IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cami Santor
7/14/24

Outline

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

Executive Summary

- Launch data was collected using SpaceX's python API
- Falcon 9 and Falcon Heavy launch records were scraped from Wikipedia
- Data wrangling was completed to classify successful and unsuccessful landings
- Exploratory data analysis was done with SQL and visualization
- Interactive analysis was done by creating an interactive dashboard with Plotly Dash and interactive maps with Folim
- Predictive analysis was done by creating four predictive machine learning models
- All 4 models had an accuracy score of 83.33%

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62 million, while competitors cost upward of $165. Much of SpaceX's savings is due to the fact that SpaceX can reuse the first stage.

If we can predict the likelihood of an unsuccessful first stage landing we can predict the true cost of SpaceX's launch and use that to our advantage.

Various data collection, scraping, wrangling, and analysis was done to formulate a model to make this prediction.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Launch data was collected using SpaceX's python API and scraped from Wikipedia.
- Perform data wrangling
  - Data wrangling was completed to classify successful and unsuccessful landings using python's pandas library.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four predictive analysis models were created and tuned. The optimal parameters for each model were found using python's library scikit-learn's GridSearchCV.

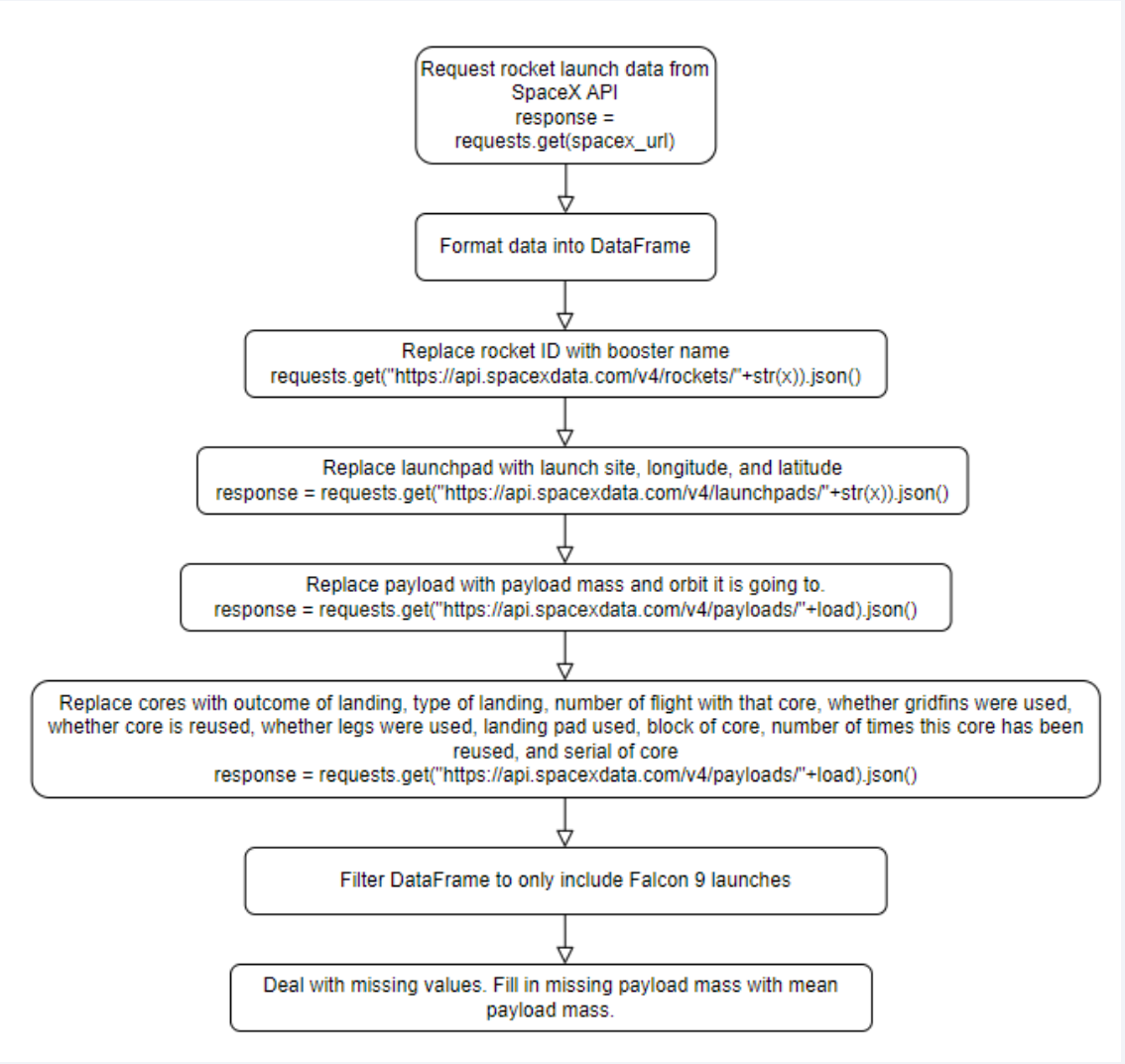
Data Collection

- Data was collected from the SpaceX website and from Wikipedia.
- To collect data from SpaceX, the SpaceX API was used. The data was formatted into a Pandas DataFrame. This DataFrame contained many IDs for each launch including IDs for rocket, payloads, launchpad, and cores. The SpaceX API was used again to fill in these IDs with meaningful values. For example, the rocket ID was filled in with the name of the booster, the payload ID was used to retrieve the mass of the payload and the orbit that it was going to, etc.
- Then the data was filtered to only include data on Falcon 9 launches. Any missing values were dealt with.
- Then launch success rates were retrieved by web scraping Wikipedia.

Data Collection – SpaceX API

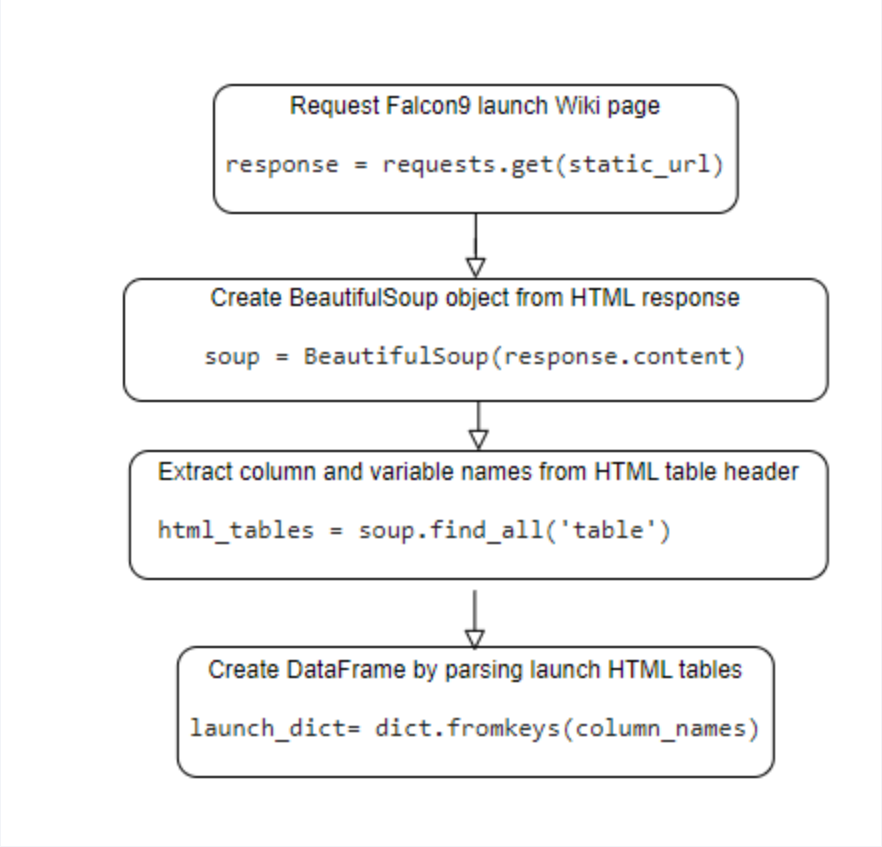
- Data collection with SpaceX REST calls flowchart

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/camisantor/ibm-course/blob/main/labs/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scraping process flowchart
- GitHub URL of the completed web scraping notebook: https://github.com/camisantor/ibm-course/blob/main/labs/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data wrangling was done to find patterns in the data
- Value counts were taken of each launch site, orbit
- Landing outcomes were labeled as good or bad
- The success rate was calculated
- GitHub URL of data wrangling related notebooks: https://github.com/camisantor/ibm-course/blob/main/labs/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

The following scatter charts were plotted

- Flight number vs. payload mass
- Flight number vs. launch site
- Payload mass vs. launch site
- Flight number vs. orbit
- Payload mass vs. Orbit

The following other charts were plotted

- A bar chart of orbit type versus mean success rate
- A line graph of mean success rate vs. Year

These charts were plotted to see how the different variables affected the landing outcome.

