IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gautam VR
30/1/2022

Outline

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

Executive Summary

- Summary of methodologies:
  - The various methodologies used to arrive at the conclusion were different types of visual analysis with Scatter plot, bar chart, pie chart, line plot, etc and we used various classification models to find the best performing model on the given dataset to provide the best accuracy.
- Summary of all results
  - The results are found out to be the best model with the highest accuracy is Decision Tree classifier with the accuracy being 88.88%
  - The various visual analysis have been plotted as screenshots

Introduction

- Space X 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 Space X can reuse the first stage.

- Objectives:

  1. To determine if the first stage will land
  2. To determine the cost of the launch
  3. To use this information if an alternate company wants to bid against SpaceX

Landing outcomes

Successful Landing

Failed Landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using “requests” to get data from SPACEXAPI
  - Web Scrapping by using BeautifulSoup package to get public information
- Perform data wrangling
  - Dealt with missing values and NAN values, cleaning the data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compared various classification models – (KNN, Logistic Regression, Decision Trees)
  - Used GridSearchCV to find the perfect hyperparameters needed for the models

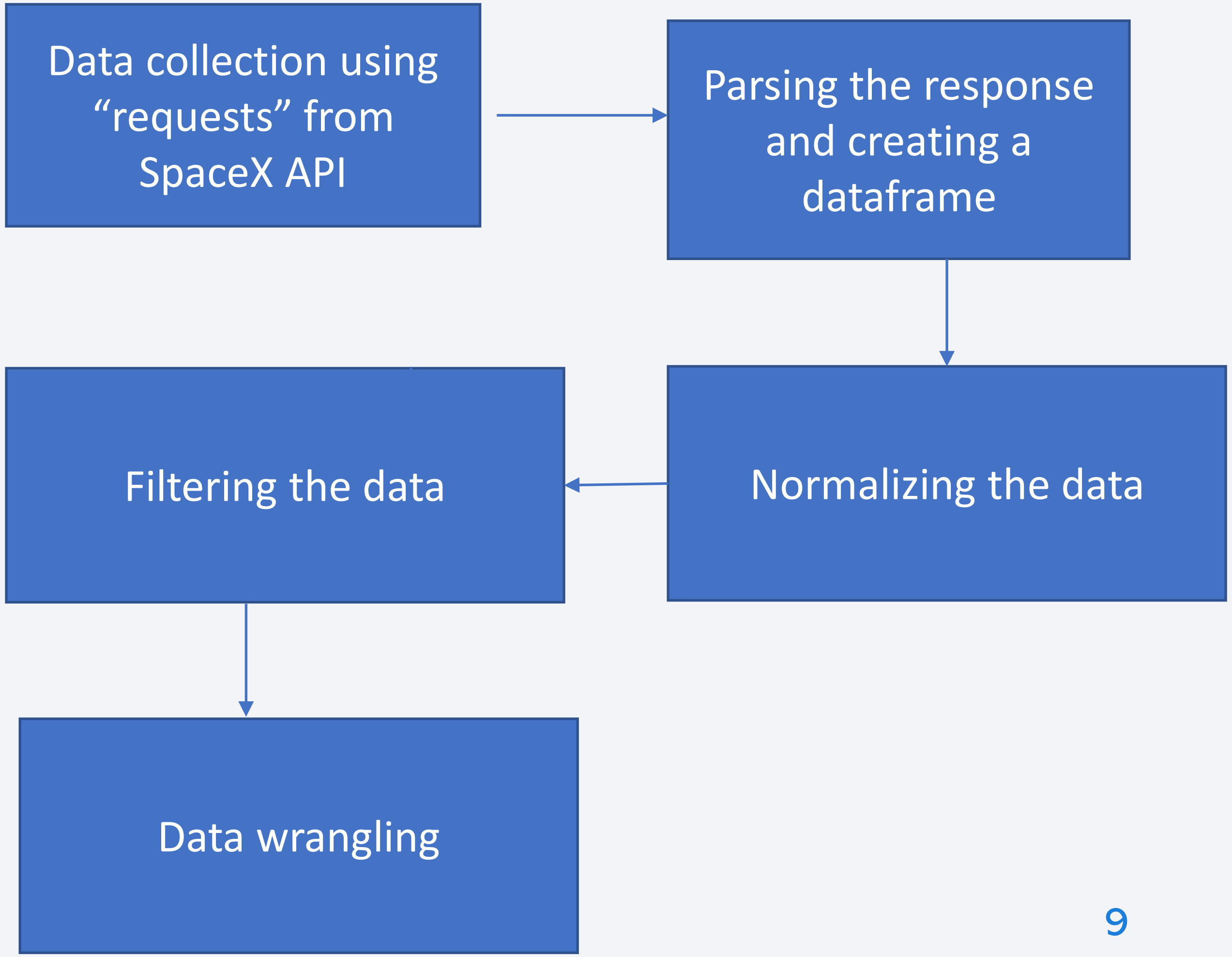
Data Collection

- Data sets were collected using the “requests” package to get the various data from the SPACEX API which is available to public that has information about the booster versions, landing outcome, launch sites, etc
- We then use the BeautifulSoup package to get the data from Wikipedia where the information about various boosters, it’s payload information, etc are present.

Data Collection – SpaceX API

- Collected the data using requests with the help of SPACEX API, by parsing the response from the api call, normalizing the data, filtering the data, wrangling.

