Data Science Capstone project

Ayushi Dadhich

Wednesday 25 , 2021

Outline

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

Executive Summary

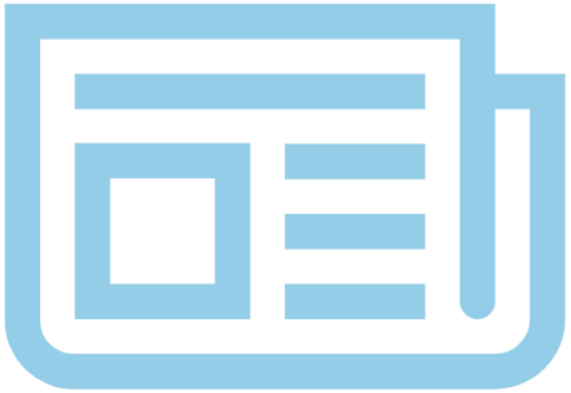
- Summary of methodologies
- This project explored the use of machine learning methodologies on predicting the outcome of landing of SpaceX's first stage. The data of spaceX launches were obtained from two sources, i.e., SpaceX REST API, and related Wikipedia pages. The collected data were converted into pandas dataframe for data wrangling to improve the quality. Missing data were inspected and addressed. Different outcomes of landing were reclassified into 1 (success) and 0 (failure). SQL queries were used to gather insights on the dataset. Data visualizations using were then carried out using Python libraries, pyplot and Seaborn, to examine the relationships of different variables. For interactive visualization, Folium was used to better understand geospatial aspect of the data, whilst Dash was used to build a web application. Data were then split into groups defined by categorical variables and four classification algorithms (Logistic Regression, SVM, Decision Tree, and kNN) were applied. These models were trained, validated, and tested.
- Summary of all results
- Overall accuracy and confusion matrix were used to compare the performance of these algorithms. Of the four classification methods, Decision Tree model, with an overall accuracy of 88.9% and F1 score of 0.917, performed best in this application.

Introduction

- ## Project background and context
- The commercial space transportation industry is booming and SpaceX being one of the most successful company currently. Major achievements of SpaceX are in the reuse of orbital-class launch vehicles. Most notable of these being the continued landings and relaunches of the first stage of Falcon 9 following a multi-year program to develop the reusable technology. This results in cost reduction in the space launch industry. However, successful landing of Falcon 9's first stage is not guaranteed. Sometimes it crashed. Other times, SpaceX sacrificed the first stage due to the mission parameters like payload, orbit, and customer.Hence, if we can determine if the first stage will land, the cost of a launch can be determined.

- ## Problems you want to find answers
- In this project, we will study the use of machine learning to predict the success of landing of SpaceX's first stage.

Methodology



- Data collection methodology:
  - SpaceX REST API ● Web scraping related Wikipedia page (Beautiful Soup)
- Perform data wrangling
  - Missing ‘PayloadMass’ data are replaced by mean value
  - Different ‘Outcomes’ are reclassified into 1 (success) and 0 (failure)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Python machine learning library: scikit-learn
  - Four (supervised) classification algorithms: Logistic Regression, SVM, Decision Tree, and k-NN
  - Hyperparameter tuning using Grid Search Cross-Validation
  - Confusion matrix is used to evaluate the accuracy of each algorithm

