IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Neerav Gala
04/11/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data to create success/fail outcome variable
- Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- Analyze the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- Explore launch site success rates and proximity to geographical markers
- Visualize the launch sites with the most success and successful payload ranges
- Build Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

Introduction

Project background and context

We predicted if the Falcon 9 first stage will land successfully. 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Problems you want to find answers

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)

Methodology

Executive Summary

- **Data Collections:** Using SpaceX REST API and web scraping techniques
- **Data Wrangling:** By filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- **EDA** with SQL and data visualization techniques
- Visualize the data using Folium and Plotly Dash
- Build Models to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection – SpaceX API

- Request data from SpaceX API (rocket launch data)
- Decode response using .json() and convert to a dataframe using .json_normalize()
- Using Custom functions request information about the launches from SpaceX API
- From the collected data, create a dictonary and then a dataframe
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass with calculated .mean()
- Export data to csv file

Github URL

Data Collection - Scraping

- Request Falcon 9 launch data from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from HTML table header
- From the collected data, create a dictonary and then a dataframe
- Create dataframe from the dictionary
- Export data to csv file

Github URL

Data Wrangling

- The type (numerical or categorical) of attribute is determined
- The number of launches in each site is listed:
  - CCAFS SLC 40 : 55
  - KSC LC 39A : 22
  - VAFB SLC 4E : 13
- Create a landing outcome variable (dependent variable) based on the landing outcomes of the first stage:
  - True Ocean means the mission outcome was successfully landed to a specific region of the ocean: 1
  - False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean: 0
  - True RTLS means the mission outcome was successfully landed to a ground pad : 1
  - False RTLS means the mission outcome was unsuccessfully landed to a ground pad: 0
  - True ASDS means the mission outcome was successfully landed to a drone ship: 1
  - False ASDS means the mission outcome was unsuccessfully landed to a drone ship: 0
  - None ASDS and None None these represent a failure to land: 0
- Export data as a csv file

Github URL

EDA with Data Visualization

- Scatter Plots
  - Flight Number VS. Payload Mass
  - Flight number vs Launch Site
  - Payload vs Launch Site
  - Flight Number vs Orbit type
  - Payload vs Orbit type

Scatter plots show how much one variable is affected by another

- Bar Plots
  - Success rate of each orbit type

A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes.

- Line Plot
  - Launch success over the years

Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded

Github URL

EDA with SQL

Performed SQL queries to gather the following information about the dataset:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Github URL

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates

Colored Markers of Launch Outcomes

- Colored markers of successful (green) and unsuccessful (red)
- Launches at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Added colored lines to show distance between launch site CCAFS SLC- 40 and its proximity to the nearest coastline, railway, highway, and city

Github URL

Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites to allow user to select all launch sites or a certain launch site
- Pie Chart Showing Successful Launches to allow user to see successful and unsuccessful launches as a percent of the total
- Slider of Payload Mass Range to allow user to select payload mass range
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version allow user to see the correlation between Payload and Launch Success

Github URL

Predictive Analysis (Classification)

BUILDING MODEL

- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset

EVALUATING MODEL

- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix

Github URL

IMPROVING MODEL

- Feature Engineering
- Algorithm Tuning

FINDING THE BEST PERFORMING CLASSIFICATION MODEL

- The model with the best accuracy score wins the best performing model
- In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook

Results

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

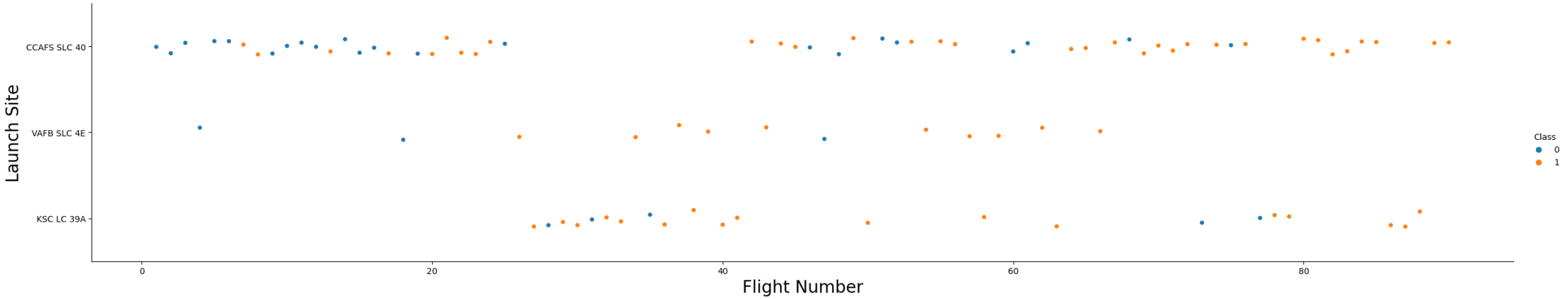
- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

