A Data Science Space Race

William Patton

09-2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Summary of Methodologies

The research attempts to identify the factors for a successful rocket landing. In order to make this an accurate conclusion, the following methodologies are used:

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data to create a success/fail outcome variable
- **Explore** data with data visualization techniques, and consider the following factors: payload, launch site, flight number and yearly trend
- **Analyze** data with SQL, calculating the following statistics: total payload, payload range of 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)

Results

Exploratory Data Analysis:

- Launch success has improved over time
- With data, we can prove with landing and orbits have the best success rates

Visualization Analytics:

- Show how most launch sites are near the equator and coastlines

Predictive Analytics:

- Models all perform similarly on the test set, but the tree model has better accuracy

Introduction

- SpaceX, a leader in the spade industry. Stiving to make space travel more affordable. They have continued to make many accomplishments in their path to make space travel more accessible. Some accomplishments include sending spacecrafts to the International Space Station, launching a satellite constellation that provides internet access, and sending manned missions into space. SpaceX is able to achieve all of this because they are the first to be able to reuse the first stage launch. This process cost approximately $62 million per launch with SpaceX, though other providers would be upwards of $165 million per launch.
- If we can use data science to determine if the first stage will land, we can determine the price of the launch as well. In order to do this, we can use public data of previous launches and machine learning models to predict whether SpaceX can reuse its first stage.

Section 1

Methodology

Methodology

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

Request data from SpaceX API (rocket launch data)

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

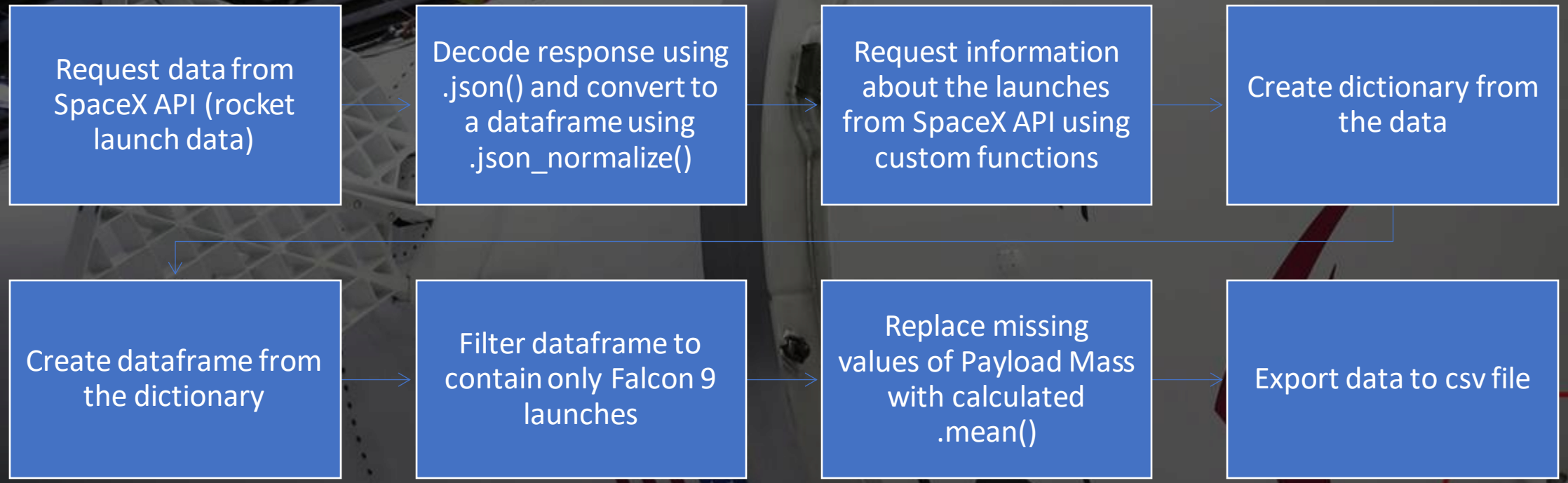
Create dataframe from the dictionary

Filter dataframe to contain only Falcon 9 launches

Replace missing values of Payload Mass with calculated .mean()

Export data to csv file

Data Collection – SpaceX API



Github - SpaceX - Collection

Data Collection – Web Scraping

Request	•Request data (Falcon 9 launch data) from Wikipedia
Create	•Create BeautifulSoup object from HTML response
Extract	•Extract column names from HTML table header
Collect	•Collect data from parsing HTML tables
Create	•Create dictionary from the data
Create	•Create dataframe from the dictionary
Export	•Export data to csv file

Github - SpaceX - Webscraping

Data Wrangling

Part 1

[Github - SpaceX - 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

Data Wrangling – Landing Outcomes

Part 2

Github - SpaceX - Wrangling

- 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

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** 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

[Github - SpaceX - EDA Data Visualization]

EDA with SQL

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 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)

Github - SpaceX - EDA SQL

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

All these markers were added with the aim to finding an optimal location for building a launch site

Github - SpaceX - Folium

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

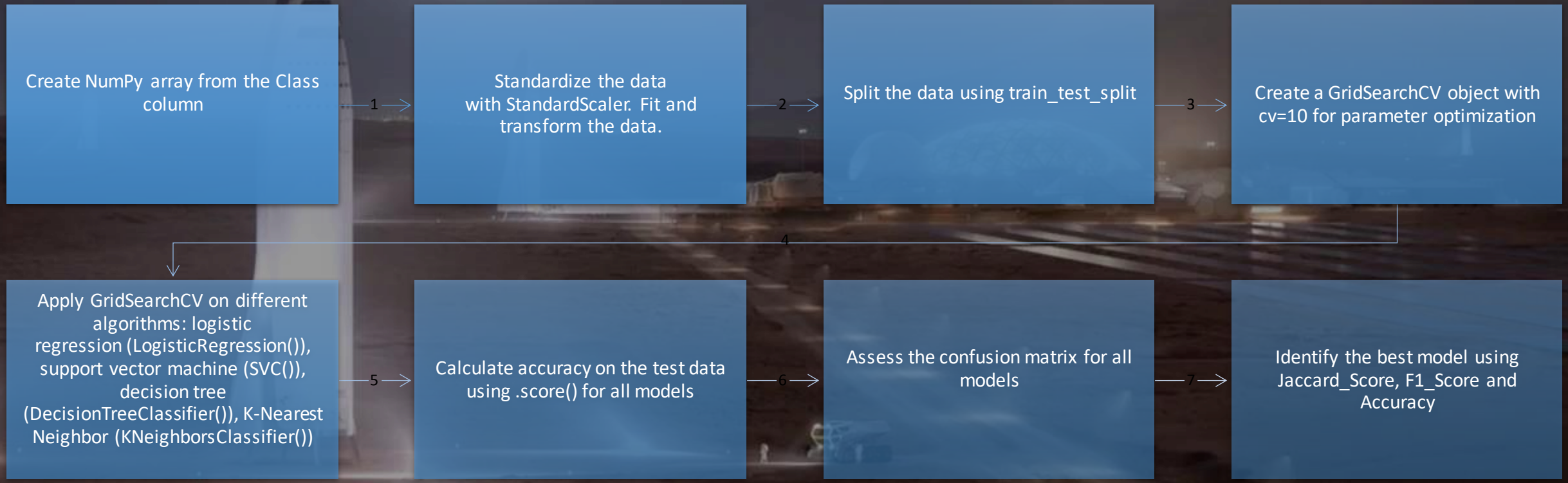
- Allow user to see successful and unsuccessful launches as a percent of the total

