Data Science Capstone project

<ZHAO XINYAN>

<2021-08-26>

Outline

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

Executive Summary

- To predict if the Falcon 9 first stage will land successfully, I consider to build some classification models such as SVM, Decision Trees, Logistic Regression and KNN and find the best hyperparameters from training set.

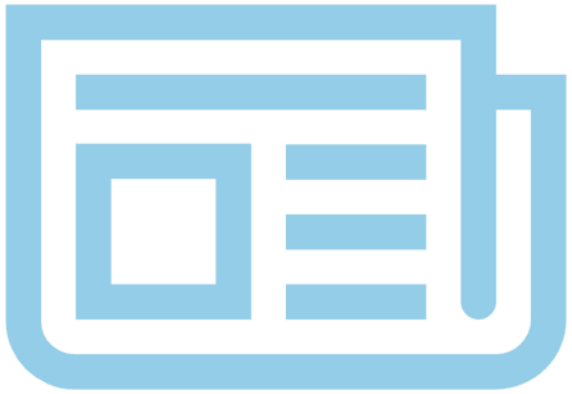
- According to the results of prediction, all models show the accuracy of 83.3% in the testing set, which means they can predict the landing in a high accuracy relatively.

Introduction

- SpaceX is a company that designs, manufactures and launches advanced rockets and spacecraft. It advertises Falcon 9 rocket launches, 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.

- The problems we need to answer include what factors influence the success rate of landing and the importance of every relevant factor.

Methodology



- Data collection methodology:
  - Request the data from SpaceX API
  - Parsing requested data into Dataframe
- Perform data wrangling
  - Filter the Dataframe to only include Falcon 9 launches
  - Replace the missing values with mean value
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Visualize the relationship between different factors using scatter point chart
  - Create dummy variables to categorical columns and cast the entire Dataframe to type float64
- Perform interactive visual analytics using Folium and Plotly Dash
  - Mark success/failed launches for each site on a map and calculate the distances between a launch site to its proximities
  - Use pie charts and scatter points chart to find relationship between different factors in an interactive way
- Perform predictive analysis using classification models
  - Split the data into training and testing set, use grid search to find the best hyperparameters for several classification models
  - Evaluate models with confusion matrix

Methodology

Data collection

- First, I consider using SpaceX API to collect needed data.

- Second, I consider using Web scraping with Beautiful Soup library to collect needed data.

Data collection – SpaceX API

GitHub URL:
https://github.com/SupGillllll/IBM-DATA-SCIENCE/blob/b1bbece4ae1142ef95212ec1824e215a20fa09cf/Applied%20Data%20Science%20Capstone/spacex-data-collection.ipynb

Import necessary Libraries and Define Auxiliary Functions

Request and parse the SpaceX launch data into a Pandas Dataframe using the GET request

Reuse the API to get information about the launches using the IDs and construct a new Dataframe

Filter the Dataframe to only include Falcon 9 launches, replace the missing values with mean value

Data collection – Web scraping

GitHub URL:
https://github.com/SupGillllll/IBM-DATA-SCIENCE/blob/b1bbece4ae1142ef95212ec1824e215a20fa09cf/Applied%20Data%20Science%20Capstone/Webscraping.ipynb

Import necessary Libraries and Define Auxiliary Functions

Request the Falcon9 Launch Wiki page from its URL and extract all column/variable names from the header

Parse the data from launch HTML tables and then transform into a pandas Dataframe

Data wrangling

GitHub URL:
https://github.com/SupGillllll/IBM-DATA-SCIENCE/blob/b1bbece4ae1142ef95212ec1824e215a20fa09cf/Applied%20Data%20Science%20Capstone/Spacex-Data-wrangling.ipynb

Import necessary Libraries and load the data

Calculate the number of launches on each site, the number and occurrence of each orbit and the number and occurrence of mission outcome per orbit type

Create a landing outcome label of binary code from Outcome column

EDA with data visualization

- Scatter point chart, bar chart and line chart have been plotted.
  - Scatter point chart with category can help to explore the relationship between different factors.
  - Bar chart can help to make comparisons with success rate among discrete categories.
  - Line chart can help to visualize the launch success trend with the change of time.

GitHub URL:

https://github.com/SupGillllll/IBM-DATA-SCIENCE/blob/b1bbece4ae1142ef95212ec1824e215a20fa09cf/Applied%20Data%20Science%20Capstone/EDA-dataviz.ipynb

EDA with SQL

- Performed SQL queries
  - SELECT statement with string patterns and key word ORDER BY and GROUP BY
  - SELECT statement with build-in functions like max() and avg()
  - SELECT statement with sub-queries and nested SELECT statement

GitHub URL:

https://github.com/SupGillllll/IBM-DATA-SCIENCE/blob/b1bbece4ae1142ef95212ec1824e215a20fa09cf/Applied%20Data%20Science%20Capstone/EDA-SQL.ipynb

Build an interactive map with Folium

- Add markers, circles, marker cluster and polylines to a folium map object
  - Markers can show every launch sites on the map
  - Circles can highlight every launch sites with a circle area
  - Markers can simplify the map containing many markers having the same launch site
  - Line can mark the distance between launch site with selected area

