IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yujin Kim
June 26, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data collection using API and web scraping
  - Data wrangling
  - Exploratory data analysis
  - Machine learning classification prediction
- Summary of all results
  - Exploratory data analysis results
  - Predictive analysis result

Introduction

- Falcon 9 rocket from SpaceX costs 62 million dollars, while other providers cost up to 165 million dollars each. This saving is because SpaceX can reuse the first stage.
- The project aims to predict if the Falcon 9 first stage will land successfully, and therefore, determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - We applied one-hot encoding to categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built classification models using logistic regression, SVM, decision-tree, and K-NN
  - Compared the model performance to determine the best classification model

Data Collection

- The data was collected by SpaceX API.
  - Task 1: Request and parse the SpaceX launch data using the GET request
  - Task 2: Filter the dataframe to only include Falcon 9 launches
  - Task 3: Missing values – replace missing PayloadMass with the mean
- Additionally, Falcon 9 historical launch records were scrapped from a Wikipedia Page.

Data Collection – SpaceX API

- We 1) collected data from SpaceX API, 2) filtered data with Falcon 9 data only, and 3) treated missing data

- https://github.com/yujin3467/dscapstone/blob/main/1-spacex-data-collection-api.ipynb

```
response = requests.get(static_json_url)
response.status_code
```

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

```
data_falcon9 = df[df['BoosterVersion']!='Falcon 1']
data_falcon9.head()
```

```
# Calculate the mean value of PayloadMass column
data_falcon9['PayloadMass'].mean()

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, 6123.5476)
```

Data Collection - Scraping

- We applied web scraping to get Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into Pandas dataframe

- https://github.com/yujin3467/dscapstone/blob/main/2-webscraping.ipynb

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

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

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
# Use soup.title attribute
soup.title
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

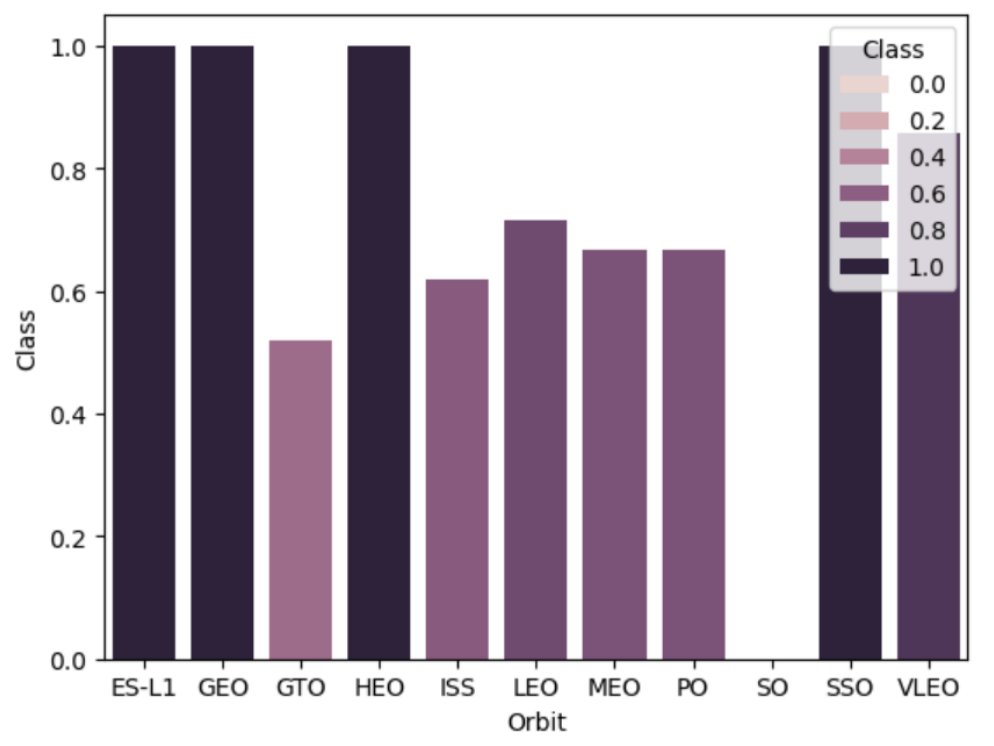
Data Wrangling

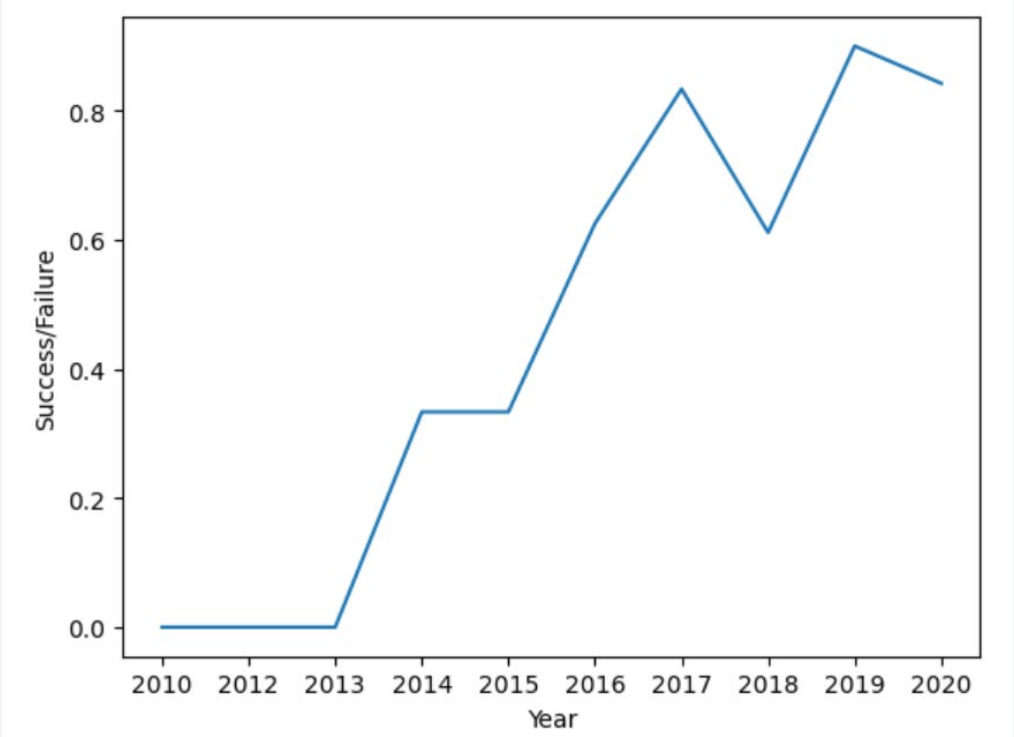
- We performed exploratory data analysis to determine the training labels.
  - Launch site
  - Orbit
  - Landing outcome
- We created a landing outcome label for supervised training

