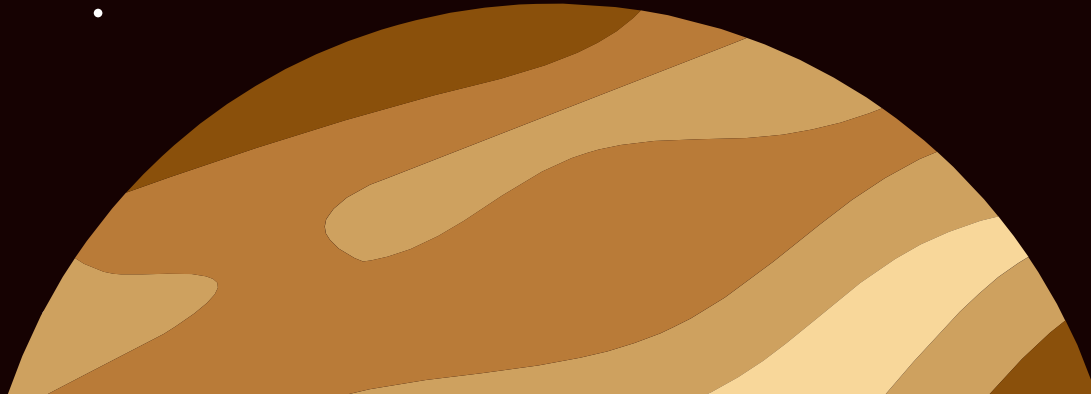


Winning Space Race with Data Science

Muhammad Nabil
Sep 2021



“The exploration of space will go ahead, whether we join in it or not, and it is one of the great adventures of all time, and no nation which expects to be the leader of other nations can expect to stay behind in the race for space.”

—John F. Kennedy



Table of contents

01 EXECUTIVE SUMMARY

02 INTRODUCTION

03 METHODOLOGY

04 RESULTS

05 CONCLUSION





01

EXECUTIVE SUMMARY

Executive Summary

Methodology

- 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)

Results

- Exploratory Data Analysis Results
- Interactive Analytics Demo in Screenshot
- Predictive Analysis Results





02

INTRODUCTION

Introduction



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.

Introduction

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.

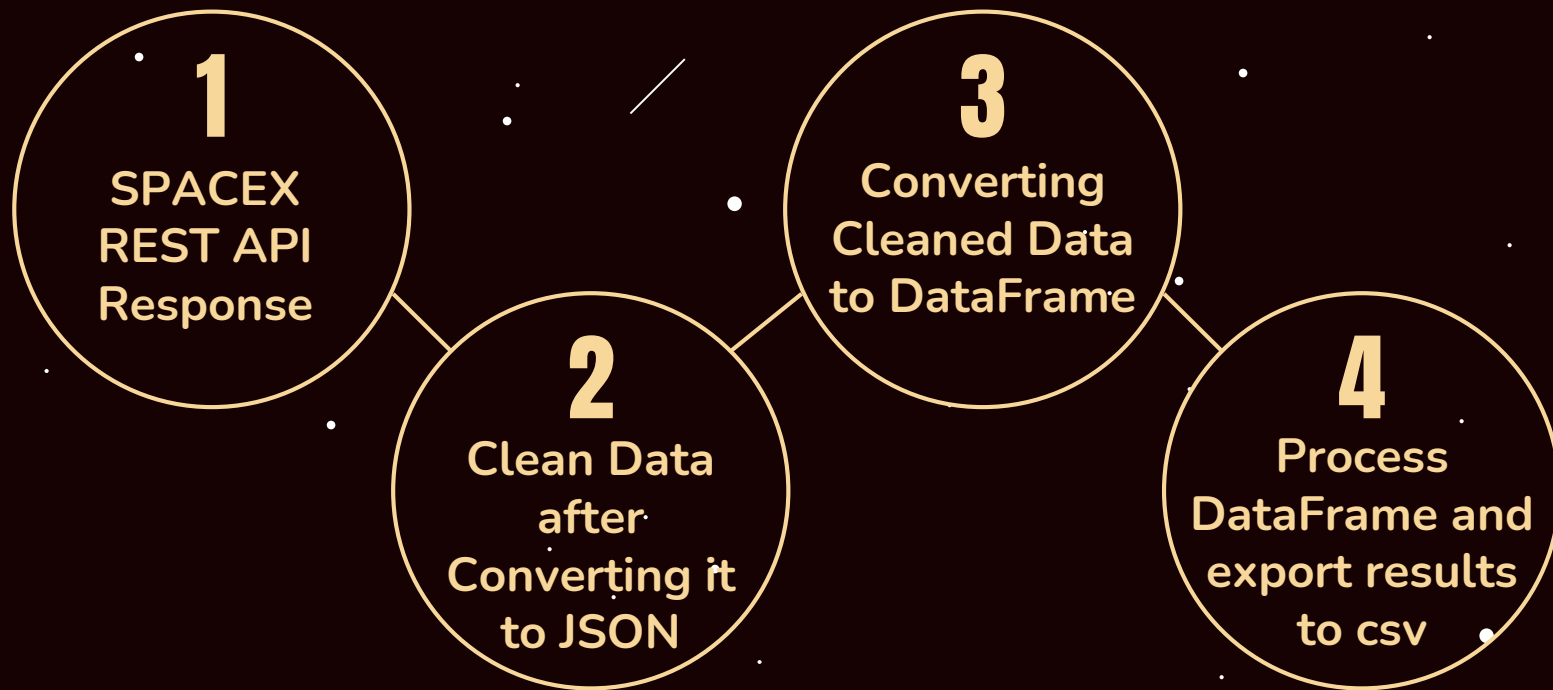




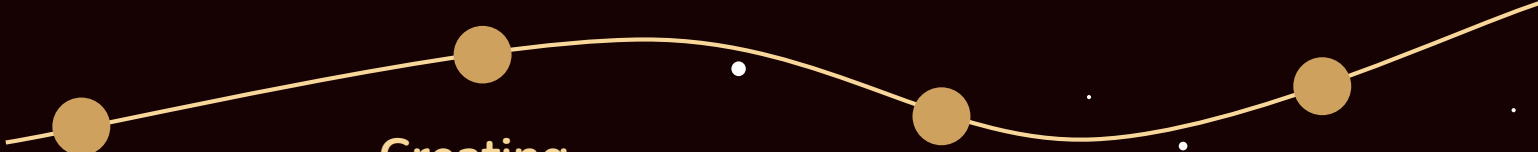
03

METHODOLOGY

Data Collection SpaceX API



Data Collection - Web Scrapping



Convert HTML
Response to
BeautifulSoup
Object

Creating
Dictionary from
all tables
(keys are
column names)

