IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Timothy Wanless
7/4/2023

Executive Summary

- Data from SpaceX was pulled and exploratory analytics were done to show the differences in the types of rockets, orbits, and launches. These were then visualized and predictive analytics were performed to show correlations between relevant factors for rocket success.
- The results for this report show the best outcomes by rocket type, orbit, launch site and payload size. This can be helpful for determining the highest success of a SpaceX launch

Outline

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

Introduction

- Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- The problem we are trying to analyze in this project is if the Falcon 9 rocket will land successfully. This will be used to factor cost for bidders of these rockets. The Falcon 9 rocket costs 62 million dollars. If competitors can figure out its success rate they can then bid competitively against Space X.
- To find this out we will analyze a set of given variables of each of the launches from 2010 to 2020 taking in variables such as: orbit, launch site, payload mass, etc.

Section 1

Methodology

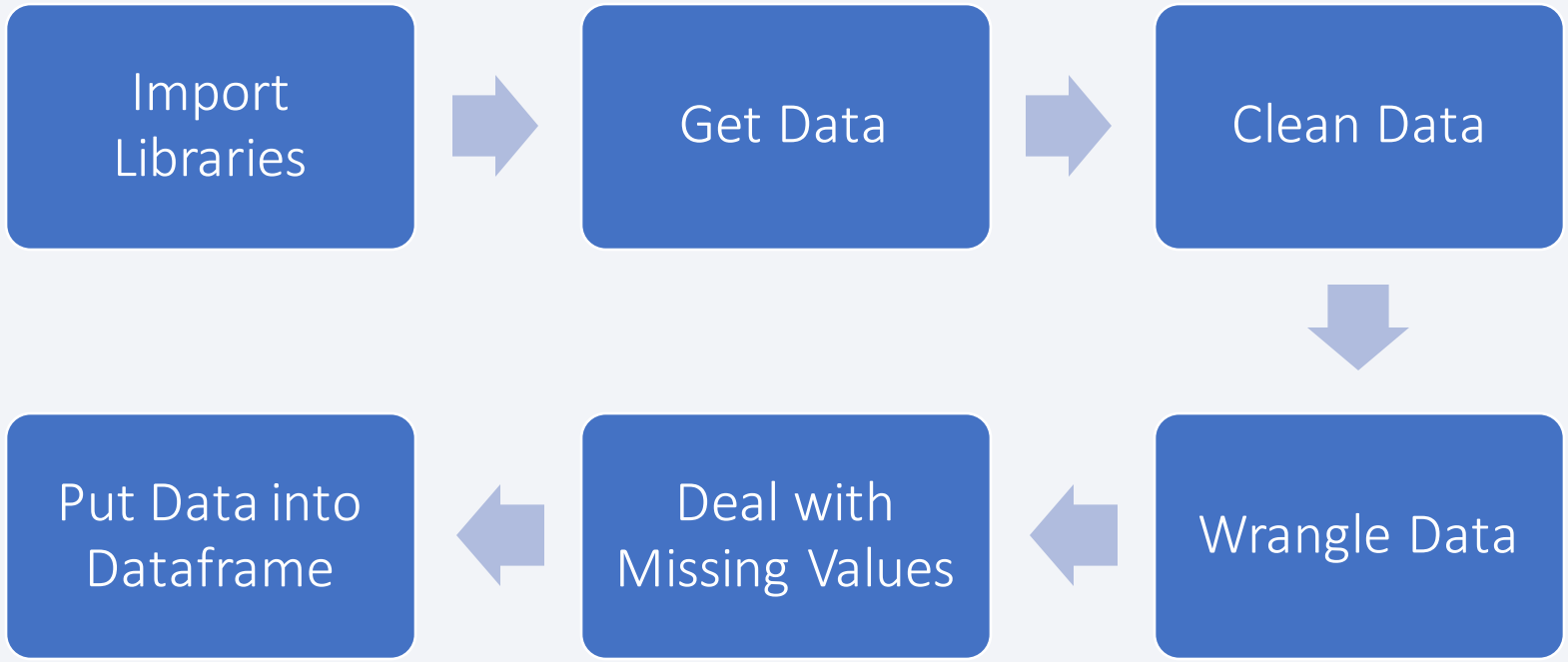
Methodology

Executive Summary

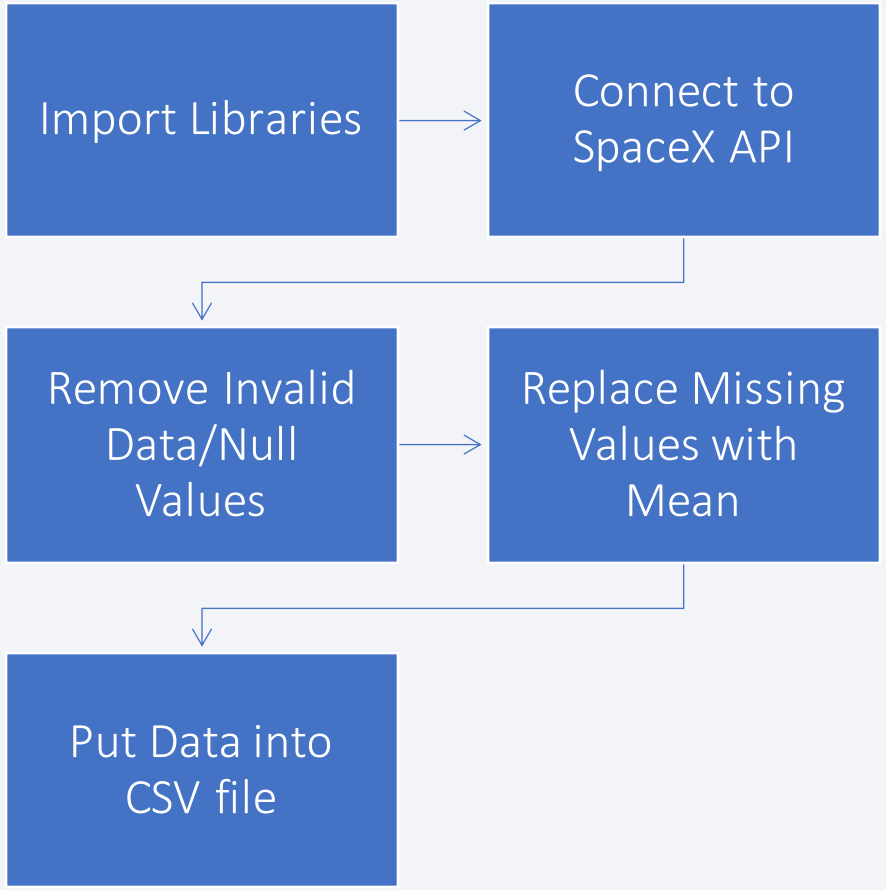
- Data collection methodology:
  - Date was collected from the SpaceX API
  - Data was collected by Webscraping the Wikipedia of SpaceX launch data.
- Perform data wrangling
  - The date was cleaned, formatted, and put into an organized dataframe
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Regression, SVM, and Decision Tree models were all used

Data Collection

- The launch data information was collected by connecting to SpaceX's API and webscraping. The data was restricted to dates before 11/13/2020. Irrelevant variables and null values were removed from the dataset.

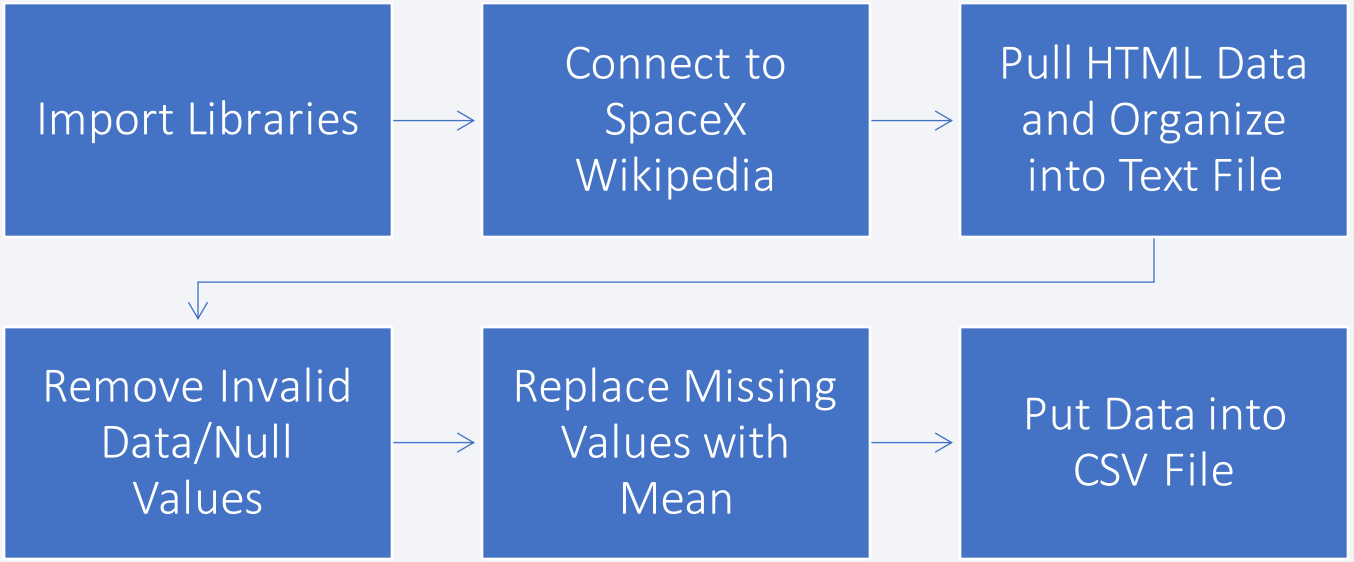


Data Collection – SpaceX API



https://github.com/TimWanless/DataScienceCapstoneProject/blob/main/1.Data_Collection_API.ipynb

