IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

PAVITHRAN. S. N

30 DEC 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **<u>Summary of methodologies:</u>**
- Data Collection using Web scraping and REST API queries
- Data Wrangling to Classify Launches based on Success and transform data into standardized numeric form
- Exploratory Data Analysis using SQL and Visualization packages for Python
- Interactive Plotly Web App to visualize payload and success launch data at each Launch Site
- Exploring Launch Sites using interactive Folium Maps
- Predictive analysis for classification of Rocket Landing Success
- **<u>Summary of all results:</u>**
- Exploratory Data Analysis Results
- Predictive Analysis Results

Introduction

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. Thus it is advantageous to be able to predict whether the Falcon 9 first stage will land successfully for new missions.

To make valuable predictions we must solve the following:

- What factors of a mission influence Falcon 9 launch success?
- What conditions must be met by SpaceX to ensure the highest probability of success for a given mission?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using Space X API and web scraping from wikipedia.
- Perform data wrangling
  - Clean the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Creating various Machine learning models and validate by score.

Data Collection

The Data sets are collected by

- SpaceX API request.
- Web Scraping

Enter the URL of the page you want to analyze for this project

Request and parse the SpaceX launch data using the GET request

decode the response content as a JSON and turn it into a Pandas data frame

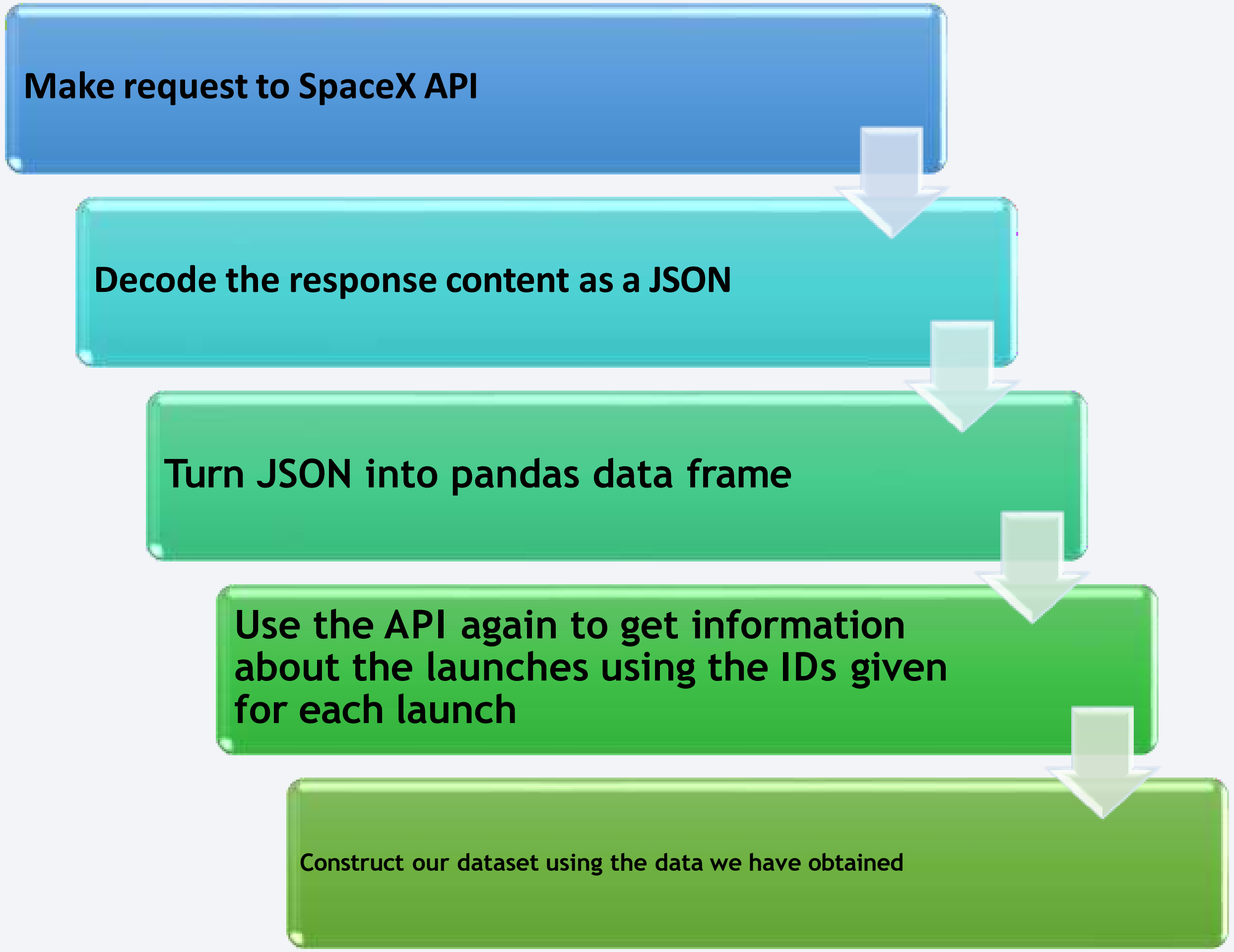
now use the API again to get information about the launches using the IDs given for each launch

Filter the data frame to only include Falcon 9 launches and replace null values and get required output

