IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

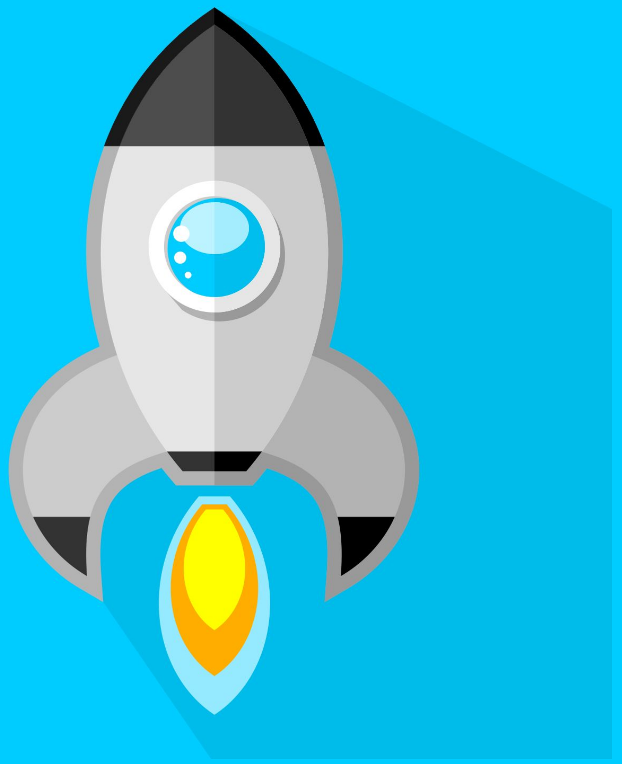
Jordan Piper
January 30, 2025

Outline

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

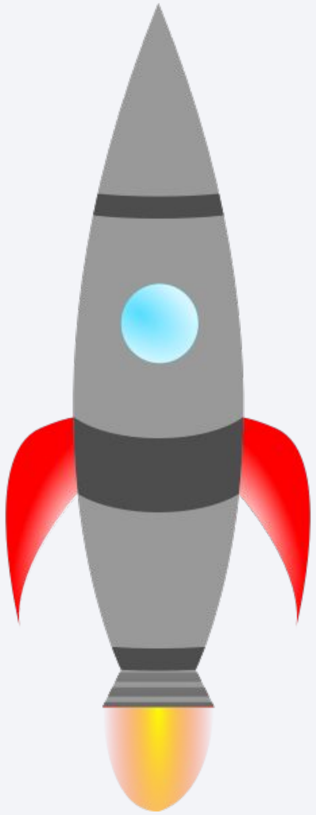
Executive Summary

- Summary of methodologies
  - 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
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics Demo in Screenshots
  - Predictive Analysis Results



Introduction

- Project background and context
  - SpaceX has revolutionized commercial space travel by pushing boundaries and significantly reducing costs. Their rockets are less than half the price of those from other space travel companies, thanks largely to their ability to reuse the first stage. Predicting whether the first stage will successfully land can help estimate the overall cost of a launch. Space Y, a rising player in the space travel industry, aims to compete with SpaceX by adopting similar innovations.
- Problems you want to find answers
  - How much will each launch cost?
  - How do payload mass, launch sit, number of flights, and orbits impact the success of the first stage landing?
  - Do successful landings increase over the years?



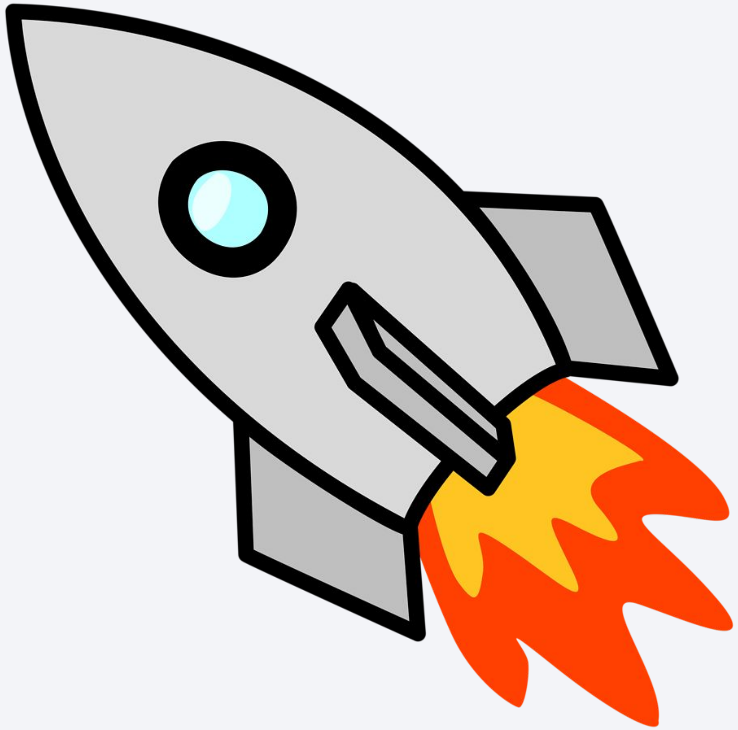
Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Utilized SpaceX Rest API
  - Utilized Web Scraping from Wikipedia
- Perform data wrangling
  - Data was filtered
  - Isolated and attended to missing values
  - Used One Hot Encoding to prepare the data to a 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 results



Data Collection

- Data was collected by utilizing the SpaceX Rest API and web scraping historical launch records from a Wikipedia page titled “List of Falcon 9 and Falcon Heavy Launches.”
- Utilized get request to receive JSON content from SpaceX Rest API to receive: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Utilized Beautiful Soup Python package to scrape launch records stored in an HTML table to receive: Flight No, Launch Site, Payload, Payload mass, Orbit, Customer, Launch Outcome, Version Booster, Date, Time

Data Collection – SpaceX API

Used request object to get SpaceX Rest API data

Decoded the response as a JSON utilizing .json() then turned it into a Pandas datafram using .json_normalize()

Used the API further to get information about launches using the ID numbers given in the json object

Put the new collected data into individual list

Put that columns into a dictionary

Created a Pandas dataframe from the dictionary

Filtered the data to only include Falcon 9 launches

Missing values were found in PayloadMass so the mean of all PayloadMass was used to replace the missing values

Data was exported to a csv

[Github: Data Collection]

Data Collection - Scraping

Used a requests.get() to get a response from the Wikipedia page.

