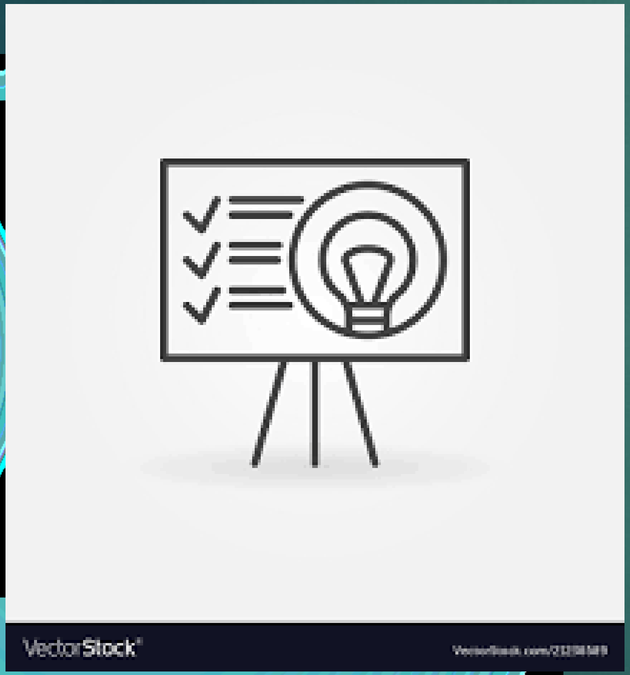
Data Science Capstone Project
Jimmy Madadi
05.03.2023

SPACEX

Outline



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

Executive Summary

Summary of methodology

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all the results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context

SpaceX is the most successful company of the commercial space age making the space travel affordable. The company advertises Falcon rocket launches on its website, with a cost of 62 million dollars whereas other providers cost up to 1625 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 launch. Based on the public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered

How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

What is the best algorithm that can be used for binary classification in this case?

Methodology

Data collection methodology:

- Using SpaceX REST API
- Using Web Scrapping from Wikipedia

Performed Data Wrangling

- Cleaning the data
- Dealing with missing values
- Using one hot encoding to prepare the data to a binary classification

Conducted exploratory data analysis (EDA) using visualization and SQL

Performing interactive visual analytics using Folium and Plotly Dash

Performing predictive analysis using classification methods

- Building, tuning and evaluation of classification models to ensure the best results

Data Collection

This process involved a combination of API requests from SpaceX REST API and Web Scrapping data from a table in SpaceX's Wikipedia entry.

Both the data collection were used in order to get complete information about the launches for a more detailed analysis.

Data columns obtained using SpaceX REST API are as follows:

FlightNumbers, Date, BoosterVersion, Orbit, PayloadMass, Customer, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data collection – SpaceX API

Requesting rocket launch data from SpaceX API

Decoding the response content using .json() and turning it into a dataframe using .json_normalize()

Requesting needed information about the launches from SpaceX API by applying custom functions

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Filtering the dataframe to only include Falcon 9 launches

Replacing missing values of Payload Mass column with calculated.mean() for this column

Exporting the data to CSV

GitHub link: Data Collection

Data collection – Web Scraping

1. Requesting Falcon 9 launch data from Wikipedia
2. Creating BeautifulSoup object from HTML response
3. Extracting all column names from the HTML table header
4. Collecting the data by parsing HTML tables
5. Constructing data we have obtained into a dictionary
6. Creating a dataframe from the dictionary
7. Exporting the data to CSV

[Github link: Data Collection Web scraping]

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful

GitHub link: Data Wrangling

EDA with data visualization

Charts were plotted: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type and Success Rate Yearly Trend Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value. Line charts show trends in data over time (time series)

GitHub link: Data Visualization

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA
- 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 when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship 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 failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

GitHub link: SQL

Build an interactive map with Folium

Markers of all Launch Sites

- Ranking 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

Colored Markers of the launch outcomes for each Launch Site

- Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities

- Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

GitHub link: Interactive Map with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List

- Dropdown list for Launch site selection

Pie Chart Showing Success Launches (All Sites/Certain Site)

- Used a scatter plot to showcase the total successful launches cunt for all sites and the Success vs. Failed counts for the site, if a specific Launch Site is selected

Slider od Payload Mass Range

- Slider is used for selecting payload range

Scatter Chart of PayloadMass vs. Success Rate for the different Booster Versions

- Used a Scatter chart to show the correlation between Payload and Launch Success

GitHub link: Dashboard with Plotly Dash

Predictive analysis (Classification)

Creating a NumPy array from the column 'Class' in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a dictionary with cv=10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree and KNN model

Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

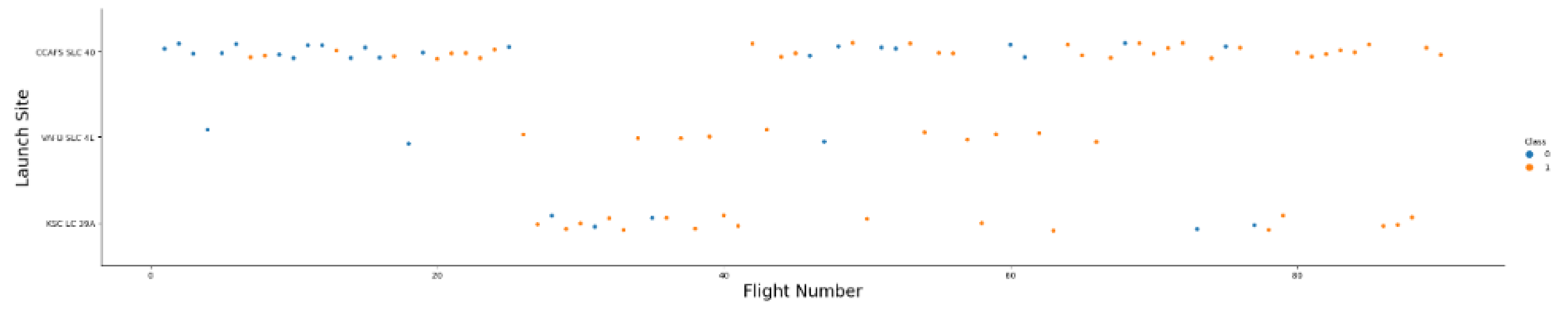
[GitHub link: Predictive analysis]

Results

1. Exploratory data analysis results
2. Interactive analytics demo in screenshots
3. Predictive analysis results

EDA with Visualization

Flight Numbers vs. Launch Site



Explanation:

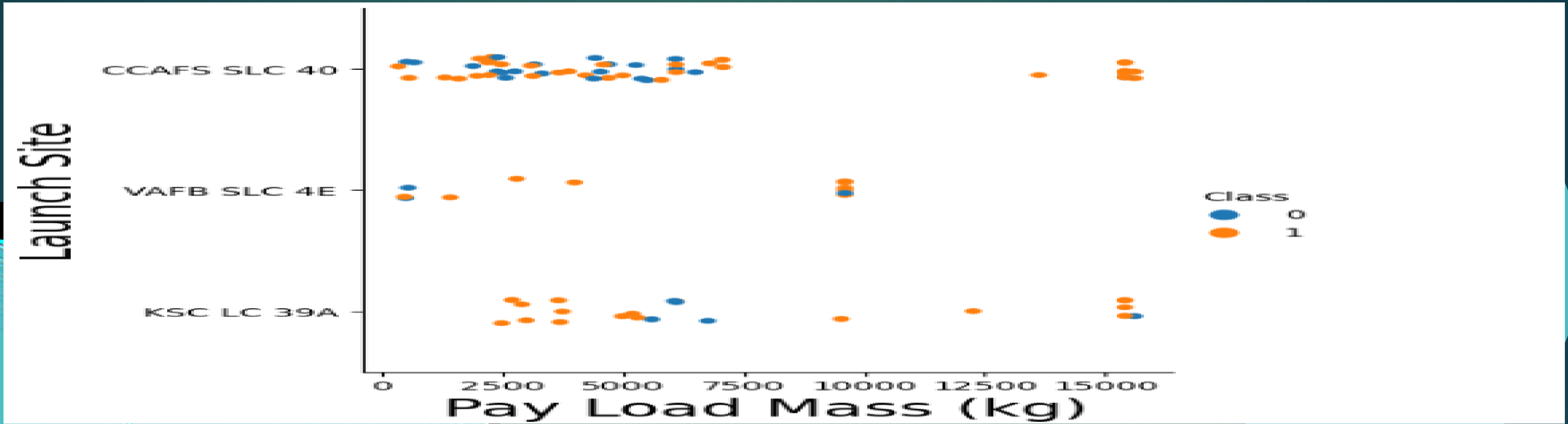
The CCAFS SLC 40 launch site has about a half of all launches

