
IBM Developer
SKILLS NETWORK
Winning Space Race
with Data Science
Michel Douglas Piamou
12/08/2023
M. D. Piamou - IBM Data Science Capstone Project - August 2023

Executive Summary

- SpaceX is able to propose rocket launch at inexpensive price because they can reuse the 1st stage. Thus, to predict the price of a rocket launch, we use public data on SpaceX past launches to train a machine learning model which can predict if the 1st stage will land successfully.
- We proceed with the following methodology :
  - We first collect data on spaceX past launches using SpaceX API and Web Scrapping Beautiful python package, then we clean the data obtained and perform data wrangling to produce dataset ready for further analysis.
  - We continue with Exploratory Data Analysis using Pandas, SQL, visualization (with matplotlib and seaborn), interactive visualization (with Folium) to understand the data and chose the best features.
  - We build a dash board using Plotly Dash allowing users to visualize relationships between the booster versions, the payload mass, the launch sites, and the success rate of 1st stage landing on past launches data.
  - We perform prediction analysis with 3 classification models : Support Vector Machine(SVM), Decision Tree and K Nearest Neighbors(KNN).
- Following the EDA, we observe that the success of the 1st stage landing is related to the Launch Site, The Payload mass, the Orbit, the Booster, Flight Number, the landing pad, …
- The Decision Tree Model has the best performance in predicting the success of 1st stage landing with an accuracy score of 94,4% compare to SVM and KNN.

Table of Content

Introduction

Context

- SpaceX space travel company success is mainly base on its ability to provide inexpensive rocket launches at 65 million$ each launch versus cost upward to 165 million$ for its competitors. SpaceX rocket launch low cost is possible because they can recover the first stage.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Objectives

- The project objectives are :
  - to determine the price of a rocket launch and
  - to provide a dash board allowing user to interactively explore past launches data.
- We'll will train a machine learning model and use public information to predict if SpaceX will reuse the first stage, thus allowing to propose a competitive price of rocket launch.

Section 1

Methodology

Methodology

- Data collection methodology
- Perform data wrangling
- 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

Two methods used to collect the data of SpaceX Falcon 9 past launches:

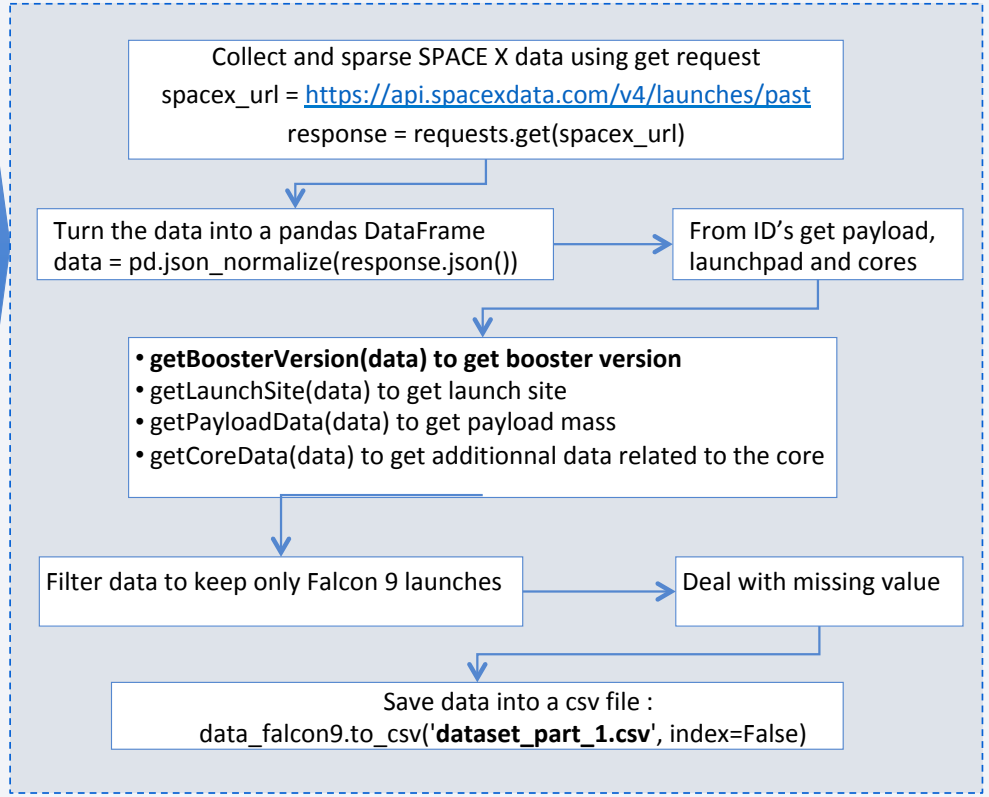
DATA COLLECTION WITH SPACEX REST API

WEBSCRAPPING WITH BEAUTIFUL SOUP PYTHON PACKAGE

Data Collection with SpaceX API

- **We use the SpaceX REST API** to get data on past launches, including rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- We derive the relevant information on rocket, payload, launchpad and cores given their IDs.
- We then use the API and defined functions to get additional information from rocket, payload, launchpad and cores data : booster version, payload mass, core related information.

***GitHub URL SpaceX API calls notebook** : https://github.com/PMDOUGLAS23/Applied-DS-C/blob/9e25d4d106880cb5bb3520d1b4fe6b14c34f289d/01-%20Collecting%20the%20data/jupyter-labs-spacex-data-collection-api_2.ipynb*

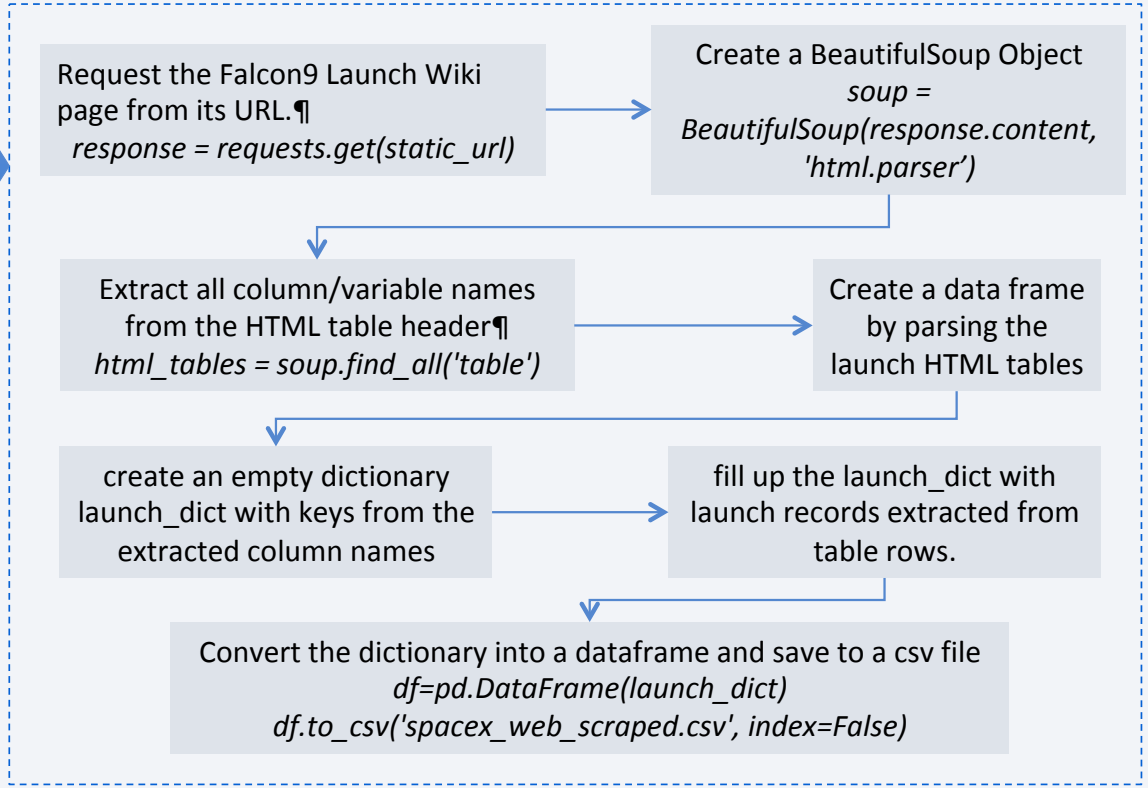


Data Collection – Web Scraping

Web Scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia using BeautifulSoup python package

