Winning Space Race with Data Science

Noor Hasan
24/12/2021

Outline

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

Executive Summary

- SpaceY is a new company that compete SpaceX. For the new a missions that they have they must study Falcon9. Falcon9 is the newest rocket from SpaceX, so the company need to study and product the Falcon9 case before lunching. First of all determine the lunching cost. Second defined if SpaceX will be reuse Falcone9 first stage or not . The method used is machine learning to predict the results.

Introduction

- Project background and context

- SpaceY is a newly formed startup that will compete with SpaceX. They must study Falcon9 in order to prepare for the new tasks they have. Because Falcon9 is SpaceX's newest rocket, the business has to analyze and develop the Falcon9 casing before launching. To begin, establish the cost of lunch. Second, it was determined if SpaceX would reuse the Falcone9's first stage. Machine learning is applied to forecast the outcomes.

Introduction

- Objective

Apply different Machine learning Methods to Find the cost of the lunch and if the Falcon9 first stage will be reuse or not

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  the data was collected with two ways such as SpaceX api, web scrubbing with Beautifulsoup

- Perform data wrangling

  Processing the data to be clear an meaningful

- 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

- The data collected using two ways
  - SpaceX Rest API

  - Web scraping from Wikipedia using BeautifulSoup

Data Collection – SpaceX API

GitHub

Data was collected by making a request to the SpaceX API

Parse data from json to dataframe

Data processing and normalizing

Convert data to CSV

Data Collection - Scraping

GitHub

- HTML from Wikipedia
- Extract using Beautiful soup
- Normalize Data
- Convert data to CSV

Data Wrangling

The Objective of data wrangling is to find patterns in the data and determine what would be the features for training supervised model

GitHub

- Load Data
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome per orbit type
- Determine the success rate

EDA with Data Visualization

GitHub

Read Data

Create dummy variables to categorical columns

Cast all numeric columns to float64

EDA with SQL

Using SQL analyzed the data to find these outcomes:

GitHub

Display	Display the names of the unique launch sites in the space mission
Display	Display 5 records where launch sites begin with the string 'CCA'
Display	Display the total payload mass carried by boosters launched by NASA (CRS)
Display	Display average payload mass carried by booster version F9 v1.1
List	List the date when the first successful landing outcome in ground pad was acheived.
List	List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
List	List the names of the booster_versions which have carried the maximum payload mass.
List	List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
Rank	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

Build an Interactive Map with Folium

GitHub

- To visualize the Launch Data into an interactive map:
- The Latitude and Longitude Coordinates at each launch site are determined and added a Circle Marker around each launch site with a label of the name of the launch site.
- the dataframe launch_outcomes(failures, successes) transformed to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- Using Haversine's formula the distance calculated from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns.
- Lines are drawn on the map to measure distance to landmarks

Build a Dashboard with Plotly Dash

GitHub

- I used two graphs Pie chart and Scatter
- Pie Chart showing the total launches by a certain site/all
  - Pie chart is a clear way to show ratio
- Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions
  - The best sway to show relationship between two variable

Predictive Analysis (Classification)

GitHub

- Summary of how the model was built:
  - Load the dataset
  - Standardize the data
  - Split data into training and test sets
  - Decide which type of machine learning algorithms to use
  - Set the parameters and algorithms to GridSearchCV
  - Fit out datasets into the GridSearchCV and train our model
- Evaluation of the model:
  - Check the accuracy of each model
  - Tune the hyperparameters of each model
- Feature Engineering
  - Finding the best Performing Classification Model:
  - The model with the highest accuracy score is chosen as the algorithm

Results

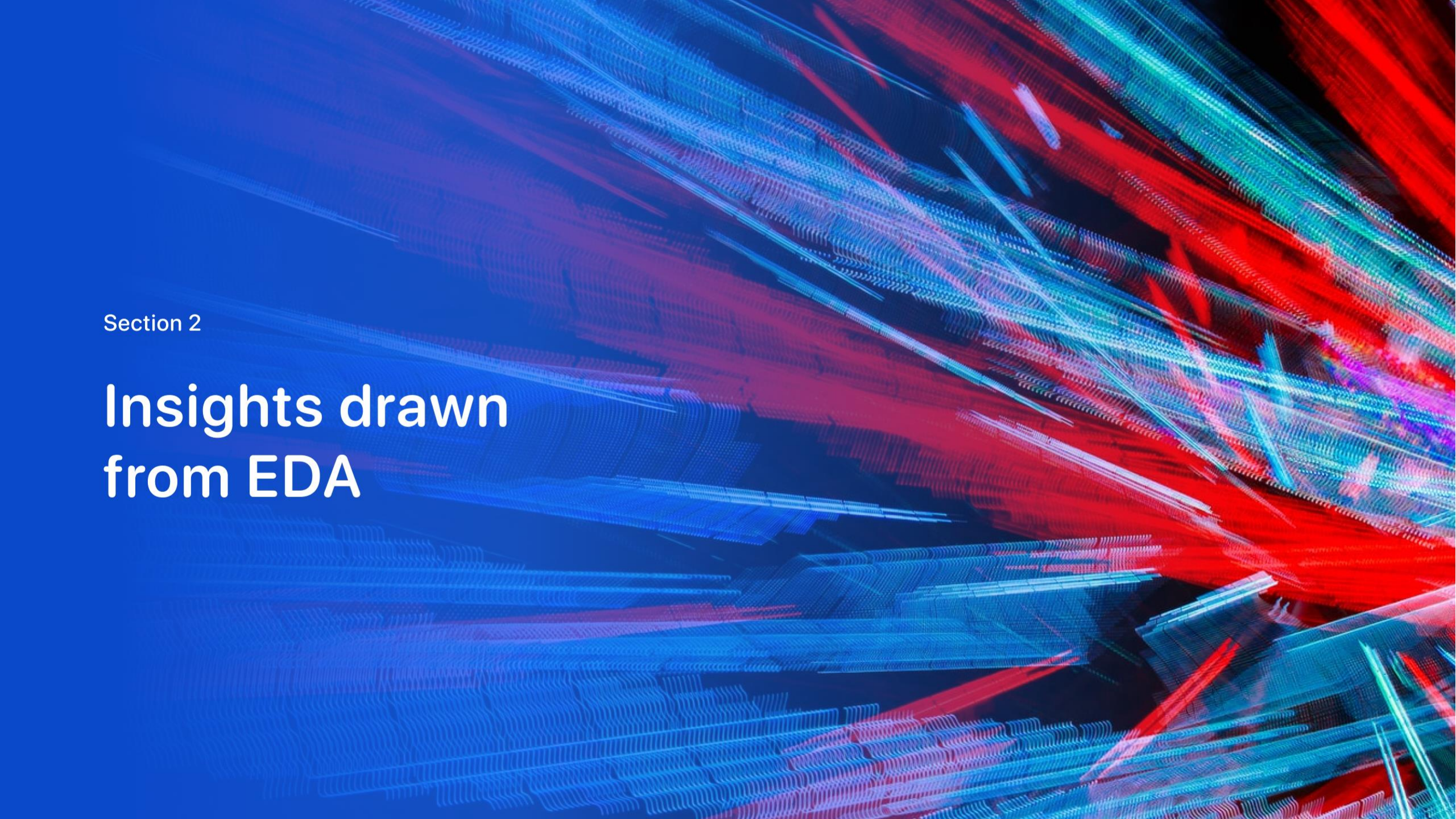
EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

