IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wiley Cubic
11-12-2025

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis and Data Visualization
- Build an interactive Folium map
- Build a Plotly Dash Dashboard
- Predictive Analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive Analysis results
- Predictive Analysis results

Introduction

Project background and context

SpaceX is an aerospace company with the goal of reducing space transportation costs and colonizing Mars. The company advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, and 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. Based on publicly available data, we are going to predict if SpaceX will reuse the first stage of its rockets.

Problems you want to find answers

- How to variables such as number of flights, launch site, payload mass, and orbit affect the success rate of the Falcon 9 first stage landing.
- What is the best algorithm to use when predicting the outcome

Section 1

Methodology

Methodology

Executive Summary:

Data collection methodology

- Using SpaceX REST APU
- Using Web Scraping

Perform data wrangling

- Cleaned missing values
- Created numeric classification for launch outcomes
- Use one hot encoding to prepare data for binary classification

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Build, tune, and evaluate classification models to ensure accurate results

Data Collection

Describe how data sets were collected

The data collection process involved a combination of API requests from SpaceX's REST API and Web Scraping from SpaceX's Wikipedia page.

Data from both of these sources are used for a more detailed analysis.

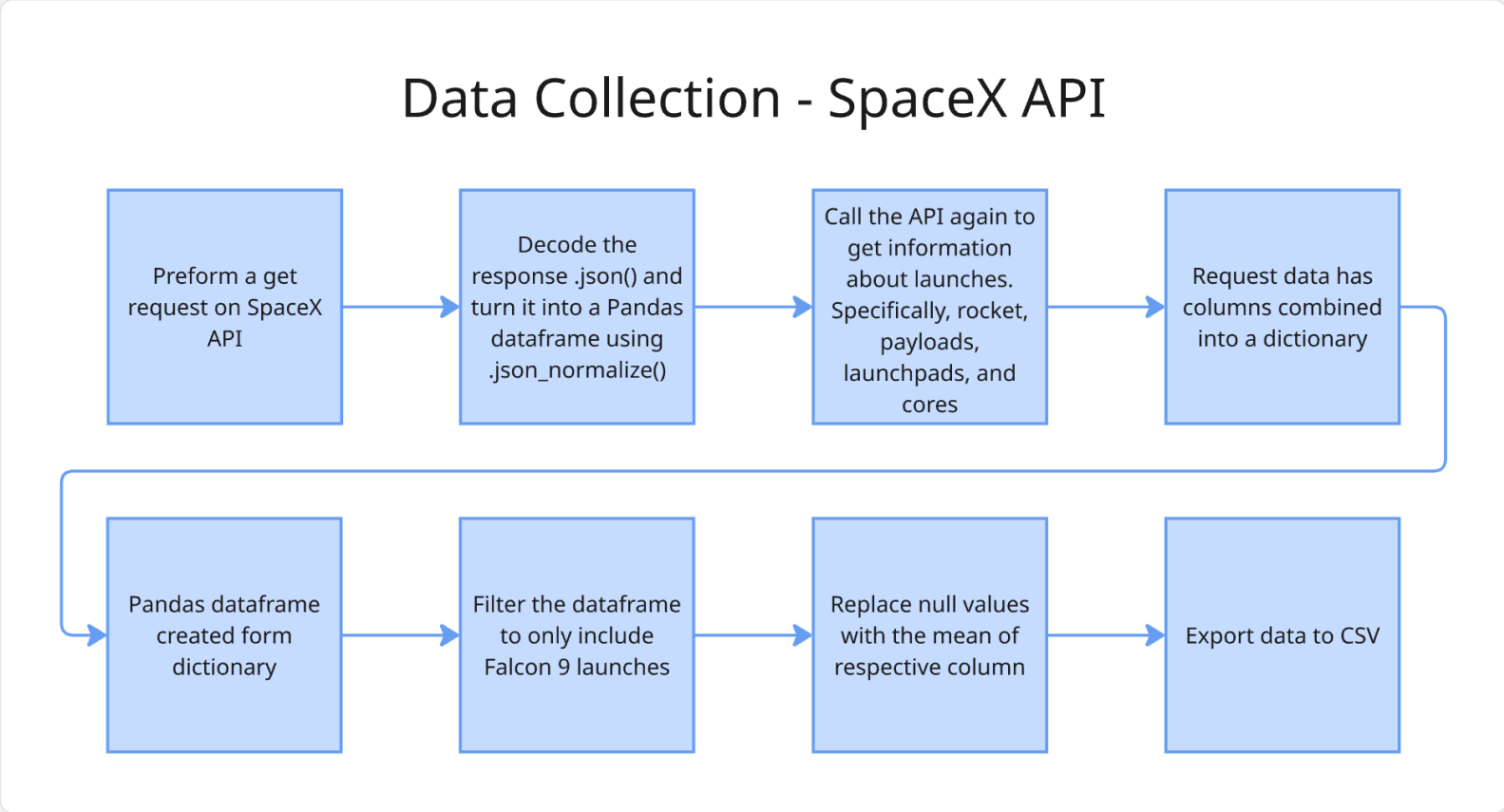
Data collected from REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, Launchsite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data collected from Web Scraping:

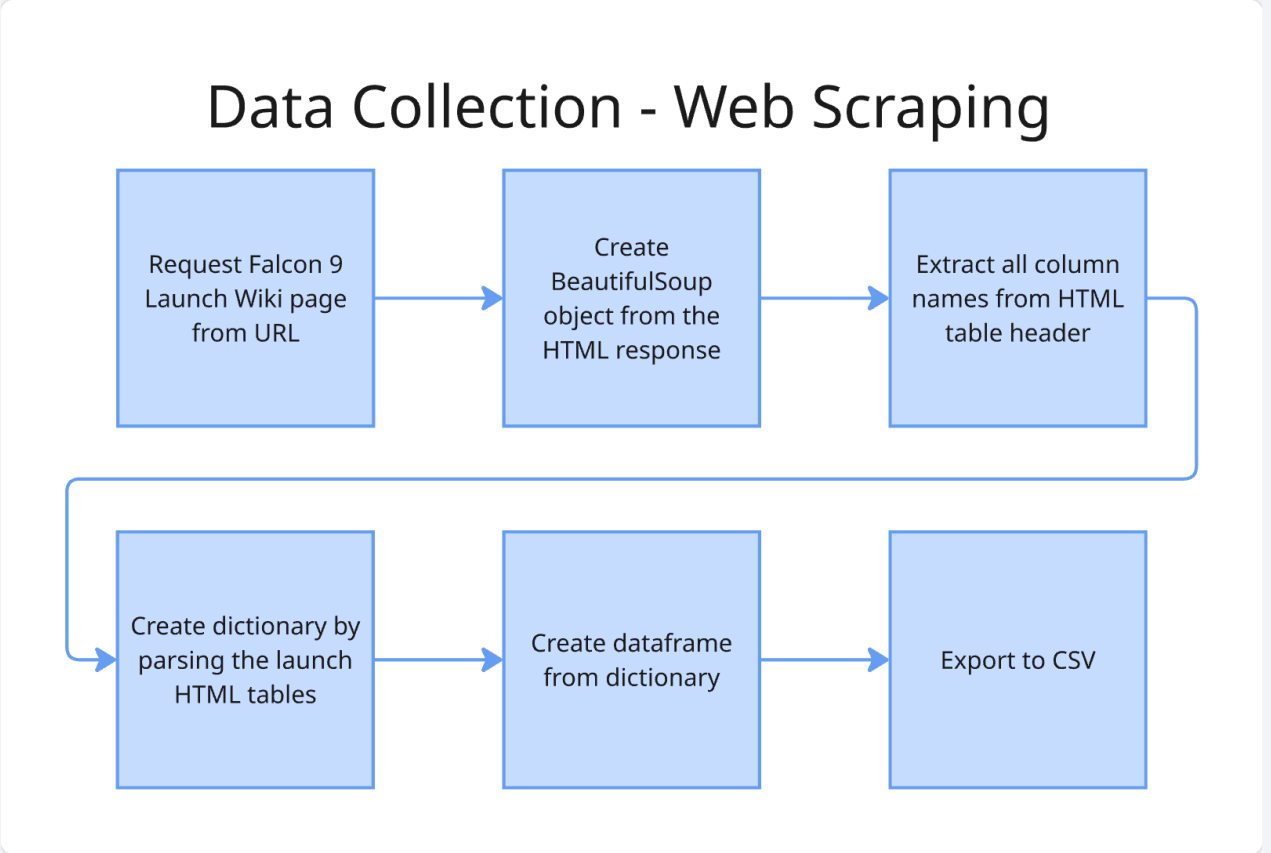
Flight Number, Launch site, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub url: SpaceX API

