IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Robson Magalhães
May 20th, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Ploty Dash
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis Visualization Results
  - Predictive Analysis

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars.
  - Other providers cost upward of 165 million dollars each.
  - Much of the savings is because SpaceX can reuse the first stage.
  - If its possible to determine if the first stage will land, the cost of a launch could be determined and the savings for the reuse of the stage predicted;

- Problems you want to find answers
  - It is possible to determine if the first stage will land successfully?
  - This information could be used for costs prediction if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data Collection:
  - Request to the SpaceX API
  - Web Scrapping from Wikipedia
- Data Wrangling
  - Data were adjusted cleaning null, missing and irrelevant values.
  - The behavior of the data were observed including classifications for site launches, orbit types and mission outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictive models were used to evaluate the best classifier (Linear Regression, KNN, Decision Tree and SVM)

Data Collection

- How data sets were collected:
  - Using SpaceX Rest API
  - Using Web Scraping

Data Collection – SpaceX API

Requesting rocket launch data from SpaceX API with the URL

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

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

Decode the response content as a Json and turn it into a Pandas dataframe

```
data=pd.json_normalize(response.json())
```

Use of custom funcions to cleaen data using the IDs given for each launch. Specifically the columns rocket, payloads, launchpad, and cores were used.

- Data get from the Columns:
  - From the rocket:
    - booster name
  - From the payload
    - mass of the payload and the orbit that it is going to
  - From the launchpad:
    - the name of the launch site being used, the longitude, and the latitude.
  - From cores:
    - outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to seperate version of cores, the number of times this specific core has been reused, and the serial of the core.

Combine the columns into a dictionary.

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

Filter the dataframe to only include Falcon 9 launches

```
data_falcon9= df_launch[df_launch.BoosterVersion != 'Falcon 1']
```

https://github.com/robsonvps/IBM_Data_Science_Professional_Certificate/blob/main/1.1%20Complete%20the%20Data%20Collection%20API%20Labv1.ipynb

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from its URL

```
html_data = requests.get(static_url).text
```

Create a *BeautifulSoup* object from the HTML

```
soup = BeautifulSoup(html_data,'html.parser')
```

Collect all relevant column names from the HTML table

```
html_tables=soup.find_all('table')
```

```
column_names = []
for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if (name != None and len(name) > 0):
        column_names.append(name)
```

Extract column name

```
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

Create a data frame by parsing the launch HTML tables

```
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders
    # 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
        if flag:
            extracted_row += 1
            # Flight Number value
            # TODO: Append the flight_number into launch_dict with key `Flight No.`
            launch_dict['Flight No.'].append(flight_number)
            print(flight_number)
            datatimelist=date_time(row[0])
```

Fill up the launch dictionary with launch records extracted from table rows

Create a dataframe

