IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniyah Qahtani

10\01\2024

Outline

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

Executive Summary

- To collect the data from website using API and Web Scraping method ,Then I performed some Exploratory Data Analysis some attributes could be used to determine if the first stage can be reused, and I perform interactive visual analytics using Folium and Plotly Dash and perform predictive analysis using classification models
- The launch success rate through years shows that it kept increasing since 2013 till 2020
- The relationship between Payload and Orbit type shows that with heavy payloads the successful landing or positive landing rate are more for Polar ,LEO and ISS.
- I found the KSC LC-39A launch site have the highest success rates
- The Decision Tree Classifier model has the best accuracy.

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a low cost, other providers cost upward of 165 million dollars each, that is because SpaceX can reuse the first stage. Therefore if I can determine if the first stage will land, I can determine the cost of a launch, so I will predict if the Falcon 9 first stage will land successfully by using machine learning .

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - data was collected by using API and Web Scraping method
- Perform data wrangling
  - I performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - I performed some Exploratory Data Analysis using a database and execute some SQL queries.
  - Also using visualization make me able to see directly how variables might be related to each other.

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
  - I used Folium to mark all launch sites on a map , Mark the success/failed launches for each site on the map and calculate the distances between a launch site to its proximities
  - I built a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.
- Perform predictive analysis using classification models
  - I built a machine learning to predict if the first stage of the Falcon 9 lands successfully, include: Preprocessing, Train_test_split and perform Grid Search. I also test Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.

Data Collection

Data was collected by using API and Web Scraping method

Collecting the Data with an API :

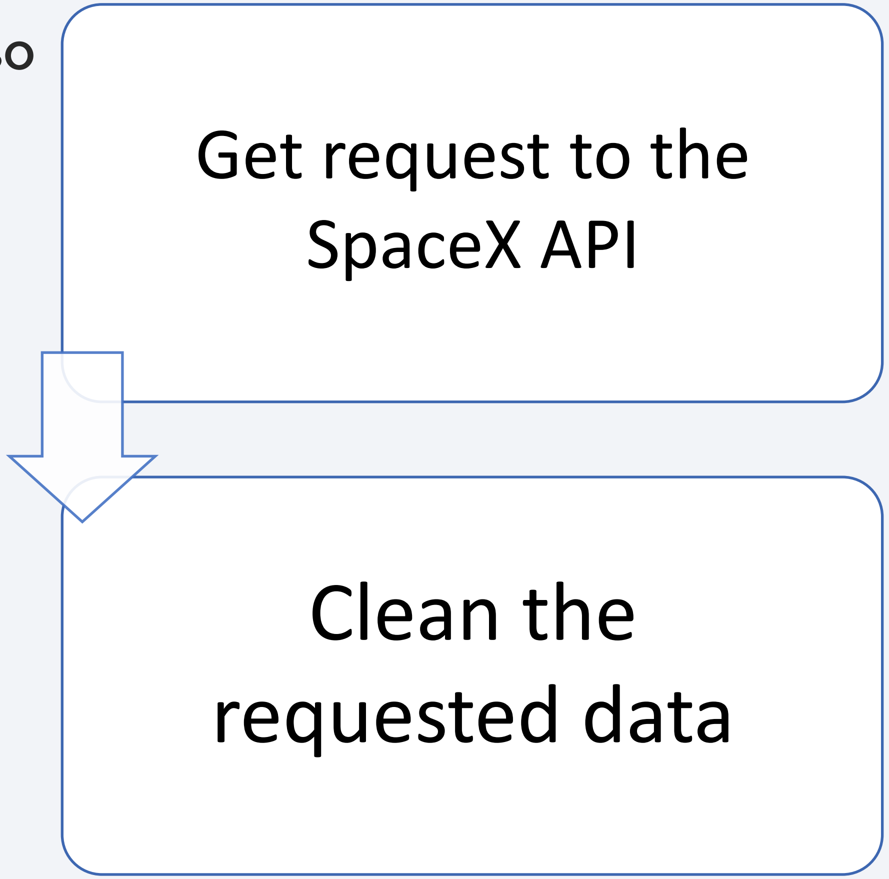
spacex_url=(https://api.spacexdata.com/v4/launches/past)

Collecting the Data with web scraping :

static_url=(https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API

- I made a get request to the SpaceX API ,and also did some basic data wrangling and formatting like dealing with missing values.



- The GitHub URL of the completed SpaceX API calls notebook (https://github.com/DaniyaQahtani/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb).

Data Collection - Scraping

- perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.
- Extract all column/variable names from the HTML table header
- Then collect all relevant column names from the HTML table header
- After that create a data frame by parsing the launch HTML tables

Extract a Falcon 9 launch records HTML table from Wikipedia

Parse the table and convert it into a Pandas data frame

- The GitHub URL of the completed web scraping notebook (https://github.com/DaniyaQahtani/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb)

Data Wrangling

- I perform exploratory Data Analysis and determine Training Labels.
- for example : True Ocean, False Ocean, True RTLS, False RTLS, True ASDS, False ASDS , I convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful

- The GitHub URL of your completed data wrangling related notebooks(https://github.com/DaniyaQahtani/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb)

EDA with Data Visualization

- I plotted a scatter point chart to visualize the relationship between Flight Number and Launch Site.
- I plotted a scatter point chart to visualize the relationship between Payload and Launch Site
- I plotted a bar chart to visualize the relationship between success rate of each orbit type
- I plotted a scatter point chart to visualize the relationship between FlightNumber and Orbit type
- I plotted a scatter point chart to visualize the relationship between Payload and Orbit type
- I plotted a line chart to visualize the launch success yearly trend

- The GitHub URL of your completed EDA with data visualization notebook (https://github.com/DaniyaQahtani/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb)

EDA with SQL

The SQL queries were 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

EDA with SQL

The SQL queries were performed:

- List the total number of successful and failure mission outcomes display 5 records where launch sites begin with the string 'CCA’
- List the names of the booster versions which have carried the maximum payload mass. Use a subquery
- 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.