Data Collection – SpaceX API

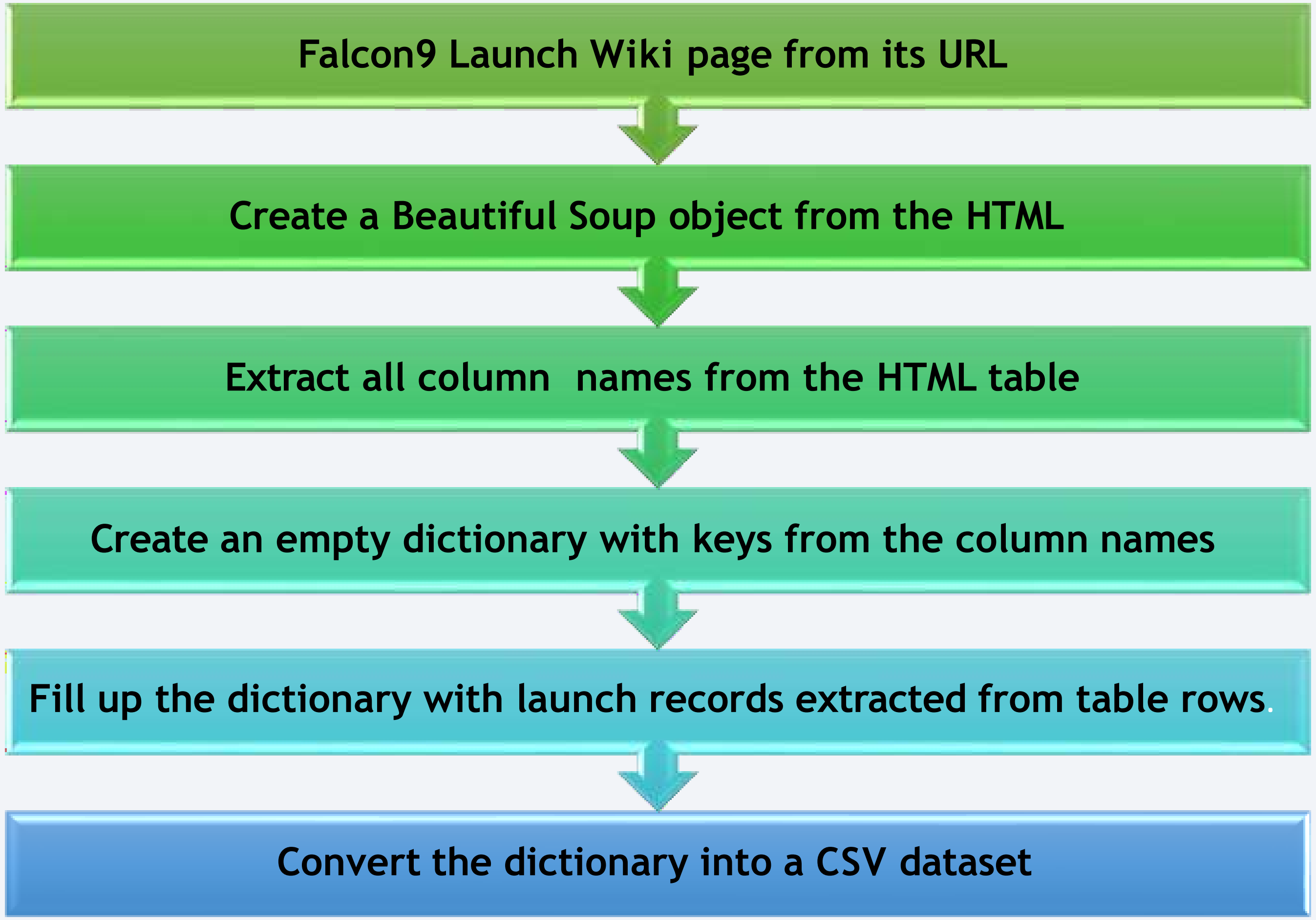
Steps

- ➢ Request data from SpaceX API (rocket launch data)
- ➢ Decode response using .json() and convert to a dataframe using .json_normalize()
- ➢ Request information about the launches from SpaceX API using custom functions
- ➢ Create dictionary from the data
- ➢ Create dataframe from the dictionary
- ➢ Filter dataframe to contain only Falcon 9 launches
- ➢ Replace missing values of Payload Mass with calculated .mean()
- ➢ Export data to csv file



Data Collection - Web Scraping

- **Steps**
  - Request data (Falcon 9 launch data) from Wikipedia
  - Create BeautifulSoup object from HTML response
  - Extract column names from HTML table header
  - Collect data from parsing HTML tables
  - Create dictionary from the data
  - Create dataframe from the dictionary
  - Export data to csv file



Data Wrangling

➢Data Wrangling process is given in a flow chart for a over view.

➢**steps**

➢ Perform EDA and determine

➢data labels

➢ Calculate:

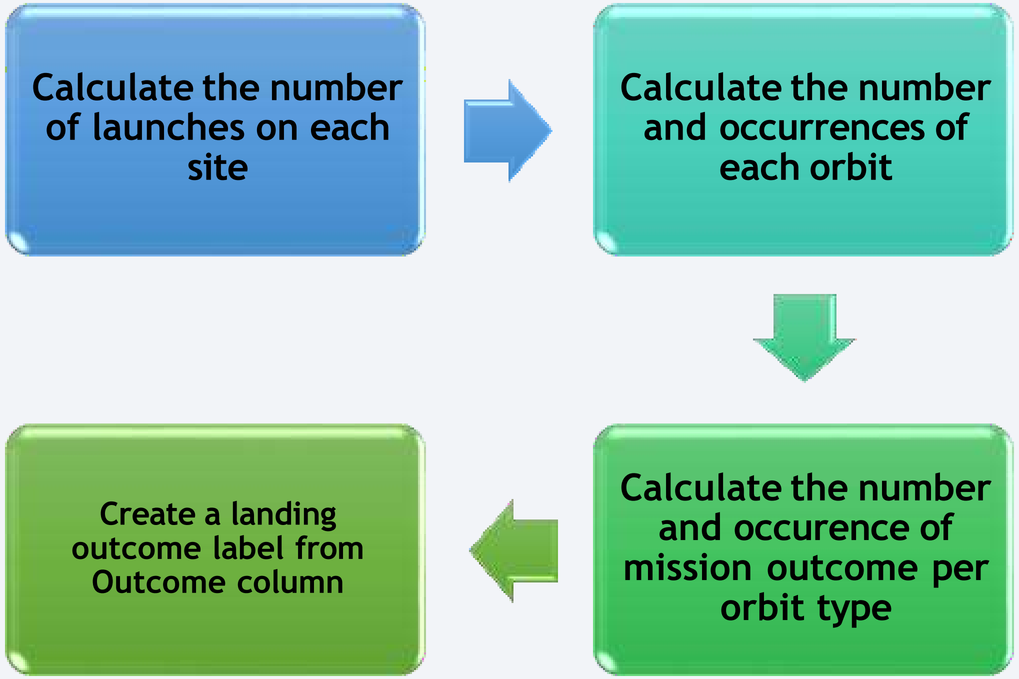
➢ # of launches for each site

➢ # and occurrence of orbit

➢ # and occurrence of mission outcome per orbit type]

➢ Create binary landing outcome column (dependent variable)

➢ Export data to csv file



EDA with Data Visualization

Types of Charts Used :

- **<u>scatter plot</u>** -   Flight Number vs Payload Mass , Flight Number vs Launch Sites , Payload and Launch Sites , Flight Number and Orbit Type , Payload and Orbit Type
- **<u>Bar chart</u>** – Success rate of each orbit
- **<u>Line plot</u>** – success rate and Date

EDA with SQL

Summary of SQL queries that were used:

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

Build an Interactive Map with Folium

- Folium Markers were used to show the Space X launch sites and their nearest important landmarks like railways, highways, cities and coastlines.
- Polylines were used to connect the launch sites to their nearest land marks.
- Red represents rocket launch Failures
- Green represents the successes.