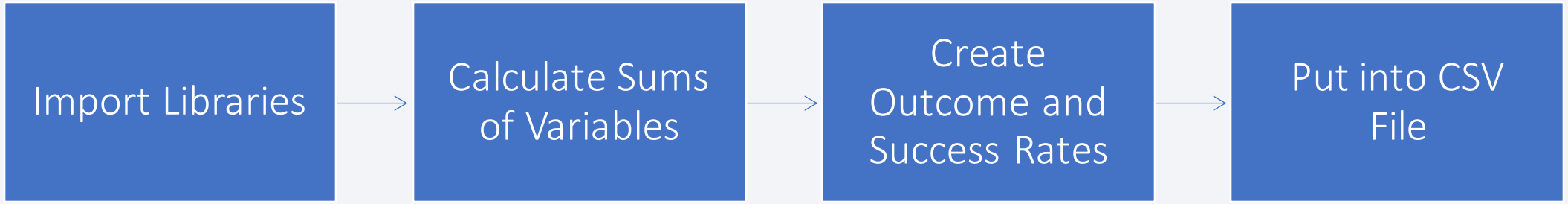
Data Collection - Scraping



https://github.com/TimWanless/DataScienceCapstoneProject/blob/main/2.Data_Collection_Webscraping.ipynb

Data Wrangling

- Data Processing
  1. Import Libraries
  2. Number of Launch Sites, Orbits, and Mission Outcomes
  3. Created Mission Outcome Label and Success Rates
  4. Put into a CSV Data



https://github.com/TimWanless/DataScienceCapstoneProject/blob/main/3.DataWrangling.ipynb

EDA with Data Visualization

- Scatterplot of Payload Mass by Flight Number
- Scatterplot of Launch Site by Flight Number
- Scatterplot of Launch Site by Payload Mass
- Barplot of Success Rate by Orbit
- Scatterplot of Orbit by Flight Number
- Scatterplot of Orbit by Payload Mass
- Line plot of the Launch Success rate over time

https://github.com/TimWanless/DataScienceCapstoneProject/blob/main/5.EDA_SQL_Queries.ipynb

EDA with SQL

- SQL Queries
  - Displayed the names of the 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 ground pad was achieved.
  - Listed the date when the first successful landing outcome in ground pad was achieved.
  - Listed the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Ranked the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

https://github.com/TimWanless/DataScienceCapstoneProject/blob/main/4.EDA_SQL_Queries.ipynb

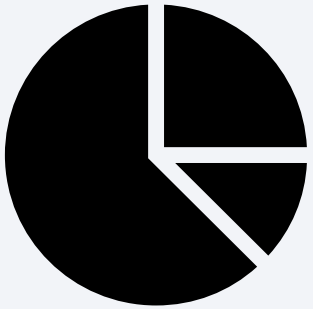
Build an Interactive Map with Folium

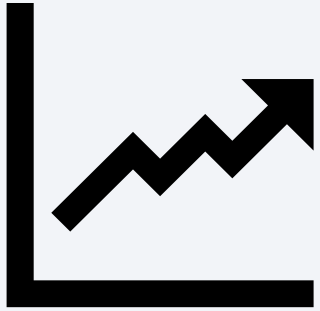
- The Interactive Folium map was made with Success/Failure markers that show geographical location. It also shows the location relative to highways, cities, coastlines and railways in order to show potential correlations between success and failure.

https://github.com/TimWanless/DataScienceCapstoneProject/blob/main/6.Interactive_Map.ipynb

Build a Dashboard with Plotly Dash

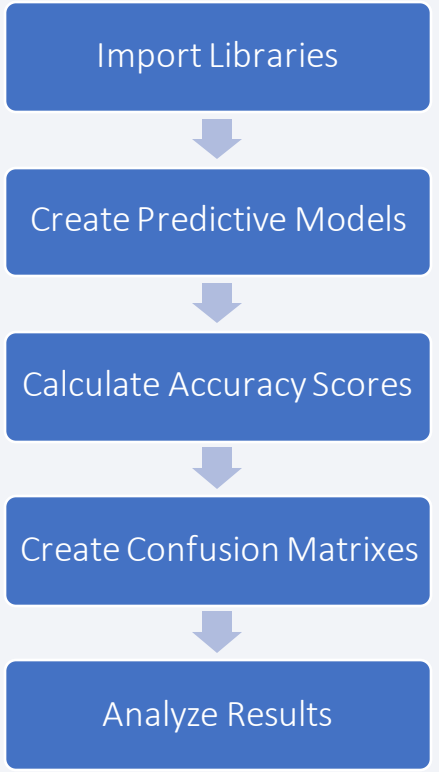
- A dropdown was added for the different type of launch locations
- Pie chart was added for the success rate
- Slider added to Payload(kg) in 100 kg segments
- Scatter plot added to show rocket type success and by payload
- Access Dashboard: **https://timwanless-8051.theiadocker-0-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai/**





Predictive Analysis (Classification)

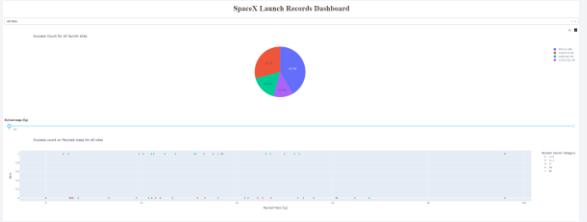
- Classification Model Selection
  1. K Nearest Neighbors Model
  2. Decision Tree Model
  3. Logistic Regression Model
  4. SVM Model
  5. Confusion Matrixes for Each Model
  6. Best Score Calculation/ Analysis

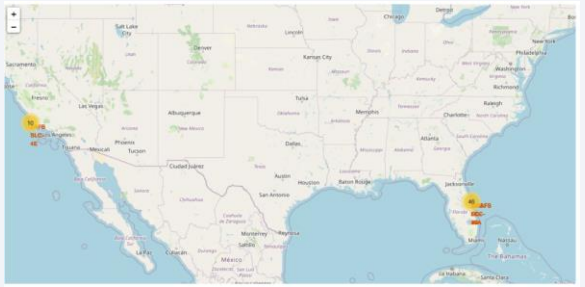


https://github.com/TimWanless/DataScienceCapstoneProject/blob/main/8.SpaceX_Predictive_Analytics.ipynb

Results

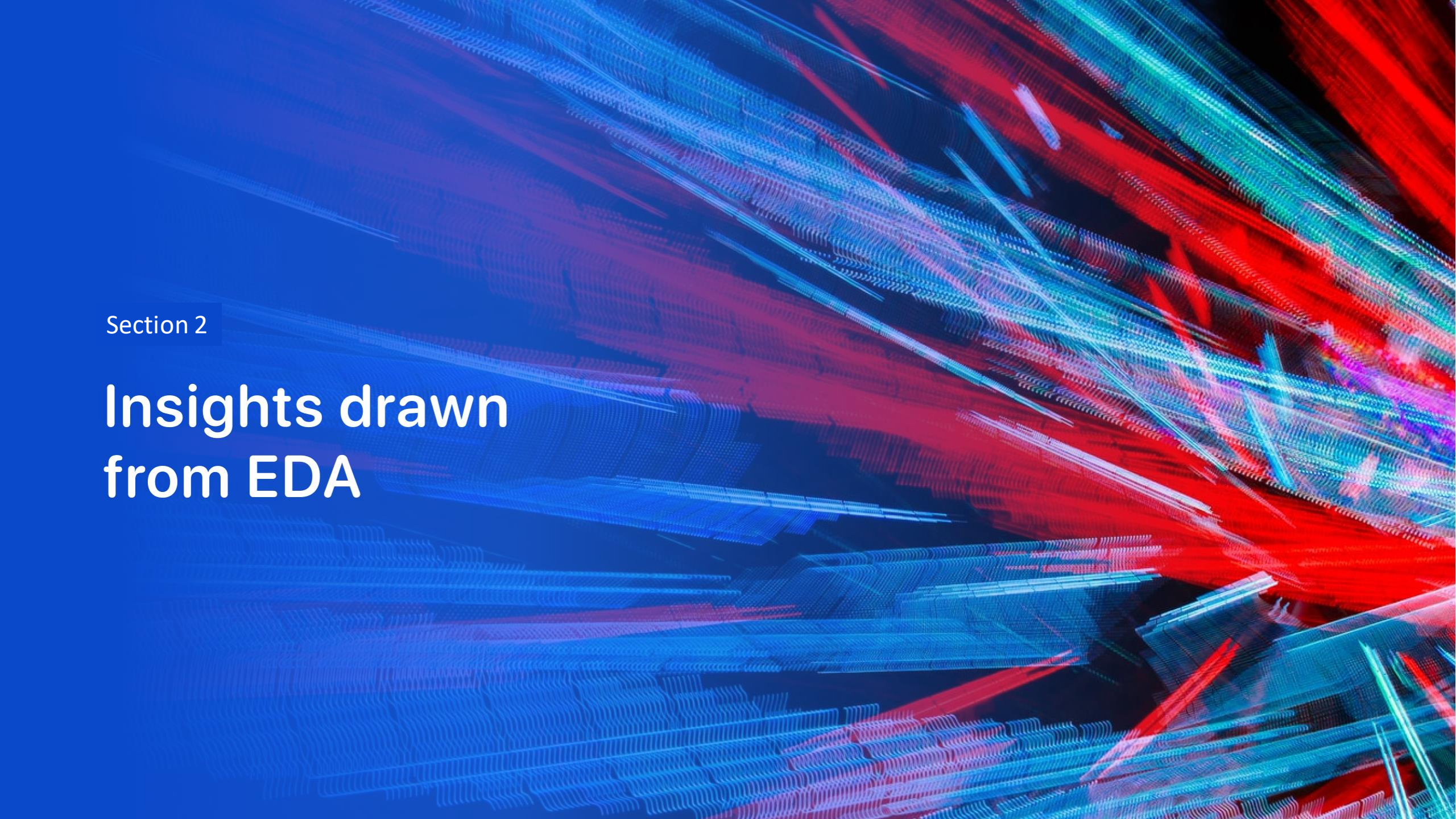
- Exploratory Data Analysis
  - The Success rate is upward trending since 2010 to 2020
  - The GTO Orbit had the most failures
- Interactive Analytics
  - The FT Booster is the most successful
  - The best launch site is KSC LC-39A
- Predictive analysis
  - The Decision Tree is the most accurate predictive model of mission success for SpaceX flights