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

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

Github - SpaceX - Dash

Predictive Analysis (Classification)



[Github - SpaceX - Predictive]

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

Visual Analytics

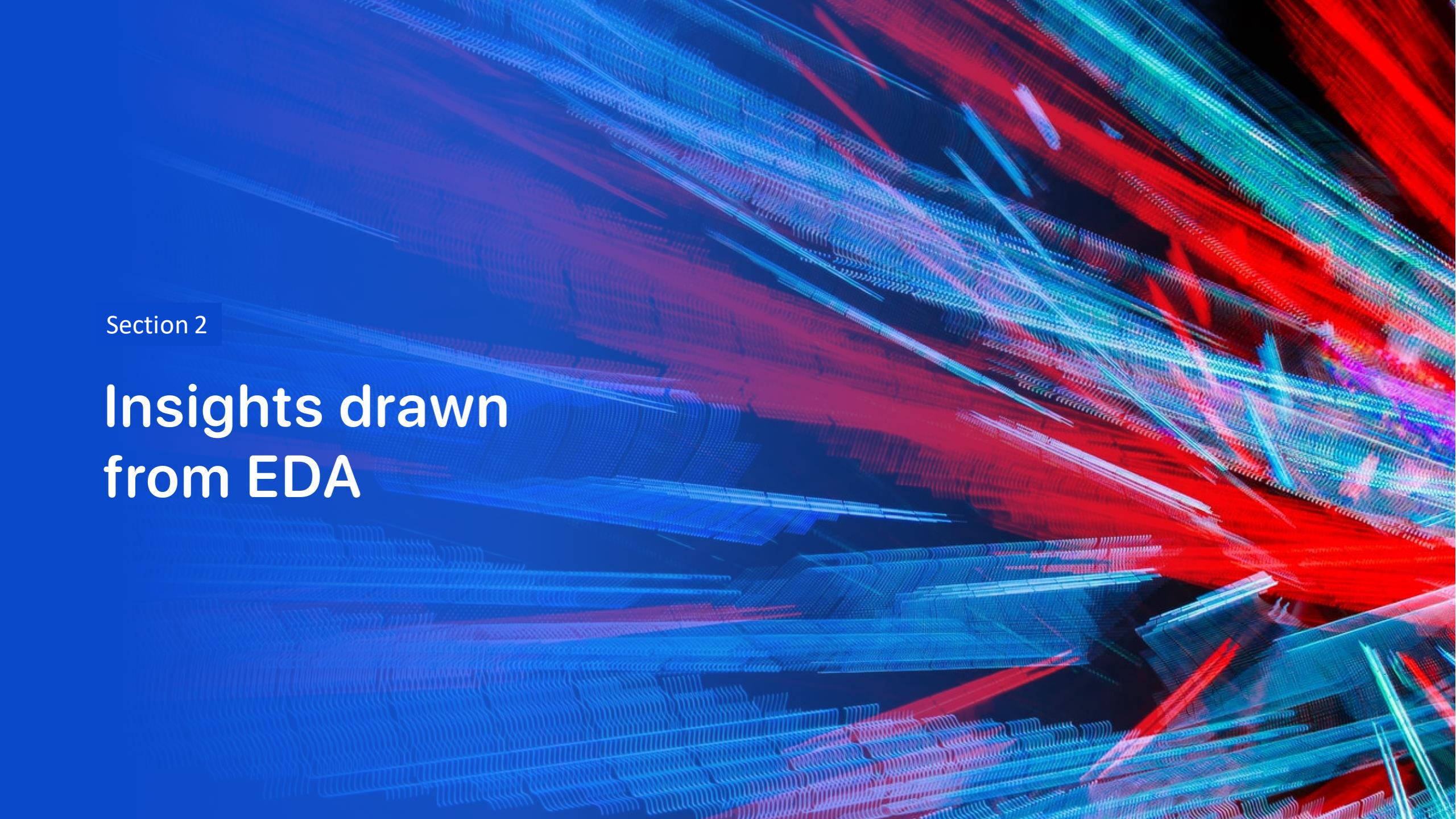
- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

