Predicting Falcon 9 First Stage Landing Success

Shahram Hakimi
August 26, 2023

OUTLINE

- Executive Summary
- Introduction
- Methodology Data collection
- Data wrangling
- EDA and interactive visual analytics
- Predictive analysis
- Results EDA with visualization
- EDA with SQL
- Interactive map with FOLIUM
- Plotly Dash Dashboard
- Predictive analysis (classification)
- Conclusion
- Appendix

Summary of Methodologies

The research aims to identify the crucial factors contributing to a successful rocket landing. To achieve this goal, the study employs the following methodologies:

1. **Data Collection:**
   Utilization of SpaceX REST API and web scraping techniques for data gathering.
2. **Data Wrangling:**
   Transformation of collected data to establish a success/failure outcome variable.
3. **Data Exploration:**
   Application of data visualization techniques to explore various factors, including payload, launch site, flight number, and yearly trends.
4. **Data Analysis:**
   Employment of SQL for data analysis, involving calculations of statistics such as total payload, payload range for successful launches, and total count of successful and failed outcomes.
5. **Launch Site Investigation:**
   Exploration of launch site success rates and their proximity to significant geographical markers.
6. **Launch Site Visualization:**
   Visualization of launch sites with the highest success rates and successful payload ranges.
7. **Predictive Modeling:**
   Construction of models for predicting landing outcomes, incorporating logistic regression, support vector machine, decision tree, and K nearest neighbor algorithms.

Results

Exploratory Data Analysis:

- Observable enhancement in launch success rates over time.
- KSC LC 39A emerges as the most successful landing site.
- Orbits E S L1, GEO, HEO, and SSO exhibit a flawless 100% success rate.

Visualization/Analytics:

- Majority of launch sites are located near the equator, with all sites positioned close to coastlines.

Predictive Analytics:

- All models demonstrate comparable performance on the test set.

.

INTRODUCTION

The commercial space age is democratizing space travel through companies like Virgin Galactic, Rocket Lab, Blue Origin, and notably SpaceX. Each offers unique services, from suborbital flights to satellite launches. SpaceX stands out with achievements like ISS missions, Starlink internet, and crewed spaceflights. Their cost advantage lies in reusing the first rocket stage, unlike others that are much more expensive. Forest Katsch's diagrams illustrate the significance of the first stage in Falcon 9 rockets, which is pivotal for the launch. Goal of this research is by gathering SpaceX data, building dashboards, and analyzing data, predict the first stage reuse using machine learning. This is in the pursuit of challenging SpaceX in the rocket industry.

METHODOLOGY

1.Exploring SpaceX Launch Data Using specifically the SpaceX REST API and Web Scraping

2.Wrangling data by cleaning, transforming, and organizing raw data into a structured and usable format for analysis.

3. One Hot Encoding and Data Cleaning: This involves preparing the data for machine learning by transforming categorical variables through one-hot encoding and addressing missing or irrelevant data.

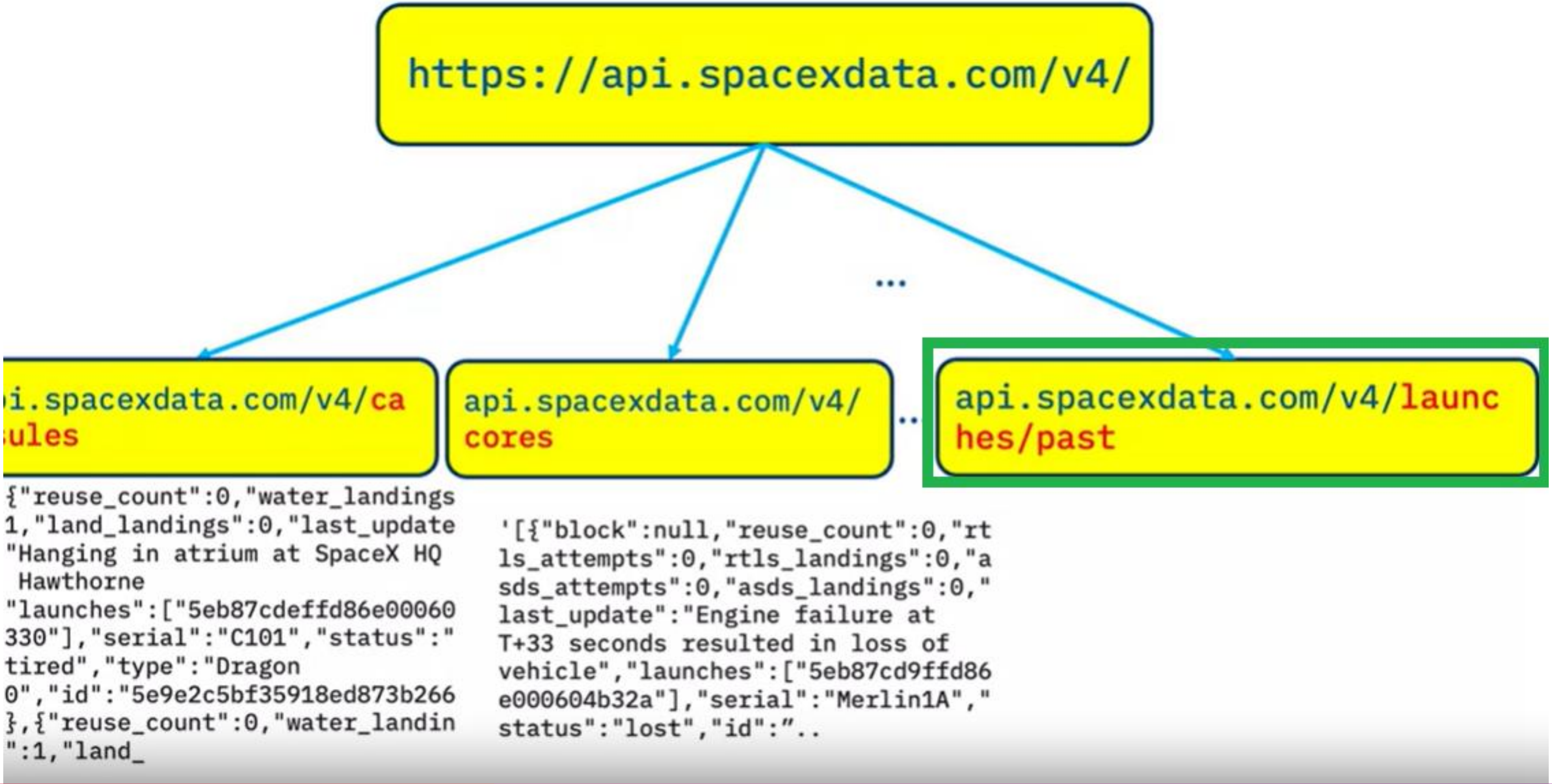
4. **Exploratory Data Analysis (EDA)**: Using both visualization techniques and SQL queries to understand and gain insights from the dataset's structure and patterns.

1.Interactive Visual Analytics with Folium and Plotly Dash: Utilizing tools like Folium and Plotly Dash to create interactive visualizations that enhance data understanding.

2.Predictive Analysis with Classification Models: Employing classification algorithms to predict outcomes based on input features.

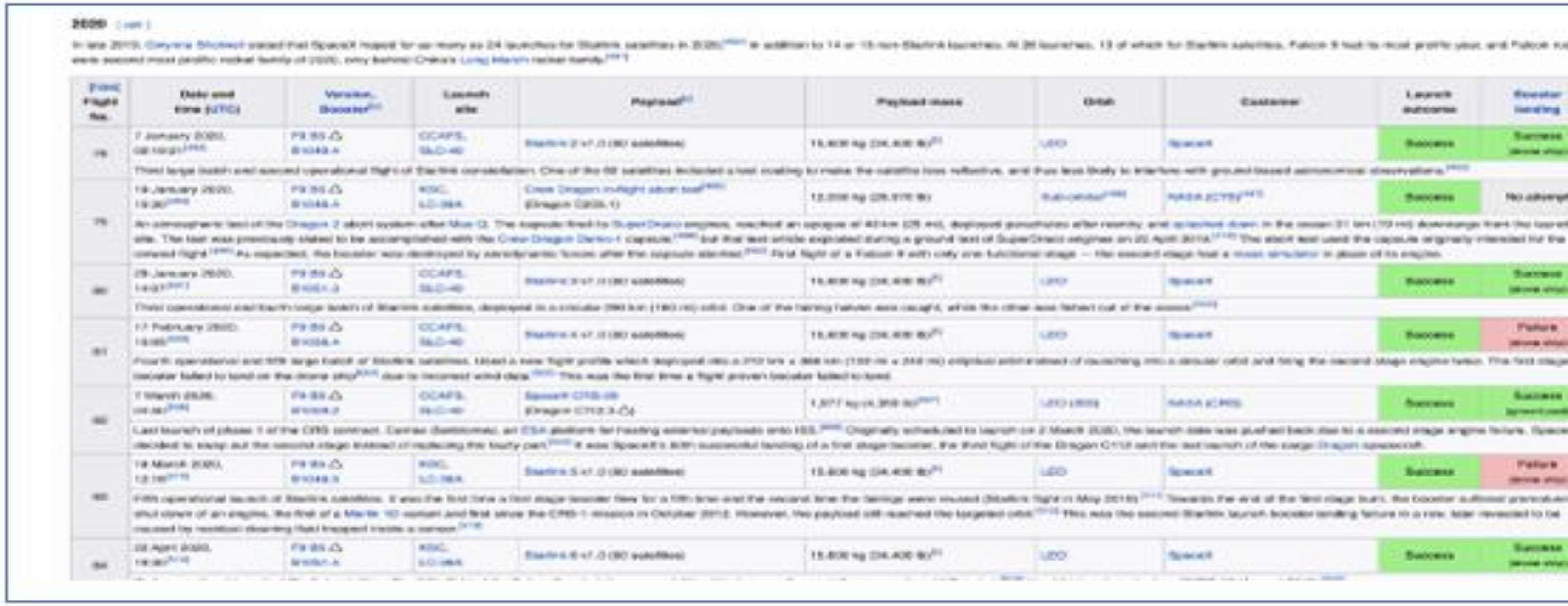
3.Building and Evaluating Classification Models: Developing models such as Logistic Regression (LR), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees (DT), then evaluating their performance to identify the best model.

