



Tumisang Fokase  
Data Science capstone project on SpaceX  
1 October 2021



# Outline

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



## Executive Summary

The objective of this project is to determine the price of each launch of SpaceX Falcon 9. I will do this by gathering information about Space X and creating dashboards for my team. I will also determine if SpaceX will reuse the first stage. I will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.





## Project background

SpaceX advertises Falcon 9 rocket 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.

## Key Questions

- What influences if the rocket will land successfully?
- Variables that impact the success of a landing?
- Conditions to consider to get the best results





# Executive Summary

- Data collection methodology:
- Request to the SpaceX API, Web Scraping from Wikipedia
- Perform data wrangling and determine training labels.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models





# Objectives

1. Gather data from the SpaceX REST API ([api.spacexdata.com/v4/](https://api.spacexdata.com/v4/))
2. Perform Web Scrapping from wikipedia using BeautifulSoup

## 1. SpaceX API



## 2. Web Scrapping





# Data Collection , SpaceX API

- ◇ Data was collected by making a request to the SpaceX API
- ◇ The GET request was made to parse the SpaceX data
- ◇ Data cleaning processes were applied to handle missing data
- ◇ Data was converted to a .CSV file

- ◇ Getting a response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url).json()
```

- ◇ Converting response to a .json file

```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

- ◇ Apply custom function to clean data
- ◇ Assign list to dictionary
- ◇ Filter dataframe and export to a csv file

[https://github.com/TumisangF/SpaceX/blob/main/1.%20Data%20Collection%20API%20Lab\\_v1.ipynb](https://github.com/TumisangF/SpaceX/blob/main/1.%20Data%20Collection%20API%20Lab_v1.ipynb)

<https://www.kaggle.com/fokase/spacex-1-data-collection>



# Data Collection, Web Scraping

- ◇ Extract a Falcon 9 launch records HTML table from Wikipedia
- ◇ Perform an HTTP GET method to request the Falcon 9 Launch HTML page
- ◇ Extract all columns/variable names from the HTML table header
- ◇ Create a dataframe by parsing the launch HTML tables
- ◇ Convert the dictionary to dataframe

```
1. Getting Response from HTML
page = requests.get(static_url)

2. Creating BeautifulSoup Object
soup = BeautifulSoup(page.text, 'html.parser')

3. Finding tables
html_tables = soup.find_all('table')

4. Getting column names
column_names = []
temp = soup.find_all('th')
for x in range(len(temp)):
    try:
        name = extract_column_from_header(temp[x])
        if (name is not None and len(name) > 0):
            column_names.append(name)
        except:
            pass

5. Creation of dictionary
launch_dict= dict.fromkeys(column_names)

# Remove an irrelevant column
del launch_dict['Date and time ( )']

launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster'] = []
launch_dict['Booster landing'] = []
launch_dict['Date'] = []
launch_dict['Time'] = []

6. Appending data to keys (refer) to notebook block 12
In [12]: extracted_row = 0
# Extract each table
for table_number, table in enumerate(
    # get table row
    for rows in table.find_all('tr'):
        # check to see if first table

7. Converting dictionary to dataframe
df = pd.DataFrame.from_dict(launch_dict)

8. Dataframe to .CSV
df.to_csv('spacex_web_scraped.csv', index=False)
```

[https://github.com/TumisangF/SpaceX/blob/main/2.%20jupyter-labs-webscraping\\_.ipynb](https://github.com/TumisangF/SpaceX/blob/main/2.%20jupyter-labs-webscraping_.ipynb)  
<https://www.kaggle.com/fokase/spacex-2-data-collection-web-scraping>



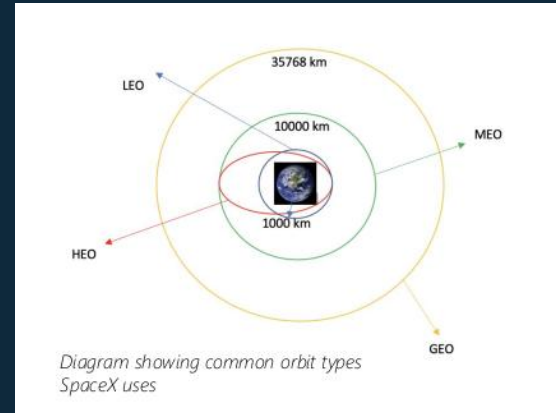
# Data Wrangling

The Objective of data wrangling is to find patterns in the data and determine what would be the features for training supervised models

Steps taken to process the data

- ◇ Calculate the number of launches on each site
- ◇ Calculate the number and occurrence of each orbit
- ◇ Calculate the number and occurrence of each mission outcome per orbit type
- ◇ Create a landing outcome label from outcome column

Each launch is dedicated of a certain orbit



<https://github.com/TumisangF/SpaceX/blob/main/3%20labs-jupyter-spacex-Data%20wrangling.ipynb>  
<https://www.kaggle.com/fokase/spacex-3-data-wrangling>



# EDA With Data Visualization

The purpose of Data visualization is to examine visually how different features affect the success rate. Different features that have been found to have an impact on the success rate will then be used in Machine Learning Algorithms to predict the Price of each launch.