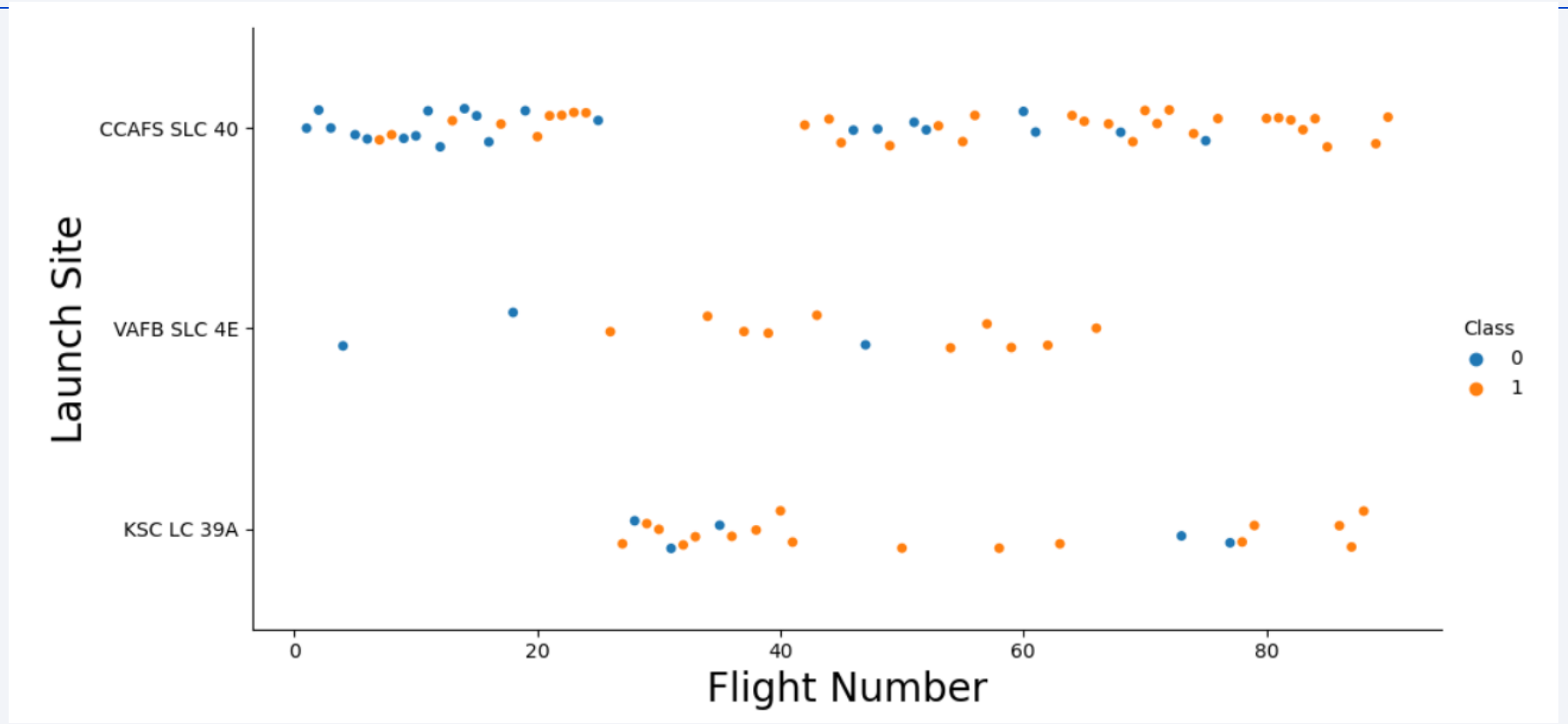


Section 2

Insights drawn from EDA

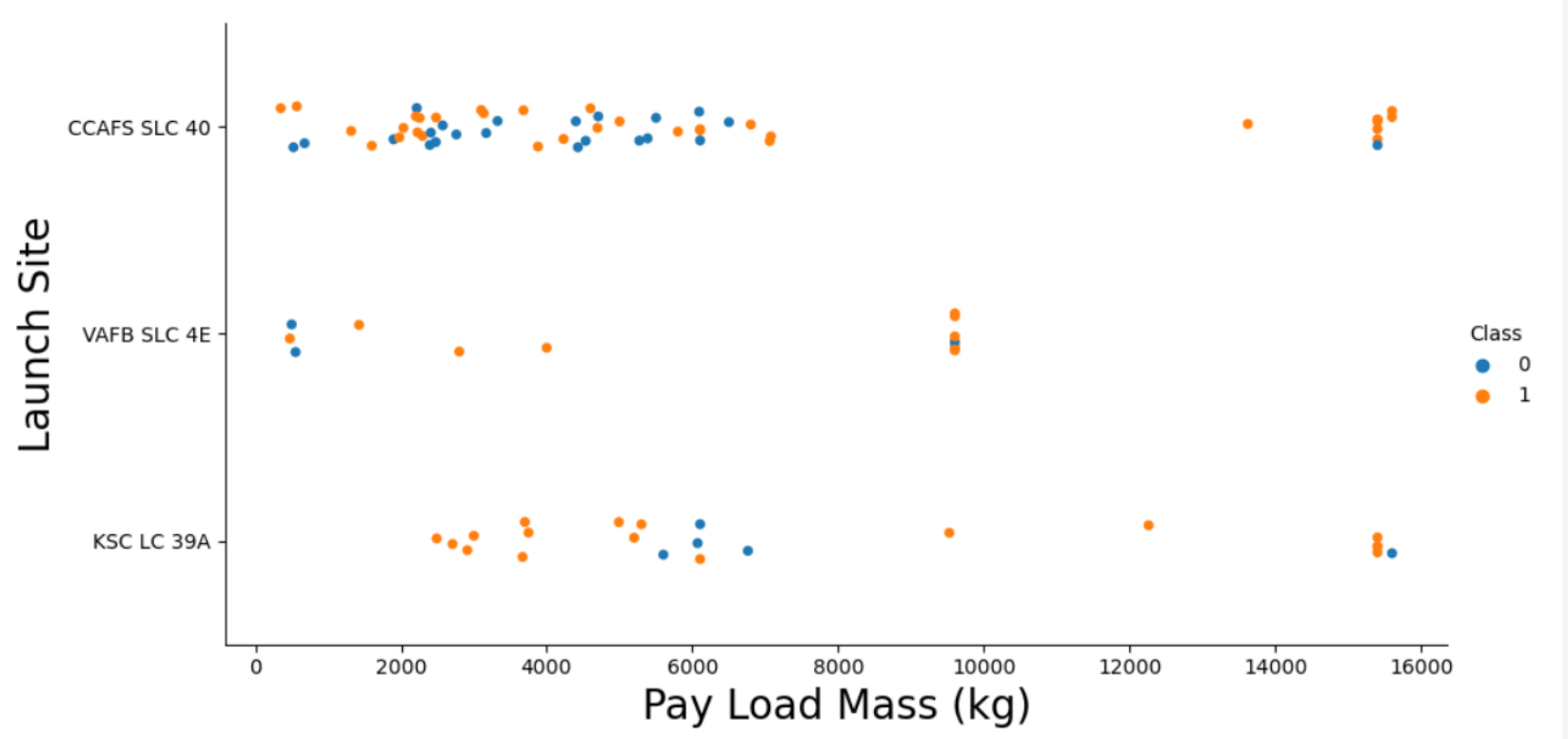


Flight Number vs. Launch Site



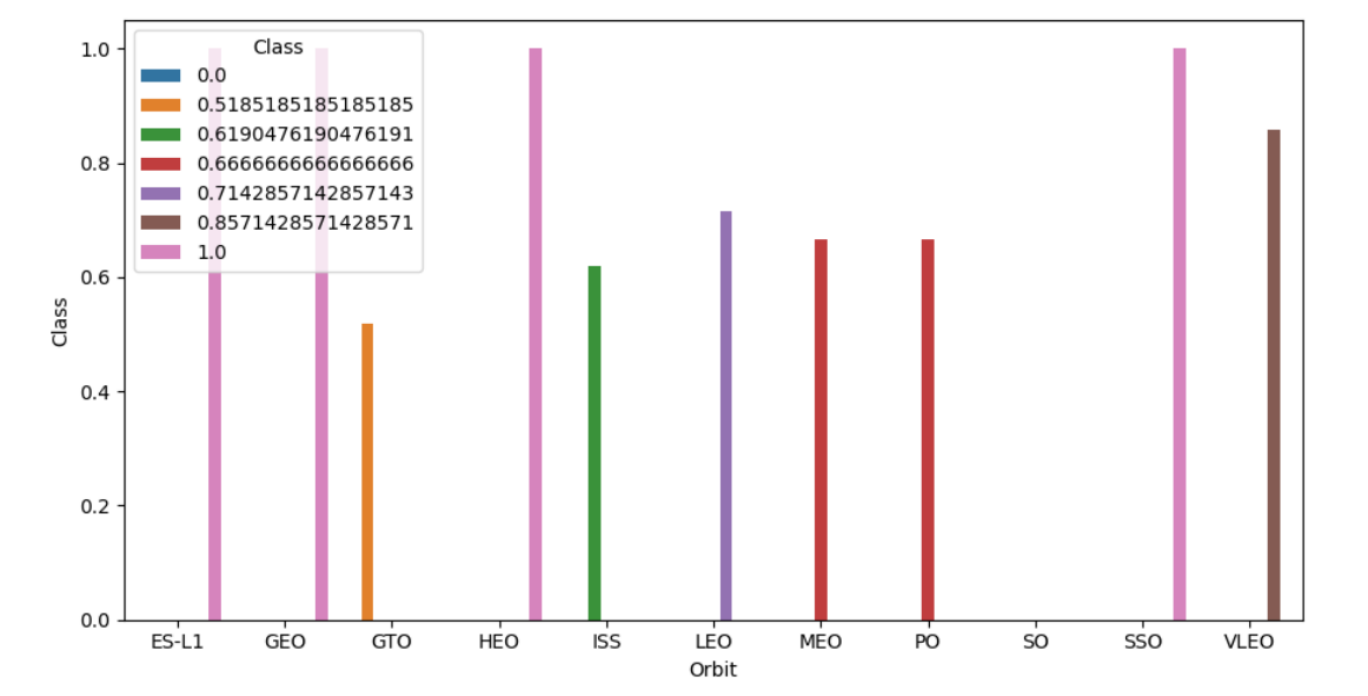
We see that there appear to be a larger number of fails in earlier flight number for CCAFS SLC 40 than other launch sights

Payload vs. Launch Site