The earliest flights all failed while the latest flight all succeeded

VAFB SLC 4E and KSC LC 39A have higher success rates

Thus, it can be assumed that each new launch has a higher rate of success

Payload vs. Launch Site



Explanation

For every launch site, the higher the payload mass, the higher the success rate

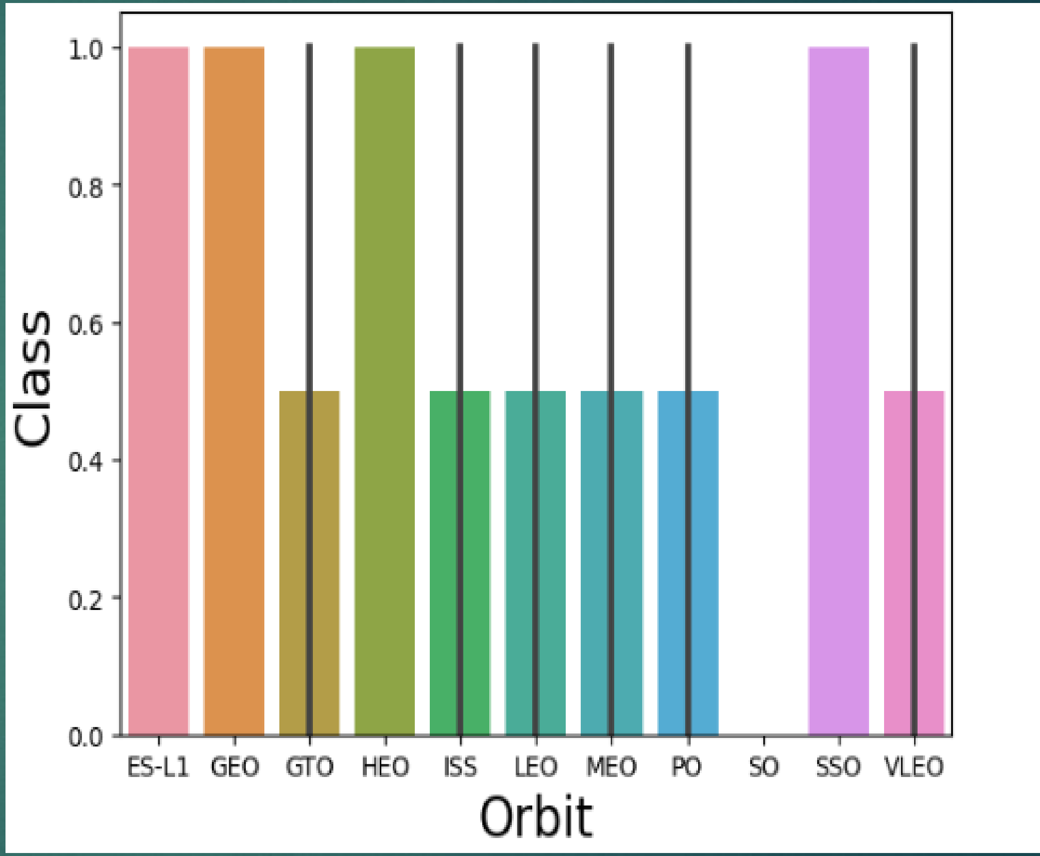
KSC LC 39A has 100% success rate for payload mass under 5500kg

Majority of the launches with payload above 7000 kg were successful

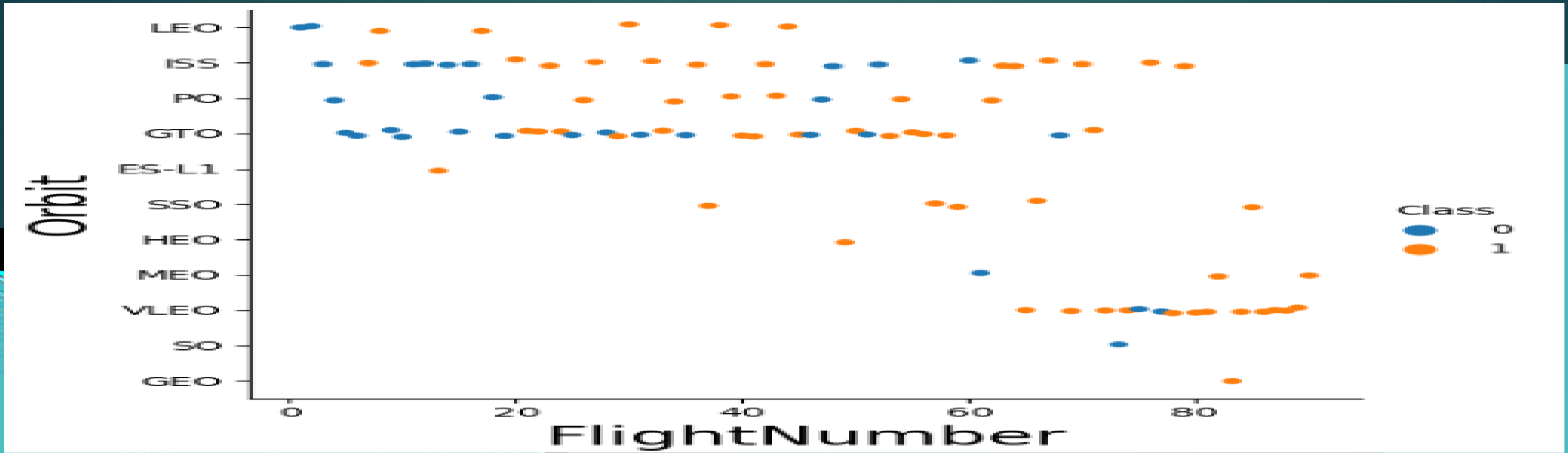
Success rate vs. Orbit type

Explanation

- Orbits with 100% success rate:
  - ✓ ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - ✓ SO
- Orbits with success rate between 50% and 85%:
  - ✓ GTO, ISS, LEO, MEO, PO



