IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Faris Al Rasyid
20/3/2024

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Machine Learning Methods

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics
- Predictive analysis results

Introduction

Project background and context:

SpaceX, founded by Elon Musk in 2002, is a leading aerospace manufacturer and space transport company known for its innovative approach to space exploration. With a focus on affordability and sustainability, SpaceX has revolutionized the industry with its Falcon rocket family and pioneering work in rocket reusability. Notable achievements include successful cargo and crew missions to the International Space Station using the Dragon spacecraft. SpaceX continues to push boundaries with projects like the Starship spacecraft, aiming to enable crewed missions to Mars and beyond.

Problems you want to find answers:

- Do different variables affect the success of first stage landing?
- Does the success rate increase over the years?
- What is the best method for binary classification?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - A combination of API requests from SpaceX and Web Scraping from Wikipedia
- Perform data wrangling
  - Filtering data and the missing values
  - Using One Hot Encoding to prepare the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluating classification models

Data Collection

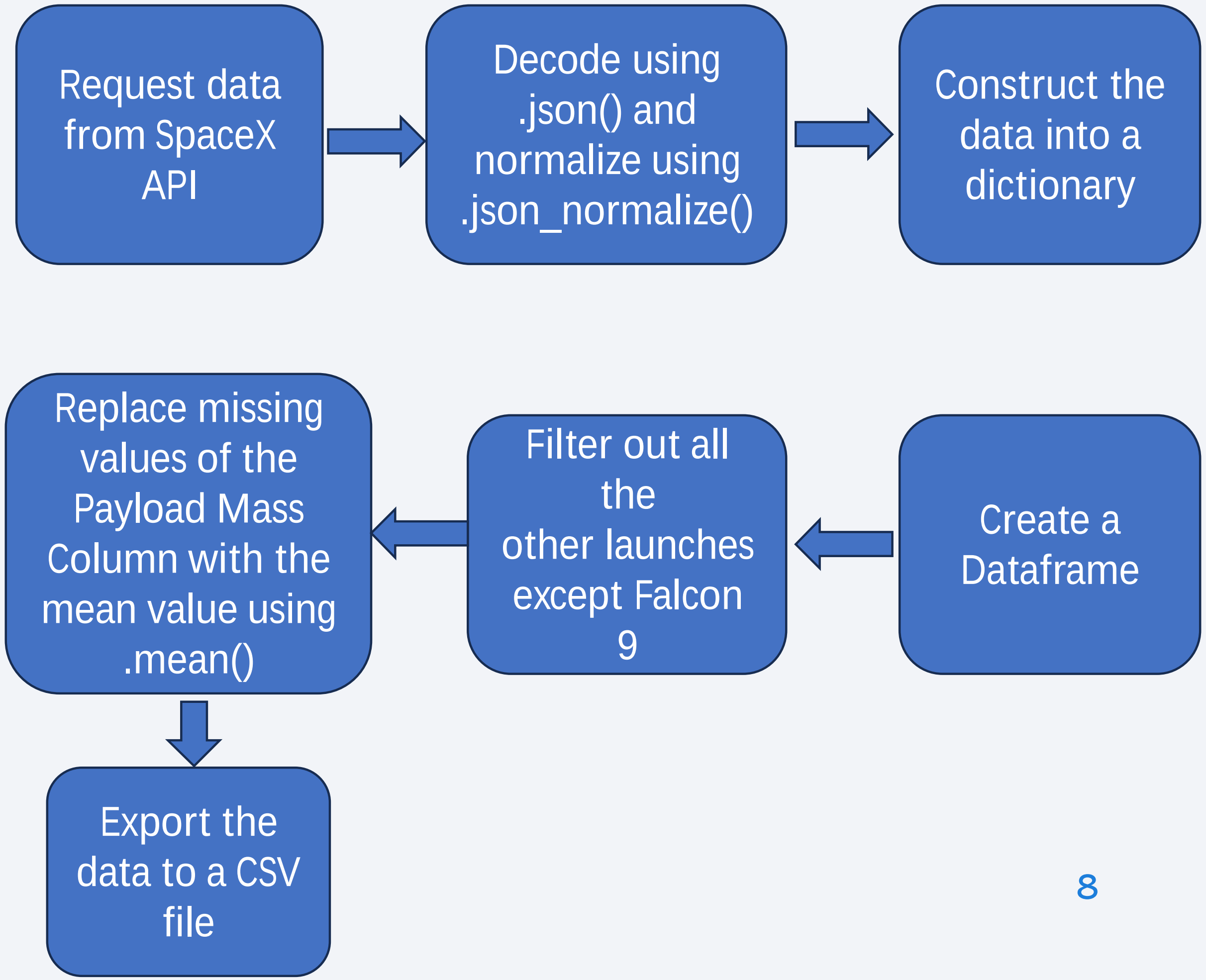
Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

- Data Columns are obtained by using SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, 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

- The Data collection was achieved using the following:

1. Requesting the Data by SpaceX API
2. Normalizing the data
3. Turn the data to a dictionary
4. Create a Dataframe
5. Filter the Data to include only Falcon 9 rockets
6. Replace the missing values using .mean() function

