IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marco Luo
1/2/25

Outline

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

Executive Summary

Summary of methodologies
- Data collection
- Data wrangling
- EDA with SQL and data visualization
- Building dashboard with Plotly Dash
- Predictive analysis with classification model

Summary of all results
- EDA results
- Interactive analytics
- Predictive analytics

Introduction

SpaceX is one of the pioneers of the commercial space industry, with innovations and advancements in space technology and affordability. Part of why SpaceX is so affordable is because they can reuse the first stage of the launch if it lands. Therefore, we'll be investigating the first stage of the Falcon 9 rocket launch to determine the cost of a launch. This information can be used to see if other companies want to bid against SpaceX rocket launches.

Our main objective is trying to predict the first launch outcome given a set of characteristics such as payload mass, orbit type, launch site, etc.

Section 1

Methodology

Methodology

Executive Summary

- Data collection using SpaceX Rest API and webscraping
- Perform data wrangling with One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

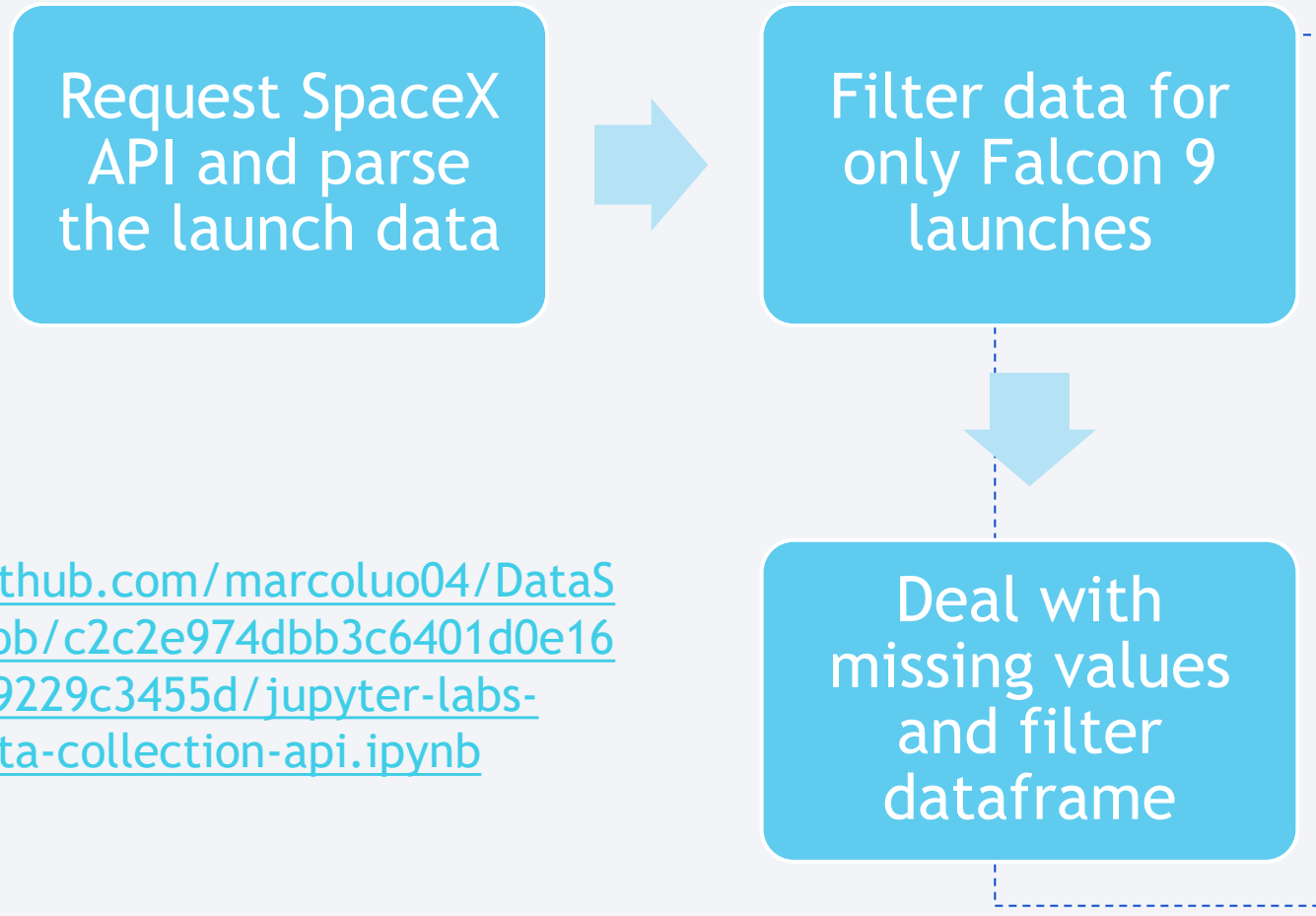
Data Collection

- Falcon 9 launch data was collected from SpaceX Rest API and webscraping

SpaceX Rest API

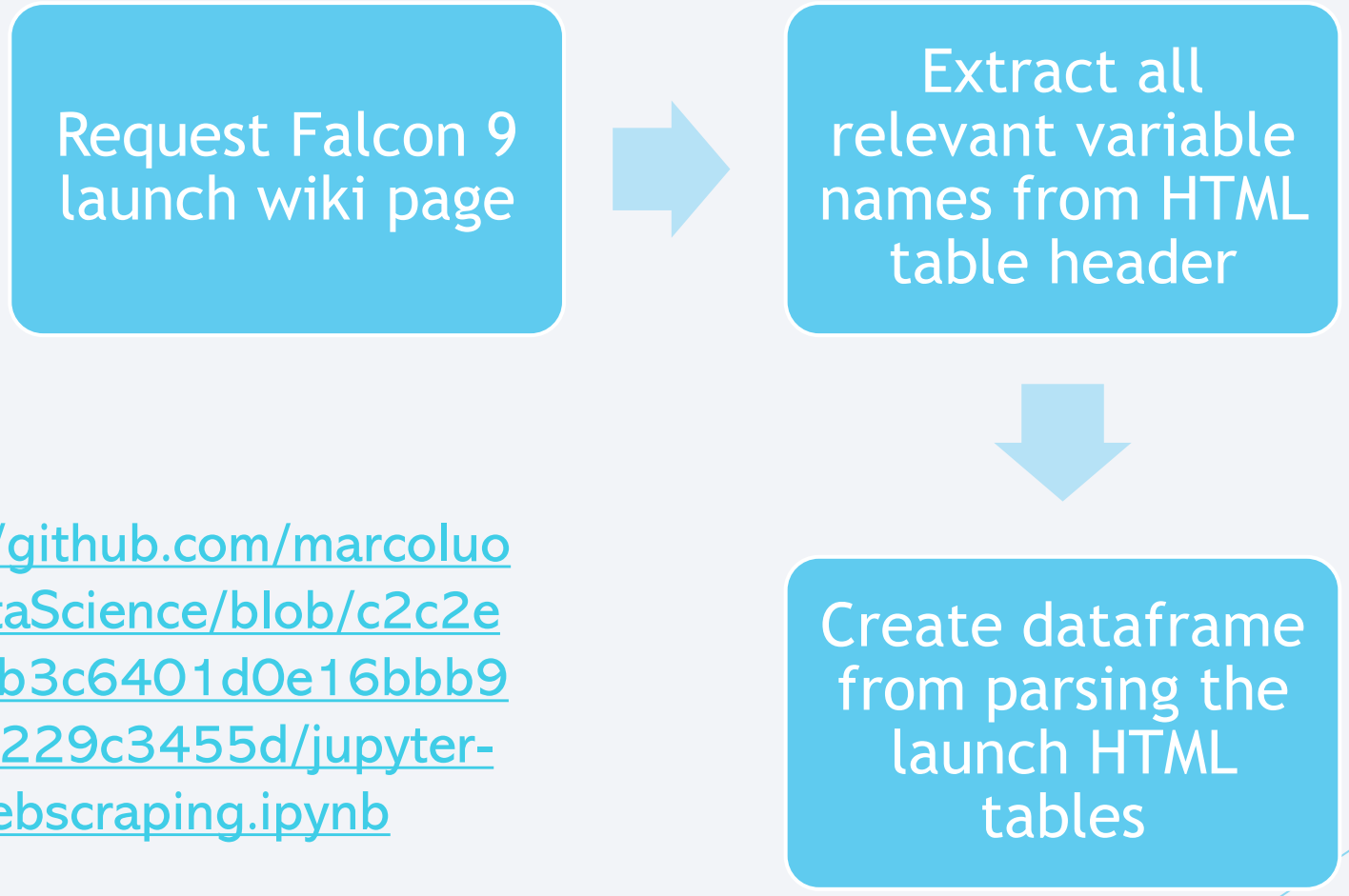
Webscraping

Data Collection – SpaceX API



- https://github.com/marcoluo04/DataScience/blob/c2c2e974dbb3c6401d0e16bbb95b399229c3455d/jupyter-labs-spacex-data-collection-api.ipynb