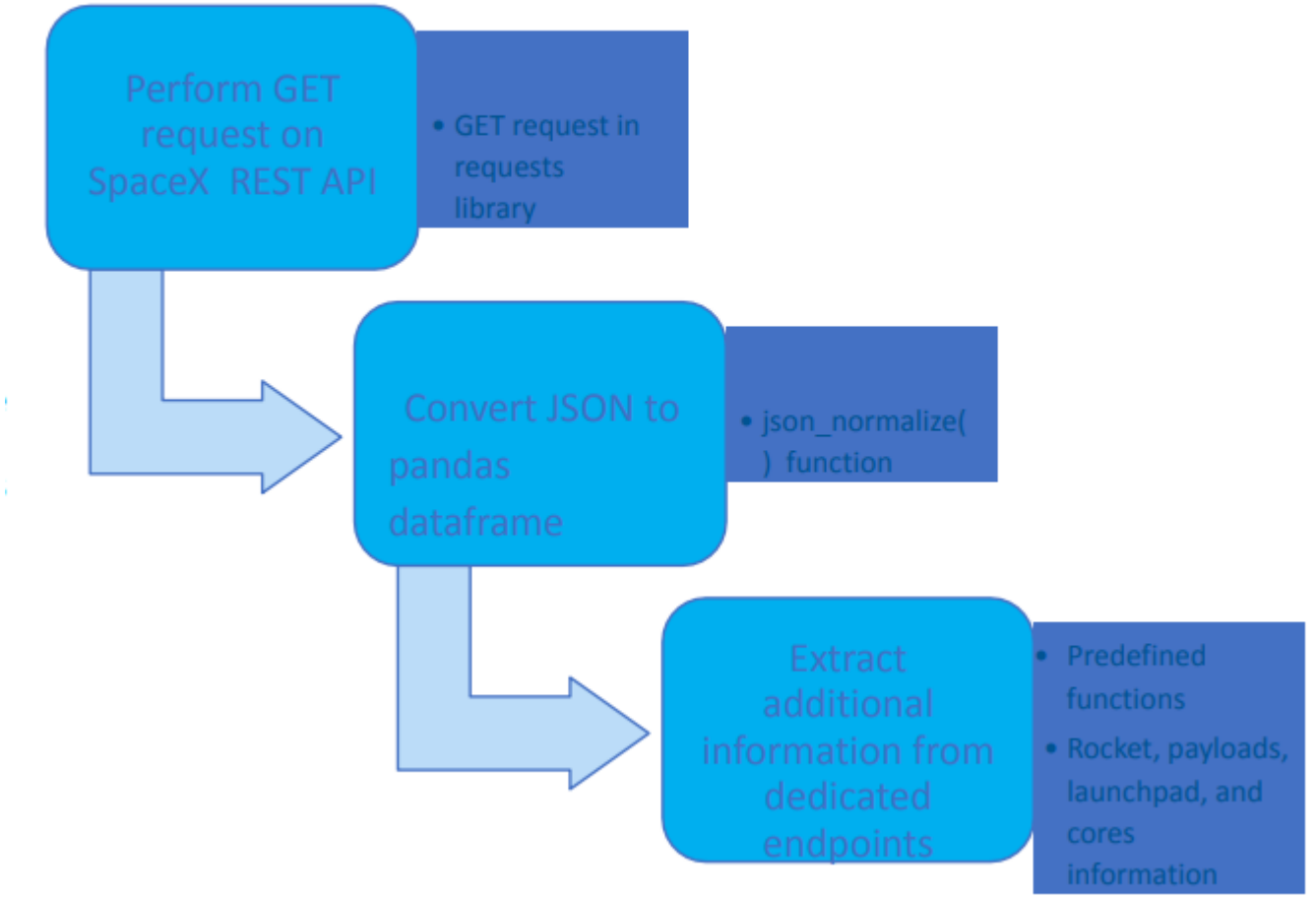
Methodology

Data collection

- The SpaceX launch data can be collected from two sources:
- • SpaceX REST API • related Wikipedia pages.
- The SpaceX REST API provides data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- These data can be accessed via different endpoints, or URLs by performing GET request using the requests library. The response will be in the form of a JSON, specifically a list of JSON objects, which can be converted into dataframe using json_normalize() function.
- The second approach is through web scraping related Wikipedia pages be using the Python BeautifulSoup package. SpaceX’s launch data from the HTML tables can be parsed and converted into dataframe for further visualization and analysis

Data collection – SpaceX API

- To extract data from SpaceX REST API, a GET request was made using the requests library to access its end point, or URL.
- The response in the form of a JSON was converted into pandas dataframe using json_normalize() function. Each row of this dataframe contains IDs for the specific launch.
- Using predefined function and this IDs for each launch, the information of rocket, payloads, launchpad, and cores were extracted for the dedicated endpoints.

