IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cristin Hendrickson
August 20th, 2024

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

Summary of Methodologies

- Data was collected with APIs, auxiliary functions, and webscraping
- Data was wrangled with Exploratory Data Analysis (EDA) to find patterns using visualizations and SQL
- Interactive visual analytics were made using Folium and Plotly Dash
- Predictive analysis was done using classification models

Summary of All Results

- KSC LC-39A has the highest success rate
- Launch sites are in close proximity to railroads, highways, and coastlines but farther away from cities
- Between 2K and 4K payload range has the highest success rate
- The FT booster version has the highest success rate
- All models have the same accuracy in predicting landings

Introduction

Project background and context:

- Space X advertises Falcon 9 rocket launches with a cost of 62 million dollars
- other providers cost upward of 165 million dollars each
- Much of the savings is because Space X can reuse the first stage
- 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 Space X for a rocket launch.

Problems you want to find answers:

- Which launch site has the highest success rate?
- Where are launch sites located?
- Which payload range has the highest success rate?
- Which booster version has the highest success rate?
- Which model has the best accuracy for predicting landings?

Section 1

Methodology

Methodology

Executive Summary

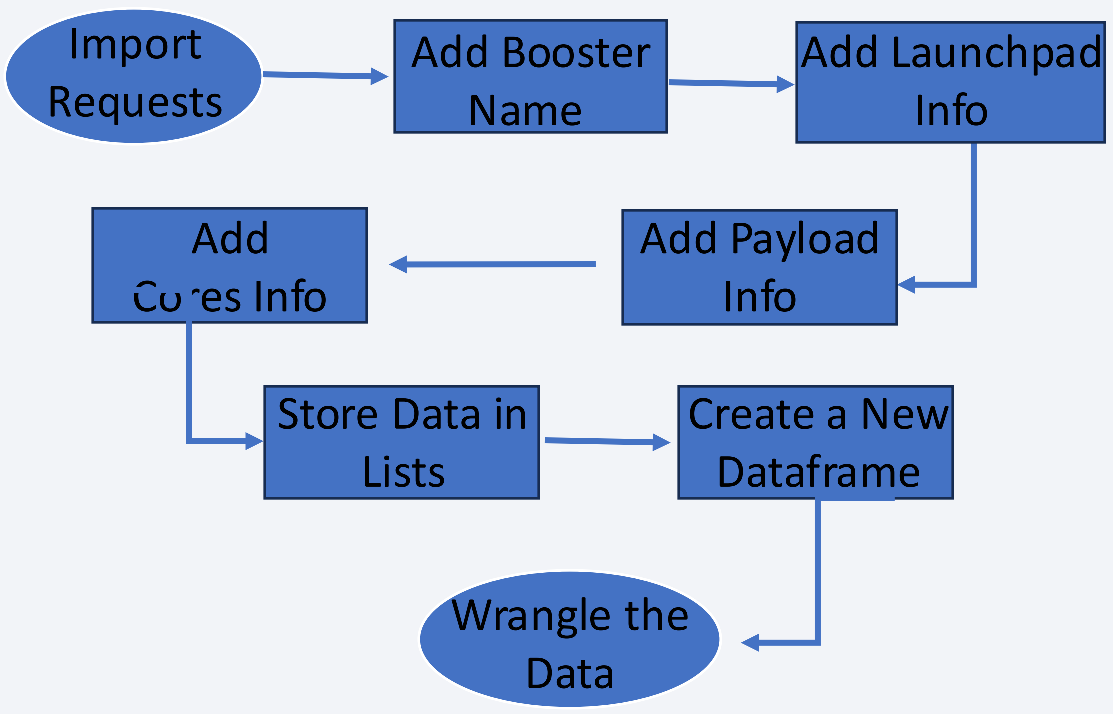
- Data collection methodology:
  - APIs and auxiliary functions were used to collect information along with webscraping
- Perform data wrangling
  - EDA found patterns in the data and determined labels for training supervised models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best hyperparameters for the models and identify the method that performs best

Data Collection

- Define a series of helper, auxiliary functions to use APIs to extract information
- Request and parse launch data using the GET request
- Convert into a Pandas dataframe using .json_normalize
- Use the APIs to get information about the launches using the IDs for each launch
- Create a new dataframe
- Filter the dataframe for Falcon 9 launches and replace missing values
- Request Wiki page from URL using a BeautifulSoup object
- Extract all column and variable names from the HTML table header
- Create a dataframe by parsing the launch HTML tables

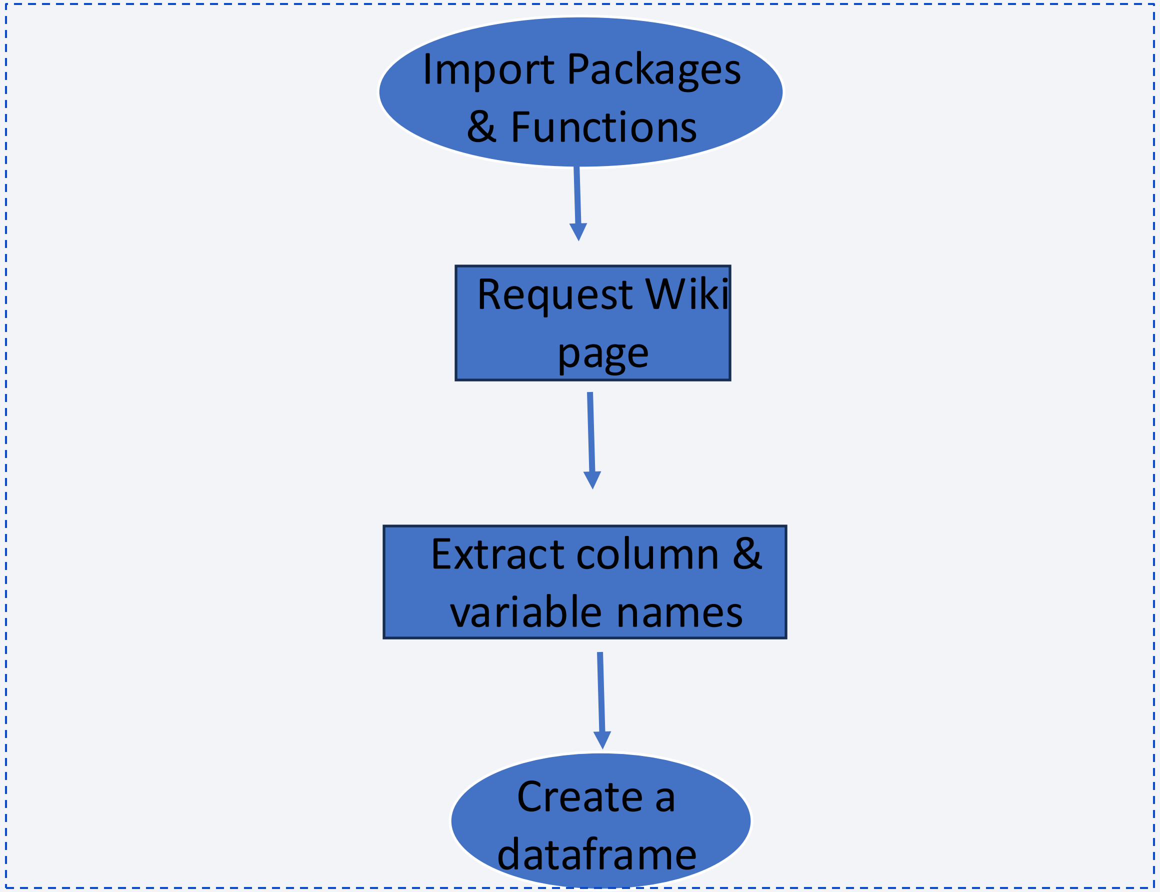
Data Collection – SpaceX API

- Define a series of functions to use APIs to extract information using identification numbers in the launch data
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Collect Falcon 9 records from Wikipedia using BeautifulSoup
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/jupyter-labs-webscraping.ipynb