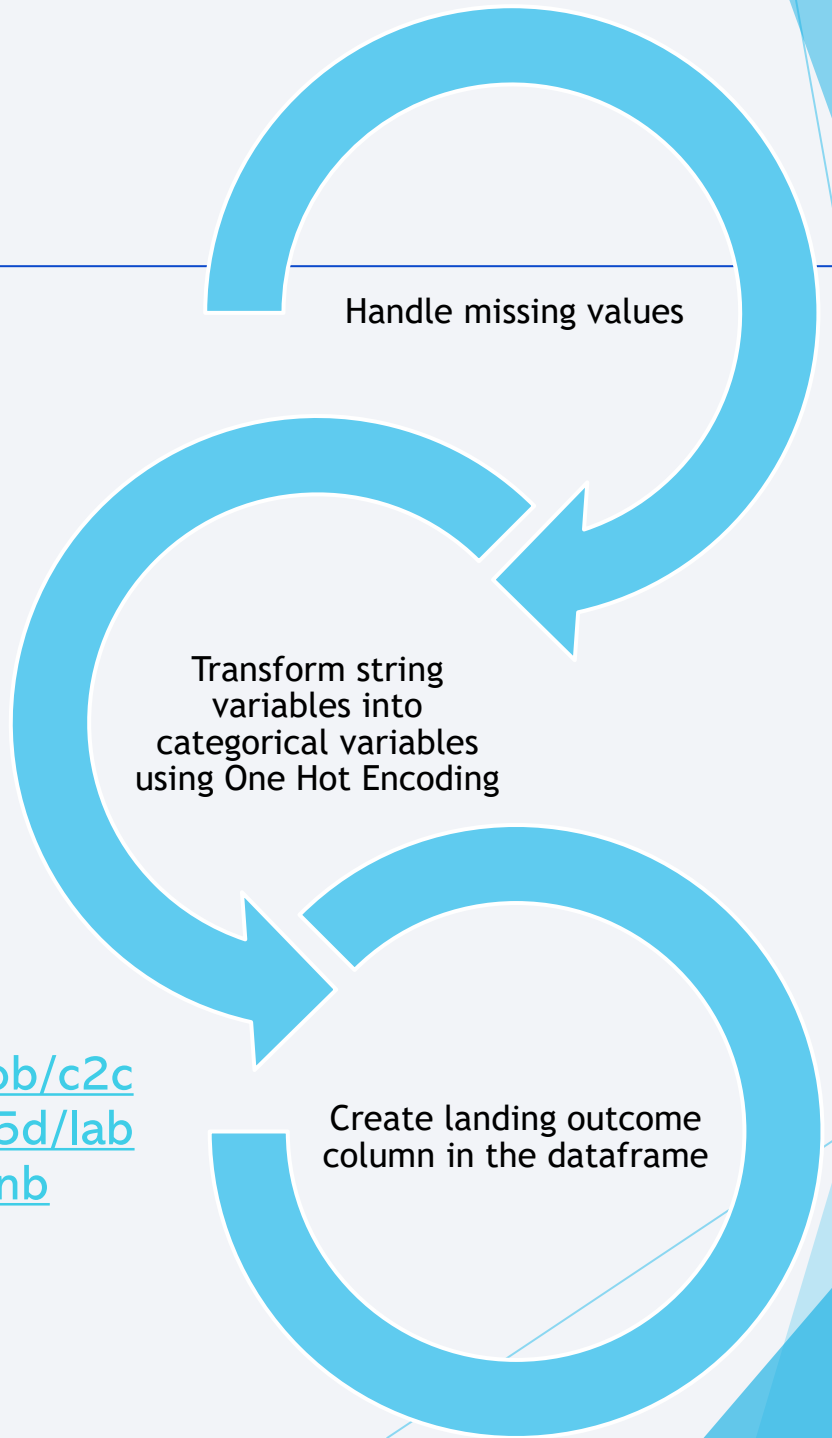
Data Collection - Scraping



- https://github.com/marcoluo04/DataScience/blob/c2c2e974dbb3c6401d0e16bbb95b399229c3455d/jupyter-labs-webscraping.ipynb

Data Wrangling

Calculate launch number for each site, number and occurrence of each orbit, number and occurrence of each mission outcome per orbit type, and then create landing outcome label



- https://github.com/marcoluo04/DataScience/blob/c2c2e974dbb3c6401d0e16bbb95b399229c3455d/labs-jupyter-spacex-Data%20wrangling%20(2).ipynb

EDA with Data Visualization

Scatter Graphs

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload vs Launch Site
- Orbit vs Flight Number
- Payload vs Orbit Type
- Orbit vs Payload Mass

Bar Graph

- Success rate vs Orbit

Line Graph

- Success rate vs Year

EDA with SQL

- Performed EDA with SQL queries to better understand our data
- Display total payload mass carried by boosters launched by NASA
- Display average payload mass carried by booster version F9 v1.1
- List dates of first successful landing outcome
- List total number of successful and failure mission outcomes
- List the names of the booster versions which carried the maximum payload mass
- List the records which will display the months, landing outcomes, booster version, and launch site for the year 2015
- Rank the count of successful landing outcomes between 6/4/2010 and 3/20/2017 in descending order
- https://github.com/marcoluo04/DataScience/blob/c2c2e974dbb3c6401d0e16bbb95b399229c3455d/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium map object is a map centered on NASA Johnson Space Center at Houston, Texas
- Red circle at NASA Johnson Space Center's coordinate with label showing its name
- Red circles at each launch site coordinates with label showing launch site name
- Grouping of points in a cluster to display different information for the same coordinates
- Markers to show successful and unsuccessful landings
- Markers to show distance between launch site to key locations with lines between them

- These objects help us better visualize and understand the context of the problem, allowing us to see all the launch sites, their surroundings and the number of successful and unsuccessful landings

Build a Dashboard with Plotly Dash

- Created a dashboard using Plotly Dash with the following components
- Dropdown which allows users to select desired launch site(s)
- Pie chart that shows the total success and total failure for the selected launch site(s)
- Rangeslider that allows a user to select a payload mass in a fixed range
- Scatter charts that show the relationship between variables

