IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

S. Ayaan Danish
11th Sept 2023

Outline

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

Executive Summary - Methodologies

Collected Launch Data

- Using SpaceX API
- By web scraping Wikipedia records

Cleaning and Prepared the Dataset

- Analyzed the dataset
- Labelled the data by outcome

Performed EDA on the Dataset

- Visualized the data
- Queried the data with SQL
- Mapped launch site locations

Predicted landing outcomes using ML

- Split and standardized the data
- Trained multiple models
- Compared each model's results

Executive Summary - Results

Logistic Regression Classifier	Accuracy: 84.6%	Precision: 80.0%	Recall: 100.0%	F1-Score: 88.0%
SVM Classifier	Accuracy: 84.8%	Precision: 80.0%	Recall: 100.0%	F1-Score: 88.0%
Decision Tree Classifier	Accuracy: 87.5%	Precision: 80.0%	Recall: 100.0%	F1-Score: 88.0%
KNN Classifier	84.8%	Precision: 80.0%	Recall: 100.0%	F1-Score: 88.0%

Introduction

- SpaceX rocket launches are substantially cheaper than other providers.
- They advertise a launch at $62 million, compared to the rest at over $165 million each.
- The reason for this low cost is that SpaceX can reuse their rockets (by landing the first stage).

Previous launch data + Machine learning models = Predicting successful landings

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Making API calls to the SpaceX data repository
  - Web scraping Wikipedia for launch records
- Perform data wrangling
  - Removed null values in the data by replacing them with mean
  - Assigned class labels for successful and unsuccessful landings
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Define our models, perform Grid Search to find best accuracy, and compare results

Data Collection – SpaceX API

- **Make API Call for data** → Extract required features → Store in pandas dataframe → Reformat the data
- **Filter out** → Non-Falcon 9 launches → Multiple payload launches → Multiple core rockets
- **Clean the data** → Check for null values → Calculating mean for missing feature → Replacing null values with mean

Notebook link: Here

Data Collection – Scraping

Request Wikipedia page
- Make request to page and store response
- Create Beautiful Soup object from response
- Verify integrity of Soup

Extract column names
- Find all HTML Tables in the page
- Iterate through appropriate table
- Extract column headings from table

Store the data
- Iterate through appropriate table
- Extract data for each column header
- Create Pandas dataframe

Notebook link: [Here]

Data Wrangling

- **Analyzing the data**
  - Check proportion of null values
  - Check data types for features
- **Calculate insights about the data**
  - Frequency of launches from each site
  - Frequency of launches for each orbit
  - Frequency of each landing outcome
- **Create prediction label**
  - Define positive and negative landing outcomes
  - Create Boolean feature to represent outcome
  - Check overall launch success rate

Notebook link: Here

EDA with Data Visualization

Charts Plotted

- Categorical Plot – To precisely show the relationship between numerical and categorical features
- Bar Plot – To show the general relationship between numerical and categorical features
- Line Chart – To relate two numerical features, often to show the trend of a feature over time

Features Analyzed

- Payload Mass
- Orbit Type
- Launch Site
- Flight Number
- Landing Success Rate

Notebook link: [Here]

EDA with SQL

- All launch sites, as well as launch sites filtered by name
- Total and Average Payload Mass for a specific Booster Version
- First successful ground landing date
- Successful landings with a certain range of Payload Mass
- Count of all mission outcomes
- List of Boosters that carried a specific Payload Mass
- Failed launch records for a specific year
- Ranking frequency of Landing Outcomes within a certain date range

Notebook link: [Here]

Build an Interactive Map with Folium

Marked SpaceX Launch Site locations on US Map

- Launch Site Name to identify the site
- Circle around Site premises to identify bounds

Marked successful and unsuccessful launches on Launch Site Map

- Launch Site Name to identify the site
- Circle around Site to identify bounds
- Marker Cluster at each Launch Site to group the markers for each launch
- Icon for each launch, colored green or red to indicate successful or unsuccessful launches respectively

Calculated and marked the distance between Launch Site to proximities

- Launch Site Name to identify the site
- Circle around Site to identify bounds
- Line from Launch Site to nearest coastline
- Label to indicate distance between Launch Site and coastline

Notebook link: Here

Build a Dashboard with Plotly Dash

Inputs	Outputs

• Dropdown Menu
  • For selecting the Launch Site to analyze, or to choose All Sites
• Range Slider
  • For selecting the range of Payload Masses to display scatter plot for
	• Pie Chart
	  • For viewing success rate of specific Launch Sites
	  • For viewing distribution of successful launches by Launch Site
	• Scatter Plot
	  • For viewing launch outcomes for various ranges of Payload Masses in all Sites
	  • For viewing launch outcomes for various ranges of Payload Mass in specific Sites

Notebook link: Here

Predictive Analysis (Classification)

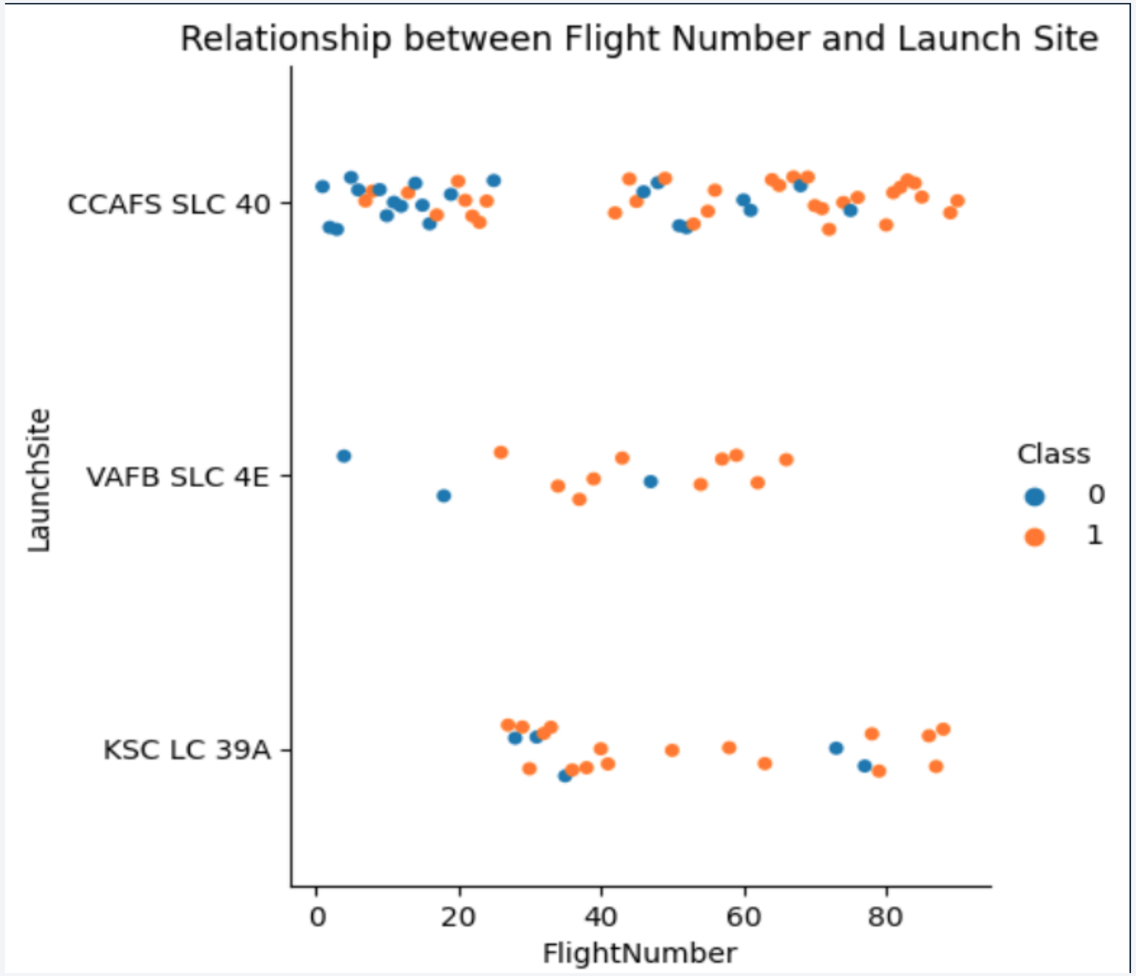
Data Preparation	Load the dataset	Prepare input features	Prepare target features	Create train-test split
Train Models	Fit Logistic Regression	Fit SVM	Fit Decision Tree	Fit KNN
Evaluate the models	Calculate accuracy	Plot confusion matrix	Plot ROC Curve	Decide best model