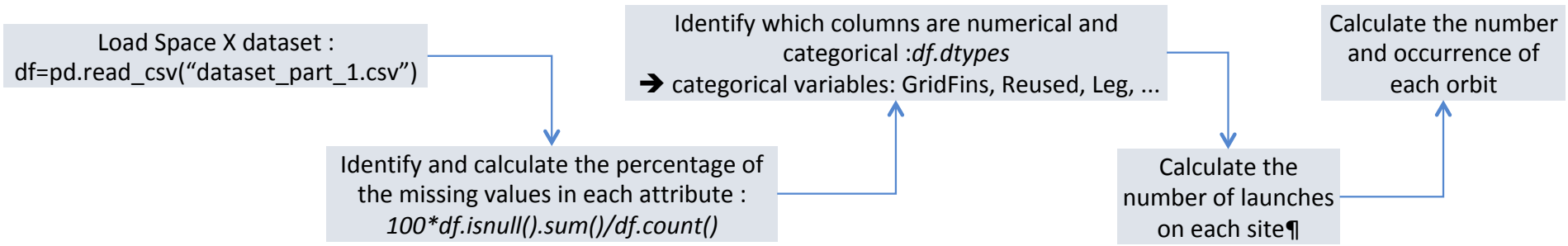
GitHub URL of web scraping notebook :

https://github.com/PMDOUGLAS23/Applied-DS-C/blob/9e25d4d106880cb5bb3520d1b4fe6b14c34f289d/01-%20Collecting%20the%20data/jupyter-labs-webscraping_3.ipynb



Data Wrangling

1 Analyze the data collected



2 Determine the training labels

Create a landing outcome label 'Class' from Outcome column. This variable will be the classification variable that represents the outcome for each launch.

- **Class = 0 if the first stage did not land successfully**
- **Class = 1 if the first stage landed Successfully**

Data Wrangling

- Save the dataframe df into a csv file : *pd.to_csv ("dataset_part_2.csv", index=False)*
- Overview of df with the training label :

```
1 df.head(5)
```

erVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

GitHub URL of data wrangling notebook *: https://github.com/PMDOUGLAS23/Applied-DS-C/blob/9e25d4d106880cb5bb3520d1b4fe6b14c34f289d/02%20-%20Data%20wrangling/labs-jupyter-spacex-Data%20wrangling_1.ipynb*

EDA with Data Visualization

Objectives :

Data visualization

to visualize the relationships between the variables and the outcome of 1st stage landing using matplotlib and seaborn

Feature Engineering

to select the best features that will be used in the prediction of the 1st stage landing success

GitHub URL of EDA visualization notebook *: https://github.com/PMDOUGLAS23/Applied-DS-C/blob/1ea872123cce9417db0b45a49ba3bbc21fa499d9/04%20-%20EDA%20Using%20Pandas%20and%20Matplotlib/jupyter-labs-eda-dataviz.ipynb*

EDA with Data Visualization

Data Visualization :

1. Plot out the FlightNumber vs. PayloadMass and overlay the outcome of the launch
2. Visualize the relationship between Flight Number and Launch Site : 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
3. Visualize the relationship between Payload and Launch Site : scatter point chart with x axis to be Payload Mass (kg) and y axis to be the launch site, and hue to be the class value
4. Visualize the relationship between success rate of each orbit type : `bar chart` for the success rate of each orbit

EDA with Data Visualization

Data Visualization :

5. Visualize the relationship between FlightNumber and Orbit type : scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
6. Visualize the relationship between Payload and Orbit type : scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
7. Visualize the launch success yearly trend: line chart with x axis to be the extracted year and y axis to be the success rate

EDA with Data Visualization

Features engineering :

1. Select the features that will be used in the 1st stage landing success prediction :

   Features : 'FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial'

2. Create dummy variables to categorical columns

   features_one_hot = pd.get_dummies(features)

3. Cast all numeric columns to `float64` :

   features_one_hot = features_one_hot.astype('float64', copy = False)

4. Export to a CSV file :

   features_one_hot.to_csv('dataset_part_3.csv', index=False)

EDA with SQL

Objective

Load the dataset into the corresponding table in a Db2 database(SPACEXTBL) and analyze the Spacex data set using SQL queries

Queries performed (1/2) :

- 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

EDA with SQL

- **Queries performed (2/2) :**
  - 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.
  - List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months 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.

***GitHub URL of EDA with SQL notebook** : https://github.com/PMDOUGLAS23/Applied-DS-C/blob/0554044cd727ed7be6ef39b4affa4e90741f7fc2/03%20-%20EDA%20Using%20SQL/jupyter-labs-eda-sql-coursera_sqllite_2.ipynb*

Build an Interactive Map with Folium

Objective

- Analyze existing launch site locations in order to uncover some factors involved in finding an optimal location for building a launch site …
- … using Interactive Map built with Folium.

GitHub URL of Interactive map with Folium notebook *: https://github.com/PMDOUGLAS23/Applied-DS-C/blob/c117f274942e14012f6baaebdf0c0b5daa5946d4/05%20-%20Interactive%20Visual%20Analytics%20and%20Dashboards/lab_jupyter_launch_site_location_2.ipynb*

Build an Interactive Map with Folium

1 Mark all launch sites on a map

- Create **a Folium Map object** with an initial center location to be NASA Johnson Space Center at Houston, Texas
- Add a circle for each launch site : Create and add `**folium.Circle**` and `**folium.Marker**` for each launch site on the site map
- Use `folium.Circle` to add a highlighted circle area with a text label on a specific coordinate

2 Mark the success / Failure launches for each site on the map

- Assign a color to launch_outcome : 'green' for 'success' and 'red' for 'failure'
- Add **marker_cluster** to the site_map
- For each launch record, create a **Marker object** with its coordinate and customize the **Marker's icon** property to indicate if this launch was successful or a failure

3 Calculate the distances between a launch site to its proximities

- Add **Mouse Position** to get the coordinate (Lat, Long) for a mouse over on the map
- Mark down a point on the closest coastline and calculate the distance between the coastline point and the launch site and create a `folium.Marker` to show the distance
- Draw a 'PolyLine' between a launch site to the selected coastline point : Create a `**folium.PolyLine**` object using the coastline coordinates and launch site coordinate
- draw a line between a launch site to its closest city, railway, highway, etc. Find `MousePosition coordinates and Create a marker with distance to a closest city, railway, highway, etc.

Build a Dashboard with Plotly Dash

- **The objective is to build for users to perform interactive visual analytics on SpaceX launch data**
- **The dash board is composed of a drop down list and a range slider interacting with :**
  - A pie chart showing the success launch by site
  - A scatter point chart showing the relationship between the payload and the success launch by site
- **Elements use to build the dash board :**
  - Launch Site **Drop-Down component**➔ to visualize data for all launch site or a specific one
  - **Callback function to render launch success-pie-chart** based selected site dropdown
  - **Range slider** to Select Payload➔ to visualize data for a selected range of payload mass.
  - **Callback function to render the success-payload-scatter-chart scatter plot** ➔ to show the relationship between the payload mass and the launch success

***GitHub URL of Dashboard – Plotly Dash notebook** : https://github.com/PMDOUGLAS23/Applied-DS-C/blob/81b7fc729b066e5cc5a0640035bad23a8664000d/05%20-%20Interactive%20Visual%20Analytics%20and%20Dashboards/SpaceX_launch_Record_Dashboard.ipynb*

Predictive Analysis (Classification)

1. Perform exploratory Data Analysis and determine Training Labels

Define the target Variable Y from column 'Class' :
Y = data['Class'].to_numpy()

→

Standardize the features data X (from dataset_part_3.csv)
preprocessing.StandardScaler()
.fit(X).transform(X)

→

split the data X and Y into training and test data
X_train, X_test, Y_train, Y_test =
train_test_split(X, Y, test_size = 0.2,
random_state = 2)

2. Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

3. Calculate the accuracy of SVM, classification Tress and Logistic Regression models using the method *score* on the test data, and compare to find the best model

