IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fatih Çağlar

5/17/2024

PERFECTING PROPULSIVE LANDING

OUTLINE

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

EXECUTIVE SUMMARY

- Our objective in this study is to conduct a predictive analysis on SpaceX Falcon-9 launches on 4 different sites in US. We collected data through SpaceX API and web scraping. After preprocessing the data and exploratory data analysis, we applied machine learning algorithms on data to be able to predict the outcomes of future launches.
- Our model needs some improvements regarding the predictions about launches that lands back. It is better on negative cases at hand. The tests on out-source data didn't be conducted yet.

INTRODUCTION

In this project, we will try to predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

If we can determine if the first stage will land, we can determine the cost of a launch.
This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

METHODOLOGY

- Data collection methodology:
  - SpaceX launch data has been collected via SpaceX API and scraping data from Wikipedia.
- Performed data wrangling
  - After data wrangling, we ended up with a EDA-ready dataset consisting of 18 different features including location, equipment, outcome, payload mass and datetime information for 90 different launches.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - We trained four different classification model on the dataset and proceeded with Decision Tree which holding highest accuracy.

DATA COLLECTION

- Through SpaceX RESTAPI:

- SpaceX launch data has been requested and parsed using GET request.

- 17 different features are stored in a dataframe using customized functions to get data from json object easily.

- Data got filtered to only include Falcon-9 launches.

DATA COLLECTION

- Through Web Scraping:

- Data got scraped from Falcon-9 launch Wiki page using BeautifulSoup.

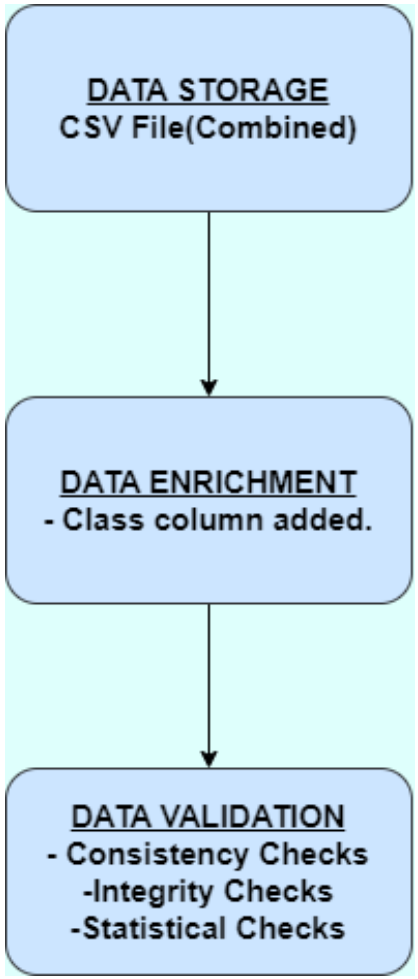
- HTML content got handled using loops and customized functions

- There were 2 columns with missing values. One imputed by its mean and the other one leaved as it is since it was an intentional missing.

DATA COLLECTION

- Code to collect data through web scraping.
- Code to collect data using API.

DATA WRANGLING



- In this stage, distinct launch sites, orbits and landing outcomes and their number of appearances were explored. "Class" feature regarding landing outcome was added to dataframe for machine learning purposes. The mean of class column is approximately 0.67 ie. 67% of launches were successful.
- Code

EDA WITH DATA VISUALIZATION

- Scatterplots (all colored by outcomes):

- Payloads vs. Flight Numbers (Do payloads change over time? Are there any significant fluctuations?)
- Launch sites vs. Flight Number (Which launch site hosts most flights? Any change by time?)
- Launch Sites vs. Payload Mass (Which launch sites are used for small/large payloads? Any pattern?)
- Orbits vs Flight Number (What is the effect of orbits to which rockets launched on outcome? What is the distribution over flight numbers?)
- Payloads vs. Orbit (What are the mean payloads & outcomes for different orbits?)

EDA WITH DATA VISUALIZATION

- Success Rates of Different Orbits – Bar Chart (How success rate changes through orbits?)
- Success Rate Through Years – Line Chart (How success rate change through years?)
- Code

EDA WITH SQL

- Queries:

  - Total payload mass carried by boosters
  - Average payload mass carried by specific boosters
  - The date when the first succesful landing outcome in ground pad was achieved
  - Boosters which have success in drone ship with specific payload range
  - Total number of successful and failure mission outcomes
  - Some other observations for certain time periods

- Code

BUILD AN INTERACTIVE MAP WITH FOLIUM

- At this stage, we created a folium map centered at NASA Johnson Space Center and added circles for each of tree launch sites with a marker with their names. We showed launches within these circles using colored icons to denote outcomes.
- We calculated the distance to nearest coastline using haversine formula and added lines between launch sites and coastlines.

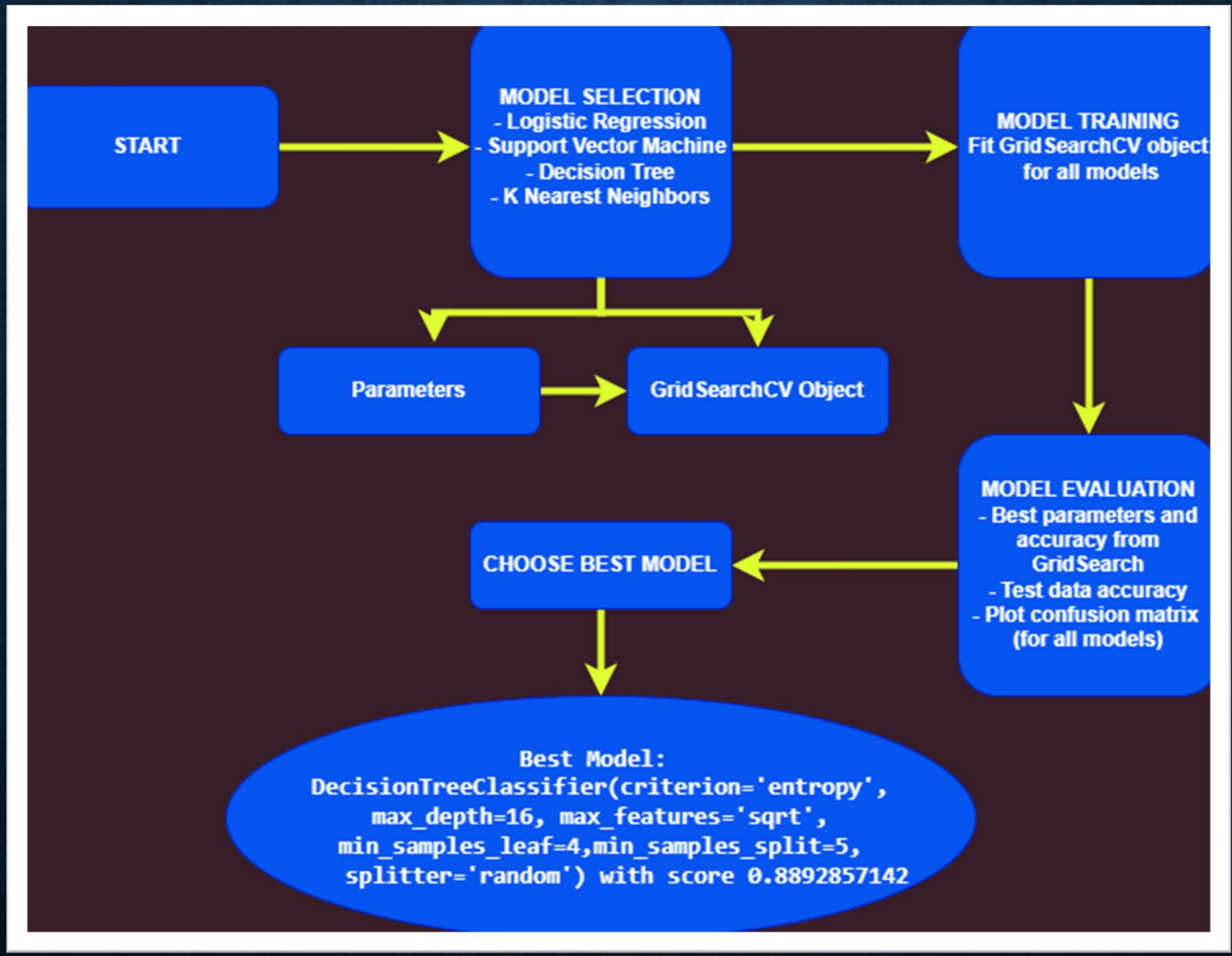
- Code

BUILD A DASHBOARD WITH PLOTLY DASH

- SpaceX Launch Records Dashboard consists of a dropdown menu to select a launch site (or all), a pie chart that shows outcome rates for launch sites and a scatter plot that shows relation between outcomes and payloads. There is also a slider to select a range for payload.
- This dashboard gives stakeholders the opportunity of exploring data more deeply themselves and provides a space without a crowd of plots thanks to its interactivity.
- Code

PREDICTIVE ANALYSIS (CLASSIFICATION)

- To find the best model for Falcon-9 first stage landing prediction, we selected 4 different models, and passed the parameter grid for these models to GridSearchCV and fitted on our training data which is 80% of all dataset.
- We evaluated the models using best parameters and best score grid search gave and using our test data. Decision tree gave the best score which is approximately 0.89.
- Code


