IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rahul Thakur
27-Feb-2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Summary of all results

- EDA Results
- Interactive analytics
- Predictive analysis

Introduction

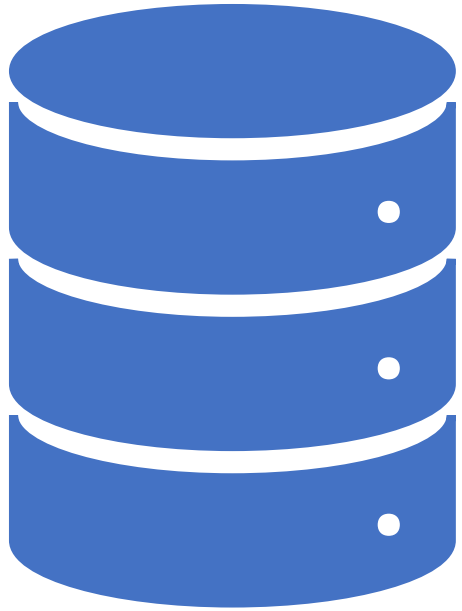
- **Project background and context**
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; Other providers cost upwards of 165 million dollars each, and much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. The goal of the project is to create a machine-learning pipeline to predict if the first stage will land successfully
- **Problems you want to find answers**
  - The project task is to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully
  - What are the variables that the landing depends upon

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and data cleaning of null values and irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, KNN, SVM, DT models have been built and evaluated for the best classifier.



Data Collection

Data was collected using the SpaceX API

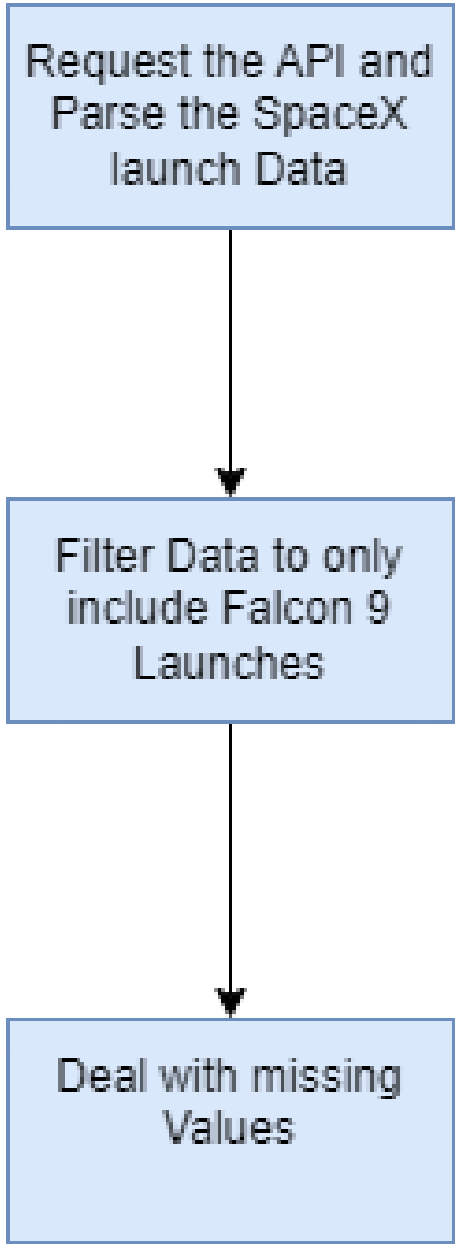
We converted the data into a data frame, using pandas.

We then cleaned the data, checked for any abnormal, missing/NaN values, and filled the values.

We performed web scraping from Wikipedia for the Falcon 9 launch Records using the Python module called BeautifulSoup.

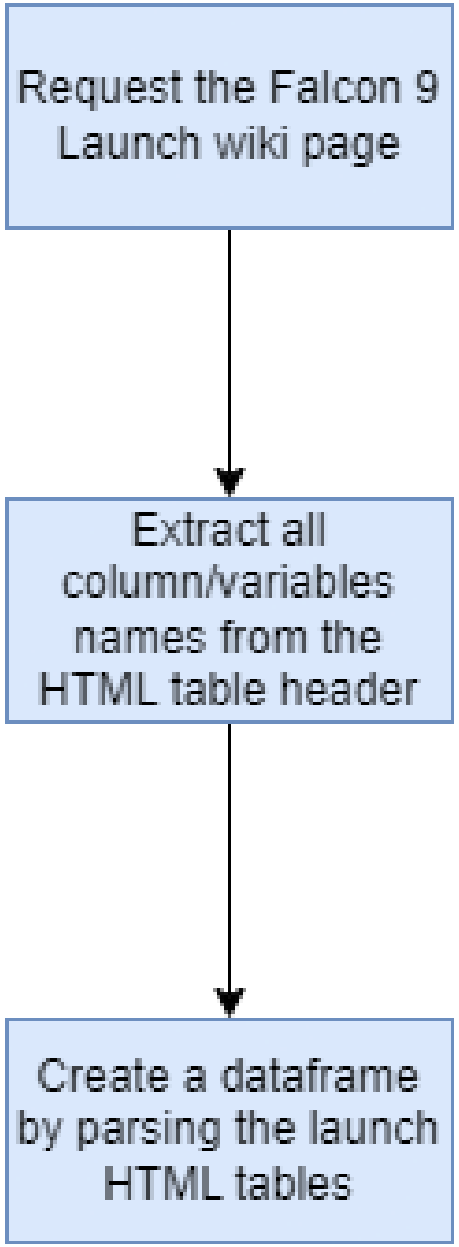
Data Collection – SpaceX API

- We collected the data using the SapceX API and the preprocessed the data(i.e clean, formatting and basic data wranging)
- The link to the notebook is https://github.com/therealzykrix/IBM-Capstone-Project/blob/b00d83c01a541f832465de933535d3049cd0d7c1/Data%20Collection.ipynb



Data Collection - Scraping

- We used webcrapping techniques such as BeautifulSoup to get the Falcon 9 Launch Records
- We then parsed the tables and converted it into a dataframe using Pandas.
- The link to the notebook is https://github.com/therealzykrix/IBM-Capstone-Project/blob/b00d83c01a541f832465de933535d3049cd0d7c1/Data%20Collection%20with%20Web%20Scaping.ipynb

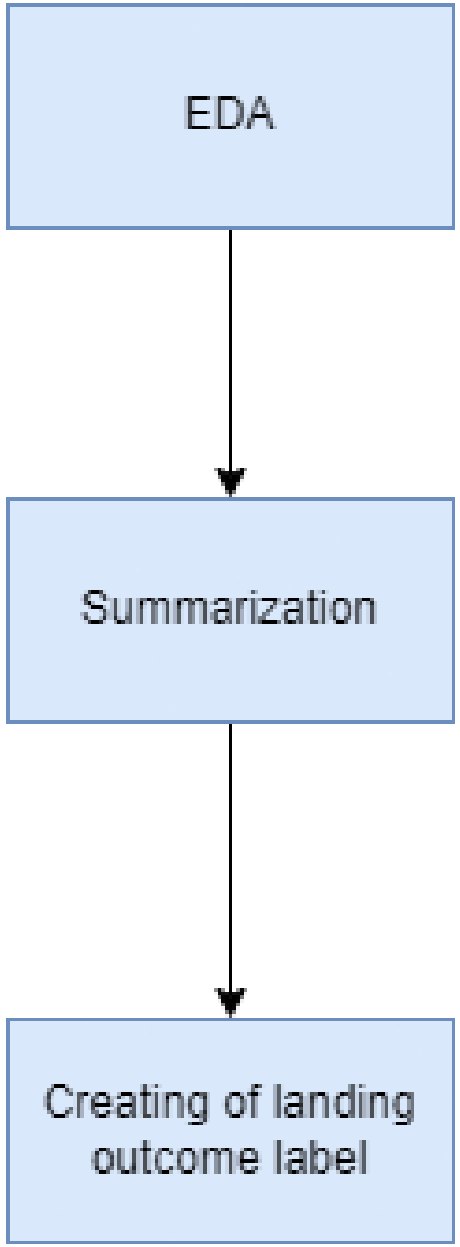


Data Wrangling

Using exploratory Data Analysis techniques, we determined the training Labels.

