IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nagalekshmi N K
26-09-2022

Outline

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

Executive Summary

- **Methodologies**:
  1. Data collection using Webscraping
  2. Data Wrangling
  3. Exploratory Data Analysis(EDA) with SQL
  4. EDA with Data Visualization
  5. Interactive Visual Analytics using Folium
  6. Interactive Dashboard with Plotly Dash
  7. Machine Learning Prediction (Model Development)

- **Result summary**:
  i. Data was successfully collected.
  ii. Retrieved data are pre-processed and set to analyze.
  iii. By doing EDA on the data, valuable outputs regarding the launch were obtained.
  iv. Interactive representation of the data was provided through series of folium maps and plotly dash- dashboard.
  v. The best model for predicting the features required for the better way to approach the launch was initiated by using ML Prediction.

Introduction

- **Objective:**

  To find all possible ways to give results for the better development of the given company Space Y.

- **Problems to be answered:**

  1. Cost estimation (i.e.) price of each launch.
  2. Best place for the launch
  3. If the first stage will be reused.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data was collected from sources like

  - https://api.spacexdata.com/v4/launches/past
  - Wikipedia page titled List of Falcon 9 and Falcon Heavy launches: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Perform data wrangling

  - The collected data was preprocessed and was properly labeled which helped in further analysis and prediction.

Continues…

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The collected data after further cleaning and summarizing was brought to create a model.
  - The data was normalized and split into train and test data randomly.
  - Created four classification models namely: Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbor.
  - Then accuracy of each model was compared for the best result.

Data Collection

Data sources:

1. Space X API:

https://api.spacexdata.com/v4/launches/past

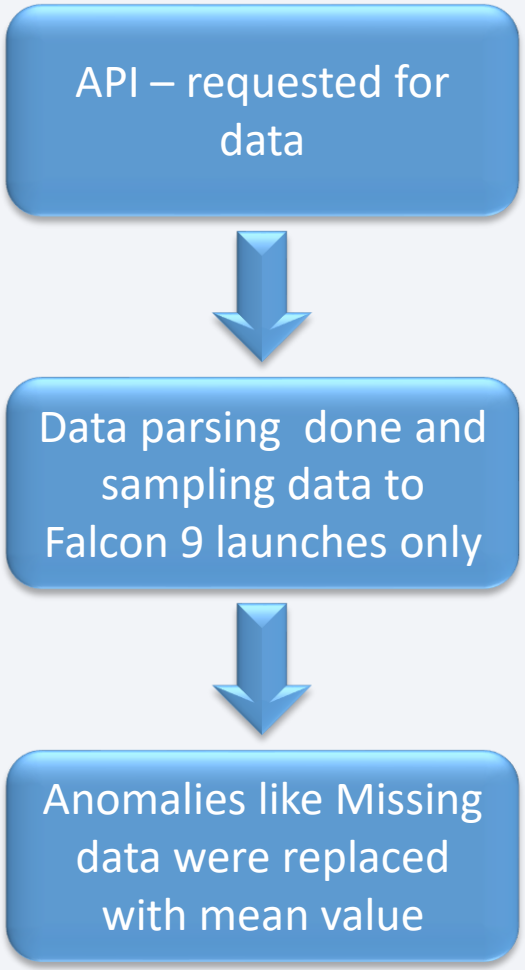
2. Wikipedia page: ‘List of Falcon 9 and Falcon Heavy launches’:

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- The data was collected from a public Space X API.
- The data was being requested through API and being parsed.
- Then, the data was sampled as per the need and dealt with anomalies
- Github link for Data Collection – SpaceX API:

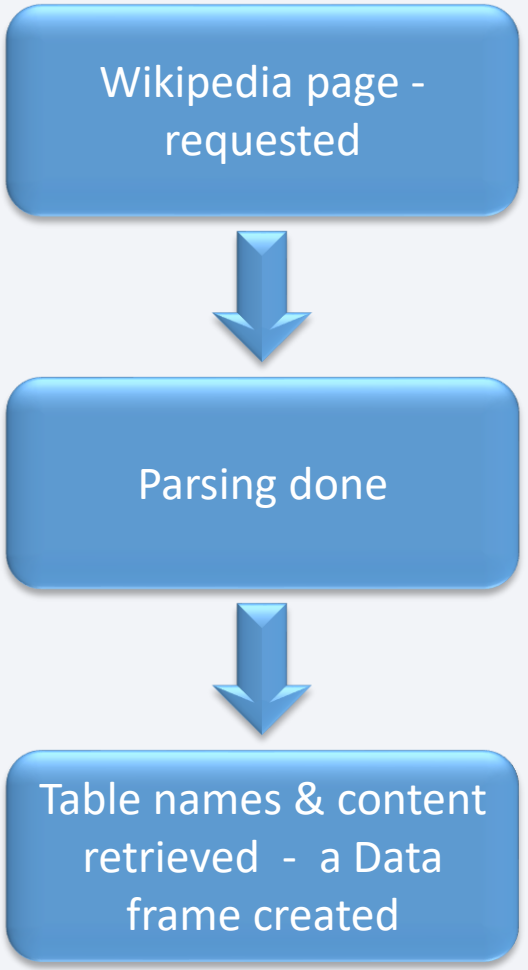
https://github.com/NagalekshmiNK/Applied-Datascience-Capstone-NAG/blob/master/Data%20Collection%20API-NAG.ipynb



Data Collection - Scraping

- The data was web scraped from the Wikipedia page: 'List of Falcon 9 and Falcon Heavy launches'
- Beautiful Soup was used.
- Github link for Data Collection – Web scraping:

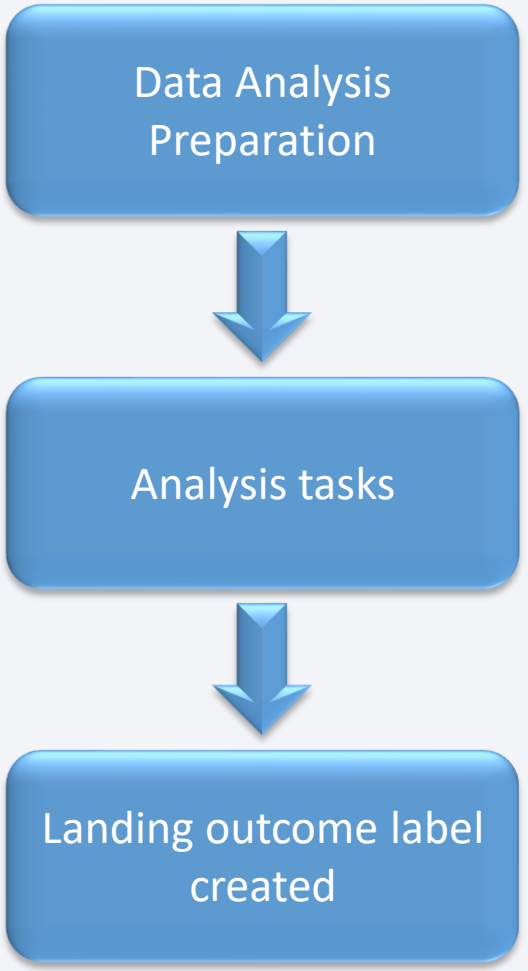
https://github.com/NagalekshmiNK/Applied-Datascience-Capstone-NAG/blob/master/Data%20Collection%20with%20Web%20Scraping-NAG.ipynb



Data Wrangling

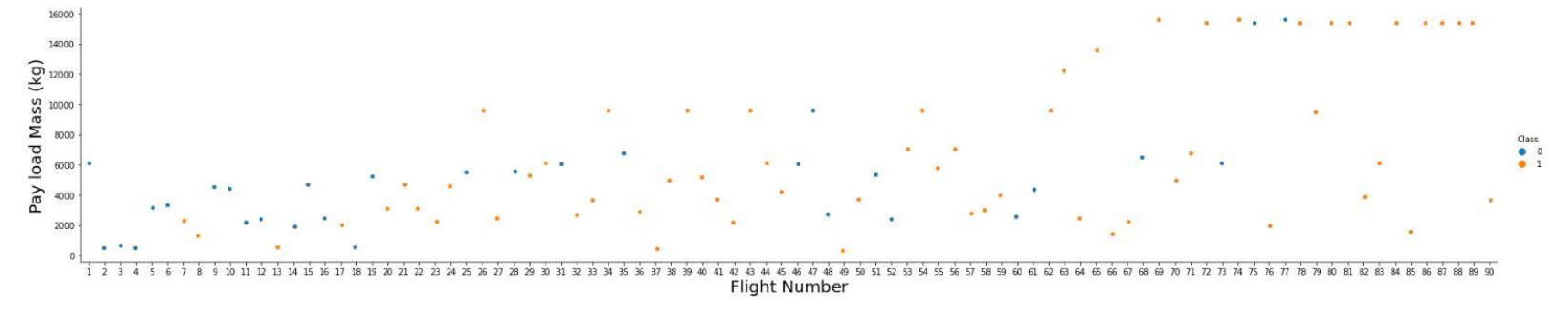
- Exploratory Data Analysis was done.
- Some training labels were determined.
- The data frame was analyzed for obtaining results like, launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type.
- Landing Outcome label was being created from Outcome column.
- Github link for Data Wrangling:

https://github.com/NagalekshmiNK/Applied-Datascience-Capstone-NAG/blob/master/Data%20Wrangling-NAG.ipynb



