IBM Developer
SKILLS NETWORK

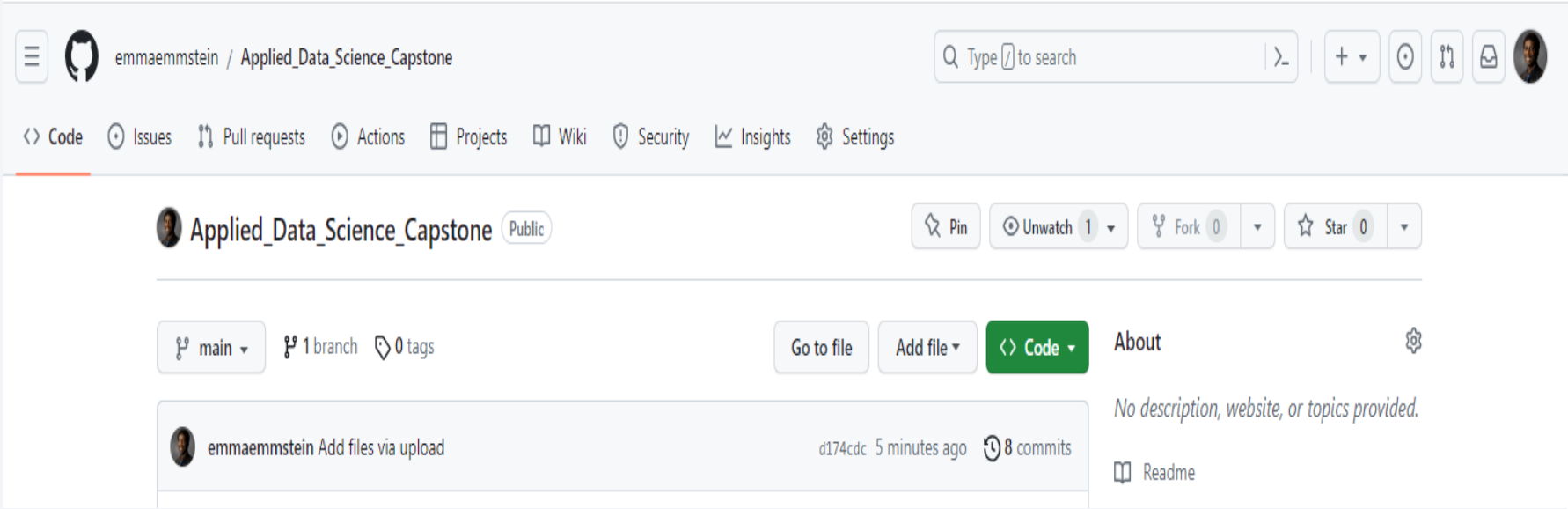
Winning Space Race with Data Science

Emmanuel Chukwuemeka

08/20/2023

Outline

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



https://github.com/emmaemmstein/Applied_Data_Science_Capstone

Executive Summary

- Summary of methodologies - Within this document, we provide a comprehensive data analysis of Space9's launch data, following the guidelines outlined in the "Applied Data Science Capstone" course by IBM / Coursera. During this examination:
  - Web scraping methods to get relevant data for the present analysis
  - We conducted initial analyses using direct SQL queries, along with charts and maps' visualizations.
  - We generated an interactive dashboard, enabling end users to conduct their own analysis.
  - We employed a range of ML Algorithms to forecast the likelihood of successfully recovering the rocket's initial stage.
- Summary of all results
  - Up to date data were gotten using the web scraping method
  - Data was then further cleaned for analyses
  - A dashboard was set up to help users analyze launch records interactively
  - The models in this report were able to predict an 83.3% success rate for the booster landing. Using this information competing companies can better adjust their cost predictions.

Introduction

- Project background and context

  On its website, SpaceX promotes Falcon 9 rocket launches at a price of 62 million dollars, a significantly lower figure compared to other providers whose charges exceed 165 million dollars per launch. This cost discrepancy mainly stems from SpaceX's ability to recycle the initial stage of the rocket. Consequently, by assessing the likelihood of a successful first stage landing, we can ascertain the overall launch cost.

- Problems you want to find answers

  We are tasked with using Data Science methods to predict if the first stage of a given launch will be recovered using past, public data, therefore predicting the launch cost. To do this we have to find key factors in first stage landing success or failure, and determine the accuracy of our prediction models.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

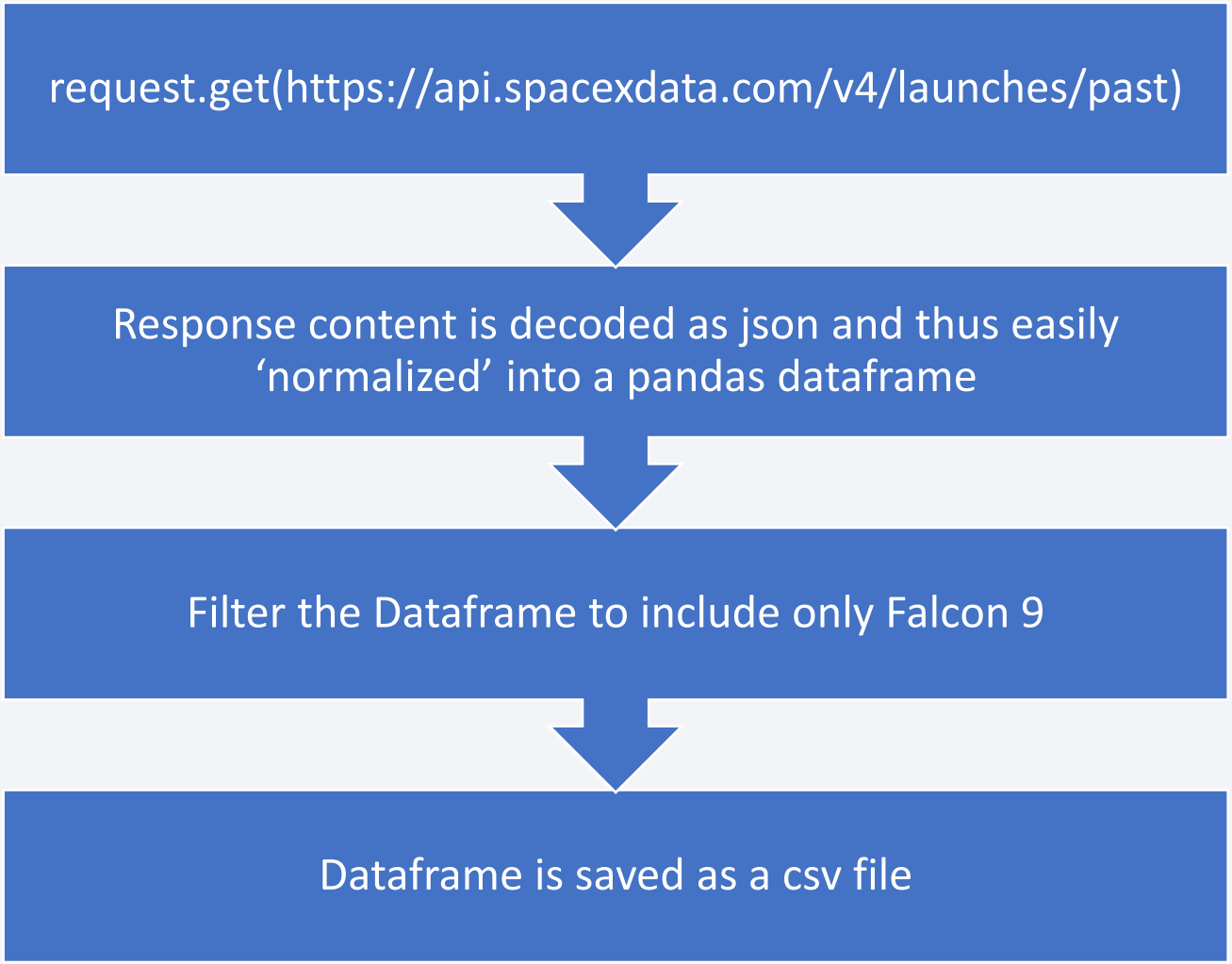
Data Collection

- Acquired rocket launch data from SpaceX API
  - Requested data using a GET
  - Filtered data into a dataframe to include only Falcon 9

Data Collection – SpaceX API

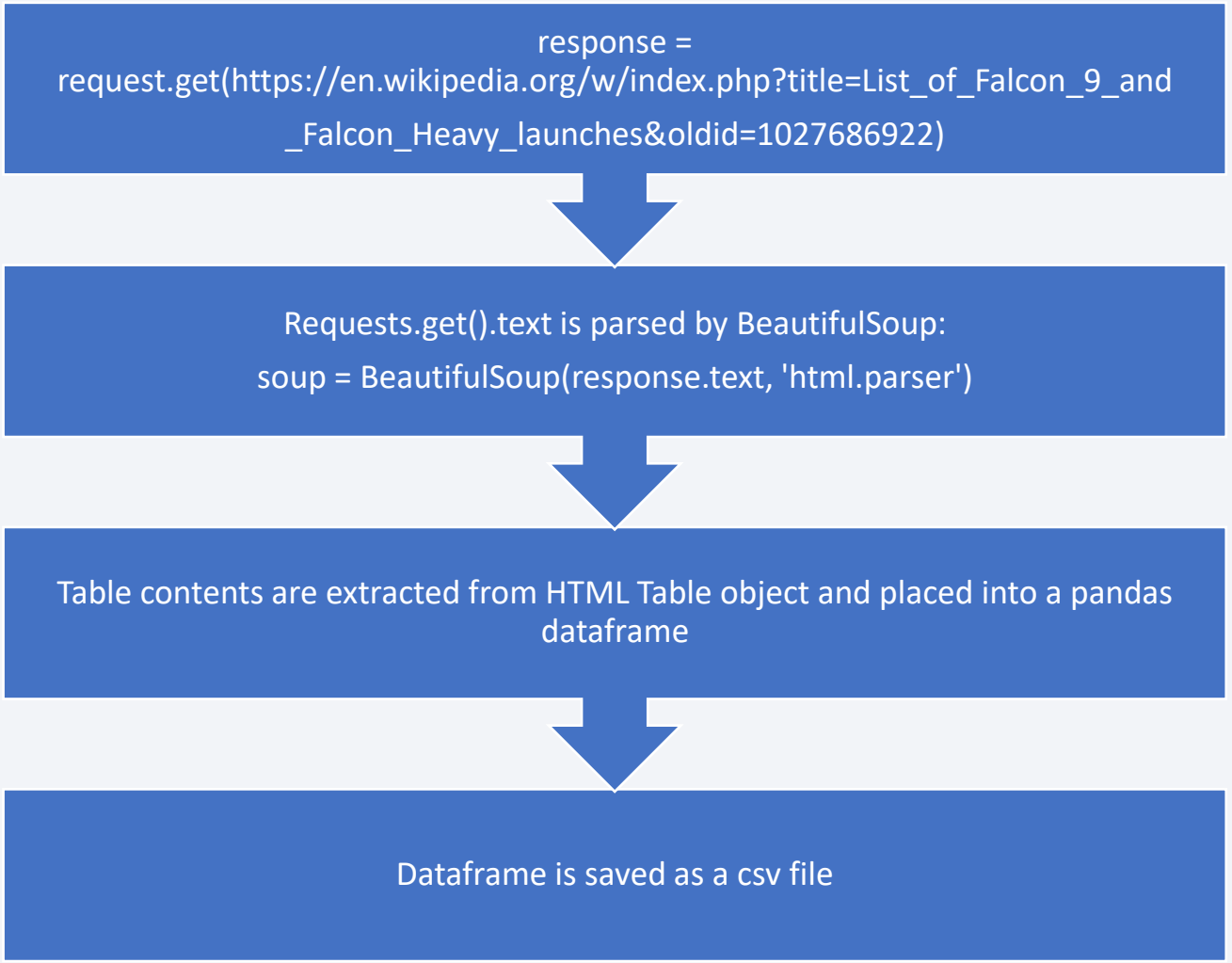
- SpaceX Launch data was requested and parsed using the GET request. Next, we decode the data as Json and normalized it before converting it to a dataframe. We get more information from the API using launch IDs – rockets, launchpads, payloads, and cores. We filter the data to include only Falcon 9 launches. The dataframe is then saved as csv

