
# Winning Space Race with Data Science

Yale Vervloet

06/03/2024

# CONTENTS

# Executive Summary

## Methodologies

With the goal of establishing a predictive model of successful launches, the following steps were taken:

- Collecting data from SpaceX  REST API & Webscraping
- Wrangling data to create success/fail outcome variables
- Exploring patterns with data visualization techniques, considering relationships between payload, launch site, flight number and yearly trend
- Analyzing data with SQL to calculate payload range for successful launches as well as the number of successful and failed outcomes in various constellations
- Exploring launch site success rates basis proximity to geographical markers via Folium maps
- Evaluating various predictive models through Machine Learning techniques such as SVM, decision tree and K-nearest neighbor (KNN)

## Results

We established that:

- Launch success ratio increased over time
- KSC LC-39A has highest success rate of landing sites
- Most launch sites are near the equator
- All models performed similarly on the test data set
- Decision tree model was slightly better than the others




Classification: Public

# 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)




Classification: Public

Section 1

# Methodology




Classification: Public

# METHODOLOGY

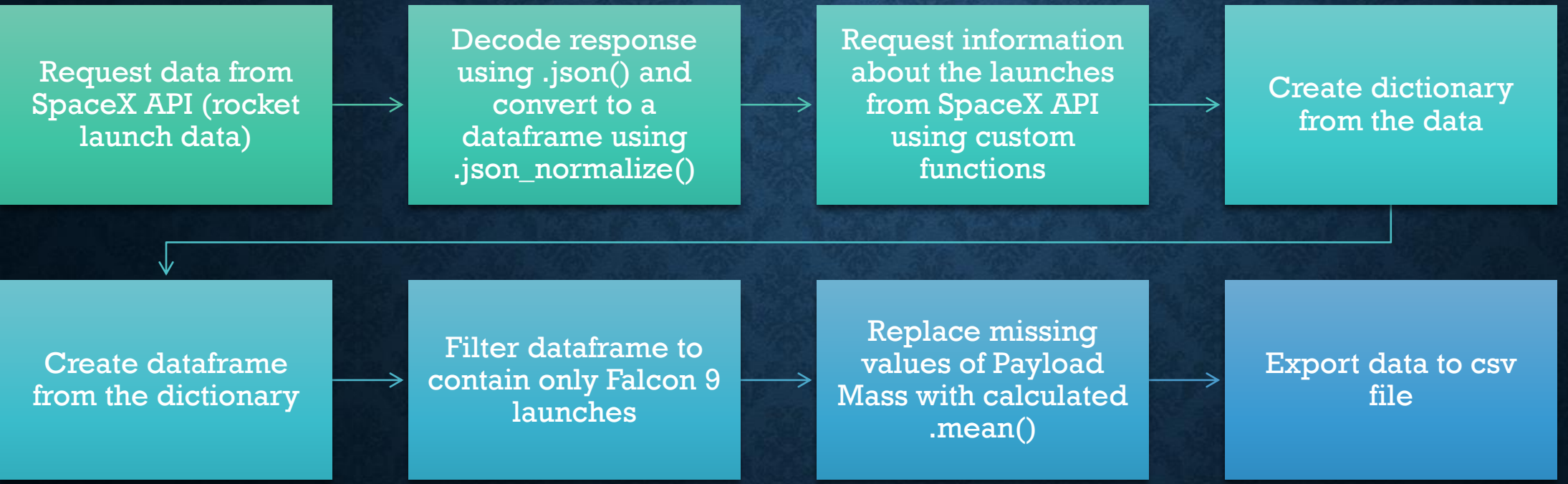
With the goal of establishing a predictive model of successful launches, the following steps were taken:

- Collecting data from SpaceX  REST API & Webscraping
- Wrangling data to create success/fail outcome variables
- Exploring patterns with data visualization techniques, using Matplotlib, Seaborn and Plotly Dash
- Analyzing data with SQL to calculate payload range for successful launches as well as the number of successful and failed outcomes in various constellations
- Exploring launch site success rates basis proximity to geographical markers via Folium maps
- Evaluating various predictive models through Machine Learning techniques such as SVM, decision tree and K-nearest neighbor (KNN)




Classification: Public

# DATA COLLECTION – SPACEX API




https://github.com/YVE014/Capstone-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb




Classification: Public

# DATA COLLECTION – WEB SCRAPING

| 01 | 02 | 03 | 04 | 05 | 06 | 07 |
|---|---|---|---|---|---|---|
| Request data (Falcon 9 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 dataframe from the dictionary | Export data to csv file |

https://github.com/YVE014/Capstone-project/blob/main/jupyter-labs-webscraping.ipynb




Classification: Public

# DATA WRANGLING

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
- 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: meant the 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

https://github.com/YVE014/Capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

# 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

## Goals

- **View relationship** by 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



https://github.com/YVE014/Capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

# EDA WITH SQL

**Using SQL, following queries were looked at:**

- 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.
- Date of first successful landing on ground pad
- 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)



https://github.com/YVE014/Capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

