Winning Space Race with Data Science

Harmanvir Singh
27.11.2021

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data wrangling
  - EDA with data visualization
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis
- Summary of all results
  - Exploratory data analysis results
  - Classification Results

Introduction

- Project background and context
  - We predicted if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, 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.
- Problems you want to find answers
  - Data Collection
  - Data wrangling
  - Visualization
  - Classification

Section 1

Methodology

Methodology

Executive Summary

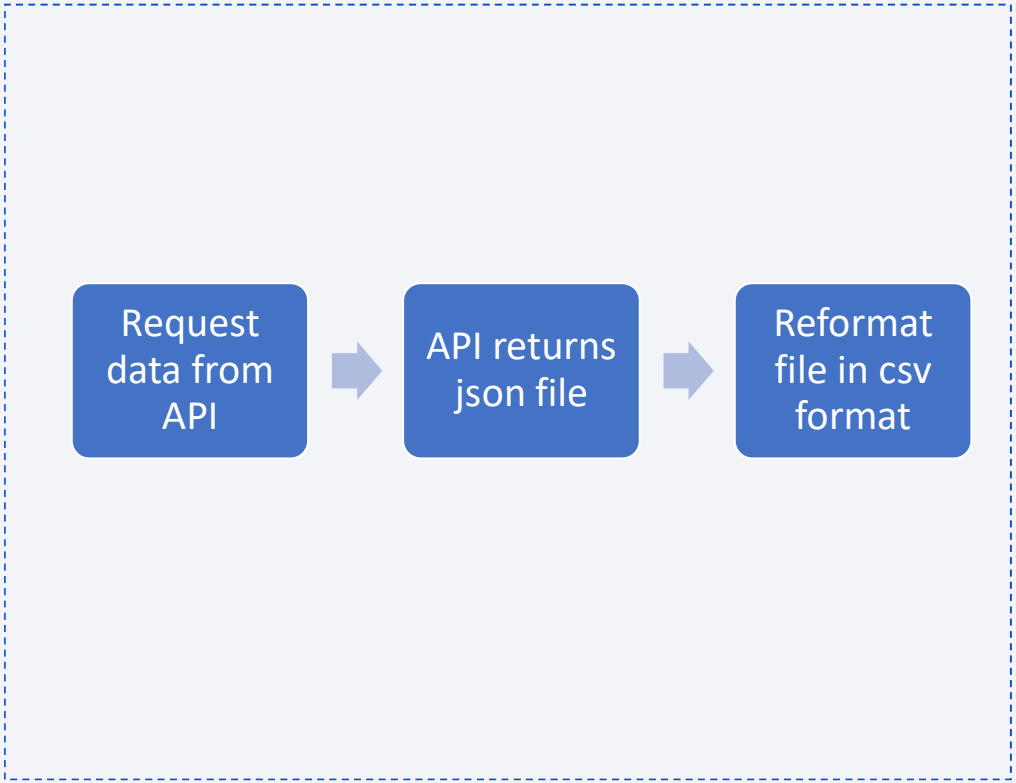
- Data collection methodology:
  - SpaceX API
  - Web Scrapping
- Perform data wrangling
  - One Hot Encoding of relevant data
  - Dropping irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Following data was collected
  - SpaceX launch data via SpaceX REST API
    - Data about launches (date, rocket used, payload delivered, launch specifications, landing information, and landing outcome)
  - Falcon 9 Launch Data of Wikipedia
    - Data about the rocket specifications
- SpaceX API and Web Scrapping Flowcharts

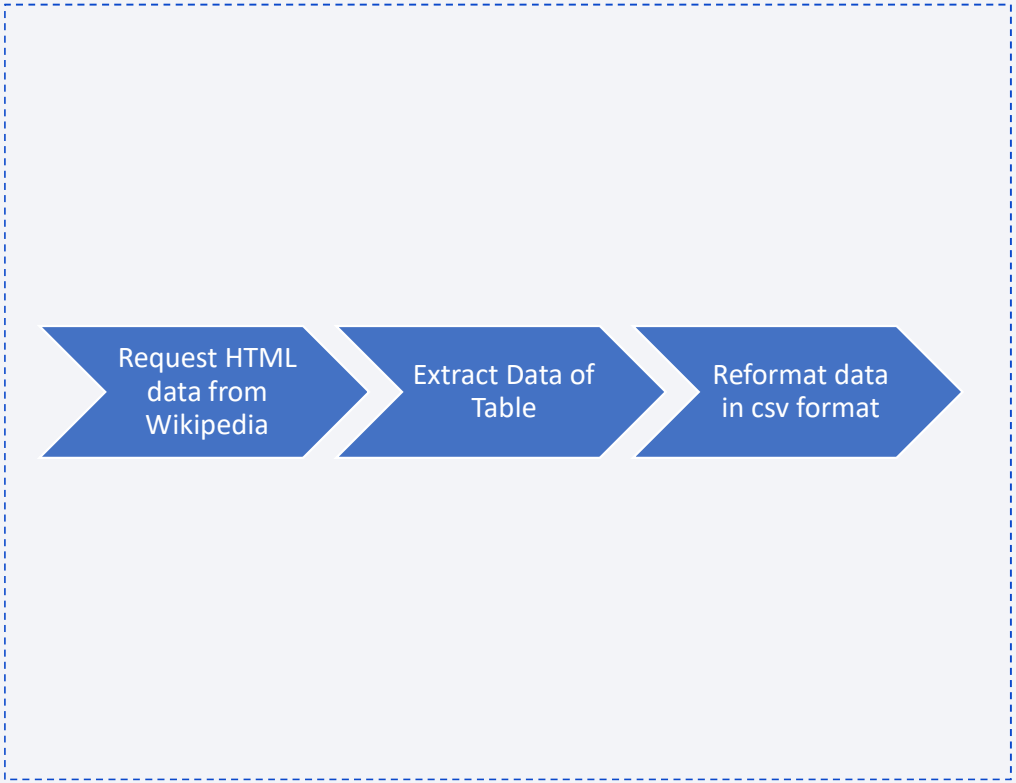
Data Collection – SpaceX API

- Request Data from API
  - response = request.get(url).json()
- Load Data in Panda DataFrame
  - pd.json_normalize(response)
- Clean Data with Custom functions
  - getLaunchSite(data), getPayloadData(data), getCoreData(data), getBoosterVersion(data)
- Assign to dictionary then dataframe
- Filter dataframe and export to flat file
- GitHub Link



Data Collection - Scraping

- Request HTML data from Wikipedia
  - Page = request.get(url)
- Create BeautifulSoup object
  - Soup = BeautifulSoup(Page.text, 'html.parser')
- Get column names
- Create dictionary
- Append data to dictionary
- Contvert to dataframe and save
- [GitHub Link]



Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- Process of EDA
  - Calculate Number of launches at each site
  - Calculate the number and occurrence of each orbit
- GitHub Link

EDA with Data Visualization

- Scatter Plots for identifying correlations
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload vs Launch Site
  - Orbit vs Flight Number
  - Payload vs Orbit Type
  - Orbit vs Payload Mass
- Bar Graph for comparisons
  - Mean vs Orbit
- Line Graph for temporal dependencies
  - Success Rate vs Year
- GitHub Link

EDA with SQL

- Following SQL Queries were done in EDA:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'KSC’ • Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date where the successful landing outcome in drone ship was achieved.
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster_versions which have carried the maximum payload mass.
  - Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
  - Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

- GitHub Link

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- Using Haversine's formula we calculated the distance 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
- Example of some trends in which the Launch Site is situated in.
  - Are launch sites in close proximity to railways? No
  - Are launch sites in close proximity to highways? No
  - Are launch sites in close proximity to coastline? Yes
  - Do launch sites keep certain distance away from cities? Yes
- GitHub Link

Build a Dashboard with Plotly Dash

- Used Python Anywhere to host the website live 24/7 so your can play around with the data and view the data - The dashboard is built with Flask and Dash web framework.
- Graphs - Pie Chart showing the total launches by a certain site/all sites - display relative proportions of multiple classes of data. - size of the circle can be made proportional to the total quantity it represents.
- Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions
  - It shows the relationship between two variables.
  - It is the best method to show you a non-linear pattern.
  - The range of data flow, i.e. maximum and minimum value, can be determined.
  - Observation and reading are straightforward.
- GitHub Link

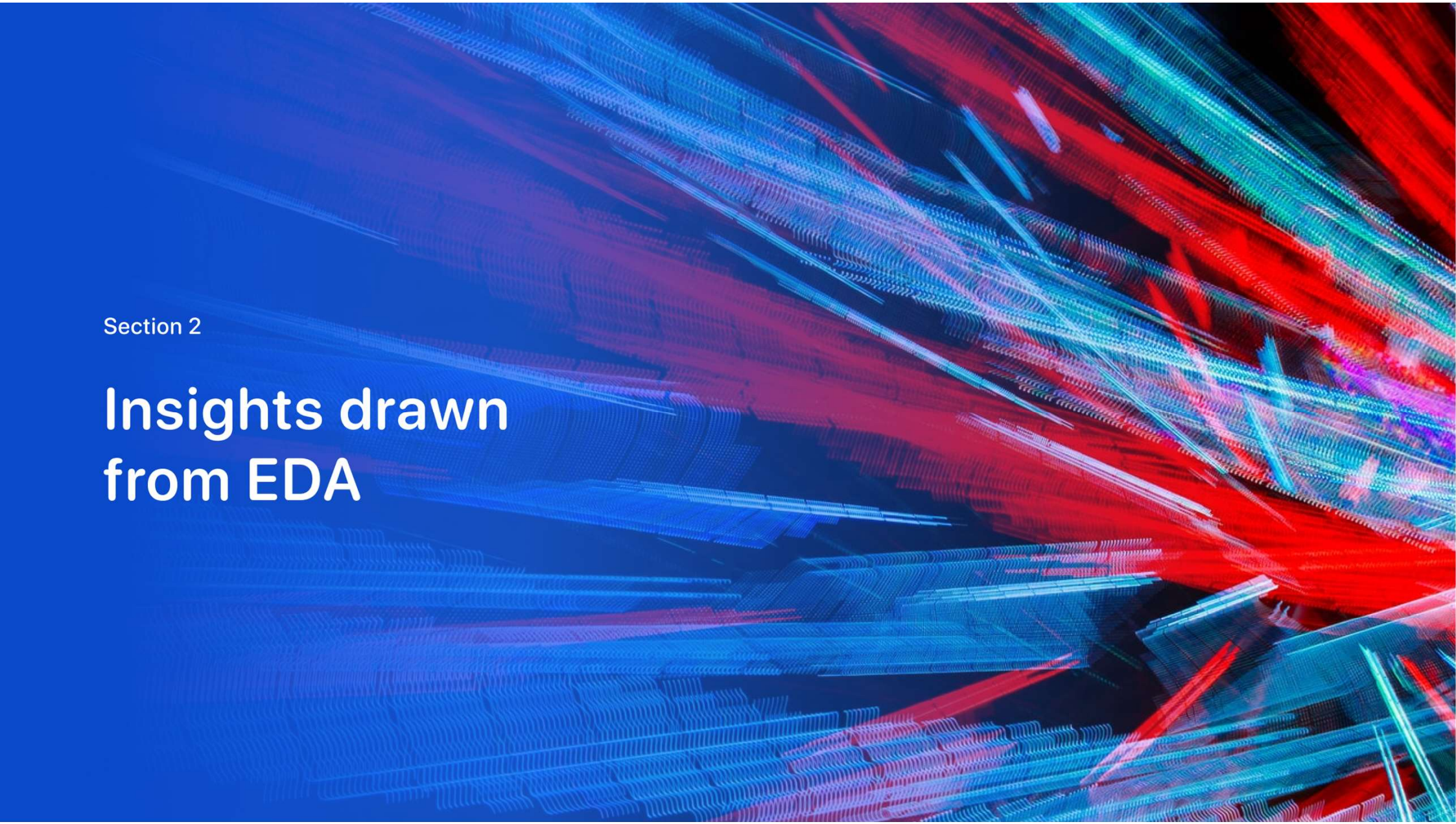
Predictive Analysis (Classification)

- BUILDING MODEL
  - Load our dataset into NumPy and Pandas
  - Transform Data
  - Split our data into training and test data sets
  - Check how many test samples we have
  - Decide which type of machine learning algorithms we want to use
  - Set our parameters and algorithms to GridSearchCV
  - Fit our datasets into the GridSearchCV objects and train our dataset.
- EVALUATING MODEL
  - Check accuracy for each model
  - Get tuned hyperparameters for each type of algorithms
  - Plot Confusion Matrix
- IMPROVING MODEL
  - Feature Engineering
  - Algorithm Tuning
- FINDING THE BEST PERFORMING CLASSIFICATION MODEL
  - The model with the best accuracy score wins the best performing model
  - In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

