IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

H S Gowri Yaamini
20 March 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
    1. Data Collection
    2. Data Wrangling
    3. EDA with SQL
    4. EDA with Data Visualization
    5. Visual Analytics with Folium
    6. Visual Analytics with Dashboard
    7. Predictive Modeling
- Summary of all results
    1. EDA results
    2. Visual Analysis
    3. Predictive Analysis

Introduction

- Background

SpaceX has achieved major milestones, such as sending spacecraft to the ISS, launching Starlink, and manned missions to space.

A key factor in their cost-efficiency is the reuse of the Falcon 9's first stage, which significantly reduces launch costs ($62 million compared to competitors' $165 million). However, the first stage doesn't always land successfully, and its reuse depends on mission parameters like payload and orbit. Predicting whether the first stage will land is crucial for determining the cost of each launch.

- Business Problem

SpaceY founded by Allon Musk, an emerging company competing with SpaceX needs to draw actionable insights from predicting whether SpaceX will reuse the Falcon 9’s first stage using machine learning, based on public data. This will help estimate launch costs and create SpaceY's competitive strategy.

Section 1

Methodology

Methodology

Executive Summary

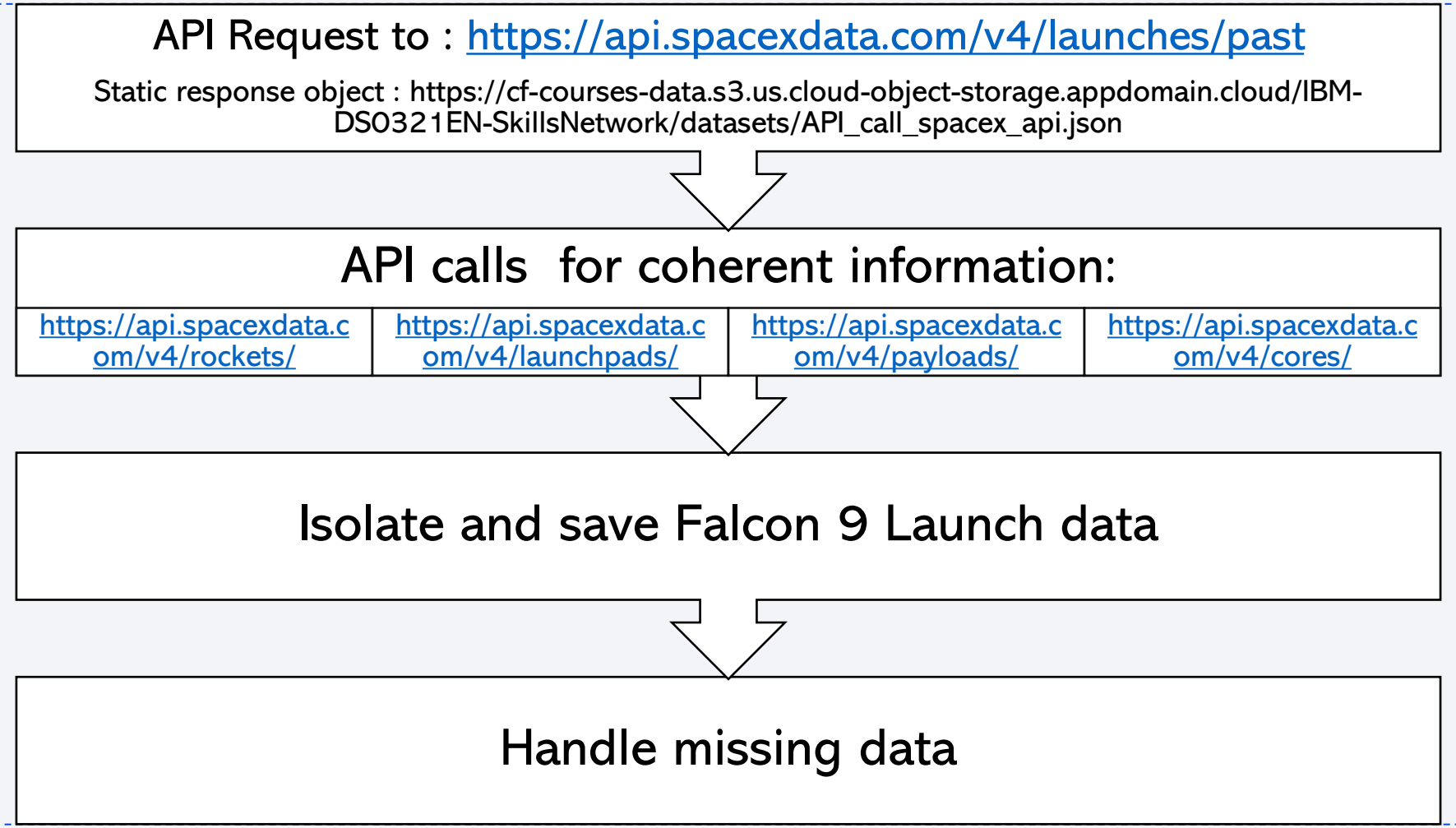
- Data collection methodology:
  - Open-Source SpaceX API
  - Web-Scraping Falcon-9 Heavy Launch records from Wikipedia
- Perform data wrangling
  - Handling missing values, feature engineering, label encoding and standardizing data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models - Logistic Regression, SVM. Decision Tree, KNN with Grid Search Cross Validation
  - Evaluation for most effective predictive model

Data Collection

Data collection is done using two main steps

1. Open-source SpaceX API
   - Collection of SpaceX launch data through their API
   - Retrieval of rocket launch data and information such as rockets, payloads, launchpad and cores using IDs with corresponding API GET requests
   - Filtering only Falcon-9 launch data
   - Handling of missing values
2. Web-scraping from Wikipedia
   - Gathering data from ‘List of Falcon 9 and Falcon Heavy launches` Wikipedia page
   - Extraction of column names from HTML header
   - Parsing data to pandas data frame after handling inconsistent and incomplete data

Data Collection – SpaceX API



GitHub : Module1 01-spacex-Data-Collection-api.ipynb

Data Collection - Scraping

Collect data from Wikipedia Page:
'List of Falcon 9 and Falcon Heavy launches`

Extract column names from HTML header

'Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome'

Parse launch HTML table to Pandas data frame

Parser for – data inconsistency, data incompleteness, data formatting

GitHub : Module1 02 - spacex- webscraping.ipynb

Data Wrangling

1. Handling Missing Values: Mean Imputation
2. Feature Engineering: Target class ‘class’ from launch outcomes
3. Categorical data encoding: One Hot Encoder
4.
   - Data type conversion: Numerical values to ‘float’
   - Scaled values: Standard Scaler

GitHub :

Module1 01-spacex-Data-Collection-api.ipynb

Module1 03 - spacex-Data wrangling.ipynb

Module2 02 - EDA-pandas matplotlib.ipynb

EDA with Data Visualization

Scatterplot (Seaborn Catplot)

- Plot 1: Flight Number v/s Payload Mass
- Plot 2: Flight Number v/s Launch Site
- Plot 3: Payload Mass v/s Launch Site
- Plot 4: Flight Number v/s Orbit Type
- Plot 5: Payload Mass v/s Orbit Type

➔ VAFB SLC 4E – no launch with heavy payload

➔ LEO, ISS, SSO, VLEO orbits– positive success rates with increase in payload number of flights

Bar Chart (Matplotlib Bar Chart)

- Plot 1: Success rate of each Orbit

➔ ES-L1, DEO, HEO and SSO have highest success rates

Line Chart (Matplotlib plot)

- Plot 1: Average Success trend over the years

➔ From 2013 the success rate has increased

GitHub : Module2 02 - EDA-pandas matplotlib.ipynb

EDA with SQL

SQL queries performed to draw insight:

- Display the names of unique launch sites in the space mission
- Display 5 records where launch site begins with “CCA”
- Display total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of boosters which have success in drone ship and have payload mass between 4000kg and 6000kg
- List total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the records which will display month names, failure landing outcomes in drone ship, booster versions and launch site for the months in year 2015
- Rank of count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

GitHub : Module2 01 - EDA-SQL.ipynb

Build an Interactive Map with Folium

Map Objects used:

Circle

- To highlight launch site area
- With marker cluster the highlighted area would be colour coded based on relative success rate

Marker

- To mark launch sites
- With marker cluster the highlighted area would include colour coded icons for each or success or failure of launch
- To mark the distance between two coordinates

PolyLine

- To draw a path/line between two coordinates e.g.: between a launch site and coastline/ railroad/ nearest city

GitHub : Module3 01 - launch_site_location.ipynb

Rendered Folium Maps : Visual Analytics with Folium

Build a Dashboard with Plotly Dash

Plots

Pie Chart

- Launch site all : Percentage share of success rate of each launch site
- Specific launch site : Percentage share of Success or Failure of launch