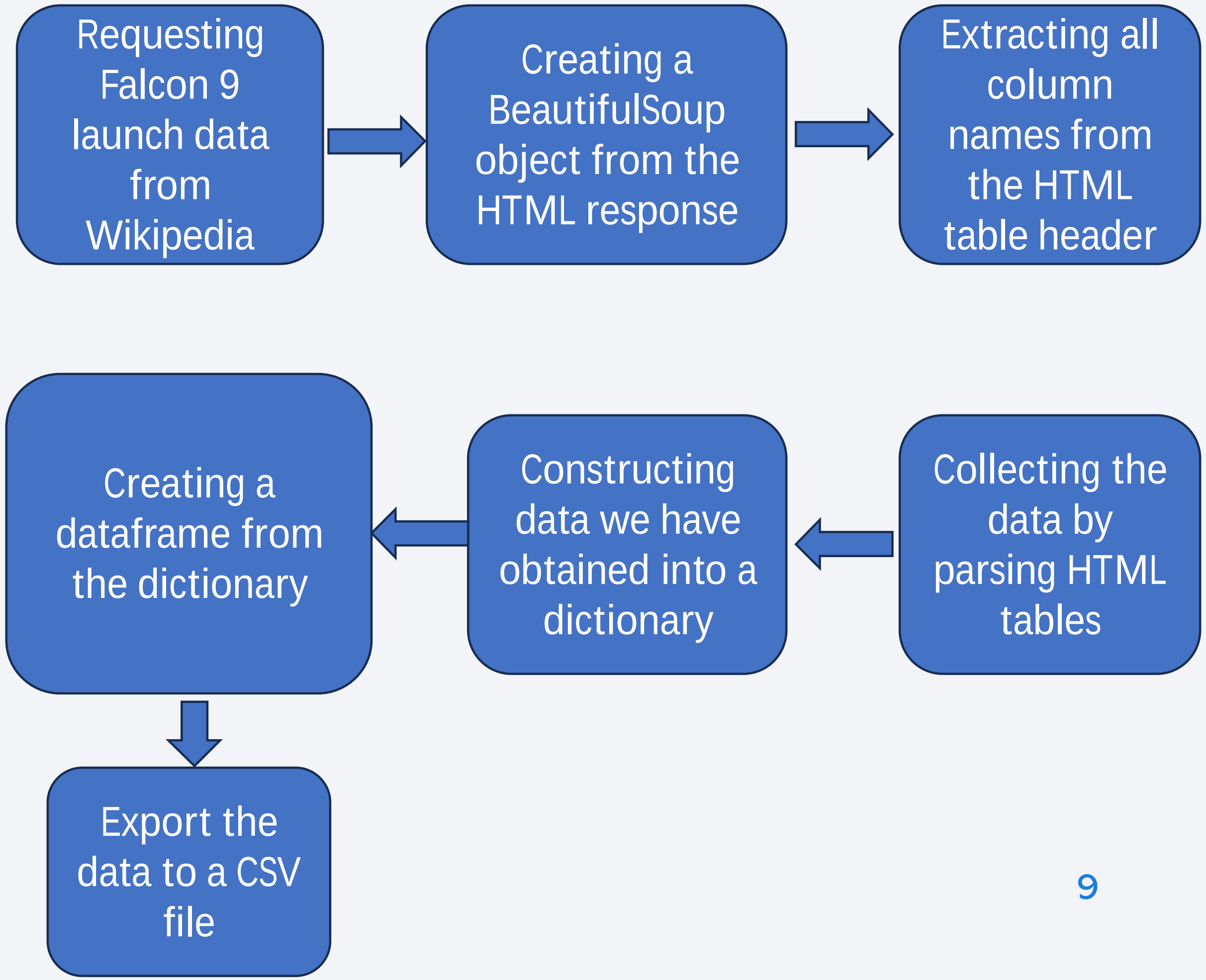


[Github Data Collection API]

Data Collection - Scraping

- The Data collection was achieved using the following:

1. Requesting the Data from Wikipedia
2. Create a BeautifulSoup object
3. Extract column names
4. Collect the data by parsing HTML tables
5. Turn into a dictionary
6. Create a Dataframe



[Github url Data Collection with Web Scraping]

Data Wrangling

From the dataset, we had different cases for the booster's landing. Below are the cases:

- True Ocean means the mission outcome was successfully landed to a specific region of the ocean, False Ocean meant unsuccessful landing.
- 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.

During the processing of the data, the outcomes where converted into 1 for successful landing and 0 for unsuccessful.

- Github URL 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 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 EDA with SQL]

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Launch Analysis with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List: -

- Added a dropdown list to enable Launch Site selection. Pie Chart showing Success Launches (All Sites/Certain Site): -
- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected. Slider of Payload Mass Range: -
- Added a slider to select Payload range. Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions: -
- Added a scatter chart to show the correlation between Payload and Launch Success.

Build a Dashboard Application 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 GridSearchCV object with cv = 10 to find the best parameters
- Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
- 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

[Machine Learning Prediction]

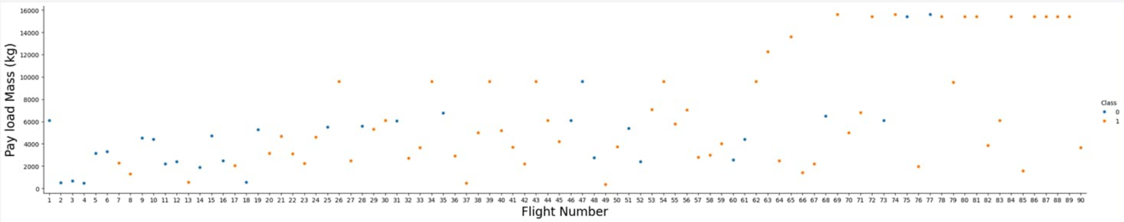
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