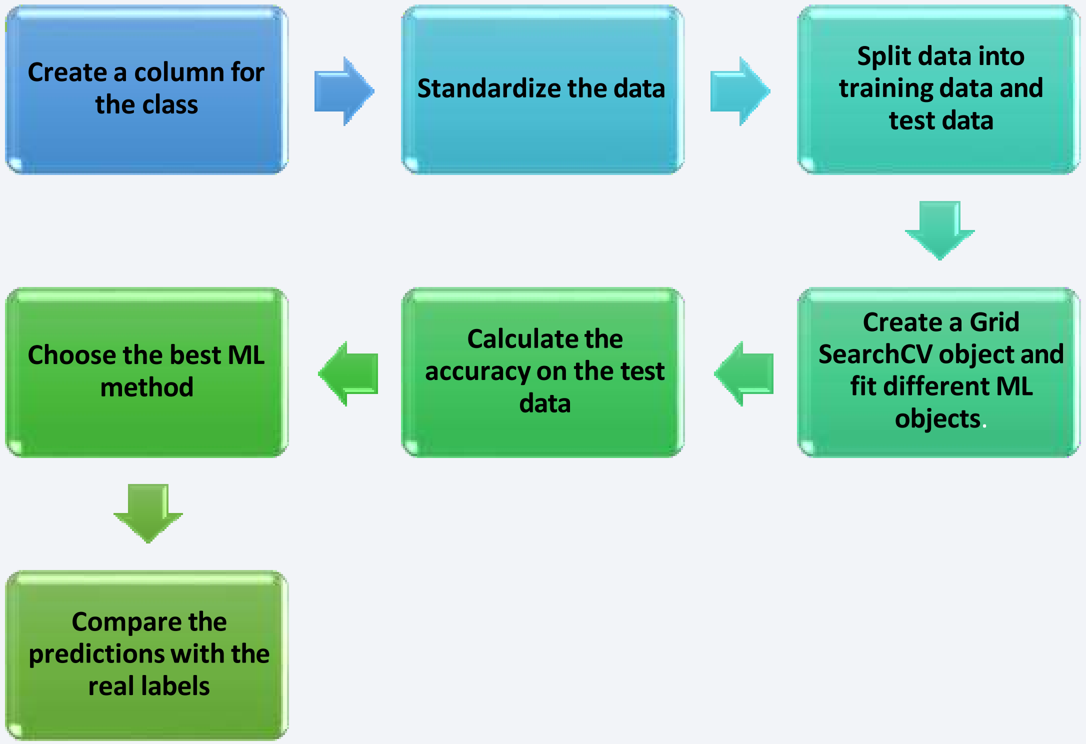
Build a Dashboard with Plotly Dash

- Pie charts and scatter charts were used to visualize the launch records of Space X.
- These charts displayed the rocket launch success rate per launch site. We are were able to get an understanding of the factors that may have been influencing the success rate at each site. Such as the payload mass and booster versions.
- Successful launches were represented by 1 while failures were represented by 0.

Predictive Analysis (Classification)

Scikit-learn is Machine Learning library that was used for predictive analysis. The following took place:

- Created a machine learning pipeline to predict if the first stage will land given the data.



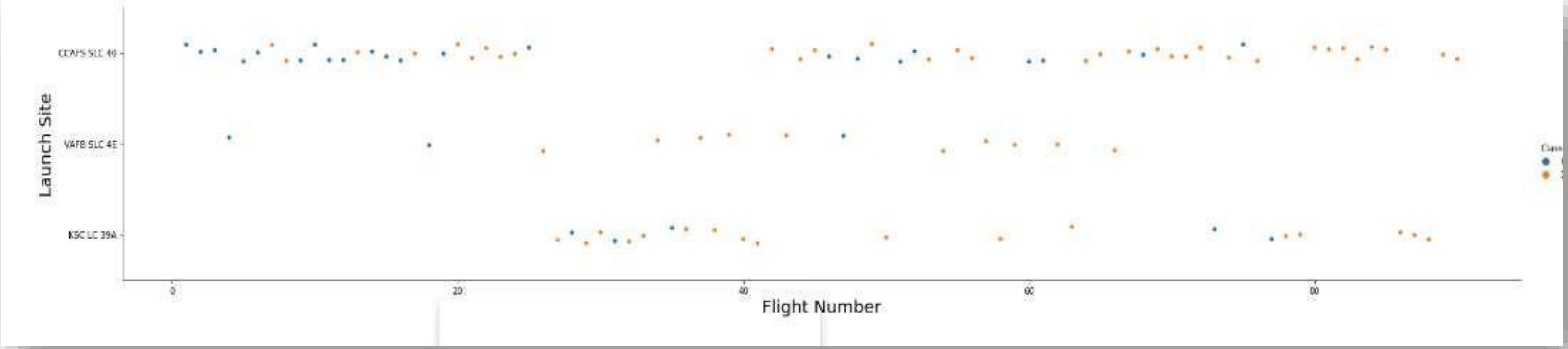
Results

- The exploratory data analysis has shown us that successful landing outcomes are somewhat correlated with flight number. It was also apparent that successful landing outcomes have had a significant increase since the year 2015.
- All launch sites are located near the coast line. Perhaps, this makes it easier to test rocket landings in the water.
- sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- The machine learning were able to predict the landing success of rockets with an accuracy score of 83.33%.

