WINNING SPACE RACE
| Data Science

Herman Pardo
4/20/23

IBM Developer
SKILLS NETWORK

Table Of Contents

Executive Summary

Summary of Methodologies:

The objective of the research was to identify the factors that contribute to a successful rocket landing. The following methodologies were employed for this purpose:

- Data Collection
  - SpaceX API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) with SQL & Data visualization
- Interactive Map with Folium
- Dashboard with Plotly Dash
- Machine Learning Prediction

Summary of Results:

- EDA
  - Improvement in launch success over time
  - KSC LC-39A having the highest success rate among landing sites
  - Orbits GEO, HEO, ES-L1, and SSO having the best success rate

- Visualization and analytics
  - All launch sites are on a coast and most launch sites are located near the equator. Core factors to consider for future lunch sites.

- Predictive Analytics
  - The Tree Classifier Algorithm is the best method for machine learning for this dataset.

Introduction

Background

- SpaceX is a pioneer in the space industry and has set its sights on making space travel accessible and affordable to the masses. Their triumphs consist of sending spacecraft to the International Space Station, launching a massive satellite constellation that provides internet access to even the most remote locations on Earth, and sending astronauts on manned missions to space.

- One of the main reasons SpaceX has been so successful is due to significantly being able reduce the cost of rocket launches. The innovative reusing of the first stage of the Falcon 9 rocket has dramatically reduced the cost of each launch to just $62 million compared to other providers that costs upwards of $165 million per launch.

- The key to determining the price of a launch is whether the first stage will land. Using public data and machine learning models, it is possible to predict whether SpaceX will be able to reuse the first stage. This data-driven approach helps to make space travel more affordable and accessible to everyone, enabling more people to explore the cosmos and achieve their dreams of space exploration.

Discover

- The relationship between launch site, mass, number of flights, and orbit types affect the success of landing for the first stage rocke

- The landing success rate over the lifetime of the company

- conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

METHODOLOGY

Methodology

Collect
- SpaceX REST API
- Web Scraping Techniques

Wrangle
- Clean data for handling
- One Hot Encoding: prepare data for machine learning modeling

Explore
- Exploratory Data Analysis (EDA) with SQL and data visualization
- Variable relationships through scatter, bar, and line graphs

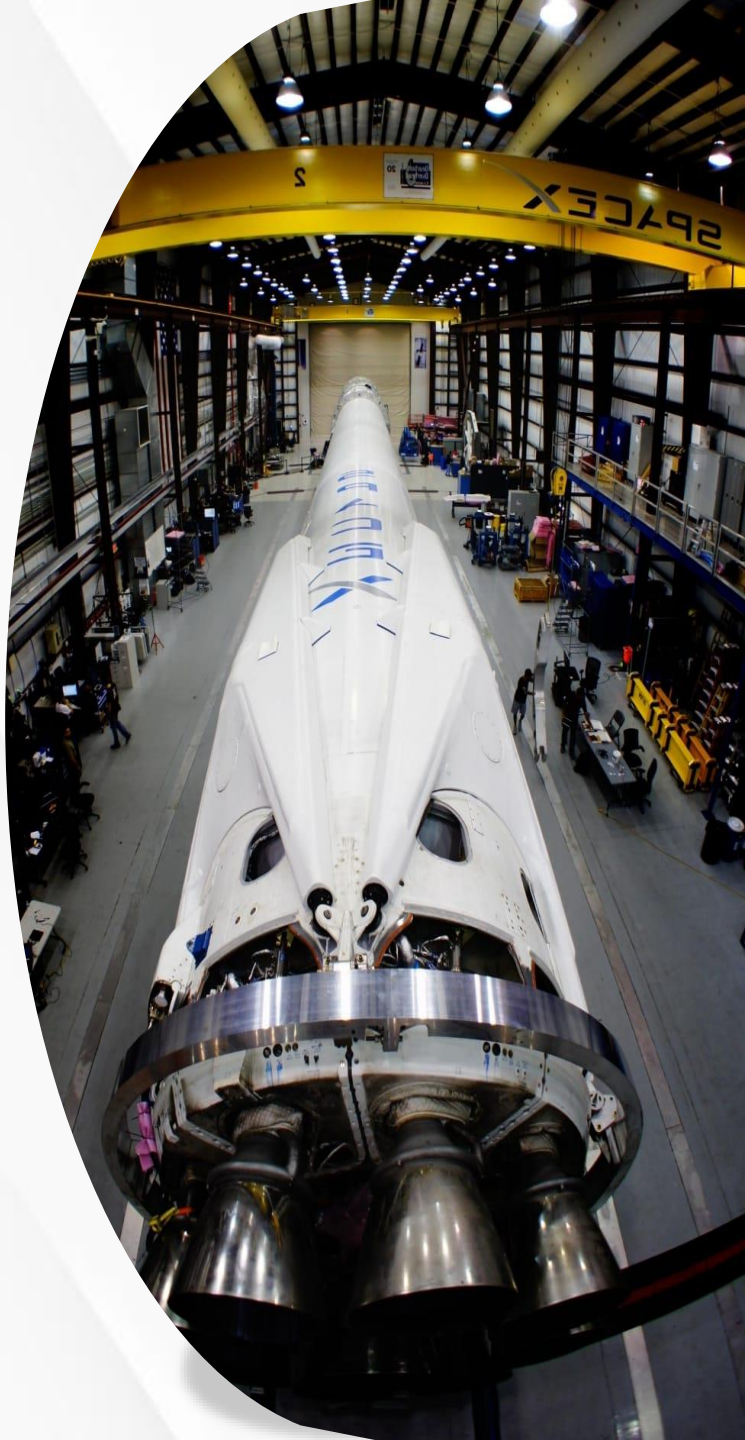
Visualize
- Folium
- Plotly Dash

Machine Learning
- Predict landing outcomes through machine learning
- Build, tune, and evaluate classification models

Data Collection – SpaceX API

- Request SpaceX data with REST API
- API fetches SpaceX rocket launch data in .json format
- Use custom functions to get data lists needed
- Assign list to dictionary
- Assign dictionary to dataframe
- Filter dataframe for Falcon 9 launches only
- Replace missing values
- Export data to .csv file

GitHub Notebook URL



Data Collection – Web Scrapping

- Request SpaceX launch data from Wikipedia
- Create BeautifulSoup object from HTML response
- Find tables in the site
- Extract column names from table header
- Collect data from parsing tables
- Create dictionary from the data
- Create dataframe from dictionary
- Export dataframe to csv file

