IBM Developer
SKILLS NETWORK

Data Science Capstone Project

Ísar Daði Pálsson
07.05.2024

Outline

Executive Summary

Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
  - Data Visualization
  - SQL
- Interactive Visual Analytics
  - Folium Map
  - Plotly Dashboard
- Machine Learning Prediction

Summary of Results

- EDA Results
- Interactive Analytics Demo (Screenshots)
- Predictive Analysis Results

Introduction

Project Background

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions

- How do different variables such as payload mass, launch site and orbits affect the success of first stage landings ?
- Does the rate of successful landings change increase with each year?
- What is the best prediction algorithm we can use for binary classification to predict successful landings?

Section 1

Methodology

Methodology

Data Collection Methodology:

- Using SpaceX REST API
- Using Web Scraping from Wikipedia with BeautifulSoup

Performed Data Wrangling

- Filtering the data
- Remove or replace missing values
- Using One Hot Encoding to prepare the data for prediction analysis

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Data Collection

The data collection process involved both API requests from the SpaceX RESTful API and Web Scraping SpaceX´s Wikipedia entry for data.

I had to use both of these data collection methods in order to get more detailed information about the launches for a more detailed and accurate analysis

SpaceX REST API Data Columns

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

Wikipedia Web Scraping Data Columns

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

→ Parse response with .json() and turn into a Data frame with .json_normalize()

→ Request needed information from SpaceX API and apply custom functions

→ Create a Dictionary from obtained Data

→ Create a Data frame from Dictionary

→ Filter Data frame to only include Falcon 9 launches

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

→ Export the Data frame to CSV

[Github URL: Data Collection - API]()

Data Collection - Scraping

Web Scrape SpaceX Falcon 9 launch records with BeautifulSoup

→ Extract column names from BeautifulSoup HTML table header

→ Extract data from HTML tables

→ Construct dictionary from extracted data

→ Create a Data frame from Dictionary

→ Export the Data frame to CSV

[Github URL: 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.

In this lab we will mainly convert those outcomes into Training Labels with ‘1’ means the booster successfully landed ‘0’ means it was unsuccessful. This will make the data viable for Machine Learning (ML) prediction algorithms.

1. Perform exploratory data analysis and determine training labels
2. Calculate the number of launches on each launch site
3. Calculate the number and occurrence of each orbit
4. Calculate the number and occurrence of mission outcome of the orbits
5. Create a landing label from the Output column
6. Export the newly labelled data to CSV

Github URL: Data Wrangling

EDA - Data Visualization

These Charts Were Plotted:

- Flight Number vs. Payload Mass (Scatter)
- Flight Number vs. Launch Site (Scatter)
- Payload Mass vs. Launch Site (Scatter)
- Orbit Type vs. Success Rate (Bar)
- Flight Number vs. Orbit Type (Scatter)
- Payload Mass vs Orbit Type (Scatter)
- Success Rate Yearly Trend (Line)

What kind of Plots:

Scatter plots reveal correlations between two continuous numerical variables. Strong positive or negative correlations, if significant, can be learned by machine learning models.

Bar charts compare discrete categories and show the relationship between a category and a measured value.

Line charts show data trends over a specified time period.

Github URL: EDA – Data Visualization

EDA - SQL

These SQL queries were performed:

- Names of the unique launch sites used for Falcon 9 launch missions.
- Records where launch sites begin with the string ‘CCA’ .
- Total payload mass carried by boosters launched by NASA (CRS).
- Average payload mass carried by booster version F9 v1.1.
- Date when the first landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have a payload mass greater than 4000kg and less than 6000kg.
- Total number of successful and failure mission outcomes.
- Name of the booster versions which have carried the maximum payload mass.
- Failed landing outcomes in drone ship, their booster versions and launch site names for the months of year 2015.
- 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 URL: EDA - SQL

Interactive Visual Analysis – Folium Map

- Added marker with a circle and both text and popup labels for all launch sites, using their latitude and longitude coordinates.

- Added markers for success (Green) and failures (Red) using the built-in Marker Cluster to identify launch site success/failure ratio.

- Added lines to show distances between launch site and relative points of interest such as a Railway, Highway, Coastline and closest City.

Github URL: Interactive Visual Analysis – Folium Map

Interactive Visual Analysis - Dash Dashboard

Launch Site Dropdown List:

- Added a dropdown list to enable launch site selection of all launch sites or a specific launch site.

Pie Chart showing Launch Site success rate:

- Added a pie chart that shows the success rate of all launch site or the success/failure rate of a specific launch site, depending on which is selected in the dropdown list.

Payload Mass Slider:

- Added a slider that allows you to select a range of Payload Mass between 0kg to 10.000kg, such as 2.000kg to 8.000kg and 6.000kg to 10.000kg.

Scatter chart of Payload Mass vs. Launch Success for different Booster Versions:

- Added a Scatter chart the shows the correlation between successful launches and payload mass, for each booster versions used by the Falcon 9.

Github URL: Interactive Visual Analysis – Dash Dashboard

Machine Learning Prediction

1. Create a NumPy Array with the method .to_numpy() from the column ‘Class’ in the Data
2. Standardize, fit and transform the data with a StandardScaler()
3. Use the function train_test_split() to split the data into training and testing data
4. Create a GridSearchCV object to find the best parameters
5. Apply the GridSearchCV object to LogReg, SVM, Tree and KNN models
6. Calculate the test data accuracy for each model using the method .score()
7. Create a confusion matrix for each model and examine it
8. Find the best performing model with the Jaccard Score and F1_Score metrics

[Github URL: Machine Learning Prediction]()