- Logistic model is the best predictive model for the dataset

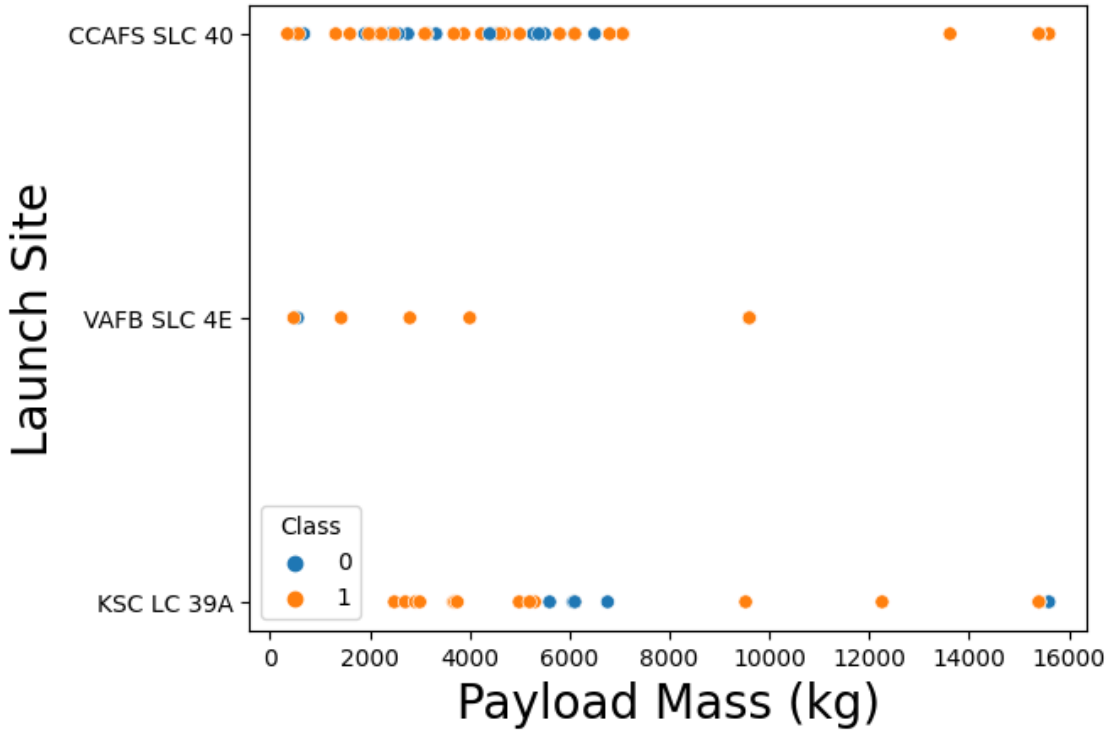
Flight Number vs. Launch Site

- We can infer that new launches have a higher success rate
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates



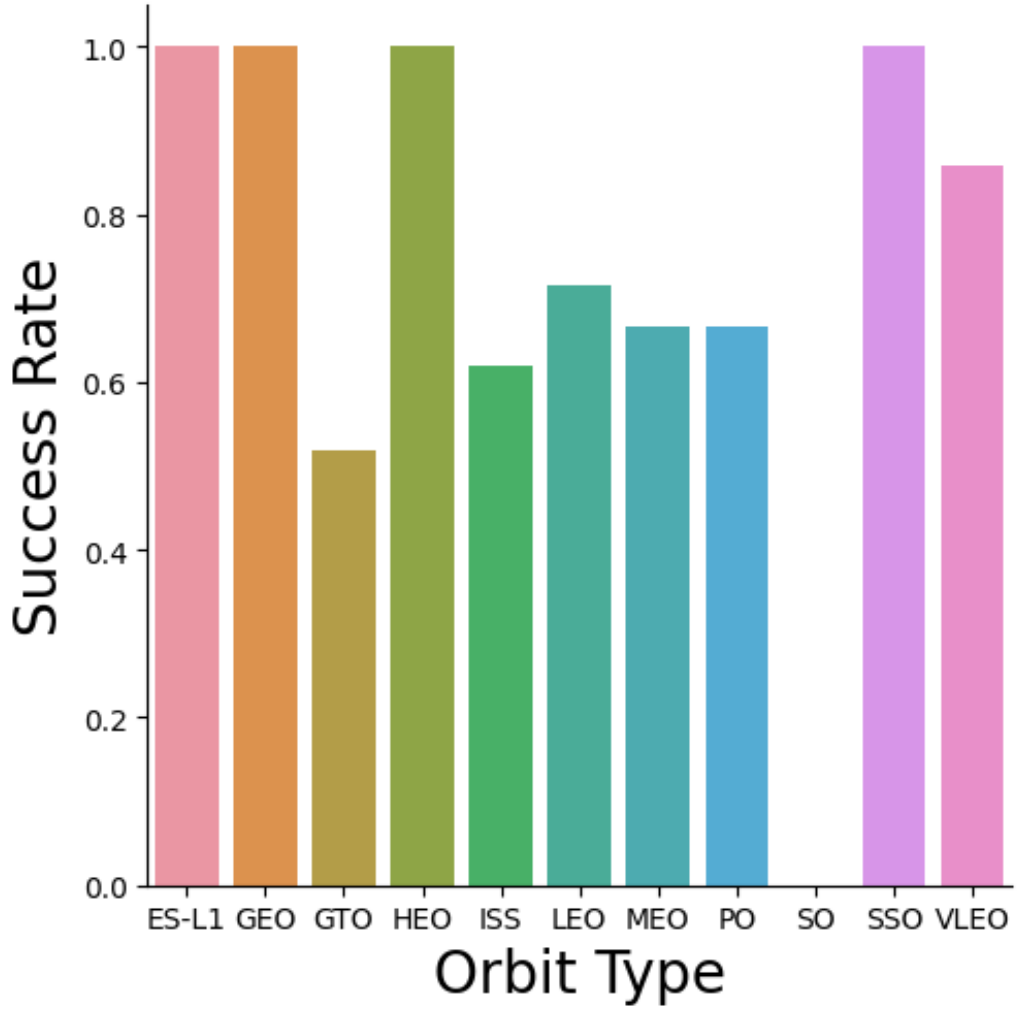
Payload vs. Launch Site

- We can infer that higher the payload, higher the success rate
- VAFB SLC 4E and KSC LC 39A has not launched a rocket with a payload greater than ~10,000 kgs
- KSC LC 39A has high success rates with low payloads