START
MODEL SELECTION
- Logistic Regression
- Support Vector Machine
- Decision Tree
- K Nearest Neighbors
MODEL TRAINING
Fit GridSearchCV object
for all models
Parameters
GridSearchCV Object
MODEL EVALUATION
- Best parameters and
accuracy from
GridSearch
- Test data accuracy
- Plot confusion matrix
(for all models)
CHOOSE BEST MODEL
Best Model:
DecisionTreeClassifier(criterion='entropy',
max_depth=16, max_features='sqrt',
min_samples_leaf=4,min_samples_split=5,
splitter='random') with score 0.8892857142

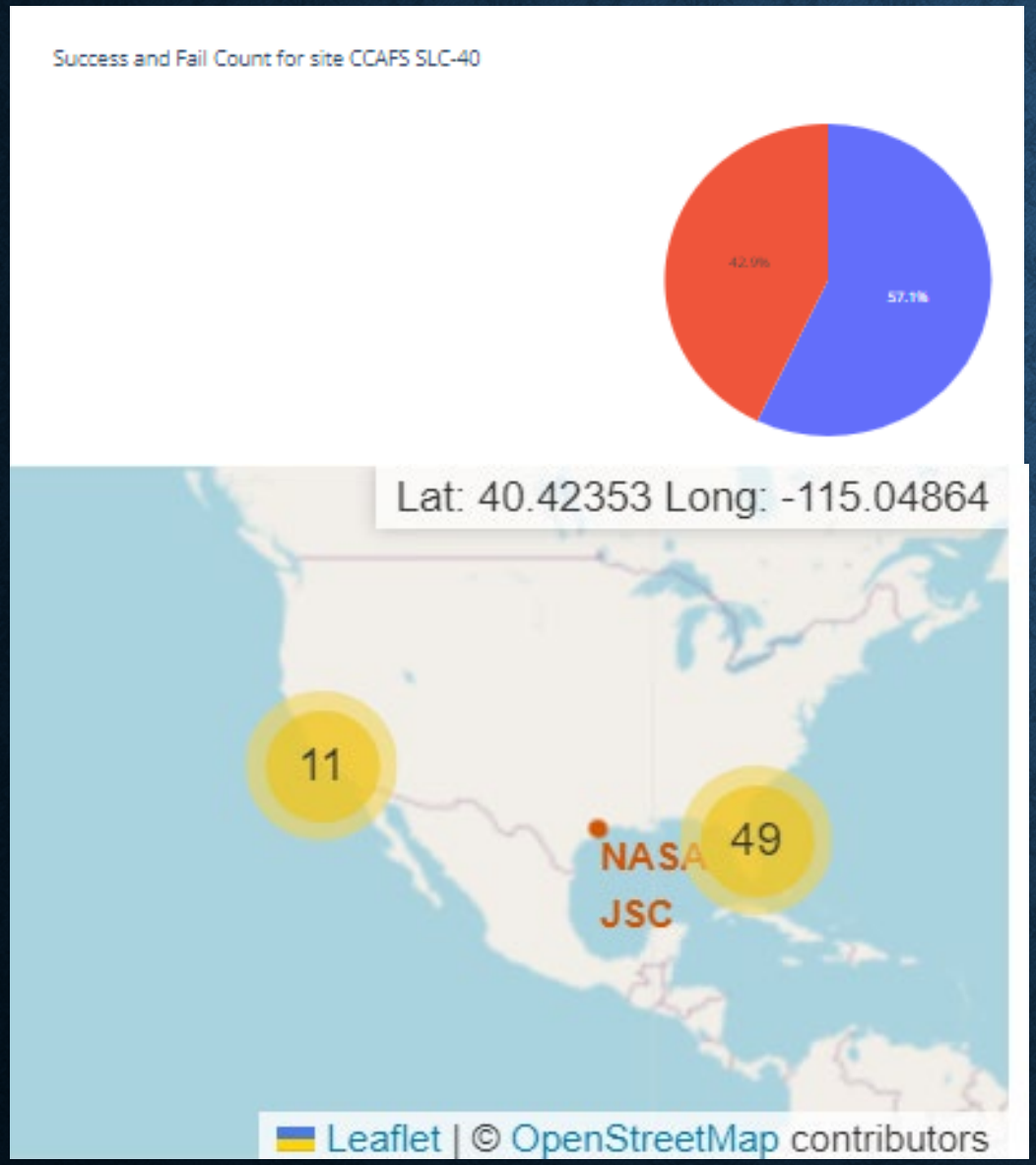
RESULTS

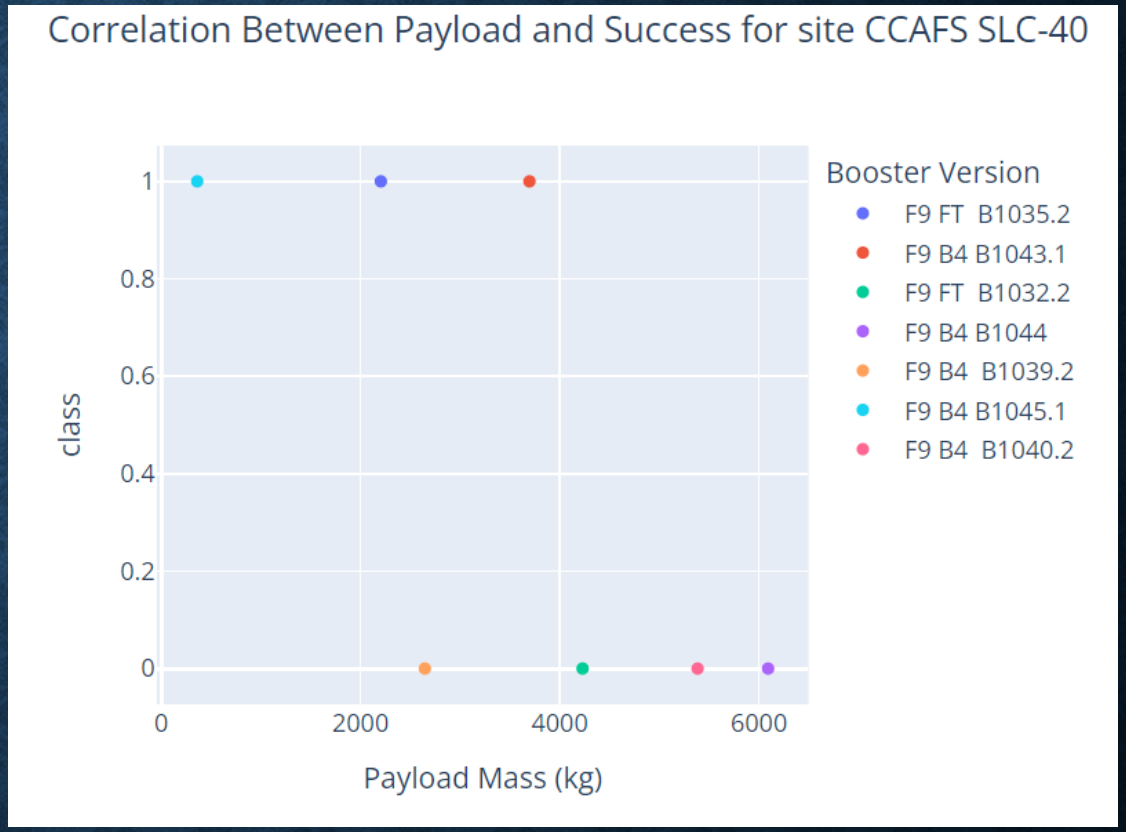
- Some results from EDA
- Interactive analytics demo in screenshots
- Predictive analysis results

INSIGHTS

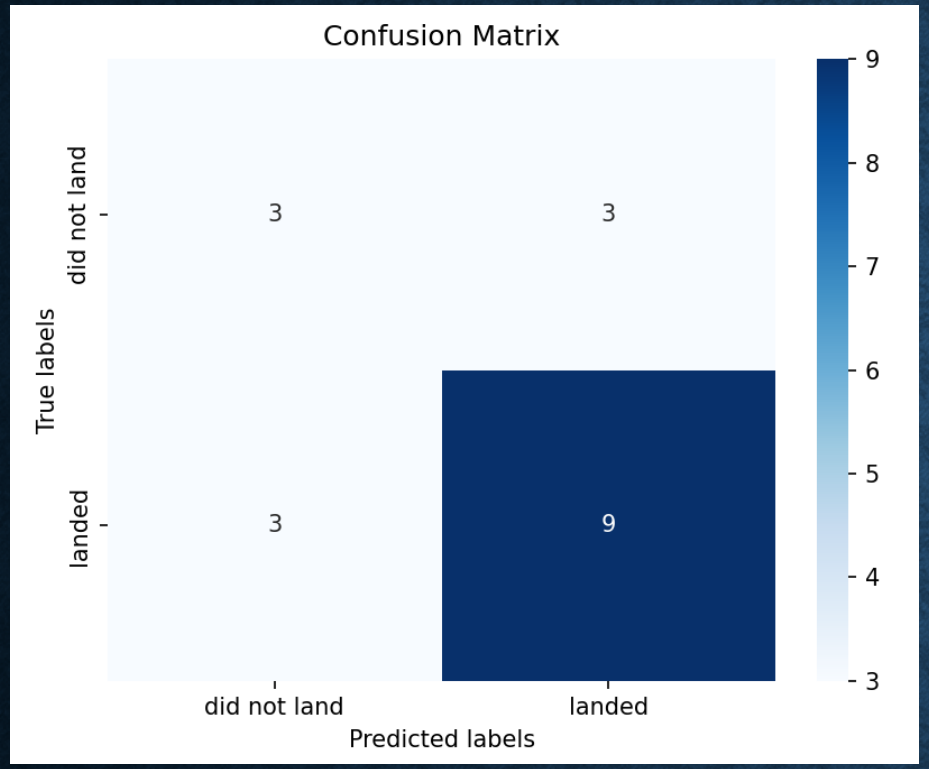
- Total success rate is 0.67.
- There are 99 successful missions. One mission had failed and payload status of one successful mission is unclear.
- CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Out of 11 unique orbits, 4 ones have a success rate of 100% and orbit SO has not any successful landing.
- The success rate since 2013 kept increasing till 2020 except a down of 25% in 2018.
- For approximately 75% of flights, payload mass is less than 10 tones.