- https://github.com/yujin3467/dscapstone/blob/main/3-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We visualized the success/failure rate by orbit type and by year.





https://github.com/yujin3467/dscapstone/blob/main/5-edadataviz.ipynb

EDA with SQL

- We executed the SQL queries to explore data.
  - Inquiry the names of the unique launch sites
  - Launch site names with specific letters
  - Total pay load mass carried by boosters launched by NASA
  - Average pay load mass carried by a specific booster version
  - The date of the first successful landing outcome in ground pad
  - The name of boosters with condition (date and landing outcome)
  - Total number of success/ failure mission outcomes
- https://github.com/yujin3467/dscapstone/blob/main/4-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We marked Launch sites with the outcome (i.e., success/ failure). Additionally, we marked the closest coastline, railway, and highway with distance.
- https://github.com/yujin3467/dscapstone/blob/main/6-launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We plotted the interactive pie chart that shows the successful launch in a specific site.
- We plotted the interactive scatter plot that shows the outcome and pay load mass (kg).

Predictive Analysis (Classification)

- We performed multiple classification methods by splitting the data into training and test sets.
- The best classification model was decision tree with score .875.

- https://github.com/yujin3467/dscapstone/blob/main/7-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

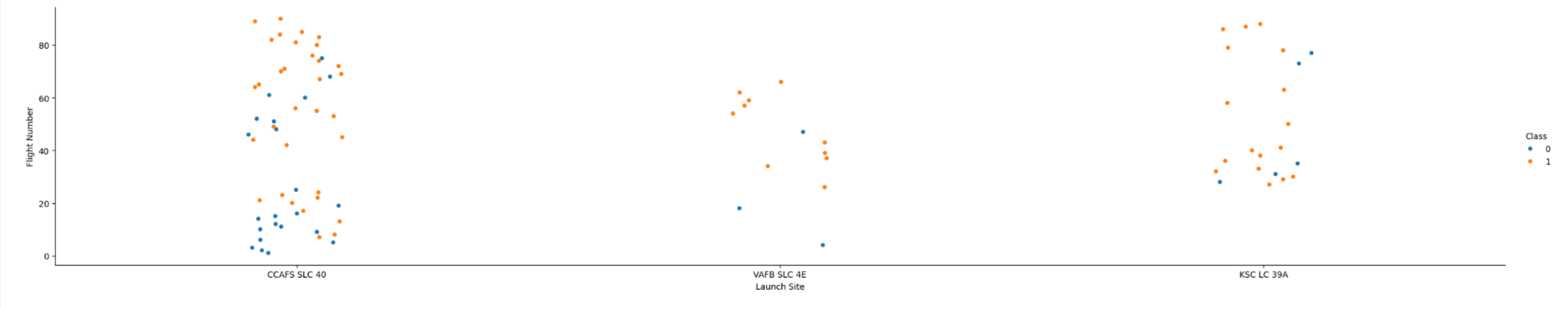
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

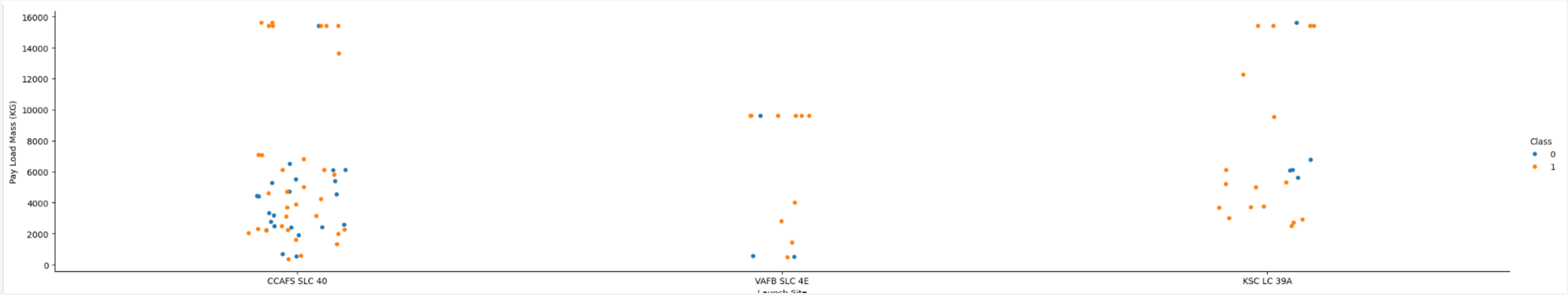
Flight Number vs. Launch Site

- We found that VAFB SLC 4E launch site shows the flight number of less than 70.
- KSC LC 39A showed all flight number of higher than 20.



