Winning Space Race with Data Science

Arwa Abdulrahman Break

August 03, 2023

IBM Developer SKILLS NETWORK

Outline

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

Executive Summary

Methodologies

In this capstone, we attempt to assist a new rocket company 'Space Y' that would like to compete with SpaceX. We will use the learned Data Science methodology to determine the price of each launch by:

- Data Collection: gathering data about Space X using REST API and Web Scraping
- Data Wrangling: cleaning the gathered data for success/failure outcomes
- EDA: exploring data with SQL queries and data visualization techniques on payload, launch site, successful launches, … etc
- Interactive Visual Analytics: building dashboards with Folium and Plotly Dash to visualize launch sites and related details on a map
- Model Development: performing predictive analysis to determine landing outcomes

Results

As a result of the aforementioned methodology implementation, the following has been determined:

- EDA: the first successful landing outcome in ground pad was achieved on 22-12-2015
- Interactive Visual Analytics: most launch sites are close to coastlines
- Predictive Analytics: best model is the Decision Tree with a score of 87%

Introduction

Background

The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets.

Perhaps the most successful is SpaceX. SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Objective

We need to explore the relationships between the first stage landing success rate and other features such as launch site, no of flights, payload, … etc. In addition, we will analyze landing success rate trends over the years.

Methodology

Methodology

- Data collection methodology:
  - gathered data about Space X using REST API and Web Scraping
- Perform data wrangling
  - cleaned the gathered data for success/failure outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
  - explored data with SQL queries and data visualization techniques on payload, launch site, successful launches, … etc
- Perform interactive visual analytics using Folium and Plotly Dash
  - built dashboards with Folium and Plotly Dash to visualize launch sites and related details on a map
- Perform predictive analysis using classification models
  - Developed model to perform predictive analysis to determine landing outcomes, … etc

Data Collection

SpaceX API

request rocket launch data from SpaceX API → decode the response content as a Json → turn it into a Pandas dataframe → deal with missing values

Web Scraping

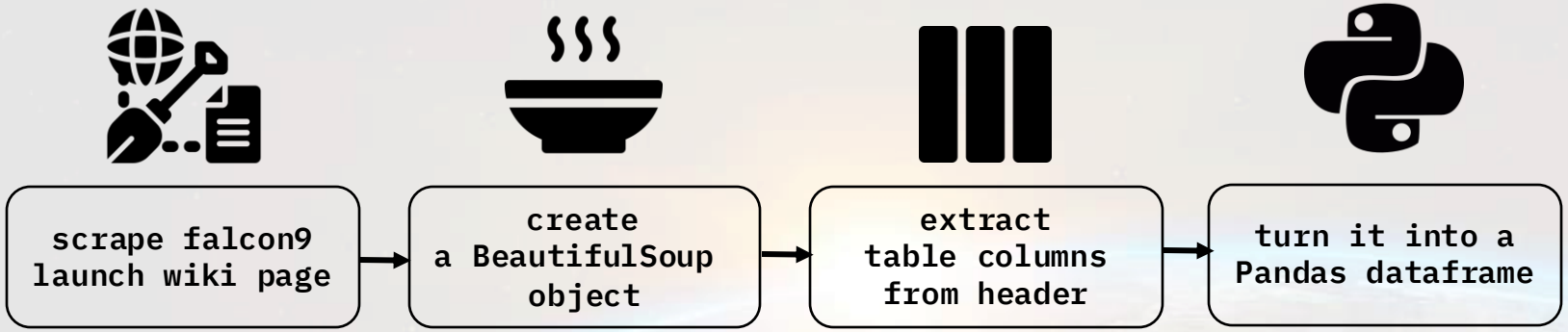
scrape falcon9 launch wiki page → create a BeautifulSoup object → extract table columns from header → turn it into a Pandas dataframe

Data Collection – SpaceX API

request rocket launch data from SpaceX API → decode the response content as a Json → turn it into a Pandas dataframe → deal with missing values

[1-spacex-data-collection-api.ipynb]

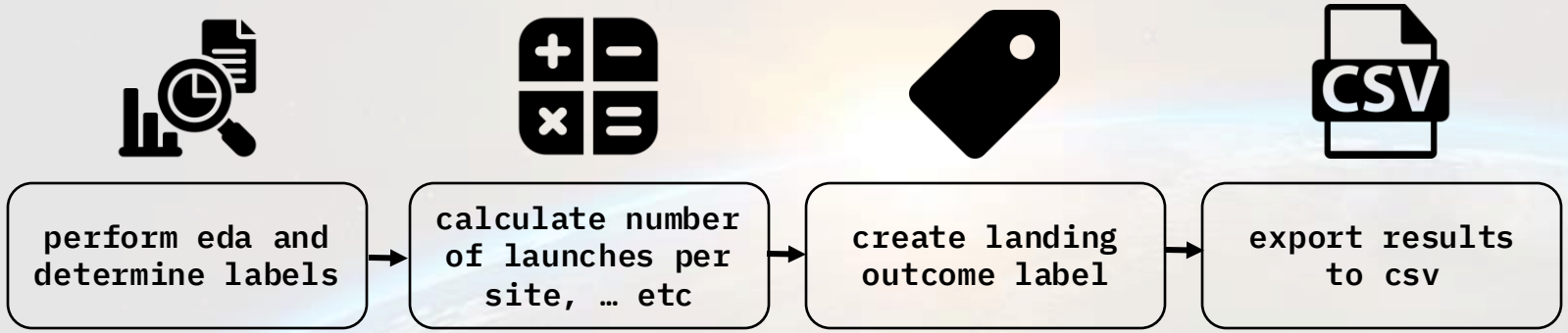
Data Collection – Web Scraping



[2-spacex-data-collection-webscraping.ipynb]

Data Wrangling

Mainly calculated: the no. of launches on each site, no. and occurrences of each orbit, no. and occurrences of mission outcome of each orbit type, and create a landing outcome label from Outcome column where 1 means the booster successfully landed and 0 means it was unsuccessful



[3-spacex-data_wrangling.ipynb]

EDA with SQL

- Queries Used:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank 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.

4-spacex-eda-sql.ipynb

EDA with Data Visualization

- Charts Used:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Success Rate and Orbit Type
  - Flight Number and Orbit Type
  - Payload and Orbit Type
  - Launch Success Yearly Trend

- Types of Charts Used:
  - Scatter Plots: to examine relationships between variables
  - Bar Chart: to compare between discrete categories and measure their values
  - Line Chart: to examine trends over time

5-spacex-eda-dataviz.ipynb

Build Interactive Map with Folium

Object	Description	Reason
Marker with Circle	Blue: NASA Johnson Space Center coordinates Red: All launch sites coordinates	Mark all launch sites on a map
Colored Markers	Green: Successful launch outcome Red: Unsuccessful launch outcome	Mark the success/failed launches for each site
Colored Lines	CCAFS SLC-40 launch site and its proximity to any railway, highway, coastline, etc.	Calculate the distances between a launch site to its proximities