Created a BeautifulSoup object from the response

Extracted the column names from the table and put them in a list

Created an empty dictionary with keys from the extracted column names

Created a dataframe from the dictionary

Exported data to CSV

[Github: Webscraping]

Data Wrangling

- load SpaceX data from data collection step
- identify missing values
- identify which columns are numerical and categorical
- calculate the number of launches for each site
- calculate number and occurrence of each orbit
- calculate number and occurrence of mission outcome of the orbuits
- loop over keys to get possible outcomes
- split outcomes into classification variables
- export dataframe to csv

 10

[Github: Data Wrangling]

EDA with Data Visualization

- Payload Mass v. Flight Number. A scatter plot was used to look at payload mass over time. As flight number increased, the first stage was more likely to land successfully.
- Launch Site v. Flight Number. A scatter plot was used to look at when each launch site was used in respect to flight number. Launch site seemed to change randomly as flight number progressed.
- Payload Mass v. Launch Site. A scatter plot was used to see if there was a correlation between launch sites and payload mass. VAFB-SLC launch site never launches rockets with a payload greater than 10000
- Orbit type v. Success Rate. A bar chart was used to see if certain orbits had greater success rates. Certain orbit types have significantly more success rate and certain orbit types didn’t have any success.
- Flight Number v. Orbit type. A scatter plot was used to see if there was a correlation between orbits as flight numbers increased. In some orbit types, success seemed to be related to number of flights, some orbits didn’t have correlations.
- Payload Mass v. Orbit type. A scatter plot was used to analyze if certain orbits had a tendency to have higher payload masses. Heavy payloads had more success with landing for certain orbit types than others.
- Success yearly Trend. A line chart was used to look at success rate over time. The graph showed that over time landings became more successful.

[Github: EDA Data Visualization]

EDA with SQL

- display the unique launch site names
- display 5 records where launch sites begin with 'CCA'
- total payload mass carried by boosters launched by NASA
- average payload mass carried by booster version F9 v1.1
- data when the first successful landing outcome in ground pad was achieved
- names of boosters which have success in drone ship and have payload mass between 4000 and 6000
- total number of successful and failure mission outcomes
- names of booster versions which have carried max payload mass
- records that display the month names, failure landing outcomes in drone ships, booster version, launch site for the month in year 2015
- rank the count of landing outcomes between date 2010-06-04 and 2017-03-20 in descending order

Github: EDA SQL

Build an Interactive Map with Folium

- Used blue folium circle and marker with a text label to mark the NASA Johnson Space Center.
- Used a folium marker and circle to mark each launch site
- Used red and green folium markers and circles to make marker clusters to show success rate of each launch site.
- Used a mouse position object to easily get distance between points on the map.
- Used a PolyLine to draw distance to nearest coastline with distance data included.

Github: Folium Map

Build a Dashboard with Plotly Dash

- Added a launch site drop-down component in order to view the four launch sites success rate comparatively and then be able to select one and see specific success rate details
- Added a callback function to render based on selected site dropdown which allows the interactivity of the launch site drop down to work.
- Added a range slider to see if variable payload is correlated to mission outcome to see if we select different payload ranges if it has some correlating visual patterns.
- Added a callback function to render the scatter plot to add interactivity so we can visually observe how payload may be correlated with mission outcomes for selected site(s).

Github: Plotly Dash

Predictive Analysis (Classification)

created a NumPy array from the column ‘Class’ in data

Standardized the data with StandardScaler() then fit it

Split the data into training and testing groups with train_test_split function

Created a logistic Regression object. Created a support vector machine object. Created a decision tree classifier object. Created a K nearest neighbors object. Created a GridSearchCV on all the above tests.

Calculated the accuracy on the test data using the method score on all tests.

Found the method that performed the best based on performance score and accuracy.

[Github: Classification]