EDA with Data Visualization

- Greater understanding are made through visualization. Hence, we use visualization methods like scatterplot and barplot to visualize the data.
- The relationship with various features are being visualized. Let us see a sample output:



- Github link for EDA with Data Visualization:

https://github.com/NagalekshmiNK/Applied-Datascience-Capstone-NAG/blob/master/EDA%20with%20Data%20Visualization-NAG.ipynb

EDA with SQL

- There were 10 queries performed on the data

1. Names of the unique launch sites in the space mission
2. Five records where launch sites begin with the string 'CCA'
3. Total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. Date when the first successful landing outcome in ground pad was achieved
6. Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
7. Total number of successful and failure mission outcomes
8. Names of the booster versions which have carried the maximum payload mass
9. Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

- Github link for EDA with SQL:

https://github.com/NagalekshmiNK/Applied-Datascience-Capstone-NAG/blob/master/EDA%20with%20SQL-NAG.ipynb

Build an Interactive Map with Folium

- Folium was used to create Maps with Markers, circles, lines and marker clusters.
- They were used for:
    a) Markers - launch sites, etc.
    b) Circles - highlighted areas around coordinates like, NASA JSC
    c) Lines – distances between two coordinates
    d) Marker clusters - launches in a launch site, etc.
- Github link for Interactive Map with Folium:

https://github.com/NagalekshmiNK/Applied-Datascience-Capstone-NAG/blob/master/Interactive%20Visual%20Analytics%20with%20Folium-NAG.ipynb

Build a Dashboard with Plotly Dash

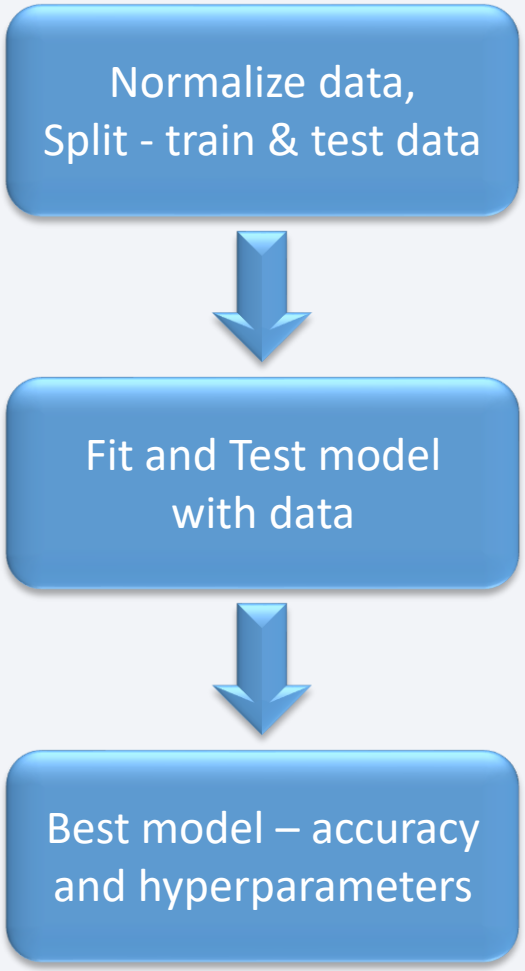
- Using Plotly Dash, pie-charts and scatter plots had been added.
- Pie-chart was used to indicate the percentage of launches by site
- Scatter plot was used to visualize the difference occurring across payload range.
- Thus, an easy understanding of relationship between Launch sites, payloads and booster version is obtained by using both the plots.
- Github link for Dashboard with Plotly:

https://github.com/NagalekshmiNK/Applied-Datascience-Capstone-NAG/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Built FOUR Classification models:
    I. Logistic Regression
    II. Support Vector Machine
    III. Decision Tree
    IV. K Nearest Neighbors
- The data was normalized and split randomly into train and test data
- Then, fit into the model for evaluation and the best model based on accuracy was determined.
- Github link for Dashboard with Predictive Analysis :

https://github.com/NagalekshmiNK/Applied-Datascience-Capstone-NAG/blob/master/Machine%20Learning%20Prediction-NAG.ipynb