- The GitHub URL of your completed EDA with SQL notebook(https://github.com/DaniyaQahtani/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb)

Build an Interactive Map with Folium

- I used folium.Circle to add a highlighted circle area with a text label on a specific coordinate, and folium.Marker for each launch site on the site map
- I also used marker_cluster to easily identify which launch sites have relatively high success rates.
- I draw a line between a launch site to its closest city.

- The GitHub URL of your completed interactive map with Folium map (https://github.com/DaniyaQahtani/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb)

Build a Dashboard with Plotly Dash

I have added a pie chart and scatter point chart to a dashboard.

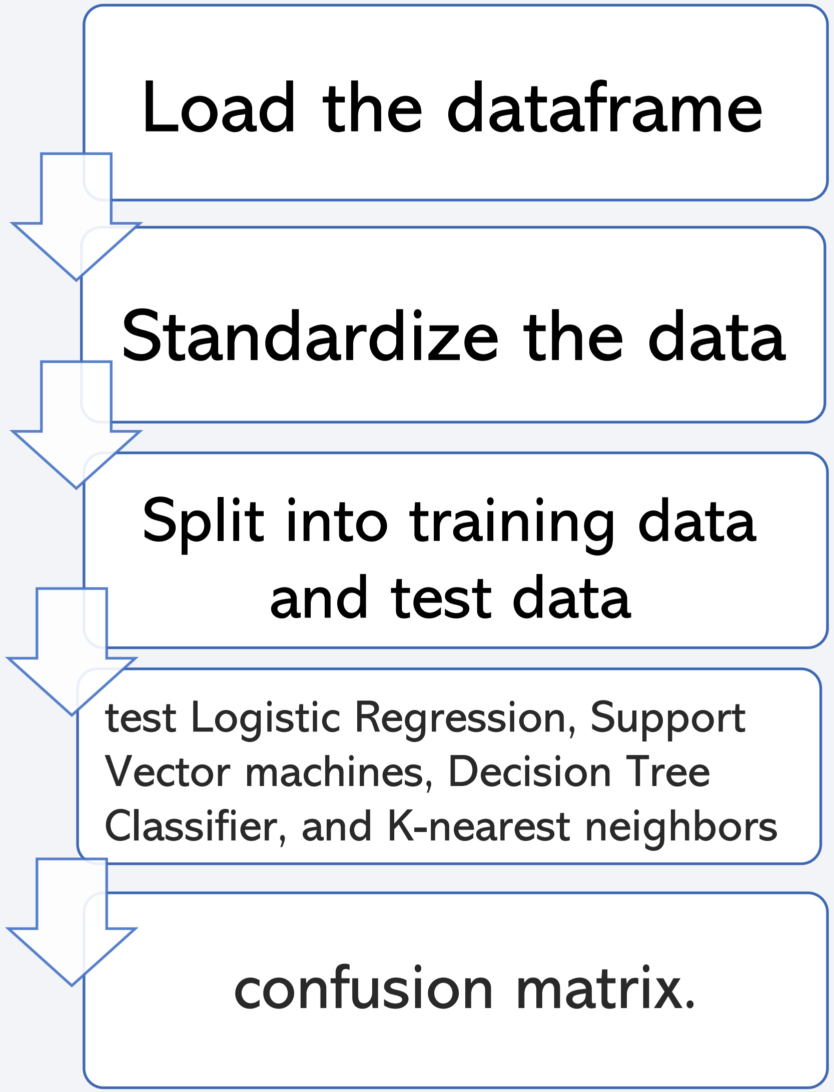
This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart ,and that help to find more insights from the SpaceX dataset more easily than with static graphs.

- Add the GitHub URL of your completed Plotly Dash lab, (https://github.com/DaniyaQahtani/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py)

Predictive Analysis (Classification)

I built a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully, included: Preprocessing, Train_test_split, I also trained the model and performed Grid Search to found the hyperparameters. Using the best hyperparameter values, I will determine the model with the best accuracy using the training data. Then I will test Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors. Finally, I will output the confusion matrix.

- Add the GitHub URL of your completed predictive analysis lab (https://github.com/DaniyaQahtani/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb)



Results

Exploratory data analysis results:

- 2015-12-22 Is the date when the first successful landing outcome in ground pad was achieved.
- F9 B5 B1048.4 Is the names of the booster versions which have carried the maximum payload mass.
- the relationship between Payload and Launch Site it is shows that for the KSC LC 39A launchsite there are no rockets launched for lightpayload mass (less than 2000)
- these orbit type GEO ,ES-L1 ,HEO and SSO have the highest success rate
- the relationship between Payload and Orbit type shows that with heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- the launch success rate through years shows that it kept increasing since 2013 till 2020

Results

Interactive analytics demo in screenshots:

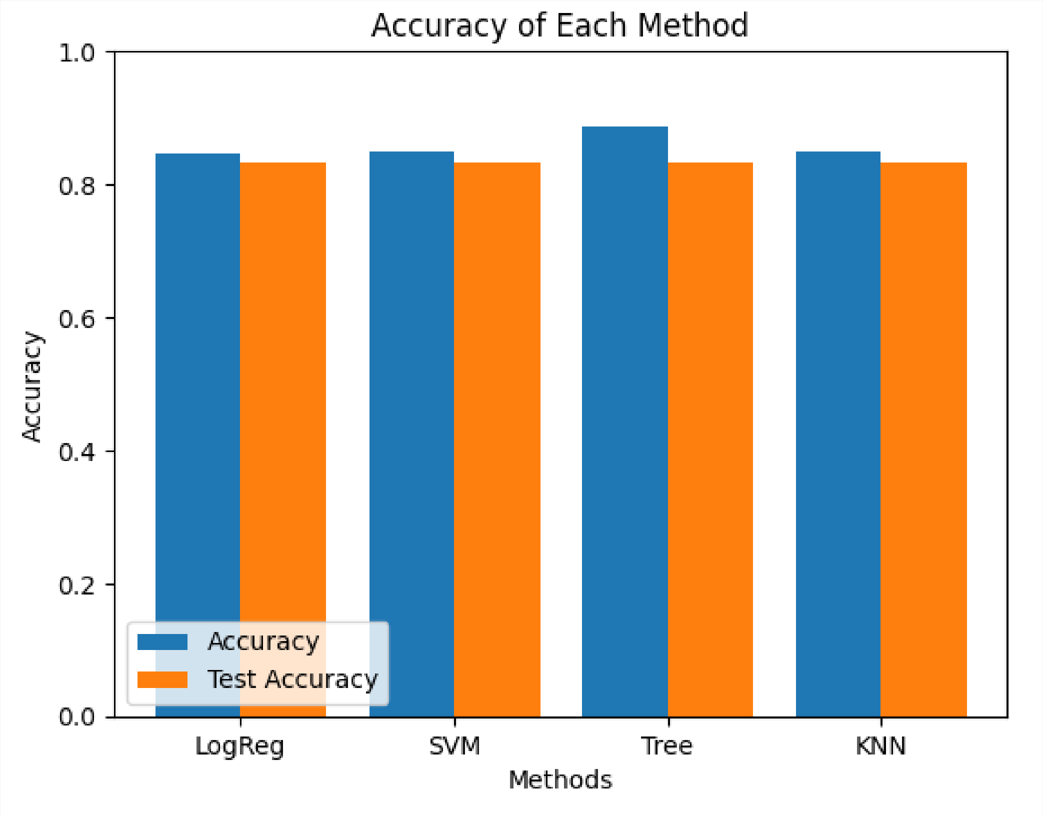
When I highlighted each launch site on the site map I found that (CCAFS LC-40, CCAFS SLC-40 , KSC LC-39A) located in the same side

Results

Predictive analysis results:

Here the bar chart shows that the Decision Tree Classifier model has the best accuracy.

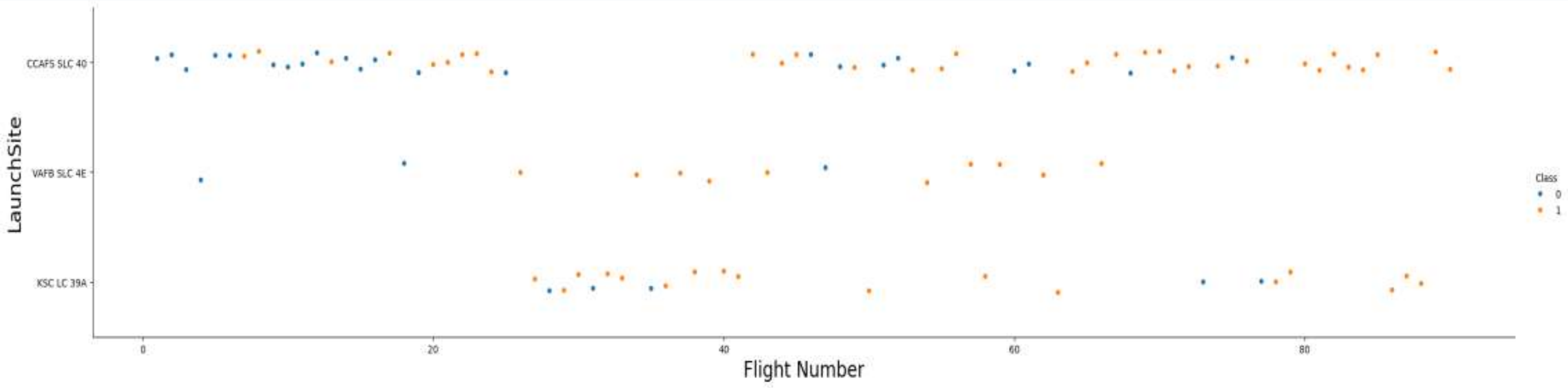
In the other hand Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors they have the same test accuracy value.



Section 2

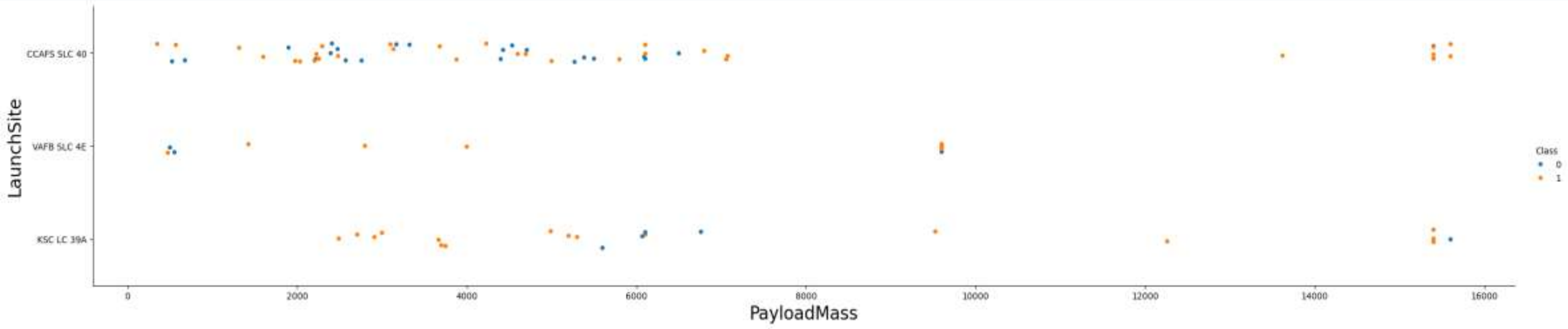
Insights drawn from EDA

Flight Number vs. Launch Site



- The plot shows that the general success rate improved over time and the best launch site is CCAFS SLC 40