- Jupyter Notebook here : https://github.com/emmaemmstein/Applied_Data_Science_Capstone/blob/main/Week_1_Data_Collection_API.ipynb



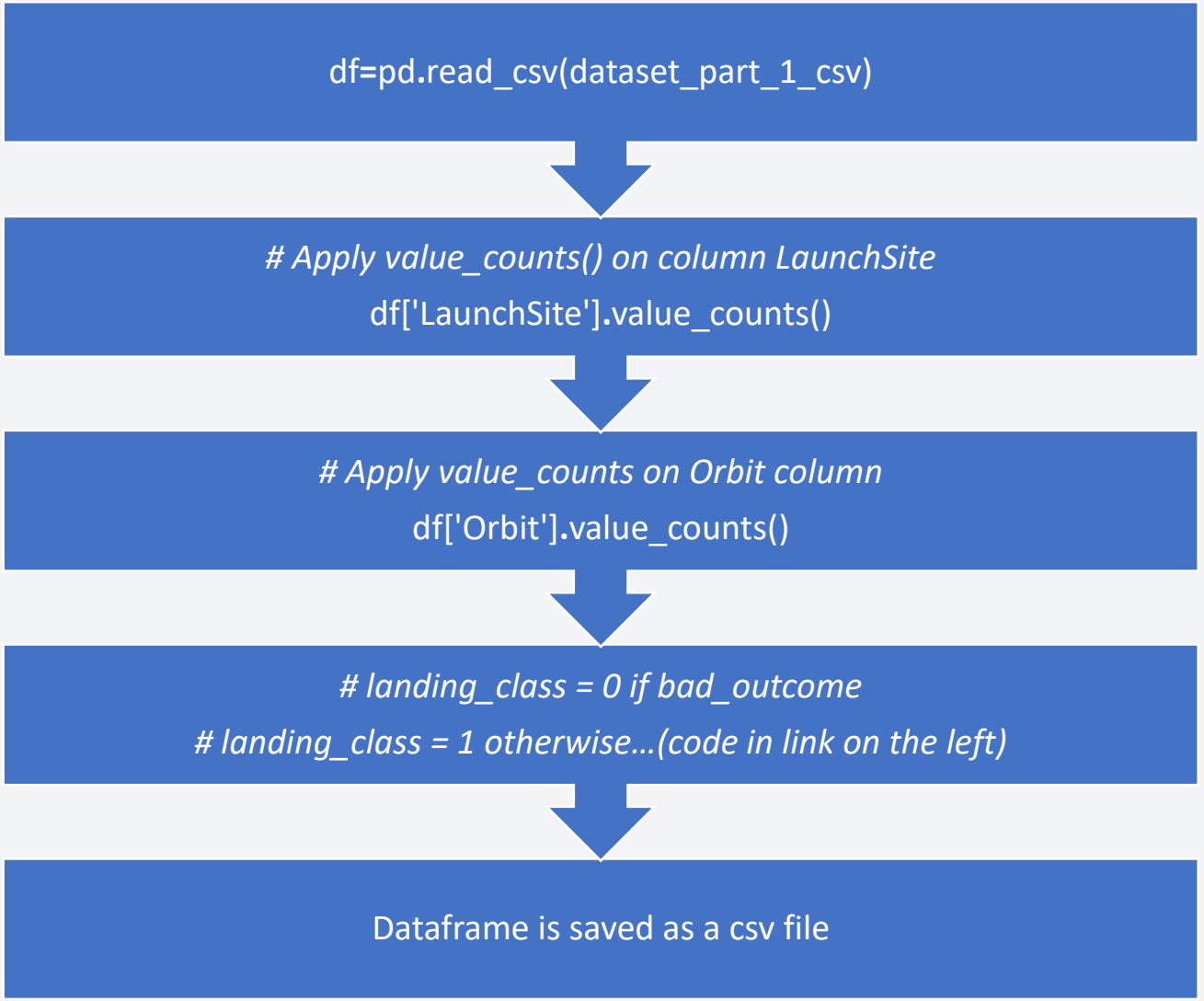
Data Collection - Scraping

- The Wikipedia page was accessed using the ‘requests’ library, the content of the HTML was then parsed using ‘BeautifulSoup’ library. This is then used to extract table object, this table is iterated and relevant contents is placed in a pandas dataframe. The dataframe is then saved as a csv file
- Jupyter Notebook : https://github.com/emmaemmstein/Applied_Data_Science_Capstone/blob/main/Week_1_Data_Collection_Webscrapping.ipynb

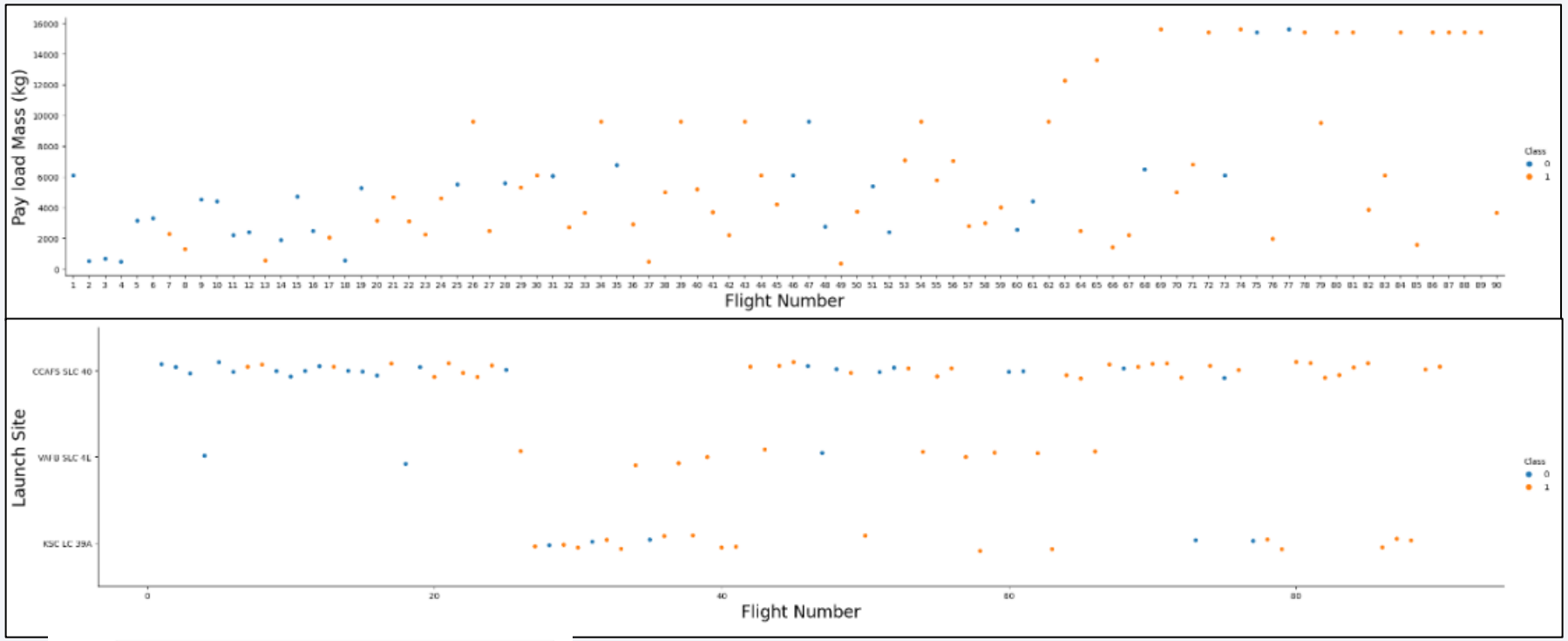


Data Wrangling

- Data from previous steps was loaded into a pandas dataframe, Identified missing attributes, some preliminary calculations were performed, and A landing outcome label from landing outcome column was created to ease analysis. The updated table is then saved as a csv file
- Jupyter Notebook : https://github.com/emmaemmstein/Applied_Data_Science_Capstone/blob/main/Week_1_SpaceX_Data_Wrangling.ipynb

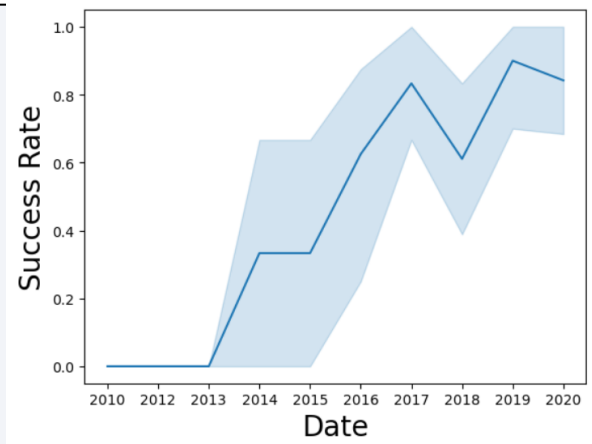


EDA with Data Visualization



it seems the more massive the payload, the less likely the first stage will return

We can see that 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%



We can see how success rate of 1st stage rocket recovery has evolved over time

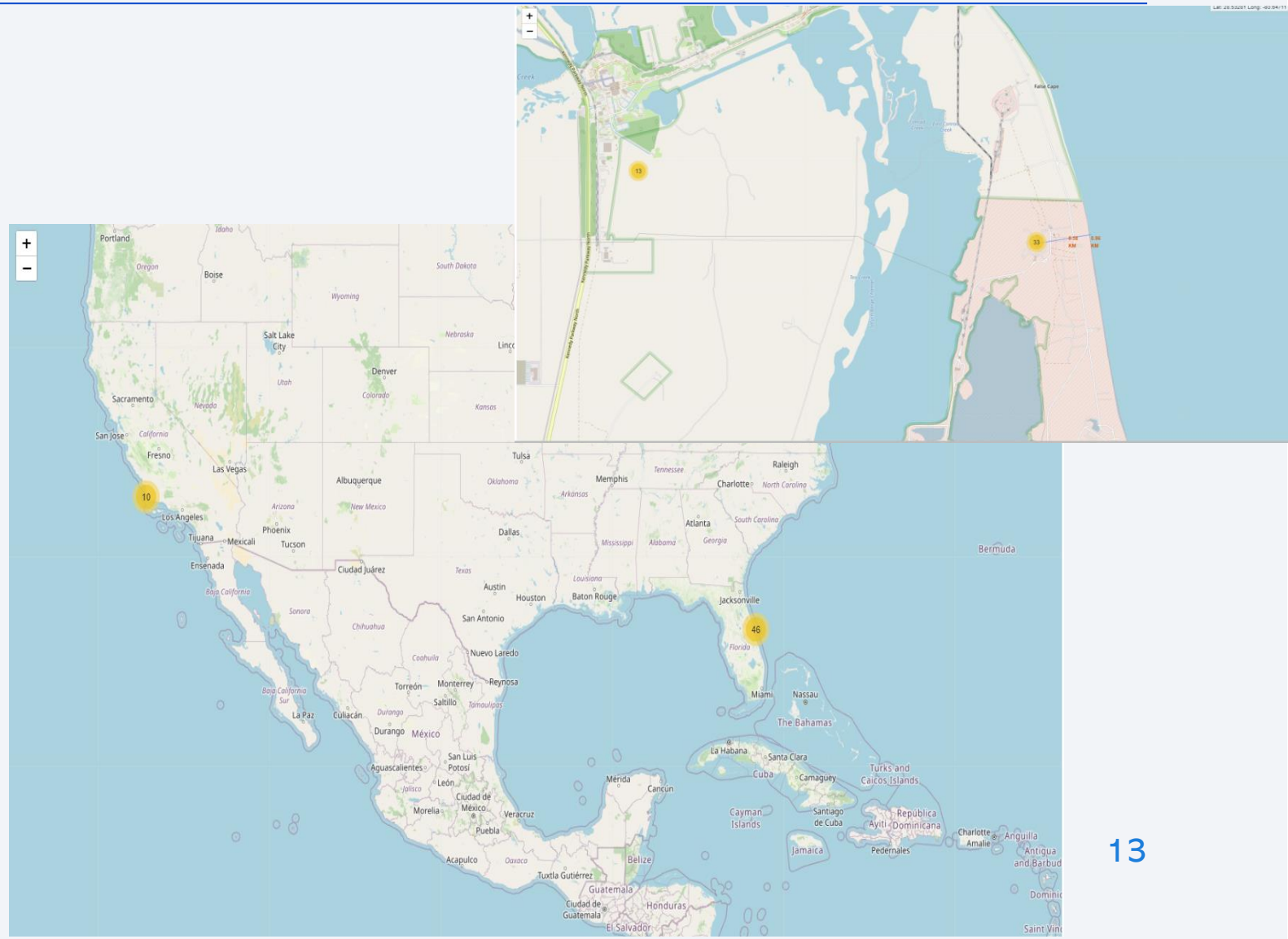
Jupyter Notebook : https://github.com/emmaemmstein/Applied_Data_Science_Capstone/blob/main/Week_2_EDA_with_Visualization_Lab.ipynb

