IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- In this project, our goal is to predict if the Falcon 9 first stage will land successfully. We collected the necessary data by either making a get request to the SpaceX API or extracting Falcon 9 launch records HTML table from Wikipedia. We then cleaned the data and preliminarily made a exploratory data analysis via descriptive statistics and visualization including charts, maps and dashboard. Finally, we found the optimal model by testing logistic regression, support vector machines, decision tree classifier, K-nearest neighbors.
- After testing logistic regression, support vector machines, decision tree classifier, K-nearest neighbors, the accuracy performed by 4 algorithms was the same, so we can use the model trained by one of these 4 algorithms to predict if the Falcon 9 first stage will land successfully.

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.In this project, we will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

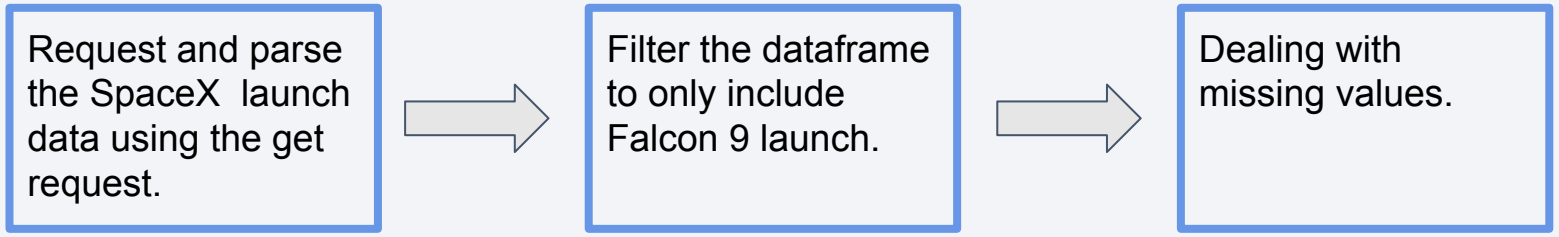
Executive Summary

- Data collection methodology:
  - We collected the necessary data by either making a get request to the SpaceX API or extracting Falcon 9 launch records HTML table from Wikipedia.
- Perform data wrangling
  - We mainly converted those outcomes into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We created a column for the class, standardized the data and split into training data and test data. Then, we performed grid search to find best hyperparameter for SVM, Classification Trees, Logistic Regression, KNN and further used the confusion matrix to determine the performance of each models.

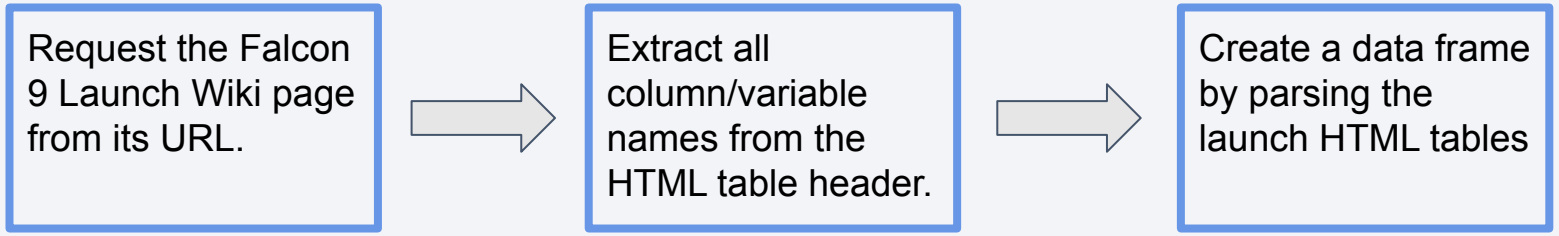
Data Collection

- The necessary data was collected by either making a get request to the SpaceX API or extracting Falcon 9 launch records HTML table from Wikipedia.

