IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Likhitha VK
26-04-2025

Outline

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

Executive Summary

- **Summary of methodologies**

  Data collection.

  Data wrangling.

  EDA with data visualization and sql.

  Building an interactive map with folium.

  Building a dashboard with plotly dash.

  Predictive analysis (correlation).

- **Summary of all results**

  Exploratory data analysis results.

  Interactive analytics screenshots.

  Predictive analysis results.

Introduction

- **Problems you want to find answers**

We predicted if the Falcon9 first stage will land successfully. SpaceX advertises Falcon9 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 the cost of launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Project background and context**

what influences if the rocket will land successfully?

The effect each relationship with certain rocket variables will impact determining the success rate of a successful landing.

What conditions does SpaceX have to achieve to get the best results and ensure the best results and ensure the best rocket success landing rate.

Section 1

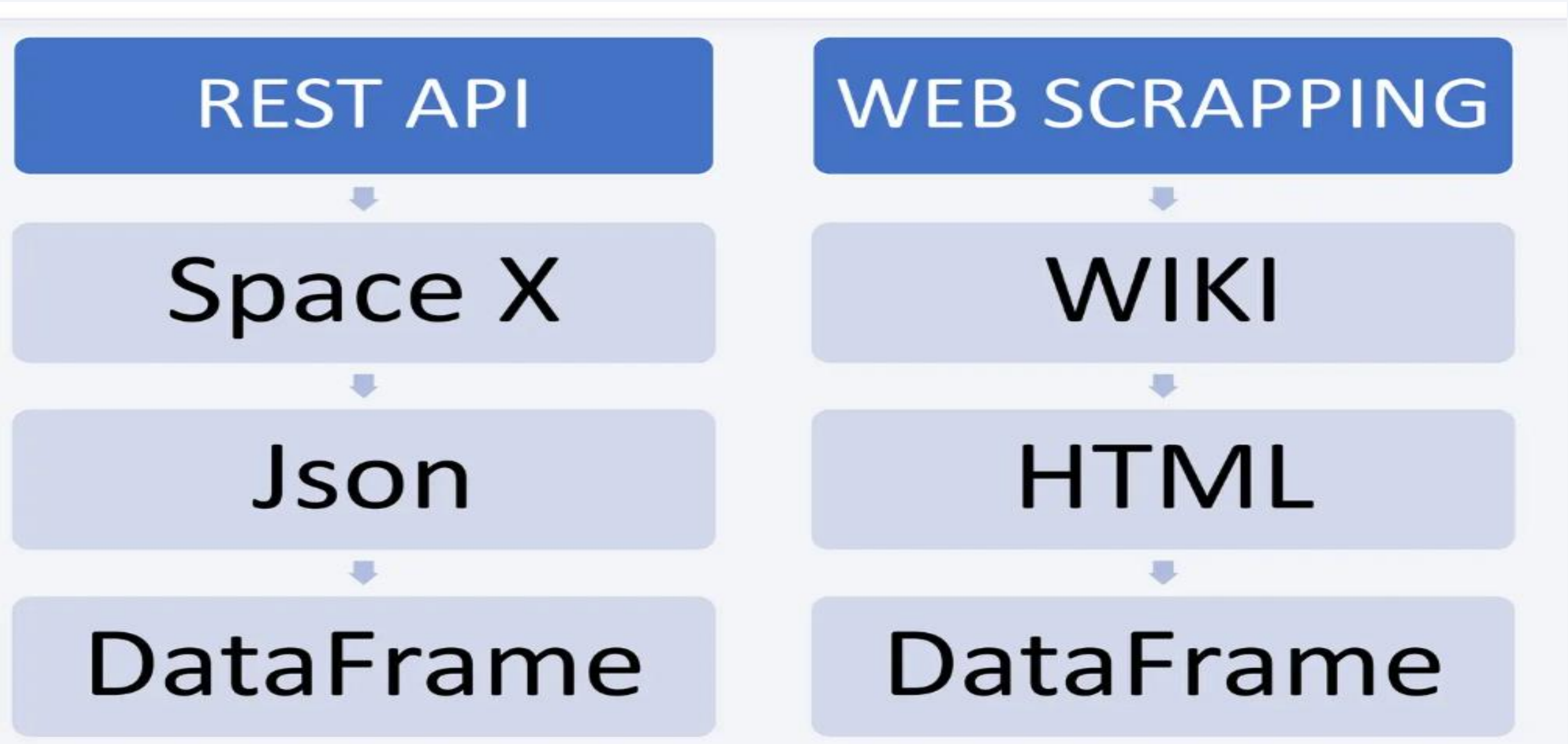
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX rest api
  - Web scraping
- Perform data wrangling
  - One hot encoding data fields for ML and dropping irrelevant columns.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection



Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

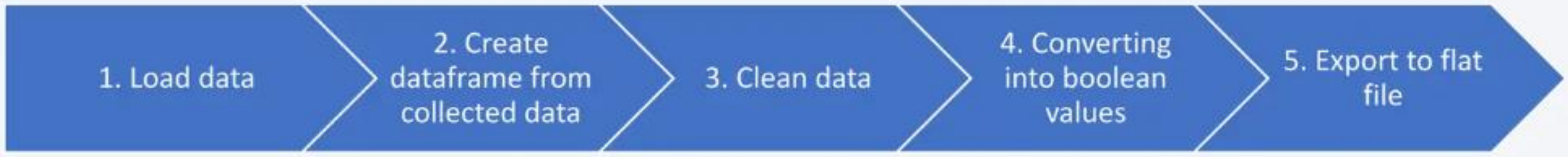
Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