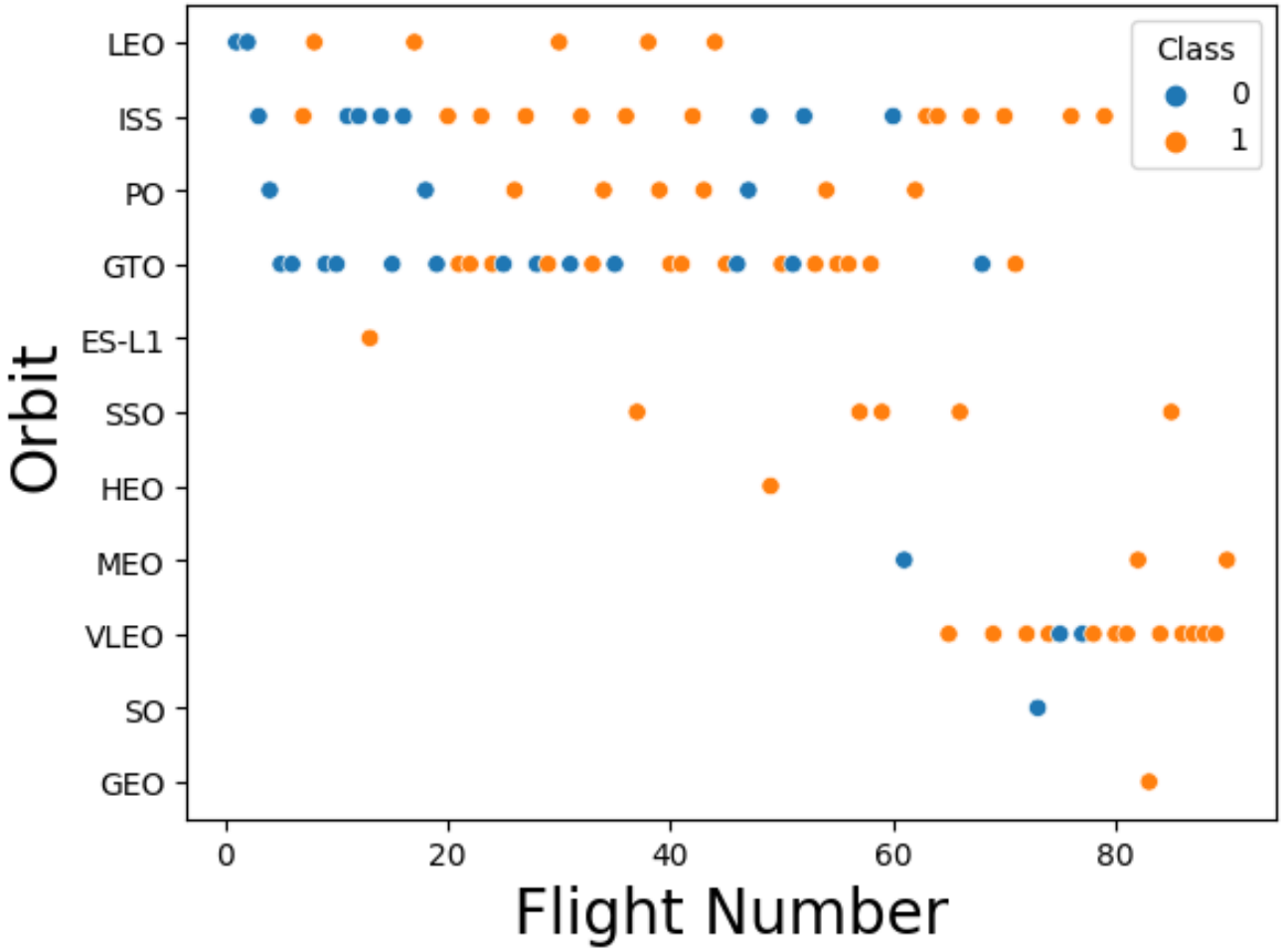
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have 100% success rates
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO have success rates between 50%-80%
- SO did not have a successful launch