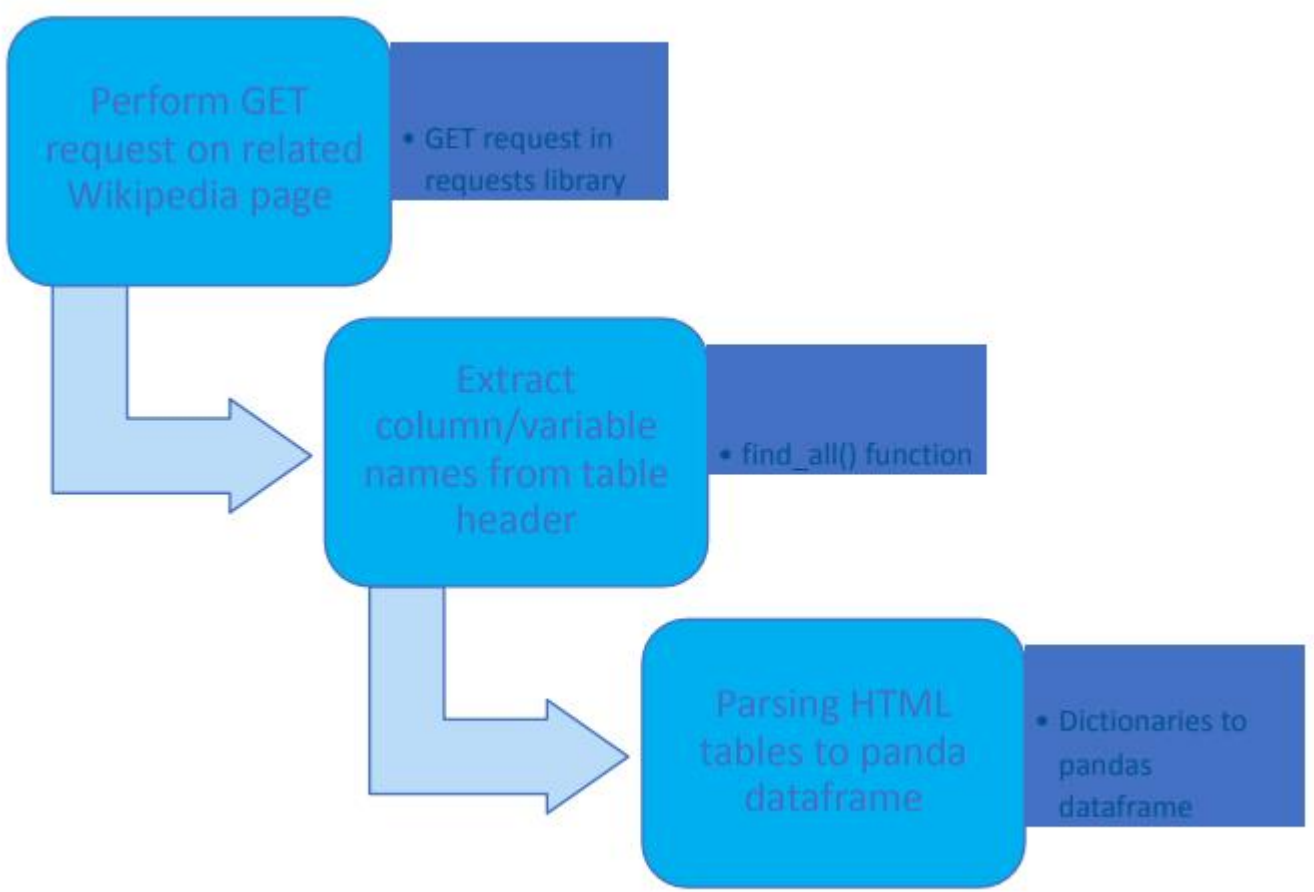


Data collection – Web scraping

To extract data from related Wikipedia page, a GET request was made using the requests library to related HTML page.