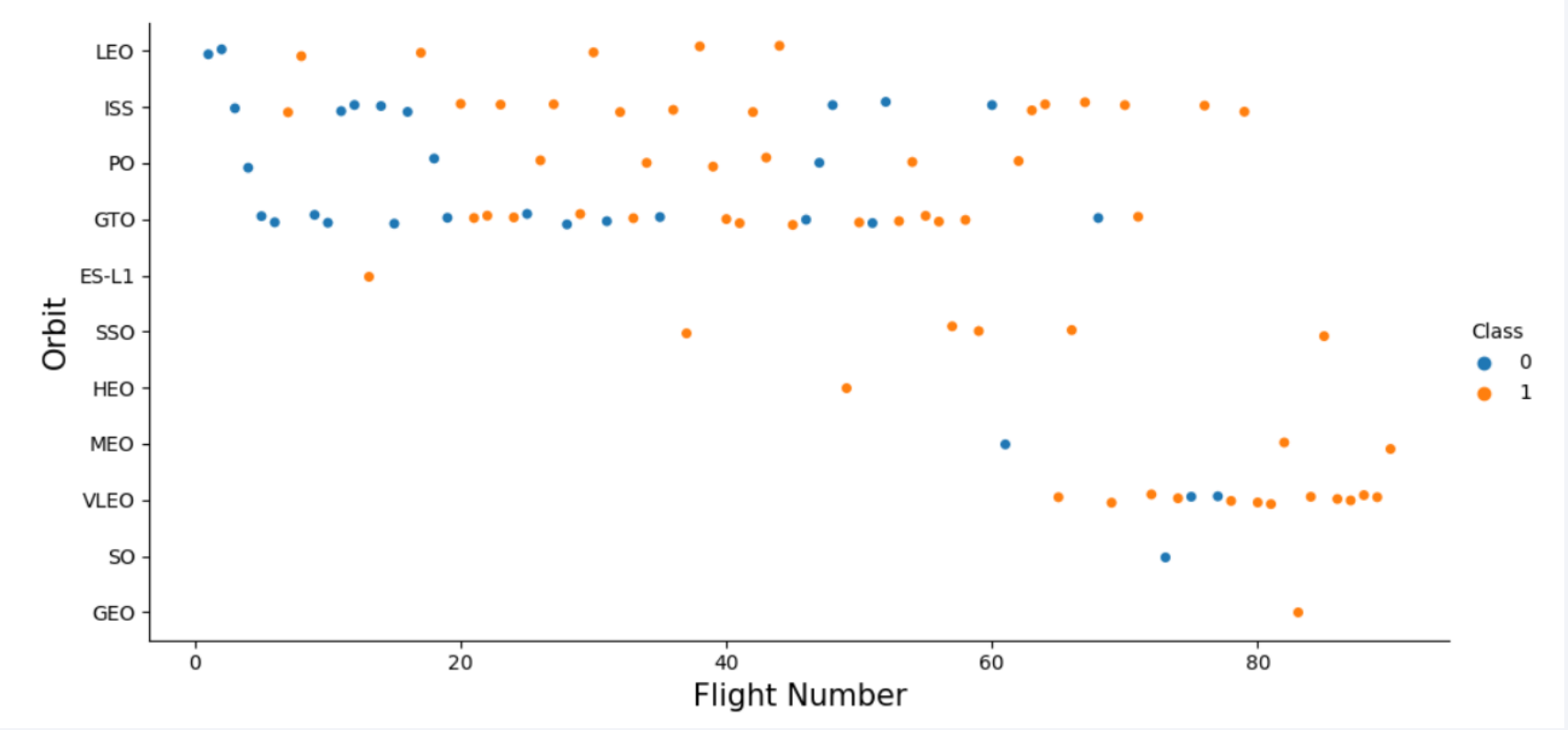
There are a number of smaller payloads that failed at every launch site, especially CCAFS SLC 40.

Success Rate vs. Orbit Type



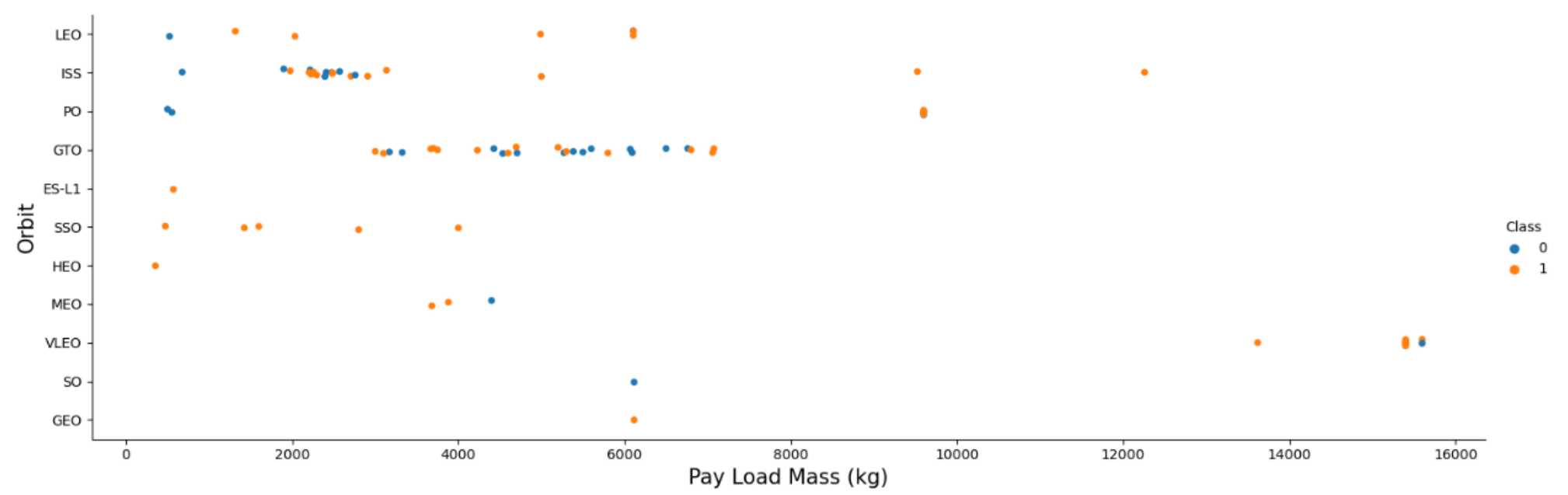
We see that ES-L1, GEO, HEO, and SSO orbits have a success rate of 100% and the worst success rate is for SO and GTO

