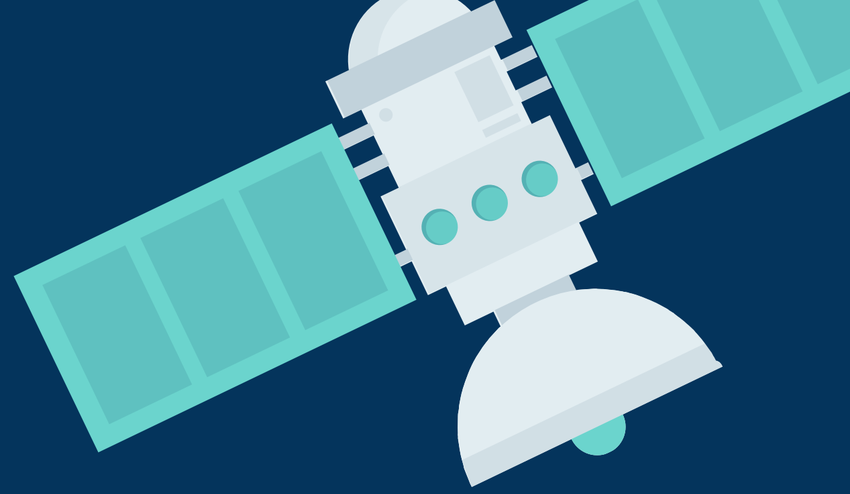
SPACE Y

WINNING SPACE RACE WITH DATA SCIENCE

Prepared by Yung-Han Kuo
06/27/2023

Outline

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



Executive Summary

Summary of Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA) with Data Visualization
- Exploratory Data Analysis (EDA) with SQL
- Interactive Map Building with Folium
- Dashboard Building with Plotly Dash
- Predictive Analysis using Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbor (KNN)

Summary of All Results

- Exploratory Data Analysis
  - launch demonstrates improvement progressively
- Visualization/InteractiveAnalysis
  - launch sites are in proximity to the Equator line and the coast
- Predictive Analysis
  - all models perform very similarly with an 0.8333 accuracy score

Introduction

Background & Context

SpaceX, an industry leader in space exploration, is dedicated to democratizing space travel by making it accessible to all. The company has achieved remarkable milestones, including successfully delivering spacecraft to the international space station, launching a constellation of satellites that enable global internet connectivity, and conducting manned missions into space. The key to SpaceX's ability to achieve this lies in its innovative approach of reusing the first stage of its Falcon 9 rocket, which significantly reduces the cost of rocket launches to a mere $62 million per launch. In contrast, other providers who do not employ this reusability feature incur expenses of over $165 million for each launch. Consequently, by assessing the probability of successfully landing and reusing the first stage, we can ascertain the cost of the launch. This determination can be made by leveraging publicly available data and employing advanced machine learning models to predict the feasibility of first stage reuse, whether it be by SpaceX or a competing company.

Introduction

Questions to Answered

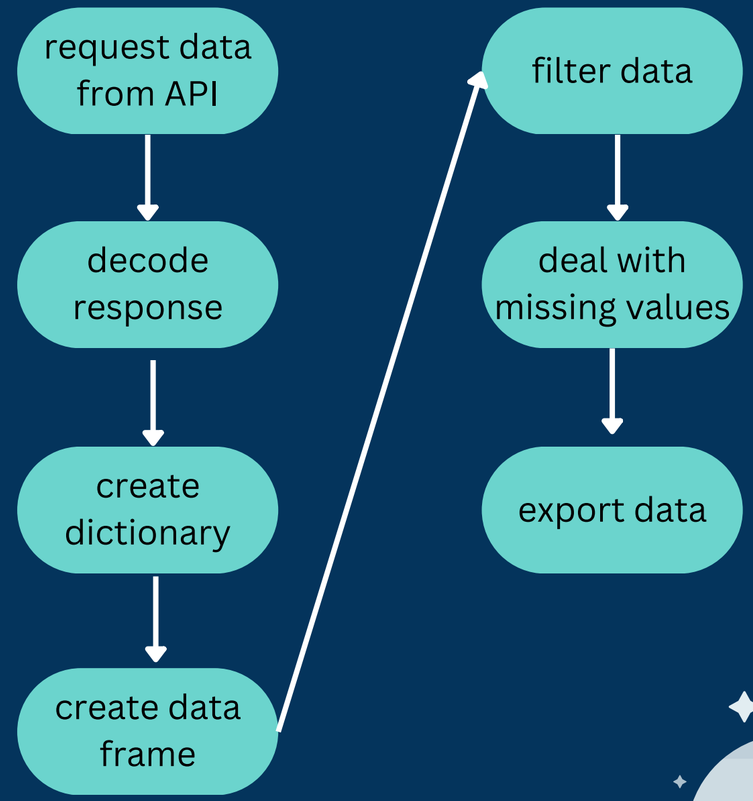
- How do features like payload mass, launch site, number of flights, and orbits influence the success of the first stage landing?
- Does the success rate of landings increase over time?
- What is the best model that can be used for binary classification to predict successful landing?

Methodology

Data Collection – SpaceX API

Steps

- request and parse the SpaceX launch data using the GET request
- decode the response with .json()
- turn the response into dataframe with .json_normalize()
- create a dictionary with the data collected
- create a data frame using the dictionary created previously
- filter the data frame to only include Falcon 9 launches
- deal with missing values in 'Payload Mass' by replacing them with its mean
- export data to a .csv file for later use

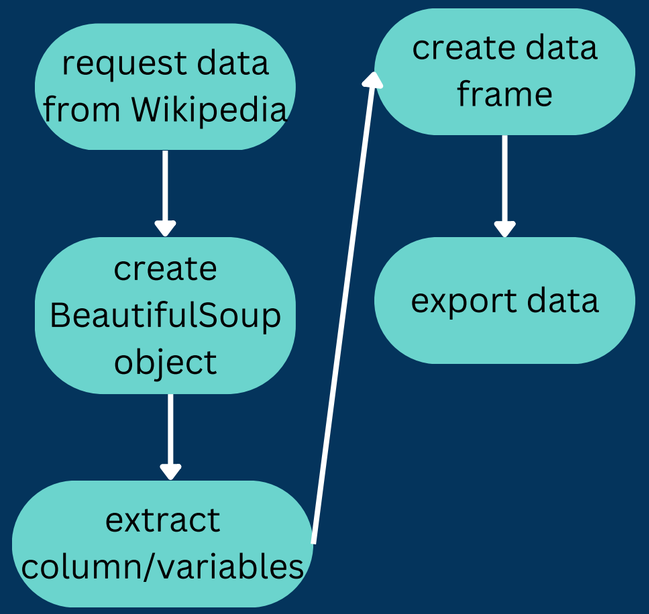


*Github link

Data Collection – Scraping

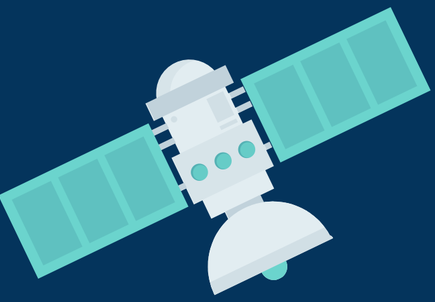
Steps

- request data of Falcon 9 launch from Wikipedia
- create a BeautifulSoup object from the HTML response
- extract all column/variable names from the HTML table header
- create a data frame by parsing the launch HTML tables
- export data to a .csv file for later use



*Github link

Data Wrangling



Goal

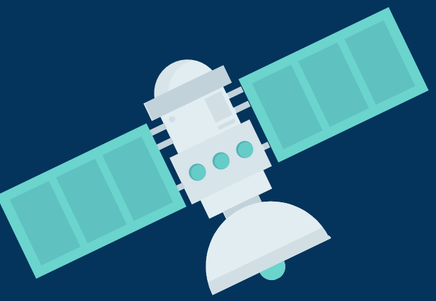
- perform EDA to find patterns in the data and determine what would be the label for training supervised models

Steps

- calculate
  - the number of launches on each site
  - the number and occurrence of each orbit
  - the number & occurence of mission outcome per orbit type
- create a landing outcome label from 'Outcome' column
- export data to a .csv file for later use

*Github link

Data Wrangling (cont.)



