IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Francisco Gaztañaga
2024/02/29

Outline

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

Executive Summary

- Methodologies used to analyze the data:
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of results
  - The data was collected from public sources without any inconvenient.
  - EDA was a great tool for identifying features and predicting the success of lunches. It was also helpful to predict a success launch
  - Through Machine Learning Prediction it was possible to find the best model to predict successes and failures in rocket launches.

Introduction

- The savings in SpaceX launches, compared to its competitors, are found in the reuse of the first stage of the rocket. Therefore, if we can determine whether the first stage will land, we can determine the cost of a launch. This information is required by the company SpaceY in order to compete in the aerospace market. The goal of the project is to create a machine learning pipeline to predict whether the first stage will be successful.

- Questions to be answered:

  Where is the launch center where most launches were successful?

  Which characteristics are the most relevant for a correct prediction of a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 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
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Data collected were normalized, dividing in training and test data sets and evaluating them by four different classification models, being the accuracy of each model analyzed using different techniques.

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API (https://api.spacexdata.com/v4/rockets/) and Web Scraping data from a table in SpaceX's Wikipedia entry (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches). We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

Data Collection – SpaceX API

(Source code)

Steps followed:

1. **Request data** from SpaceX API (rocket launch data)
2. **Decode response** using .json() and convert to a dataframe using .json_normalize()
3. **Request information** about the launches from SpaceX API using custom functions
4. **Create dictionary** from the data
5. **Create dataframe** from the dictionary
6. **Filter dataframe** to contain only Falcon 9 launches
7. **Replace missing values** of Payload Mass with calculated .mean()
8. **Export data** to csv file

Data Collection – Web Scraping

(Source code)

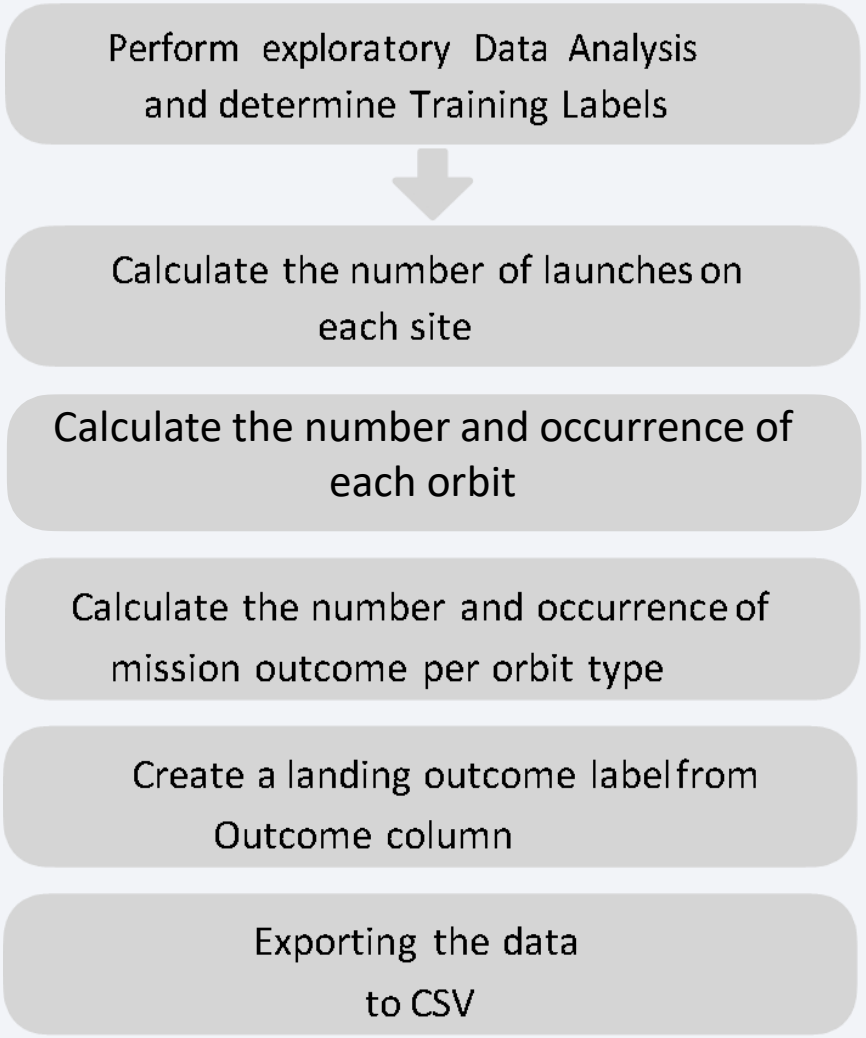
Steps followed:

1. **Request data** (Falcon 9 launch data) from Wikipedia
2. **Create BeautifulSoup object** from HTML response
3. **Extract column names** from HTML table header
4. **Collect data** from parsing HTML tables
5. **Create dictionary** from the data
6. **Create dataframe** from the dictionary
7. **Export data** to csv file

Data Wrangling

(Source code)

The process for data wrangling could be summarized in the flowchart:



EDA with Data Visualization

(Source code)

For the Exploratory Data Analysis, were plotted different charts, summarized as follows:

- Flight Number vs. Payload (scatter plot)
- Flight Number vs. Launch Site (scatter plot)
- Payload Mass (kg) vs. Launch Site (scatter plot)
- Orbit type vs. Class (bar chart)
- Flight number vs. Orbit type (scatter plot)
- Payload Mass (kg) vs. Orbit type (scatter plot)
- Launch Year vs. Success Rate (line plot)

The objective of those charts was to analyze different pair of variables in order to discover which features are actually related with the success or failure of a mission. Scatter plots were useful for a first sight analysis about relationship between variables. Bar charts show relationships measured values from different categories.

EDA with SQL

(Source code)

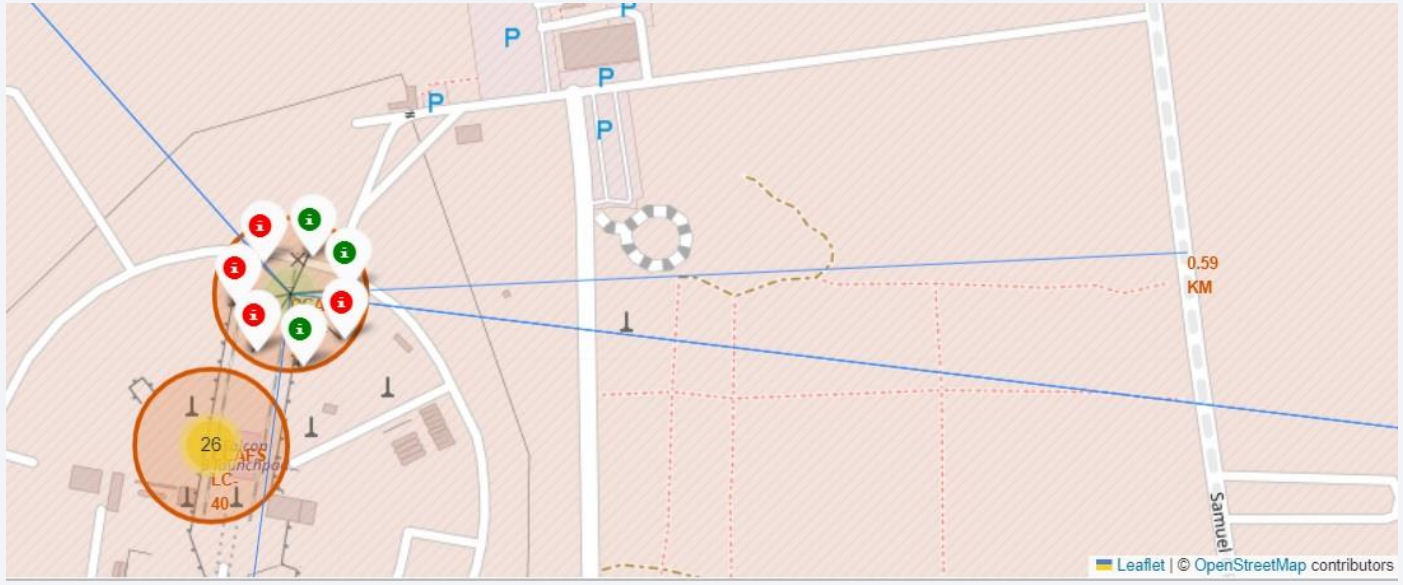
Queries performed in order to display:

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.
- The total number of successful and failed mission outcomes
- The failed landing outcomes in drone ship, their booster version and launch site names.

Build an Interactive Map with Folium

(Source code)

There were added to the interactive map: labeled circles in launch sites, colored markers for successful (green) and unsuccessful (red) lunches, and lines, labeled with the length of it, indicating distances between different places and launch sites. This way, it is easier to analyze where the most of successful launches occurred.



