Winning Space Race with Data Science

Sneha KK

July 6 2022

Outline

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

Executive Summary

- Methodologies used in this project:
- Data Collection: SpaceX API and web Scraping.
- Data Wrangling.
- Exploratory Data Analysis: SQL, data visualization using matplotlib, folium and dash.
- Predictive Analysis: Classification using Machine Learning algorithms.
- Results:
- Successful data collection, cleaning and feature engineering.
- Visualizations helped in deriving insights.
- Derived the best machine learning model for prediction based on accuracy of performance of the tested models.

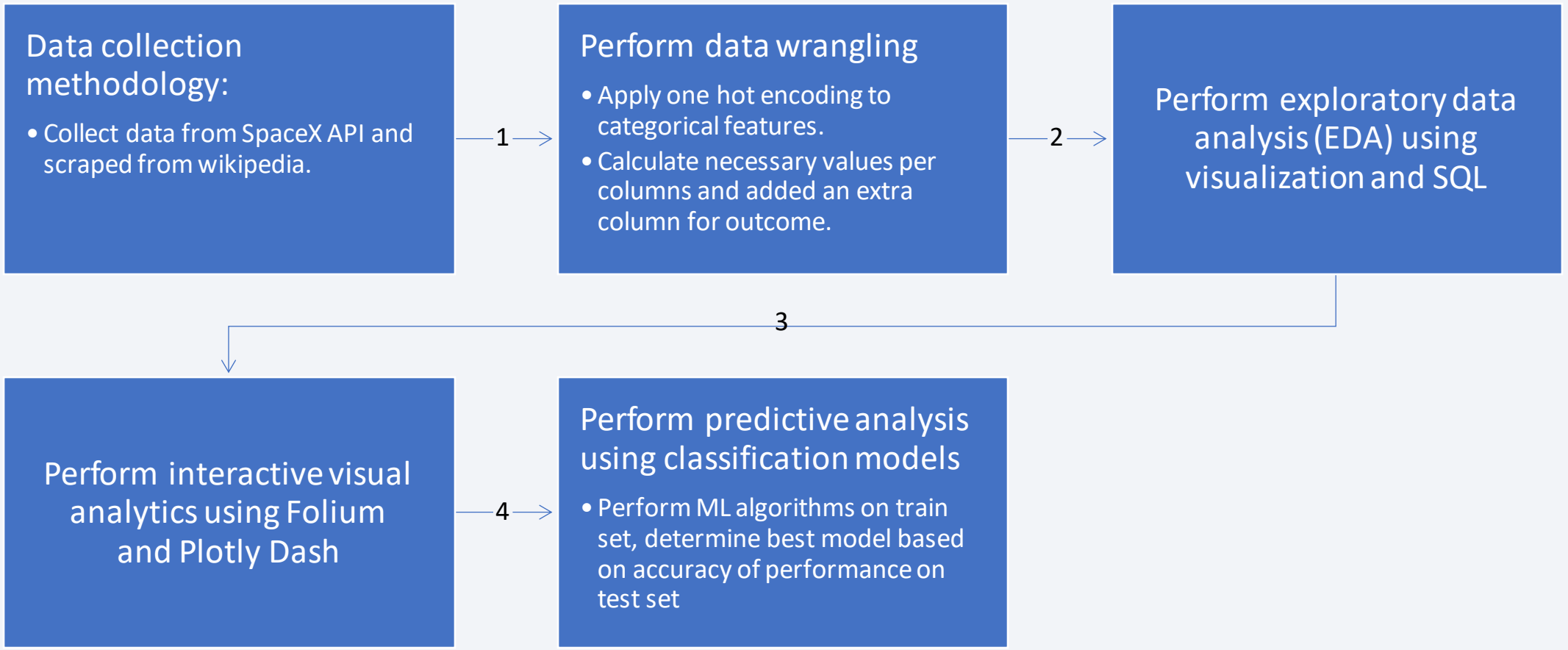
Introduction

- SpaceX is the most successful spacecraft engineering company till date.
- The purpose of this project is to assist a new spacecraft company Space Y to compete with Space X by analyzing datasets and making predictions.
- SpaceX saves most of its money by reusing the first stage of the rocket, and therefore if we can predict if the first stage will land, we can determine the cost of a launch.
- Considering these factors, our main goals will be :
- To understand the deterministic factors of success rate.
- To determine the success rate of landing.

Section 1

Methodology

Methodology



Data Collection

1. SpaceX API:

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json

Collection using get requests.

2.Wikipedia:

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

Scaping using BeautifulSoup.

Data Collection – SpaceX API

Code: https://github.com/Sneha-K-K/IBMDataSciencecapstone/blob/main/Data%20Collection.ipynb

01

1. Build necessary helper functions to get information

02

2.Request and parse the SpaceX launch data using the GET request

03

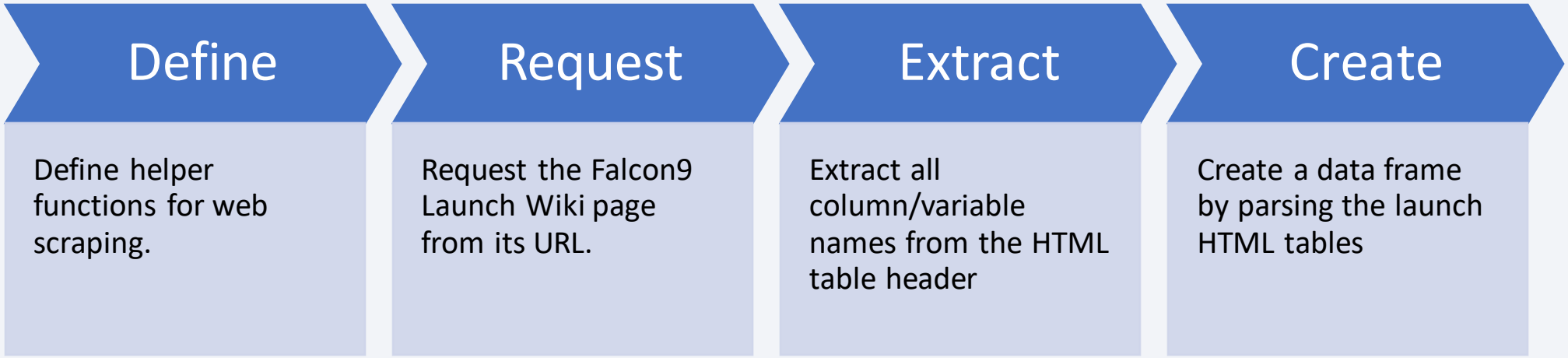
3. Filter the dataframe to only include Falcon 9 launches

04

4.Deal with missing values

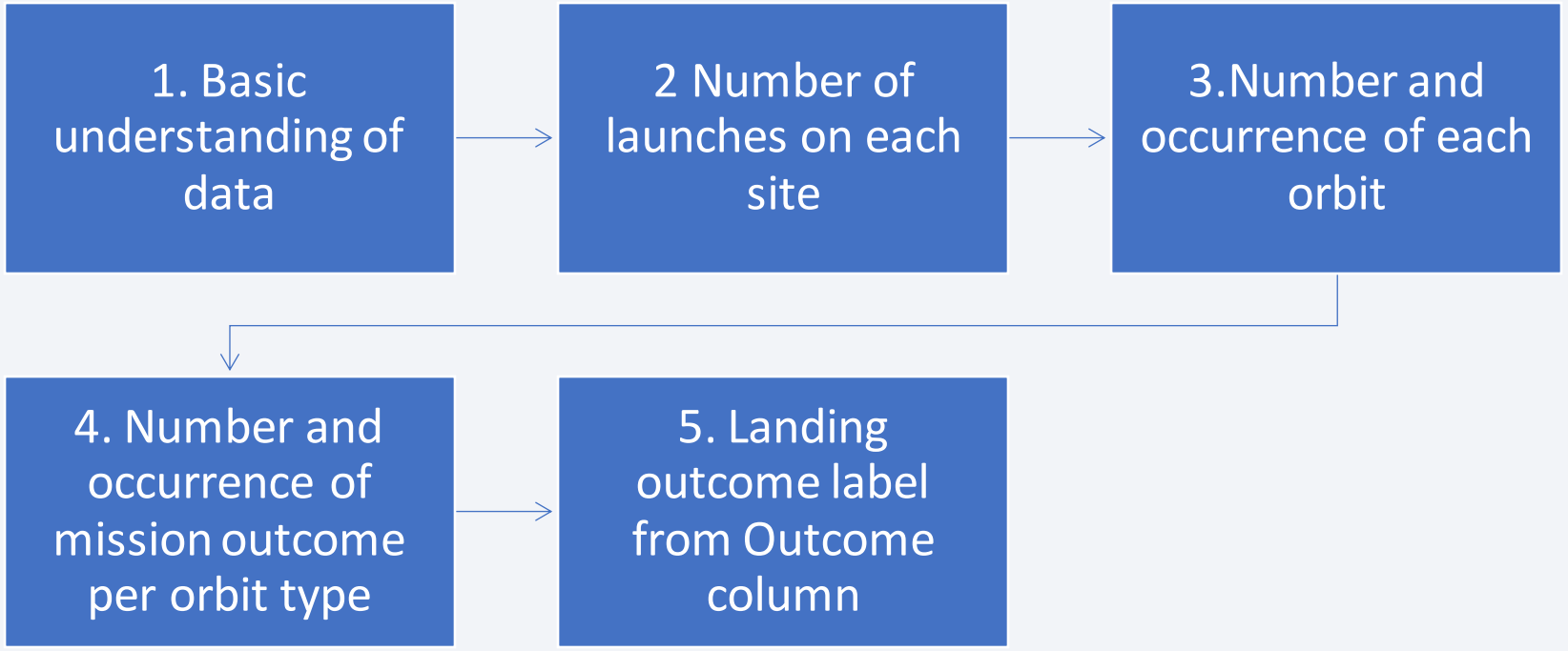
Data Collection - Scraping

Code: https://github.com/Sneha-K-K/IBM-DataScience-capstone/blob/main/Web%20scraping.ipynb