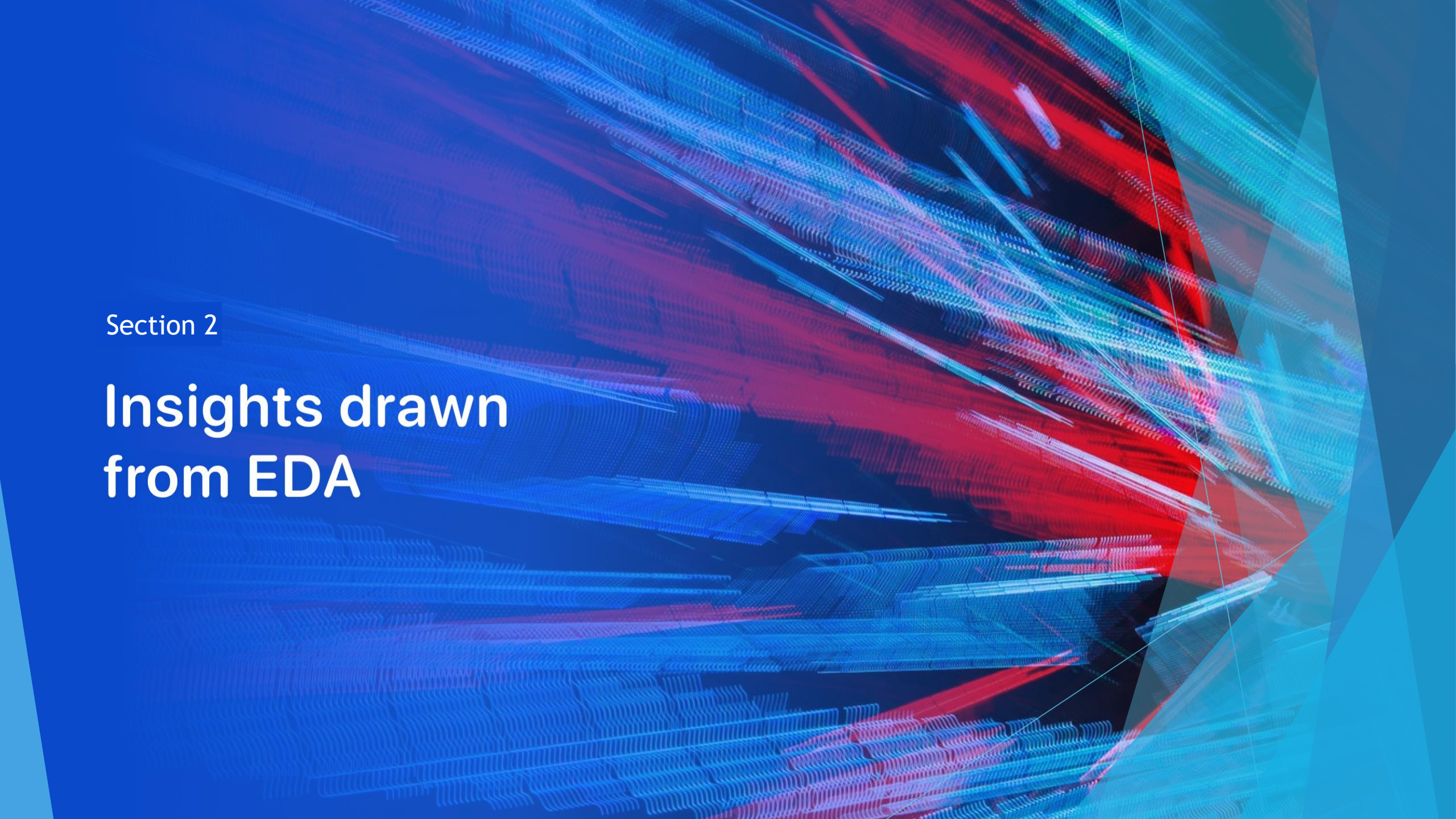
- https://github.com/marcoluo04/DataScience/blob/c2c2e974dbb3c6401d0e16bbb95b399229c3455d/spacex_dash_app.py

Predictive Analysis (Classification)

- Data Preparation
- Load dataset, normalize data, split data into training and test sets for cross validation
- Model Preparation
- Selection of machine learning algorithms, set parameters for each algorithm to GridSearchCV, training GridSearch models with training set
- Model Evaluation
- Looking for best hyper parameters for each model, compute accuracy for each model using test dataset, plot confusion matrix
- Model Comparison
- Compare models based on their accuracy, choosing the model with the best accuracy
- https://github.com/marcoluo04/DataScience/blob/c2c2e974dbb3c6401d0e16bbb95b399229c3455d/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

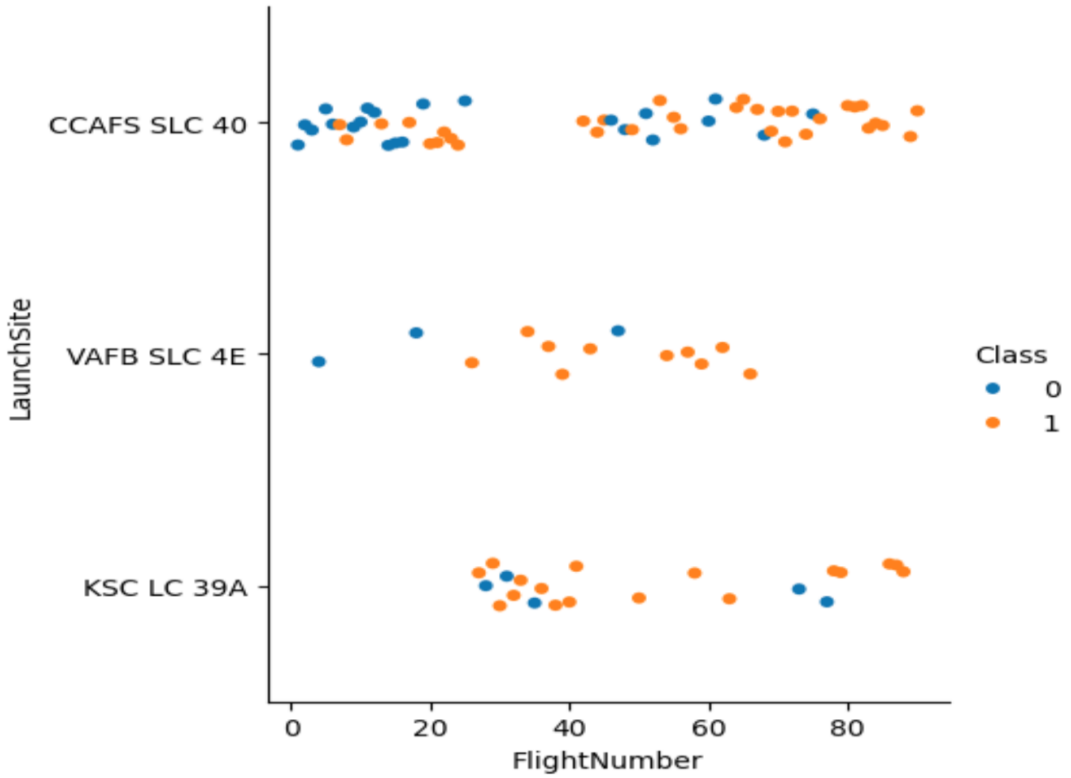
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

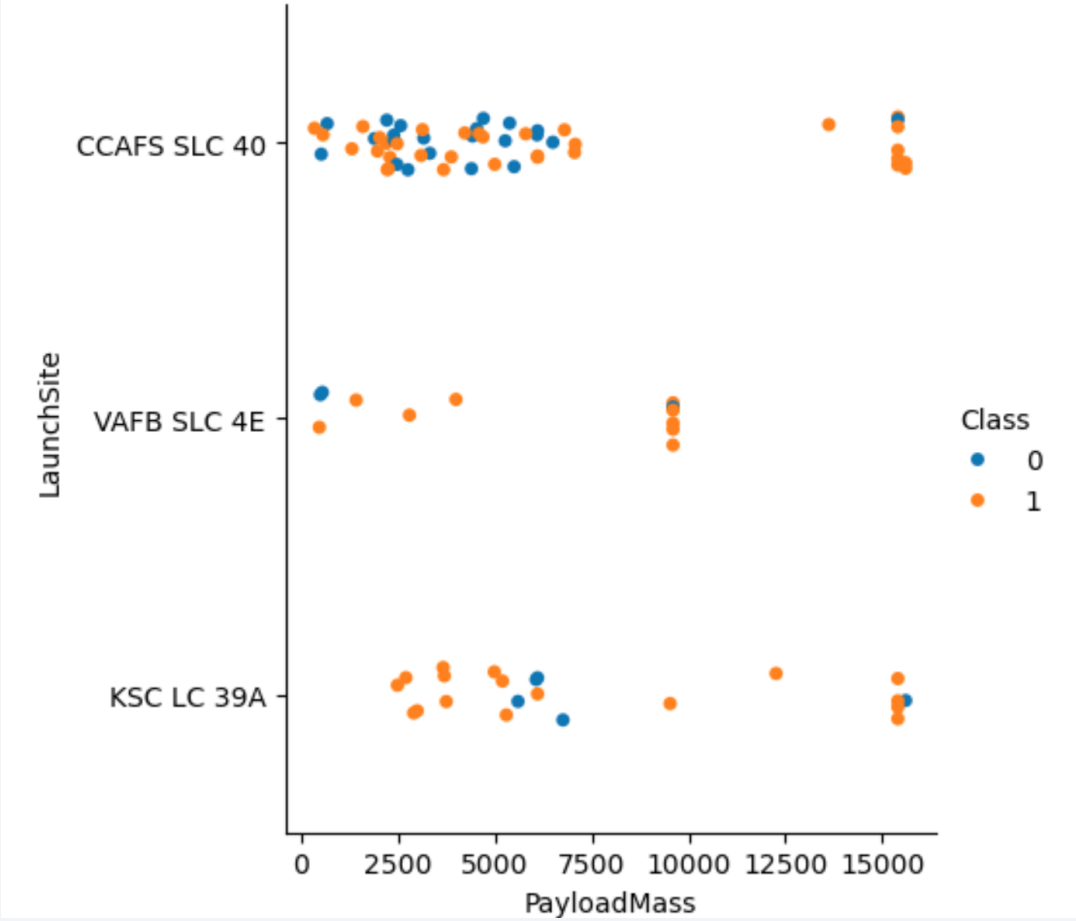
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df)
```