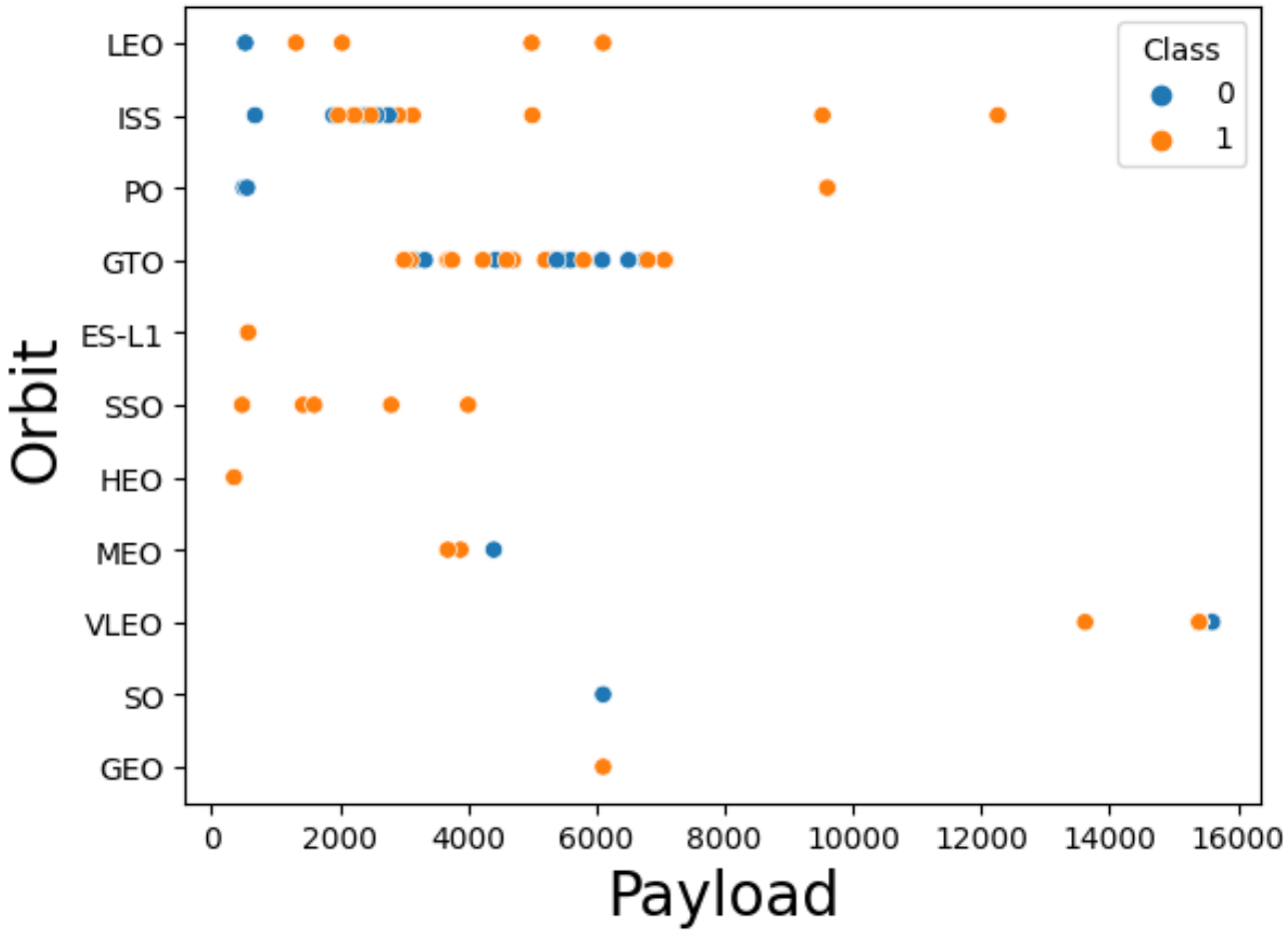
Flight Number vs. Orbit Type

- In each orbit type, generally, the chance of a success increases with flight number