Method 1: API

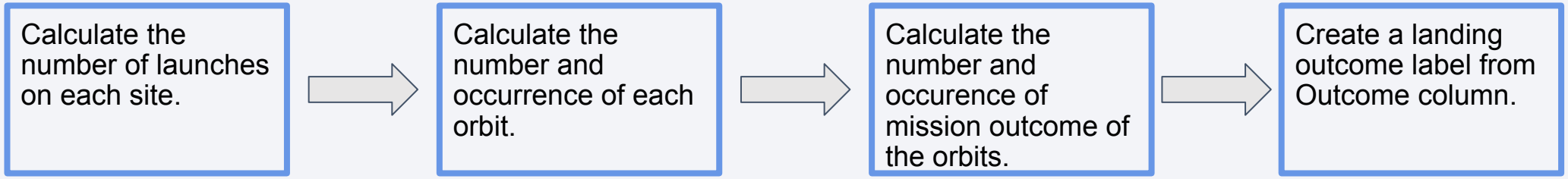


Method 2: Web Scraping



Data Wrangling

- The data were processed through below steps:
  1. Calculate the number of launches on each site.
  2. Calculate the number and occurrence of each orbit.
  3. Calculate the number and occurence of mission outcome of the orbits.
  4. Create a landing outcome label from Outcome column.



Source code:
https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore the data via visualization,
  1. The scatter plot was used to visualize the relationship between Flight Number and Launch Site, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type.
  2. The bar chart was plotted to visualize the relationship between success rate of each orbit type.
  3. The line chart was plotted to visualize launch success yearly trend.

Source code:
https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed
  1. Display the names of the unique launch sites in the space mission
  2. Display 5 records where launch sites begin with the string 'CCA'
  3. Display the total payload mass carried by boosters launched by NASA (CRS)
  4. Display average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome in ground pad was achieved.
  6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  7. List the total number of successful and failure mission outcomes
  8. List the names of the booster_versions which have carried the maximum payload mass by using a subquery.
  9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  10. 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.

Source code:
https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

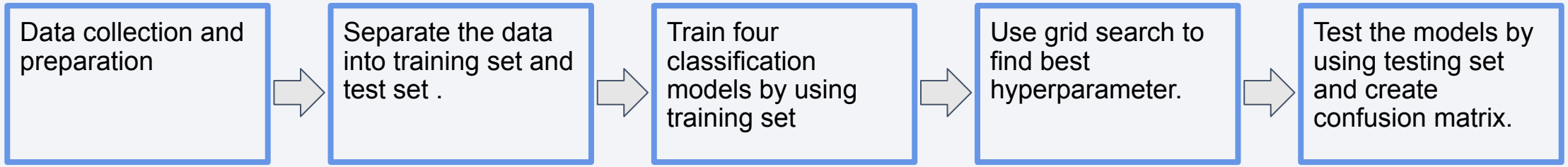
- The followings were created to perform launch sites locations analysis with Folium:
  1. Mark all launch sites on a map with adding folium.Circle and folim.Marker
  2. Mark the success/failed launches for each site on the map.
  3. Calculate the distances between a launch site to its proximities

Source code:

https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

- Four classification models: SVM, Classification Trees, Logistic Regression, KNN, were trained and grid search was performed on these four models to find best hyperparameter. The confusion matrix was created for each model to determine the performances.



Source code:

https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
  1. There are four launch sites in the mission.
  2. Total payload mass carried by boosters launched by NASA (CRS) is 45,596 kg.
  3. Average payload mass carried by booster version F9 v1.1 is 2,928.4 kg.
  4. The first successful landing outcome in ground pad was achieved on 12/22/2015.
  5. There is no relationship between flight number and launch site.
  6. For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass(greater than 10000).
  7. Four orbit types: ES-L1, GEO, HEO and SSO have 100% success rate.