GitHub URL of EDA with data visualization notebook: https://github.com/camisantor/ibm-course/blob/main/labs/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Exploratory data analysis was performed using SQL
- SQL was used to:
  - Get SpaceX's unique launch sites
  - Get the average payload mass carried by different booster versions
  - Get the date of when the first successful landing outcome in drone ship was achieved
  - List the the names of the boosters which have success in ground pad and have a payload mass between 4000- 6000 kg
  - List the total number of successful and failure mission outcomes
  - List the names of booster versions which have carried the maximum payload mass
  - List the information about launches by year
  - List the ranked count of landing outcomes during a given time period
- GitHub URL of completed EDA with SQL notebook: https://github.com/camisantor/ibm-course/blob/main/labs/jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

- Folium was used to create visual analytics and explore how geographical variables affected launch outcomes
- The following objects were plotted on the folium graphics:
  - Launch sites were marked with a circle
  - A marker cluster with each launch, marked green for successful outcome and red for failed outcome
  - A mouse position was added to get the latitude and longitude of nearby points of interest to analyze
  - A line and distance marker were added between launch sites and points of interest to look for a trend in proximity to points of interest
- GitHub URL of completed interactive map with Folium map: https://github.com/camisantor/ibm-course/blob/main/labs/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- The following were added to the plotly dash app for data analysis:
  - A launch site drop-down to enable analysis of all sites collectively or individually
  - A pie chart to show success rate
  - A scatter chart to show how payload mass and booster version category affect success rate
    - A payload range slider was added to enable interactive customization of the payload range the chart showed
- GitHub URL of completed Plotly Dash app: https://github.com/camisantor/ibm-course/blob/main/labs/spacex_dash_app.py

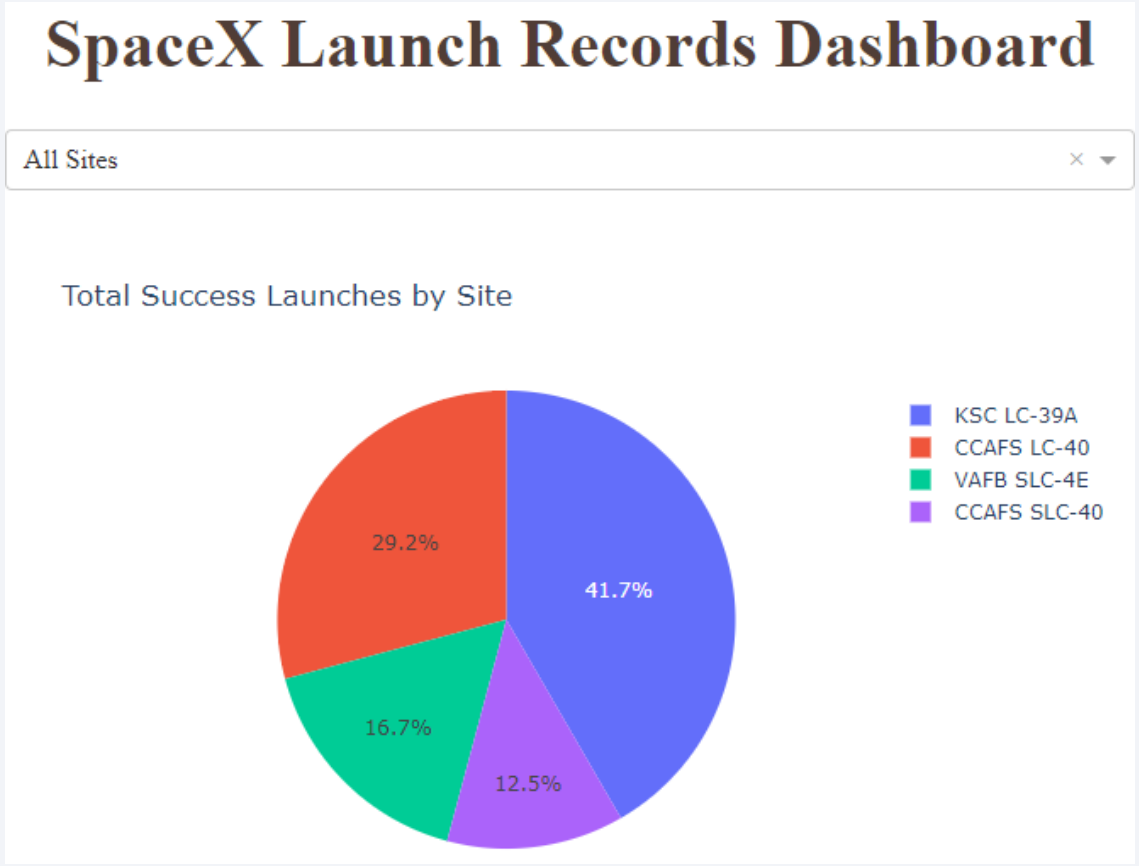
Predictive Analysis (Classification)

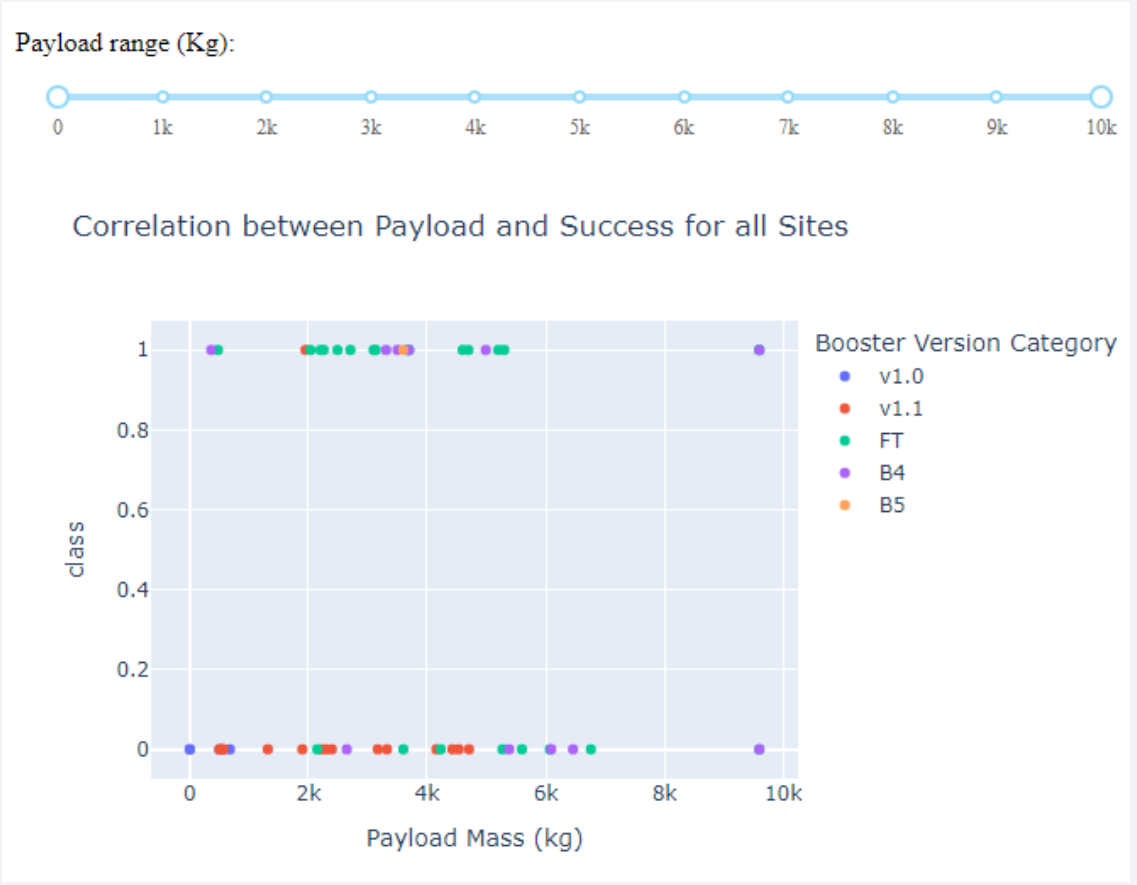
- Data was standardized and split into training and testing sets
- The following models were created:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree Classifier
  - K Nearest Neighbor
- Each model was tuned using a GridSearchCV to find the parameters that achieved the most accurate model
- The models were evaluated by calculating the accuracy score and plotting the confusion matrix
- GitHub URL of predictive analysis: https://github.com/camisantor/ibm-course/blob/main/labs/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results: Exploratory Data Analysis

- Exploratory data analysis results showed some trends in the data
- The following trends were observed:
  - Different launch sites had different success rates
    - CCAFS LC-40 has a success rate of about 60%
    - KSC LC-39A and VAFB SLC 4E have a success rate of about 77%
  - Different orbits had different success rates and variables affected the success rates differently
    - LEO orbit had a positive correlation of flight number to success
    - Heavy payloads had a higher successful landing rate for Polar, LEO, and ISS orbits
  - Success rate increased from 2013 to 2020

