IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Haochen Li
04/05/2023

Outline

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

Executive Summary

- To prepare the data, we used API and web scraping to extract the data, and performed data wrangling such as replacing nulls or adding new columns.
- In our initial data exploration analysis, we found that there were 4 different launch sites, most of the payload mass was below 8000 kg, and launch success increased with increasing number of years and experience.
- Using Folium data visualization, we explored the geography of different launch sites. All of the launch sites are close to the coastline and are away from the cities. KSC LC39A have the most amount of successful launches compared to other sites.
- KSC LC39A having the highest success rate is corroborated through our pie chart using Plotly. In addition, we identified that Booster version FT has had the highest success rates and that most successful launches tends to be between 2k – 4k kg payload mass.
- Finally we used predictive analysis to test the accuracy of launches and found that Decision Tree has the highest accuracy of 87.7%. The main issue with the models is due to false positives.

Introduction

In this capstone project, we are working as data scientists for Space Y who would like to compete with Space X. Space X is one of the most successful company of the commercial rocket space and this is largely due to their success in reusing the first stage of rocket launches. As a result of reduced costs, their Falcon 9 rocket costs 62 million dollars in comparison to other providers of 165 million dollars.

To understand what makes Space X successful, we will gather information about Space X and uses data science to gain insights. In addition, we will create dashboards to communicate this to our team.

We will determine factors such as price of each launch, predicting successful landings of the first stage, and predict if Space X will re-use the first stage. We will also determine which is the best classification model for this data science project.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected through two different paths:

  1. Using an API (Application Programming Interface). This returns the data in the JSON format. The source of the API is (https://api.spacexdata.com/v4)

  2. Webscraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

- Perform data wrangling
  - The data was converted in to a pandas dataframe.
  - Replacing/dealing with any missing values.
  - A new column 'Class' was created which returns 1 if the first stage landed successfully and 0 if the first stage did not land successfully (one hot encoding).

Methodology

Executive Summary

- 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 standardised and divided into training and testing data using the function train_test_split.
  - We then create a logistic regression object to find the best parameters.
  - We then calculate the accuracy on the test data using the method score.
  - This is repeated for the following classification models: SVM (support vector machine), decision tree, and KNN (K-Nearerst Neighbours).

Data Collection

We obtained two sets of data by using two different sources:

1. API from the Space X Rest API (api.spacexdata.com/v4)
   - The dataset relates to Falcon 9 launches.
   - The following columns were captured: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcomes, Flights, Frid Fins, Reused, Legs, Landing Pad, Block, Serial, Longitude, Latitude.
2. Web Scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
   - The dataset relates to Falcon 9 historical launch records.
   - The following columns were captured: Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster landing, Date, Time.

Data Collection – SpaceX API

API standards for Application Programming Interface and it is how two computers talk to each other. In this case, Space X offers a public API in which we, the third party, can request data from.

Link: Github – 1. Data Collection with API

Select a REST API url. In this case, it was an Open API and therefore we do not need an API key.

↓

Write custom requests to pull the relevant data columns.

↓

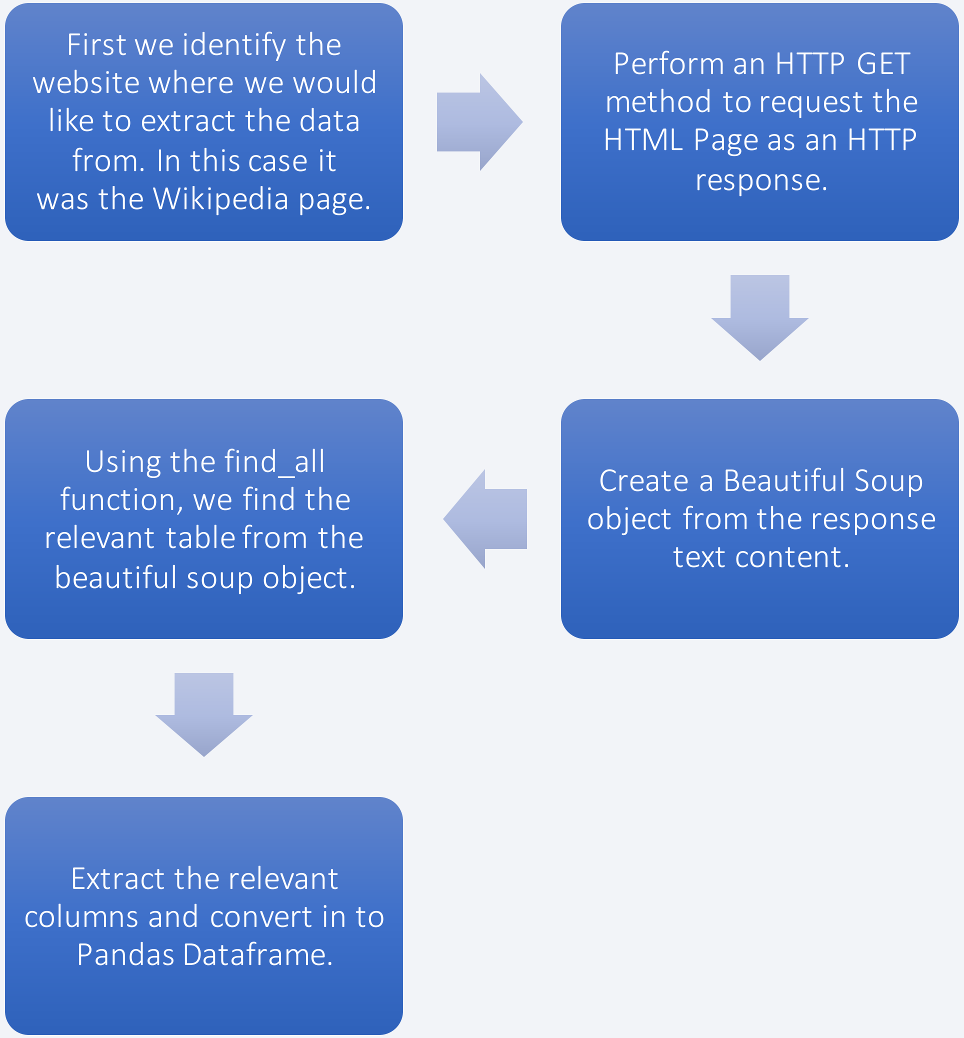
The pulled data is converted in to a pandas data frame.

↓

Finally we filter the dataset to include only Falcon 9 launches and replace any missing values with its mean value.

Data Collection – Web Scraping

Web scraping refers to extracting data from a website. In this case, we used wikipedia to extract Flacon 9 launch information.

