IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joe Davis
May 28, 2025

Outline

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

Executive Summary

Summary of methodologies utilized:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an Interactive Map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis

- Summary of all results from above methodologies:
  - Exploratory Data Analysis Results
  - Interactive Analytics via Screenshots
  - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX advertises on its website that the Falcon 9 rocket cost 62 million dollars, with other providers costing as much as 165 million dollars.
  - SpaceX saves $s in the resuse of the first stage engine. IF a determination can be made when the first stage will land safely, the true cost of a launch and ability to minimize launch risks can be determined and these findings can be applied to Space Y's own program.
- Problems you want to find answers
  - For a given set of features related to Falcon 9 rocket launches, will the first stage of the rocket land successfully?

Section 1

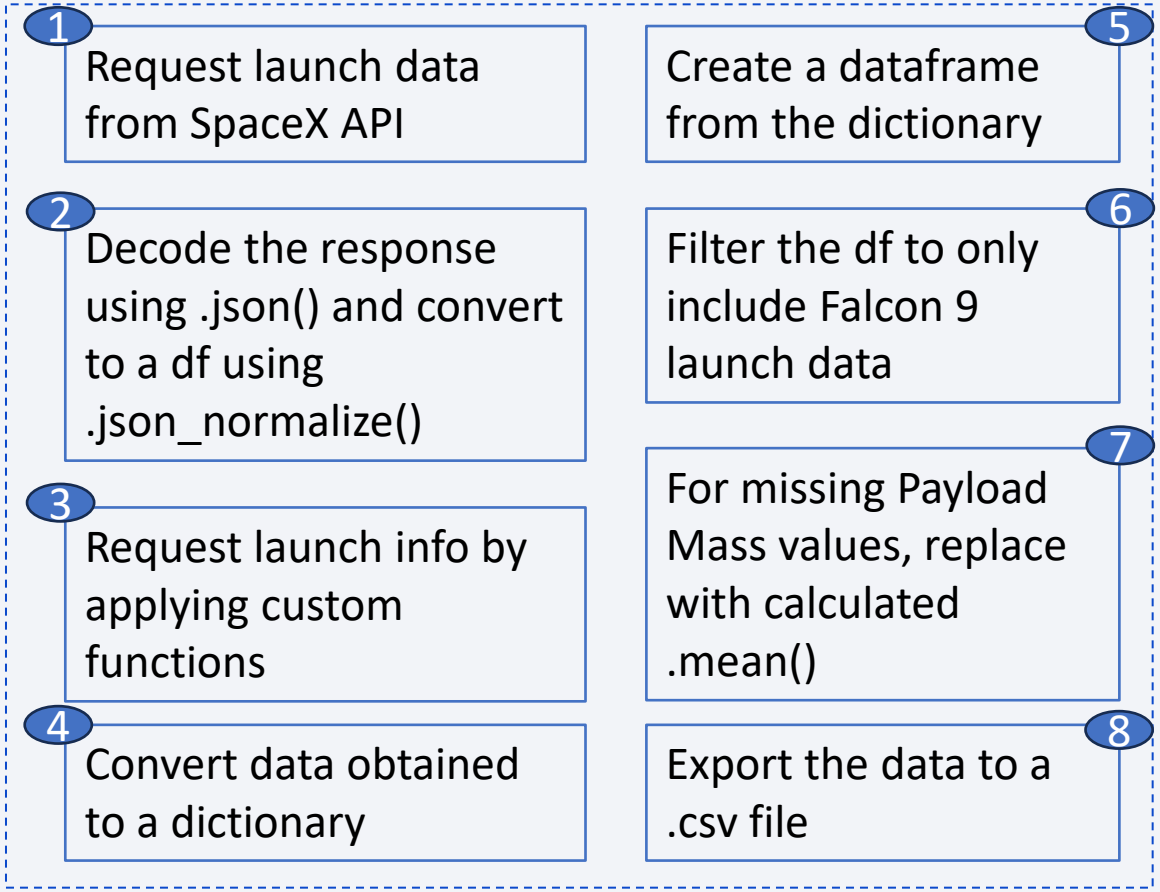
Methodology

Methodology

- Data collection methodology:
  - Used SpaceX Rest API
  - Used Web Scraping from Wikipedia
- Perform data wrangling
  - Filtered the data
  - Dealt with missing values
  - Used One Hot Encoding to prepare the 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
  - Built, tuned, and evaluated classification models to ensure the best findings

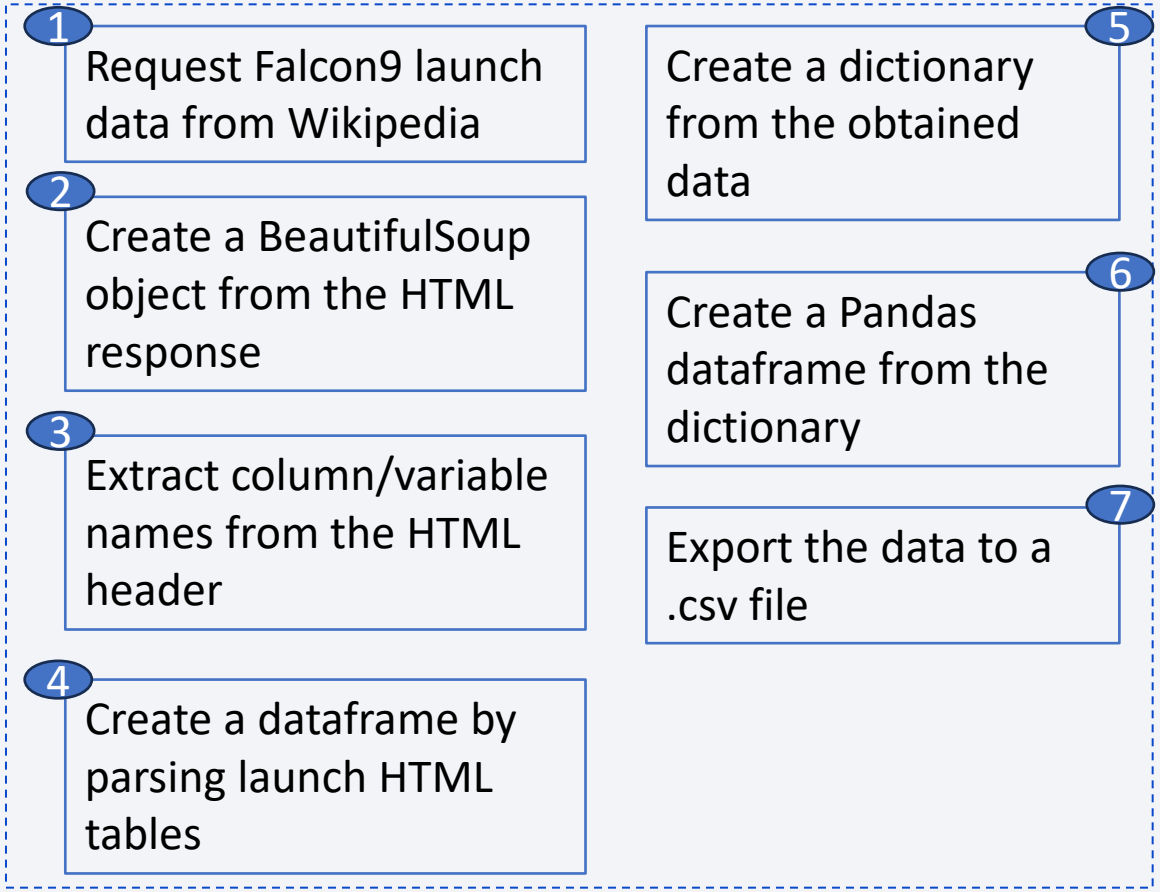
Data Collection – SpaceX API

- API
  - Acquired historical data from Open Source REST API for SpaceX
    - API utilized is https://api.spacexdata.com/v4/rockets/
    - Requested and parsed the launch data using the GET request
    - Filtered the dataframe to only include Falcon 9 launches
    - Replaced missing payload mass values from classified missions with mean
- GitHub URL: Data Collection API