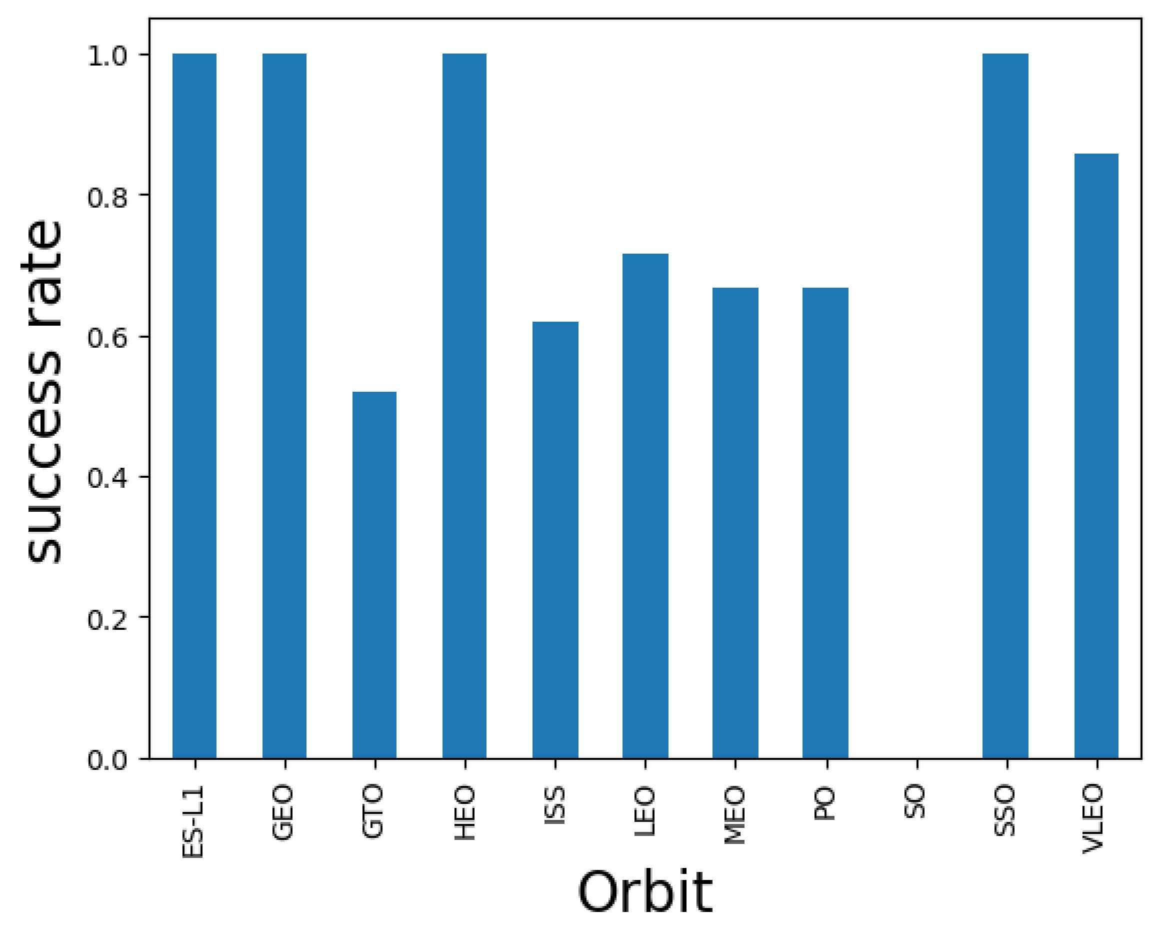
Payload vs. Launch Site



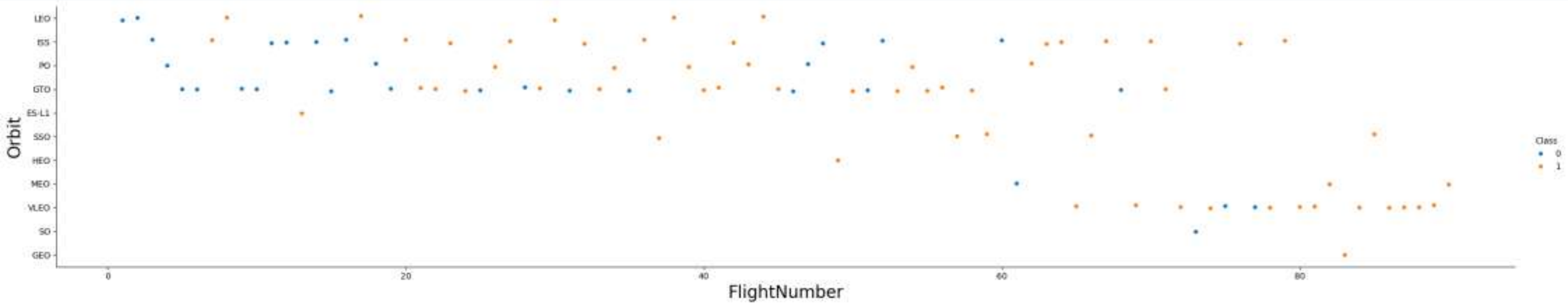
- The plot shows that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

The plot shows that the SO orbit has lowest success rate and the GEO ,ES-L1 ,HEO and SSO have the highest success rate.