Potential variables explored are:

- ◇ Flight Number vs Payload
- ◇ Flight Number vs launch site
- ◇ Payload vs launch site
- ◇ Orbit vs flight number
- ◇ Payload vs orbit type
- ◇ Orbit vs payload mass
- ◇ Mean vs orbit
- ◇ success rate vs year

[https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz\\_.ipynb](https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz_.ipynb)

<https://www.kaggle.com/fokase/spacex-5-data-viz>





# EDA With SQL

## Summary of the SQL queries you performed on the Db2 database

- ◇ Displaying the names of the unique launch sites in the space mission
- ◇ Displaying 5 records where launch sites begin with the string 'KSC'
- ◇ 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 where the successful landing outcome in drone ship was achieved.
- ◇ Listing the names of the boosters which have success in ground pad and have payload mass greater than 400 but less than 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 records which will display the month names, successful landing\_outcomes in ground pad ,booster
- ◇ versions, launch\_site for the months in year 2017
- ◇ Ranking the count of successful landing\_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

[https://github.com/TumisangF/SpaceX/blob/main/4.%20jupyter-labs-eda-sql-coursera\\_.ipynb](https://github.com/TumisangF/SpaceX/blob/main/4.%20jupyter-labs-eda-sql-coursera_.ipynb)

<https://www.kaggle.com/fokase/spacex-4-data-wrangling-with-sql>



# Predictive Analysis





# Predictive analysis (Classification)

Summary of how the model was built:

- ◇ Load the dataset
- ◇ Standardize the data
- ◇ Split data into training and test sets
- ◇ Decide which type of machine learning algorithms to use
- ◇ Set the parameters and algorithms to GridSearchCV
- ◇ Fit out datasets into the GridSearchCV and train our model

Evaluation of the model:

- ◇ Check the accuracy of each model
- ◇ Tune the hyperparameters of each model
- ◇ Feature Engineering

Finding the best Performing Classification Model:

- ◇ The model with the highest accuracy score is chosen as the algorithm

[https://github.com/TumisangF/SpaceX/blob/main/7.%20%20SpaceX\\_Machine%20Learning%20Prediction\\_Part\\_5.ipynb](https://github.com/TumisangF/SpaceX/blob/main/7.%20%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)

<https://www.kaggle.com/fokase/spacex-7-predictive-analysis>



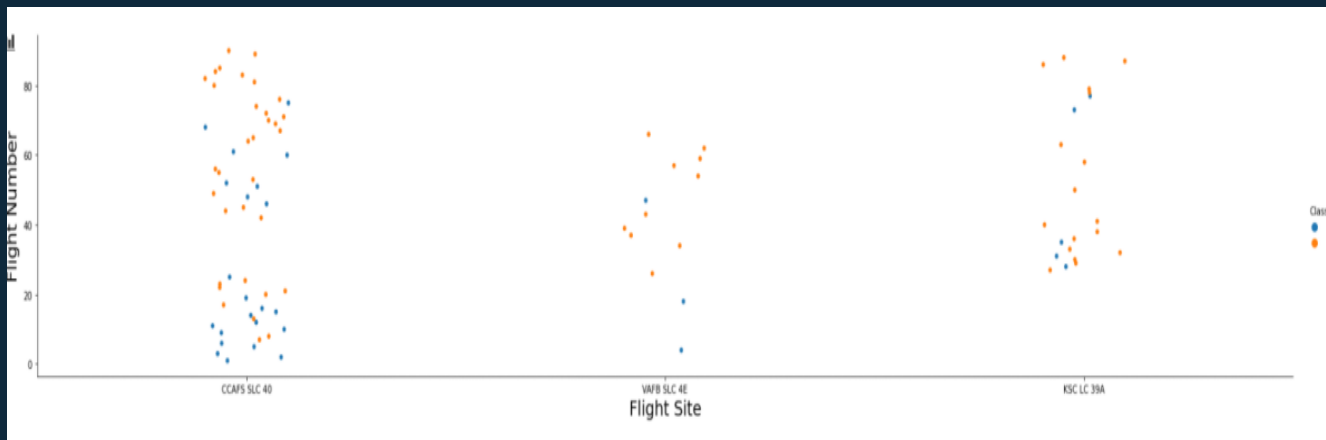


# A picture is worth a thousand words

I have added plots/graphs and interactions dashboards to visualize data.