GitHub URL:

https://github.com/SupGillllll/IBM-DATA-SCIENCE/blob/b1bbece4ae1142ef95212ec1824e215a20fa09cf/Applied%20Data%20Science%20Capstone/launch_site_location.ipynb

Build a Dashboard with Plotly Dash

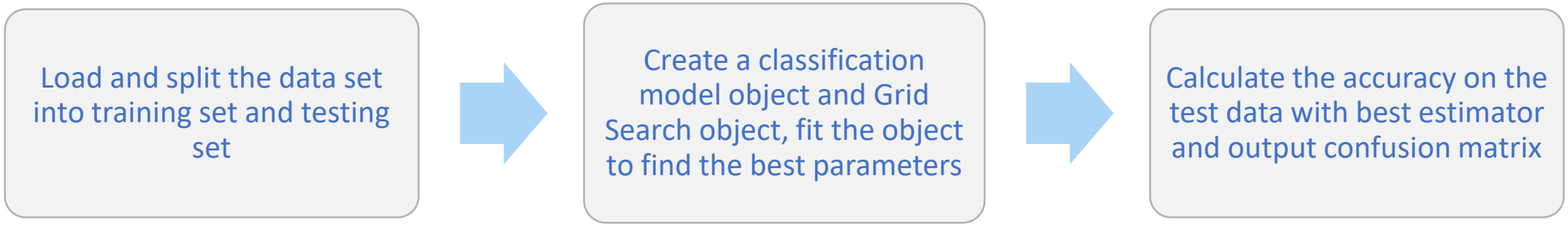
- Add pie chart, scatter point chart, dropdown list and slider to a dashboard with callback function.
  - Dropdown list can let us select different launch sites and interact with pie chart. Pie chart can visualize launch success counts with selected sites.
  - Slider can let us select the range of payload and interact with scatter point chart. Scatter point chart can visualize how payload may be correlated with mission outcomes for selected sites with different booster version.

GitHub URL:

https://github.com/SupGillllll/IBM-DATA-SCIENCE/blob/b1bbece4ae1142ef95212ec1824e215a20fa09cf/Applied%20Data%20Science%20Capstone/spacex_dashboard.py

Predictive analysis (Classification)

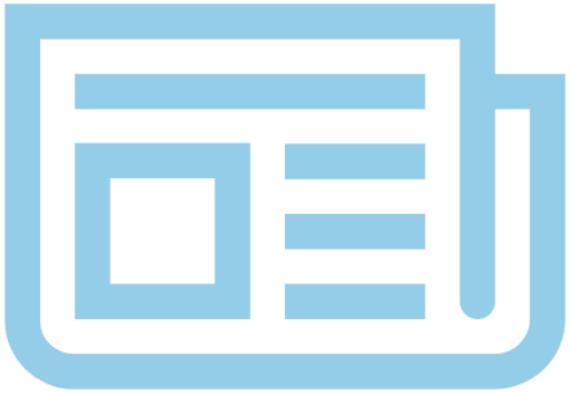
- I build the classification models including SVM, Decision Trees, Logistic Regression and KNN, I split the data set into training set and testing set and find the best hyperparameters using Grid Search. I evaluate models with accuracy score and confusion matrix.



GitHub URL:

https://github.com/SupGillllll/IBM-DATA-SCIENCE/blob/b1bbece4ae1142ef95212ec1824e215a20fa09cf/Applied%20Data%20Science%20Capstone/SpaceX_Machine_Learning_Prediction.ipynb