Section 2

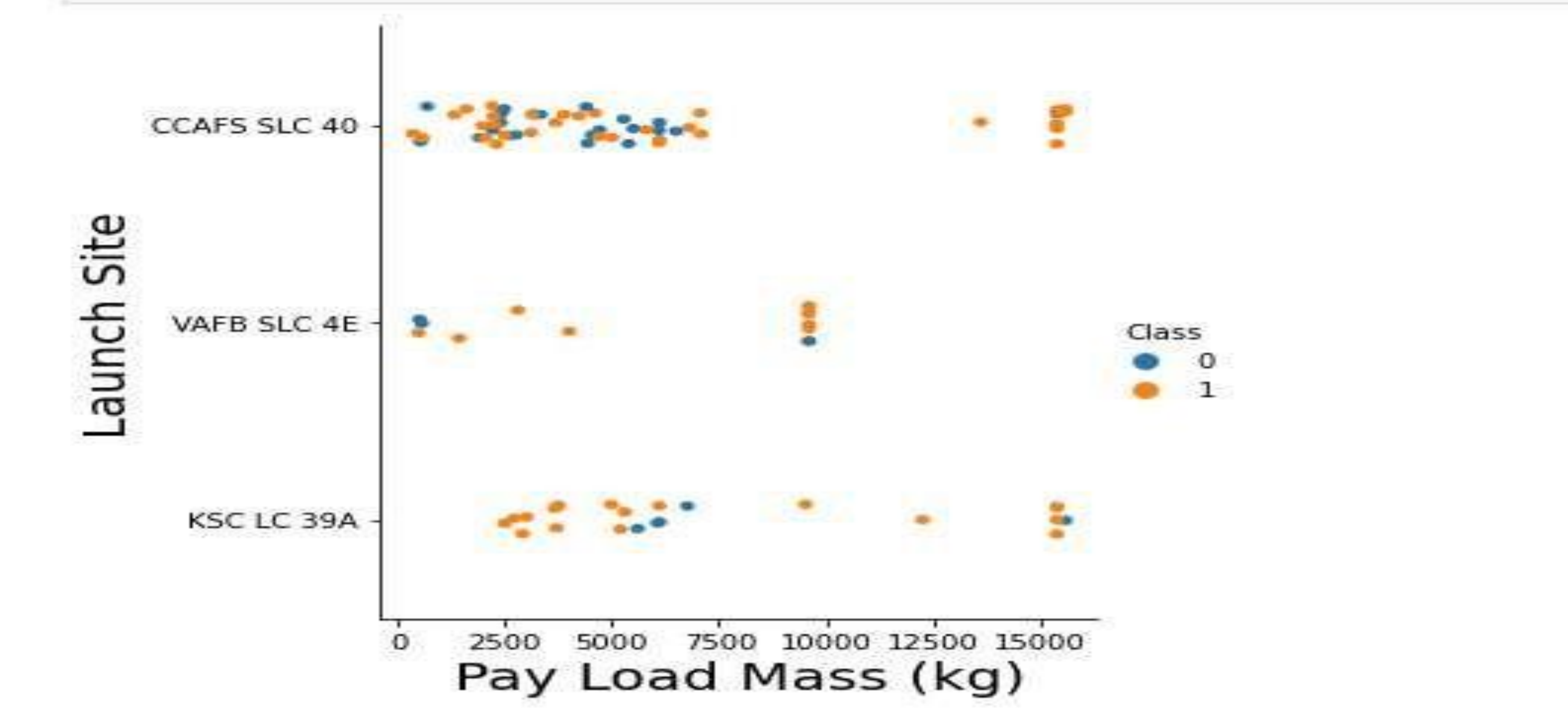
Insights drawn from EDA

Flight Number vs. Launch Site



- It appears that there were more successful landings as the flight numbers increased. launch site **CCAFS SLC 40** had the most number of landing.

Payload vs. Launch Site

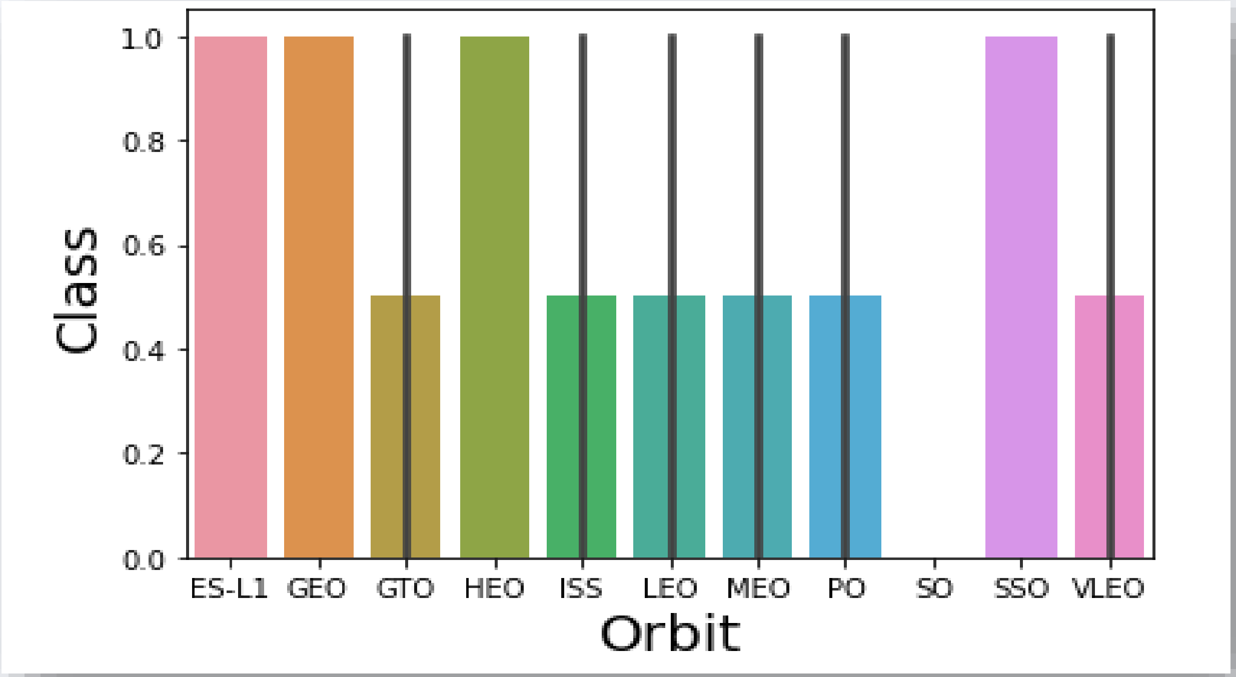


- Now if you observe the scatter point chart, you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

