IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Amol Shelke

12.09.2022

Outline

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics (MAP) with Folium
  - Predictive analysis (classification)
- Summary of all results
  - Reviewing the exploratory data analysis result
  - Demo the interactive analytics
  - Reviewing predictive analysis result

Introduction

- Project background and context
  - Space X 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 Space X can reuse the first stage.

- Problems you want to find answers
  - The main task of project is to predict if first stage rocket of SpaceX Falcon 9 rocket will land successfully. If we cab figure out if the first stage will land, we can figure out how much a launch will cost. Also, this information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

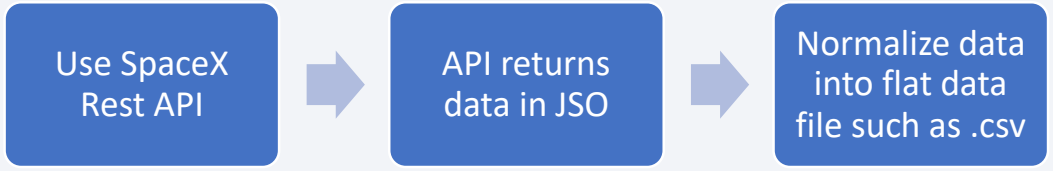
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features, for Machine Learning and dropping irrelevant columns.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Plotting : Scatter Graphs, Bar Graphs to show relationships between variables to show patterns of data.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Perform exploratory Data Analysis and determine Training Labels
  - Find best Hyperparameter for SVM, KNN neighbor, Classification Trees and Logistic Regression

Data Collection

- The data was collected using various methods
  - Data collection was done using get request to the SpaceX API.
  - Next, we decoded the response content as a json using .json() function call and turn it into a pandas data frame using .json normalize().
  - We then cleaned the data, checked for missing values and fill in missing values where necessary.
  - In addition, we performed web scraping from Wikipedia for Falcon9 launch records with BeautifulSoup.
  - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas data frame for future analysis.

SpaceX API

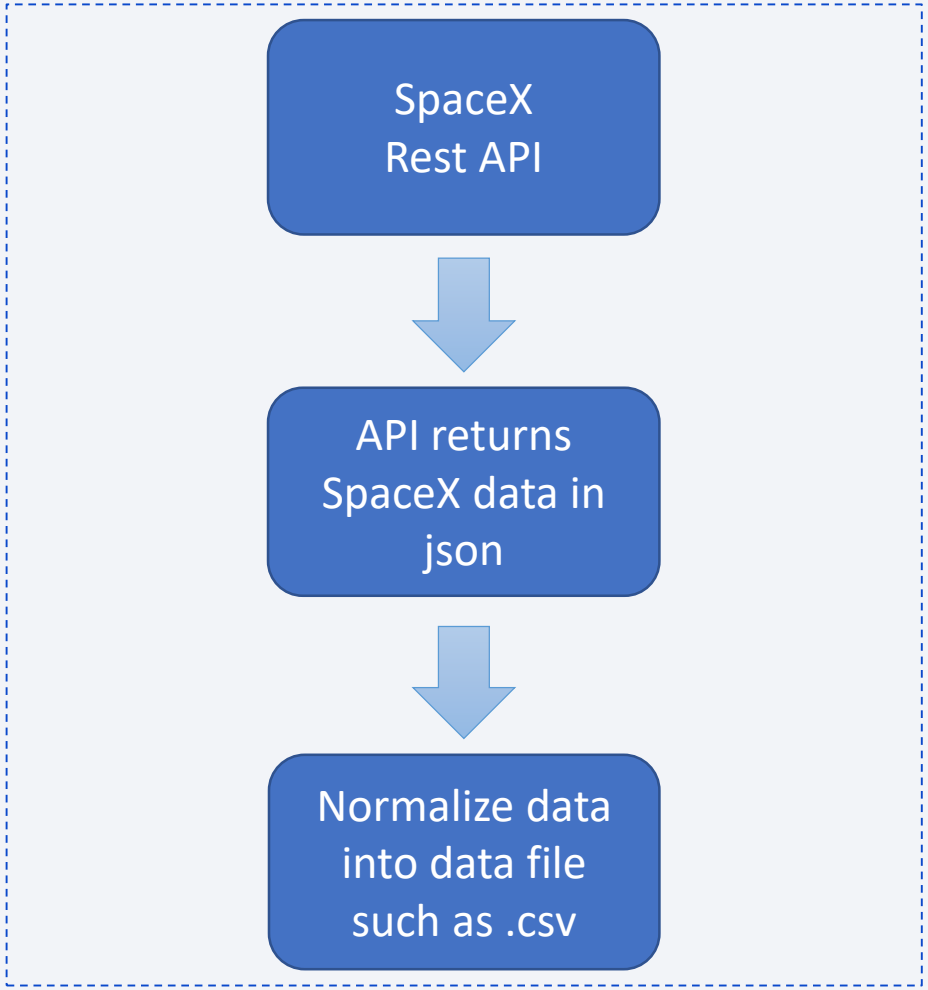


Web Scrapping

Data Collection – SpaceX API

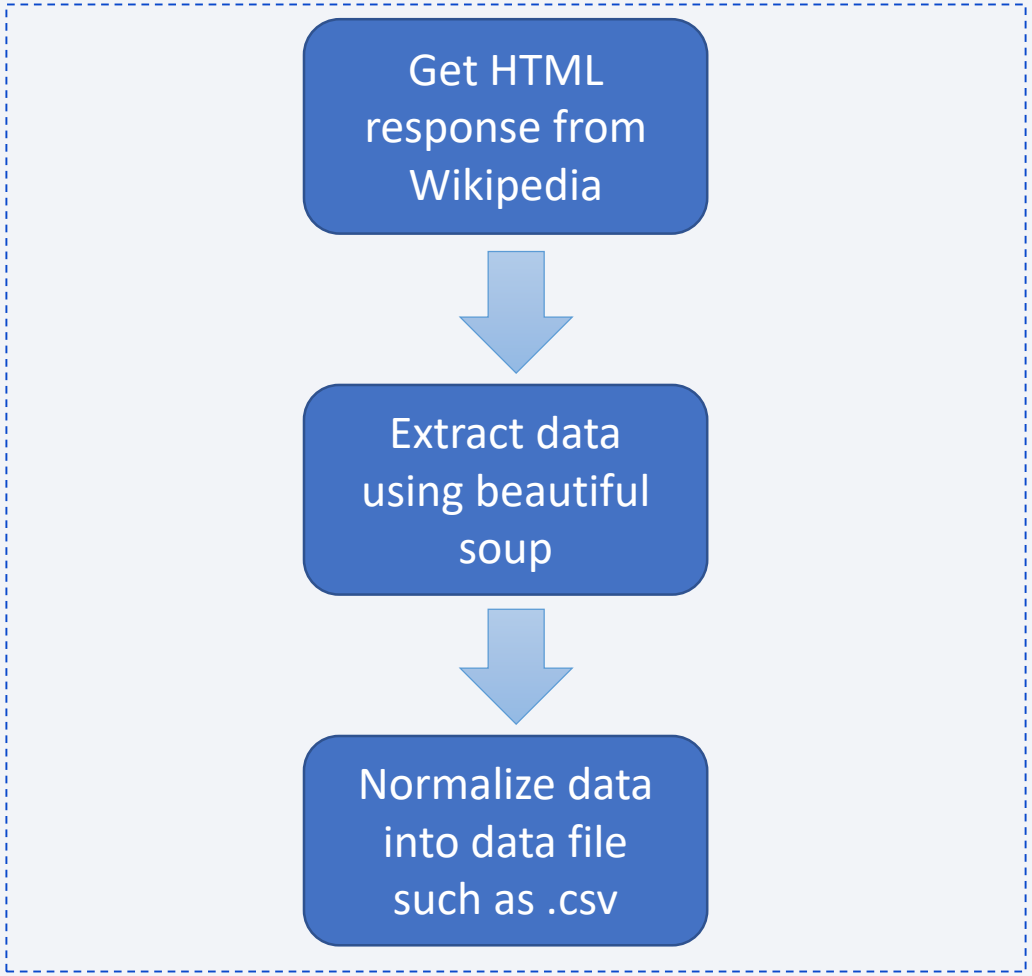
- SpaceX provides a public API from where data can be obtained, analyzed and used.
- This API was used according to the flow chart shown.

