IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shi Chuanhua

17 Jun 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

Project background and context

SpaceX advertises Falcon 9 rocket launches on its website at a cost of 62 million dollars; other providers cost upward of 165 million dollars each, and 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 you want to find answers

- How payload mass, launch site, number of flights, and orbits affect the first-stage landing success
- Rate of successful landings in different circumstances.
- Which is the best predictive model for successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected via Space X API as well as Wikipedia web scraping
- Perform data wrangling
  - Data was transformed to the formats that allow data analytics, additional steps were taken to deal with missing values and create an outcome column for binary classification.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data collected were normalized, split into test and train data, evaluated using different techniques.

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

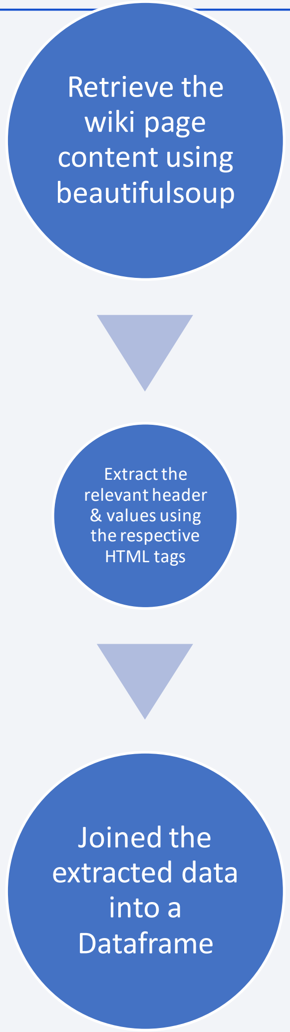
- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/Chuanhua512/IBMDataScience/blob/main/Module%201/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

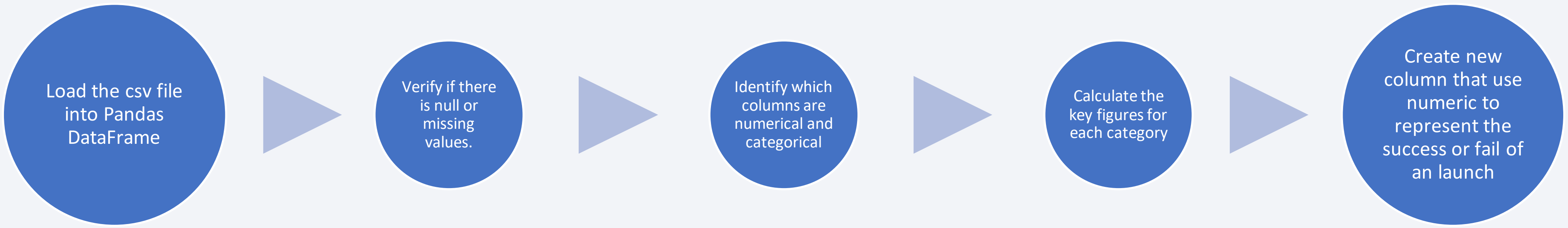
- Present your web scraping process using key phrases and flowcharts

- https://github.com/Chuanhua512/IBMDataScience/blob/main/Module%201/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data wrangling was done using Pandas



- https://github.com/Chuanhua512/IBMDataScience/blob/main/Module%201/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

I made use of scatter charts to observe if the following pairs are related

  - flight numbers vs success rate of the launch sites
  - payload and the success rate of the launch sites.
  - Orbit and the Flight number
  - Orbit and the payload mass

Then I use a bar chart to illustrate the success rate of each orbit, a line chart to show the trend over the year

https://github.com/Chuanhua512/IBMDataScience/blob/main/Module%202/edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
    - Retrieve the unique launch sites in the space mission
    - Display 5 records where launch sites begin with the string 'CCA'
    - 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 succesful 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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months 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.
- https://github.com/Chuanhua512/IBMDataScience/blob/main/Module%202/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

Markers are used to indicate the launch sites, circles are used to highlight the surroundings of NASA JSC & the launch sites, lines were used to display the distance between the launch site & specified coordinates.

- Explain why you added those objects

They were used for ease of identification for the objects.

https://github.com/Chuanhua512/IBMDataScience/blob/main/Module%203/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

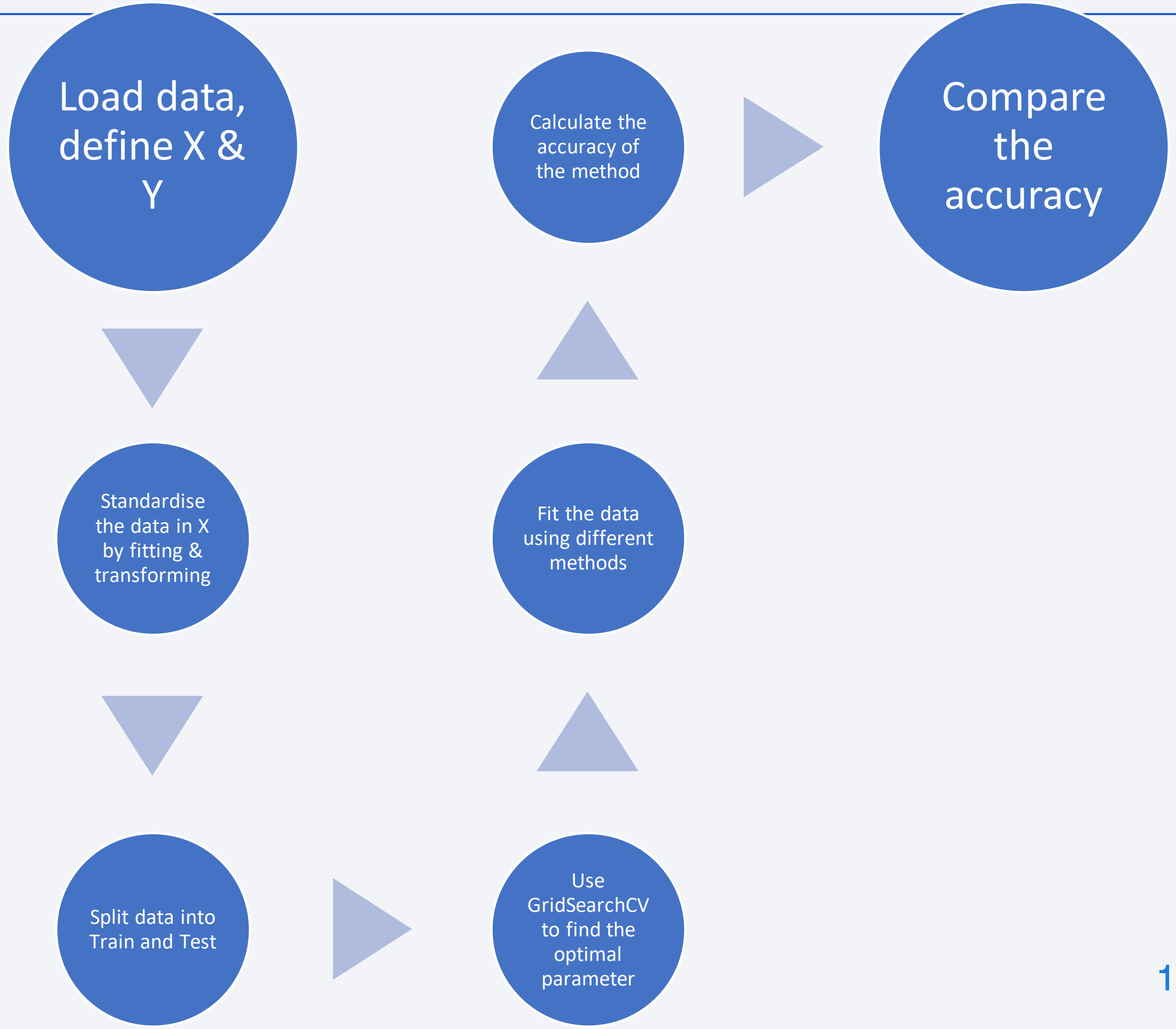
- I have created a dashboard with Pie charts that show the success count of each site, and within each site, there's another Pie chart that displays the success rate of the respective site. This allows us to interpret which site has the highest success rate.
- I have also created a scatter plot that compares the Payload mass and the success rate, the plots are differentiated by booster versions using different colours, so we can judge which booster tends to have a higher rate of success.

