IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

In this capstone, we tried to determine if Falcon 9 first stage rocket will land successfully by training a machine learning model and use public information to predict if SpaceX will reuse the first stage. The methodologies performed are;

- Data Collection through both API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL and Data Visualization
- Interactive Visual Analytics with Folium and and Plotly Dash
- Machine Learning Prediction

Based on the results launch success rate started to increase in 2013 till 2020. The orbits, ES-L1, GEO, HEO and SSO, have high success rates. KSC LC-39A had the most successful launches in any sites. Decision tree classifier is the best model that can distinguish between the different.

Introduction

SpaceX 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 SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Problems we want to find answers

- What are the factors that affects the successful landing of first stage rocket?
- Are there any relationship between these factors impacting the success rate?
- Will the first stage land?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by using SpaceX API and web scraping Falcon 9 launch records from Wikipedia
- Perform data wrangling
  - Missing values was replaced and OneHotEncoder was applied
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data sets were collected using the following method,

  (a) Making a get request to the SpaceX Rest API and

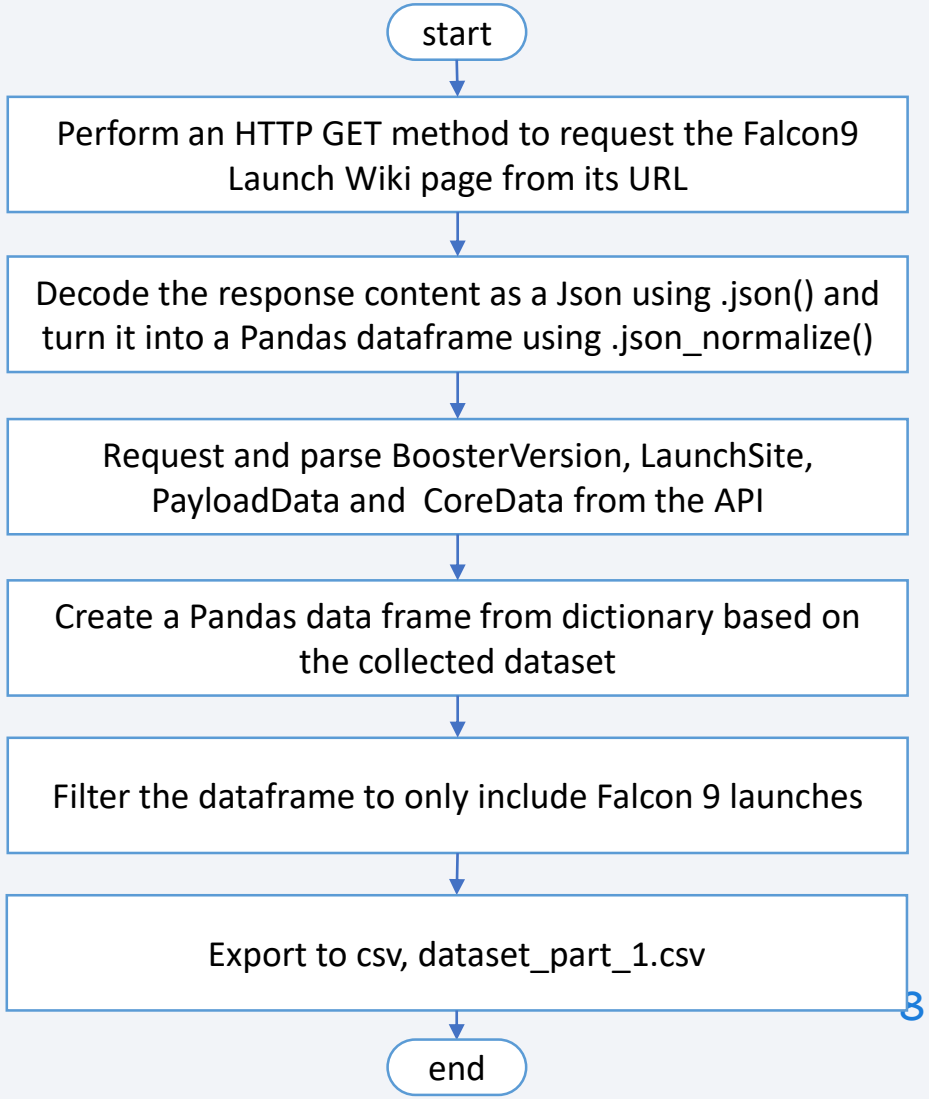
  (b) Web scraping of Falcon 9 launch records from Wikipedia

Data Collection – SpaceX API

- To construct the *dataset_part_ 1.csv*, we made a get request to the SpaceX API using below URL,

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

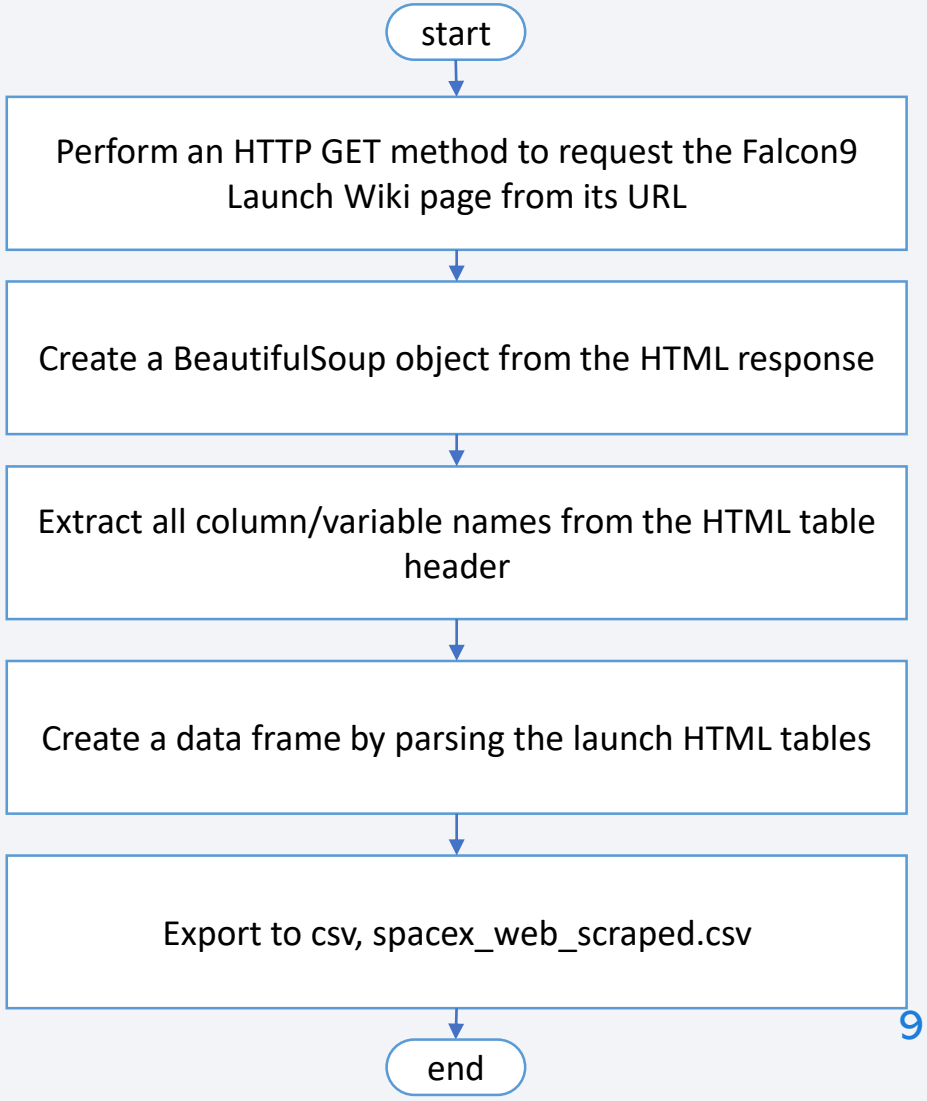
- Completed SpaceX API calls notebook is https://github.com/bellamom/IBM-Applied-Data-Science-Capstone/blob/main/notebook_Collecting_the_data.ipynb



Data Collection - Scraping

- To construct the *spacex_web_scraped.csv*, we did web scraping of Falcon 9 launch records from Wikipedia using below URL, *https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches*

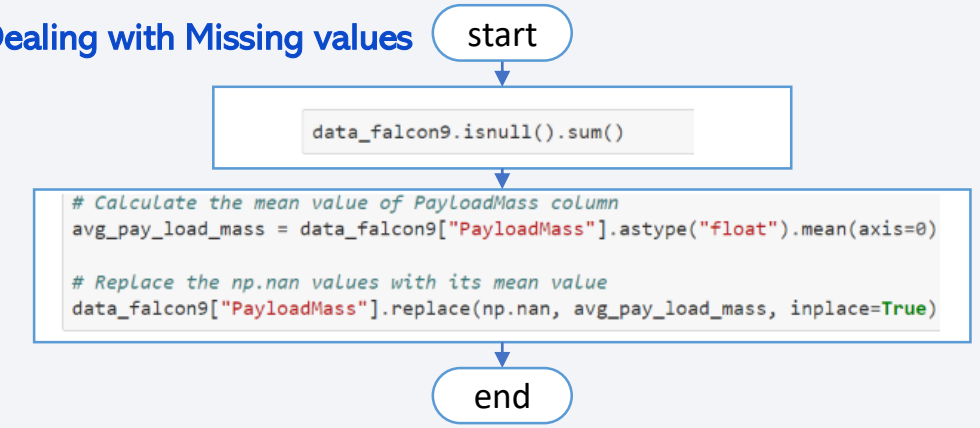
- Completed web scraping notebook is https://github.com/bellamom/IBM-Applied-Data-Science-Capstone/blob/main/notebook_Webscraping.ipynb



