

## IBM DATA SCIENCE CAPSTONE PROJECT - SPACEX

MUHAMMAD FACHRURROZY





## OUTLINE

**Executive Summary** 

Introduction

Methodology

Results

Conclusion

Appendix



## **EXECUTIVE SUMMARY**

#### SUMMARY OF METHODOLOGIES

- Data collection
- Data wrangling
- EDA with data visualization
- EDA 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

#### HOW WE DID IT

#### **Project background and context**

We predicted if 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 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

#### Common problems that needed solving.

- What influences if the rocket will land successfully?
- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.



## METHODOLOGY

#### **Data collection methodology:**

- SpaceX Rest API
- (Web Scrapping) from Wikipedia

#### **Performed data wrangling (Transforming data for Machine Learning)**

One Hot Encoding data fields for Machine Learning and dropping irrelevant columns

## Performed exploratory data analysis (EDA) using visualization and SQL

• Plotting: Scatter Graphs, Bar Graphs to show relationships between variables to show patterns of data.

Performed interactive visual analytics using Folium and Plotly Dash

#### Performed predictive analysis using classification models

• How to build, tune, evaluate classification models

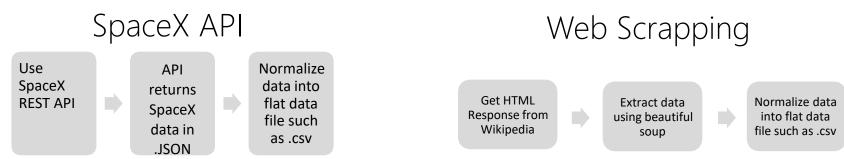


## METHODOLOGY

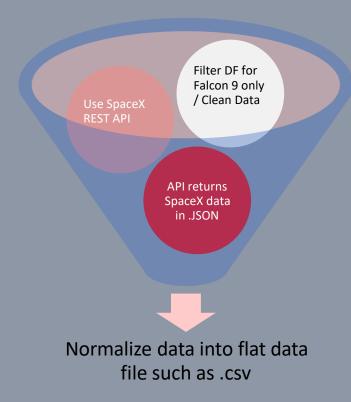
LET'S DIVE IN

## METHODOLOGY

- The following datasets was collected by
  - We worked with SpaceX launch data that is gathered from the SpaceX REST API.
  - This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.
  - The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.
  - Another popular data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup.