'Outcome'

- True Ocean: means the mission outcome was successfully landed to a specific region of the ocean
- 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

*Github link

EDA with Data Visualization

Required Libaray

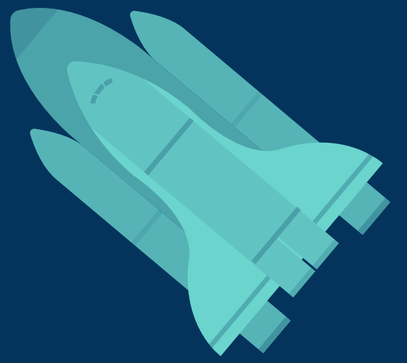
- use Matplotlib

Charts Plotted

*Github link

- Scatterplot - demonstrates the relationship between variables
  - Flight Number vs. Payload
  - Flight Number vs. LaunchSite
  - Flight Number vs. Orbit Type
  - Payload Mass (kg) vs. Launch Site
  - Payload Mass (kg) vs. Orbit Type
- Bar Chart - demonstrates comparisons among discrete categories
  - Orbit Type vs. Success Rate
- Line Chart - demonstrates trends in data over time
  - Year vs. Success Rate

EDA with SQL



Execute SQL Queries for Following Tasks

*Github link

- display the names of the unique launch sites in the space mission
- display 5 records where launch sites begin with the string 'CCA'
- display the total payload mass carried by boosters launched by NASA (CRS)
- display average payload mass carried by booster version F9 v1.1
- list the date when the first succesful landing outcome in ground pad was acheived
- list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- list the total number of successful and failure mission outcomes
- list the names of the booster_versions which have carried the maximum payload mass
- 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
- 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.

Build an Interactive Map with Folium

Mark All Launch Sites

- create circles at NASA Johnson Space Center's coordinate with a popup label showing its name
- create circles at all launch sites coordinates with a popup label showing its name using its name

Mark Success/Failed Launches for Each Site

- add green colored markers for successful at each site
- add red colored markers for unsuccessful at each site

Calculate Distances Between A Launch Site To Its Proximities

- create colored lines to show distance between launch site its proximity to the nearest coastline, railway, highway, and city

*Github link

Build a Dashboard with Plotly Dash

Add a Launch Site Drop-down Input Component

- enable selection for each launch site or all launch sites

Generate Pie Chart and Scatterplot

- pie chart - provide visualization for successful launch vs. failed launch
- scatterplot - provide visualization for relationship between payload mass and launch outcome

Add a Range Slider to Select Payload

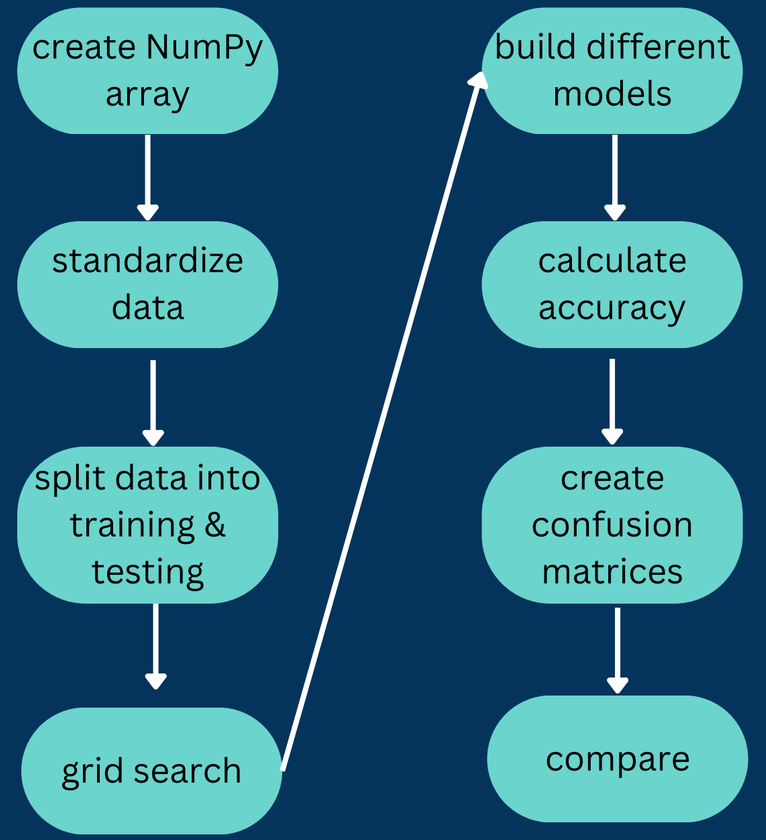
- enable range selection for payload mass

*Github link

Predictive Analysis (Classification)

Steps

- create a NumPy array from the column Class in data
- standardize the data
- split the data X and Y into training and test data
- create a GridSearchCV object with cv = 10
- apply GridSearchCV to Logistic Regression, SVM, Decision Tree, and KNN models
- calculate the accuracy for all models
- create confusion matrix for all models
- compare and contrast the models