Section 2

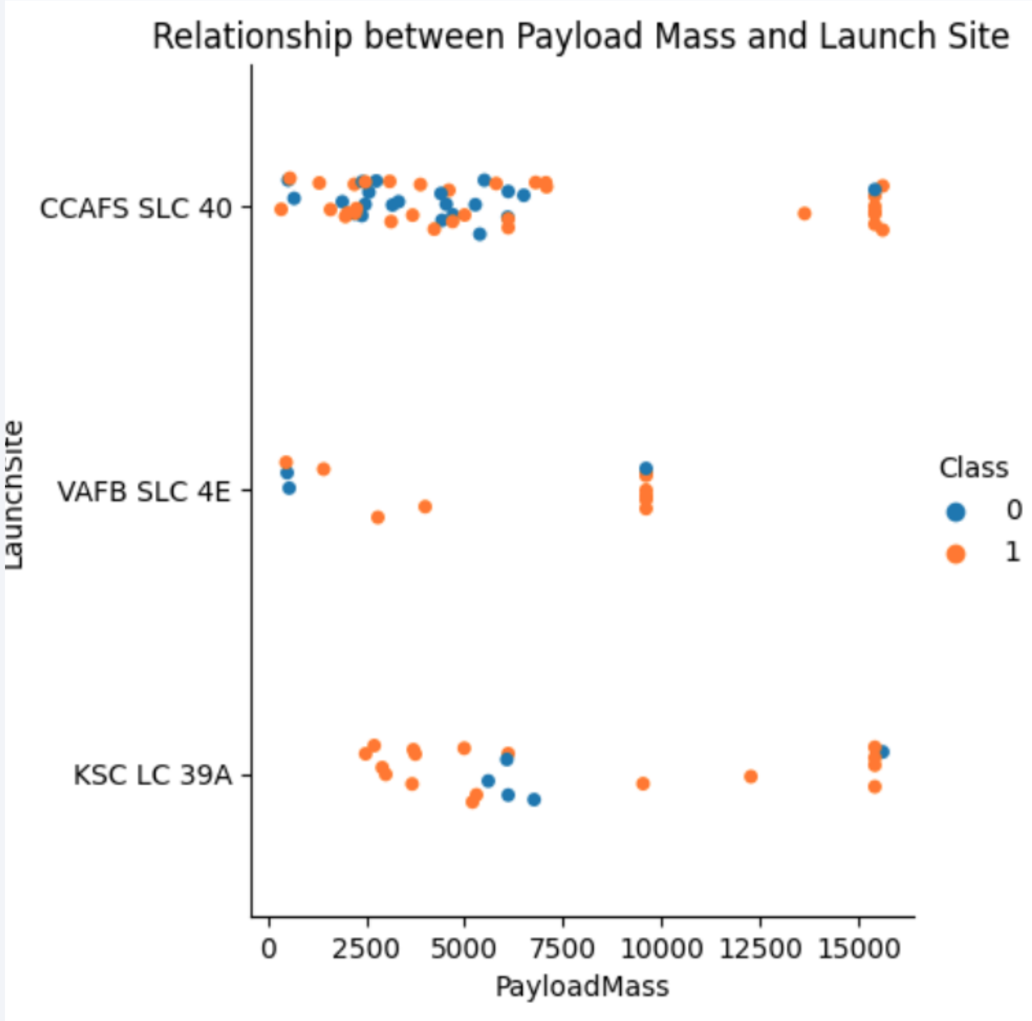
Insights drawn from EDA

Flight Number vs. Launch Site

- Majority launches are from the CCAFS SLC-40 Site
- We can see that the successful landings increase as Flight Number increases
- Indicates SpaceX's improvement over time



