IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Rohan B. Salunkhe
26|07|2022

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

# Executive Summary

- Summary of methodologies
  - Data collection, wrangling, visualization, building interactive maps and charts, and predictive analysis is done using SpaceX API, Jupyter Notebook, Python IDE.
  - Programming languages used for the projects are Python and SQL with the help of python libraries such as Plotly, NumPy, Folium, matplotlib and many more.
- Summary of all results
  - During this projects I came across some amazing results. You will find them in the later slides.
  - These results include EDA results, interactive  analytics and predictive analysis.
  - Classification method is used for predictive analysis.

# Introduction

- Project background and context
  - SpaceX has made a clever decision. They produced technology that lets them reuse the rocket boosters by safely landing them on the launch sites. By using this clever technique, they reduced cost of space exploration.
  - Here is small comparison SpaceX advertises Falcon 9 rocket launch cost of $62Mn; where other providers cost upward of $165Mn.
  - That is cost saving of 250+ percent
- Problems you want to find answers
  - With every launch there is probability of successful landing. In this project we will predict if the Falcon 9 first stage will land successfully or not.

Section 1

# Methodology

# Methodology

## Executive Summary

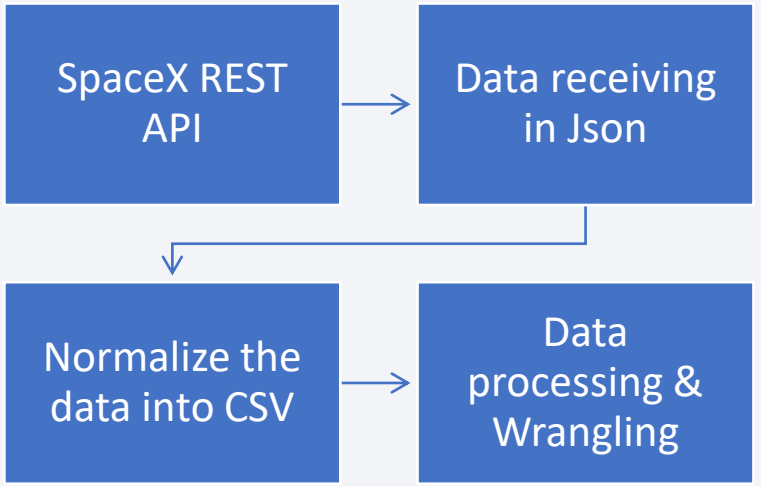
- Data collection methodology:
  - Data collection is done using SpaceX official REST API. By using we can collect various data points such as launch details, booster details, landing site details.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models such as LR, KNN, SVM and DT are used for analysis.

# Data Collection

- Data collection was done using SpaceX REST API.
- Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
- We then cleaned the data, checked for missing values and fill in missing values where necessary.
- Data scrapping is performed from Wikipedia for more details.
- The objective was to extract the launch records as HTML table and convert it to a pandas dataframe for future analysis.

# Data Collection – SpaceX API

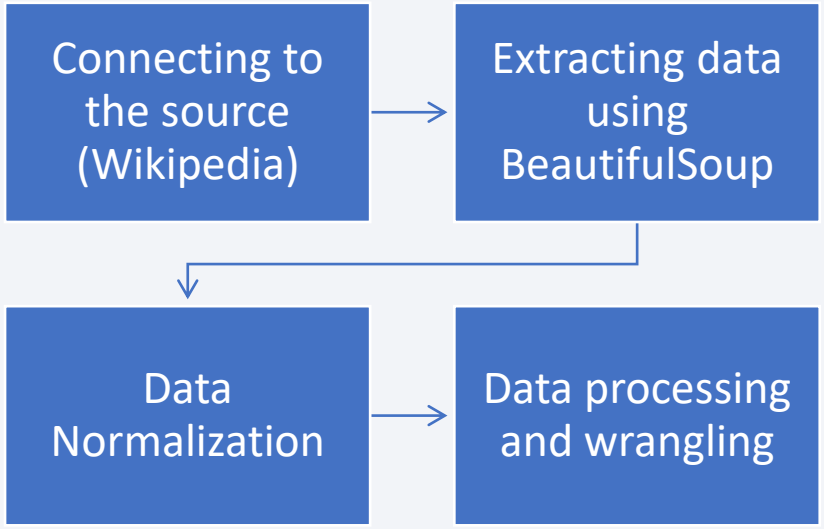
- Data collection is done through SpaceX REST API.
- Data received in Json format which is then normalized and converted to CSV format for Data processing and Wrangling.




- GitHub URL for data collection notebook

# Data Collection - Scraping

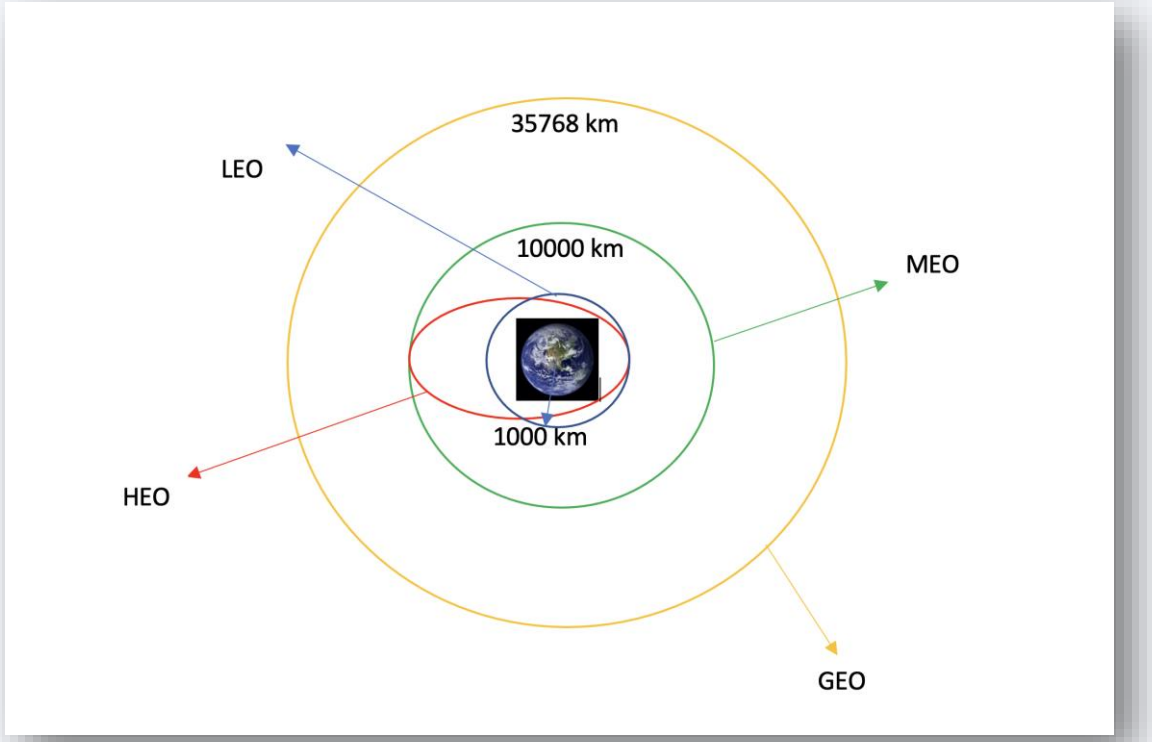
- Web scraping is done from Wikipedia.
- Data is then extracted in the form of table using BeautifulSoup object.
- Finally, data normalization is done and further data processing is done.




- GitHub URL

# Data Wrangling

- Exploratory data analysis is performed to determine the training labels.
- Calculation of number of launches at each site, & occurrence of each orbits is performed.
- Created landing outcome label from outcome column and exported the results to CSV.