Flight Number vs. Orbit Type



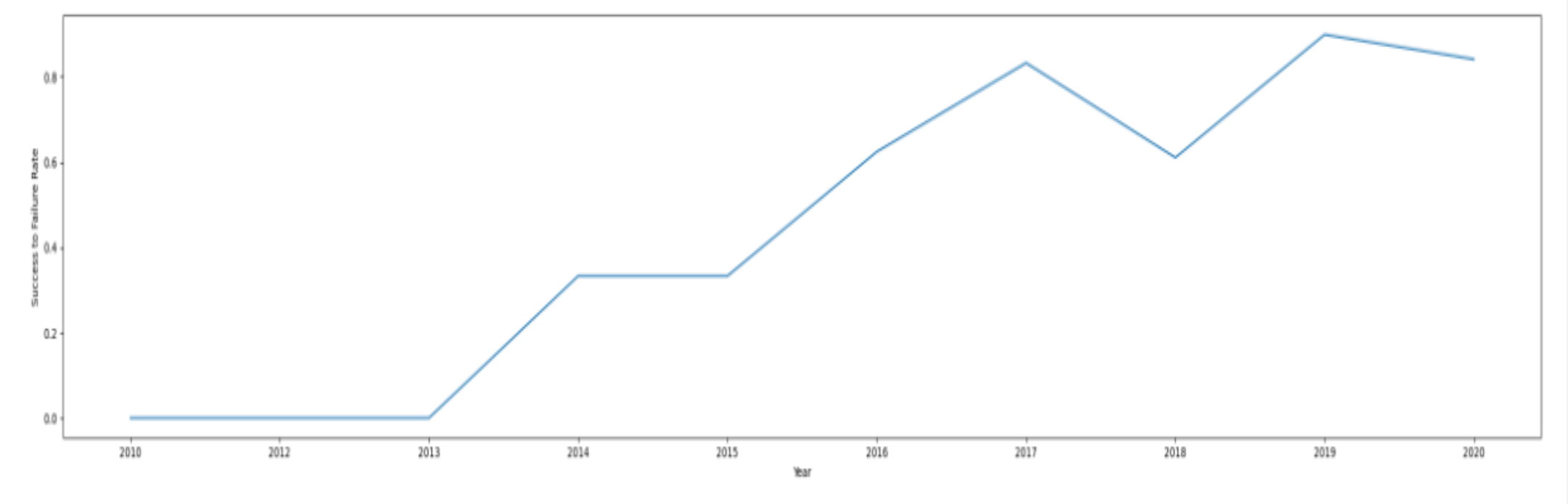
Earlier flight numbers failed at a higher rate with the most happening in the GTO orbit.

Payload vs. Orbit Type



There is a cluster of failures in the GTO orbit with payloads in the range or 4000 kgs to 6000 kgs, and there is another in the ISS orbit with payloads aroun 2000 kgs to 3000 kgs.

Launch Success Yearly Trend



There is an upward trending success rate for the launches from 2010 to 2020 with a slight dip in 2018.

All Launch Site Names

- Querying the data for all Launch site names came with 4 sites
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

The following query shows 5 instances of Launch Site names that begin with CCA with the following variables.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No atte
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No atte
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

This query sums the total payload of kg for NASA using the SUM function in SQL

SUM(PAYLOAD_MASS__KG_)
107010.0

Average Payload Mass by F9 v1.1

This query finds the average of payload mass for the F9 v1.1 booster using the AVG function in SQL

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

This query found the first successful ground landing date using the MIN function where the landing_outcome was successful

MIN(Date)
01/08/2018

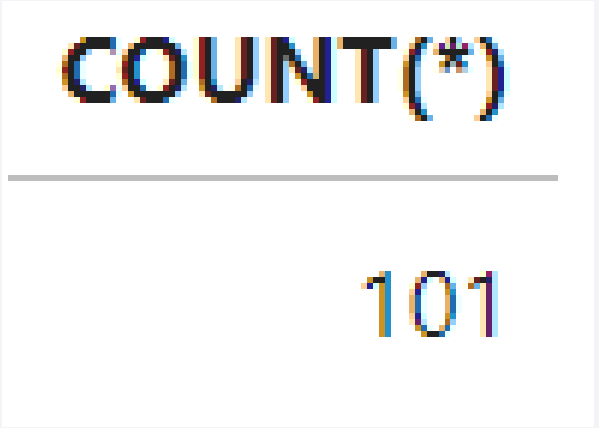
Successful Drone Ship Landing with Payload between 4000 and 6000

This query lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 using the DISTINCT and WHERE functions using Payload Mass

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

Total Number of Successful and Failure Mission Outcomes

- The query used the COUNT function where Mission_Outcome is LIKE %Failure% to calculate the total number of successful and failure mission outcomes
- There is a total number of 101 outcomes



Boosters Carried Maximum Payload

This query finds a list of the names of the booster which have carried the maximum payload mass. It uses a subquery using the MAX function of payload to find the boosters

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

This query gives a list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 using a SQL subquery using the date 2015 and finding the following data.

Month	Landing_Outcome	Booster_Version	Launch_Site
None	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
None	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

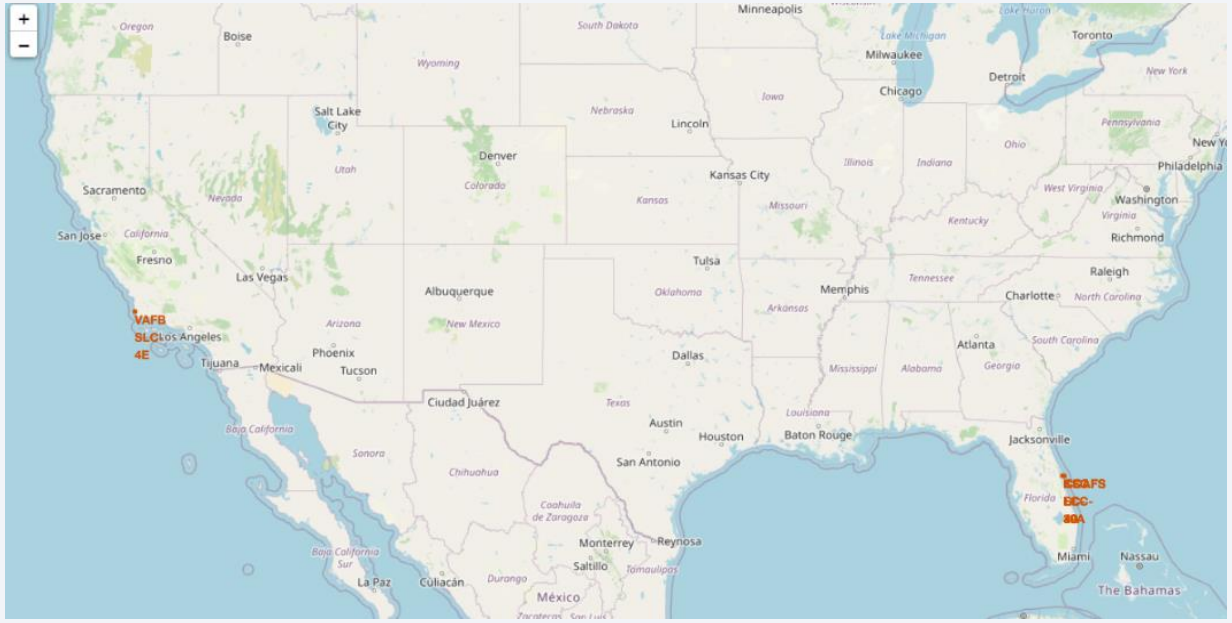
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

This query finds 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 using the COUNT and BETWEEN for the above dates.