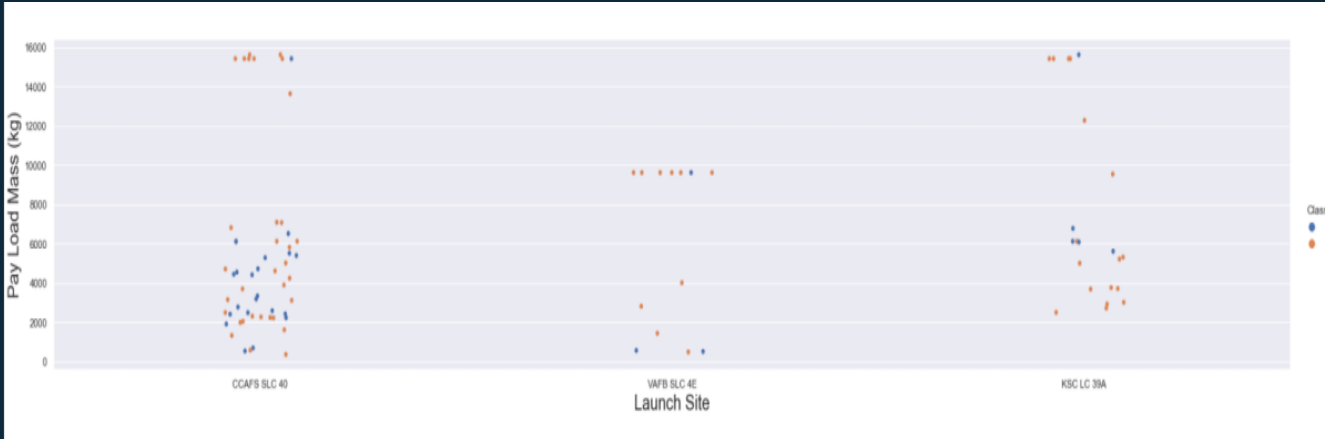
# Flight Number VS Launch Site



Key Observation is that the more the amount of flights at a launch site the greater the success rate at a launch site

[https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz\\_.ipynb](https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz_.ipynb)

# Payload vs Launch Site



Key Observation is that the more the amount of flights at a launch site the greater the success rate at a launch site

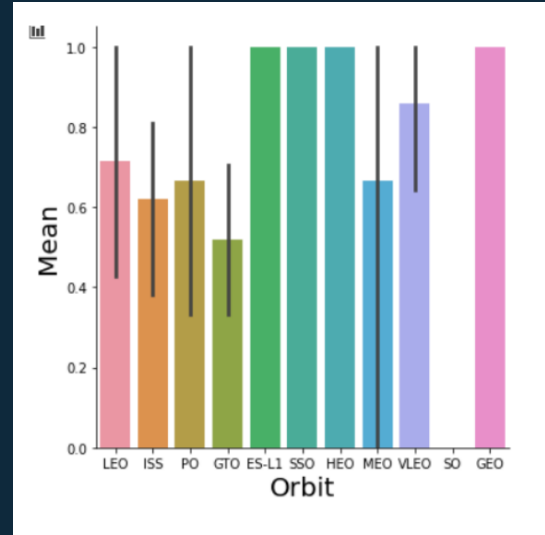
[https://github.com/TumisangF/SpaceX/blob/main/5.%20jupyter-labs-eda-dataviz\\_.ipynb](https://github.com/TumisangF/SpaceX/blob/main/5.%20jupyter-labs-eda-dataviz_.ipynb)



# Success Rate VS Orbit Type

Orbits that have high success rate:

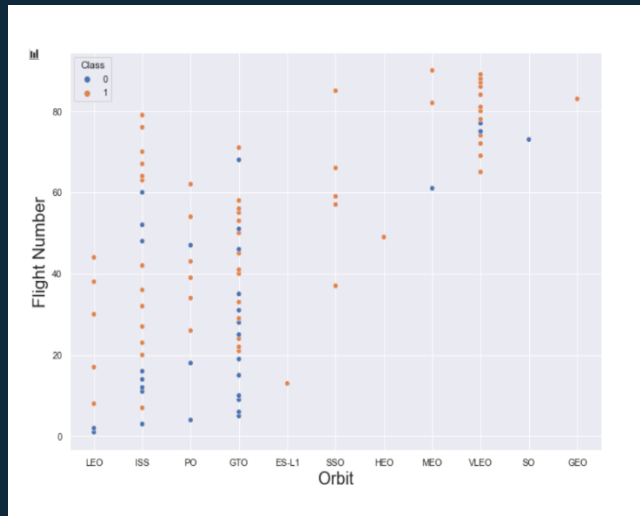
- ◇ GEO
- ◇ HEO
- ◇ SSO
- ◇ ES-L1



[https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz\\_.ipynb](https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz_.ipynb)

