IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yusuf Ardahan Dogru
23 September 2024

Outline

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

Executive Summary

- Data collection was done on SpaceX data with the help of SpaceX API and Wikipedia scraping. Exploratory Data Analysis was done to get an initial understanding of the data with scatter plots, bar and line charts.
- An interactive map and dashboard was created with Folium and Plotly Dash. Lastly, predictive analysis was done with the help of various machine learning models for the classification of launch outcomes given the other data.
- The models performed with high accuracy on the test set, supplemented by the optimization of hyperparameters and the understanding of the data from the data science methodology

Introduction

- SpaceX offers a significantly lower cost of $62 million per Falcon 9 rocket launch compared to the $165 million charged by its competitiors. A key factor in this cost difference is the reuse of the first stage of the rocket.
- By accurately predicting the landing of the first stage, the overall launch cost can be estimated. This data is valuable in various scenarios, such as when a competitor seeks to bid against SpaceX for a rocket launch contract.
- We wanted to find if there's a way to predict first stage landing by looking at the data provided by SpaceX.
- The factors which affect the probability of the landing is also an import question to answer.

Section 1

Methodology

Methodology

Executive Summary

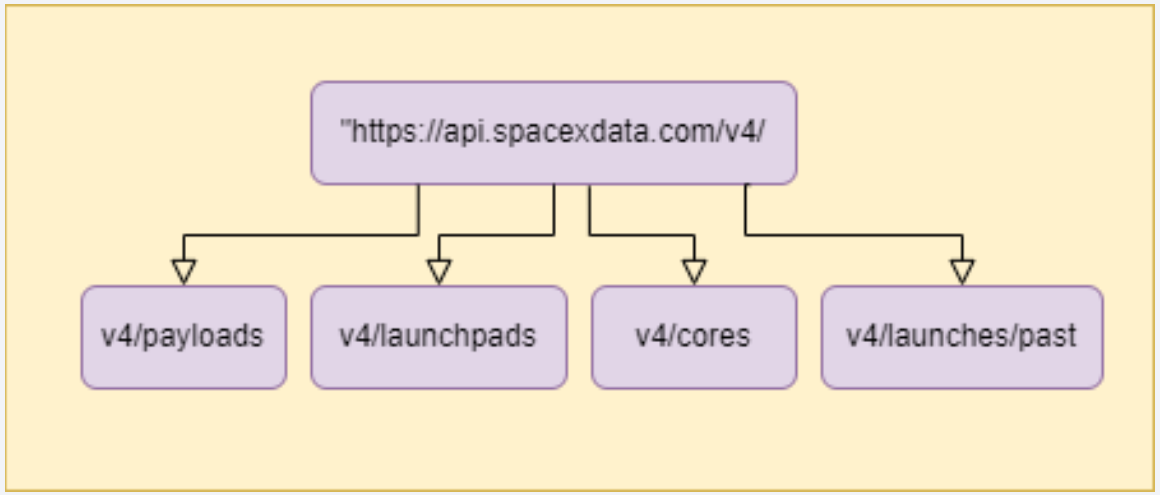
- Data collection methodology:
- Perform data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection

- SpaceX launch data was gathered from the SpaceX REST API.
- The API contained data about launches, including but not limited to rocket information, delivered payload, launch and landing specifications, and landing outcome.
- Differing data needed for the analysis was obtained through endpoints stemming from ‘https://api.spacexdata.com/v4/’

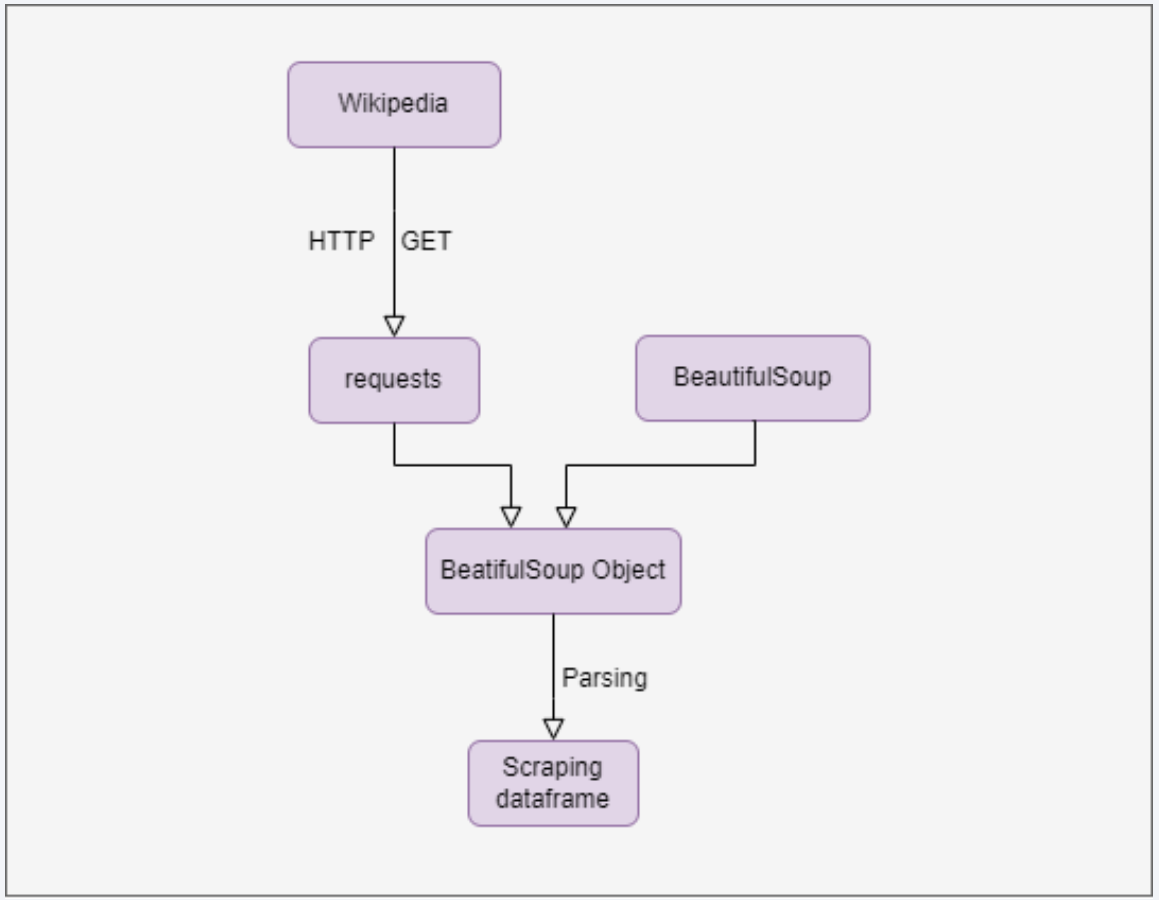
Data Collection – SpaceX API

- Booster name was extracted from v4/rockets, payload mass and destination orbit information was extracted from v4/payload, launch site data was extracted from v4/launchpads, etc.
- The following link can be used to access the GitHub URL: SpaceX API Calls



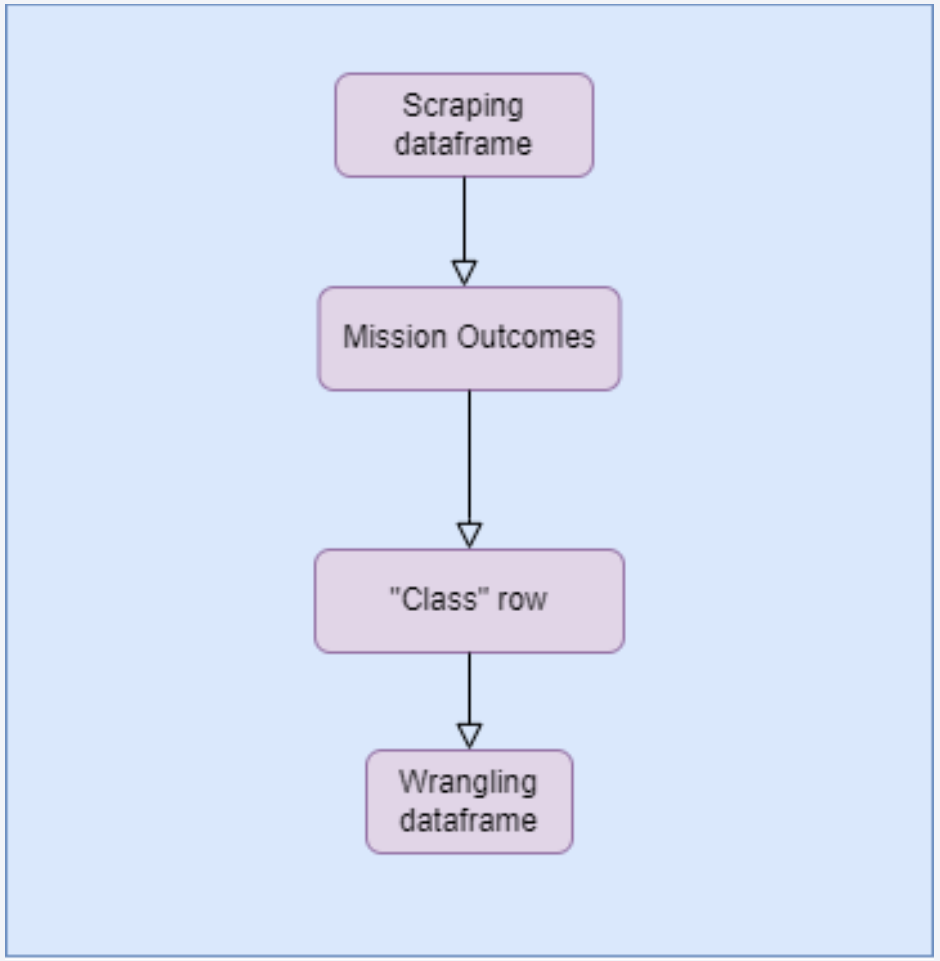
Data Collection - Scraping

- BeautifulSoup library was used to extract Falcon 9 launch records from Wikipedia, which was stored in a HTML table.
- Information parsed from the HTML table was stored in a dataframe for the next step.
- Access to the notebook for reference: [Web Scraping from Wikipedia]



