IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zemelak Goraga

4/28/2023

Outline

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

Executive Summary

Summary of methodologies

- Data Collection – API & Web Scraping
- Data Wrangling
- Exploratory Data Analysis – SQL & Data Visualization
- Interactive Visual Analytics - Folium
- Machine Learning Prediction

Summary of all results

- Exploratory Data Analysis
- Interactive analytics
- Predictive Analytics

Introduction

Project background and context

- This project focused on a rocket launch case of Space X. It 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.

Problems you want to find answers

In this capstone, we will predict if the Falcon 9 first stage will land successfully?

- What factors determine if the rocket will land successfully?
- The interaction amongst various features
- What operating conditions needs to be in place for successful landing

Section 1

Methodology

Methodology

- Executive Summary

Data collection methodology:

- Data collection Method

 - SpaceX API and web scraping from Wikipedia was applied for data collection

- 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

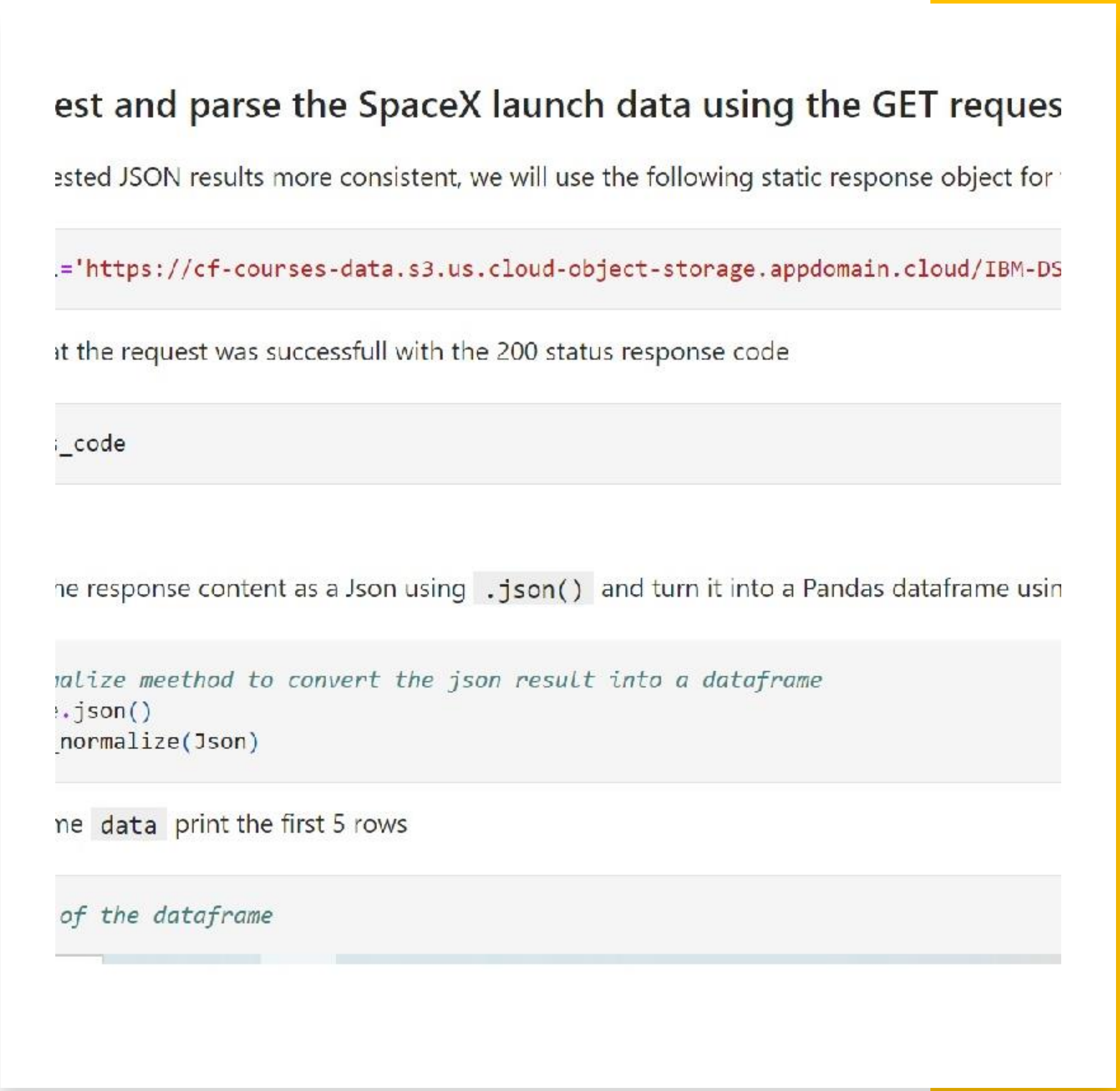
Data Collection

- The data was collected using various methods:
  - Data was first collected using SpaceX API by making a get request
  - In the Next step, the response content was decoded as a Json using .json() function call and turned it into a pandas dataframe using .json_normalize().
  - Then the data was cleaned, checked for missing values and fill in missing values
  - Also performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page

Data Collection – SpaceX API

- The get request was applied to the SpaceX API to collect data, clean it and did data wrangling and formatting.

- The link to the notebook is:
- https://github.com/ZemelakGoraga/IBM_Data_Science/blob/main/Complete%20the%20Data%20Collection%20API-Space%20X%20Falcon%209%20rocket%20launche.ipynb



Data Collection - Scraping

- Web scraping was applied to collect Falcon 9 historical launch records from a Wikipedia using BeautifulSoup and request, to extract the Falcon 9 launch records from HTML table of the Wikipedia page

- Then created a data frame by parsing the launch HTML
- The link to the notebook is:
- https://github.com/ZemelakGoraga/IBM_Data_Science/blob/main/Complete%20the%20Data%20Collection%20with%20Web%20Scraping-Space%20X%20Falcon%209%20rocket%20launche.ipynb

TASK 1: Request the Falcon9 Launch Wiki page from it

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML p

In [5]:

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

r = requests.get(static_url)

# assign the response to a object

data = r.text
```

Create a `BeautifulSoup` object from the HTML `response`

In [6]:

```
# Use BeautifulSoup() to create a BeautifulSoup object from a respo

soup = BeautifulSoup(data,"html.parser")
```

Print the page title to verify if the `BeautifulSoup` object was created properl

In [7]:

```
# Use soup.title attribute

print(soup.title)
```

Data Wrangling

- Data was filtered using the *BoosterVersion* column to only keep the Falcon 9 launches, then dealt with the missing data values in the *LandingPad* and *PayloadMass* columns.
- For the *PayloadMass* , missing data values were replaced using mean value of column.
- Exploratory data analysis was performed and training labels were determined.
- The number of launches at each site were calculated , and the number and occurrence of each orbits too.
- The link to the notebook is
- https://github.com/ZemelakGoraga/IBM_Data_Science/blob/main/Data%20wrangling-Space%20X%20Falcon%209%20rocket%20launche.ipynb

TASK 3: Calculate the number and occurence of mission outcome of the orbits

Use the method `.value_counts()` on the column `Outcome` to determine the number of `landing_outcomes` .Then assign it to a variable landing_outcomes.

In [7]:

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

Out[7]:

```
Outcome
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: count, dtype: int64
```

`True Ocean` means the mission outcome was successfully landed to a specific region of the ocean while `False Ocean` means the mission outcome was unsucces
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 outcor
unsuccessfully landed to a ground pad. `True ASDS` means the mission outcome was successfully landed to a drone ship `False ASDS` means the mission outcome
unsuccessfully landed to a drone ship. `None ASDS` and `None None` these represent a failure to land.

In [8]:

```
for i,outcome in enumerate(landing_outcomes.keys()):
    print(i,outcome)
```

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

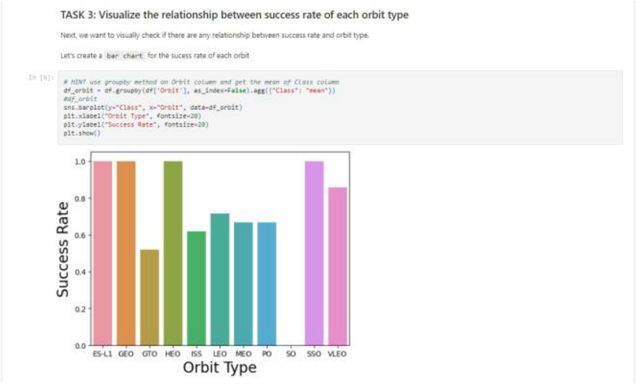
EDA with Data Visualization

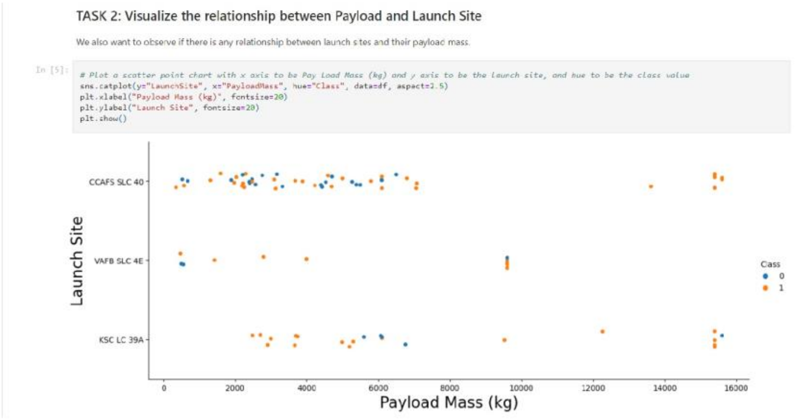
- Exploratory data analysis was performed by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.

- Scatter plots was used to Visualize the relationship between Flight Number and Launch Site, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type.

- Bar chart was used to Visualize the relationship between success rate of each orbit type; Whereas, Line plot was used to Visualize the launch success yearly trend.

- The link to the notebook is:
- https://github.com/ZemelakGoraga/IBM_Data_Science/blob/main/EDA%20with%20Data%20Visualization-Space%20X%20Falcon%209%20rocket%20launche.ipynb





EDA with SQL

- SQL queries were performed for EDA in order to find answer for the names of unique launch sites, the total payload mass carried by boosters launched by NASA (CRS), the average payload mass carried by booster version F9 v1.1, and the total number of successful and failure mission outcomes .
- The link to the notebook is
- https://github.com/ZemelakGoraga/IBM_Data_Science/blob/main/Execute%20sql%20queries-Space%20X%20Falcon%209%20rocket%20launche.ipynb

- Query for displaying names of unique launch sites

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

- Query to show 5 records of launch sites that begin with 'CCA'

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

- Query to show the total payload mass carried by boosters launched by NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)
```

