SPACEX Rocket Launch

Descriptive and Predictive Analysis

Shailja Raghuvanshi
16/03/2025

Outline

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

Executive Summary

Summary of methodologies

- Data Collection (Through SpaceX API and Web Scrapping)
- Data Wrangling (Cleaning the data, handling the missing values in dataset)
- Exploratory Data Analysis with Data Visualization.
- Exploratory Data Analysis with SQL.
- Performing interactive visual analytics with Folium and Plotly Dash.
- Predictive analysis (classification models such as Logistic Regression, Decision Tree and others).

Summary of all results

- Exploratory Data Analysis results.
- Static and interactive Data visualization such as bar charts, scatter plot, maps and dashboard.
- Predictive data analysis with the help of classifier models such as Logistic Regression and others.

Introduction

- **Project background and context**

- SpaceX , a successful commercial space trips provider, advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; while other providers cost upward of 165 million dollars each, the much difference 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. Based on public information available on the launched rockets in the past and their successful/unsuccessful landing, we can analyze the factor suppose to determine the outcome of a launch and also predict the outcome of the landing based on machine learning models.

- **Problems you want to find answers**

- How do different variables such as payload mass, launch site, number of flights , booster version , orbits affect the success of first stage landing.

- The best classification machine learning model which predicts the actual outcome of past landings with minimum error.

Section 1

Methodology

Methodology

Data collection methodology:

Using SPACEX API

Using web scrapping of data available on Wikipedia.

Perform data wrangling

Filtering the dataset to get the required data

Dealing with missing values present in our filtered data.

Using One Hot Encoding to prepare data to make it suitable for binary classification.

Performed exploratory data analysis (EDA) using different visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

Building, tuning and finally evaluating different classification model to get the best model for landing outcome classification.

Data Collection

- To obtain the required data on SpaceX rocket launches and their landing outcomes, both **SpaceX API** and **web scrapping** from Wikipedia were used in order to get complete needed information about the launches to carry on our detailed and predicative analysis.

- After calling the SpaceX API we get our dataset of consisting columns as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude and Latitide.

- By Web scrapping the Wikipedia page on past SpaceX launches, we received additional information in column as Payload, Customer and others.

Data Collection – SpaceX API

Requesting rocket launch data from SpaceX API.

As the response received from the API call is in json format so create a dataframe out of this response object by .json_normalize() method.

Applied some custom function to get the desired column values as the response we received having IDs as rocket Id, launchpad ID etc.

After calling the function on the data received through API call, we converted the data into dictionary and then created a dataframe ‘df’ from that dictionary.

As we are only interested in the landing outcome of Falcon 9 rocket so we filtered our newly created dataframe to have only records related to Falcon9.

We dealt with the missing values as and when needed, for example replacing the missing values in PayloadMass column with the mean value of that particular column.

GitHub URL: https://github.com/Shailjasharma28/SpaceX-Analysis--Coursera

Data Collection – Web Scraping

Requesting Falcon 9 rocket launch data from Wikipedia

Creating a BeautifulSoul object from the HTML response.

Extracting the column names from the received HTML table header.

Collecting the data by parsing the HTML tables.

Create a dictionary out of the data received.

Create a Dataframe from the dictionary.

Export the data to a CSV file.

GitHub URL: https://github.com/Shailjasharma28/SpaceX-Analysis--Coursera

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

Now we will convert those outcomes into Training Labels with **1 means the booster successfully landed 0 means it was unsuccessful.**

Perform Exploratory Data Analysis →

Calculate the number of launches on each site.

Calculate the number of occurrence of each orbit.

Calculate the number and occurrence of mission outcome of all the orbits.

Determining training labels →

Create a landing outcome label from outcome column.

↓

Exporting the data to CSV..

GitHub URL: https://github.com/Shailjasharma28/SpaceX-Analysis--Coursera

EDA with Data Visualization

Several charts were plotted for exploratory data analysis.

- Scatter plot to visualize the relationship between Payload Mass and Launch site.
- Bar chart to visualize the success rate of each orbit.
- Scatter plot to visualize the relationship between FlightNumber and Orbit type.
- Scatter plot for Payload Mass and orbit type.
- Yearly trend line chart for success.

All these visualization for numerical and categorical variable helped us in identifying the relationship between our columns and yearly wise trend of success rate.

GitHub URL: https://github.com/Shailjasharma28/SpaceX-Analysis--Coursera/blob/main/Exploratory%20Visulization.ipynb

EDA with SQL

SQL queries:

- Display the name of the unique launch sites.
- Displaying top 5 records of the launch sites with name begins with ‘CCA’.
- Displaying the total payload mass carried by booster launched by NASA.
- Displaying average payload mass carried by booster version F9 v1.1.
- Listing the date where the first successful landing outcome was in ground pad.
- List the name of the booster which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure outcome.
- Listing the name of the booster version which have carried the maximum payload mass.
- Listing all the failed landing outcomes in drone ship, their booster version and launch site name.
- Ranking the landing outcomes between the date 2010-06-04 and 2017-03-20.