- **DataCollection by : API**
- Obtain the response and decode it using the appropriate method. Convert the decoded data into a dataframe using the .json_ function.
- Use custom functions to request specific information about the launches from the SpaceX API.
- Generate a dictionary from the obtained data.
- Create a dataframe using the generated dictionary.
- Apply a filter to the dataframe to include only launches associated with the Falcon 9 rocket.
- Fill in any missing values in the Payload Mass column using a calculated method.
- Export the processed data to a CSV file.


https://api.spacexdata.com/v4/
i.spacexdata.com/v4/ca
ules
api.spacexdata.com/v4/
cores
...
api.spacexdata.com/v4/launc
hes/past
{"reuse_count":0,"water_landings
1,"land_landings":0,"last_update
"Hanging in atrium at SpaceX HQ
Hawthorne
"launches":["5eb87cdeffd86e00060
330"],"serial":"C101","status":"
tired","type":"Dragon
0","id":"5e9e2c5bf35918ed873b266
},{"reuse_count":0,"water_landin
":1,"land_
'[{"block":null,"reuse_count":0,"rt
ls_attempts":0,"rtls_landings":0,"a
sds_attempts":0,"asds_landings":0,"
last_update":"Engine failure at
T+33 seconds resulted in loss of
vehicle","launches":["5eb87cd9ffd86
e000604b32a"],"serial":"Merlin1A","
status":"lost","id":"..

Steps for getting Data by Web Scraping

- Retrieve Falcon 9 launch data from Wikipedia.
- Generate a BeautifulSoup object from the HTML response.
- Extract column names from the header of the HTML table.
- Gather data by parsing HTML tables.
- Construct a dictionary from the collected data.
- Generate a dataframe using the created dictionary.
- Export the data to a CSV file.



```
df.to_csv('spacex_web_scraped.csv', index=False)
```

In [18]:

```
df=pd.read_csv('spacex_launch_geo.csv')
df.head()
```

Out[18]:

	Flight Number	Date	Time (UTC)	Booster Version	Launch Site	Payload	Payload Mass (kg)	Orbit	Customer	Landing Outcome	class	Lat
0	1	6/4/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Failure (parachute)	0	28.562302
1	2	12/8/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel o...	0.0	LEO (ISS)	NASA (COTS) NRO	Failure (parachute)	0	28.562302
2	3	5/22/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2+	525.0	LEO (ISS)	NASA (COTS)	No attempt	0	28.562302
3	4	10/8/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	No attempt	0	28.562302
4	5	3/1/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	No attempt	0	28.562302

Data Wrangling

I performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
Also defined "Class" Column to convert outcomes into Training Labels with '1' means the booster successfully landed '0' means it was unsuccessful
Then export data as a CSV file

In [21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90 entries, 0 to 89
Data columns (total 18 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   FlightNumber    90 non-null     int64
 1   Date            90 non-null     object
 2   BoosterVersion  90 non-null     object
 3   PayloadMass     90 non-null     float64
 4   Orbit           90 non-null     object
 5   LaunchSite      90 non-null     object
 6   Outcome         90 non-null     object
 7   Flights         90 non-null     int64
 8   GridFins        90 non-null     bool
 9   Reused          90 non-null     bool
 10  Legs            90 non-null     bool
 11  LandingPad      64 non-null     object
 12  Block           90 non-null     float64
 13  ReusedCount     90 non-null     int64
 14  Serial          90 non-null     object
 15  Longitude       90 non-null     float64
 16  Latitude        90 non-null     float64
 17  Class           90 non-null     int64
dtypes: bool(3), float64(4), int64(4), object(7)
memory usage: 10.9+ KB
```

In [15]:
```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class=[]
for i in df['Outcome']:
    if i in (bad_outcomes):
        i=0
        landing_class.append(i)
    else:
        i=1
        landing_class.append(i)
```

In [16]:
```
landing_class_1 = landing_class.count(1)
landing_class_0 = landing_class.count(0)