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

Section 2

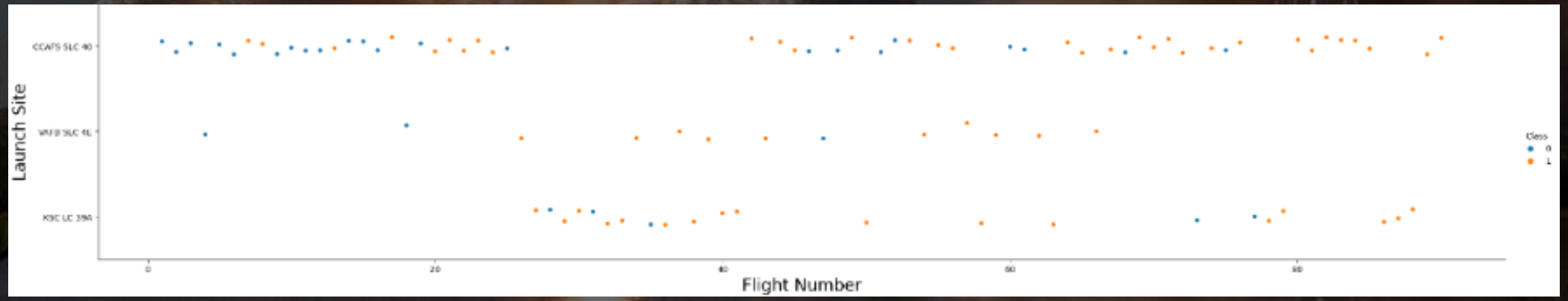
Insights drawn from EDA



Flight Number vs. Launch Site

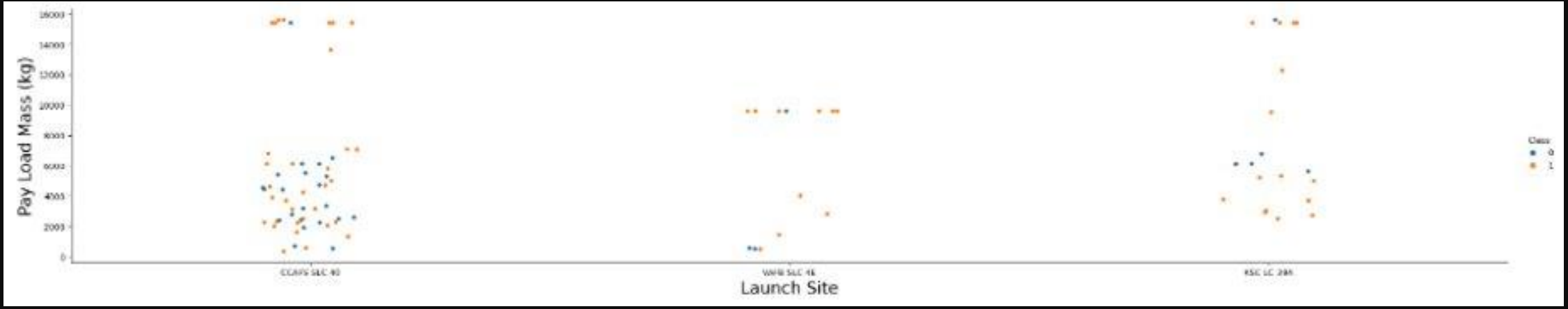
- **Exploratory Data Analysis:**

- 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



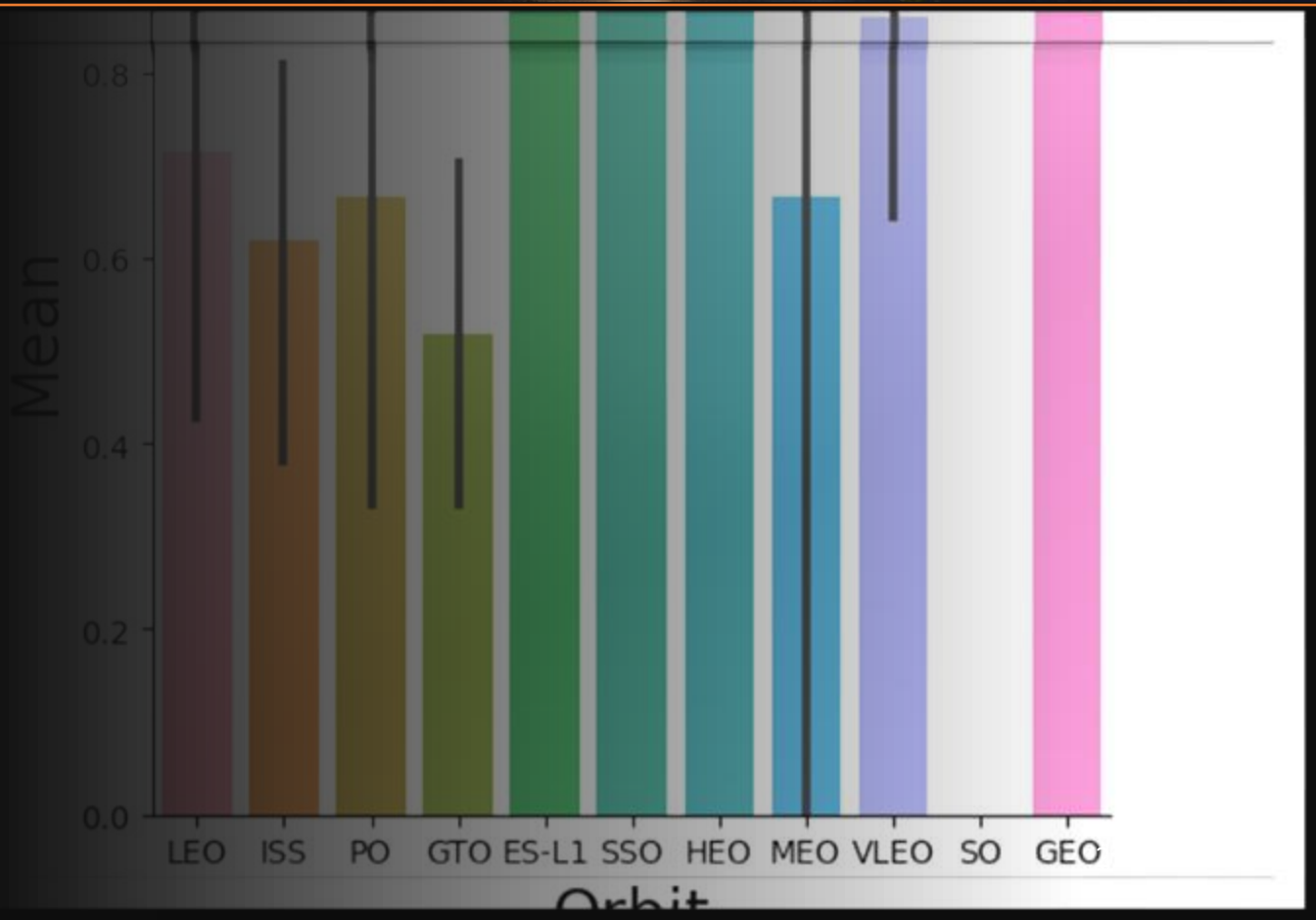
Payload vs. Launch Site

- **Exploratory Data Analysis:**
- 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 k



Success Rate vs. Orbit Type

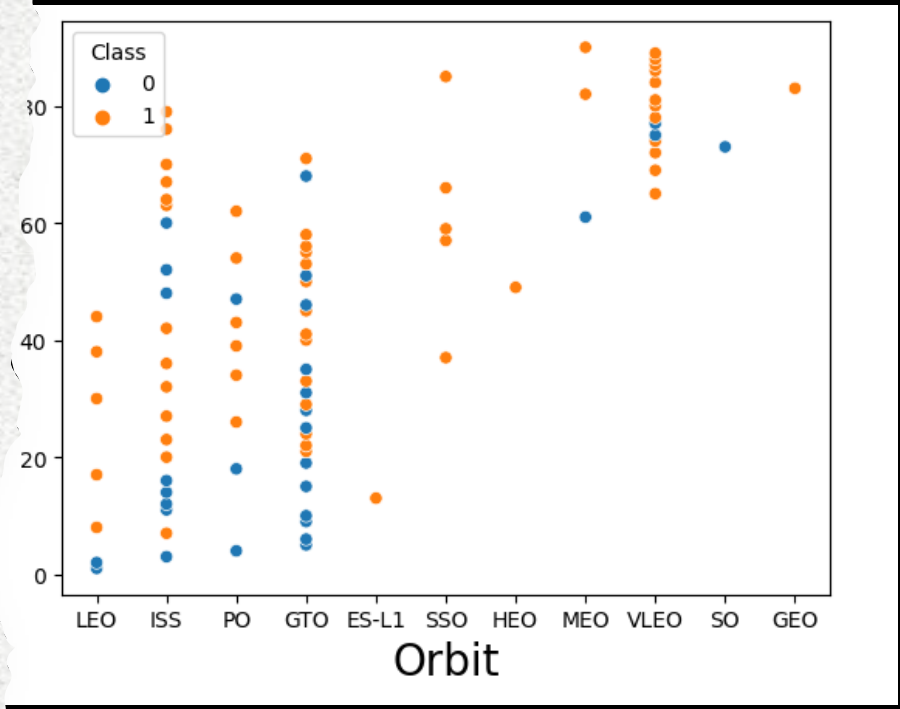
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 80%-50% Success Rate: GTO, ISS, LEO, MEO PO
- 0% Success Rate: SO



Flight Number vs. Orbit Type

Analysis:

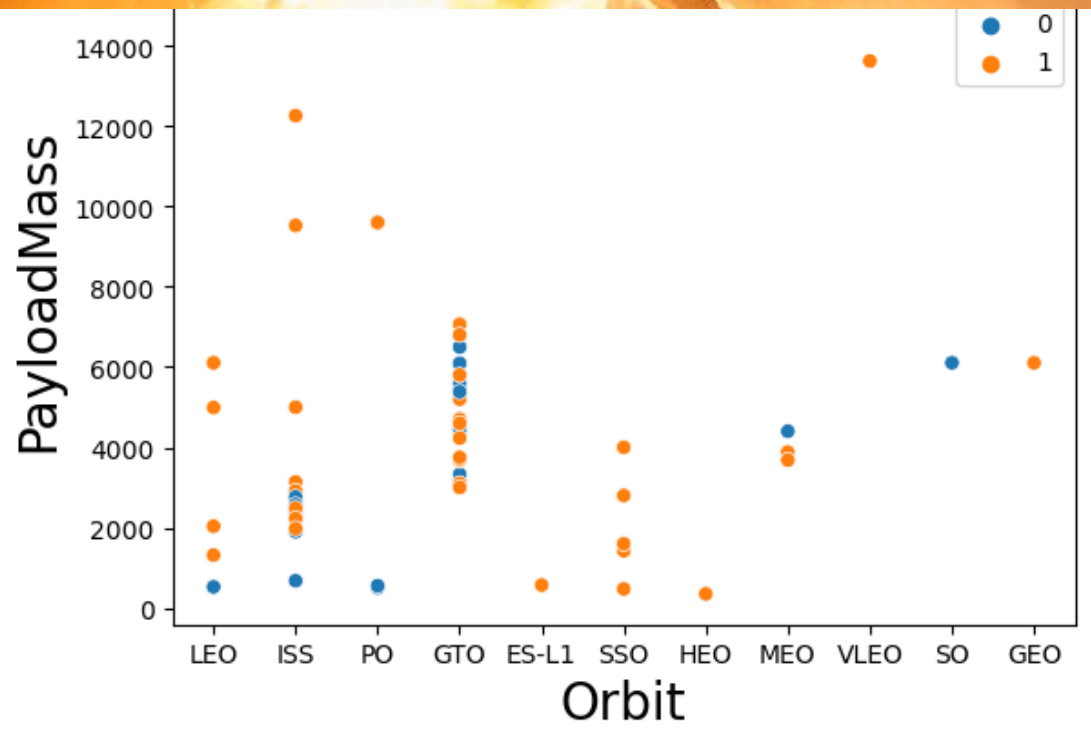
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, does not follow this trend though

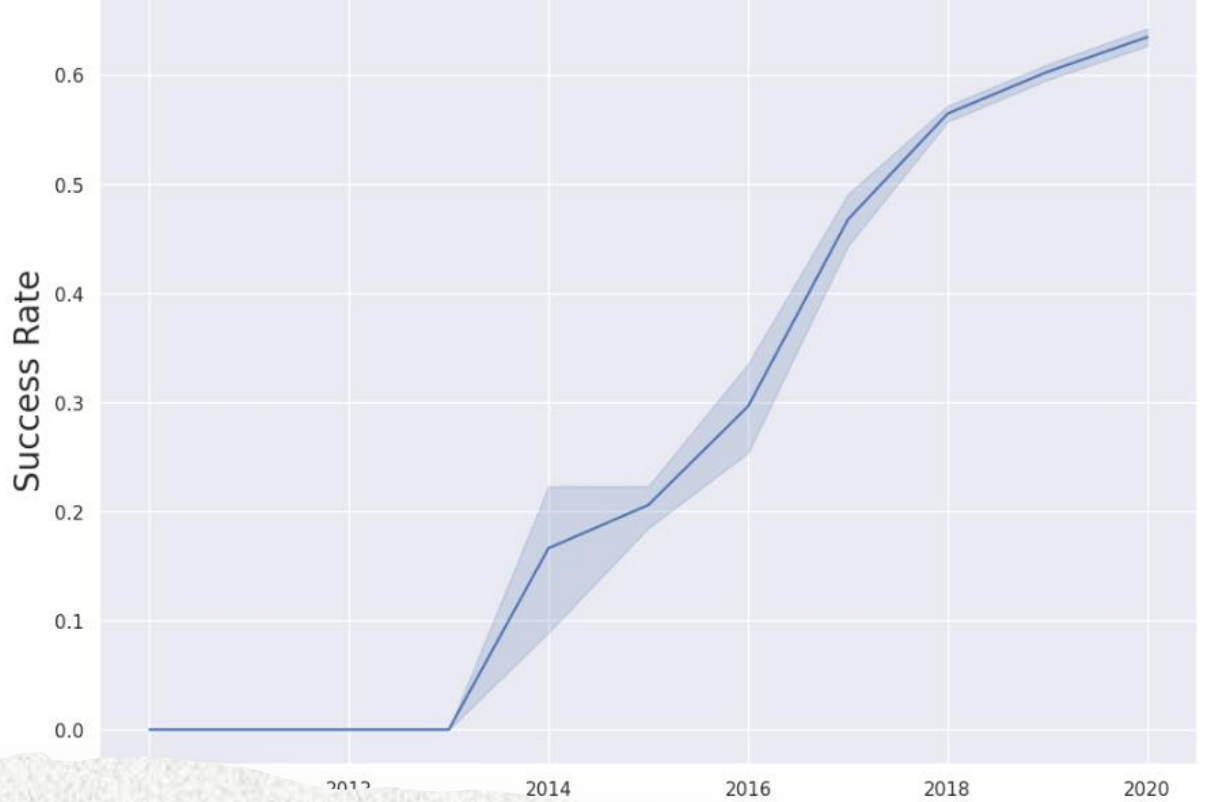


Payload vs. Orbit Type

Analysis:

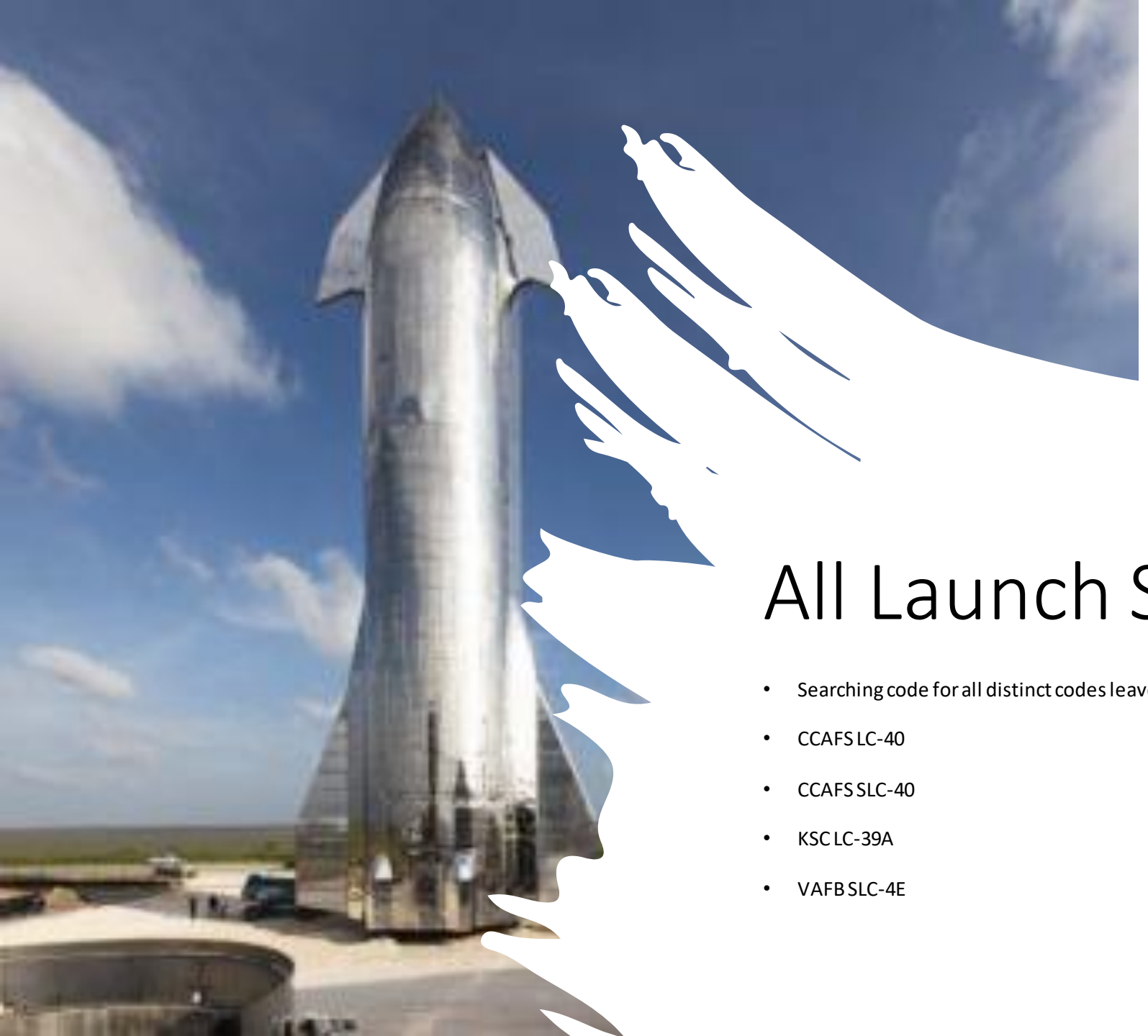
- Heavy payloads are better with the LEO, ISS and PO orbits
- The GTO orbit has had mixed success with heavier payloads