GitHub URL: https://github.com/Shailjasharma28/SpaceX-Analysis--Coursera/blob/main/Exploratory%20Visulization.ipynb

Build an Interactive Map with Folium

Markers of all Launch sites:

- Added marker with circle, popup label and text of NASA Johnson space center using its longitude and latitude as a starting coordinates in the folim map.
- Added marker with popup lebel and text label of all the four launch sites by using their coordinates.

Coloured markers of the launch outcomes for each launch site.

- Added colored markers of success (green) and failure (red) landing and also identified about which site has the maximum successful rate.

Distance between a launch site and its closest proximities like railway, highway , coastline and closet city.

GitHub URL:https://github.com/Shailjasharma28/SpaceX-Analysis/Coursera/blob/main/lab_jupyter_launch_site_location_with_Folium.ipynb

Build a Dashboard with Plotly Dash

Launch Site dropdown List:

- Added a dropdown list to enable launch site selection between all the four sites or all sites.
- Added a pie chart to show the successful launches count for all sites and also success vs failed counts for the site , if a specific site were chosen.

Slider of payload mass:

- Added a slider to select payload mass.

Scatter plot of payload mass and success rate for the different version.

GitHub URL:https://github.com/Shailjasharma28/SpaceX-Analysis--Coursera/blob/main/SpaceX%20Dashboard%20with%20Dash.py

Predictive Analysis (Classification)

Creating a NumPy array from the column “Class” in dataframe.

Standardizing the data with StandardScaler, then fitting and transforming the values.

Splitting the dataset into training and testing set with train_test_split function.

Creating a GridSearchCv object with cv=10 to find the best parameters for various model..

-Applying GridSearchCV on all the classifier models- LogReg, SVM, DecisionTree and KNN model.

-Exmining the confusion matrix for all four trained model and Jaccard_score and F1_score metrices.

Calculating the accuracy score on the test data using the method.best_score_ for all models.

https://github.com/Shailjasharma28/SpaceX-Analysis--Coursera/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

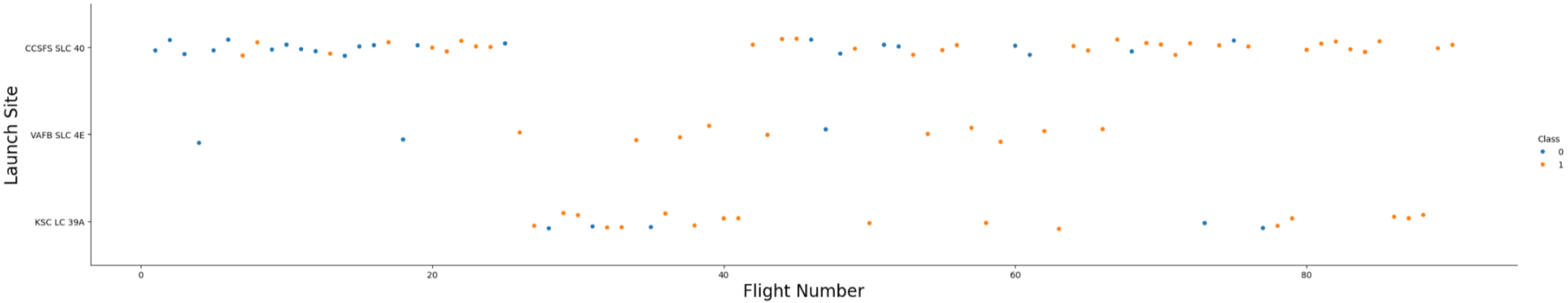
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

With more flights there are more successful outcomes for all three launch sites.

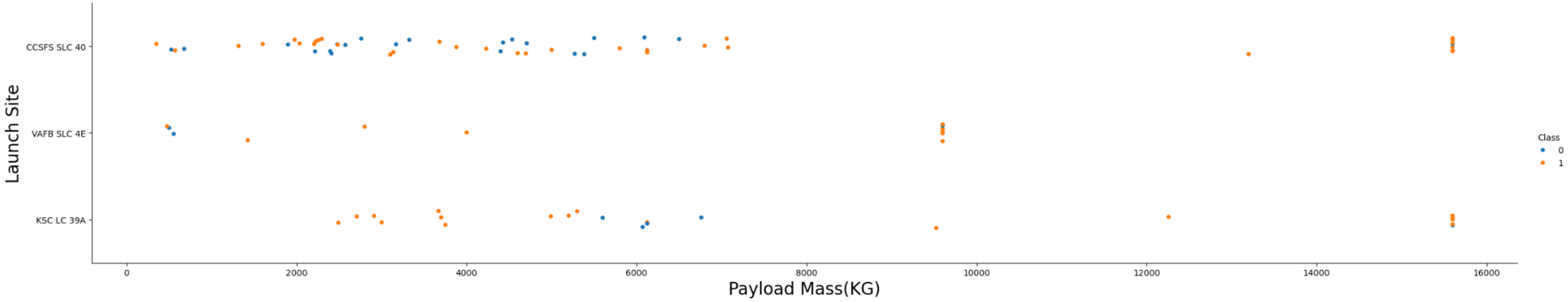
```python
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=data_falcon9, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



Payload vs. Launch Site

- For payload mass of 10,000 kg and more there are more successful launches from all three sites.
- The maximum payload mass launched from VAFB SLC 4E is 10,000 Kg.

