Data Science Capstone project

Jacopo Castellani

Sept. 26th 2021

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
- Predictive analysis (Classification)

Summary of all 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 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.

Problems you want to find answers

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- 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
- **Perform data wrangling**
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- **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 evaluation of classification models to ensure the best results

Methodology

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.

We had to use both these data collection methods in order to get complete information about the launches for a more detailed analysis.

The following 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.

The following 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 from SpaceX API the rocket launch data
- Decoding the response content with .json()
- Turning the response content into a dataframe with .json_normalize()
- Requesting information about the launches from SpaceX API with custom functions
- Organising our data into a dictionary
- Creating a dataframe from the dictionary
- Filtering the dataframe including Falcon 9 launches only
- Replacing missing values of Payload Mass column with the .mean () of the column
- Exporting the data into CSV

GitHub URL 1

Data collection – Web scraping

- Requesting from Wikipedia the Falcon 9 launch data
- Creating a BeautifulSoup object from the HTML response
- Extracting all column names from the HTML table header
- Collecting the data by parsing HTML tables
- Organising data into a dictionary
- Creating a dataframe from the dictionary
- Exporting the data into CSV

GitHub URL 2

Data wrangling

- Perform exploratory Data Analysis
- Determine training labels
- Compute the number of launches on each site
- Compute the number and occurrence of each orbit
- Compute the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from column “Outcome”
- Exporting the data into CSV

In the dataset, there are several different cases where the booster did not land successfully.

Those outcomes were converted into training label; namely, *0* means the launch was unsuccessful, *1* means the booster landed successfully.

GitHub URL 3

EDA with data visualization

The following 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
- 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 URL 4

EDA with SQL

The following SQL queries were performed:

- 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 between 4000 and 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 between 2010-06-04 and 2017-03-20 in descending order

GitHub URL 5

Build an interactive map with Folium

The following markers for all launch sites were added:

- Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- 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.

The following coloured markers for the launch outcomes for each launch site were added:

- Coloured markers for 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 were displayed with:

- Coloured lines to show distances between a given launch site and its proximities (like Railway, Highway, Coastline and Closest City).

GitHub URL 6

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:

  A dropdown list allows Launch Site selection.

- Pie Chart showing Success Launches (All Sites/Certain Site):

  A pie chart shows 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:

  A slider allows to select Payload range.

- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

  A scatter chart shows the correlation between Payload and Launch Success.

GitHub URL 7

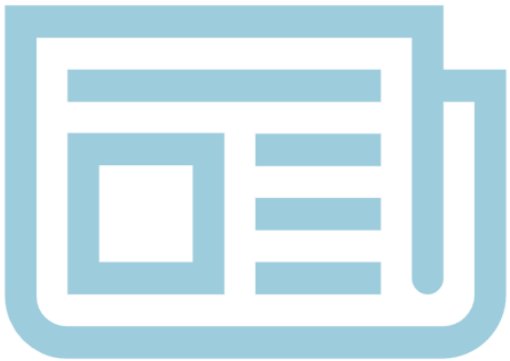
Predictive analysis (Classification)

- Creating a NumPy array from the column “Class”
- Standardizing the data with StandardScaler
- Fitting and transforming the data
- 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 metric

GitHub URL 8

Results

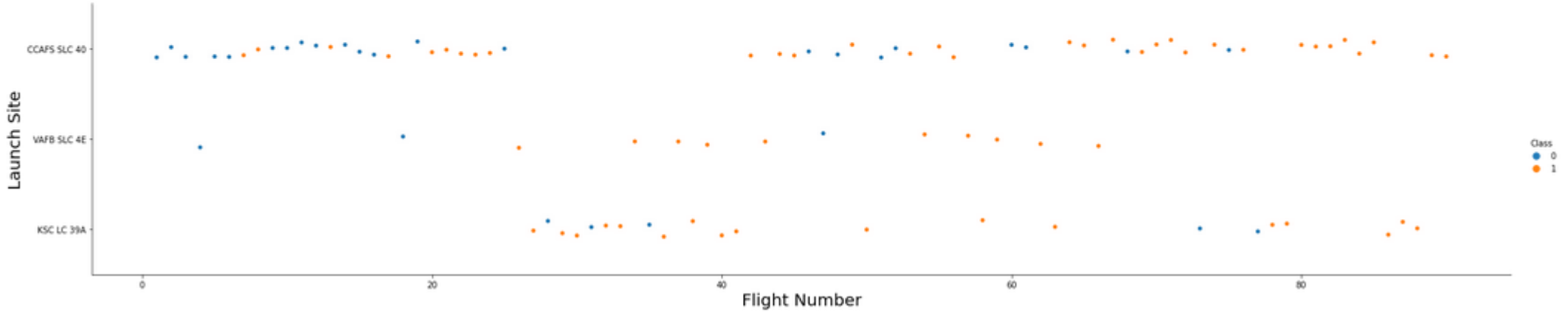
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



