IBM Data Science and Machine Learning Capstone Project

Winning Space Race with Data Science

Yan Jun Chua

23rd Jan 2022

00.

Outline

01 EXECUTIVE SUMMARY

Methodology & Results

Methodologies

- Data Collection & Data Wrangling
- Exploratory Data Analysis (EDA) with Data Visualisation & SQL
- Interactive Visual Analytics using Folium & Plotly Dash
- Predictive Analysis using Classification Models

Results

- EDA results
- Interactive Analytics: Maps & Charts on success launches
- Predictive Analysis results

02 INTRODUCTION

Project Background

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.

Key Question

To find out what would have affected the Falcon 9 landing. In order to answer this question, in this project, we will utilise previous data to identify the factors of the successful landing of Falcon 9, then predict the success rate and cost of landing for stakeholders to decide whether they want to bid for a rocket launch in the future.

Methodology: Data Collection

COLLECT SPACEX LAUNCH DATA

- Request the SpaceX launch data via API
- Filter the data to keep only data for Falcon 9
- After dealing with missing values, convert the data to a CSV file.

Github URL: Data collection

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
```

```
# Calculate the mean value of PayloadMass column
print(df['PayloadMass'].mean())
# Replace the np.nan values with its mean value
df['PayloadMass'].fillna(value=df['PayloadMass'].mean(), inplace=True)
print(df['PayloadMass'].head())

data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

COLLECT DATA FROM WIKIPEDIA

- Web scraping data for Falcon 9 and Falcon Heavy launch records from Wikipedia via URL
- Extract and parse data from the HTML table
- After creating an empty dictionary with keys from extracted data, convert the dataframe to a CSV file.

Github URL: Web scraping

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

```
# use requests.get() method with the provided static_url
page = requests.get(static_url)
```