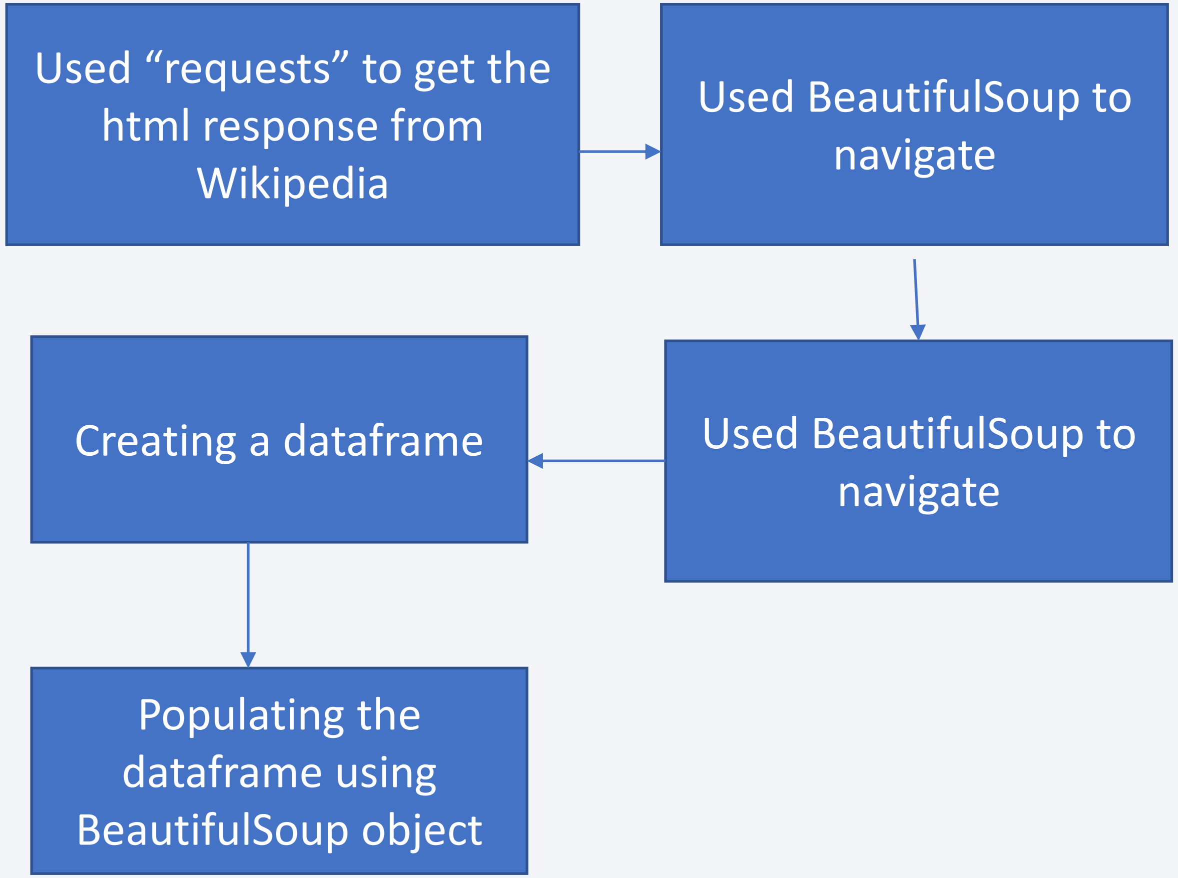
- GitHub URL for DataCollection:
    - https://github.com/gautamvr/SpaceY/blob/master/Space_DataCollection.ipynb



Data Collection - Scraping

- Used requests to get the response from the Wikipedia link.
- Used beautifulSoup package to navigate through the tables, rows and populate the data frame.

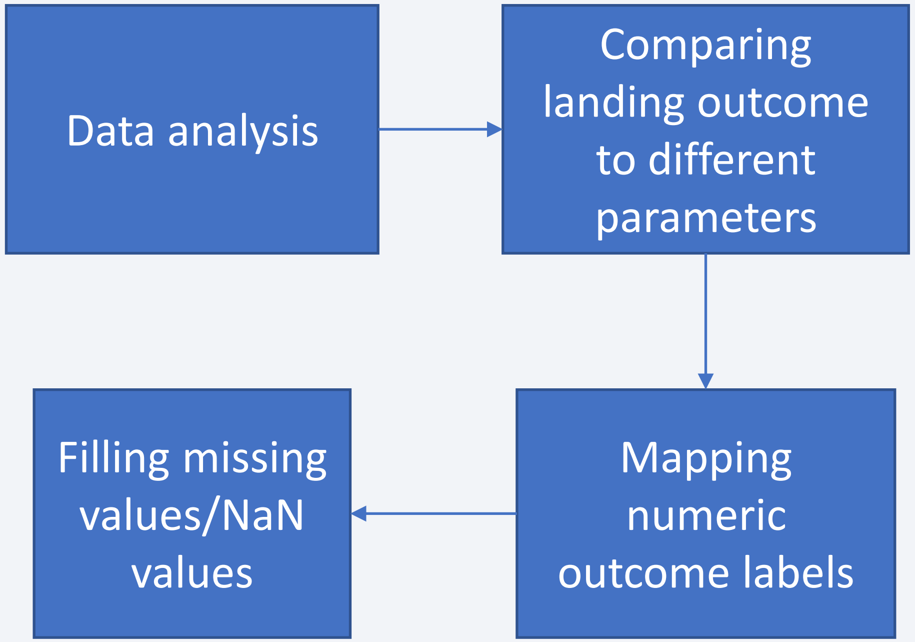
- GitHub URL for web scraping:
  - https://github.com/gautamvr/SpaceY/blob/master/WebScrapping_SpaceData.ipynb



Data Wrangling

- Data wrangling was process on our collected data by performing data analysis and finding the landing outcomes related to different parameters, mapping numeric outcome label for each outcome and filling out missing values or NAN values

- GitHub URL of data wrangling:
  - https://github.com/gautamvr/SpaceY/blob/master/DataWrangling_EDA.ipynb



EDA with Data Visualization

- Plotted various charts to determine the required features for the classification. (Scatter Plot, Line Plot, Bar Plot)
- These charts were plotted and feature engineering was performed to arrive at the few important features that would affect the success rate. Those features were found out to be:
  - FlightNumber, PayloadMass, Orbit, LaunchSite, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial
- GitHub link to the DataVisualization project:
  - https://github.com/gautamvr/SpaceY/blob/master/EDA_Visualization.ipynb

EDA with SQL

- Used various SQL queries to perform data analysis and gather sensible knowledge from the collected data. The queries include:
    - SELECT
    - SUM,AVG,MIN,MAX
    - GROUP BY
    - ORDER BY
    - UNIQUE
- GitHub URL for EDA with SQL:
    - https://github.com/gautamvr/SpaceY/blob/master/EDA_withSQL.ipynb

Build an Interactive Map with Folium

- Various map objects were used to plot the coordinates of the Launch sites in the map to analyze insights from the map.
- The map objects include:
    - Markers – To denote the name of the particular Launch site using a label
    - Circles – To indicate the region of the coordinate in the defined circle radius
    - MarkerCluster – To indicate the cluster of the launch outcomes at a particular launch site (Whether the landing was successful or not)
    - Lines – To measure the distance from the launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- GitHub URL of the interactive map project:
    - https://github.com/gautamvr/SpaceY/blob/master/InteractiveVisualAnalytics.ipynb

Build a Dashboard with Plotly Dash

- PieCharts and Scatter Plot was used in the Dashboard to get insights for the various payload weights and the launch site inputs
- These plots were used to collect the data based on the interactive inputs that was provided in the dashboard for the launch sites and the payload weights that the used can choose.
- GitHub URL of the Plotly Dash lab:
    - https://github.com/gautamvr/SpaceY/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- The data collected were first normalized using ScalarTransform()
- The normalized data was then split into training data and testing data
- The data was then passed to various models to find the best accuracy in the training set
- Each model's training was fit using GridSearchCV to find the best hyperparameters that yielded high accuracy.
- flowchart
- GitHub URL of the predictive analysis lab:
  - https://github.com/gautamvr/SpaceY/blob/master/MachineLearningLab.ipynb

