IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohamed Bayoudh
29/11/2024

Outline

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

Executive Summary

Summary of Methodologies
Collecting Data: Gathering relevant datasets for analysis.
Data Wrangling: Cleaning and transforming data for usability.
Exploratory Data Analysis (EDA):
- Utilizing data visualizations for insights.
- Leveraging SQL for in-depth analysis.

Interactive Visualizations:
- Creating dynamic maps using Folium.
- Developing a dashboard with Plotly Dash.

Predictive Modeling: Implementing classification techniques for predictions.
Summary of Results
Key insights derived from EDA.
Screenshots showcasing interactive analytics tools.
Results and evaluation of predictive models.

Introduction

Project Background and Context

SpaceX has revolutionized the commercial space industry by making space travel more affordable. The Falcon 9 rocket, advertised on the company's website, costs $62 million per launch—significantly less than the $165 million charged by other providers. This cost efficiency is largely due to SpaceX's ability to reuse the rocket's first stage.

Predicting whether the first stage will successfully land is key to estimating the cost of a launch. Using publicly available data and machine learning models, this project aims to forecast the likelihood of SpaceX reusing the first stage.

Key Questions to Address

How do variables like **payload mass**, **launch site**, **number of flights**, and **orbit type** impact the success of first-stage landings?

Has the **rate of successful landings** improved over time?

Which **binary classification algorithm** is best suited for predicting first-stage landing success?

Section 1

Methodology

Data collection methodology: - Using SpaceX Rest API - Using Web Scrapping from Wikipedia

Performed data wrangling - Filtering the data - Dealing with missing values - Using One Hot Encoding to prepare the data to a binary classification

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models - Building, tuning and evaluation of classification models to ensure the best

results

Data Collection

Data Collection ProcessThe data for this project was gathered using two complementary methods:SpaceX REST API: This provided detailed launch information.Web Scraping: Data was extracted from a table on SpaceX's Wikipedia entry to fill in gaps and ensure comprehensive coverage.By combining these approaches, we obtained a complete dataset for in-depth analysis of SpaceX launches.Data CollectedFrom SpaceX REST API:FlightNumberDateBoosterVersionPayloadMassOrbitLaunchSiteOutcomeFlightsGridFinsReusedLegsLandingPadBlockReusedCountSerialLongitudeLatitudeFrom Wikipedia Web Scraping:Flight No.Launch sitePayloadPayloadMassOrbitCustomerLaunch outcomeVersion BoosterBooster landingDateTimeThis dual-source collection ensured a robust dataset for a thorough exploration of SpaceX launch metrics.

Data Collection – SpaceX API

Requesting rocket launch data from SpaceX AP

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

https://github.com/mohamedbayoudh1/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

- Requesting Falcon 9 launch data from Wikipedia
- Creating a BeautifulSoup object from the HTML response
- Extracting all column names from the HTML table header
- Collecting the data by parsing HTML tables
- Constructing data we have obtained into a dictionary GitHub URL: Data Collection with Web Scrapin
- Creating a dataframe from the dictionary
- Exporting the data to CSV

https://github.com/mohamedbayoudh1/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

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

Perform exploratory Data Analysis and determine Training Labels

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Exporting the data to CSV

https://github.com/mohamedbayoudh1/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

Data Visualization Overview
Various charts were created to explore the relationships and trends in the dataset:
Plotted Charts:
1. **Flight Number vs. Payload Mass**
2. **Flight Number vs. Launch Site**
3. **Payload Mass vs. Launch Site**
4. **Orbit Type vs. Success Rate**
5. **Flight Number vs. Orbit Type**
6. **Payload Mass vs. Orbit Type**
7. **Yearly Trend of Success Rate**

Visualization Techniques
1. **Scatter Plots**
    1. Highlight the relationships between variables.
    2. Useful for identifying potential patterns that can inform machine learning models.
2. **Bar Charts**
    1. Compare values across discrete categories.
    2. Provide insight into relationships between specific categories and their associated metrics.
3. **Line Charts**
    1. Show trends over time, particularly in time series data.
    2. Ideal for analyzing changes, such as the yearly trend in success rates.

https://github.com/mohamedbayoudh1/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

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

https://github.com/mohamedbayoudh1/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Markers for Launch Sites

NASA Johnson Space Center (Start Location)

- Marker with a **circle**, **popup label**, and **text label** placed using latitude and longitude coordinates.

All Launch Sites

- Markers with **circles**, **popup labels**, and **text labels** added for each launch site to showcase their geographical locations.
- Highlights proximity to the **Equator** and **coastlines** for better understanding of site placement.

Coloured Markers for Launch Outcomes

Green Markers: Represent successful launches.

Red Markers: Indicate failed launches.

Markers are grouped using a **Marker Cluster**, providing a clear view of success rates for each launch site.