GitHub URL of predictive analysis lab notebook *: https://github.com/PMDOUGLAS23/Applied-DS-C/blob/9e25d4d106880cb5bb3520d1b4fe6b14c34f289d/06%20-%20ML%20Prediction/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite_3.ipynb*

Results

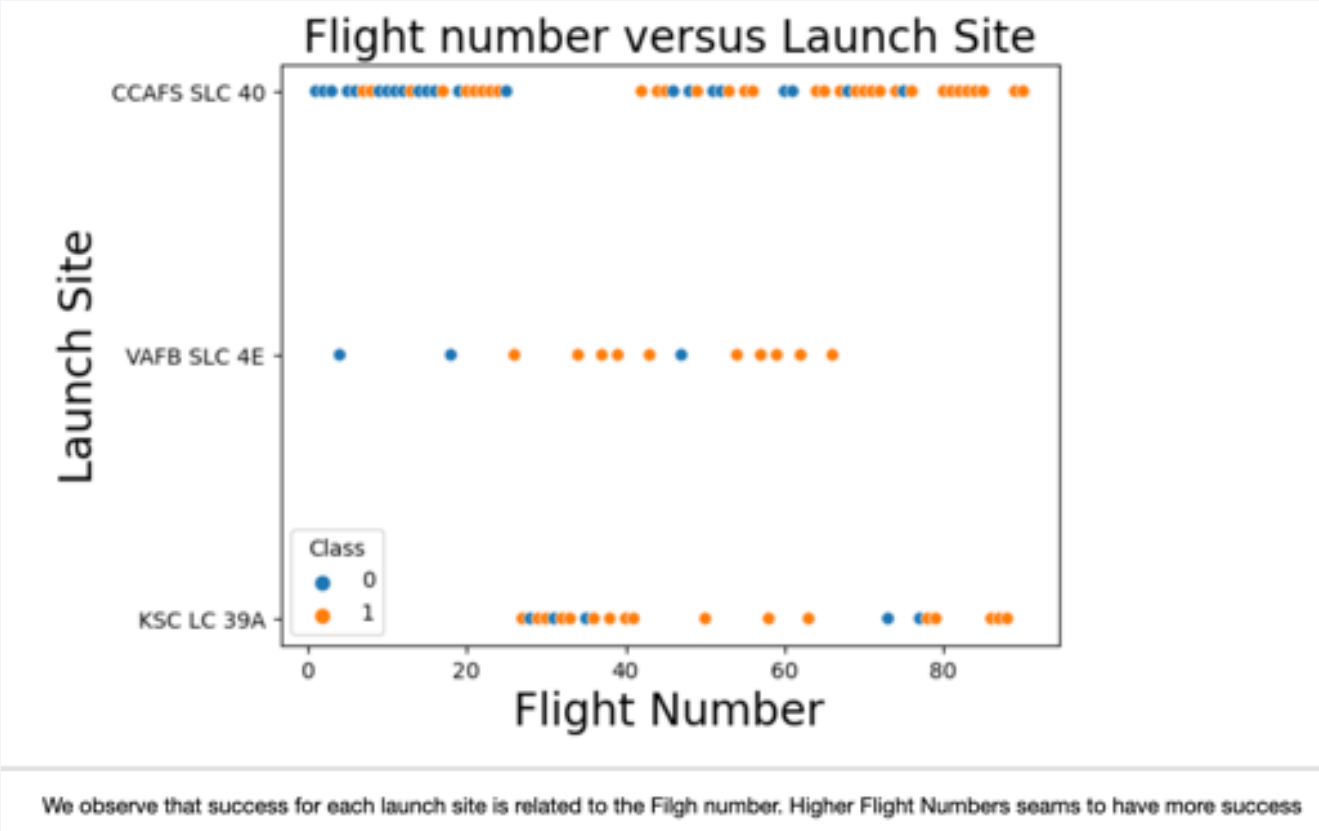
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

We observe that on each launch site, the Success is related to the Flight number.

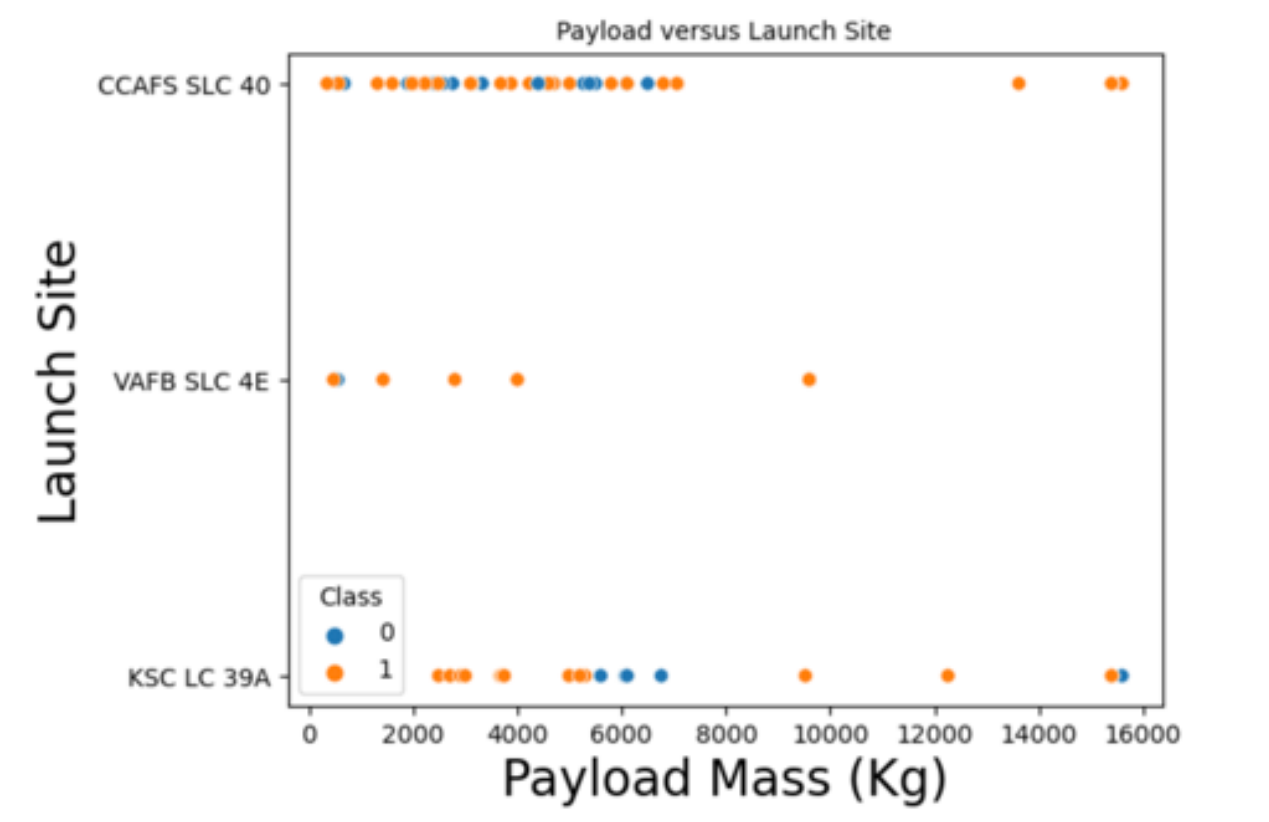


We observe that success for each launch site is related to the Filgh number. Higher Flight Numbers seams to have more success