EDA with Visualization

Flight Number vs. Launch Site

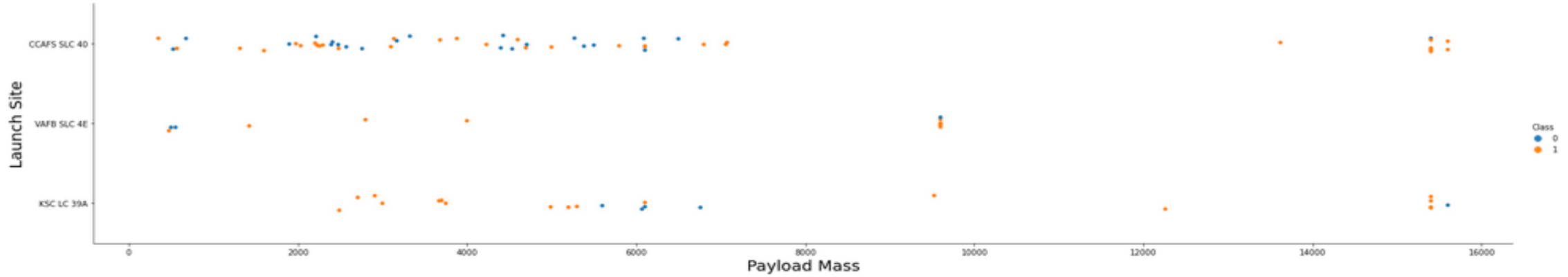
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(x='FlightNumber', y='LaunchSite', hue='Class', data=df, aspect=5)
plt.xlabel('Flight Number', fontsize=20)
plt.ylabel('Launch Site', fontsize=20)
plt.show()
```



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC40 launch site has about one half of all launches, while VAFBSLC4E and KSCLC39A have higher success rates.
- We can assume that newer launches have a higher probability of success.

Payload vs. Launch Site

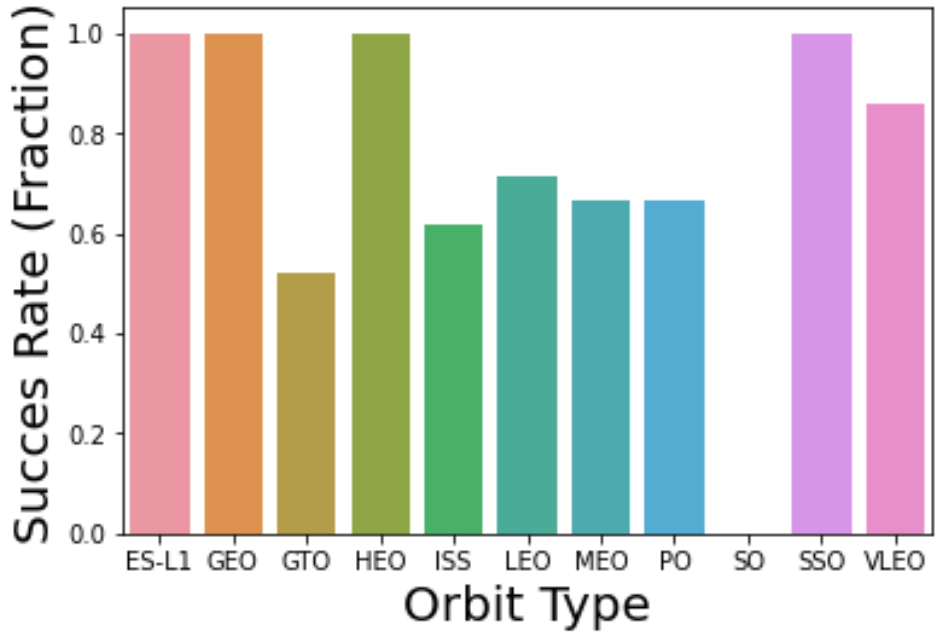