Results

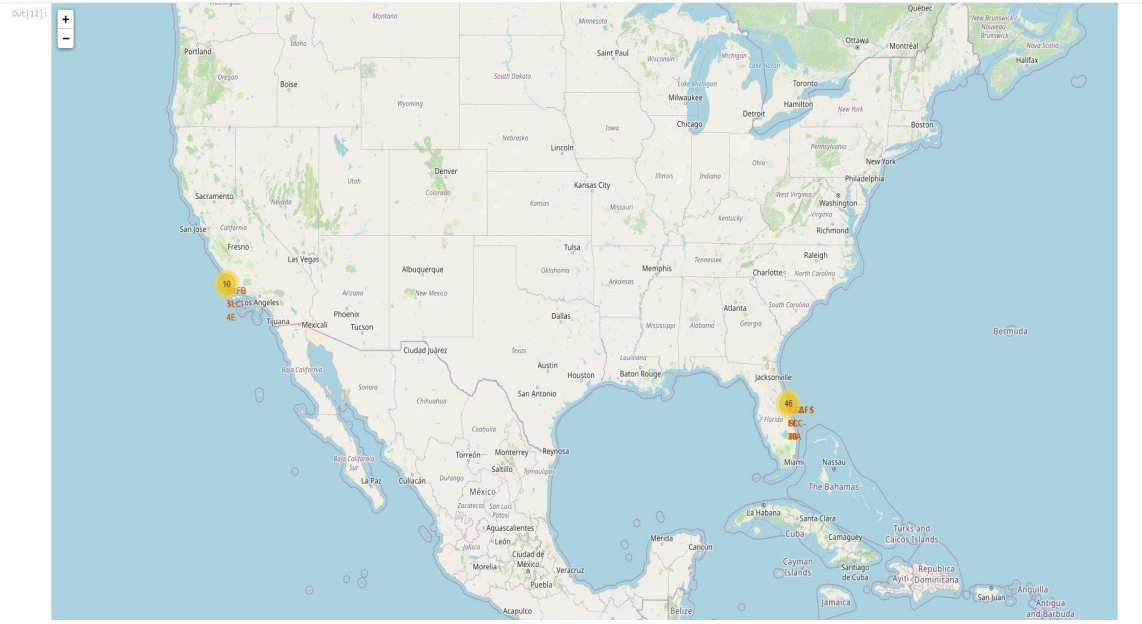
Exploratory data analysis results

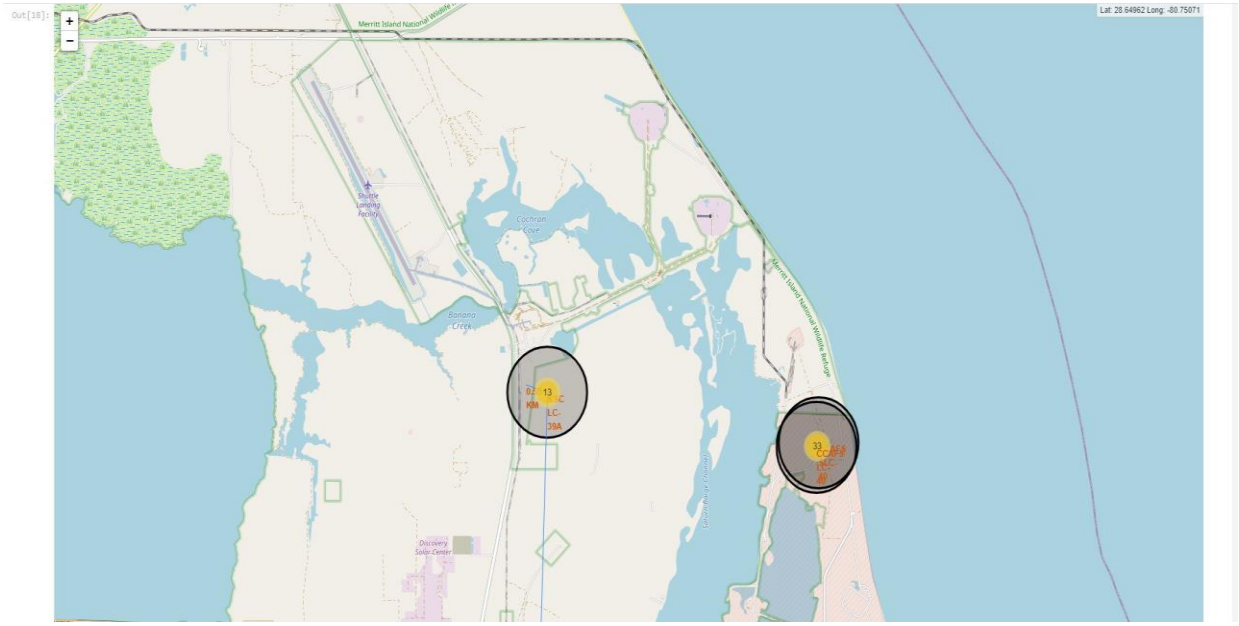
1. CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E are the four different launch sites of Space X.
2. The first launches were done to Space X by itself and NASA.
3. The average payload of F9 v1.1 booster is 2,928 kg.
4. The first success landing outcome in ground-pad happened on 22-12-2015, five year after the first launch.
5. Many Falcon 9 booster versions were successful at landing in drone ships having payload between 4000 and 6000 kg.
6. Almost 100% of mission outcomes were successful.
7. Two booster versions failed at landing in drone ships in 2015 are F9 v1.1 B1012 and F9 v1.1 B1015, both at the launch site CCAFS LC-40.
8. The number of landing outcomes became as better as years passed

Results

Interactive analytics Results

- Interactive analytics helped in ruling out the safety of the launch sites by locating them near safe places like sea, also states that they are away from cities.
- The view of distance between launch sites and means of transport were given.
- Most launches happens at east cost launch sites.



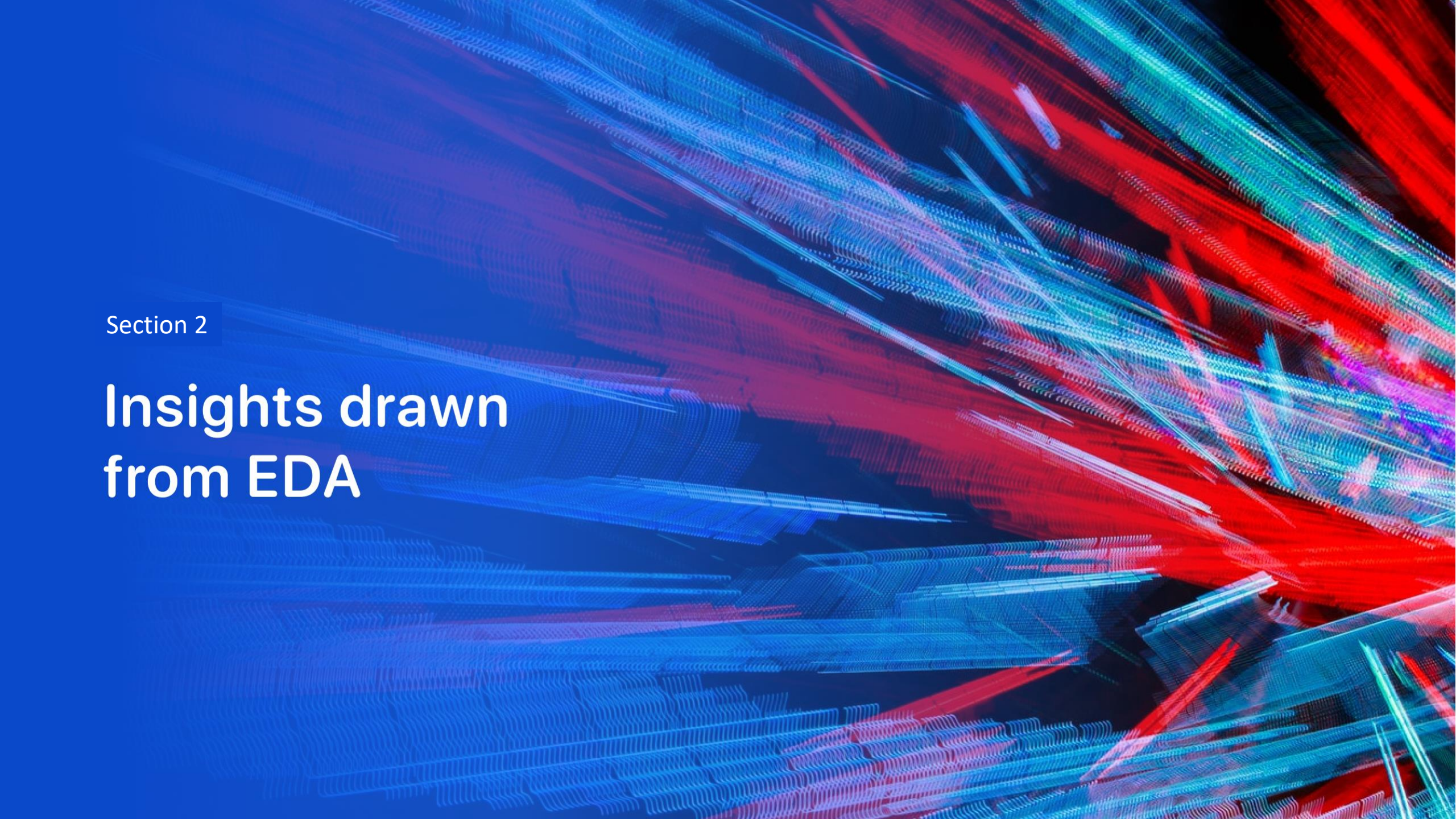


Results

Predictive analysis results

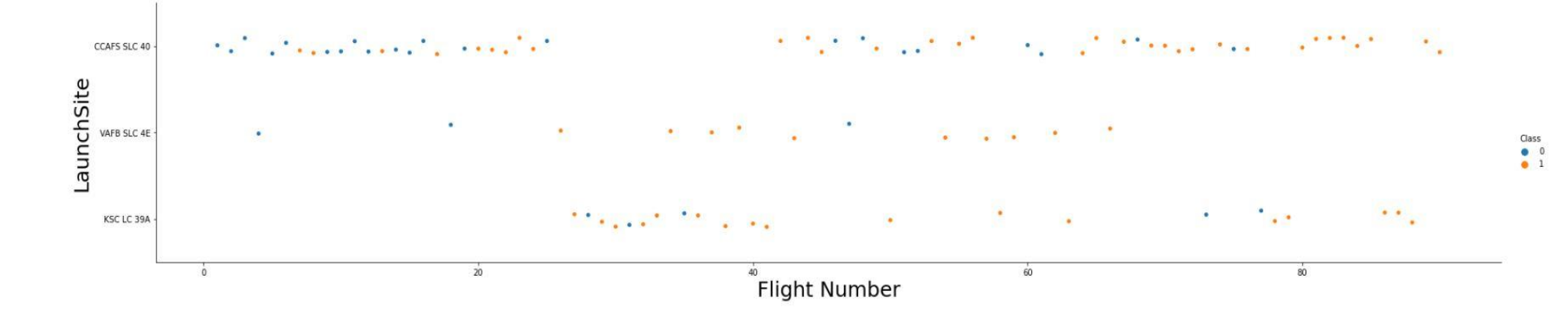
- Decision Tree was finalized as the best model for the prediction of successful landings by Machine Learning Predictive Analysis with an accuracy of 89% and accuracy for test data is 83%
- In the below table, LR, SVM, DT, KNN denotes Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbor respectively.