## Data collection -SpaceX API



#### 1.Getting Response from API

simplified flow chart

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

#### 2. Converting Response to a .json file

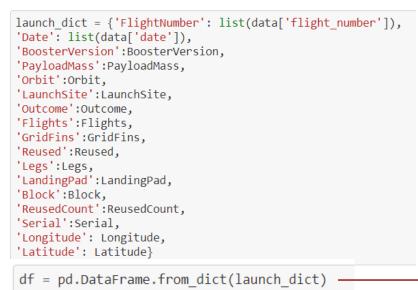
```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

#### 3. Apply custom functions to clean data

getLaunchSite(data)
getPayloadData(data)
getCoreData(data)

getBoosterVersion(data)

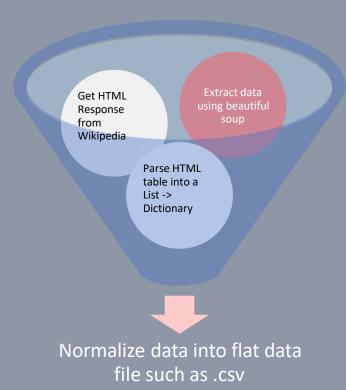
#### 4. Assign list to dictionary then dataframe



5. Filter dataframe and export to flat file (.csv)

data\_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
data falcon9.to csv('dataset part 1.csv', index=False)

## Data collection-Web Scrapping



#### 1 .Getting Response from HTML

page = requests.get(static\_url)

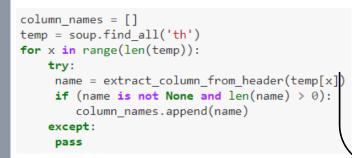
#### 2. Creating BeautifulSoup Object 4

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

#### 3. Finding tables

html\_tables = soup.find\_all('table') •

#### 4. Getting column names



#### 5. Creation of dictionary

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant 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

#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)

## DATA WRANGLING

#### Introduction

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.

We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

#### **Process**

Perform Exploratory Data Analysis EDA on dataset

Calculate the number of launches at each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Export dataset as .CSV

Create a landing outcome label from Outcome column

Work out success rate for every landing in dataset

Each launch aims to an dedicated orbit, and here are some common orbit types:

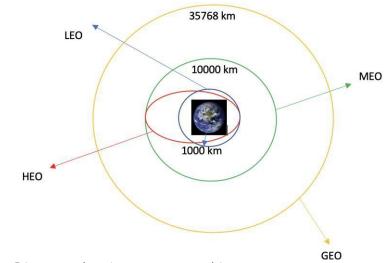


Diagram showing common orbit types SpaceX uses

