IBM Developer
SKILLS NETWORK

SpaceX Data Collection and Analysis

Winning Space Race with Data Science

Carlos Pamias Mora
24/09/2024
https://github.com/CarlosPamias/IBM_Coursera_Final

Presentation Contents

- Executive Summary.
- Introduction.
- Methodology.
- Results.
  - EDA with Visualization.
  - EDA with SQL.
  - Interactive Maps with Folium.
  - Plotly Dash Dashboard.
  - Predictive Analytics.
- Conclusion.

Executive Summary

- Summary of all results

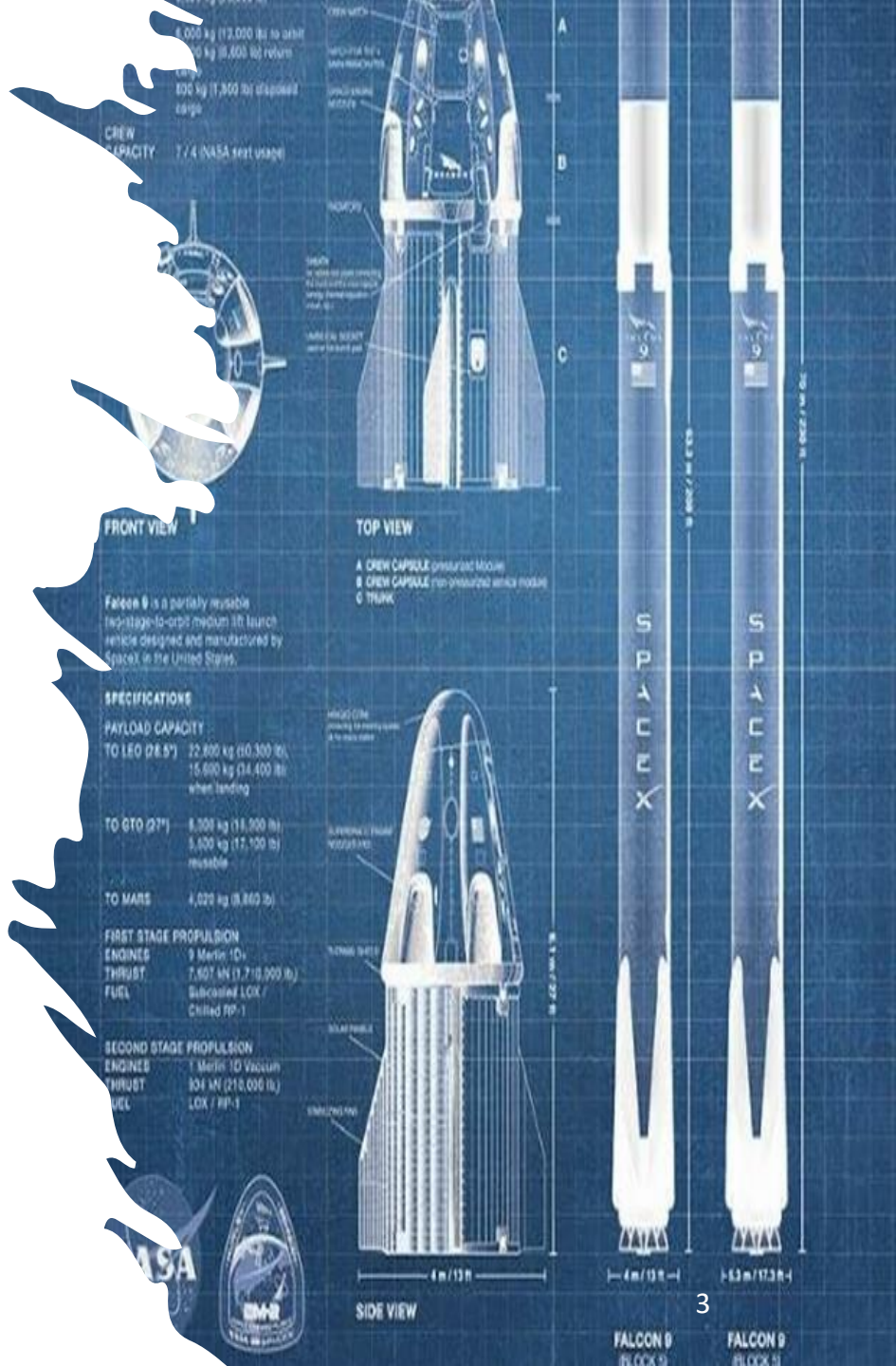
- **Collect**: Data using SpaceX REST API and web scraping techniques.
- **Explore**: Data using data visualization techniques, taking into account factors such as payload, launch site, flight number, and annual trend.
- **Analyze**: Data using SQL, statistically calculating total payload, payload range for successful launches, and total number of successful and failed outcomes.
- **Explore and visualize**: Launch site success rate and proximity to geographic markers, showing the most successful launch sites and successful payload ranges.
- **Build Models**: To be able to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN).

- Summary of all results

Exploratory Data Analysis:
Launch success has improved over time
KSC LC-39A has the highest success rate among landing sites.
Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.

Visualization/Analytics:
Most launch sites are near the equator, and all are close to the coast.

Predictive Analytics:
All models performed similarly on the test set.
The decision tree model slightly outperformed.



Introduction

• Project Description

The commercial space age is booming, with companies like SpaceX leading the way by reusing rockets, which reduces launch costs.
SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each.
This project analyzes SpaceX data, specifically Falcon 9 rocket launches, to determine if the first stage will be reused. This is key, as it allows SpaceX to offer launches at lower prices than other providers.

• Key objectives

- Determine the cost of each launch.
- Predict first stage reusability.
- Use machine learning to optimize the competitiveness of Space Y, a new rocket company.

• Expected results

Development of interactive panels to visualize data and training of predictive models to improve the sustainability and profitability of launches.

Methodology

Data Collection – SpaceX API

Steps using Python:

- **Request** data from SpaceX API (rocket launch data).
- **Decode** the response using . json() and convert it to a data frame using . json_normalize.
- **Request** launch information from SpaceX API using custom functions.
- **Create** a dictionary from the obtained data.
- **Create** a data frame from the dictionary.
- **Filter** the data frame to contain only Falcon 9 launches.
- **Replace** missing payload mass values with calculated .mean().
- **Export** data to a csv file.

View SpaceX API details

Data Collection - Web Scraping

Steps using Python:

- **Requesting** data (Falcon 9 launch data) from Wikipedia.
- **Creating** a BeautifulSoup object from an HTML response.
- **Extracting** column names from an HTML table header.
- **Collecting** data from HTML table parsing.
- **Creating** a dictionary from the data.
- **Creating** a data frame from the dictionary.
- **Exporting** data to a csv file.

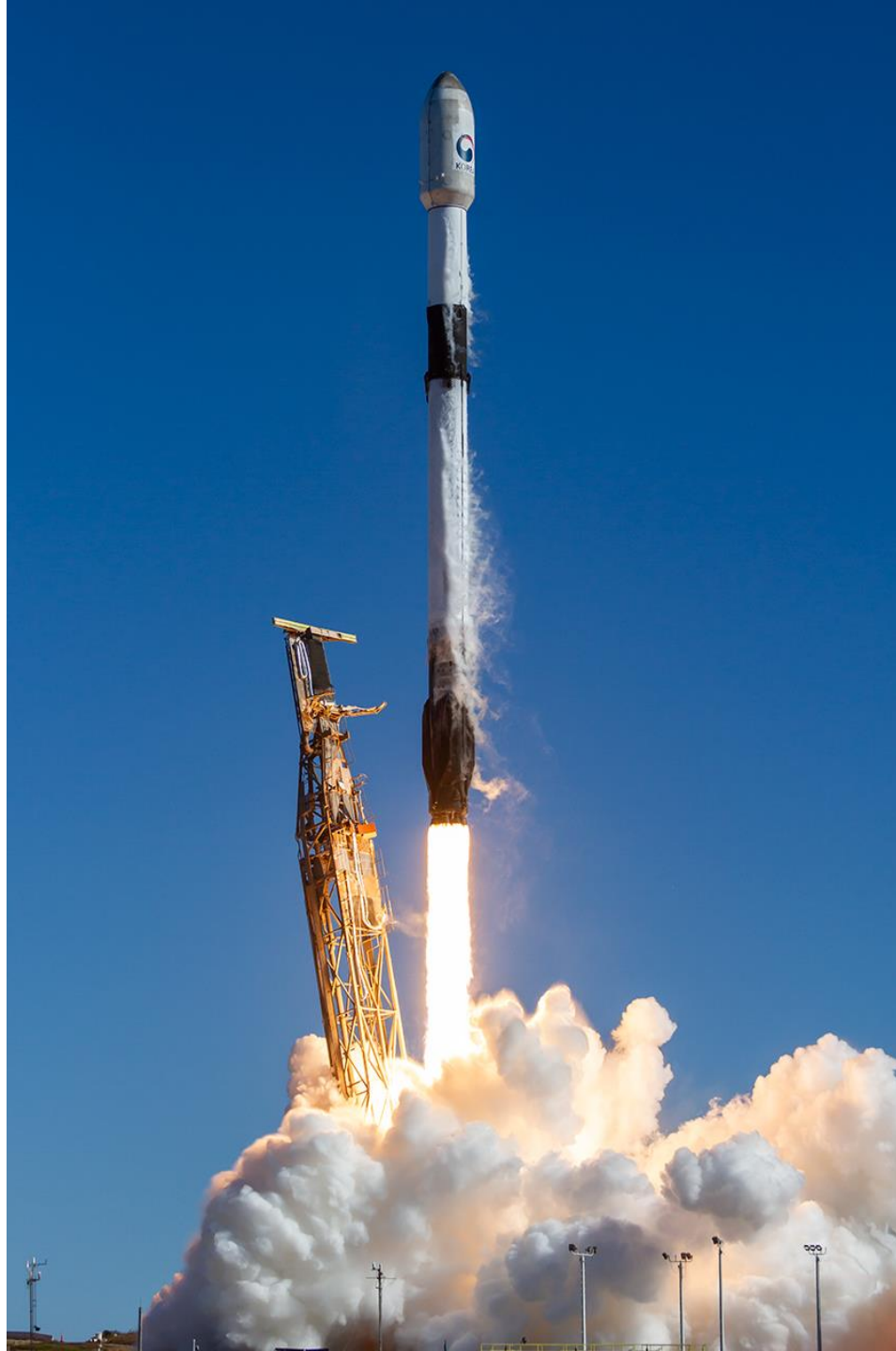
View Web Scraping details

Data Wrangling

Steps using Python:

- **Perform** EDA and determine data labelsCalculate:
  - Number of launches for each site.
  - Number and occurrence of orbits.
  - Number and occurrence of mission results by orbit type.
- **Create** a binary column of landing results.
- **Differentiate** landing results by location and outcome.
  - True Ocean.
  - False Ocean.
  - True RTLS.
  - False RTLS.
  - True ASDS.
  - False ASDS**.**
  - Outcomes converted.
- **Export** data to a csv file.