calculated the number of launches at each site, and the number and occurrence of each orbits • We created landing outcome label from outcome column and exported the results to csv

The link to the notebook is https://github.com/therealzykrix/IBM-Capstone-Project/blob/b00d83c01a541f832465de933535d3049cd0d7c1/Data%20Wrangling.ipynb



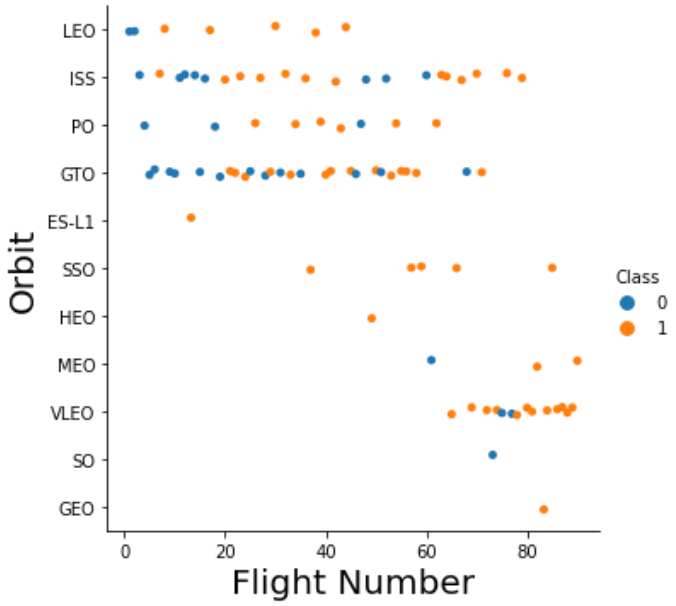
EDA with Data Visualization

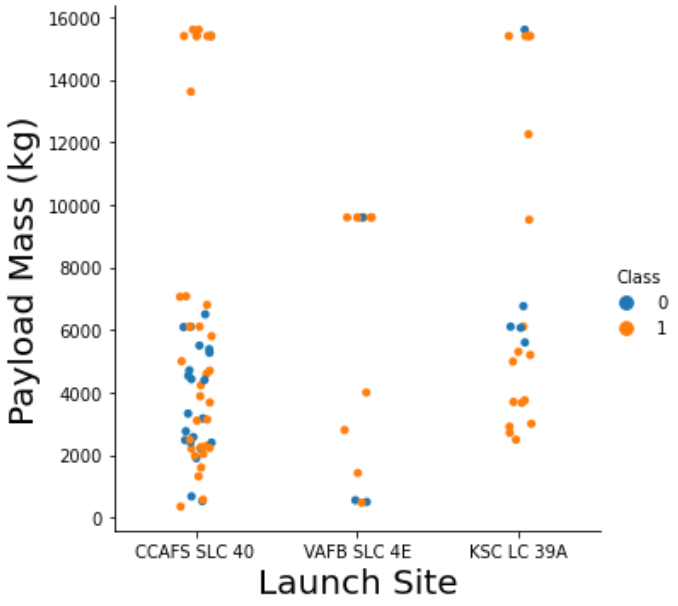
- I used scatterplots to visualize the relationship between flight number and launch site, payload and launch site, flight number and orbit type, and payload and orbit type.

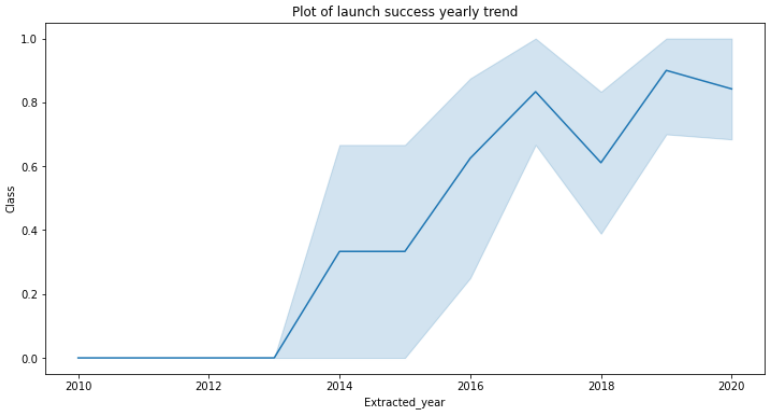
- I used a bar chart to visualize the relationship between the success rate of each orbit type.
-
- Line plot to visualize the launch success yearly trend.

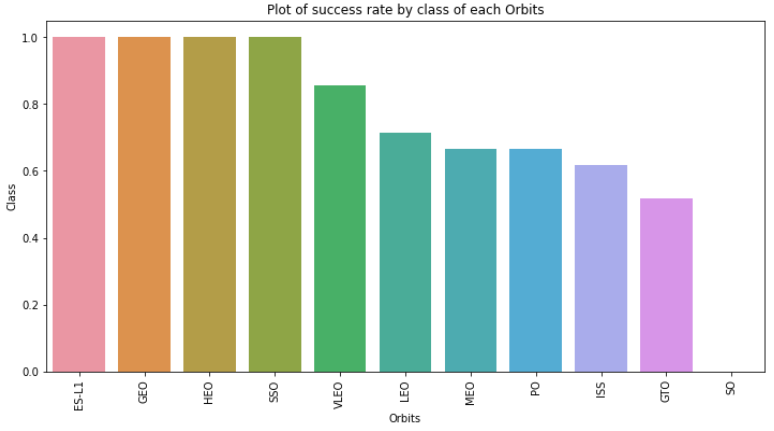
- Link: https://github.com/therealzykrix/IBM-Capstone-Project/blob/b00d83c01a541f832465de933535d3049cd0d7c1/EDA%20with%20Data%20Visualization.ipynb

EDA with Data Visualization









EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in the ground pad was achieved.
- List the names of the boosters which have success in drone ships and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failed mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records that will display the month names, failure landing_outcomes in drone ship, booster versions, and launch_site for the months in the year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order.
- Link: https://github.com/therealzykrix/IBM-Capstone-Project/blob/b00d83c01a541f832465de933535d3049cd0d7c1/EDA%20WITH%20SQL.ipynb

Build an Interactive Map with Folium

- We marked the launch sites
- We added map objects like markers, circles, and lines to mark the success or failure of launches on each location
- We assigned classes to launches to distinguish between failed and successful(1= success and 0 = fail)
- We used color labels to identify which launch site has a relatively high success rate.
- We also answered some questions:
  - Are launch sites near railways, highways, and coastlines?
  - Do launch sites keep a certain distance away from
- Link: https://github.com/therealzykrix/IBM-Capstone-Project/blob/b00d83c01a541f832465de933535d3049cd0d7c1/Launch%20Site%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- A launch site drop-down input component
- A success pie chart based on the selected site dropdown
- A range slicer to select payload
- A success-payload-scatter-chart scatter plot based on the selected site dropdown

- Link: https://github.com/therealzykrix/IBM-Capstone-Project/blob/3d74492432869092ce407d52b3219d1dce70eb88/app.py