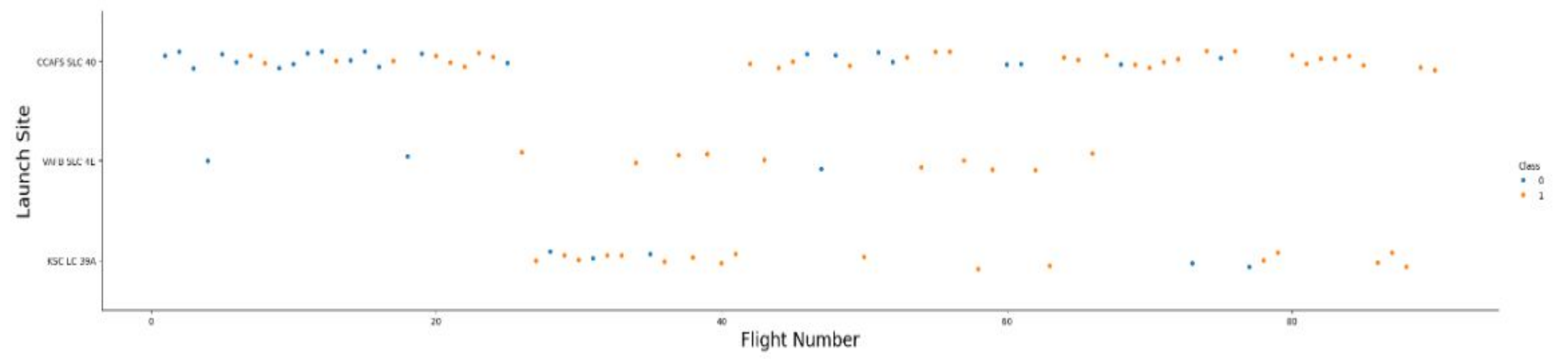
Results

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

Section 2

Insights drawn from EDA

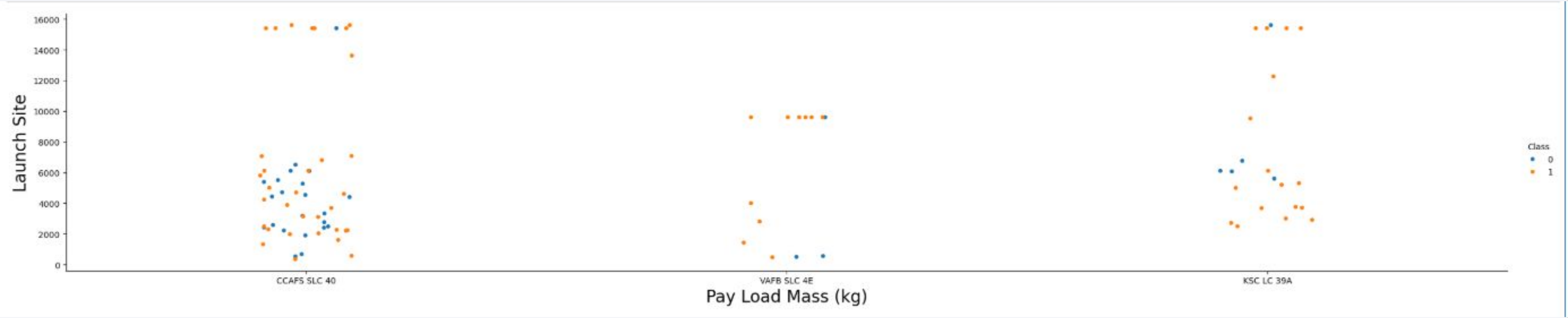
Flight Number vs. Launch Site



Explanation:

- CCAFS SLC 40 had more launches than the other 2 sites and had significantly more successes later on.
- VFAB SLC 4E had fewer launches than both the other sites and had more successes than failures.
- KSC LC 39A seemed to fill in the launch gaps for CCAFS SLC 40 and had more successes than failures.
- All launch sites grew more successful with each progressive flight number

Payload vs. Launch Site



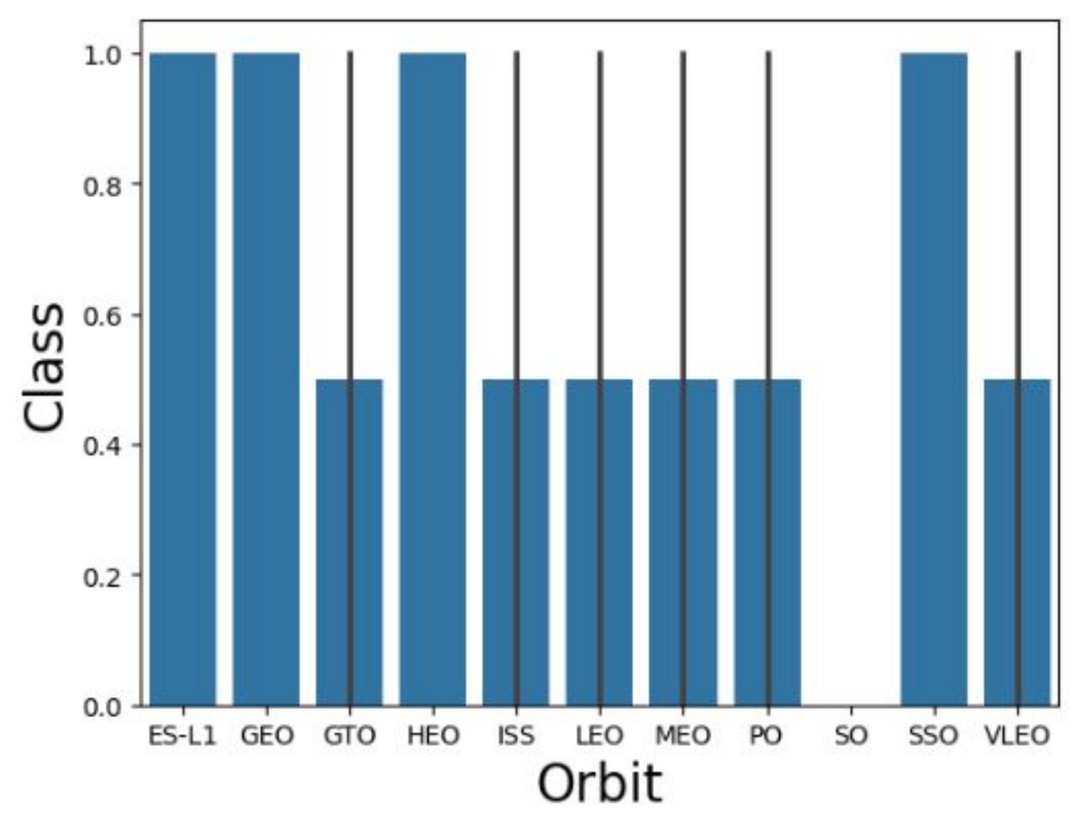
Explanation:

- VAFB SLC 4E payload mass never went above 10000
- Both CCAFS SLC 40 and KSC LC 39A had payload masses in the 160000 with successful landings

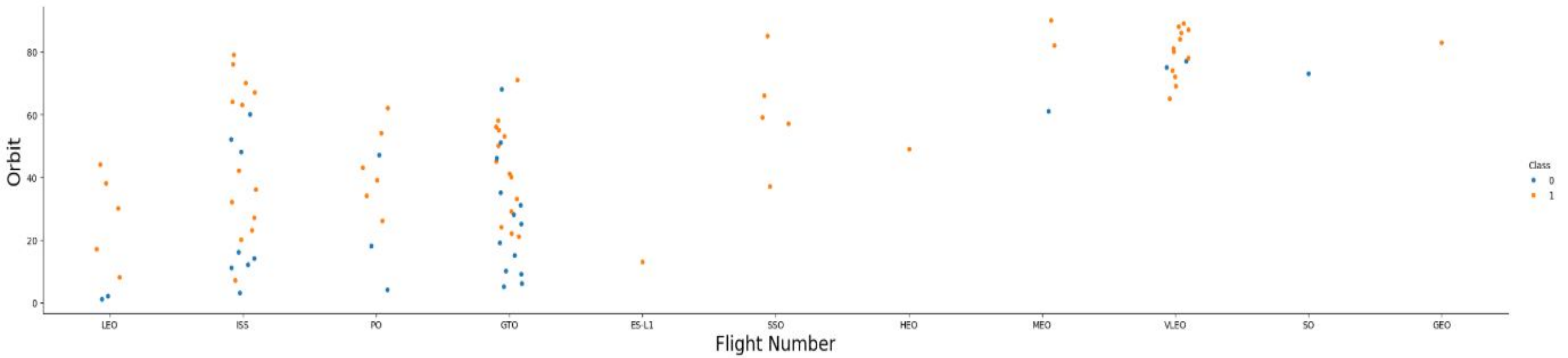
Success Rate vs. Orbit Type

Explanation:

- ES-L1, GEO, HEO, and SSO orbits had 100% success rate.
- SO orbit had no success rate.
- All other orbits had about 50% success rate.