View Data Wrangling details



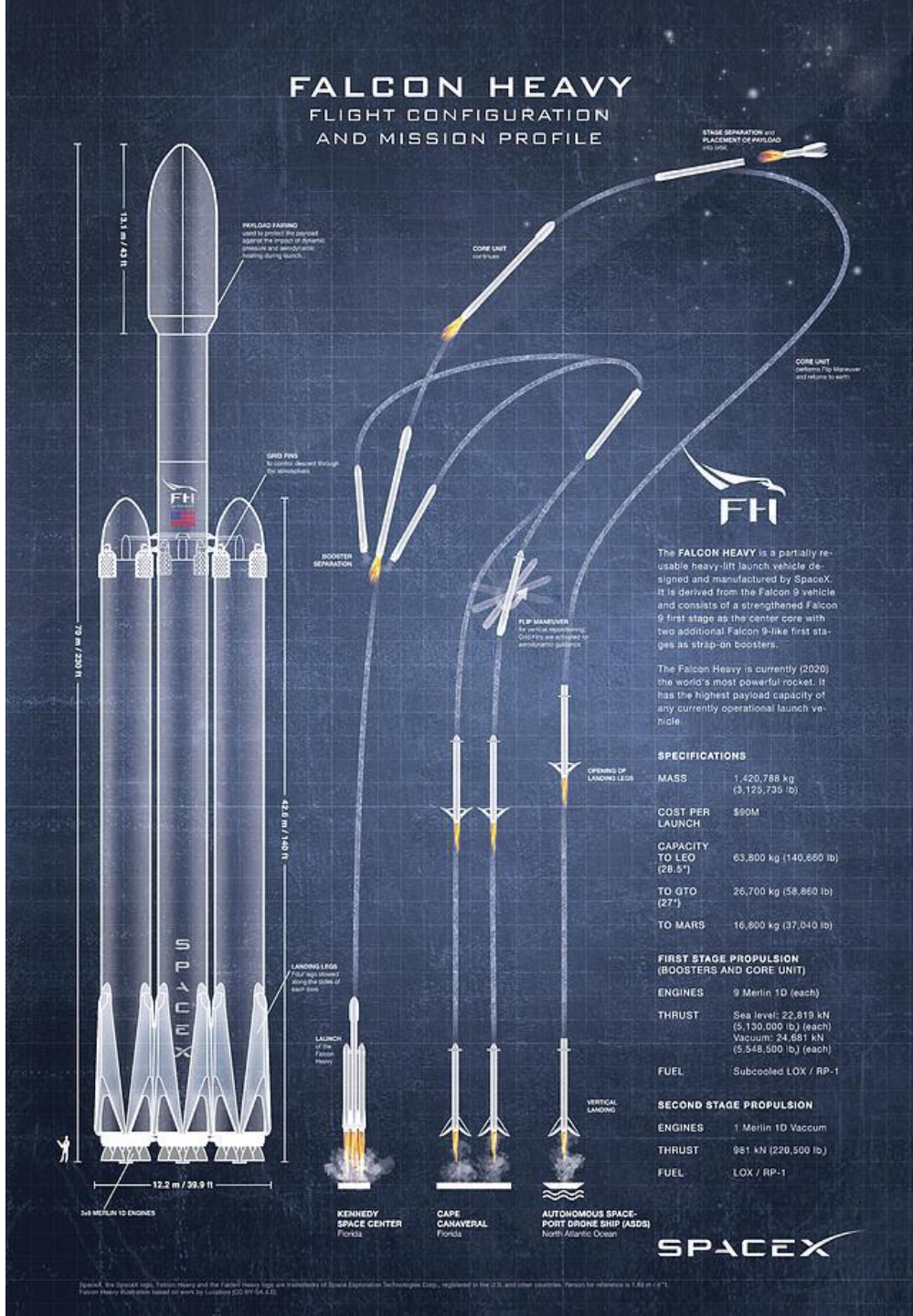
EDA with Data Visualization

Steps using Python:

- **Generated Charts:**
  - Flight number vs. payload.
  - Flight number vs. launch site.
  - Payload mass (kg) vs. launch site.
  - Payload mass (kg) vs. orbit type.

- **Analysis**
  - **View** the relationship using **scatter plots.** Variables could be useful for machine learning if a relationship exists.
  - **Show** comparisons between discrete categories with **bar charts.** Bar charts show relationships between categories and a measured value.

View [EDA with Data Visualization] details



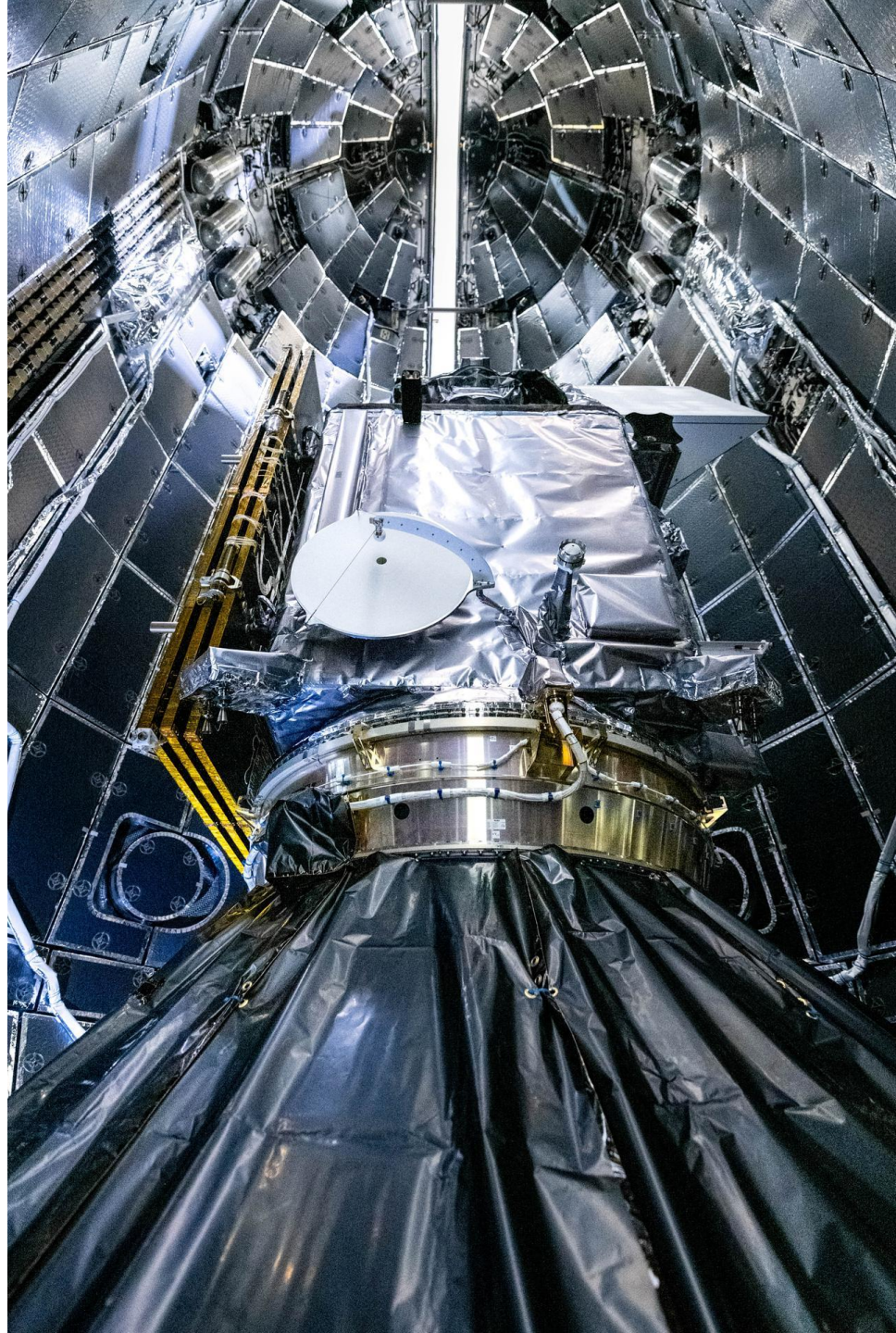
EDA with SQL

Queries

- **Display:**
  - Unique launch site names.
  - 5 records where launch site starts with 'CCA’.
  - Total payload mass carried by NASA-launched boosters (CRS).
  - Average payload mass carried by booster version F9 v1.1.

- **List:**
  - Date of first successful landing on ground platform.
  - Names of boosters that were successful in landing on an unmanned spacecraft and have a payload mass greater than 4.000 but less than 6.000.
  - Total number of successful and failed missions.
  - Names of booster versions that have carried the maximum payload Results of failed landings on an unmanned spacecraft, their booster version, and launch site during the months of the year 2015.
  - Count of landing results between 2010/06/04 and 2017/03/20 (desc).

View EDA with SQL details



Map with Folium

Markers Indicating Launch Sites

- Added **blue** circle at **NASA Johnson Space Center's coordinate** with a **popup label** showing its name using its latitude and longitude coordinates.
- Added **red** circles at **all launch sites coordinates** with a **popup label** showing its name using its name using its latitude and longitude coordinates.

Markers Indicating Launch Sites

- Added **colored markers** of **successful green** and **unsuccessful red** launches at each launch site to show which launch sites have high success rates.

Distances Between a Launch Site to Proximities

- Added **colored lines** to show distance between launch site **CCAFS SLC 40** and its proximity to the nearest coastline, railway, highway, and city.

View Map with Folium details

Dashboard with Plotly Dash

Dropdown List with Launch Sites.

- Allow user to select all launch sites or a certain launch site.

Pie Chart Showing Successful Launches.

- Allow user to see successful and unsuccessful launches as a percent of the total.

Slider of Payload Mass Range.

- Allow user to select payload mass range.

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version.

- Allow user to see the correlation between Payload and Launch Success.