Launch Success Yearly Trend

- Success rate improved significantly from 2013-2017 and 2018-2019
- Success rate decreased from 2017-2018 and from 2019-2020
- But overall the success rate has continued improving since 2013



```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

All Launch Site Names

- Searching code for all distinct codes leave us with these 4 launch sites
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Code to find 5 records where launch sites begin with `CCA`

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Get the sum of all values for NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- Using MIN to find first date, as it is the minimum date
- 2015-12-22


```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The payload mass data was taken by those greater than 4000 but less than 6000, with the outcome determined to be "success drone ship"


```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AN
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

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

```
%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

Boosters Carried Maximum Payload

- Using MAX to see which booster carried Maximum Payload Mass

```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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
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

2015 Launch Records

- 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)

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

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

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

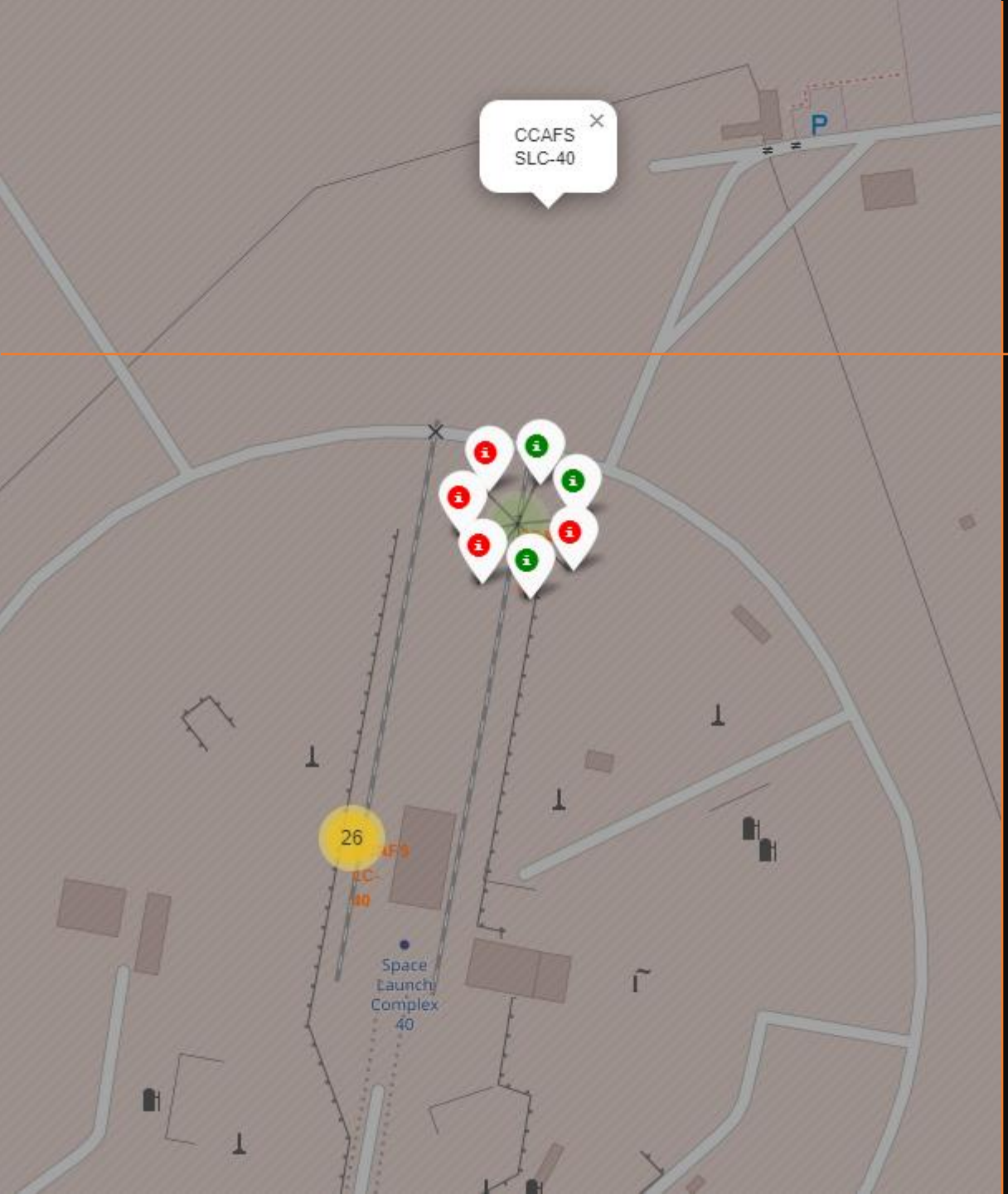
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

Launch Sites

With Markers

- Near Equator: the closer the launch site is 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

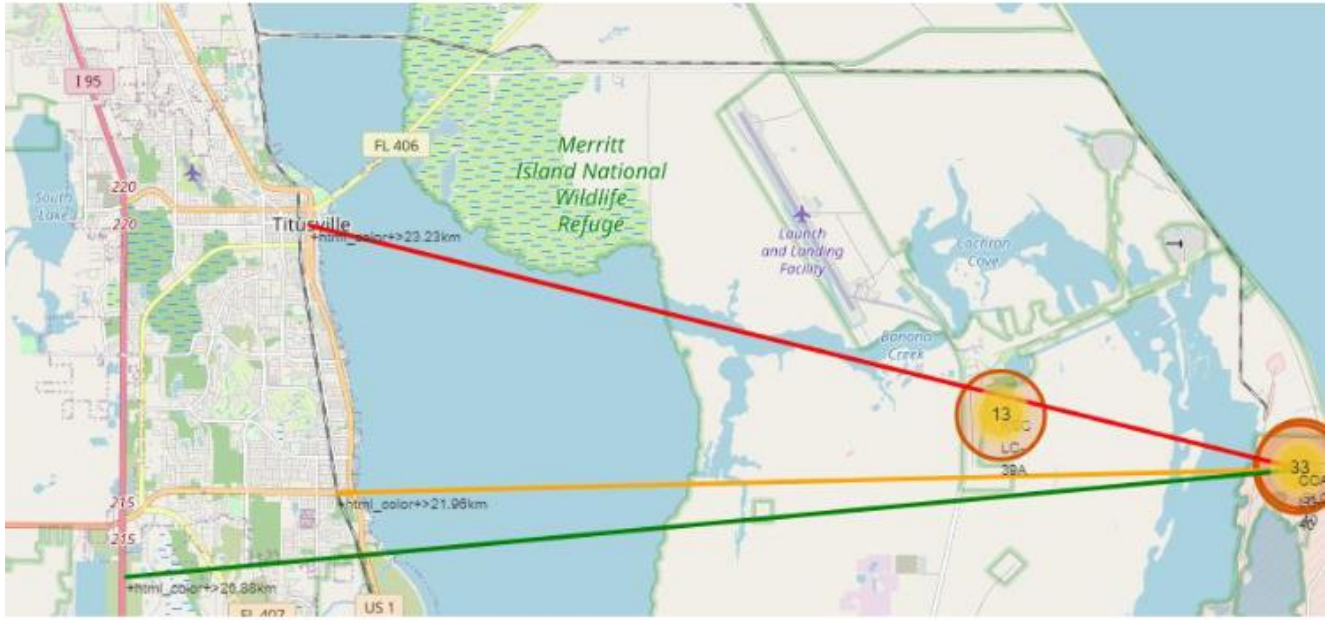


Launch Outcomes

At Each Launch Site

Outcomes:

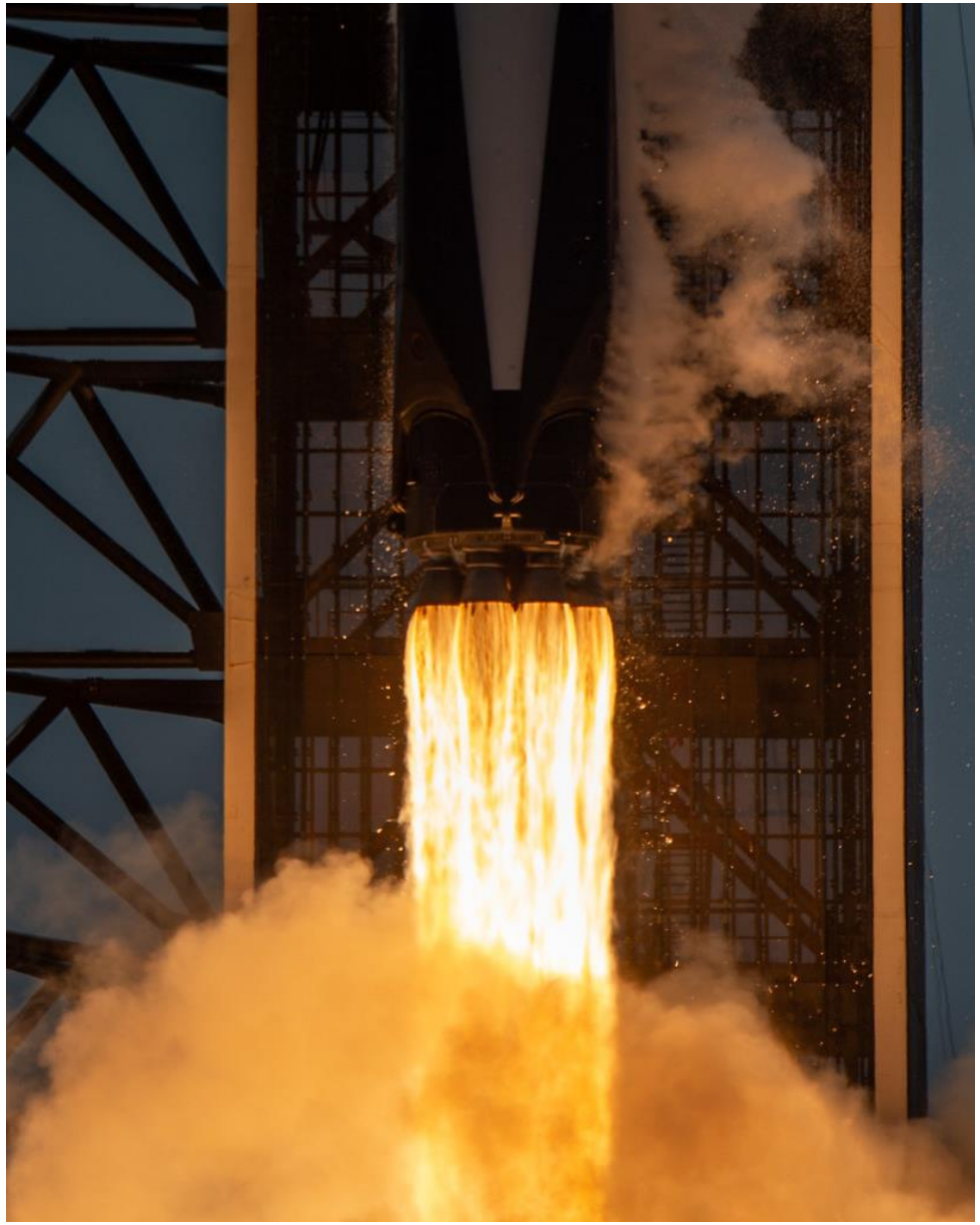
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



Distance to Proximities

CCAFS SLC-40

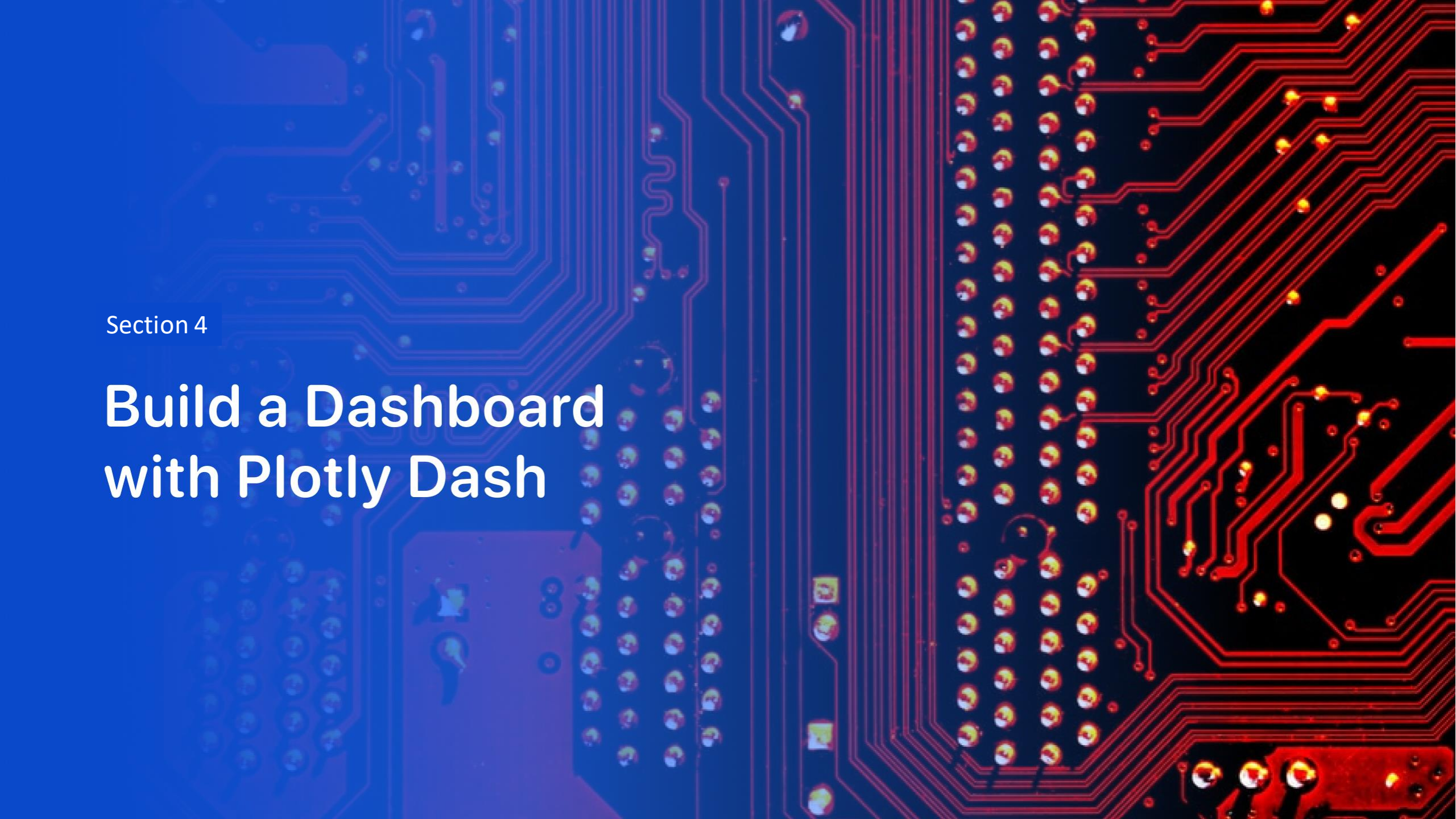
- **.86 km from the nearest coastline**