```
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=data_falcon9, aspect = 5)
plt.xlabel("Payload Mass(KG)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



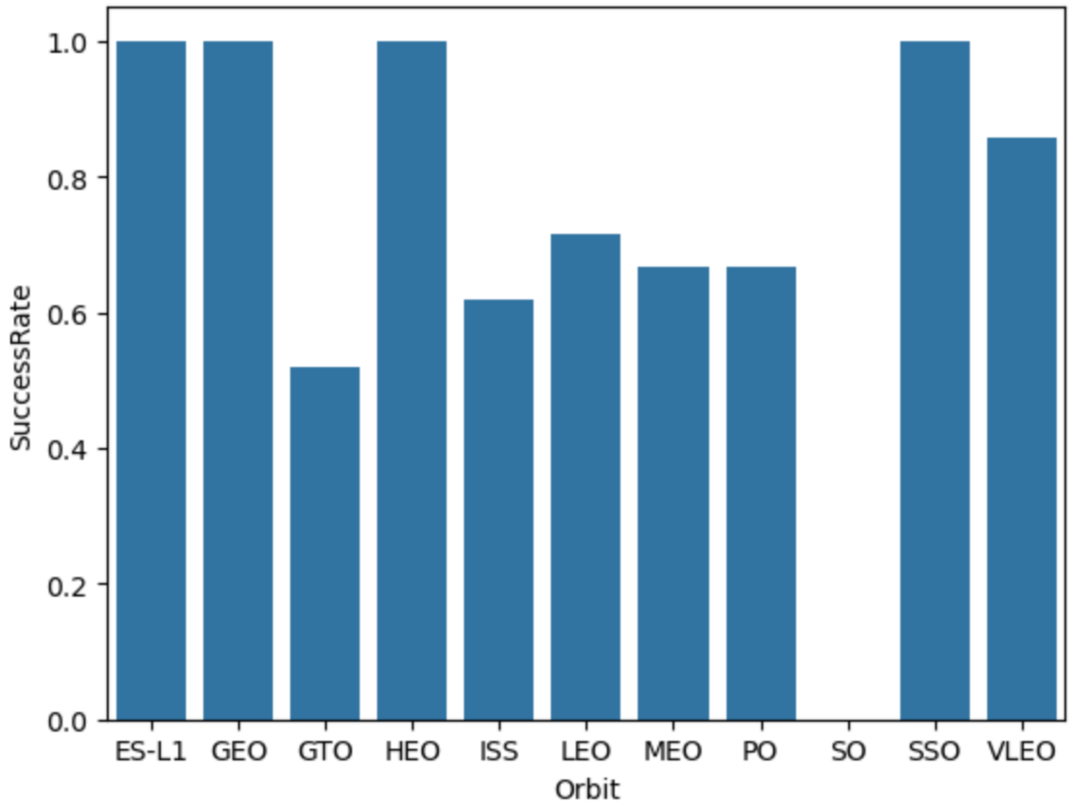
Success Rate vs. Orbit Type

- As from the graph we can say that Success rate for orbits ES-L1, GEO, HEO and SSO is maximum.
- For orbit type SO the success rate is minimum ~ 0.

```
[ ]: mean_success_rate = data_falcon9.groupby('Orbit')['Class'].mean()
     result_df = mean_success_rate.reset_index()
     result_df.columns = ['Orbit', 'SuccessRate']
     result_df
```

```
[326]: sns.barplot(data=result_df, x='Orbit', y="SuccessRate")
```

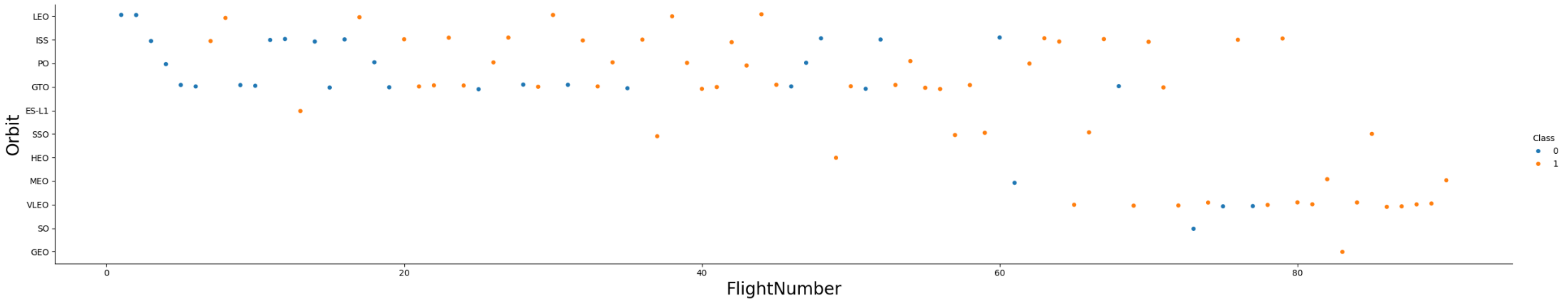
```
[326]: <Axes: xlabel='Orbit', ylabel='SuccessRate'>
```



Flight Number vs. Orbit Type

- After flight number 80 there are only successful outcomes for all the launches for all orbits.
- After flight number 80 most rockets were launched for orbit VLEO.