Build a Dashboard with Plotly Dash

(Source code)

There were made:

- **Dropdown List with Launch Sites**

Allow user to select all launch sites or a certain launch site

- **Pie Chart Showing Successful Launches**

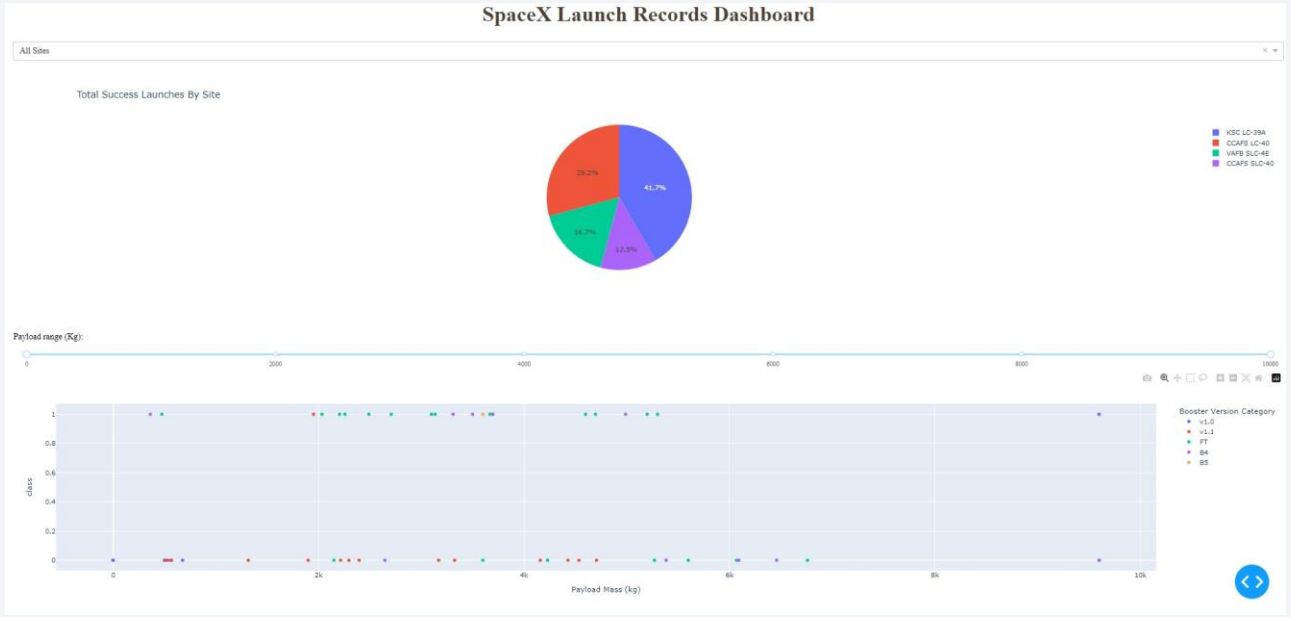
Allow user to see successful and failed launches as a percent of the total

- **Slider of Payload Mass Range**

Allow user to select payload mass range

- **Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version**

Allow user to see the correlation between Payload and Launch Success



Predictive Analysis (Classification)

(Source code)

The tools used were:

1. **Create** NumPy array from the Class column
2. **Standardize** the data with StandardScaler. Fit and transform the data.
3. **Split** the data using train_test_split()
4. **Create** a GridSearchCV object with cv=10 for parameter optimization
5. **Apply** GridSearchCV on different algorithms: Logistic Regression (*LogisticRegression()*), Support Vector Machine (*SVC()*), Decision Tree (*DecisionTreeClassifier()*), K-Nearest Neighbor (*KNeighborsClassifier()*)
6. **Calculate** accuracy on the test data using .score() for all models
7. **Assess** the confusion matrix for all models
8. **Identify** the best model using Jaccard_Score, F1_Score and Accuracy

Results summary

Exploratory Data Analysis

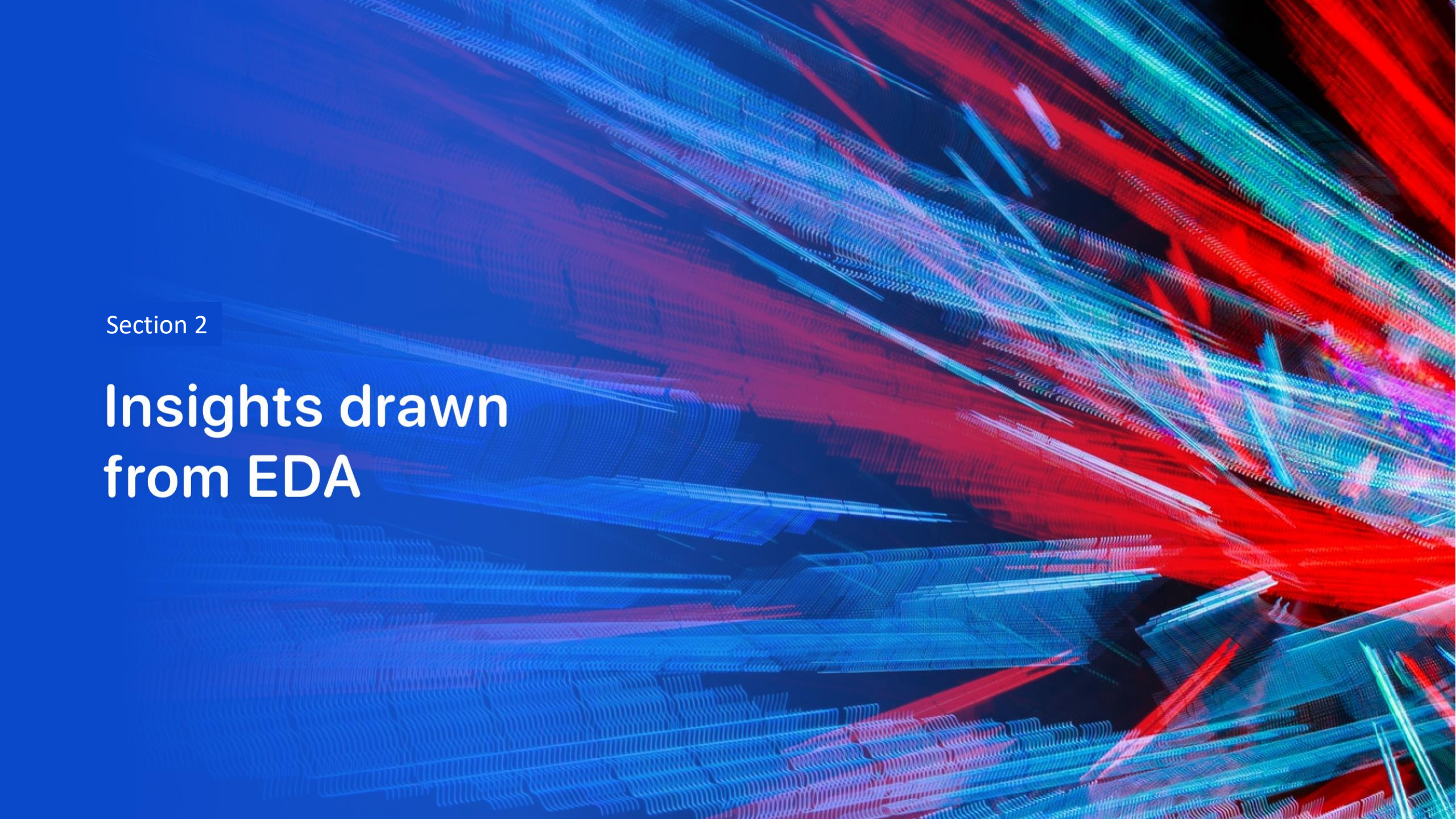
- Launch success has improved over time.
- KSC LC-39A has the highest success rate among landing sites.
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate.

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything can be damaged in a fail launch (city, highway, railway), while are still close enough to bring people and material to support launch activities.