```
<seaborn.axisgrid.FacetGrid at 0x7c98920>
```



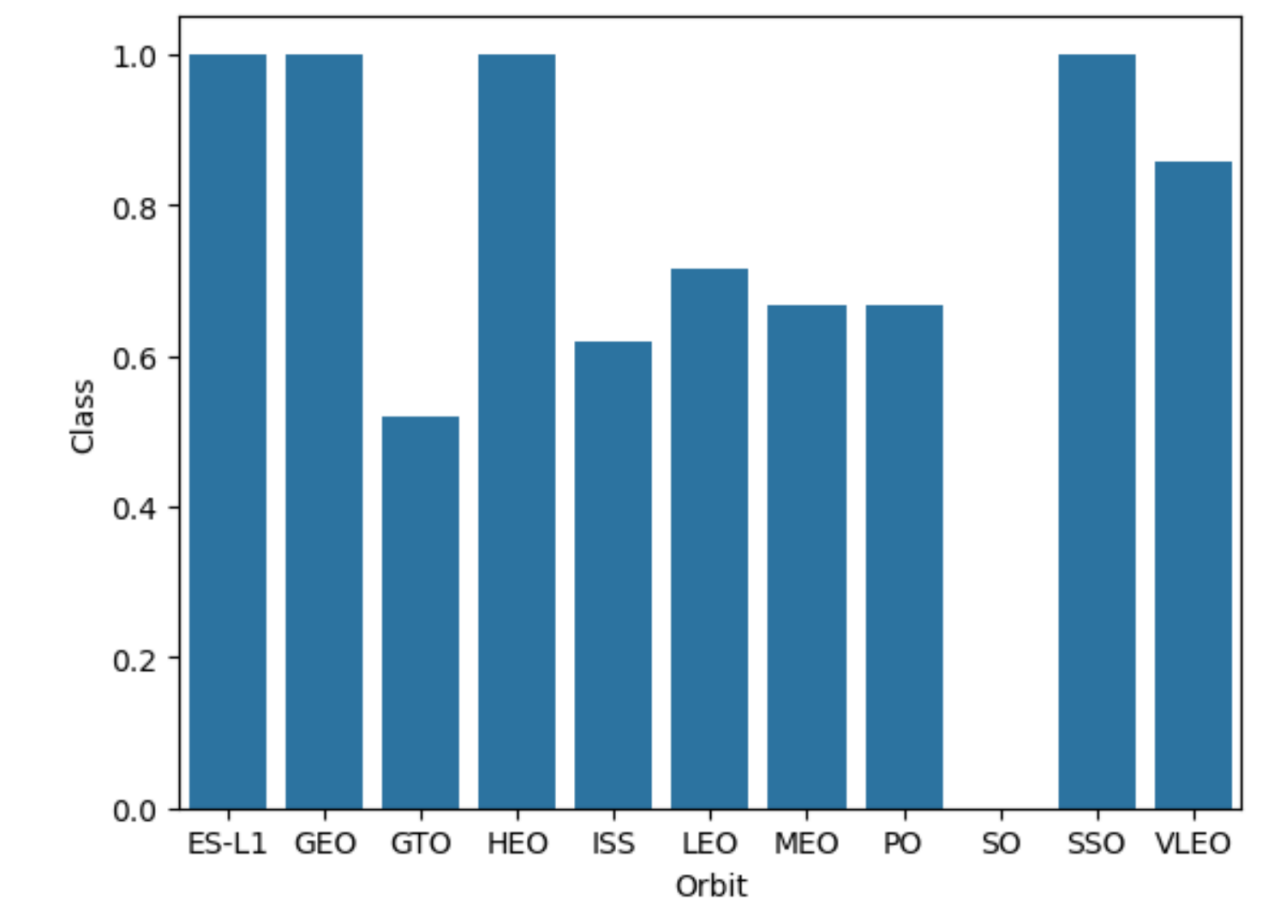
We see that the success rate for the launch sites is increasing

Payload vs. Launch Site



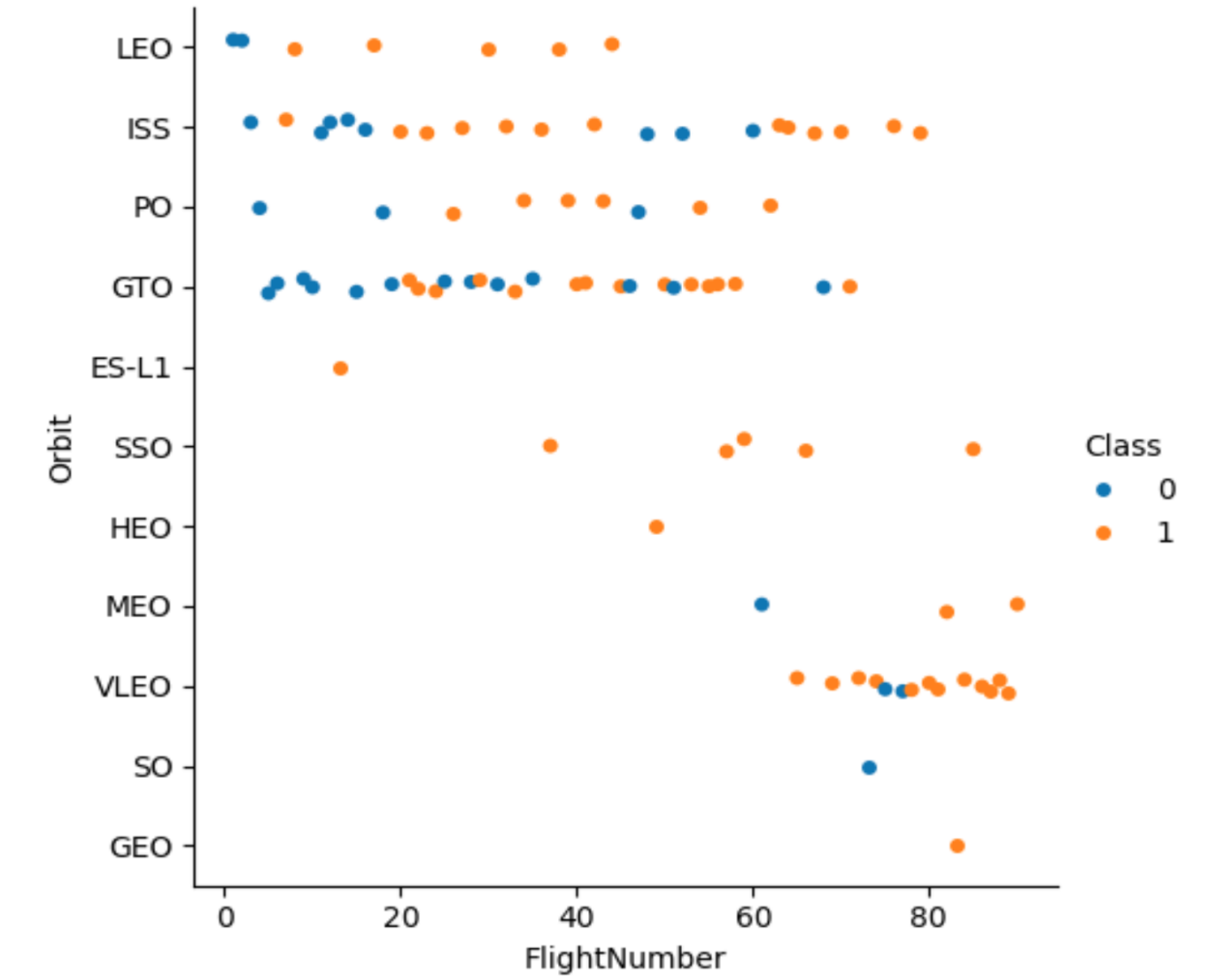
Heavier payloads may indicate a better success rate, but too heavy of payloads can lead to landing failure.

Success Rate vs. Orbit Type



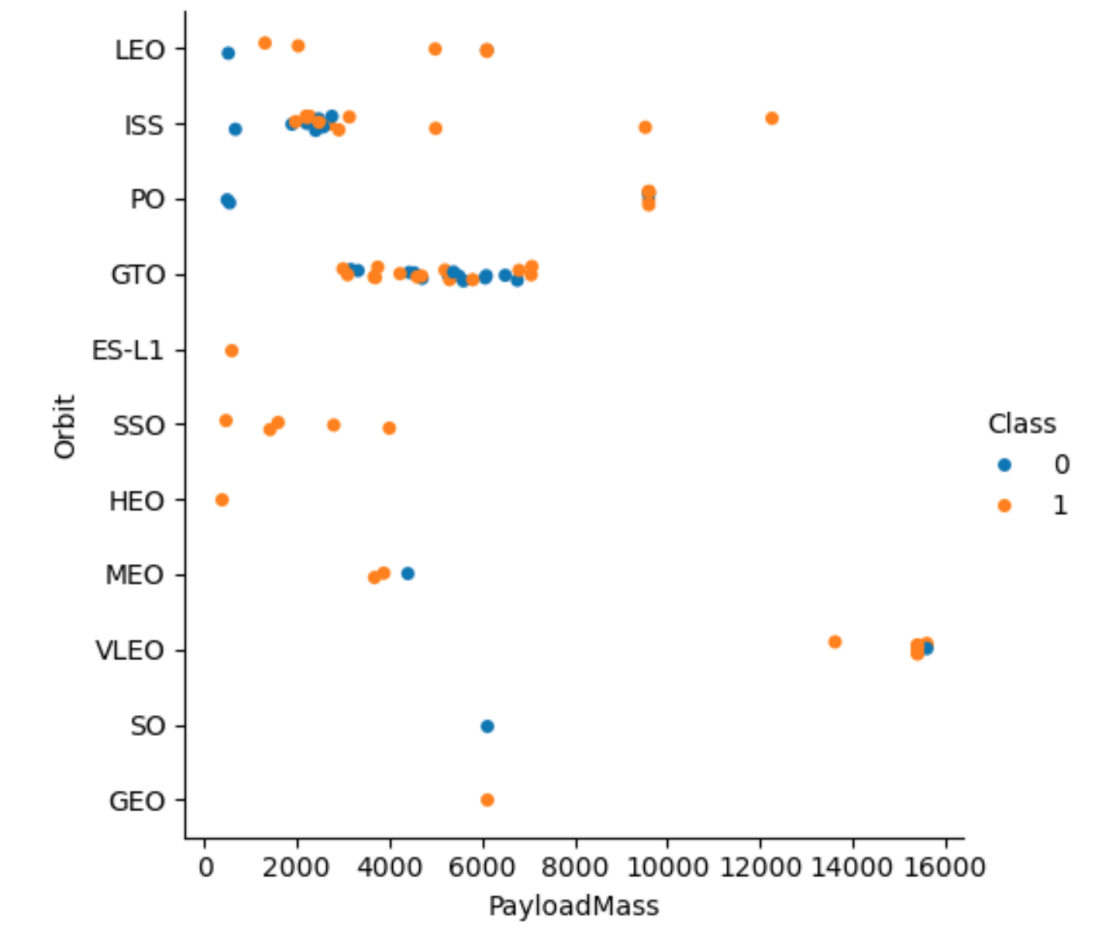
We can see that ES-L1, GEO, HEO, and SSO orbit types have the highest success rate

Flight Number vs. Orbit Type



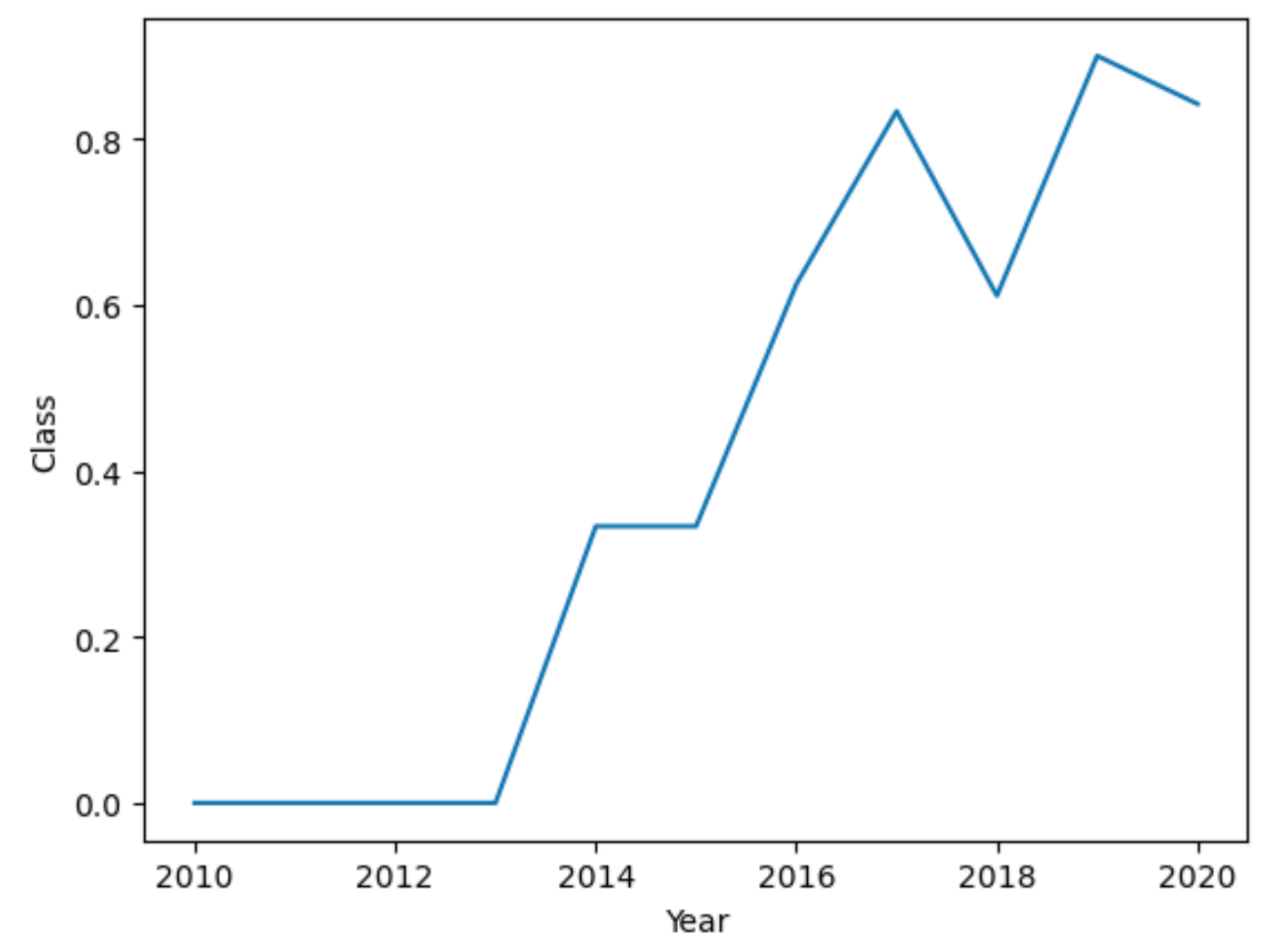
We can see the success rate for LEO orbit types increase as flight number increases. For most of the orbit types we see a similar trend.

Payload vs. Orbit Type



Payload mass can have varying effects on success rate for different orbit types. For example, heavier payload mass indicates higher success rate for LEO orbits but don't seem to have an effect for GTO orbits.

Launch Success Yearly Trend



We can see that since 2013, the success rate for launches has been increasing.

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
q = pd.read_sql('select distinct Launch_Site from SPACEXTBL', con)
q
```