- https://github.com/Chuanhua512/IBMDataScience/blob/main/Module%203/spacex_dash_app.py

Predictive Analysis (Classification)

- The data was split into training and testing data, it was then tested with 4 different methods such as KNN, Log regression, SVM and decision tree.
- I then compare the accuracy of the 4 models against the test data.

- https://github.com/Chuanhua512/IBMDataScience/blob/main/Module%204/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

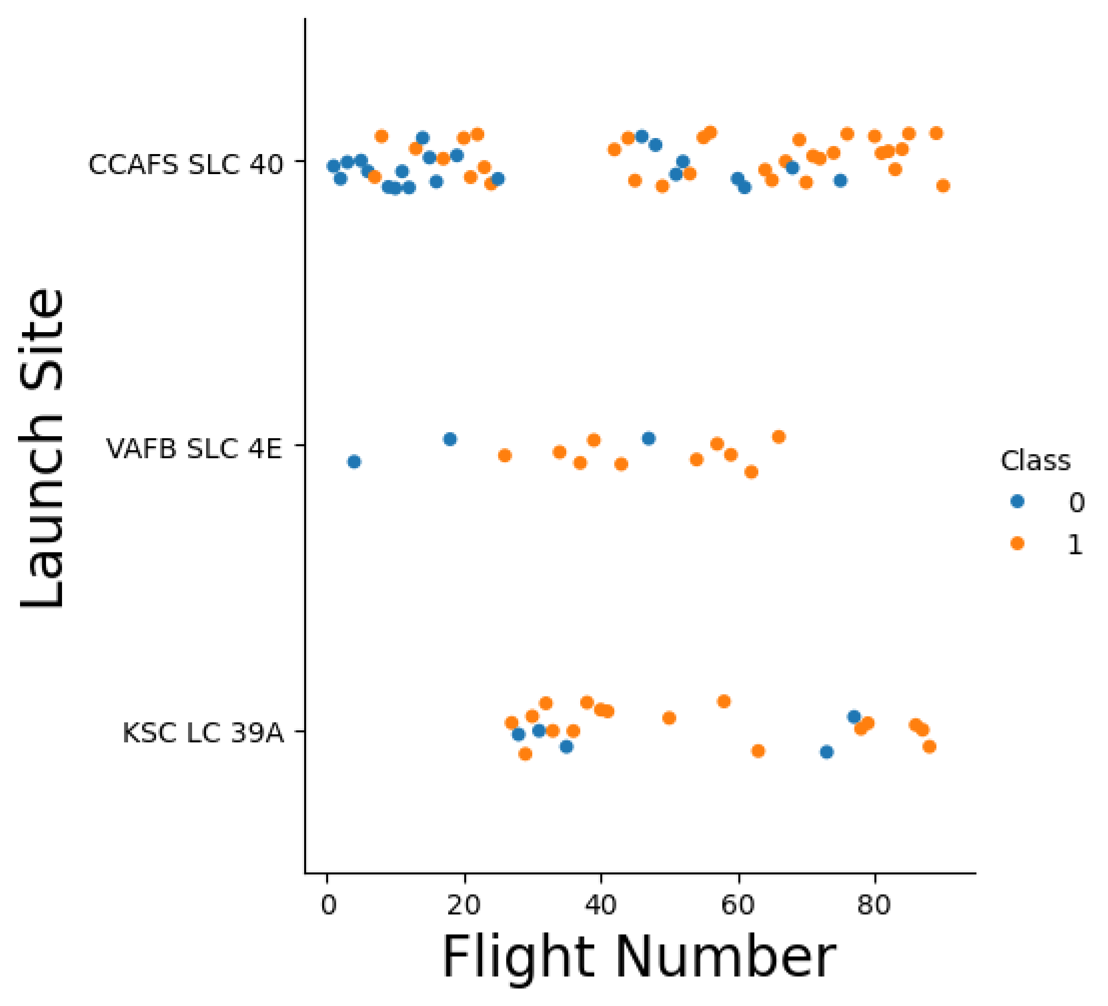
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