GitHub URL to Notebook

## EDA WITH DATA VISUALIZATION

Scatter Graphs being drawn:

Flight Number VS. Payload Mass

Flight Number VS. Launch Site

Payload VS. Launch Site

Orbit VS. Flight Number

Payload VS. Orbit Type

Orbit VS. Payload Mass

Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation . Scatter plots usually consist of a large body of data.

GitHub URL to Notebook



Mean VS. Orbit



A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.

## Line Graph being drawn:

Success Rate VS. Year



Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded

## EDA WITH SQL

### Performed SQL queries to gather information about the dataset.

For example of some questions we were asked about the data we needed information about. Which we are using SQL queries to get the answers in the dataset:

- 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 4000 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.



## BUILDING AN INTERACTIVE MAP WITH FOLIUM

To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a *Circle Marker around each launch site with a label of the name of the launch site.* 

We assigned the dataframe launch\_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()

**Using Haversine's formula we calculated the distance** from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. **Lines** are drawn on the map to measure distance to landmarks

#### Example of some trends in which the Launch Site is situated in.

- •Are launch sites in close proximity to railways? No
- •Are launch sites in close proximity to highways? No
- •Are launch sites in close proximity to coastline? Yes
- •Do launch sites keep certain distance away from cities? Yes

## PREDICTIVE ANALYSIS (CLASSIFICATION)

#### **BUILDING MODEL**

- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.