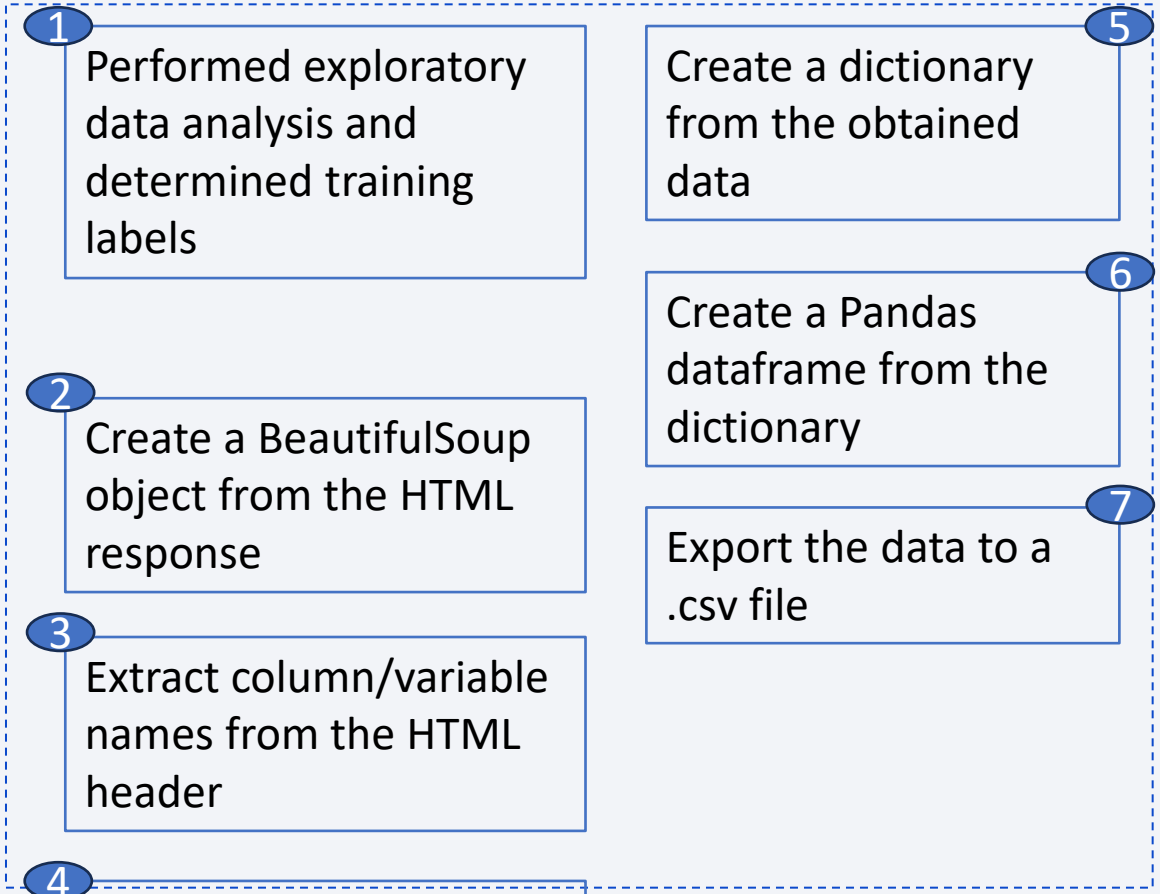
Data Collection – Web Scraping

- Web Scraping
  - Acquired historical data from Wikipedia for “List of Falcon 9 and Falcon Heavy launches”
    - Wikipedia data link utilized: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - Extracted all columns/variable names from the HTML table header
    - Parsed the table and converted it to a Pandas dataframe
- GitHub URL: Data Collection via Web Scraping



Data Collection – Data Wrangling

- Web Scraping
  - Acquired historical data from Wikipedia for “List of Falcon 9 and Falcon Heavy launches”
    - Wikipedia data link utilized: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - Extracted all columns/variable names from the HTML table header
    - Parsed the table and converted it to a Pandas dataframe
- GitHub URL: Data Wrangling



EDA with Data Visualization

- Various charts were plotted:
  - Flight Number vs Payload Mass (Scatter Plot), Flight Number vs Launch Site (Scatter Plot), Payload Mass vs Launch Site (Scatter Plot), Success Rate for Each Orbit Type (Bar Chart), Flight Number vs Orbit Type (Scatter Plot), Payload Mass vs Orbit Type (Scatter Plot), Launch Success Rate by Year (Line Chart)
- The majority of charts created were of the Scatter variety. These show the relationship (potential or concrete) between variables. If a solid relationship, can be utilized for machine learning.
- Bar charts are used to show comparisons within discrete categories. Shows measure values within a category (or categories).
- Line charts show data trends over time.
- GitHub URL: EDA with Data Visualization

EDA with SQL

- SQL queries performed:
  - Names of unique launch sites in the space mission
  - Displayed 5 records where launch sites begin with the string ‘CCA’
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1
  - Listed the date when the first successful landing outcome in a ground pad was achieved
  - Listed the names of the boosters with success in drone ship and payload mass >4k and <6k
  - Listed the total number of successful and failed mission outcomes
  - Listed all the Booster Versions that carried the maximum payload mass
  - Listed records displaying month names, failure landing outcomes in drone ship, booster versions, and launch site for year = 2015
  - Ranked the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

- GitHub URL: EDA with Data Visualization

Build an Interactive Map with Folium

- Markers for Launch Sites
  - Added marker with a Circle, Popup Label, and Text Label for NASA Johnson Space Center using the center's lat and long coordinates as the start location
  - Added markers with Circle, Popul Labels, and Text Labels for all Launch Sites using their respective lat/long coordinates to show their locations relative to the Equator and nearest coasts
- Markers colored to designate success and failures were added (Green and Red) to show which Launch Sites have high success rates
- Colored lines were added between Launch Sites and nearest landmarks (e.g. Railways, Highways, Coastlines, Cities) to show proximity to these locations
- GitHub URL: Interactive Visualizations with Folium

Build a Dashboard with Plotly Dash

- Created a Launch Sites Dropdown List
  - A dropdown list to enable Launch Site user selection
- Created a Pie Chart showing Successful Launches
  - Ability to select for All Sites or Individual Site at user discretion
- Created a Slider of Payload Mass Range
  - Enables the user to see results with a quick slider function for various Mass values
- Created a Scatter Chart showing Payload Mass vs Success Rate
  - Shows correlation between Payload Mass and Launch Success Rate for different booster versions
- GitHub URL: Plotly Dash

