IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mehmet Ozkan Ceylan

15.03.2023

Outline

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

Executive Summary

- Summary of methodologies

Our data science project aims to predict the success of the Falcon 9 first stage landing to determine the cost of a launch. We'll use data science techniques such as data collection, wrangling, and visualization, as well as data classification to obtain predictions and insights. This information will help us make informed decisions and support our business goals.

- Summary of all results

Visualizations are built for clear understanding

Several machine learning algorithm applied and selected best resulted one

Introduction

- Project background and context

SpaceX declare that the Falcon9 can make a space travel %260 cheaper than other rockets because of the re-usable first stage. We are going to predict success landing for re-usable first stage.

- Problems to find answers

1. Will SpaceX land successfully?
2. Which factors are behind the failure of landing?
3. What is the accuracy of a successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was gathered with API and webscraping
- Perform data wrangling
  - There was a several types of outcomes, it is reduced to 1/0 to clear understanding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Linear Regression, SVM, Decision Tree, KNN methods were applied

Data Collection

Datasets are collected 2 different way

1. By using API
2. By using WebScraping from wikipedia

Data Collection – SpaceX API

- Data was collected by using ‘request’ to gather data from API
- GitHub URL for API

1. Getting response from API by using ‘request’
2. Converting the response to a .json file
3. Apply custom functions to the columns for cleaning
4. Assign list to dictionary and then a dataframe
5. Filter the dataframe with only Falcon9

Export as a .csv file

Data Collection - Scraping

- Data is collected by using Web Scrapping from Wikipedia

- GitHub URL for webscrapping

1. Getting response from HTML
2. Creating BeatifulSoup Object
3. Finding tables
4. Getting Column Names
5. Dictionary creation
6. Appending data to keys
7. Dictionary to Dataframe conversion

Dataframe to .csv

Data Wrangling

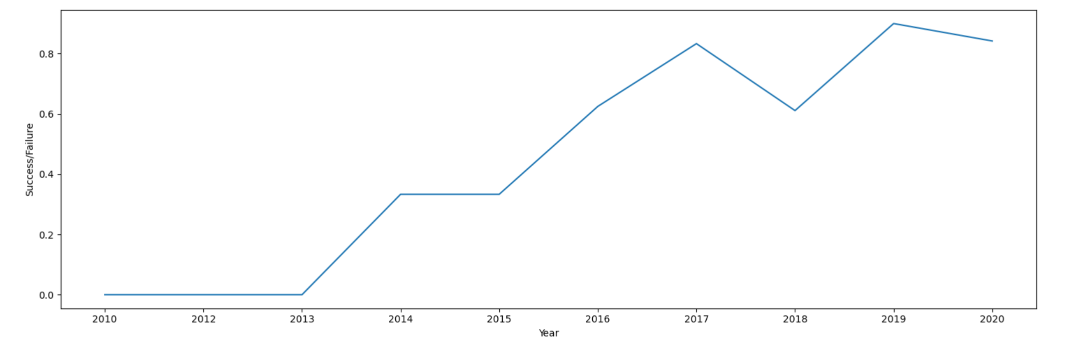
- Collected data was examined and prepared for the analysis. The NaN values replaced, and obtained landing class column from Outcome column

- GitHub URL for Data Wrangling

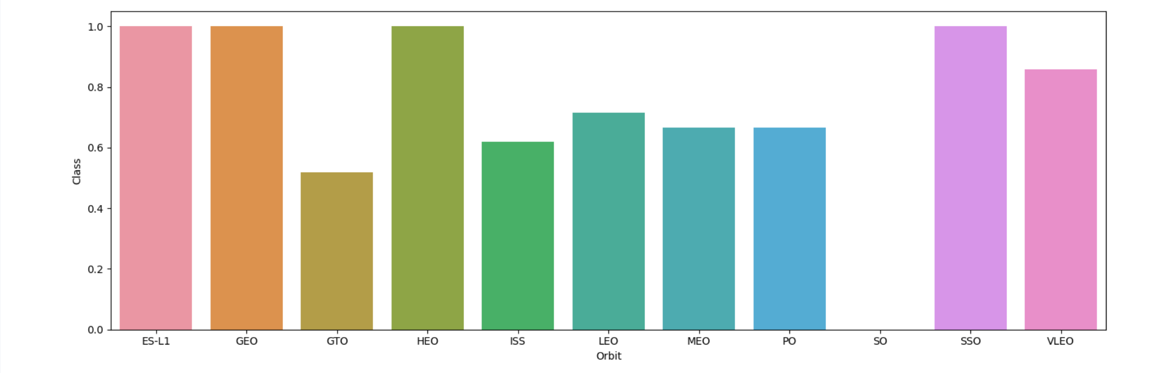
1. Read the Collected Data
2. Calculate the number of launches on each site
3. Calculate the number and occurrence of each orbit
4. Calculate the number and occurence of mission outcome per orbit type
5. Create a landing outcome label from Outcome column
5. Extract .csv file

EDA with Data Visualization

- For visualized the wrangled data Scatter Plot, Bar Plot and Line Charts was used. Thus, the relation between variables can observe clearly. Some Examples:



Line Chart for Yearly Success/Failure Rates



Relationship between success rate of each orbit type

- [GitHub URL]() for EDA with Visualization

EDA with SQL

- For better understand the Spacex DataSet, It was load into the corresponding table in a Db2 database. Then, by execute SQL queries to see relationships between variables.
- SqlAlchemy was used for DB connection
- Some examples for queries:

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

Average payload mass carried by booster version F9 v1.1

Landing _Outcome	Date	COUNT_LAUNCHES
Success	07-08-2018	20
No attempt	08-10-2012	10
Success (drone ship)	08-04-2016	8
Success (ground pad)	18-07-2016	6
Failure (drone ship)	10-01-2015	4
Failure	05-12-2018	3
Controlled (ocean)	18-04-2014	3
Failure (parachute)	04-06-2010	2
No attempt	06-08-2019	1

Ranking the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- GitHub URL for EDA with SQL

Build an Interactive Map with Folium

- Launch Sites Locations are analyzed with Folium. This lab contains the following tasks:
- TASK 1: Mark all launch sites on a map
- TASK 2: Mark the success/failed launches for each site on the map
- TASK 3: Calculate the distances between a launch site to its proximities

- GitHub URL for interactive map

Build a Dashboard with Plotly Dash

- 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.
- TASK 1: Add a Launch Site Drop-down Input Component
- TASK 2: Add a callback function to render success-pie-chart based on selected site dropdown
- TASK 3: Add a Range Slider to Select Payload
- TASK 4: Add a callback function to render the success-payload-scatter-chart scatter plotExplain why you added those plots and interactions
- GitHub URL for the dashboard