Landing_Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	7

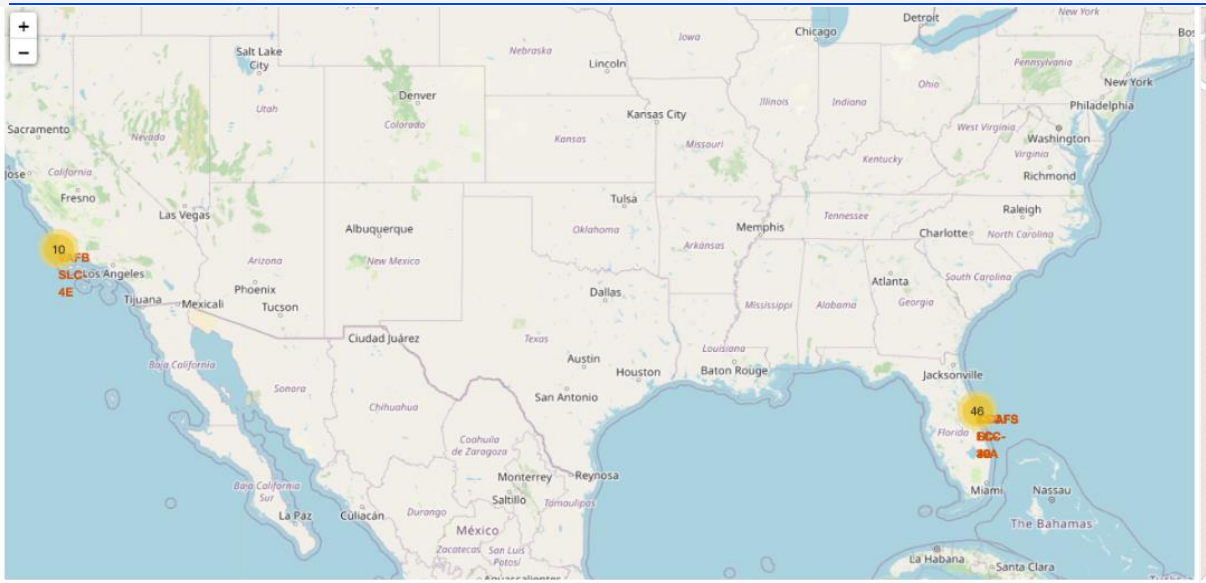

Section 3
Launch Sites
Proximities Analysis

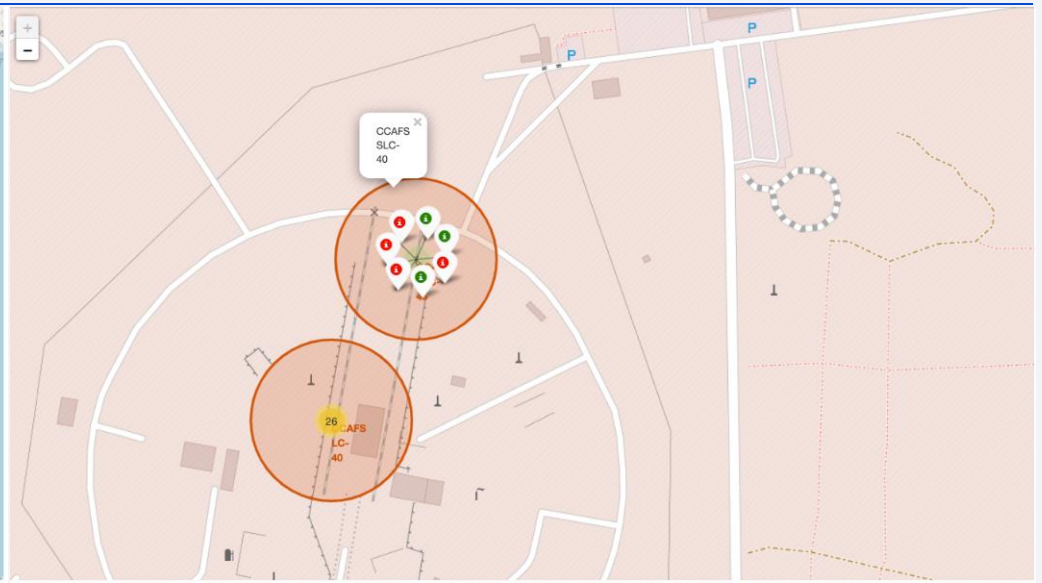
Launch Locations on Map



The launch locations are listed geographically using the identification codes.
The launch locations are above in California and Florida.

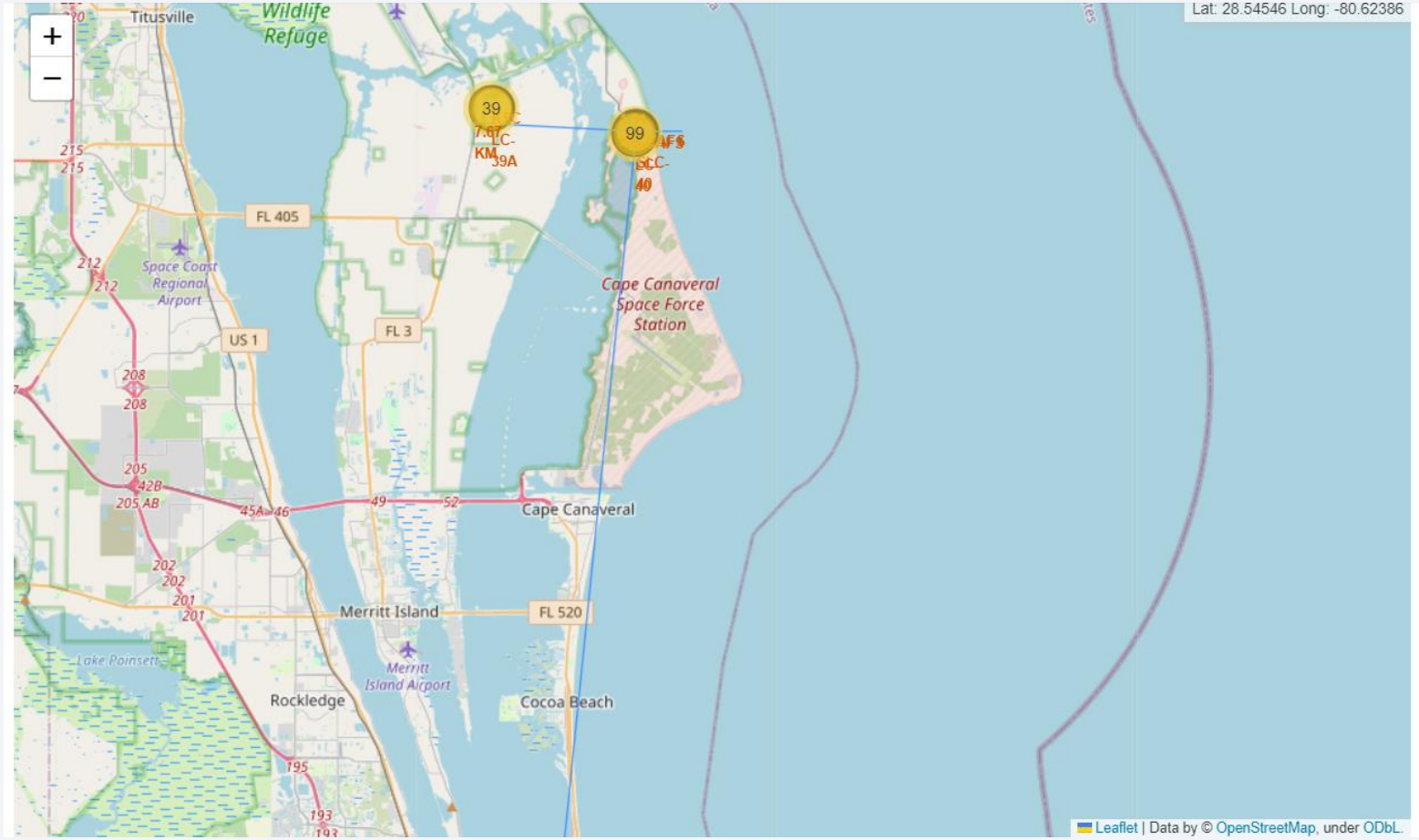
Geographical Markers of Success/Failure Launch Outcomes