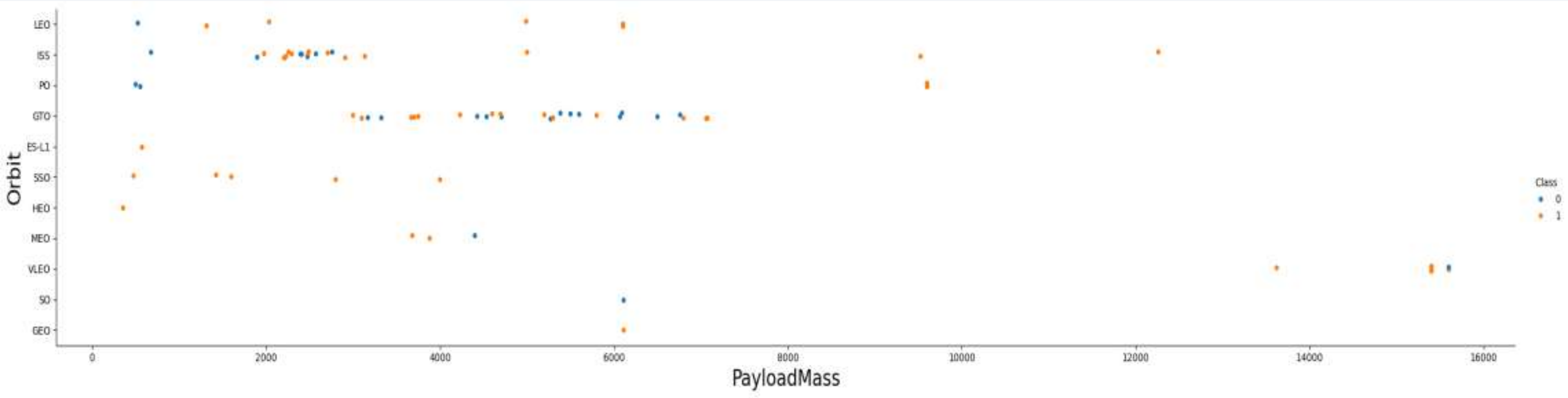


Flight Number vs. Orbit Type



- The plot shows 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.

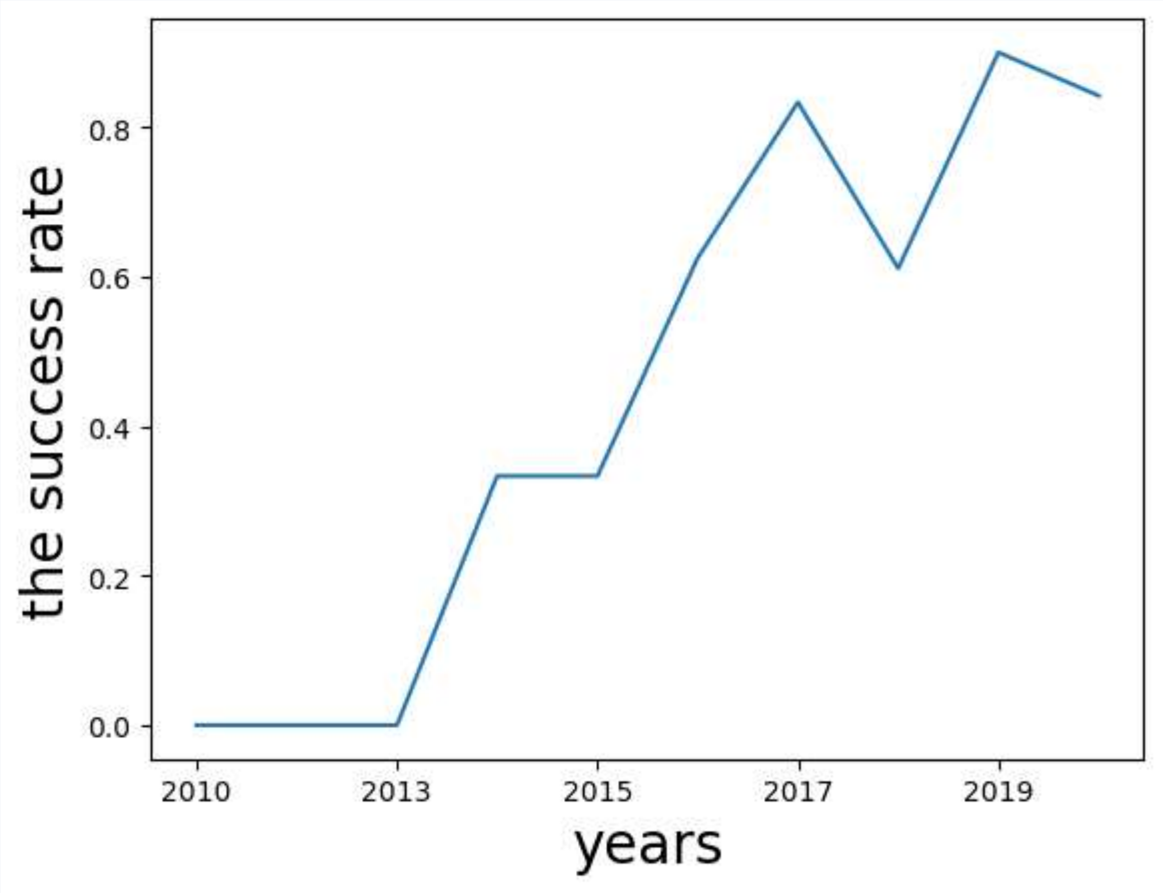
Payload vs. Orbit Type



- The plot shows that with light payloads the failure landing rate are more for Polar,LEO and ISS

Launch Success Yearly Trend

The launch success rate through years shows that it kept increasing since 2013 till 2020



All Launch Site Names

- Here are the names of the unique launch sites
- I used DISTINCT in the SQL query to find the unique launch sites

In [8]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

- Here the 5 records where launch sites begin with `CCA` ,the LIKE operator is used in where clause to search for specified pattern in column.

[9]:
```
%sql SELECT *  FROM SPACEXTBL WHERE Launch_Site LIKE'CCA%'LIMIT(5)
```

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

[9]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

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

```
In [10]: %sql SELECT SUM (PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Payload like '%CRS%'

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

Out[10]:

SUM (PAYLOAD_MASS__KG_)
111268

- the total payload carried by boosters from NASA is 111268 kg

Average Payload Mass by F9 v1.1

In [11]:
```
%sql SELECT AVG (PAYLOAD_MASS__KG_)  FROM SPACEXTBL WHERE Booster_Version='F9 v1.1'
```

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

Out[11]:

AVG (PAYLOAD_MASS__KG_)
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg

First Successful Ground Landing Date

In [12]:
```
%sql SELECT  MIN (Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Out[12]:

MIN (Date)
2015-12-22

- 2015-12-22 Is the date when the first successful landing outcome in ground pad was achieved .