6-spacex-interactive-va-folium.ipynb

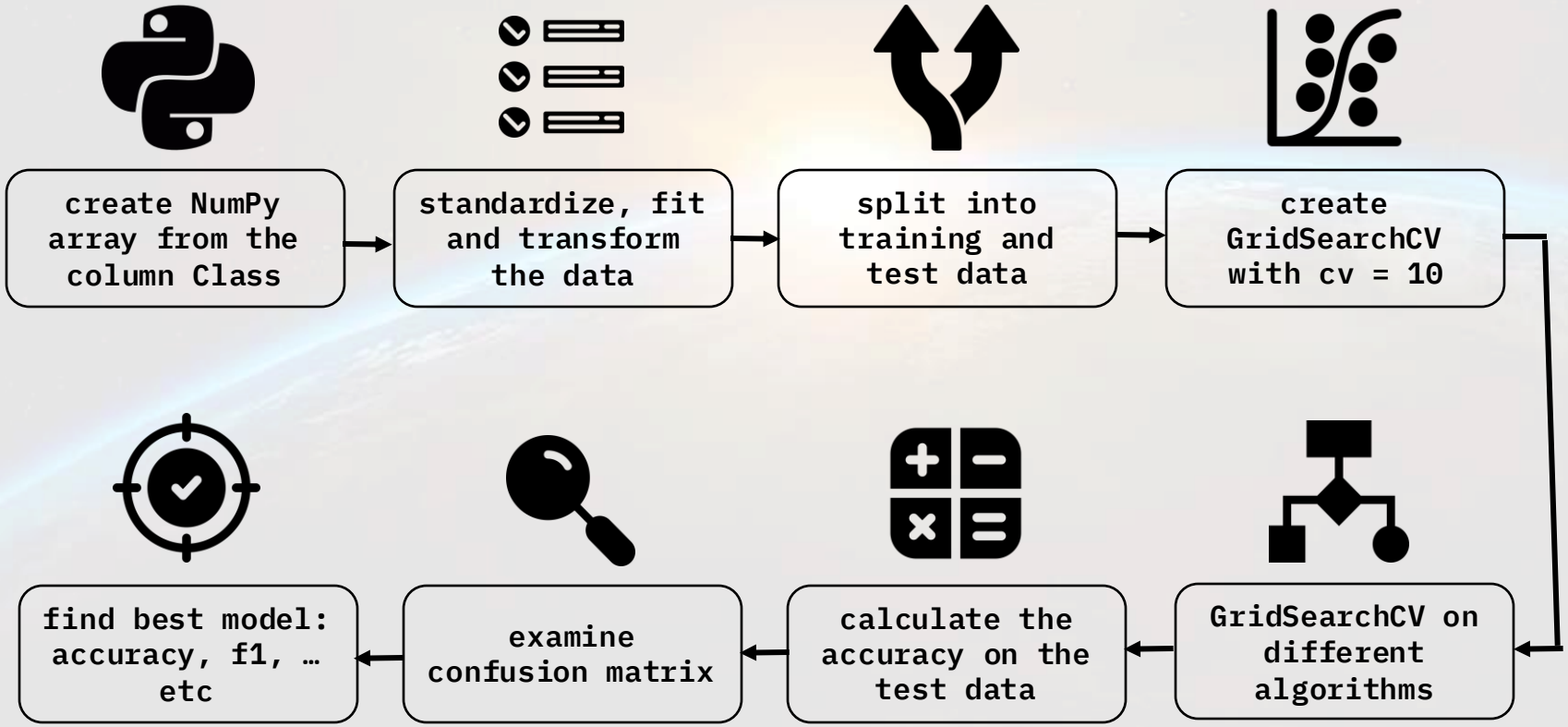
Build Dashboard with Plotly Dash

- Dashboard Graphs/Interactions Used:
  - Launch Site Selection (Dropdown): to filter results for a specific launch site or all sites
  - Success vs. Failed Launches Pie Chart: to show the total successful/failed launches count for a specific launch site or all sites (as per the dropdown selection)
  - Payload Range Slider: to select payload range
  - Payload and Launch Success Scatter Plot: to show the correlation between payload and launch success

7-spacex_dash_app.py

Predictive Analysis (Classification)

Started by standardizing, fitting and transforming the data before splitting it into train/test, then applied logistic regression, svm, decision tree and KNN and calculated the accuracy on test data, examines the confusion matrix and identified the best model.



[8-spacex-ml-prediction.ipynb]

Results

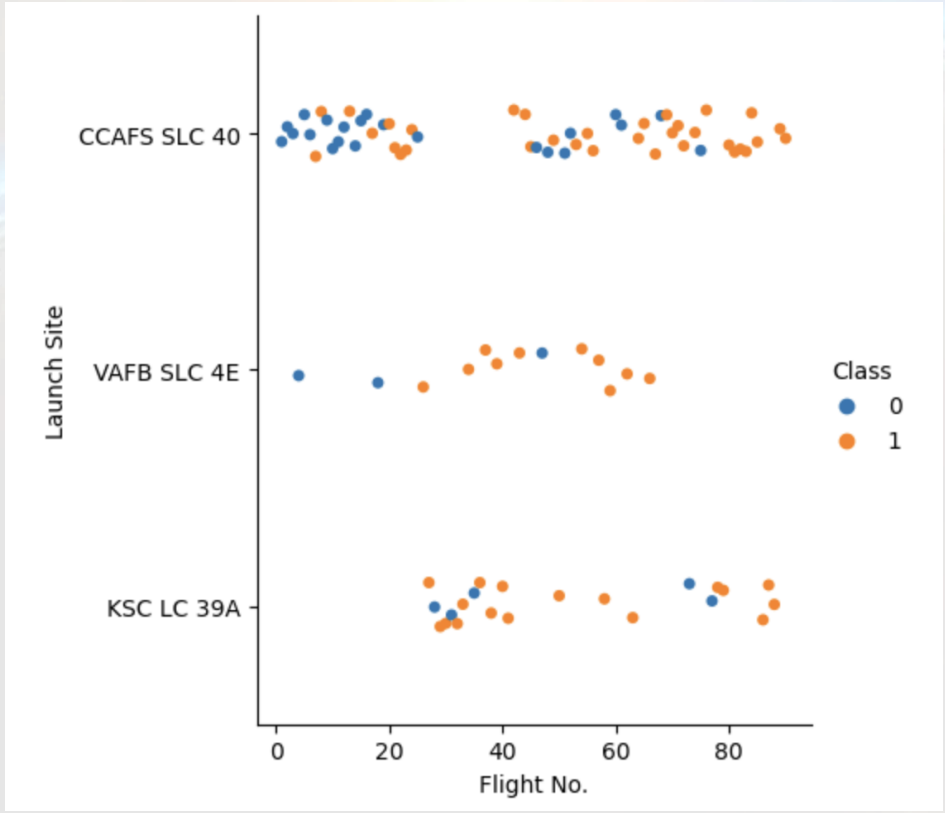
Results Summary

- Exploratory Data Analysis
  - First successful landing outcome in ground pad was achieved in 22-12-2015
  - ES-L1, GEO, HEO and SSO Orbits have 100% success rate while SO orbit has the lowest
  - Since 2013m the success rate kept increasing until 2020
- Interactive Analytics
  - Almost all launch sites in proximity to the Equator line
  - Almost all launch sites in very close proximity to the coast
  - Launch site CCAFS SLC-40 has 3/7 success rate
- Predictive Analysis
  - Examining the confusion matrix, we see that the major problem is false positives
  - Best model is Decision Tree with a score of 87%

Insights Drawn From EDA

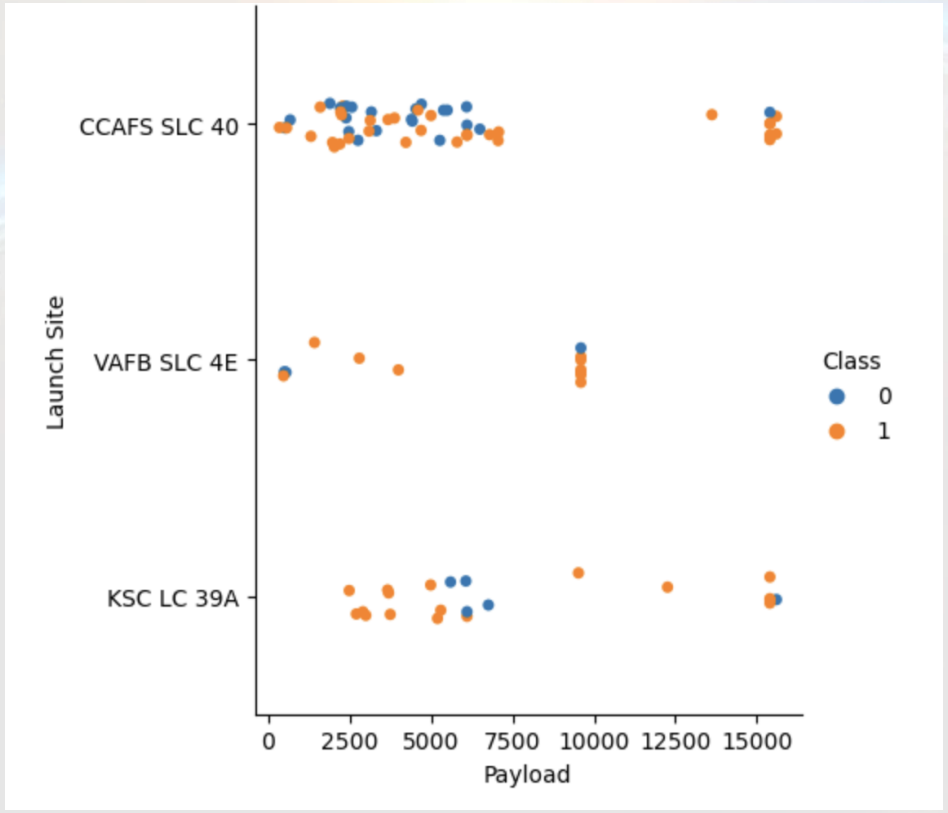
Flight Number vs. Launch Site

- Exploratory Data Analysis
  - Most earlier flights had lower success rate (Class 0 = Failure)
  - Most later flights had higher success rate (Class 1 = Success)
  - KSC LC-39A and VAFB SLC 4E have higher success rate
  - Newer launches may have higher success rate