Results

The following slides contain the results of the below information:

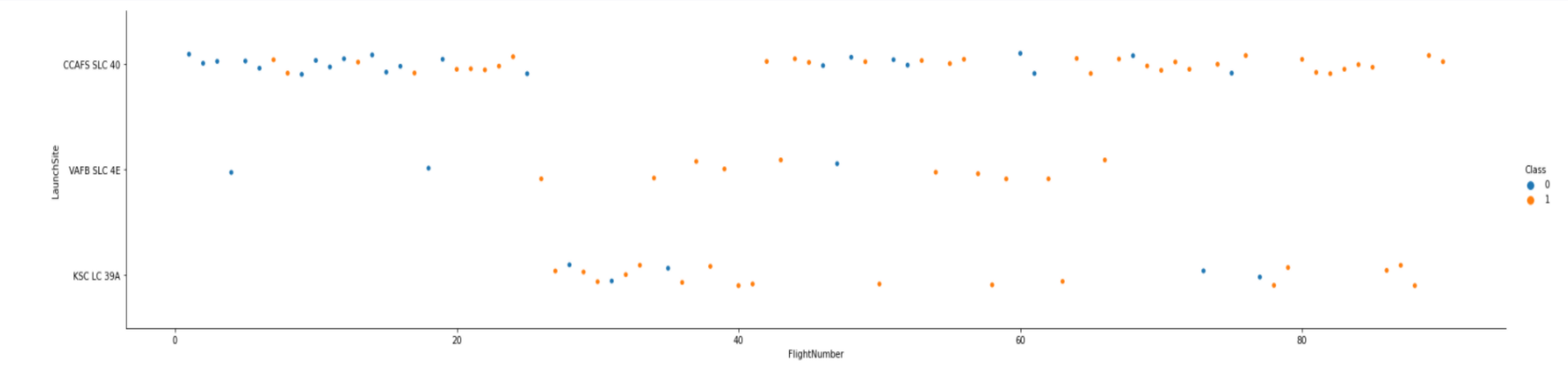
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

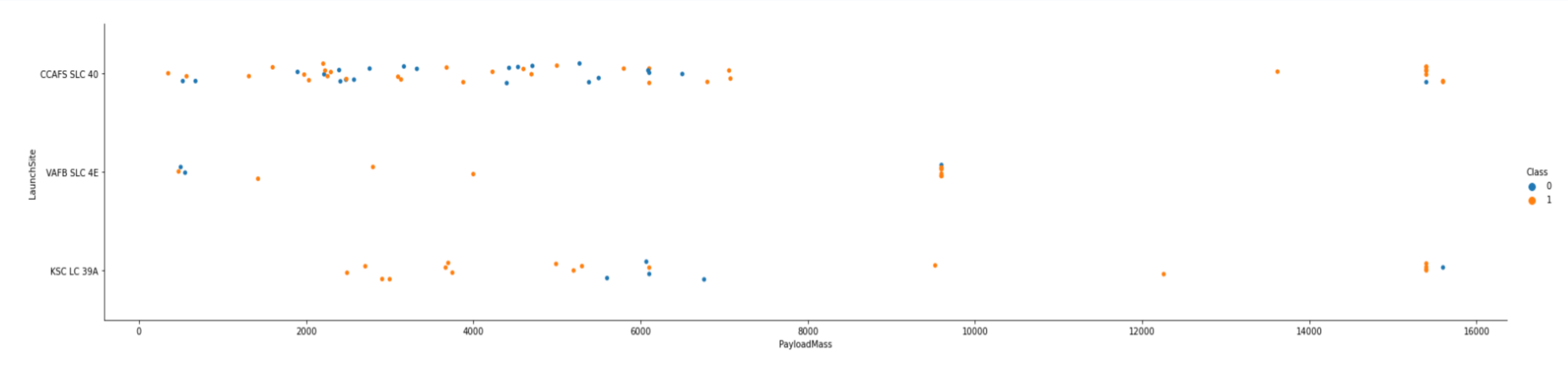
- Scatter plot of Flight Number vs. Launch Site



We see that as the flight number increases, the success rate is good in the CCAFS SLC 40 Launch site.

Payload vs. Launch Site

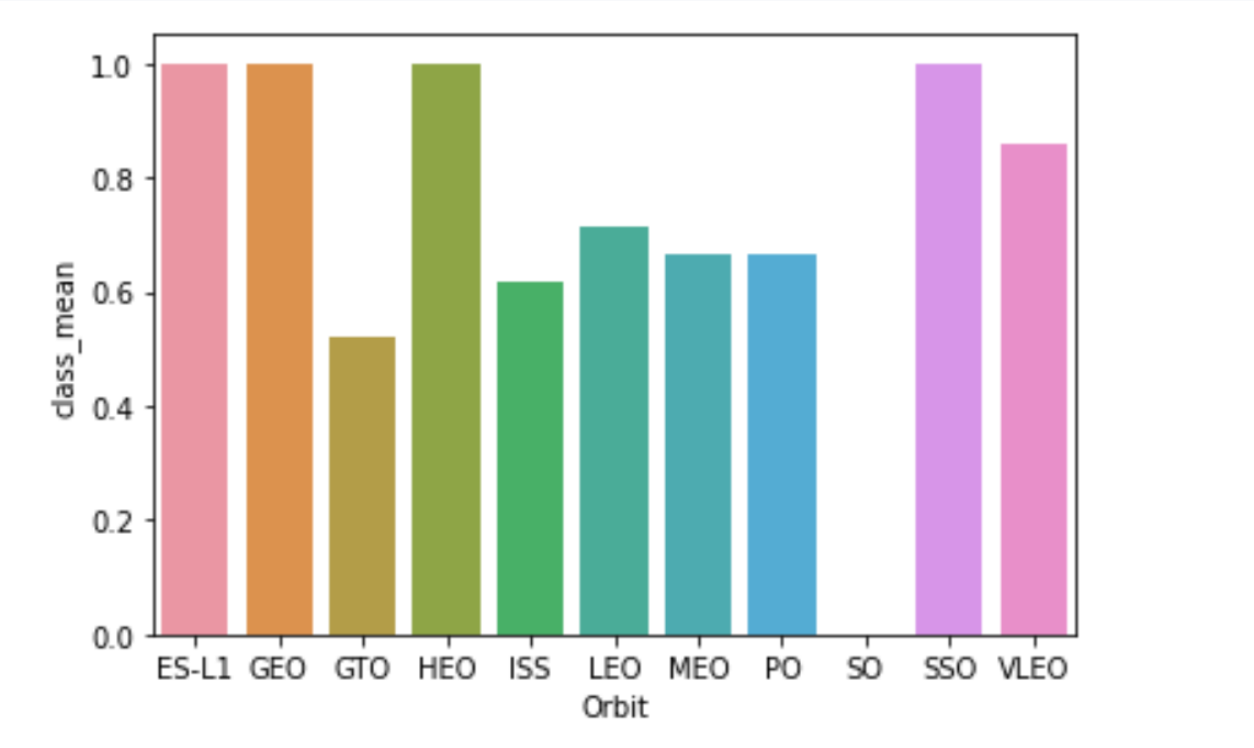
- Scatter plot of Payload vs. Launch Site