Results: Interactive Analytics



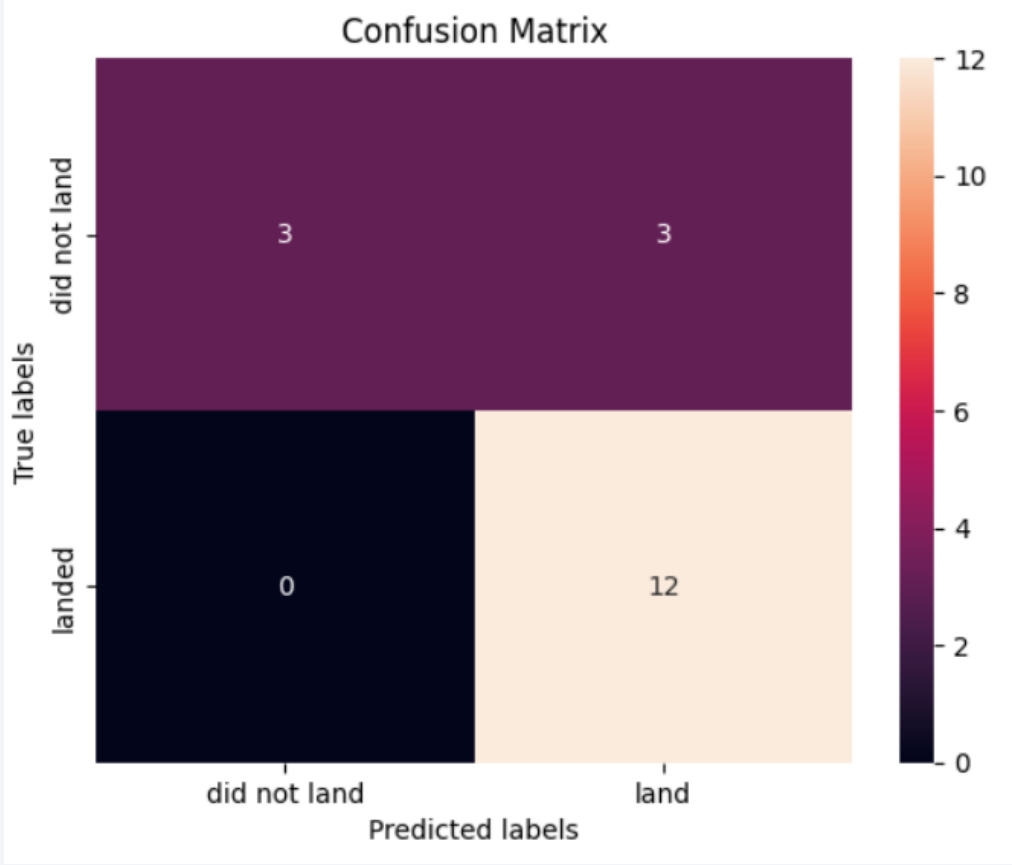


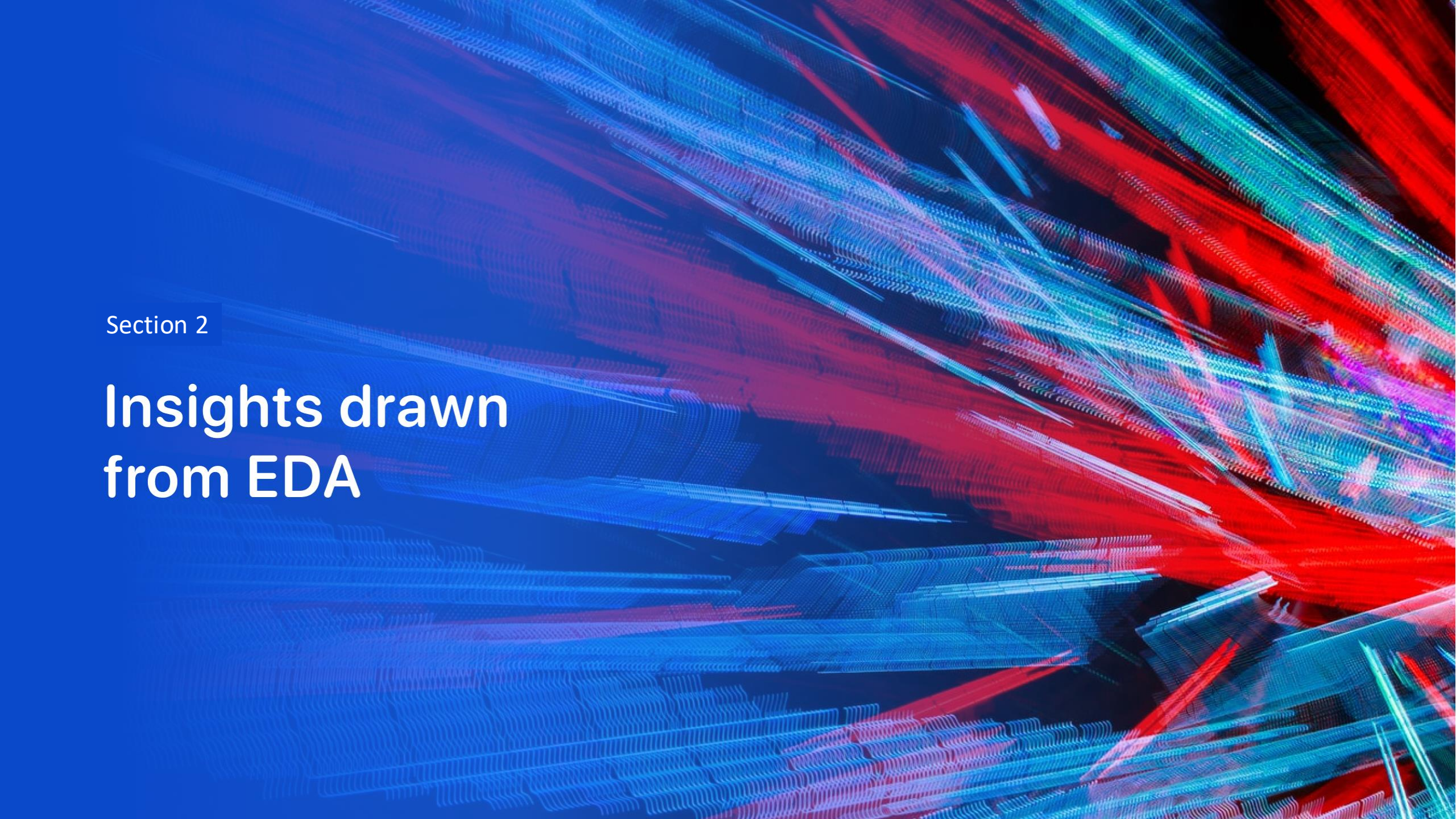
Results: Predictive Analysis

- All models once tuned to the optimal parameters had the same accuracy score of 0.8333 and confusion matrix values
- Each confusion matrix had 12 true positives and 3 false positives

```
knn_cv.score(X_test, Y_test)
```