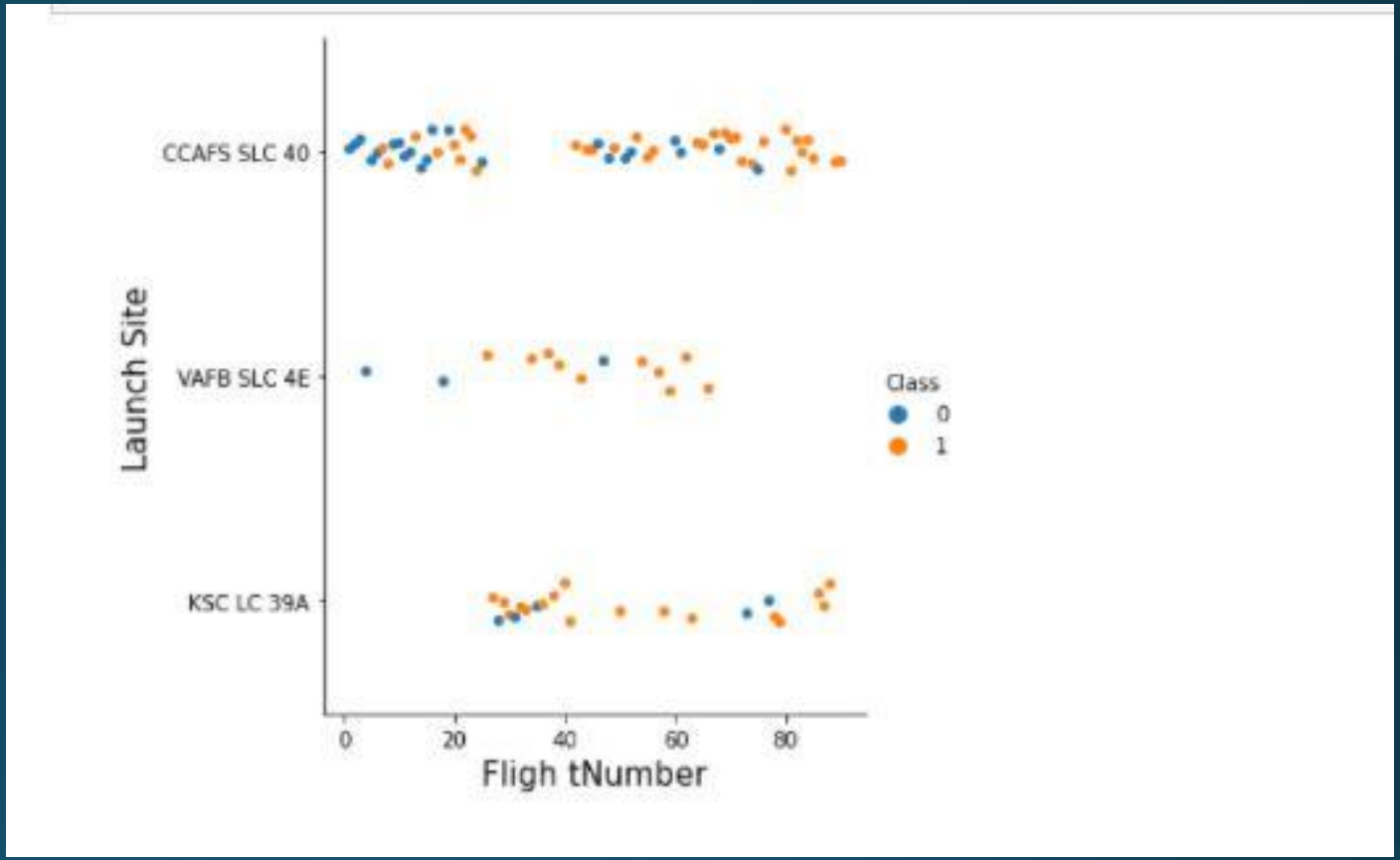
PREDICTIVE ANALYSIS RESULTS

Section 2

Insights drawn from EDA

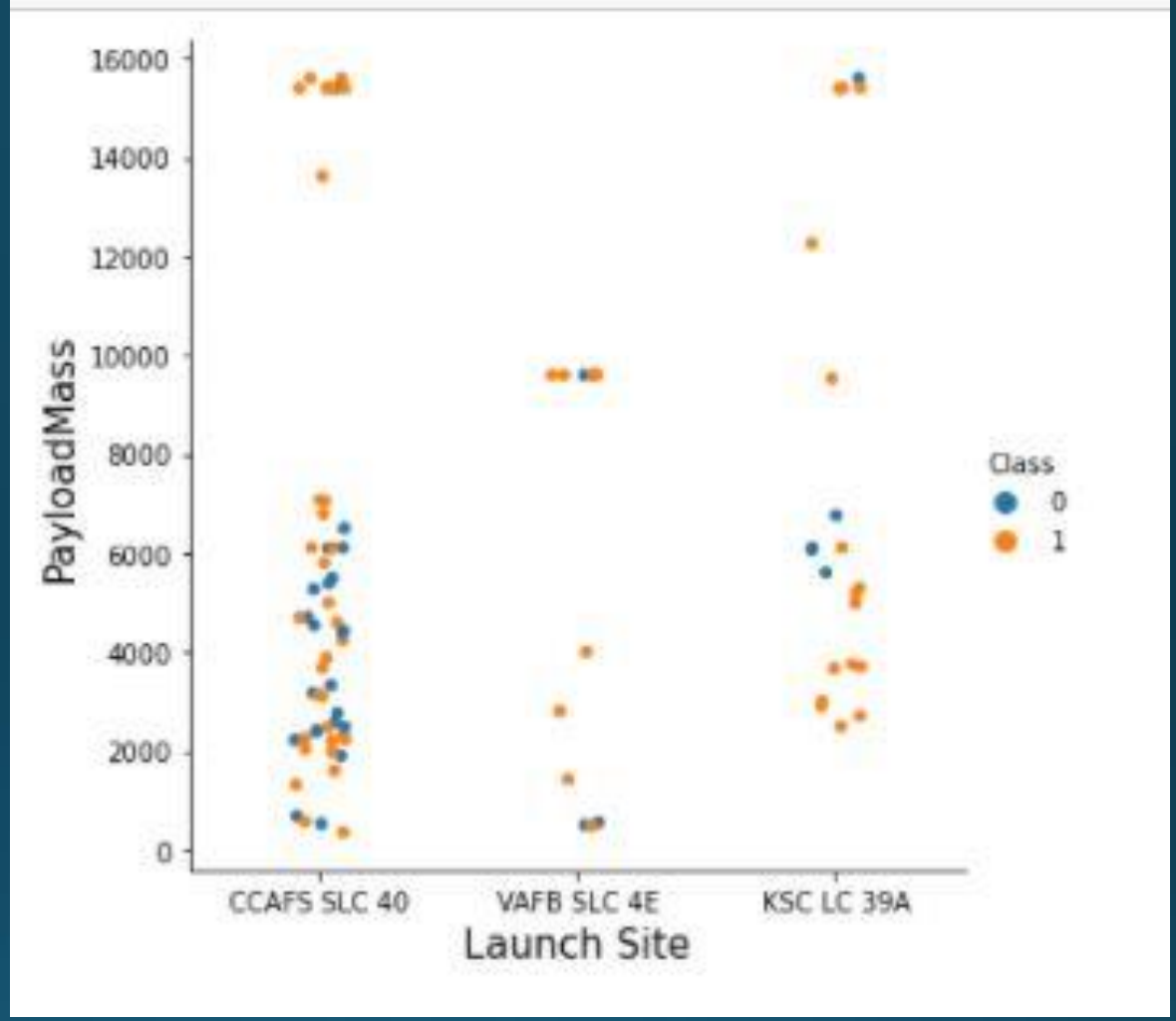


Flight Number vs. Launch Site



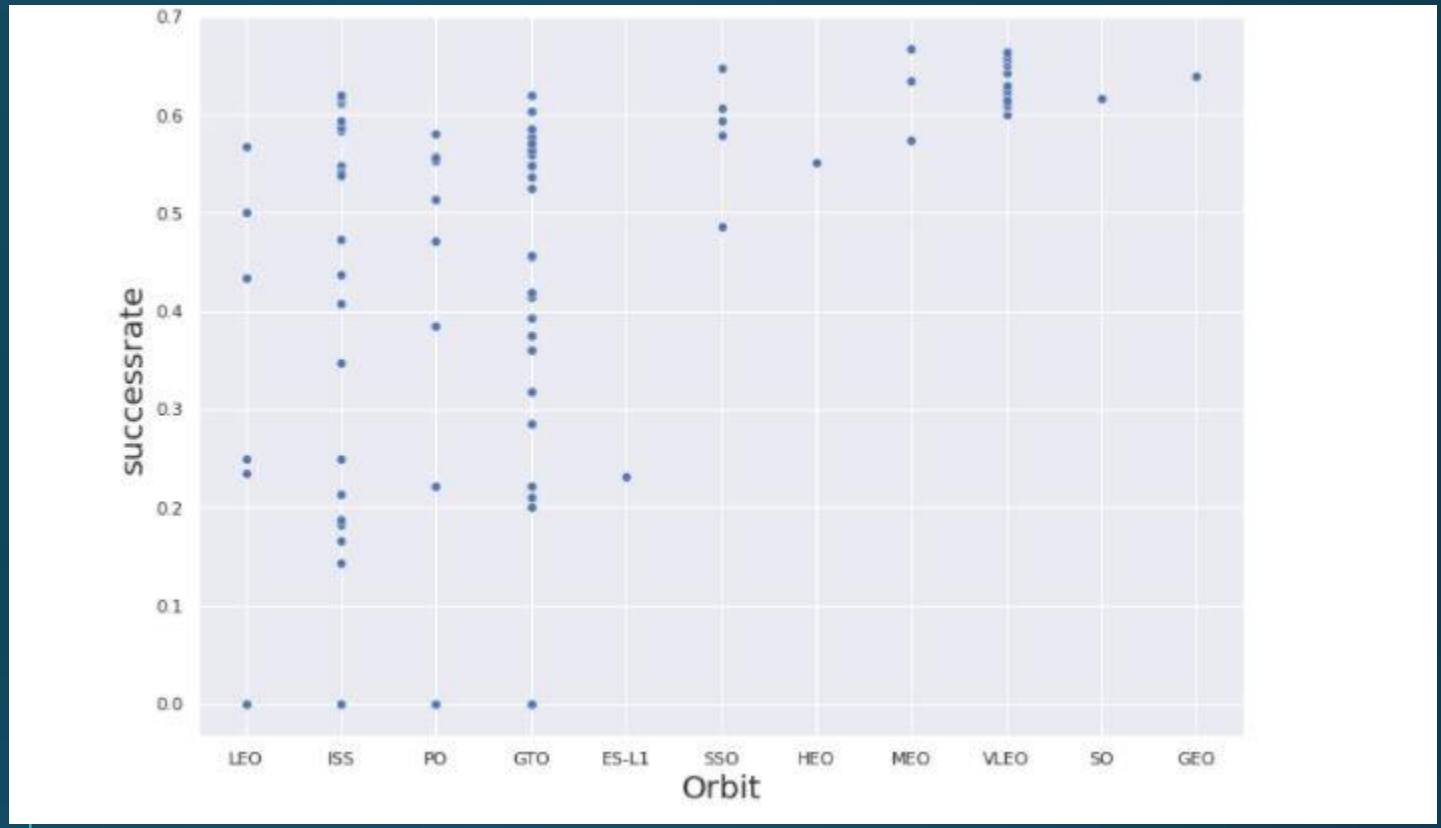