Distance Visualization

Coloured Lines illustrate distances from the launch site **KSC LC-39A** to nearby points of interest:

- **Railways**
- **Highways**
- **Coastlines**
- **Closest City**

https://github.com/mohamedbayoudh1/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

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

https://github.com/mohamedbayoudh1/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

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

https://github.com/mohamedbayoudh1/IBM-Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

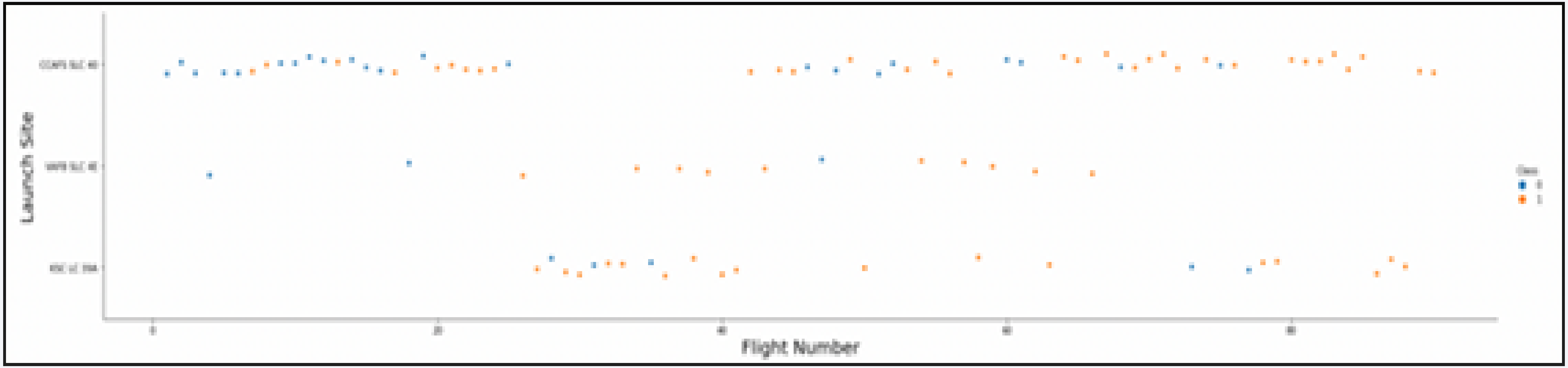
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

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.



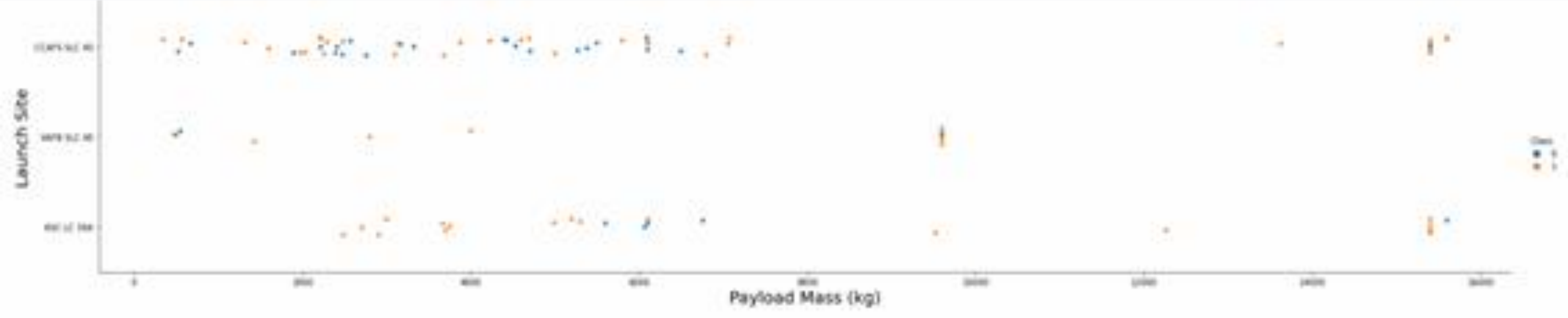
Payload vs. Launch Site

Explanation:

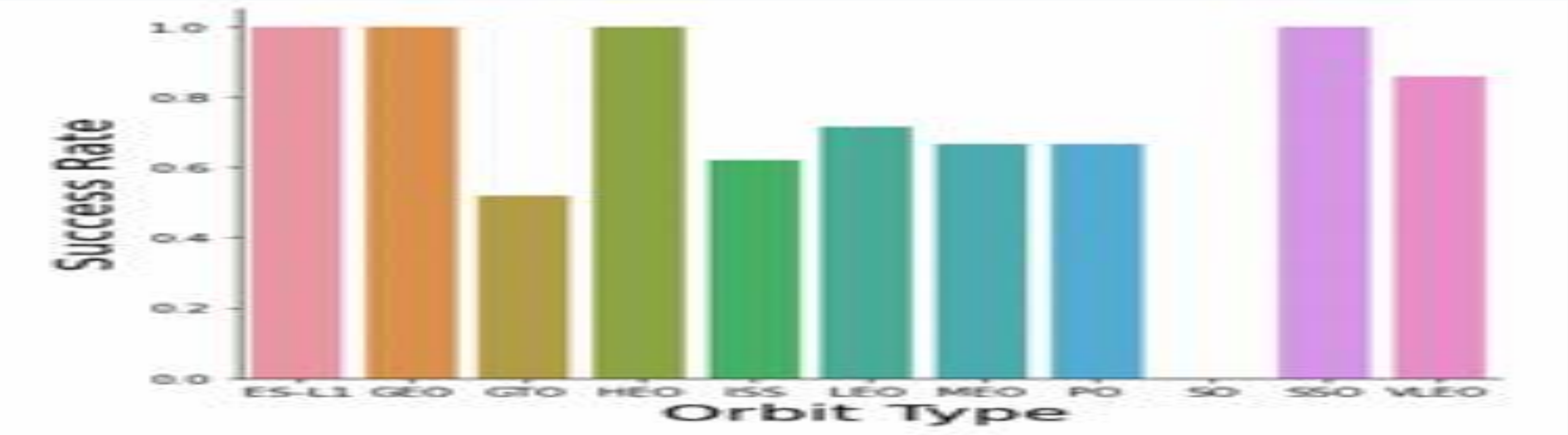
For every launch site the higher the payload mass, the higher the success rate.

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:

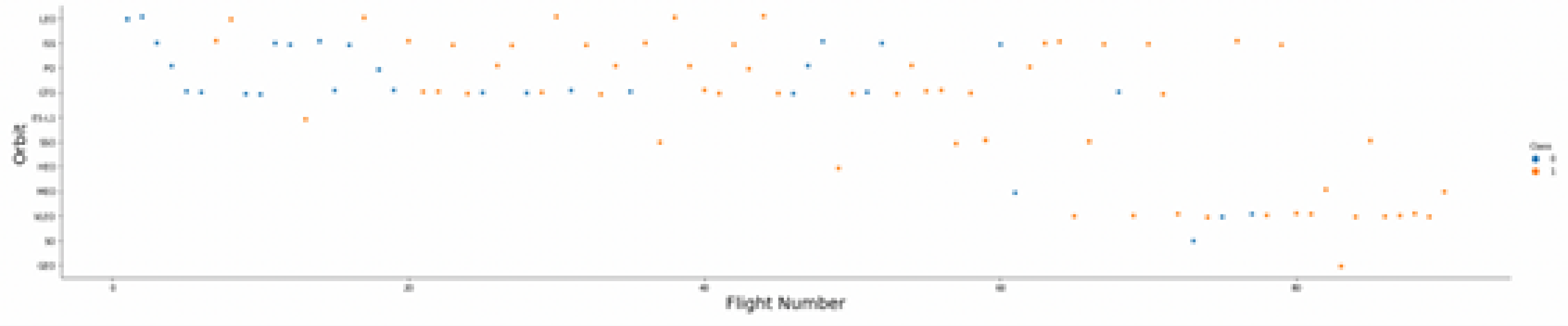
Orbits with 100% success rate:

- 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:

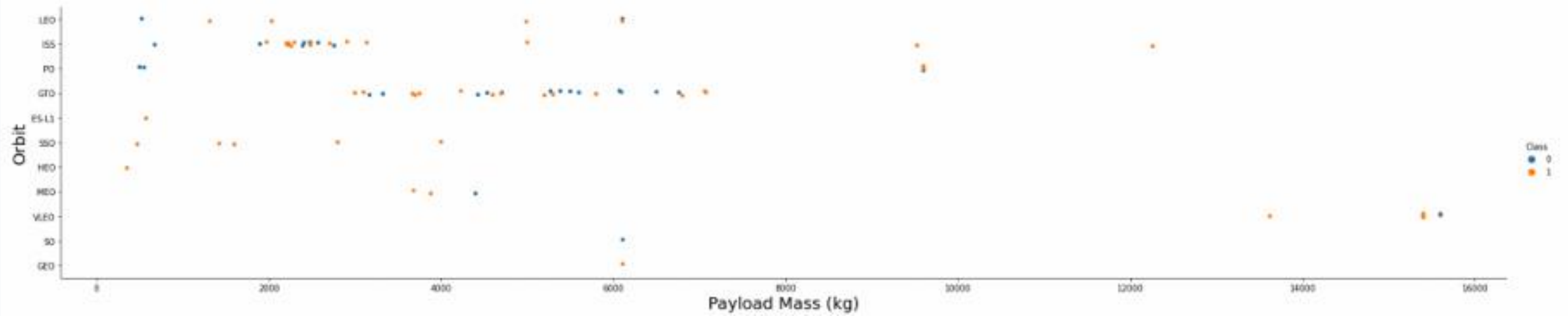
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

Explanation:

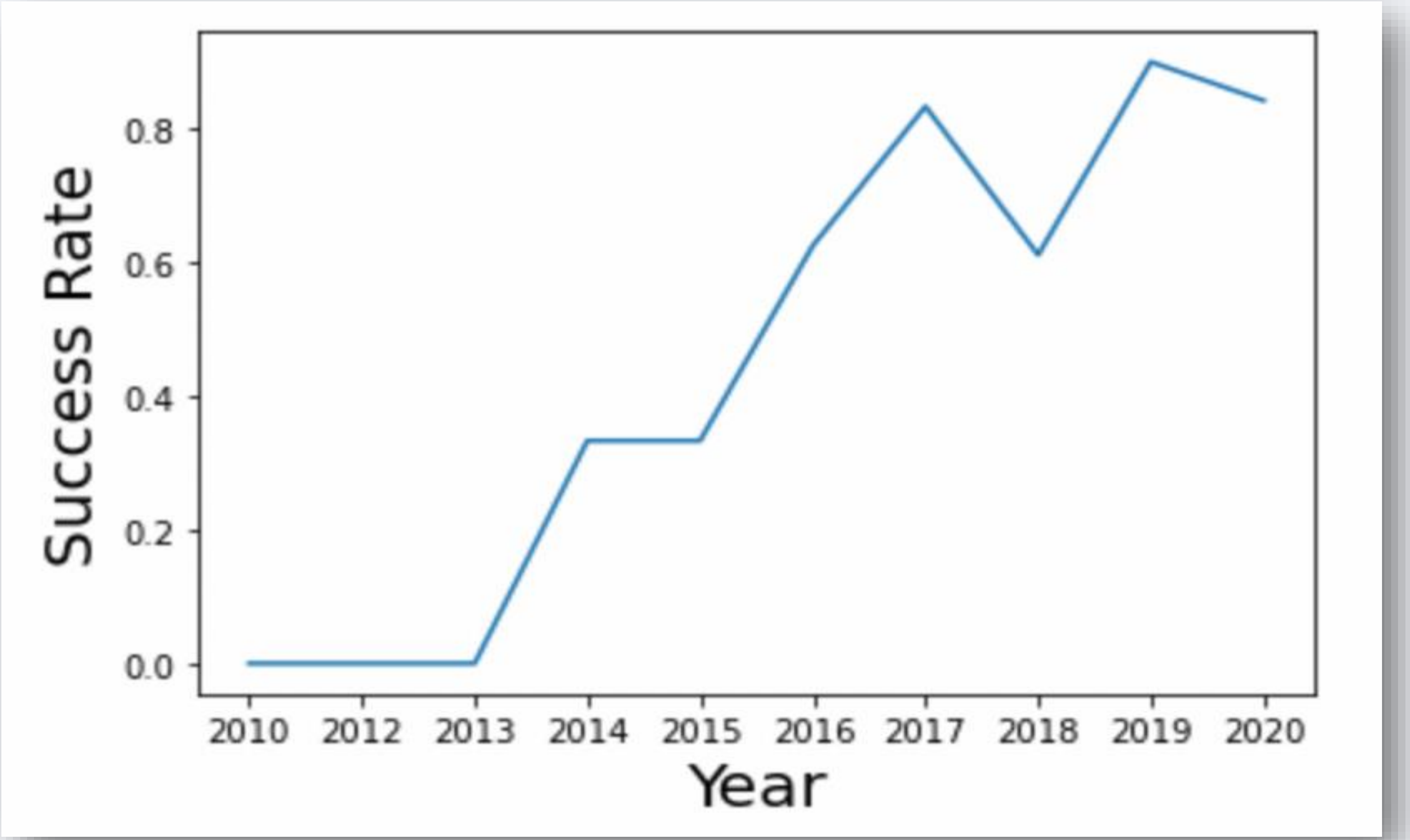
Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

Explanation:

The success rate since 2013 kept increasing till 2020.



All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

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

Out[4]:

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

Explanation:

Displaying the names of the unique launch sites 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:
Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

```
In [6]: %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.
Out[6]:
```

total_payload_mass
45596

Explanation: • Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

Explanation: • Displaying average payload mass carried by booster version F9 v1.1.

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:3
Done.
Out[7]:
```

average_payload_mass
2534

First Successful Ground Landing Date

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

```
In [8]: %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.
Out[8]:
```

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[9]:
```

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

Total Number of Successful and Failure Mission Outcomes

Explanation:

Listing the total number of successful and failure mission outcomes.

In [10]:
```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

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

Out[10]:

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

Boosters Carried Maximum Payload

Explanation:

Listing the names of the booster versions which have carried the maximum payload mass.

