Winning Space Race with Data Science

Ammar Zidan
<Aug 16, 2022>

Outline

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

Executive Summary

- Summary of methodologies

Data Collection throughout (web scraping and SpaceX API)

Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics

Machine Learning's test, split, and prediction.

- Summary of all results

Data was collectable using (web scraping and SpaceX API) even though was from a public source.

Exploratory Data Analysis (EDA) helped with the data processing phase and data approaching and manipulation.

Machine Learning model showed the data tested, and amount of data trained and predicted best launches possible based on the model results.

Introduction

- Project background and context

A company Space Y, are competitors of Space X and using their public rocket lunching data to determine few questions the management were considering.

- Problems you want to find answers

To estimate the total cost for launches, and predicting successful landings of the first stage of rockets;
locate the best place for successful launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Data from Space X was collected from two different sources:

- Space X API (https://api.spacexdata.com/v4/rockets/)
- WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Perform data wrangling
  - With data acquisition and collection, there was a need for data cleansing and manipulation, which created a landing more usable outcome, based on SpaceX data as Income after summarizing and analyzing data outputs.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Always facing a problem connecting my Jupyter Notebook to my IBM cloud, never the less when after establishing connection, working with SQL to get make our data set more clear and use it later in a artistic visualization serves the purpose of the presentation.

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Collected dataset till this step weas normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different sets of parameters.

Data Collection

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

Data sets were collected from the following websites, and were scrapped using web scraping command lines and techniques.

- **Space X API**

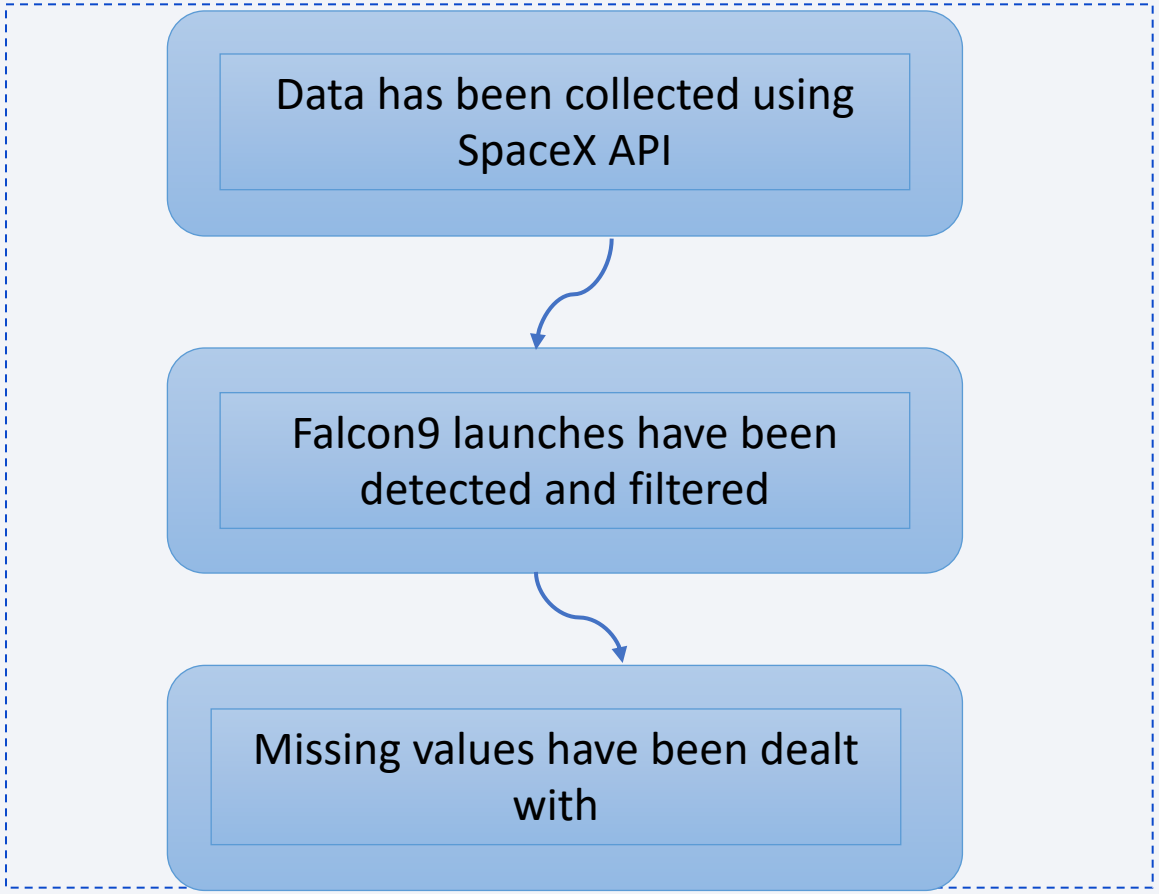
url: (https://api.spacexdata.com/v4/rockets/).

- **Wikipedia**

url: (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches).