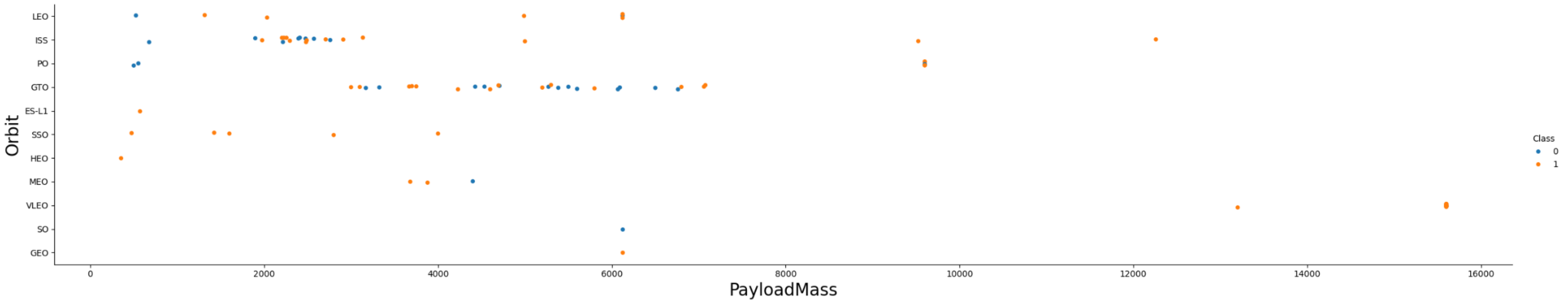
```
[328]: sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=data_falcon9, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Payload vs. Orbit Type

- For payload mass of 10,000 kg or more there are only successful outcomes for all orbits.
- For payload mass over 10,000 kg only for ISS, PO, VLEO orbits rockets have been launched.

[330]:
```
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=data_falcon9, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

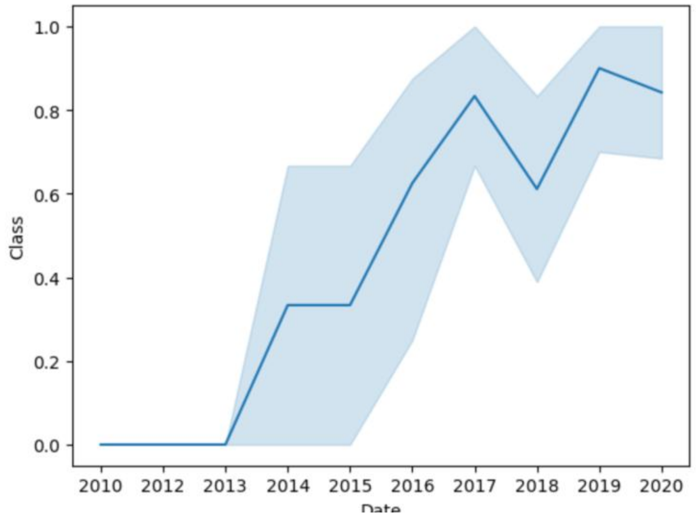


Launch Success Yearly Trend

- As we can see from the line chart that after 2013 there is increased success rate for launch outcomes.

```
[334]:
# Function to extract years from the date.
# Convert the 'Date' column to strings
data_falcon9['Date'] = data_falcon9['Date'].astype(str)
year=[]
def Extract_year():
    for i in data_falcon9["Date"]:
        year.append(i.split("-")[0])
    return year
Extract_year()
data_falcon9['Date'] = year

sns.lineplot(data=data_falcon9, x="Date", y="Class")
```

[334]: <Axes: xlabel='Date', ylabel='Class'>



All Launch Site Names

- The four launch sites present in our dataset are:
- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40

All Launch Sites name:

[336]:
```
%sql select distinct("Launch_Site") from SPACEXTABLE;
```

```
 * sqlite:///my_data2.db
Done.
```

[336]:

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

Launch Site Names Begin with 'CCA'

- Given here are the first five records of launch site name starting with ‘CCA’.

Launch Site name begins with 'CCA'

[340]:
```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5;
```

```
 * sqlite:///my_data2.db
Done.
```

[340]:

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 carried from NASA

The total payload mass carried by boosters from NASA as customer is equal to 45,596 kg.