# BUILD AN INTERACTIVE MAP WITH FOLIUM

**Markers Indicating Launch Sites**

- Added **bluecircle** 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**



https://github.com/YVE014/Capstone-project/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

# BUILD A DASHBOARD WITH PLOTLY DASH

**Dropdown List with Launch Sites**

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

**Slider of Payload Mass Range**

- Allow user to select payload mass range

**Pie Chart Showing Successful Launches**

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

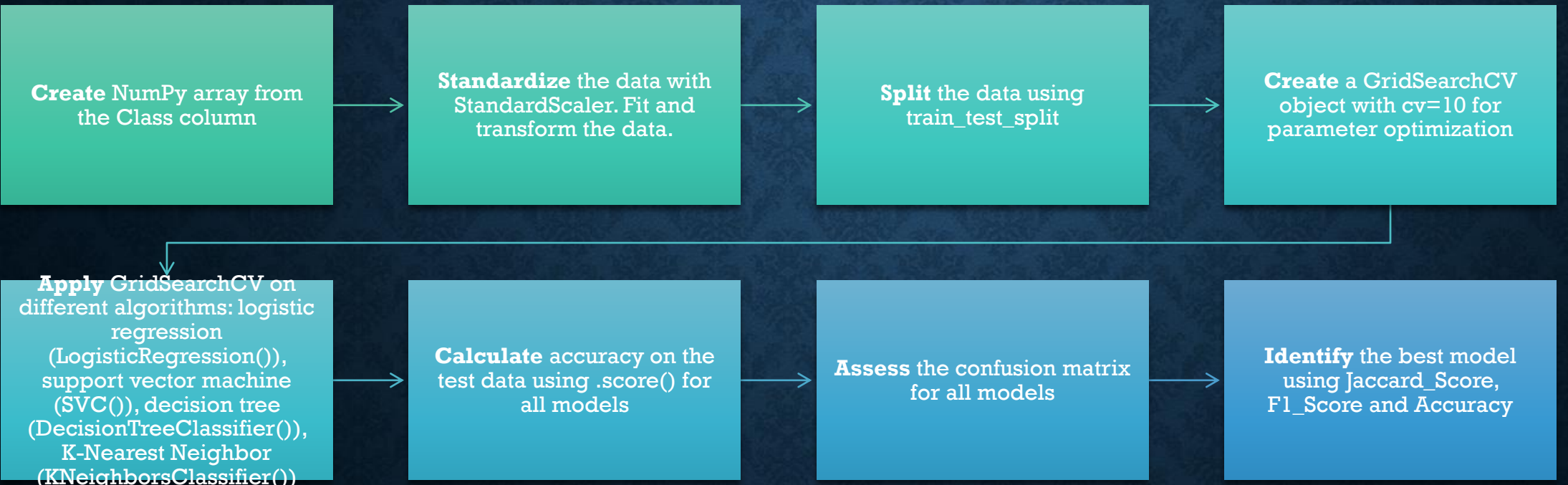
**Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version**

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



https://github.com/YVE014/Capstone-project/blob/main/DASH%20Pie.PNG

# PREDICTIVE ANALYSIS (CLASSIFICATION)






https://github.com/YVE014/Capstone-project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