GitHub Notebook URL

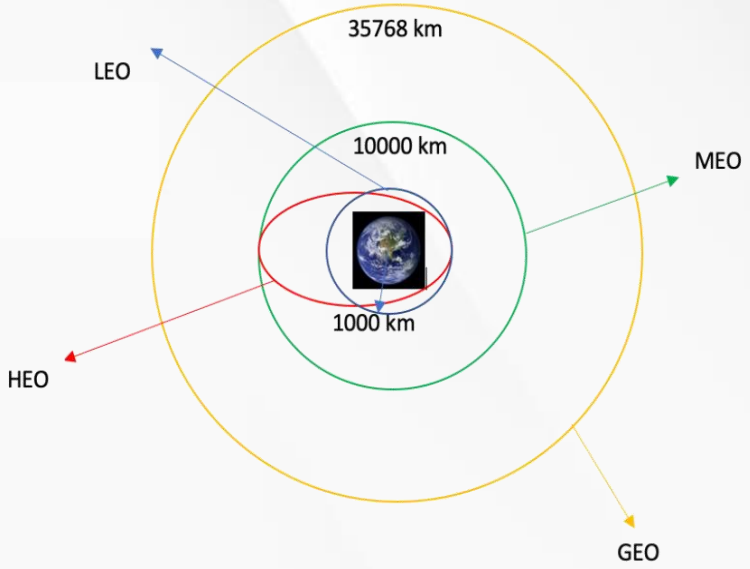
Data Wrangling

Process:

- Perform exploratory data analysis and determined the training labels.
- Calculate:
  - Number of launches at each site
  - number and occurrence of each orbits and outcome
- Create landing outcome label from outcome column and exported the results to a csv file
- Convert outcomes into binary:
  - 1 for successful landing
  - 0 for unsuccessful landing

Landing Outcomes	Description
True Ocean	Mission outcome had a successful landing to a specific region of the ocean
False Ocean	Unsuccessful landing to a specific region of ocean
True RTLS	Successful landing on a ground pad
False RTLS	Unsuccessful landing on a ground pad
True ASDS	Successful landing on a drone ship
False ASDS	Unsuccessful landing on drone ship

Common orbit types for rocket launches:



GitHub Notebook URL

EDA – Visualization

Scatter Graphs

Visualize the correlation between two variables and how much one variable is affected by another. These variables may aid machine learning. The following scatter graphs will be presented:

- Flight Number VS. Payload Mass
- Flight Number VS. Launch Site
- Payload VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type
- Orbit VS. Payload Mass

Bar Graph

Easy way to observe the comparison between different groups of data. The following bar graph will be presented:

- Mean vs. Orbit

GitHub
Notebook URL

Line Graph

Method that connects a series of data points to show trends and data variables that help make predictions. The following line graph will be presented:

- Success Rate vs Year

EDA – SQL

10 SQL queries were executed to display and list results as follows:

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.

GitHub Notebook URL

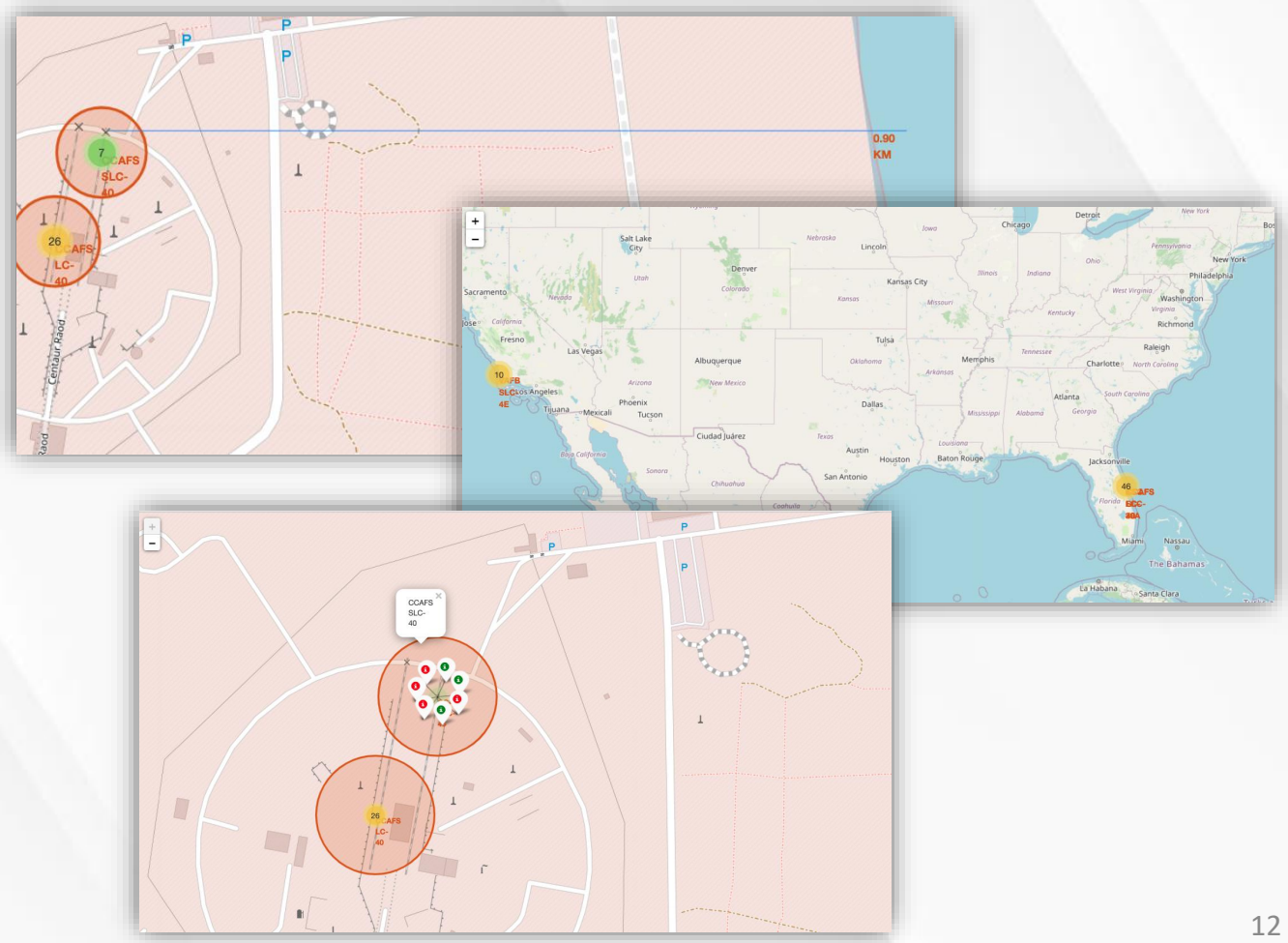
List

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship 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 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)

Map – Folium

Using Folium in Python, the following objects were placed on a map to illustrate the extracted launch data with insight:

- Red circle markers are placed around each launch site using the GPS coordinates. Also added labels to each site to better identify locations.
- Green and Red markers are used at the launch site to represent the successful and failed launch outcomes
- Lines are used on the map to measure distance from CCAFS SLC-40 launch site to landmarks using Haversine's formula. These lines show the nearest city, highway, railway, and coastline in order to easily visualize the surrounding environment. Such data can be used to create requirements for building a rocket launch facility from a safety point of view.



