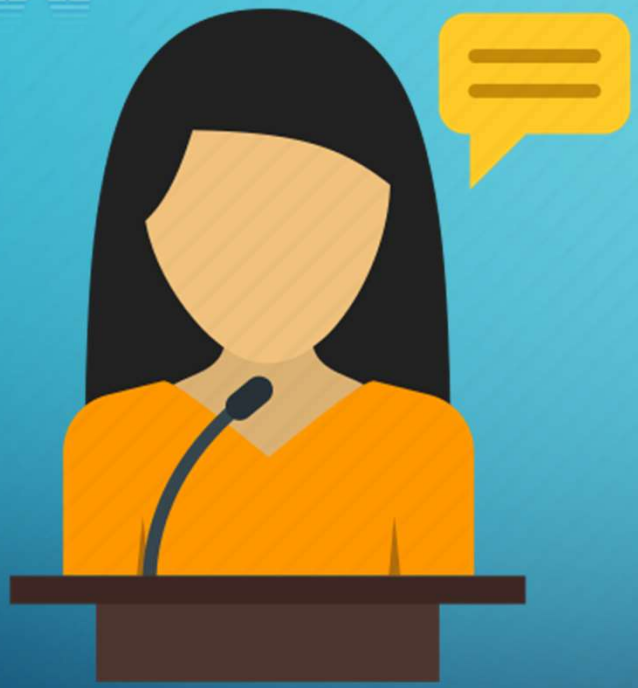

IBM
DATA SCIENCE CAPSTONE PROJECT
SPACEX
M.Sc. Marisely Urdaneta
https://github.com/marisely/IBM_Data_Science-Capstone-SpaceX.git

OUTLINE



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

EXECUTIVE SUMMARY

Summary of Methodologies

1. Data collection
2. Data wrangling
3. Exploratory data analysis (EDA) using SQL
4. Exploratory analysis using Pandas and Matplotlib
5. Interactive visual analytics and Dashboard using Folium and Plotly Dash
6. Predictive analysis using classification models

Summary of Results

1. Exploratory data analysis results
2. Interactive visual analytics results
3. Predictive analysis results

INTRODUCTION

Project Background and context

SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company capable of returning a spacecraft from low-Earth orbit, and in 2012 its Dragon spacecraft became the first commercial spacecraft to deliver cargo to and from the International Space Station. And in 2020, SpaceX became the first private company to take humans there as well. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if I can determine if the first stage will land, I can determine the cost of a launch. Spaces X's Falcon 9 launch like regular rockets. My objective is to determine the price of each launch. Also, I will predict if SpaceX will reuse the first stage by training a machine learning model.

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?

- 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

- SpaceX Rest API
- Web scraping related Wiki pages

Data wrangling

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

Exploratory data analysis (EDA) using SQL

Exploratory analysis using Pandas and Matplotlib

Interactive visual analytics and Dashboard using Folium and Plotly Dash

Predictive analysis using classification models

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

METHODOLOGY: DATA COLLECTION

During this stage, data is gathered from an API and web scraping. Firstly, using specifically the SpaceX REST API. This API will give me data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

Secondly, another popular data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. I will be using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.

METHODOLOGY: DATA COLLECTION SPACEX WITH API

- Request and parse the SpaceX launch data using the GET request
- Call the .json() method to review results
- Use the json_normalize function to convert this JSON to a dataframe
- Apply custom functions to request needed information about the launches from SpaceX API
- Construct dataset using the data obtained by combining columns into a dictionary
- Request and parse the SpaceX launch data using the GET request
- Create a Pandas data frame from the dictionary launch_dict
- Filter the dataframe to only include Falcon 9 launches
- Data Wrangling, dealing with missing values by replacing payloadmass with the payloadmass mean
- Save the data to csv file

METHODOLOGY: DATA COLLECTION WEB SCRAPING

- Request and parse the SpaceX launch data using the GET request
- Create a BeautifulSoup object from the HTML response
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- Construct dataset using the data obtained by combining columns into a dictionary
- Create a Pandas data frame from the dictionary launch_dict
- Filter the dataframe to only include Falcon 9 launches
- Save the data to csv file

METHODOLOGY: DATA WRANGLING

During this stage, I will perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

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.

I will 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**

METHODOLOGY: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.

Therefore, if I can determine if the first stage will land, I can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

This dataset includes a record for each payload carried during a SpaceX mission into outer space.

