
  - Most launch sites are near the equator
  - and all are close to the coast
- Predictive Analytics:
  - All models performed similarly on the test set.
  - The decision tree model has outperformed

Introduction

- **Project background and context**
  - This project circles around the successful landing prediction of Falcon 9.
  - 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.

- Problems to address
  - What are the factors for the success of rocket landing?
  - How payload mass, launch site, number of flights, and orbits affect first-stage landing success ?
  - How much is the rate of successful landings over time ?
  - Can we predict some model for successful landing (binary classification) ?

Section 1

Methodology

Methodology (Executive Summary)

- **Data collection methodology**
  - data is collected by REST API and web scraping techniques
- **Perform data wrangling**
  - Ignore irrelevant columns, applying one hot encoding
- **Perform exploratory data analysis (EDA) using visualization and SQL**
  - Various graphs and charts are constructed to depict relationship between variables to dig hidden facts in data
- **Perform interactive visual analytics**
  - Use of API Folium and Plotly Dash to promote understanding
- **Perform predictive analysis using classification models**
  - Build models and calculate accuracy of model

Data Collection – (Steps Performed)

Data is collected by Rest API and web scraping technique

by using REST API

1 call specialized SpaceX REST API

2 receive response in JSON

3 Normalize data into flat file

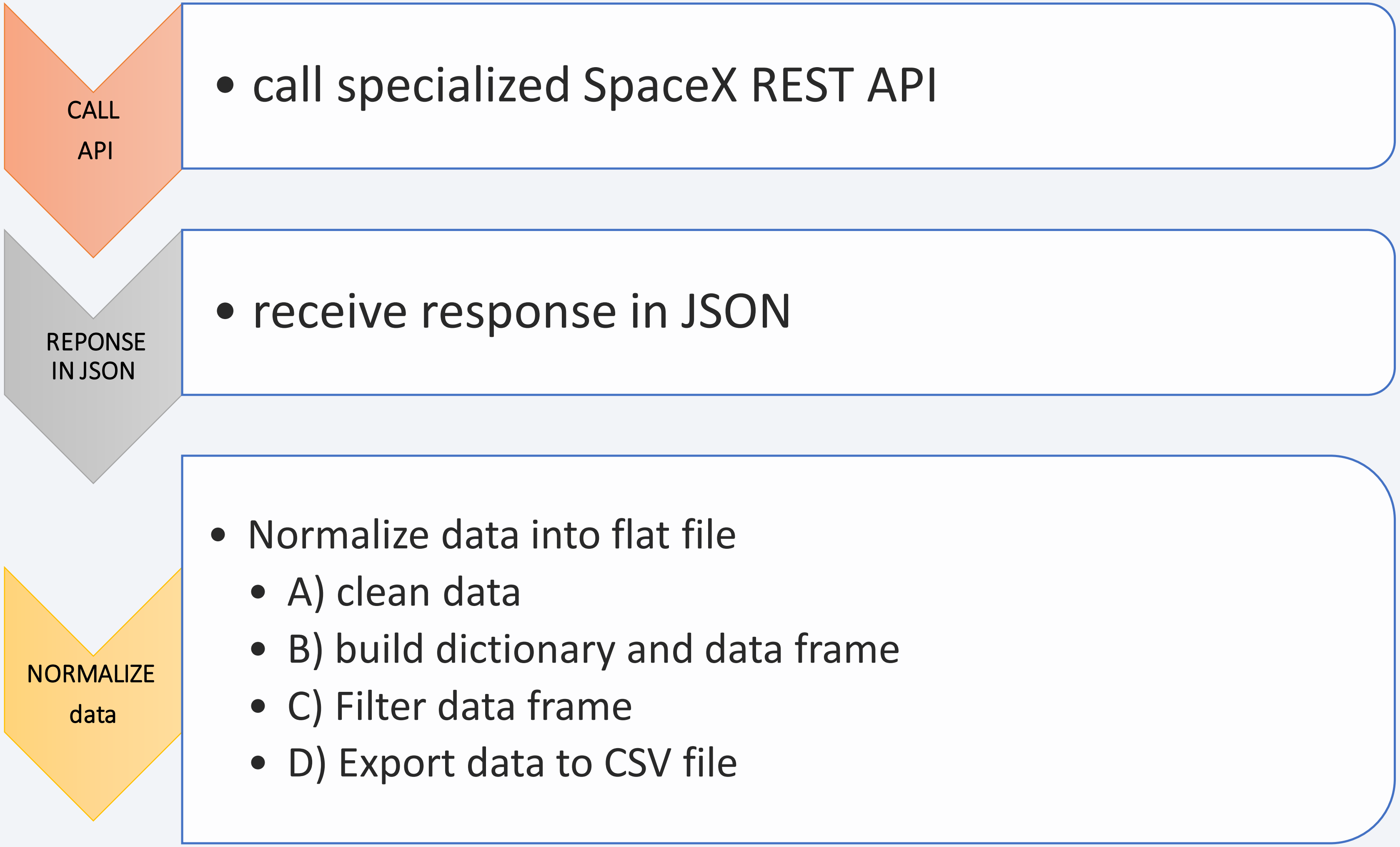
by web scraping techniques

1 get HTML response from Wiki

2 extract data using beautiful soup

3 Normalize data into flat file

Data Collection SpaceX API



CALL API

- call specialized SpaceX REST API

REPONSE IN JSON

- receive response in JSON

NORMALIZE data

- Normalize data into flat file
  - A) clean data
  - B) build dictionary and data frame
  - C) Filter data frame
  - D) Export data to CSV file

Git link

Data Collection - Scraping

By using web scraping techniques

1 get HTML response from Wiki

2 Build beautifulSoup object

3 Build Data from soup

A. Find required table

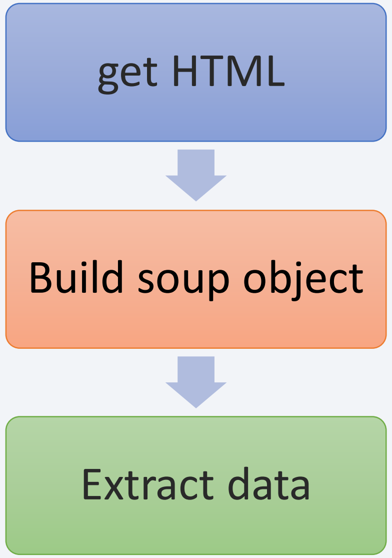
B. Get desired columns

C. Create Dictionary of colums

D. Append data to keys

E. Assign Dictionary to DataFrame

F. Register DataFrame to CSV file


get HTML

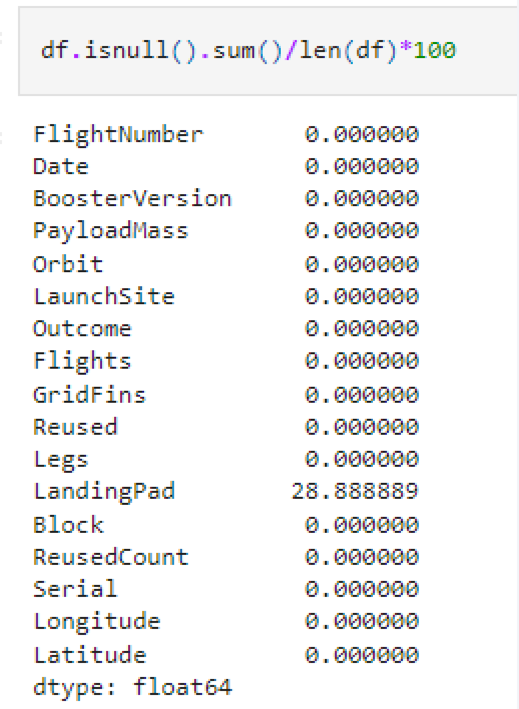
Build soup object

Extract data

git link

Data Wrangling

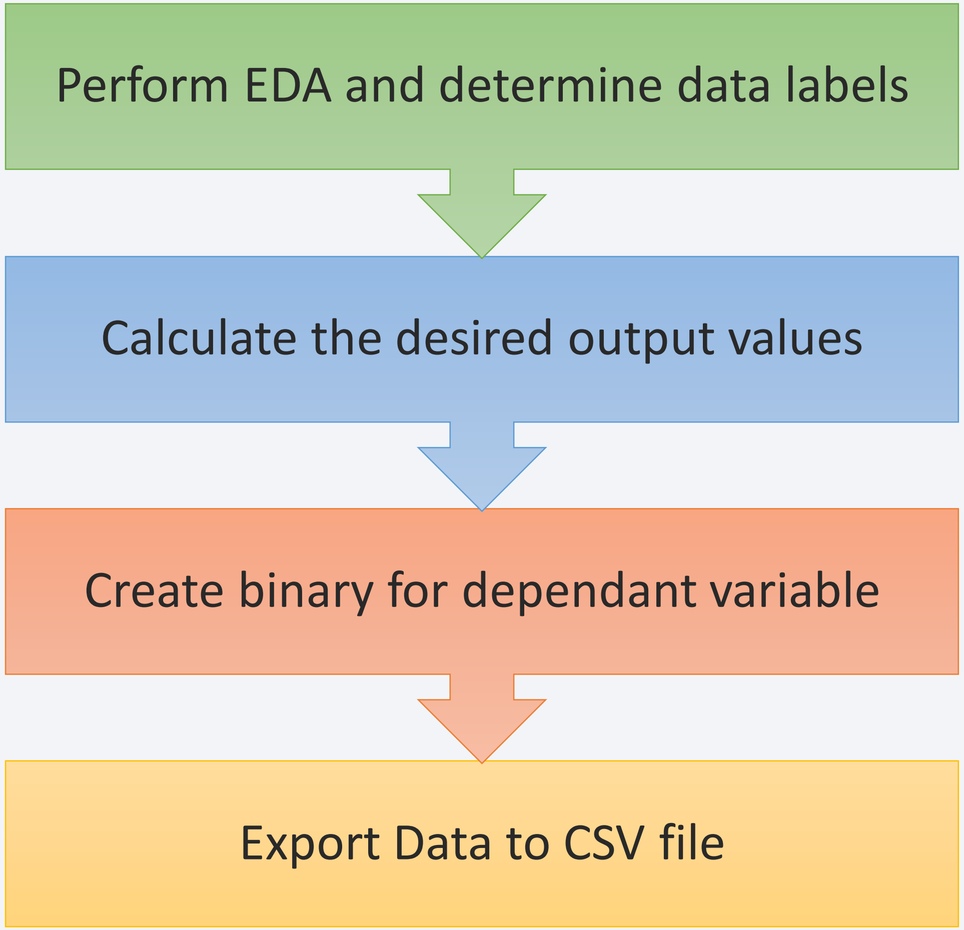
Example from case study:
Identify and calculate the percentage of the missing values in each attribute



```
df.isnull().sum()/len(df)*100
```

```
FlightNumber       0.000000
Date               0.000000
BoosterVersion     0.000000
PayloadMass        0.000000
Orbit              0.000000
LaunchSite         0.000000
Outcome            0.000000
Flights            0.000000
GridFins           0.000000
Reused             0.000000
Legs               0.000000
LandingPad        28.888889
Block              0.000000
ReusedCount        0.000000
Serial             0.000000
Longitude          0.000000
Latitude           0.000000
dtype: float64
```

git link



Perform EDA and determine data labels

Calculate the desired output values

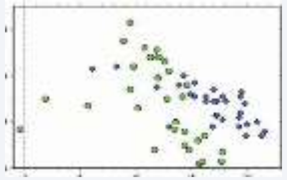
Create binary for dependant variable

Export Data to CSV file

EDA with Data Visualization

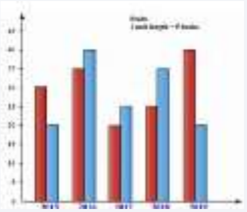
Scatter Graphs

- Various scatter graphs are build to visualize Flight_no Vs Payload Mass, Flight_no Vs Launch Site, Payload vs Launch Site, Orbit vs Payload Mass, etc.



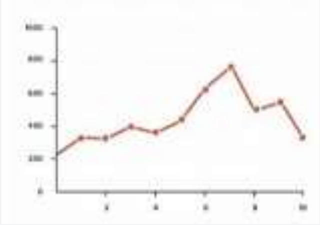
Bar Graph

- Mean vs. Orbit



Line Graph

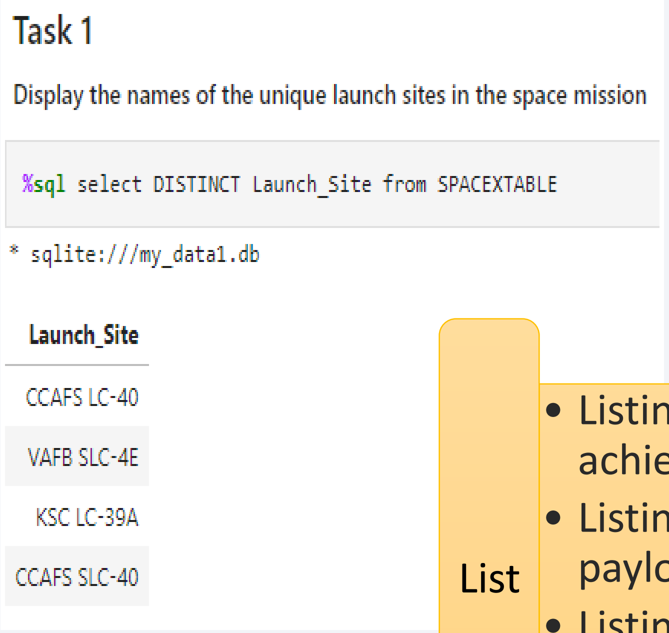
- Success Rate vs Year



git link

EDA with SQL

Example from case study



Task 1

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

```
%sql select DISTINCT Launch_Site from SPACEXTABLE
```

```
* sqlite:///my_data1.db
```

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

Display

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1 …etc

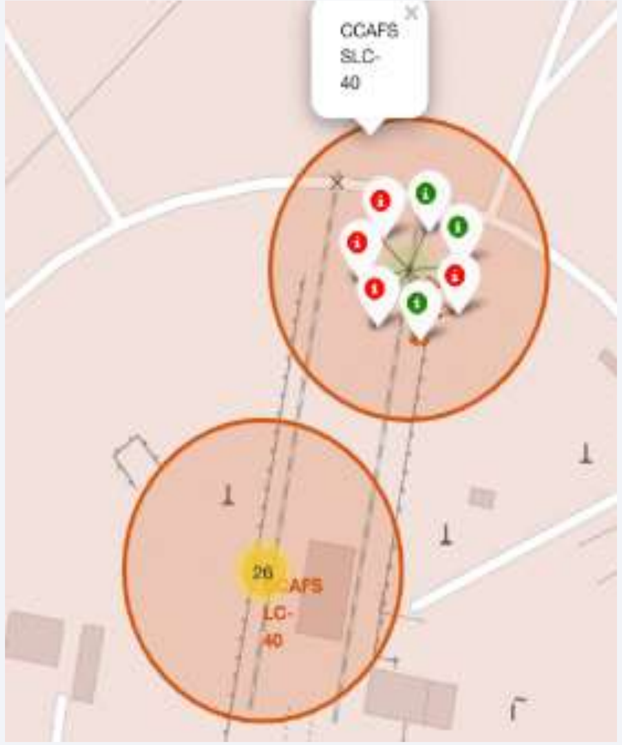
List

- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass ….etc

[Git link]