Data Wrangling

- Data was processed to generate a new column in the dataframe.
- The new column, Class, is a binary variable that represent the classification variable that represents the outcome of each launch.
- Access to the notebook for reference: Data Wrangling



EDA with Data Visualization

- Scatter plots demonstrating the relationships between pairs of variables were generated.
- Flight number, launch site, payload mass, orbit, landing outcome, were compared with each other, to get an idea of how each of these factors influenced each other.
- Class column was plotted against year in a line graph to get a clear visualization of how the success rate of landings changed throughout the years.
- A bar chart was plotted using orbit and class data to identify which orbits have the highest success rates.
- For reference: EDA with Data Visualization

EDA with SQL

- Various SQL queries were performed to get a better idea about the data:
- The total number of successful and failed mission outcomes were queried.
- Failed landing outcomes in 2015 were queried, shown month by month, also displaying booster version and launch site information.
- The count of successful and failed landings between 4 June 2010 and 28 March 2017 were ranked in descending order.
- The total payload mass carried by boosters launched by NASA was queried.
- Average payload mass carried by version F9 v1.1 booster was displayed.
- For reference: EDA with SQL

Interactive Map with Folium

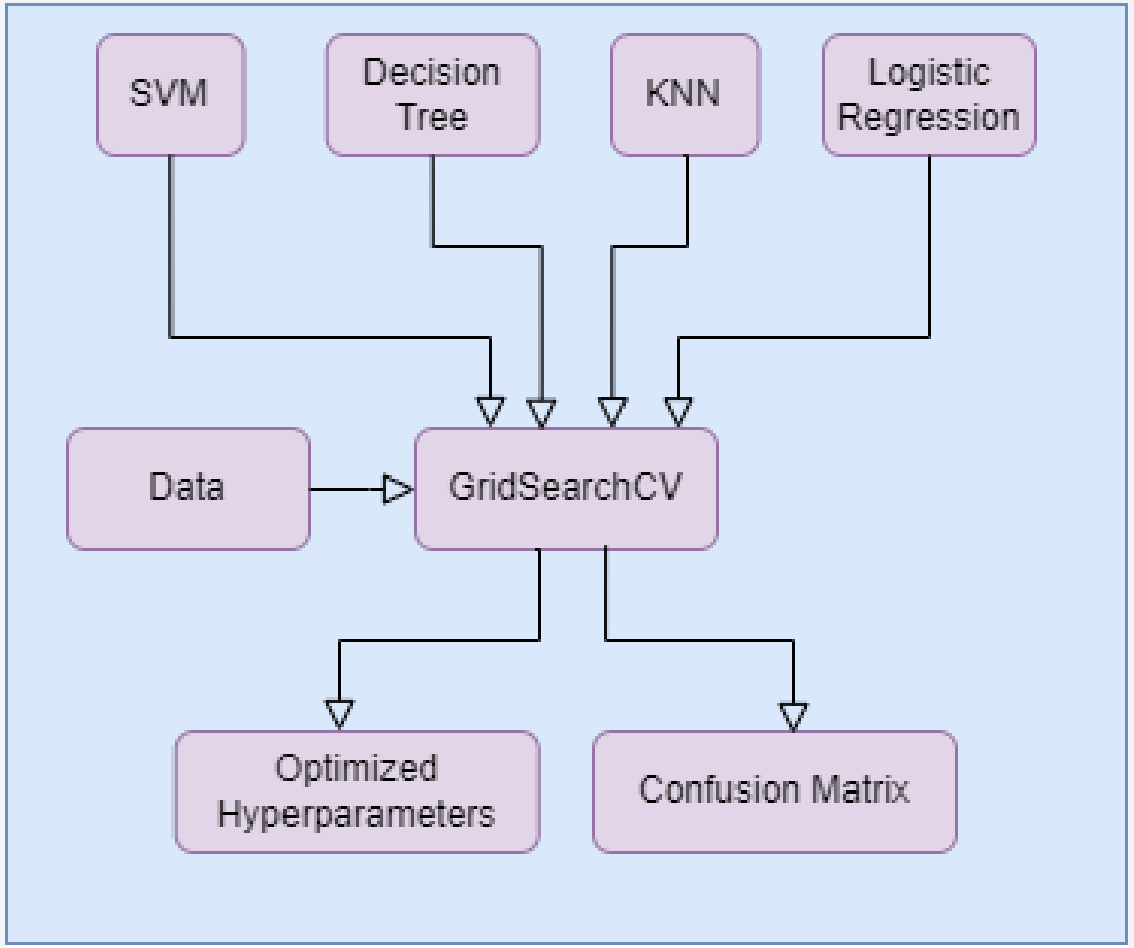
- All launch sites existing in the data was marked with a small circle, along with text and pop-ups to display where the launch sites are located on the map.
- For each launch site, successful and failed launches were marked to clearly show which sites had better success rate. Successful launches were marked with green and failed launches were marked with red, emphasizing the outcome.
- Lines were drawn from launch sites to proximities such as coastline, to give a sense of distance in the map.
- For reference: Interactive Folium Map

Dashboard with Plotly Dash

- The dashboard contains a pie graph, a scatter plot, a dropdown menu for selecting the launch site, and a slider for determining the payload mass limits.
- The pie graph shows the number of successful launches for every launch site, or a specific launch site chosen by the dropdown menu. This allowed a quick way of comparing launch numbers and success percentage between launch sites.
- The success of the launches was plotted against the payload mass in the scatter plot, and colored with the booster type to display the role of payload mass and booster type in the success percentage of rocket launches.
- For reference: Plotly Dash Dashboard

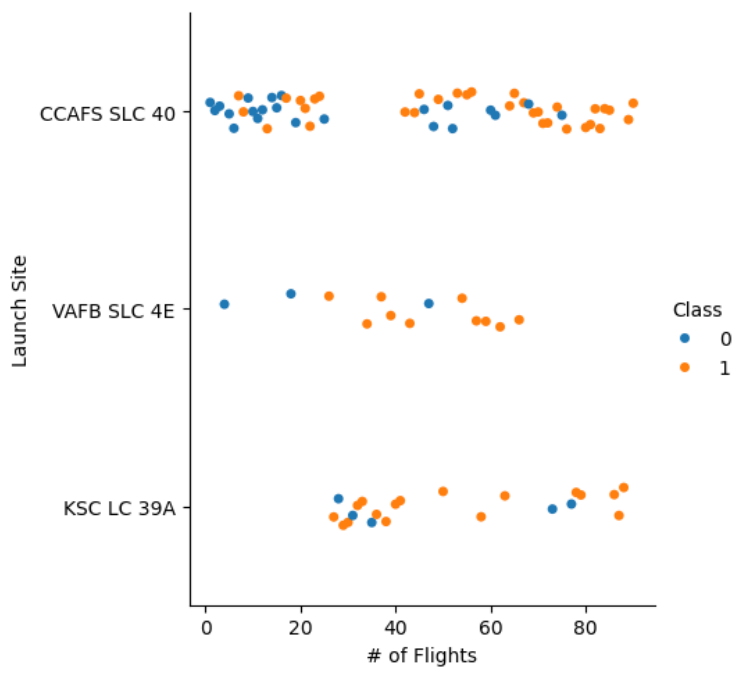
Predictive Analysis (Classification)

- Class column from the spaceX dataframe was designated as the target variable as the model was designed to predict the outcome of a launch.
- Data was standardized and split into train and test sets.
- Logistic regression, SVM, decision tree and KNN models were trained using the train data while using GridSearch to optimize the hyperparameters.
- Model performance was assessed using test accuracy and confusion matrix.
- For reference: Machine Learning Predictions

