IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Passavit Apinanthawachpong
September 17th, 2023

Outline

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

Executive Summary

Methodology

- Data collection
- Data wrangling
- Exploratory Data Analysis
- Interactive Visual Analytics
- Predictive Analysis

Results

- Exploratory Data Analysis results
- Interactive Analytics demo
- 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.

- The problems include :
  - What are the factors which can affect the landing outcome?
  - How can those factors be modified to increase the probability of the successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - API, Web Scraping
- Perform data wrangling
  - One-hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, evaluate classification models

Data Collection

- We will be working with SpaceX launch data that is gathered from an API, specifically the SpaceX REST API. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

- Another popular data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. We will also be using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.

Data Collection – SpaceX API

- Request rocket launch data from SpaceX API.
- Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize().
- Perform data cleaning.
- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_data-collection-api.ipynb

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

```
# Lets take a subset of our dataframe keeping only the features we want and the flight
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc']]

# We will remove rows with multiple cores because those are falcon rockets with 2 extra
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]

# Since payloads and cores are lists of size 1 we will also extract the single value in
data['cores'] = data['cores'].map(lambda x : x[0])
data['payloads'] = data['payloads'].map(lambda x : x[0])

# We also want to convert the date_utc to a datetime datatype and then extracting the d
data['date'] = pd.to_datetime(data['date_utc']).dt.date

# Using the date we will restrict the dates of the launches
data = data[data['date'] <= datetime.date(2020, 11, 13)]
```

Data Collection - Scraping

- Perform an HTTP GET method to request the Falcon9 Launch HTML page.
- Create a BeautifulSoup object from the HTML response.
- Extract all column/variable names from the HTML table header.
- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_webscraping.ipynb

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

```
# Use BeautifulSoup() to create a BeautifulSoup object fro
soup = BeautifulSoup(response)
```

```
extracted_row = 0
#Extract each table
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
   # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is as number corresponding to launch a number
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
        else:
            flag=False
        #get table element
        row=rows.find_all('td')
        #if it is number save cells in a dictonary
```

Data Wrangling

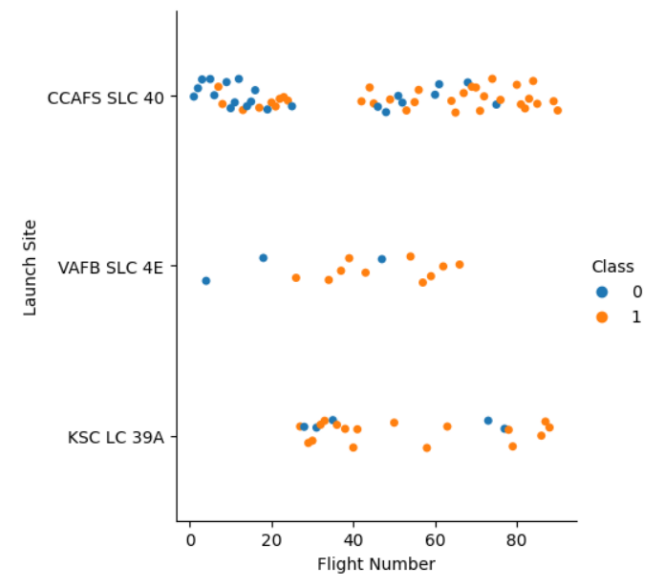
- We would like landing outcomes to be converted to either 0 or 1 and assign them to the column “class”. 0 is a bad outcome, that is, the booster did not land. 1 is a good outcome, that is, the booster did land. The variable Y will represent the classification variable that represents the outcome of each launch.

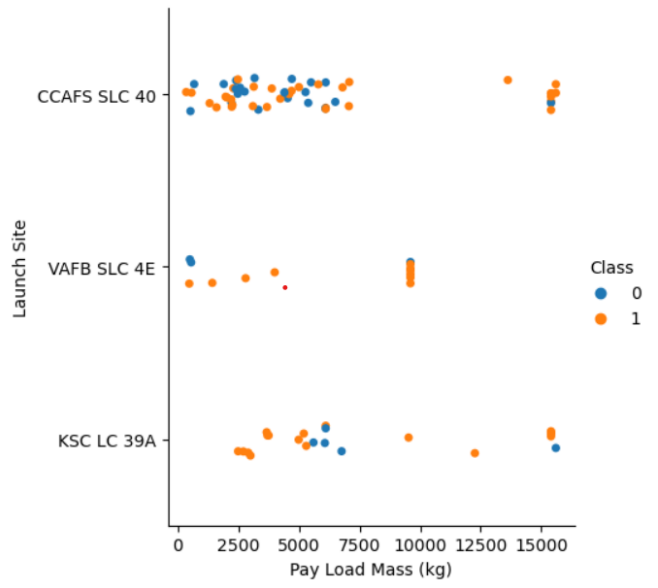
- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_data_wrangling.ipynb