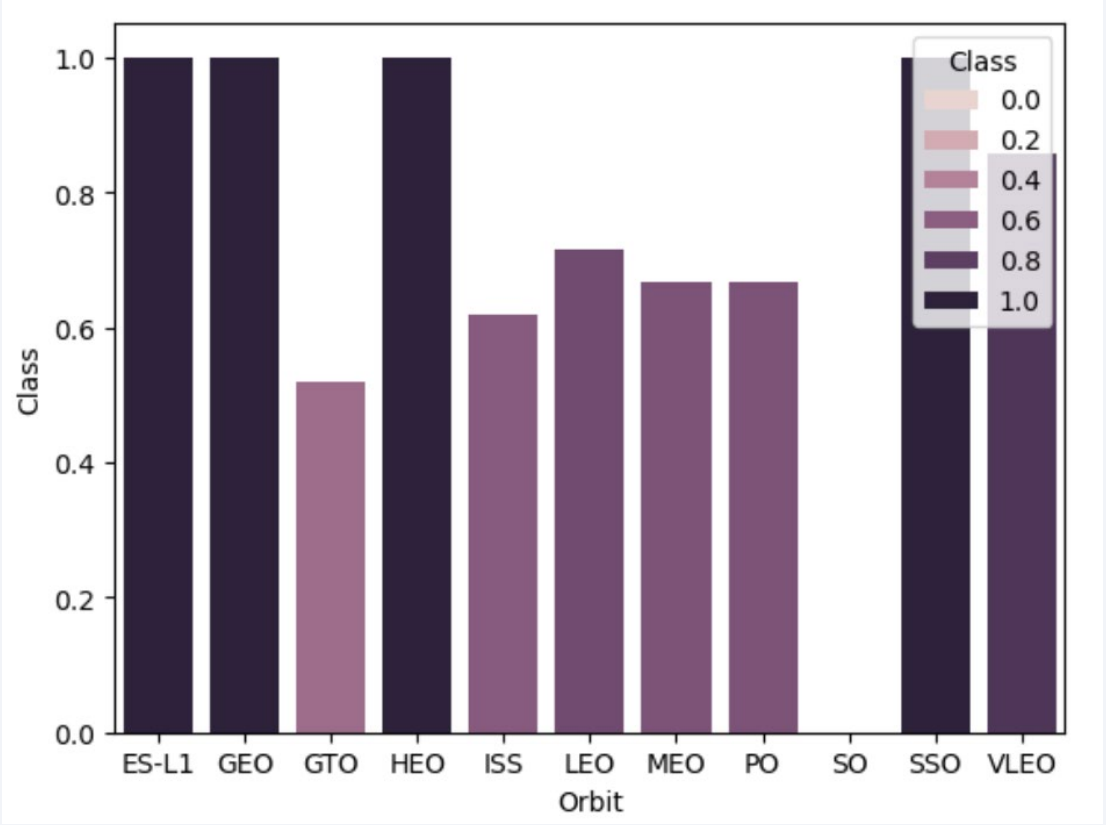
Payload vs. Launch Site

- for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



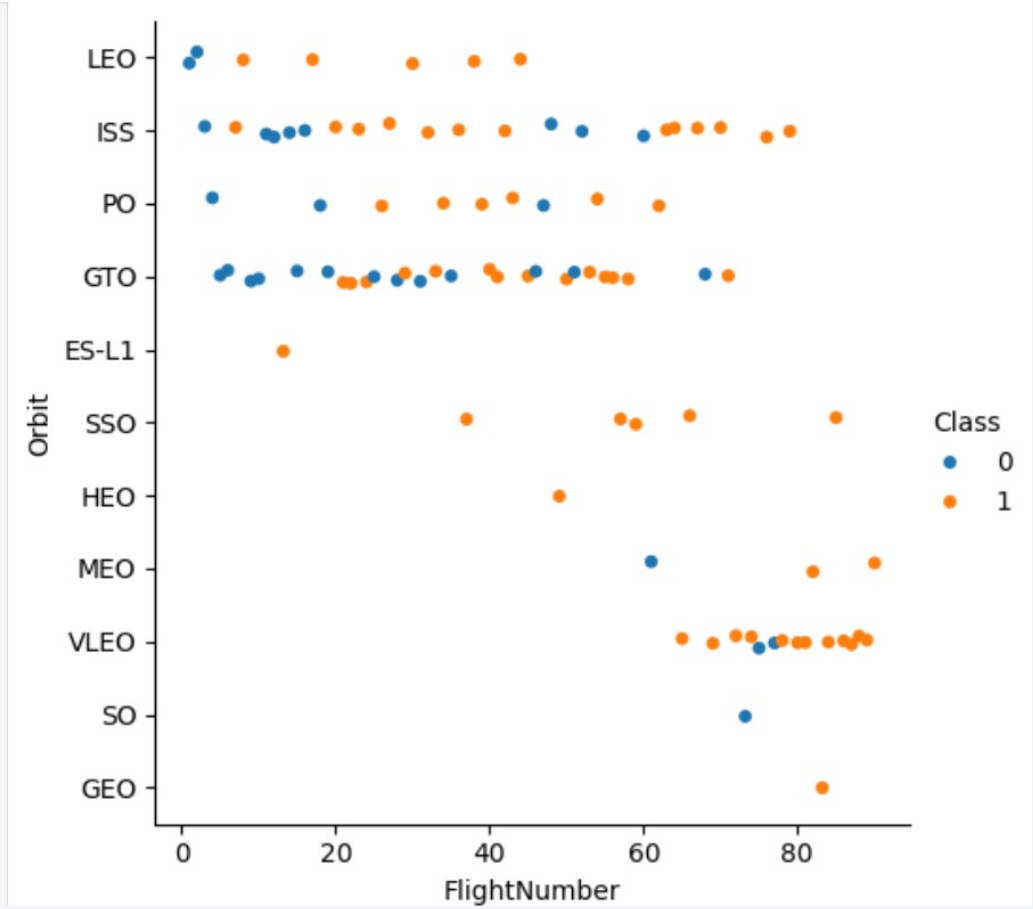
Success Rate vs. Orbit Type