```
df=pd.DataFrame(launch_dict)
```


https://github.com/robsonvps/IBM_Data_Science_Professional_Certificate/blob/main/1.2%20Complete%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

Deal with **Missing Values**

Calculate the **number of launches** on each site

Calculate **the number and occurrence** of each orbit

Calculate **number and occurrence of mission outcome** per orbit type

Create a **landing outcome label** from Outcome column

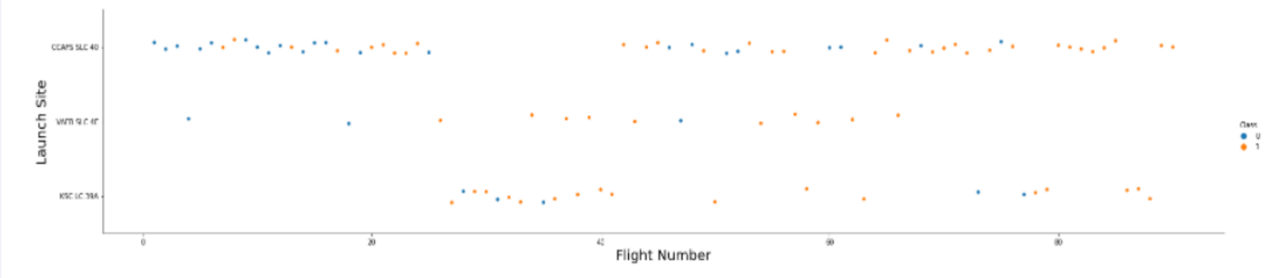
This variable is the classification that represents the outcome of each launch.
If the value is :

- **zero**, the first stage did not land successfully;
- **one** means the first stage landed Successfully

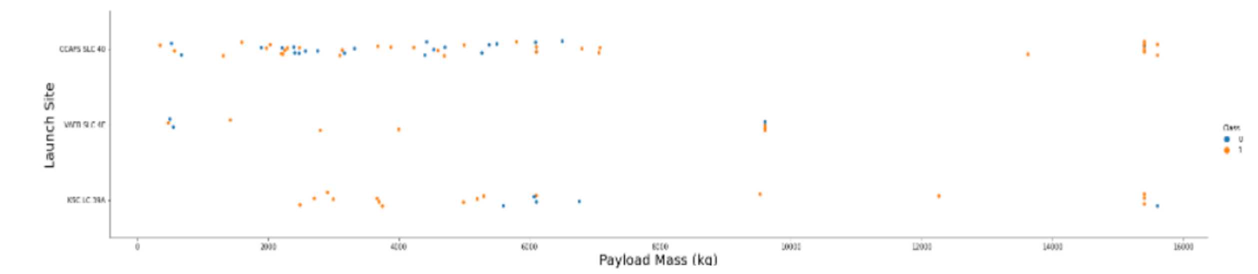
https://github.com/robsonvps/IBM_Data_Science_Professional_Certificate/blob/main/1.3%20Data%20wrangling.ipynb

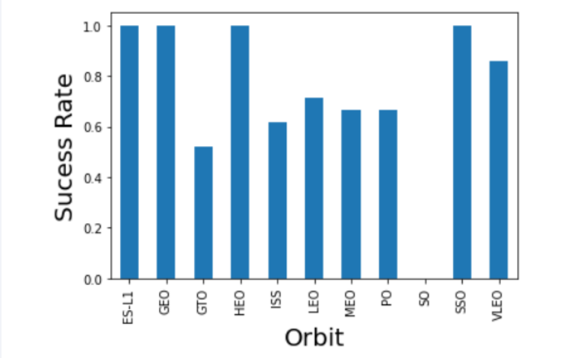
EDA with Data Visualization

Visualize the relationship between <u>Flight Number and Launch Site</u>



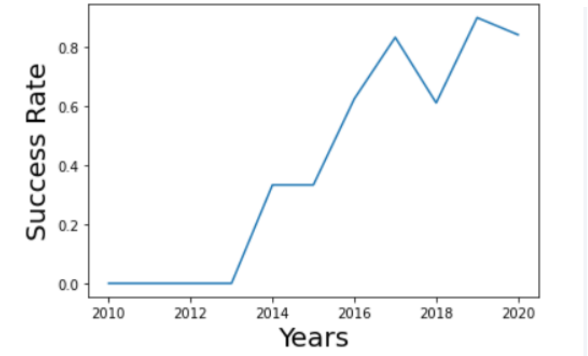
Visualize the relationship between <u>Payload and Launch Site</u>

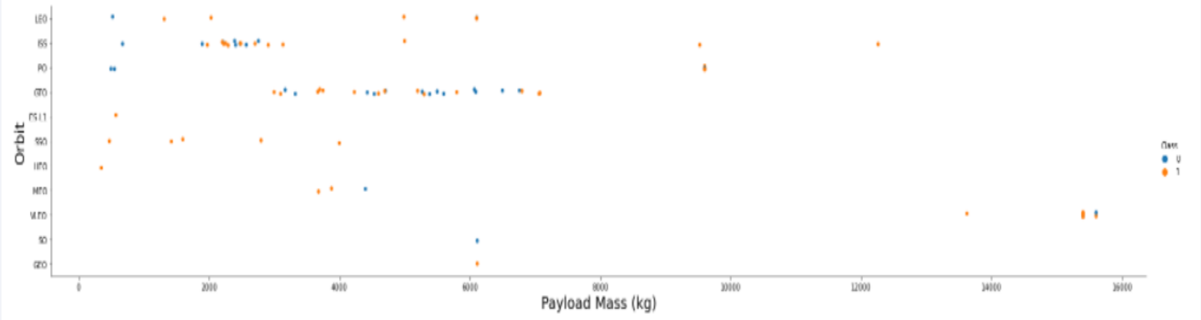




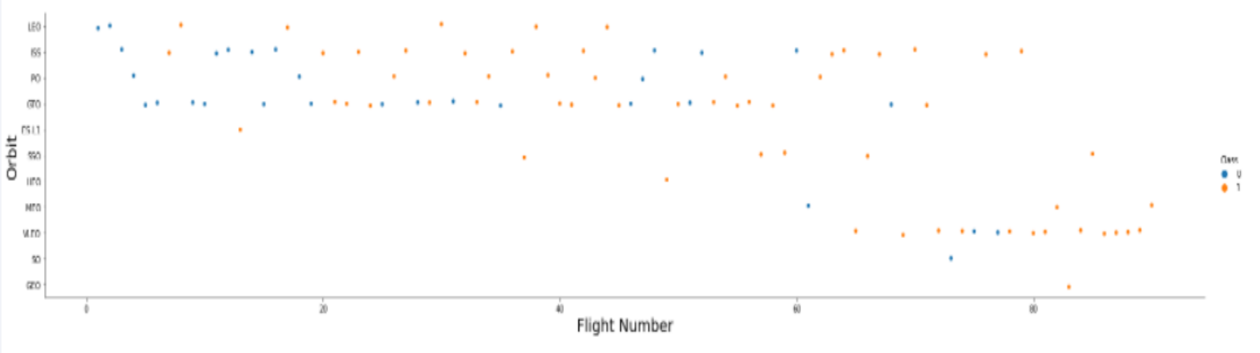
Visualize the relationship between <u>success rate of each orbit type</u>

Visualize the <u>launch success yearly trend</u>





Visualize the relationship between <u>Payload and Orbit type</u>



Visualize the relationship between <u>FlightNumber and Orbit type</u>

https://github.com/robsonvps/IBM_Data_Science_Professional_Certificate/blob/main/2.2%20Exploratory%20Data%20Analysis%20Using%20Pandas%20and%20Matplotlib.ipynb

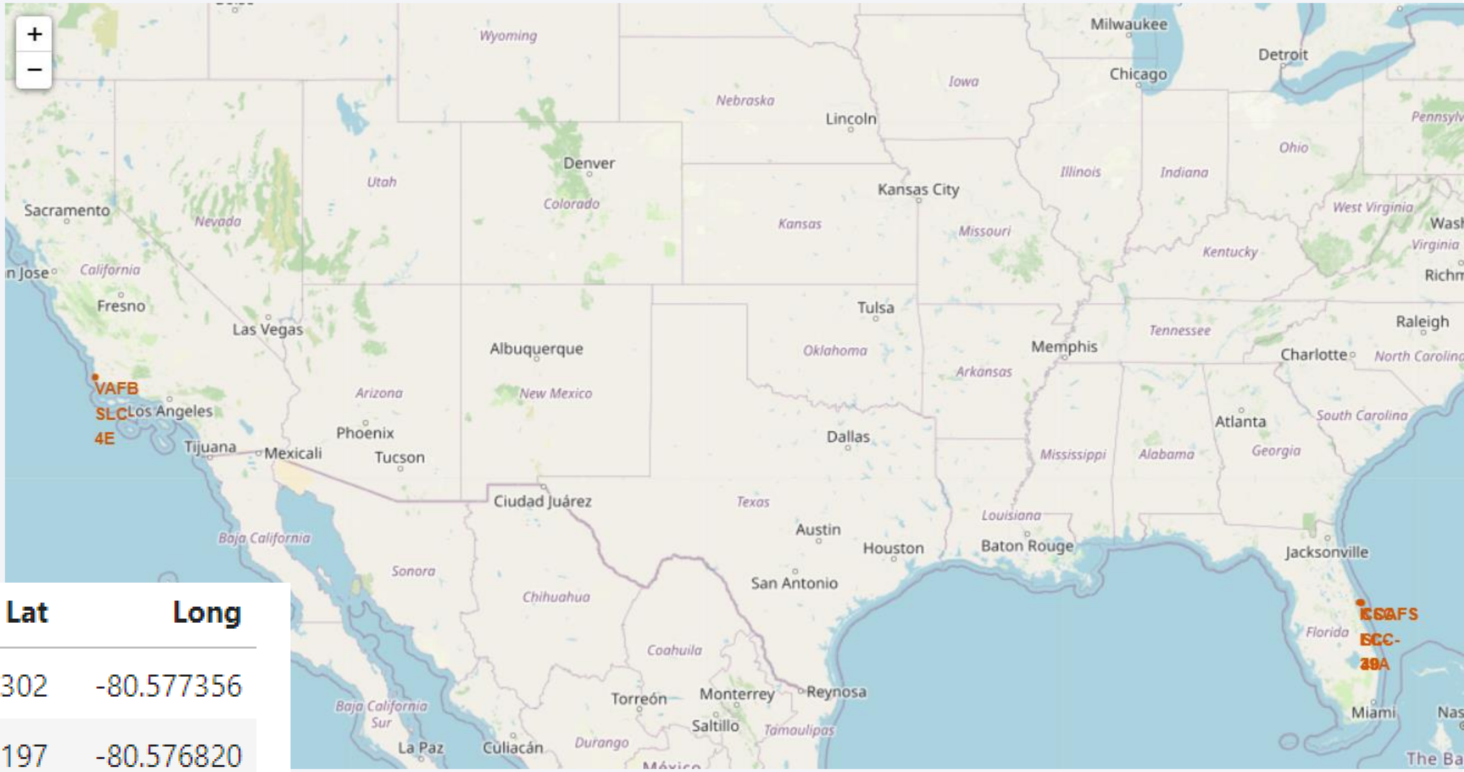
EDA with SQL

- SQL queries performed:
  - **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 achieved;**
  - **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.**

https://github.com/robsonvps/IBM_Data_Science_Professional_Certificate/blob/main/2.1%20Exploratoy%20Data%20Analysis%20Using%20SQL.ipynb

Build an Interactive Map with Folium

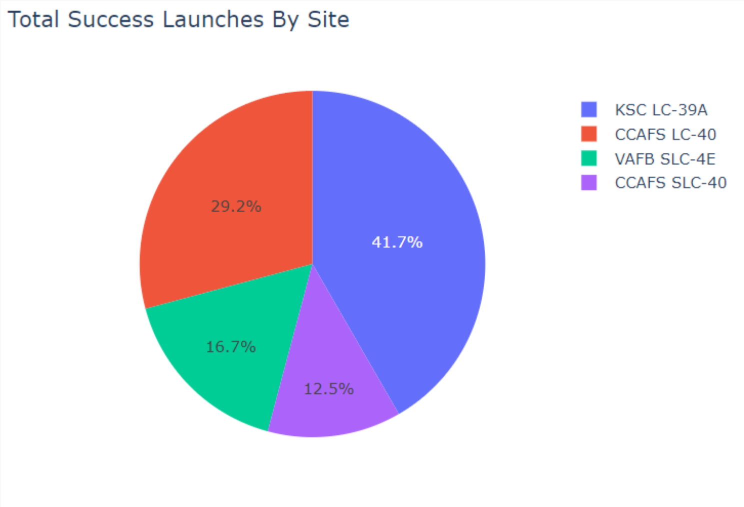
- Folium Circle and Marker were included for each launch site on the map;