#### **EVALUATING MODEL**

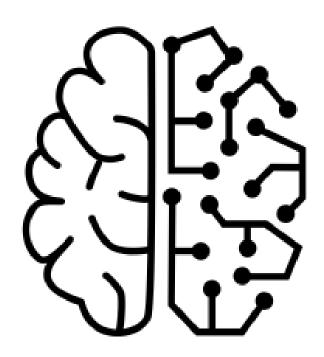
- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix

#### **IMPROVING MODEL**

- Feature Engineering
- Algorithm Tuning

#### FINDING THE BEST PERFORMING CLASSIFICATION MODEL

- The model with the best accuracy score wins the best performing model
- In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.



GitHub Link to source code



## RESULT

LOOKING AHEAD

## RESULT

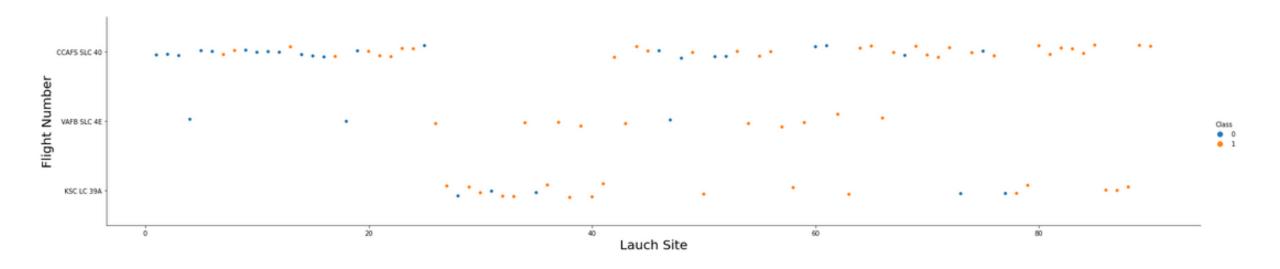


- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



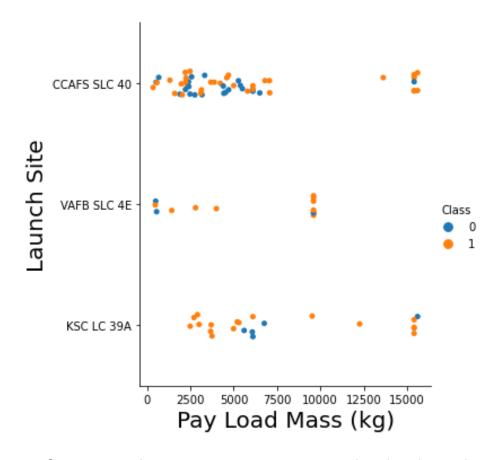
# EDA WITH VISUALIZATION

## FLIGHT NUMBER VS FLIGHT SITE



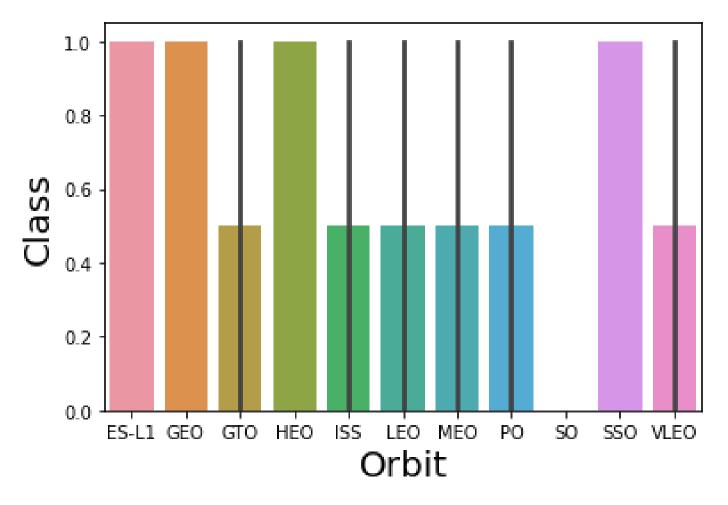
The more amount of flights at a launch site the greater the success rate at a launch site.

## PAYLOAD MASS VS LAUNCH SITE



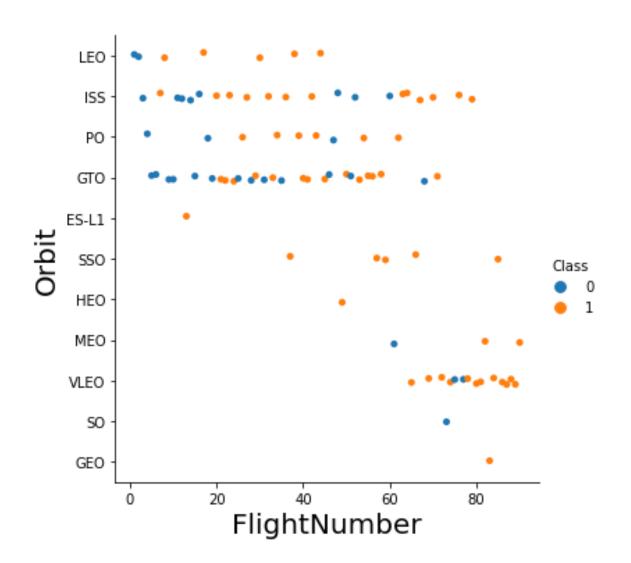
The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket. There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependent on Pay Load Mass for a success launch.

## SUCCESS RATE VS ORBIT TYPE



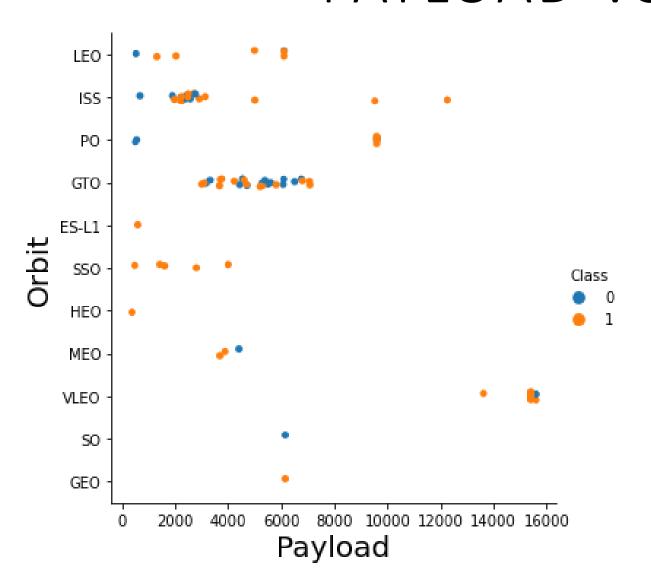
Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

## FLIGHT NUMBER VS ORBIT TYPE



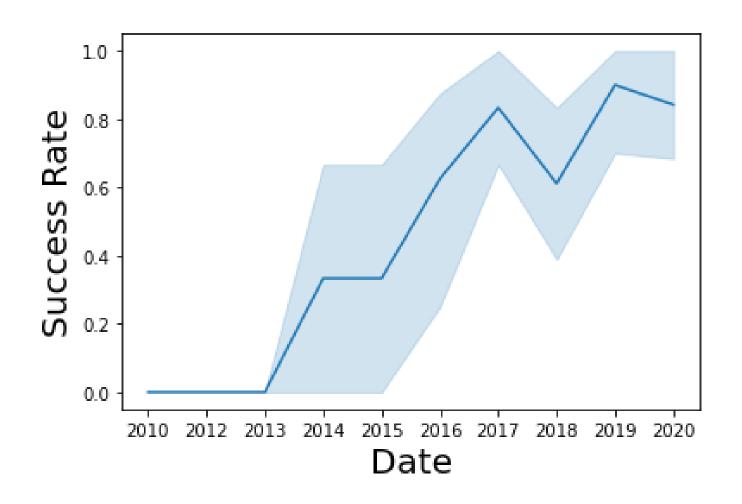
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

## PAYLOAD VS ORBIT



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

## LAUNCH SUCCESS YEARLY TREND



you can observe that the success rate since 2013 kept increasing till 2020



# EDA WITH SQL

## UNIQUE LAUNCH SITE

## **SQL QUERY**

SELECT DISTINCT LAUNCH\_SITE as "Launch\_Sites" FROM SPACEX;



## Launch\_Sites

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

#### **QUERY EXPLAINATION**

Using the word **DISTINCT** in the query means that it will only show Unique values in the **Launch\_Site** column from **SpaceX** 