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
```

```
df=pd.DataFrame(launch_dict)
df.to_csv('spacex_web_scraped.csv', index=False)
```

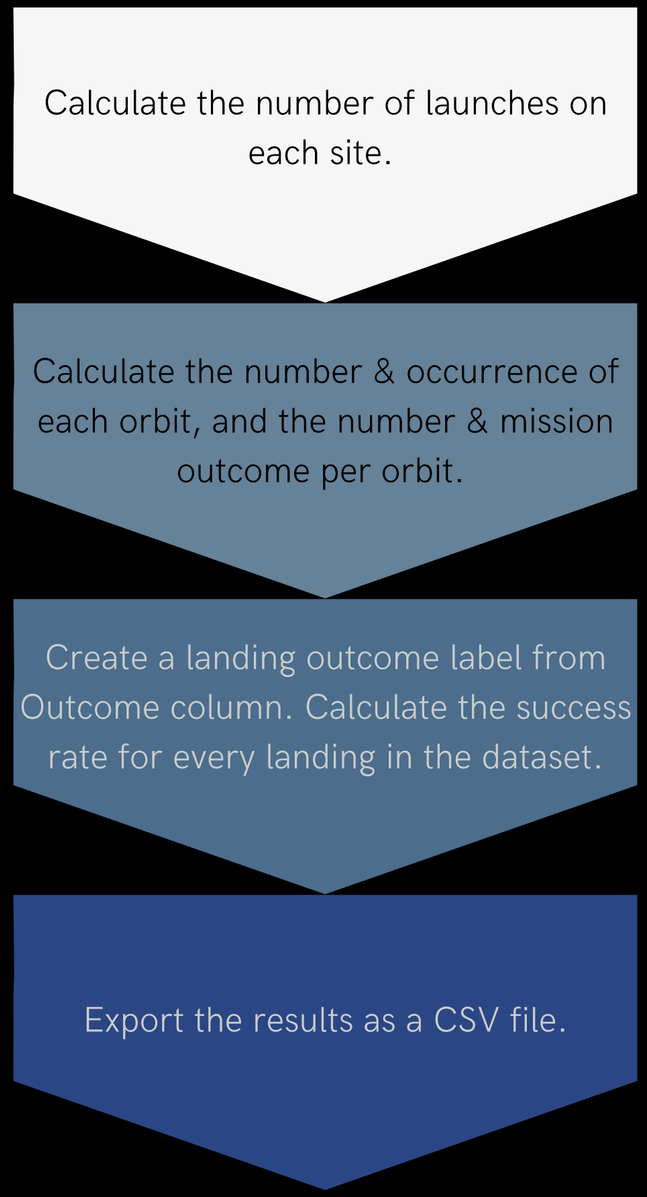
Methodology: Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

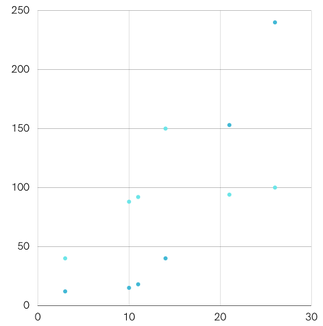
TO DO:

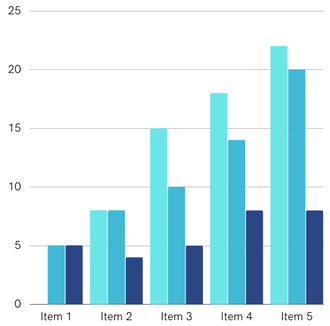
Use Exploratory Data Analysis (EDA) to find patterns in the data.

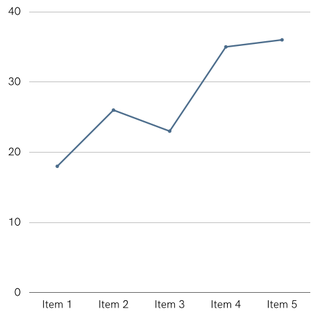
Github URL: Data Wrangling



03







EDA with Data Visualisation

Github URL: EDA with Data Visualisation

Scatter plots:

1. Payload Mass vs. Flight Number
2. Launch Site vs. Flight Number
3. Launch Site vs. Payload Mass
4. Orbit Type vs. Flight Number
5. Payload Mass vs. Orbit Type

To observe possible correlations between these variables.

Bar chart:

1. Success Rate vs. Orbit Type

To identify which orbit has the highest success rate.

Line Graph:

1. Success Rate vs. Year

To show the average success rate by year (annual trend).

SOME QUERIES:

1. Names of unique launch sites
2. Display 5 records where launch sites begins with 'KSC'
3. Total payload mass carried by boosters launch by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. Date where the 1st successful landing outcome in the drone ship
6. Names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
7. Total number of successful and failure mission outcomes
8. Names of the booster_versions which have carried the maximum payload mass
9. Records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
10. Count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20.

EDA with SQL

Using SQL queries to obtain some specific info we needed from the dataset.

03

Github URL: EDA with SQL

Github URL: SQL file as a workaround

03

INTERACTIVE ANALYTICS WITH FOLIUM & PLOTLY DASH

Folium: To view the launch data in an interactive map

Github URL: Folium

- Focus on location - longitudes and latitudes for each launch sites;
- Add a circle marker around each launch site, with a label of name;
- Classify the launch outcomes by failure/success with Red and Green markers, respectively;
- Calculate the distance between launch sites to other landmarks (highways, railways, coastlines, cities) so we can observe the trend of geolocations of these launch sites.

Plotly Dash: To view the launch data in every launch site

Github URL: Plotly Dash

- A dropdown list that shows all the launch sites;
- Pie charts to present the success rate by launch sites;
- A scatter plot to indicate the relationship between Payload Mass and launch outcomes;