Predictive Analysis (Classification)

Model Building

- Creating Pandas DataFrame and Numpy array
- Splitting data to training and test data
- Standartization&Normalization of the data

Model Evaluating

- Check accuracy and precision with several method
- Get tuned the hyperparameters for each method

Model Improvement

- Feature engineering and model tuning

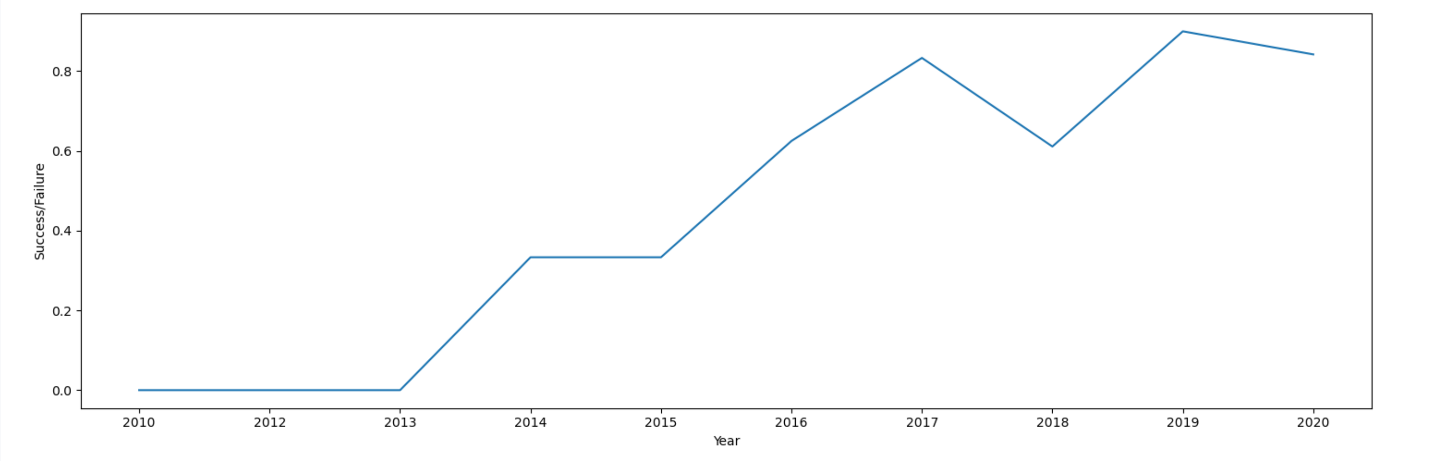
Getting the best results

- Comparing the models and selecting best accuracy score and best performing parameters as well as variables

- [GitHub URL] for predictive analysis

Results

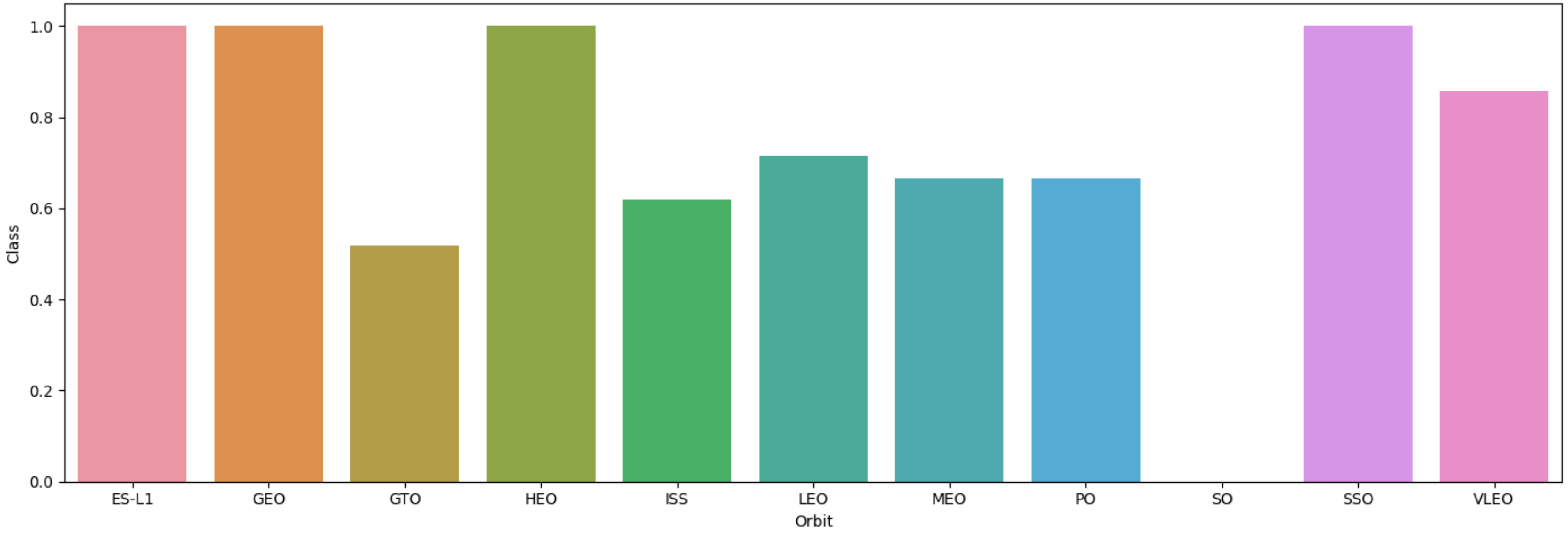
- Exploratory data analysis results



- The upward trend was observed for success rate in the past 10 years.