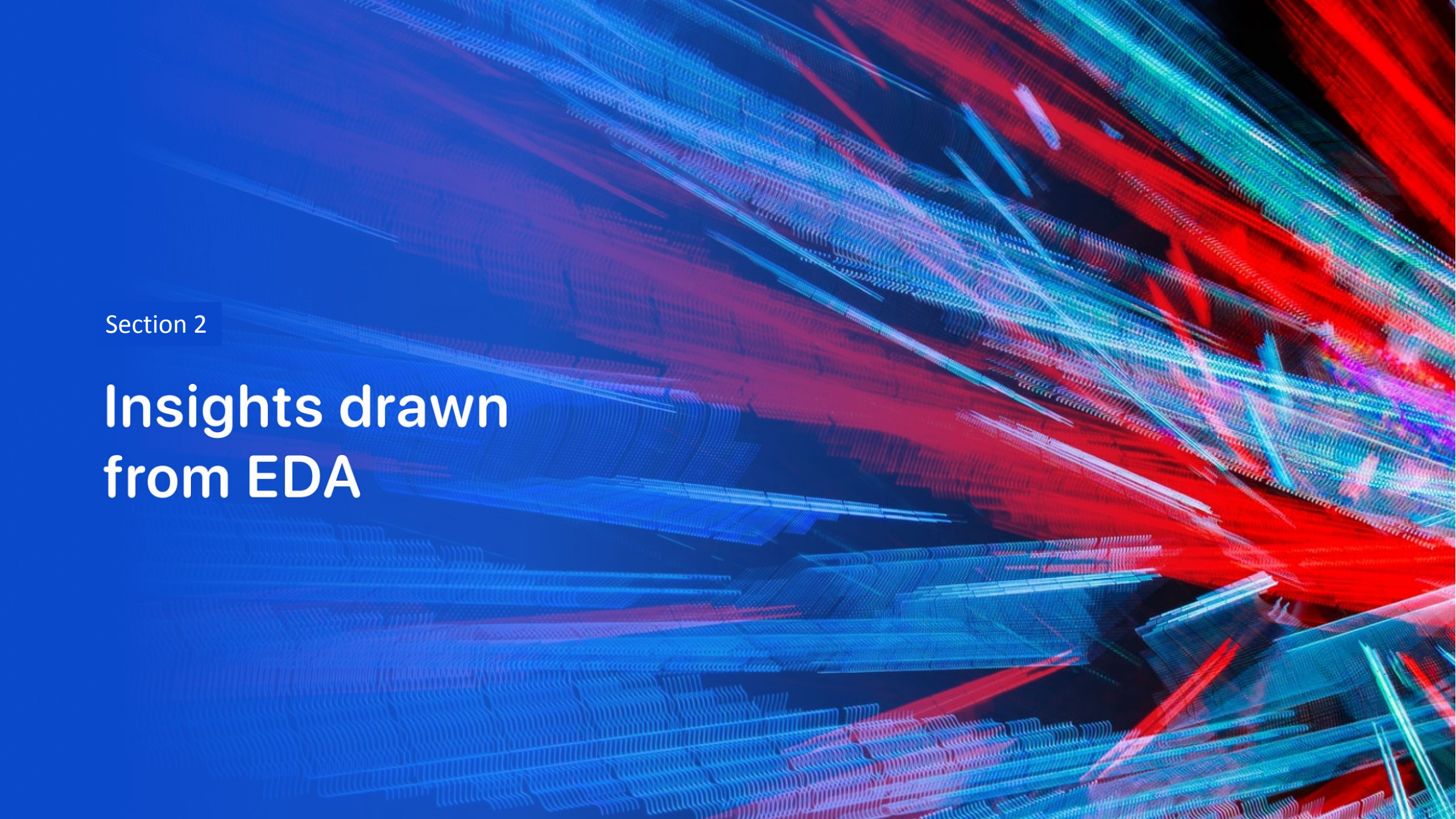
- Predictive analysis results
  1. The best hyperparameters for logistic regression model are: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}.
  2. The best hyperparameters for SVM model are: {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}.
  3. The best hyperparameters for decision tree classifier are: {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}.
  4. The best hyperparameters for KNN model are: {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
  5. Accuracy for Logistics Regression method, SVM model, decision tree classifier and KNN model is: 83.33%