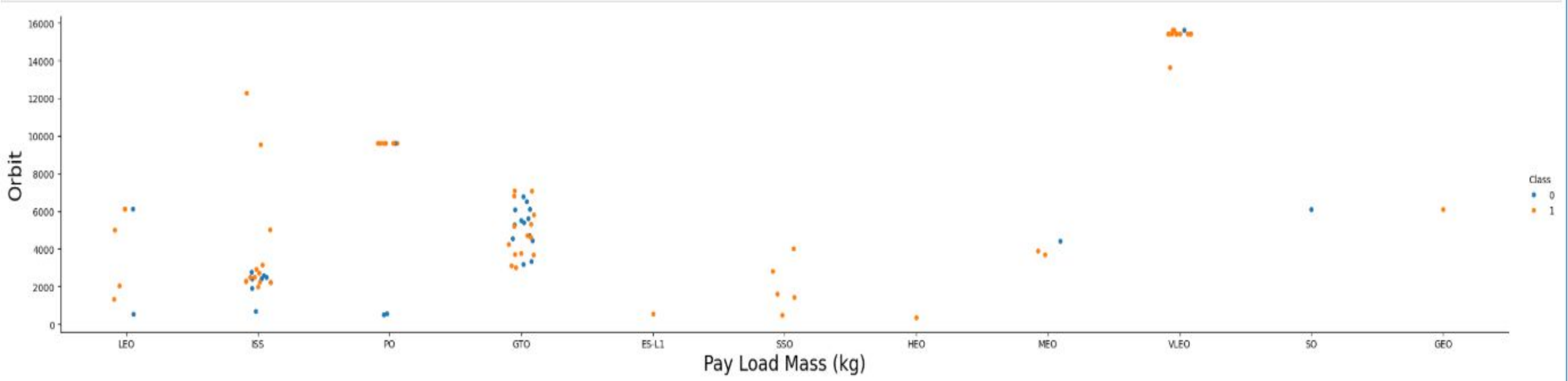
Flight Number vs. Orbit Type



Explanation:

- LEO, ISS, PO, GTO and ES-L1 orbits had earlier flight number launches.
- SSO, HEO, MEO, VLEO, SO, GEO orbits had later flight number launches.
- ISS, PO, GTO had significantly more launches than the other orbits.
- ES-L1, HEO, SO, and GEO only had 1 launch each and only SO orbit's launch failed
- SSO, ES-L1, HEO, and GEO had 100% launch success every time but only SSO had more than 1 launch to orbit.

Payload vs. Orbit Type



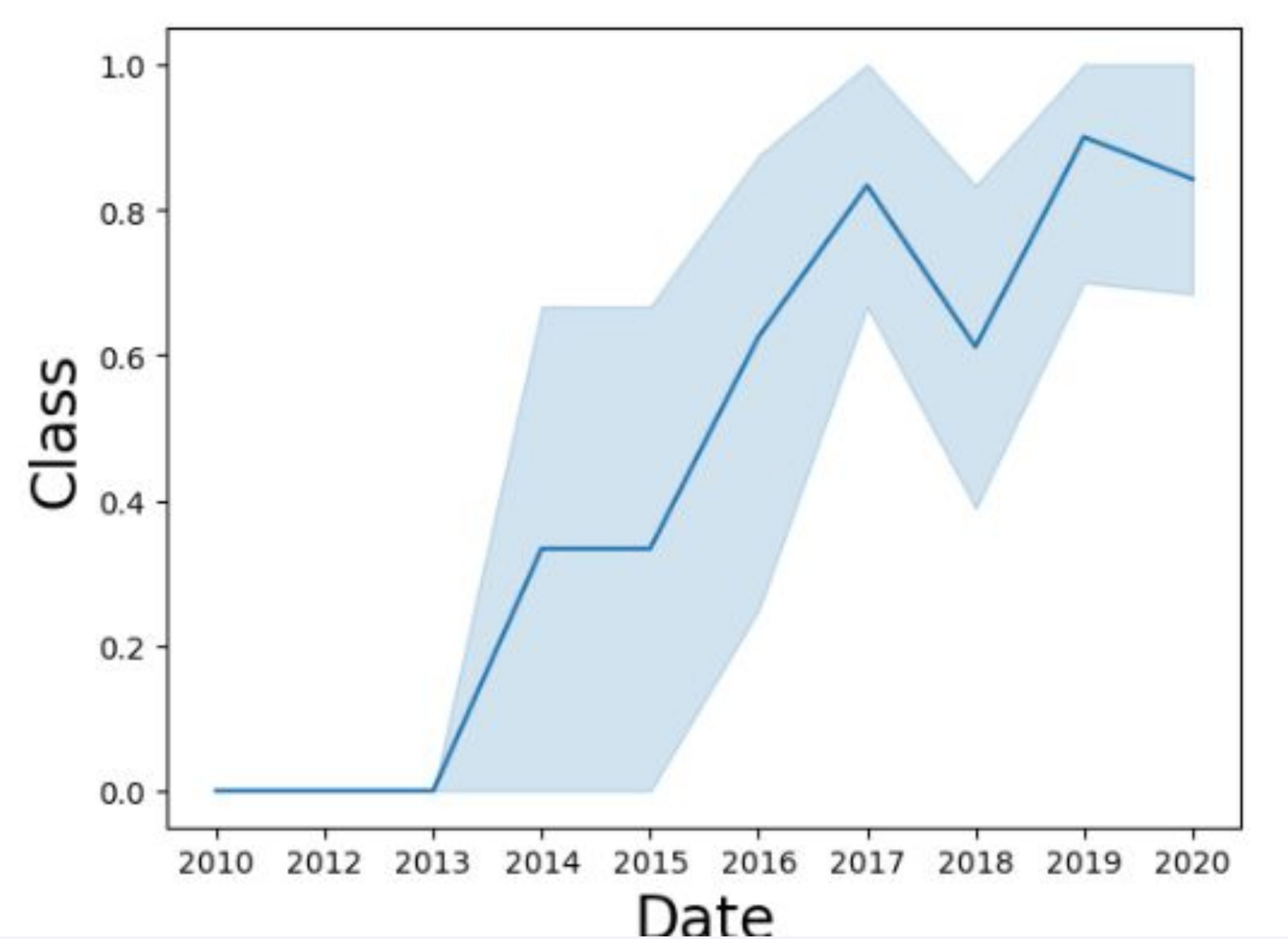
Explanation:

- Heavy payloads had higher successful landing rates for Polar, LEO, ISS, and VLEO.
- VLEO only have heavy payload mass sent to orbit
- HEO and ES-L1 had only 1 launch each and both had light payload masses
- ES-L1, SSO, HEO, and GEO had 100% successful launches but all only had payload masses below 10000.

Launch Success Yearly Trend

Explanation:

- Success rate since 2013 kept increasing till 2020
- Little decline in success in 2018



All Launch Site Names

Explanation:

Selected all unique launch site names with sql from table

Done.

[31]:

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

Launch Site Names Begin with 'CCA'

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:

Selected first 5 records where launch site name begins with ‘CCA’

Total Payload Mass

Explanation:

Selected sum of all payload masses carried by boosters launches by NASA

SUM(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

Explanation:

Selected average payload mass carried by booster F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Explanation:

Selected first successful landing outcome in ground pad

5]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Explanation:

Selected list of boosters in which drone ship was successful and had a payload mass between 4000 and 6000.

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Explanation:

Selected total number of successful and failed mission outcomes

Successful

count(*)
100

Failed

count(*)
1

Boosters Carried Maximum Payload

Explanation:

Selected names of booster versions which have carried the max payload mass

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

Explanation:

Listed records which display the months, failure landing outcomes in drone ships, booster versions, and launch sites for the months in 2025

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	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 Between 2010-06-04 and 2017-03-20

Explanation:

Rank the count of landing outcomes between the data of 2010-06-04 through 2017-03-20 in descending order

Landing_Outcome	count_outcomes
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

Section 3

Launch Sites Proximities Analysis

Launch Sites on map

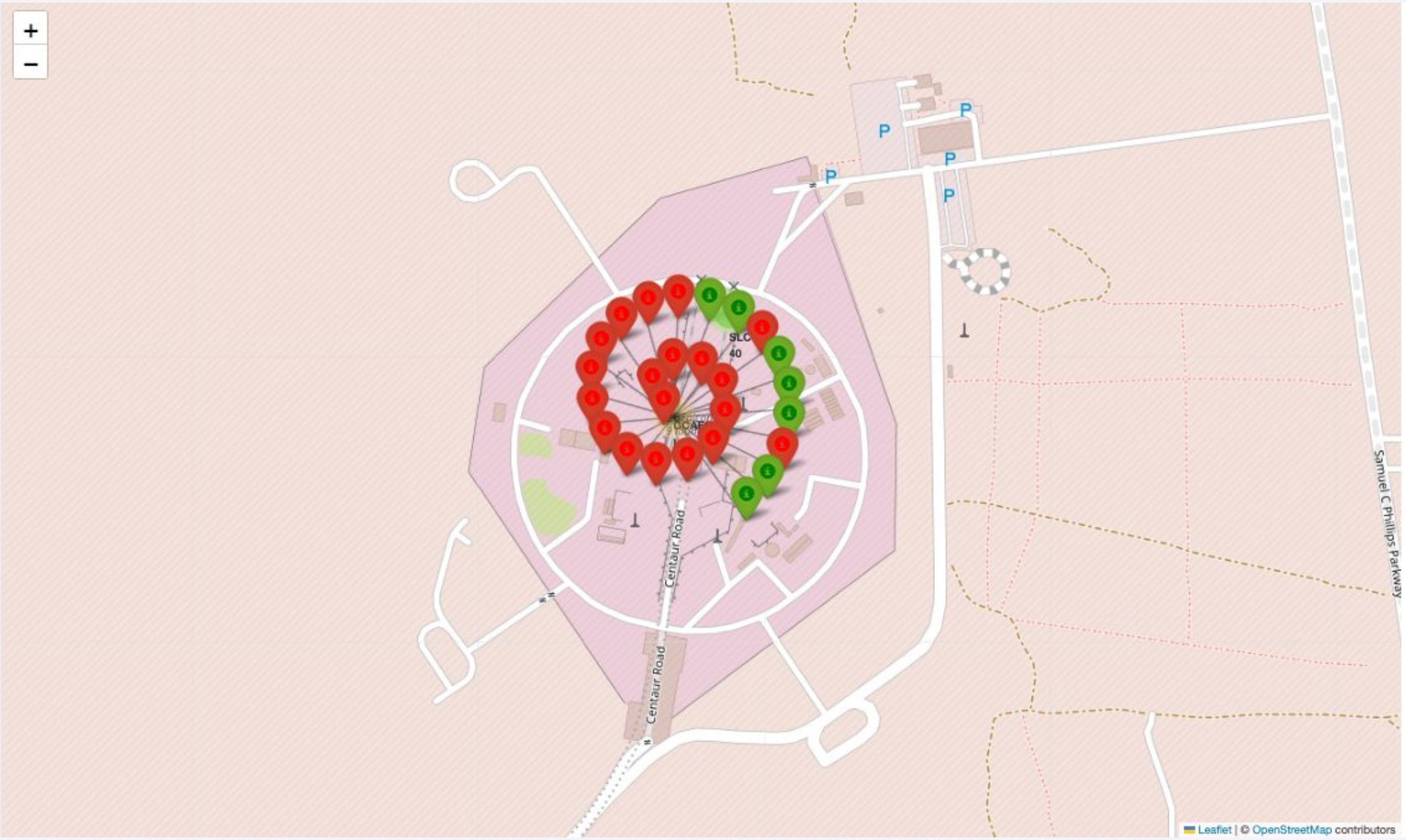
Explanation:

- Launch sites tend to exist near the coastline so that it minimizes the risk of injuring people when faults happen during the launch.
- Launch sites also tend to be near the equator, where the earth is moving faster, to aid in getting rockets up to speed during launch.