Build an Interactive Map with Folium

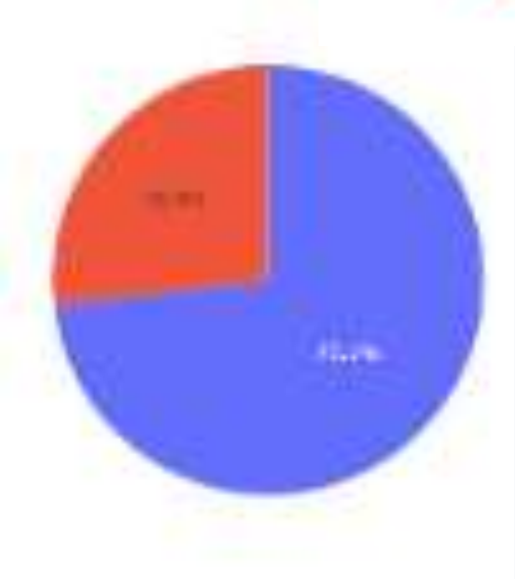
- Launch success rate may depend on the location and proximity of a launch site. Folium Interactive Map was used for visualizing and analyzing SpaceX Launch Sites.
  - **TASK 1:** Mark all launch sites on a map
  - **TASK 2:** Mark the success/failed launches for each site on the map
  - **TASK 3:** Calculate the distances between a launch site to its proximities



git link

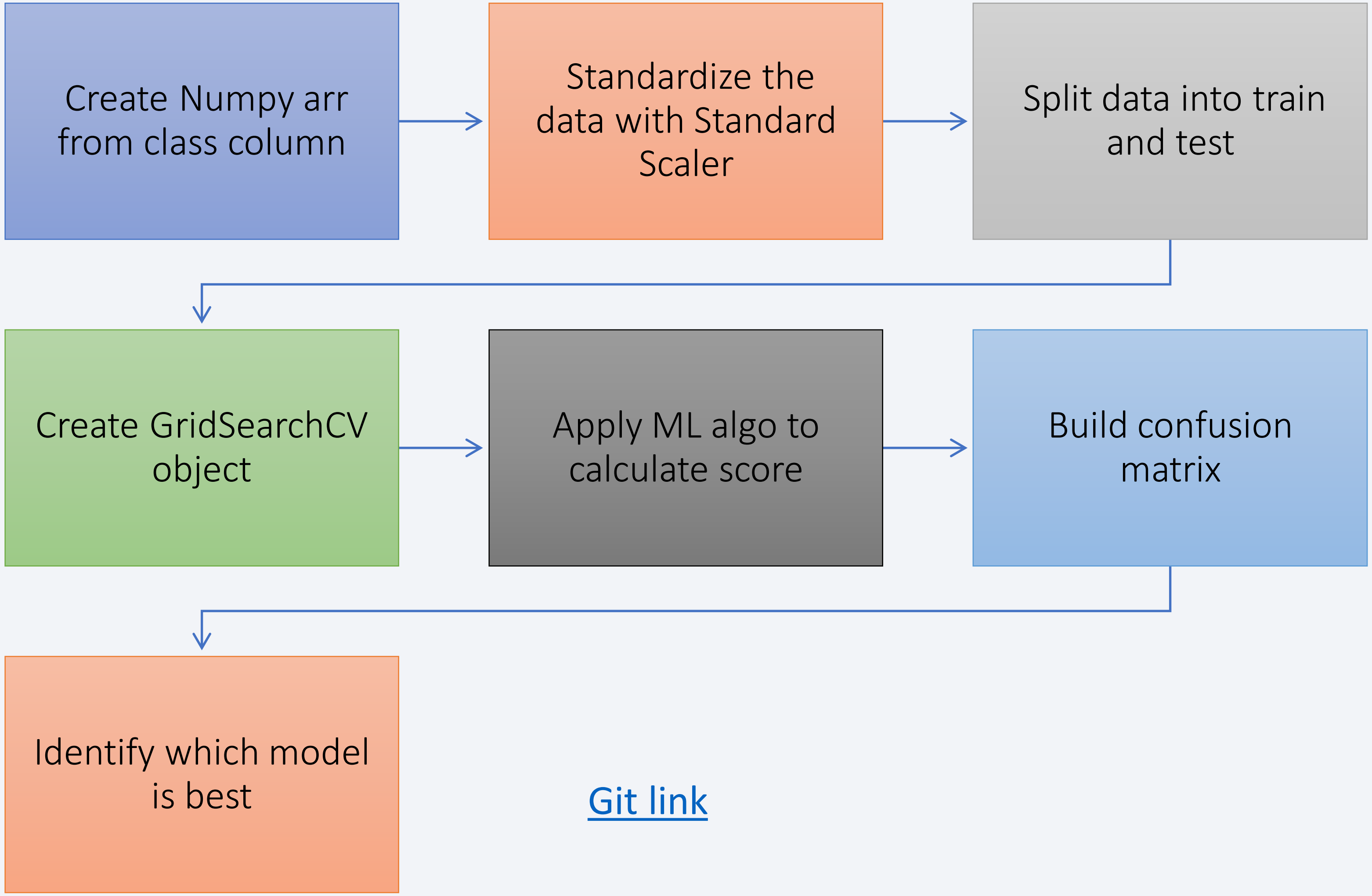
Build a Dashboard with Plotly Dash

- After visual analysis using the dashboard, we are able to obtain some insights to answer the following five questions:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?



git link

Predictive Analysis (Classification)



Create Numpy arr from class column

Standardize the data with Standard Scaler

Split data into train and test

Create GridSearchCV object

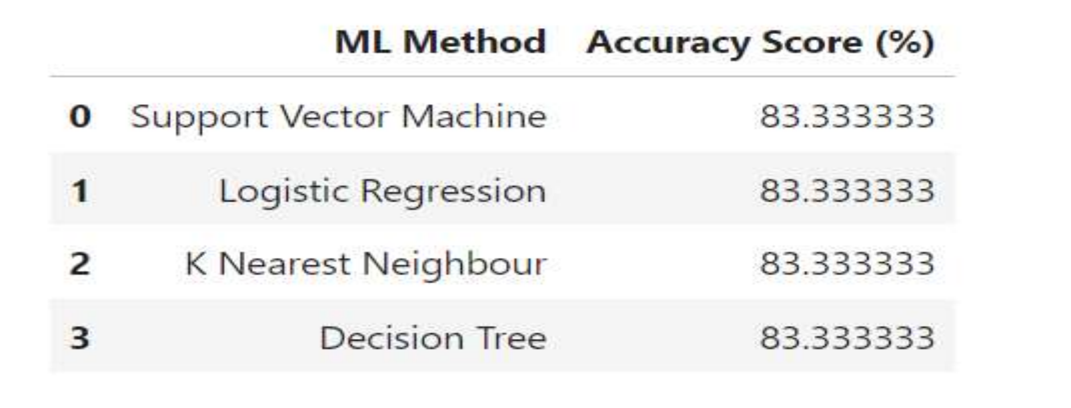
Apply ML algo to calculate score

Build confusion matrix

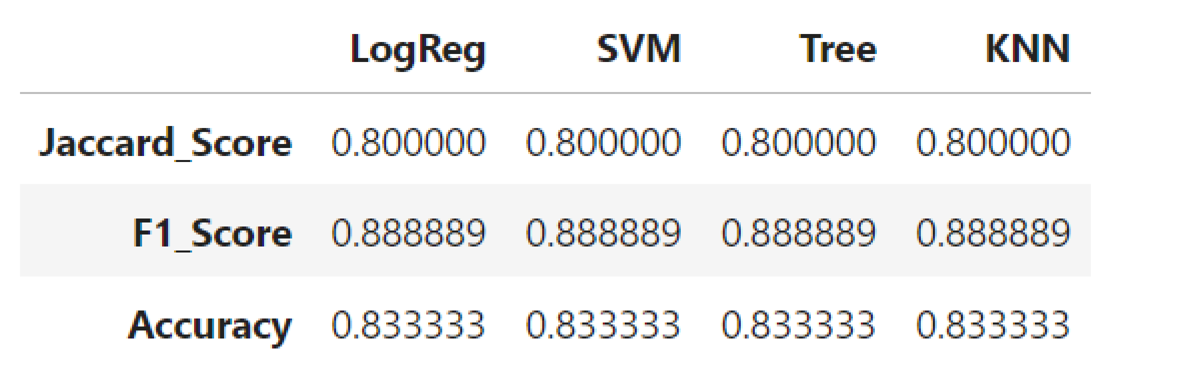
Identify which model is best

Git link

Results



	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

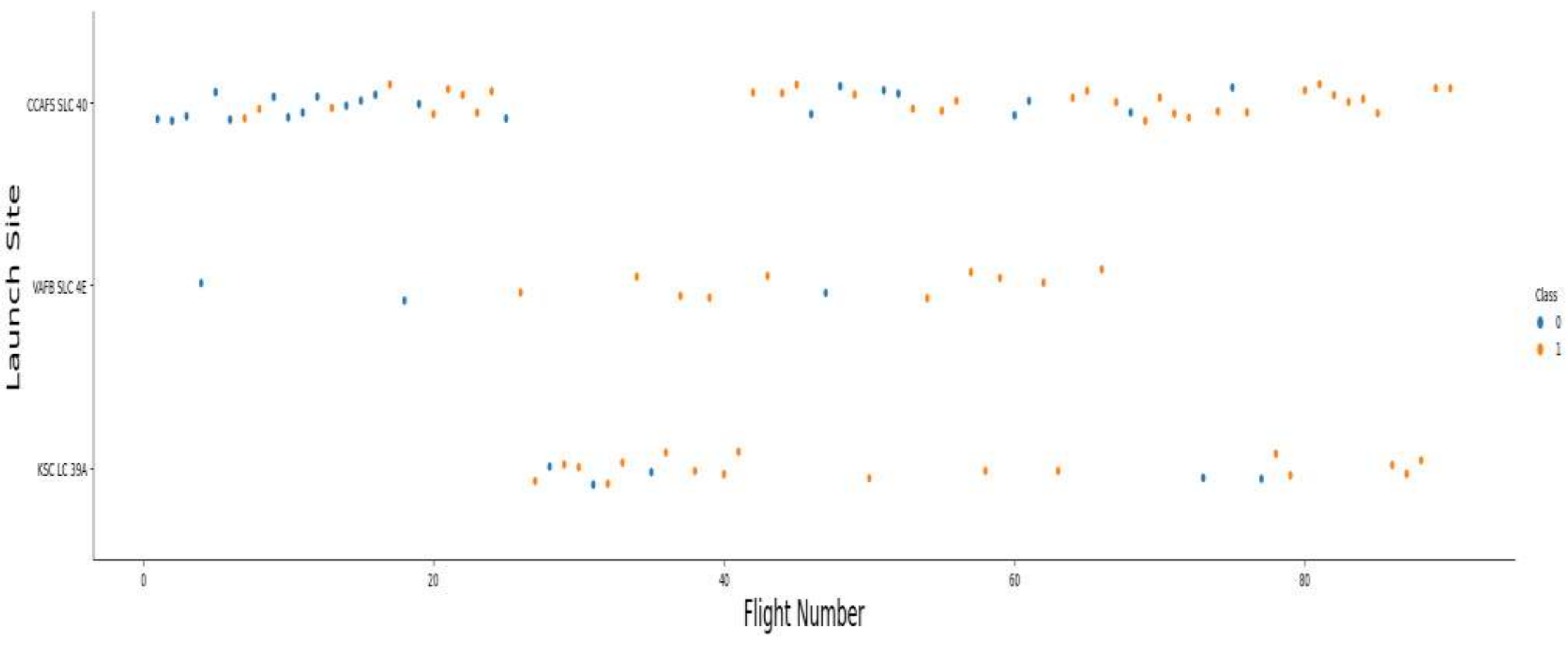
- Best model is DecisionTree with a score of 0.875
- Best params is : {'criterion': 'entropy', 'max_depth': 2, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}

Section 2

Insights drawn from EDA

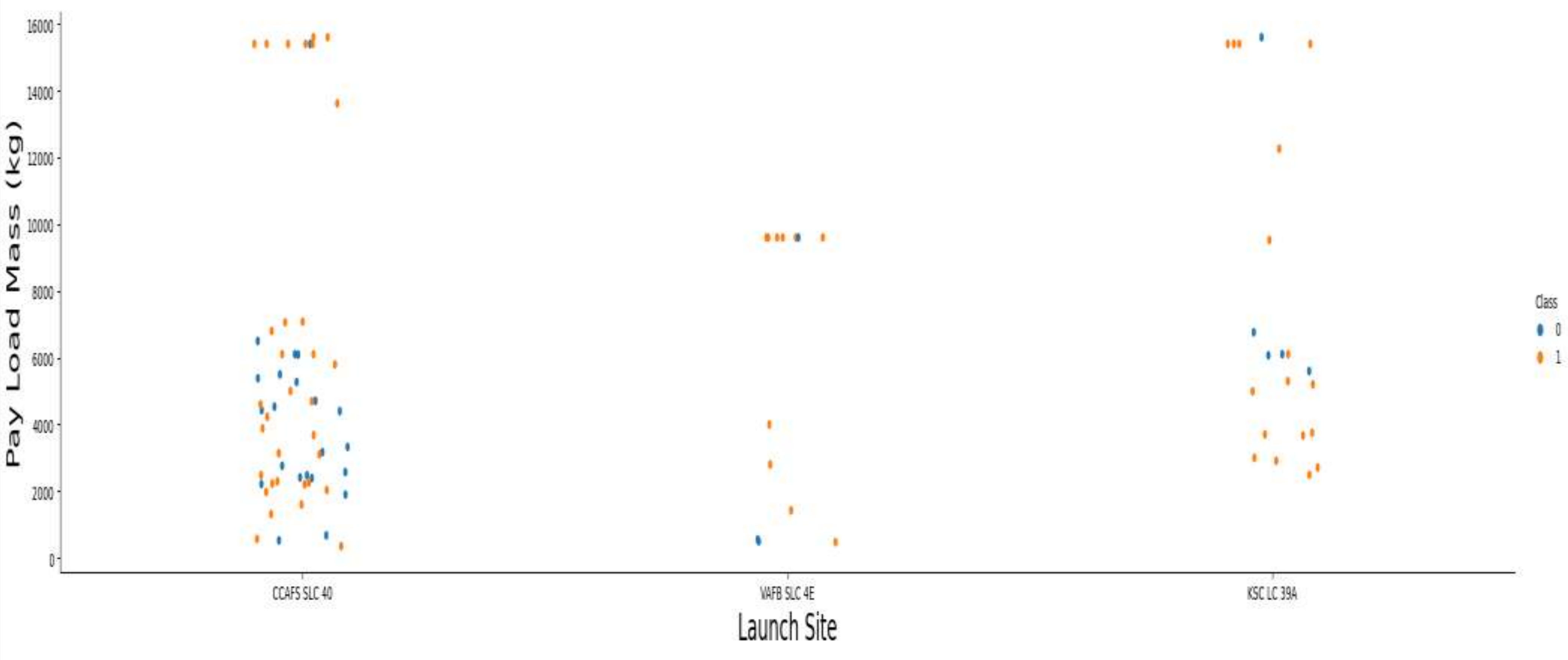
Flight Number vs. Launch Site

By using the function catplot to plot FlightNumber vs LaunchSite, the parameter x FlightNumber, the y Launch Site and the parameter hue as 'class'