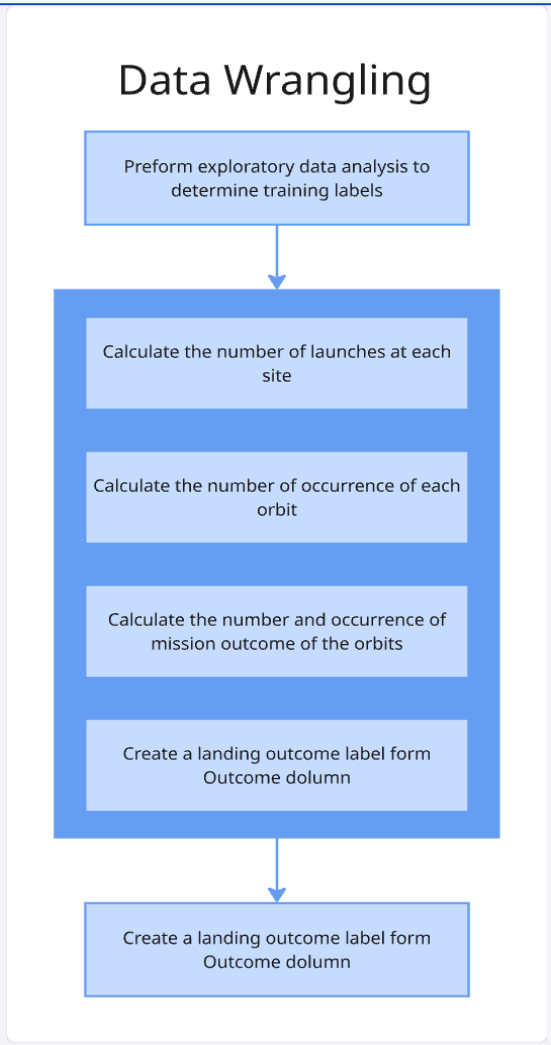
Data Collection - Scraping



GitHub url: Web Scraping

Data Wrangling

Through data wrangling, we explore the data to identify useful parameters for training. We find that launch site, payload mass, and orbit are all useful parameters. We also process our outcomes column into classification labels (1 for success and 0 for failure). This is because the original outcomes are formatted [(true, false, none) (ASDS, RTLS, Ocean, none)]. For example, True Ocean would indicate a successful landing in a specific region of the ocean, while False ASDS indicated a failed landing on a drone ship. The addition of classification variables is essential for use in a binary classification model.



GitHub url: Data Wrangling

EDA with Data Visualization

Charts Created:

Flight Number vs. Payload Mass (Scatter), Flight Number vs. Launch Site (Scatter), Launch Site vs. Payload Mass (Scatter), Success Rate vs. Orbit (bar), Orbit vs. Flight Number (Scatter), Orbit vs. Payload Mass (Scatter), Success Rate over time (line).

Scatter plots:

Visualize the relationship between variables. We are using this to determine if there is a relationship and, if so, to what extent.

Bar charts:

Shows the comparison among categories. The purpose is to show orbits with differing levels of success.

Line charts:

Show a trend in the data. In this case, there is an incremental improvement in the success rate over time.

GitHub url: EDA Data Visualization

EDA with SQL

SQL Queries Performed:

- Display the names of the unique launch sites
- Display 5 records where launch sites begin with 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CSR)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome on a ground pad was achieved
- List the names of the boosters that have been successful in drone ship and have a payload mass greater than 4000 but less than 6000
- List the total number of successful and failed mission outcomes
- List all the booster versions that have carried the maximum payload mass
- List the failed landing outcomes in drone ships, their month, booster version, and launch site, for the year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order

GitHub url: EDA with SQL

Build an Interactive Map with Folium

Map Markers of Launch Sites:

Added a marker with a circle, a pop-up label, and a text label of NASA JSC using longitude and latitude coordinates.

Added markers with circle, popup label, and text label for all launch sites using longitude and latitude to show the location and proximity to surroundings.

Colored Launch Outcome Markers for each Launch Site:

Added marker cluster to each launch site that shows the number of launches as well as whether or not they were a success (green) or a failure (red).

Distance lines from Launch Site to proximities:

Added blue lines to site VAFB SLC-4E showing proximity distance (KM) to a coastline, railway, highway, and city.

GitHub url: Interactive Map with Folium

Build a Dashboard with Plotly Dash

Launch Site Dropdown List:

Added a dropdown for launch site selection.

Pie Chart showing Success Launches (All/Select Sites):

Added a pie chart to visualize the successful launch counts for all sites and the success and failure counts for select sites.

Slider of Payload Range:

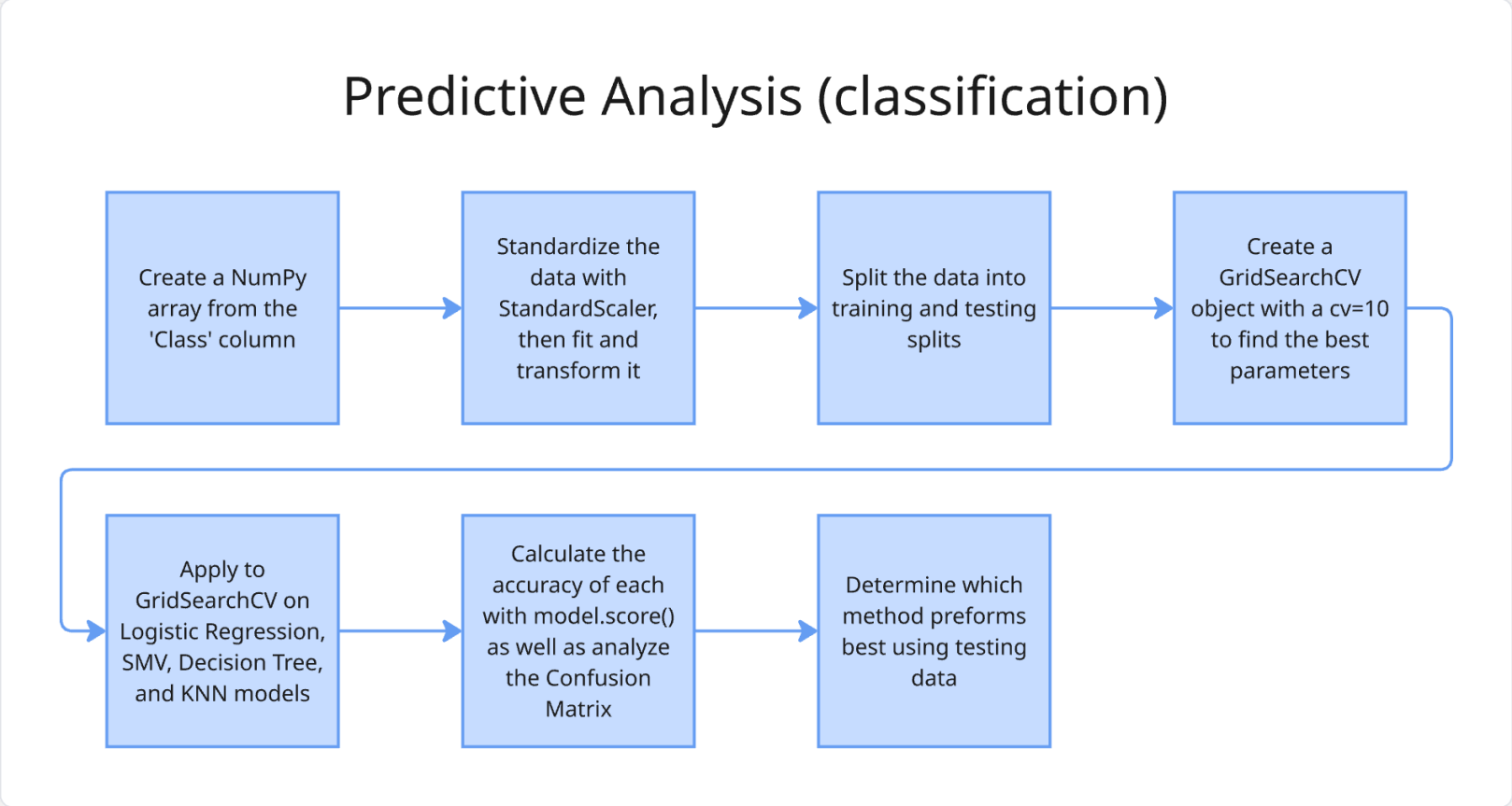
Added a slider to take in an input range for the scatter plot.