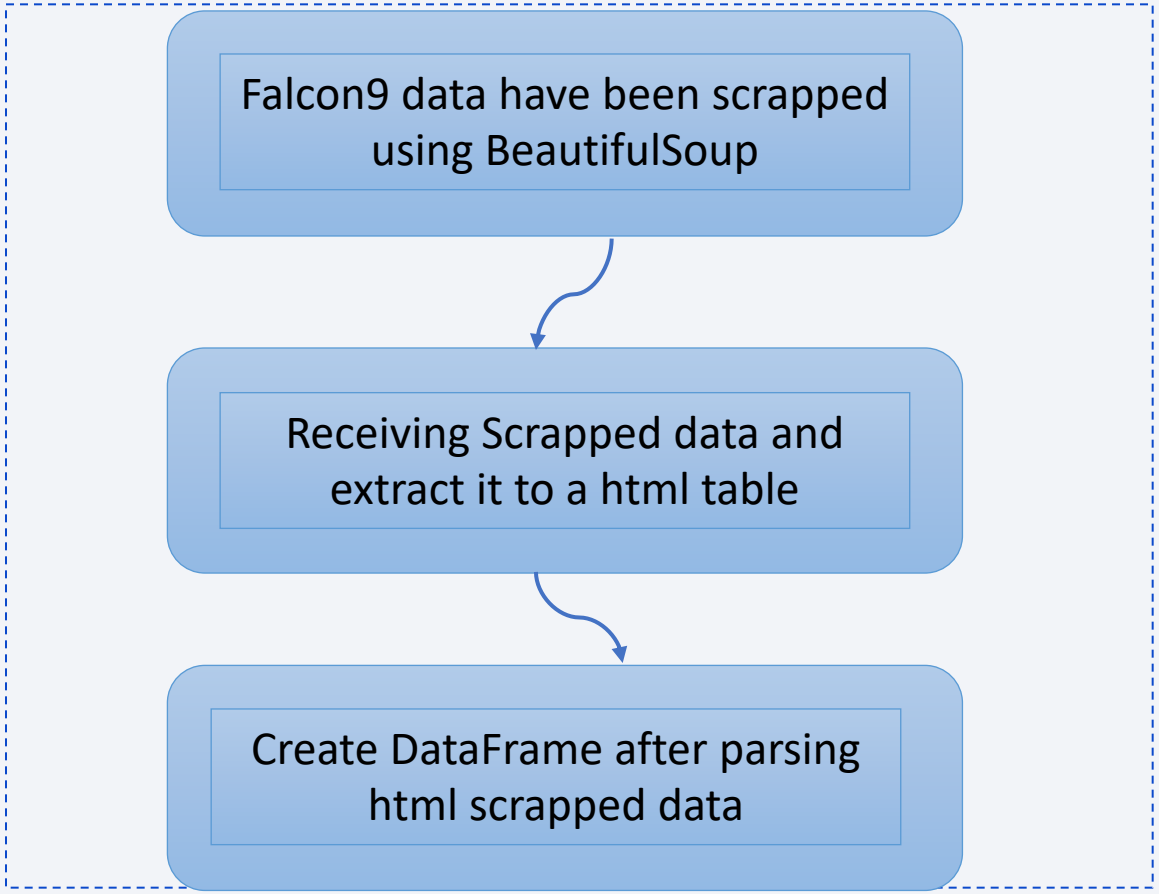
Data Collection – SpaceX API

- SpaceX API has been used in order to collect the dataset we need. With receiving dataset,
- data has been dealt with and checked
  launching areas checked
  falcon 9 has been chosen
  and after we will be taking care of missing values.
- GitHub URL (https://github.com/ammoryindub/FinalProject_DS-IBM/blob/master/spacex_data_collection_api.ipynb)



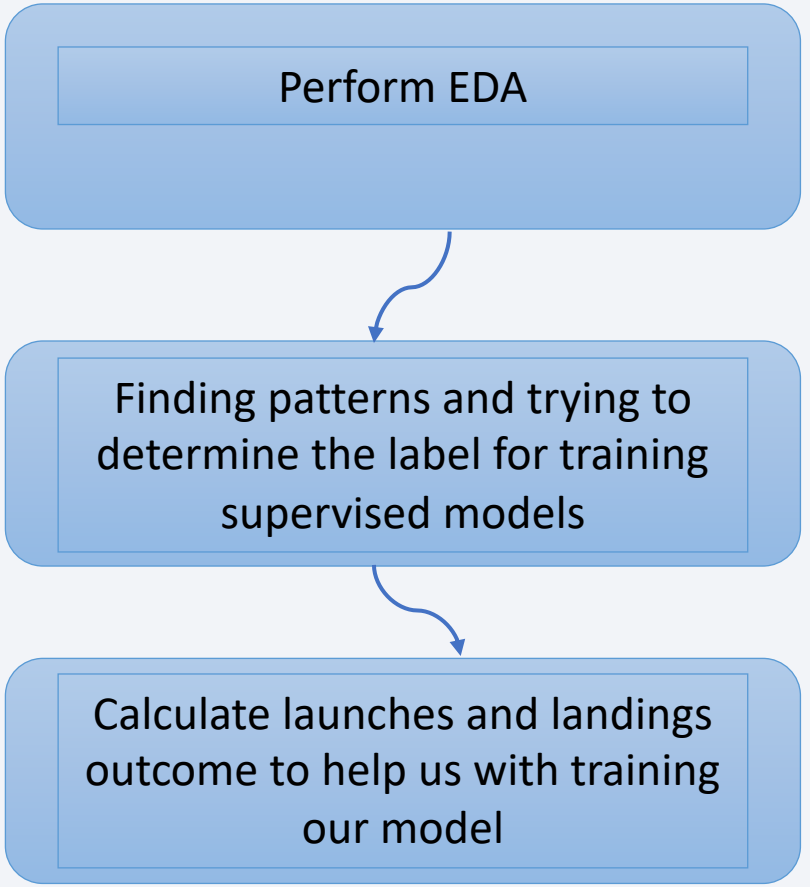
Data Collection - Scraping

- Data of falcon 9 launches has been scrapped using different scrapping methods most famous of all BeautifulSoup
- Page been provided and used for data scrapping was Wikipedia
- GitHub URL (https://github.com/ammoryindub/FinalProject_DS-IBM/blob/master/webscraping.ipynb)



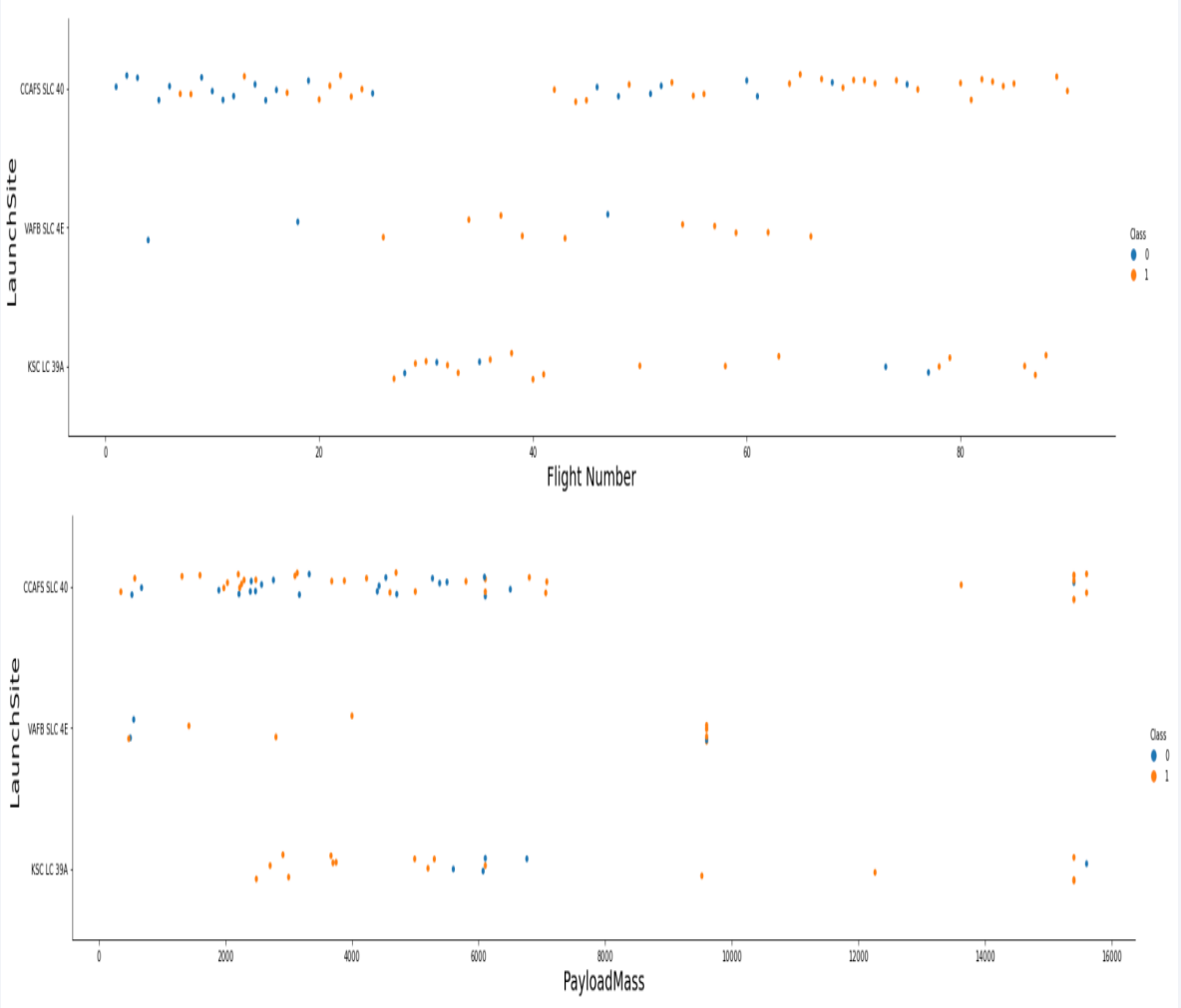
Data Wrangling

- Exploratory Data Analysis (EDA) has been performed in order to find some patterns in the dataset.
- Calculate number of launches and get the landing outcome, to make comparisons.
- GitHub URL (https://github.com/ammoryindub/FinalProject_DS-IBM/blob/master/spacex_data_wrangling.ipynb)