Scatter Plot

- Payload vs Success rate for each launch site or a specific launch site opted
- Payload mass between 0Kg – 10000Kg, tuned according to user interaction

Interactions

Dropdown menu

- Options for selecting all launch sites or a specific launch site

Range Slider

- Ranges from 0Kg – 10000Kg at 2500Kg intervals

GitHub : Module3 02 - spacex_dash_app.py

Predictive Analysis (Classification)

Preprocessing

- Features and Target is separated
- Features are standardized using Standard Scaler
- Data corpus is split into train set 80% and test set 20%

Model Building

Selected models:

- Logistic Regression
- SVM
- Decision Tree
- KNN

Each model is built with Grid Search Cross Validation using training data

Each with 10 folds and predefined hyperparameters

Hyperparameter Tuning

Each model is retrained with best hyperparameters given by Grid Search model

Model Evaluation using Metrics

Performance evaluation based on:

- Precision
- Recall
- F1-Score
- Accuracy

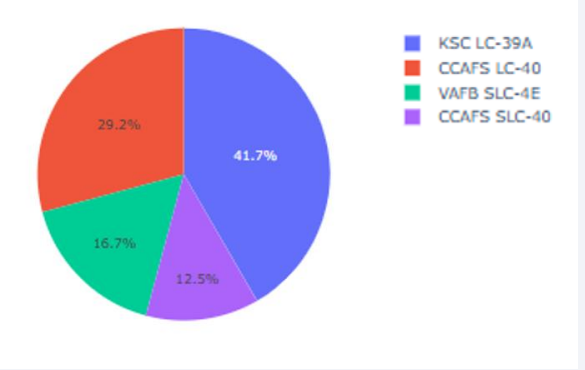
LR, SVM, KNN: 83% test accuracy

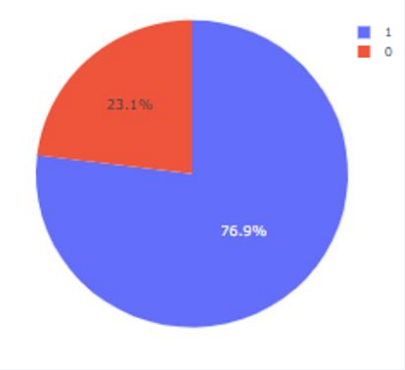
Decision Tree: 78% test accuracy

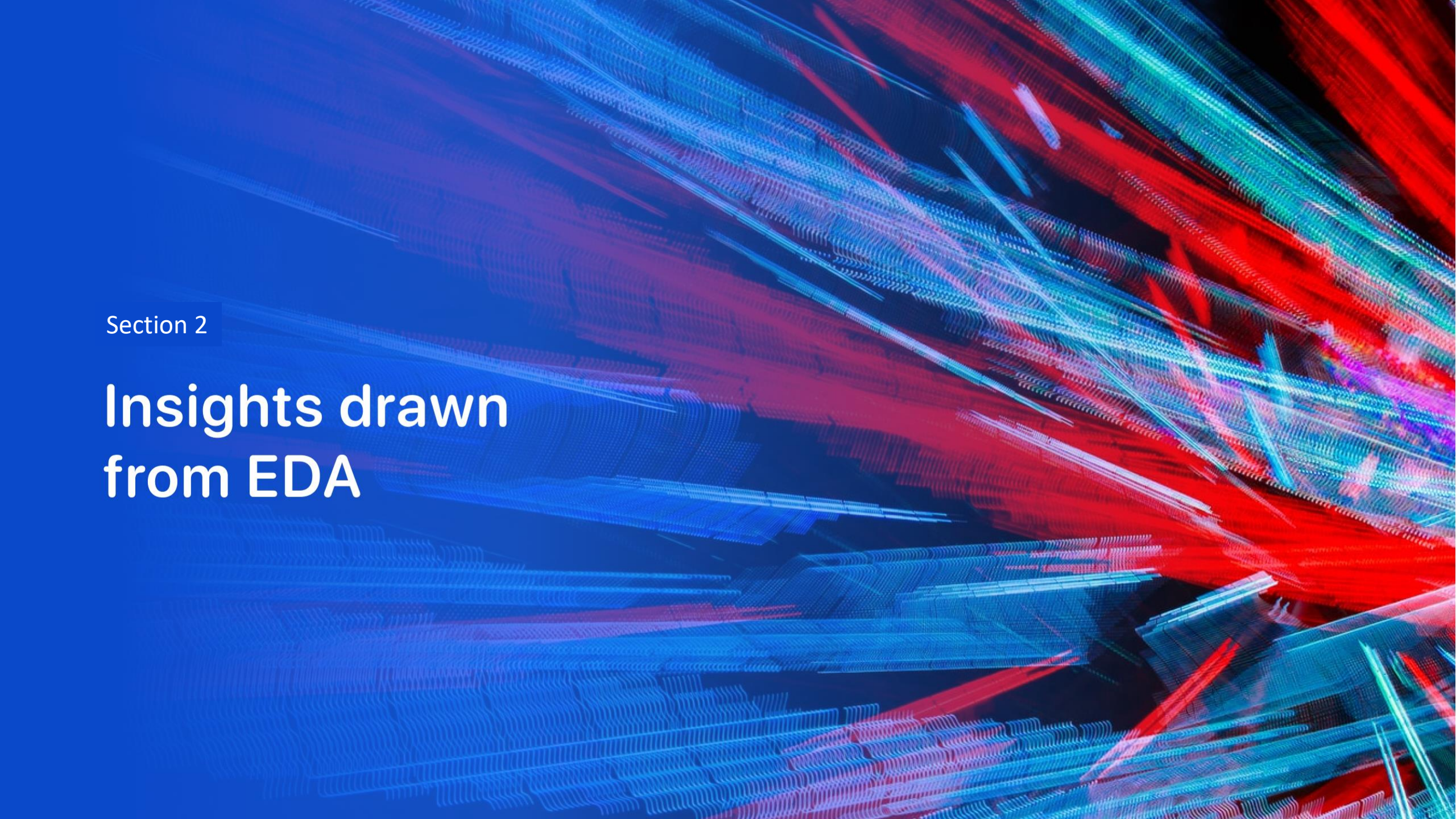
GitHub : Module4 - SpaceX_Machine Learning Prediction.ipynb

Results

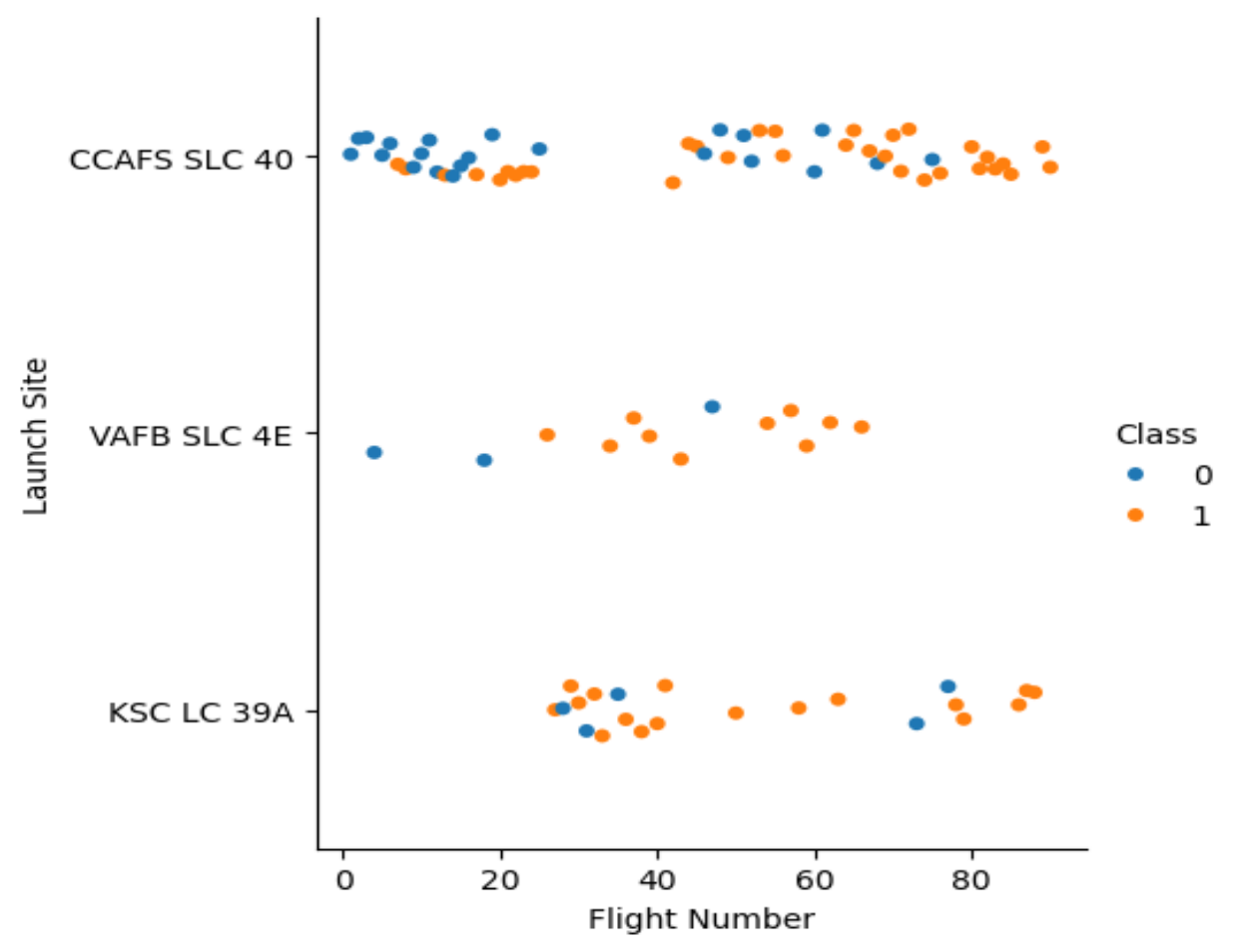
- Exploratory data analysis results
  - VAFB SLC 4E launch site has positive success rate as the number of flights increases and it has no rocket launched with heavy payload
  - ES-L1, DEO, HEO and SSO orbits have highest success rates
  - LEO, ISS, SSO, VLEO orbits have positive success rates with increase in payload and an increase in number of flights
- Interactive analytics
  - KSC LC 39A has highest success rate
  - 79.5% successful launch
- Predictive analysis results
  - Logistic Regression, SVM and KNN have performed equally well with testing accuracy of 83%