- We found ES-L1, GEO, HEO, and SSO had higher success rate.



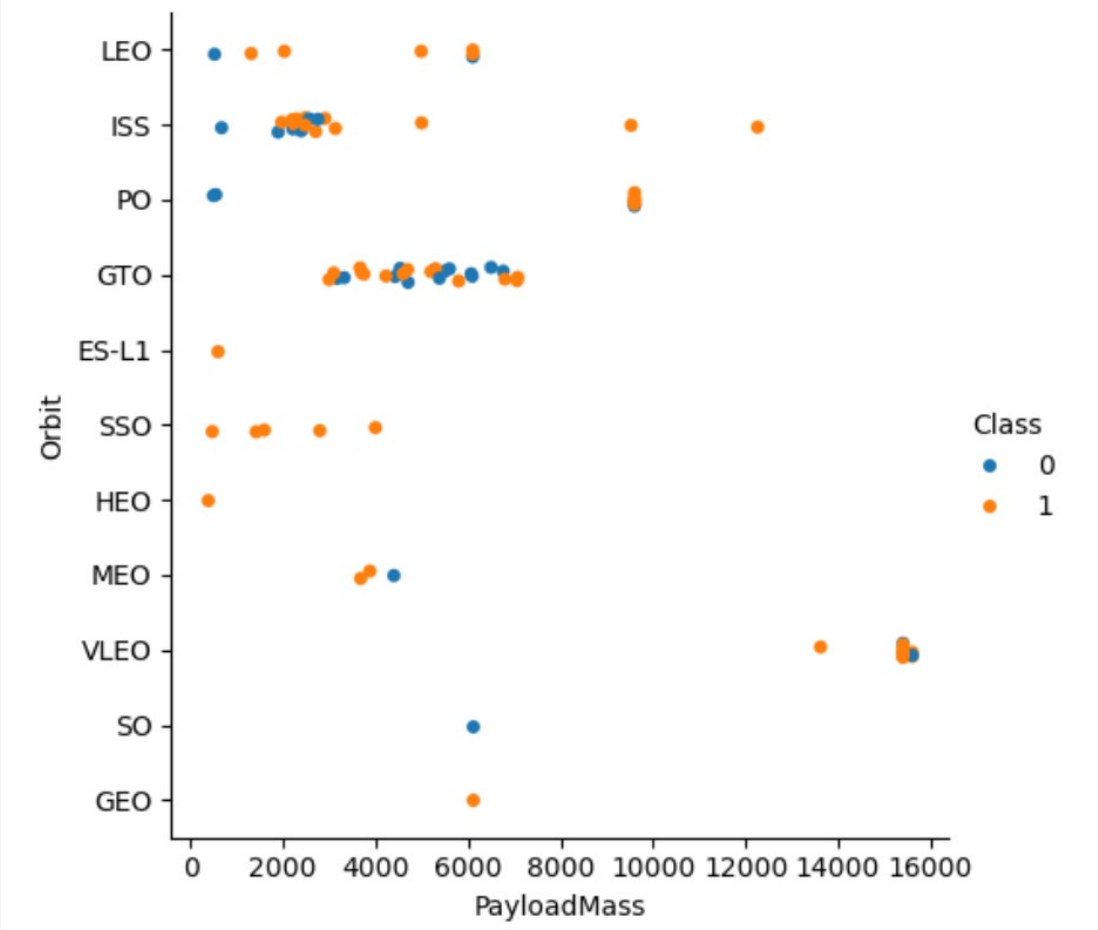
Flight Number vs. Orbit Type

- The number of flights was different according to the orbit.
- For MEO, VLEO, SO, and GEO, the number of flight was high.