Using the Beautiful Soup package, data from the HTML tables were parsed. Column or variable names from the HTML table header were located using the find_all() function and then extracted.

Then, dictionaries of extracted data were created by parsing the relevant HTML tables. A panda dataframe was created by converting the dictionaries.



Data wrangling

- Upon inspection, there are rows with missing values found in ‘PayloadMass’ column and ‘LandingPad’ column. The ‘LandingPad’ column retained ‘None’ values to represent when landing pads were not used. As for the ‘PayloadMass’ column, the missing values were replaced with the mean of the ‘PayloadMass’ column.

- The attributes of data (e.g., launch sites, orbits) were then reviewed to gain better understanding of the dataset. For the ‘Outcome’ column, there are eight different value to indicate if the first stage has successfully landed at different sites. Since the landing site information is embedded in the ‘LandingPad’ column, the ‘Outcome’ column can be reclassified into 1 (success) and 0 (failure) without losing any information.

EDA with data visualization

- Scatter plots overlaid with the outcome of landing were used to visualize the relationships of flight number, payload mass, and launch site. These plots allow trends and pattern to be observed.
- As the different orbit type are involved, the landing success rate of each orbit type were compared in the form of bar chart.
- Again, scatter plots overlaid with the outcome of landing were used to investigate the relationships of orbit type, payload mass, and launch site.
- Lastly, a line plot of yearly landing success rate was plotted to visualize the trend.

EDA with SQL

- Display the names of the unique launch sites
- 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 first successful landing outcome in ground pad was achieved
- List the names of boosters which have success in drone ship and have payload mass between 4000 and 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 using subquery
- List the records which will display the month names, failure landing outcomes in drone ships, booster versions, launch sites for the months in year 2015
- Rank the count of successful landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

Build an interactive map with Folium

- Folium enables launch sites to be visualized on a map.
- Circles and markers were added to Folium map to mark all launch sites.
- Successful and failed launches for each site on the map were marked using marker clusters to simplify the map which contains many markers of the same coordinates.
- MousePosition was added to the map to get coordinate for a mouse over a point on the map. This enables coordinates of any point of interest can easily be obtained.
- Polyline can be added to display the distance of the launch site and the point of interest.