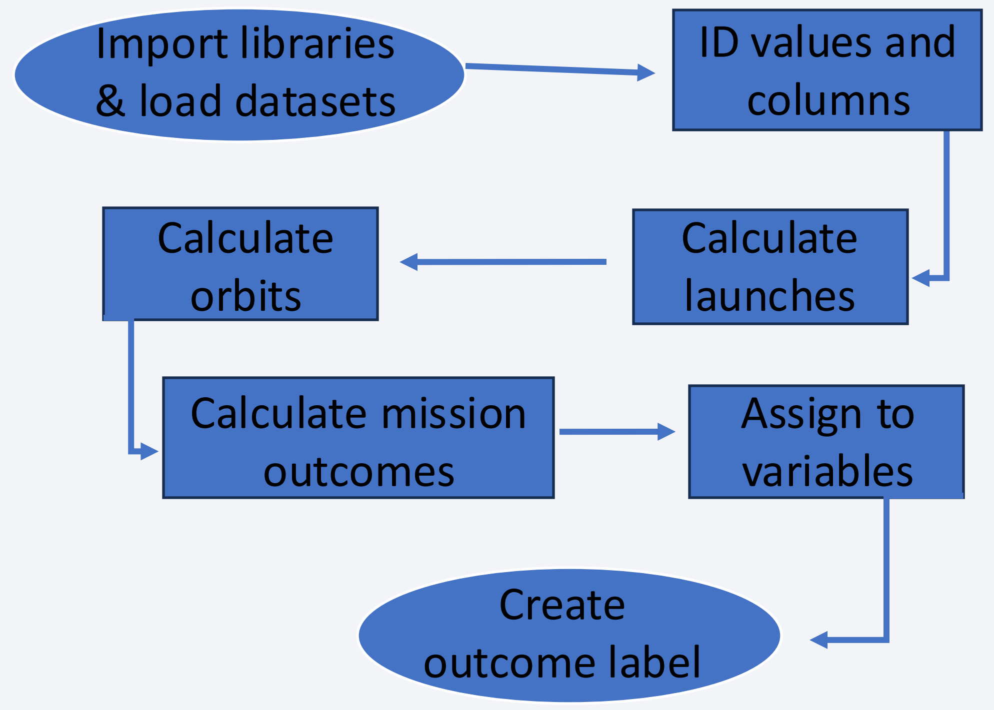
Data Wrangling

- Perform Exploratory Data Analysis to find patterns in the data and determine labels for training supervised models
- Import libraries and load datasets
- Identify missing values and numerical and categorical columns
- Calculate the number of launches at each site using value_counts()
- Calculate the number and occurrence of each orbit using value_counts()
- Calculate the number and occurrence of mission outcomes using value_counts() and assign to variables
- Create a landing outcome label from the previous variables
- Calculate success rate with .mean()

Data Wrangling Continued

- Perform Exploratory Data Analysis to find patterns in the data and determine labels for training supervised models

- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Use charts to obtain insights on each variable's effect on success rate
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/edadataviz.ipynb

- Plot the following graphs:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Success Rate and Orbit Type
  - Flight Number vs. Orbit Type
  - Payload Mass vs. Orbit Type
  - Success Rate over Time

EDA with Data Visualization Continued

- Select the features that will be used in predicting success in future models
- Chosen features include:
  - FlightNumber
  - PayloadMass
  - Orbit
  - LaunchSite
  - Flights
  - GridFins
  - Reused
  - Legs
  - LandingPad
  - Block
  - ReusedCount
  - Serial

EDA with SQL

The following SQL queries were written and executed on the dataset:

- Display the names of the unique launch sites
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was achieved

https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL Continued

Additional SQL queries written and executed on the dataset:

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- 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

Build an Interactive Map with Folium

- Map objects were created to find geographical patterns about launch sites and identify success rates of sites
- Circles were added to show the location of each launch site
- Markers were added for all launch records to show if the launch was successful or failed
- Coastlines, railways, highways and cites were marked on maps and lines were drawn to show distances between them and launch sites
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

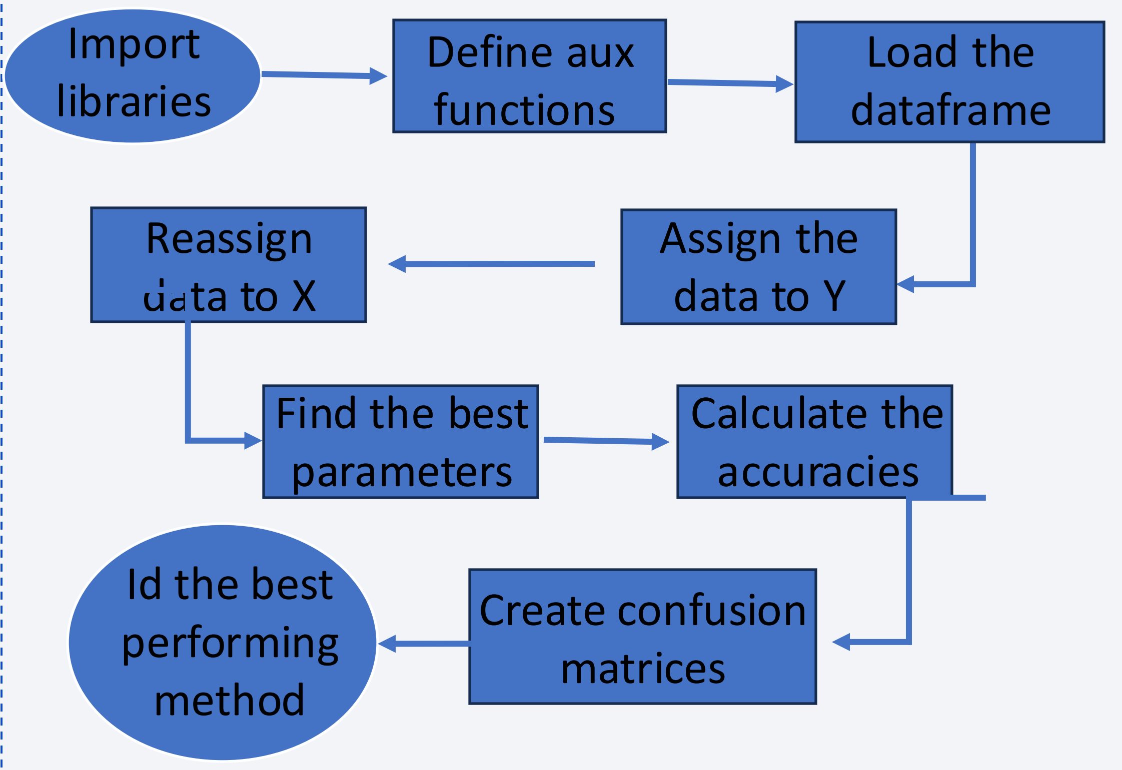
- Perform interactive visual analytics on launch data in real time
- Dispaly a pie chart of successful launches and a scatter chart to show the correlation between payload, launch sites and success
- Filter results based on launch site and payload range
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/Dashboard%20Application%20with%20Plotly%20Dash