Converting
Dictionary to
DataFrame

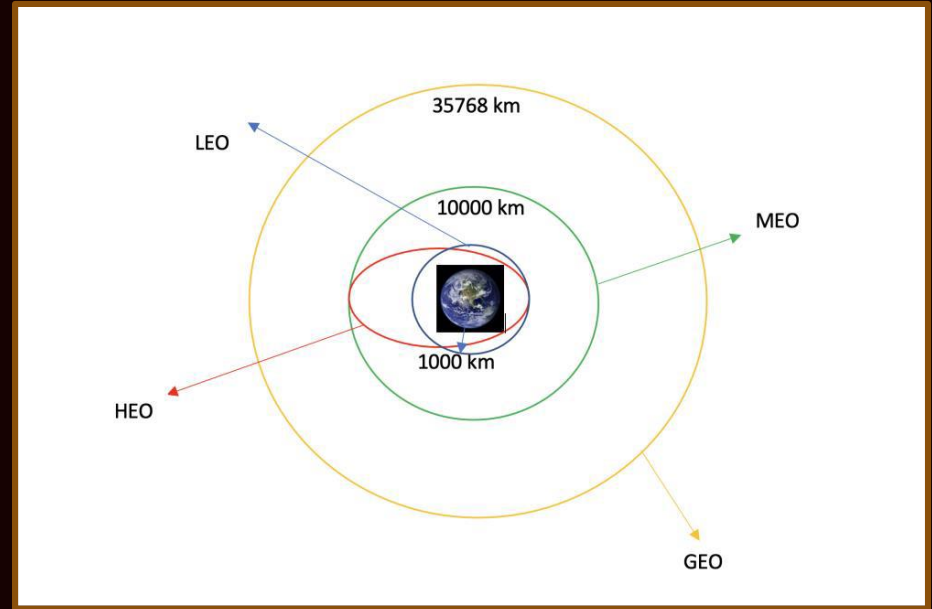
Process
DataFrame
and export
results to csv

[Open Notebook](#)

Data Wrangling

Process phases:

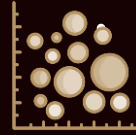
- Perform Exploratory Data Analysis EDA on dataset
- Calculate the number of launches at each site
 - Calculate the number and occurrence of mission outcome per orbit type
 - Export dataset as .CSV
- Calculate the number and occurrence of each orbit
 - Create a landing outcome label from Outcome column
 - Work out success rate for every landing in dataset



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

[Open Notebook](#)

EDA with Data Visualization



Scatter Graphs



Bar Graphs



Line Graphs

EDA with SQL

Queries Performed

- 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.

[Open Notebook](#)

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

Building an Interactive Dashboard with Dash

Used Python Anywhere to host the website live 24/7 so you can play around with the data and view the data

-The dashboard is built with Flask and Dash web framework.

Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions

-It shows the relationship between two variables.

-It is the best method to show you a non-linear pattern.

-The range of data flow, i.e. maximum and minimum value, can be determined.

-Observation and reading are straightforward.

Graphs

-Pie Chart showing the total launches by a certain site/all sites

-display relative proportions of multiple classes of data.

-size of the circle can be made proportional to the total quantity it represents.

[Open Notebook](#)

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

Predictive Analysis (Classification)

IMPROVING MODEL

- Feature Engineering
- Algorithm Tuning

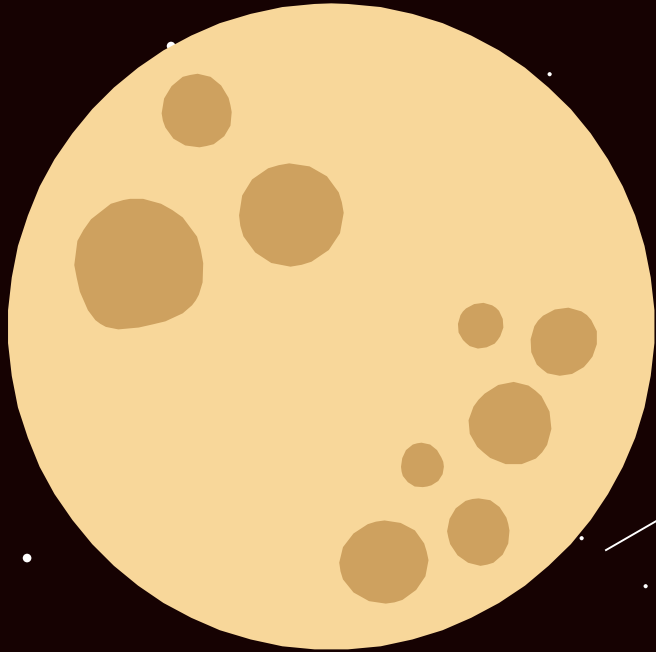
FINDING THE BEST PERFORMING CLASSIFICATION MODEL

- 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.