Model	Accuracy	TestAccuracy
LR	0.84643	0.83333
SVM	0.84821	0.83333
DT	0.88929	0.83333
KNN	0.84821	0.83333


Section 2
Insights drawn from EDA

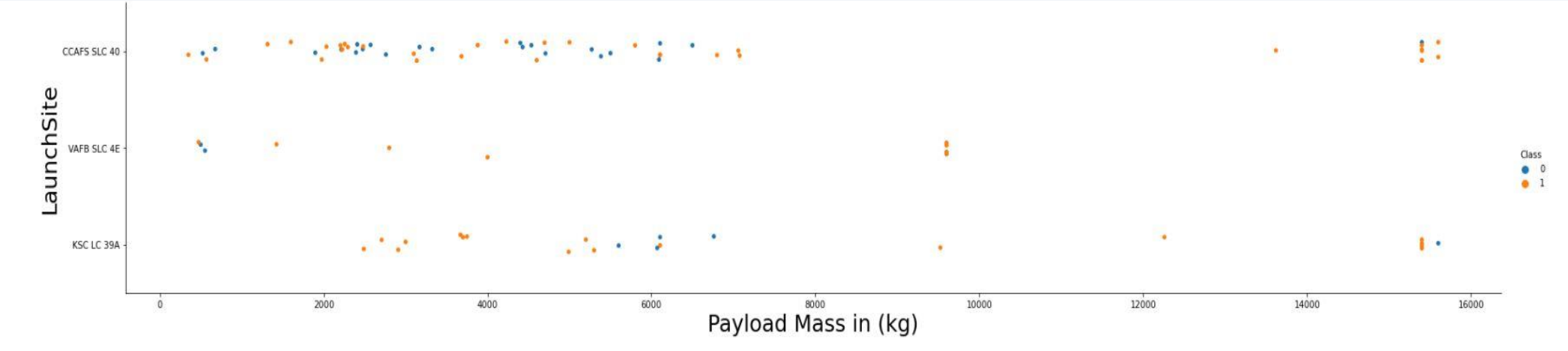
Flight Number vs. Launch Site

- The below plot says that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful. VAFB SLC 4E and KSC LC 39A take the second and third places respectively
- Note that the general success rate is improved over time



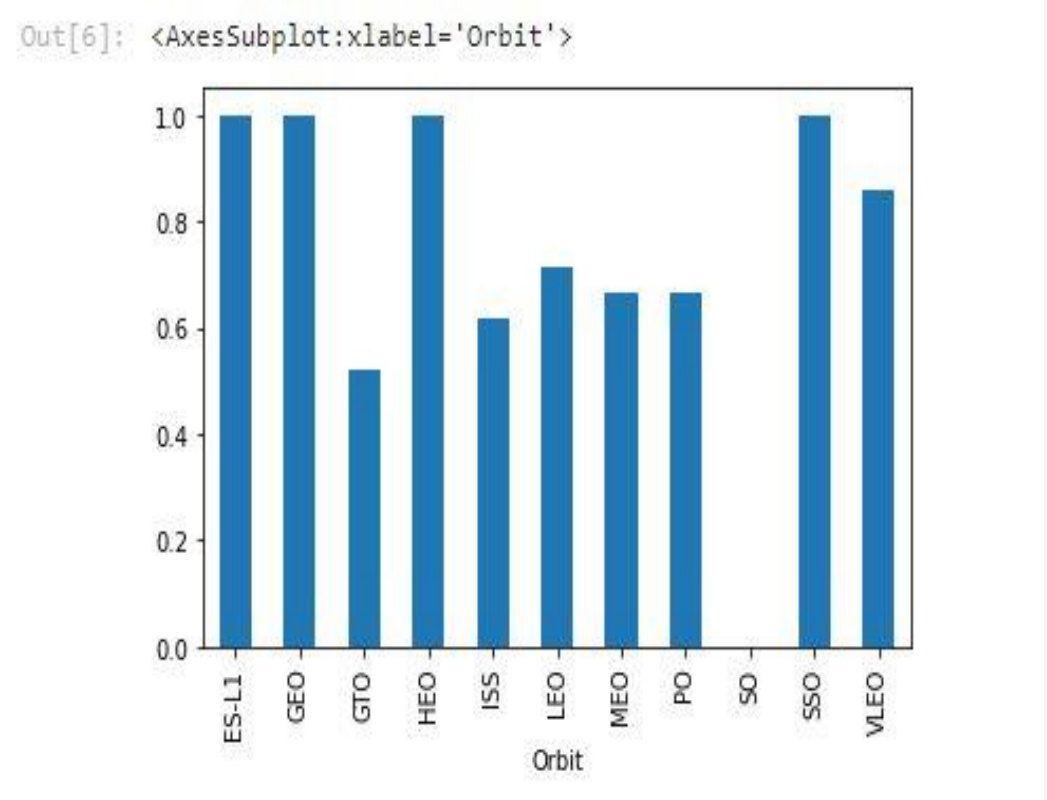
Payload vs. Launch Site

- Payloads over 9,000kg have excellent success rate
- Payloads over 12,000kg seems to have success rate only at CCAFS SLC 40 and KSC LC 39A launch sites



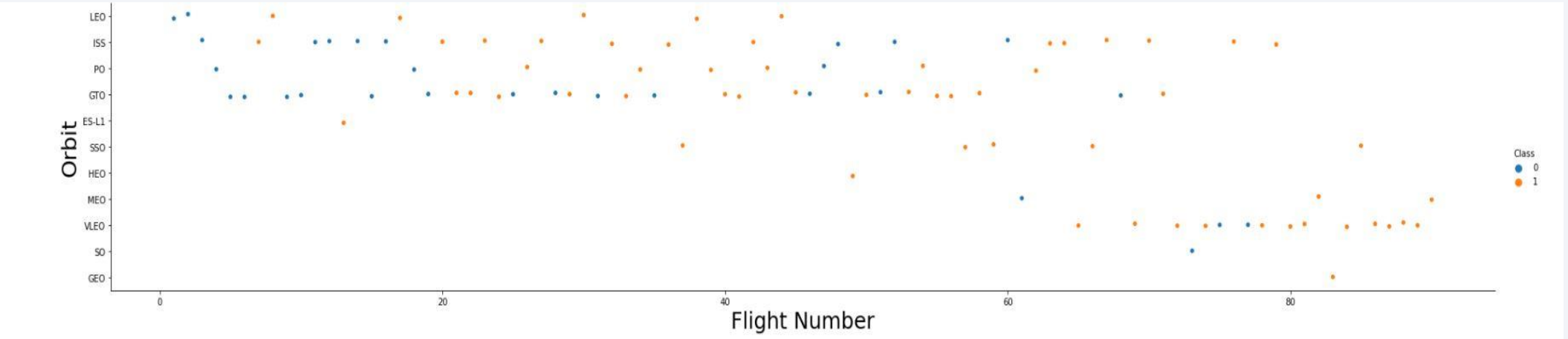
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by:
  - VLEO (above 80%)
  - LEO (above 70%)



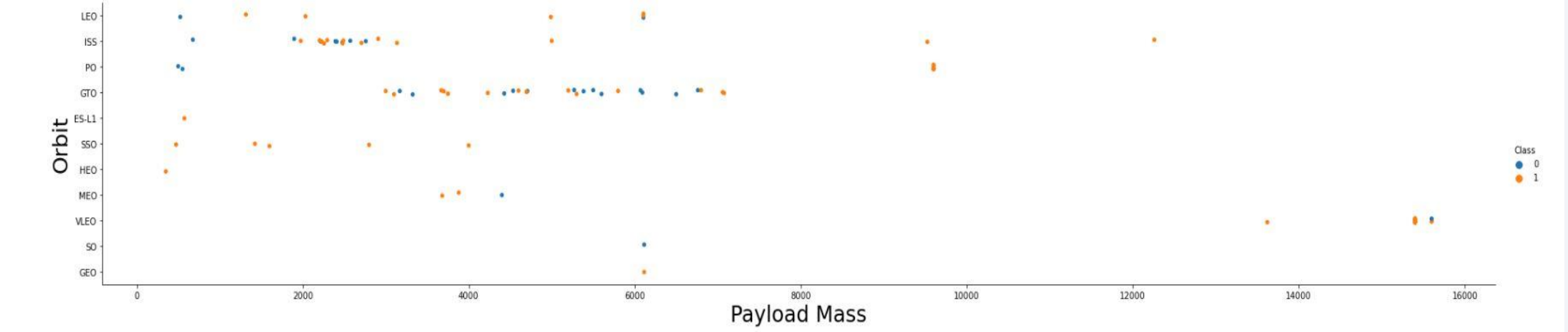
Flight Number vs. Orbit Type

- The success rate improved over time for all the orbits
- Orbit VLEO has a recent increase in success frequency