Display the total Payload Mass carried by boosters from NASA

[343]:
```
%sql select sum("Payload_Mass__KG_"), "Customer" from SPACEXTABLE group by Customer having "Customer" = 'NASA (CRS)';
```

```
 * sqlite:///my_data2.db
Done.
```

[343]:

sum(Payload_Mass__KG_)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

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

Display the average Payload Mass carried by booster version F9 v1.1

[346]:
```
%sql select avg("Payload_Mass__KG_"), "Booster_Version" from SPACEXTABLE group by "Booster_Version" having "Booster_Version" like 'F9 v1.1';
```

```
 * sqlite:///my_data2.db
Done.
```

[346]:

avg(Payload_Mass__KG_)	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

- As per our dataset on 22 -12-1015, first successful landing in ground pad took place.

Find the date of first succesful landing outcome in ground pad ¶

```
[349]: %sql select min("Date") from SPACEXTABLE where "Landing_Outcome" like '%ground pad%';
```

```
 * sqlite:///my_data2.db
Done.
```

[349]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Here is the List of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Boosters which have Successful Drone ship landing with payload between 4000 and 6000:

[352]:
```
%sql select "Booster_Version", "Landing_Outcome" from SPACEXTABLE where ("Landing_Outcome" = 'Success (drone ship)') & ("PAYLOAD_MASS__KG_" > 4000 & "PAY
```

```
 * sqlite:///my_data2.db
Done.
```

[352]:

Booster_Version	Landing_Outcome
F9 FT B1021.1	Success (drone ship)
F9 FT B1022	Success (drone ship)
F9 FT B1023.1	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1029.1	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1029.2	Success (drone ship)
F9 FT B1036.1	Success (drone ship)
F9 FT B1038.1	Success (drone ship)
F9 B4 B1041.1	Success (drone ship)
F9 FT B1031.2	Success (drone ship)
F9 B4 B1042.1	Success (drone ship)
F9 B4 B1045.1	Success (drone ship)
F9 B5 B1046.1	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- As per the query result 100 mission outcomes have been successful and only 1 mission outcome has been unsuccessful.

Total number of successful and failure mission outcomes

[355]:
```
%sql select count("Mission_Outcome"), "Mission_Outcome" from SPACEXTABLE group by "Mission_Outcome" ;
```

```
 * sqlite:///my_data2.db
Done.
```

[355]:

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

Boosters Carried Maximum Payload

- Here is the names of booster version which have carried the maximum payload (15,600 kg) to space.s

Name of booster_versions which have carried maximum payload mass.

```
[358]: %sql select "Booster_Version", "PAYLOAD_MASS__KG_" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

 * sqlite:///my_data2.db
Done.
```

[358]:

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

- Here the records of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- As we can see that both outcome which involved landing on drone ship were unsuccessful.

List the records displaying month names, failure landing_outcomes in drone ship, booster versions and launch sites for the months in 2015

[361]:

```
%sql select substr(Date,6,2) as month, substr(Date,0,5) as year, "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where ("Landing_Out
```

```
 * sqlite:///my_data2.db
Done.
```

[361]:

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

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

- As we can see maximum number of launches (21) were not attempted to land successfully.

Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20

[364]:

```
%sql select count("Landing_Outcome") as cnt, "Landing_Outcome" from SPACEXTABLE group by "Landing_Outcome" having date between '2010-06-04' and '2017-03-
```

```
 * sqlite:///my_data2.db
Done.
```

[364]:

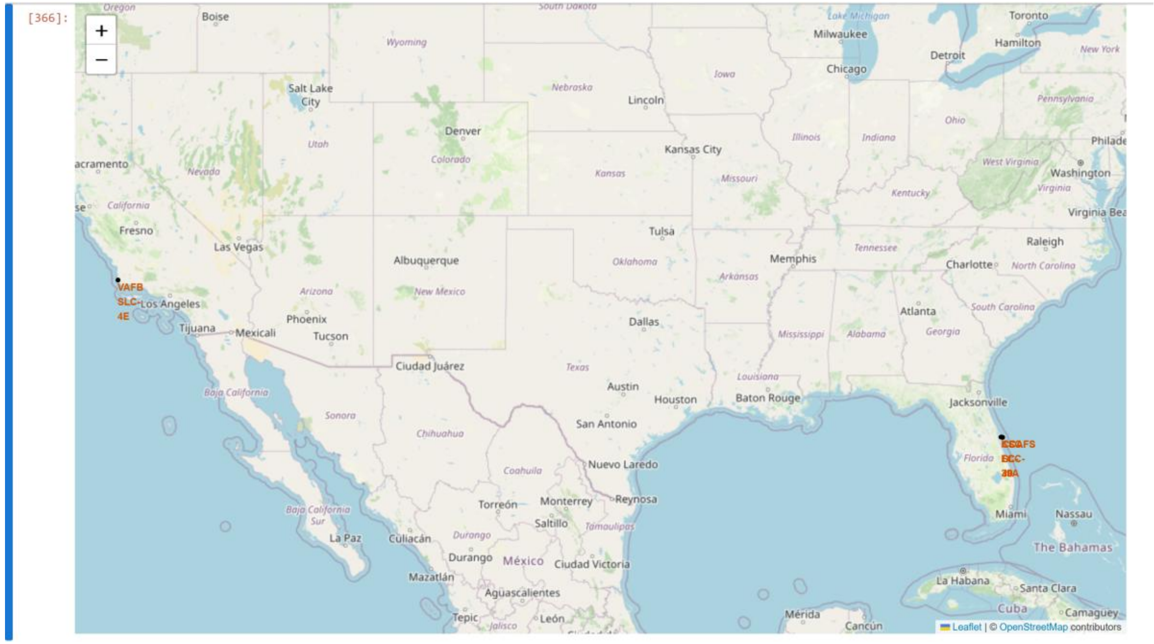
cnt	Landing_Outcome
21	No attempt
14	Success (drone ship)
9	Success (ground pad)
5	Failure (drone ship)
5	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