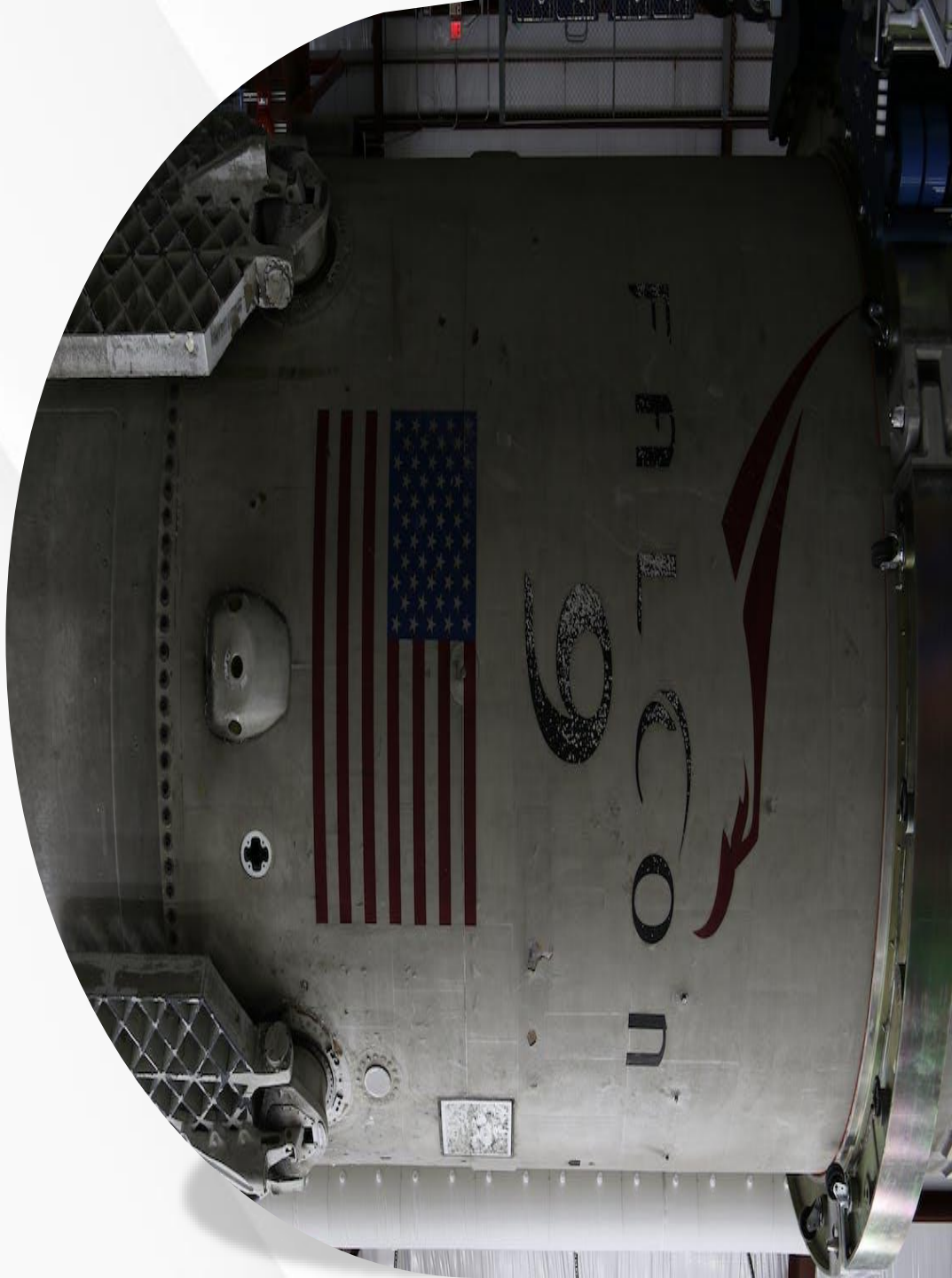
GitHub
Notebook URL

Dashboard – Plotly

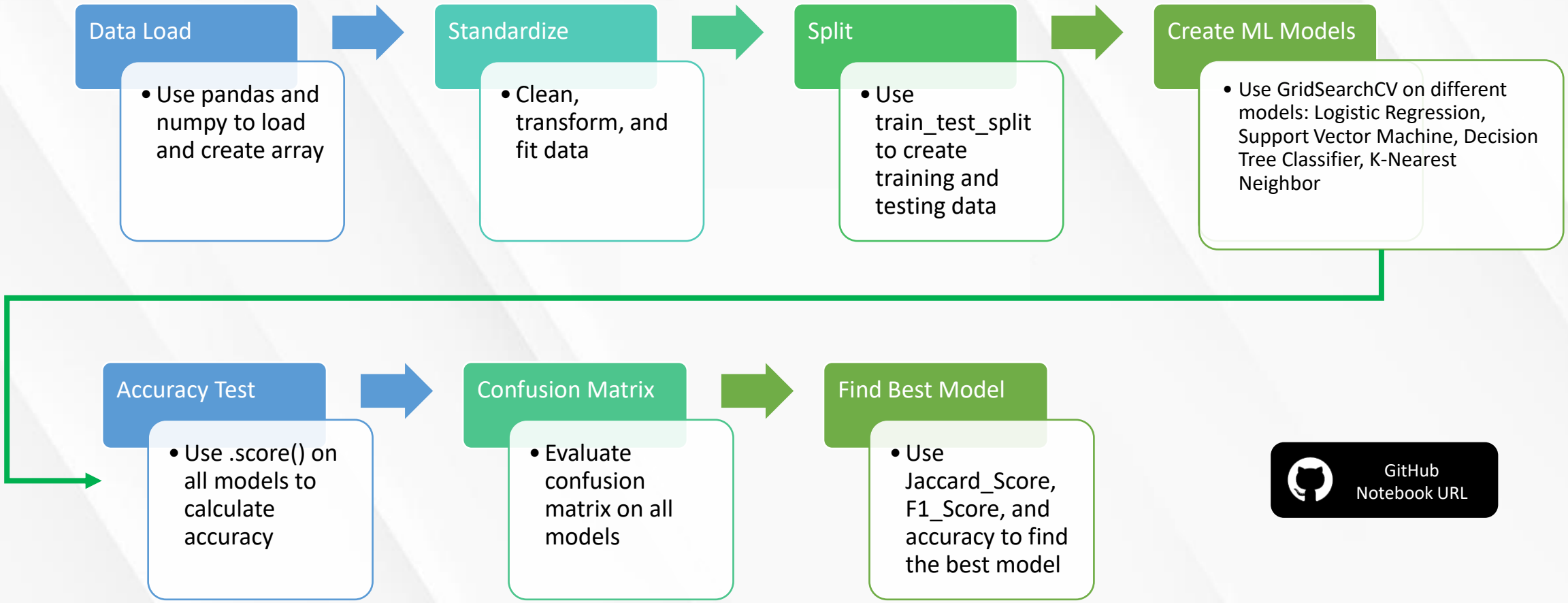
Interactive dashboard was built using Plotly:

- Dropdown menu with launch sites available to be selected
- Pie charts show the total launches by a certain sites
- Interactive slider object allows the selection of payload mass change
- Scatter graph shows the success rate relationship between Outcome and Payload Mass by booster version

GitHub
Notebook URL



Predictive Analytics (Classification)



GitHub Notebook URL

Results Summary

Exploratory Data Analysis

- Falcon 9 rocket launch success has improved over time
- KSC LC-39A has the highest success rate of all sites
- Orbits ES-L1, GEO, HEO, and SSO all have a 100% success rate