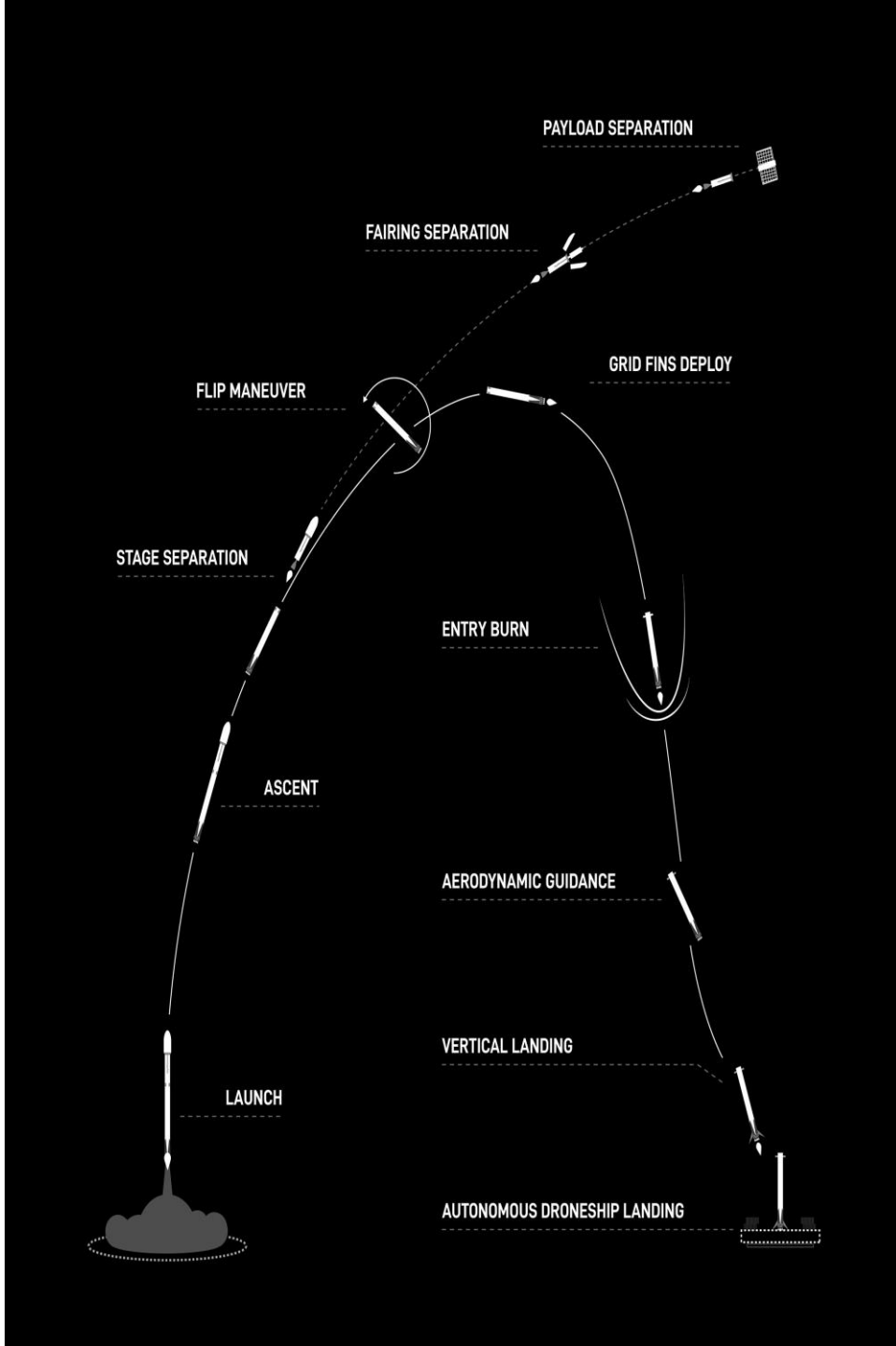
View Dashboard with Plotly Dash details

Predictive Analytics

Charts:

- **Create** a NumPy array from the Class column.
- **Standardize** the data with StandardScaler. Fit and transform the data.
- **Split** the data with train_test_split**.**
- **Create** a GridSearchCV object with cv=10 for parameter optimization.
- **Apply GridSearchCV in different algorithms:**
  - Logistic Regression (LogisticRegression()).
  - Support Vector Machine (SVC()).
  - Decision Tree (DecisionTreeClassifier()).
  - K Nearest Neighbor (KNeighborsClassifier()).
- **Calculate** the accuracy of the test data with .score() for all models.
- **Evaluate** the confusion matrix for all models.
- **Identify** the best model with Jaccard_Score, F1_Score and Accuracy.

View Predictive Analytics details



Results Summary

Exploratory Data Analysis

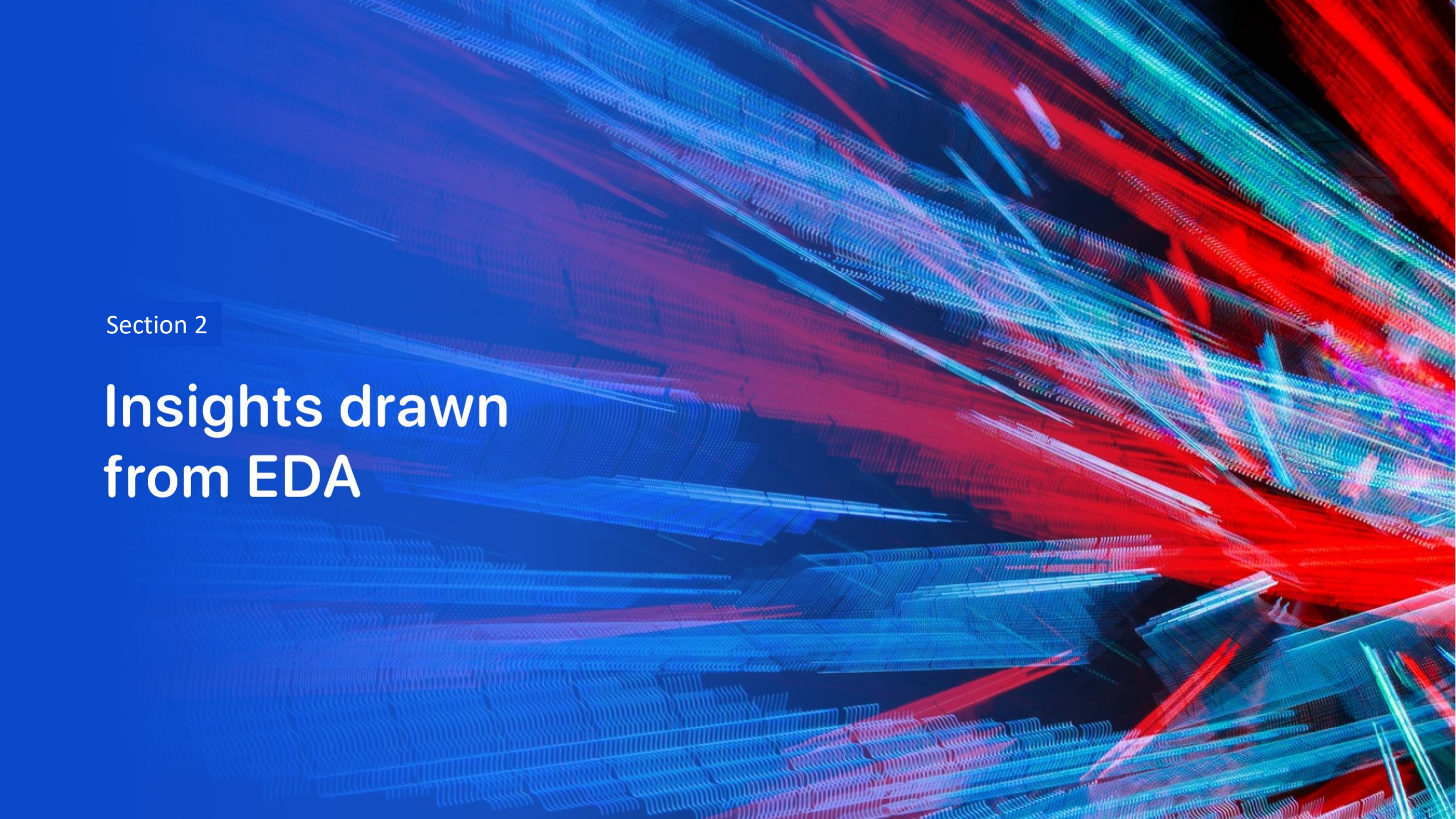
- Launch success has improved over time.
- KSC LC39A has the highest success rate among landing sites.
- ES-L1, GEO, HEO, and SSO orbits have a 100% success rate Summary of Results.

Visual Analysis

- Most launch sites are near the equator, and all are near the coast.
- Launch sites are far enough away from anything that a failed launch could damage (city, highway, railroad), but close enough to carry people and material to support launch activities.

Predictive Analysis

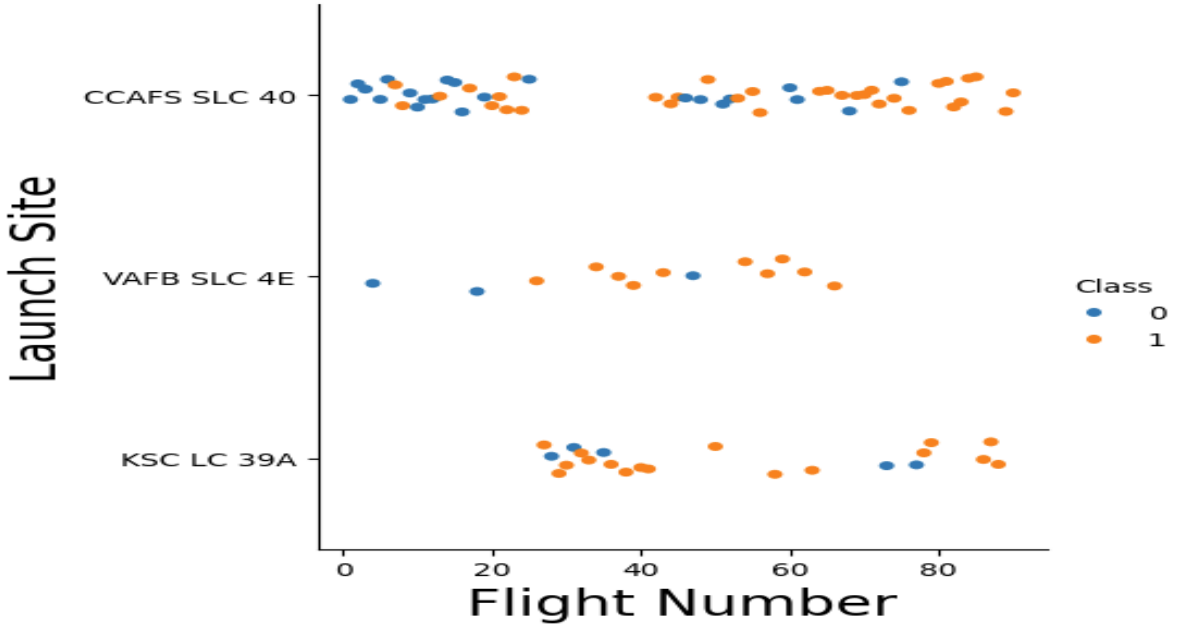
- Decision tree model is the best predictive model for the data set.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