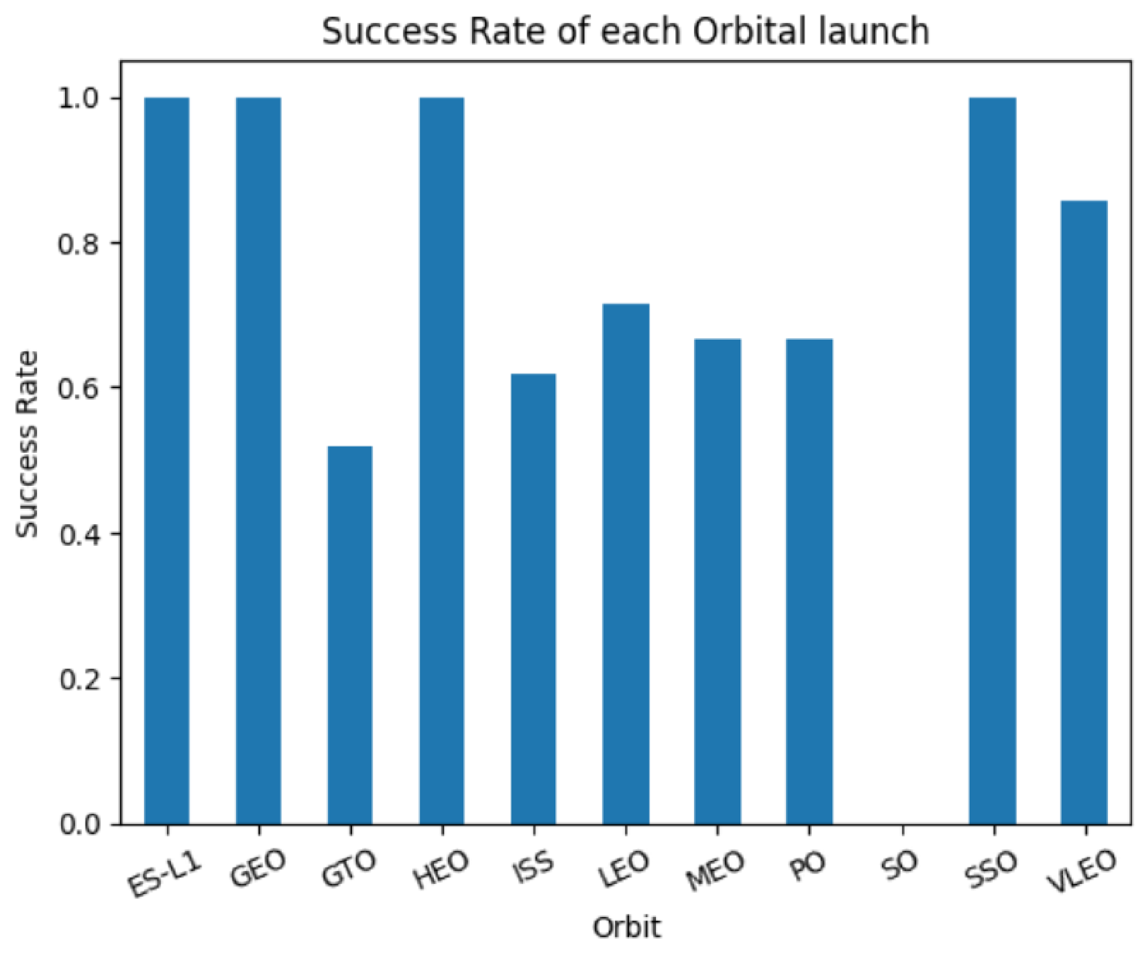
Payload vs. Launch Site



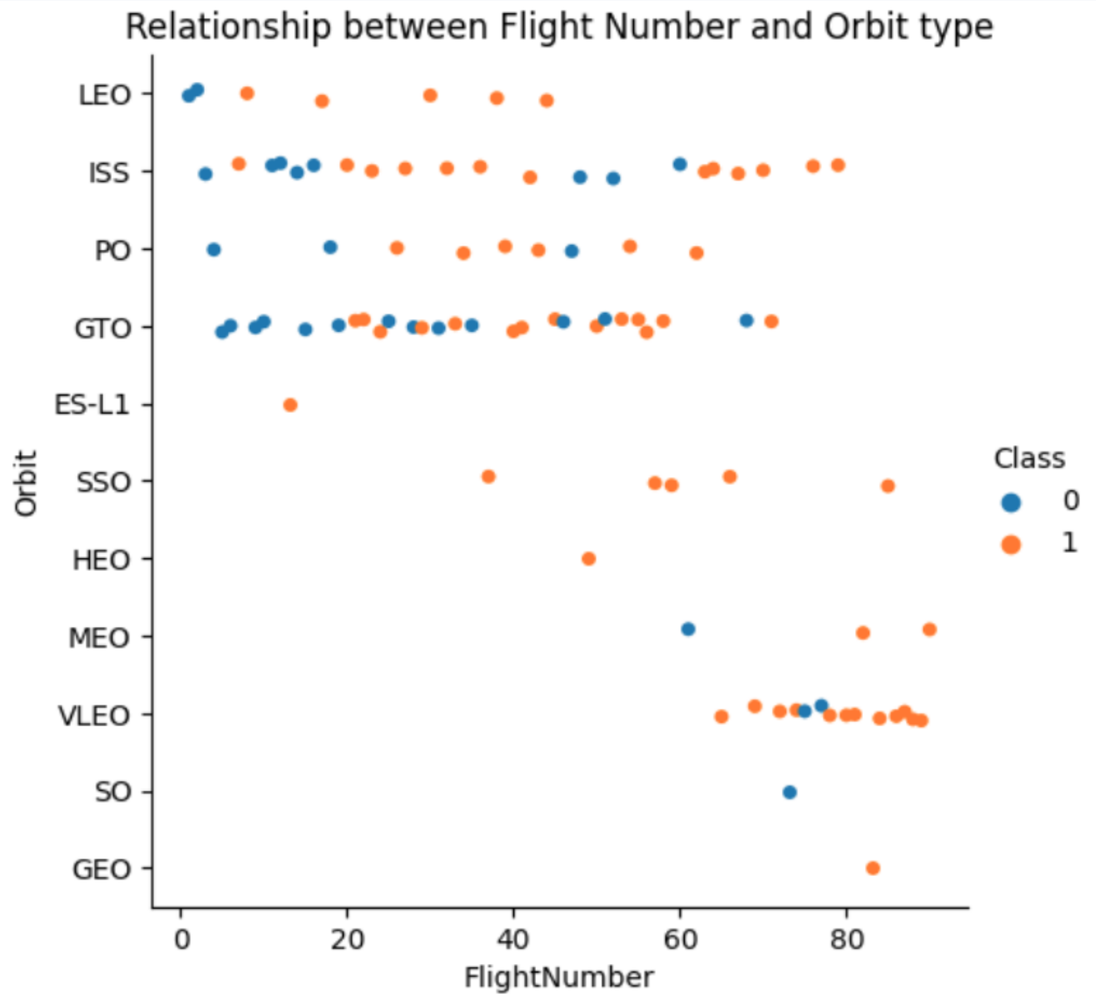
- The VAFB SLC 4E Site does not launch any rockets with payload above 10,000kg
- We can clearly see that successful landings (in blue) decrease sharply for all Launch Sites when Payload Mass increases
- This means lighter payloads are easier to land!