CHARTS AND MAPS





Check more at [Live app]



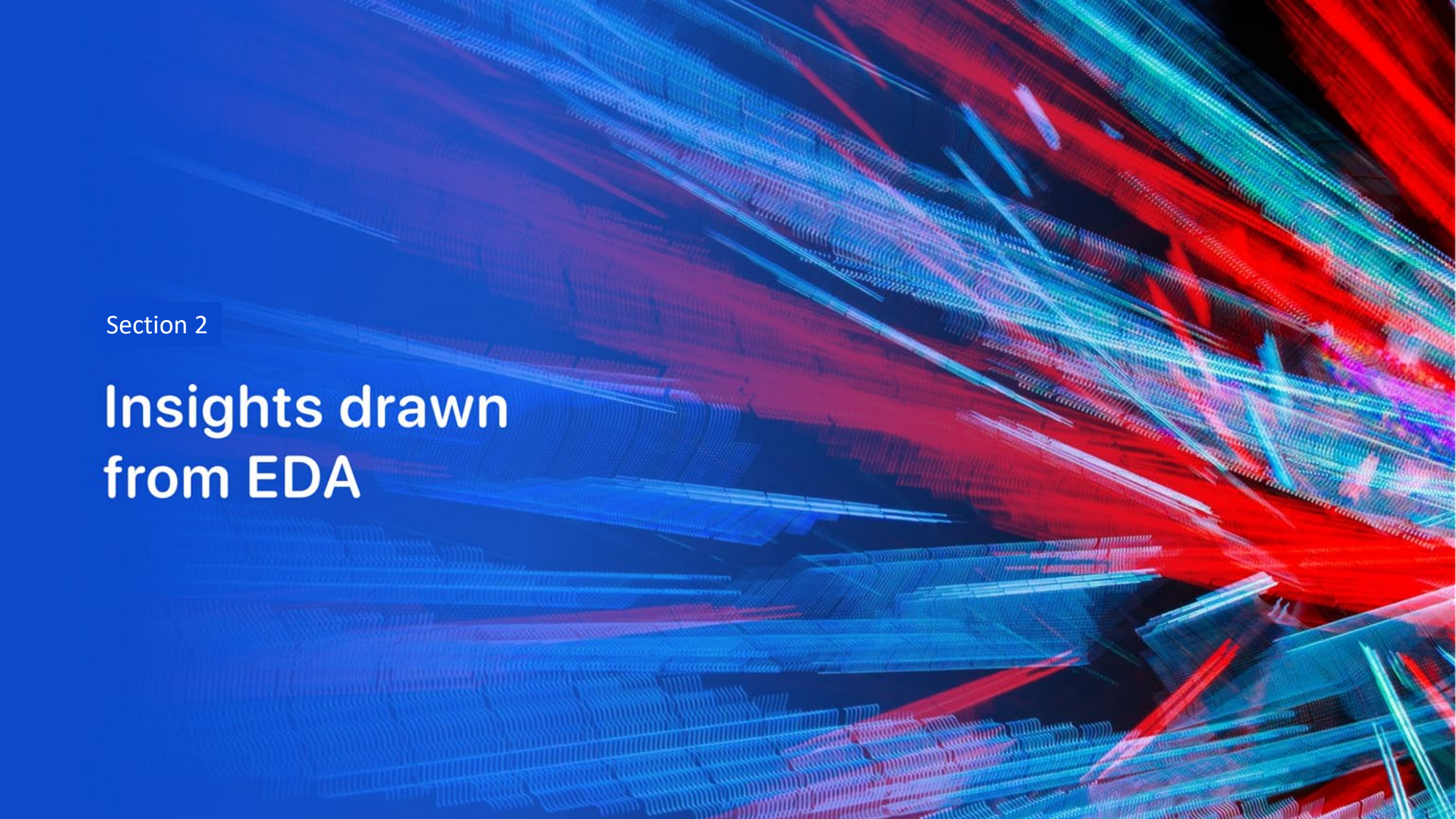

```
model_grid_searches = [logreg_cv, svm_cv, tree_cv, knn_cv]
best_model = None
best_score = 0
for search in model_grid_searches:
    model_score = search.best_score_
    print(f"{search.best_estimator_}: {model_score}")
    if model_score > best_score:
        best_score = model_score
        best_model = search.best_estimator_

print(f"Best Model: {best_model} with score {best_score}")
```

```
LogisticRegression(C=0.01): 0.8464285714285713
SVC(gamma=0.03162277660168379, kernel='sigmoid'): 0.8482142857142856
DecisionTreeClassifier(criterion='entropy', max_depth=16, max_features='sqrt',
                       min_samples_leaf=4, min_samples_split=5,
                       splitter='random'): 0.8892857142857142
KNeighborsClassifier(n_neighbors=10, p=1): 0.8482142857142858
Best Model: DecisionTreeClassifier(criterion='entropy', max_depth=16, max_features='sqrt',
                       min_samples_leaf=4, min_samples_split=5,
                       splitter='random') with score 0.8892857142857142
```

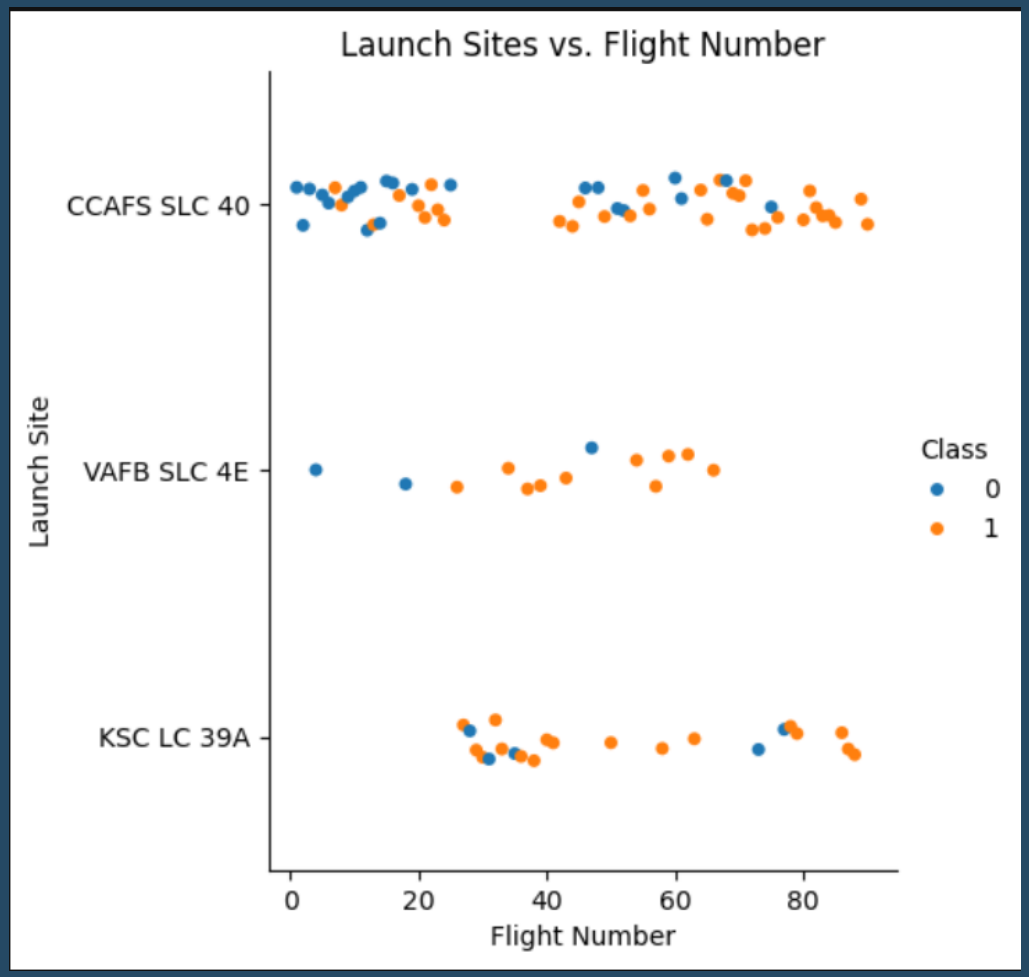
PREDICTIVE ANALYSIS RESULTS

- Best machine learning model for the dataset is Decision Tree.
- Its accuracy is 0.889.