- The range of payload mass can be adjusted accordingly to read the scatter plot in detail, i.e. the launch outcomes by booster versions in different payload mass.

Predictive Analysis (Classifications)

03

1. BUILD THE MODEL

1. Load the data to Numpy & Pandas, transform it to fit our purpose;
2. Split the data into training vs. test data;
3. Using GridSearch method, create the below objects:
    a. Logistic Regression
    b. Support Vector Machine (SVM)
    c. Decision Tree Classifier ('Tree')
    d. K Nearest Neighbours (KNN)

2. EVALUATE & IMPROVE

1. Create GridSearch object, run the model with training data;
2. Calculate the accuracy rate of the model with test data;
3. Predict the model with test data and plot a confusion matrix;
4. Repeat steps 1 to 3 for different GridSearch objects.

3. IDENTIFY THE BEST MODEL

According to the calculated accuracy rate, choose the model that has the highest accuracy rate.

Github URL: ML Prediction

04

RESULTS

EDA Data Visualisations

EDA with SQL

Interactive Analytics:

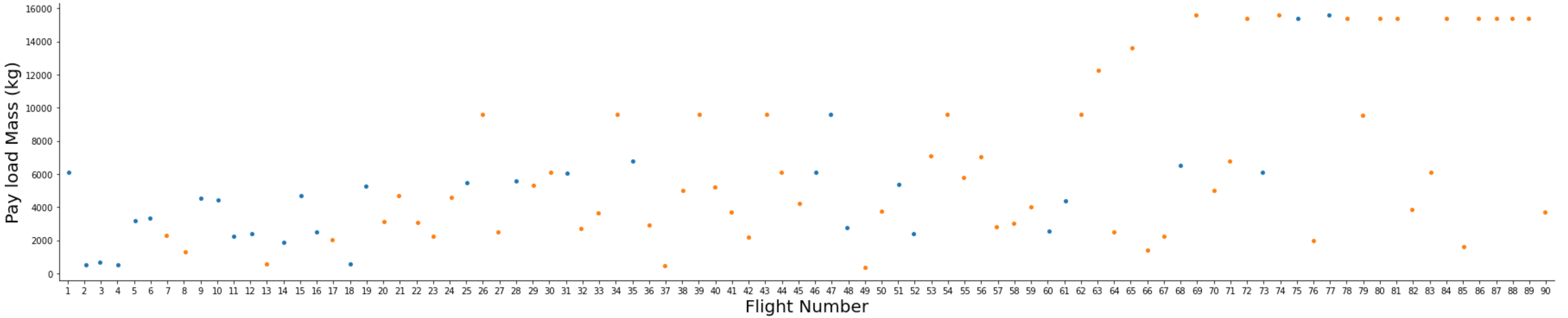
Folium Map & Plotly Dash

Predictive Analysis

Results

04 EDA

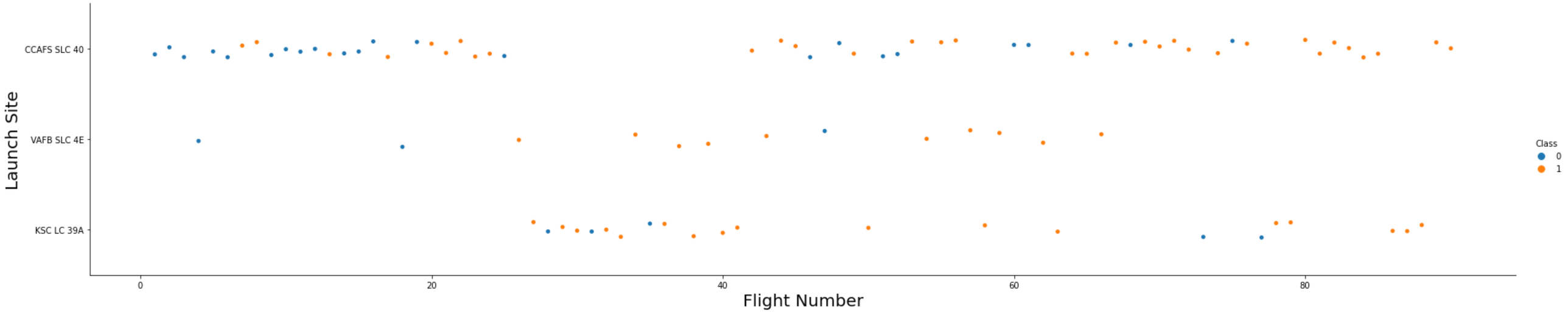
Payload Mass vs. Flight Number



As the flight number increases, the first landing would be more likely to succeed. We can also see that those with 10,000 kgs of payload mass (and above) have on average, a higher success rate.

04 EDA

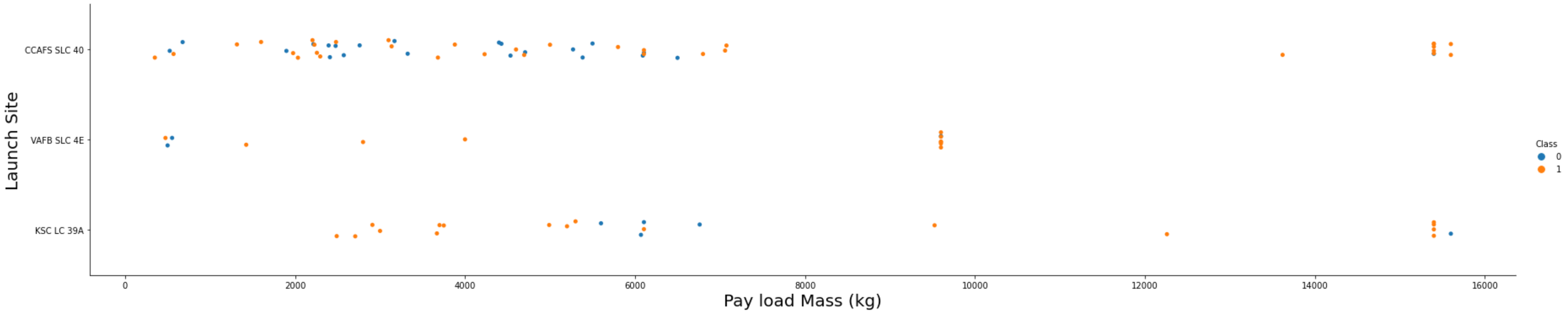
Flight Number vs. Launch Sites



Different launch sites have different success rates. CCAPS SLC-40 has the most launches with a success rate of 60%; while KSC LC-39A and VAFB SLC-4E both have a higher success rate at around 70%.

04 EDA

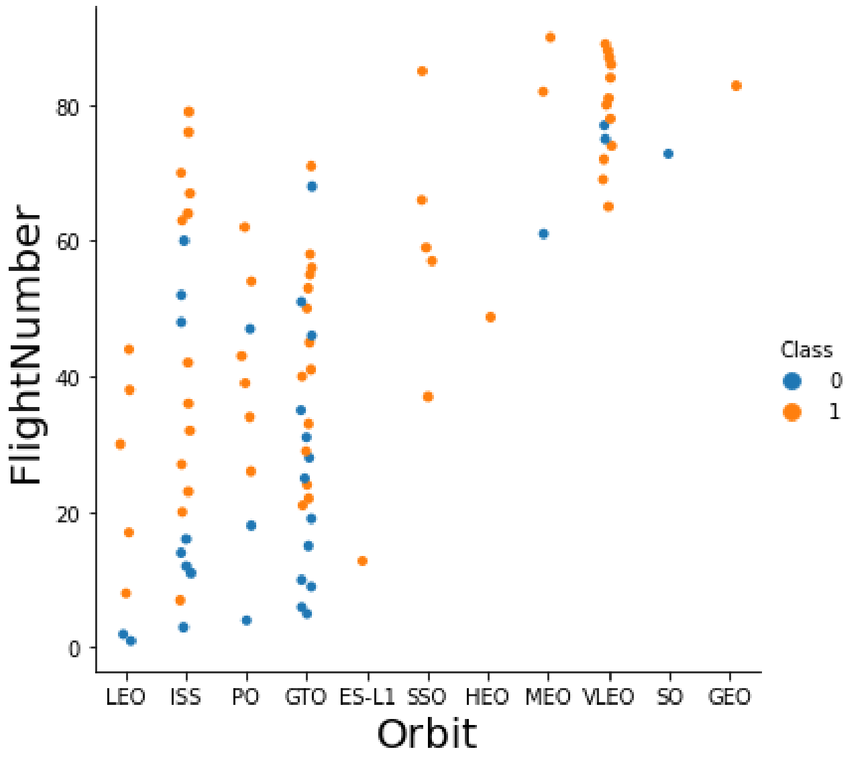
Payload Mass vs. Launch Sites



For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10,000 kgs).

04 EDA

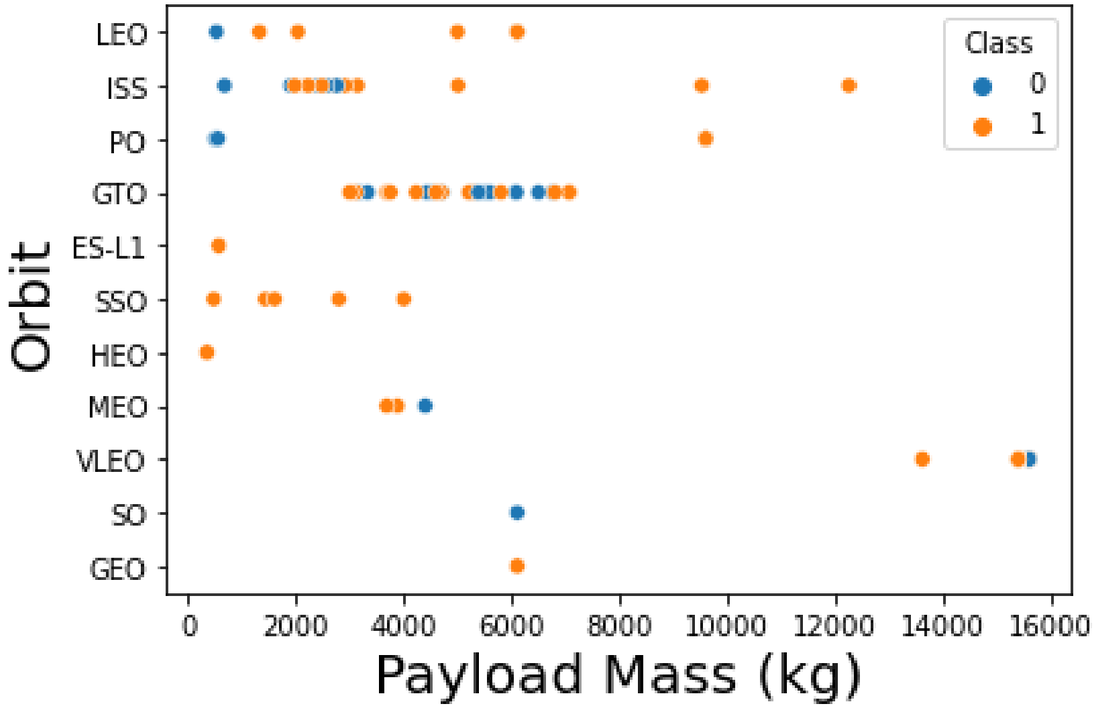
Flight Number vs. Orbit Type



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 numbers when in GTO orbit.

04 EDA

Payload Mass vs. Orbit Type

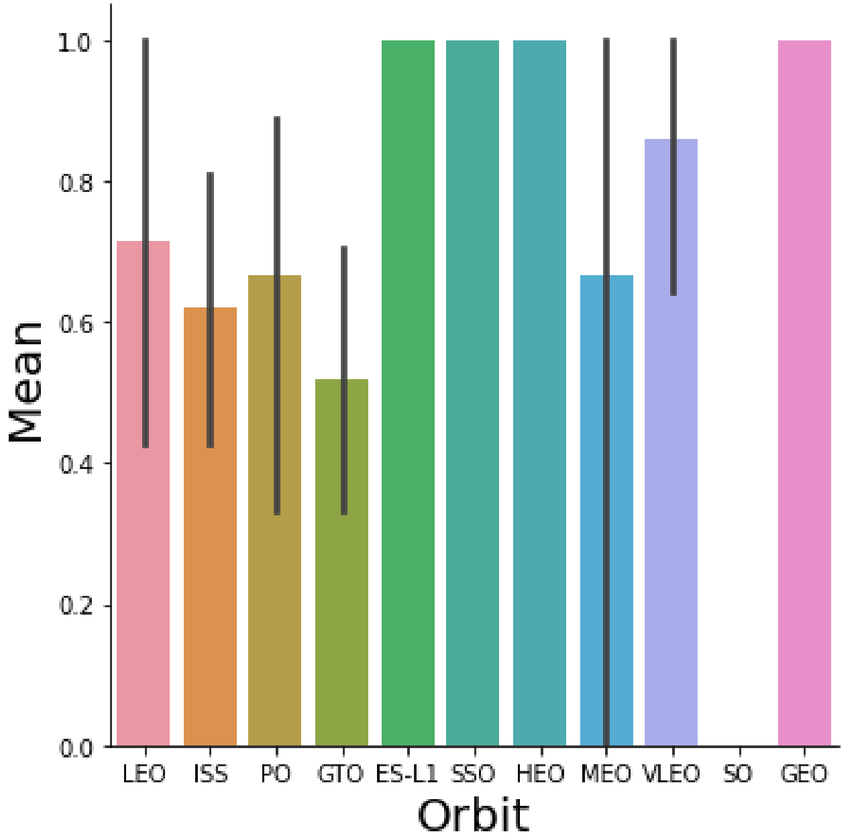


With heavy payloads, the successful landing or positive landing rate is more for Polar, LEO, and ISS.

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

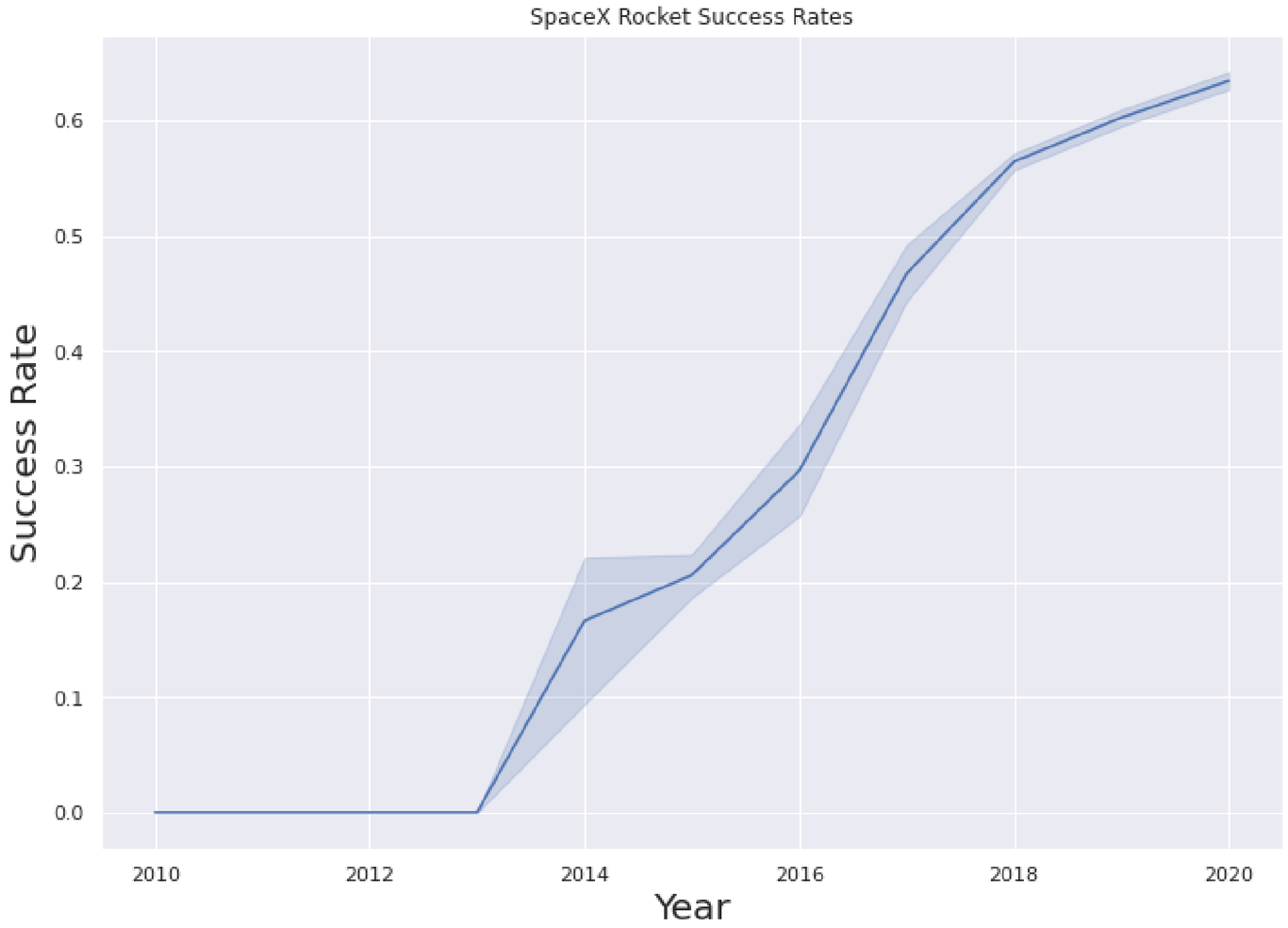
04 EDA

Average Success Rate vs. Orbit Type



On average, ES-L1, SSO, HEO, and GEO have the highest success rate; GTO has the average lowest success rate.

04 EDA



SpaceX Rocket Launch Success Rate by year

The success rate has dramatically increased in 2013, presenting an uprising trend until 2020.