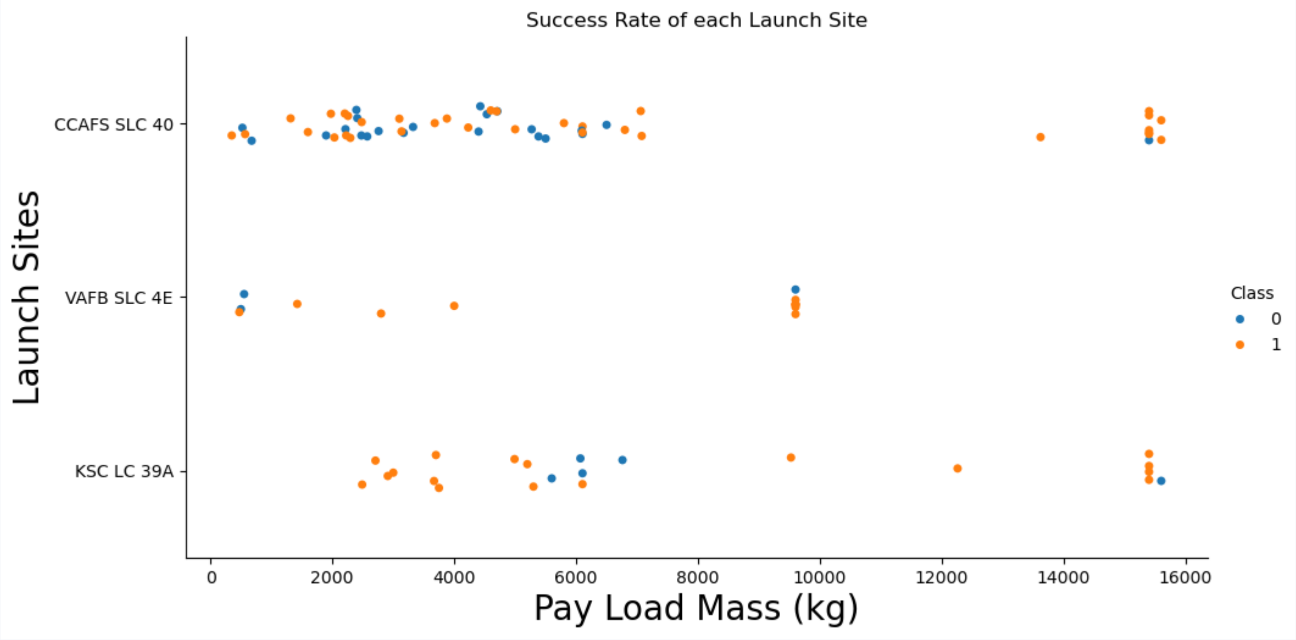
Flight Number vs. Launch Site

- We can deduce that VAFB SLC 4E & KSC LC 39A tend to have a higher success rate



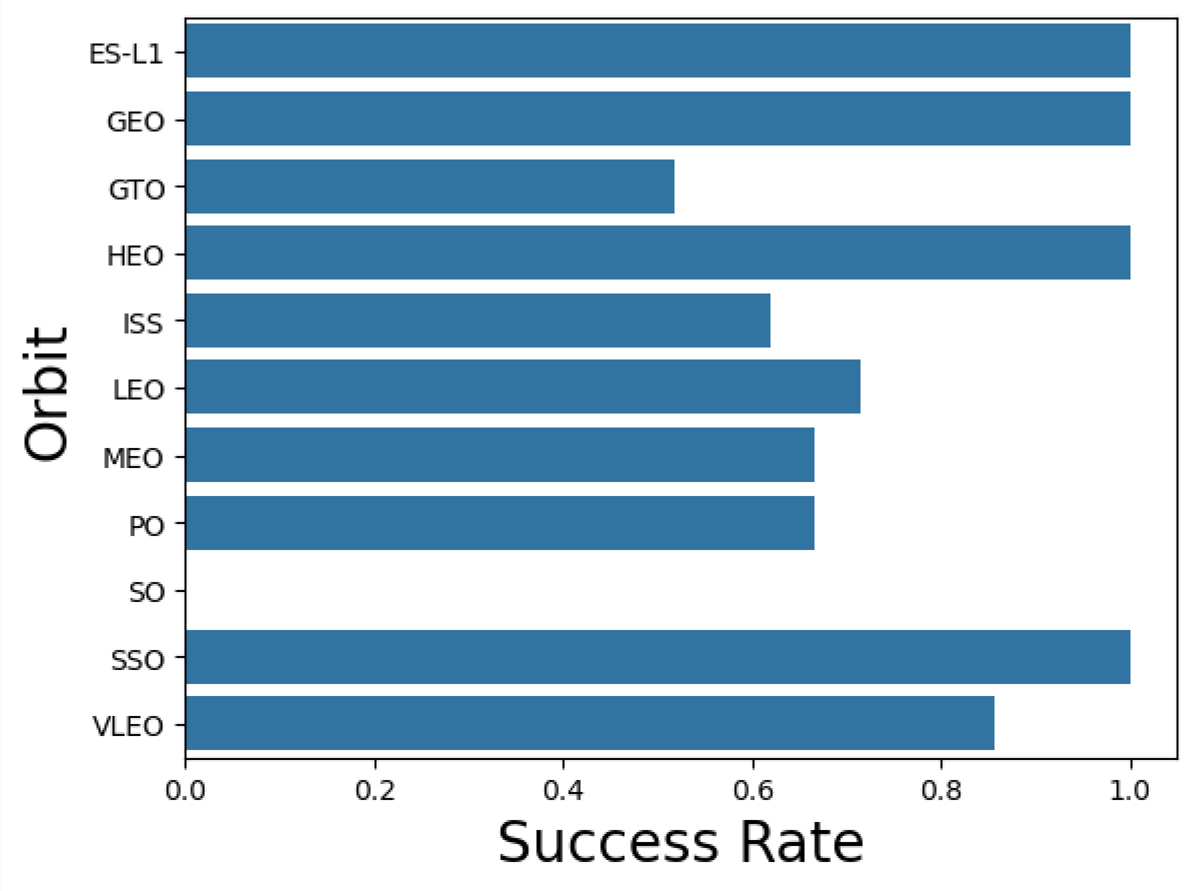
Payload vs. Launch Site

- Payload & Launch site don't seem to have a direct link, we can also tell that VAFB SLC 4E does not have launches with payload higher than 10k