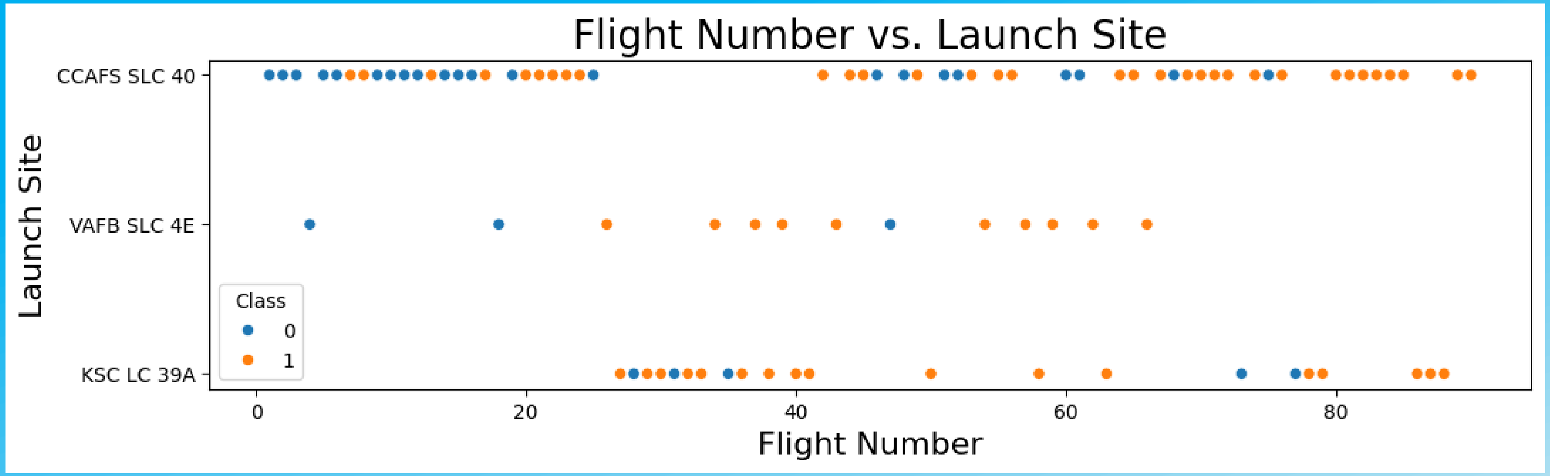
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Machine Learning Prediction results

Section 2

EDA – Data Visualization

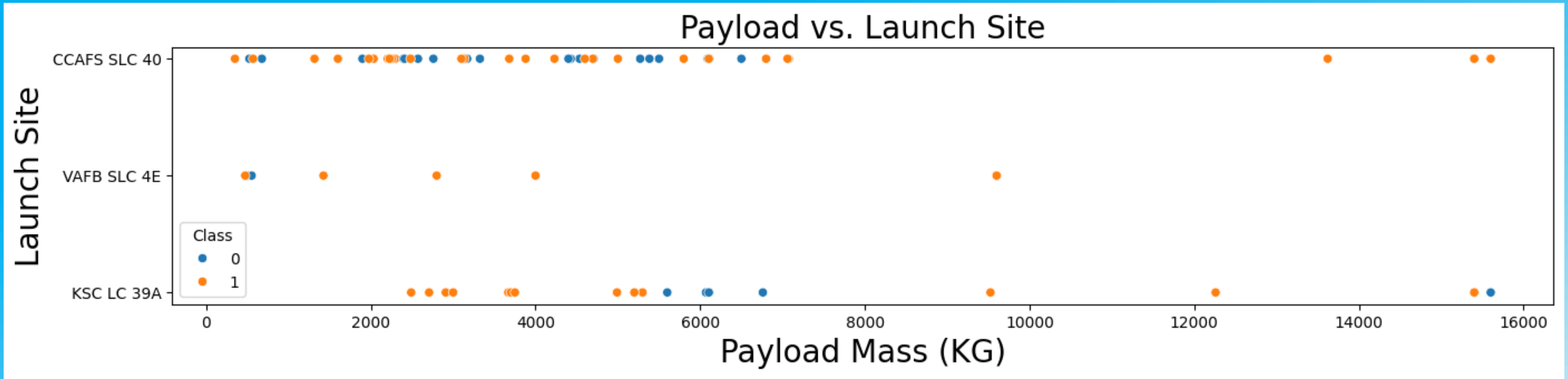
Flight Number vs. Launch Site



Observations

- Launch Site CCAFS SLC 40 is responsible for the largest amount of launches
- Launch Sites VAFB SLC 4E and KSC LC 39A have a higher success rate than Launch Site CCAFS SLC 40
- Earlier flights (0-20) have a high failure rate and later flights (60+) have a high success rate, indicating a steady increase in success as more launches are made.
- With these observations, it is reasonable to infer upcoming flights will have a higher success rate than previous flights.

