IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

TOLIVA SOLOMON
IWUTUNG
February 7, 2022

Outline

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization and SQL
  - Interactive visual analytics with Folium and Plotly Dash
  - Predictive analysis using Classification
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- In this project we determine the cost of a rocket launch for a new rocket company, Space Y, founded by Allon Musk.

- We use information from Space X, which is perhaps the most successful among all the companies in the commercial space race. The success of Space X is largely due to the relatively low cost of each launch as a result of the reusable first stage of the rocket.

- We therefore examine Space X data, and use it to create a machine learning model to predict if Space X will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Creating a column that classifies landing outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- SpaceX API
- Web scraping from Wikipedia

Data Collection – SpaceX API

- We work with data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome, that is gathered from the SpaceX REST API.
- We use the URL, https://api.spacexdata.com/v4/launches/past, to target a specific endpoint of the API to get past launch data. We perform a get request using the requests library to obtain the launch data.
- Reference notebook

...

Request rocket launch data from SpaceX API → Decode the response content as a JSON → Apply custom functions to clean data → Convert the data to a dataframe → Filter the dataframe to contain only falcon 9 launches → Deal with null values and export the dataframe as csv

Data Collection - Scraping

- We also use the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.
- We then parse the data from those tables and convert them into a Pandas data frame for further visualization and analysis.
- [Reference notebook]

Request the Falcon9 Launch Wiki page from its URL → Extract all column/variable names from the HTML table header → Create a data frame by parsing the launch HTML tables → Export the dataframe as csv

Data Wrangling

- We perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- The dataset contains several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident.
- We convert those outcomes into Training Labels where 1 means the booster successfully landed and 0 means it was unsuccessful.
- [Reference notebook]

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

EDA with Data Visualization

- With the outcome of the launch as category, we examine the following relationships using categorical scatter plots:
  - How the Flight Number and Payload variables would affect the launch outcome.
  - The relationship between Flight Number and Launch Site
  - The relationship between Payload mass and Launch Site.
  - The relationship between Flight Number and Orbit type.
  - The relationship between Payload mass and Orbit type.
- Using a bar chart, we examine the relationship between success rate of each orbit type.
- Using a line chart, we visualize the launch success yearly trend.
- Reference notebook

EDA with SQL

- We perform the following queries using SQL:
    - Display the names of the unique launch sites in the space mission
    - Display 5 records where launch sites begin with the string 'CCA'
    - Display the total payload mass carried by boosters launched by NASA (CRS)
    - Display average payload mass carried by booster version F9 v1.1
    - List the date when the first successful landing outcome in ground pad was achieved.
    - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - List the total number of successful and failure mission outcomes
    - List the names of the booster versions which have carried the maximum payload mass using a subquery
    - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
    - 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

- Reference notebook

Build an Interactive Map with Folium

- We create a folium map and and mark all the launch sites , color coded according to whether the launch was a success or failure
- We add a Mouse Position feature on the map to get coordinates for a mouse over a point on the map, and a PolyLine between a selected launch site to selected proximities
- We also calculate the distances between a selected launch site to its proximities, and add the distance marker to the map
- [Reference notebook]

Build a Dashboard with Plotly Dash

- We a dropdown list to enable Launch Site selection for the data to be visualized.
- We add a pie chart to show the total successful launches count for all sites if all sites were chosen from the drop down, or the Success vs. Failed counts for the site if a specific launch site was selected.
- We add a slider to select payload range to be displayed on a scatter chart which shows the correlation between payload and launch success
- Reference Dash Lab

Predictive Analysis (Classification)

- We create NumPy arrays from the data to be used in modeling, standardize the feature set, and split the data into training and test sets
- Using Grid search, we find best Hyperparameter for for building SVM, Classification Trees and Logistic Regression models, and fit each of these models using the best hyperparameters, and then evaluate the models to determine the best performing model
- Reference notebook

Standardize the data and split into training and test sets → Build the models using the best parameter → Evaluate each model → Plot the confusion matrix → Find the best performing model