EDA with SQL

- Established a connection to the database using SQLAlchemy
- Filter out blank rows from record
- Find the names of all unique launch site
- Find total Payload mass by NASA (CRS)
- Find average Payload mass carried by F9v1.1
- (……)
- 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.
- Jupyter Notebook: https://github.com/emmaemmstein/Applied_Data_Science_Capstone/blob/main/Week_2_EDA_SQL.ipynb

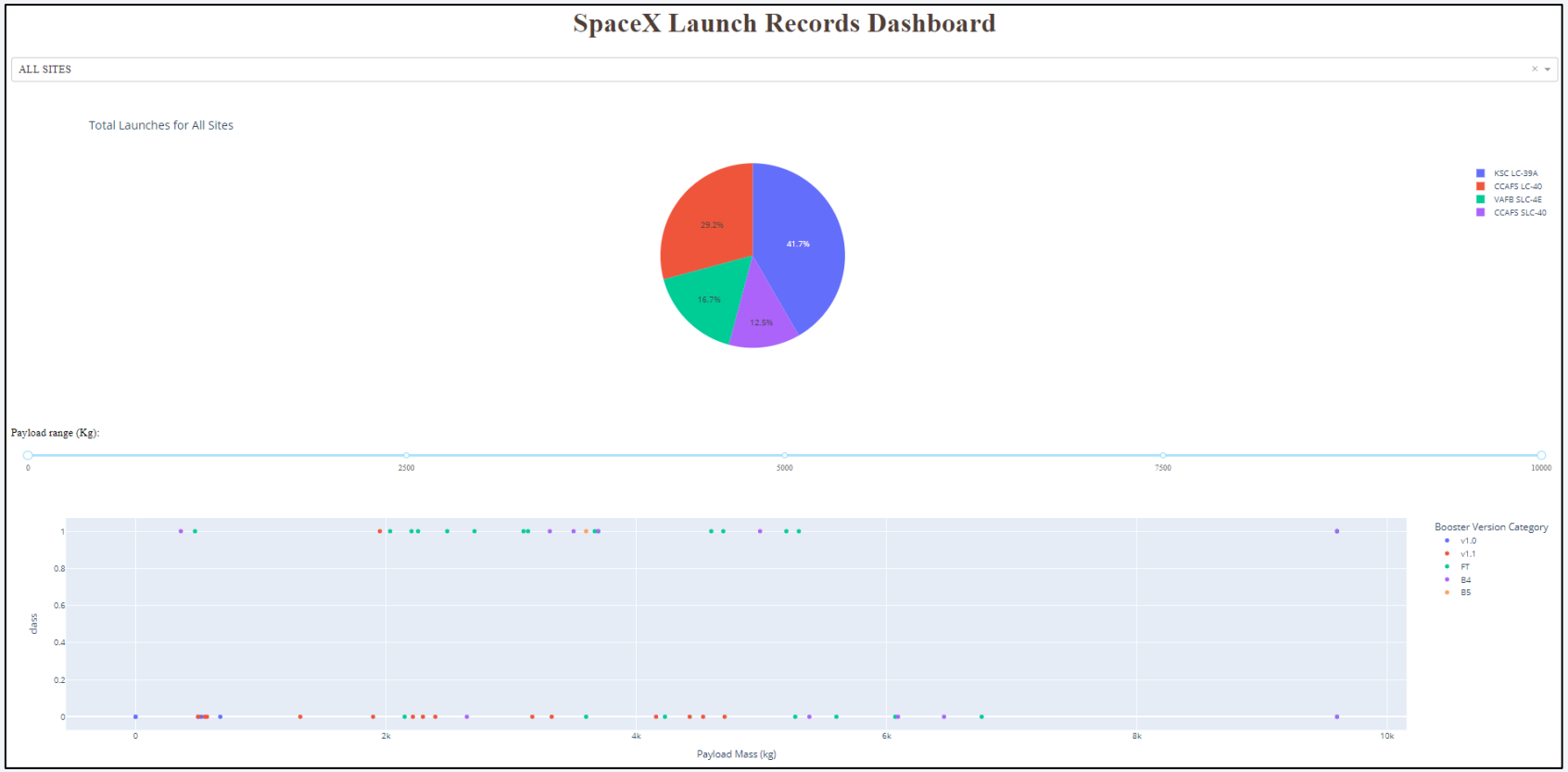
Build an Interactive Map with Folium

- Launch site locations were added on to a folium map. Booster markers for successful/failed launches for each site were created as well
- By zooming in, you can view further details
- Jupyter Notebook: https://github.com/emmaemmstein/Applied_Data_Science_Capstone/blob/main/Week_3_Interactive-Visual_Analytics_Folium.ipynb



Build a Dashboard with Plotly Dash

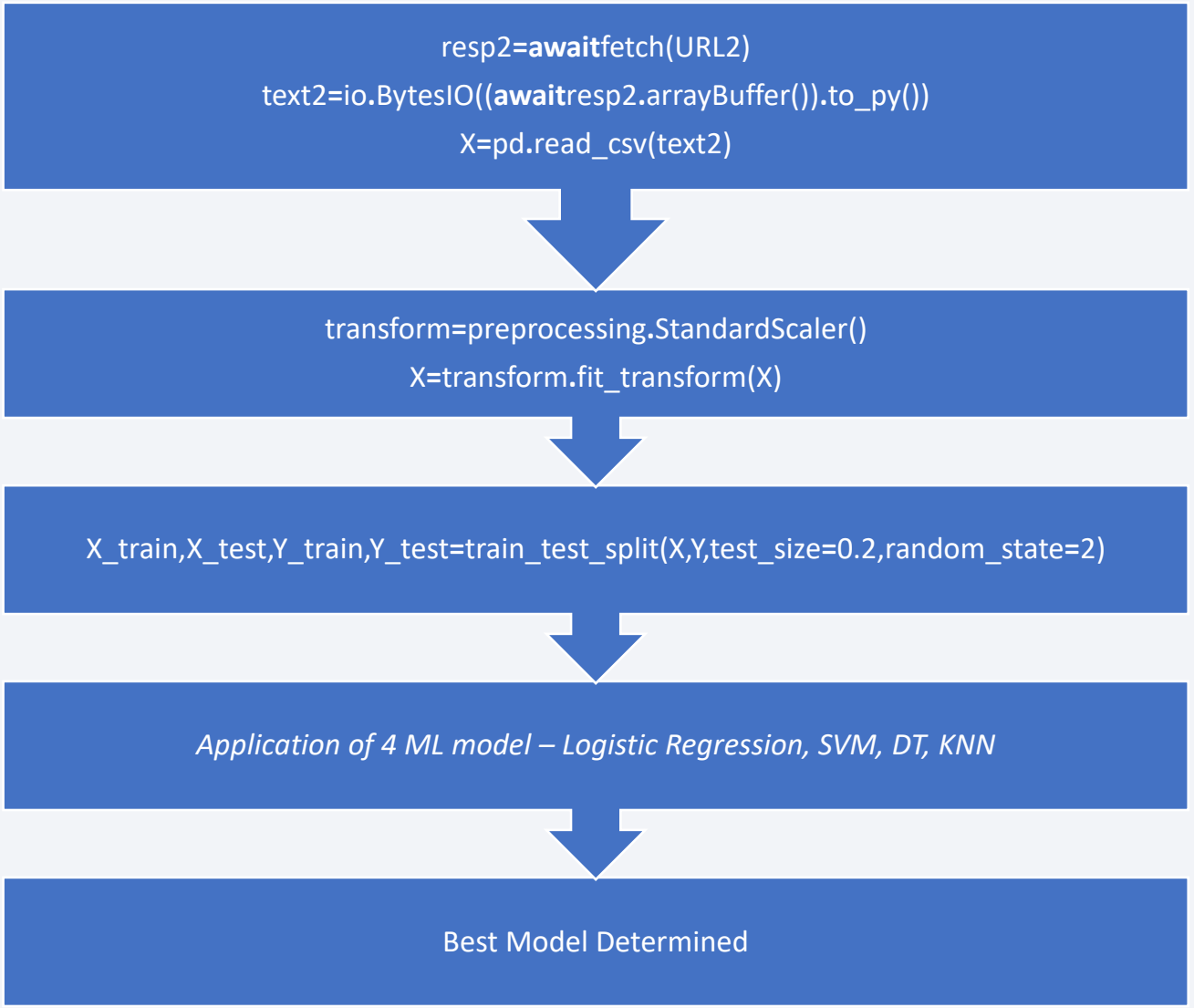
- A pie chart showing the successful launches by launch site is developed. A scatter plot is shown also, displaying the Payload mass with respect to its success/failure. These plots would aid decision by the users.