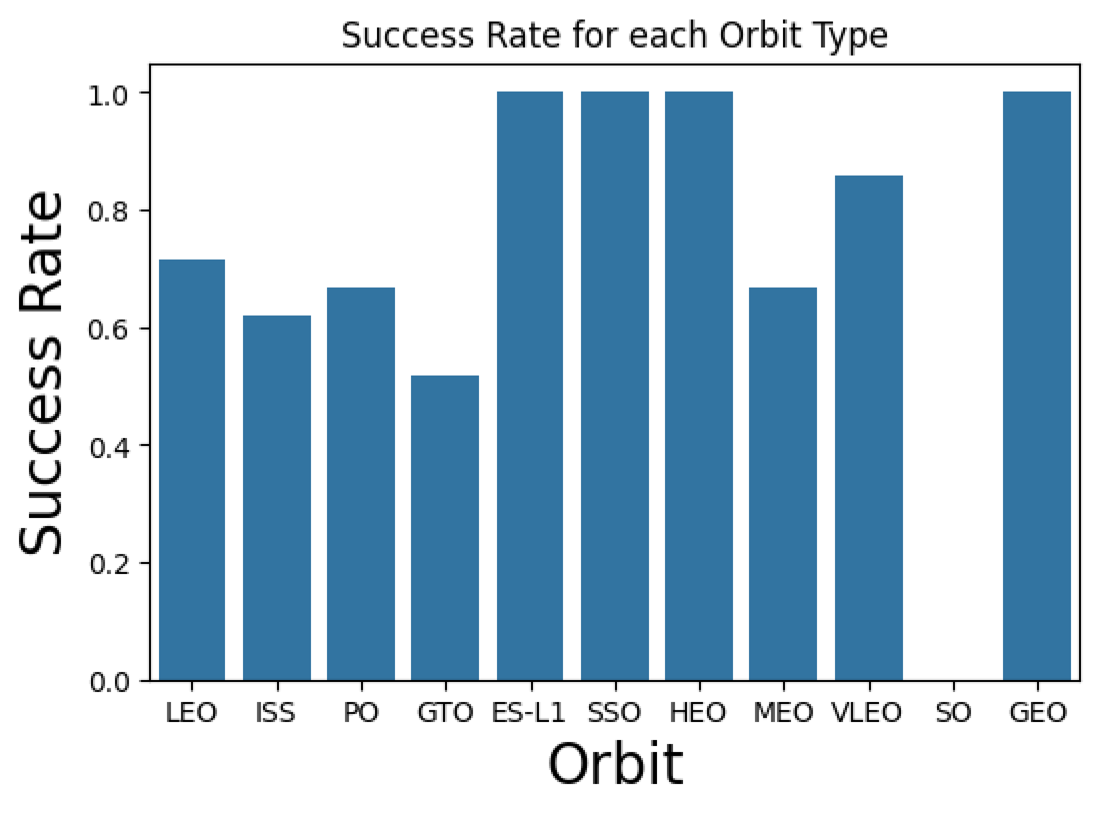
Payload vs. Launch Site



Observations

- Most launches have a payload mass under 8.000kg.
- Most launches over 8.000kg were successful launches.
- All launches under 5.000kg at launch site KSC LC 39A were successful launches.
- All launches over 1.000kg at launch site VAFB SLC 4E were successful launches.

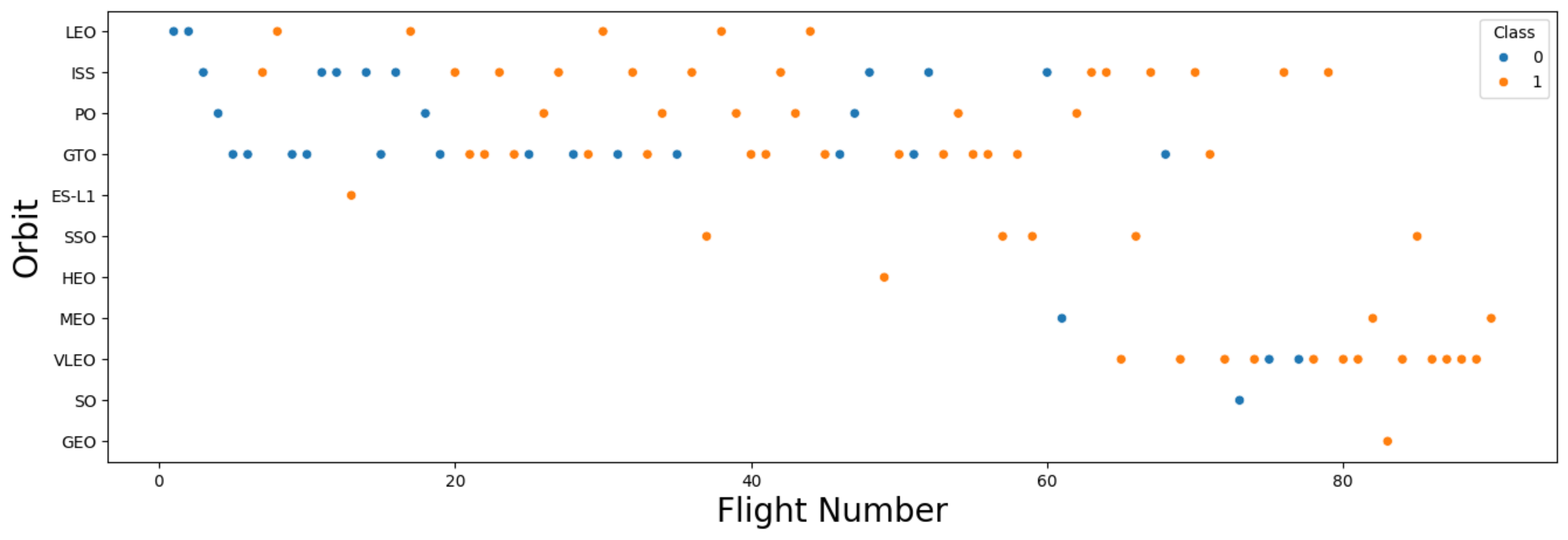
Success Rate vs. Orbit Type



Observations

- Launches for ES-L1, SSO, HEO, GEO were 100% successful
- Launches for SO had a 0% success rate
- Launches for LEO, ISS, PO, GTO, MEO and VLEO had success rates ranging from 50% to 85%