Success Rate vs. Orbit Type

- ES-L1, GEO, and SSO orbit launches have always succeeded
- Conversely, SO orbit has never succeeded
- The lowest success rate after that is of GTO at 50%



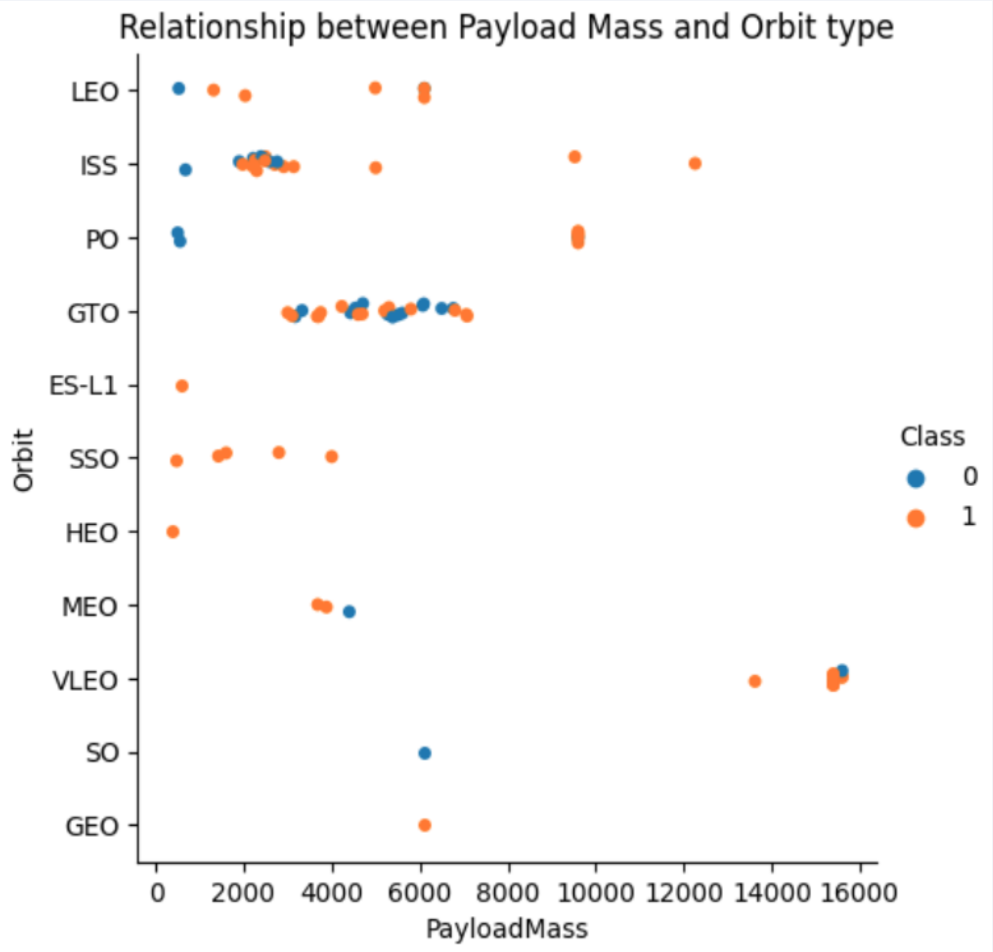
Flight Number vs. Orbit Type



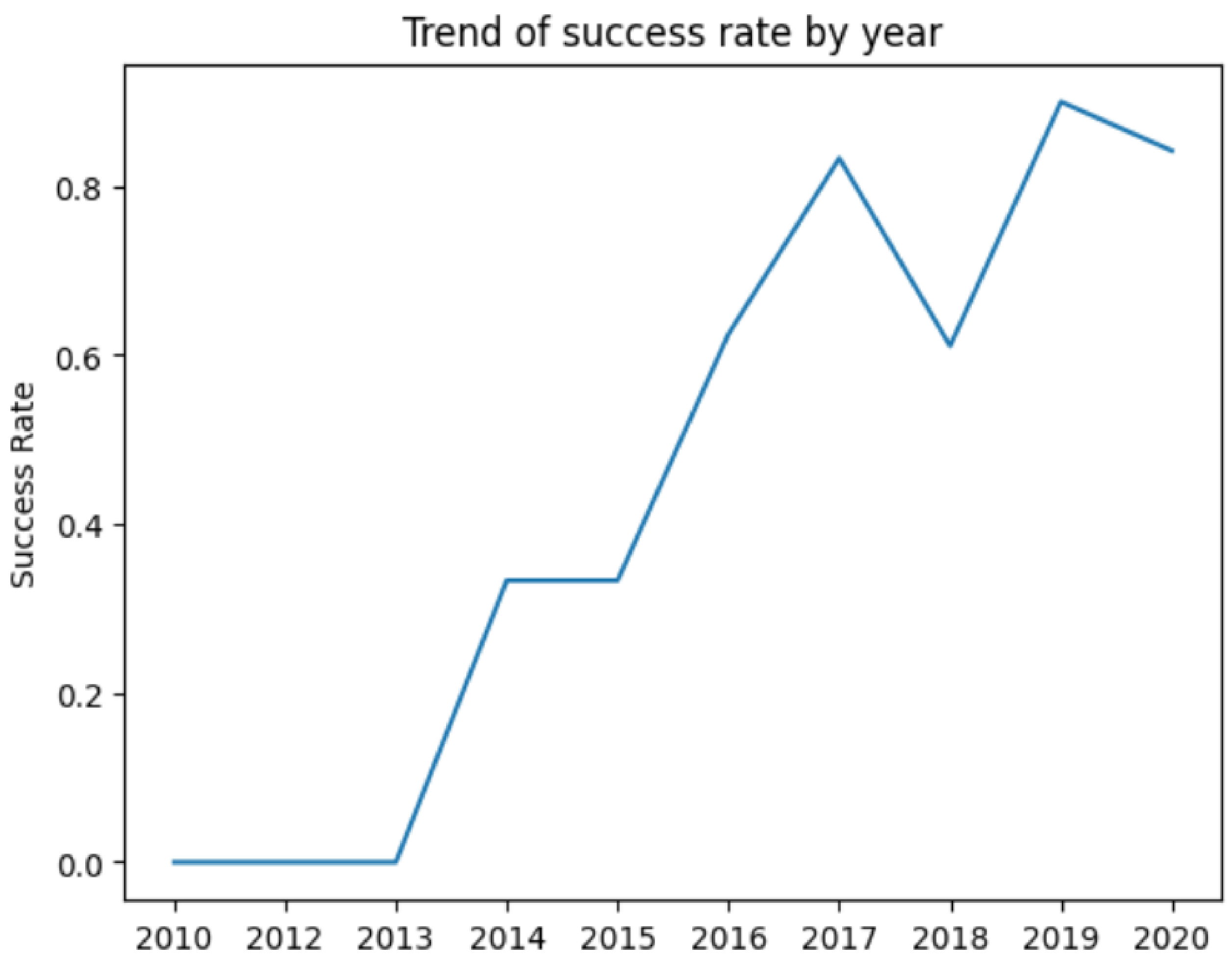
- Once again we see the success rate increase with Flight Number
- But this relationship doesn't seem to hold for all orbits
- It's true for LEO but doesn't hold in GTO orbit