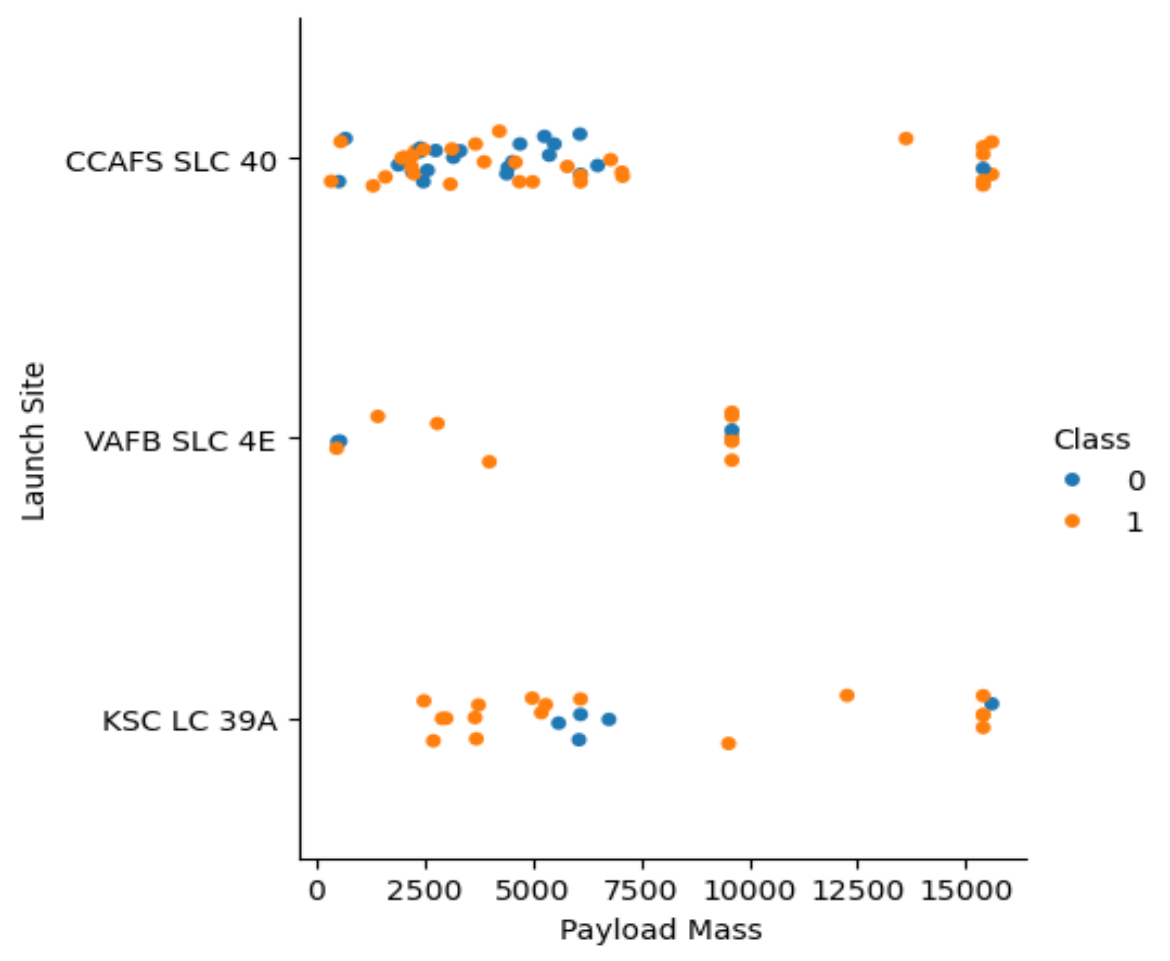

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



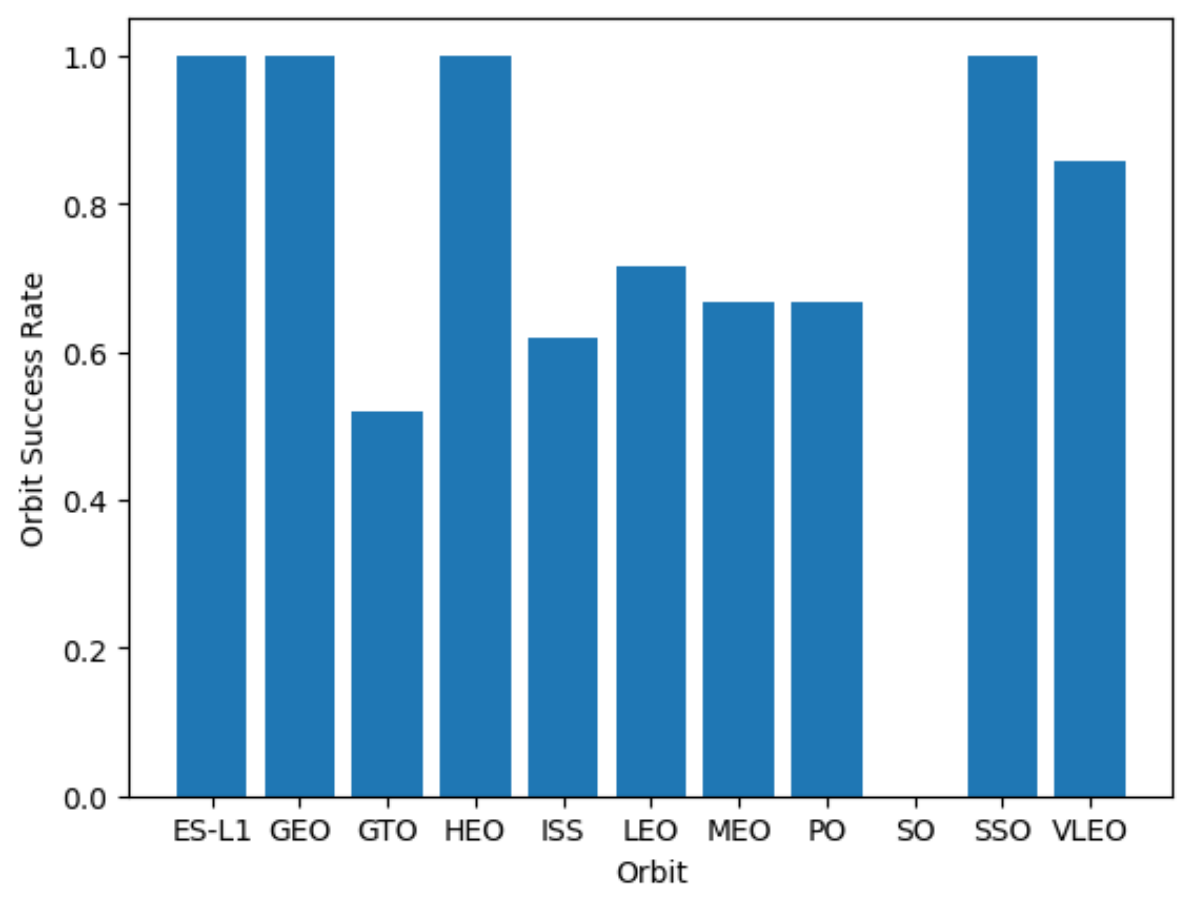
- Success rate increases as the number of flights is increased
- CCAFS SLC 40 launch site has more failed launches than others and the number of launches is higher at this site