print("Successes:", landing_class_1)
print("Failures:", landing_class_0)
```

```
Successes: 60
Failures: 30
```

In [17]:
```
df['Class']=landing_class
df[['Class']].head(8)
```

Out[17]:

	Class
0	0
1	0
2	0
3	0

In [23]:
```
df[['Outcome','Class']]
```

Out[23]:

	Outcome	Class
0	None None	0
1	None None	0
2	None None	0
3	False Ocean	0

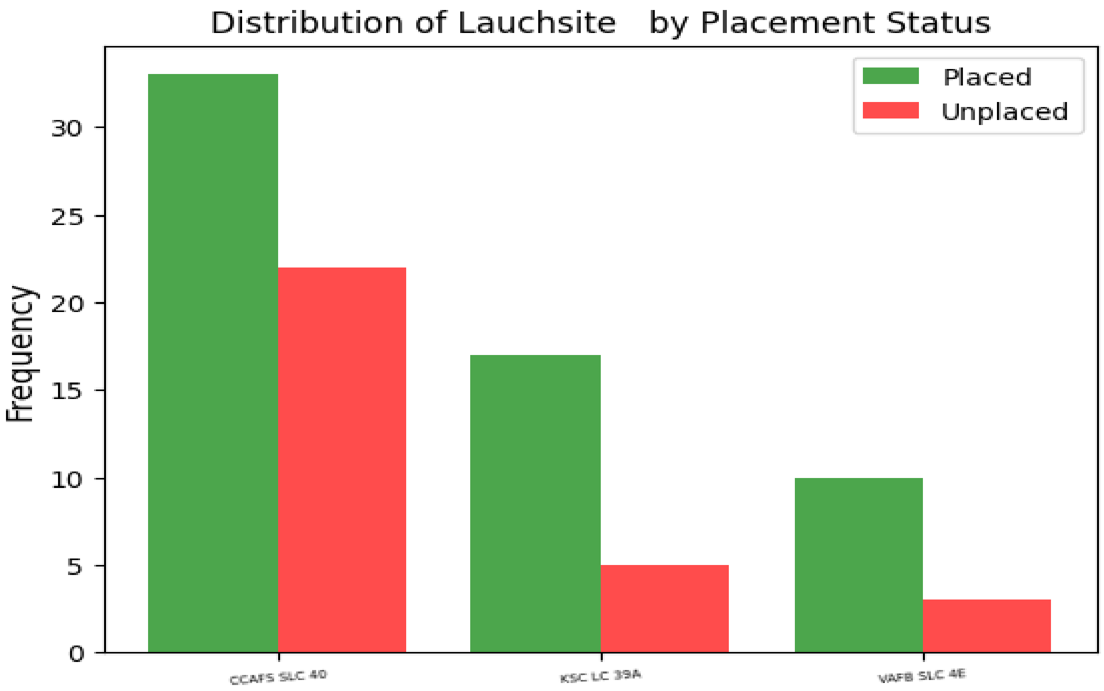
EDA WITH DATA VISUALIZATION

1. We first started by using scatter graph to find the relationship between the attributes such as between:

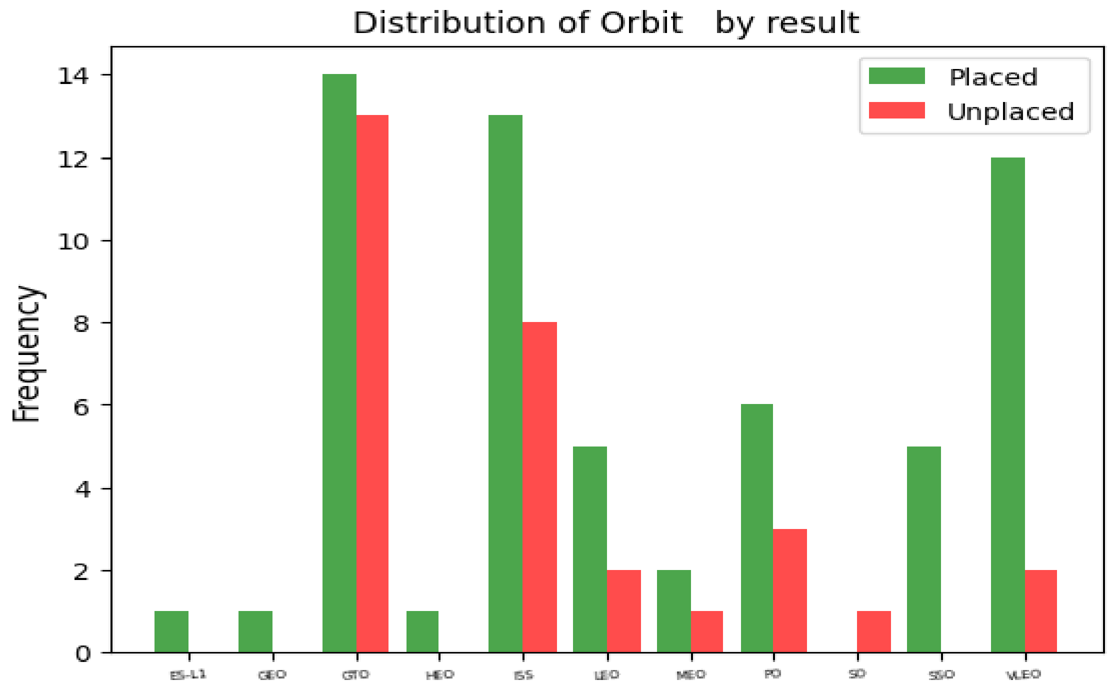
- Payload and Flight Number.
- Flight Number and Launch Site.
- Payload and Launch Site.
- Flight Number and Orbit Type.
- Payload and Orbit Type.

and find relationship by using bar charts between Class types(successful =Placed, unsuccessful =Unplaced)

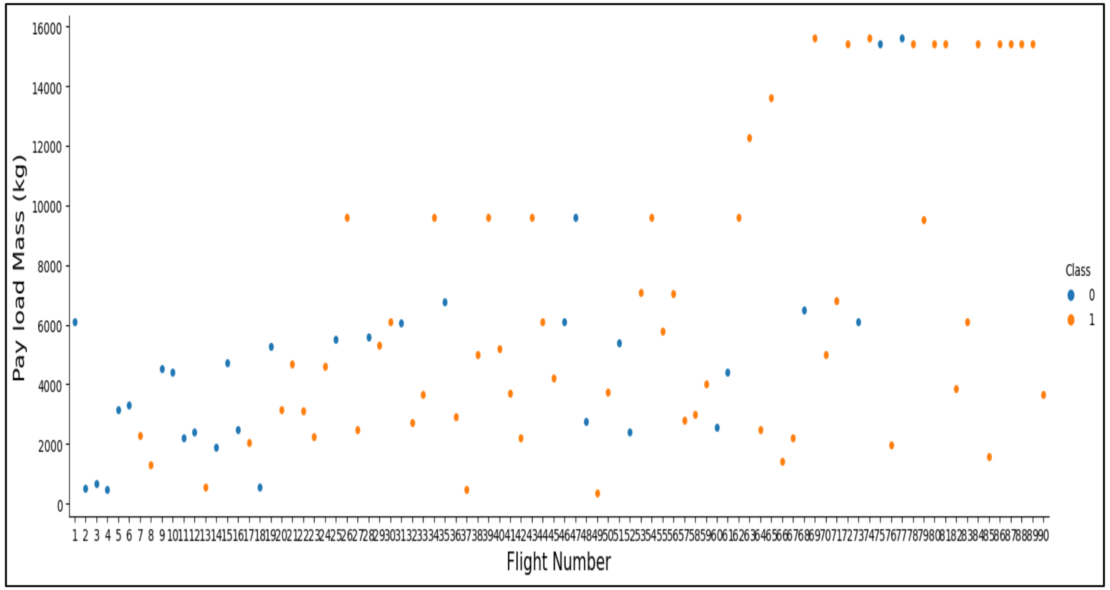
orbit and Launchsite

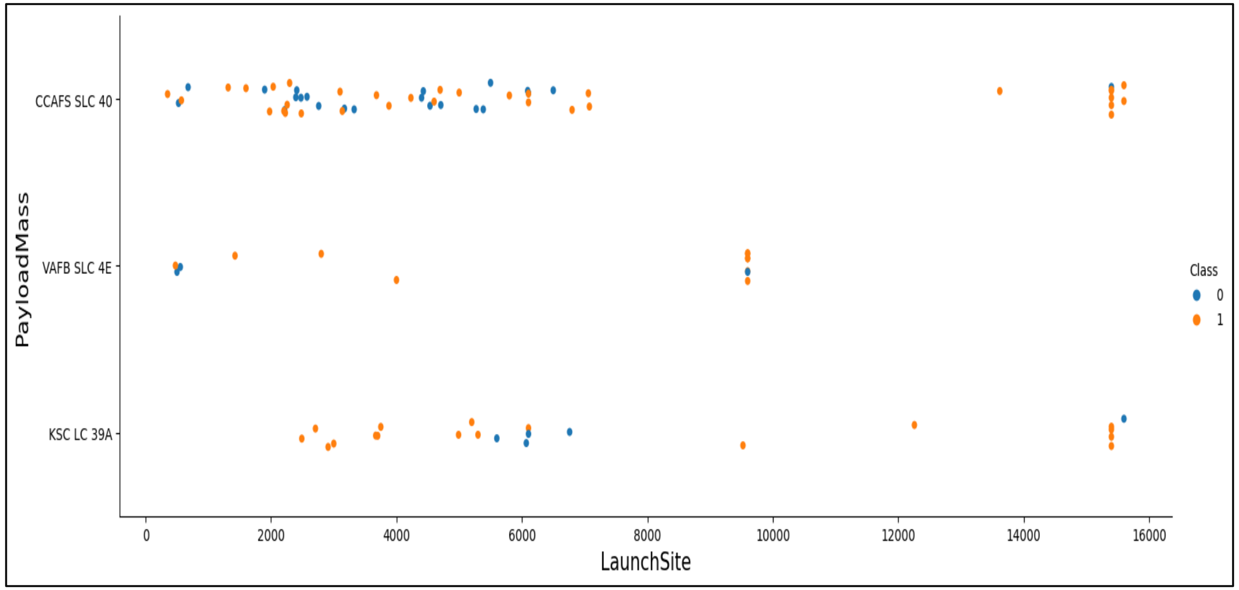


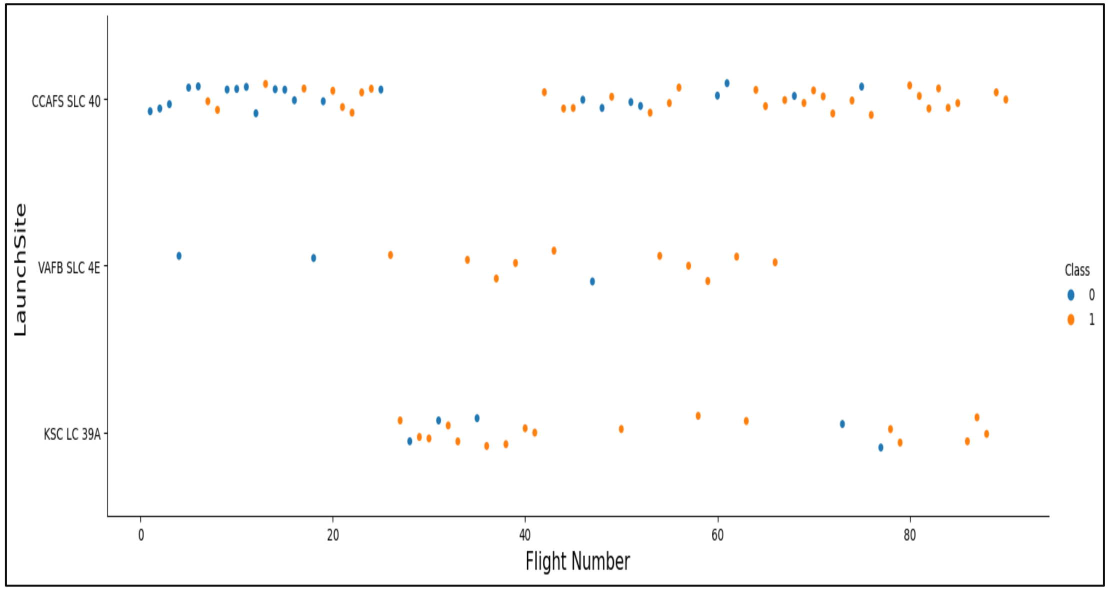
LaunchSite	Placed	Unplaced
CCAFS SLC 40	33	22
KSC LC 39A	17	5
VAFB SLC 4E	10	3

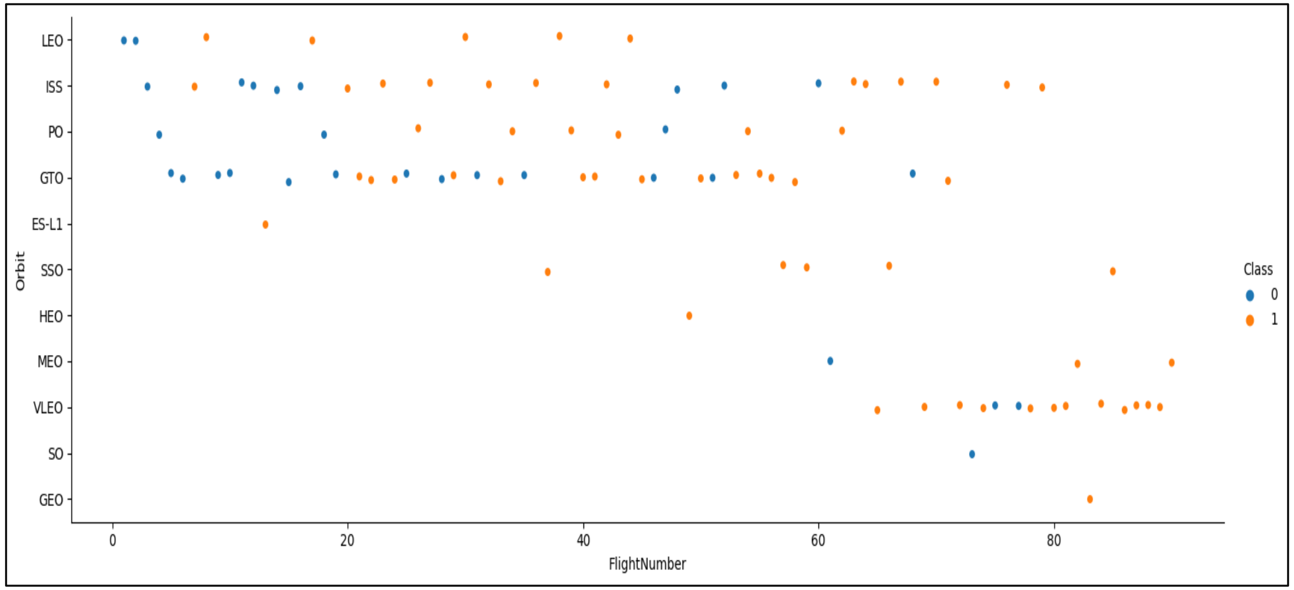


Orbit	Placed	Unplaced
ES-L1	1	0
GEO	1	0
GTO	14	13
HEO	1	0
ISS	13	8
LEO	5	2
MEO	2	1
PO	6	3
SO	0	1
SSO	5	0
VLEO	12	2