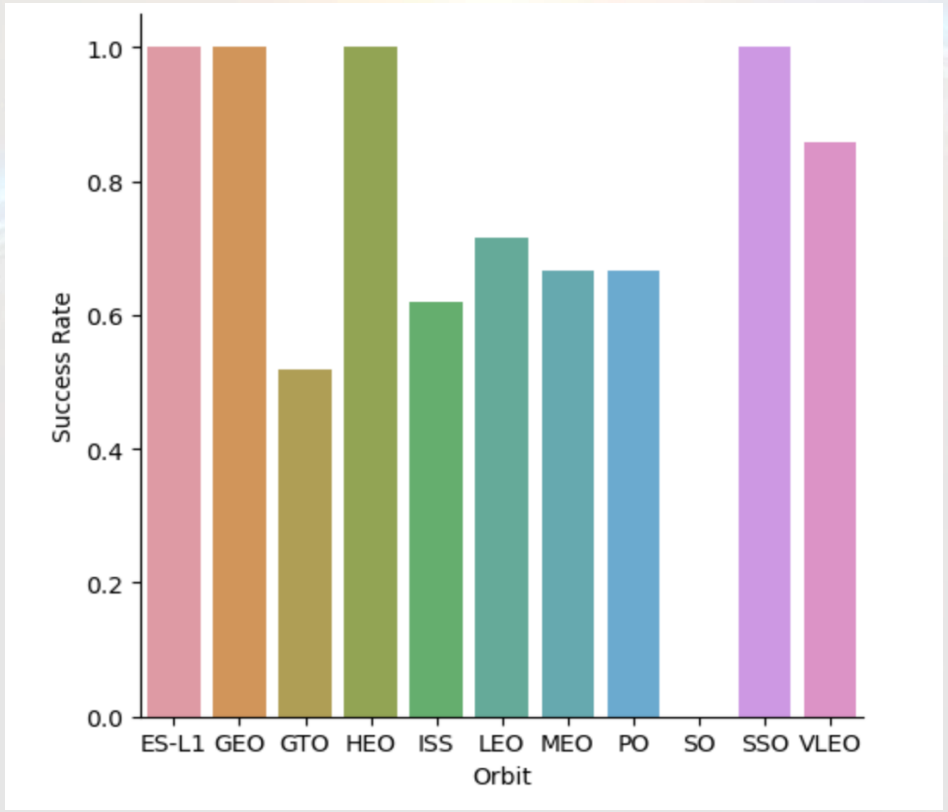
Payload vs. Launch Site

- Exploratory Data Analysis
  - VAFB SLC 4E has only launched payload of 10000 kg or less
  - Most launches with payload higher than 7500 kg had success rate (Class 1 = Success)
  - KSC LC-39A has 100% success rate for payload 5000 kg or less
  - The higher the payload, the higher the success rate



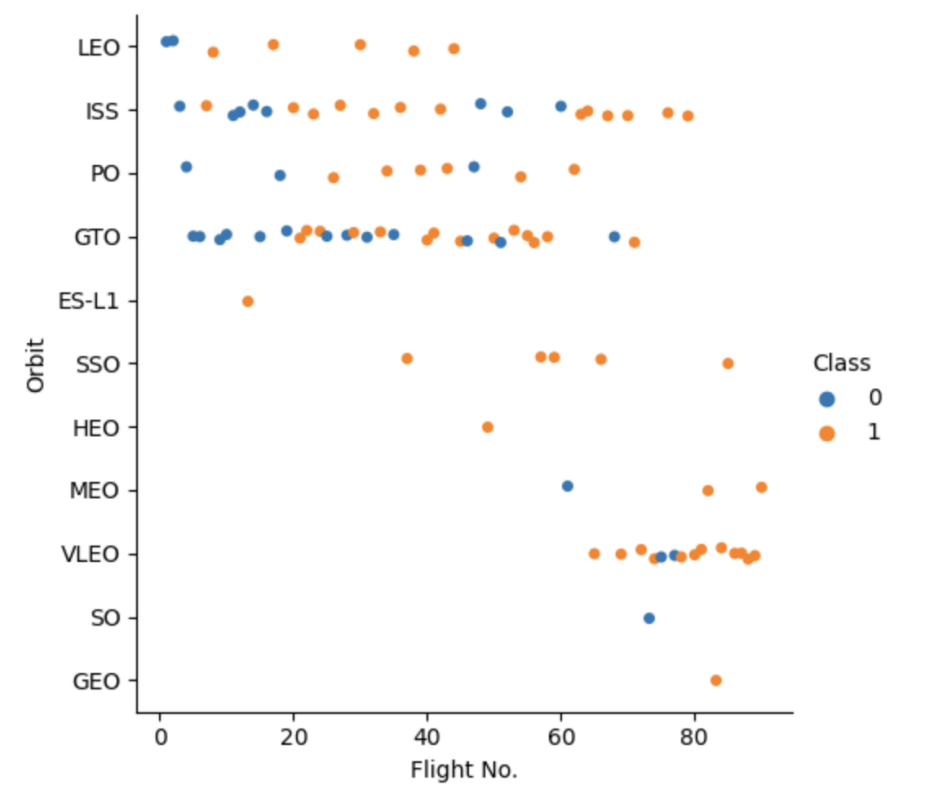
Success Rate vs. Orbit Type

- Exploratory Data Analysis
  - Highest success rate orbits types are ES-L1, GEO, HEO and SSO (100%)
  - Medium success rate orbits types are GTO, ISS, LEO, MEO, PO and VLEO (50%-85%)
  - Lowest success rate orbits type is SO (0%)



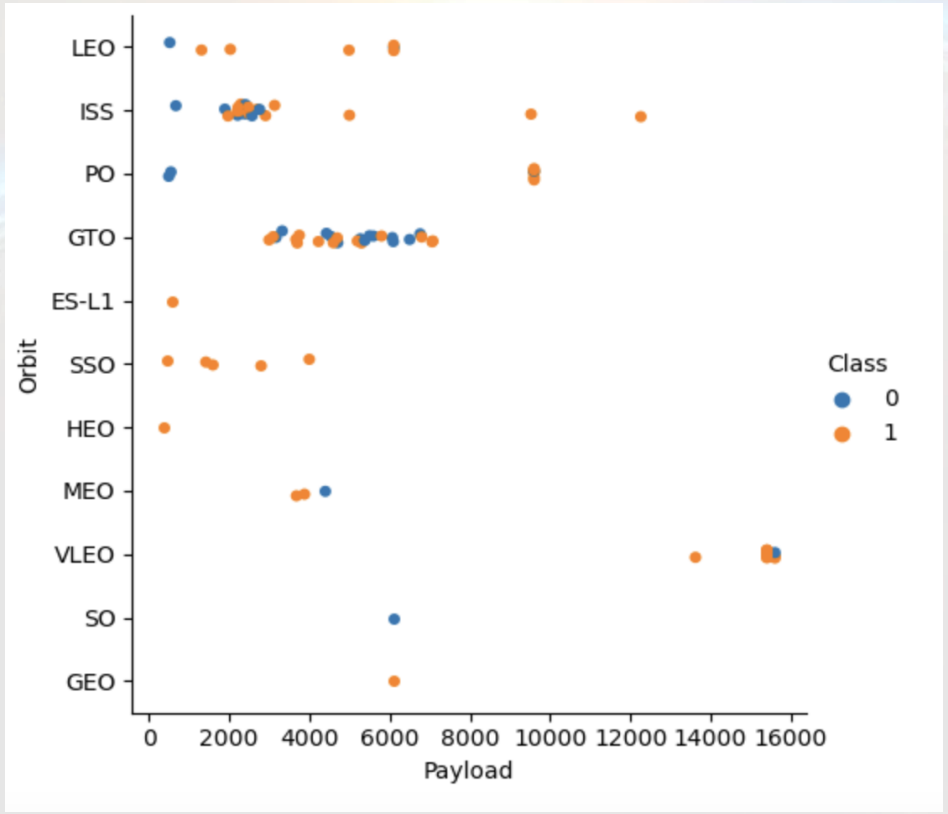
Flight Number vs. Orbit Type

- Exploratory Data Analysis
  - In the LEO orbit the Success appears related to the number of flights
  - There seems to be no relationship between flight number when in GTO orbit