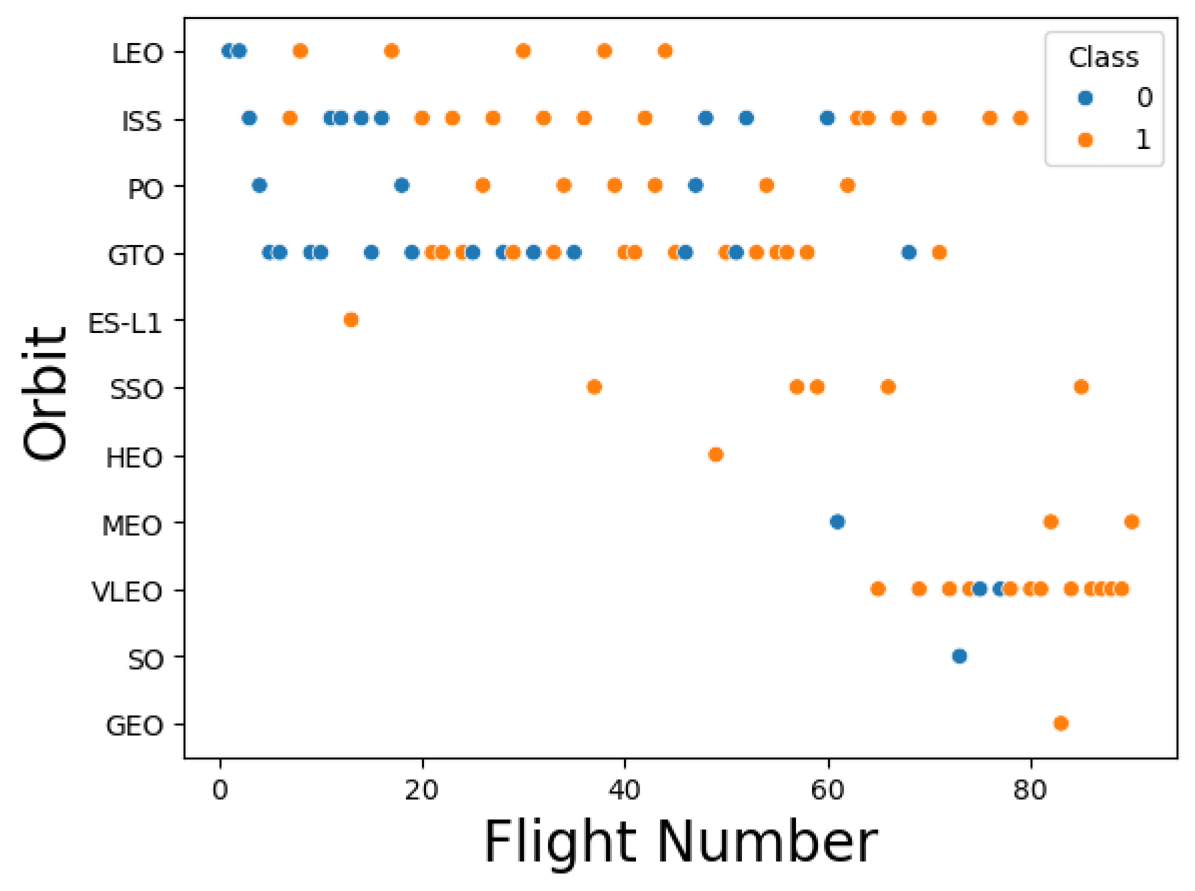
Success Rate vs. Orbit Type

- SO orbit has the lowest rate of success, ES-L1,GEO,GTO,HEO & SSO has the highest rate of success.



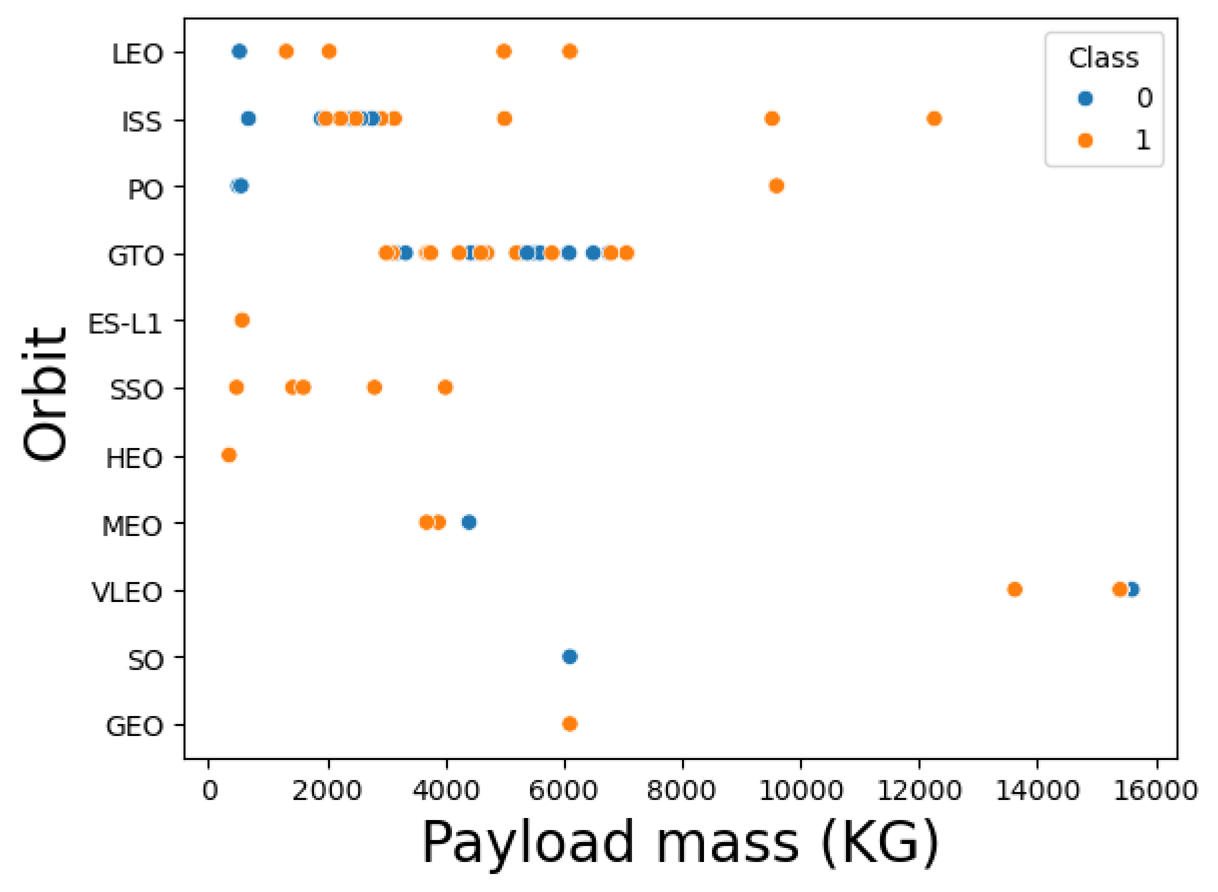
Flight Number vs. Orbit Type

For some of the Orbits like LEO & GTO the rate of success does seem to increase as the flight number increases.