04

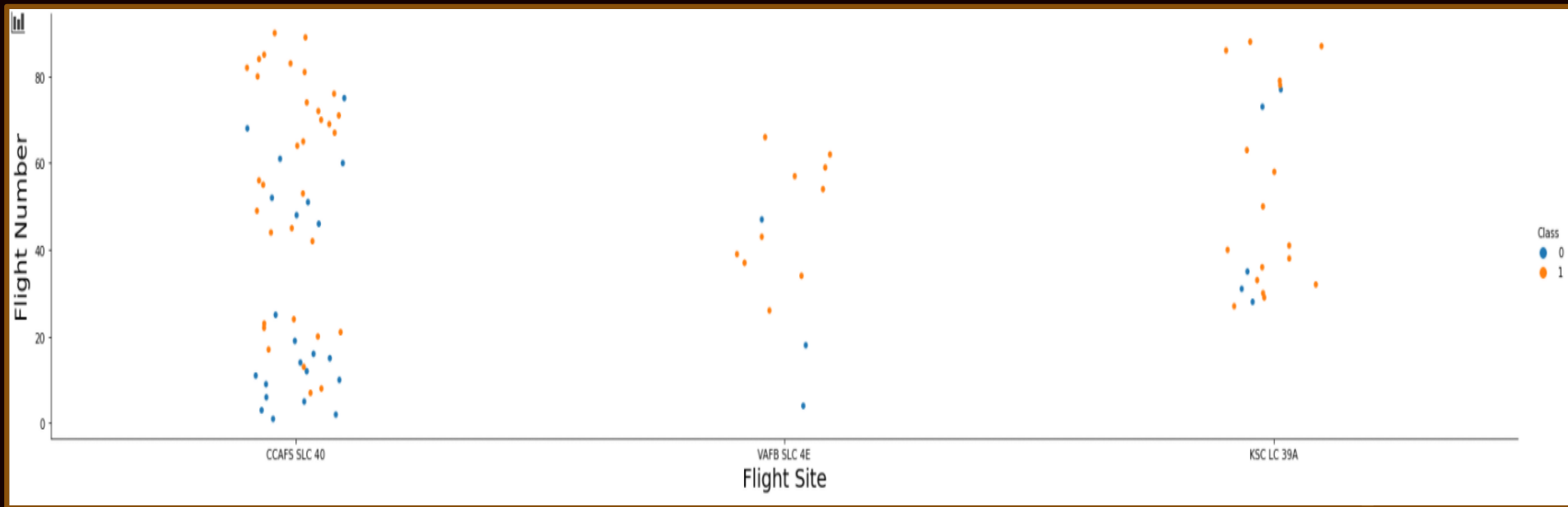
RESULTS!



Results!

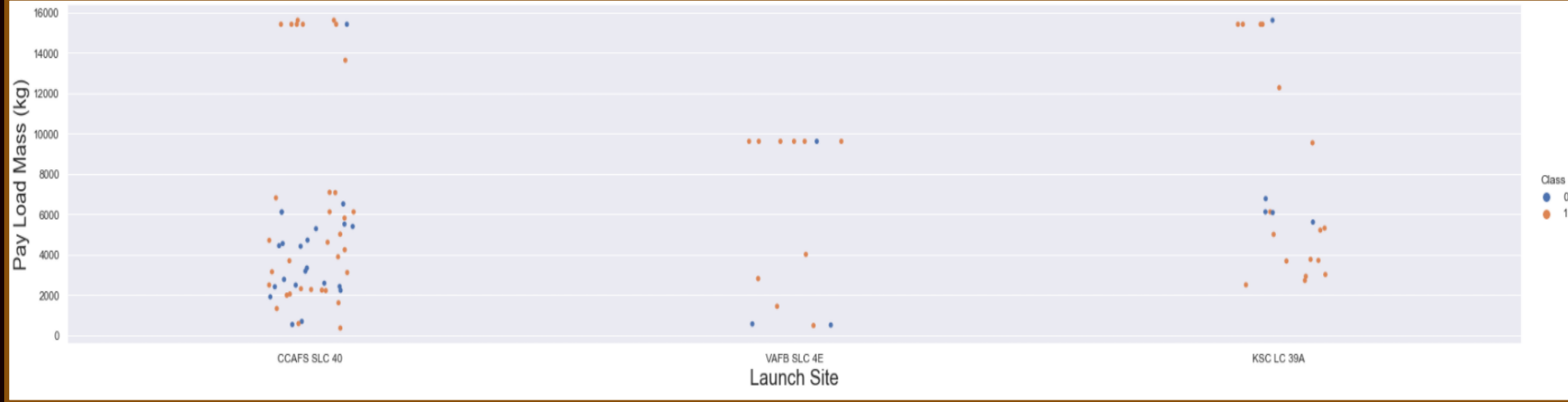
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