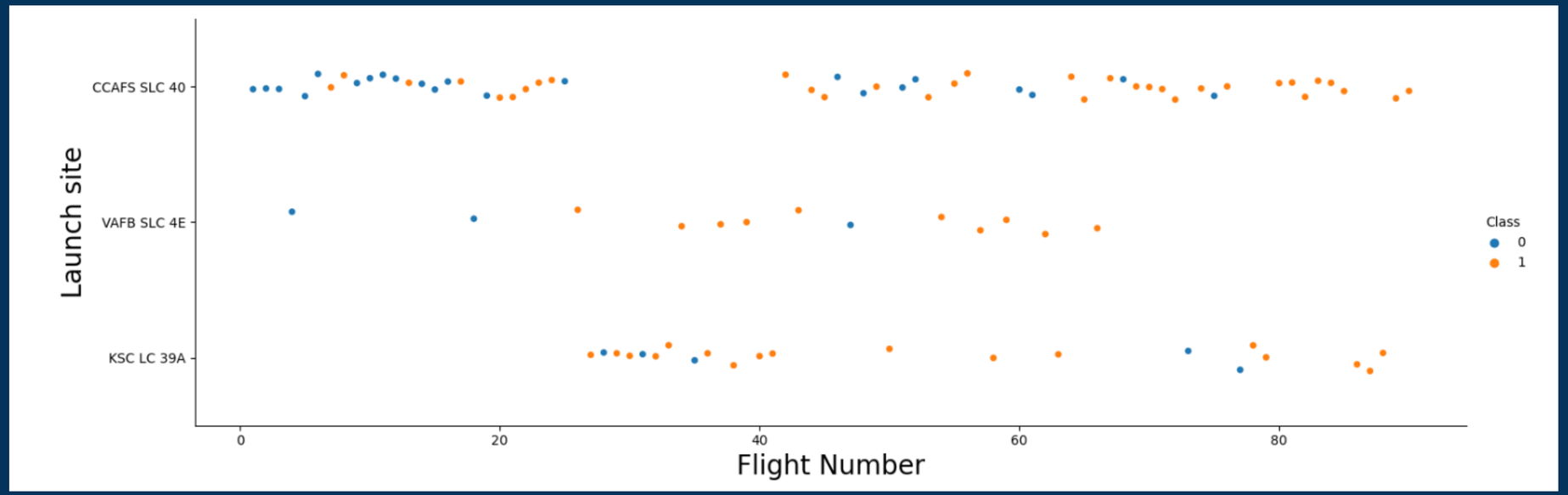


*Github link

Insights Drawn from EDA

EDA with Data Visualization

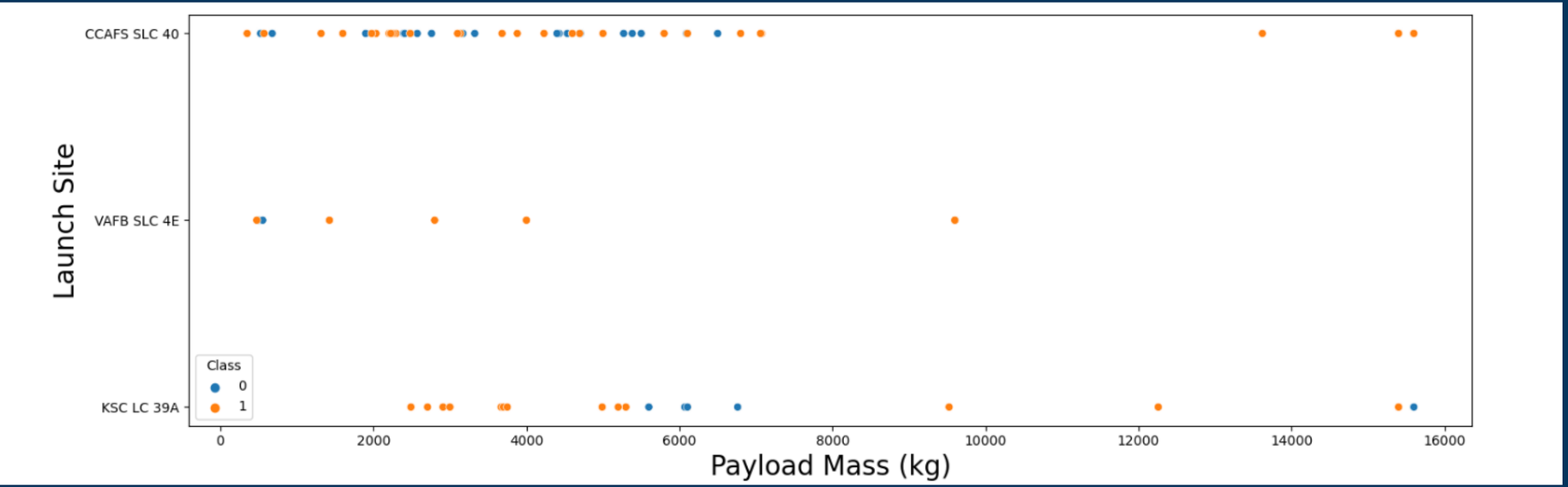
Flight Number vs. Launch Site



Explanation

- blue represents failed launch while orange represents successful launch
- earlier flights mostly failed and the latter flights mostly succeeded
- most launches are from CCAFS SLC 40 site
- VAFB SLC 4E and KSC LC 39A sites have higher success rates

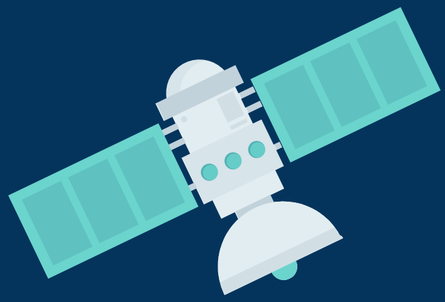
Payload vs. Launch Site

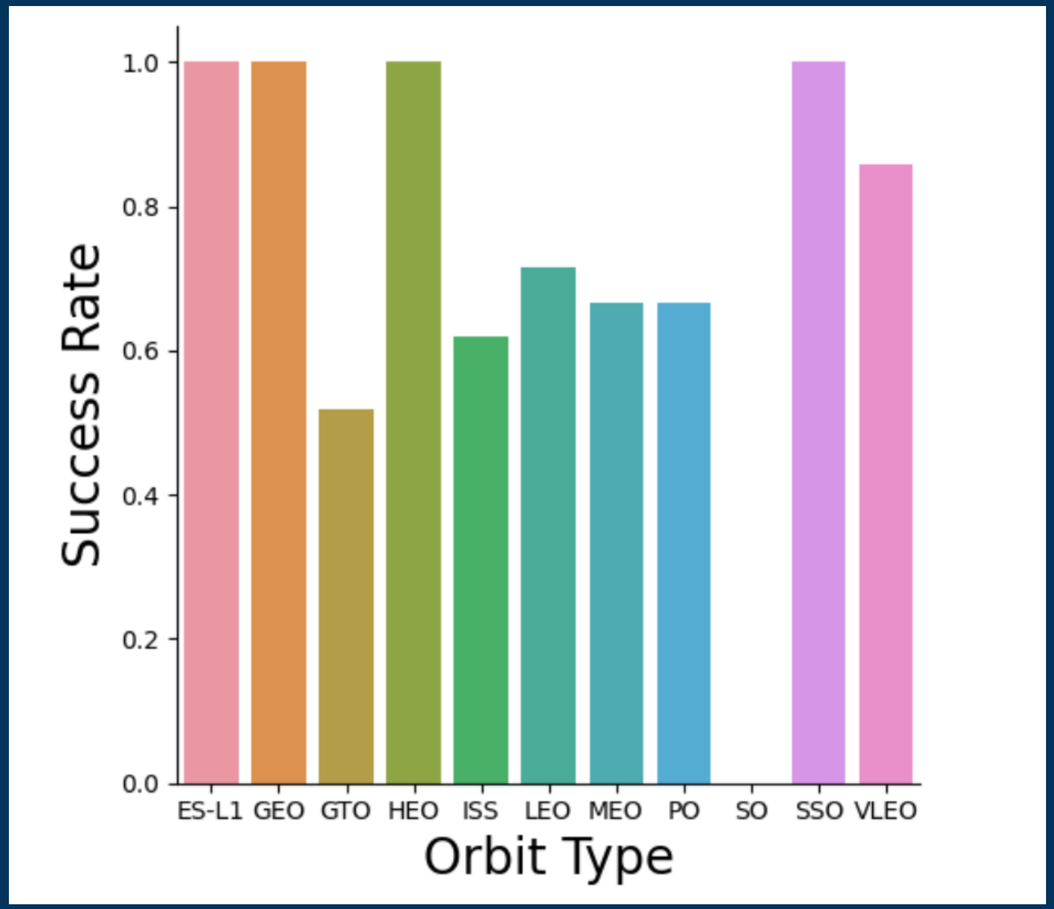


Explanation

- generally, when the payload mass is higher, the success rate is also higher
- for the VAFB-SLC site, there are no rockets launched for heavy payload mass (greater than 10000 kg).
- most launches with a payload greater than 7000 kg were successful
- KSC LC 39A site has a 100% success rate for launches with payload less than about 5500 kg

Success Rate vs. Orbit Type

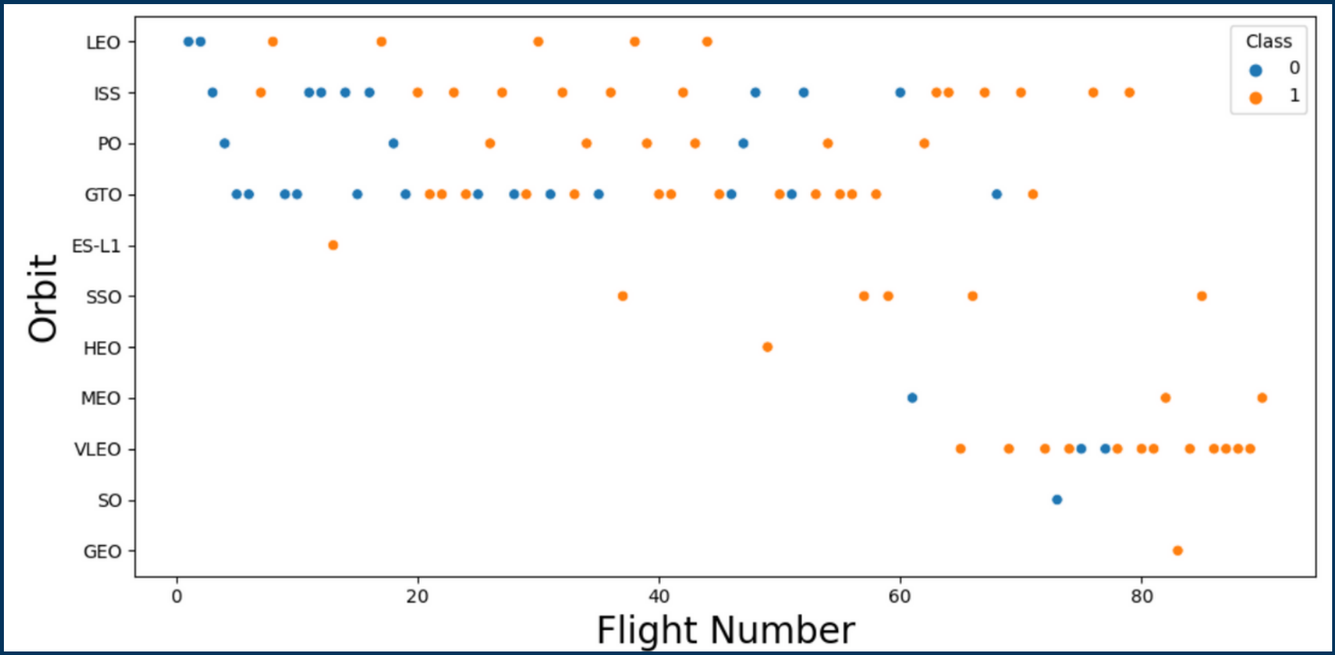




Explanation

- 100% SuccessRate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- about 90% Success Rate: VLEO
- between 50-75% Success Rate:
  - GTO
  - ISS
  - LEO
  - MEO
  - PO
- 0% Success Rate: SO