- **21.96 km from the nearest railway**
- **23.23 km from the nearest city**
- **26.88 km from the nearest highway**



Distance to Proximities Part 2

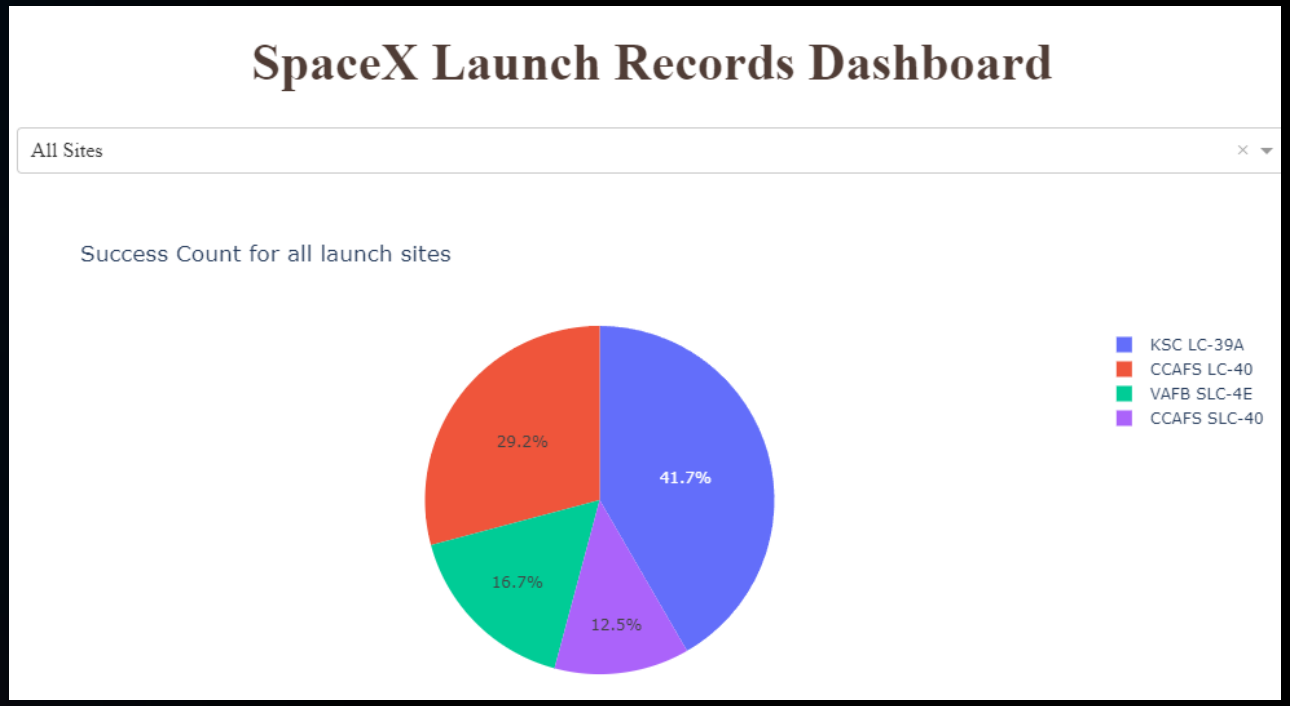
CCAFS SLC-40

- Coasts: Because it helps ensure that the spent stages drop along the launch path or failed launches don't fall on people or property.
- Safety/Security: Because there needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.
- Transportation/Infrastructure and Cities: Because site needs to be away from anything a failed launch could damage, but still close enough to roads/rails/docks to be able to bring people and material to or from site in support of launch activities.


Section 4
Build a Dashboard with Plotly Dash

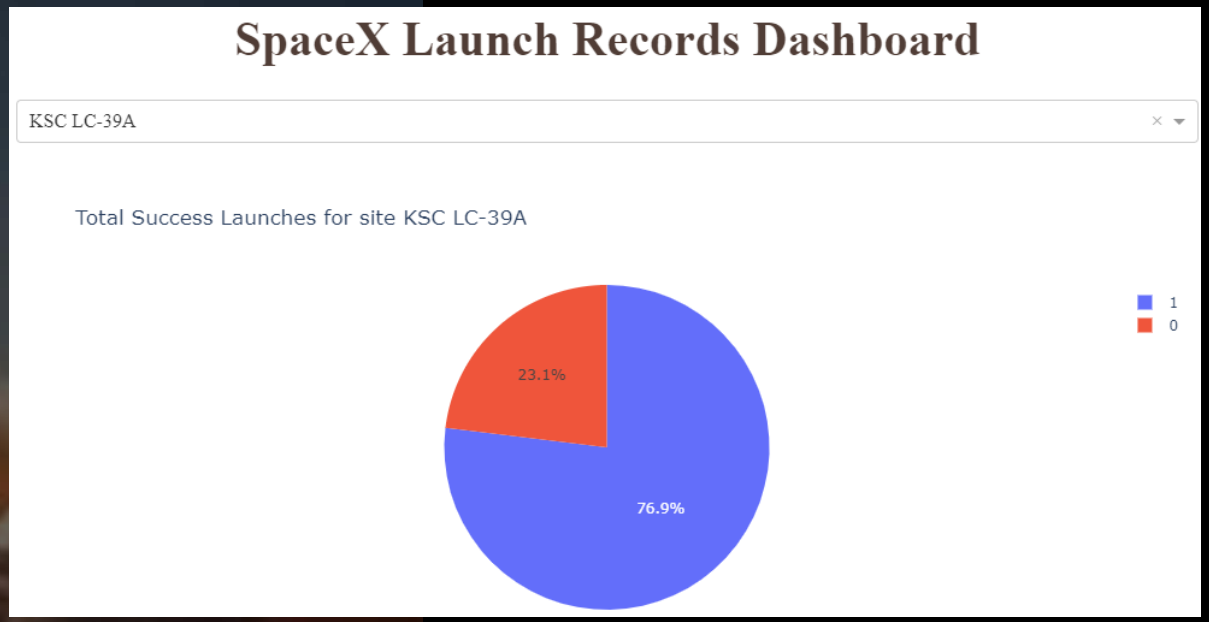
Launch Success by Site

- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success (KSC LC-29A)

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

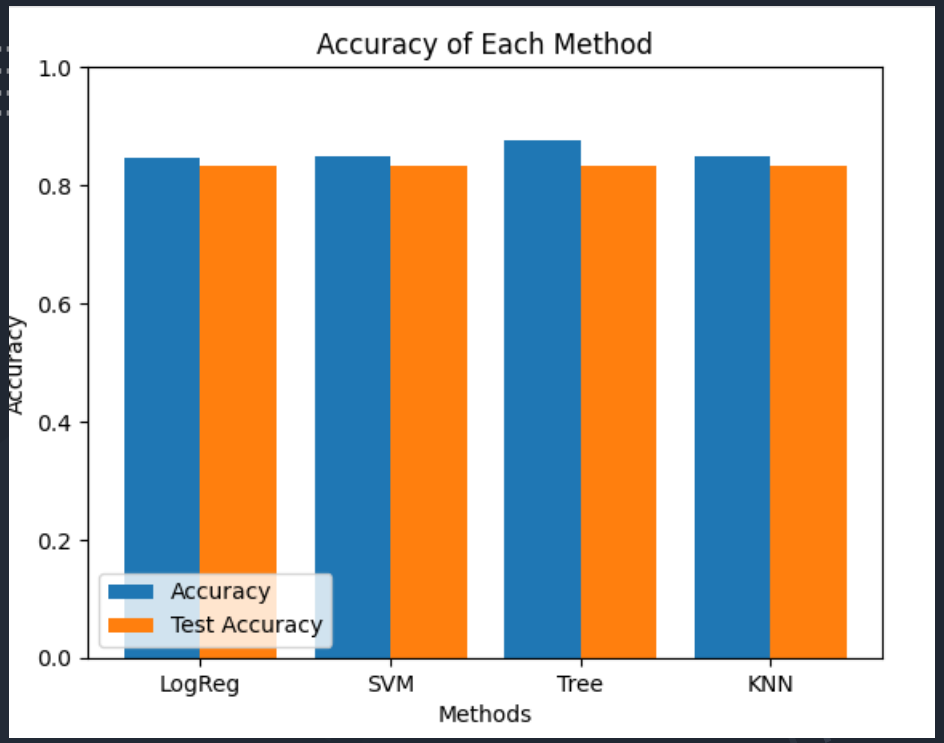


Payload Mass and Success

- 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

Section 5

Predictive Analysis (Classification)