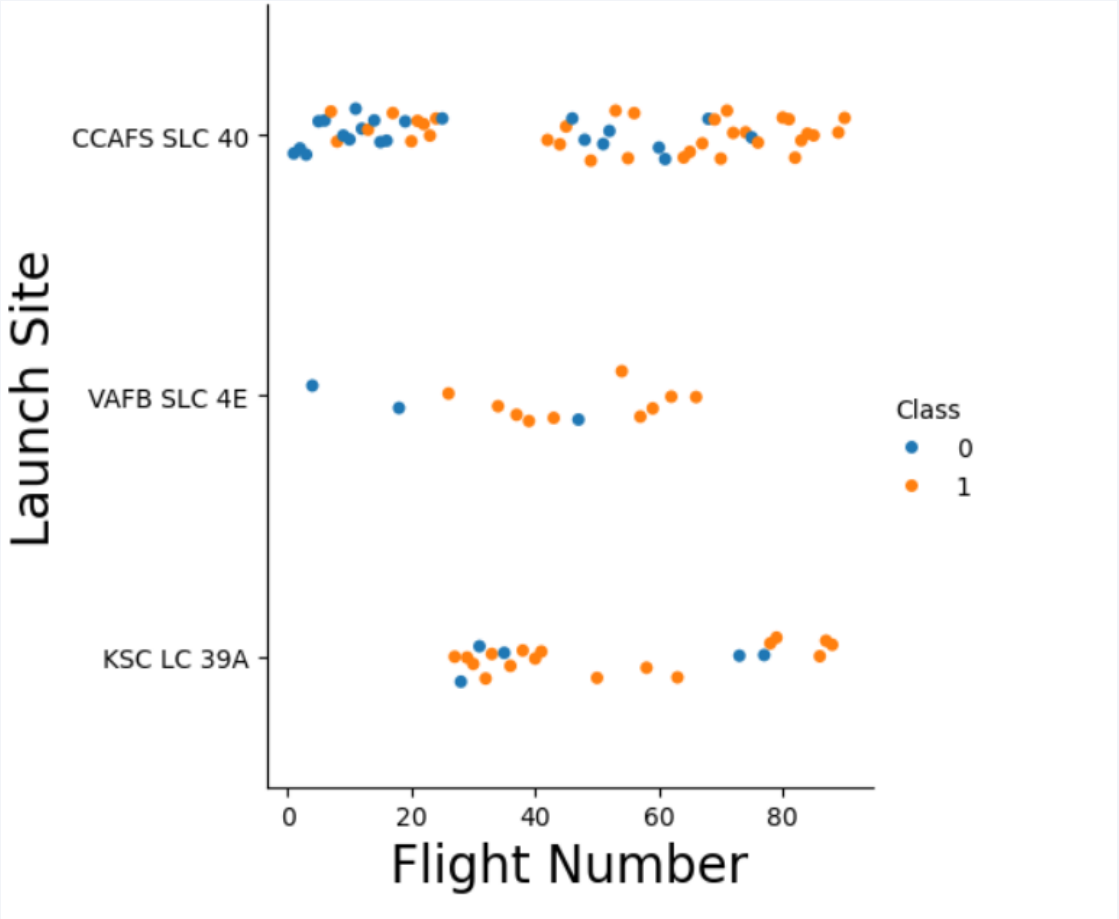
```
0.8333333333333334
```




Section 2
Insights drawn from EDA

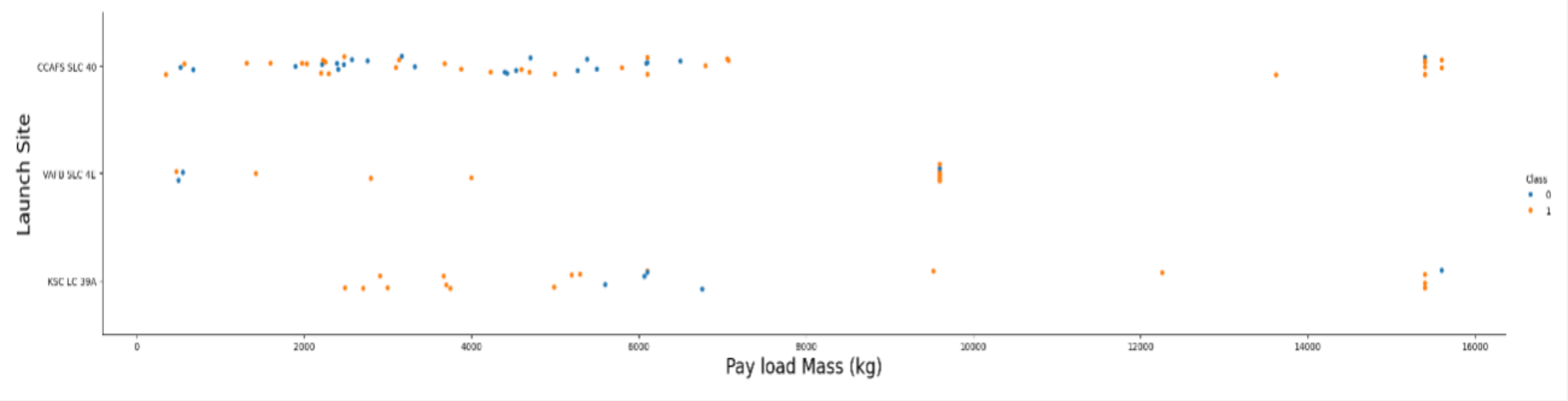
Flight Number vs. Launch Site

- This plot shows:
  - Higher success rates at KSC LC 39A and VAFB SLC 4E
  - Higher success rates with higher flight numbers
  - More early flight numbers were done at CCAFS SLC 40



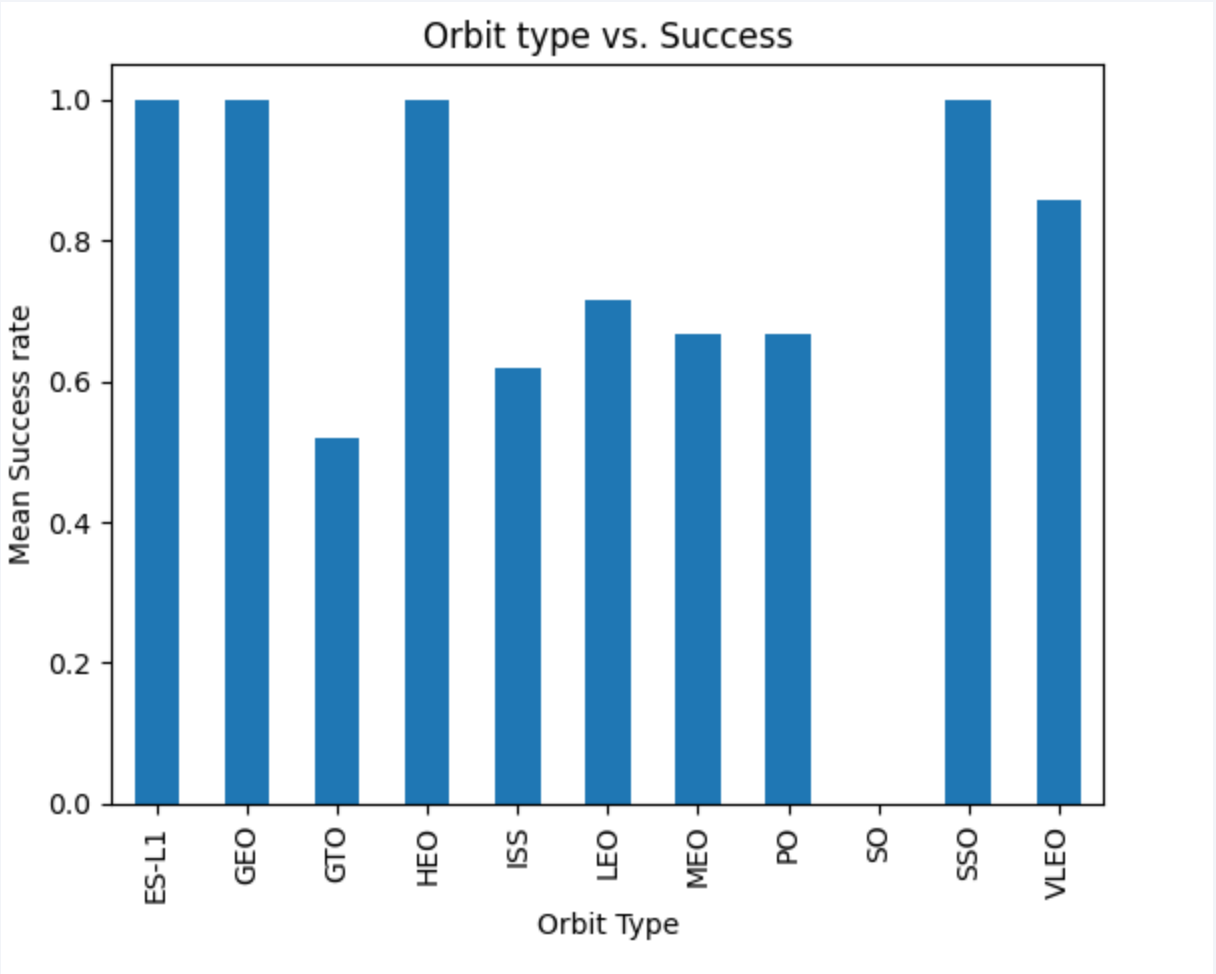
Payload vs. Launch Site

- The scatter plot of Payload vs. Launch Site shows
  - There are no rockets launched for heavy payload mass (greater than 1000) from VAFB SLC launch site
  - KSC LC 39 A launch site has a high success rate with payload mass less than 5000 kg