Conclusions

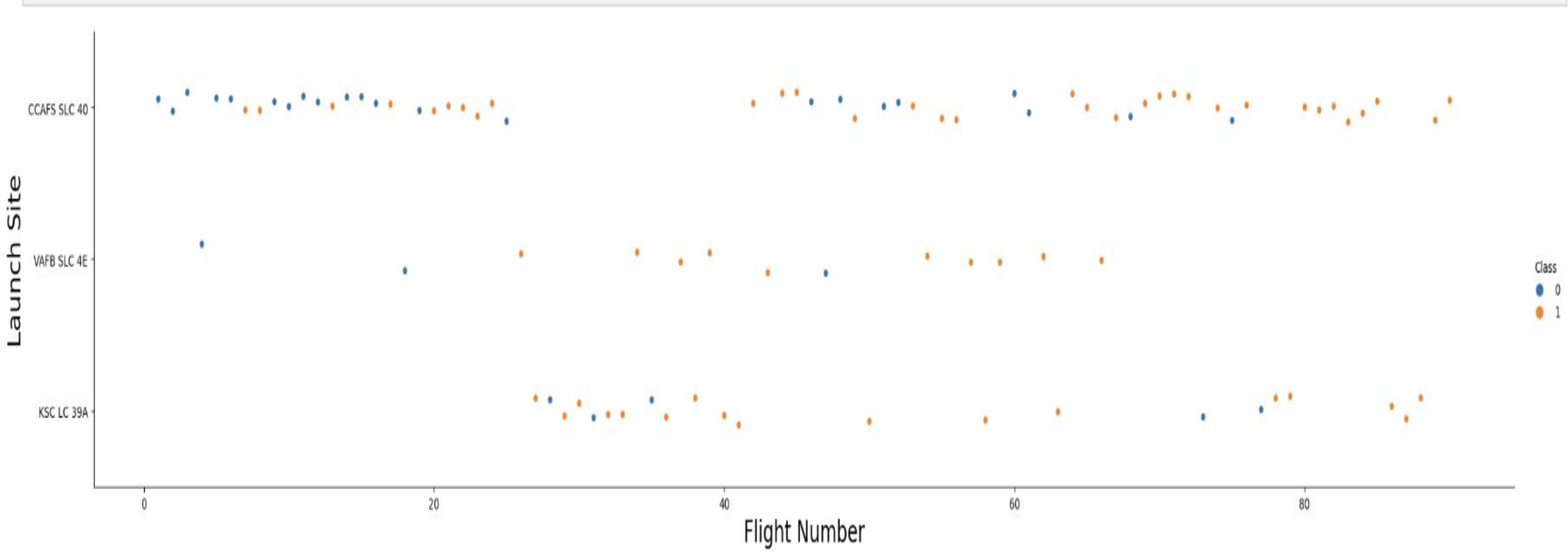
- After testing logistic regression, support vector machines, decision tree classifier, K-nearest neighbors, the accuracy performed by 4 algorithms was the same, so we can use the model trained by one of these 4 algorithms to predict if the Falcon 9 first stage will land successfully.