Predictive Analysis (Classification)

- Import libraries, define auxiliary functions, and load the dataframe
- Use to_numpy() to assign the data to the variable Y
- Use a transform to standardize the data and resassign it to the variable X
- Use the function train_test_split to split the data X and Y into training and test sets
- Fit a GridSearchCV object to a logistic regression object to find the best parameters
- Repeat for support vector machine, decision tree classifier, and K nearest neighbor
- Calculate the accuracy on the test data using the method score and create a confusion matrix
- Identify the method that performs best

Predictive Analysis (Classification) Continued

- Perform exploratory data analysis and determine training labels
- Find best hyperparameter for SVM, Classification Trees, KNN and Logistic Regression
- Identify the method that performs best using test data
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

Exploratory Data Analysis Results

Interactive Analytics Demo in Screenshots

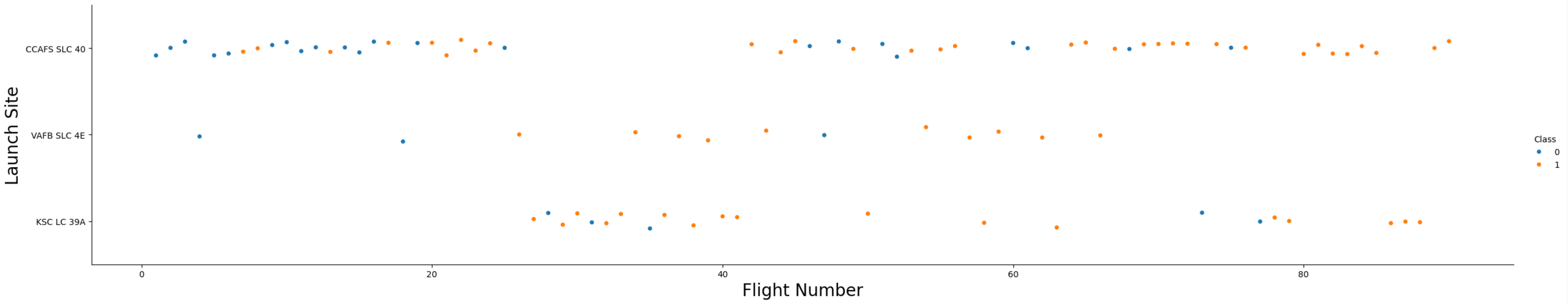
Predictive Analysis Results

Section 2

Insights drawn from EDA

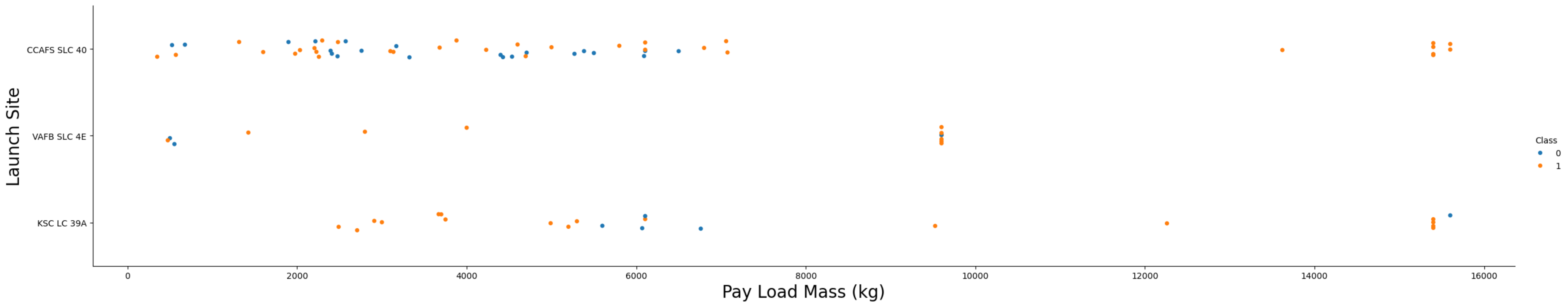
Flight Number vs. Launch Site

- CCAFS SLC launch site has a higher success rate at flight numbers greater than 80
- VAFB SLC launch site has most of its flight numbers between 20 and 60
- KSC LC launch site has all flight number above 20 and mixed success throughout