# Payload vs Orbit Type

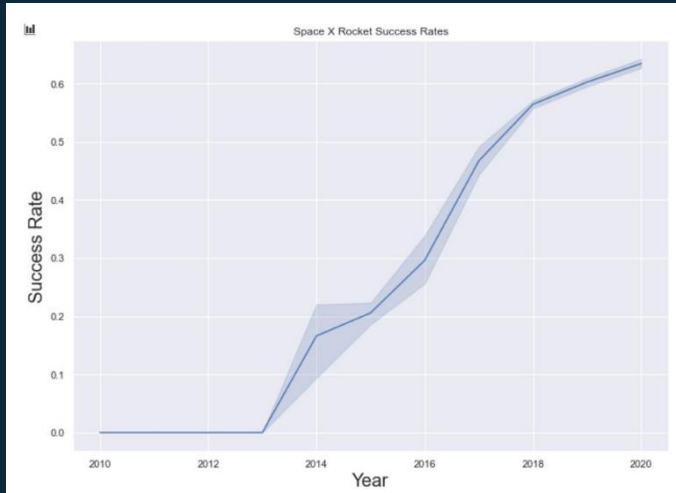
The higher the number of flights  
the more chances of success



[https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz\\_.ipynb](https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz_.ipynb)

# Payload vs Orbit Type

The success rate has been increasing 2013



[https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz\\_.ipynb](https://github.com/TumisangF/SpaceX/blob/main/5.%20%20jupyter-labs-eda-dataviz_.ipynb)



## Data Analysis with SQL

Summary of all SQL queries  
made to the db2 database.





## All Launch site names

SQL Query:

SELECT

DISTINCT(Launch\_Site)

FROM

SpaceX

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

The 'DISTINCT' in the query is used to return only unique value in the 'Launch\_Site' column





# Total Payload mass

SQL Query:

```
SELECT SUM(Payload_Mas_kg) as  
Total_Payload_mass
```

```
FROM SpaceX
```

```
WHERE Customer = 'NASA (CRS)'
```

TOTAL_PAYLOAD_MASS
45596

The 'SUM' function summates all values in the column 'Payload\_mass\_kg'  
The 'WHERE' clause filters in 'NASA (CRS)' customer.





# Average payload mass

```
SELECT AVG(Payload_mass_kg) as  
  
    AVERAGE_PAYLOAD_MASS  
  
FROM SpaceX  
  
WHERE Booster_Version = 'F9 V1.1'
```

AVERAGE_PAYLOAD_MASS
2928

The 'AVG' function calculates the average of all the values in the 'Payload\_mas\_kg' column  
The 'WHERE' clause filters in all booter versions with the name 'F9 v1.1'





# Date of the first success landing


```
SELECT MIN(DATE) AS  
FIRST_SUCCESSFUL_LANDING_GROUD_PAD  
  
FROM SpaceX  
  
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

AVERAGE_PAYLOAD_MASS
2928

The 'MIN' function finds the minimum date in the 'DATE' column  
The 'WHERE' clause filters in successful landing outcomes







# Successful drone ship landing with payload mass between 4000 and 6000

```
SELECT DISTINCT(BOOSTER_VERSION)

FROM SpaceX

WHERE LANDING__OUTCOME = 'Success (ground ship)'
AND PAYLOA_MASS_KG > 4000 AND
PAYLOA_MASS_KG < 6000
```

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

The 'WHERE' clause filters in only mass values between 4000 and 6000 whose landing outcome was successful





# Total number of successful and failed outcome


SELECT

```
COUNT(LANDING_OUTCOME) AS  
NUMBER_OF_SUCSESFUL_OUTCOMES,  
(SELECT COUNT(LANDING_OUTCOME) AS  
NUMBER_OF_FAILED_OUTCOMES
```

FROM SpaceX

WHERE LANDING\_OUTCOME LIKE 'Success%'

NUMBER_OF_SUCCESSFUL_OUTCOMES	NUMBER_OF_FAILED_OUTCOMES
61	10



Two select queries are used to find the number of successful and failed outcomes

## Booster carries by maximum payload mass

```
SELECT  
  
    DISTINCT (BOOSTER_VERSION),  
    MAX(Payload_mass)  
  
FROM SpaceX  
  
GROUP BY BOOSTER_VERSION  
  
ORDER BY MAX(Payload_mass)
```

MAX function is used to select only Booster\_version that carries the maximum load of 15600kg

BOOSTER_VERSION	2
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600

Result set is truncated, only the first 97 rows have been loaded. Select on the right top of the result to view all loaded rows.

# Launches in 2015

```
SELECT
    LANDING_OUTCOME, BOOSTER_VERSION,
    LAUNCH_SITE

FROM SpaceX

WHERE
    LANDING_OUTCOME = 'Failed (drone ship)'
    AND DATA LIKE '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