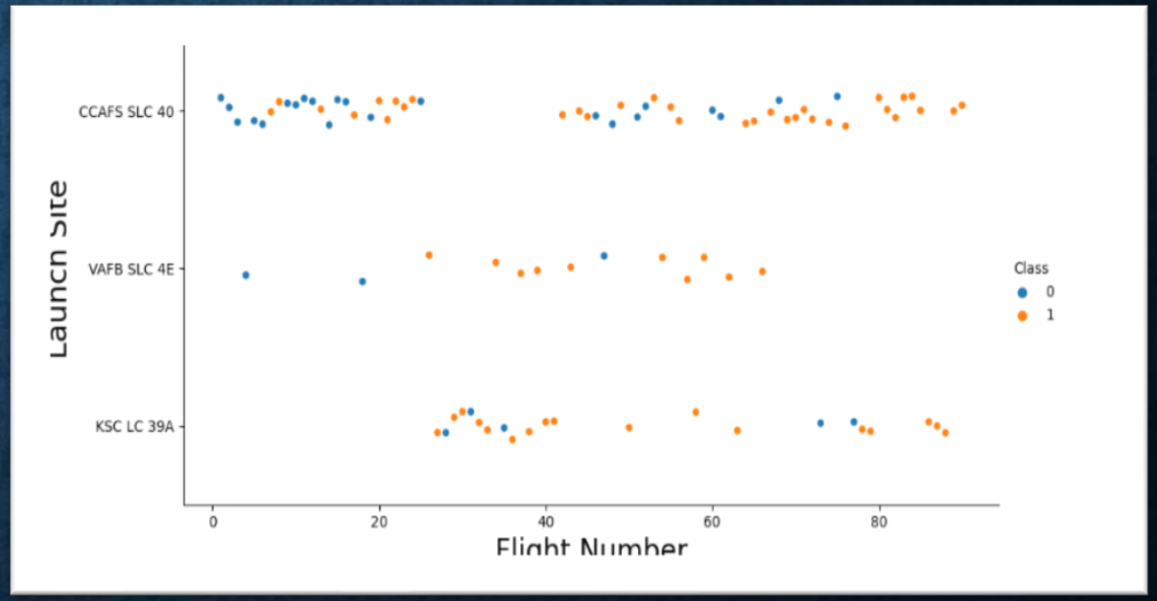
# Section 2

# Insights drawn from EDA


Classification: Public

# 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



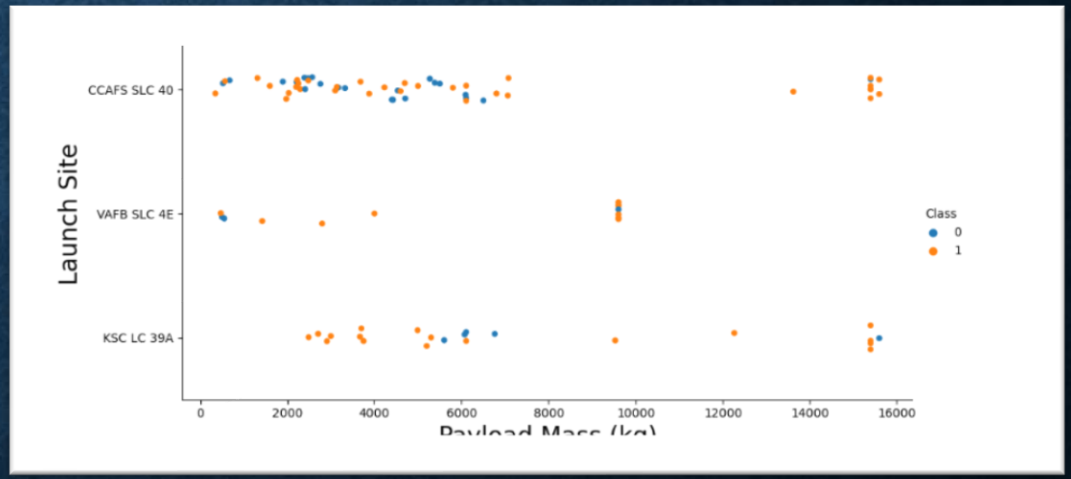




Classification: Public

# PAYLOAD VS. LAUNCH SITE

- Typically, the **higher** the **payload mass** (kg), the **higher** the **success rate**
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