Successful Drone Ship Landing with Payload between 4000 and 6000

In [13]:

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Su
```

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

Out[13]:

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

- Here are the names of boosters which have successfully landed on drone ship and had payload mass between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

In [14]:

```
%sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTBL GROUP BY Mission_Outcome ORDER BY Mission_Outcome;
```

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

Out[14]:

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

- Here is the total number of successful and failure mission outcomes
- The total number of successful mission outcomes is 100
- The total number of failure mission outcomes is 1

Boosters Carried Maximum Payload

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

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

2015 Launch Records

```
In [54]:  %sql SELECT SUBSTR(Date, 6,2)as month,SUBSTR(Date,0,5)as year, Booster_Version,Launch_Site,Landing_Outcome FROM SPACEXTBL wl

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

Out[54]:

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

- This are the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 in January and April

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

- Here is 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

Out[36]:

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

Section 3

Launch Sites Proximities Analysis

All launch sites on a map

The map shows that CCAFS LC-40, CCAFS SLC-40 , KSC LC-39A launch sites located in the same side of the map where is the VAFB SLC-4E launch site in the other side of the map

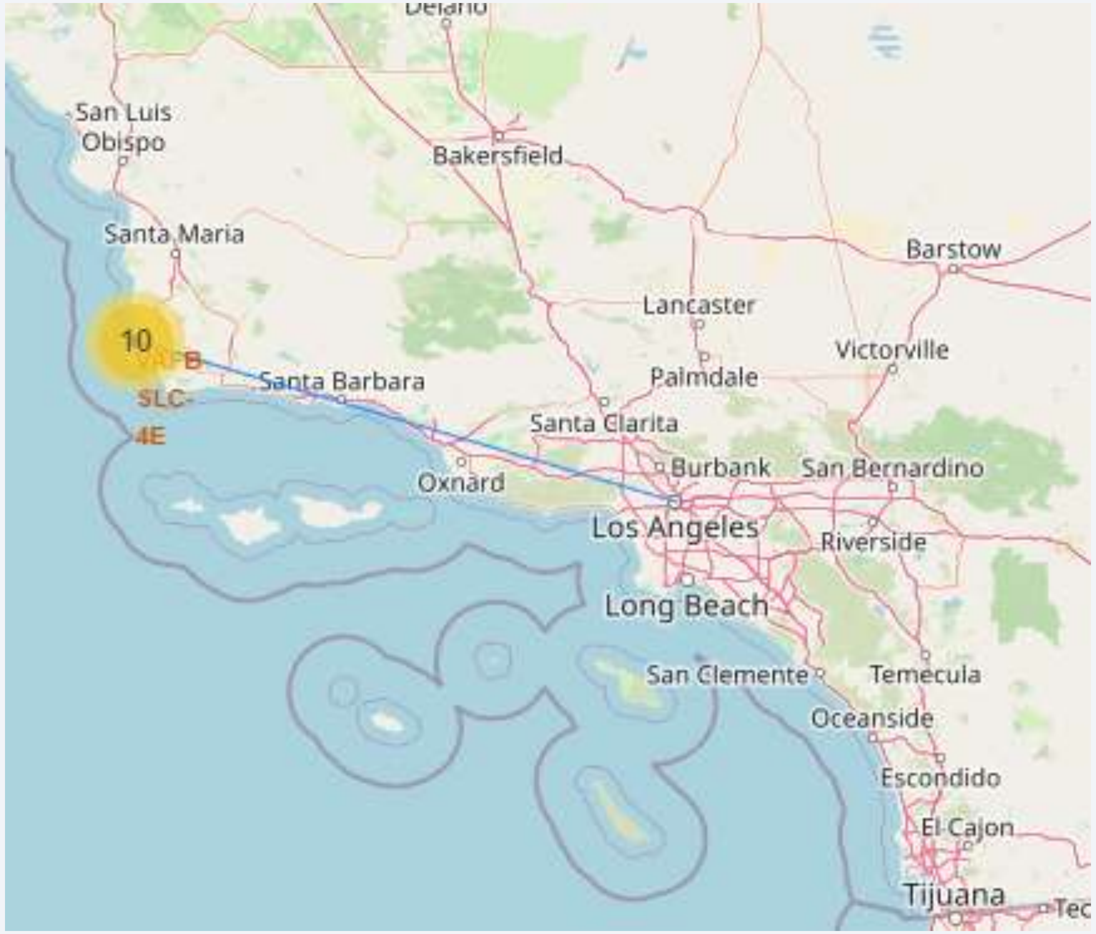
The success/failed launches for each site

After adding the launch outcomes for each site, I found the KSC LC-39A launch site have the highest success rates

The green marker means launch was successful ,and the red marker means launch was failed

launch sites and its proximities

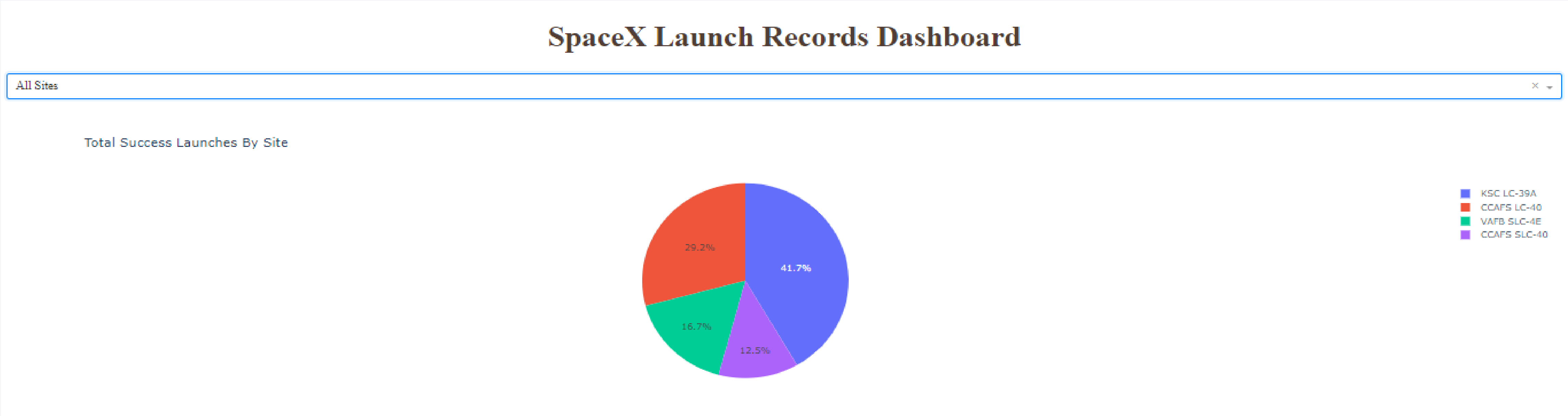
- I draw a line between a VAFB SLC-4E launch site to its closest city Los Angeles.