Where clause is used to select failed outcomes that occurred in the year 2015

# Ranking landing outcomes between 2010-06-04 and 2017-03-20

SELECT

COUNT(LANDING\_OUTCOME),  
LANDING\_OUTCOME

FROM SpaceX

WHERE

DATE > 2010-06-04 AND DATE < '2017-03-20'

NUMBER_ATTEMPTS	OUTCOME
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)

Result set is truncated, only the first 8 rows have been loaded. Select ["View all loaded data"](#) on the right top of the result to view all loaded rows. [More](#)

This query gives a summary of the description of the outcome as well as the number of attempts



# All launch sites on a Global map

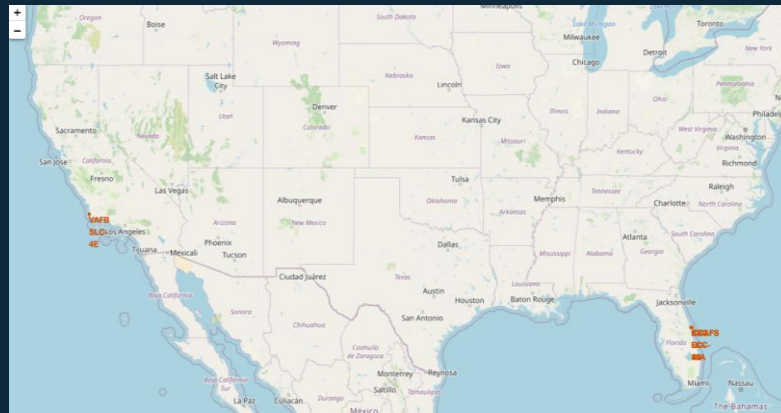




# All launch sites on a global map

The map shows that all launch sites are located in two places in the United States:

- ◆ California
- ◆ Florida



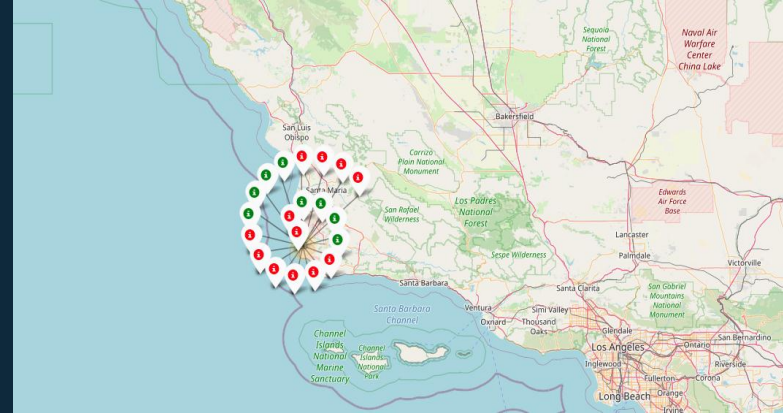
Further observation about the launch site

- ◆ All launch sites are in close proximity to the Equator line because the land at equator is moving 1670 km per hour so launching from the equator makes the spacecrafts move almost 500km/hour faster once it is launched
- ◆ All launch sites are in close proximity to the coast to minimize the risk of having any debris dropping or exploding near



# California launch sites success rate

- ◇ Green markers show successful launches
- ◇ Red markers show unsuccessful launches



Further observation about the launch site

- ◇ 40% of the launches were unsuccessful
- ◇ 60% of the launches were successful



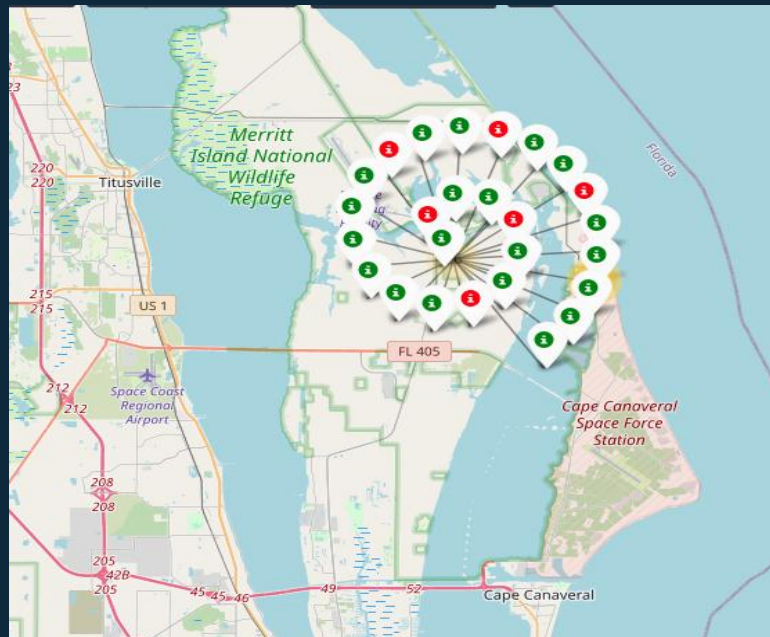


# Florida launch sites success rate

- ◇ Green markers show successful launches
- ◇ Red markers show unsuccessful launches

