Winning Space Race with Data Science

DOUGLAS RAFAEL SILVA SOUSA

21 June 2023

Outline

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

Executive Summary

The reuse of the first stage in launching a rocket makes its cost much more economical. In that work it was predicted whether the first stage of a SpaceX Falcon 9 Rocket would land successfully. Data science methodology was followed, going through data collection, data manipulation, exploratory data analysis, visualization and model development. The KNN, Decision Tree, Logistic Regression and SVM models were built, which at the end of the evaluation, presented the same statistic of accuracy of 83%.

Introduction

Taking into account the savings that can be realized by reusing the first stage of a launched rocket and being able to determine whether the first stage will land successfully, it is feasible to determine the cost of a launch, allowing an alternative company to make an offer against SpaceX, which uses this strategy to make its launches cheaper.

This work has as main objective to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully or not. As secondary objectives, we can cite the study of relationships between different variables.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API data request and web scraping throw data from a Wikipedia page.
- Perform data wrangling
  - Use of Python Pandas library and various methods to make necessary adjustments to the database.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The models (KNN, SVM, Decision Tree and Logistic Regression) were built with GridSearchCV to find the best hyperparameters and evaluated by error metrics.

Data Collection

- Data were collected in two different ways:
  - In the first, a GET request was made to the SpaceX API, through the requests library, the content of the response was decoded as a JSON. In addition, a basic formatting of the data was performed, replacing null values and including only Falcon 9.
  - In the second way, an HTML table of Falcon 9 launch records from Wikipedia was extracted, using the BeautifulSoup library, in addition to the conversion process being carried out in a Pandas data frame.

Data Collection – SpaceX API

1. Request the SpaceX API launch data using the GET method of request library;
2. Normalize JSON response into a dataframe;
3. Adjustment and extraction of the necessary columns;
4. Create new pandas dataframe from dictionary;
5. Filter dataframe to only include Falcon 9 launches;
6. Handle missing values;
7. Export to CSV file.

- GitHub URL: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/blob/master/1.%20Space-X%20Data%20Collection%20API.ipynb

Data Collection - Scraping

1. Request launch data from the Wikipedia page;
2. Instantiate the BeautifulSoup object;
3. Extract the column or variable names from the HTML table header;
4. Create a Pandas data frame by parsing the launch HTML tables;
5. Export to CSV file.

- GitHub URL: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/blob/master/2.%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

Data Wrangling

1. Check % of null values and data type;
2. Verification of information (count of each category, for example) about the features;
3. Adjustment of the dataframe class/target, by creating a landing outcome label from Outcome column.

- GitHub URL: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/blob/master/3.%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Scatter plots are constructed to visualize the relationship between two variables. In this work, this type of graph was constructed for the variables: Flight Number and Launch Site, Payload Mass and Launch Site, FlightNumber and Orbit type, Payload Mass and Orbit type;
- Bar graphs are constructed to facilitate the comparison of values between various categories of a feature. In this work, bar graphs were used to compare the Success Rate for different Orbit Types;
- A line graph was used to show Success rate over the years. It can demonstrate the trend over time.

- GitHub URL: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/blob/master/5.%20EDA%20with%20Visualization%20using%20Matplotlib%20and%20Seaborn.ipynb

EDA with SQL

- The following SQL queries were performed for EDA:
  - 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 succesful landing outcome in ground pad was achieved;
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;
  - List the total number of successful and failure mission outcomes;
  - List the names of the booster_versions which have carried the maximum payload mass;
  - 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;
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- To view the query commands, follow the GitHub link: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/blob/master/4.%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Several objects were created and added to a Folium map, such as markers, used to mark launch locations on the map (including success/failure). Lines were also drawn to show the distances between a launch site and landmarks.

- GitHub URL: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/blob/master/6.%20Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- A pie chart and a scatter chart were implemented.
- Interactions were handled by a Dropdown menu, which allowed you to select a specific launch location, or all. In the pie chart, it allowed you to view the success rate of launches in a specific location or general information about all locations.
- In addition, a slider for selecting the payload has also been added. In this case, the selection interacted in the scatter plot between the landing outcomes (class) and PayloadMass variables, to visualize the changes observed in different situations.

- GitHub URL: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/blob/master/7.%20Space-X%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

- Several models were built and evaluated against different error metrics in order to identify the one with the best performance.