## LAUNCH SITE NAMES BEGIN WITH CCA

## **SQL QUERY**

SELECT \* FROM SPACEX WHERE LAUNCH\_SITE LIKE 'CCA%' LIMIT 5;



#### **QUERY EXPLAINATION**

Using the word **TOP 5** in the query means that it will only show 5 records from **SpaceX** and **LIKE** keyword has a wild card with the words **'CCA%'** the percentage in the end suggests that the Launch\_Site name must start with KSC.

DATE	time_utc_	booster_version	launch_site	payload	payload_masskg_	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

## TOTAL PAYLOAD MASS BY CUSTOMER NASA (CRS)

## **SQL QUERY**

SELECT SUM(PAYLOAD\_MASS\_\_KG\_) AS "Total Paylod Mass by NASA" FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';



Total Paylod Mass by NASA

45596

#### **QUERY EXPLAINATION**

Using the function **SUM** summates the total in the column **PAYLOAD\_MASS\_KG\_** 

The **WHERE** clause filters the dataset to only perform calculations on **Customer NASA (CRS)** 

# AVERAGE PAYLOAD MASS CARRIED BY BOSSTER VERSION TO V.1.1

## **SQL QUERY**

SELECT AVG (PAYLOAD\_MASS\_\_KG\_) AS "Average Payload Mass Carried" FROM SPACEX WHERE BOOSTER\_VERSION = 'F9 v1.1';



## Average Payload Mass Carried

2928

#### **QUERY EXPLAINATION**

Using the function **AVG** works out the average in the column **PAYLOAD\_MASS\_\_KG\_** 

The **WHERE** clause filters the dataset to only perform calculations on **Booster\_version F9 v1.1** 

# THE DATE WHERE THE SUCCESSFUL LANDING OUTCOME IN GROUND PAD

## **SQL QUERY**

SELECT MIN(DATE) AS "The First Successful Landing Outcome in Ground Pad" FROM SPACEX WHERE LANDING\_OUTCOME = 'Success (ground pad)';



The First Successful Landing Outcome in Ground Pad

2015-12-22

#### **QUERY EXPLAINATION**

Using the function *MIN* works out the minimum date in the column *Date* 