Payload vs. Orbit Type

- For heavier payloads, VLEO, ISS and PO orbit are the safest
- With GTO it's hard to tell since the results are mixed for all payloads launched
- For lighter payloads SSO seems a good option as it has always succeeded



Launch Success Yearly Trend



- This clearly illustrates SpaceX's improvement over time
- Despite a slight hiccup in 2018, their success rate has constantly improved since 2013

All Launch Site Names

- There are 4 distinct launch sites
- Matches the map visualization we saw earlier

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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- There are two launch sites whose name begin with ‘CCA’
- Both of these launch sites only launch into LEO orbit with F9 v1.0 Boosters
- All missions at these sites have been successful
- Both NASA and SpaceX have commissioned launches at these sites

Total Payload Mass

Customer	Total Payload Mass (kg)
NASA (CRS)	45596

- The total payload mass launched by NASA is 45,596 kg

Average Payload Mass by F9 v1.1

- The average Payload Mass carried by F9 v1.1 rockets was 2,928.4kg

Booster_Version	Average Payload Mass (kg)
F9 v1.1	2928.4

First Successful Ground Landing Date

Date of 1st successful ground pad landing
2015-12-22

- The first successful ground landing was in 2015
- It took SpaceX five years to manage a successful landing!

Successful Drone Ship Landing with Payload between 4000 and 6000

- There have been five successful drone ship landings with a payload between 4000kg and 6000kg
- We can see that the F9 FT Boosters are well-suited to drone ship landings

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

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

- There have been 99 successes and only 2 failures total!
- Almost all launches had a successful outcome with a few minor exceptions. Amazing!

Boosters Carried Maximum Payload

- The heaviest payload carried was 15,600kg
- We can see that the F9 B5 Boosters are well-suited to such a massive payload

Booster_Version	Max Payload Mass (kg)
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

- There were 2 launches in the year 2015 that failed to land on a drone ship
- It seems that the F9 v1.1 Boosters may not be very well suited to drone ship landings

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

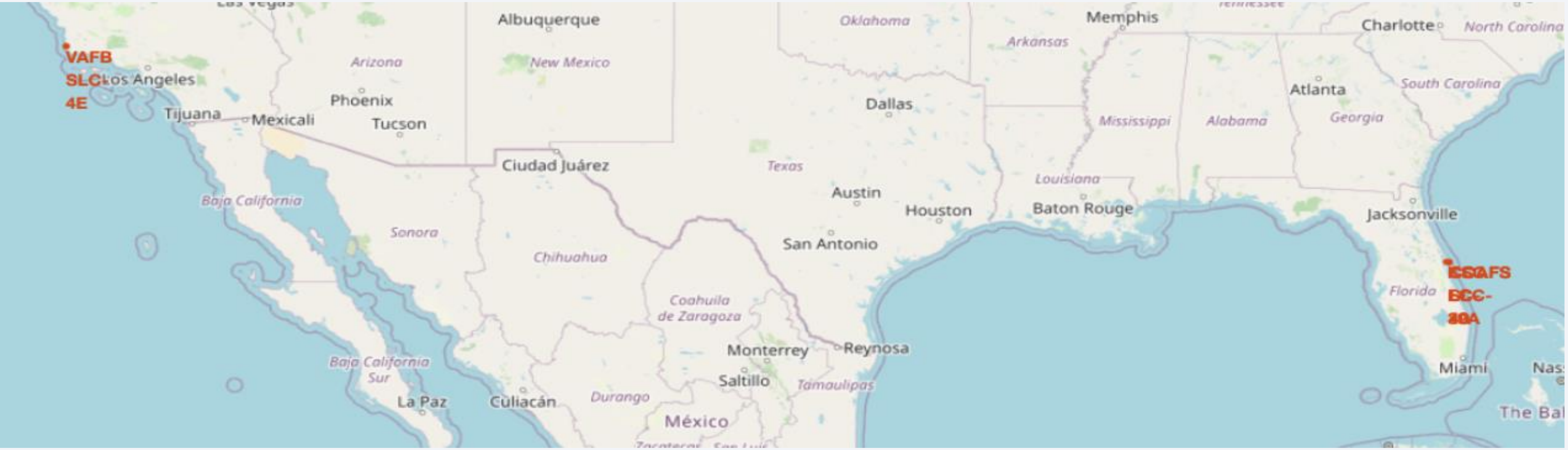
- Majority of launches between these dates made no attempt to land the rocket
- Those that did, however, were mostly successful
- Only a handful of failures from all the launches