1. Creation of the output variable (target) from the 'Class' column of the data;

2. Data standardization;

3. Division of data into training and test sets;

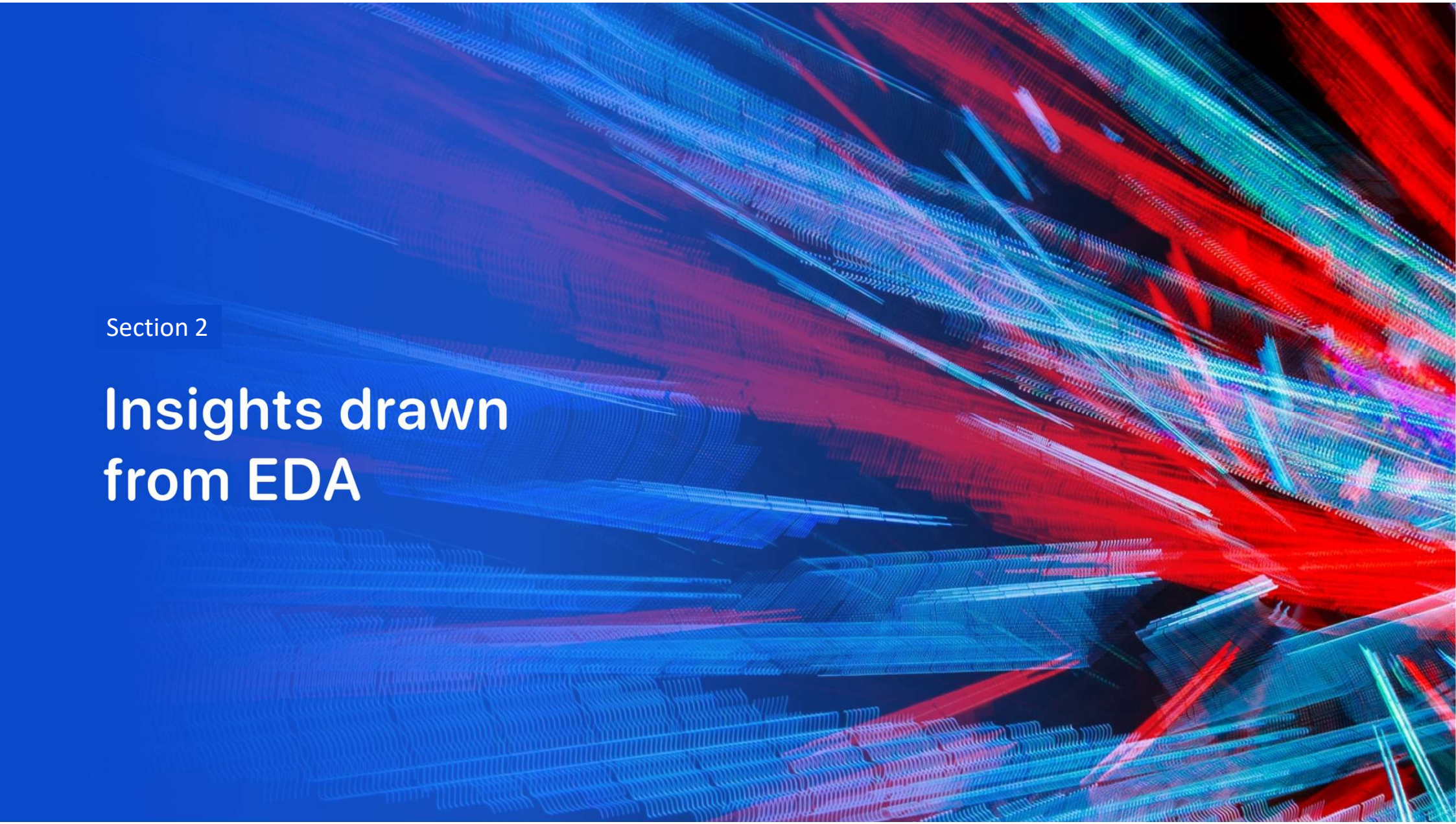
4. Using GridSearchCV to find the best hyperparameters among those available for each model (Logistic Regression, SVM, Decision Tree and KNN);

5. Evaluation of each model based on the accuracy obtained in the test set and in the confusion matrix.

- GitHub URL: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/blob/master/8.%20SpaceX%20Machine%20Learning%20Prediction.ipynb