All launch site location marker

- Here is a screenshot of all launch sites marker on the map.
- As we can see in the map that there are four distinct sites visible as red circles.

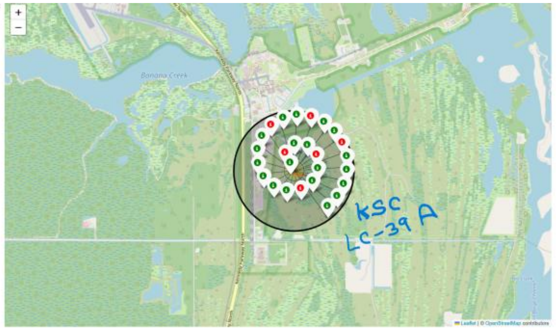


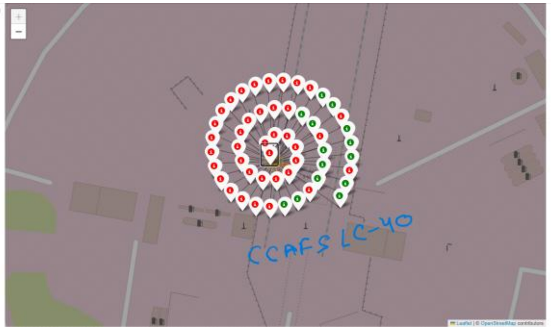

```
[366]: # Initial the map
site_map = folium.Map(location=nasa_coordinate, zoom_start=5)
# For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name as a popup label
for index, row in launch_sites_df.iterrows():
    circle = folium.Circle([row['Lat'], row['Long']], radius=1000, color='#000000', fill=True).add_child(folium.Popup(row['Launch Site']))
    marker = folium.Marker([row['Lat'], row['Long']], icon=DivIcon(icon_size=(20,20), icon_anchor=(0,0), html='<div style="font-size: 12; color: #d35400;

    site_map.add_child(circle)
    site_map.add_child(marker)


site_map
```

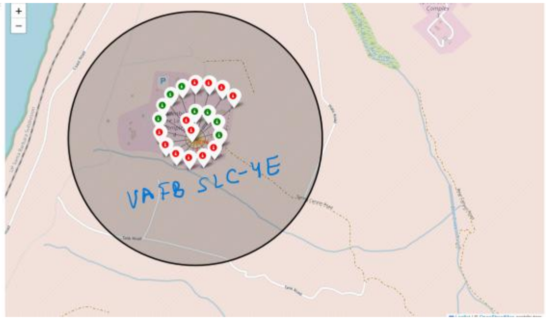
Color labeled (green-successful & red-unsuccessful) outcomes on the map

As seen on the map we can say that for KSC LC-39A launch site, there are more failure outcomes that success.

- The highest number of successful outcomes of landing is at CCAFS LC-40.