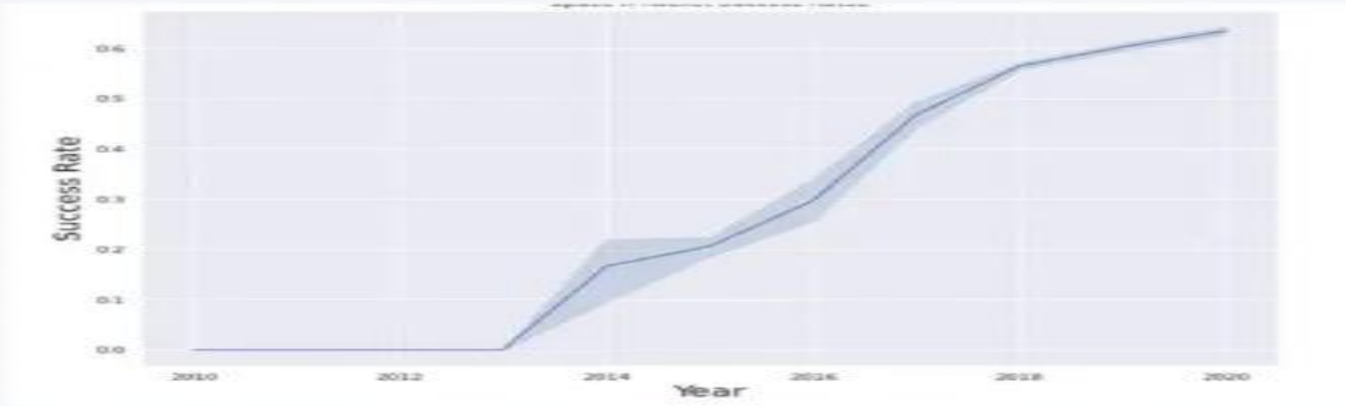
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

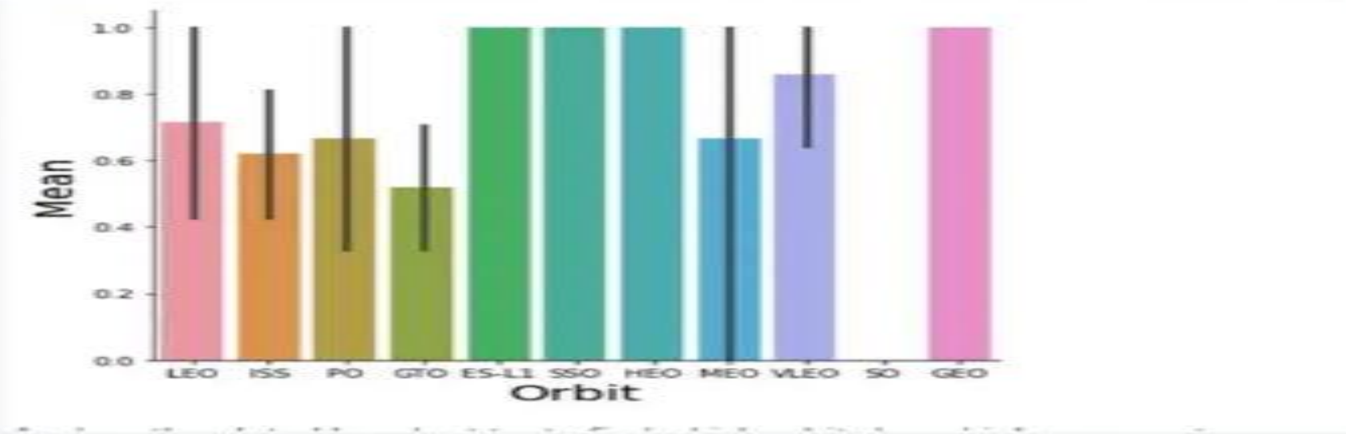
Data Wrangling

- After collecting the data we check the missing data and data types and do the following to clean the data : Replace the missing data with one- Using mean or so. Change data type of the data. Represent categorical data using integer or float dummy numbers –one hot encoding



EDA with Data Visualization





EDA with SQL

- 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
- Displaying average payload mass carried by booster version F9 v1.1 listing the date whre the successful landing outcomes in drone ship was achived.
- 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 records which will display the month names. Successful landing_outcomes in ground pad booster vaersions , launch_site for the months of the year 2017.

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance: - Are launch sites near railways, highways and coastlines. - Do launch sites keep certain distance away from cities.

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version
- The link to the notebook is

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.
- The link to the notebook is