- The o represent the first stage failed to land and the one is success to land. The more flights in a site the more successful landing

Payload vs. Launch Site



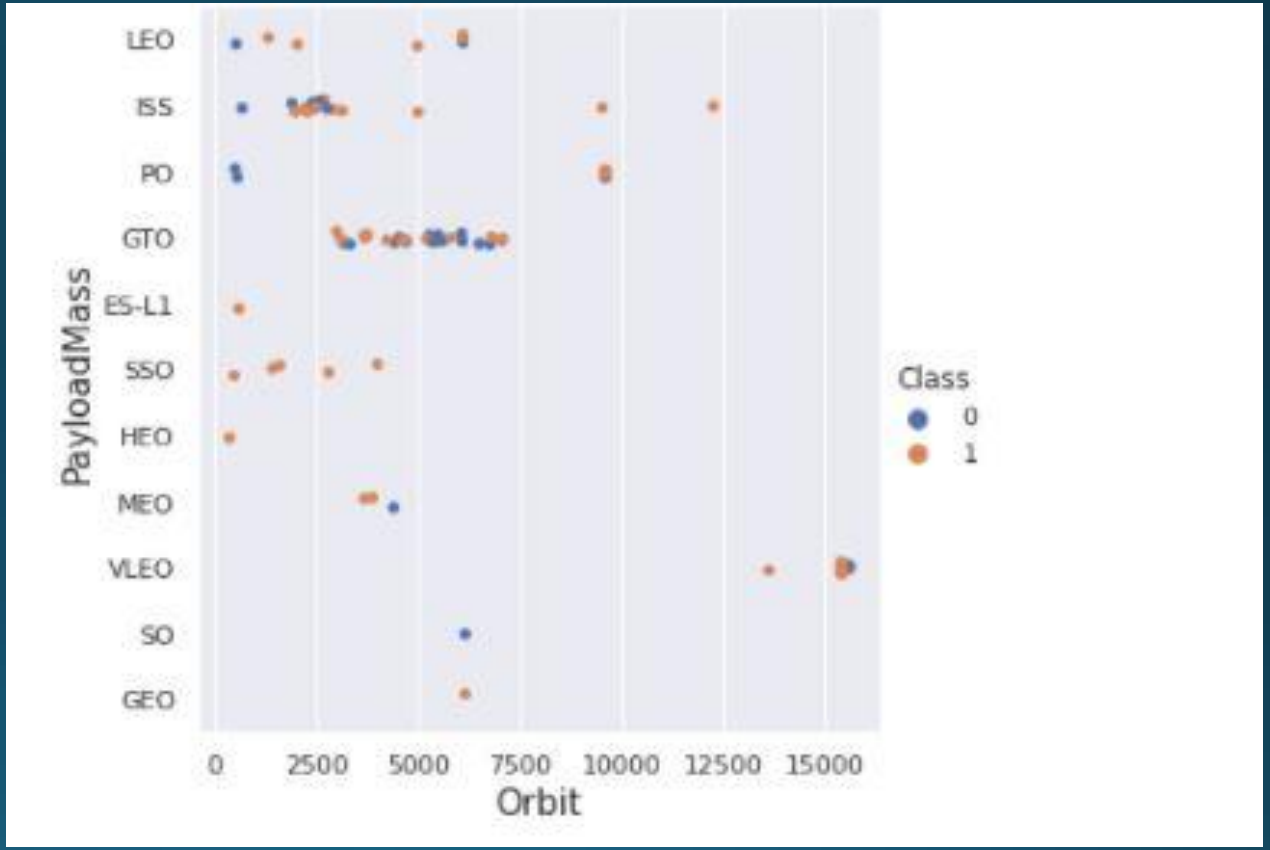
The greater play load has property to success landing in all sites