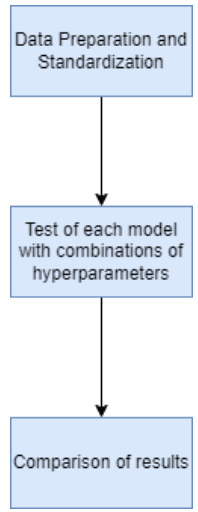
Predictive Analysis (Classification)

- We used 4 classifiers: Decision Tree, Logistic, Regression, Support Vector Machines, and K Nearest Neighbour to check which of these models gives us the highest accuracy.
- We tuned different hyperparameters using GridSearchCV
- We found that Decision Tree Classifier had the best accuracy
- Link: https://github.com/therealzykrix/IBM-Capstone-Project/blob/b00d83c01a541f832465de933535d3049cd0d7c1/Machine%20Learning.ipynb

TASK 12

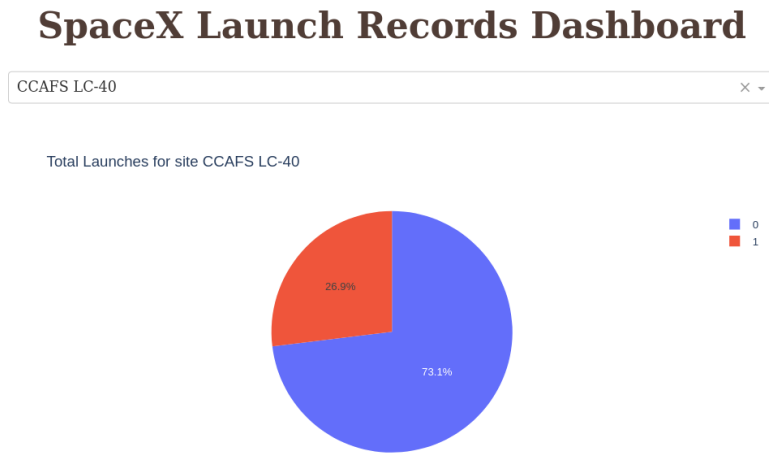
In [26]:
```
models = {'KNeighbors':knn_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'LogisticRegression':logreg_cv.best_score_,
          'SupportVector': svm_cv.best_score_}

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_)
```

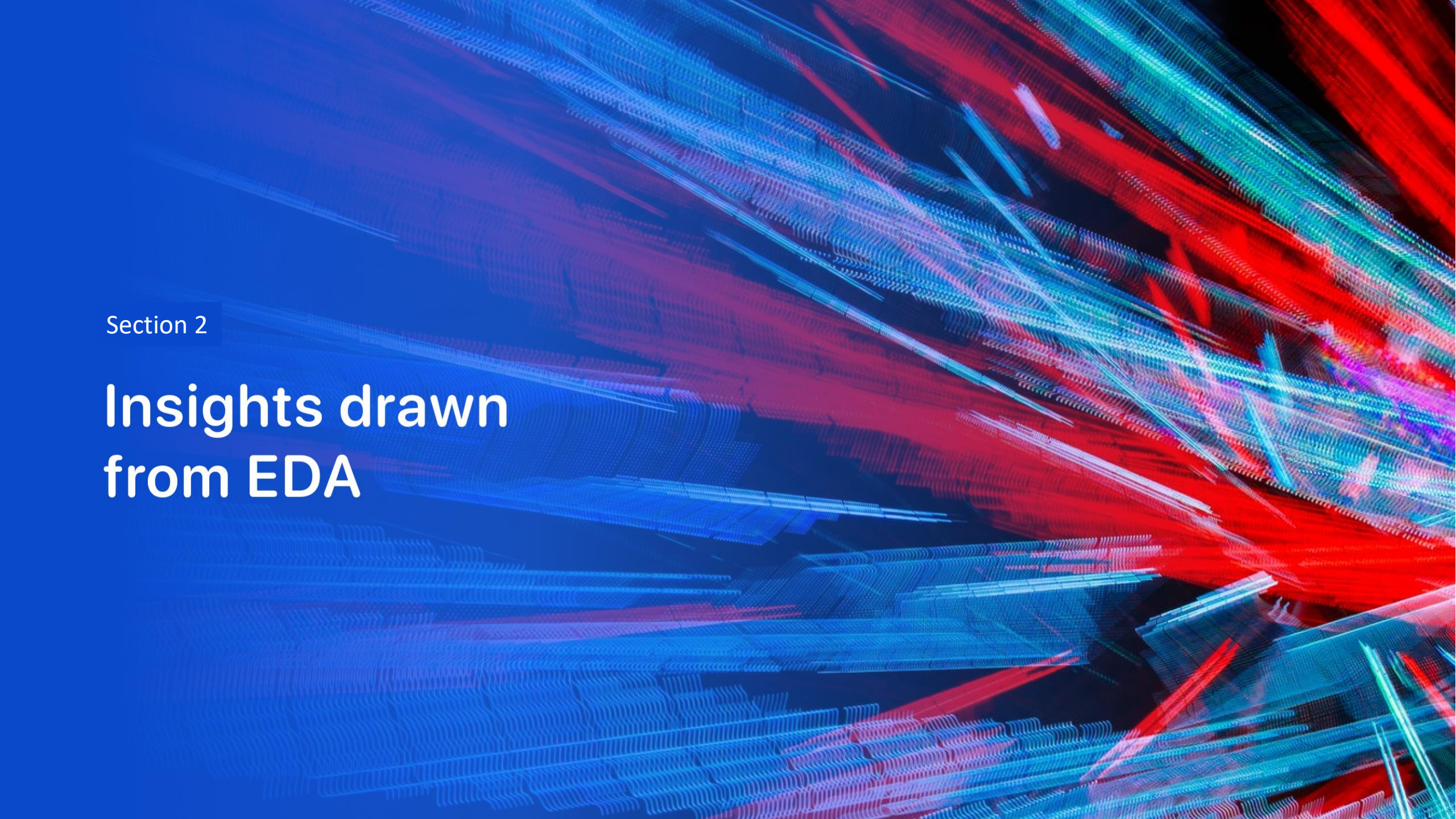
```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split':
5, 'splitter': 'random'}
```