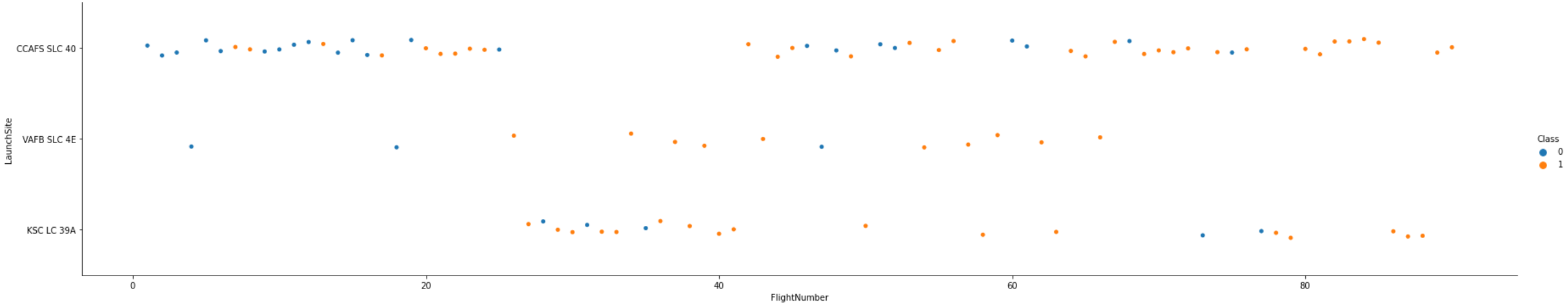
Results



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

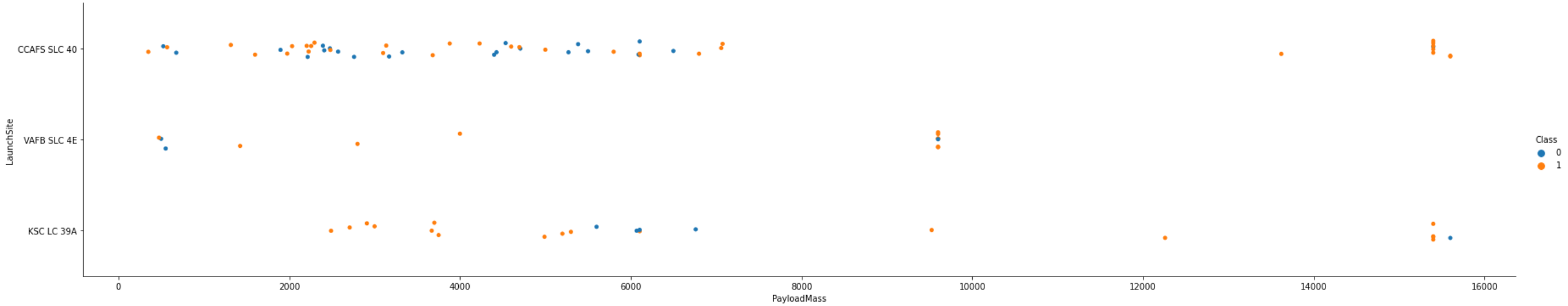
EDA with Visualization

Flight Number vs. Launch Site



- As the flight number increases, the success rate increases.
- KSC and VAFB have higher success rate than CCAFS.

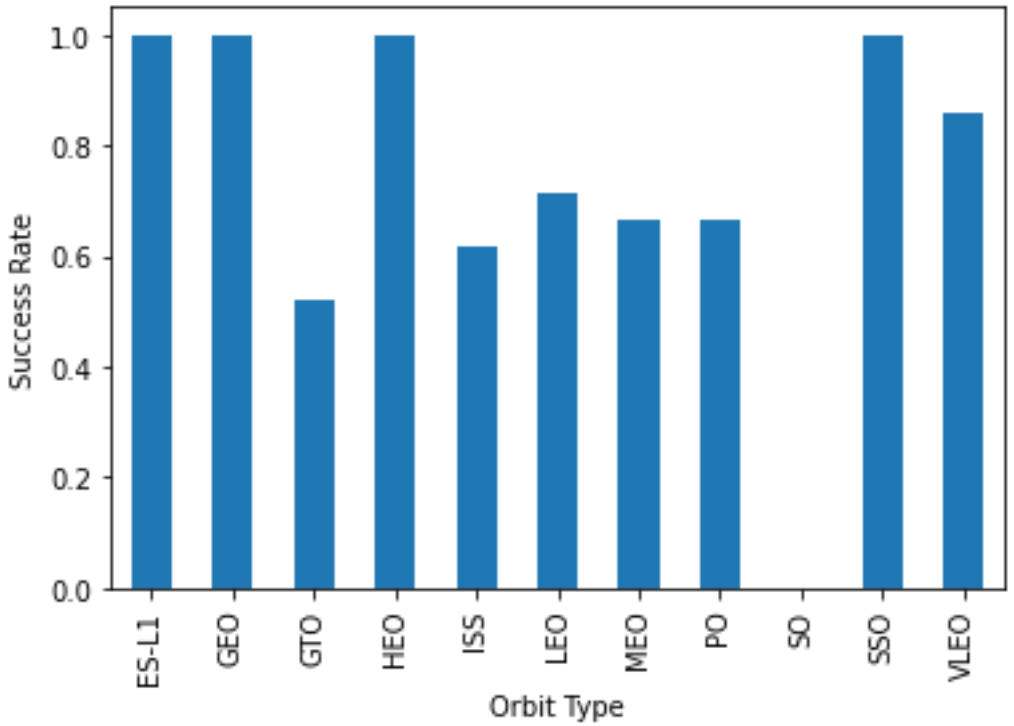
Payload vs. Launch Site



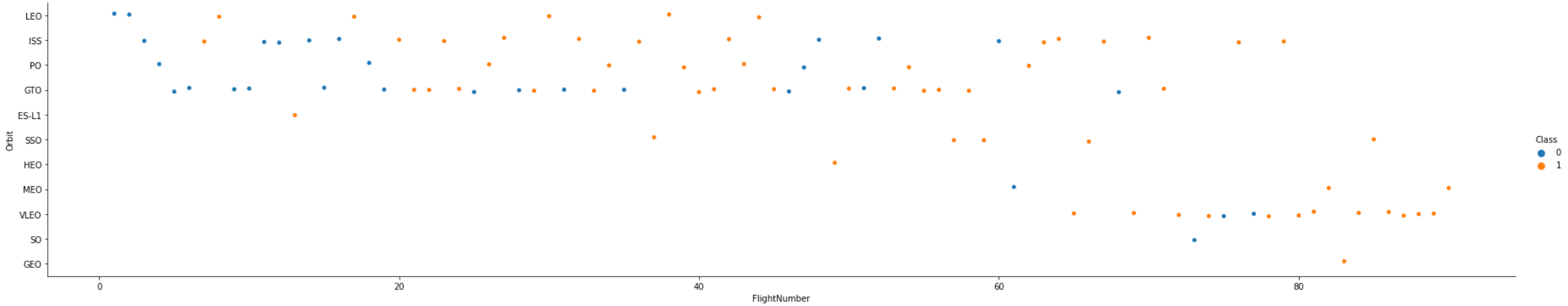
- The more massive the payload, the less likely the first stage will return.
- KSC and VAFB have higher success rate than CCAFS.