Success Rate vs. Orbit Type

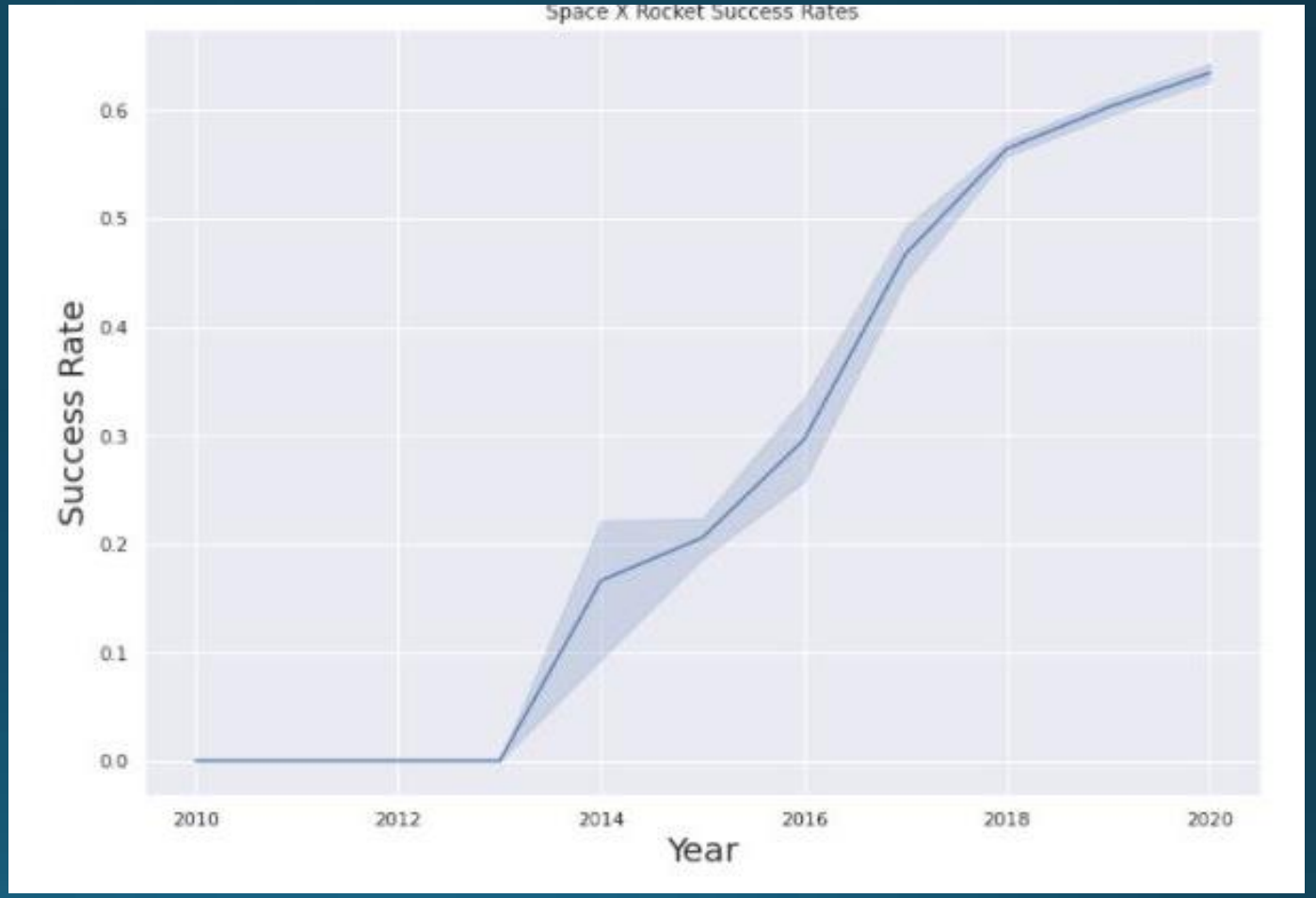


- GTO is the most type has success rate but the higher us VLEO

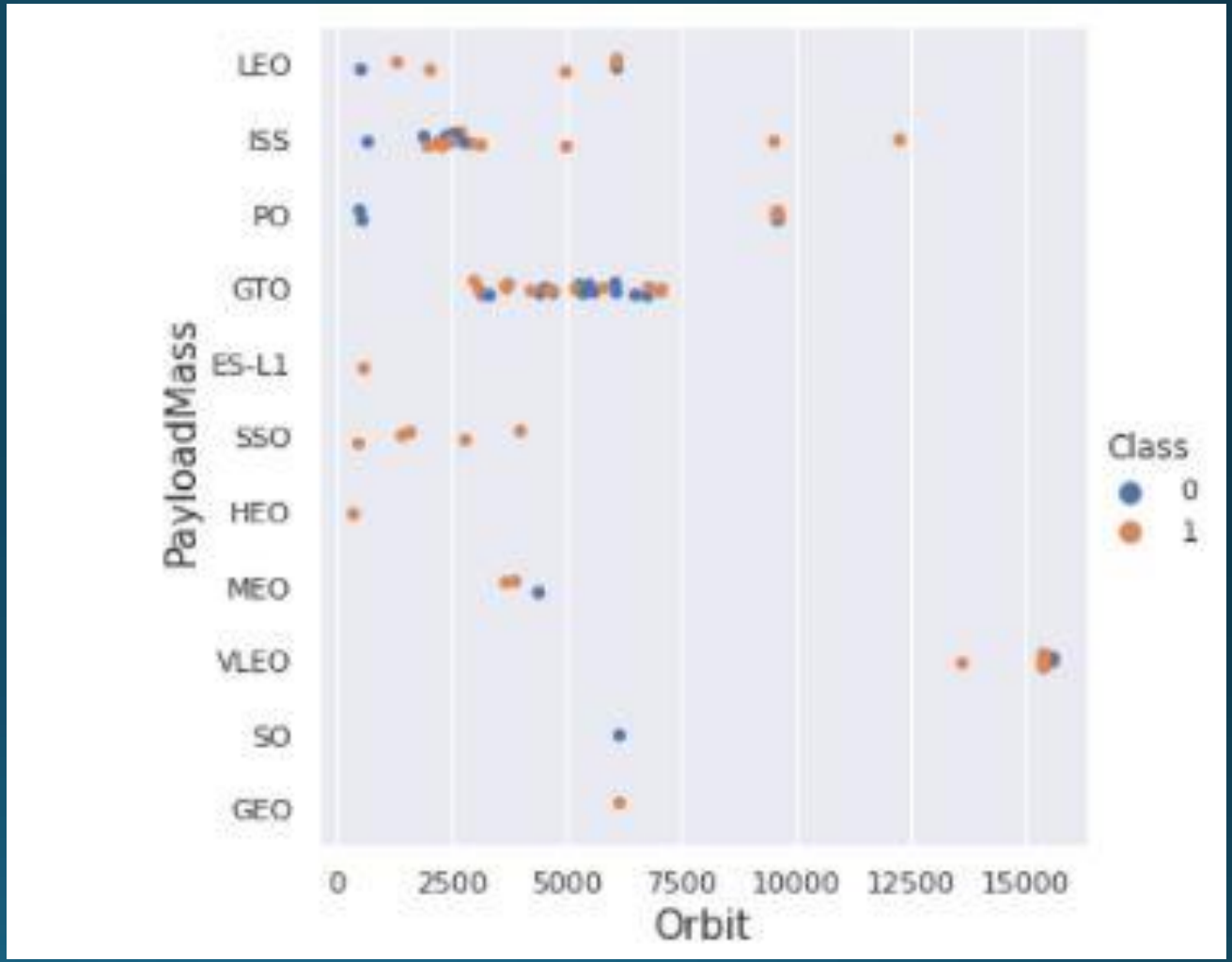
Flight Number vs. Orbit Type



Launch Success Yearly Trend



Payload vs. Orbit Type



All Launch Site Names

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

%sql select distinct(LAUNCH_SITE) from SPACEX2

To show launch sites without duplication I used distinct .

Launch Site Names Begin with 'CCA'

launch_site
CCAFS LC-40
CCAFS SLC-40

%sql select distinct(LAUNCH_SITE) from SPACEX2 where LAUNCH_SITE like 'CCA%' limit 5

To find the launch sites names begin with ‘CCA’ I use ‘like’ them limit will give first 5 results. But here only two lunch sites found begin with ‘CCA’

Total Payload Mass

45596

%sql select sum(PAYLOAD_MASS__KG_) from SPACEX2 where CUSTOMER = 'NASA (CRS)'

‘sum’ is used to calculate the total payload for the customer ‘NASA(CRS)’

Average Payload Mass by F9 v1.1

2928

%sql select AVG(PAYLOAD_MASS__KG_) from SPACEX2 where BOOSTER_VERSION = 'F9 v1.1'

'AVG' is used to find the average of the payload for version 'F9 v1.1'

First Successful Ground Landing Date

2015-12-22

%sql select min(DATE) from SPACEX2 where LANDING__OUTCOME = 'Success (ground pad)'

‘min’ is used to find the smallest date that had success landing