- Python code: https://github.com/emmaemmstein/Applied_Data_Science_Capstone/blob/main/dash_interactivity_1.py

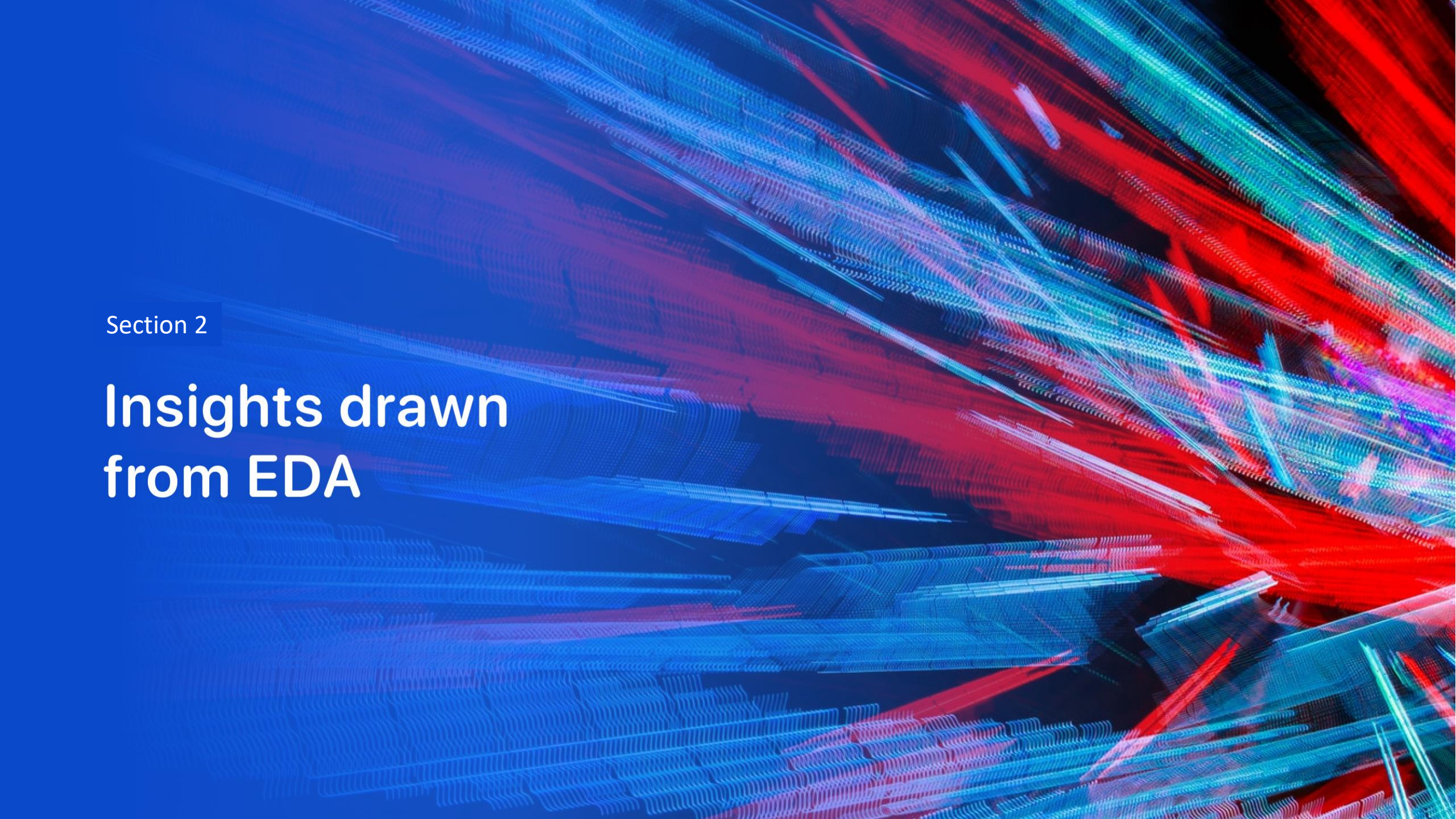
Predictive Analysis (Classification)

- Data was loaded into a Pandas dataframe from a csv file
- The data was then standardized
- Data is then split into train and test data
- Application of ML models on the data
- Determination of model with best accuracy
- Jupyter Notbook: https://github.com/emmaemmstein/Applied_Data_Science_Capstone/blob/main/Week_4_SpaceX_Machine_Learning_Prediction.ipynb

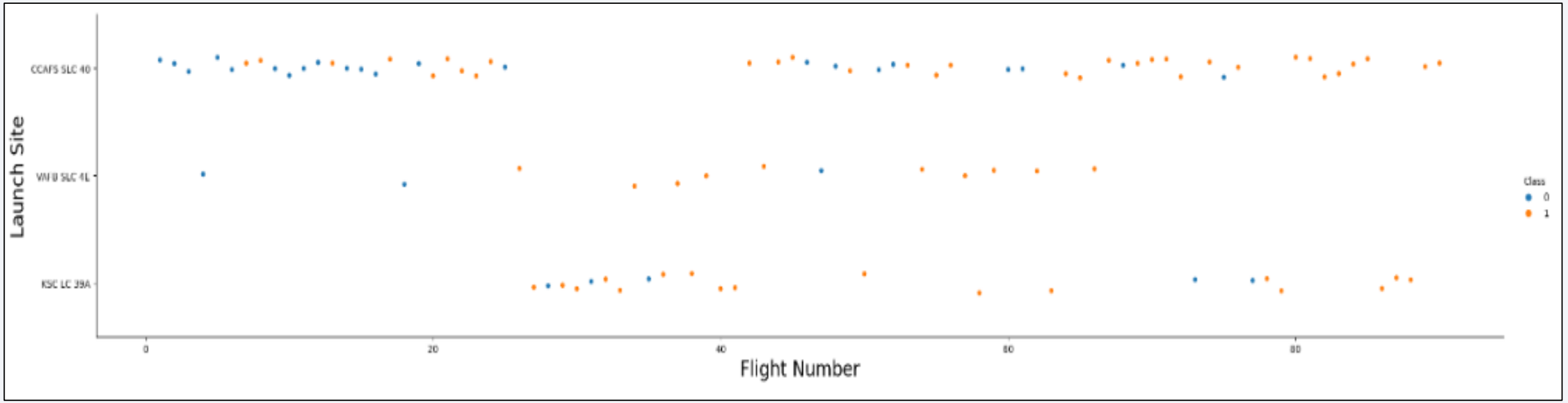


Results

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

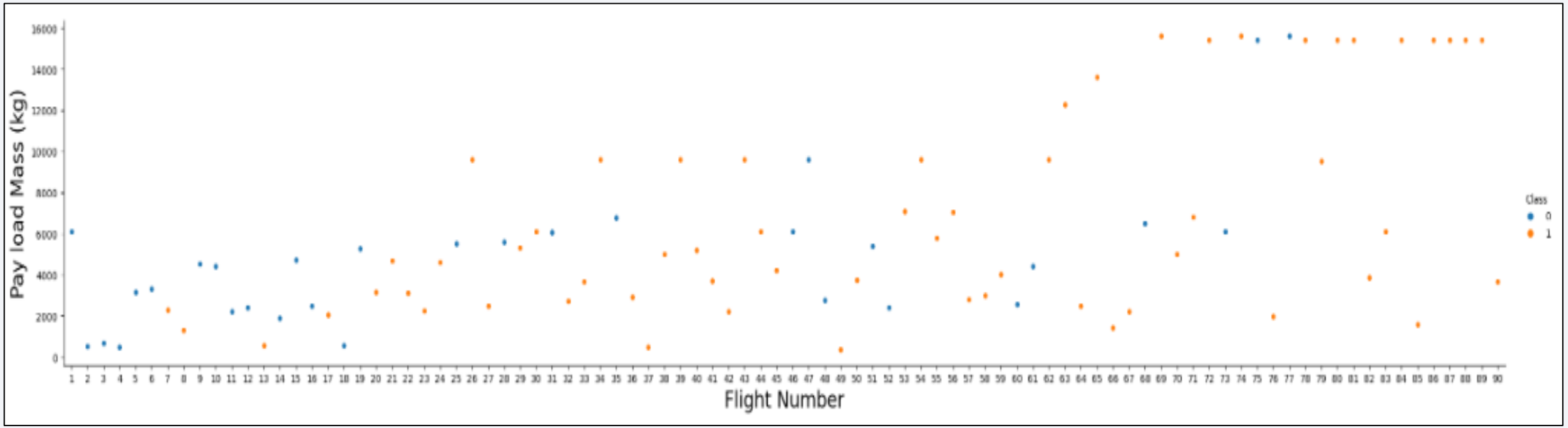

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



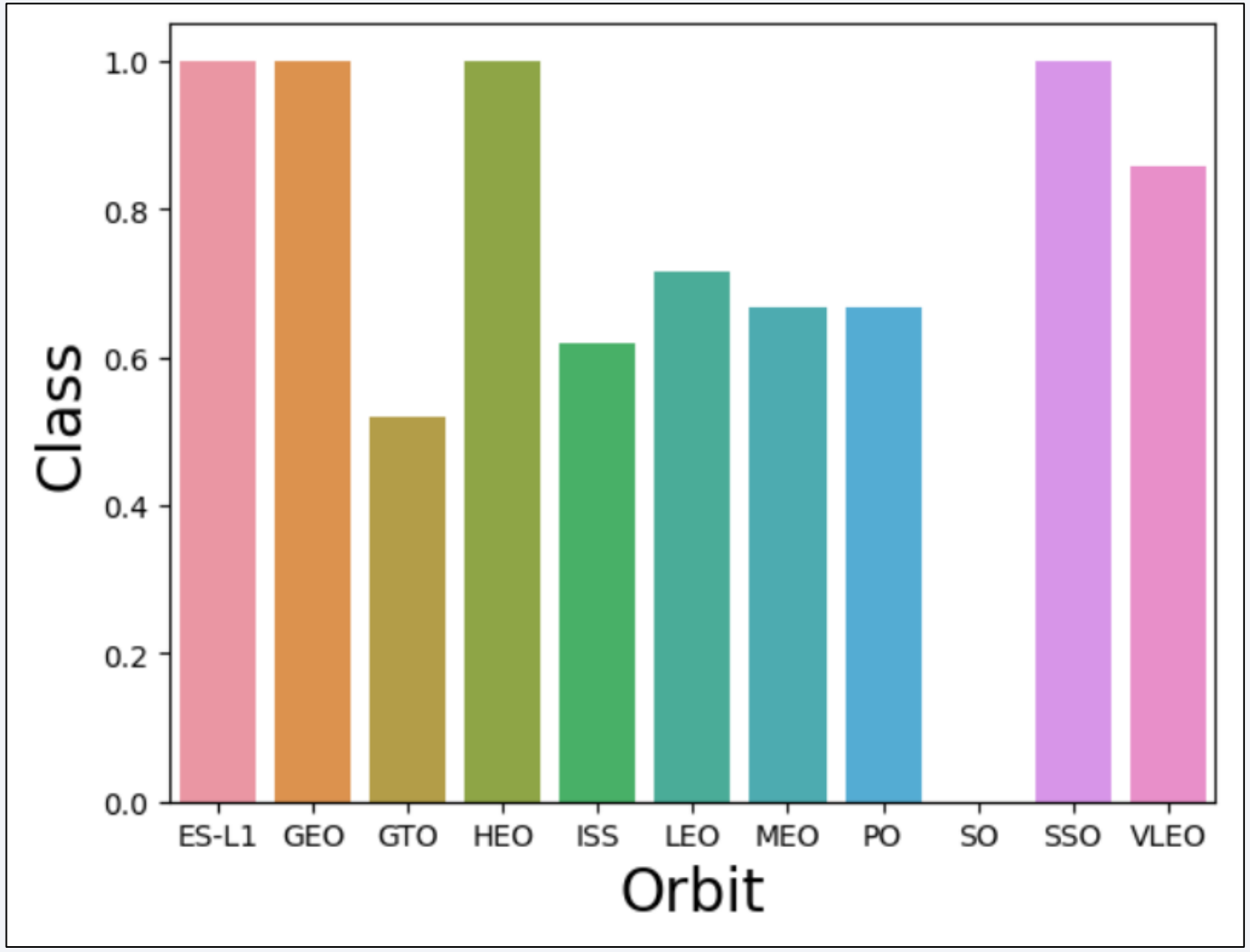
- We can observe that CCAFS LC-40 was mostly used, the reason we may have to find out from an expert. Also, we can observe that more success were recorded on later flights than at the earlier ones.