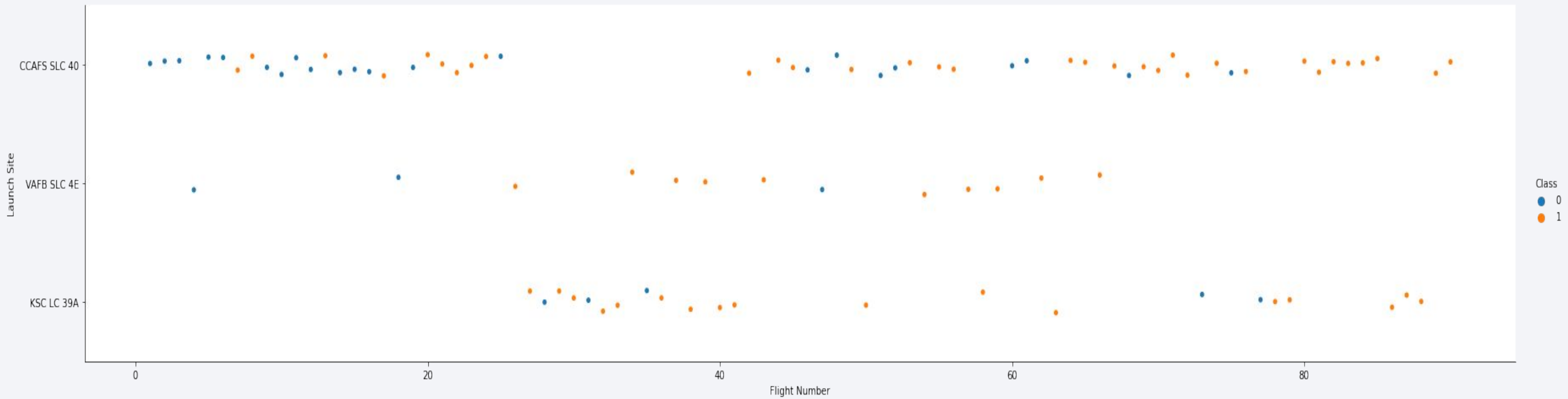
Results

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

Section 2

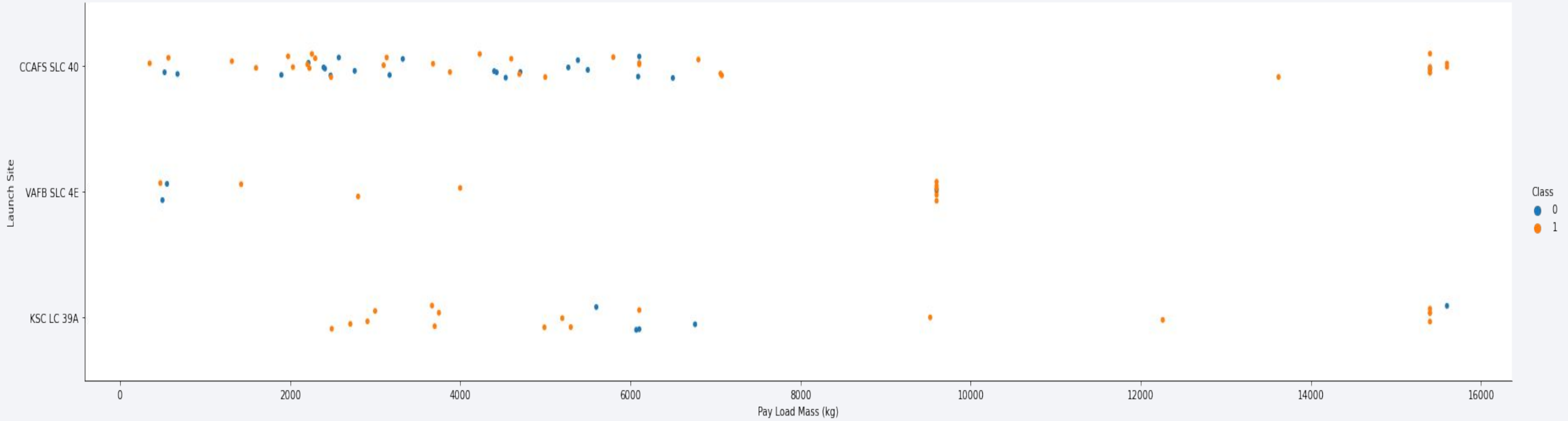
Insights drawn from EDA

Flight Number vs. Launch Site



- We see that as the number of flights for different launch sites increases, the first stage appears more likely to land successfully.