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select BOOSTER_VERSION from SPACEX2 where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

'between' is to defined the rage from 4000 to 6000 of payload mass

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

Total Number of Successful and Failure Miss ion Outcomes

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

%sql select MISSION_OUTCOME, count(BOOSTER_VERSION) from SPACEX2 group By MISSION_OUTCOME

'group by' dived records to categories bass on the column chosen 'count' is used to calculate total of records for each category

Boosters Carried Maximum Payload

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

```
%sql select booster_version from spaceX2
where payload_mass__kg_ =(select
max(payload_mass__kg_) from spacex2)
```

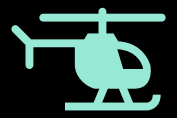
The second query is nested query that find the maximum of all records then we compare it with the first query to find the versions that had mass equal to the maximum

2015 Launch Records

```
%sql select landing__outcome ,count(*) total from spacex2 where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by 2 desc;
```

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

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

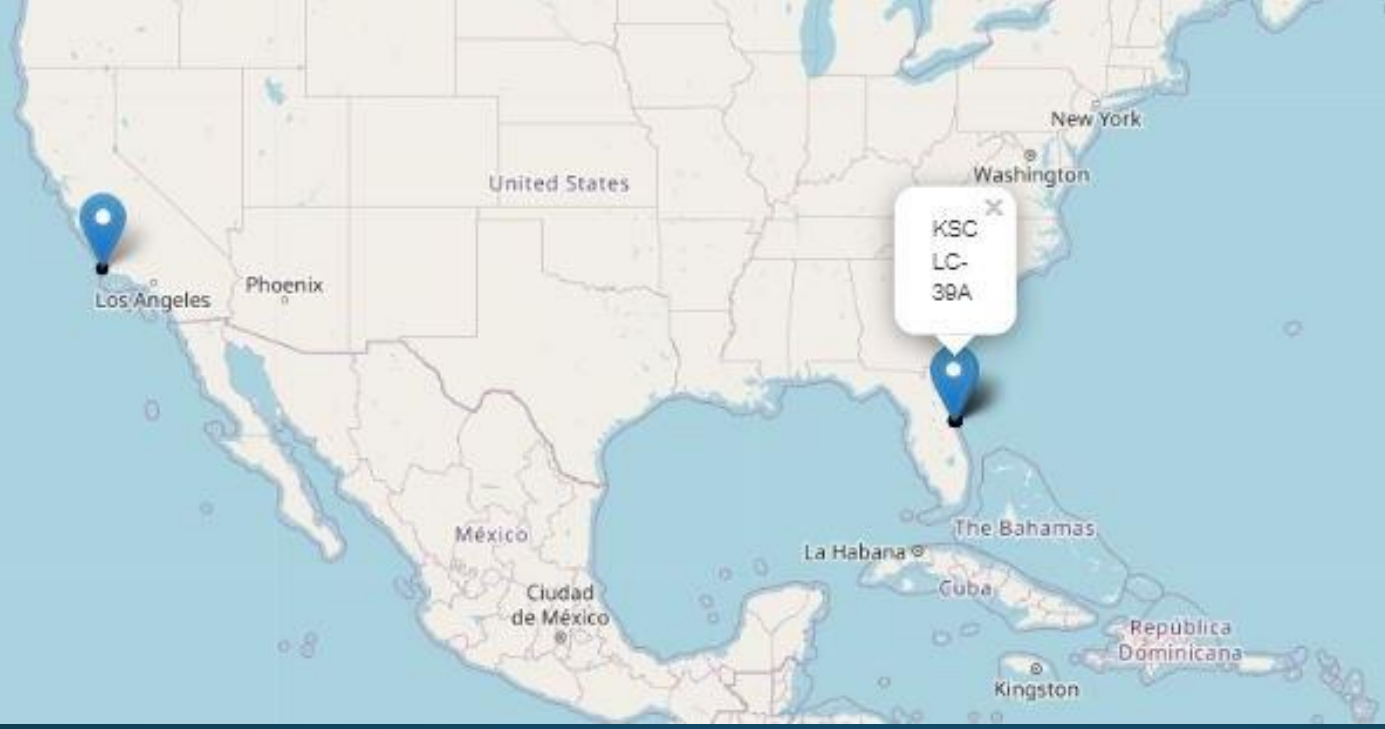


```
%sql select landing__outcome ,count(*) total from spacex2 where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by 2 desc;
```