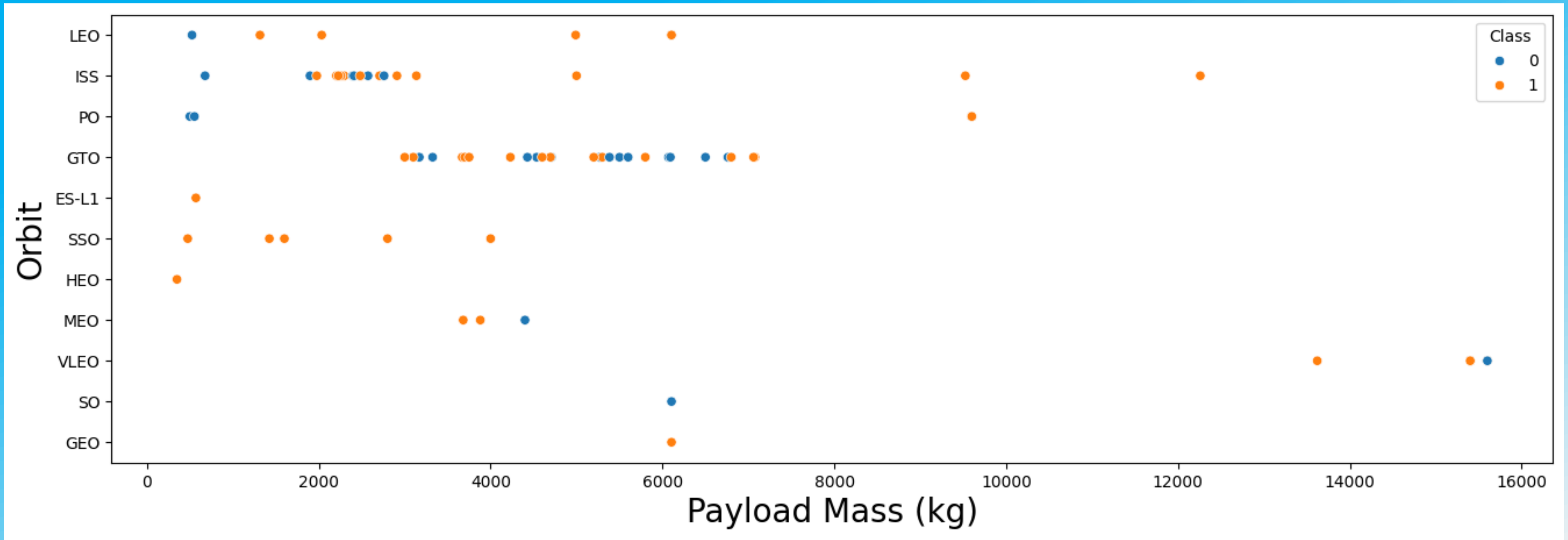
Flight Number vs. Orbit Type



Observations

- The Orbits with 100% success rates had very few launches, these results could be skewed to a small sample size. Same with orbit SO, which has a 0% success rate with only a single launch.
- Successful launches into orbit LEO seems related to number of flights.

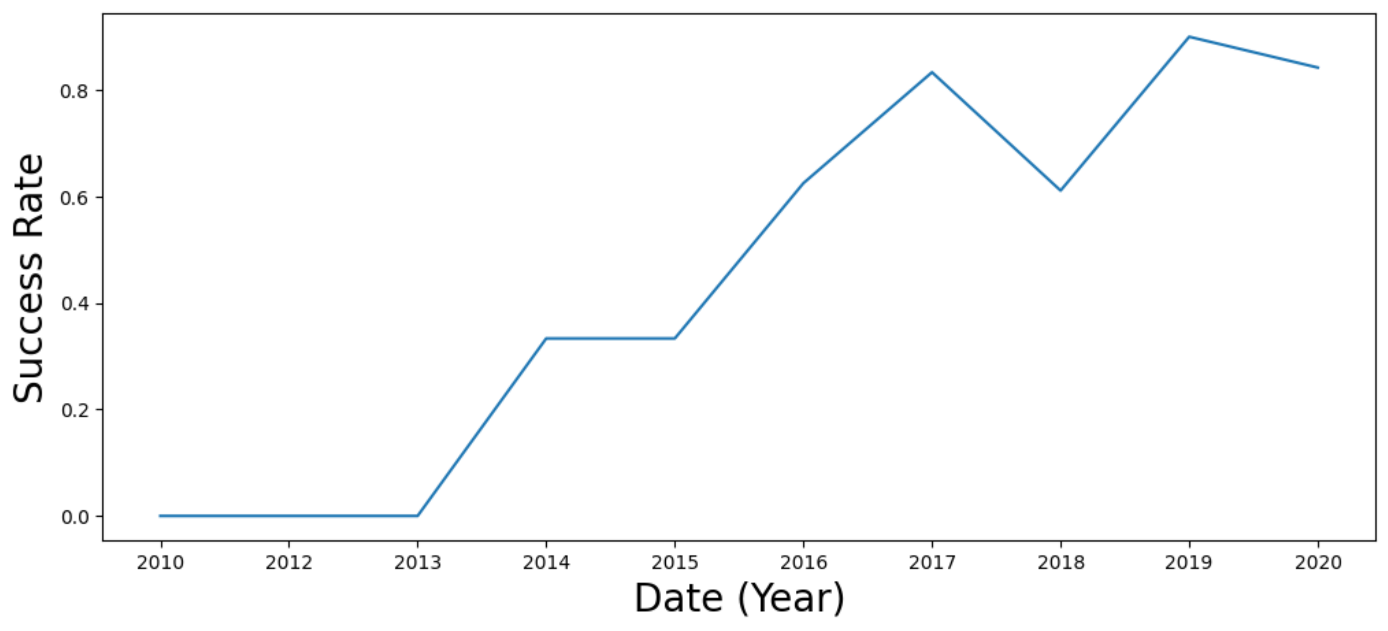
Payload vs. Orbit Type



Observations

- Only orbits ISS, PO and SO had launches with over 9.000kg payloads.
- Only orbit GTO seems to show a negative relation between success rates and increasing payload mass.

Launch Success Yearly Trend



Observations

- Success rates started increasing from 2013 up until 2020, with the only drop in the years 2017 - 2018

Section 3

EDA – SQL

All Launch Site Names

```
%%sql
SELECT DISTINCT Launch_Site
FROM SPACEXTABLE
```

[13] ✓ 0.0s

* 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
```

✓ 0.0s 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

```
%%sql

SELECT Sum(PAYLOAD_MASS__KG_) as 'Total Payload Mass', Customer
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
```

[15] ✓ 0.0s

* sqlite:///my_data1.db
Done.

Total Payload Mass	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as 'Average Payload Mass'
FROM SPACEXTABLE
WHERE Booster_Version Like '%F9 v1.1%'
```

✓ 0.0s

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

Average Payload Mass
2534.6666666666665

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date) as 'First Successful Landing'
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)'
```

✓ 0.0s

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

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000
```

✓ 0.0s

```
* 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

```
%%sql

SELECT Mission_Outcome, COUNT(*) as Total
FROM SPACEXTABLE
GROUP BY Mission_Outcome
```

✓ 0.0s

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

Mission_Outcome	Total
Failure (in flight)	1
Success	99
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
)
```

✓ 0.0s

```
* 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 substr(Date, 7,1) AS Month, *
FROM SPACEXTABLE
WHERE substr(Date,0,5) = '2015'
AND Landing_Outcome = 'Failure (drone ship)'
GROUP BY substr(Date, 6,2)
```

✓ 0.0s

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

Month	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
1	2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
4	2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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

```
%%sql

SELECT Landing_Outcome, COUNT(Landing_Outcome) as outcome_count
FROM SPACEXTABLE
WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY outcome_count DESC
```