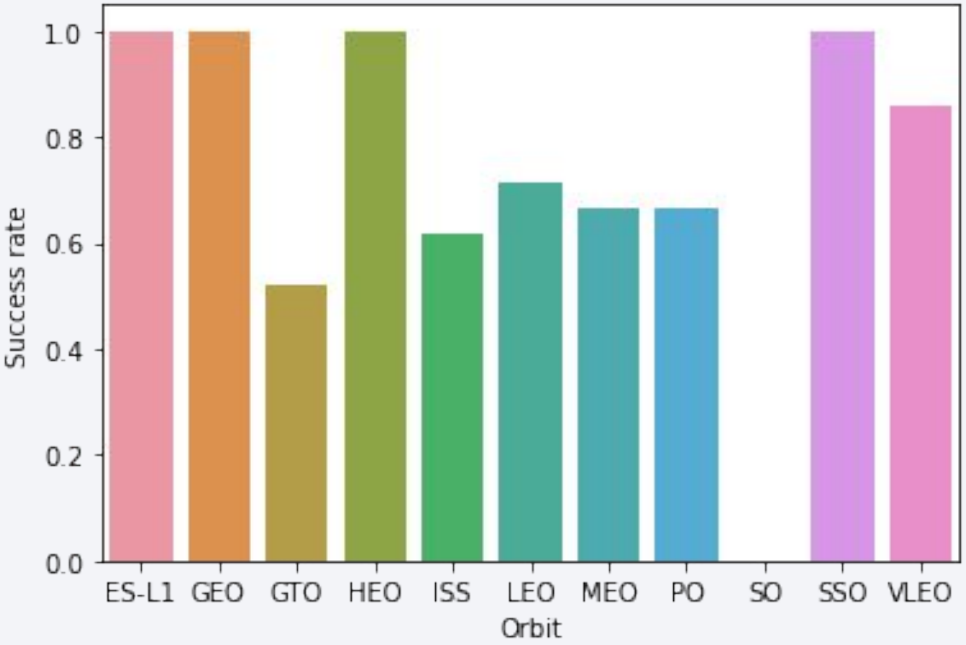
Payload vs. Launch Site



- While for the VAFB-SLC launch site there are no rockets launched for heavy payload mass, the success rate increases for launch site CCAFS SLC 40 as the payload increases. There is no clear indication for launch site KSC LC 39A.

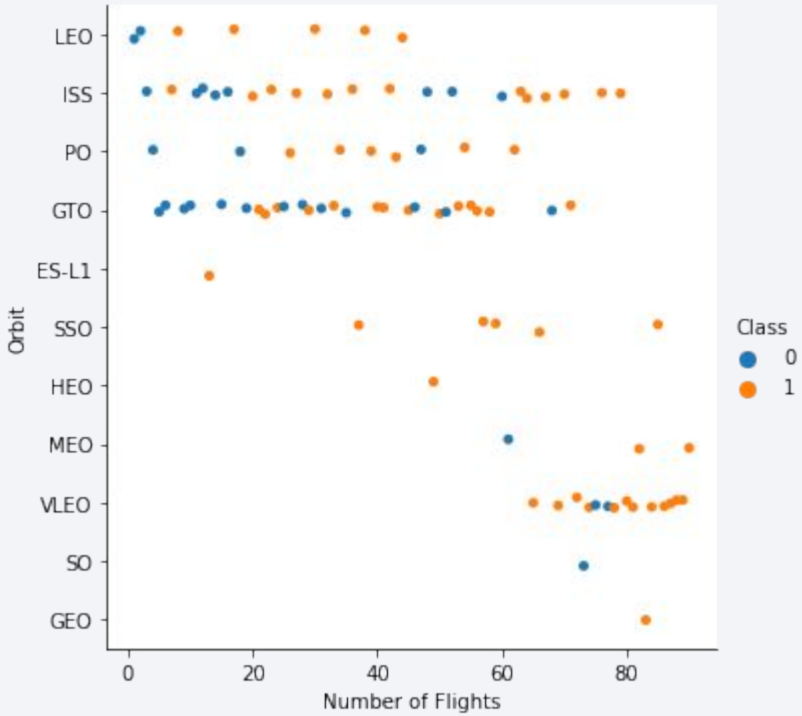
Success Rate vs. Orbit Type

- We can see that ES-L1, GEO, HEO and SSO have the highest success rate while SO on the other hand has the lowest success rate.