Results

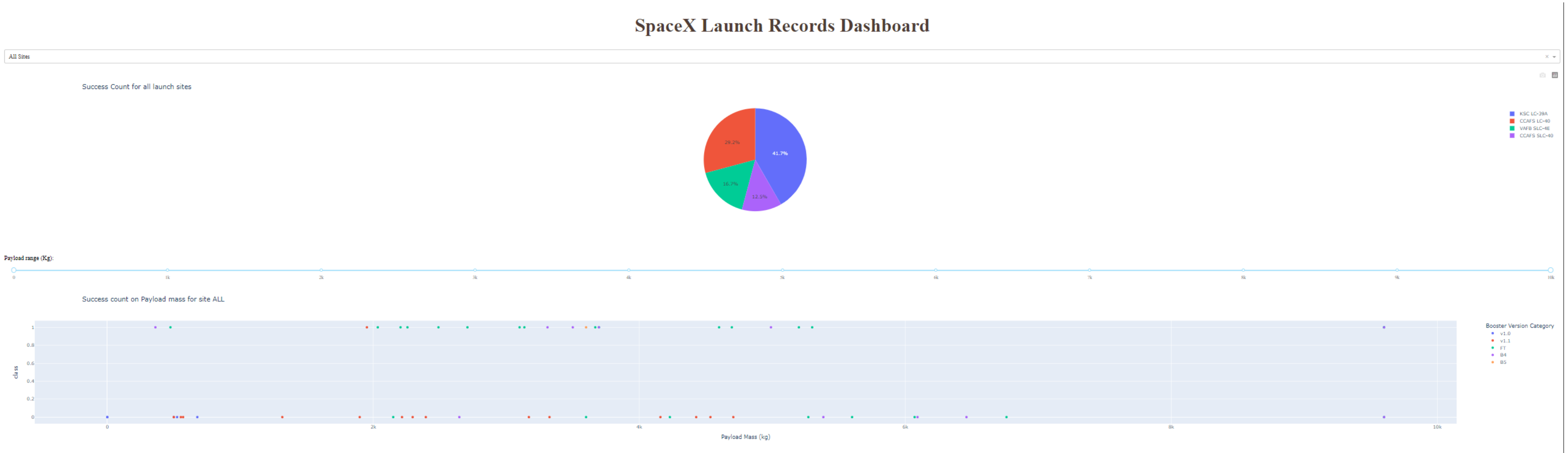
- Exploratory data analysis results



- The success rates of orbit types was determined.

Results

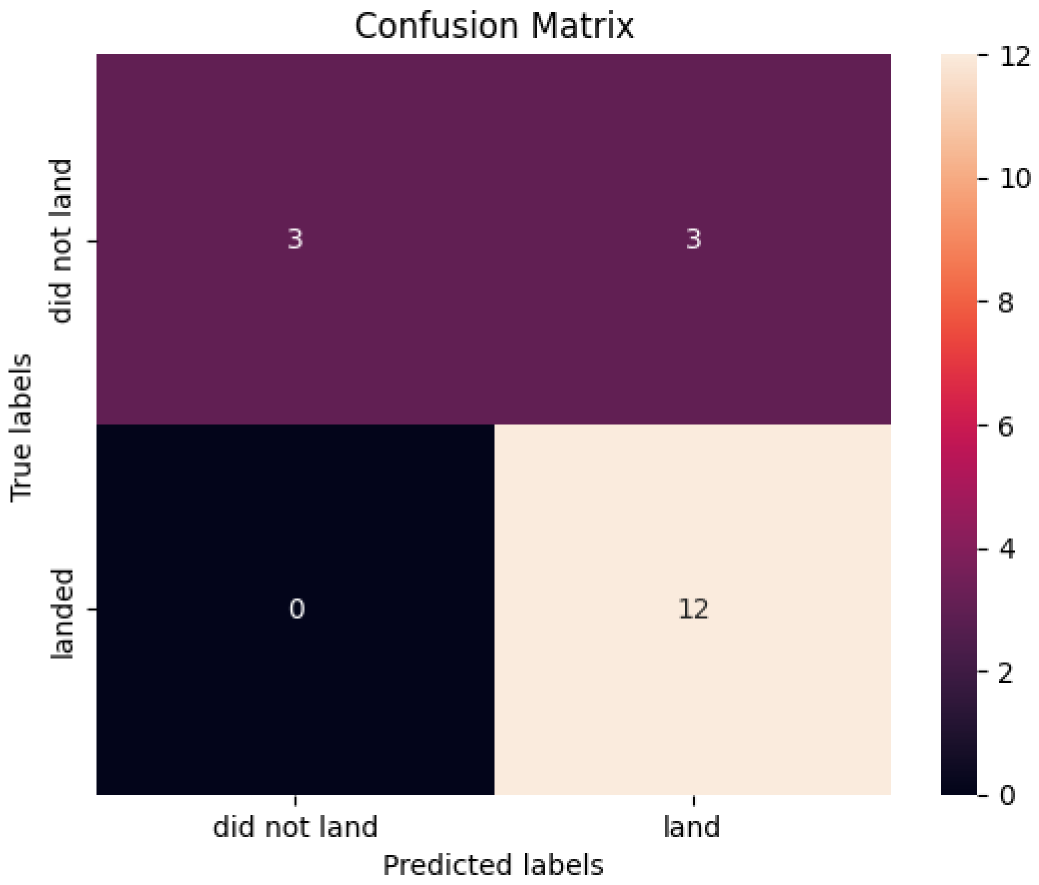
- Interactive analytics demo



- Predictive analysis results

Results

- Predictive analysis results



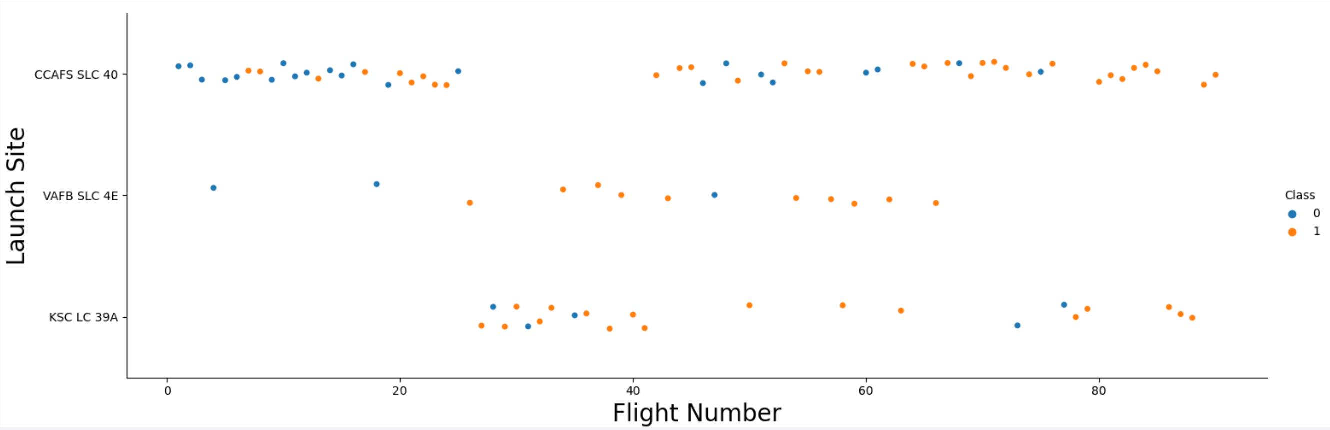

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.7777777777777778
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