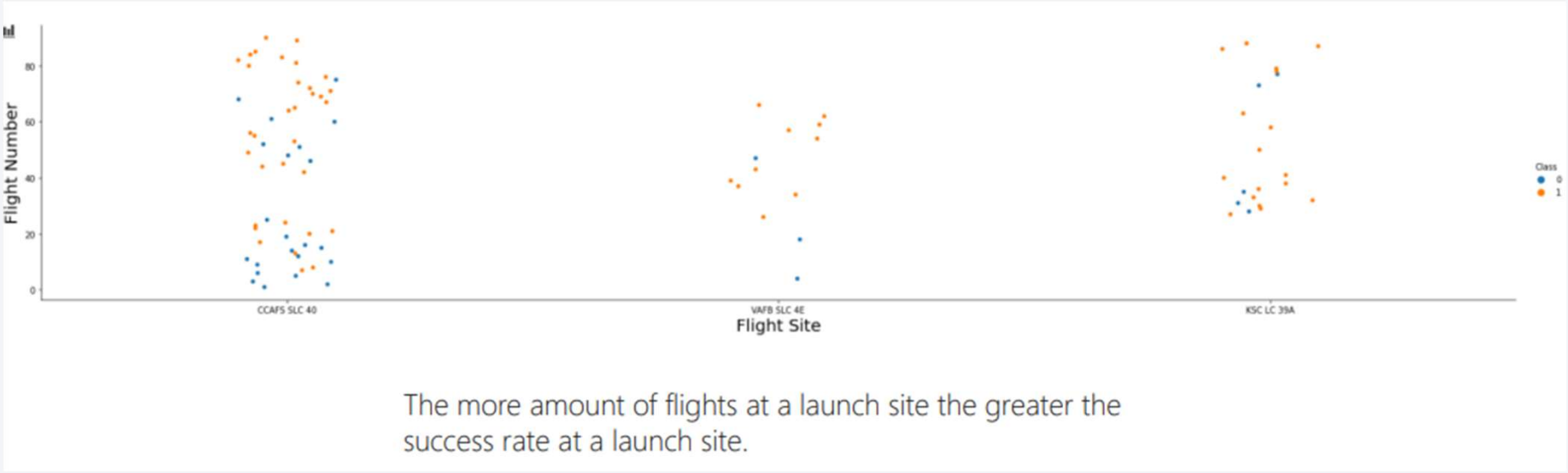
GitHub Link

Results

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

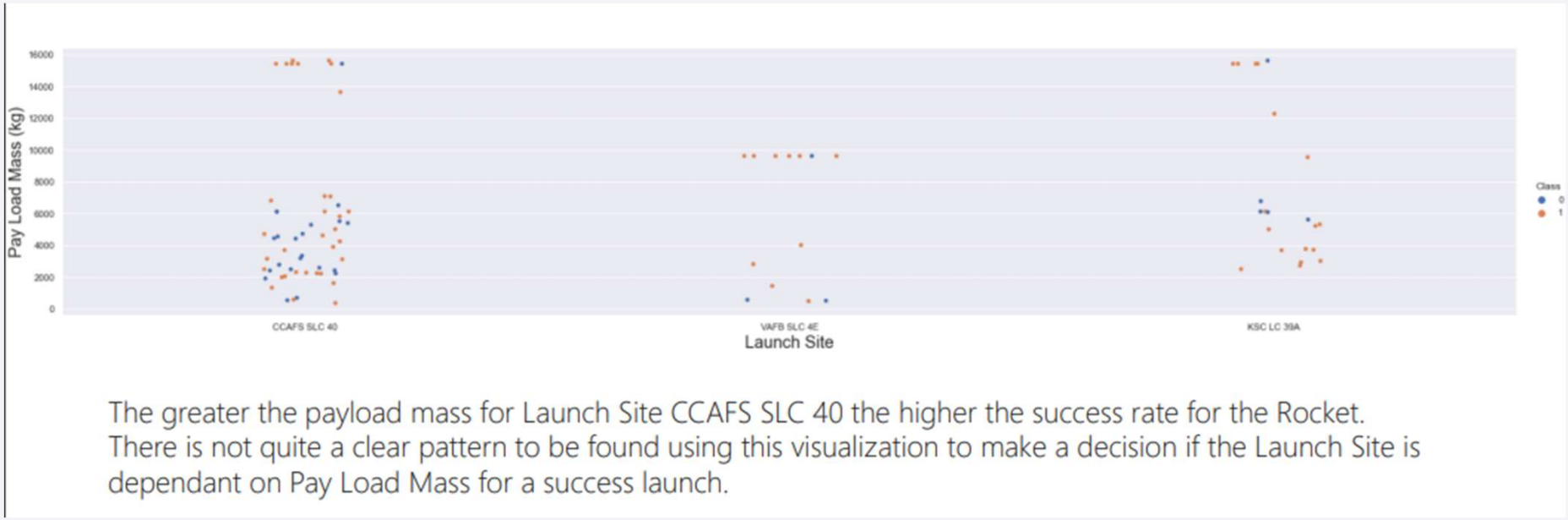

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



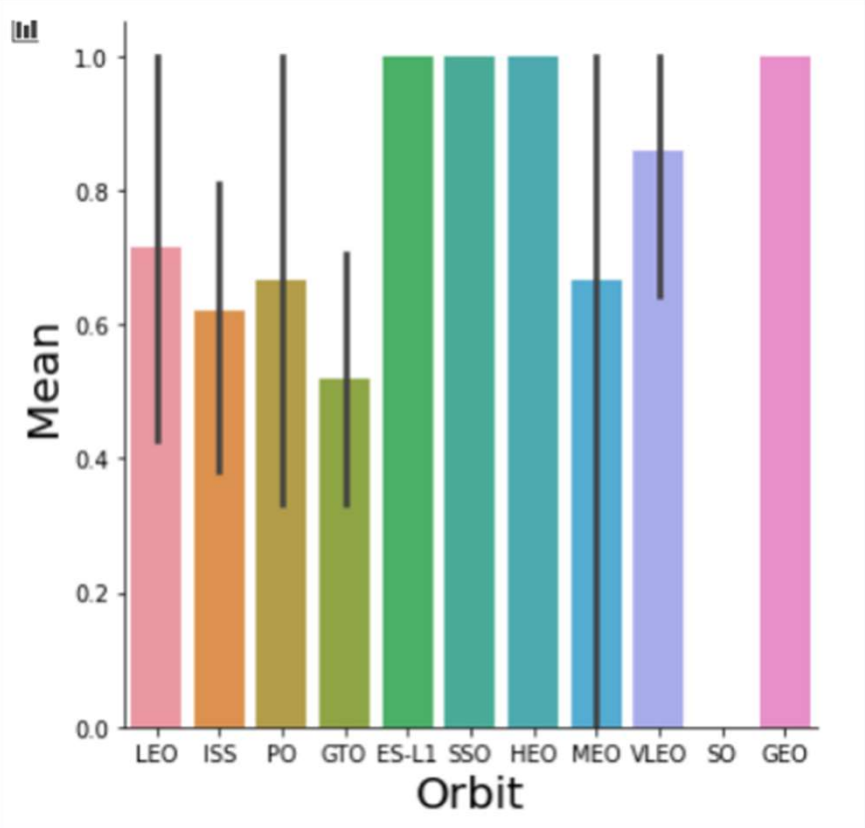
The more amount of flights at a launch site the greater the success rate at a launch site.