- Source Code : https://github.com/macpatil/Capstone-Project-IBM-Data-Science/blob/main/Capstone%20project.ipynb



Data Collection - Scraping

- Data for SpaceX launches can be obtained from Wikipedia.
- Data downloaded from Wikipedia were processed according to the flowchart shown.

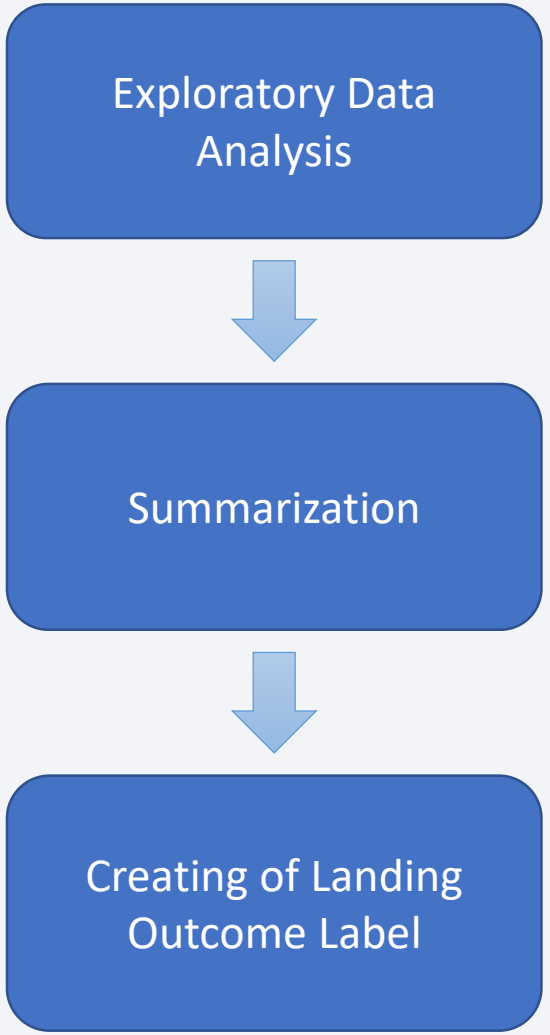


- Source Code : https://github.com/macpatil/Capstone-Project-IBM-Data-Science/blob/main/Castone%20Project%20Data%20collection%20web%20scraping%20W1(1).ipynb

Data Wrangling

- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits.
- We created landing outcome label from outcome column and exported the results to csv.

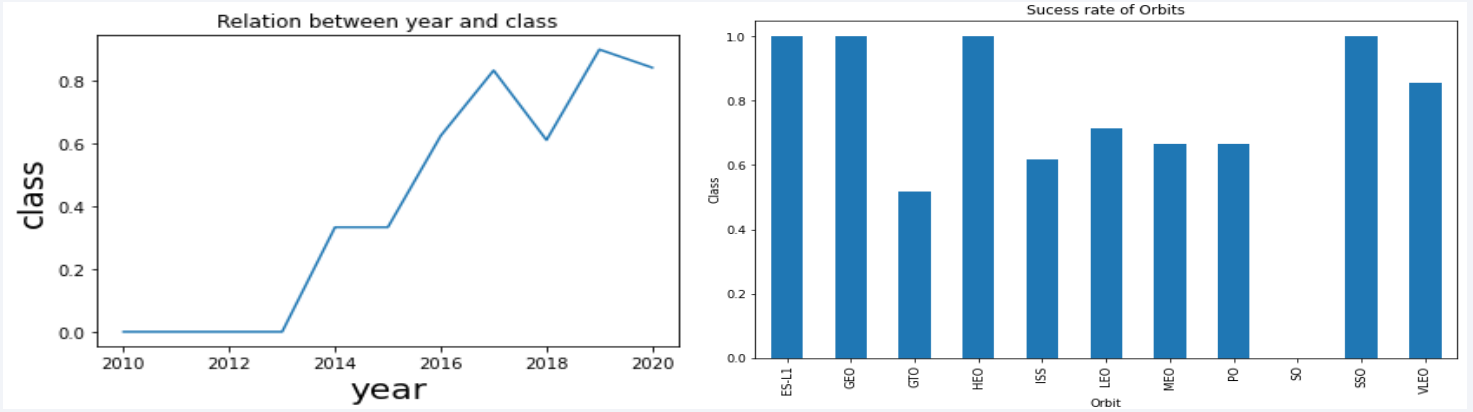
- Source Code : https://github.com/macpatil/Capstone-Project-IBM-Data-Science/blob/main/Capstone%20Project%20Data%20Wrangling%20W1.ipynb

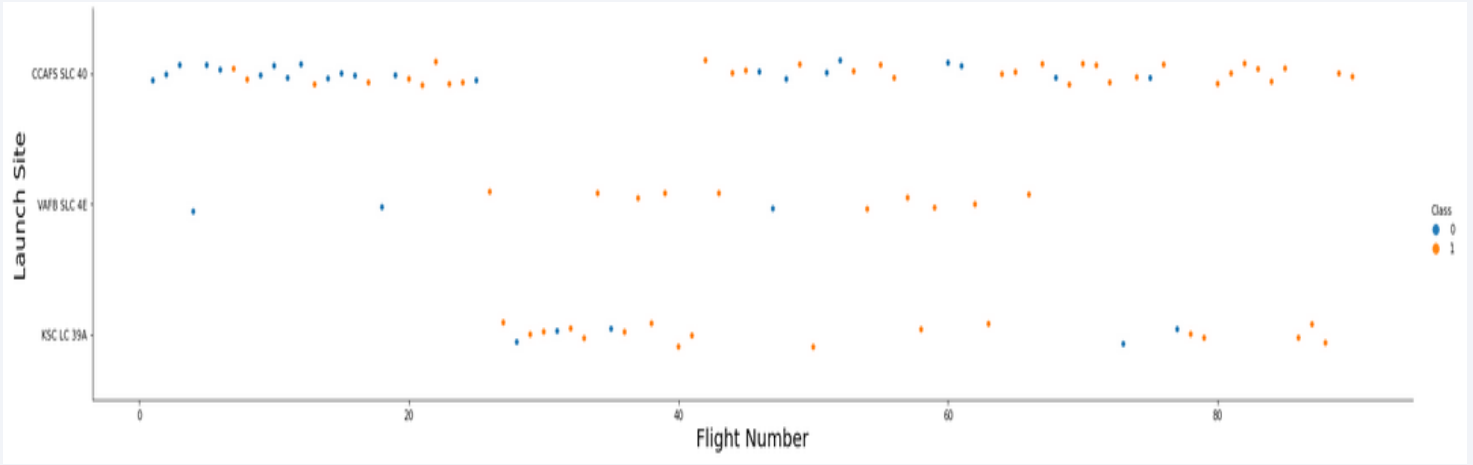


EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.