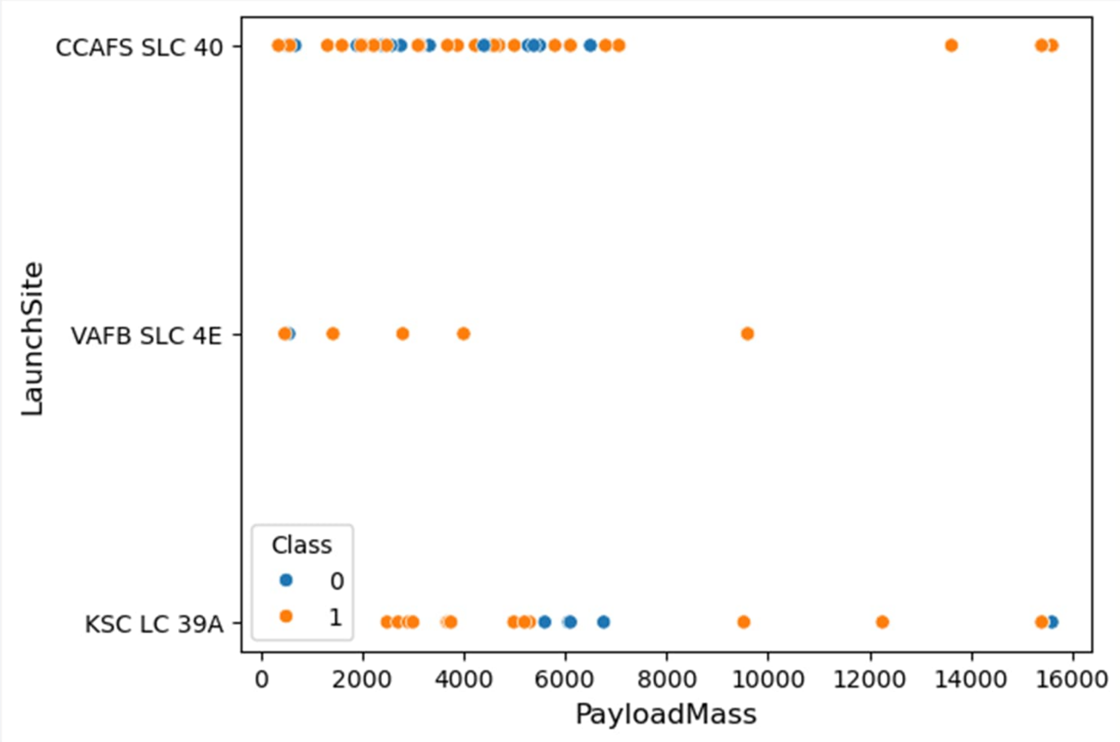


Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches. VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

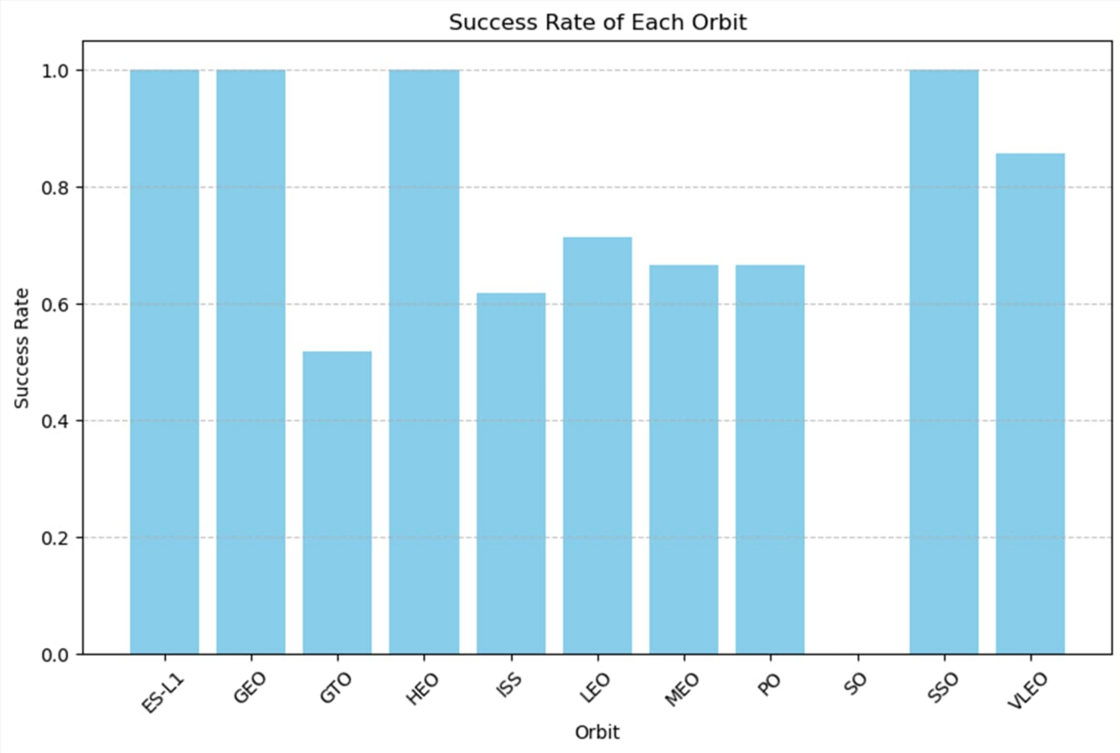
Payload vs. Launch Site

- Explanation:
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.