Payload vs. Launch Site



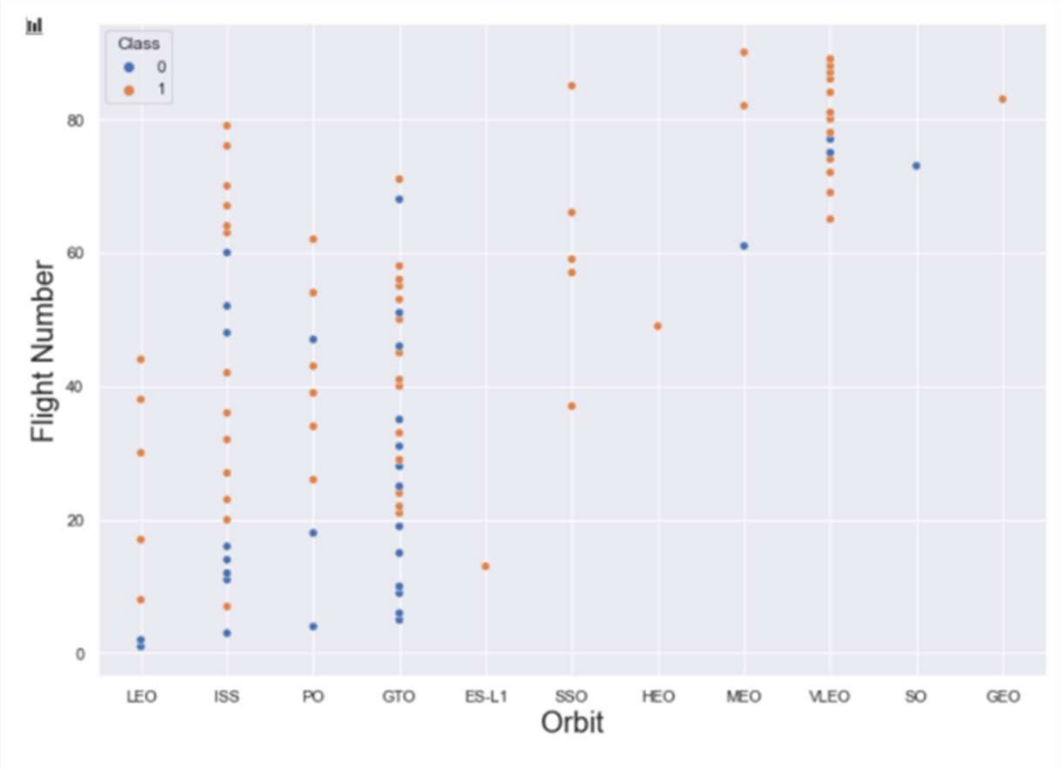
The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket.
There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependant on Pay Load Mass for a success launch.

Success Rate vs. Orbit Type



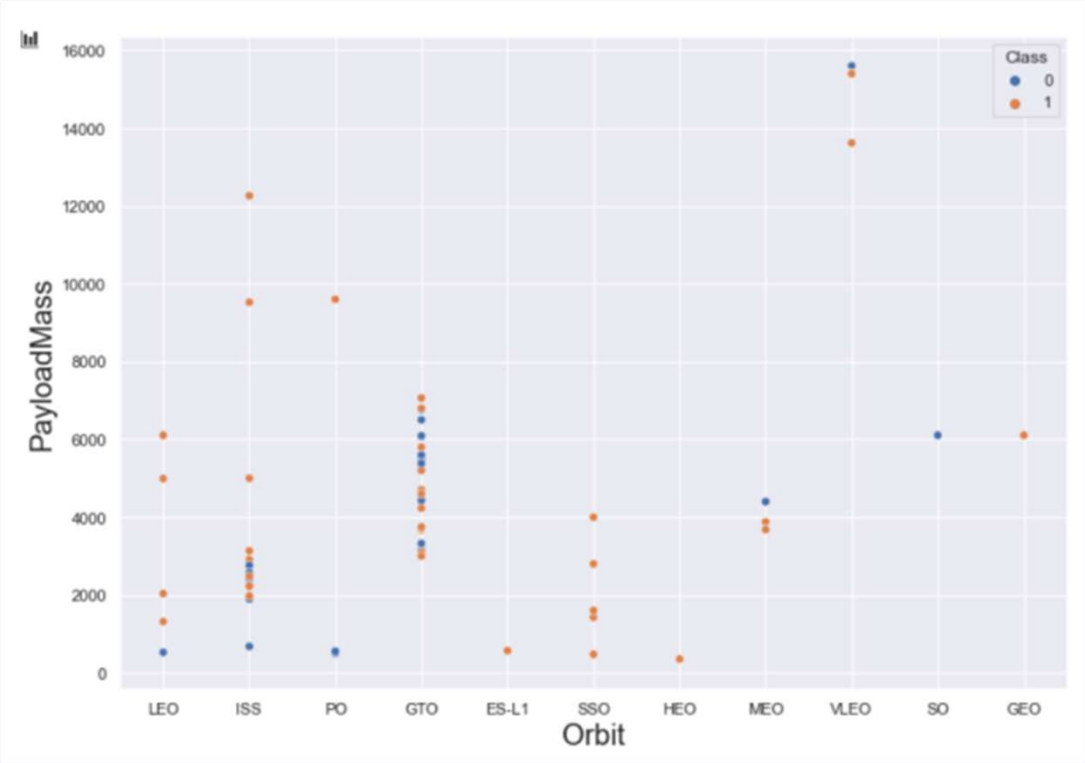
Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Flight Number vs. Orbit Type