- https://github.com/macpatil/Capstone-Project-IBM-Data-Science/blob/main/Capstone%20project%20EDA%20with%20visualization%20W2.ipynb





Relationship between Flight Number and Launch Site

EDA with SQL

We loaded the SpaceX dataset into a SQLite database without leaving the jupyter note book. Load CSV to SQLite with Create New Table and Perform Analysis on Table Name SpaceX

- Name of the unique launch sites in the space mission
- 5 records where launch sites begin with the string ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Name of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Name of the booster version which have carried the maximum payload mass.
- Failed landing outcomes in drone ship, their booster versions, and launch site names in the year 2015
- Count of successful landing outcome between the date 2010-06-04 and 2017-03-20 in descending order

- Source Code : https://github.com/macpatil/Capstone-Project-IBM-Data-Science/blob/main/Capstone%20EDA%20with%20SQL%20W2.ipynb

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 (Green and Red) 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 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
- Source code : https://github.com/macpatil/Capstone-Project-IBM-Data-Science/blob/main/Capstone%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

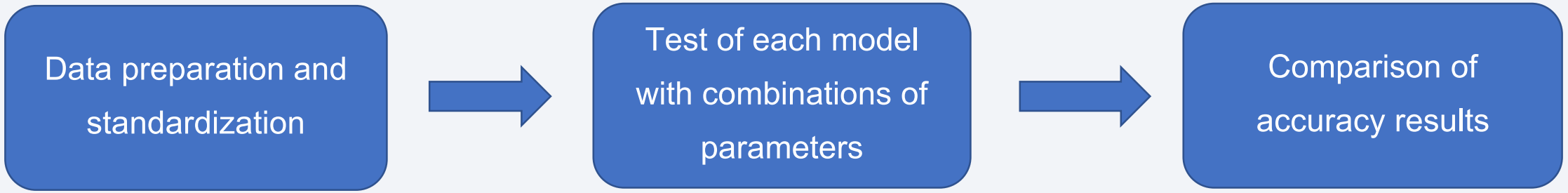
Build a Dashboard with Plotly Dash

- 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.

- Source Code : https://github.com/macpatil/Capstone-Project-IBM-Data-Science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

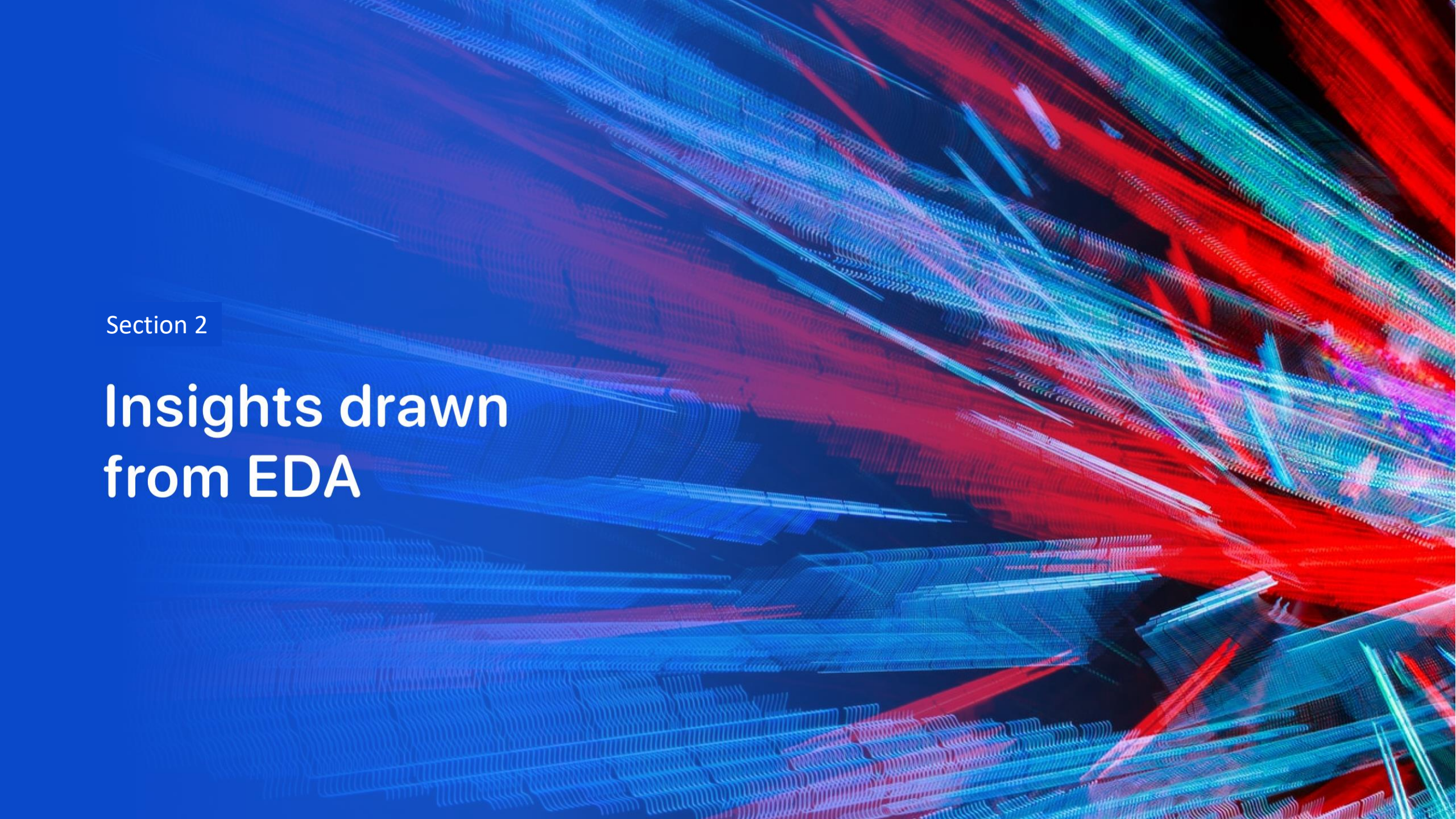
- Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbor were the Classification models used and compared to find which Classification model will give more accurate result.



- Source Code : https://github.com/macpatil/Capstone-Project-IBM-Data-Science/blob/main/Capstone%20Machine%20Learning%20Lab.ipynb

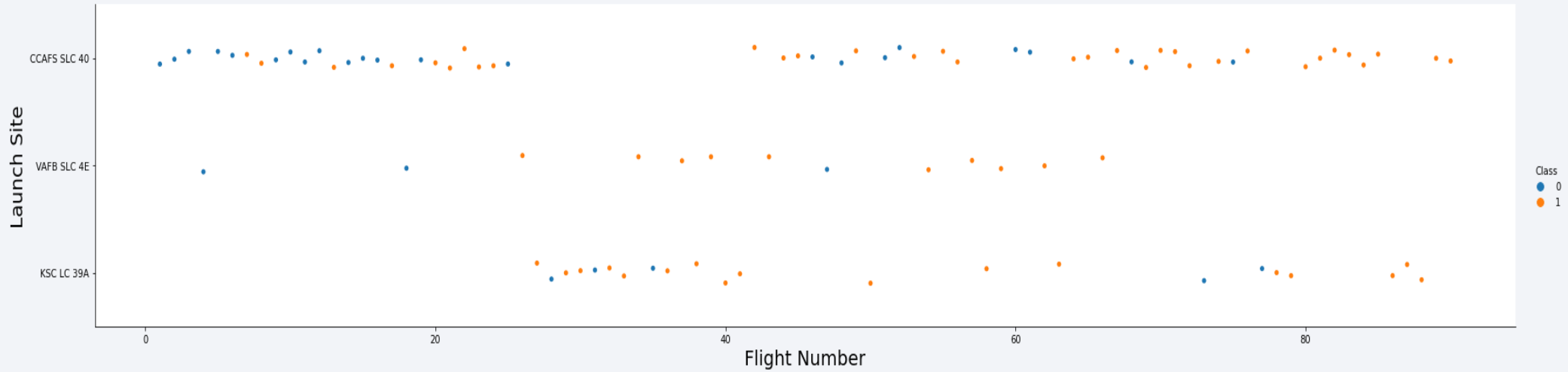
Results

- Exploratory data analysis results
  - Average payload of F9 v1.1 booster is 2,928 kg.
  - The first launch were done in SpaceX and NASA.
  - The first successful landing happened in 2015,5 years after the first launch.
  - The number of landing outcome be came better as years passed.
  - Almost 100%of mission outcome were successful.
- Predictive analysis results showed that the Decision Tree Classifier is the best model to predict successful landing, having an accuracy over 87%.
- Interactive analytics
  - Using interactive analytics, it was possible to identify that launch site were having good logistic location, near the coastal line and far away from populated area.