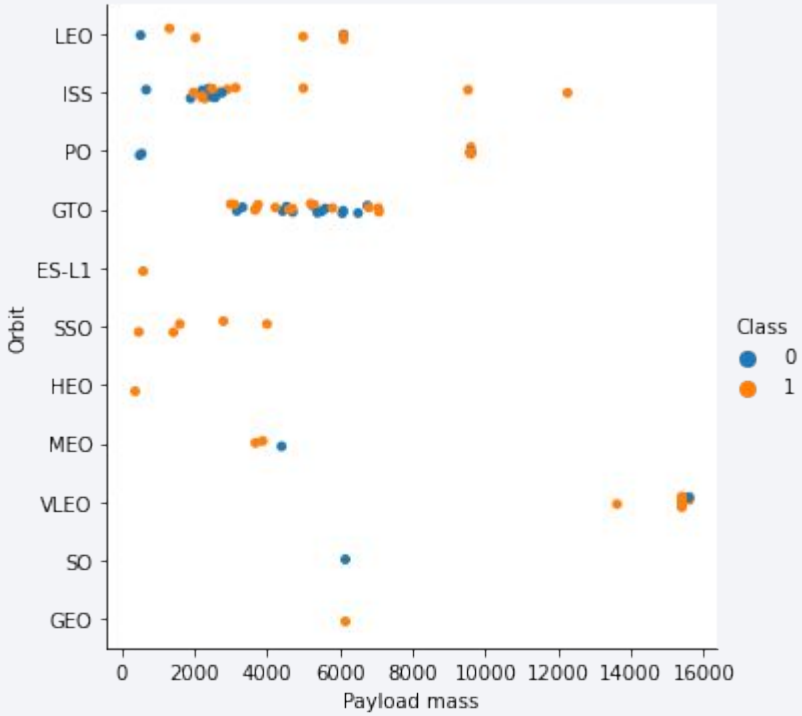
Flight Number vs. Orbit Type

- We see that 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.



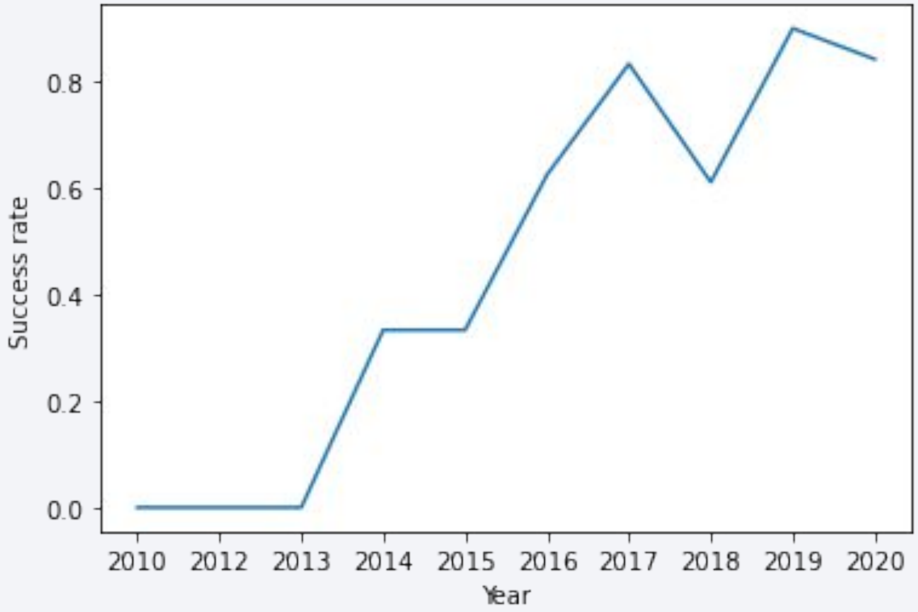
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

- **SELECT DISTINCT Launch_Site AS 'Unique launch sites in the space mission' FROM SPACEXTBL;**
- This query returns the unique launch site names

Unique launch sites in the space mission
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04/06/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08/12/2010	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08/10/2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01/03/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- **SELECT * FROM SPACEXTBL WHERE Launch_Site like 'CCA%’ LIMIT 5;**
- This query returns the first 5 launch sites with the name starting with CCA.