EDA with Data Visualization

- Scatter Plot, Bar plot have been used at this stage in order to explore data, and trying to find patterns.
  “Launch site” been compared against “FlightNumber” and “PayloadMass”
- GitHub URL (https://github.com/ammoryindub/FinalProject_DS-IBM/blob/master/eda-dataviz.ipynb)



EDA with SQL

- Found names of the unique launch sites in the space mission
- Locate launch sites begin with the string 'CCA'
- Payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Landing outcomes (such as Failure or Success) between the date 2010-06-04 and 2017-03-20

- GitHub URL (https://github.com/ammoryindub/FinalProject_DS-IBM/blob/master/eda_sql_coursera.ipynb)

Build an Interactive Map with Folium

- Markers to indicate points (launch sites).
- Circles to indicate highlighted areas around specific coordinates (NASA Johnson Space Center)
- Marker clusters to indicate groups of events in each coordinate (launches in a launch site)
- Lines used to indicate distances between two coordinates
- GitHub URL (https://github.com/ammoryindub/FinalProject_DS-IBM/blob/master/launch_site_location.ipynb)

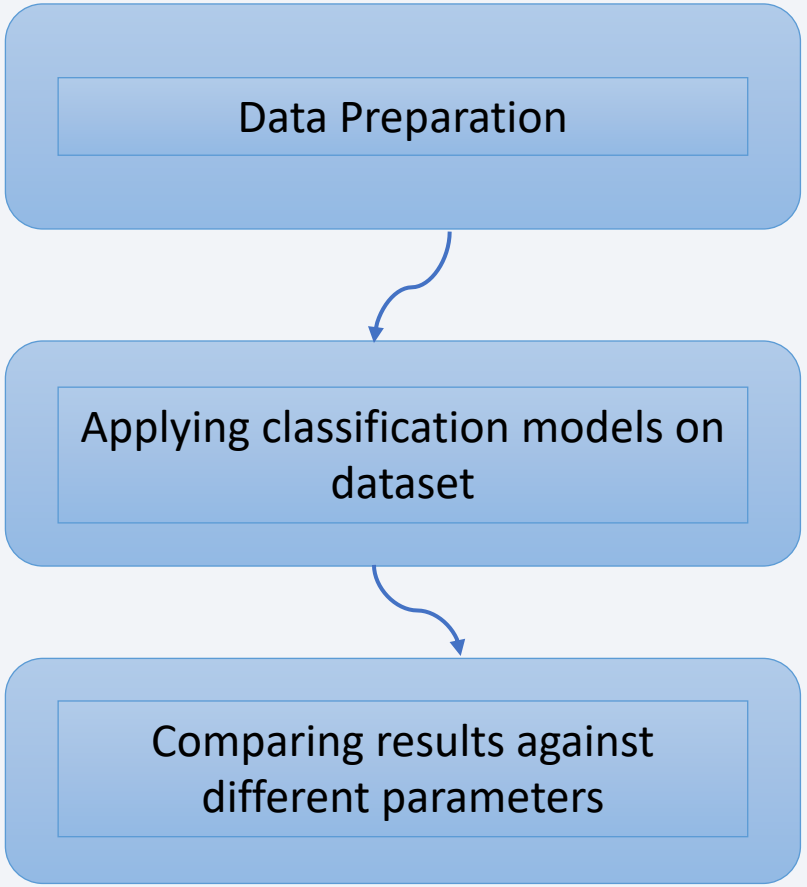
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
- Percentage of launches by site
- Payload range
- This comparison showed us the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.
- GitHub URL (https://github.com/ammoryindub/FinalProject_DS-IBM/blob/master/dash.txt)

Predictive Analysis (Classification)

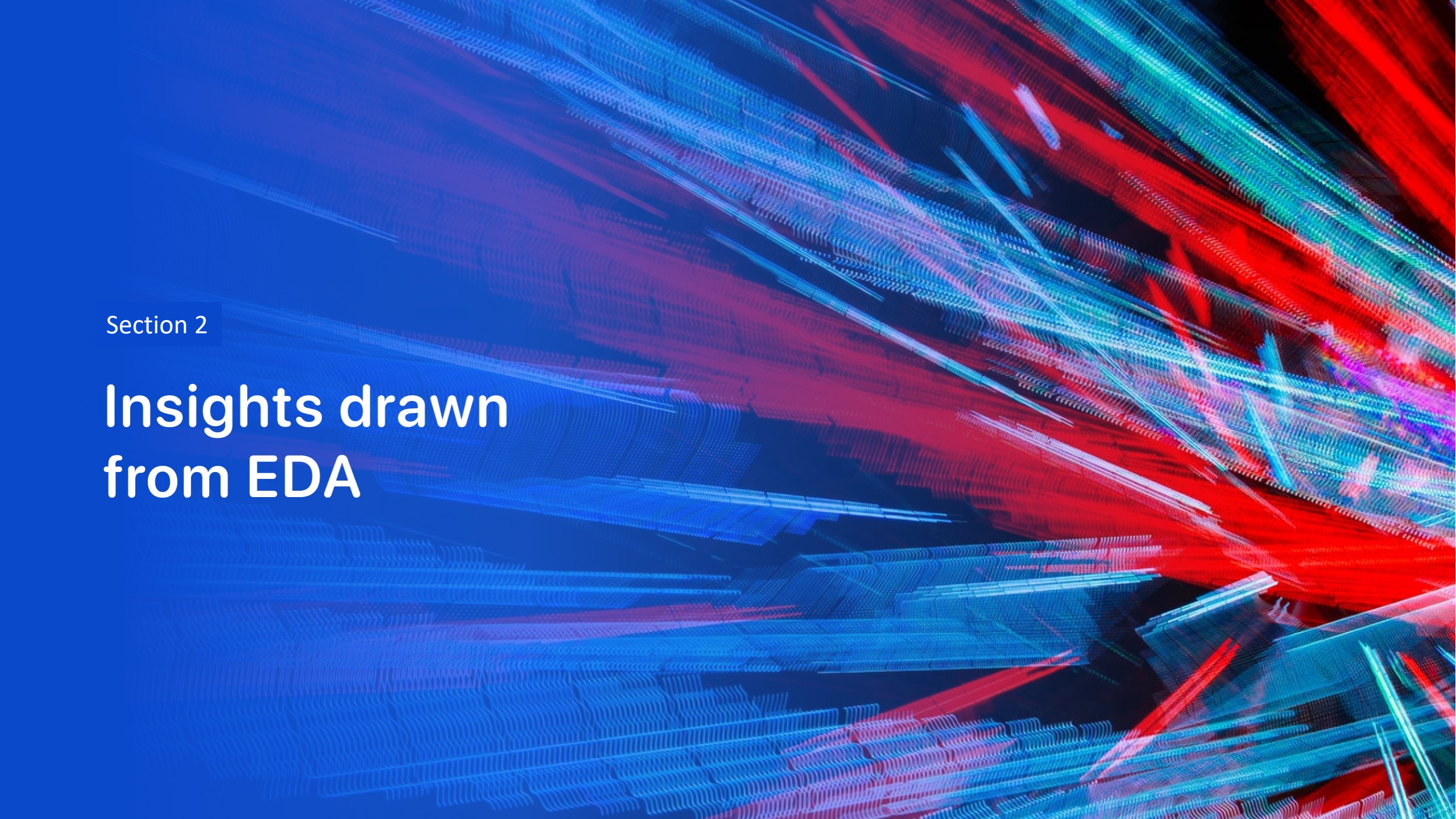
- classification models were compared:
  - logistic regression
  - support vector machine
  - decision tree
  - k nearest neighbors.