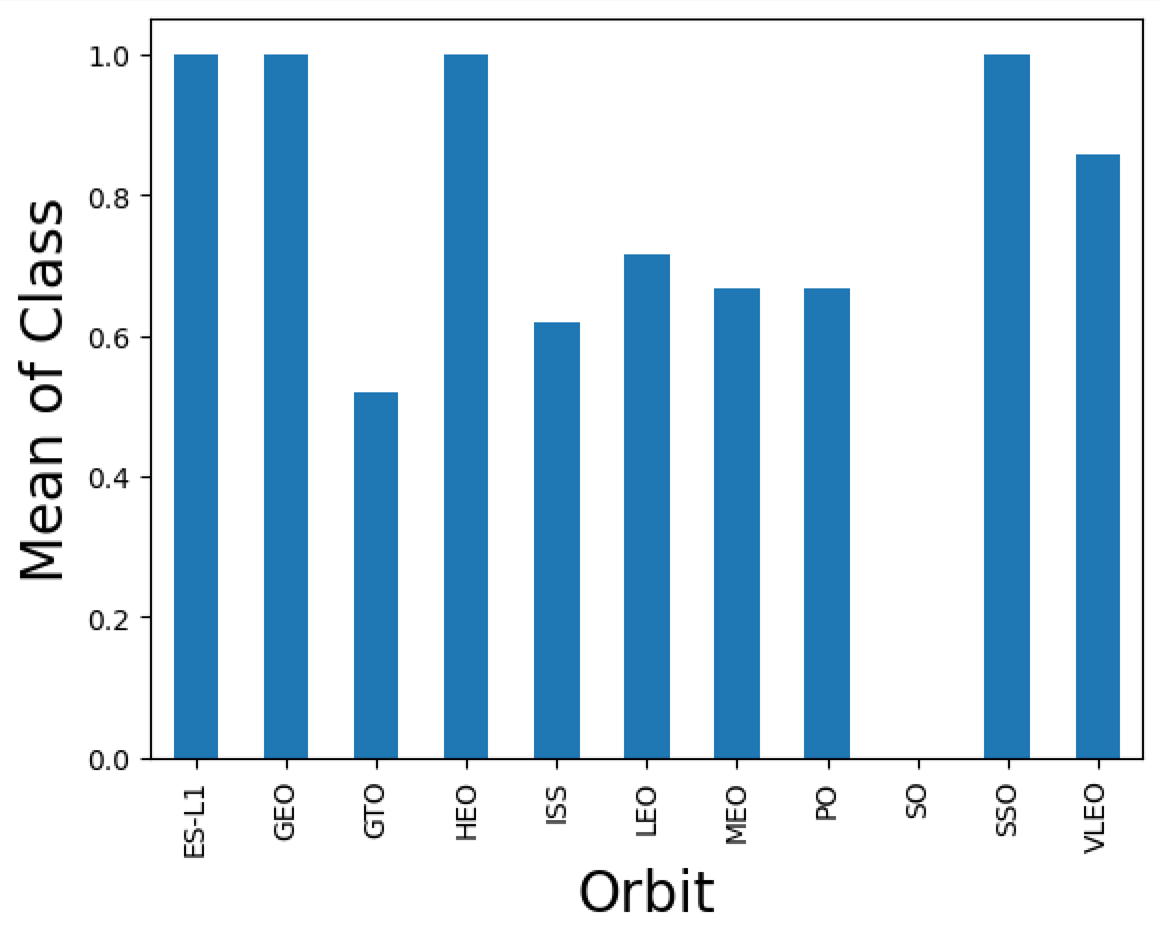
Payload vs. Launch Site

- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000)



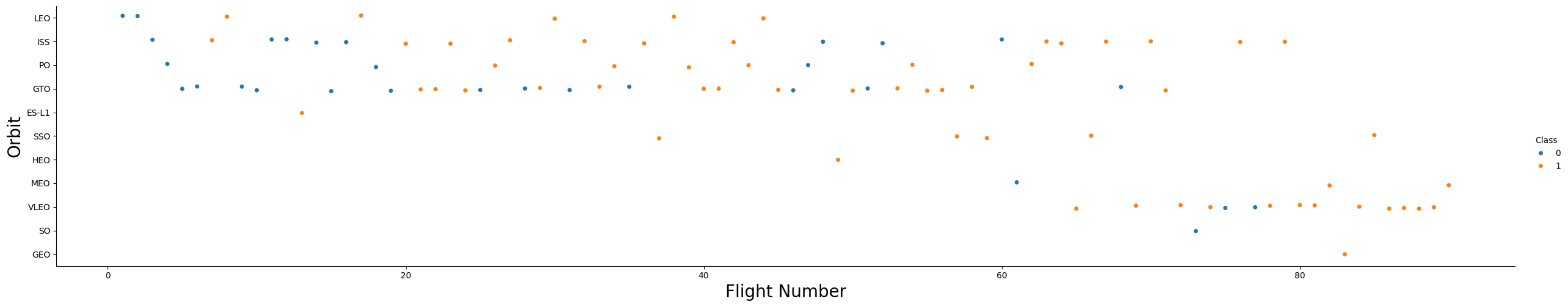
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have the highest success rates
- VLEO also has a high success rate
- GTO has the lowest success rate



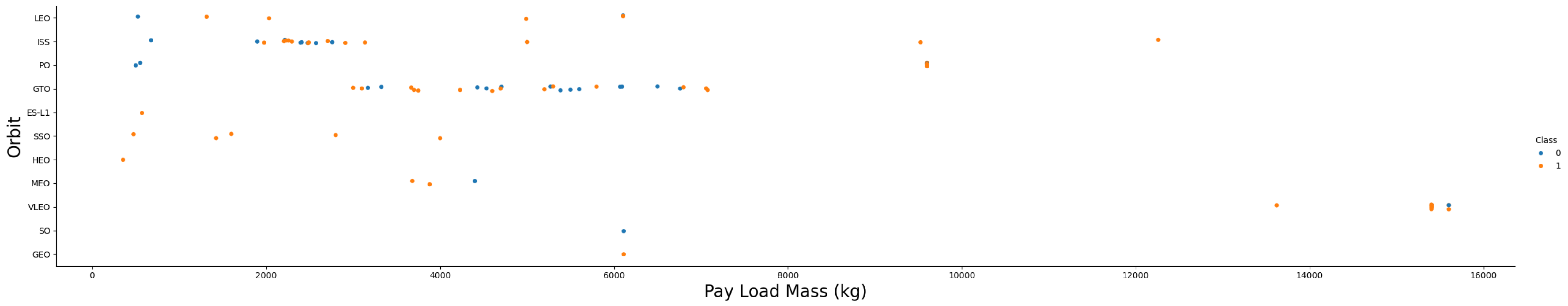
Flight Number vs. Orbit Type

- In the LEO orbit, success seems to be related to the number of flights
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success



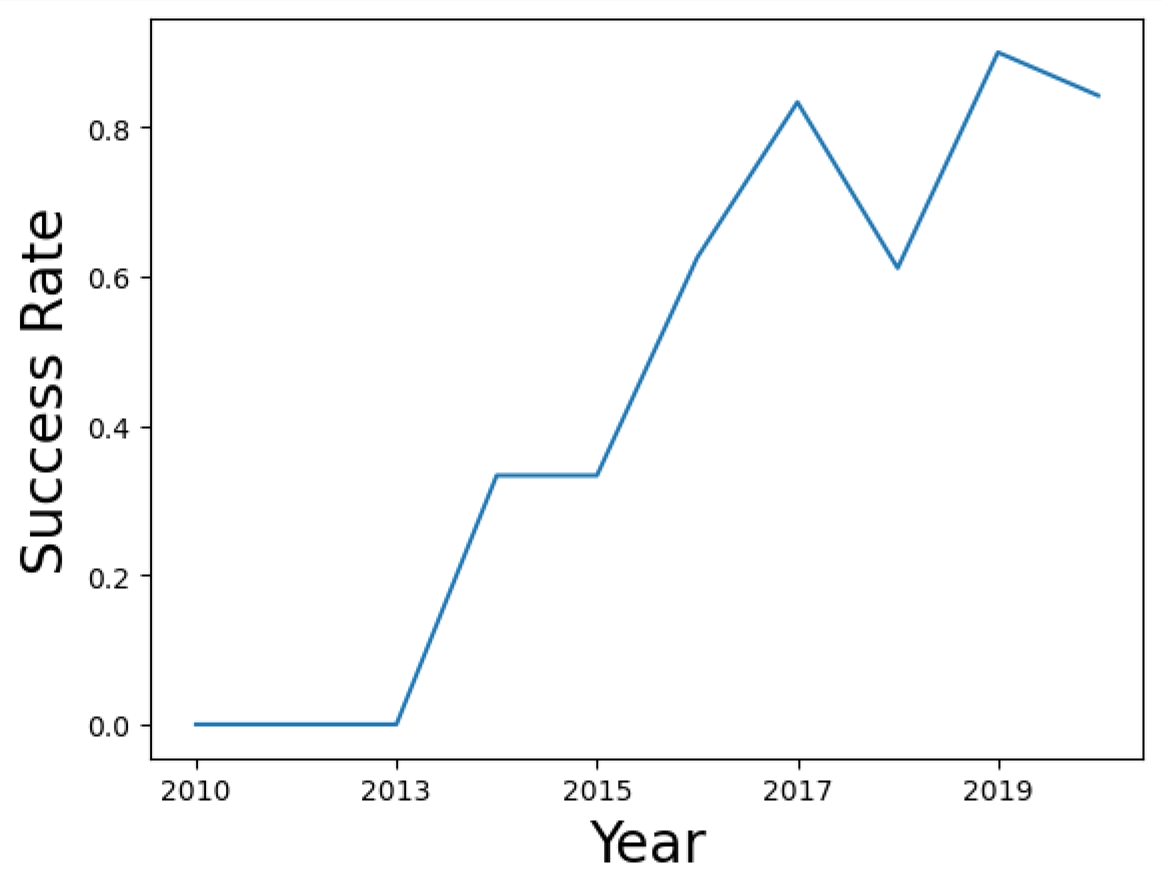
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rates are more for Polar, LEO and ISS
- For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present