Section 2
Insights drawn
from EDA

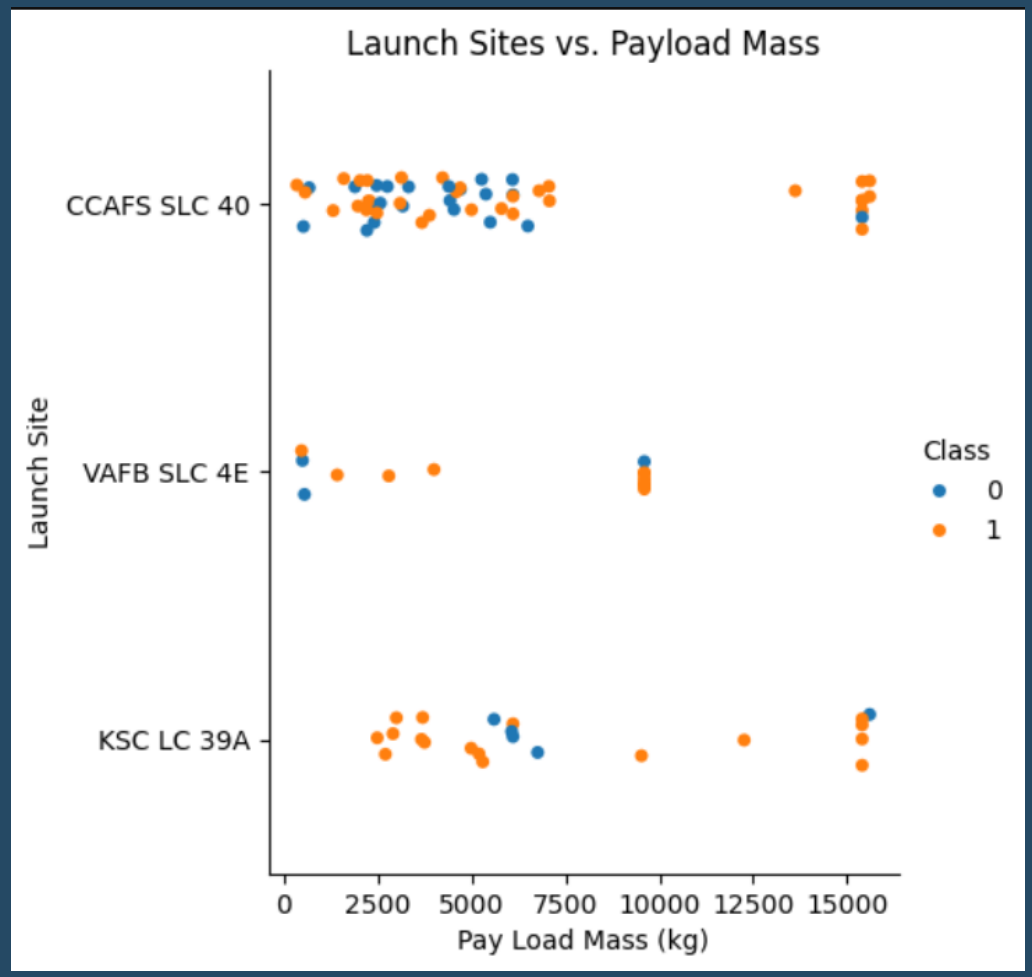
FLIGHT NUMBER VS. LAUNCH SITE

- Sites VAFB SLC 4E and KSC LC 39A seem to have high success rates. CCAFS SLC 40 hosts large number of launches but it's success rate is not that much.



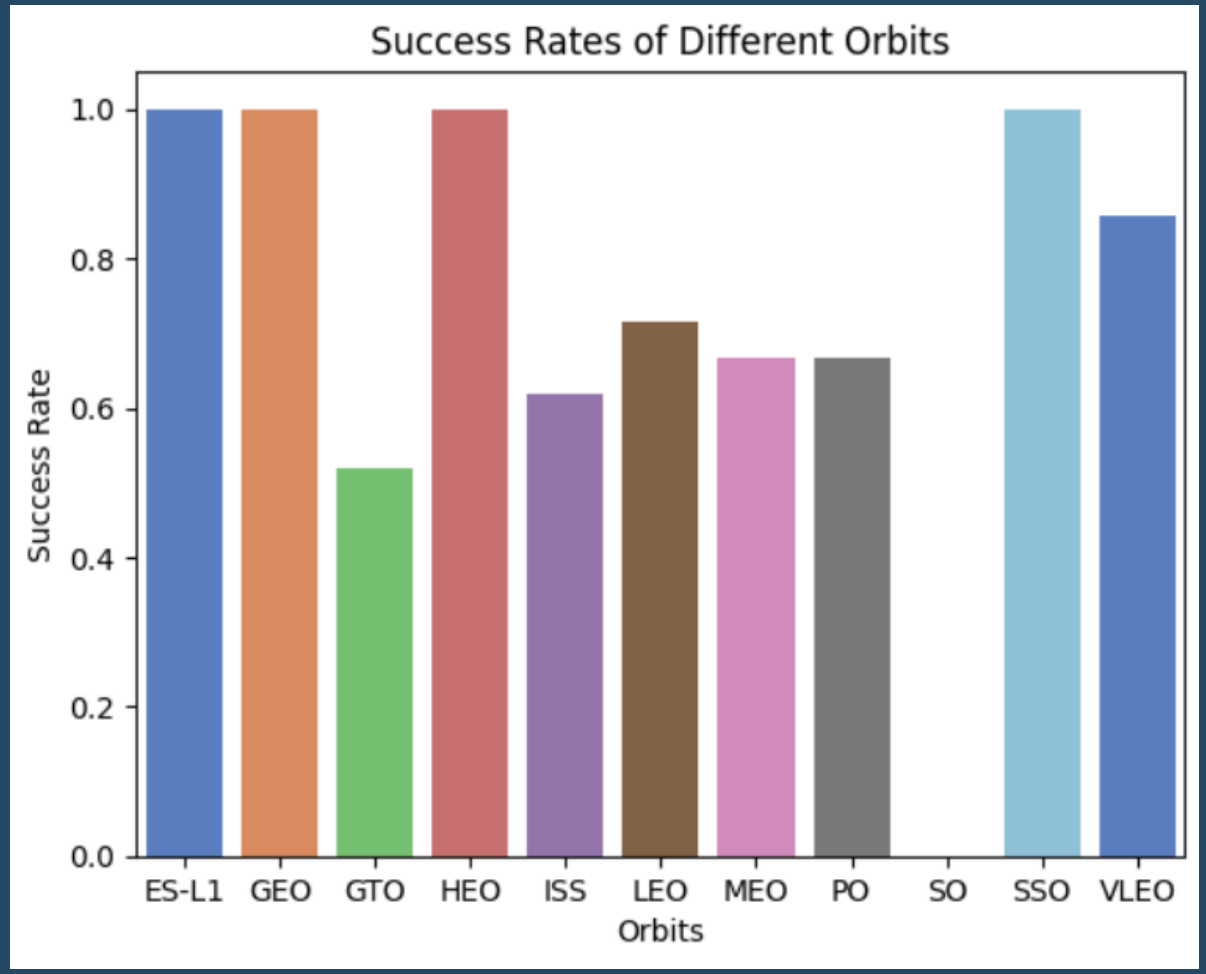
PAYLOAD VS. LAUNCH SITE

- Most of launches are with payloads less than 10000 kg.
- Launches with payloads greater than 10000 are from CCAFS SLC 40 & KSC LC 39A and of very high success rate.