```
In [11]: %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.
```

Out[11]:

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

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

```
In [12]: %%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.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

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)

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

Explanation:
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.

```
In [13]: %%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.
```

Out[13]:

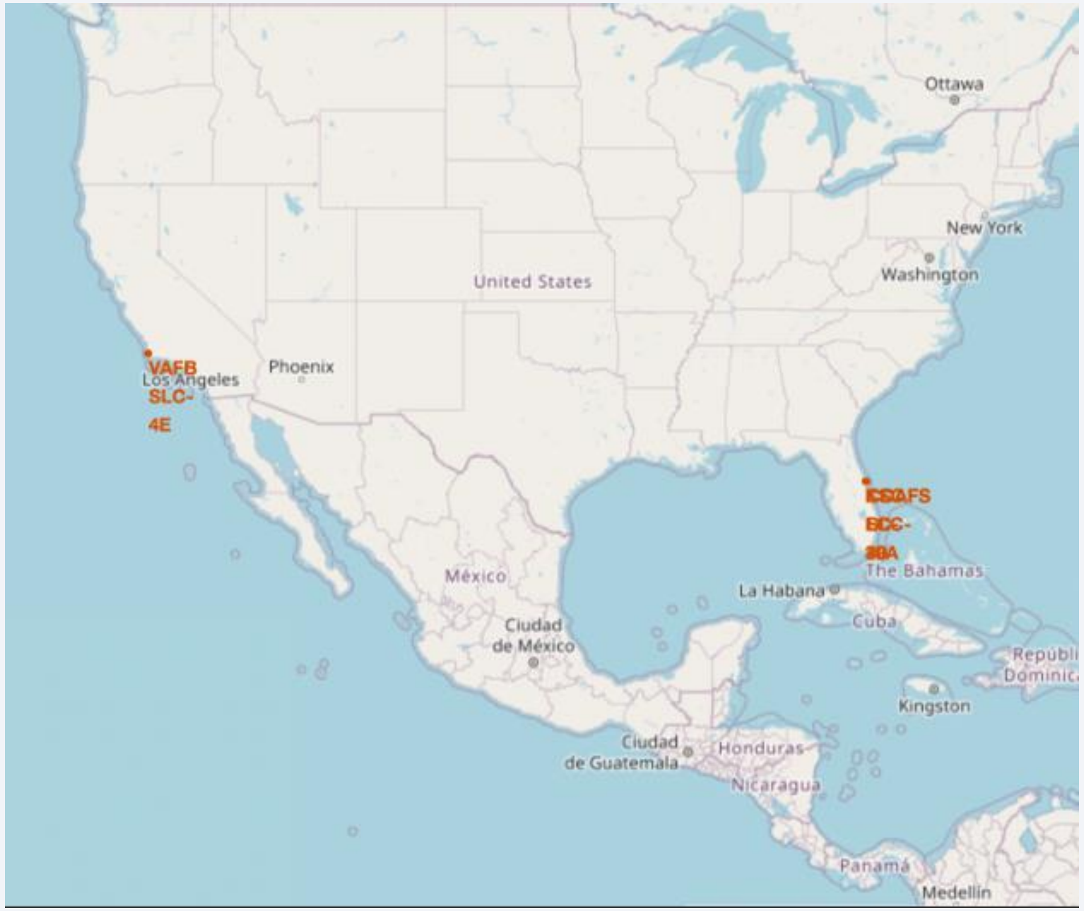
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on a global map

Explanation:
Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a 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. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit. • All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



Colour-labeled launch records on the map

Explanation:

• From the colour-labeled markers

we should be able to easily

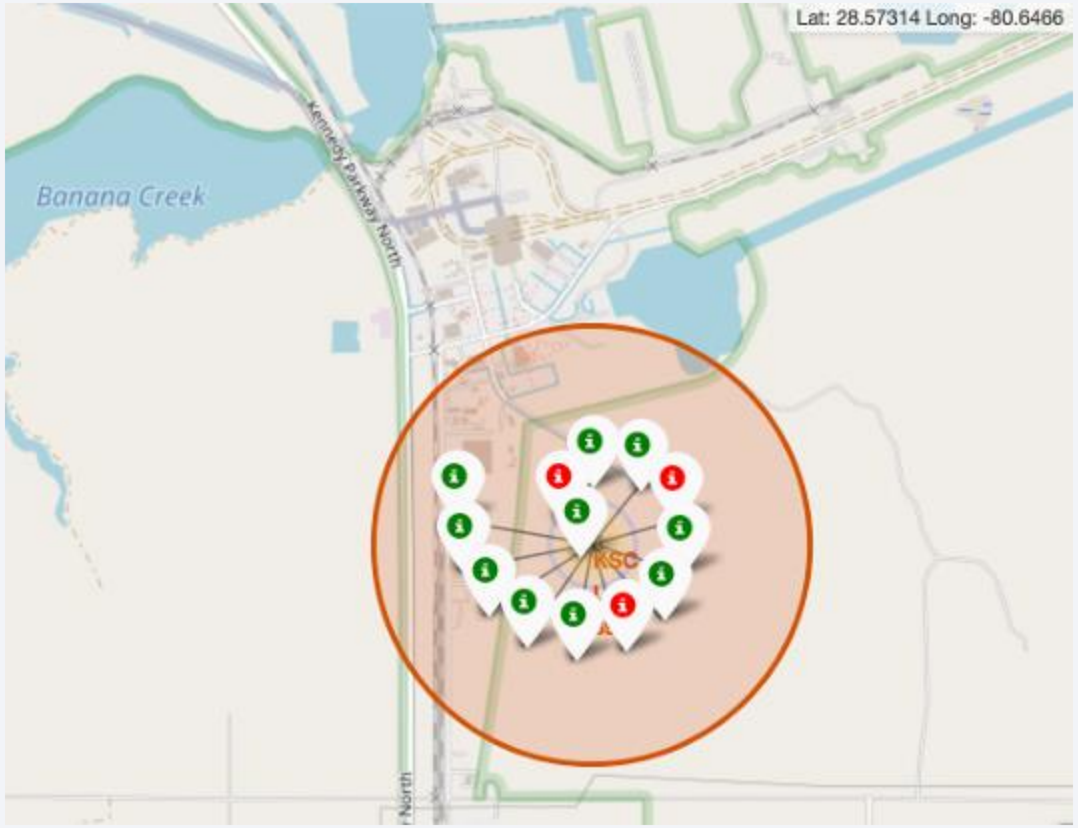
identify which launch sites have

relatively high success rates. - Green Marker = Successful

Launch - Red Marker = Failed Launch

• Launch Site KSC LC-39A has a

very high Success Rate



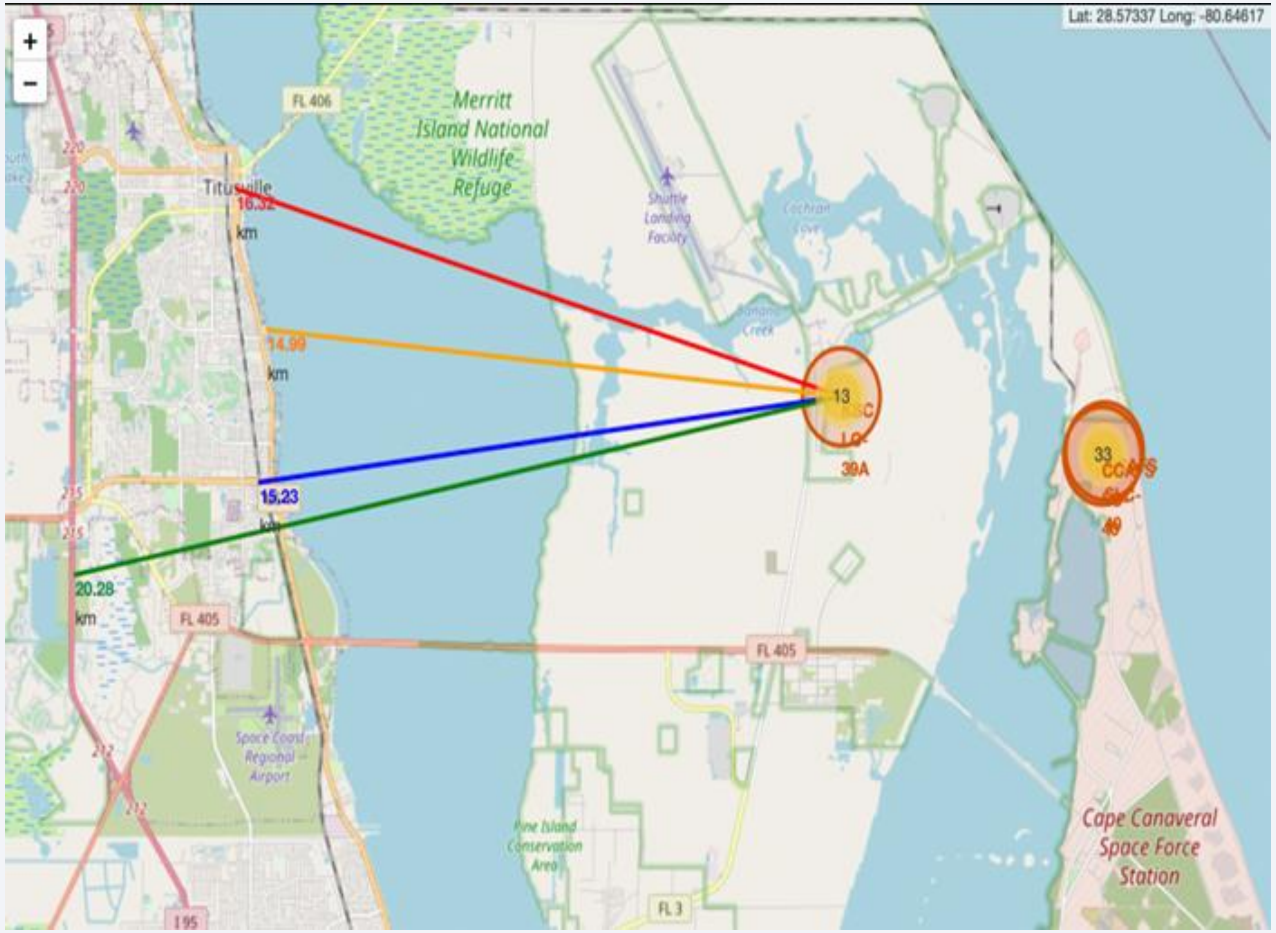
Distance from the launch site KSC LC-39A to its proximities