Further observation about the launch site

- ◇ Florida has a relatively high rate of successful launches

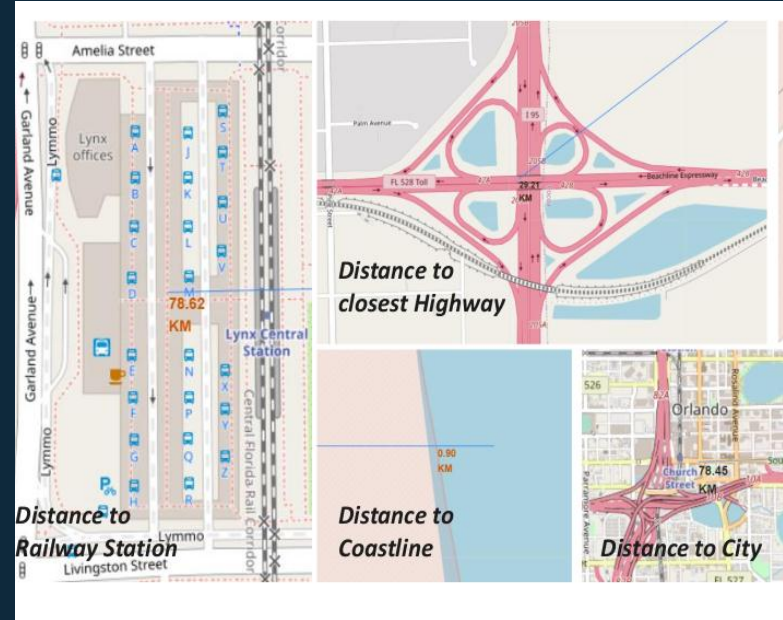




# Florida launch sites success rate

## Observations

- ◆ Launch sites are not in close proximity to the railways, 78.62 km away
- ◆ Launch sites are not in close proximity to the highways, 28.21 km away
- ◆ Launch sites are in close proximity to the coastlines
- ◆ Launch sites do keep a certain distance from the citie