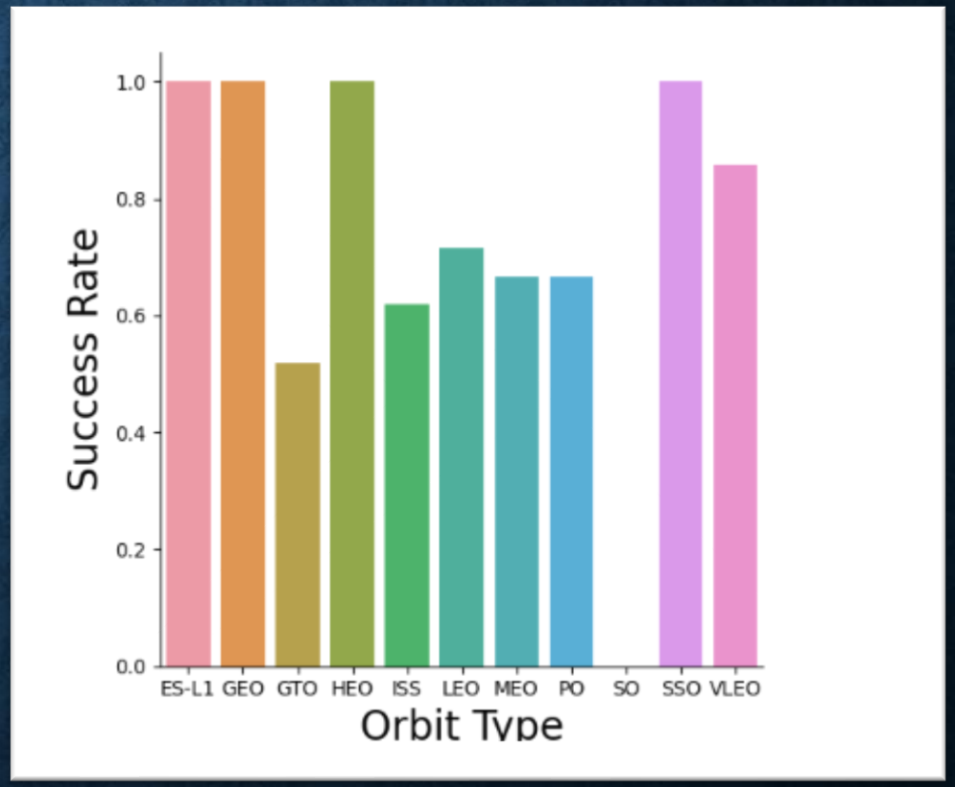




Classification: Public

# SUCCESS RATE VS. ORBIT TYPE

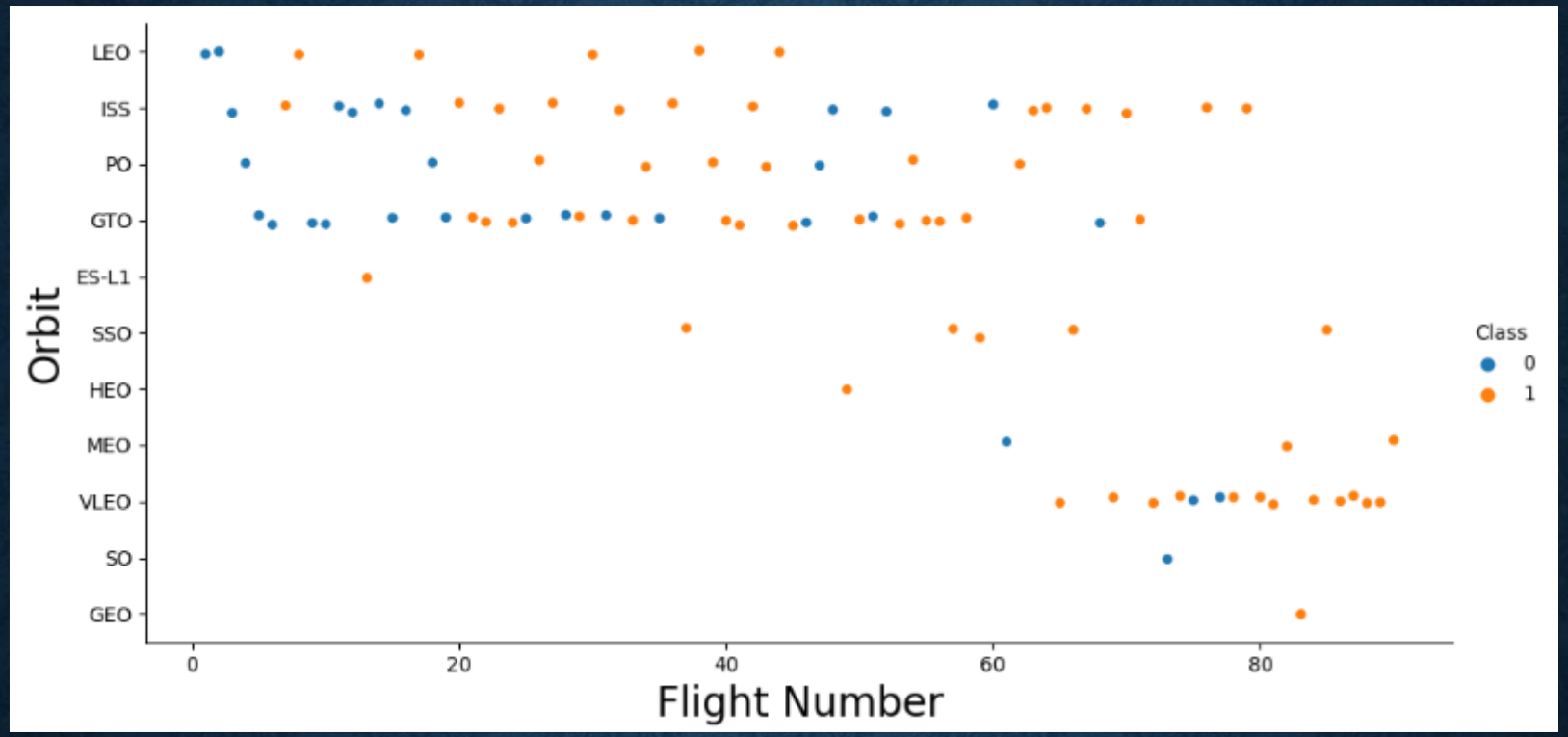
- **100% Success Rate**: ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO







Classification: Public




# FLIGHT NUMBER VS. ORBIT TYPE

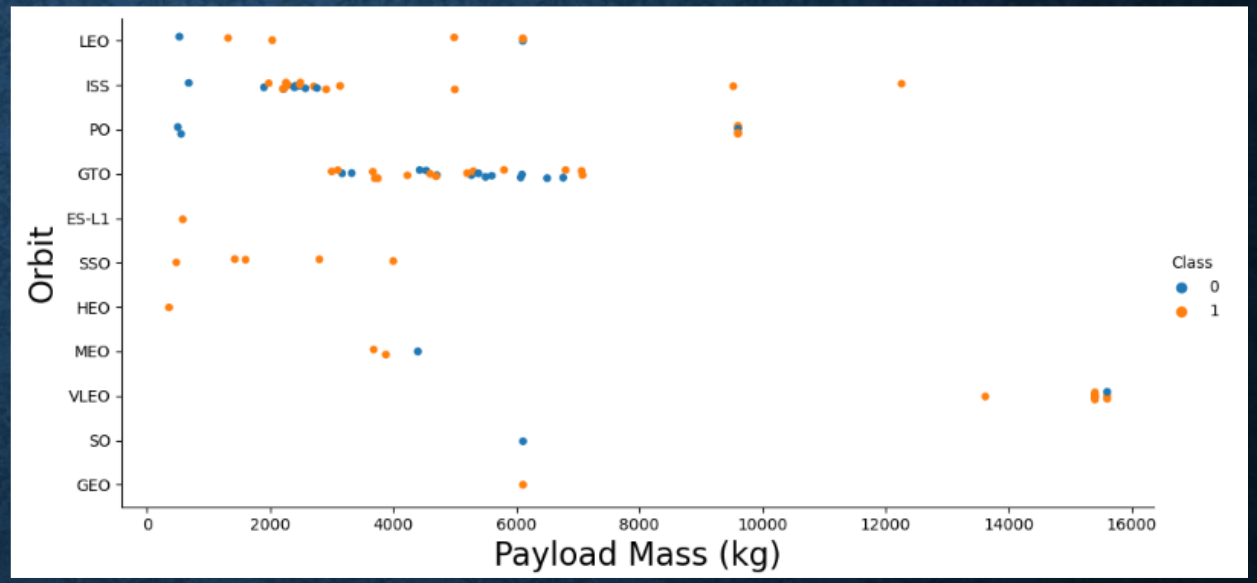
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend




Classification: Public

# PAYLOAD VS. ORBIT TYPE

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



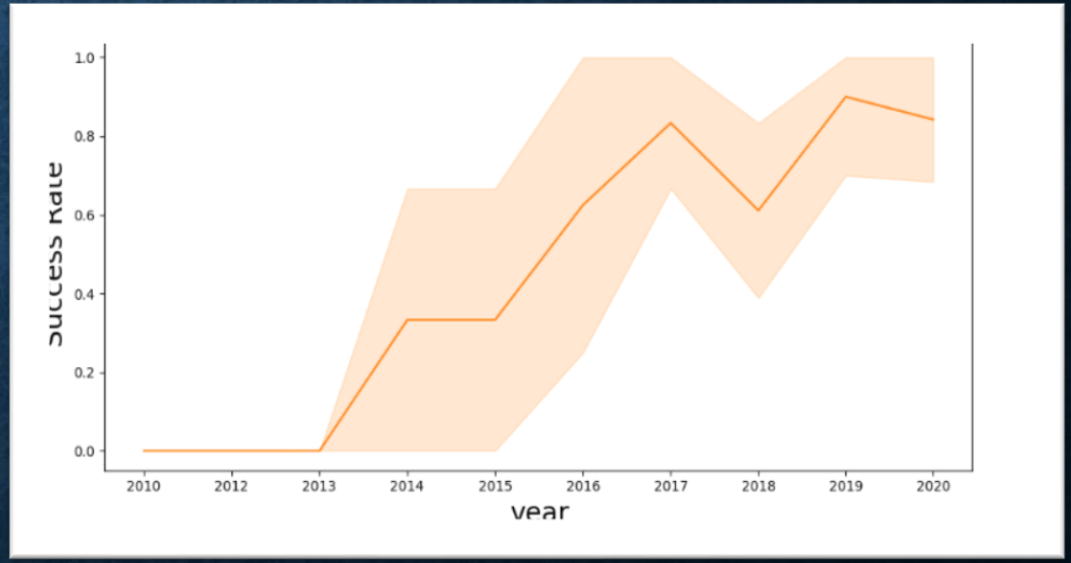




Classification: Public

# 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







Classification: Public

# ALL LAUNCH SITE NAMES

```
[30]: %sql ibm_db_sa://yyy33800:dwNKg8J3L0IBd6CP@1bbf73c5
      %sql SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL;

       * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9
         sqlite:///my_data1.db
      Done.
```

[30]:

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

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


Classification: Public

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
   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 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 00:35:00 | 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'