Data Wrangling

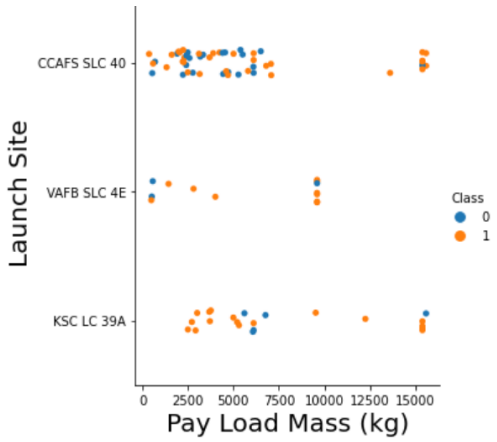
Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

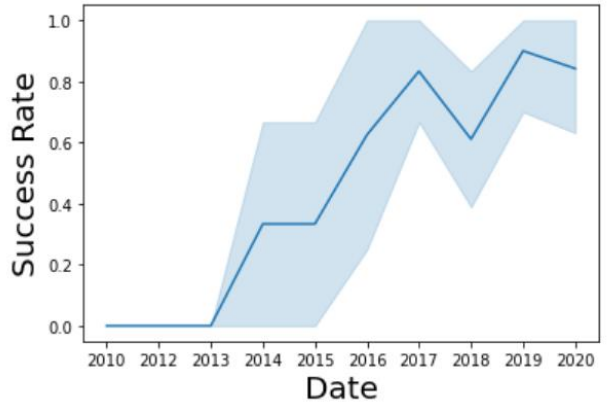


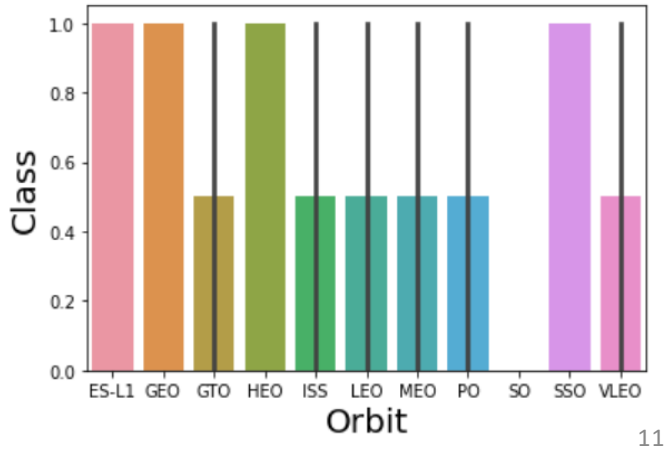
Code: https://github.com/Sneha-K-K/IBM-DataScience-capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- Basic bar plots, scatter plots and line plots were used to visualize the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type and the annual launch success.







Code: https://github.com/Sneha-K-K/IBM-DataScience-capstone/blob/main/Data_Visualization.ipynb

EDA with SQL

After loading the Postgre SQL database in jupyter notebook, some SQL queries were conducted to understand the data, like:

- -The names of unique launch sites in the space mission.
- -The total payload mass carried by boosters launched by NASA (CRS)
- -The average payload mass carried by booster version F9 v1.1
- -The total number of successful and failure mission outcomes
- The failed landing outcomes in drone ship, their booster version and launch site names

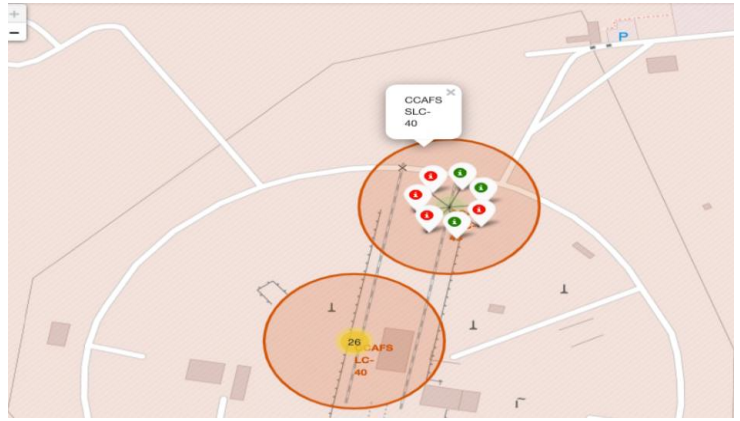
The resultant tables helped us to understand and analyze the data values from the dataset.

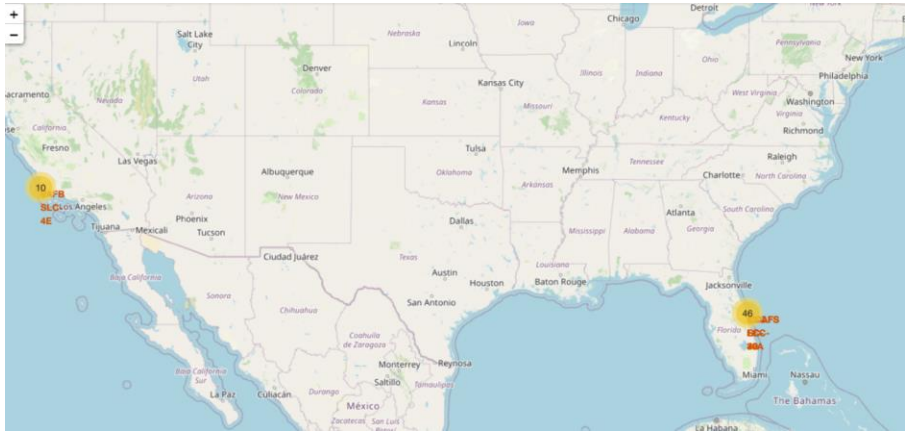
Code: https://github.com/Sneha-K-K/IBM-DataScience-capstone/blob/main/EDA%20with%20SQL.ipynb

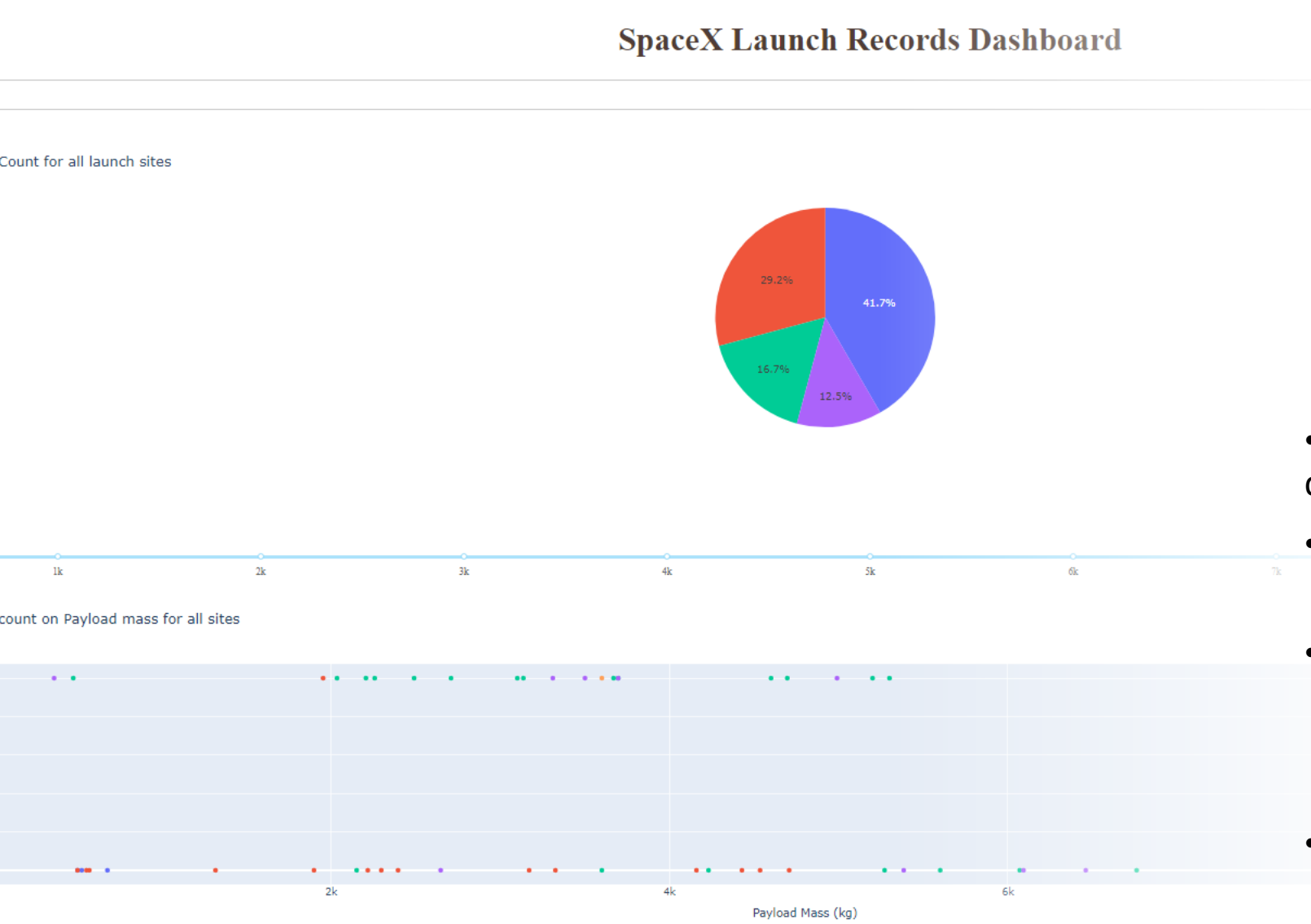
Interactive Map with Folium

- We marked all launch sites and used map objects like markers, circles, lines to indicate launch sites, coordinates and distances, launches in a launch site and success/failure.
- We calculated the distances between a launch site to its proximities and checked if launch sites near railways, highways and coastlines and if launch sites keep certain distance away from cities.