Section 2

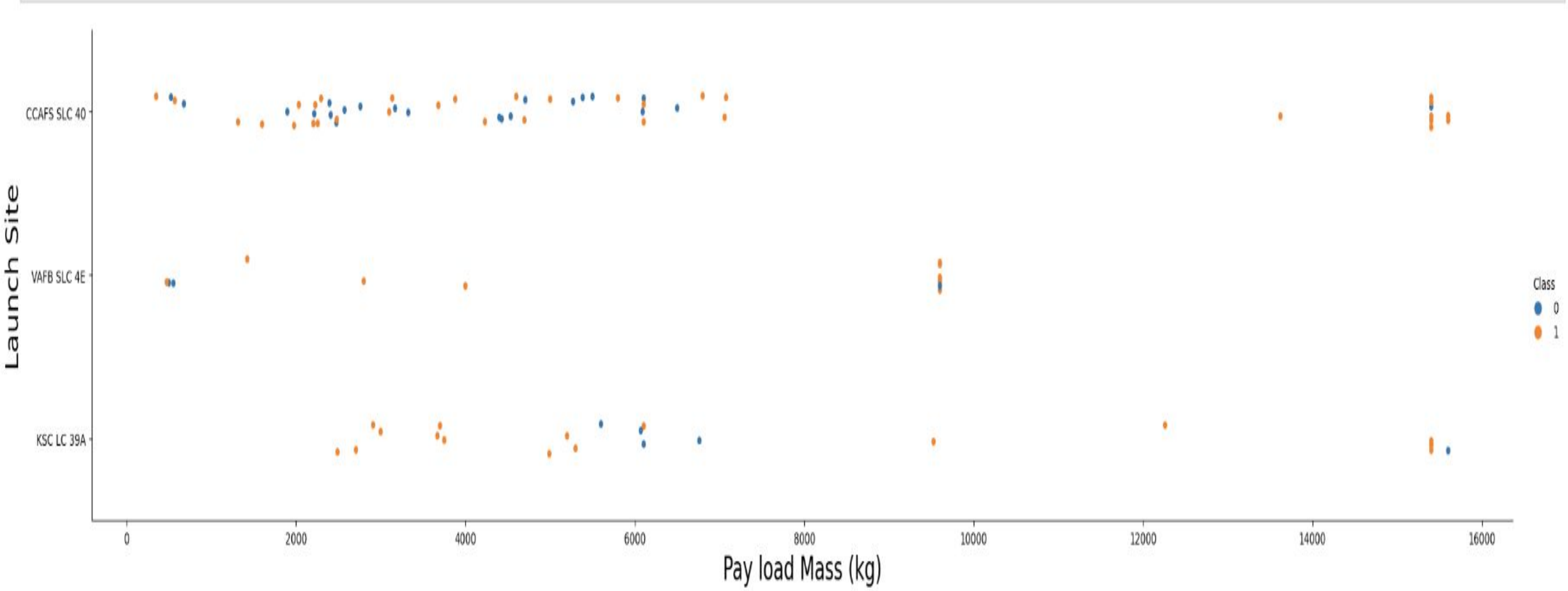
Insights drawn from EDA



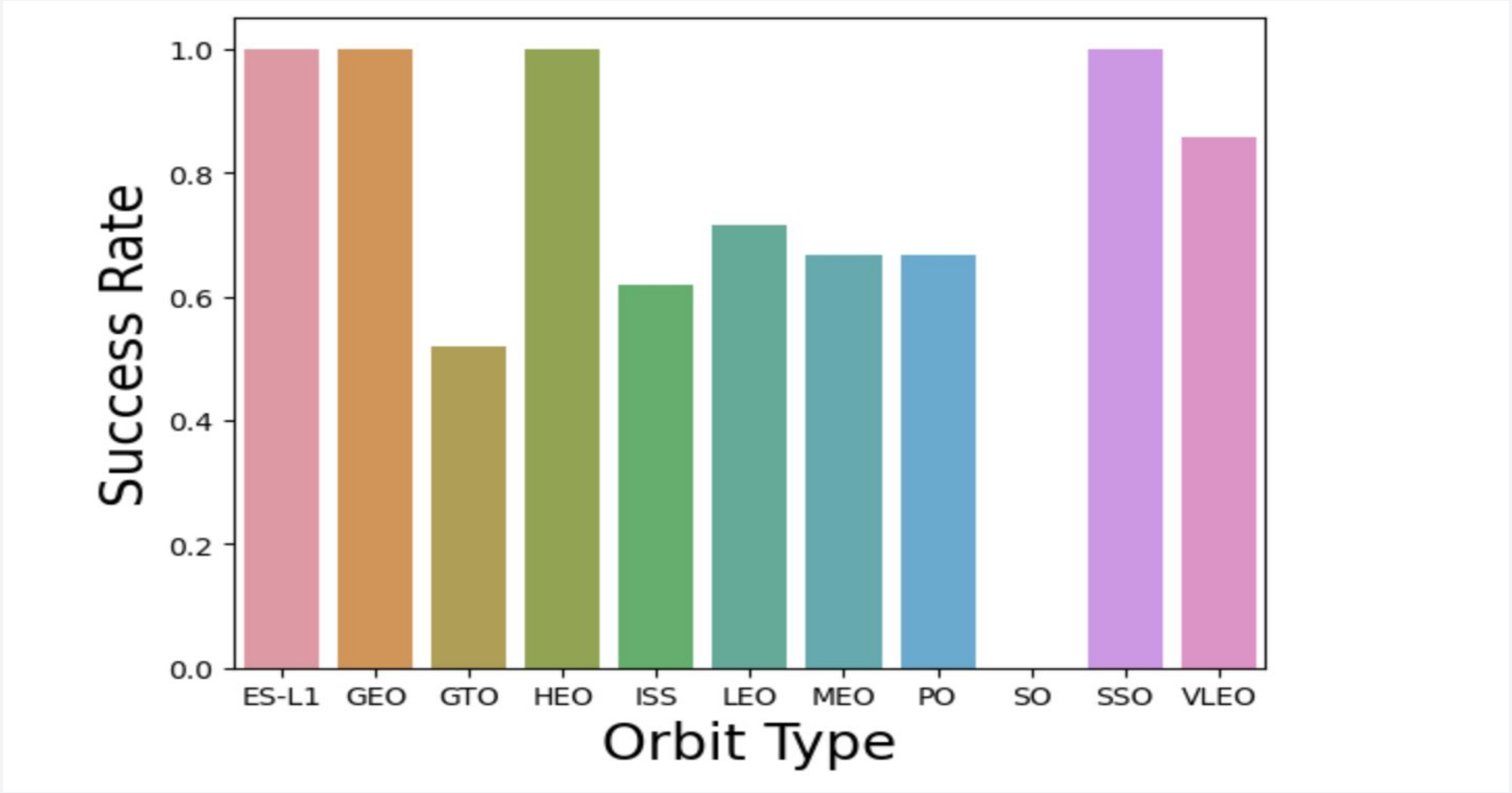
Flight Number vs. Launch Site



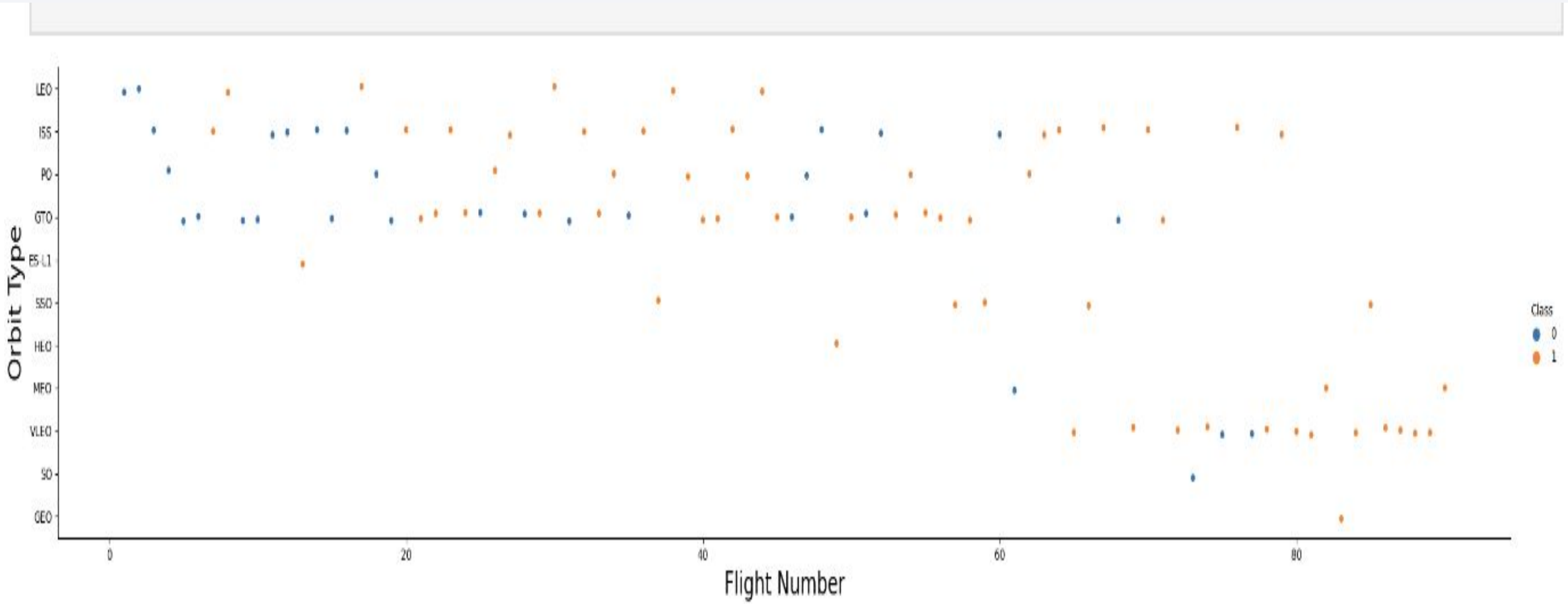
Payload vs. Launch Site



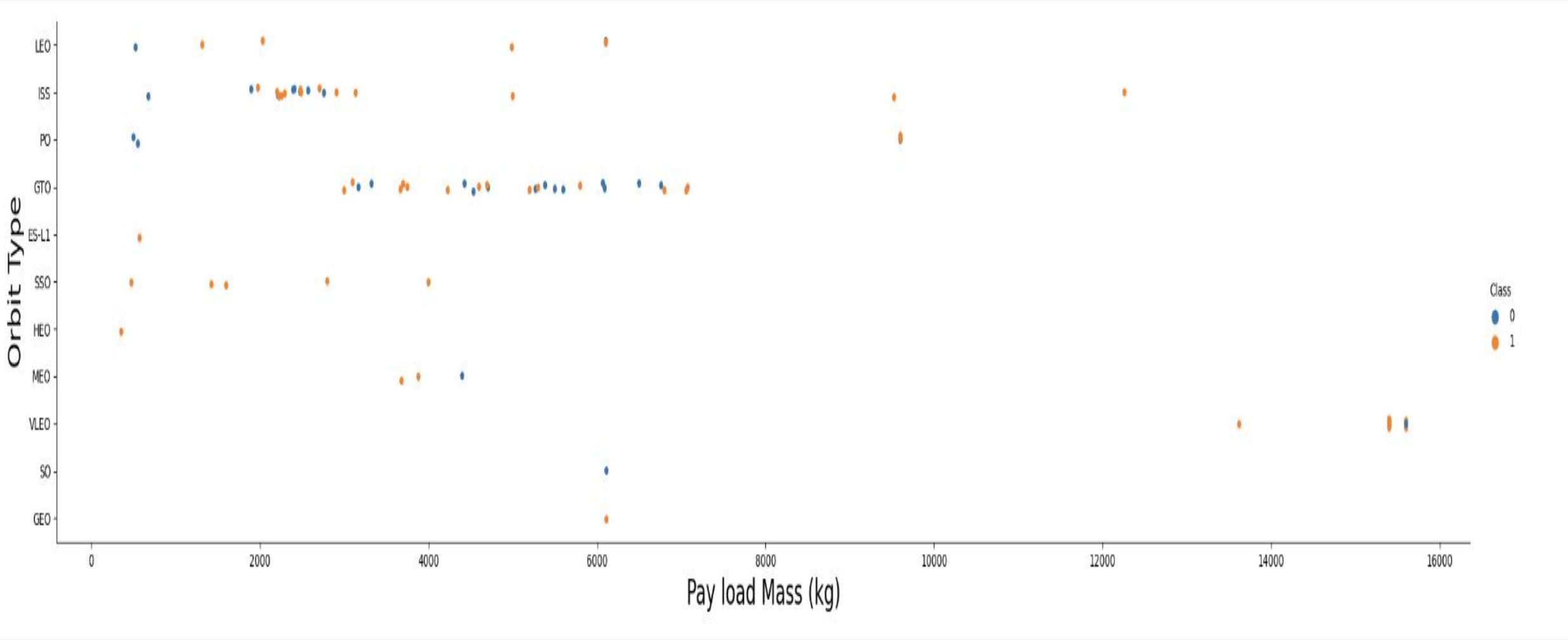
Success Rate vs. Orbit Type



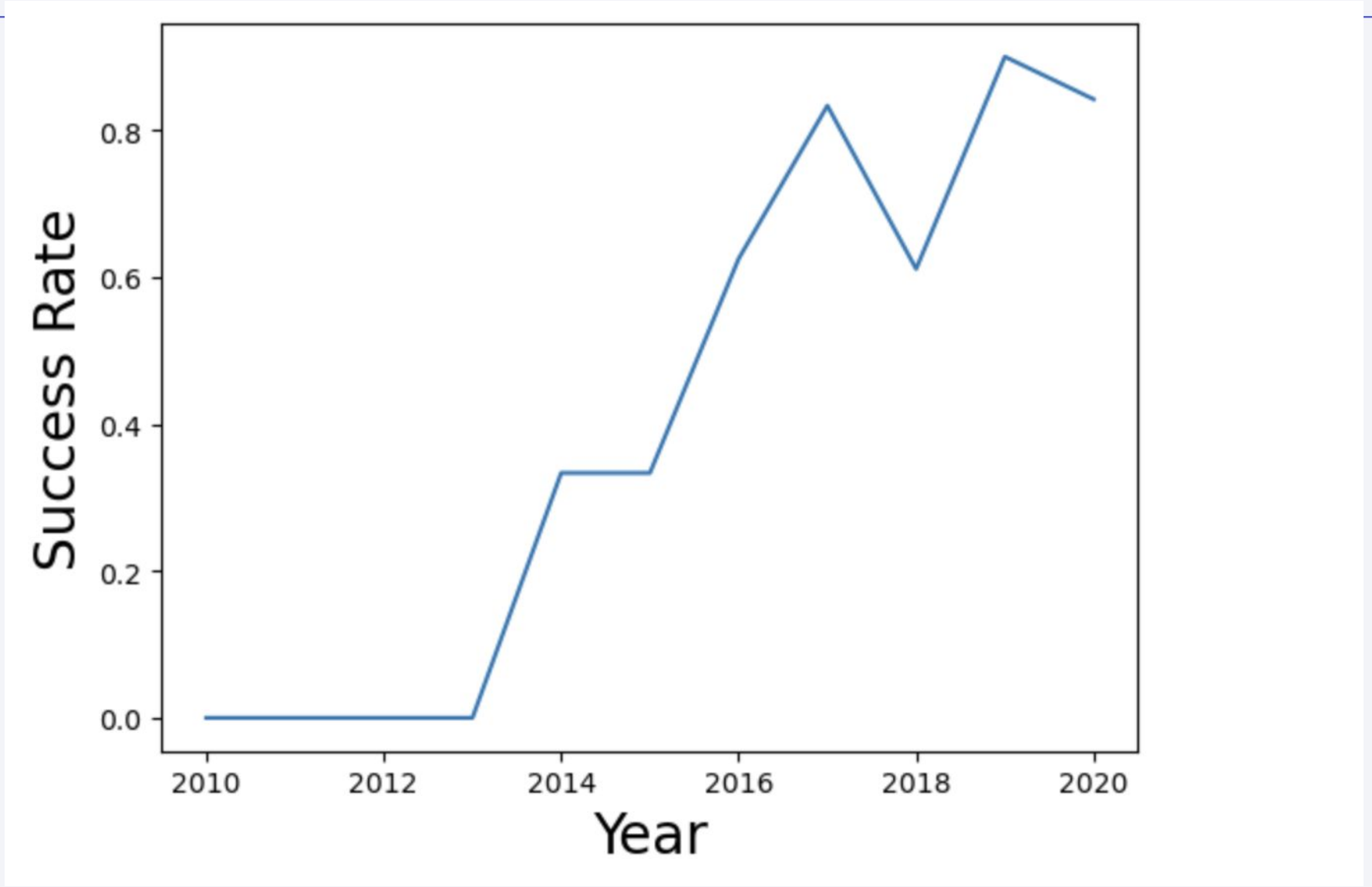
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- Four launch sites are as follows:
  1. CCAFS LC-40
  2. VAFB SLC-4E
  3. KSC LC-39A
  4. CCAFS SLC-40

```
In [8]: %sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[8]:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` are as follows:

```
In [9]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE '%CCA%'LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- The total payload carried by boosters from NASA is 45,596 kg

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

In [10]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

Out[10]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg

In [11]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

Out[11]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 12/22/2015

In [13]:
```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Out[13]:

MIN(Date)
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 are:
  1. F9 FT B1022
  2. F9 FT B2016
  3. F9 FT B1021.2
  4. F9 FT B1031.2

In [14]:
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_
```

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

Out[14]:

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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes are 100 and 1, respectively.

In [17]:
```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS count FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

Out[17]:

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are as follows:

```
In [18]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX
 * sqlite:///my_data1.db
Done.
```

Out[18]:

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are as follows:

In [19]:
```
%sql SELECT substr(Date, 6, 2) AS month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Lar
```

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

Out[19]:

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

In [20]:
```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND
```

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

Out[20]:

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

Conclusions

- After testing logistic regression, support vector machines, decision tree classifier, K-nearest neighbors, the accuracy performed by 4 algorithms was the same, so we can use the model trained by one of these 4 algorithms to predict if the Falcon 9 first stage will land successfully.

Appendix

- https://github.com/Superhero0706/IBM_Applied_Data_Science_Capstone/tree/main

Thank you!