- Objects added to make possible the visualization of geographic characteristics of each launch site. E.g., proximity to the Equator line or to the coast;

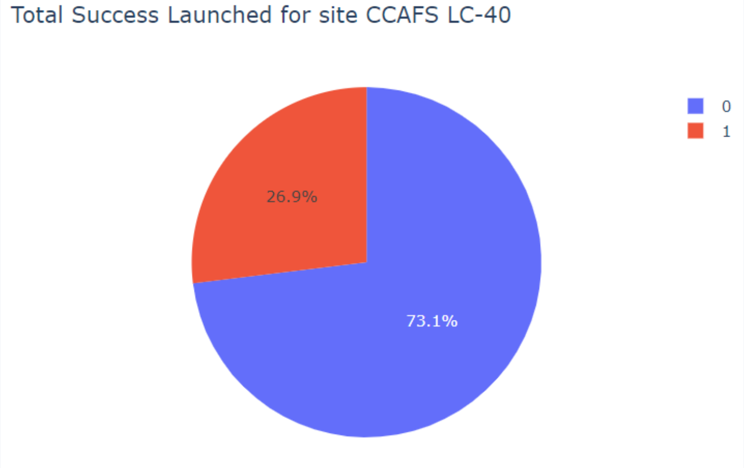


Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

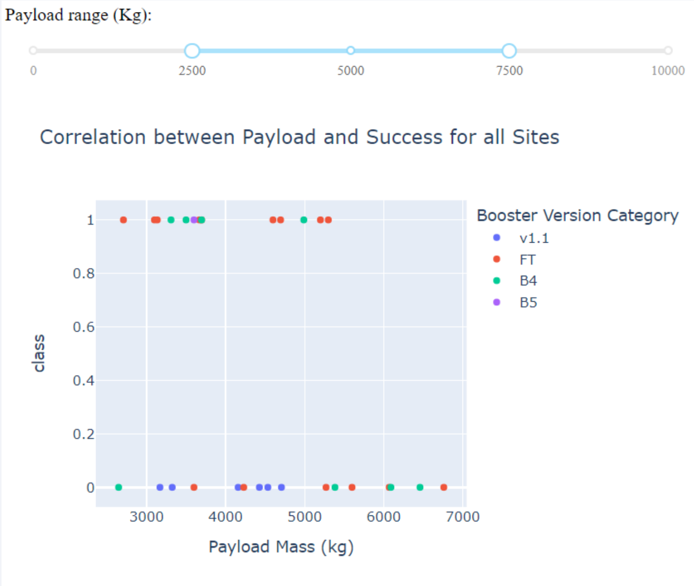
https://github.com/robsonvps/IBM_Data_Science_Professional_Certificate/blob/main/3.1%20Complete%20Data%20Visualization%20Using%20Folium.ipynb

Build a Dashboard with Plotly Dash





- Using Plotly Dash was possible to create an Interactive view of the data;



- As an example, we can observe the pie chart with success/failure for CCAFS LC 40 launch site and graphics separating the booster versions for medium payload mass (2500 to 7500 kg));

https://github.com/robsonvps/IBM_Data_Science_Professional_Certificate/blob/main/3.2%20Build%20an%20Interactive%20Dashboard%20with%20Plotly%20Dash.ipynb

Predictive Analysis (Classification)

Load the dataframe

Standardize the data in X

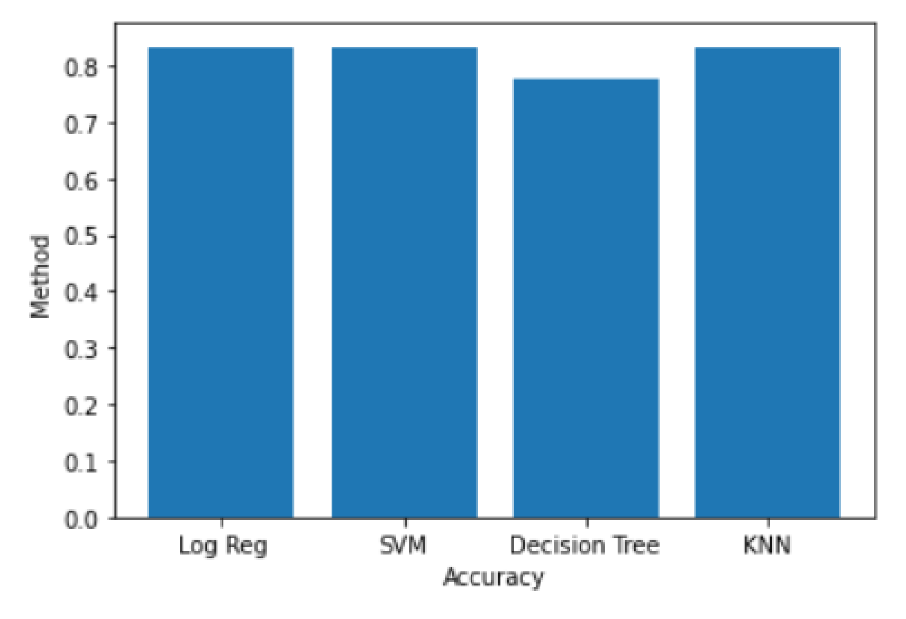
Split the data into training and testing data

Create a logistic regression object

Create a support vector machine object

Create a decision tree classifier object

Create a k nearest neighbor's object



Logistic Regression, SVM and KNN achieved the highest accuracy with 83,33%

https://github.com/robsonvps/IBM_Data_Science_Professional_Certificate/blob/main/4%20Predictive%20Analysis.ipynb

Results

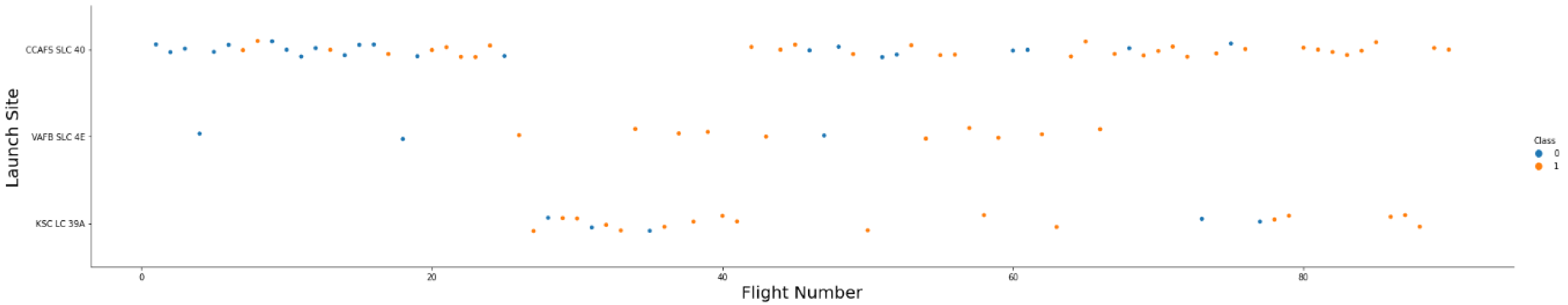
- Exploratory data analysis and Interactive analytics results of Space X launches:
  - The launch site keep a minimum safe distance from the surrounds inhabited area (city, highway, railroad and coastline);
  - Orbits ES L1, GEO, HEO, SSO have more success;
  - The last flights were launched to VLEO orbit;
  - KSC LC 39A launch site had more success launches;

- Predictive analysis results
  - Logistic Regression, SVM and KNN methods achieved the highest accuracy with 83,33% of success in the prediction. These methods are the **recommendation** for the dataset used.

Section 2

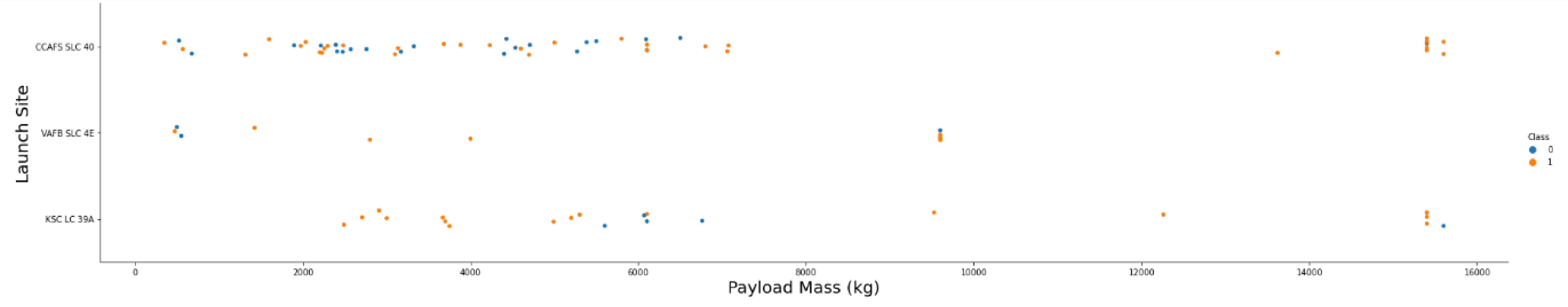