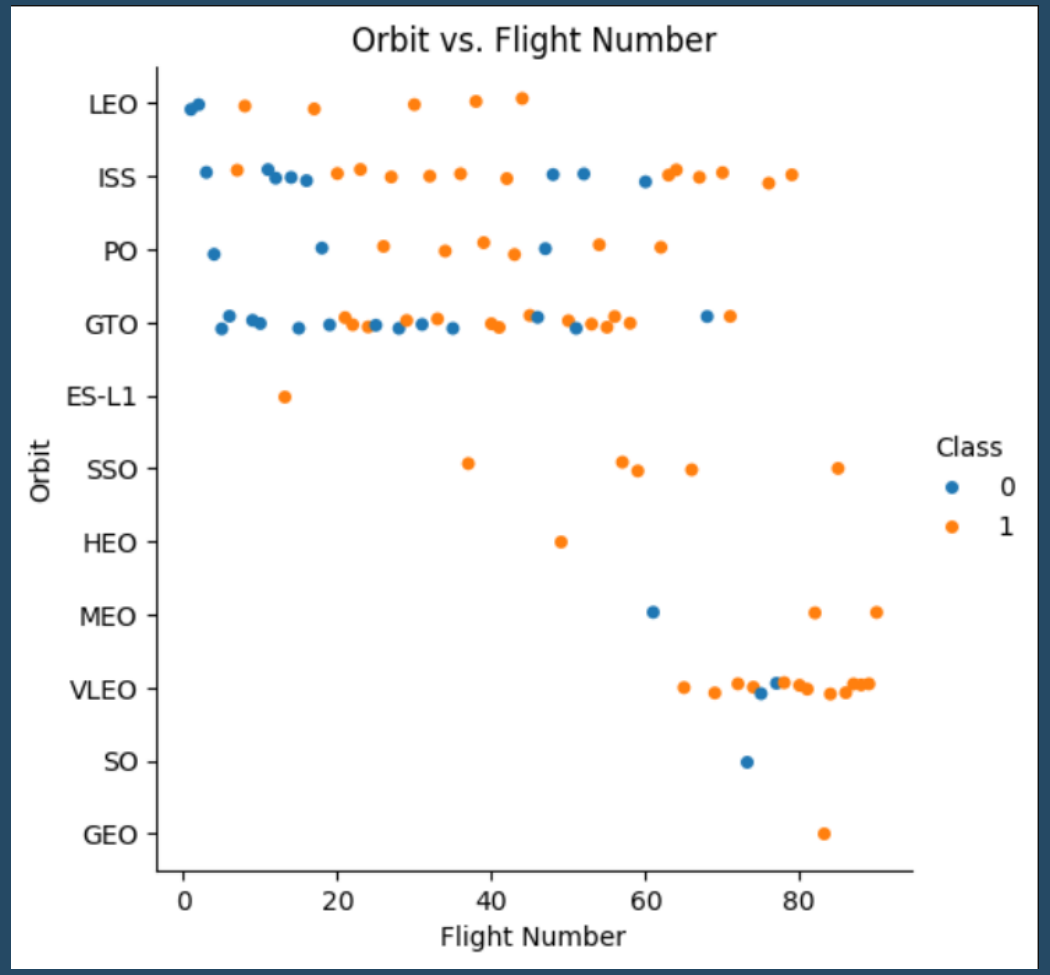
SUCCESS RATE VS. ORBIT TYPE

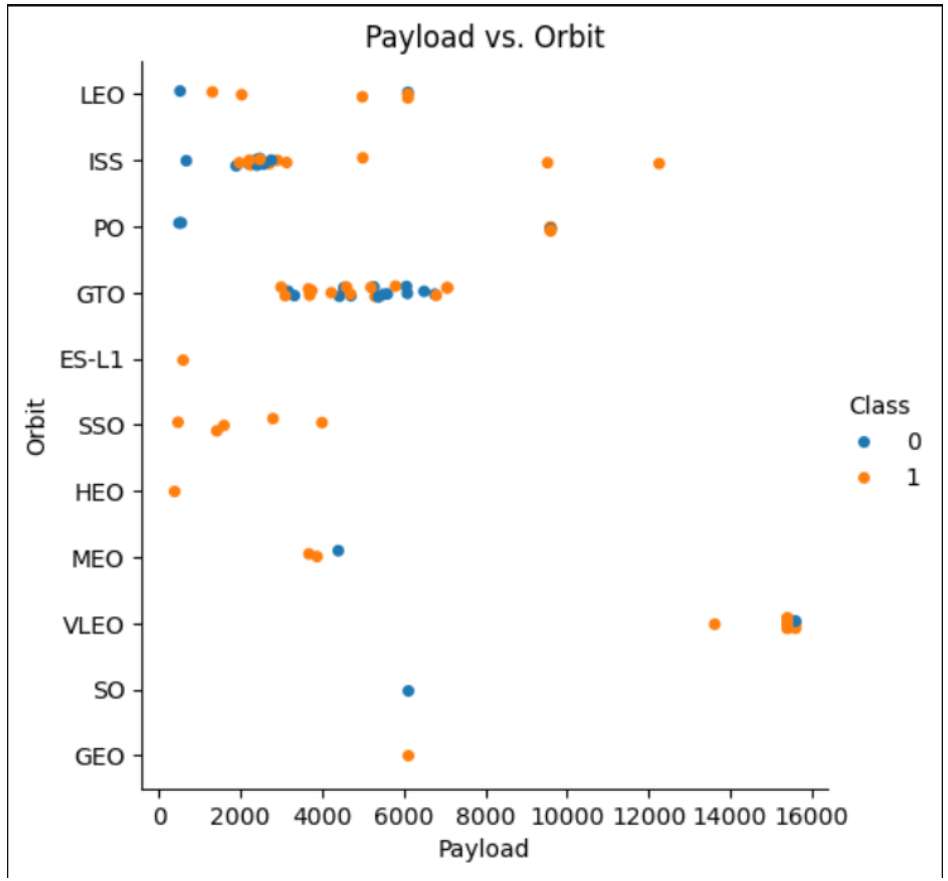
- There are 11 different orbits and 4 of them have full success rate. 6 orbits have success rate between 0.5 and 0.8 and launches to one orbit are all failed.



FLIGHT NUMBER VS. ORBIT TYPE

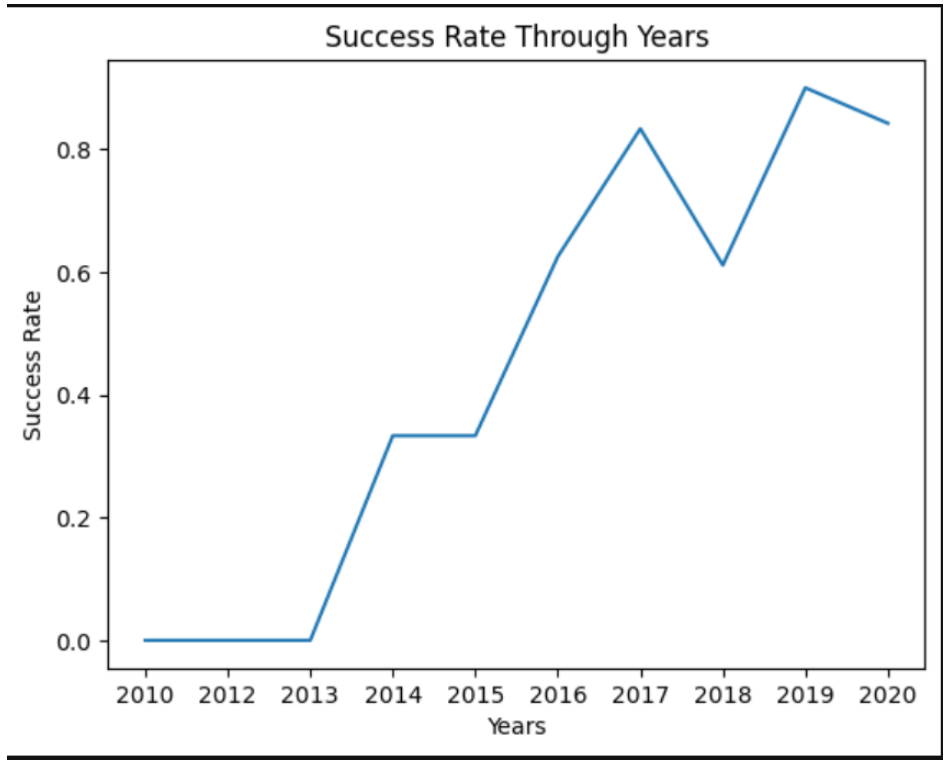
- First 20 launches have very low success rate and after 20th, success rate is very high. 7 of orbits are not preferred much until 60th launch.





Payload vs. Orbit Type

Highest number of launches are to GTO and their payload is in range of 2000-8000.



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020 except a down of 25% in 2018.

```
%sql select distinct Launch_Site from spacextable
```

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

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

All Launch Site Names

There are 4 unique launch sites.

```
%sql select * from spacextable 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	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

Launch Site Names Begin with 'CCA'

See the similarity between orbits and customers.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Total Payload Mass Carried by Boosters from NASA

Total payload mass carried by boosters from NASA is 45596 kg.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
AVG(PAYLOAD_MASS__KG_)
2928.4
```

Average Payload Mass by F9 v1.1

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

```
%sql SELECT MIN(Date) AS 'First Successful Landing Date' FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

First Successful Landing Date
2015-12-22

First Successful Ground Landing Date

First successful ground landing date is 2015-12-22.

```
%%sql
SELECT DISTINCT Booster_Version FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND (PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000)
```

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

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
--there are 2 different 'Success' one of which has extra spaces
SELECT TRIM(Mission_Outcome) AS Clean_Mission_Outcome, COUNT(*)
FROM SPACEXTABLE
GROUP BY Clean_Mission_Outcome;
```

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

Clean_Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
ORDER BY Booster_Version
```

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

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

Boosters Carried Maximum Payload

Failed Launches in 2015

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
%%sql
SELECT
    CASE SUBSTR(Date, 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS Month_Name,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM
    SPACEXTABLE
WHERE
    SUBSTR(Date, 0, 5) = '2015'
    AND Landing_Outcome = 'Failure (drone ship)'
```

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

```
%%sql
SELECT Landing_Outcome,COUNT(*) AS 'Count' FROM SPACEXTABLE
WHERE Date > '2010-06-04' AND Date < '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT(*) DESC
```

