

The SpaceX logo is displayed prominently in the center of the slide. It consists of the word "SPACEX" in a bold, white, sans-serif font. A thick, light blue curved swoosh starts from the top right of the letter "P" and sweeps across the letters "A", "C", and "E", ending near the top of the letter "X".

SPACEX

Troy Amegashie

8/22/2024

Outline

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

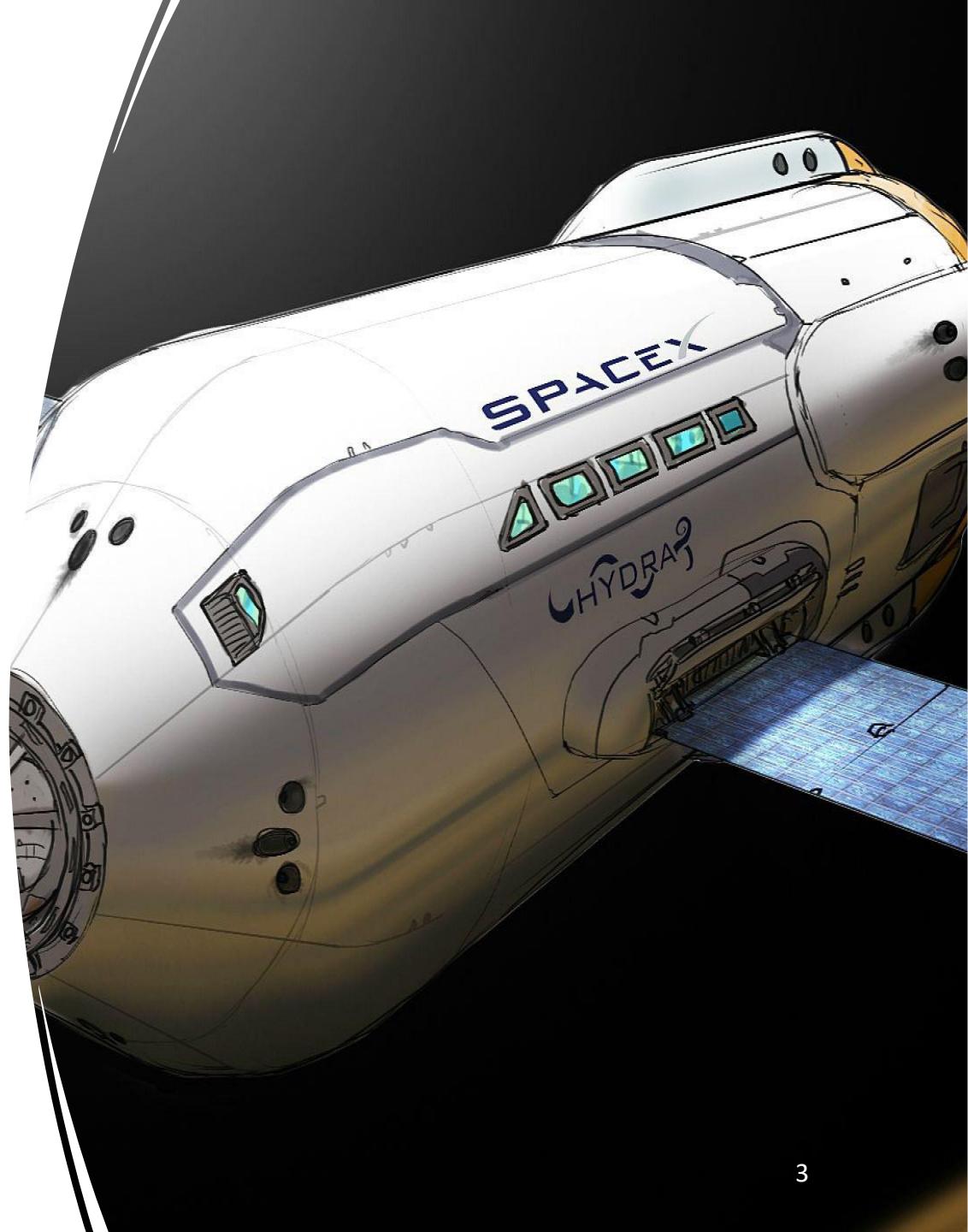


Executive Summary

Summary of methodologies

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies were used:

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data to create success/fail outcome variable
- **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- **Analyze** the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the launch sites with the most success and successful payload ranges
- **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)



Executive Summary

Results

Exploratory Data Analysis:

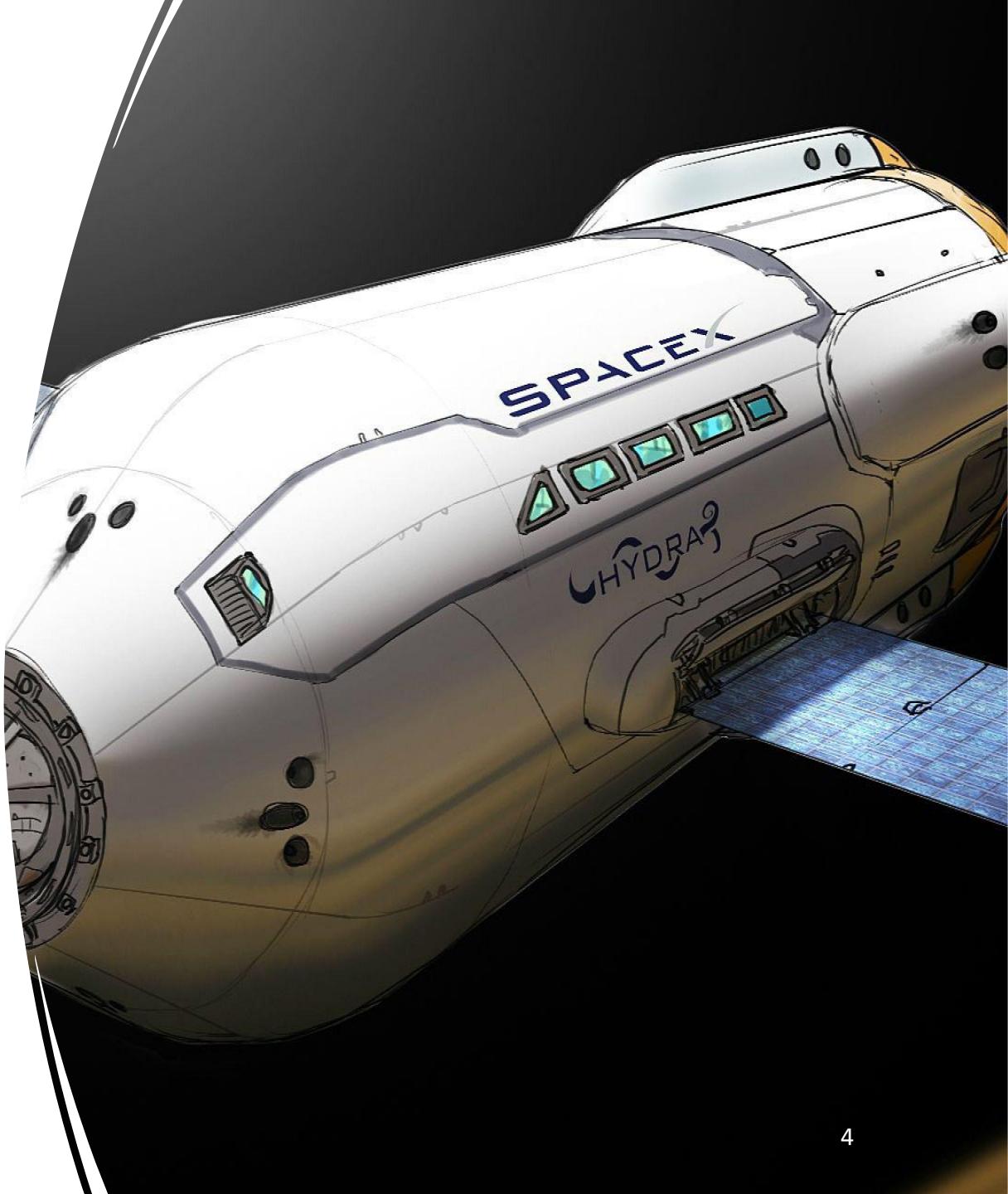
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbit types ES-L1, GEO, HEO, and SSO have a 100% success rate

