IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

John Ever Vino Duran
2024-11-27

Outline

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

Executive Summary

- Summary of methodologies
  - Colect Data, using SpaceX API and scraping its web page
  - Wrangle, create a manageable dataframe
  - EDA, Explore data with plots
  - Query with sql, analyze data
  - Plot in maps, explore lauch sites in the globe
  - Vizualize data in a interactive way to check success from lauch sites, payloadmass booster version
  - Build models to make predictions, models used: Logref, SVM, KNN, Tree
- Summary of all results
  - KSJ LC 39A site has more success rate
  - All of the launches sites are close to the coast
  - The best model to predict is Tree

Introduction

- Project background and context

SpaceX strives to make space travel affordable for everyone. It has rocket launches relative inexpensive because of its reusable rockets (Falcon 9). They published its data and it is interested in determine if their rocket will do a successful landing.

- Problems you want to find answers

How payload mass, launch site, orbit affects landing success

find a way to predict success

Section
1
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Two methods were used: consulting the SPACEX API and performing web scraping
- Perform data wrangling
  - For data wrangling, Python and the Numpy and Pandas libraries were used. To process the data first It was removed nan values, and creating new data columns from the existing ones to easily do the EDA
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - From the total data we select train data and test data we test with different models and select the model with best result on predict test data

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

Retrieve data from
https://api.spacexdata.com
Parsing the results, cleaning, construct the launch_dict and filtering by launches using Flacon 9 and save it.

- GitHub URL https://github.com/EverVino/data-science-capstone/blob/main/01-jupyter-labs-spacex-data-collection-api-v2.ipynb

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1
...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2

Data Collection - Scraping

Retrieve data from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Parsing the results with BeautifulSoup, dealing with missing values, construct the launch_dict and filtering by launches by Flacon 9 and save it.

GitHub URL: https://github.com/EverVino/data-science-capstone/blob/main/02-jupyter-labs-webscraping.ipynb

```
[9]: # Let's print the third table and check its content
     first_launch_table = html_tables[2]
     print(first_launch_table)

     <table class="wikitable plainrowheaders collapsible" style="width: 100%;">
     <tbody><tr>
     <th scope="col">Flight No.
     </th>
     <th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coo
     rdinated Universal Time">UTC</a>)
     </th>
     <th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Fal
     con 9 first-stage boosters">Version,<br/>Booster</a> <sup class="reference" id="cite_re
     f-booster_11-0"><a href="#cite_note-booster-11"><span class="cite-bracket">[</span>b<spa
     n class="cite-bracket">]</span></a></sup>
     </th>
     <th scope="col">Launch site
     </th>
     <th scope="col">Payload<sup class="reference" id="cite_ref-Dragon_12-0"><a href="#cite_n
     ote-Dragon-12"><span class="cite-bracket">[</span>c<span class="cite-bracket">]</span></
     a></sup>
     </th>
     <th scope="col">Payload mass
     </th>
     <th scope="col">Orbit
     </th>
     <th scope="col">Customer
     </th>
     <th scope="col">Launch<br/>outcome
     </th>
     <th scope="col"><a href="/wiki/Falcon_9_first-stage_landing_tests" title="Falcon 9 firs
     t-stage landing tests">Booster<br/>landing</a>
     </th></tr>
     <tr>
     <th rowspan="2" scope="row" style="text-align:center;
```

Would you like to get notified about official

Data Wrangling

- After getting the data we start to convert it into manageable data, first we load data, convert it in to a pandas dataframe, verify the missing values, dealing with null data, We focus in analize the outcome and create a new column called “class” which will be 1 if the landing outcome was successful otherwise 0, after data wrangling, the resulting dataframe was saved into a csv file

GitHub URLs:

https://github.com/EverVino/data-science-capstone/blob/main/03-labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

- Charts summary

FlightNumber vs PayloadMass (To check the evolucion the PayloadMass)
FlightNumber vs LaunchSite (To check which site has more success and has more Launches)
PayloadMass vs LaunchSite (To check Which site has more success with the PayloadMass)
Barchart in success rate for Orbit (To check Which orbits has more successes)
FlightNumber vs Orbit type (To check success over the number of flights)
PayloadMass vs Orbit type (to check in which orbit could be launched differente PayloadMass)
Date vs Class (To check the evolucion of success in launches)
GitHub URL:
https://github.com/EverVino/data-science-capstone/blob/main/05-edadataviz.ipynb

EDA with SQL (1)

Performed SQL queries

- select * from SPACEXTABLE limit 5;
- select distinct Launch_Site from SPACEXTABLE;
- select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5;
- select sum("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer"="NASA (CRS)";
- select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" like "F9 v1.1";
- select min("Date") from SPACEXTABLE where "Landing_Outcome" == "Success";
- select "Booster_Version" from SPACEXTABLE where "Mission_Outcome" == "Success" and "PAYLOAD_MASS__KG_" between 4000 and 6000;

GitHub URL:
https://github.com/EverVino/data-science-capstone/blob/main/04-jupyter-labs-eda-sql-coursera_sqlite.ipynb

EDA with SQL(2)

Performed SQL queries

- select "Mission_Outcome", count(*) from SPACEXTABLE group by "Mission_Outcome";
- select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE);
- select substr(Date,6,2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where "Landing_Outcome" = "Failure (drone ship)" and substr(Date,0,5) = "2015";
- select "Landing_Outcome", count(*) as Count from SPACEXTABLE group by "Landing_Outcome" having "Date" between "2010-06-04" and "2017-03-20" order by Count desc;

GitHub URL:
https://github.com/EverVino/data-science-capstone/blob/main/04-jupyter-labs-eda-sql-coursera_sqlite.ipynb

Build an Interactive Map with Folium

Objects added to the map

- In the Map we adder marker, circle and icon with Folium for NASA location and Launch sites locations, to check if they were near to equator line and if they were near to coastline, cities, highline and railroads
- And added lines to identify proximity to a coastline, city highline and railroads

GitHub URL: https://github.com/EverVino/data-science-capstone/blob/main/06-lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard

In the dashboard we added a piechart and a scatter plot with a drop down to select the launch site and a Mass Slider to select the PayloadMass
When the site selected is “ALL” in the pie chart is showed the relation of success from all launch sites, and in the scatter plot is showed payloadMass vs class (success) differentiated by Launch Site. But if the we select a specific launch site we will see in the piechart the percentage of success and failure, and in the scatter plot PayloadMass vs the class (success) differentiated by Booster version.