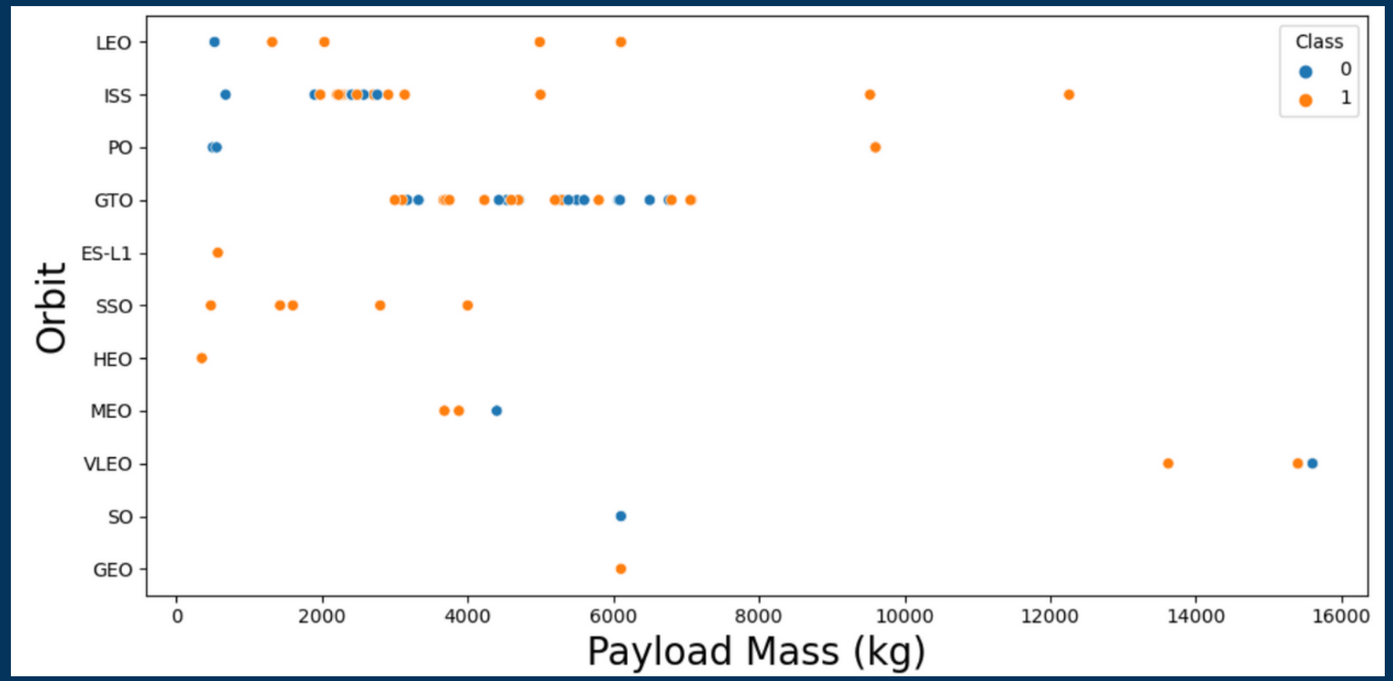
Flight Number vs. Orbit Type



Explanation

- as the number of flights increases, the success rate generally increases as well
- LEO orbit the Success appears related to the number of flights
- no relationship between flight number when in GTO orbit

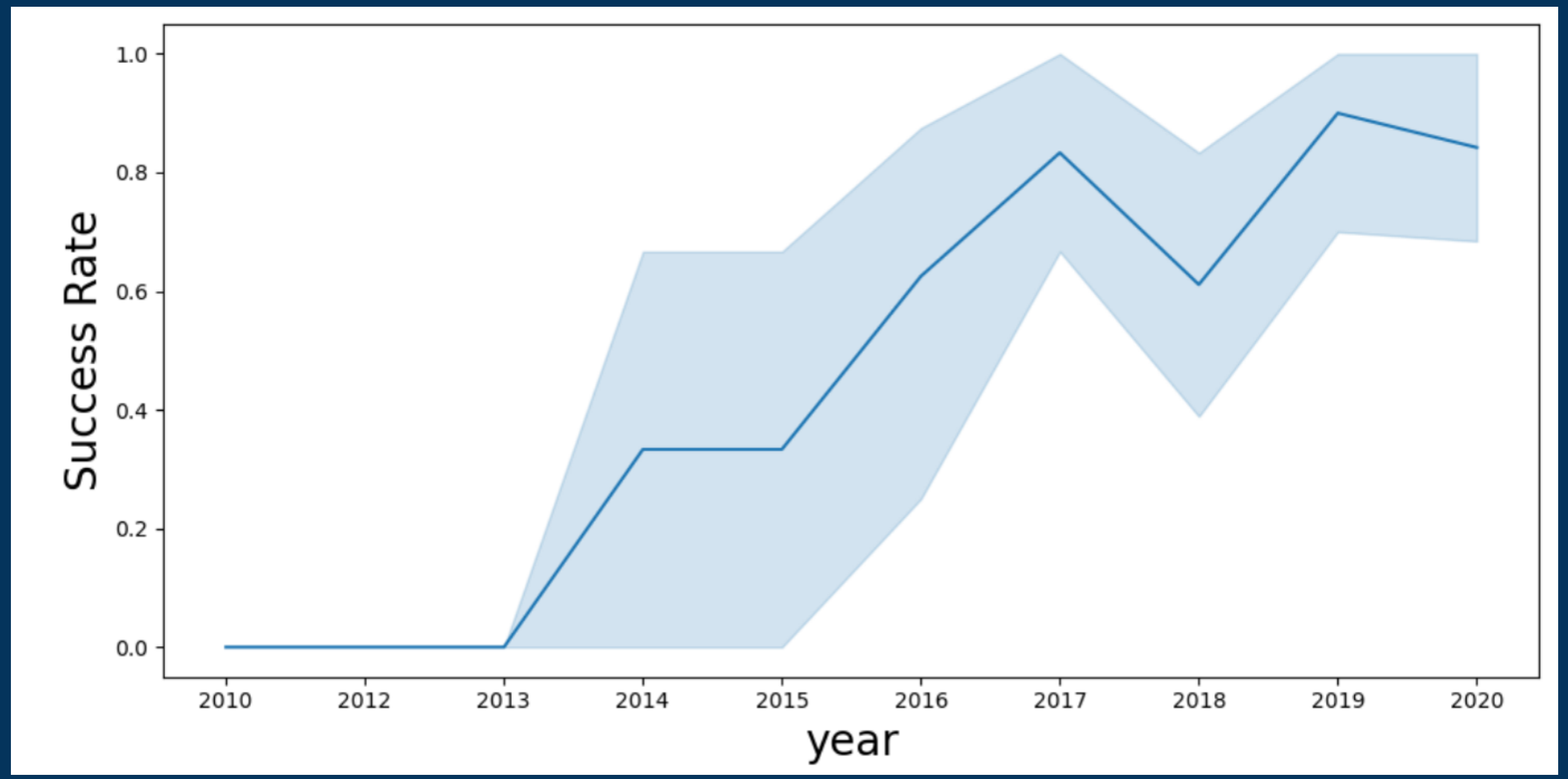
Payload vs. Orbit Type



Explanation

- with heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS
- for GTO, there are both positive landing rate and negative landing rate

Launch Success Yearly Trend

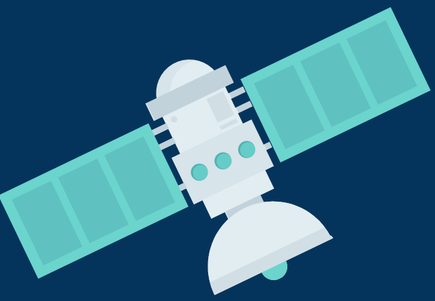


Explanation

- the success rate since 2013 kept increasing till 2020

EDA with SQL

All Launch Site Names



```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
```

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

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

Using Distinct() To Show Unique Sites

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

Launch Site Names Begin with 'CCA'

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fail
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fail
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

Using Like() To Show Site Name Starts With 'CCA'