Payload vs. Launch Site

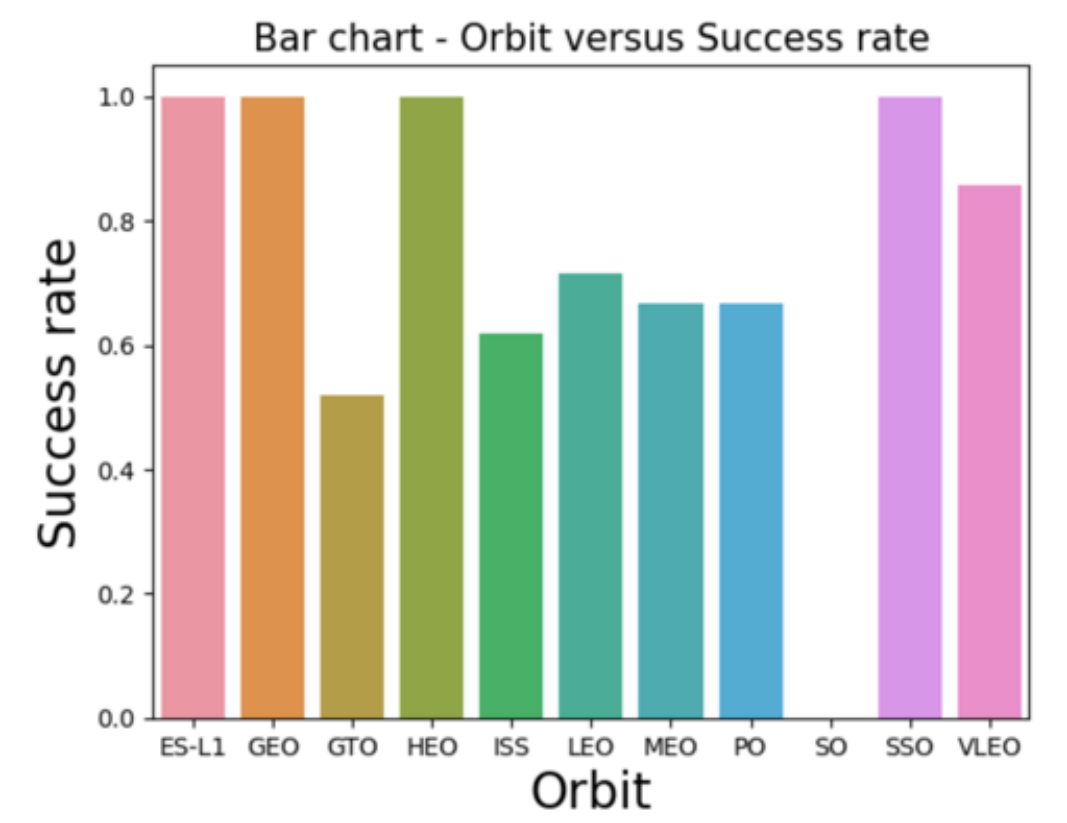
The scatter plot shows a relationship between payload mass and landing success :

- On VAFB SLC 4E Site, no heavy launch, Success for all launch with payload mass over 1000 kg
- On CCAFS SLC 40 Site, success likely for launch with payload mass over 7500 kg
- On KSC LC 39A Site, Failure likely for launch with paylaod mass between 5500 kg and 7000 kg



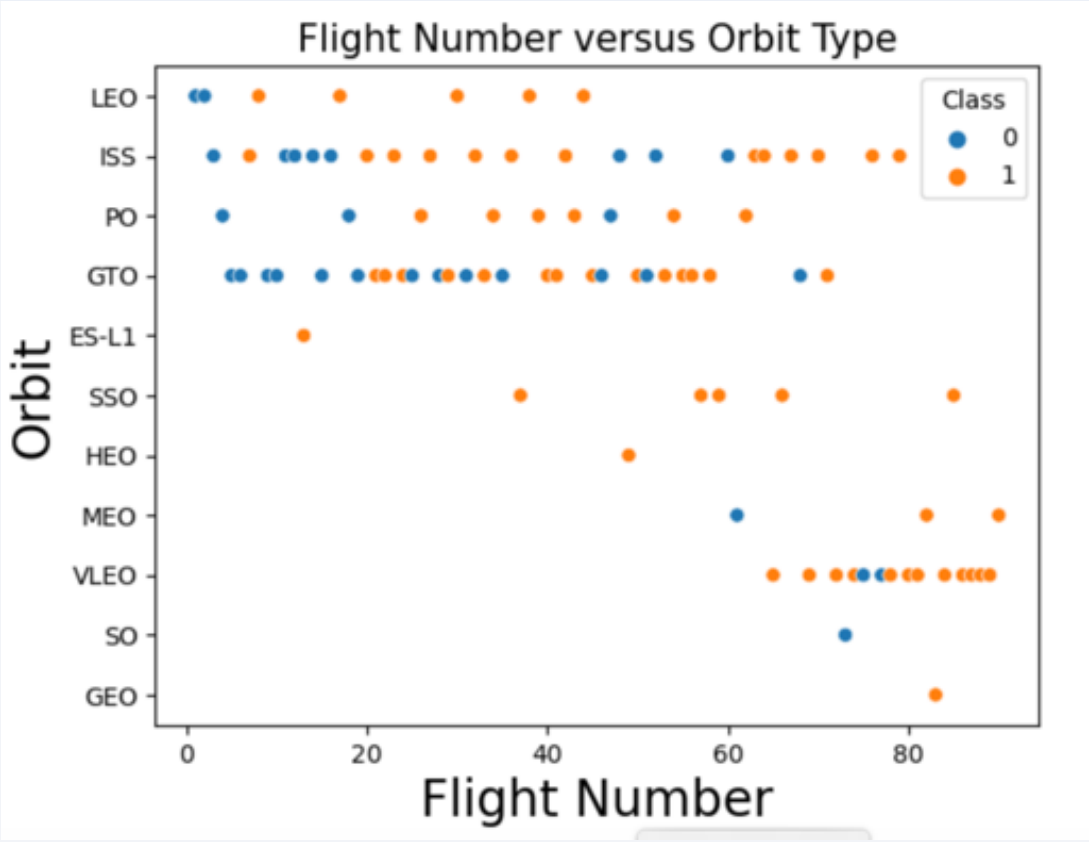
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO have the highest success rate 100%.
- We also note note that VLEO orbit has a notable success around 85%
- There is no launch on SO orbit
- GTO orbit has the lowest success rate (success rate around 50%)



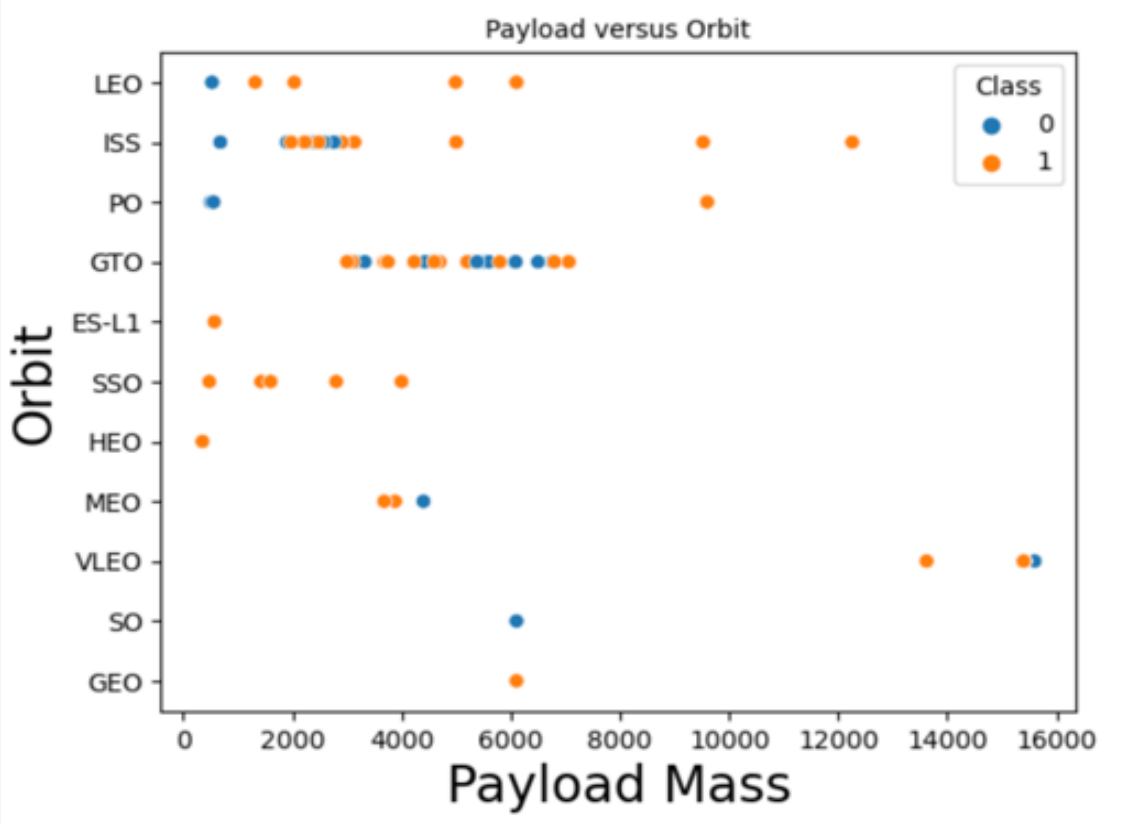
Flight Number vs. Orbit Type

- 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.
- We observe also success for all SSO landings