✓ 0.0s

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

Landing_Outcome	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

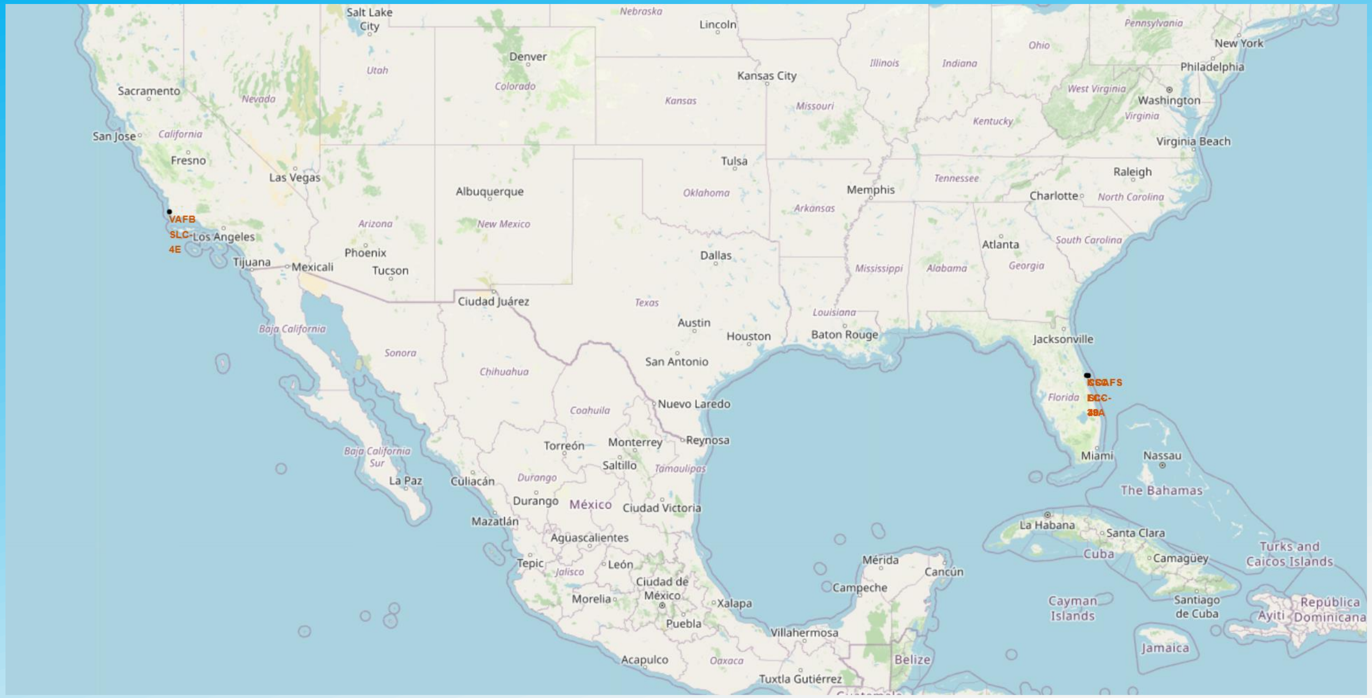
Section 4

Interactive Visual Analytics
Folium Map

All Launch Sites on a Global Map

Observations

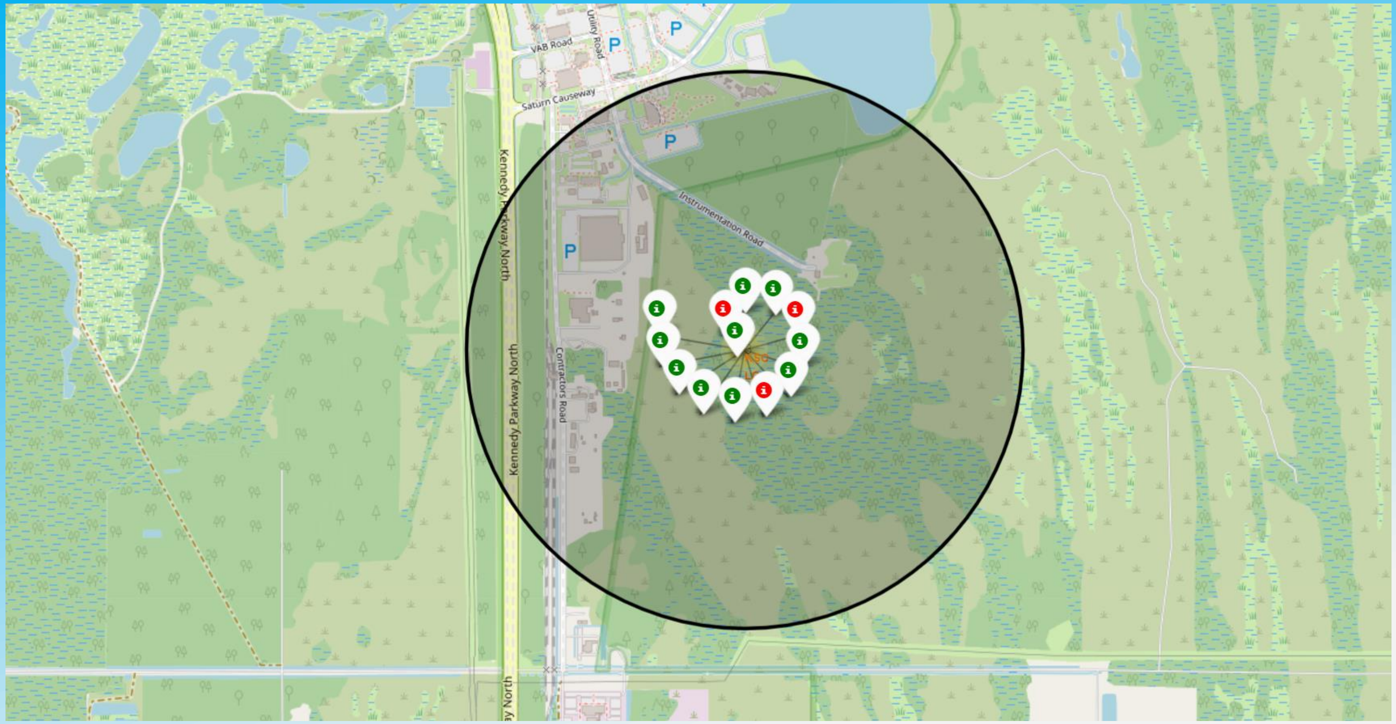
- Most of the launch sites are aligned to the Equator line.
  The speed of Earth's rotation is 1670 km/h and because of inertia, the launch rocket will maintain the same speed after launch.
  This speed helps the spacecraft stay in orbit.
- Damage is minimized due to the launch sites being stationed along the coastlines, increasing chances of a crash happening on a body of water, not inland.
- Launch sites are stationed away from largely populated areas but still nearby populated cities, such as Los Angeles and Miami.
- We can infer that if another launch station were built, it would be along a coastline, near a largely populated city and near the Equator line, most likely along the coastline near Houston.