Results

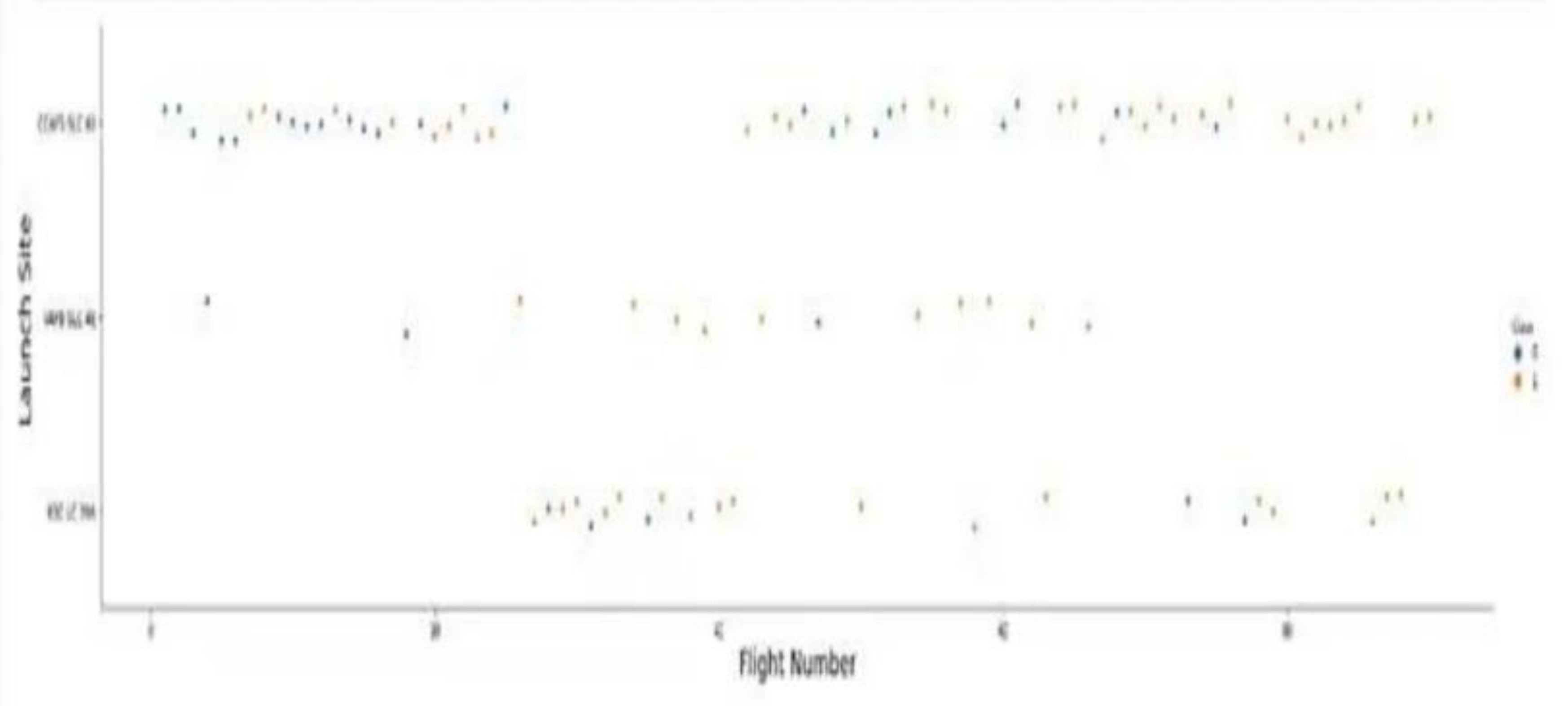
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

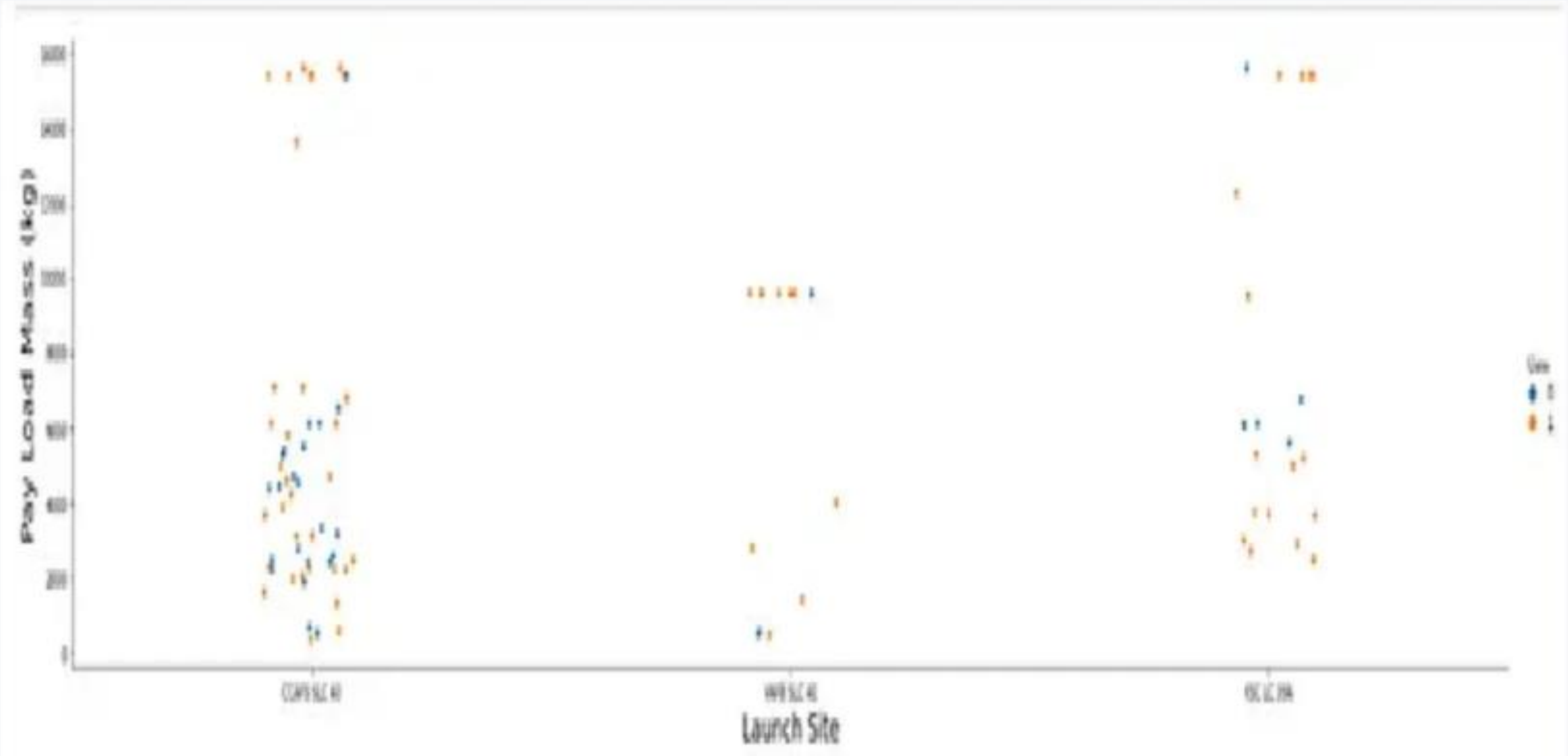
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