Pay load Mass (kg)
16000
14000
12000
10000
8000
6000
4000
2000
0
Flight Number
Class
0
1


PayloadMass
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
0
2000
4000
6000
8000
10000
12000
14000
16000
LaunchSite
Class
0
1


LaunchSite
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
0
20
40
60
80
Flight Number
Class
0
1


Orbit
LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
0
20
40
60
80
FlightNumber
Class
0
1


Yearly Success Rate Trend
Average Success Rate
Year
0.0
0.2
0.4
0.6
0.8
1.0
2010
2012
2014
2016
2018
2020

EDA WITH SQL

SQL queries performed include:-

Displaying the names of the launch sites.

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

- Displaying the total payload mass carried by the booster launched by NASA (CRS).

45596 kg

- Displaying the average payload mass carried by booster version F9 v1.1.

2928.4

-

EDA with SQL

Listing the date when the first successful landing outcome in the ground pad was achieved.

First_Successful_Landing_On_Ground_Pad
2015-12-22

- List the names of the boosters which have success in drone ships and have payload mass greater than 4000 but less than 6000.

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

- Listing the total number of successful and failed mission outcomes.

Failure (in flight)	**1**
Success	**98**
Success	**1**
Success (payload status unclear)	1

EDA with SQL

- Listing the names of the booster_versions which have carried the maximum payload mass.

F9 B5 B1048.4	**F9 B5 B1049.4**	**F9 B5 B1051.3**
F9 B5 B1056.	**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**

- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for the year 2015.

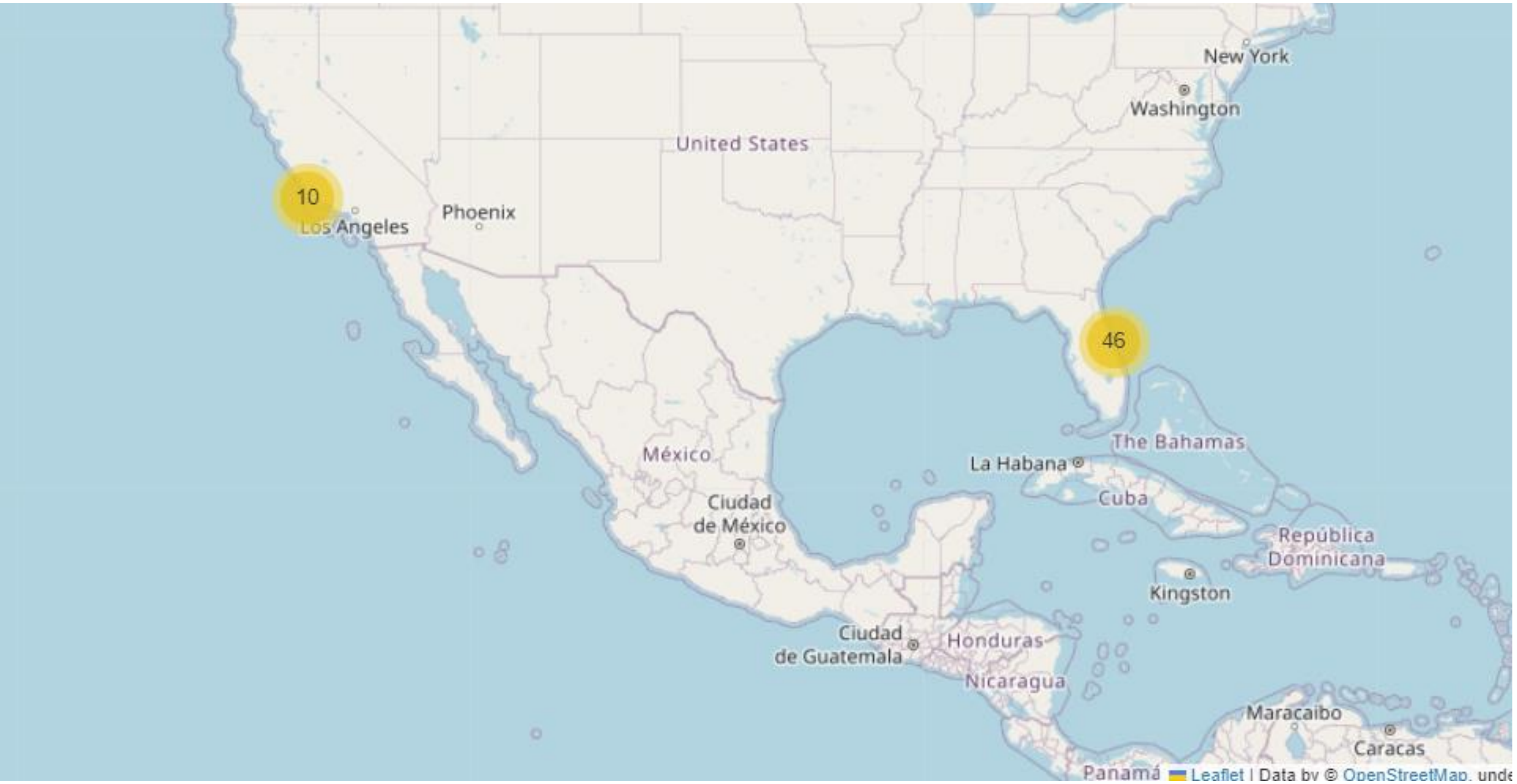
October	**Failure (drone ship)**	**F9 v1.1 B1012**	**CCAFS LC-40**
April	**Failure (drone ship)**	**F9 v1.1 B1015**	**CCAFS LC-40**

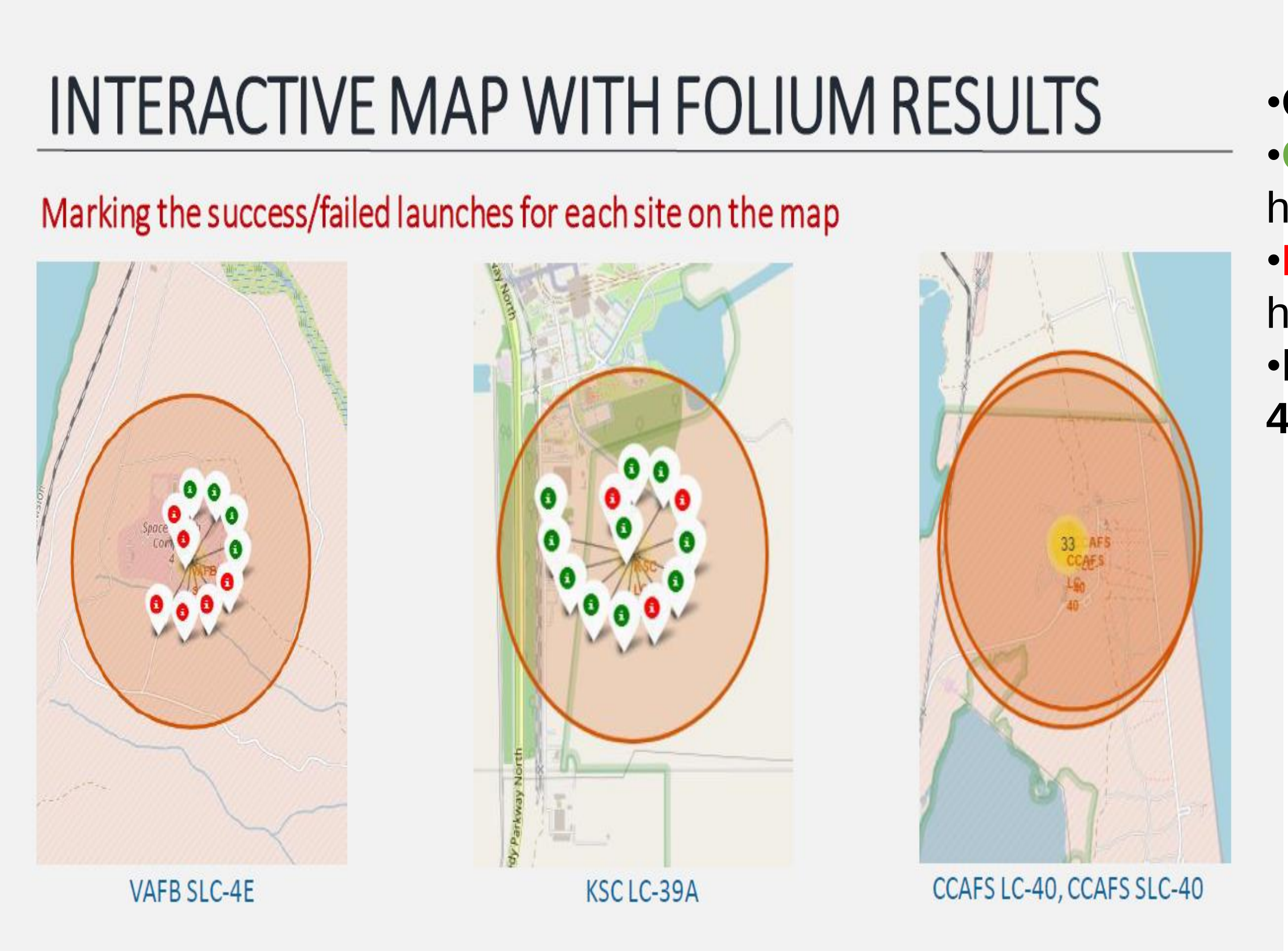
- Rank the count of landing outcomes or success between the dates 2010-06-04 and 2017-03-20, in descending order.

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**

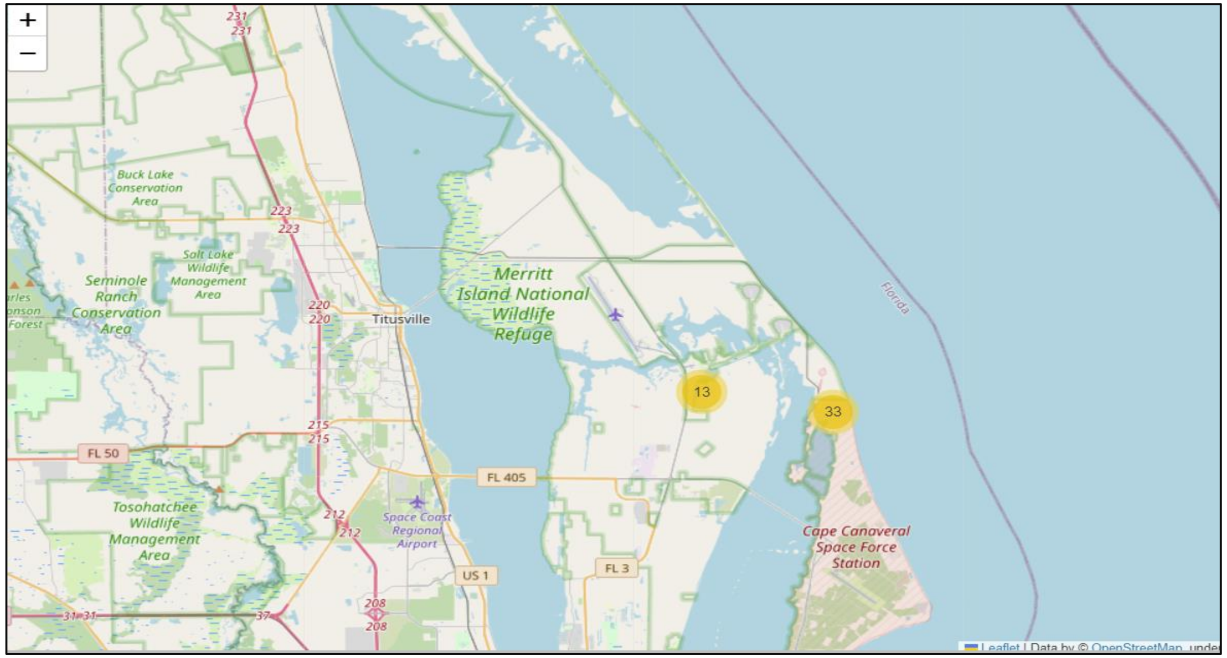
Build an Interactive Map with Folium

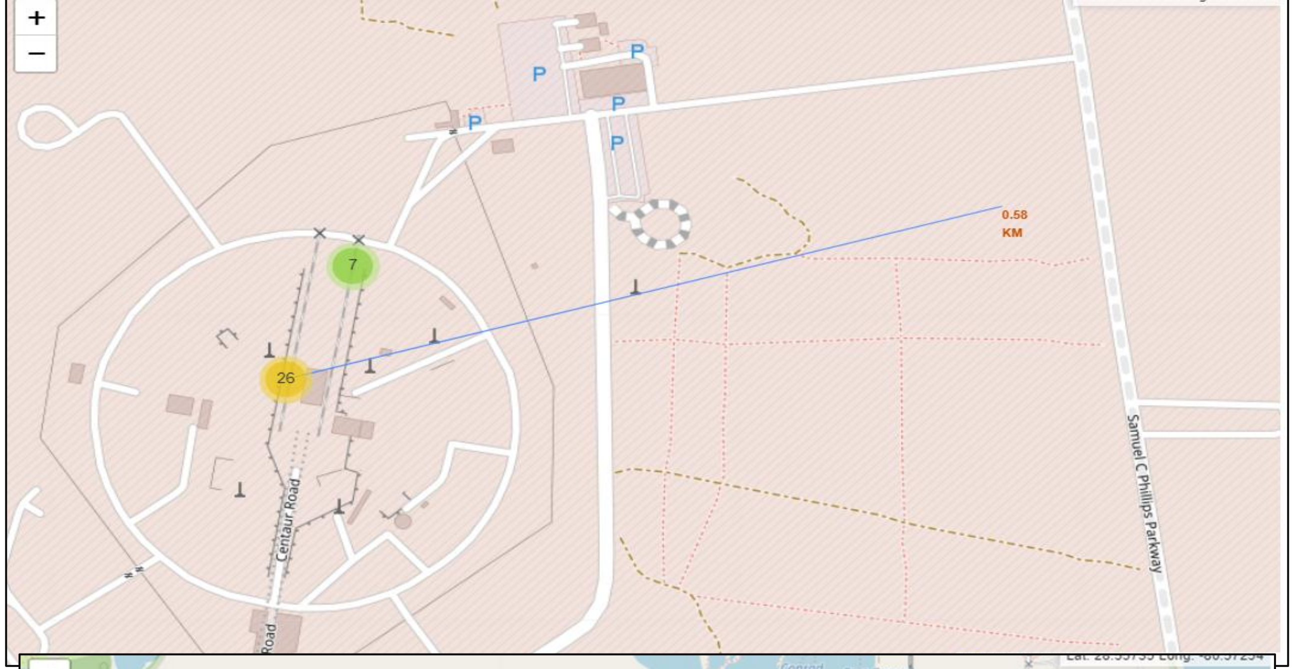
Marking the success/failed launches for each site on the map

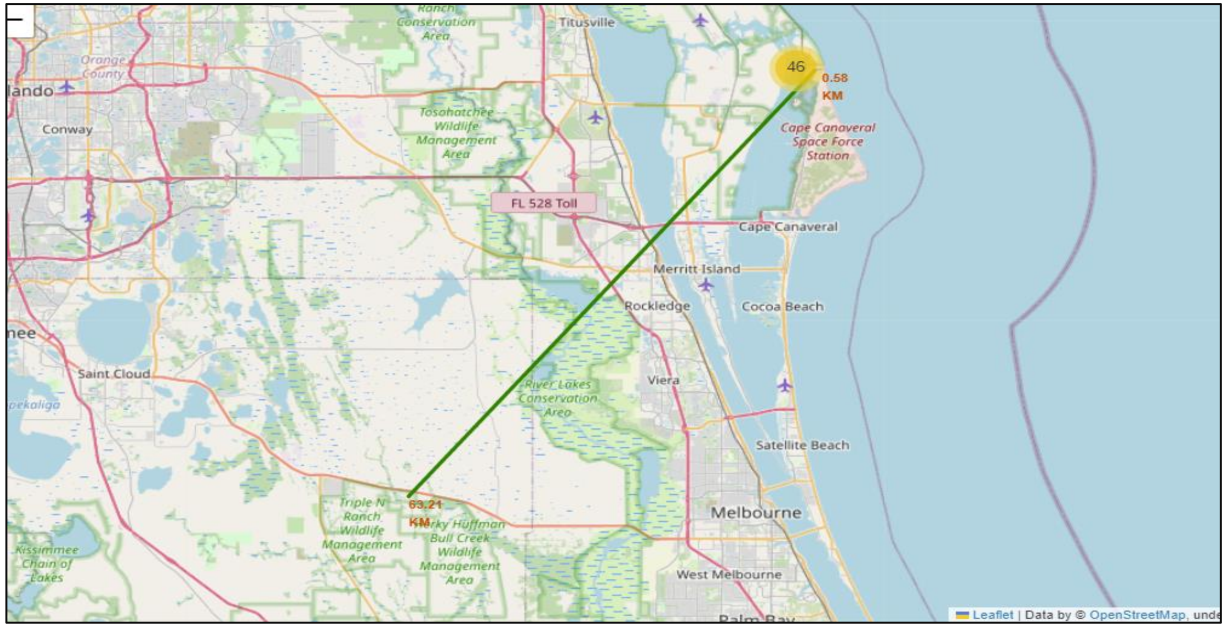


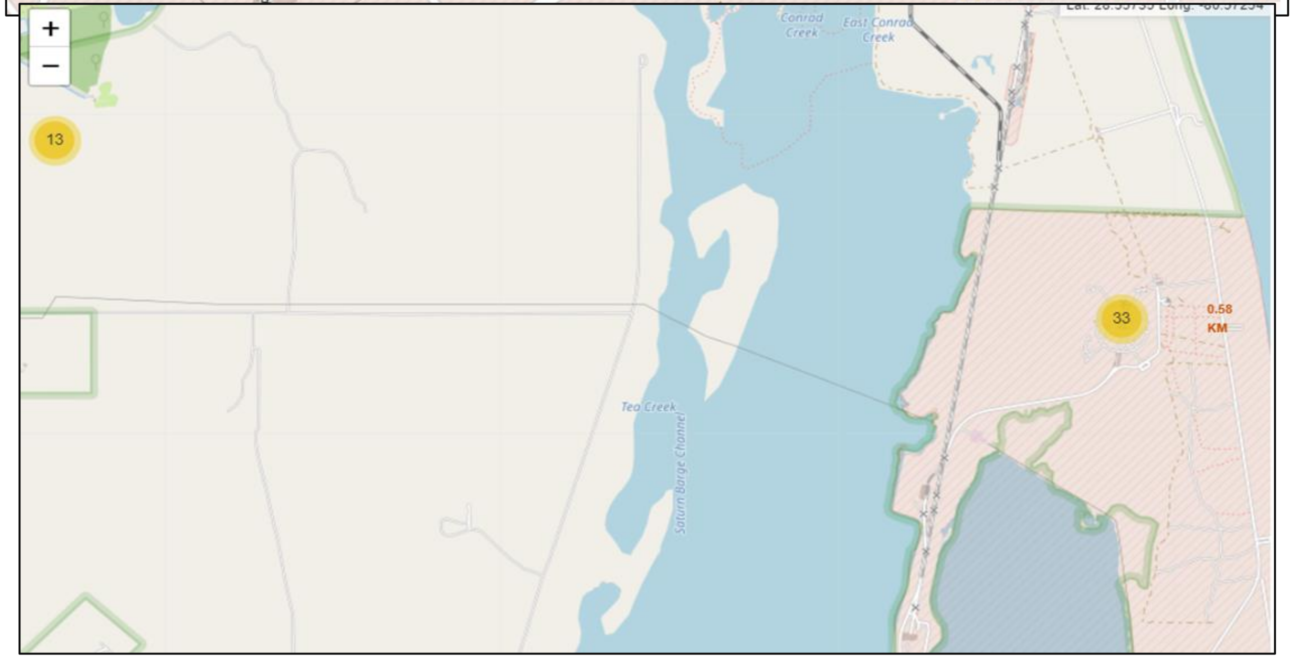


- **Outcomes**:
- Greenmarkersforsuccessfullaunches
- Redmarkersforunsuccessfullaunches