The **WHERE** clause filters the dataset to only perform calculations on **Landing\_Outcome Success (ground pad)** 

# SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

## **SQL QUERY**

SELECT PAYLOAD FROM SPACEX WHERE LANDING\_OUTCOME = 'Success (drone ship)' AND PAYLOAD\_MASS\_KG\_ BETWEEN 4000 AND 6000;



#### payload

JCSAT-14

JCSAT-16

SES-10

SES-11 / EchoStar 105

#### **QUERY EXPLAINATION**

Selecting only **Booster\_Version** 

The **WHERE** clause filters the dataset to **Landing\_Outcome** = **Success (drone ship)** 

The **AND** clause specifies additional filter conditions **Payload\_MASS\_KG\_ BETWEEN 4000 AND 6000** 

# TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

## **SQL QUERY**

SELECT COUNT(LANDING\_OUTCOME) AS "Number Of Success" FROM SPACEX WHERE LANDING\_OUTCOME LIKE 'Success%';



Number Of Success

61

SELECT COUNT(LANDING\_OUTCOME) AS "Number Of Failure" FROM SPACEX WHERE LANDING\_OUTCOME LIKE 'Failure%';



Number Of Failure

10

## BOOSTERS CARRIED MAXIMUM PAYLOAD

## **SQL QUERY**

SELECT DISTINCT BOOSTER\_VERSION AS "Booster Versions" FROM SPACEX
WHERE PAYLOAD\_MASS\_\_KG\_ = (SELECT MAX(PAYLOAD\_MASS\_\_KG\_) FROM SPACEX);



#### **QUERY EXPLAINATION**

Using the word **DISTINCT** in the query means that it will only show Unique values in the **Booster\_Version** column from **SpaceX** 

#### **Booster Versions**

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

# LIST THE FAILED LANDING OUTCOMES IN DRONE SHIP

## **SQL QUERY**

SELECT BOOSTER\_VERSION,
LAUNCH\_SITE FROM SPACEX WHERE
LANDING\_OUTCOME LIKE 'Failure (drone ship)' AND DATE LIKE '2015-%'



launch_site	booster_version
CCAFS LC-40	F9 v1.1 B1012
CCAFS LC-40	F9 v1.1 B1015

## RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

## **SQL QUERY**

SELECT LANDING\_OUTCOME, COUNT(LANDING\_OUTCOME)