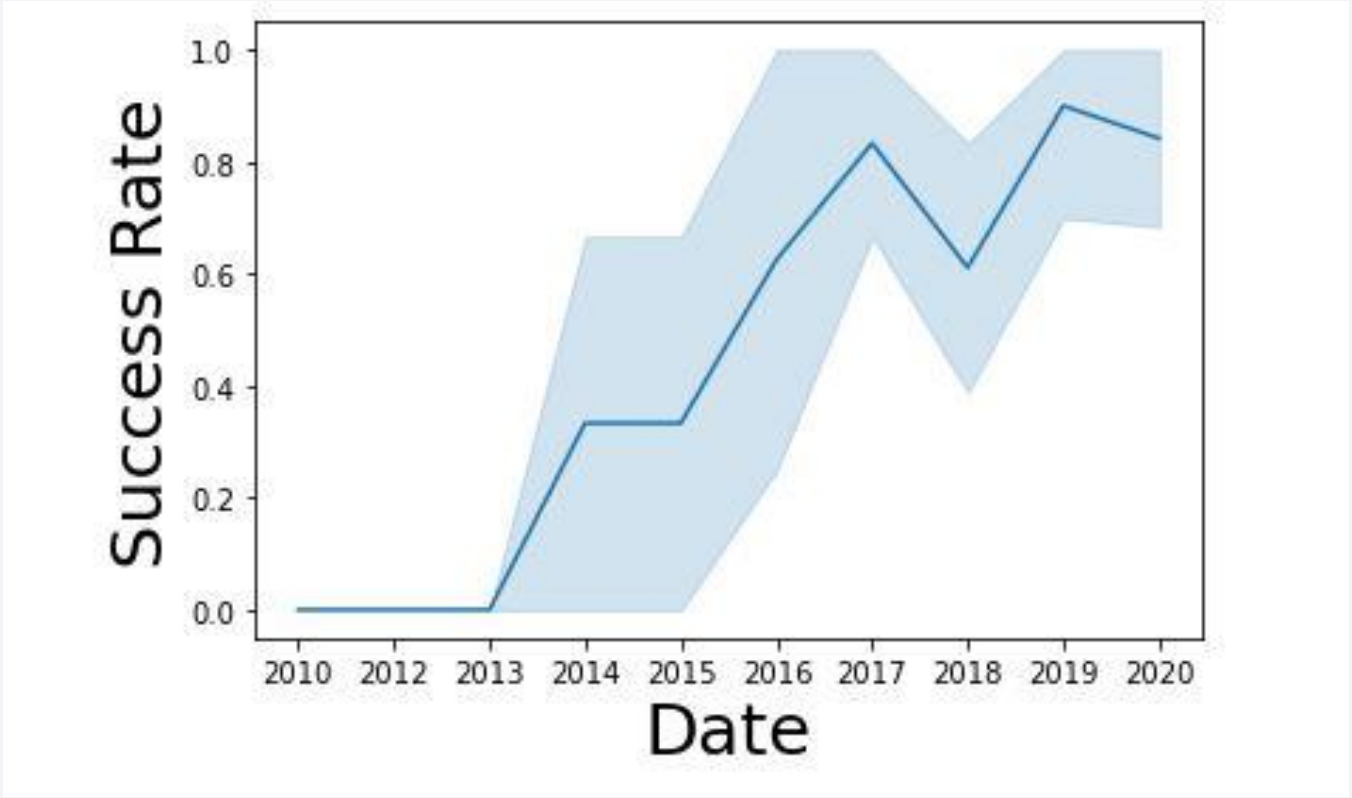
Payload vs. Orbit Type

- There is no relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and a good rate of success
- There are few launches to the orbits SO and GEO



Launch Success Yearly Trend

- The success rate from 2010-2020 is being displayed here.
- Success rate started increasing in 2013 and kept rising until 2020.
- The earlier stages were like a training and testing period for advancing the technology and the needed features.



All Launch Site Names

The four launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

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

In [5]:
```
%%sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[5]:

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

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

In [6]:
```
%%sql
 SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[6]:

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	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total payload carried by boosters from NASA(CRS): 45596 kg

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

In [10]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[10]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1: 2928 kg

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

In [11]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[11]:

average_payload_mass
2928

First Successful Ground Landing Date

Date of the first successful landing outcome on ground pad: 22-12-2015

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

Hint:Use min function

In [12]:
```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[12]:

first_success_gp
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 kg are listed using below query

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

In [13]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[13]:

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

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes: 99 success, I success (payload status unclear) and 1 failure

List the total number of successful and failure mission outcomes

In [14]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL
FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[14]:

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass are listed below

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

In [15]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
ORDER BY BOOSTER_VERSION;
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[15]:

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

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are
  - F9 v1.1 B1012 and CCAFS LC-40
  - F9 v1.1 B1015 and CCAFS LC-40

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

In [16]:
```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[16]:

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

Ranking of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 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