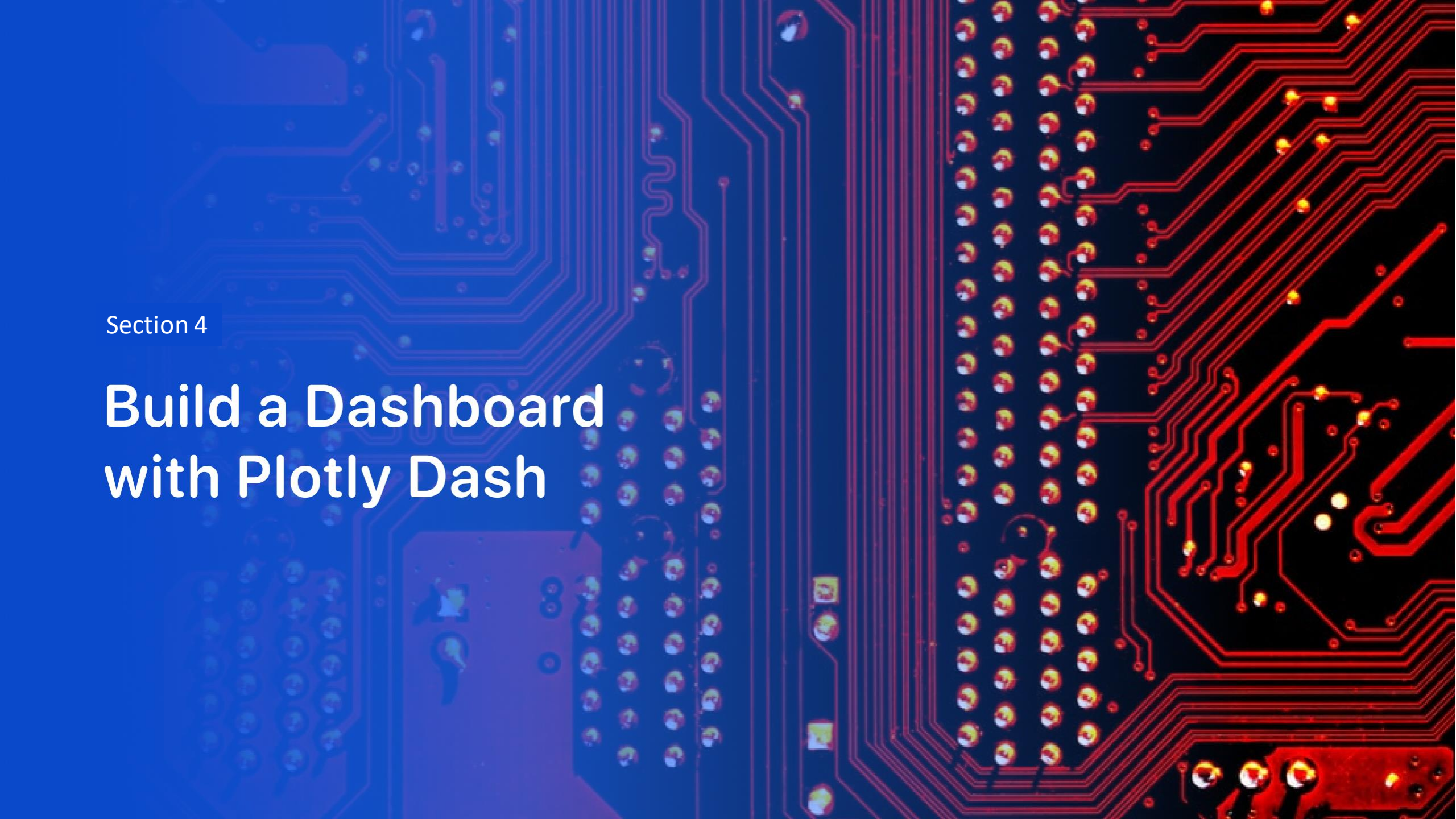


- The map has color coordinated markers to show the Failed(red), Successful(green), and Other(yellow) outcomes of the mission by geographical location

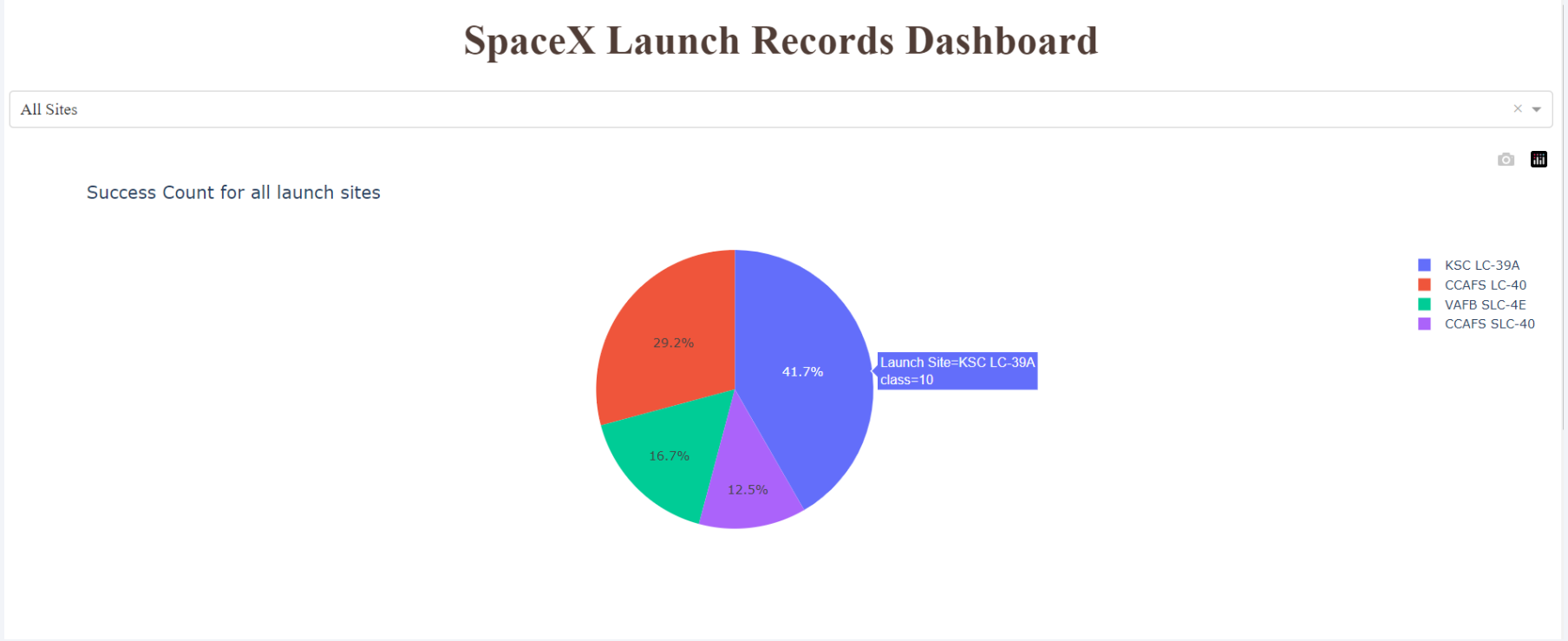
Distance of Launch sites to Geographical Markers



The successful launch sites are near the railway and coastline and fairly far from the nearest highway and city.