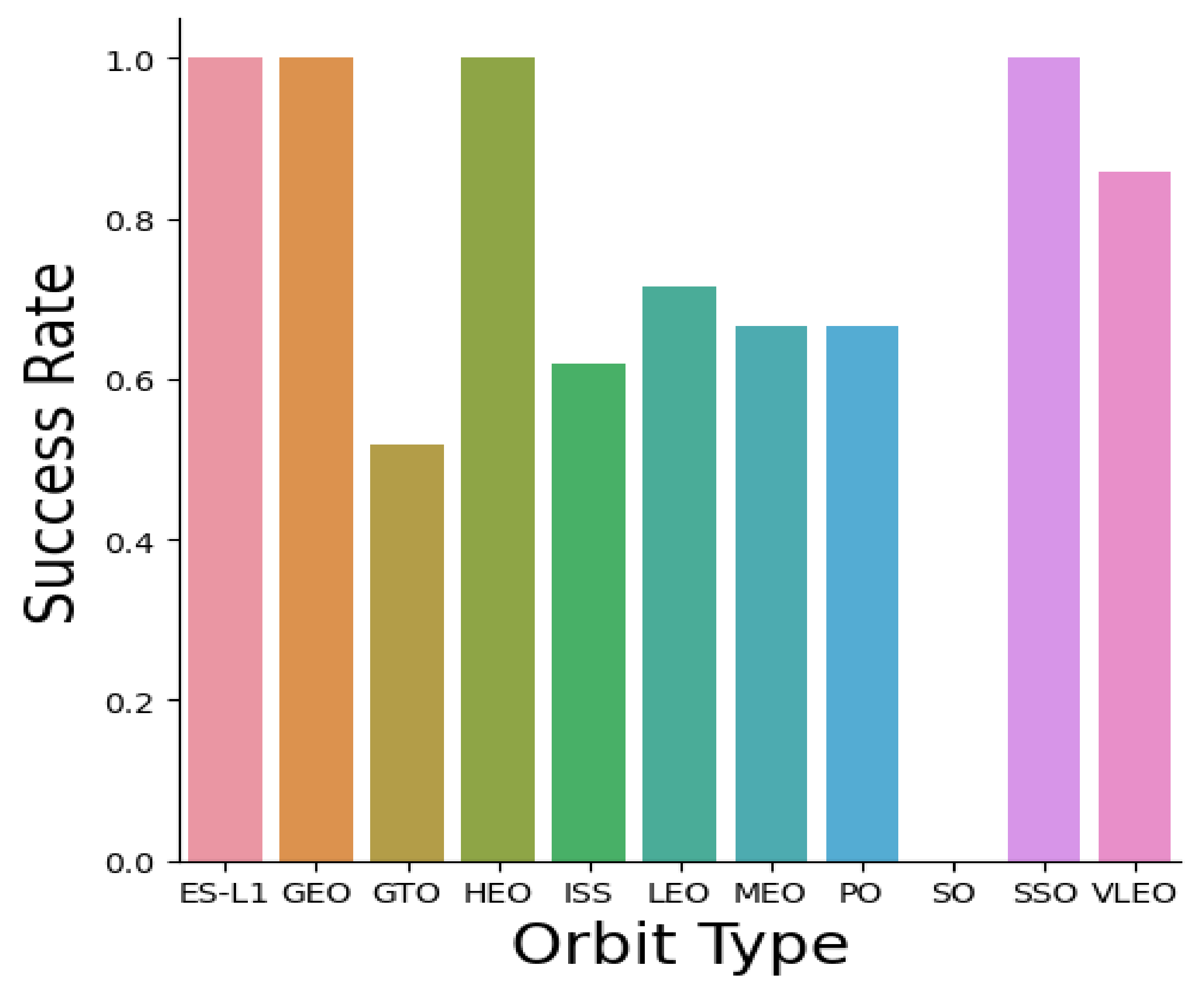
Payload vs. Launch Site

a scatter point chart with x axis Pay Load Mass (kg) and y axis the launch site, and hue to be the class value

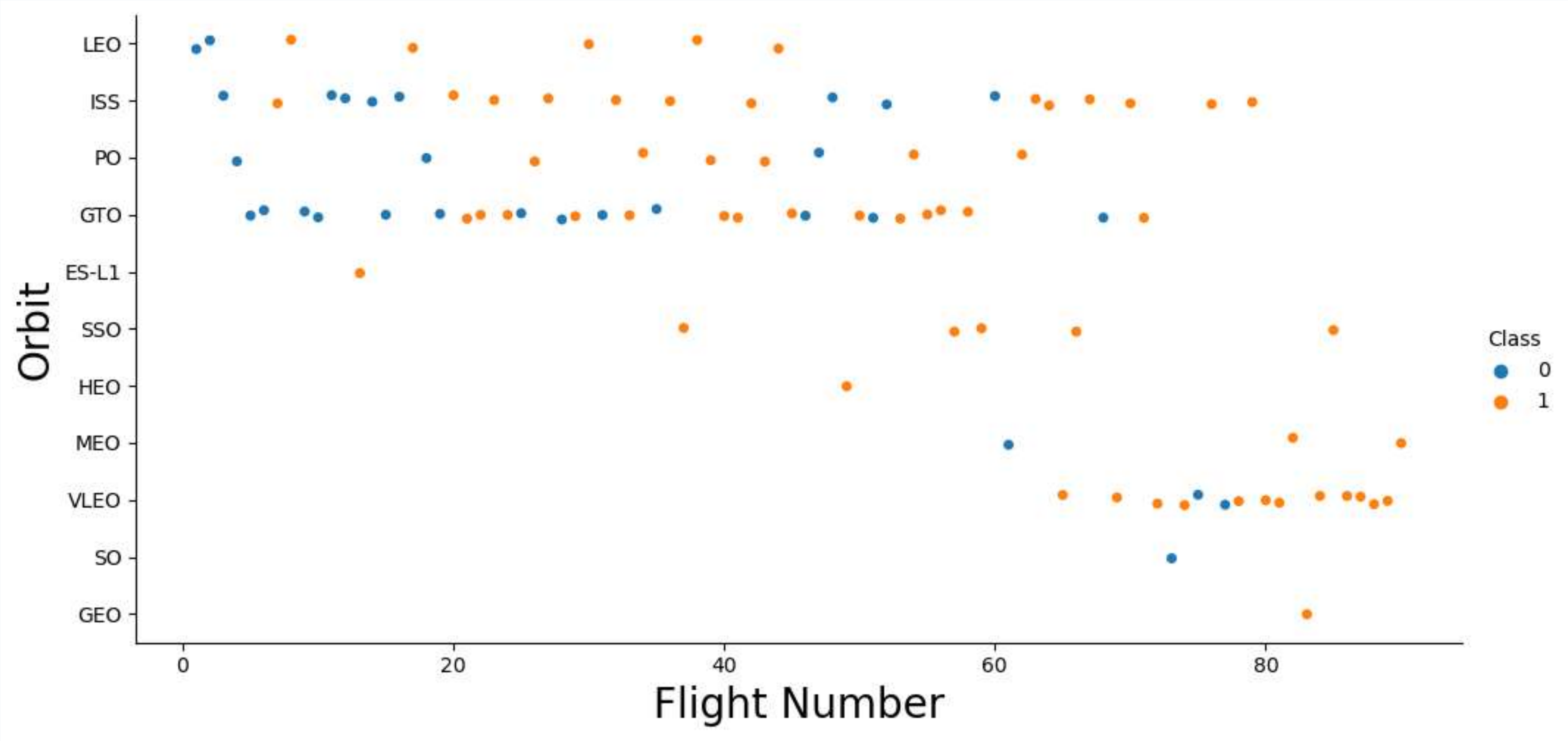


Success Rate vs. Orbit Type

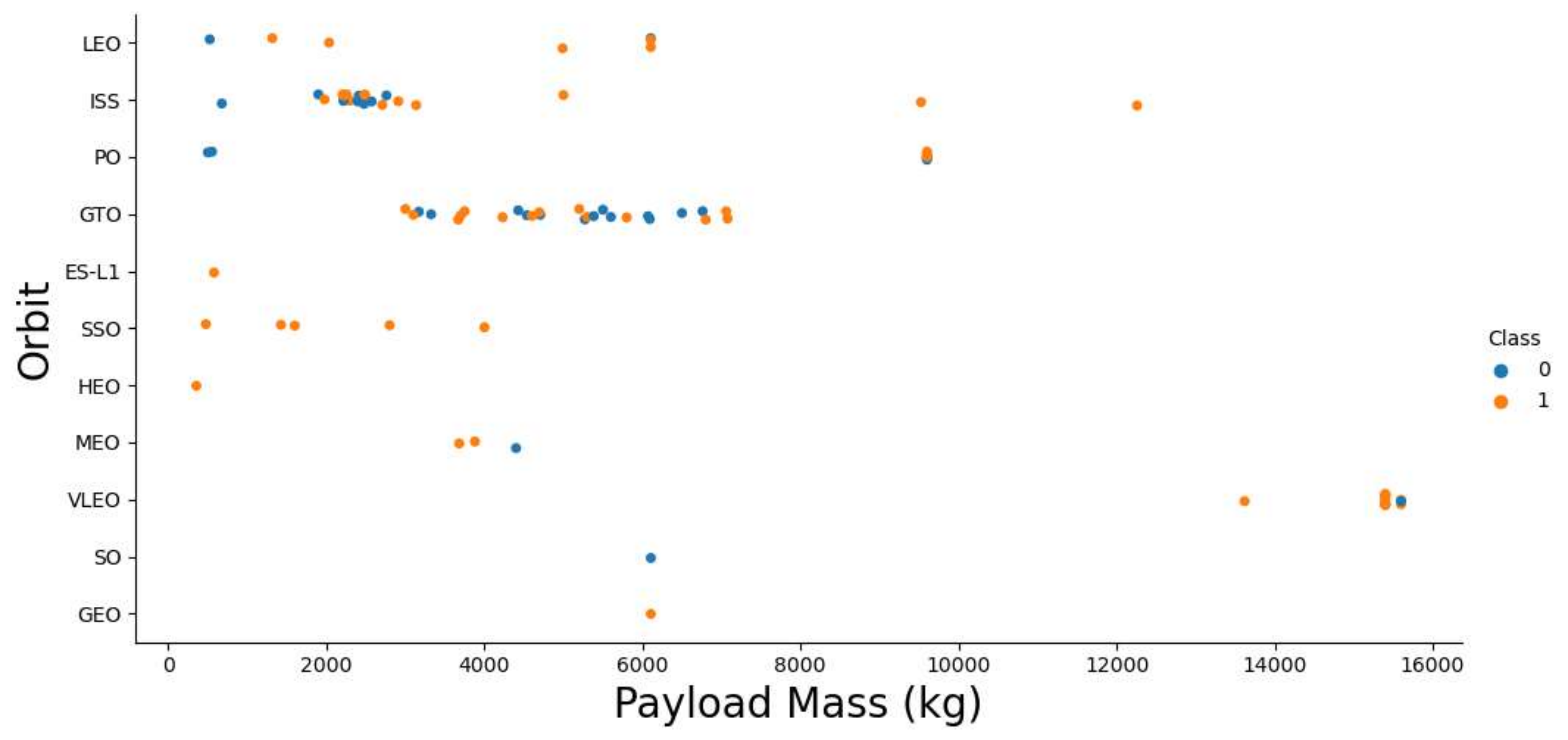
relationship between success rate and orbit type



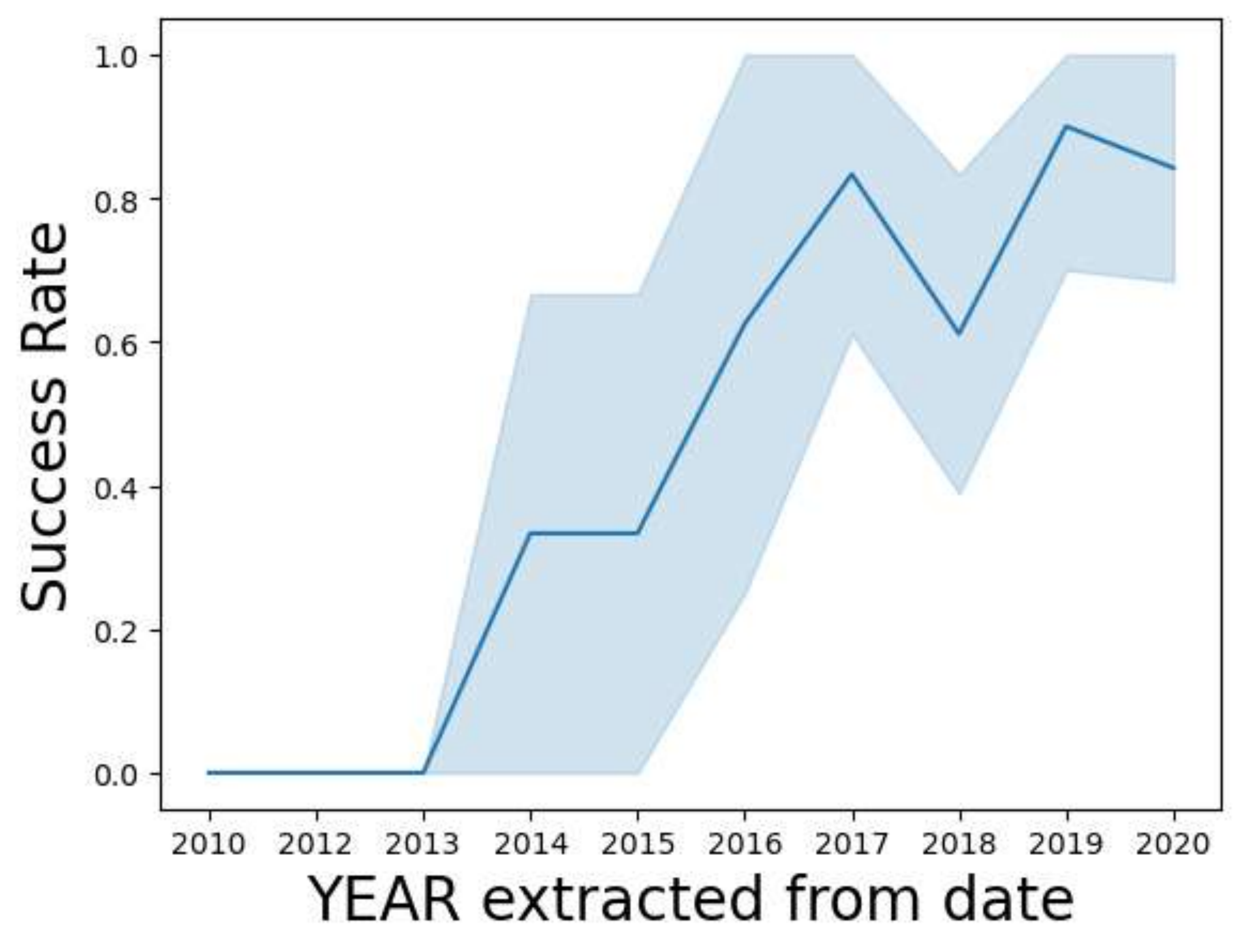
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend

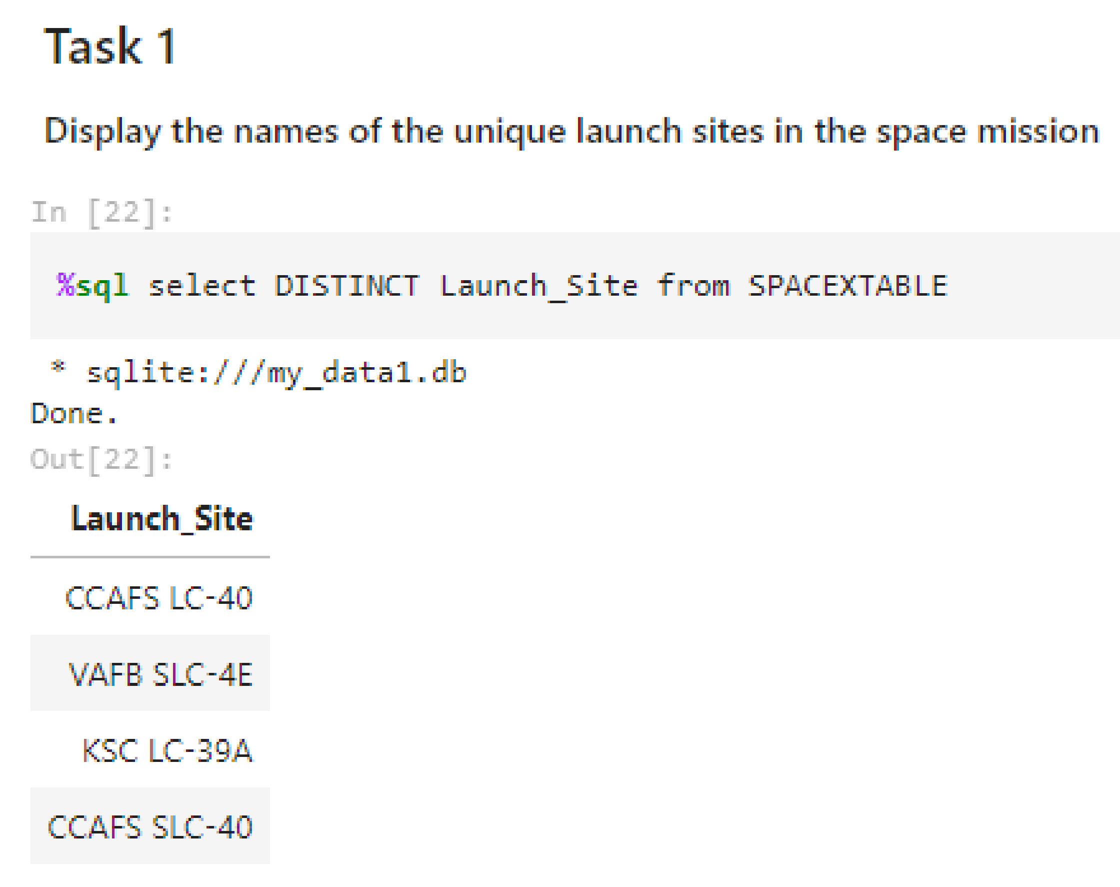


All Launch Site Names

Task Description=>

Sql query =>

Result =>



Task 1

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

```
In [22]:
%sql select DISTINCT Launch_Site from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
Out[22]:
```

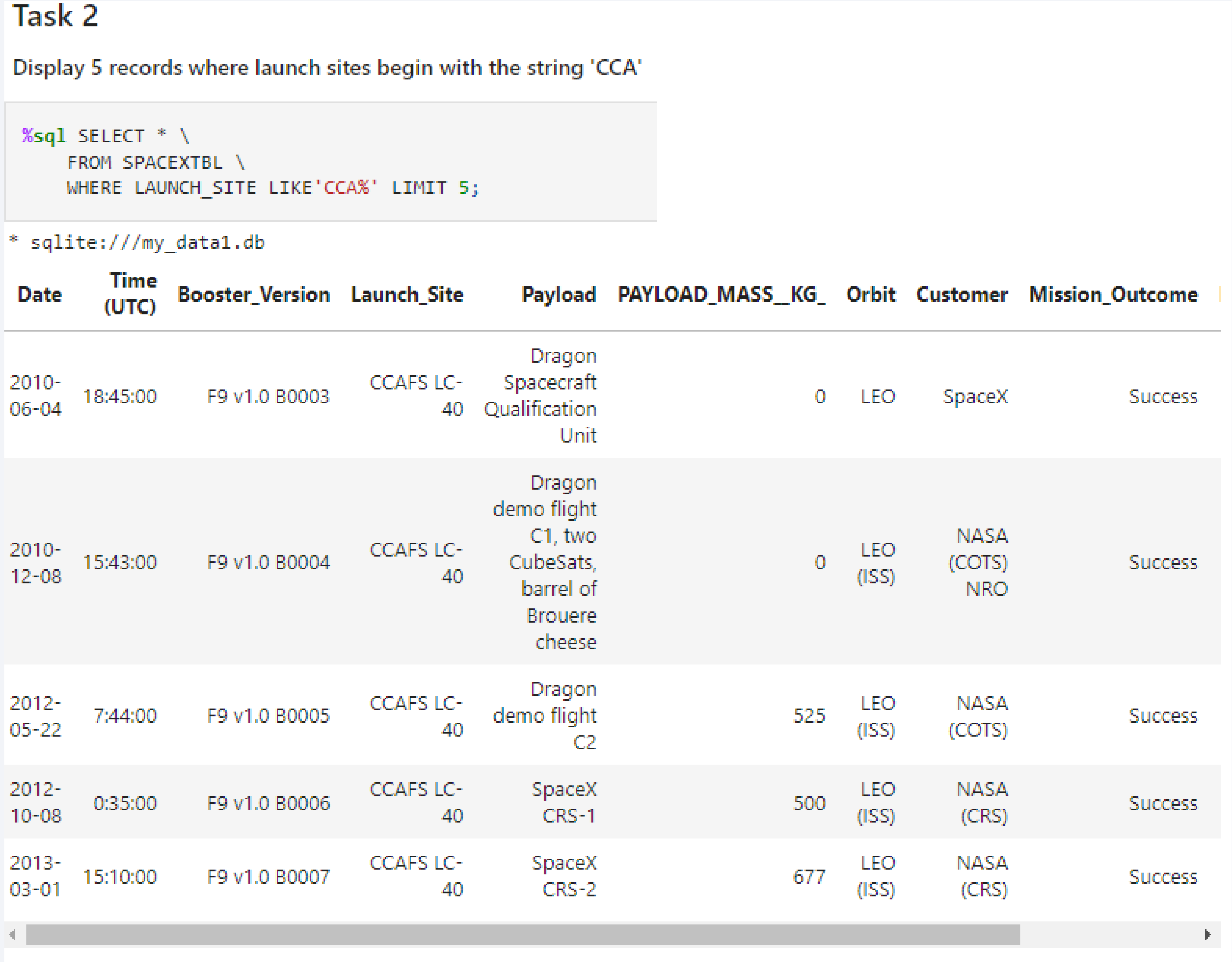
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Task Description=>

Sql query =>

Result =>



Task 2

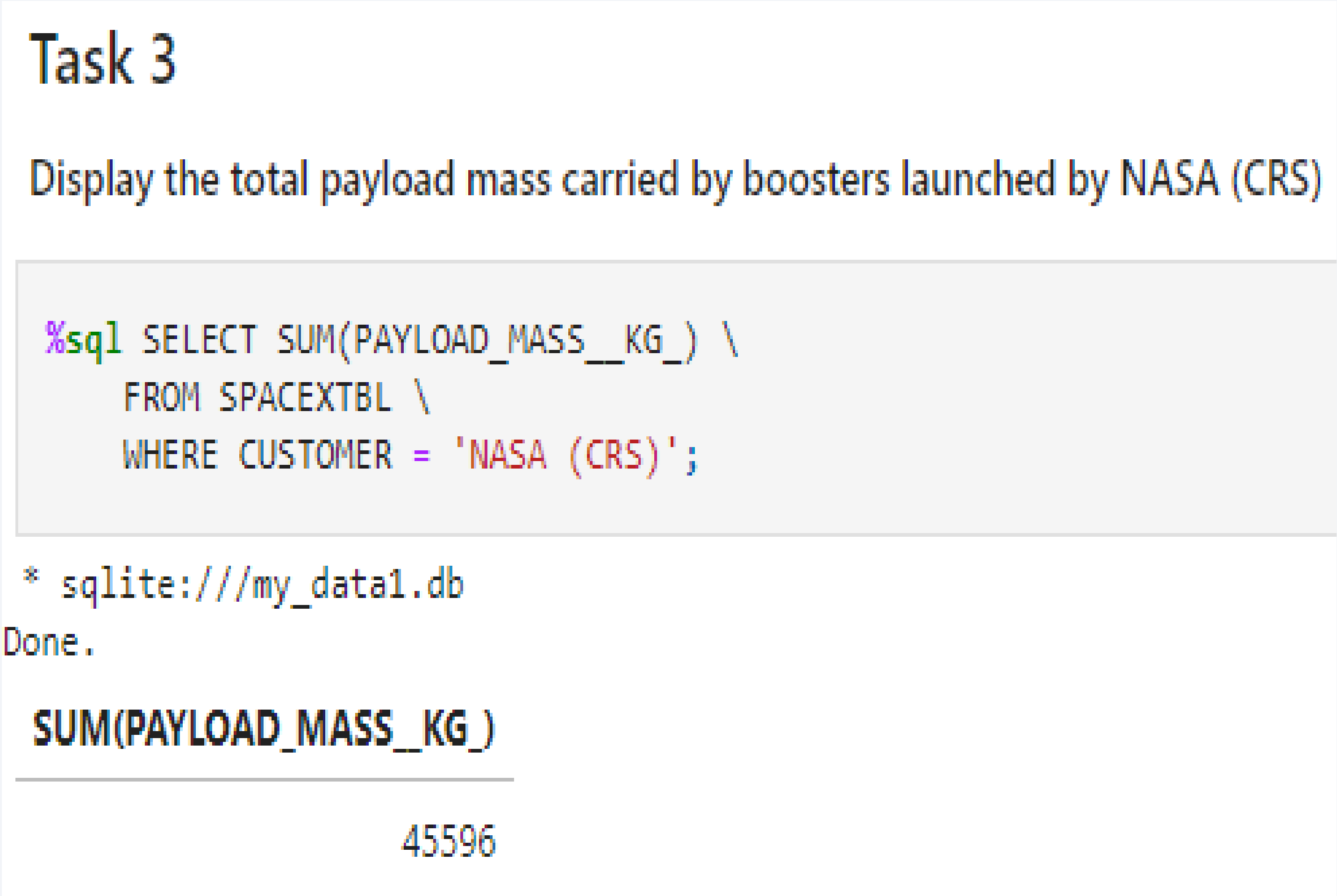
Display 5 records where launch sites begin with the string 'CCA'

