IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Younes Mahdavi
Apr 19, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.
- Summary of all results
  - Successful landings occurred for Payload Masses under 8000 kg. More precisely, more successful landings occurred for Payload Masses under 7000 kg at GTO and ISS launch sites
  - Four orbits of ES-L1, GEO, HEO, and SSO acquired the highest success rate among other orbits.
  - The success rate kept increasing from 2013 until 2020; however, it had a slight drop in 2018.
  - The total payload mass launched by NASA was 107010 kg.
  - The first successful landing outcome on ground pad occurred on the 4th of June 2010.
  - All predicting models had high accuracy, however, the Decisions Tree model gained the highest accuracy of % 94 among other models.

Introduction

- Project background and context
- Problems you want to find answers

  This project aims to predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Raw data was acquired from Space X API in a JSON format and converted into a Data Frame.
- Perform data wrangling
  - A lot of the data were IDs or links to external resources and websites like YouTube or Wikipedia. We cleaned that raw data by removing unnecessary columns in the initial Data Frame by focusing on columns “rocket”, “payloads”, “launchpad”, and “cores” and those related to these features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook:

  https://github.com/Mahdavi-y/YM-testrepo-Data-Scince-Capstone/blob/614812116a4db0eaf4ddbe2467854a8518bfa769/SpaceX%20Falcon%209%20first%20stage%20Landing%20Prediction-collection-api.ipynb

requesting rocket launch data

SpaceX API from https://api.spacexdata.com/v4

↓

Requesting and parsing JSON file of the SpaceX launch data using the GET request

↓

Using columns, **rocket**, **payloads**, **launchpad** and **cores**.

↓

Filtering the dataframe to only include Falcon 9 launches

↓

Dealing with Missing Values and replacing them with mean values.

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook:

  https://github.com/Mahdavi-y/YM-testrepo-Data-Scince-Capstone/blob/7df61120ec29a5b6b950c715b9a50ab964d0f388/Space%20X%20Falcon%209%20First%20Stage%20Landing%20Prediction_webscraping.ipynb

Extracting a Falcon 9 launch records HTML table from Wikipedia from its URL

↓

Parsing the table and converting it into a Pandas data frame using BeautifulSoup

↓

Extracting all column/variable names from the HTML table header

↓

Creating a data frame by parsing the launch HTML tables

Data Wrangling

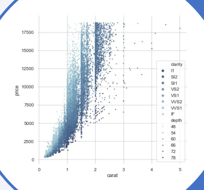
- In this lab, we performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- GitHub URL of completed data wrangling notebooks:
  https://github.com/Mahdavi-y/YM-testrepo-Data-Scince-Capstone/blob/8d6651b54575db82f0e56a75d86541c9045c2b17/Space%20X%20Falcon%209%20First%20Stage%20Landing%20Prediction-Data%20wrangling.ipynb

1. Loading Space X dataset and Calculating the number of launches on each site
2. Calculating the number and occurrence of each orbit
3. Calculating the number and occurrence of mission outcome of the orbits
4. Creating a landing outcome label from Outcome column ("landing class")

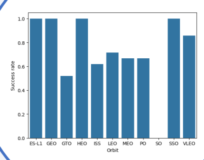
EDA with Data Visualization

- Scatter plots, Bar charts, and Line plots were used to show the effect of multiple features on the launch outcome.
- GitHub URL of completed EDA with data visualization notebook:

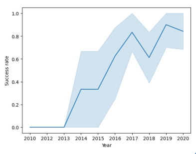
  https://github.com/Mahdavi-y/YM-testrepo-Data-Scince-Capstone/blob/b6313fd63de441b6307fda5294ea1a5d54b89f85/SpaceX%20Falcon%209%20First%20Stage%20Landing%20Prediction_EDA%20with%20Data%20Visualization.ipynb



Scatter plots were used to show the effect of multiple features on the launch outcome.



A bar plot shows the success rate of each orbit.



The line plot shows the launch success yearly trend.

EDA with SQL

- SQL queries to display and list:
  - the names of the unique launch sites in the space mission;
  - five records where launch sites begin with the string 'CCA';
  - the date when the first successful landing outcome in the ground pad was achieved;
  - the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;
  - the total number of successful and failed mission outcomes;
  - the names of the booster versions which have carried the maximum payload mass;
  - the records which will display the month names, failure landing outcomes in drone ship, booster versions, and launch site for the months in the year 2015;
  - the count of landing outcomes (Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order.
- GitHub URL of the completed EDA with SQL notebook:

  https://github.com/Mahdavi-y/YM-testrepo-Data-Scince-Capstone/blob/5d2060ba88f3121c43171f2c93468ea77c163a7b/SQL%20Notebook%20for%20Peer%20Assignment%20-%2010%20Tasks.ipynb

Build an Interactive Map with Folium

- We used markers, circles, lines, Marker Clusters, Mouse Position, etc. to display coordinates with relevant popup names.
- We added these objects to:
  - Mark all launch sites on a map
  - Mark the successful/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities
- GitHub URL of the completed interactive map with Folium map:
  - https://github.com/Mahdavi-y/YM-testrepo-Data-Scince-Capstone/blob/05ff1c30216f2e240c817daf8d08f85afd935b0f/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- We used a pie chart to show the total successful launches count for each site, and a scatter chart to show the correlation between payload and launch success for each site.
- The chart aimed to enable users to perform interactive visual analytics on SpaceX launch data in real time.
- GitHub URL of the completed Plotly Dash lab:

  https://github.com/Mahdavi-y/YM-testrepo-Data-Scince-Capstone/blob/e970156252ad582220f655e27998f68a8ea32bba/spacex_dash_app.py

Predictive Analysis (Classification)

- First, we created a column for the class that shows the launch outcome, then we Standardized the data and split it into training data and test data to find the best Hyperparameter for SVM, Classification Trees, and Logistic Regression. Finally, we found the method performs best using test data.
- GitHub URL of the completed predictive analysis lab:
  - https://github.com/Mahdavi-y/YM-testrepo-Data-Scince-Capstone/blob/7be9c69f42a029a2397bb7fc8731758acbcf7fa9/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

1. Load the dataframe and create a NumPy array from column **Class**
2. Standardize the data in dataframe
3. split the data into training and testing data using the function **train_test_split**
4. hyperparameters are selected using the function **GridSearchCV**
5. Create a **Logistic Regression** object and output the **GridSearchCV** object for that and fit the data.
6. Calculate the **best_params_** and **best_score_** and the accuracy on the test data the **score** method and display the Confusion Matrix.
7. Create an **SVM**, a **Classification Tree**, and a **KNN** object and apply numbers 5 and 6 to them.
8. By comparing the accuracy of each method, we will find the best one.