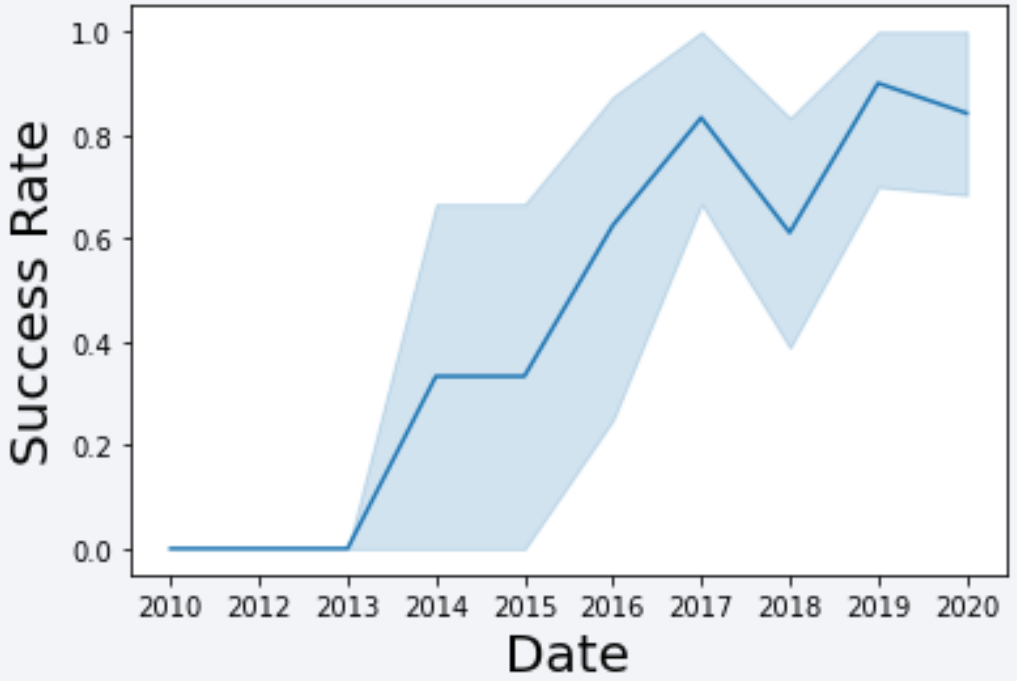

Earth Orbits

- GitHub URL

# EDA with Data Visualization

- Scatter plots, Line plots, bar graphs were plotted based for various points.
- In the process dummy variables created

- GitHub URL




Landing success rate date wise

# EDA with SQL

- SQL queries performed to find out unique launch sites
- Total payload carried
- To find total number of successful missions
- Which booster version carried max payload
- To find out successful landings between set dates.

- GitHub URL

# Build an Interactive Map with Folium

- Marked all launch sites, & added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- Assigned the feature launch outcomes (failure or success).
- Using the color-labeled marker clusters identified which launch sites have high success rate.
- Calculated the distances between a launch site to its proximities.
- Are launch sites near railways, highways &coastlines.

- GitHub URL

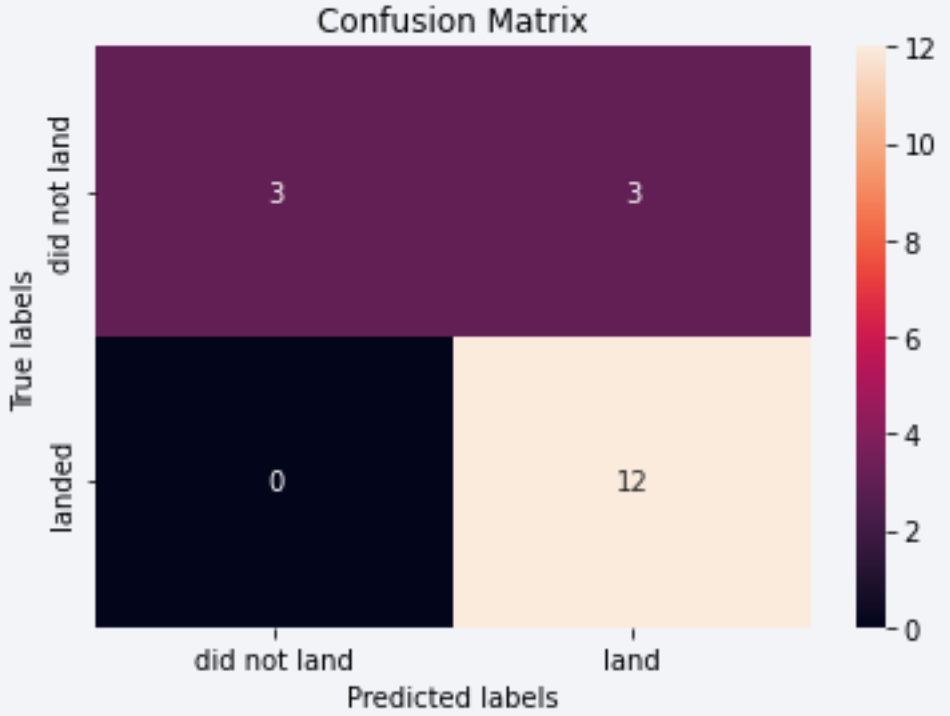
# Build a Dashboard with Plotly Dash

- Built an interactive dashboard with Plotly dash
- Plotted pie charts showing the total launches by a certain sites
- Plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version Explain why you added those plots and interactions

- GitHub URL

# Predictive Analysis (Classification)

- Loaded the data using Numpy and pandas. Transformed the data, split our data into training and testing.
- Built different machine learning models and tune different hyperparameters using GridSearchCV.
- Used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- Classification models such as LR, KNN, SVM and DT are used for analysis.