- GitHub URL (https://github.com/ammoryindub/FinalProject_DS-IBM/blob/master/SpaceX_machine_learning_prediction.ipynb)

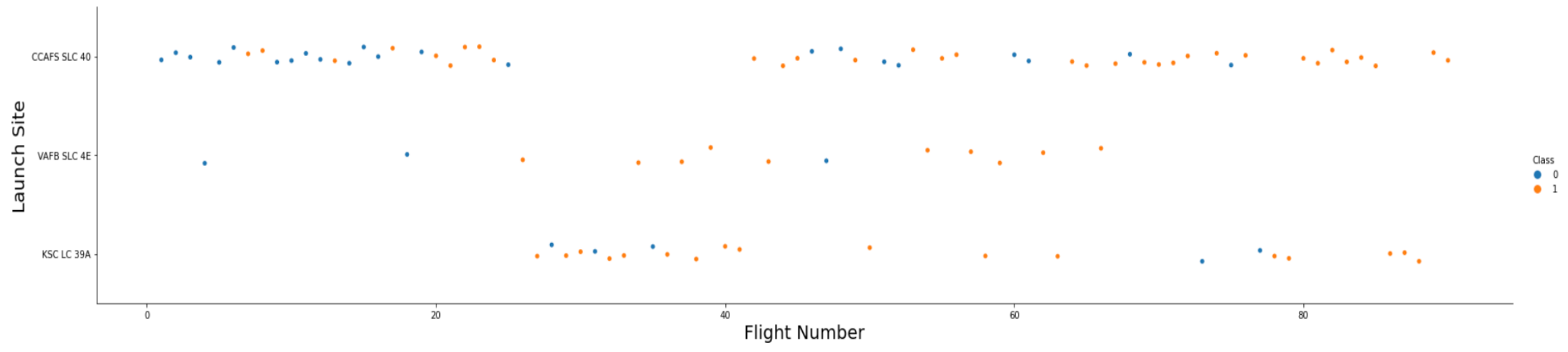


Results

- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

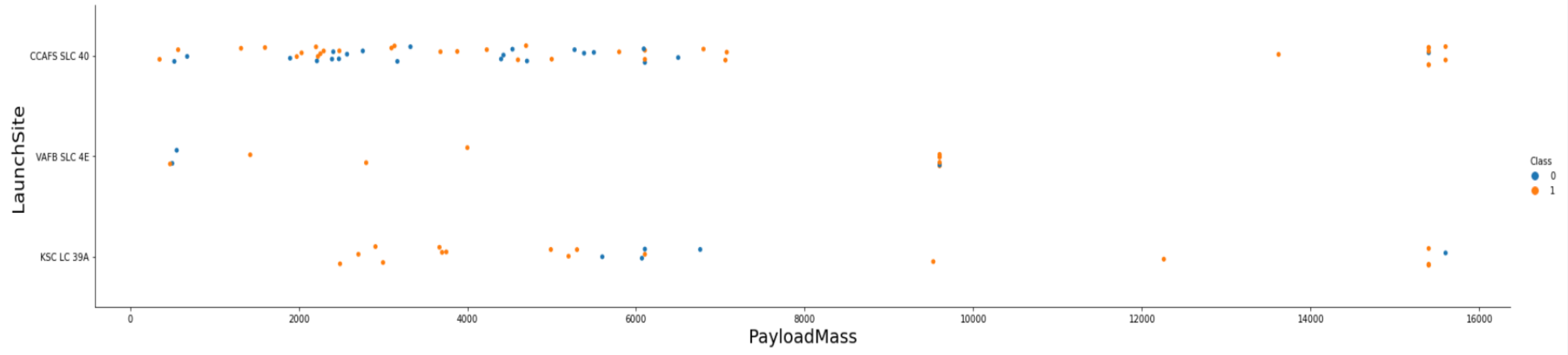

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Best launching site is CCAFS SLC 40, and we can see the they got improved with time.
- VAFB SLC 4E has also improved with time but not enough launches initiated on this location to build a strong decision with it
- KSC LC 39A comes in third place of what we see in the scatterplot.