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

Background

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. 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. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Explore

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
 - Best predictive model for successful landing (binary classification)



Methodology

Methodology

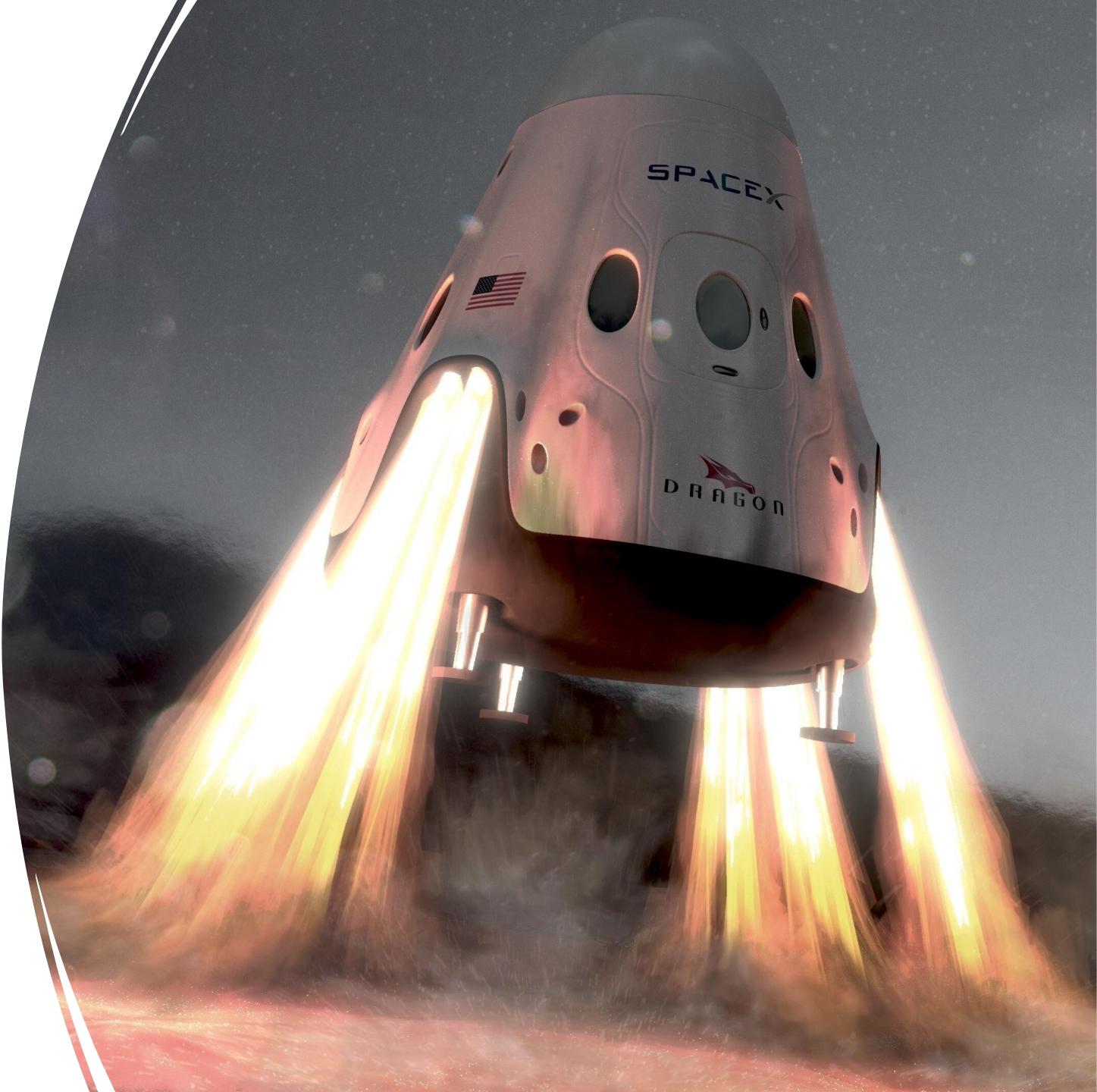
Summary

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data – by filtering the data, handling missing values and applying
 - one hot encoding – to prepare the data for analysis and modeling
- **Explore** data via EDA with SQL and data visualization techniques
- **Visualize** the data using Folium and Plotly Dash
- **Build Models** to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters



Data Collection - API

- **Request data** from SpaceX API(rocketlaunchdata)
- **Decode response** using .json() and convert to a dataframe using .json_normalize()
- **Request information** about the launches from SpaceX API using custom functions
- **Create dictionary** from the data
- **Create dataframe** from the dictionary
- **Filter dataframe** to contain only Falcon9 launches
- **Replace missing values** of Payload Mass with calculated .mean()
- **Export data** to csv file



Data Collection

- Web Scraping

- **Request data**(Falcon9 launch data) from Wikipedia
- **Create BeautifulSoup object** from HTML response
- **Extract column names** from HTML table header
- **Collect data** from parsing HTML tables
- **Create dictionary** from the data
- **Create data frame** from the dictionary
- **Export data** to csv file



Data Wrangling

Steps

- **Perform EDA** and determine data labels
- **Calculate:**
 - #of launches for each site
 - #and occurrence of orbit
 - #and occurrence of mission
 - outcome per orbit type]
- **Create binary** landing outcome column (dependent variable)
- **Export data** to csv file

Landing Outcome

- Landing was not always successful
- **True Ocean:** mission outcome had a successful landing to a specific region of the ocean

Landing Outcome Cont.