Interactive Analytics

- Visual mapping reveals that all launch sites are close to a coast as well as most near the equator.
- Due to the explosive nature of rocket launches, launch sites are positioned just far enough away to not disrupt civilian structures

Predictive Analysis

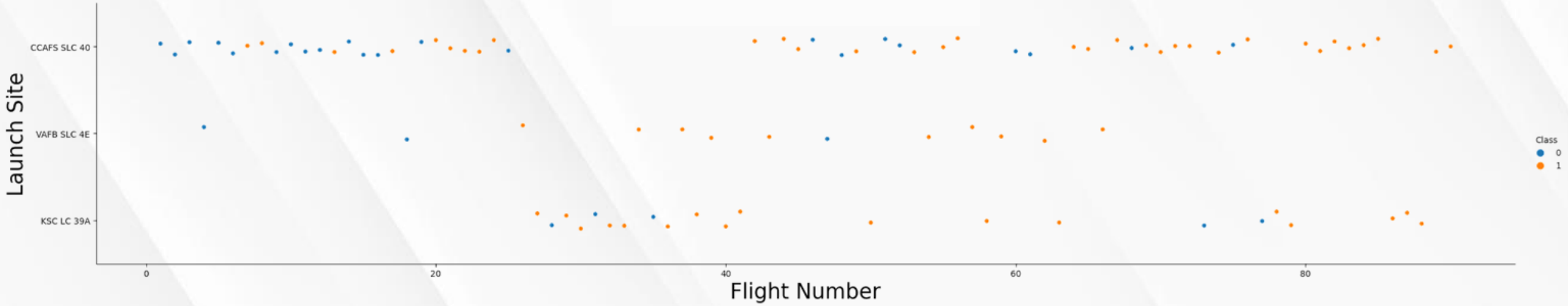
- Evaluation of ML models shows Decision Tree model to be the best predictive model for this data

Section 2

INSIGHTS FROM EDA

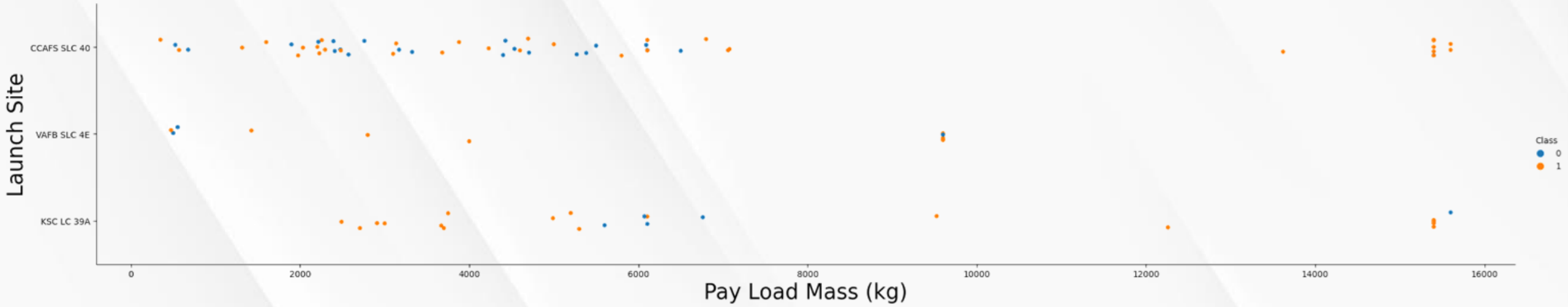
Flight Number vs. Launch Site

- Earlier flights (•) had a lower success rate compared to the later flights (•)
- Launches improved in success rate over time
- Launch site CCAFS SLC-40 has almost 2x amount of the other two combined
- CCAFS SLC-40 has the least success rate of all the launch sites



Payload vs. Launch Site

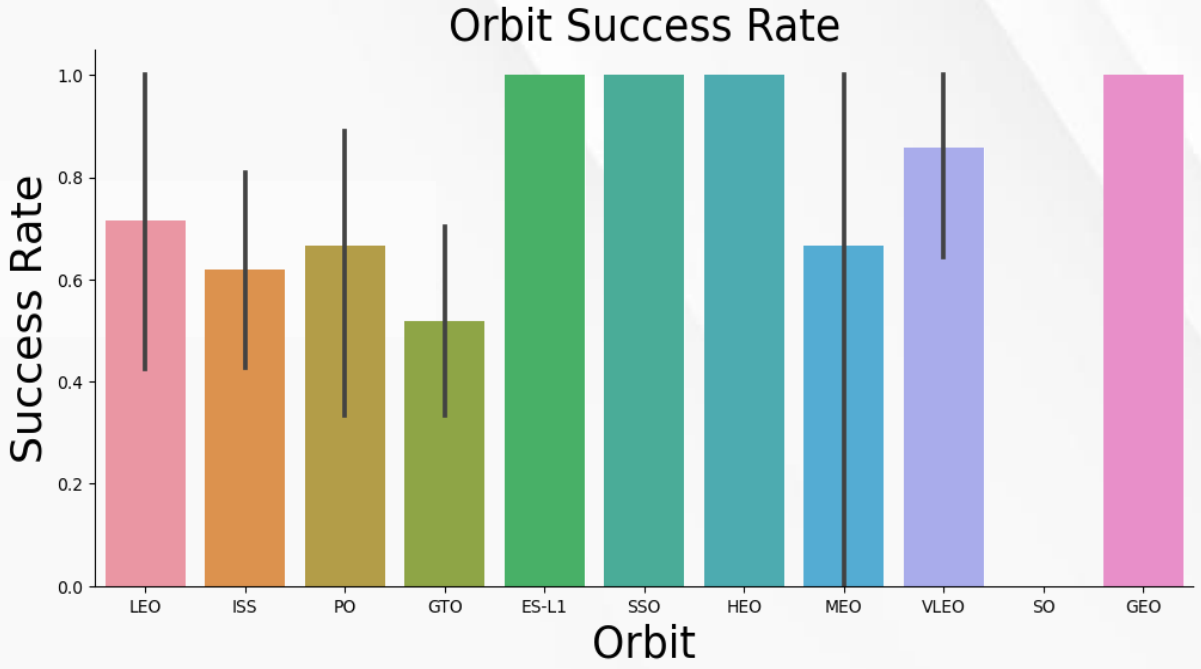
- KSC LC 39A has a launch success rate of 100% for payloads less than 5,500kg
- Higher payloads have greater chance of success rate
- VAFB SLC 4E doesn't launch anything over 10,000kg
- CCAFS SLC-40 has the least success rate of all the launch sites