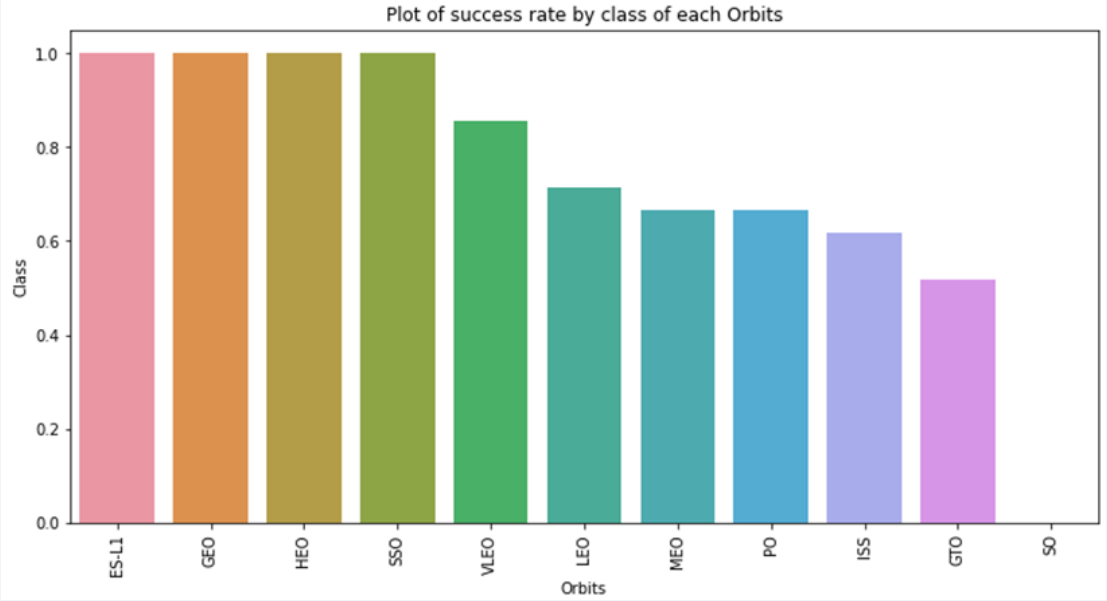
Payload vs. Launch Site



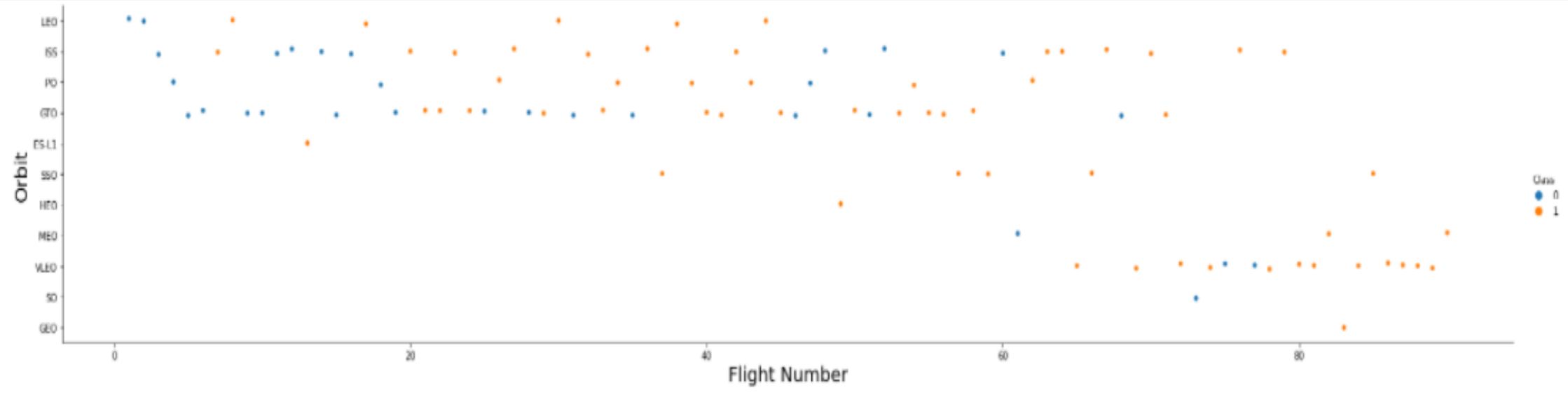
- Best launching site is CCAFS SLC 40, with loads bigger than 12000kg.
- VAFB SLC 4E has a successful launch at different load ranges, but we still need more than one launch to build a stronger decision.
- KSC LC 39A we see high success launches with loads smaller than 6000kg.

Success Rate vs. Orbit Type

- Here we can notice that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

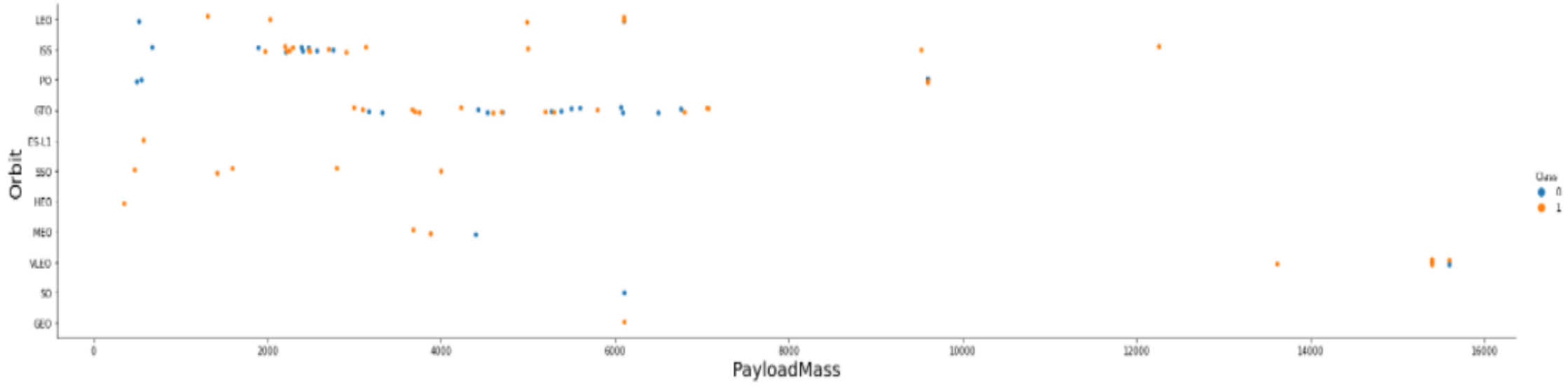


Flight Number vs. Orbit Type



- The plot (Flight Number vs. Orbit type). We notice that LEO orbit, success is related to the number of flights, while GTO orbit there is no relationship between flight number and the orbit.