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

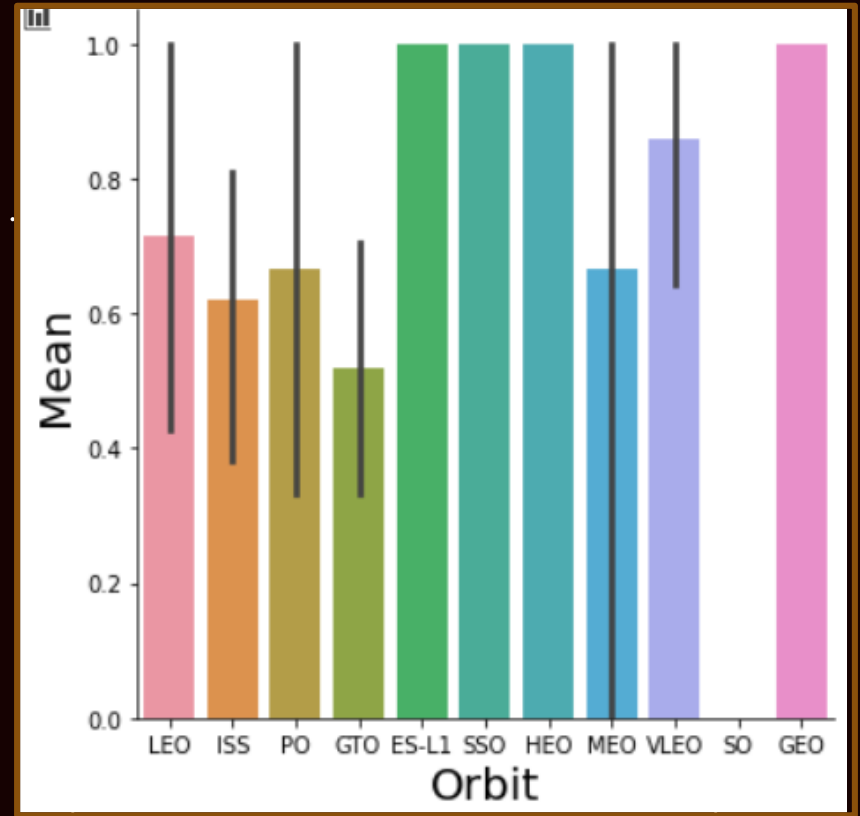


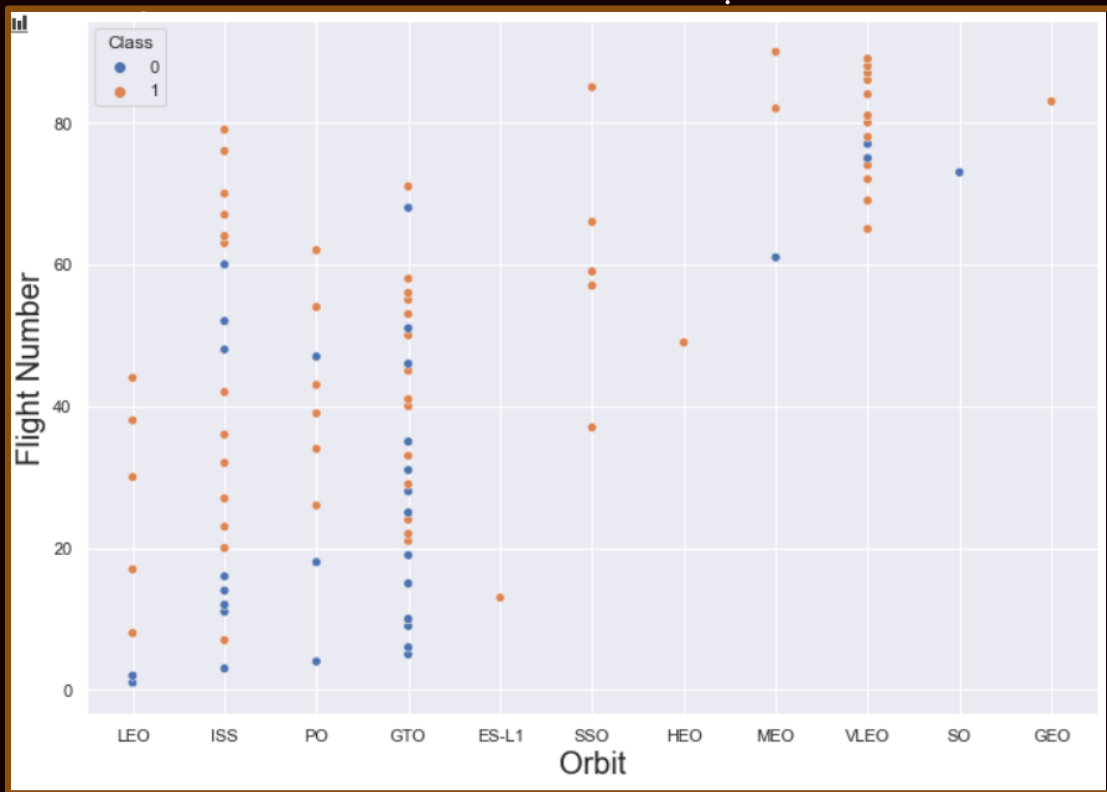
The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket.

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

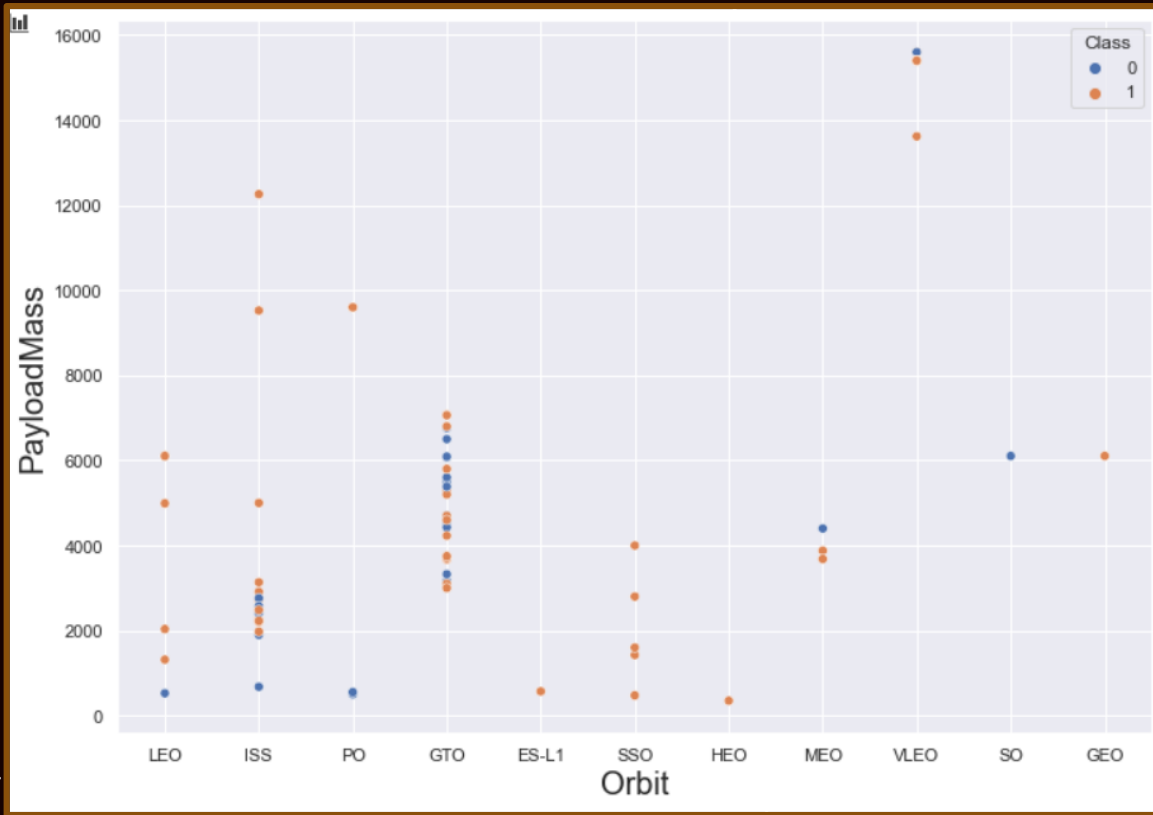
Orbit GEO,HEO,SSO,ES-L1
has the best Success Rate





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.