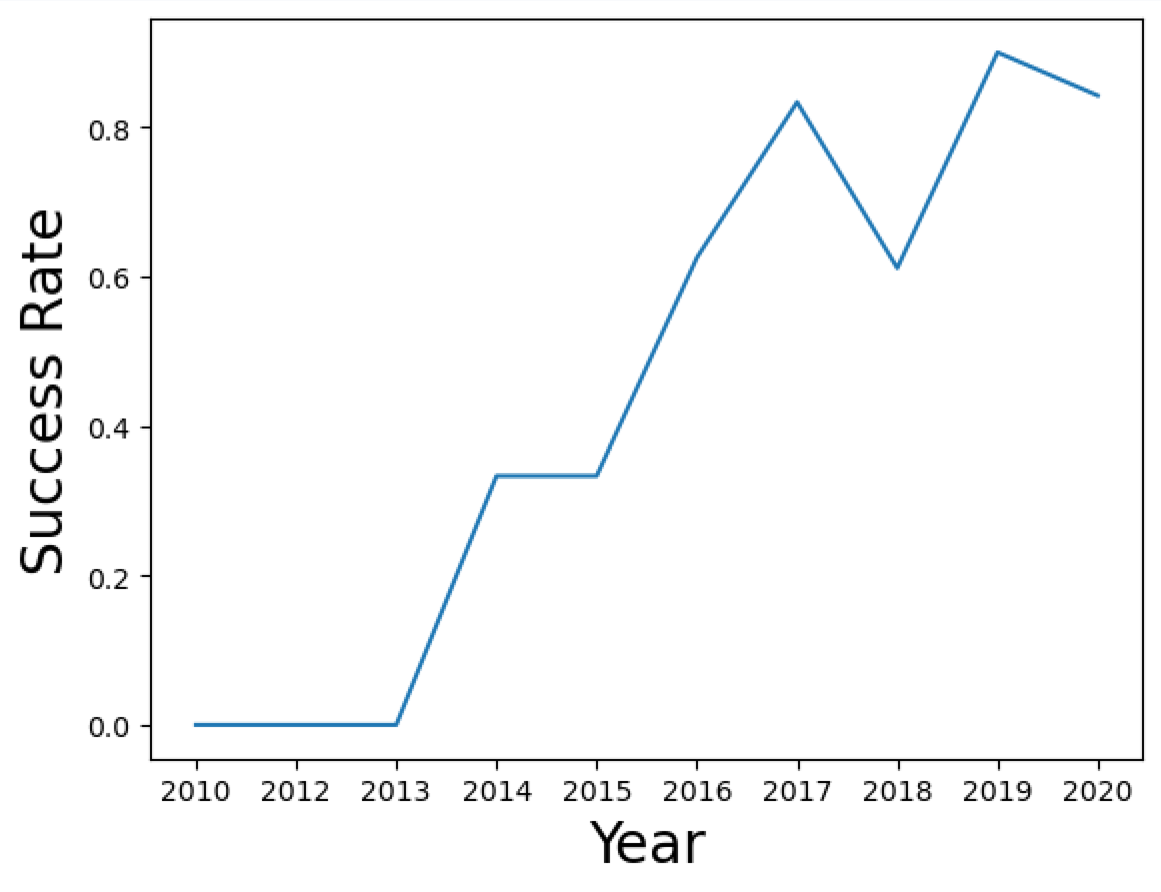
Payload vs. Orbit Type

- The Payload mass and Orbit type don't seem to have a direct link.
- For LSS & LEO the heavier the payload the higher the success rate.



Launch Success Yearly Trend

- As the year increases, the launch success rate increases. Likely due to more advance technology.



All Launch Site Names

- Simply select distinct from the Launch_Site column.

Task 1

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

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'

- Select all columns from table where Launch_Site like CCA%

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

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	Mis
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Total Payload Mass

- Use SQL to sum the payload mass column where customer is NASA (CRS)

Task 3

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

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Select the average of payload mass column where booster version is F9 V1.1

Task 4

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

[7]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

[7]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Select the date where the landing outcome is success gound pad and order by date. Get the first date.

Task 5

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

Hint:Use min function

```
[20]: %sql select Date from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)' order by Date Limit 1
```

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

[20]:

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Use the SQL query below for the result

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

```
]: %sql select distinct(Booster_Version) from SPACEXTABLE where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ <6000 and Landi
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Use the query below to get the count of outcome.

Task 7

List the total number of successful and failure mission outcomes

```
[23]: %sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
```

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

[23]:

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

Boosters Carried Maximum Payload

- List the names of the boosters which have carried the maximum payload mass
- Use the SQL query below to first get the max payload mass, then find the booster version where the payload mass is equals to the max payload mass.

Task 8

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

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

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

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

- Select the respective columns where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'

months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql select substr(Date, 6,2) as Month,Landing_Outcome,Booster_Version,Launch_Site from SPACEXTABLE where Landing_Outcome =
```

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

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

- 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

- Select outcome and count where Date >= '2010-06-04' and Date <= '2017-03-20' Group by Landing_Outcome Order by occurrence desc

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.

```
%sql select Landing_Outcome, count(*) as occurrence from SPACEXTABLE where Date >= '2010-06-04' and Date <= '2017-03-20' G
```