Total Payload Mass

- **SELECT SUM(PAYLOAD_MASS__KG_) AS 'Total payload mass carried by boosters launched by NASA (CRS)' FROM SPACEXTBL GROUP BY Customer HAVING Customer = 'NASA (CRS)';**
- This query returns the total payload mass carried by boosters launched by NASA (CRS)

Total payload mass carried by boosters launched by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- **SELECT AVG(PAYLOAD_MASS__KG_) AS 'Average payload mass carried by booster version F9 v1.1’ FROM SPACEXTBL GROU BY Booster_Version HAVING Booster_Version = 'F9 v1.1’;**
- This query returns the average payload mass carried by booster version F9 v1.1

Average payload mass carried by booster version F9 v1.1
2928.4

First Successful Ground Landing Date

- **SELECT MIN(Date) AS 'Date when the first successful landing outcome in ground pad was achieved’ FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)’;**
- This query returns the dates of the first successful landing outcome on ground pad

Date when the first successful landing outcome in ground pad was achieved
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- **SELECT Booster_Version AS 'Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000' FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;**
- This query returns the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- **SELECT (SELECT Count(Mission_Outcome) FROM SPACEXTBL WHERE Mission_Outcome LIKE '%Success%') AS Success, (SELECT Count(Mission_Outcome) FROM SPACEXTBL WHERE Mission_Outcome LIKE '%Failure%') AS Failure;**
- This query returns the total number of successful and failure mission outcomes

Success	Failure
100	1

Boosters Carried Maximum Payload

- **SELECT Booster_Version AS 'Booster_versions which have carried the maximum payload mass’ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);**
- This query returns the names of the booster which have carried the maximum payload mass

Booster_versions which have carried the maximum payload mass
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

- **SELECT "Landing _Outcome" AS 'Landing outcome', Booster_Version AS 'Booster version', Launch_Site AS 'Launch site' FROM SPACEXTBL WHERE "Landing _Outcome"='Failure (drone ship)' AND Date LIKE '%2015';**
- This query returns the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Landing outcome	Booster version	Launch site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

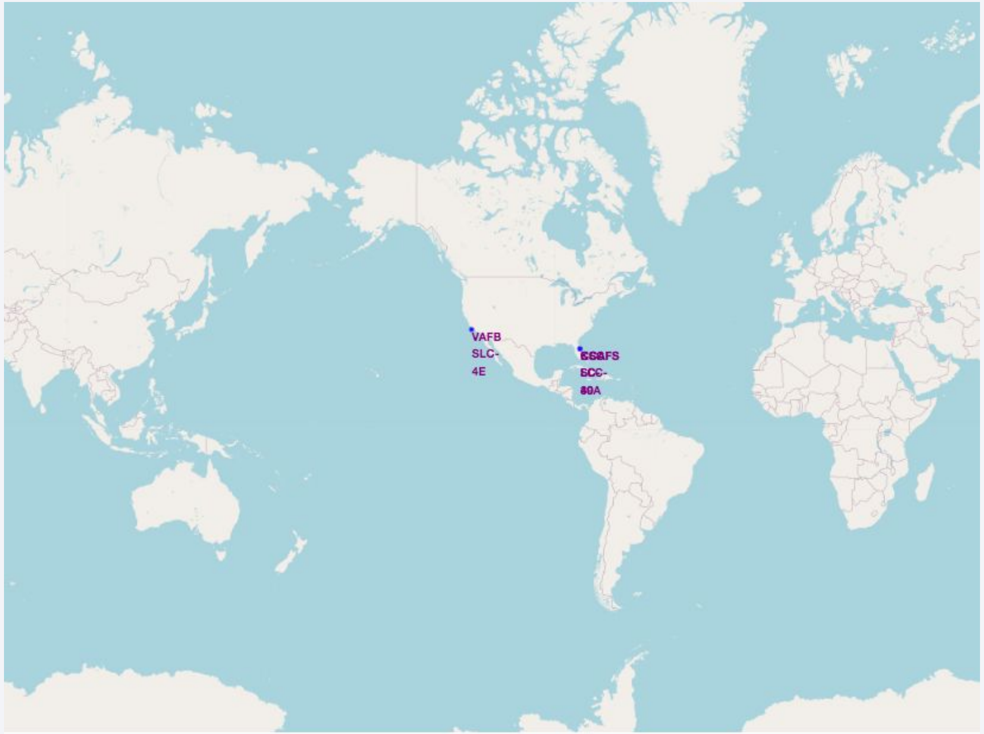
- **SELECT "Landing _Outcome" AS 'Landing outcome', COUNT("Landing _Outcome") AS Count FROM SPACEXTBL GROUP BY "Landing _Outcome" ORDER BY 2 DESC;**
- This query ranks the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing outcome	Count
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