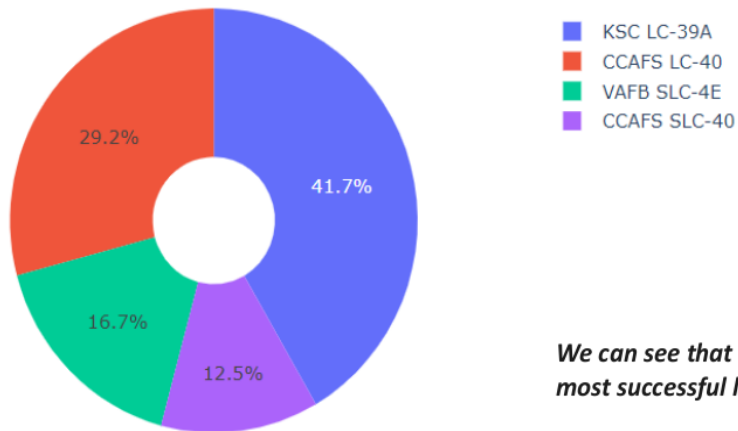




A picture is worth a  
thousand words,  
Data visualization  
results

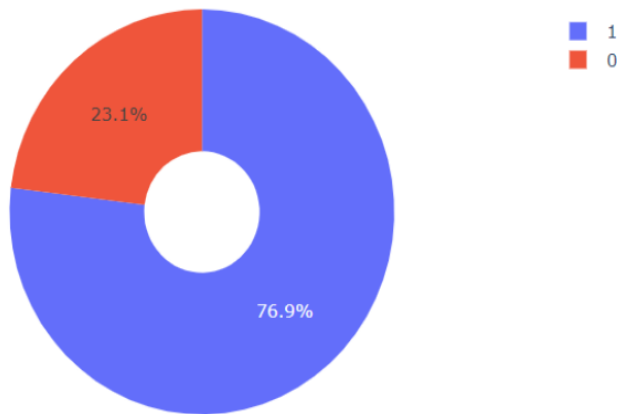


## Success rate at each launch site



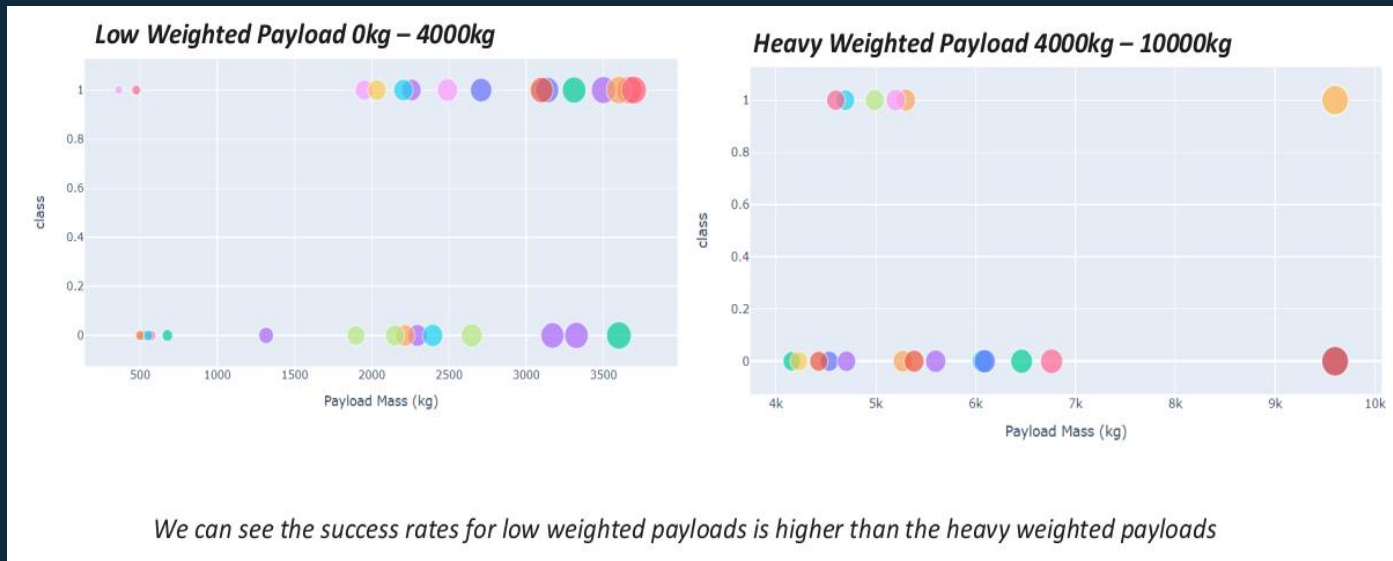
*We can see that KSC LC-39A had the most successful launches from all the sites*