Scatter plot of Success Rates vs Payload Mass for all Booster Versions:

Added a scatter plot to show the correlation between Success and Payload for each booster. The scatter plot can also be filtered by launch site.

GitHub url: SpaceX Dash App

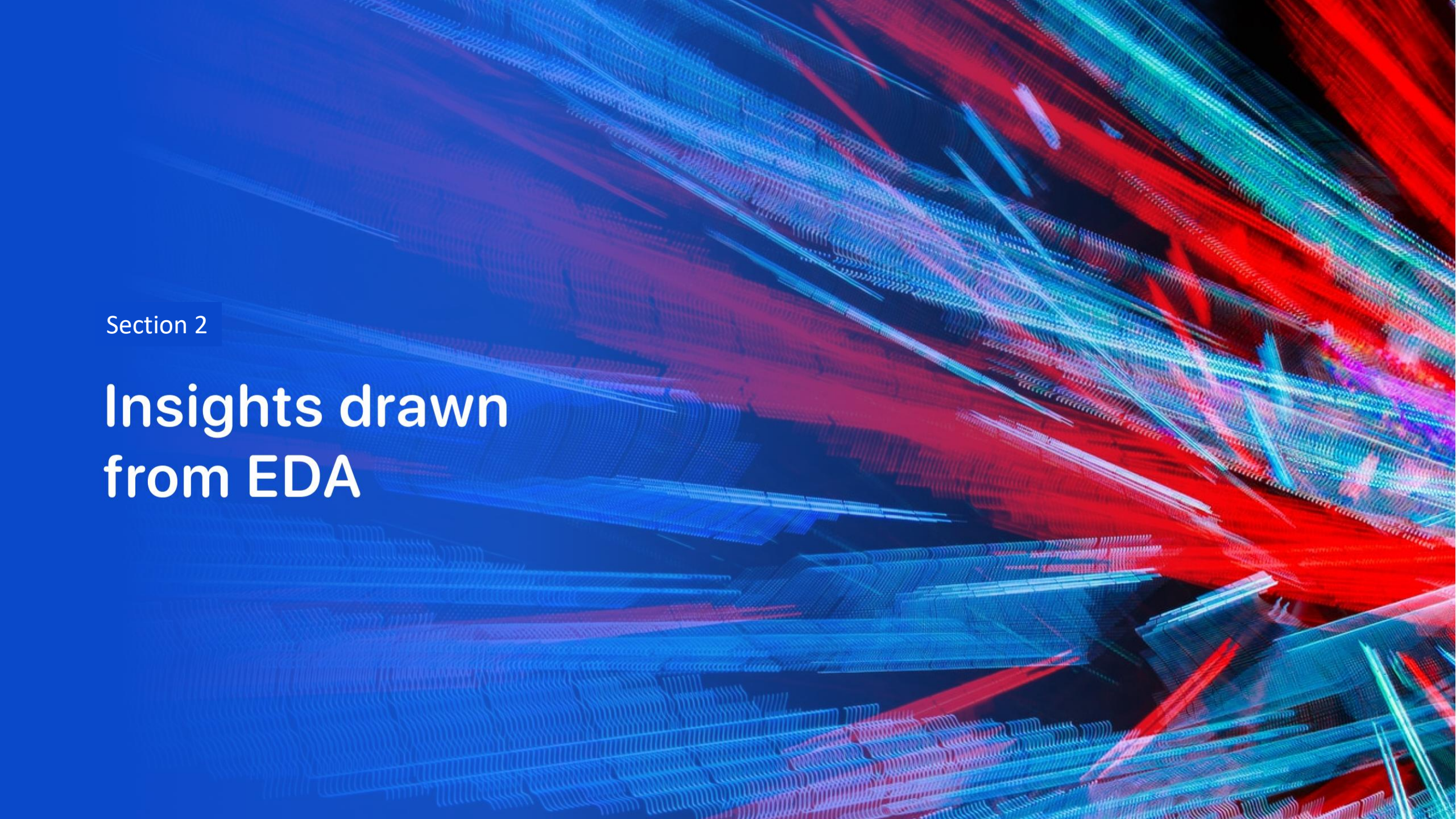
Predictive Analysis (Classification)



[Github url: Predictive Analysis (Classification)]

Results

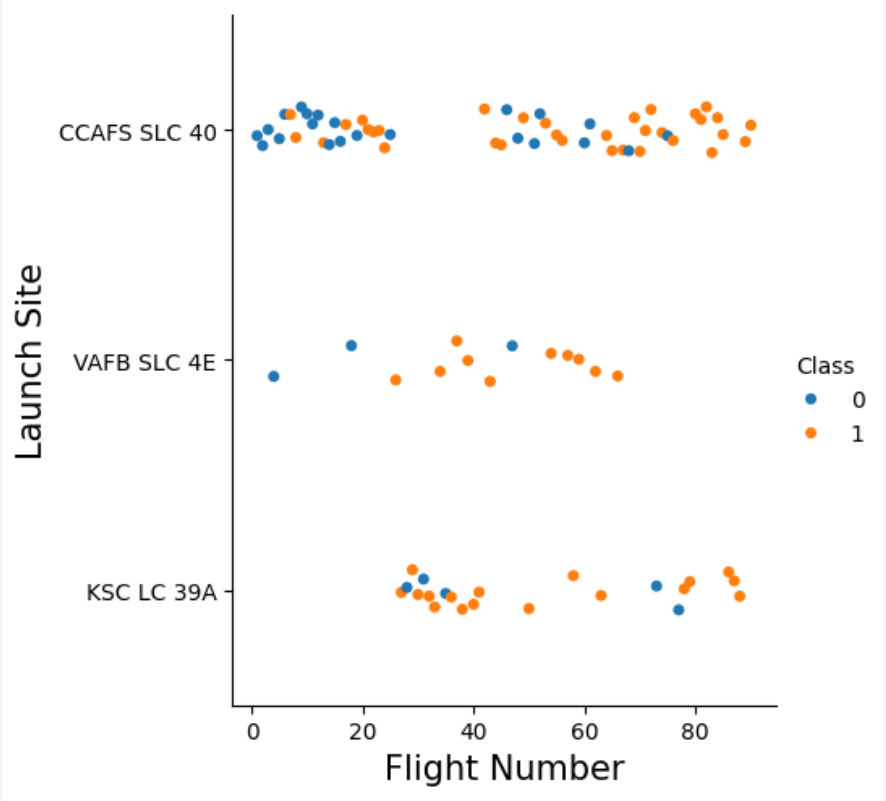
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Explanation:

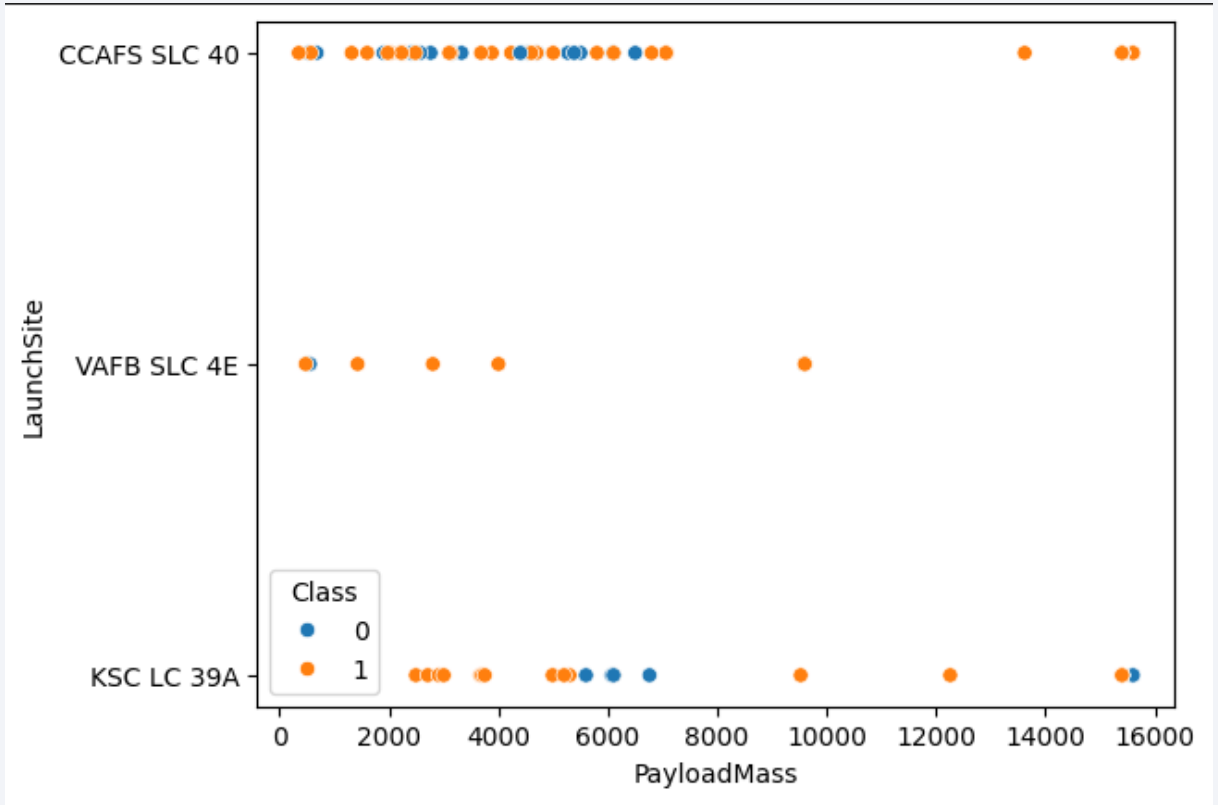
- The earlier flight tended to have more failures than the later one.
- CCAFS SLC 40 has had substantially more launches than other locations.
- VAFB SLC 4E and KSC LC 39A both have high success rates.