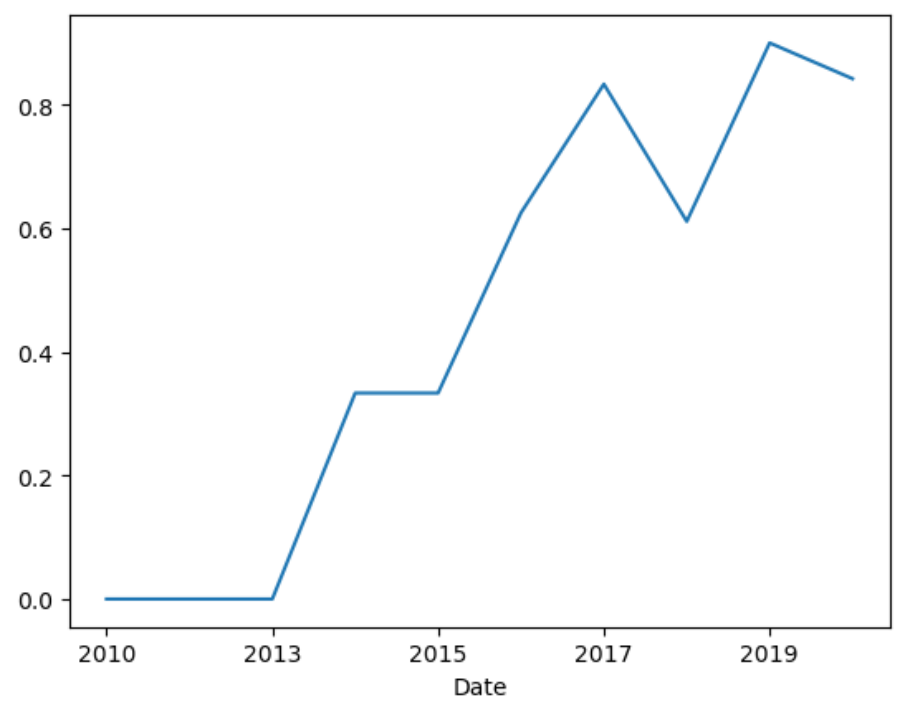


Results - EDA

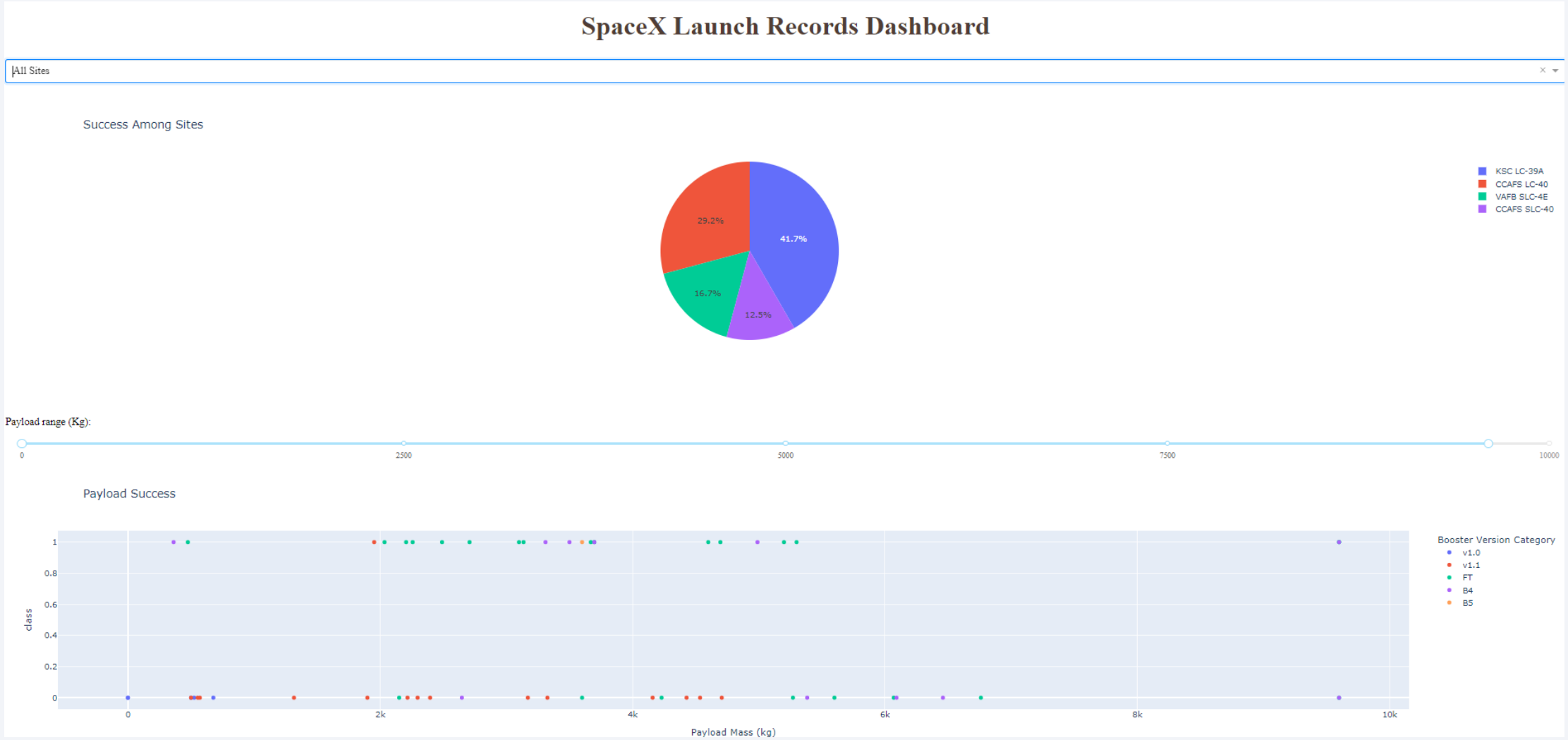
- A trend of increasing success with higher flight numbers was observed for each launch site(orange dots are successes)



- Success rate has kept increasing since 2013

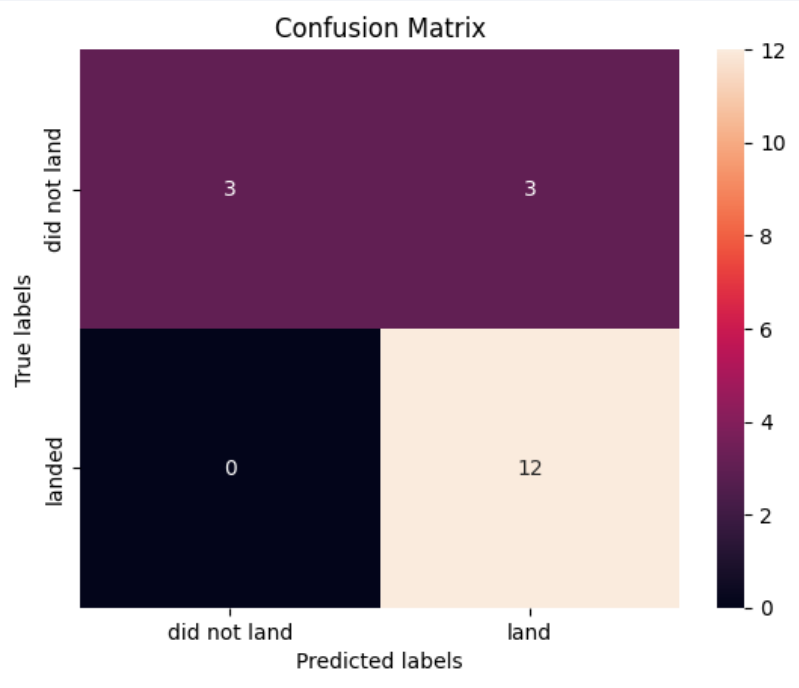


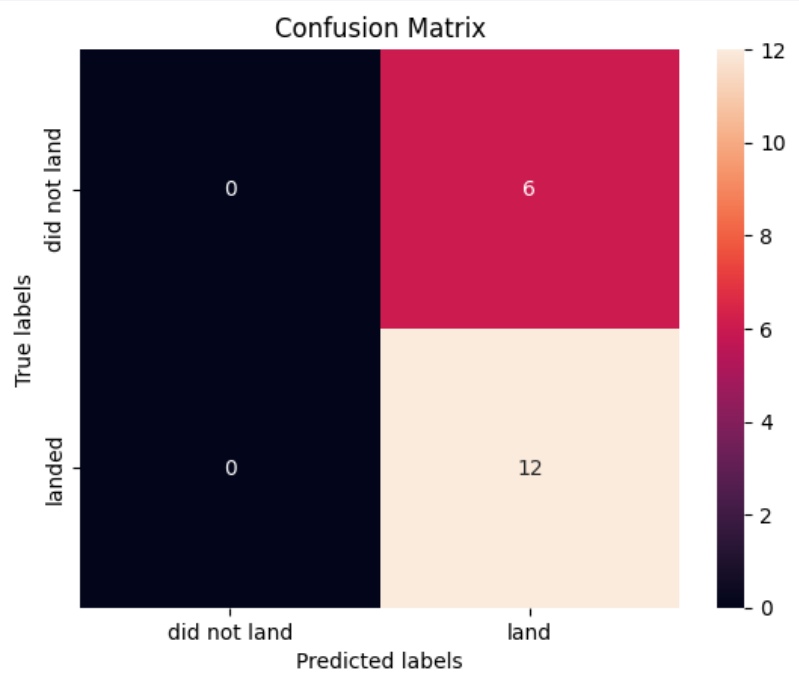
Results – Interactive Analytics



Results – Predictive Analysis

- Logistic regression, SVM, decision tree and KNN models were trained and tuned to achieve good accuracy on the test data. Some confusion matrices gained as a result are:



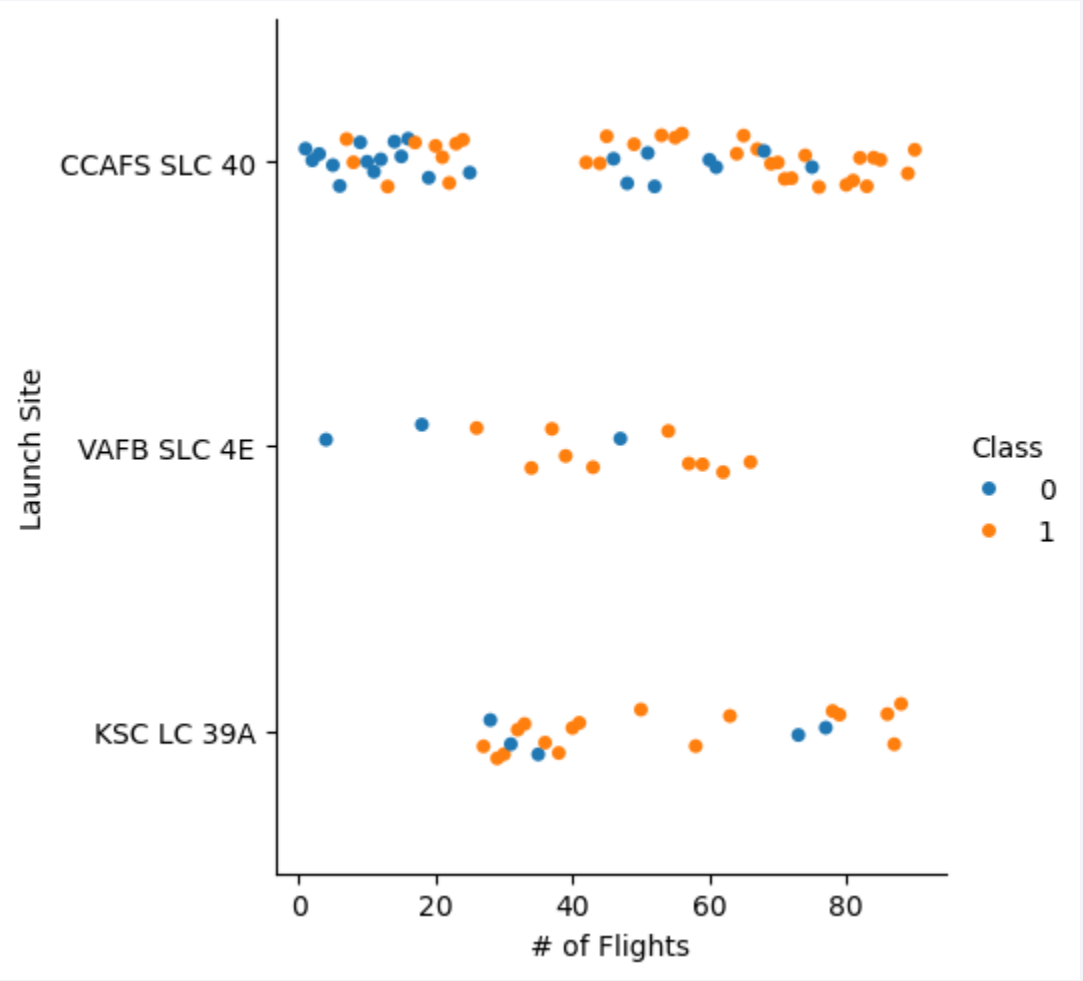


Section 2

Insights drawn from EDA

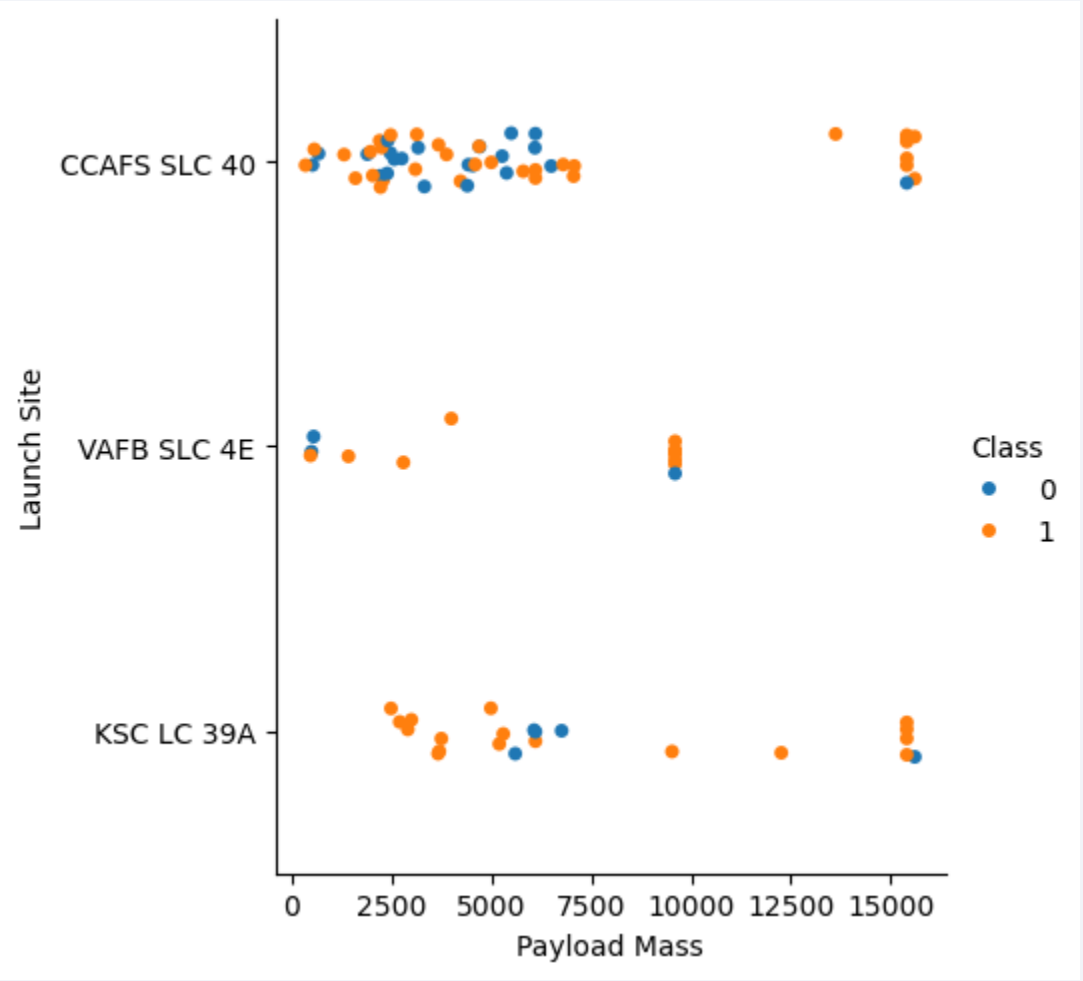
Flight Number vs. Launch Site

- For various launch sites, there was a clear affect of the number of flights on the outcome of mission success.
- In general, a trend of increasing success with increasing flight numbers were observed.



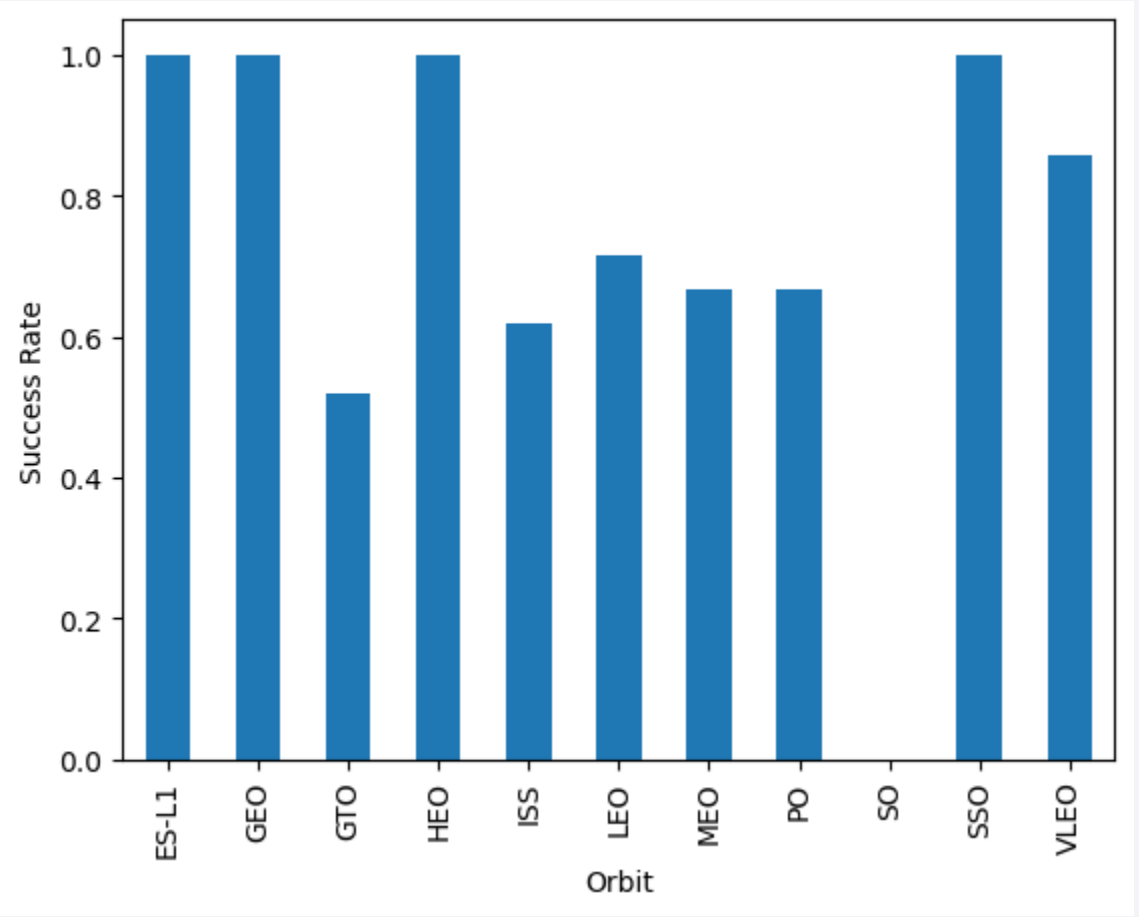
Payload vs. Launch Site

- The payload mass' effect was not so clear on the mission outcome, at least for the three launch sites in the graph.
- Especially for KSC LC 39A and CCAFS SLC 40, there were many successes with both low and high payload mass