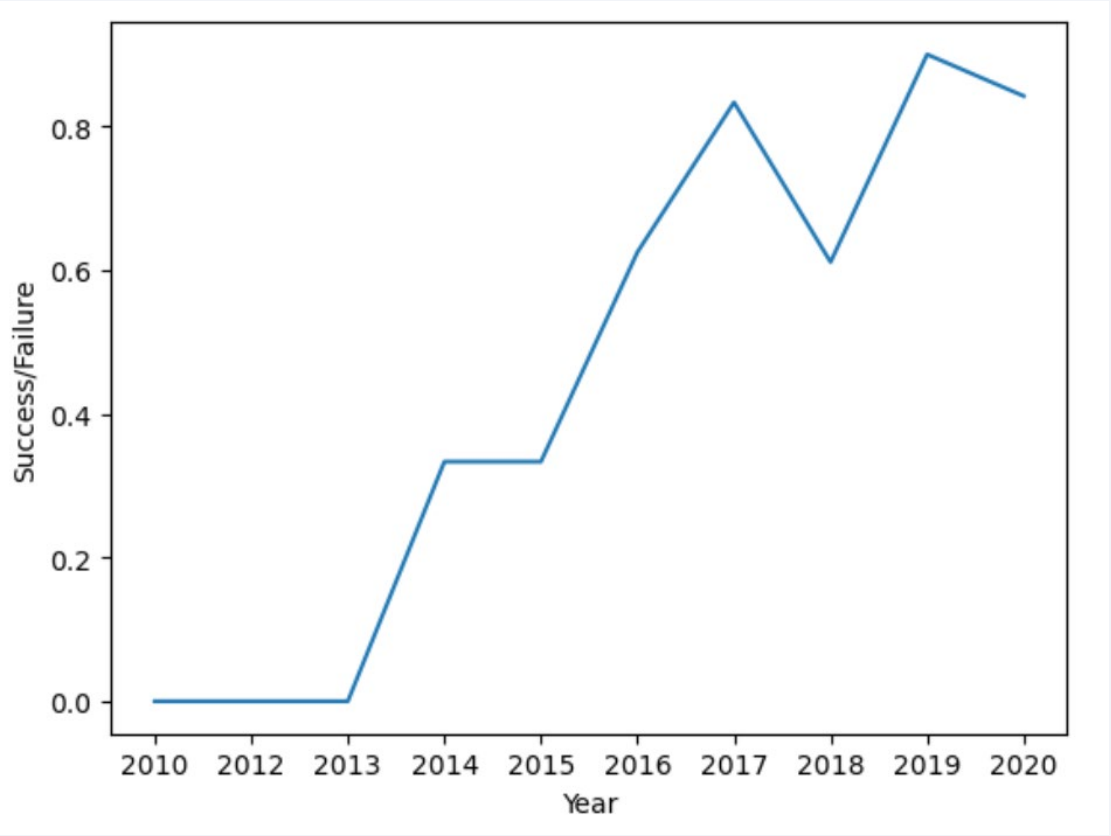
Payload vs. Orbit Type

- VLEO showed the heaviest pay load mass.



Launch Success Yearly Trend

- The success to failure rate went higher over years.



All Launch Site Names

- There were four different launch site.

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

Launch Site Names Begin with 'CCA'

- The launch site starting with CCA was ‘CCAFS LC-40’.
- SpaceX used once while NASA used four times.

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

- The total payload carried by boosters from NASA was 45,596 kg.

Average Payload Mass by F9 v1.1

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

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad was 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are given in the table.

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

Total Number of Successful and Failure Mission Outcomes

- The mission outcome shows 100 success and 1 failure.

```
%sql select count(Mission_Outcome) as Success from SPACEXTBL where Mission_Outcome Like '%success%';
```

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

Success
100

```
%sql select count(Mission_Outcome) as Failure from SPACEXTBL where Mission_Outcome Like '%failure%';
```

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

Failure
1

Boosters Carried Maximum Payload

- There were multiple boosters carried maximum payload.

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

2015 Launch Records

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

- Present your query result with a short explanation here

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

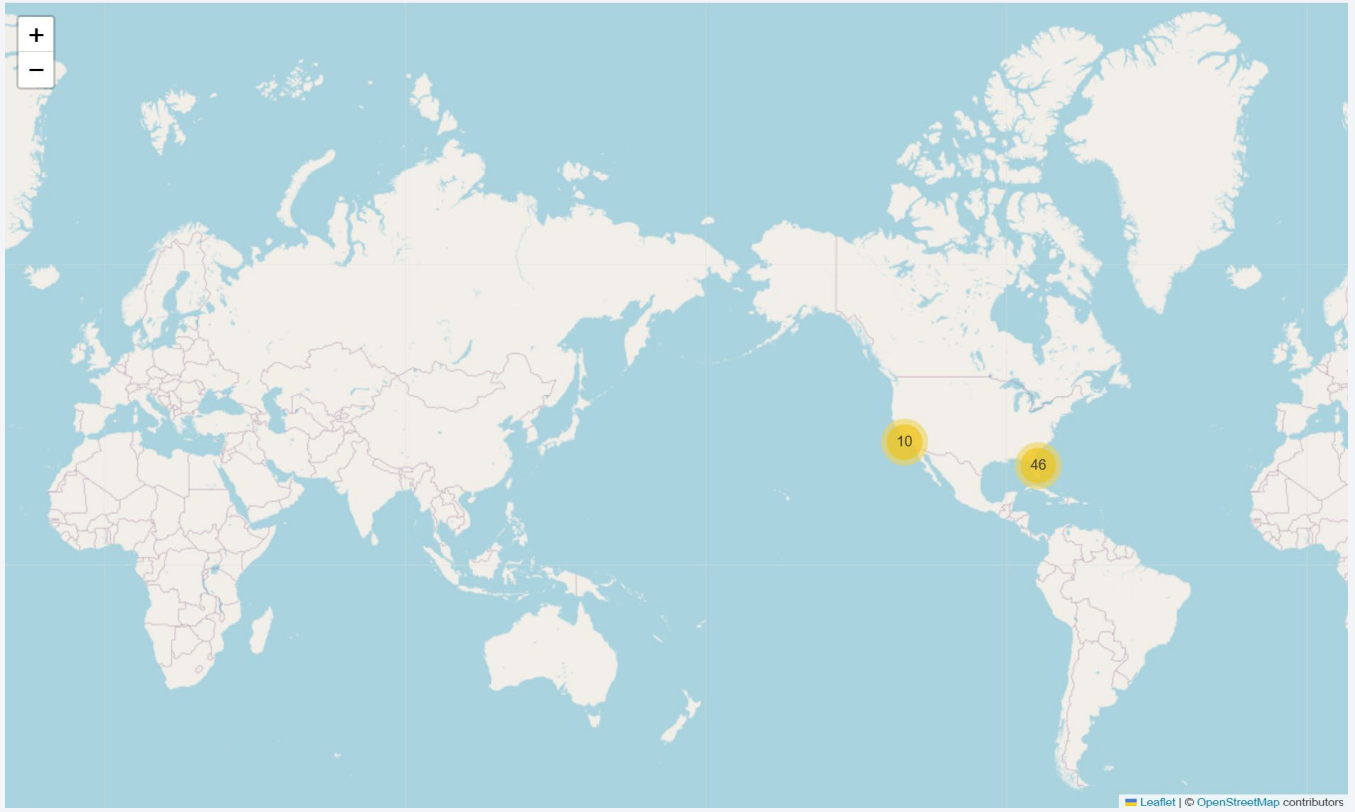
- the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing_Outcome	count(landing_outcome)	Date
Success (drone ship)	5	2016-04-08
Success (ground pad)	3	2015-12-22
Precluded (drone ship)	1	2015-06-28
Failure (drone ship)	5	2015-01-10
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
No attempt	10	2012-05-22
Failure (parachute)	2	2010-06-04