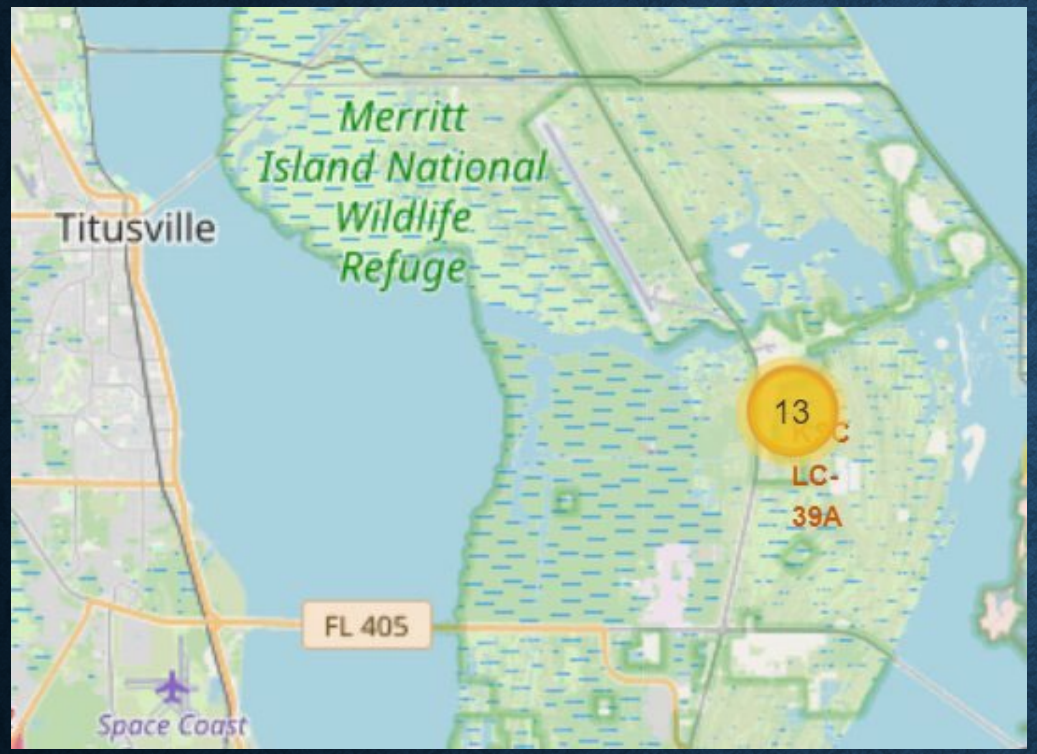
```
 * sqlite:///my_data1.db
Done.
```

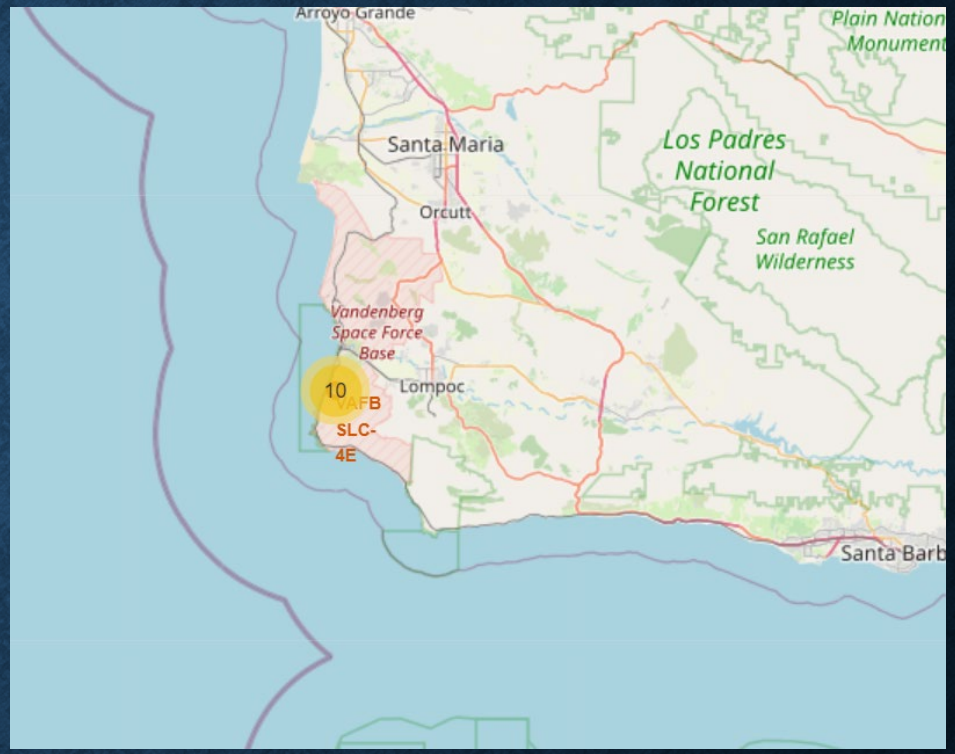
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

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

Section 3

Launch Sites Proximities Analysis



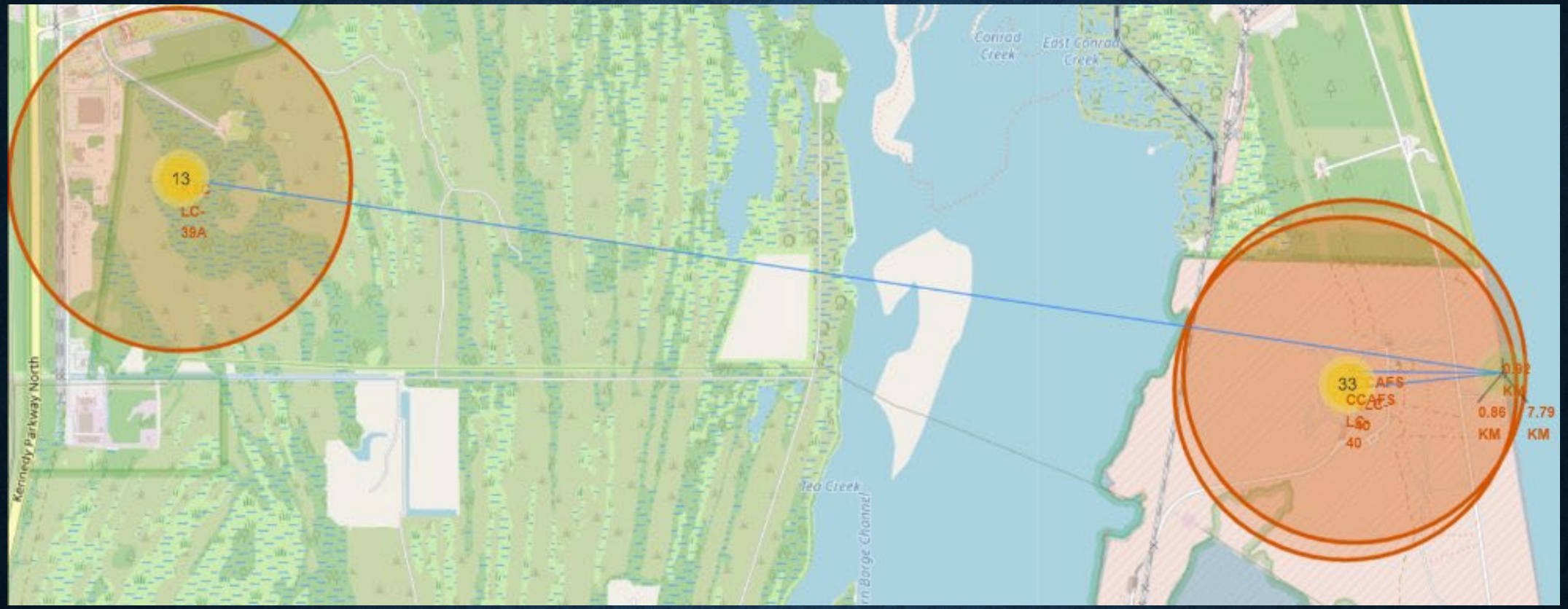


Locations of Launch Sites

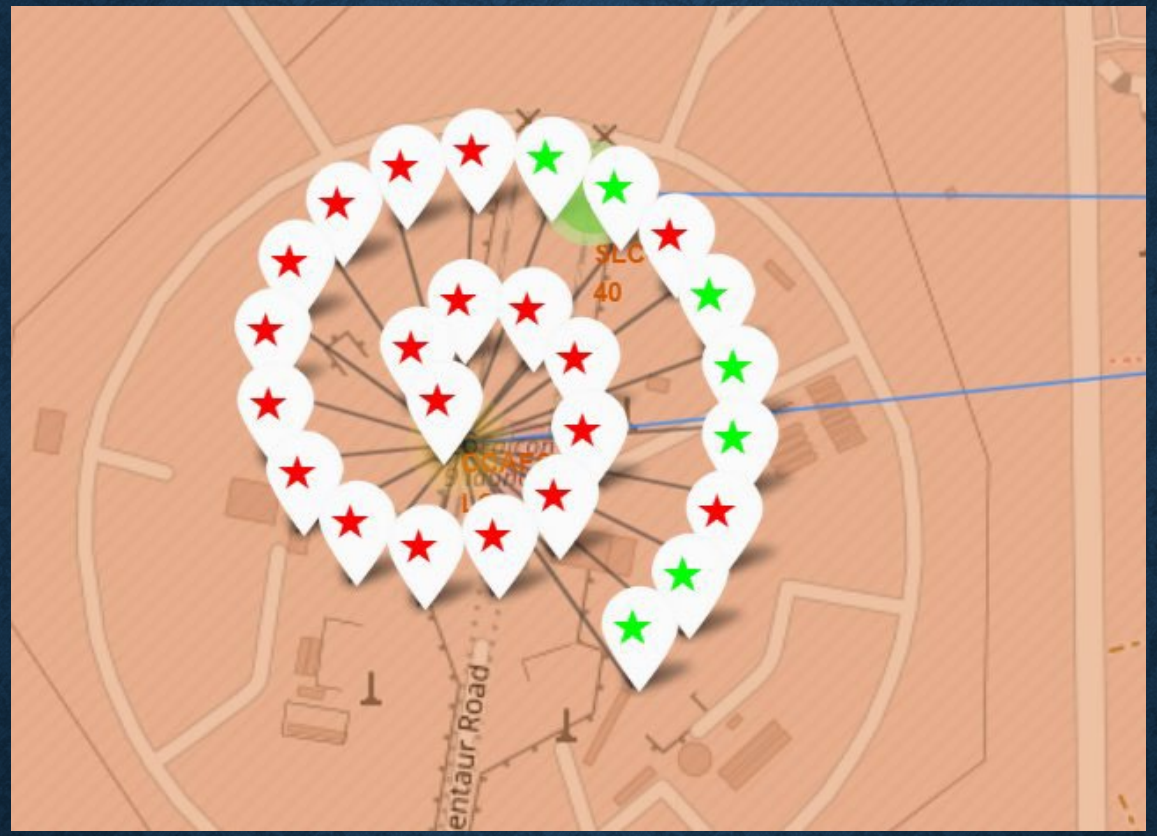
Locations of Launch Sites

VAFB SCL-4E Launch Outcomes & Coastline Proximity

Green stars denote successes and red stars denote failures.