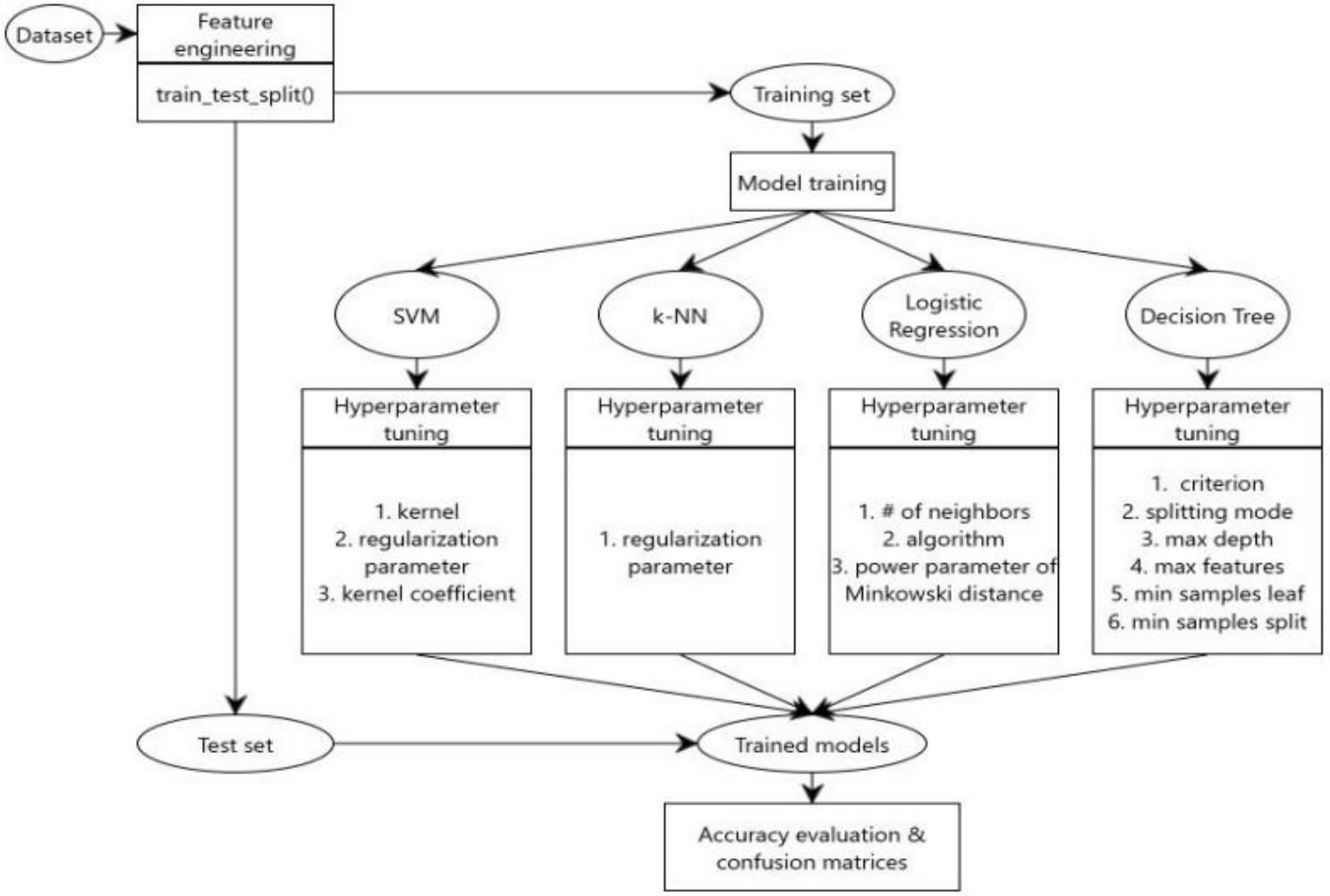
Build a Dashboard with Plotly Dash

- A pie chart and a scatter plot were added to the dashboard for visualization.
- The scatter plot enable the correlation between success of landing and the payload mass to be visualized.
- A dropdown bar was added to enable different launch site to be selected for display on the pie chart and the scatter plot.
- An ‘All Sites’ option was included so that the overall success rate to be analyzed.
- A range slider for payload mass was also added to allow customization of the scatter plot.
- These interaction features of the dashboard enable insights to be discovered more easily.

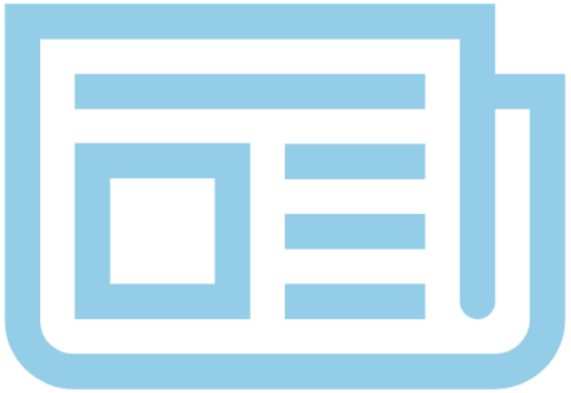
Predictive analysis (Classification)

- Feature engineering was implemented to convert nominal categorical data using one-hot encoding.
- Four classification algorithms, namely Logistic Regression, SVM, Decision Tree, and k-NN, were compared.
- The dataset was split into training set (80%) and test set (20%).
- The hyperparameters of each algorithm were tuned using Grid Search Cross-Validation.
- Overall accuracy and confusion matrix were used to evaluate the performance of each algorithm on the test set.