Section 2
Insights drawn from EDA

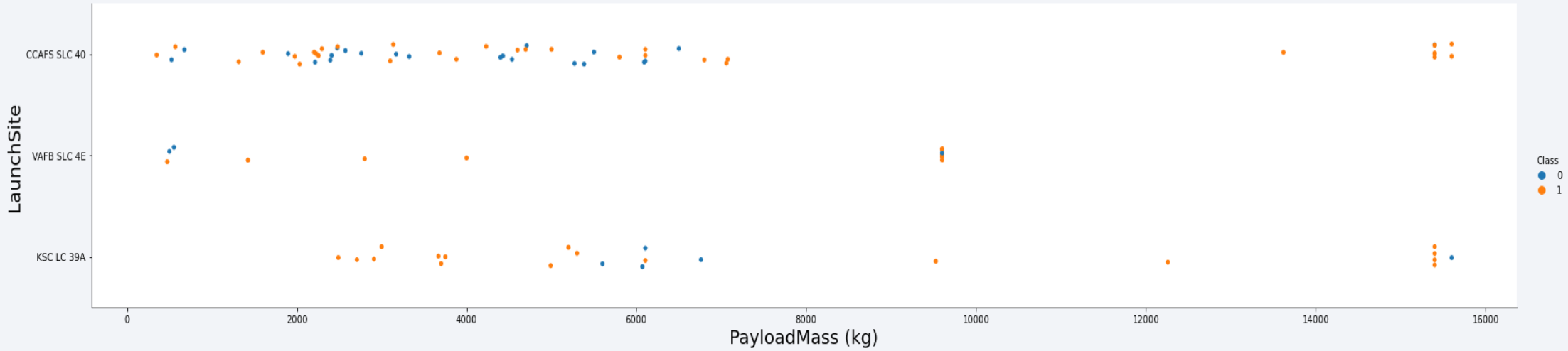
Flight Number vs. Launch Site

- The more flights at a launch site, the greater the success rate it gets.
- The plot shows that the best launch siteisCCAF5 SLC 40, where most of recent launches were successful.



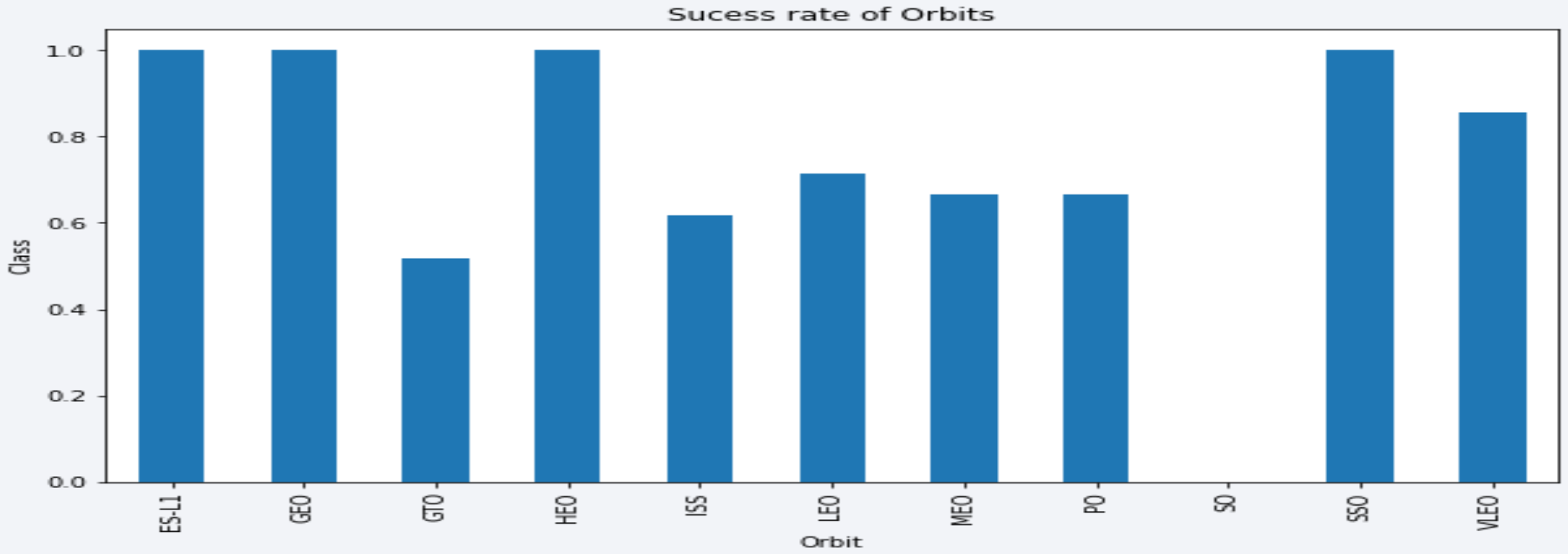
Payload vs. Launch Site

- Payloads over 9000 kg have excellent success rate.
- Payloads over12000 kg seems to be possible only on CCAFSSLC 40 and KSC LC 39A launch sites.



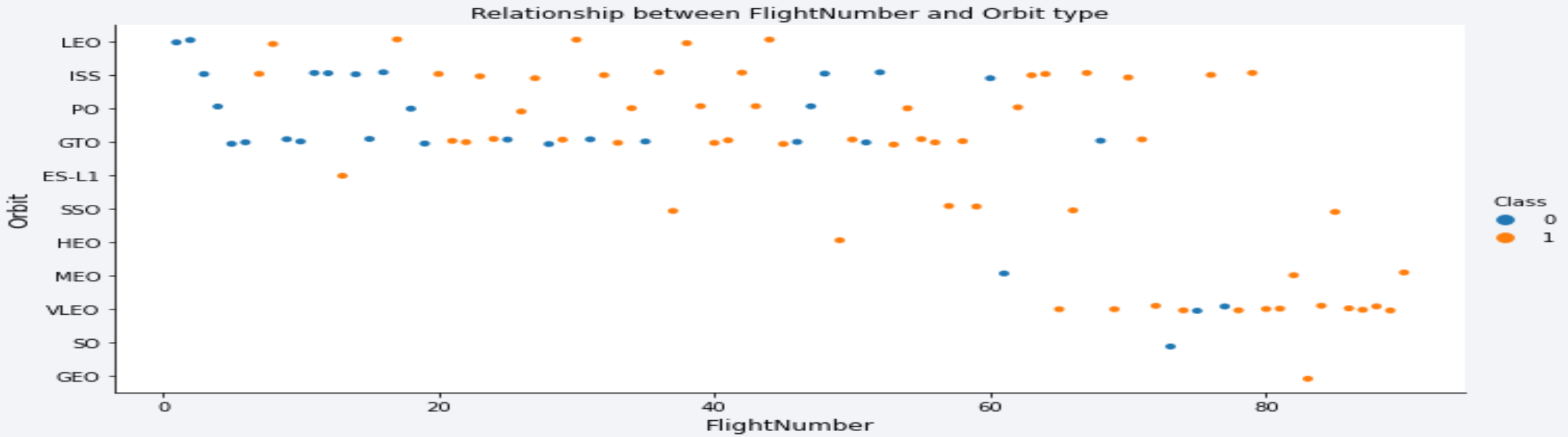
Success Rate vs. Orbit Type

- Orbit GEO, HEO, SSO, ES-L1 has the best success rate and SO having 0 success rate.



