IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yi Sheng Goh
29 Jul 23

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection through web scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL & Data Visualisation
  - Interactive Visual Analysis with Folium
  - Machine Learning Predictions
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive Analysis result
  - 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. 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 space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers

1. What factors determine if the first stage will land
2. Interaction between the factors to determine the success rate
3. What factors to be in place to have a higher success rate

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected using SpaceX API and web scraping through wikipedia
- Perform data wrangling
    - One-hot encoding was applied to categorical features
- 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

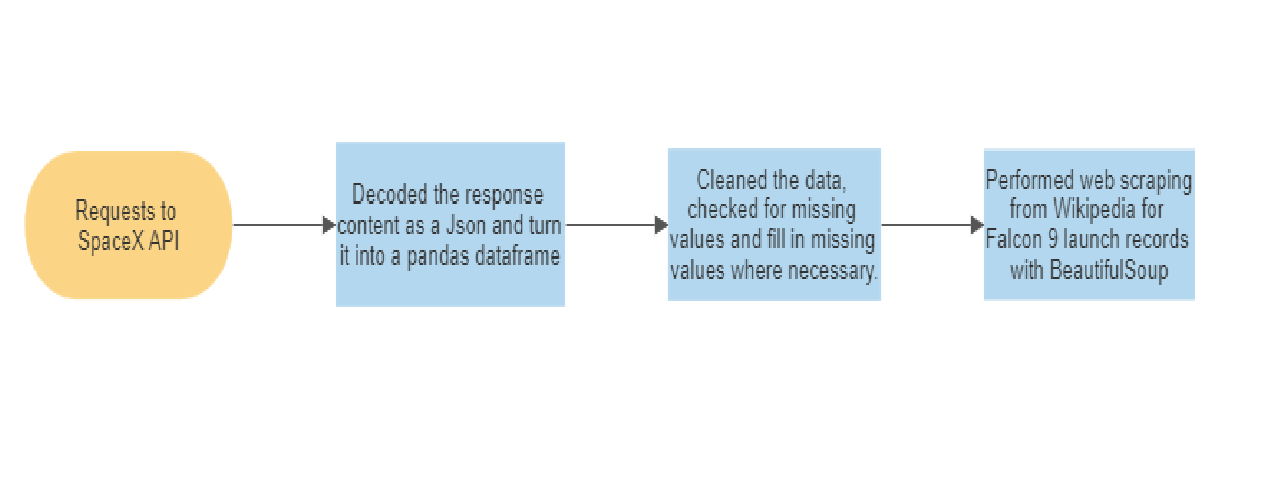
Data Collection

- Data collection was done using requests to SpaceX API
- Decoded the response content as a Json and turn it into a pandas dataframe
- Cleaned the data, checked for missing values and fill in missing values where necessary.
- Performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.

Data Collection – SpaceX API

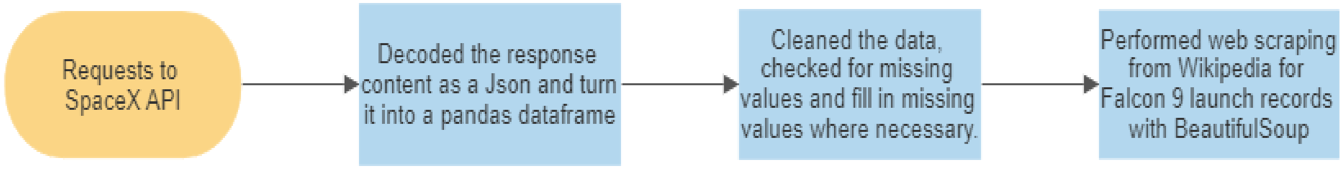
- SpaceX API & Web scraping from SpaceX's Wikipedia page

- https://github.com/yishenggoh/Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Webscrap Falcon 9 launch records with BeautifulSoup
- Parsed the table and converted it into a pandas dataframe



- https://github.com/yishenggoh/Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

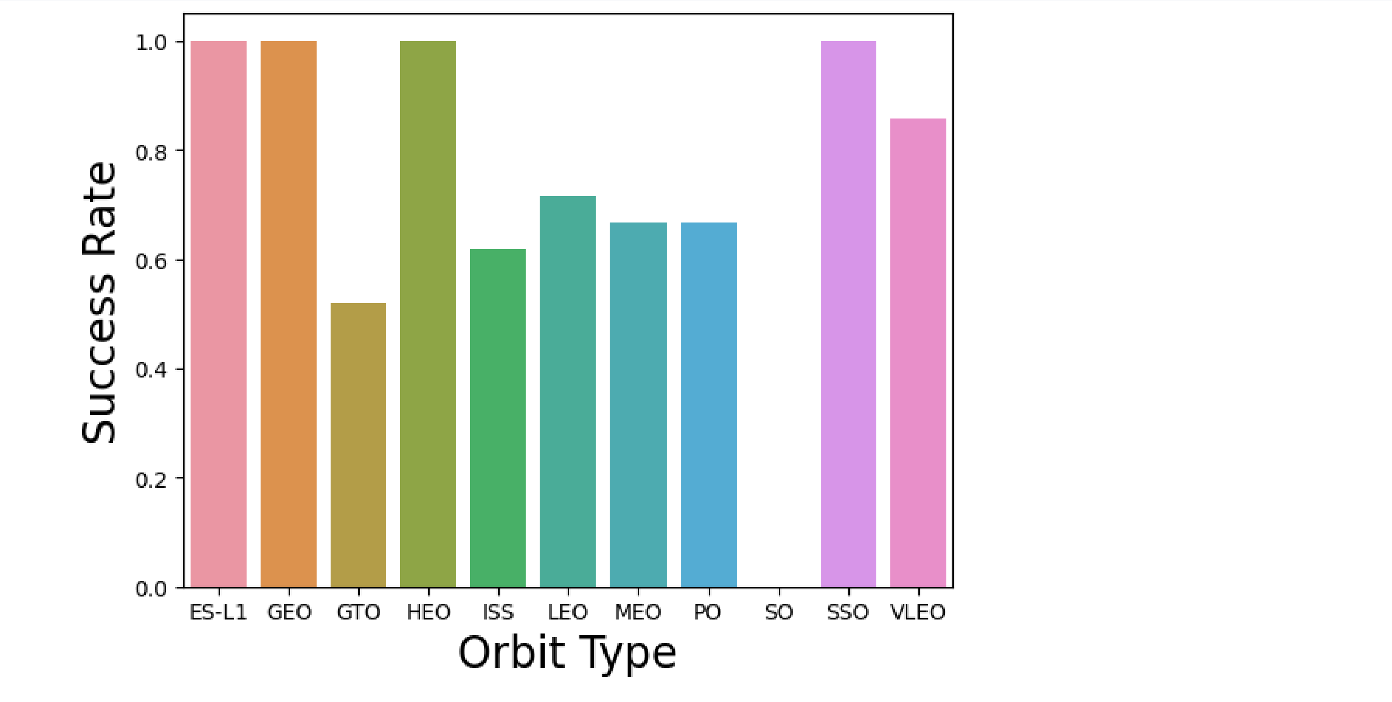
- Performed exploratory data analysis and determined the training labels.
- Calculated the number of launches at each site, and the number and occurrence of each orbits
- Created landing outcome label from outcome column and exported the results to csv.