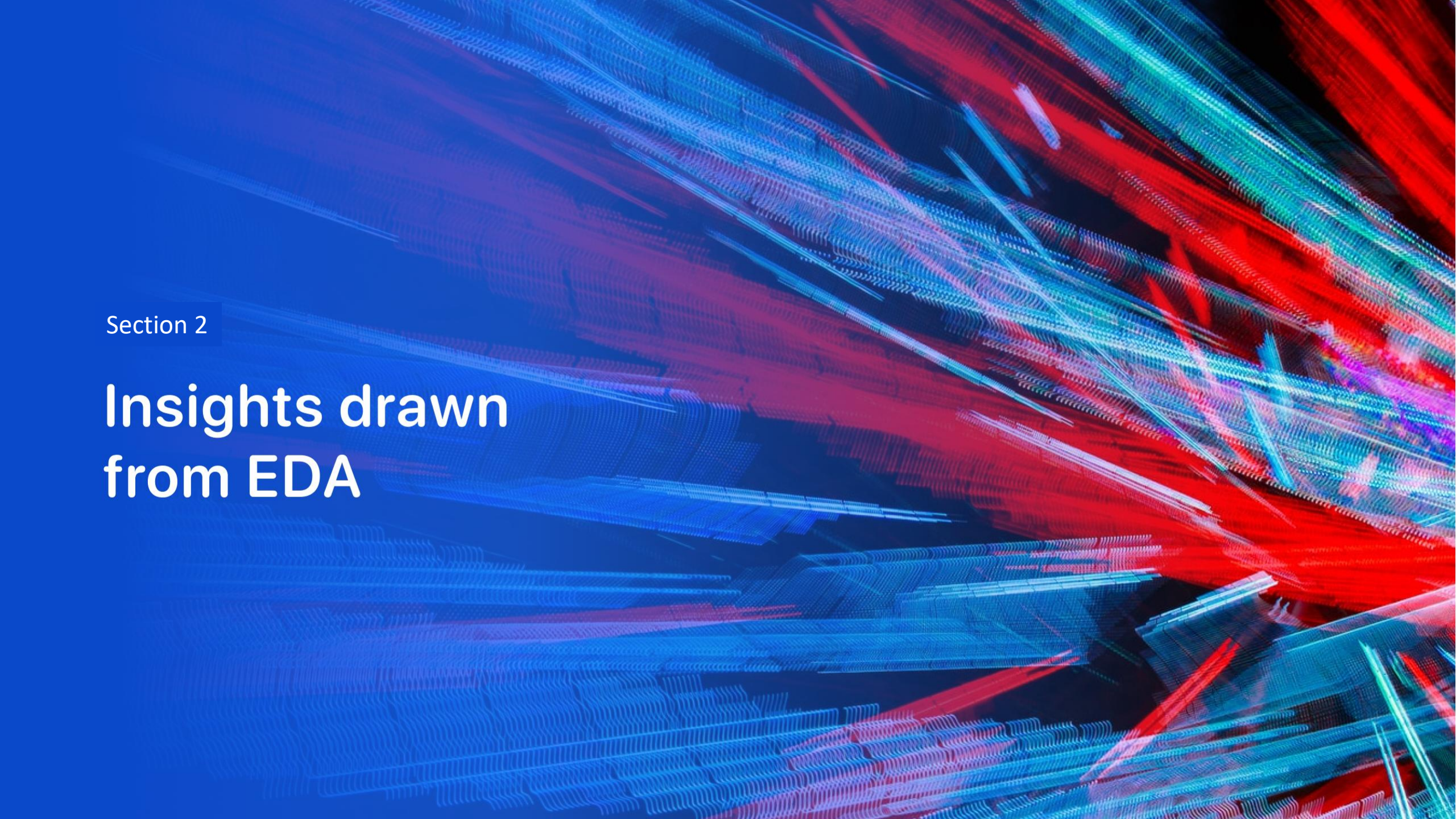
Predictive Analysis (Classification)

- Utilized Train/Test machine learning to analyze the data using various methods.
- GitHub URL: Machine Learning

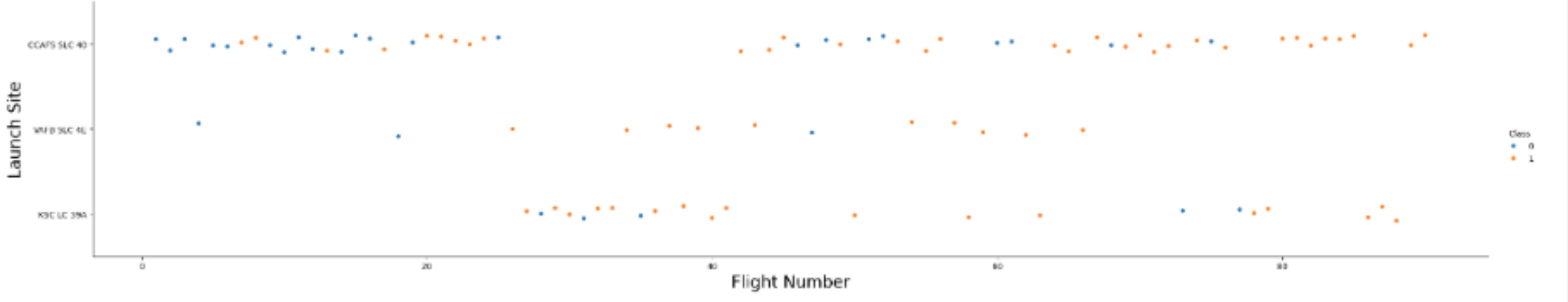
1. Created a NumPy arrach from the column “Class” in the dataset
2. Standarized the data in “X” and reassigned it to variable “X” using StandardScaler
3. Split data into training and test sets using train_test_split function
4. Created logistic regression object then a GridSearchCV object (cv=10)
5. Calculated accuracy of test data using “score” method
6. Examined confusion matrices
7. Determined best performing method: all performed well. Training data slightly better.

Results

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

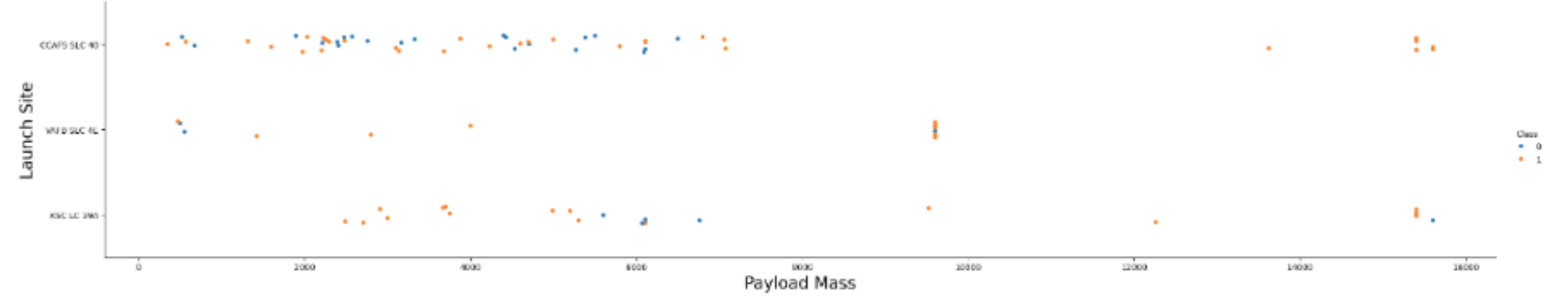

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



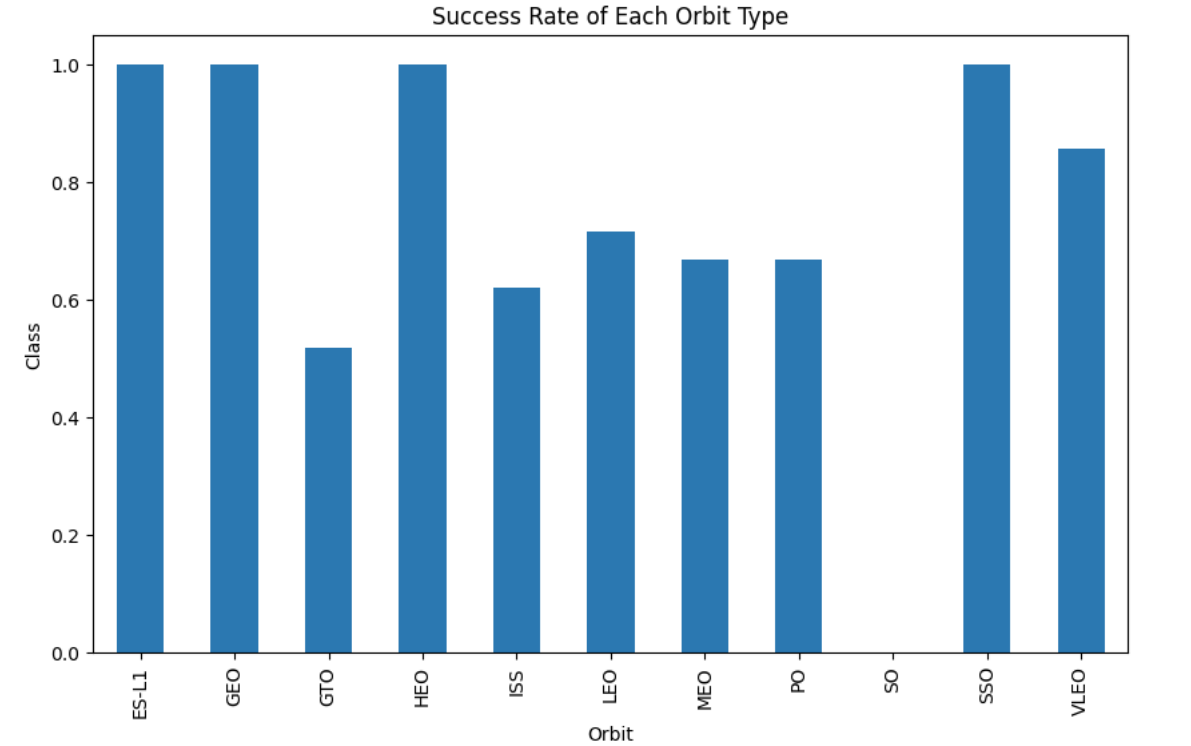
- Explanations:
  - Mixed success rates for earlier flights.
  - ~50% of launches from CCAFS SLC 40 site
  - Higher success rates at other 2 locations
  - Over time, ratio of success to failure has improved. Net new launches show “success” as a probable outcome