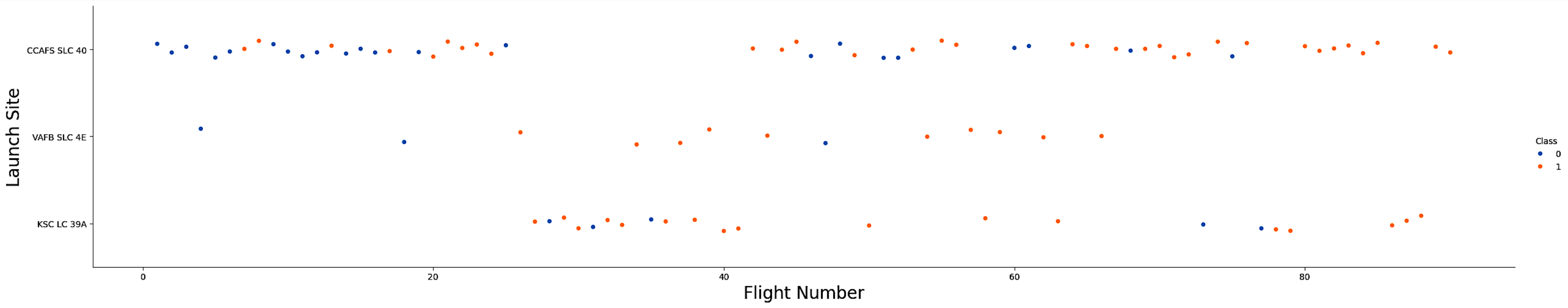
Results

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

Section 2

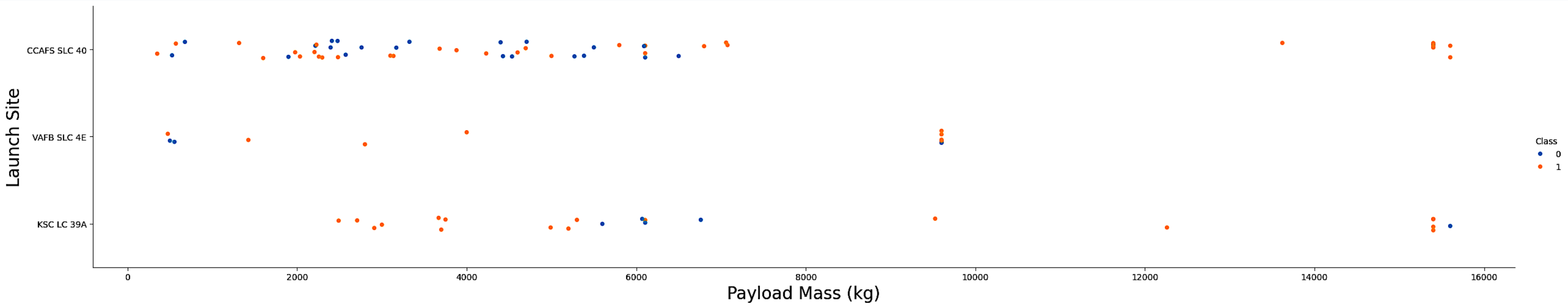
Insights drawn from EDA

Flight Number vs. Launch Site



- The chart shows that successful landings of Falcon 9 first stage, belong to CCAFS SLC 40 launch site for flight numbers below 27.

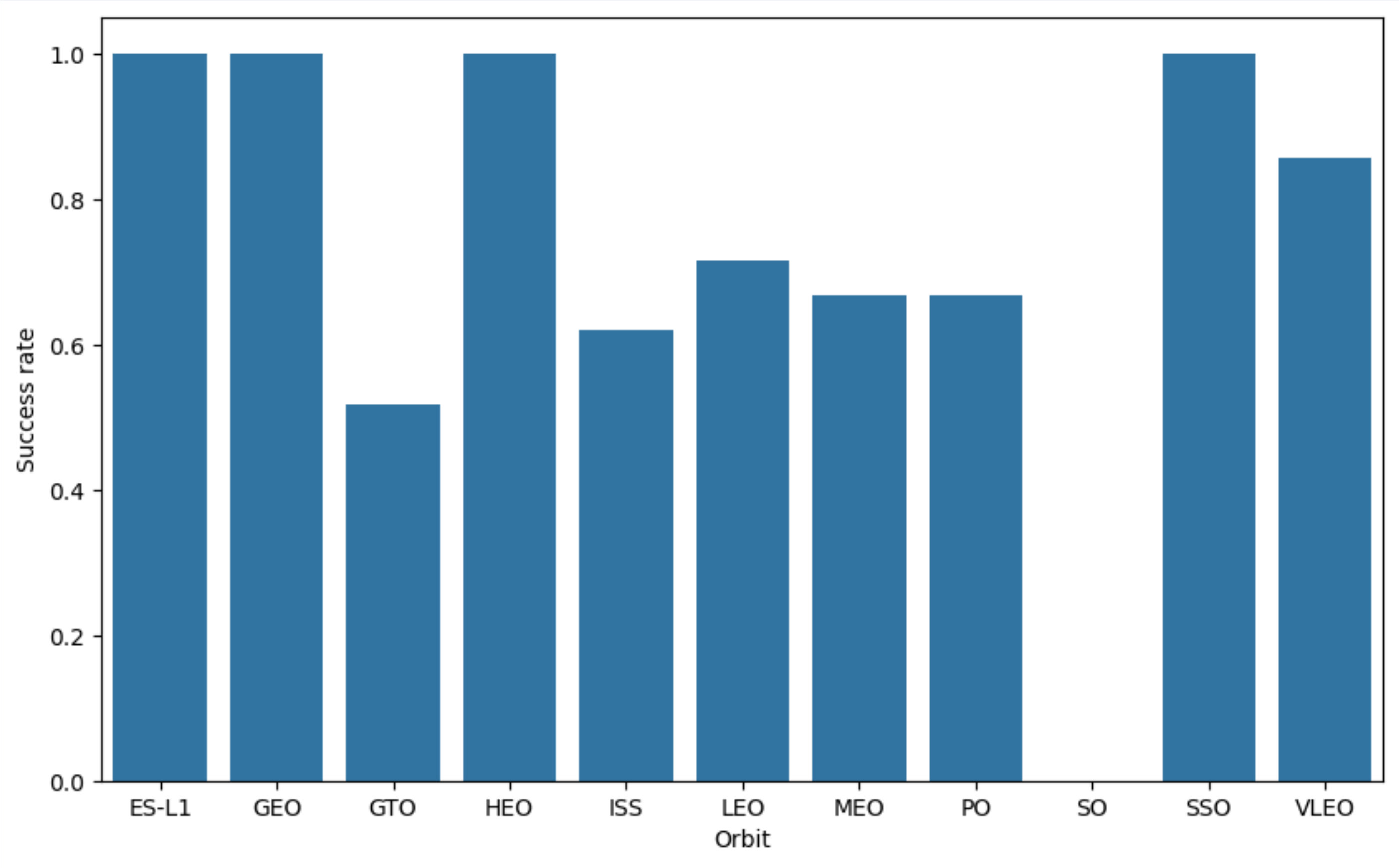
Payload vs. Launch Site



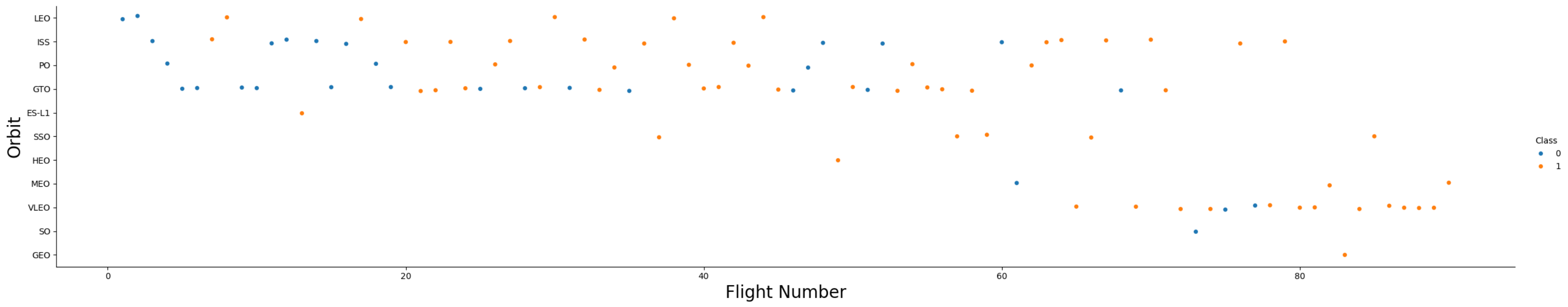
- Successful landings occurred for Payload Masses under 8000 kg at the CCAFS SLC 40 launch site.