Payload vs. Launch Site



- CCAFS SLC 40 has highest number of launches in the payload range 0Kg-7500Kg
- VAFB SLC 4E has no rocket launch with heavy payload
- CCAFS SLC 40 and KSC LC 39A has greater success rate with heavy payload rocket launch

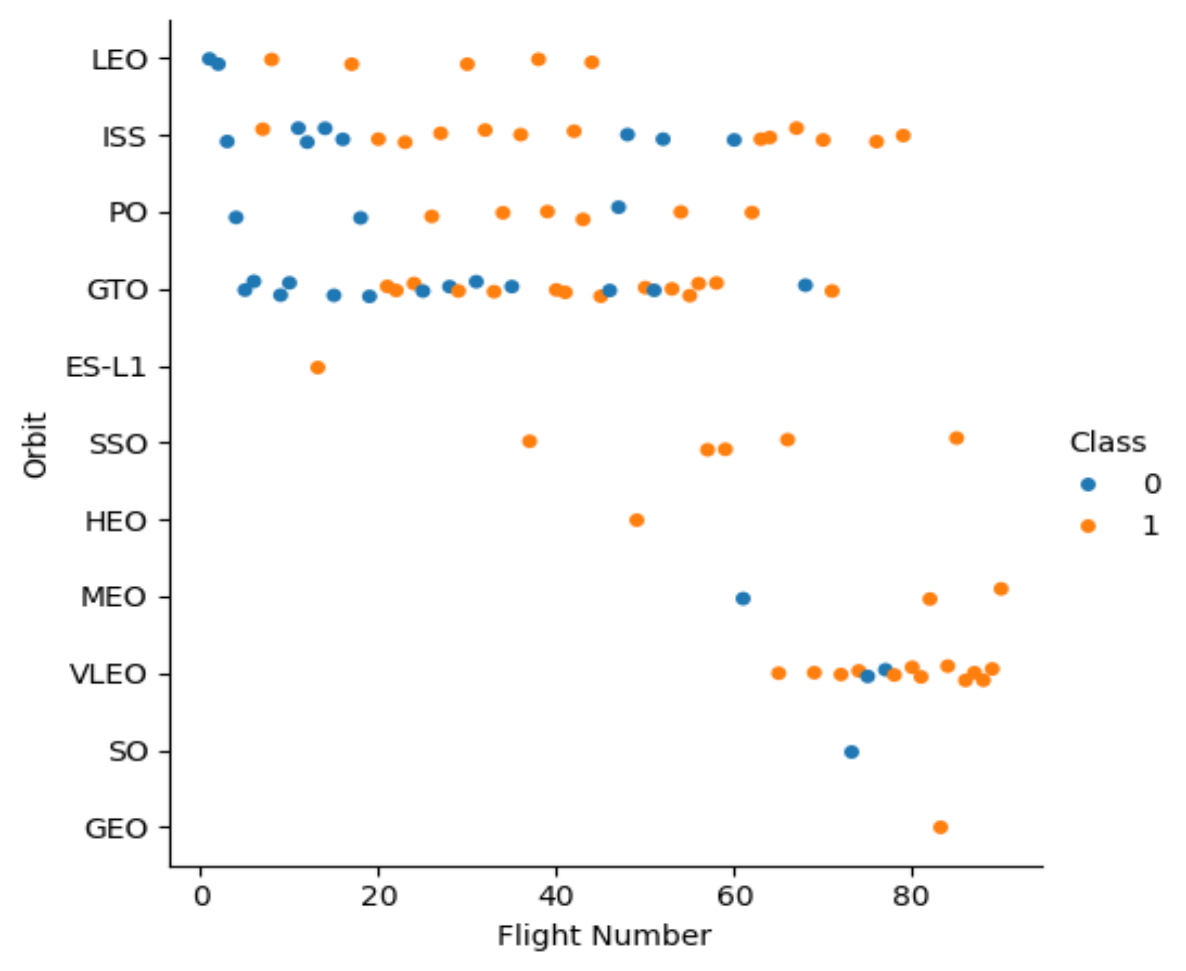
Success Rate vs. Orbit Type



- ES-L1, DEO, HEO and SSO orbits have highest success rates
- Every orbit is observed to have success rate of at least 50%

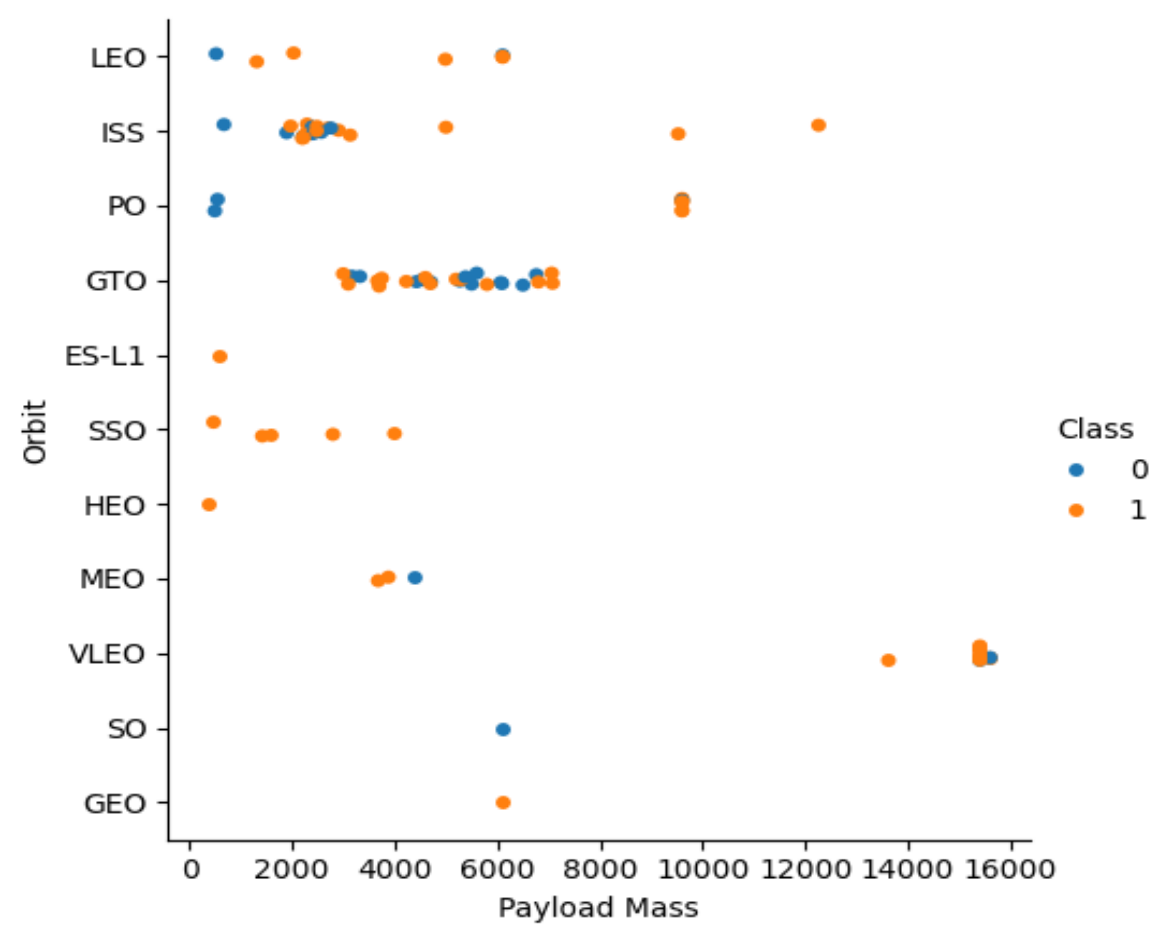
Note: SSO and SO are the same Sun-synchronous orbit