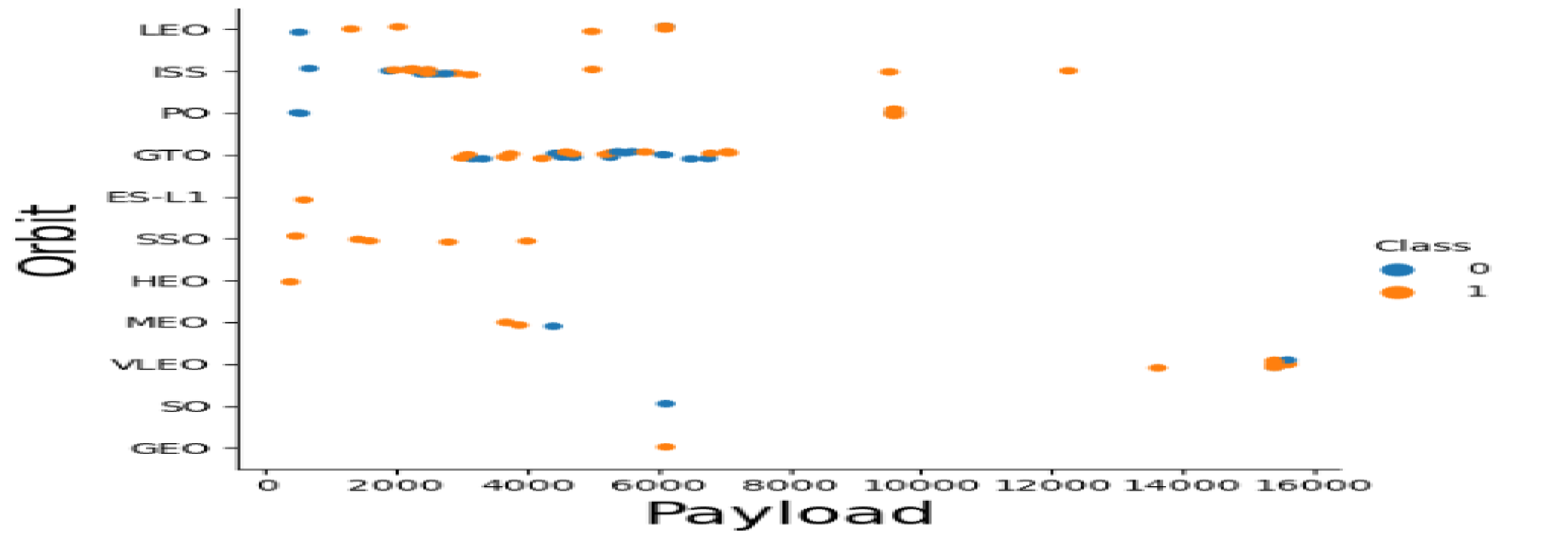
Flight Number vs. Orbit type



Explanation:

In the LEO orbit, the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload Mass vs. Orbit type



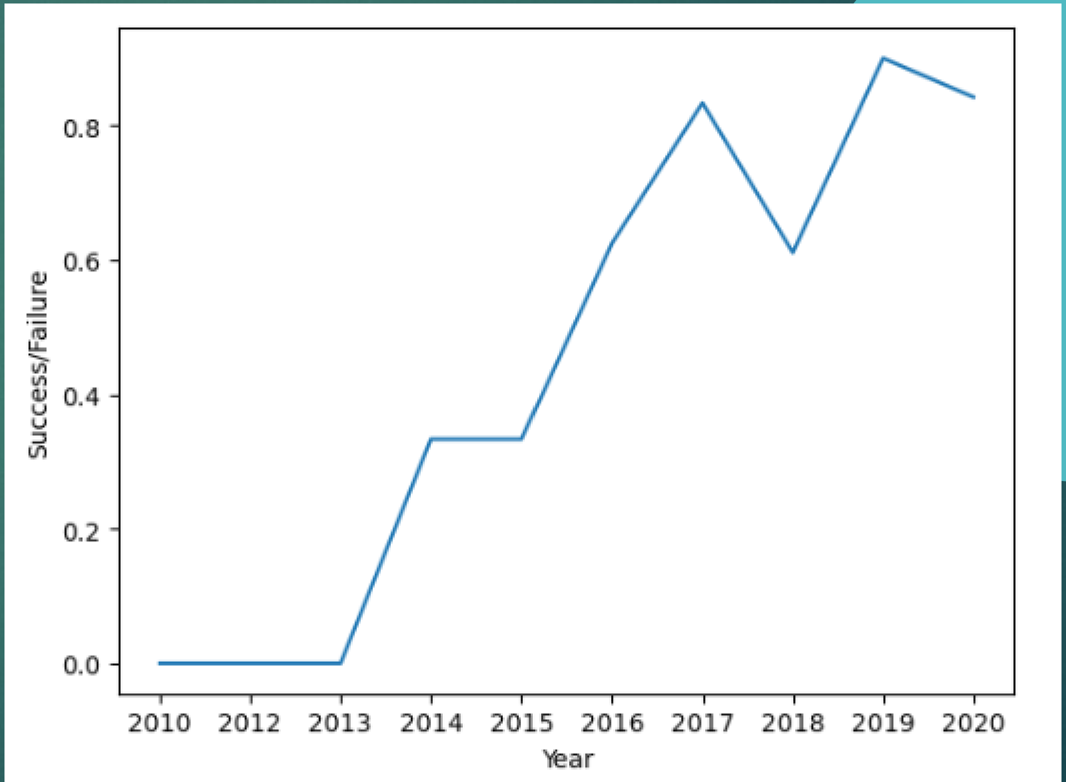
Explanation:

Heavy payloads have a negative impact on GTO orbit and positive on GTO and Polar LEO(ISS) orbits.

Launch Success Yearly trend

Explanation:

The success rate since 2013 kept increasing till 2020.



EDA with SQL

All launch site names


```
%sql select distinct launch_site from SPACEXDATASET;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Explanation:

Highlight of the names of the unique launch site in the space mission.

Launch site names begin with 'CCA'


```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Explanation:

Display of 5 records where launch site begin with the string 'CCA'

Total payload mass


```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

total_payload_mass
45596

Explanation:

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

Average payload mass by F9v1.1


```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

average_payload_mass
2534

Explanation:

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

First successful ground landing date


```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

first_successful_landing
2015-12-22

Explanation:
listing the date when the first successful landing outcome in ground pad was achieved.

Successful drone ship landing with payload between 4000 and 6000


```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Explanation:

Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

List of the number of successful and failure mission outcomes

Boosters carried maximum payload

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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

Explanation:

The list of names of the booster versions which have carried the maximum payload.

2015 launch records

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:
list of the failed landing outcomes in drone ships, their booster versions and launch site names for the months in the year 2015.

Rank success count between 2010-06-04 and 2017-03-20

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Explanation:

Ranking the count of landings outcomes (such as failure or success between 2010-06-04 and 2017-03-20 in descending order

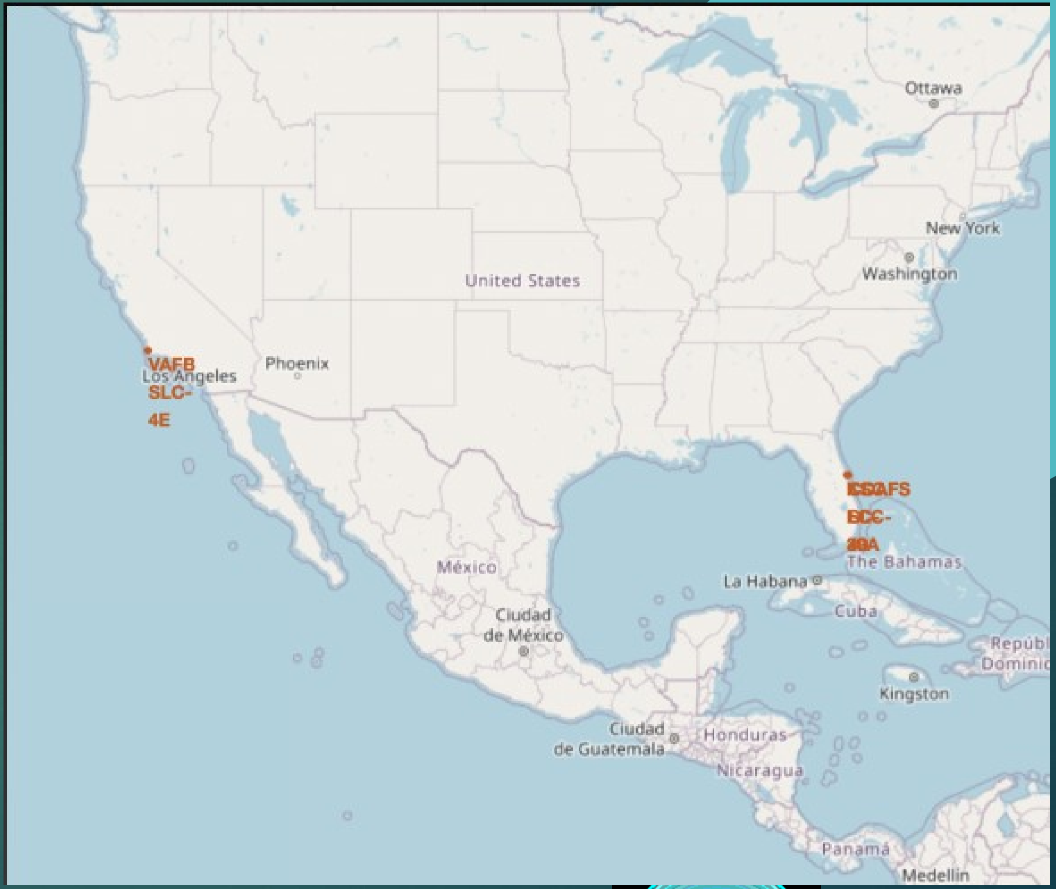
Interactive map with Folium

All launch sites location markers on global map

Explanation:

Majority of launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place of the earth. If the ship is launched from the equator it goes up into space and it is also moving around the earth at the same speed it was moving before launching. Anything on the surface of the earth at the equator is already moving at 1670 km/hr. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to spay in the orbit.