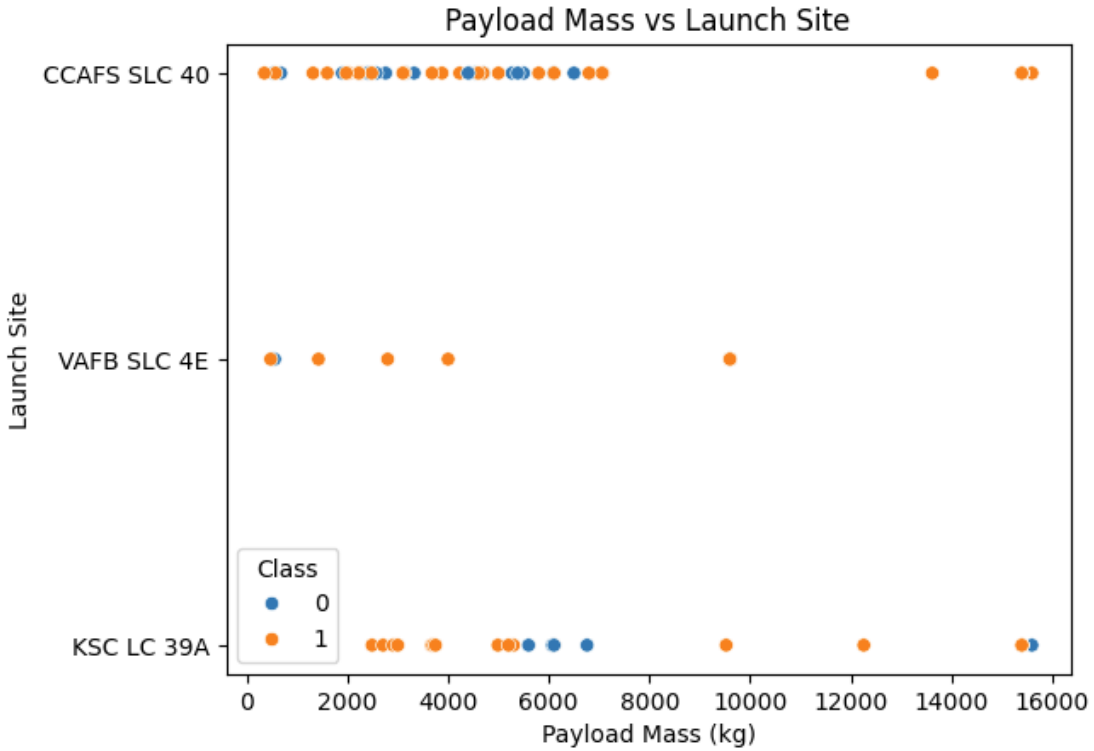
- Earlier flights had a lower success rate **blue = failure.**
- Later flights had a higher success rate **orange = success.**
- About half of the launches were conducted from CCAFS's SLC 40 launch site.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- We can infer that newer launches have a higher success rate.



Payload vs. Launch Site

Exploratory Data Analysis

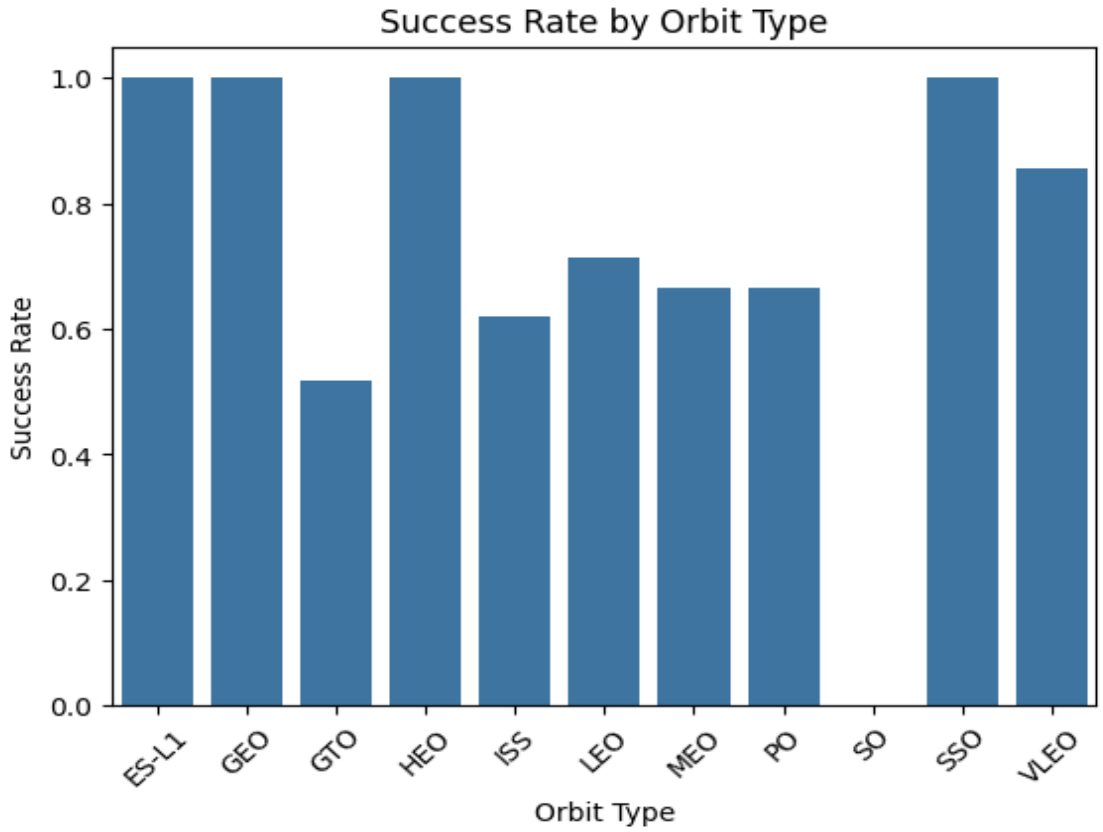
- Generally, the **higher** the **payload mass** (kg), the higher the **success rate.**
- Most launches with a payload over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for launches under 5500 kg.
- VAFB SKC 4E has not launched anything weighing more than ~10,000 kg.



Success Rate by Orbit

Exploratory Data Analysis

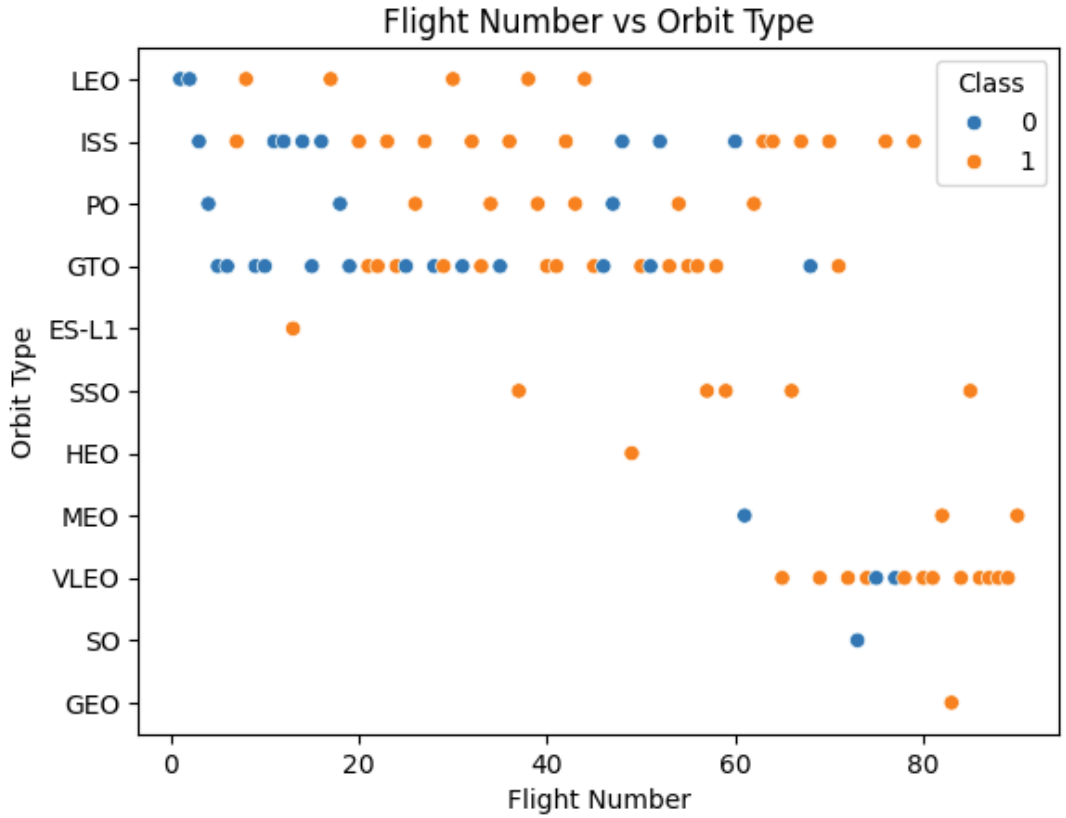
- 100% success rate: ES L1, GEO, HEO, SSO.
- 50% to 85% success rate: GTO, ISS, LEO, MEO, PO, WLEO.
- 0% success rate: SO.