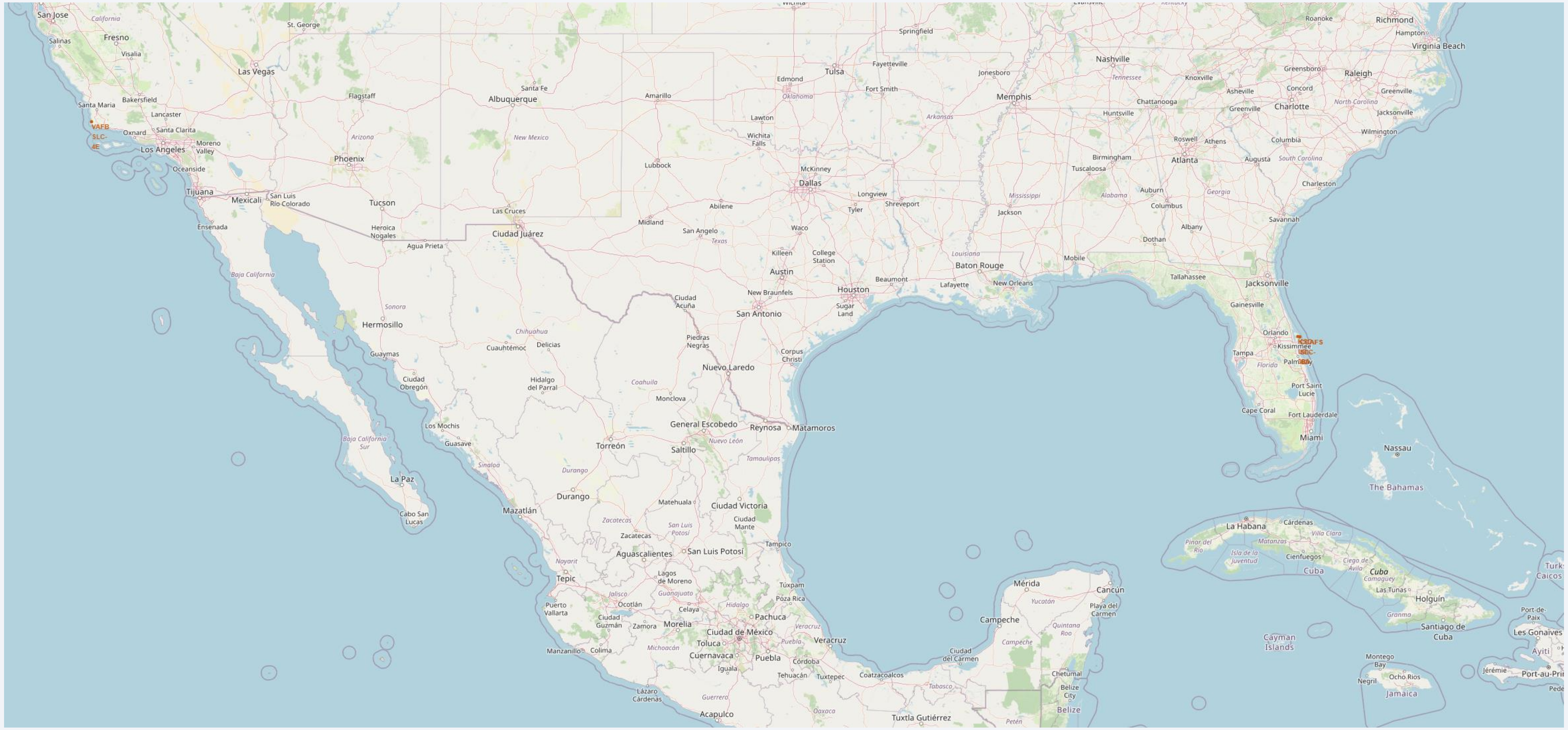
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	occurrence
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