Classification: Public

# TOTAL PAYLOAD MASS

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
    1
45596
```

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




Classification: Public

# AVERAGE PAYLOAD MASS BY F9 V1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

| 1 |
|---|
| 2928 |

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


Classification: Public

# FIRST SUCCESSFUL GROUND LANDING DATE

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85
   sqlite:///my_data1.db
Done.
```

| 1 |
|---|
| 2015-12-22 |

**1st Successful Landing in Ground Pad**

- 12/22/2015




Classification: Public

# SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-
   sqlite:///my_data1.db
Done.
```

| payload |
|---|
| JCSAT-14 |
| JCSAT-16 |
| SES-10 |
| SES-11 / EchoStar 105 |

Booster mass greater than 4,000 but less than 6,000

- JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105


Classification: Public

# TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

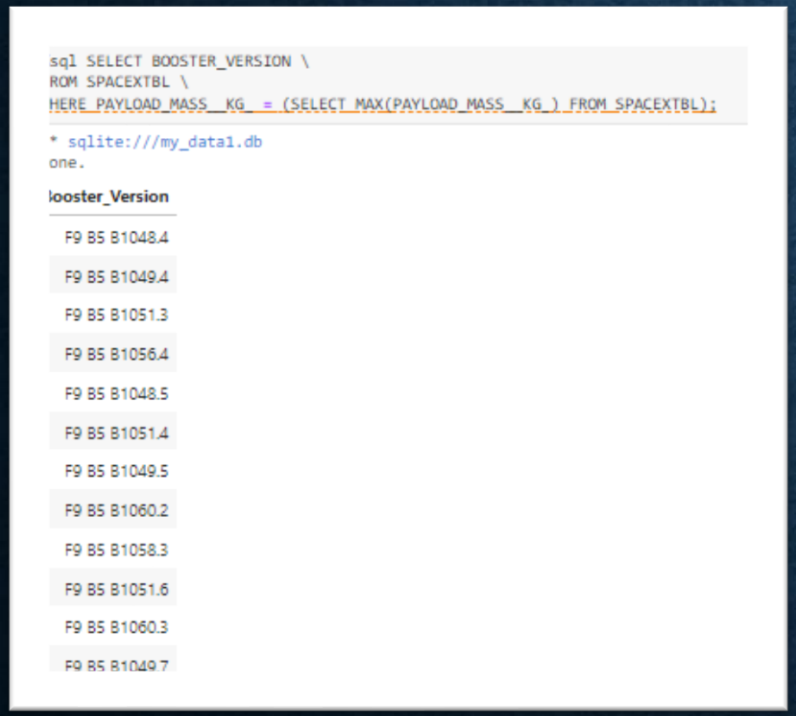
```
 * sqlite:///my_data1.db
Done.
```

| Mission_Outcome | total_number |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

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


Classification: Public

# BOOSTERS CARRIED MAXIMUM PAYLOAD



```
sql SELECT BOOSTER_VERSION \
ROM SPACEXTBL \
HERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

| 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 |

- **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




Classification: Public

# 2015 LAUNCH RECORDS

- **In 2015**
- Showing month, date, booster version, launch site and landing outcome

```
%sql SELECT substr(Date,4,2) as month, DATE,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 | Date | Booster_Version | Launch_Site | Landing _Outcome |
|---|---|---|---|---|
| 01 | 10-01-2015 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 04 | 14-04-2015 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |




Classification: Public

# RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