The number of each outcome between 2010-06-04 and 2017-03-20 ordered from more to less

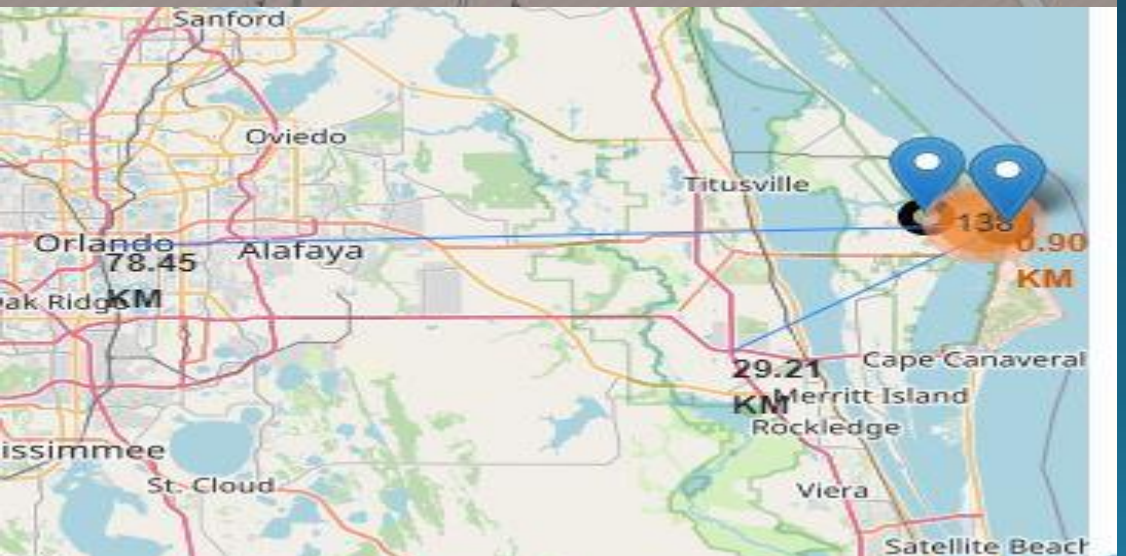
Section 4

Launch Sites Proximities Analysis



Show all the lunch sites

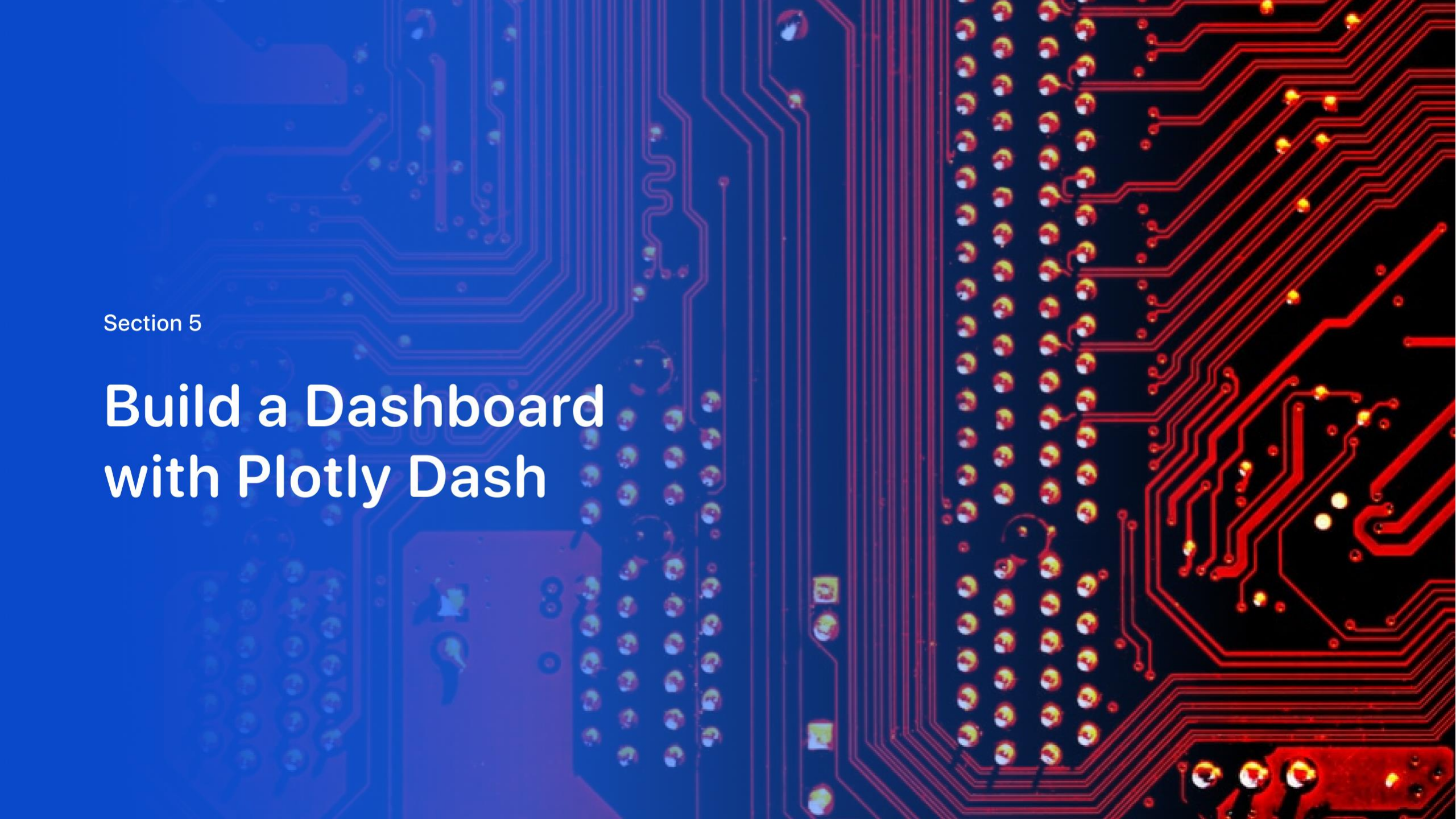
Lunch sites



- show the color-labeled launch outcomes on the map