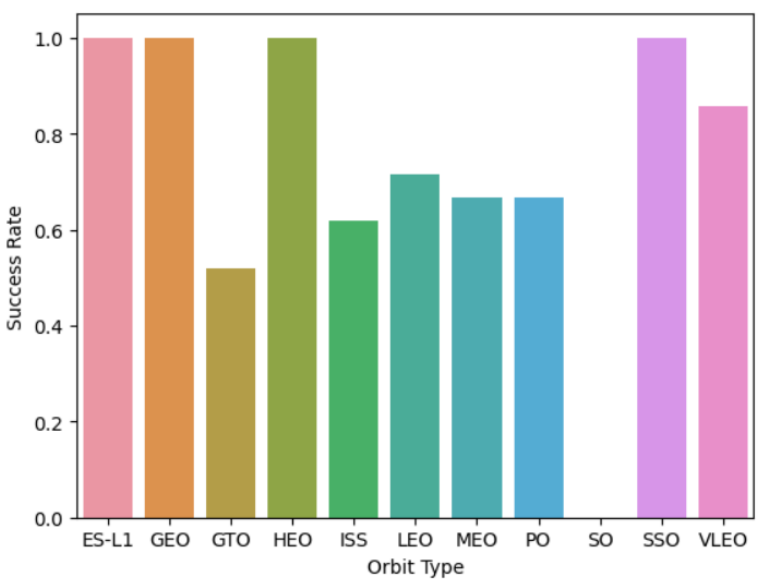
```
df['Class']=landing_class
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

EDA with Data Visualization



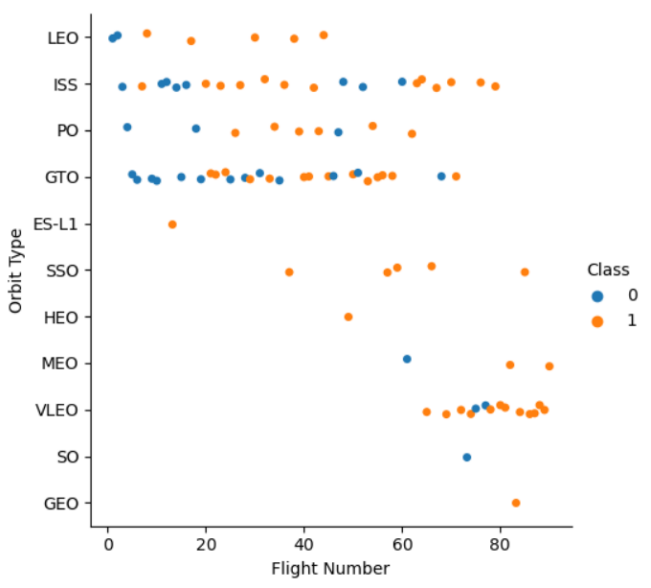


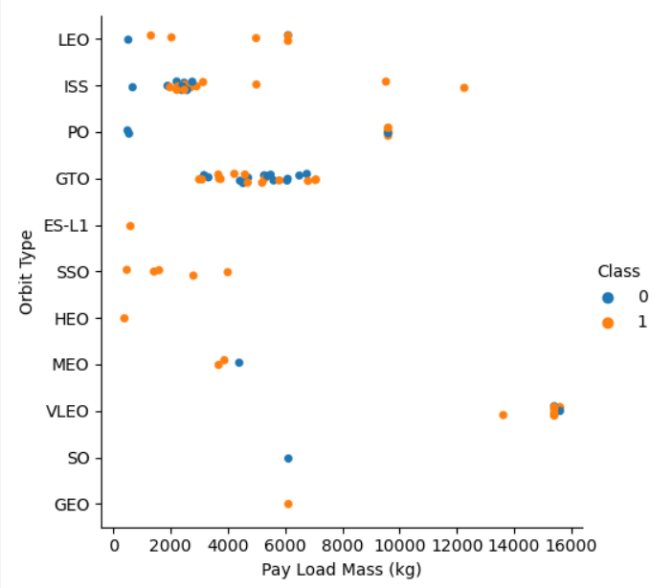


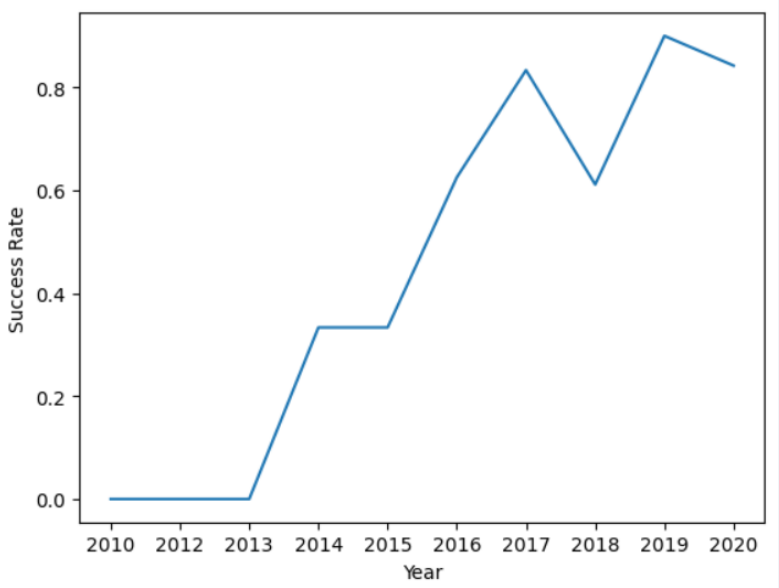
- Visualize the relationship between Flight Number and Launch Site.
- Visualize the relationship between Payload and Launch Site.
- Visualize the relationship between success rate of each orbit type.

- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_eda-dataviz.ipynb

EDA with Data Visualization







- Visualize the relationship between Flight Number and Orbit type.
- Visualize the relationship between Payload and Orbit type.
- Visualize the launch success yearly trend

- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_eda-dataviz.ipynb

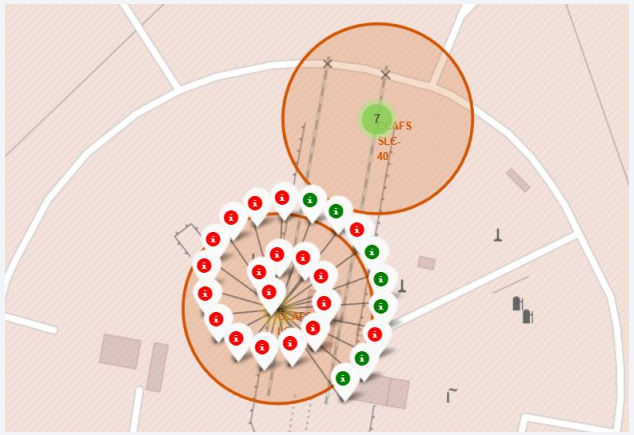
EDA with SQL

- 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 succesful 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.
- 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.
- 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.

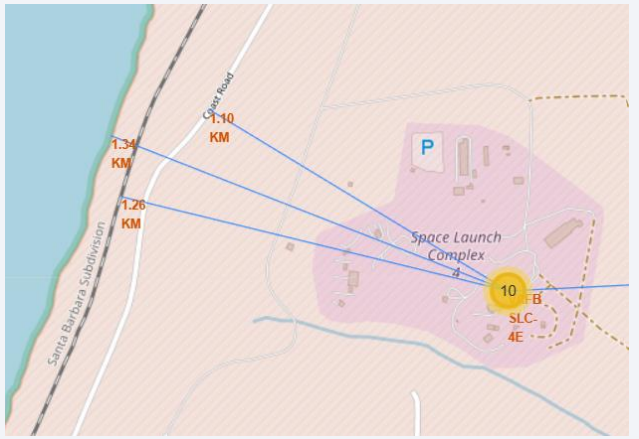
- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_eda-sql.ipynb

Build an Interactive Map with Folium

- For each launch site, add a Circle object based on its coordinate values. In addition, add Launch site name as a popup label.



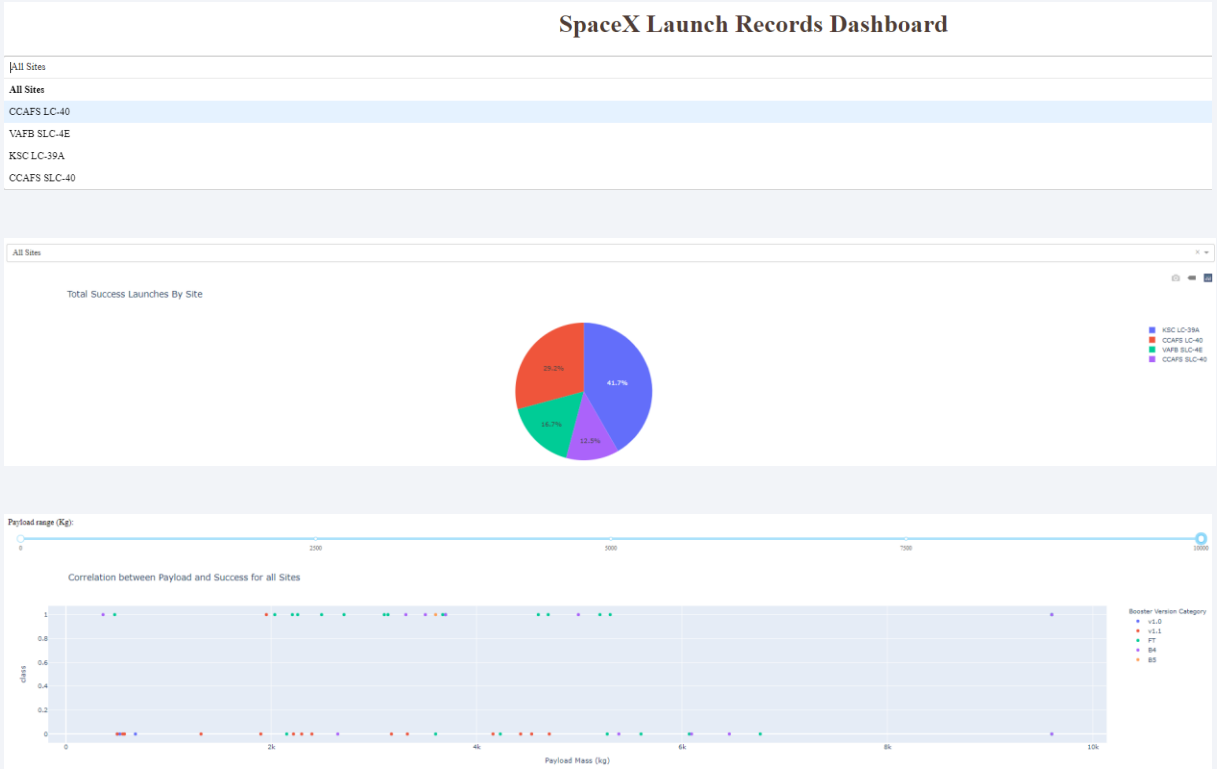
- Create markers for all launch records. If a launch was successful, then we use a green marker and if a launch was failed, we use a red marker.



- Create a marker with distance to a closest city, railway, highway, and coastline.

- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_launch_site_location.ipynb

Build a Dashboard with Plotly Dash



- Add a Launch Site Drop-down Input Component.
- Add a callback function to render success-pie-chart based on selected site dropdown.
- Add a Range Slider to Select Payload.
- Add a callback function to render the success-payload-scatter-chart scatter plot.

- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_dash_app.py

Predictive Analysis (Classification)

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```
knn_cv = GridSearchCV(KNN, parameters, cv=10).fit(X_train, Y_train)
```