Payload vs. Launch Site

Explanation:

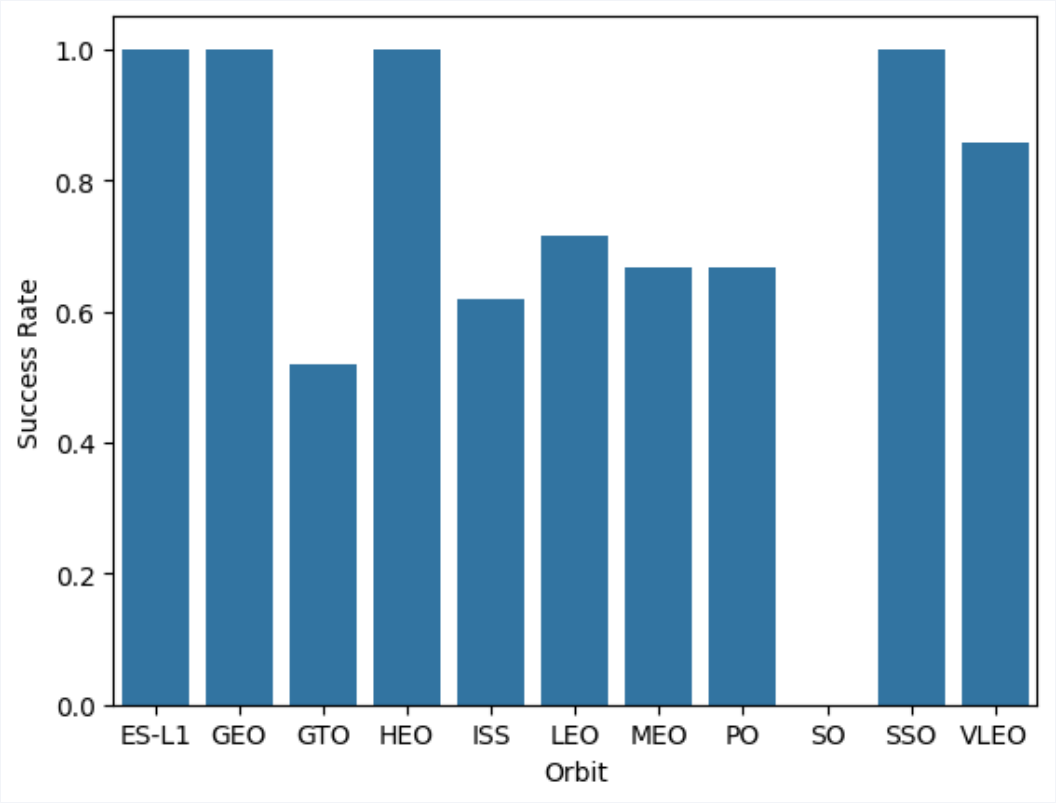
- A majority of the launches have been under 8000 kg.
- For each site, as the payload increases, the successful launches increase.
- KSC LC 39A has a 100% success rate when the payload is under roughly 5500 kg.



Success Rate vs. Orbit Type

Explanation:

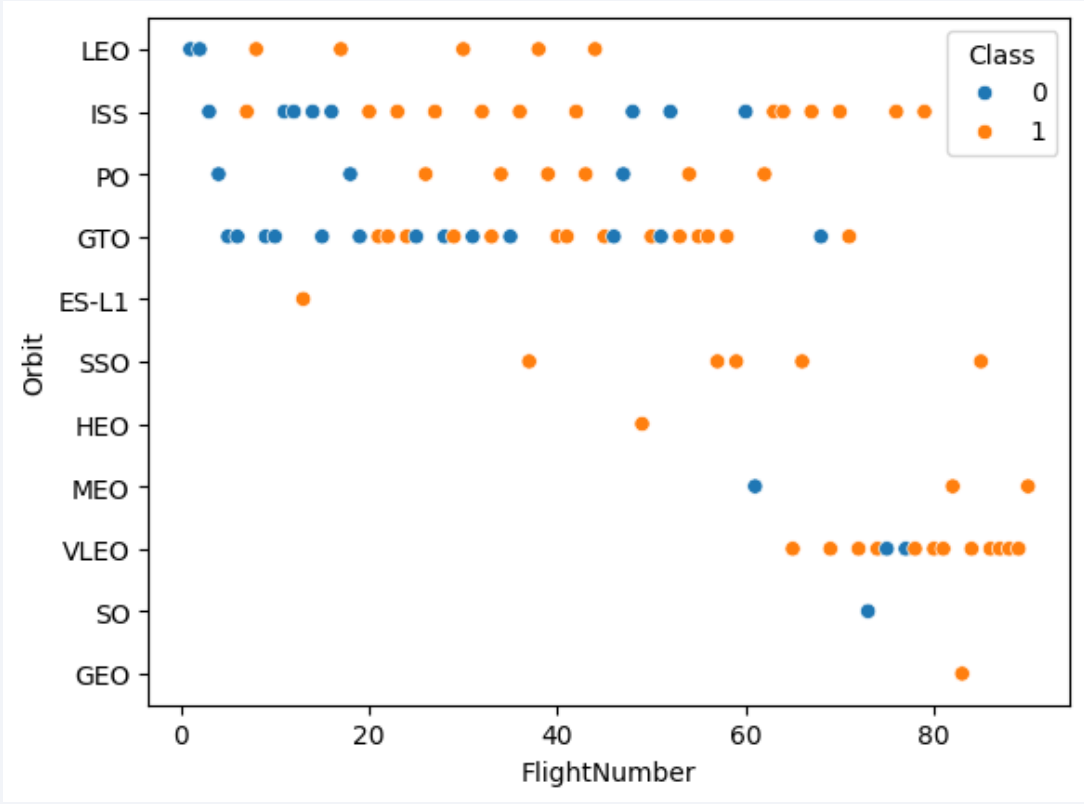
- Orbits with a success rate of 100%
  - ES-L1, GEO, HEO, SSO (SO)
- Orbits with a 50 to 85% success rate
  - GTO, ISS, LEO, MEO, PO, VLEO



Flight Number vs. Orbit Type

Explanation:

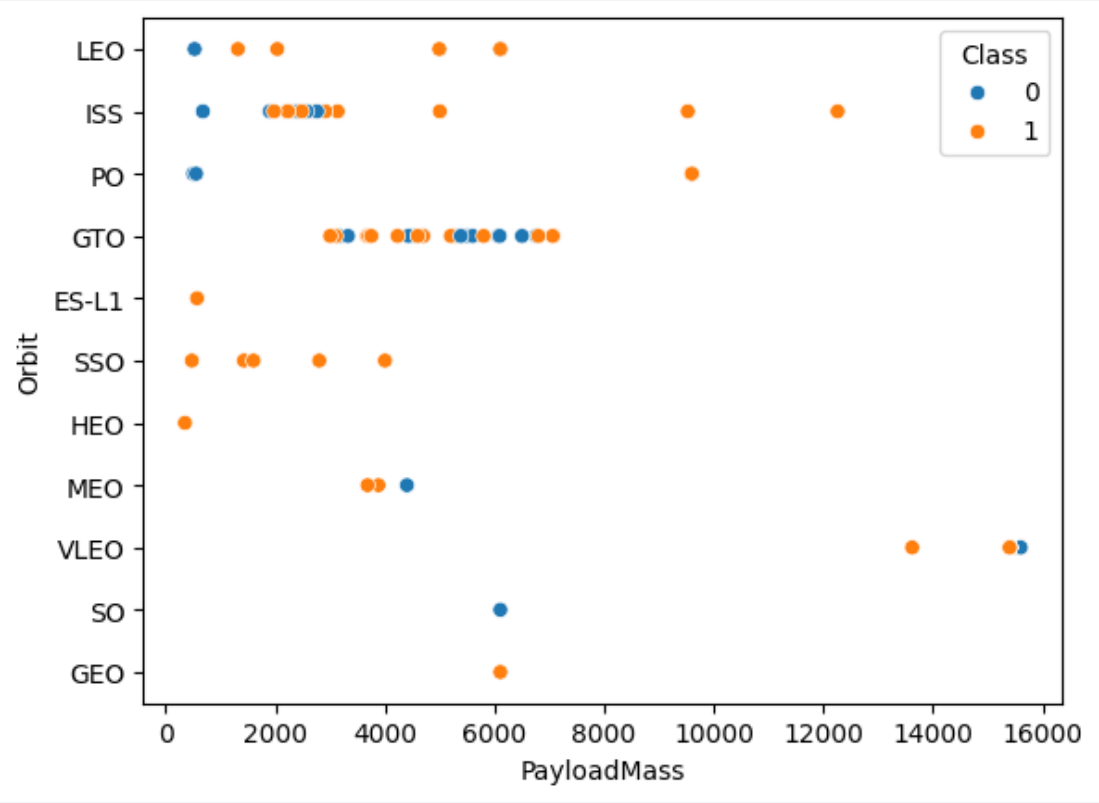
- LEO and MEO success rate increased with fights.
- ISS, PO, and VLEO go through cycles of success.
- GTO does not appear to have a relationship between success and the number of flights.



Payload vs. Orbit Type

Explanation:

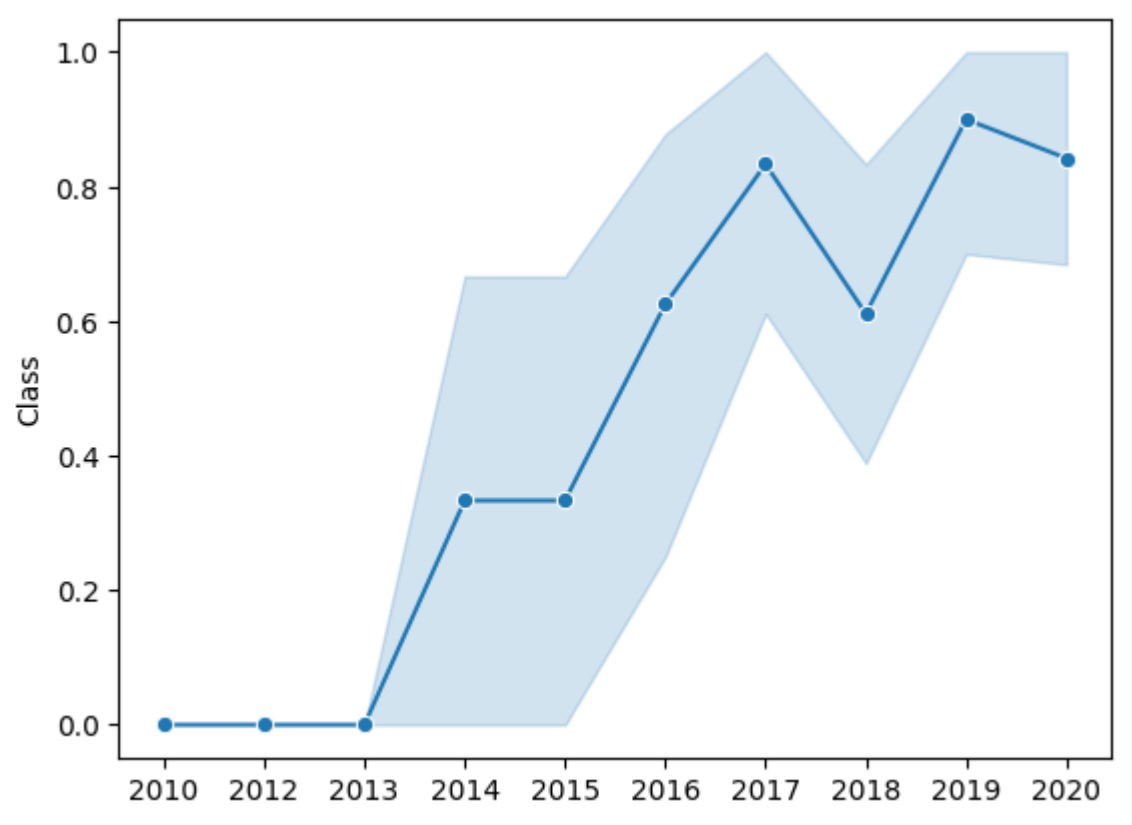
- Heavier payloads have higher success for LEO, ISS, PO.
- GTO has lower success as the payload increases.
- ISS has the most launch payloads between 2000 and 3500 kg.
- GTO has all launch payloads between roughly 3000 and 7500 kg.



Launch Success Yearly Trend

Explanation:

- The success rate has had a consistent increase from 2013 to 2020.



All Launch Site Names

```
query1 = '''
select distinct Launch_Site
from SPACEXTABLE
'''
%sql $query1
```

Python

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

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

Explanation:

- Display the distinct names of all launch sites

Launch Site Names Begin with 'CCA'

```
query2 = '''
select *
from SPACEXTABLE
where Launch_Site like 'CCA%'
limit 5
'''
%sql $query2
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation:

- Display five records where the launch site starts with ‘CCA’

Total Payload Mass

```
query3 = '''
select sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass
from SPACEXTABLE
where Customer='NASA (CRS)'
'''
%sql $query3
```

Python

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

Total_Payload_Mass
45596

Explanation:

- Display the total payload carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
query4 = '''
select avg(PAYLOAD_MASS__KG_) as Average_Payload_Mass
from SPACEXTABLE
where Booster_Version = 'F9 v1.1'
'''
%sql $query4
```

Python

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

Average_Payload_Mass
2928.4

Explanation:

- Display the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
query5 = '''
select min(Date) as First_Successful_Ground_Pad_Landing
from SPACEXTABLE
where Landing_Outcome = 'Success (ground pad)'
'''
%sql $query5
```

Python

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

First_Successful_Ground_Pad_Landing
2015-12-22

Explanation:

- Display the date of the first successful landing on a ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
query6 = '''
select distinct Booster_Version
from SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000
'''
%sql $query6
```

Python

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

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

Explanation:

- List the names of boosters that have successfully landed on drone ships and have a payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
query7 = '''
select
    sum(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 ELSE 0 END) as Success_Count,
    sum(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) as Failure_Count
from SPACEXTABLE
'''
%sql $query7
```

Python

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

Success_Count	Failure_Count
100	1

Explanation:

- List the total number of successful and failed mission outcomes

Boosters Carried Maximum Payload

```
query8 = '''
select distinct Booster_Version
from SPACEXTABLE
where PAYLOAD_MASS__KG_ =
    (select max (PAYLOAD_MASS__KG_)
    from SPACEXTABLE
    where PAYLOAD_MASS__KG_ is not null)
'''
%sql $query8
```

Python

```
 * 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

Explanation:

- List the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
query9 = '''
Select substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTABLE
where Landing_Outcome = 'Failure (drone ship)'
and substr(Date,1,4) = '2015'
'''
%sql $query9
```

Python

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

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

Explanation:

- List the failed landing outcomes on drone ships, their booster versions, and launch site names for the year 2015

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