```
select distinct Launch_Site from dbo.Spacex$
```

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

- DISTINCT clause to remove duplicates from the result set. In our data, the duplicates are the launch sites.
- We can see that there are 4 launch sites in this dataset.

1.

Names of the unique launch sites in the space mission

04 EDA-SQL

```
select top 5 * from dbo.Spacex$ where Launch_Site like 'KSC%'
```

2.

5 records where launch sites begin with the string 'KSC'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
2017-02-19 00:00:00.000	1899-12-30 14:39:00.000	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16 00:00:00.000	1899-12-30 06:00:00.000	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30 00:00:00.000	1899-12-30 22:27:00.000	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01 00:00:00.000	1899-12-30 11:15:00.000	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15 00:00:00.000	1899-12-30 23:21:00.000	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- Using WHERE clause to look for results, subject to certain conditions. In this case, we focus on the launches specifically from the launch site where its name contains 'KSC'.

04 EDA-SQL

```
select sum(PAYLOAD_MASS__KG_) as TotalPayloadMass from
dbo.SpaceX$ where Customer = 'NASA (CRS)'
```

	TotalPayloadMass
1	45596

- SUM clause to return the sum of the variable 'Payload mass', specifically for NASA (CRS).
- The total payload mass for NASA (CRS) is 45,596 kgs.

3.

Total payload mass carried by boosters launched by NASA (CRS)

04 EDA-SQL

```
select AVG(PAYLOAD_MASS__KG_) as AveragePayloadMass
from dbo.SpaceX$ where Booster_Version = 'F9 v1.1'
```

4.

Average payload mass carried
by booster version F9 v1.1

	AveragePayloadMass
1	2928.4

- AVG clause to calculate the average of payload mass specifically for Booster version F9 v1.1.
- The average payload for this booster version is 2,928.40 kgs.

```
select MIN(Date) as SuccessLandingOutcome from dbo.Spacex$
where [Landing _Outcome] = 'Success (drone ship)'
```

	SuccessLandingOutcome
1	2016-04-08 00:00:00.000

- MIN clause literally means 'the minimum value' in a set. In this case, we are looking for the 'minimum date' - the earliest day where the landing outcome was succeeded (drone ship).
- The very first successful landing in drone ship was on 8th Apr 2016.

5.

The date where the first successful landing outcome in drone ship was achieved.