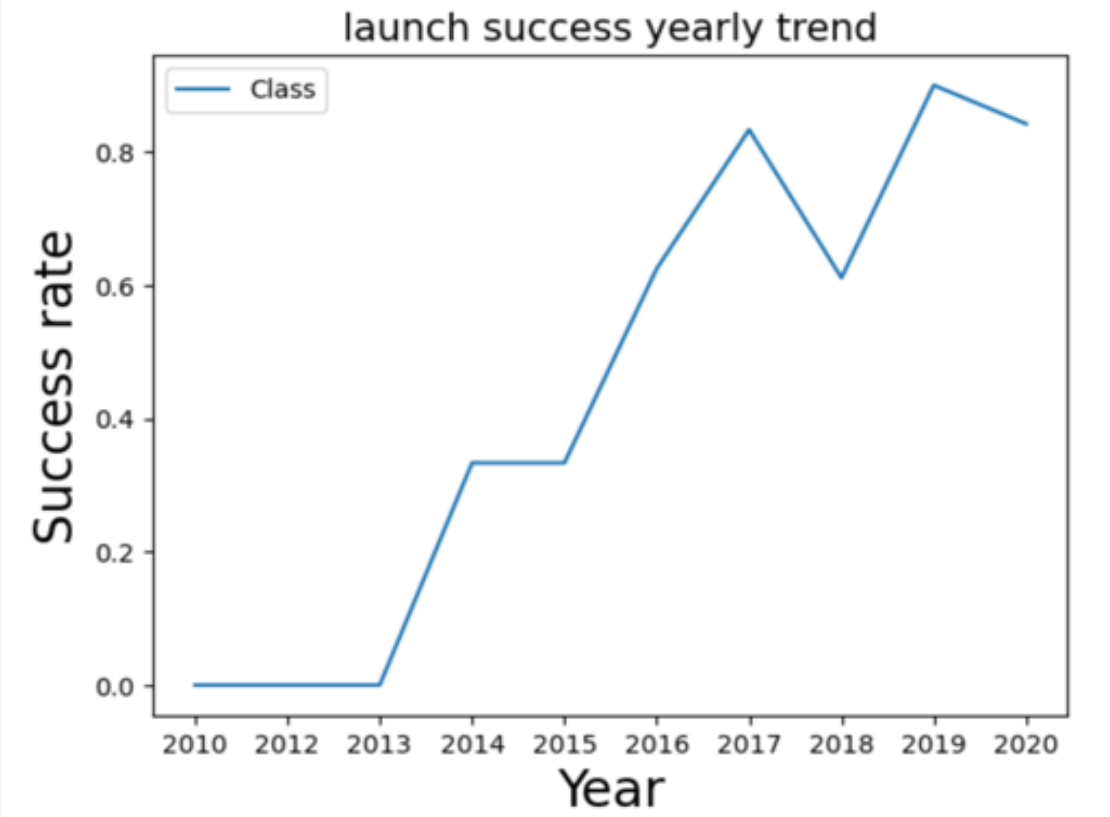
Payload vs. Orbit Type

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



All Launch Site Names

- There are 4 distinct launch sites

%sql select distinct Launch_site from SPACEXTBL

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

*%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5*

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	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	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- For the 5 records above on CCAFS LC-50 , either there is no landing attempt or the landing attempt is a failure

Total Payload Mass

- the total payload carried by boosters launch by NASA CRS is : 45 596 kg

total_payload_mass
45596.0

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.66 kg
- Booster F9 v1.1 carries in average small paylaods

avg_payload_mass
2534.6666666666665

%sql select avg(PAYLOAD_MASS__KG_) as avg_payload_mass from SPACEXTBL where Booster_Version like 'F9 v1.1%'

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was achieved on august, 1st 2018 :

min(Date)
01/08/2018

```
%sql select min(Date) from SPACEXTBL where
Landing_Outcome = 'Success (ground pad)'
```

- It took 8 years (from 2010 to 2018) to achieve the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000kg and 6000 kg

Names of the 23 boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 :

- 4 booster wwith F9 V1.1 …
- 3 booster with F9 B4
- 8 boosters with F9 FT …
- 8 boosters with F9 B5 …

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

We observe that the missions have a 99% success rate

success_mission_outcomes
100

failure_mission_outcome
1

Boosters Carried Maximum Payload

- 12 boosters carried the maximum payload mass
- All of them are F9 B5 booster version series

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

Failed landing_outcome in drone ship, their booster versions, and launch site names in year 2015 :

Year	Monthname	Landing_Outcome	Booster_Version	Launch_Site
2015	october	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	april	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Two failed landing_outcome in drone ship in april and october in year 2015 with F9 V1.1 booster versions category

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

We observe that between 2010-06-04 and 2017-03-20 most of the attempt landing where done on Drone ship and ground pad

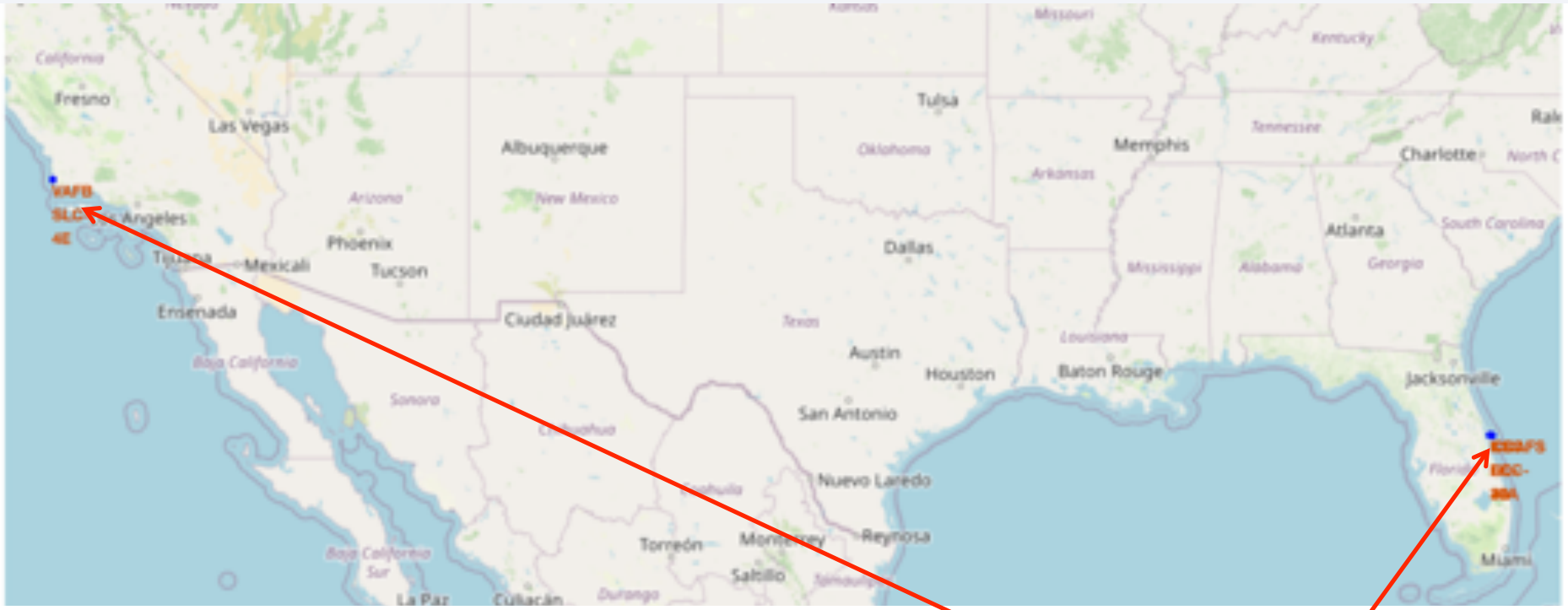
- Most of the attempt landing where on Drone Ship with 5 successful landing and 5 failures
- All the 5 landing on ground pad where successful