Explanation:

• From the visual analysis of the launch site KSC LC-39A we can clearly see that it is: - relative close to railway (15.23 km) – relative close to highway (20.28 km) - relative close to coastline (14.99 km)

• Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

• Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



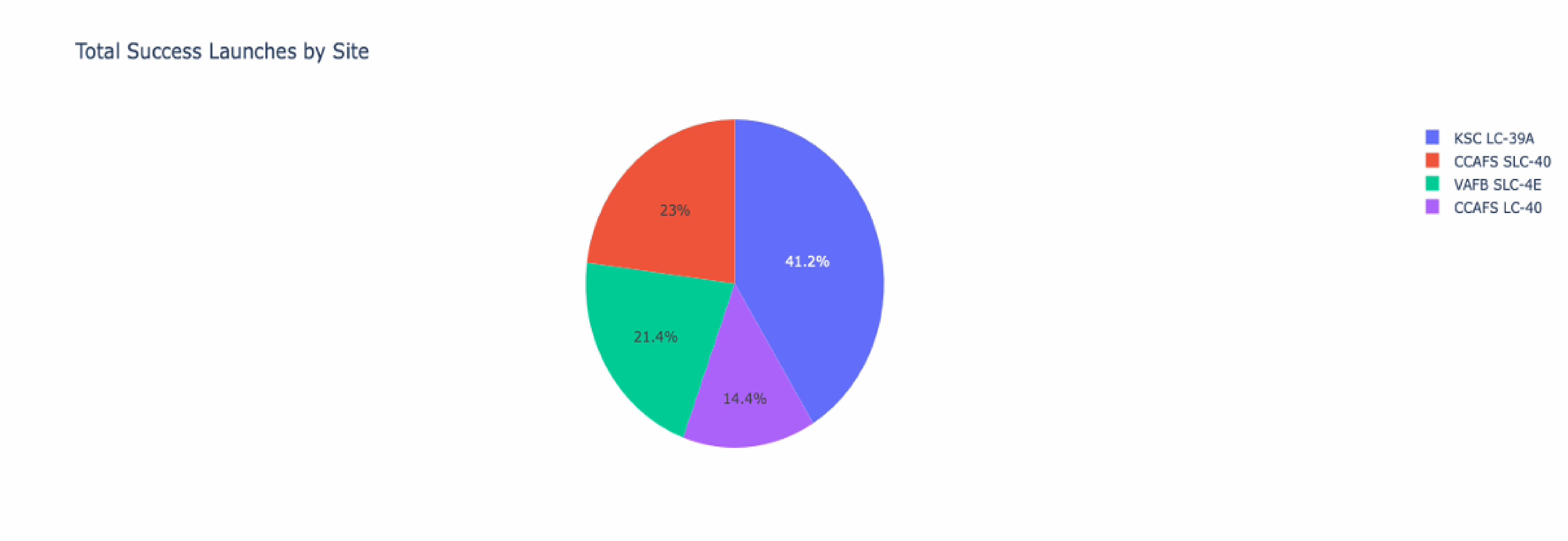
Section 4

Build a Dashboard with Plotly Dash

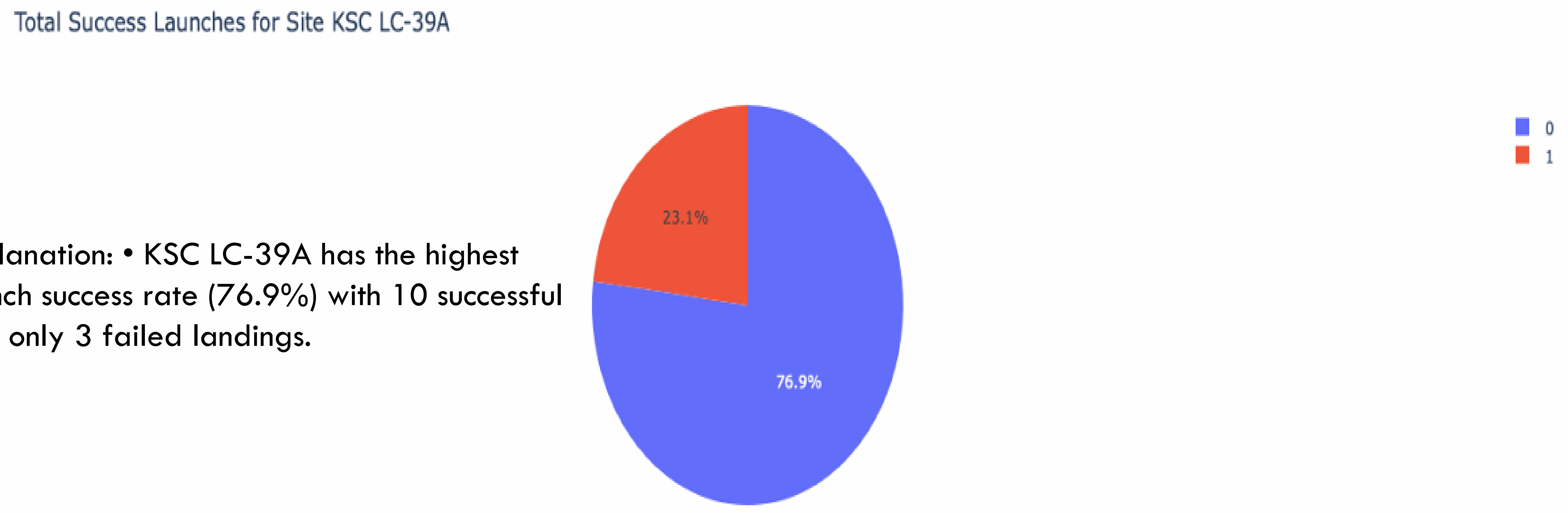
Launch success count for all sites

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%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites

Explanation:

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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

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 confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Scores 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