Payload vs. Launch Site



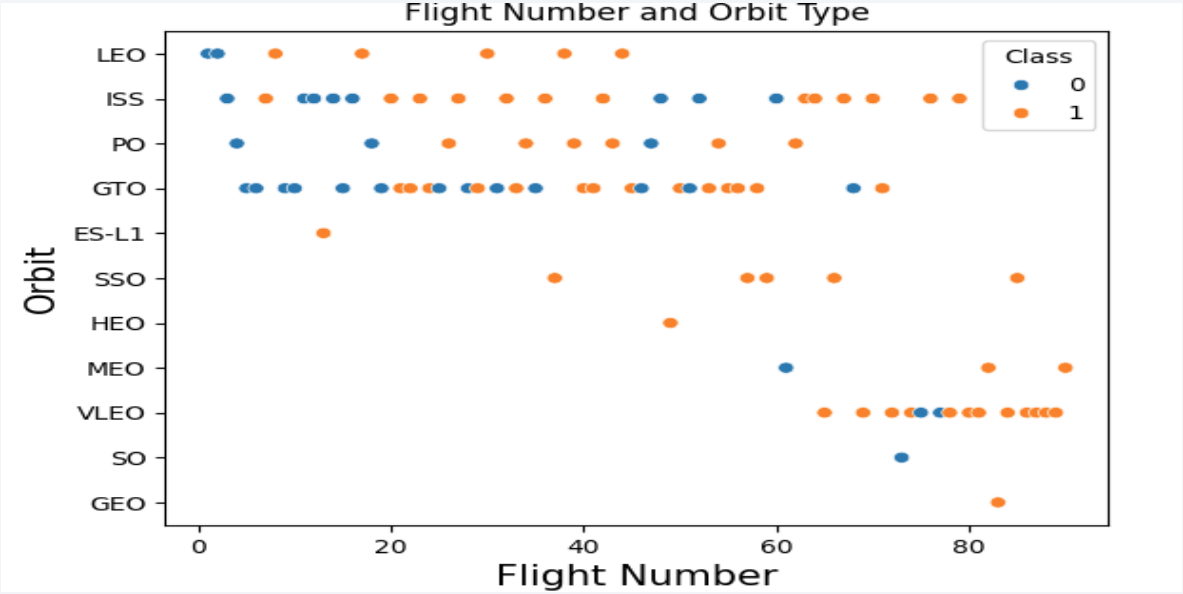
- Explanations:
  - Majority of payloads over 2k were successful
  - The higher the payload (across all launch sites), the higher the success rate
  - KSC LC 39A shows a near 100% success rate for loads <6k

Success Rate vs. Orbit Type



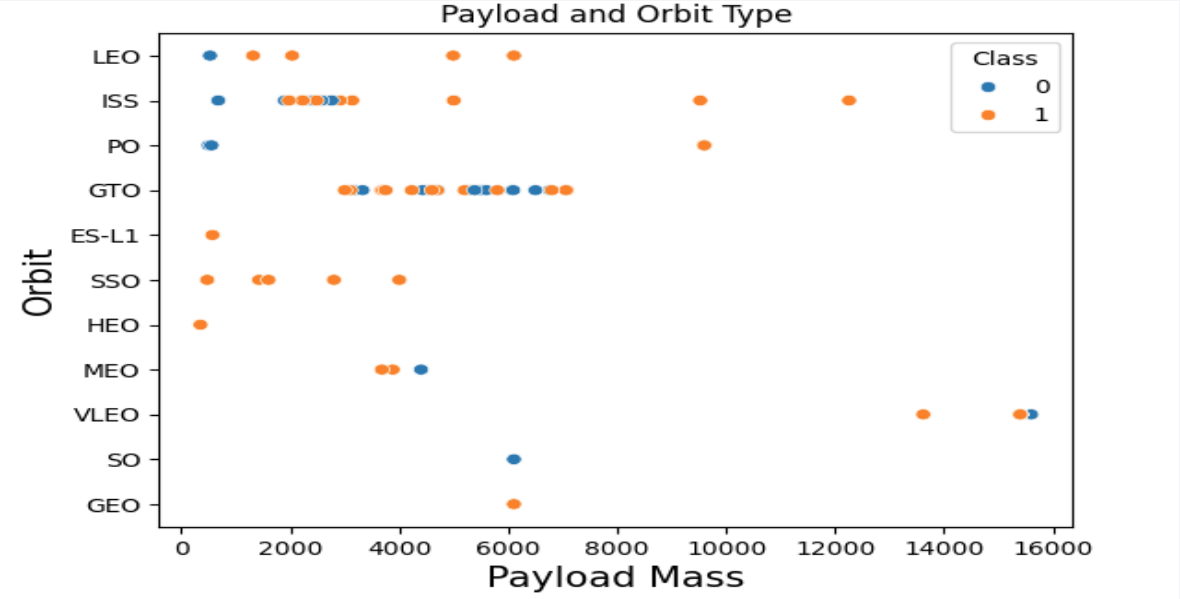
- Explanations:
  - Highest success rates (100%) for orbit types ES-L1, GEO, HEO, SSO
  - SO shows a 0% success rate
  - Intermediary success (>50%) with remaining orbit types (GTO, ISS, LEO, MEO, PO, VLEO)

Flight Number vs. Orbit Type



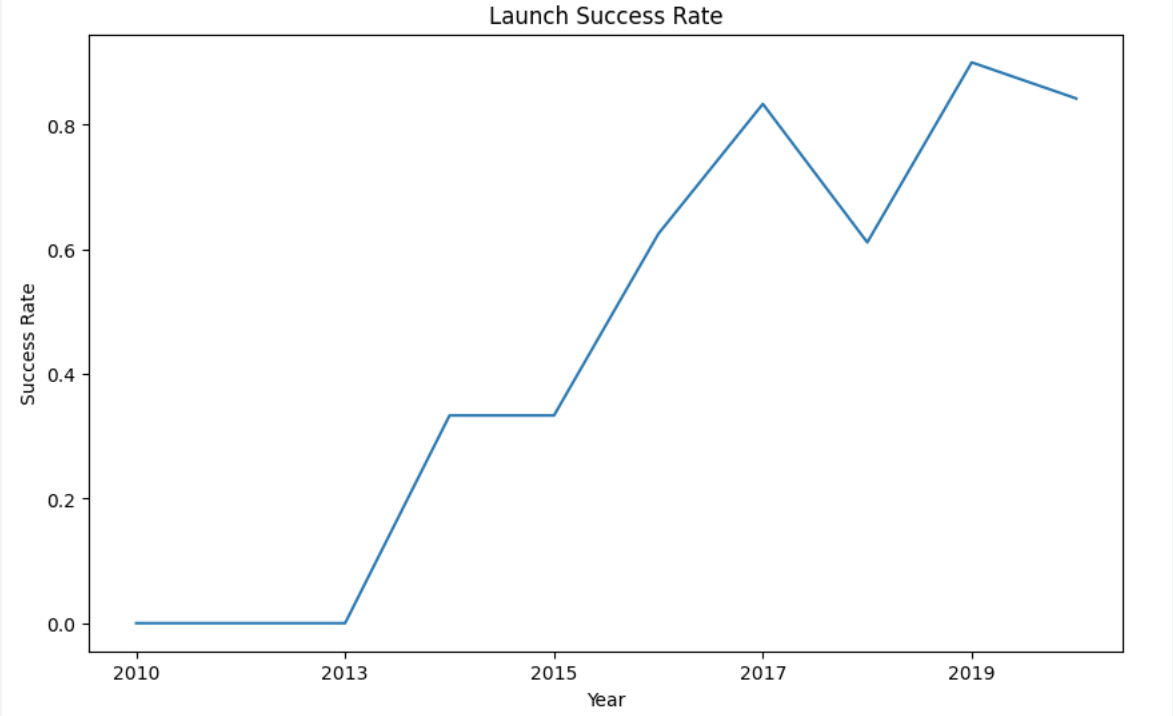
- Explanations:
  - LEO/VLEO orbit types have a high success rate based on number of launches
  - GTO does not seem to have a direct success correlation based on flight numbers