- Query showing average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%'
```

Build an Interactive Map with Folium

- Folium map was created to marked all the launch sites
- Map objects such as markers, circles, and lines were created to mark the success or failure of launches for each launch site.
- Feature launch outcomes (failure or success) were assigned to class 0 and 1, where 0 for failure, and 1 for success.
- which launch sites have relatively high success rate was identified using the color-labeled marker clusters
- The distances between a launch site to its proximities were calculated.
- Here is the GitHub URL of the map:
- https://github.com/ZemelakGoraga/IBM_Data_Science/blob/main/Interactive%20visual%20analytics%20using%20Folium-Space%20X%20Falcon%209%20rocket%20launche.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard application with Plotly dash was built by adding a Launch Site Drop-down Input Component, adding a callback function , adding a Range Slider and adding a callback function

- Pie charts and Scatter graph were plotted for showing the total launches by a certain sites and the relationship with Outcome and Payload Mass (Kg) for the different booster version, respectively.

The link to the notebook is:

- https://github.com/ZemelakGoraga/IBM_Data_Science/blob/main/Interactive%20Dashboard-Space%20X%20Falcon%209%20rocket%20launche

Predictive Analysis (Classification)

- First the data was loaded as a Pandas Datafram using numpy and pandas
- Next, the data was transformed and splited into training and testing.
- Afterwards, different machine learning models were built
- We used accuracy as the metric for our model,
- Finally, the models was improved using feature engineering and algorithm tuning and the best model was identified.
- The link to the notebook is:
- https://github.com/ZemelakGoraga/IBM_Data_Science/blob/main/Predictive%20Analysis-SpaceX-Predictive%20Analysis-Space%20X%20Falcon%209%20rocket%20launche.ipynb

Results

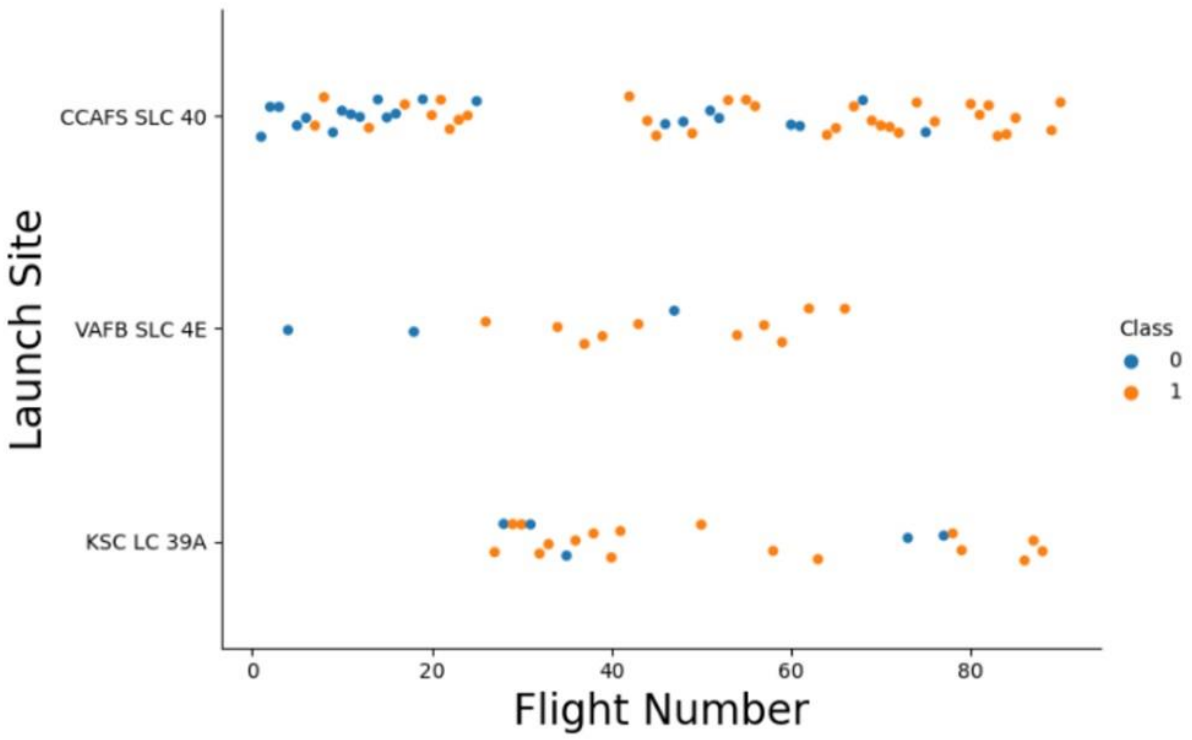
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

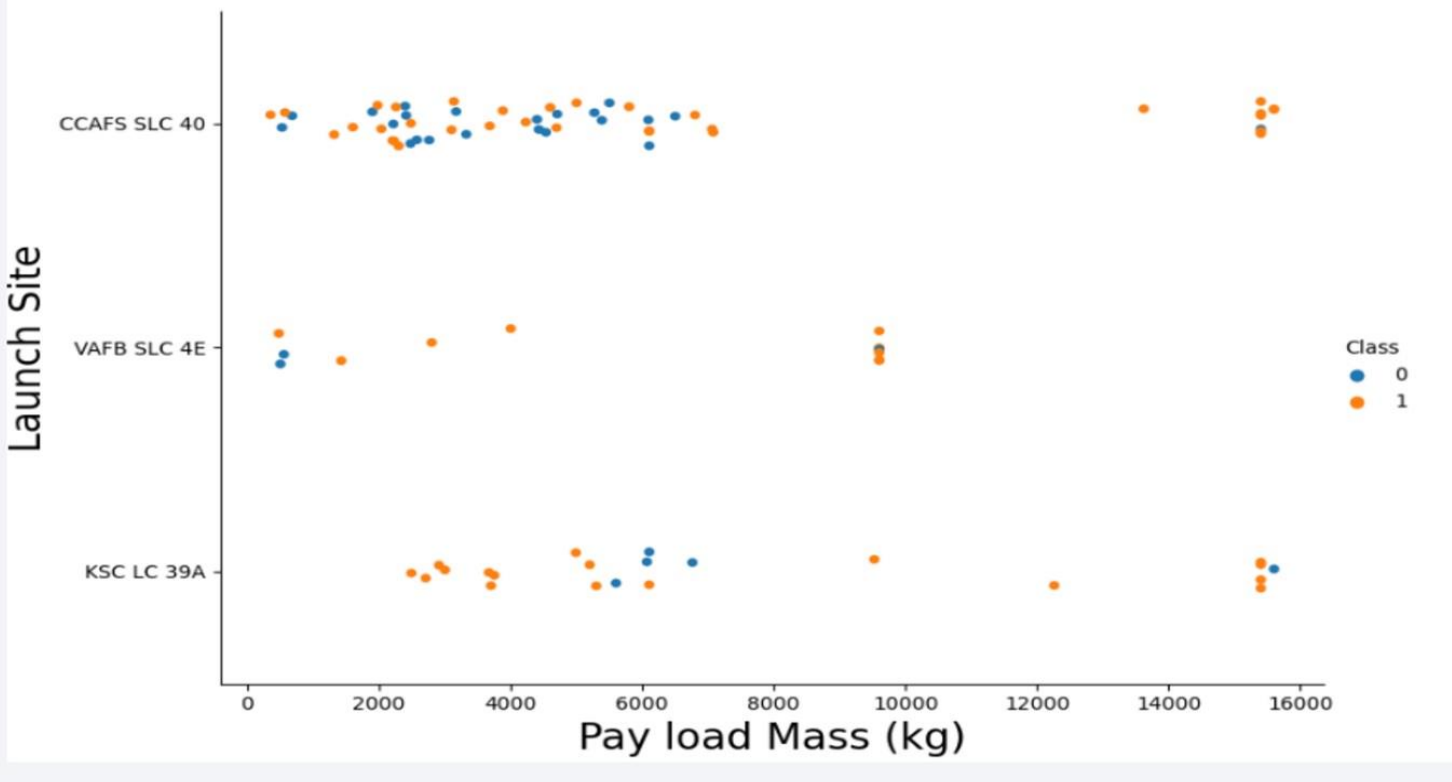
Insights drawn from EDA

Flight Number vs. Launch Site

Insight: the success rate increased as number of flights increased



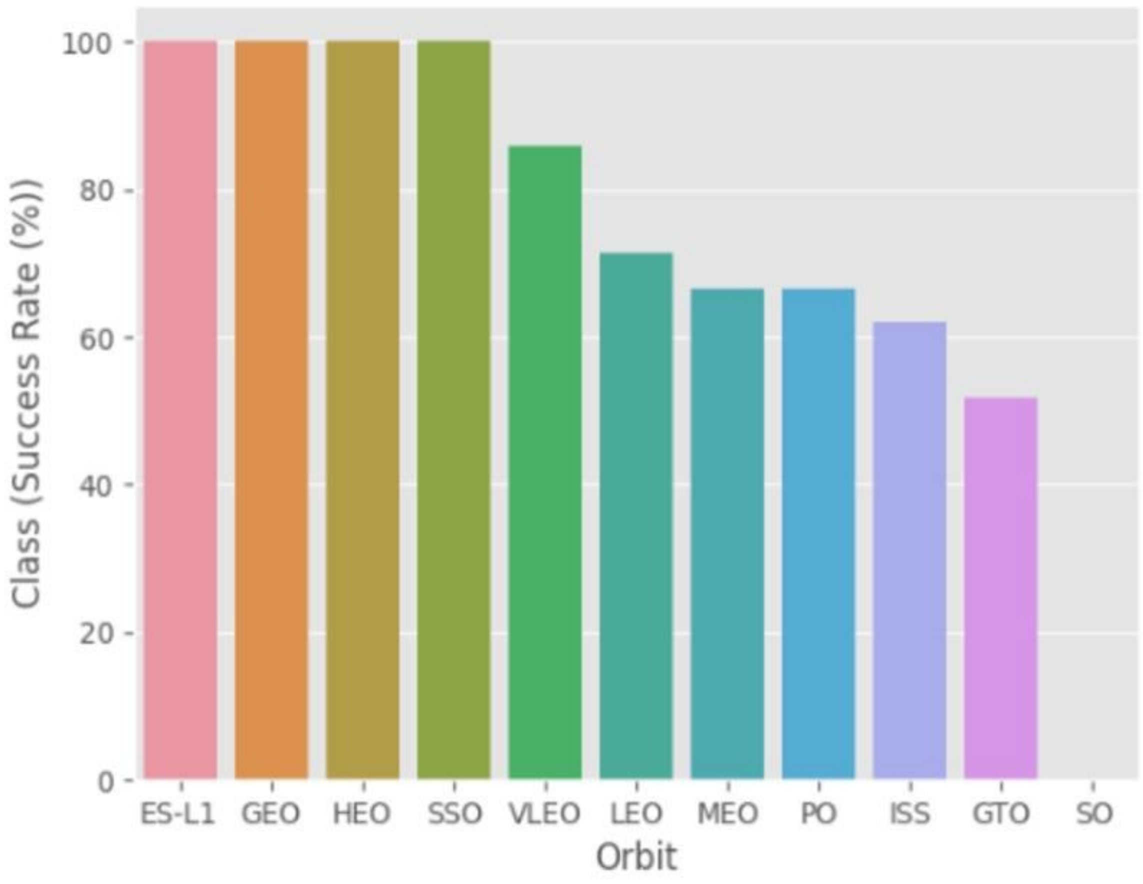
Payload vs. Launch Site