Total Payload Mass

```
%sql SELECT sum(PAYLOAD_MASS__KG_) AS sum_payloadmass FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

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

sum_payloadmass
45596.0

Using Sum() and Where Clause

- 45596 kg

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(PAYLOAD_MASS__KG_) AS avg_payloadmass FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

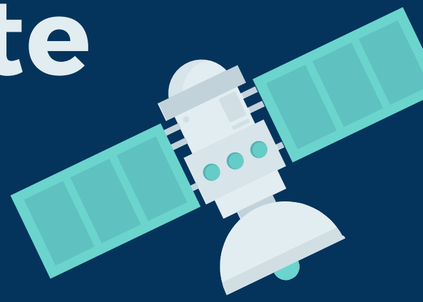
```
 * sqlite:///my_data1.db
Done.
```

avg_payloadmass
2928.4

Using Avg() and Where Clause

- 2928.4 kg

First Successful Ground Landing Date



```
%sql SELECT min(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

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

min(DATE)
01/08/2018

Using Min() and Where Clause

- 01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE Landing_Outcome ='Success (drone ship)' AND PAYLOAD_MASS__KG_ BE
```

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

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

Using Where Clause and Between

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT (MISSION_OUTCOME) AS total_count FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

Mission_Outcome	total_count
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

```
%sql SELECT MISSION_OUTCOME, COUNT (MISSION_OUTCOME) AS total_count FROM SPACEXTBL WHERE (MISSION_OUTCOME LIKE 'S
```

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

Mission_Outcome	total_count
Success	100

```
%sql SELECT MISSION_OUTCOME, COUNT (MISSION_OUTCOME) AS total_count FROM SPACEXTBL WHERE (MISSION_OUTCOME LIKE 'F
```

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

Mission_Outcome	total_count
Failure (in flight)	1

Using Count() and Group By()

- Success: 100
- Failure: 1

* it somehow showed up weird so I tried using Like() to show better results

Boosters Carried Maximum Payload

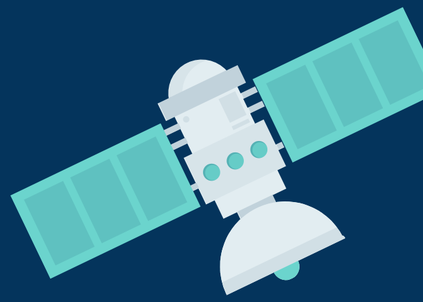
```
%sql SELECT BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG
```

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

boosterversion
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

Using Where Clause and Between

2015 Launch Records



```
%%sql SELECT substr(Date,4,2) AS month, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND substr(Date,7,4)='2015';
```

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

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

Using Where Clause and And

- 10: October
- 04: April

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