Payload vs. Orbit Type



- Explanations:
  - GTO orbit type displays mixed (trending towards positive) results between 2.5k and 7k
  - LEO, MEO, and VLEO show positive results with a negative outlier VLEO at high mass

Launch Success Yearly Trend



- Explanations:
  - Success rate shows an overall positive increase over time

All Launch Site Names

```
In [10]: %sql select distinct Launch_Site from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[10]:

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

- Explanations:
  - 4 unique Launch Site name values from SpaceX dataset

Launch Site Names Begin with 'CCA'

```
In [11]: %sql select * from SPACEXTABLE where Launch_site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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

- Explanations:
  - 5 records displayed from dataset with string “CCA”

Total Payload Mass

```
In [15]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[15]: sum(PAYLOAD_MASS__KG_)
                          45596
```

- Explanations:
  - SUM of PAYLOAD_MASS_KG field from SpaceX data table. Mass carried by boosters.

Average Payload Mass by F9 v1.1

```
In [16]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
Out[16]: avg(PAYLOAD_MASS__KG_)
         2928.4
```

- Explanations:
  - Calculation of Average Payload Mass for specific booster version F9 v1.1

First Successful Ground Landing Date

```
In [17]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
Out[17]: min(Date)
         2015-12-22
```

- Explanations:
  - Date selection for when the first landing outcome for SpaceX for Ground Pad was successful

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [18]: %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.
Out[18]:
```

Out[18]:

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

- Explanations:
  - Date selection (Booster Versions) for successful Drone Ship landings where Payload Mass is between 4k and 6k

Total Number of Successful and Failure Mission Outcomes

```
In [24]: %sql select distinct Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
 * sqlite:///my_data1.db
Done.
```

Out[24]:

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

- Explanations:
  - Date selection for total of Success and Failure outcomes listed by “Mission Outcome” type

Boosters Carried Maximum Payload

In [32]:
```
%%sql
select Booster_Version
from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

Out[32]:

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

- Explanations:
  - Booster Type selection for those that have carried the maximum payload mass

2015 Launch Records

In [36]:
```
%%sql
select Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTABLE
where Landing_Outcome = 'Failure (drone ship)'
and substr(date,0,5) = '2015'
```

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

Out[36]:

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

- Explanations:
  - Listing records for failed landing outcomes for drone ship and booster versions

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

In [37]:
```
%%sql
select Landing_Outcome, 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.
```

Out[37]:

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

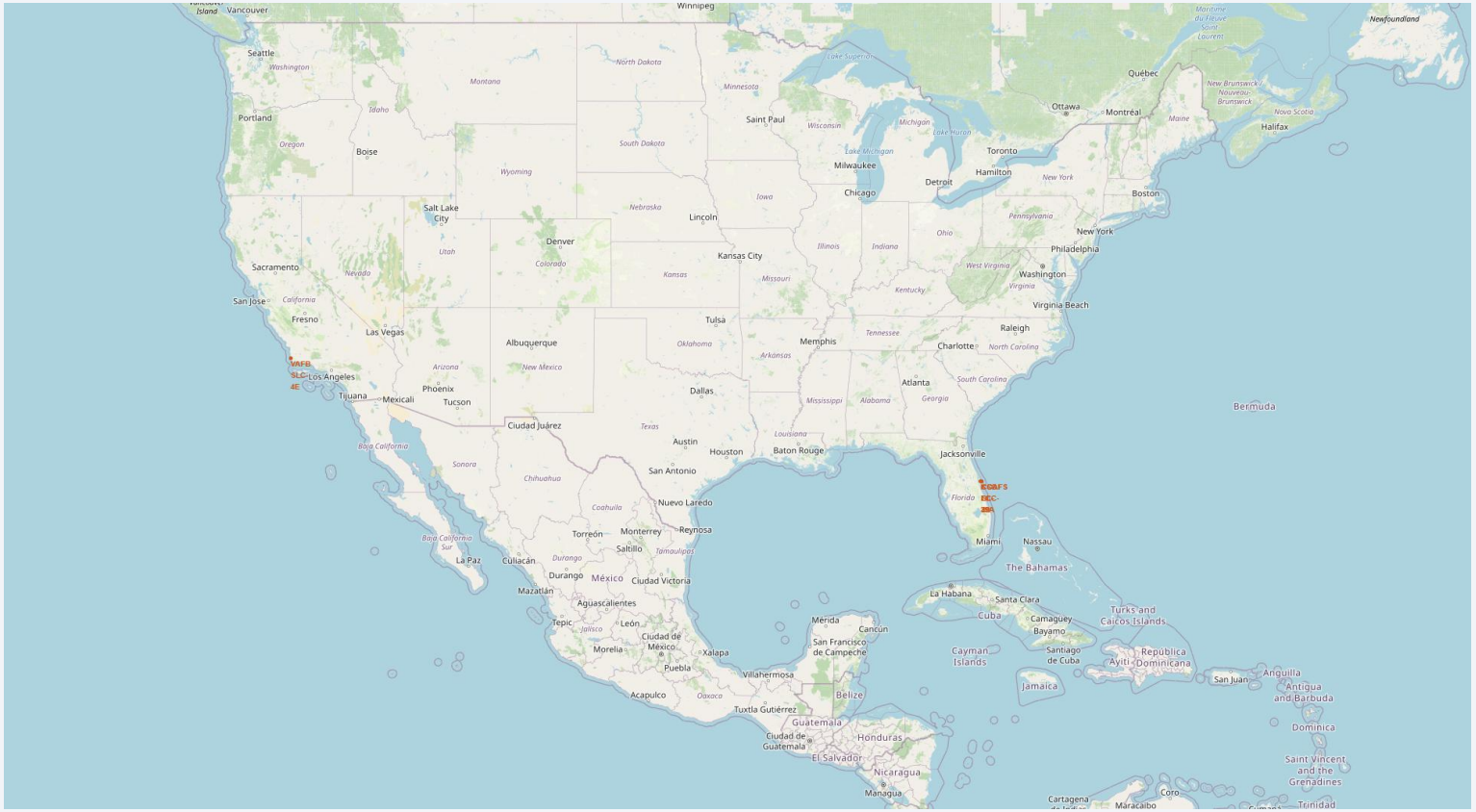
- Explanations:
  - Ranked outcomes for launches between 2010-06-04 and 2017-03-20 presented in descending order