- **False Ocean:** represented an unsuccessful landing to a specific region of ocean
- **True RTLS:** meant the mission had a successful landing on a ground pad
- **False RTLS:** represented an unsuccessful landing on a ground pad
- **True ASDS:** mission outcome had a successful landing on a drone ship
- **False ASDS:** represented an unsuccessful landing on drone ship
- **Outcomes converted** into 1 for a successful landing and 0 for an unsuccessful landing



EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass(kg) vs. Orbit type

Analysis

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



EDA with SQL

Queries

Display:

- Names of unique launch sites
- 5 records where launch site begins with 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of first successful landing on groundpad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)



Build an Interactive 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

Colored Markers of Launch Outcomes

- 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**



Build a 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



Predictive Analysis (Classification)

- **Create** NumPy array from the Class column
- **Standardize** the data with StandardScaler. Fit and transform the data.
- **Split** the data using train_test_split
- **Create** a GridSearchCV object with cv=10 for parameter optimization
- **Apply** GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- **Calculate** accuracy on the test data using .score() for all models
- **Assess** the confusion matrix for all models
- **Identify** the best model using Jaccard_Score, F1_Score and Accuracy



Results



Results Summary

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- OrbitES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

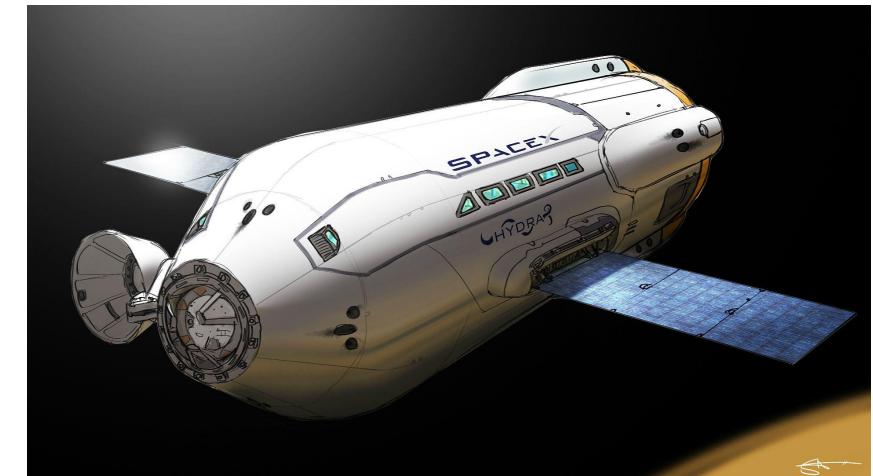
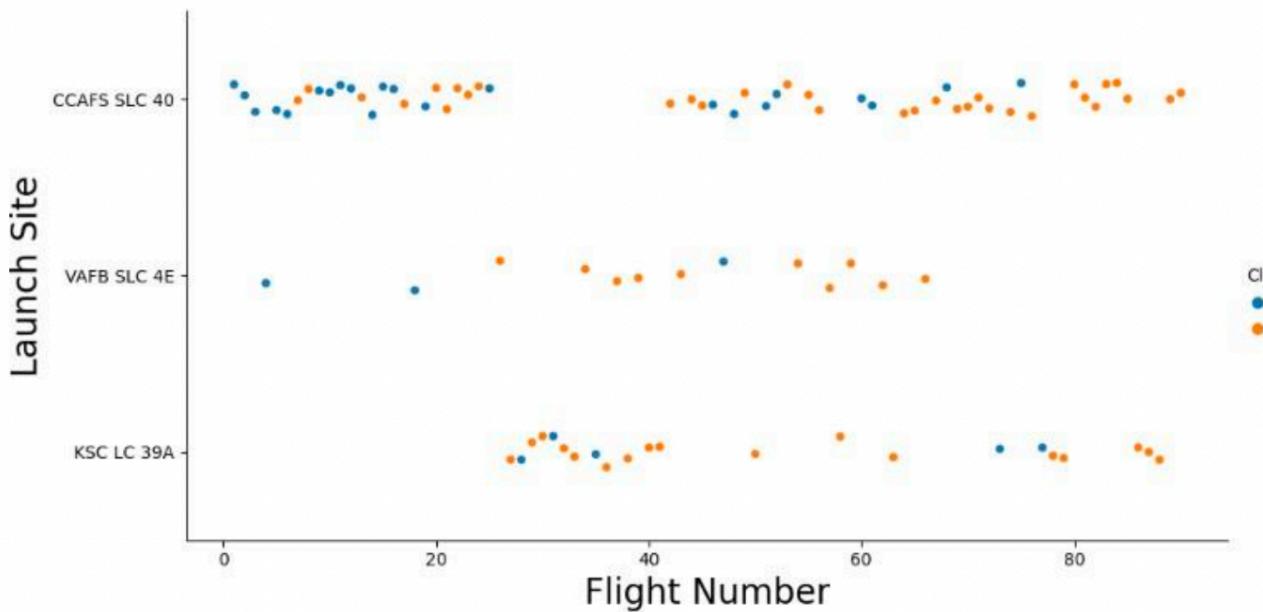
Predictive Analytics