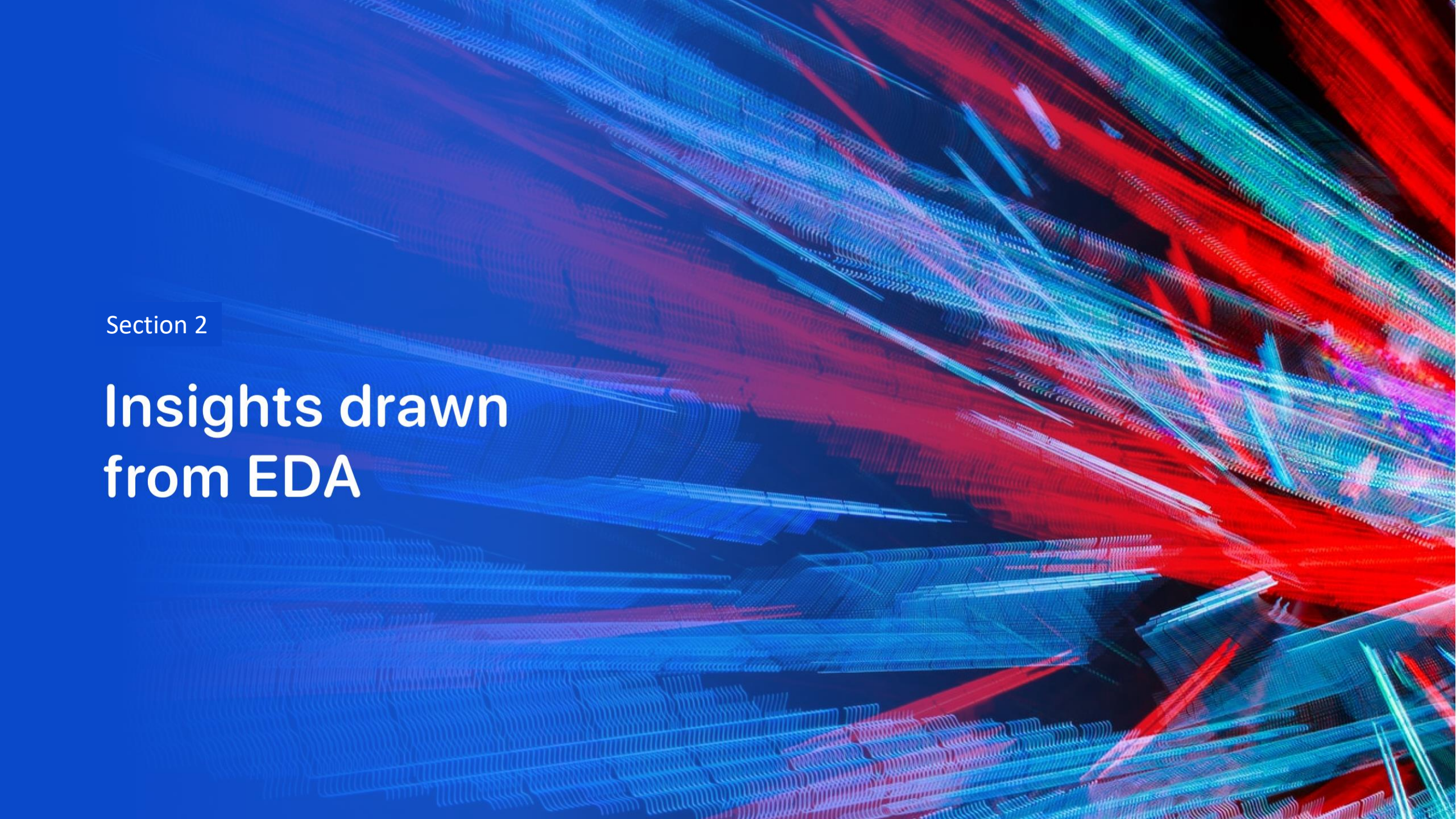



KNN Model Analysis

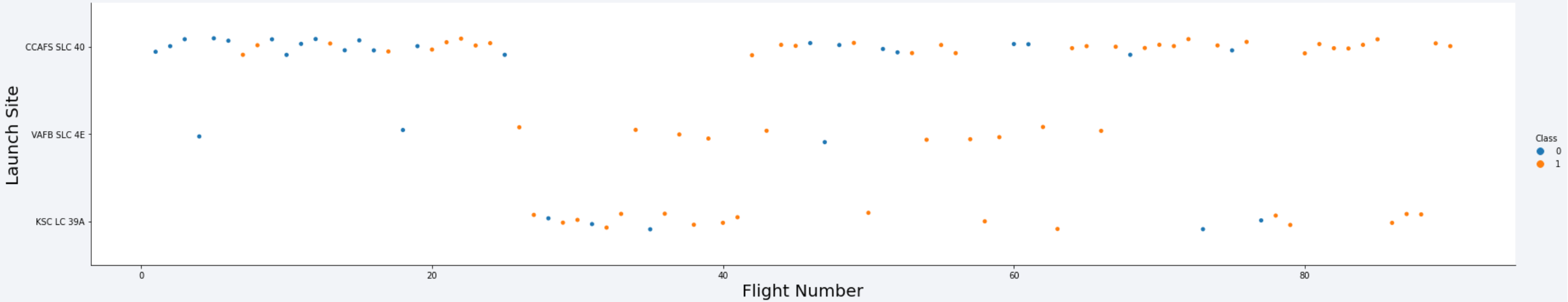
- [GitHub URL]

# Results

- Low payloads are better than heavier payloads.
- KSC LC 39A has highest success rate.
- If payload goes below 7500kg. It has significantly greater launches.
- Success rate has increased greatly with respect to the time.
- KNN, SVM and Logistic Regression models are well suited with KNN model giving high efficiency.