Flight Number vs. Orbit

Exploratory Data Analysis

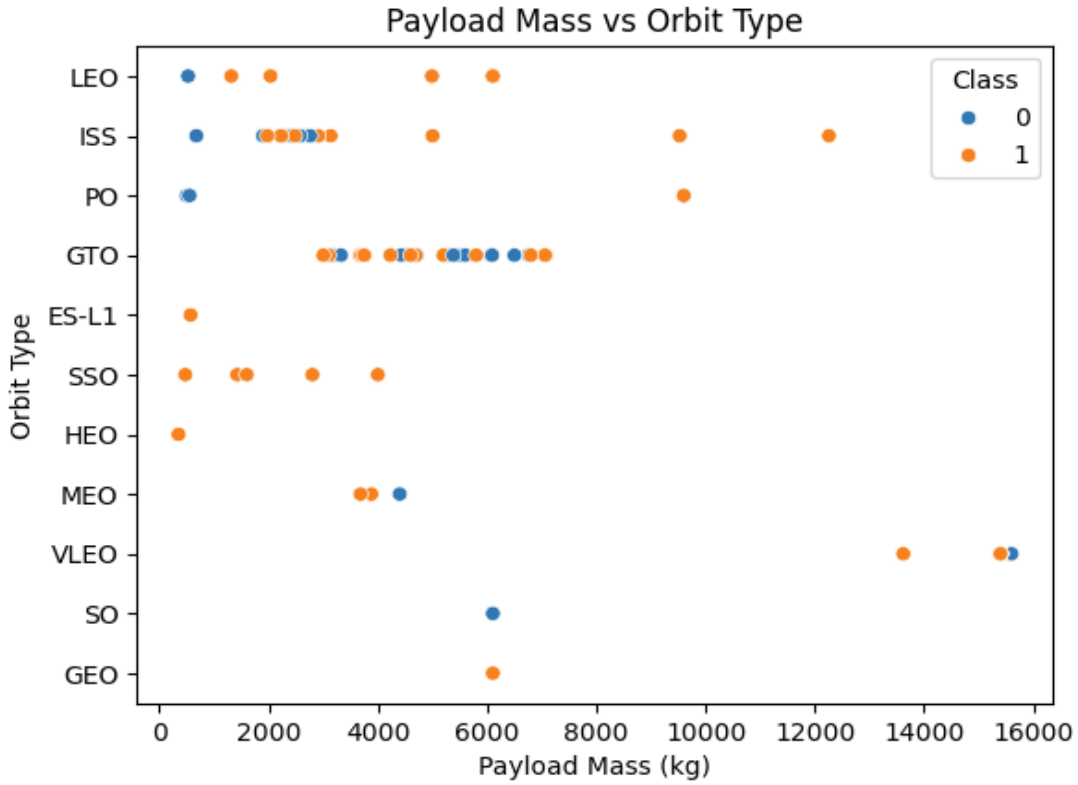
- The success rate typically increases with the number of flights for each orbit.
- This relationship is very evident in the case of the LEO orbit.
- However, the GTO orbit does not follow this trend.



Payload vs. Orbit

Exploratory Data Analysis

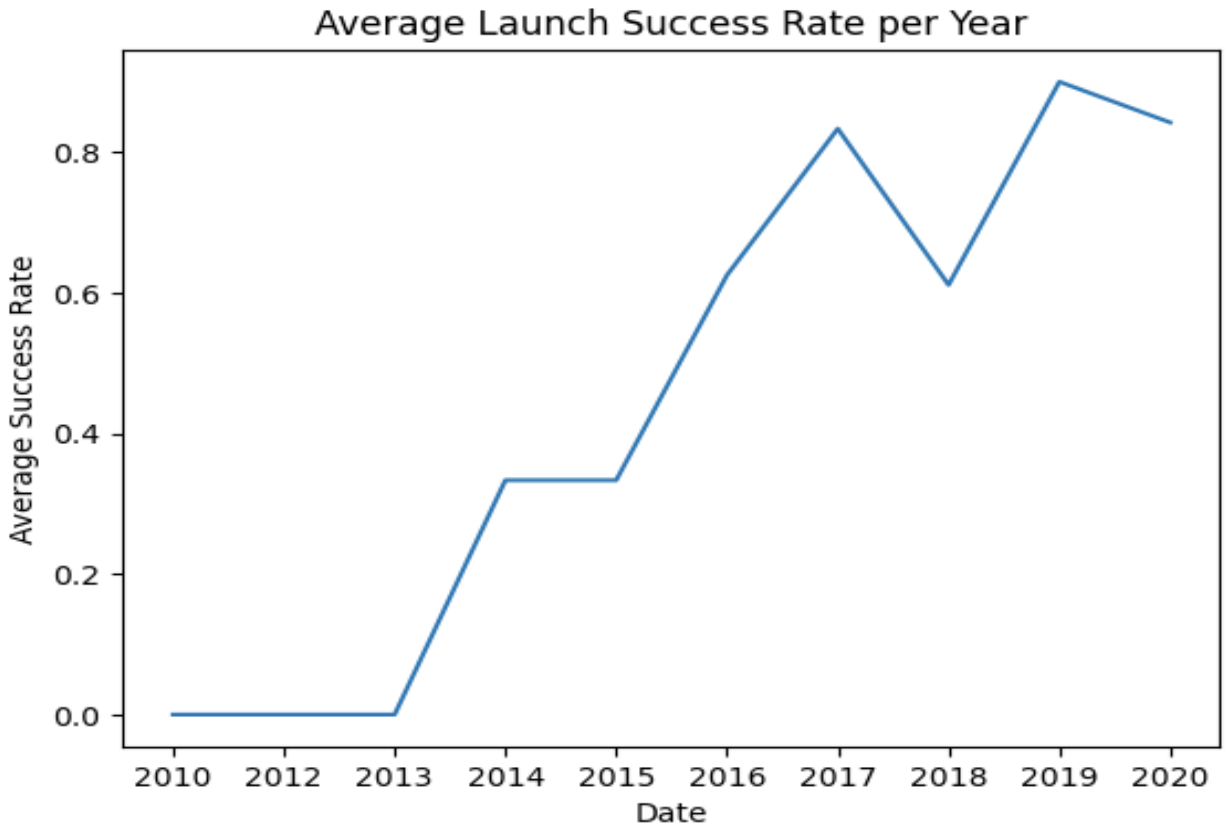
- Heavy payloads are better in LEO, ISS and PO orbits.
- GTO orbit has mixed success with heavier payloads.

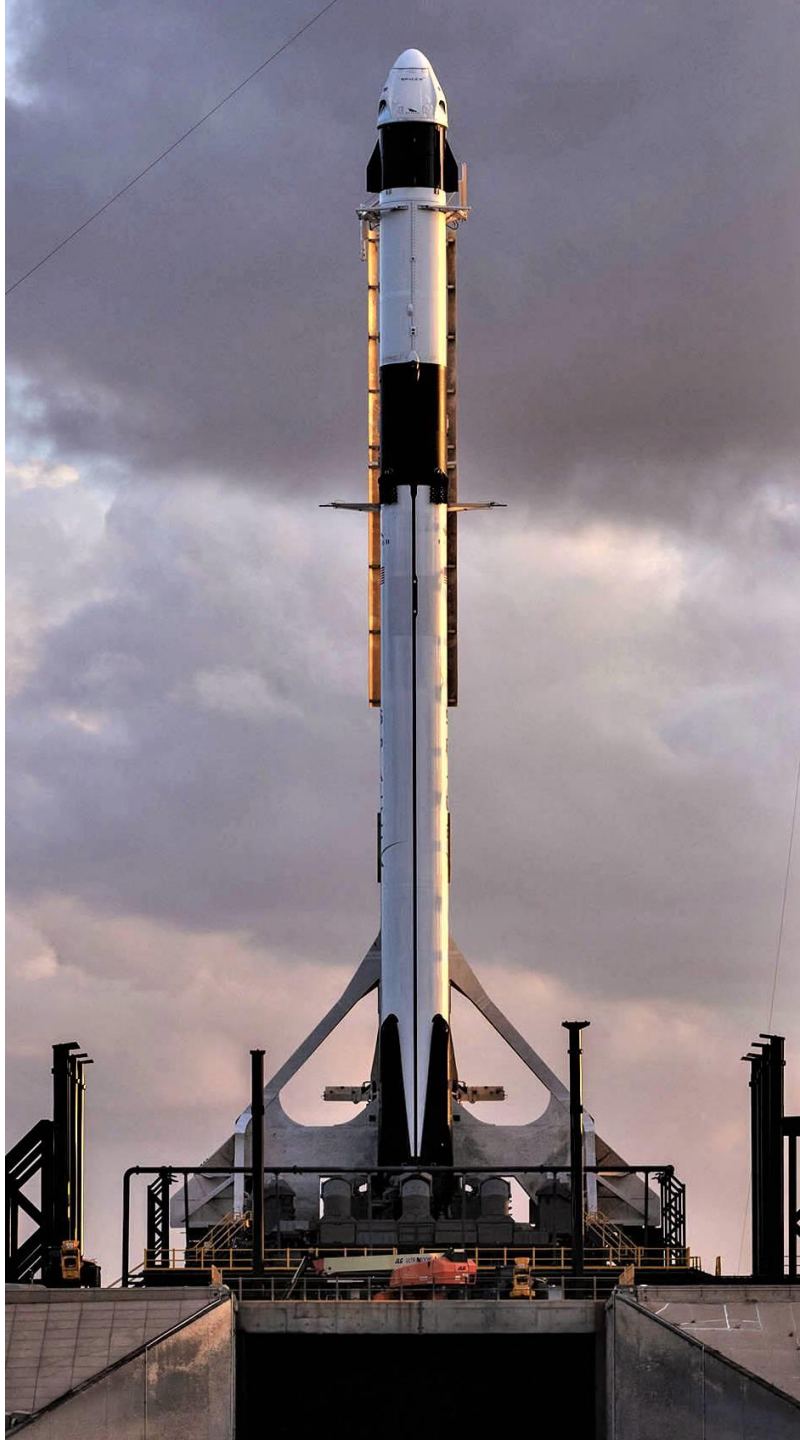


Launch Success over Time

Exploratory Data Analysis

- The success rate improved between 2013 - 2017 and between 2018 – 2019.
- The success rate decreased between 2017 - 2018 and between 2019 – 2020.
- Overall, the success rate has improved since 2013.





Launch Site Information

Launch site names

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

Landing Outcome Cont.

In [10]:
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

Out[10]:

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

Records with Launch Site Starting with CCA.