AS "Total Count" FROM SPACEX WHERE DATE BETWEEN '201006-04' AND '2017-03-20'
GROUP BY LANDING\_OUTCOME
ORDER BY COUNT(LANDING OUTCOME) DESC;

#### **QUERY EXPLAINATION**

Function *COUNT* counts records in column *WHERE* filters data

landing_outcome	Total Count
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



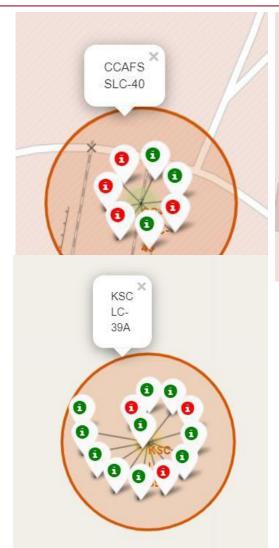
# INTERACTIVE MAP WITH FOLIUM

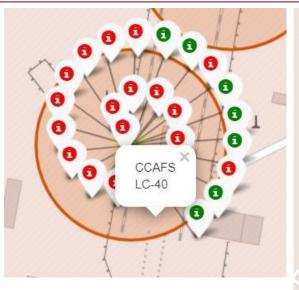
## MARK ALL LAUNCH SITES ON MAP



We can see that the SpaceX launch sites are in the United States of America coasts.
Florida and California

# MARK THE SUCCESS/FAILED LAUNCHES FOR EACH SITE ON THE MAP







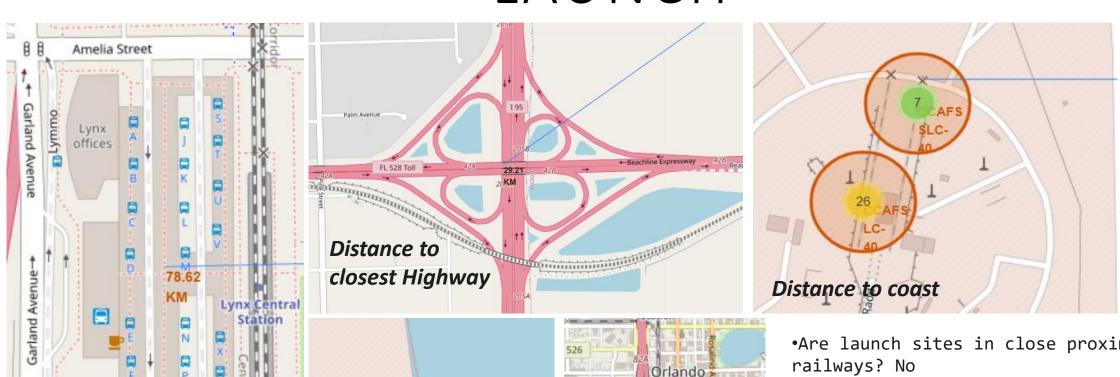


Green Marker shows successful Launches and Red **Marker** shows Failures



California Launch Site

# CALCULATE DISTANCE BEETWEEN A LAUNCH



0.90 Distance to Coastline

Distance to

Railway Station

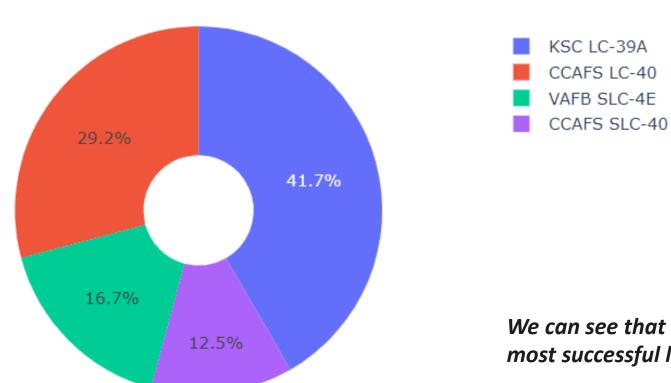
- Distance to City
- •Are launch sites in close proximity to
- •Are launch sites in close proximity to highways? No
- •Are launch sites in close proximity to coastline? Yes
- •Do launch sites keep certain distance away from cities? Yes



# DASHBOARD WITH PLOTLY DASH

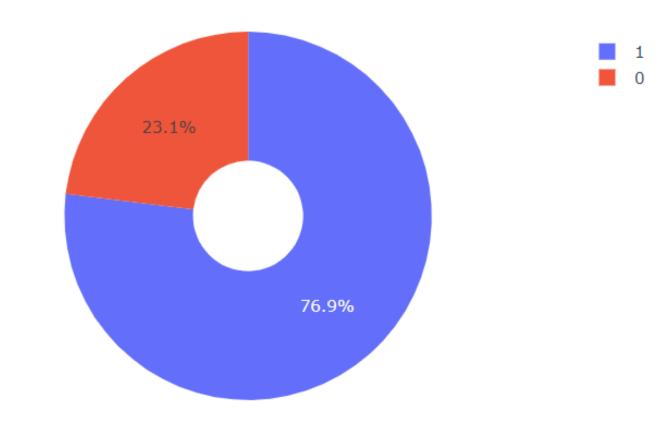
# DASHBOARD – PIE CHART THE SUCCESS PERCENTAGE

Total Success Launches By all sites



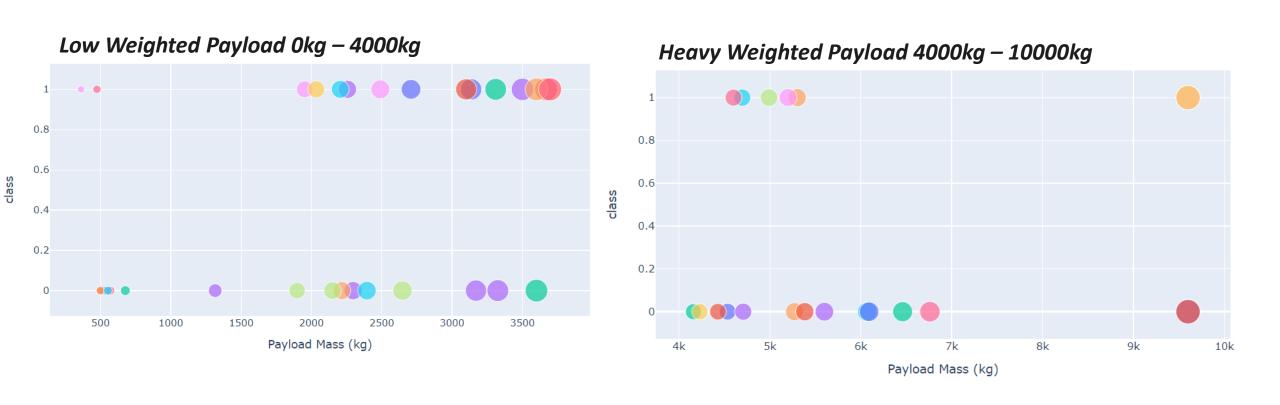
We can see that KSC LC-39A had the most successful launches from all the sites

# DASHBOARD — PIE CHART FOR THE LAUNCH SITE WIH HIGHEST LAUNCH SUCCESS RATIO



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

## DASHBOARD - PAYLOAD VC LAUNCH OUTCOME SCATTER PLOT



We can see the success rates for low weighted payloads is higher than the heavy weighted payloads



# PREDICTION ANALYSIS (CLASSIFICATION)

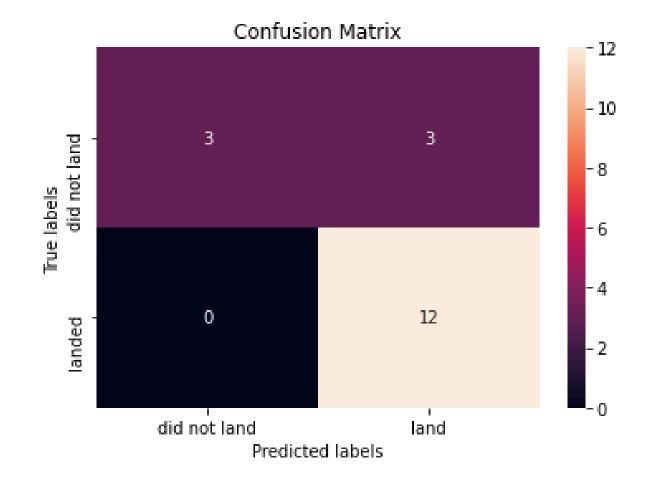
# CLASIFICATION ACCURACY USING TRAINING DATA