Landing_Outcome	Count
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

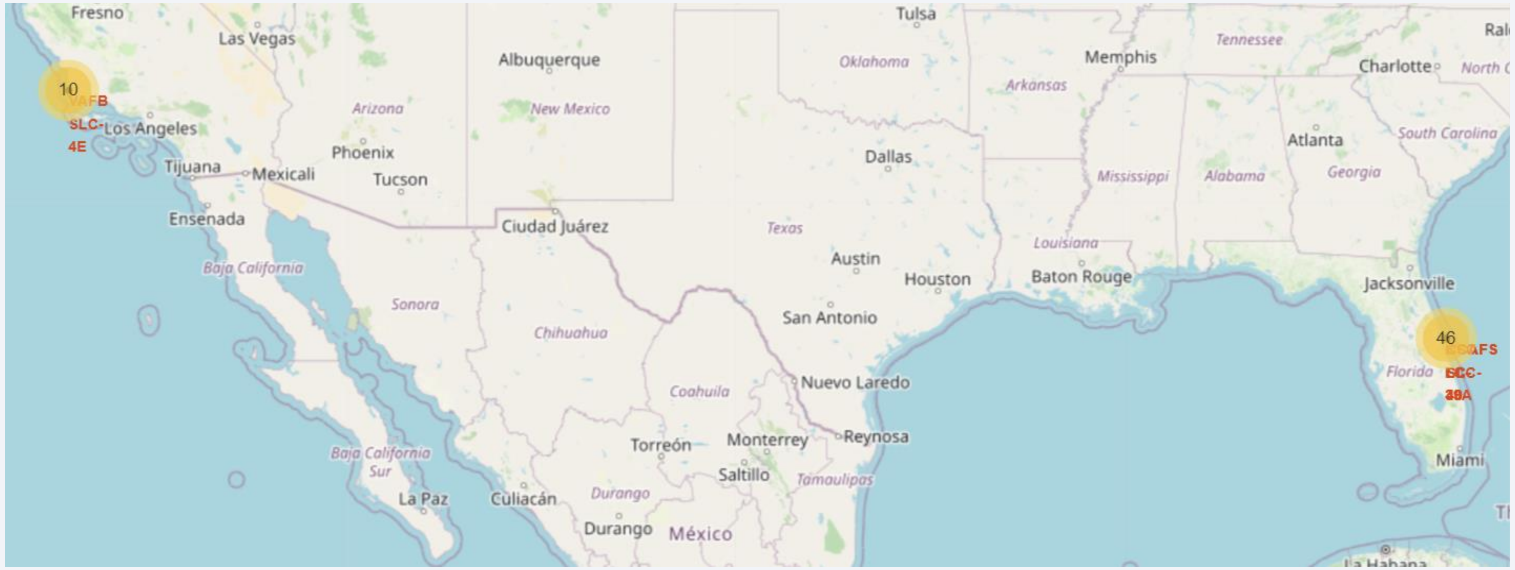
Launch Sites Proximities Analysis

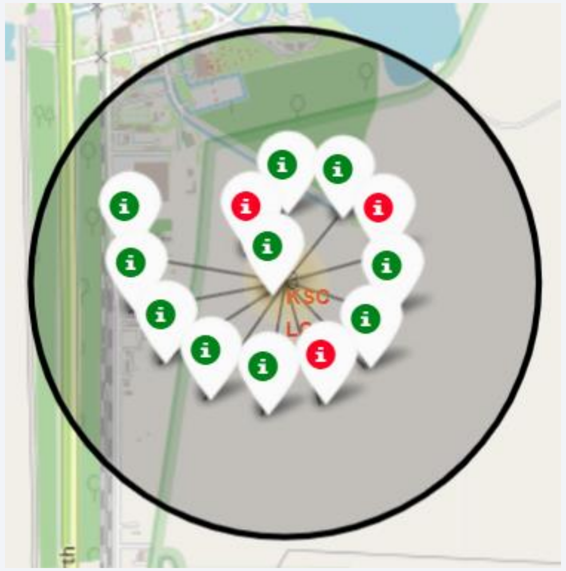
SpaceX Launch Site Locations Mapped



- Maps the locations of the SpaceX Launch Sites
- We can see that the VAFB SLC 4E site is located on the West Coast while the others are on the East Coast of the United States

Landing Outcomes Mapped by Site





- Displays clusters of markers at each Launch Site
- Each individual marker represents a launch, colored by whether it was a successful outcome or not
- Quickly illustrates the successes and failures of each Launch Site

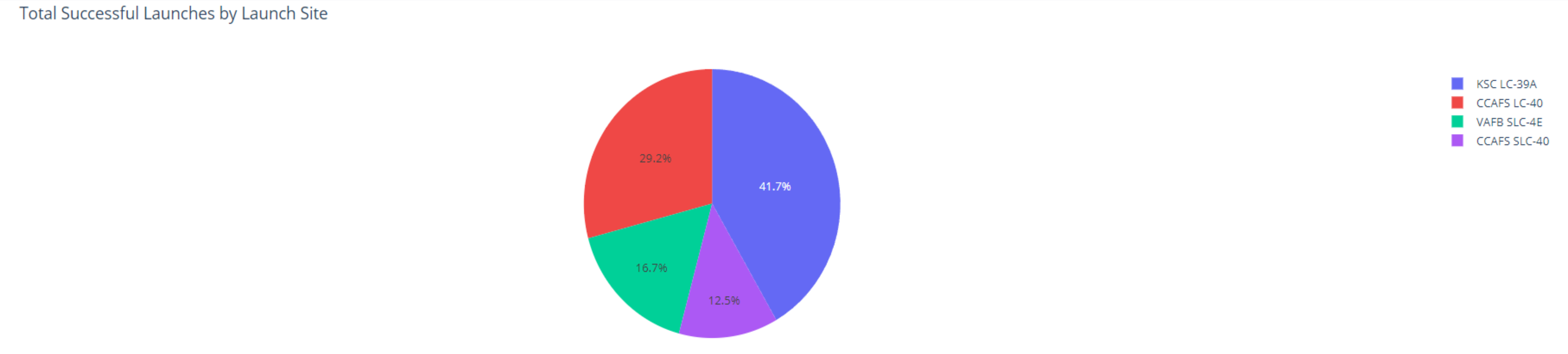
Launch Site Distance to Proximity Mapped

- All launch sites have a few specific structures in their proximities
- These include railways, coastlines, and highways
- This map shows the distance of launch site KSC LC-39A to its closest proximity, which is the nearby coastline 0.90Km away.
- This shows that all launch sites maintain a close distance to coastlines

Section 4

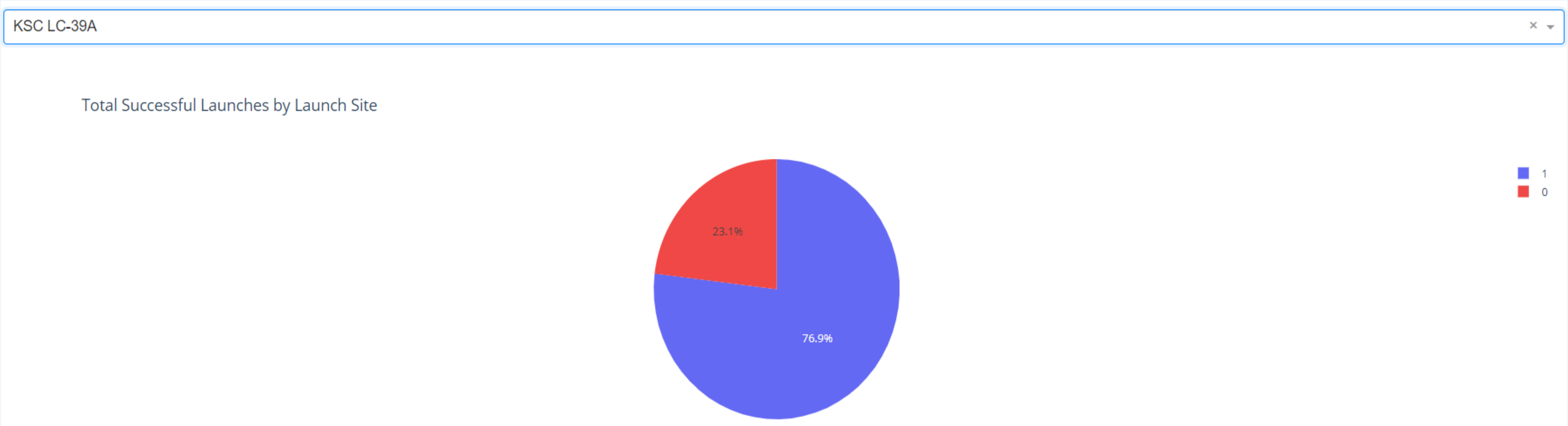
Build a Dashboard with Plotly Dash

All Sites Launch Success Rate



- Shows the proportion of successful launches for each launch site
- We can see that majority of successful launches come from the KSC LC-39A site
- The least successful launch site has been CCAFS SLC-40