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

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```

```
print('Accuracy of Logistic Regression :', logreg_cv.score(X_test, Y_test))
print('Accuracy of Support Vector Machine :', svm_cv.score(X_test, Y_test))
print('Accuracy of Decision Tree :', tree_cv.score(X_test, Y_test))
print('Accuracy of K-Nearest Neighbors :', knn_cv.score(X_test, Y_test))
```

```
Accuracy of Logistic Regression : 0.8333333333333334
Accuracy of Support Vector Machine : 0.8333333333333334
Accuracy of Decision Tree : 0.8333333333333334
Accuracy of K-Nearest Neighbors : 0.8333333333333334
```

- Use the function train_test_split to split the data X and Y into training and test data.
- Fit the object to find the best parameters from the dictionary parameters.
- Calculate the accuracy on the test data for each model using the method score.

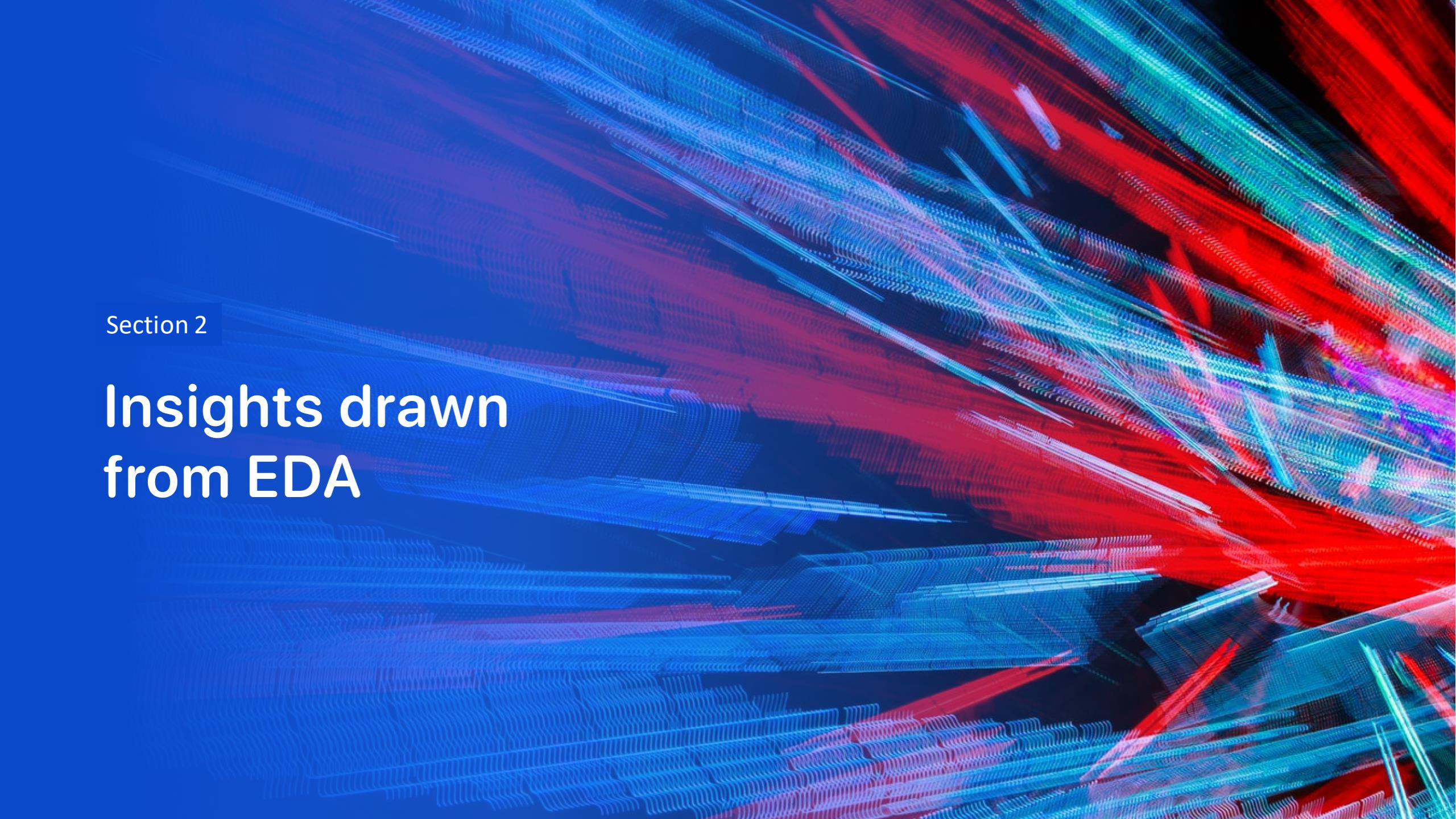
- https://github.com/PassavitA/IBM-Data-Science/blob/e487a4a37cdaf22bad3526811b4480383cd18f44/spacex_machine_learning_prediction.ipynb

Results

- Exploratory data analysis results
  - Insights drawn from EDA
- Interactive analytics demo in screenshots
  - Proximities Analysis, Interactive Dashboard
- Predictive analysis results
  - Predictictive Analysis (Classification)