- https://github.com/yishenggoh/Capstone/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.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.

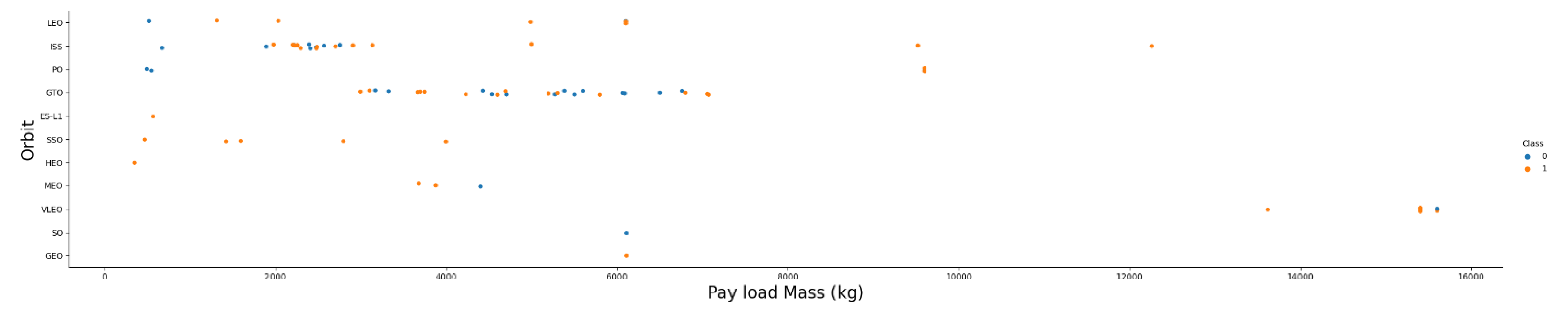
- From the bar chart, we can see the success rate for each orbit type by categorization



- https://github.com/yishenggoh/Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization

- Using the scatter plot, we can see how the 2 factors interact with each other.
- From this scatter plot, we can infer with heavy payloads, the successful landing or positive landing rate are increased for PO ,LEO and ISS.



- https://github.com/yishenggoh/Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

We applied EDA with SQL to get insight from the data. Queries as shown:

- The names of unique launch sites in the space mission
- The total payload mass carried by boosters
- The average payload mass carried by booster version F9 v1.1
- The total number of successful and failure mission outcomes
- The failed landing outcomes in drone ship, their booster version and launch site names.

- https://github.com/yishenggoh/Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marked all the launch sites, assigned launch outcomes and color labeled clusters to know the success rate of the sites.
- Calculated and marked the distance between launch site with nearest coastline, highway, railway and city
- Added these objects to visualize and see if these variables affect the success rate of the boosters landing. In addition, to analysis which site has the highest success rate and why.

- https://github.com/yishenggoh/Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Interactive dashboard with Plotly dash, dropbox with different launch site selectiions.
- Pie charts showing Total launches and success rates by each site.
- Scatter plot showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

- https://github.com/yishenggoh/Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Create Numpy array from Class column
- Standardize the data with StandardScaler. Fit and transform data
- Split the data using train_test_split
- Create a GridSearchCV
- Apply GridSearchCV for Logistic Regression, SVC, Decision Tree & KNN
- Calculate accuracy for all models
- Assess confusion matrix for all models
- Identify the best model

- https://github.com/yishenggoh/Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results