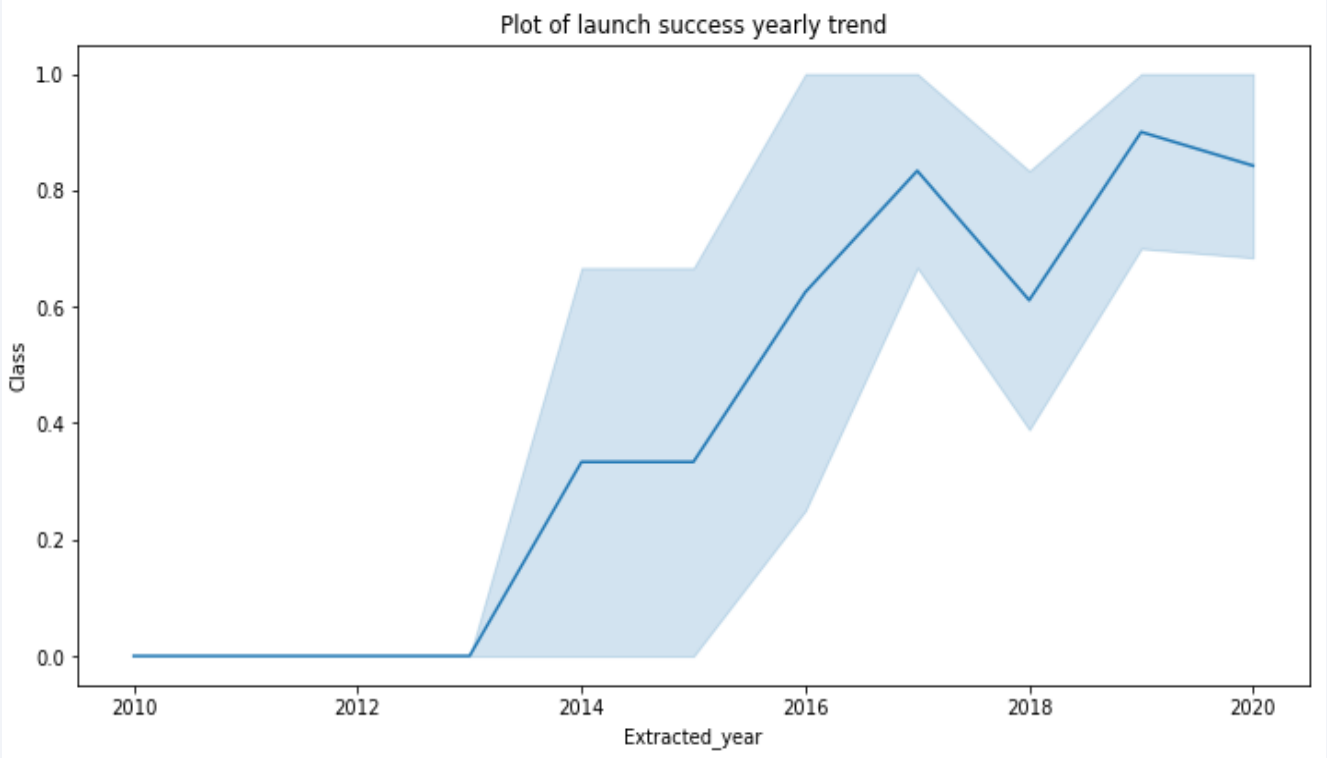
Payload vs. Orbit Type



- We notice that with heavier payloads, the successful landing rate gets higher in PO, LEO and ISS orbits.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and till 2020; the end of our dataset values



All Launch Site Names

- After running the query we have the list of launch sites
- They are obtained by selecting unique occurrences of “launch site” values from the dataset.

Out[10]:

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

Launch Site Names Begin with 'CCA'

Out[11]:

	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

- 5 records where launch sites begin with `CCA` and we can notice that all 5 launches were successful

Total Payload Mass

- Total payload calculated, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA

- So the total payload mass result was 45596kg.

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

In [12]:
```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

Out[12]:

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

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

In [13]:
```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

Out[13]:

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

- Looking after dates dates of the first successful landing outcome on ground pad
- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

In [14]:
```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

Out[14]:

	firstsuccessfull_landing_date
0	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

In [15]:
```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculating the total number of successful and failure mission outcomes
- We can notice that number of failed outcome is 1
  number of success outcome is 100

List the total number of successful and failure mission outcomes

In [16]:
```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

Out[16]:

The total number of successful mission outcome is:

	successoutcome
0	100

The total number of failed mission outcome is:

	failureoutcome
0	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
- Boosters which have carried the maximum payload mass registered in the dataset and we can notice that the max load is 15600kg for all.

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

In [17]:
```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Results shows 2 failed landings

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

In [18]:
```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- 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
- As we notice that including failure and success attempts, we find (10 No attempt) as well, which could be a planned launch but never been launched.

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))

In [19]:
```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

Out[19]:

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

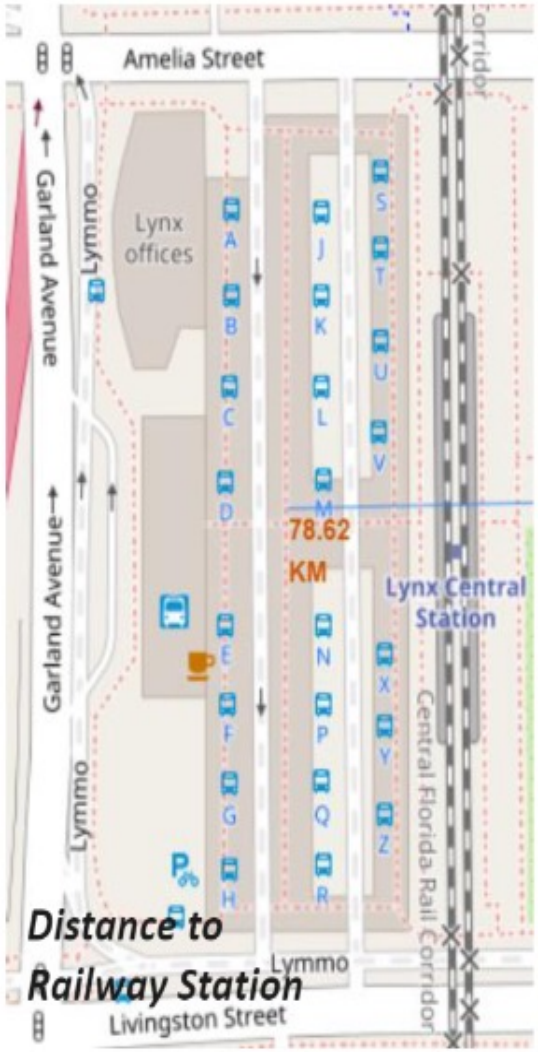

Section 3
Launch Sites
Proximities Analysis

All launch sites global map markers

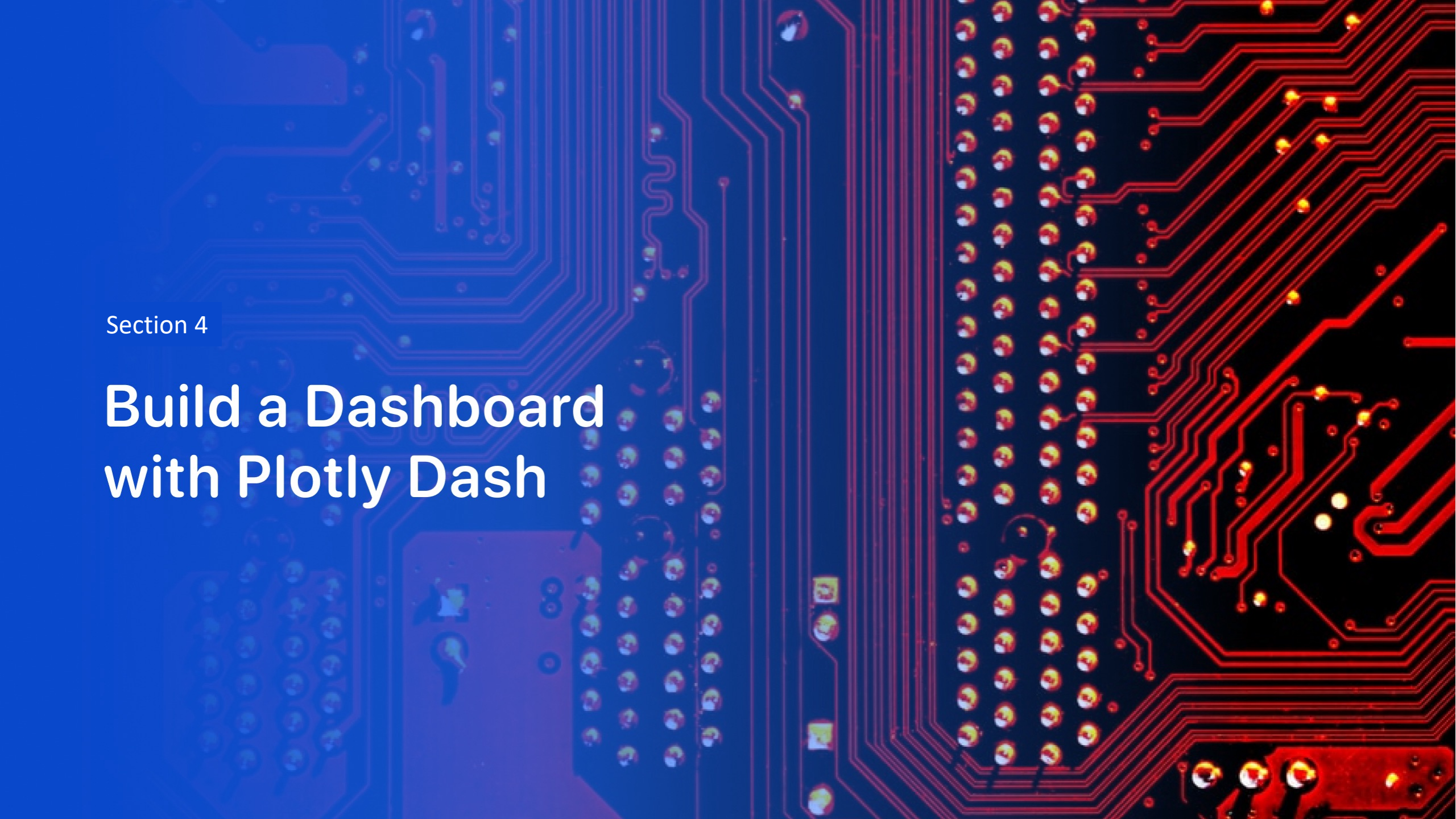
We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