Success Rate vs. Orbit Type

- Four orbits of ES-L1, GEO, HEO, and SSO acquired the highest success rate among other orbits.

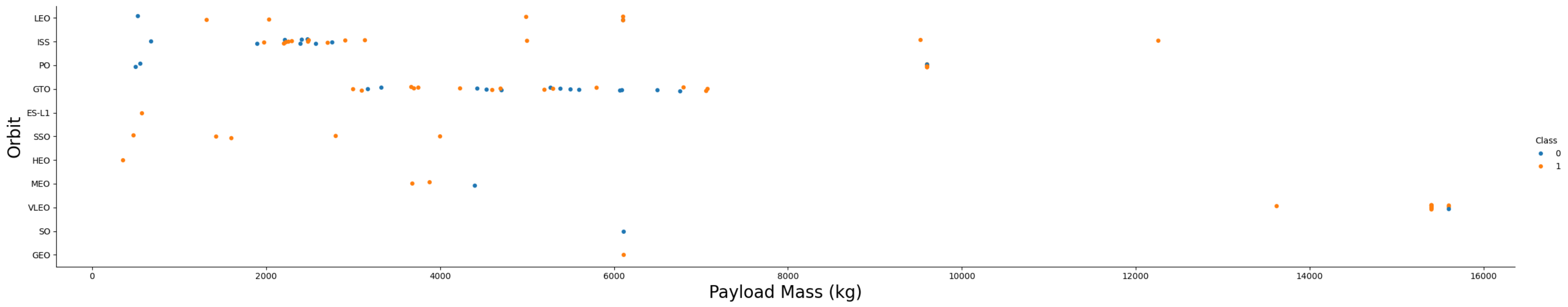


Flight Number vs. Orbit Type



- The plot shows that a high number of flights belong to GTO, ISS, and VLEO launch sites respectively, while ES-L1, GEO, and SO had only one flight.

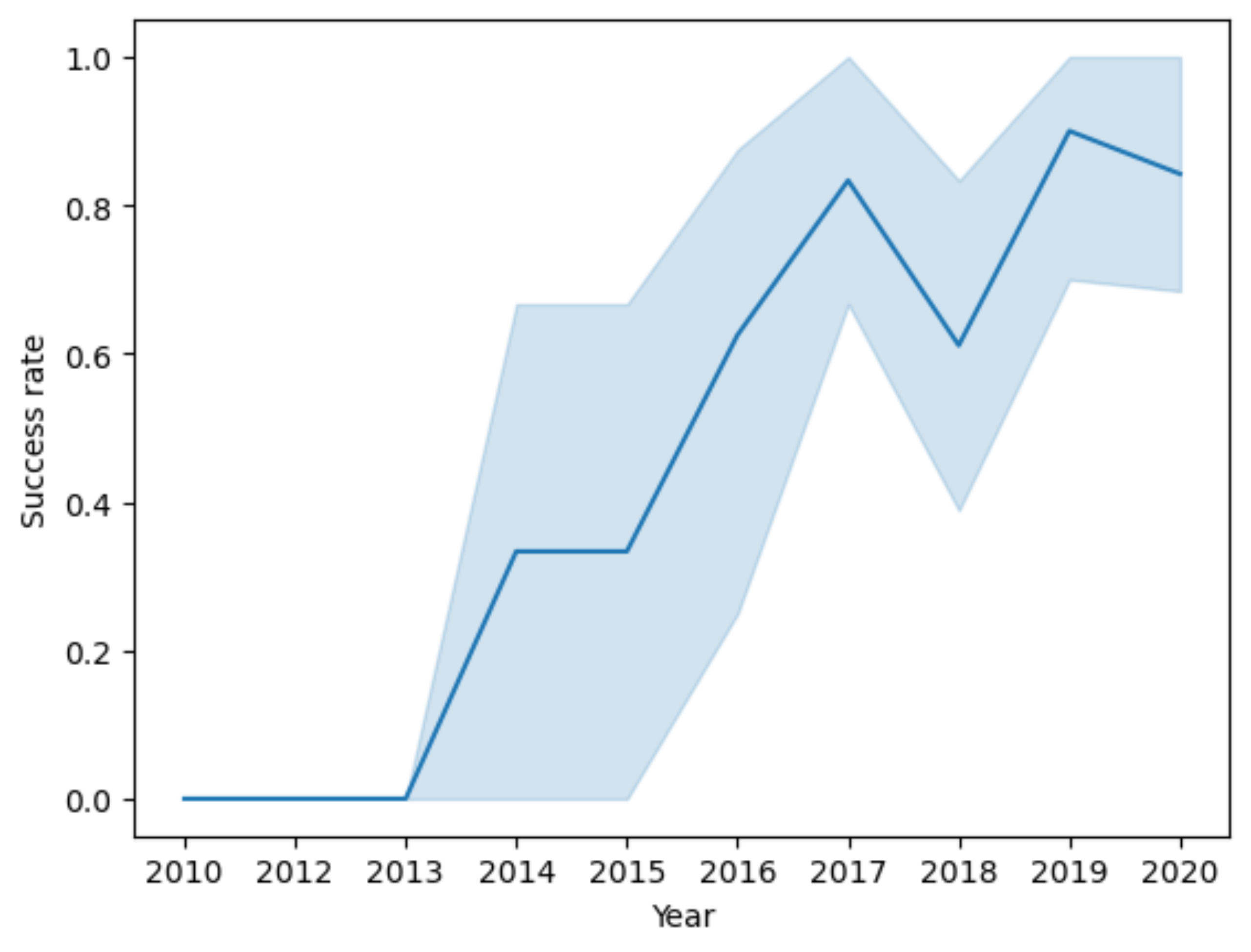
Payload vs. Orbit Type



- The plot shows more successful landings occurred for Payload Masses under 7000 kg at GTO and ISS launch sites.

Launch Success Yearly Trend

- It is obvious that the success rate kept increasing from 2013 until 2020; however, it had a slight drop in 2018.



All Launch Site Names

- Names of the launch sites:
  - CCAFS
  - LC-40VAFB
  - SLC-4EKSC
  - LC-39ACCAFS
  - SLC-40
- There are 5 unique launch sites.

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

- The first five launch site names beginning with ‘CCA’ gained successful outcomes.

Total Payload Mass

Totol Payload Mass	Customer
107010	NASA (COTS) NRO

- The total payload mass launched by NASA was 107010 kg.

Average Payload Mass by F9 v1.1

Average Payload Mass	Booster_Version
2534.67	F9 v1.1 B1003

- The average payload mass carried by booster version F9 v1.1 was 2534.67 kg.

First Successful Ground Landing Date

Landing Date	Mission_Outcome
2010-06-04	Success

- The first successful landing outcome on ground pad occurred on the 4th of June 2010.

Successful Drone Ship Landing with Payload between 4000 and 6000

- There were 23 successful Drone ship landing with payload mass between 4000 and 6000.