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

[11]:
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

[11]:

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

Payload Mass

Total Payload Mass

45,596 kg (total) carried by rockets launched by NASA (CRS).

```
[11]: %%sql SELECT SUM("PAYLOAD_MASS__KG_") AS total_payload_mass
          FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

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

[11]:

total_payload_mass
45596

Average Payload Mass

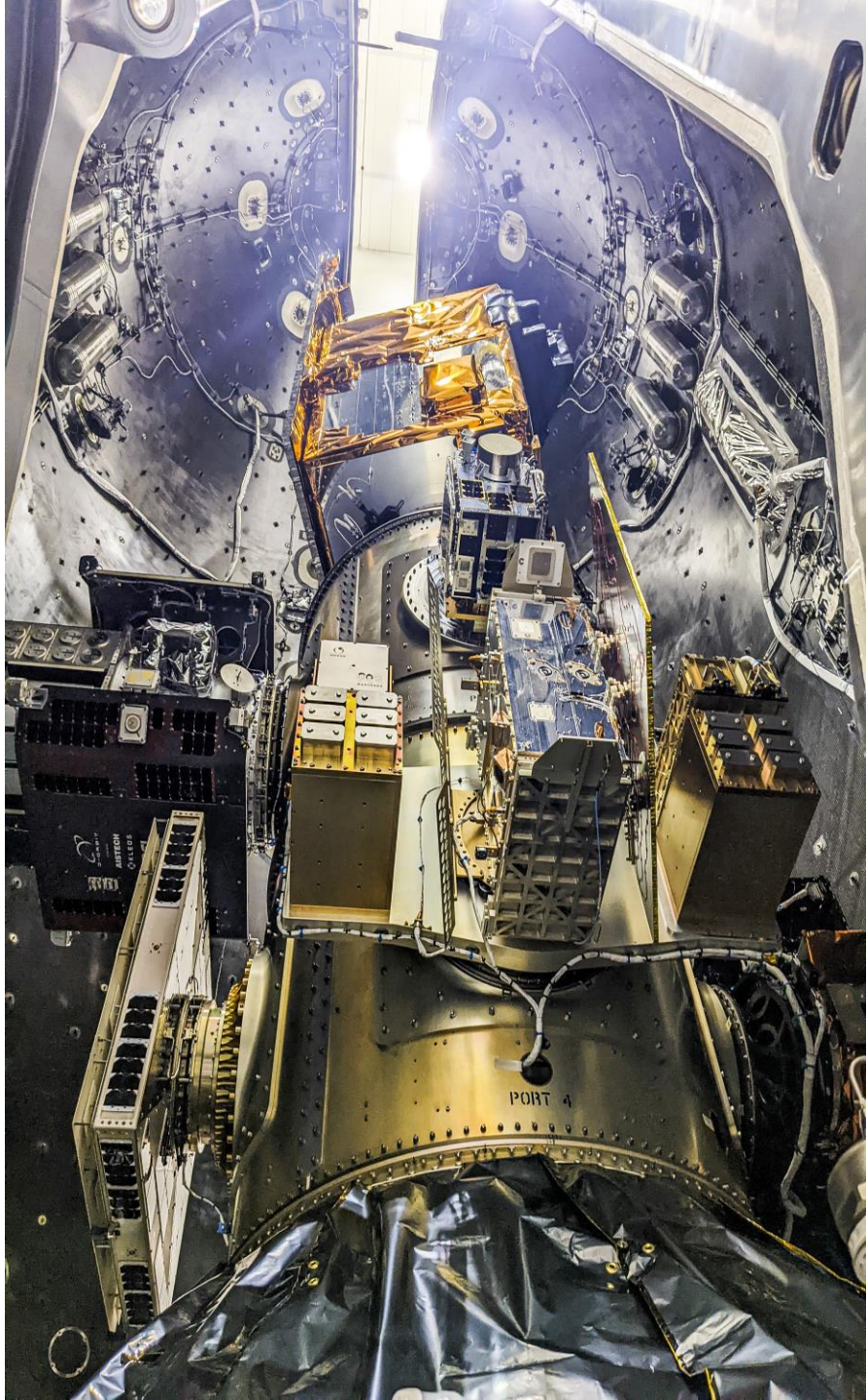
2,928 kg (average) carried by the F9 v1.1 rocket version.

```
[12]: %%sql SELECT AVG("PAYLOAD_MASS__KG_") AS average_payload_mass
          FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

[12]:

average_payload_mass
2928.4



Landing & Mission Info

1st Successful Landing in Ground Pad

12/22/2015

```
[13]: %%sql SELECT MIN("Date") AS first_successful_landing
          FROM SPACEXTABLE
          WHERE "Landing_Outcome" = 'Success (ground pad)';

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

[13]:

first_successful_landing
2015-12-22

Booster Drone Ship Landing

- Booster rocket mass greater than 4000 but less than 6000.
- JSCAT-14, JSCAT-16, SES-10, SES-1 / EchoStar 105.

```
[16]: %%sql SELECT "Payload" FROM SPACEXTABLE
          WHERE "Landing_Outcome" = 'Success (drone ship)'
          AND "PAYLOAD_MASS__KG_"
          BETWEEN 4000 AND 6000

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

[16]:

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failed Mission Outcomes

1 In-flight failure

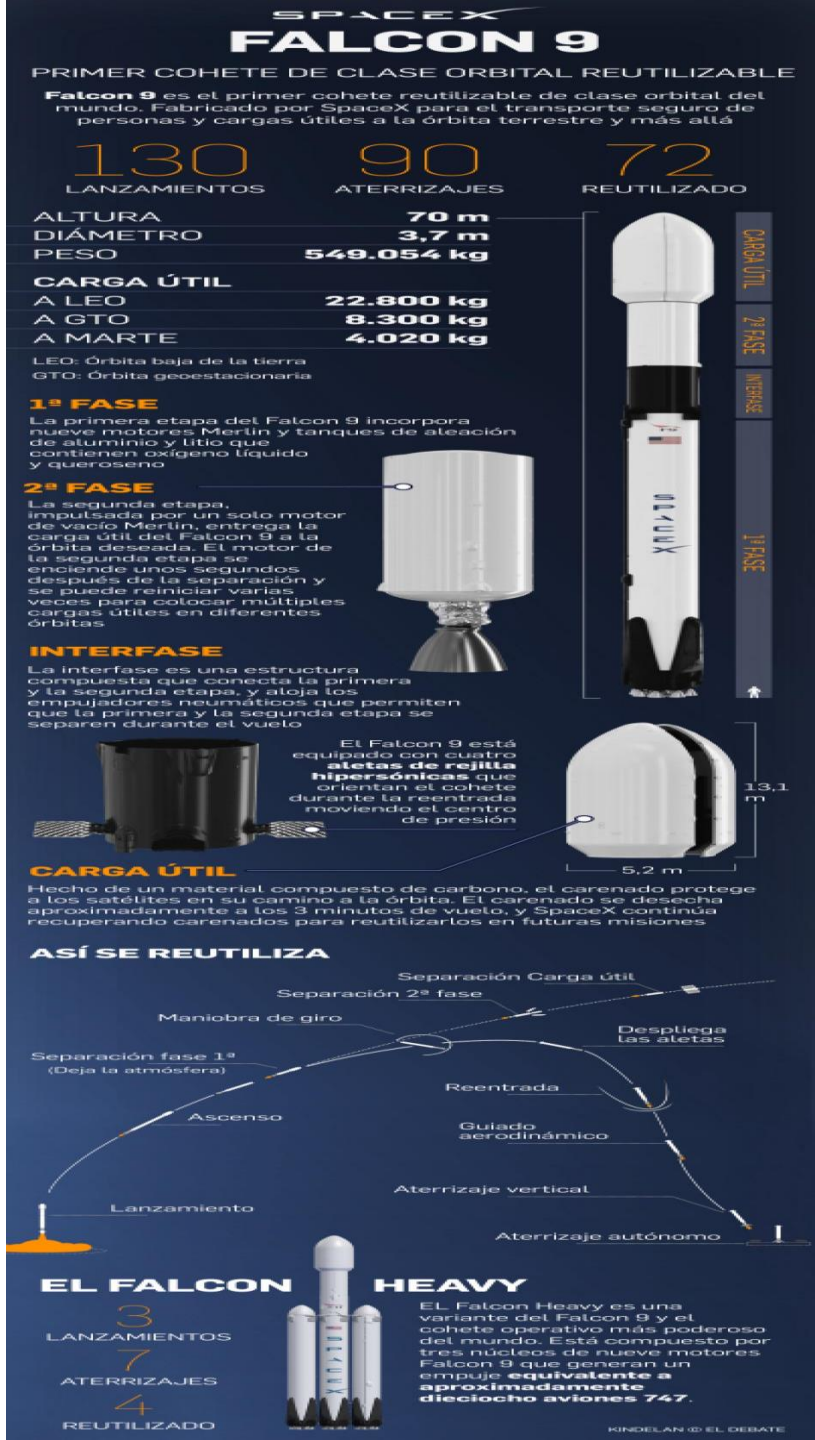
99 Success

1 Success (payload status unclear)

```
[17]: %%sql SELECT "Mission_Outcome", COUNT(*) AS total
          FROM SPACEXTABLE GROUP BY "Mission_Outcome";

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

[17]:

Mission_Outcome	total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1



Boosters

Carrying Max Payload

- 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