Clustered Launch Records for each Launch Site

Explanation:

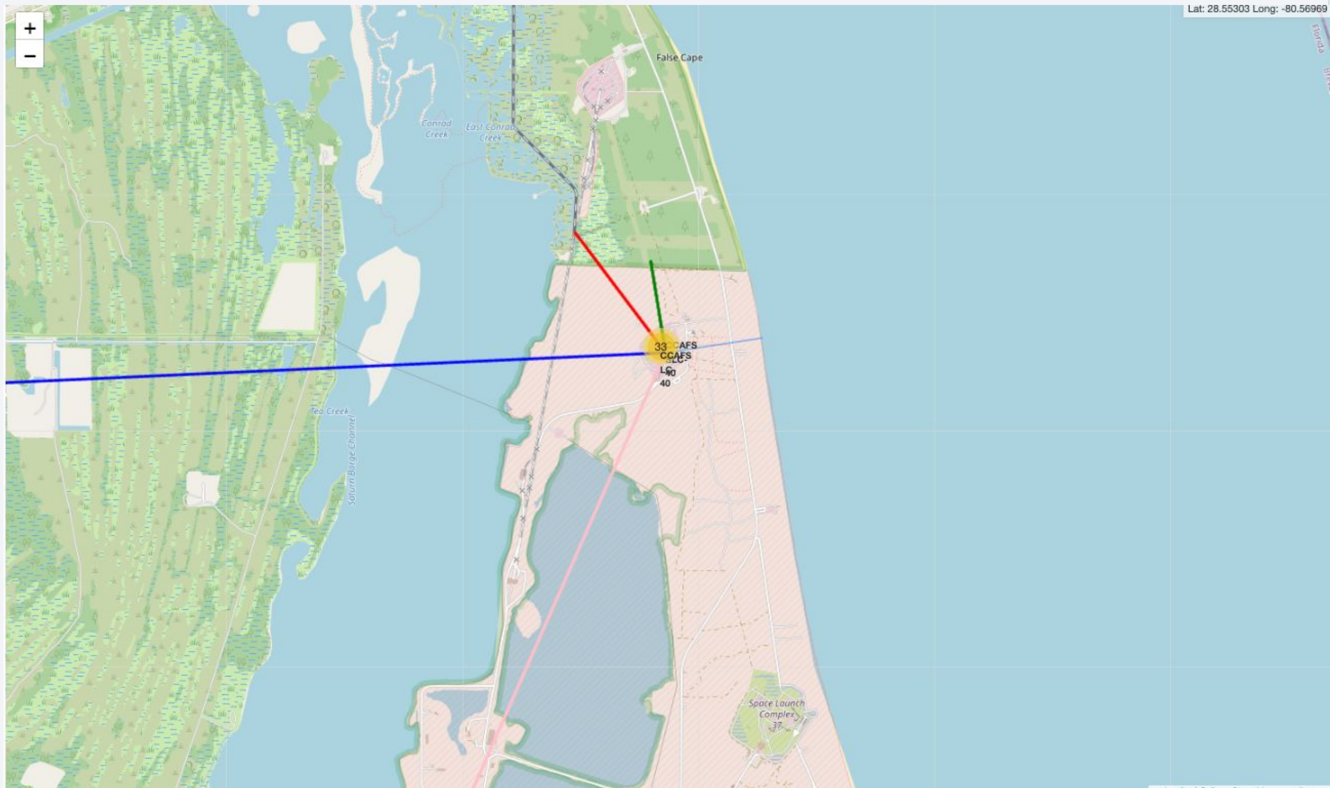
- Green Marker is a successful launch
- Red Marker is a failed launch.
- Launch site CCAFS LC-40 has had more failed launches than successful launches as indicated by the little circles spiraling out from the center.



Lines to Nearest Feature of Note

Explanation:

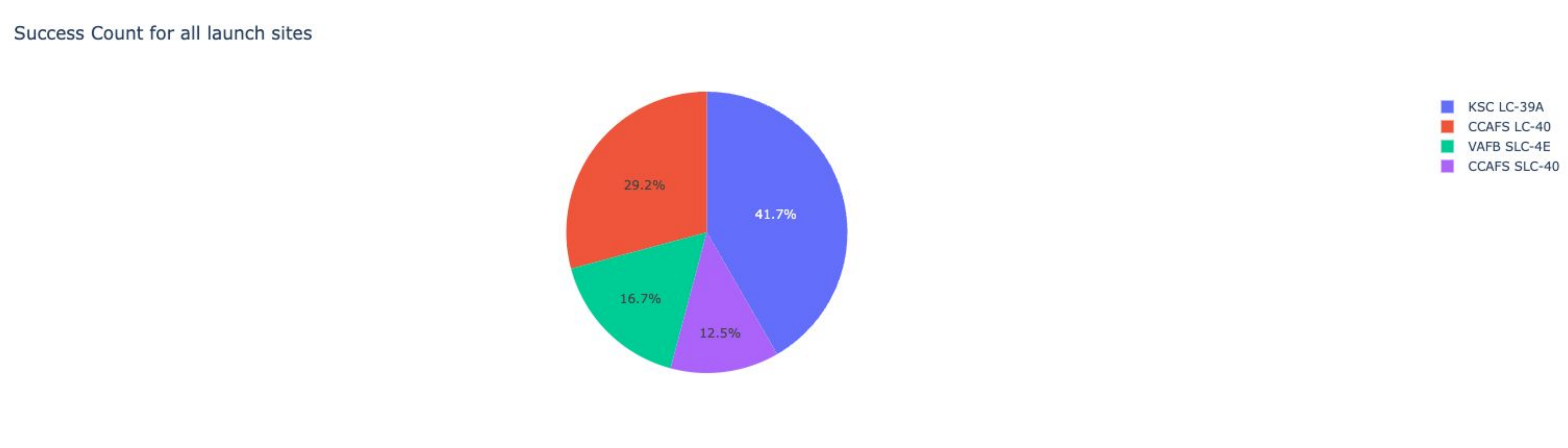
- Visualizing distances to nearest features of note is good to see what kind of impact launch sites are having on neighboring areas.
- CCAFS LC-40 is fairly close to major cities, wildlife refuge areas, airports, and railroads.