If we observe Payload Vs. Launch Site scatter point chart, we can find for the VAFB-SLC launch site there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

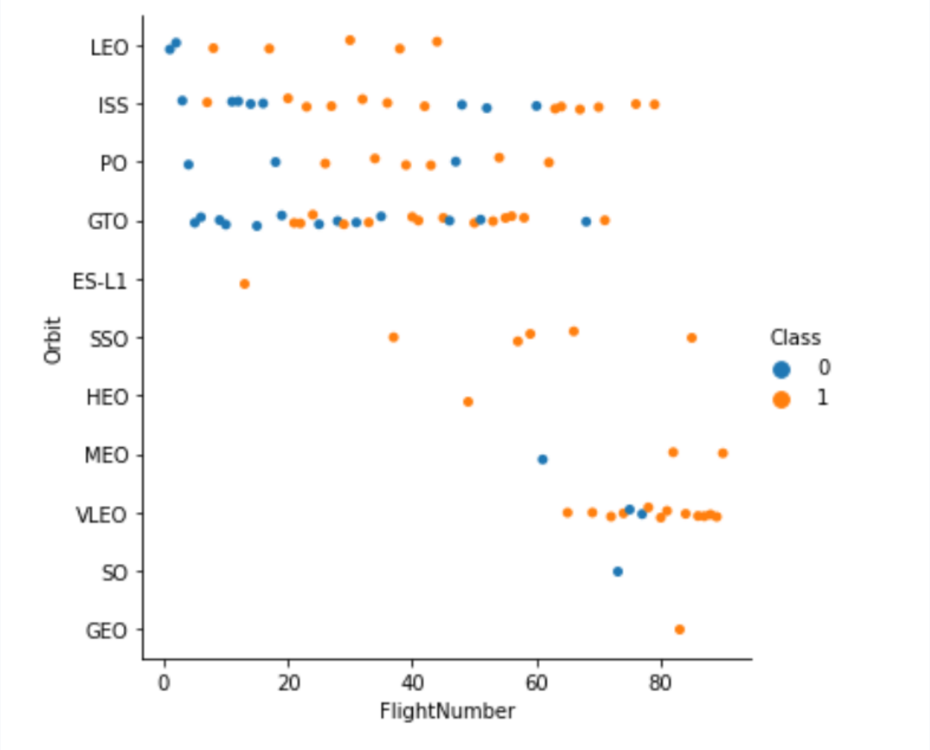
- Bar chart for the success rate of each orbit type



From this data, we see that orbits – [ES-L1,GEO, HEO, SSO] have the highest success rate compared to the other orbits

Flight Number vs. Orbit Type

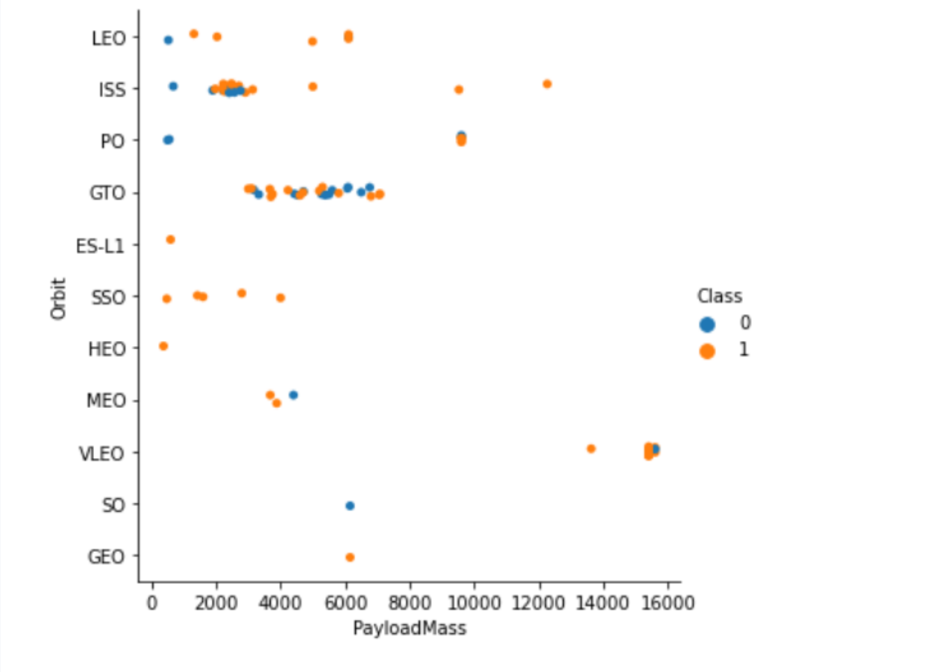
- Scatter point of Flight number vs. Orbit type