Predictive Analytics

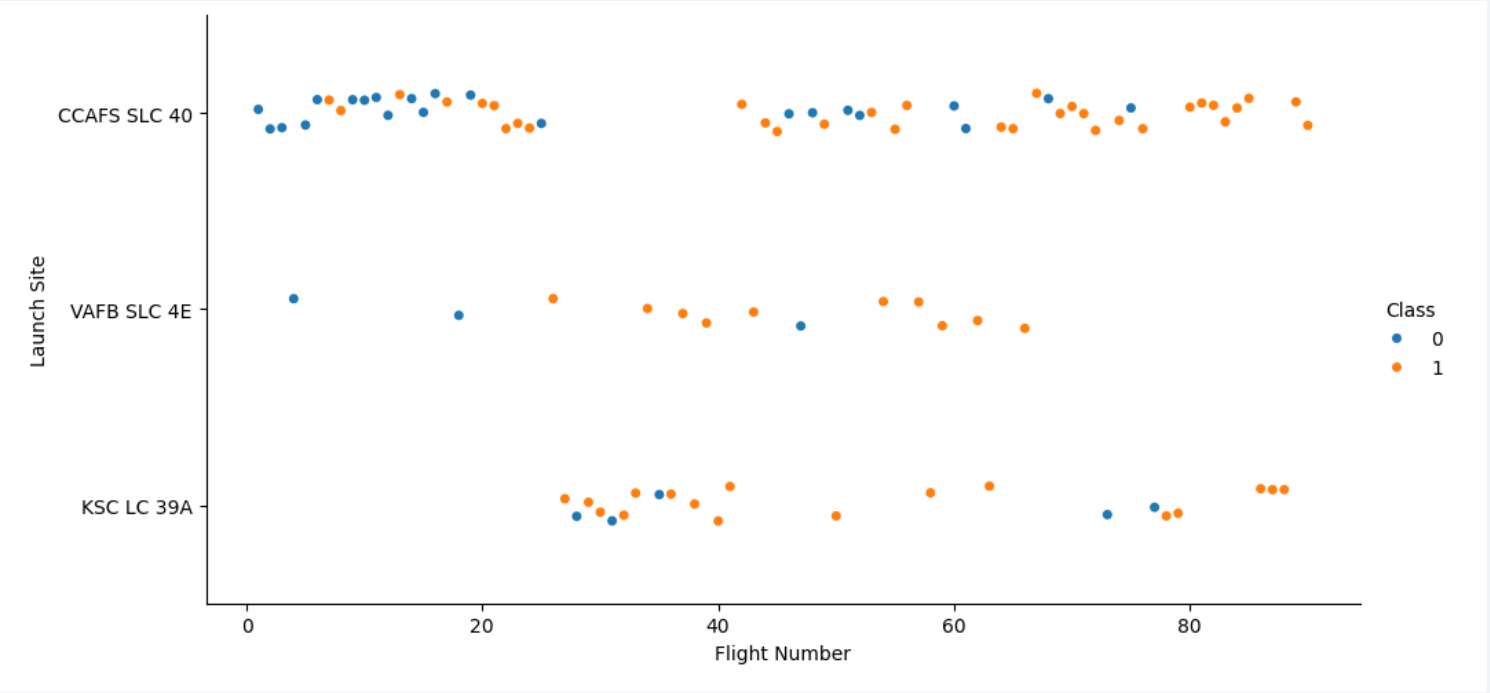
- Decision Tree model is the best predictive model for the dataset with a test accuracy of 0.94.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

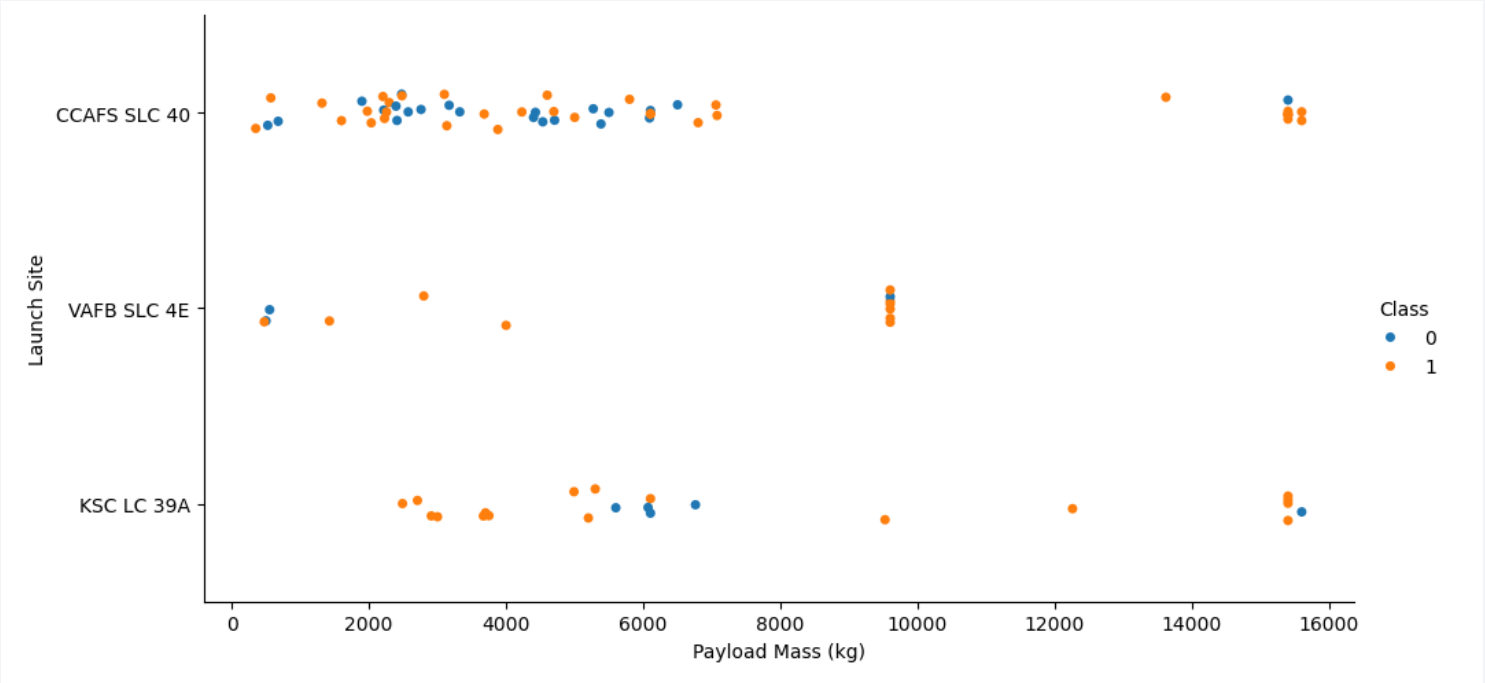
It can be seen that:

- Higher flight number is related with more successful (class = 1)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates



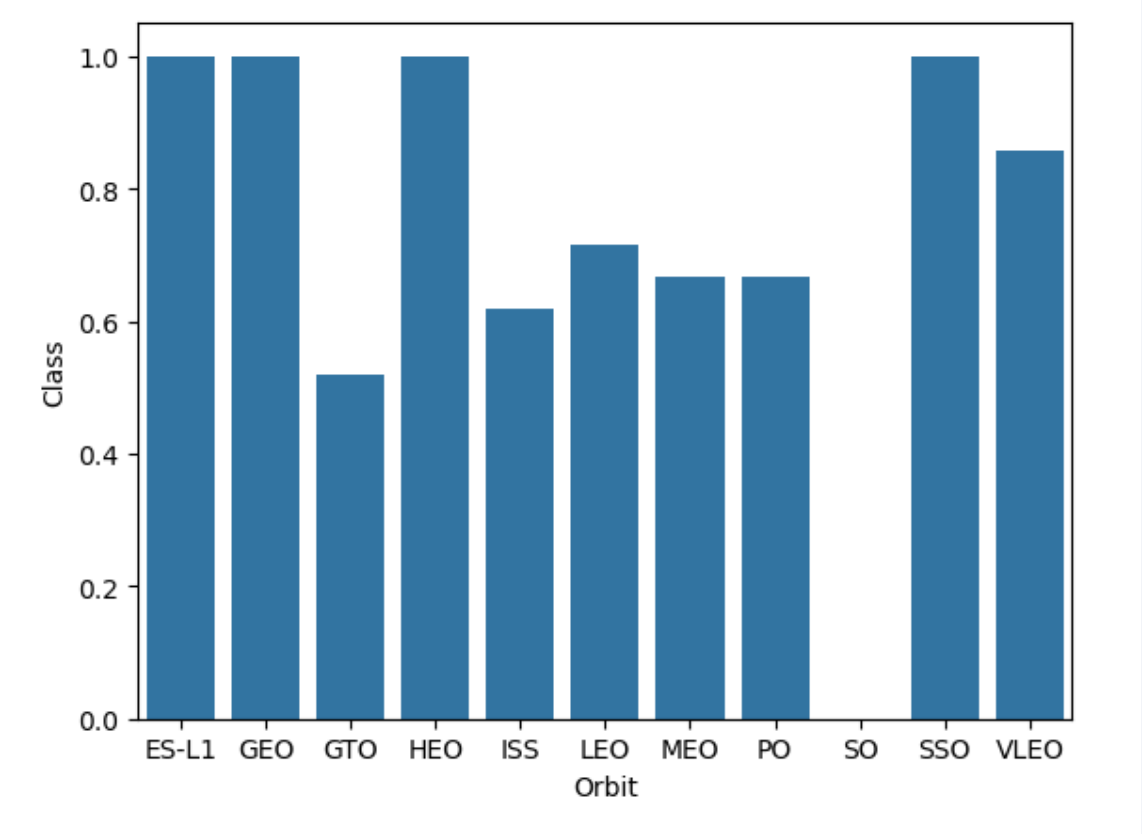
Payload vs. Launch Site

- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



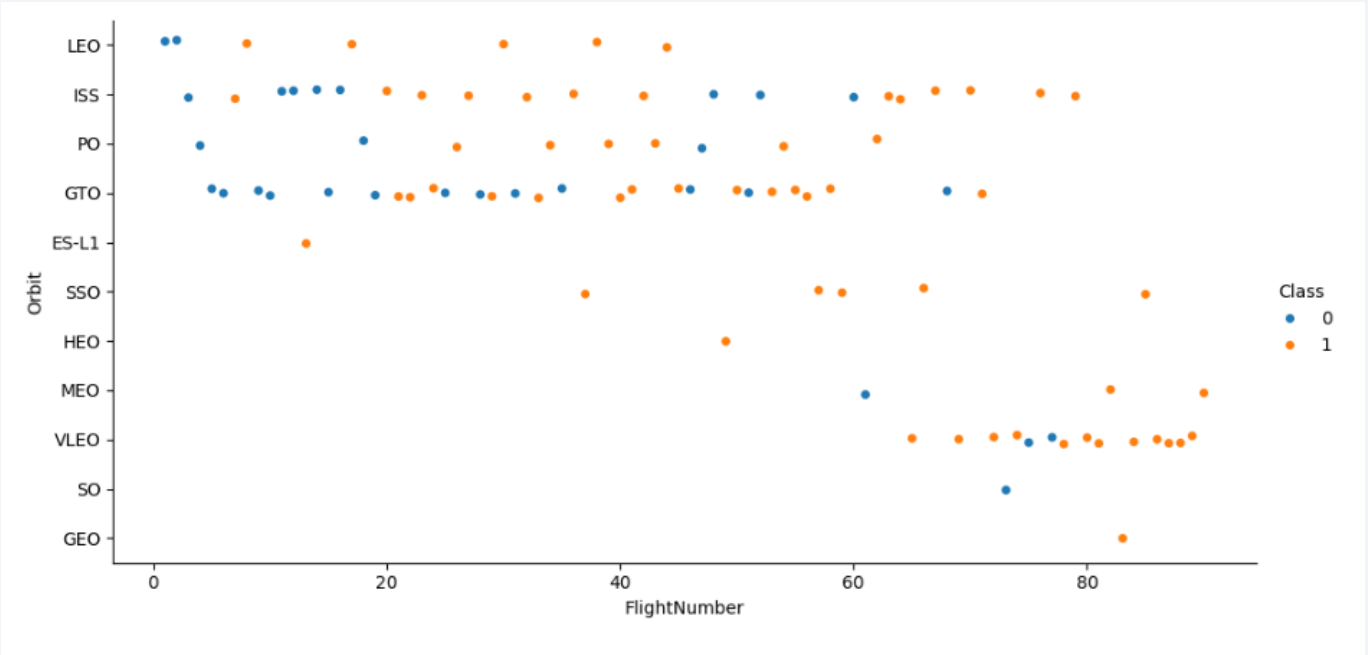
Success Rate (Class) vs. Orbit Type

- Understanding the class (1 to 0) as the succession rate percentage over 100, it can be seen:
- **100% Success Rate:** ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate:** GTO, ISS, LEO, MEO, PO
- **0% Success Rate:** SO



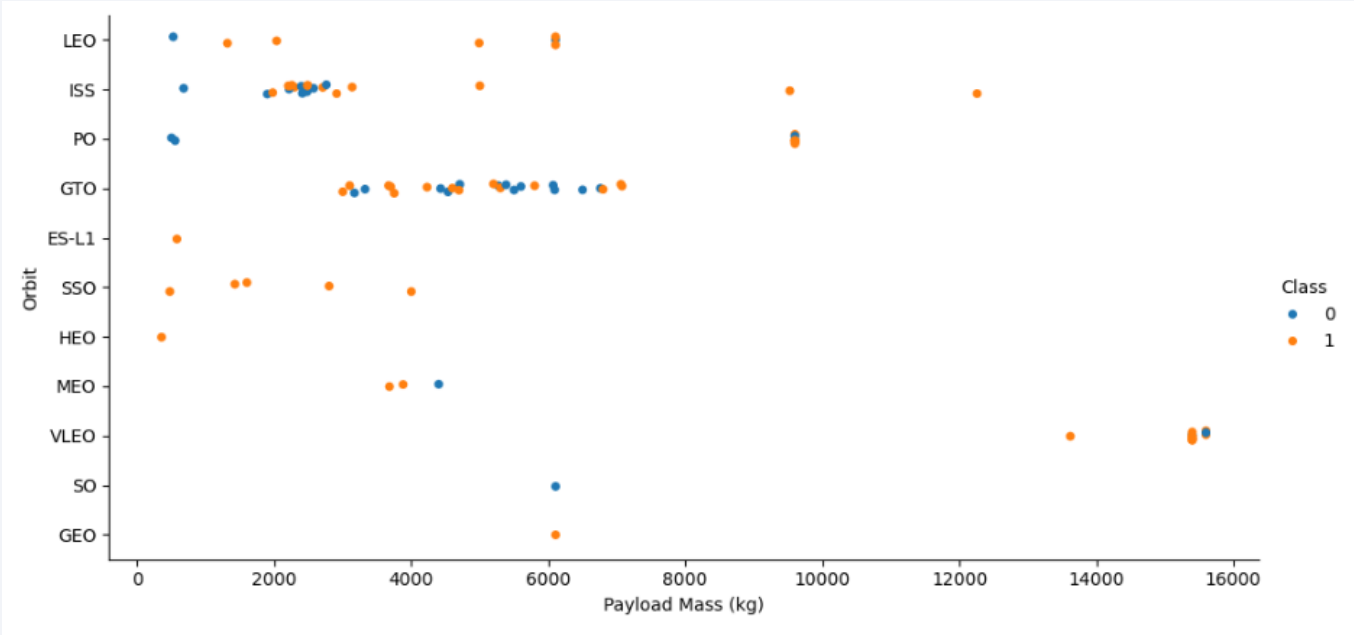
Flight Number vs. Orbit Type

- The success rate usually increases with the number of flights for each orbit
- There are orbits highly related with flight number: HEO, MEO, VLEO, SO and GEO orbits were studied later the first 50 flights
- Last flight numbers are mor successful than first ones.