Launch Success Yearly Trend

- The success rate since 2013 kept increasing until 2020



All Launch Site Names

- The names of the launch sites are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40
- The query produced the following results:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` are shown below

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_O
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (pa
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 (pa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No

Total Payload Mass

- The total payload carried by boosters from NASA is 48213 KG
- The query produced the following results:

SUM(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.667 KG
- The query produced the following results:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing outcome on a ground pad is 2015-12-22
- The query produced the following results:

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2
- The query produced the following results:

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

Total Number of Successful and Failure Mission Outcomes

- There were 100 successful outcomes and one failure
- The query produced the following results:

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

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass are 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, and F9 B5 B1049.7
- The query produced the following results:

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for the year 2015 are shown below:

MONTH_IN_2015	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

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

- No attempt is the most common landing outcome with success (drone ship) and failure (drone ship) as the runner ups
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order is shown here

Landing_Outcome	COUNT("Landing_Outcome")
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

Section 3

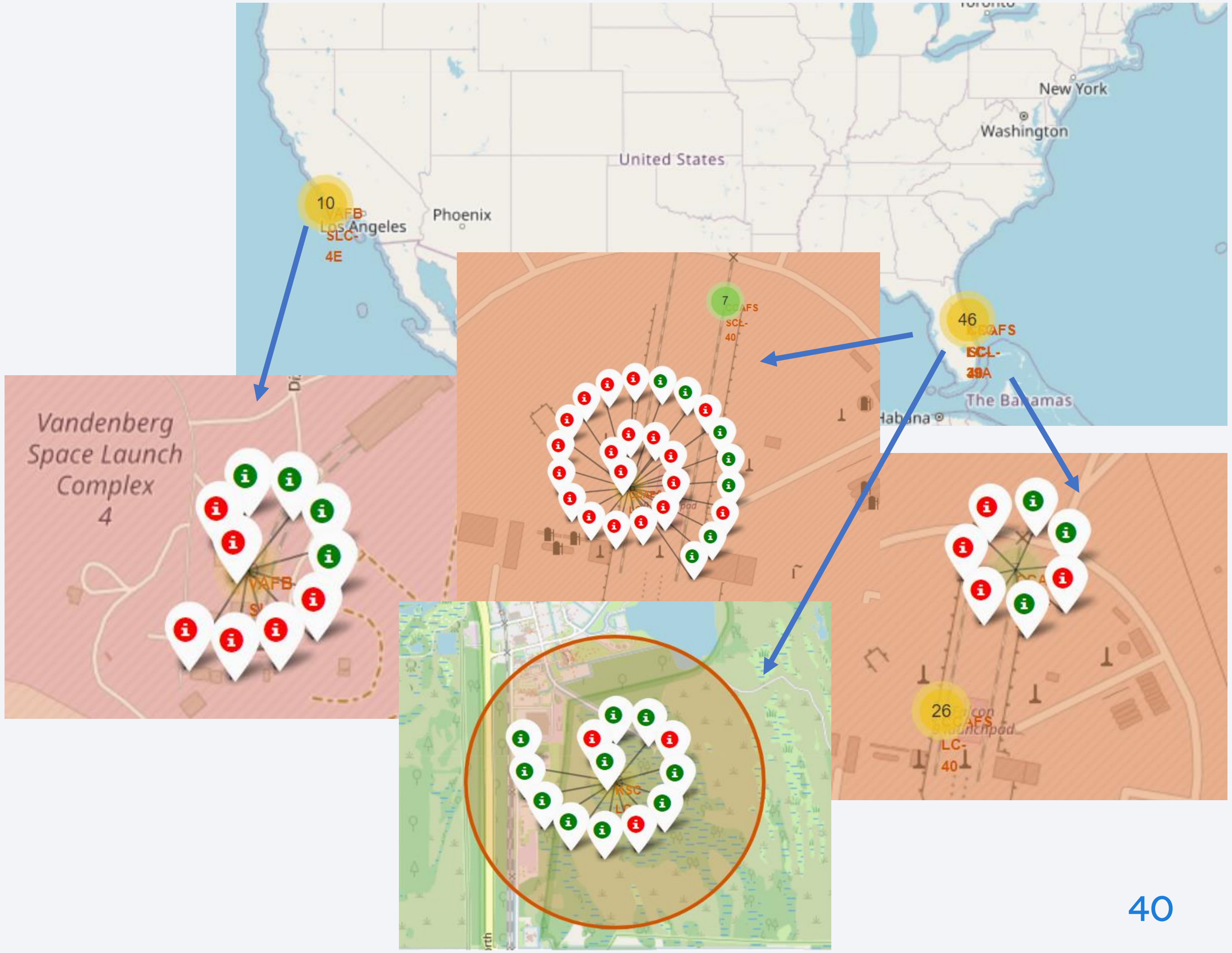
Launch Sites Proximities Analysis

Launch Site Locations

- There are four launch sites
- One launch sites is located on the coast of California
- Three launch sites are located on the coast of Florida