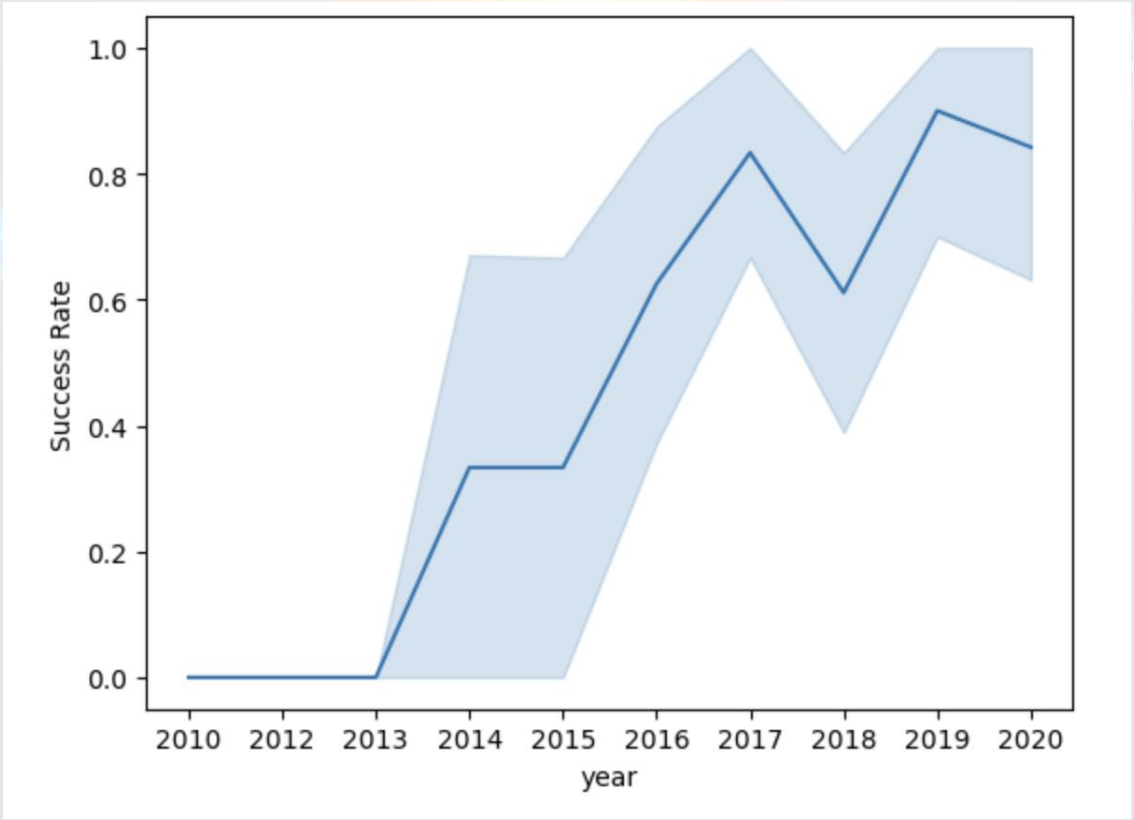
Payload vs. Orbit Type

- Exploratory Data Analysis
  - With heavy payloads the successful landing rate are more for Polar, LEO and ISS
  - For GTO however, there is an obvious mix of success and failure



Launch Success Yearly Trend

- Exploratory Data Analysis
  - The success rate started to increase from 2013 to 2020
  - It dropped in 2017 but started to go back up the next year (2018)
  - The success rate was at its lowest for launches between 2010 and 2013
  - Overall, there is an obvious positive success rate trend over time



All Launch Site Names

- Exploratory Data Analysis
  - Display the names of the unique launch sites in the space mission

```
]: %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

Launch Site Names Begin with 'CCA'

- Exploratory Data Analysis
  - Display 5 records where launch sites begin with the string '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
2010-04-06	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

- Exploratory Data Analysis
  - Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql select sum(PAYLOAD_MASS__KG_) as "Total Payload Mass" from SPACEXTABLE
        where CUSTOMER = "NASA (CRS)";
```

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

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

- Exploratory Data Analysis
  - Display average payload mass carried by booster version F9 v1.1

```
%%sql select avg(PAYLOAD_MASS__KG_) as "Average Payload Mass" from SPACEXTABLE
        where BOOSTER_VERSION = "F9 v1.1";
```

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

Average Payload Mass
2928.4

First Successful Ground Landing Date

- Exploratory Data Analysis
  - List the date when the first successful landing outcome in ground pad was achieved

```
%%sql select min(DATE) as "First Succesful Landing" from SPACEXTABLE
        where LANDING_OUTCOME = "Success (ground pad)";
```

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

First Succesful Landing
2015-12-22

Successful Drone Ship Landing

- Exploratory Data Analysis
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql select BOOSTER_VERSION from SPACEXTABLE
        where LANDING_OUTCOME = "Success (drone ship)"
        and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

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

Total Successful/Failure Mission

- **Exploratory Data Analysis**
  - **List the total number of successful and failure mission outcomes**

```
%%sql select MISSION_OUTCOME, count(*) as "Total" from SPACEXTABLE
        group by MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- Exploratory Data Analysis
  - List the names of the booster_versions which have carried the maximum payload mass

```
%%sql select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEXTABLE
        where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

Booster_Version	PAYLOAD_MASS__KG_
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

- Exploratory Data Analysis
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

```
%%sql select substr(Date,6,2) as Month, DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME
        from SPACEXTABLE
        where LANDING_OUTCOME like "Failure%"
        and substr(Date,1,4) = "2015";
```

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

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes

- **Exploratory Data Analysis**
  - **Rank 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**

```
%%sql select LANDING_OUTCOME, DATE, COUNT(*) AS "Count" from SPACEXTABLE
        where DATE between "2010-06-04" and "2017-03-20"
        group by LANDING_OUTCOME
        order by count(*) desc
```

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

Landing_Outcome	Date	Count
No attempt	2012-05-22	10
Success (ground pad)	2015-12-22	5
Success (drone ship)	2016-08-04	5
Failure (drone ship)	2015-10-01	5
Controlled (ocean)	2014-04-18	3
Uncontrolled (ocean)	2013-09-29	2
Precluded (drone ship)	2015-06-28	1
Failure (parachute)	2010-08-12	1

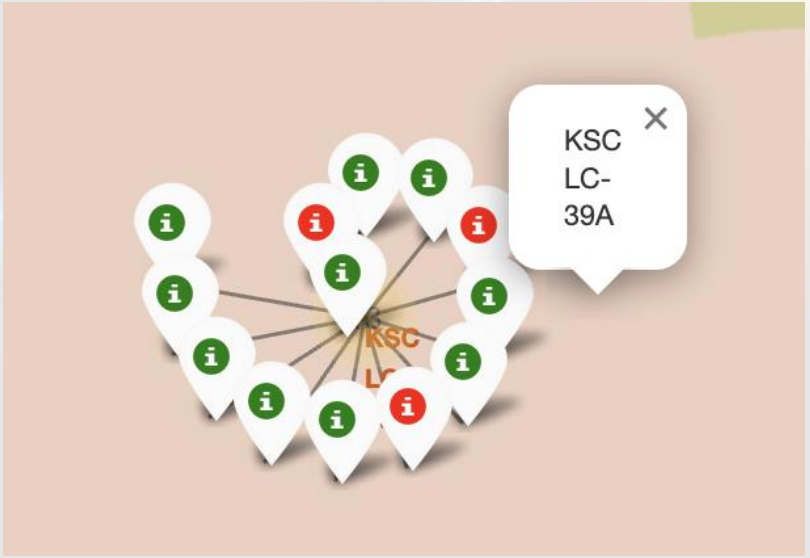
Launch Sights Proximities Analysis

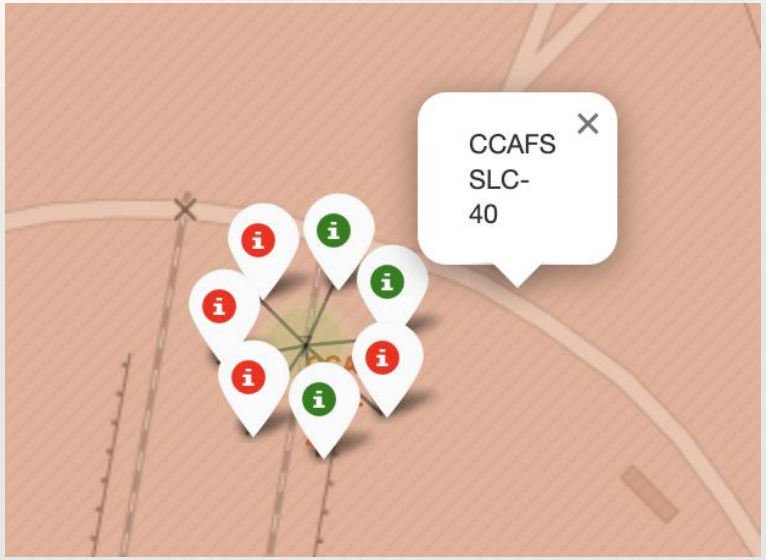
All Launch Sites Location Markers

- Visual Analytics - Folium
  - All launch sites locations are marked with circles on the map
  - All in proximity to the Equator line
  - All in very close proximity to the coast