Best Launch Site Success Rate

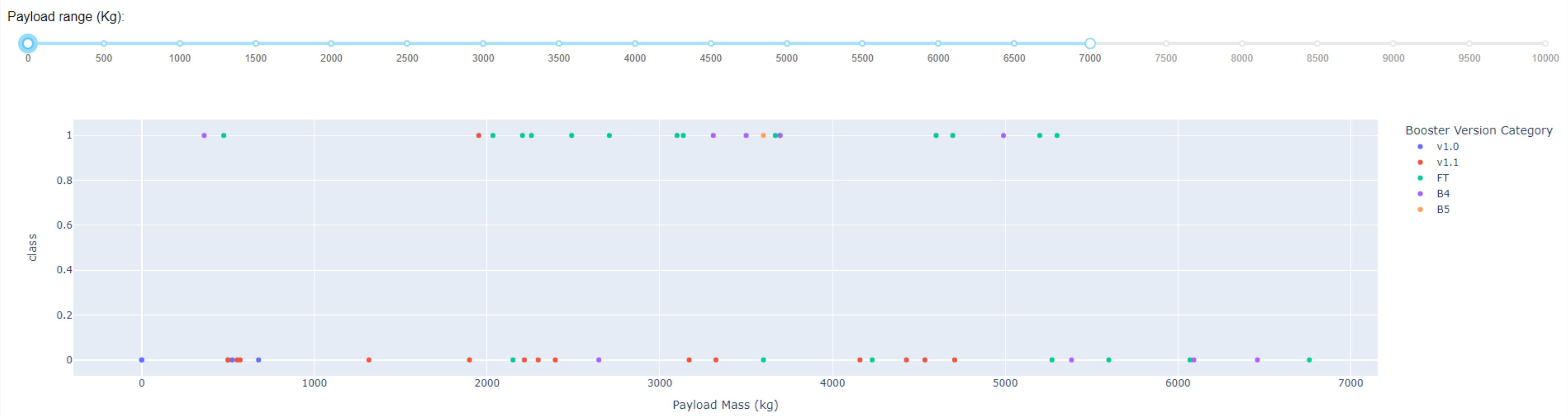


- The most successful launch site is KSC LC-39A
- The success rate of this site is very high, at 76.9%

Launch Outcome by Payload Mass (Best Range)

- The scatter plot illustrates the Launch Outcome against the Payload Mass in kg
- This particular plot is for the Payload Mass range 2000kg to 5000kg
- This range of payload masses contains the majority of successful launches

Launch Outcome by Payload Mass (Worst Range)

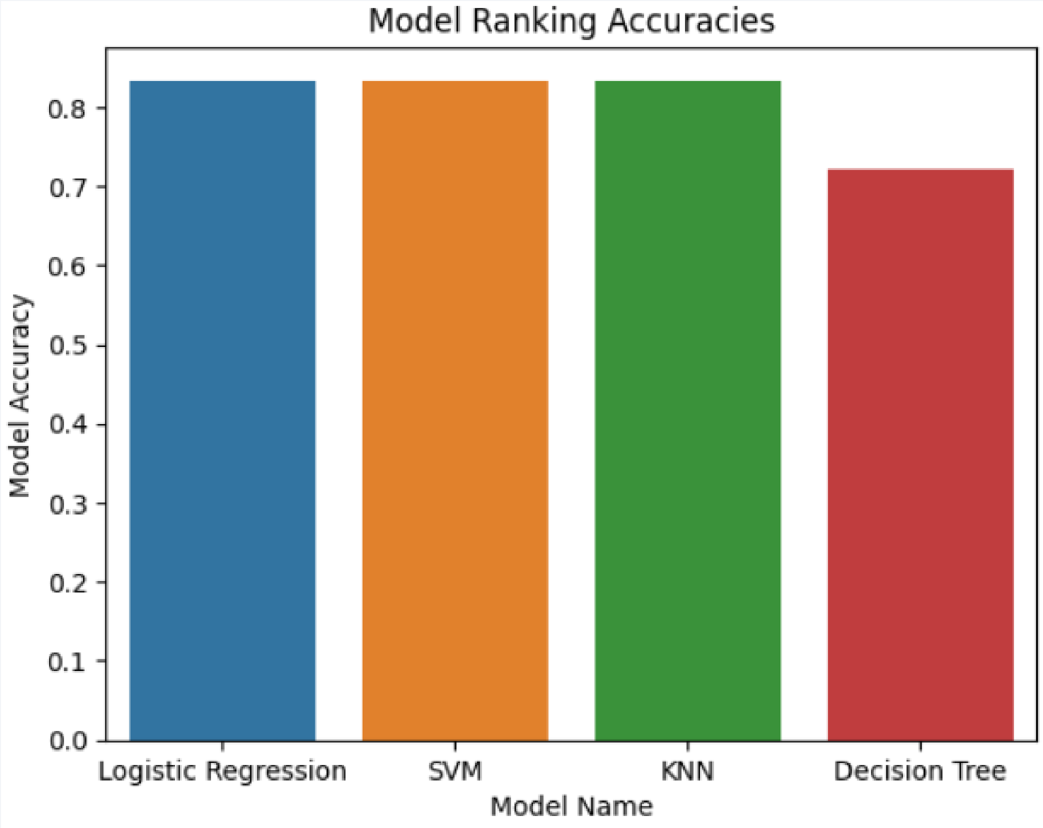


- Like the previous plot, the scatter plot illustrates the Launch Outcome against the Payload Mass in kg
- This particular plot is for the Payload Mass range 0kg to 7000kg
- This range of payload masses contains the majority of unsuccessful launches

Section 5

Predictive Analysis (Classification)