Section 3

Launch Sites Proximities Analysis

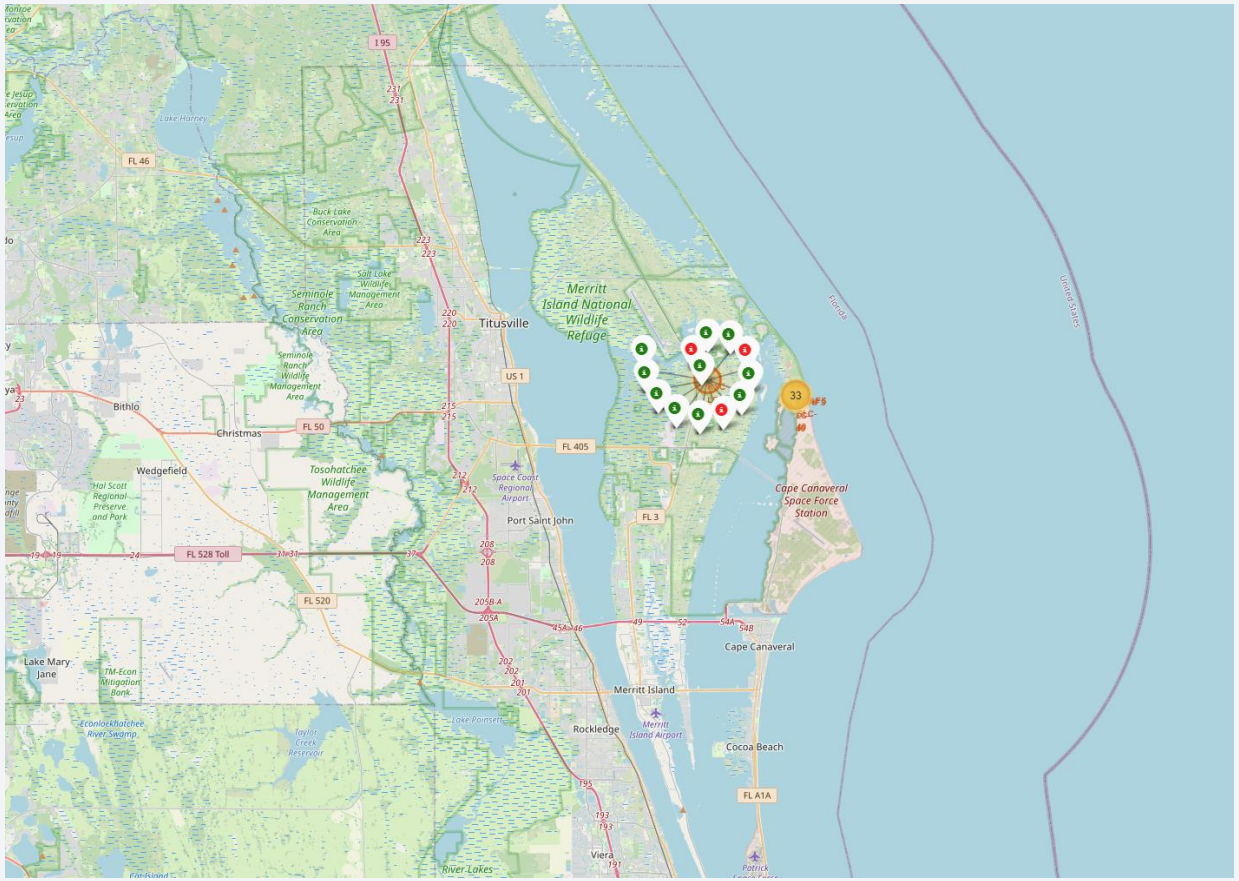
All Launch Site Locations - Global

- Findings:
  - All launch sites are located near coastal areas. This provides a safe launch environment away from population centers to minimize risk of failure.