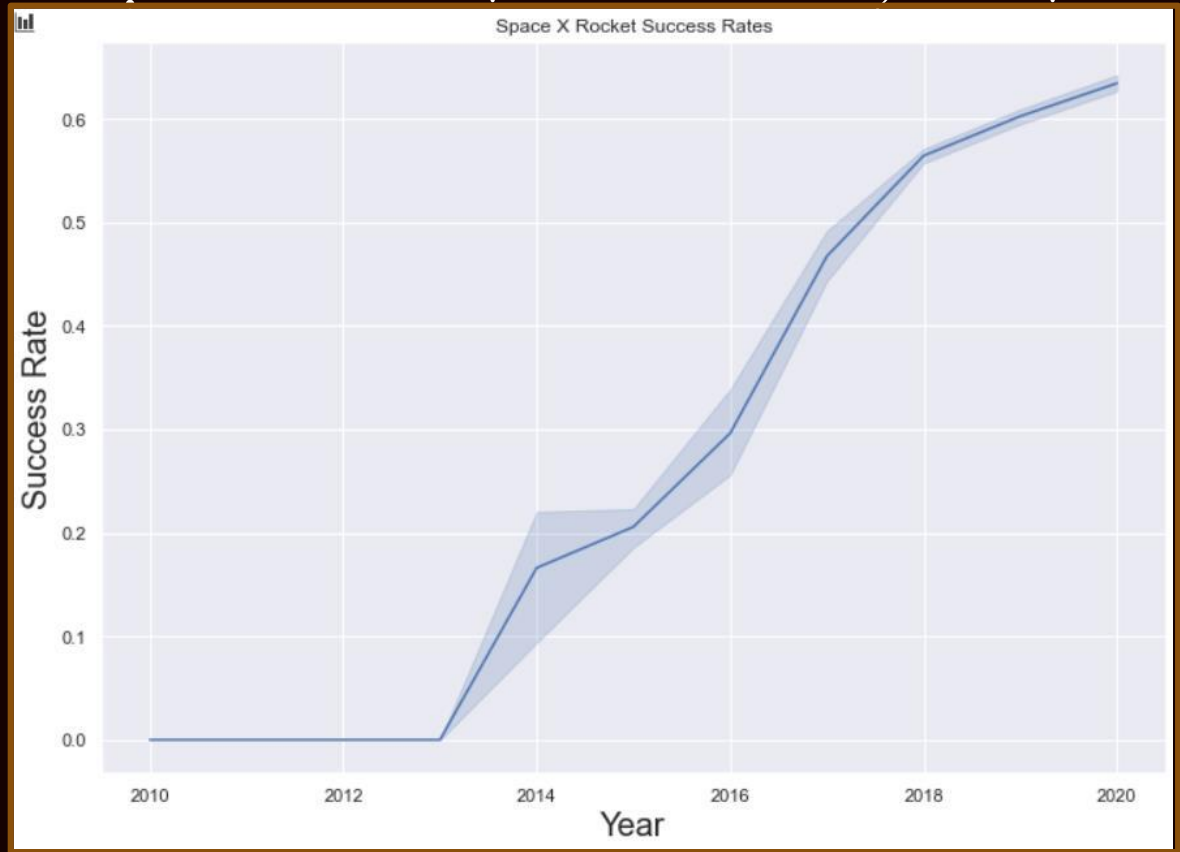


Payload Mass vs Orbit Type

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

Launch Success Rate Yearly Trend

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



Unique Launch Sites Names

```
select distinct Launch_Site from tblSpaceX
```

Unique Launch Sites
CCAFS LC-40
CCAFS SLC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Using the word ***distinct*** in the query means that it will only show Unique values in the ***Launch_Site*** column from ***tblSpaceX***

Launch Site Names Begin with 'KSC'

Select top 5 * from tblSpaceX where Launch_Site like 'KSC%'

	Date	Time_UTC	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	19-02-2017	2021-07-02 14:39:00.0000000	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
1	16-03-2017	2021-07-02 06:00:00.0000000	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2	30-03-2017	2021-07-02 22:27:00.0000000	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
3	01-05-2017	2021-07-02 11:15:00.0000000	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
4	15-05-2017	2021-07-02 23:21:00.0000000	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Using the word **top 5** in the query means that it will only show 5 records from **tblSpaceX** and **like** keyword has a wild card with the words **'KSC%'** the percentage in the end suggests that the Launch_Site name must start with KSC.