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535
F9 v1.1 B1011	4428
F9 v1.1 B1014	4159
F9 v1.1 B1016	4707
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1030	5600
F9 FT B1021.2	5300
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 FT B1031.2	5200

Booster_Version	PAYLOAD_MASS__KG_
F9 B4 B1043.1	5000
F9 FT B1032.2	4230
F9 B4 B1040.2	5384
F9 B5 B1046.2	5800
F9 B5 B1047.2	5300
F9 B5B1054	4400
F9 B5 B1048.3	4850
F9 B5 B1051.2	4200
F9 B5B1060.1	4311
F9 B5 B1058.2	5500
F9 B5B1062.1	4311

Total Number of Successful and Failure Mission Outcomes

Number of Missions	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

- There are 101 successful missions and one failed mission.

Boosters Carried Maximum Payload

- There are 12 boosters that carried the maximum payload mass of 15600 kg.

Booster_Version	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

- The table shows that there are 10 failed landing outcomes along with booster versions and launch site names for in the year 2015.

Month	Landing_Outcome	Booster_Version	Launch_Site
06	Failure (parachute)	F9 v1.0 B0003	CCAFS LC-40
12	Failure (parachute)	F9 v1.0 B0004	CCAFS LC-40
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
01	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
03	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
06	Failure (drone ship)	F9 FT B1024	CCAFS LC-40
12	Failure	F9 B5B1050	CCAFS SLC-40
02	Failure	F9 B5 B1056.4	CCAFS SLC-40
03	Failure	F9 B5 B1048.5	KSC LC-39A

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

- The count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order. More successful landing belongs to 2016 and more failed ones in 2012.

Date	Count	Landing_Outcome
2016-04-08	5	Success (drone ship)
2015-12-22	3	Success (ground pad)
2015-06-28	1	Precluded (drone ship)
2015-01-10	5	Failure (drone ship)
2014-04-18	3	Controlled (ocean)
2013-09-29	2	Uncontrolled (ocean)
2012-05-22	10	No attempt
2010-06-04	2	Failure (parachute)

Section 3

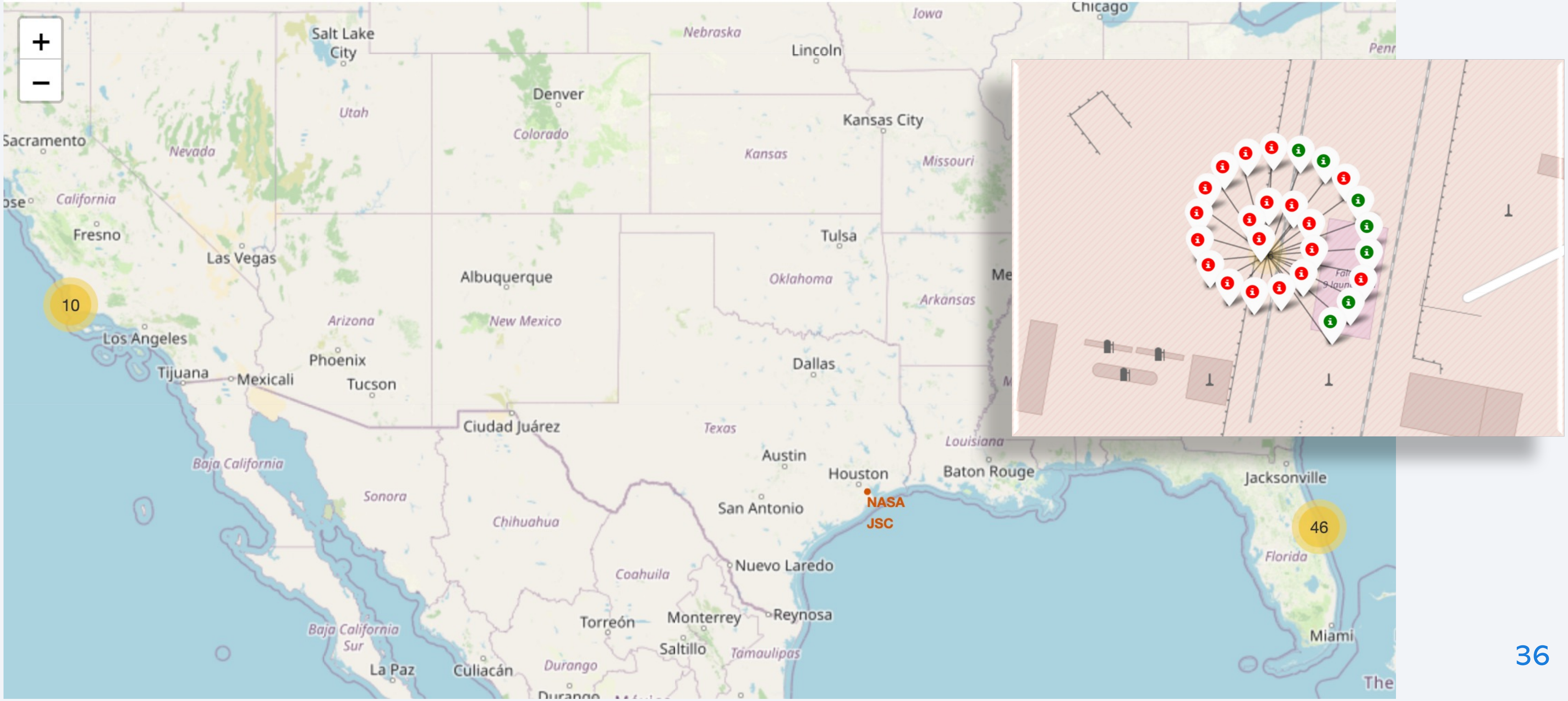
Launch Sites Proximities Analysis

Map of all launch sites

- All launch sites are in the states of California and Florida near the coastline areas.

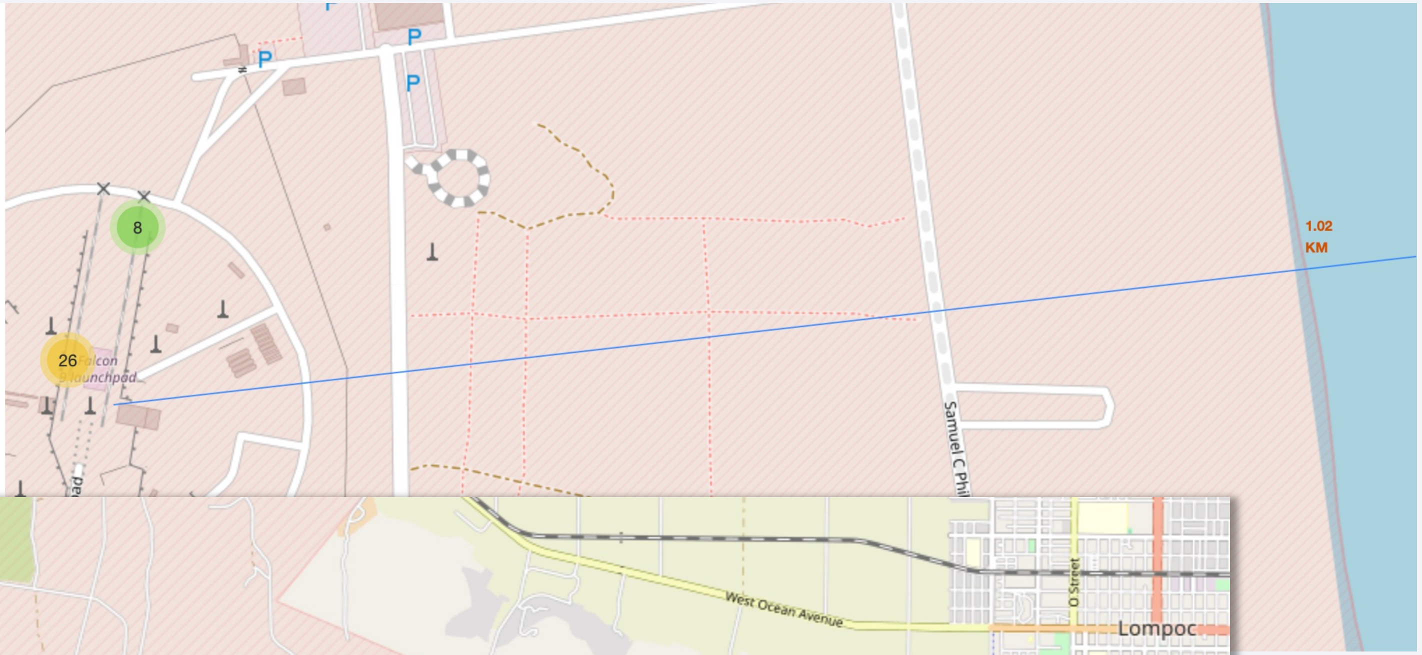
Map of launch sites marked and colored for success and failure