Success rate vs. Orbit type

- Orbit type ES-L1, GEO, HEO and SSO have higher success rate comparing with other orbits.

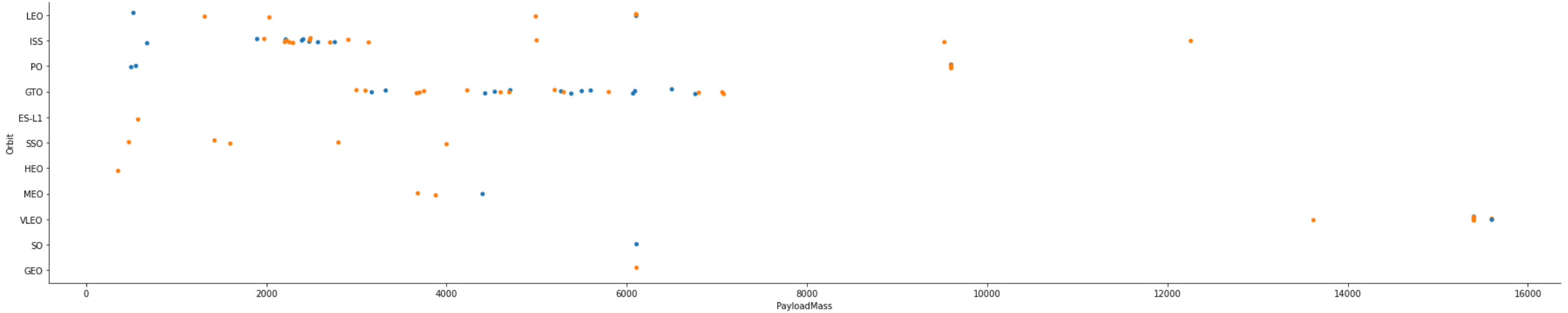


Flight Number vs. Orbit type



- As the flight number increases, the success rate increases.
- Orbit type ES-L1, GEO, HEO and SSO have higher success rate comparing with other orbits.

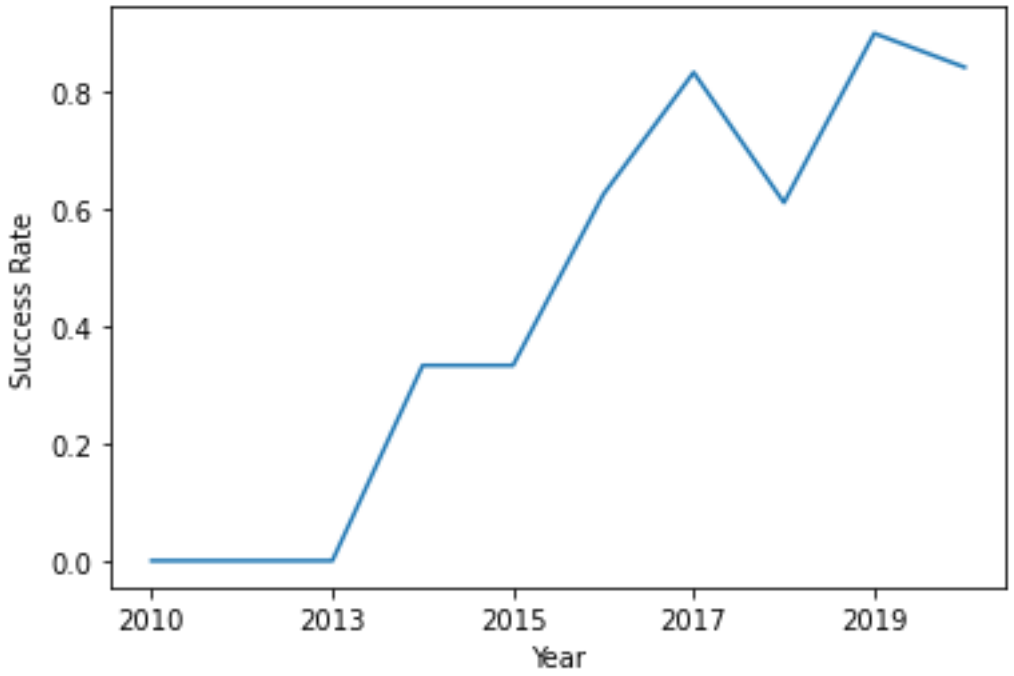
Payload vs. Orbit type



- The more massive the payload, the less likely the first stage will return.
- Orbit type ES-L1, GEO, HEO and SSO have higher success rate comparing with other orbits.