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value

sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



- 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

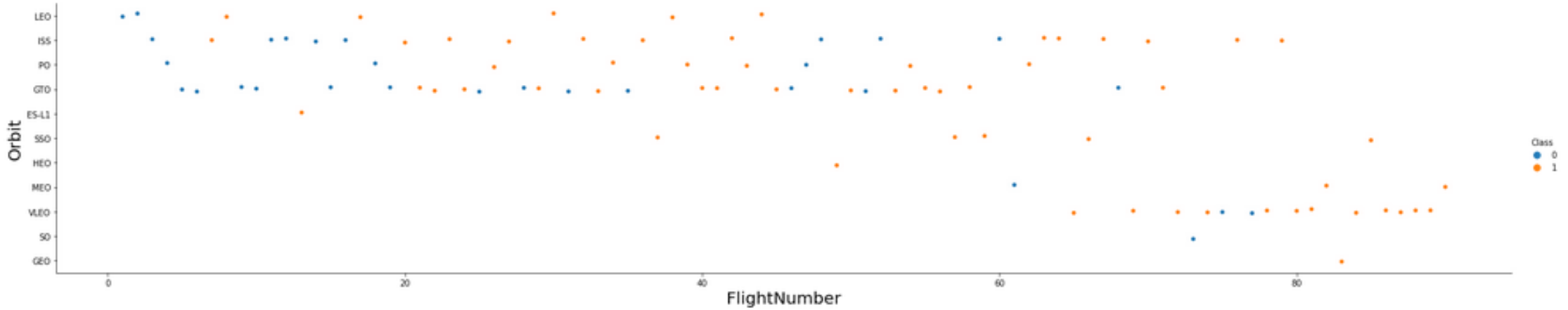
```
sns.barplot(x=results_df.index, y="Class", data=results_df)
plt.xlabel("Orbit Type",fontsize=20)
plt.ylabel("Succes Rate (Fraction)",fontsize=20)
plt.show()
```



- Orbits ES-L1, GEO, HEO, SSO have a 100% success rate
- Orbit SO has a 0% success rate
- Orbits GTO, ISS, LEO, MEO, PO have a success rate between 50% and 85%

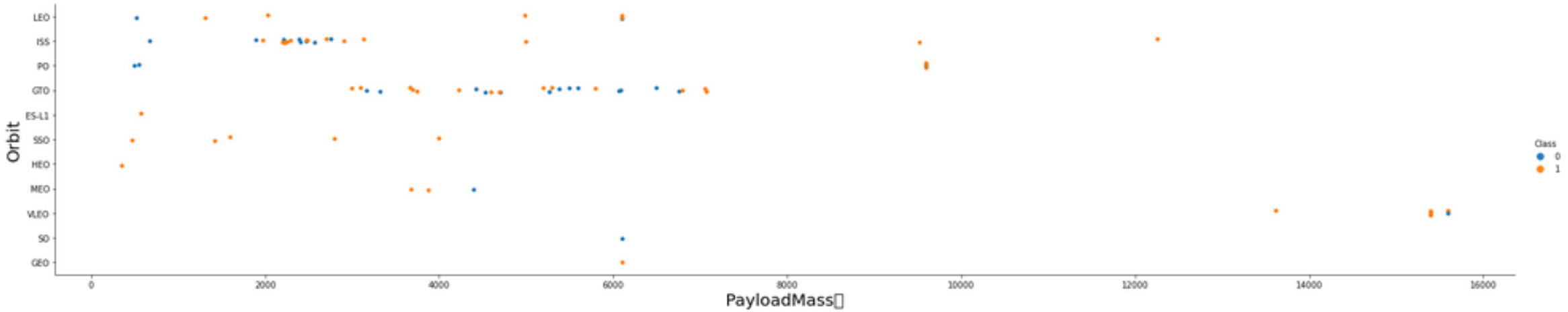
Flight Number vs. Orbit type