Flight Number vs. Orbit Type



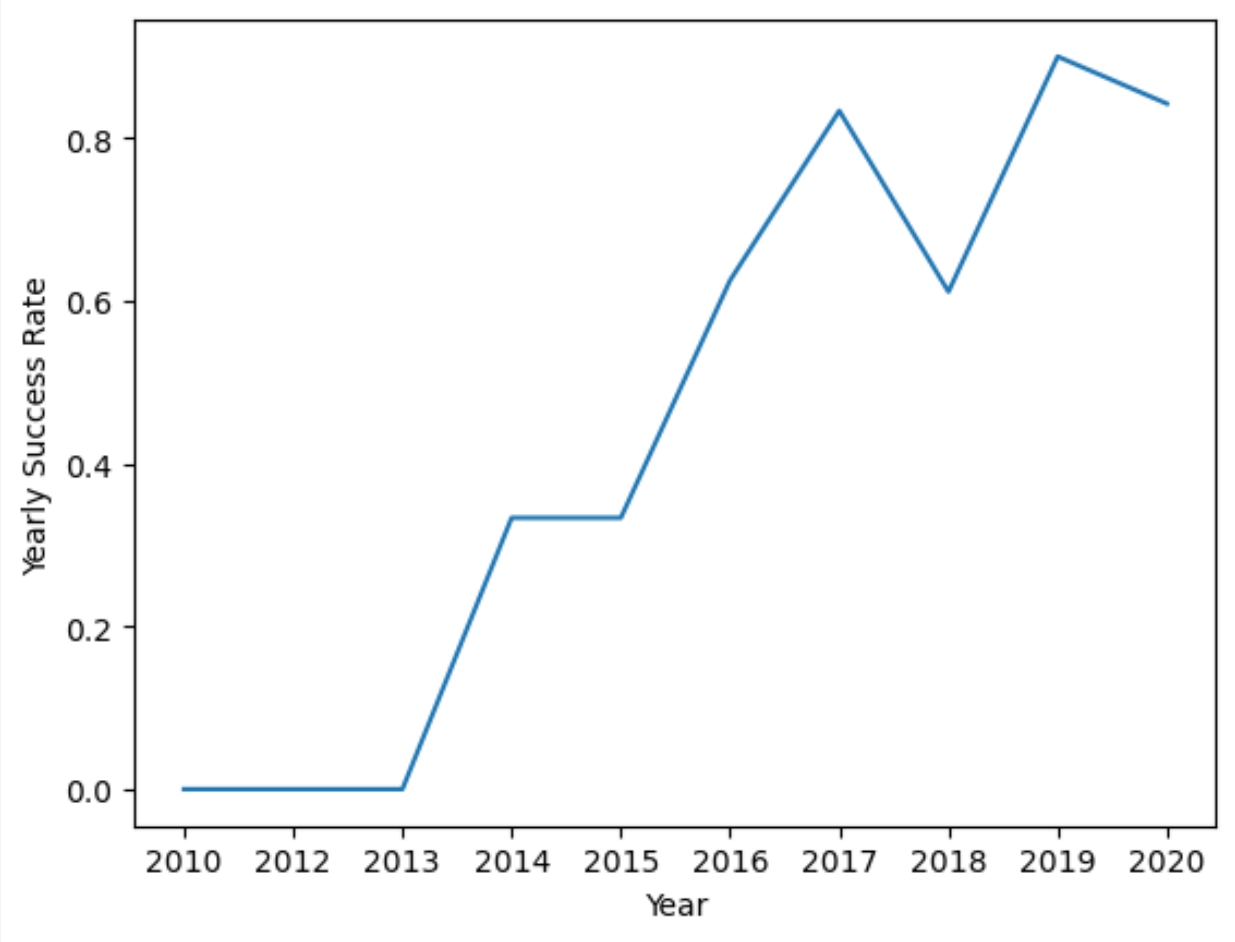
- LEO, ISS, SSO, VLEO orbits have positive success rates with increase number of flights
- Surge of launches to VLEO in the observed recent years

Payload vs. Orbit Type



- LEO, ISS, SSO orbits have positive success rates with increase in payload; however, LEO and SSO has no rocket launch with heavy payloads
- GTO has launches only between the range of 3000Kg-8000kg payload and the result is 50-50
- VLEO only has heavy payload launches and has a success of 75%

Launch Success Yearly Trend



- There has been constant increase in success rate since 2013
- A drop in 2017-2018 and improved performance since that point indicates a technological shift

All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

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

There are four distinct launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

There are two launch sites which begin with `CCA` and among two, the first five records are of CCAFS LC-40 launch site.

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_   ) as 'Total_Payload_Mass' from SPACEXTABLE where Customer='NASA (CRS)'
```

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

Total_Payload_Mass
45596

The total payload carried by boosters from NASA is 45,596Kg.

Average Payload Mass by F9 v1.1

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

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

Average_Payload_Mass
2534.6666666666665

The average payload mass carried by booster version F9 v1.1 is 2534.66 Kg.

First Successful Ground Landing Date

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome='Success (ground pad)'
```

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

min(Date)
2015-12-22

The first successful landing outcome on ground pad was on 22nd December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version, Payload from SPACEXTABLE where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000
```

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

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

There are four boosters which have successfully landed on drone ship and had payload mass greater than 4000Kg and less than 6000Kg.

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome,count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

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

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

The total number of successful and failure mission outcomes:

- Success -100
- Failure - 1

Boosters Carried Maximum Payload

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ >= (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * 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

There are twelve booster which have carried the maximum payload mass.

2015 Launch Records

```
%sql select substr(Date, 6,2) as 'Month', (select Landing_Outcome from SPACEXTABLE where Landing_Outcome='Failure (drone ship)') as 'Landing_Outcome', Booster_Version, Launch_Site from SPACEXTABLI
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Failure (drone ship)	F9 v1.1 B1013	CCAFS LC-40
03	Failure (drone ship)	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1016	CCAFS LC-40
06	Failure (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Failure (drone ship)	F9 FT B1019	CCAFS LC-40

There seven records of Failure (drone ship) in the year 2015 and all the launches are from CCAFS LC-40 launch site. The month of April has two failures and July-November has no launch failures.

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

```
%sql select Landing_Outcome, count(Landing_Outcome) as 'counts' from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order
```

Python

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

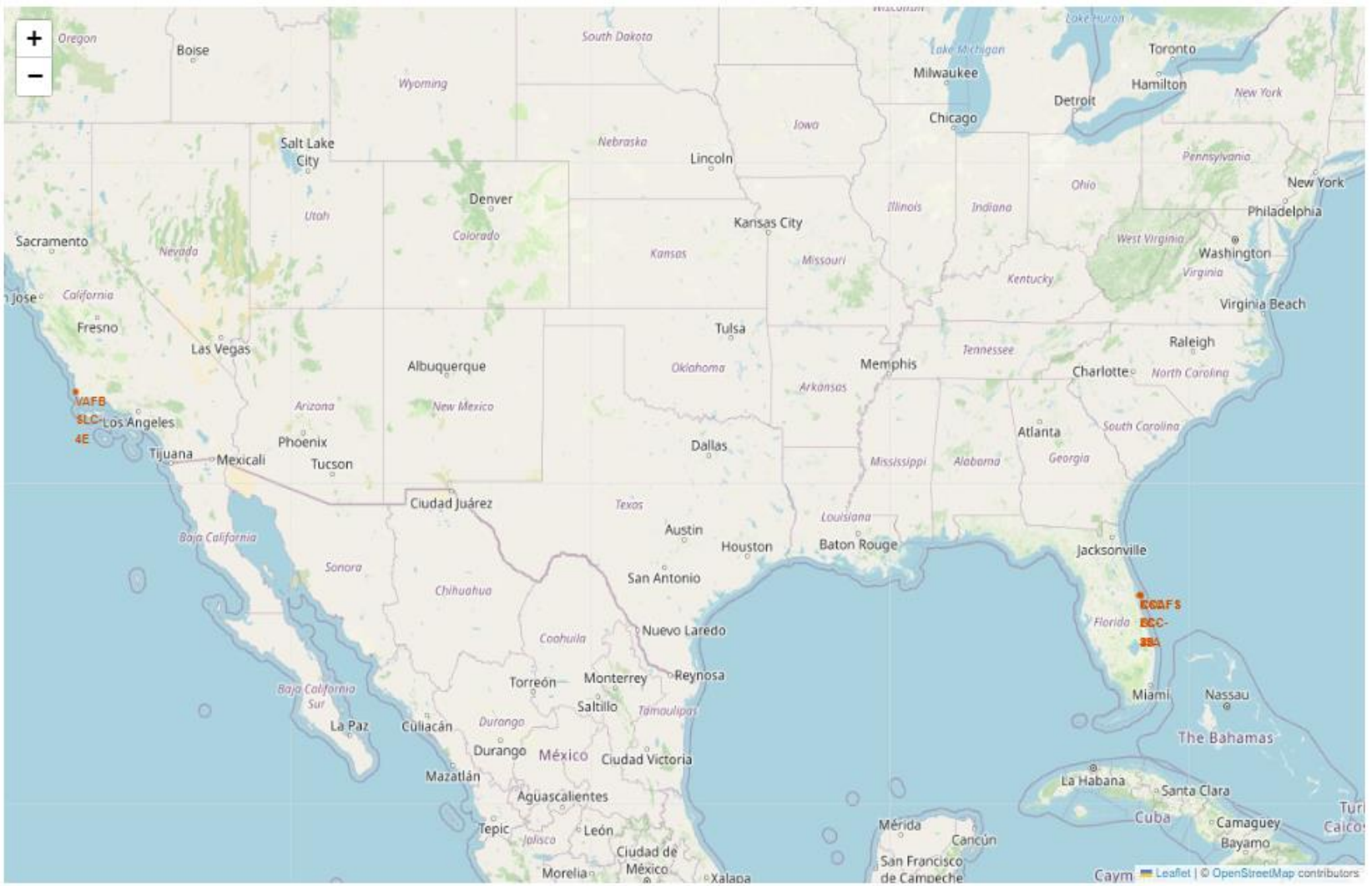
Landing_Outcome	counts
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Between the date 2010-06-04 and 2017-03-20 the highest rank of landing outcomes is for ‘No attempt’ with 10 records. Which indicates that 10 of the launched rockets have made no attempt to land in those 7 years.