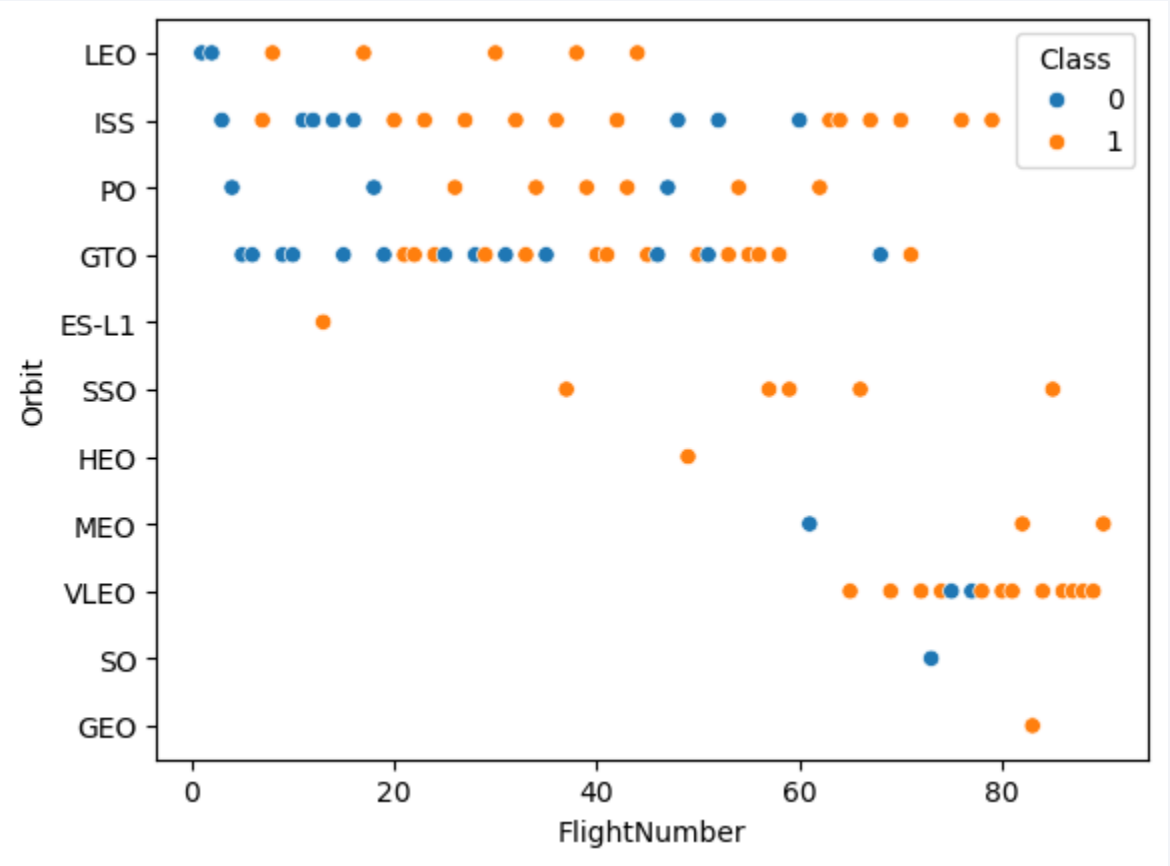
Success Rate vs. Orbit Type

- The orbit SO had no success for some reason.
- Some orbits, GTO, ISS, LEO, MEO, and PO had non-adequate success rate, between 50 and 70 percent.
- Other orbits like ES-L1, GEO, HEO, SSO, VLEO showed good success rates between 80 and 100 percent



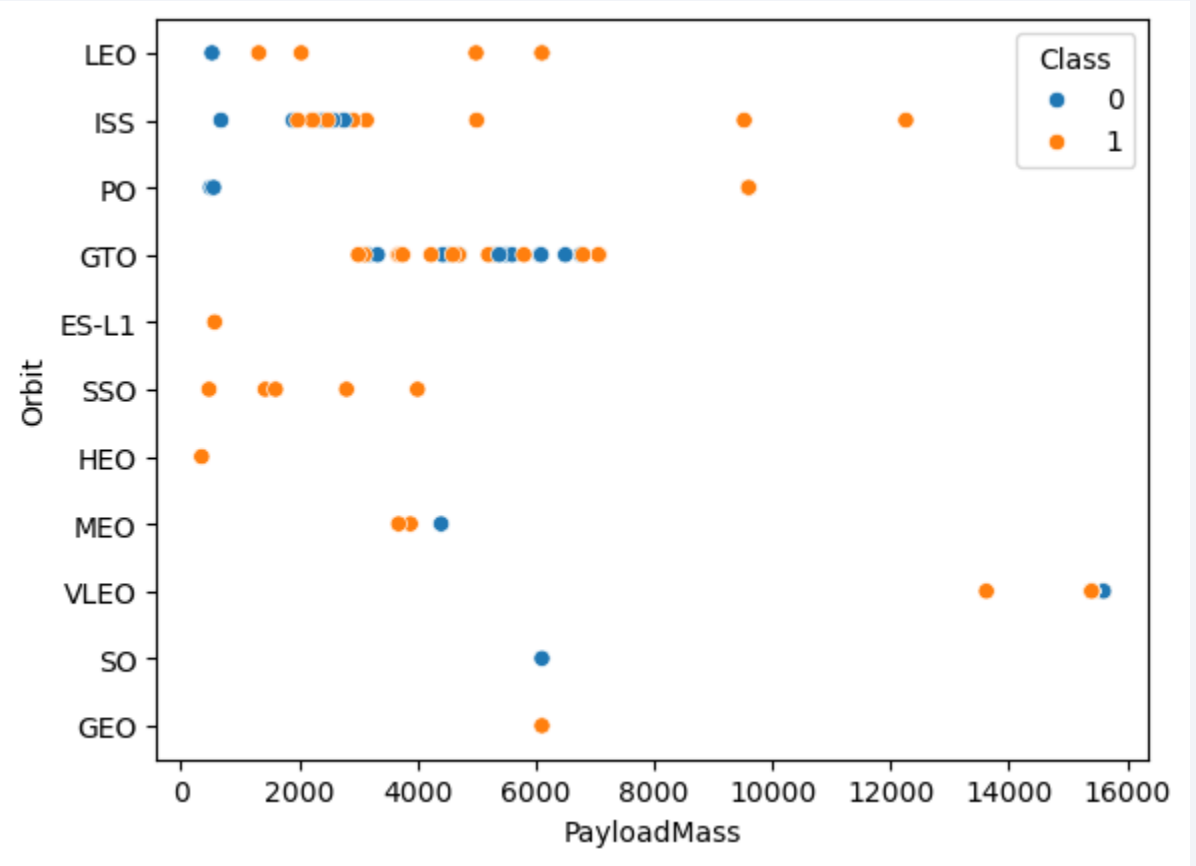
Flight Number vs. Orbit Type

- A general trend of increasing success with increasing flight number was observed among most of the Orbit types.
- Some orbits had a small amount of data and cannot said to have a relation with the flight number.



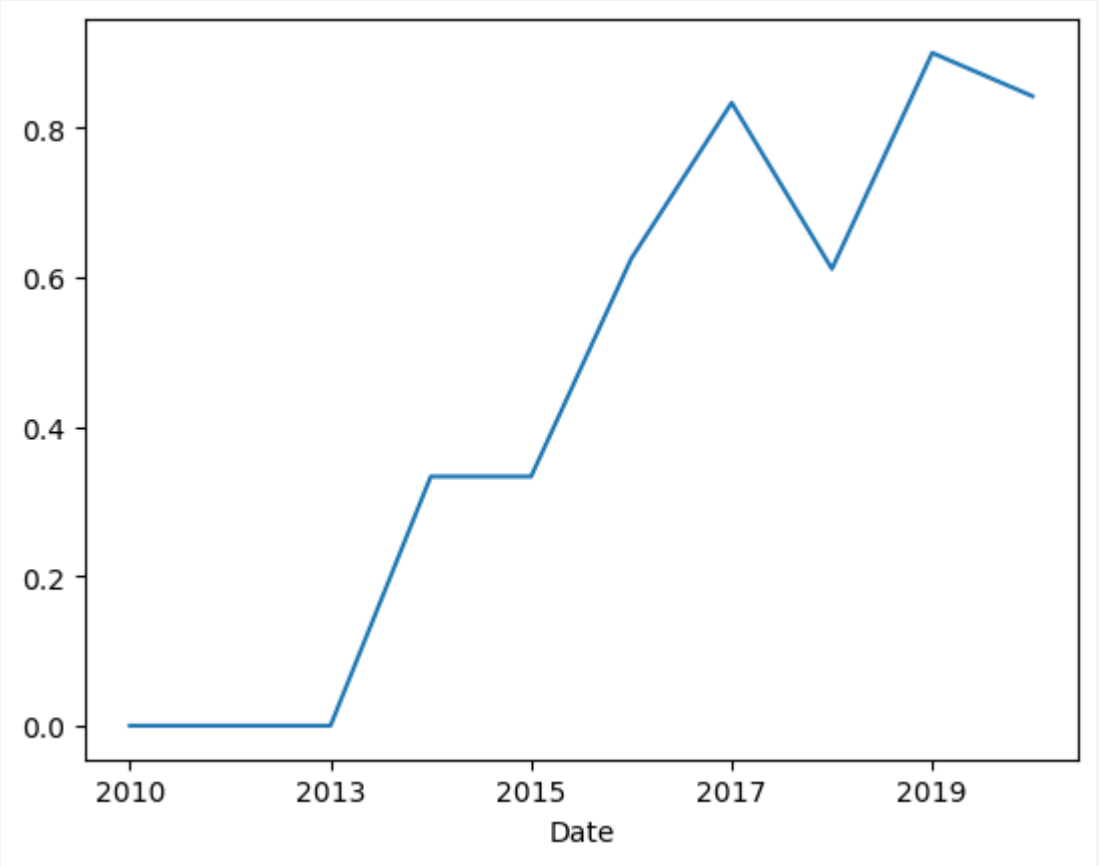
Payload vs. Orbit Type

- A general trend of increasing success with increasing payload mass was observed among most of the Orbit types.
- Some orbits had a small amount of data and cannot said to have a relation with the payload mass.