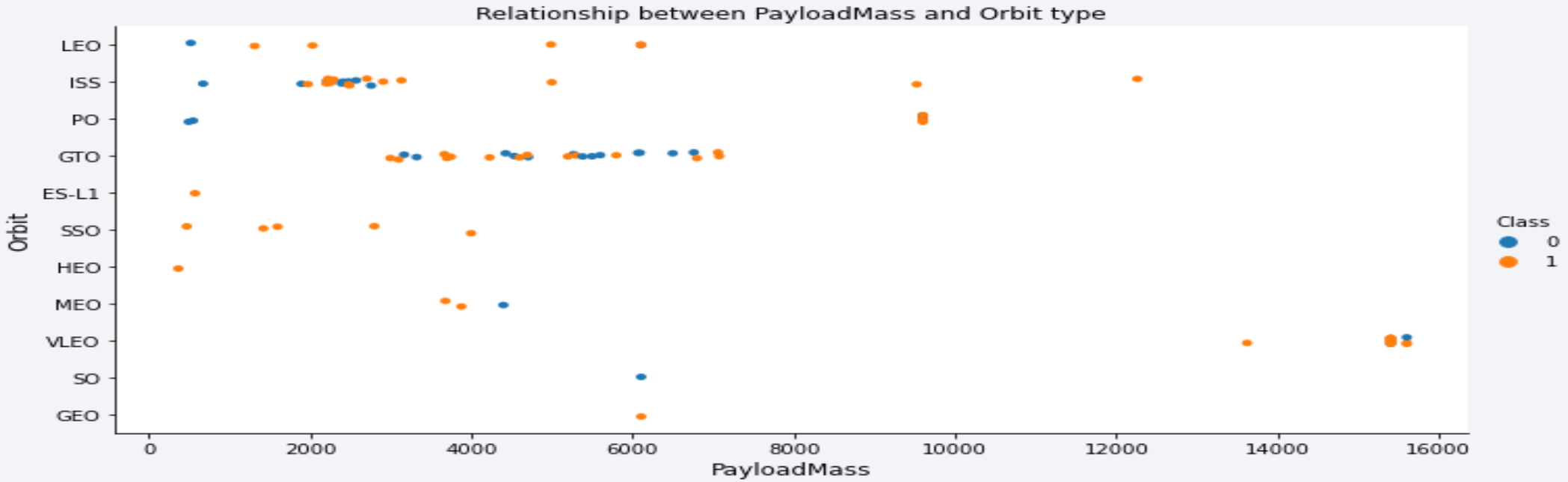
Flight Number vs. Orbit Type

- It is evident that in the LEO orbit the success rate appears to be related to the number of flights. On the other hand, there seems to be no relationship between flight number when in GTO orbit.



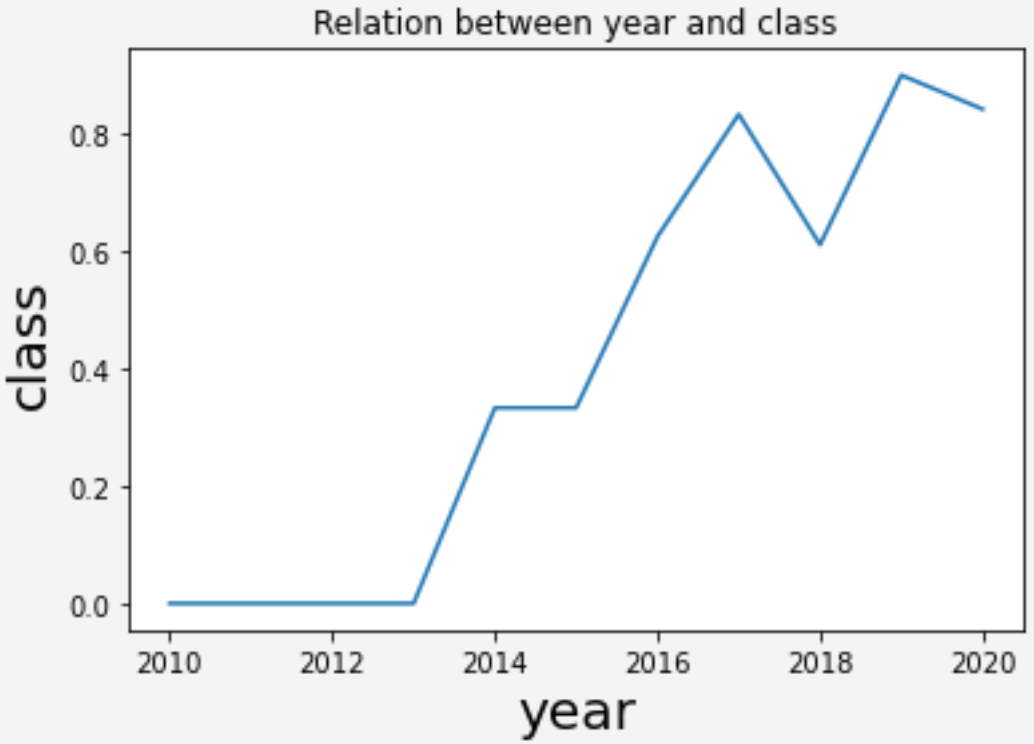
Payload vs. Orbit Type

- The graph shows no relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and good rate of success.
- Few launches are evident to the orbits SO and GEO.



Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- Success in recent years at around 80%.



All Launch Site Names

- The data showed that there are 4 launch sites as follows below:
- Query was obtained by selecting unique occurrences of "launch_site" values from the database.

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

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

- There were 5 recorded launch sites that begin with `CCA`.
- Here we can see 5 samples of launches that had landing failure or no attempt at all.

Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

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

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

- Total payload were calculated below by summing up all payloads whose having 'CRS', which correlate with NASA.

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Data were filtered by the booster version and calculated the average payload mass as shown below.

Display average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Data were filtered by successful landing outcome on ground pad and getting the minimum date value possible to identify the very first occurrence as shown below.

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [36]:
```
%%sql
select min(DATE) as "First Successful Landing Outcome on Ground Pad" from SPACEXTBL
where LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://byx44847:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg
Done.
```

Out[36]:

First Successful Landing Outcome on Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query returns the 4 booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 non-inclusively.

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

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 and  "Landing _Outcome" = 'Success (drone ship)'
```

```
 * 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

- This query returns a count of each mission outcome. SpaceX appears to achieve its mission outcome nearly 99%of the time.

List the total number of successful and failure mission outcomes

```
%%sql
SELECT count(Mission_Outcome)
FROM SPACEXTBL
WHERE Mission_Outcome like 'Success%'
```

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

count(Mission_Outcome)
100

```
%%sql
SELECT count(Mission_Outcome)
FROM SPACEXTBL
WHERE Mission_Outcome like 'Failure%'
```

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

count(Mission_Outcome)
1

Boosters Carried Maximum Payload

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar and all are of the F9 B5B10xx.x variety.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%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

- This query returns the Month, Landing Outcome, Booster Version, Payload Mass and Launch site of 2015 launches where stage 1 failed to land on a drone ship which had two such occurrences.