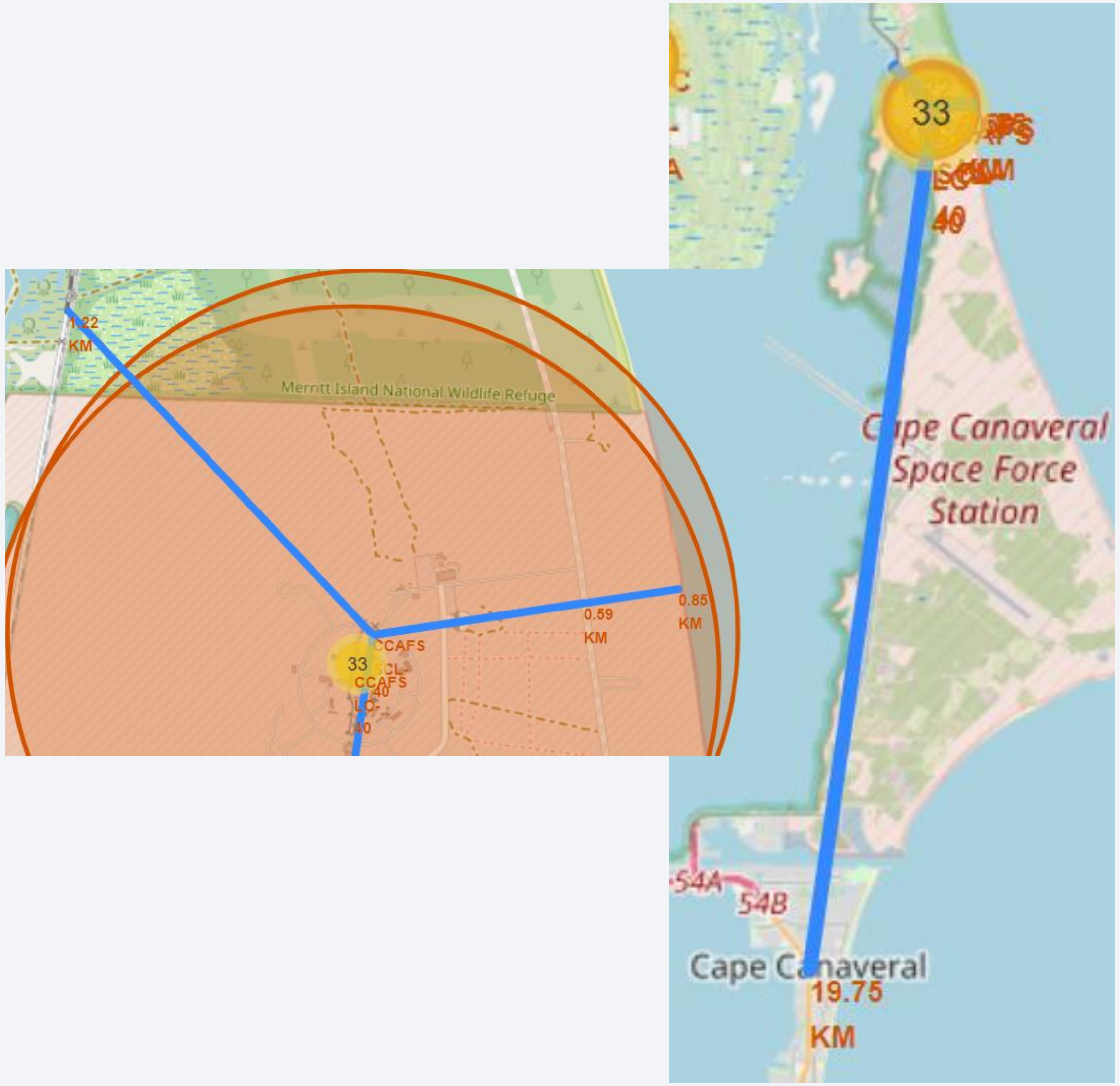
Launch Record Markers

- Green markers show successful launches while red markers show failed launches
- Most launches take place in Florida
- KSC LC-39A has the highest success rate while CCAFS LC-40 has the lowest success rate



Launch Site Proximities

- Launch sites are in close proximity to railroads, highways, and coastlines
- CCAFS SCL-40 is 1.22 km away from a railway and less than 1 km away from a highway and coastline
- Launch sites are farther away from cities
- Cape Canaveral is 19.75 km away from CCAFS SCL-40