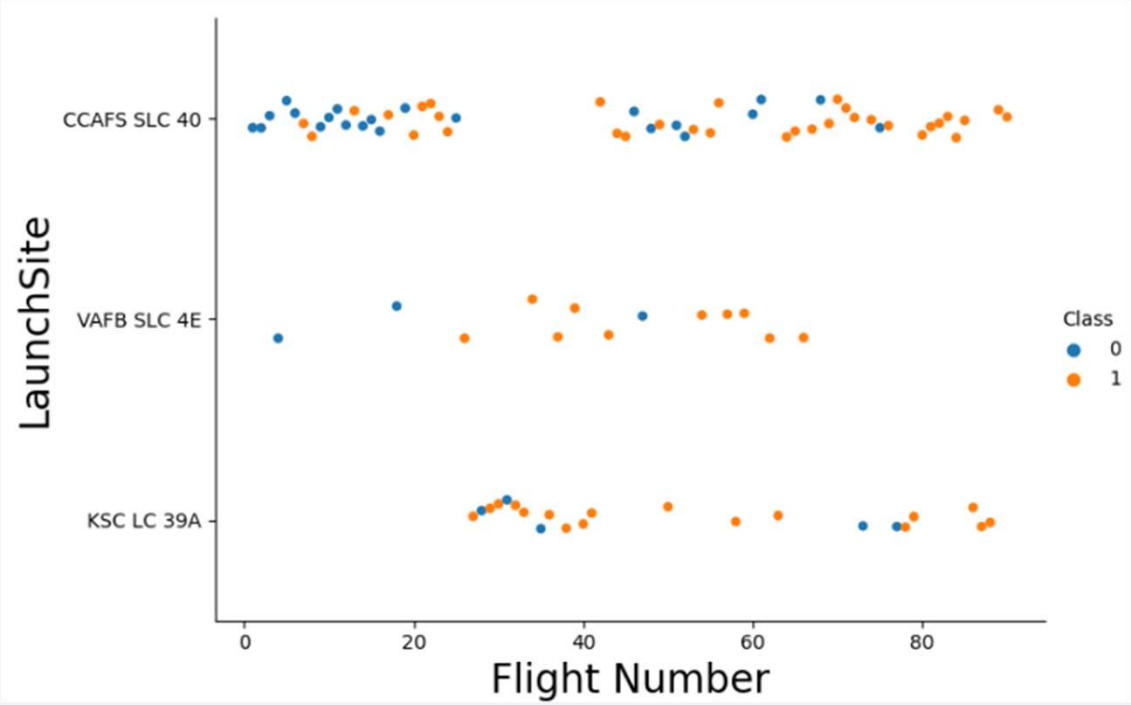
Results

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


Section 2
Insights drawn
from EDA

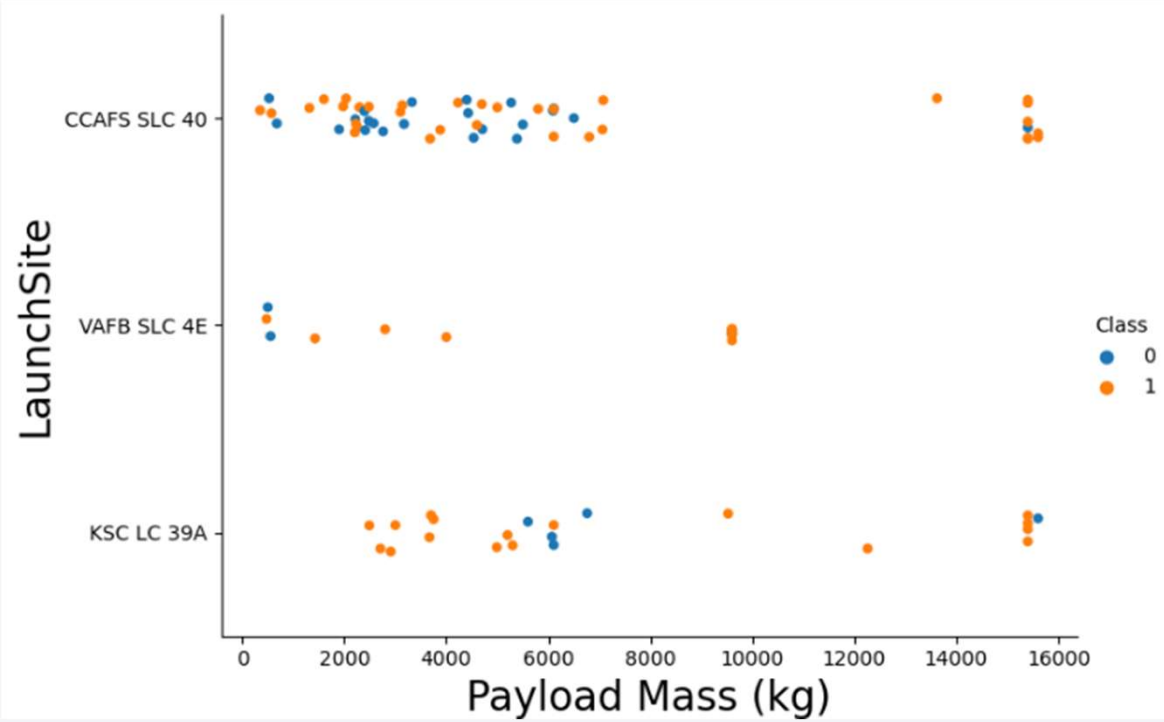
Flight Number vs. Launch Site

- Class 0 (blue) indicates failure, class 1 (yellow) indicates success;
- This figure shows that the success rate increased as the number of flights increased, class 1.



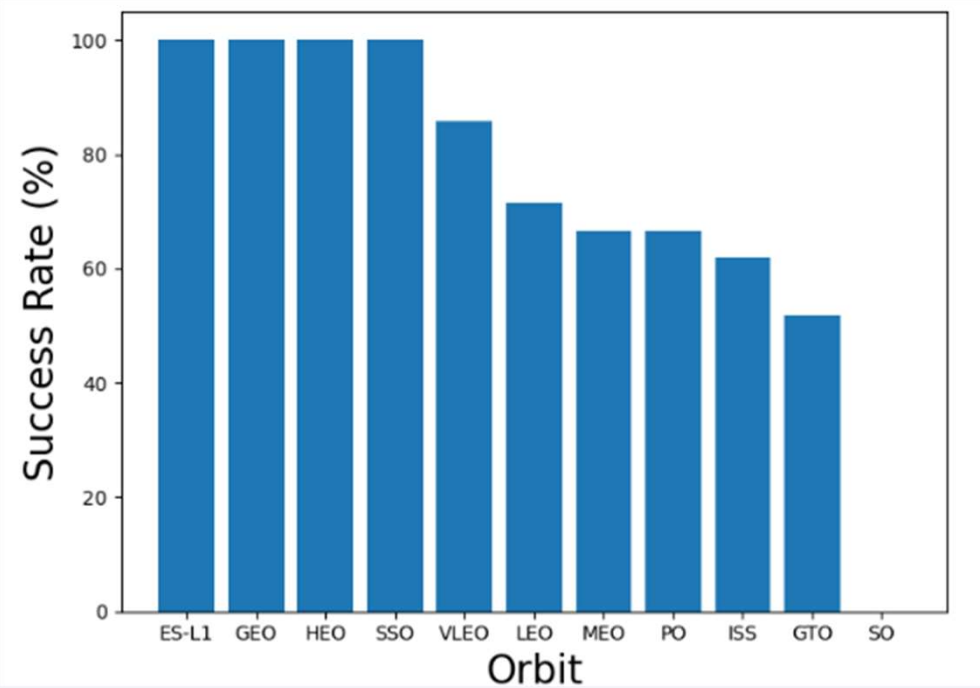
Payload vs. Launch Site

- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass, i.e., greater than 10000 kg.



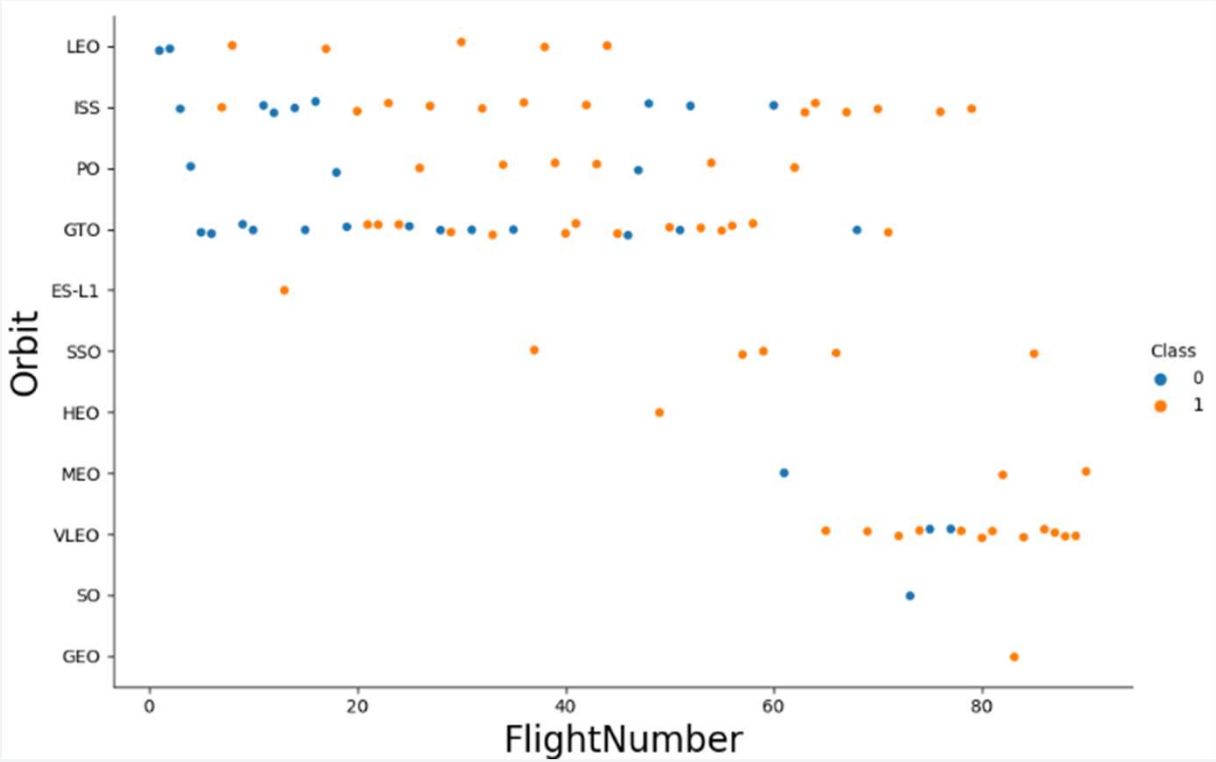
Success Rate vs. Orbit Type

- Orbits SSO, HEO, GEO, and ES-L1 have 100% success rates, while SO orbit have 0% success rate.