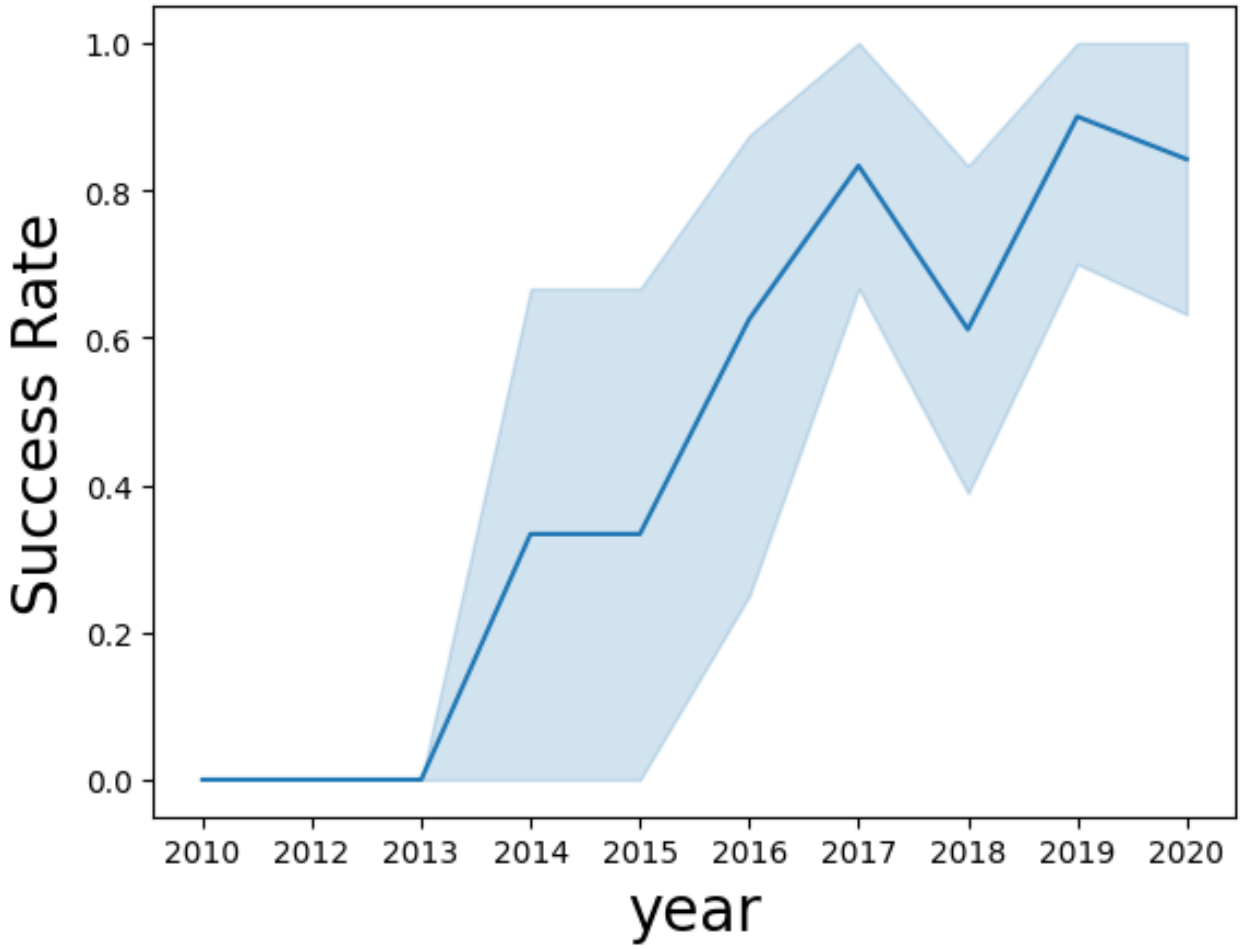
Payload vs. Orbit Type

- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

- Overall, the success rate has improved since 2013



All Launch Site Names

```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5
```

[12]

```
 * sqlite:///my_data1.db
Done.
```

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

+ Code + Markdown

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS) was 45,496 kgs

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 was 2,534.67 kg

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like "F9 v1.1%"
```

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was on 2015-12-22

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome like "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

```
%%sql select Booster_Version
from SPACEXTABLE
where
PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
AND
Landing_Outcome like "Success (drone ship)"
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Mission outcomes
  - Success : 99
  - Failure : 1
  - Success (payload status unclear) : 1

```
%%sql
select
Mission_Outcome, count(*) as totals
from SPACEXTABLE
GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

```
%%sql
SELECT
DISTINCT(Booster_Version)
FROM
SPACEXTABLE
WHERE
PAYLOAD_MASS__KG_ =
(SELECT
max(PAYLOAD_MASS__KG_)
FROM
SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

- List of the failed landing_outcomes in drone ship, their booster versions, and launch site names in year 2015

```
%%sql
SELECT
substr(DATE, 6, 2) as Month,
Landing_Outcome,
Booster_Version,
Launch_Site
from SPACEXTABLE
where
Landing_Outcome like "Failure (drone ship)"
AND
substr(DATE, 1, 4) like "2015"
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
SELECT
Landing_Outcome,
count(Landing_Outcome) as Count
from SPACEXTABLE
WHERE strftime('%Y-%m-%d', DATE) BETWEEN "2010-06-04" AND "2017-03-20"
GROUP BY
Landing_Outcome
Order by
Count DESC
```