Payload vs. Launch Site



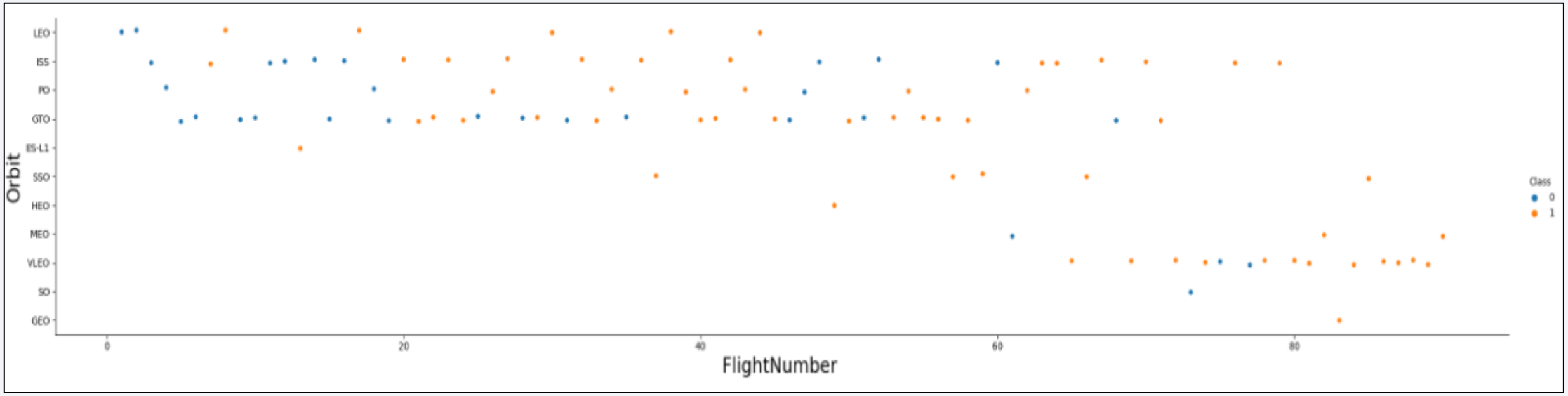
- This plot demonstrates the improvement on launches, we see that in later flights the payload mass is increasing, and the launches are being more successful as well.

Success Rate vs. Orbit Type



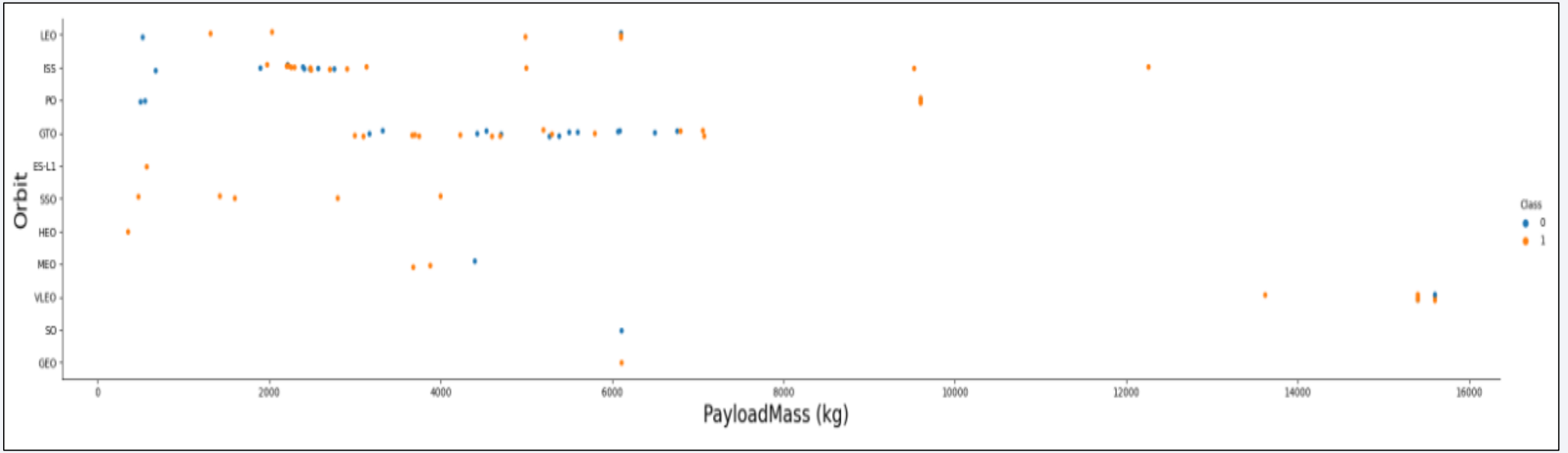
- The plot shows ES-L1, GEO, HEO and SSO orbit have a 100% successful landing. The least being SO orbit followed by GTO

Flight Number vs. Orbit Type



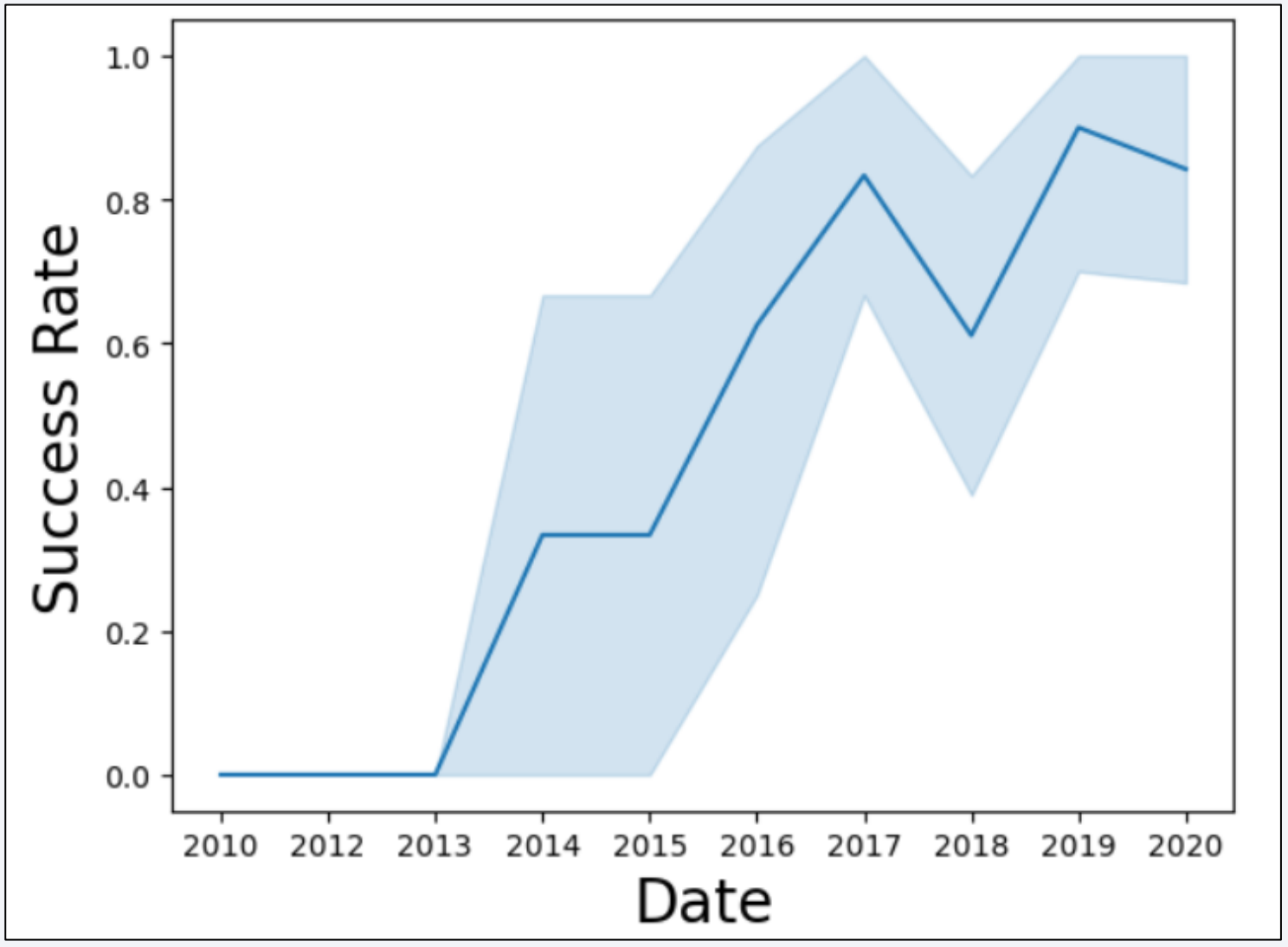
- GEO, PO, ISS, LEO were the orbit targeted for the first 15 flights. They were also the orbits most targeted during the period considered in this project. We could see SO orbit was targeted only once so the failure rate in the previous slide doesn't tell the complete story. VLEO orbit was targeted mostly for later flight with success

Payload vs. Orbit Type



- We can see that the VLEO orbit was the target for the bigger payloads

Launch Success Yearly Trend



- We see that no record of success for the first 3 years, then successful launches were demonstrated with a linear progression over the years.

All Launch Site Names

Task 1

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

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

Task 2

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

```
%sql select * from SPACEXTBL where Launch_site like 'CCA%' limit 5
```

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

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

Total Payload Mass

Task 3

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

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

sum(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version = 'F9 v1.1'
```

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

avg(payload_mass__kg_)
2928.4

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

```
%sql select min(date) from SPACEXTBL where landing_outcome = 'Success (ground pad)'
```

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

min(date)
2015-12-22

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

```
%sql select booster_version from SPACEXTBL where (landing_outcome = 'Success (drone ship)' and 4000<payload_mass__kg_ and
```

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

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql select count(mission_outcome) from SPACEXTBL
```

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

count(mission_outcome)
101

Boosters Carried Maximum Payload

Task 8

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

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_= (select max(payload_mass__kg_) from SPACEXTBL)
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

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.

```
%sql select substr(date,6,2),landing_outcome, booster_version, launch_site from SPACEXTBL where substr(Date,1,4) = '2015'
```

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

substr(date,6,2)	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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.

```
%sql select landing_outcome, count(landing_outcome) from SPACEXTBL where date between '2010-06-04' and '2017-03-20' group
```