```
algorithms = {'KNN':knn_cv.best_score_,'Tree':tree_cv.best_score_,'LogisticRegression':logreg_cv.best_score_}
   bestalgorithm = max(algorithms, key=algorithms.get)
   print('Best Algorithm is', bestalgorithm, 'with a score of', algorithms[bestalgorithm])
   if bestalgorithm == 'Tree':
      print('Best Params is :',tree_cv.best_params_)
   if bestalgorithm == 'KNN':
      print('Best Params is :',knn_cv.best_params_)
   if bestalgorithm == 'LogisticRegression':
      print('Best Params is :',logreg_cv.best_params_)
Best Algorithm is Tree with a score of 0.8767857142857143
Best Params is : {'criterion': 'entropy', 'max depth': 8, 'max features': 'auto', 'min sampl
es leaf': 2, 'min samples split': 2, 'splitter': 'random'}
```

## CONFUSION MATRIX FOR THE TREE

# Confusion Matrix for the Tree

Examining the confusion matrix, we see that Tree can distinguish between the different classes. We see that the major problem is false positives.





# CONCLUSION

LOOKING AHEAD

## CONCLUSION



- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time in years they will
- eventually perfect the launches

  We can see that KSC LC-39A had the most
- successful launches from all the sites
   Orbit GEO,HEO,SSO,ES-L1 has the best Success
   Rate

### APPENDIX

- Haversine formula
- ADGGoogleMaps Module (not used created)
- Module sqlserver (ADGSQLSERVER)
- PythonAnywhere 24/7 dashboard

# ADD GOOGLEMAPS MODULE

#### Introduction

This is a python class I designed to make my life easier and others when getting coordinates from addresses and producing a Folium Python Map in a Jypyter Notebook. This class processing is powered by the Google Geocoding API.

#### **Prerequisites**

- •Sign up for a Google Geocoding API Set up a Google API Key
- •Head to library under APIs & Services and add Geocoding API

#### Requirements

Google API Secret (API Key)

#### **Python Usage Documentation**

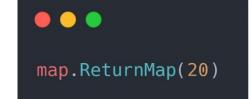
Getting Cords

Simple arguments to follow by the object \_\_init\_ (self,api\_key,address) - Look at following usage for guidance



This returns a list with your coordinates of that address example ['longitude','latitude']





- •Returning Folium Python Map in Jypyter Notebook
- assuming you declared mapclass in 1.
  .ReturnMap(size)

this function returns this Folium map in the Jypyter Notebook



## HEVERSINE FORMULA

#### Introduction

The haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes. Important in navigation, it is a special case of a more general formula in spherical trigonometry, the law of haversines, that relates the sides and angles of spherical triangles.

#### Usage

Why did I use this formula? First of all, I believe the Earth is round/elliptical. I am not a Flat Earth Believer! Jokes aside when doing Google research for integrating my ADGGoogleMaps API with a Python function to calculate the distance using two distinct sets of {longitudinal, latitudinal} list sets. Haversine was the trigonometric solution to solve my requirements above.

#### Formula

$$a = \sin^{2}(\frac{\Delta \varphi}{2}) + \cos \varphi 1 \cdot \cos \varphi 2 \cdot \sin^{2}(\frac{\Delta \lambda}{2})$$

$$c = 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

#### 

#### #Variables

d is the distance between the two points along a great circle of the sphere (see spherical distance), r is the radius of the sphere.

 $\phi 1, \ \phi 2$  are the latitude of point 1 and latitude of point 2 (in radians),

 $\lambda 1$ ,  $\lambda 2$  are the longitude of point 1 and longitude of point 2 (in radians).

# ADGSQL SERVER

#### Introduction

basically I just wanted an easier way to get my data into my python programming by coding my own module powered by ODBC to simplify my code

#### **Implementation**

Pull data from columns Extract Records Run Stored Procedures

• • • •

```
import sqlserver as ss

#(ip,portnumber,databasename,username,password)
db = ss.sqlserver('localhost','1433','CVs','','')

#(query,columnname)
db.GetRecordsOfColumn('select * from tblUsers','personid')
```

## PYTHON ANYWHERE

#### Introduction

I wanted to put my python website running 24/7 on the cloud so anyone can view it then I came across PythonAnywhere.

#### **Implementation**

We run Flask in /www/ on the docker Linux container We have two files flask\_app.py , wsgi.py

These are the files that run our website

#### **Pricing**

Free but we are restricted to hitting the renew button every 3 months and we cannot link the domain up to our own private domain. We only can run one instance of a website per month.