Section 4

Launch Sites Proximities Analysis

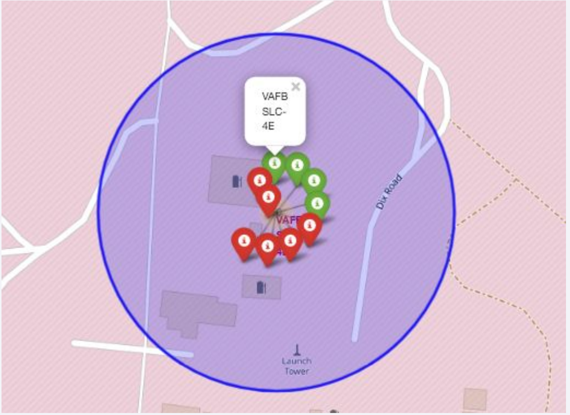
All Launch sites



- We see that the launch sites are all located along the coast of the United states of America

Color Labeled Launch outcomes for all sites

Florida Launch sites



California Launch sites

- Successful launches are represented by a green marker, while failed launches are represented by red markers.

Launch site proximities

- We see that while the selected Launch site is relatively close to the coastline, highway, and railway line, there is a significant amount of distance between the Launch site and the nearest city.

Section 5

Build a Dashboard with Plotly Dash

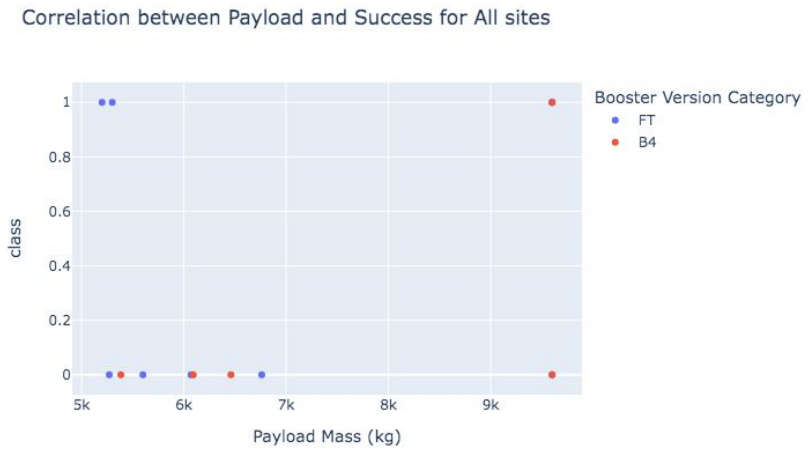
Launch success count for all sites

- We see that site KSC LC-39A has the largest number of successful launches

Launch site with the highest Launch success ratio

- We see that KSC LC-39A had a success ratio of 76.9%, which was the highest among all Launch sites