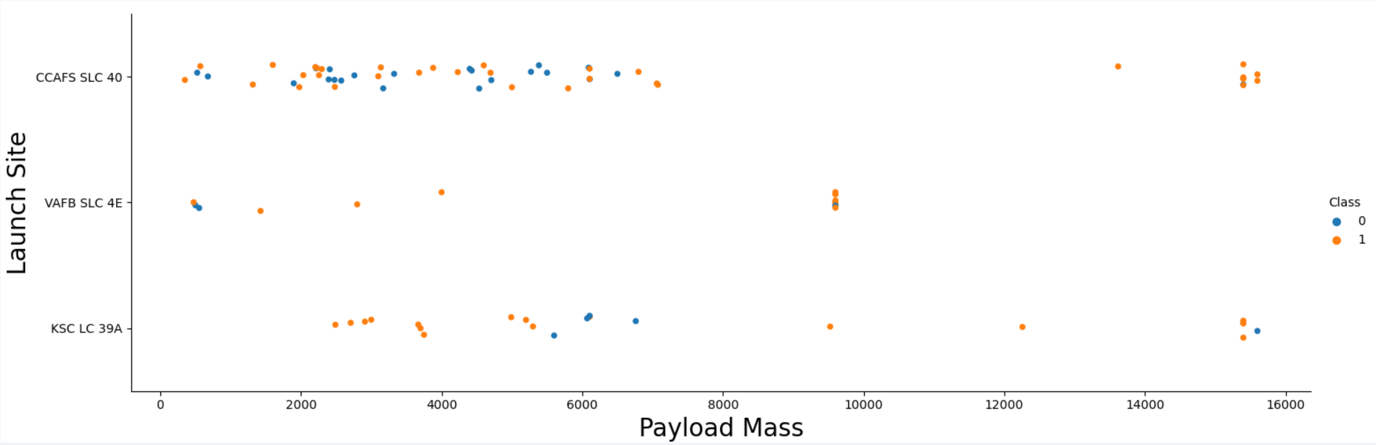
- Scatter plot of Flight Number vs. Launch Site




```
### TASK 1: Visualize the relationship between Flight Number and Launch Site
sns.catplot(x="FlightNumber", y="LaunchSite", hue="Class", data=df, aspect = 3)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

Payload vs. Launch Site

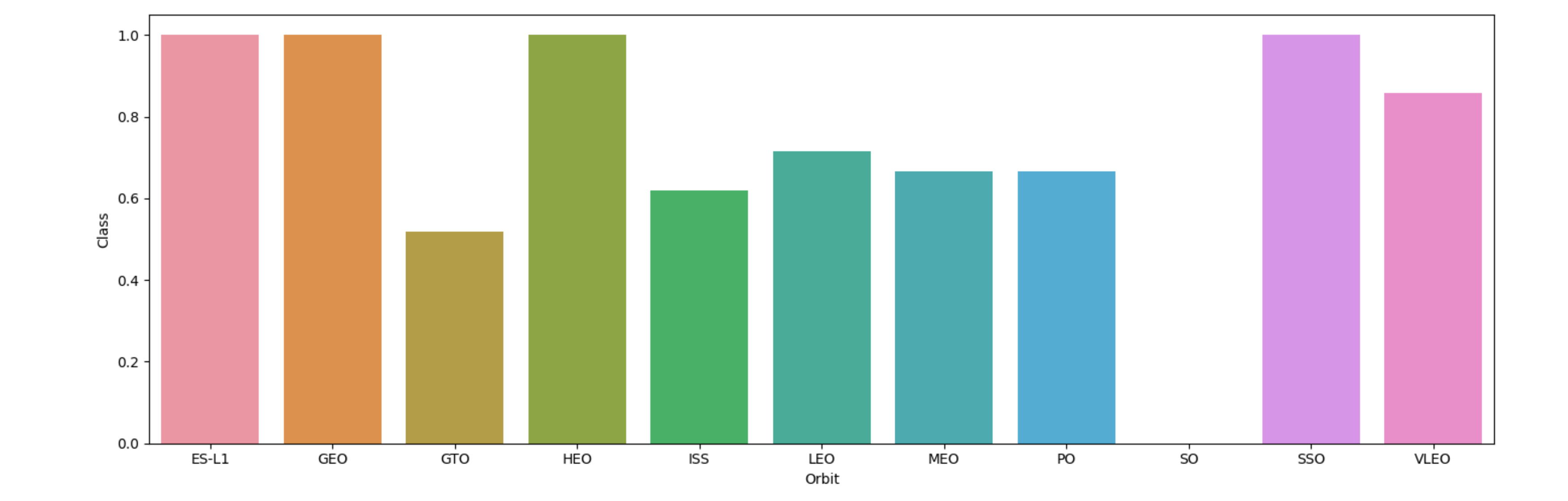
- Scatter plot of Payload vs. Launch Site




```python
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 3)
plt.ylabel("Launch Site",fontsize=20)
plt.xlabel("Payload Mass",fontsize=20)
plt.show()
```

Success Rate vs. Orbit Type

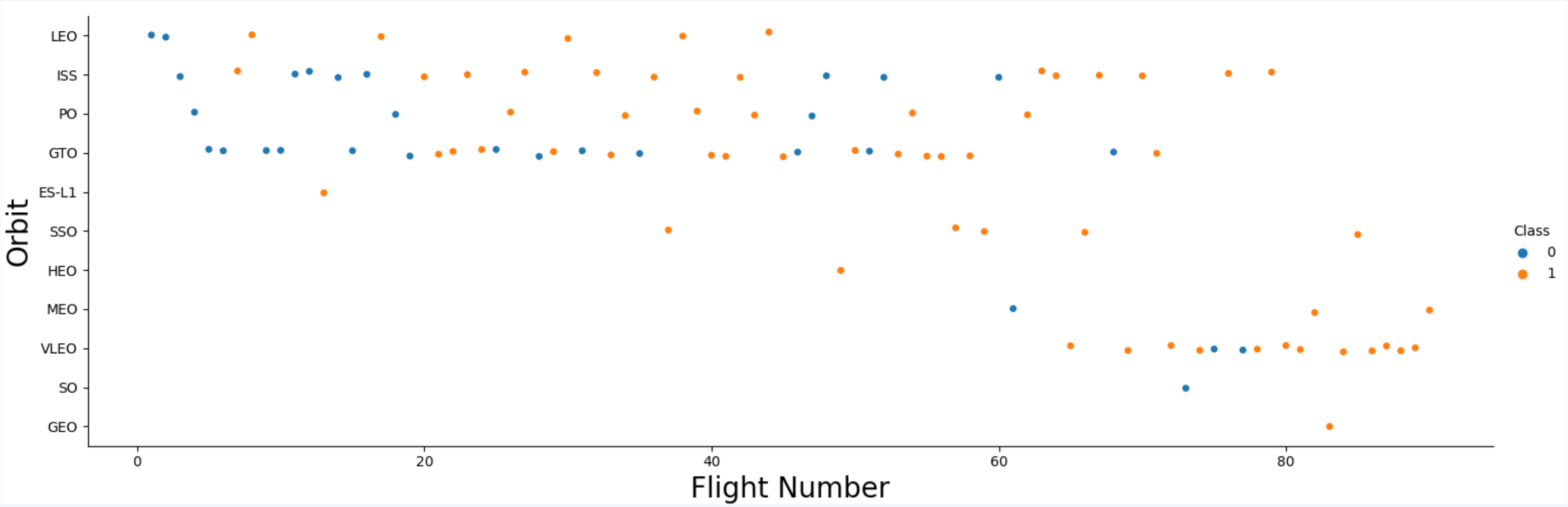
- Bar chart for the success rate of each orbit type