GitHub URL:

rawcode:
https://github.com/EverVino/data-science-capstone/blob/main/07-spacex_dash-raw-code.txt

screenshots :
https://github.com/EverVino/data-science-capstone/tree/main/07-dash-screenshots

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- To perform a classification we start selecting independent variables and dependant variable, Y (dependent variable) = class column (success or failure) and the other columns like Flight Number, PayLoadMass Orbit type … were the X (independent variables), we standarize the data in X and from the data we choose a test size of 0.2.
- We try different models: Logistic Regression, SVM, tree and k nearest neighbor, after that we calculate the scores and accuracy predicting with the test sample

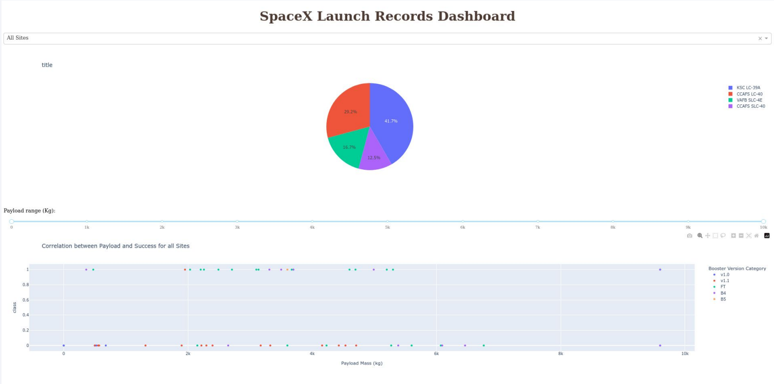
GitHub URL:
https://github.com/EverVino/data-science-capstone/blob/main/08-SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

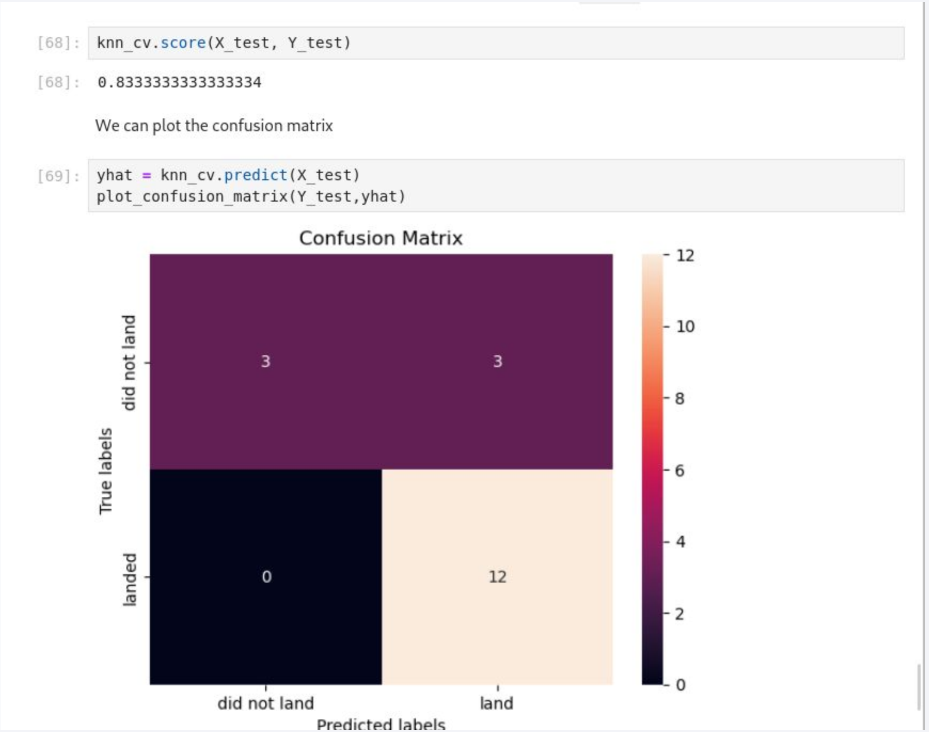
Results

- Exploratory data analysis results
    - From the data collect it was retrieved enough data to perform a data analysis. It has been noticed that as Flight Number increases the success rate increases as well
    - Some key results: average success 0.67 over 1
    - Most of success were in the launch site CCAFS SLC 40
    - Most successes were in orbit GTO