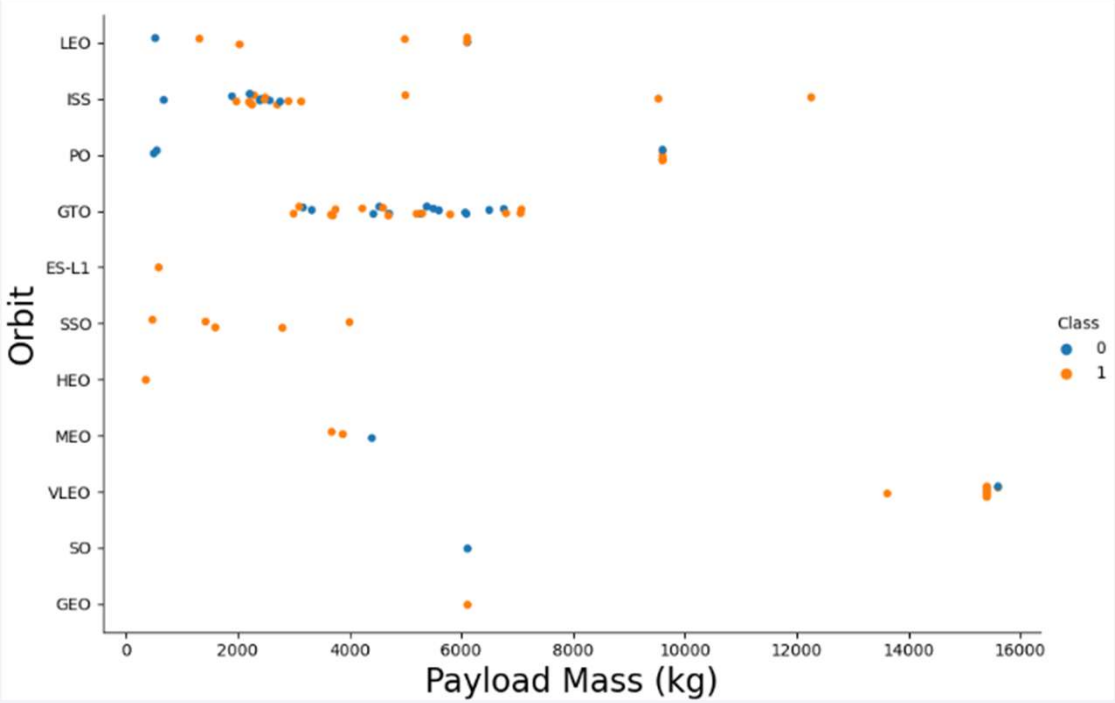
Flight Number vs. Orbit Type

- In LEO orbit the success appears related to the number of flights;
- There seems to be no relationship between flight number when in GTO orbit.



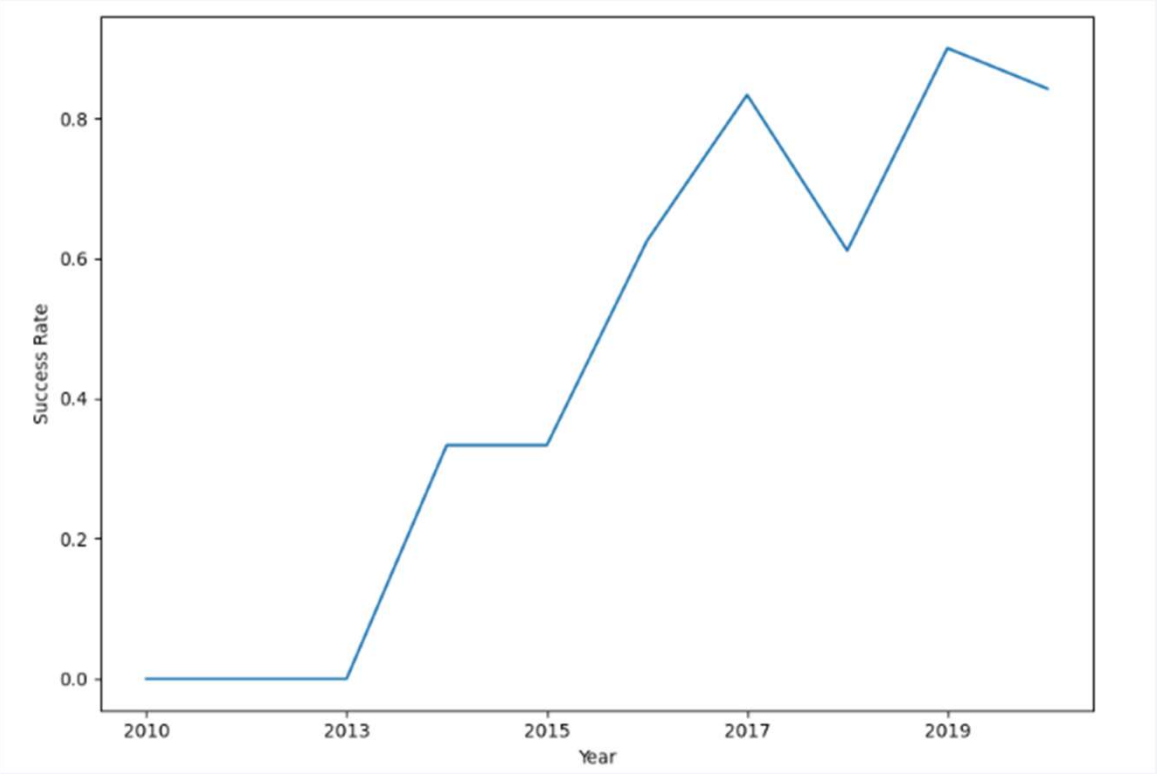
Payload vs. Orbit Type

- With heavy payloads the successful landing are more for Polar, LEO and ISS;
- As for the GTO orbit, it is not possible to clearly distinguish this relation.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

- Find the names of the unique launch sites
  - The distinct command was used to return only the unique values of the required column.

```
[ ]  %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'

- Find 5 records where launch sites begin with `CCA`
  - Use of the Where clause together with Like to search for the desired pattern and limitation to 5 records.