Launch success yearly trend

- Generally, the success rate increases with the change of year.



EDA with SQL

All launch site names

```
In [12]: %sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

```
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[12]:

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

- Use key word DISTINCT to find all launch site names

Launch site names begin with `CCA`

In [13]:
```
%sql select LAUNCH_SITE from spacextbl where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[13]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- Use string patterns to find launch site names begin with ‘CCA’

Total payload mass

```
In [14]: %sql select sum(PAYLOAD_MASS__KG_) from spacextbl where customer='NASA (CRS)'
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
Out[14]:
```

1
45596

- Use function sum() to calculate total payload mass

Average payload mass by F9 v1.1

```
In [15]: %sql select avg(PAYLOAD_MASS__KG_) from spacextbl where booster_version='F9 v1.1'
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
Out[15]:
```

1
2928

- Use function avg() to calculate average payload mass by F9 v1.1

First successful ground landing date

```
In [29]: %sql select min(date) from spacextbl where "Landing _Outcome"='Success (ground pad)'
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[29]:

1
2015-12-22

- Use function min() to find first successful ground landing date

Successful drone ship landing with payload between 4000 and 6000

In [30]:
```
%%sql
select booster_version, "Landing _Outcome", PAYLOAD_MASS__KG_ from spacextbl where "Landing _Outcome"='Success (drone ship)' and
PAYLOAD_MASS__KG_ between 4000 and 6000;
```

```
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[30]:

booster_version	Landing _Outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- Use key word BETWEEN AND to find Successful drone ship landing with payload between 4000 and 6000

Total number of successful and failure mission outcomes

In [37]:
```
%%sql
select "Landing _Outcome", count(*) as times from spacextbl group by "Landing _Outcome"
```

```
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[37]:

Landing _Outcome	times
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	22
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

- Use key word GROUP BY to find Total number of successful and failure mission outcomes

Boosters carried maximum payload

In [33]:
```
%%sql
select booster_version,PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from spacextbl)
```

```
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[33]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- Use a subquery to find the Boosters carried maximum payload

2015 launch records

In [36]:
```
%%sql
select "Landing _Outcome", booster_version, LAUNCH_SITE, date
from spacextbl where "Landing _Outcome"='Failure (drone ship)' and date between '2015-01-01' and '2015-12-31'
```

```
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[36]:

Landing _Outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- Use key word BETWEEN AND to find the 2015 launch records