```
query10 = '''
select Landing_Outcome, count(*) as Outcome_Count
from SPACEXTABLE
where Date between '2010-0604' and '2017-03-20'
group by Landing_Outcome
order by Outcome_Count desc
'''
%sql $query10
```

Python

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

Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Explanation:

- Ranking 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

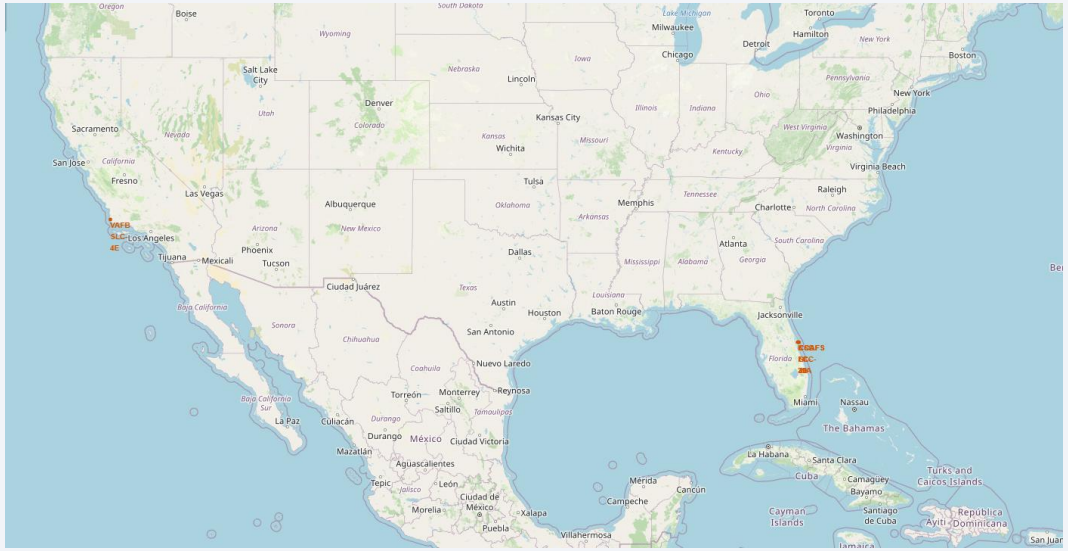
Section 3

Launch Sites Proximities Analysis

Launch Site Locations

Explanation:

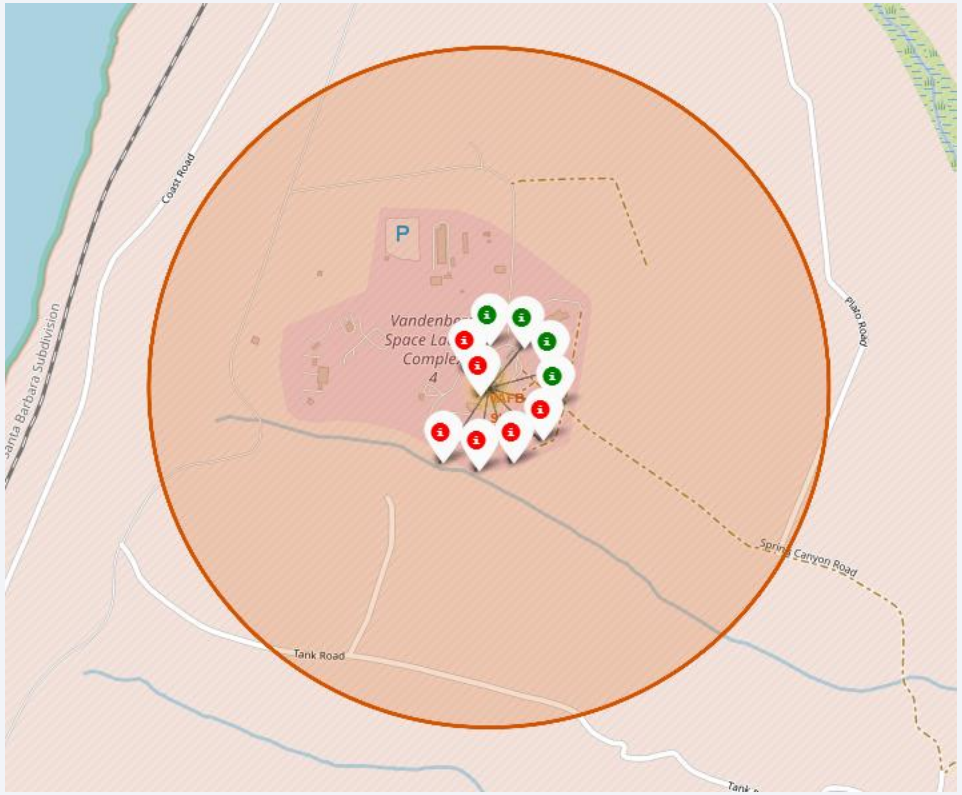
- There are four launch sites in total. Three of which are located within a couple of miles of each other in Florida, while there is one in California. All sites are within 6 degrees of latitude of each other, with the southernmost being 28 degrees from the equator.
- All launch sites are in very close proximity to a coastline. This allows for rockets to be tested over water, where the risk to civilian life can be minimized in the event of failure.



Color-coded Launch Markers

Explanation:

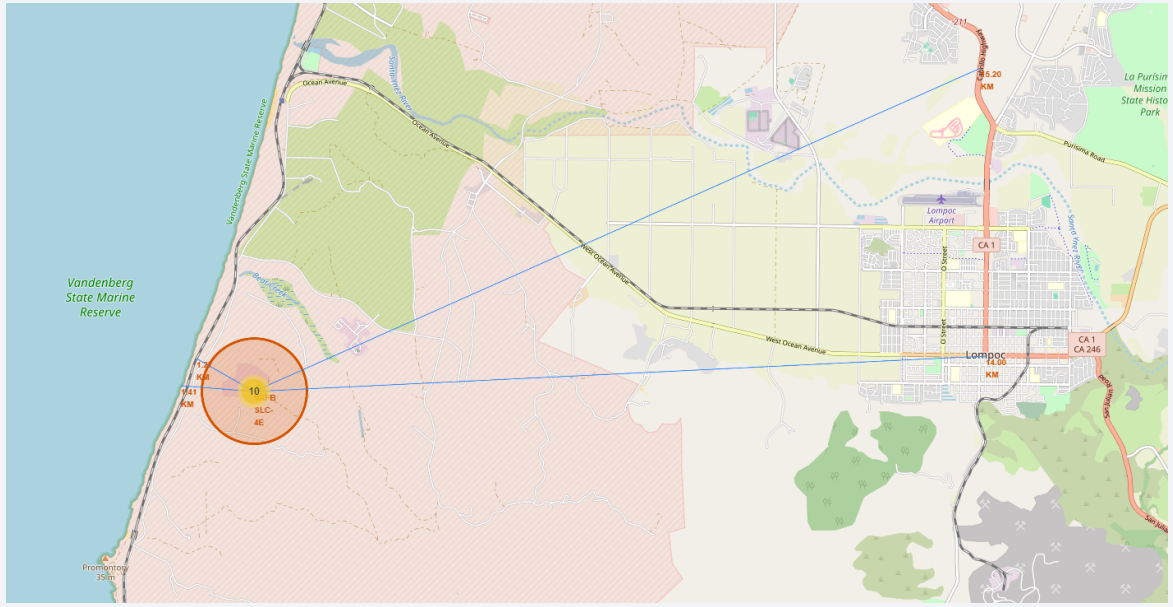
- From the addition of color-coded markers, we can quickly and easily identify the quantity of launches and their outcomes.
  - Green Markers for Success
  - Red Markers for Failure
- Launch site VAFB SLC-4E has had a total of 10 launches, with 4 of them successful.

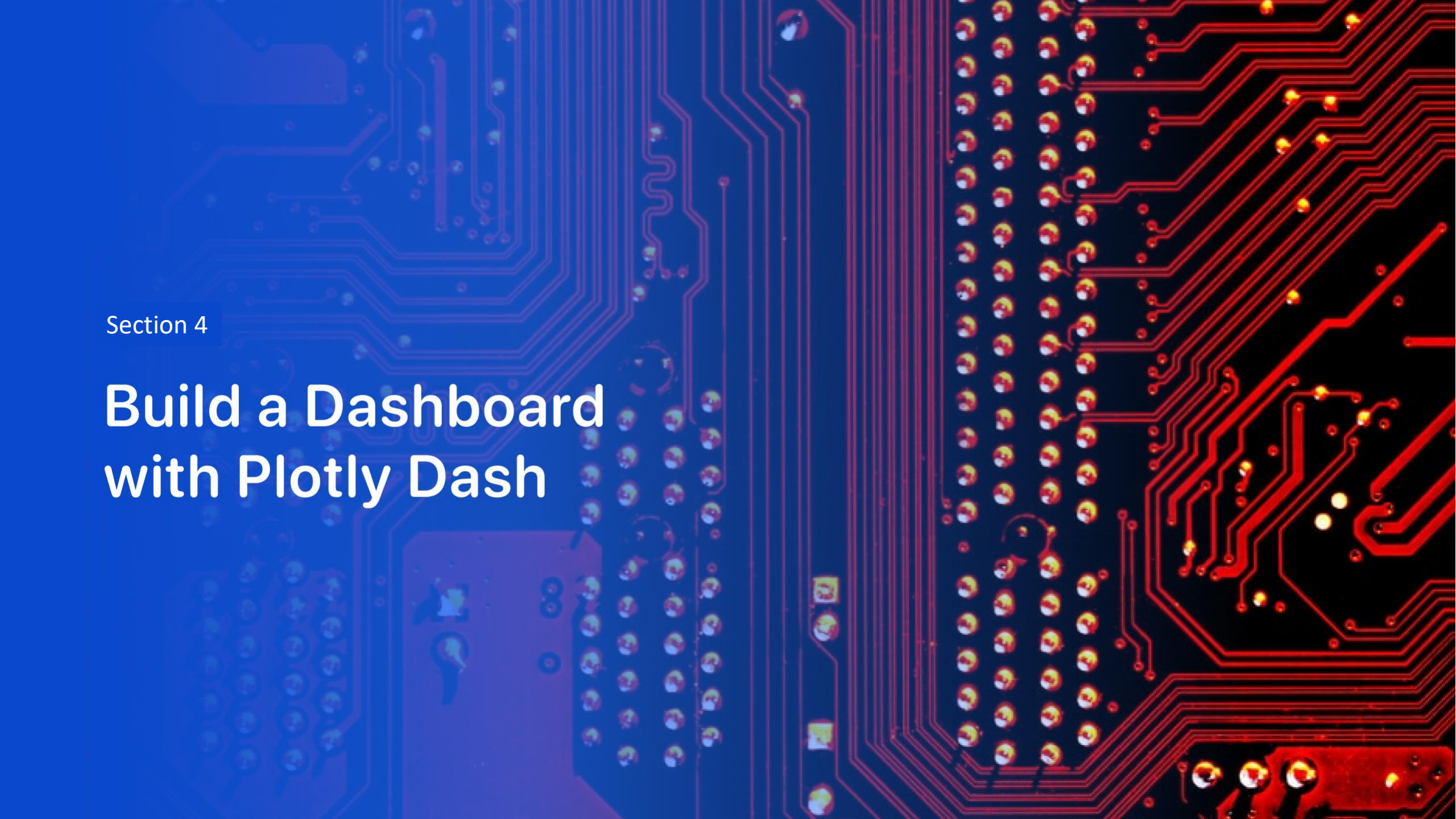