Code: https://github.com/Sneha-K-K/IBM-DataScience-capstone/blob/main/Interactive%20visual%20analytics%20with%20folium.ipynb





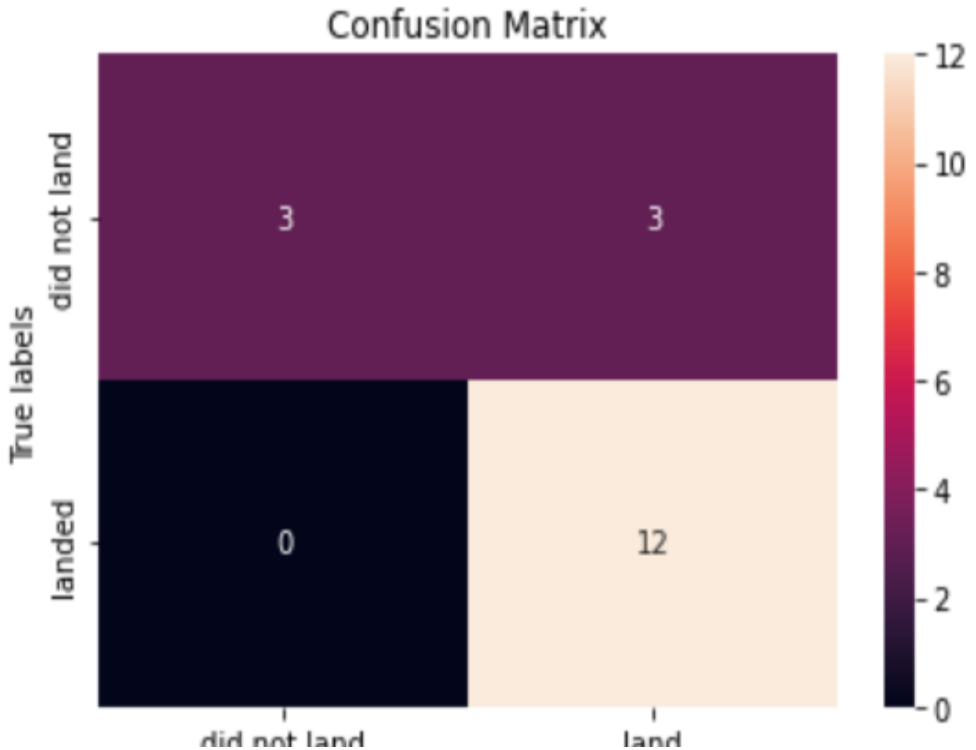


Dashboard with Plotly Dash

- An interactive dashboard with Plotly dash was built.
- Pie charts showing the total launches by a certain sites were plotted
- Scatter graph showing the relationship with Outcome and Payload Mass for the different booster version was plotted.
- Code: https://github.com/Sneha-K-K/IBM-DataScience-capstone/blob/main/Spacex%20dash%20app

Predictive Analysis (Classification)

- Data was loaded and split into train and test sets/
- Four different machine learning models: Logistic regression, support vector machines ,decision tree classifier and K nearest neighbor were used and compared
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- The best performing classification model was identified.
- Code: https://github.com/Sneha-K-K/IBM-DataScience-capstone/blob/main/Machine%20Learning%20Prediction.ipynb