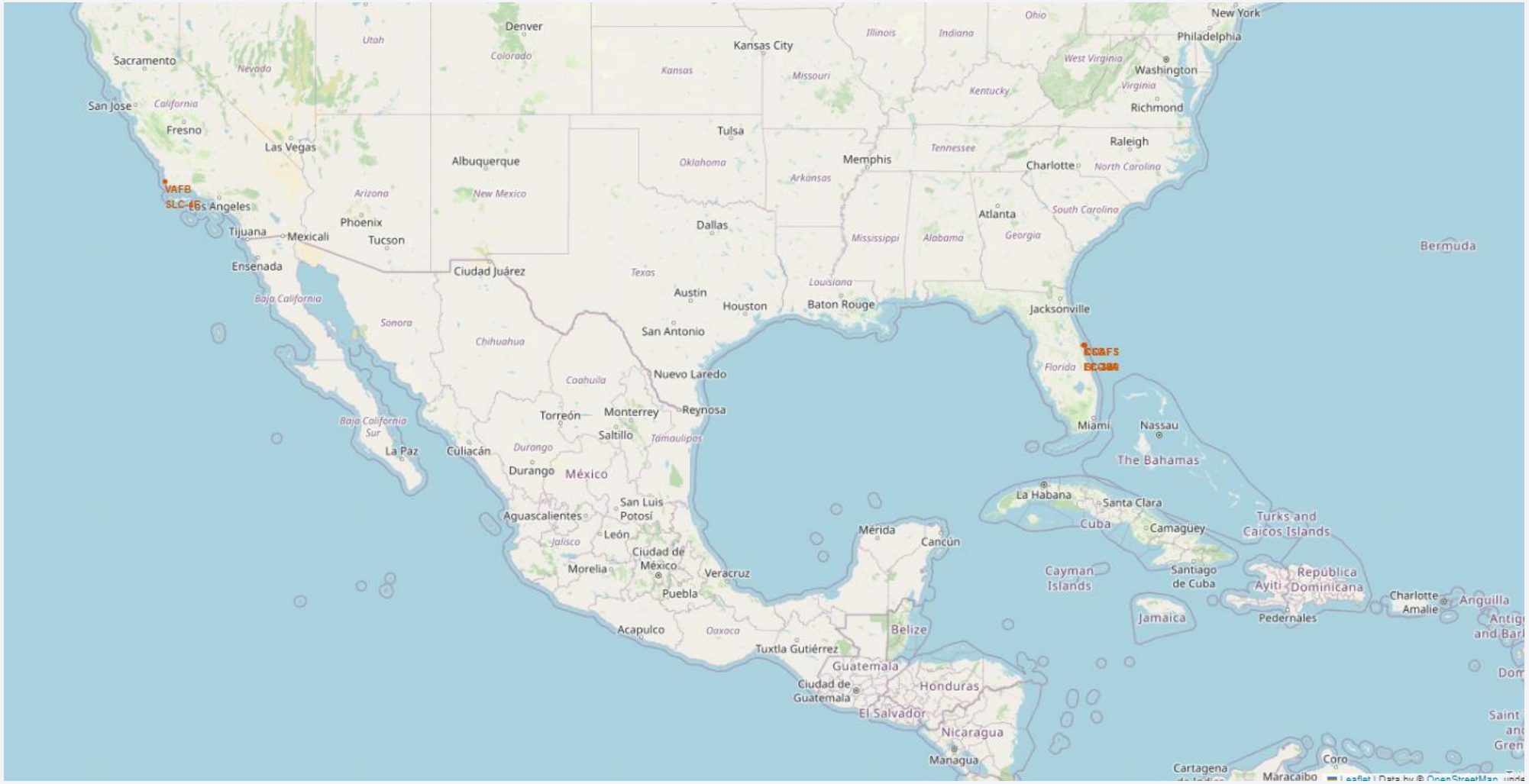
```
* sqlite:///my_data1.db
Done.
```

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

Section 3

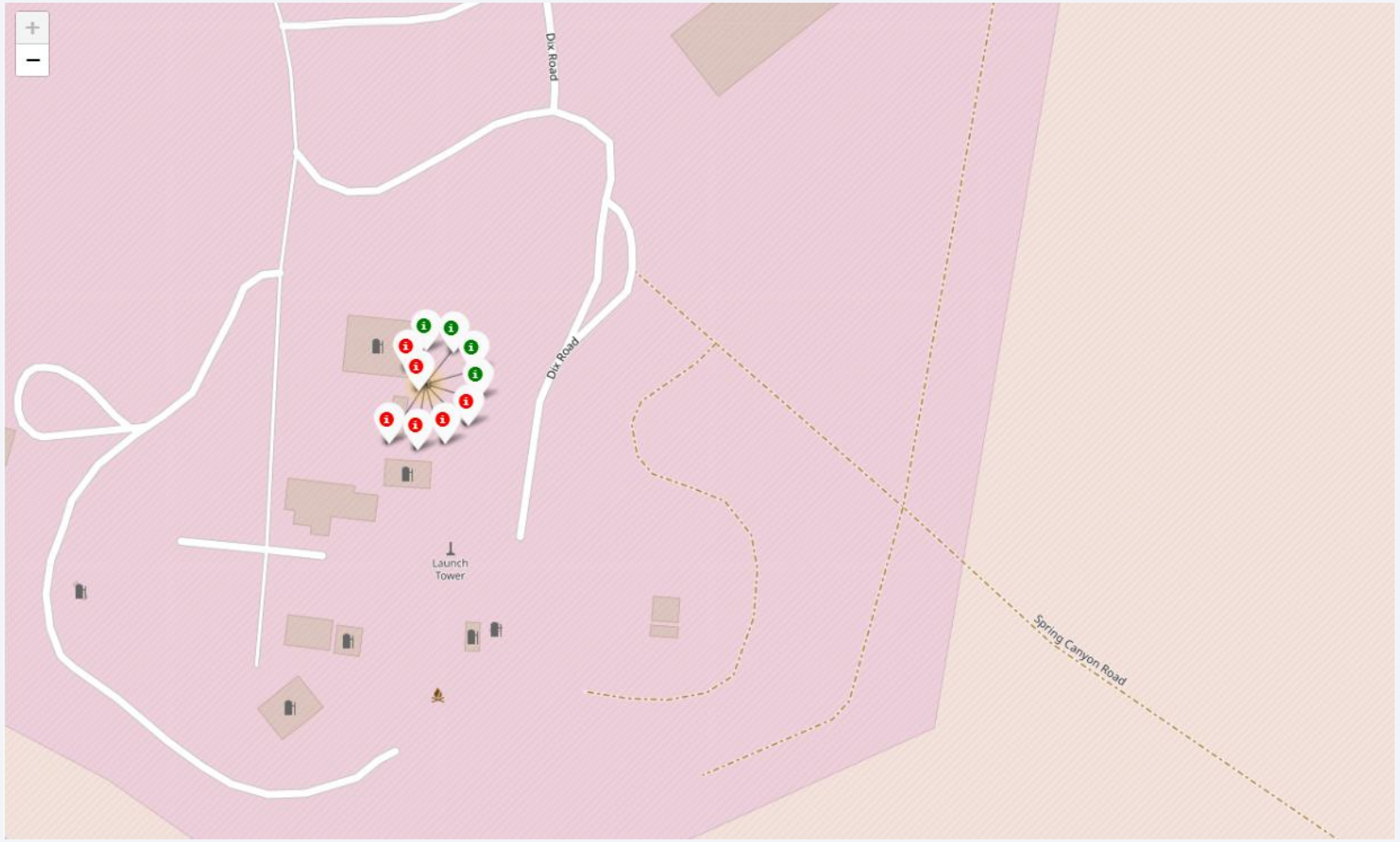
Launch Sites Proximities Analysis

Launch Sites' Locations



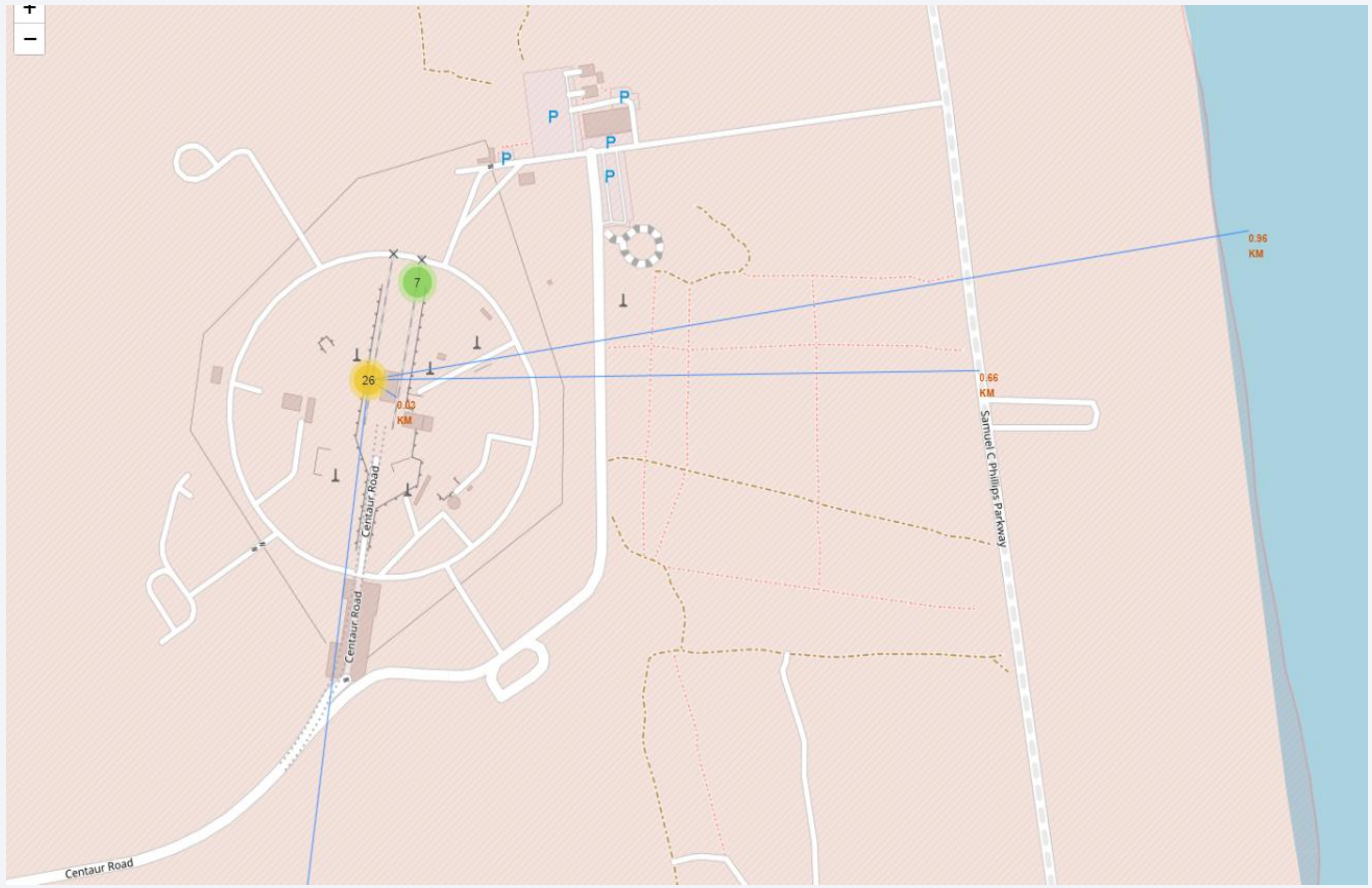
- We can see that all launch sites are located near the coastline