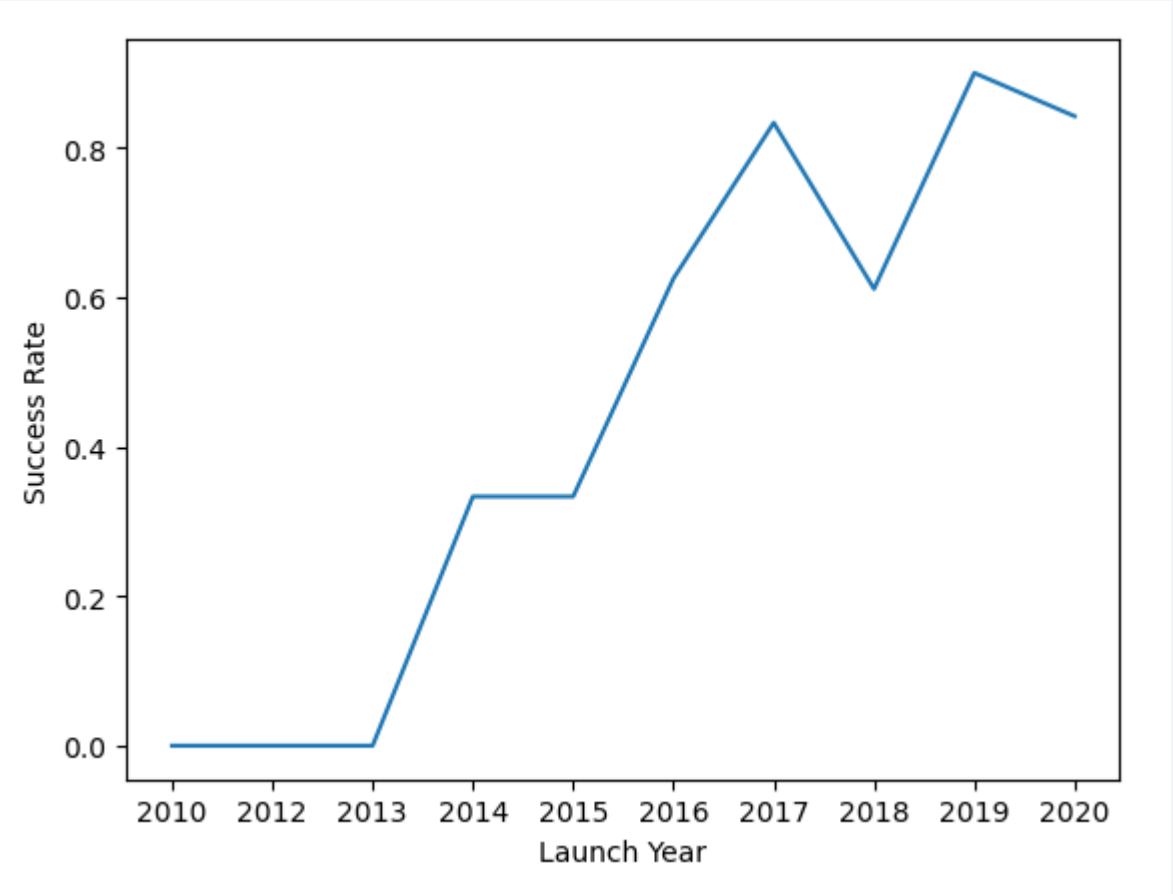
Payload vs. Orbit Type

- Heavy payloads worked better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with medium payloads
- Some orbits were studied with a narrow set of Payload mass while others were analyzed with a wider range of values.



Launch Success Yearly Trend

- It can be observed that, through the years, succession rate increased steadly.
- In the first 3 years not even one of the launches was successful



All Launch Site Names and starting with ‘CCA’

- On the left, a query to show the 4 different Launch sites.
- On the Right, it show the query to access to Records with Launch sites starting with ‘CCA’.

```
[8]: %%sql select distinct("Launch_Site")
     from SPACEXTABLE
```

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

[8]:

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

```
[31]: %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
```

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

[31]:

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

Payload Mass

- On the left, it is shown the calculation of the total payload carried by boosters from NASA and the query to access to it.
- On the right, the average payload mass carried by booster version F9 v1.1 is presented with its corresponding query.

```
[10]: %%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTABLE
      where Customer like 'NASA (CRS)'

       * sqlite:///my_data1.db
      Done.

[10]: sum("PAYLOAD_MASS__KG_")
      45596
```

```
[11]: %%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE
      where Booster_Version like "F9 v1.1"

       * sqlite:///my_data1.db
      Done.

[11]: avg("PAYLOAD_MASS__KG_")
      2928.4
```

First Successful Ground Landing Date

- With this query it is possible to see the date of the first successful landing outcome on ground pad: the 22th of December 2015,

```
[12]: %%sql select min(Date) from SPACEXTABLE
      where Landing_Outcome like 'Success (ground pad)'

       * sqlite:///my_data1.db
      Done.

[12]:  min(Date)
      ----------
       2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 kg and 6000 kg

- This query shows a list of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 6000 kg.

```
[13]: %%sql select Booster_Version from SPACEXTABLE
      where Landing_Outcome like 'Success (drone ship)' and 4000<"payload_mass__kg_"<6000
```

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

[13]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failed Mission Outcomes

- It is shown how to calculate the total number of successful and failed mission outcomes.

The result is:

- 100 successes
- 1 failures

```
[14]: %%sql SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL FROM SPACEXTBL
      GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

[14]:

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

Boosters Carried Maximum Payload

- Here, the query to list the names of the booster which have carried the maximum payload mass.

```
[18]: %%sql select Booster_Version from SPACEXTABLE
      where "Payload_Mass__kg_" = (select max("Payload_mass__kg_") from SPACEXTABLE);
```

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

[18]:

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

- Here, the query to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

[117]:
```
%%sql select substr(Date, 6,2) as Month, substr(Date,0,5) as Year,
"Landing_Outcome", "Launch_Site", "Booster_Version"
from SPACEXTABLE
where substr(Date,0,5)='2015' and "Landing_outcome" = "Failure (drone ship)";
```

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

[117]:

Month	Year	Landing_Outcome	Launch_Site	Booster_Version
01	2015	Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1012
04	2015	Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1015

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

- With this query it is ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

[130]:
```
%%sql select Landing_Outcome, count(*) as 'quantity'
from SPACEXTABLE
where "Date" between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by quantity desc;
```

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

[130]:

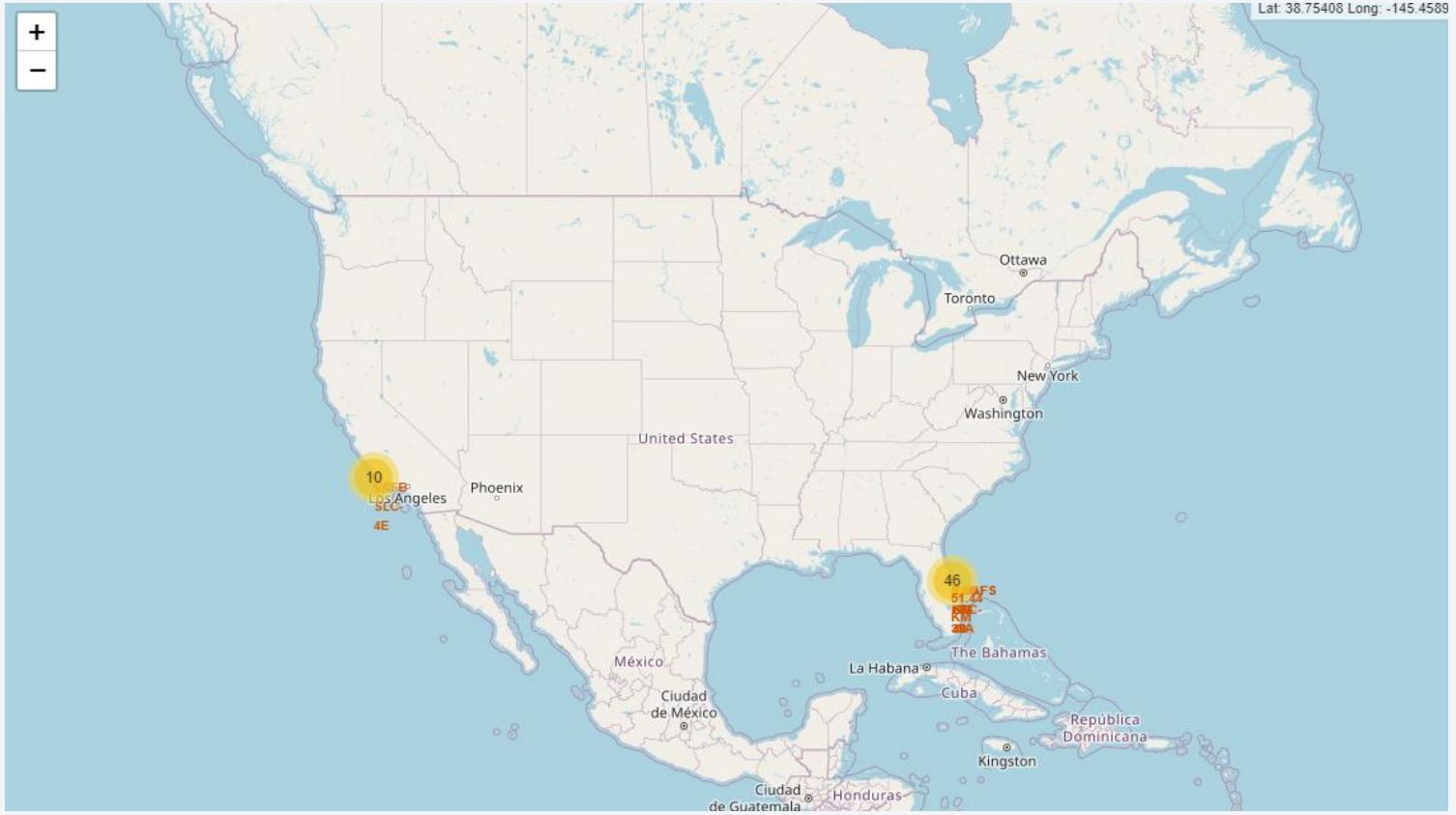
Landing_Outcome	quantity
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