```
%%sql SELECT LANDING_OUTCOME, COUNT(*) AS count_outcomes FROM SPACEXTBL
WHERE DATE between '04-06-2010'AND '20-03-2017' GROUP BY LANDING_OUTCOME ORDER BY count_outcomes DESC
```

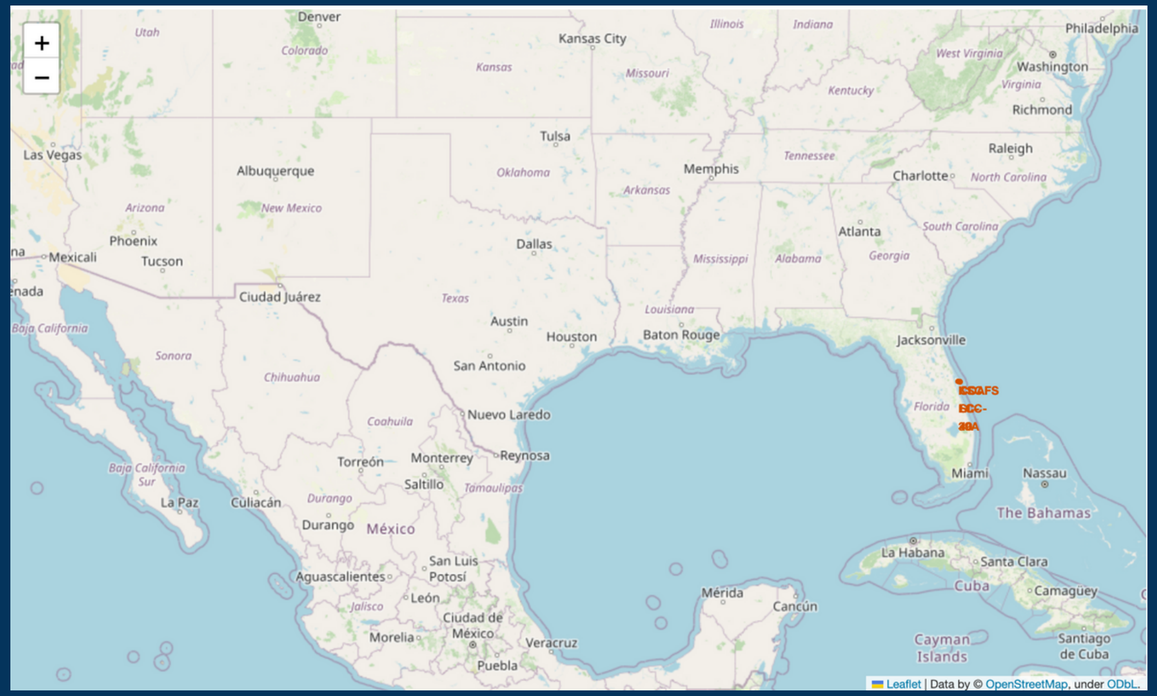
```
 * sqlite:///my_data1.db
Done.
```

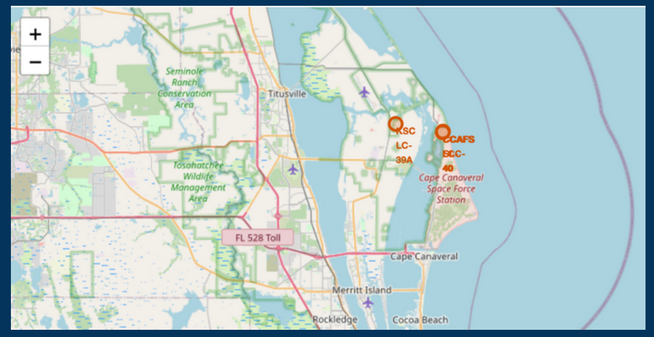
Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

- Using Where Clause, Count, Between, And, Group By, Order By

Launch Sites Proximities Analysis

All Launch Sites On Map





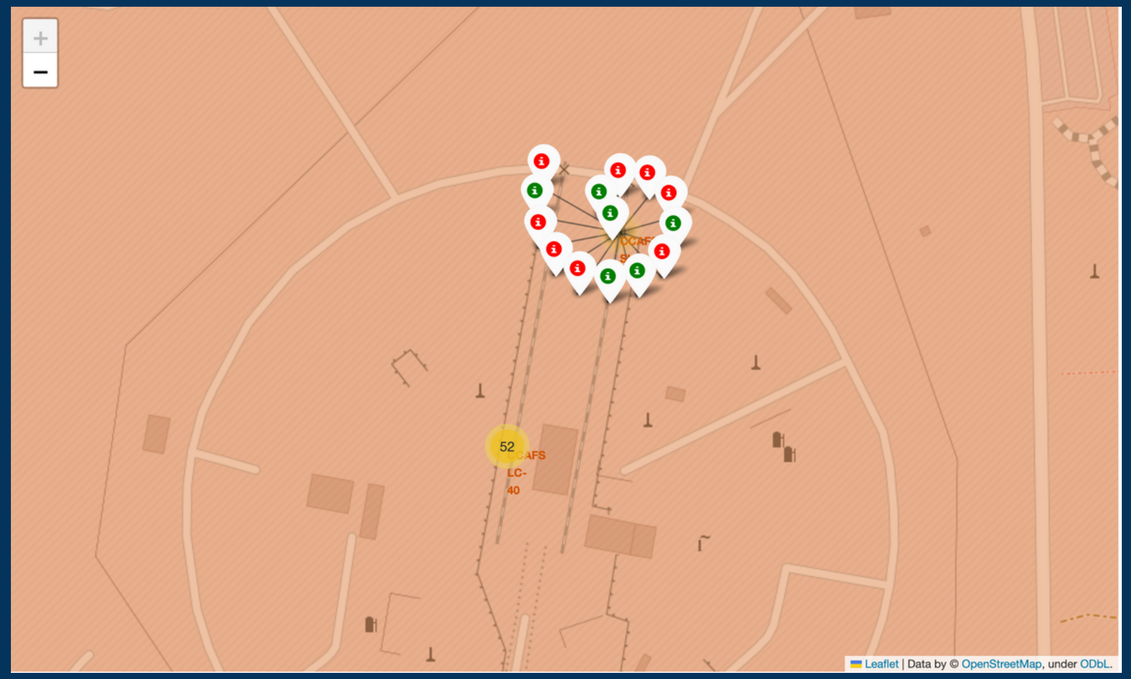
Elements

- create a folium Map object
- use folium.Circle and folium.Marker to add a highlighted circle area with a text label on a specific coordinate

Findings

- all launch sites are in proximity to the Equator line - help with the launch and orbit
- all launch sites are in very close proximity to the coast - to prevent failures inland

Success/Failed Launch For Each Site On Map



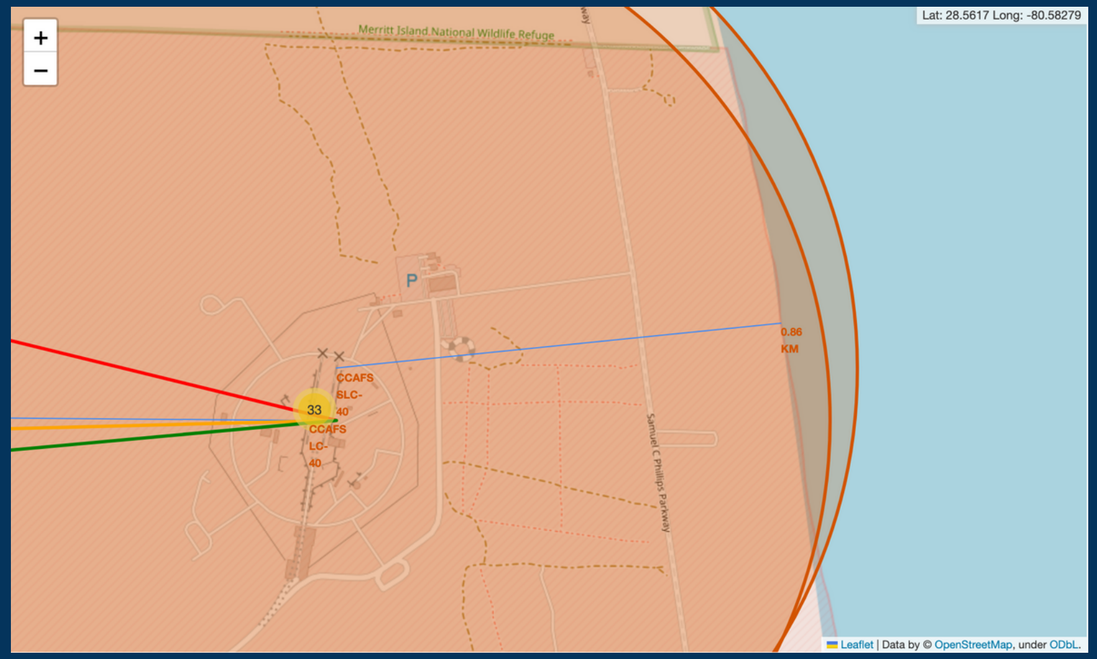
Elements

- create a MarkerCluster object
- create a new column in launch_sites data frame called marker_color to store the marker colors: successful launch use green marker and failed launch use red marker
- add a folium.Marker to marker_cluster

Findings

- KSC LC-39A: 10/13 success rate
- CCAFS SLC-40: 3/7 success rate
- CCAFS LC-40: 7/26 success rate
- VAFB SLC-4E: 4/10 success rate

Distances Between Launch Site to Its Proximities On Map



Elements

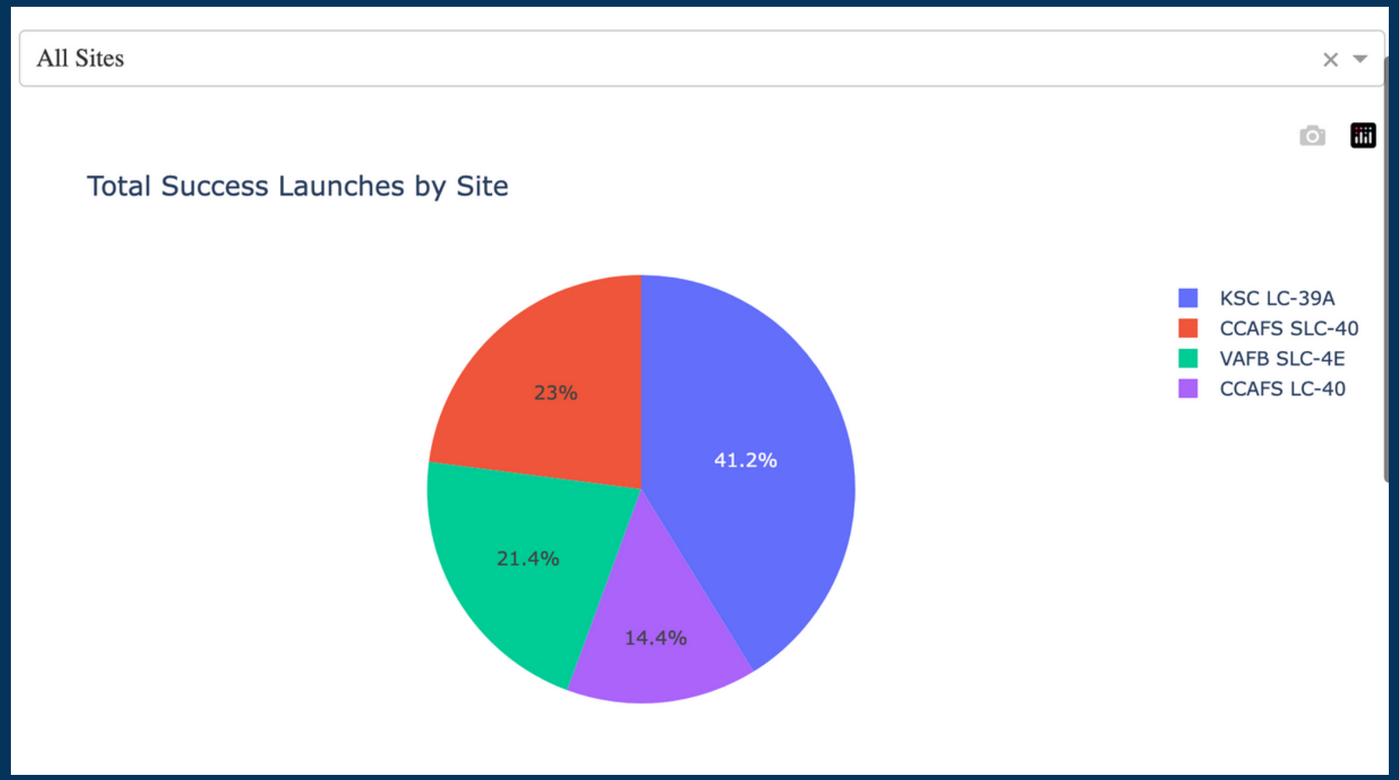
- add a MousePosition on the map to get coordinate