Section 3

Launch Sites Proximities Analysis

All Launch Sites on the Map

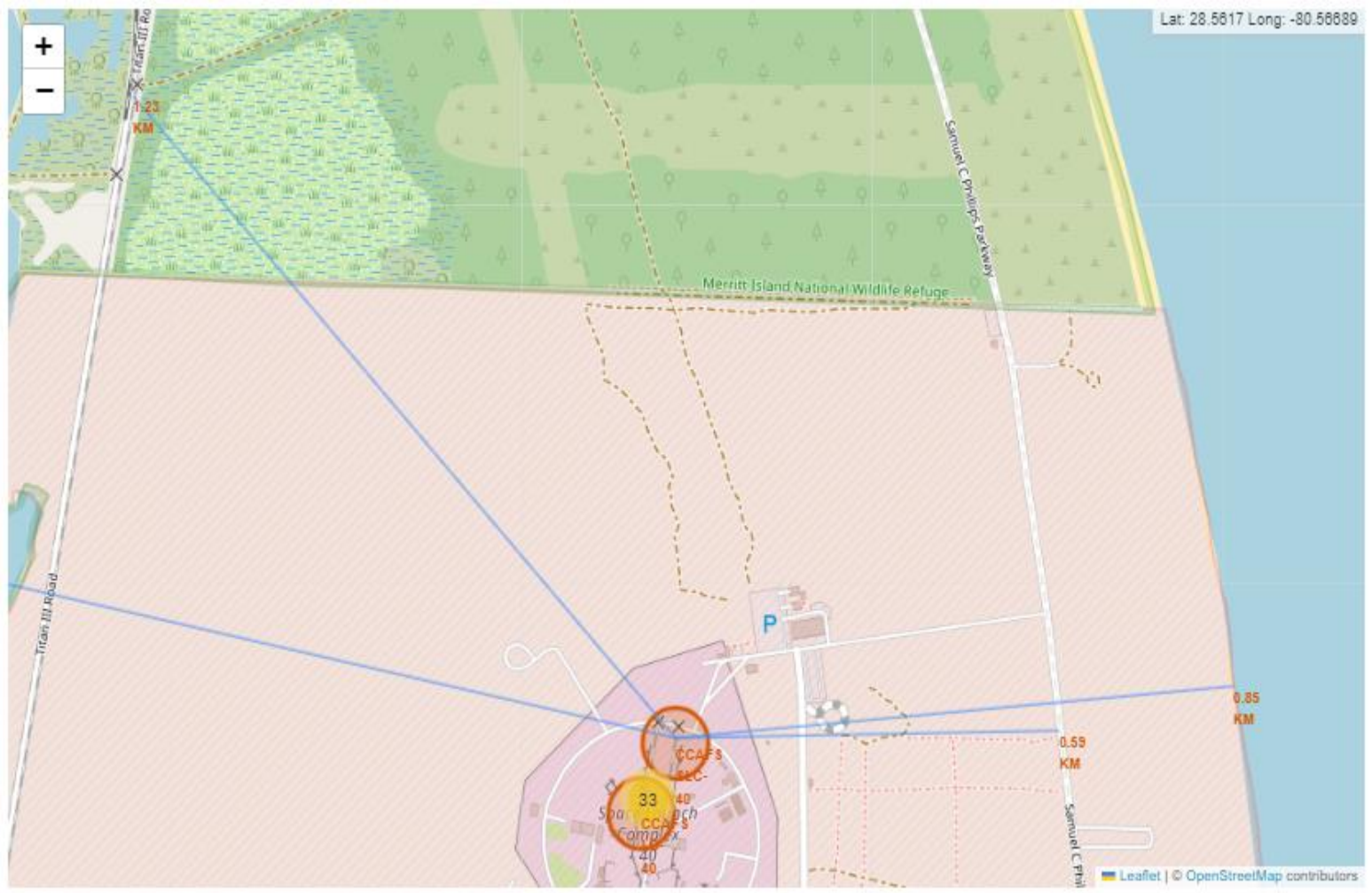


Out of four launch sites, one is in the West coast of California and the other three are in the East coast of Florida

Success/Failure launches marked on the Map

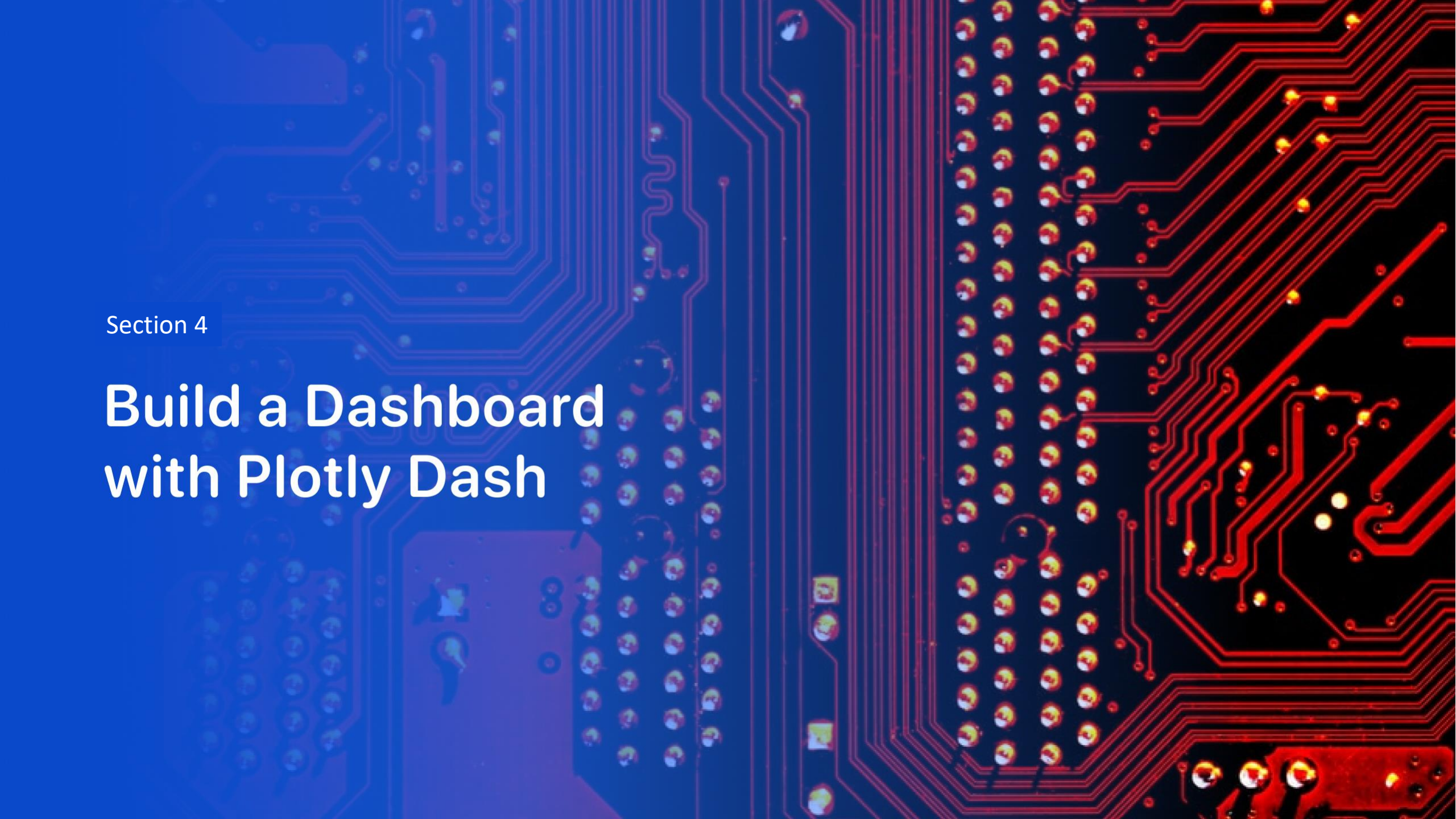
Out of 26 launches at CCAFS LC-40, only 7 launches are successful