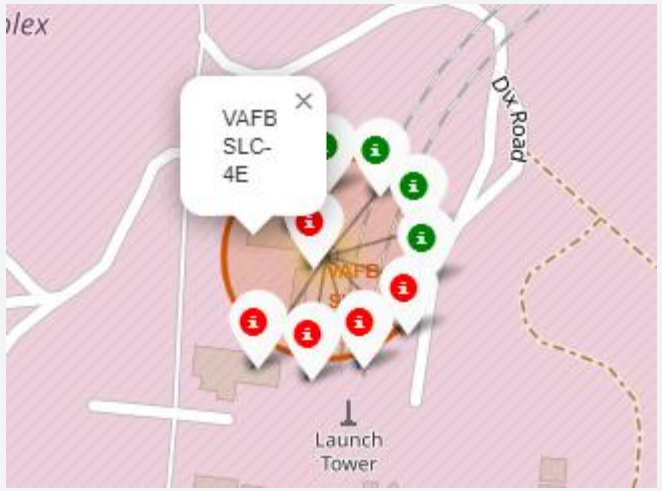
Launch Sites

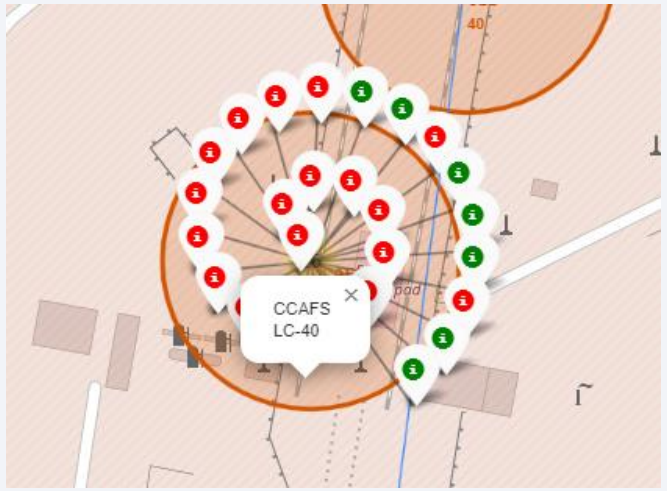
- It can be seen the largest number of launches were carried out in the East coast, in Florida, while in California, only 10 launches took place. All of them, in de USA.

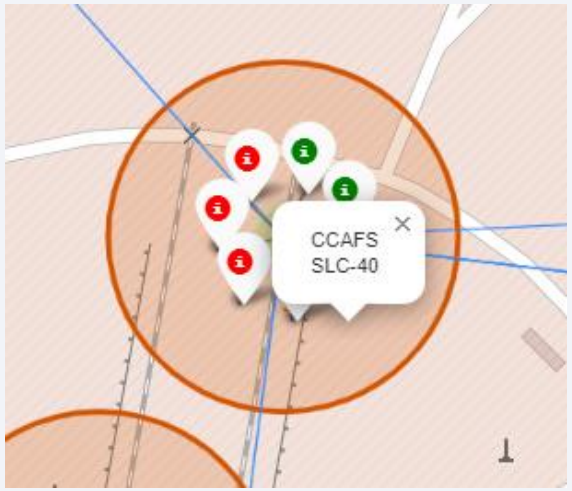


Color-labeled launch outcomes

- Here it is possible to see 3 different locations, with its own name labels and a marker for each successful (green) and failed (red) launch.
- The one on the left, represents the location placed in California. The other two, are located in Florida.

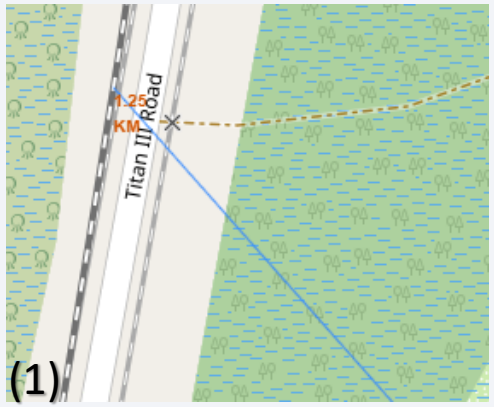


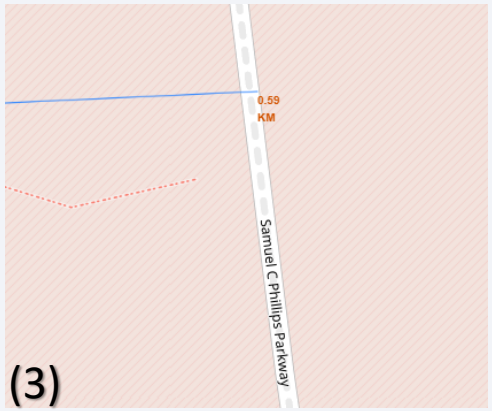


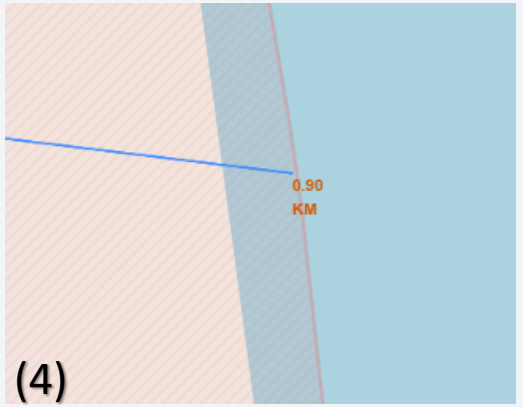


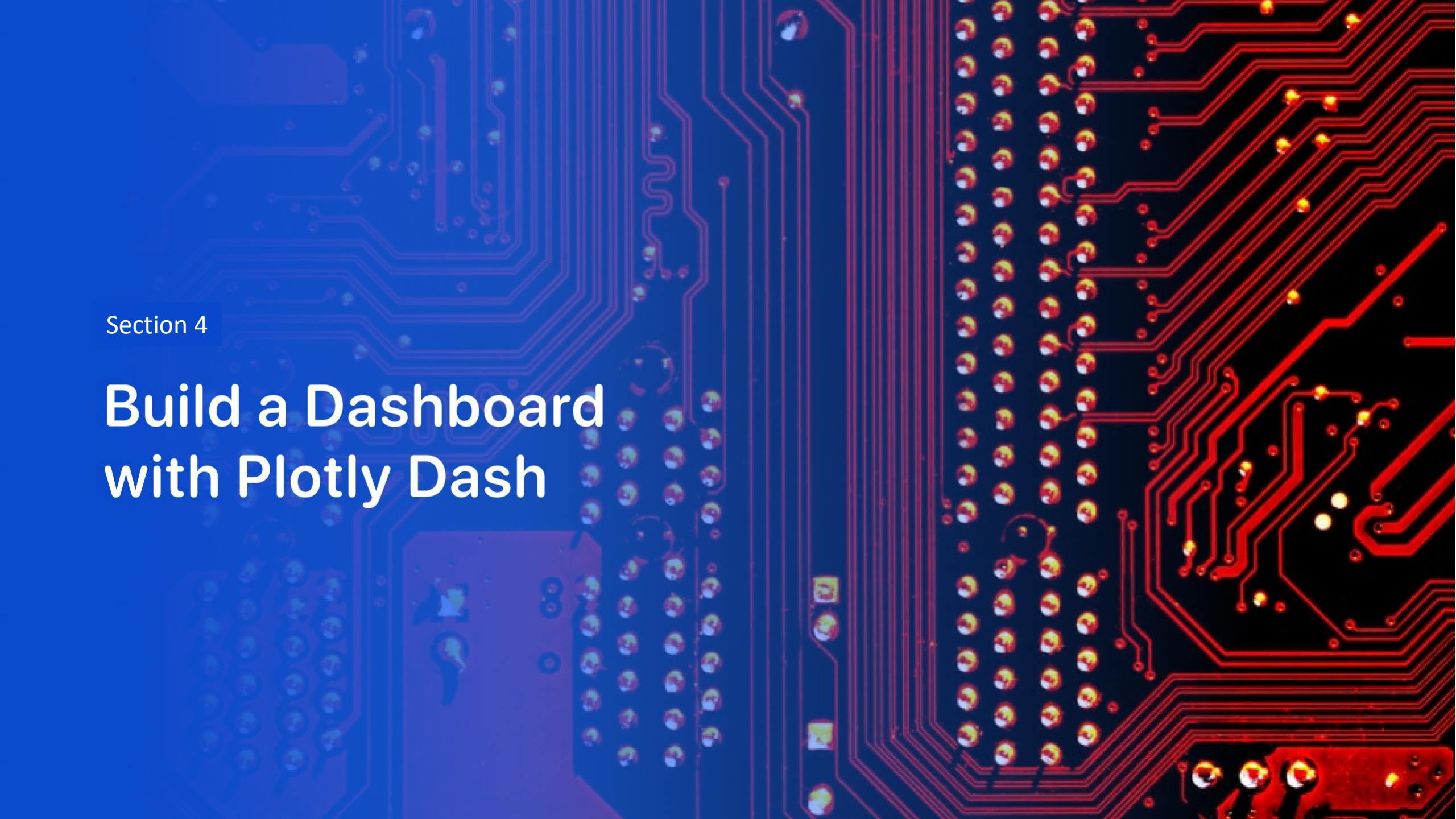
Distances from CCAFS SLC-40

It is shown screenshots of some selected places -(1) a railway, (2) a pathway, (3) a coastline, and (4) Melbourne- and its respective distances to the launch site CCAFS SLC-40.