In [20]:
```
%%sql
SELECT LANDING__OUTCOME, COUNT(*) AS TOTAL FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY TOTAL DESC;
```

```
 * ibm_db_sa://qxs86371:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[20]:

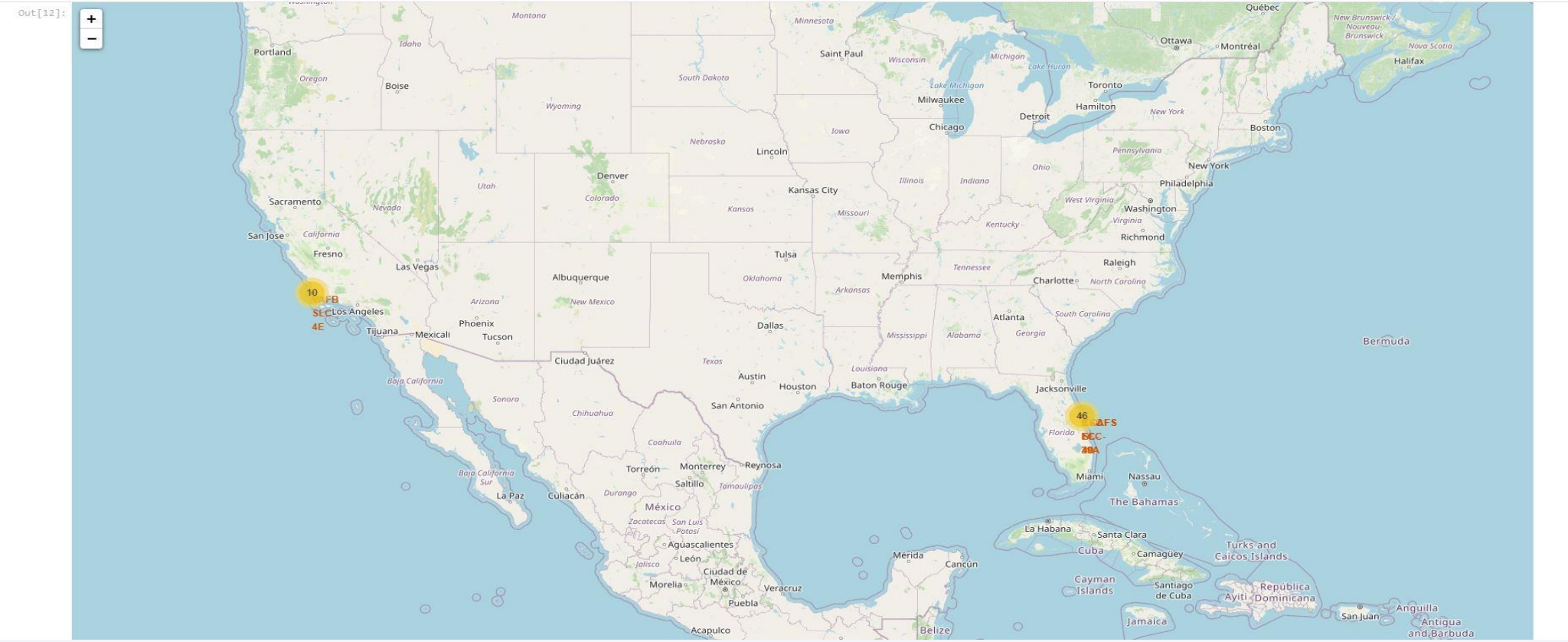
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

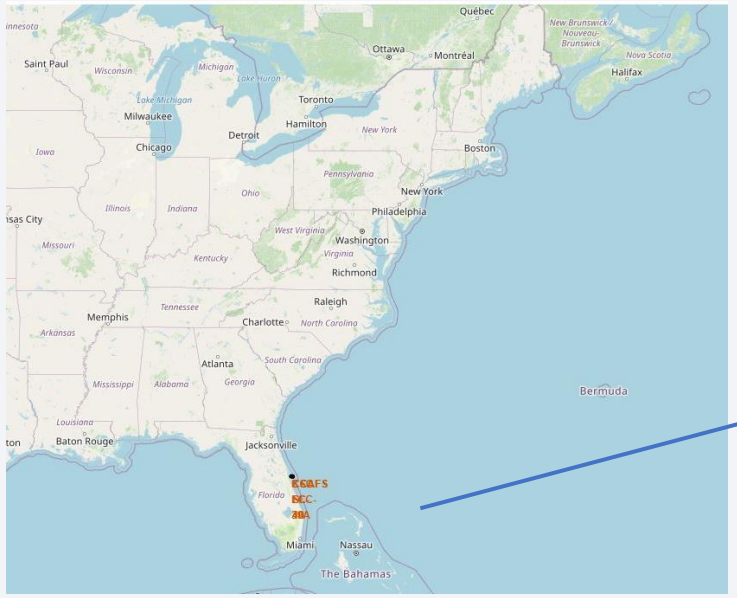
Launch sites

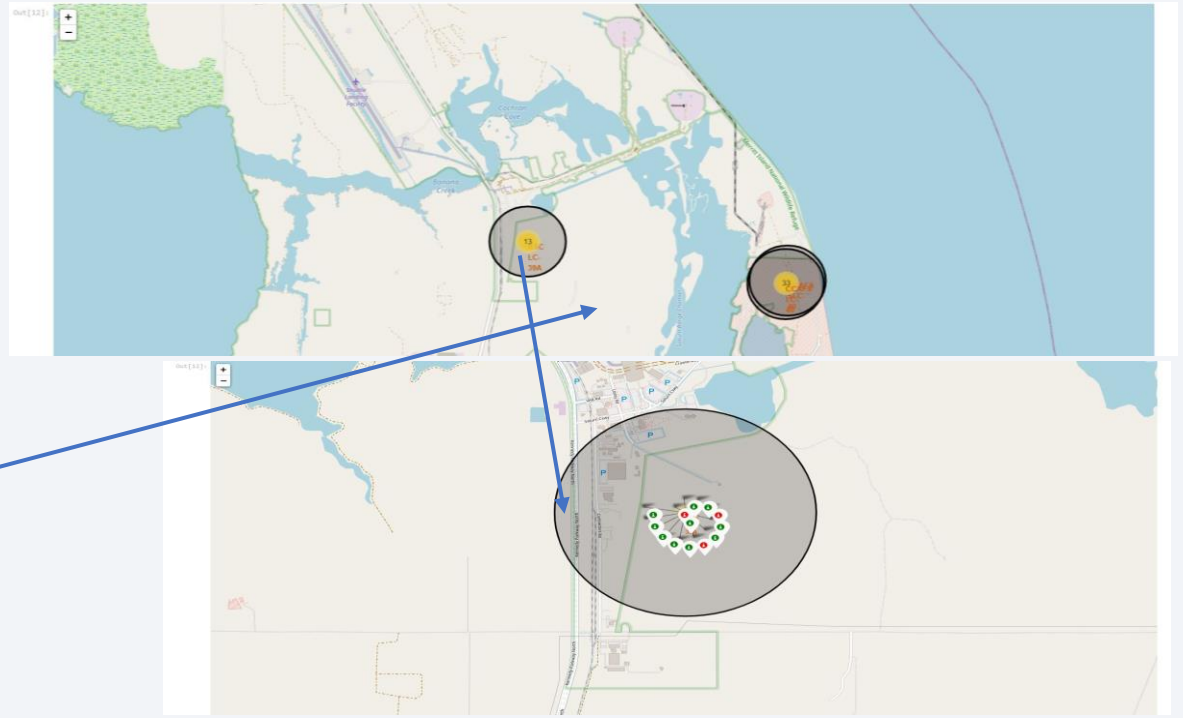
The below image displays all launch sites and shows that they are located at safer distance from the cities and also very near to the sea



Closer view Launch site

- As we zoom into the folium map, the launch sites are located at their accurate locations
- The green and red markers indicate Success and Failure of launches.

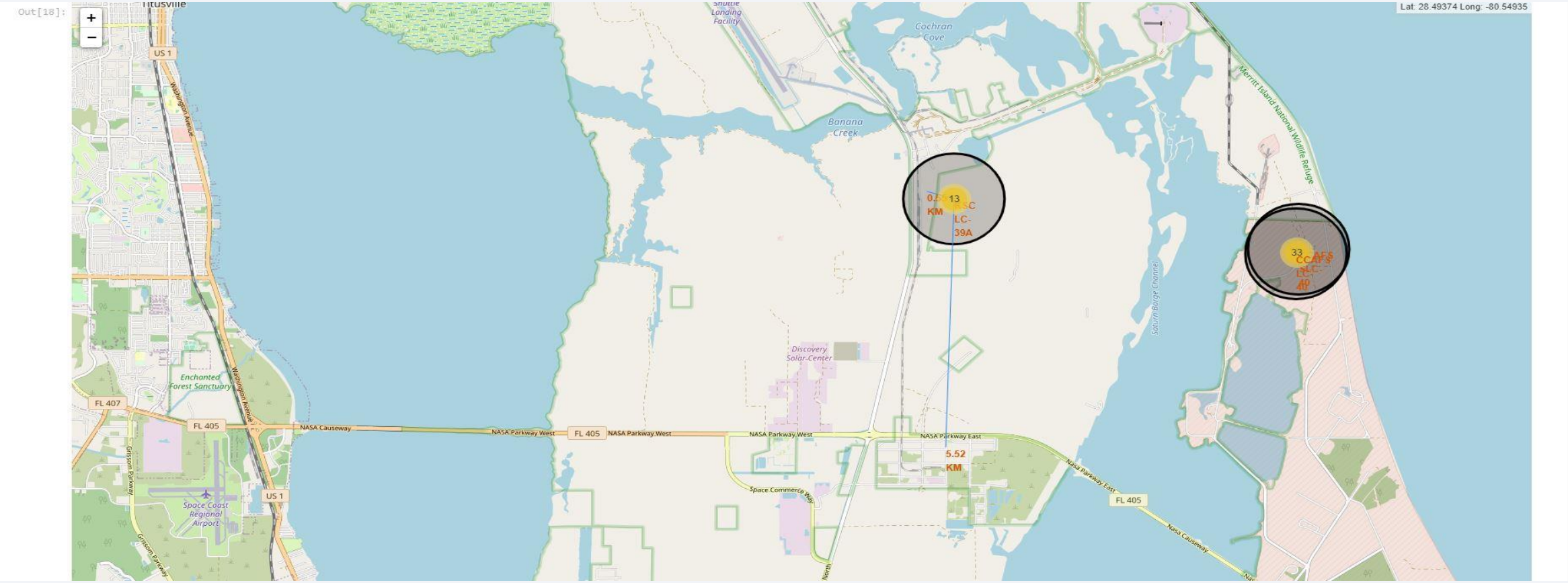


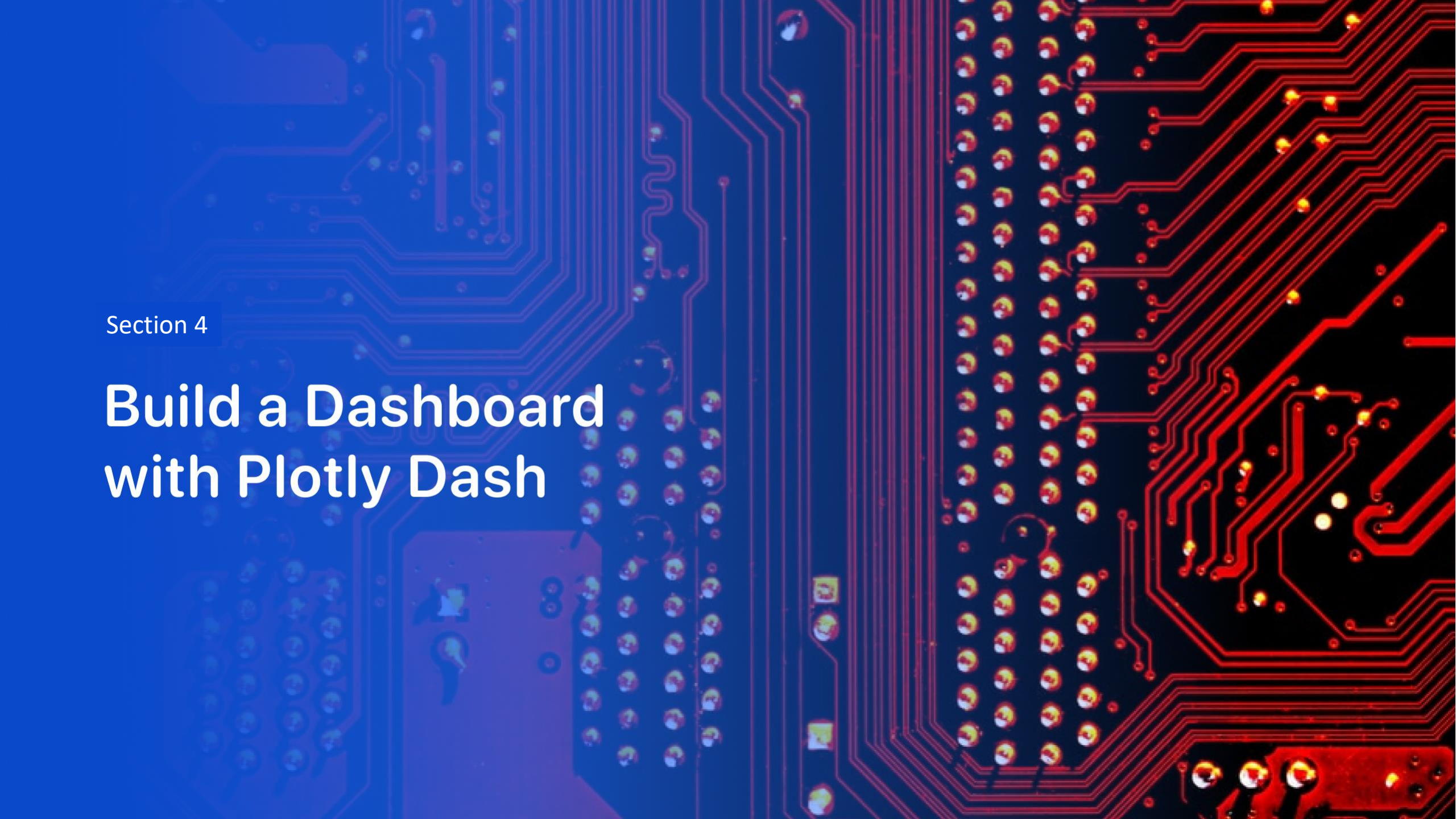


Safety and Proximity to Transport

The below generated folium map explores the proximity of a selected launch site to railway, highway, coastline and the distance is also calculated.

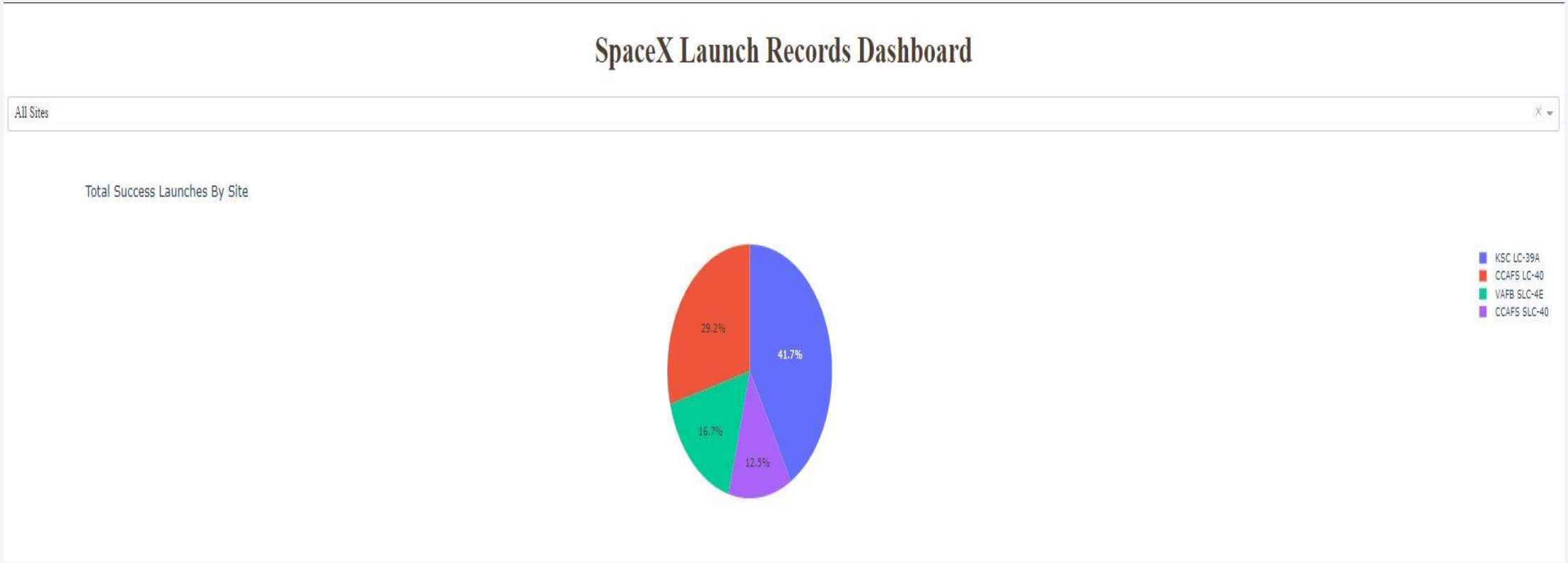
Out[18]:




Section 4
Build a Dashboard with Plotly Dash