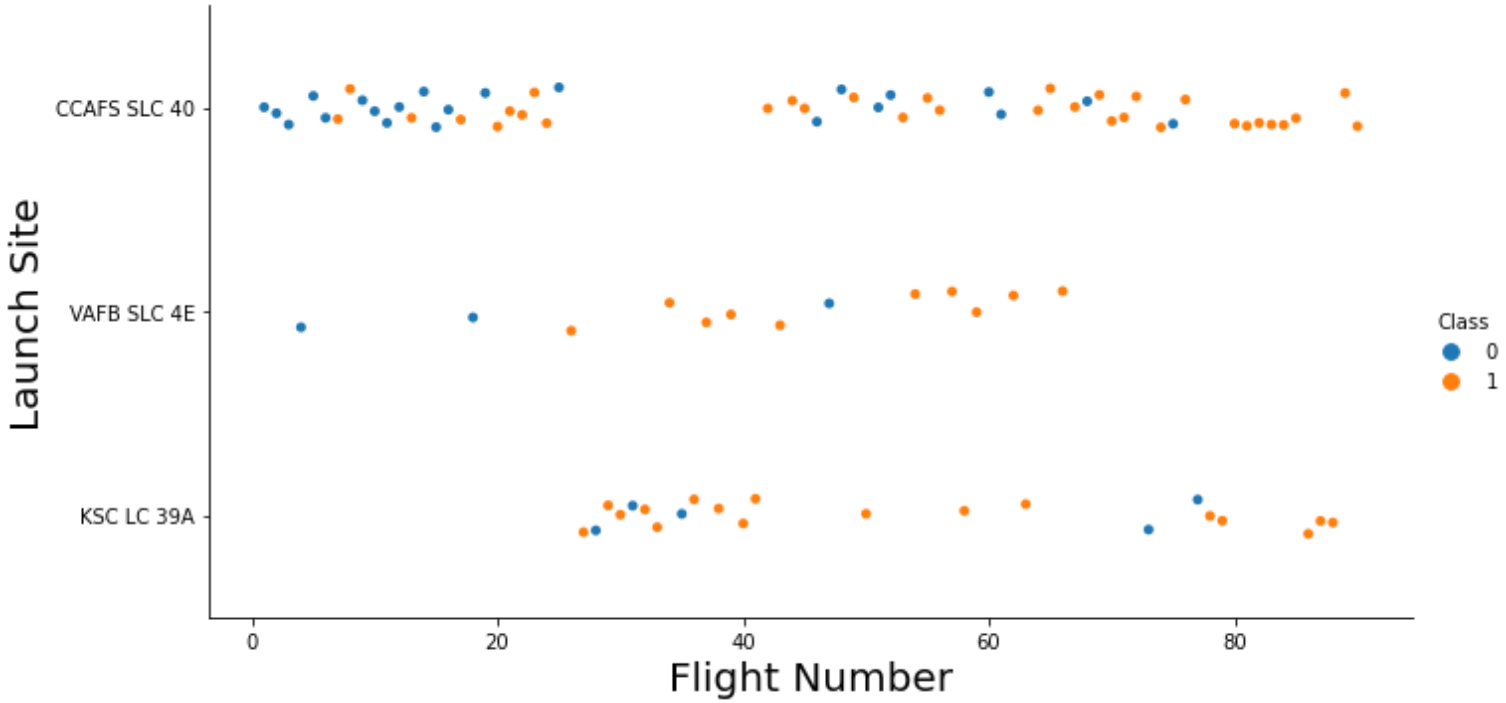
Results

- We saw that SpaceX uses 4 different launch sites.
- The first successful landing outcome happened in 2015 (5 years after the first launch)
- The number of landing outcomes gradually increased over the years.



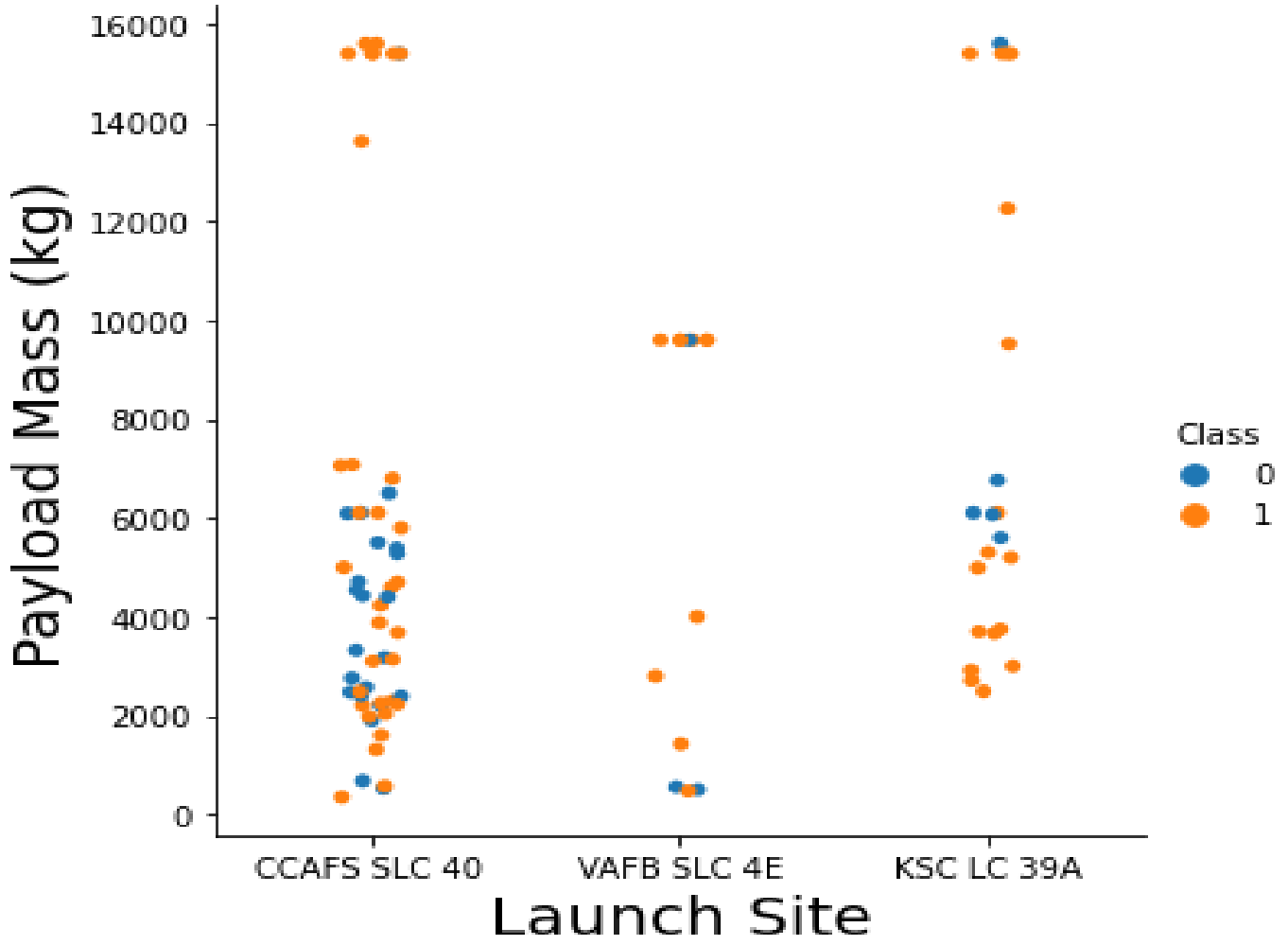

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



According to the plot, we can say that the most successful launch site is CCAF5 SLC 40