We 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

- Scatter point of payload vs. orbit type

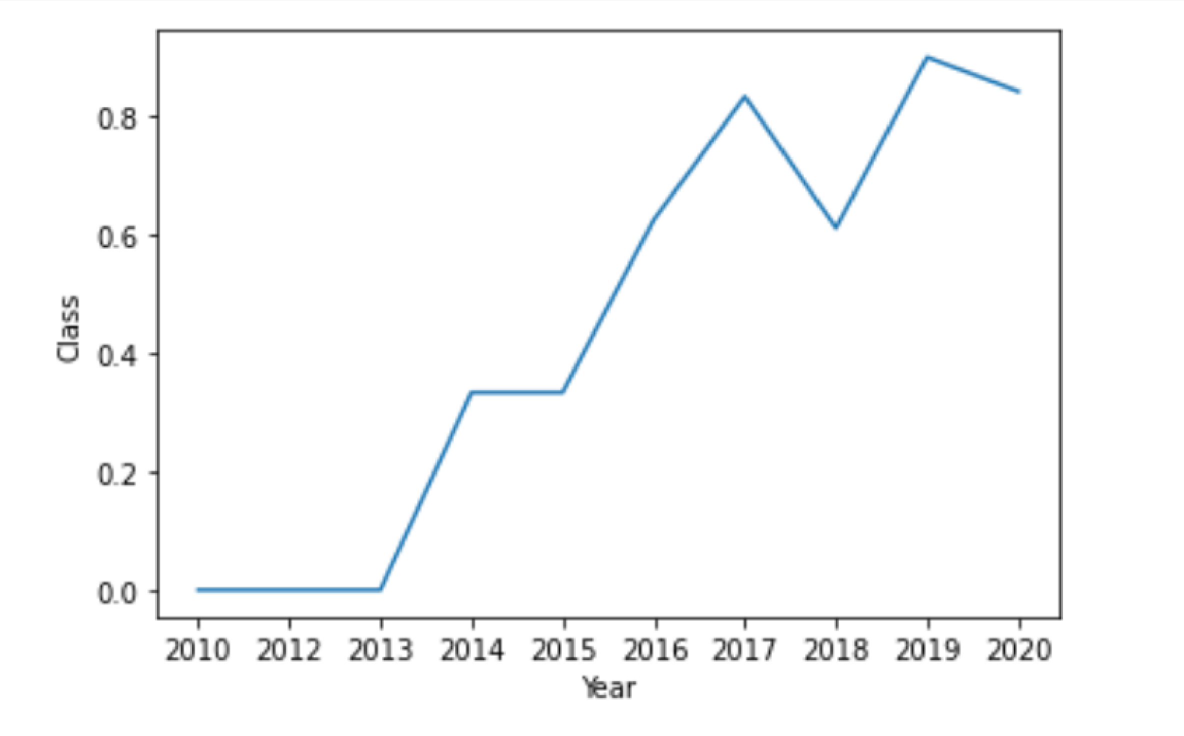


With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Line chart of yearly average success rate



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

All Launch Site Names

- The names of all the launch sites

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

These were the unique launch sites names that was present in the column, which was extracted using UNIQUE keyword through SQL.

Launch Site Names Begin with 'CCA'

- The 5 records where launch sites begin with `CCA`

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

These are the 5 of the records where the launch site's name begin with 'CCA'. This was collected using WHERE conditional operation and LIKE keyword, and limited to 5 using LIMIT keyword.

Total Payload Mass

- The total payload carried by boosters from NASA

```
%%sql
Select Sum(payload_mass__kg_) from SPACEXDATASET
Where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://zgr22430:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
45596

The total payload carried by all the boosters from NASA are 45596 kgs

Average Payload Mass by F9 v1.1

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

```
%%sql
Select AVG(payload_mass__kg_) from SPACEXDATASET
Where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://zgr22430:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
2534