I will write and execute SQL queries to solve these tasks

METHODOLOGY: EXPLORATORY ANALYSIS USING PANDAS AND MATPLOTLIB

I will predict if1 the Falcon 9 first stage will land successfully. SpaceX 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 due to the fact that SpaceX can reuse the first stage.

I will perform Exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib.

- **Visualize the relationship between Flight Number and Launch Site**
- **Visualize the relationship between Payload and Launch Site**
- **Visualize the relationship between success rate of each orbit type**
- **Visualize the relationship between FlightNumber and Orbit type**
- **Visualize the relationship between Payload and Orbit type**
- **Visualize the launch success yearly trend**

Features Engineering

- **Create dummy variables to categorical columns**
- **Cast all numeric columns to float64**
- **Save the data to csv file**

METHODOLOGY: INTERACTIVE VISUAL ANALYTICS AND DASHBOARD USING FOLIUM AND PLOTLY DASH

The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

METHODOLOGY: INTERACTIVE VISUAL ANALYTICS AND DASHBOARD USING FOLIUM AND 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

METHODOLOGY: PREDICTIVE ANALYSIS USING CLASSIFICATION MODELS

I will create a machine learning pipeline to predict if the first stage will land given the data from the preceding steps.

-Perform exploratory Data Analysis and determine Training Labels:

- create a column for the class
- Standardize the data
- Split into training data and test data

-Find best Hyperparameter for SVM, Classification Trees and Logistic Regression:

- Find the method performs best using test data

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

Display the names of the unique launch sites in the space mission

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

7]:

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

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

Display 5 records where launch sites begin with the string 'CCA'

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing__Outcome
4/6/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
8/12/2010	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)
22-05-2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
8/10/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
1/3/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

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

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

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

9]:

TOTAL_PAYLOAD
111268

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

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

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

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

1]:

FIRST_SUCCESS_GP
1/5/2017

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

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 DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Succ
```

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

2]:

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

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

List the total number of successful and failure mission outcomes ¶

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

5]:

Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) USING SQL

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

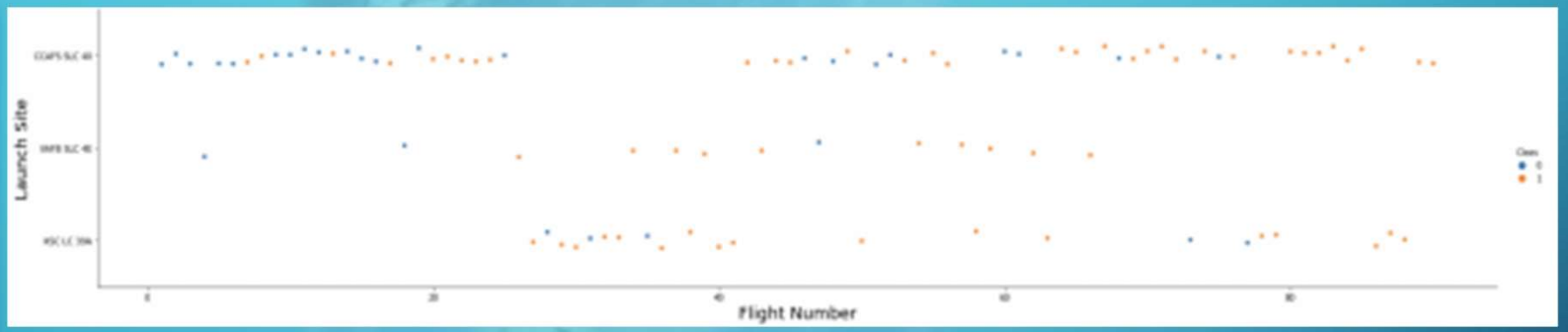
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

7]:

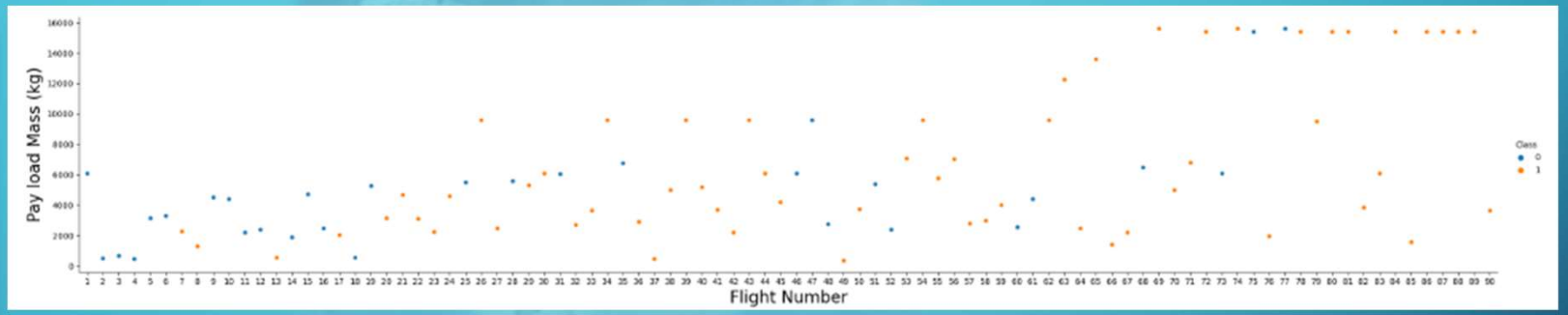
Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) THROUGH VISUALIZATION



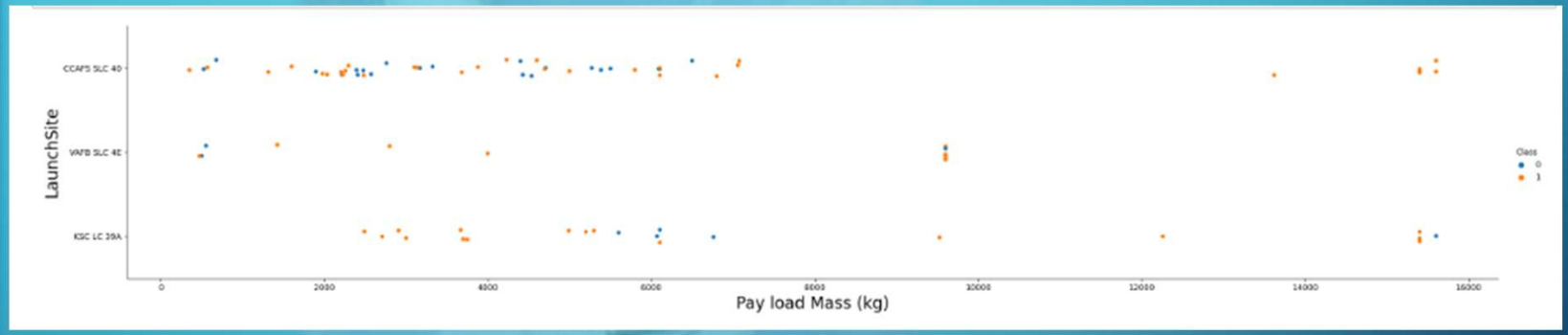
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

RESULTS: EXPLORATORY DATA ANALYSIS (EDA) THROUGH VISUALIZATION