Color-Labeled Launch Outcomes

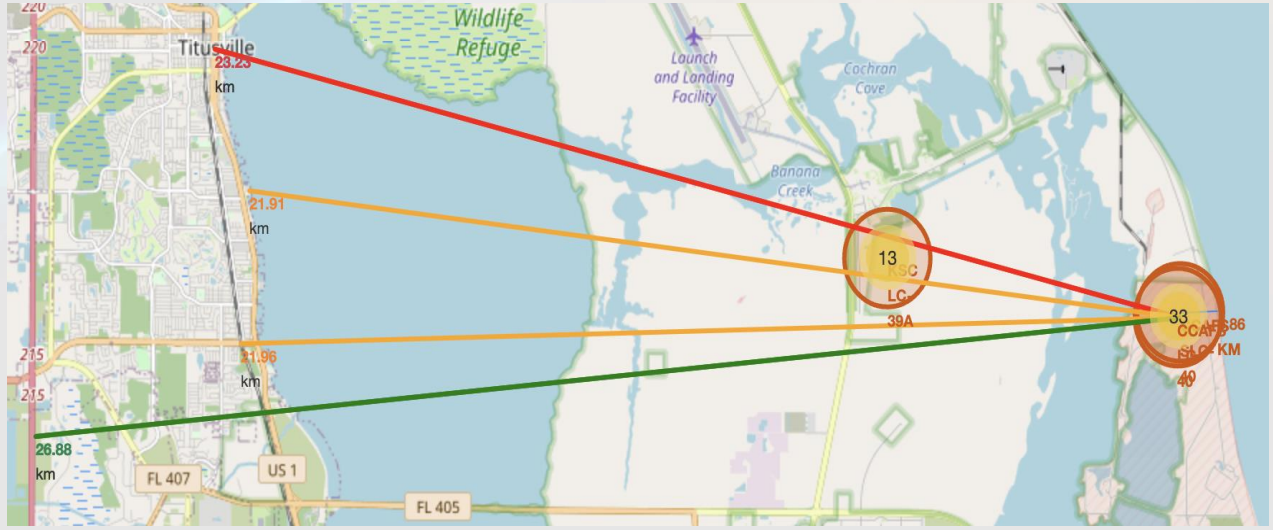
- Visual Analytics - Folium
  - Each launch outcome is color-labeled with Green for successful outcomes and Red for failed launch outcomes
  - The success rate can visually be calculated as (no. of success outcomes)/(total outcomes) for each site
  - CCAFS SLC-40 has a success rate of 3/7 (42.9%)
  - KSC LC-39A has a success rate of 10/13 (76.9%)

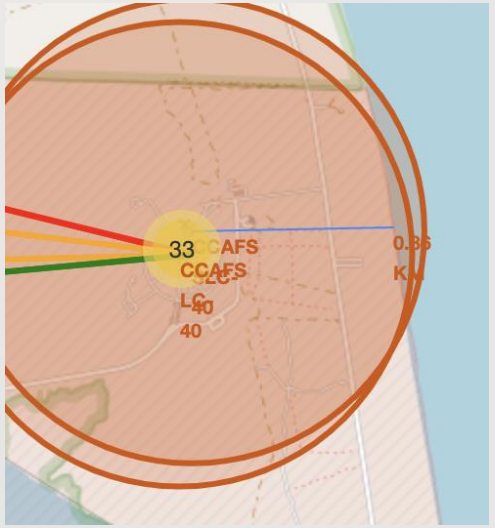


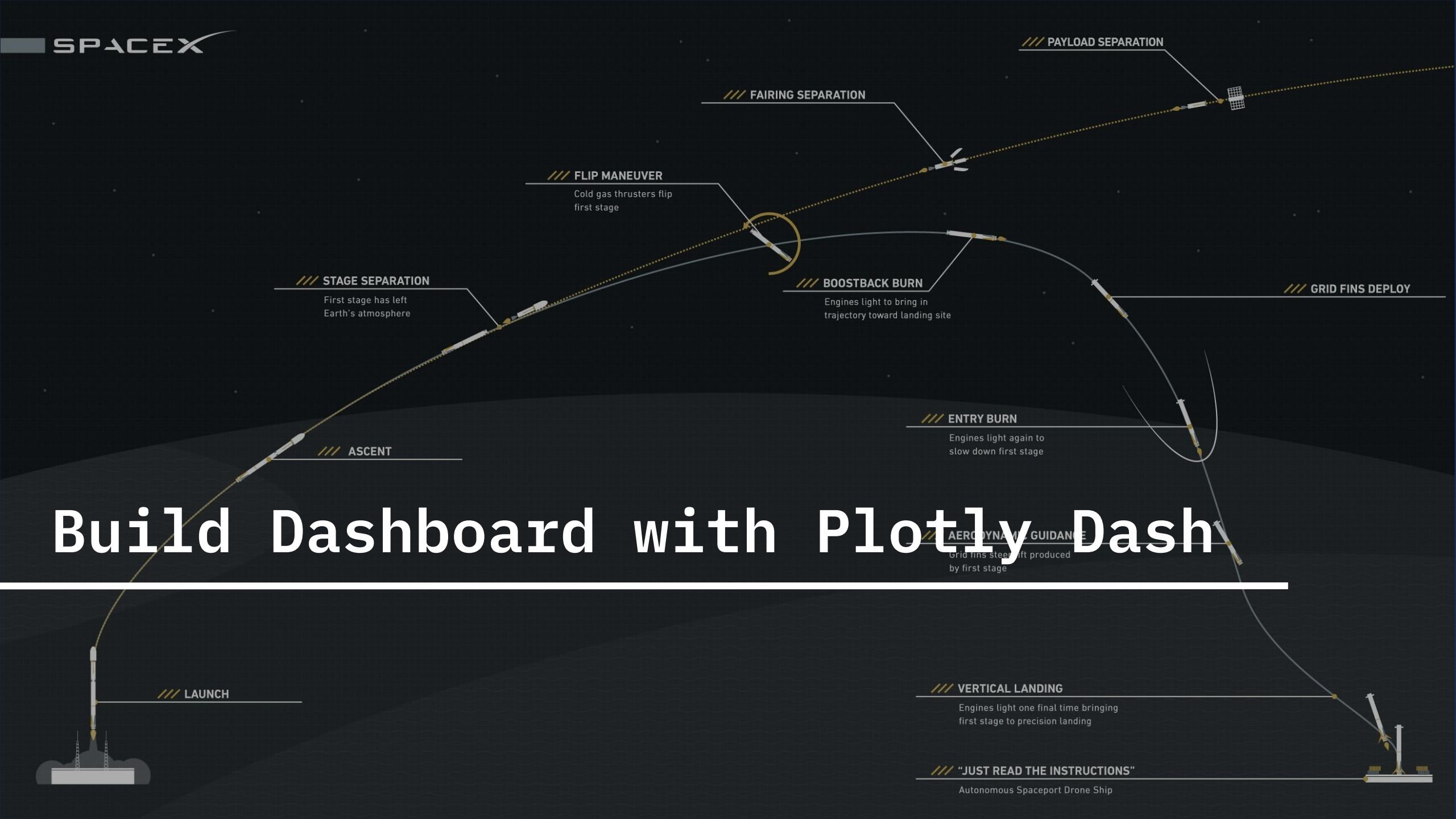


Launch Site Proximities

- Visual Analytics - Folium
  - CCAFS SLC-40 launch site proximities to closest city, railway, highway can be observed
  - 0.86 km from nearest coastline
  - 21.91 km from nearest railway
  - 23.23 km from nearest city
  - 26.88 km from nearest highway