Success Rate vs. Orbit Type

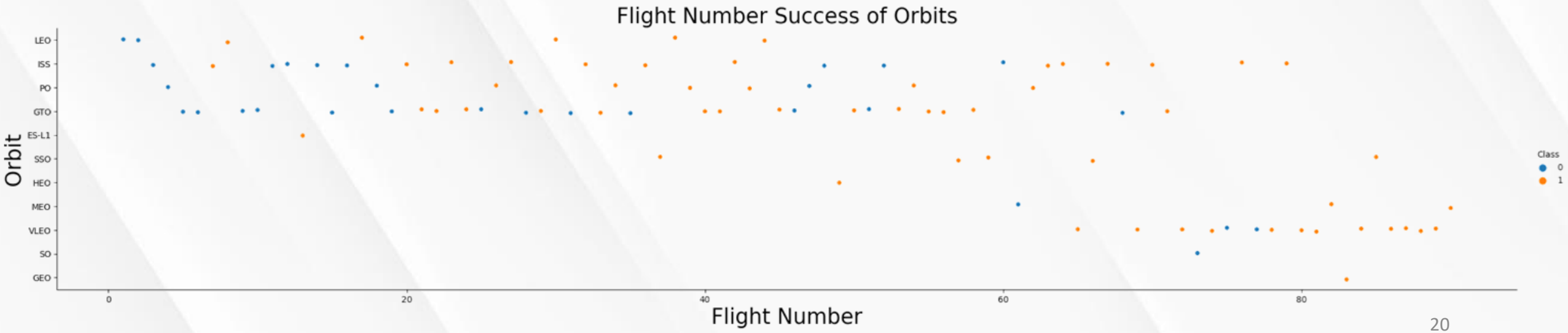
Success Rate:

- 100%: ES-L1, SSO, HEO, GEO
- 90% - 70%: LEO, VLEO
- 70% - 50%: ISS, PO, GTO, MEO
- 0%: SO



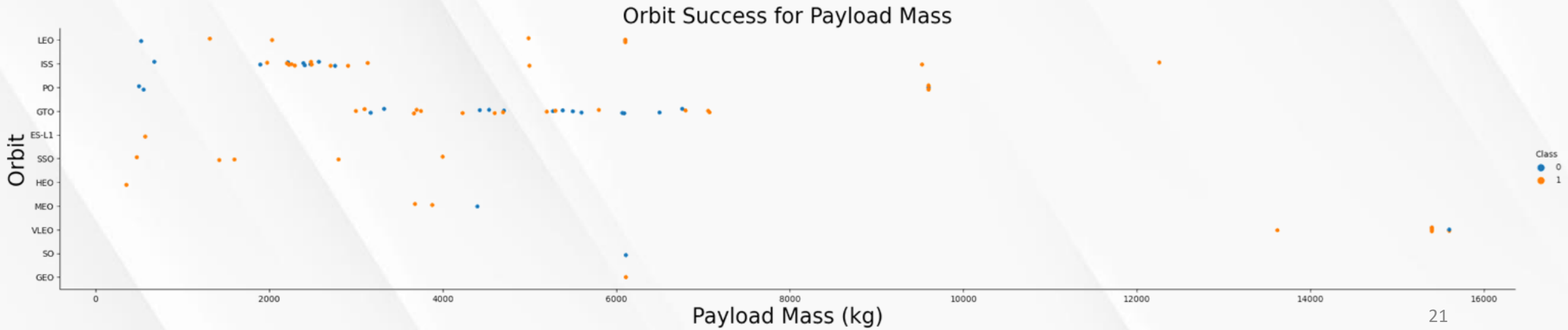
Flight Number vs. Orbit

- Success rate improves with number of launches at all sites
- SO has 100% failure due to only having 1 launch
- GTO is the slowest rate of improvement
- Second half of all flights show significant improvement over first half of all flights



Payload vs. Orbit Type

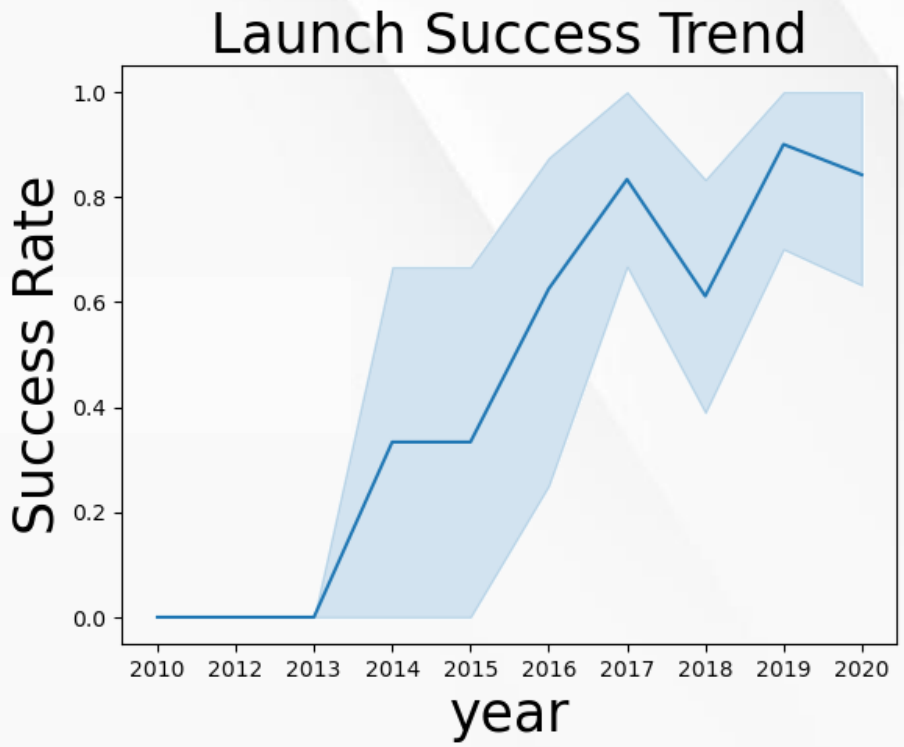
- GTO success rate lowers with heavier payloads
- SSO, MEO, HEO have 100% success rate with payload less than 4,000kg



Launch Success Yearly Trend

EDA:

- Overall rate of success has improved since 2013
- 2017 & 2019 should be examined to improve common points of failures
- Failure rate decreases over time on failed years.



All Launch Site Names

Using SQL in python, all the site names were able to be retrieved:

- There are 4 total launch sites
  - 2 at Cape Canaveral Air Force Station
  - 1 at Kennedy Space Center
  - 1 at Vandenberg Air Force Base

```
%sql SELECT UNIQUE(LAUNCH_SITE) "Launch Sites" FROM SPACEX;
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

Launch Site Names Beginning With ‘CCA’

Using SQL in python, we found 5 site names beginning with ‘CCA’:

- Narrow down the list of launches to only those beginning with ‘CCA’ for Cape Canaveral Air Force Station

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

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
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

Total Payload Mass

Using SQL in python, we calculated the total payload carried by boosters from NASA:

- NASA's total payload mass from launches is 45,596 kg

```
%sql SELECT SUM(PAYLOAD_MASS_KG_) "Total Payload Mass" FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

Using SQL in python, we calculated the Average Payload Mass by F9 v1.1:

- The average payload from booster version 1.1 of a Falcon 9 rocket is 2,928 kg

```
%sql SELECT AVG(PAYLOAD_MASS_KG_) "Average Payload Mass" FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

Average Payload Mass
2928

First Successful Ground Landing Date

Using SQL in python, we calculated the First Successful Ground Landing Date:

- On December 22, 2015, SpaceX had it's first successful landing on ground pad