Colored Launch Site Success Rates

Observations

- Every single launch on the launch sites has been marked so that you can visually indicate which the success rate of each launch site
  - Green = Successful
  - Red = Failure
- This launch site, KSC LC-39A has a high launch success rate as we can see with the amount of green markers compared to red markers.



Launch Site Proximities

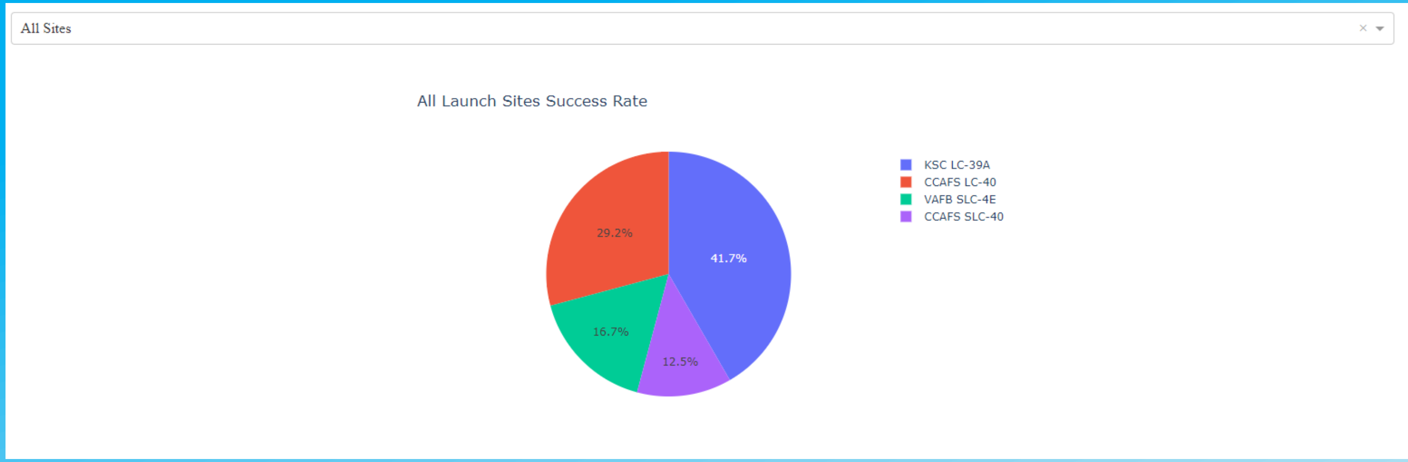
Observations

- Launch Site KSC LC-39A has a relatively short distance from
  - Titusville (16.35 KM)
  - A Coastline (7.12 KM)
  - A Railway (15.47 KM)
  - A Highway (19.96KM)
- These distances are relatively short for a rocket intended for interorbital travel and could therefore be at risk from falling debris from a failed launch.

Section 5

Interactive Visual Analytics Dash Dashboard

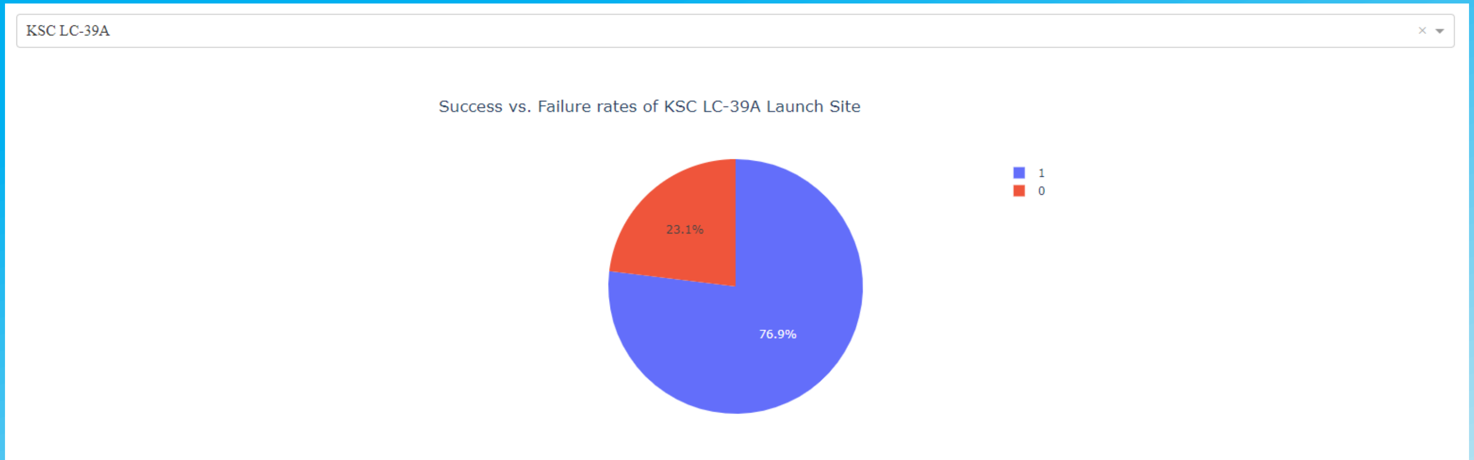
All Launch Sites Success Rate



Observations

- According to the success rate, Launch Site KSC LC-39A is responsible for more than a third of all successful launches, at a percentage of 41.7%
- The Launch Site CCAFS SLC-40 is responsible for the least of all successful launches, at a percentage of 12.5%
- The percentages are not to be confused with direct success rates of each launch site, just their contribution to the total of all successful launches