Launch Success Yearly Trend

- Success rate has increased along the years with increasing flight numbers



All Launch Site Names

- Names of the unique launch sites in the dataset were queried.
- There were 4 unique sites home to rocket launches

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` were queried.
- No ordering was done.

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

Total Payload Mass

- Total payload carried by boosters from NASA were calculated.

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated.

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad were found.

Date
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were listed.

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes were calculated.

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass were listed.

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 in year 2015 were listed.

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

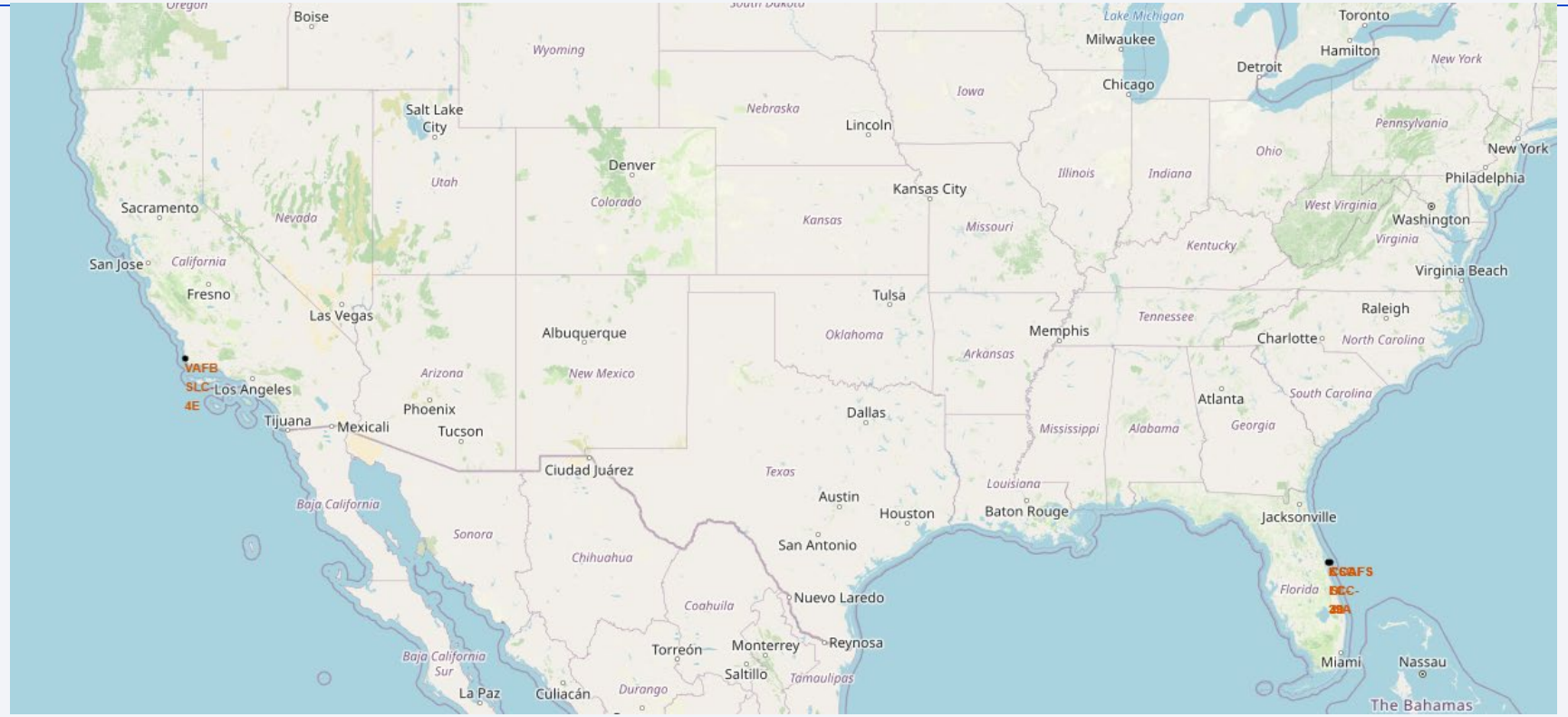
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, were ranked in descending order.

Landing_Outcome	count
Success (drone ship)	5
Success (ground pad)	3
Precluded (drone ship)	1
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
No attempt	10
Failure (parachute)	2

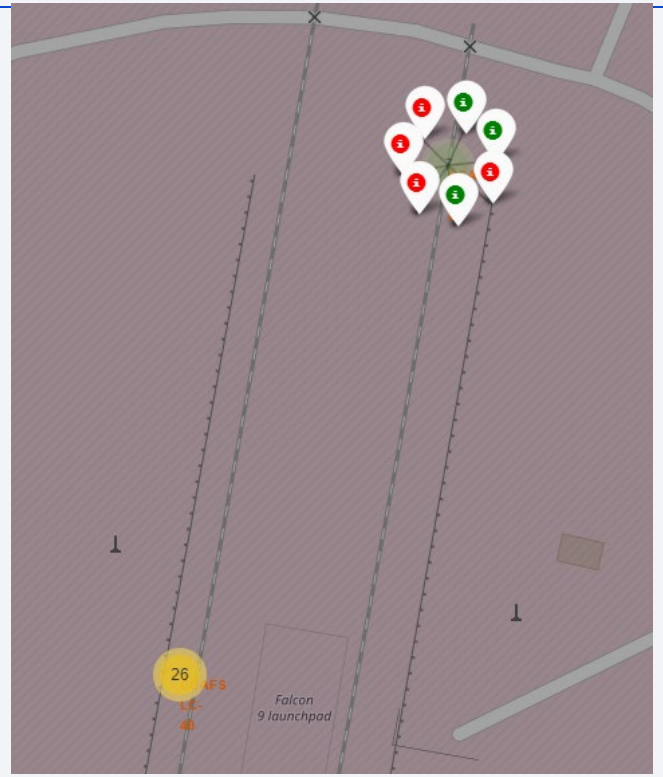

Section 3
Launch Sites
Proximities Analysis

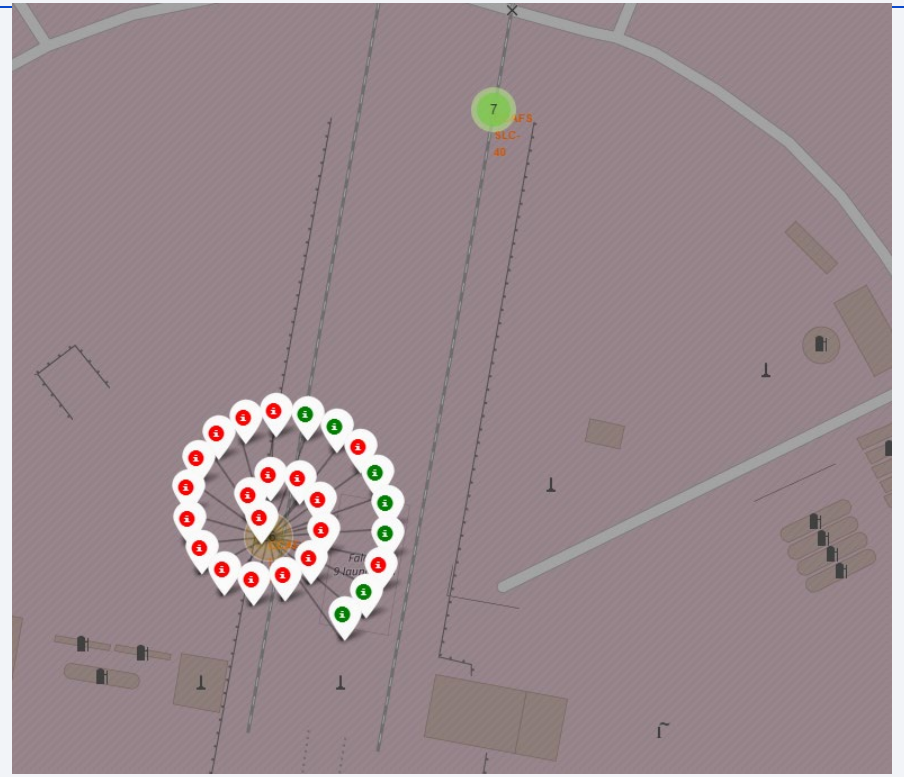
Launch Sites Marked on Map



- The unique launch sites in the dataset was marked on the Folium map with circle markers and pop-ups, labeled with their name. Three sites were nearly in the same place, which is the reason for the cramped text on the right.