All launching site are close to the coast since launching the rocket towards the ocean helps minimize the risk of having any debris falling or exploding near people.



Color- labeled launch records on the map

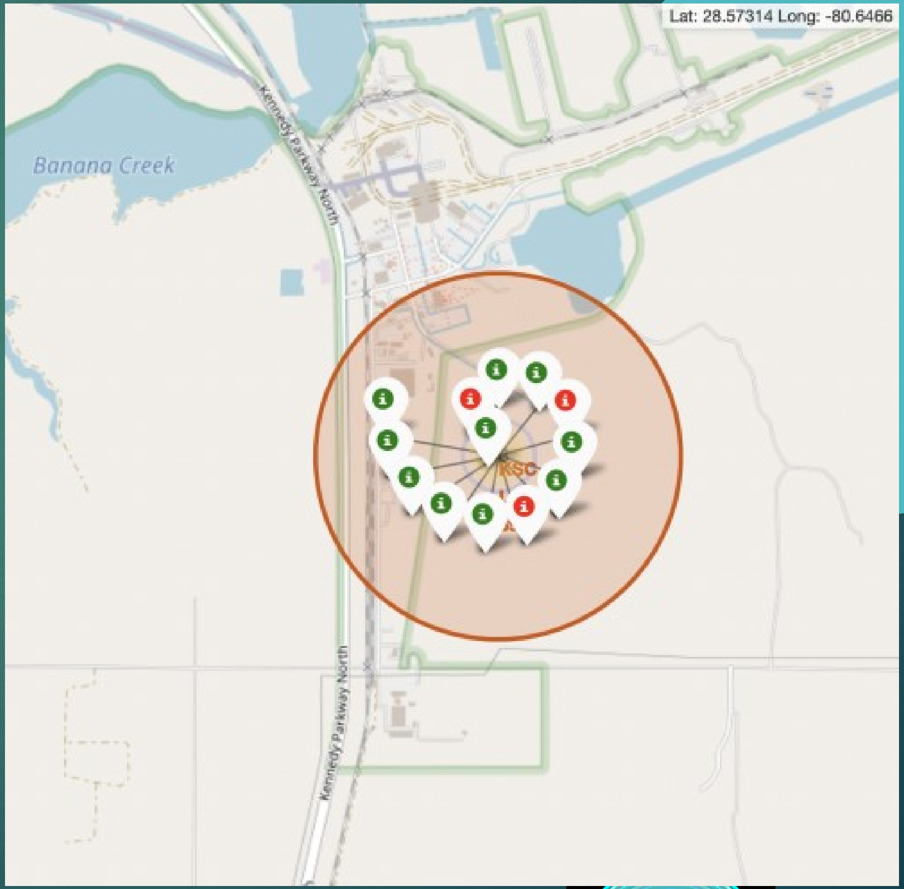
Explanation:

We should be able to easily identify which launch site have success rates.