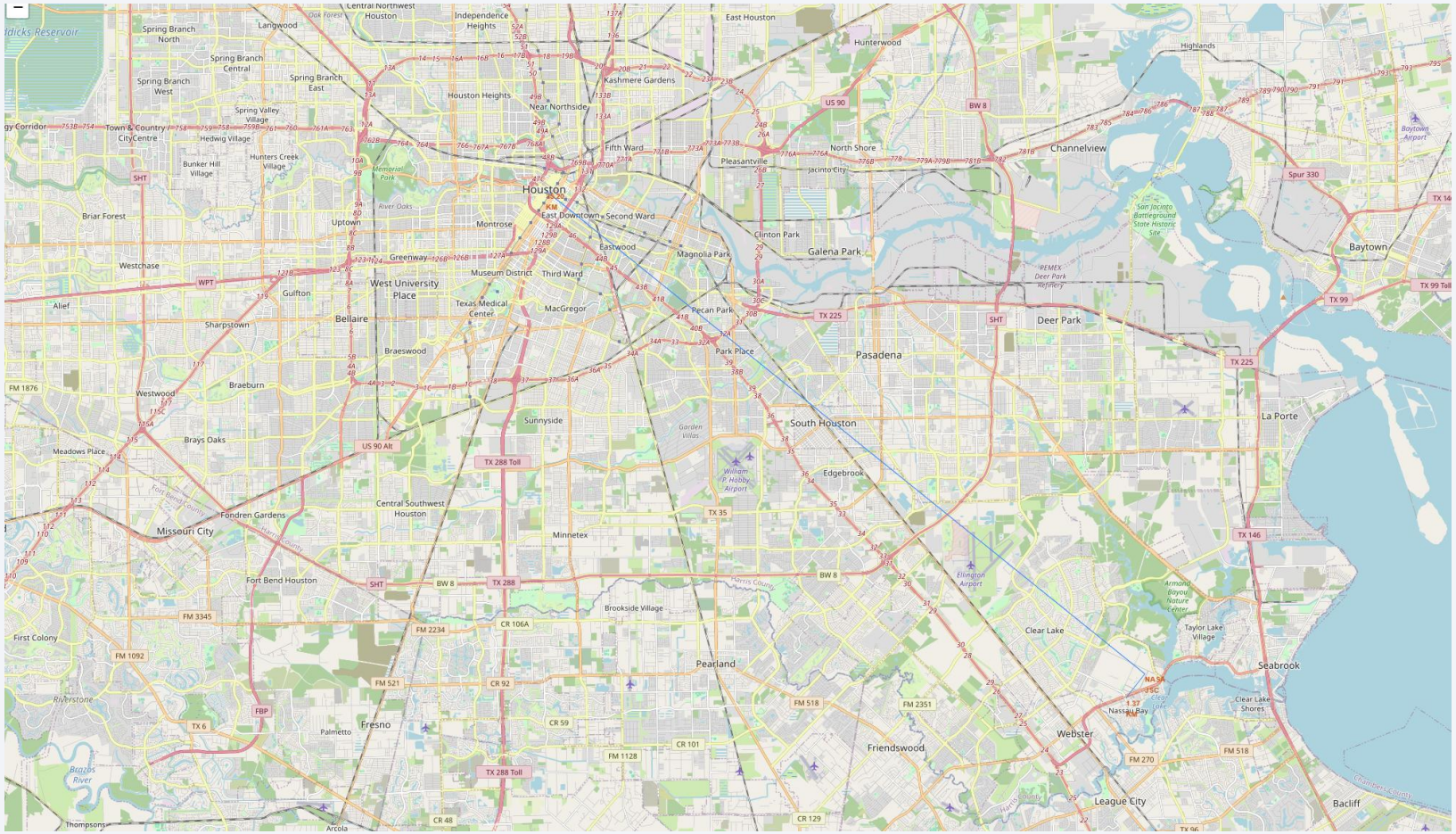
Launch Outcomes – Color Coded

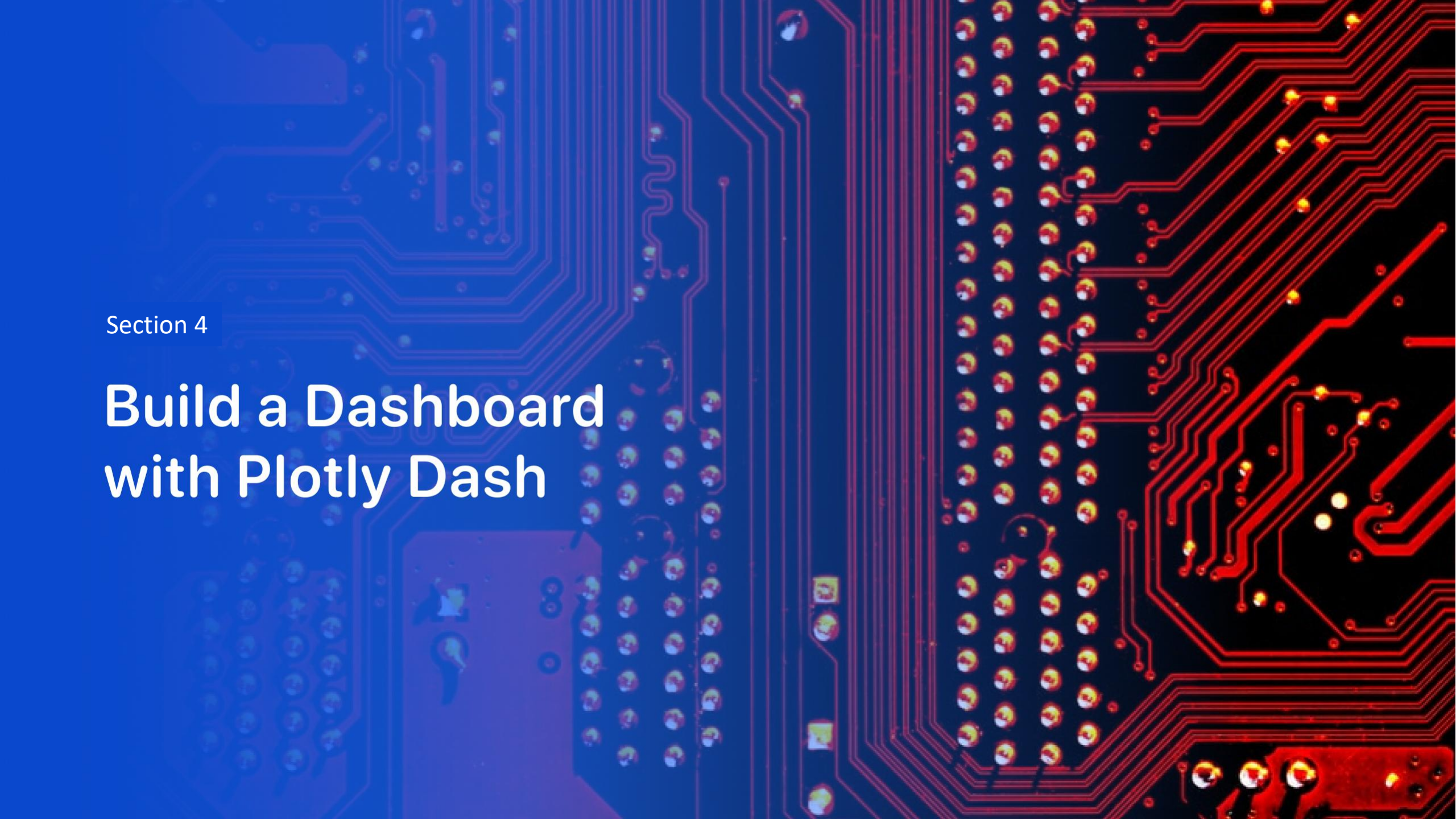
- Findings:
  - Drilling into site specific information, color coding allows the viewer a quick scan of the data outcomes
  - In this case, Red = Failed Launch, Green = Successful Launch.
  - This is for Site KSC LC-39A



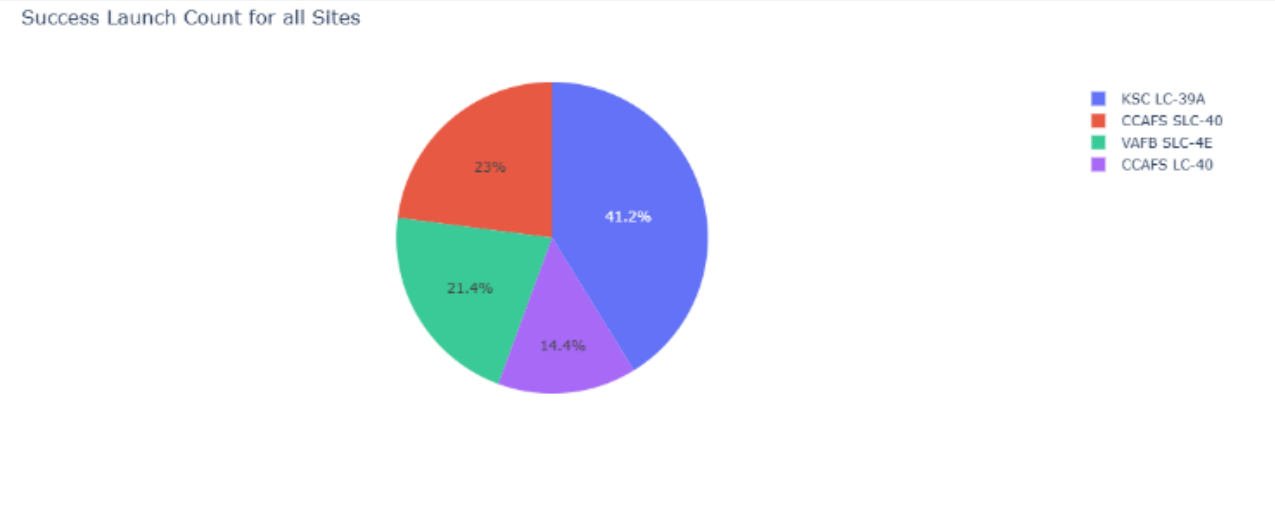
Launch Site Proximity to Landmarks

- Findings:
  - For this example, launch site proximity (NASA JSC) to nearest large city, Houston, is 25.2 km (blue line)
  - Proximity to population centers and areas of concern are necessary to calculate the risk involved for a failed launch