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

Section 3

Launch Sites Proximities Analysis

Launch sites locations on Folium Map

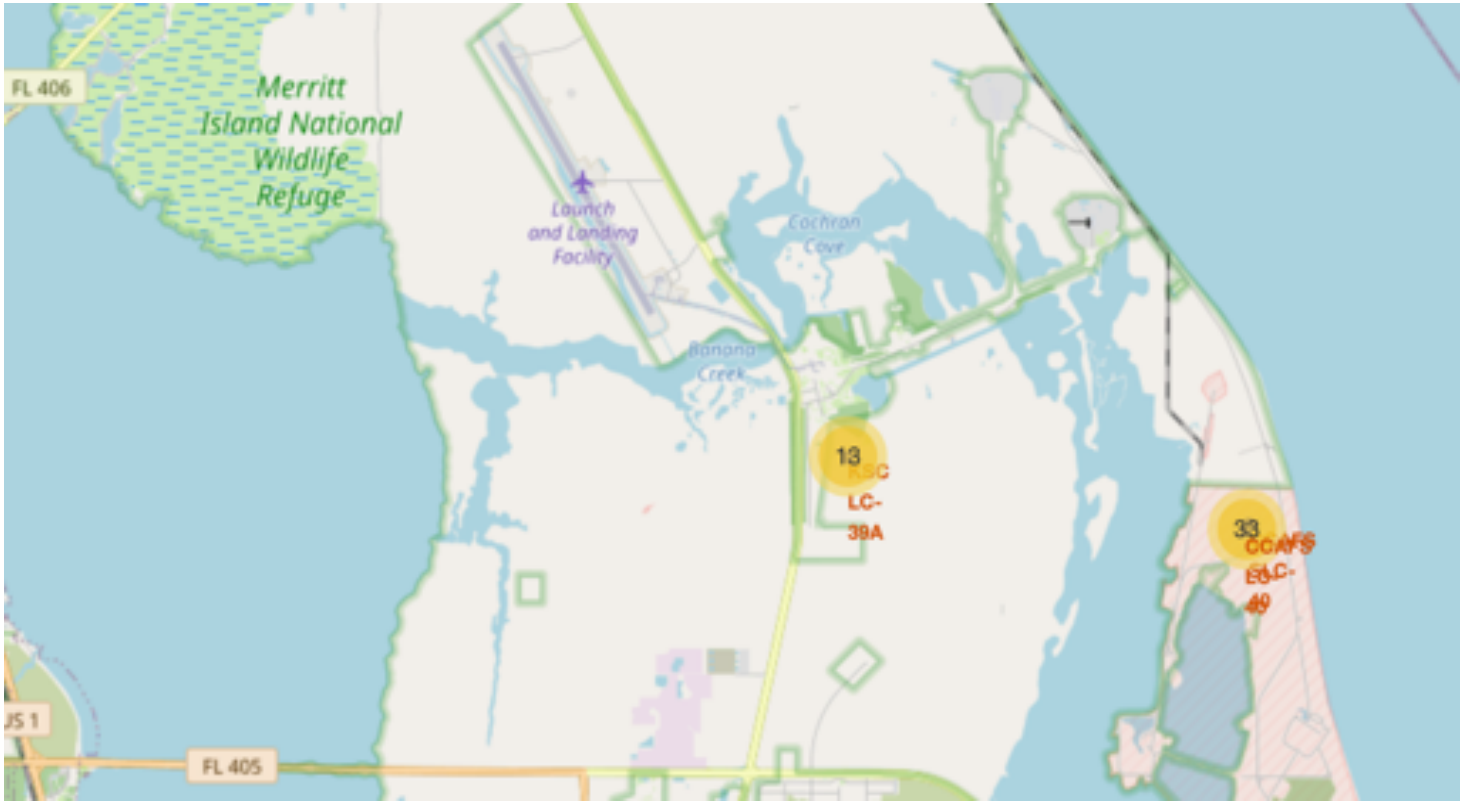


The launch sites are located not very far from the coast : California west coast and Florida east coast

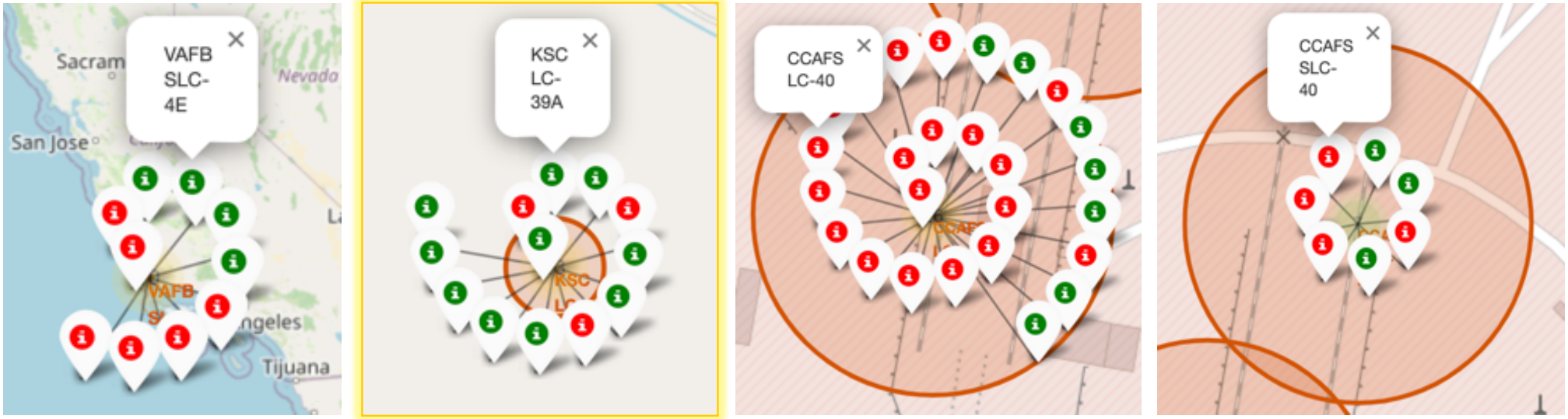
Launch Sites Locations and Landing outcome map

The maps shows that there were :

- 13 launch on KSC LC 39A site
- 33 launch on the distributed between CCAFS SLC 40 Site (7 launch) and CCAFS LC 40 Sites (26 launch)



Launch Sites Locations and Launch outcome map

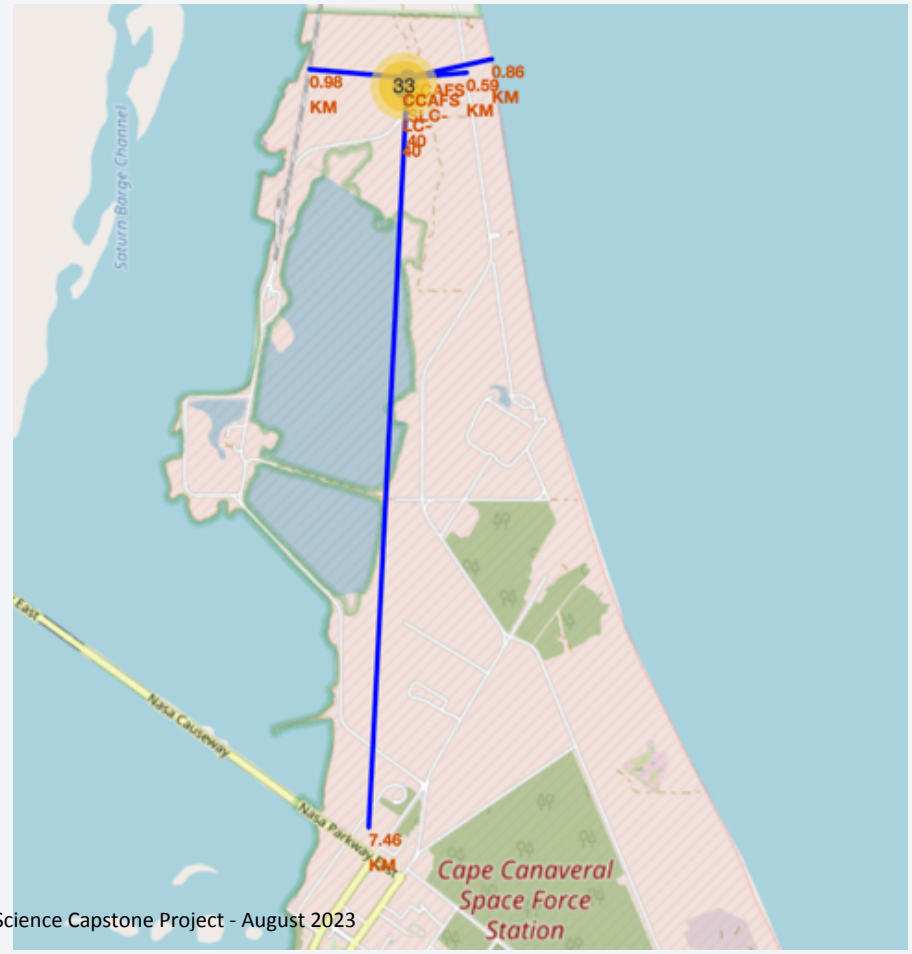


When we zoom in, we can visualize the 1st stage landing outcome (success in green and failure in red) :