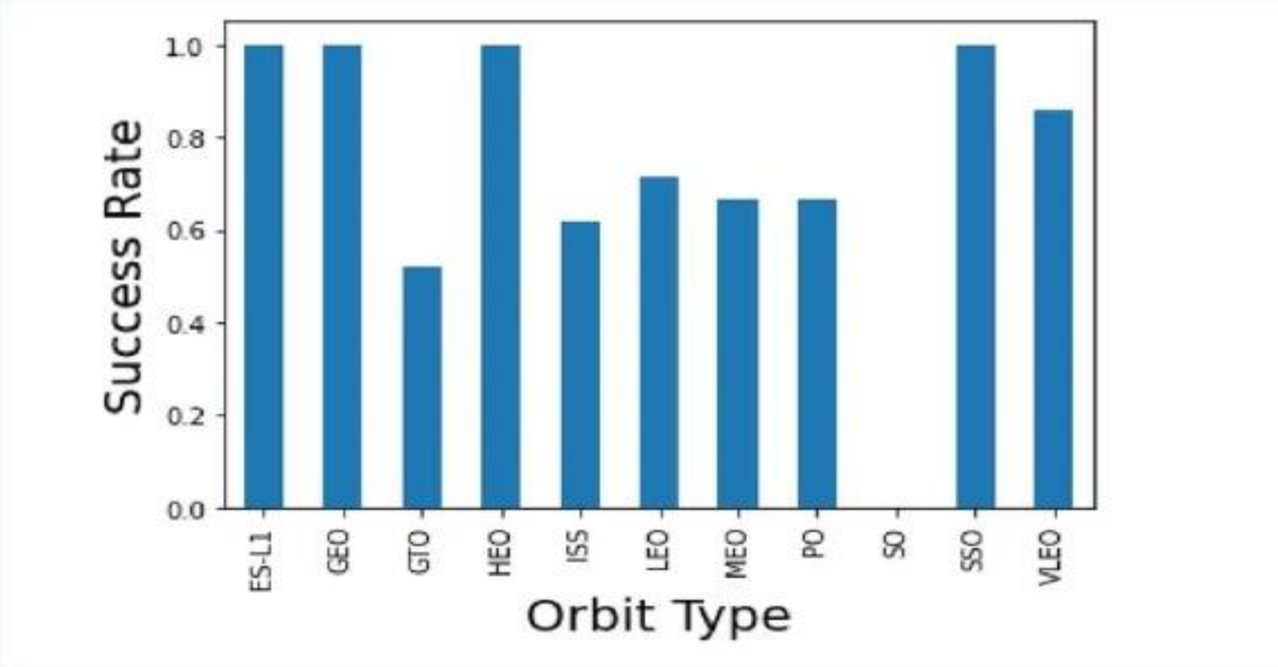
Payload vs. Launch Site

- The greater the payload_mass for launch site CCAFS SLC 40 the higher the success rate for the rocket



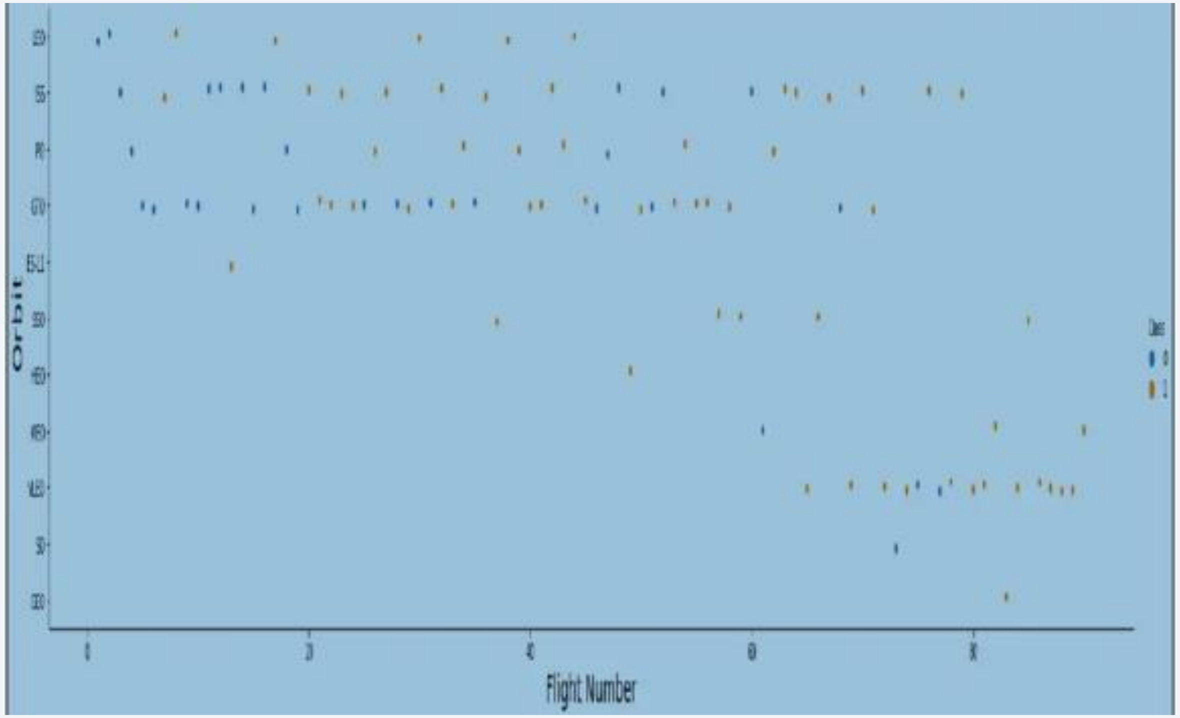
Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.