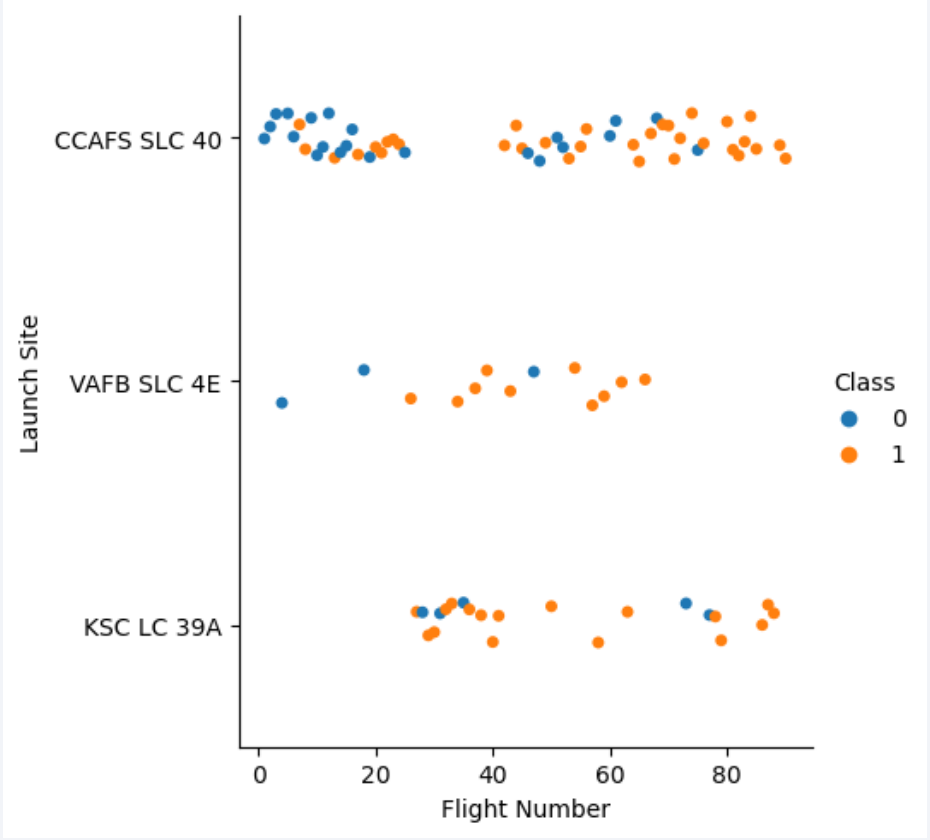
Section 2

Insights drawn from EDA



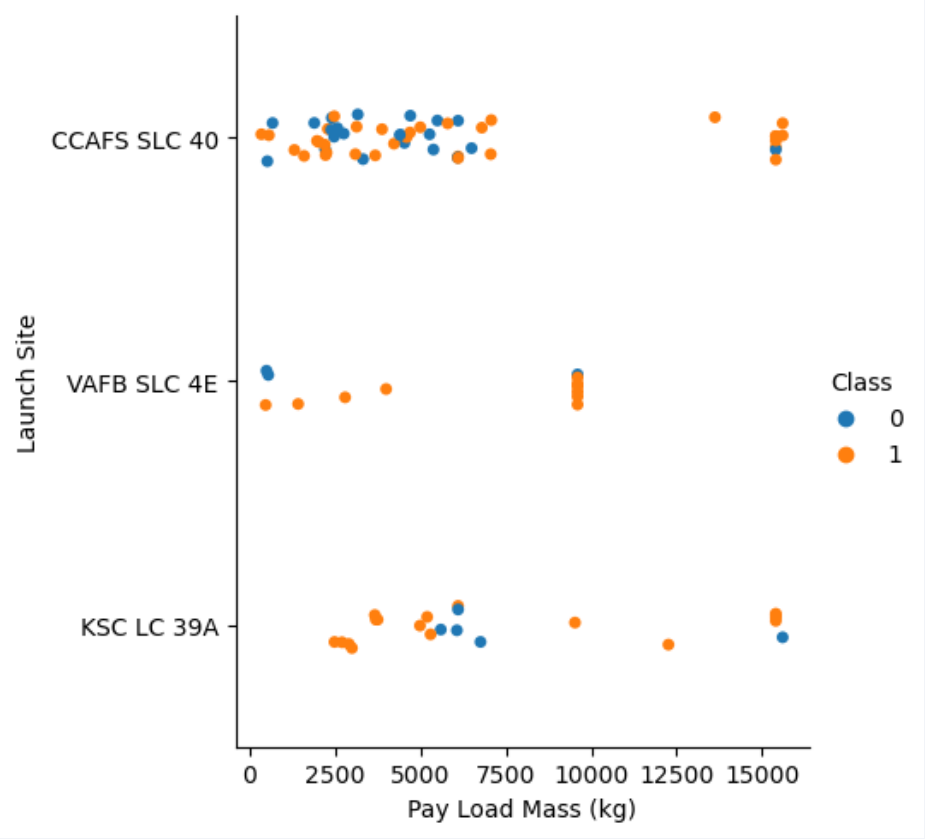
Flight Number vs. Launch Site

- For CCAFS SLC, the unsuccessful landing decreases as the flight number increases.
- For VAFB SLC 4E, the unsuccessful landing also decreases as the flight number increases.
- For KSC LC 39A, the landing outcome doesn't seem the be correlated with the flight number.



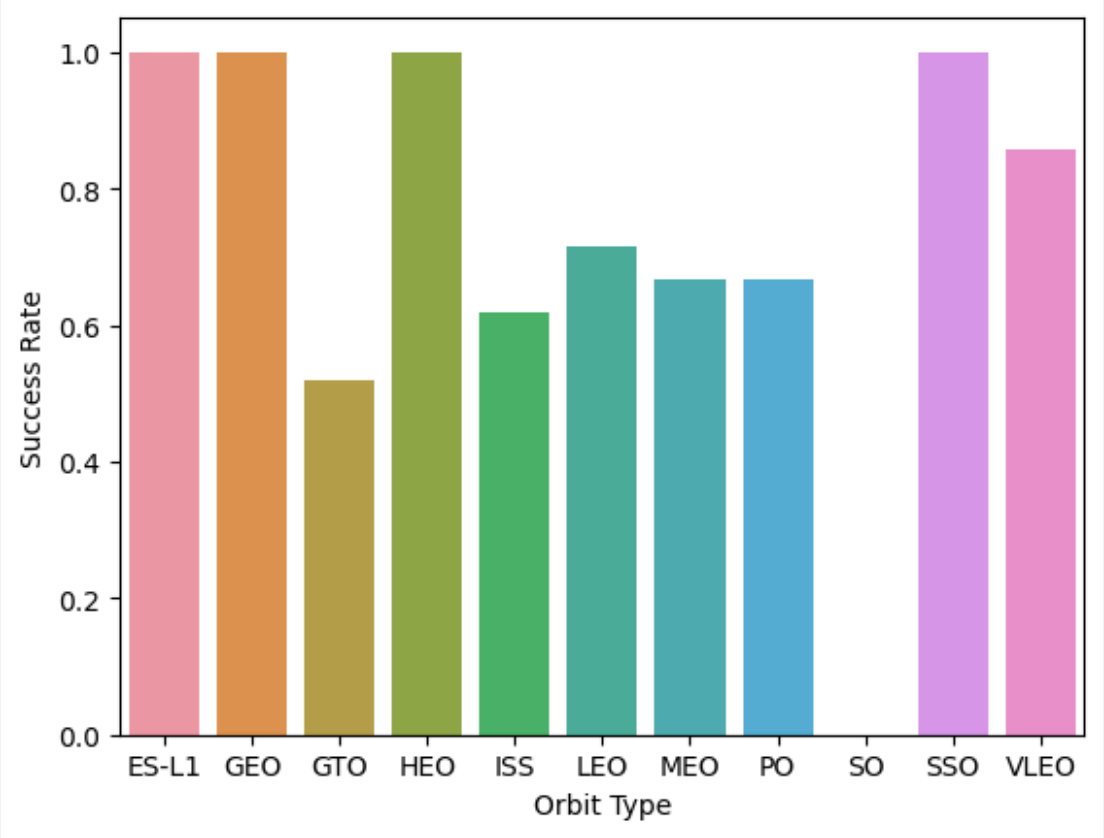
Payload vs. Launch Site

- For all the 3 sites, the landing outcome doesn't seem to be correlated with the payload mass.