Results

- Interactive analytics demo in screenshots





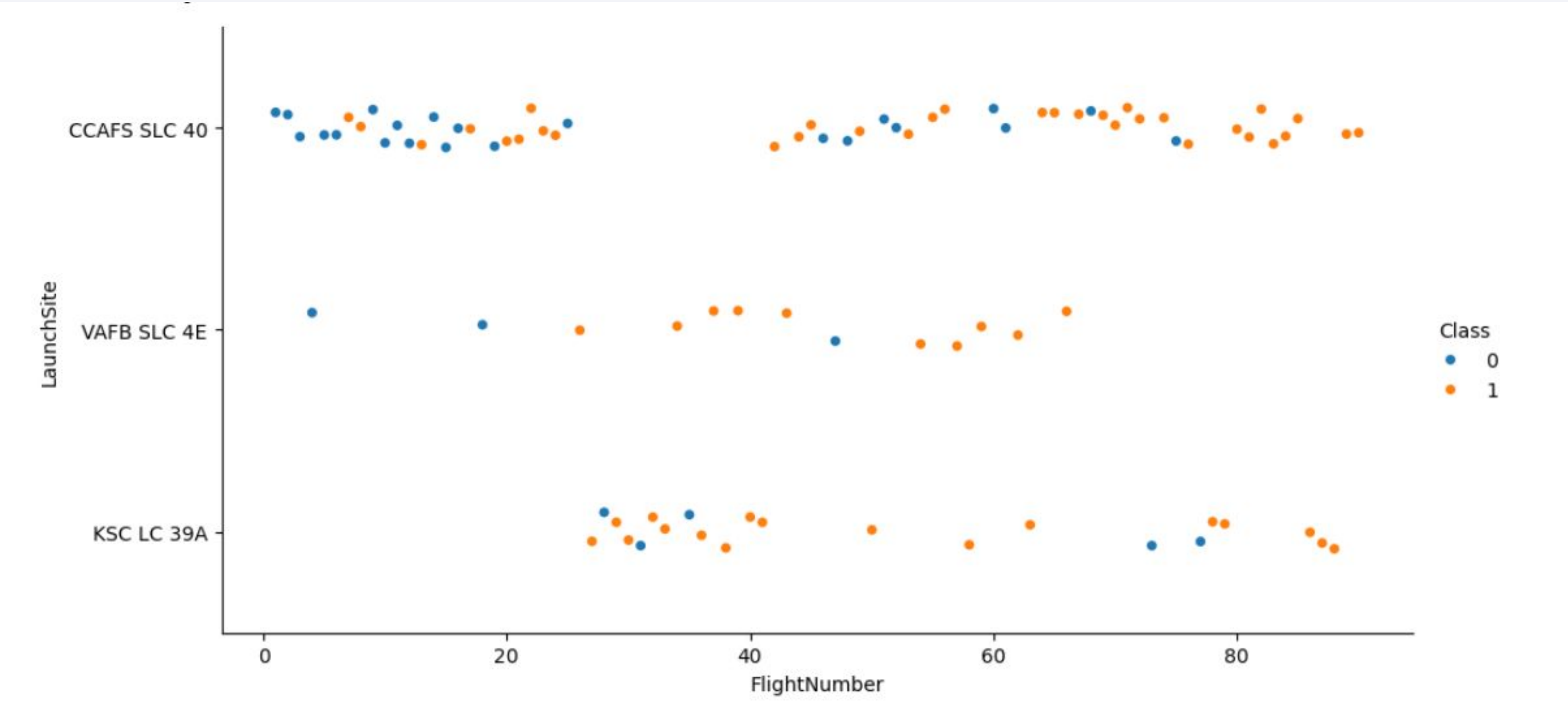
Results

- Predictive analysis results
    - Best model: Tree
    - Accuracy: 0.875
    - Score: 0.9444444444444444