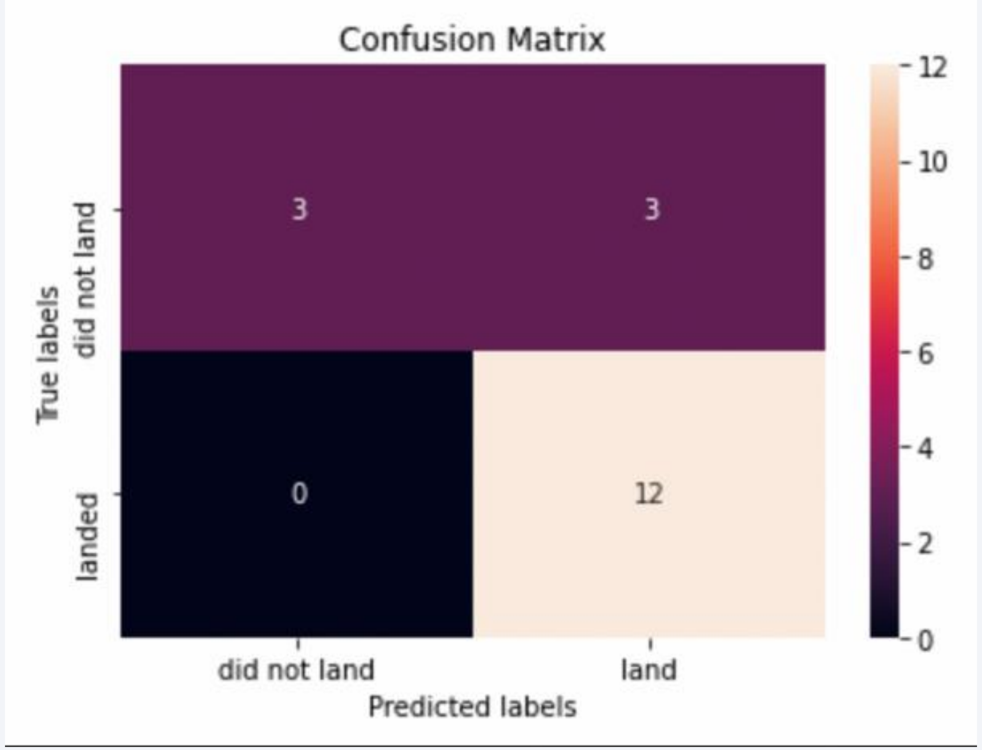
Scores and Accuracy of the Entire Data Set

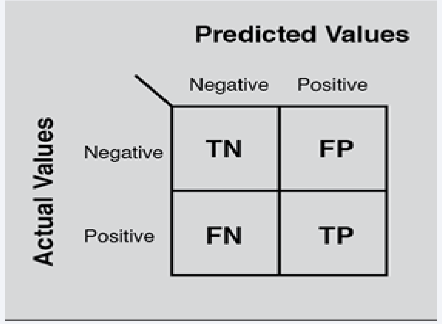
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

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.





Conclusions

The analysis revealed that the Decision Tree Model is the most effective algorithm for predicting launch outcomes in this dataset. Launches with lower payload masses are more likely to succeed, and most launch sites are strategically located near the Equator and coastlines to optimize efficiency and safety. Over the years, the success rate of launches has steadily increased, with KSC LC-39A emerging as the most reliable site. Additionally, certain orbits, including ES-L1, GEO, HEO, and SSO, demonstrate a 100% success rate, underscoring their dependability for successful missions.

Thank you!