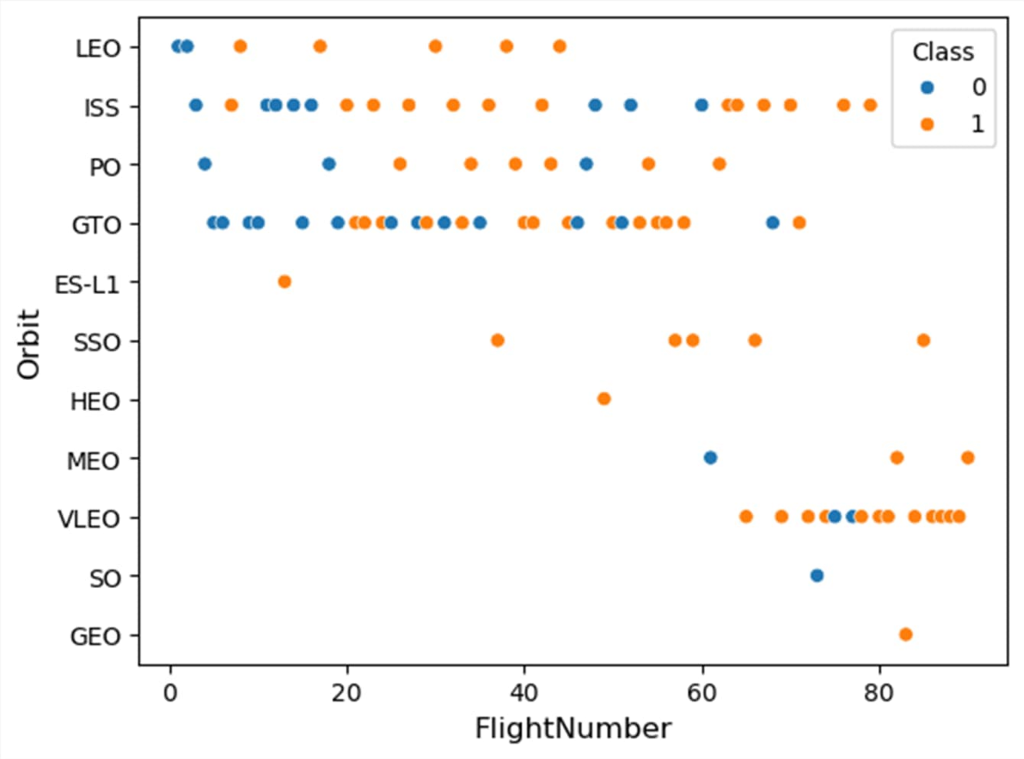
Success Rate vs. Orbit Type

- 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

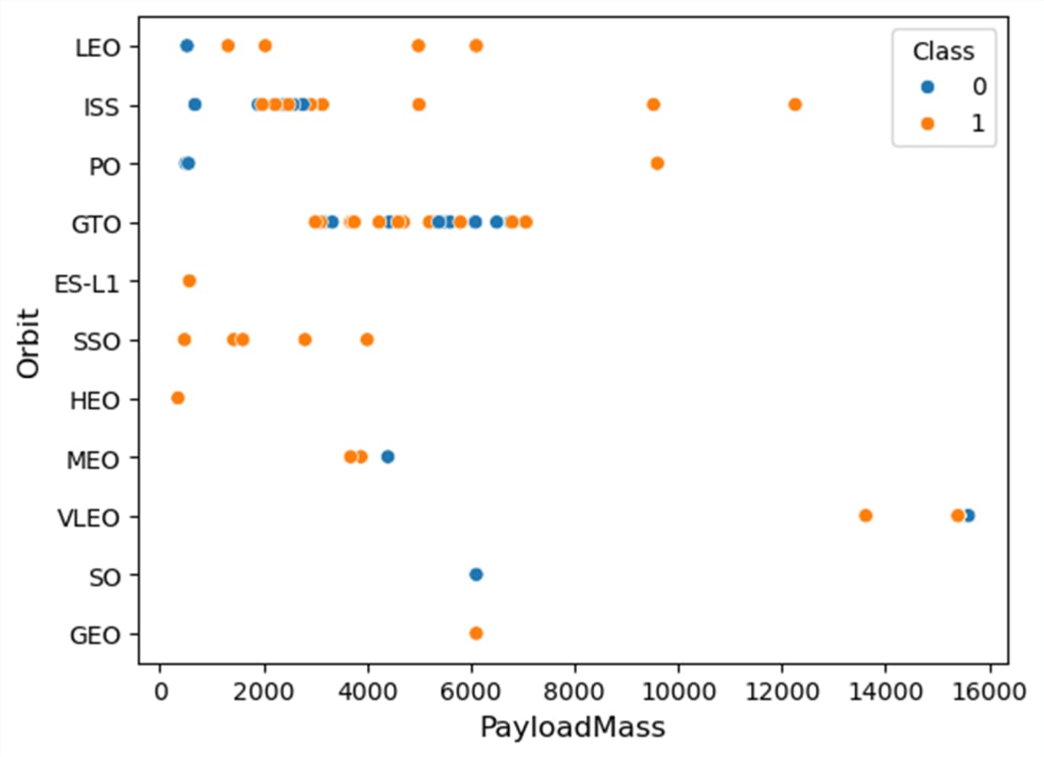
- Explanation:
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit



Payload vs. Orbit Type

Explanation:

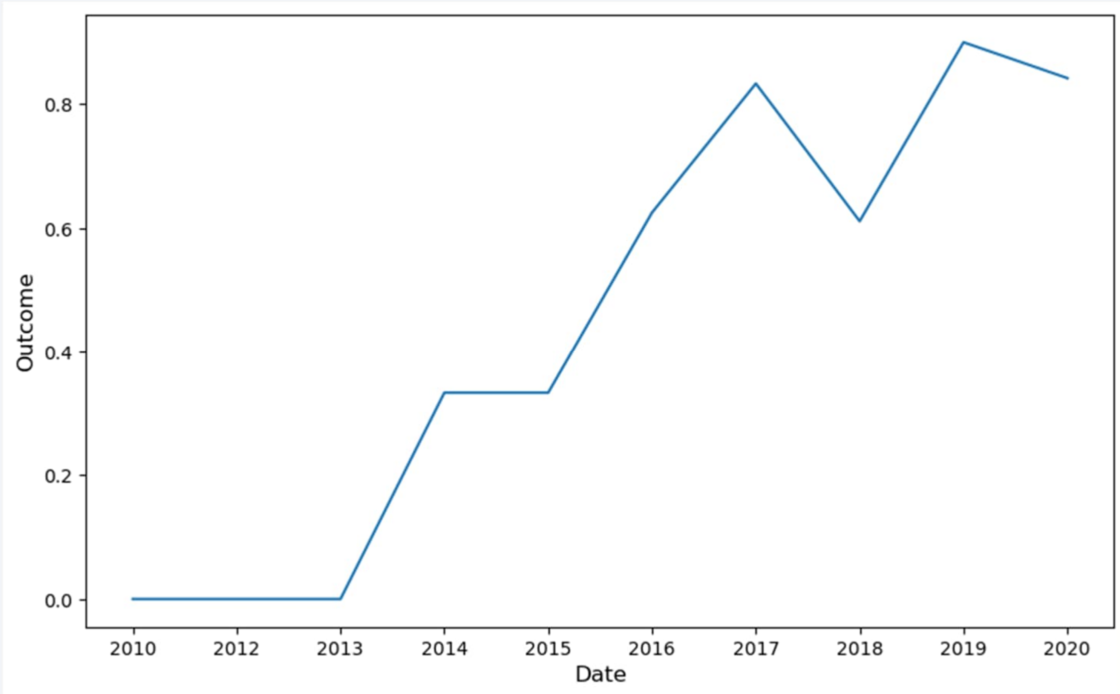
- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