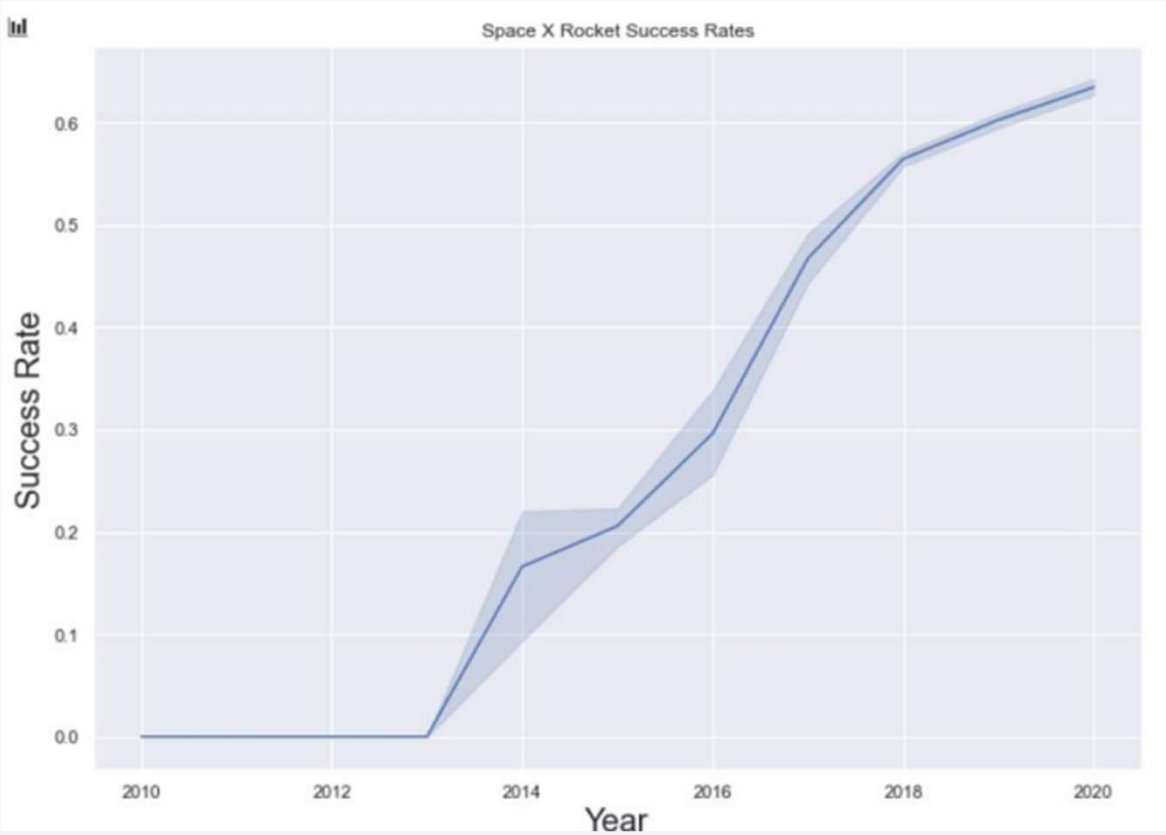
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



you can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

- SQL Query
  - SELECT DISTINCT Launch_Site from SaceX_DATA
- Explanation
  - DISTINCT → only shows unique values
- Result

Unique Launch Sites
CCAFS LC-40
CCAFS SLC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

select * from SPACEXDATASET where launch_site like 'CCA%' limit 5

Launch Site Names Begin with 'CCA'

- SQL Query
  - select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
- Explanation
  - where launch_site like 'CCA%‘ → only shows values where launch_site begins with ‘CCA’
  - Limit 5 → only shows the first 5 values
- Result

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

- SQL Query
  - select sum(payload_mass__kg_) from SPACEXDATASET where customer like 'NASA (CRS)'
- Explanation
  - sum(payload_mass__kg_) → calculates the sum of the payload mass
- Result

Average Payload Mass by F9 v1.1

- SQL Query
  - select avg(payload_mass__kg_) from SPACEXDATASET where booster_version like 'F9 v1.0%'
- Explanation
  - avg(payload_mass__kg_) → calculates the average of the payload mass
- Result

1
340

First Successful Ground Landing Date

- SQL Query
  - select min(date) from SPACEXDATASET where landing__outcome like 'Success (ground pad)'
- Explanation
  - Min(date) → selects the first date in date
- Result

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL Query
  - select booster_version from SPACEXDATASET where payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000 and landing__outcome like 'Success (drone ship)'
- Explanation
  - Where condition→ selects data which fulfills the condition
- Result

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

Total Number of Successful and Failure Mission Outcomes

- SQL Query
  - select count(*) from SPACEXDATASET where mission_outcome like 'Success%' or mission_outcome like 'Failure%'
- Explanation
  - Count(*) → counts all entries
- Result

Boosters Carried Maximum Payload

- SQL Query
  - select unique(booster_version) from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET)
- Explanation
  - Sub Query for condition
- Result

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

2015 Launch Records

- SQL Query
  - select DATE, booster_version, launch_site, landing__outcome from SPACEXDATASET where landing__outcome like 'Failure (drone ship)' and year(DATE)=2015
- Explanation
  - simple
- Result

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

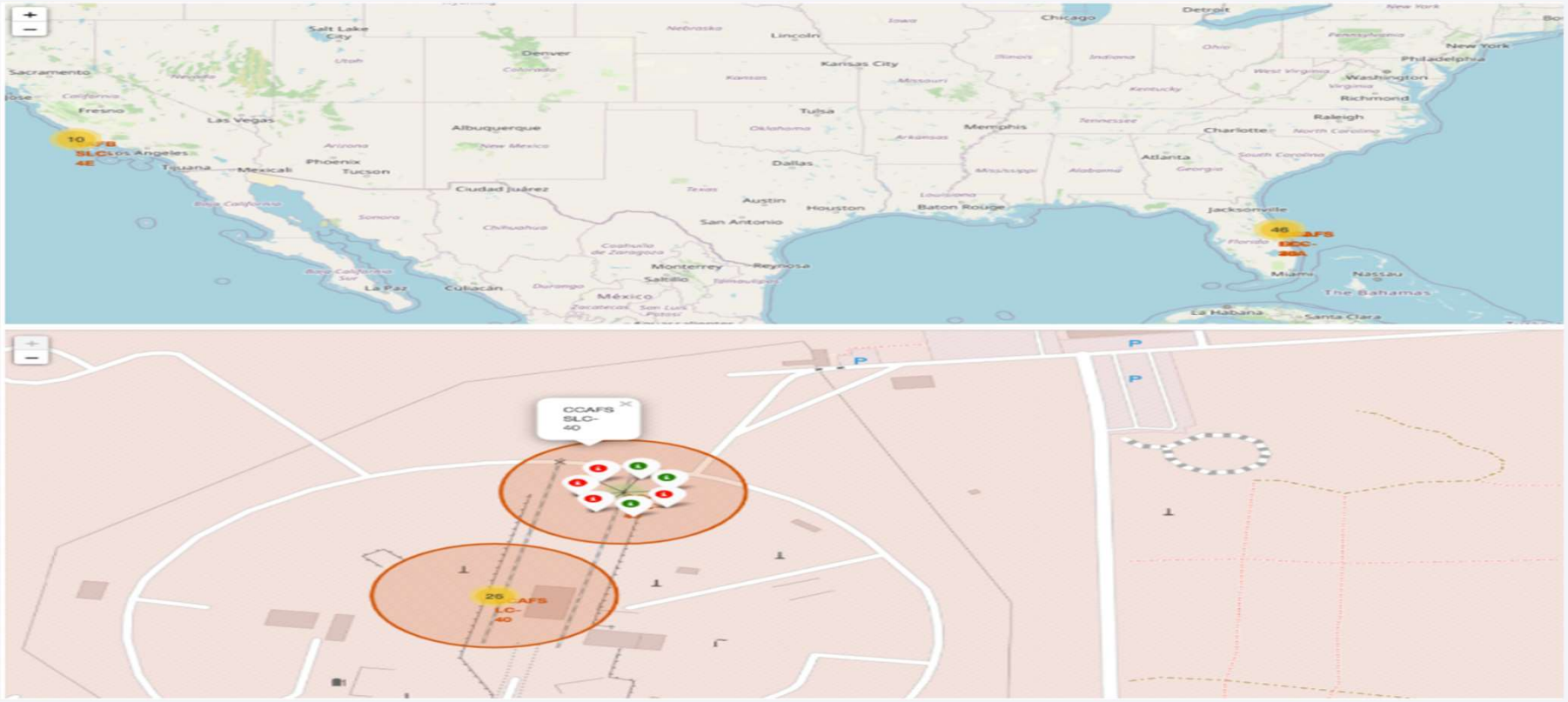
- SQL Query
  - select landing__outcome, count(*) as count from SPACEXDATASET where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome order by count desc
- Explanation
  - simple
- Result