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 was more used for launches during the period analyzed;
- CCAFS SLC 40 and KSC LC 39A are constantly being used;
- VAFB SLC 4E is not being used during the last quarter of the period analyzed.

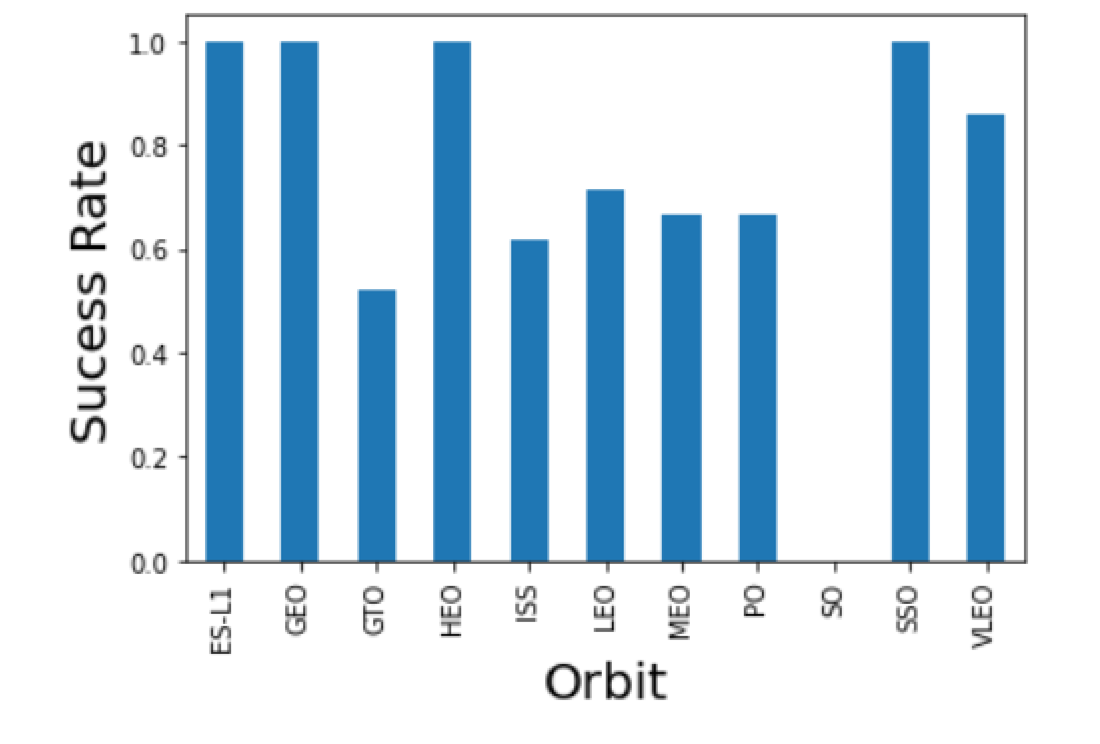
Payload vs. Launch Site



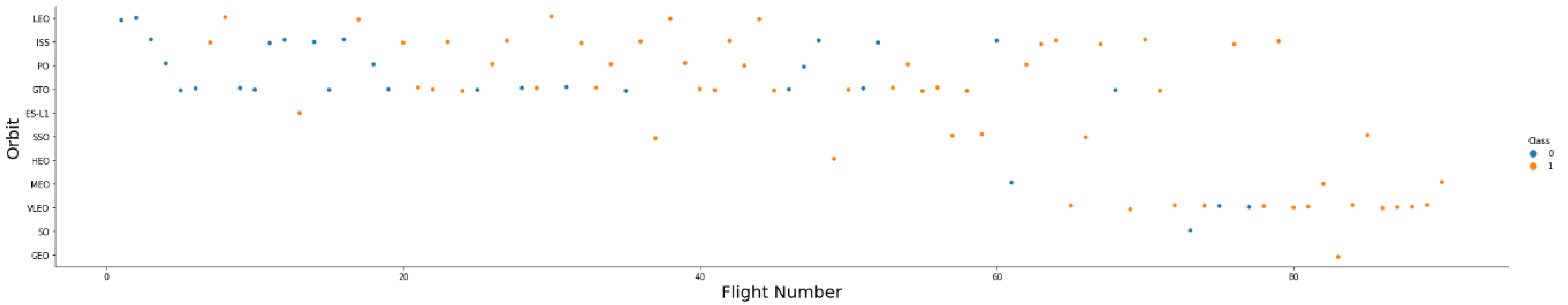
- VAFB SLC 4E has no record of launches with high payload mass;
- CCAFS SLC 40 has the bigger part of occurrence of launches for low payload mass.
- Low payload (2000kg to 4000kg) launched to KSC LC 39A orbit has high success rate;
- Medium payload (around 6000kg) launched to KSC LC 39A orbit has very low success rate
- High payload (around 15000kg) launched to KSC LC 39A and CCAFS SLC 40 orbits has high success rate

Success Rate vs. Orbit Type

- Orbits ES L1, GEO, HEO, SSO have more success;
- Orbit GTO has fewer success, followed close by orbits ISS, MEO and PO;
- No record for orbit SO.

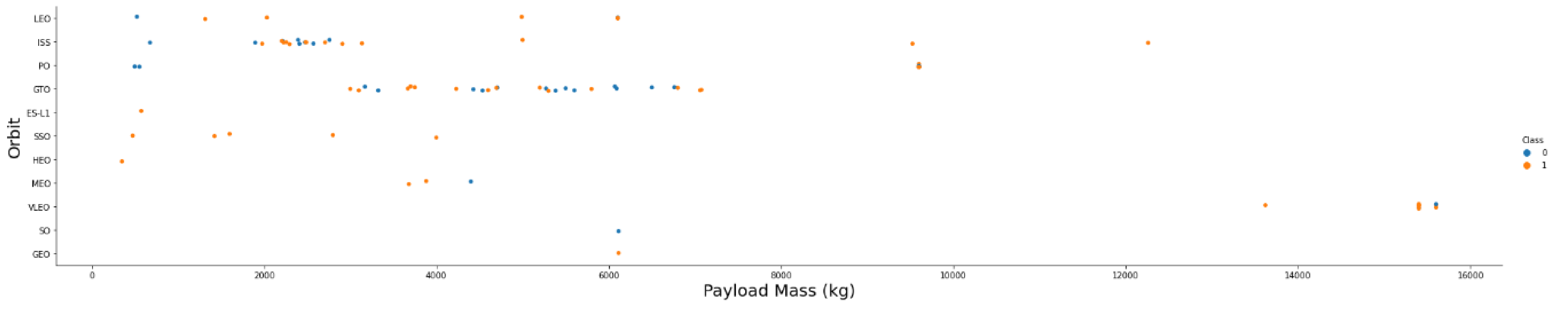


Flight Number vs. Orbit Type



- The beginning of the flights in the dataset were launched to the orbit LEO, ISS, PO and GTO;
- The last flights were launched to VLEO orbit;
- A change of behavior can be observed for launches from LEO, ISS, PO and GTO to VLEO
- The bigger concentration of blue dots in the left part than in the right part of the chart represents a reduction in the failure rate.

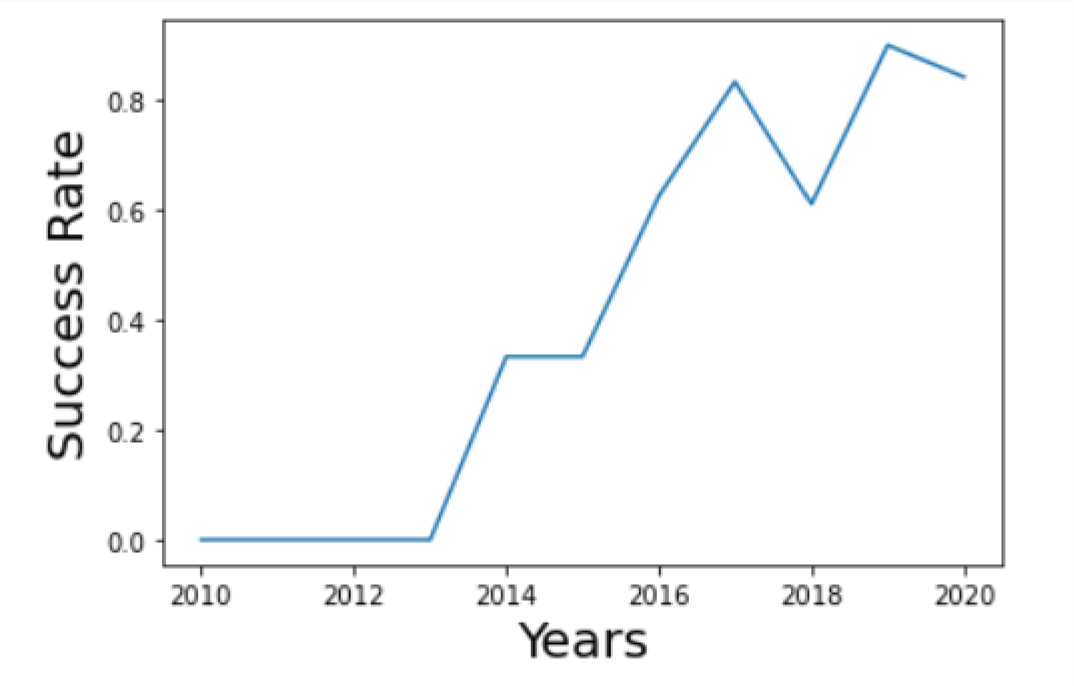
Payload vs. Orbit Type



- The correlations can be observed:
  - Orbit ISS with payload around 2000 kg;
  - Orbit GTO with payload around 3000 kg to 7000 kg;
  - Orbit VLEO with payload around 15000kg.

Launch Success Yearly Trend

- Low success rate until 2013;
- Increases from 2013 to 2017;
- Keep close to 0.8 from 2017 until 2020;
- The increment of the success rate can point to the maturation of the project based in the experienced launches, the advance of the methodology and technology used;
- The stabilization of the success rate near 0.8 can be associated with the saturation of the process used.



All Launch Site Names

- The query was used to find the names of the unique launch sites:

```
%sql select distinct launch_site from SPACEXTBL
```

- Result:

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

Launch Site Names Begin with 'CCA'

- The query was used to list the launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

- Result:

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

Total Payload Mass

- The query was used to calculate the total payload carried by boosters from NASA

```
%sql select customer, sum(payload_mass__kg_) from SPACEXTBL where customer='NASA (CRS)' group by customer
```

- Result:

NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The query was used to calculate the average payload mass carried by booster version F9 v1.1

```
%sql select booster_version, avg(payload_mass__kg_) from SPACEXTBL where booster_version ='F9 v1.1' group by booster_version
```

- Result:

F9 v1.1	2928

First Successful Ground Landing Date

- The query was used to find the dates of the first successful landing outcome on ground pad:

```
%sql select min(DATE) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)'
```

- Result:

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query was used to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select booster_version, payload_mass__kg_, "Landing _Outcome" from SPACEXTBL where "Landing _Outcome" = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 600
```

- Result:

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)

Total Number of Successful and Failure Mission Outcomes

- The query was used to calculate the total number of successful and failure mission outcomes:

```
%sql select mission_outcome, count(mission_outcome) from SPACEXTBL group by mission_outcome
```

- Result:

Failure (in flight)	1
Success	99

Boosters Carried Maximum Payload

- The query was used to list the names of the booster which have carried the maximum payload mass

```
%sql select distinct booster_version, payload_mass__kg_ from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

- Result:

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

2015 Launch Records

- The query was used to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

```
%sql select date, booster_version, launch_site, "Landing _Outcome" from SPACEXTBL where "Landing _Outcome" = 'Failure (drone ship)' and DATE like '2015%'
```

- Result:

DATE	booster_version	launch_site	Landing _Outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- The query was used to 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") from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' group by "Landing _Outcome" order by count("Landing _Outcome") desc
```

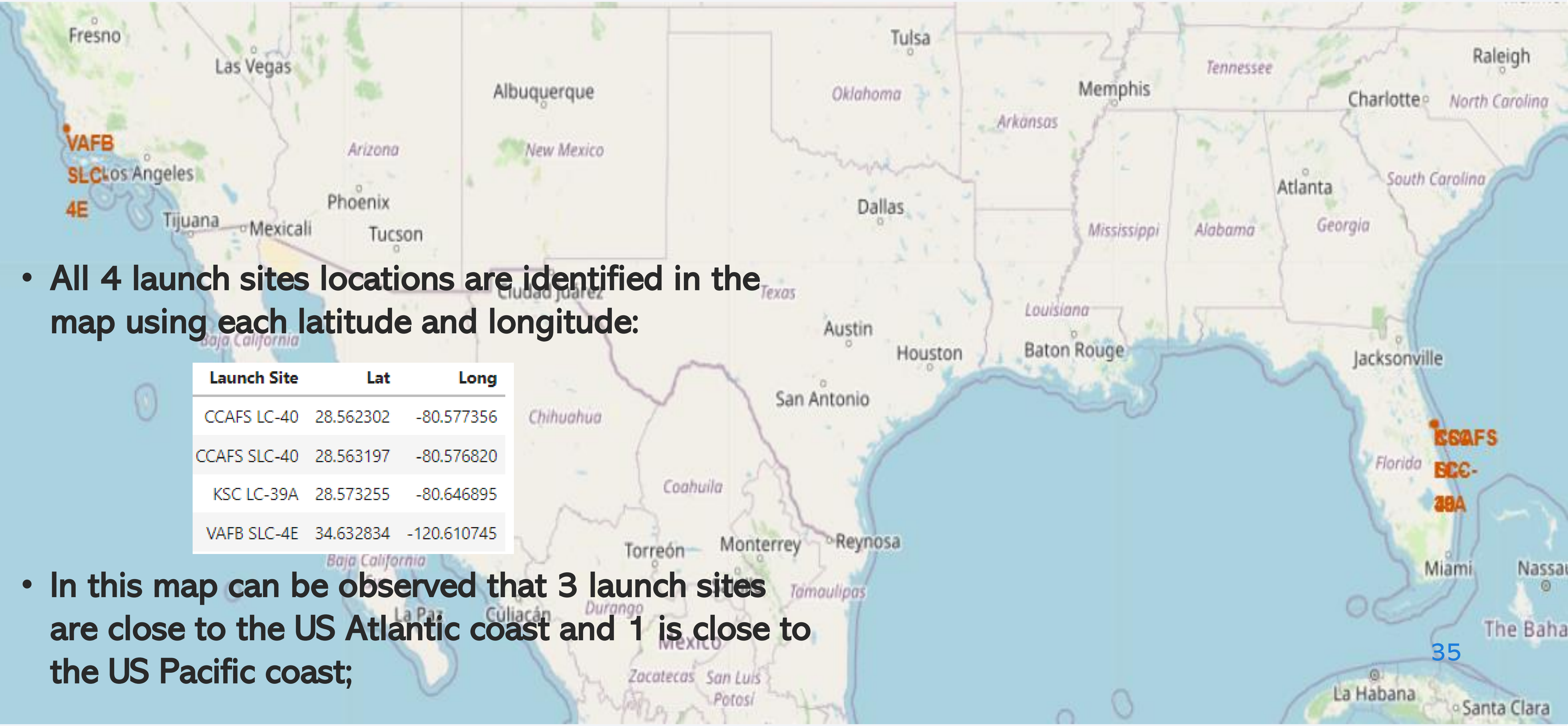
- Result:

Landing _Outcome	2
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

Section 3

Launch Sites Proximities Analysis

Launch site's location

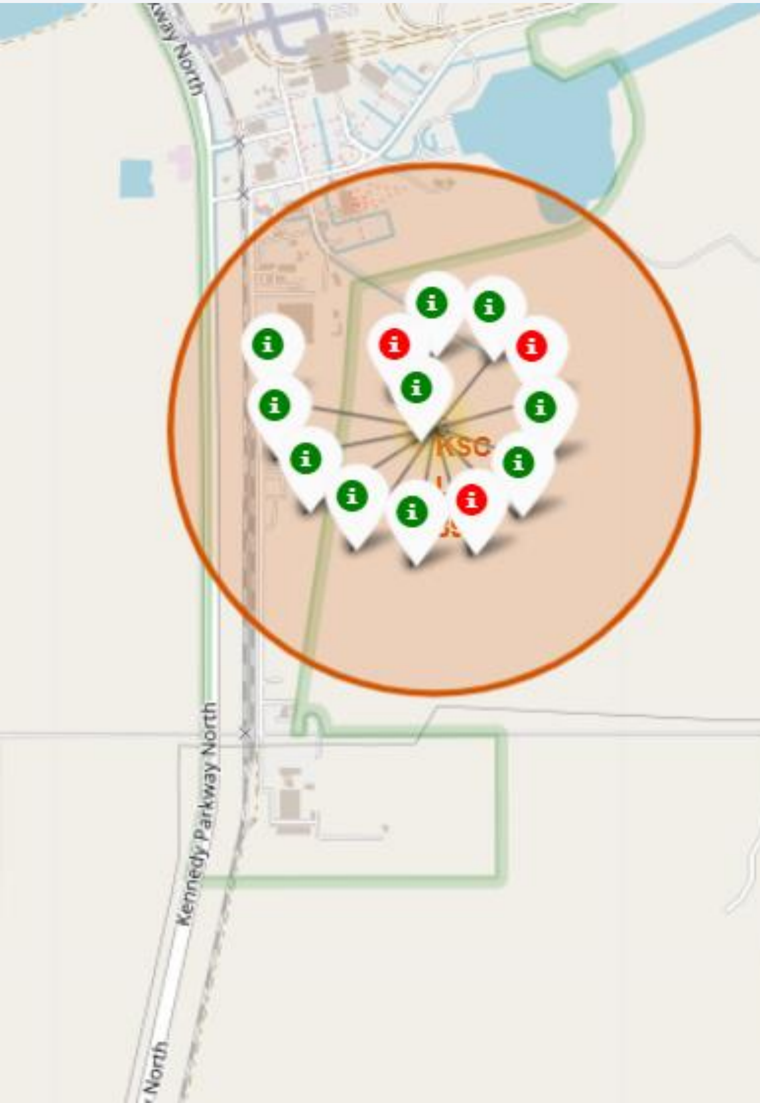


- **All 4 launch sites locations are identified in the map using each latitude and longitude:**