Rank success count between 2010-06-04 and 2017-03-20

In [47]:
```
%%sql
select "Landing _Outcome", count(*) as times from spacextbl where date between '2010-06-04' and '2017-03-20' group by "Landing _Outcome" having count
(*)< 8 order by times desc
```

```
 * ibm_db_sa://ydx99284:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

Out[47]:

Landing _Outcome	times
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

- Use key word GROUP BY and ORDER BY to find the Rank success count between 2010-06-04 and 2017-03-20

Interactive map with Folium

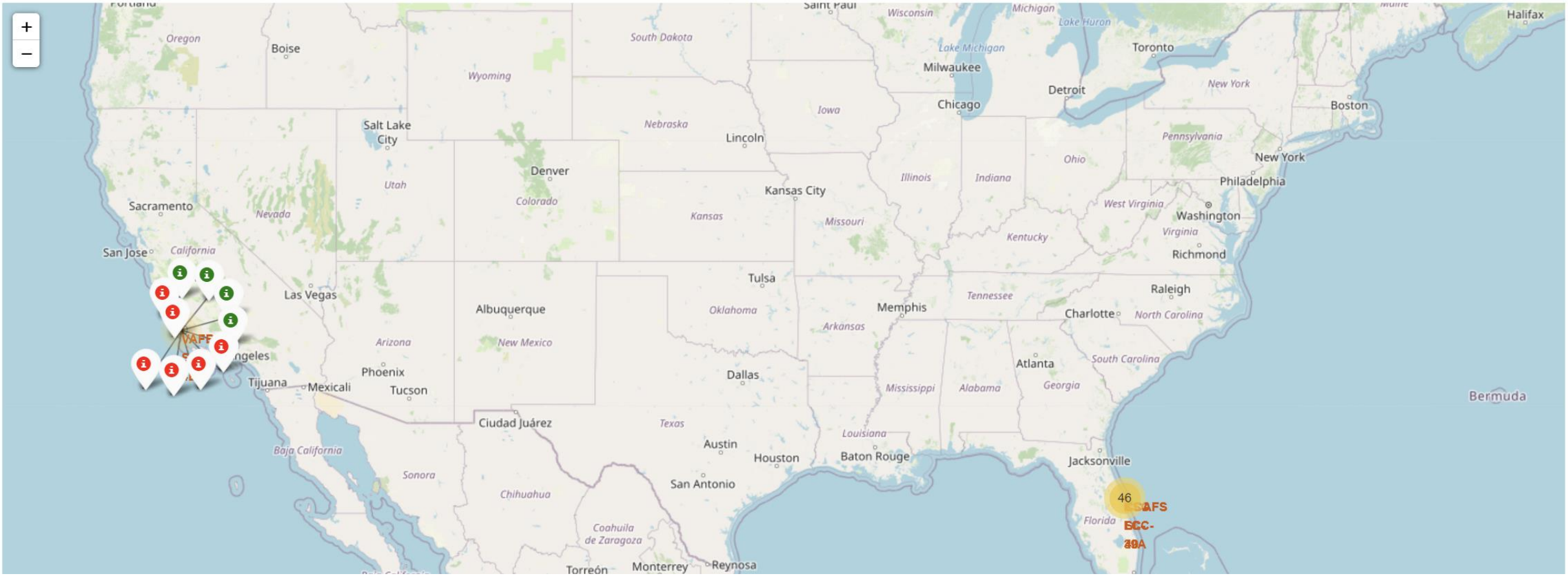
ALL launch sites on a global map

Out[8]:

- There are total four launch sites represented by circles and markers according to the records.

Color-labeled launch records on the map

Out[12]:



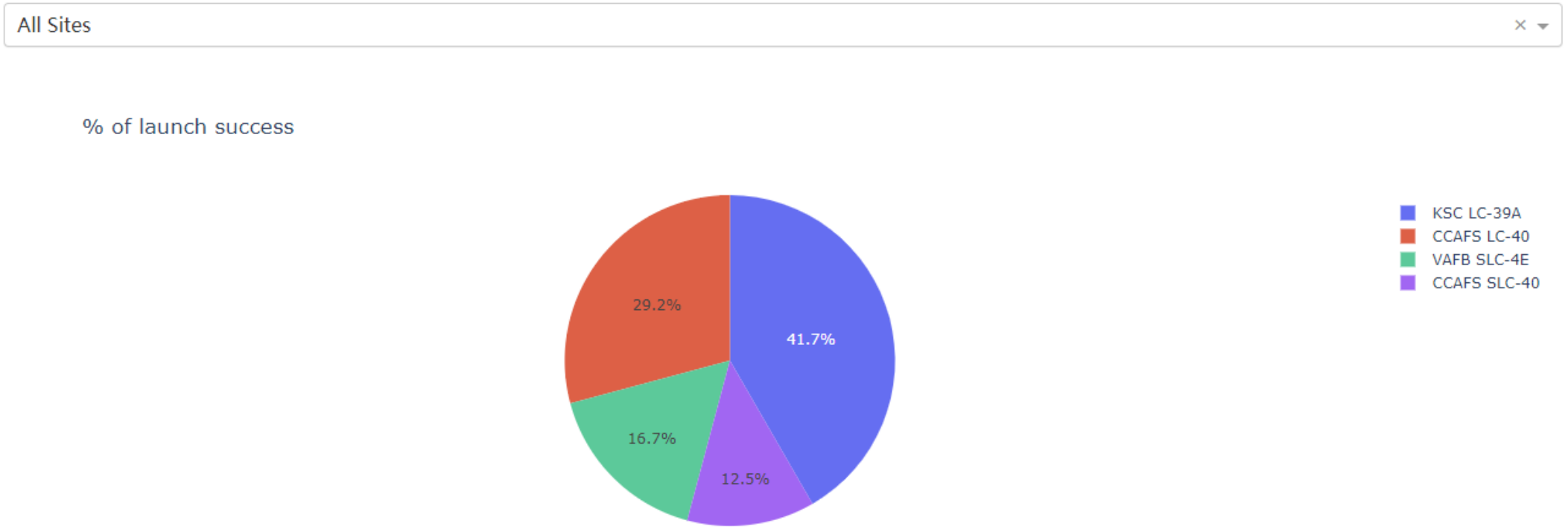