```
%sql SELECT MIN(DATE) "First Successful Landing Outcome in Ground Pad" FROM SPACEX WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

First Successful Landing Outcome in Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload Between 4,000 kg & 6,000 Kg

Using SQL in python, we found the names of boosters which have successfully landed on drone ship and had payload mass greater than 4,000 kg but less than 6,000 kg:

- There are 4 booster versions that have been able to land successfully on a drone ship after launching with a payload between 4,000 kg and 6,000 kg

```
%sql SELECT BOOSTER_VERSION "Successful Landing Boosters Between 4000 & 6000 Payload" FROM SPACEX WHERE LANDING_OUTCOME = 'Success (drone ship)' AND P
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

Successful Landing Boosters Between 4000 & 6000 Payload
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Using SQL in python, we calculated the total number of successful and failure mission outcomes:

- There have been a total of 100 successful missions (1 where payload status is unclear) and 1 Failure

```
%sql SELECT MISSION_OUTCOME "Mission Outcome", COUNT(*) "Total Number" FROM SPACEX GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

Mission Outcome	Total Number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Using SQL in python, we found the names of the booster which have carried the maximum payload mass:

- There are 12 boosters that have carried the max payload of 15,600 kg

```
%sql SELECT BOOSTER_VERSION "Booster Version", PAYLOAD_MASS_KG_ "Max Payload Mass" FROM SPACEX WHERE (PAYLOAD_MASS_KG_ = (SELECT MAX(PAYLOAD_MASS_KG_)
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

Booster Version	Max Payload Mass
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Using SQL in python, we found the failed landing outcomes on drone ship, their booster versions, and launch site names for in year 2015:

- There are 2 failed landing outcomes on a drone ship in 2015. Both are from the same launch site in almost exactly 3 months apart.

```
%sql SELECT DATE, BOOSTER_VERSION "Booster Version", LAUNCH_SITE "Launch Site", LANDING_OUTCOME "Landing Outcome" FROM SPACEX WHERE LANDING_OUTCOME =
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

DATE	Booster Version	Launch Site	Landing Outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Using SQL in python, we ranked the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