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

```
* sqlite:///my_data1.db
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass



Task Description=>

Task 3

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

Sql query =>

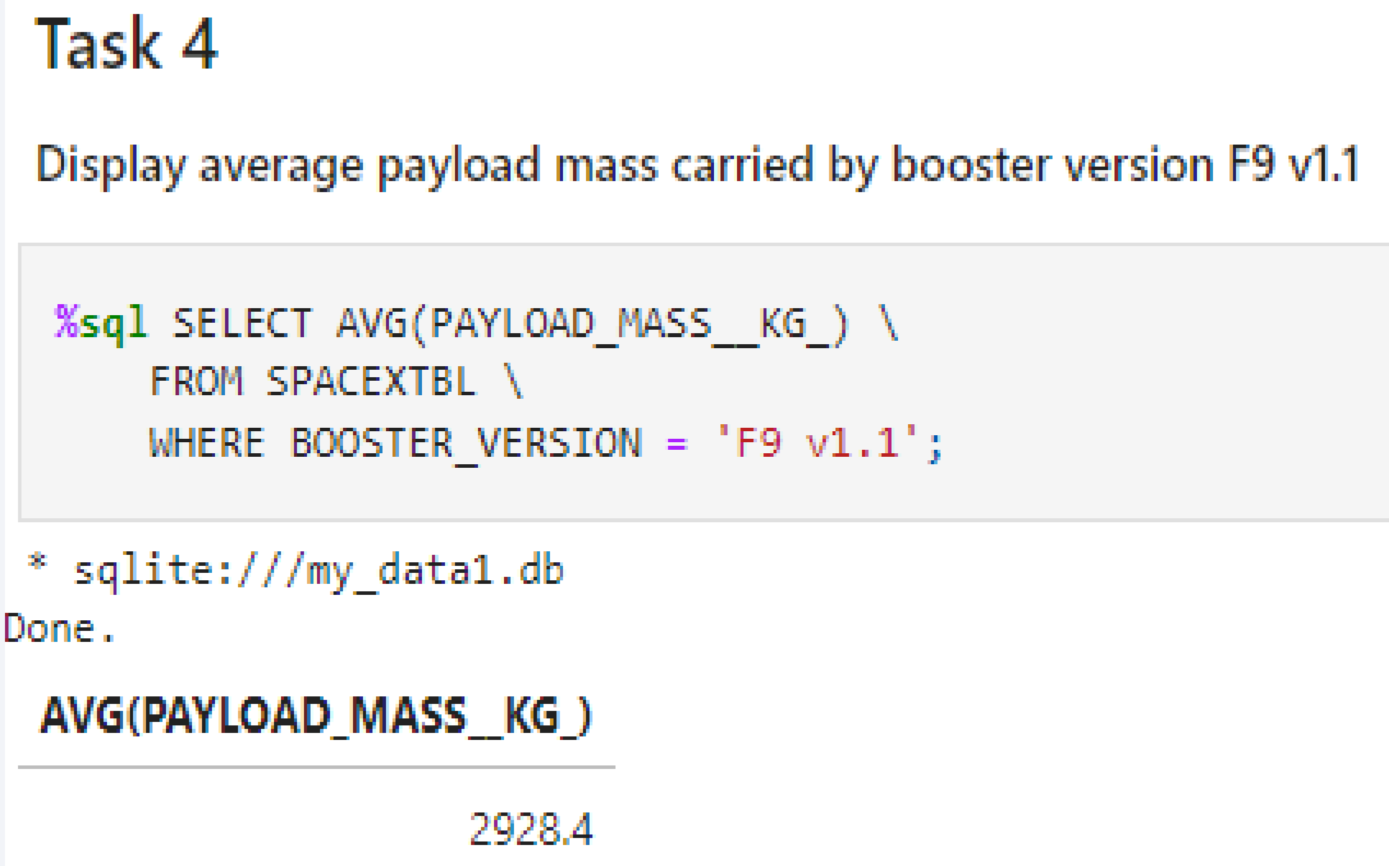
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

Result =>

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1



Task Description=>

Task 4

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

Sql query =>

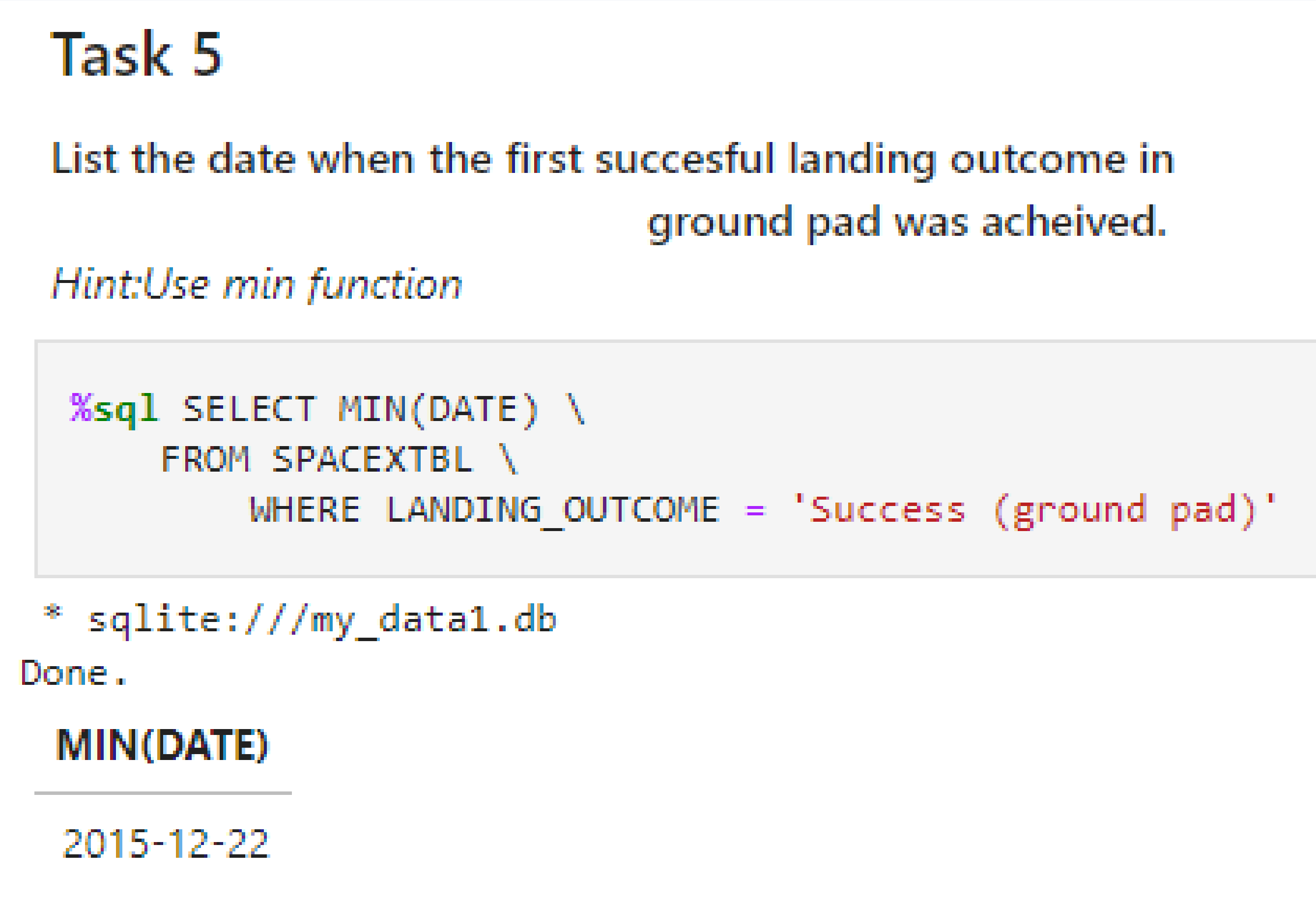
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

Result =>

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date



Task Description=>

Task 5

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

Hint:Use min function

Sql query =>

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

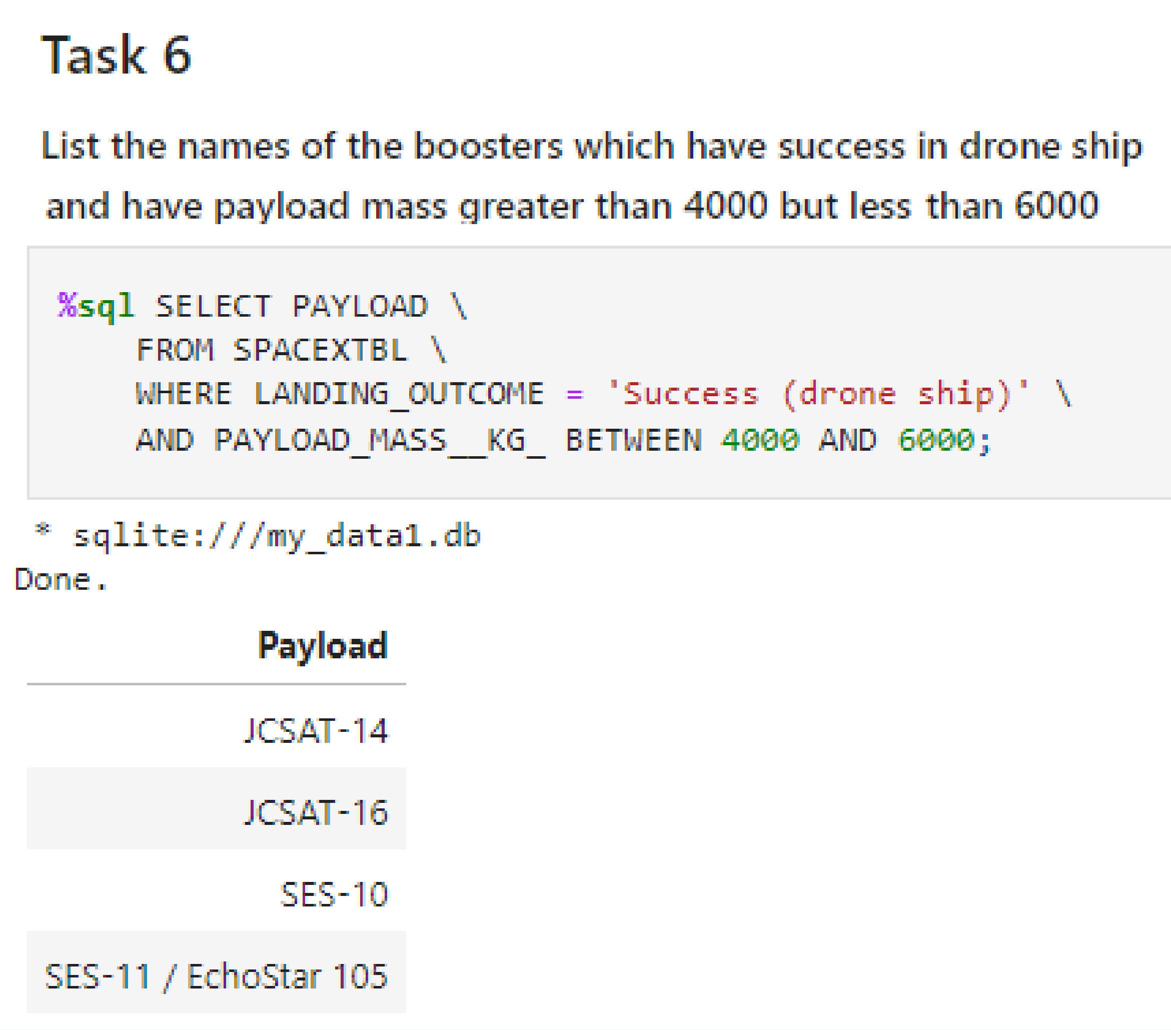
```
 * sqlite:///my_data1.db
Done.
```

Result =>

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task Description=>



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 query =>

```
%sql SELECT PAYLOAD \
    FROM SPACEXTBL \
    WHERE LANDING_OUTCOME = 'Success (drone ship)' \
    AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

Result =>

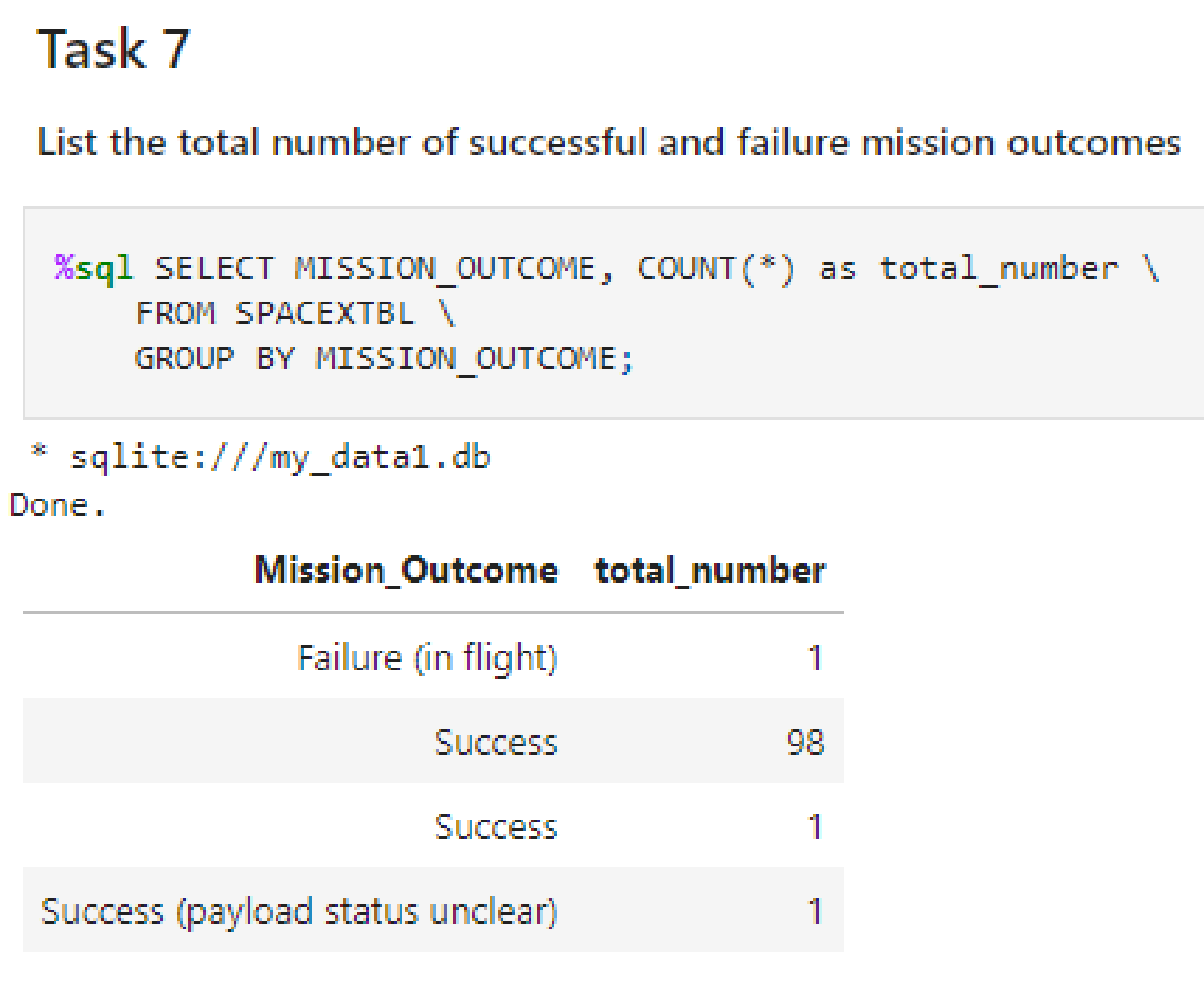
Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