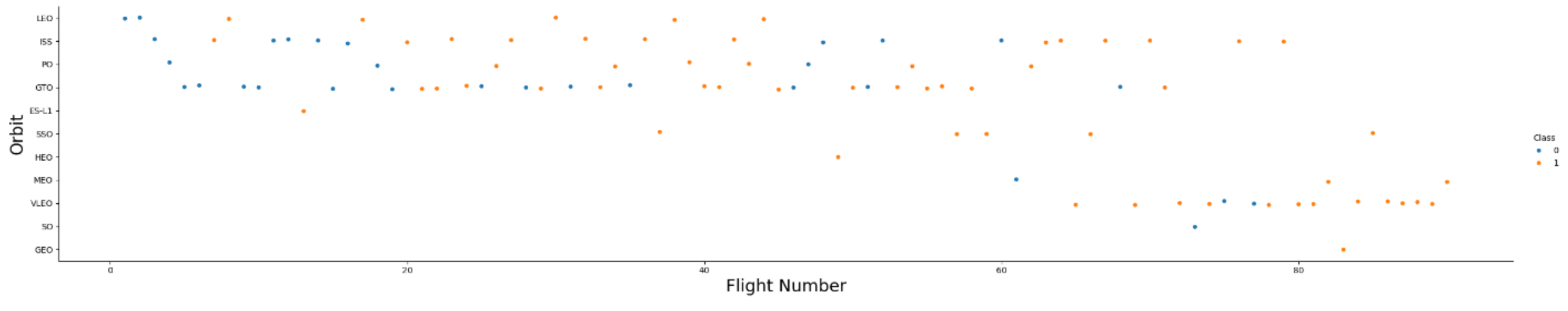
Success Rate vs. Orbit Type

- The bar chart for the success rate of each orbit type shows
  - ES-L1, GEO, HEO, and SSO have the highest success rates
  - GTO has the lowest success rate
  - There is no launch data from a launch to a SO orbit



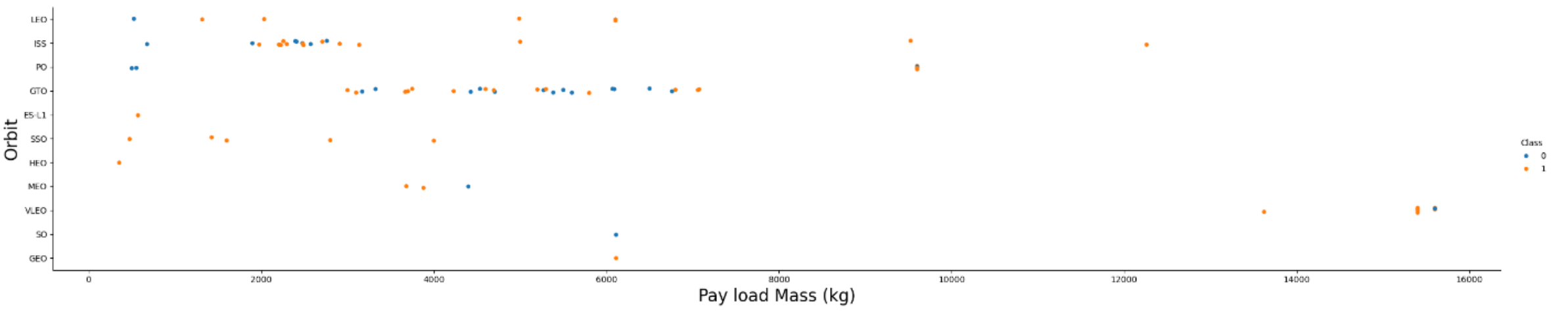
Flight Number vs. Orbit Type

- The scatter plot of Flight number vs. Orbit type shows:
  - The LEO orbit success rate correlates positively to flight number
  - There is no correlation between flight number and success rate in GTO orbit



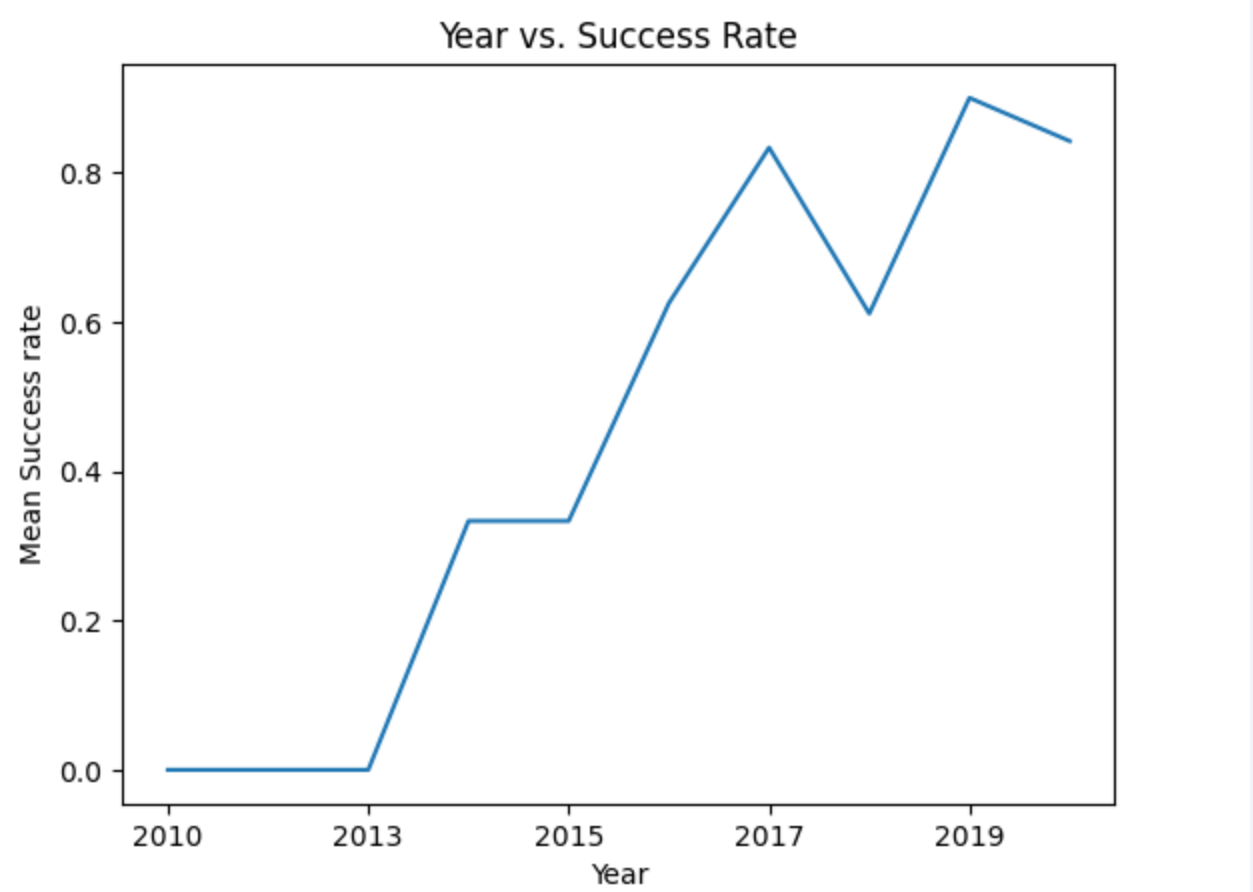
Payload vs. Orbit Type

- The scatter plot of payload vs. orbit type shows
  - The success rate for heavy payloads is higher for Polar, LEO, and ISS orbits
  - There appears to be no correlation between payload mass and success rate for GTO orbit



Launch Success Yearly Trend

- The line chart of year vs. average success rate shows the success rate has increased steadily from 2013 to 2020



All Launch Site Names

- A list of the names of the unique launch sites were found by querying the SQL database
- Query result:

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

Launch Site Names Begin with 'KSC'

- The SQL database was queried for 5 records where launch sites' names start with `KSC`
- Query result:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- The SQL database was queried to calculate the total payload carried by boosters from NASA
- Query result:

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- The SQL database was queried to calculate the average payload mass carried by booster version F9 v1.1
- Query result:

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- The SQL database was queried to find the dates of the first successful landing outcome on drone ship.
- Query result:

MIN("Date")
2016-04-08

Successful Drone Ship Landing with Payload between 4000 and 6000

- The SQL database was queried to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query result:

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- The SQL database was queried to calculate the total number of successful and failure mission outcomes
- Query result:

Count(*)
71

Boosters Carried Maximum Payload

- The SQL database was queried to list the names of the booster which have carried the maximum payload mass

Query result:

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

2017 Launch Records

- The SQL database was queried to list the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

- Query result:

monthnames	Landing_Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

- The SQL database was queried to 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
- Query result:

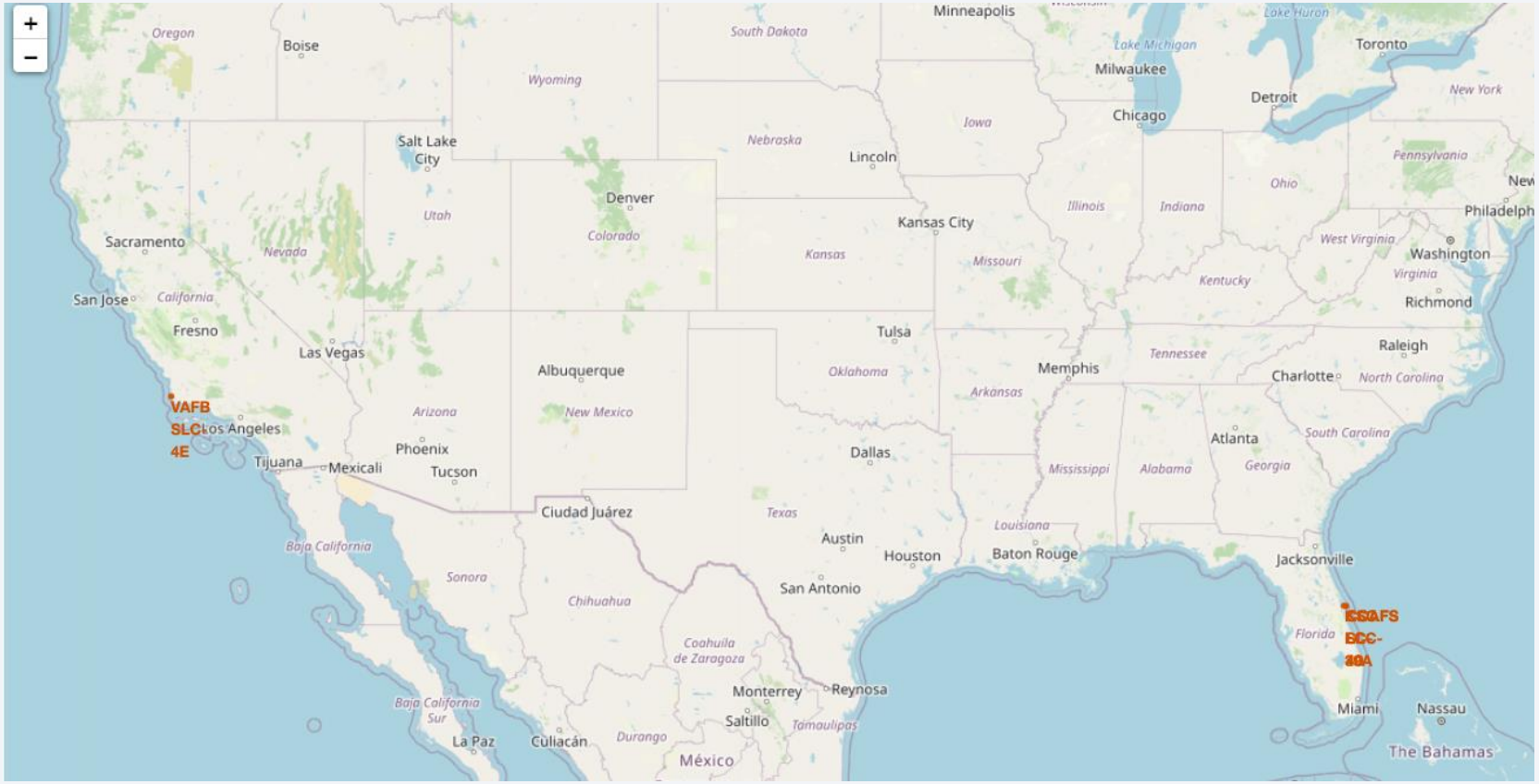
Date	Landing_Outcome	Landing_Outcome_rank
2015-02-11	Controlled (ocean)	1
2014-07-14	Controlled (ocean)	2
2014-04-18	Controlled (ocean)	3
2016-06-15	Failure (drone ship)	1
2016-03-04	Failure (drone ship)	2
2016-01-17	Failure (drone ship)	3
2015-04-14	Failure (drone ship)	4
2015-01-10	Failure (drone ship)	5
2010-12-08	Failure (parachute)	1
2017-03-16	No attempt	1
2015-04-27	No attempt	2
2015-03-02	No attempt	3
2014-09-07	No attempt	4
2014-08-05	No attempt	5
2014-01-06	No attempt	6
2013-12-03	No attempt	7
2013-03-01	No attempt	8
2012-10-08	No attempt	9
2012-05-22	No attempt	10
2015-06-28	Precluded (drone ship)	1
2017-01-14	Success (drone ship)	1
2016-08-14	Success (drone ship)	2
2016-05-27	Success (drone ship)	3
2016-05-06	Success (drone ship)	4
2016-04-08	Success (drone ship)	5
2017-02-19	Success (ground pad)	1
2016-07-18	Success (ground pad)	2
2015-12-22	Success (ground pad)	3
2014-09-21	Uncontrolled (ocean)	1
2013-09-29	Uncontrolled (ocean)	2

Section 3

Launch Sites Proximities Analysis

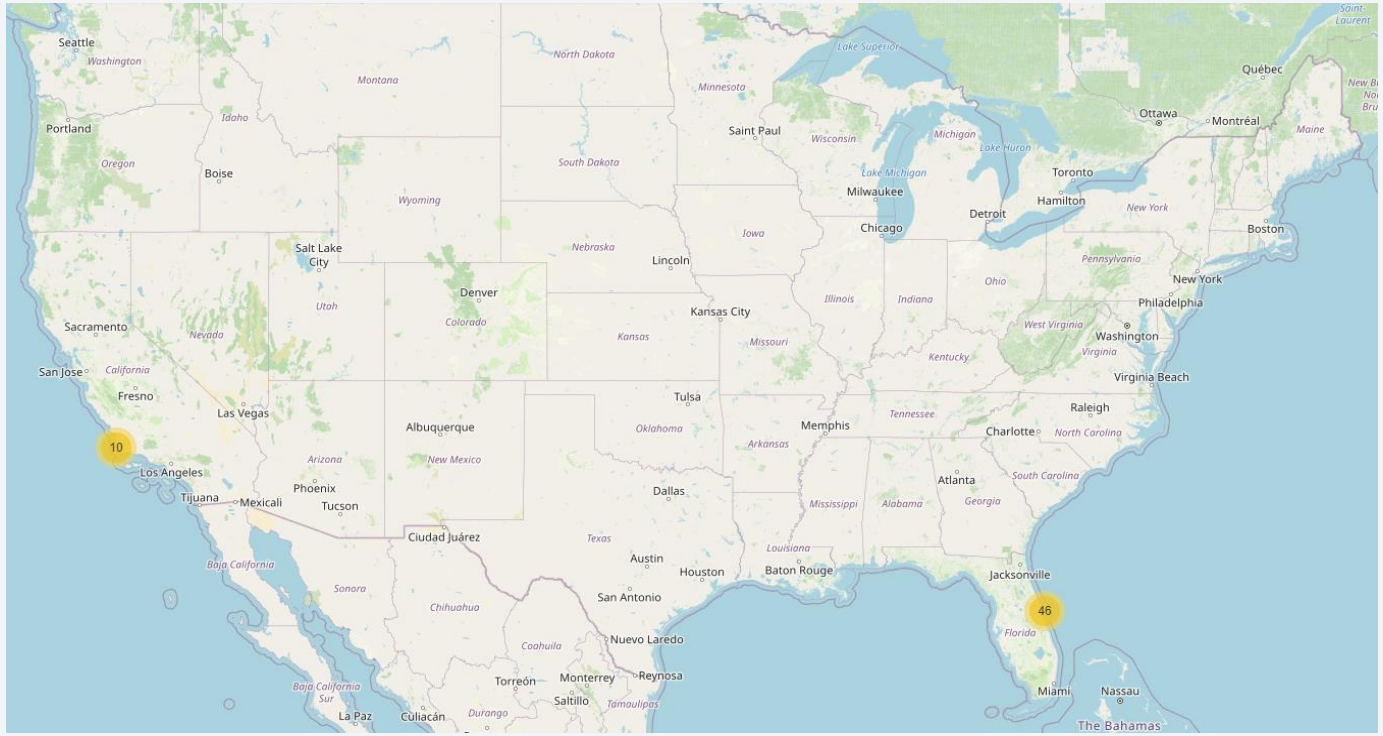
Map of Launch Sites

Map of SpaceX Launch Sites shows all launch sites are in the southern part of the U.S. and near the coast