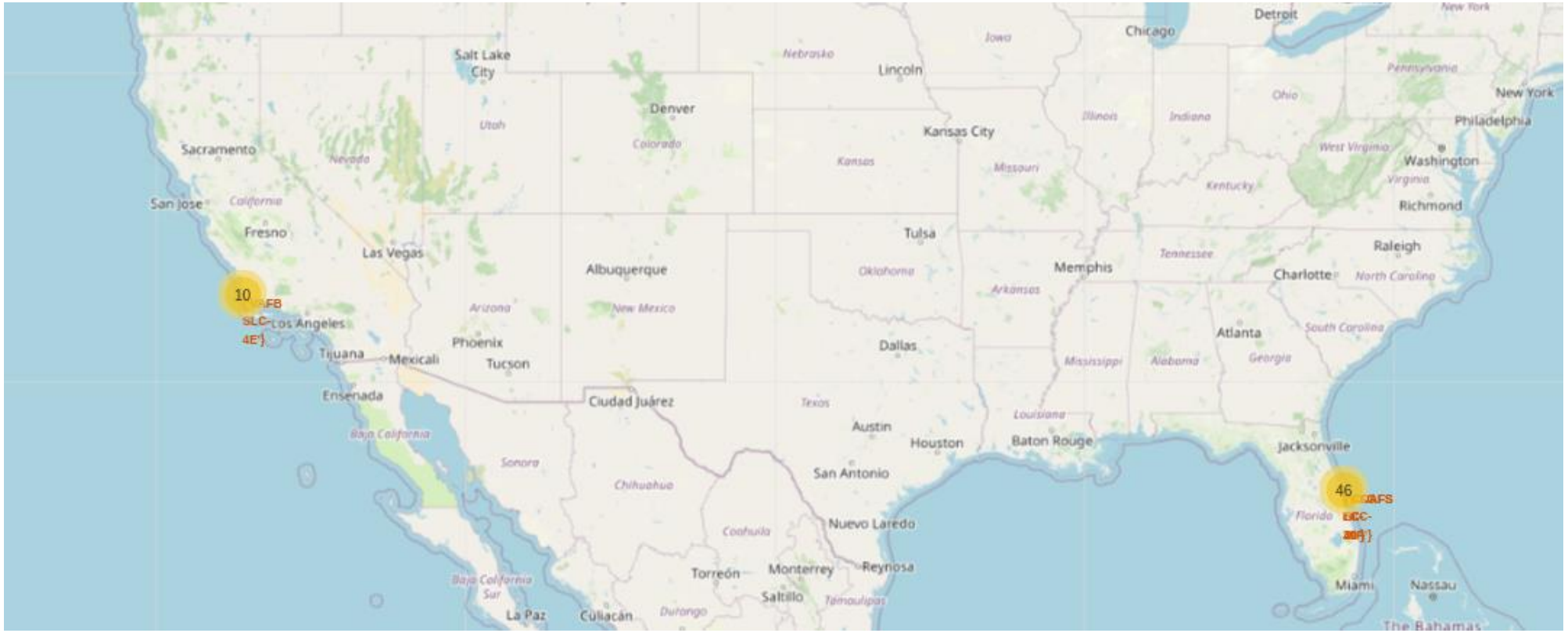
```
 * sqlite:///my_data1.db
Done.
```

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

Launch Sites

We can see that the SpaceX launch sites are in the United States of America coasts.