```
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



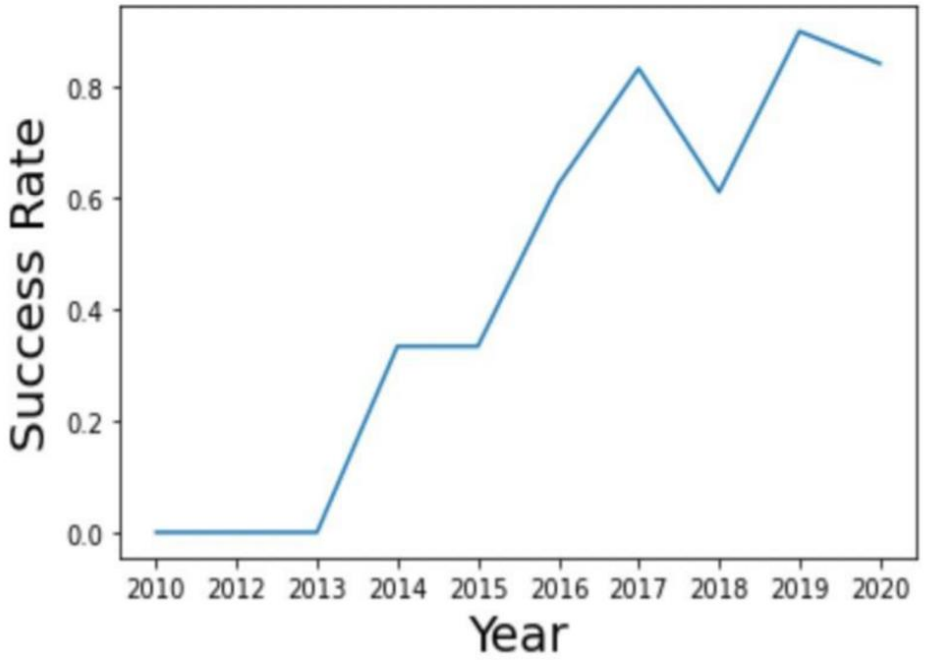
- In the LEO orbit the success seems related to the flight number.
- There seems to be no relationship between flight number in GTO orbit.

Payload vs. Orbit type



- Heavy payloads have a negative influence on GTO orbit
- Heavy payloads have a positive influence on ISS orbit

Launch success yearly trend



- The success rate was 0 until 2013
- Afterwards it shows an increasing trend (with a couple of exceptions, i.e. in 2018 and 2020)

EDA with SQL

All launch site names

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

- Displaying the names of the unique launch sites in the space mission

Launch site names begin with `CCA`

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

- Displaying 5 records where launch sites begin with the string 'CCA'

Total payload mass

total_payload_mass
45596

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

Average payload mass by F9 v1.1

average_payload_mass
2534

- Displaying the average payload mass carried by booster version F9 v1.1

First successful ground landing date

first_successful_landing
2015-12-22

- Displaying the date when the first successful landing outcome in ground pad was achieved

Successful drone ship landing with payload between 4000 and 6000

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

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

Boosters carried maximum payload

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

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

2015 launch records

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)