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

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
q = pd.read_sql("select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5", con)
q
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcom
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...	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 attemp
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attemp
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

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

```
q = pd.read_sql("select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer='NASA (CRS)'", con)
q
```

	sum(PAYLOAD_MASS__KG_)
0	45596

The total payload mass carried by boosters launched by NASA is 45596kg

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
q = pd.read_sql("select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version='F9 v1.1'", con)
q
```

	avg(PAYLOAD_MASS__KG_)
0	2928.4

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

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
q = pd.read_sql("select min(Date) from SPACEXTBL where Landing_Outcome='Success (ground pad)'", con)
q
```

	min(Date)
0	2015-12-22

Using the min function we can see the first successful landing outcome in ground pad was 12/22/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
q = pd.read_sql("select distinct Booster_Version from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_M
q
```

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

Names of the boosters which have success outcomes and a payload mass between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
q = pd.read_sql("select substr(Mission_Outcome,1,7) as Mission_Outcome, count(*) from SPACEXTBL group by 1", con)
q
```

	Mission_Outcome	count(*)
0	Failure	1
1	Success	100

Total number of successful and failure mission outcomes: 1 failure and 100 success missions

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
q = pd.read_sql("select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) f
q
```

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

Names of the booster version that carried the maximum payload mass using a subquery

2015 Launch Records

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
q = pd.read_sql("select distinct Date, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome=
q
```

	Date	Landing_Outcome	Booster_Version	Launch_Site
0	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Dates, landing outcomes, booster versions, and launch sites for the months in 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.

```
q = pd.read_sql("select Landing_Outcome, count(*) from SPACEXTBL where Date between '2011-06-04' and '2017-03-20' group by
q
```

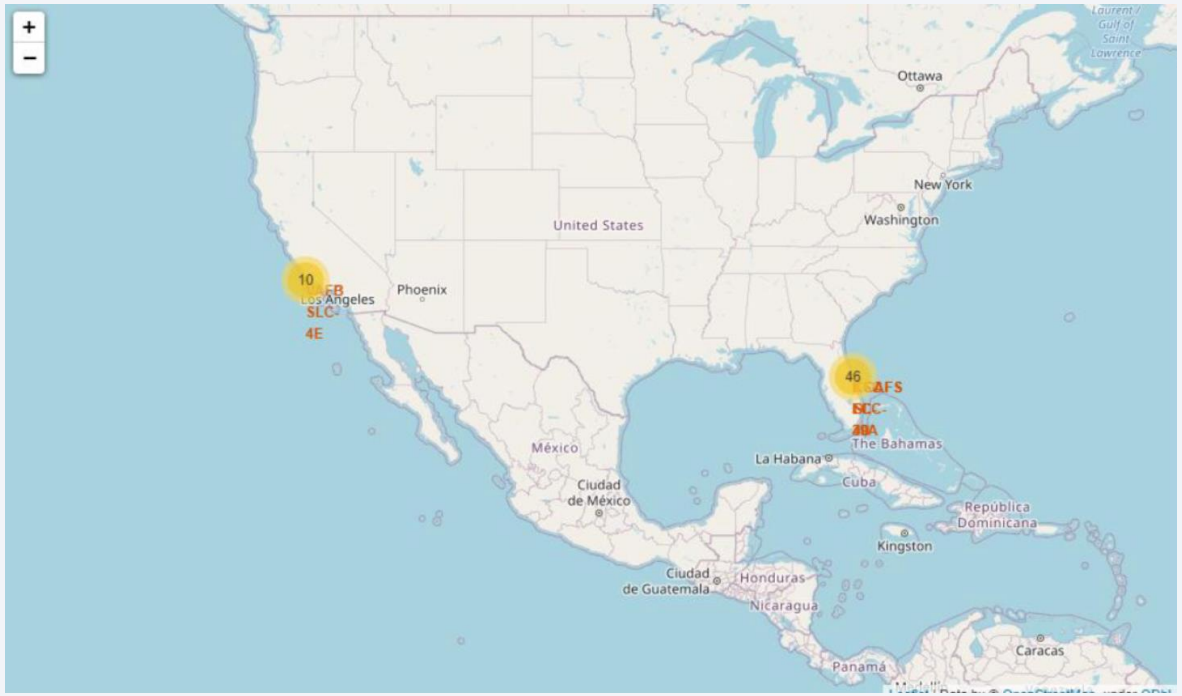
	Landing_Outcome	count(*)
0	No attempt	10
1	Success (drone ship)	5
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1

Ranking of landing outcomes frequency from 2010 to 2017

Section 3

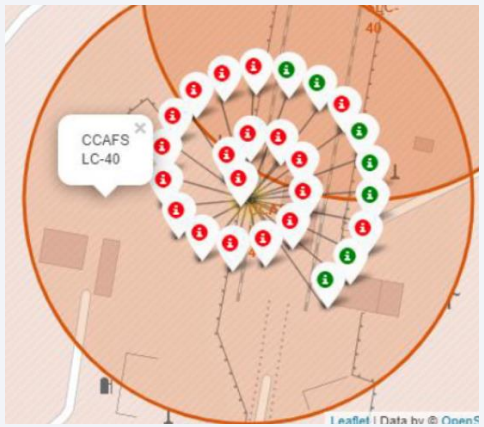
Launch Sites Proximities Analysis

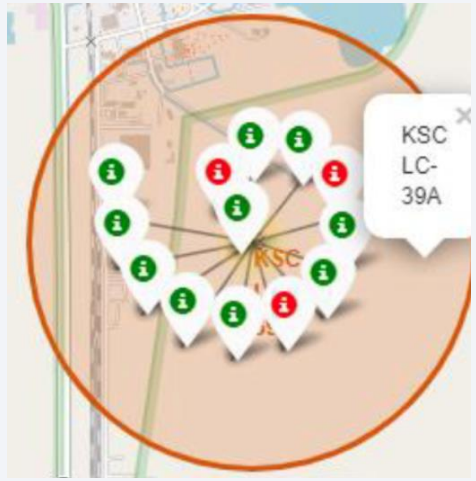
Folium map Ground Stations



SpaceX launch sites are located near the coasts of the US, specifically California and Florida