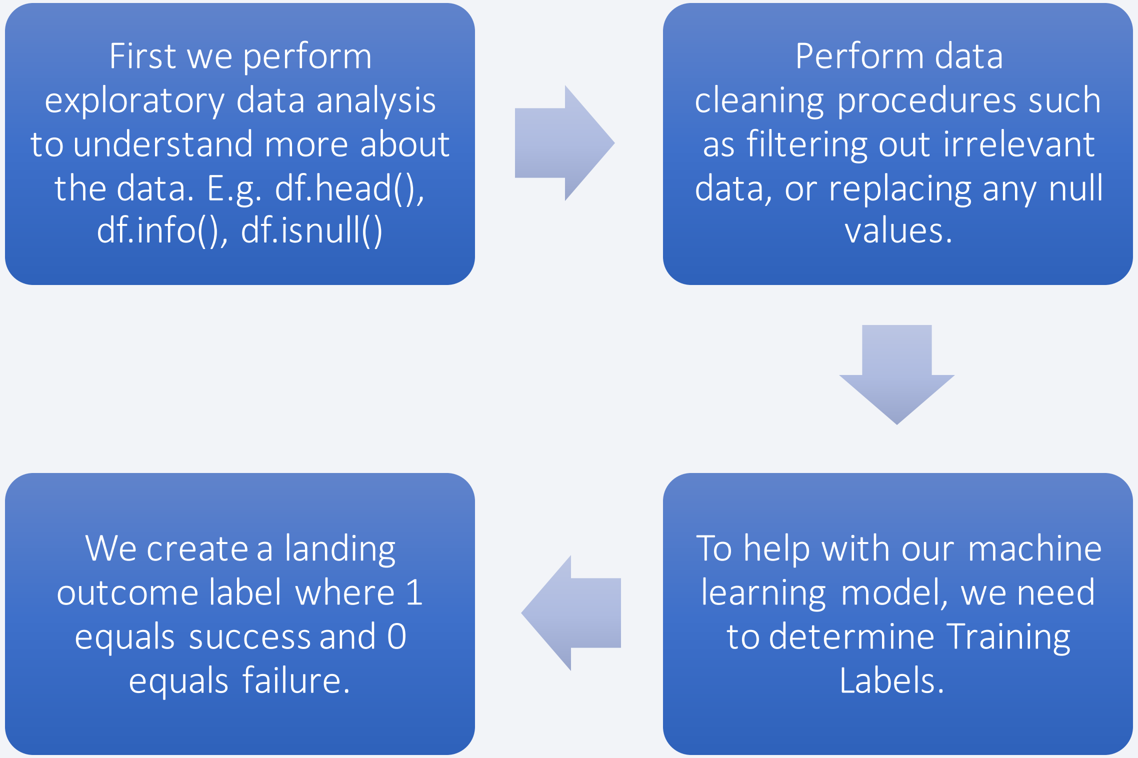


Link: Github – 2. Data collection with web scraping

Data Wrangling

With the datasets we have collected, we will look at it and perform some data wrangling activities. Data wrangling refers to transforming the data from its raw format in to a readily and usable format.

Link: Github – 3. Data Wrangling



EDA with Data Visualization

Scatter graphs were used to identify the relationship between the following:

- How flight number and payload variables would affect the launch outcome
- How flight number and launch site affect the launch outcome
- How flight number and orbit affect the launch outcome
- How payload mass and orbit affect the launch outcome

Bar charts were used to identify the relationship between continuous variable success rate and categorical variable orbit type.

Line charts were used to identify the relationship between the success rate and the years.

Link: Github – 4. EDA Data Visualisation

EDA with SQL

- Using select distinct query to find the unique names of launch sites
- Using the LIKE wild card statement to display 5 records where launch sites begin with the string 'CCA'
- Using the SUM statement to display the total payload mass
- Using the AVG statement to display the average payload mass carried by a specific booster version
- Using the MIN(Date) statement to display the date when the first successful landing outcome in groud pad was achieved
- Using numerous WHERE statements to list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Using the COUNT and GROUP BY statements to list the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass using a subquery.
- Using the SUBSTR query to display month names where the landing outcome is failure (drone ship) and the year is 2015.
- Ranking the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Link: Github – 4.1 EDA with SQL

Build an Interactive Map with Folium

- Circles were added to highlight a circled area with a pop up text label.
- Markers were added to indicate certain points such as launch sites. Markers were also used to indicate whether a launch was successful (green) or if it was a failure (red).
- MousePosition was added to get coordinate for a mouse over a point on the map.
- PolyLine was added to draw a line between a launch site to the selected coastline point.

Link: Github – 5. Launch Site Location Analysis with Folium

Build a Dashboard with Plotly Dash

A pie chart showing the percentage of successful launches by sites were added to quickly profile the launches and helps us to identify which sites had the most/least success.

A Payload range slider was also added to identify if the variable is correlated to mission outcomes.

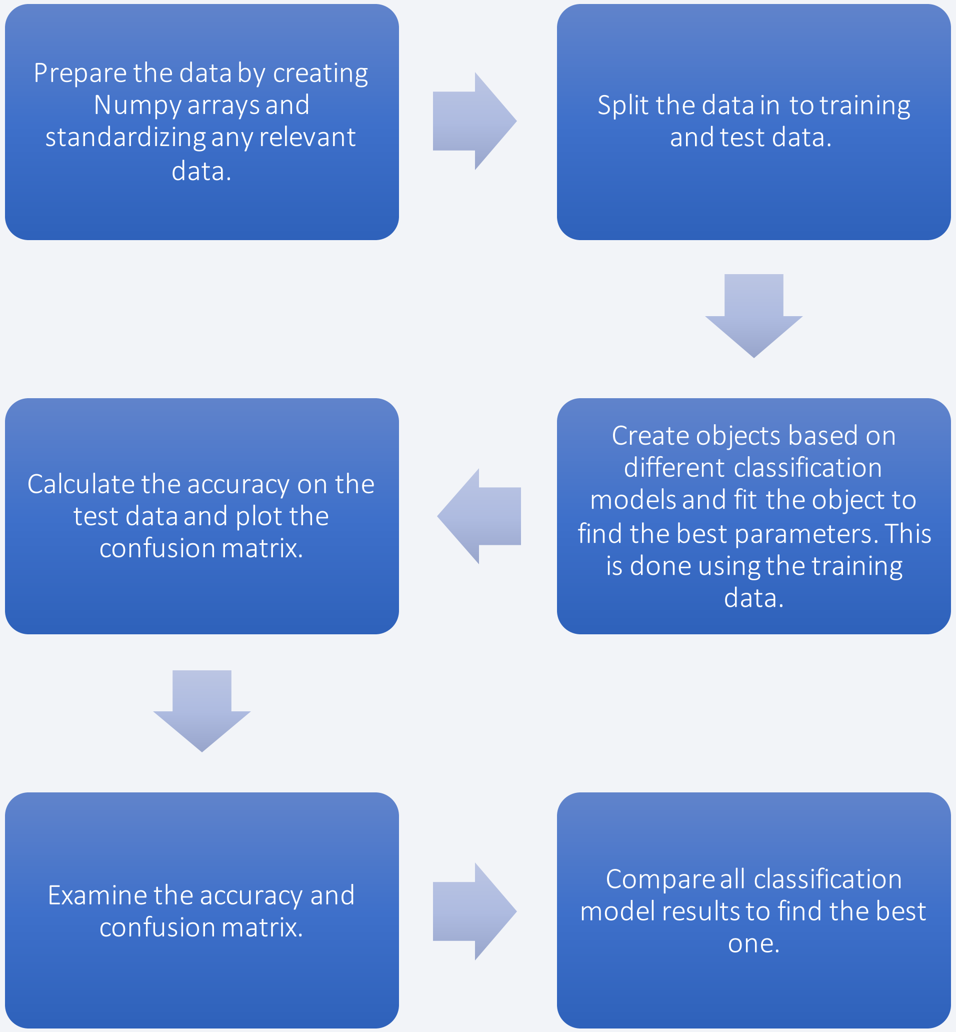
A scatter graph was added to identify correlation between payload and launch outcome. As such, we can observe how payload may be correlated with mission outcomes for selected sites.

Link: Github – spacex_dash_app.py

Predictive Analysis (Classification)

Four types of classification models were used: logistic regression, decision tree, K-Nearest Neighbours, and support vector machine. For each we tested the accuracy and looked at the confusion matrix to find the method that performed the best.

<u>Link: Github – 6. Machine Learning Predictions</u>



Results

Exploratory data analysis results

- CCAFS SLC-40 had the most amount of flights
- CCAFS LC-40 has a success rate of 60% while KSC LC-39A and VAFB SLC 4E has a success rate of 77%
- ES-L1, GEO, HEO, SSO have the highest success rates at 100%
- Success rate since 2013 have been increasing, with the first successful landing in 2015.