Distance to Proximities for Launch Site VAFB SLC-4E

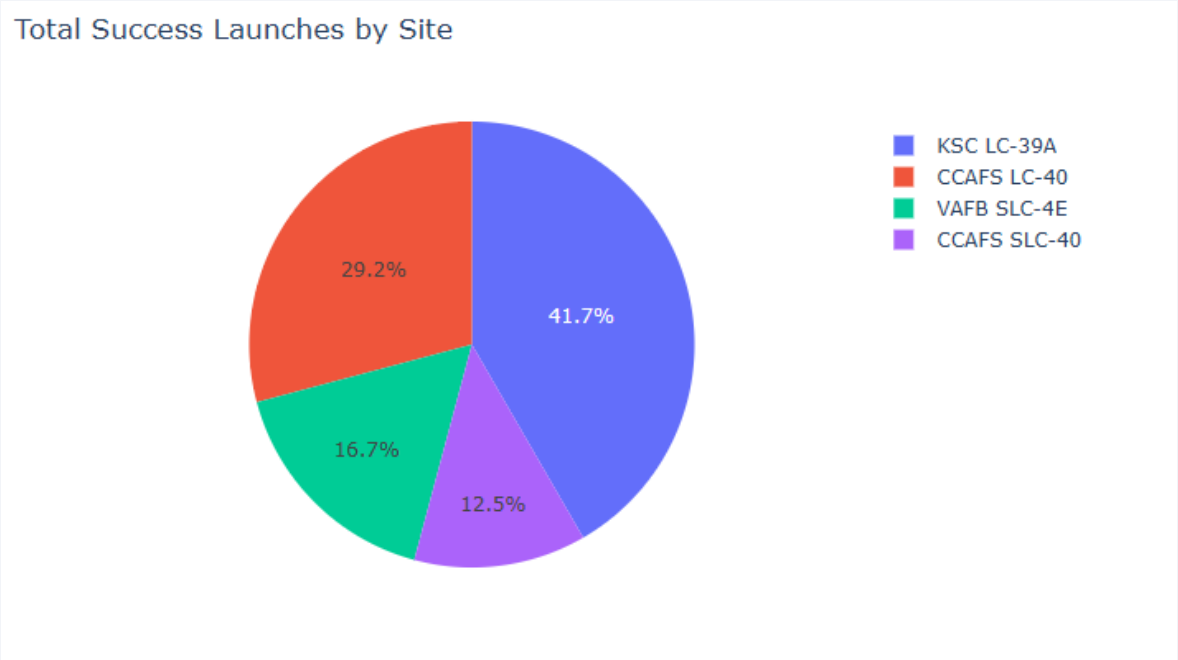
Explanation:

- For the map of launch site VAFB SLC-4E, we can see
  - Railway is 1.27 km away
  - Coastline is 1.41km away
  - City (Lompoc) is 14 km away
  - Highway is 15.2 km away
- The launch site is in proximity to civilian cities and infrastructure. The railway is used for civilian transport, and the city could be affected by a failure given its proximity.




Section 4
Build a Dashboard with Plotly Dash

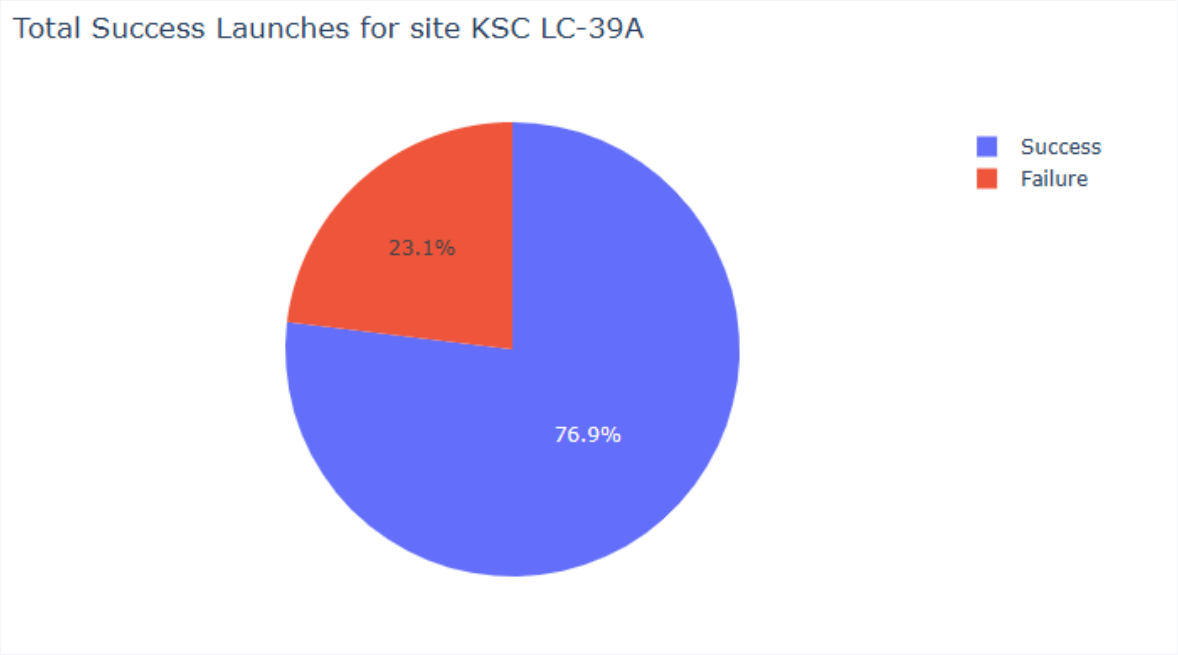
Launch Success for all Sites



Explanation:

The chart shows that out of all sites, KSC LC-39A has the highest success with 41.7% of all successful launches coming from this site.

Launch Site with the Highest Success



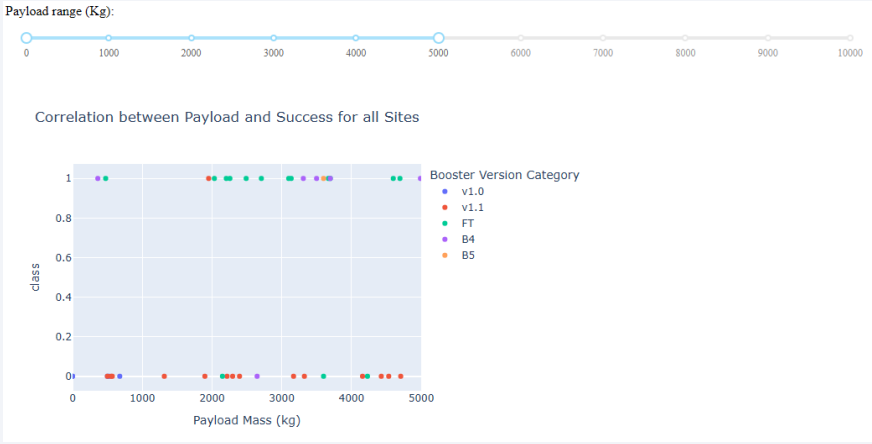
Explanation:

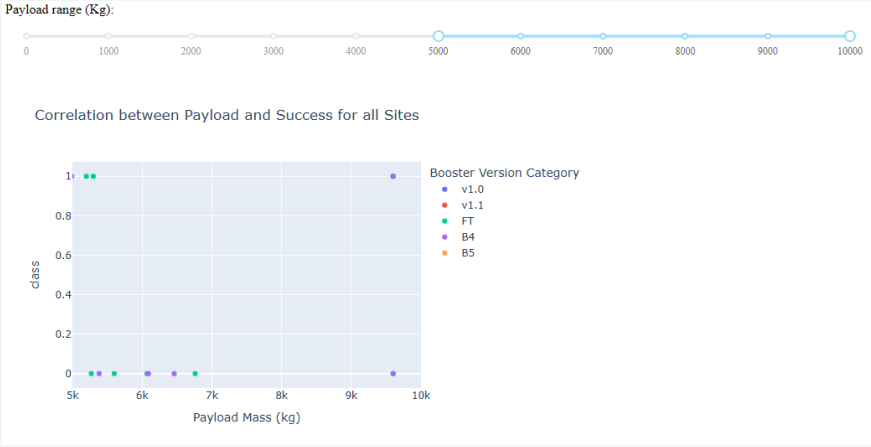
KSC LC-39A has the highest success rate of 76.9 % with 10 successful and 3 failed launches.

Launch Outcome vs. Payload Mass for all Sites

Explanation:

- The graphs show successful flights for different payload masses and booster versions.
- Payloads between 2000 and 5500 kg, along with booster versions B4 and FT have the most success.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- Based on the scores from the test set of data, we can see that the decision tree model performs better than the others.