Explanation:

- The success rate since 2013 kept increasing till 2020.



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
```

Python

* sqlite:///my_data1.db
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

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

```
%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTABLE \
        where "Customer" like '%NASA%' and "Mission_Outcome" like '%Success%'
```

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

sum("PAYLOAD_MASS__KG_")
105058

Average Payload Mass by F9 v1.1

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

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 "Mission_Outcome" like 'Success%'
```

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

min("Date")
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select "Booster_Version" from SPACEXTABLE \
        where "Landing_Outcome" like '%Success (drone ship)%' \
        AND "PAYLOAD_MASS__KG_" > 4000 \
        AND "PAYLOAD_MASS__KG_" < 6000
```

Pyth

```
 * 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

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS "Total" \
    FROM SPACEXTABLE \
    WHERE "Mission_Outcome" IN ('Success', 'Failure') \
    GROUP BY "Mission_Outcome"
```

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

Mission_Outcome	Total
Success	98

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT "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

```
%sql SELECT CASE substr(Date, 6, 2) \
        WHEN '01' THEN 'January'\
        WHEN '02' THEN 'February'\
        WHEN '03' THEN 'March'\
        WHEN '04' THEN 'April'\
        WHEN '05' THEN 'May'\
        WHEN '06' THEN 'June'\
        WHEN '07' THEN 'July'\
        WHEN '08' THEN 'August'\
        WHEN '09' THEN 'September'\
        WHEN '10' THEN 'October'\
        WHEN '11' THEN 'November'\
        WHEN '12' THEN 'December'\
    END AS Month,\
    "Landing_Outcome",\
    "Booster_Version",\
    "Launch_Site"\