  - Launch success rates have increased over time

  - KSC LC-39A has the highest success rate among the sites

  - Orbits ES-L1, GEO, HEO & SSO have a 100% success rate

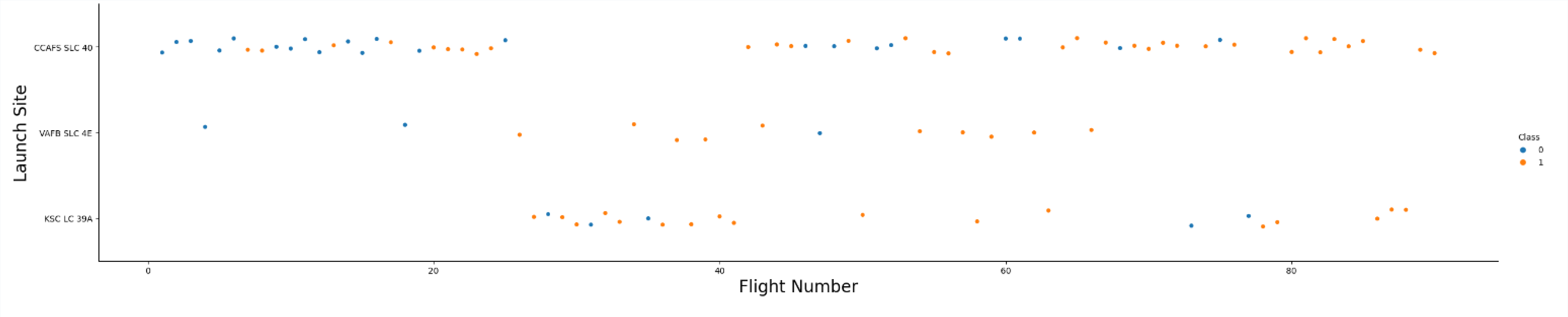
- Predictive analysis results

  - K-Nearest Neighbour is the best predictive model for the dataset

Section 2

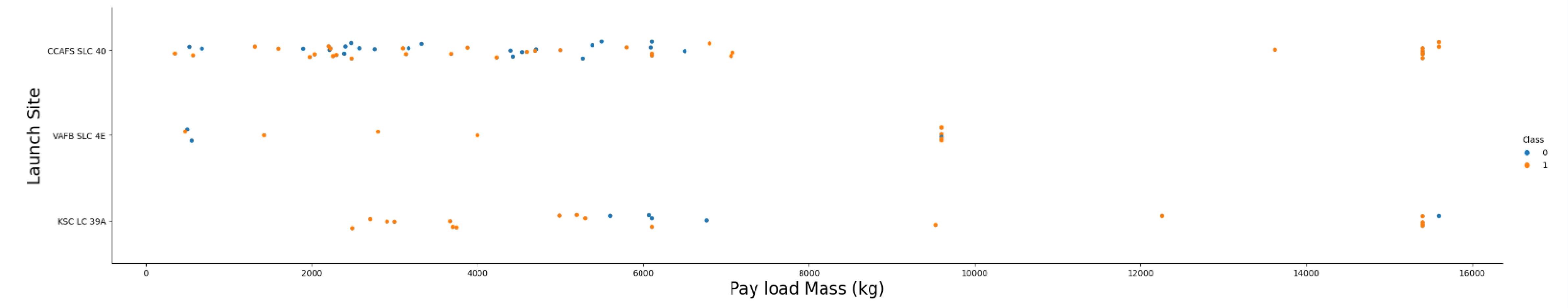
Insights drawn from EDA

Flight Number vs. Launch Site



- Earlier flights have lower success rate
- Over half the launches were from SLC40
- SLC 4E and LC39A has higher success rate
- Newer launches have higher success rates

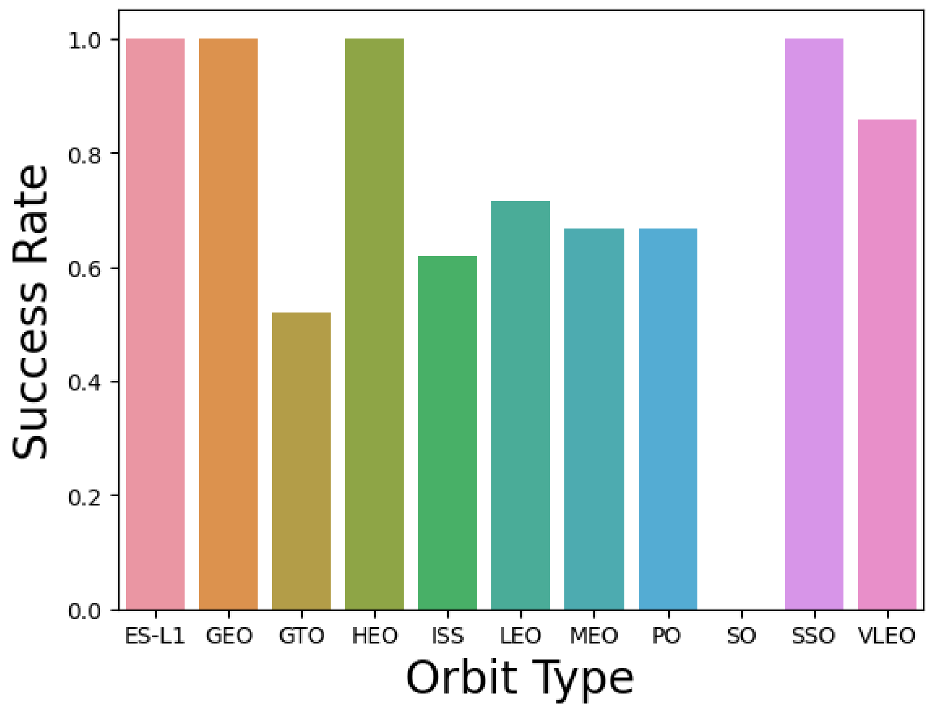
Payload vs. Launch Site



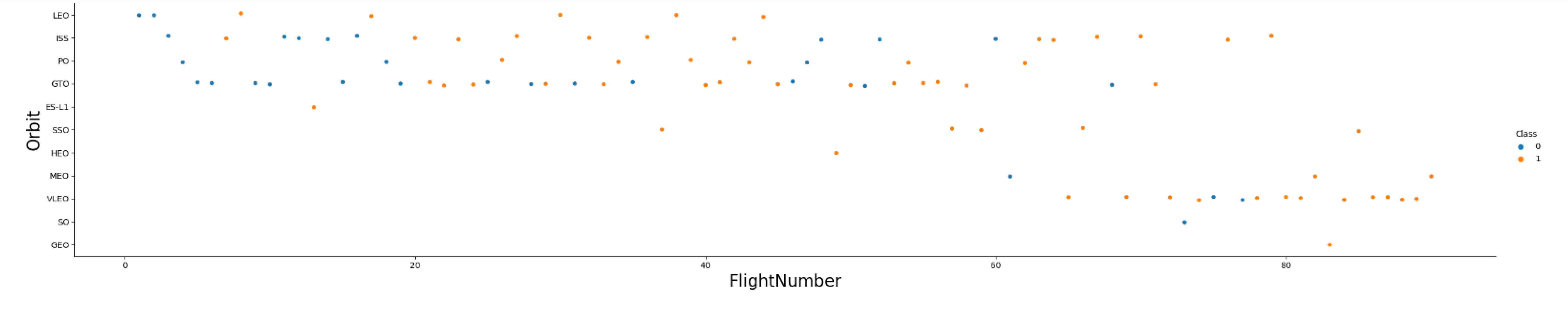
- The higher the payload, the higher the success rate
- Most launches with payload above 8000kg were successful
- Site LC39A has 100% success rate below 5000kg