Total Payload Mass by Customer NASA (CRS)

```
select sum(PAYLOAD_MASS_KG_) as TotalPayloadMass from tblSpaceX  
where Customer = 'NASA (CRS)'
```

Total Payload Mass	
0	45596

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 booster version F9 v1.1

```
select avg(PAYLOAD_MASS_KG_) as AveragePayloadMass from tblSpaceX  
where Booster_Version = 'F9v1.1'
```

Average Payload Mass	
0	2928

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

First Successful Ground Landing Date

```
select min(Date) as  
    DateWhichFirstSuccessfulLandingOutcomeInDroneShipWasAcheived  
from tblSpaceX where  
    Landing_Outcome = "Success(drone ship)"
```

Date which first Successful landing outcome in drone ship was acheived.	
0	06-05-2016

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 (drone ship)*

Successful drone ship landing with payload between 4000 and 6000

Select `Booster_Version` from `tblSpaceX` where
`Landing_Outcome = 'Success (ground pad)'` and
`Payload_MASS_KG_ > 4000` and
`Payload_MASS_KG_ < 6000`

0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

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_ > 4000 and
Payload_MASS_KG_ < 6000

Total Number of Successful and Failure Mission Outcomes

```
select(select count(Mission_Outcome) from tblSpaceX where  
Mission_Outcome like '%Success%') as Successful_Mission_Outcomes,  
(select count(Mission_Outcome) from tblSpaceX where Mission_Outcome  
like '%Failure%') as Failure_Mission_Outcomes
```

a much harder query I must say, we used subqueries here to produce the results. The **LIKE** **'%foo%'** wildcard shows that in the record the **foo** phrase is in any part of the string in the records for example.

Successful_Mission_Outcomes	Failure_Mission_Outcomes
0	100
	1

Boosters Carried Maximum Payload

```
select distinct Booster_Version, max(payload_mass_kg_) as  
MaximumPayloadMass from tblSpaceX groupby Booster_Version  
orderby MaximumPayloadMass desc
```

	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0

97 rows x 2 columns

Using the word ***distinct*** in the query means that it will only show Unique values in the ***Booster_Version*** column from ***tblSpaceX***

group by puts the list in order set to a certain condition.

desc means its arranging the dataset into descending order

2017 Launch Records

```
select datename(month, dateadd(month, month(convert(date, Date, 105)), 0) -1)
      as Month, Booster_Version, Launch_Site, Landing_Outcome
      from tblSpaceX
where (Landing_Outcome like N'%Success%') and (year(convert(date, Date,
      105)) = '2017')
```

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
March	F9 FT B1021.2	KSC LC-39A	Success (drone ship)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1029.2	KSC LC-39A	Success (drone ship)
June	F9 FT B1036.1	VAFB SLC-4E	Success (drone ship)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
August	F9 FT B1038.1	VAFB SLC-4E	Success (drone ship)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
October	F9 B4 B1041.1	VAFB SLC-4E	Success (drone ship)
October	F9 FT B1031.2	KSC LC-39A	Success (drone ship)
October	F9 B4 B1042.1	KSC LC-39A	Success (drone ship)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

a much more complex query as I had my **Date** fields in SQL Server stored as **nvarchar** the **month** function returns name month. The function **convert** converts **nvarchar** to **Date**.

where clause filters **Year** to be 2017

Rank success count between 2010-06-04 and 2017-03-20

```
select count(Landing_Outcome) from tblSpaceX where (Landing_Outcome  
like '%Success%')  
and (Date > '04-06-2010')  
and (Date < '20-03-2017')
```