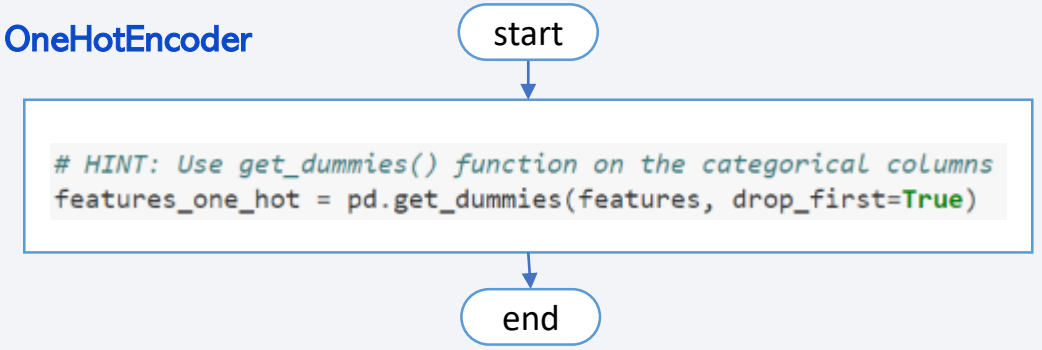
Data Wrangling

- During the data wrangling process of missing values were checked and managed by replacing with calculated mean value
- OneHotEncoder was applied in column Orbits, LaunchSite, LandingPad, and Serial
- Completed data wrangling related notebook, https://github.com/bellamom/IBM-Applied-Data-Science-Capstone/blob/main/notebook_EDA.ipynb

Dealing with Missing values



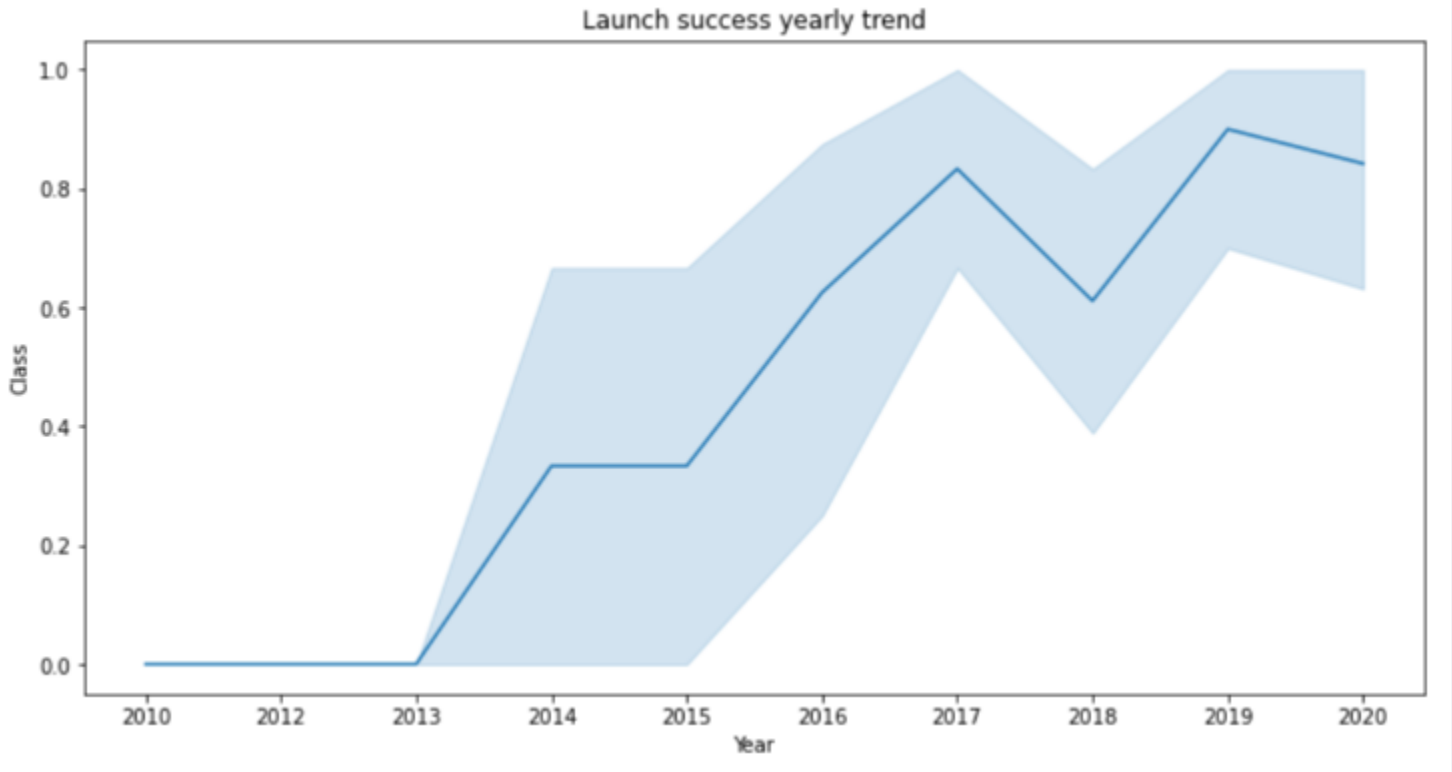
OneHotEncoder



EDA with Data Visualization

EDA by visualization was performed on the following:

- FlightNumber (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome
- Relationship between Flight Number and Launch Site
- relationship between Payload and Launch Site
- relationship between success rate of each orbit type
- relationship between FlightNumber and Orbit type
- launch success yearly trend



Completed EDA with data visualization notebook, https://github.com/bellamom/IBM-Applied-Data-Science-Capstone/blob/main/notebook_Exploratory_Analysis_Using_Pandas_and_Matplotlib.ipynb

EDA with SQL

- To understand the SpaceX dataset, it was loaded into a Db2 database
- SQL queries were performed to,
  - Display the names of the unique launch sites in the space mission
  - Display the 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 acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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
- Completed EDA with SQL notebook, https://github.com/bellamom/IBM-Applied-Data-Science-Capstone/blob/main/notebook_EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Map objects were added to a folium map,
  - To identify each launch site, a circle object based on its coordinate (Lat, Long) values with Launch site name as a popup label using marker object
  - To mark the success/failed launches for each site, MarkerCluster object was used
  - To show the distances between a launch site to its proximities, PolyLine was used

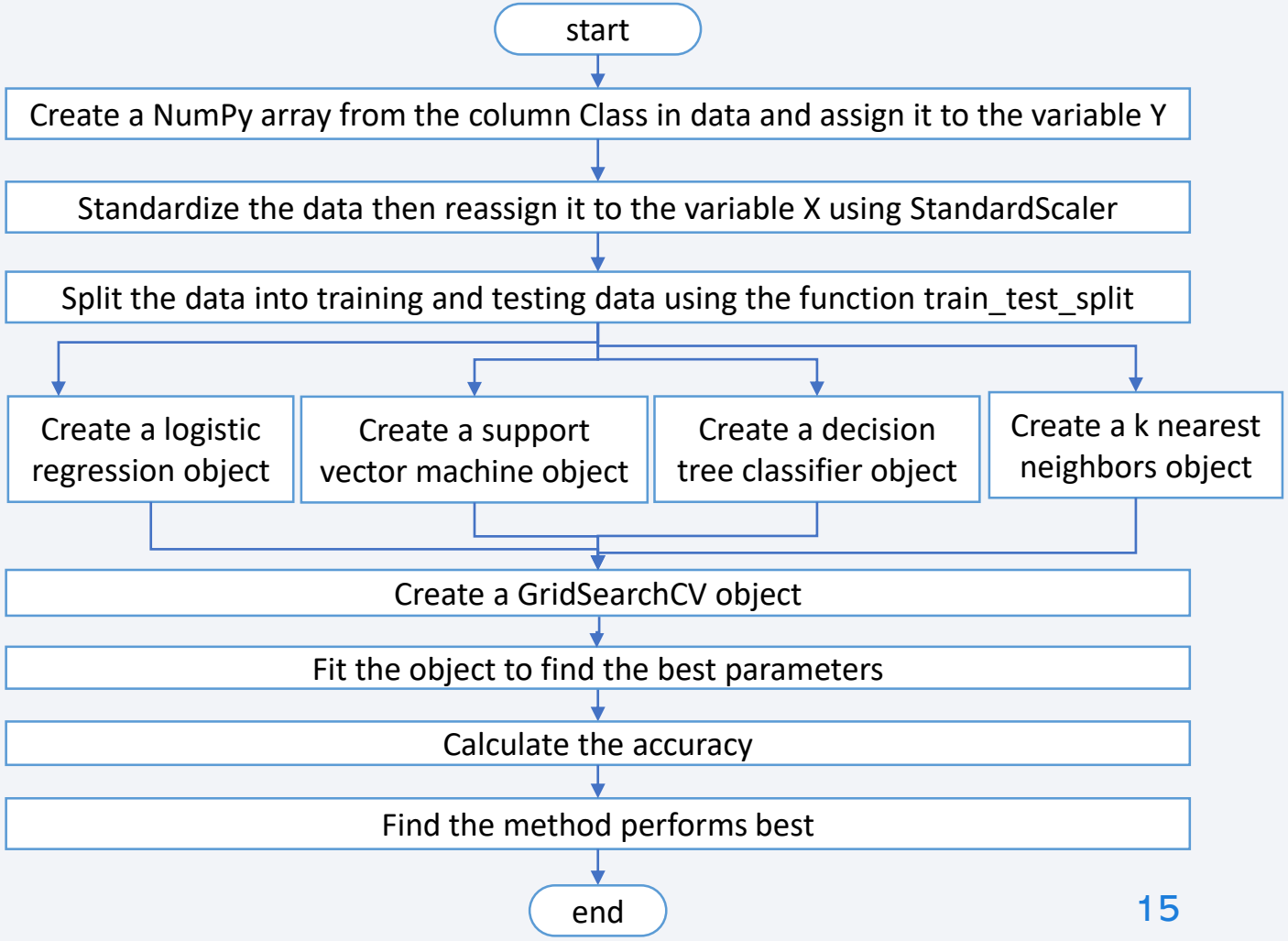
- Completed interactive map with Folium map, https://github.com/bellamom/IBM-Applied-Data-Science-Capstone/blob/main/notebook_Interactive_Visual_Analytics_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- Below components were added to the dashboard
  - Launch Site Drop-down Input Component
  - Callback function to render success-pie-chart based on selected site dropdown
  - Range Slider to Select Payload
  - Callback function to render the success-payload-scatter-chart scatter plot
- Above items were to assist in answering the following questions:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
- Completed Plotly Dash lab, https://github.com/bellamom/IBM-Applied-Data-Science-Capstone/blob/main/dash_app.py