Section
2

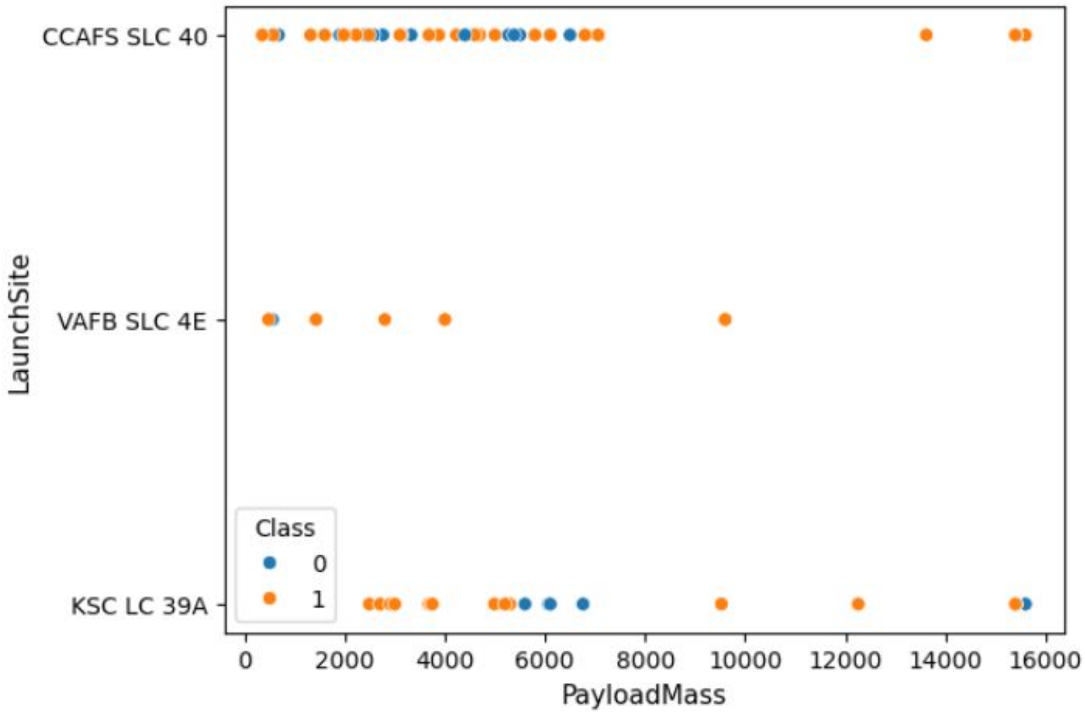
Insights drawn from EDA

Flight Number vs. Launch Site



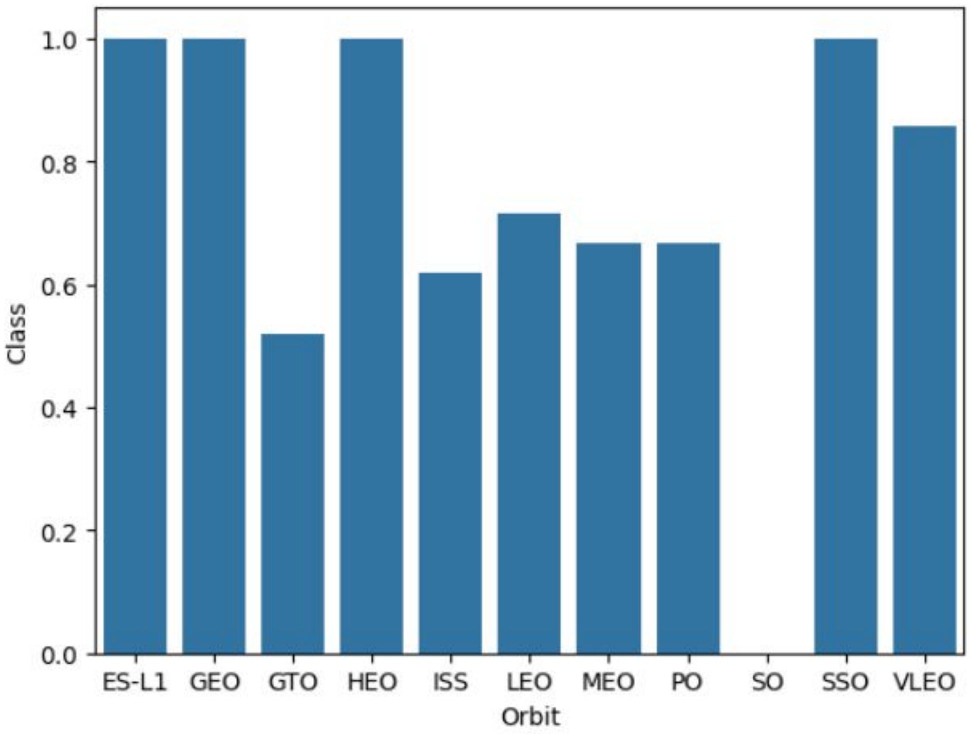
We have more launches from CCASD SLC 40 Site, Last launches are comming from KSC LC 39A and CCASD SLC 40 and the success increase with the flight number for every launch site

Payload vs. Launch Site



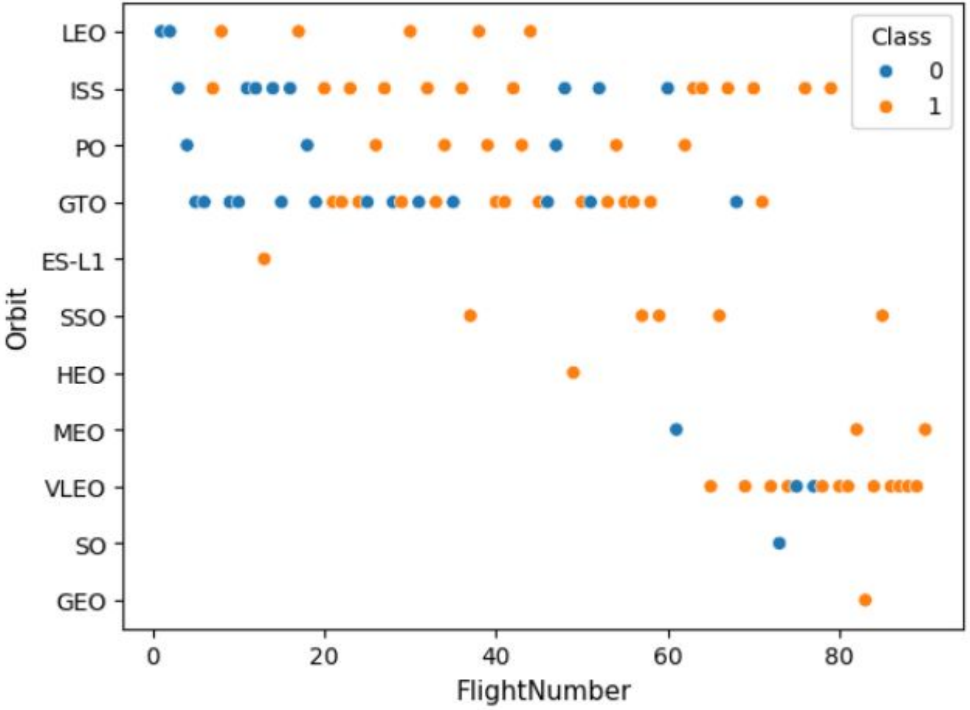
Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



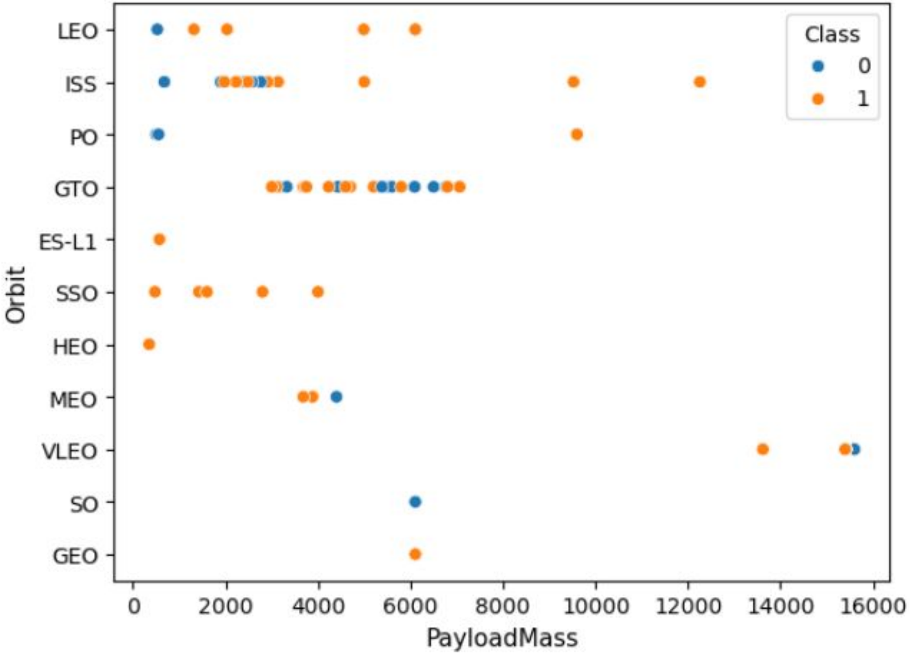
Orbit with highest rate = ES-L1, GEO, HEO, SSO, VLEO

Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

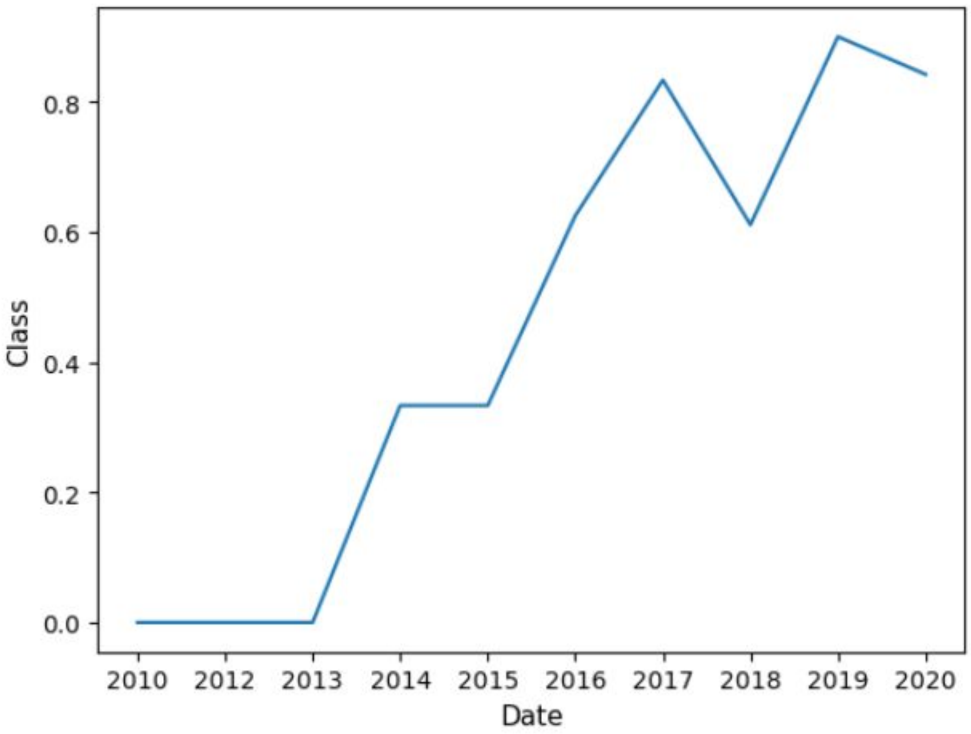
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



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

All Launch Site Names

- Unique launch sites
    - CCAFS LC-40
    - VAFB SLC-4E
    - KSC LC-39A
    - CCAFS SLC-40

- Using sql queries we get these launch sites

Launch Site Names Begin with 'CCA'

- Launch sites begin with `CCA`

[9]:
```
%%sql
select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5;
```

Running query in 'sqlite:///my_data1.db'

[9]:

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

The results shows that all belongs to the same launch site

Total Payload Mass

Total payload carried by boosters from NASA
45596 Kg

The total payload mass is the sum of all mass carried out in the data

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
2928.4 Kg

- That mass is the average of payloadmass per launch

First Successful Ground Landing Date

- First successful landing outcome on ground pad

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2018-07-22	5:50:00	F9 B5B1047.1	CCAFS SLC-40	Telstar 19V	7075	GTO	Telesat	Success	Success

- To do the query we search for min(date) where LandingOutcome is Success

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

F9 v1.1, F9 v1.1 B1011, F9 v1.1 B1014, F9 v1.1 B1016, F9 FT B1020, F9 FT B1022, F9 FT B1026, F9 FT B1030F9, FT B1021.2, F9 FT B1032.1, F9 B4 B1040.1, F9 FT B1031.2, F9 FT B1032.2, F9 B4 B1040.2, F9 B5 B1046.2, F9 B5 B1047.2, F9 B5 B1046.3, F9 B5 B1048.3, F9 B5 B1051.2, F9 B5B1060.1, F9 B5 B1058.2, F9 B5B1062.1

- The query filter in range of payloadmass and has an Mission_Outcome =Success

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Running query in 'sqlite:///my_data1.db'

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

The query group and count “Mission_outcome” results

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
- To list the booster names we have to do a subquery searching for the max value in payloadmass

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

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

Query was made filtering by date

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

- Ranking 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

query was made grouping by landing outcome, filtering by date range and order by count