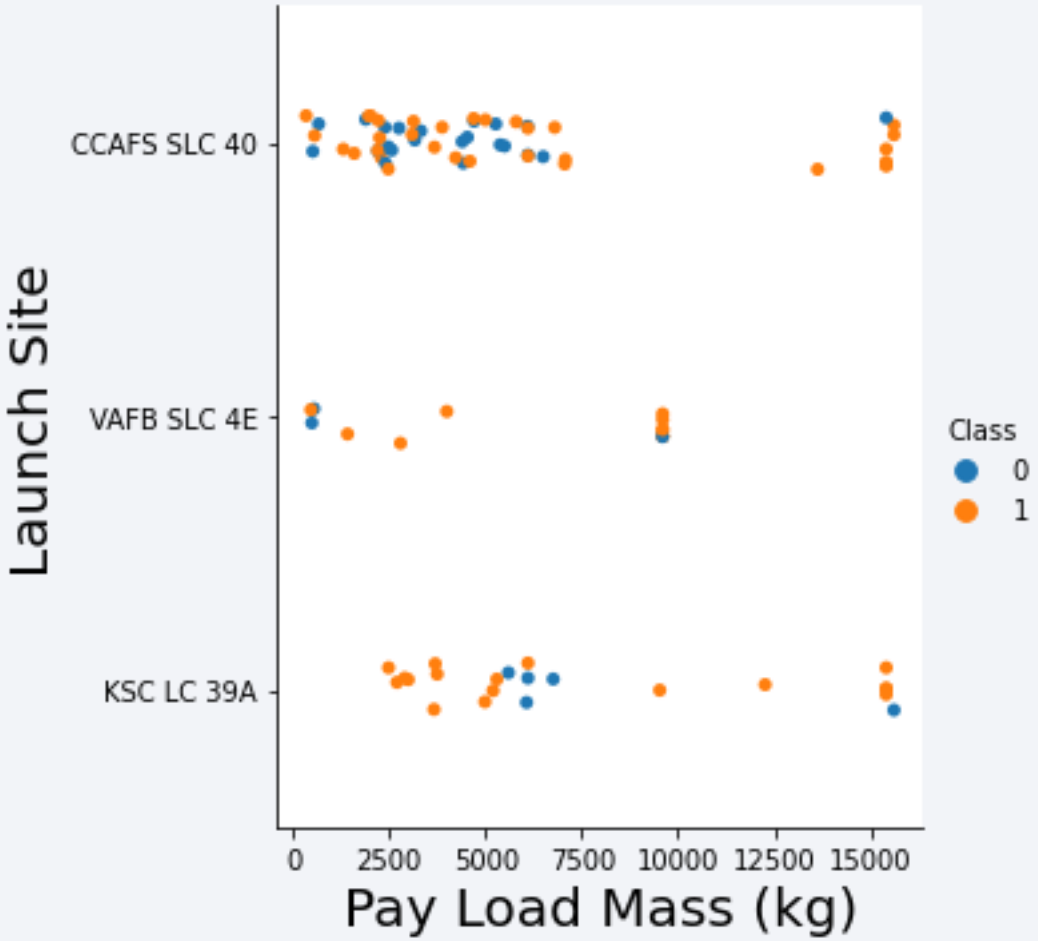

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site




- CCAFS SLC 40 is showing significantly more launches compared to others

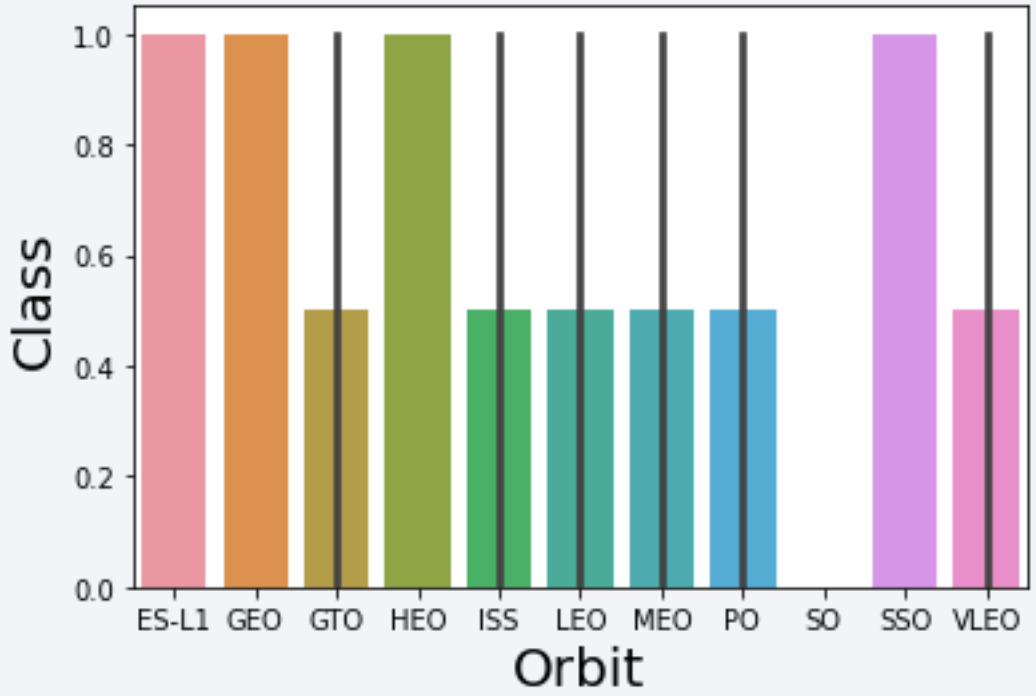
# Payload vs. Launch Site



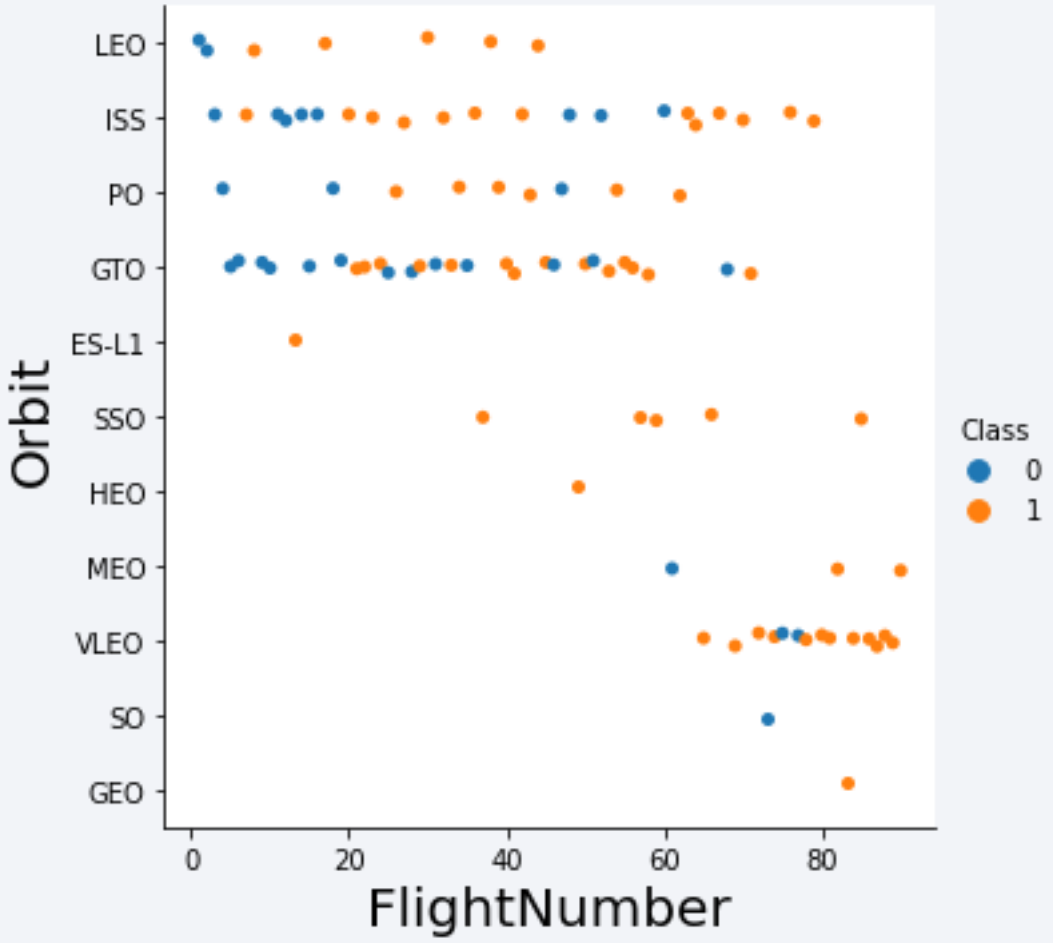

- Pay load of less than 7500 kg has significantly hire launches.
- Majority of the CCAFS SLC 40 launches are of low pay loads.

# Success Rate vs. Orbit Type

- ES-L1, GEO, HEO & SSO has highest success rates



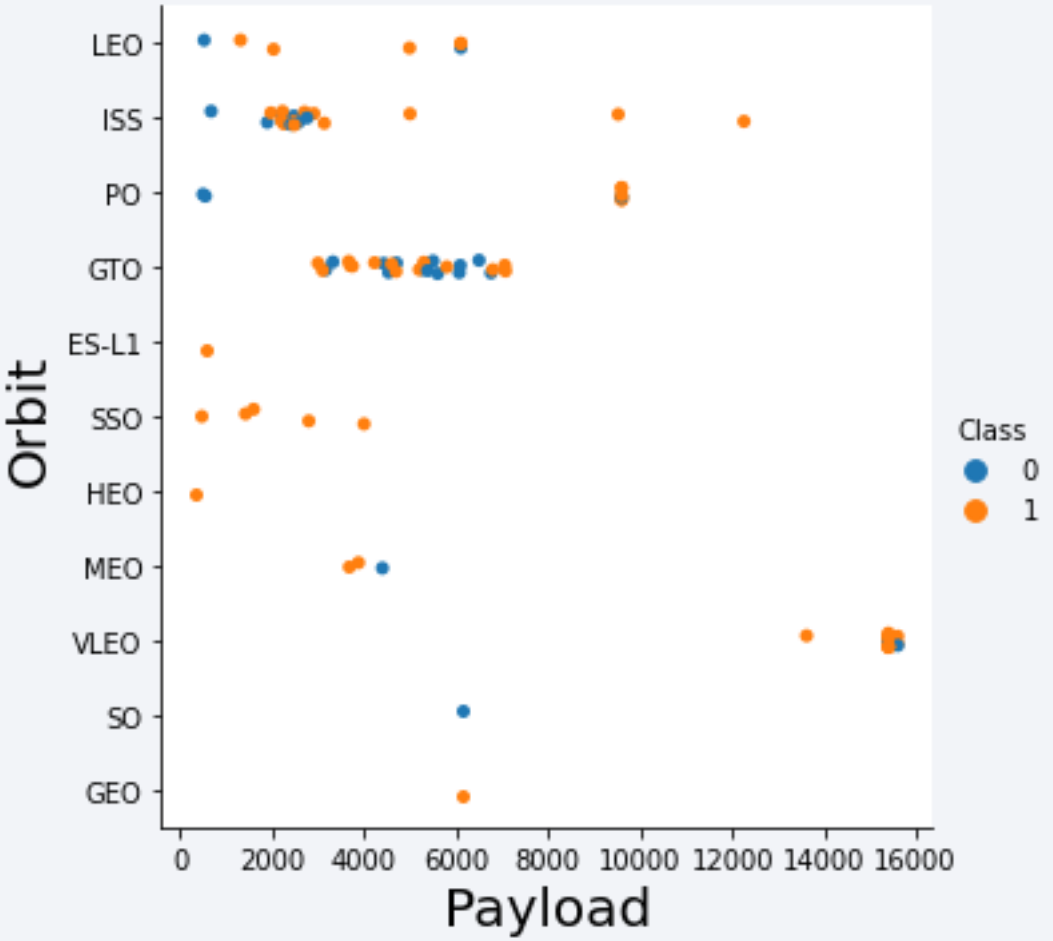
# Flight Number vs. Orbit Type




- VLEO has seen significantly hire flight numbers
- This trend is due to SpaceX new low earth orbit satellite internet project