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

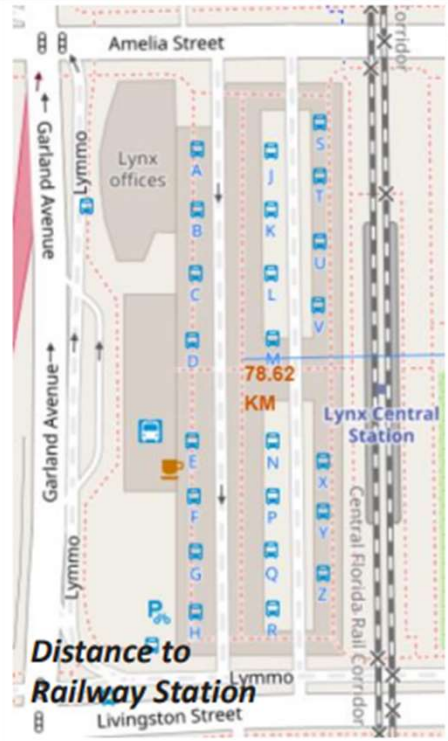

Section 4
Launch Sites
Proximities Analysis

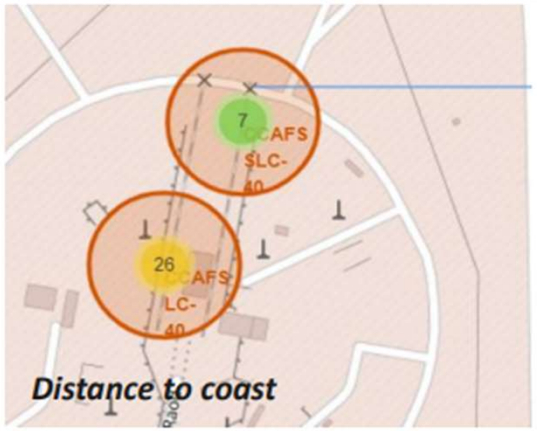
All launch sites global map

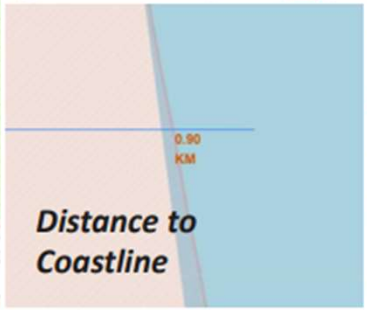
Launch locations



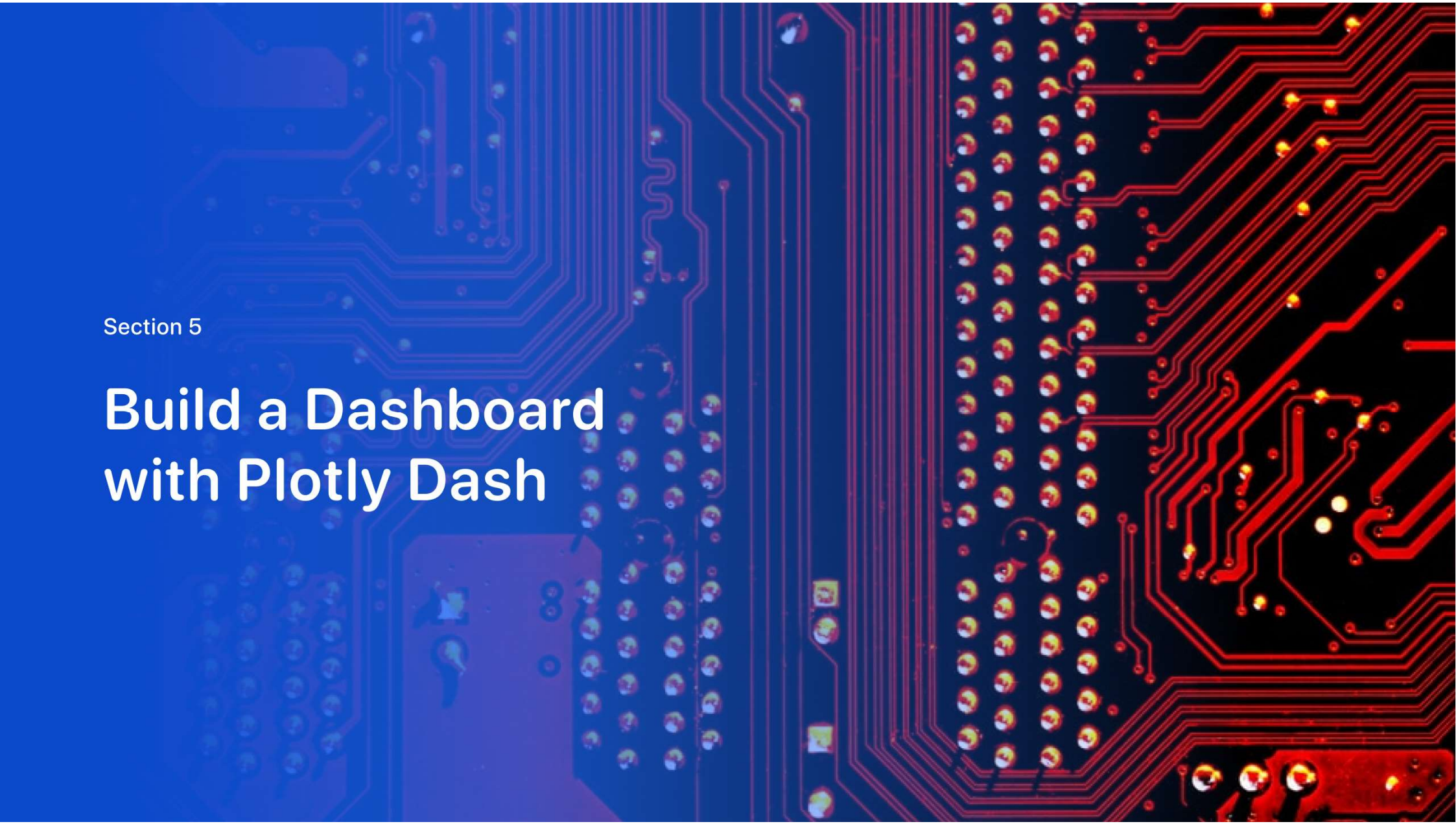
Working out Launch Sites distance to landmarks to find trends with Haversine formula using CCAFS-SLC-40 as a reference