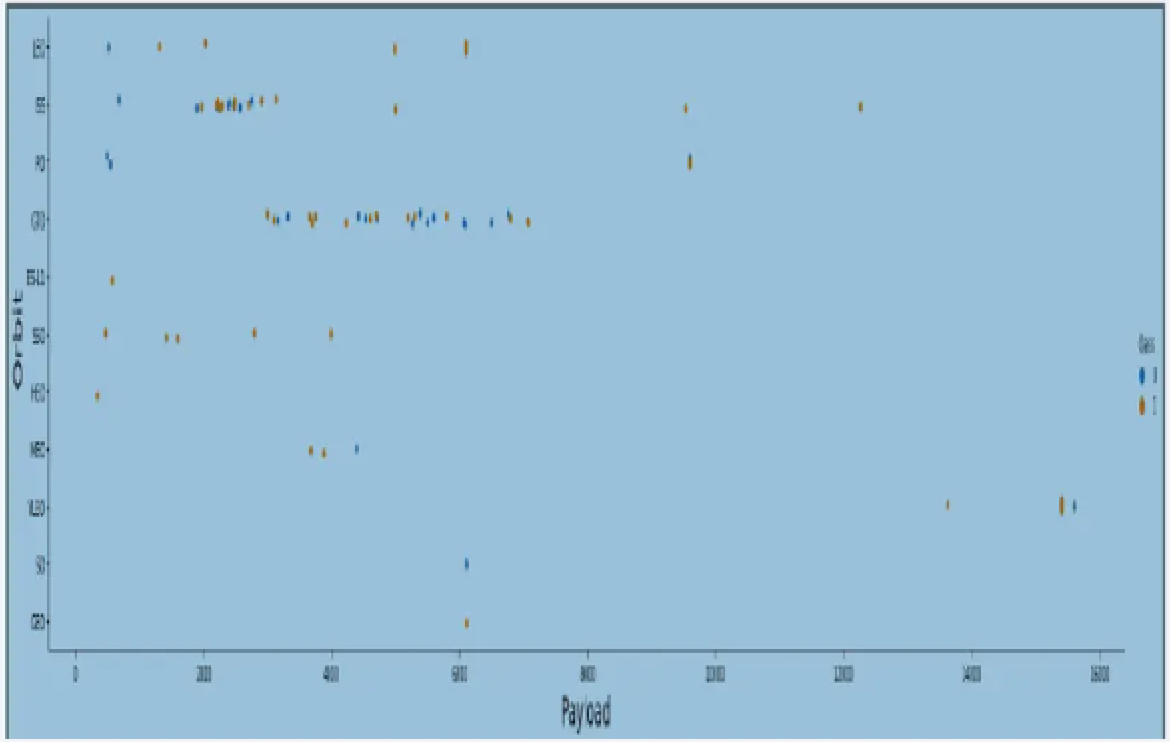
Flight Number vs. Orbit Type

- The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



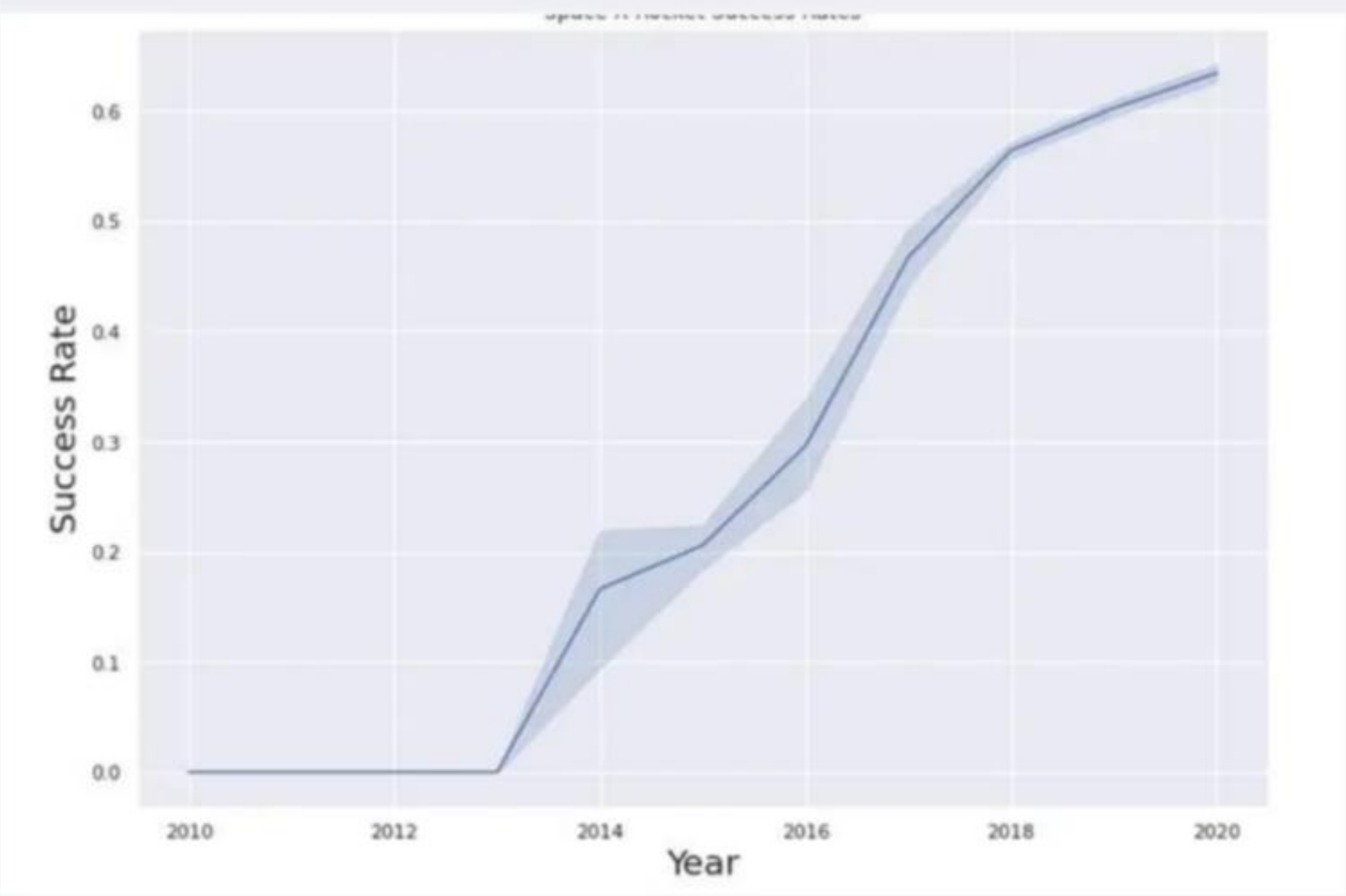
Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



Launch Success Yearly Trend

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020.