Predictive analysis (Classification)



Results



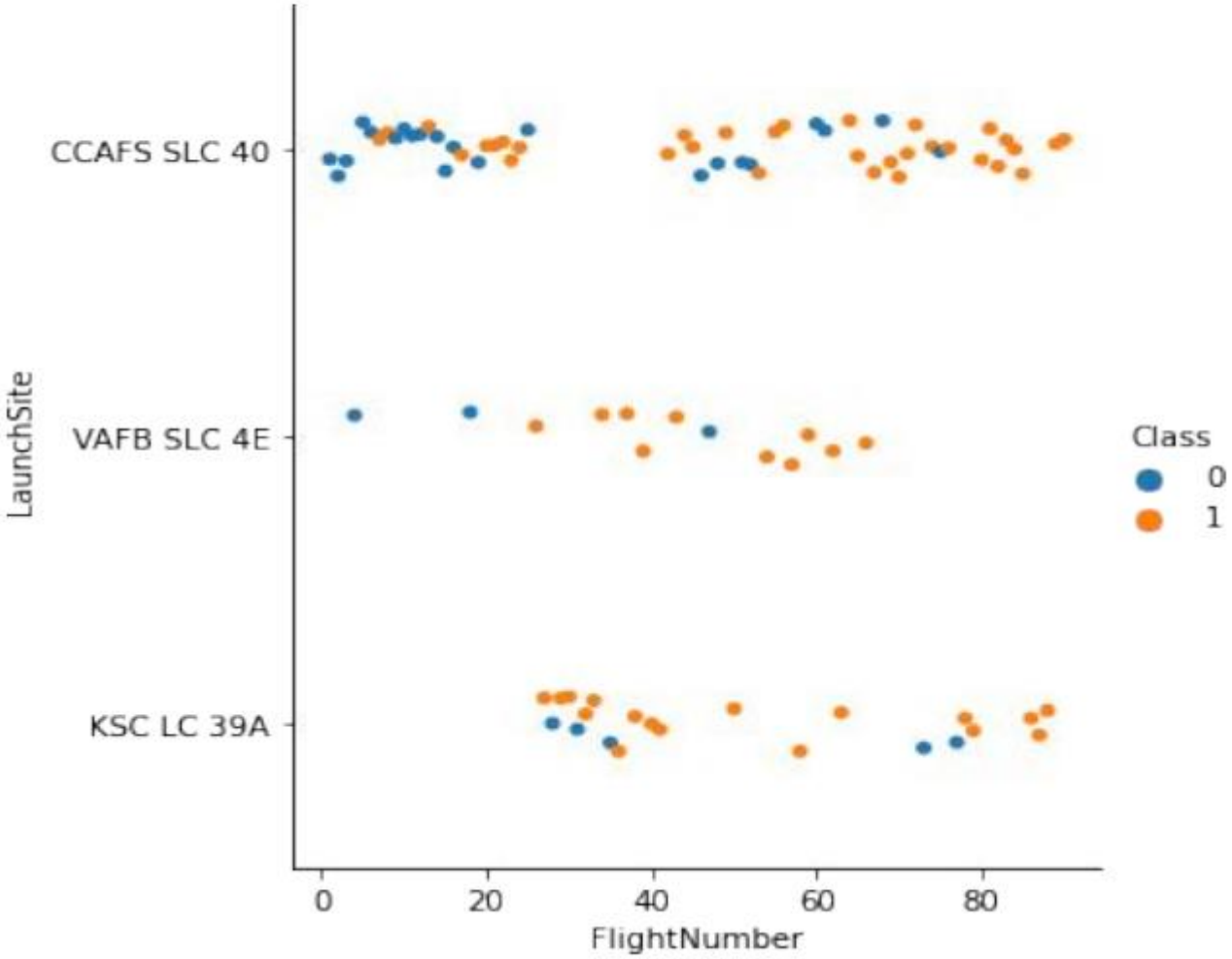
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

Flight Number vs. Launch Site

Launch site CCAFS SLC 40 (Cape Canaveral Space Launch Complex 40) has the lowest success rate. This is due high failure rate due to earlier launches (flight number less than 30).