```
sql SELECT [Landing _Outcome], count(*) as count_outcomes \
ROM SPACEXTBL \
HERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC
 * sqlite:///my_data1.db
one.
```

| Landing _Outcome | count_outcomes |
|---|---|
| Success | 20 |
| No attempt | 10 |
| Success (drone ship) | 8 |
| uccess (ground pad) | 6 |
| Failure (drone ship) | 4 |
| Failure | 3 |
| Controlled (ocean) | 3 |
| Failure (parachute) | 2 |

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

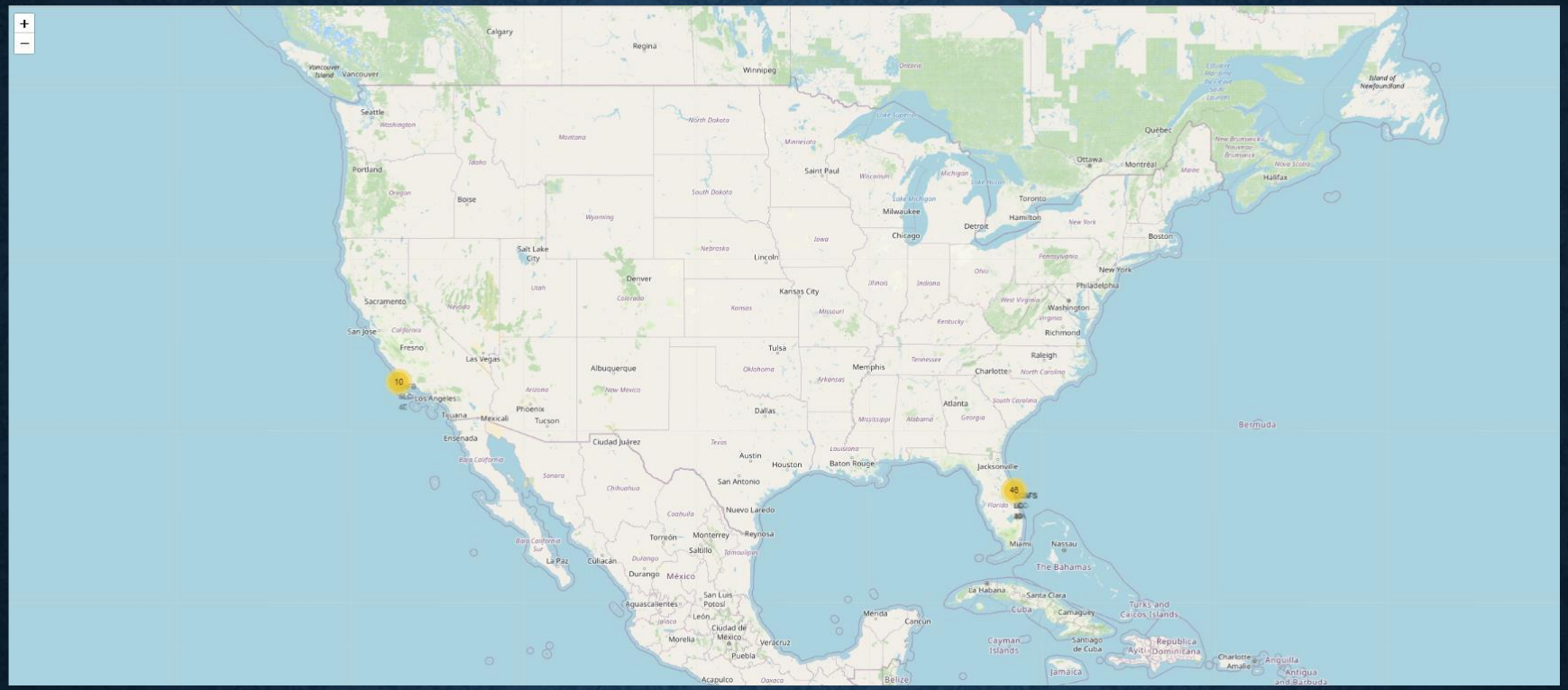



Classification: Public

# LAUNCH SITE PROXIMITY ANALYSIS

Section 3


Classification: Public




# LAUNCH SITE LOCATIONS

- **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.




Classification: Public


# LAUNCH OUTCOMES VIA FOLIUM

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




Classification: Public

# DISTANCE TO PROXIMITIES

- **CCAFS SLC-40**
- **.86 km** from nearest coastline
- **21.96 km** from nearest railway
- **23.23 km** from nearest city
- **26.88 km** from nearest highway





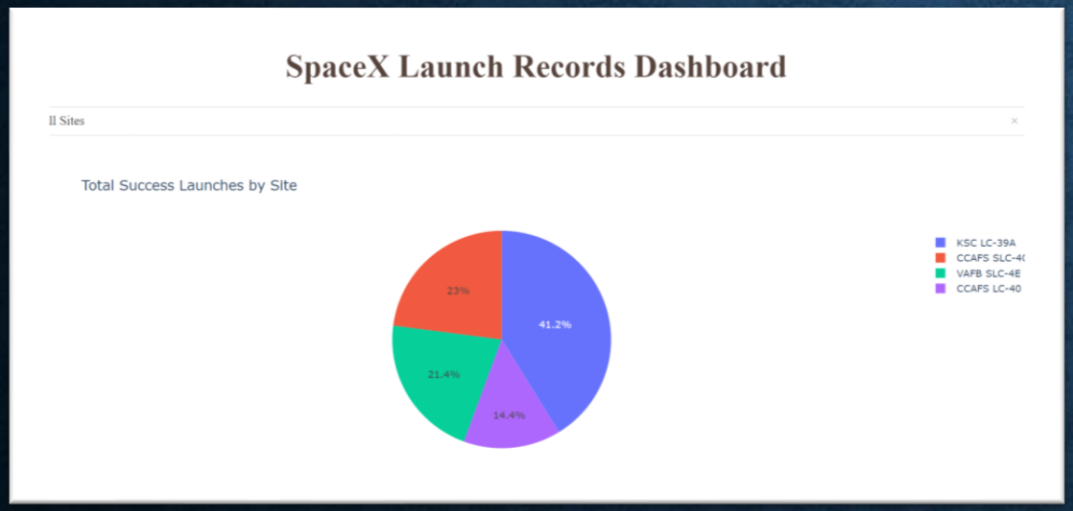
Classification: Public

Section 4

# BUILD A DASHBOARD WITH PLOTLY DASH


Classification: Public

# LAUNCH SUCCESS PER SITE

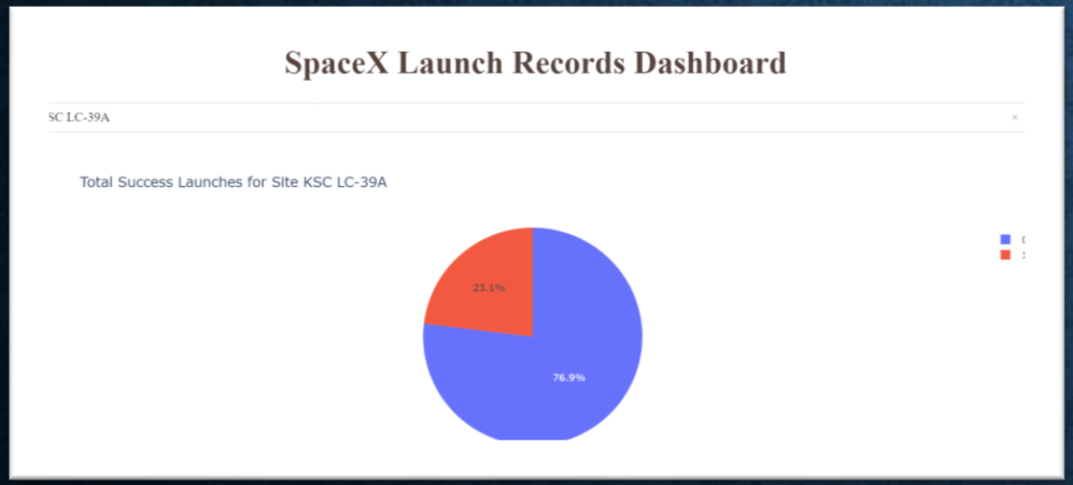



- **Success as Percent of Total**
- **KSC LC-39A** has the **most successful launches** amongst launch sites (**41.2%**)




Classification: Public