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

```
%%sql
SELECT
        Booster_Version,
        Launch_site
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Failure (drone ship)' and substr(Date,7,4)='2015'
```

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

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

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- Present your query result with a short explanation here

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
select LANDING__OUTCOME, count(LANDING__OUTCOME) as "Count" from SPACEXTBL
where DATE between '2010-06-04' and '2017-03-20'
group by LANDING__OUTCOME order by count(LANDING__OUTCOME) desc ;
```

```
 * ibm_db_sa://byx44847:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

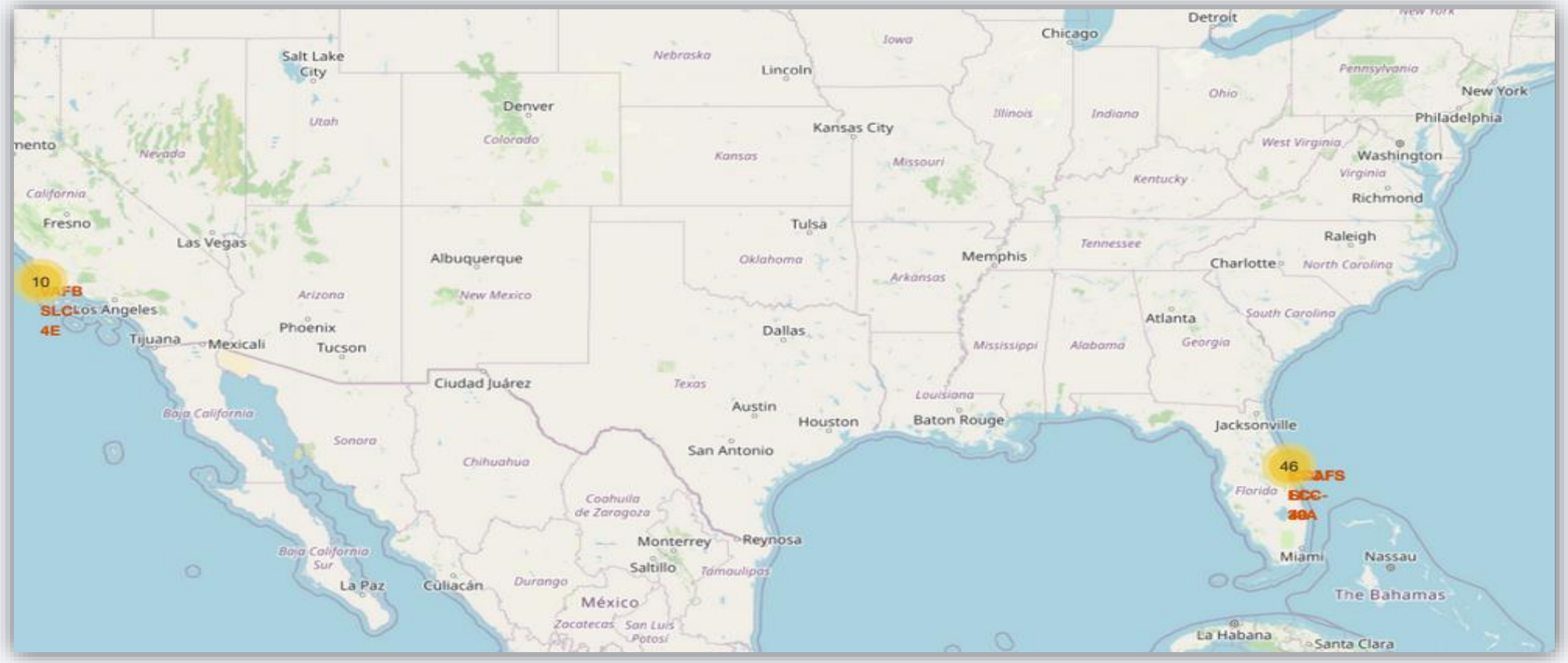
landing__outcome	Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

- Launch sites are situated logistically sound, not far from roads and railroads and mostly near the ocean but far away from inhabited areas.

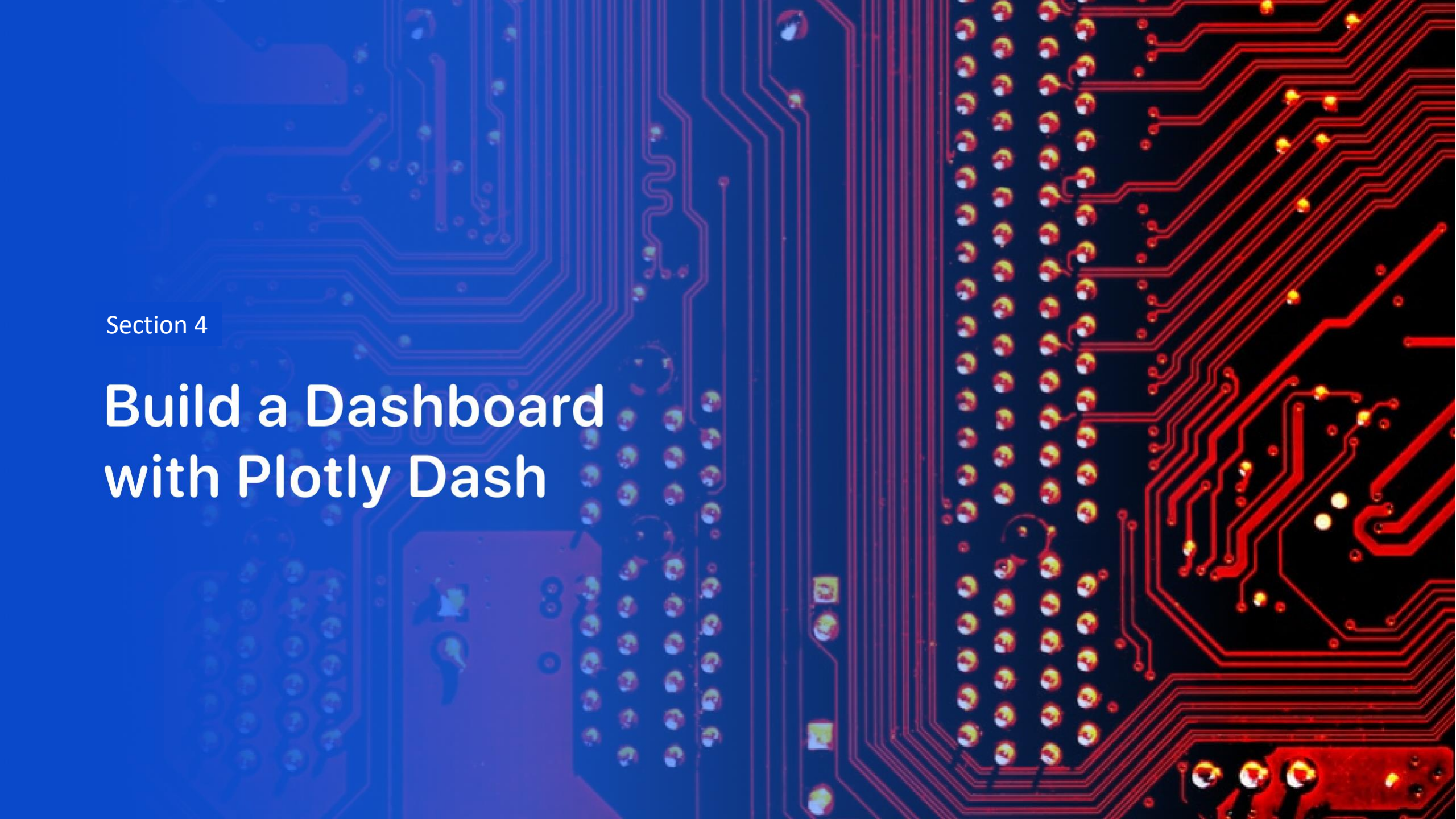


Colored-coded Launch Outcomes

- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon).