Results

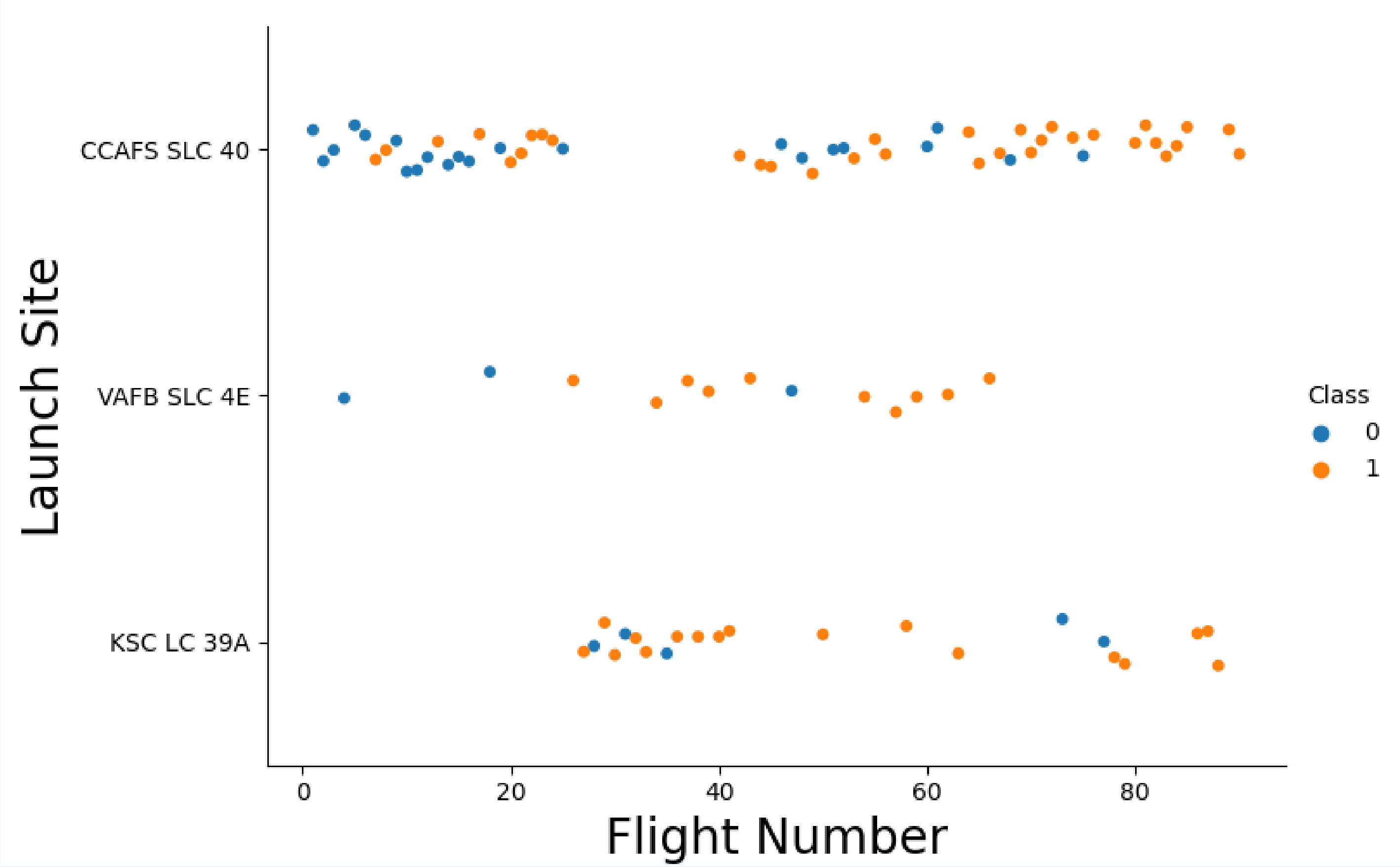
Predictive analysis results

Model	Accuracy	Test Accuracy
Logistic Regression	84.64%	83.33%
SVM	84.82%	83.33%
Decision Tree	88.75%	83.33%
KNN	84.82%	83.33%