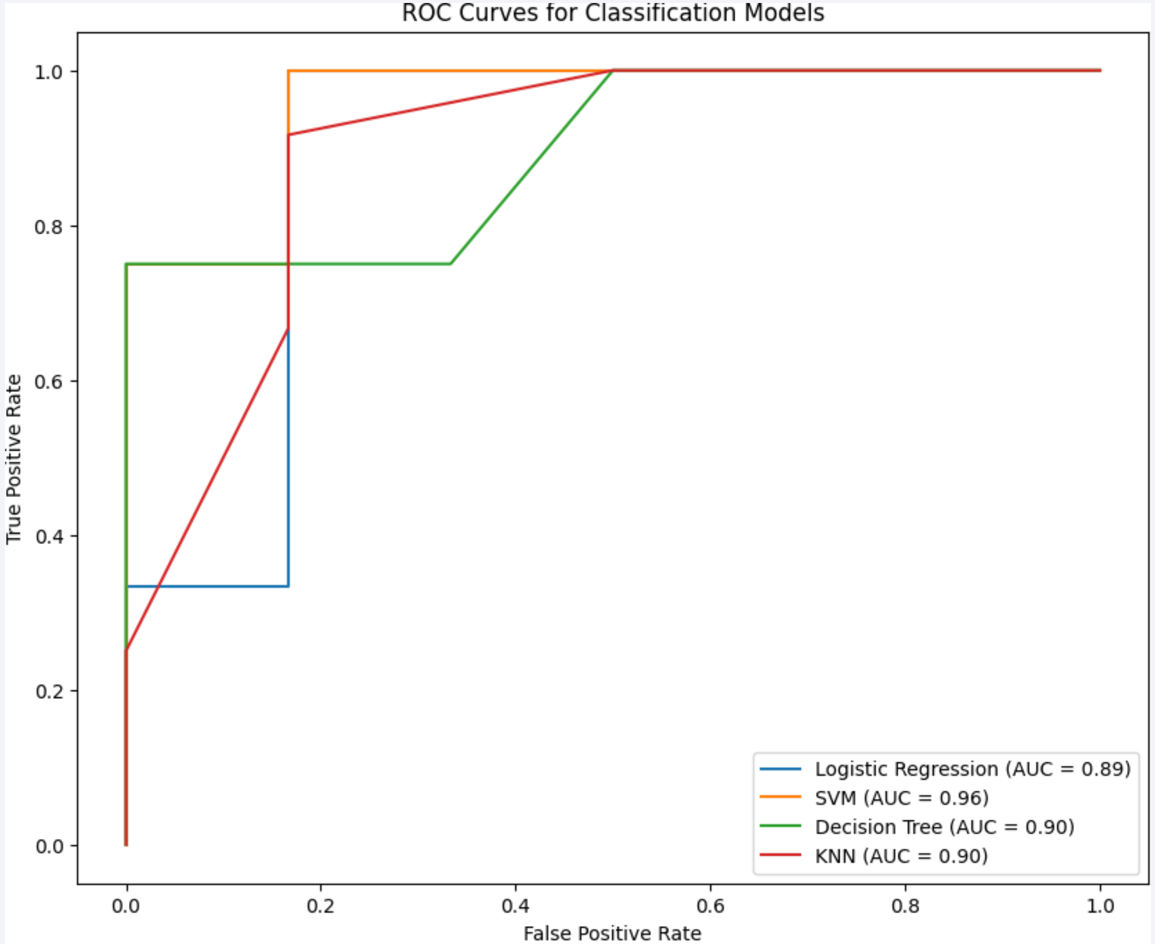
Classification Accuracy – Model Comparison



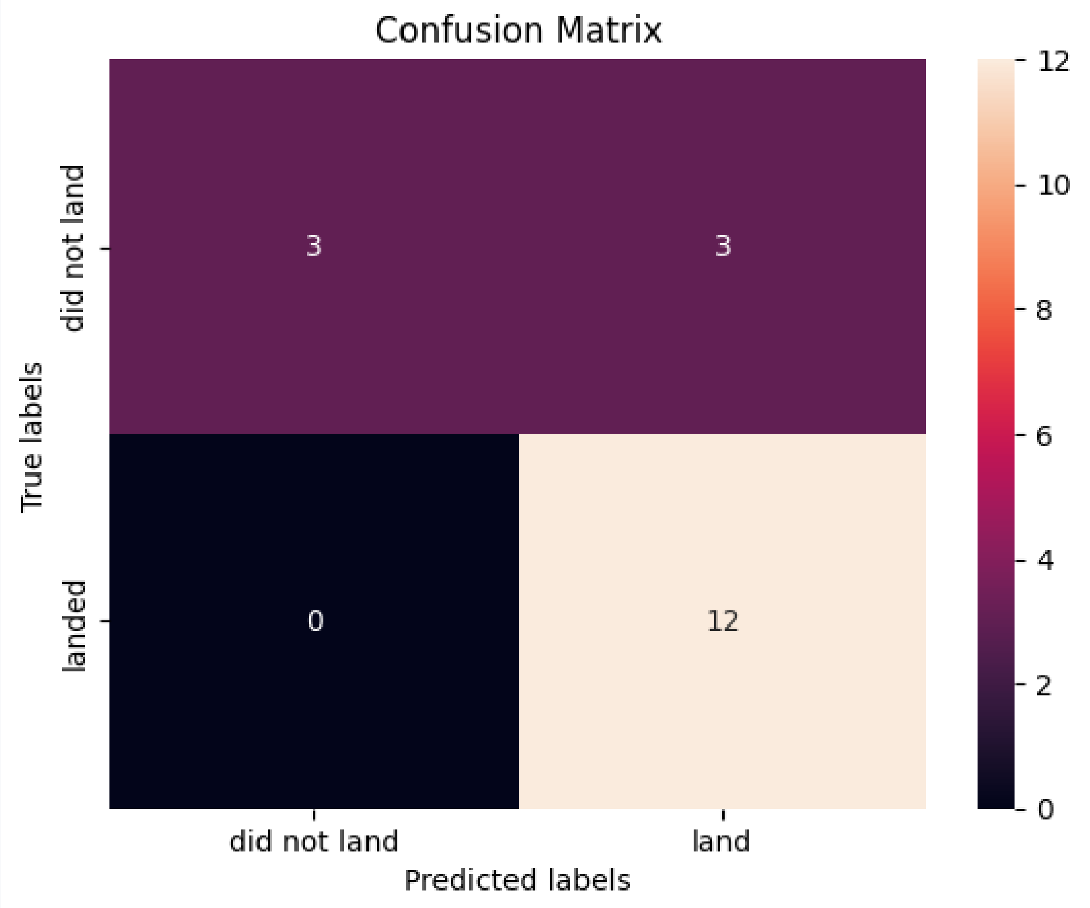
- The bar chart plots the accuracies of each model
- We can see that Logistic Regression, SVM, and KNN all performed similarly with an accuracy of around 0.8
- Decision Tree performed worse than the rest, at around 0.7 accuracy

Classification Accuracy – Picking Best Model

- The following ROC curve helps break the tie between the similarly performing models.
- We can see that the SVM Classifier has the highest area under curve (AUC) value
- This means it best balances the specificity and sensitivity of its predictions
- Making SVM the best model for this problem



Confusion Matrix



- This is the confusion matrix for the SVM Classifier Model
- We can see that majority of the predictions are true
- However, the model seems to predict some false positives
- Fortunately, the model does not predict any false negatives

Conclusions

- In conclusion, we were able to use historical SpaceX launch data to create powerful machine learning models
- These models turned out capable of predicting the launch outcome for a SpaceX rocket given its features
- The deployment and implementation of these models can greatly aid SpaceX in their launch prospects
- By helping them identify factors that deter the success rate of a launch
- And allowing them to simulate the results of a launch before it even occurs!

Thank you!