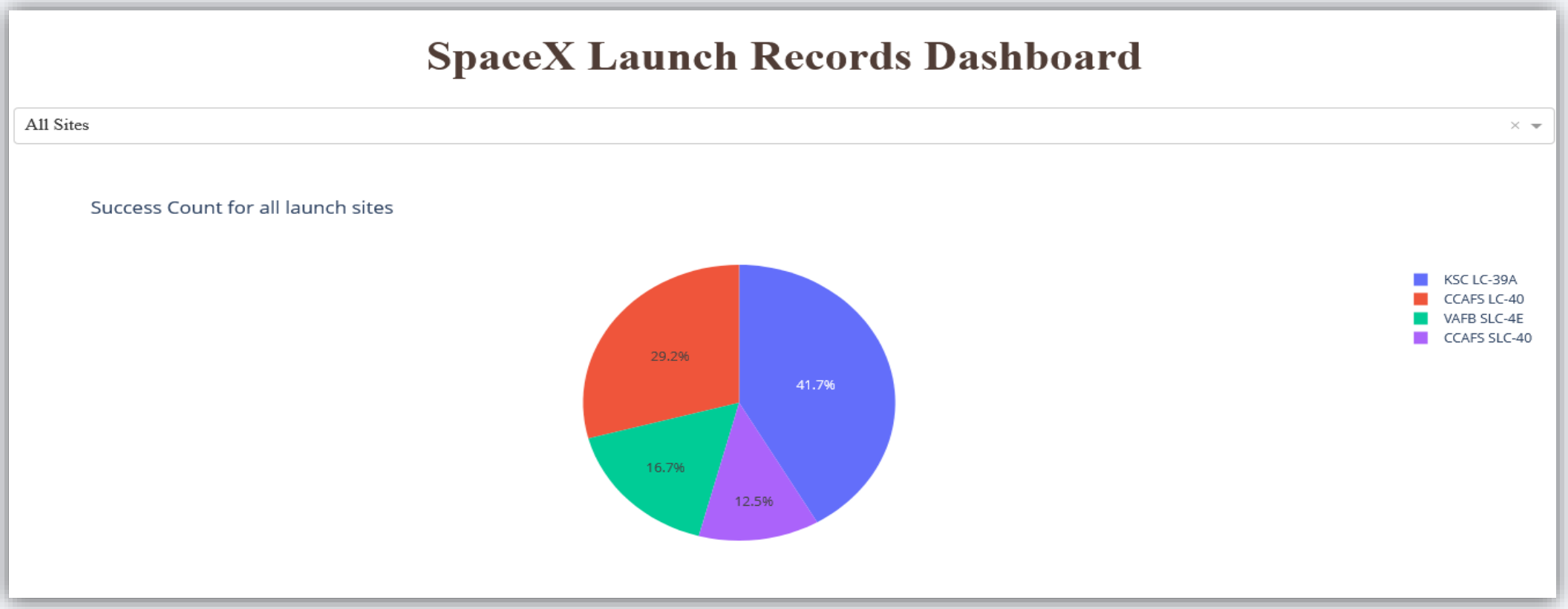
Logistics and Safety

- Launch site KSC LC-39A has very logistics, being near both railroad and roads and relatively far from inhabited areas.


Section 4
Build a Dashboard with Plotly Dash

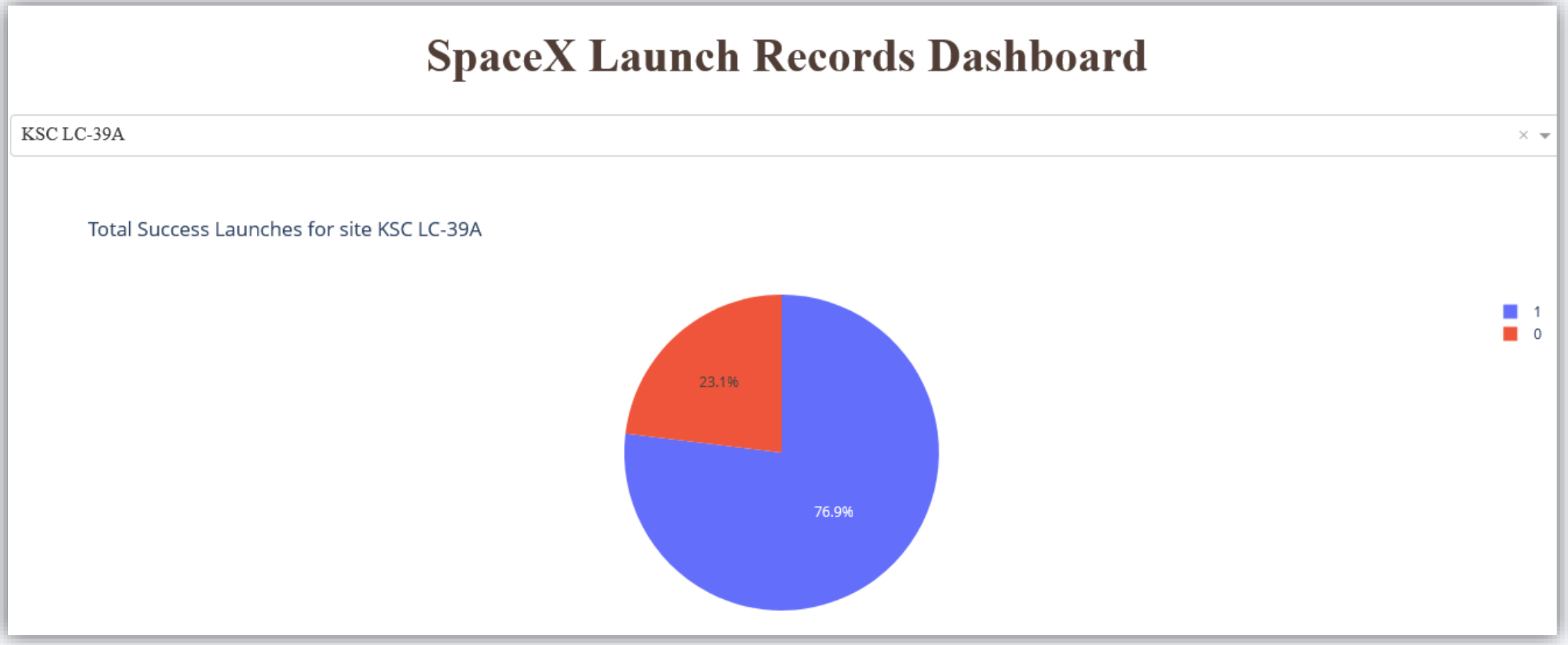
Successful Launches Across Launch Sites

- This is the distribution of successful landings across all launch sites. CCAFS and KSC have the same amount of successful landings. VAFB has the smallest percentage of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.