Insight: As the payload mass increased, the success rate also increased

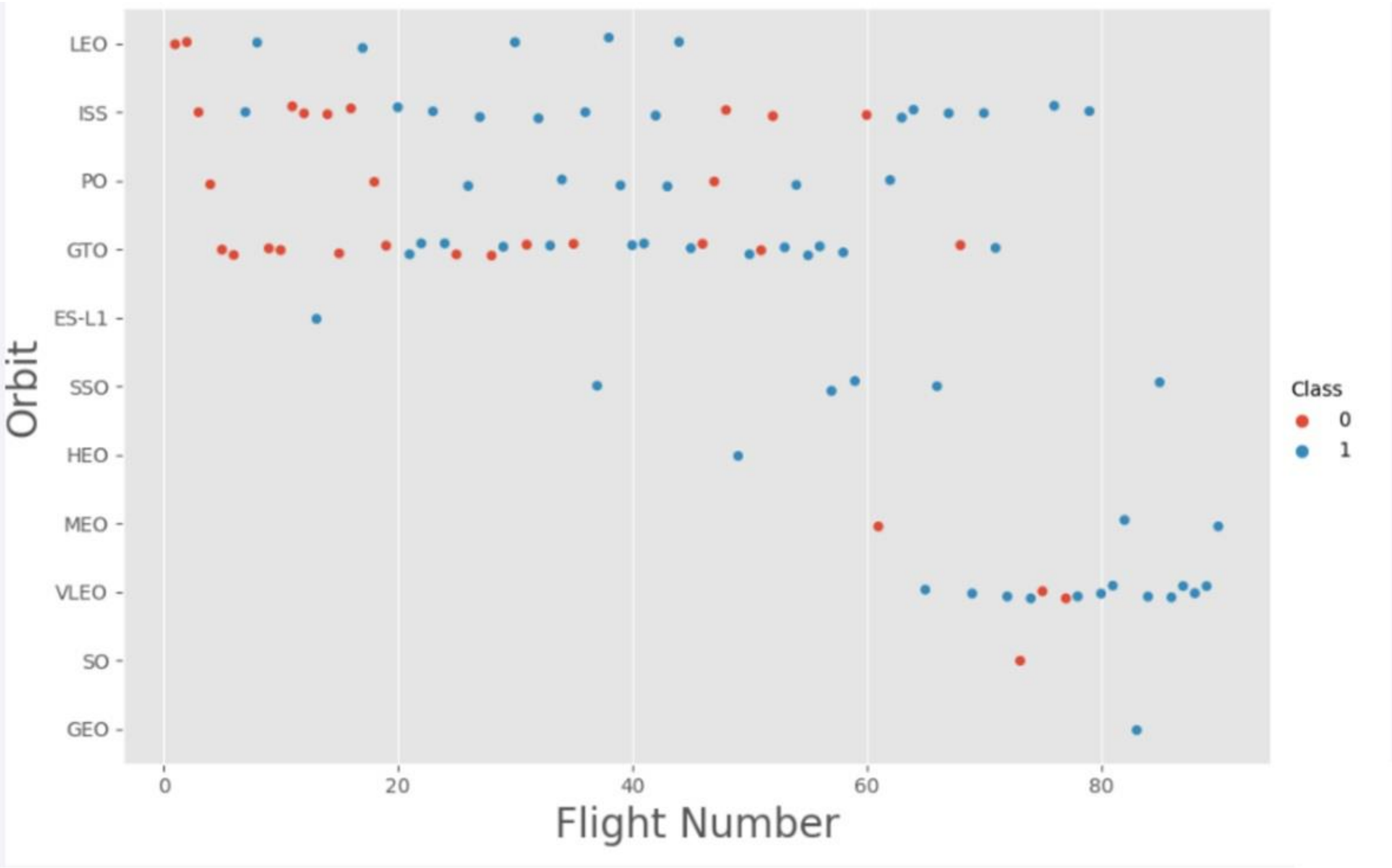
Success Rate vs. Orbit Type

As can be seen from the chart, orbit ES-L1, GEO, HEO, and SSO had the highest success rate.

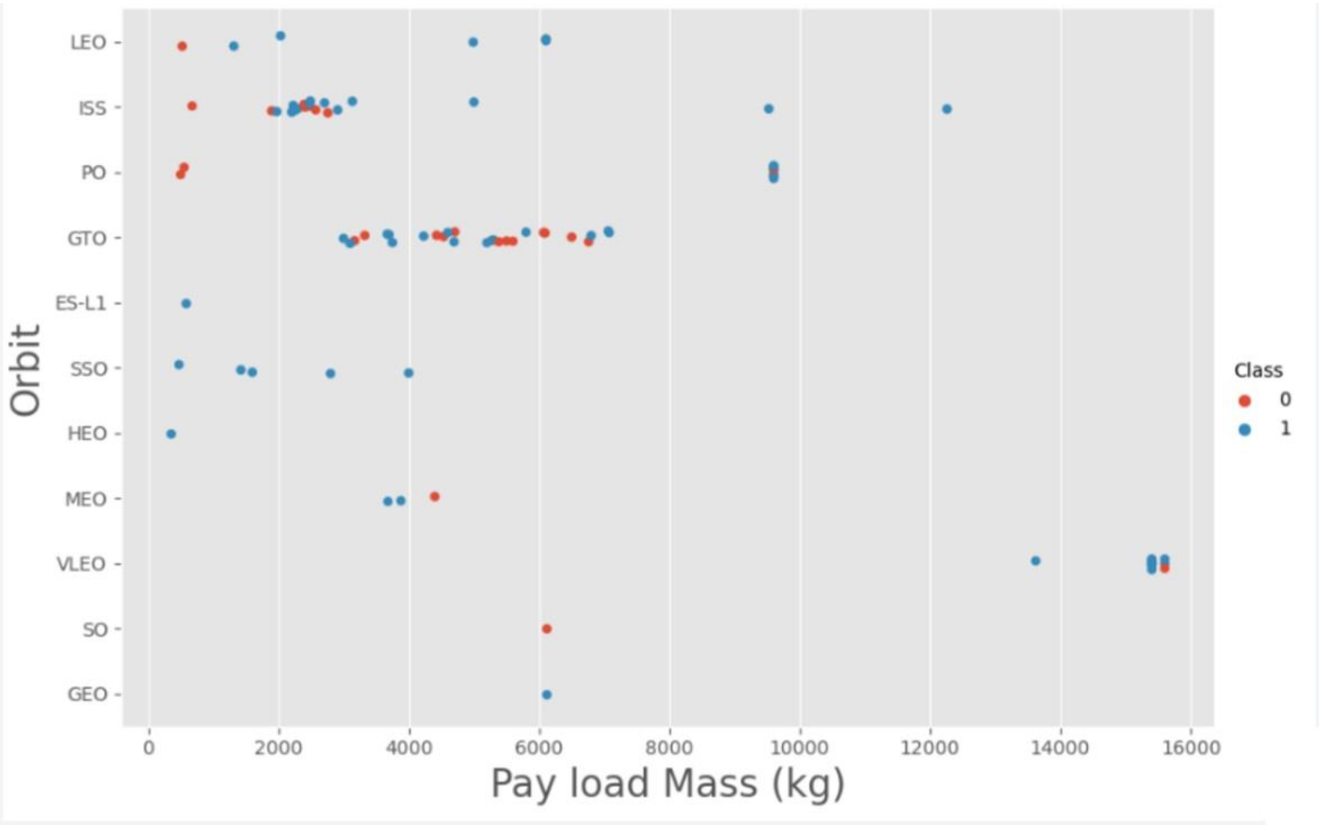


Flight Number vs. Orbit Type

- As can be seen from the plot, the more the number of flights(Flight Number) at Orbit LEO, ISS, ES-L1, and VLEO, there is frequent success. This might imply positive relationship between target and independent variables at those Orbits.



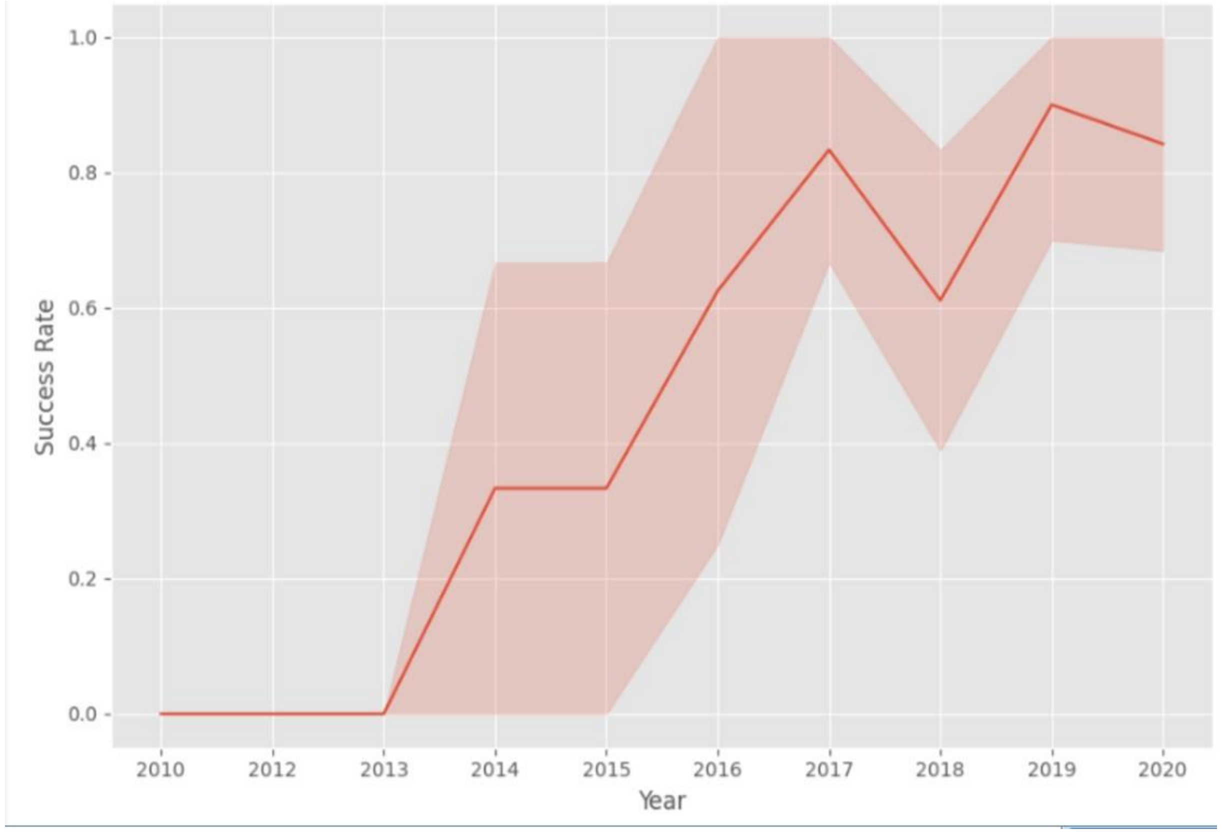
Payload vs. Orbit Type



- As can be seen from the plot, the more the pay load mass, there is most frequent success. This is true for PO, LEO and ISS Orbits. So, it seems that there is a positive r/ship between target and independent variables at those three Orbits.

Launch Success Yearly Trend

The plot shows an increasing success rate from the year 2013 on wards. The highest success rate was observed in the year 2019.



All Launch Site Names

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

In [10]:
```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