Section 4

Build a Dashboard with Plotly Dash

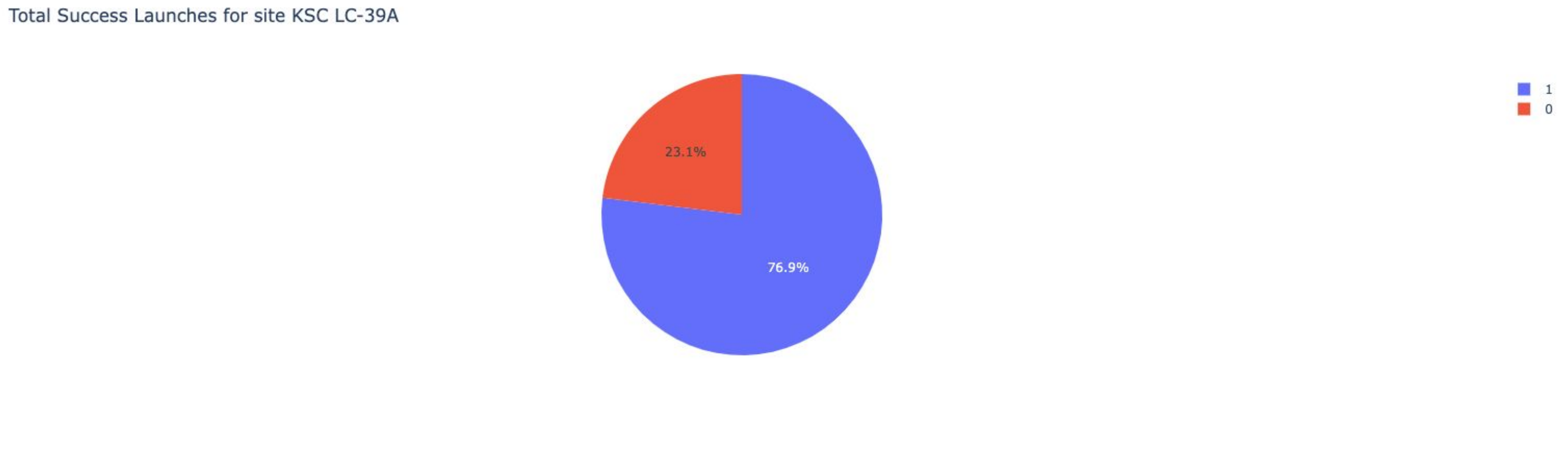
Success Count for all Launch Sites



Explanation:

The pie chart shows that launch site KSC LC-39A had the most successful launches.

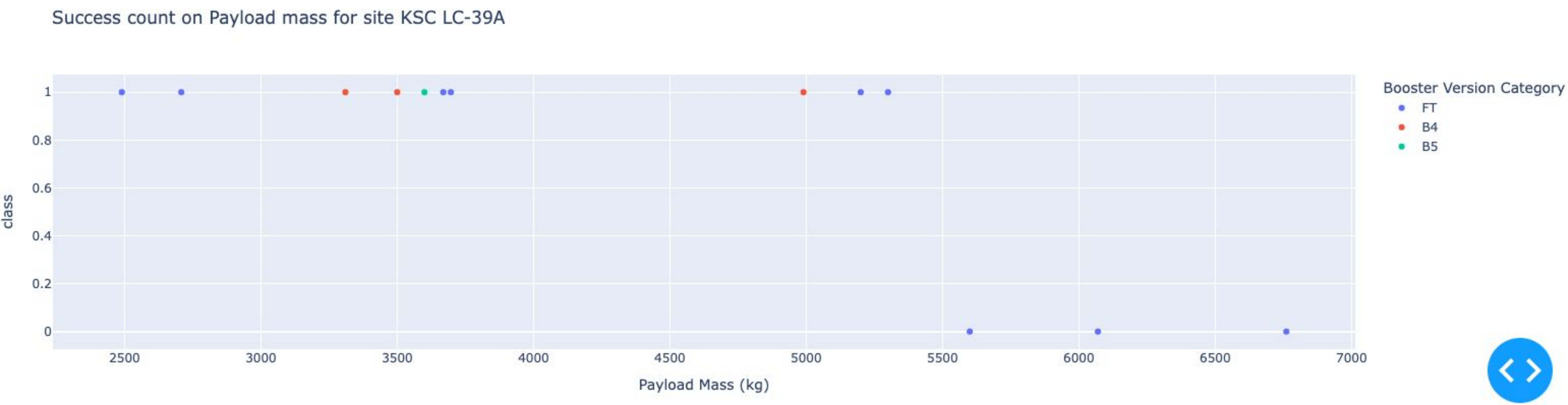
Total Success Launches for site KSC LC-39A



Explanation:

The pie chart shows that the launch site KSC LC-39A had 76.9% successful launches and 23.1% failed launches.

Success count on Payload Mass for site KSC LC-39A



Explanation:

The scatter plot chart shows that launches with a payload mass less than 5500 were more successful.