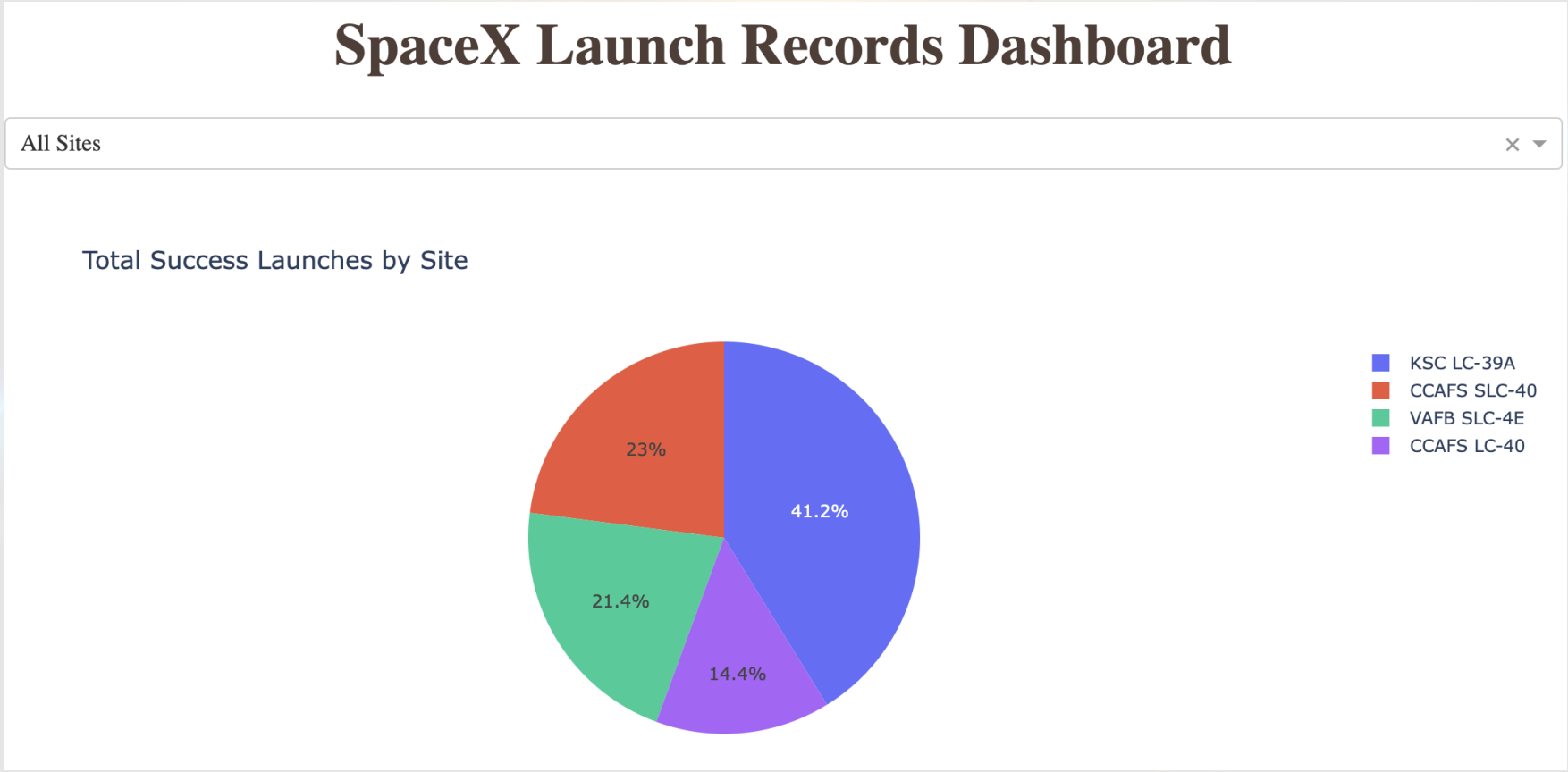


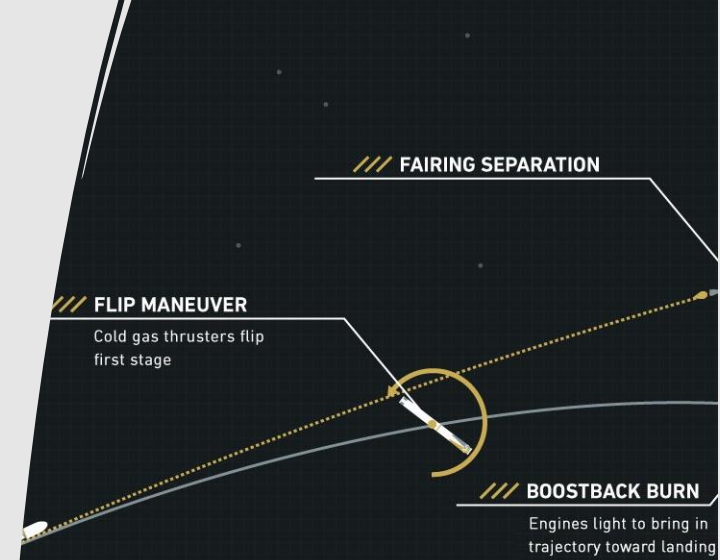


Build Dashboard with Plotly Dash

Launch Success Count For All Sites

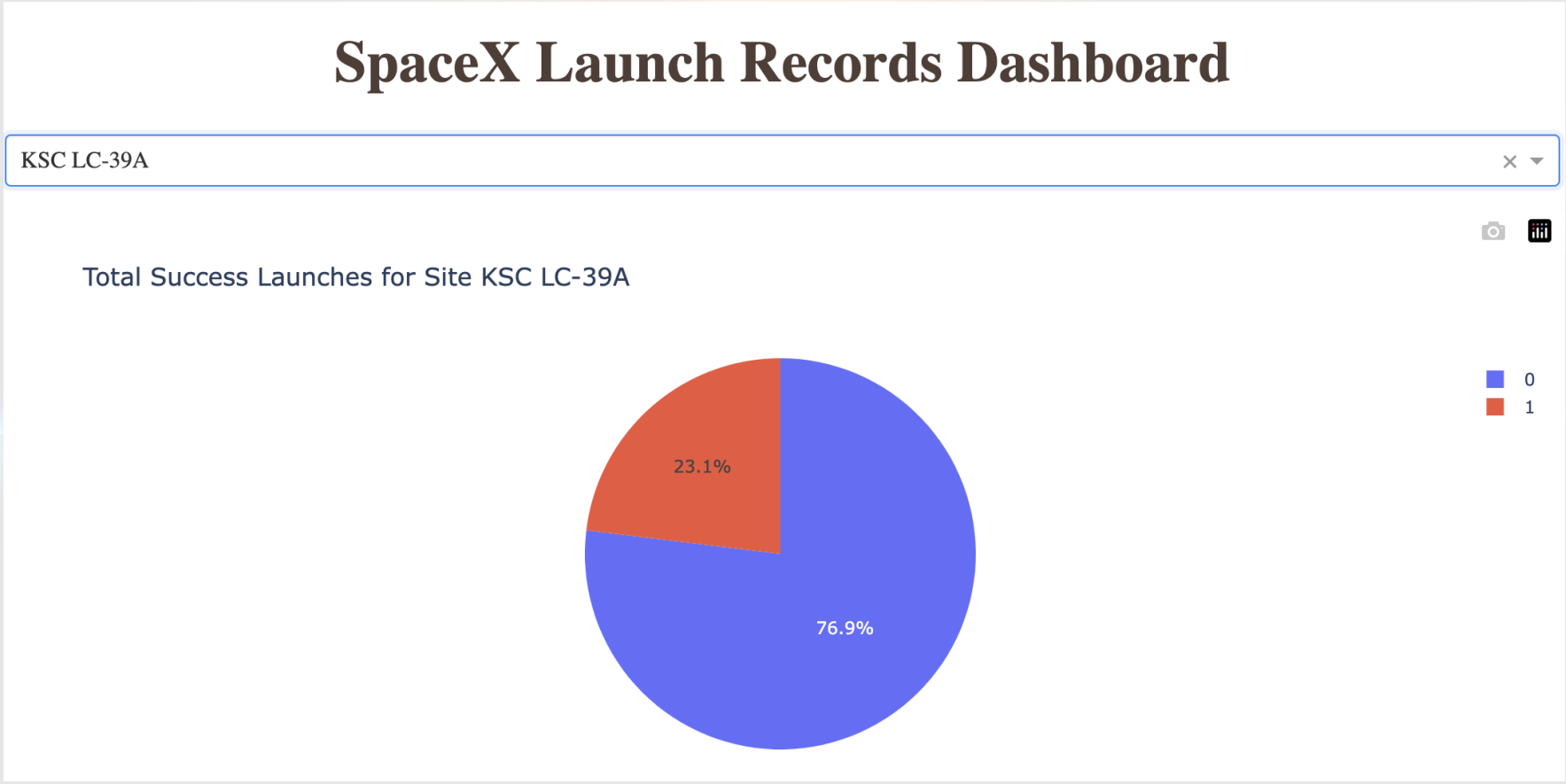
- Visual Analytics – Plotly Dash
  - KSC LC-39A has the highest the successful launches (41.2%)
  - CCAFS SLC-40 has the least successful launches (14.4%)