The average payload mass carried by the F9 V1.1 version of booster is 2534 kgs.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad

```
%%sql
Select MIN(DATE) from SPACEXDATASET
Where mission_Outcome = 'Success'
```

```
 * ibm_db_sa://zgr22430:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
2010-06-04

The first successful landing on the ground pad was on 4th June 2010

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- The data provided is the names of the boosters which had payload mass between 4000kg and 6000kg. This was found using WHERE, AND, BETWEEN keywords from SQL.

booster_version
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 B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

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

We see that total number of successful mission outcomes are 99 + 1(payload status unclear) and only 1 is failure.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

These are the booster version that have carried 15600 kgs of payload mass, which is the highest from the data.

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

Landing _Outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

These are failed landing outcomes in the year 2015 collected from SQL

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.

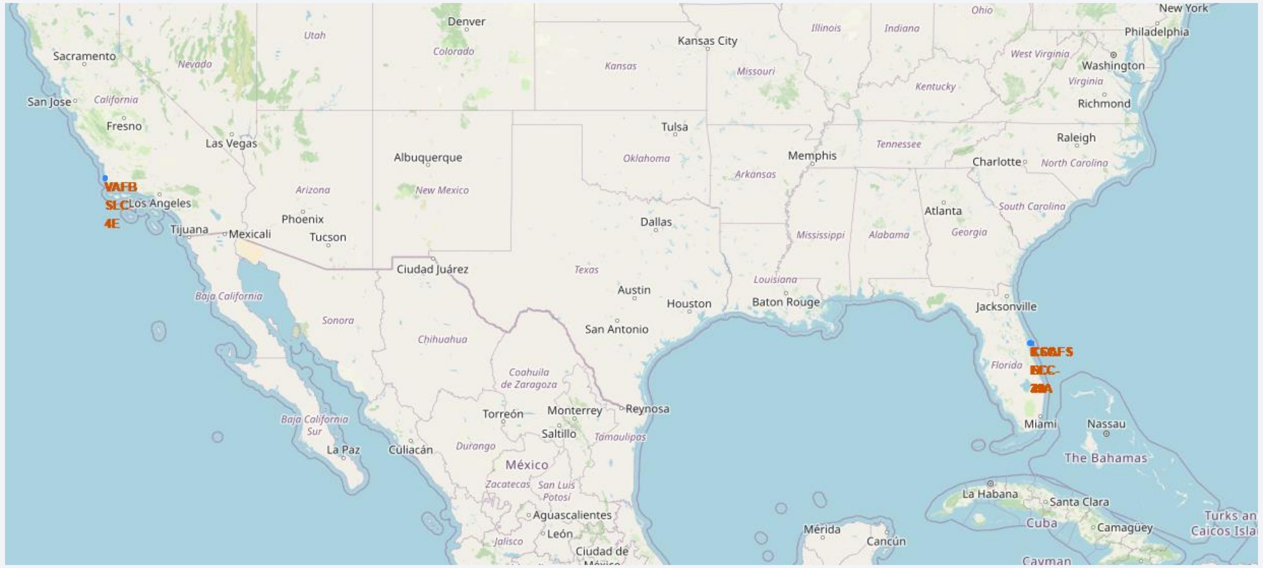
These are counts of each landing outcomes between the given data, in descending order – Collected using GROUP BY, DESC keywords through SQL