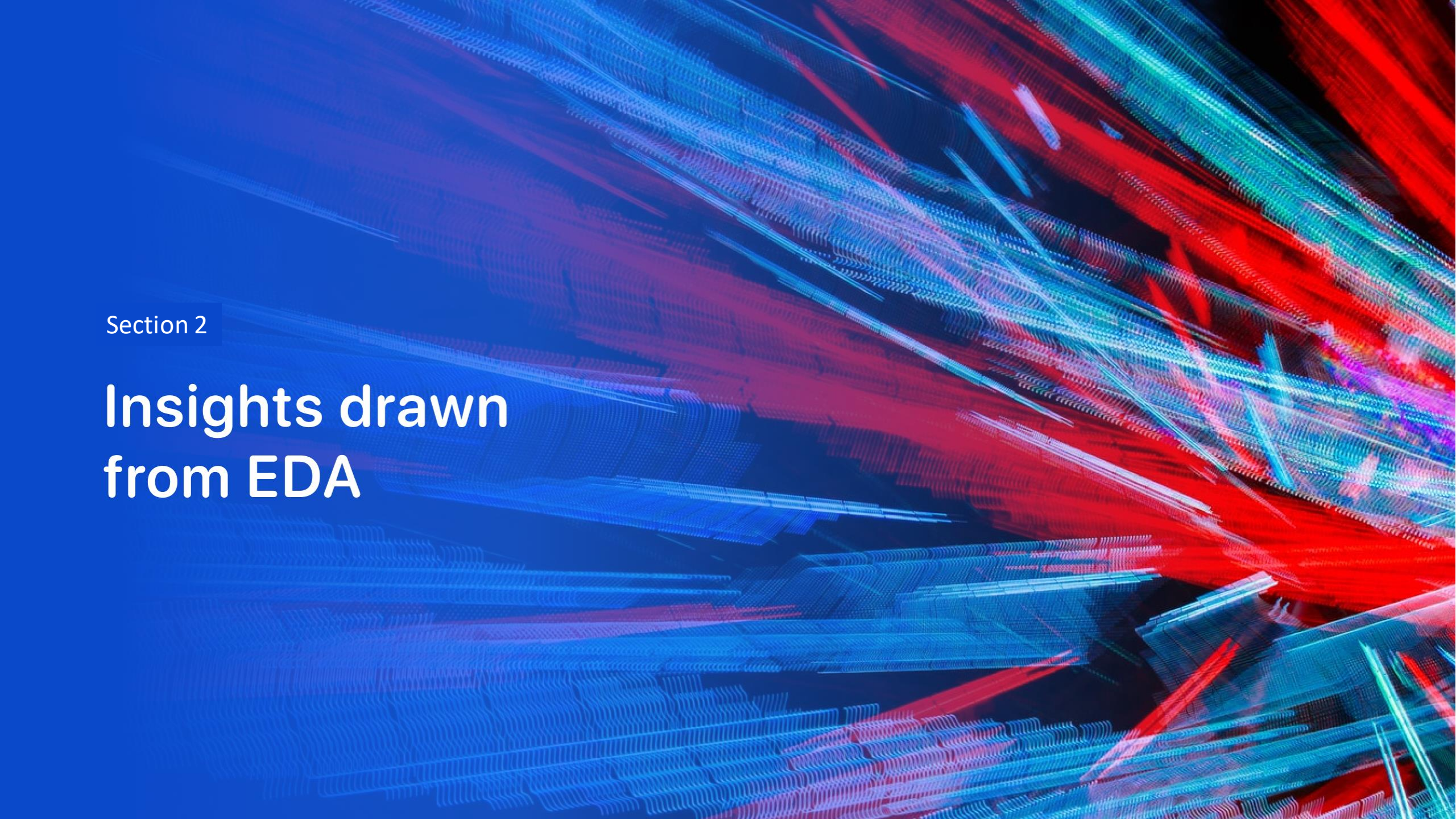


Results

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

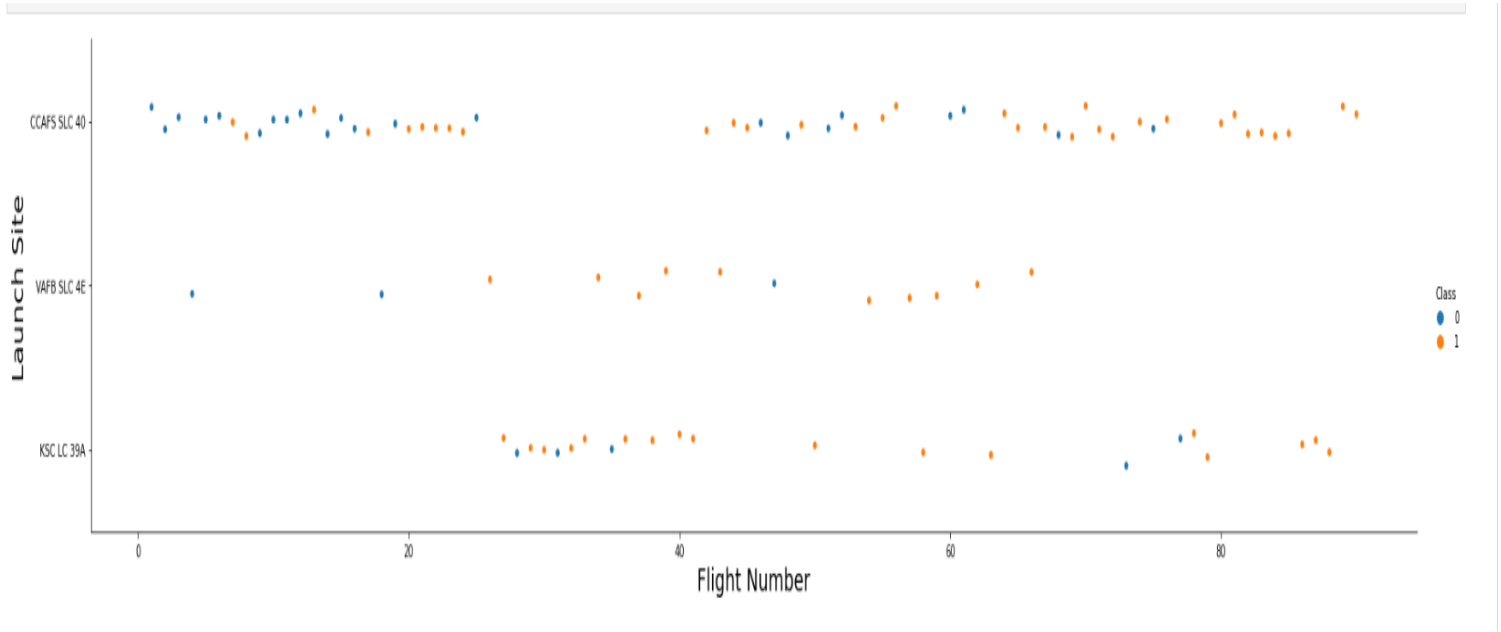
Section 2

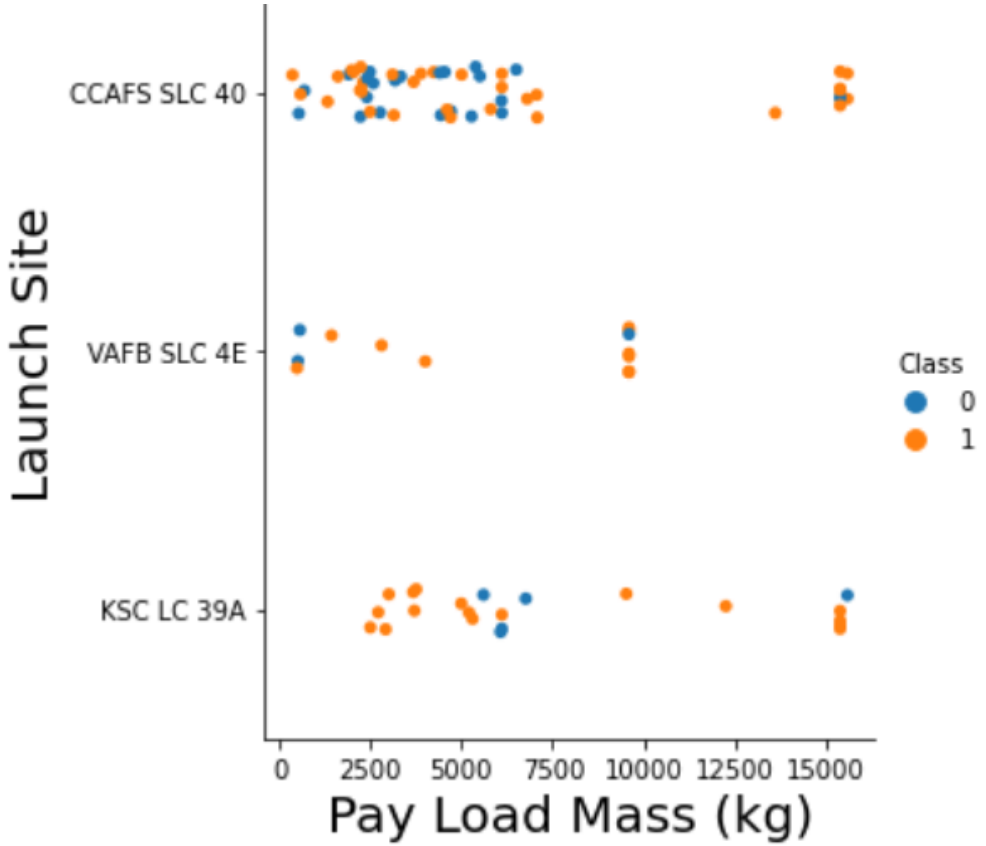
Insights drawn from EDA



Flight Number vs. Launch Site

- From the plot we can see that the success rate at a launch site is proportional to the flight amount.



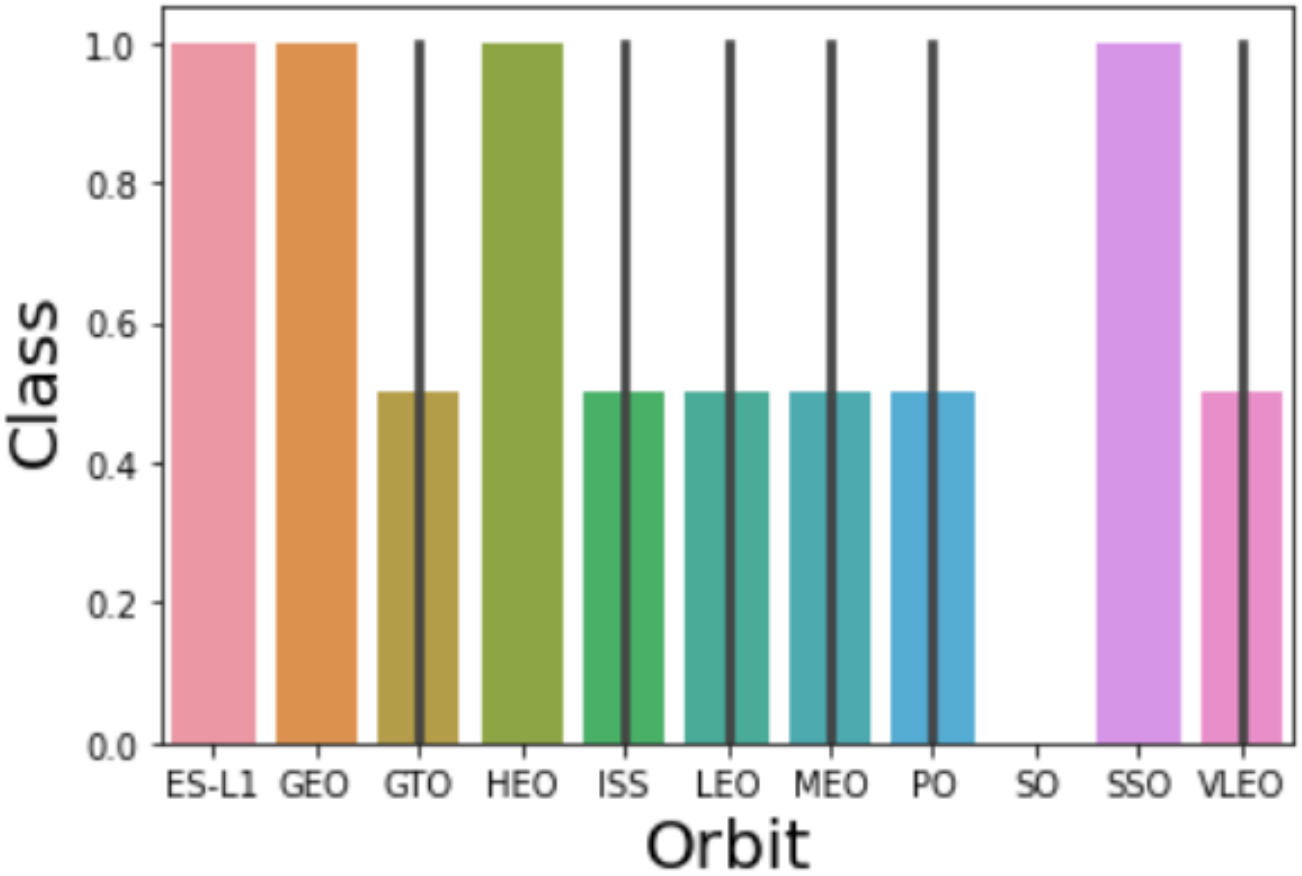


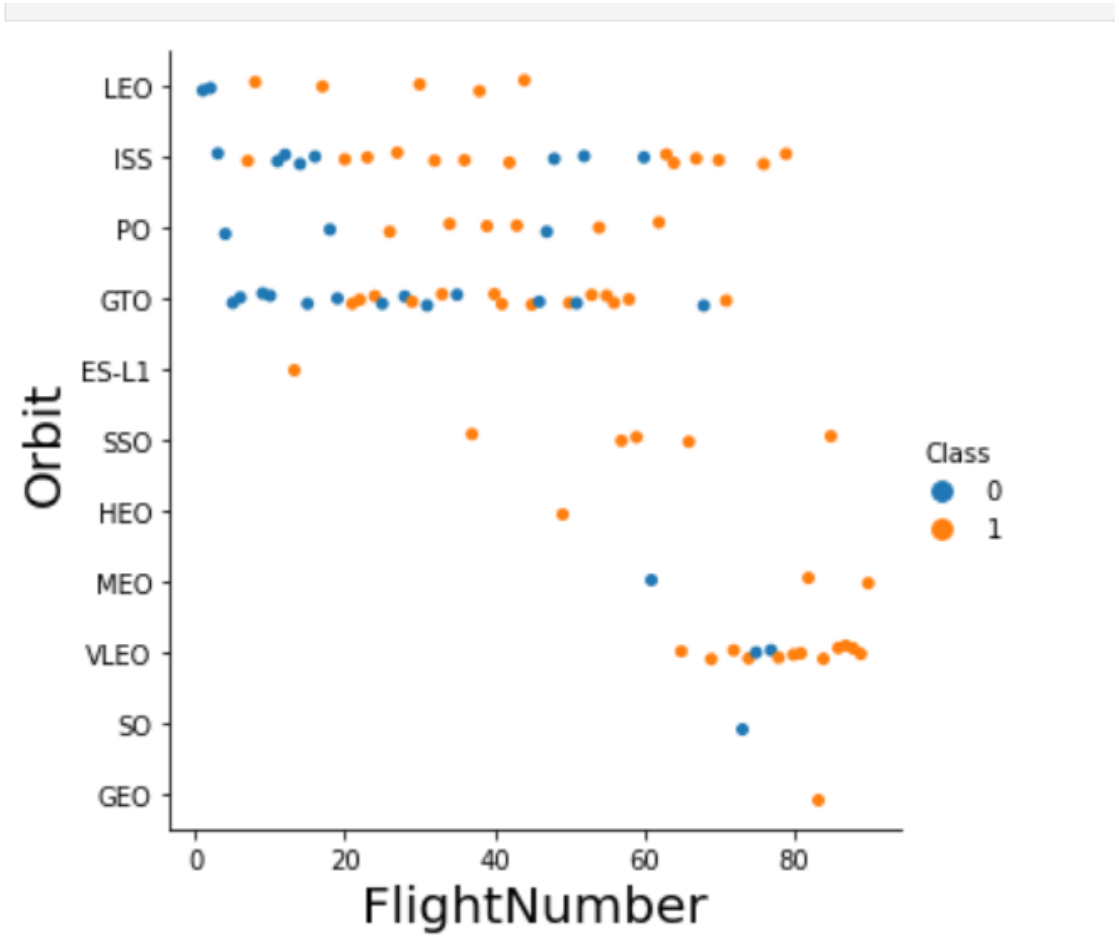
Payload vs. Launch Site

- Payloads less than 4000kg are more successful in general, except at CCAFS SLC 40 where heavier launches are more successful.
- Payloads more than 10000kg are less frequent and are launched only at CCAFS SLC 40 and KSC LC 39A

Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO have more success rate with almost all the launches being successful



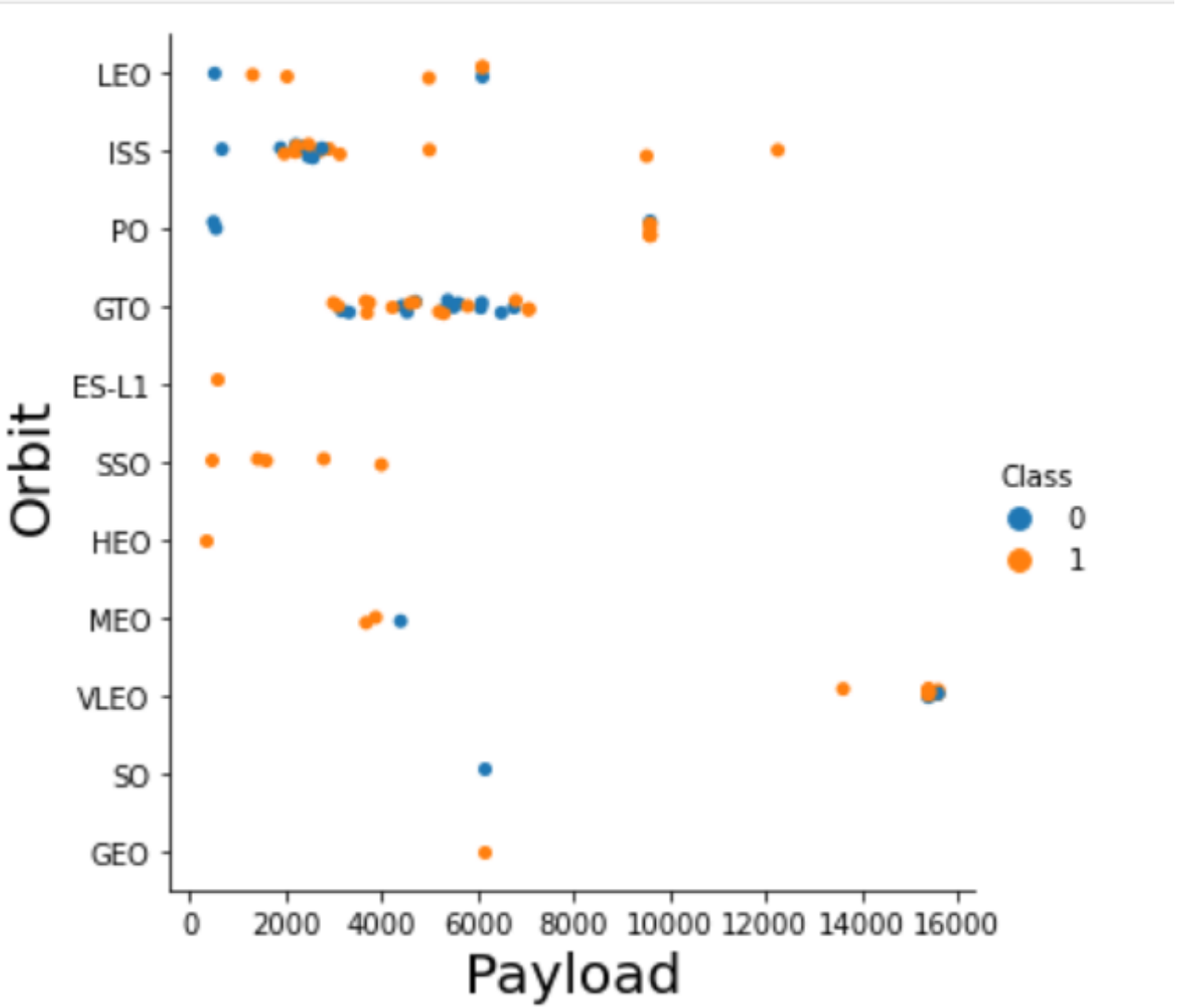


Flight Number vs. Orbit Type

- In LEO,ISS,PO orbits the success rates are proportional to number of launches.
- VLEO has been more successful in launching many rockets.
- GTO orbit's success rate does not seem to depend on number of launches.

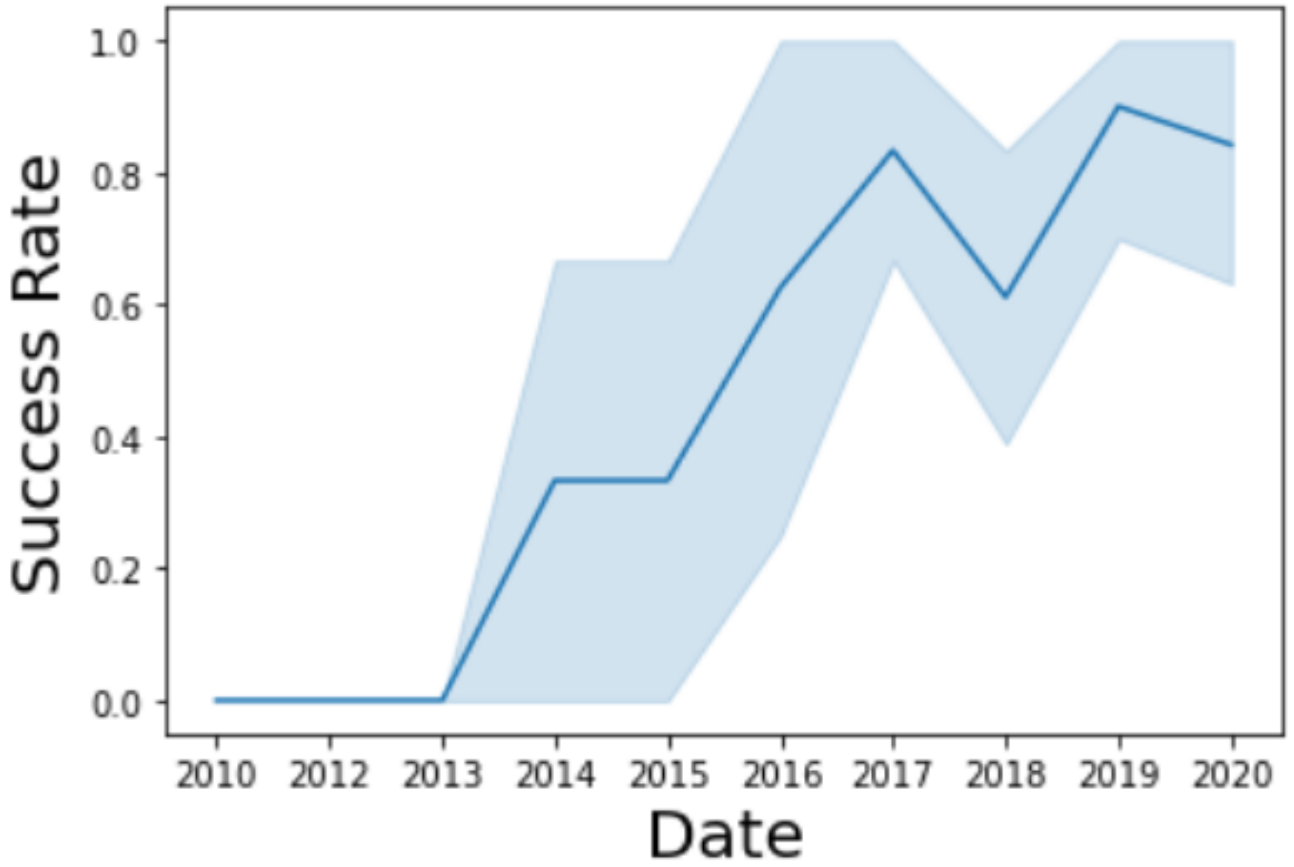
Payload vs. Orbit Type

- ISS orbit has the widest range of payload masses and GTO the narrowest.
- With heavy payloads, the successful landing are more for PO, LEO and ISS orbits.
- Only heavy load rockets are placed at VLEO.



Launch Success Yearly Trend

- Success rates started increasing since 2013 and launches are more successful with time.