Section 2

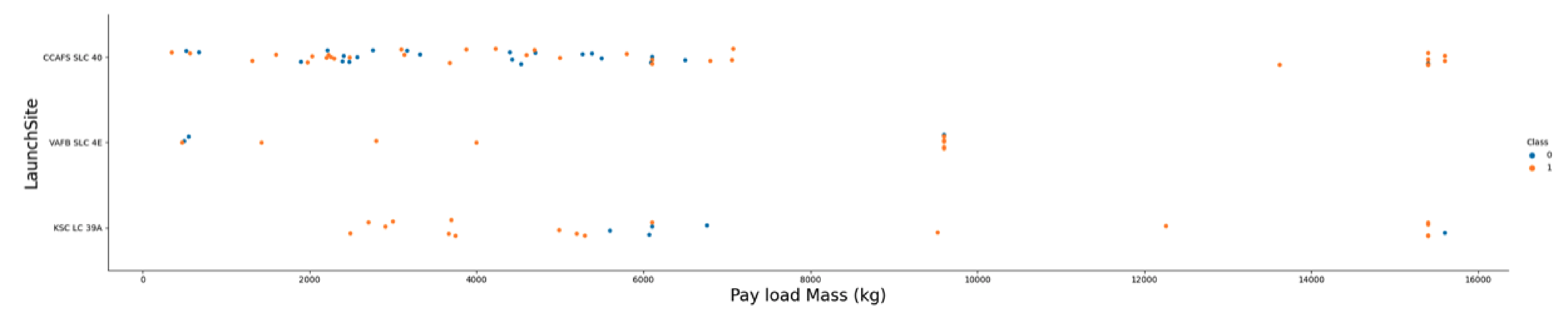
Insights drawn from EDA

Flight Number vs. Launch Site



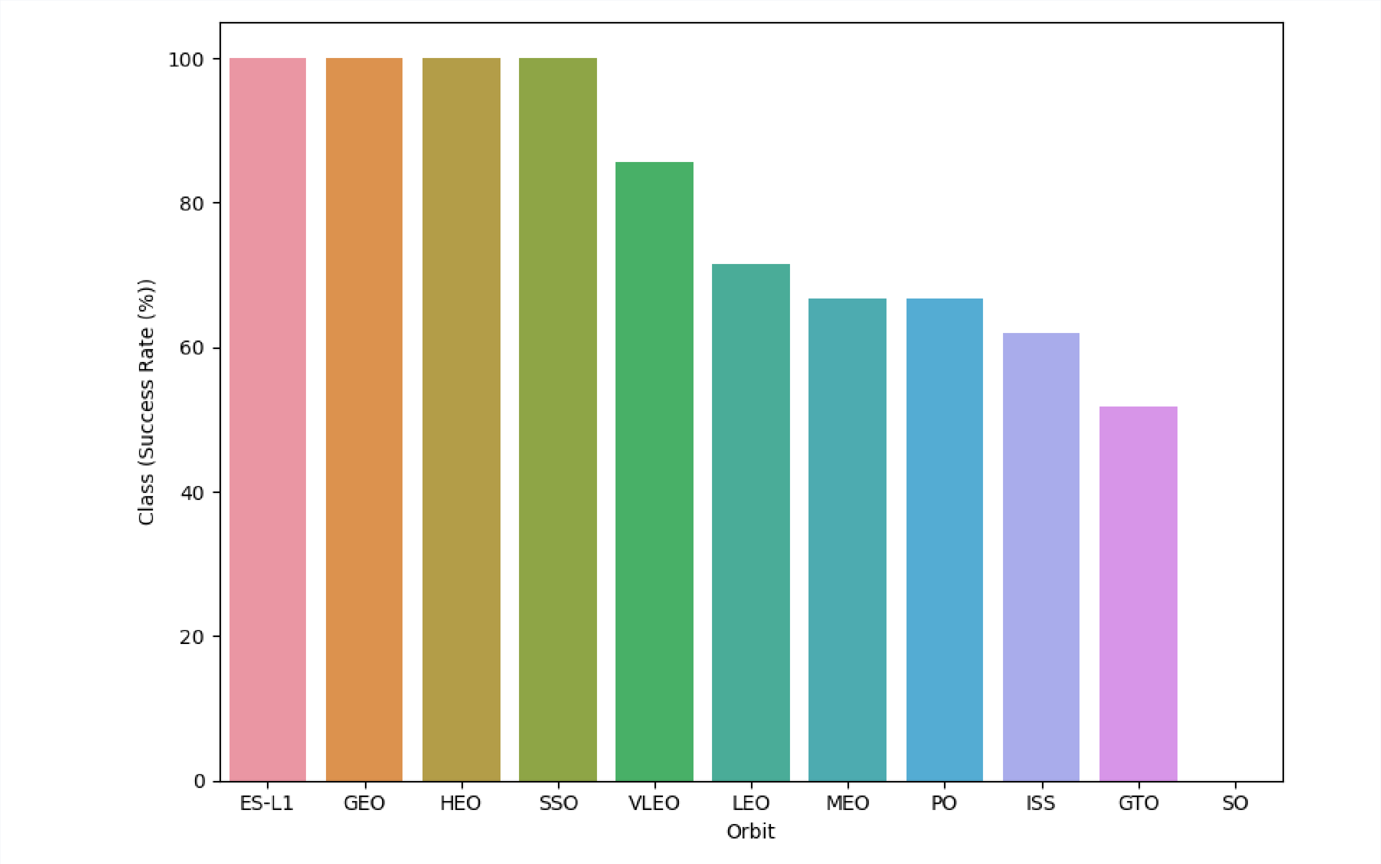
- CCAFS SLC40 has had the most launches
- The success rate increases with increased number of launches
- Overall it looks like KSC LC 39A has the highest percentage of success rate

Payload vs. Launch Site



- Most of the launches have a payload mass of below 8000 kg.
- The highest payload mass is around 15,000 kg for CCAFS SLC 40 and KSC LC 39A
- The success rate between launch site and payload mass is unclear.

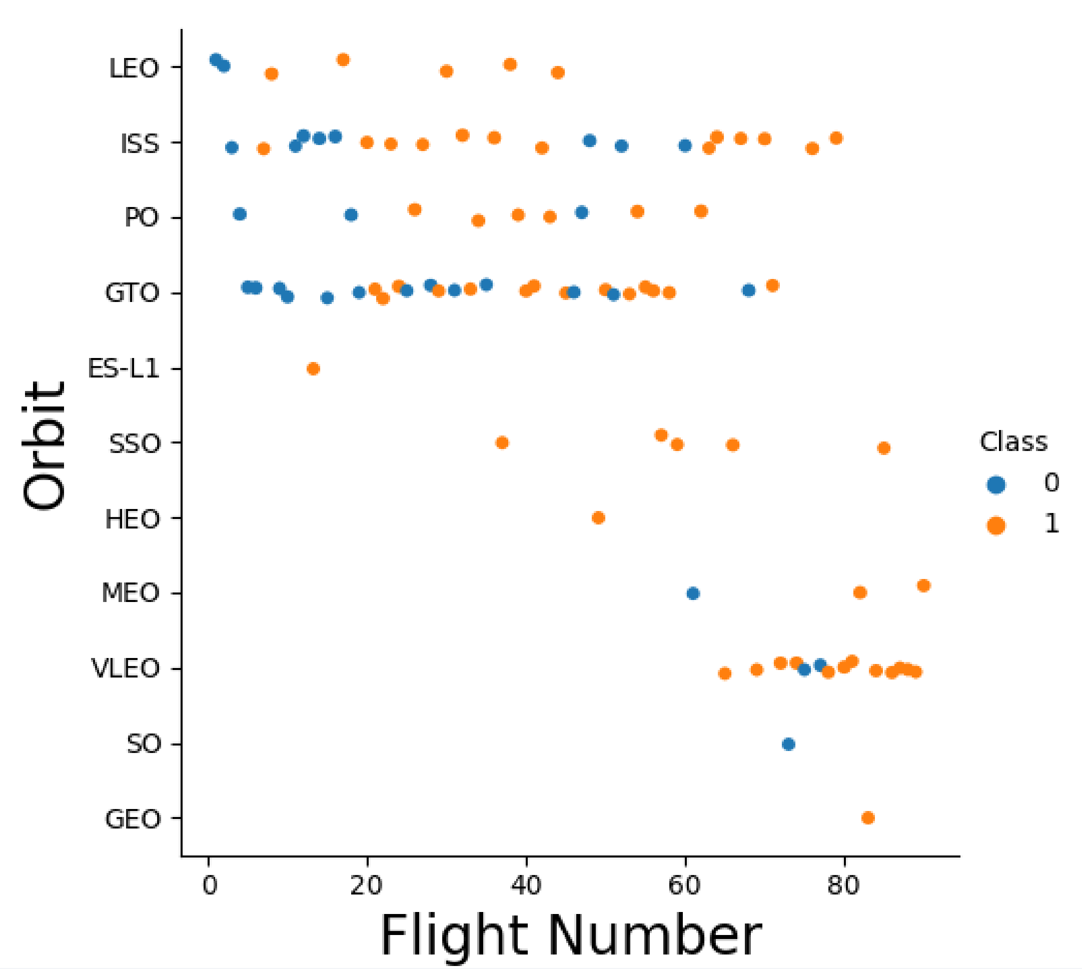
Success Rate vs. Orbit Type



- ES-L1, GEO, HEO, and SSO all have 100% success rate.
- GTO has the lowest success rate at around 58%