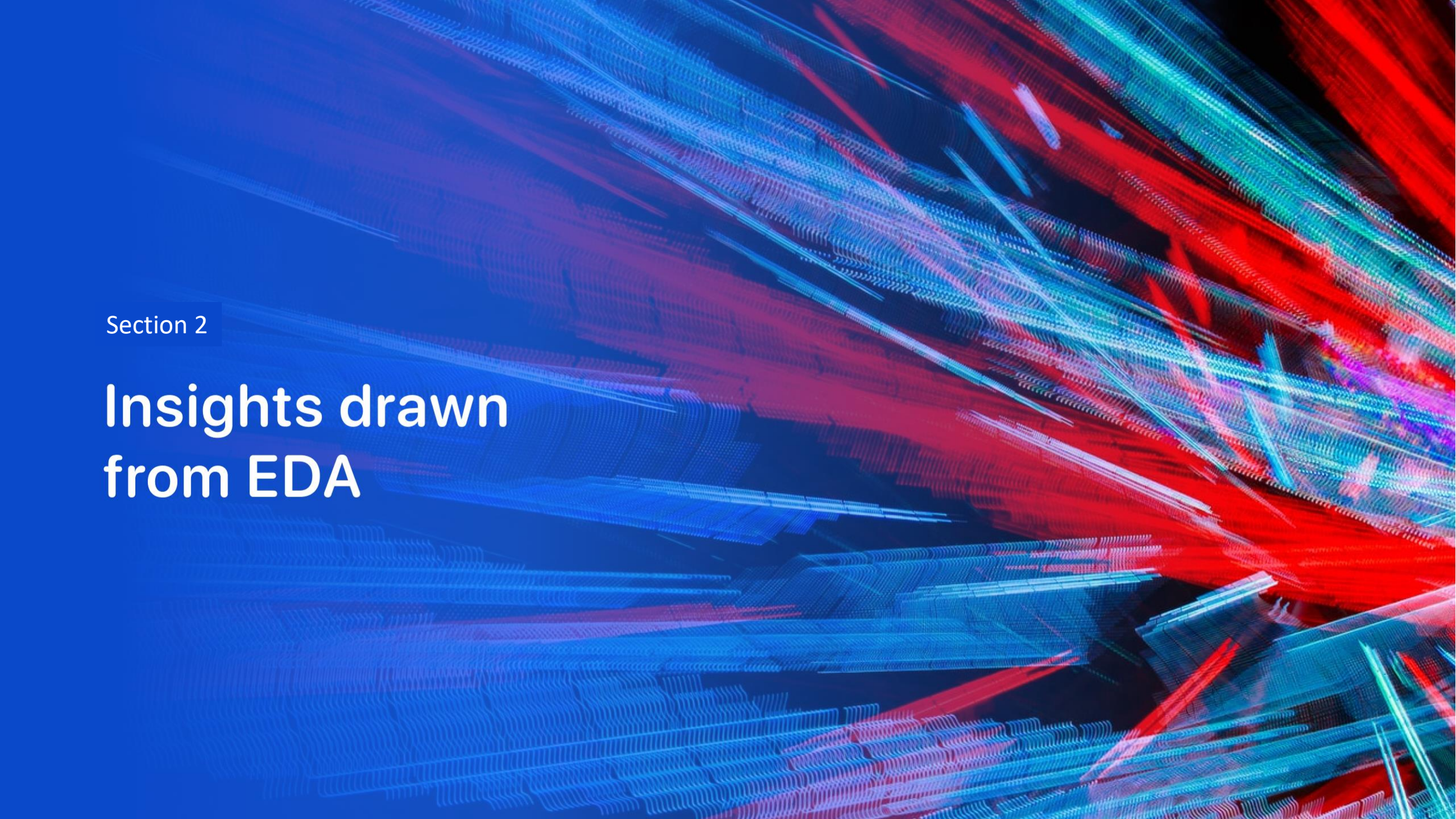
Predictive Analysis (Classification)

- In order to find the best performing method/model for the main task, GridSearchCV was applied to find the best Hyperparameter for Logistic Regression, SVM, Decision Tree Classifier and KNearest Neighbors together with the training and test data.
- Completed predictive analysis lab, https://github.com/bellamom/IBM-Applied-Data-Science-Capstone/blob/main/notebook_Machine_Learning_Prediction.ipynb