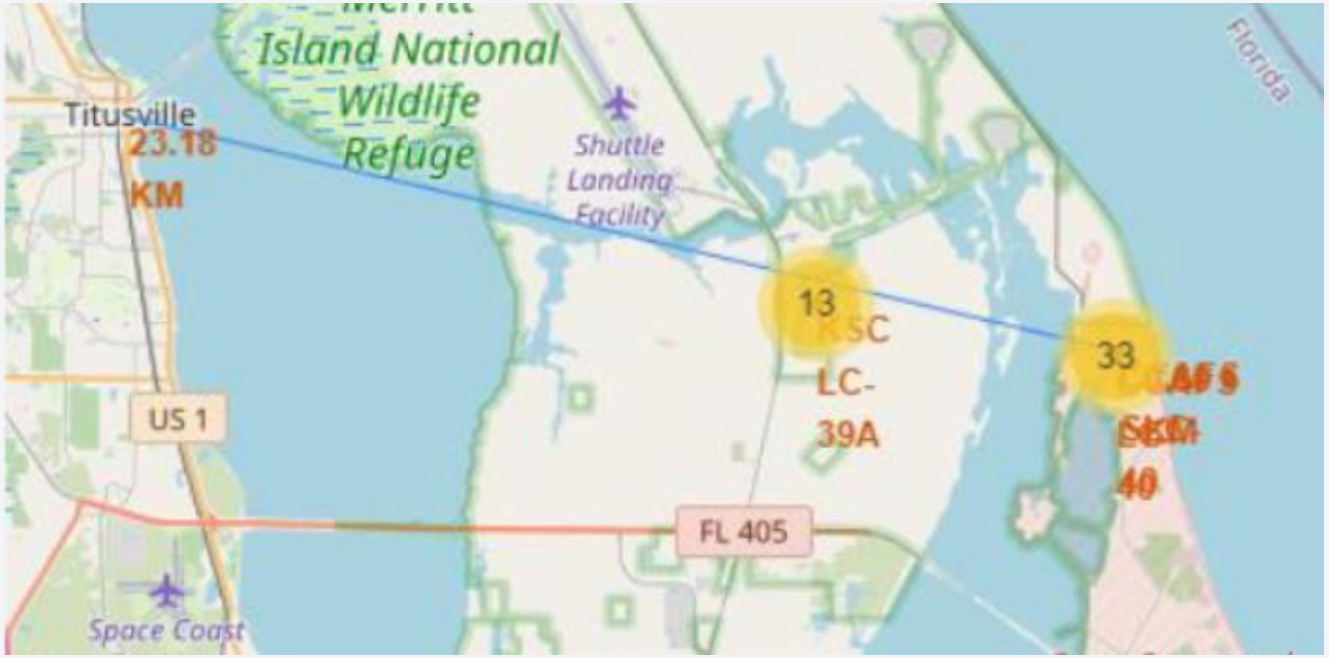
Folium map colored markers



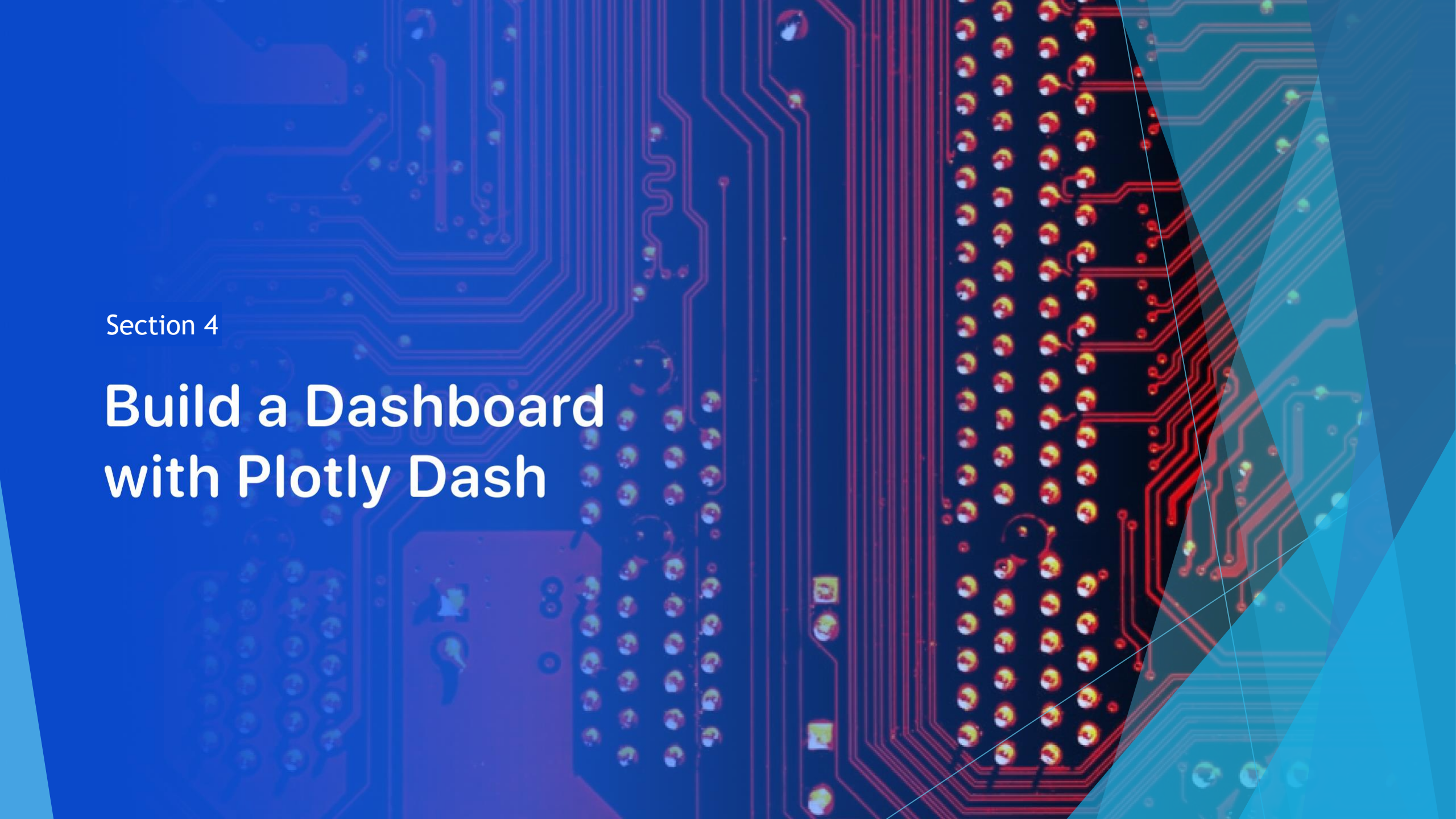


Green colored markers represents successful launches and red colored markers represents failed launches

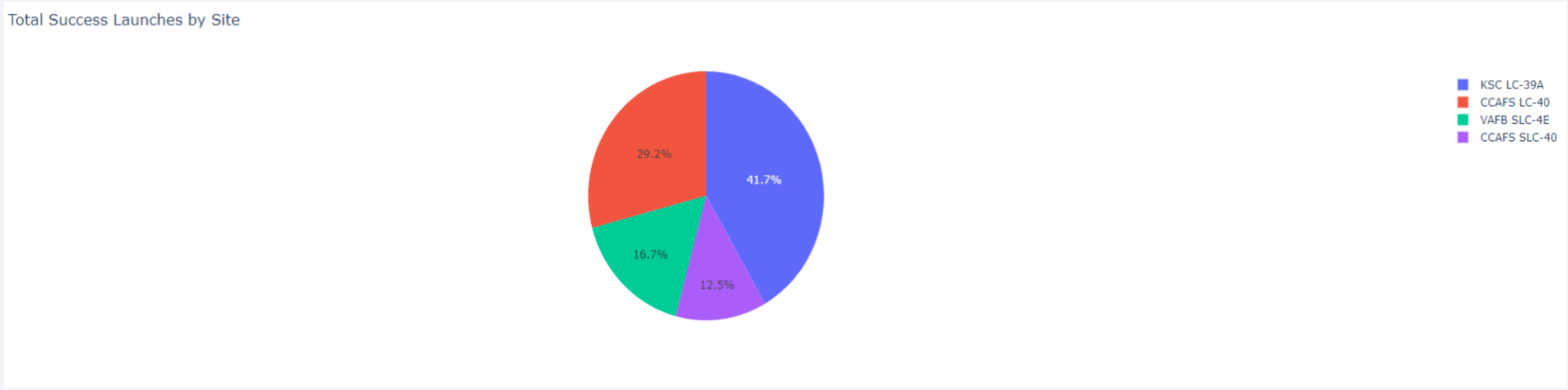
<Folium Map Screenshot 3>



CCAFS SLC-40 is close to coastline, highways, and railways

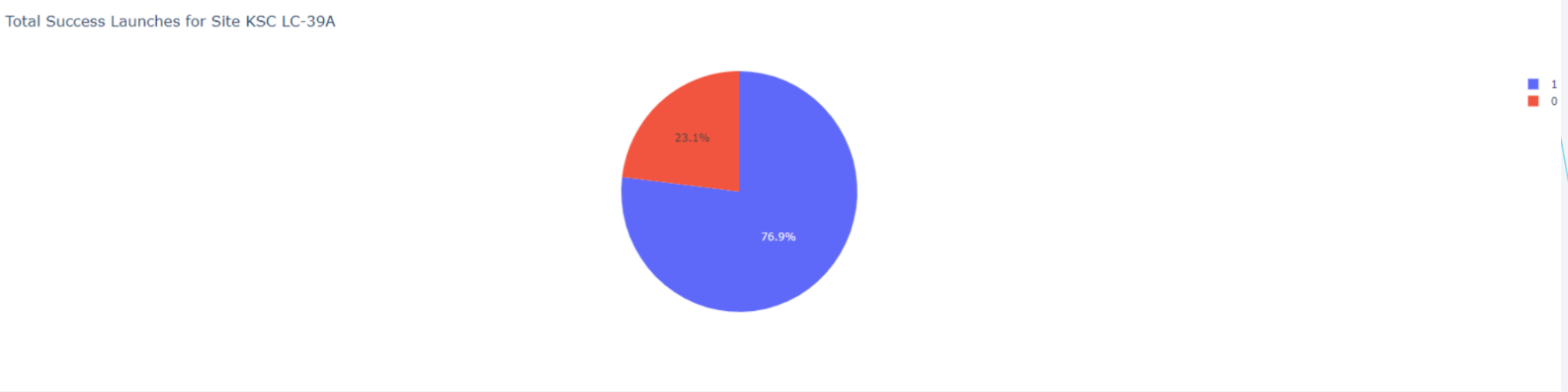

Section 4
Build a Dashboard with Plotly Dash

Plotly Dashboard success launches by site



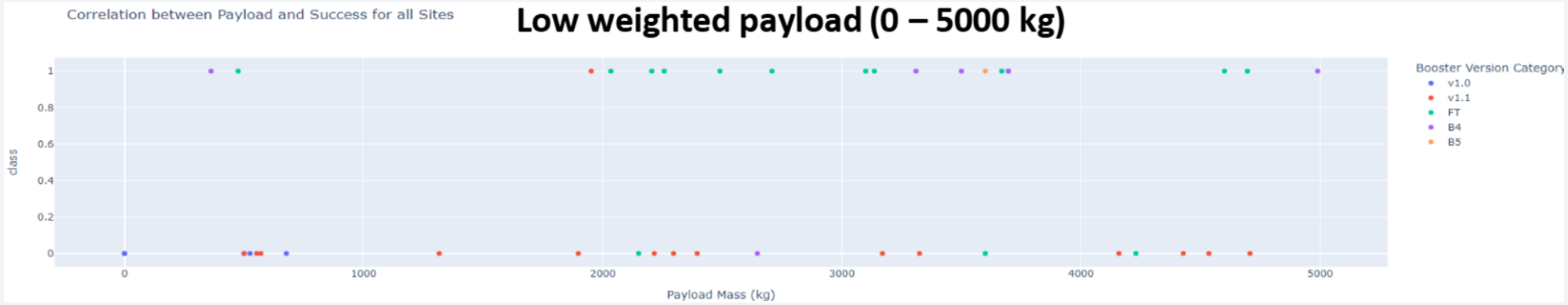
We see that KSC LC-39A has the highest success rate