- mark down a point on the closest coastline using MousePosition and calculate the distance between the coastline point and the launch site
- draw a PolyLine between a launch site to the selected coastline point

Findings

- launch sites are usually pretty close to the coastline
- launch sites are usually a far from the railway and highway

Build A Dashboard with Plotly Dash

Launch Success Count for All Sites



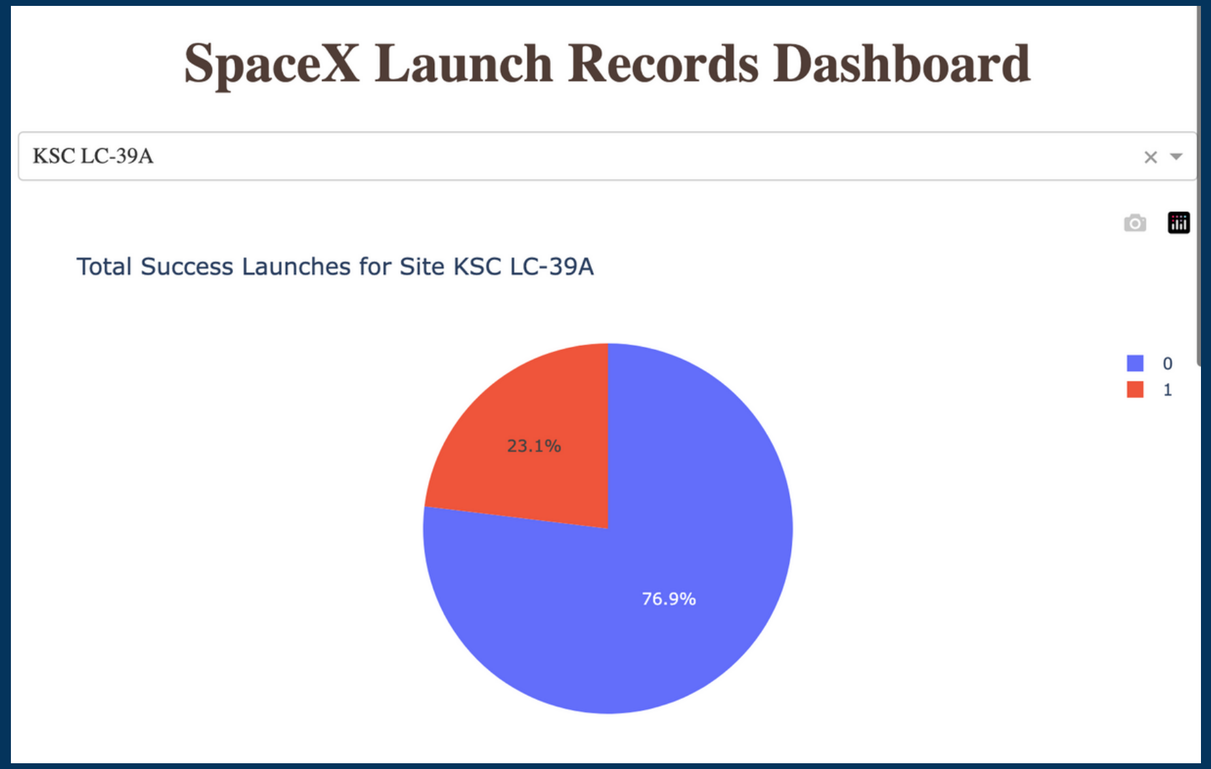
Elements

- generate a pie chart to demonstrate launch success amongst sites

Findings

- KSC LC-39A site has the most successful launches of all launch sites, taking up 41.2%

Launch Site with Highest Success Ratio



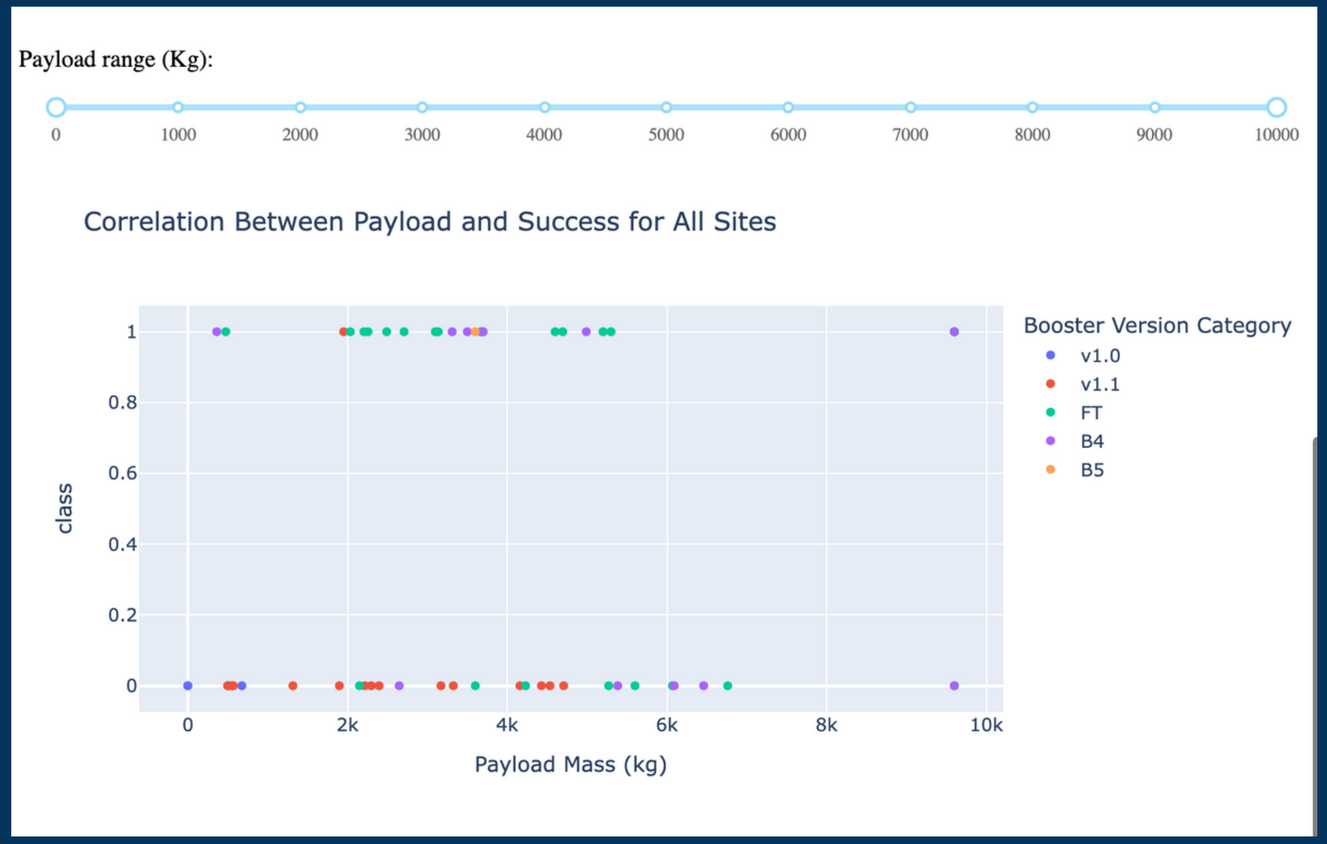
Elements

- generate a pie chart to demonstrate launch success vs. failure

Findings

- KSC LC-39A site has the highest success ratio which is 76.9%

Payload vs. Launch Outcome



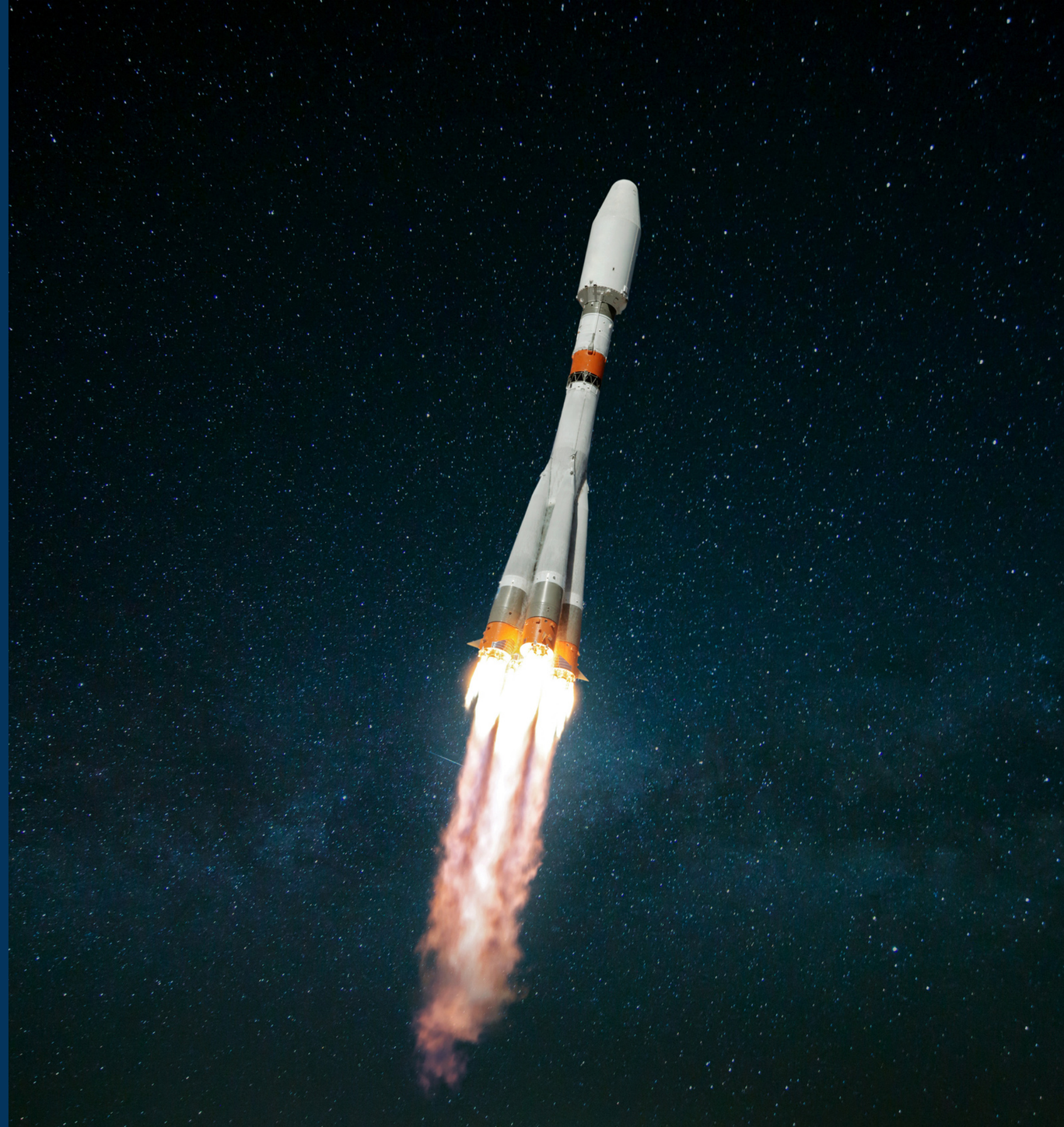
Elements

- generate a scatterplot to payload vs. launch outcome
- class 1 means successful launch
- class 2 means failed launch

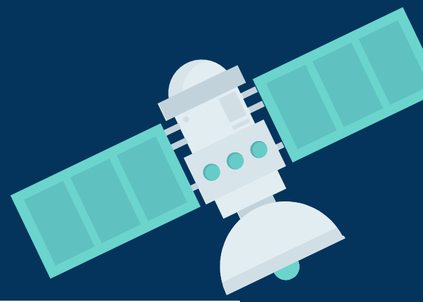
Findings

- payload mass between 2000 to 6000 kg has the highest success launch rate

Predictive Analysis (Classification)



Classification Accuracy



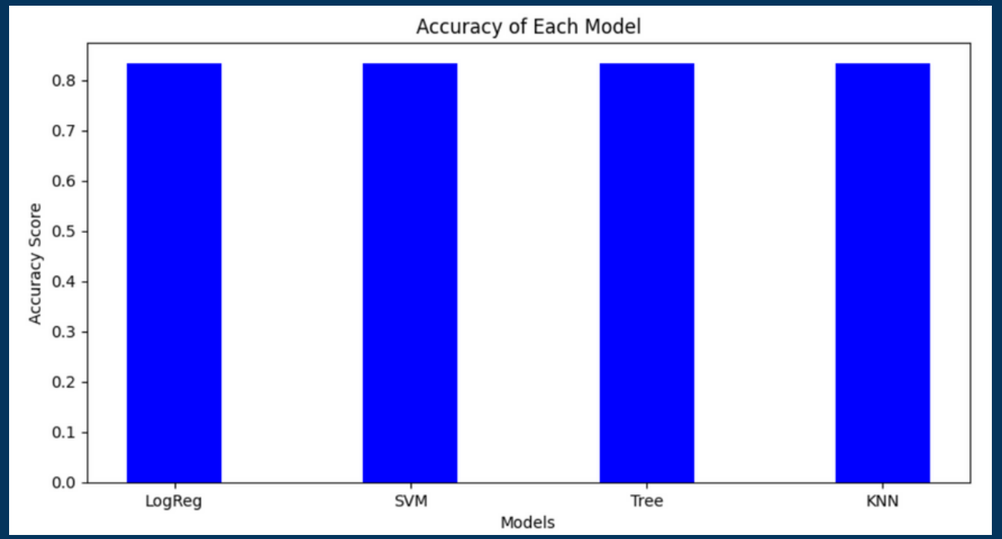
```
# test sets
jaccard_scores = [
                  jaccard_score(Y_test, logreg_yhat, average='binary'),
                  jaccard_score(Y_test, svm_yhat, average='binary'),
                  jaccard_score(Y_test, tree_yhat, average='binary'),
                  jaccard_score(Y_test, knn_yhat, average='binary'),
                 ]