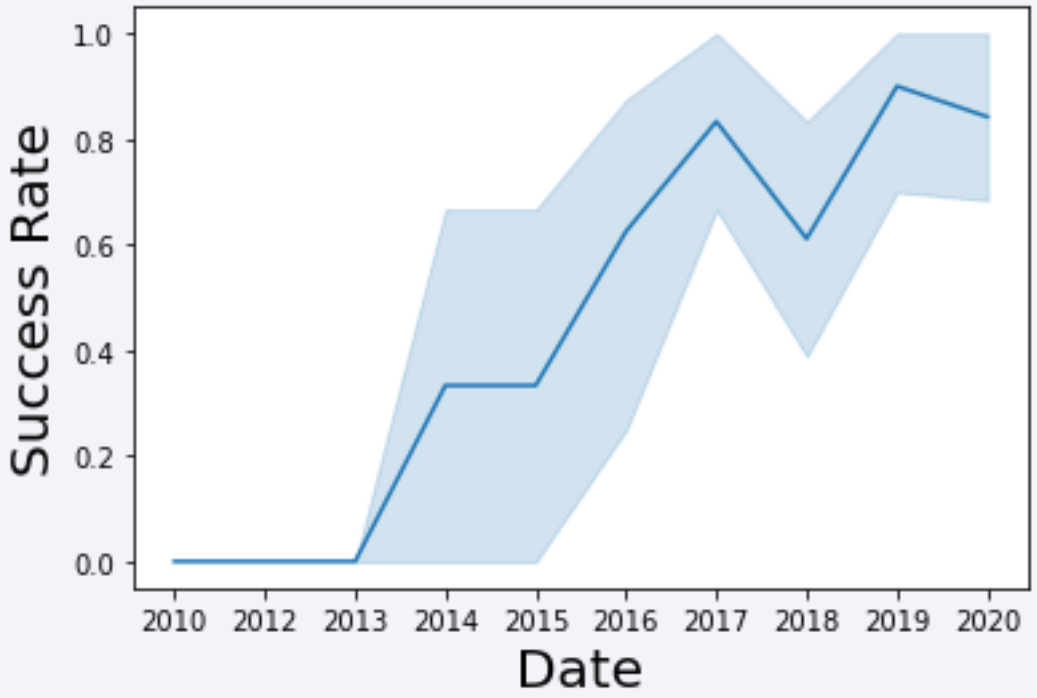
# Payload vs. Orbit Type

- GTO has seen launches with the payload of  2000kg to 8000kg
- ISS has launches with payload concentration around 2000kg



# Launch Success Yearly Trend

- Launch success rate has increased significantly from 2013 to 2020



# All Launch Site Names

In [7]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

Out[7]:

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

Using distinct operator in SQL table is formed

# Launch Site Names Begin with 'CCA'

- Special query with % sign at end of CCA enabled to find the results

In [10]:
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

Out[10]:

| 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

```
In [11]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[11]: sum(PAYLOAD_MASS__KG_)
                          45596
```

- Total payload is around 45596 kg

# Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2928.4 kg

```
In [19]: %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_version == 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
Out[19]: AVG(PAYLOAD_MASS__KG_)
         2928.4
```

# First Successful Ground Landing Date

```
In [43]: %sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'

 * ibm_db_sa://sdk38546:***@dashdb-txn-sbox-yp-lon02-07.services.eu-gb.bluemix.net:50000/BLUDB
Done.

Out[43]:        1

        2015-12-22
```

- First successful Landing is done on 22-12-2015

# Successful Drone Ship Landing with Payload between 4000 and 6000

In [46]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * ibm_db_sa://sdk38546:***@dashdb-txn-sbox-yp-lon02-07.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[46]:

| 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 |

- List of the names. Used subquery in SQL for the results

# Total Number of Successful and Failure Mission Outcomes

- Used count object to find the results

```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

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

| count(MISSION_OUTCOME) |
|---|
| 99 |

# Boosters Carried Maximum Payload

```
%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 |

- Given above is the list of boosters with max payload

# 2015 Launch Records

- Total of 2 outcomes has come from running the query

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

In [18]:
```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

| | boosterversion | launchsite | landingoutcome |
|---|---|---|---|
| 0 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 1 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

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

- With the help of count and order by results were calculated