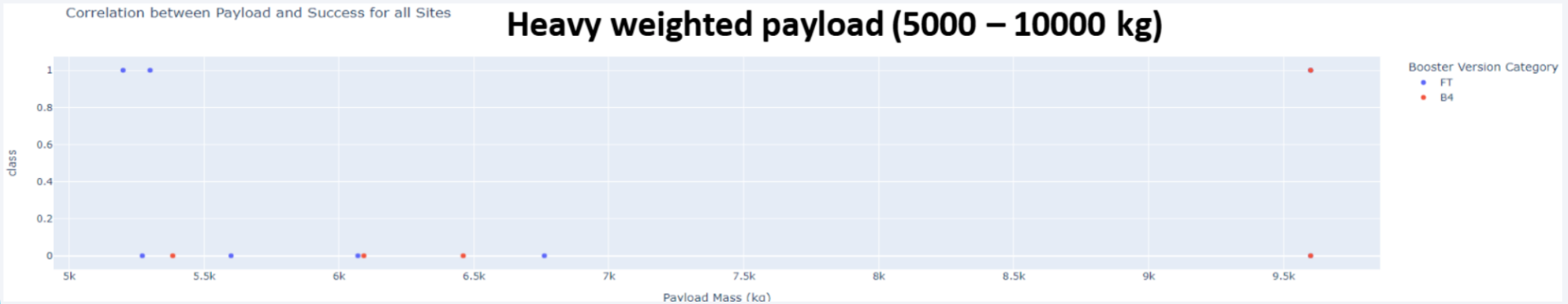
Success launches for site KSC LC-39A



76.9% of launches at the KSC LC-39A site were successes

Success rates by payload mass



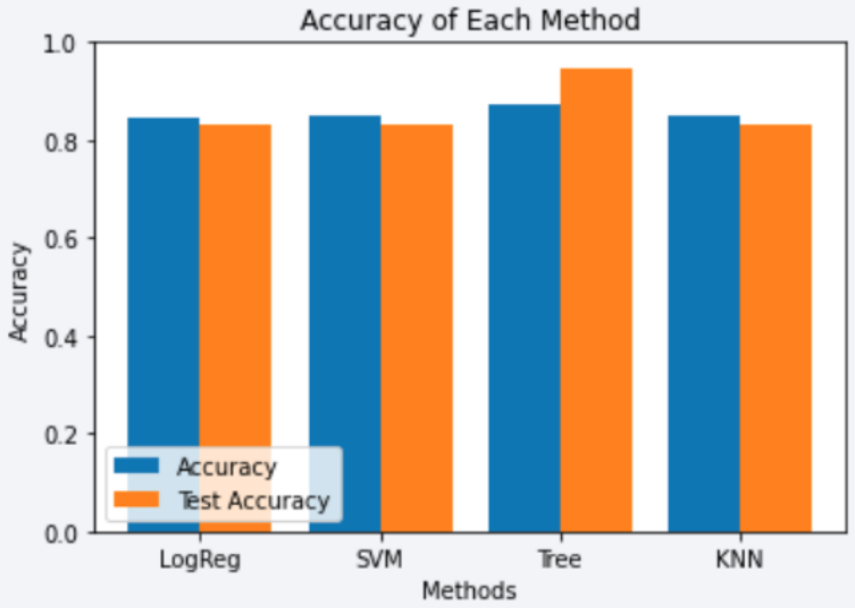


Lower weighted payloads tend to have higher success rates

Section 5

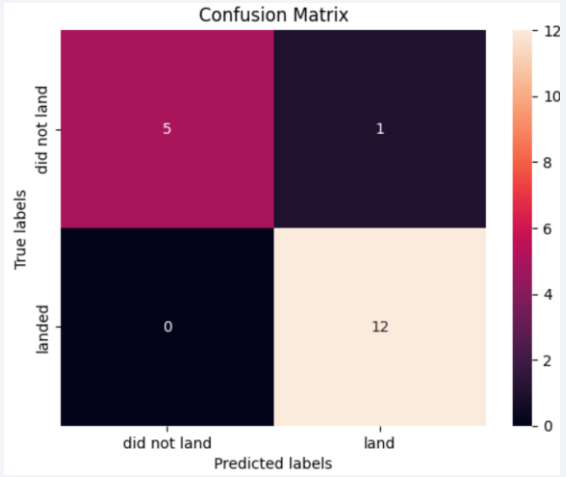
Predictive Analysis (Classification)

Classification Accuracy

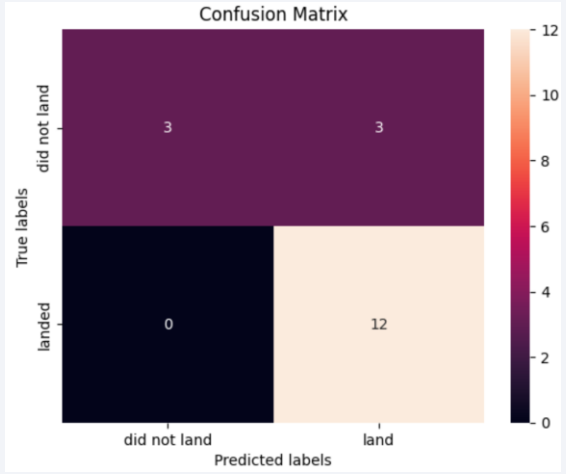


The best choice is the decision tree as it performed the best on the training data and all the methods had equal accuracy on the test data

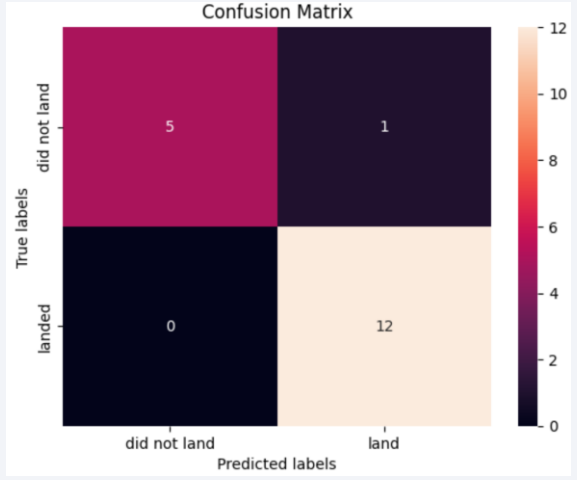
Confusion Matrix



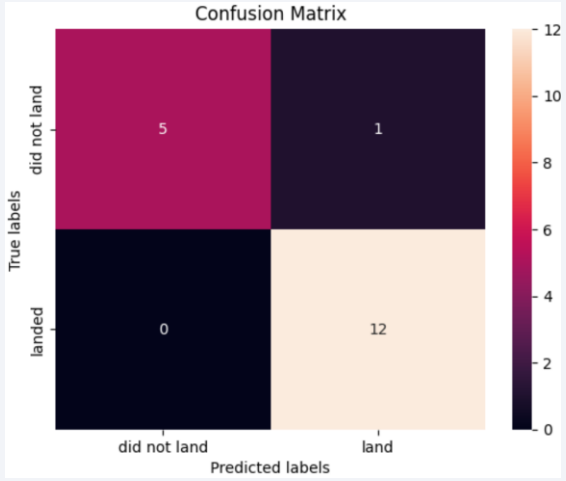
KNN



Tree



SVM



Logistic regression

Conclusions

- The success rate of launches can be determined by various factors such as orbit type, launch site, and payload mass
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1
- Payload masses with lower weights tend to have higher success rates
- The most successful launch site was KSC LC-39A
- For this dataset, we've determined that the decision tree classifier was the best performing model to predict successful launches

Thank you!