f1_scores = [
              f1_score(Y_test, logreg_yhat, average='binary'),
              f1_score(Y_test, svm_yhat, average='binary'),
              f1_score(Y_test, tree_yhat, average='binary'),
              f1_score(Y_test, knn_yhat, average='binary'),
             ]

accuracy = [logreg_accuracy, svm_accuracy, tree_accuracy, knn_accuracy]

scores = pd.DataFrame(np.array([jaccard_scores, f1_scores, accuracy]), index=['Jaccard_Score', 'F1_Score', 'Accur
scores
```

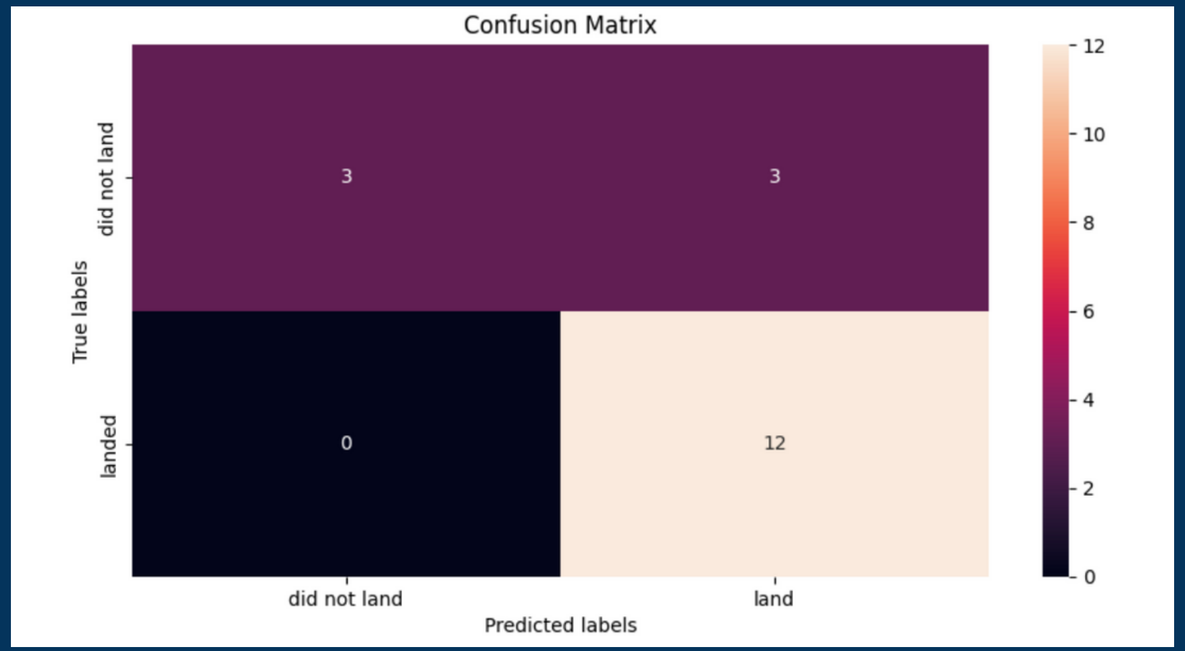
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333



Findings

- all models' classification accuracy scores are the same
  - Jaccard_Score = 0.800
  - F1_Score = 0.8889
  - Accuracy = 0.8333

Confusion Matrix



Findings

- all models' confusion matrices are the same
  - 12 True positive
  - 3 True negative
  - 3 False positive
  - 0 False Negative
- 3 False positives could be an issue as the launch is predicted to have a successful launch but didn't

Conclusion

Conclusion

- Falcon 9 launches have improved progressively
- As payload mass increases, launch success rate also increase - indicating a positive relationship
- Orbit ES-L1, GEO, HEO, and SSO have a 100% launch success rate
- Launch sites are usually near the Equator line since the spacecraft can take advantage of the Earth's rotational speed
- Launch sites are also in close proximity to the coast to ensure safety
- KSC LC-39A has the highest launch success rate of all sites
- All models, logistic regression, SVM, decision tree, and KNN have very similar performance

Appendix

- Relevant Assets
  - Python
  - Jupyter Notebook
  - Coursera
  - IBM
  - Skills Network Labs


THANK YOU!