- The scores from the entire dataset continue to prove this, with the decision tree model having the highest scores and accuracy.

Scores for testing set:

	F1 Score	Precision Score	Recall Score	Accuracy Score
Logistic Regression	0.814815	0.866667	0.833333	0.833333
SVM	0.814815	0.866667	0.833333	0.833333
Decision Tree	0.888889	0.888889	0.888889	0.888889
KNN	0.814815	0.866667	0.833333	0.833333

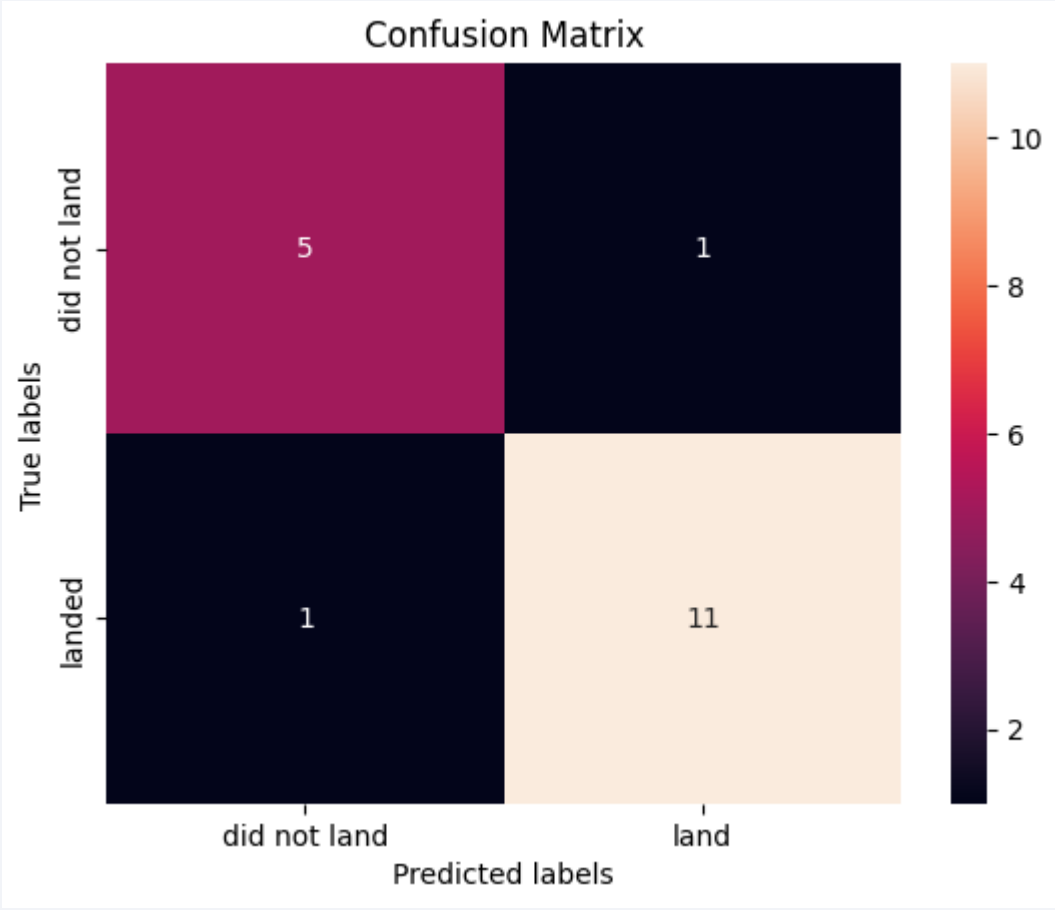
Scores for prediction:

	F1 Score	Precision Score	Recall Score	Accuracy Score
Logistic Regression	0.856061	0.888889	0.866667	0.866667
SVM	0.869190	0.896714	0.877778	0.877778
Decision Tree	0.923077	0.926103	0.922222	0.922222
KNN	0.845407	0.869780	0.855556	0.855556

Confusion Matrix

Explanation:

- From looking at the Matrix, it can be seen that the model predicted 16 out of the 18 launches correctly. The model's primary issue is that it predicts minimal false positives and negatives.

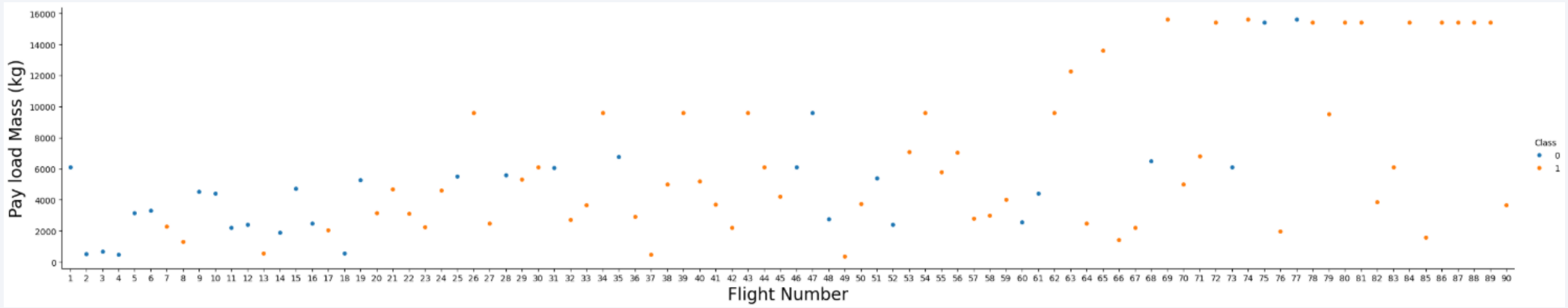


Conclusions

- The decision tree model is the best for classifying that data with a 92% accuracy
- All launch sites are within very close proximity to a coastline
- KSC LC-39A has the highest success rate out of all launch sites
- Most successful rockets have a payload mass between 2000 and 5500 kg, with a booster version of B4 or FT
- Orbits ES-L1, GEO, HEO, and SSO have a success rate of 100%

Appendix

Additional Graph showing Payload vs. Flight Number



Thank you!