- Of the 31 launches, there are only 8 fully successful missions, a 25.8% success rate.

```
%sql SELECT LANDING_OUTCOME "Landing Outcome", COUNT(*) as "Count Outcomes" FROM SPACEX WHERE DATE BETWEEN '2010-06-04' and '2017-03-20' group by LAND
```

```
 * ibm_db_sa://vhb27331:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

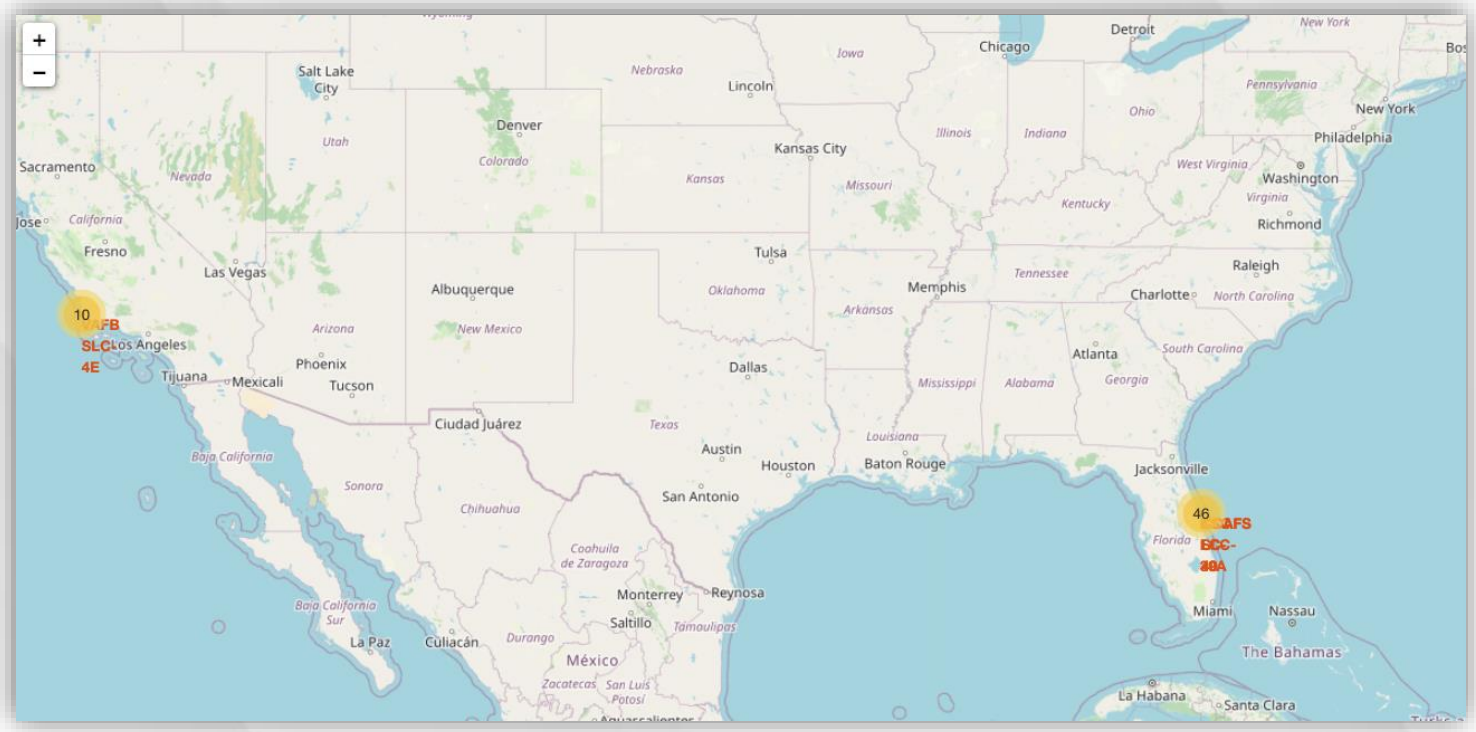
Section 3

LAUNCH SITES PROXIMITIES ANALYSIS

Global Launch Site Map

- Using Folium in Python, a global map of earth is abled to be rendered with the locations of all launch sites.

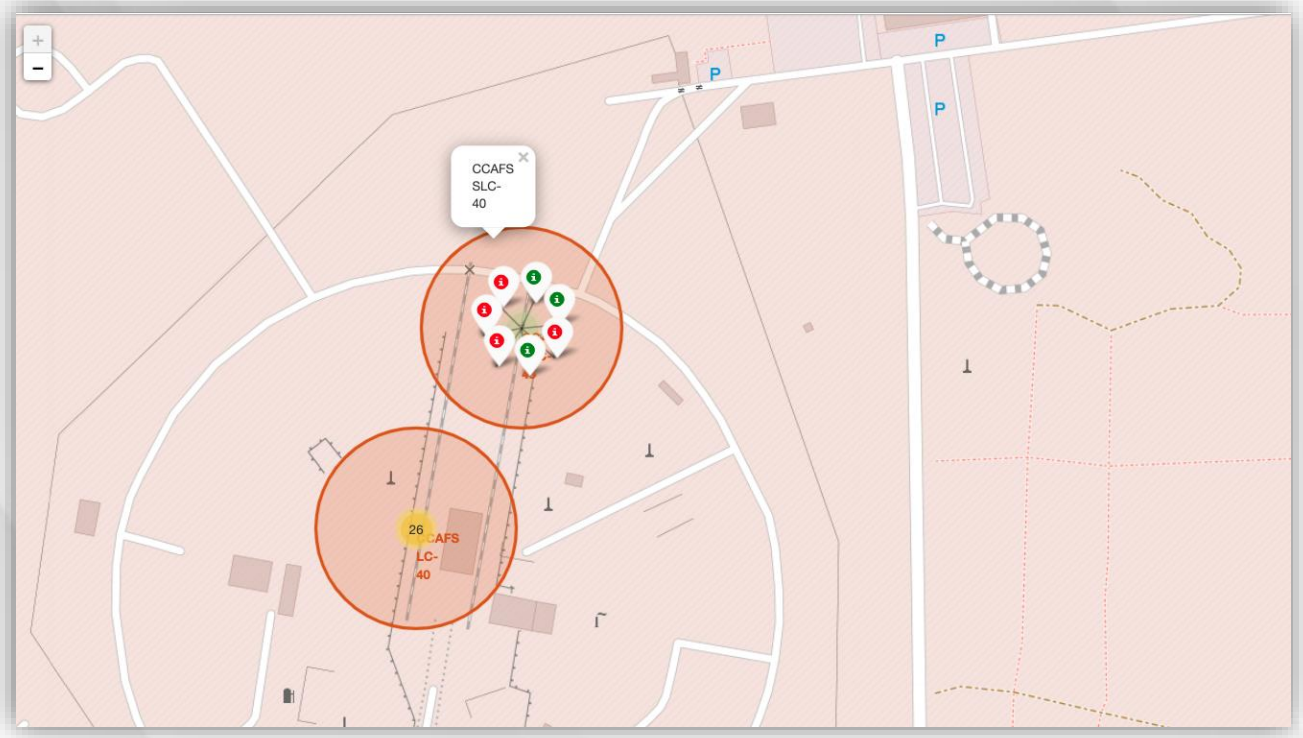
- Common trait between them all is that they are built near coastal waters.



Site Launch Outcome Results Map

- Using Folium in Python, each launch site is outlined by a red circle that when selected, it displays the launch outcomes in green (success) and red (fail) markers

- In the example below, CCAFS SLC-40 launch site has 7 launches, with 3 succeeding and 4 failing



Launch Site Proximities Map

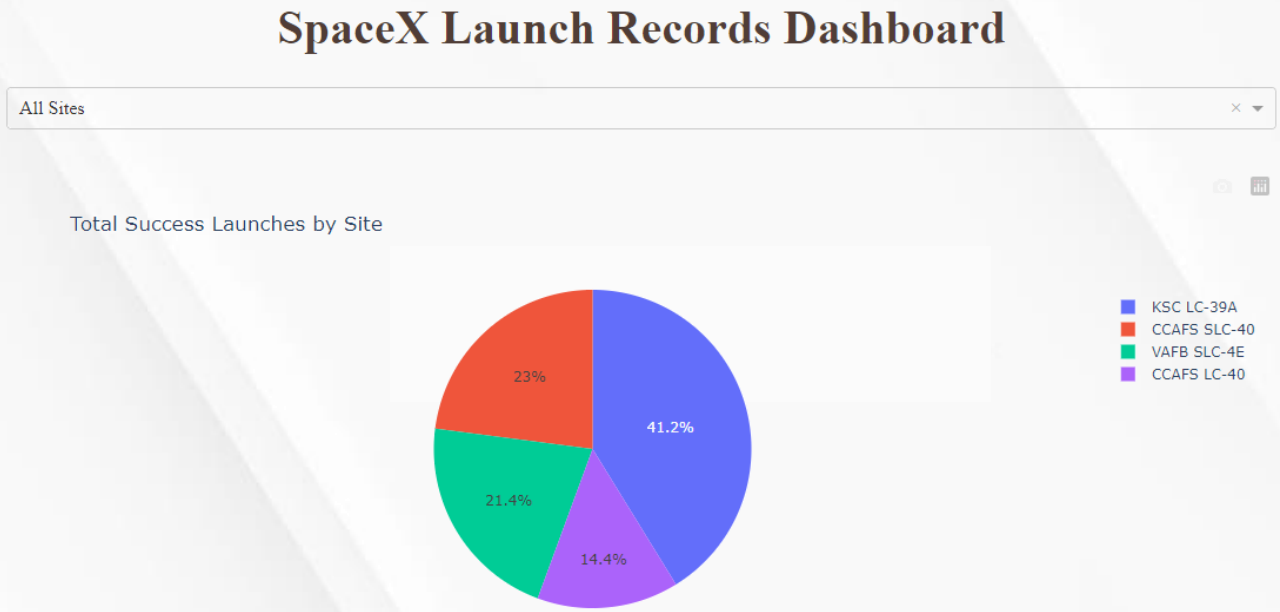
- Using Folium in Python, each launch site shows its proximities to nearest railway, highway, coastline along with the calculated distance displayed

- In this example, the nearest coastline to launch site CCAFS SLC-40 is the Atlantic Ocean located 0.90 KM

Section 4

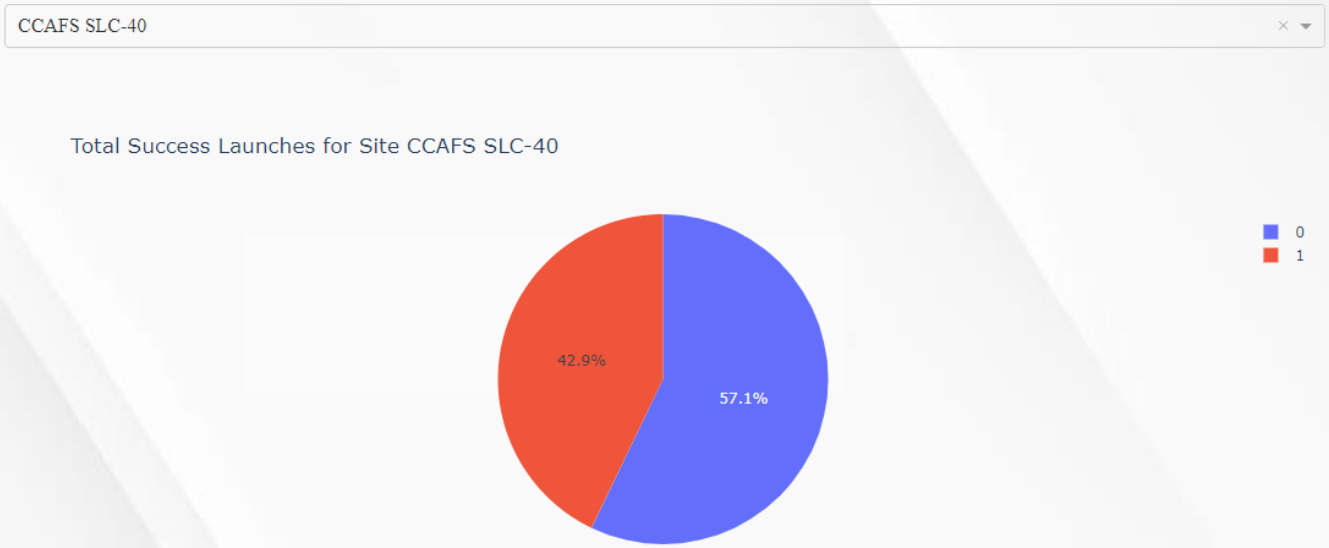
PLOTLY DASHBOARD

Successful Launch Count of All Sites



- Above: Rates of success per launch site indicate that KSC LC-39A has the highest total success rate at 41.2%

Highest Success Launch Site



- Above: CCAFS SLC-40 launch site has the highest launch success ratio of all sites at 42.9%

Payload vs. Launch Scatter Plot

- The interactive slider allows the user to adjust the payload range to compare figures
- From figure B and C it can be concluded that there's a significant difference in higher success rate of lighter loads (B), compared to that of heavier loads (C)
- Scatter plot can compare the success rate of all launches of up to a payload of 10,000 kg.


PREDICTIVE
ANALYTICS

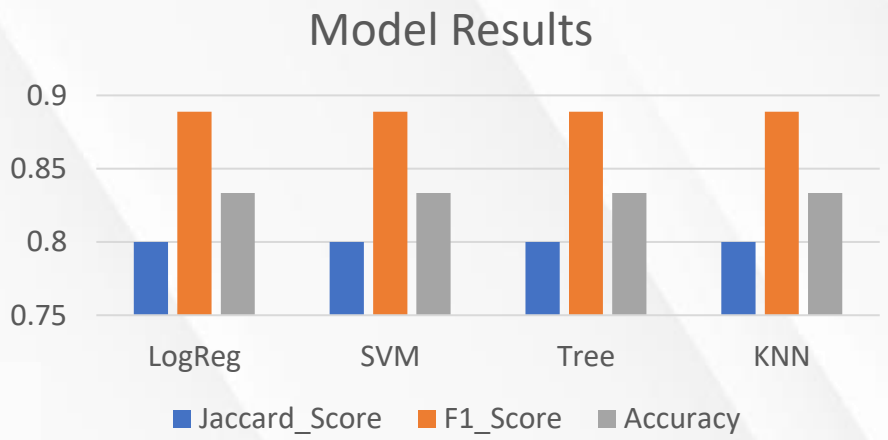
Classification Accuracy

- From chart and bar graph we can see that all models are relatively close in results.

- Using parameters and testing, we found the following model to be the best:
  - Decision Tree
  - Score: 0.87321

- The parameters that best suited the model are as follows:
  - Criterion: entropy
  - Max_depth: 10
  - Max_features: auto
  - Min_samples_leaf: 4
  - Min_samples_split: 10
  - Splitter: random

	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