```
orbit_success = df.groupby('Orbit').mean()
orbit_success.reset_index(inplace=True)
sns.barplot(x="Orbit",y="Class",data=orbit_success)
plt.show()
```

Flight Number vs. Orbit Type

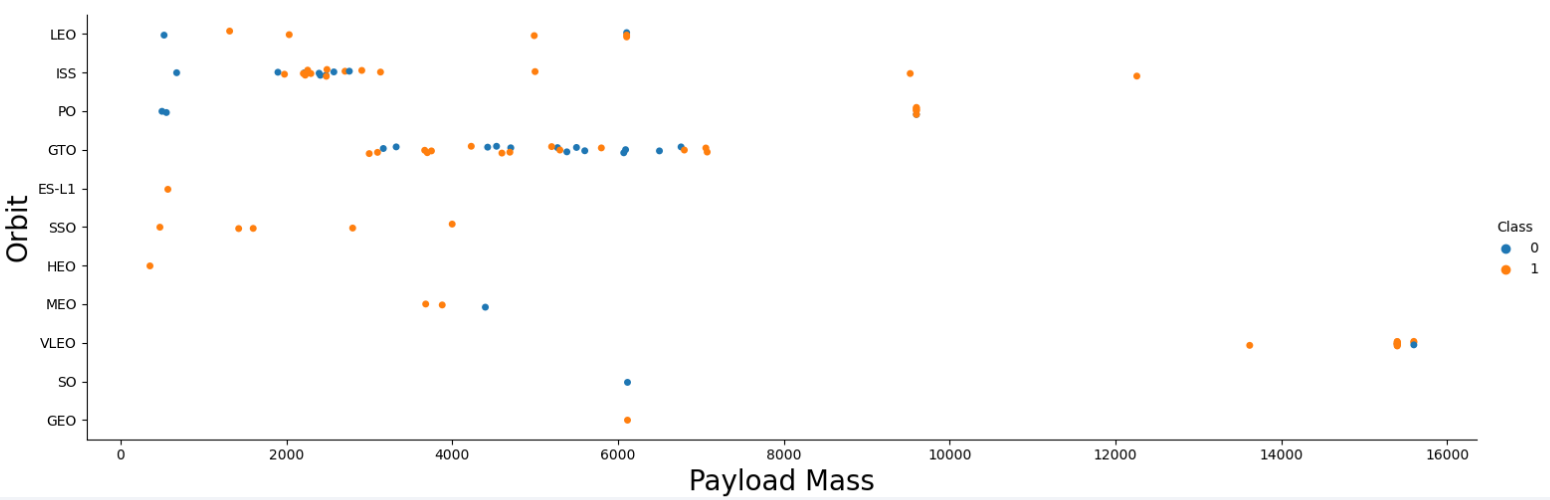
- Scatter point of Flight number vs. Orbit type




```
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 3)
plt.ylabel("Orbit",fontsize=20)
plt.xlabel("Flight Number",fontsize=20)
plt.show()
```

Payload vs. Orbit Type

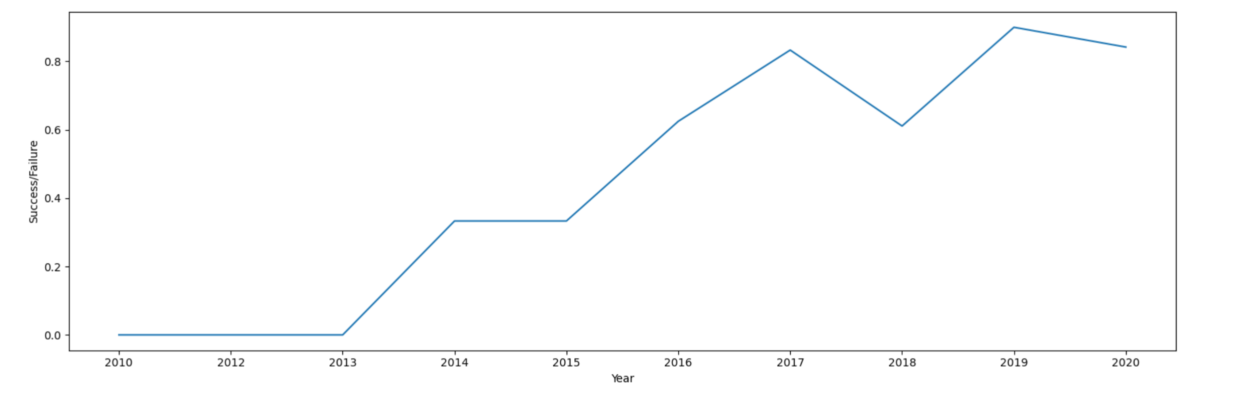
- Scatter point of payload vs. orbit type




```
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 3)
plt.ylabel("Orbit",fontsize=20)
plt.xlabel("Payload Mass",fontsize=20)
plt.show()
```

Launch Success Yearly Trend

- Line chart of yearly average success rate




```
plt.plot(average_by_year["Year"],average_by_year["Class"])
plt.xlabel("Year")
plt.ylabel("Success/Failure")
plt.show()
```

All Launch Site Names

- The names of the unique launch sites

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

It can be obtained by using distinct formula:

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL;
```

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

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

- It can be find with WHERE and LIKE functions of SQL

```
%sql select LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Total Payload Mass

- Total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

- It can be find with SUM command

```
%%sql

select SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

Average Payload Mass by F9 v1.1

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- It can be find by using SQL AVG() command

```
%sql select AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

min(date)
01-05-2017

- It can be find with SQL MIN() command

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

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

- SQL – BETWEEN() command is a perfect fit for this process

```
%sql select BOOSTER_VERSION from SPACEXTBL where "Landing _Outcome"='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 6000
```

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

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