Launch Site - Highest Success Rate



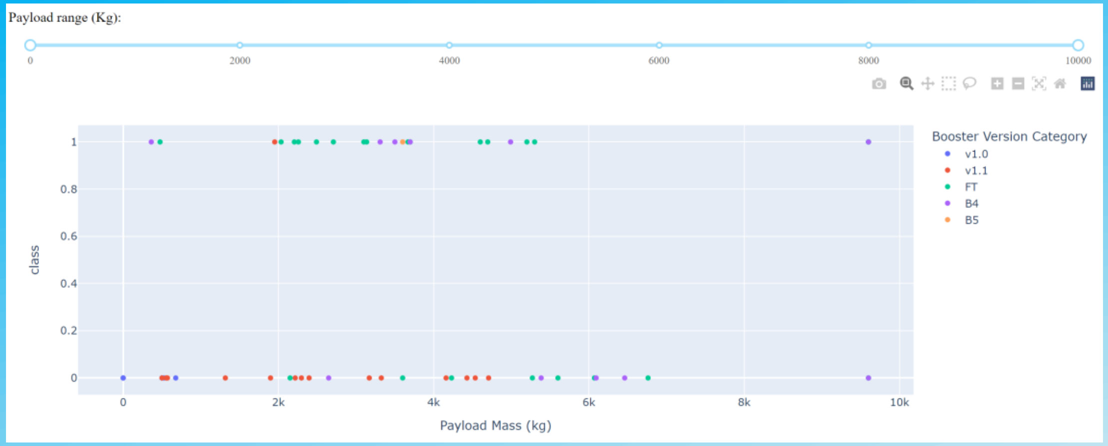
Observations

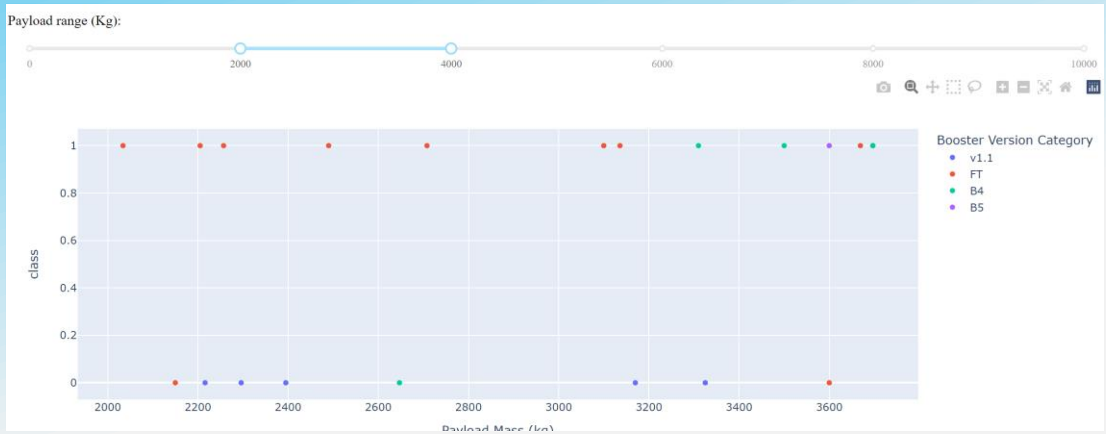
- Launch Site KSC LC-39A has a success rate of 76.9%, with 10 successful launches and 3 failed launches

Payload Mass vs. Success Rate

Observations

- The Booster Version that has the highest number of successful launches is the Booster Version FT
- The highest number of launches happen with payloads in the range of 2.000Kg to 4.000Kg, numbering 20 launches out of the 48 total launches in range of 0Kg to 10.000Kg
- The number of successful launches are 12 in the range of 2.000Kg to 4.000Kg out of 21 total in the range of 0Kg to 10.000Kg, accounting for more than half of all successful launches, at a percentage of 51.7%