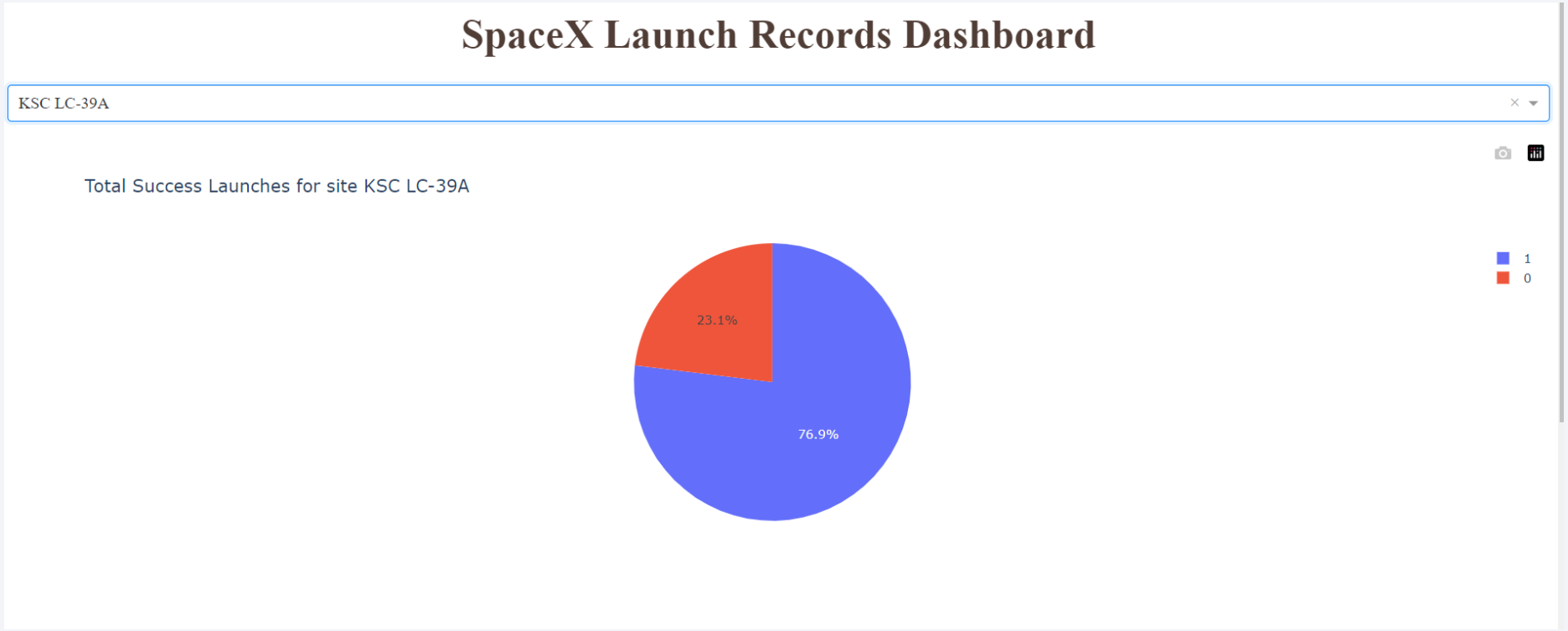

Section 4
Build a Dashboard with Plotly Dash

Launch Success by Location



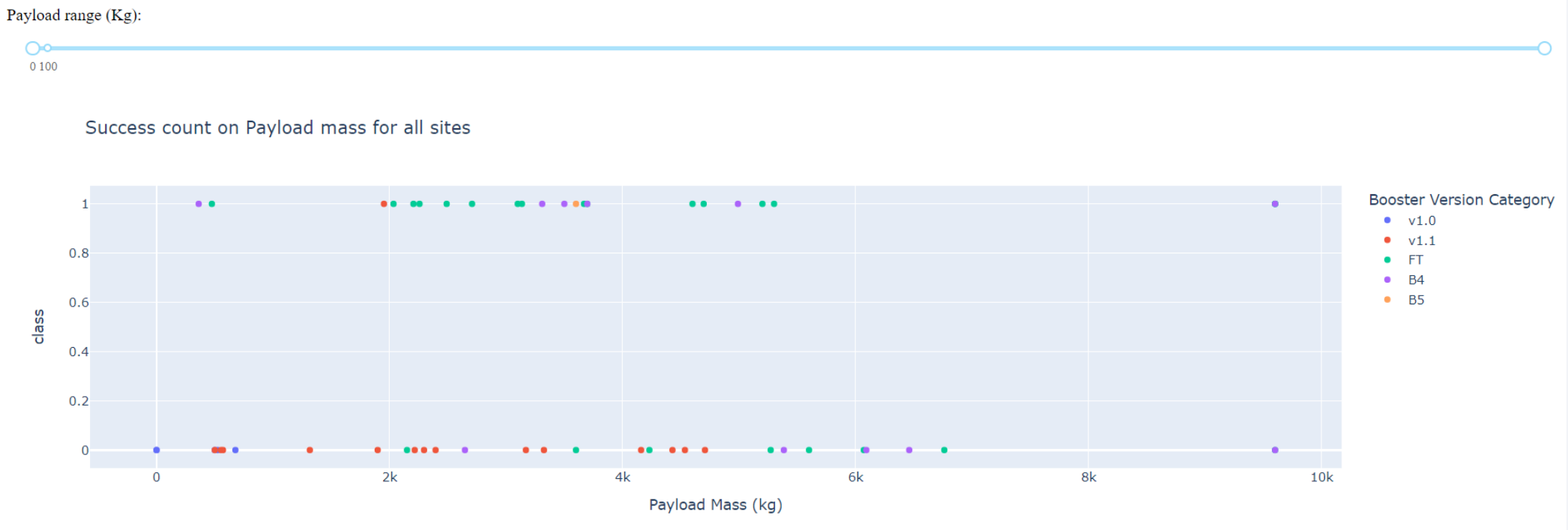
- Launch site KSC LC-39A accounts for the most successful launches
- Launch site CCAFS SLC-40 accounts for the least amount of launches

Best Performing Launch Site



- KSC LC-39A is the most successful launch site with a success rate of 76.9%

Payload vs Launch Success



- The FT booster version is most successful
- The v1.1 booster has the least successful

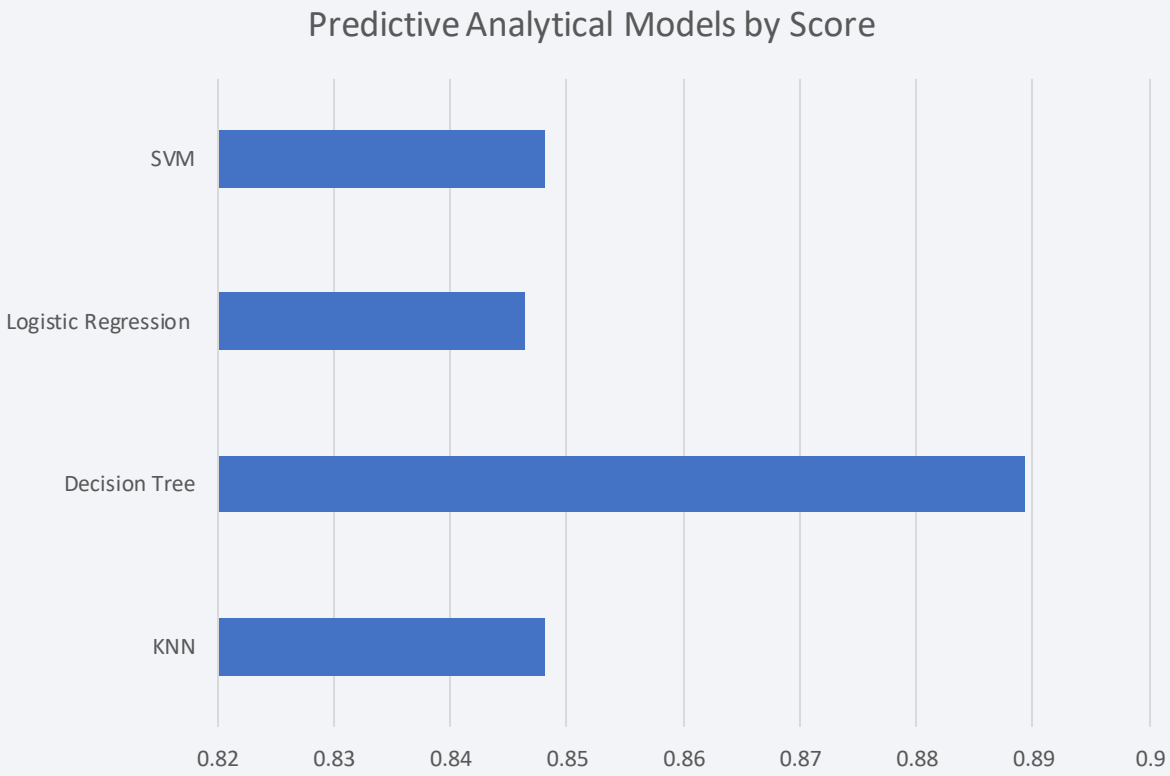
Section 5

Predictive Analysis (Classification)

Classification Accuracy

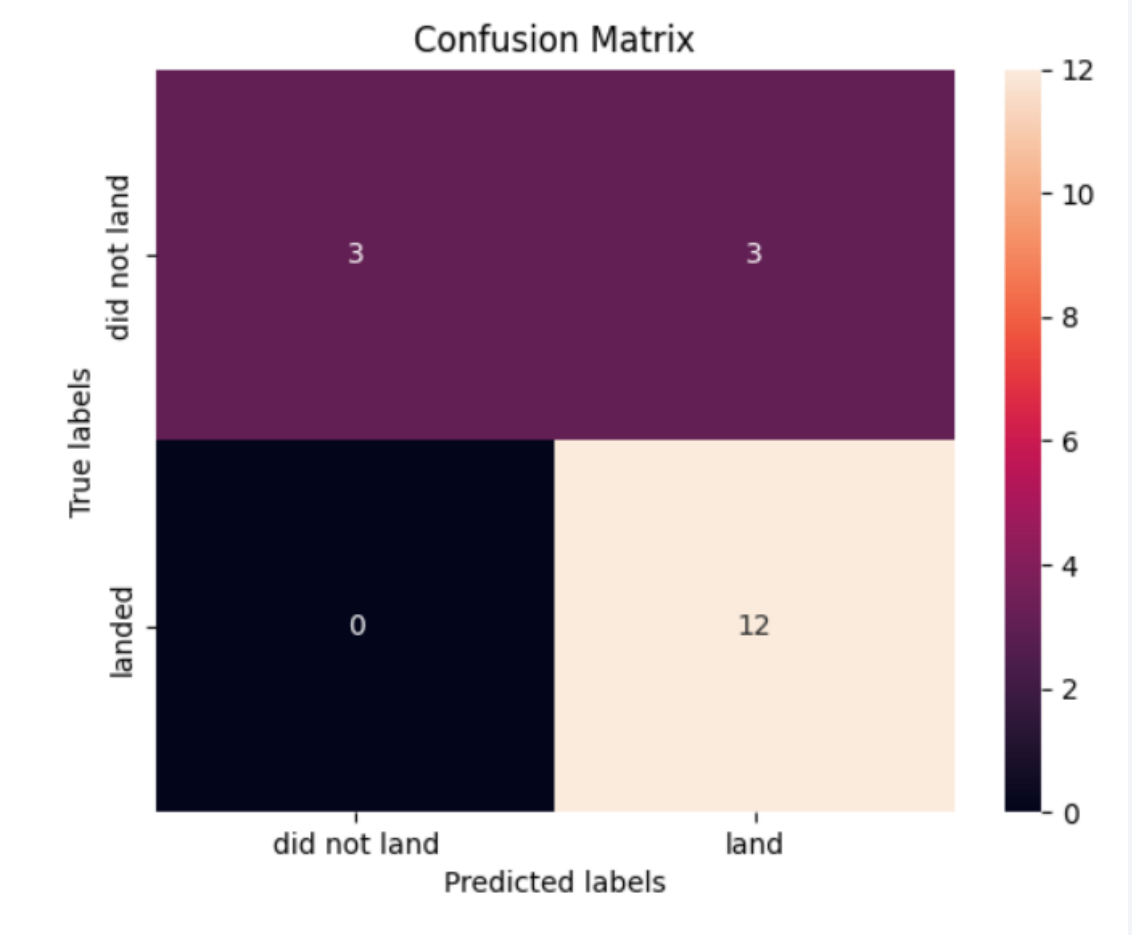
Several Predictive Classification models were used in determining future classification of mission success.

The Decision Tree Model has the highest score with 0.89 compared to the other models that had scores around 0.85



Confusion Matrix

The Decision Tree model performed best with correctly predicting all instances where the rocket landed and the results of the rocket not landing were split.



Conclusions

- The success rate of launches is upward trending
- The GTO orbit has the largest amount of failed launches
- The FT Rocket has the highest success rate
- The launch site has the highest success rate
- A decision tree model can be used to predict the success rate of a rocket launch with the best accuracy

Appendix

All code can be found at https://github.com/TimWanless/DataScienceCapstoneProject

Thank you!