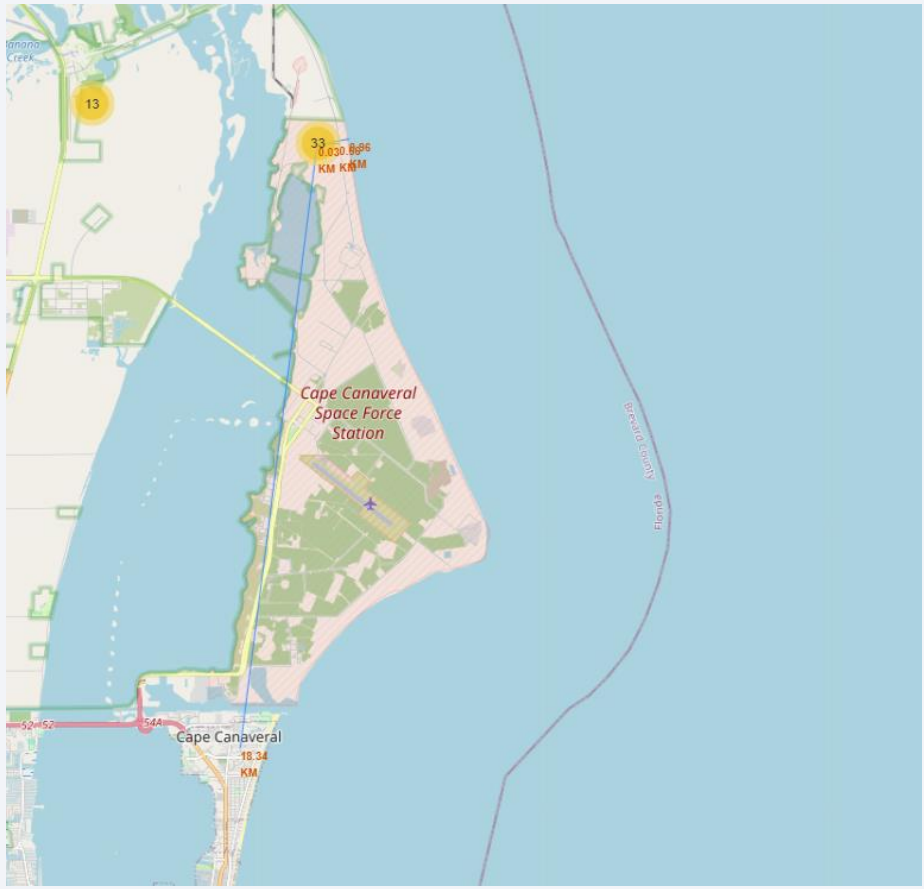
Launch outcomes by Labels



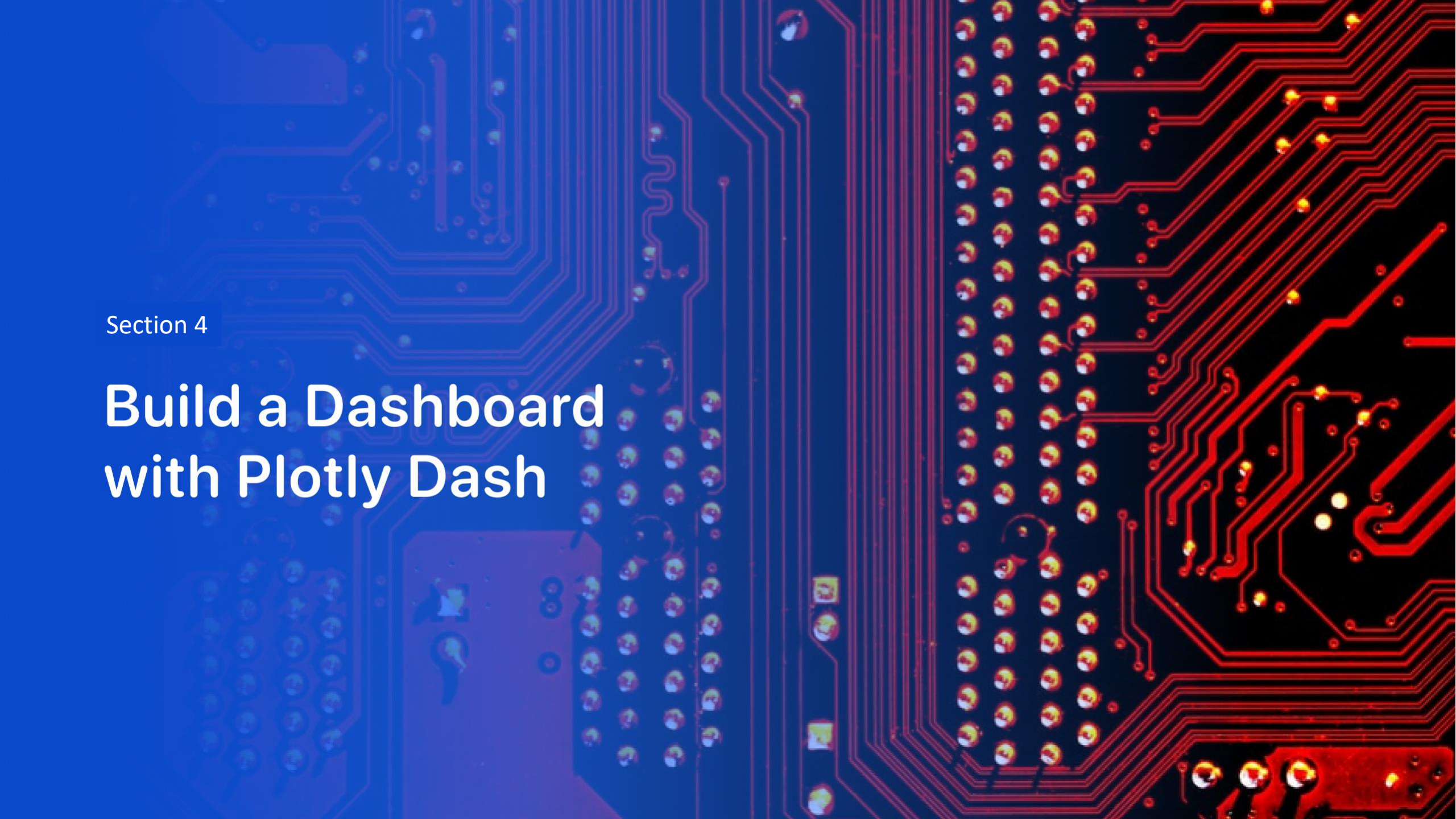
- Launch outcome of either success or failure is shows by two different colors: Green for success and red for failure

Launch Site proximity to Land marks

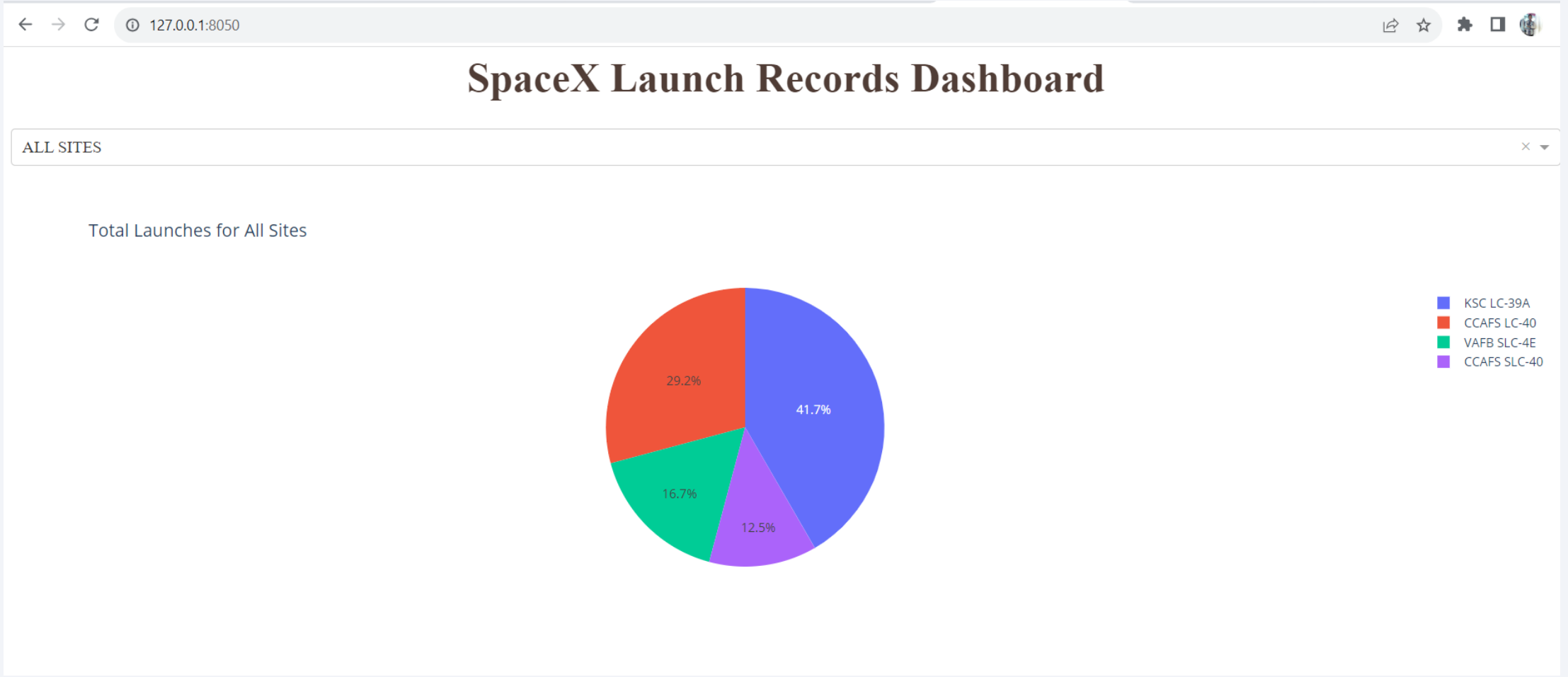




- The closest city to this particular site is about 18 km (from the Left zoom out picture), distance from coast (0.96 km), railway (0.03 km), highway (0.66 km)