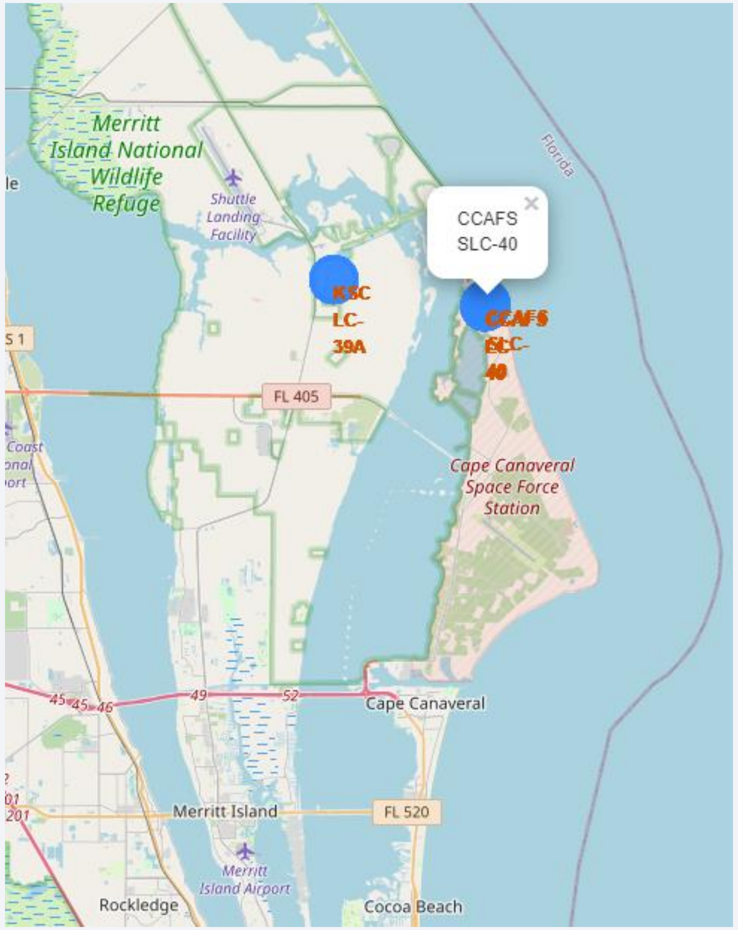
Landing _Outcome	**2**
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

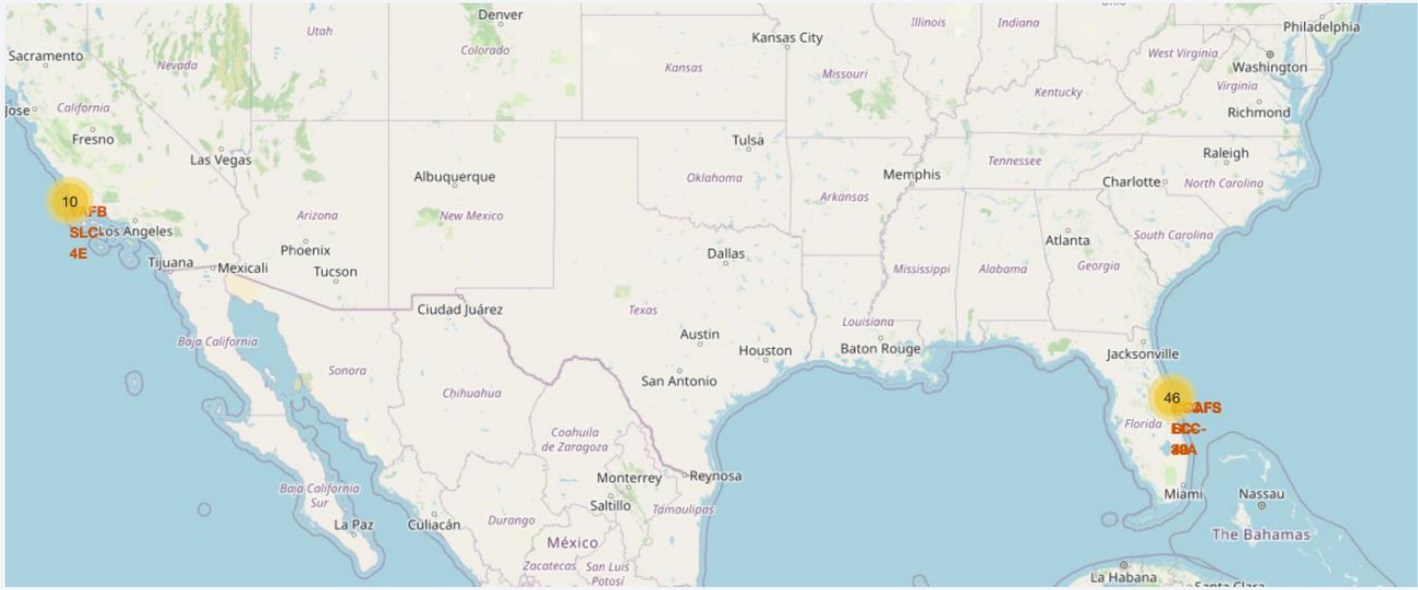
Map of the launch site's location markers

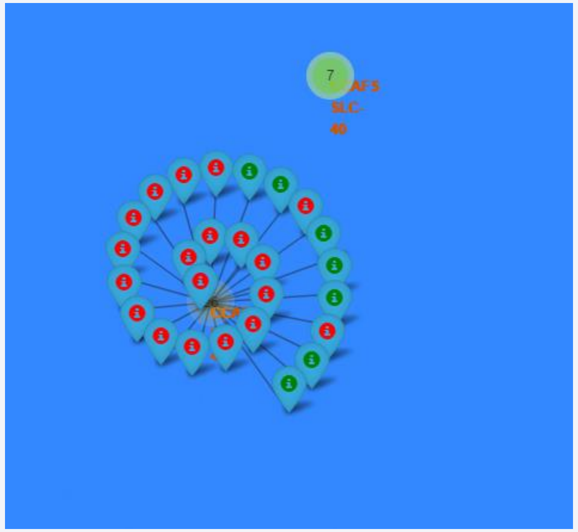




The location of the launch sites, marked on the map with interactive markers which shows the names of each launch site's name in the pop up.

Map of Launch sites and it’s outcomes





The maps which has color labelled landing outcomes for each Launch site

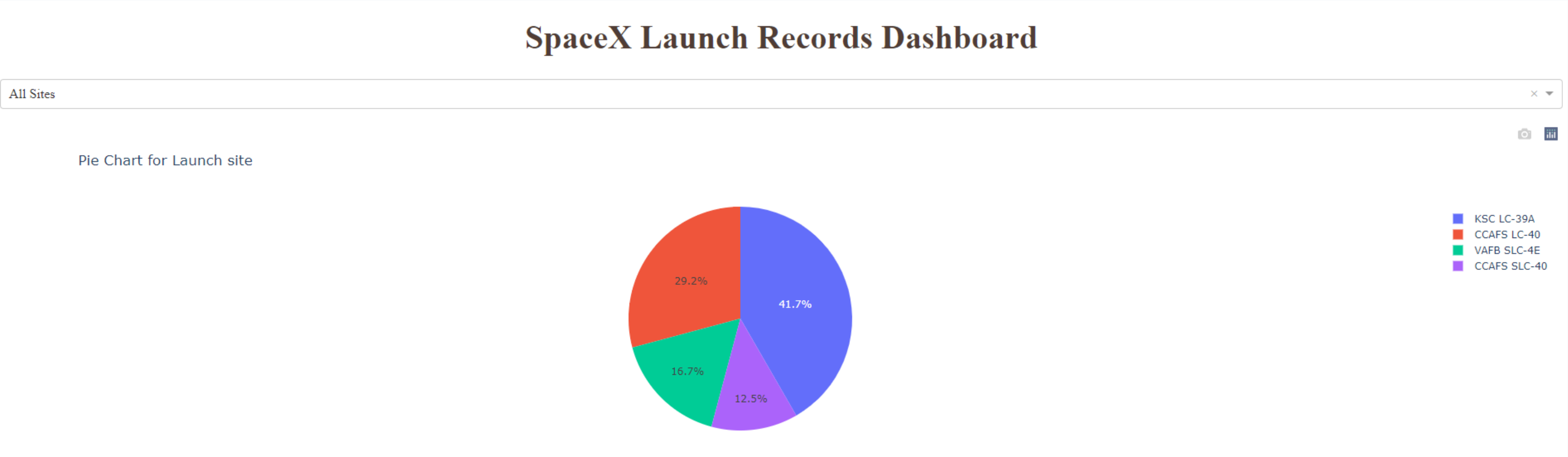
Map with Launch sites to it’s proximities

The map shows the distance between one of the launch to its closest proximity, Sea. This shows that the launch site is 0.90 Kms away from the closest distance from the sea.

Section 4

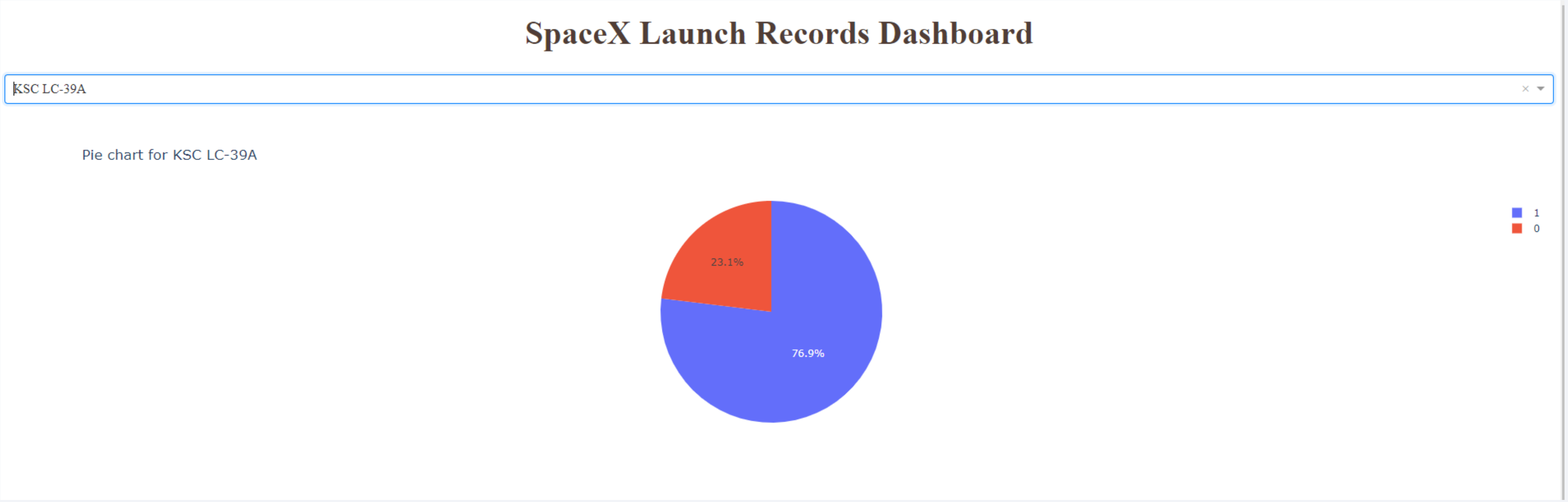
Build a Dashboard with Plotly Dash

Dashboard for all Launch site's success count



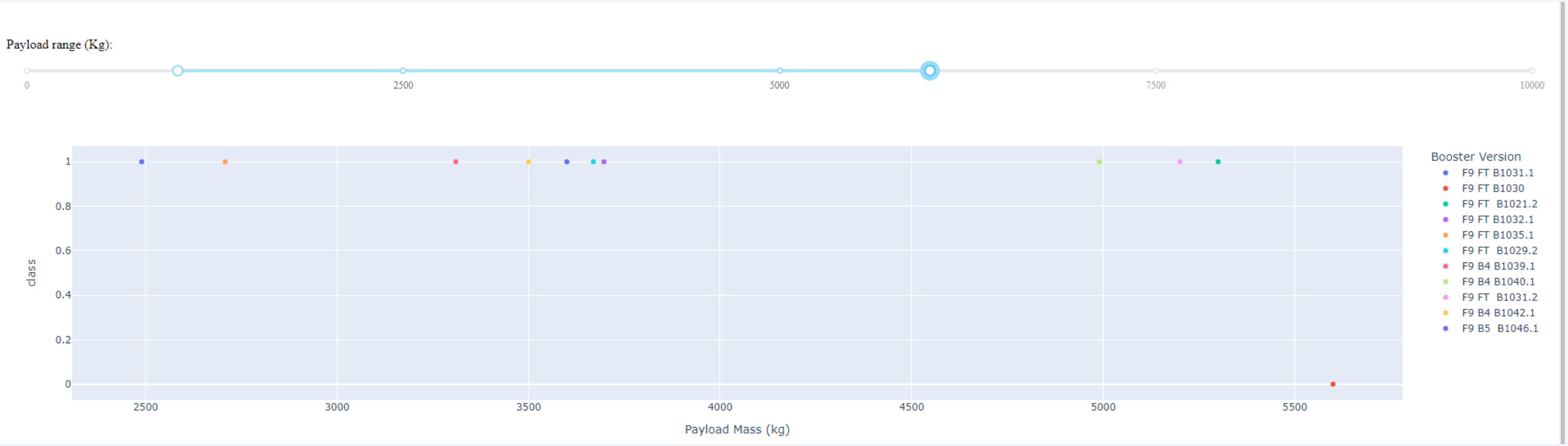
This pie chart provides success ratio for all the launch sites. We can easily see that the launch site with the highest success ratio is KSC-LC-39A with 41.7%.

Dashboard – Pie chart of best launch site



We can see the pie chart of the specific selected launch site (with the successful landing outcome deduced from the previous slide). It shows that 76.9% of it's landings are successful.