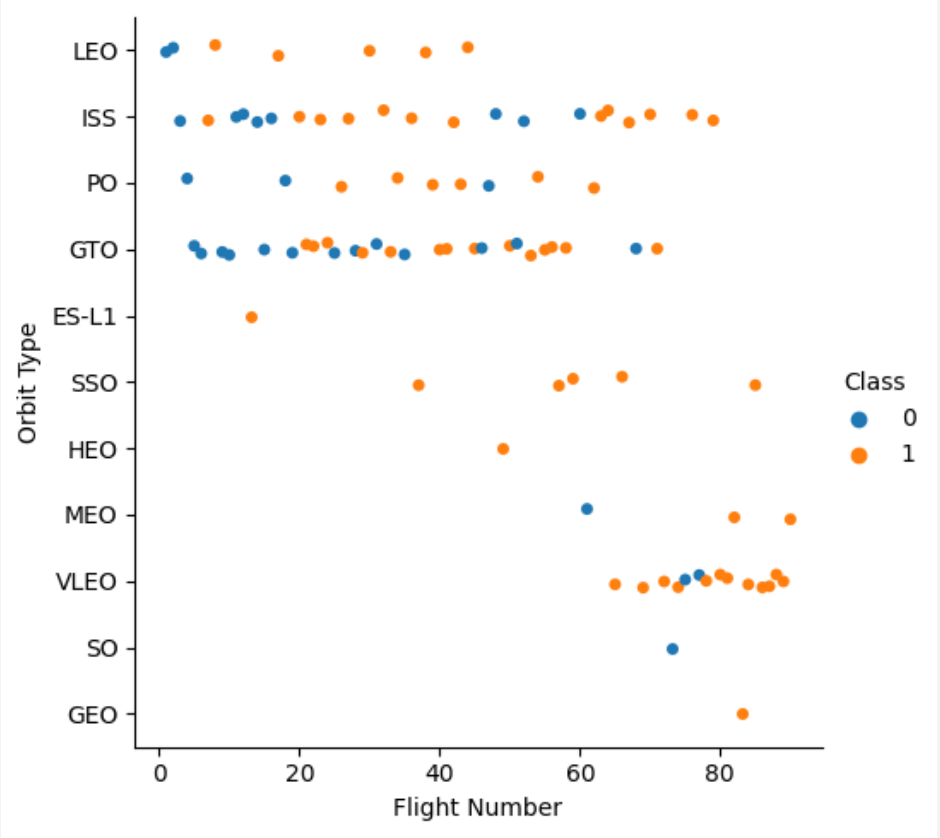
Success Rate vs. Orbit Type

- Orbit type with the highest success rates are ES-L1, GEO, HEO, and SSO with the 100% success rates.



Flight Number vs. Orbit Type

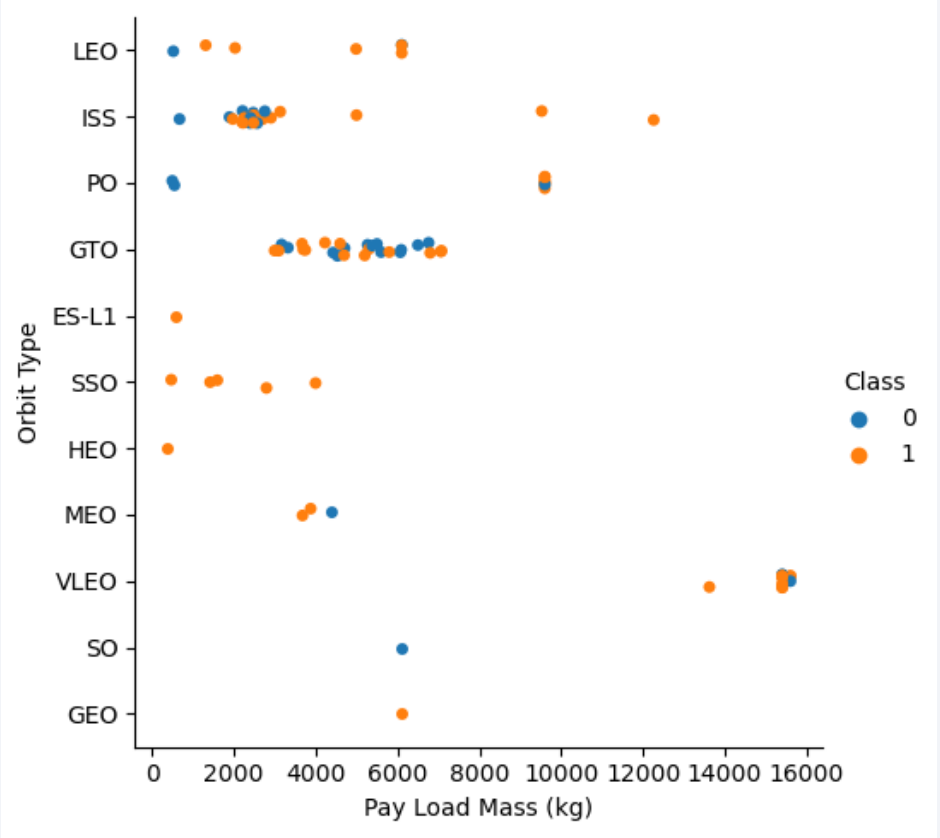
- For LEO, the success rate increases as the flight number increases.
- However, for other orbit types, the flight number doesn't seem to have an impact on the landing outcome.



Payload vs. Orbit Type

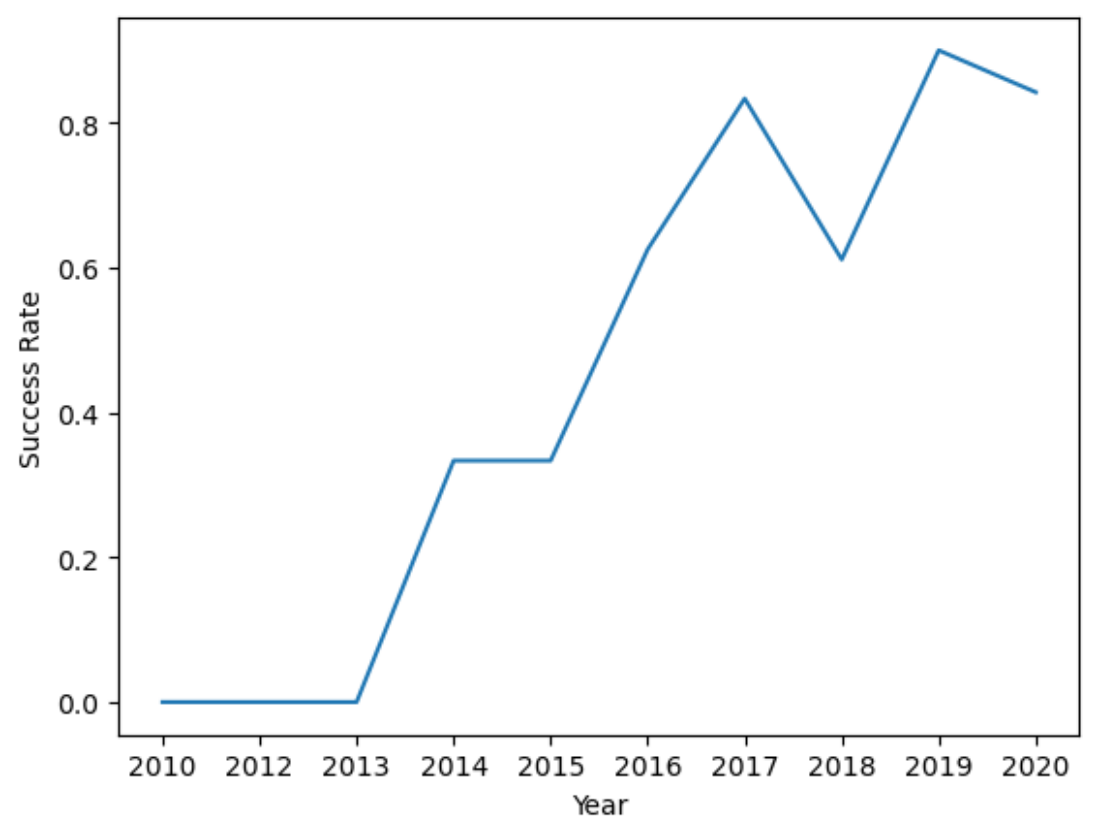
- With heavy payloads the successful landing or positive landing rate are more for PO, LEO, and ISS.