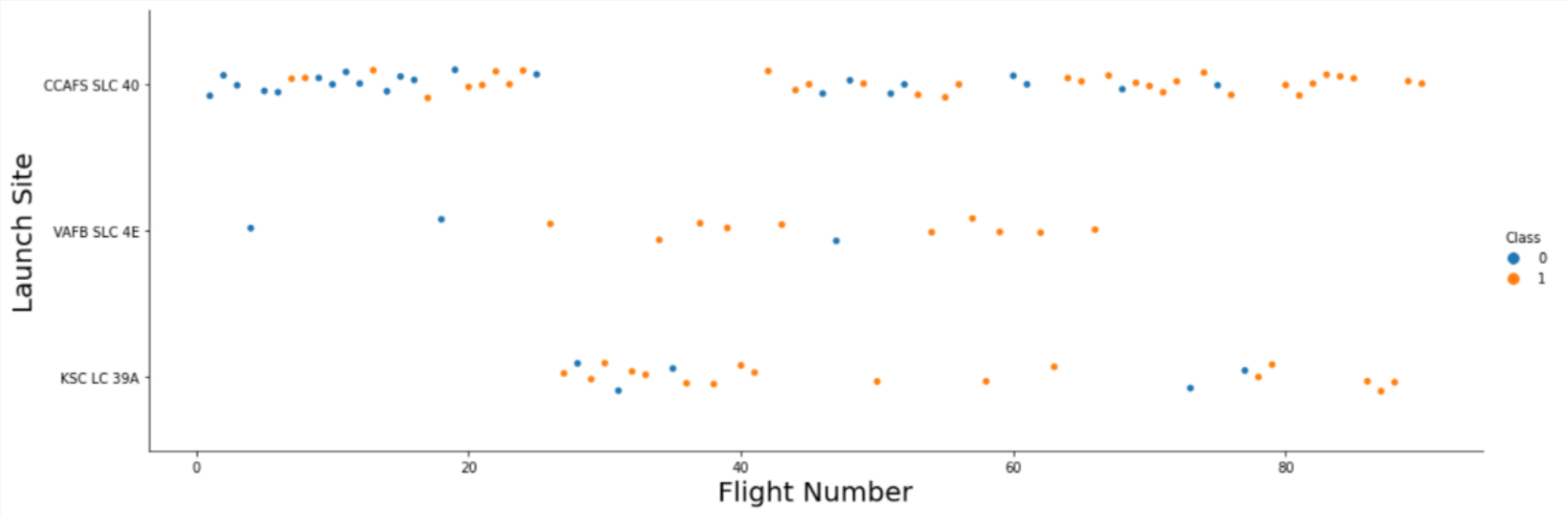


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

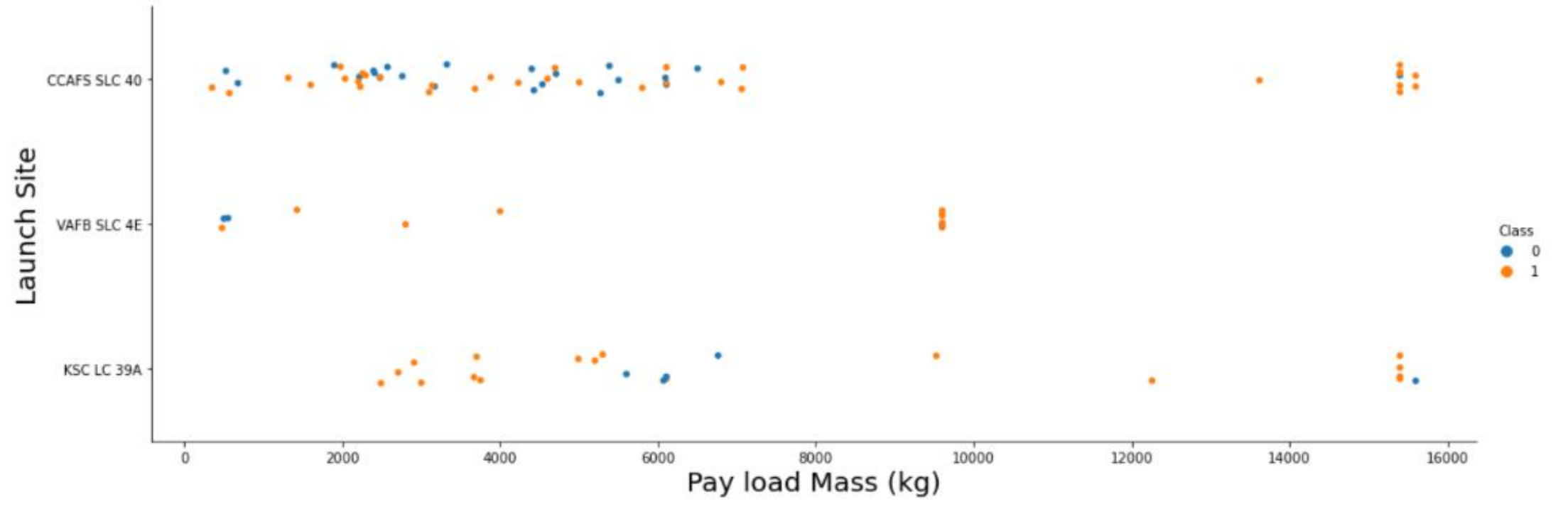

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



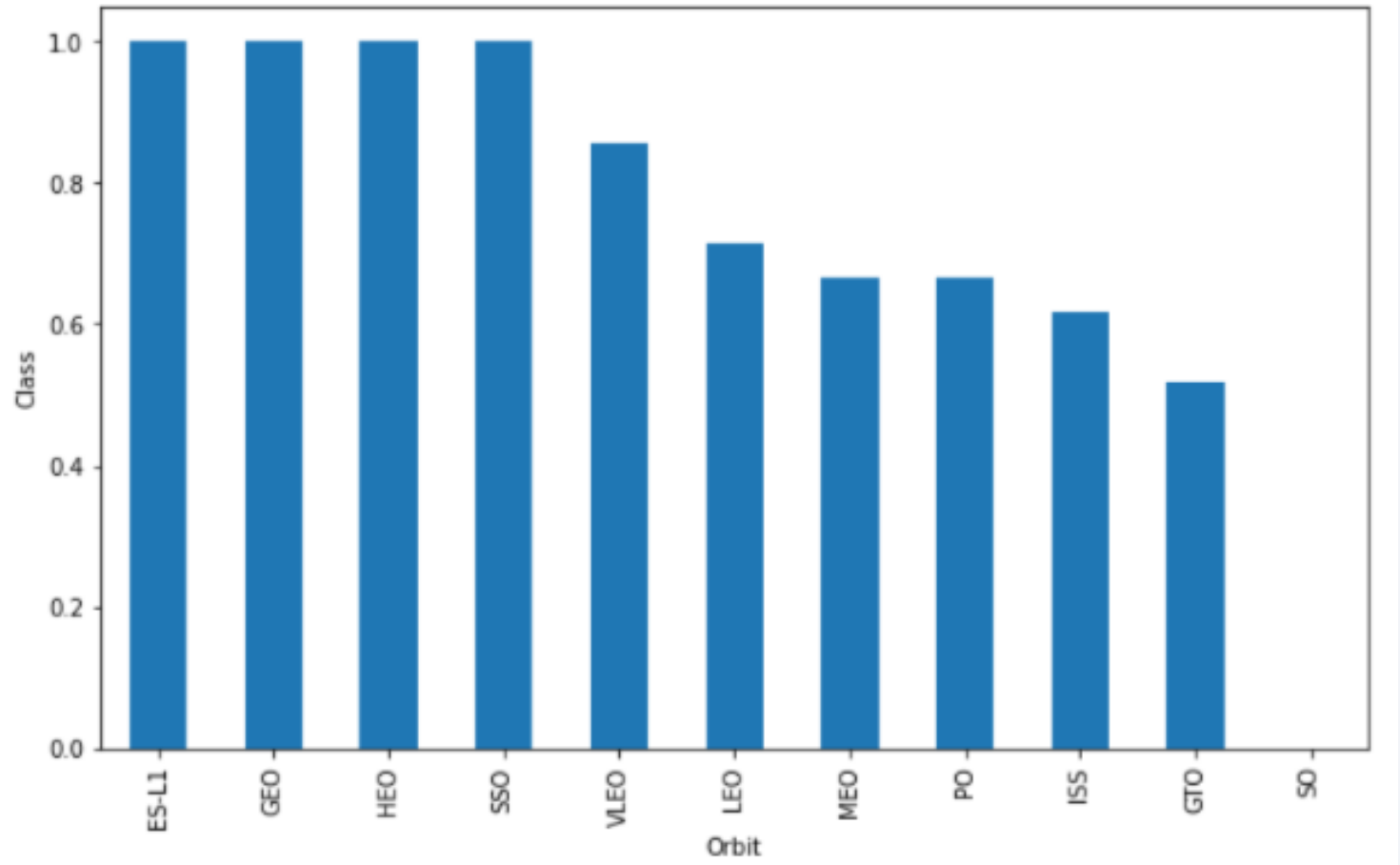
The plot shows that the most recent flights for all launch sites have higher success rate.

Payload vs. Launch Site



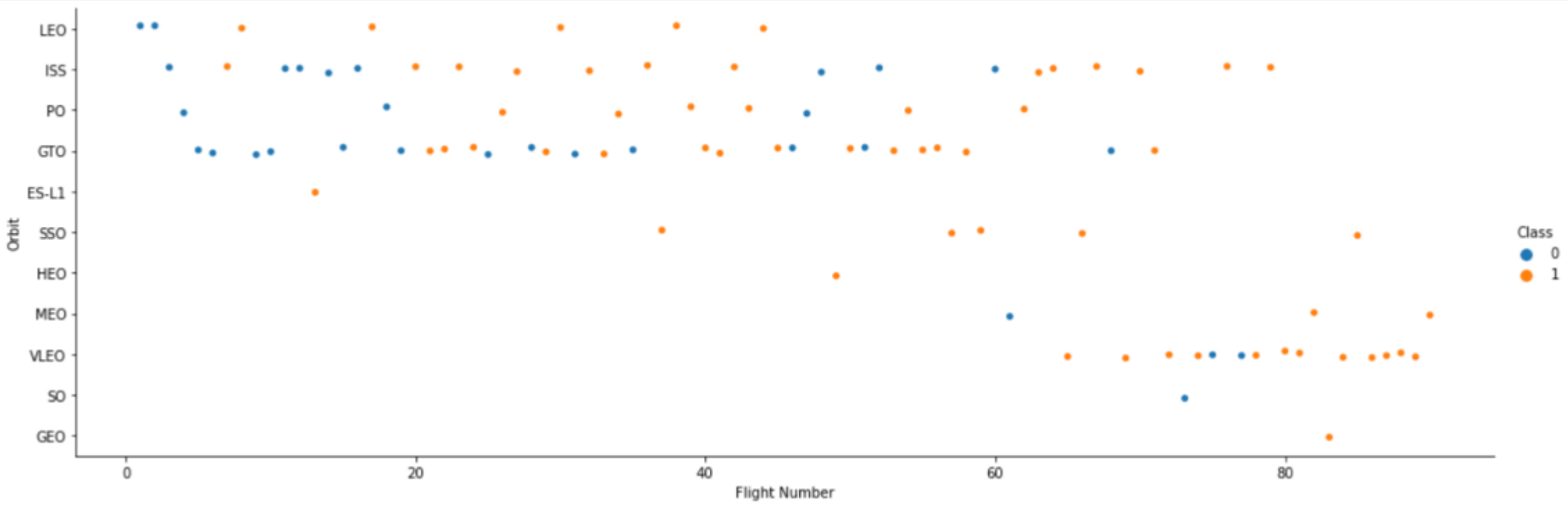
Payload Vs. Launch Site scatter point chart shows that in VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



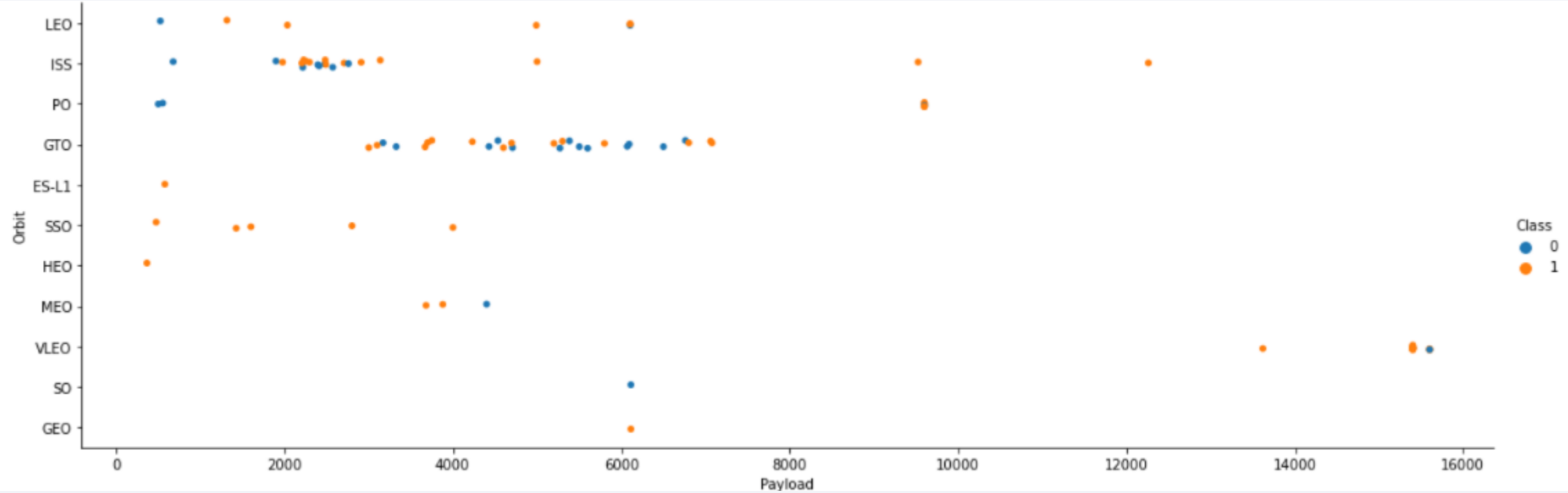
This plotted bar chart shows orbits: ES-L1, GEO, HEO and SSO have high success rate.