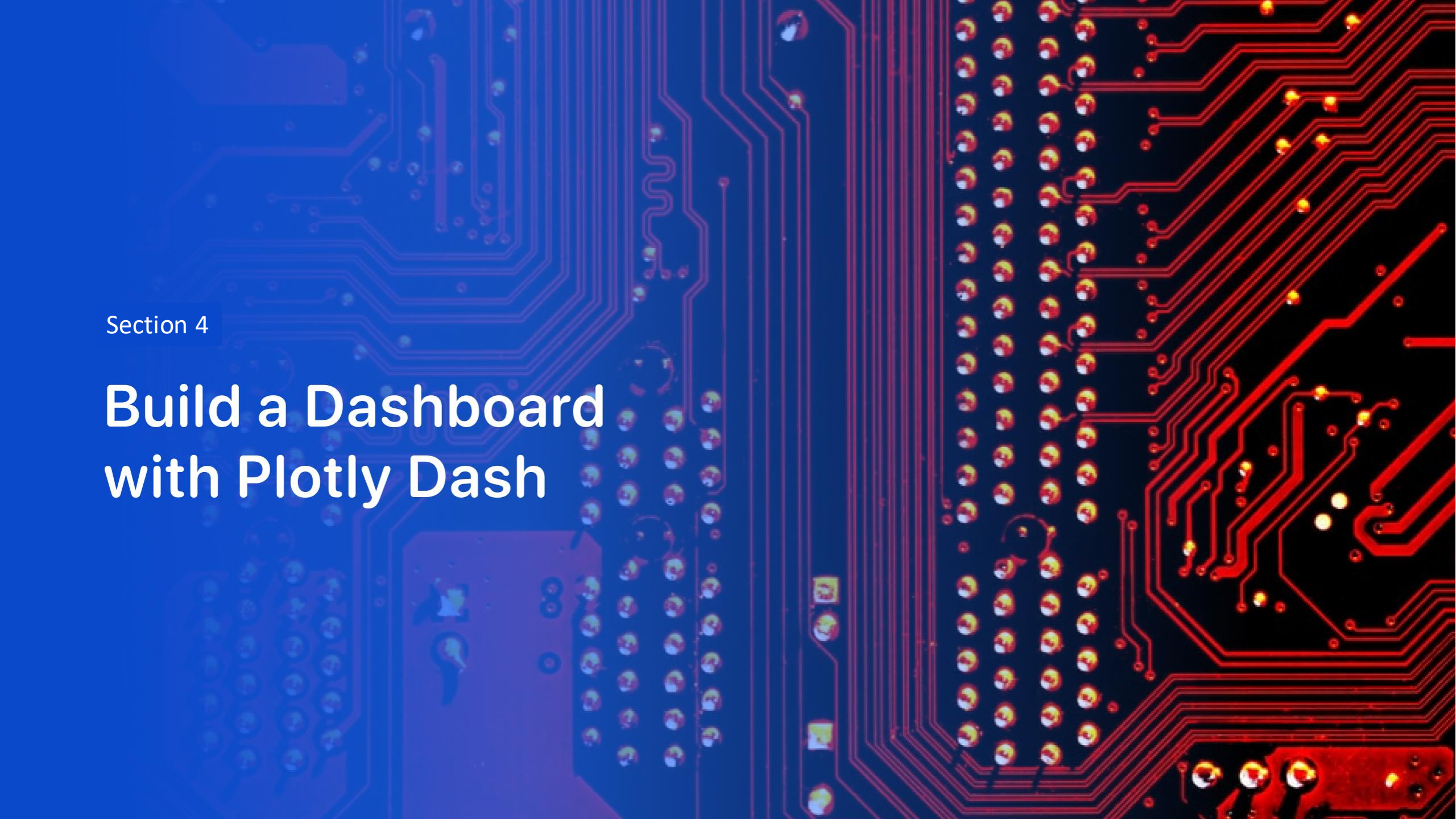
SpaceX Launch Outcomes for each Site

- A marker cluster of launch outcomes was added to the maps to show which launch sites have relatively high success rates



Distance from Launch Site to Coastline

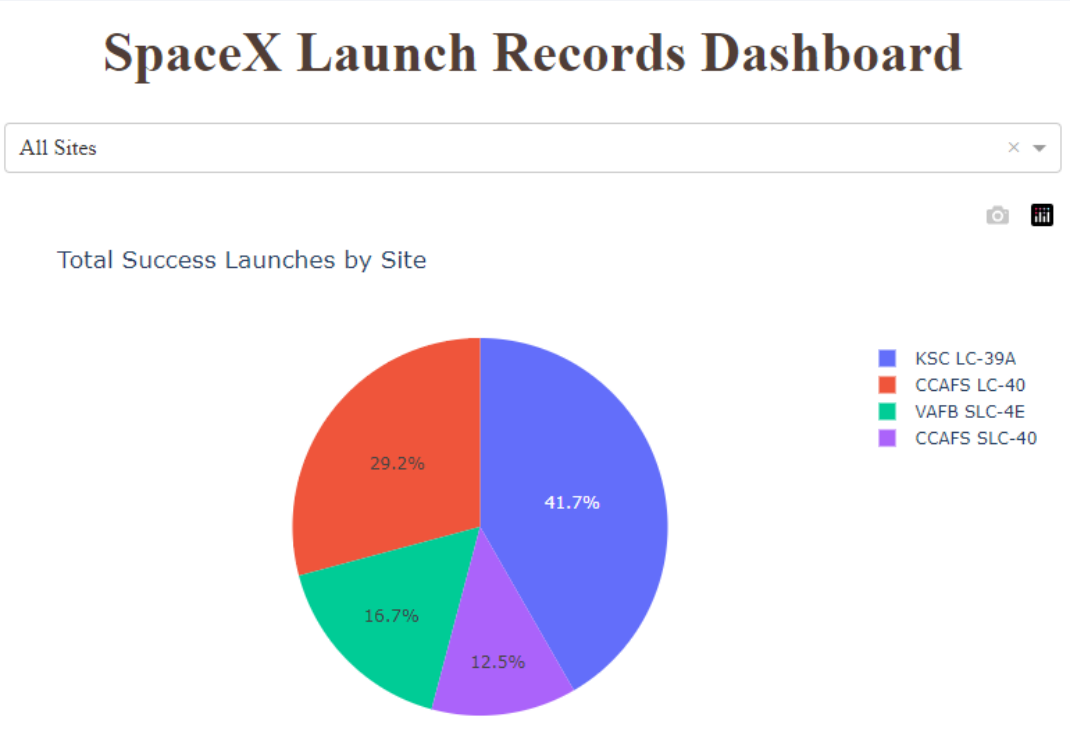
- A line with a label of the distance from the launch site to the coastline was added to the map to analyze the proximities of launch sites to coastline


Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site

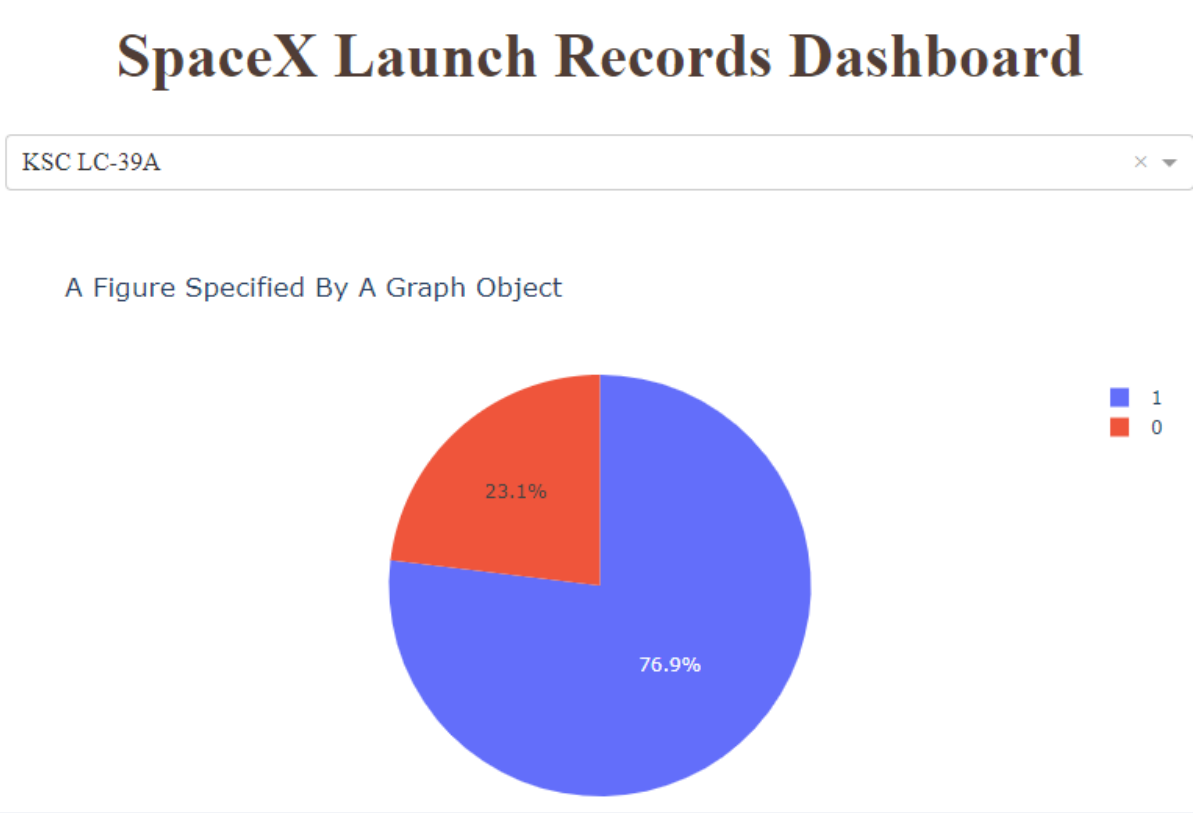
The pie chart of launch success count for all sites shows

- Launch site KSC LC-39A has the highest success rate
- Launch site CCAFS SLC-40 has the lowest success rate