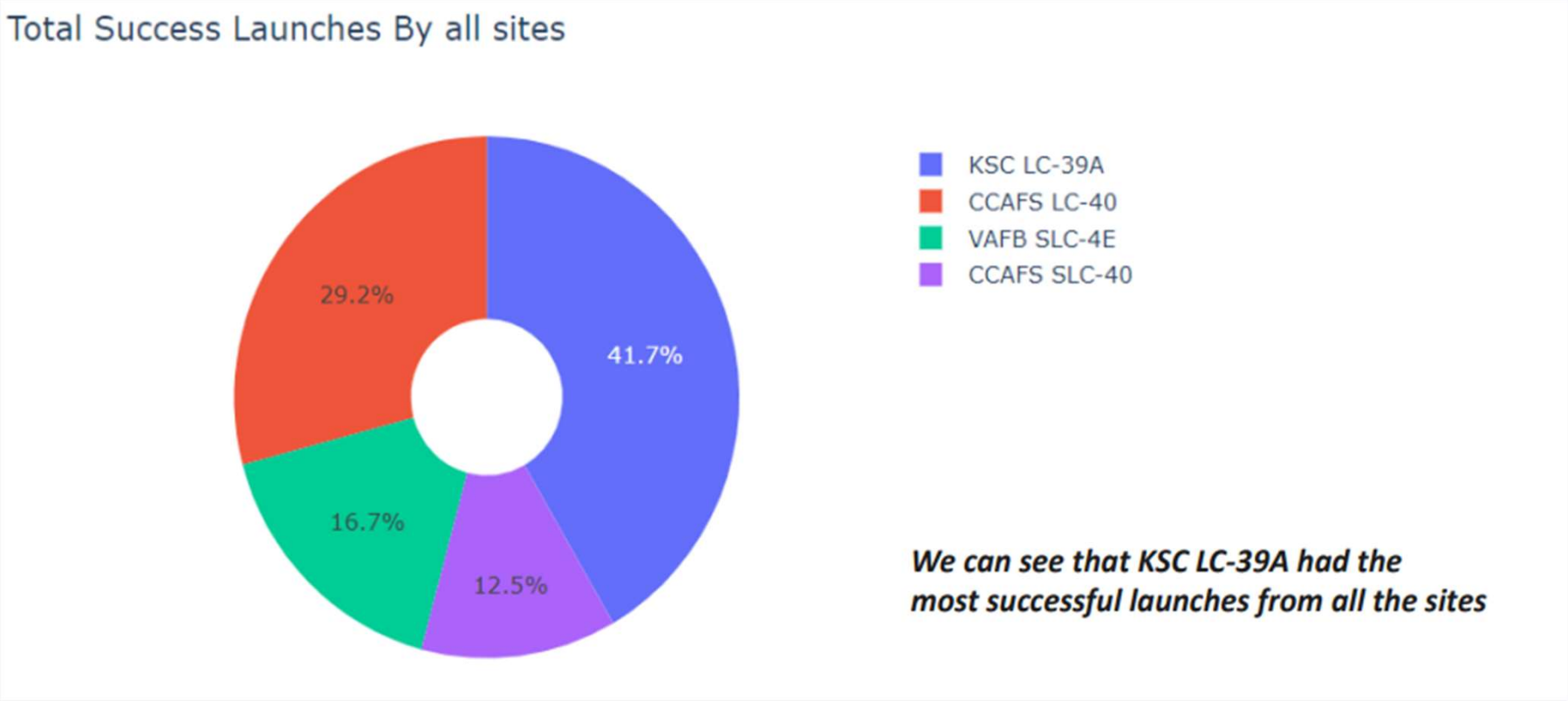




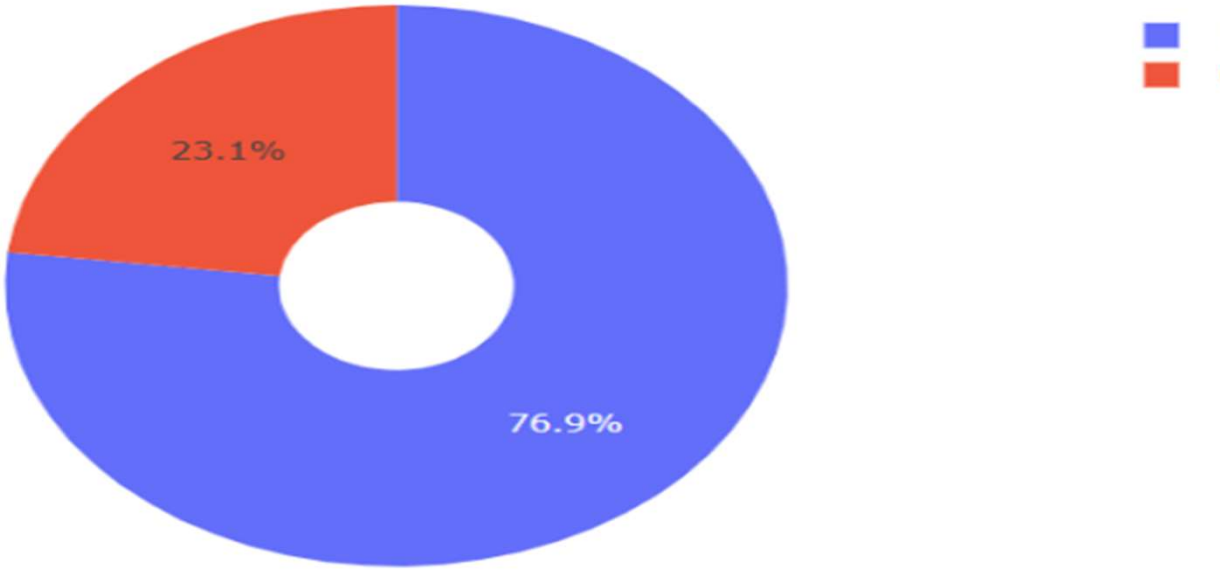
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 5
Build a Dashboard with Plotly Dash

shot 1> 40 DASHBOARD – Pie chart showing the success percentage achieved by each launch site