- Show distance

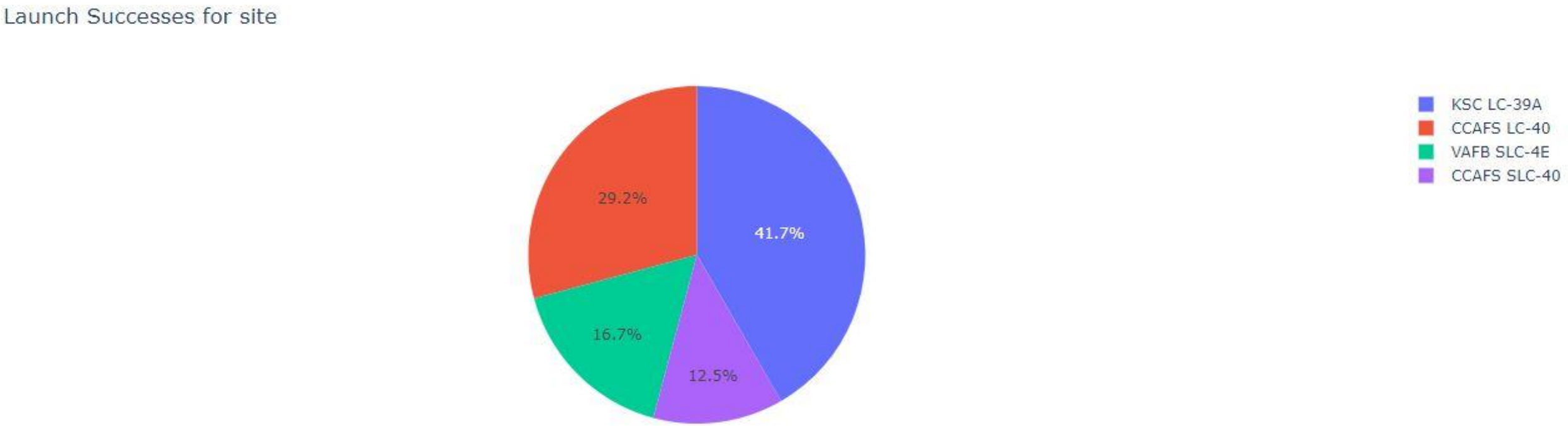
Section 5

Build a Dashboard with Plotly Dash



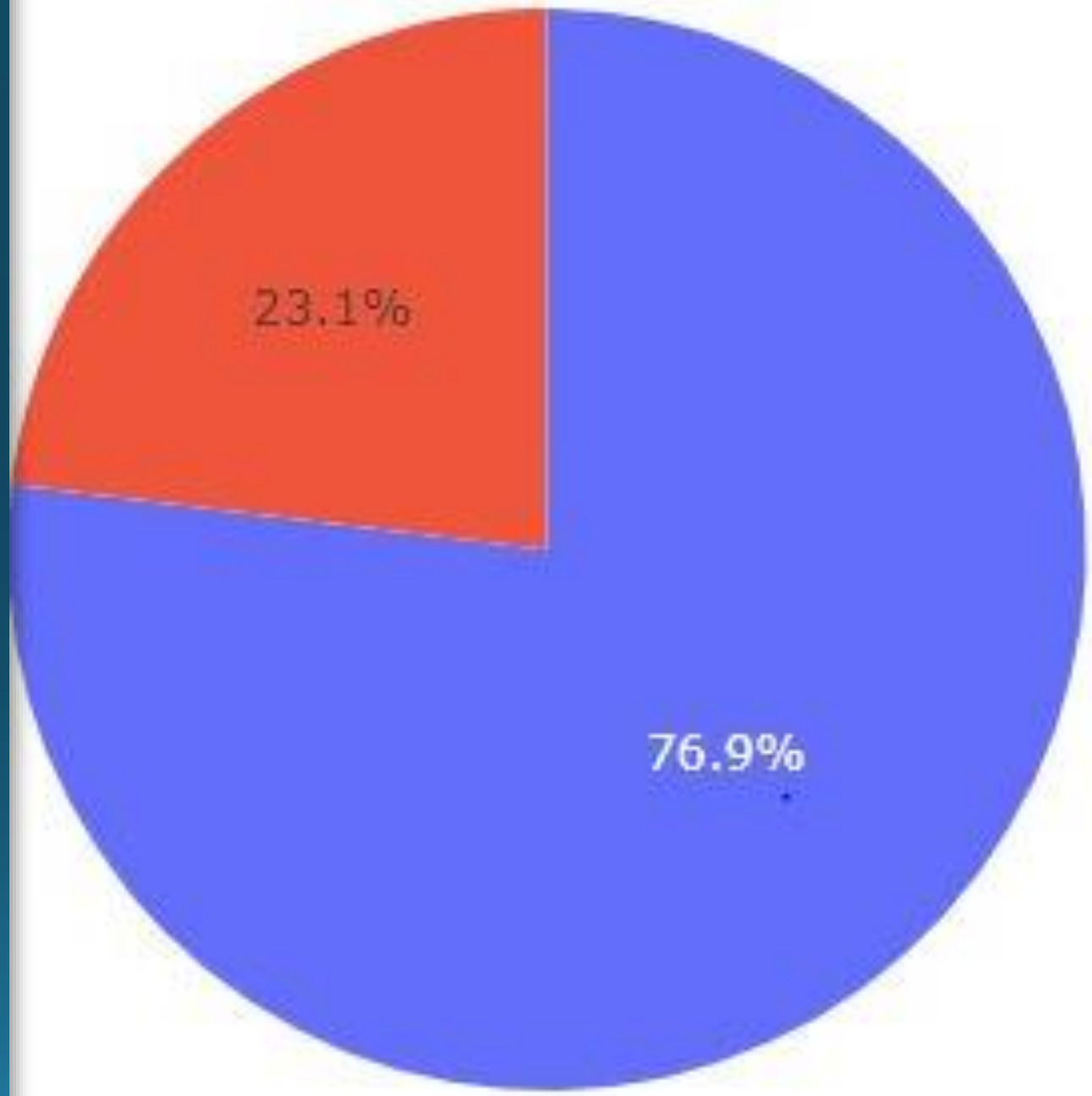
launch success count for all sites - Dashboard

- We can see the name of launch sites on the right as a key graph.
- As shown in the graph the most success site is KSC LC-39A