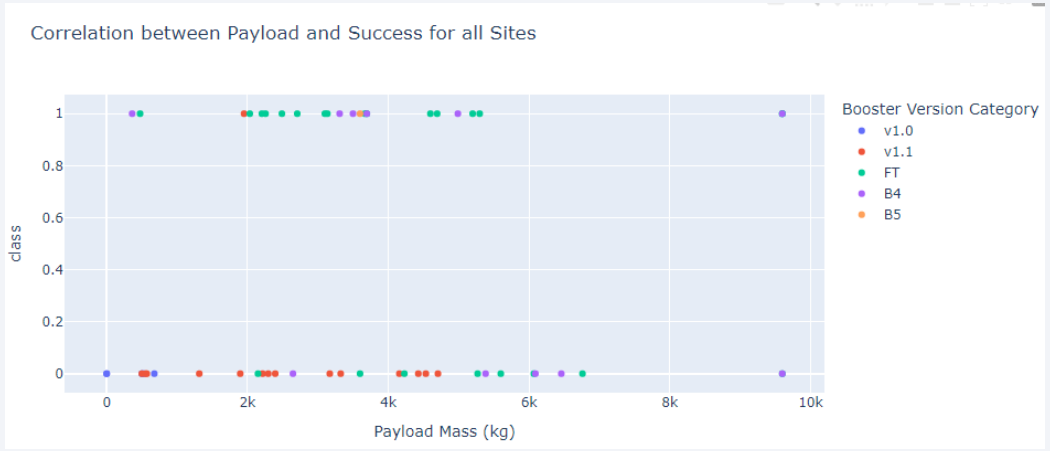
KSC LC-39A Success Rate

The launch site KSC LC-39A has the highest success rate of 76.9%



Correlation between Payload and Success

- The top chart shows the correlation between payload and success for all payload values
- The bottom chart shows the correlation between payload and success for payload values between 2000 and 4000 kg
- The bottom chart shows a high success rate for this range for the booster version category FT



Section 5

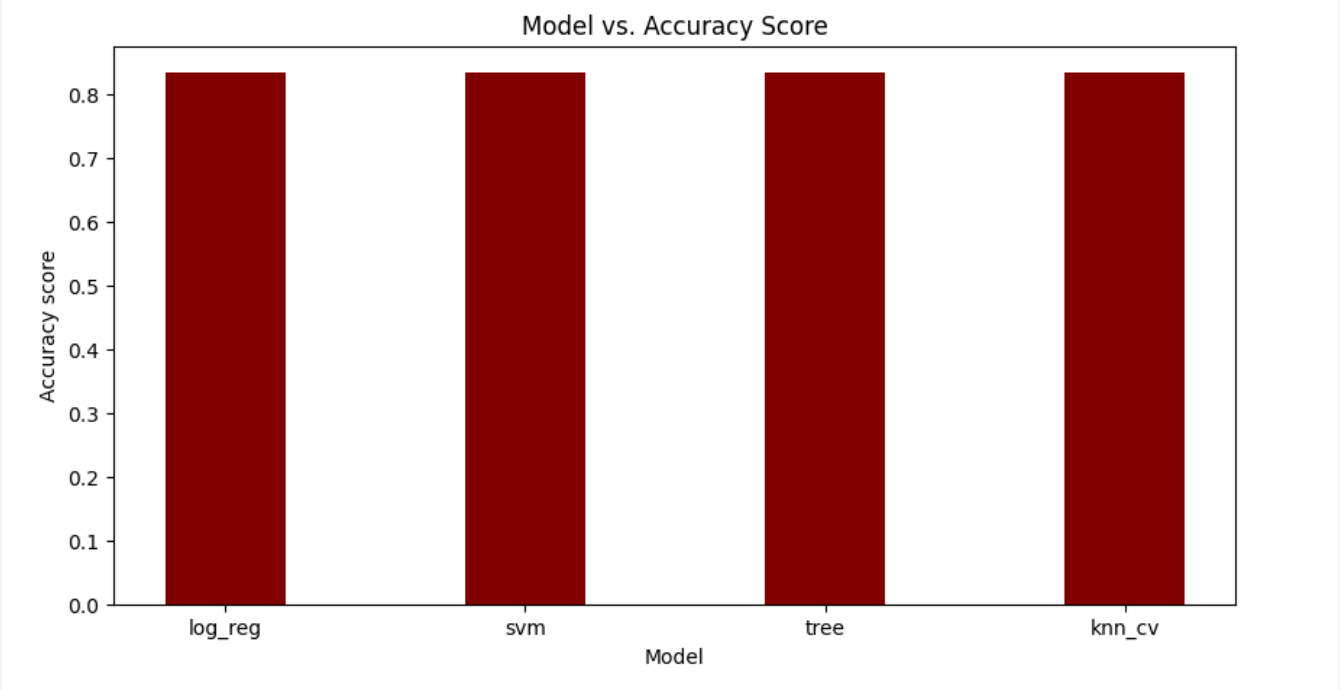
Predictive Analysis (Classification)

Classification Accuracy

All models exhibited the same accuracy score of 0.8333

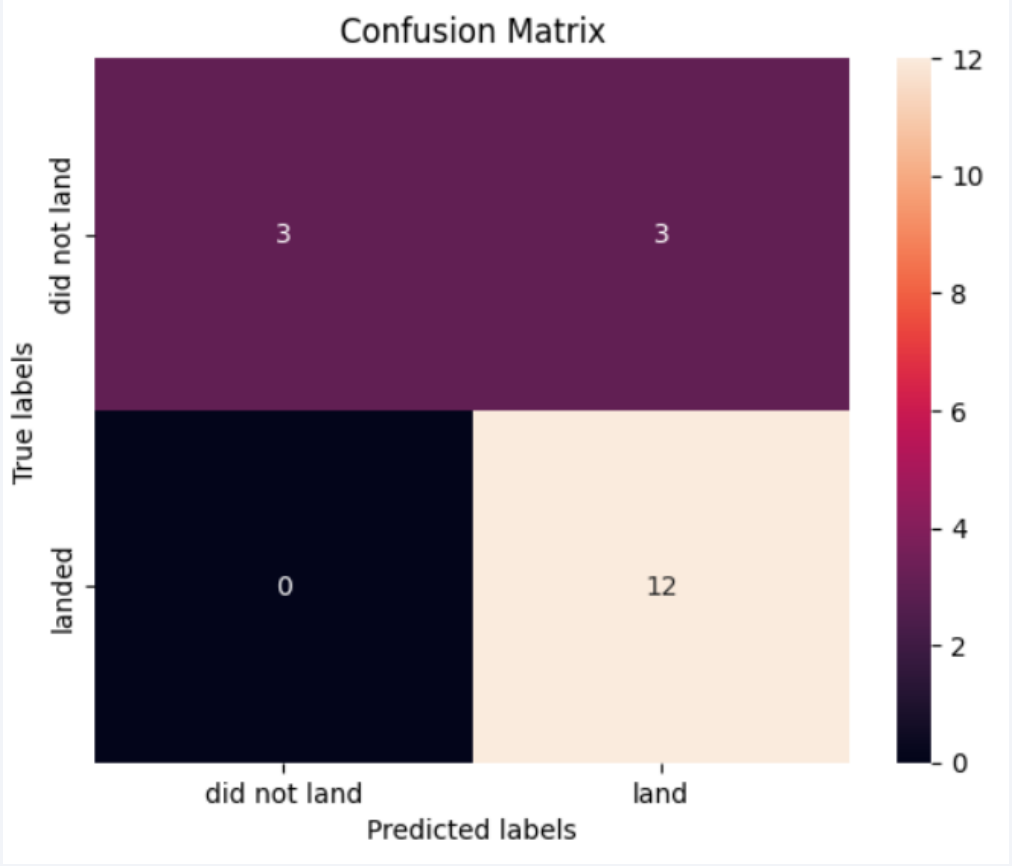
```
knn_cv.score(X_test, Y_test)
```

```
0.8333333333333334
```



Confusion Matrix

All models created the same confusion matrix with 12 true detects and 3 false positives.



Conclusions

There were some trends in the data that showed how different variables would affect the probability of a successful launch.

There were also trends that showed how one or more variables affected the result depending on the category of another variable.

Four models were developed with an accuracy score of 83.33% that can be used to predict weather a launch will have a successful outcome or not.

Appendix

Data collection preliminary results

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details	crew	ships	capsules	payloads	launchpad	flight_number
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	Engine failure at 33 seconds and loss of vehicle	[]	[]	[]	[5eb0e4b5b6c3bb0006eeb1e1]	5e9e4502f5090995de566f86	1
1	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 301, 'altitude': 289, 'reason': 'harmonic oscillation leading to premature engine shutdown'}]	Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage	[]	[]	[]	[5eb0e4b6b6c3bb0006eeb1e2]	5e9e4502f5090995de566f86	2

Thank you!