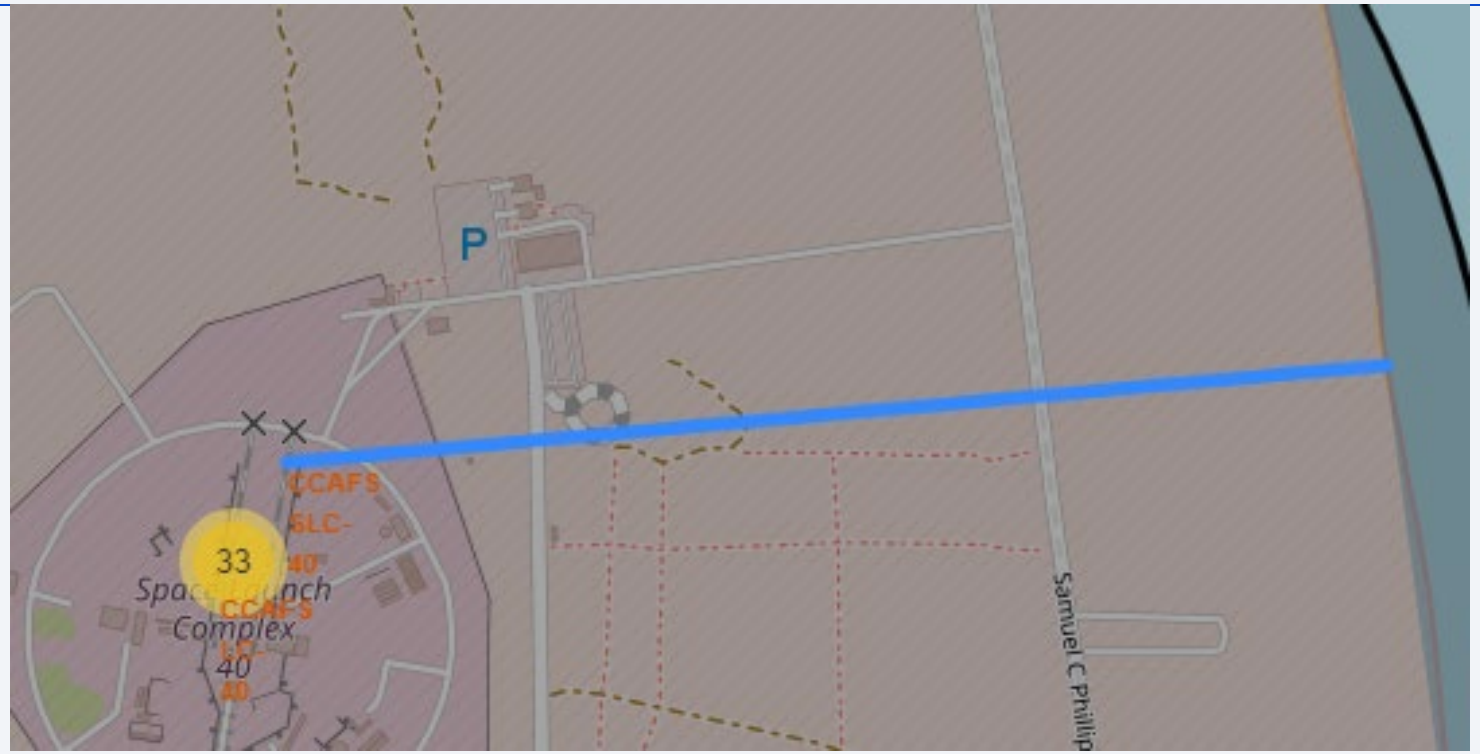
Success of the Launch Outcomes Marked on Map



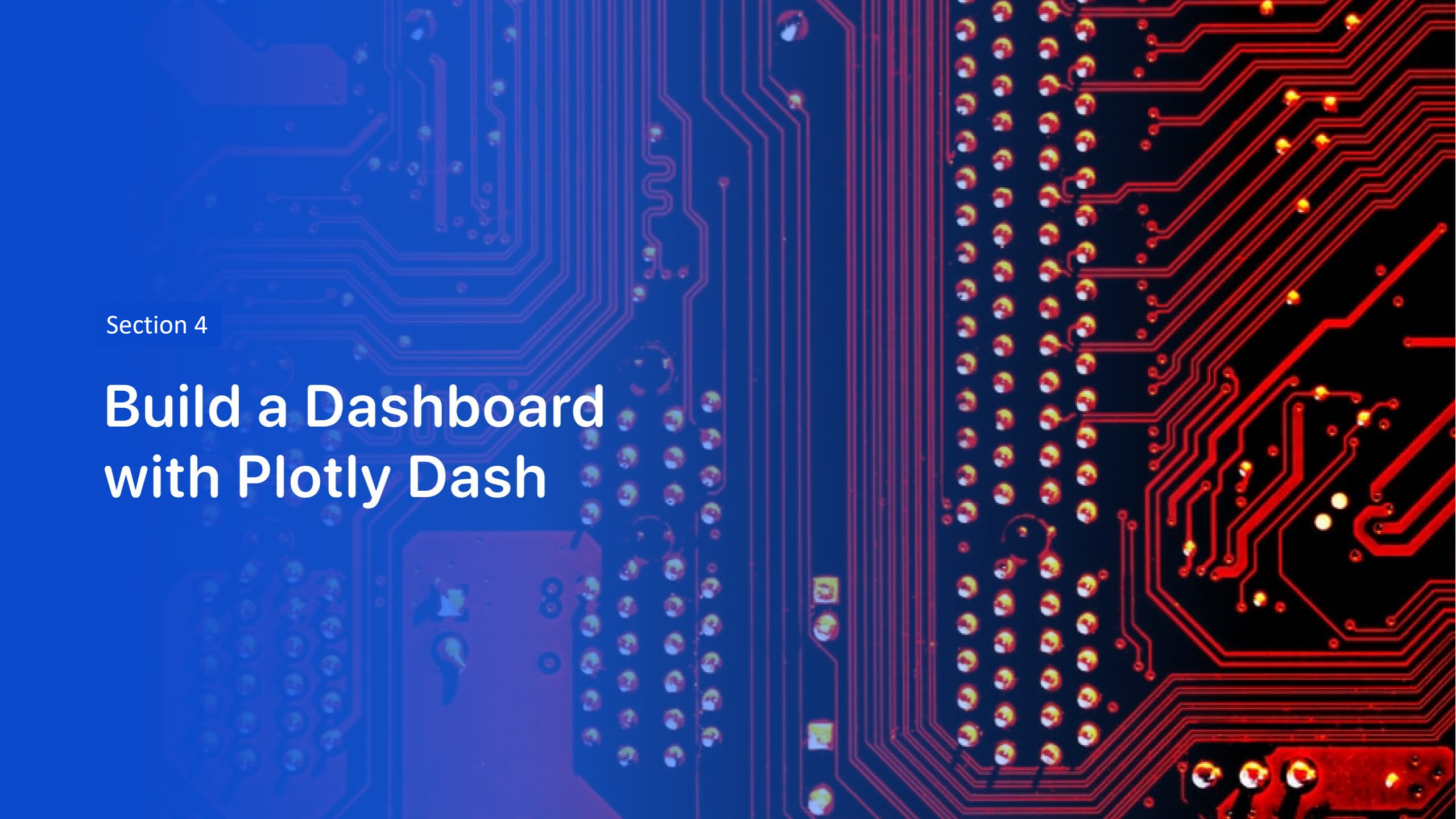


- Each launch was marked with red for an unsuccessful launch and with green for a successful launch. The marks can be seen by clicking on the launch site markers.

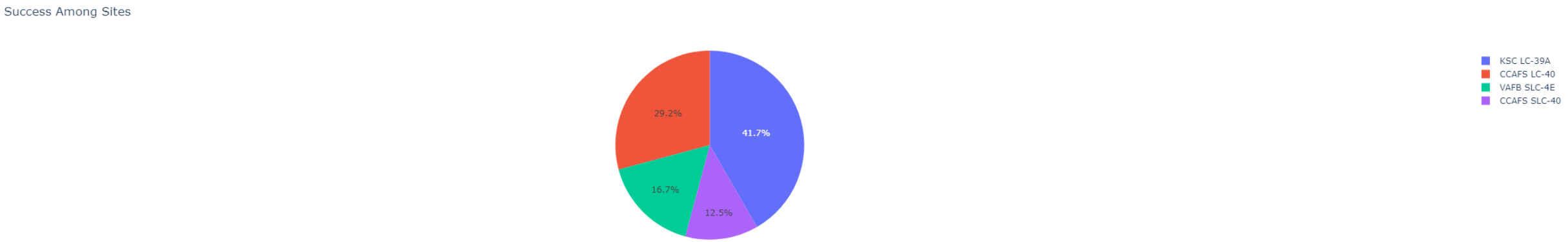
Lines to Proximities from Launch Sites



- The generated folium map was marked with lines that show the distances to its proximities such as railways and coastlines.