Task Description=>

Sql query =>

Result =>



Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
    FROM SPACEXTBL \
    GROUP BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
```

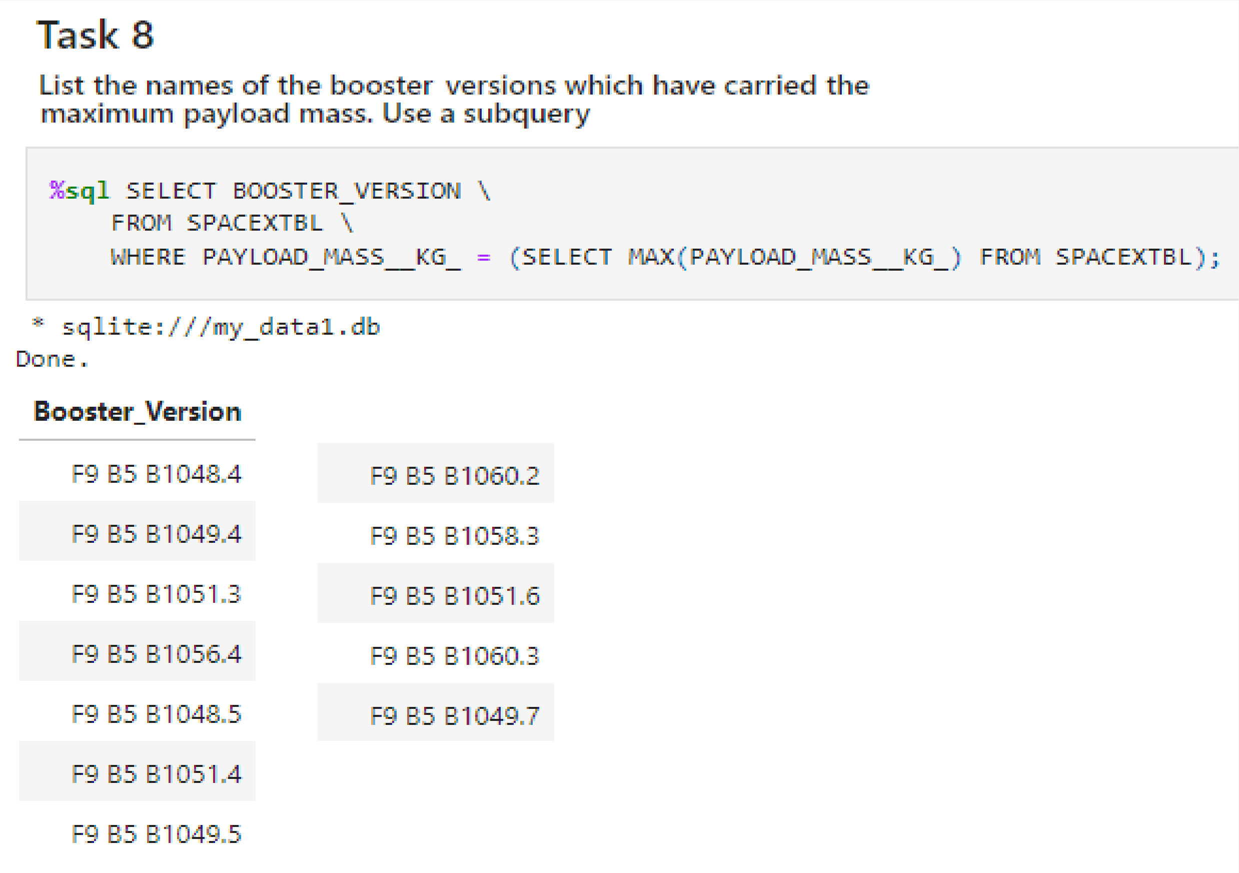
Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Task Description=>

Sql query =>

Result =>



Task 8

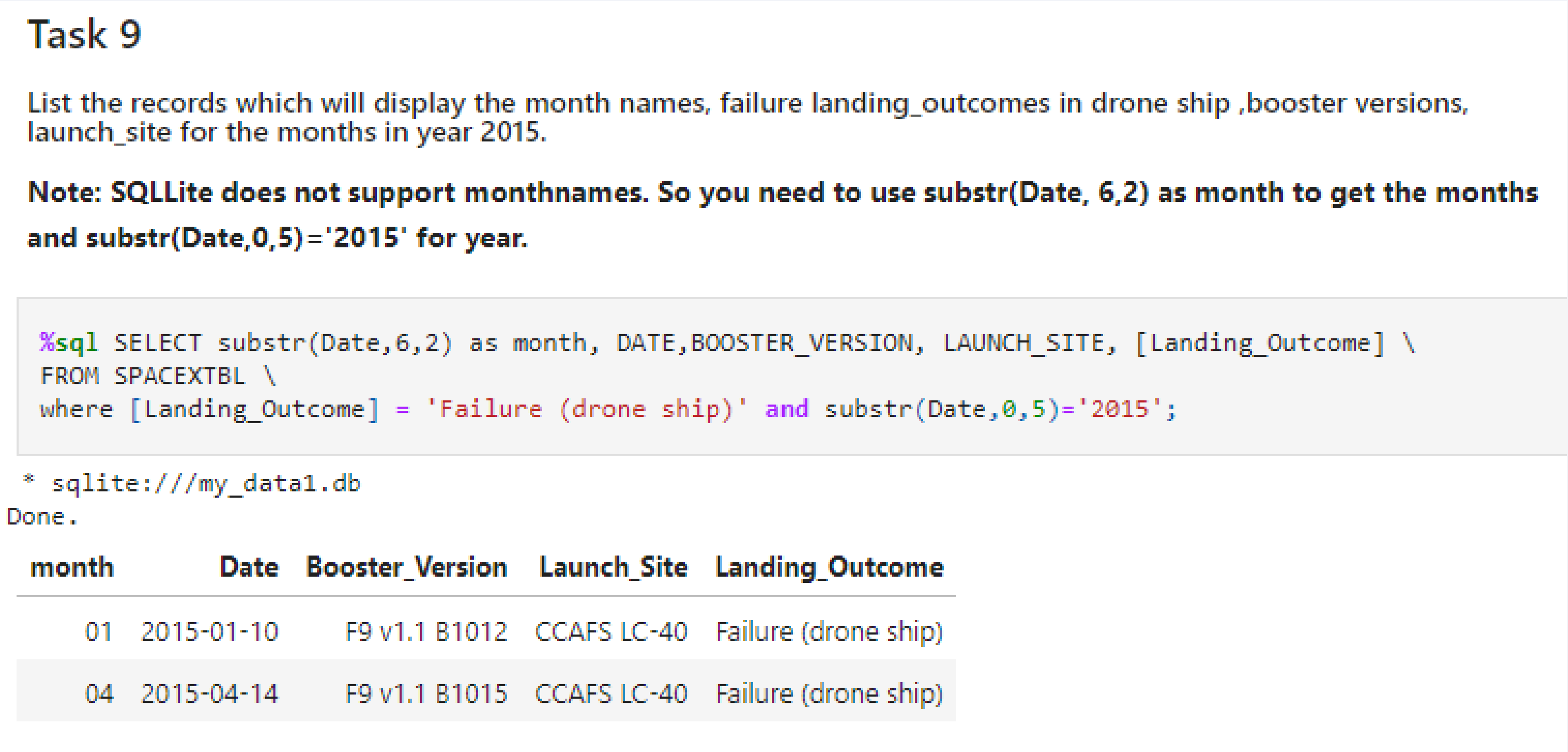
List the names of the booster versions which have carried the maximum payload mass. Use a subquery

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

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records



Task Description=>

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

Sql query =>

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

Result =>

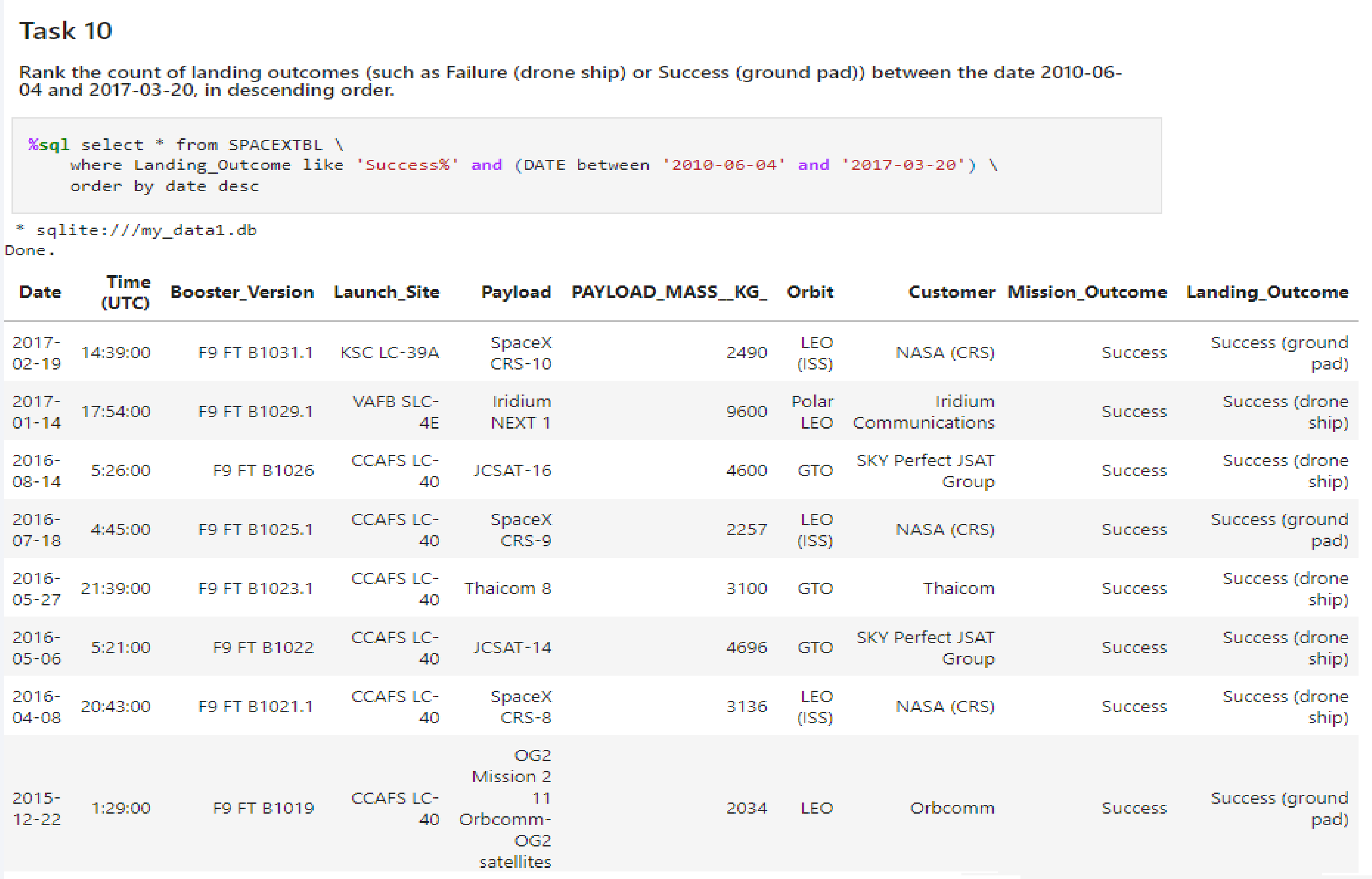