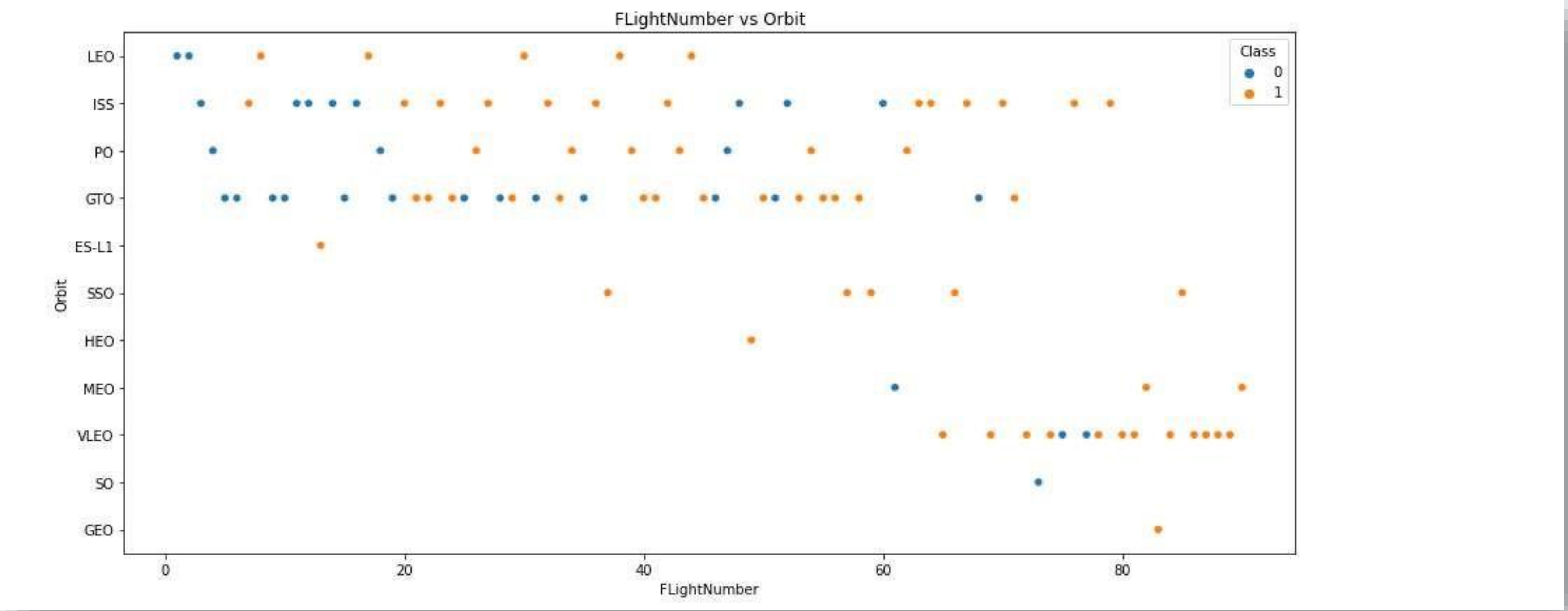
Success Rate vs. Orbit Type

The highest success rate ORBITS are

ES-L1 GEO SSO HEO

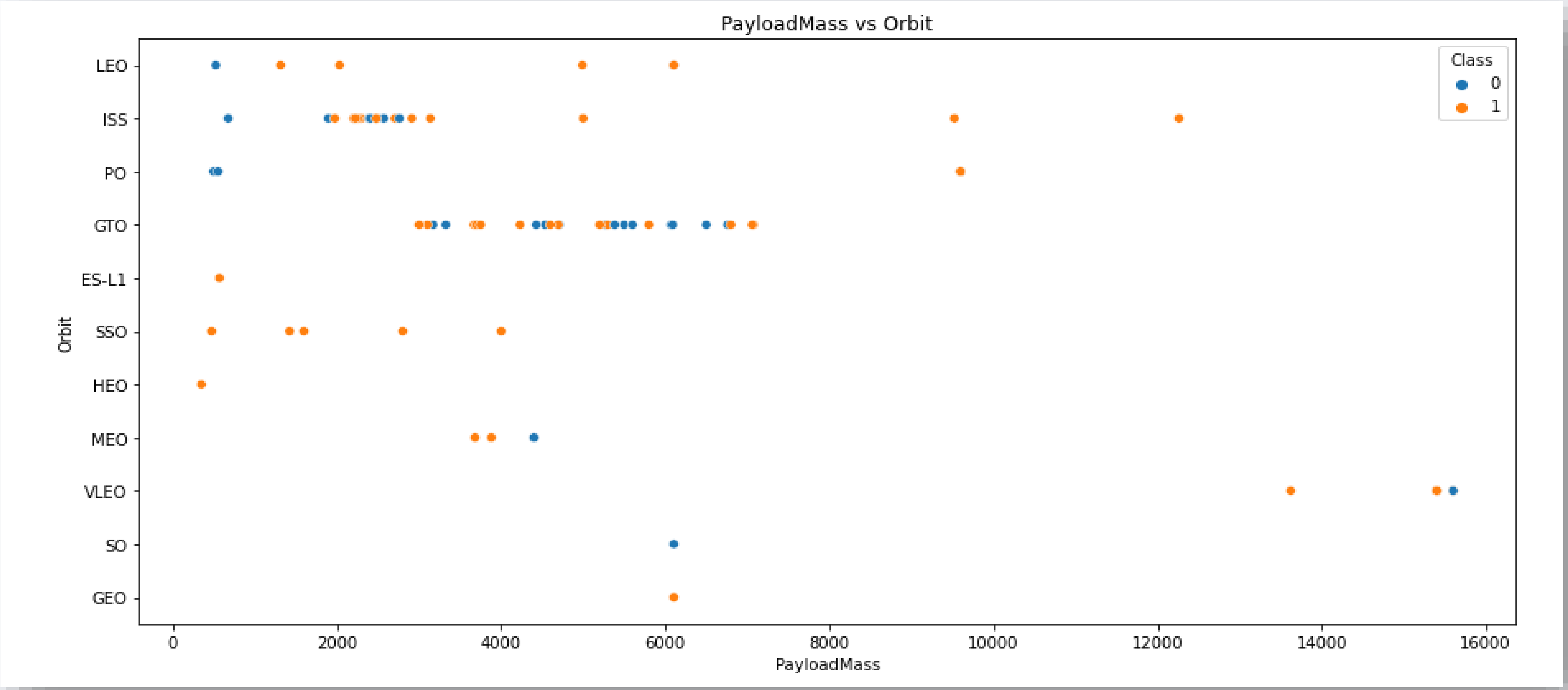


Flight Number vs. Orbit Type



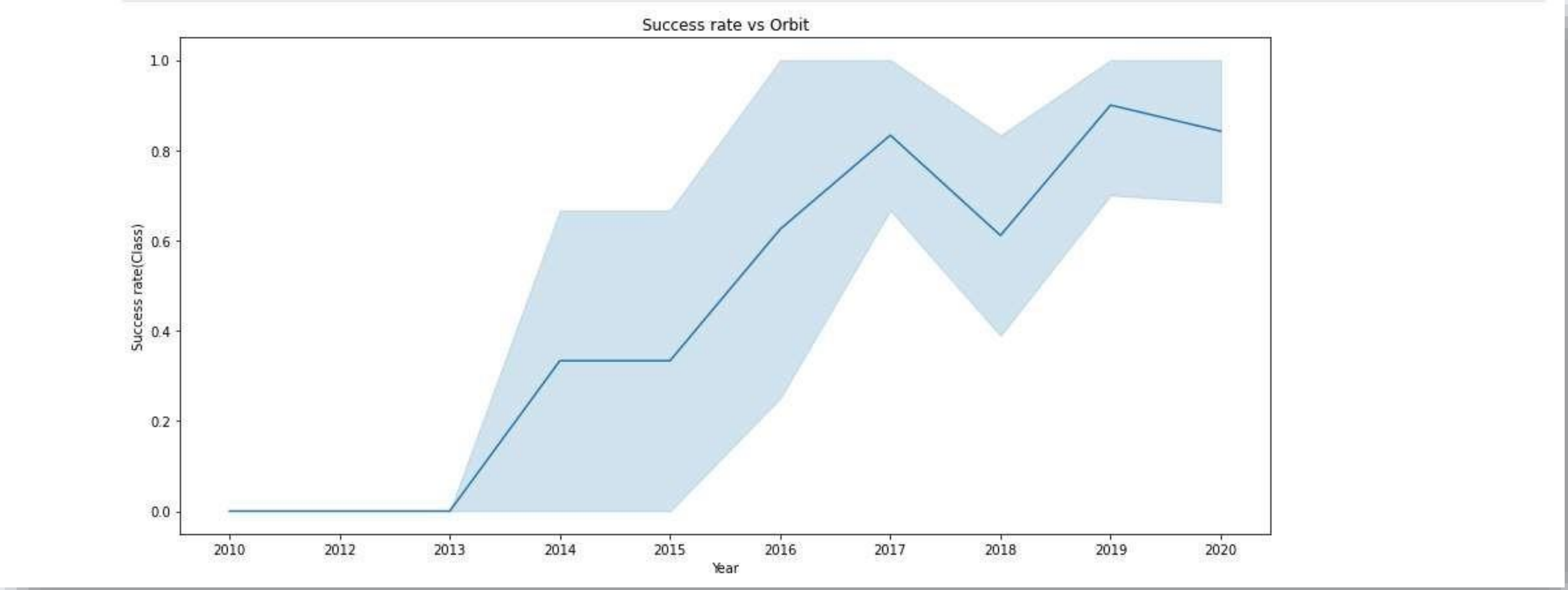
You can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there.

Launch Success Yearly Trend



It is apparent that the success rate has significantly increased from 2013 to 2020.

All Launch Site Names

Given the data, these are the names of the launch sites where different rocket landings where attempted:

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

Launch Site Names Beginning with 'CCA'

In [18]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[18]:

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

These are 5 records where launch sites begin with the letters 'CCA'. As we can see, there are other organizations besides Space X that were testing their rockets.