```
Select Booster_Version from dbo.Spacex$
where [Landing _Outcome] = 'Success (ground pad)' AND
Payload_MASS__KG_ > 4000 AND Payload_MASS__KG_ <
6000
```

	Booster_Version
1	F9 FT B1032.1
2	F9 B4 B1040.1
3	F9 B4 B1043.1

- Using comparison operators such as '>' or '<' to restrict our results.

6.

The names of the boosters which have success in the ground pad and have payload mass greater than 4000 but less than 6000.

04 EDA-SQL

```
Select
 (Select Count(Mission_Outcome) from dbo.Spacex$ where
Mission_Outcome LIKE '%Success%') as
Successful_Mission_Outcomes,
 (Select Count(Mission_Outcome) from dbo.Spacex$ where
Mission_Outcome LIKE '%Failure%') as
Failure_Mission_Outcomes
```

	Successful_Mission_Outcomes	Failure_Mission_Outcomes
1	100	1

- Using subqueries to count ('Count' clause) the number of successes and failures in landing.
- After grouping the mission outcomes, we know that there was only one failed mission!

7 .

The total number of successful and failure mission outcomes.

04 EDA-SQL

```
Select distinct Booster_Version, MAX(PAYLOAD_MASS__KG_) as [Maximum Payload Mass]
 from dbo.Spacex$ group by Booster_Version order by [Maximum Payload Mass] desc
```

	Booster_Version	Maximum Payload Mass
1	F9 B5 B1048.4	15600
2	F9 B5 B1048.5	15600
3	F9 B5 B1049.4	15600
4	F9 B5 B1049.5	15600
5	F9 B5 B1049.7	15600
6	F9 B5 B1051.3	15600
7	F9 B5 B1051.4	15600
8	F9 B5 B1051.6	15600
9	F9 B5 B1056.4	15600
10	F9 B5 B1058.3	15600
11	F9 B5 B1060.2	15600
12	F9 B5 B1060.3	15600
13	F9 B5 B1049.6	15440
14	F9 B5 B1059.3	15410
15	F9 B5 B1051.5	14932
16	F9 B5 B1049.3	13620

- The GROUP BY clause allows us to combine the booster versions into groups and the DESC clause to order the result in a descending order - the booster version with the heaviest payload ranked at 1st.

8.

The names of the booster_versions which have carried the maximum payload mass.

```
Select
 DateName( month , DateAdd( month ,
MONTH(CONVERT(date,Date, 105)) , 0 ) - 1 )  as Month,
 Booster_Version, Launch_Site, [Landing _Outcome] from
dbo.Spacex$
 where  ([Landing _Outcome] = 'Success (ground pad)') and
YEAR(CONVERT(date,Date, 105)) = '2017'
```

	Month	Booster_Version	Launch_Site	Landing _Outcome
1	February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
2	May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
3	June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
4	August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
5	September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
6	December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

- CONVERT clause to convert the date and time variables into a more readable and more consistent format. Filtering the result set with date/time variables.
- There were 6 successful landings in the ground pad during the year 2017.

9.

The records which will display the month names, successful landing_outcomes in the ground pad, booster versions, launch_site for the months in the year 2017.

04 EDA-SQL

10.

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