Section 3
Launch Sites
Proximities Analysis

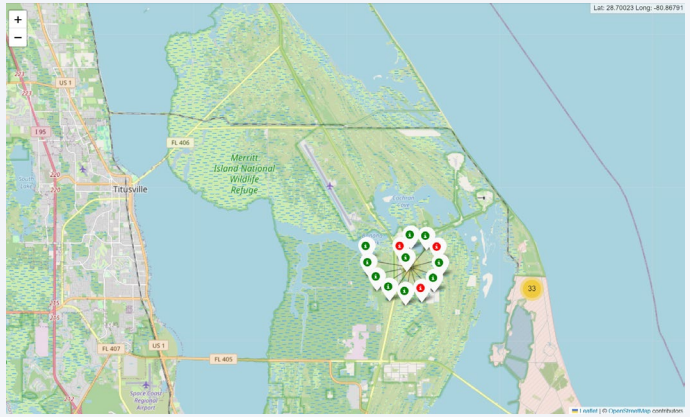
All launch sites' location markers on a global map

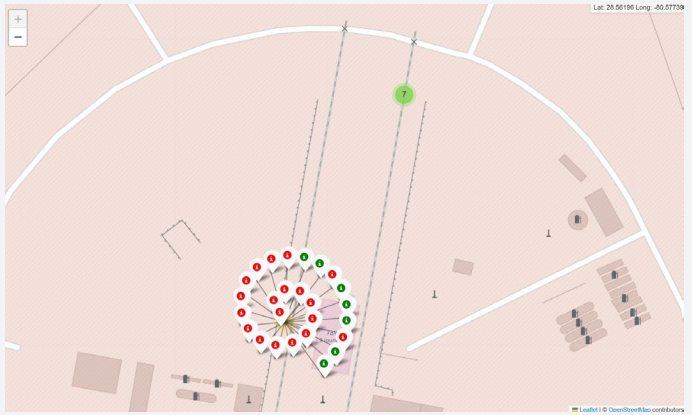
- The launch sites are located in CA and FL in the USA.



Launch outcomes on the map

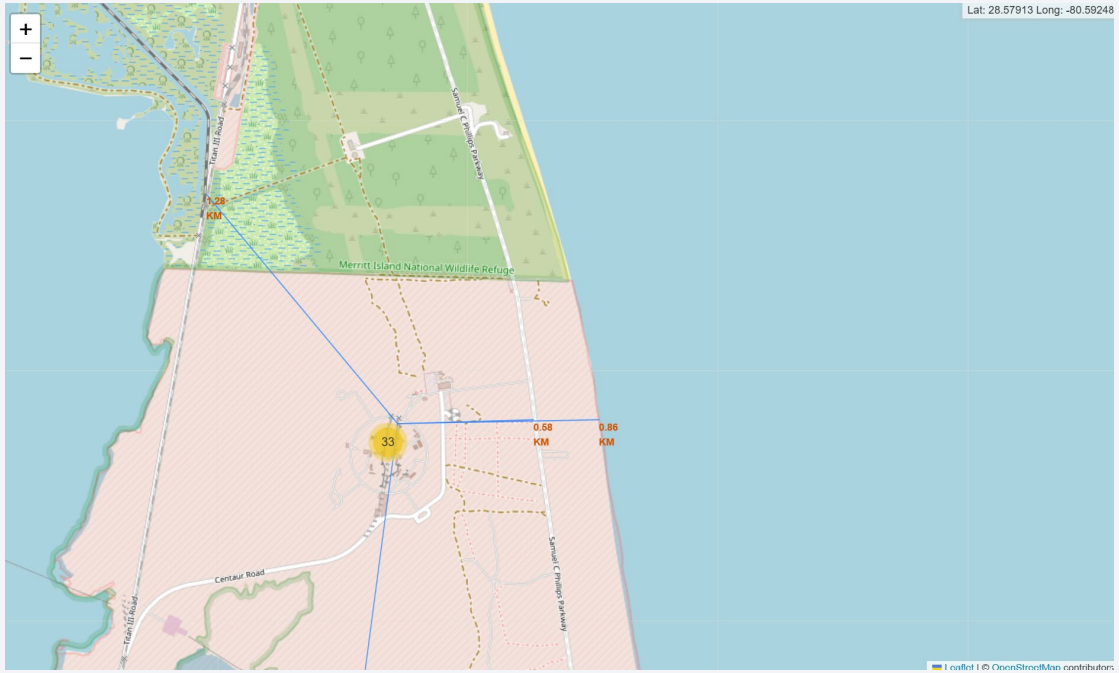
- The following figures shows the launch outcomes in Florida.
- Red mark shows failure; Green mark shows success.