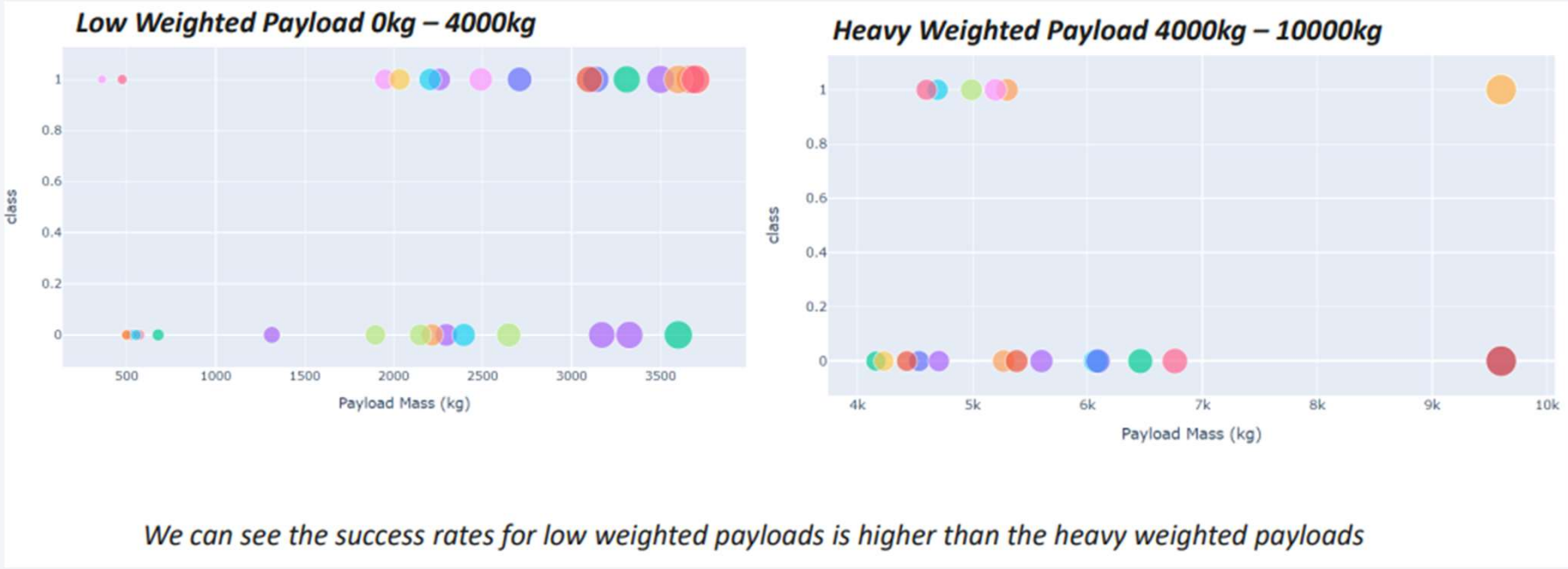


DASHBOARD – Pie chart for the launch site with highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

DASHBOARD – Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



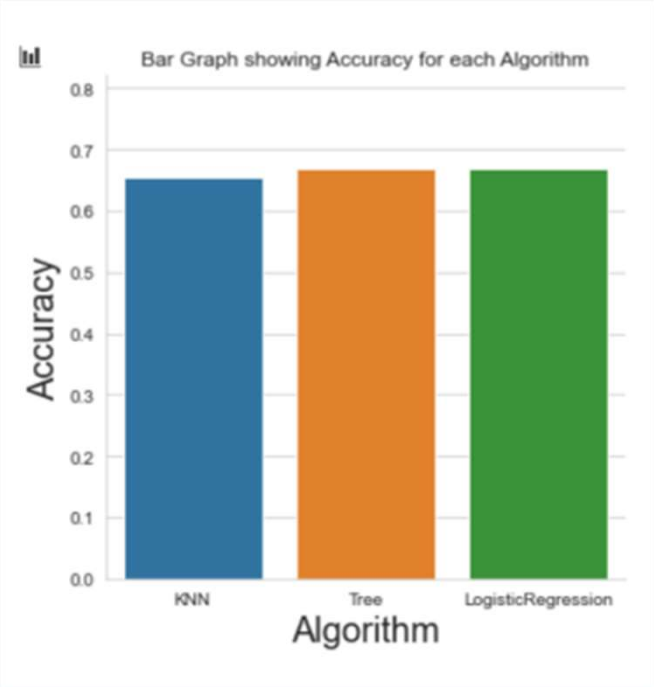
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 6

Predictive Analysis (Classification)

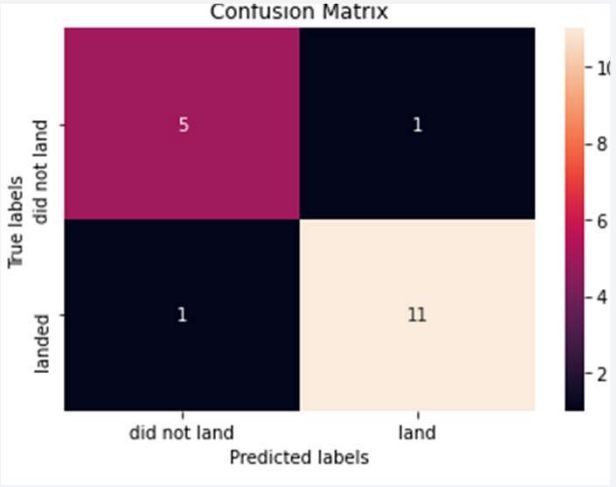
Classification Accuracy

- Tree model has the highest accuracy.



Confusion Matrix

- Examining the confusion matrix, we see that Tree can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- We can see that 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!