```
%sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

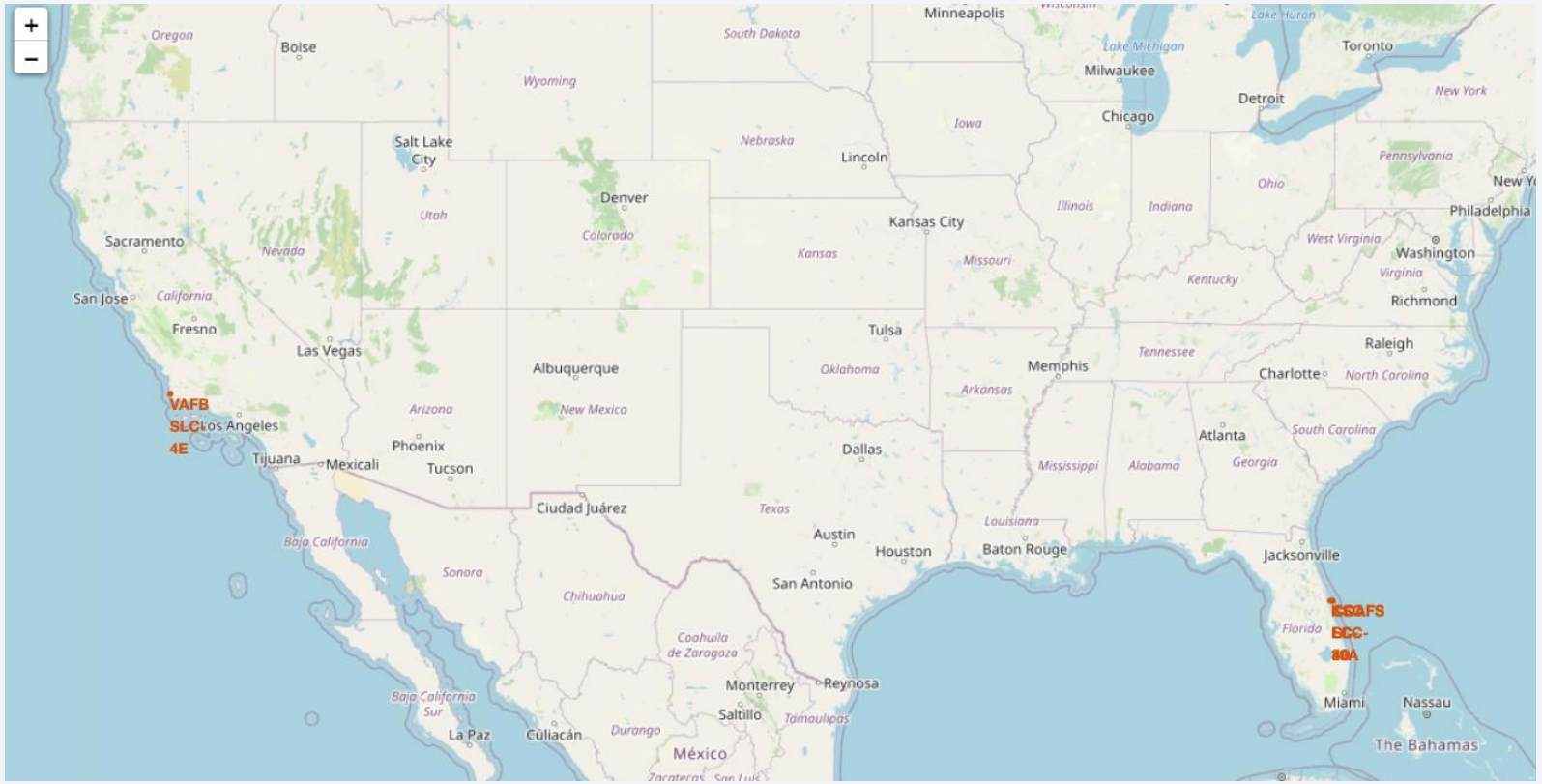
```
 * ibm_db_sa://sdk38546:***@dashdb-txn-sbox-yp-lon02-07.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

| DATE | time__utc_ | booster_version | launch_site | payload | payload_mass__kg_ | orbit | customer | mission_outcome | landing__outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2017-02-19 | 14:39:00 | F9 FT B1031.1 | KSC LC-39A | SpaceX CRS-10 | 2490 | LEO (ISS) | NASA (CRS) | Success | Success (ground pad) |
| 2017-01-14 | 17:54:00 | F9 FT B1029.1 | VAFB SLC-4E | Iridium NEXT 1 | 9600 | Polar LEO | Iridium Communications | Success | Success (drone ship) |
| 2016-08-14 | 05:26:00 | F9 FT B1026 | CCAFS LC-40 | JCSAT-16 | 4600 | GTO | SKY Perfect JSAT Group | Success | Success (drone ship) |
| 2016-07-18 | 04:45:00 | F9 FT B1025.1 | CCAFS LC-40 | SpaceX CRS-9 | 2257 | LEO (ISS) | NASA (CRS) | Success | Success (ground pad) |