- Launchsite**CCAFSSLC-40**hasa**3/7successrate**(**42.9%)**


Buck Lake Conservation Area
Salt Lake Wildlife Management Area
Seminole Ranch Conservation Area
Titusville
Merritt Island National Wildlife Refuge
13
33
FL 50
FL 405
Tosohatchee Wildlife Management Area
Space Coast Regional Airport
Cape Canaveral Space Force Station
US 1
FL 3
Leaflet | Data by © OpenStreetMap


7
26
0.58 KM
Centaur Road
Samuel C Phillips Parkway


Titusville
Conway
Tosohatchee Wildlife Management Area
46
0.58 KM
Cape Canaveral Space Force Station
FL 528 Toll
Cape Canaveral
Merritt Island
Rockledge
Cocoa Beach
Saint Cloud
Viera
River Lakes Conservation Area
Satellite Beach
Triple N Ranch Wildlife Management Area
63.21 KM
Bull Creek Wildlife Management Area
Melbourne
West Melbourne
Leaflet | Data by © OpenStreetMap


13
33
0.58 KM
Tea Creek
Saturn Barge Channel
East Conrad Creek

Calculating the proximity of each site to a coastline, railway, highway and city

The closest launch sites to a coastline are:CCAFS LC-40 andCCAFS SLC-40

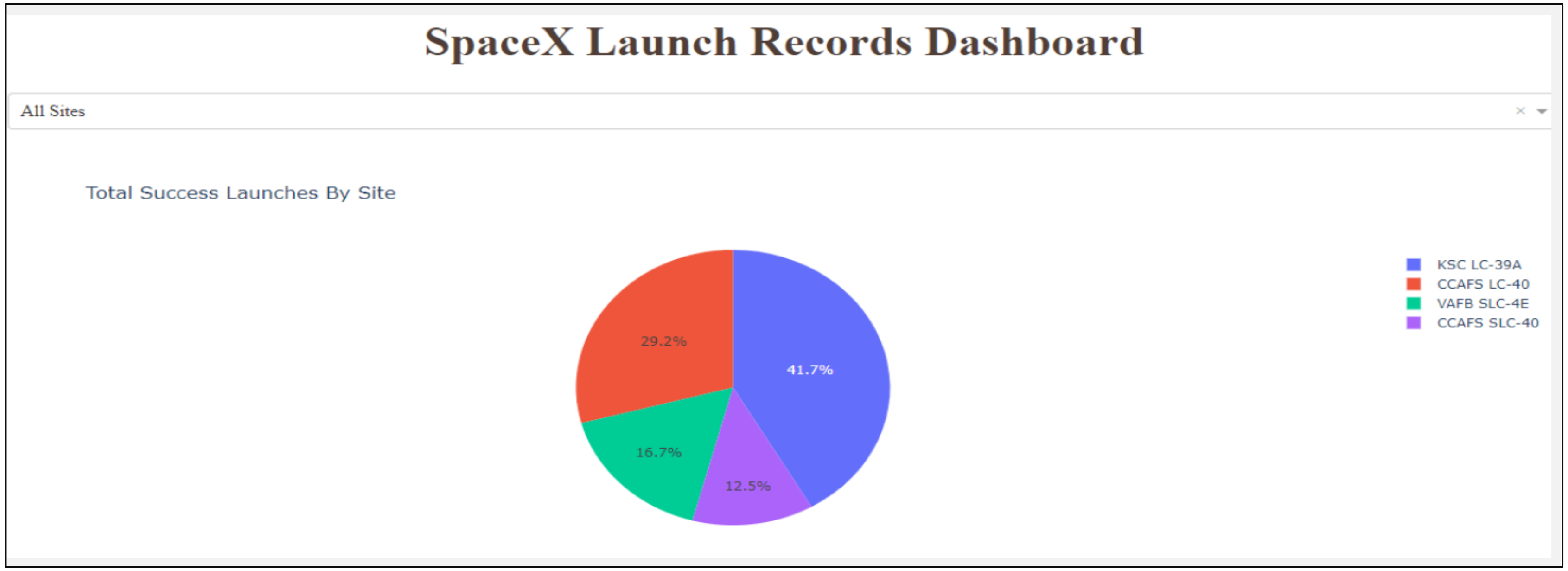
All launch sites have an average proximity to a railway of 1,15 km

VAFB SLC-4E have a longest distance to a highway (8,85 km)