Flight Number vs. Orbit Type



In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

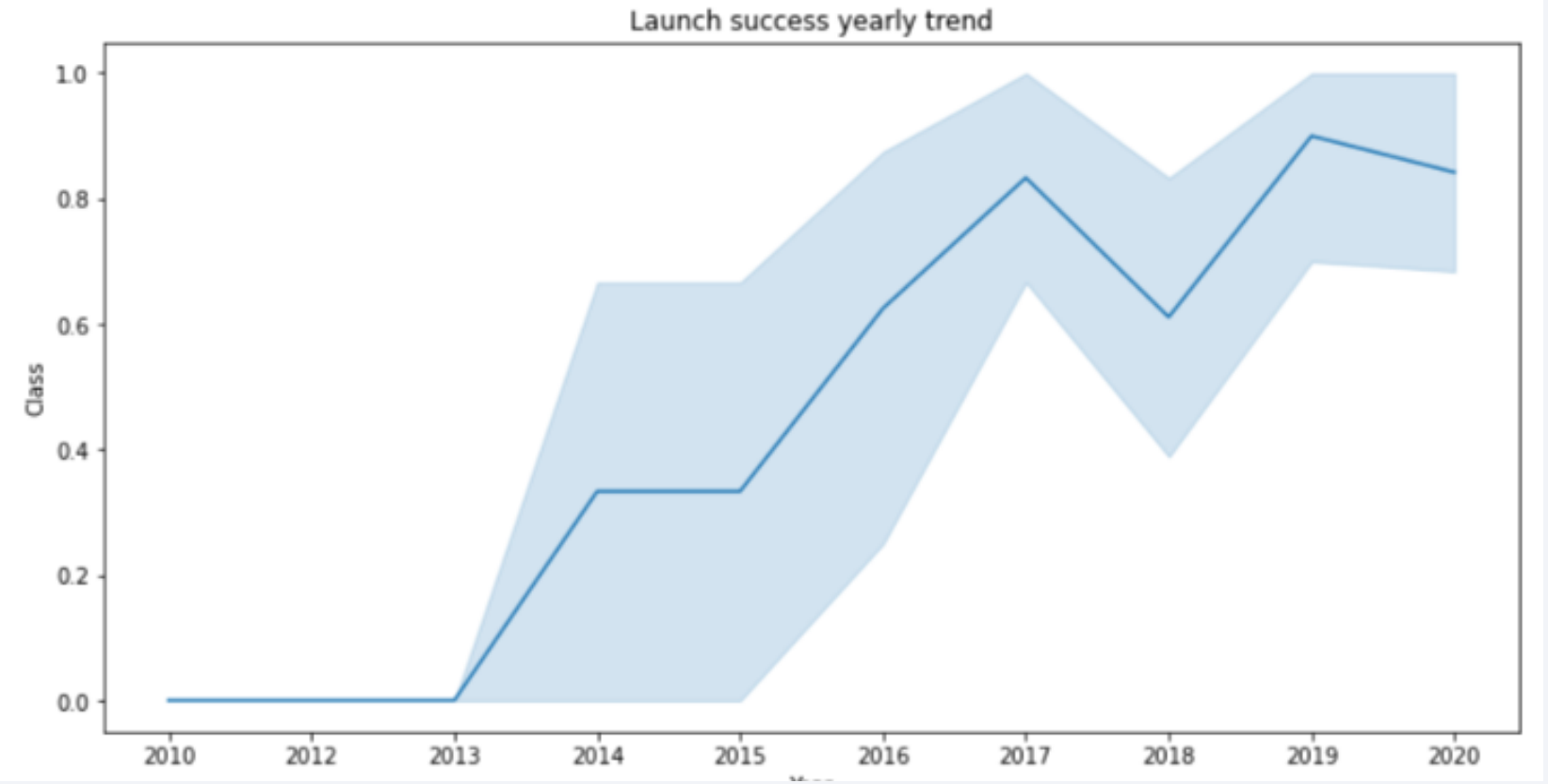
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



The trend shows that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

To get the unique launch sites DISTINCT was used.

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

```
In [6]: %sql select distinct launch_site from SPACEXTBL
```

```
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b
Done.
```

Out[6]:

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

Launch Site Names Begin with 'CCA'

To find 5 records where launch sites begin with `CCA`, LIMIT set to 5 together with filter LIKE was used.

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

In [7]:
```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

Out[7]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

To calculate the total payload carried by boosters from NASA, function SUM() on column payload_mass_kg was used with filter customer = ‘NASA (CRS)’.

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

```
In [8]: %sql select sum(payload_mass__kg_) as "NASA (CRS)" from SPACEXTBL where customer = 'NASA (CRS)'
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.
Done.
```

Out[8]:

NASA (CRS)
45596

Average Payload Mass by F9 v1.1

Function AVG() on column payload_mass_kg for booster_version F9 v1.1 was used to calculate the average payload mass carried by booster version F9 v1.1.

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

In [9]:
```
%sql select avg(payload_mass__kg_) as "F9 v1.1" from SPACEXTBL where booster_version like 'F9 v1.1'
```

```
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appc
Done.
```

Out[9]:

F9 v1.1
2928

First Successful Ground Landing Date

To find the dates of the first successful landing outcome on ground pad, function MIN() on column date was used.

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

Hint:Use min function

```
In [10]: %sql select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)'
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00
Done.
```

Out[10]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, below SQL query was used.

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

In [16]:
```
%sql select distinct booster_version from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
```

```
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

Out[16]:

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

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failure mission outcomes, count with group by column mission_outcome was used

List the total number of successful and failure mission outcomes

In [12]:
```
%sql select mission_outcome, count(*) from SPACEXTBL group by mission_outcome
```

```
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0
Done.
```

Out[12]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

To list the names of the booster which have carried the maximum payload mass, a subquery with function MAX() was used.

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

In [13]:
```
%sql select distinct booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

```
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

Out[13]:

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

To list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, below query was used.

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

In [14]:
```
%sql select landing__outcome, booster_version, launch_site from SPACEXTBL where landing__outcome = 'Failure (drone ship)' and EX
TRACT(YEAR FROM DATE) = '2015'
```

```
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

Out[14]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	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

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 [18]:
```
%sql select landing__outcome, count(landing__outcome) as total from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' g
roup by landing__outcome order by total desc
```

```
 * ibm_db_sa://rxd49332:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

Out[18]:

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

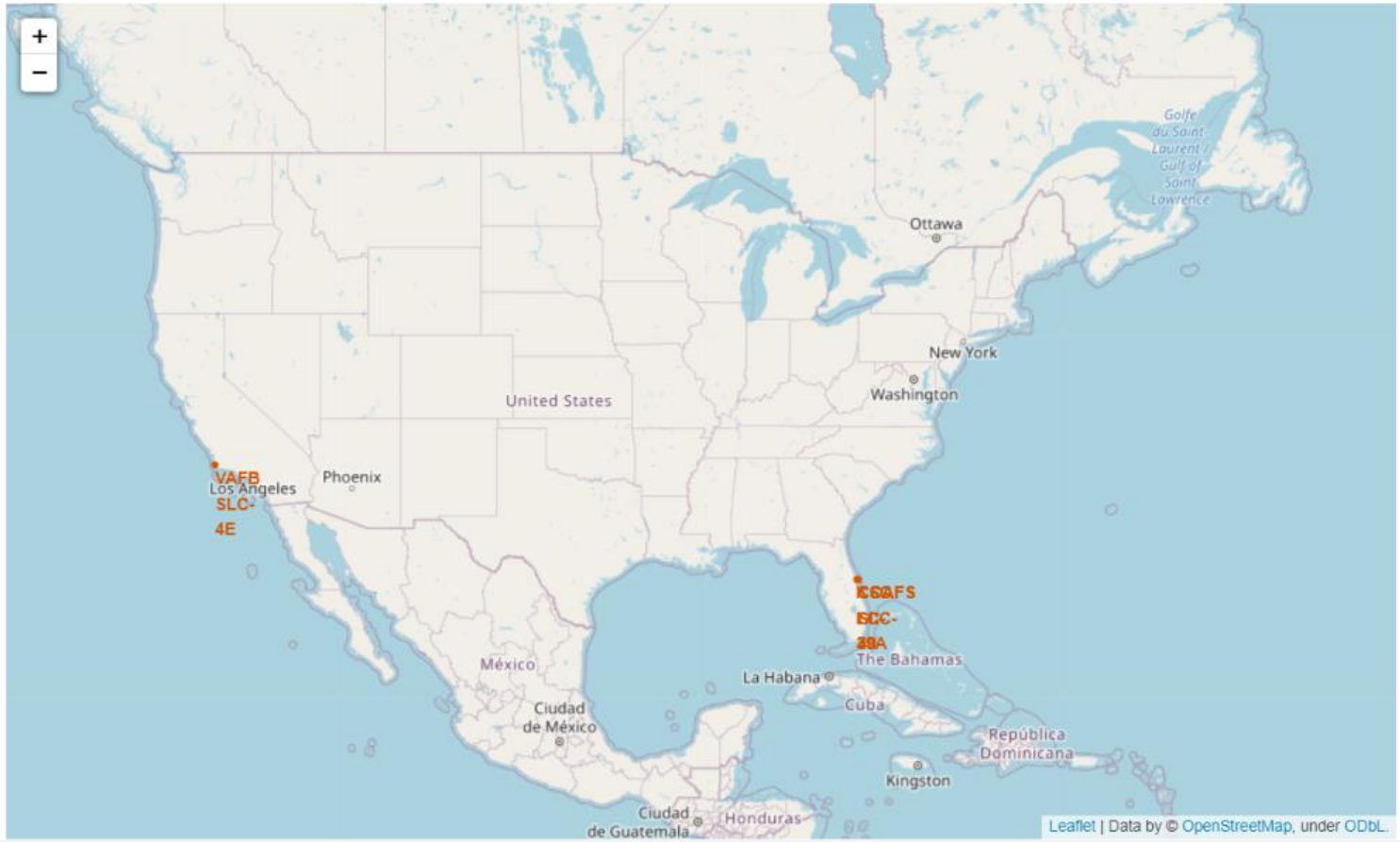
Section 3

Launch Sites Proximities Analysis

Map with Marked Launch Sites

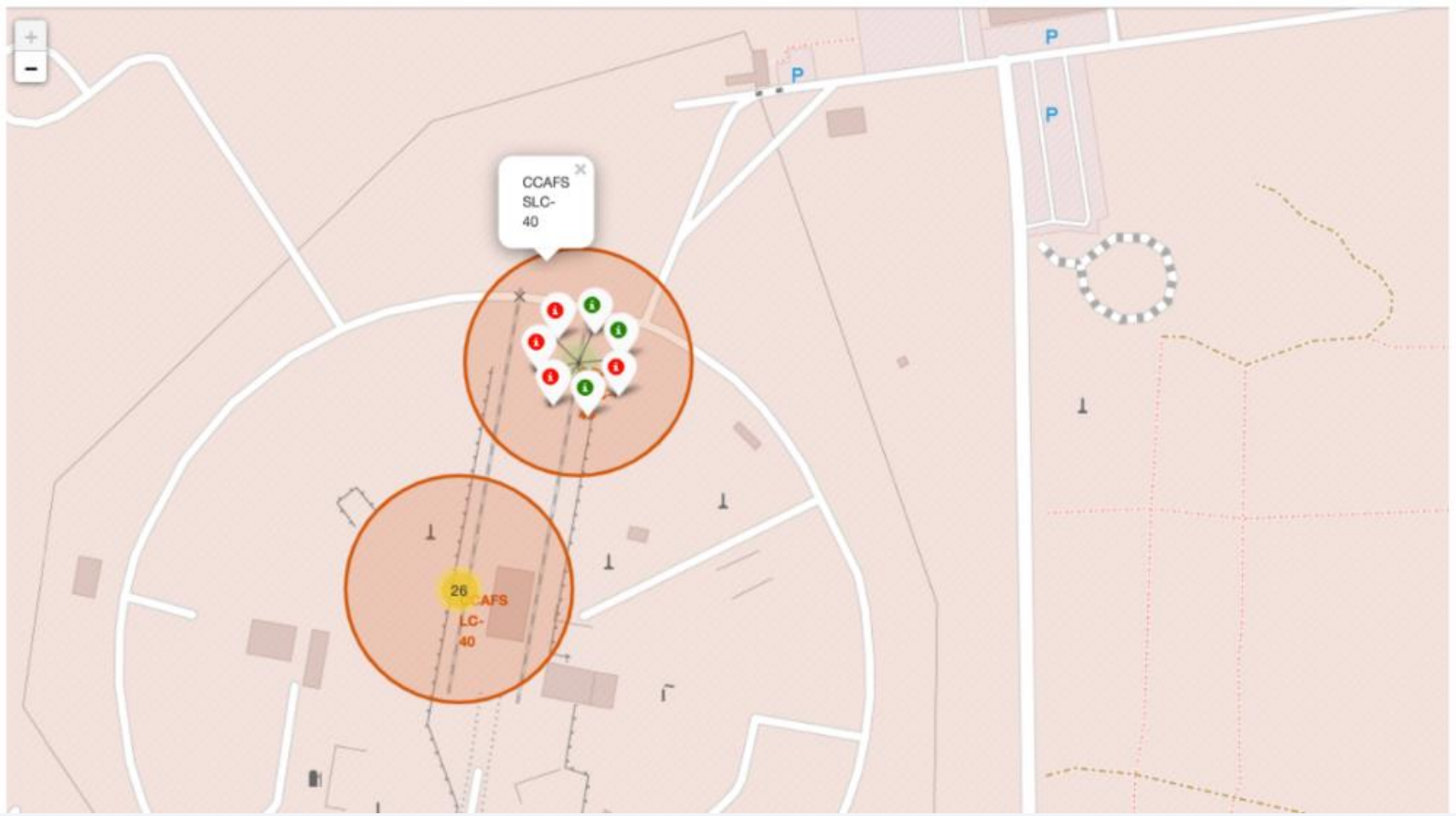
All launch sites are in proximity to the Equator lineand in very close proximity to the coast.

Out[8]:



Color-Labeled Markers

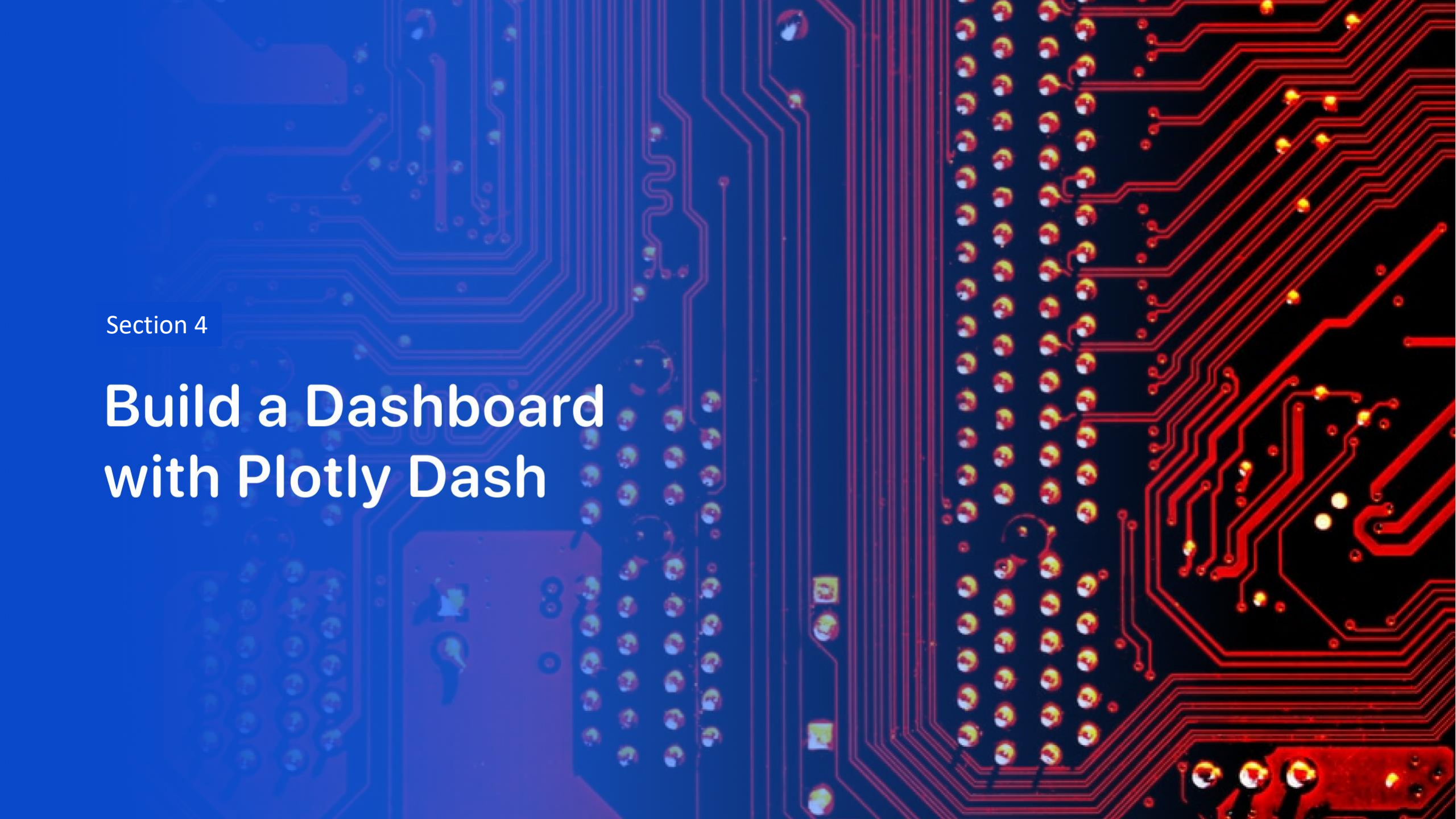
From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.