```
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: ', bestalgorithm,'. Score:', 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_)
```

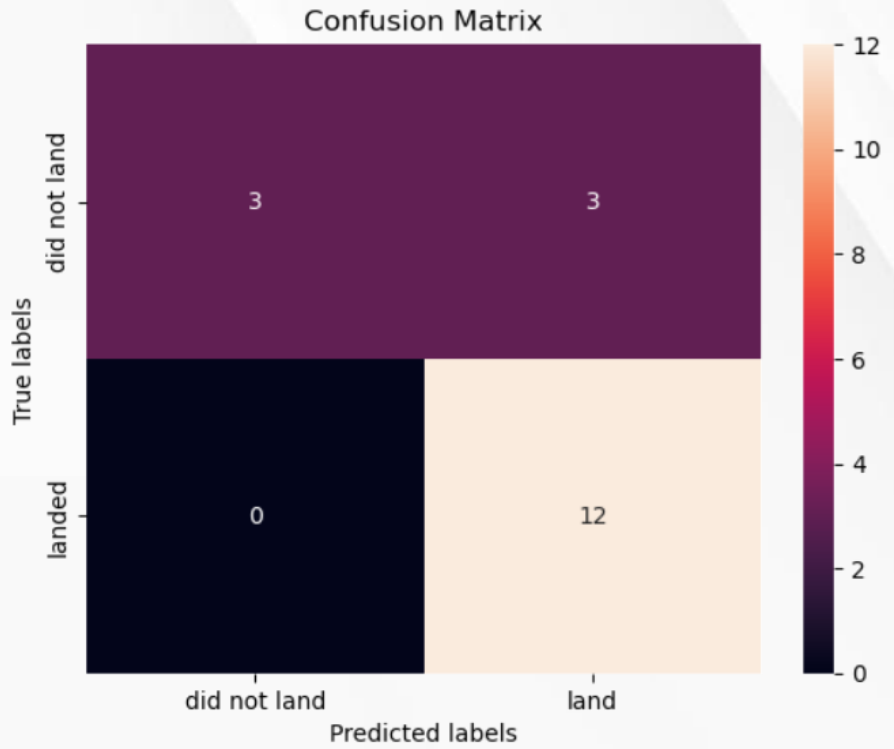
```
Best model:  DecisionTree . Score: 0.8732142857142857
Best params is : {'criterion': 'entropy', 'max_depth': 10, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}
```

Confusion Matrix

Performance Summary:

- All model confusion matrices were identical
- Confusion Matrix Outputs:
  - 12 True positive
  - 3 True negative
  - 3 False positive
  - 0 False Negative
- Having 3 false positives is the only Type 1 error that should be taken into consideration since those are unsuccessful landing marked as successful landing by the classifier

Precision TP / (TP + FP)	Recall TP / (TP + FN)	F1 Score 2 * (Precision * Recall) / (Precision + Recall)	Accuracy (TP + TN) / (TP + TN + FP + FN)
12 / 15	12 / 12	2 * (.8 * 1) / (.8 + 1)	(12+3)/(12+3+3 +0)
.80	1	.89	.833



Conclusion

Summary Outlook:

- Overall launch success rate is directly proportional to the time SpaceX has started as it has learned greatly from its failures and will one day perfect them
- Launch bases all follow the same location scheme:
  - Always on coastline and away from civilian infrastructure for safety
  - Most near equator due to the increase in boost from rotation of the earth that help in the savings of fuel
- Machine Learning model with the best performance is Decision Tree, which is recommended for the studying of these results in the future.
- Having 3 false positives is the only Type 1 error that should be taken into consideration

Data Number Results:

- Orbits ES-L1, GEO, HEO, SSO have a 100% success rate
- CCAFS SLC-40 has the highest launch success ratio of all sites at 42.9%
- KSC LC-39A has the highest total success rate at 41.2% and 100% success rate for payloads less than 5,500 kg
- Decision Tree offers the best success rate of 87.3%, and can be improved with more data

Recommendations

- Use launch site location data to further improve locations for future bases
- Increase the dataset as time progresses to increase model and output data to further refine results
- Apply these models and practices to Falcon Heavy and Starbase as launches continue
- Research or develop new tools and models, such as XGBoost, to not only reassure calculations, but also to find new data and answers, as well as keeping up with the latest in computing


SPACEX
THE END
BOEING
STARLINER
SPACEX