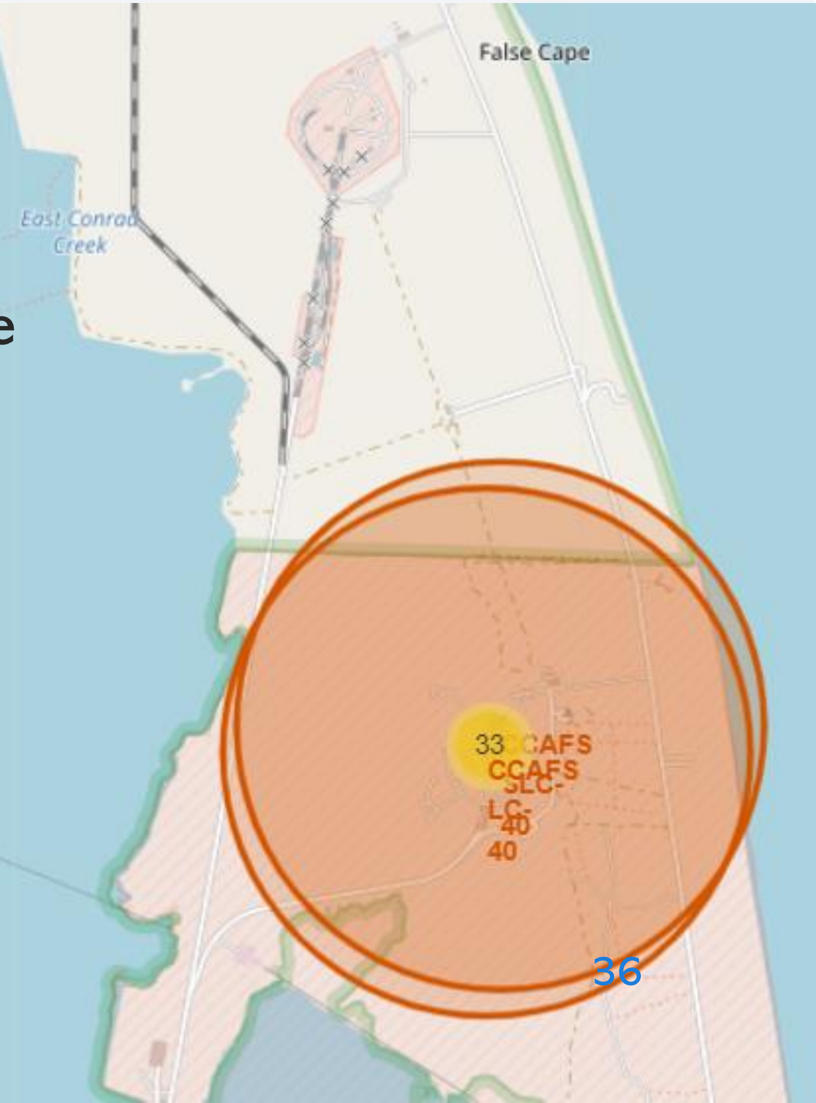
Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

- **In this map can be observed that 3 launch sites are close to the US Atlantic coast and 1 is close to the US Pacific coast;**

Color-labeled launch outcomes

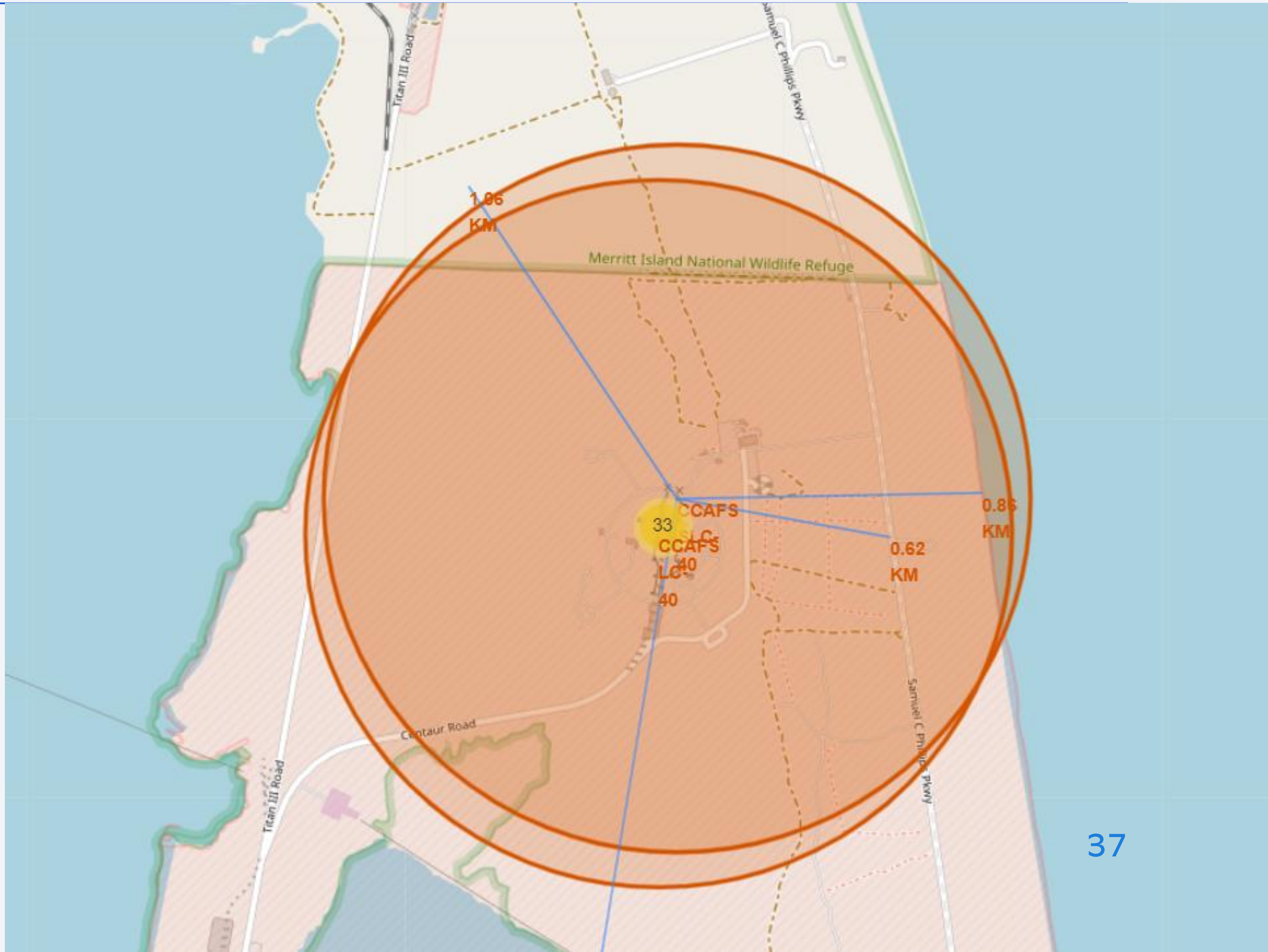


- In the map are identified the KSC LC-39, CCAFS LC-40 and SLC-40 launch sites;
- Around the KSC LC-39 launch site the launch outcomes are identified as follows:
    - Success (class=1), marker color green
    - Failure (class=0), marker color red
- This color-labeled method helps to visually identify the success and failure outcome in each launch sites.



Launch sites proximities

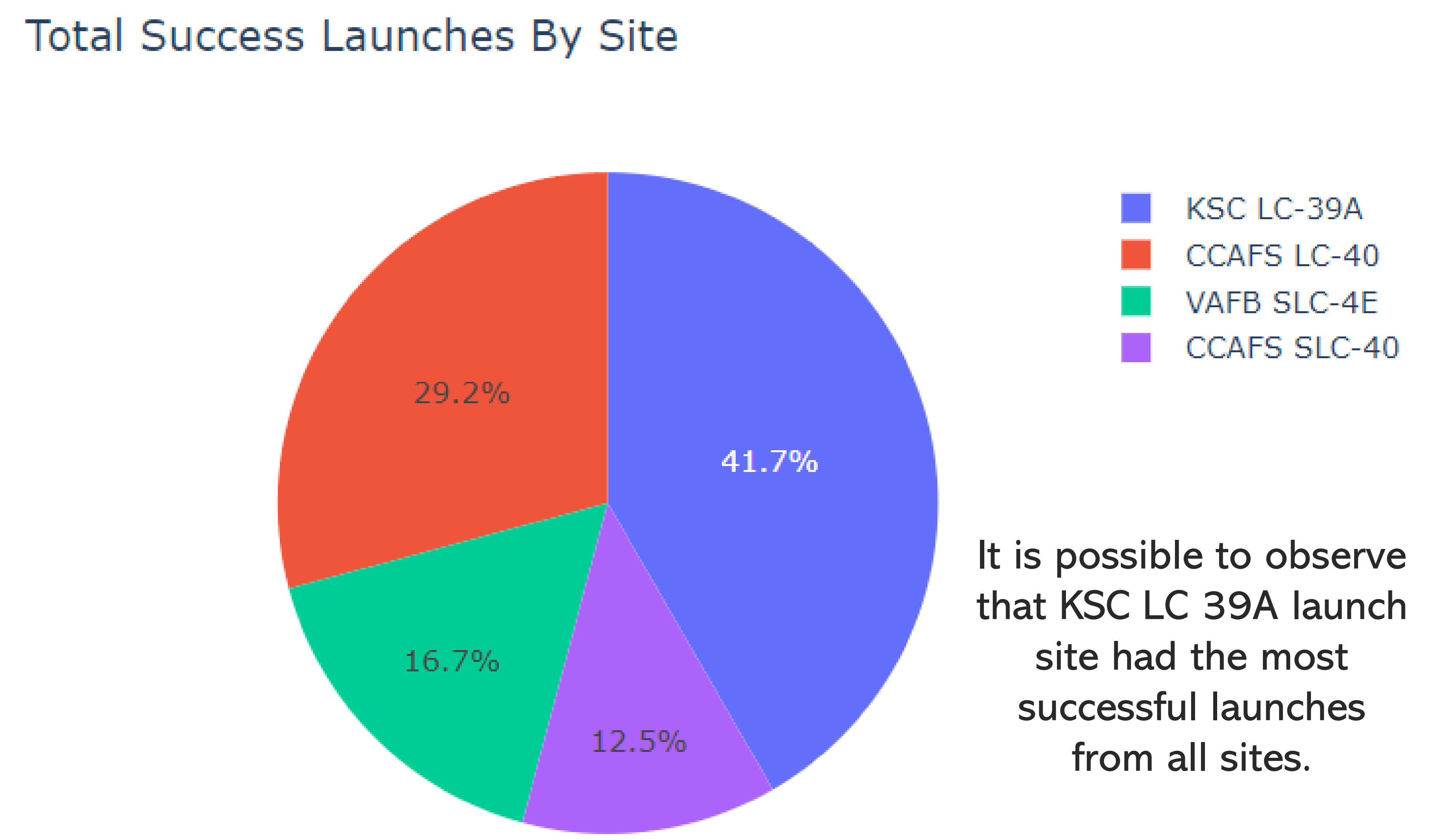
- In the map the surrounding of the CCAFS LC-40 and SLC-40 launch sites can be observed:
- The following nearby surrounding items were identified:
  - Coastline 0.86 km
  - Highway 0.62 km
  - Railroad 1.06 km
  - City 17.65 km
- The launch site keep a minimum safe distance from the items observed.