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

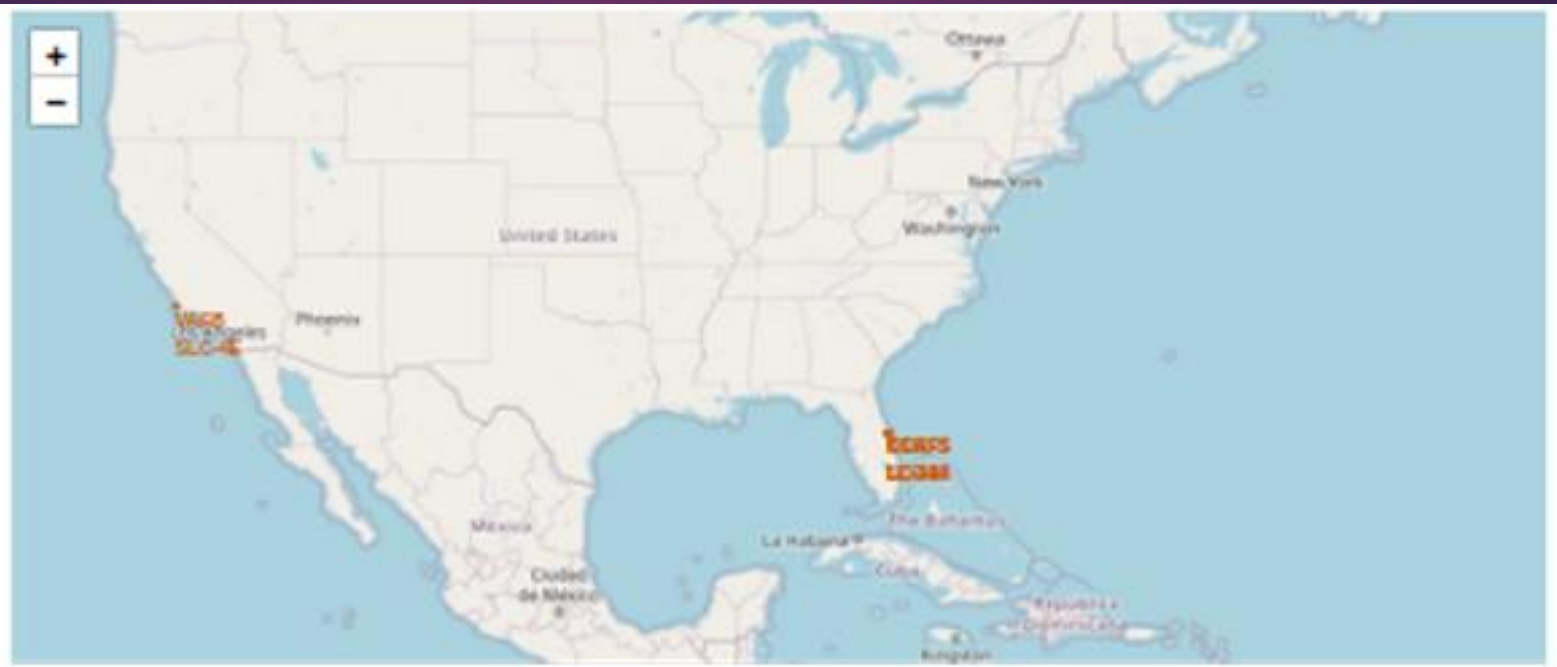
Rank success count between 2010-06-04 and 2017-03-20