Success Rate vs. Orbit Type

- Orbit type ES-L1, GEO, HEO & SSO has 100% success rate

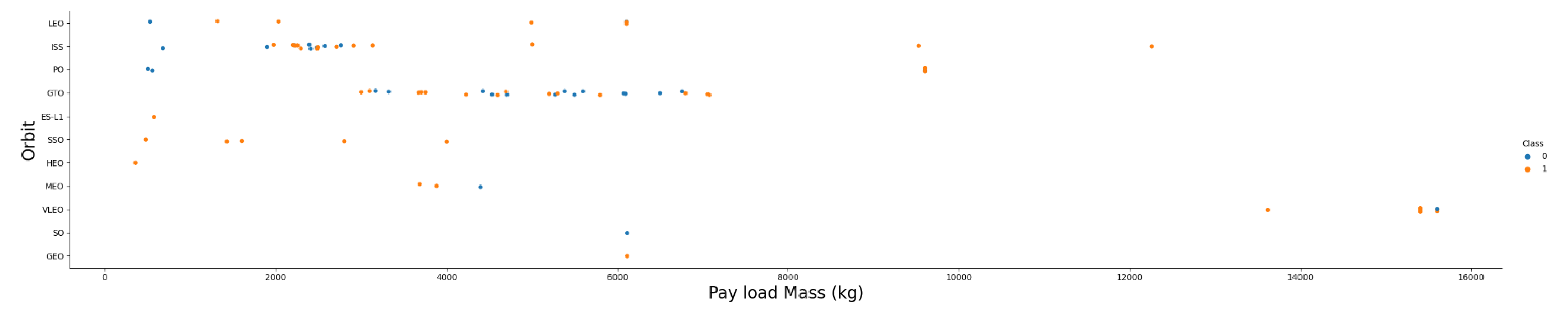


Flight Number vs. Orbit Type



- The newer the flights, the higher the chance of success

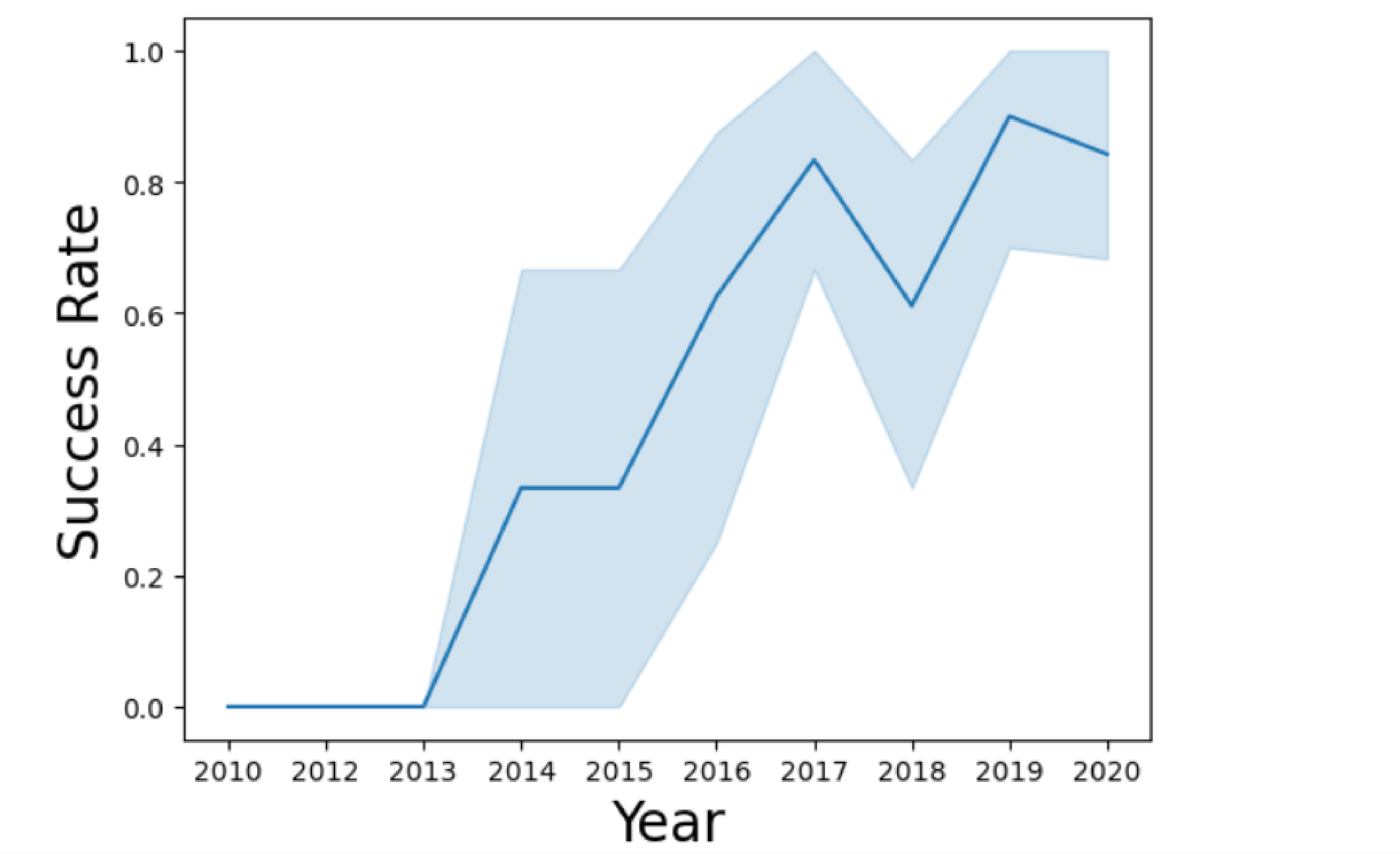
Payload vs. Orbit Type



- We can see that for heavier payloads for LEO, PO & LSS, the success rate is higher compared to a lower payload.