Total Payload Mass

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

In [23]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA(CRS)';
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[23]:

1

- The information in the picture displays the total payload mass carried by boosters launched by NASA

Average Payload Mass by F9 v1.1

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

In [24]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[24]:

1
2928

- The average payload mass carried by F9 v1.1 was 2928.4 kg.

First Successful Ground Landing Date

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

Hint:Use min function

In [28]:
```
%sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[28]:

1
2015-12-22

- From the picture given above you can see that the first successful ground pad was in 22 December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [32]:

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

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[32]:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- It appears that there only 4 Boosters with a payload mass between 4000 and 6000 they are
- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [33]:
```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[33]:

1
100

- The Above picture show the total number of successful and failure mission outcomes

Boosters That Carried the Maximum Payload Mass

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

In [34]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[34]:

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

- From the above picture it shows that 12 boosters have carried the maximum payload mass of 15600 kg.

2015 Launch Records - Failed Landing Outcomes

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)

- 2 boosters **F9 v1.1B1012_CCAFS LC-40** and **F9v1.1B1015 CCAFS LC-40** failed to land at 2015

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

In [42]:
```
%sql select * from SPACEXTBL where Landing__Outcome = 'Success (ground pad)' or and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

```
 * ibm_db_sa://gfd86828:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

Out[42]:

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)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

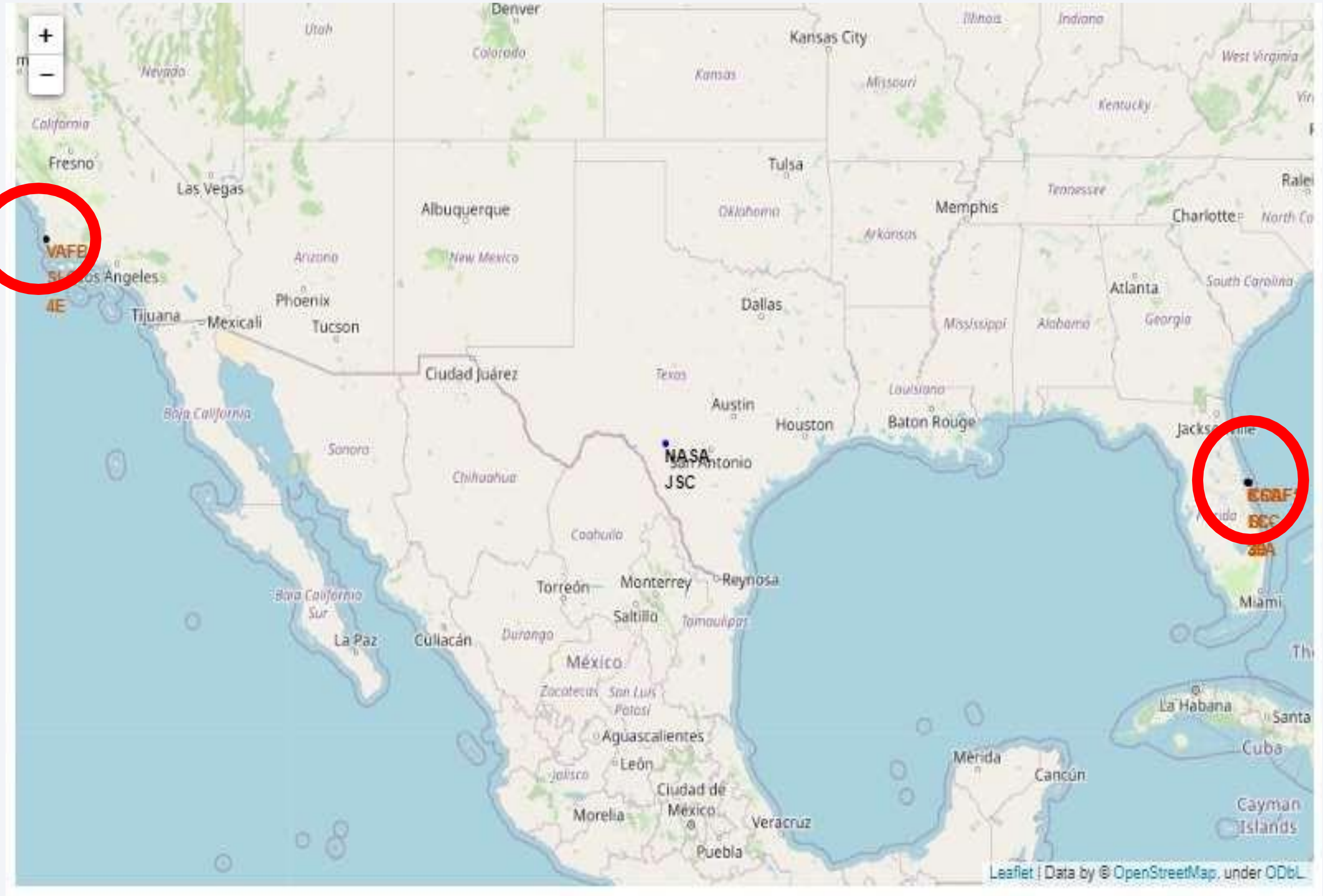
- The number of successful landings have increased since 2015.