Markers showing launch sites with color labels

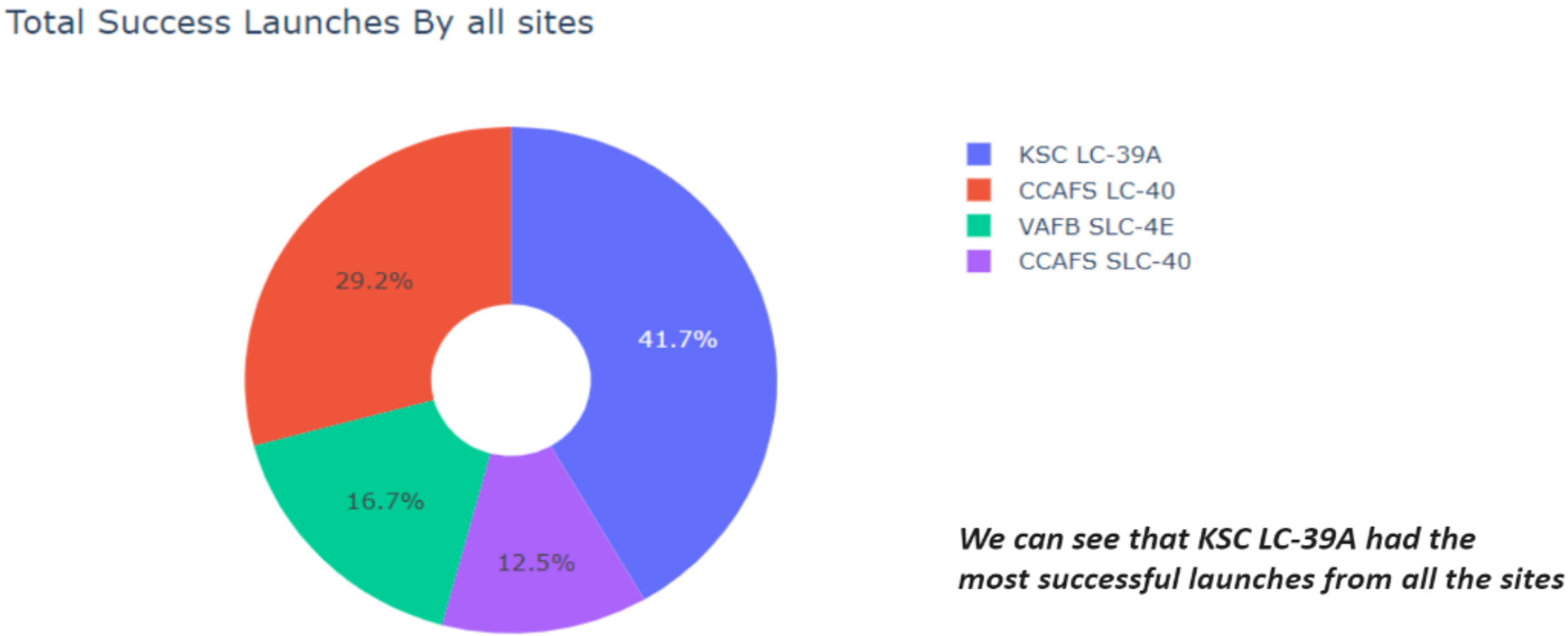
Launch Site distance to landmarks



- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

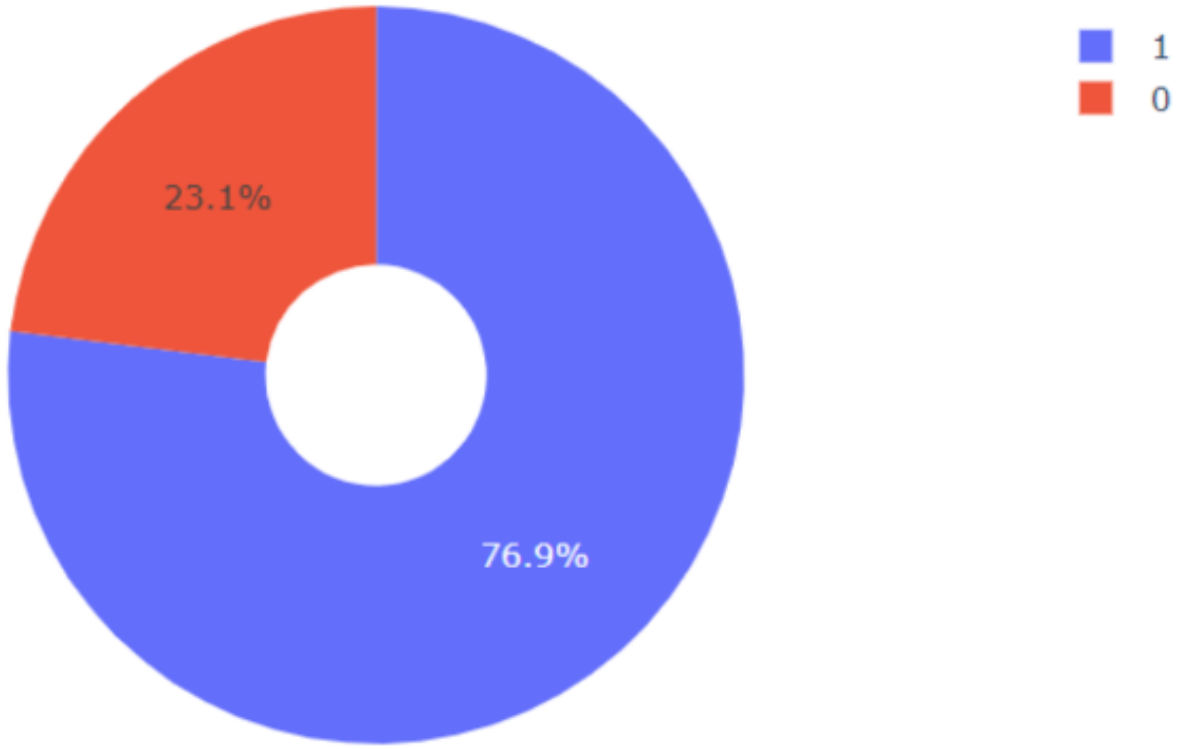

Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



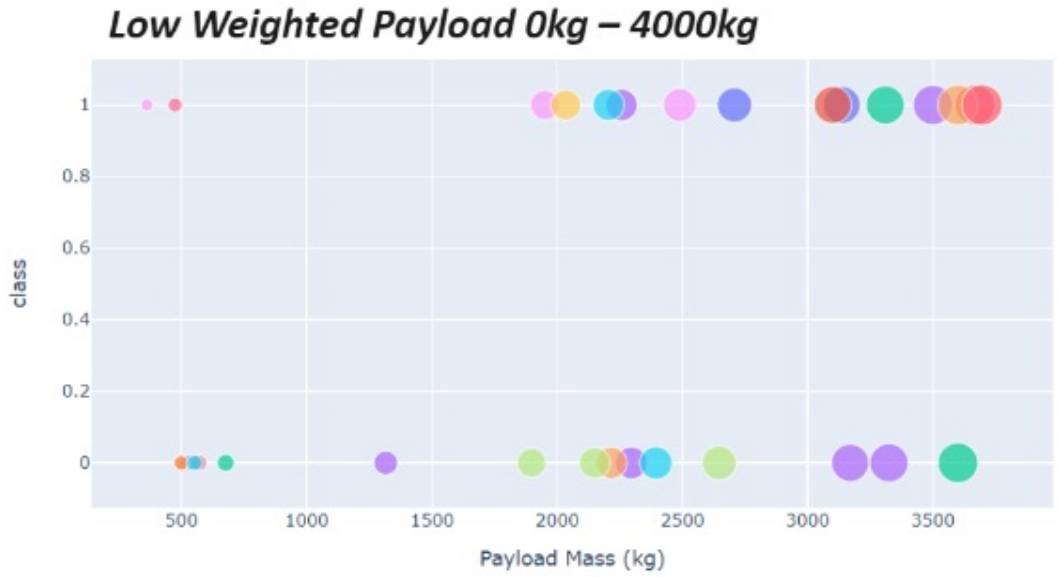
We can see that KSC LC-39A had the most successful launches from all the sites

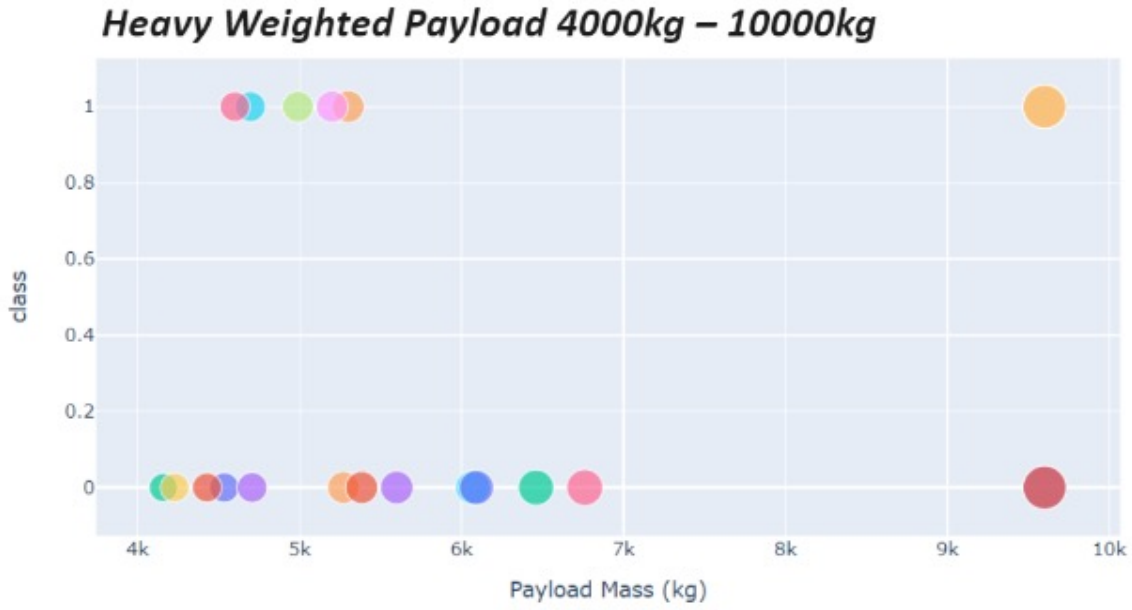
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

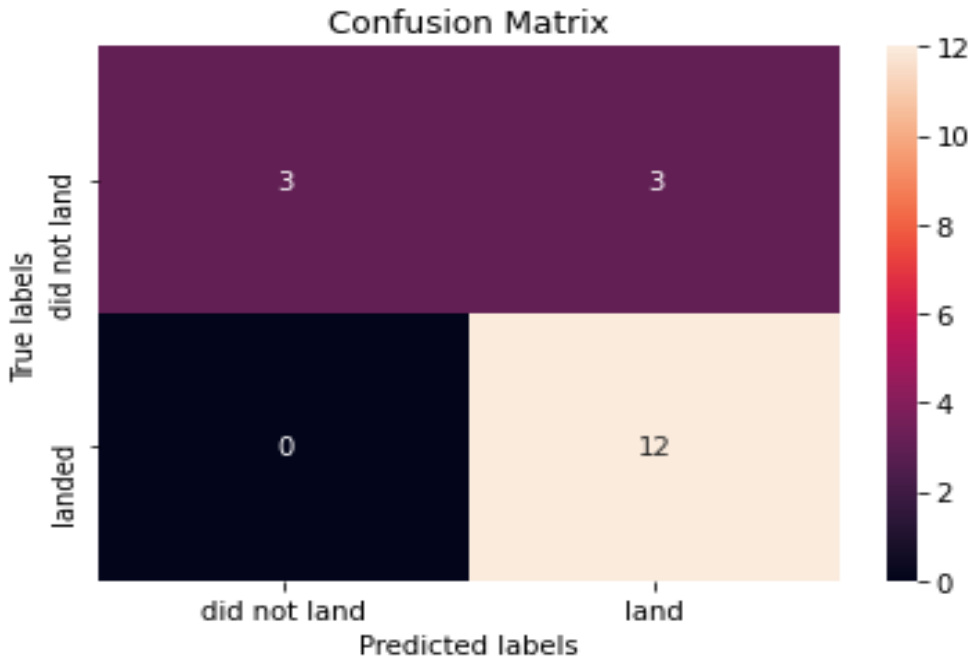
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

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_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

- Four classification models were tested (Decision Tree, Kneighbors, Logistic Regression, Support Vector)
- Decision Tree Classifier is the highest classification accuracy , with an accuracy over than 87%.

Confusion Matrix

- The confusion matrix for the decision tree classifier proves its accuracy, it shows big numbers of true positive and true negative compared to the false ones.



Conclusions

We can conclude that:

- The best launch site is KSC LC-39A.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- The Decision tree classifier is the most suitable machine learning algorithm for this task.
- Launches above 7,000kg are less risky;

Thank you!