Landing_Outcome	Count
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

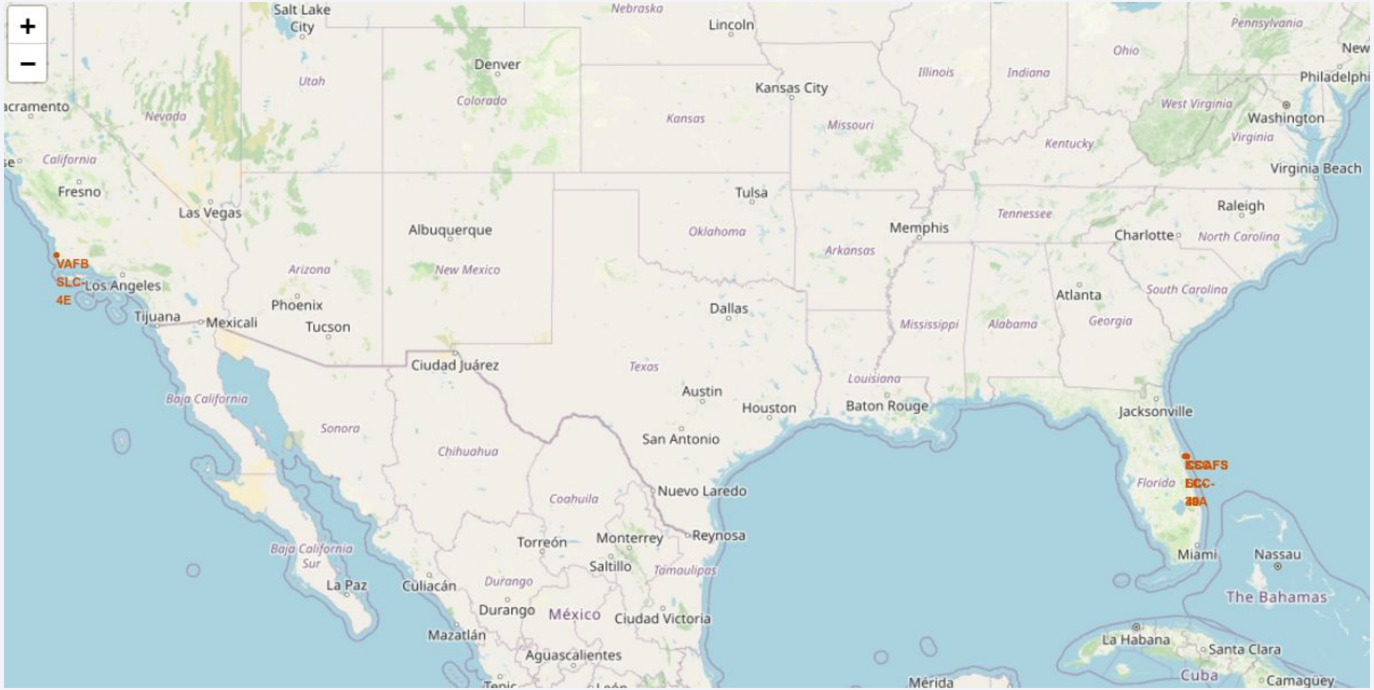
Section
3

Launch Sites Proximities Analysis

Launch Site Location

Launch sites are near to the equatorial line inside USA

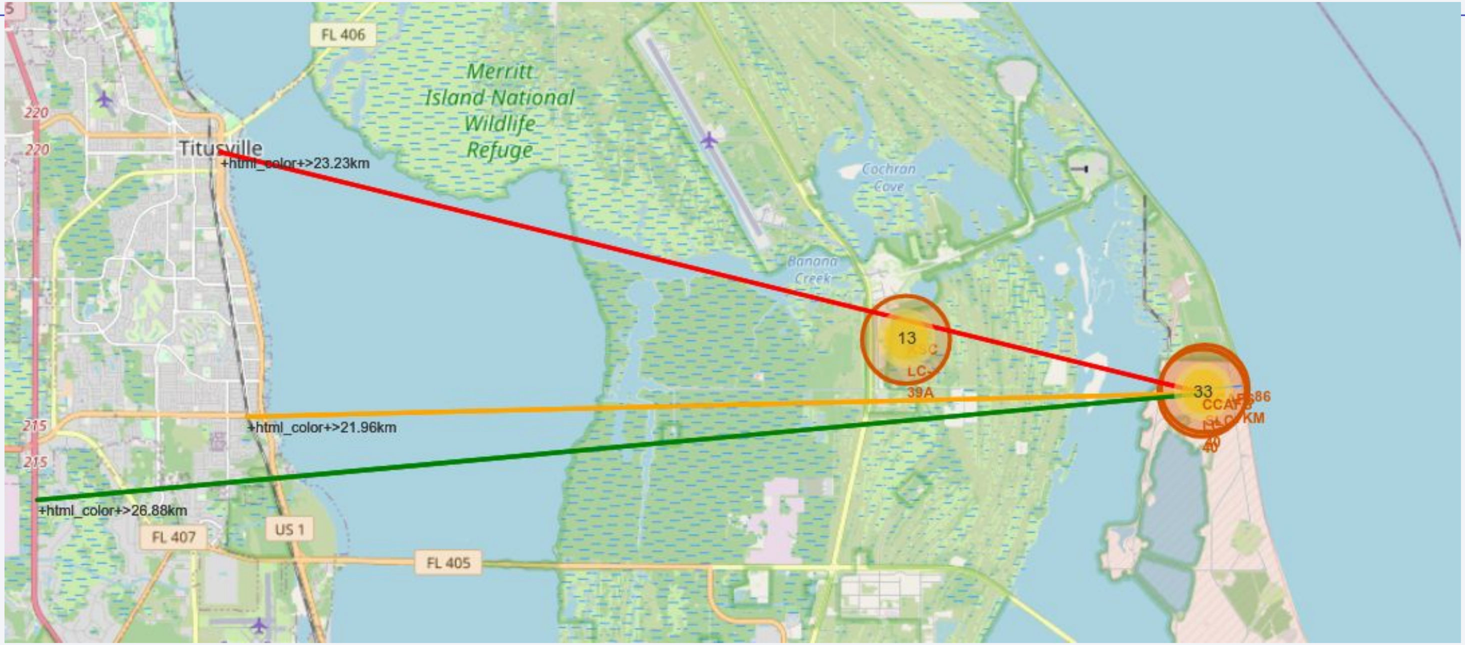
All launch sites are near to the coastline