Distance between a Launch Site to it's Proximities

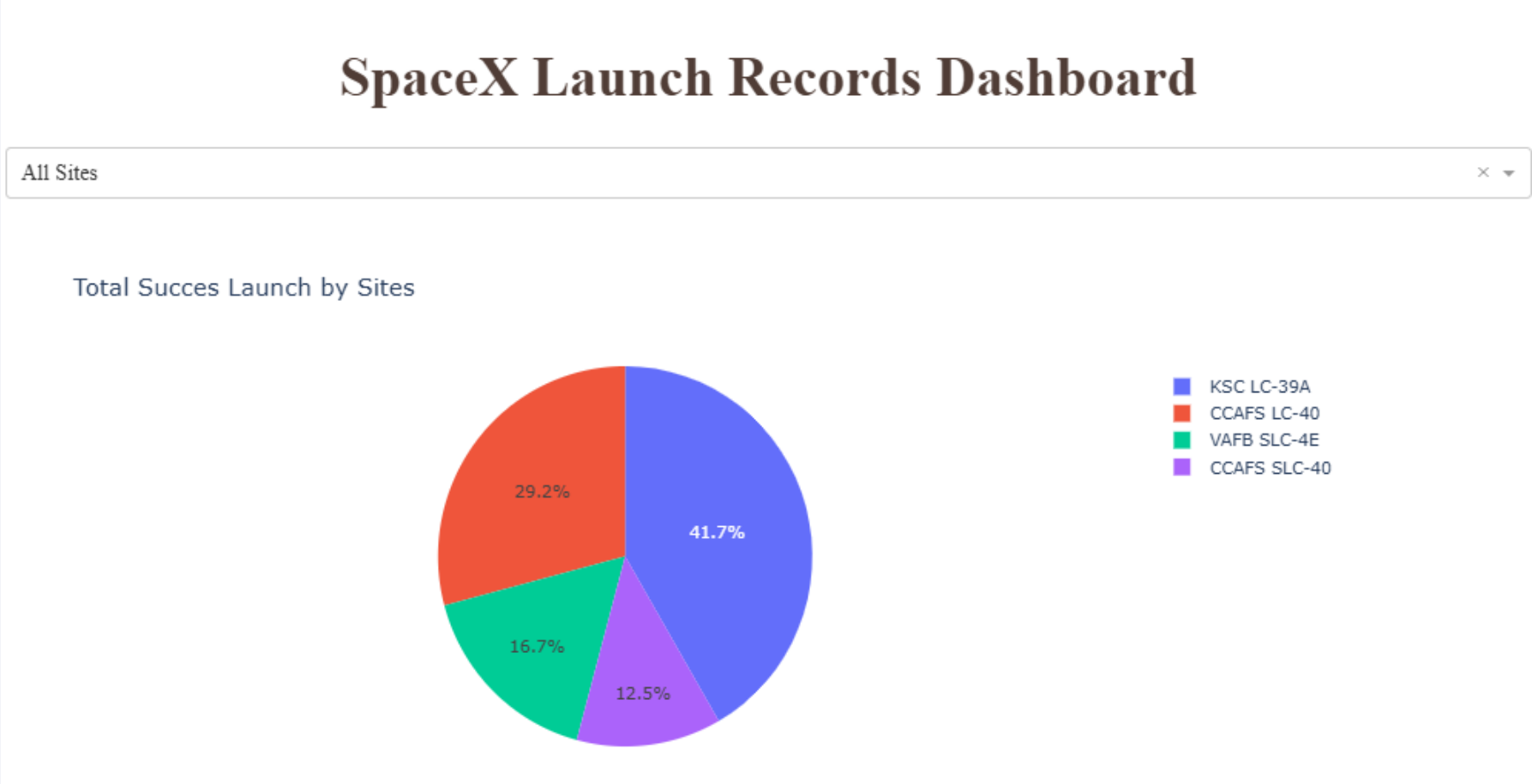


Distance between CCAFS SLC-40 launch site to nearest

- Highway - 0.59 KM
- Railway - 1.23 KM
- Coastline - 0.85 KM
- City - Titusville 24.11 KM

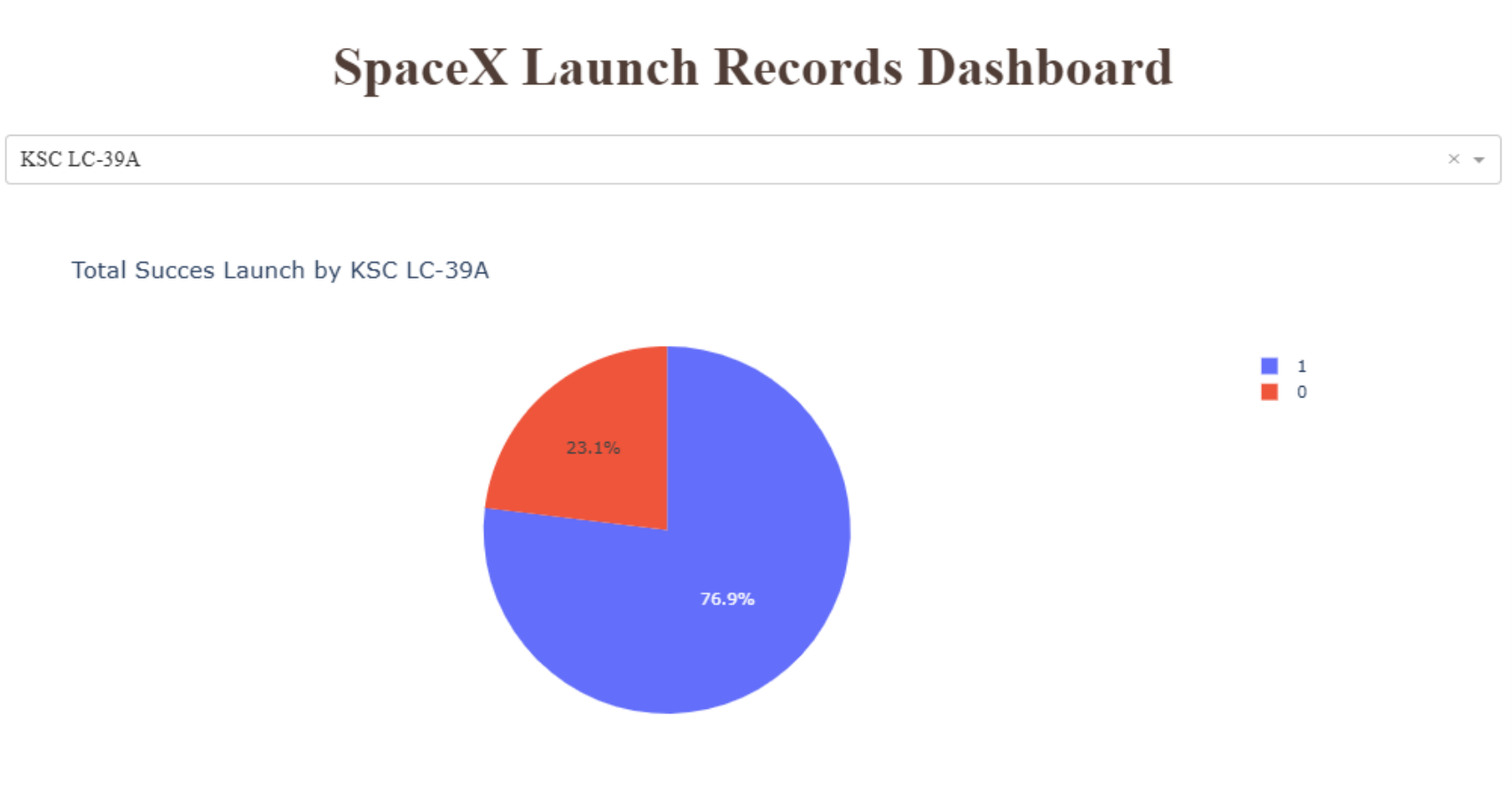

Section 4
Build a Dashboard with Plotly Dash

Total Success Rate by All Launch Sites



Out of all the launch sites, KSC LC-39A has the highest success rate of 41.7% and CCAFS SLC-40 has the lowest success rate with 12.5%