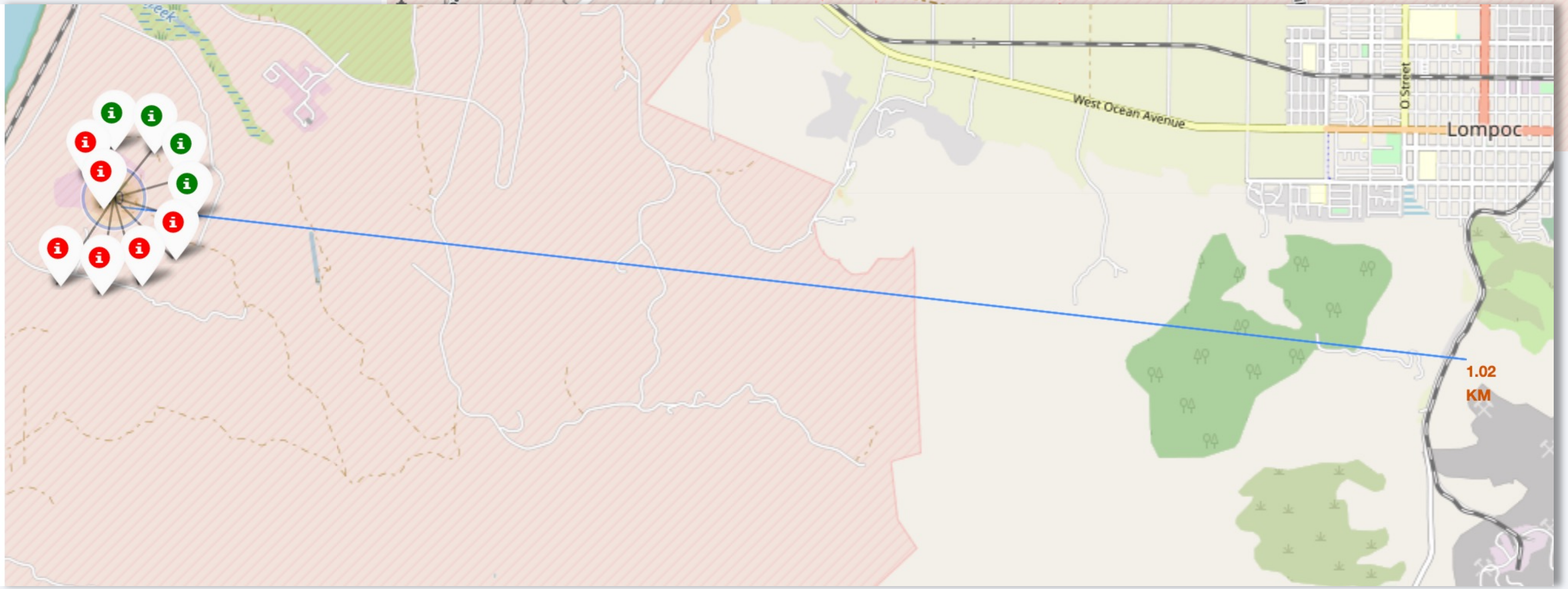
- All launch sites on the map are marked by their successful and failed landings.



Exploring Launch Sites Proximities

- The map displays locations like railway, highway, and coastline, with distance calculated for each to two of the launch sites. This means there are railways and coastlines in the proximity of the launch site as close as 1 km.