Success/failed launches for each site

Most of the launches come from the east coast

Proximity to coastline, city, railroad and highlines

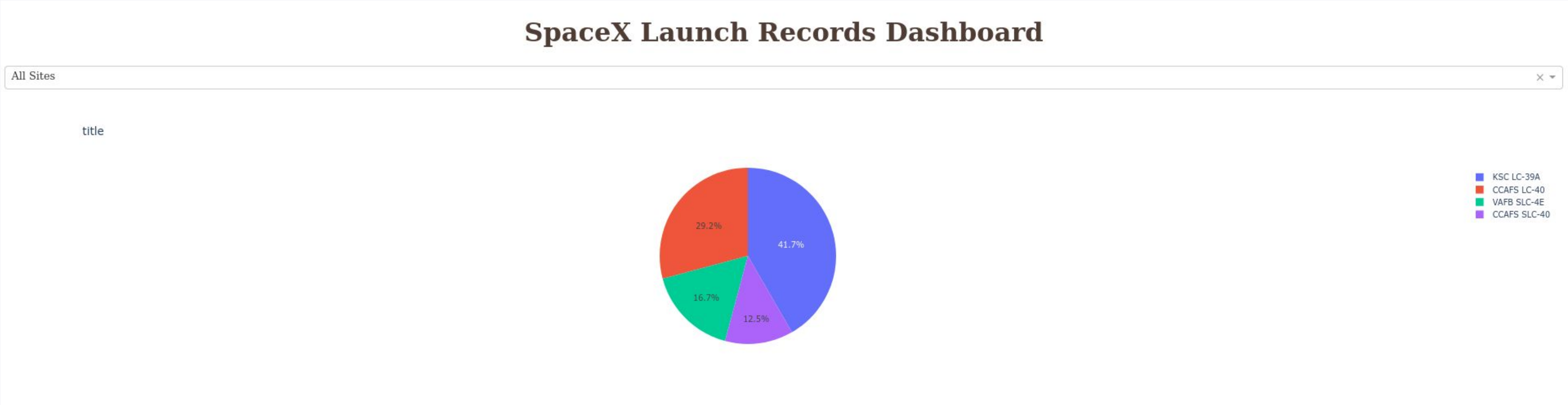


All launches sites are near to railways and highways and coastline but all of them as well is not too near from the cities. Having a transportation medium decrease expenses for material transportation and comunication, Keeping distance from cities reduce risk from population but at the same time being not to far allow help, human resources and services when needed.