- Decision Tree model is the best predictive model for the dataset



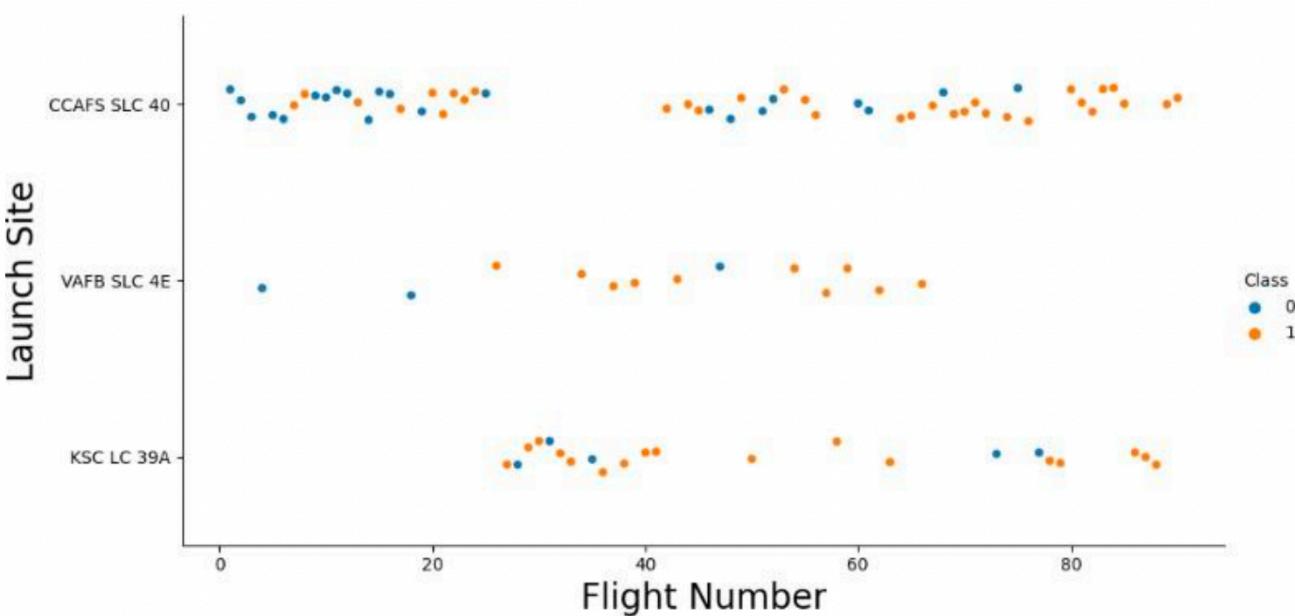
Flight Number vs. Launch Site

- Earlier flights had a **lower success rate**(blue=fail)
- Later flights had a **higher success rate**(orange=success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



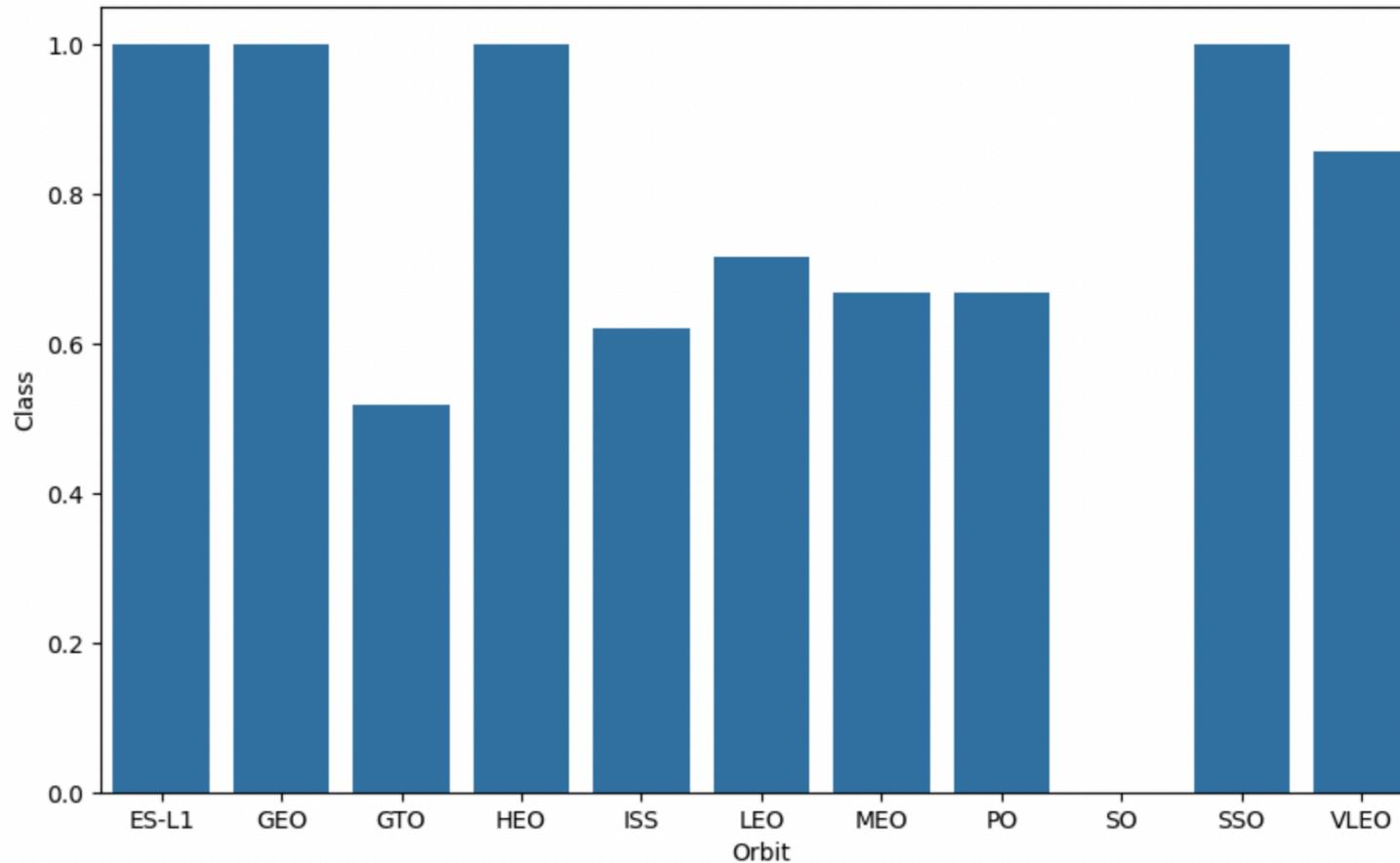
Payload vs. Launch Site

- Typically, the **higher** the **payloadmass(kg)**,the **higher** the **success rate**
- Most launches with a payload greater than 7,000 kg were successful
- KSCLC39A has a 100% success rate for launches less than 5,500kg
- VAFBSKC4E has not launched anything greater than~10,000kg



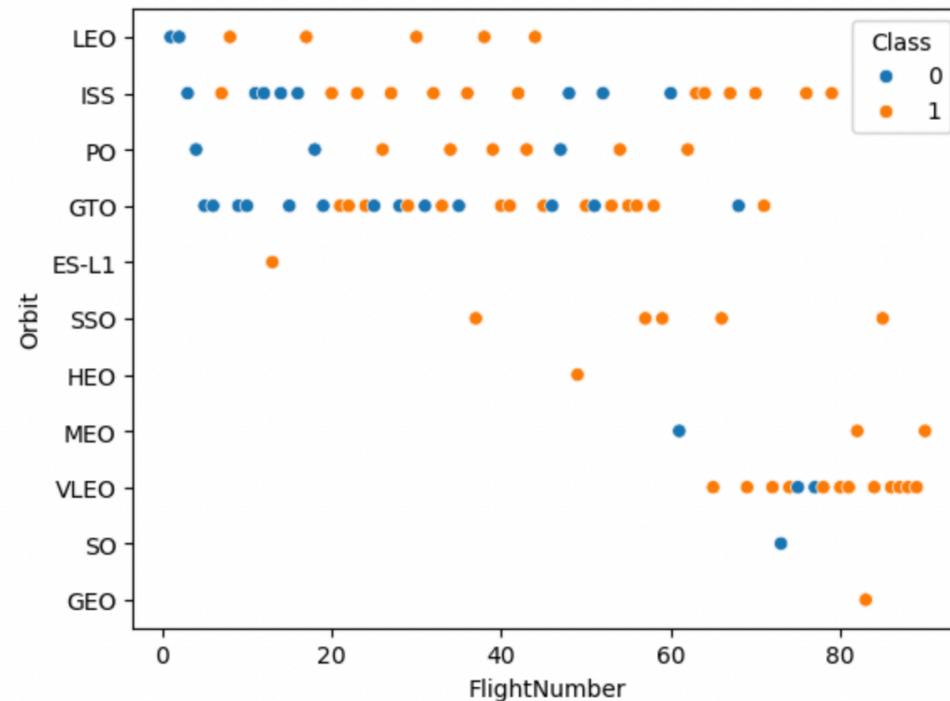
Success Rate vs. Orbit Type

- **100% Success Rate:** ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate:** GTO, ISS, LEO, MEO, PO
- **0% Success Rate:** SO