- KSC LC 39 launch site in Merritt Island has the highest launch success rate (77%)
- CCAFS LC 40 site holds the highest number of launch (26) but with success rate only around 27%

Map with distance to a closest city, railway, highway, ...

- Launch site are in proximities to railways and roads (less than 1 km) probably to ease the transport of equipment
- The launch sites are not from from from coastline
- The launch cities keep certain distance away from cities probably for safety and environmental concerns

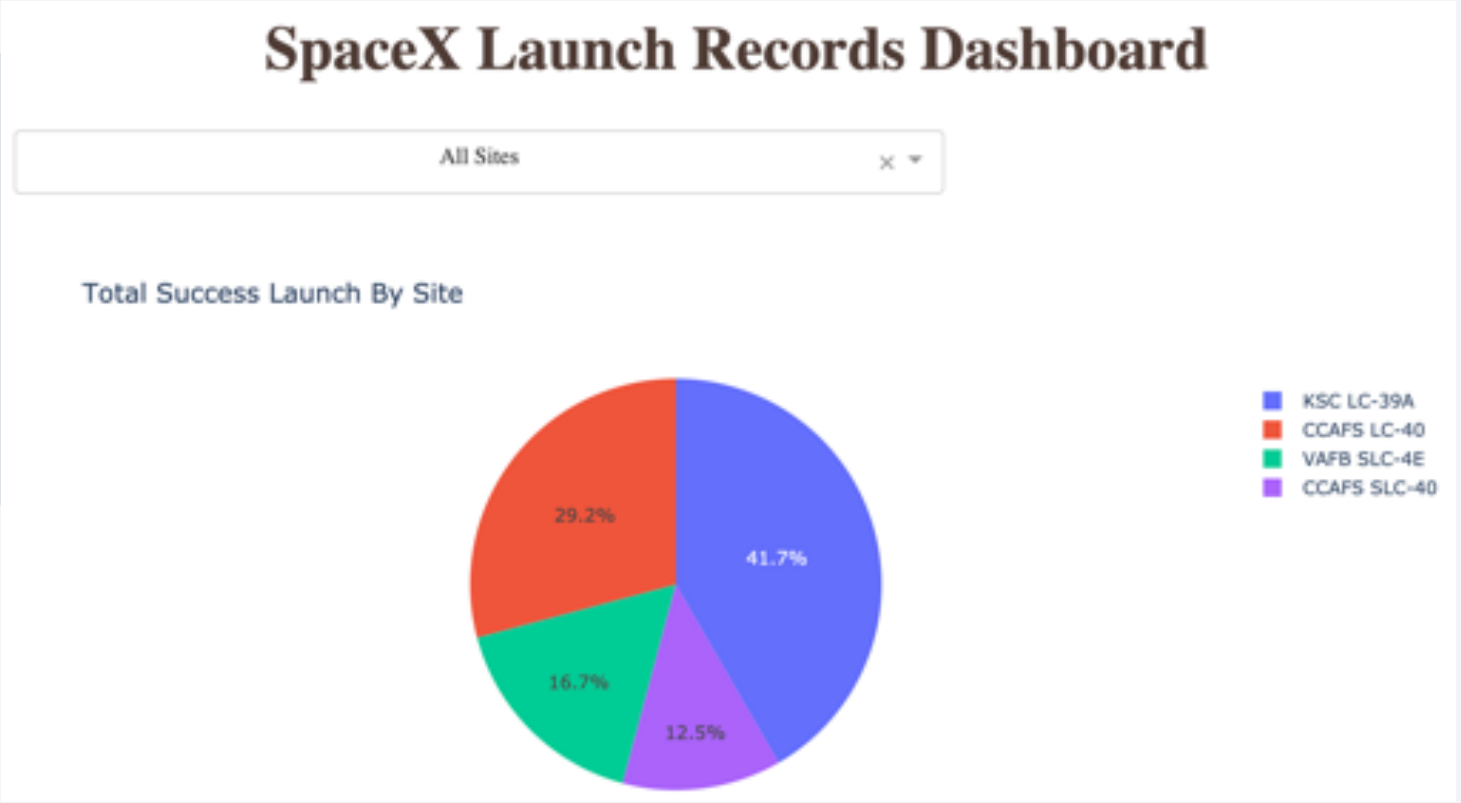


Section 4

Build a Dashboard with Plotly Dash

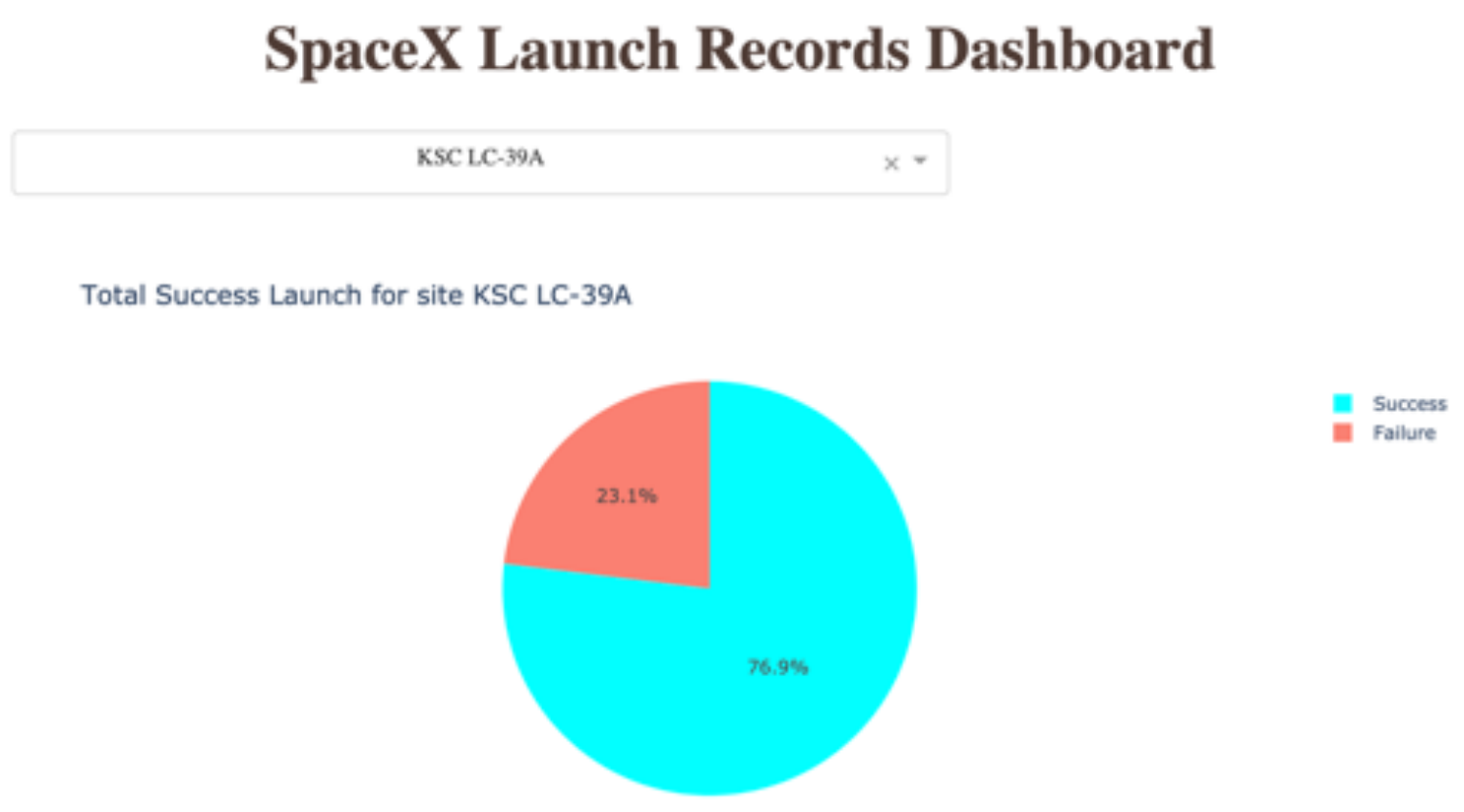
Launch success count for all site

- KSC LC-39A has the highest count of success launch at 41.7%
- Followed by the CCAFS LC-40 site with 29.2%