- The launch records in a site can be represented by the label with different colors and we can easily identify which launch sites have relatively high success rates.

A selected launch site to its proximities

- The launch site is close to the railway, approximately 1.27km.

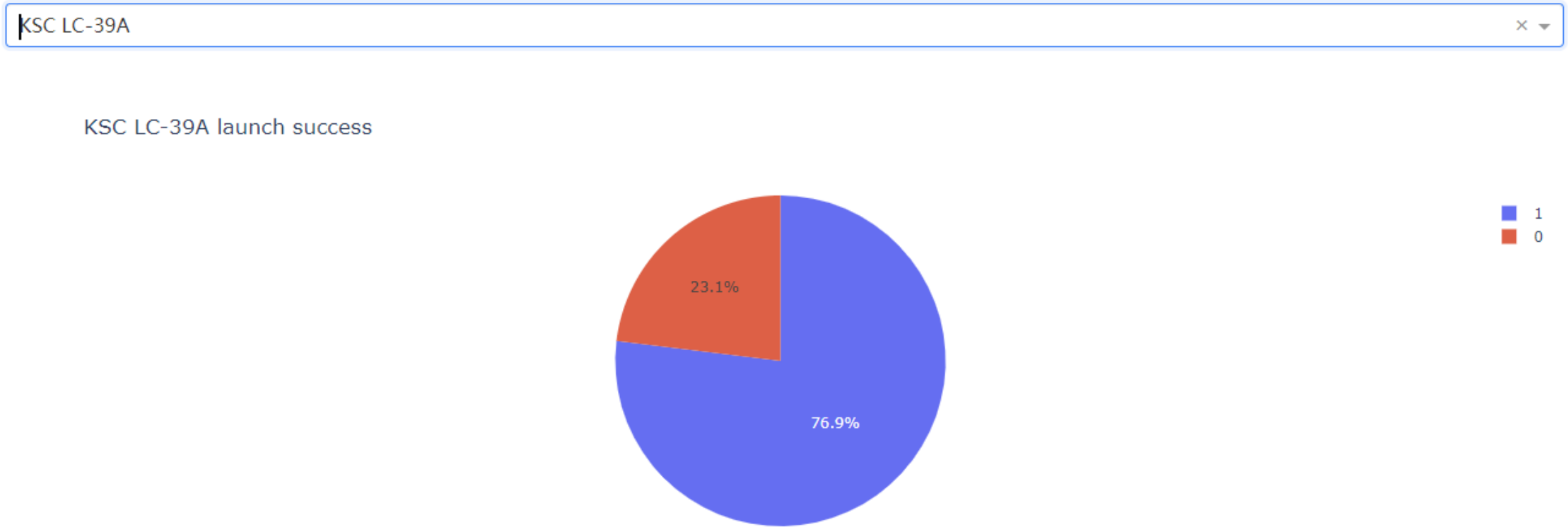
Build a Dashboard with Plotly Dash

Launch success count for all sites



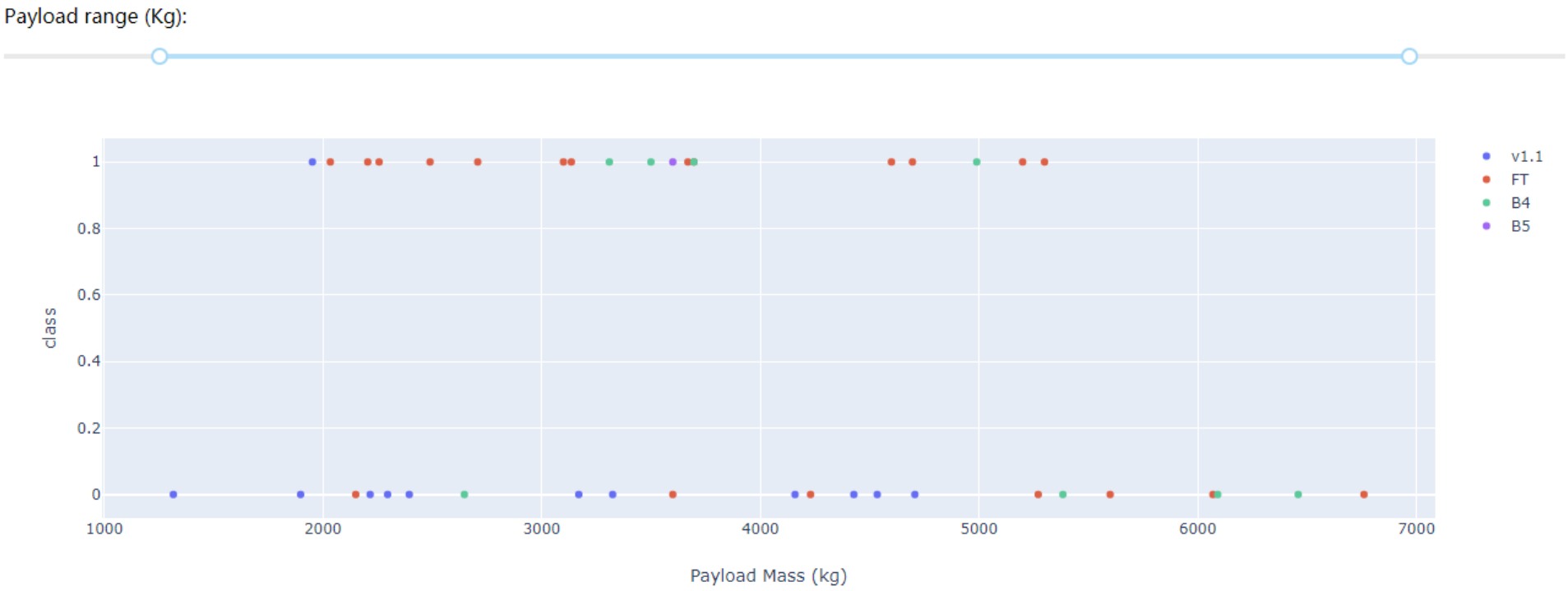
- The pie chart shows the launch success count for all sites, KSC site have the largest launch success count.

Pie chart for KSC



- The pie chart shows the launch success count for KSC site, which have the 76.9% success rate.

Payload vs. Launch Outcome scatter plot

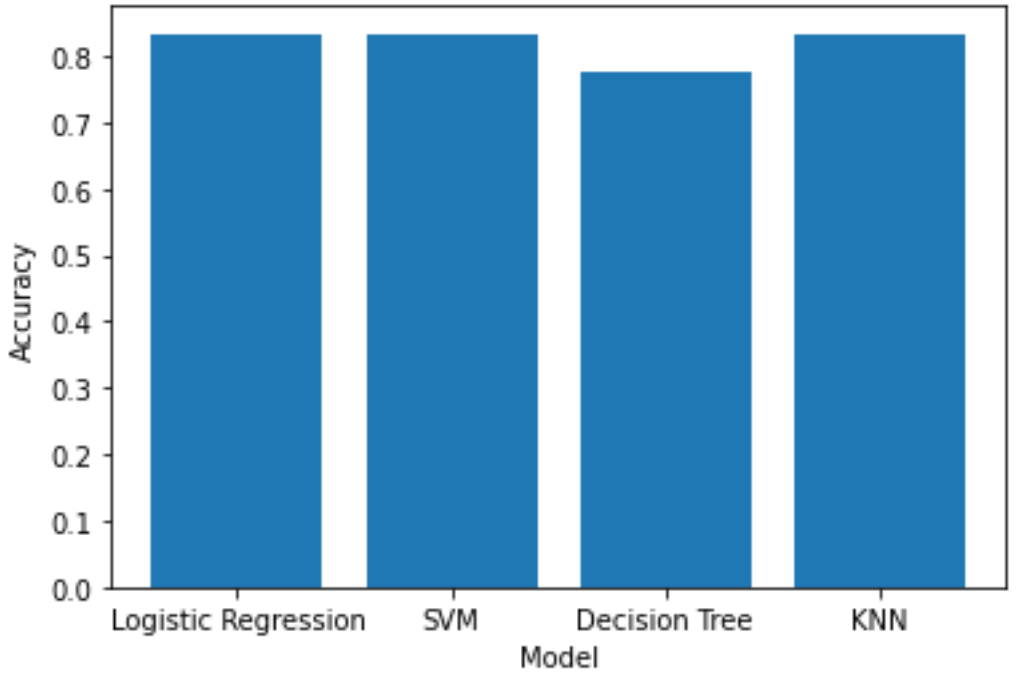