Section 4

Build a Dashboard with Plotly Dash

Total Success Launches By Site

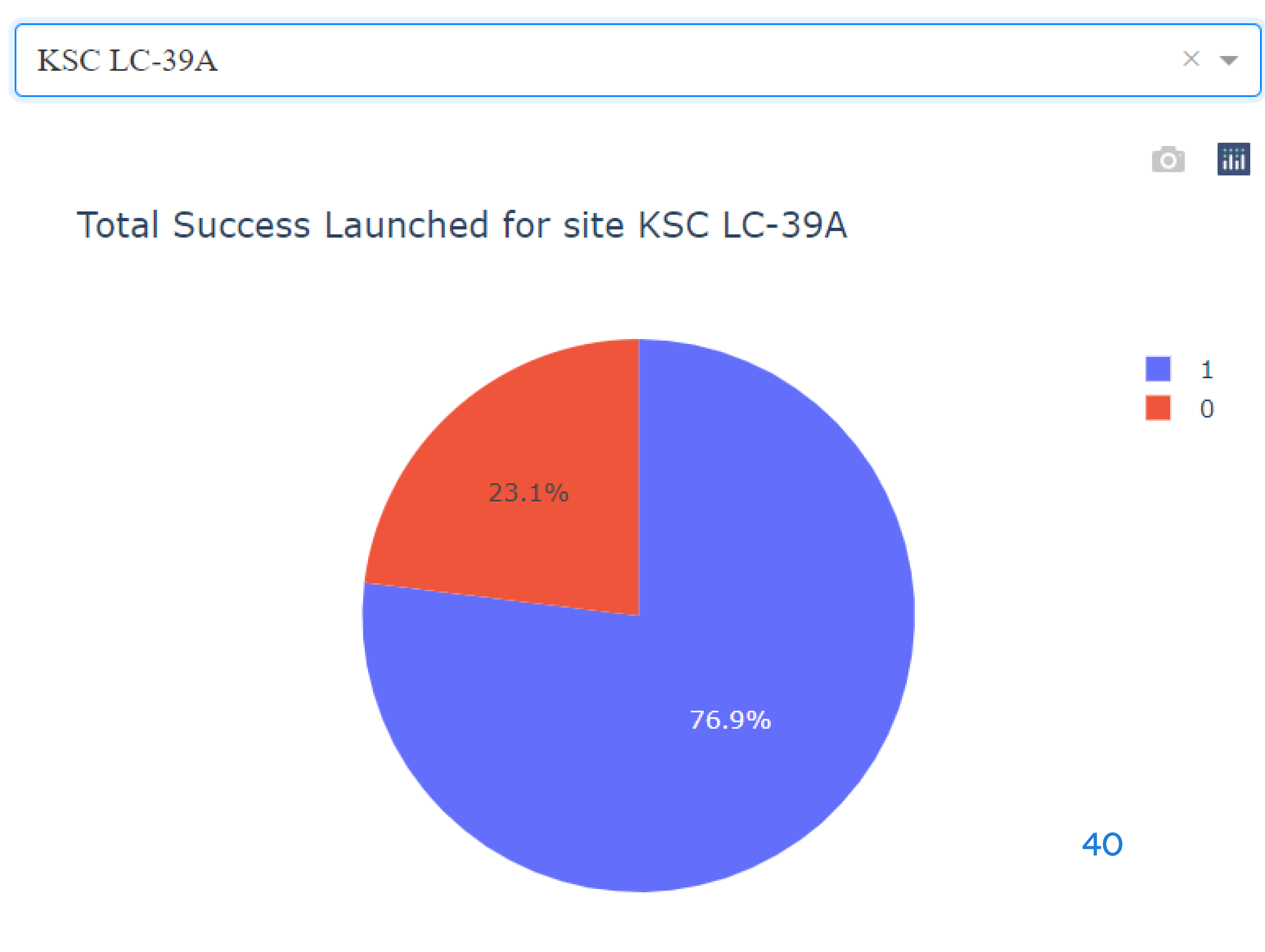


It is possible to observe that KSC LC 39A launch site had the most successful launches from all sites.

Total Success Launched for each site

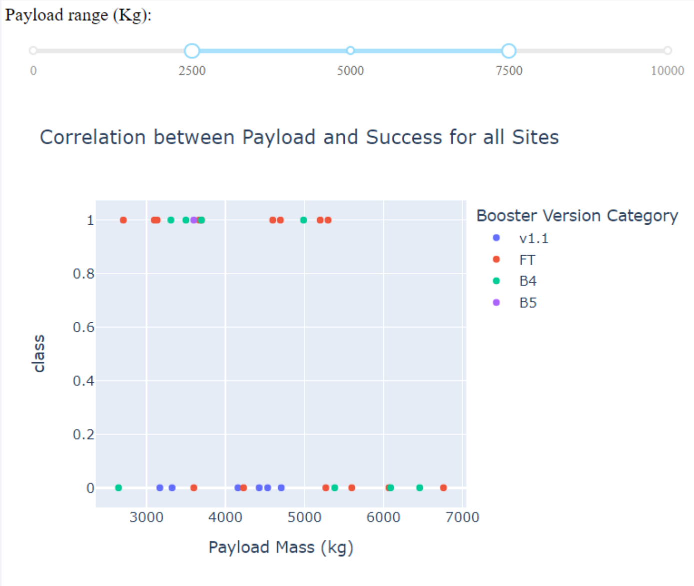
- Examining the pie chart with interactive behavior is possible to visually identify the high success ratio;

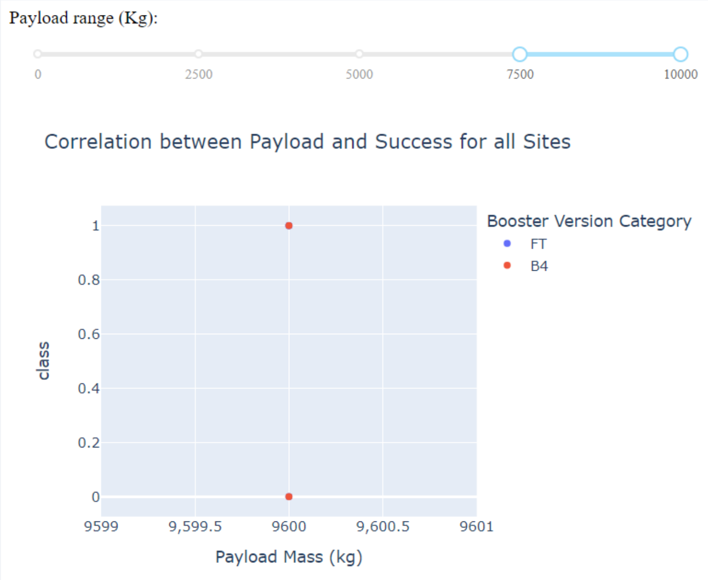
- The highest launch success ratio if for site KSC LC 39A.