| 2016-05-27 | 21:39:00 | F9 FT B1023.1 | CCAFS LC-40 | Thaicom 8 | 3100 | GTO | Thaicom | Success | Success (drone ship) |
| 2016-05-06 | 05:21:00 | F9 FT B1022 | CCAFS LC-40 | JCSAT-14 | 4696 | GTO | SKY Perfect JSAT Group | Success | Success (drone ship) |
| 2016-04-08 | 20:43:00 | F9 FT B1021.1 | CCAFS LC-40 | SpaceX CRS-8 | 3136 | LEO (ISS) | NASA (CRS) | Success | Success (drone ship) |
| 2015-12-22 | 01:29:00 | F9 FT B1019 | CCAFS LC-40 | OG2 Mission 2 11 Orbcomm-OG2 satellites | 2034 | LEO | Orbcomm | Success | Success (ground pad) |

Section 3

# Launch Sites Proximities Analysis

# SpaceX launch site locations in USA

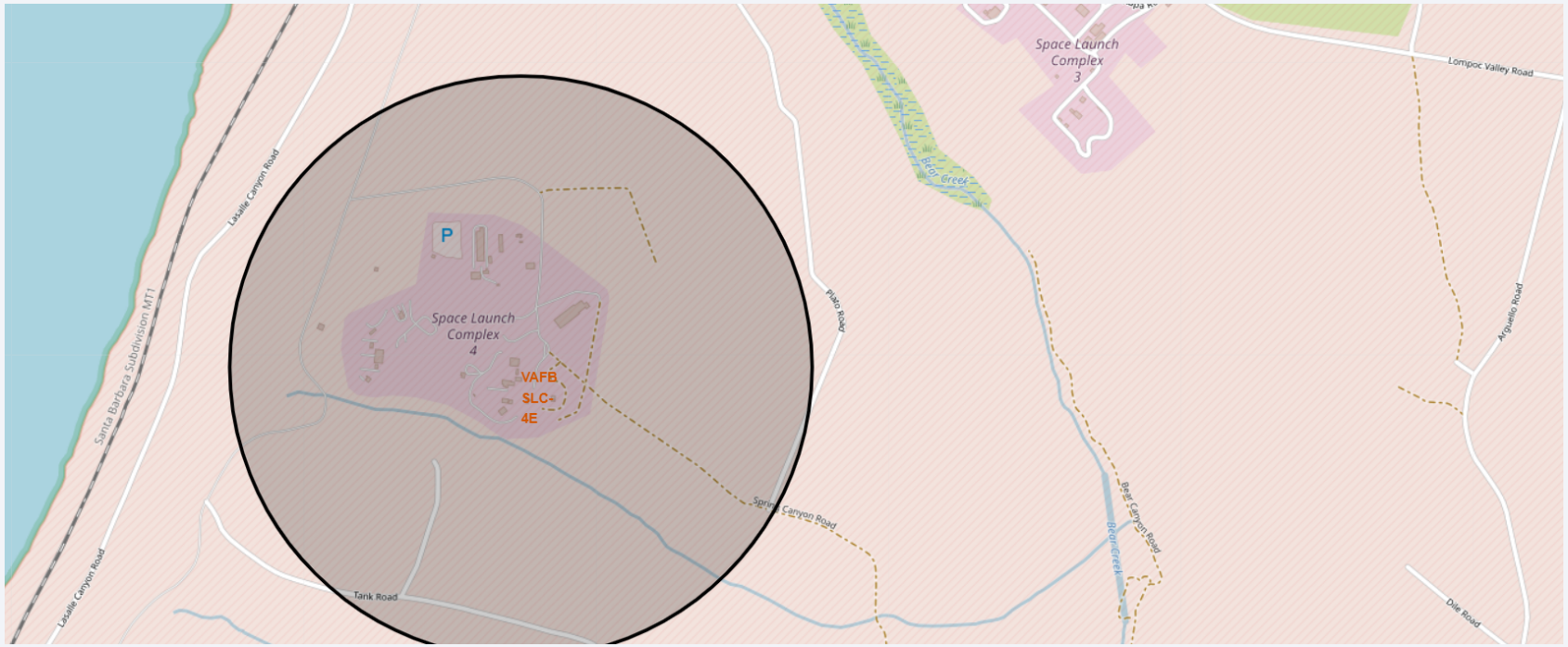



There are 2 main launch sites in USA

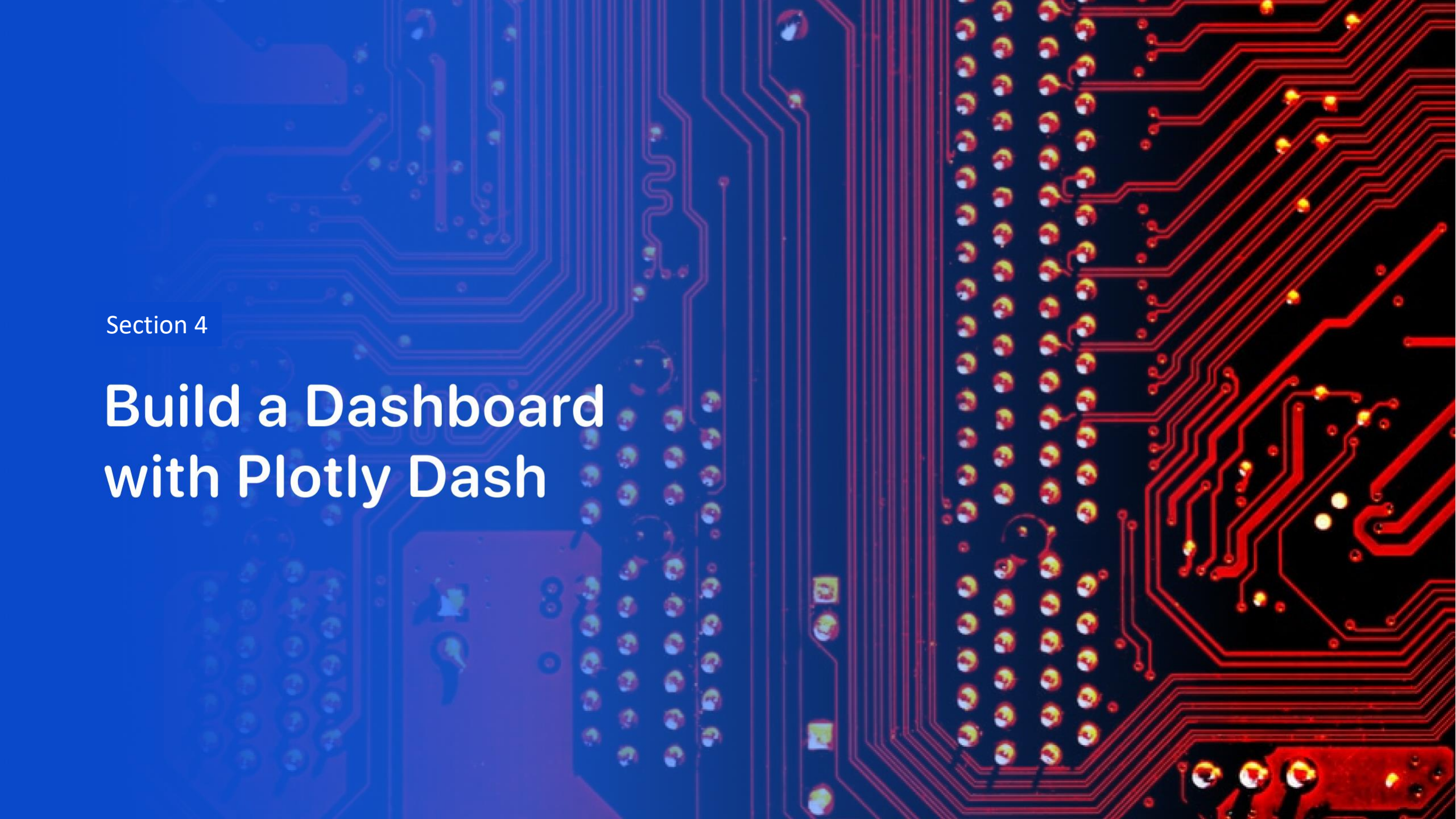
# Launch site cluster for CCAFS SLC 40


Zoomed area of launch side CCAFS SLC 40 with cluster

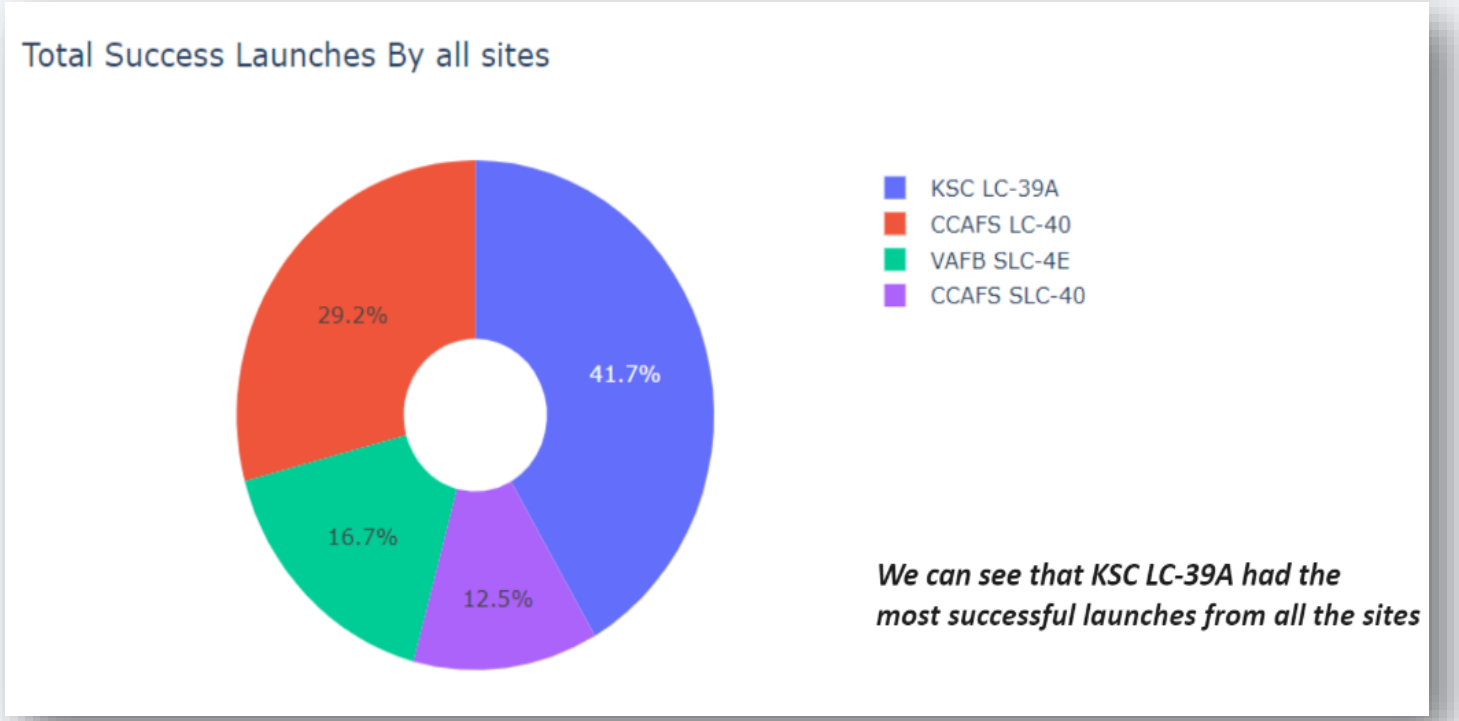
# Launch site VAFBSLC 4E with roads




Launch side VAFB SLC 4E with surrounding roads and coastal line

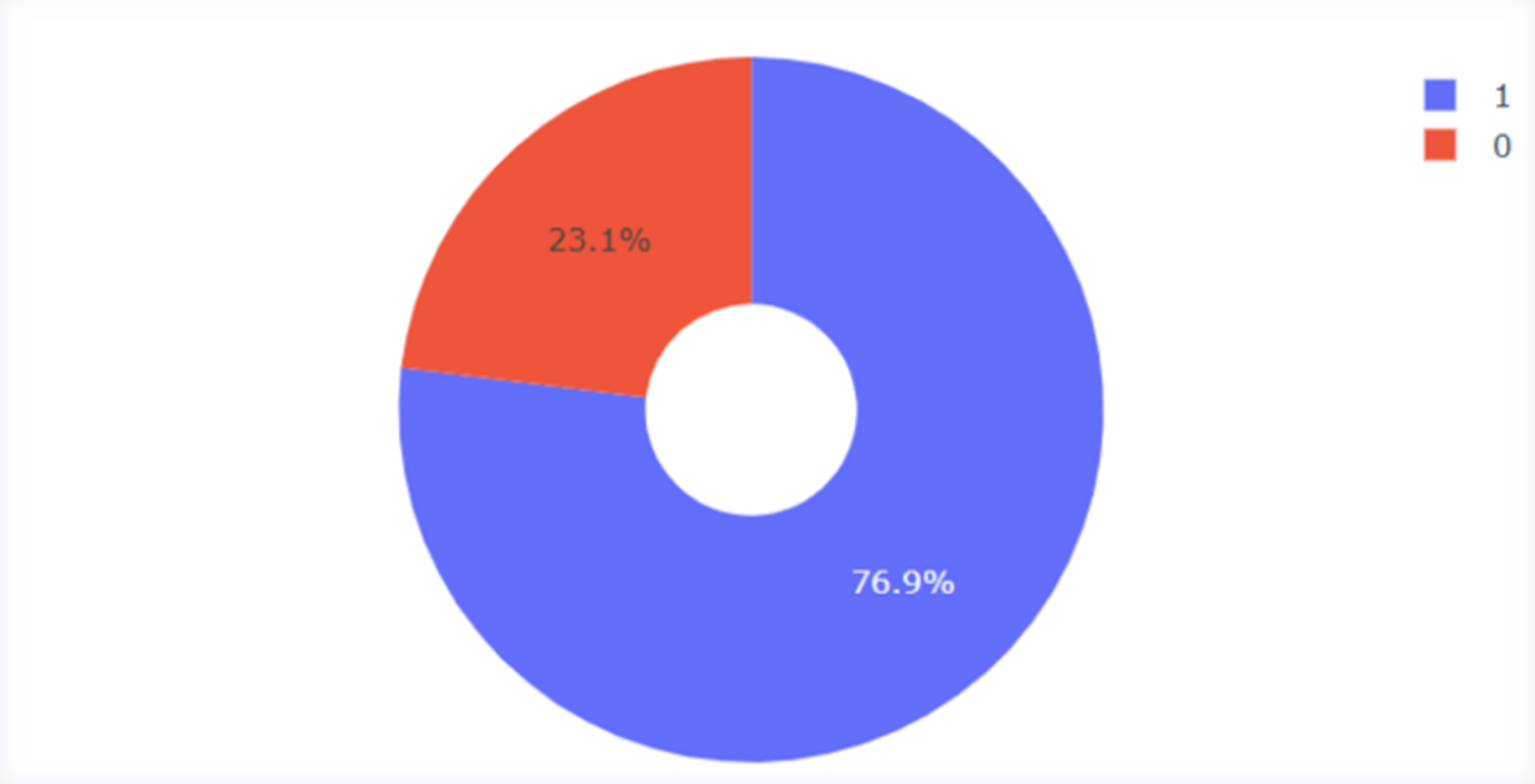

Section 4