```
%sql select Mission_Outcome, count(Mission_Outcome) as number from SPACEXTBL GROUP BY (Mission_Outcome)
```

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Date	Booster_Version	Launch_Site	Landing _Outcome
10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, "Landing _Outcome"  FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND substr(Date,7,4)='2015'
```

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

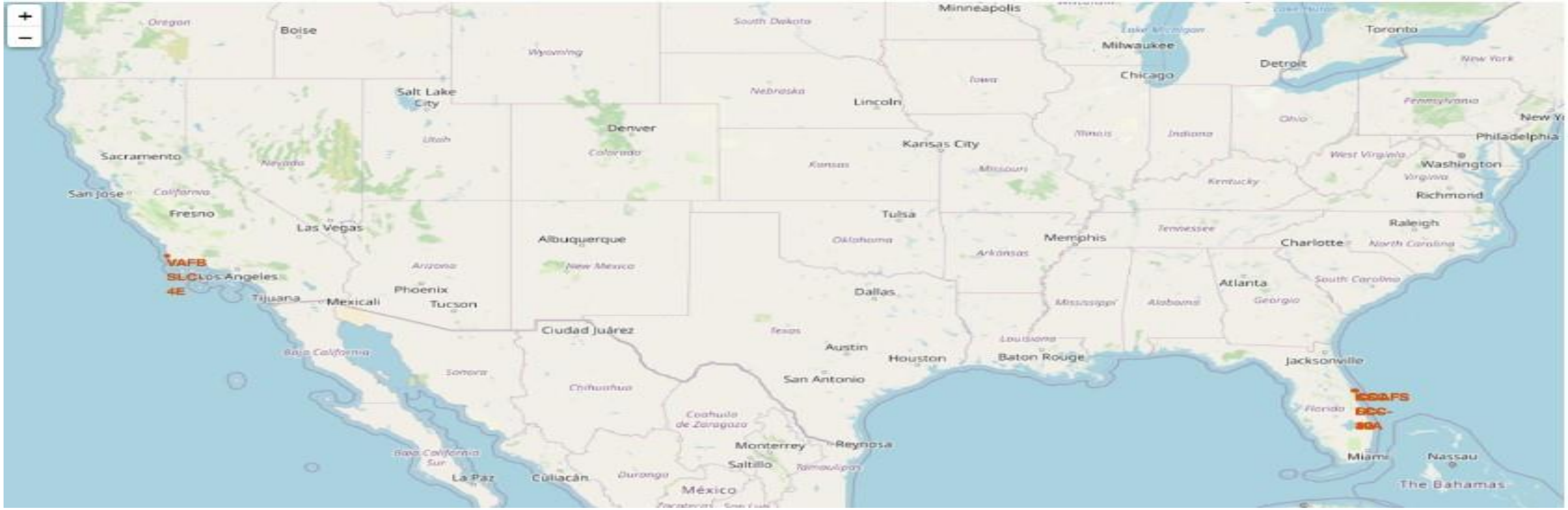
Landing _Outcome	Date	COUNT_LAUNCHES
Success	07-08-2018	20
No attempt	08-10-2012	10
Success (drone ship)	08-04-2016	8
Success (ground pad)	18-07-2016	6
Failure (drone ship)	10-01-2015	4
Failure	05-12-2018	3
Controlled (ocean)	18-04-2014	3
Failure (parachute)	04-06-2010	2
No attempt	06-08-2019	1

```
%sql SELECT "Landing _Outcome", DATE, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010'
AND '20-03-2017' GROUP BY "Landing _Outcome" ORDER BY COUNT_LAUNCHES DESC;
```

Section 3

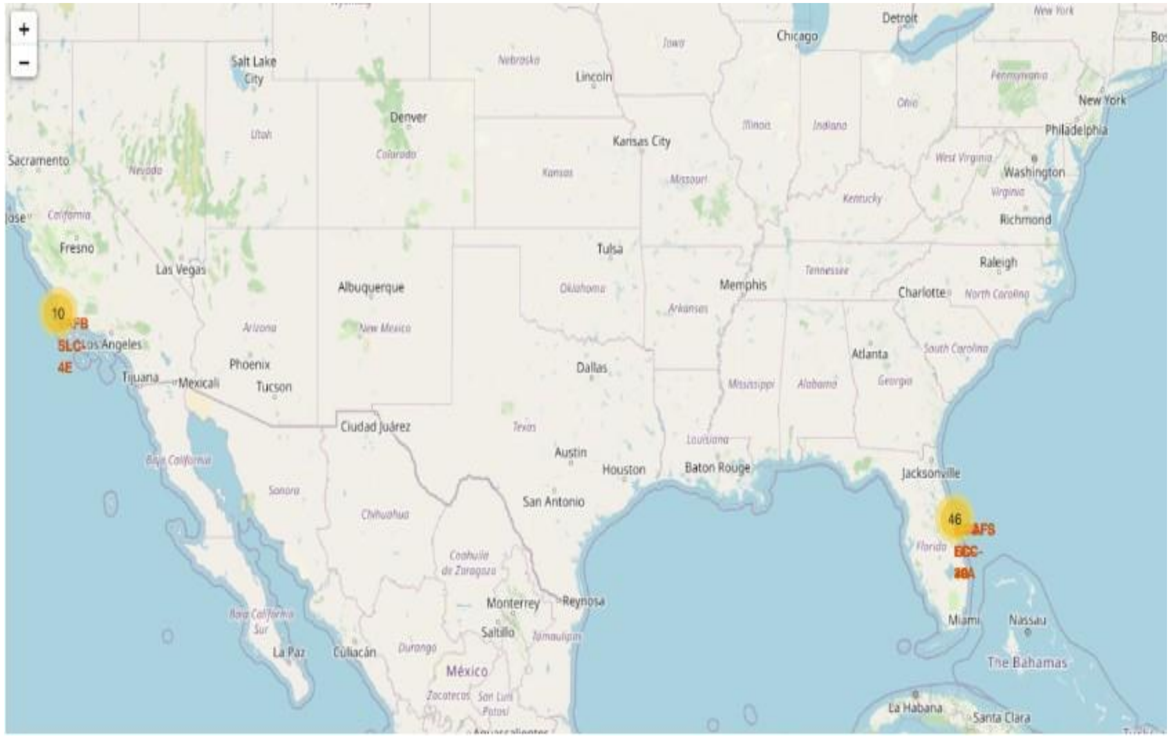
Launch Sites Proximities Analysis

Launch Sites On A Map



“SpaceX launch sites are in– Florida on the USA east coast and California on the USA west coast”

Sites And Launch Outcomes On A Map



“Green colored are the successful launches and red colored are the failed launches”