```
%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_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - Using the sum aggregation function to find the desired result.

```
[ ] %sql select Customer, sum(PAYLOAD_MASS__KG_) as 'TPM (kg)' from SPACEXTBL where Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
 Customer  TPM (kg)
NASA (CRS) 45596.0
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - Using the average aggregation function and the Where clause with the Like for the desired pattern.

```
[ ] %%sql
    select avg(PAYLOAD_MASS__KG_) as 'AVG_PM (kg)' from SPACEXTBL where Booster_Version like 'F9 v1.1%'

     * sqlite:///my_data1.db
    Done.
       AVG_PM (kg)
    2534.6666666666665
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - It was filtered by the desired pattern and ordered ascending by date, selecting only the first value (the smallest).

```
[ ] %sql select Date from SPACEXTBL where Landing_Outcome = 'Success (ground pad)' order by date(Date) limit 1;
 * sqlite:///my_data1.db
Done.
   Date
22/12/2015
```

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
  - It was filtered by the desired composite pattern with the Where clause.

```
[ ] %%sql
    select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_
    > 4000 and PAYLOAD_MASS__KG_ < 6000

     * 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

- Calculate the total number of successful and failure mission outcomes
  - Grouping was performed on the desired column, along with the count aggregation function.

```
[ ] %%sql
    select Mission_Outcome, count(Mission_Outcome) as Total from SPACEXTBL group by Mission_Outcome;
```

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

Mission_Outcome	Total
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - A subquery was used to perform the query, since the desired value of the filter was not known initially.

```
[ ] %%sql
    select distinct(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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Use of the Where clause with the desired information.

```
[ ] %%sql
    select substr(Date, 4, 2),Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where substr(Date,7,4)='2015' and Landing_Outcome = 'Failure (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
substr(Date, 4, 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

- 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
  - Use of the Where clause with the desired information, and also the count aggregator function