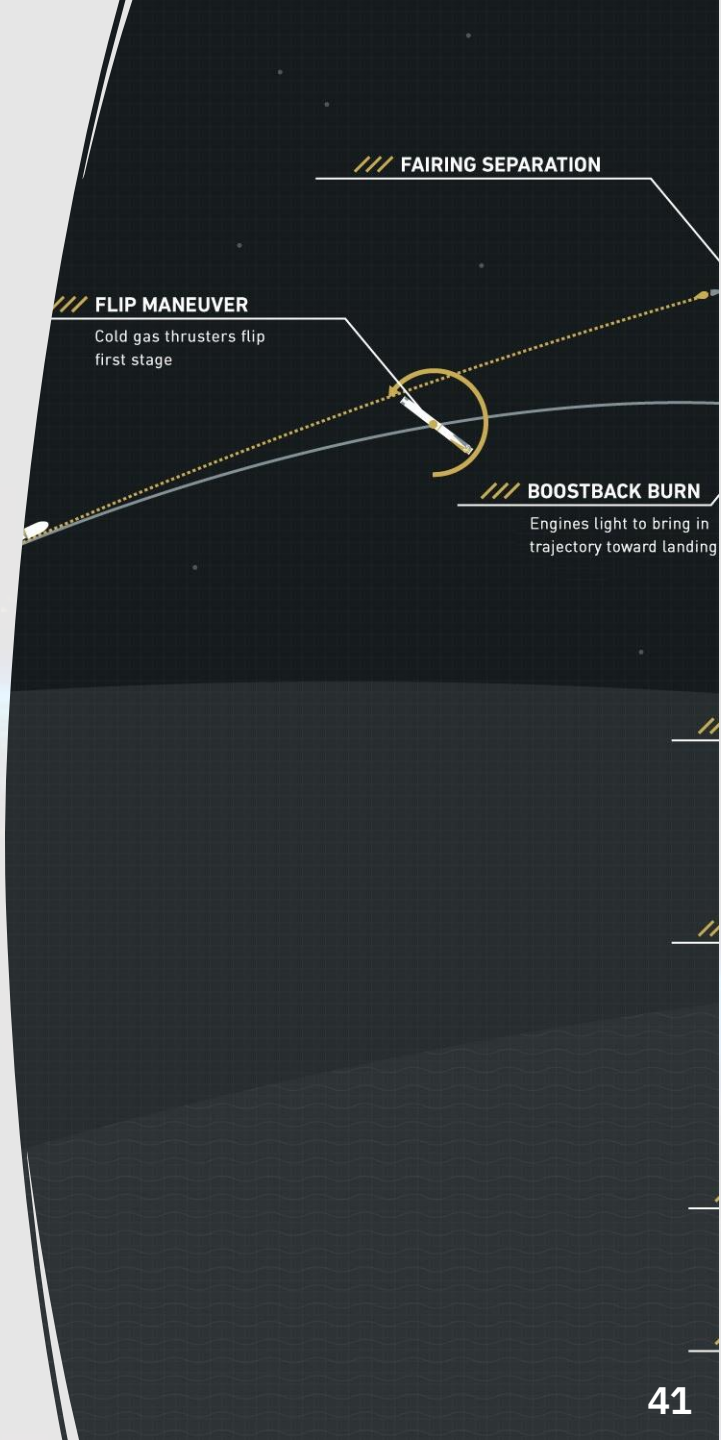




Launch Site Highest Success Ratio

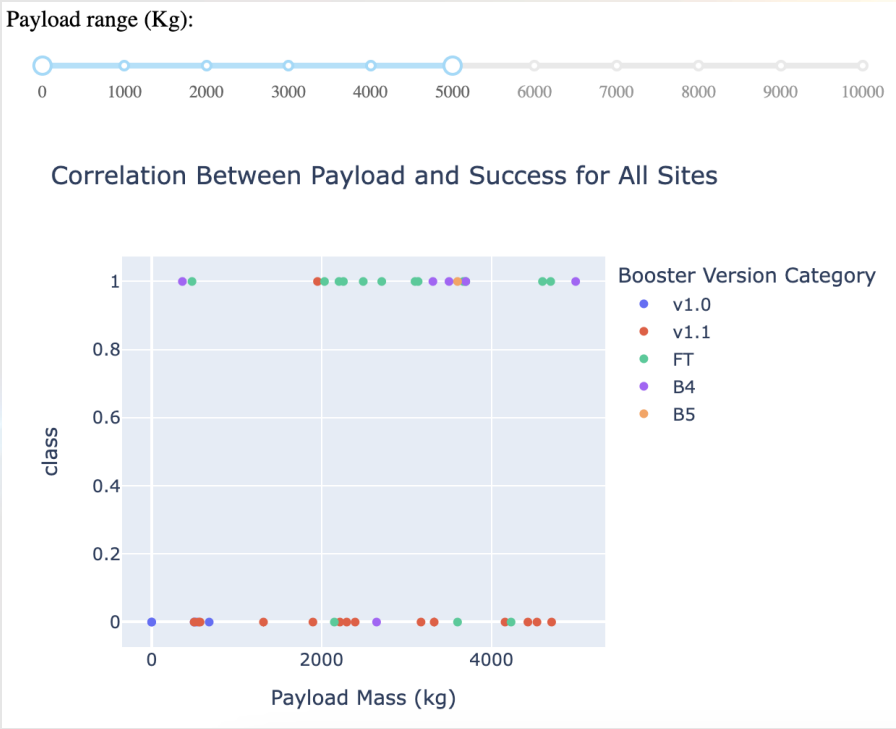
- Visual Analytics – Plotly Dash
  - KSC LC-39A has the highest the successful launches (76.9%)
  - 10 successful launches and 3 failures

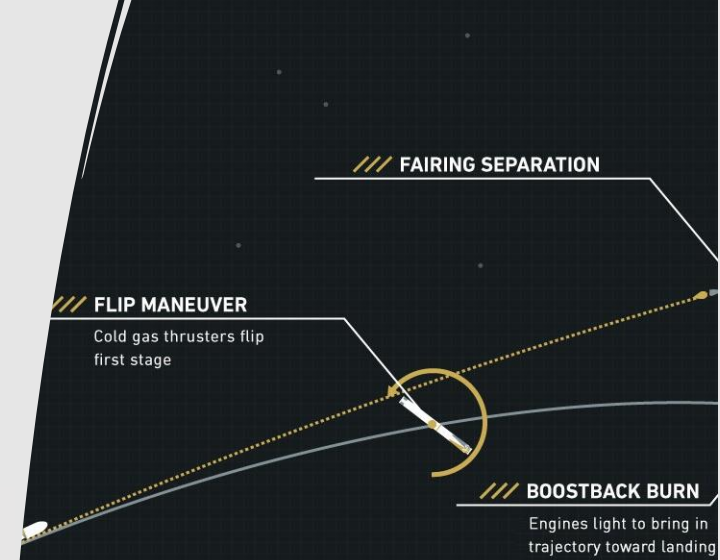




Payload vs. Launch Outcome

- Visual Analytics – Plotly Dash
  - Payloads between 2000 and 5000 have the highest success rate
  - Note that Class 0 = Failure and Class 1 = Successful

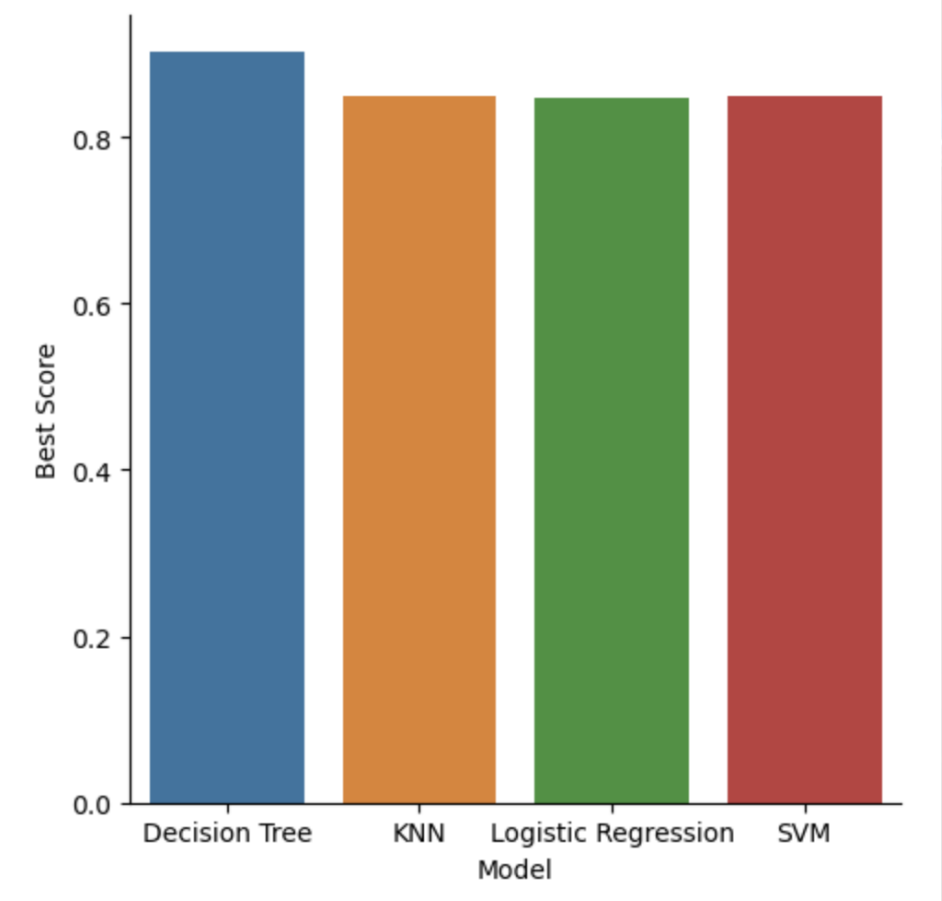




Predictive Analysis (Classification)

Classification Accuracy (1)

- Predictive Analysis
  - Best model is the Decision Tree with a score of 0.90



Classification Accuracy (2)

- Predictive Analysis
  - As can be seen below however, all models performed similarly, probably due to the small dataset.
  - The decision tree outperformed the rest when looking at .best_score_