All Launch Site Names

Find the names of the unique launch sites

```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`

```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Calculate the total payload carried by boosters from NASA

```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

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

```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_dat
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

firstsuccessfull_landing_date
2015-12-22

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

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

	failureoutcome
0	1

Boosters Carried Maximum Payload

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

```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

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

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

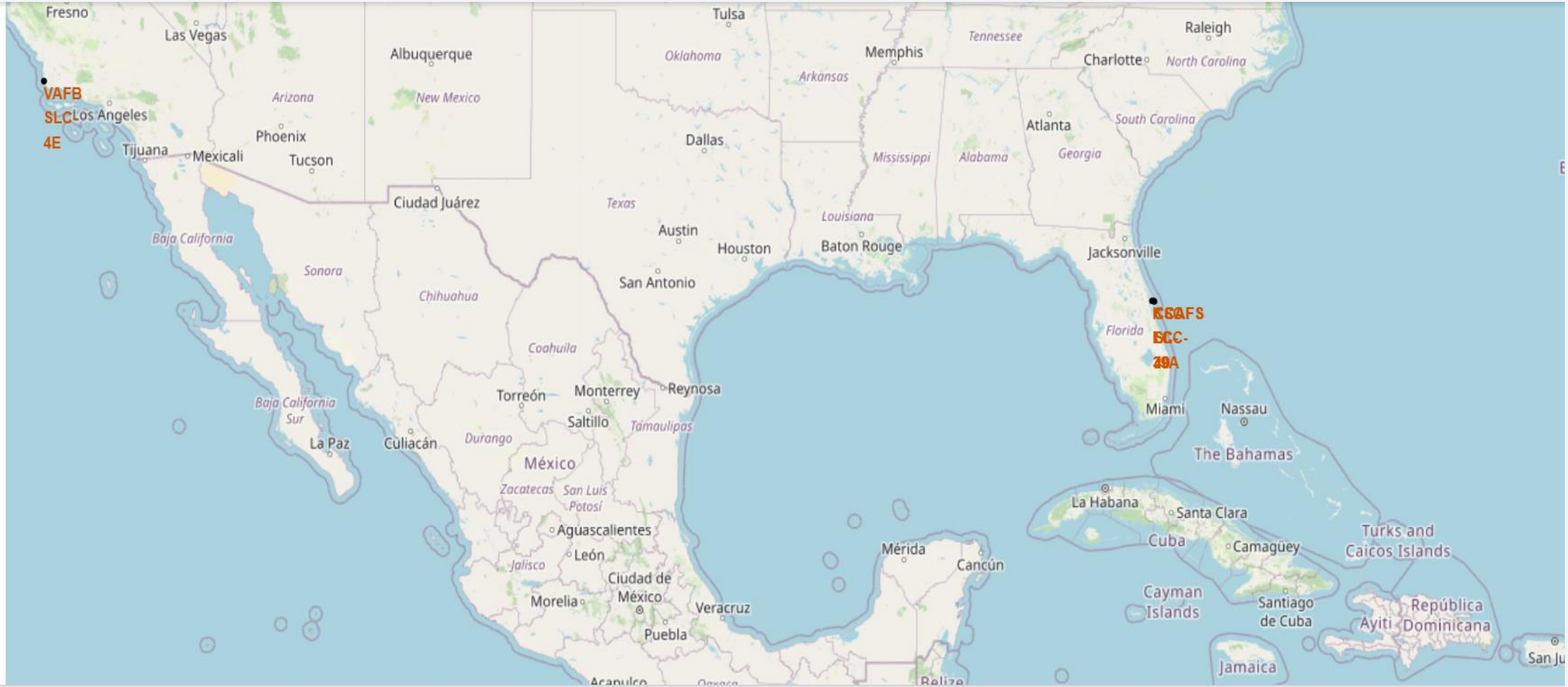
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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