Payload vs. Launch Outcome scatter plot for all sites



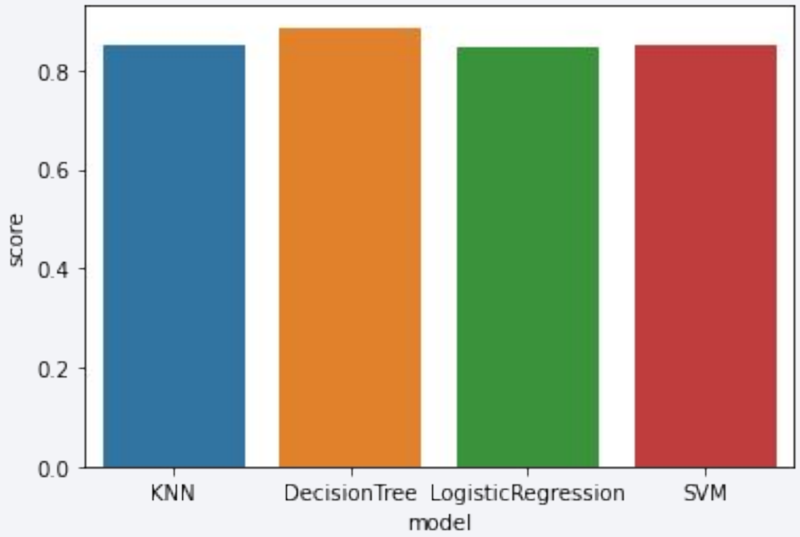
- We see that the success rate is lower for higher payloads than for lower payloads.

Section 6

Predictive Analysis (Classification)

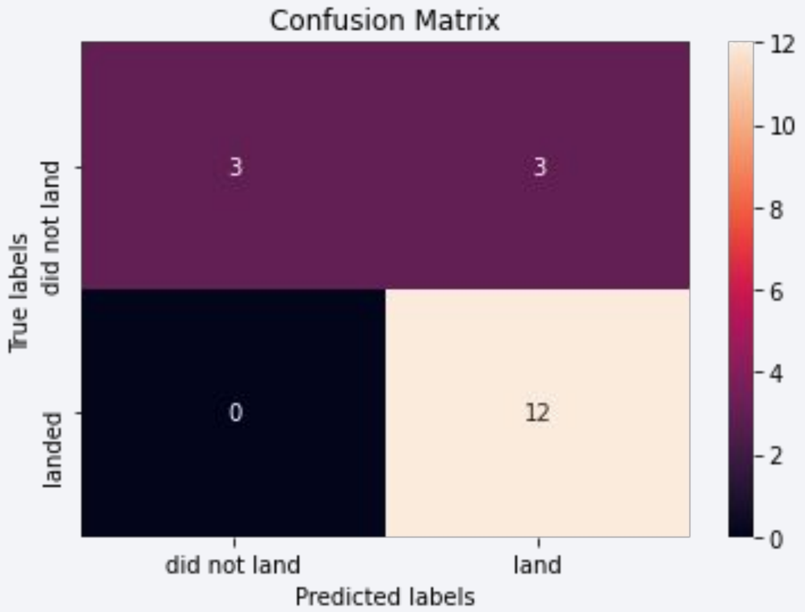
Classification Accuracy

- From the bar graph, we see that the Decision Tree model has the highest classification accuracy.



Confusion Matrix

- We see that apart from false positives, the Decision tree model can differentiate the classes of the launch outcomes



Conclusions

- As the number of flights for different launch sites increases, the first stage appears more likely to land successfully.
- ES-L1, GEO, HEO and SSO orbit types have the highest success rate while SO on the other hand has the lowest success rate.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS orbit types.
- While the selected Launch site is relatively close to the coastline, highway, and railway line, there is a significant amount of distance between the Launch site and the nearest city.
- The Decision Tree model has the highest classification accuracy.

Appendix

- Plotting the bar chart of model accuracy for all built classification models

Data used for creating the chart:

model	score
KNN	0.848214
DecisionTree	0.875
LogisticRegression	0.846429
SVM	0.848214

Code:

```
model_Score = {'model':('KNN','DecisionTree','LogisticRegression','SVM'), 'score':(knn_cv.best_score_,tree_cv.best_score_,logreg_cv.best_score_,svm_cv.best_score_)}

model_df = pd.DataFrame(model_Score)
model_df
```

```
sns.barplot(data=model_df,y='score',x='model')
```

Thank you!