```
Select Date, Booster_Version, Launch_site, [Landing
_Outcome] from dbo.Spacex$
 where [Landing _Outcome] like '%Success%' and Date between
'2010-06-04' and '2017-03-20' order by Date desc
```

	Date	Booster_Version	Launch_site	Landing _Outcome
1	2017-02-19 00:00:00.000	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
2	2017-01-14 00:00:00.000	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
3	2016-08-14 00:00:00.000	F9 FT B1026	CCAFS LC-40	Success (drone ship)
4	2016-07-18 00:00:00.000	F9 FT B1025.1	CCAFS LC-40	Success (ground pad)
5	2016-05-27 00:00:00.000	F9 FT B1023.1	CCAFS LC-40	Success (drone ship)
6	2016-05-06 00:00:00.000	F9 FT B1022	CCAFS LC-40	Success (drone ship)
7	2016-04-08 00:00:00.000	F9 FT B1021.1	CCAFS LC-40	Success (drone ship)
8	2015-12-22 00:00:00.000	F9 FT B1019	CCAFS LC-40	Success (ground pad)

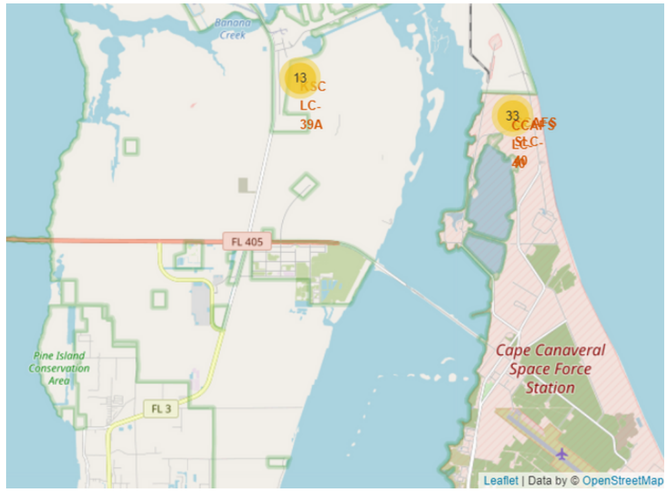
- Within this time period, there were 8 successful landings, which only 1 in 2015, 2 in 2017, and the rest was in 2016. This somehow indicates that there were no successful landings before the end of the year 2015.

LAUNCH SITES

All launch sites on a global map

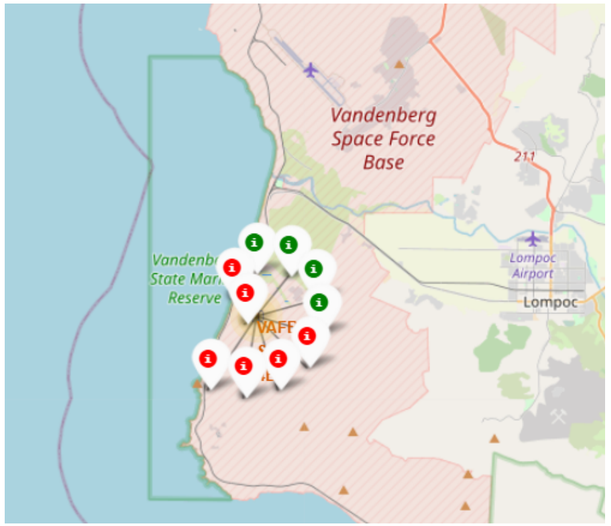
We can see that the launch sites are all located very near to the coastlines. VAFB SLC-4E is the only site located on the west, while the other 3 sites are all on the east.

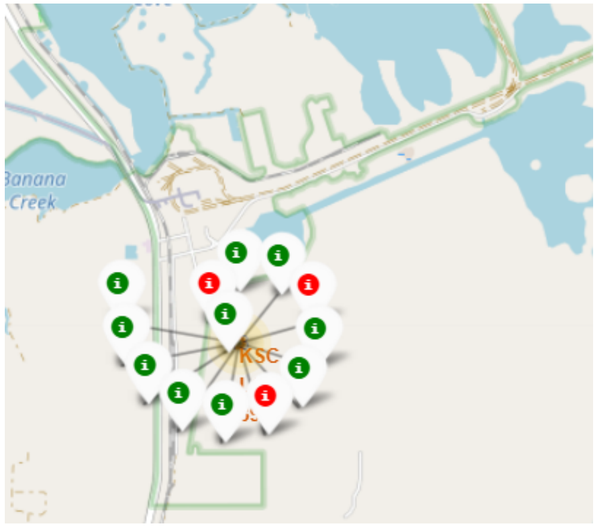
04 MAP



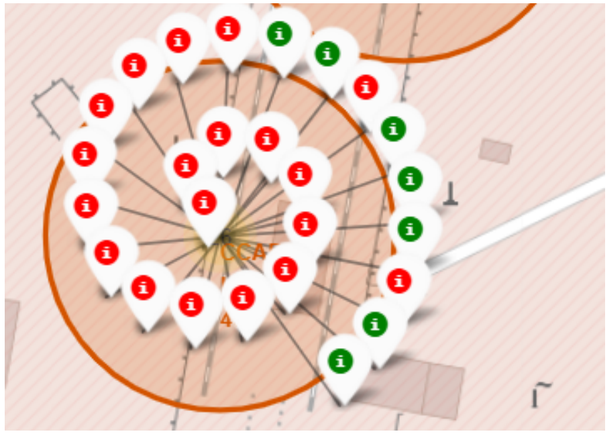
LAUNCH OUTCOMES

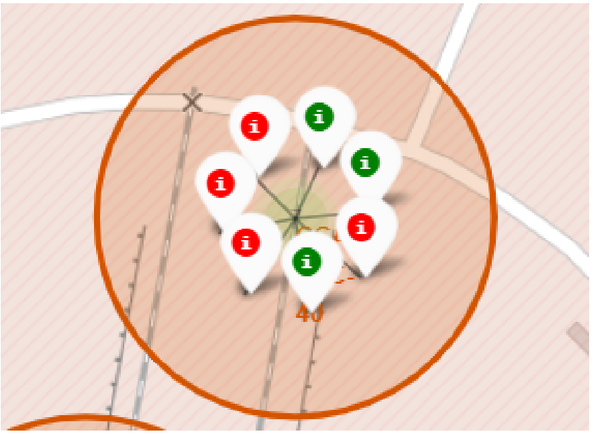