Build a Dashboard with Plotly Dash

# Success launches site wise



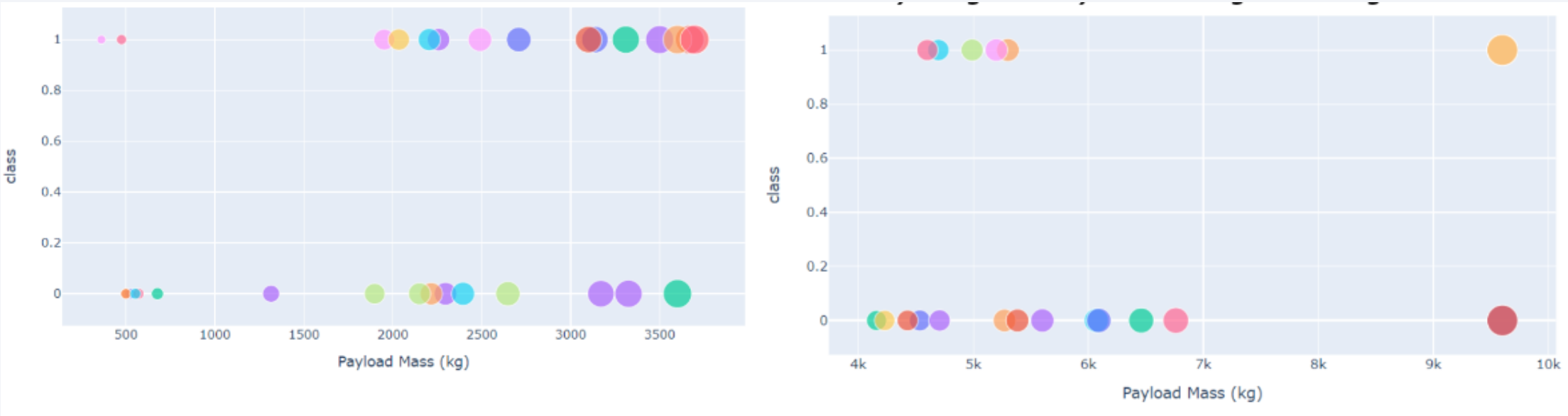

KSL LC 39A has most success launches

# Highest successful launch ratio




- KSL LC 39A has highest successful launch ratio

# Payload vs Success Rate

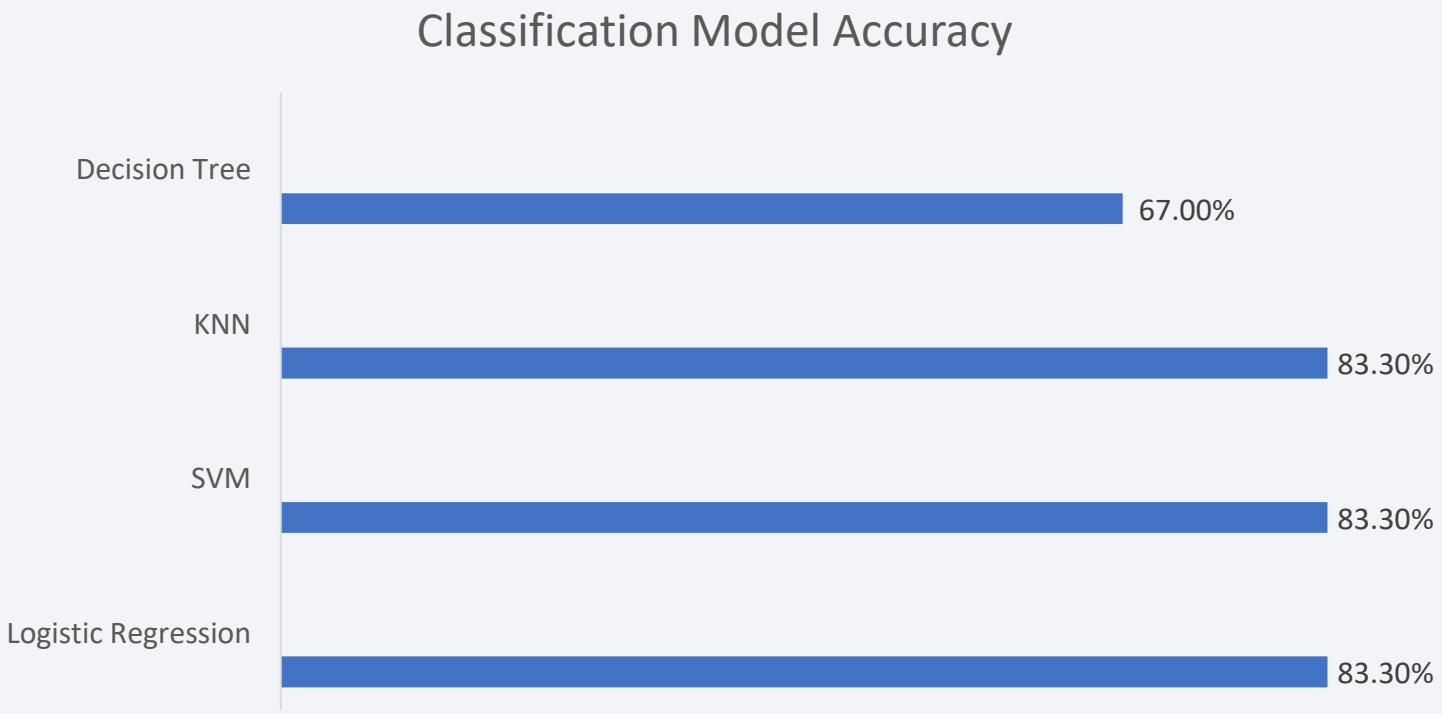



- Success rate for low payload launch rates is higher

Section 5

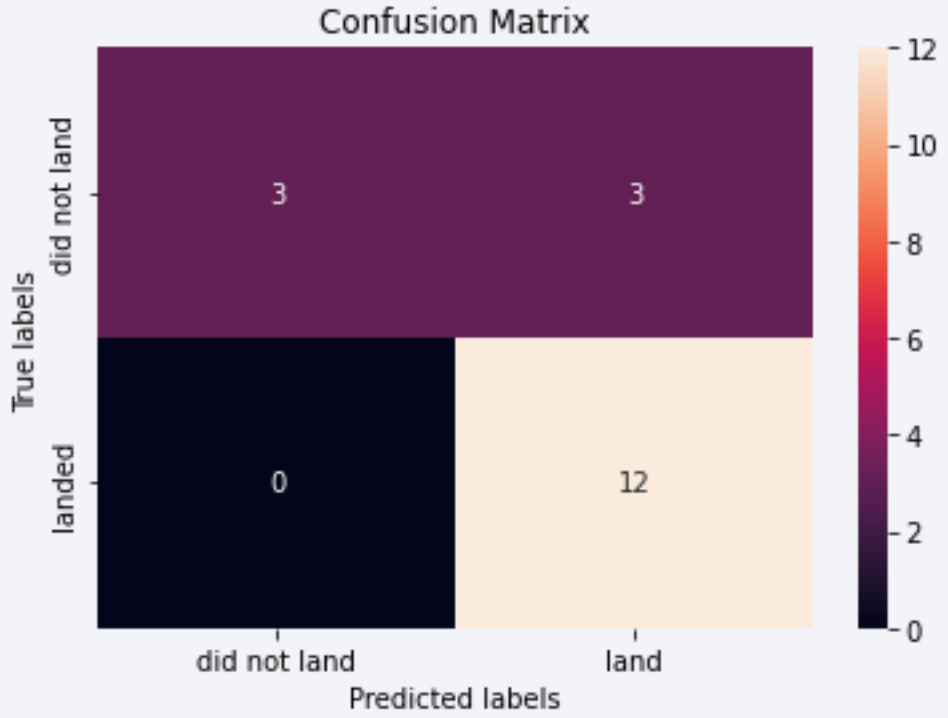
# Predictive Analysis (Classification)

# Classification Accuracy




KNN, SVM & Logistic Regression models have highest accuracy

# Confusion Matrix




Above Confusion Matrix is of KNN since it has maximum accuracy of 83.3%

# Conclusions

- Orbits GEO, HEO, SSO, ES L1has highest success rate
- Success rate of launches is directly proportional to the time
- Launches with the low payload have highest success rate compared to heavy payload
- KSC LC 39A has highest success rate
- Prediction can be performed using KNN, SVM and Logical Regression models as they have highest accuracy of more than 83%

# Thank you!