Launch Success Yearly Trend

- From the line chart, success rate has been on the increasing trend to about 80%



All Launch Site Names

- Query using DISTINCT to find all unique launch site names

Task 1

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

In [10]:
```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL
```

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

Out[10]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- Query using ‘WHERE’ and ‘LIKE’ to get records with ‘CCA’ in the launch site name.
- Query with ‘LIMIT 5’ to get the first 5 records

Task 2

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

In [16]:

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

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

Out[16]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No att
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No att
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

- Query using ‘SUM(PAYLOAD_MASS__KG_)’ to get the sum of all the payloads
- Query using WHERE CUSTOMER = ‘NASA (CRS)’ to only select only the payloads from NASA

Task 3

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

In [20]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
FROM SPACEXTBL \
WHERE CUSTOMER = 'NASA (CRS)';
```

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

Out[20]:

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- Query using AVG(PAYLOAD_MASS__KG_) to get the avg of payloads
- Query using WHERE Booster_Version = ‘F9 v1.1’ to only get the avg of payloads from booster version F9 v1.1

Task 4

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

In [21]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
FROM SPACEXTBL \
WHERE Booster_Version = 'F9 v1.1';
```

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

Out[21]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Query using MIN(Date) to get the earliest date
- Query using WHERE Landing_Outcome = 'Success (ground pad)' to get the date of successful landing on a ground pad

Task 5

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

Hint:Use min function

In [22]:

```
%sql SELECT MIN(Date) \
FROM SPACEXTBL \
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Out[22]:

MIN(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query using SELECT Payload to choose the column
- Query using WHERE Landing_Outcome = ‘Success (drone ship)’ to get successful landing on a drone ship
- Query using AND PAYLOAD_MASS__KG__ BETWEEN 4000 AND 6000; to get the payloads between 4000 and 6000

Task 6

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

In [23]:

```
%sql SELECT Payload \
FROM SPACEXTBL \
WHERE Landing_Outcome = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Out[23]:

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- SELECT MISSION_OUTCOME, COUNT(*) as total_number to count the number of occurrence in the unique mission outcome and create a total number column and to group them together

Task 7

List the total number of successful and failure mission outcomes

In [24]:
```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

Out[24]:

Mission_Outcome	total_number
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Query SELECT BOOSTER_VERSION to choose from the column
- Query WHERE PAYLOAD_MASS__KG__ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL); is a sub query to get the highest payload

Task 8

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

In [25]:

```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[25]:

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

- Query to find the month, date, booster version and launch sites in the year 2015 with failure on a drone ship

Task 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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [29]:

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

Out[29]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	01/10/2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14/04/2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Query to count the unique landing outcomes and create a count outcomes column between the 2 dates and groupby the outcomes in descending order

Task 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.