Section 6

Machine Learning Prediction

Classification Accuracy

Observations

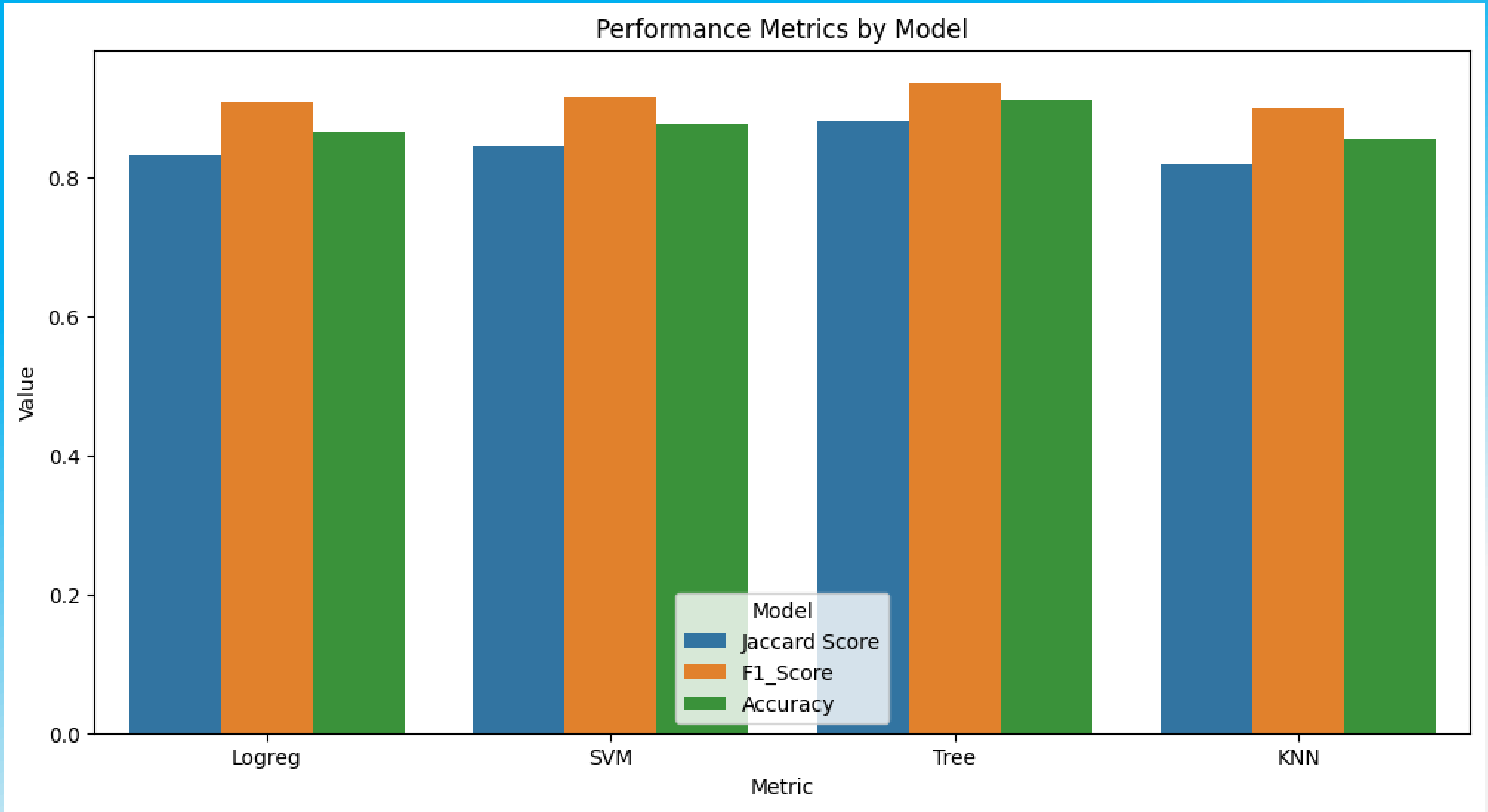
- The sample size on the Test Set was simply too small, resulting in all the scores being the same. Which classification model is the best could not be determined this way.
- I decided to test on the entire Data Set. This has a particular trade-off, the negative being that the accuracy is not fully out-of-sample but the positive being a more detailed result from each model working with a larger sample size.
- With these results, we can determine that the Decision Tree Model is the best performing model, outclassing all the other models on all metrics.
- For those more visually inclined, you can see a bar chart of the result of the testing the entire data set on the next slide.

Scores only on 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 on Entire Data Set

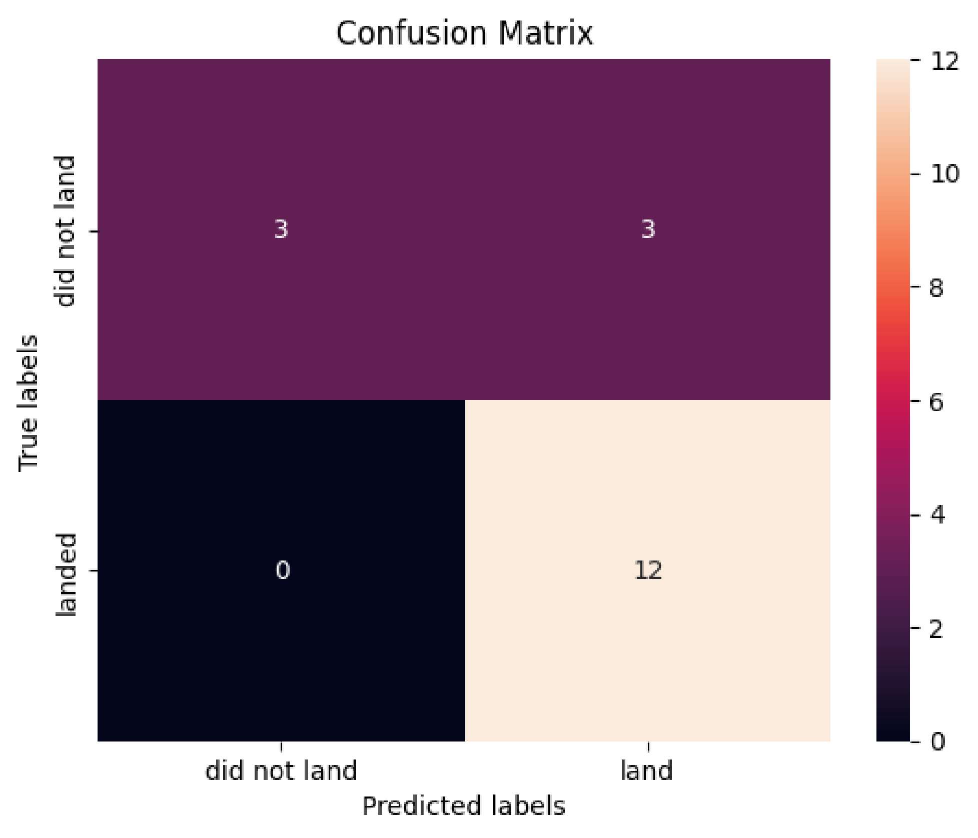
	Logreg	SVM	Tree	KNN
Jaccard Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556


Performance Metrics by Model
Value
0.0
0.2
0.4
0.6
0.8
Model
Jaccard Score
F1_Score
Accuracy
Logreg
SVM
Tree
KNN
Metric

Confusion Matrix

Observations

- The Confusion Matrix was made with the Decision Tree Model using only the Test Set, ensuring the result was fully out-of-sample.
- We can see that there are no false negatives and only three false positives. This indicates that we need to finetune the model to reduce the rate of false positives.
- This shows that the Decision Tree can properly classify the data, with a 100% accuracy on the negative class and an 80% accuracy on the positive class.



Conclusions

- Decision Tree Model is the best classification algorithm for this Data Set
- Most Launch Sites are stationed near the Equator Line
- KSC LC-39A is the launch site with the highest success rate
- Launch success rates have been increasing yearly
- Launches with payloads between 2.000kg to 4.000kg are more than half of all successful launches
- Orbits ES-L1, GEO, HEO and SSO have a 100% launch success rate
- It is very rare for a launch to have a payload mass heavier than 8.000kg

Thank you!