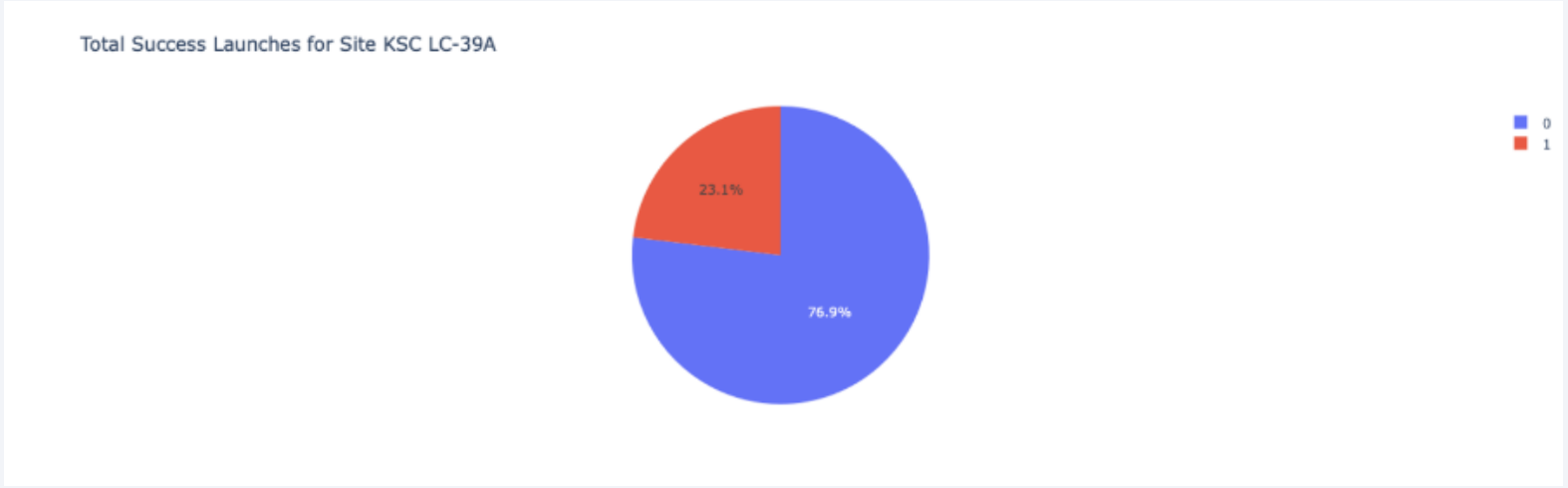

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites - Plotly



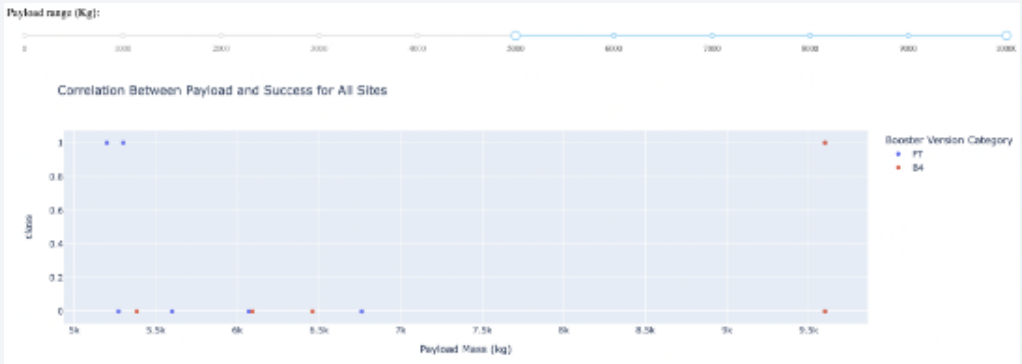
- From the pie chart, we can discern that the most successful launch site is clearly KSC LC-39A (41.2%) and the least successful is CCAFS LC-40 (14.4%)

Launch Site with Highest Launch Success Ratio



- KSC LC-39A has the highest success among the measured launch sites at 76.9%.

Payload Mass vs Launch Outcome for all Sites



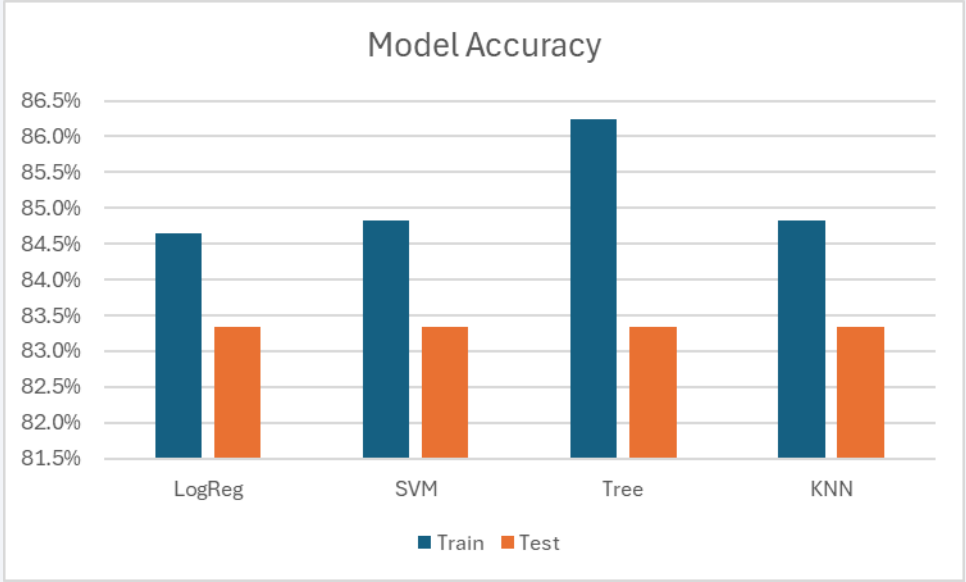
- The top chart shows all data for all sites. The Second view is for the middle of the range. Highest success from the data appears between 2kg and 5.5kg payloads.

Section 5

Predictive Analysis (Classification)

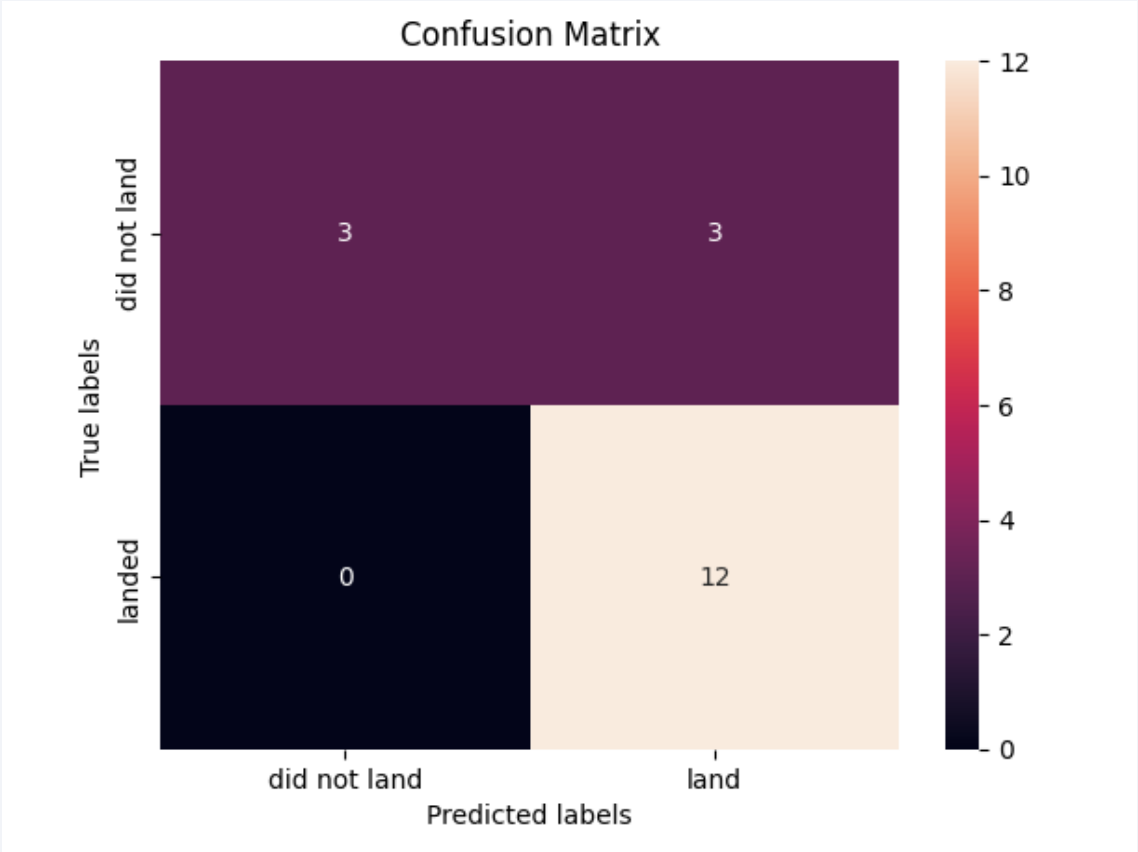
Classification Accuracy

- When viewing the Test models for accuracy, all 4 methods performed at the same level ~83.33%
- However, the Training model data performed better in all cases. The best being “Tree.”
- IF we were to combined our data sets, overall Tree would be the best indicator of accuracy



Confusion Matrix

- Displayed is the Tree Model Confusion Matrix
- We can see that false positives (upper right quadrant) with such a small dataset (18 values) create an accuracy issue



Conclusions

- As shown in the previous 2 slides, the Decision Tree method is the best machine learning algorithm to utilize for the SpaceX launch dataset.
- The highest number of successful launches occurred at site KSC LC-39A
- The success rate for launches has improved consistently over time
- Highest success rates (100%) for orbit types ES-L1, GEO, HEO, SSO
- Majority of payloads over 2k were successful

Thank you!