Section
4

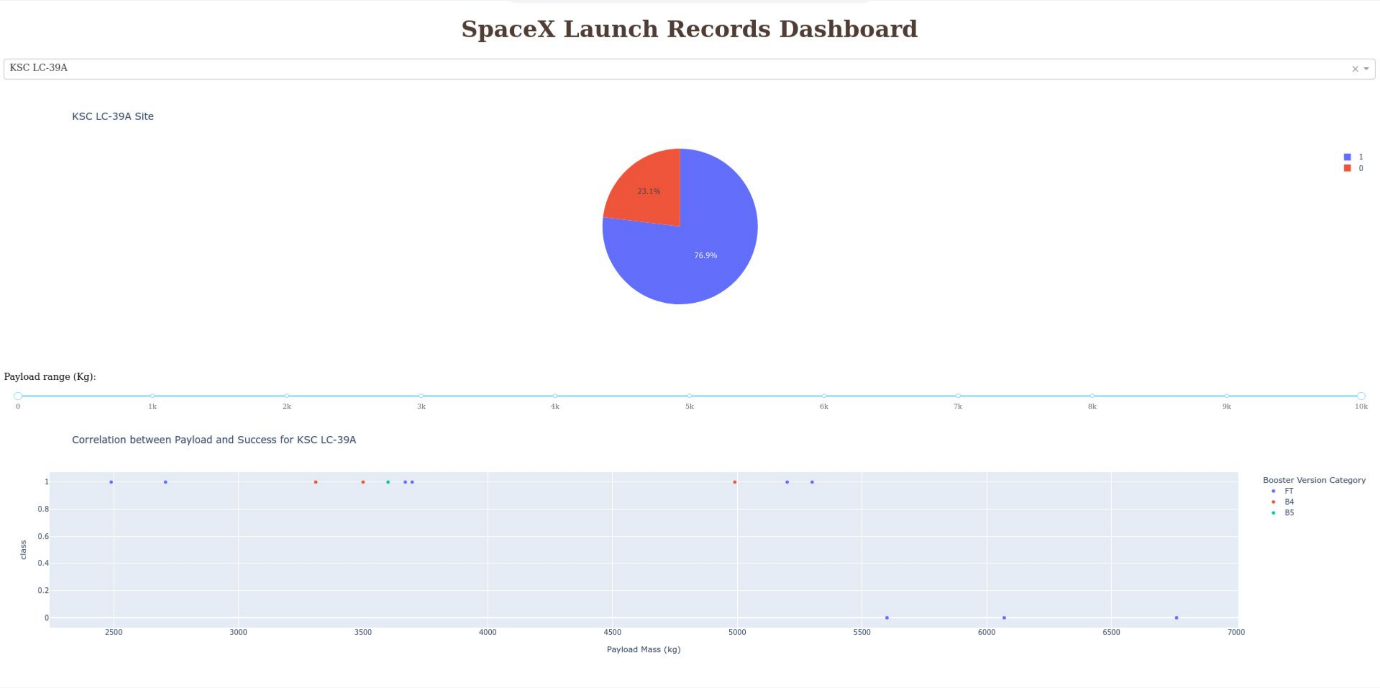
Build a Dashboard with Plotly Dash

Success rate from all sites



The site with more success is coming from the site KSC LC-39A

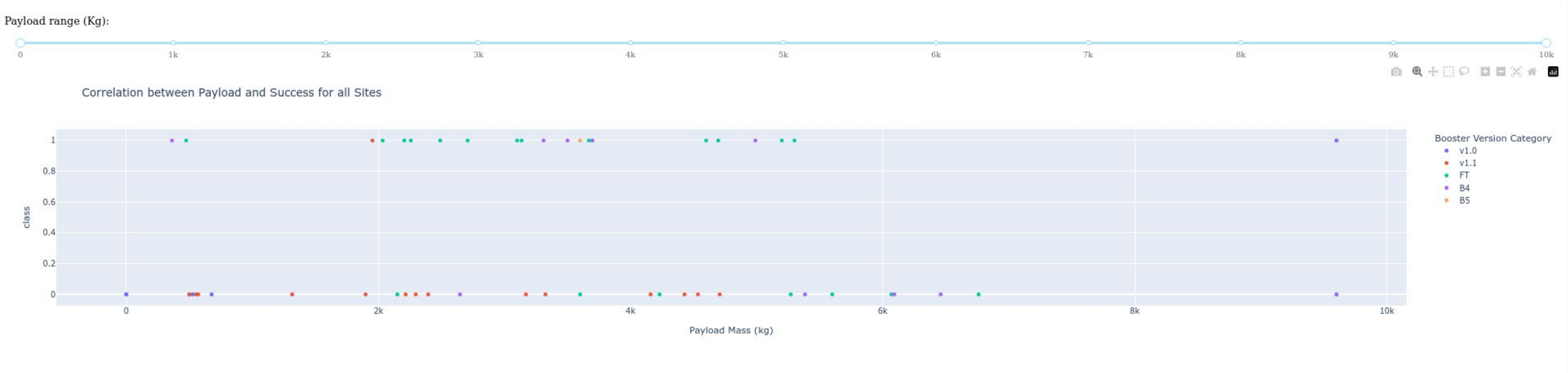
Most successful site launches



Launch site KSC LC 39A

has a success rate of 76.9 %

Payload vs Success (Booster ver)



- FT Booster has most succeeded launches

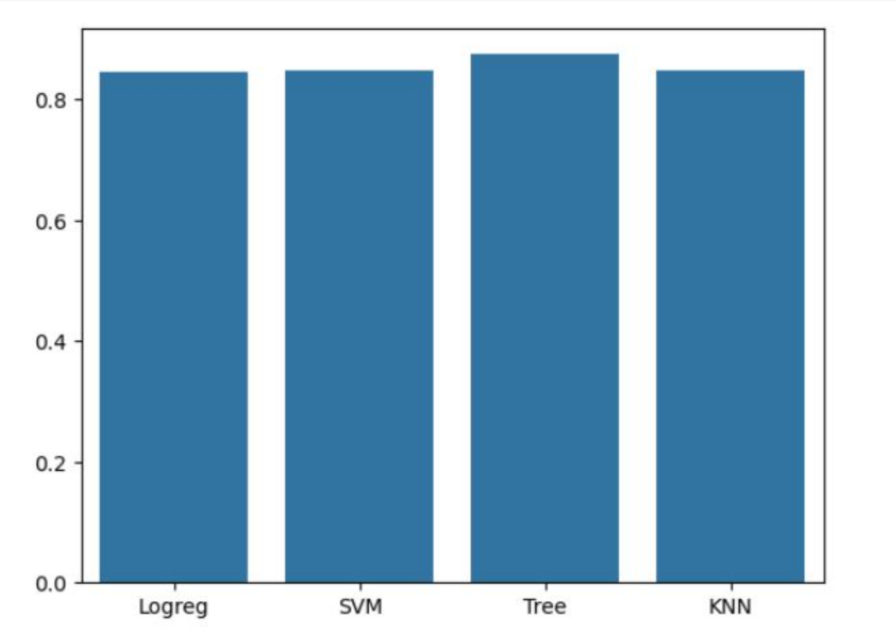
Section
5

Predictive Analysis (Classification)

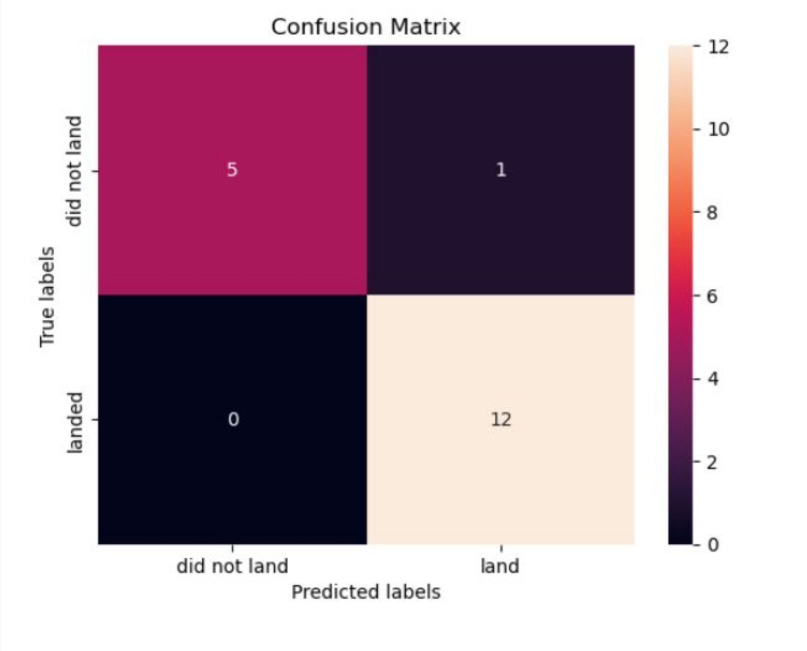
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- The highest classification accuracy if Tree model



Confusion Matrix



- True Postive - 12 (True label is landed, Predicted label is also landed)
- False Postive - 5 (True label is not landed, Predicted label is landed)

Conclusions

- KSJ LC 39A site has more success rate
- All of the launches sites are close to the coast, highline, railroads and they are not to far from a city
- Launches are improved over time
- Orbit ES l1, GEO, HEO and SSO have 100 % success rate
- The best model to predict successful launches are Tree
-

Appendix

- All notebook and saved data can be found in my repo https://github.com/EverVino/data-science-capstone

Thank you!