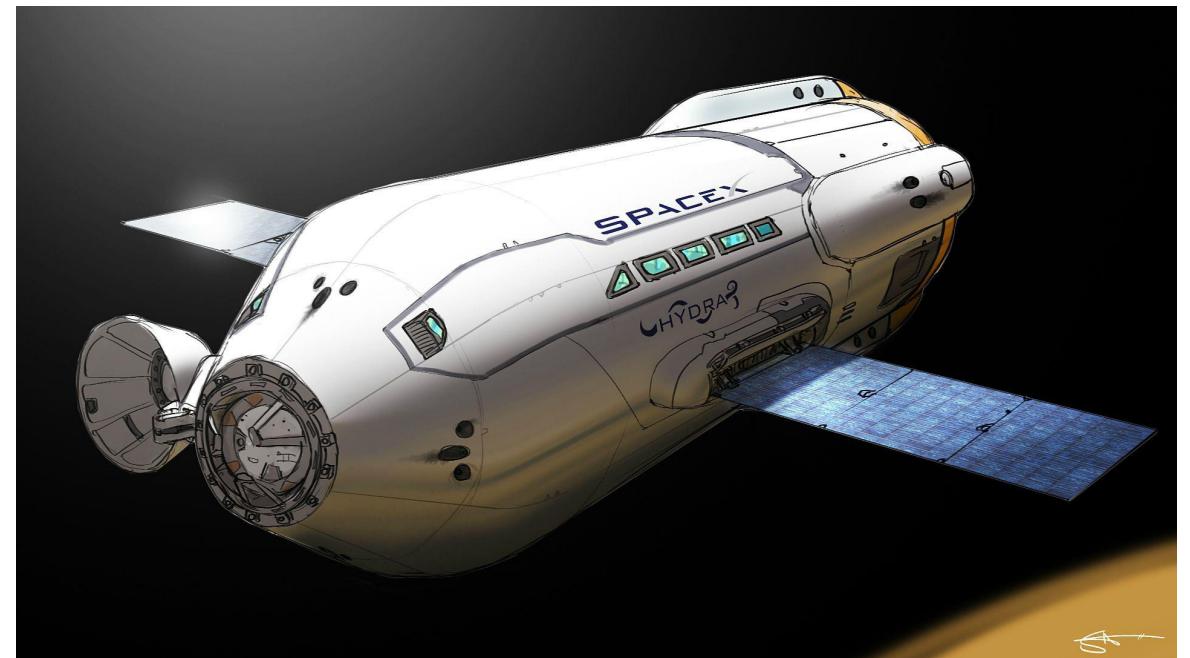
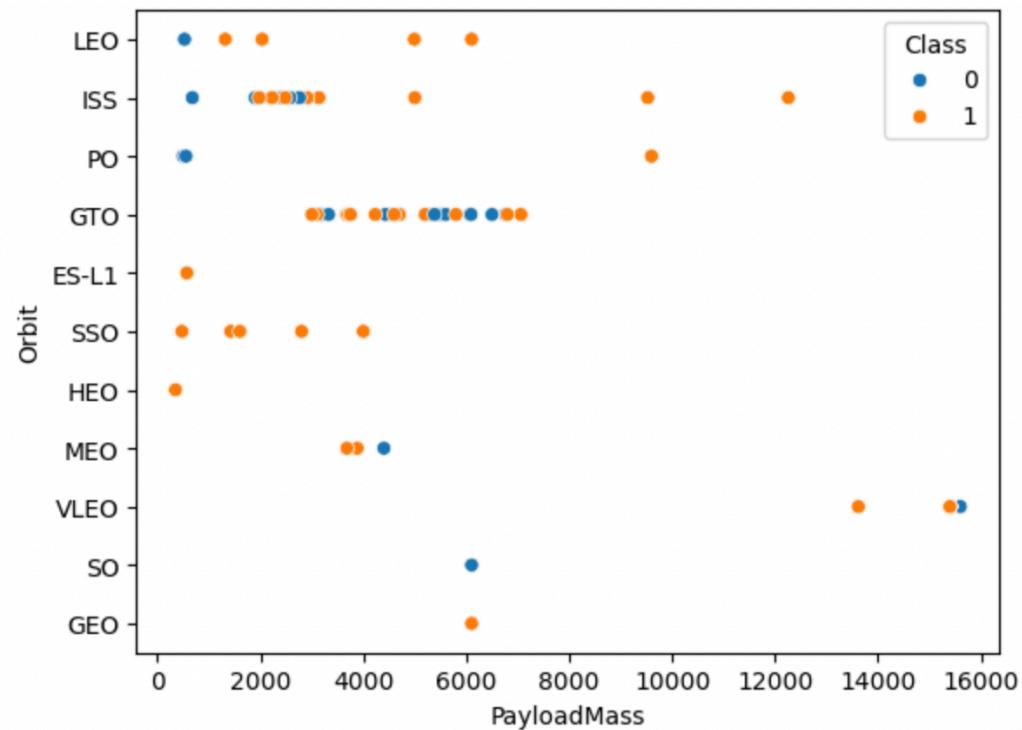
Flight Number vs. Orbit Type

- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



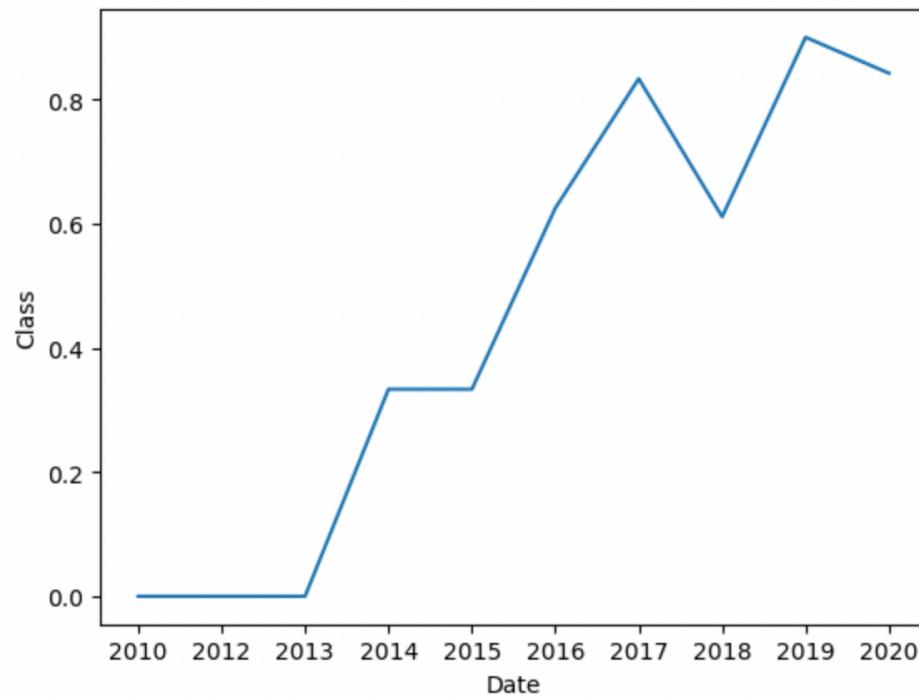
Payload vs. Orbit Type

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



Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 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

Records with Launch Site Starting with CCA

```
%%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

```
%%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

Total/Average Payload Mass

Total Payload Mass

- **45,596 kg** (total) carried by boosters launched by NASA (CRS)

```
%%sql
```

```
SELECT SUM(("PAYLOAD_MASS_KG_")) AS Total_Payload_Mass_Kg  
FROM SPACEXTABLE  
WHERE Customer = 'NASA (CRS)';
```

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

Total_Payload_Mass_Kg

45596

Average Payload Mass

- **2,928 kg** (average) carried by booster version F9 v1.1

```
%%sql
```

```
SELECT AVG("PAYLOAD_MASS_KG_") AS "Average_Payload_Mass_kg"  
FROM SPACEXTABLE  
WHERE BOOSTER_VERSION LIKE "F9 V1.1";
```

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

Average_Payload_Mass_kg

2928.4

Landing & Mission Info

1st Successful Landing in Ground Pad

- 12/22/2015

```
%%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

```
%%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



Booster Drone Ship Landing

- Booster mass greater than 4,000 but less than 6,000
- JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105

Total Number of Successful and Failed Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

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

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

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

Boosters Carried Maximum Payload

Carrying Max Payload

- F9B5B1048.4
- F9B5B1049.4
- F9B5B1051.3
- F9B5B1056.4
- F9B5B1048.5
- F9B5B1051.4
- F9B5B1049.5
- F9B5B1060.2
- F9B5B1058.3
- F9B5B1051.6
- F9B5B1060.3
- F9B5B1049.7

```
%%sql
```

```
SELECT Booster_Version  
FROM SPACEXTABLE  
WHERE PAYLOAD_MASS_KG_ = (  
    SELECT MAX(PAYLOAD_MASS_KG_)  
    FROM SPACEXTABLE  
) ;
```

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

Booster_Version

F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5

2015 Launch Records

In 2015

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

```
%%sql
```

```
SELECT
    substr("Date", 6, 2) AS Month,
    "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.
```

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

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

Ranked Descending

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

```
SELECT
    "Landing_Outcome",
    COUNT(*) AS Count_of_Outcomes
FROM
    SPACEXTABLE
WHERE
    substr(Date, 0, 5) BETWEEN '2010' AND '2017'
    AND (substr(Date, 6, 2) >= '06' OR substr(Date, 6, 2) <= '03')
GROUP BY
    "Landing_Outcome"
ORDER BY
    Count_of_Outcomes DESC;
```

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

Landing_Outcome	Count_of_Outcomes
Success (drone ship)	9
No attempt	9
Success (ground pad)	7
Failure (drone ship)	4
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2

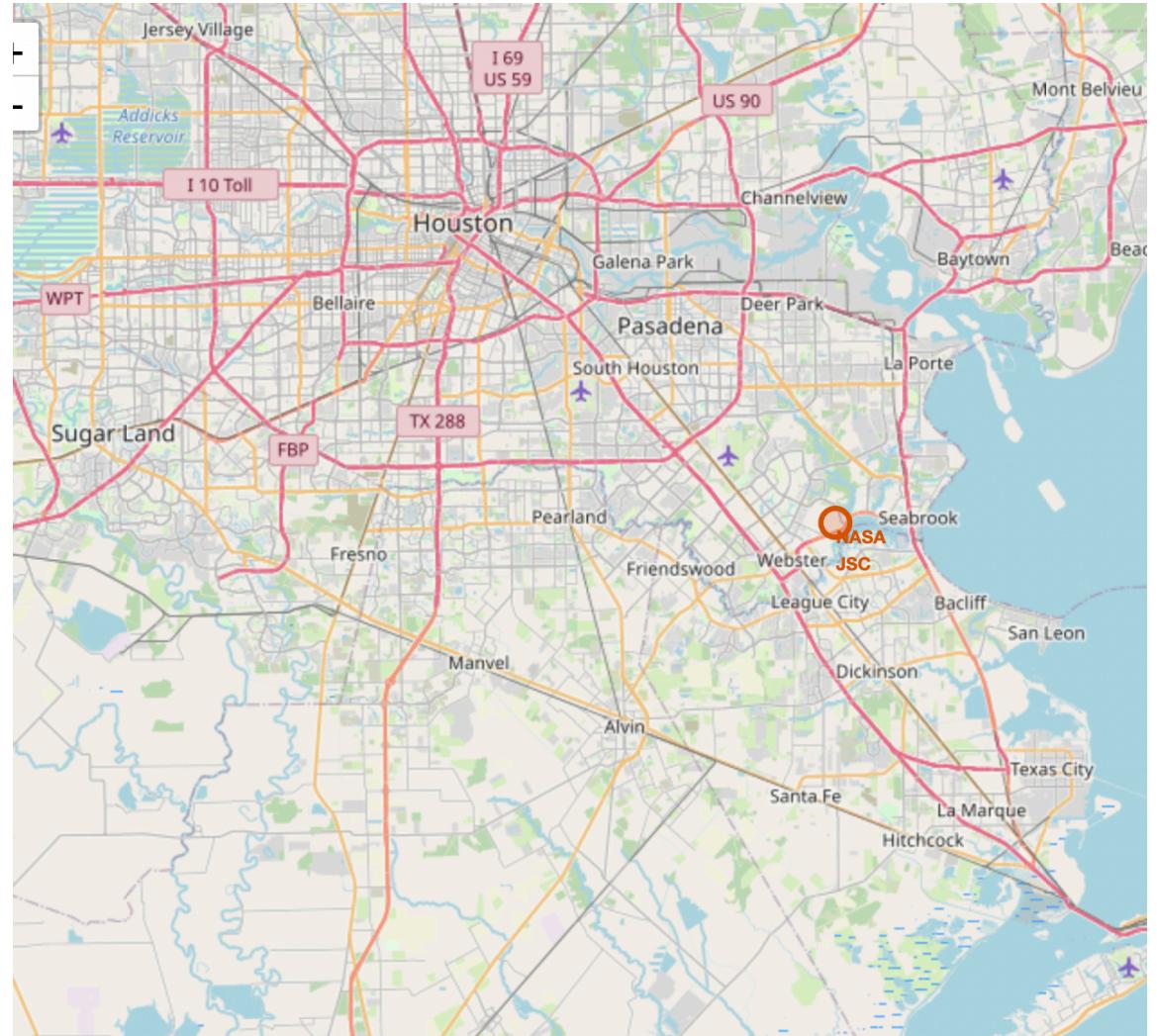
Launch Site Analysis



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:

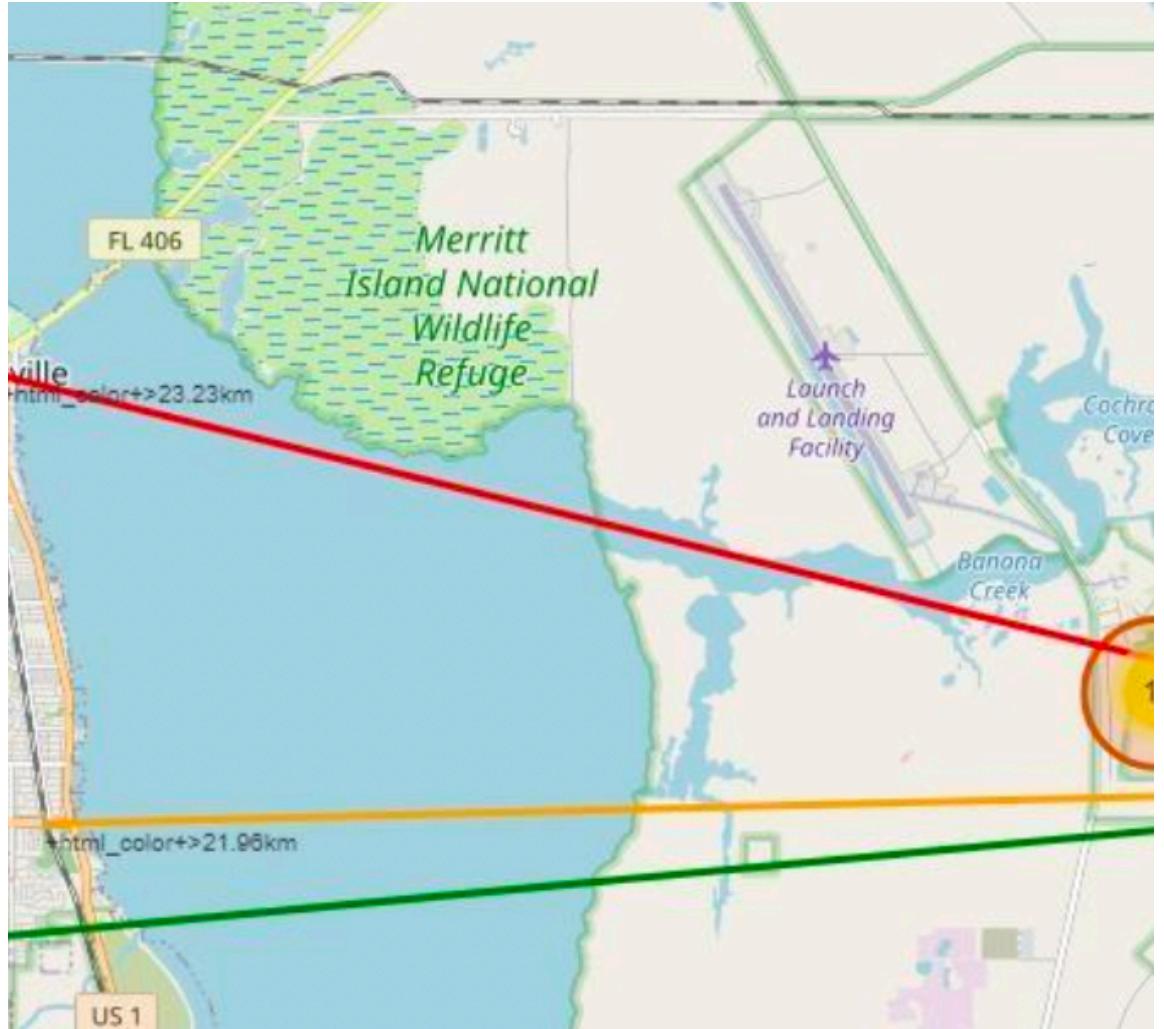
- **Green** markers for successful launches
- **Red** markers for unsuccessful launches
- Launchsite**CCAFS SLC-40**has a **3/7 success rate(42.9%)**



Distance to Proximities

CCAFS SLC-40

- **0.86km** from nearest coastline
- **21.96km** from nearest railway
- **23.23km** from nearest city
- **26.88km** from nearest highway

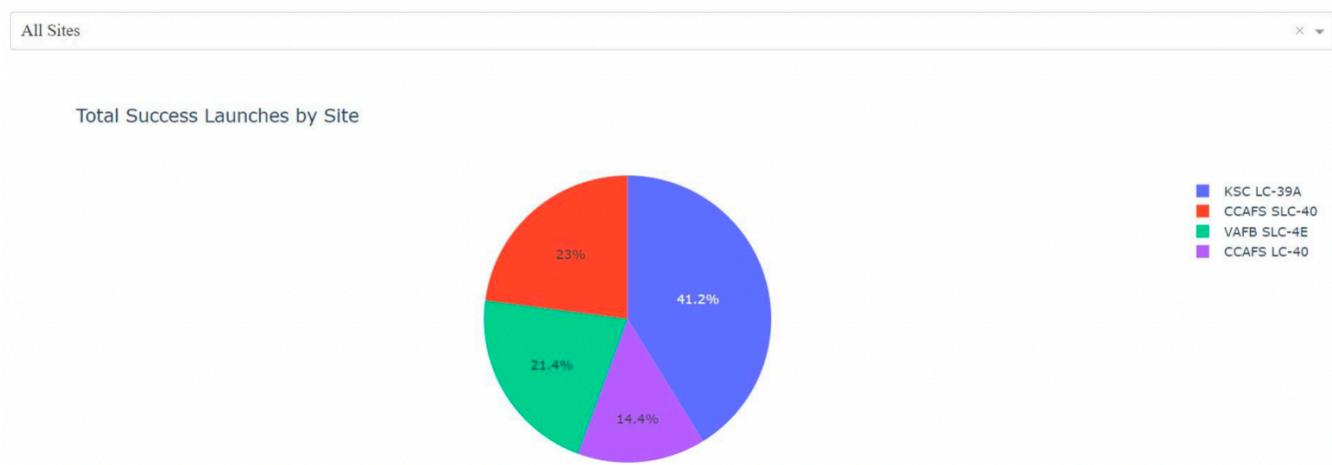


Dashboard with Plotly



Launch Success by Site

- **Success as Percent of Total**
 - **KSCLC-39A has the most successful launches amongst launch sites (41.2%)**