Coastline Proximities of Other Sites

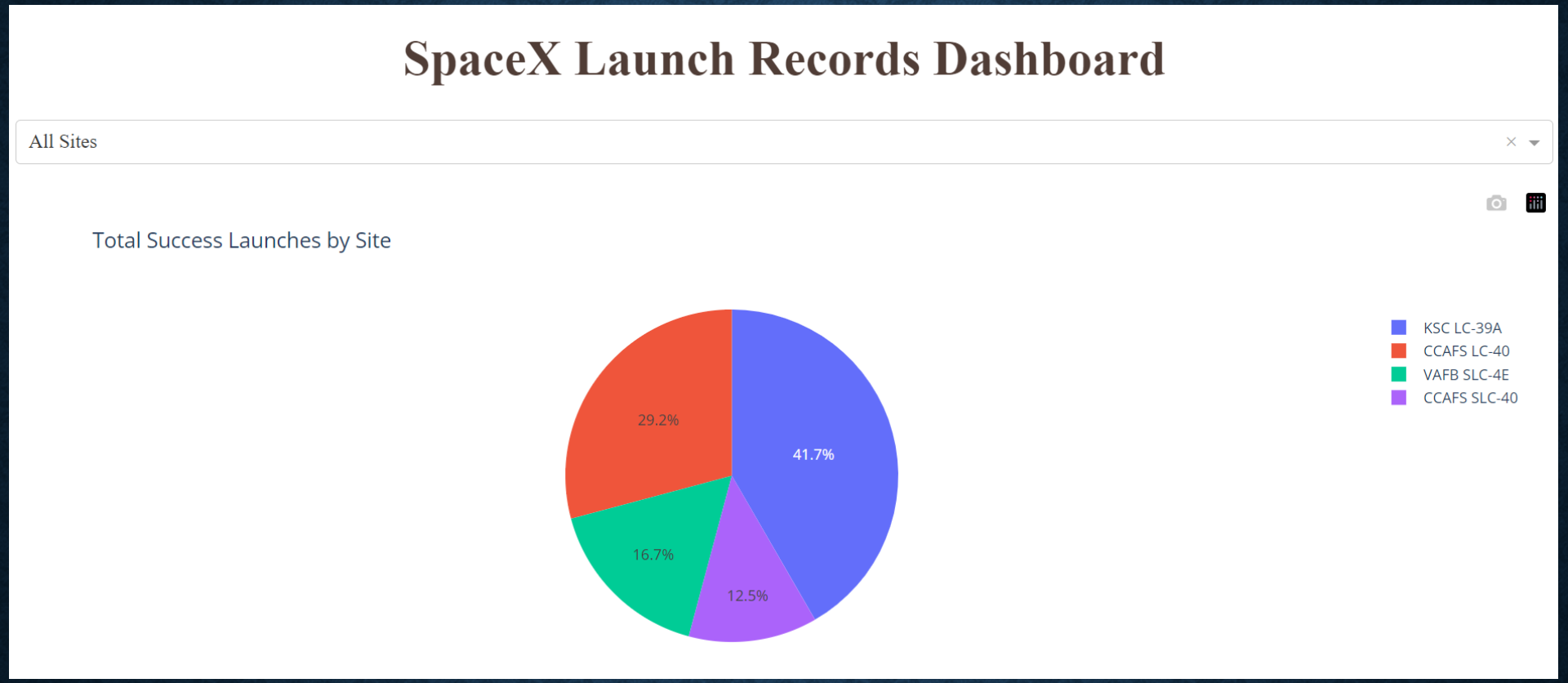


Launch Outcomes for CCAFS-LC-40

Check more at live app

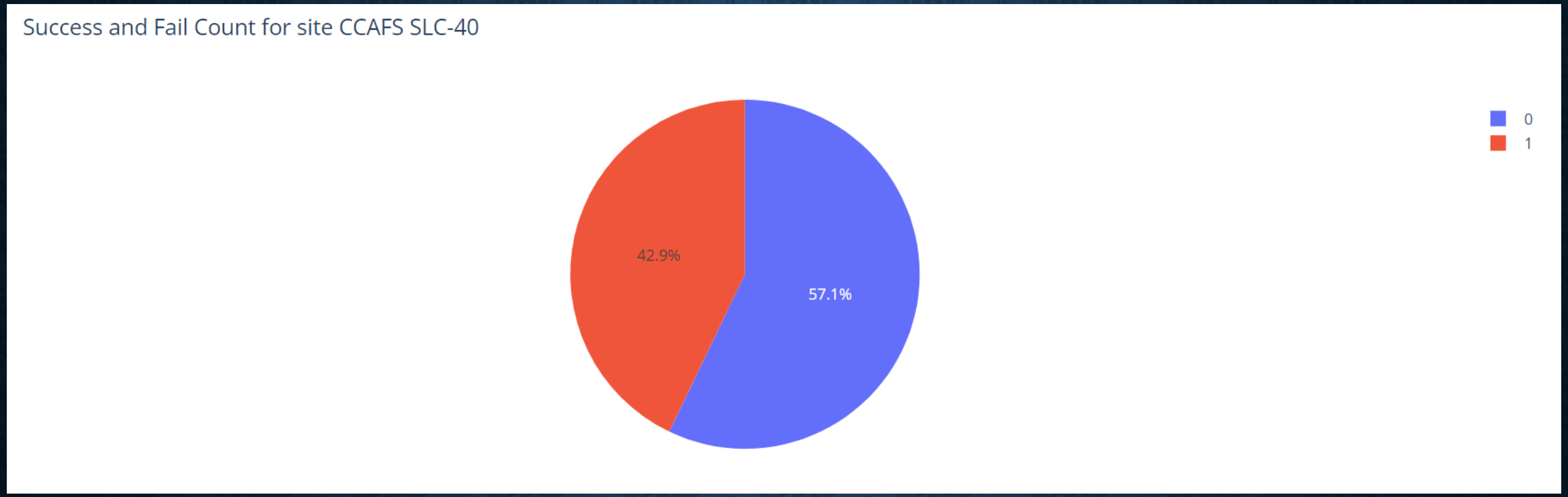
Section 4

Build a Dashboard with Plotly Dash

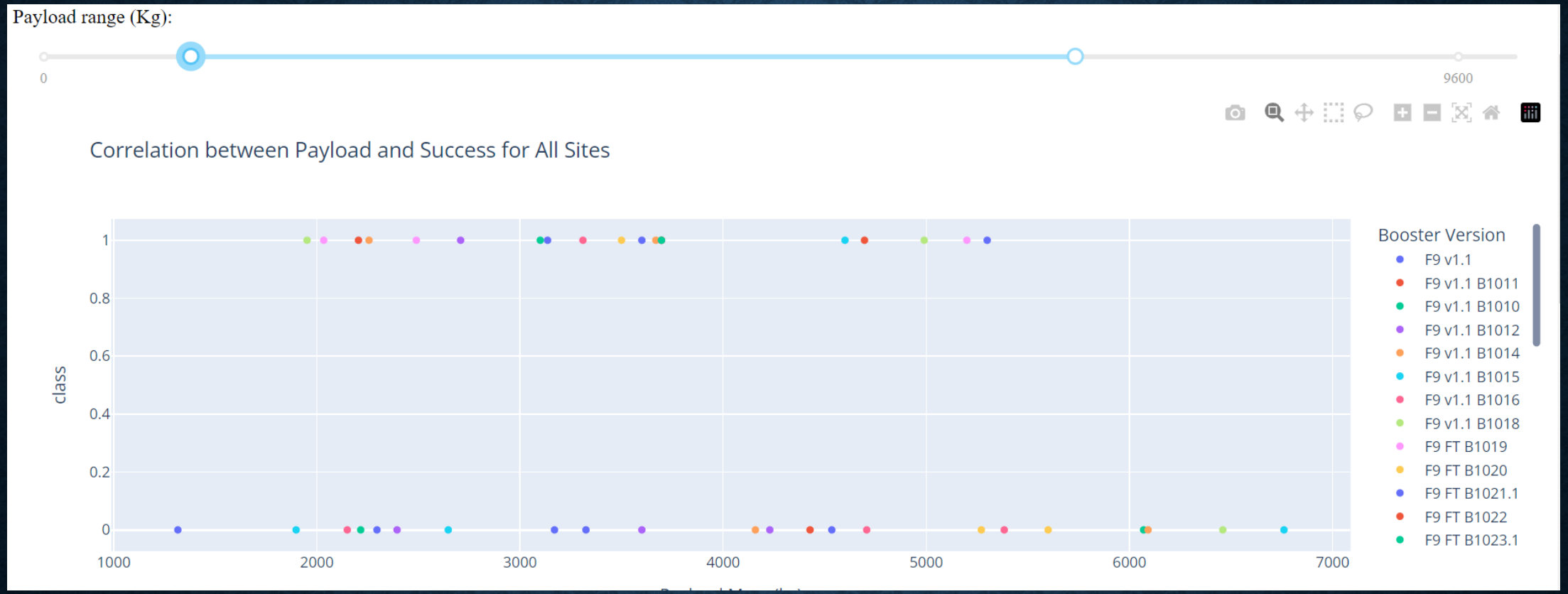


Launch Success Count for All Sites

41.7 % of launches are from KSC LC-39A which is the highest.