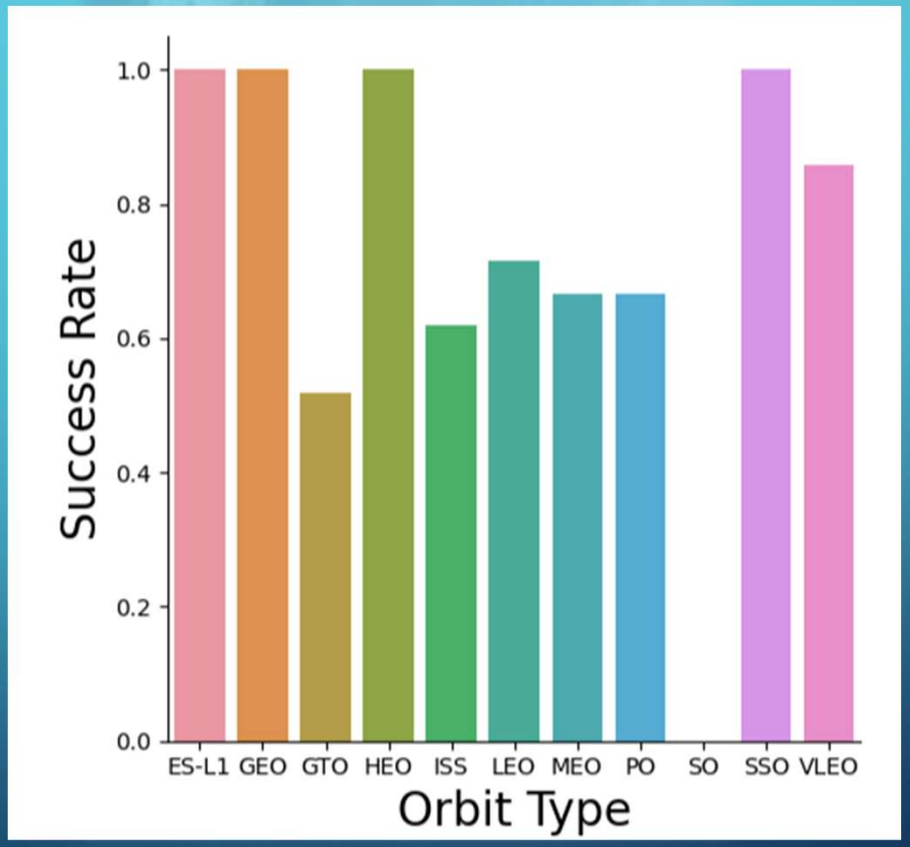
- We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

RESULTS: EXPLORATORY DATA ANALYSIS (EDA)

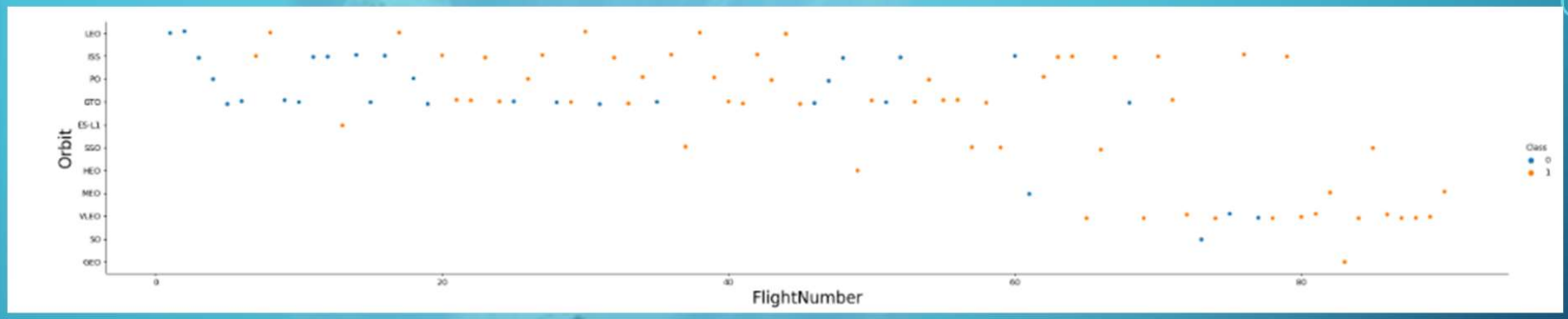


We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

RESULTS: EXPLORATORY DATA ANALYSIS (EDA)

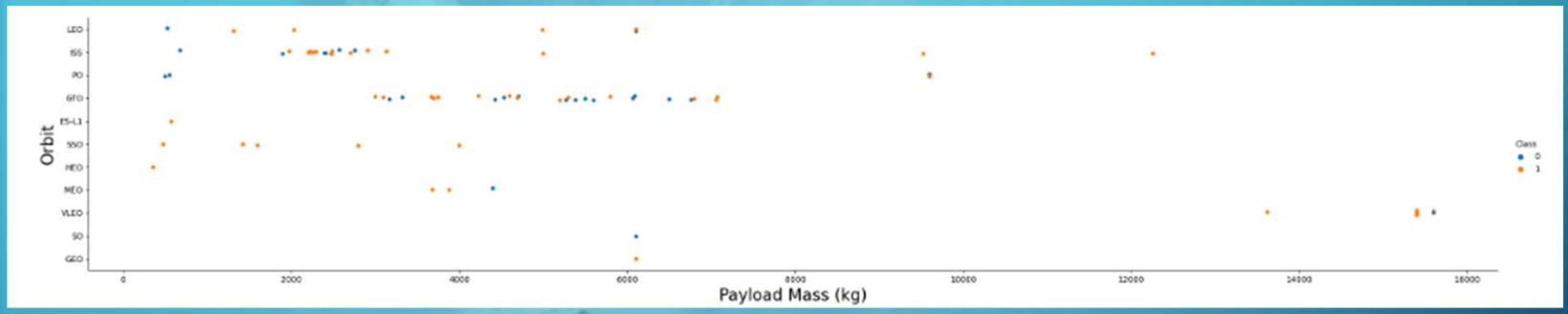


RESULTS: EXPLORATORY DATA ANALYSIS (EDA)



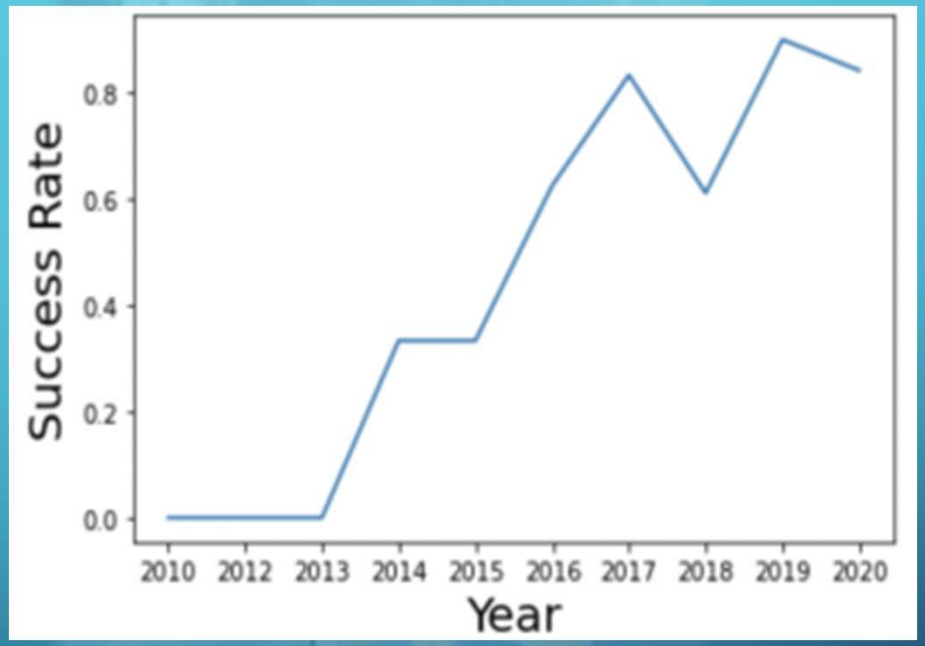
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

RESULTS: EXPLORATORY DATA ANALYSIS (EDA)