Green maker indicates success launch

Red maker indicate failed launch

Launch site KSC LC-39A has a very high success rate



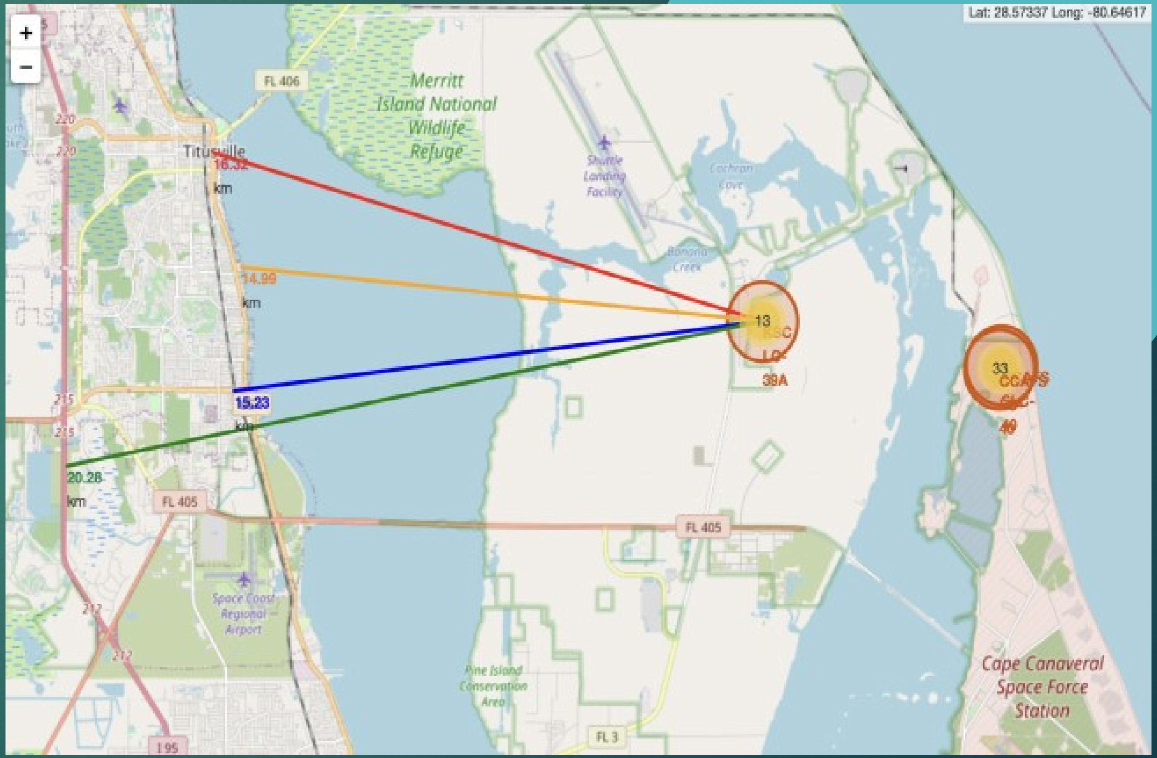
Distance from the launch site KSC LC-39A to its proximities

Explanation:

From the visual analysis of the launch site KSC LC-39A we can see that it is:

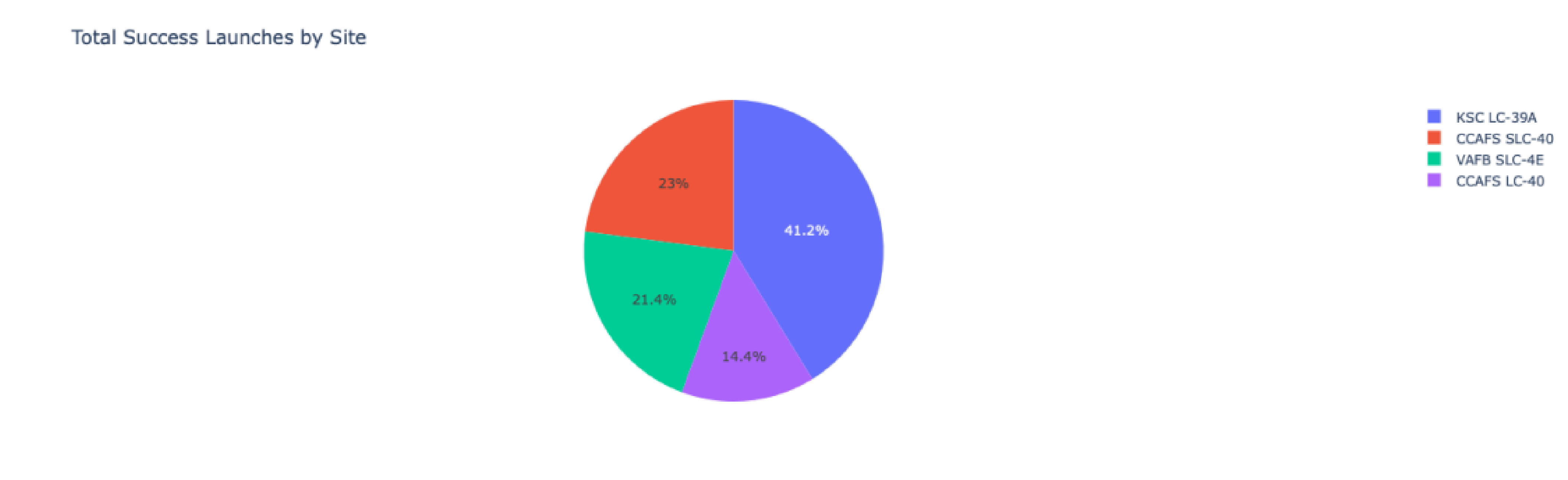
- Relatively close to railways (15.23
- Relative close to highway (20.28 km)
- Relative close to coastline (14.99 km)

Failed rocket with its highest speed can cover distances like 10 – 20 km in a few seconds thus a potential danger to populated areas.