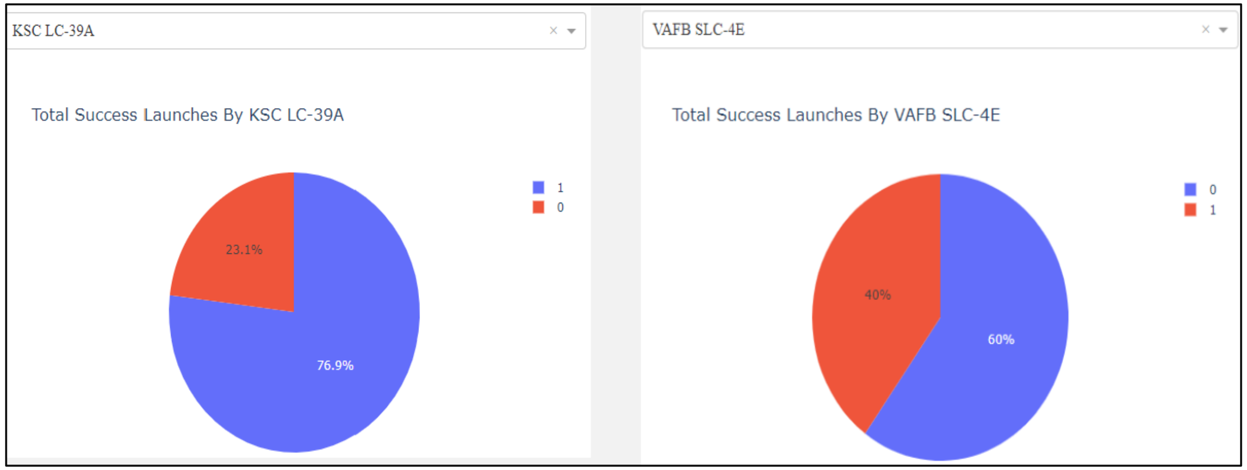
The launch sites farthest from city are: CCAFS LC-40,CCAFS SLC-40 andKSC LC-39A with a distance of 51,5 km. VAFB SLC-4E is closer to a city with 12,6 km.

KSC LC 39A had the most successful launches of any sites.

PLOTLY DASH DASHBOARD RESULTS


SpaceX Launch Records Dashboard
All Sites
Total Success Launches By Site
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
41.7%
29.2%
16.7%
12.5%

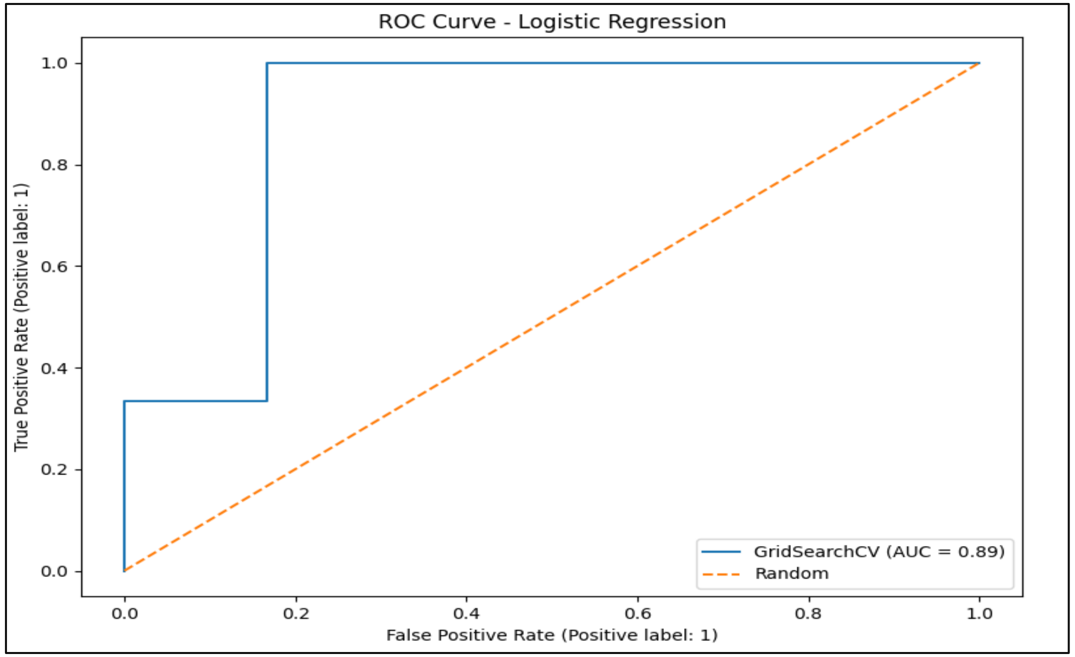

CCAFS LC-40
Total Success Launches By CCAFS LC-40
0
1
26.9%
73.1%
CCAFS SLC-40
Total Success Launches By CCAFS SLC-40
0
1
42.9%
57.1%

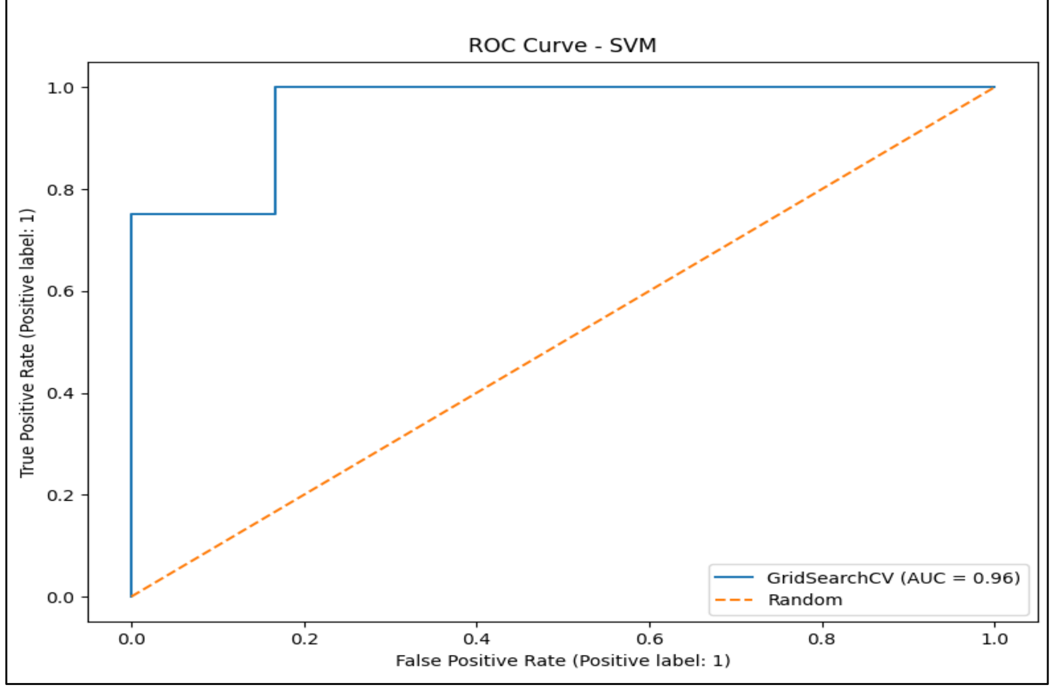

KSC LC-39A
Total Success Launches By KSC LC-39A
1
0
23.1%
76.9%
VAFB SLC-4E
Total Success Launches By VAFB SLC-4E
0
1
40%
60%

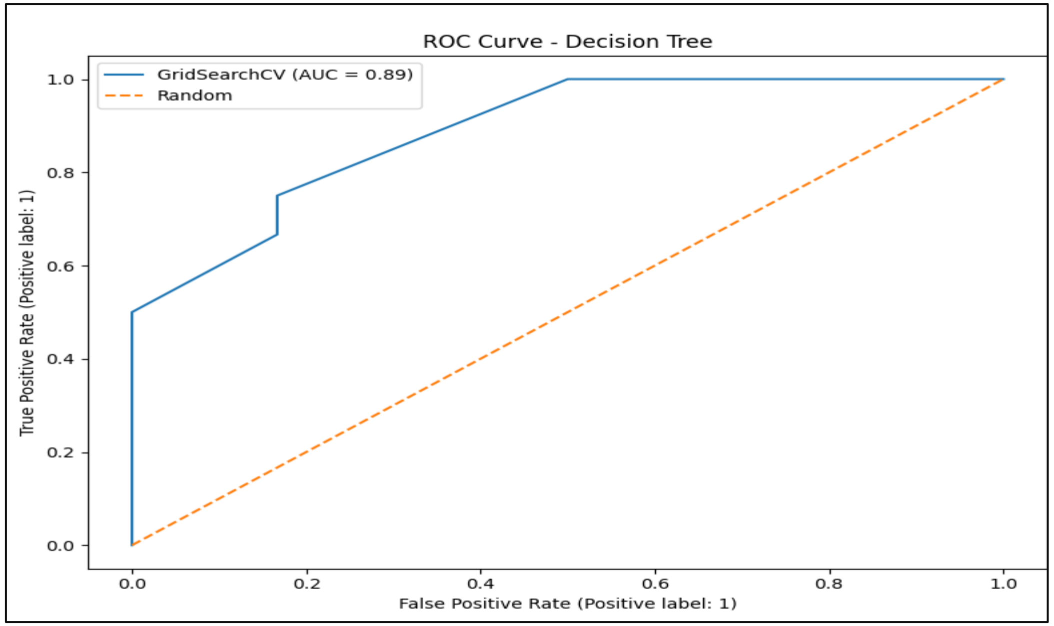
Payload range (Kg):

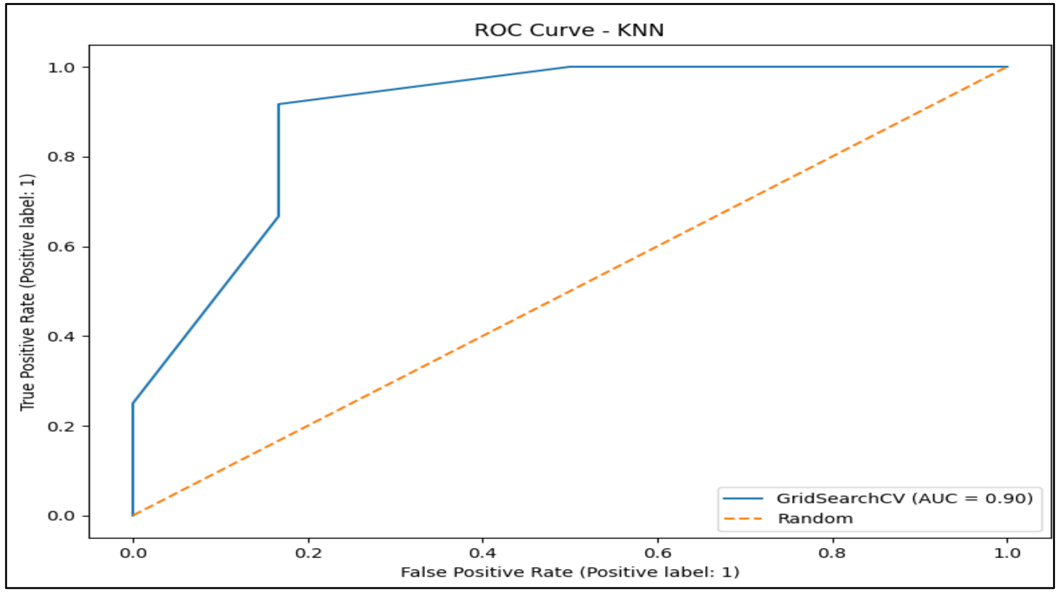
0 2000 4000 6000 8000 10000

MACHINE LEARNING PREDICTION

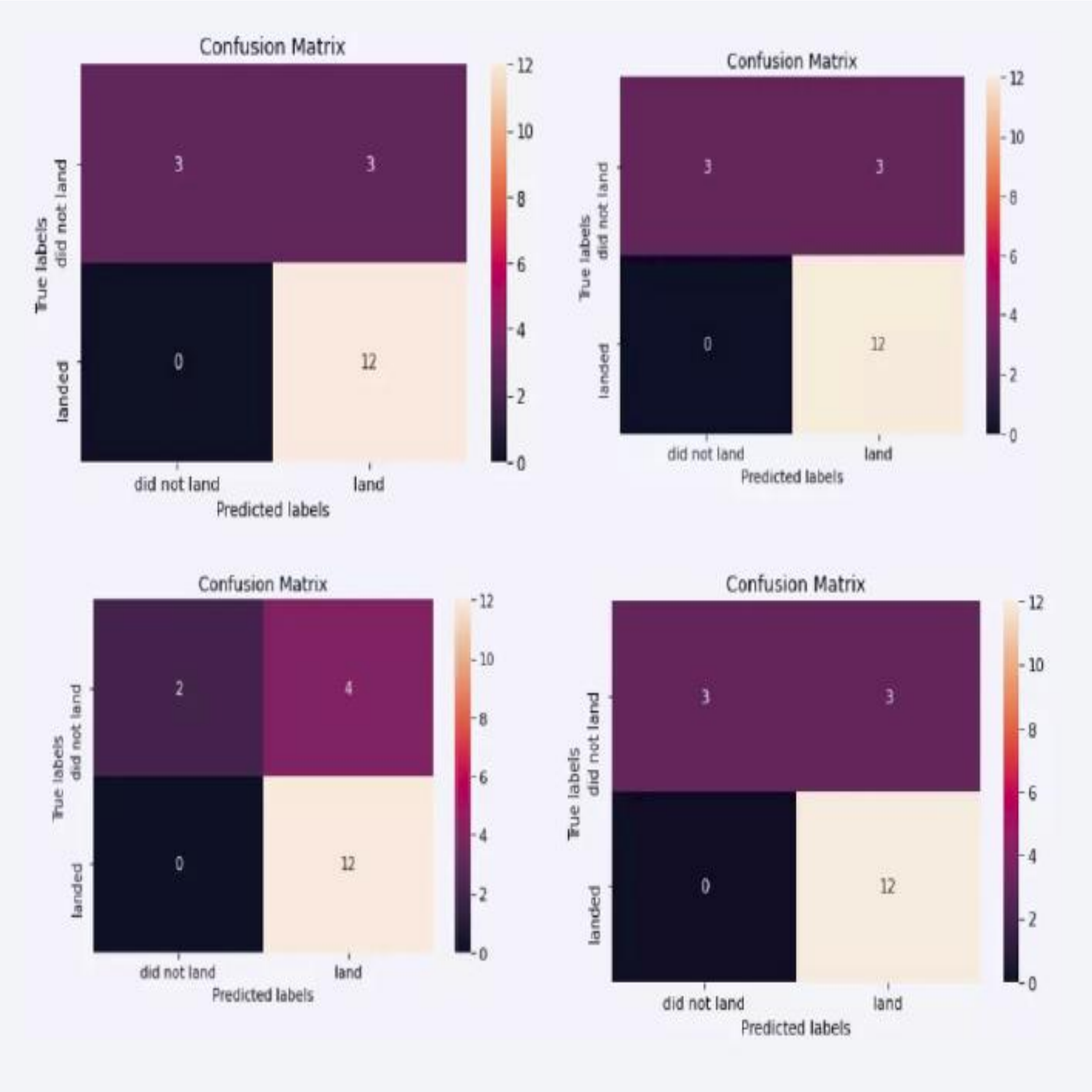

ROC Curve - Logistic Regression
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
GridSearchCV (AUC = 0.89)
Random


ROC Curve - SVM
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
GridSearchCV (AUC = 0.96)
Random


ROC Curve - Decision Tree
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
GridSearchCV (AUC = 0.89)
Random


ROC Curve - KNN
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
GridSearchCV (AUC = 0.90)
Random

CONFUSION MATRIX



- The accuracy of knn_cv = 0.8333333333333334
- The accuracy of decision tree_cv = 0.8333333333333334
- The accuracy of SVM= 0.8333333333333334
- The accuracy of Logistic Regression= 0.8333333333333334

Accuracy

All the models achieved similar performance with equivalent scores and precision. This uniformity in results is probably because of the limited dataset. Among the models, the Decision Tree slightly outperformed the others, especially when considering the best score.

CONCLUSION

1. The SVM, KNN, and Logistic Regression models have superior predictive accuracy on this dataset.
2. Payloads with lower weights outperform their heavier counterparts.
3. The rate of success in SpaceX launches increases in direct proportion to the number of years taken to refine the launches.
4. Among all the sites, KSC LC 39A recorded the highest number of successful launches.
5. The Orbit categories GEO, HEO, SSO, and ES L1 demonstrate the highest rate of success.