All Launch Site Names

- We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

Display the names of the unique launch sites in the space mission

[10]:

```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

t[10]:

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- We used the query below to display 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- **We calculated the total payload carried by boosters from NASA as 99980 using the query below**

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like 'NASA%'
```

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2534.66

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

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- We observed that the dates of the first successful landing outcome on ground pad was 1st May 2017

```
%sql select min(Date) from SPACEXTBL where "Landing _Outcome" = "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

```
%%sql

select Booster_Version from SPACEXTBL
where "Landing _Outcome" = "Success (drone ship)"
    and PAYLOAD_MASS__KG_>4000
    and PAYLOAD_MASS__KG_ < 6000
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- We used wildcard like '%' to filter for WHERE MissionOutcome was a success or a failure.

```
[10] %%sql

     select count(*) from SPACEXTBL
     where "Mission_Outcome" like "Success%"

      * sqlite:///my_data1.db
     Done.
     count(*)
     100
```

```
[11] %%sql

     select count(*) from SPACEXTBL
     where "Mission_Outcome" like "Failure%"

      * sqlite:///my_data1.db
     Done.
     count(*)
     1
```

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

```
%%sql

select Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

- We used a combinations of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

```
[14] %%sql

     select substr(Date, 4, 2) as Month, Booster_Version, Launch_Site from SPACEXTBL
     where substr(Date,7,4)='2015' and "Landing _Outcome" = "Failure (drone ship)"
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.
- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

```
%%sql

select "Landing _Outcome",
    count("Landing _Outcome") as landings
from SPACEXTBL
where Date >= "04-06-2010" and Date <= "20-03-2017"
group by "Landing _Outcome"
order by landings desc
```

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	landings
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Controlled (ocean)	3
Failure	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

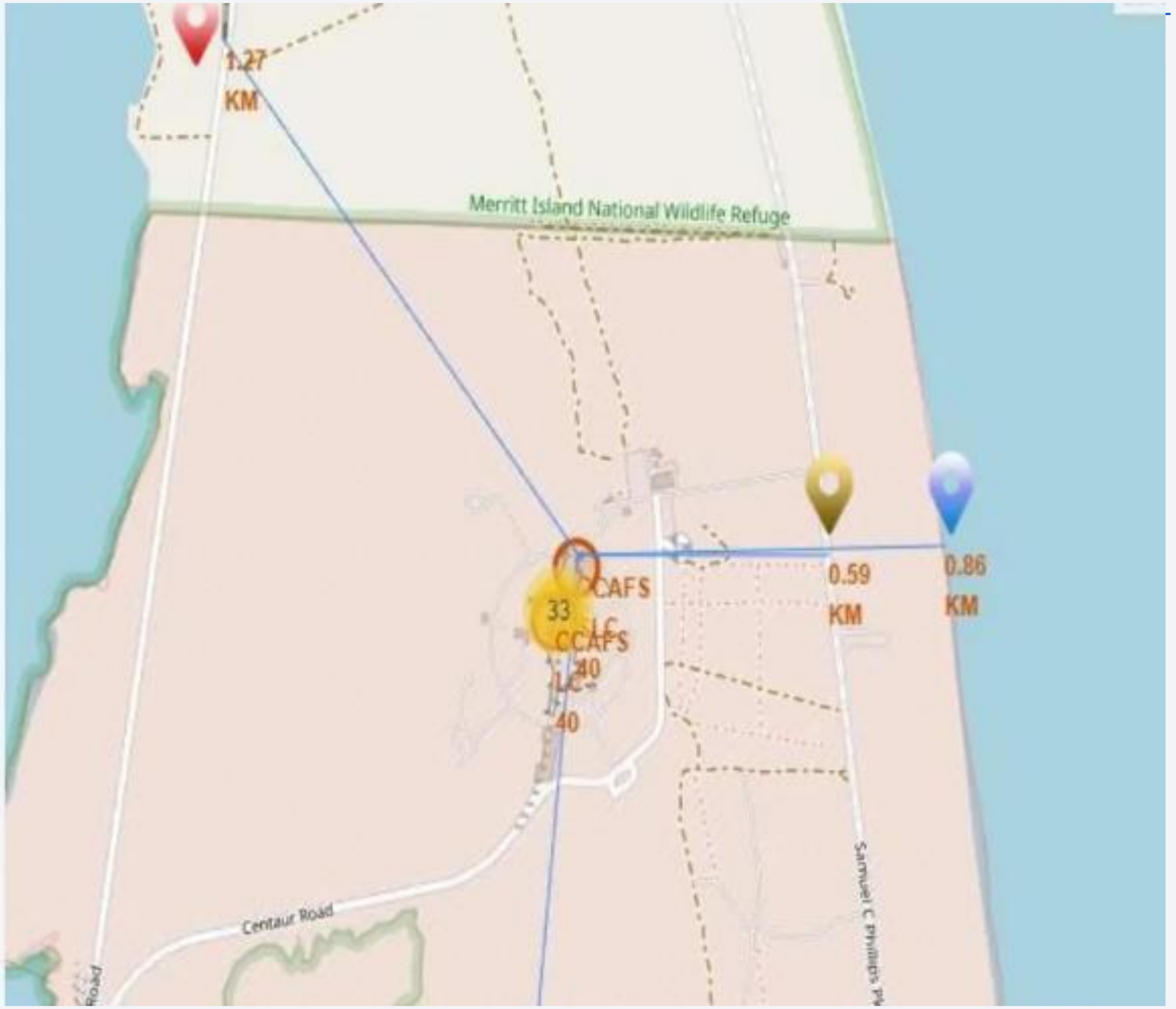
All launch sites global map markers

Markers showing launch sites with color labels

- We can see that CCAFS SLC-40 has low rate.