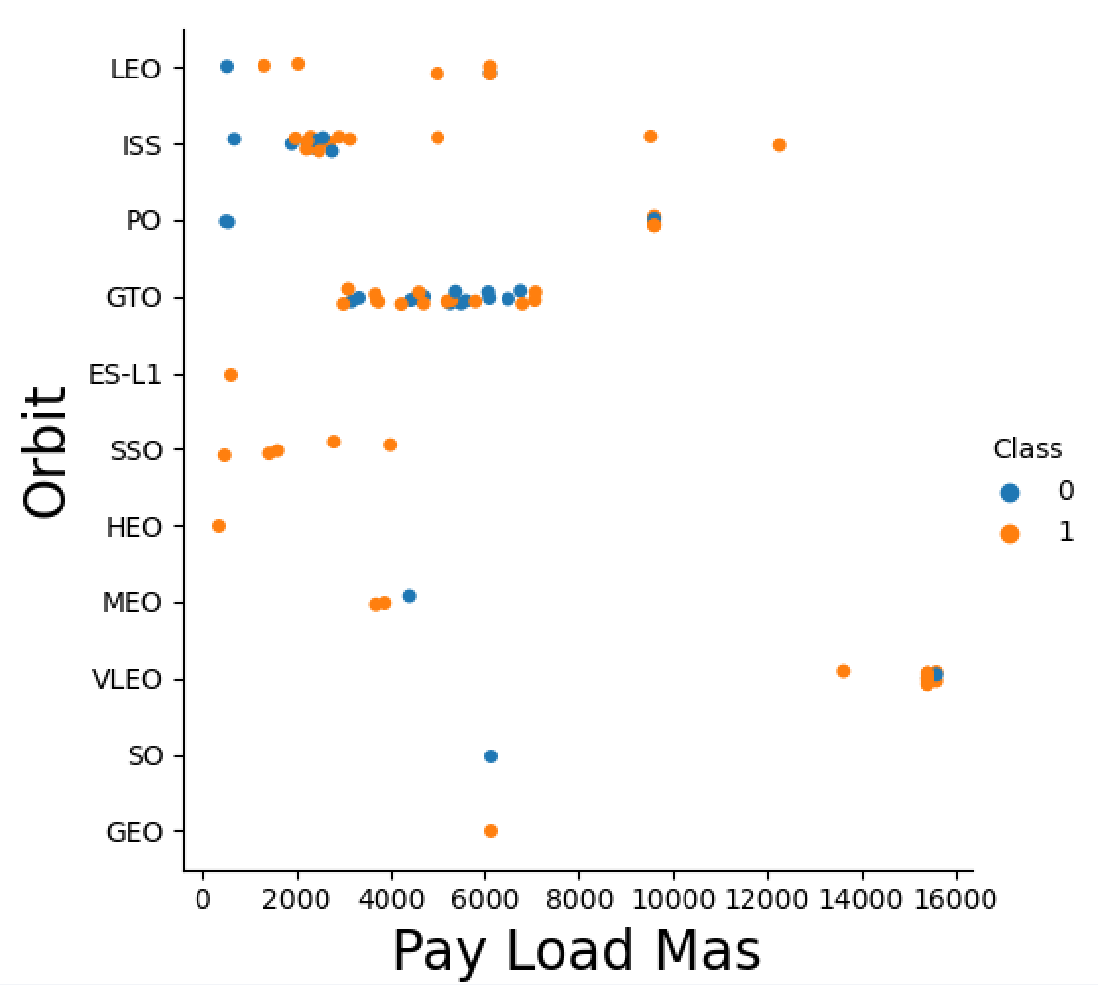
Flight Number vs. Orbit Type

- Most of the launches have been done through LEO, ISS, PO and GTO orbit type.
- Later launches/flights have used SSO, MEO, VLEO, SO, and GEO. The frequency of launches on these were much less.
- Unclear whether the success rate is due to obit type or learning from previous launches



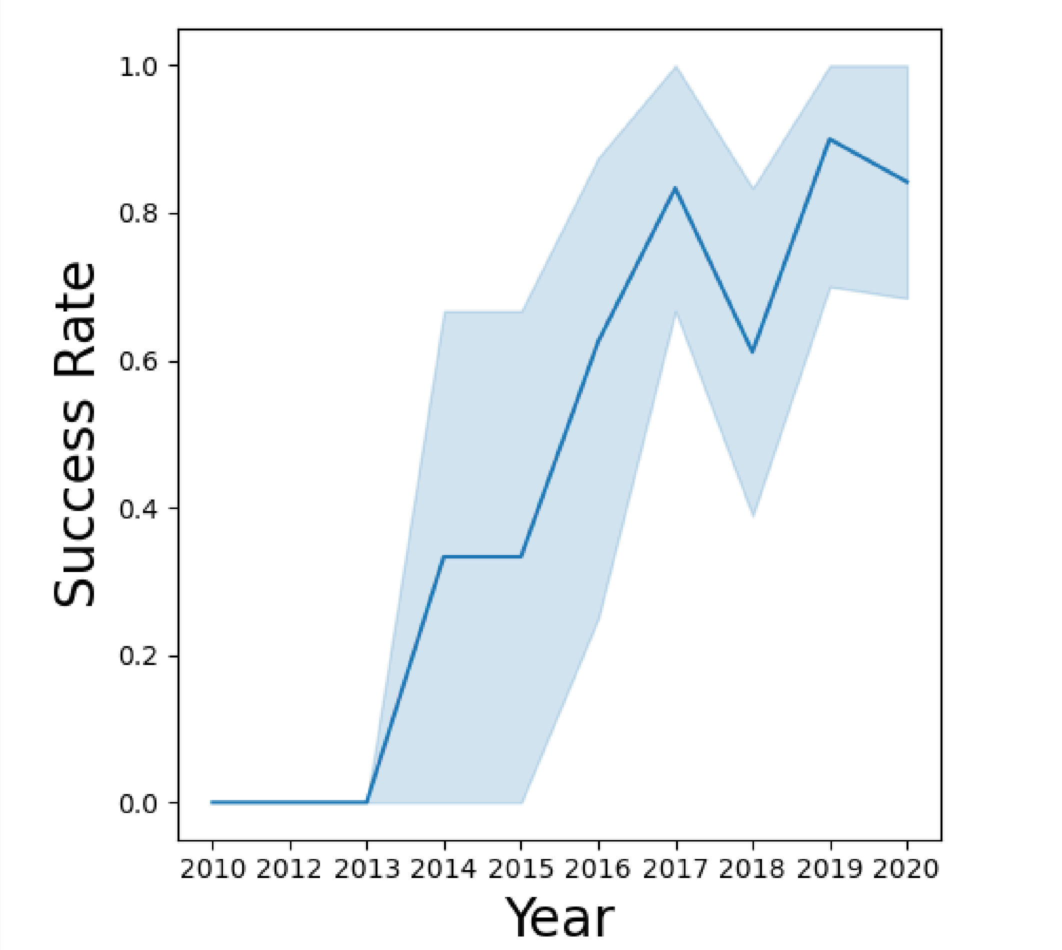
Payload vs. Orbit Type

- It seems that most of the launches were below 8000 kg.
- Heavier payload mass launches were used for VLEO orbit.
- The correlation between orbit type and payload mass and success rate is unclear.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020.
- This suggests that SpaceX have been learning from the previous launches.
- There was a slight dip in 2017 – reasons unclear.



All Launch Site Names

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- There are 4 unique launch sites for Space X's Falcon 9.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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 (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- There are 5 different records with site names beginning with CCA.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- This is the total Payload Mass by NASA (CRS)

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE '%F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- This is to be expected as the payload mass is usually below 8000kg and are on the lighter side.

First Successful Ground Landing Date

```
%sql SELECT min(DATE) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success (ground pad)%' limit 5
```

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

min(DATE)
01-05-2017

- The first success for ground pad was on 1st May 2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
CT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

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

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

- There were 4 unique booster with a success in drone ship and have payload mass of between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(MISSION_OUTCOME), MISSION_OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

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

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

Boosters Carried Maximum Payload

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

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTBL ORDER BY PAYLOAD_MASS__KG_ DESC LIMIT 1
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600

- The booster with the heaviest payload mass of 15,600 kg is F9 B5 B1048.4

2015 Launch Records

```
%sql select SUBSTR(DATE, 4, 2) AS MONTH, "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL WHERE SUBSTR(DATE, 7, 4) = '2015' AND "Landin
```

```
 * 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

- To find the month we used substring function to extract it from the date.
- In total there are two launches with the specified requirement.

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