```
%%sql
select Landing_Outcome, count(Landing_Outcome) as Number from SPACEXTBL where (substr(Date,7,4) between '2010' and '2017') and (substr(Date,1,2) between '04' and '20') group by Landing_Outcome ORDER BY count(Landing_Outc
```

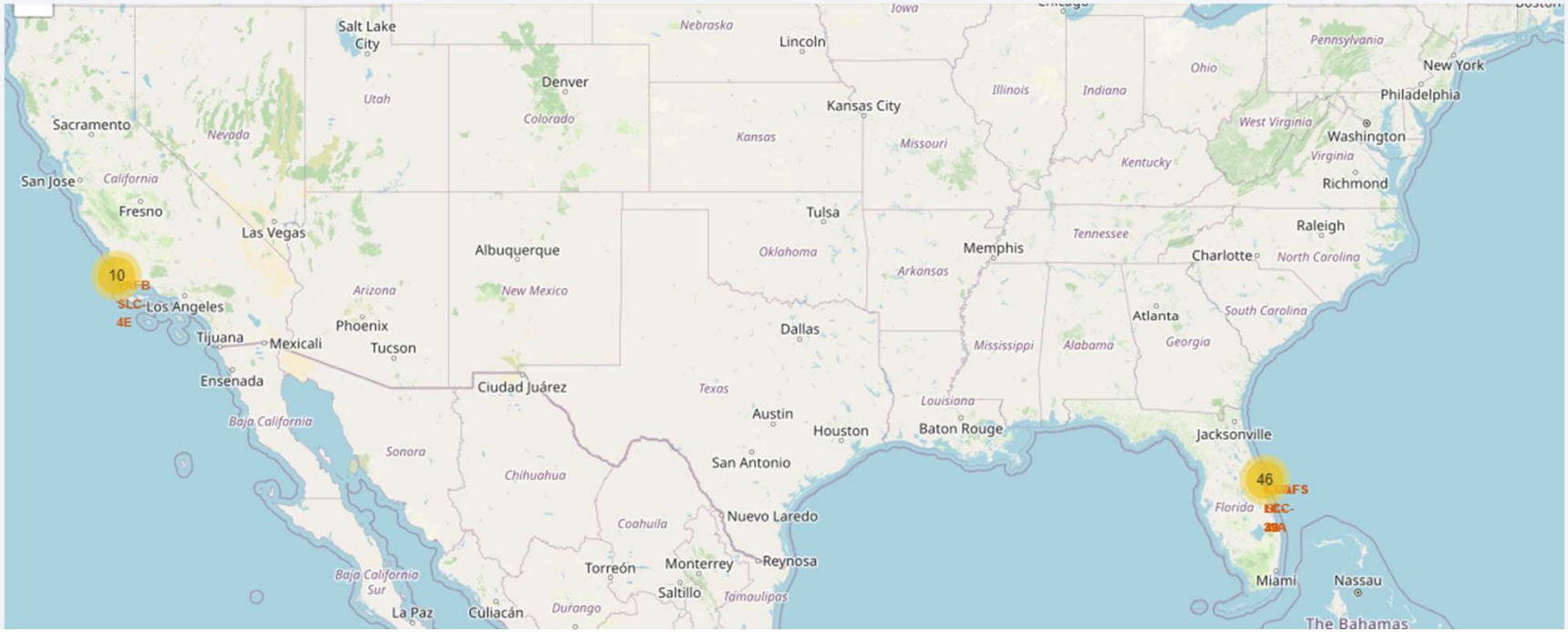
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Number
Success (ground pad)	7
No attempt	7
Success (drone ship)	6
Failure (drone ship)	3
Failure (parachute)	2
Controlled (ocean)	2


Section 3
Launch Sites
Proximities Analysis

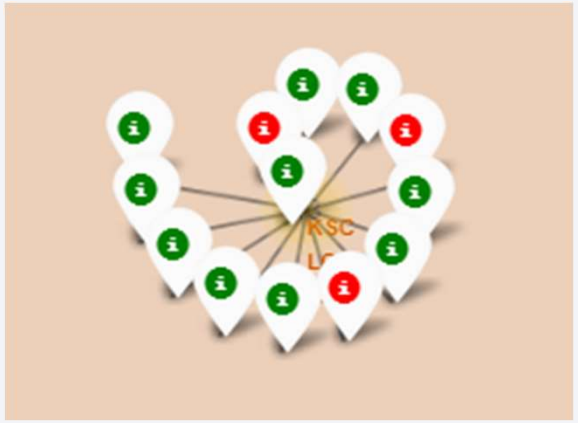
SpaceX Launch Sites Locations

- Yellow markers indicate SpaceX launch sites, as well as the amount coalesced (shown by the numeral), due to their close location.
- Launch sites are present on both the east and west coasts.



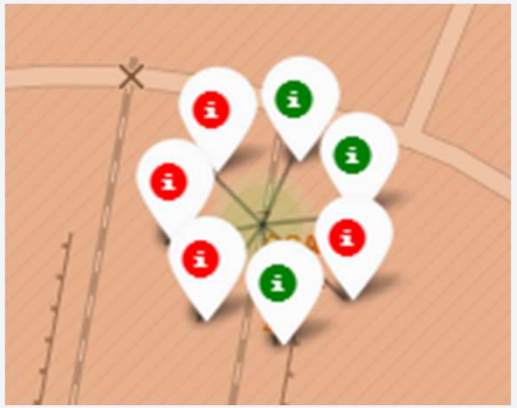
Launch outcomes for each location by colored markers

- The figure shows successful (green) and unsuccessful (red) landings at a location. This is important to get the information quickly and visually.
- In the Eastern Coast (Florida), Launch site KSC LC-39A has relatively high success rates compared to CCAFS LC-40 and CCAFS SLC-40.



KSC LC-39A