In [30]:

```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

Out[30]:

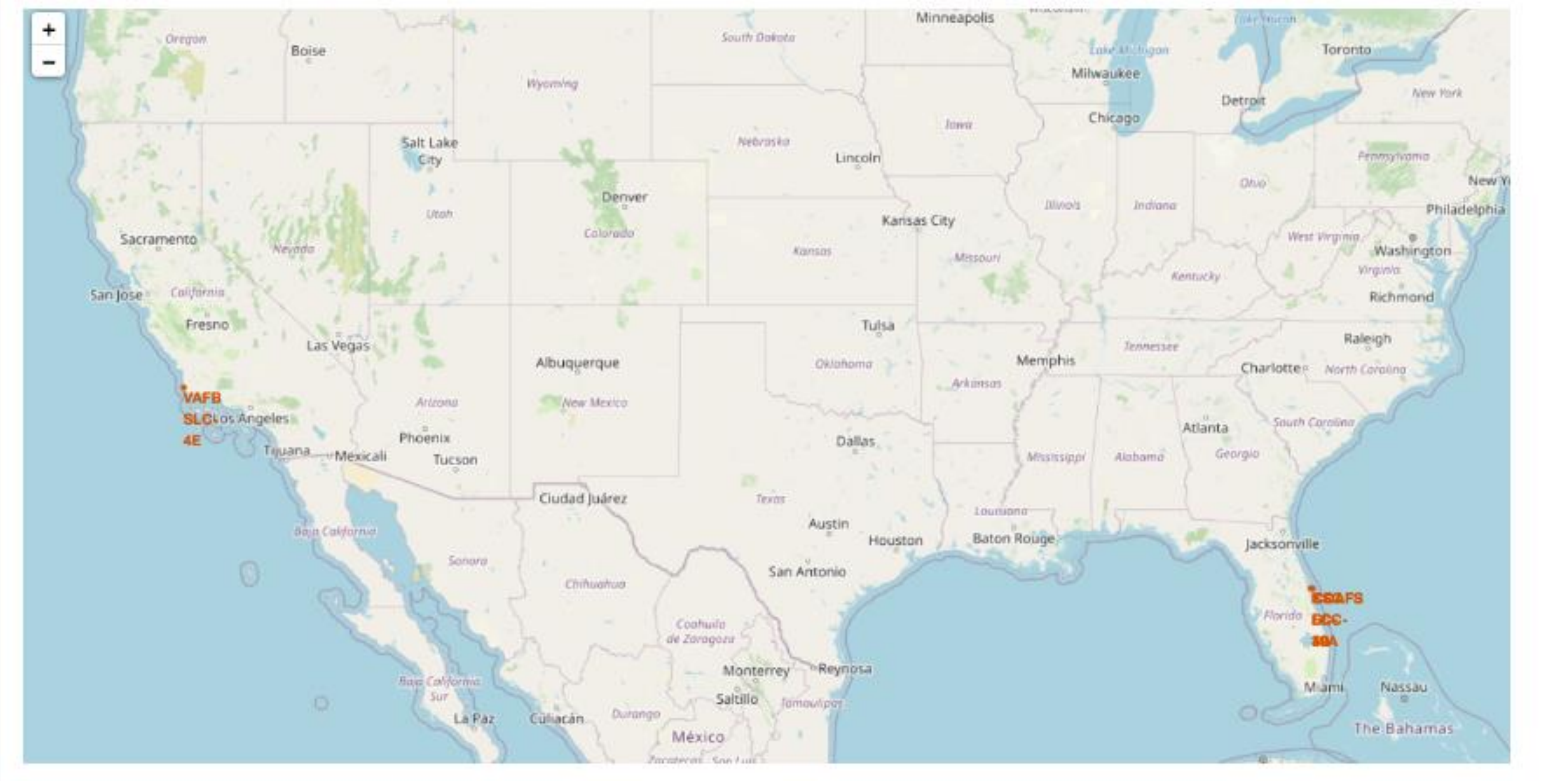
Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Launch Sites

- We can see that the launch sites are all near the coast line

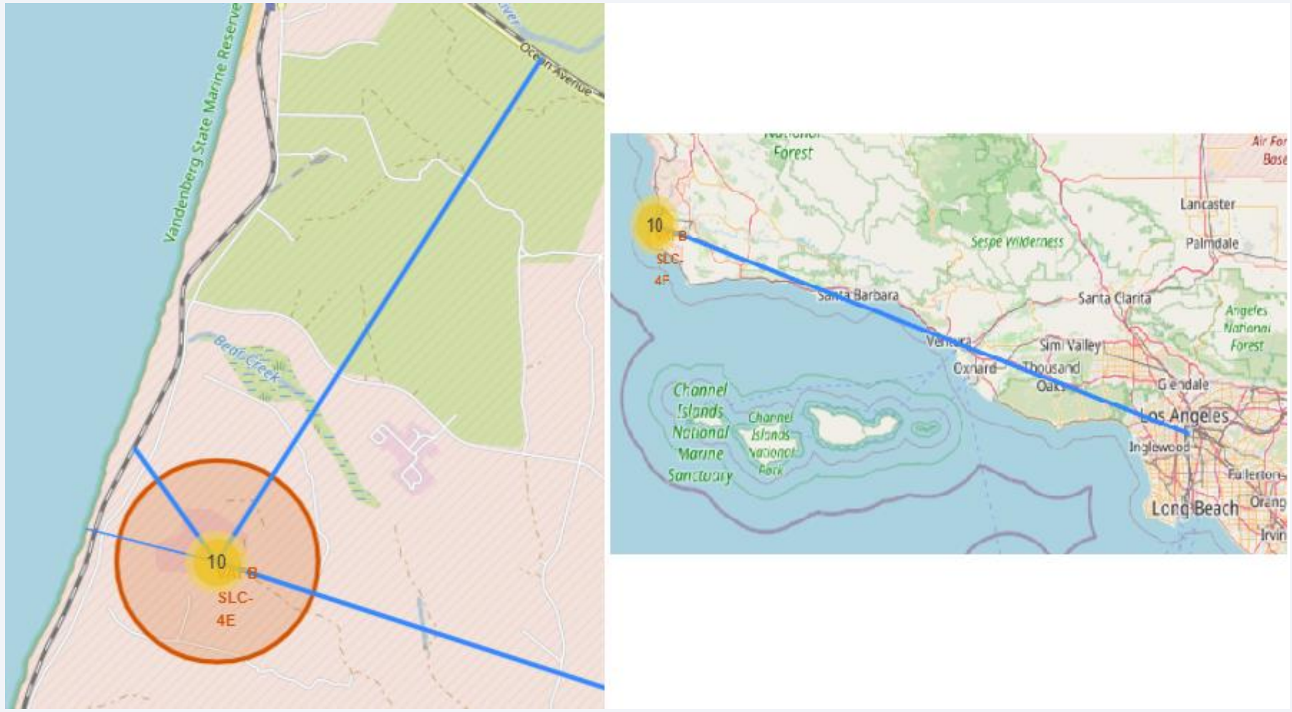


Marker Cluster

- Green Marker shows successful launches and Red marker shows unsuccessful launches

Distance to coastline, railway, highway and city

- Launch Sites are always near the coast line. It can be near a railway or a little further to a highway. But is always a large distance from a city

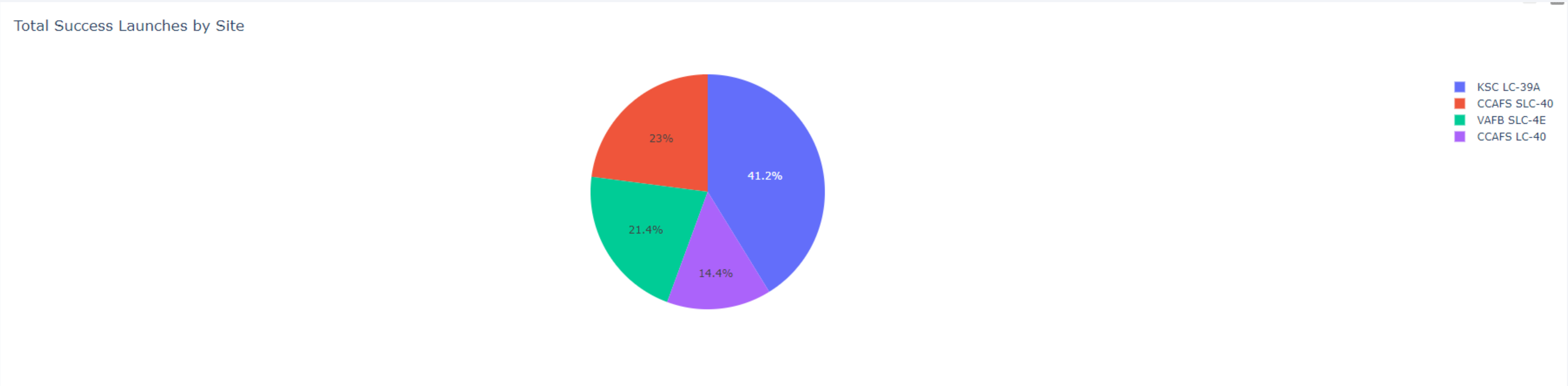


Section 4