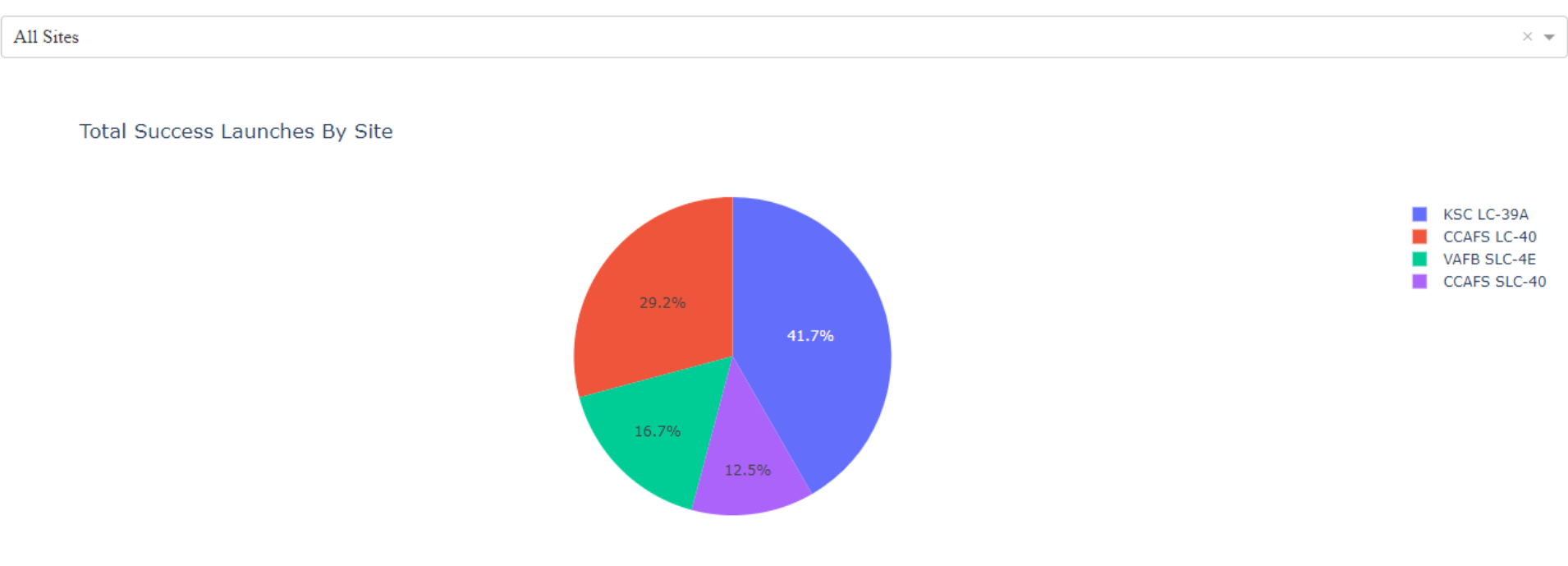





Section 4
Build a Dashboard with Plotly Dash

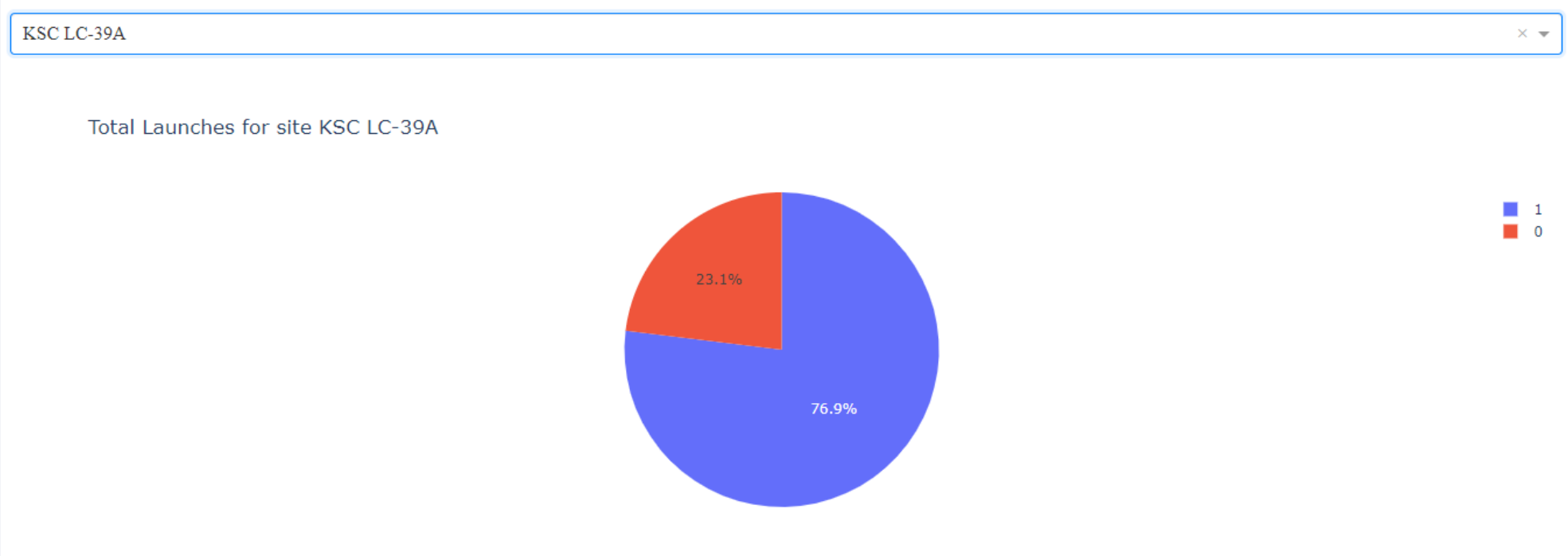
Launch Success by Site

- It can be seen a comparative of the success rate of the 4 different Launch Sites.
- It is observed that KSC LC-39A has the highest amount of success (41.7% of the total launches)



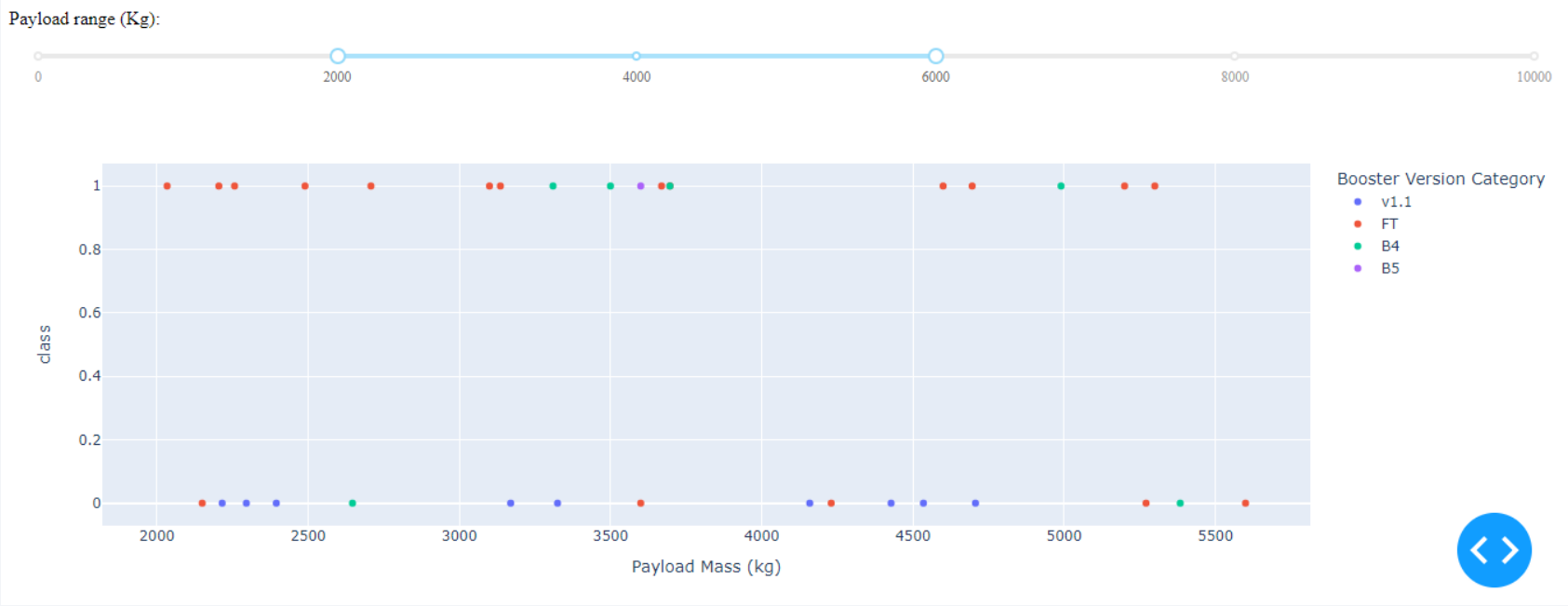
Launch Success of KSC LC-39A

- Exploring all launches occurred in KSC LC-39A, it is observed that 76.9 % of launches were successful, while 23.1 % were unsuccessful.