Section 4

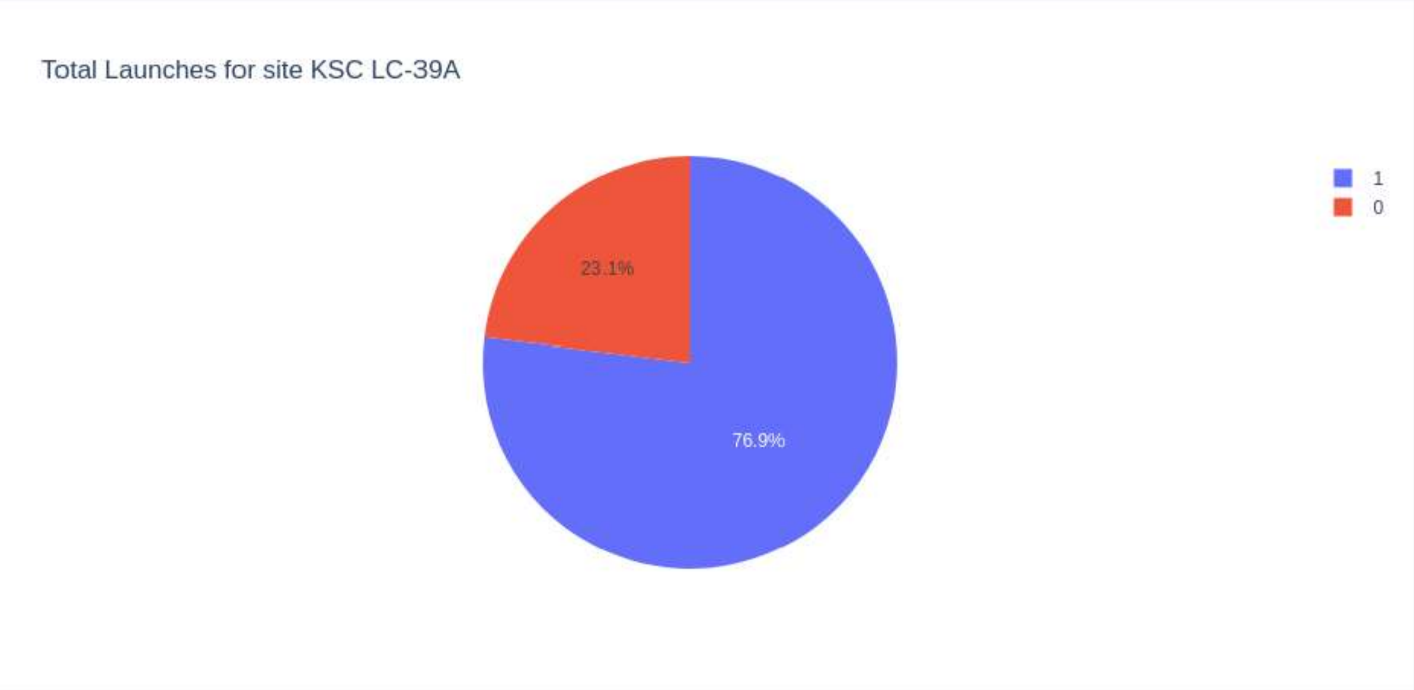
Build a Dashboard with Plotly Dash

launch success count for all sites



- Pie chart shows that KSC LC-39A launch site has 41.7% which is the highest success launch