```
LECT COUNT("Landing _Outcome"), "Landing _Outcome" FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY "Landing _Outcome" ORDER BY COUNT("Landing _Outcome") DES(
```

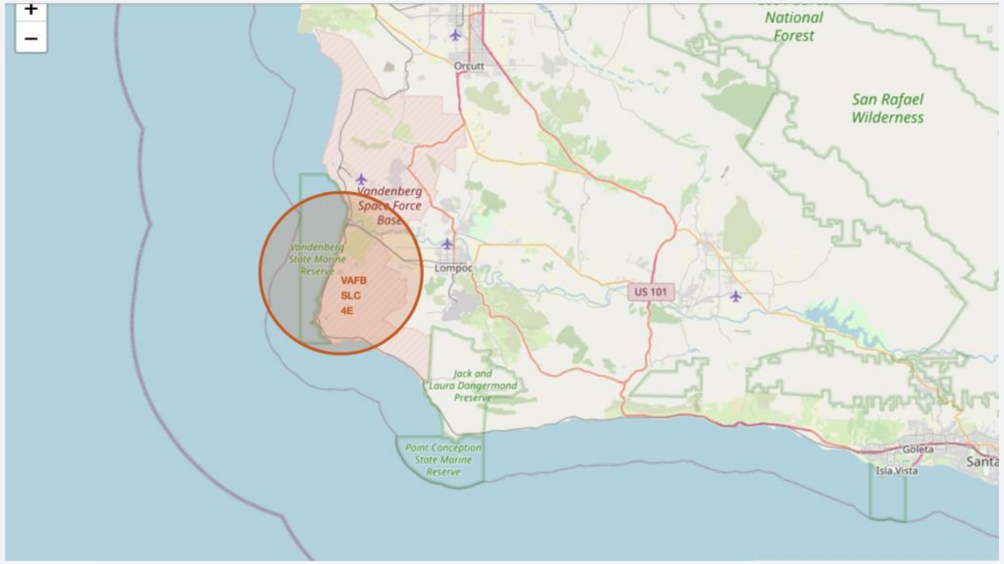
```
 * sqlite:///my_data1.db
Done.
```

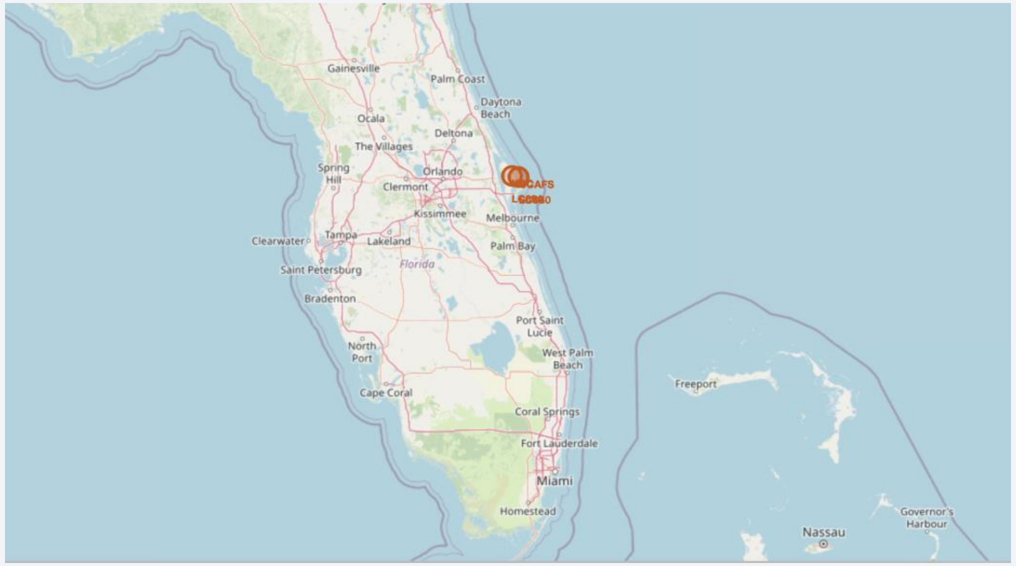
COUNT("Landing _Outcome")	Landing _Outcome
20	Success
10	No attempt
8	Success (drone ship)
6	Success (ground pad)
4	Failure (drone ship)
3	Failure
3	Controlled (ocean)
2	Failure (parachute)
1	No attempt

Section 3

Launch Sites Proximities Analysis

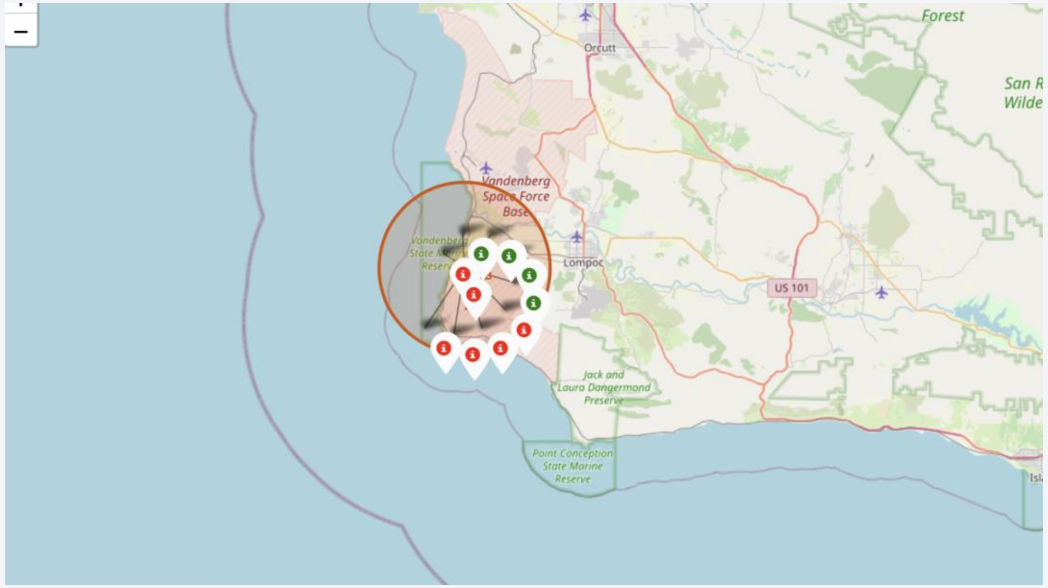
Exploring Launch Sites with Folium

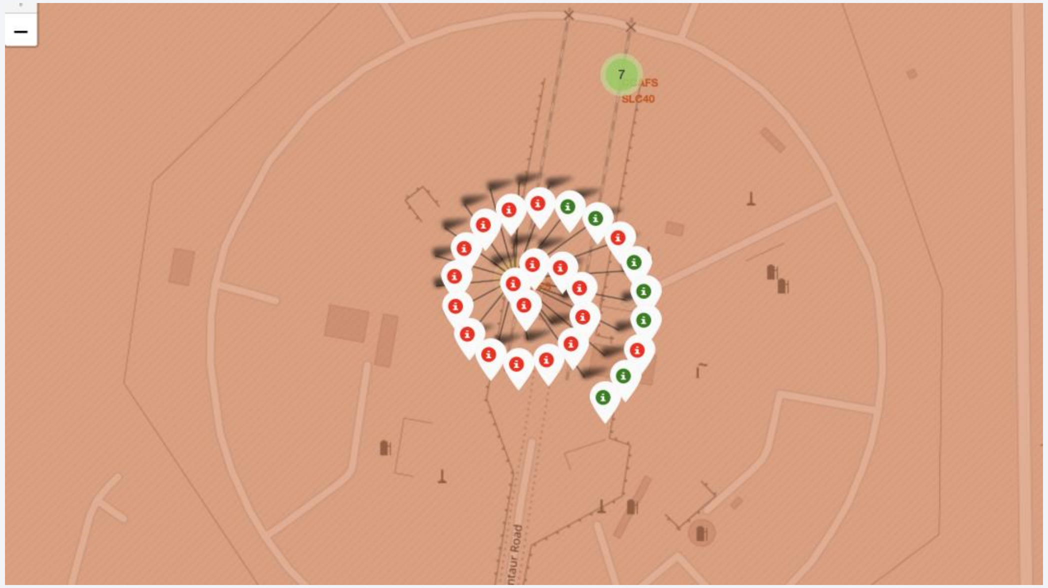




- Using the launch site coordinates and folium circles and folium markers, we have identified the different launch sites.
- All of the launch sites are very close to the sea and they are also close to airports.
- Aside from one of the launch sites, the other 3 are next to each other.

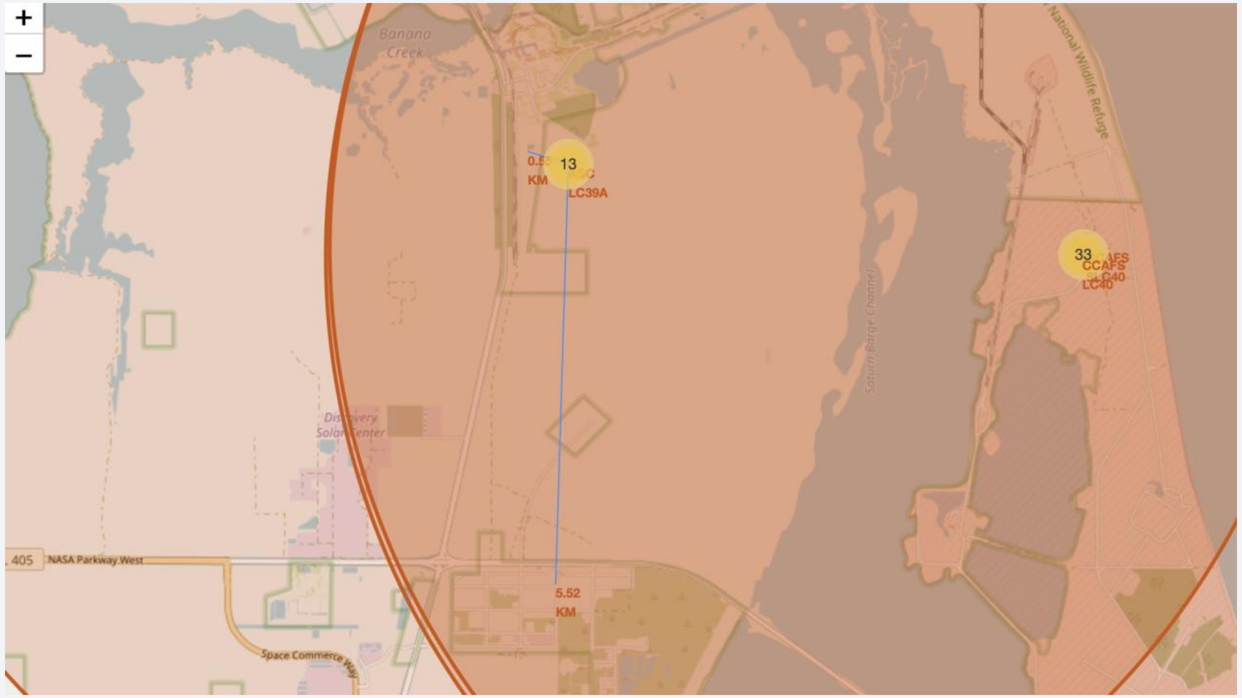
Exploring Launch Site Success/Failure





- The visual was enhanced by color coding successful launches as green and unsuccessful launches as red.
- This way we can identify which launch site had more/less launches, and which launch site has been more successful.