Launch outcomes in each sites





By looking at the color labels, we know that KSC LC-39A has the highest success rate among others (10 out of 13 launches); while CCAFS LC-40 has the lowest success rate, but the most launches (26).

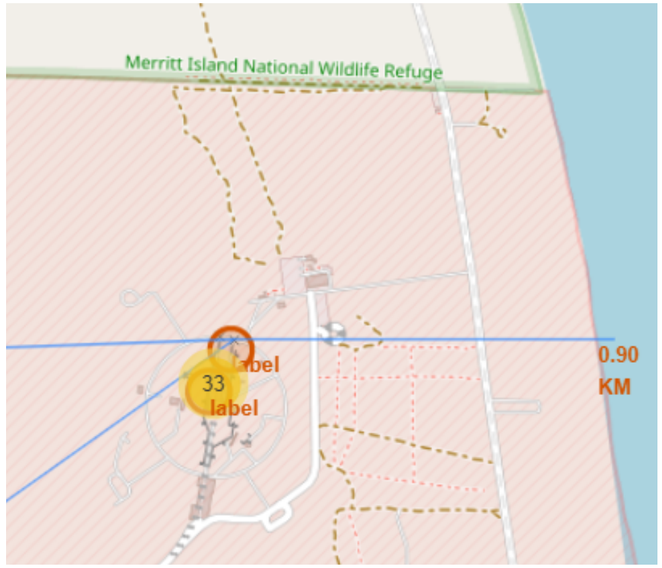


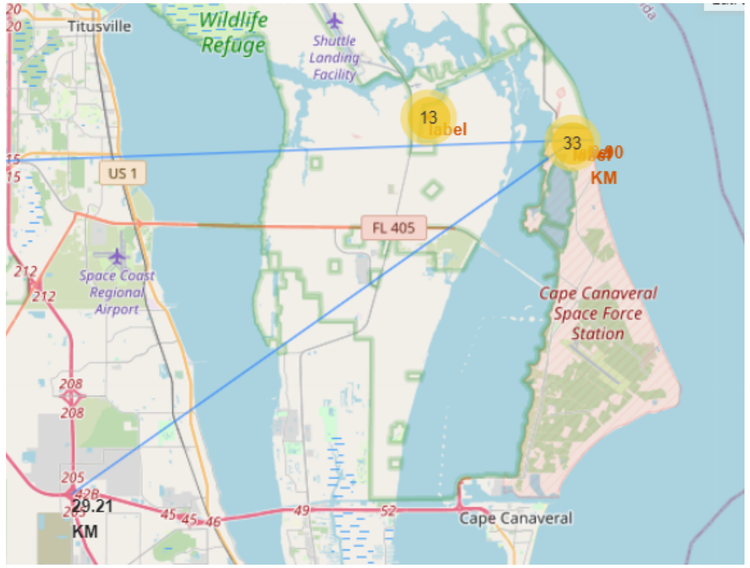


04 MAP

PROXIMITIES OF LAUNCH SITES

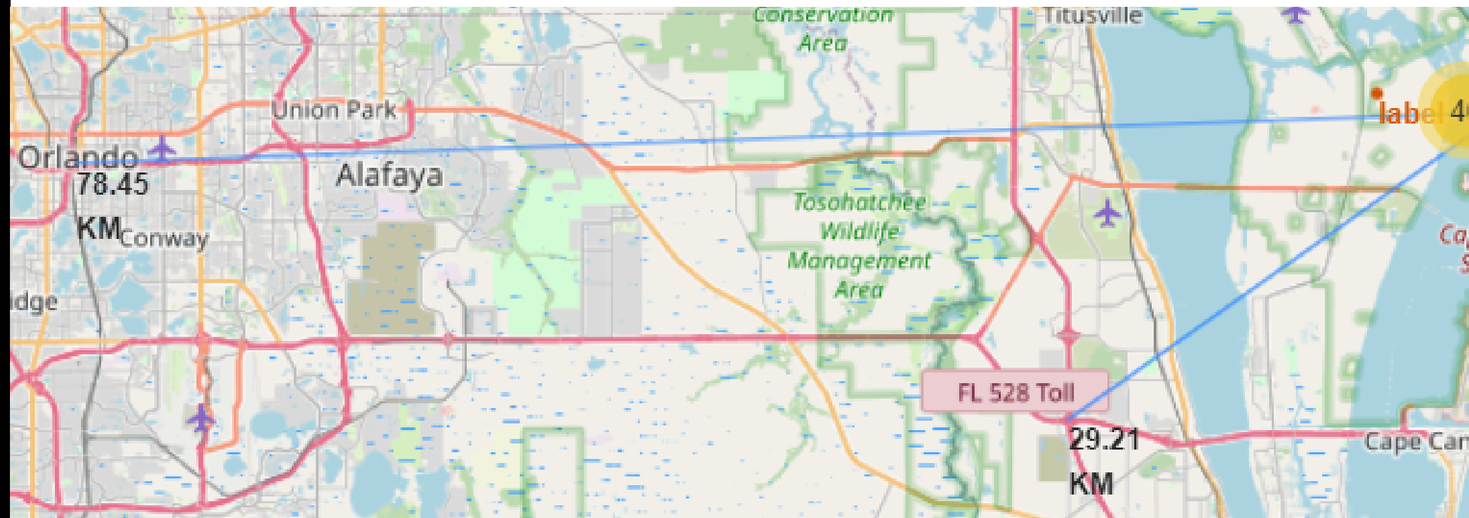
Where do launch sites usually locate?





The launch sites have these in common:

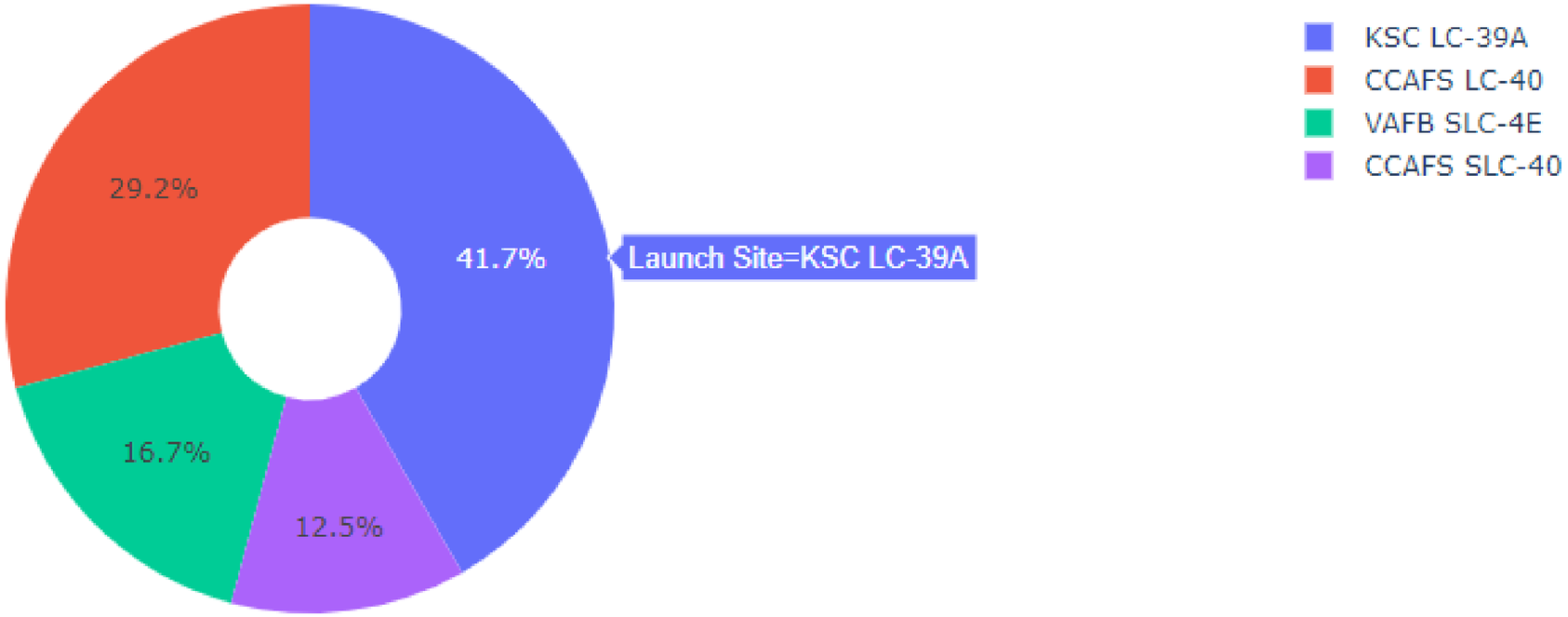
- very close to coastlines;
- far from main cities and highways.



04 MAP

04 PLOTLY DASH

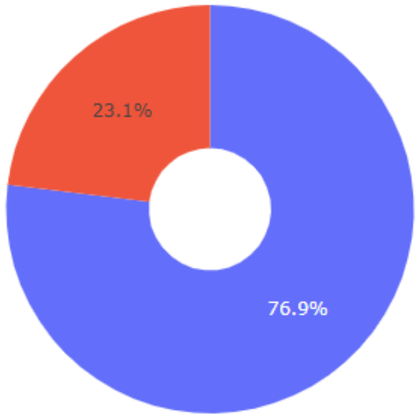
Launch Success Count for All Sites



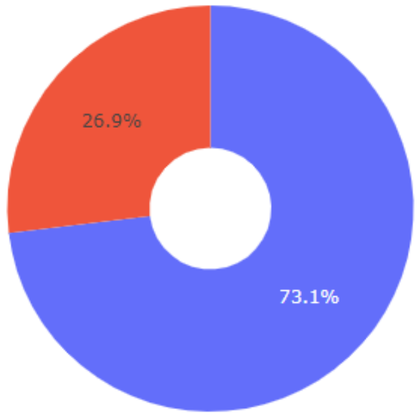
KSC LC-39A has the highest success rate in rocket launches, whereas CCAFS SLC-40 has the lowest.

04 PLOTLY DASH

Total Success Launches for site KSC LC-39A



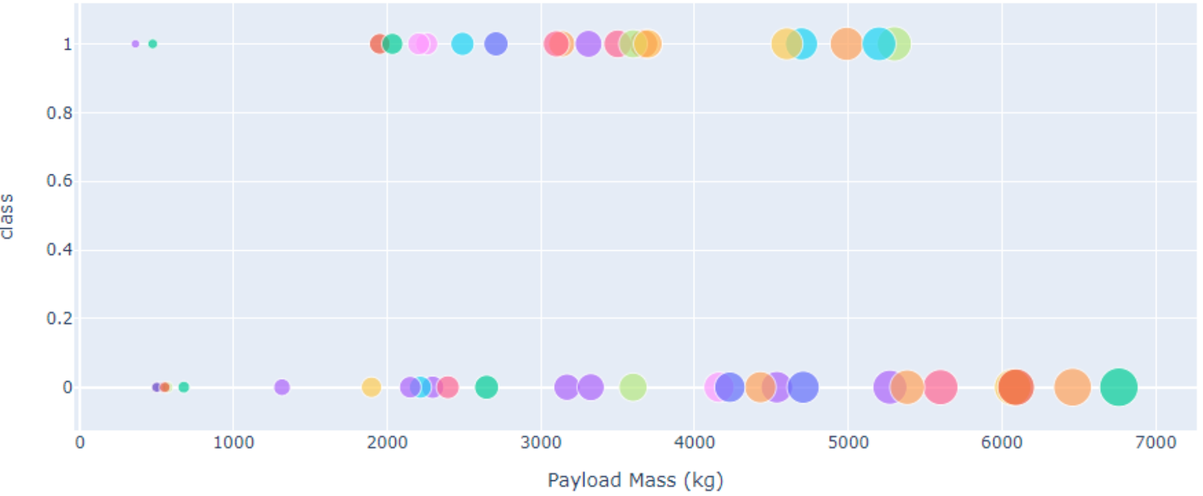
Total Success Launches for site CCAFS LC-40

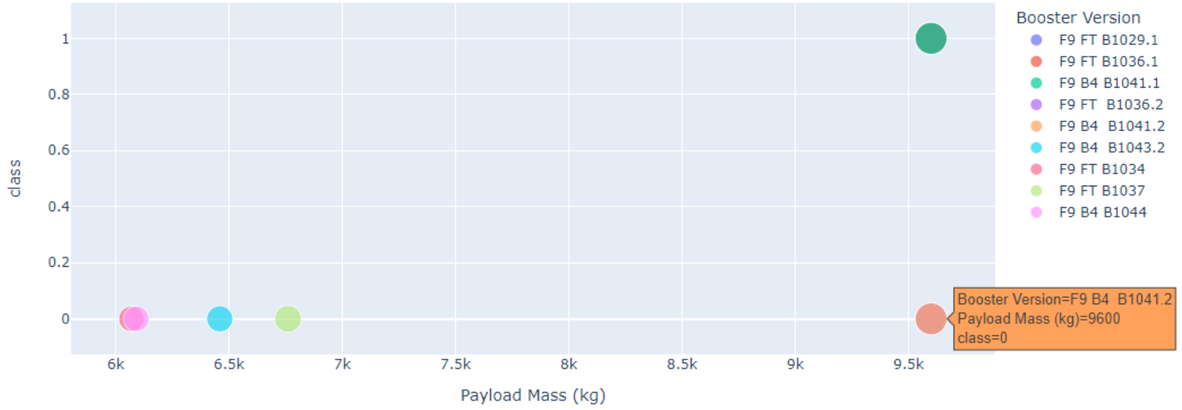


Launch Site with Highest Success Ratio

Among all the sites, KSC LC-39A, again, has the highest success ratio of rocket launches (around 77%). Additionally, CCAFS LC-40 which had the least launches compared to other sites had in fact, a 73% of success ratio, ranked 2nd place over all the sites.

04 PLOTLY DASH





Payload Mass vs. Launch Outcome

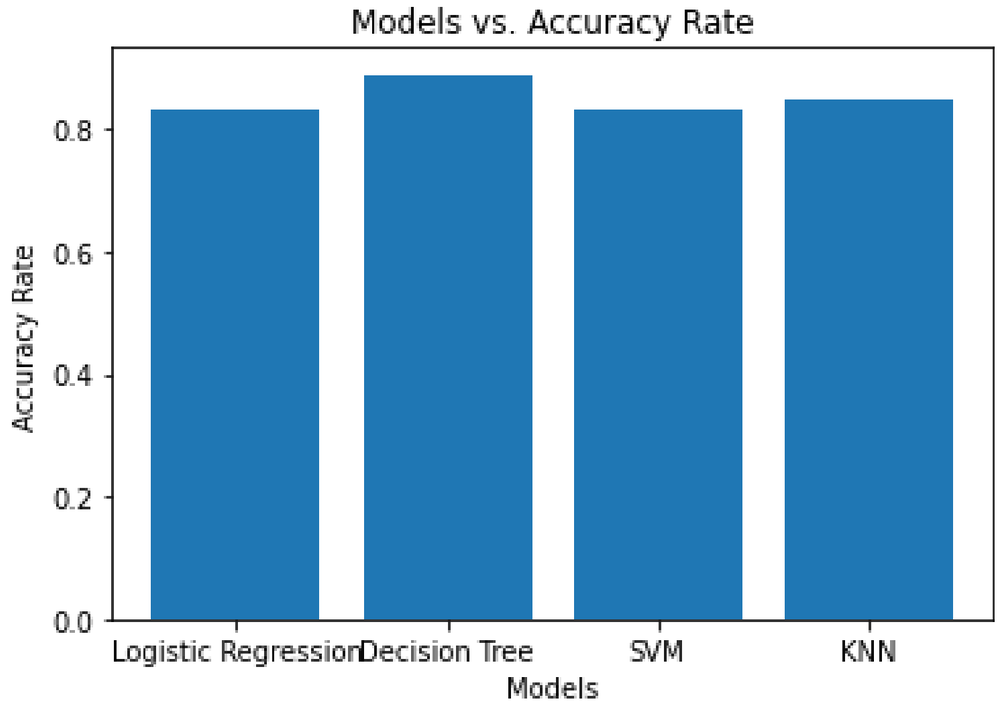
Most of the successful launches had a payload between 2,000 - 6,000 kgs. The success rate dropped significantly with relatively light-weighted (< 1,900 kgs) and heavily weighed launches (> 6,000 kgs).