```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

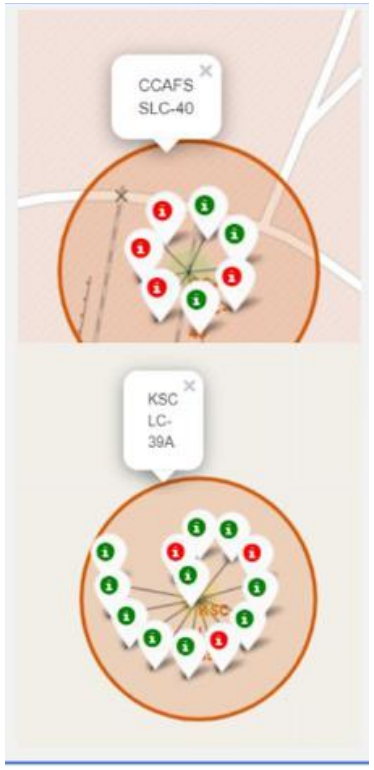

Section 3
Launch Sites
Proximities Analysis

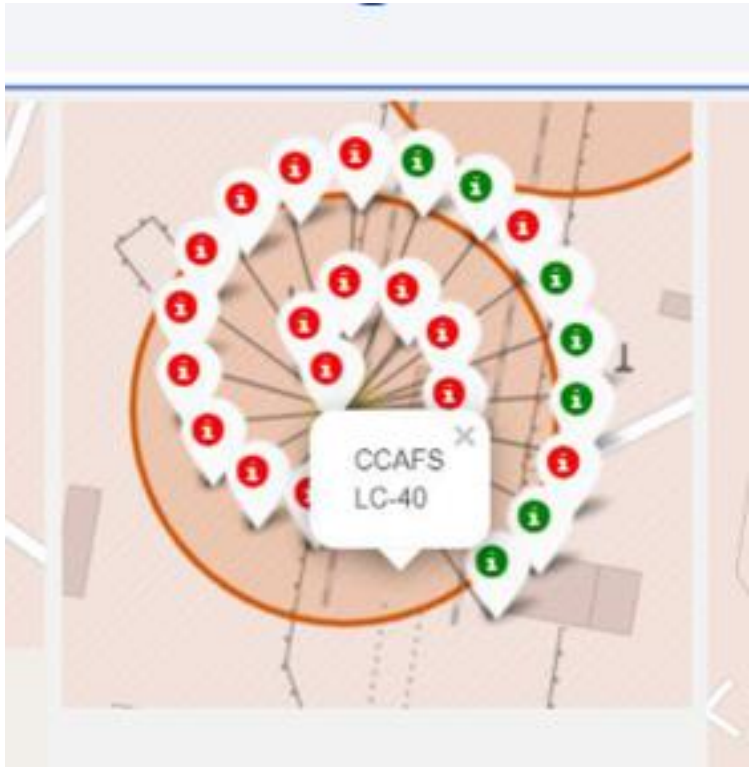


SpaceX launch sites are in California and Florida

Folium maps launch site analysis

- Green: success
- Red: Failure
- Florida and California launch sites





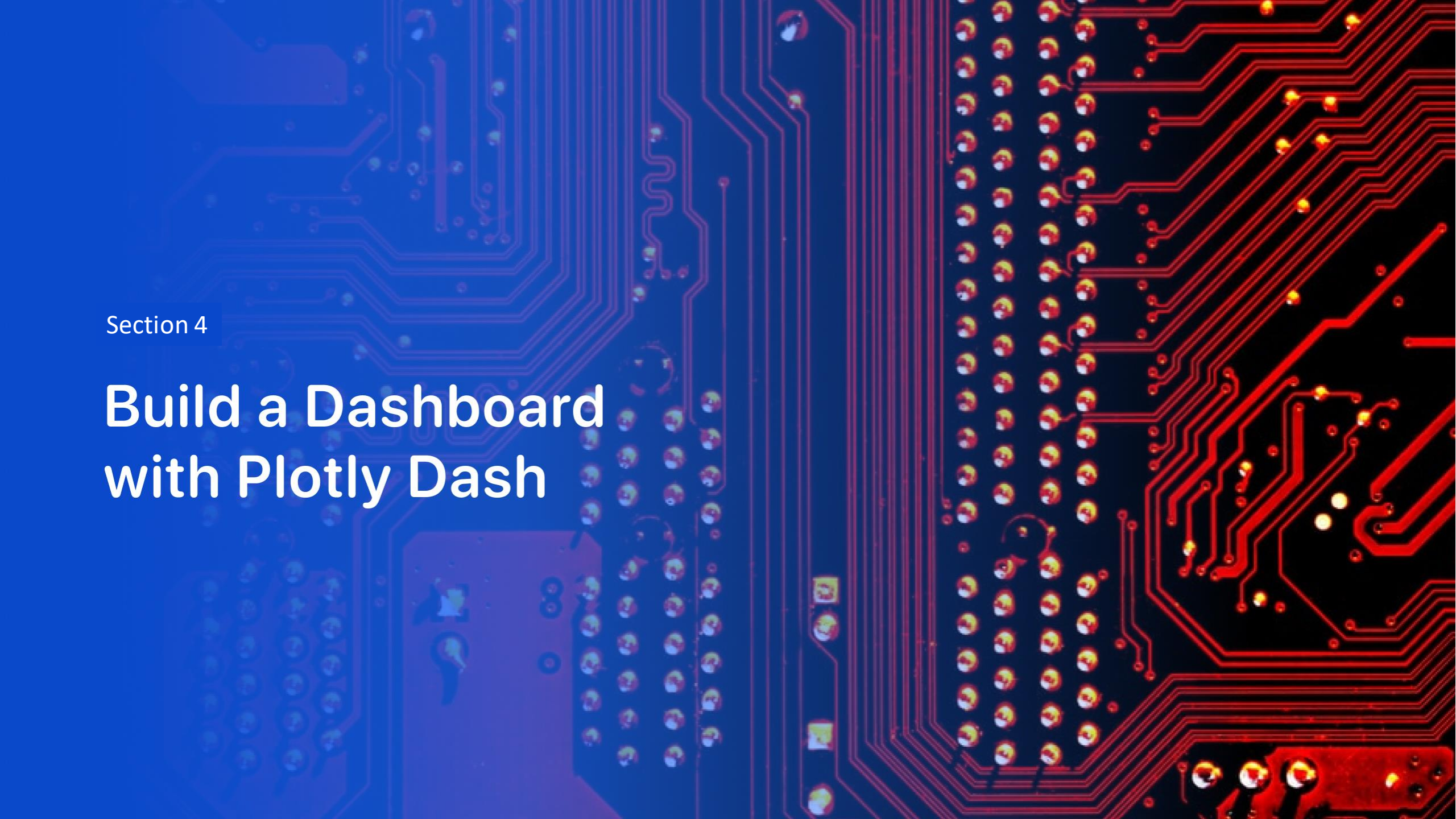
Folium maps launch site analysis

Railway Station: Considerable distance

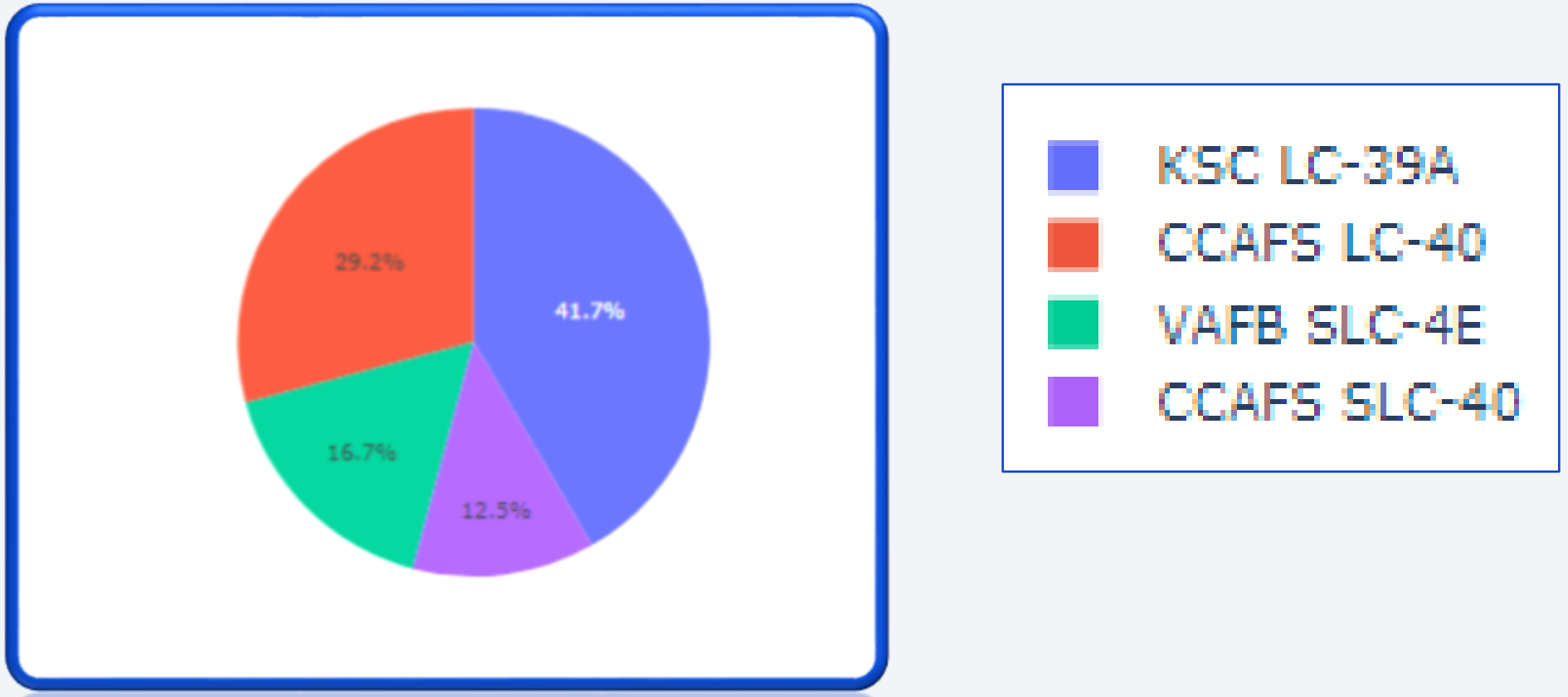
Highway : Considerable distance

Coastline :Near

City : Considerable distance

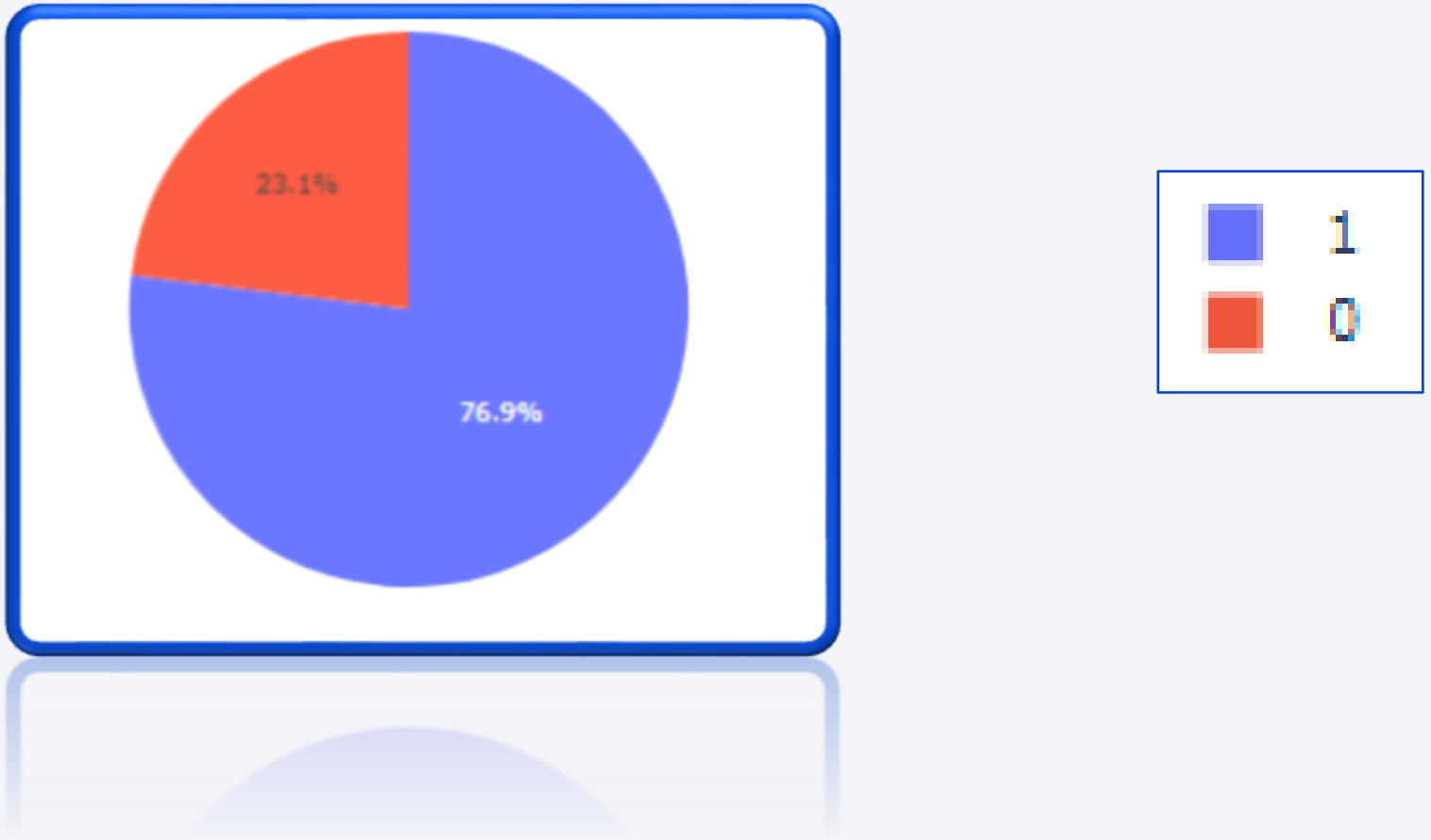

Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



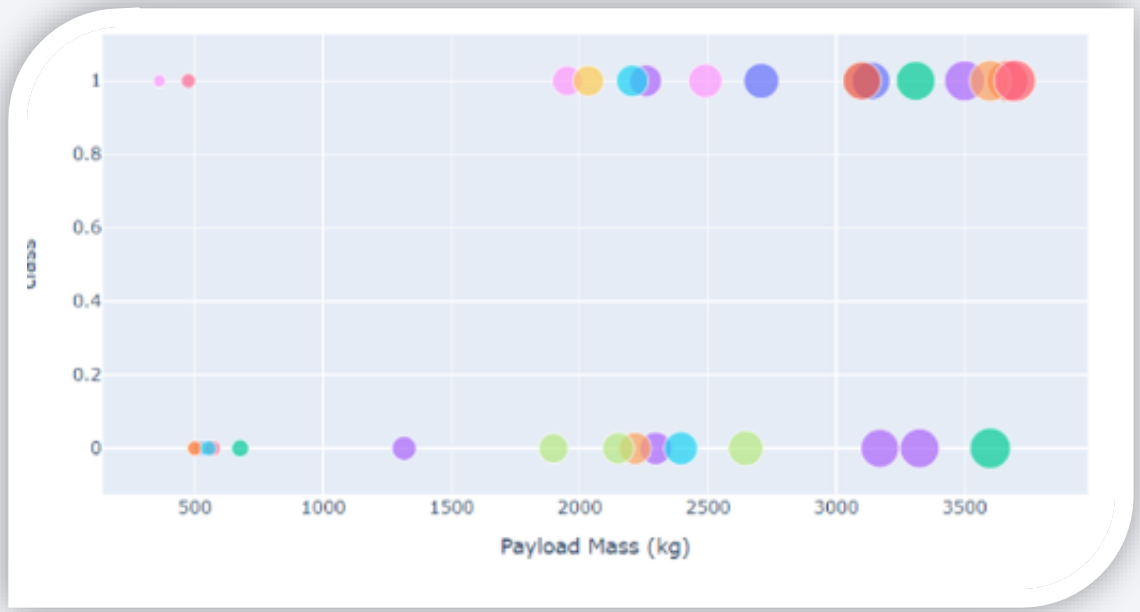
KSC LC-39A had the most successful launches while CCAFS SLC-40 had the least successful launches