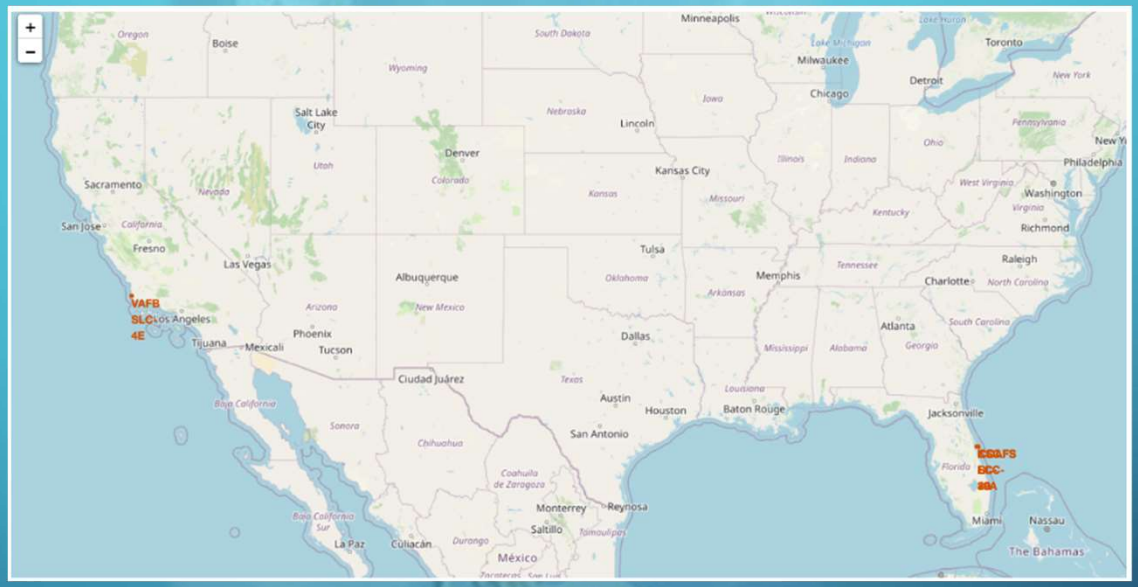
Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits

RESULTS: EXPLORATORY DATA ANALYSIS (EDA)



The success rate since 2013 kept increasing untill 2020

RESULTS: INTERACTIVE VISUAL ANALYTICS AND DASHBOARD USING FOLIUM



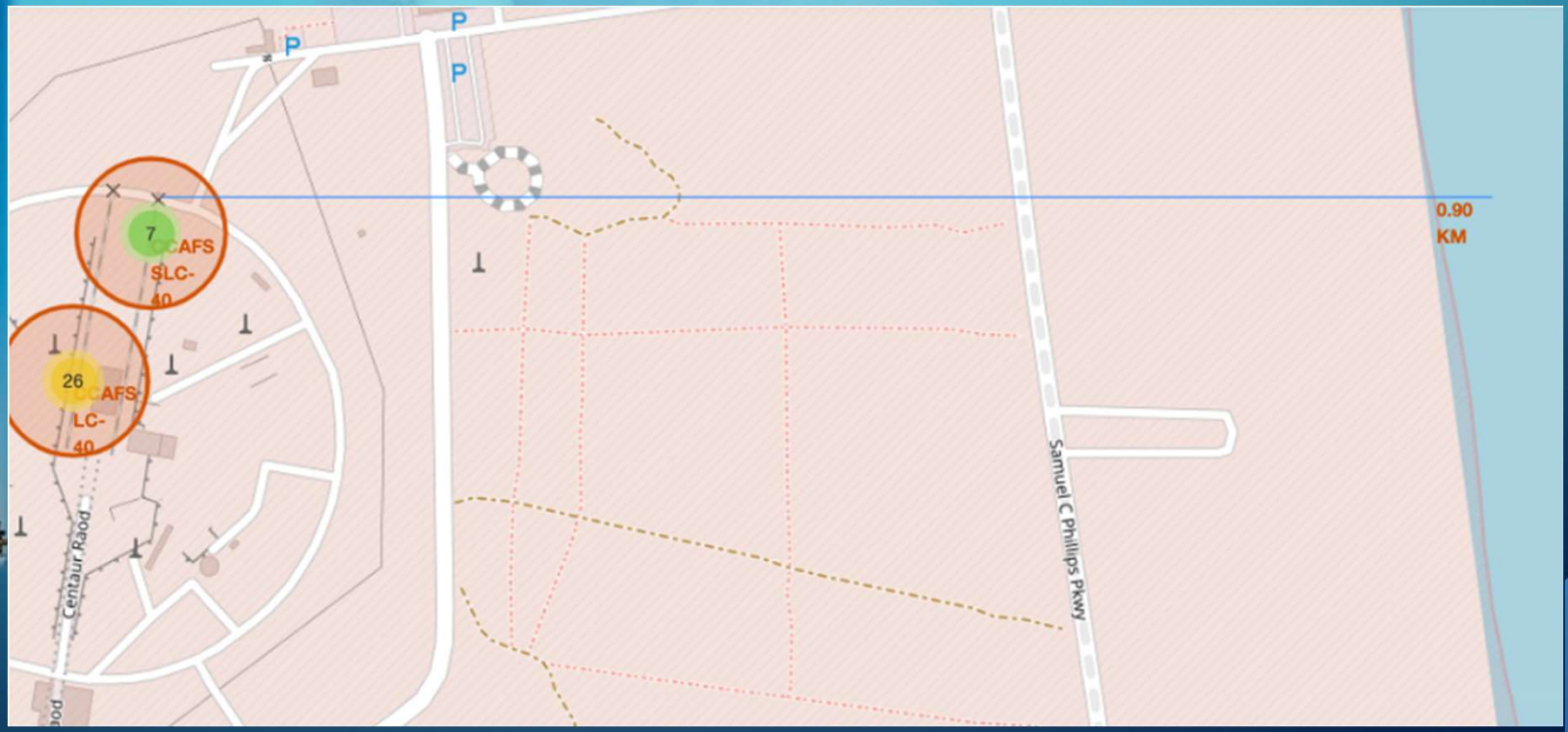
The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

RESULTS: INTERACTIVE VISUAL ANALYTICS AND DASHBOARD USING FOLIUM

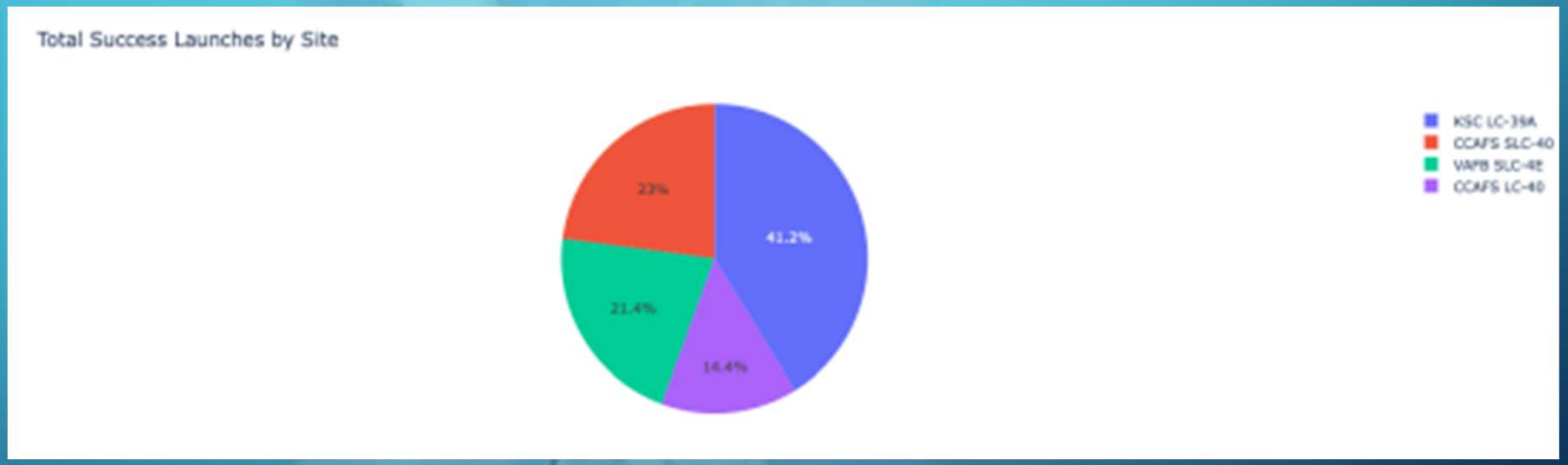
From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