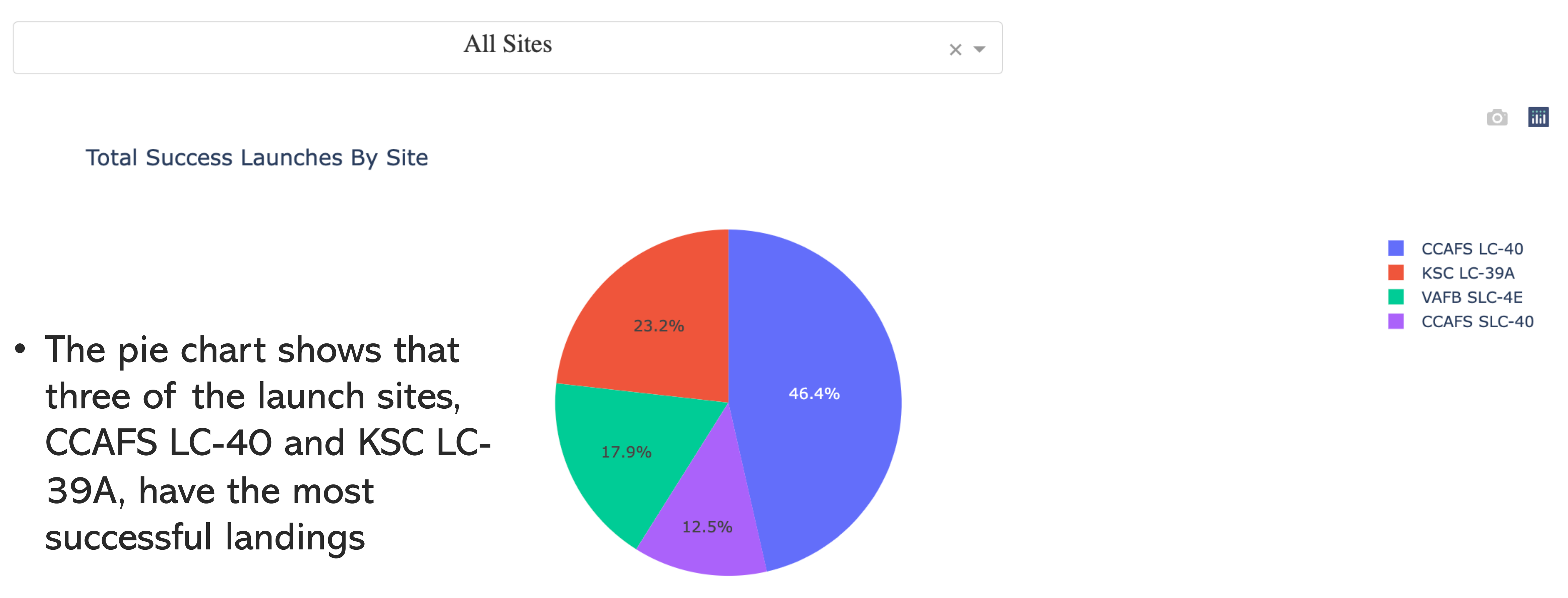




Section 4

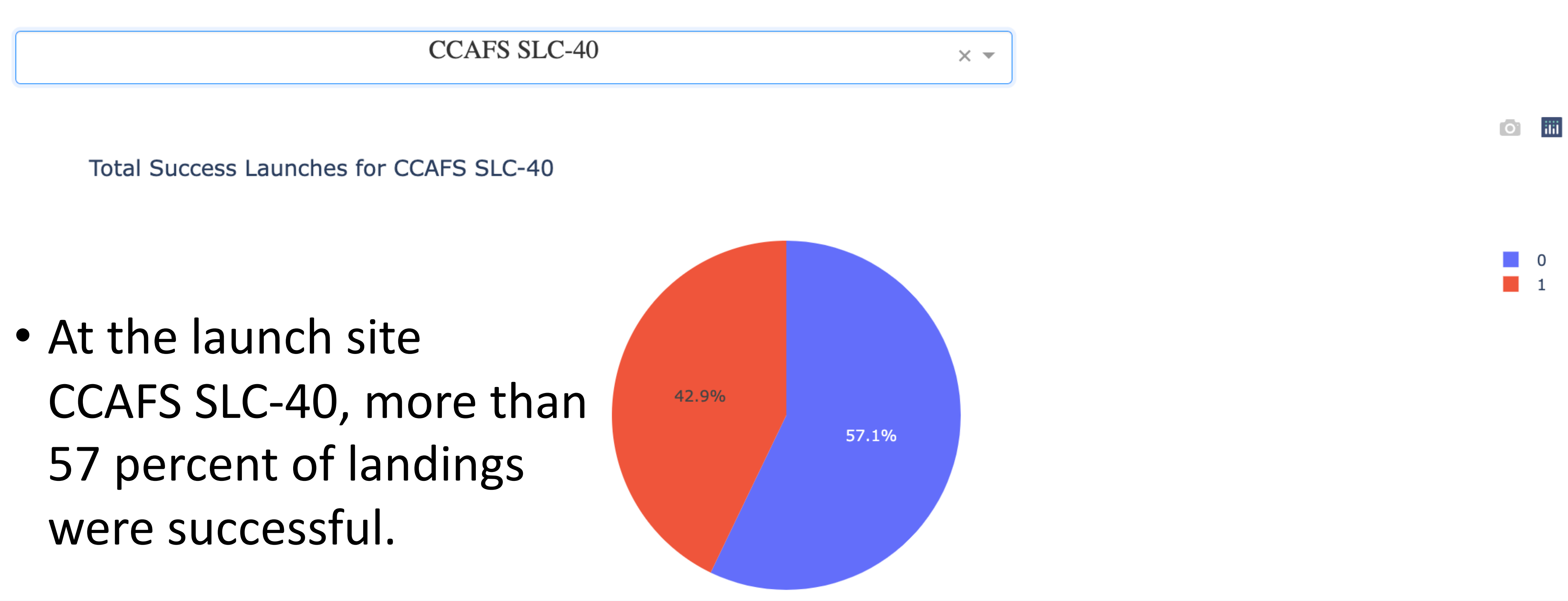
Build a Dashboard with Plotly Dash

Success Launches for all Sites



- The pie chart shows that three of the launch sites, CCAFS LC-40 and KSC LC-39A, have the most successful landings

Launch Site with the Highest Success Rate



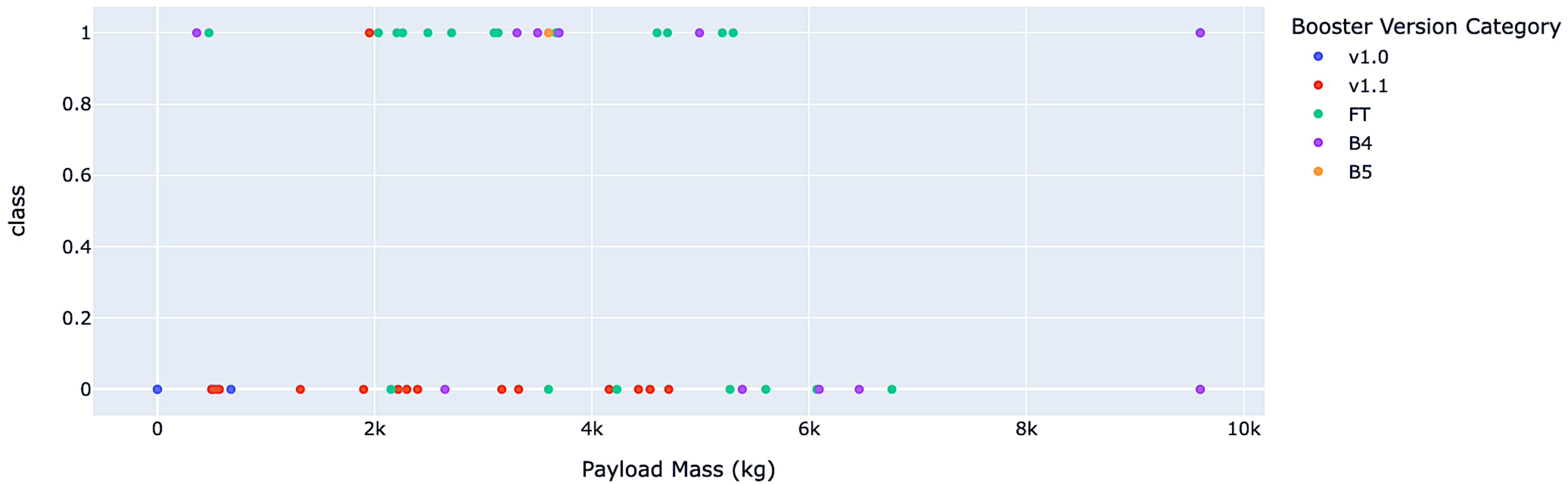
- At the launch site CCAFS SLC-40, more than 57 percent of landings were successful.

<Dashboard Screenshot 3>

Payload range (Kg):

0100

- For the payload mass between 2000 kg and 6000 kg, more successful landings were obtained.