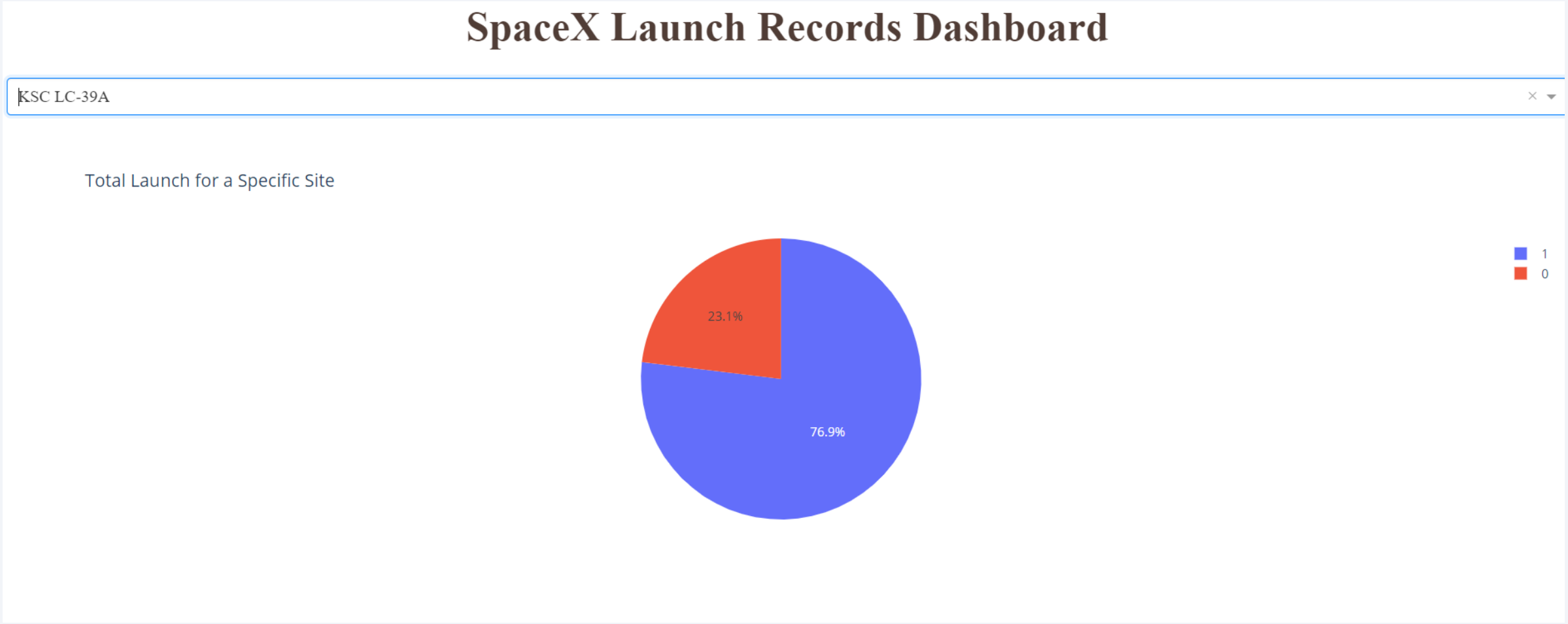

Section 4
Build a Dashboard with Plotly Dash

Dashboard for Successful Launches



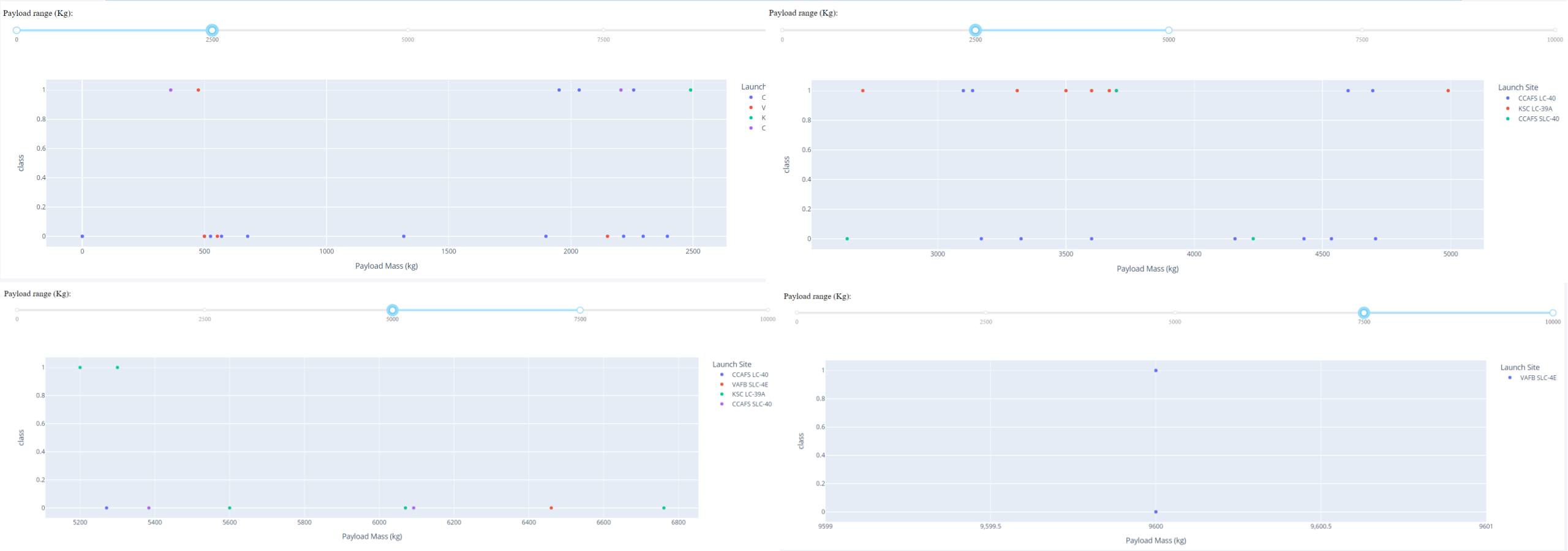
- Total success launches by all sites. KSC LC-39A shows to have the most success, and the least being CCAFS SLC-40

Launch site with Highest Launch Success (KSC_LC 39A)



- Success percentage of 76.9

PayLoad vs Launch Outcomes

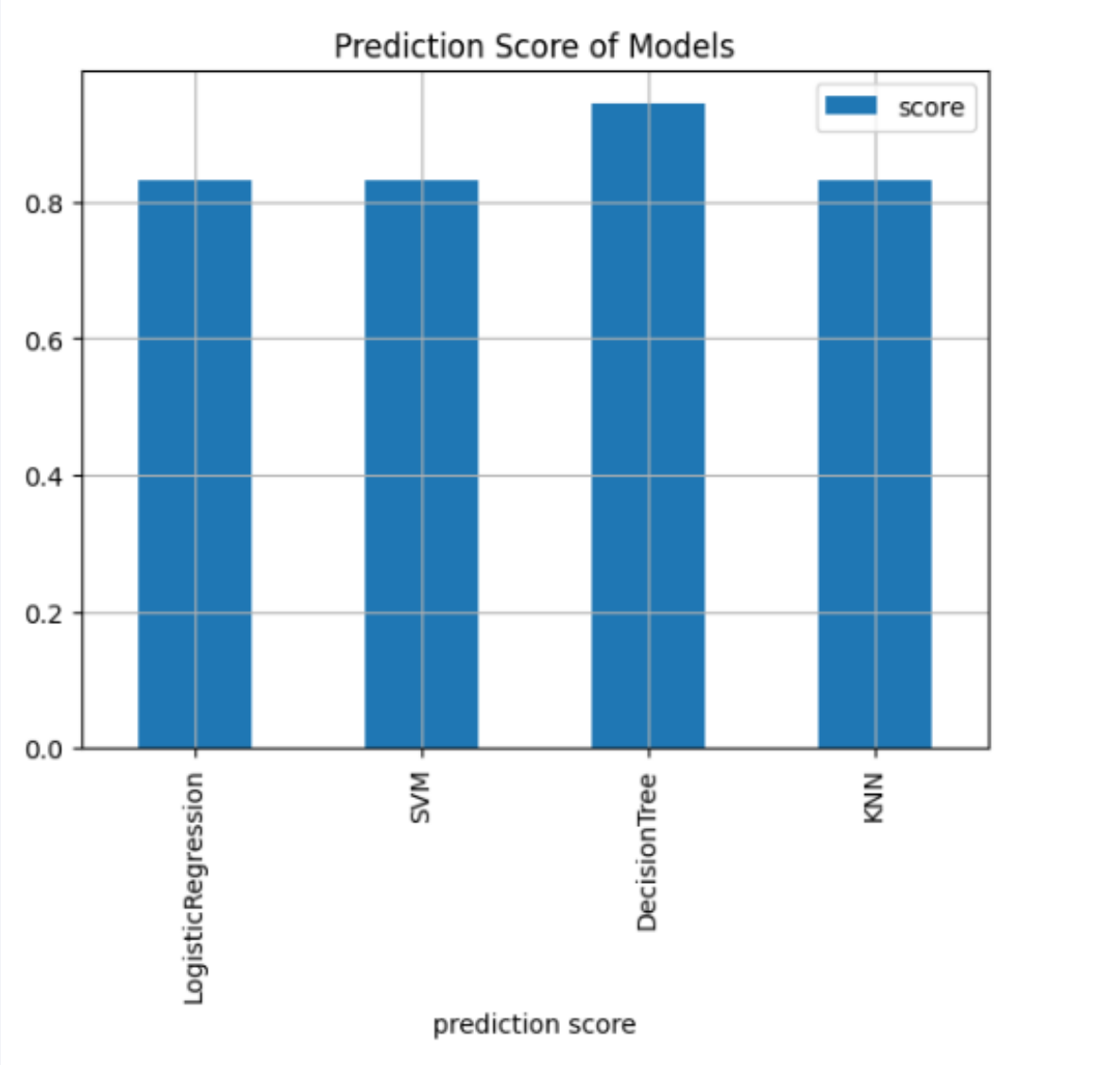


The Payload mass with respect to the launch outcome is shown for different ranges of payload mass (0-2500, 2500-5000, 5000-7500, 7500-10000 kg)

Section 5

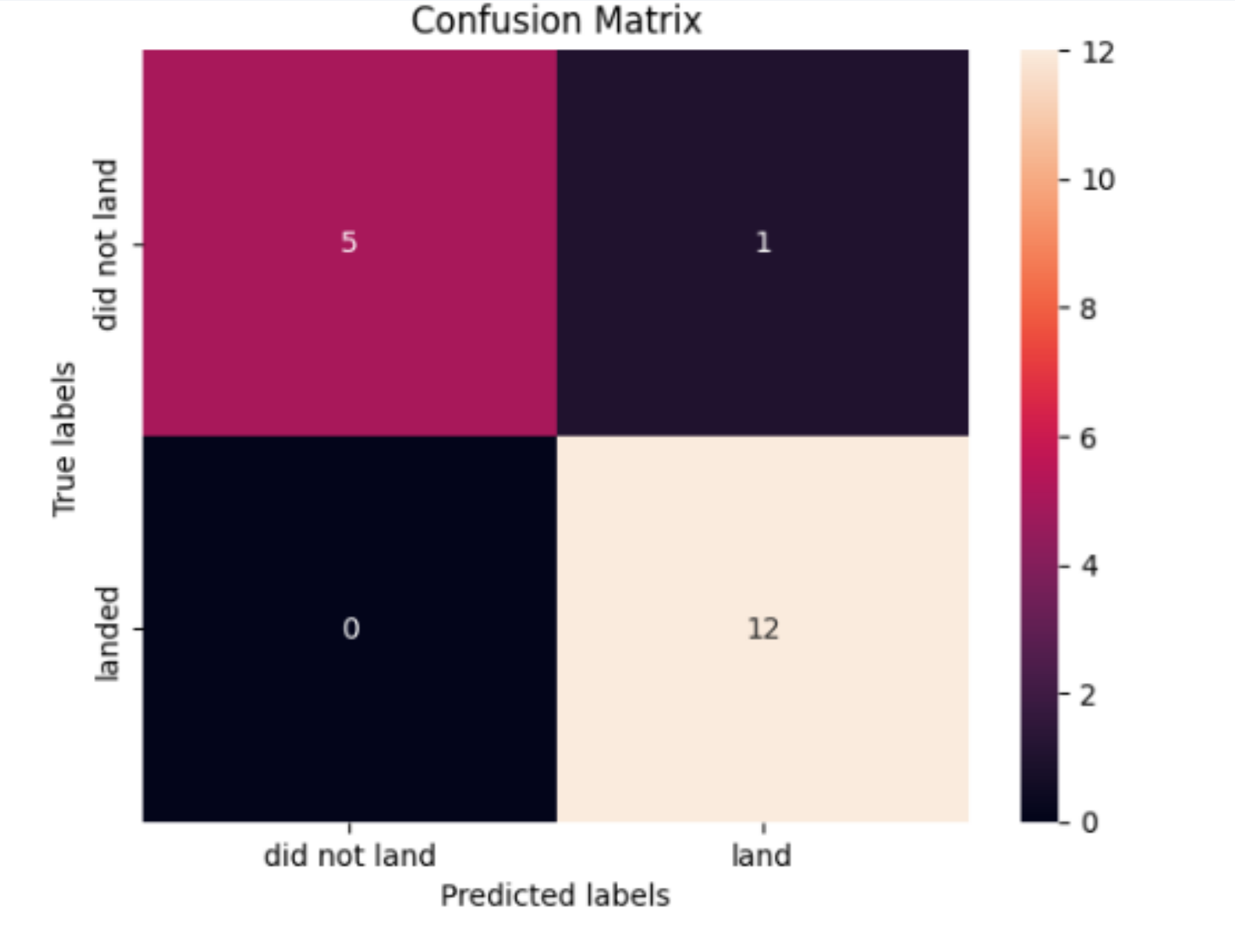
Predictive Analysis (Classification)

Classification Accuracy



The Decision Tree has the highest prediction score

Confusion Matrix



The Decision Tree Confusion Matrix

Conclusions

- We have been able to successfully demonstrate Data Science methodology in this project, from web scraping to ML prediction
- We observe that the success rate of launch for the Falcon 9 increased between 2013-2017 but kind of plateau from 2017-2020
- Decision Tree model shows to be a better model of all model compared.

Insights

- We can consider getting more data by considering data from 2021 and 2022
- We could also consider other advanced ML models for better prediction accuracy
- Checks for if factors like weather condition play a role in launch success

Appendix

- Code for prediction score

TASK 12

Find the method performs best:

```
[37]: #The method with the best predictor.test(X_test, Y_test)
model_score = {'model':['LogisticRegression','SVM','DecisionTree','KNN'],
               'score':[logreg_cv.score(X_test, Y_test),svm_cv.score(X_test, Y_test),
                        tree_cv.score(X_test, Y_test),knn_cv.score(X_test, Y_test)]}

dff = pd.DataFrame.from_dict(model_score)
dff.plot(x='model', y='score',kind = 'bar', xlabel='prediction score',\
        title = 'Prediction Score of Models',grid=True,)
```

https://github.com/emmaemmstein/Applied_Data_Science_Capstone

Thank you!

Emmanuel Chukwuemeka
08/20/2023