Launch Site with Highest Success Rate Ratio

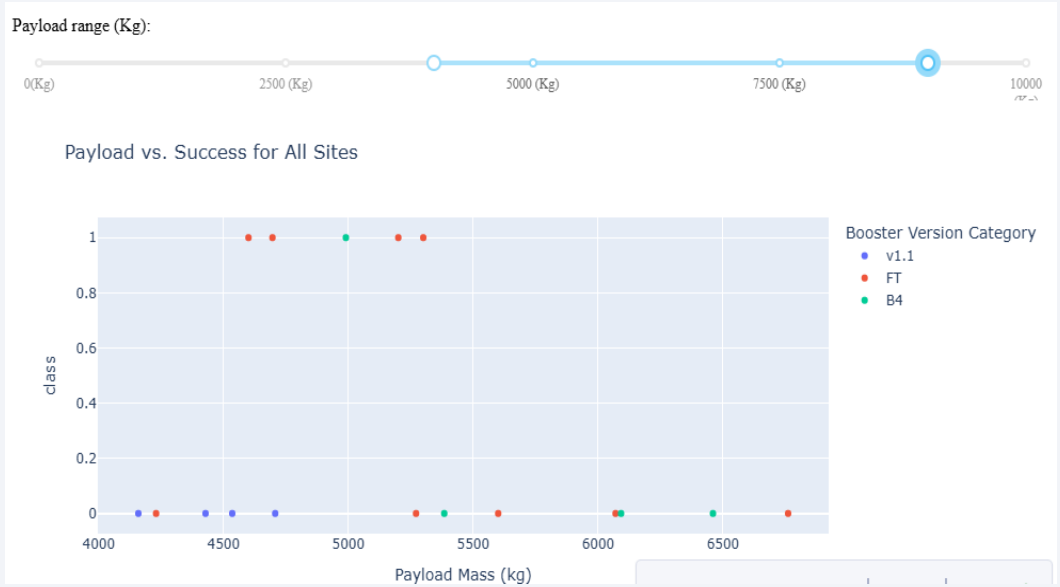


KSC LC-39A has the highest success rate ratio with 76.9% successful launches and 23.1% failures.

Payload vs. Launch Outcome for all sites

Payload range 0Kg – 10,000Kg

Payload range 4,000Kg – 9,000Kg



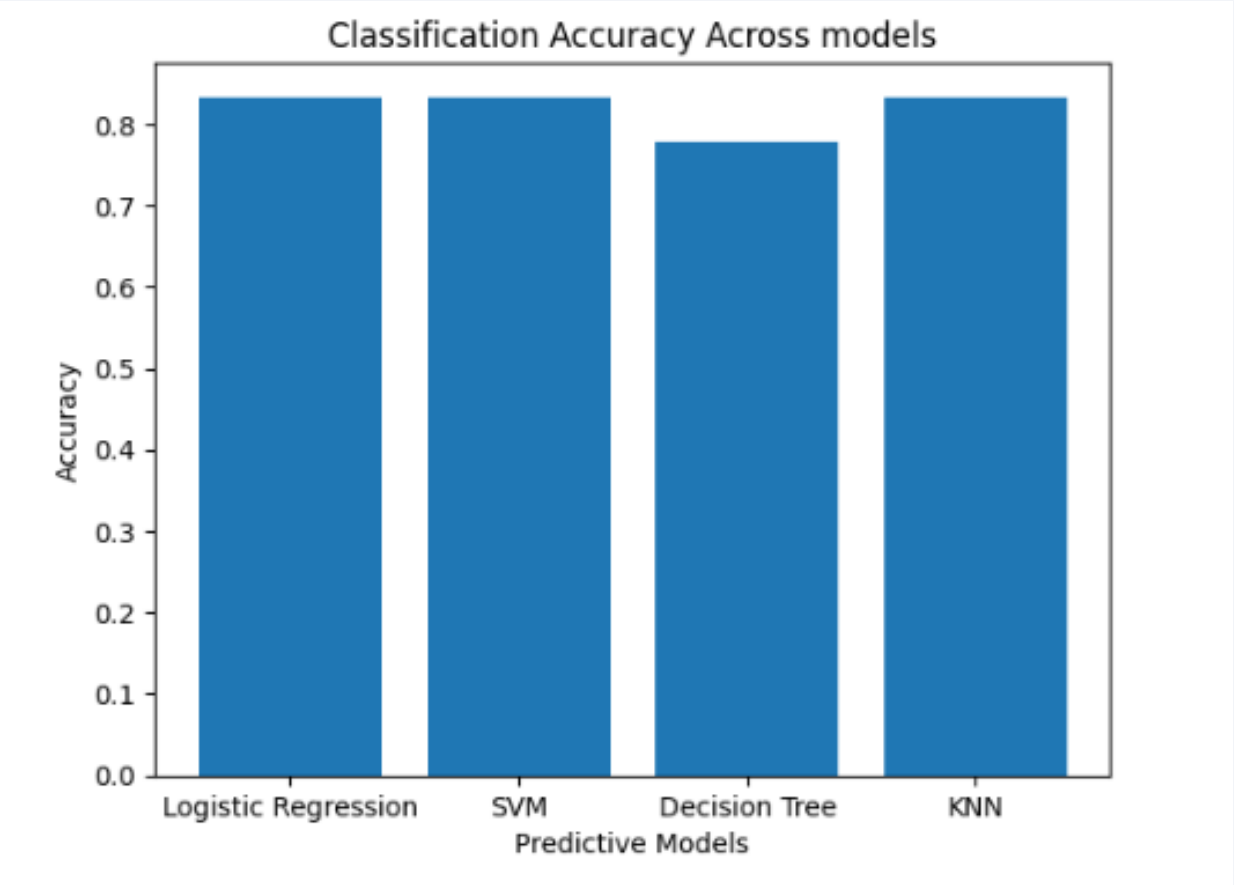
Success rate is high when the payload range is between 2000Kg and 6000Kg and the FT Booster Version has more successful launches.

Among other boosters, only FT and B4 have launched rockets with heavy payload (greater than 6000Kg) and B4 is the only one with a successful launch at 9600Kg payload

Section 5

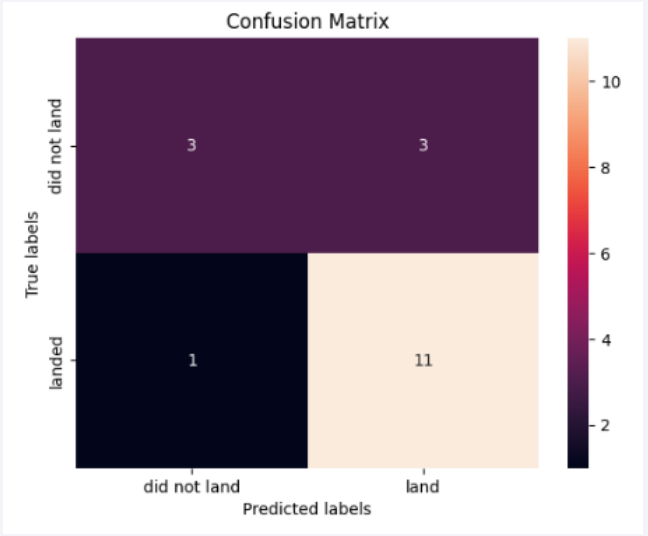
Predictive Analysis (Classification)

Classification Accuracy

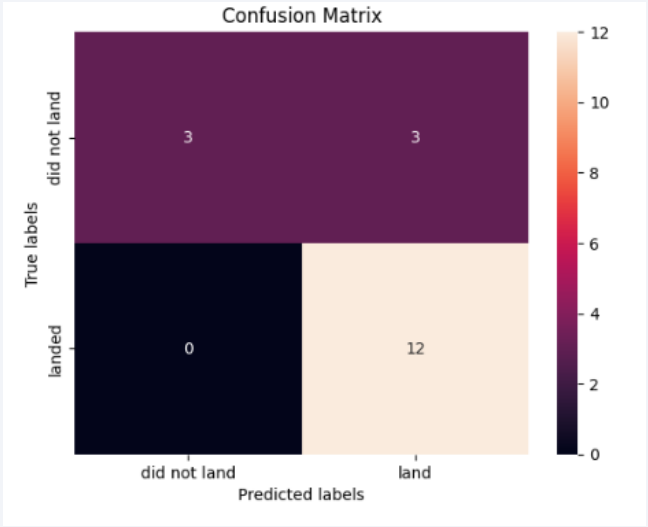


Logistic Regression, SVM and KNN all three have 83% of test accuracy as seen in the bar chart whereas, Decision Tree classifier fall behind

Confusion Matrix



Confusion matrix of Decision Tree



Confusion matrix of Logistic Regression, SVM, and KNN

There are 18 test observations and Logistic Regression, SVM and KNN has given12 correct predictions (True Positives) unlike Decision Tree with 11 TP.

Conclusions

- KSC LC 39A launch site which is situated in the East coast of Florida at Merritt Island is the one with most successful launches and it has launched rockets with variable payload from 2000Kg to 15000Kg heavy payload with positive success rate.
- Rockets launched to GEO, SSO orbits has the highest success rate however, ISS and SSO are the ones delivering positive success rate consistently over the years and VLEO has joined the lineup in the recent years with promising success.
- FT and B4 boosters are the ones with higher success rate over the variable payload scale and B4 is the only one which has succeeded in the launch with heavy payload.
- Payload range of 2000Kg-7000Kg has the highest success rate.
- Logistic Regression, SVM and KNN perform the best in predicting the successful launch of a rocket in the given SpaceX dataset.

Thank you!