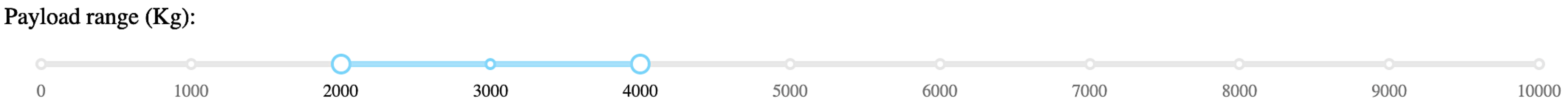


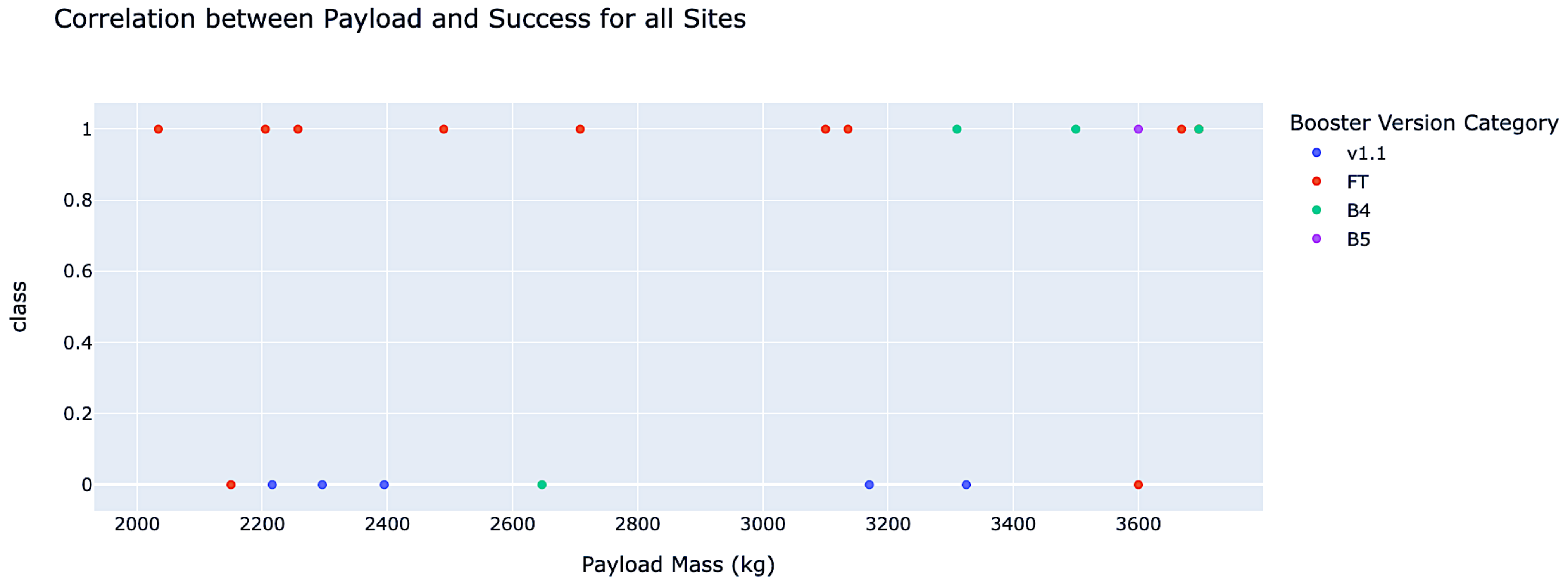
Payload vs. Launch Outcome for all Sites

The range slide shows that most of the successful landings occurred for the payloads between 2000 kg and 8000 kg.

Payload vs. Launch Outcome for all Sites

For the payloads between 2000 kg and 400 kg, most of the success landings occurred for FT and v1.1 Boosters.