Build Dashboard with Plotly Dash

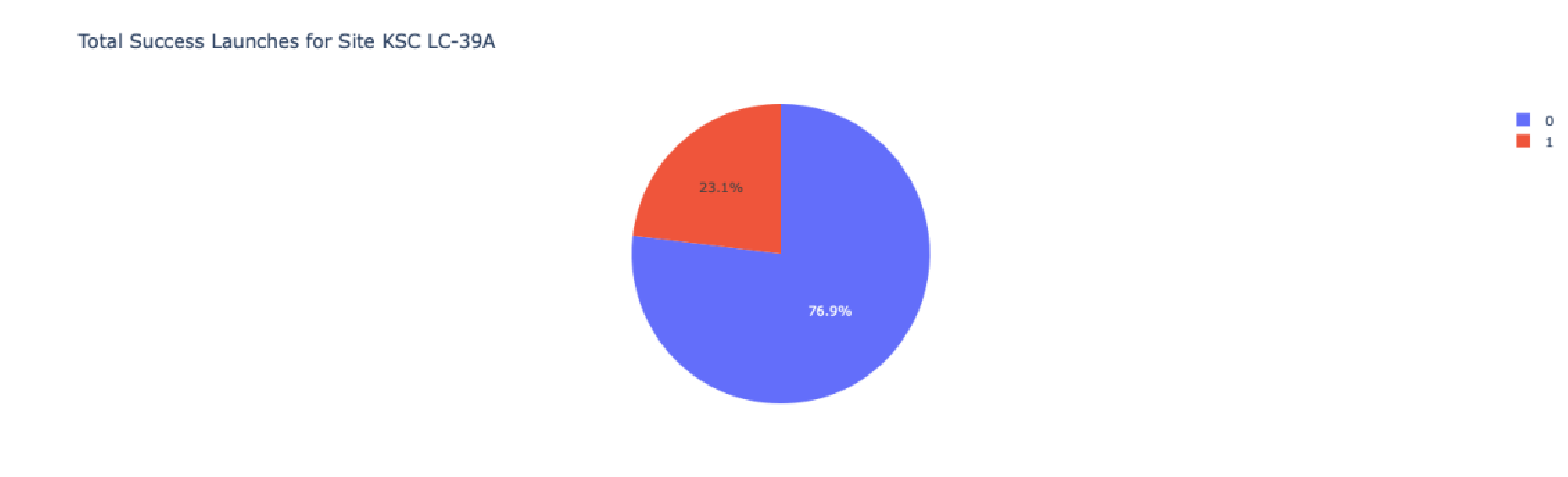
Launch success count for all sites



Explanation:

The chart showcases that across all the sites, KSC-LC 39A has the most successful launches

Launch site with the highest launch success rate



Explanation:

KSC-LC 39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landing.

Payload Mass vs. Launch Outcome for all sites

Explanation:

The charts showcase that payloads between 2000 and 5500 kg have the highest success rate

Predictive analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset affirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy

Score and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

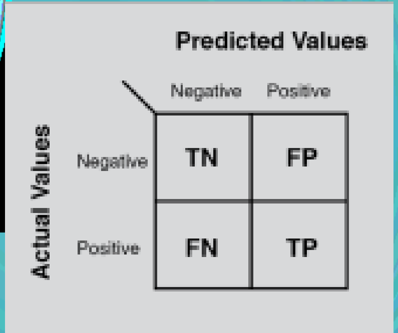
Score and Accuracy of the Entire Data Set

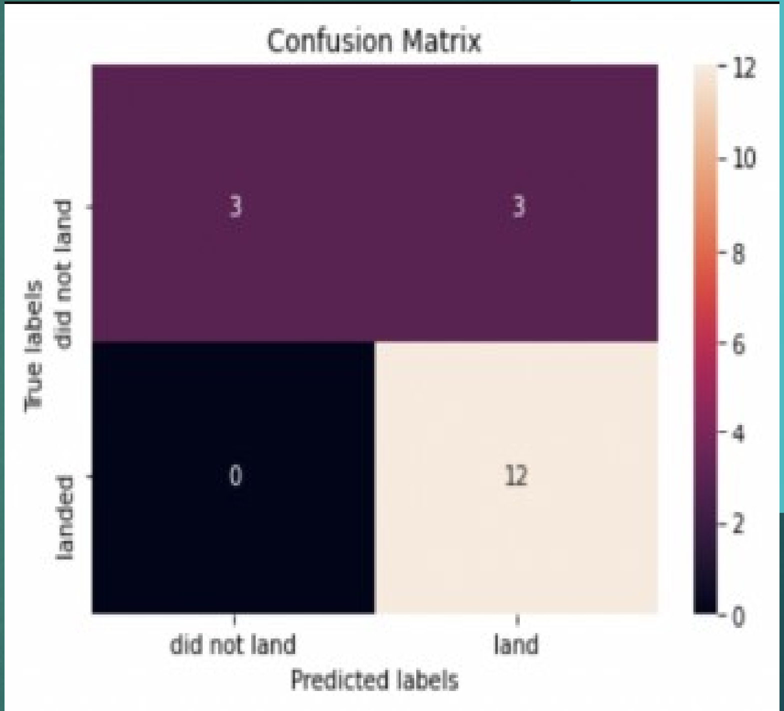
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation:

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.





Conclusion

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rat

Appendix

Special thanks to:

Instructors

Coursera

IBM