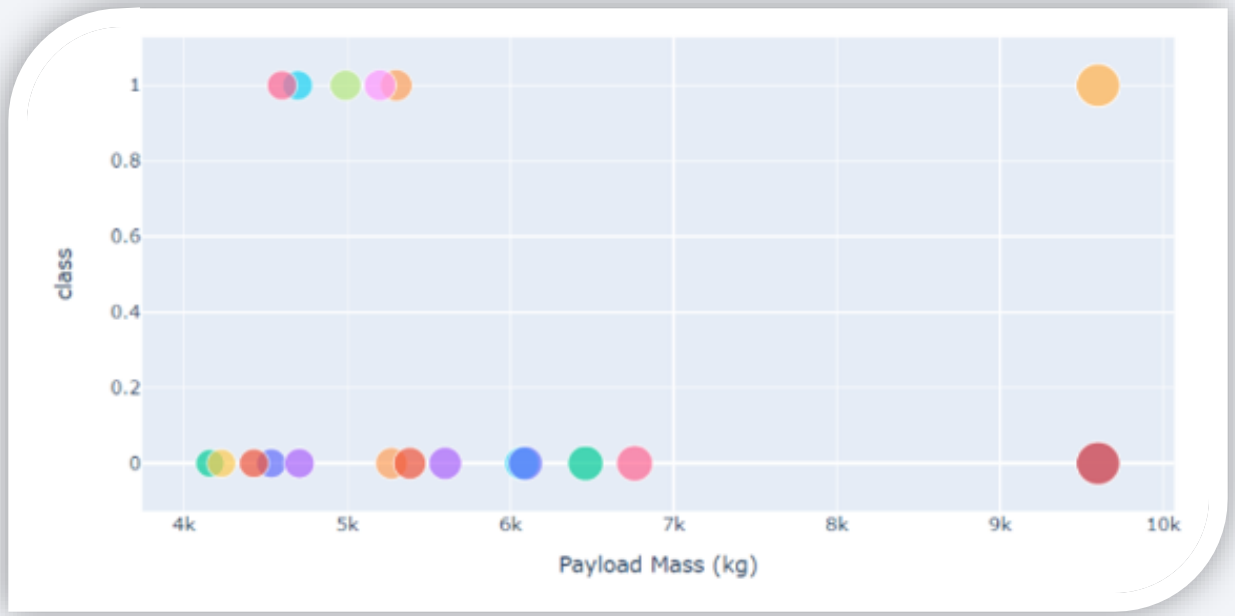
Pie chart showing the launch site with highest success ratio



KSC LC-39A had the highest success ratio of 76.9%

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider





Payload < 4000kg : more successful

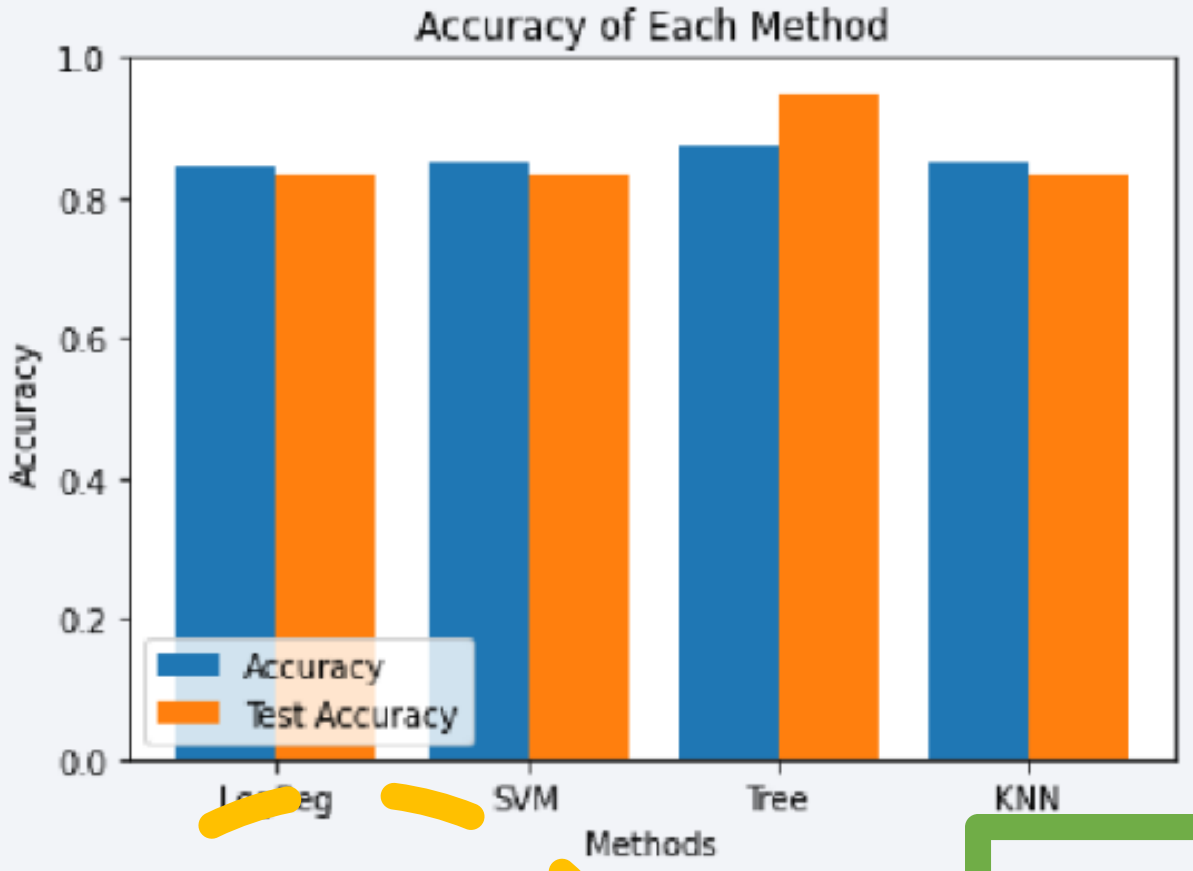
Payload>4000kg: less successful

Section 5

Predictive Analysis (Classification)

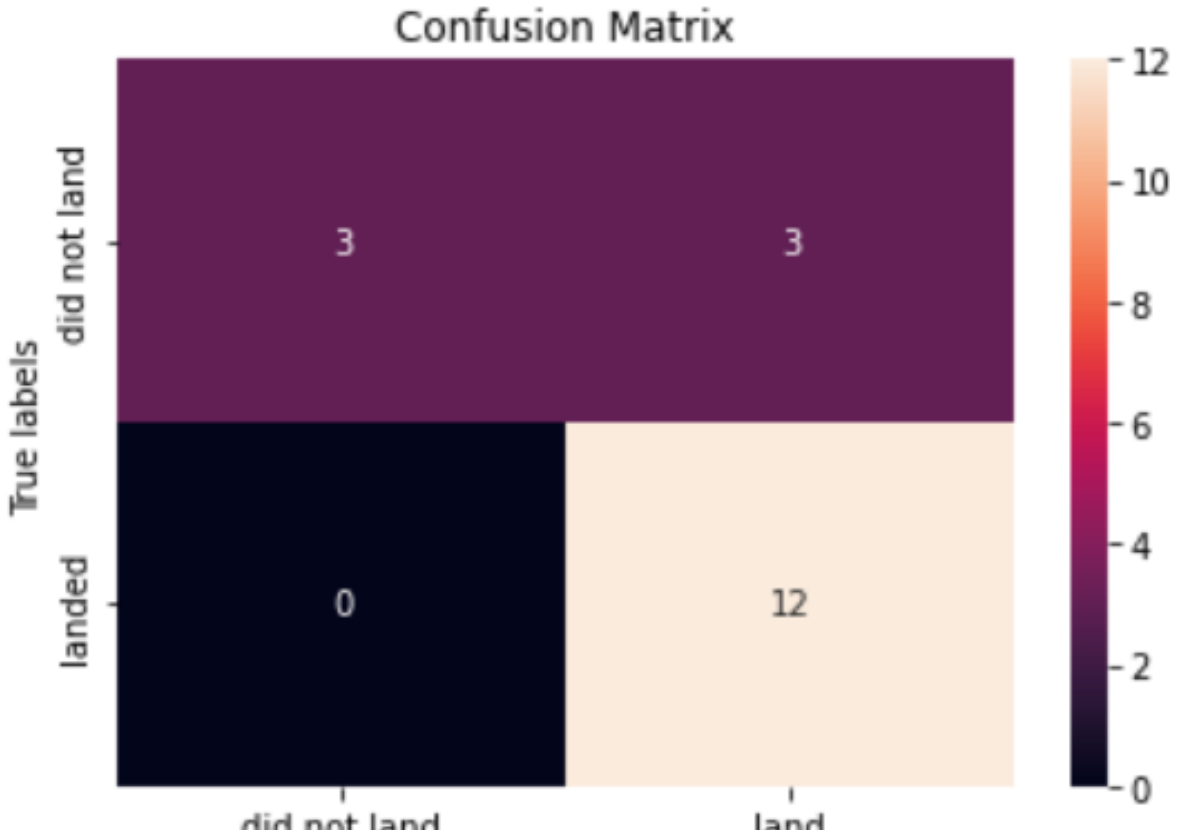
Classification Accuracy

- Four machine learning models : Logistic Regression, SVM, Decision Tree Classifier, KNN were used and their accuracy were calculated
- Decision Tree Classifier had the highest classification accuracy(>87%)



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.
- The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

- SpaceX's main budget save comes from reusing the first stage.
- Space X has launch sites in California and Florida with KSC-LC 39A being the most succesful.
- The success rate at a launch site is proportional to the flight amount.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- Launch success rates are increasing with time.
- The decision tree classifier is the best for predicting success.

Reference links

- SpaceX archives:https://web.archive.org/web/20131129020000/http://www.spacex.com/falcon9
- Wiki: https://en.wikipedia.org/wiki/Falcon_9 , https://en.wikipedia.org/wiki/Launch_vehicle
- Folium reference: https://python-visualization.github.io/folium/modules.html
- https://www.popularmechanics.com/space/rockets/a20152543/spacex-test-fire-new-falcon-9-block-5/
- https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20110015564.pdf
- https://spaceflight.nasa.gov/shuttle/reference/shutref/sts/aborts/rtls.html
- https://spacenews.com/blue-origin-reflies-new-shepard-suborbital-vehicle/
- https://web.archive.org/web/20021013181710/http://www.space.com/missionlaunches/fl_clcs_020918.html
- https://www.nasa.gov/pdf/500393main_TA01-LaunchPropulsion-DRAFT-Nov2010-A.pdf

Thank you!