Section 3

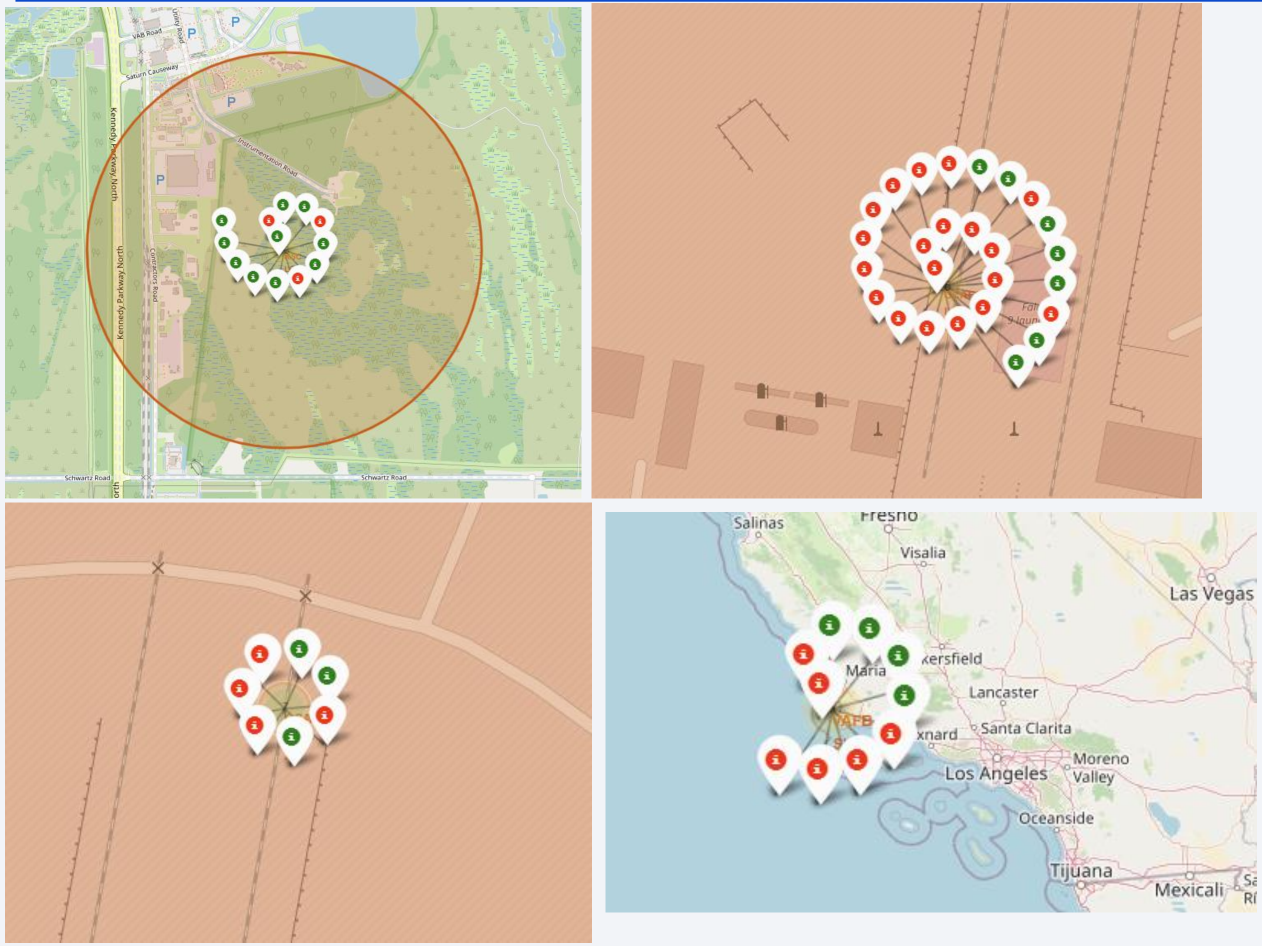
Launch Sites Proximities Analysis

Launch Site locations



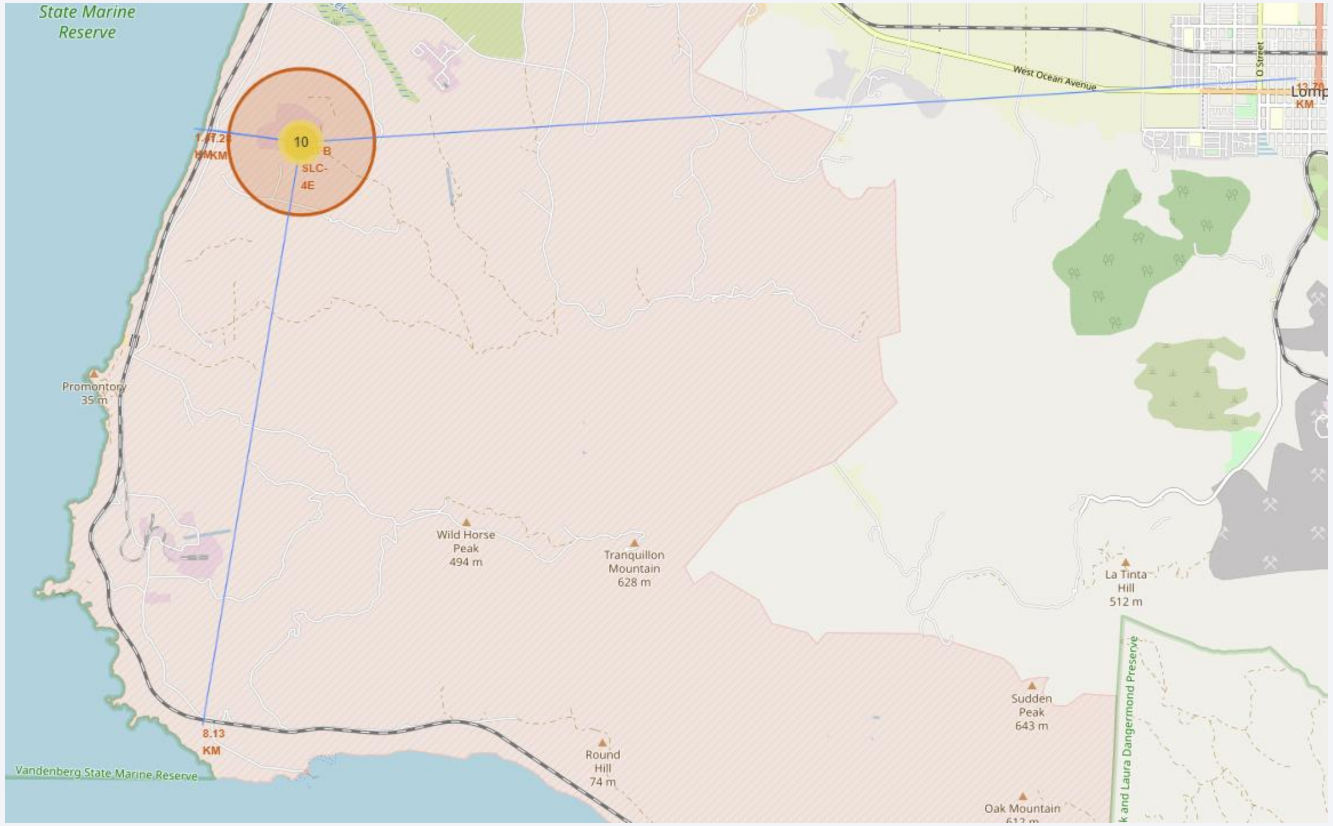
- The launch sites are located close to the coast in general. In case of a launch mishap or explosion, debris falls away from the launch pad. An ocean provides a vast, mostly unpopulated area for debris to fall into, minimizing the risk of damage or casualties on land.

Success Rate shown on map



- The visual furnished with the successful and non-successful launch attempts, it provides an overview of the launches and by one look, we can infer that base KLC LC 39A has a higher success launches.

Launch site distances to landmarks

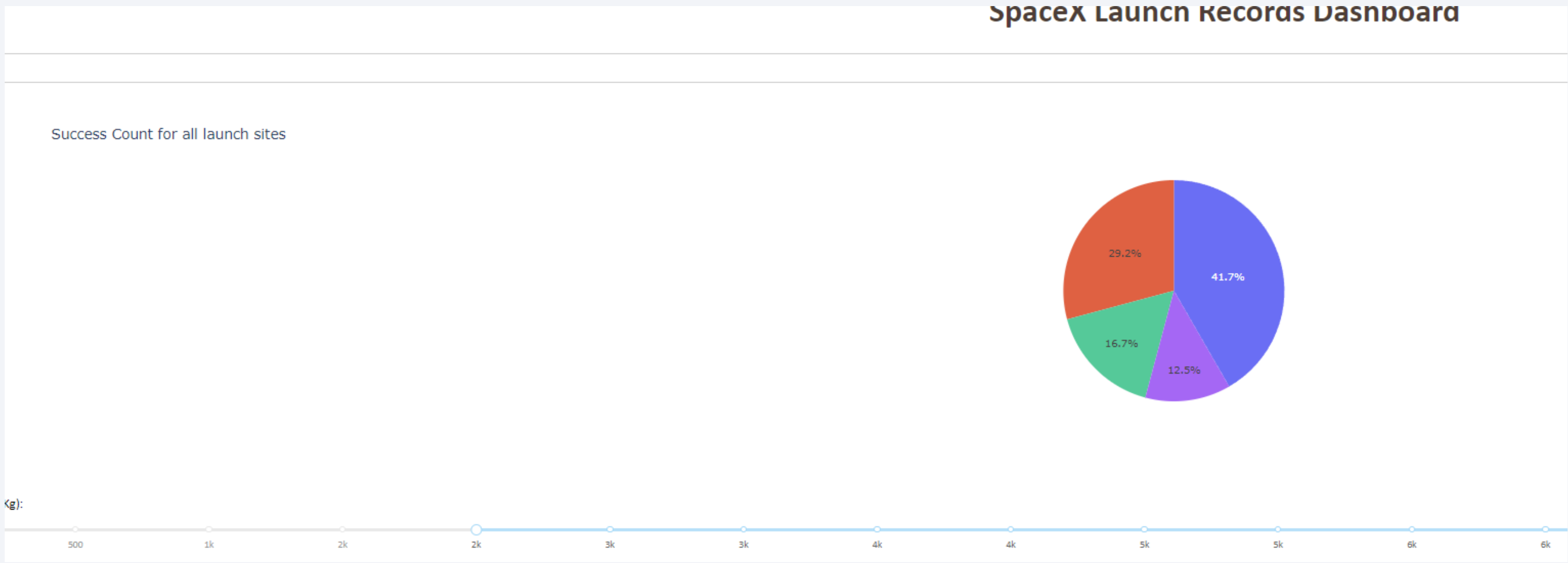


- From the map, we can determine that the launch site is generally further away from populations, and it is accessible by rail transport, although the nearest highway is 8KM away, there should be no issue with supplying necessities to the site.

Section 4

Build a Dashboard with Plotly Dash

Success count of all sites



- This chart shows the total count of successful launch of each sites. However it is important to know that this ratio doesn't mean a site has high success rate, the success rate for each site should be calculated separately.