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

- The unique launch sites from the SpaceX data were screened using the ‘SELECT DISTINCT’ statement.
- The slide on the left shows the those unique sites.

Launch Site Names Begin with 'CCA'

- As can be seen below, the query ‘LIKE’ command with ‘%’ wildcard in ‘WHERE’ , was used to display 5 records where launch site names start with `CCA`

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

```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Task 3

Display the total payload mass carried by boosters launched by NA

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

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

Out[ ]:

SUM(PAYLOAD_MASS__KG_)
45596

Total Payload Mass

As can be seen the left slide, total payload carried by boosters from NASA was calculated using the ‘SUM()’ function and the result was 45596

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster

In [ ]:

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

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

Out[ ]:

AVG(PAYLOAD_MASS__KG_)
2928.4

Using the AVE() function , the average payload mass carried by booster version F9 v1.1 was 2928.4

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcon

Hint:Use min function

```
In [ ]: %sql select min(DATE) from SPACEXTBL WHERE "La

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

Out[ ]:

min(DATE)
01-05-2017

- The ‘MIN()’ function was used to determine the date on which the first successful landing outcome was attained.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which had success are mentioned below. The **WHERE** clause was used to filter them.

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 * FROM 'SPACEXTBL'
```

```
%sql SELECT DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOA
```

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

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes were calculated using the 'COUNT()' together with the 'GROUP BY' statement

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)

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

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

The list of all the boosters that have carried the Max payload of

15600kgs were obtained by using a subquery in the **WHERE** clause

2015 Launch Records

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 [ ]:

```
%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[ ]:

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015 were obtained using a combinations of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions

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

- The landing outcomes BETWEEN 2010-06-04 to 2010-03-20 were filtered using

COUNT and WHERE clauses followed by the GROUP BY and the ORDER BY clauses

Task 10

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

In [ ]:

```
%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[ ]:

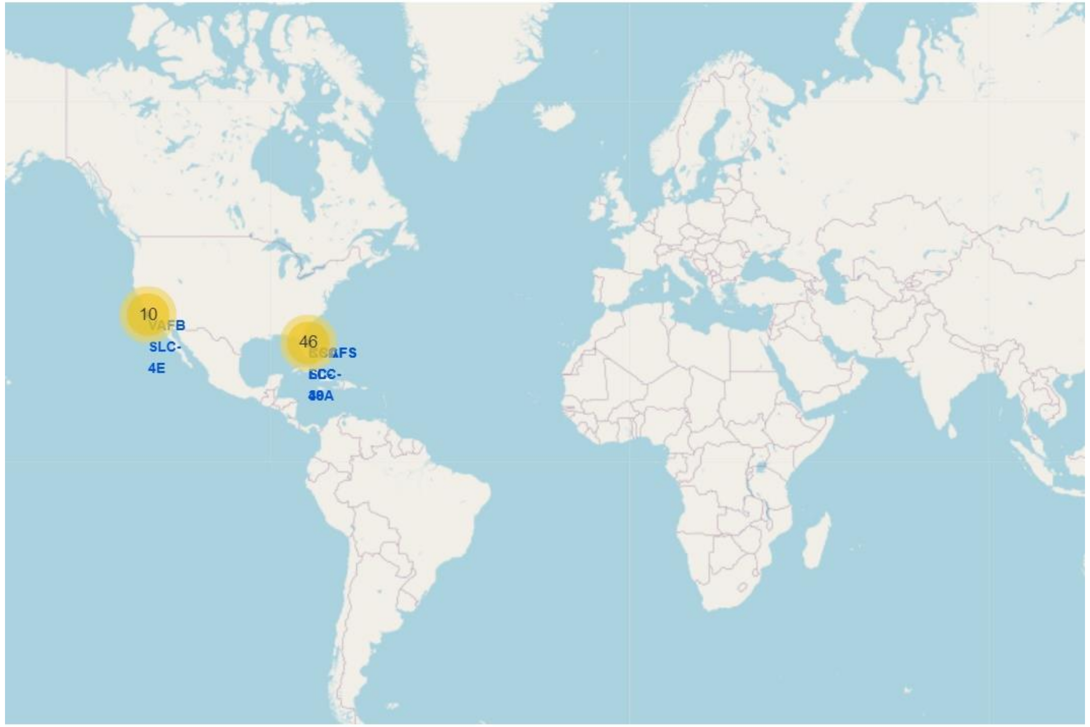
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

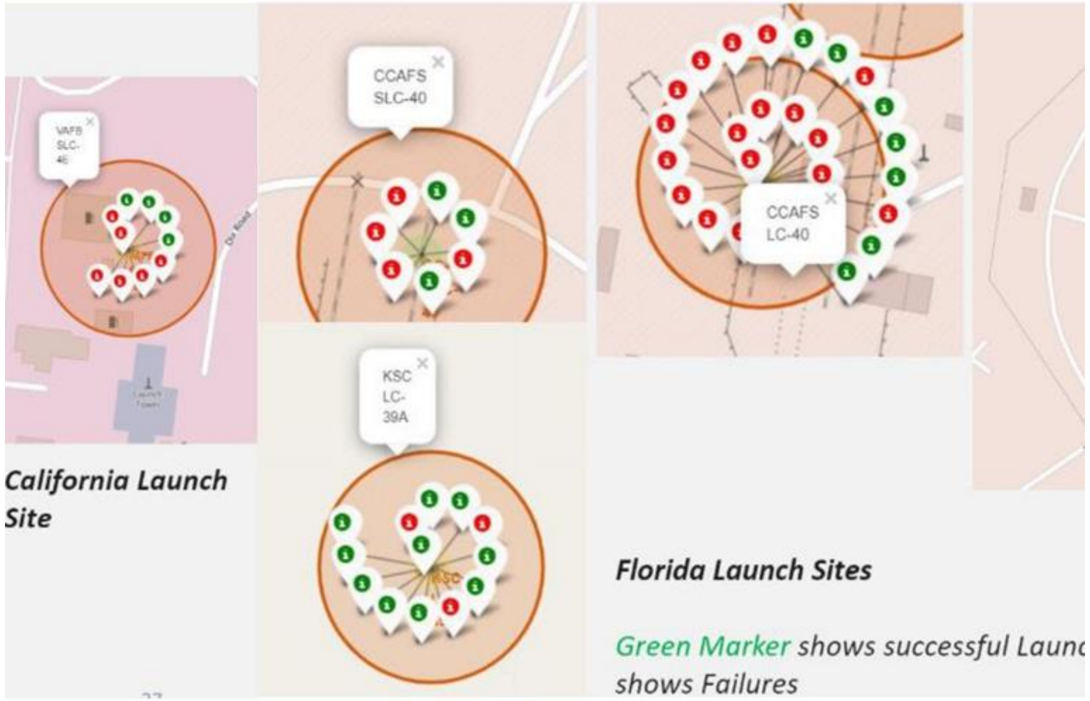
All launch sites global map markers

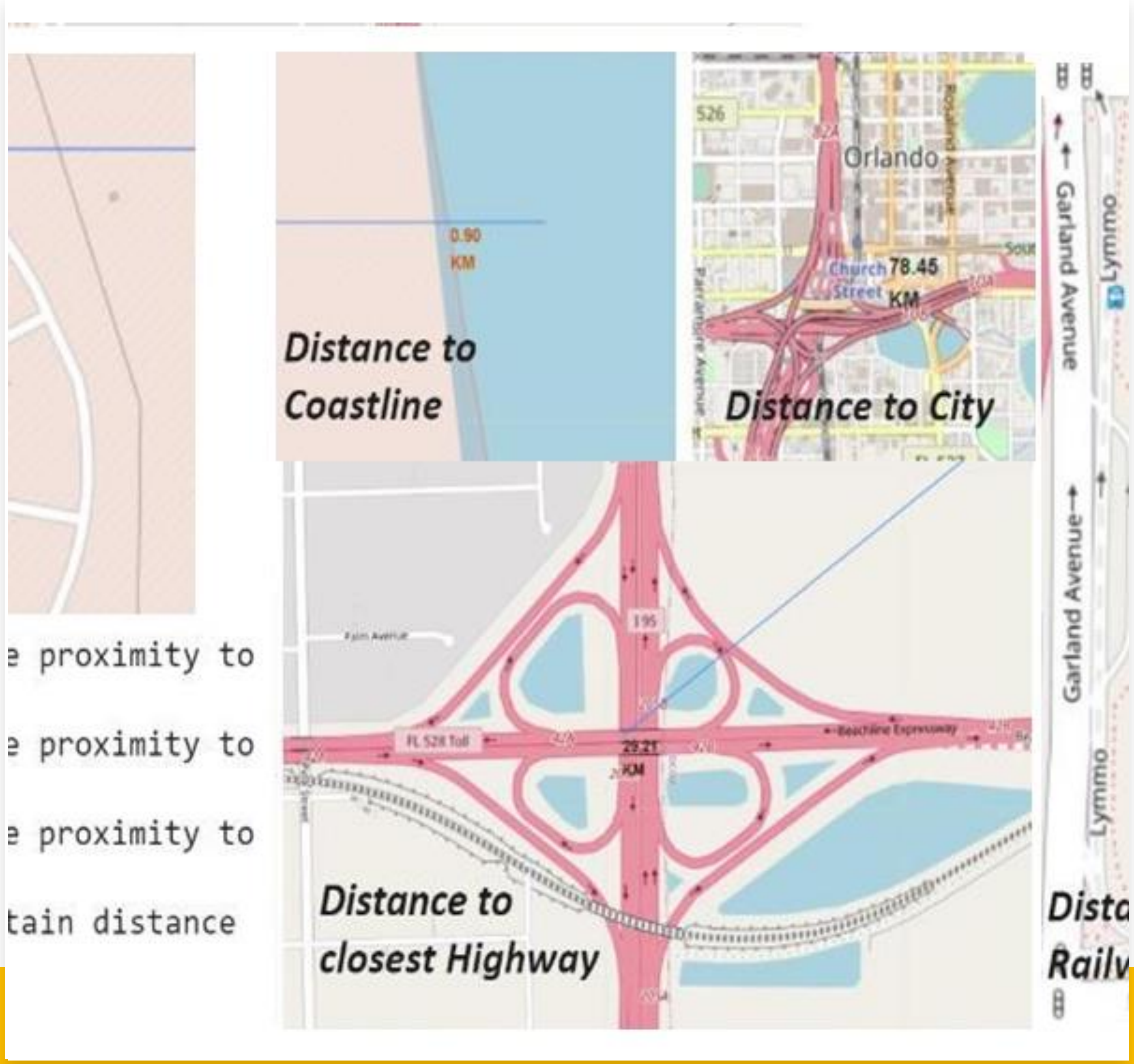
- The launch sites were in close proximity and located close to the Equator



Markers showing launch sites with color labels

- Launch site KSC LC-39A has relatively high success rates as compared with all other sites





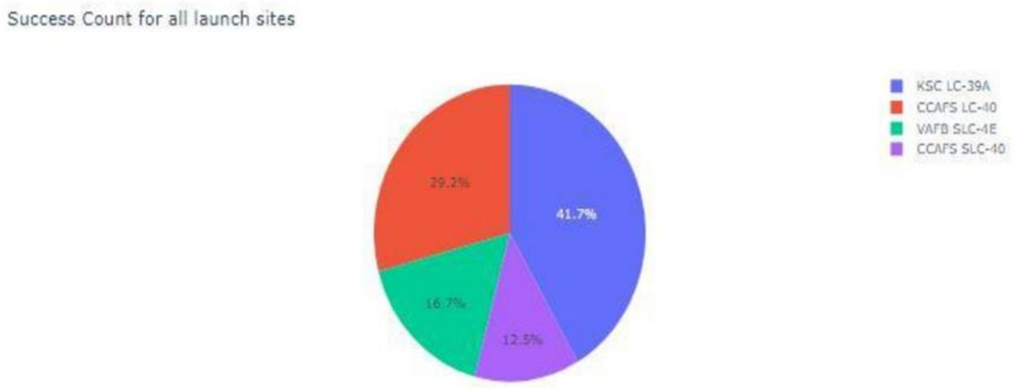
Launch Site distance to landmarks

Section 4

Build a Dashboard with Plotly Dash

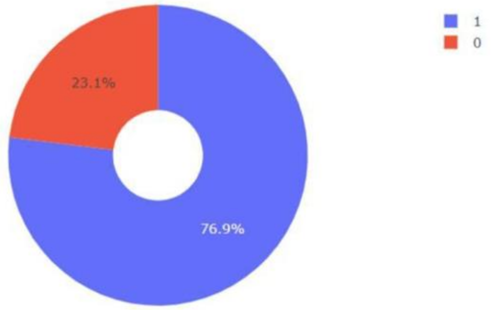
Pie chart showing the success percentage achieved by each launch site

- Among all the launch sites, the highest launch success rate (42%) was achieved by Launch site KSC LC-39, followed by CCAFS LC-40 (29%)

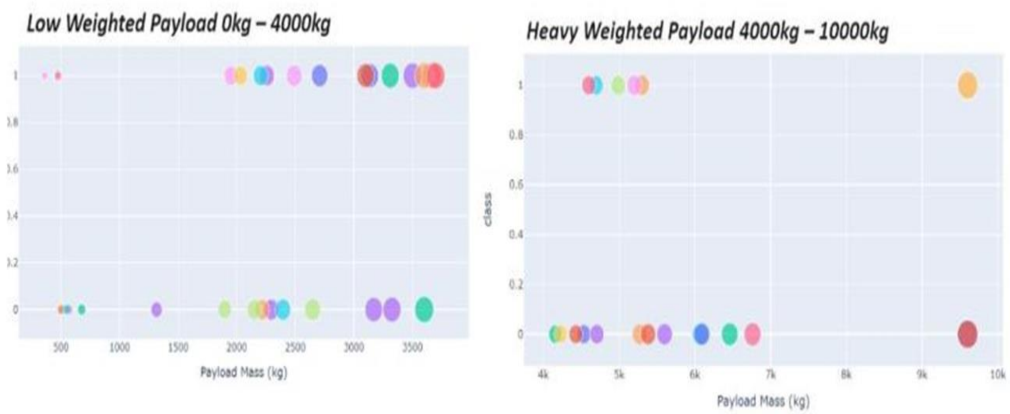