Pie chart for the launch site with highest launch success ratio

KSC LC-39A launch site has the highest launch success rate at 76,9%



Payload vs. Launch Outcome scatter plot for all sites

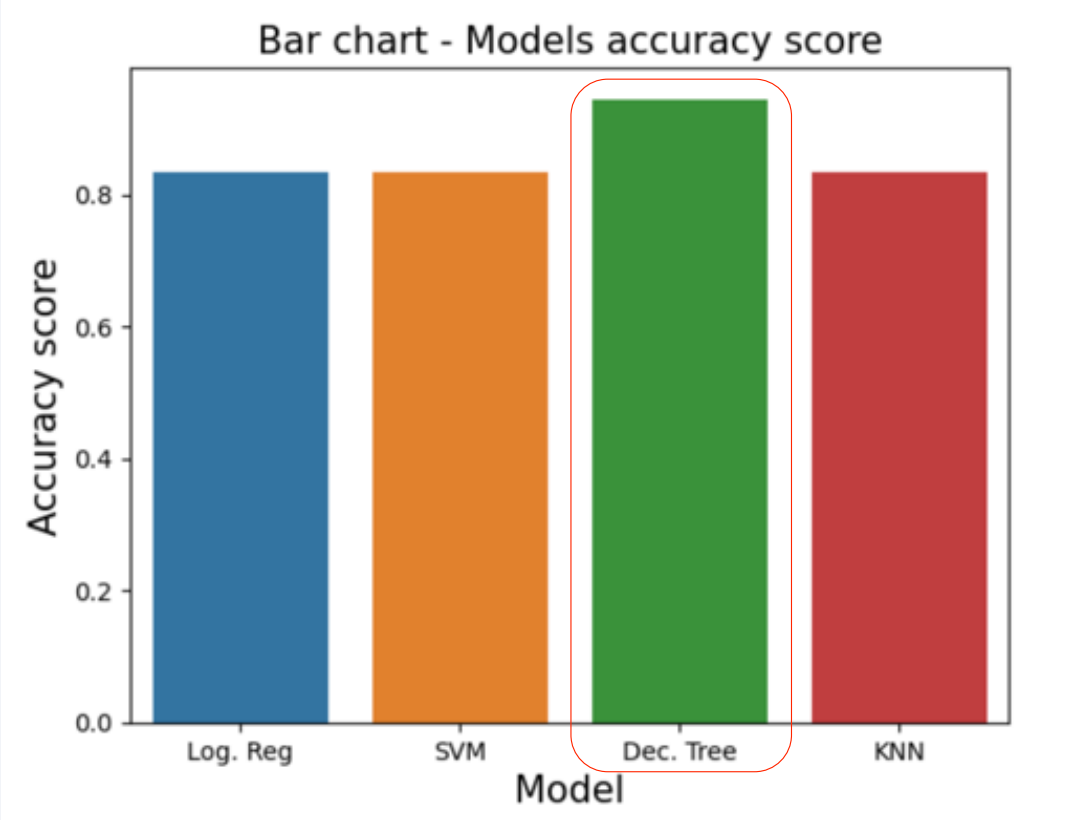
- The Booster version category FT has the largest success rate
- The payload range [2000 kg, 4000 kg] has the largest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy of Decision Tree Model

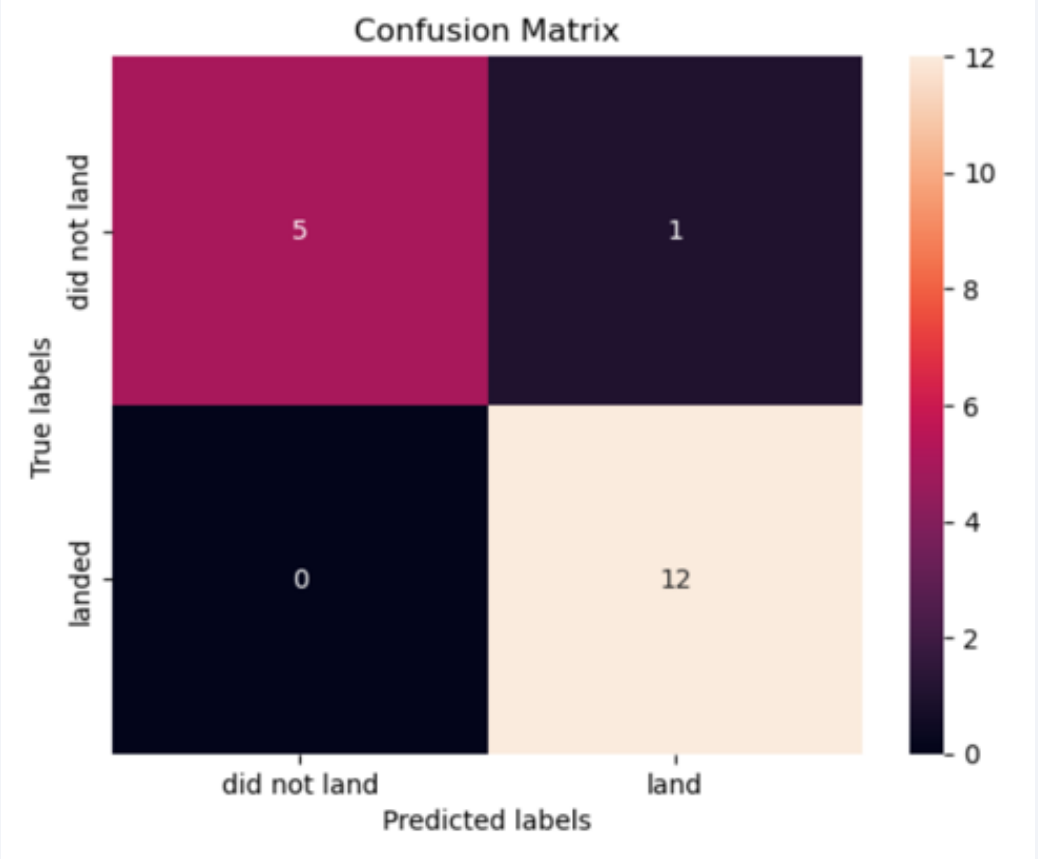
The Decision Tree model has the highest classification accuracy score : 94,44 %



Confusion Matrix of Decision Tree Model

The confusion matrix shows of the decision tree model shows :

- **Only 1 false positive** : 1 false prediction of success landing on test data
- **Zero false negative** : correct prediction of all success landing on tests data



Conclusion

- To predict the predicting of Outcome of the first stage landing (Categorical variable Class = 1 if successful landing, 0 otherwise), we've selected the following features after EDA Analysis : *FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'SerialPoint 2'.*

- We've built an interactive dashboard allowing users to interactively explore launches data and visualize the relationships between success of 1st stage landing, payload range, launch site, booster versions of the rocket.

- The Decision Tree Model performed the best, with an accuracy score of 94,4% in predicting the landing outcome of the Falcon 9 first stage, compare to SVM and KNN models.

- With this model, we have a good tool to predict with if a launch will lead to a successful 1st stage recovery landing. We will then be able to predict the price of a rocket launch.


Thank you!