Dashboard – Payload/Launch Outcome comparision



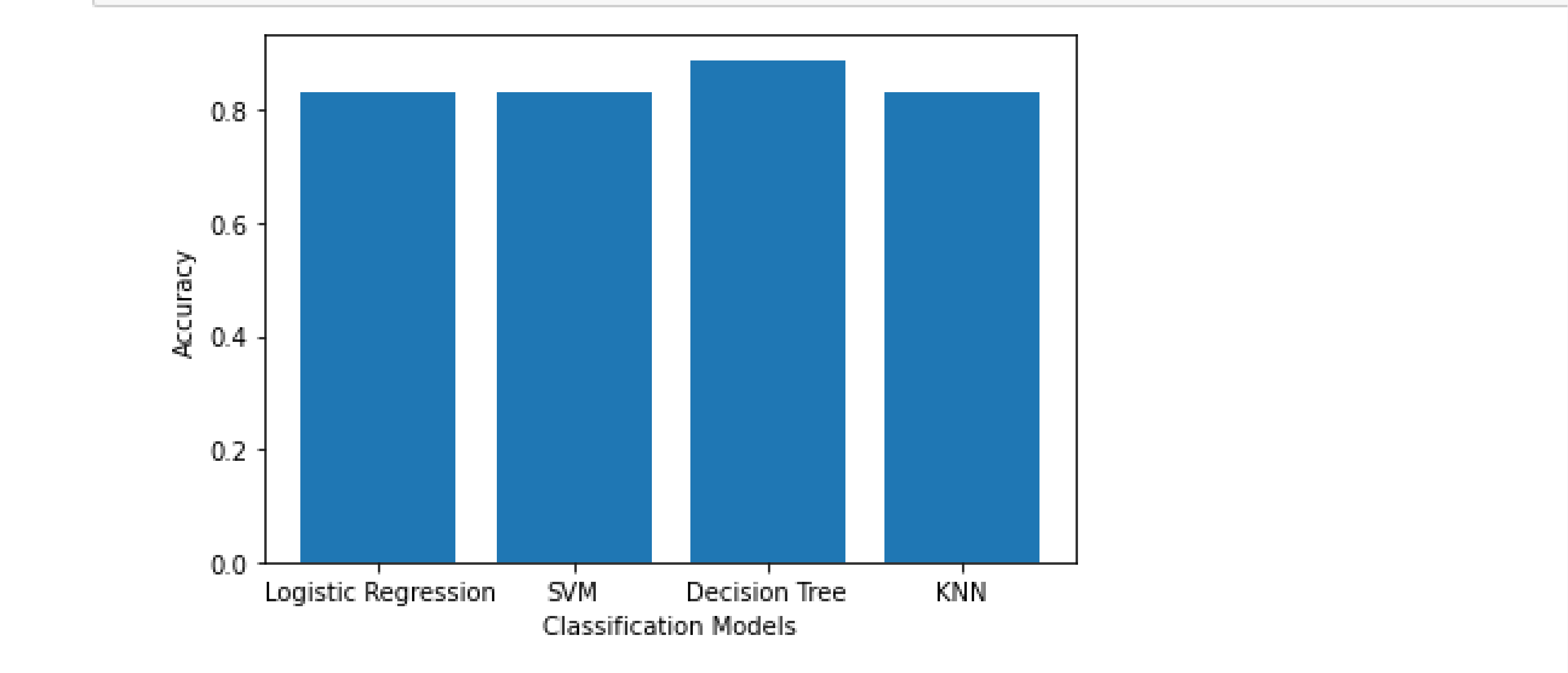
We see from adjusting the payload range, the success rate for the payload range is between 2000 to 6000 Kgs for the Launch site as KSC-LC-39A. The booster version is F9 FT B1031.1 as we can see that multiple color points for this booster version is present within the given range.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

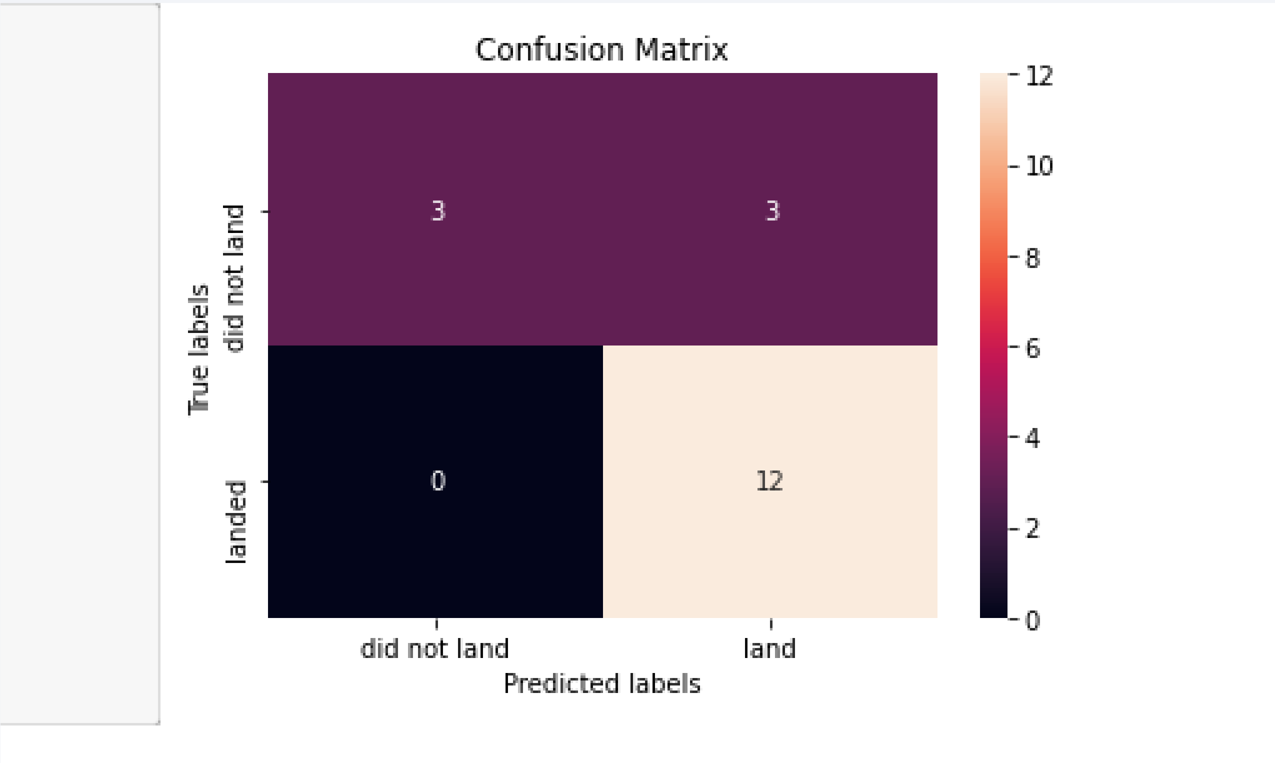
- Visualizing the built model accuracy for all built classification models, in a bar chart



We see that the decision Tree model has the highest accuracy among the other classification models, with accuracy as 88.88%

Confusion Matrix

- The confusion matrix of the best performing model (Decision Tree Classifier) with an explanation



Examining the confusion matrix, we see that decision tree model can distinguish between the different classes, by predicting the true labels.

Conclusions

- We find that the Decision tree classifier is the best model we can use for classification
- The Decision tree classifier has the accuracy of 88.88%
- Hence, we could use this model to predict the landing outcome of the next launch mission
- We can use this findings to decide the approximate cost required to decide for the next mission and use that information in the bidding.

Appendix

- GitHub URL for the repo : https://github.com/gautamvr/SpaceY/tree/master

Thank you!