Pie chart showing the Launch site with the highest launch success ratio

Launch site CCAFS LC-40 had the success ratio of 76.9%



Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider



We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

- The plot shows that the low weighted payloads had higher success rate than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

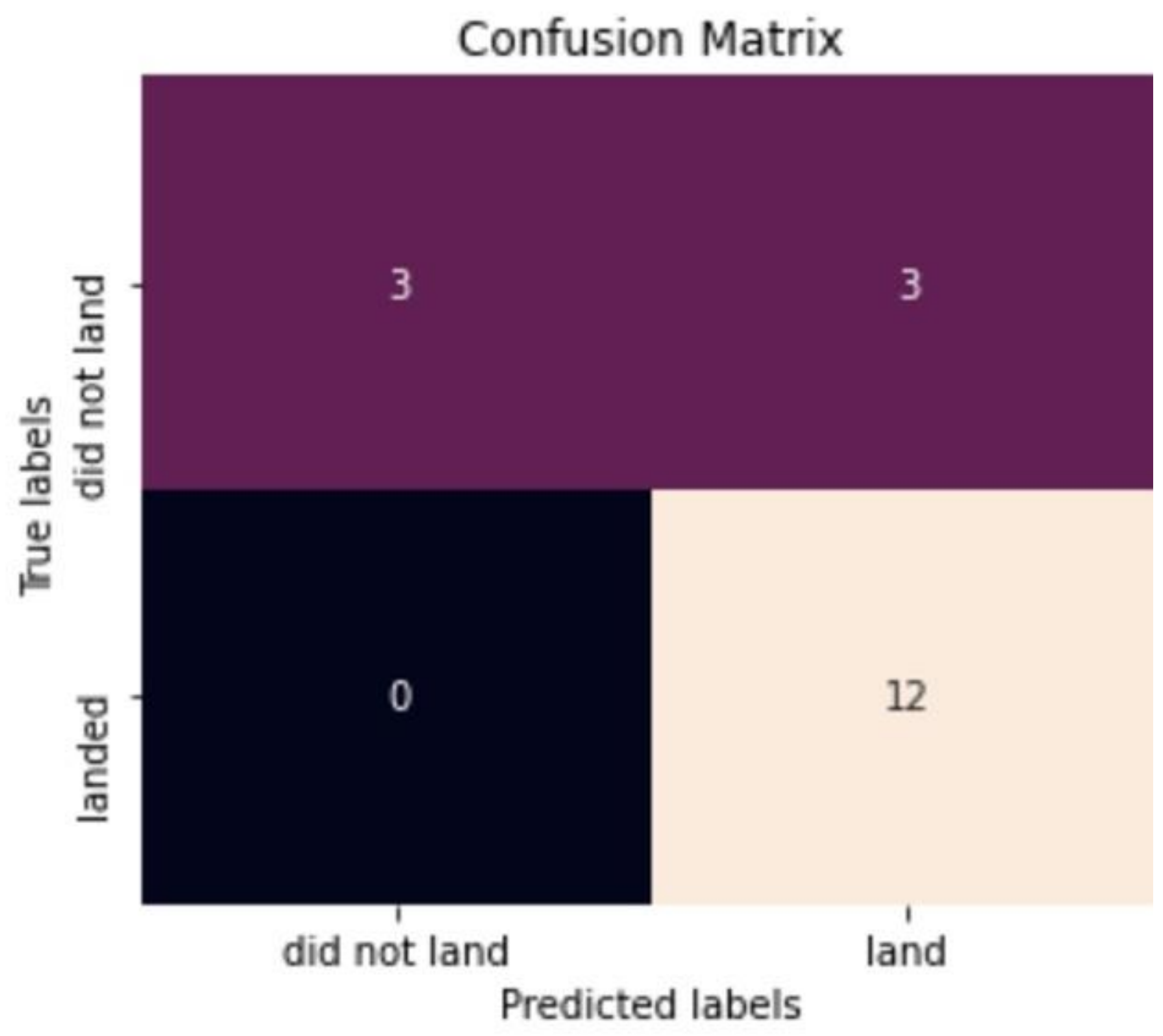
Classification Accuracy

- DecisionTree was found to be the best model with a score of 0.873

```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_s
```



Confusion Matrix

- Except, the false positives .i.e., unsuccessful landing marked as successful, . all the 4 classification model can distinguish between the different classes and had similar results.

Conclusions

- The result of data analysis showed that, the degree of successful landing outcomes differed by launch sites
- The most successful launches site was KSC LC-39A
- The most success rates were obtained at Orbit ES-L1, GEO, HEO, SSO, and VLEO
- A positive r/ship was observed between flight number and success rate., where the former increase the later too.
- Looking at the distribution of the success rate by year, success was increased from 2013 on wards.

Thank you!