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

0

34

Function **count** counts records in column
where filters data

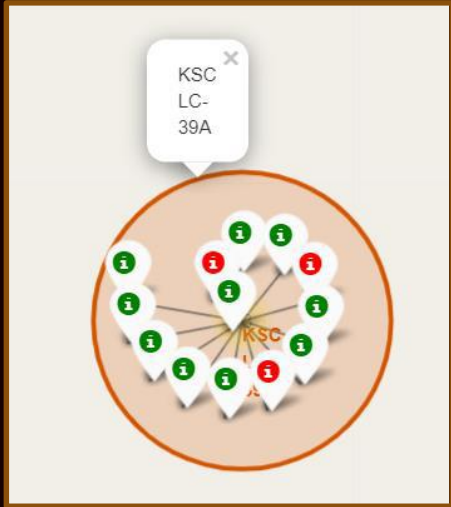
All launch sites global map markers

• VAFB
SLC-
4E

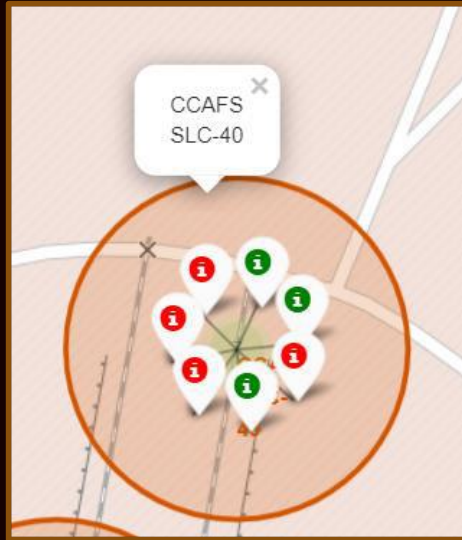
• SSCAFS
SCC-
38A

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

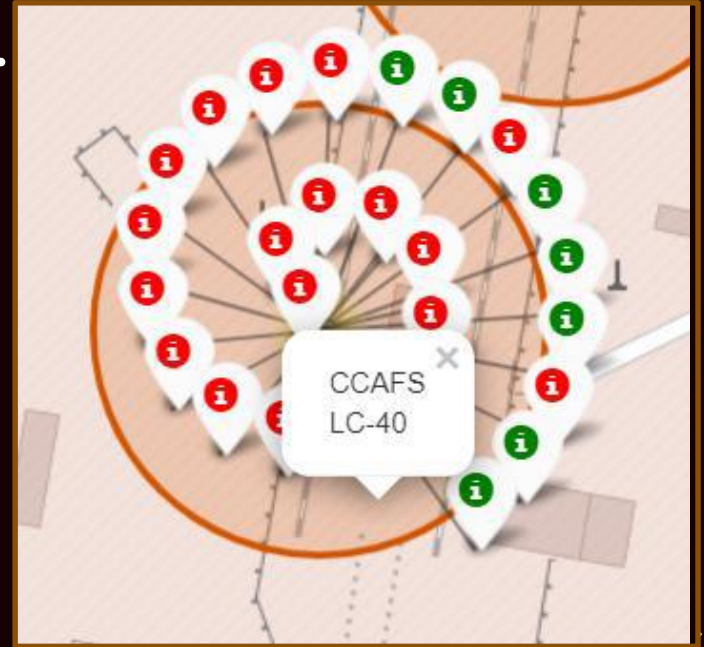
Color Labelled Markers



Florida Launch Sites



Florida Launch Sites



Florida Launch Sites

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

Color Labelled Markers



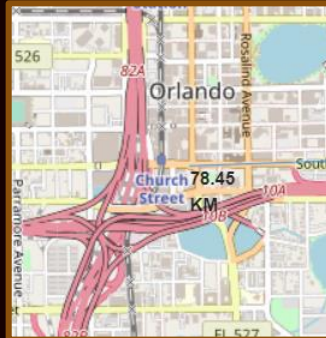
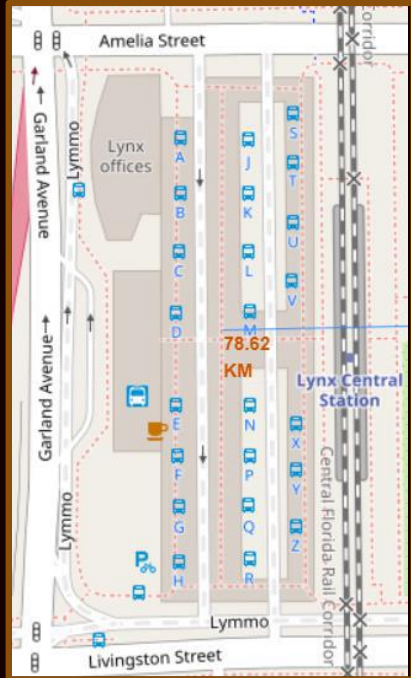
California Launch Site



Florida Launch Sites

Green Marker shows successful Launches and Red Markers shows Failures

Working out Launch Sites distance to landmarks to find trends with Haversine formula using CCAFS-SLC-40 as a reference

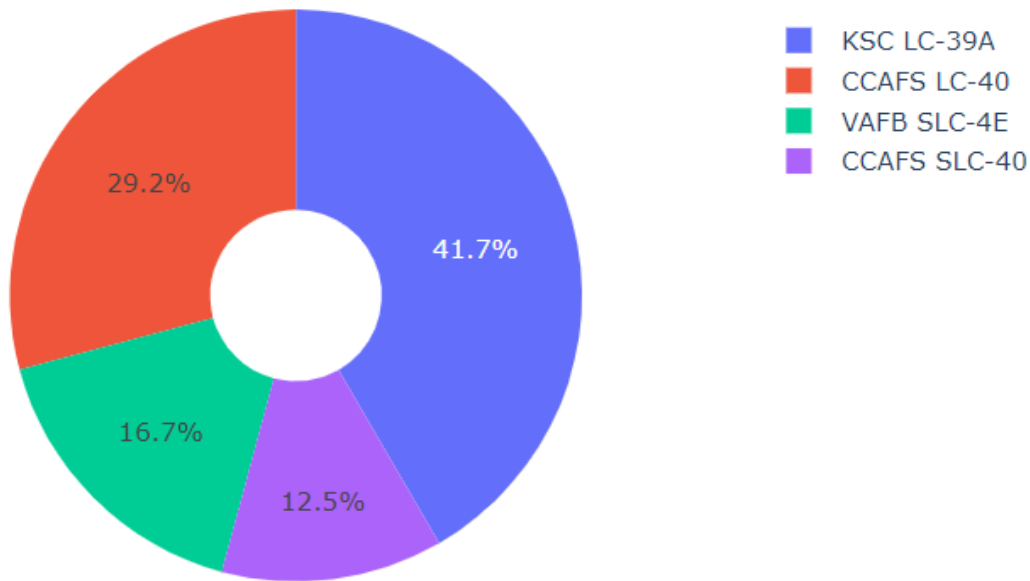


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes



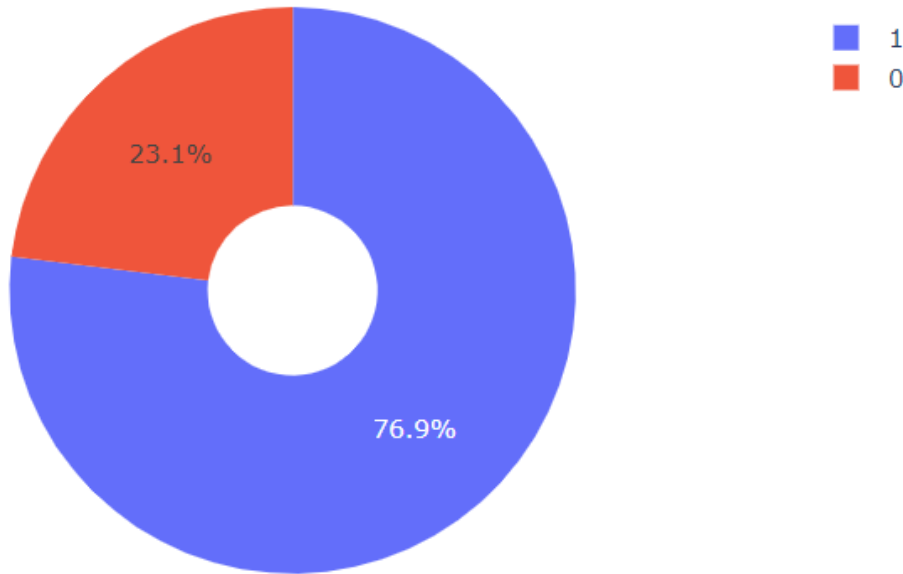
DASHBOARD-Pie chart showing the success percentage achieved by each launch site