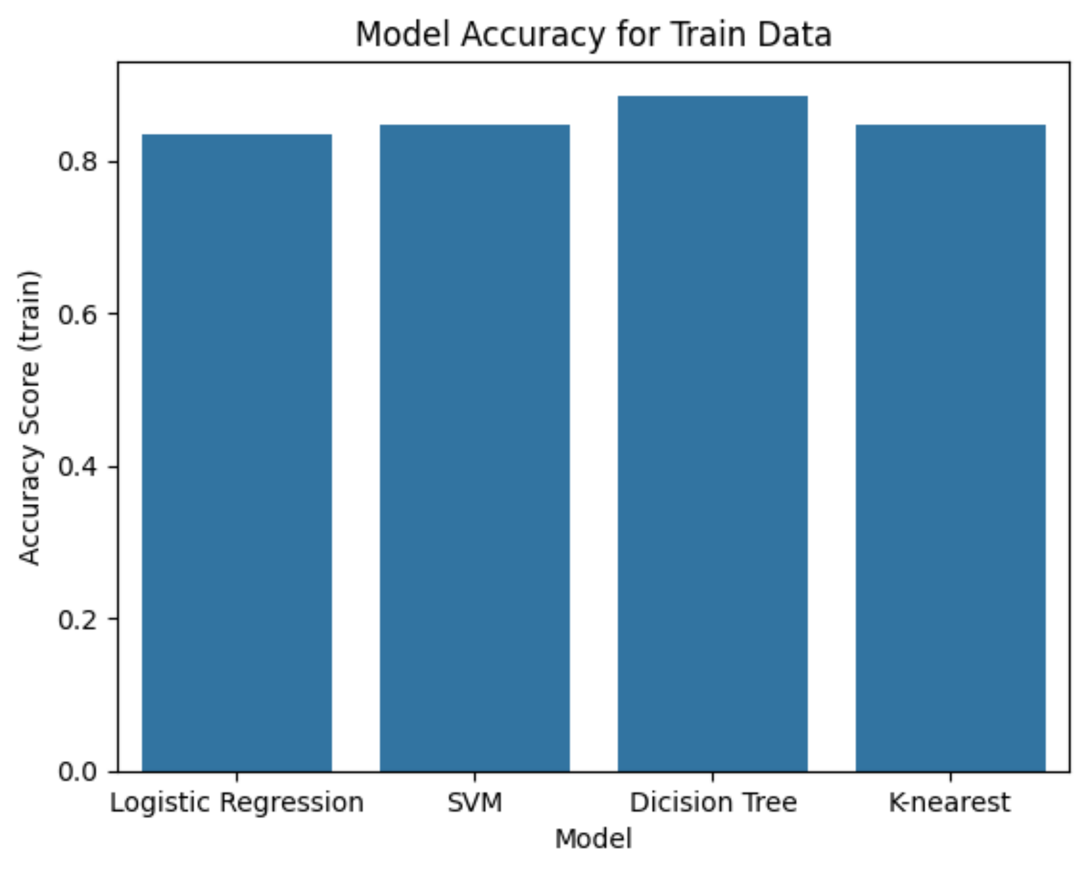


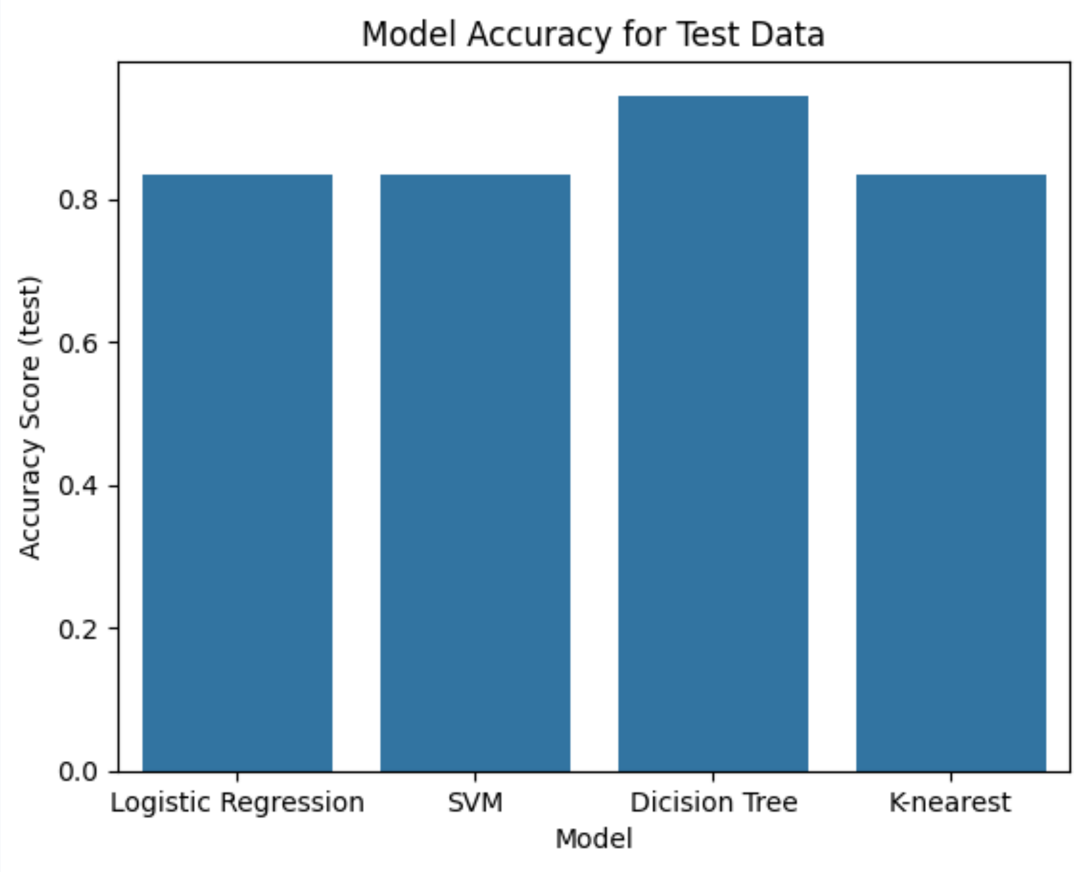


Section 5

Predictive Analysis (Classification)

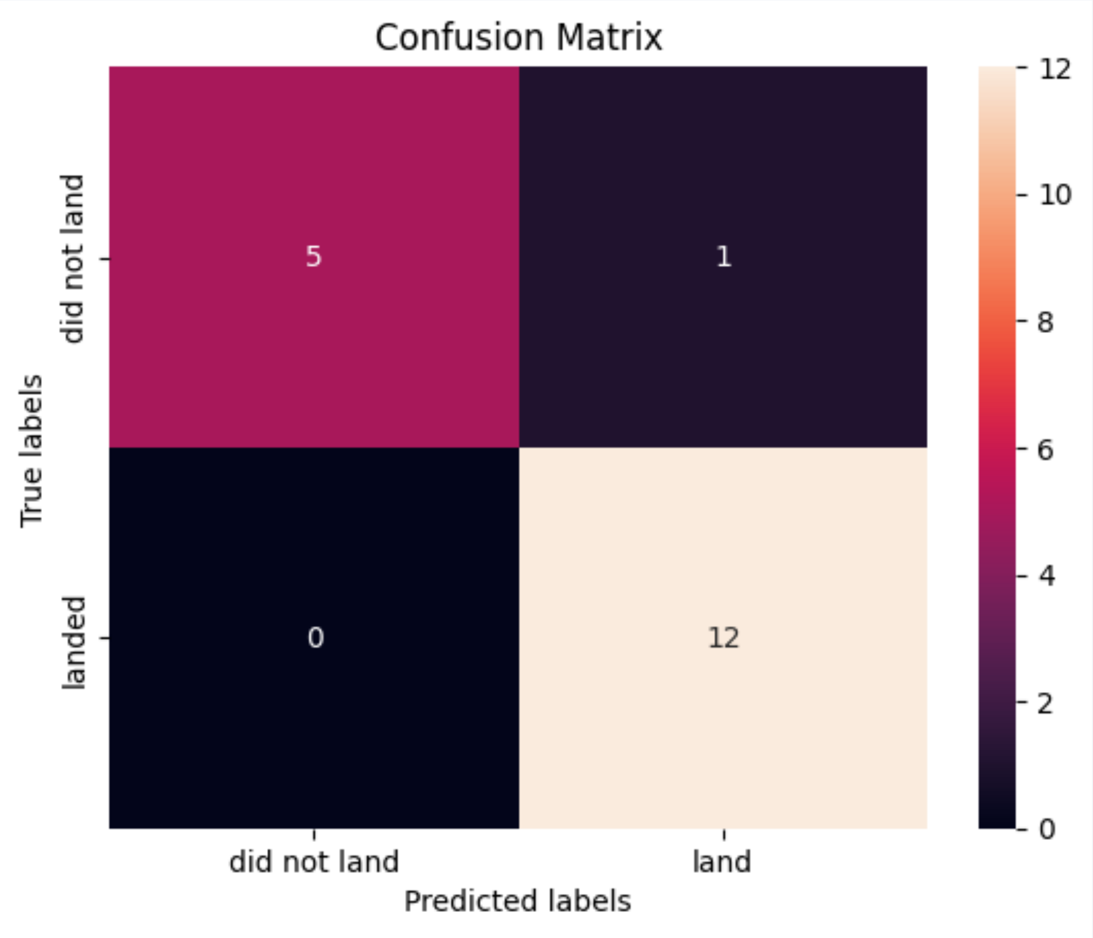
Classification Accuracy





Confusion Matrix of Classification Tree Model

- This model has the highest accuracy score and as the Confusion Matrix shows, it predicted all successful landings correctly and only missed only one out of six failed landings to predict correctly.



Conclusions

- High Payload Mass would not obtain a successful landing in many cases. Most of the successful outcomes occurred for Payload Masses under 8000 kg, particularly for Payload Masses under 7000 kg at GTO and ISS launch sites.
- The first successful landing outcome on ground pad occurred on the 4th of June 2010.
- The success rate kept increasing from 2013 until 2020; however, it had a slight drop in 2018.
- All predicting models had high accuracy, however, the Decisions Tree model gained the highest accuracy of % 94 among other models. So, for this project, it is recommended to implement more of Decision Tree models for accurate predictions.

Appendix I

- Callback function for correlation between Payload and Success Landings

```
@app.callback(
    Output(component_id='success-payload-scatter-chart', component_property='figure'),
    [Input(component_id='site-dropdown', component_property='value'),
     Input(component_id="payload-slider", component_property="value")])

def success_payload_scatter_chart(entered_site,payload):
    filtered_df = spacex_df
    new_df = filtered_df[(filtered_df['Payload Mass (kg)'] >= payload[0]) & (filtered_df['Payload Mass (kg)'] <= payload[1])]
    if entered_site == 'ALL':
        fig = px.scatter(new_df, x="Payload Mass (kg)",y="class", color="Booster Version Category",
        title='Correlation between Payload and Success for all Sites')
        return fig
    else:
        fig = px.scatter(new_df[new_df['Launch Site']==entered_site],
                         x="Payload Mass (kg)",y="class", color="Booster Version Category",
        title= str('Correlation between Payload and Success for ' + entered_site))
        return fig
```

Appendix II

- Accuracy Scores of all classification methods

```
Report = {'Accuracy Score (train)':[logreg_cv.best_score_,svm_cv.best_score_,tree_cv.best_score_,knn_cv.best_score_],
          'Accuracy Score (test)':[logreg_cv.score(X_test,Y_test),svm_cv.score(X_test,Y_test),tree_cv.score(X_test,Y_test)
                                   ,knn_cv.score(X_test,Y_test)],
         }

Report_df = pd.DataFrame(Report,index=(['Logistic Regression','SVM','Dicision Tree','K-nearest']))
Report_df
```

	Accuracy Score (train)	Accuracy Score (test)
Logistic Regression	0.834286	0.833333
SVM	0.848214	0.833333
Dicision Tree	0.885714	0.944444
K-nearest	0.848214	0.833333

Thank you!