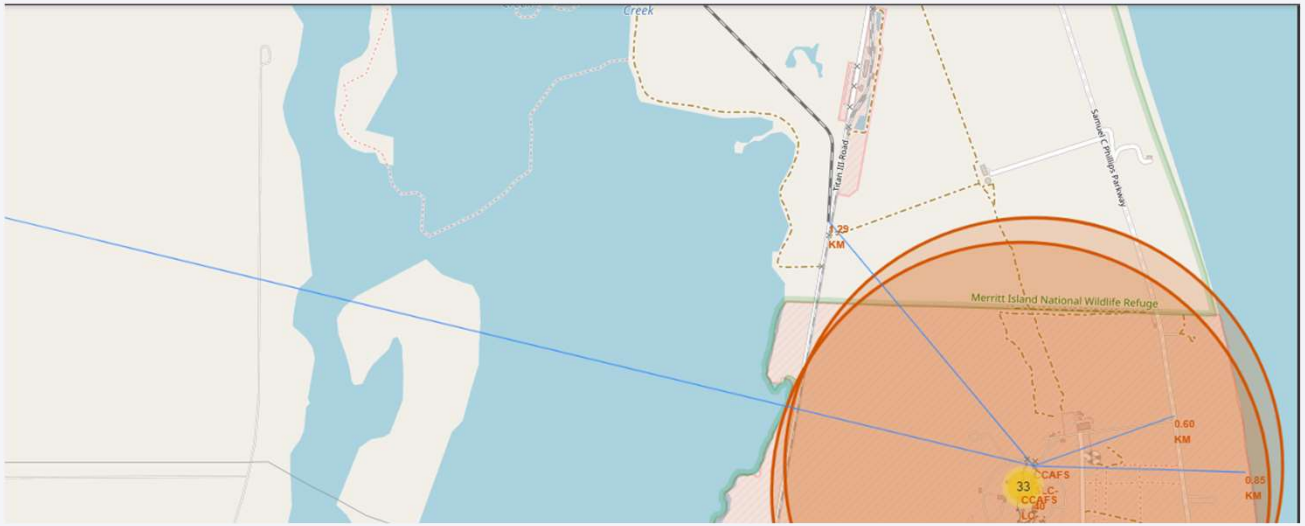
CCAFS LC-40



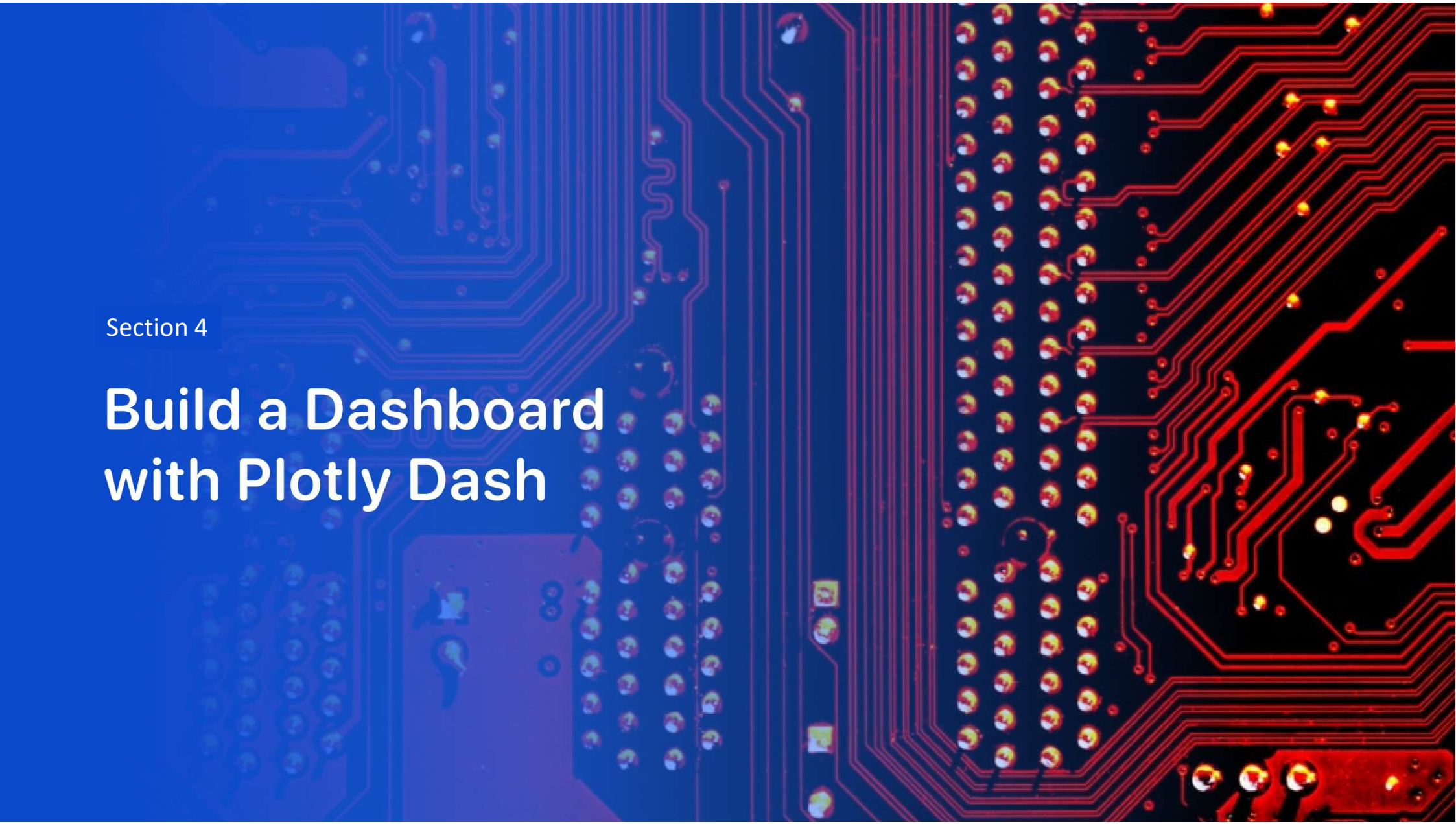
CCAFS SLC-40

Launch Site Proximities

- According to the figures, it is possible to verify that the specified launch site is relatively close to the highway, the railway and to the coast, but far from cities.

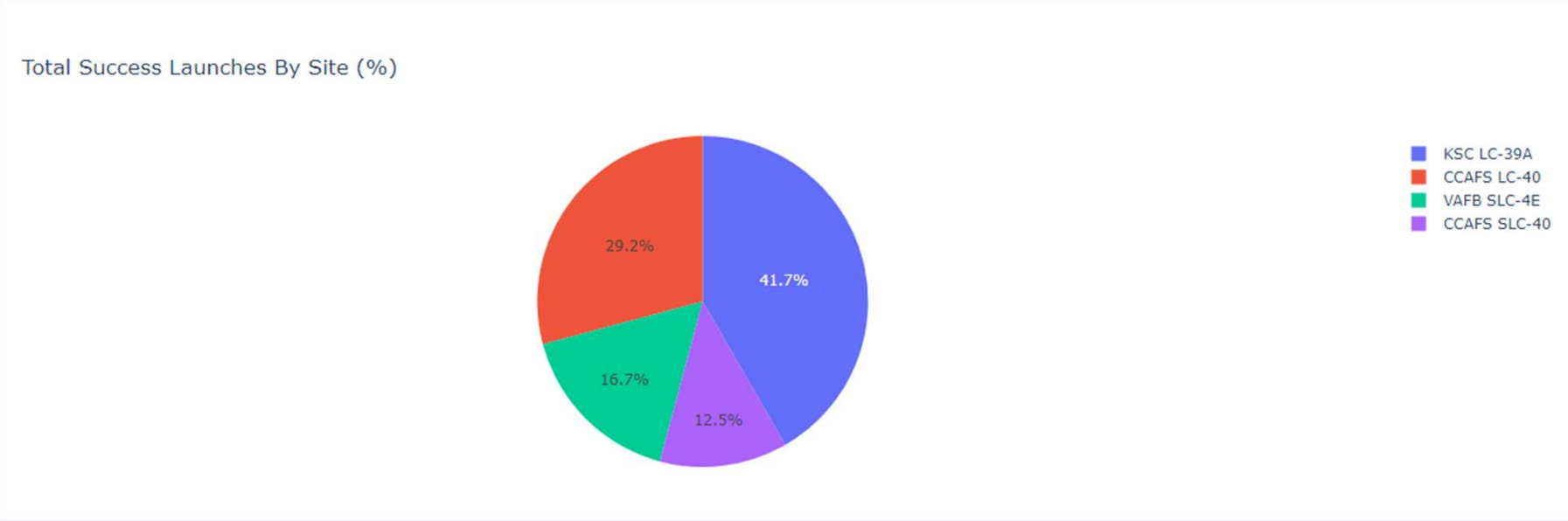


Local	Distance from CCAFS SLC-40 (km)
Highway	0.6
Railway	1.29
Coast	0.85
City	23.20


Section 4
Build a Dashboard with Plotly Dash

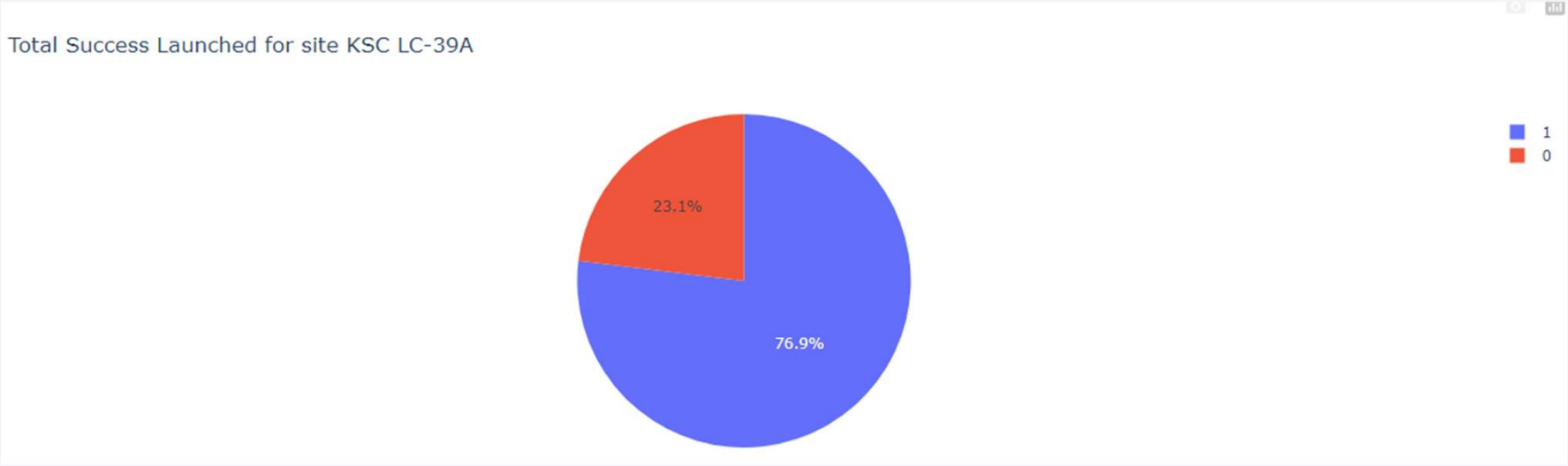
Total Successful Launches By Site

- The KSC LC-39A Launch site has the most successful launches;
- The CCAFS SLC-40 Launch site has the fewest successful launches.



Launch Site with Highest Success Ratio

- The KSC LC-39A success rate is 76.9% versus 23.1% failure.



Payloads vs Launch Outcome

- With the Payload Mass range up to 10000 kg, the booster version FT has the largest success rate.

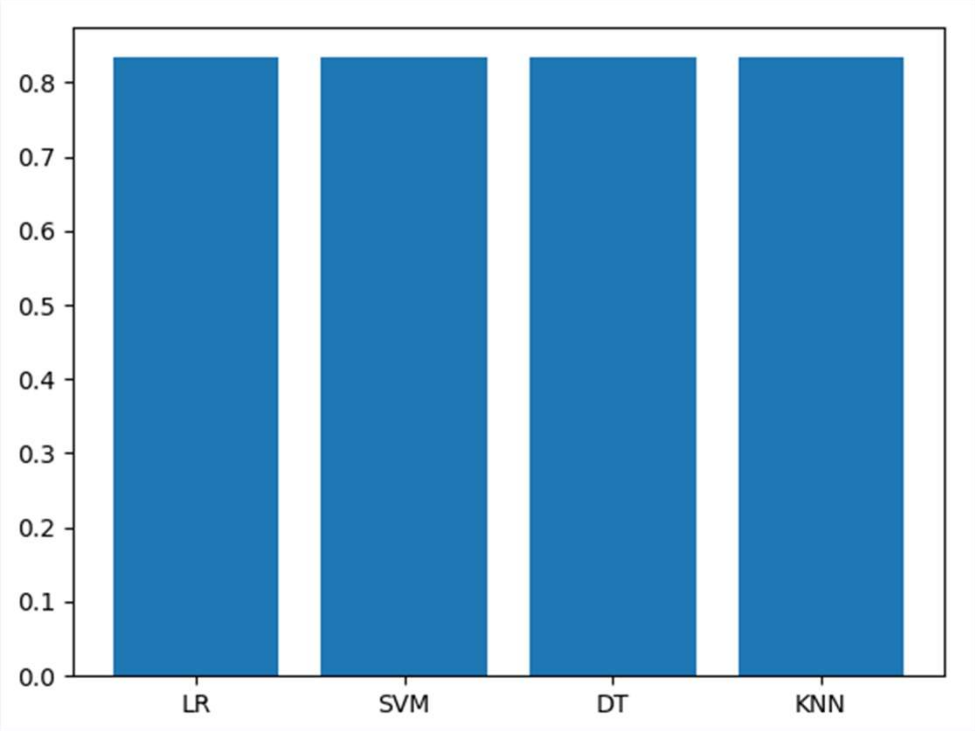
- With the variation between 0 and 2000 kg, there are more failures than success.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

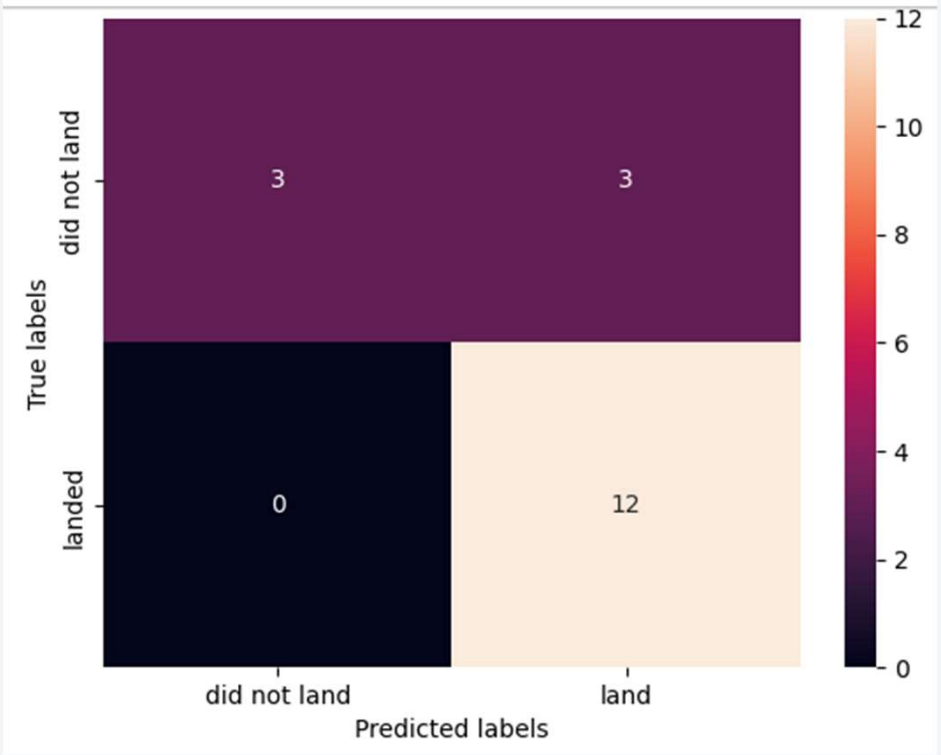
- All methods achieved the same result on the test set.



Confusion Matrix

- The model predicts the label "did not land" three times, getting it right three times (True Negative). Of the 15 predictions of the "land" label, the model was correct in a total of 12 (True Positive).

- The models showed a good result.



Conclusions

- The analysis showed that, in general, with the increase in the number of flights, there is also an increase in the success rate, for the different launch sites;
- The amount of launches at the VAFB-SLC location is lower than at the others, in addition, there are no rockets launched for heavy payload mass, i.e., greater than 10000 kg;
- The different orbits have different success rates, with the SSO, HEO, GEO, and ES-L1 orbits having the highest success rates (100%) and SO with no success at all;
- For Leo and ISS orbits, with a higher heavy payload, landings are more successful;
- Generally speaking, the success rate since 2013 kept increasing till 2020;
- All methods (Logistic Regression, SVM, Decision Tree and KNN) achieved the same result on the test set.

Appendix

- GitHub Repository: https://github.com/douglasrafa-dev/spaceX-1st-stage-landing-success-prediction/tree/master

Thank you!