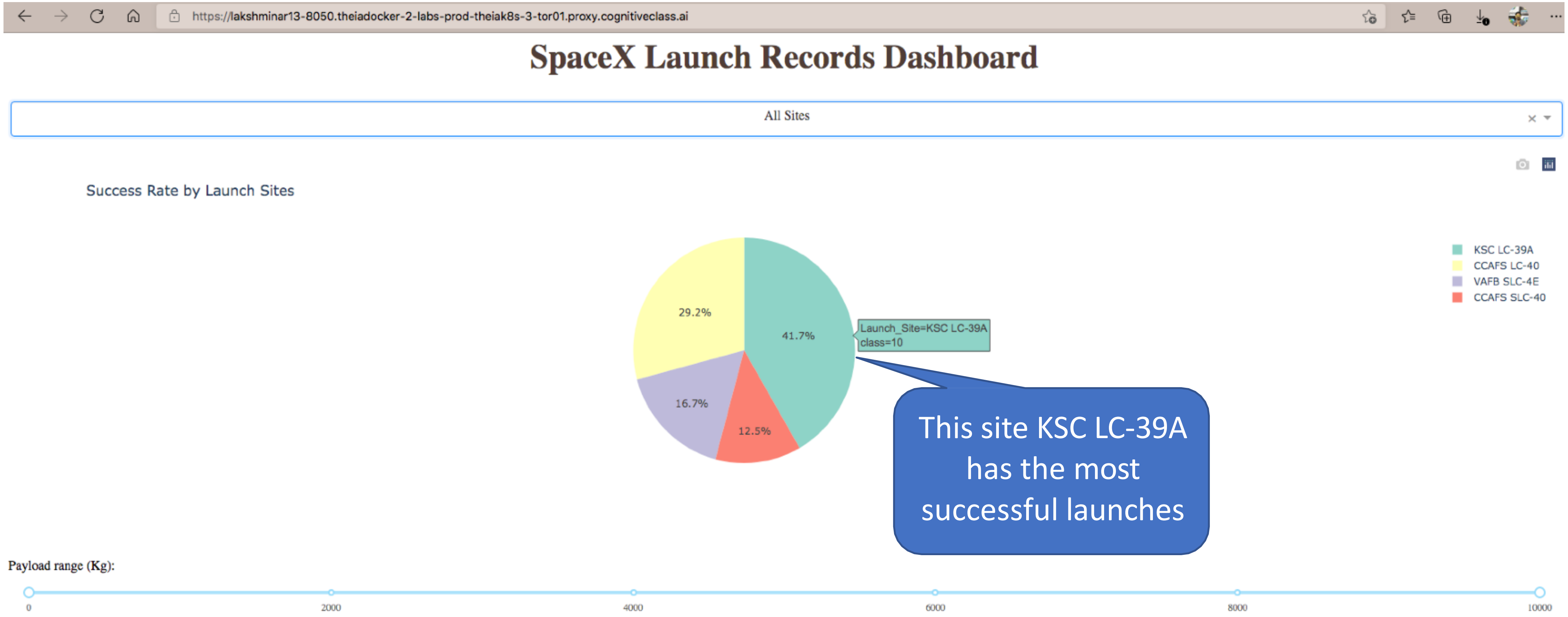
Launch Site Proximities On A Map

"Here, we can observe distance of launch sites from east coast, highways, key road, railway line visualized"