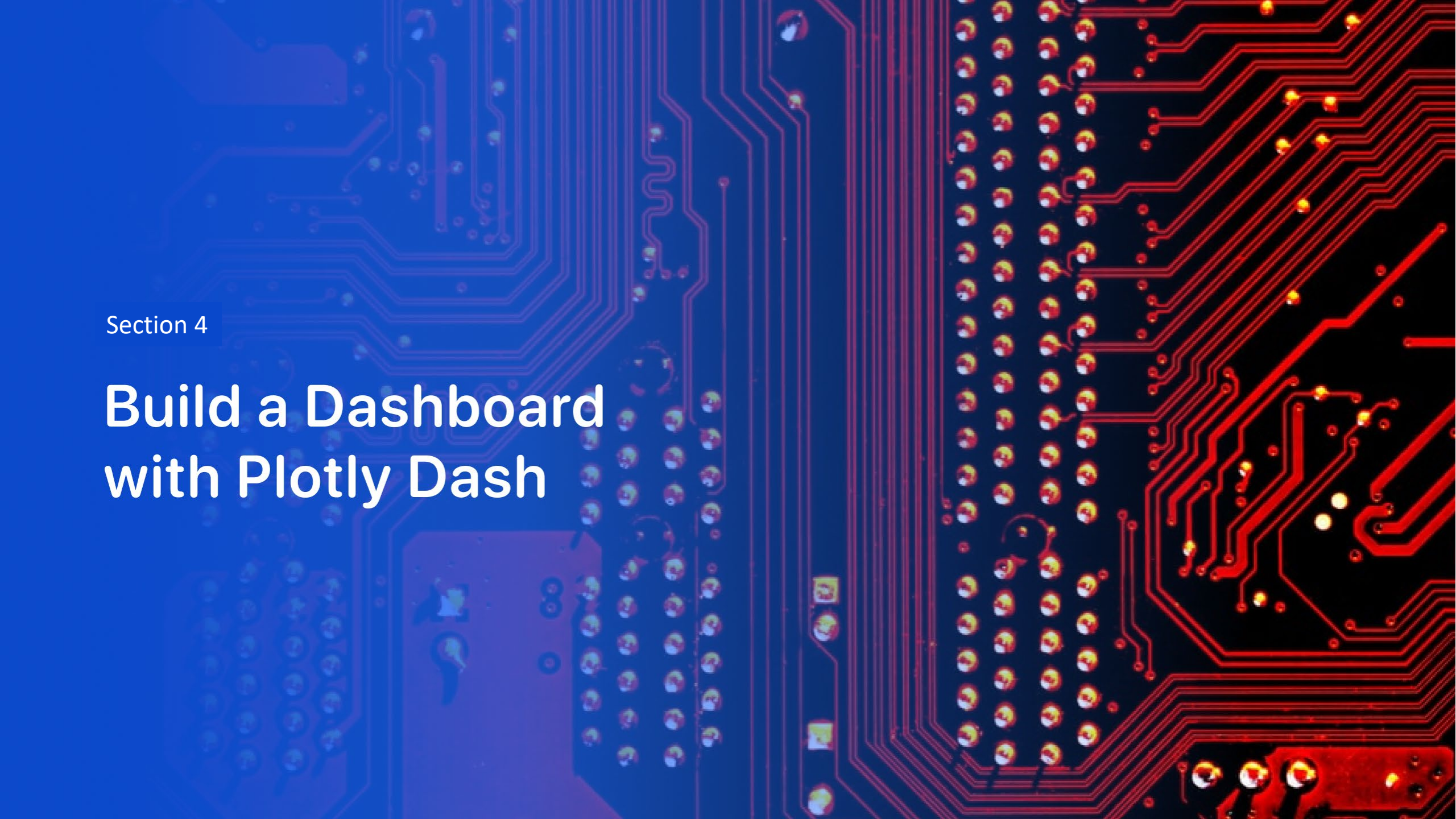




Proximities to railway, highway, and coastline

- The following map shows the proximities between a selected launch site and railway, highway, coastline, with distance calculated and displayed




Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites

Launch site with highest launch success ratio

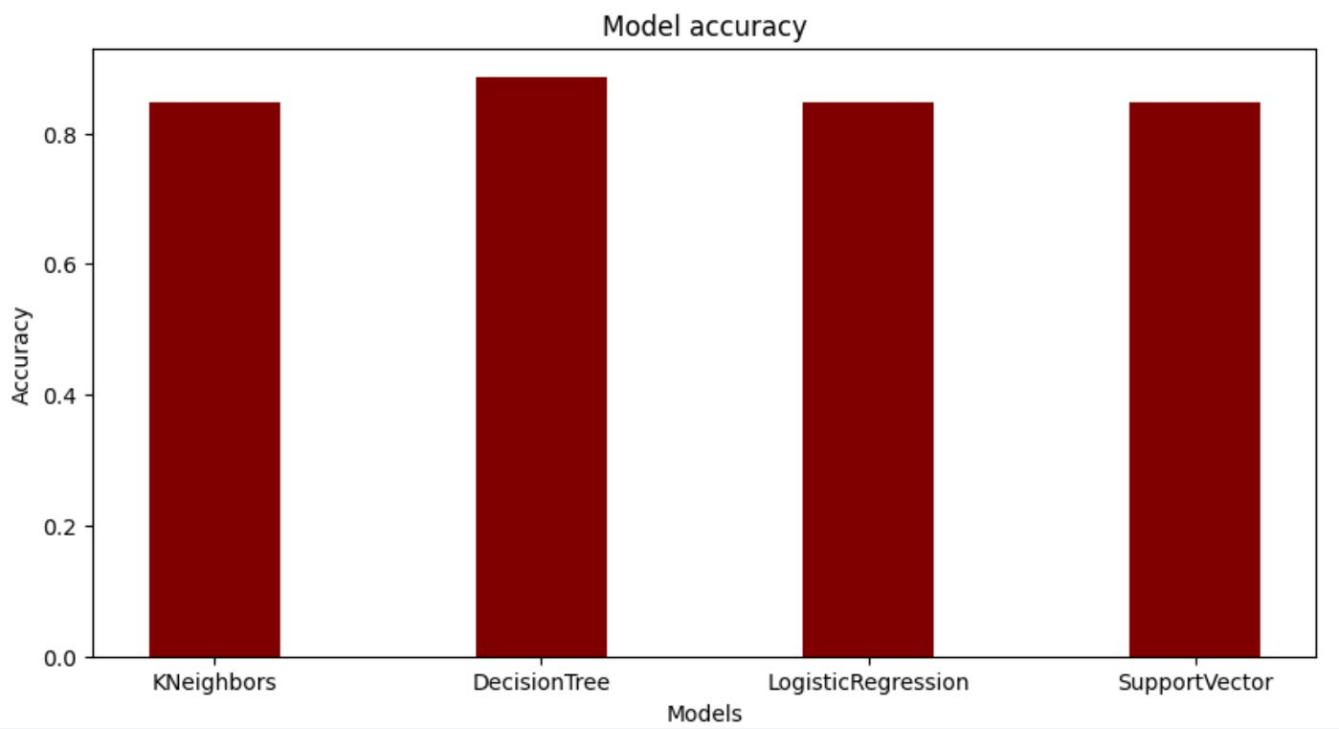
Payload vs. Launch outcome for all sites

Section 5

Predictive Analysis (Classification)

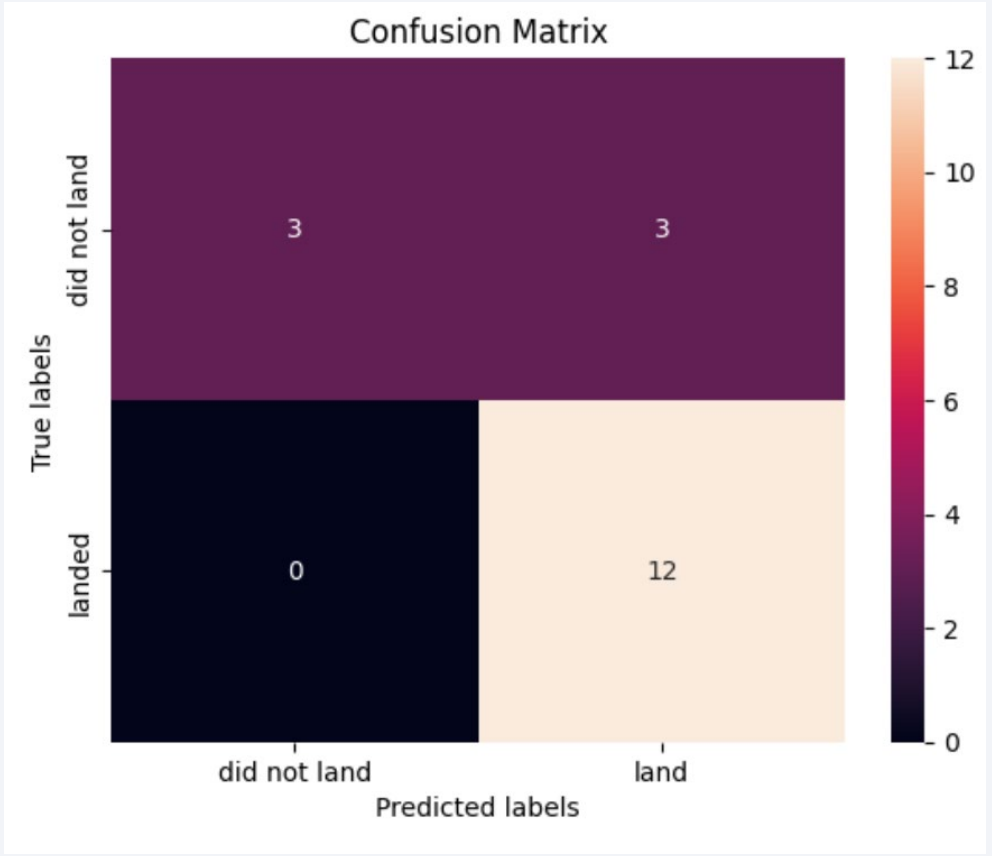
Classification Accuracy

- The decision tree model shows the highest accuracy among the classification models.



Confusion Matrix

- The accuracy of the decision tree model is .8857.



Conclusions

- Collected SpaceX rocket launch data using API
- Data wrangling
  - Filtered data to include Falcon9 data only
  - Replaced missing values
- Exploratory data analysis
  - Number of launches on launch site, orbit, and pay load mass
  - Identified trends by visualization
- Predict analysis using classification models
  - Calculated model performances of each model
  - Visualized the outcome by confusion matrix

Thank you!