Section 5

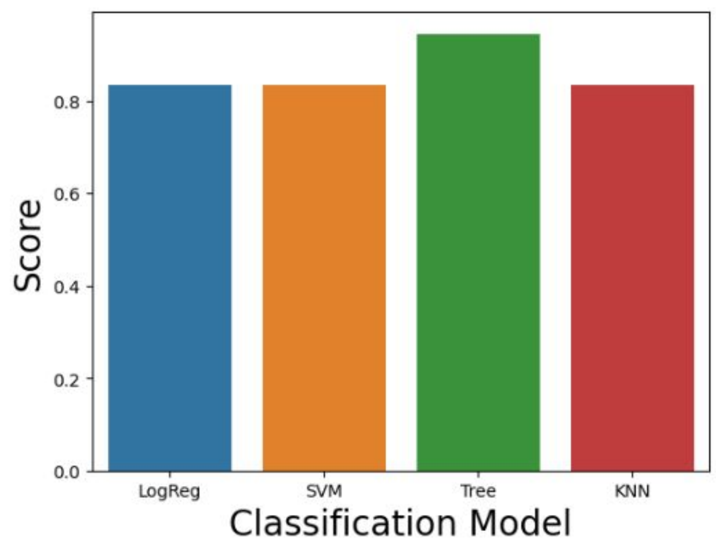
Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- It is difficult to clearly see what classification type scored the best on test data, perhaps that is due to a small test sample size, but on all the data it looks like the tree classification had the highest jaccard score, f1 score, and accuracy score.

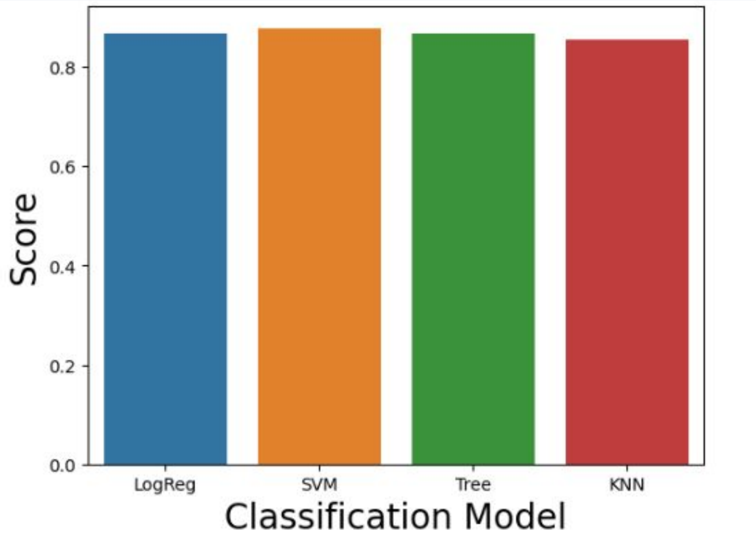
Scores based on all data



	LogReg	SVM	Tree	KNN
Accuracy	0.833333	0.833333	0.944444	0.833333

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.923077	0.800000
F1_Score	0.888889	0.888889	0.960000	0.888889
Accuracy	0.833333	0.833333	0.944444	0.833333

Scores based on test data

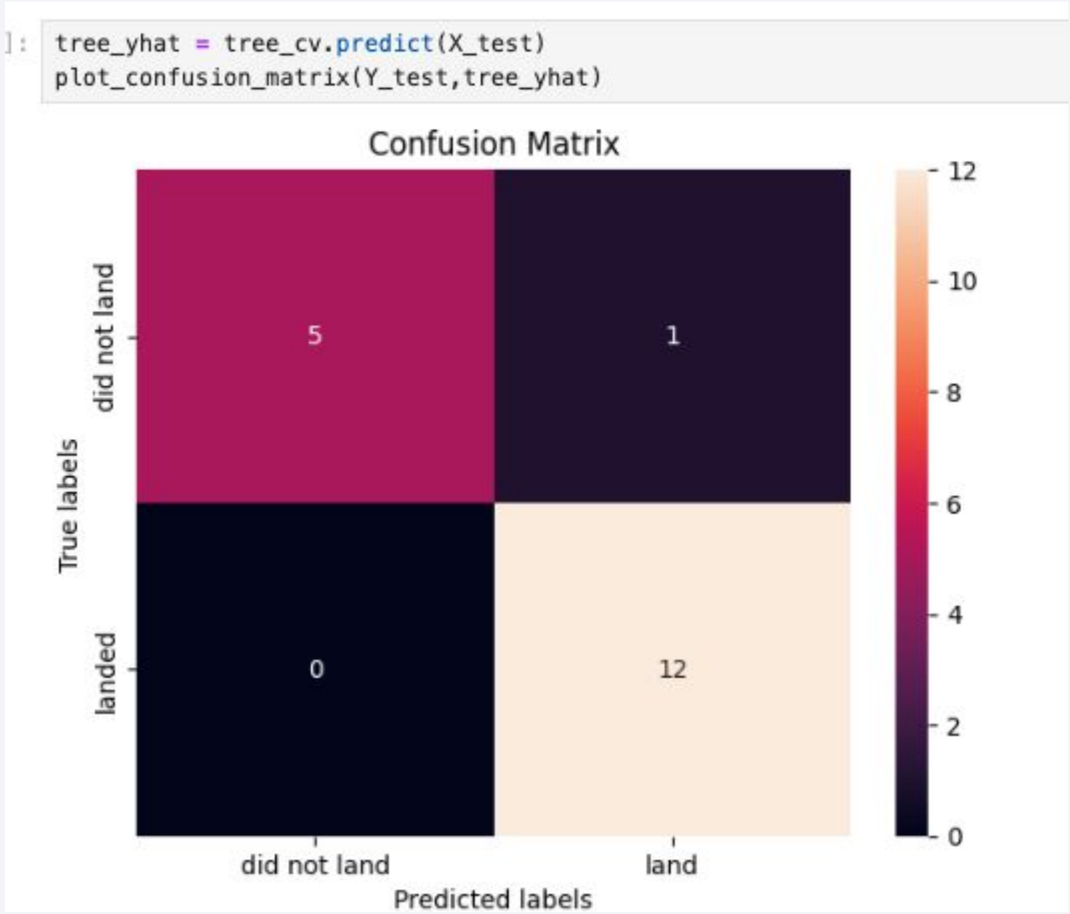


	LogReg	SVM	Tree	KNN
Accuracy	0.866667	0.877778	0.866667	0.855556

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.818182	0.819444
F1_Score	0.909091	0.916031	0.900000	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

Confusion Matrix

- The confusion matrix predicted one false positive while all other predictions were correct.



Conclusions

- The best type of classification for this dataset is the tree classification.
- Clearly there was a learning curve as launches become more successful as time and flight numbers progressed.
- Launches will lower payload masses had a better outcome than launches with higher payload masses.
- KCS LC-39A had the highest amount of successful launches.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!