Launch Site distance to landmarks

- launch sites are less than 2km from railways.
- launch sites are less than 2km from highways.
- Launch sites are less than 5km from coastline.
- It keeps 15km away from the city.

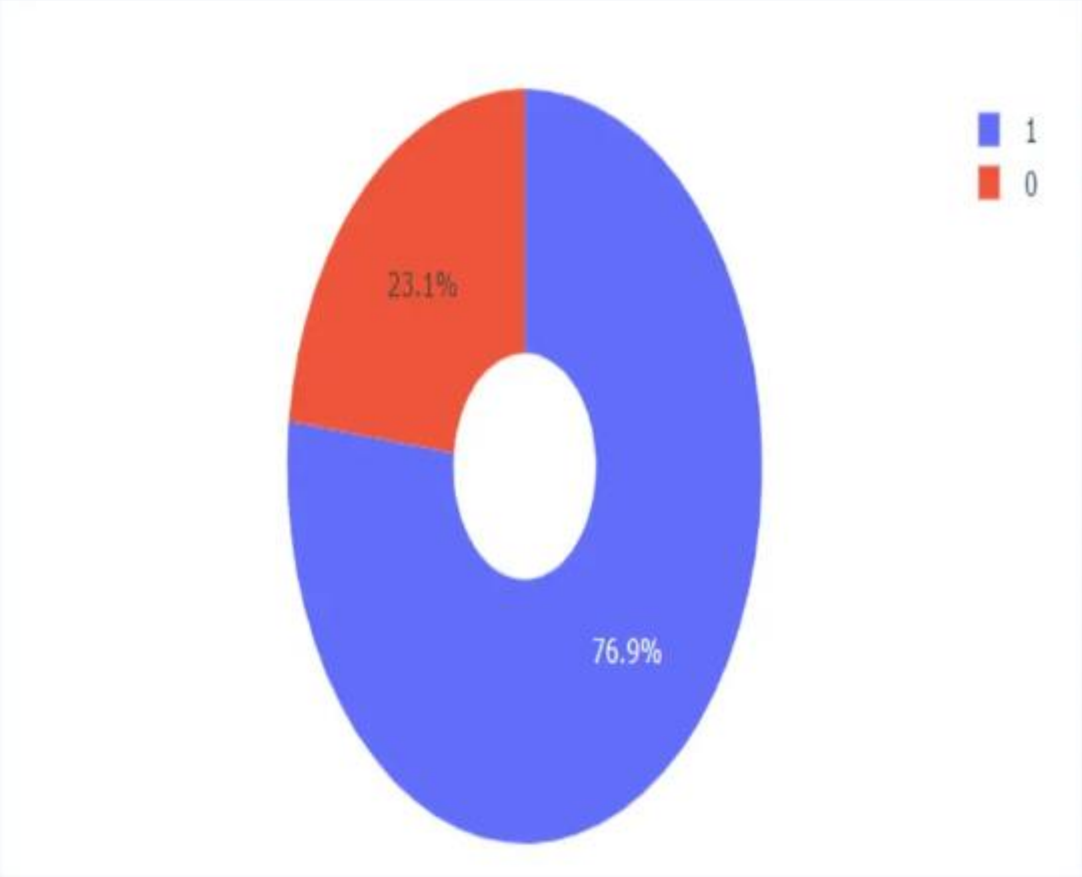


Section 4

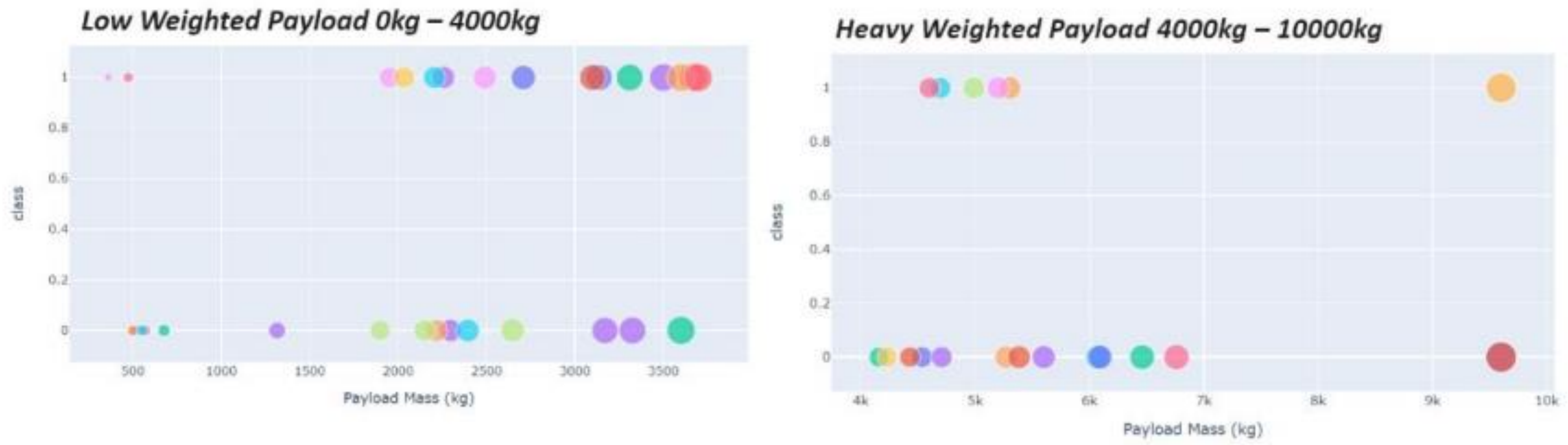
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site