Distance Lines to the Proximities

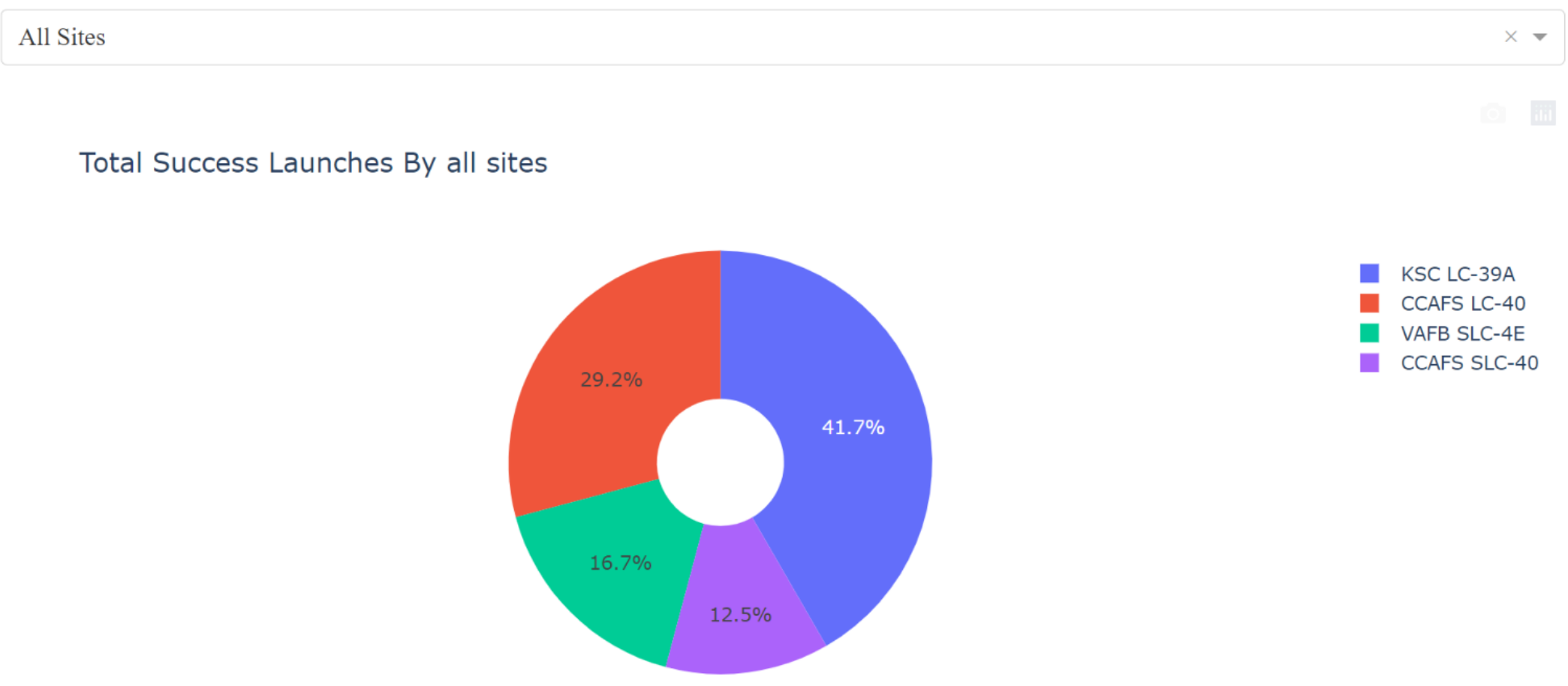
To answer below questions, distance lines to the proximities is done.

- •Are launch sites in close proximity to railways?
- •Are launch sites in close proximity to highways?
- •Are launch sites in close proximity to coastline?
- •Do launch sites keep certain distance away from cities?


Section 4
Build a Dashboard with Plotly Dash

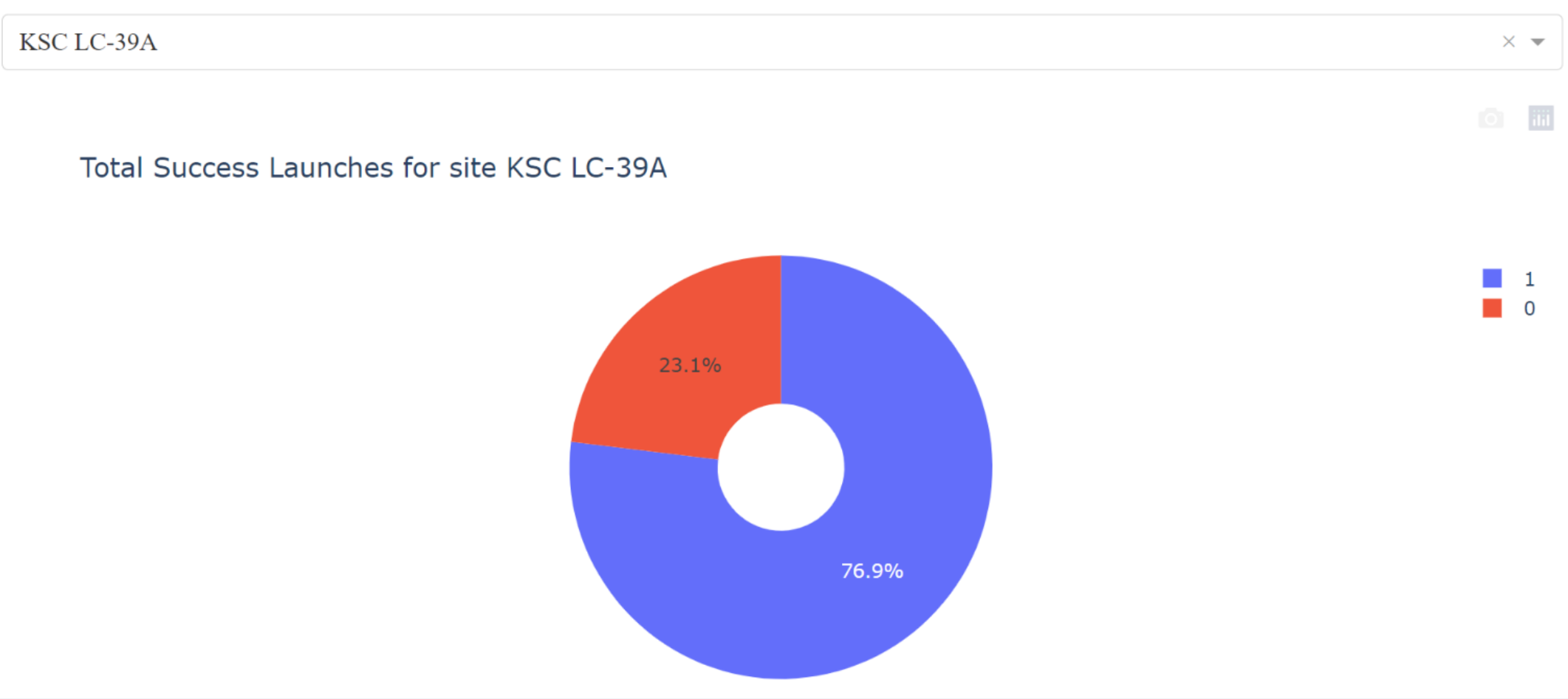
Total Success Launches by All Sites

76.% of KSC LC-39A launches were successful.

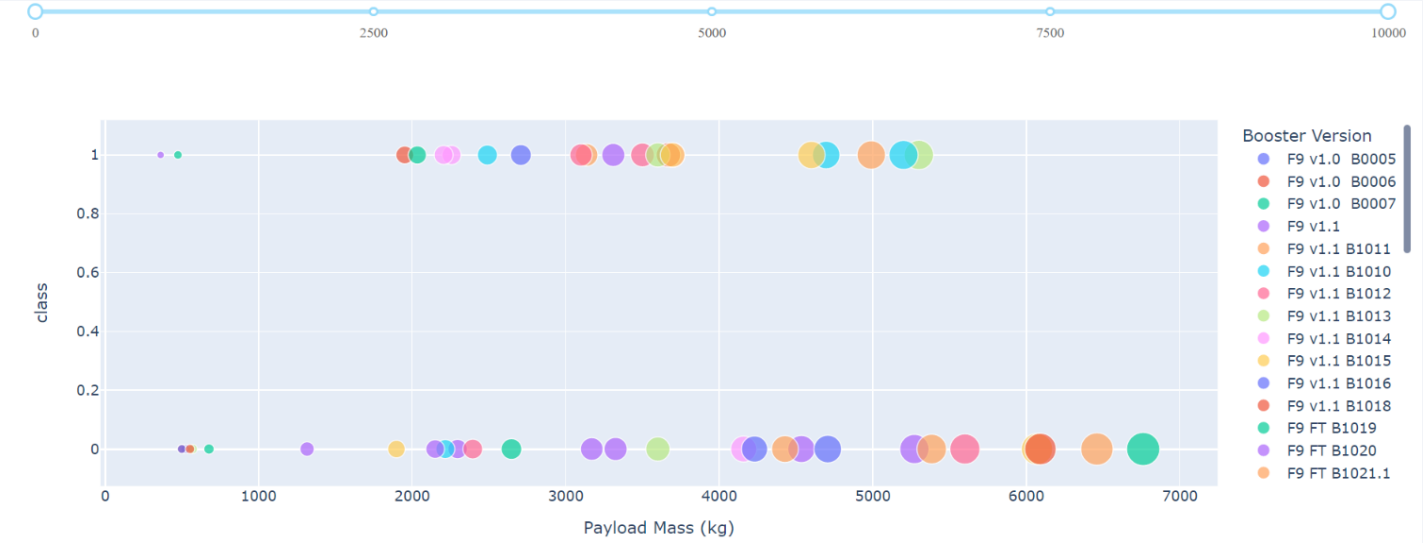


Total Success Launches for Site KSC LC-39A

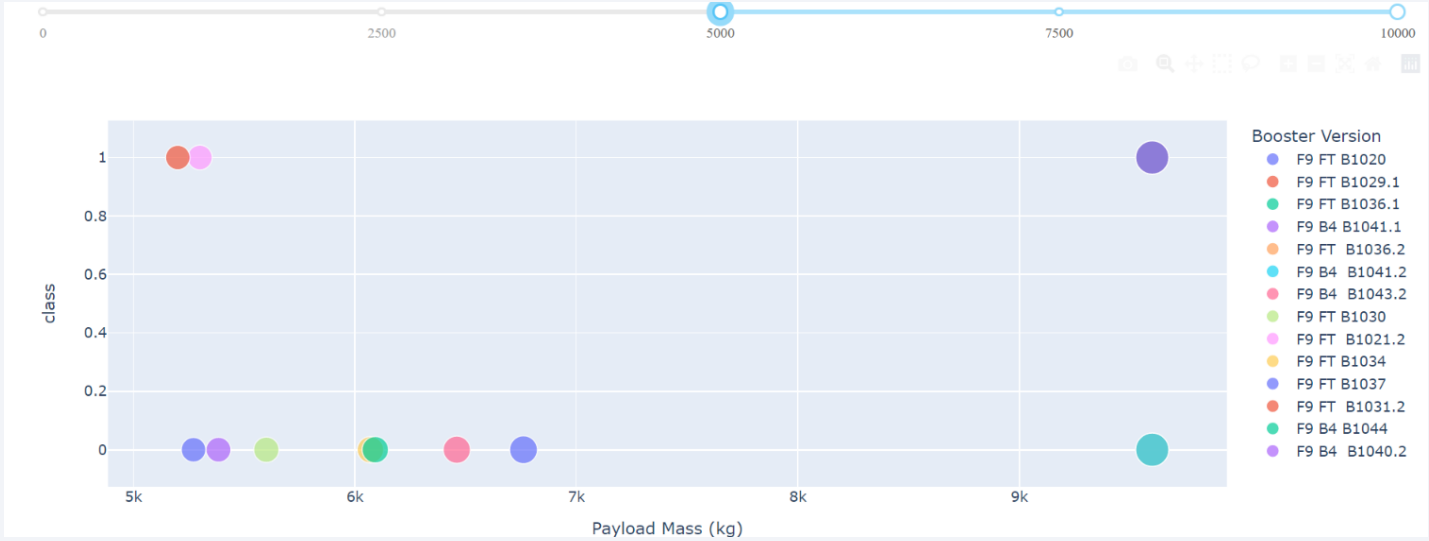
- Rep



Payload vs Launch Outcome for all sites, with different payload selected in the range slider