Payload vs. Launch Site

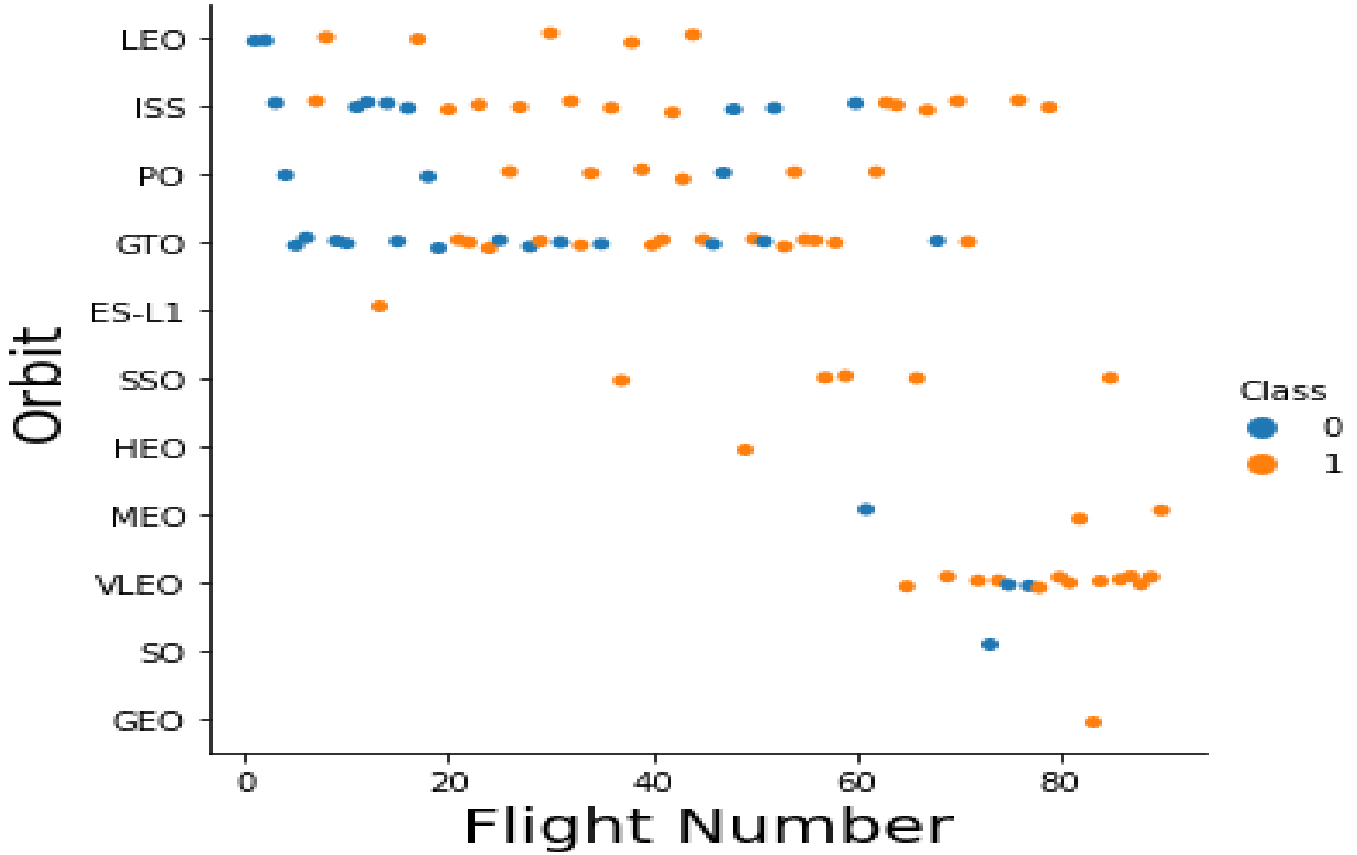


- As we can see the success rate increases as we increase the weight of the payload.
- We can also infer from the plot that heavier payloads can be launched only from CCAFS SLC 40 and KSC LC 39A

Success Rate vs. Orbit Type

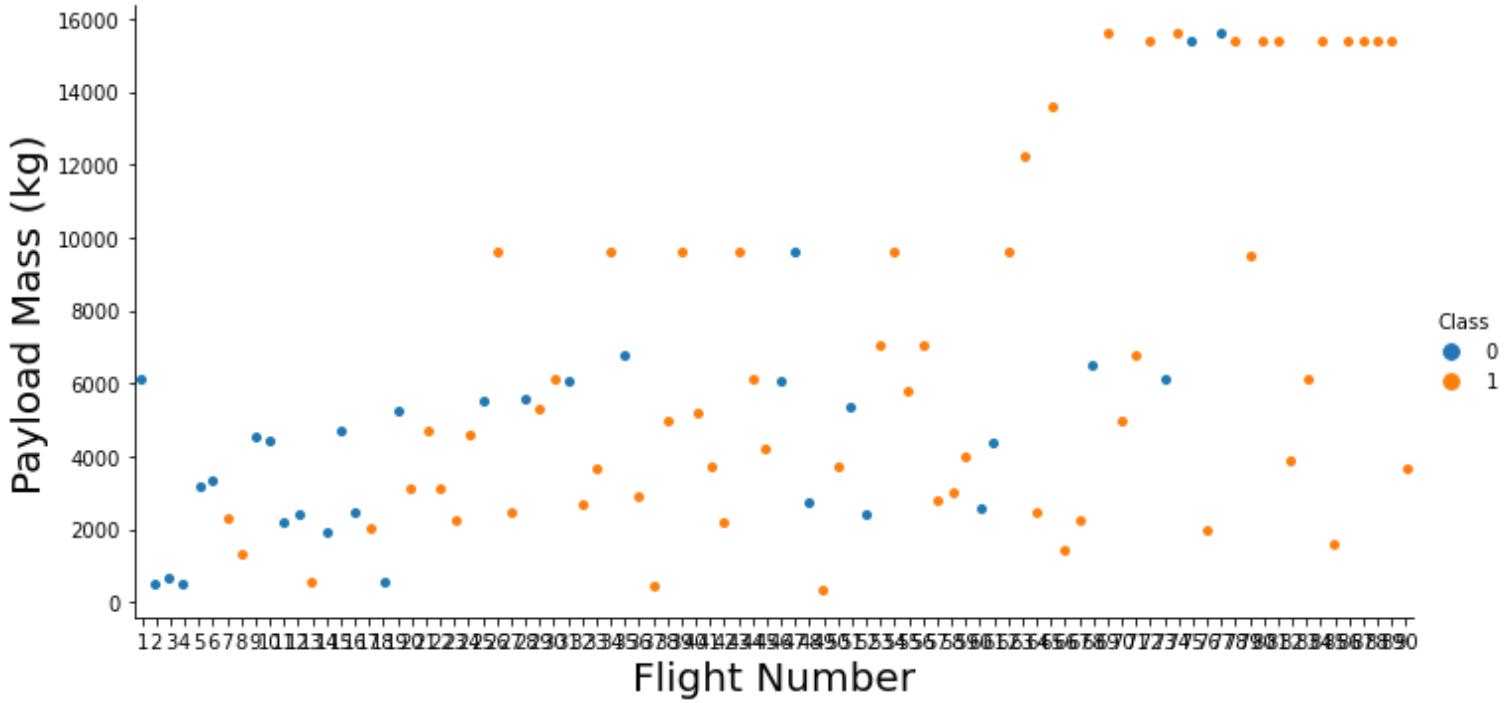
ES-L1, GEO, HEO and SSO have the highest success rates