```
print("Model\t\tAccuracy\tTestAccuracy")#,logreg_cv.best_score_)
print("LogReg\t\t{}\t\t{}".format((logreg_cv.best_score_).round(5), logreg_cv.score(X_test, Y_test).round(5)))
print("SVM\t\t{}\t\t{}".format((svm_cv.best_score_).round(5), svm_cv.score(X_test, Y_test).round(5)))
print("Tree\t\t{}\t\t{}".format((tree_cv.best_score_).round(5), tree_cv.score(X_test, Y_test).round(5)))
print("KNN\t\t{}\t\t{}".format((knn_cv.best_score_).round(5), knn_cv.score(X_test, Y_test).round(5)))

comparison = {}

comparison['LogReg'] = {'Accuracy': logreg_cv.best_score_.round(5), 'TestAccuracy': logreg_cv.score(X_test, Y_test).round(5
comparison['SVM'] = {'Accuracy': svm_cv.best_score_.round(5), 'TestAccuracy': svm_cv.score(X_test, Y_test).round(5)}
comparison['Tree'] = {'Accuracy': tree_cv.best_score_.round(5), 'TestAccuracy': tree_cv.score(X_test, Y_test).round(5)}
comparison['KNN'] = {'Accuracy': knn_cv.best_score_.round(5), 'TestAccuracy': knn_cv.score(X_test, Y_test).round(5)}
```

```
Model           Accuracy        TestAccuracy
LogReg          0.84643         0.83333
SVM             0.84821         0.83333
Tree            0.875           0.83333
KNN             0.84821         0.83333
```

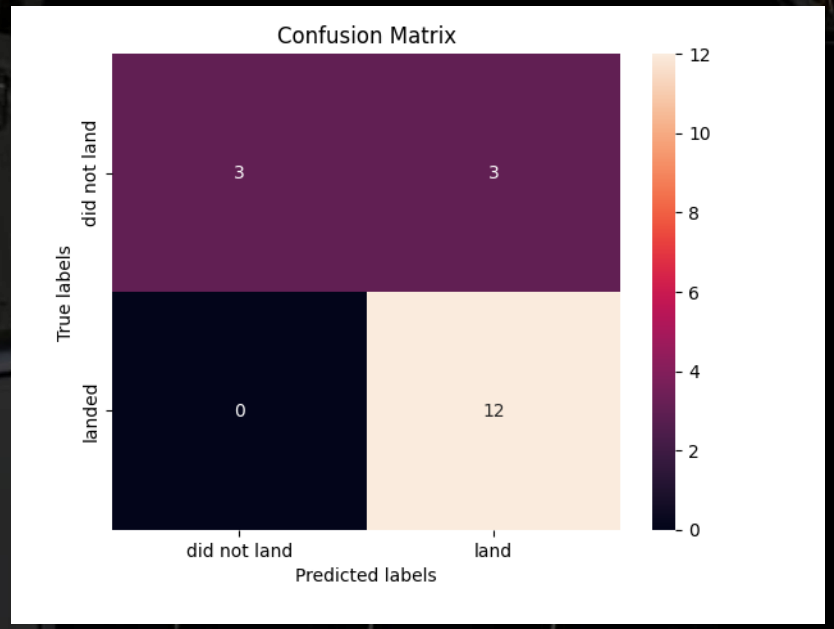
Classification Accuracy

- All models performed at a very similar level. If we go by Test Accuracy alone, all models have the same scores (0.83333). This is likely due to the small dataset.
- If you look at overall Accuracy and best scores, the Decision Tree model slightly outperforms at 0.875.
- Accuracy uses the .best_score_ function which is the average of all cv folds for a single combination of the parameters

Confusion Matrix

Summary

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that false positives (Type 1 error) are present, 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



Conclusions

Model Performance: The models performed similarly on the test set but the decision tree slightly outperformed.

Equator: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of the Earth – which helps save the cost of putting in extra fuel and boosters.

Coast: All launch sites are close to the coast.

Launch Success: Increases over time

KSC LC-39A: Has had the highest success rate among the 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

Conclusions pt 2

- Dataset: At this point a larger dataset could help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set.
- PCA: Principal component analysis should be conducted to see if it can help improve accuracy.

Thank you

To see more work please Github - Capstone Project

Thank you!