```
[18]: %%sql SELECT distinct "Booster_Version"
          FROM SPACEXTABLE
          WHERE "PAYLOAD_MASS__KG_" =
          (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
       * sqlite:///my_data1.db
      Done.
```

[18]:

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

Failed Landings on Drone Ship

In 2015

- Showing month, date, booster version, launch site and landing outcome.

[18]:
```
%%sql SELECT
    CASE substr("Date", 6, 2)
        WHEN '01' THEN 'Enero'
        WHEN '02' THEN 'Febrero'
        WHEN '03' THEN 'Marzo'
        WHEN '04' THEN 'Abril'
        WHEN '05' THEN 'Mayo'
        WHEN '06' THEN 'Junio'
        WHEN '07' THEN 'Julio'
        WHEN '08' THEN 'Agosto'
        WHEN '09' THEN 'Septiembre'
        WHEN '10' THEN 'Octubre'
        WHEN '11' THEN 'Noviembre'
        WHEN '12' THEN 'Diciembre'
    END AS month_name,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Failure (drone ship)'
AND substr("Date", 0, 5) = '2015'
```

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

[18]:

month_name	Landing_Outcome	Booster_Version	Launch_Site
Enero	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Abril	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Count of Successful Landings

Ranked Descending

- Count of landing outcomes between 2010 06 04 and 2017 03 20 in descending order.

[19]:
```
%%sql SELECT "Landing_Outcome", COUNT(*) AS outcome_count
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY outcome_count DESC;
```

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

[19]:

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

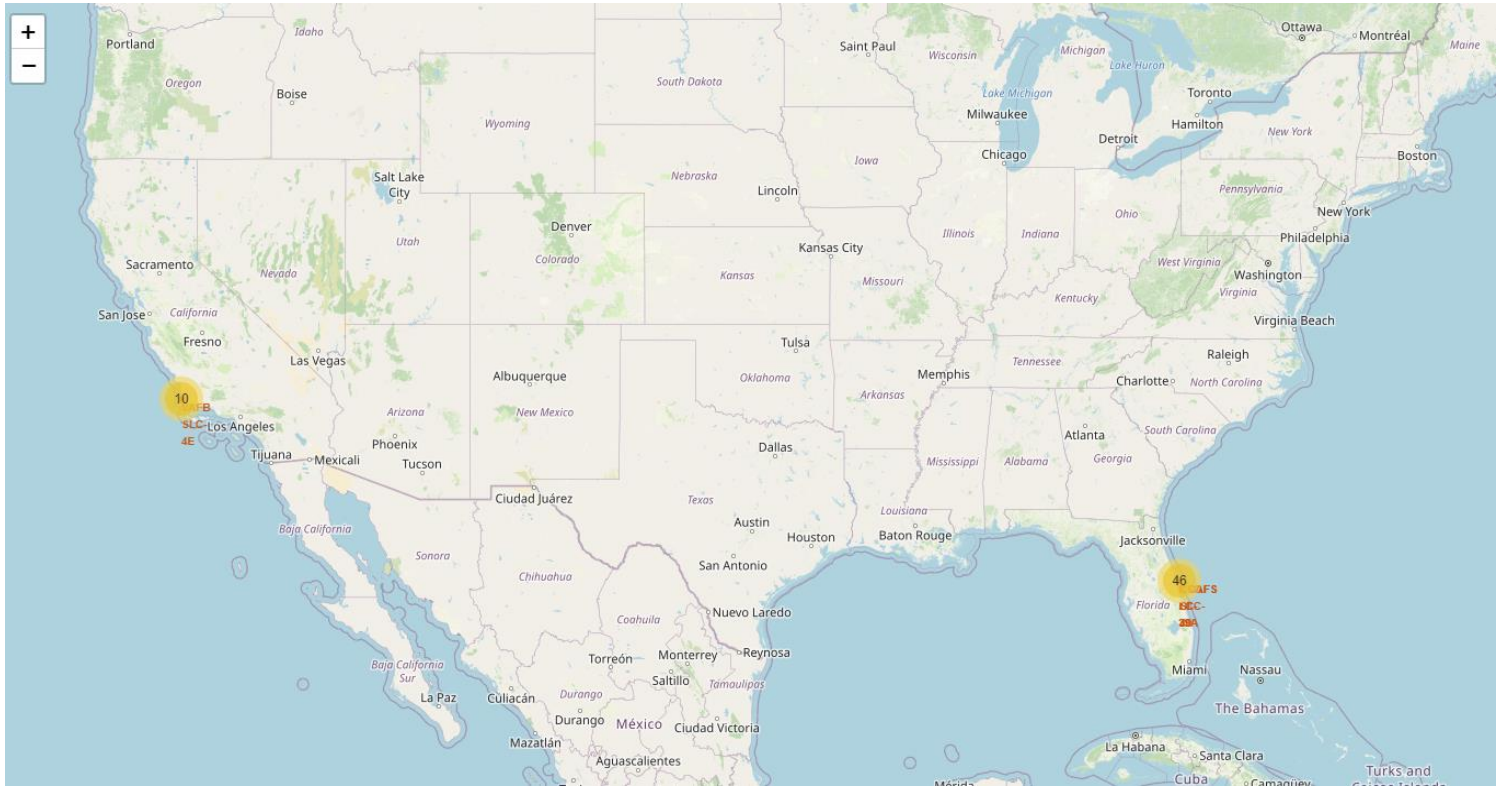
Section 3

Launch Sites Proximities Analysis

Launch Sites

With Markers

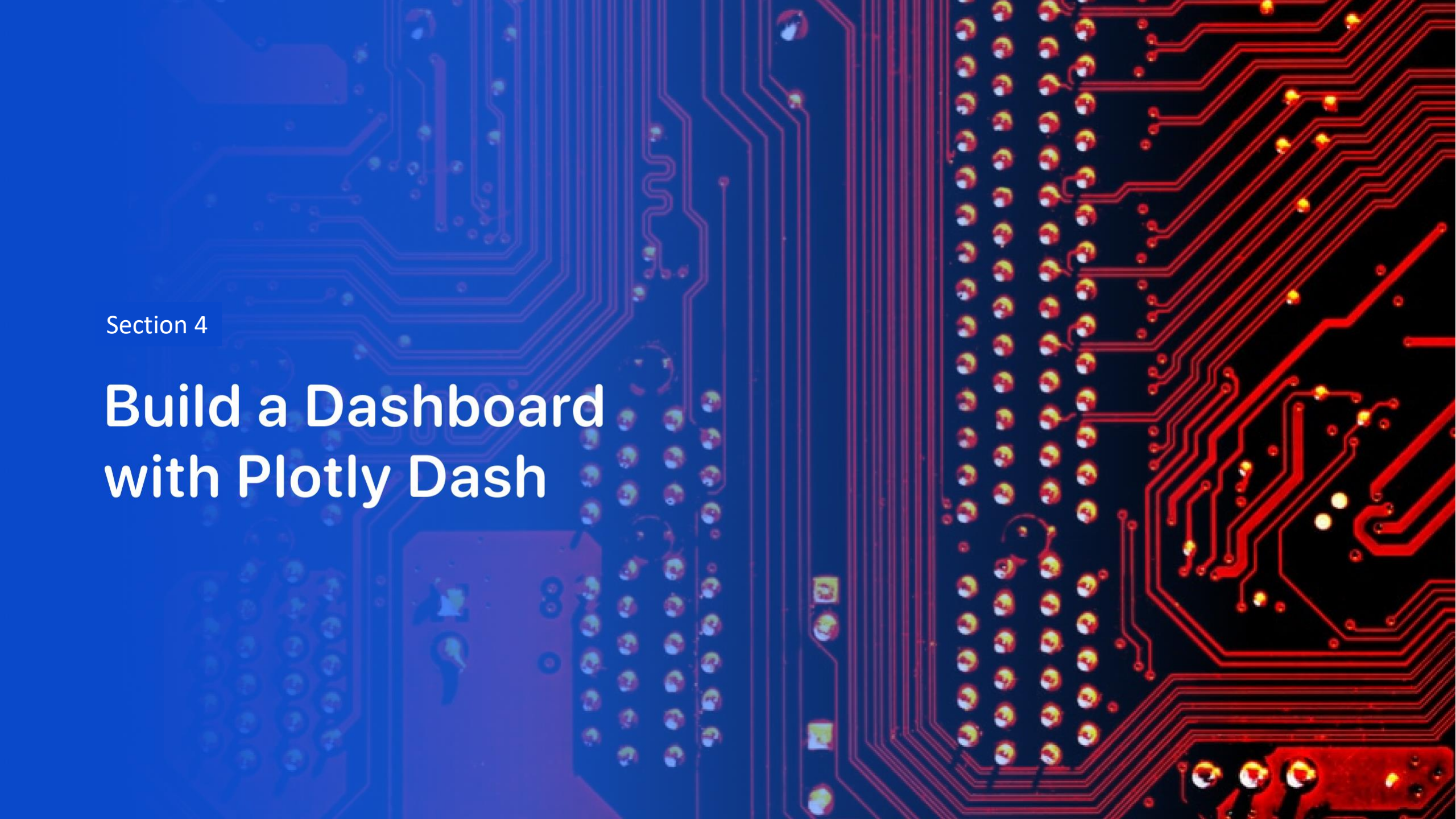
- **Near Equator**: the closer the launch site to the equator, the **easier** it is **to launch** to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an **additional natural boost** - due to the rotational speed of earth - that **helps save the cost** of putting in extra fuel and boosters.



Launch Outcomes

At Each Launch Site

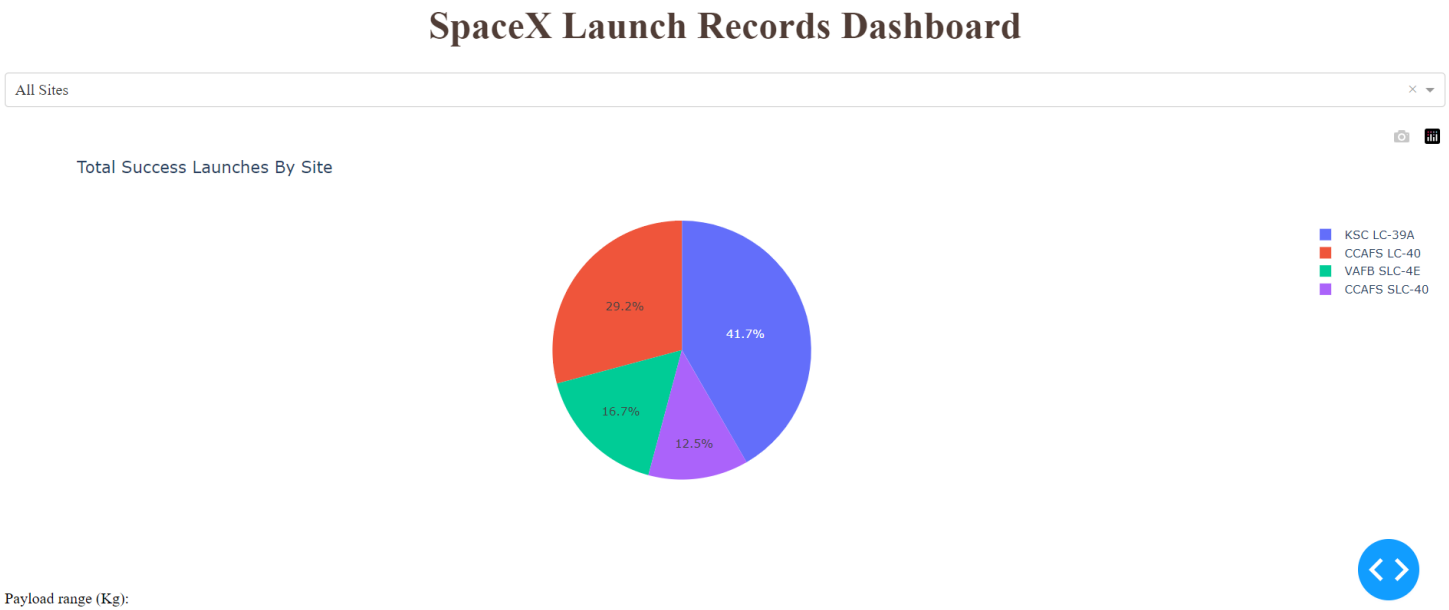
- Outcomes
- **Green** markers for successful launches.
- **Red** markers for unsuccessful launches.
- Launch site **CCAFS SLC 40** has a 3/7 success rate 42.9%).


Section 4
Build a Dashboard with Plotly Dash

Launch Success by Site

Success as Percent of Total

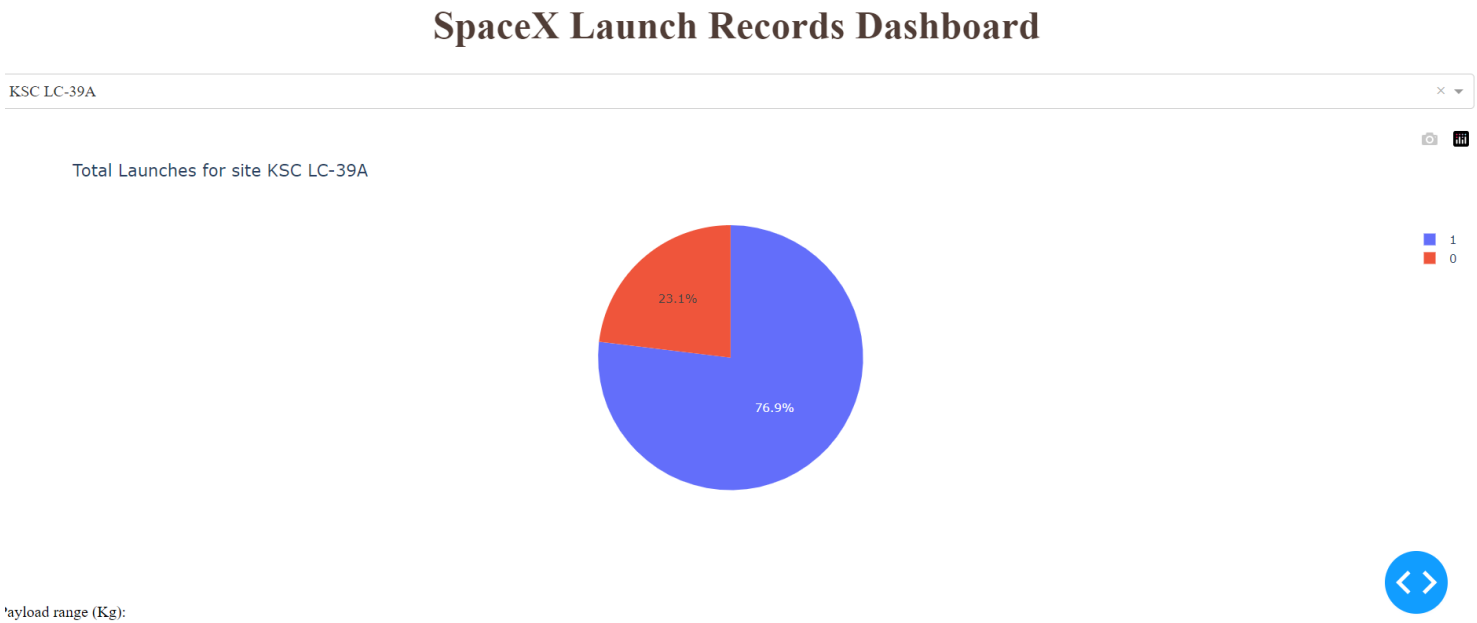
- **KSC LC-39A** has the **most successful launches** amongst launch sites 41.2%.



Launch Success (KSC LC-29A)

Success as Percent of Total

- **KSC LC-39A** has the **highest success** rate amongst launch sites (76.9%).
- 10 successful launches and 3 failed launches.



Payload Mass and Success

By Booster Version

- **Payloads between 2,000 kg** and **5,000 kg** have the **highest success rate.**
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome.

'ayload range (Kg):

0 100

All sites - payload mass between 100kg and 10,000kg

Section 5

Predictive Analysis (Classification)

Classification

Accuracy

- All models performed similarly and had the same scores and accuracy. This is likely due to the small dataset. The decision tree model slightly outperformed the rest when tested. .best_.
- .best_score_ is the average of all CV folds for a single combination of the parameters.

[33]:

	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	0.777778	0.833333

[34]:

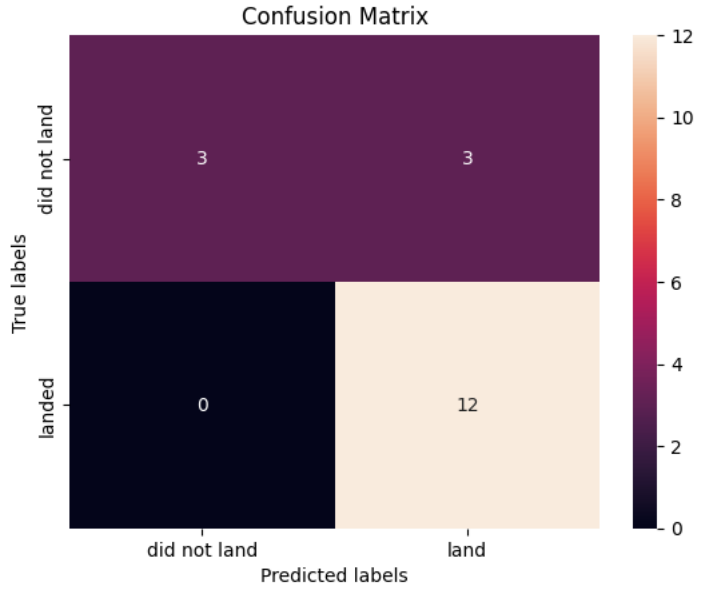
```
models = {'KNeighbors':knn_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'LogisticRegression':logreg_cv.best_score_,
          'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8607142857142855
Best params is : {'criterion': 'entropy', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrices

Performance Summary

- A confusion matrix summarizes the performance of a classification algorithm.
- All confusion matrices were identical.
- The fact that there are false positives (type 1 error) is not good.
- Confusion matrix results:
  - 12 true positives.
  - 3 true negatives.
  - 3 **false positives.**
  - 0 false negatives.



- **Precision** = TP / (TP + FP).

  12 / 15 = .80

- **Recall** = TP / (TP + FN).

  12 / 12 = 1

- **F1 score** = 2 * (Precision * Recall) / (Precision + Recall).

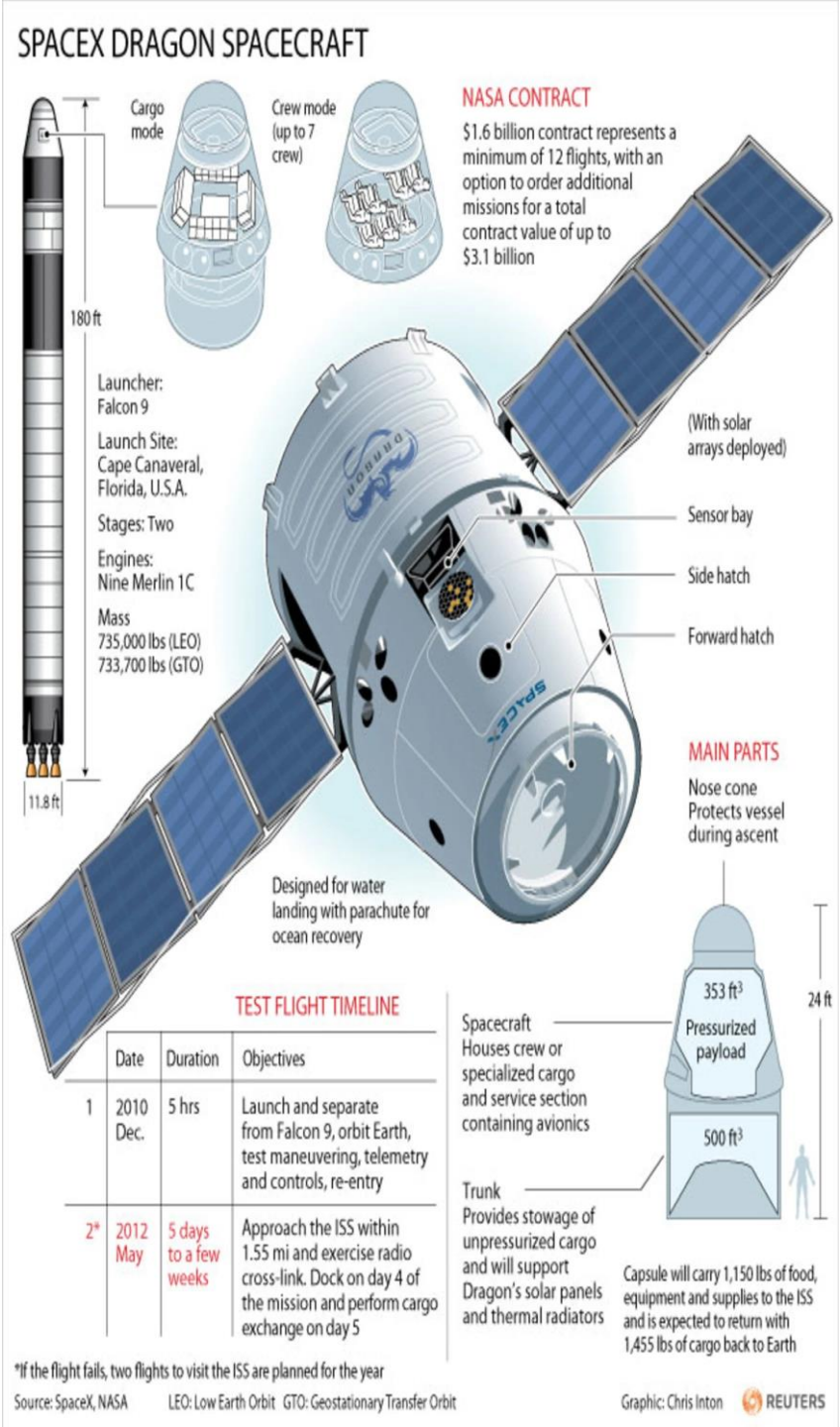
  2 * (.8 * 1) / (.8 + 1) = .89

- **Accuracy** = (TP + TN) / (TP + TN + FP + FN) = .833.

Conclusion

Research

- **Model Performance**: Models performed similarly across the test set, with the decision tree model slightly superior.
- **Equator**: Most launch sites are near the equator for an extra natural boost due to the Earth's rotation rate, which helps save the cost of putting in extra fuel and propellants.
- **Coast**: All launch sites are near the coast.
- **Launch Success**: Increases over time.
- **KSC LC-39A**: Has the highest success rate among launch sites. It has a 100% success rate for launches under 5500 kg.
- **Orbits:** ES-L1, GEO, HEO, and SSO have a 100% success rate.
- **Payload Mass**: At all launch sites, the higher the payload mass (kg), the higher the success rate.



Conclusion

Things to consider

- **Dataset**: A larger dataset will help develop the predictive analytics results to understand if the findings can be generalized to a larger dataset.
- **Feature Analysis/PCA**: Additional feature analysis or principal component analysis should be performed to see if it can help improve accuracy.
- **XGBoost**: This is a powerful model that was not used in this study. It would be interesting to see if it outperforms the other classification models.

Thank you!