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

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

Interactive map with Folium

Localising launch sites on a global map

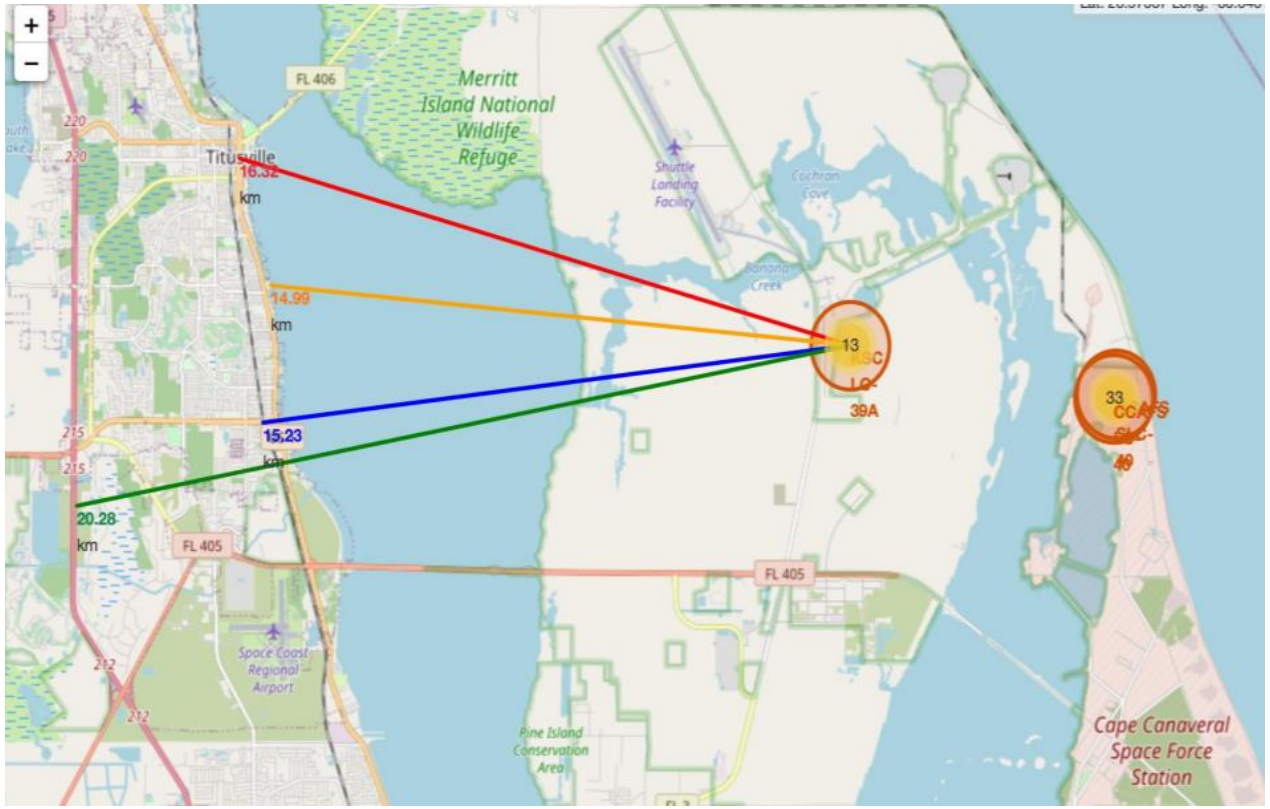


- Most launch sites are in proximity to the Equator line. The land is indeed moving faster at the equator than any other place on the surface of the Earth and this speed will help the space craft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, since launching rockets towards the ocean minimizes the risk of having any debris dropping or exploding near to populated areas

Color-labeled launch records

- The color-labeled markers allow us to easily identify which launch sites have relatively high success rates.
- Green Marker = Successful Launch while Red Marker = Failed Launch

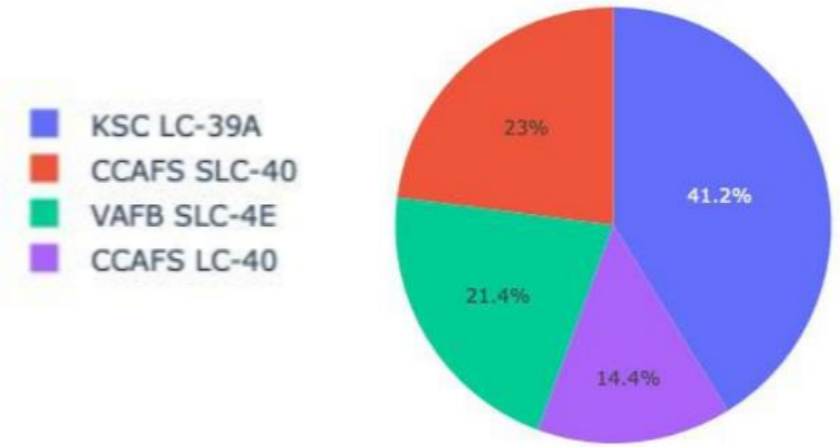
Distance between launch site KSCLC-39A and its proximities



- The launch site KSCLC-39A is:
- Close to railway (15.23 km)
- Close to highway (20.28 km)
- Close to coastline (14.99 km)
- Close to Titusville city (16.32 km).