- However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing until 2020.



All Launch Site Names

- These are the names of the unique launch sites in the space mission.

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

Launch Site Names Begin with 'CCA'

- These are the first 5 records where launch sites begin with the string 'CCA'.

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

Total_Payload_Mass	Customer
45596	NASA (CRS)

- This is the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

Average_Payload_Mass	Booster_Version
2928.4	F9 v1.1

- This is the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

First_Landing_Ground_Pad	Landing_Outcome
2015-12-22	Success (ground pad)

- This is the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- These are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

Successful_Mission	Failure_Mission
100	1

- These are the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

- These are the names of the booster which have carried the maximum payload mass.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- These are the records with the failed landing_outcomes in drone ship for in year 2015, their booster versions and launch site names are also shown.

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

- These are the count of landing outcomes between the date 2010-06-04 and 2017-03-20, ranked in descending order.

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

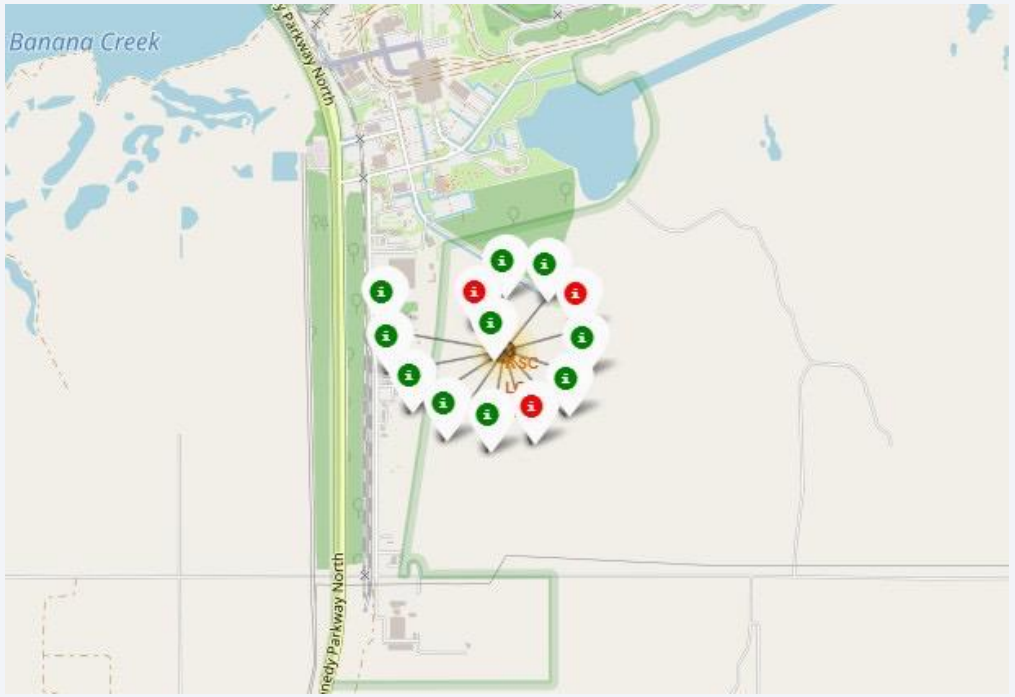

Section 3
Launch Sites
Proximities Analysis

Launch Site Locations

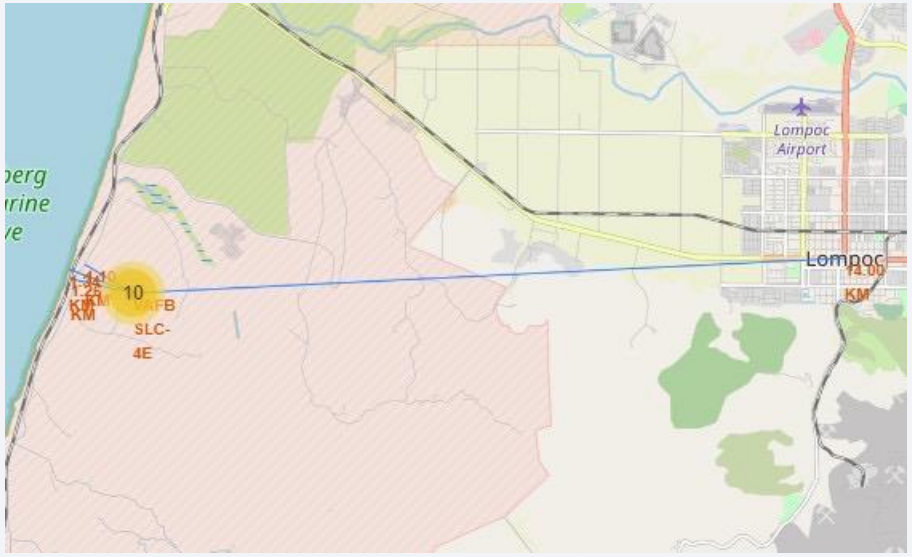
- All launch sites are not in proximity to the Equator line.
- However, all launch sites are in very close proximity to the coastline.