- The scatter plot in selected payload range shows that as the payload mass increases, the success rate will drop down and FT rocket has the highest success rate.

Predictive analysis (Classification)

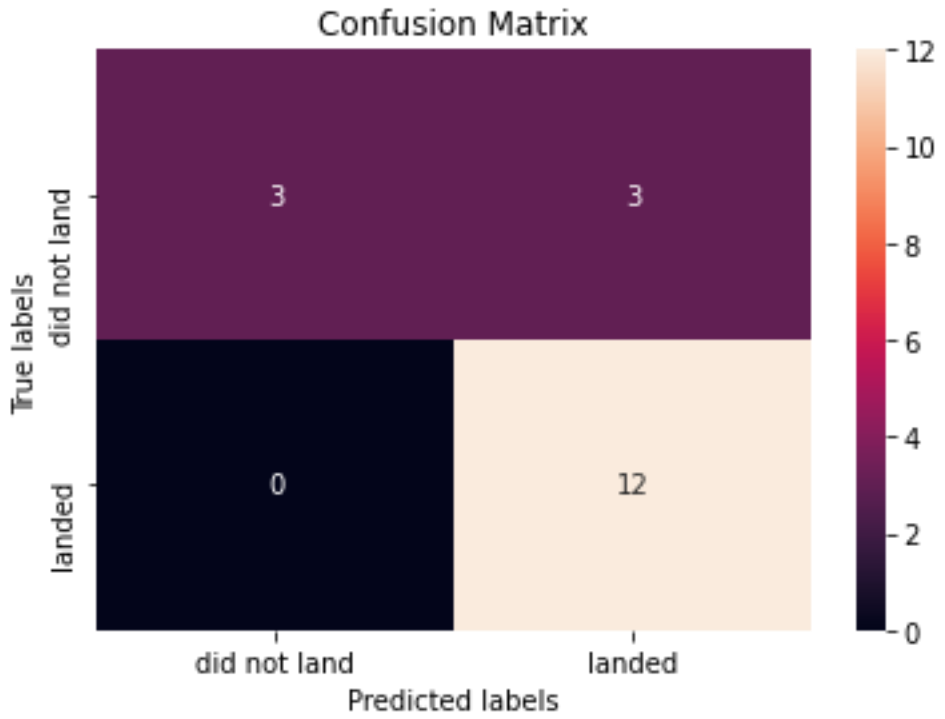
Classification Accuracy

Logistic Regression, SVM and KNN have the accuracy of 83.4%, however, Decision tree have the accuracy of 77.8%.



Confusion Matrix

Logistic Regression, SVM and KNN have the same confusion matrix, the major problem is the false positive detection.



CONCLUSION

- After a series of process, we can use the wrangled data and built machine learning model to predict if the first stage will land, which helps to solve practical problem.
- We can draw a conclusion that the land success rate is determined by many factors such as orbit and payload mass.

APPENDIX

GitHub Repository: https://github.com/SupGillllll/IBM-DATA-SCIENCE