Launch Success (KSC LC-39A)

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

SpaceX Launch Records Dashboard

KSC LC-39A x ▾

Total Success Launches for Site KSC LC-39A

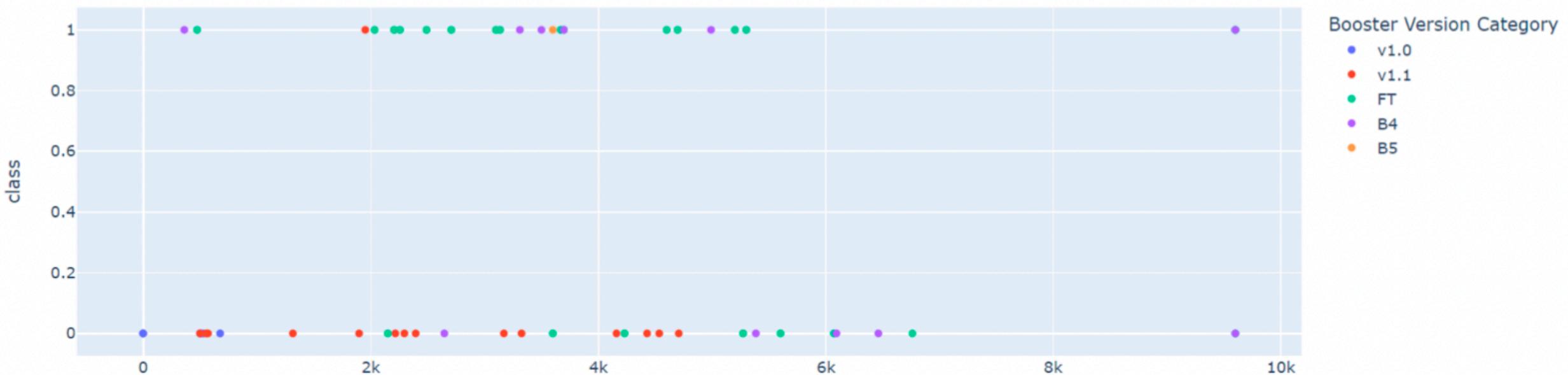


Payload Mass and Success

By Booster Version

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

Correlation Between Payload and Success for All Sites



Predictive Analytics



Classification Accuracy

Accuracy

- All the **models** performed at about the same level and had the **same scores and accuracy**. This is likely due to the **small dataset**.
- The **Decision Tree model slightly outperformed** the rest when looking at `.best_score_`
- `.best_score_` is the average of all cv folds for a single combination of the parameters

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

```
: 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 0.9017857142857142  
Best params is : {'criterion': 'gini', 'max_depth': 16, 'max_features': 'auto'}
```

Confusion Matrix

Performance Summary

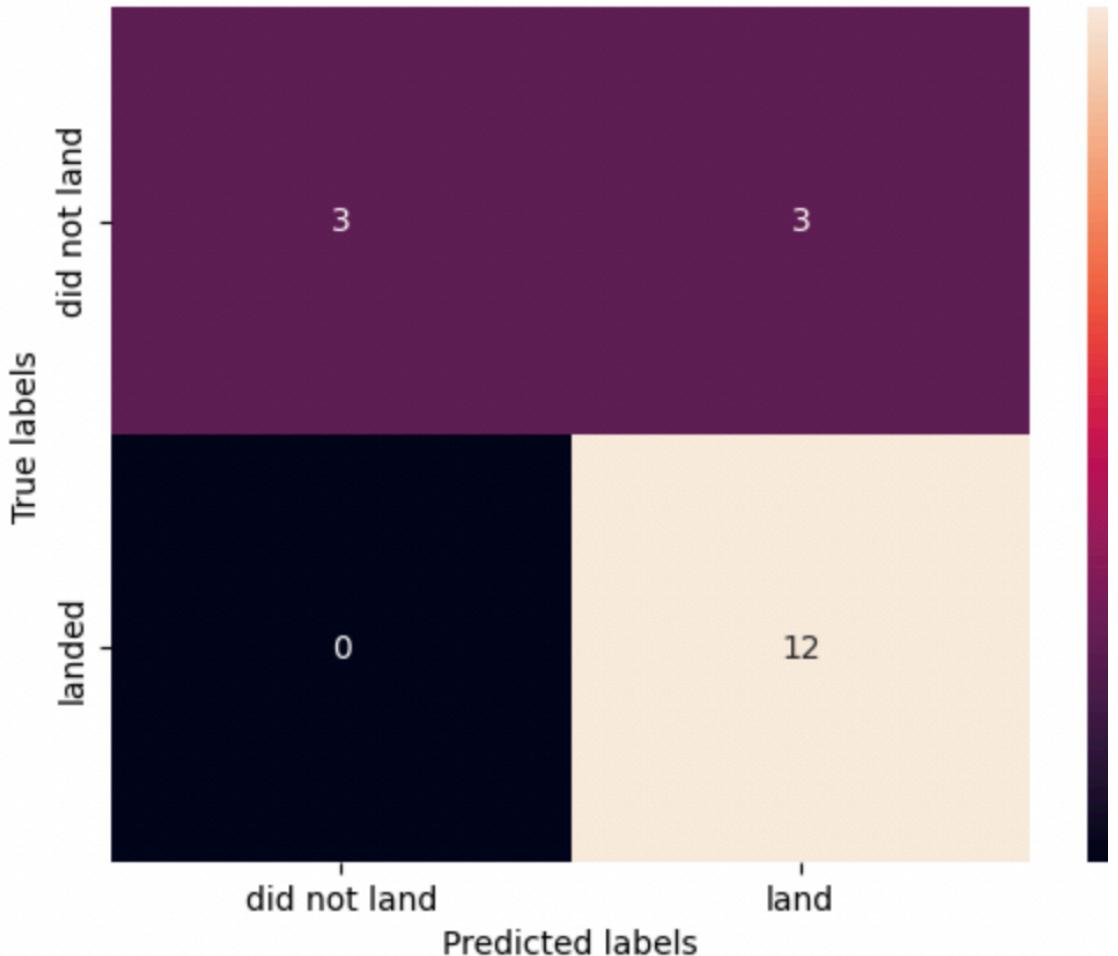
- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- 12 True positive
- 3 True negative
- **3 False positive**
- 0 False Negative

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \cdot 12/15=.80$$

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \cdot 12/12=1$$

$$\bullet \text{F1Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Confusion Matrix



Conclusions

- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator:** Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast:** All the launch sites are close to the coast
- **Launch Success:** Increases over time
- **KSCLC39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits:** ESL1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate



SPACEX