Flight Number vs. Orbit Type



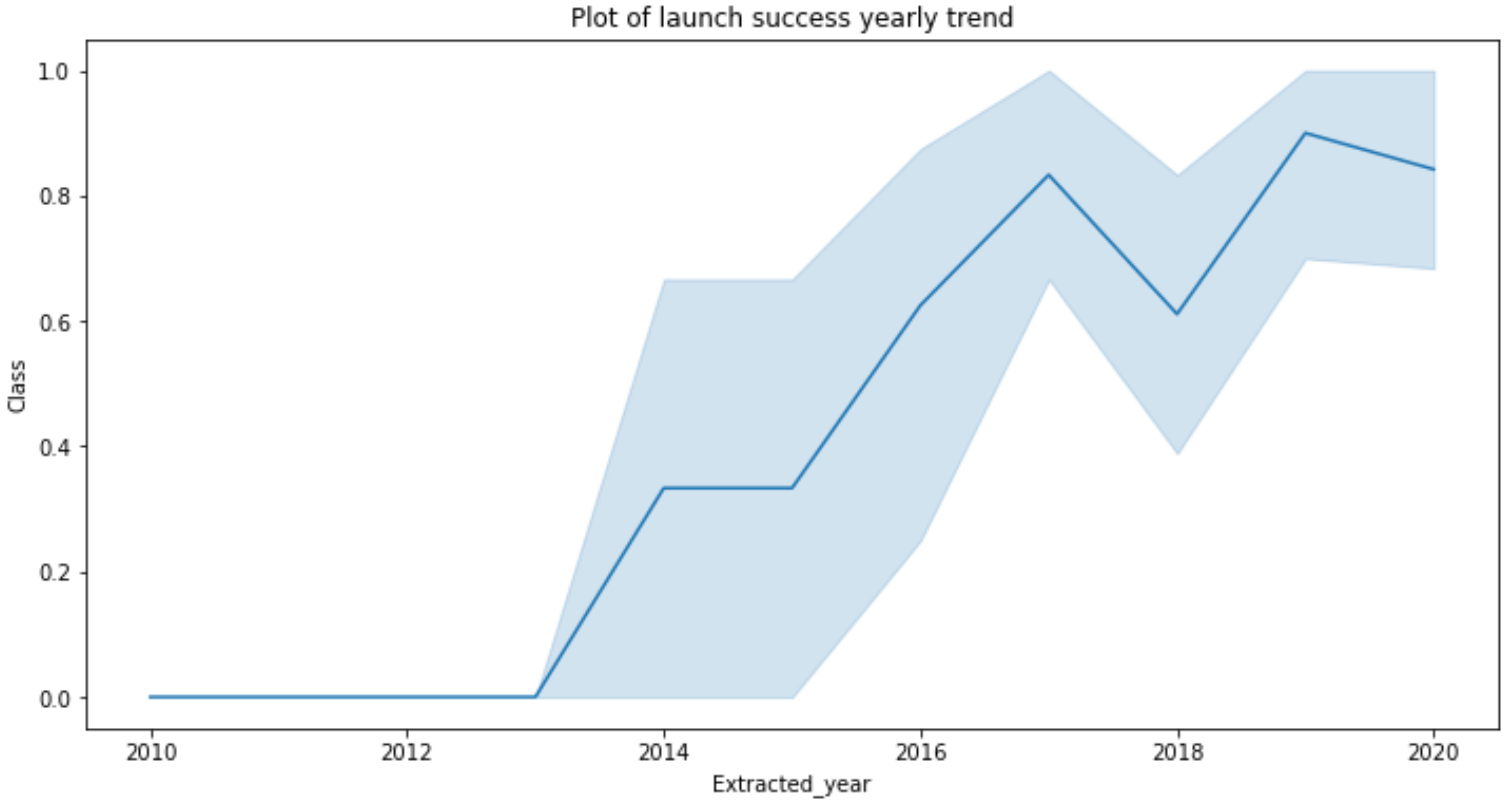
- We can that the success rates have improved over time in all the orbits
- VLEO has been the preferred orbit in recent missions

Payload vs. Orbit Type



- ISS orbit has the widest range of payload and a good rate of success
- We cannot infer anything regarding the relation between the payload and the success rate in the GTO

Launch Success Yearly Trend



We can see the success rate has been on a steady increase since 2013

All Launch Site Names

Task 1

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

In [9]:
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

Out[9]:

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

According to the data, there are 4 launch sites.

Launch Site Names Begin with 'CCA'

Task 2

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

In [28]:
```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Out[28]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- There are 5 records that start with ‘CCA’
- We can see that all the outcomes were failure

Total Payload Mass

Task 3

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

In [60]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

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

Out[60]:

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

The total payload mass for NASA is 45,596 kg

Average Payload Mass by F9 v1.1

Task 4

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

In [61]:
```
%sql SELECT avg(PAYLOAD_MASS__KG_) AS Avg_Payload FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1';
```

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

Out[61]:

Avg_Payload
2928.4

The average payload mass carried by booster version F9 v1.1 is 2,928.40 kg

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

In [19]:
```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

Out[19]:

FIRST_SUCCESS_GP
2015-12-22

The first ground landing successful is on 01.05.2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [58]:
```
%sql SELECT DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)" AND PAYLOAD_MA
```

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

Out[58]:

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

The most successful landing is by drone ship.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [26]:
```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Out[26]:

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

We can see that there is 1 failure, 1 success with unclear payload status and 99 successful mission outcome

Boosters Carried Maximum Payload

Task 8

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

In [50]:
```
%sql SELECT Booster_Version, Max_Payload FROM (SELECT Booster_Version, MAX(PAYLOAD_MASS__KG_) AS Max_Payload FROM SPACEXTBL
```

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

Out[50]:

Booster_Version	Max_Payload
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092
F9 B4 B1045.1	362
F9 B4 B1045.2	2697
F9 B5 B1046.1	3600
F9 B5 B1046.2	5800
F9 B5 B1046.3	4000
F9 B5 B1046.4	12050
F9 B5 B1047.2	5300
F9 B5 B1047.3	6500
F9 B5 B1048.2	3000

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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [56]:
```
sql SELECT substr(Date,0,5), substr(Date, 4, 2),Booster_Version, Launch_Site, Payload, PAYLOAD_MASS__KG_, Mission_Outcome, l
```

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

Out[56]:

substr(Date,0,5)	substr(Date, 4, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing_Outcome
2015	5-	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	Success	Failure (drone ship)
2015	5-	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	Success	Failure (drone ship)

- The months that had launch failures were January and April.
- Booster Versions were B1012 and B1015

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

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

In [23]:
```
sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDINC
```

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

Out[23]:

Landing_Outcome	QTY
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

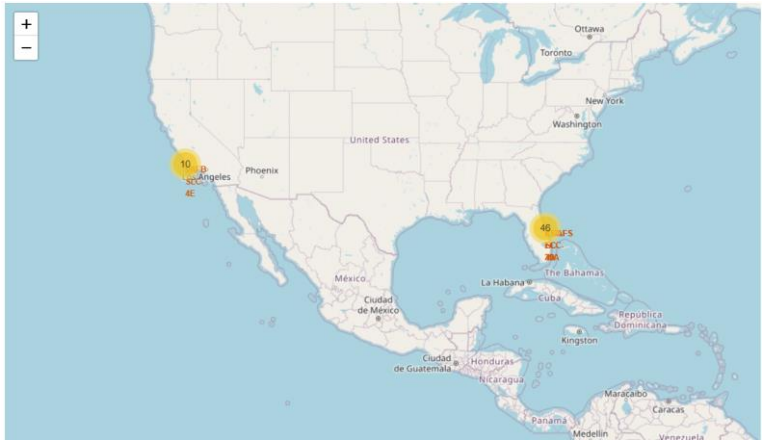