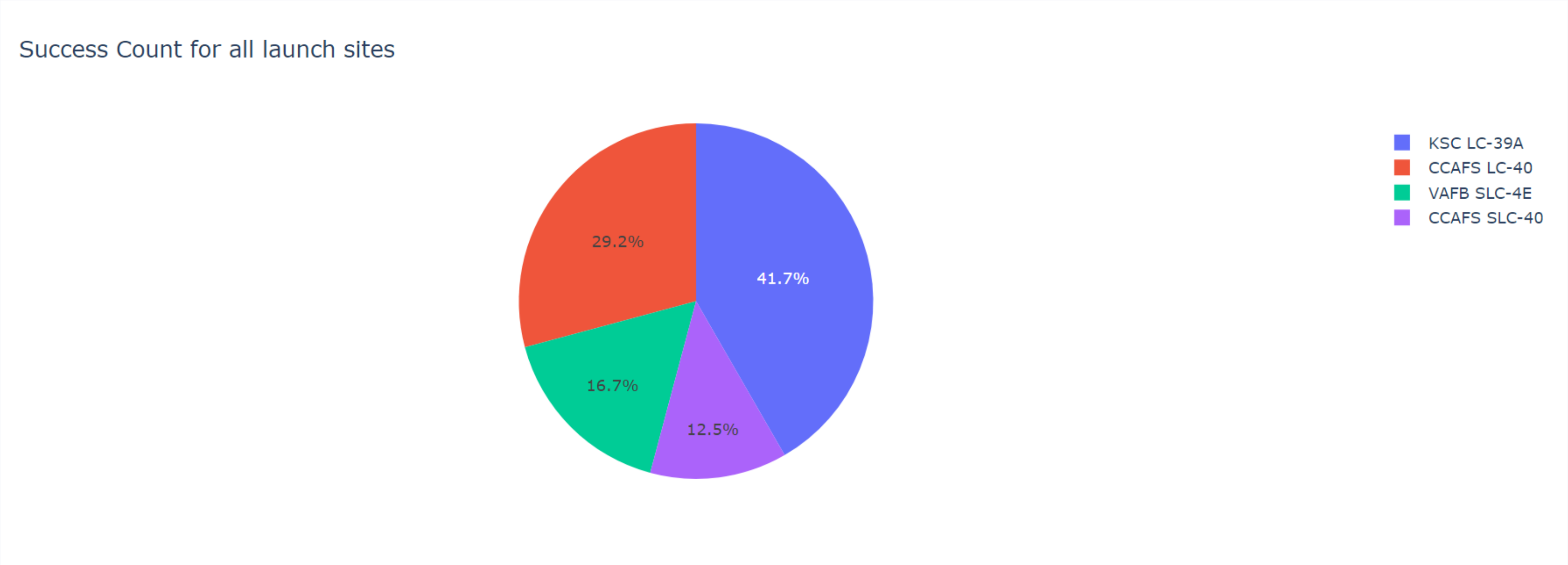


Section 4

Build a Dashboard with Plotly Dash

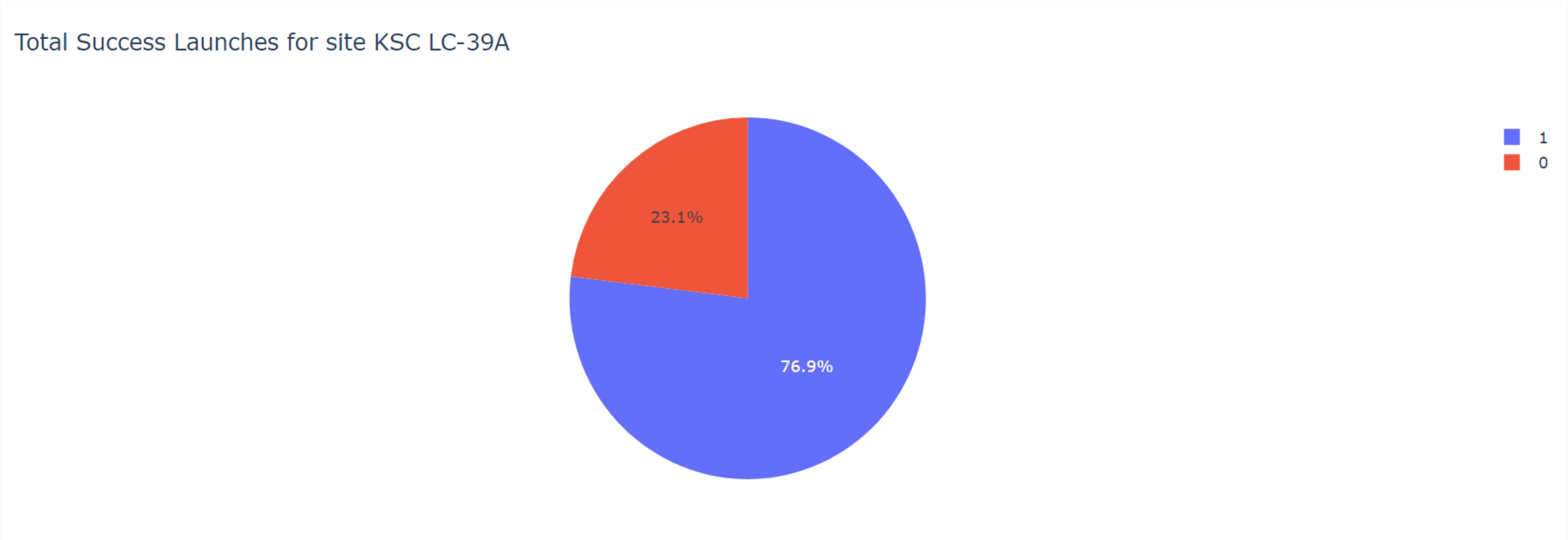
Launch Success Count for All Sites

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



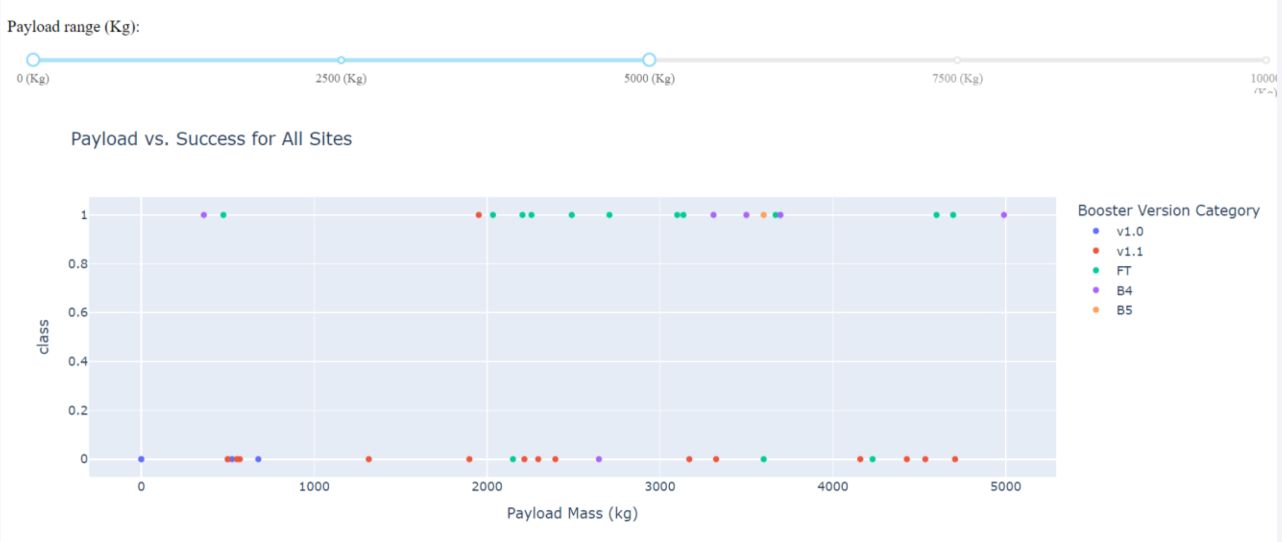
Highest Launch Success Ratio

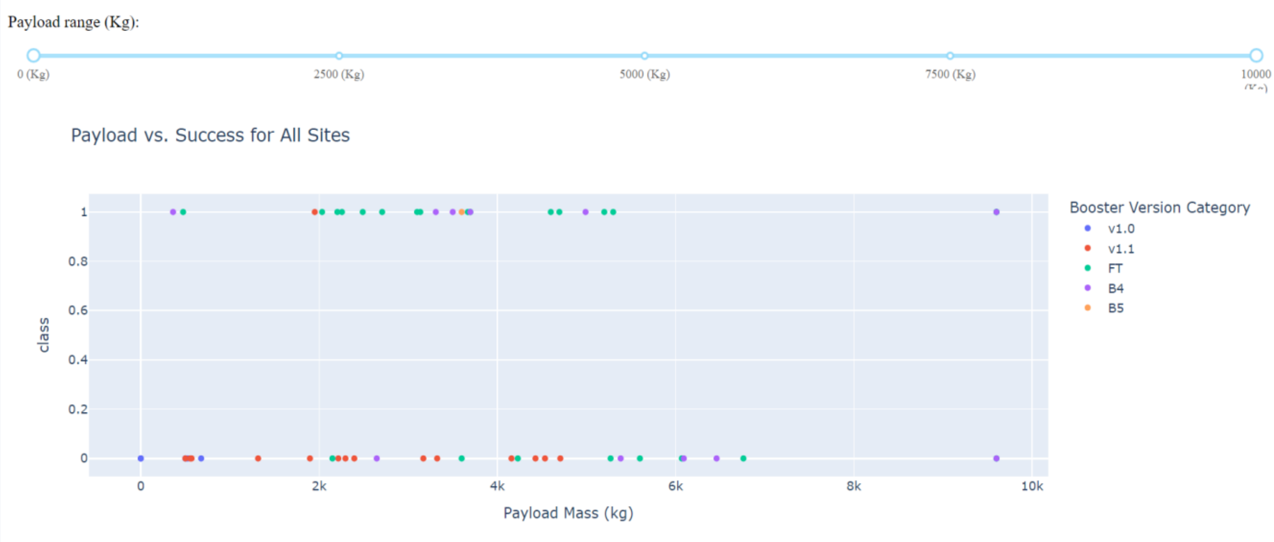
- KSC LC-39A has the highest launch success ratio
- 76.9% of launches were successful
- 26.1% of launches were failures



Payload vs. Launch Outcomes

- Between 2K and 4K payload range has the highest success rate
- Between 6K and 8K payload range has the lowest success rate
- The FT booster version has the highest success rate