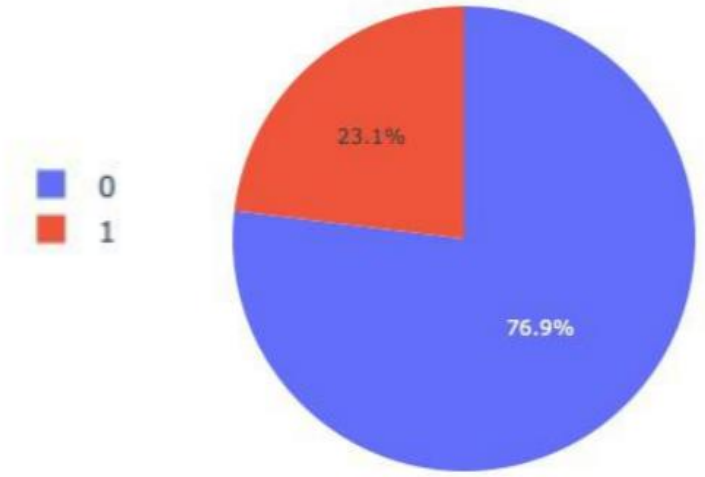
Build a Dashboard with Plotly Dash

Launch success count for all sites



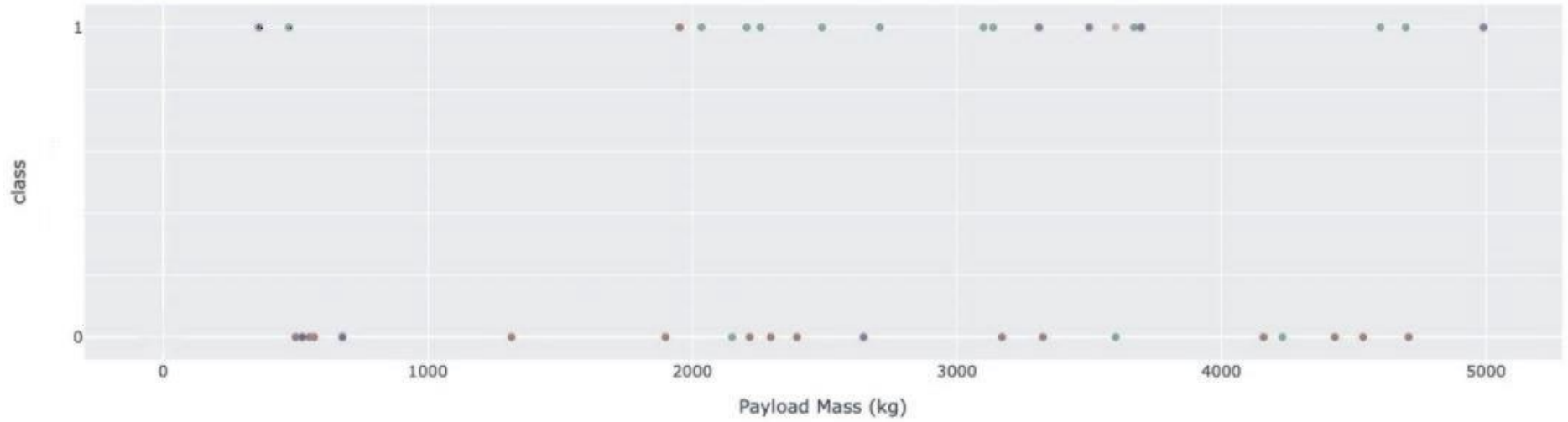
- Launch site KSC LC-39A is the one with the highest number of successful launches

Success rate for KSC LC-39A



- Launch site KSC has a success rate of 76.9%

Payload mass vs. Outcome



- Payloads between2000 and 5500 kg have the highest success rates
- Payloads above 5000 kg are not displayed here for readability, but they happen to have a very low success rate

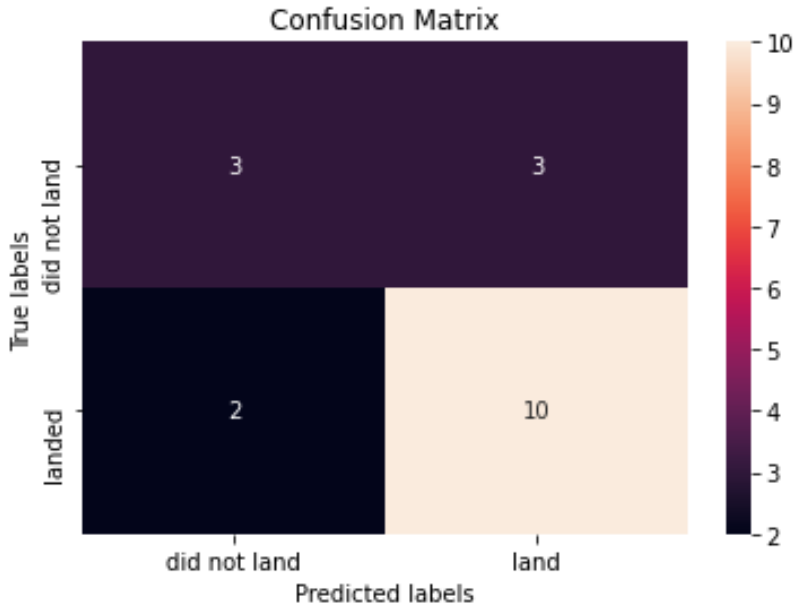
Predictive analysis (Classification)

Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.865672	0.805556	0.805556	0.788732
F1_Score	0.928000	0.892308	0.892308	0.881890
Accuracy	0.900000	0.844444	0.844444	0.833333

- The best model is the Logistic Regression. This model has not only higher scores, but also the highest accuracy.

Confusion Matrix



- We can observe that logistic regression can distinguish between the different classes.
- The major problem is false positives.

CONCLUSION

- The success rate of launches has an increasing trend over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- Launches with a payload between 2000 and 5500 show the best results.
- Logistic regression is the best algorithm for this dataset.

APPENDIX