month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Task Description=>

Sql query =>

Result =>



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 * from SPACEXTBL \
    where Landing_Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') \
    order by date desc
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

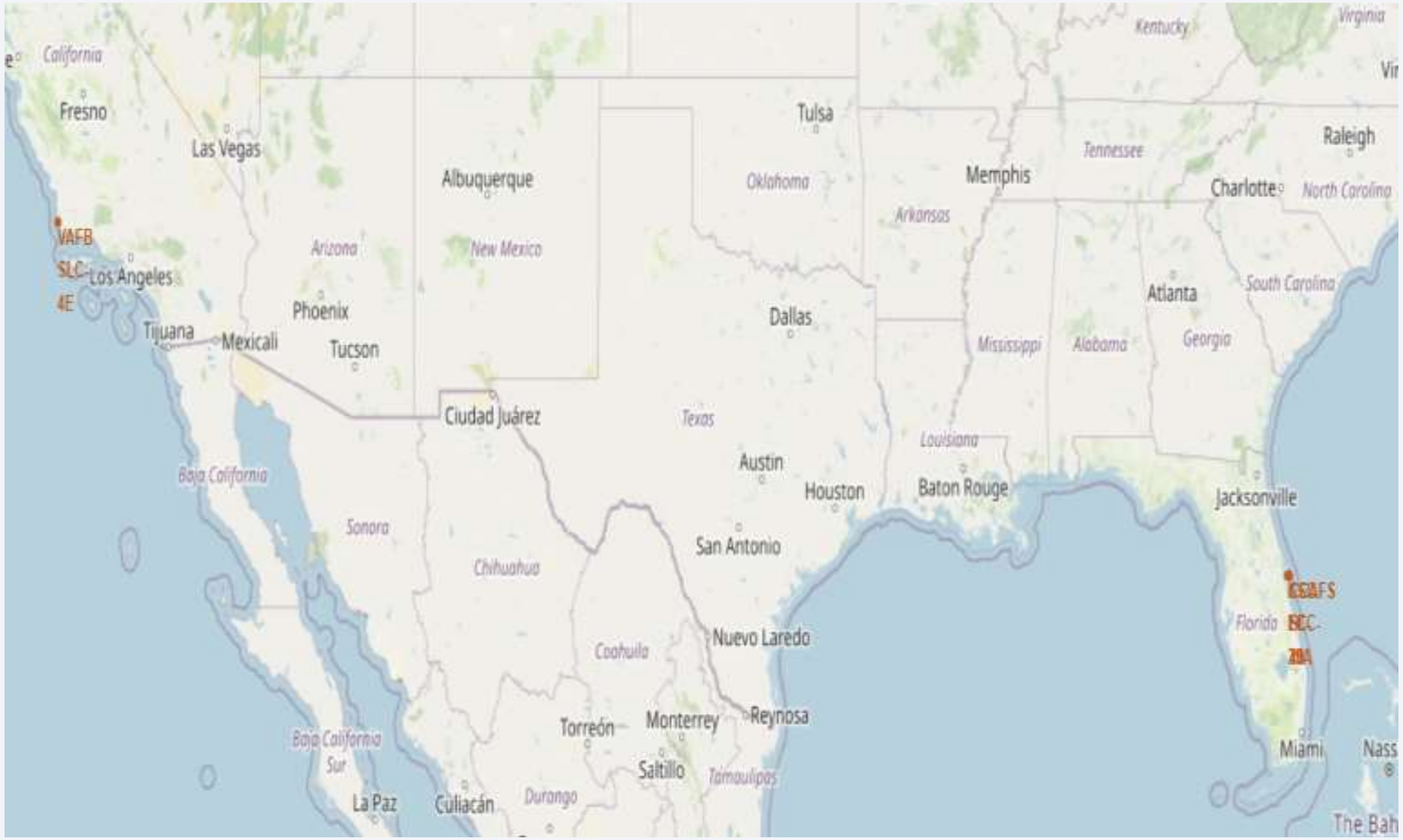
Section 3

Launch Sites Proximities Analysis

Markers – Launch Sites

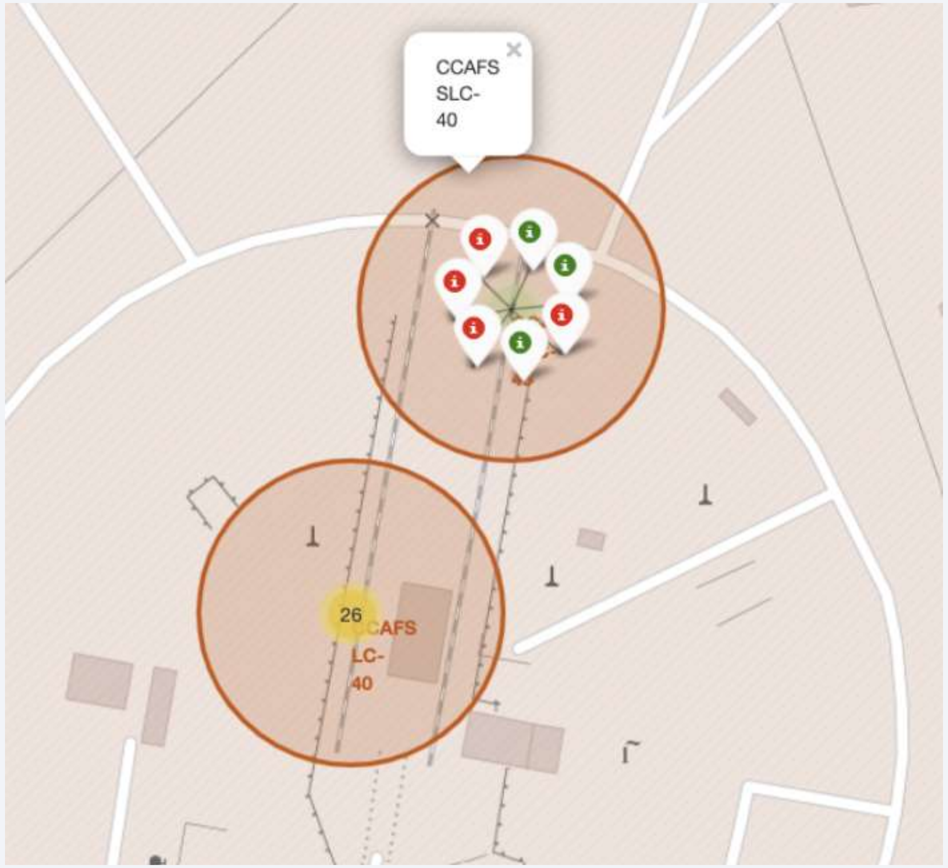
Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit.

Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.

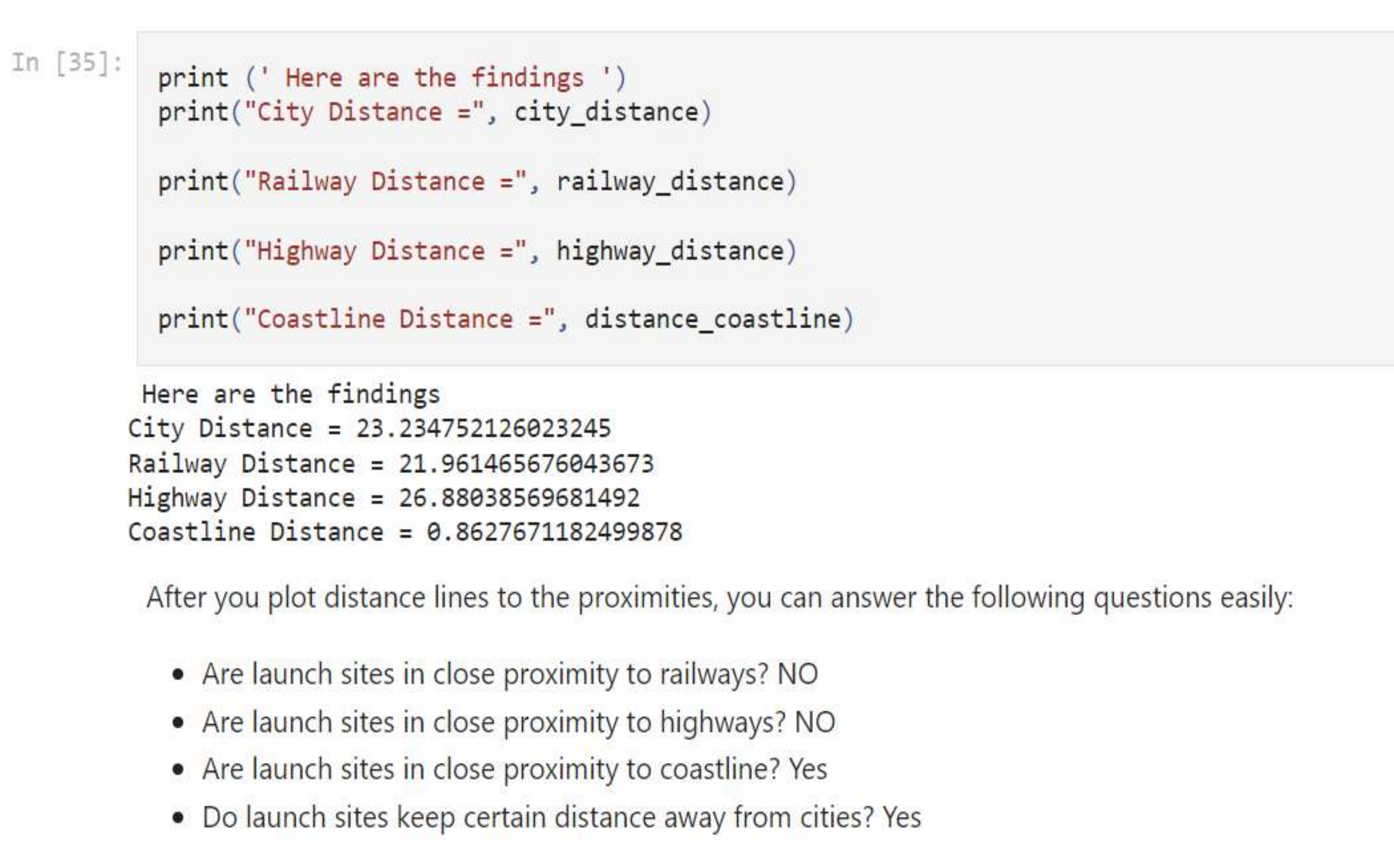


Successful and unsuccessful Launch outcomes

- Green markers for successful launches
- Red markers for unsuccessful launches



Distance Lines to the proximities and finding



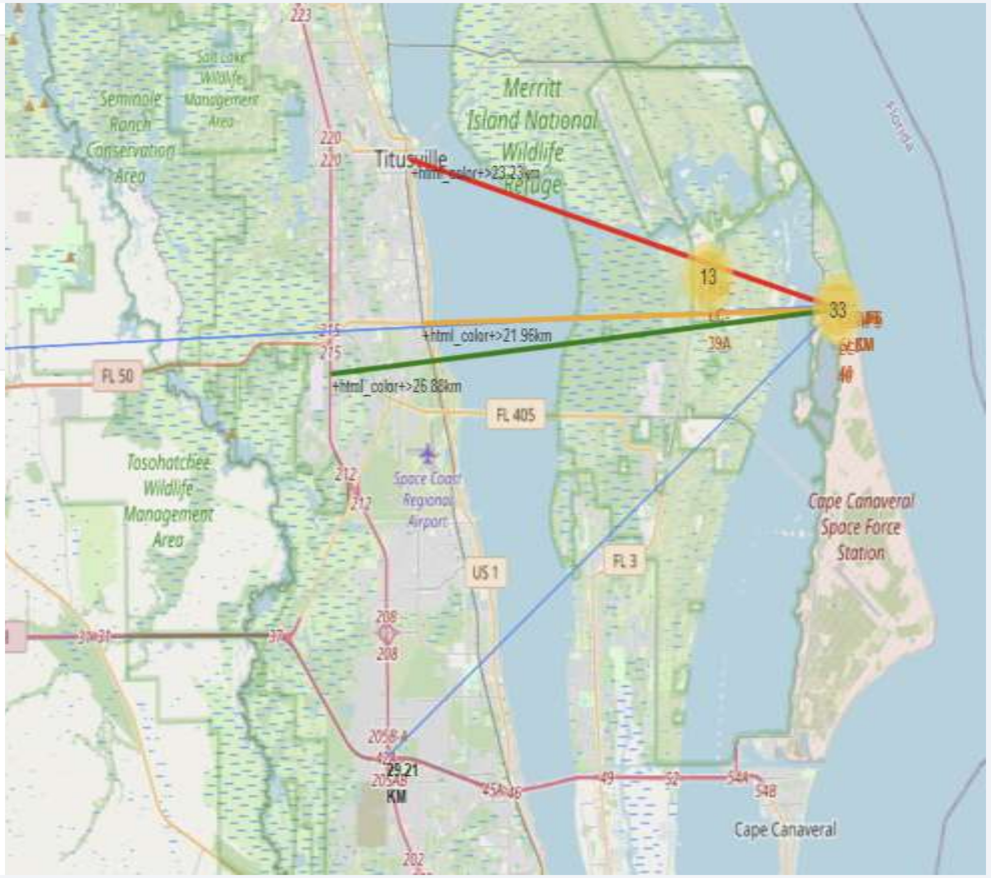
In [35]:

```
print (' Here are the findings ')
print("City Distance =", city_distance)

print("Railway Distance =", railway_distance)

print("Highway Distance =", highway_distance)

print("Coastline Distance =", distance_coastline)
```

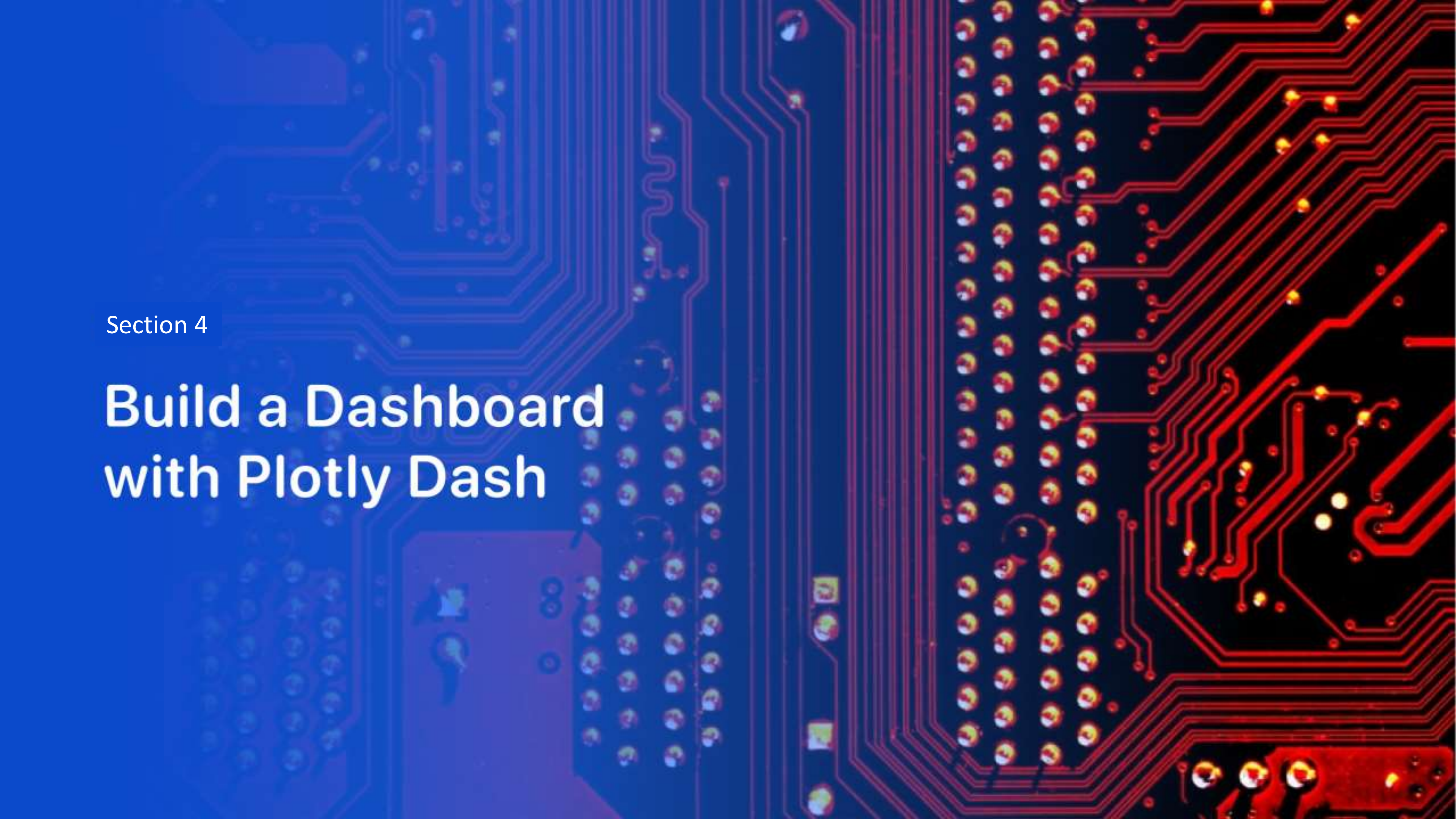
```
 Here are the findings
City Distance = 23.234752126023245
Railway Distance = 21.961465676043673
Highway Distance = 26.88038569681492
Coastline Distance = 0.8627671182499878
```

After you plot distance lines to the proximities, you can answer the following questions easily:

- Are launch sites in close proximity to railways? NO
- Are launch sites in close proximity to highways? NO
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes



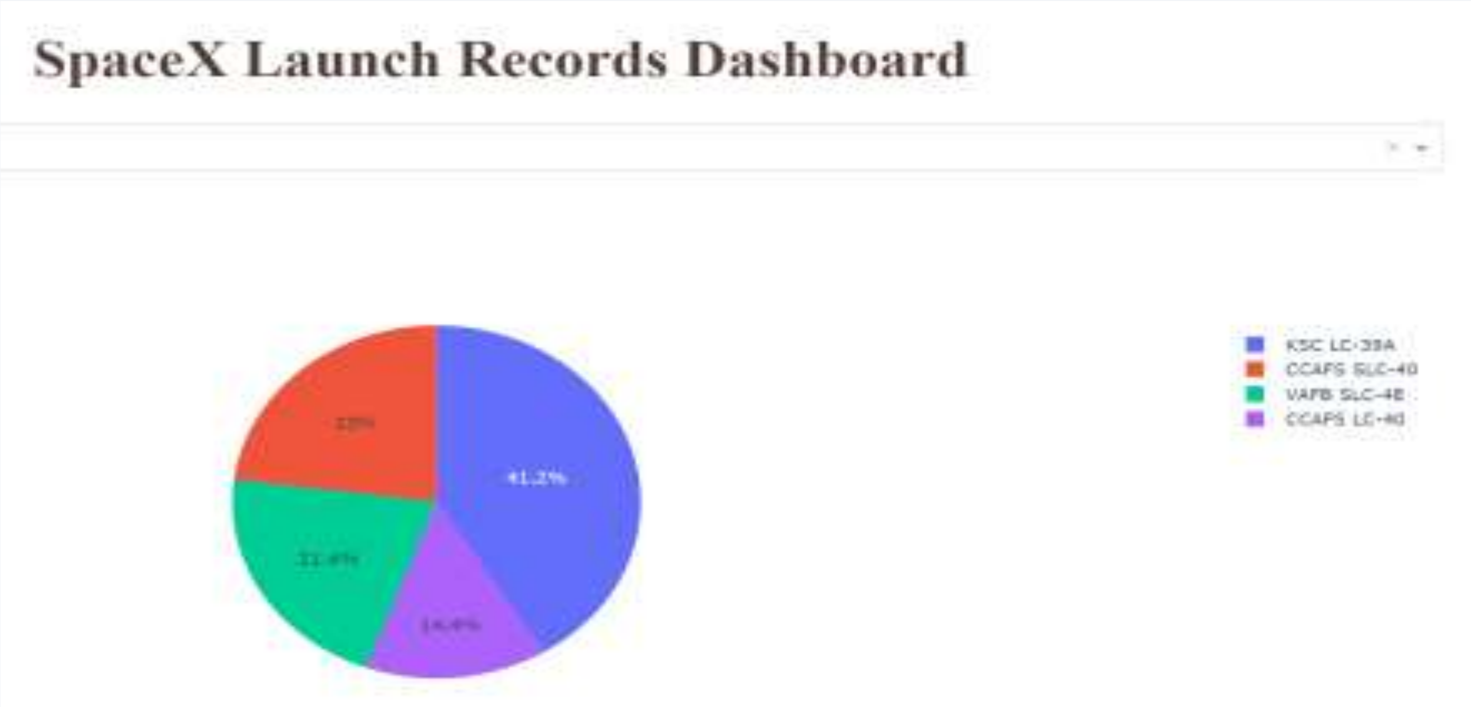
Seminole Ranch Conservation Area
Salt Lake Wildlife Management Area
Merritt Island National Wildlife Refuge
Titusville
+html_color+>23.23km
13
33
+html_color+>21.96km
+html_color+>26.88km
39A
FL 50
FL 405
Tosohatchee Wildlife Management Area
Space Coast Regional Airport
Cape Canaveral Space Force Station
US 1
FL 3
29.21 KM
Cape Canaveral



Section 4

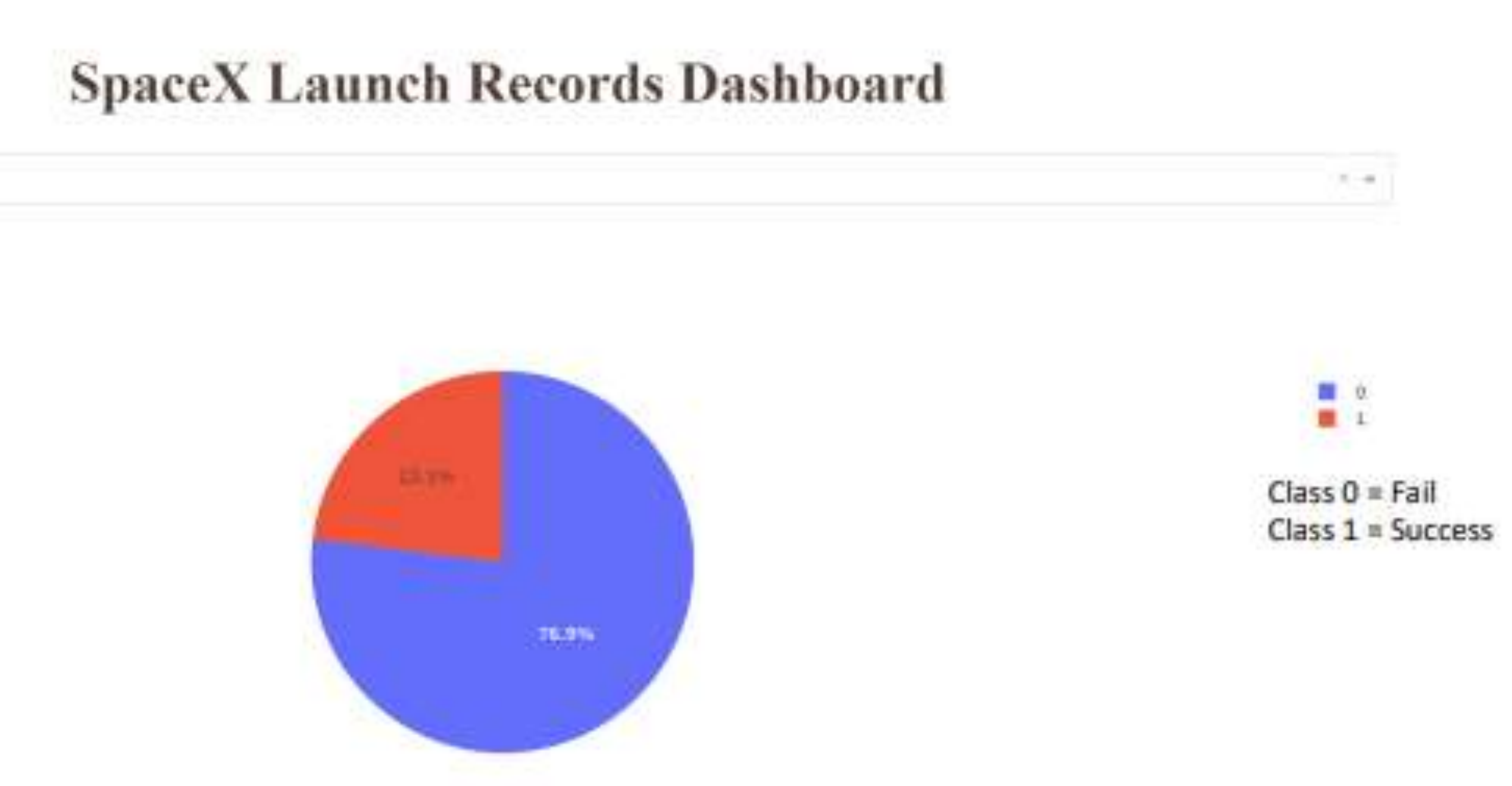
Build a Dashboard with Plotly Dash

Launch Records on Pie Chart



SpaceX Launch Records Dashboard

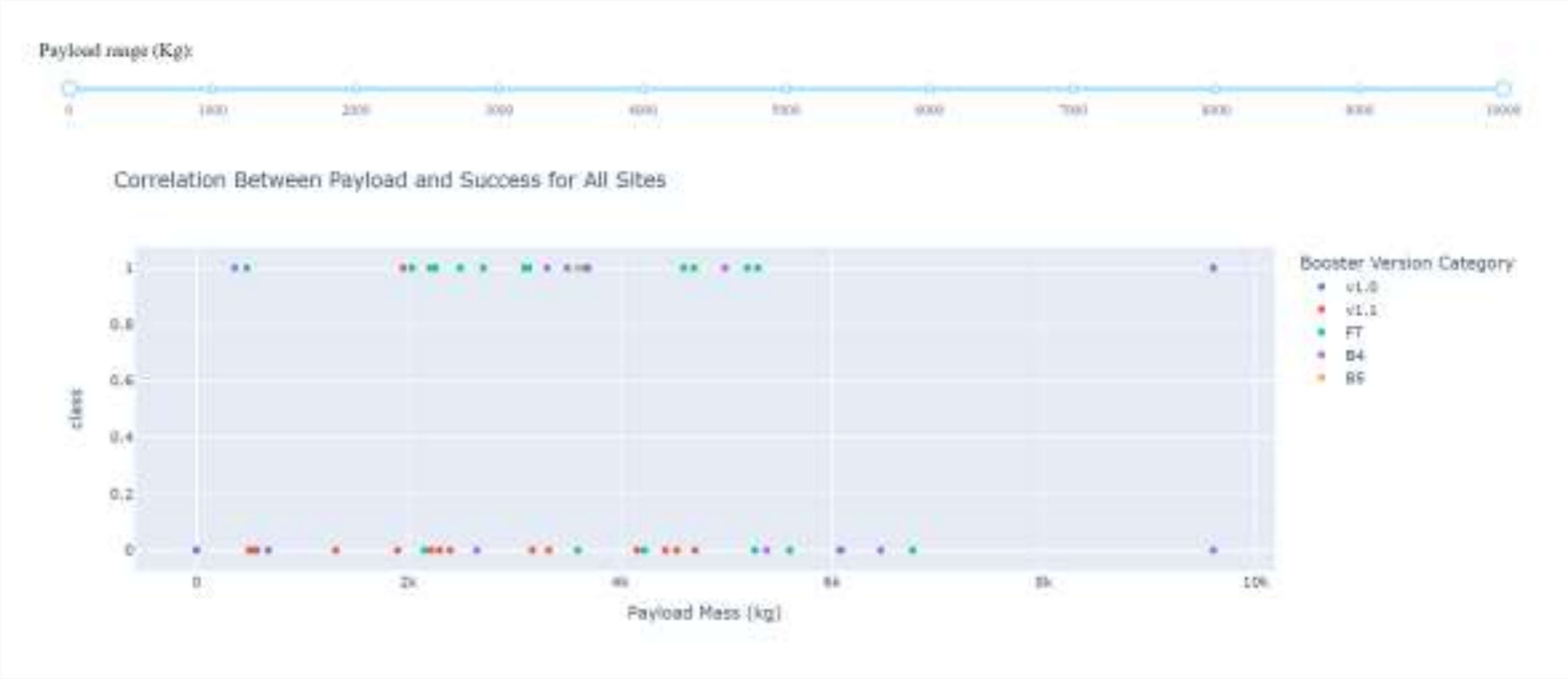
Success of KSC LC-29A



SpaceX Launch Records Dashboard

Class 0 = Fail
Class 1 = Success

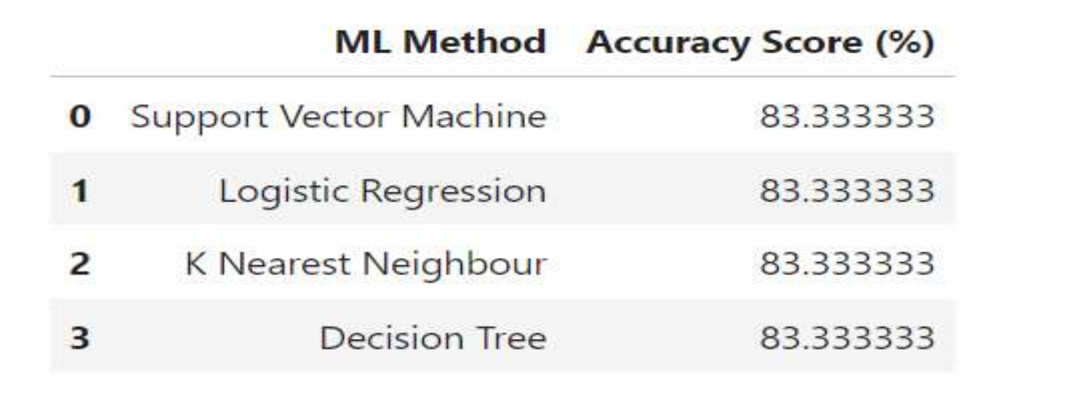
Payload Mass and Success



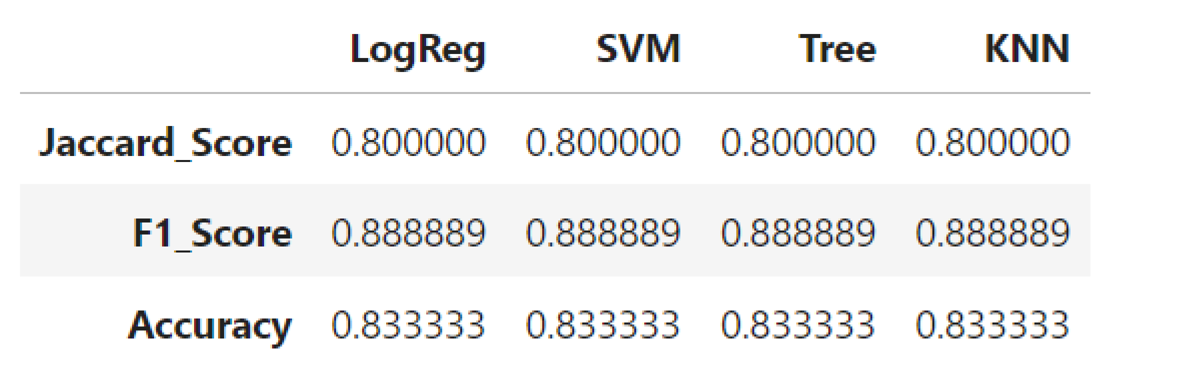
Section 5

Predictive Analysis (Classification)

Classification Accuracy



	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333



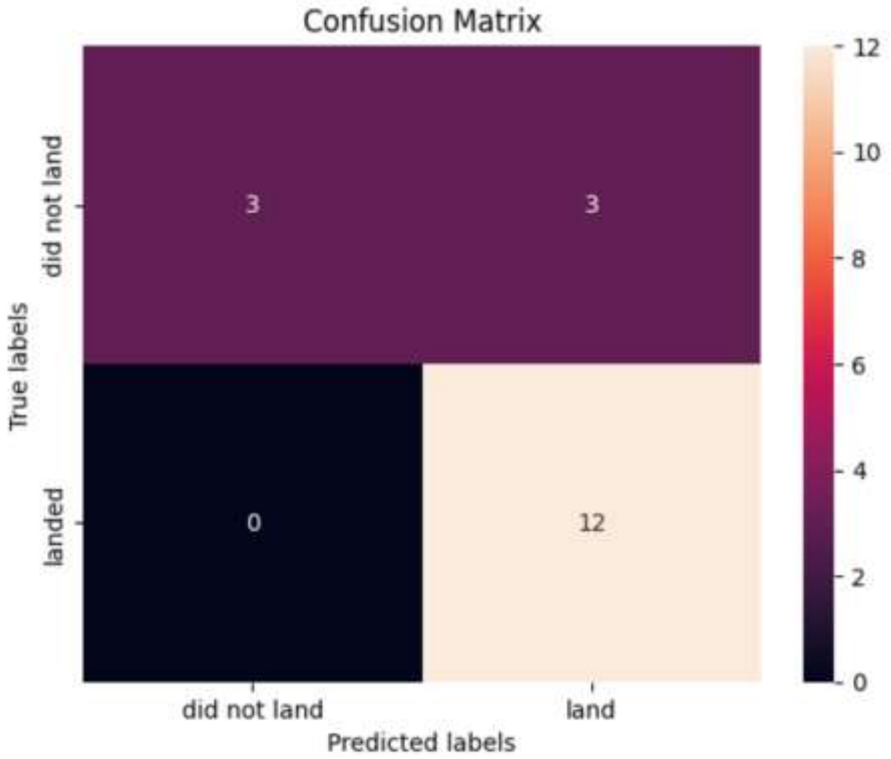
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Same accuracy score achieved

- Best model is **DecisionTree** with a score of 0.875
- Best params is : {'criterion': 'entropy', 'max_depth': 2, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}

Confusion Matrix

A confusion matrix summarizes the performance of a classification algorithm



Same results are achieved for various models under observation

Conclusions

- Distance Measurement: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters All the launch sites are close to the coast
- Launch Success: Increases over time
- KSC LC-39A:
  - Has the highest success rate among launch sites.
  - Has a 100% success rate for launches less than 5,500 kg
- Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate
- Various models are tested with similar results; but decision tree has outperformed
  - The models performed similarly on the test set with the decision tree model slightly outperforming

Thank you!