Pie chart showing the Launch site with the highest launch success ratio



Pie chart shScatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slideowing the Launch site with the highest launch success ratio



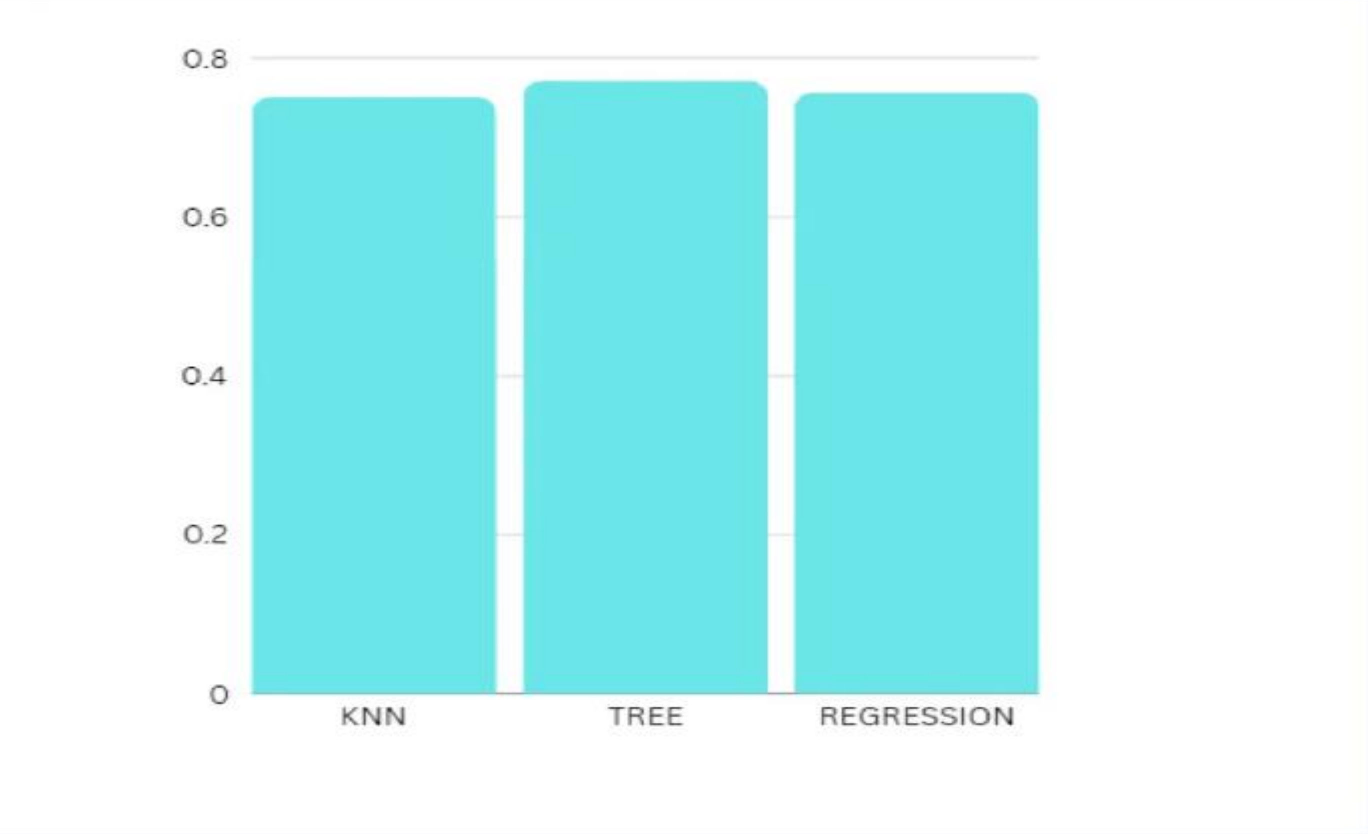
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

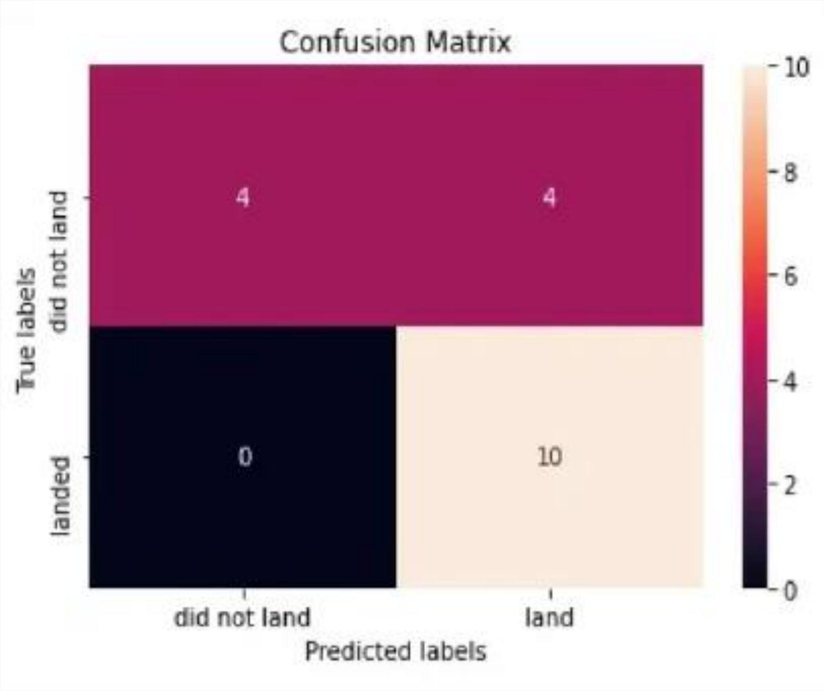
Classification Accuracy

- The decision tree classifier is the model with the highest classification accuracy



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!