Payload Mass and Success

- It can be seen successes (class 1) and failures (class 0), for all Launch sites, colored according to different booster versions. The slider above allows you to select different payload mass ranges.
- For example, if we select payload mass in 2000 kg to 6000 kg range, FT booster version has the highest number of success launches, as it is shown below.

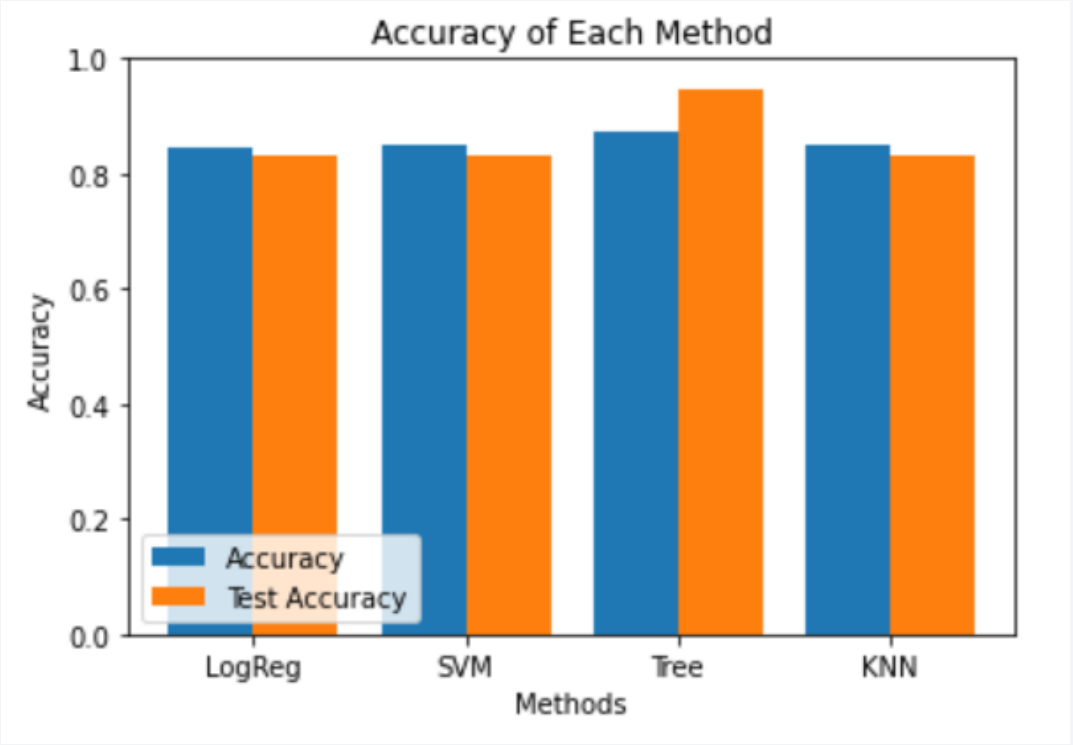


Section 5

Predictive Analysis (Classification)

Classification Accuracy

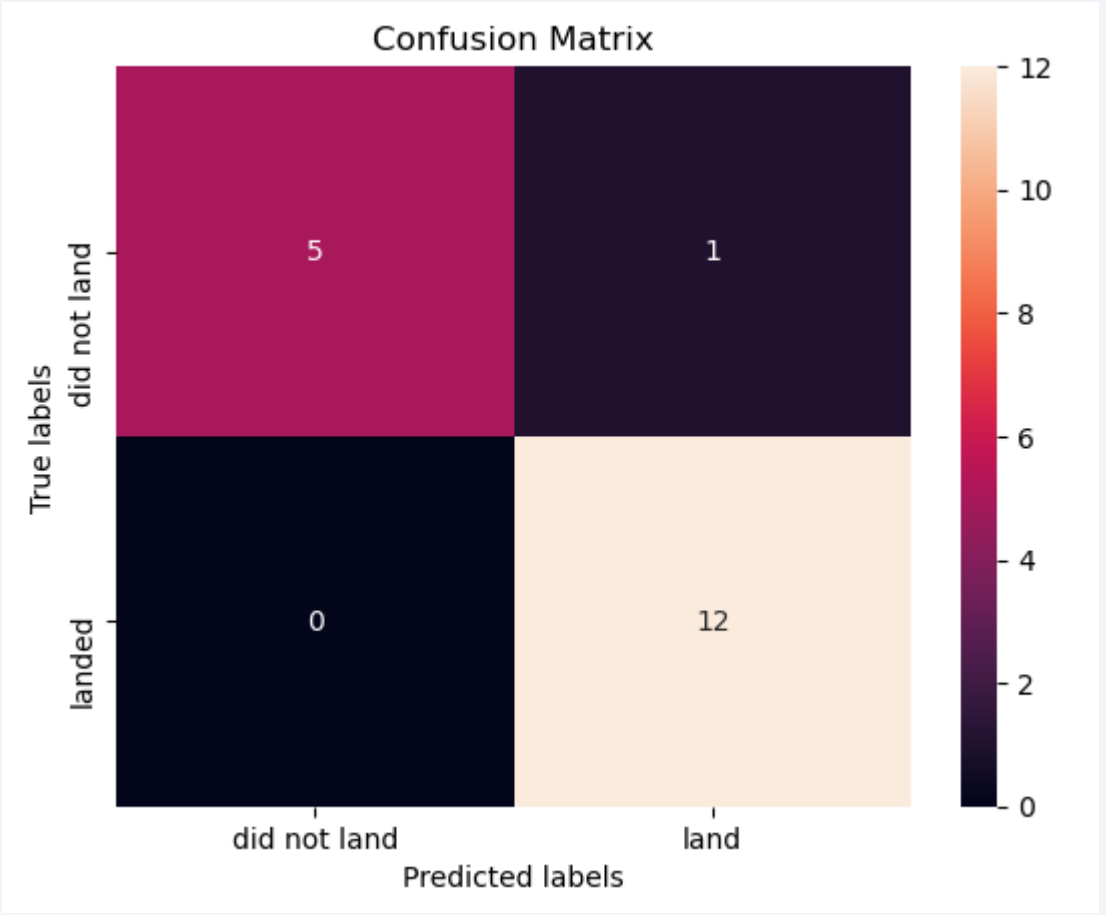
- There were compared 4 models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-nearest Neighbors.
- Decision Tree Classifier showed the highest train accuracy (R^2 =0.8889) and test accuracy (R^2 = 0.9444).



Confusion Matrix

- The confusion matrix for the Classification Tree model is shown to the right.

- It can be seen that the model classified almost correctly landed and not landed cases, but it failed just in the classification of one not landed case as a landed.



General Conclusions

- The best launch site is KSC LC-39A;
- Launches above 9,000kg trend to be more successful;
- Launch success rate starts to increase in 2013 with a steady growth.
- Orbits ES-L1, GEO, HEO, SSO, VLEO have the highest success rate.
- KSC LC-39A have the most successful launches of any sites.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- The complete work, Jupyter Notebooks and Python codes can be found in the link https://github.com/FranGazta/Data-Science-Capstone-Coursera
- As an improvement for model tests, it could be significant to randomize the seed using np.random.seed variable.
- Another improvement could be to analyze newer data and add launches from 2020 to 2024.
- It could be easier to plot graphs with dates in x-axis, instead of flight numbers. It is related, but is straightforward to see the date for analyzing the improvement through years.

Thank you!