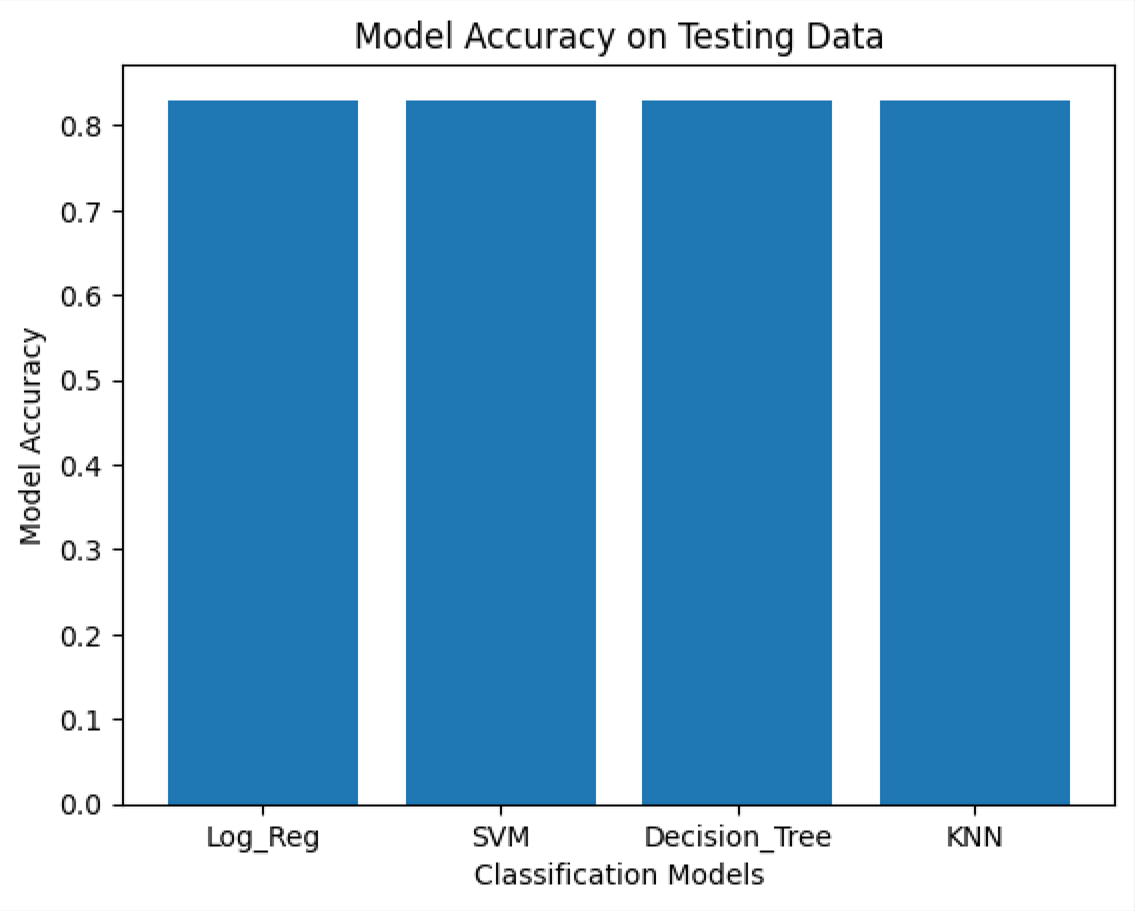


Section 5

Predictive Analysis (Classification)

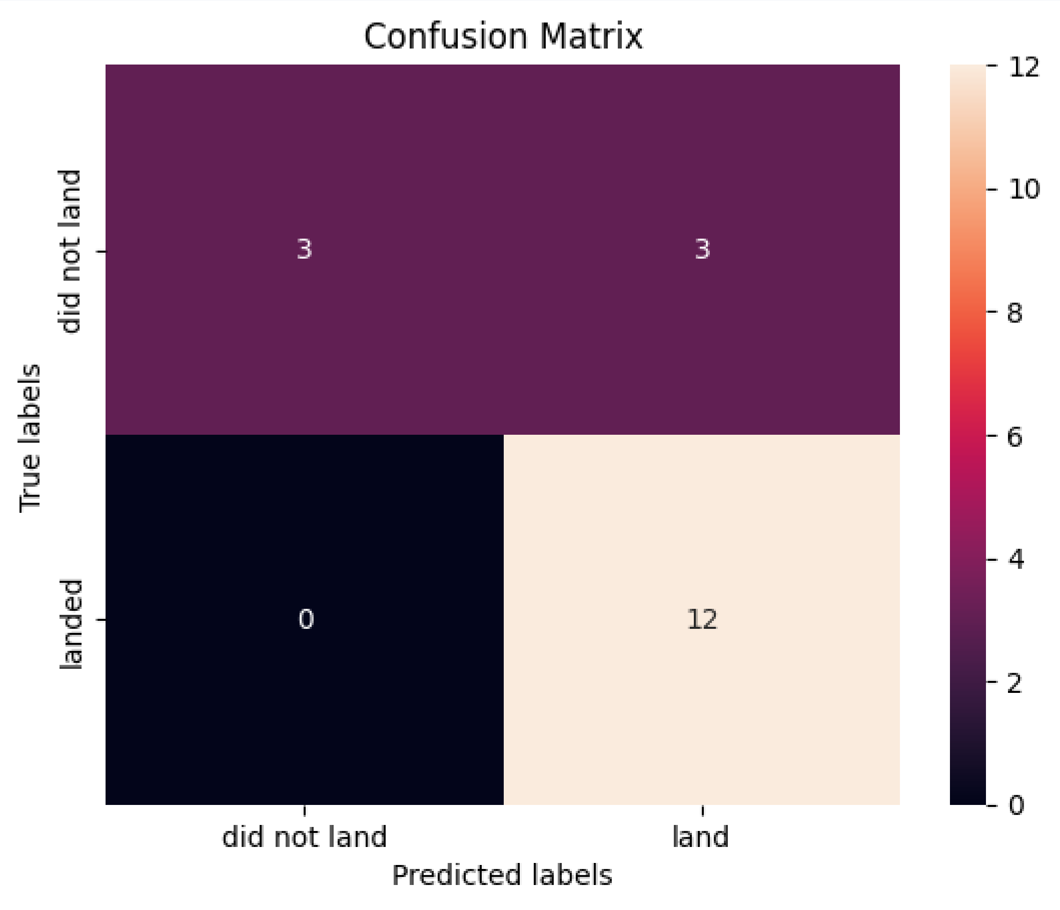
Classification Accuracy

- All models have the same accuracy



Confusion Matrix

- The confusion matrix is the same for all models
- The model can distinguish between different classes but there is a problem with false positives
- True Postive - 12 (True label is landed, Predicted label is also landed)
- False Postive - 3 (True label is not landed, Predicted label is landed)



Conclusions

- Of the four launch sites, most launches take place in Florida
- KSC LC-39A has the highest success rate and highest launch success ratio
- Launch sites are in close proximity to railroads, highways, and coastlines but farther away from cities
- Between 2K and 4K payload range has the highest success rate
- The FT booster version has the highest success rate
- Between logistic regression, support vector machine, decision tree, and K nearest neighbor, all models have the same accuracy in predicting landings
- The model can distinguish between different classes but there is a problem with false positives

Appendix

Python code and outputs can be found below:

- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/jupyter-labs-spacex-data-collection-api.ipynb
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/jupyter-labs-webscraping.ipynb
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/labs-jupyter-spacex-Data%20wrangling.ipynb
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/edadataviz.ipynb
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/lab_jupyter_launch_site_location.ipynb
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/Dashboard%20Application%20with%20Plotly%20Dash
- https://github.com/crhendr/Applied-Data-Science-Capstone/blob/c8a72f1287ec26cdf69fedd869b5697a4a0162e6/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!