Launch Site Success Rates

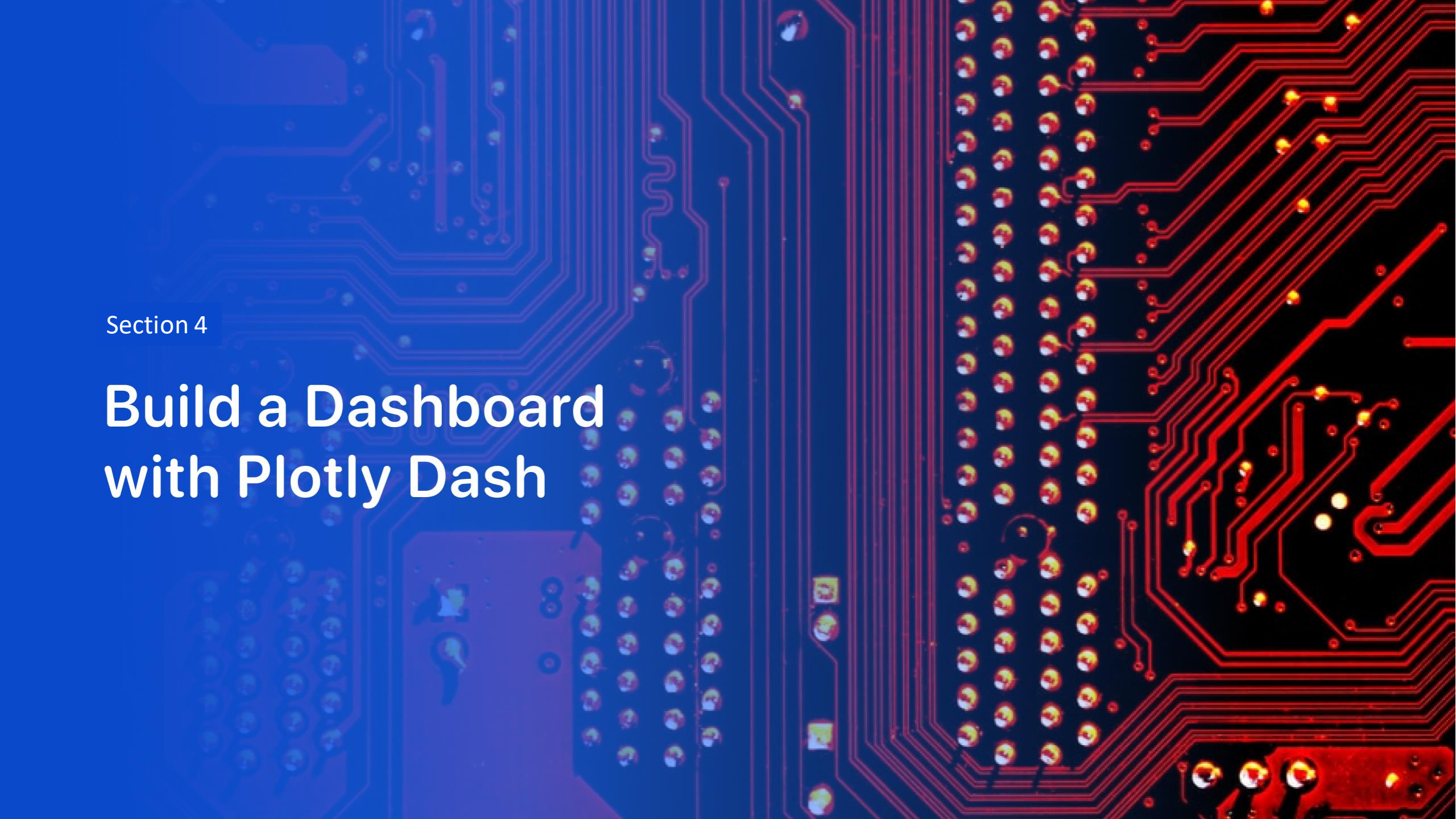
- From the color-labeled markers in marker clusters, the launch site that has the highest success rate is KSC LC-39A.



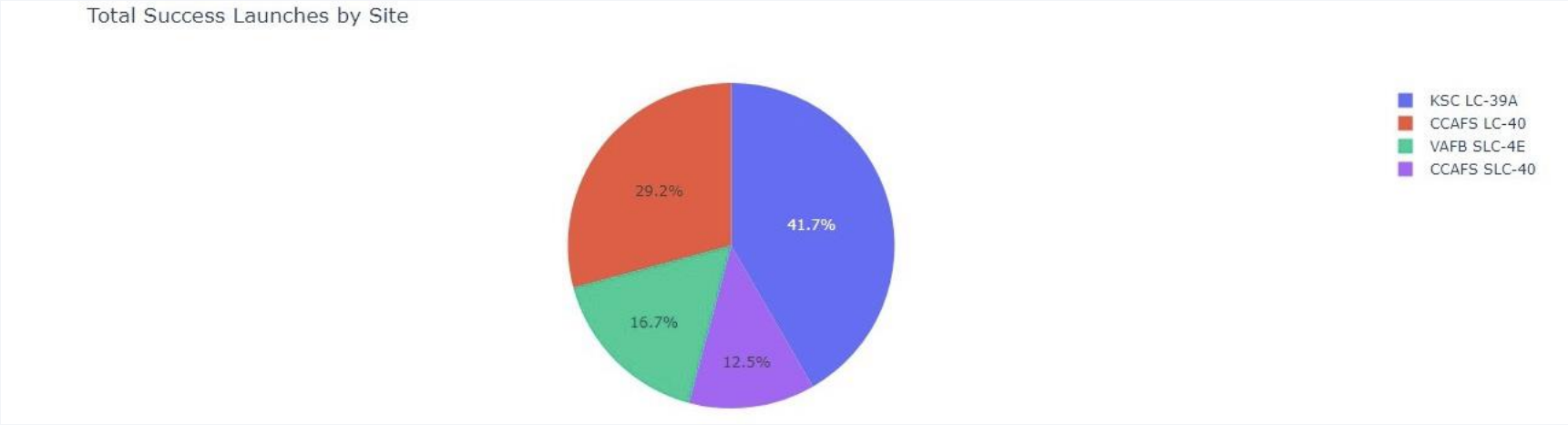
Launch Site's Proximities



- While railways, highways, coastline don't have an impact on the location of a launch site, the launch site tries to avoid being too close to the city.

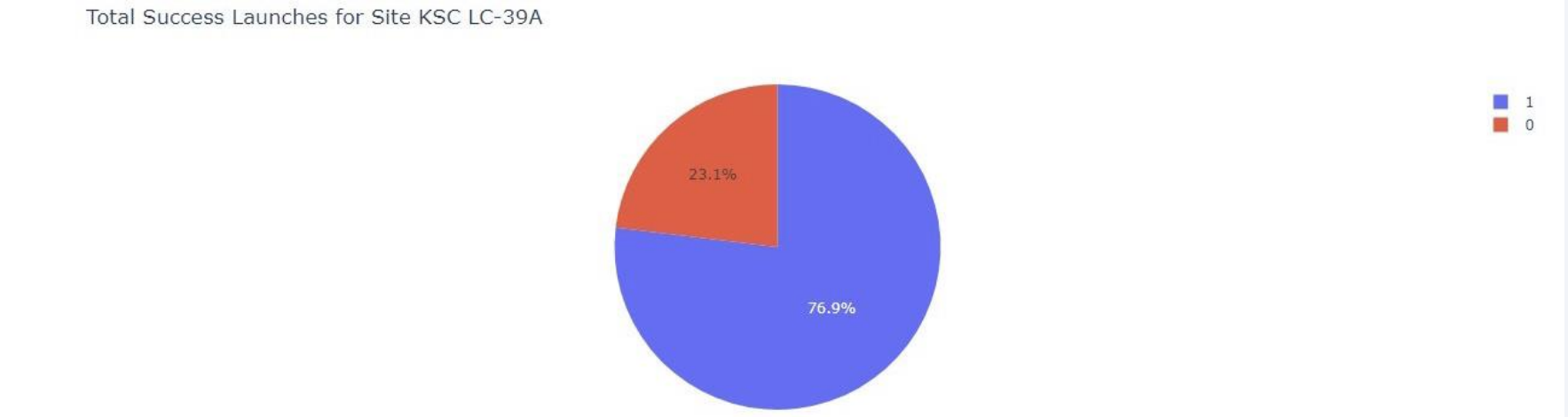

Section 4
Build a Dashboard with Plotly Dash

Success Count for Each Launch Site



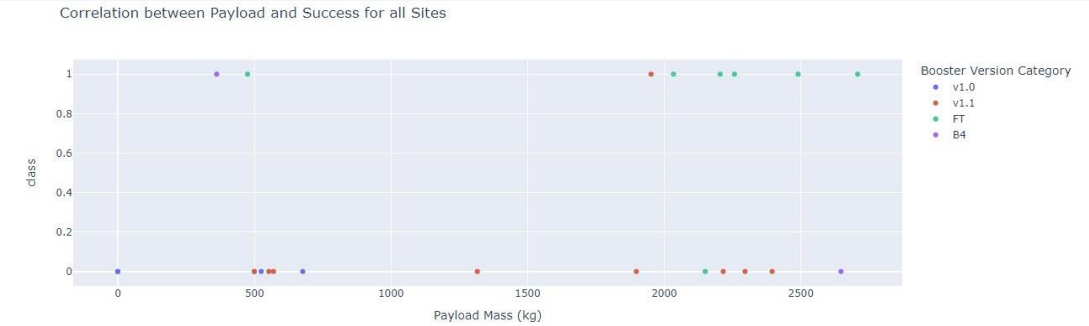
- KSC LC-39A is the launch site with the most success count among all sites.