Between 4th June 2010 and 20th March 2017, there were 31 landings out of which 10 were No attempt, 5 were Successes (drone ship), 5 were Failures (drone ship), 3 were Successes (ground pad), 3 were Controlled (ocean), 2 were Uncontrolled (ocean), 2 were Failure (parachute), 1 was Precluded (drone ship).

Section 3

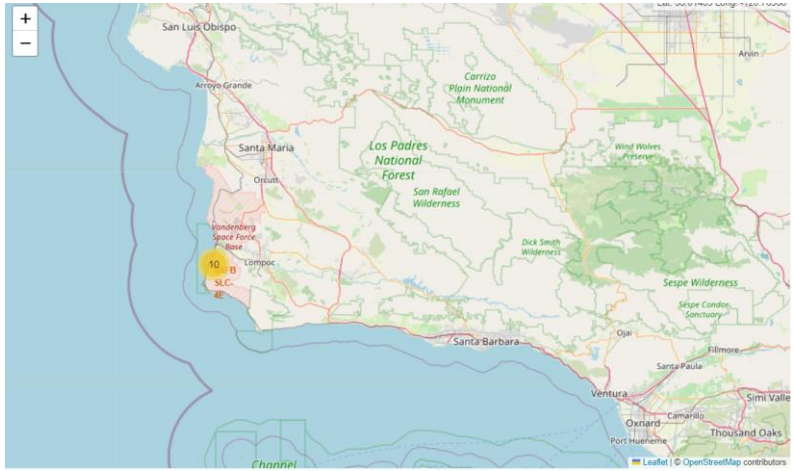
Launch Sites Proximities Analysis

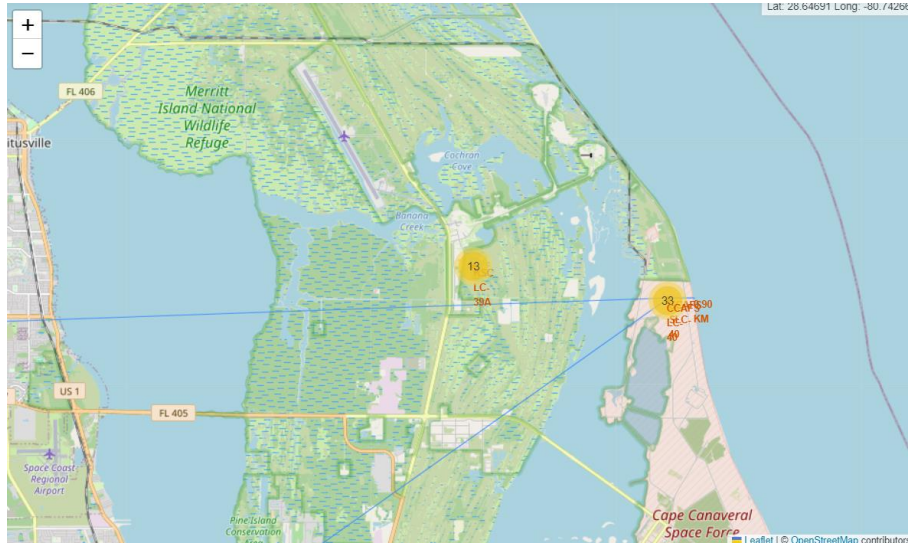
All Launch Sites

- All the launch sites are in proximity to the equator.
- All the launch sites are in proximity to the coast.

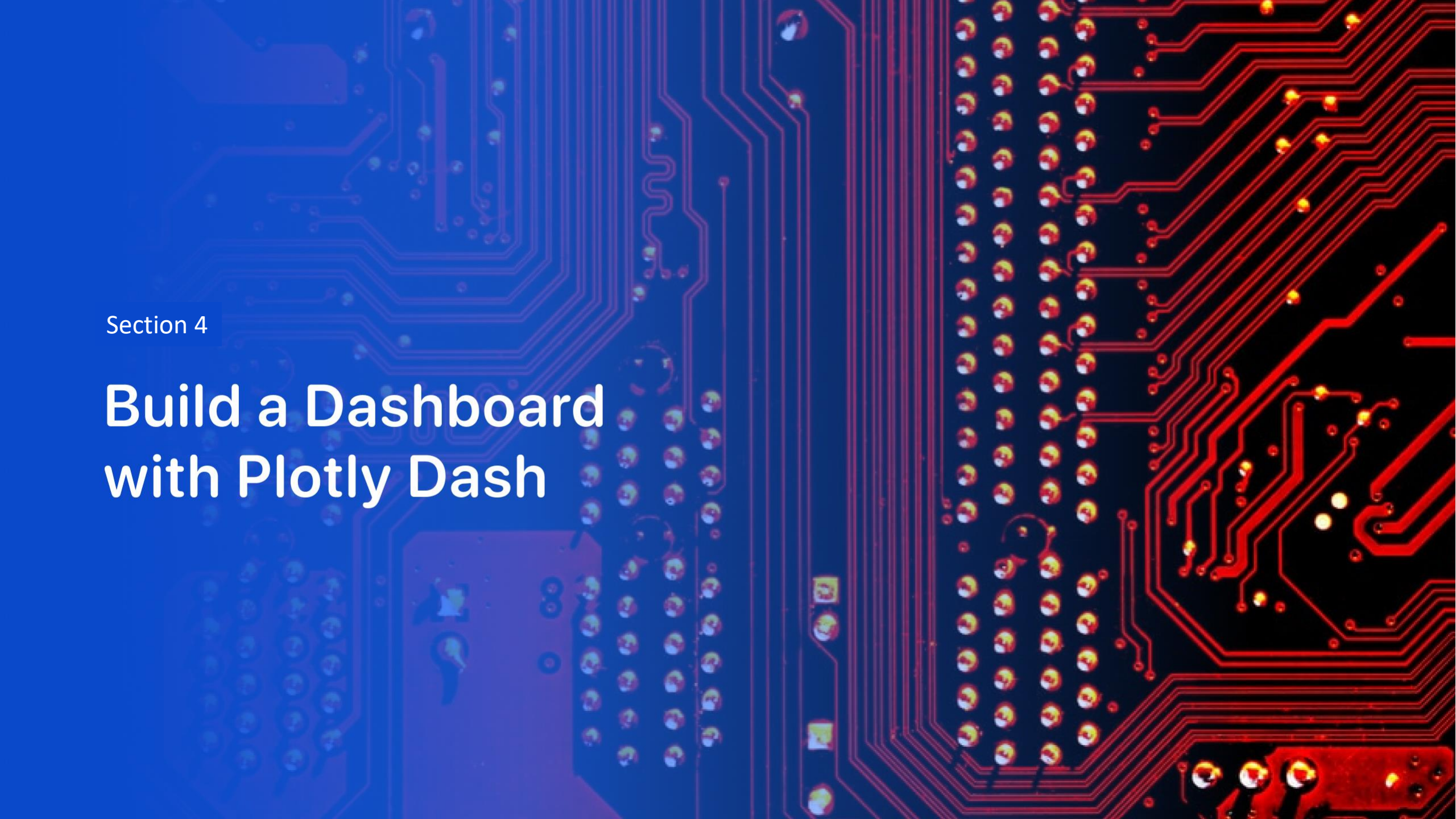


Launch Outcome by Site



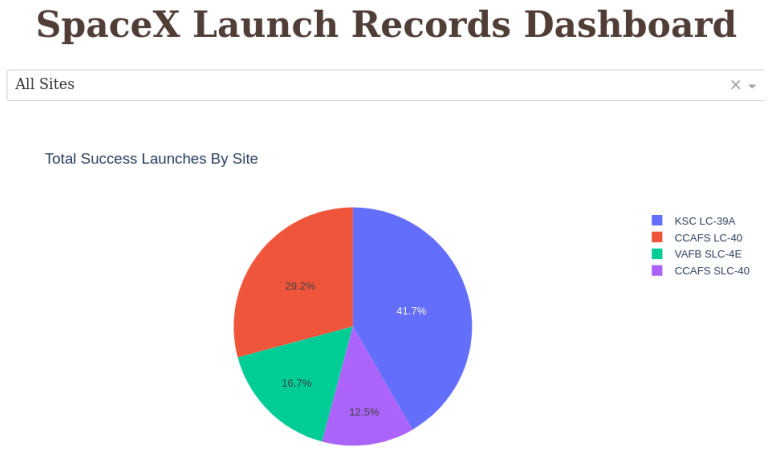


The Eastern Coast has more Launches than the Western Coast with 46 launches to 10 launches


Section 4
Build a Dashboard with Plotly Dash

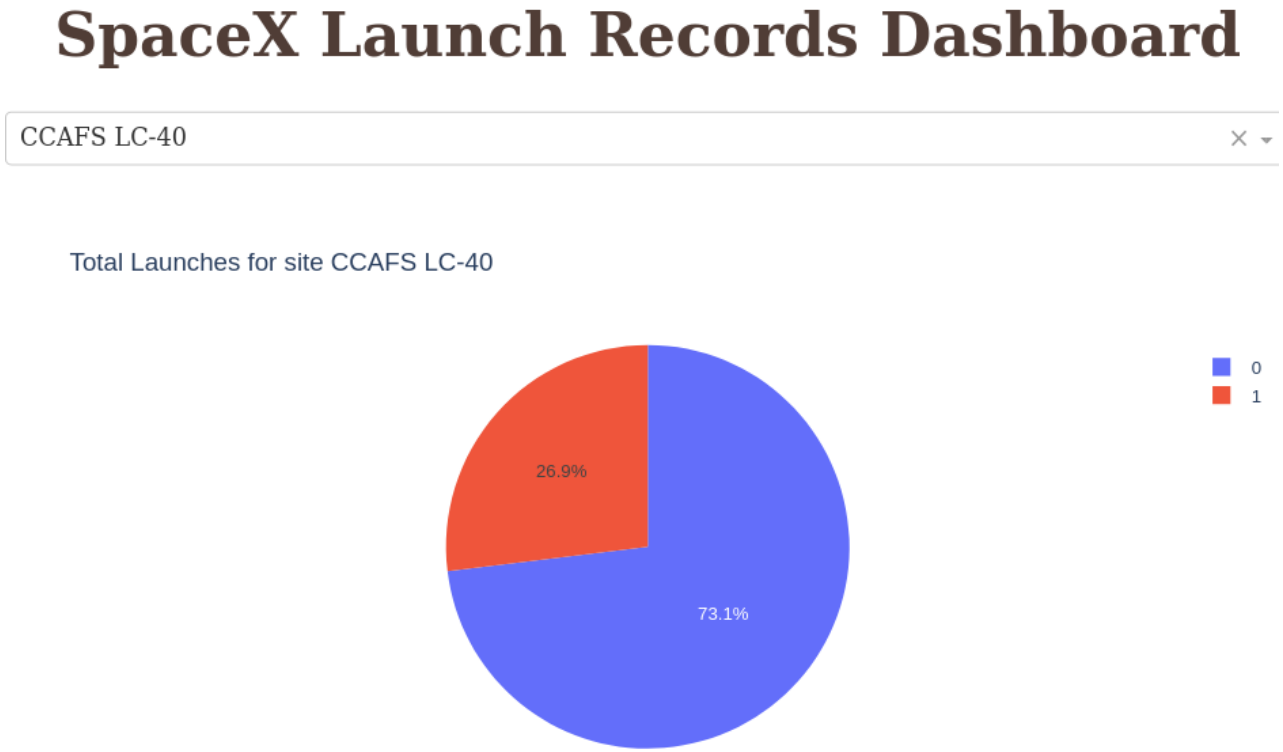
Pie-Chart for launch success count for all sites

- Launch site KSC LC-39A has the highest launch success rate at 42% followed by CCAFS LC-40 at 29%, VAFB SLC-4E at 17% and lastly launch site CCAFS SLC-40 with a success rate of 13%



Pie chart for the launch site with 2nd highest launch success ratio

- 76.9% of launches are successful in this site.



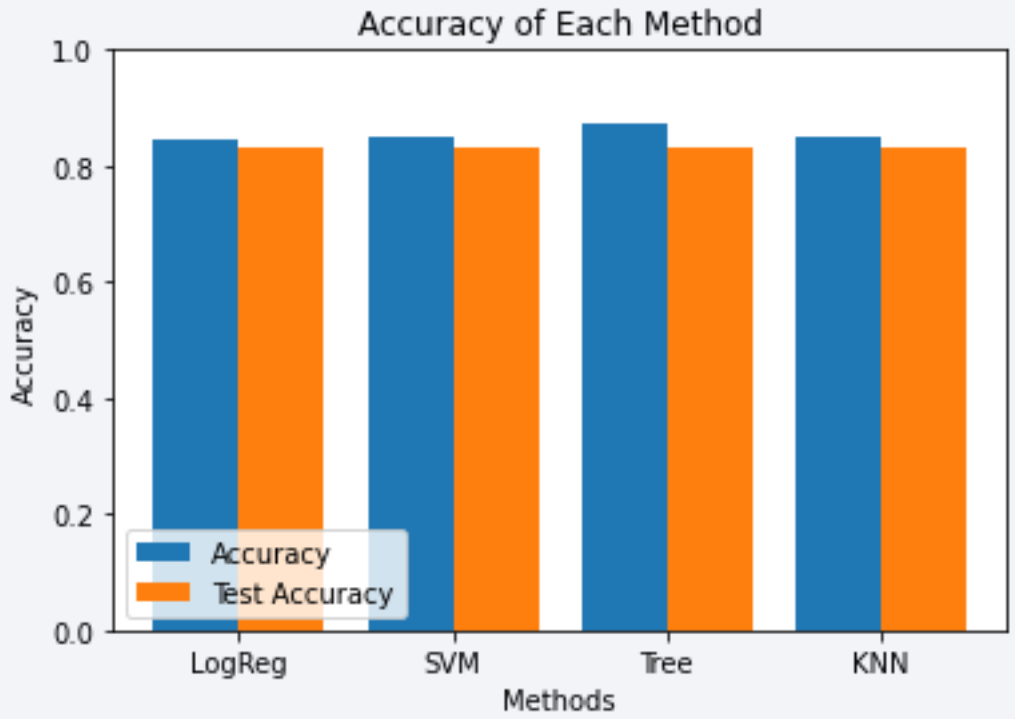
Payload vs. Launch Outcome scatter plot for all sites

- Payloads under 6,000kg and FT boosters are the most successful combination.

Section 5

Predictive Analysis (Classification)

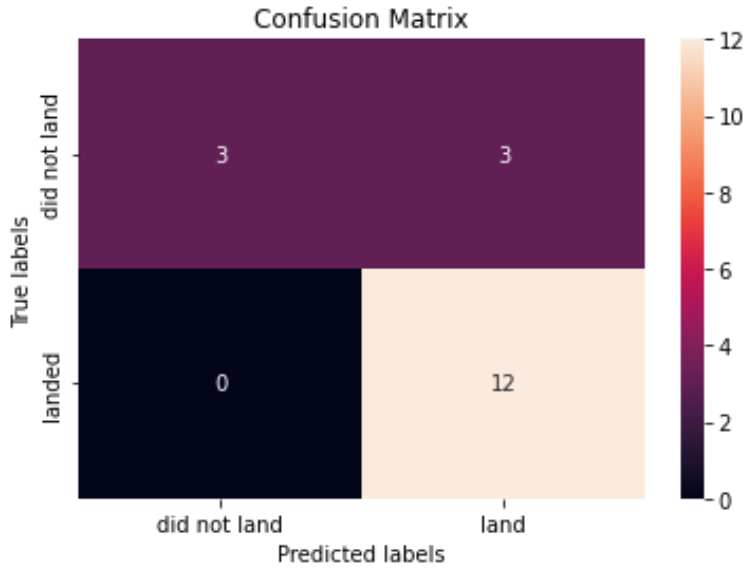
Classification Accuracy



The best model is the decision tree classifier because it has the highest classification accuracy.

Confusion Matrix

- All 4 classification models had the same confusion matrixes and were able to equally distinguish between the different classes. The major problem is false positives for all the models.



Conclusions

- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Payload of more than 8000 kgs seems to have more successful rate
- Mission outcomes have improved more time.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!