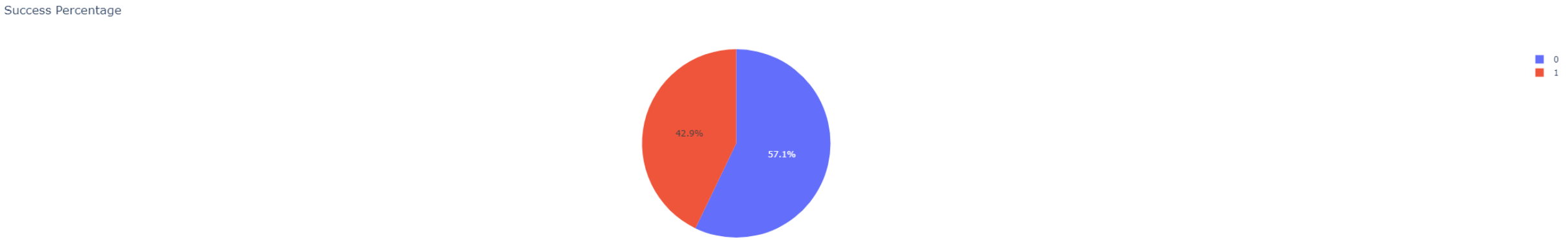

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count Pie Chart



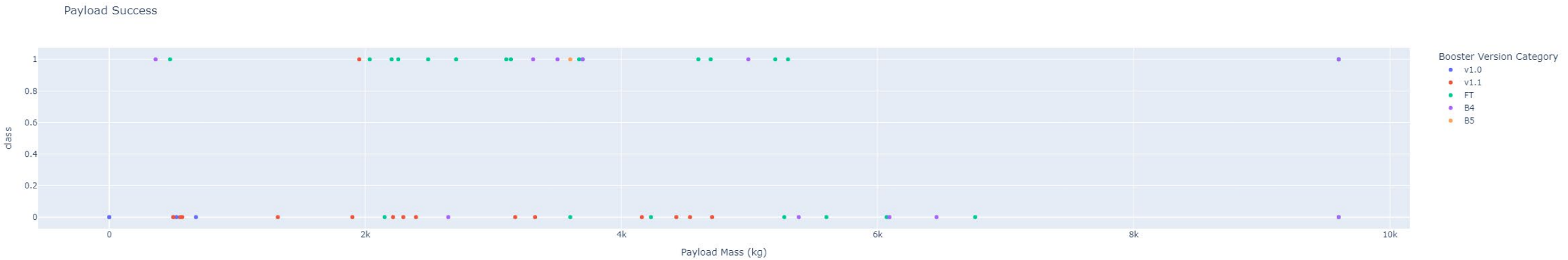
The total amount of launch successes grouped by the launch sites is generated by the Dashboard on default.

Launch Site With the Highest h success ratio



The pie chart demonstrates the launch site with the highest success ratio, CCAFS-SLC-40, selected from the dropdown menu on the dashboard.

Payload vs. Launch Outcome for all Sites



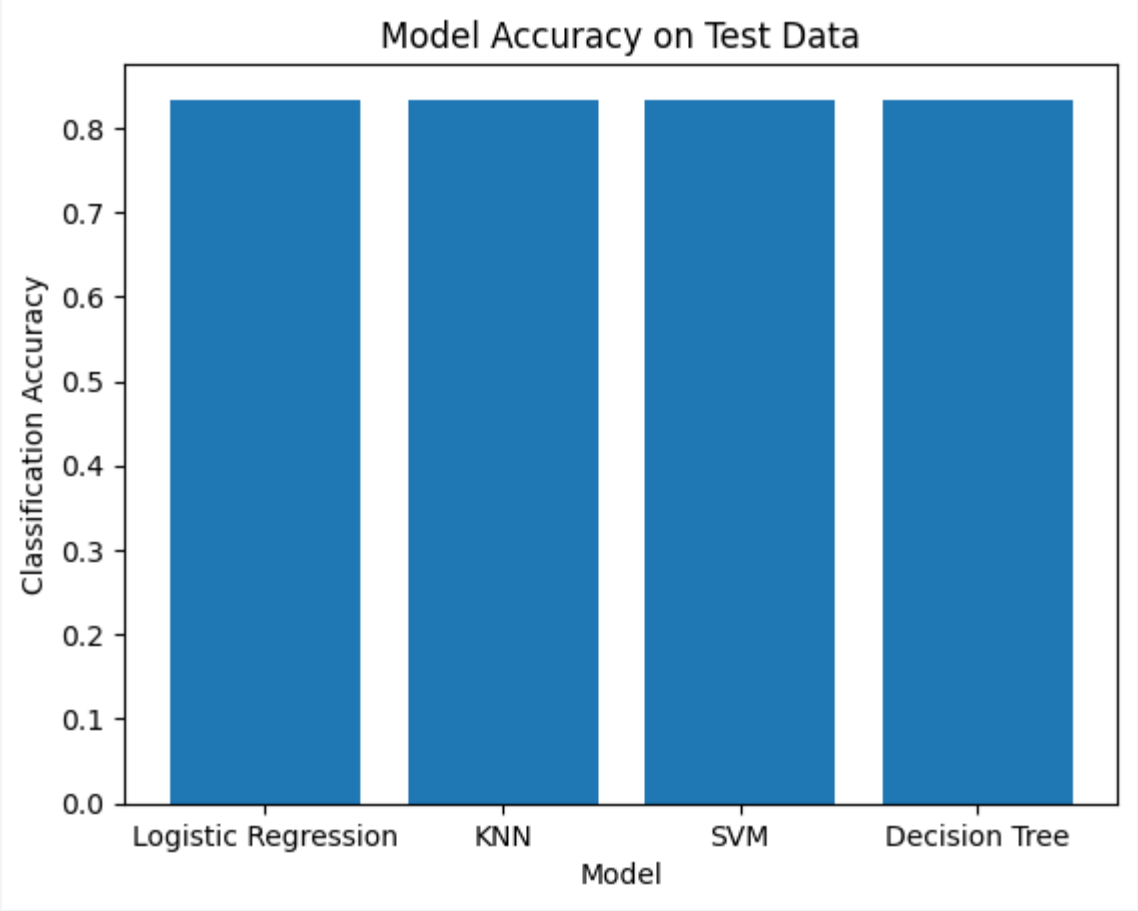
- The scatterplot is shown by default with different payload range selected in the range slider, showing outcomes for different booster versions.
- It is clear that the turquoise dots are more common for those of class 1, or those with successful launches. Therefore, it can be inferred that the booster version FT boasts the best success rate, at least for the payload mass range between 2k and 6k

Section 5

Predictive Analysis (Classification)

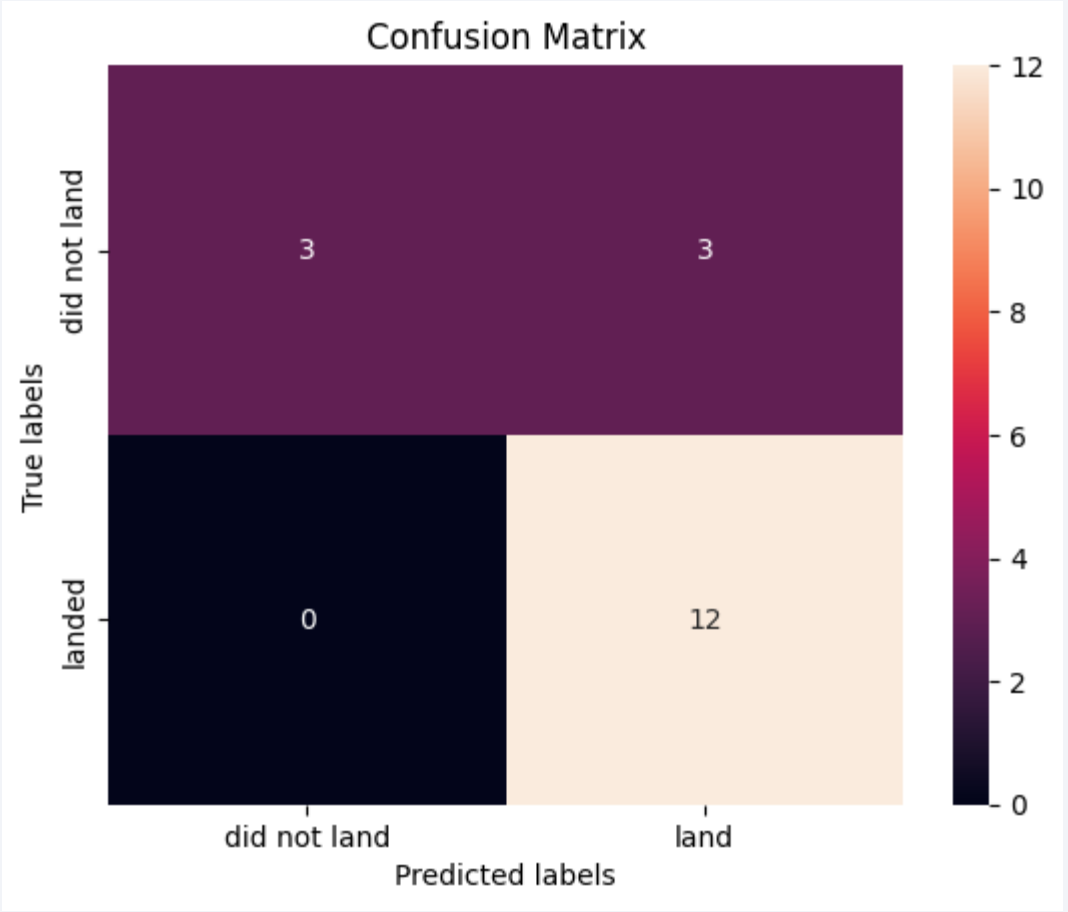
Classification Accuracy

- Built model accuracy was the same for all 4 models, and all performed with good accuracy over 80%.



Confusion Matrix

- Here is the KNN model's confusion matrix.
- It was able to predict 12 successful landings and 3 failed landings correctly.
- It classified 3 failed landings as successful but no successful landings as failed, which points to the fact that it is biased towards successful landings.



Conclusions

- The outcome of the first stage landing can be predicted with high accuracy from the given data and generated models.
- Among all the different models, all models performed similarly with high accuracy on the test set and similar accuracies on the training set.
- Since the data is made out of mostly successful landings, the models were biased to predict a landing as successful rather than failed.

Thank you!