# Pie chart for the launch site with the highest launch success ratio



*KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate*

# Payload mass vs Launch outcome scatter plot

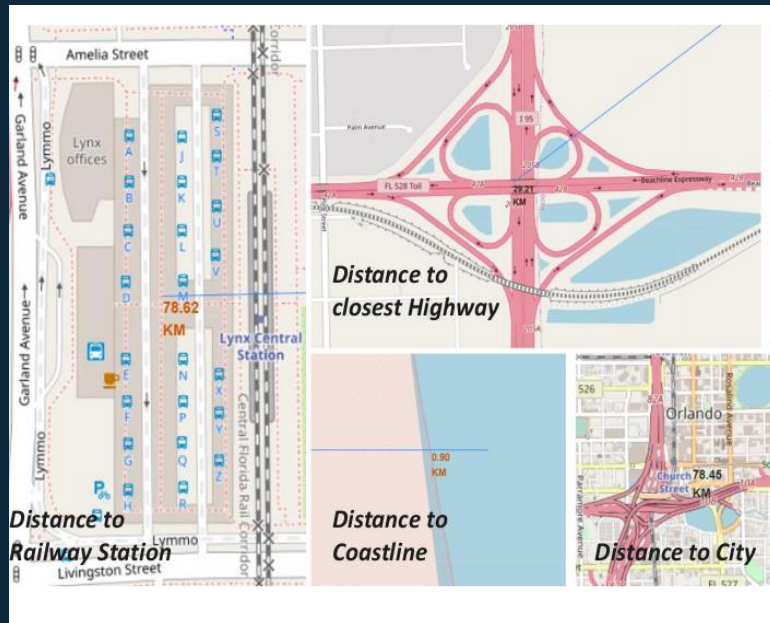




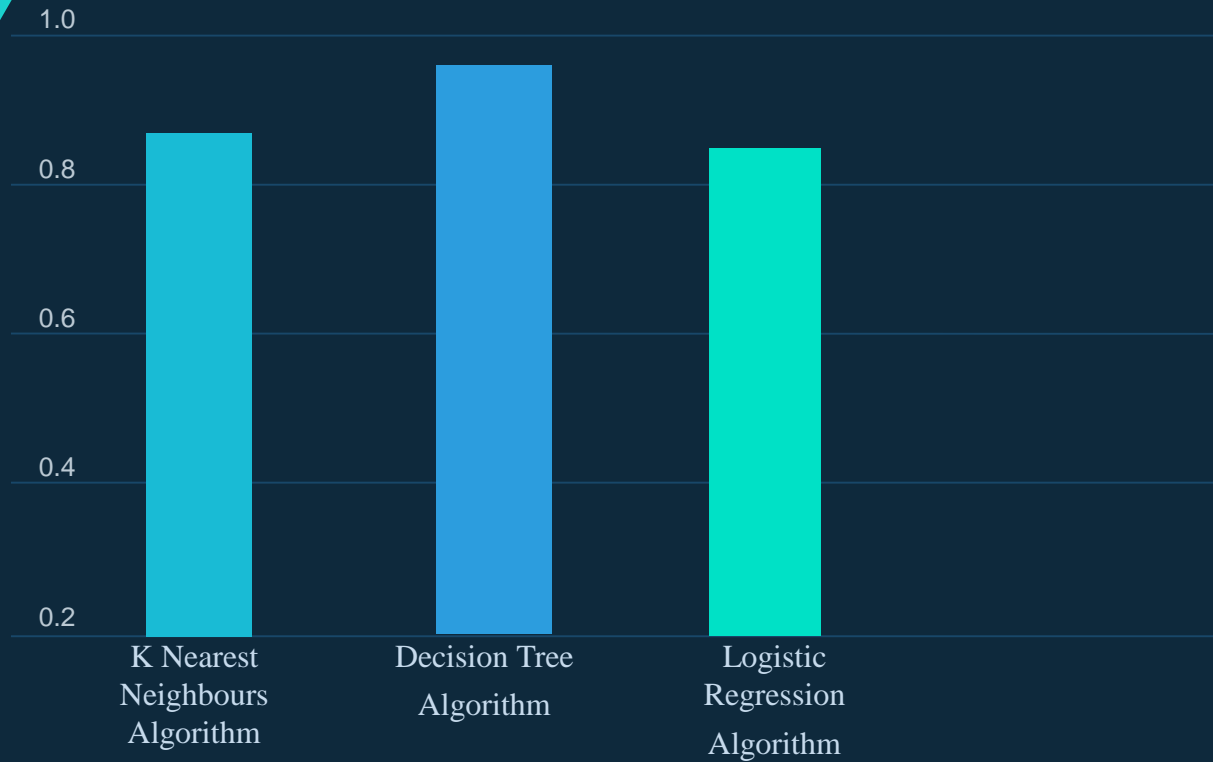
# Florida launch sites success rate

## Observations

- ◆ Launch sites are not in close proximity to the railways, 78.62 km away
- ◆ Launch sites are not in close proximity to the highways, 28.21 km away
- ◆ Launch sites are in close proximity to the coastlines
- ◆ Launch sites do keep a certain distance from the citie



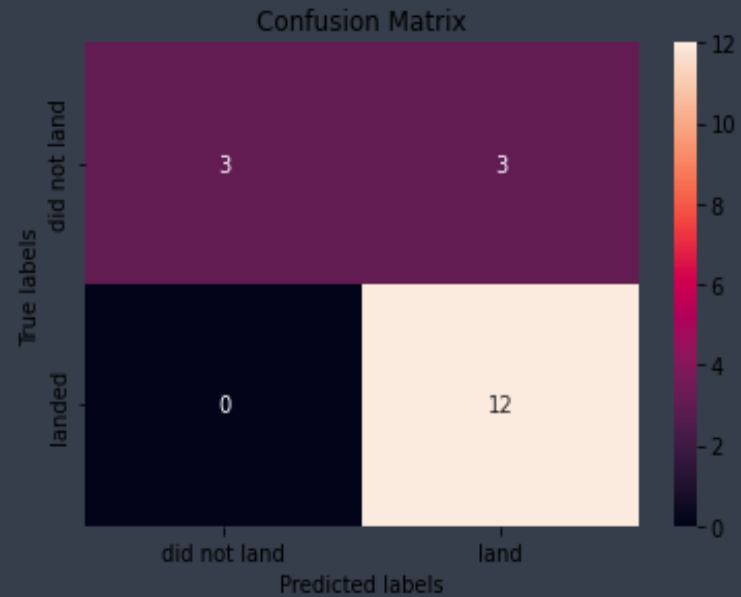
# Florida launch sites success rate





# Confusion Matrix

Examining the confusion Matrix,  
We can see that the Decision Tree  
Algorithm can distinguish between  
the different classes





# Conclusion

- ◇ The Decision Tree Model is the best algorithm because it has the highest accuracy score of 87%
- ◇ KSC LC-39A launch site has the most successful launches from all launch sites
- ◇ Orbits, GEO, HEO, SSO, ES-L1 had the highest success rate



# Appendix

- ◇ IBM capstone project full details:  
<https://www.coursera.org/learn/applied-data-science-capstone?specialization=ibm-data-science>
- ◇ GitHub URL :  
<https://github.com/TumisangF/SpaceX>
- ◇ Presentation template credits URL :
- ◇ [www.slidescarnival.com](http://www.slidescarnival.com)