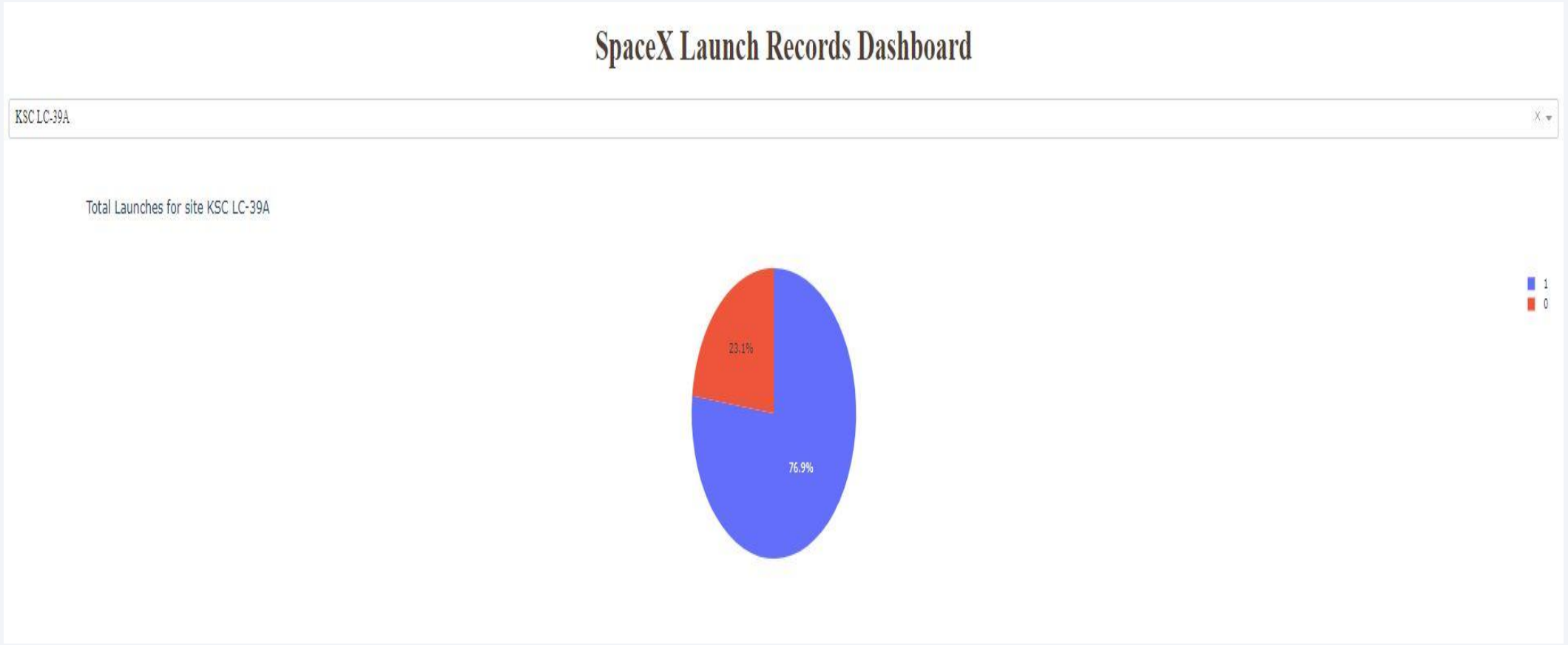
Launch success count for all sites

Launch success count for all sites is displayed using a pie-chart, indicating that launch sites influence success rate



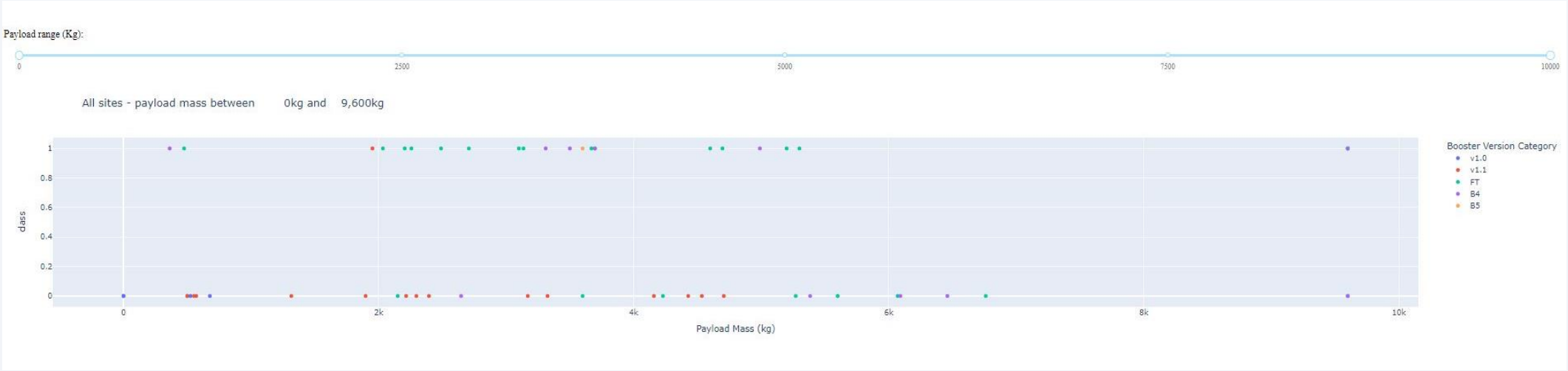
Launch site with highest launch success ratio

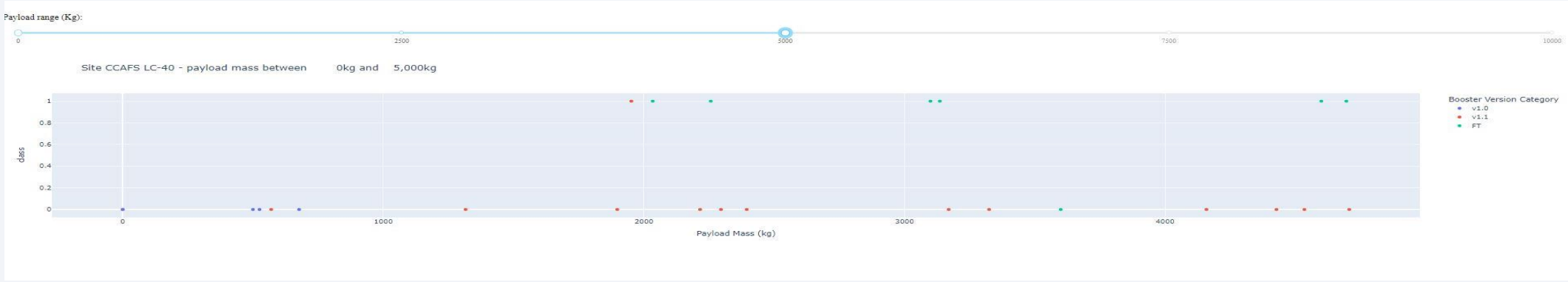
The launch site with highest launch success ratio is KSC LC-39A with 76.9% successful launches



Payload vs. Launch Outcome scatter plot for all sites

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider is displayed below. Payloads under 6,000kg and FT boosters are the most successful combination





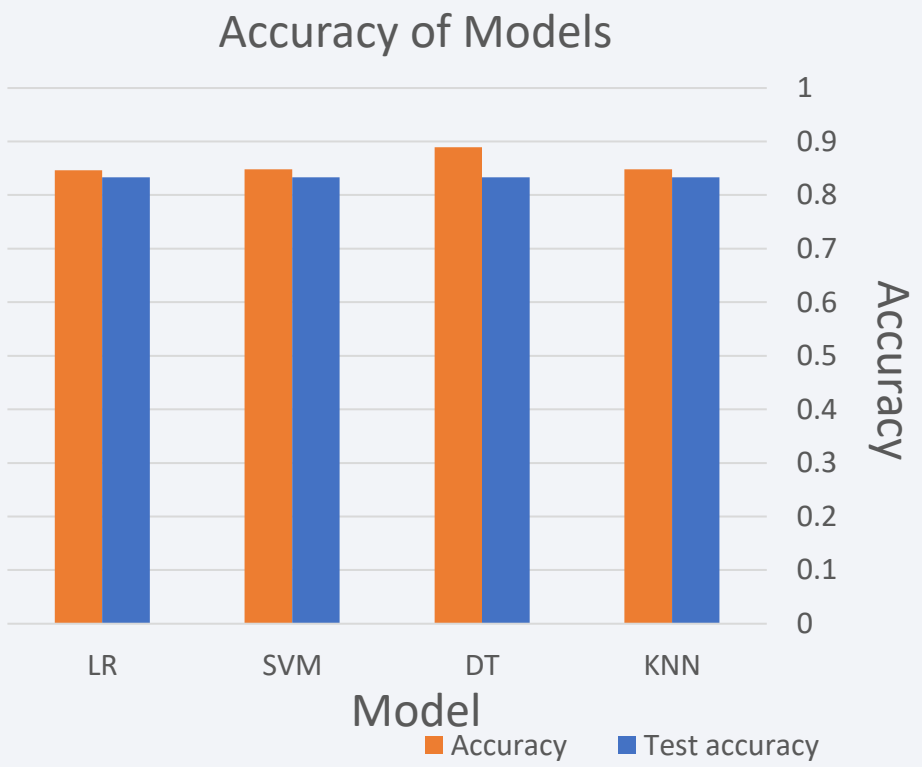
Section 5

Predictive Analysis (Classification)

Classification Accuracy

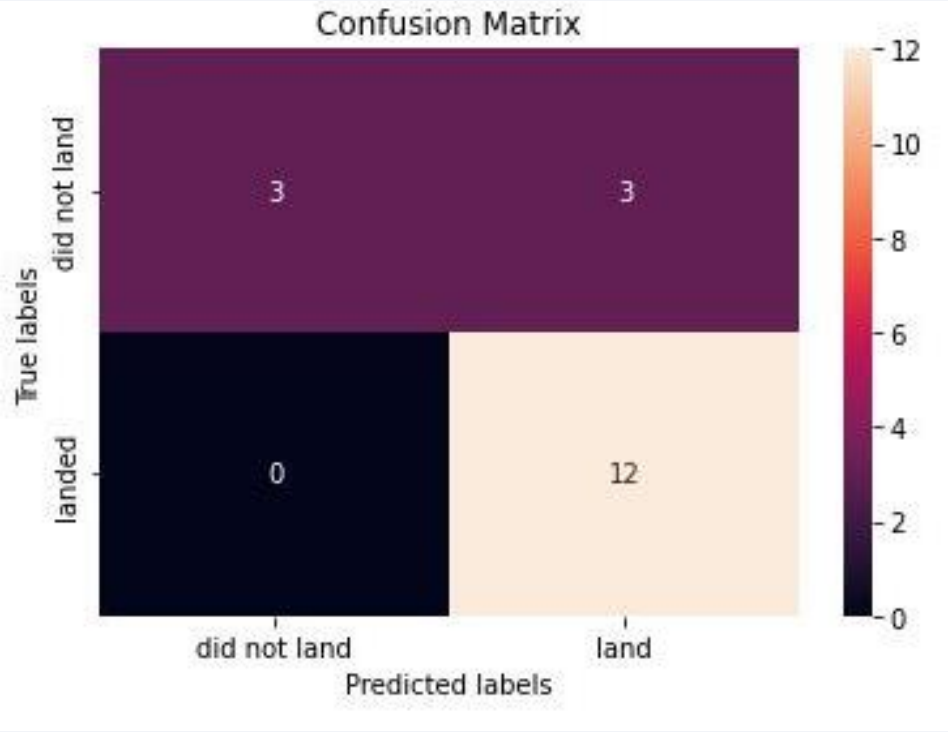
- The accuracies of all the four classification model are displayed here
- The model with the highest classification accuracy is **Decision Tree** with an accuracy of 89% and test accuracy of 83%

Model	Accuracy	TestAccuracy
LR	0.84643	0.83333
SVM	0.84821	0.83333
DT	0.88929	0.83333
KNN	0.84821	0.83333



Confusion Matrix of Decision Tree

The confusion matrix of the best performing model, Decision Tree is displayed below. It proves its accuracy by showing the larger number of true positive and true negative compared to the false ones.



Conclusions

- Data from two different data sources were collected, pre-processed and analyzed
- Launch sites and the payloads influence the success rate
- The success rate has increased from 2013 and is consistent in 2020
- The launch sites are safer locations
- It is found that the best launch site is KSC LC-39A
- Payload mass under 6,000kg and FT boosters are the most successful combination.
- Decision Tree Classifier is the best suited model to predict successful landings

The above conclusions can be used to enhance the success rate of the launches and thereby increasing the profit and win the space race.

Appendix

- Dataset created at the end of Data collection with Space X API: dataset_part_1.csv
- Dataset created at the end of Data collection with Web scraping: spacex_web_scraped.csv
- Dataset created at the end of Data wrangling: dataset_part_2.csv

Thank you!