Distance between launch site CCAFS LC-40 and the nearest parkway

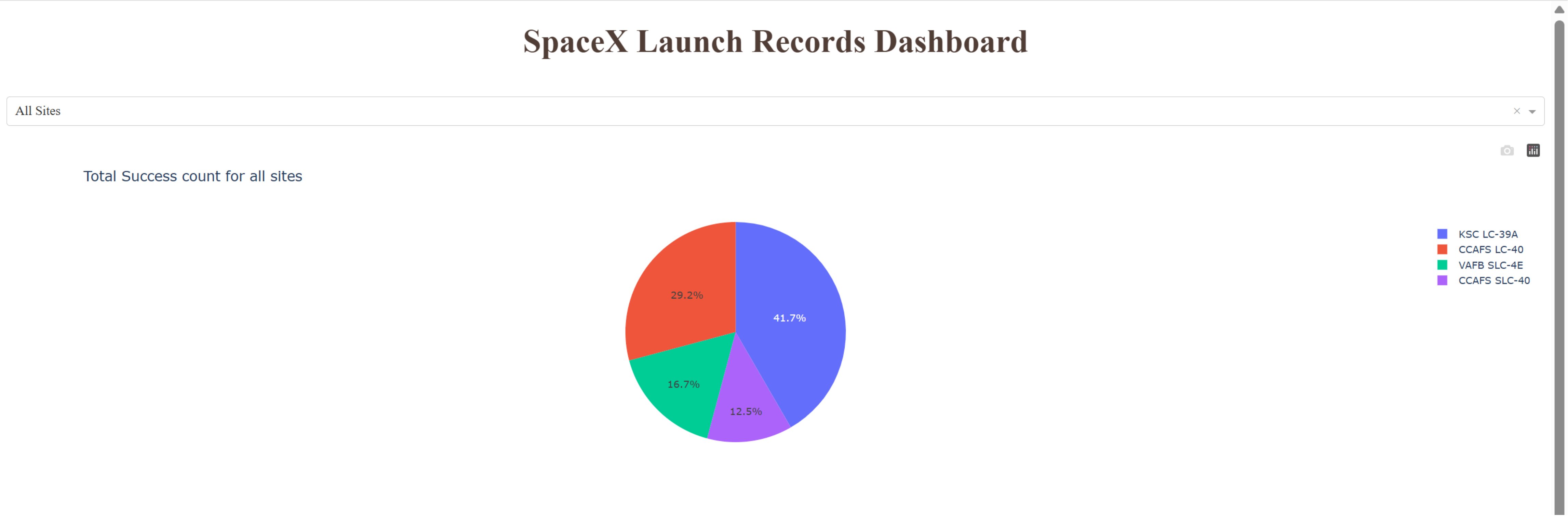
- The calculated distance (drawn by blue line) between launch site CCAFS LC-40 and the nearest parkway is 0.65 km.

Section 4

Build a Dashboard with Plotly Dash

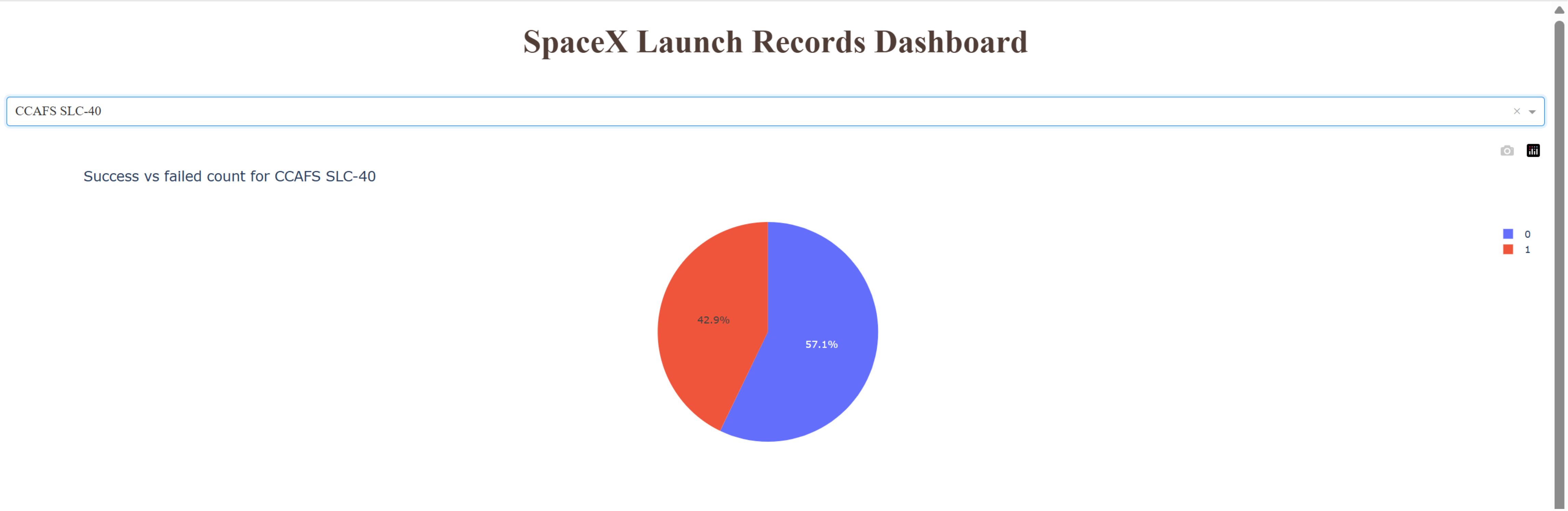
Total Success count for all sites

As we can observe from the given pie chart, the highest success percentage(41.7%) for all the successful outcomes of a landing is for the rockets launched by KSC LC-39A following by CCAFS LC-40(29.2%).



Launch site with highest launch success ratio(CCAFS SLC-40)

- As it is shown in the below pie chart, for CCAFS SLC-40 launch site, the success vs failed outcomes ratio is almost 60:40.that concludes that this particular site having more successful landing outcomes than failure.



Payload mass vs. Launch outcome scatter plot for all sites

- By observing the scatter plot between payload mass and launch outcome for all booster version for all sites, we can say that in the payload mass below 8K we are having more failure outcomes than successful ones for all the booster versions..