Build a Dashboard with Plotly Dash

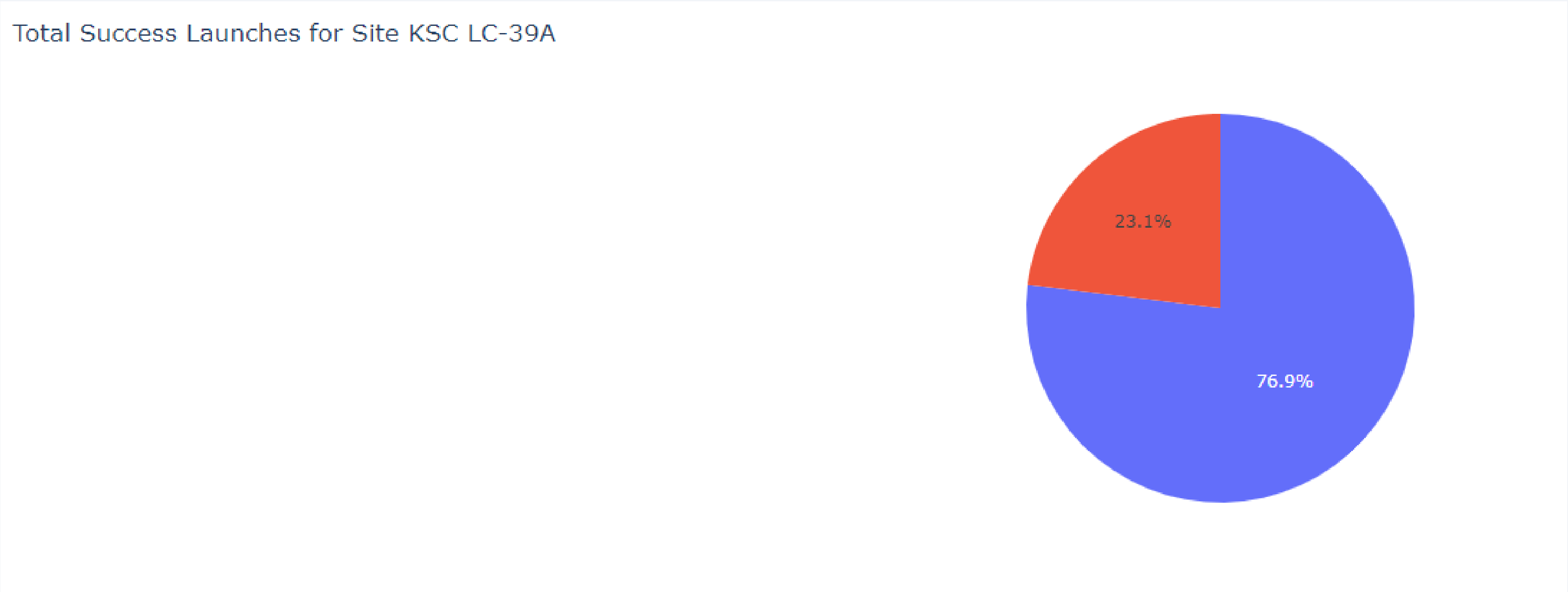
Pie Chart

- Pie chart shows the successful launches by percentage



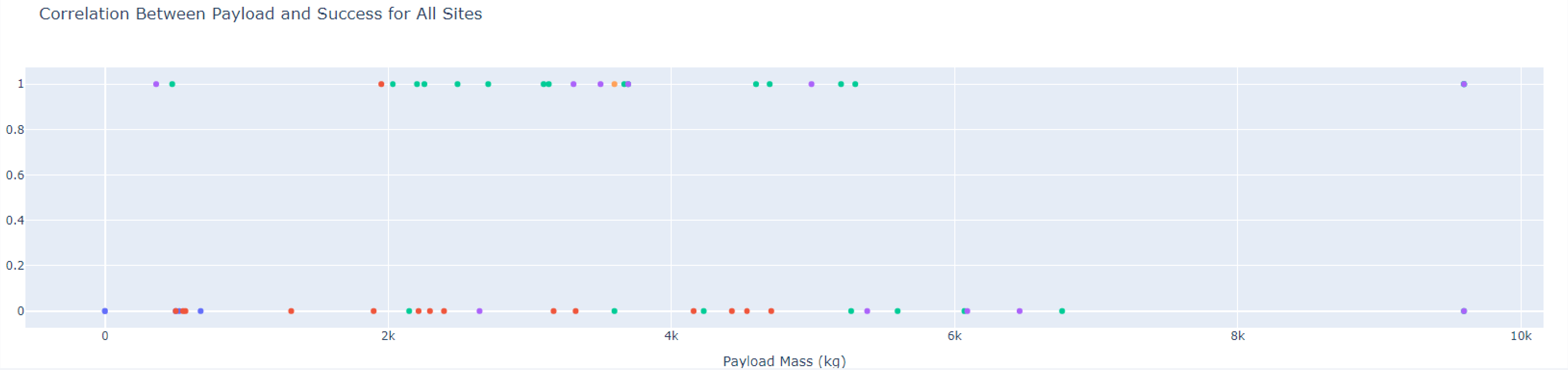
Launch Site KSC LC-39A

- KSC LC-39A has highest success launches of 76.9% and failure of 23.1%



Correlation between payload and success(scatter plot)

- For correlation between payload and success, it seems that lower payload mass below 5000kg has higher success rate than payloads above 5000kg

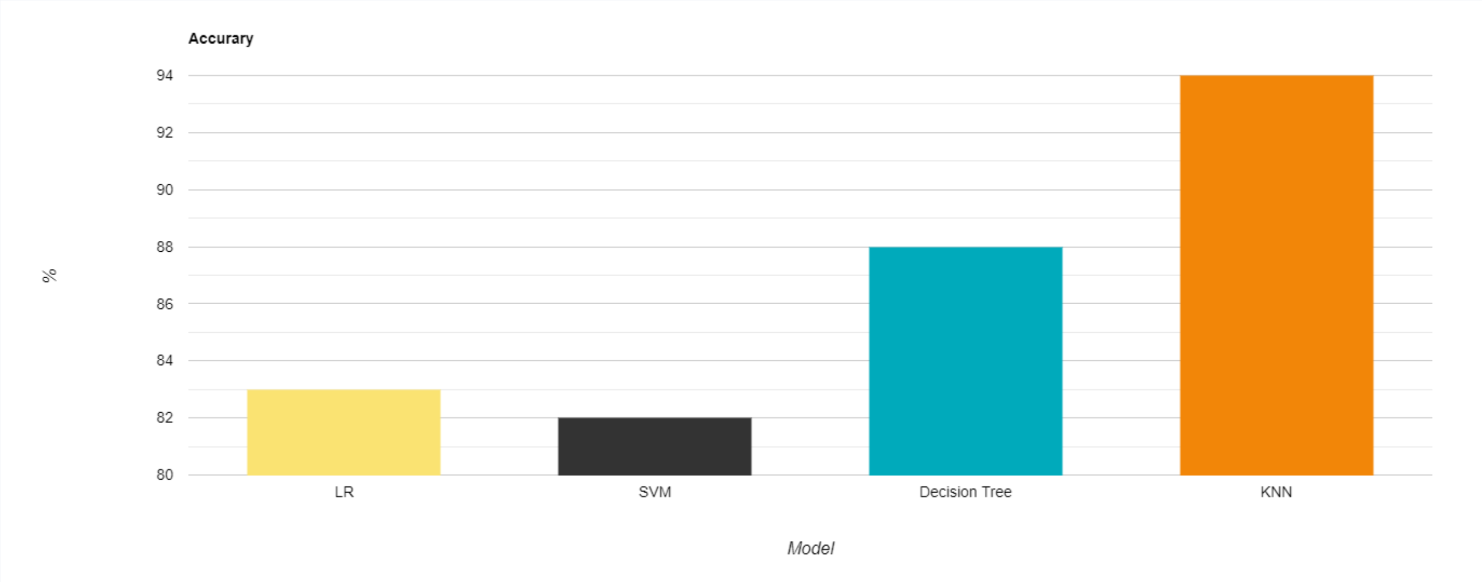


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- KNN has highest Accuracy

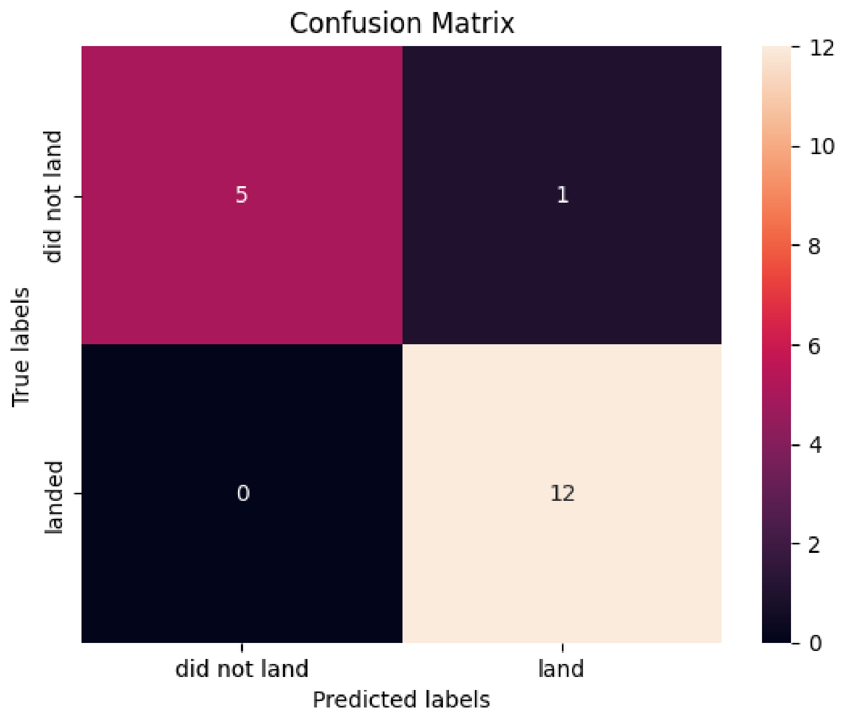


Find the method performs best:

```
Logistic Regression: 0.833
SVM: 0.8222
Decision Tree: 0.888
KNN: 0.944
```

Confusion Matrix

- It shows the number or True positives and False positives. In this case only 1 False positive in which the booster did not land.



Conclusions

- We can conclude that:
- The more the flight amount at a launch site, the greater the success at a launch site.
- Launch success increases from 2013 onwards.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success.
- KSC LC-39A had the most successful launches of any sites.
- The KNN classifier is the best machine learning algorithm for this task.

Thank you!