Success rates for low payloads are higher than heavy payloads.

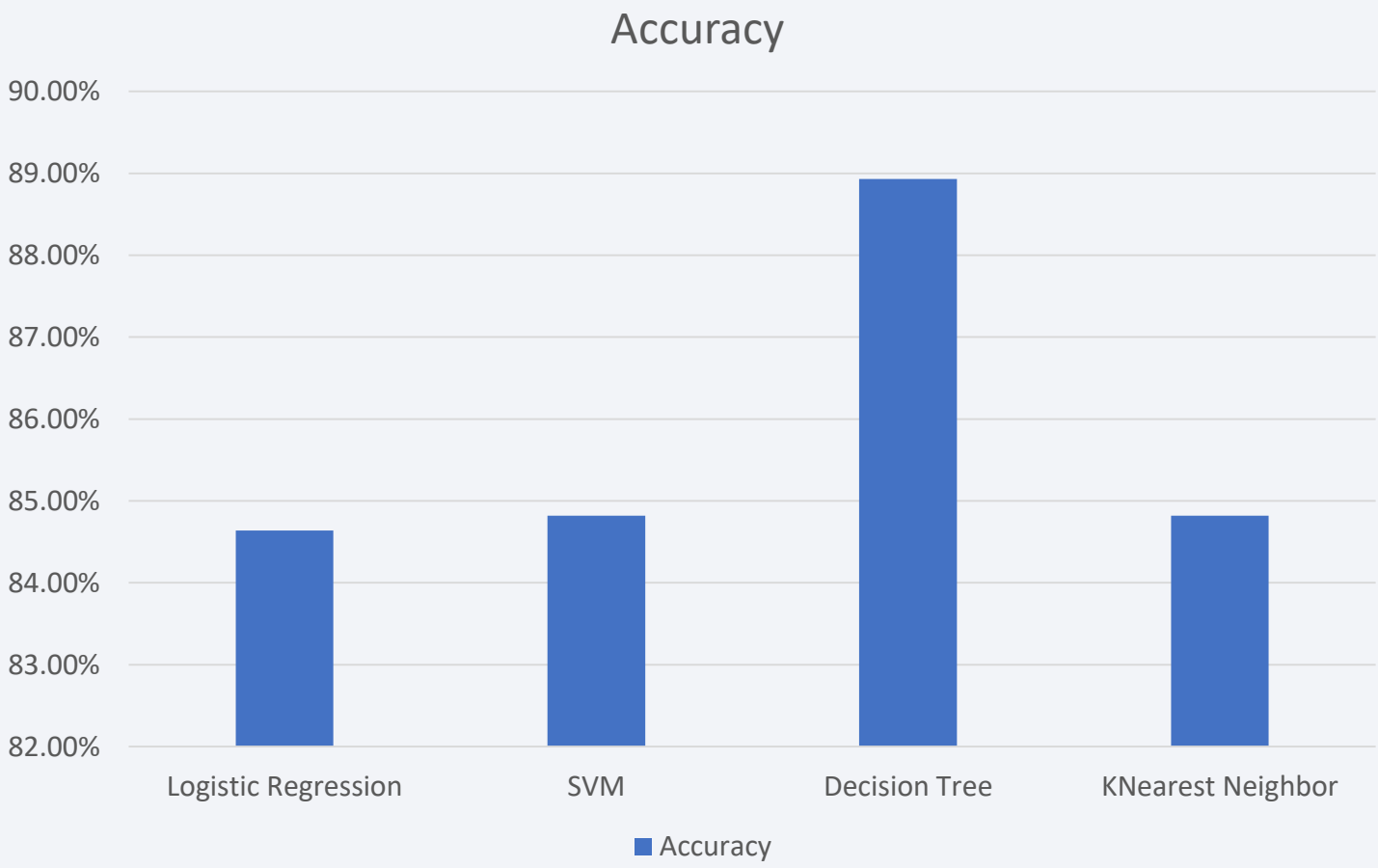


Section 5

Predictive Analysis (Classification)

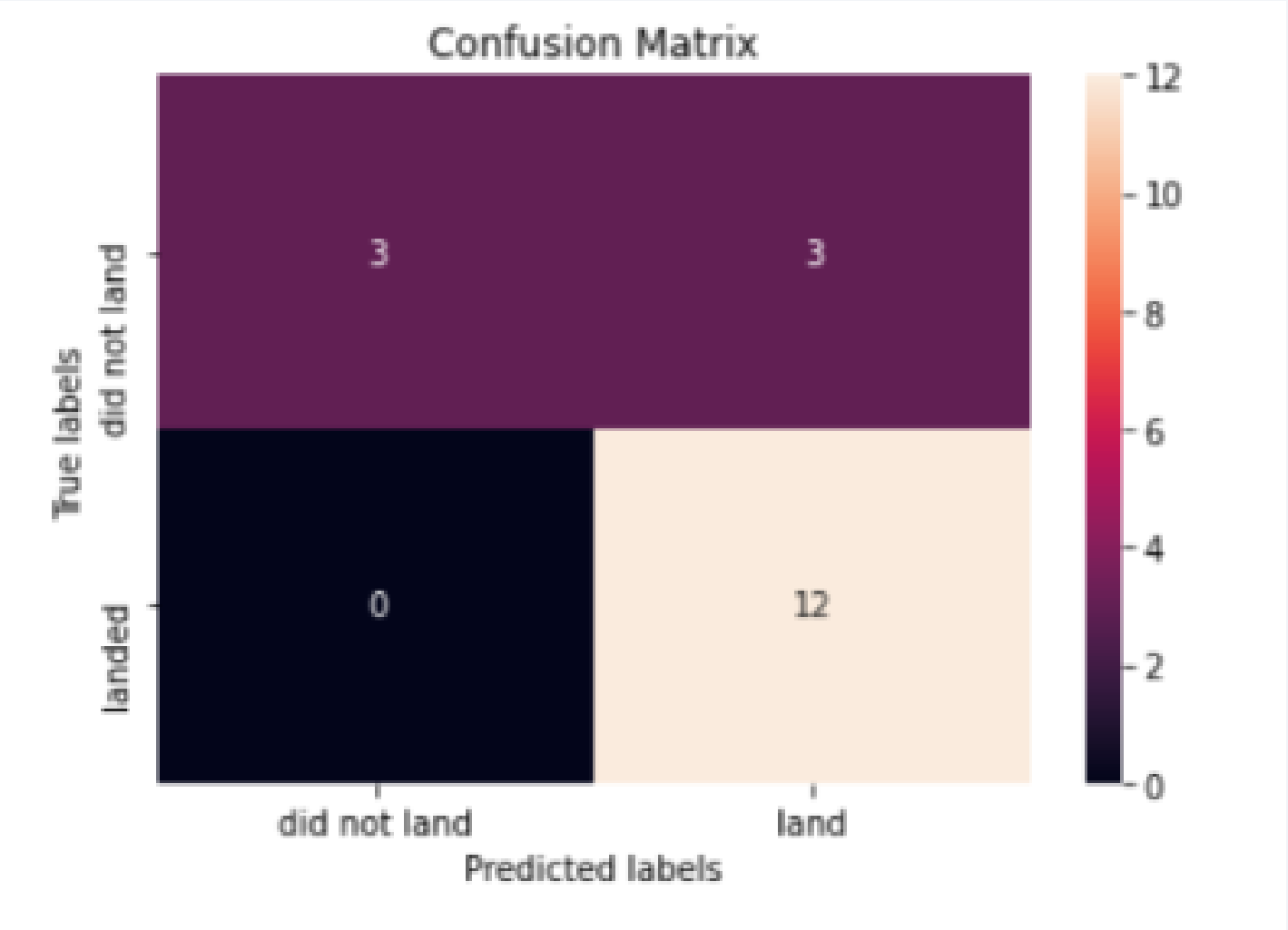
Classification Accuracy

The model which has the highest classification accuracy is **Decision Tree Classifier** with **88.93%**



Confusion Matrix

Decision Tree Classifier can distinguish between the different classes but the major problem is the false positives.



Conclusions

- Payload, Orbit and Flight number are factors to be considered for success rate
- Launch success rate started to increase in 2013 till 2020.
- ES-L1, GEO, HEO and SSO have high success rate.
- KSC LC-39A had the most successful launches in any sites.
- The Decision tree classifier is the best model that can distinguish between the different classes but the major problem is the false positives

Thank you!