Section 5

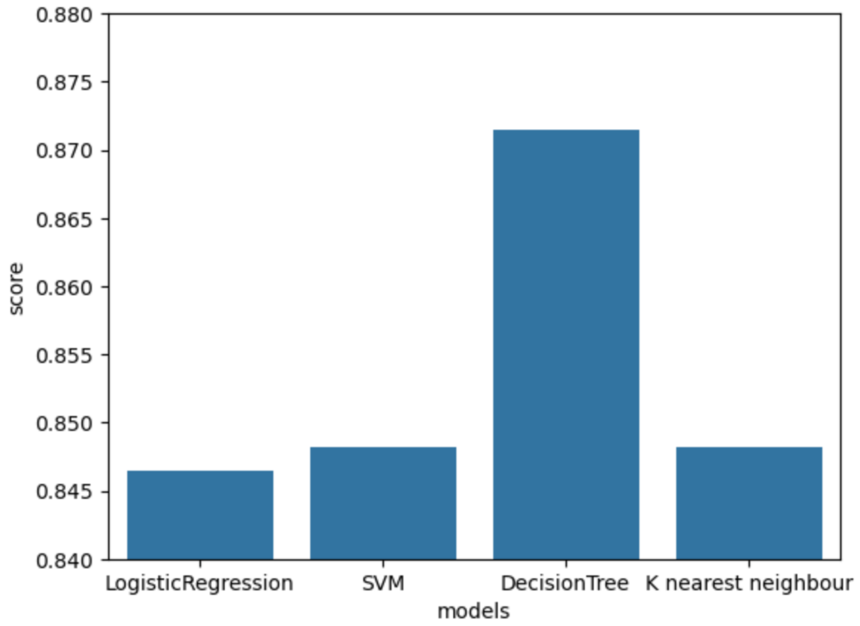
Predictive Analysis (Classification)

Classification Accuracy

- From the bar chart of four different models and their accuracy score given below we can observe that for our dataset Decision Tree classifier has worked best in term of accuracy score of 0.870 as compare to other classifier such as logistic regression and K nearest neighbour which have comparatively lower scores than Decision Tree.

[372]:
```
sns.barplot(x='models', y='score', data=df_scores)
plt.ylim(0.84, 0.88)
plt.xlabel('models', loc='center')
```

[372]: Text(0.5, 0, 'models')

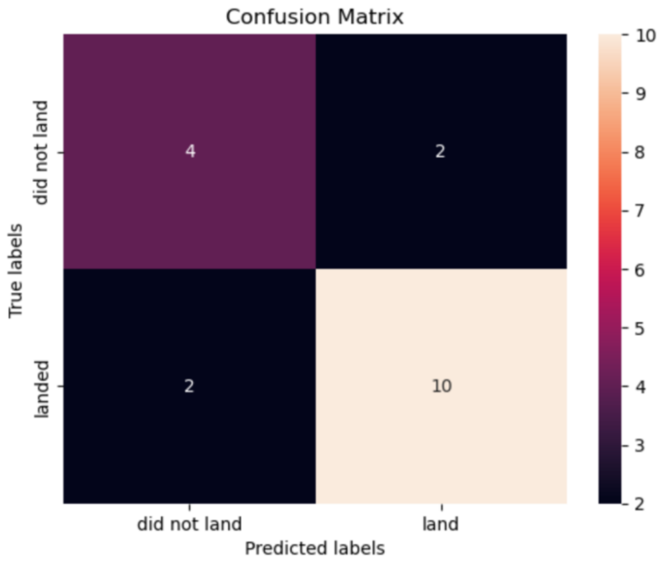


Confusion Matrix

As observed in the last slide of bar chart, Decision Tree has the best accuracy score. Here is the screenshot of confusion matrix for the Decision Tree classifier. The model has performed better than other models in classifying the ‘land’ and ‘did not land’ class more accurately with less errors.

[374]:
```
yhat_tree = tree_cv.predict(X_test)
print(classification_report(Y_test, yhat_tree))
plot_confusion_matrix(Y_test,yhat_tree)
```

```
              precision    recall  f1-score   support

           0       0.67      0.67      0.67         6
           1       0.83      0.83      0.83        12

    accuracy                           0.78        18
   macro avg       0.75      0.75      0.75        18
weighted avg       0.78      0.78      0.78        18
```



Conclusions

- From Exploratory data analysis we observe that launches with low payload mass show better results than heavy payload mass launches.
- The success rate of landing outcomes have increased over the years.
- KSC LC-39A has the highest success rate of landing outcome for all orbits and from all launch sites.
- Orbit ES-L1, GEO, HEO and SSO have almost 100% success rate of landing outcome.
- From map visualization we observed that most of the launch sites are in proximity to the Equator line compared in close proximity to the coast.
- From predictive analysis we saw that Decision Tree classifier have outperformed all the other three classification model in terms of accuracy and distinguishing between the two different classes as compared to the other models.

Appendix

Special thanks to:

- Coursera
- IBM
- Course instructors

Thank you!

Shailja Raghuvanshi