Success vs. Fail Rate for the Site with Highest Success Rate

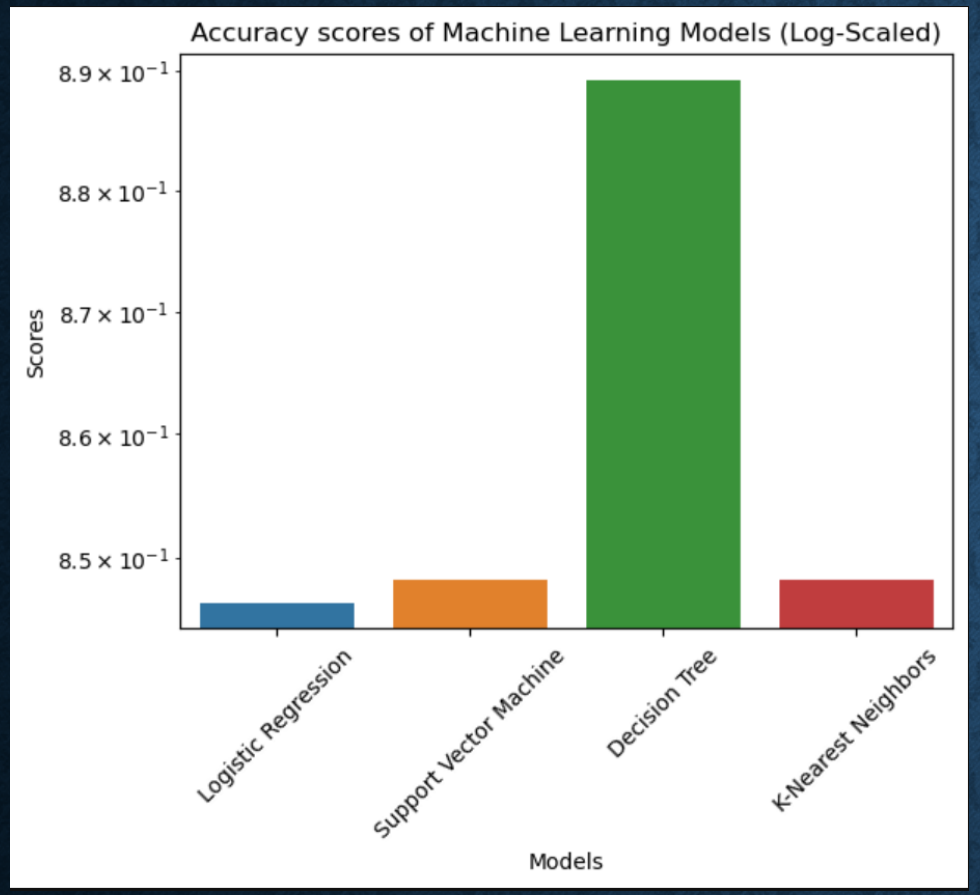


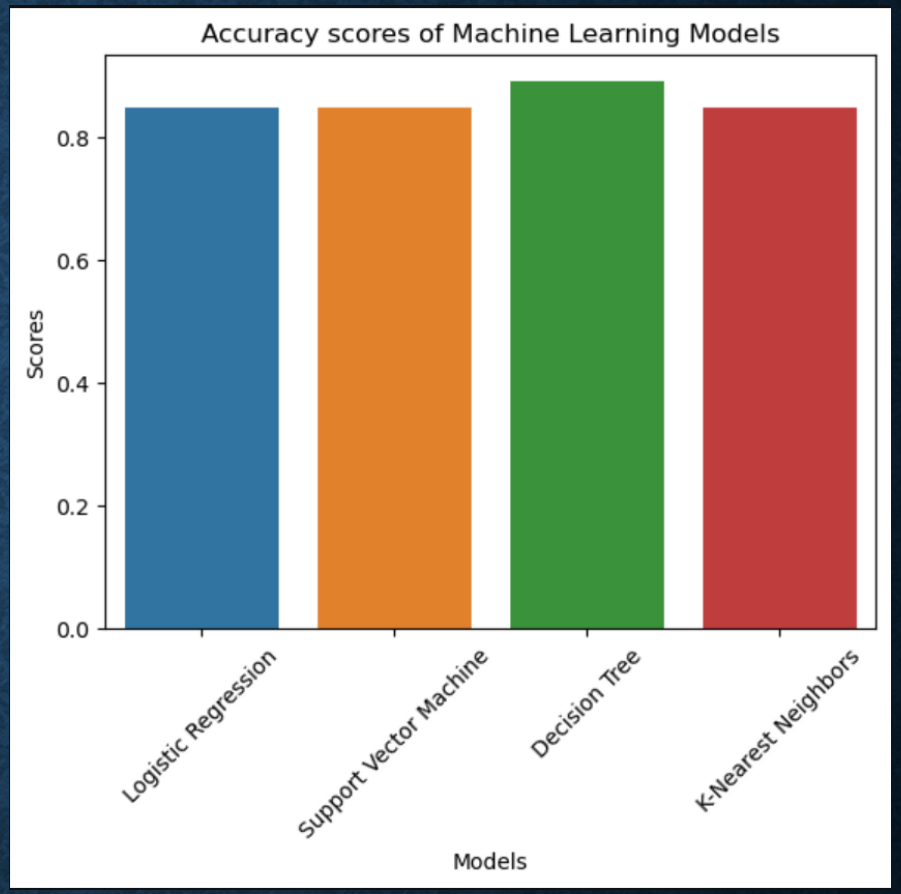
Outcome vs. Payload for All sites for Payload 1000-7000

Outcome vs. Payload for All sites for Payload 4000-9600

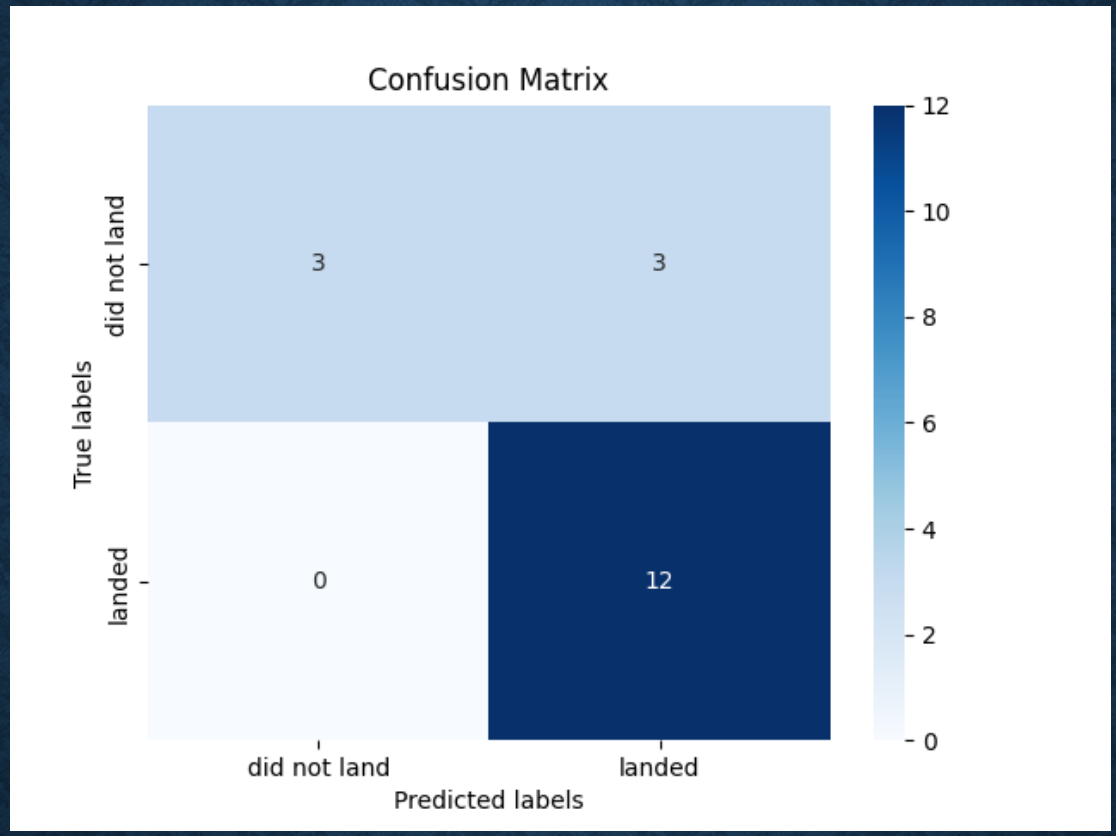
Section 5

Predictive Analysis (Classification)





Classification Accuracy



Confusion Matrix of Decision Tree

	precision	recall	f1-score	support
did not land	0.56	0.83	0.67	6
landed	0.89	0.67	0.76	12
accuracy			0.72	18
macro avg	0.72	0.75	0.71	18
weighted avg	0.78	0.72	0.73	18

Classification Report of Decision Tree

CONCLUSIONS

- Overall accuracy is 72%, indicating that it correctly predicts the outcome for 72% of launches.
- The performance on negative cases is moderate as the precision is 56%. It identifies 83% of not-landing cases as recall suggests. This is relatively high but causes a trade-off between precision.
- The performance on positive cases is relatively high. It predicts the landing returns with 89% precision and 67% recall meaning that it misses 33% of landings. This recall is not satisfying.
- *Overall:* The trade-off between recall and precision should be considered by business owners. Precision on negatives is not a real concern but bad recall on positives could be improved because it may damage monetary decisions.

APPENDIX

- Github Repo
- Some similar Studies: 1 2 3
- Code to predict the outcome of upcoming Falcon-9 Launch on May 18, 2024

RESOURCES

- SpaceX Web Page
- SpaceX API
- Falcon-9 Wiki
- Upcoming Falcon-9 Launches
- Some related news: 1 2 3
- SpaceX Stats

Thank you!