Site with Highest Success Rate

- This chart shows the site with the highest success rate, it is a fairer comparison as compared to the pie chart with all sites.

Payload & Launch outcome scatter plot

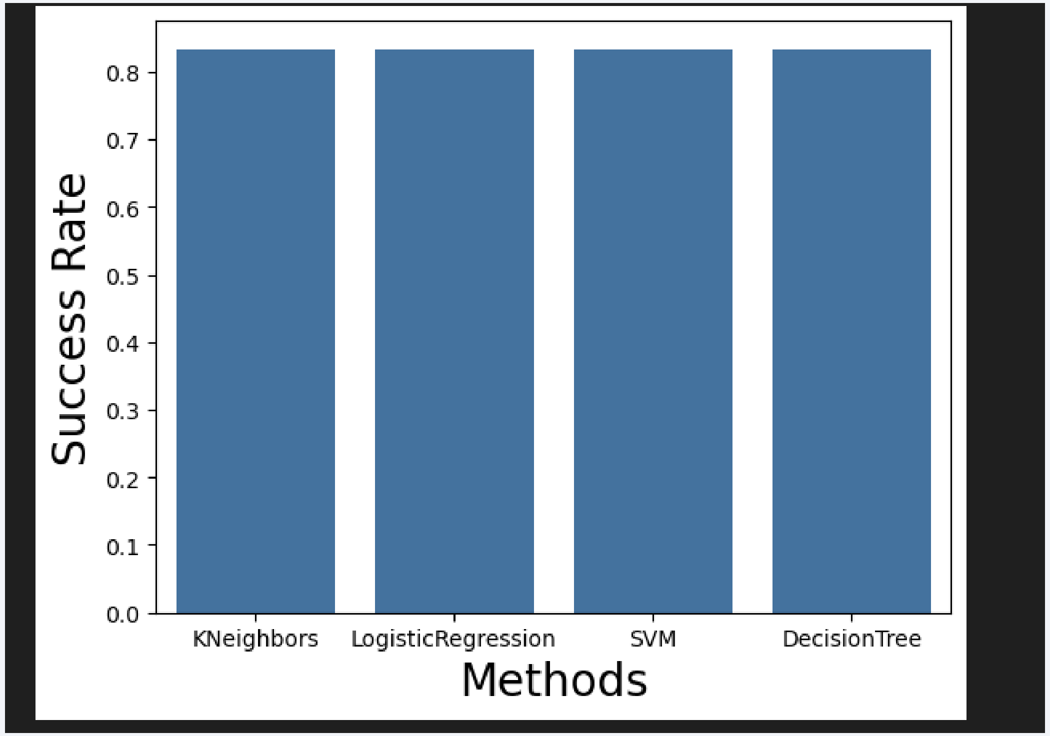
- Payload range below 6K has a higher rate of success, Booster version FT has a higher rate of success.

Section 5

Predictive Analysis (Classification)

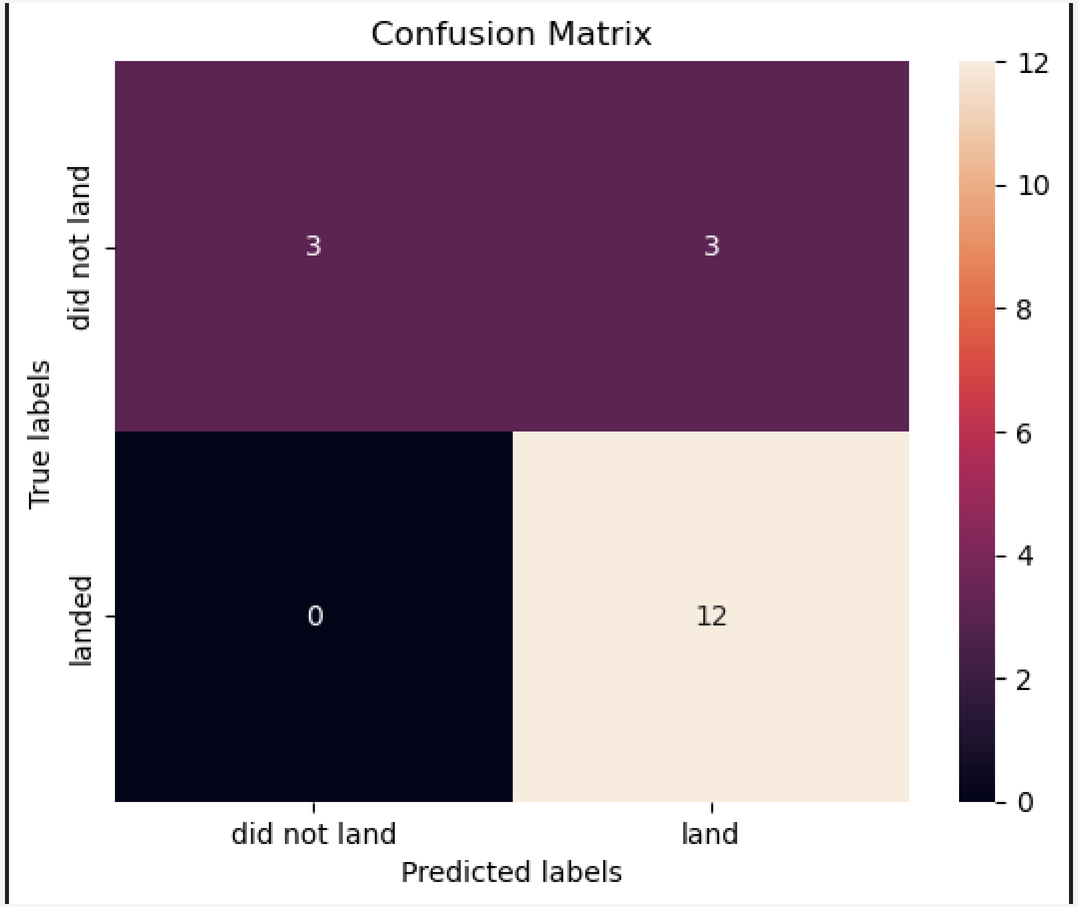
Classification Accuracy

- All model has the same accuracy



Confusion Matrix

- All models are showing the same confusion matrix, this could be due to sample size too small.



Conclusions

- As the flight number increases the likelihood of a successful landing increases
- ES-L1,GEO,GTO,HEO & SSO Orbits has the highest rate of success
- The likelihood of SpaceX being able to reuse their rocket is high. Resulting in significant cost savings, it is viable for Space X to keep the launch price low, which beats other companies in terms of price.

Thank you!