FROM SPACEXTABLE\
WHERE substr(Date, 0, 5) = '2015'\
    AND "Landing_Outcome" LIKE '%Failure (drone ship)%'
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS "Count" \
FROM SPACEXTABLE \
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY "Landing_Outcome" \
ORDER BY "Count" DESC
```

Pyth

* sqlite:///my_data1.db
Done.

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

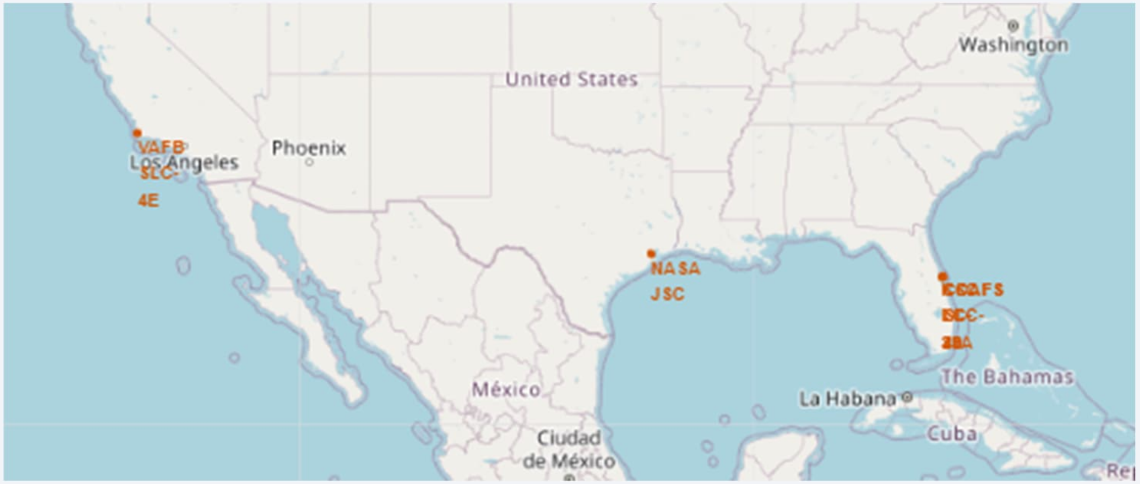
Section 3

Launch Sites Proximities Analysis

Launch sites in the global Map

Explanation:

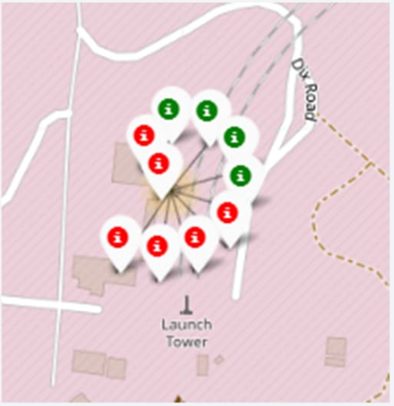
- All the launch sites are located close to the equator, since the earth moves father there.
- All the launch sites are close to the cost and away from big cities, to minimize the danger of a rocket exploding and damaging people and properties.



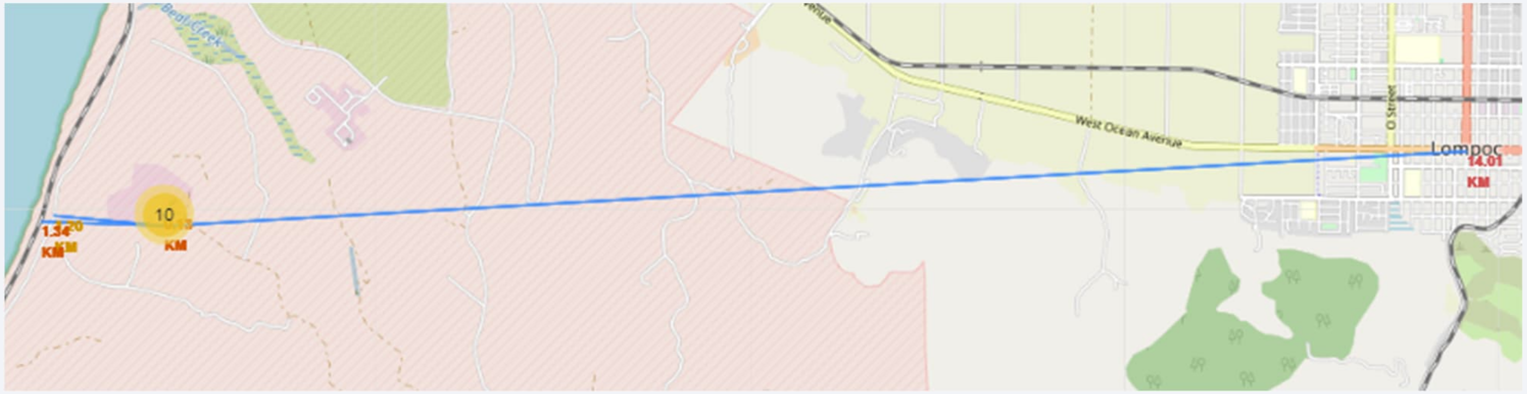
Colour-labeled launch records on the map

Explanation:
From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- ❑Green Marker = Successful Launch