Correlation Between Payload and Success for all sites

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider;



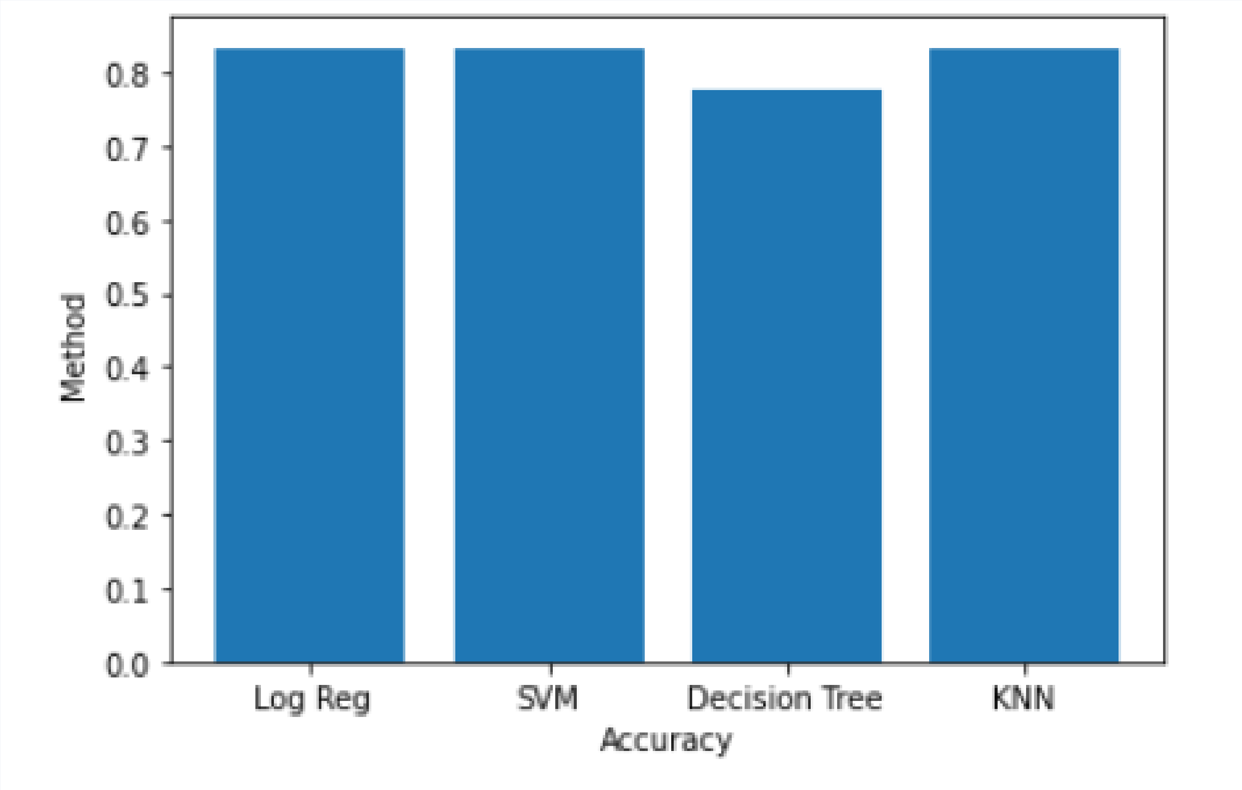


- This interactive analysis helps to identify the behavior of the dataset with the variation of the payload mass. For example, it is possible to observe that there is low data with high payload mass and success.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

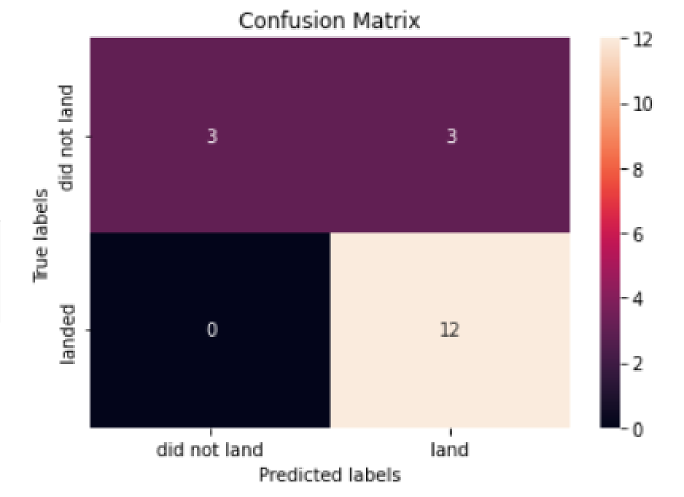


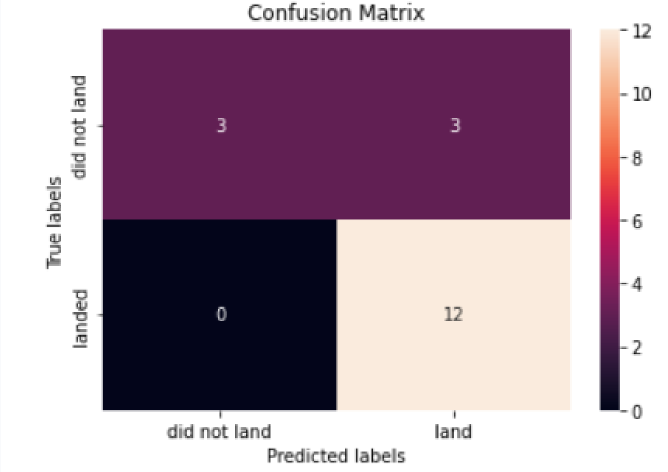

```
The score for log reg method is:  0.8333333333333334
The score for SVM method is:  0.8333333333333334
The score for Decision Tree method is:  0.7777777777777778
The score for KNN method is:  0.8333333333333334
```

Confusion Matrix

Log Reg

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



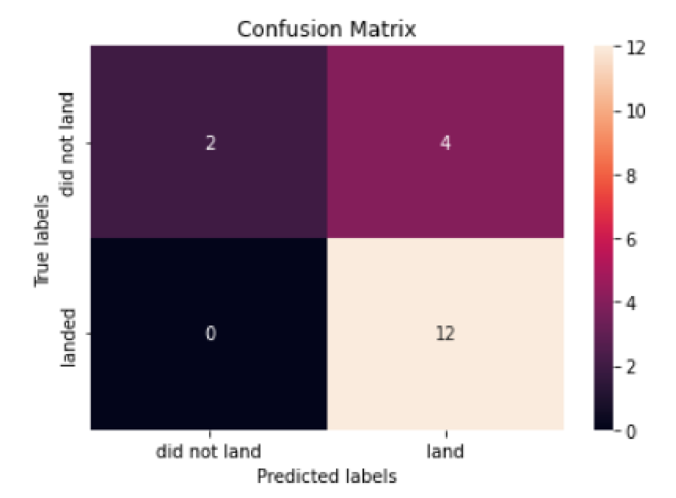


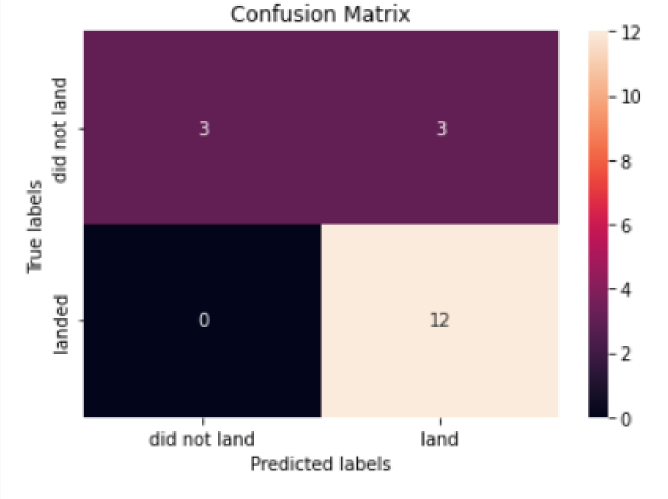
SVM

```
yhat=svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

Decision Tree

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```





KNN

```
yhat = Knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

Conclusions

- For the Exploratory Data Analysis of Space X launches we can observe:
  - The increment of the success rate can from 2013 to 2017 point to the **maturation** of the project based in the experienced launches, the advance of the methodology and technology used;
  - The stabilization of the success rate near 0.8 can be associated with the **saturation** of the process used. To keep improving this rate a review of the process could be necessary;
- Examining the confusion matrix of the predictive analysis method:
  - The methods used can distinguish well between the different classes.
  - The major problem is false positives.
- Answering the question: It is possible to determine if the first stage will land successfully?
  - **Yes**. Using Logistic Regression, SVM and KNN methods with accuracy of 83,33% of success in the prediction.
  - These methods are the **recommendation for the predictive analysis** for the dataset used.

Thank you!