RESULTS: INTERACTIVE VISUAL ANALYTICS AND DASHBOARD USING FOLIUM

Your updated map with distance line should look like the following screenshot

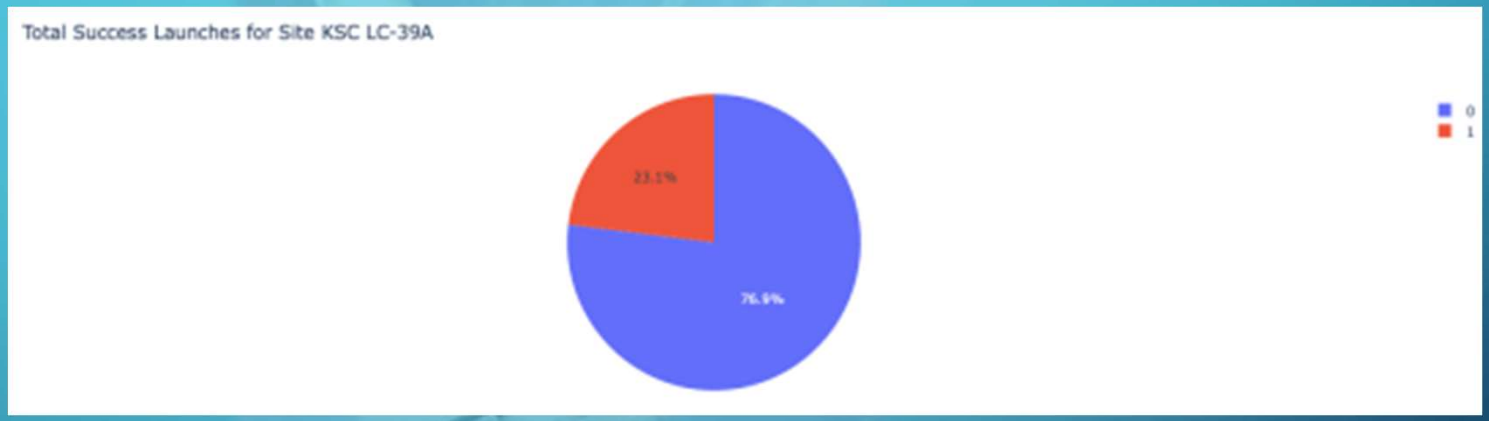


RESULTS: INTERACTIVE VISUAL ANALYTICS AND DASHBOARD USING PLOTLY DASH



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

RESULTS: INTERACTIVE VISUAL ANALYTICS AND DASHBOARD USING PLOTLY DASH



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

RESULTS: PREDICTIVE ANALYSIS USING CLASSIFICATION MODELS