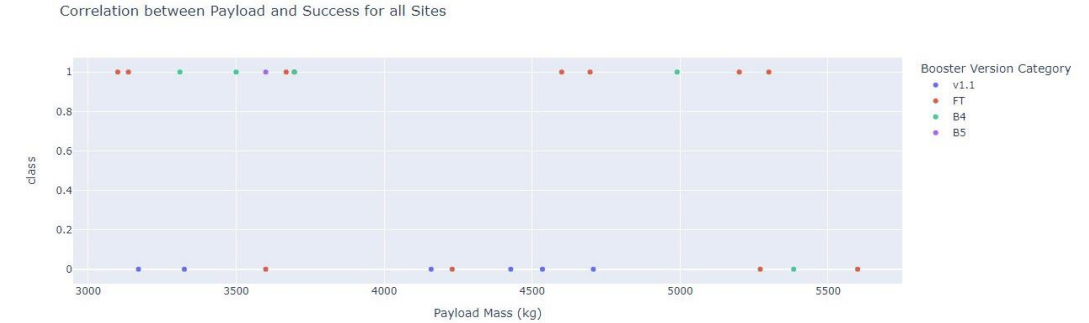
KSC LC-39A Success Rate

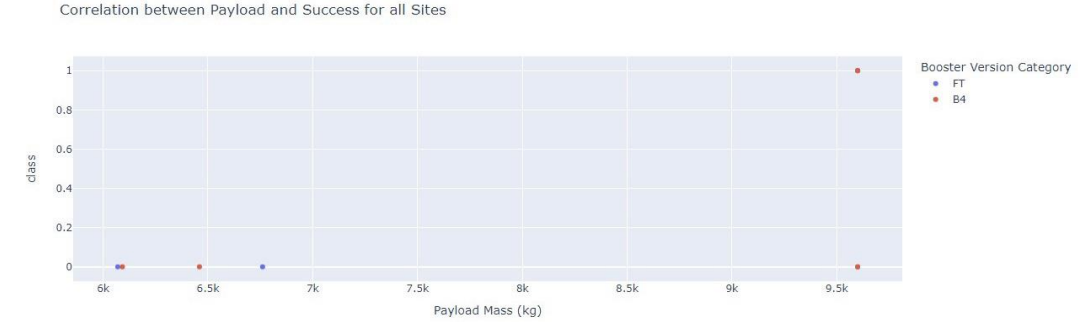


- KSC LC-39A has the success rate of 76.9%.

Payload vs Launch Outcome







- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider. The ranges are 0-3000kg, 3000-6000kg, and 6000-100000kg from the top to the bottom respectively.
- The low to medium payload mass range (0-6000kg) have more success rate when compared to the high range.
- For the booster version, FT seems to be the one with the highest success rate.

Section 5

Predictive Analysis (Classification)

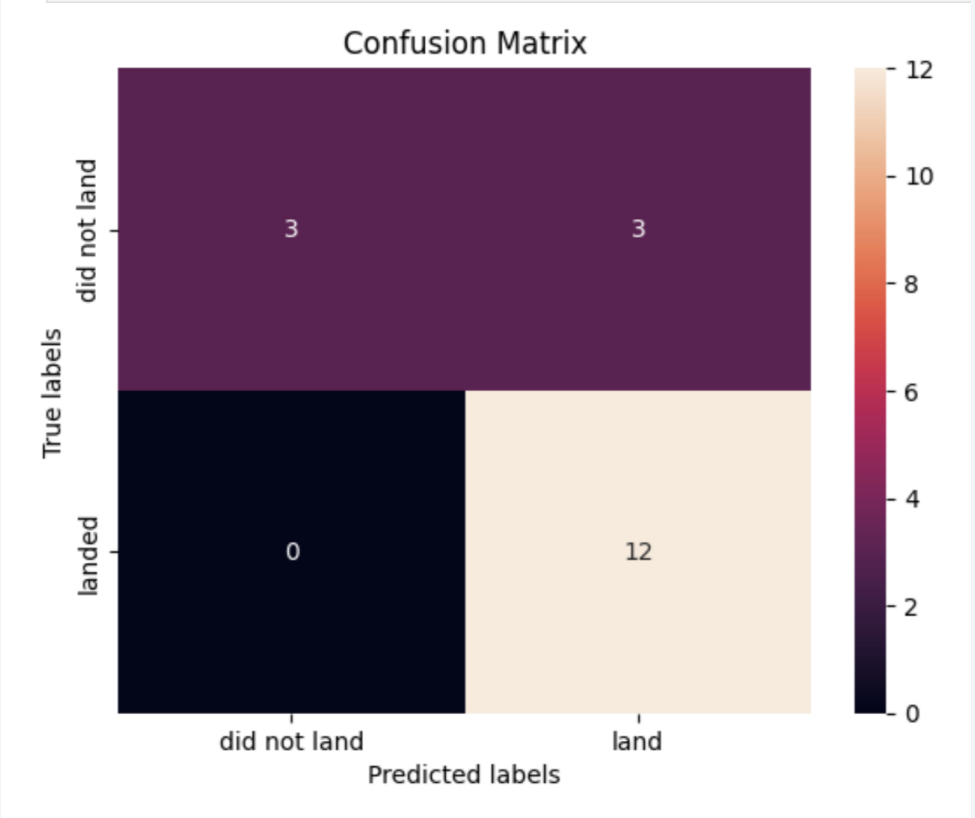
Classification Accuracy

```
Accuracy of Logistic Regression : 0.8333333333333334
Accuracy of Support Vector Machine : 0.8333333333333334
Accuracy of Decision Tree : 0.8333333333333334
Accuracy of K-Nearest Neighbors : 0.8333333333333334
```

- All the 4 classification models have the same accuracy.

Confusion Matrix

- The confusion matrix of the 4 classification models are the same with some false positive where the model wrongly predicted that the outcome to be positive.



Conclusions

- ES-L1, GEO, HEO, and SSO are orbits type with the highest success rate of 100%.
- The success rate of SpaceX launches has been increasing since the year 2013.
- Most of the successful landings are performed the the ground pad.
- KSC LC-39A is the launch site with the highest success rate of 76.9%.
- The low to medium payload mass range (0-6000kg) have more success rate when compared to the high range.
- FT is the booster version with the highest success rate.
- Any of the 4 classification models including Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors can be used to predict the landing outcome since the accuracy of the 4 are the same.

Thank you!