Section 3

Launch Sites Proximities Analysis

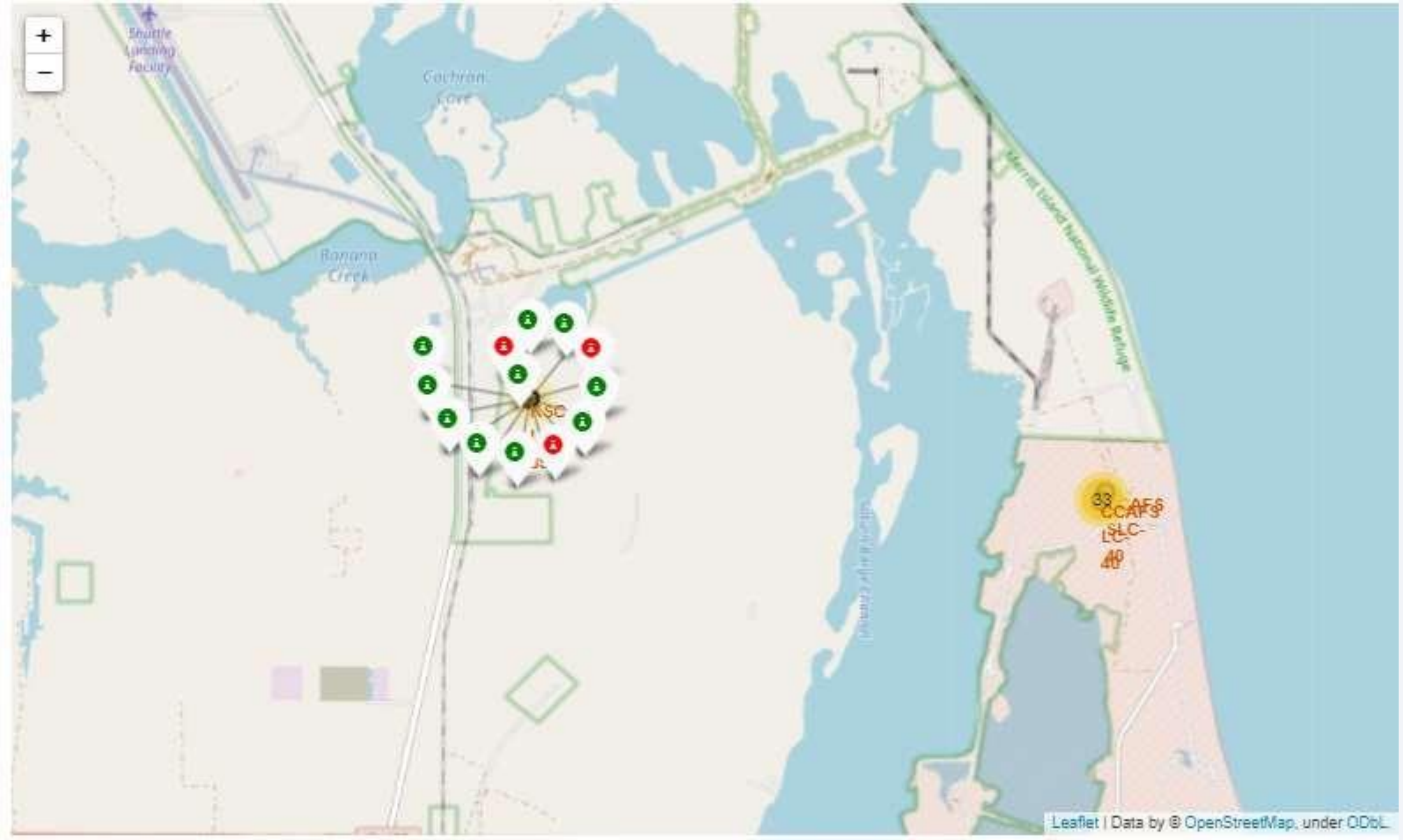
Launch Site Locations

- all launch sites are in very close proximity to the coast and they are also a couple thousand kilometers away from the equator line.



Success Rate of Rocket Launches

- The successful launches are represented by a **green** marker while the **red** marker represents failed rocket launches.



Surrounding Landmarks

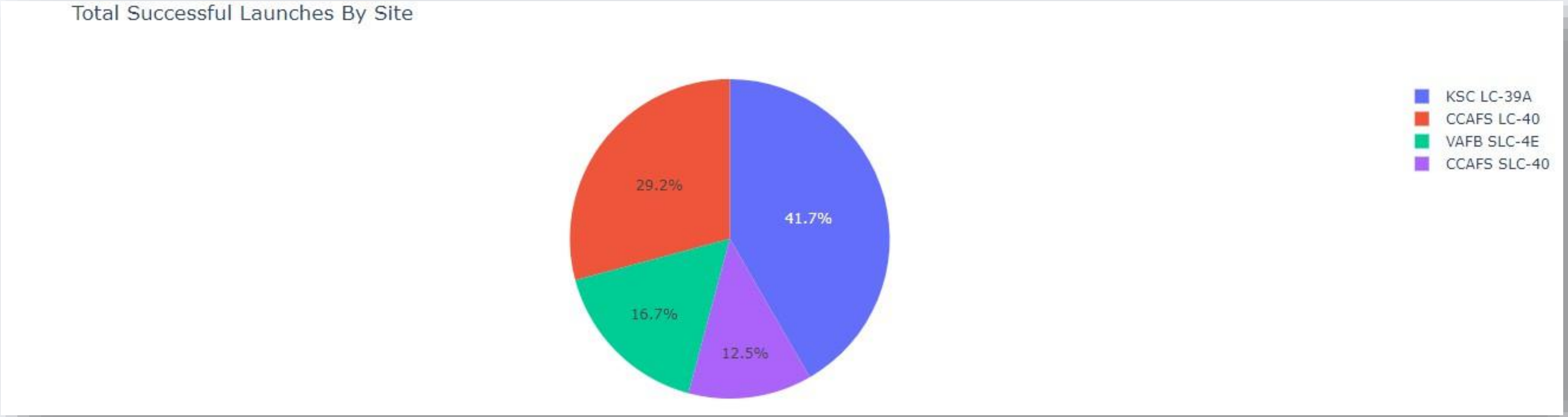
- It appears that launch sites are usually set up at least 18 km away from cities. This may be because of the desire to prevent any crashes near populated areas.
- It is also apparent that launch sites are in very close proximity to railways and highways. Perhaps, due to the necessary transportation requirements for rocket parts.
- The sites are close the coast line. This is evident with the many rocket landing tests on water bodies like the ocean.

Map Object	Colour
Nearest Highway	Green
Nearest Railway	Purple
Nearest City	Crimson
Nearest Coastline	Dark Blue