Models Accuracy – Test Set

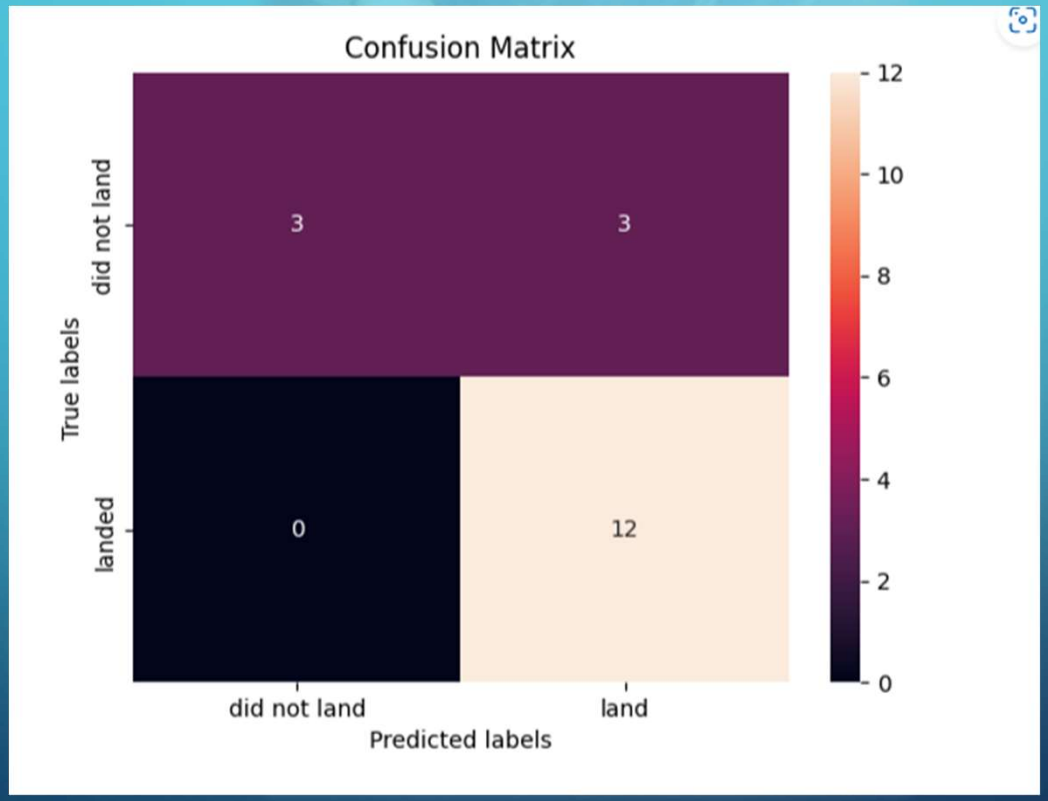
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Models Accuracy – Intire Data Set

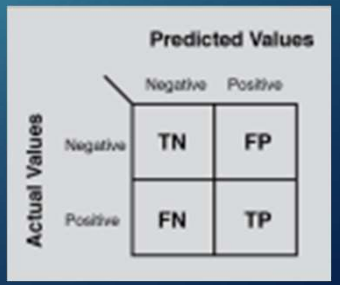
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

RESULTS: PREDICTIVE ANALYSIS USING CLASSIFICATION MODELS

Confusion Matrix



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

- 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 rate.

APPENDIX

Thanks to:

Coursera

Instructors

IBM

Upwardly Global