04 PREDICTIVE ANALYSIS

Classification Accuracy

Among the models we predicted, the Decision Tree Classifier model ('Tree') has the highest accuracy rate (88.93%) -- which is selected as the best fit model.

Followed by the K Nearest Neighbors model (KNN) with around 84% accuracy rate, whereas the Logistic Regression model and Support Vector Machine (SVM) model have the lowest (83%).

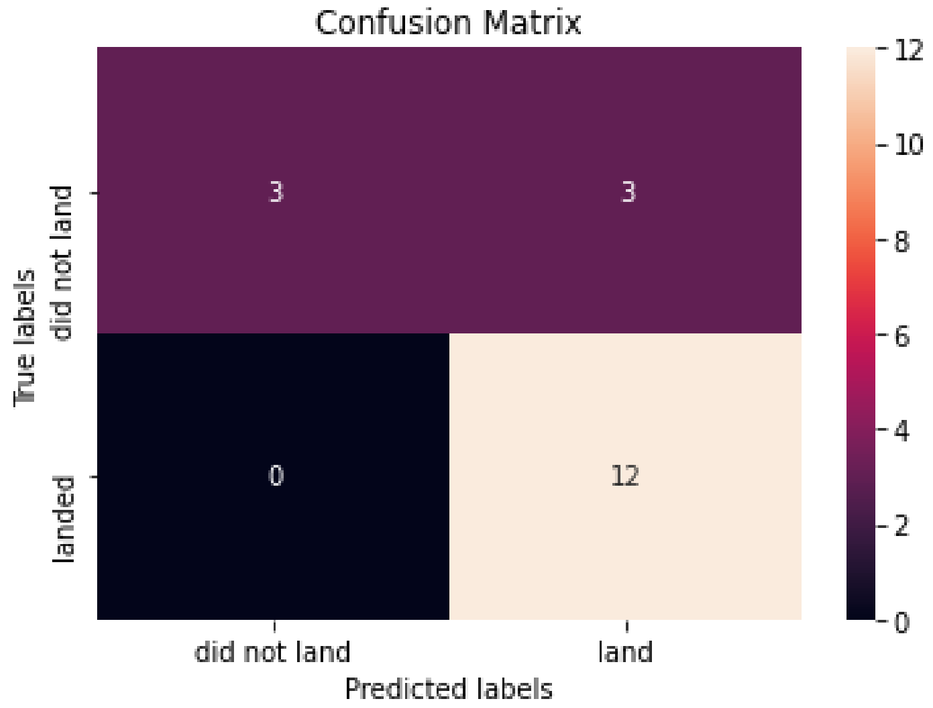



```
Best Method is Tree with a score of 0.8892857142857142
Best Params is : {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_spli
t': 5, 'splitter': 'random'}
```

04 PREDICTIVE ANALYSIS

Confusion Matrix

Looking at the Confusion matrix for the Tree model, we can see that most predictions fit with the true labels (12), whereas 3 are the 'false positive' cases.



05

CONCLUSION

1. Launches with payload between 2,000 - 6,000 kgs have a higher probability to succeed;
2. Orbit ES-L1, SSO, HEO and GEO have the highest success rate;
3. KSC LC-39A performs the best among all launch sites;
4. The Decision Tree Classifier model best fits with our research purpose for prediction.

A. APPENDIX

1. Project guidance refer to IBM Data Science and Machine Learning Capstone Project materials (edX.org).
2. Github repository: Yan-DSMLProject.
3. SpaceX Falcon 9 rocket images from NASA: NASA official website
4. SQL section workaround reference: Use VS Code to connect SQL.
5. Presentation slides created using Canva.