Florida and California

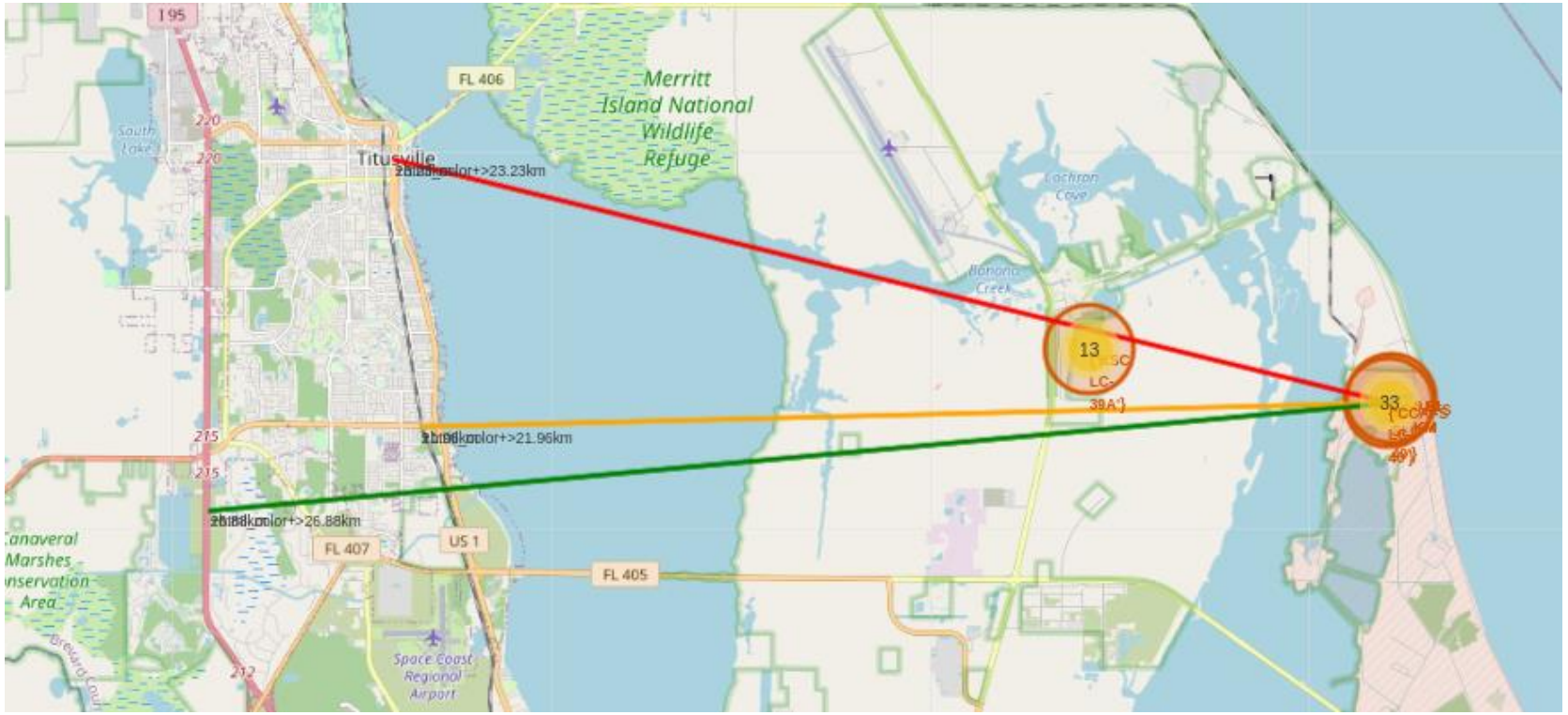


Launch Outcomes

At Each Launch Site

- Green markers for successful launches
- Red markers for unsuccessful launches

Distance to Proximities



Different Proximities

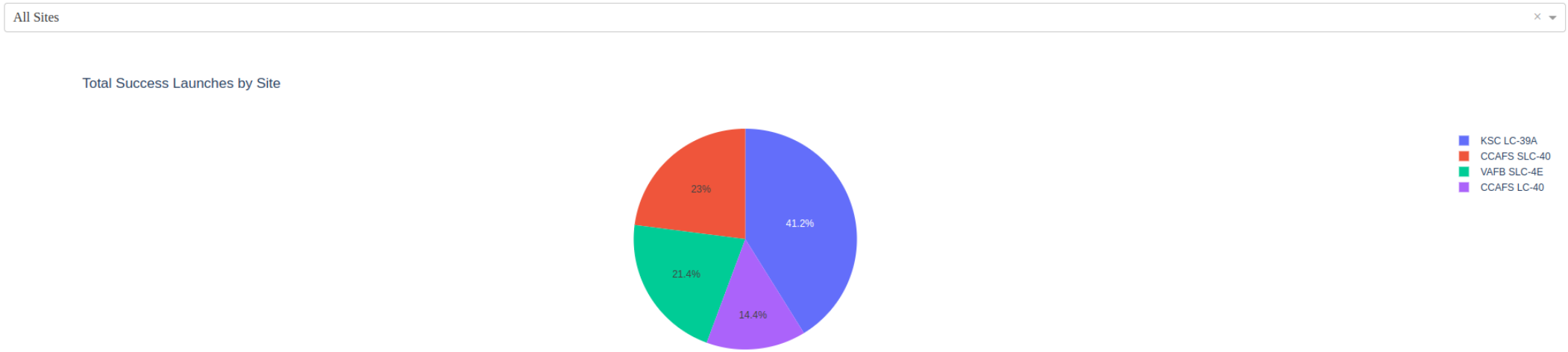
Also please try to explain your findings.

```
print("City Distance", city_distance)
print("Railway Distance", railway_distance)
print("Highway Distance", highway_distance)
print("Coastline Distance", distance_coastline)
```

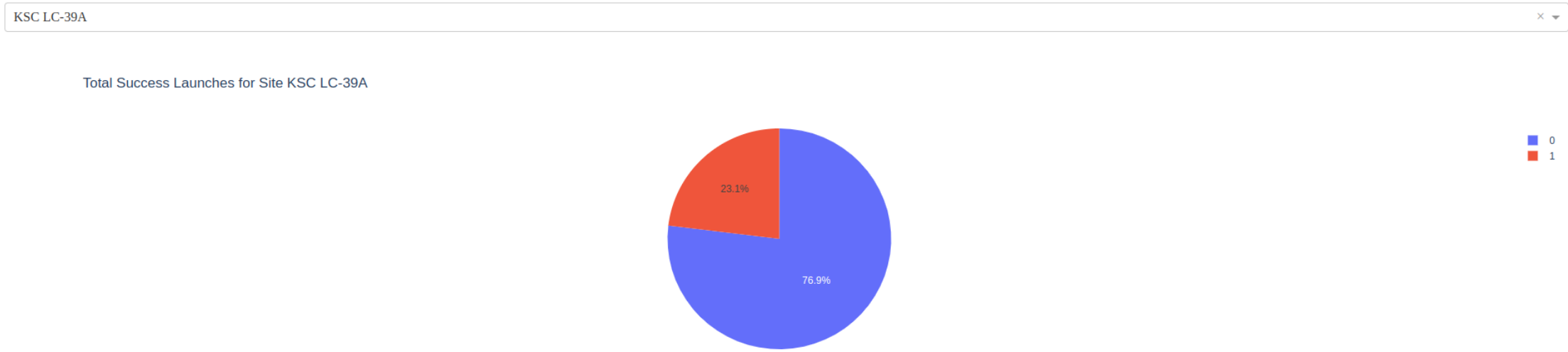
```
City Distance 23.234752126023245
Railway Distance 21.961465676043673
Highway Distance 26.88038569681492
Coastline Distance 0.5097431144955059
```

Launch Success by Site

KSC LC-39A has the most successful launches amongst launch sites (41.2%)



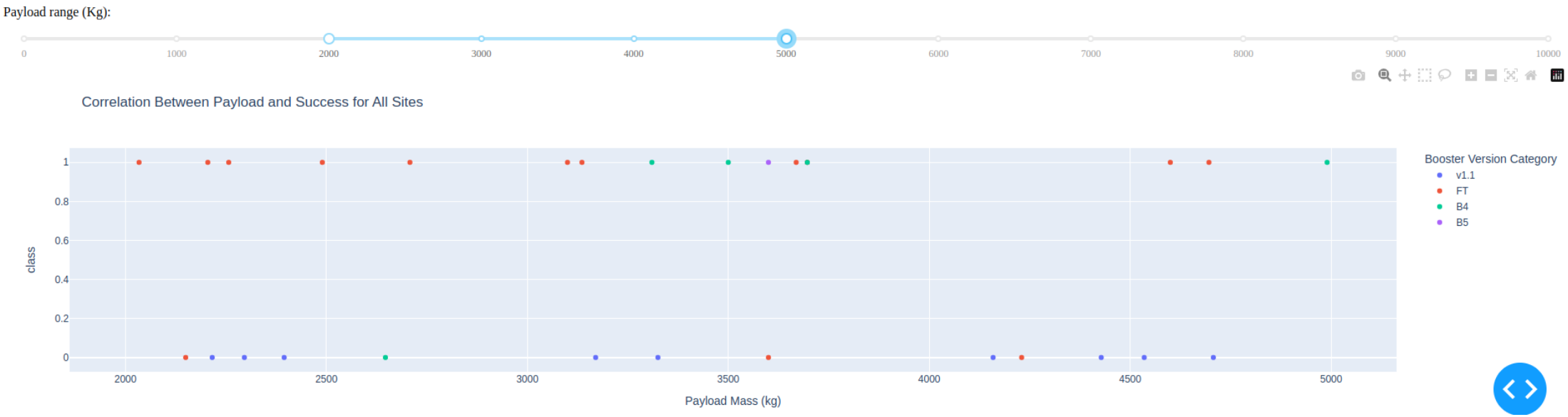
Launch Success (KSC LC-29A)



- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches

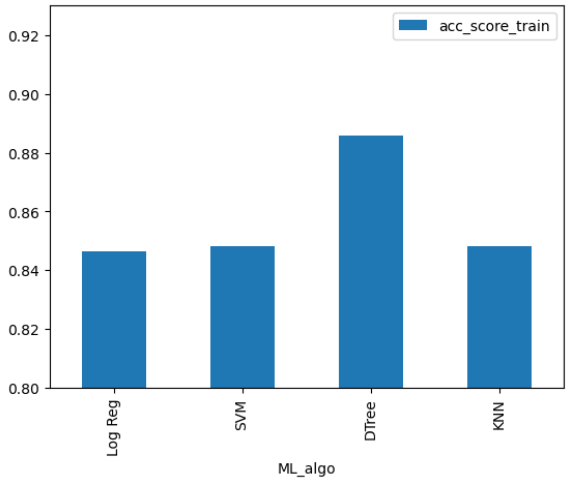
Payload Mass and Success

- By Booster Version
- • Payloads between 2,000 kg and 5,000 kg have the highest success rate
- • 1 indicating successful outcome and 0 indicating an unsuccessful outcome

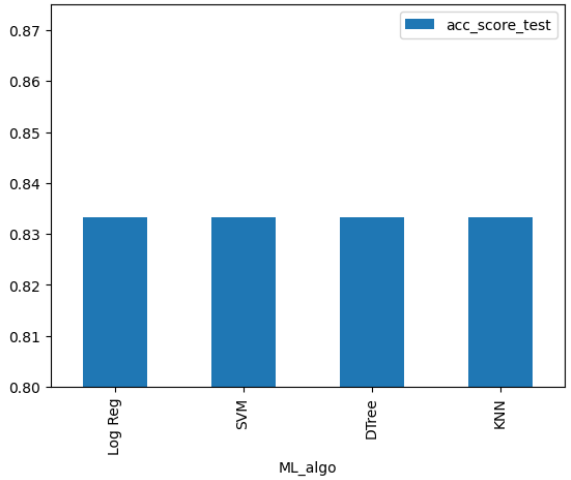


Classification Accuracy

Accuracy – Training Data



Accuracy – Testing Data

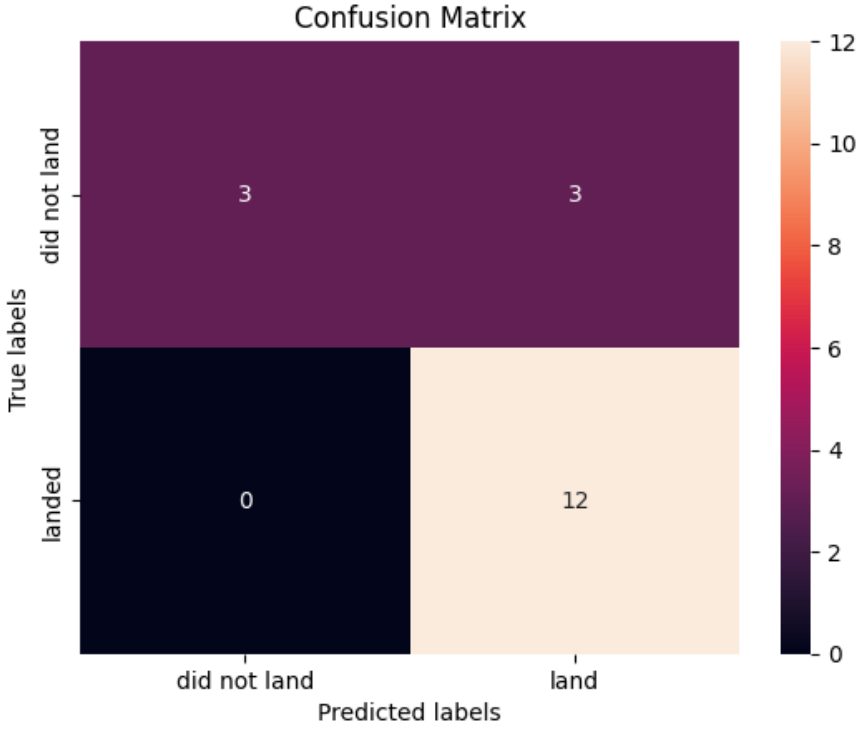


- Decision Tree has a slightly better score on the training set. However, since, Decision Trees are is non parametric and hence can lead to overfitting
- All algorithms have the same testing score
- Considering that Log Reg alogrithm with an L2 regularization is a good fit

Confusion Matrix

Best Model:

- Logistic regression
- Parameters: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}



Confusion Matrix Outputs:

- 12 True positive
- 3 True negative
- 3 False positive
- 0 False Negative

Conclusions

- Model Performance: The models performed similarly on the test set with the decision tree model slightly outperforming due to overfitting. However, the best performing model was logistic regression
- Launch Success: Increases over time
- KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Thank you!