- ❑Red Marker = Failed Launch

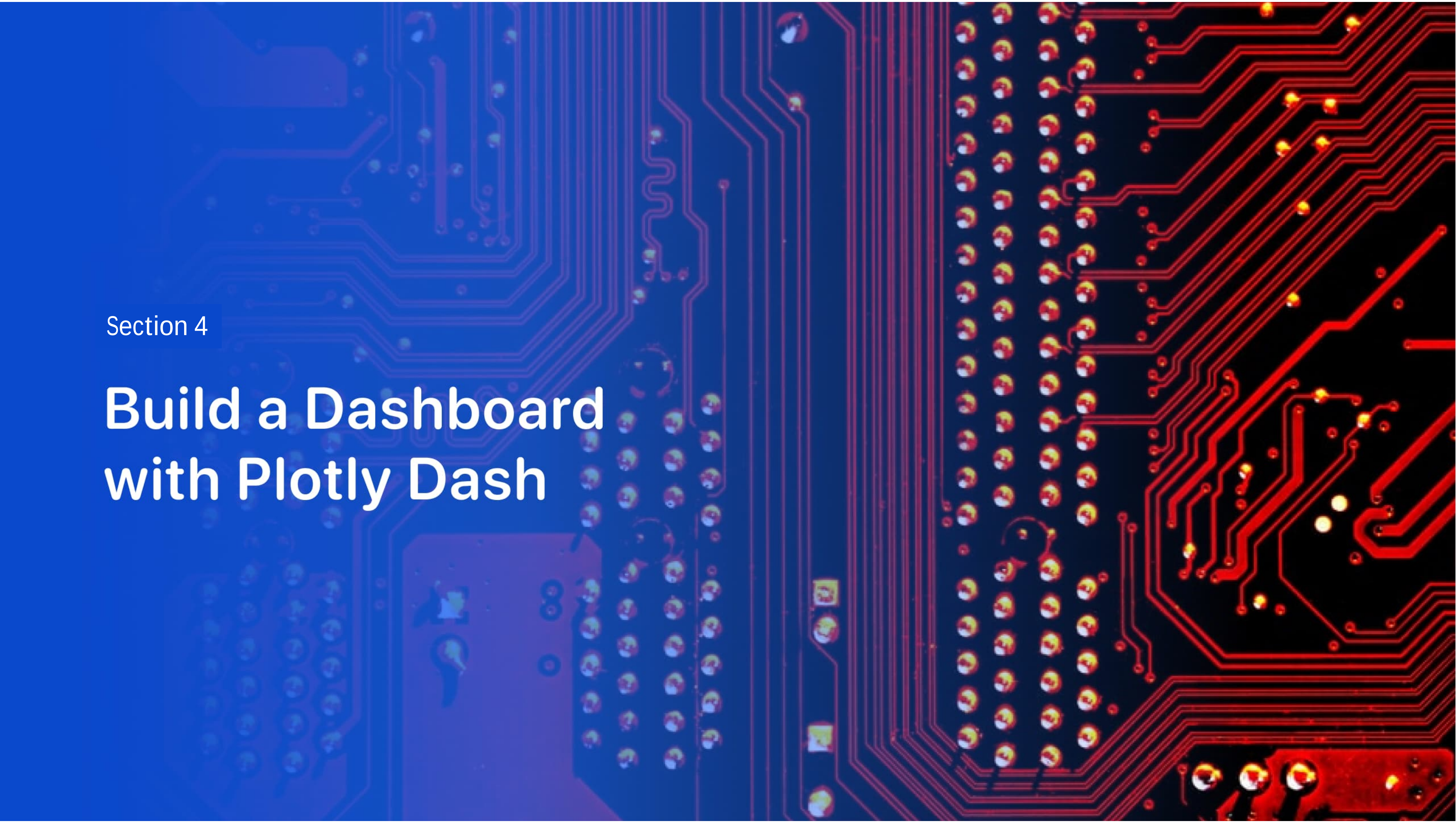


Distance from the launch site



Explanation:

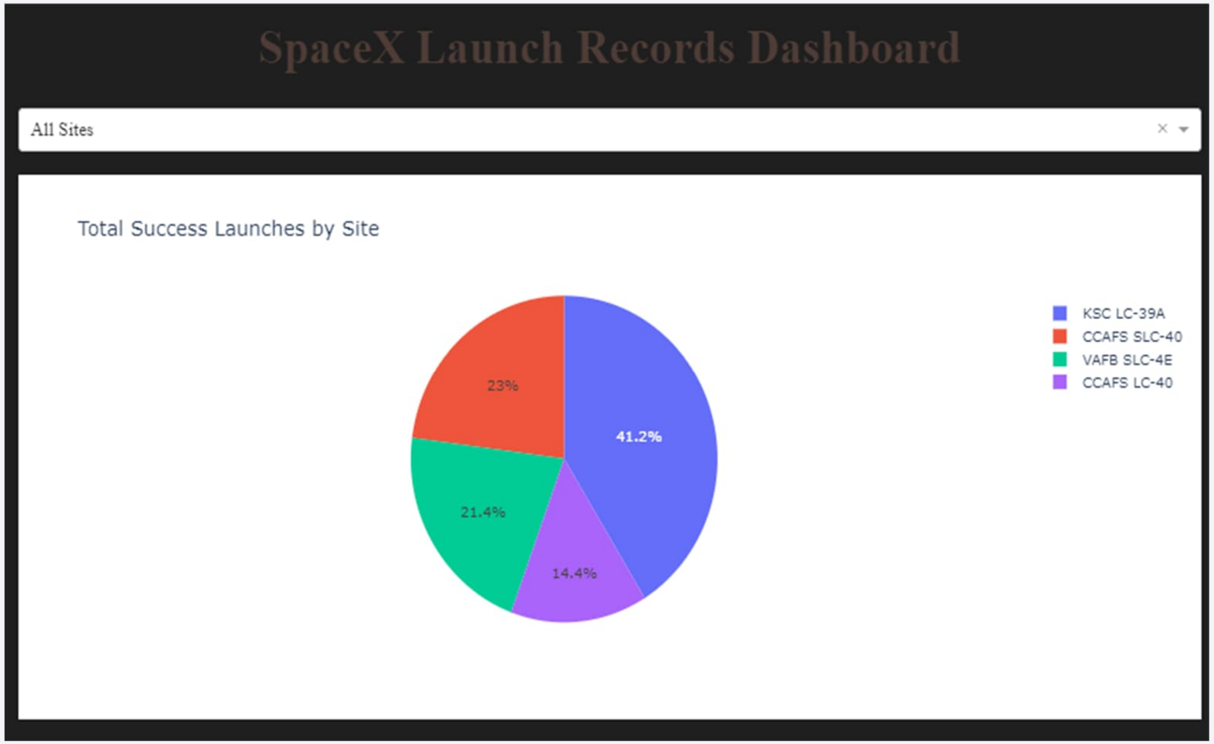
- The distance from the closest city is 14.01 km
- From the closest highway is 1.29 km
- From the railway is 1.34 km


Section 4
Build a Dashboard with Plotly Dash

Pie Chart of all the Sites

Explanation:

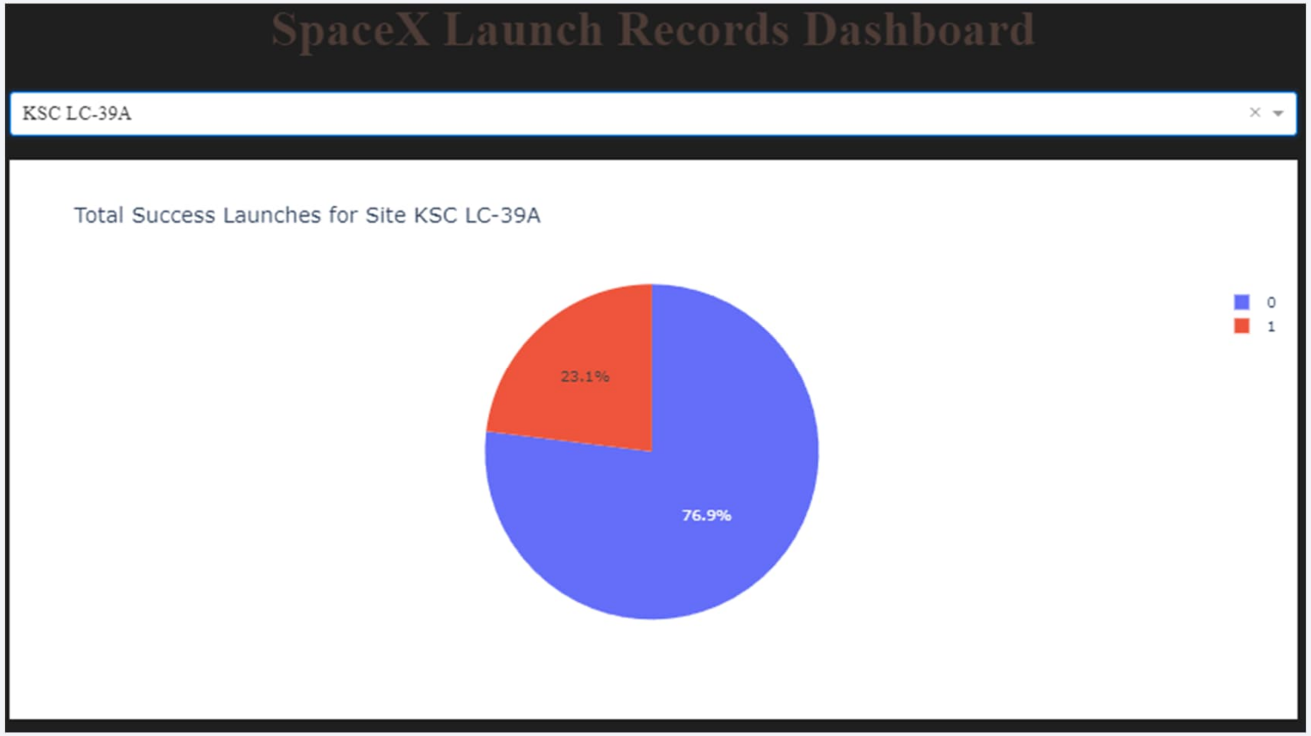
- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.



Launch site with highest launch success ratio

Explanation:

- KSC LC-39A has the highest launch success rate (76.9%)



Payload Mass vs. Launch Outcome for all sites

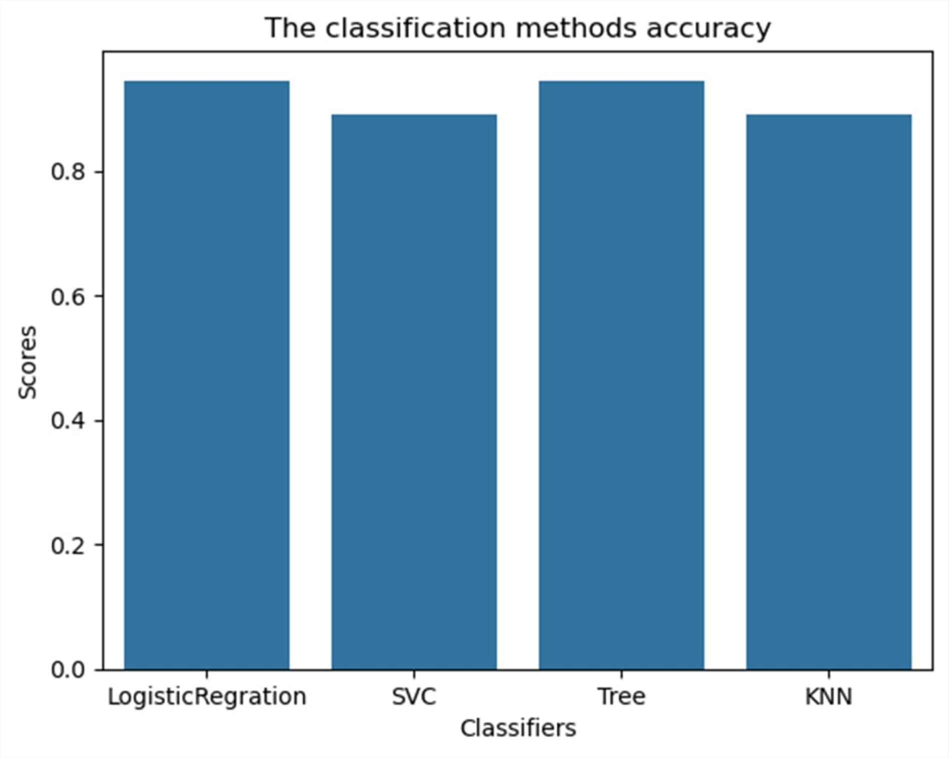
The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

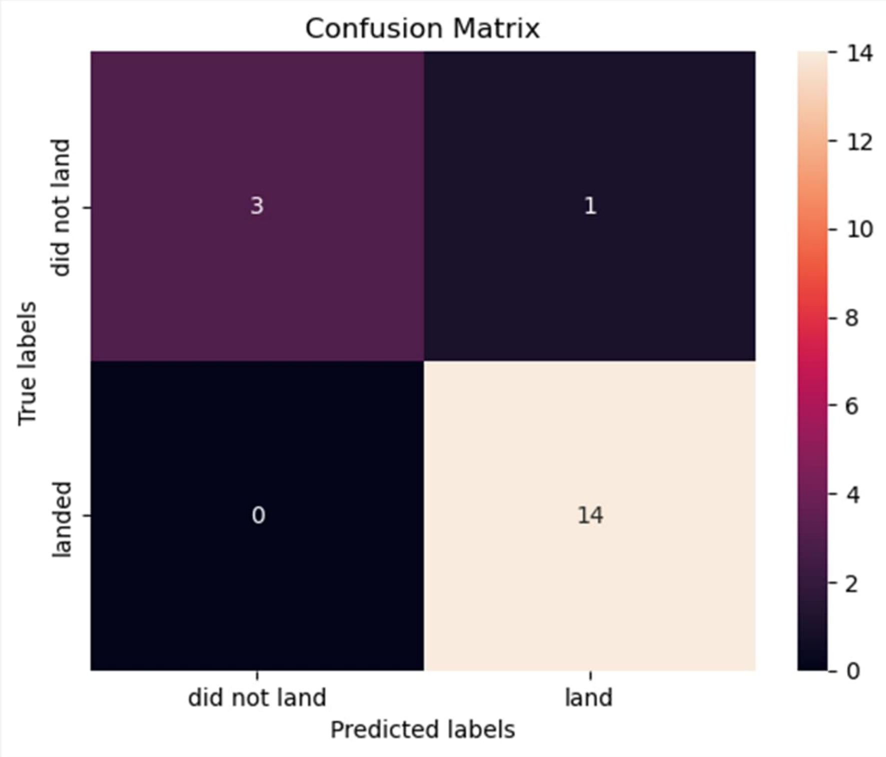
- As we can see the logistic regression method gives the best results.



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.



Conclusions

- 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.


Thank you!