DASHBOARD – Pie chart for the launch site with highest launch success ratio

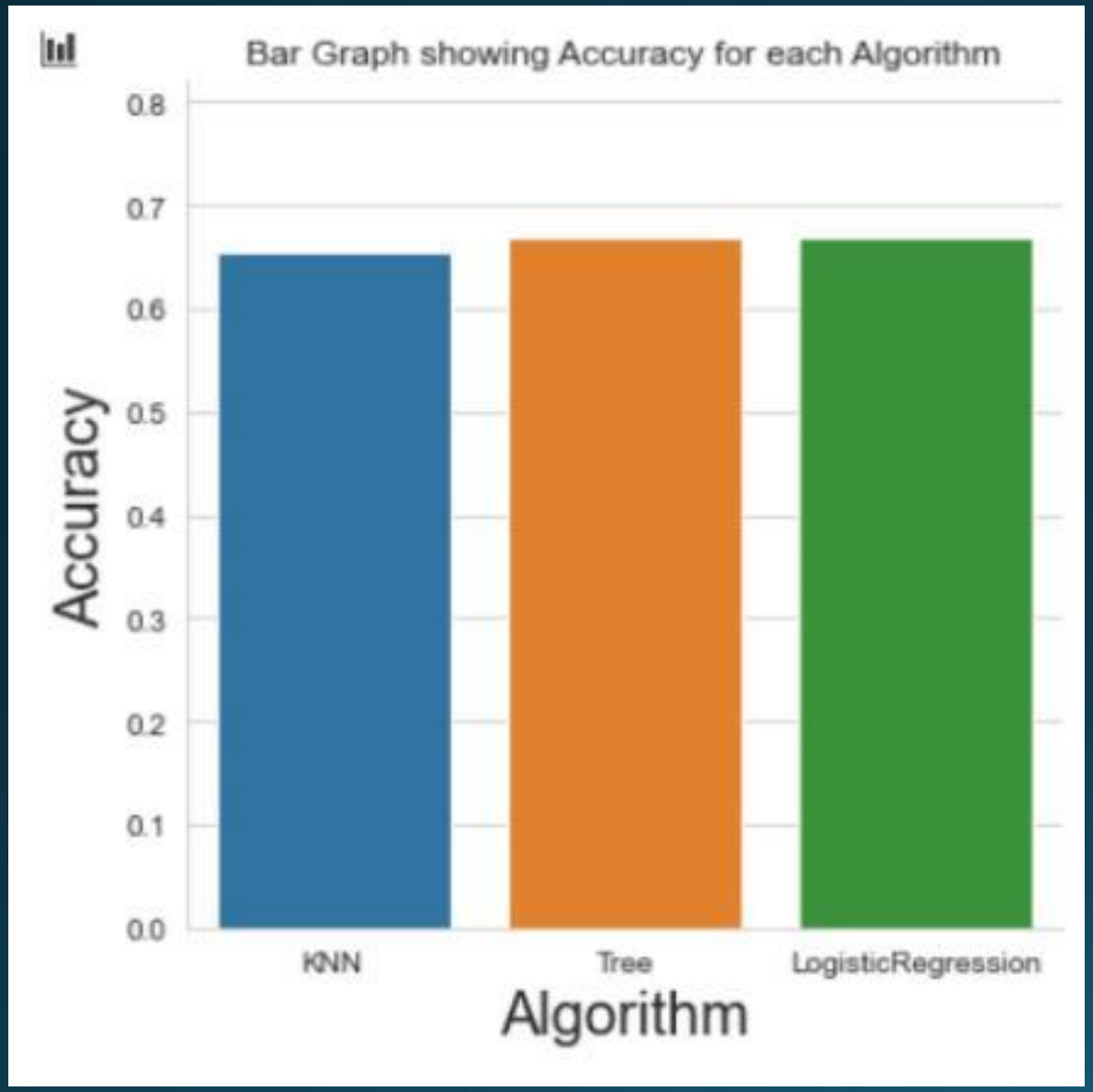
- KSC LC-39A has a success ratio of 76.9%



Payload vs. Launch Outcome scatter plot for all sites

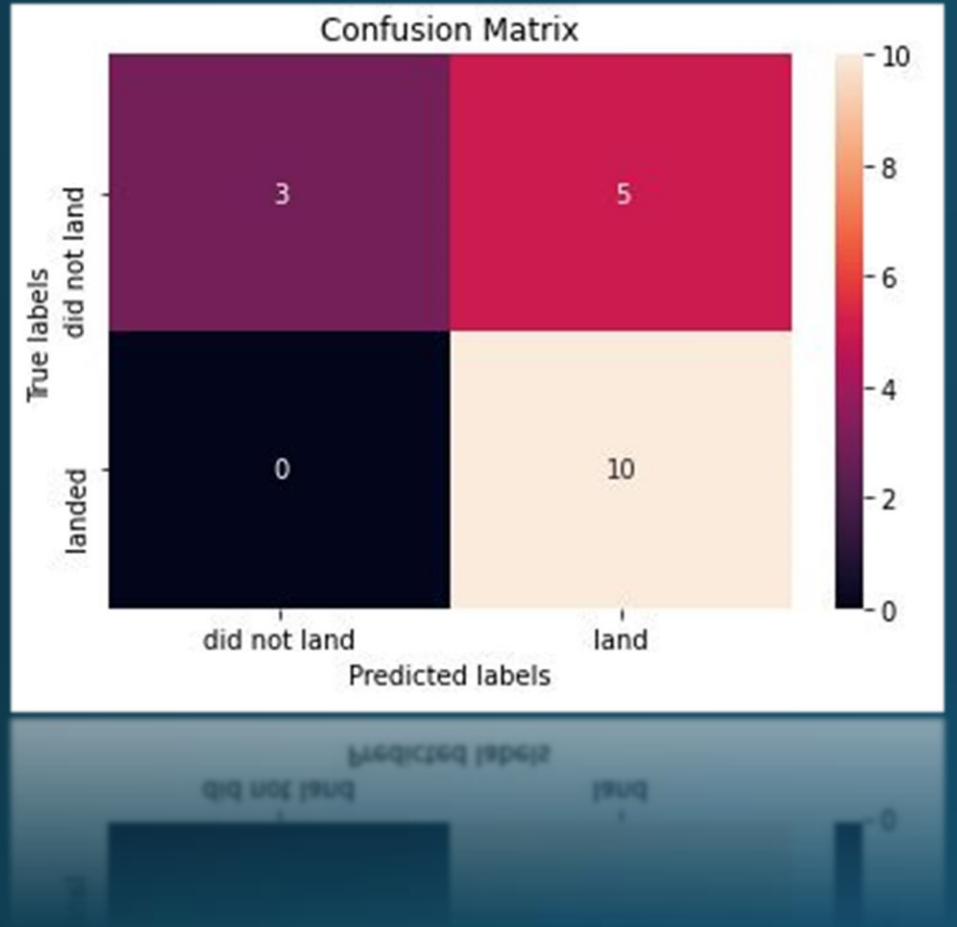
Section 6

Predictive Analysis (Classification)



Classification Accuracy

- Tree Algorithm is the best accuracy



Examining the confusion Matrix, We can see that the Decision Tree Algorithm can distinguish between the different classes

Confusion Matrix

Conclusions

- The Tree Classifier Algorithm is the best for this case.
- Low weighted payloads is better.
- The KSC LC-39A had the most successful launches from all the sites.
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Thank you!