Section 4

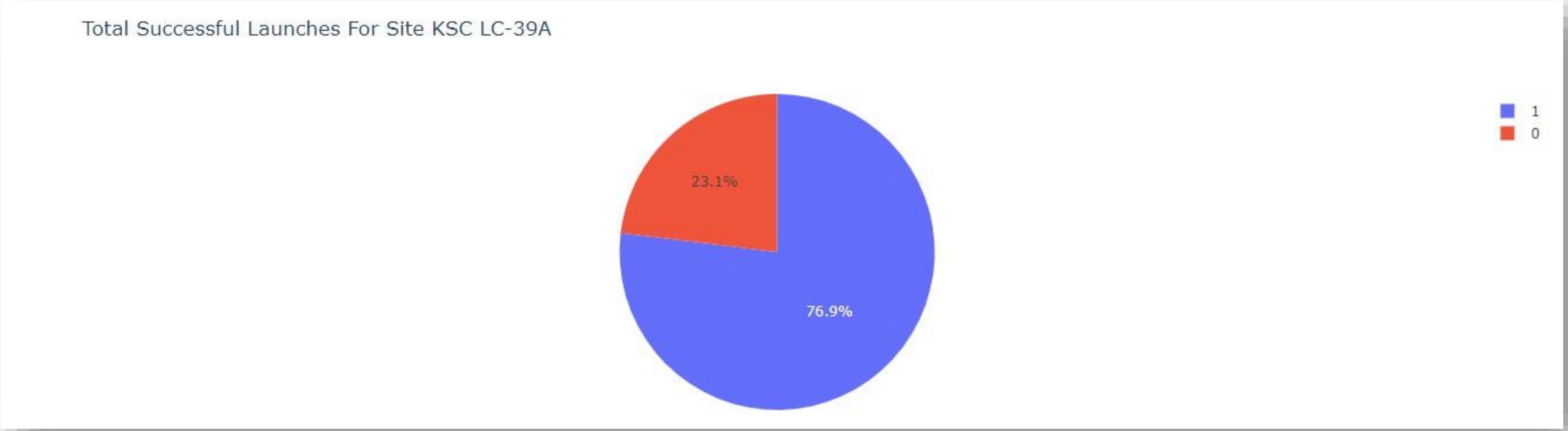
Build a Dashboard with Plotly Dash

Successful Launches by Site



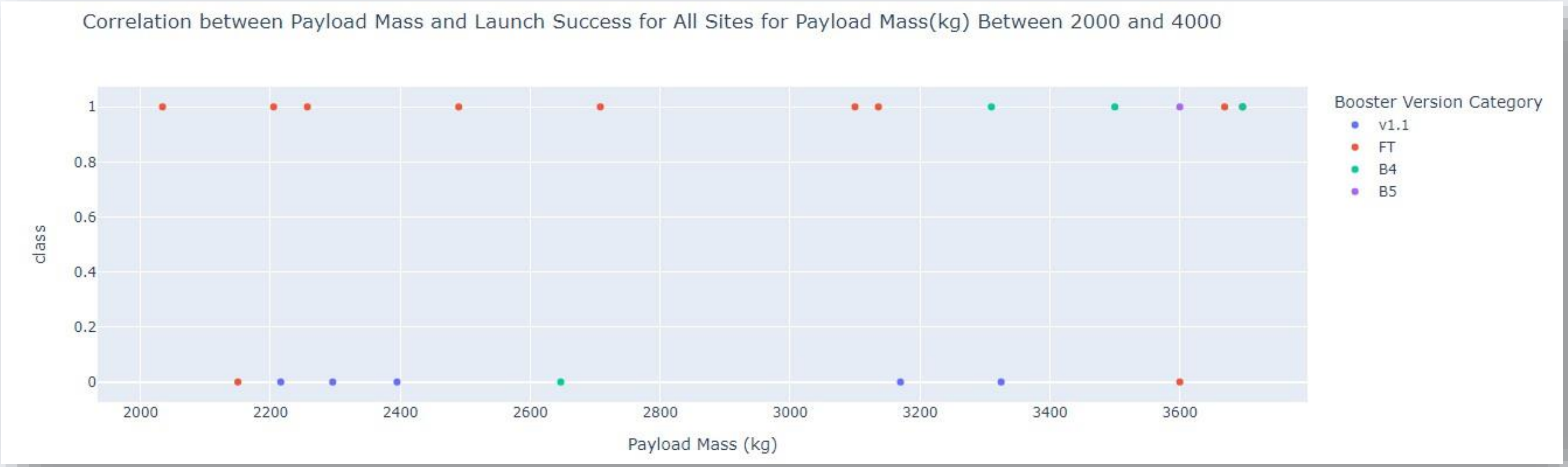
- You can see from the plot that Site KSC LC-39A has the largest successful launches as well the highest launch success rate.

Total Successful Launches for Site KSC LC-39A



- You can see that 76.9% of the total launches at site KSC LC-39A were successful. This is a the highest success rate of all the different launch sites.

Payload Mass vs. Launch Success for All Sites



- It appears that the payload range between 2000 kg and 4000 kg has the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Find the method performs best:

In [28]:

```
accuracy = [svm_cv_score, logreg_score, knn_cv_score, tree_cv_score]
accuracy = [i * 100 for i in accuracy]

method = ['Support Vector Machine', 'Logistic Regression', 'K Nearest Neighbour', 'Decision Tree']
models = {'ML Method':method, 'Accuracy Score (%)':accuracy}

ML_df = pd.DataFrame(models)
ML_df
```

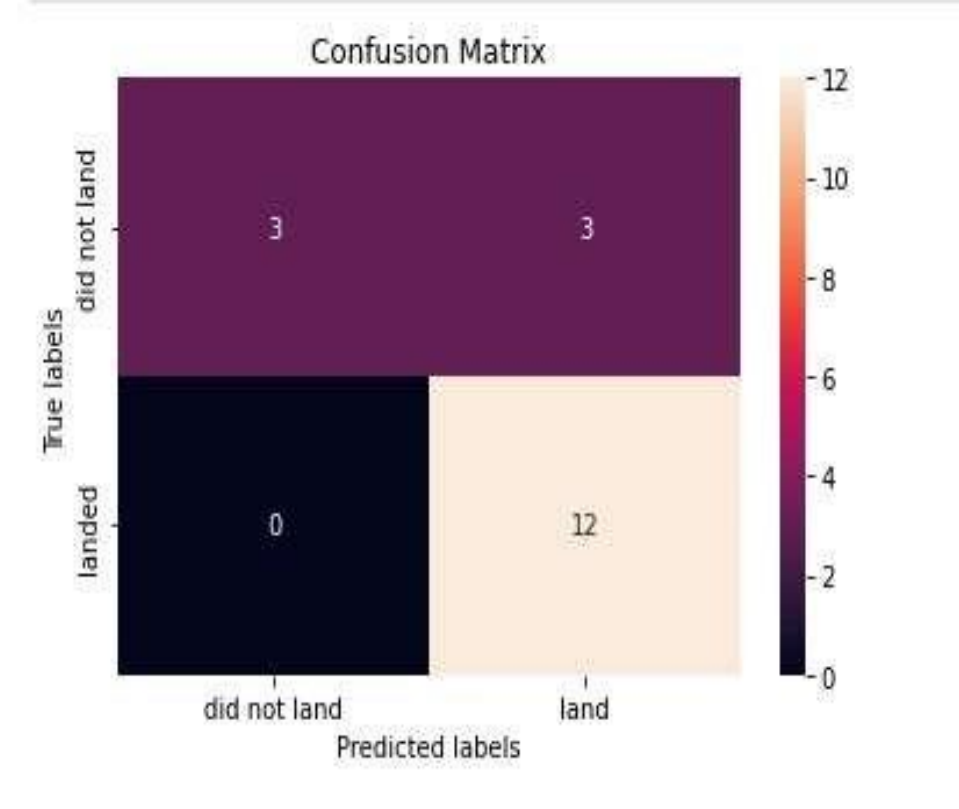
Out[28]:

	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

- You can see that All the methods have an identical accuracy score of 83.33%, so we decided to use Logistic Regression for the classification

Confusion Matrix

- The chart shows the confusion matrix of the Logistic Regression model that was chosen.
- The model only failed to accurately predict 3 labels.



Conclusions

In order to compete with Space X Through this process, a general picture of their success methods are

- All their launch sites are located near the coast, away from nearby cities. This enabled to them to test their rocket landings without much interference.
- Site KSC LC-39A had the highest launch success rate out of all the launch sites.
- From 2015 onwards, the success rate of rocket landings significantly increased. It was also apparent that landing success increased with flight number

All this data was used to train a machine learning model that is able to predict the landing outcome of rocket launches with 83.33% accuracy.

Appendix

Haversine's Formula used to calculate distances on Folium Maps

```
from math import sin, cos, sqrt, atan2, radians

def calculate_distance(lat1, lon1, lat2, lon2):
    # approximate radius of earth in km
    R = 6373.0

    lat1 = radians(lat1)
    lon1 = radians(lon1)
    lat2 = radians(lat2)
    lon2 = radians(lon2)

    dlon = lon2 - lon1
    dlat = lat2 - lat1

    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))

    distance = R * c
    return distance
```

Insights

Things to Consider

• Dataset: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set

• Feature Analysis / PCA: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy

• XGBoost: Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models

Thank you!