- For example, KSC LC39A seems to have had more successes and location of where it is could be one of the reasons.

Exploring Proximities with Folium

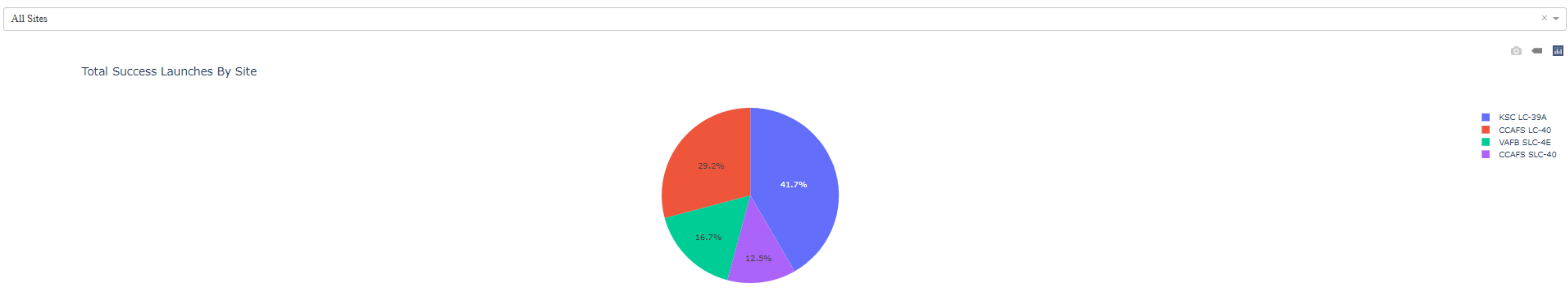


We identified that the launch sites have close proximities to highways, parking facilities, and coast lines, but is further away from cities.

Section 4

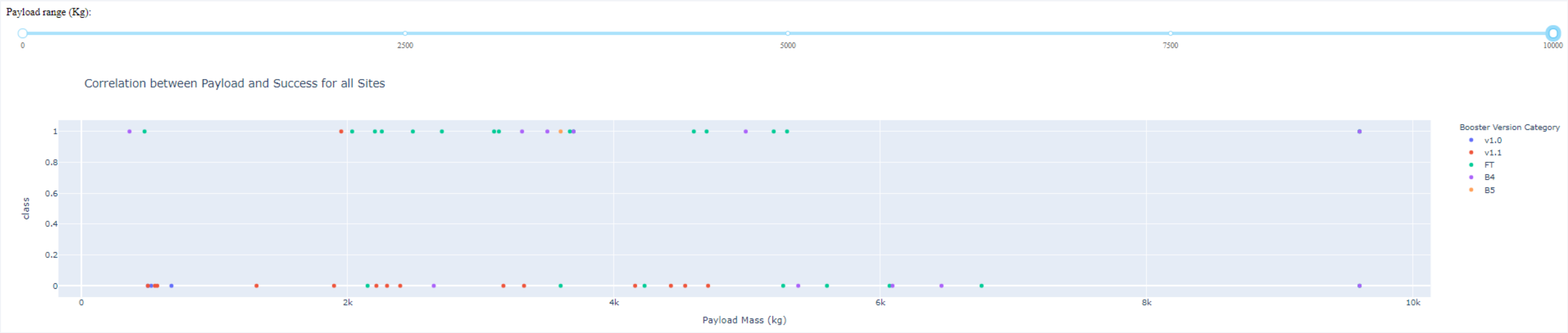
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



- KSC LC-39A has the most succesful launches in comparison to all other sites.

<Dashboard Screenshot 3>

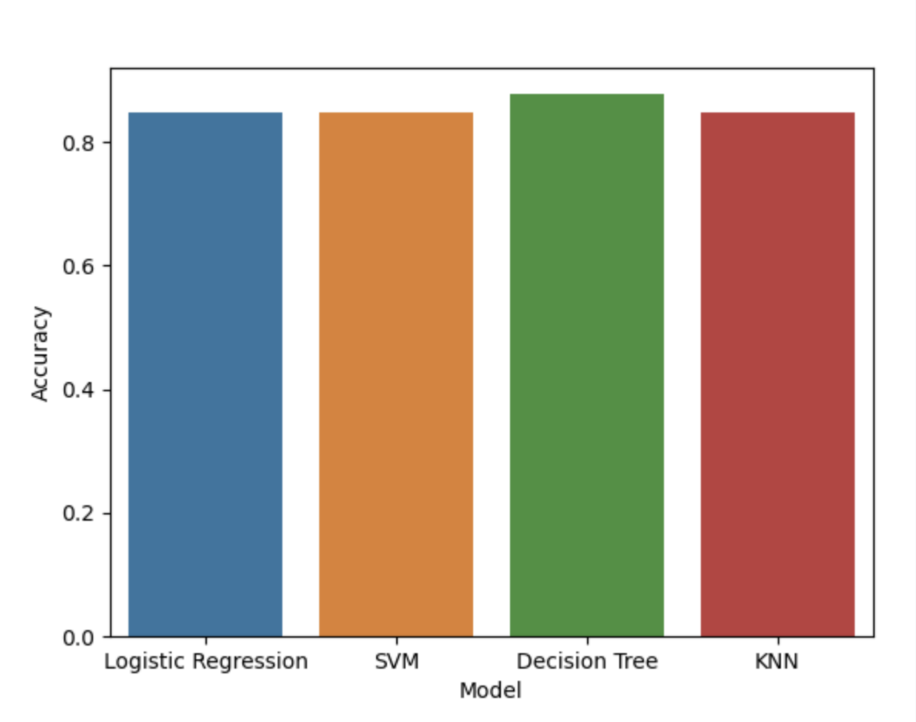


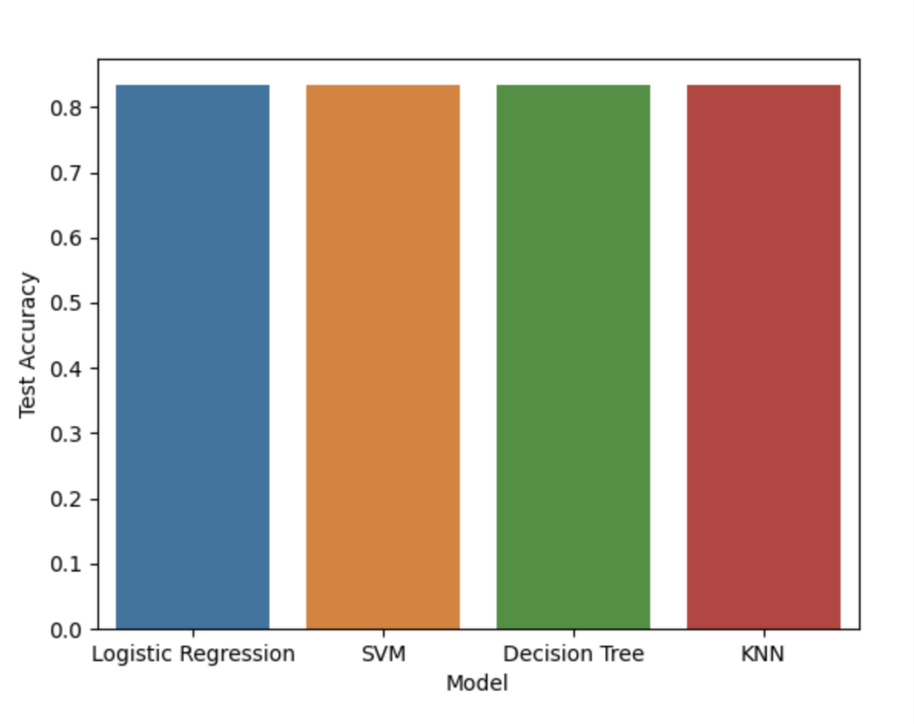
- Booster version FT (Green) has had the largest success rate in comparison to other booster versions
- For payload mass, it appears that the optimum range is between 2k and 4k and anything less than 2k or over 6k tends to be unsuccessful.

Section 5

Predictive Analysis (Classification)

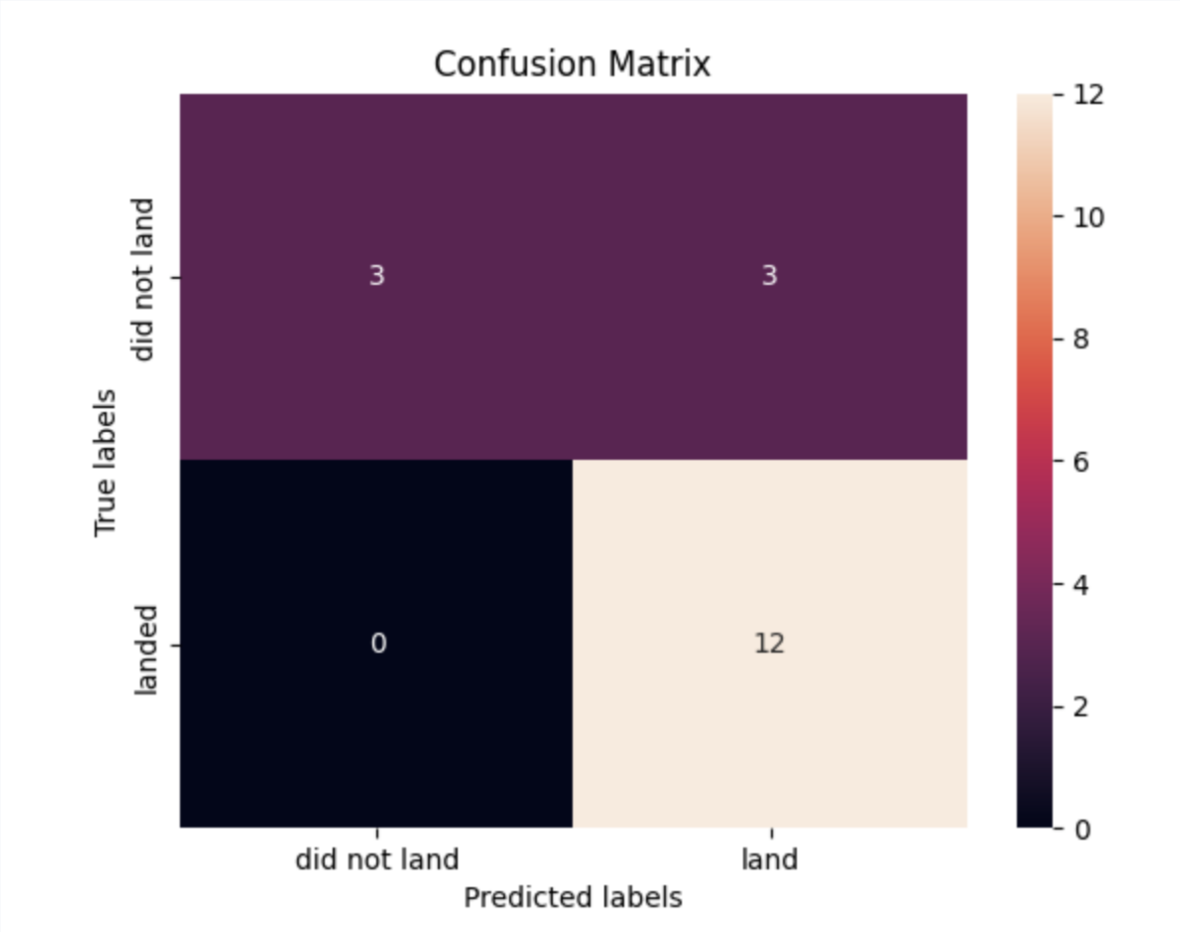
Classification Accuracy





- Decision Tree has the highest accuracy at 87.7%.
- The test accuracy for all models are the same at 83%

Confusion Matrix



- We can see that there are 3 true positives being did not land v did not land
- There are 12 true negatives being landed v land
- The biggest problem then are the false positives being did not land v land

Conclusions

- To prepare the data, we used API and web scraping to extract the data, and performed data wrangling such as replacing nulls or adding new columns.
- In our initial data exploration analysis, we found that there were 4 different launch sites, most of the payload mass was below 8000 kg, and launch success increased with increasing number of years and experience.
- Using Folium data visualization, we explored the geography of different launch sites. All of the launch sites are close to the coastline and are away from the cities. KSC LC39A have the most amount of successful launches compared to other sites.
- KSC LC39A having the highest success rate is corroborated through our pie chart using Plotly. In addition, we identified that Booster version FT has had the highest success rates and that most successful launches tends to be between 2k – 4k kg payload mass.
- Finally we used predictive analysis to test the accuracy of launches and found that Decision Tree has the highest accuracy of 87.7%. The main issue with the models is due to false positives.

Appendix

- Please refer to the Github repository: https://github.com/rikahcli/IBM-Capstone-Project

Thank you!