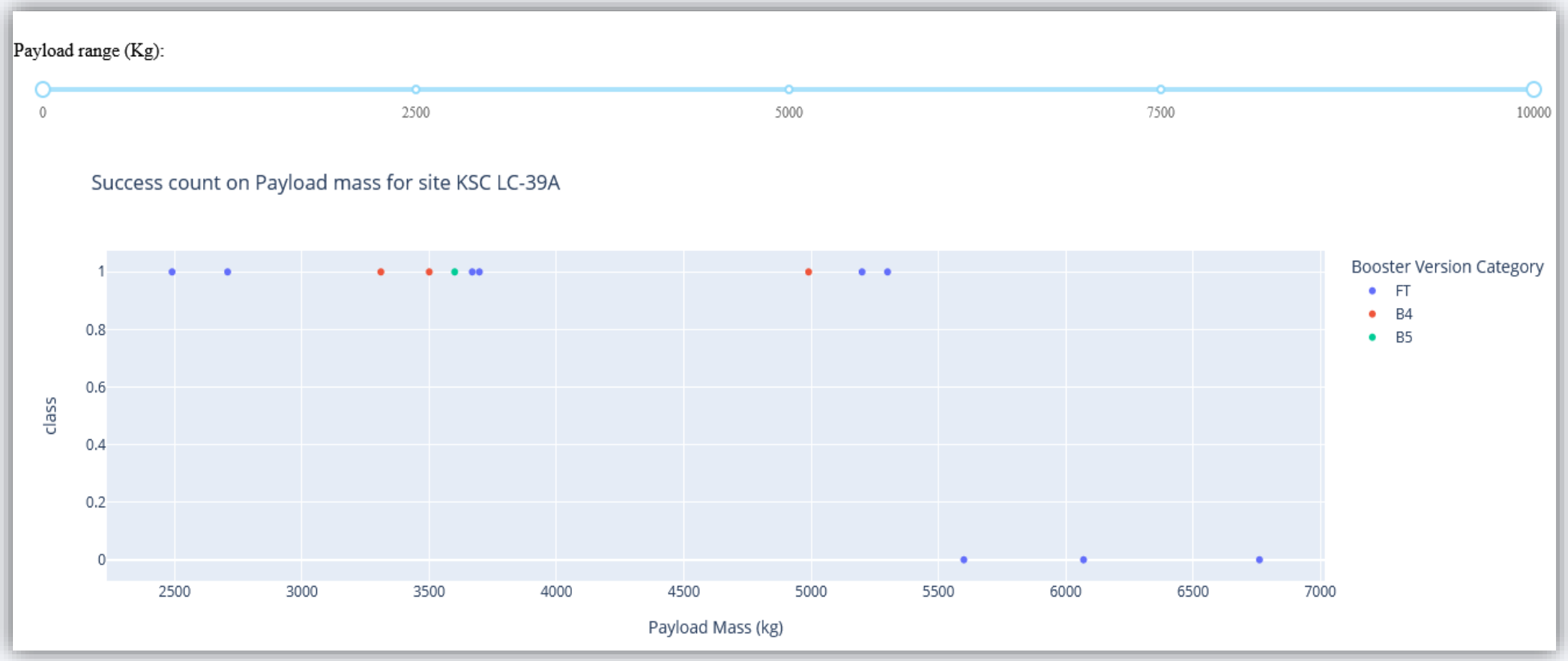
Highest Success Rate Launch Site

- KSC LC-39A has the highest success rate of 76.9% among all the launch sites.



Payload vs. Launch Outcome

- Payloads under 6,000 kg. And FT Boosters are the most successful combination as shown in the scatterplot.



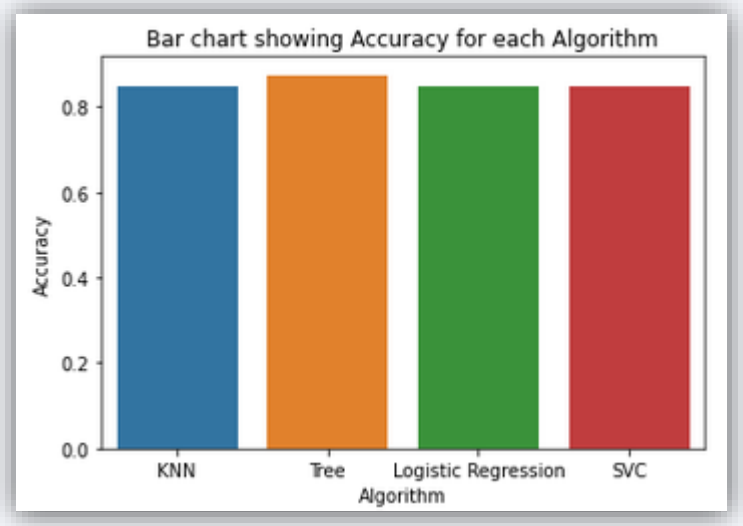
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- There are four classification models tested and the accuracies are shown in the bar chart.

- As shown in the bar graph and chart below, the model with the highest accuracy is the Decision Tree Classifier which having 87.5% accuracy and 94.4% test accuracy.

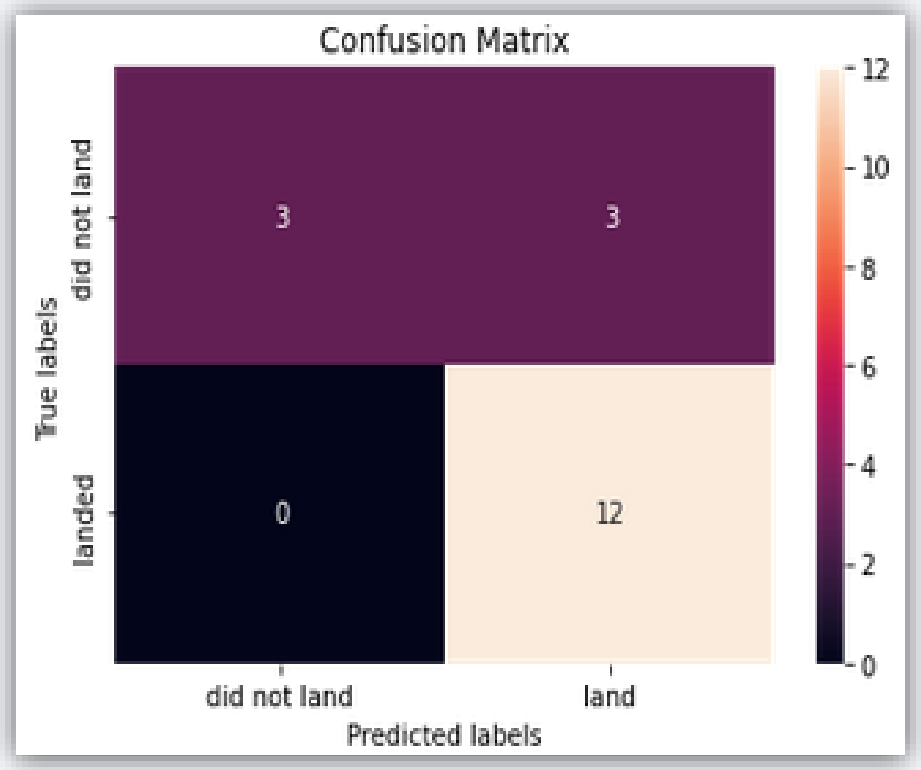


	Algorithm	Accuracy
0	KNN	0.848214
1	Tree	0.875000
2	Logistic Regression	0.846429
3	SVC	0.848214

Confusion Matrix

- Decision Tree Classifier Confusion Matrix clearly show its accuracy by the numbers shown in the true positive and true negative compared to the false one.

- Correct predictions are read diagonally from the top left down to the bottom right.



Conclusions

- Low weight payloads perform better than heavier ones.
- The success rates of SpaceX launches and landings is directly proportional to the numbers of launches and the times in years as they improve future launches.
- As the data have shown , KSC LC-39Ahad the most success rate of launches done from all the launch sites.
- Orbit GEO, HEO, SSO, ES-L1 has the best success rate.
- Created a machine learning model and come up with the best Classifier with an accuracy of87.5%.
- If possible more data should be collected to better determine the best machine learning model and improve accuracy.

Appendix

- GitHub repository URL:

  https://github.com/macpatil/Capstone-Project-IBM-Data-Science

- Data Science Instructors:

  https://www.coursera.org/professional-certificates/ibm-data-science?#instructors

Thank you!