# SUCCESS RATIO




- **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




Classification: Public

# PAYLOAD MASS AND SUCCESS


- **By Booster Version**
- **Payloads between2,000 kg** and **5,000 kg** have the **highest success rate**
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome




Classification: Public

# PREDICTIVE ANALYSIS (CLASSIFICATION)

Section 5


Classification: Public

# CLASSIFICATION ACCURACY

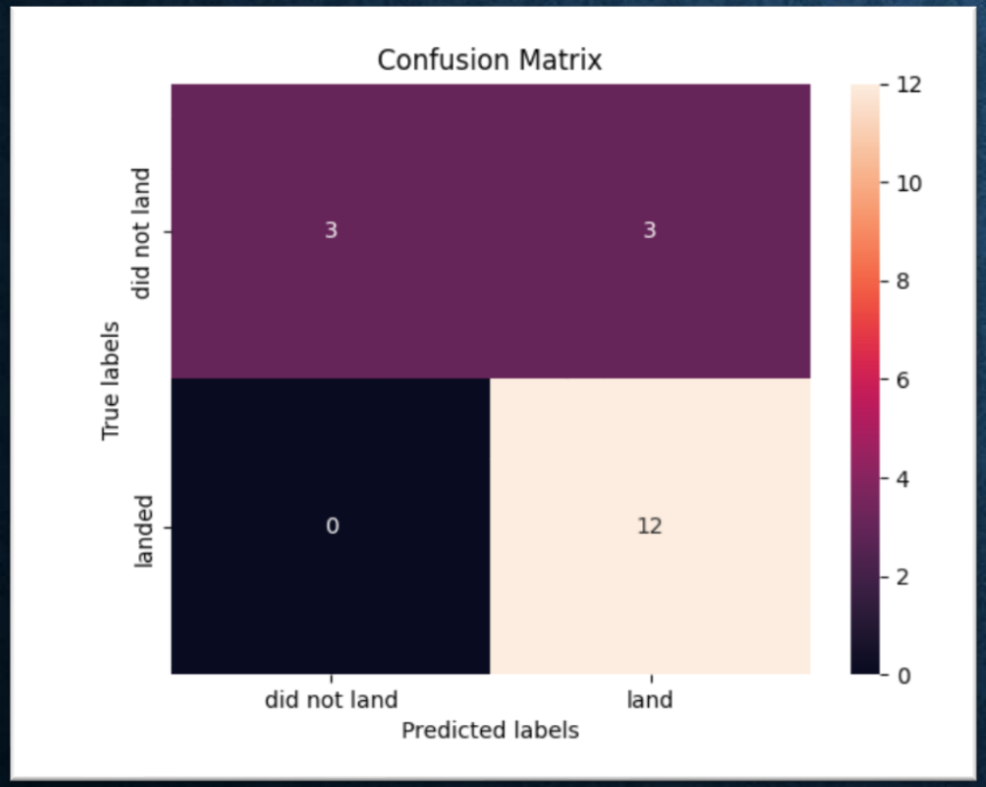
| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.800000 | 0.800000 | 0.800000 | 0.800000 |
| F1_Score | 0.888889 | 0.888889 | 0.888889 | 0.888889 |
| Accuracy | 0.833333 | 0.833333 | 0.833333 | 0.833333 |

- **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 mode 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




Classification: Public

# CONFUSION MATRIX




- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:12 True positive
- 3 True negative
- **3 False positive**
- 0 False Negative
- **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




Classification: Public

# 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
- **KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, 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




Classification: Public

# FURTHER RESEARCH

**Dataset**: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set

**Feature Analysis / PCA**: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy




Classification: Public