Section 4

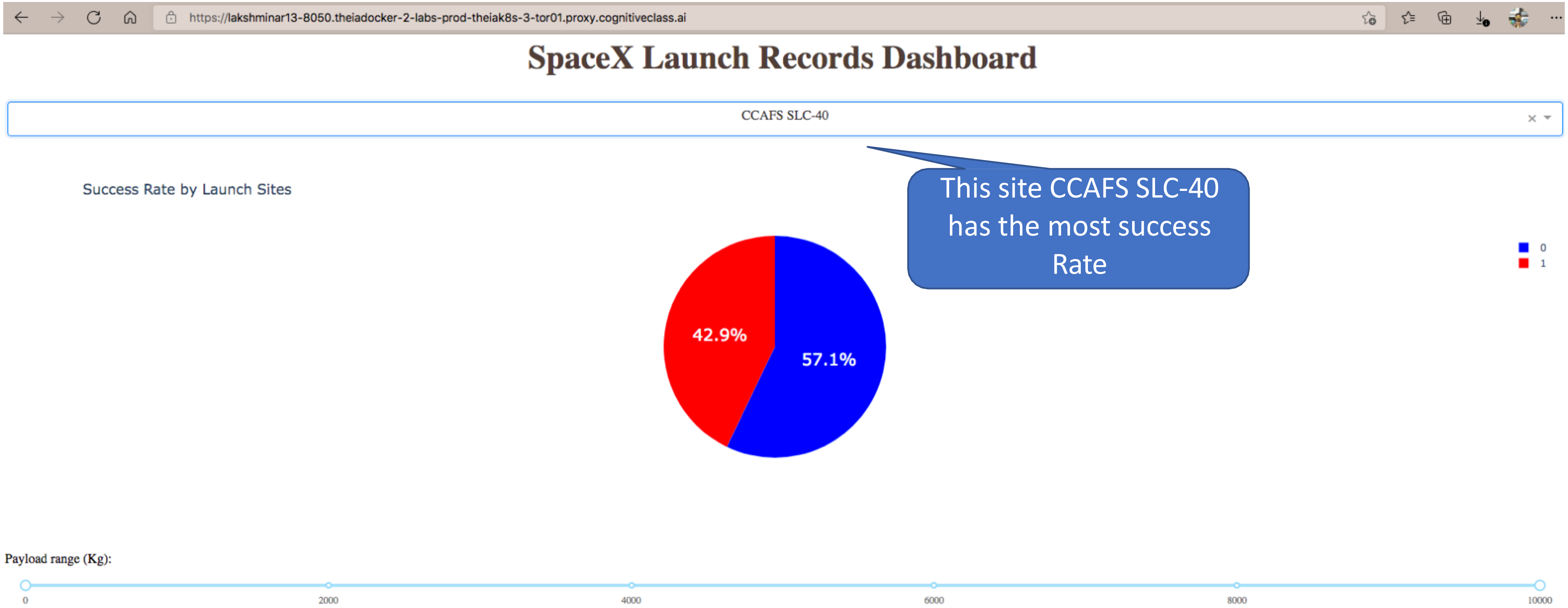
Build a Dashboard with Plotly Dash

Success Rate by Launch Sites



This site KSC LC-39A has the most successful launches

Most Successful Launch Site



Success by Payload Mass & Booster Version

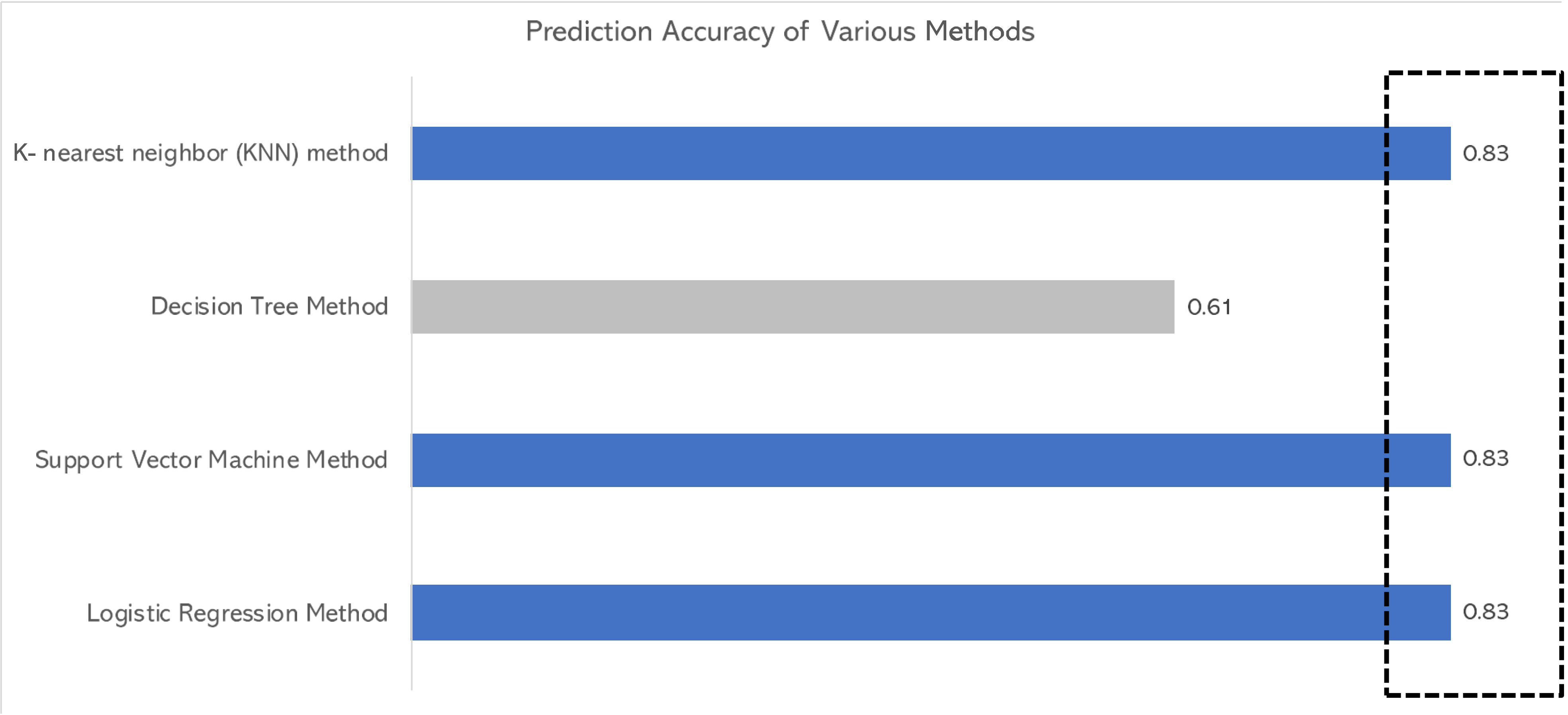
Lower payload launches are more successful

"Lower Payload launches (up to 6,000 kg) are more successful"

Section 5

Predictive Analysis (Classification)

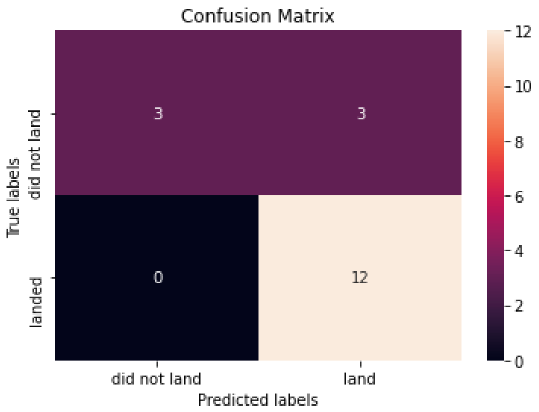
Classification Accuracy



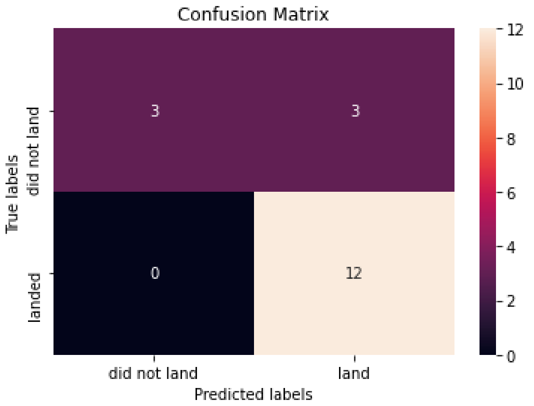
KNN, Support Vector & Logistic Regression Methods have high accuracy with same values

Confusion Matrix

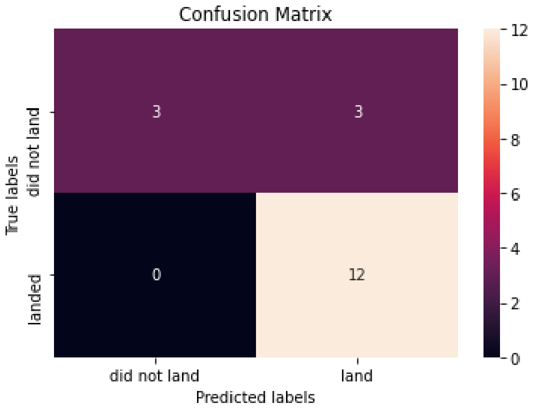
```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```




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




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



“The above confusion matrix shows that all 3 models – KNN, Logistic Regression & SVM have highest true positives and least false negatives"

Conclusions

- The landing point must be on the sea or near to the sea side
- There are a negative correlation between payload and success rate. If the payload get lighter the success rate will be increase
- There was a upward trend in success rate in past 10 years
- The F9 Booster has a highest success rate
- Decision Tree modal is not effective for this process. KNN, Logistic Regression and SVM can prefer for modelling.

Appendix

- All the codes, datasets, dashboards and notebooks are available on my Github profile. You can reach this project via this link.

- Also, my linkedin account and e-mail is below.

- E-mail: ozkannceylan@gmail.com
- LinkedIn

Thank you!