Total Success Launches By all sites



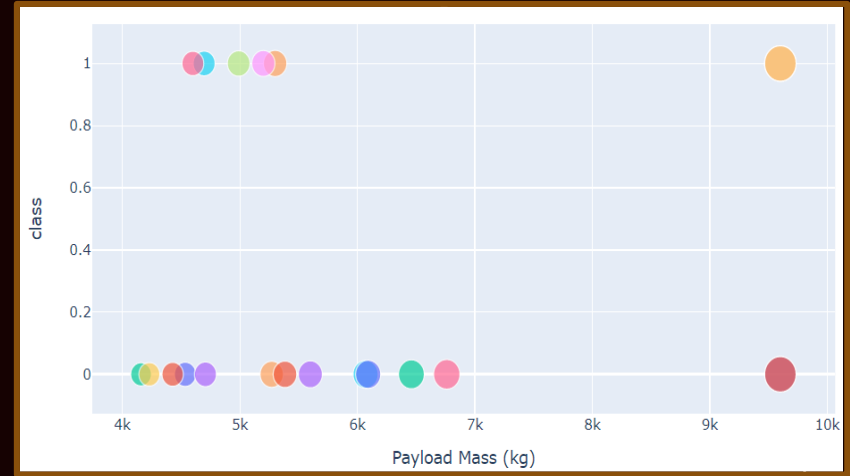
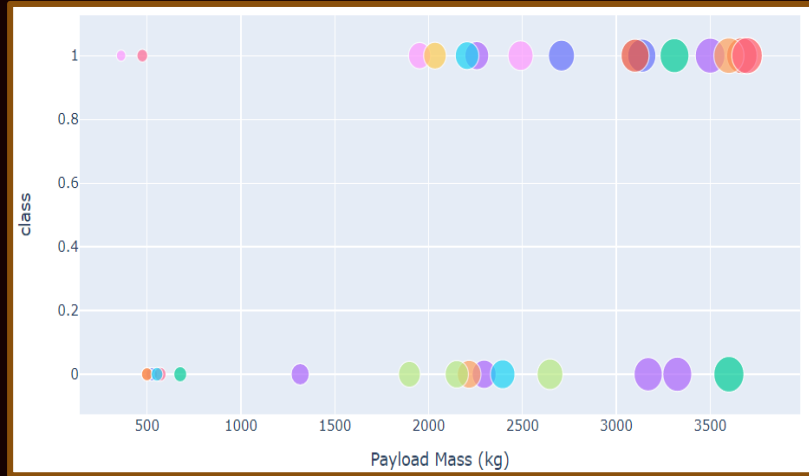
We can see that KSC LC-39A had the most successful launches from all the sites

DASHBOARD-Pie chart for the launch site with highest launch success ratio



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

DASHBOARD-Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



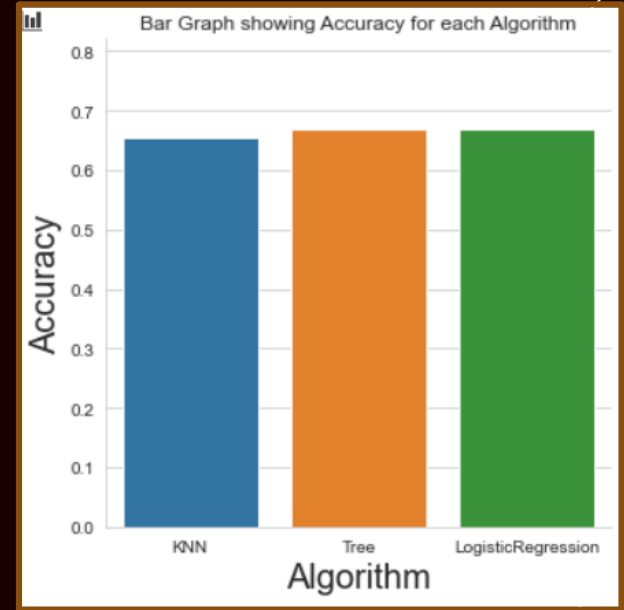
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Classification Accuracy using training data

	Accuracy	Algorithm
0	0.653571	KNN
1	0.667857	Tree
2	0.667857	LogisticRegression

The tree algorithm wins!!

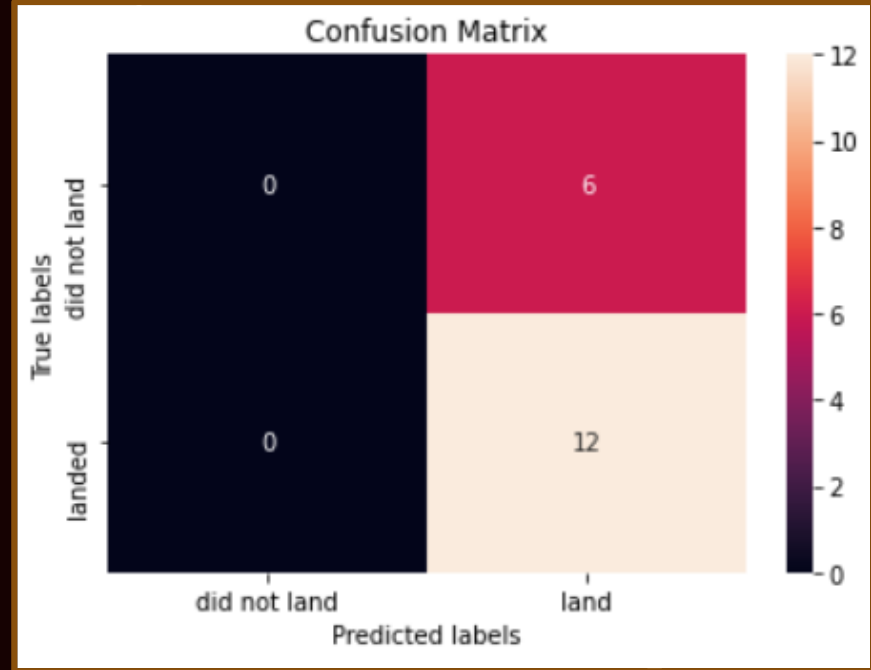
After selecting the best hyperparameters for the decision tree classifier using the validation data, we achieved 83.33% accuracy on the test data.



As you can see our accuracy is extremely close but we do have a winner its down to decimal places!

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.





05

CONCLUSION

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