	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

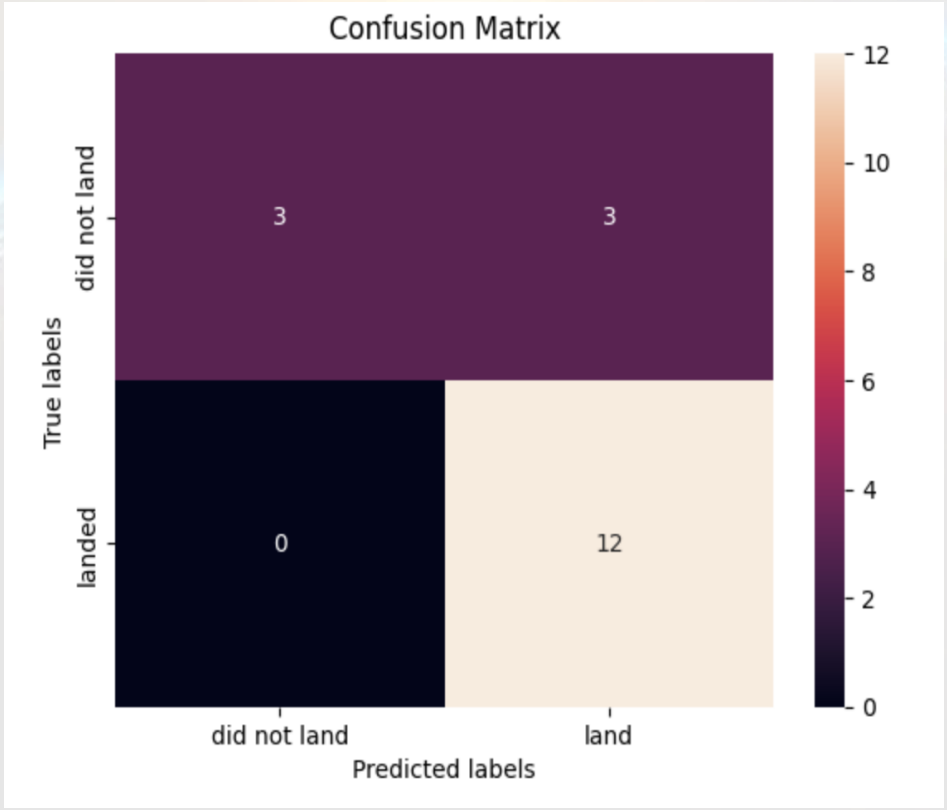
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model 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': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'random'}
```

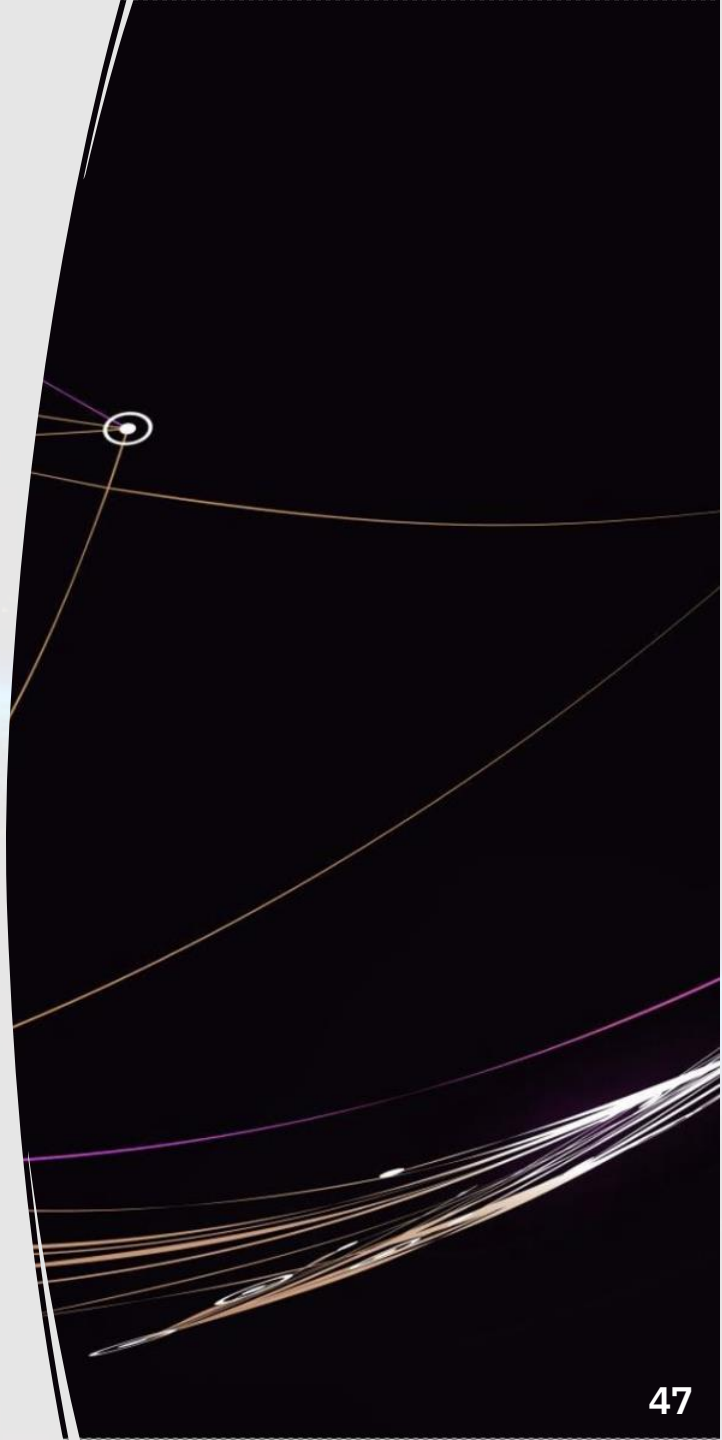
Confusion Matrix

- Predictive Analysis
  - All classifications had identical confusion matrix
  - The major problem is false positives
  - Matrix Outcome: 12 TP, 3 TN, 3 FP and 0 FN



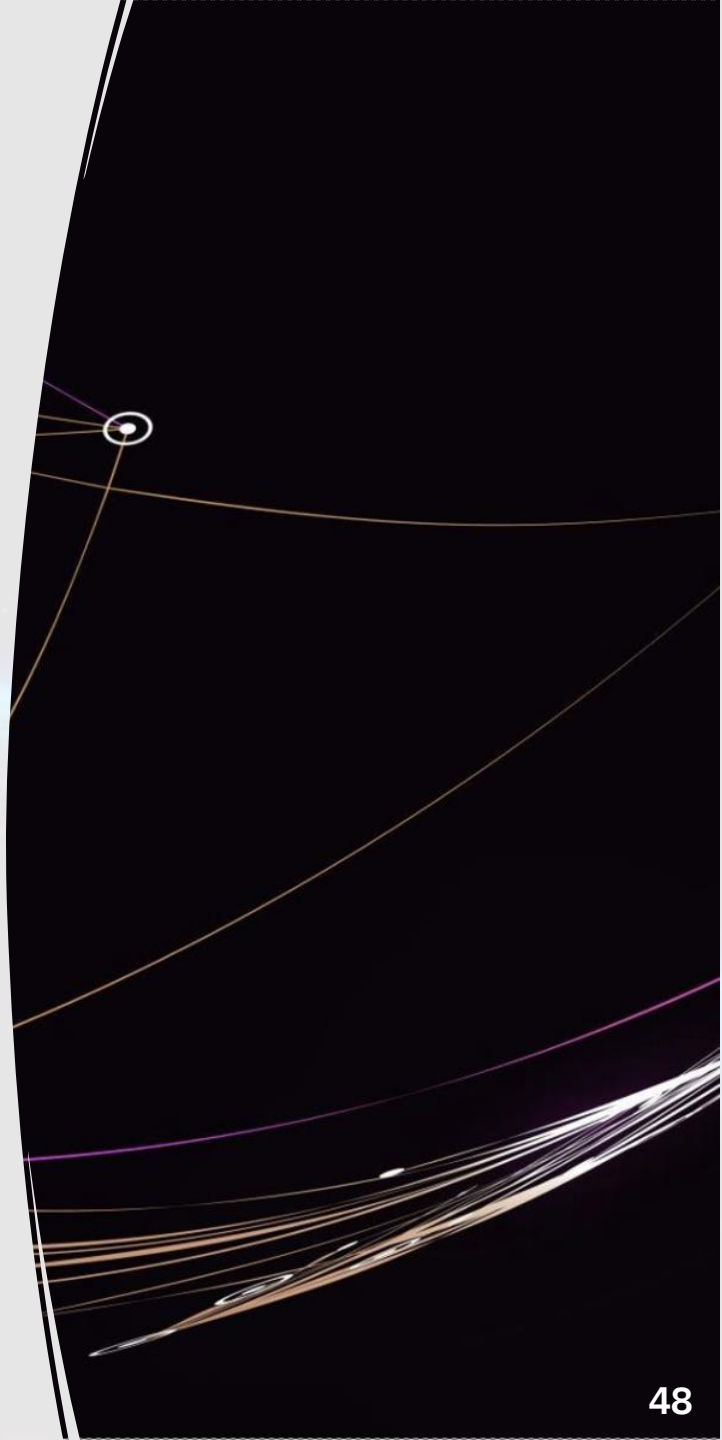
Conclusions

- Best classification model for this data set is the Decision Tree
- KSC LC-39A has the highest the successful launches (41.2%)
- Payloads between 2000 and 5000 have the highest success launches
- All launch sites are within close proximity to the Equator line and the coast
- There is a positive launch success rate trend over time
- Highest success rate orbits types are ES-L1, GEO, HEO and SSO (100%)



Appendix

- IBM-DS-Capstone
  - 1-spacex-data-collection-api.ipynb
  - 2-spacex-data-collection-webscraping.ipynb
  - 3-spacex-data_wrangling.ipynb
  - 4-spacex-eda-sql.ipynb
  - 5-spacex-eda-dataviz.ipynb
  - 6-spacex-interactive-va-folium.ipynb
  - 7-spacex_dash_app.py
  - 8-spacex-ml-prediction.ipynb




Thank You..