The launch site with highest launch success ratio



- KSC LC-39A launch site with highest launch success ratio about 76.9%

Payload vs. Launch Outcome for all sites

Payloads under 6000kg and FT booster are the most successful combination

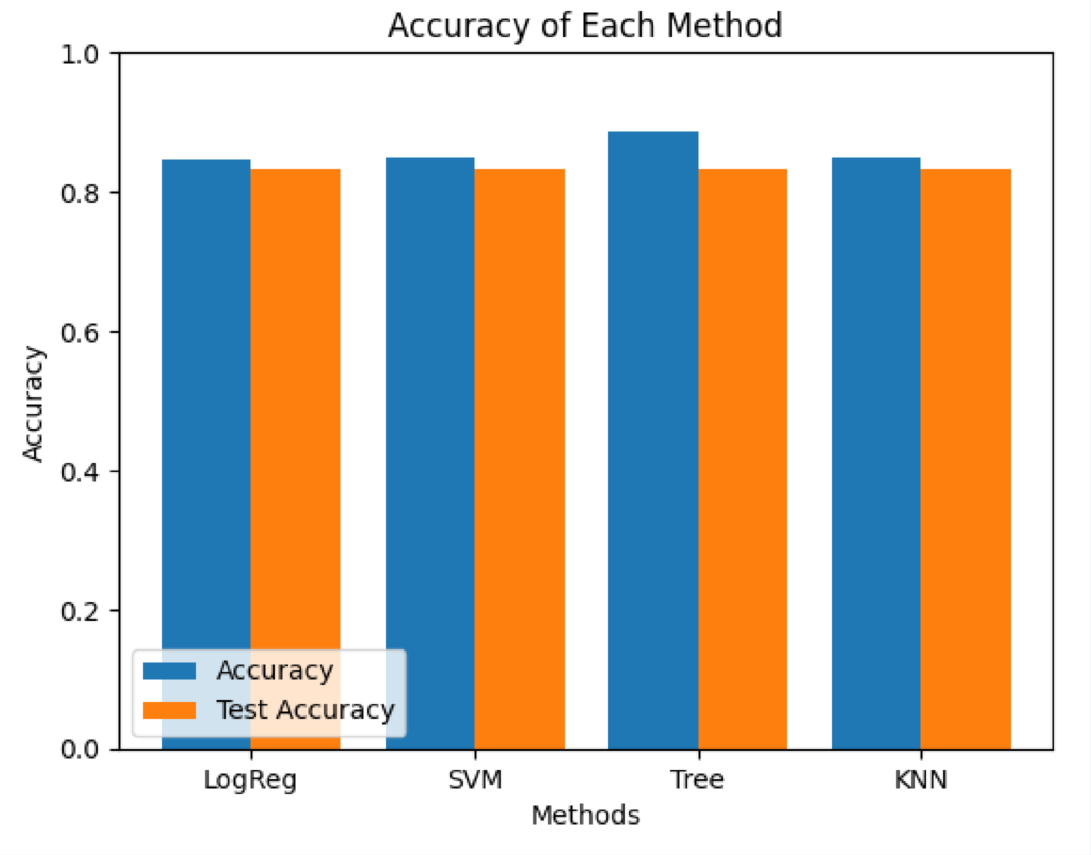
Section 5

Predictive Analysis (Classification)

Classification Accuracy

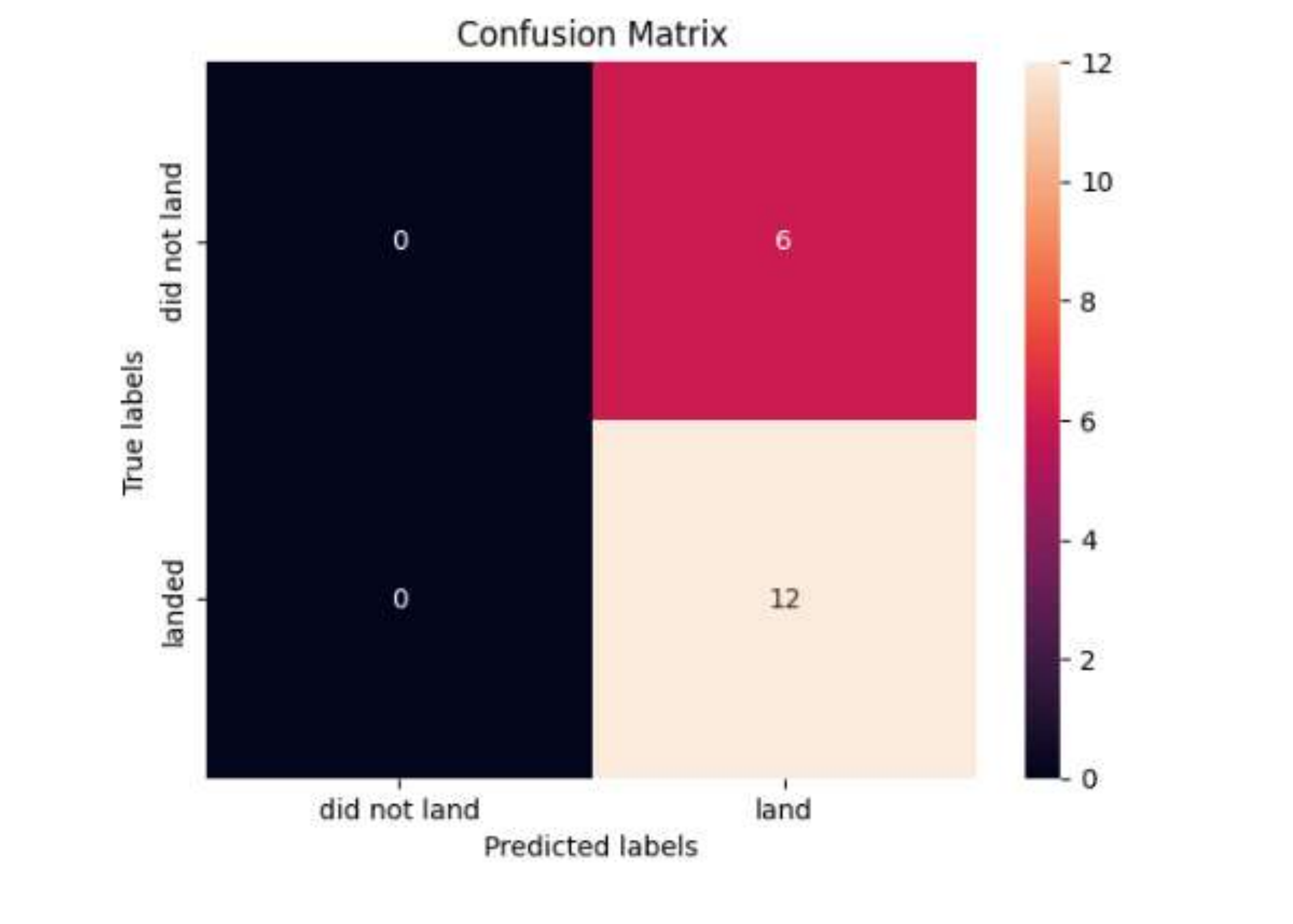
Here the bar chart shows that the Decision Tree Classifier model has the best accuracy.

In the other hand Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors they have the same test accuracy value.



Confusion Matrix

- This is the confusion matrix of the best performing model which is Decision Tree Classifier



Conclusions

- In this project I want to determine if the first stage will land successfully, and that depends on many factors such as payload mass, orbit type, and also location and proximities of a launch site and so on .

- I create a machine learning pipeline to predict if the first stage will land I find the Decision Tree Classifier model has the best accuracy.

Appendix

```python
# HINT: Use get_dummies() function on the categorical columns
features_one_hot=pd.get_dummies(features,columns=['Orbit', 'LaunchSite','LandingPad', 'Serial'], drop_first=False)
features_one_hot.head()
```

```sql
%%sql SELECT SUBSTR(Date, 6,2)as month,SUBSTR(Date,0,5)as year, Booster_Version,Launch_Site,Landing_Outcome
FROM SPACEXTBL where Landing_Outcome= 'Failure (drone ship)'and SUBSTR(Date,0,5)="2015"
```

Thank you!