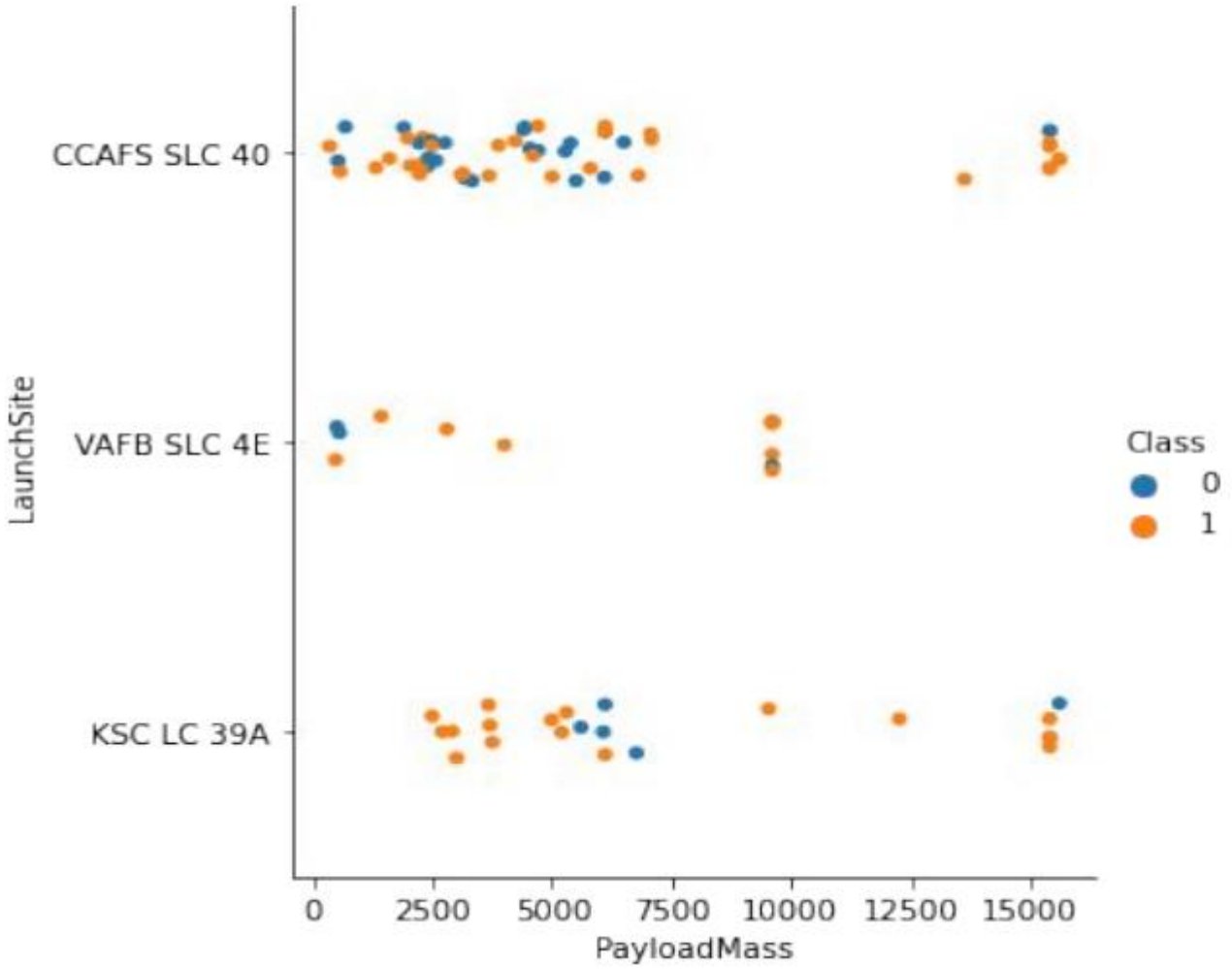
Considering only launches after flight number 30, launch site (Vandenberg Space Launch Complex 4) has the highest success rate (90%), whereas CCAFS SLC 40 and KSC LC 39A (Kennedy Space Center Launch Complex 39A) have similar success rate (75% and 78% respectively).



Payload vs. Launch Site

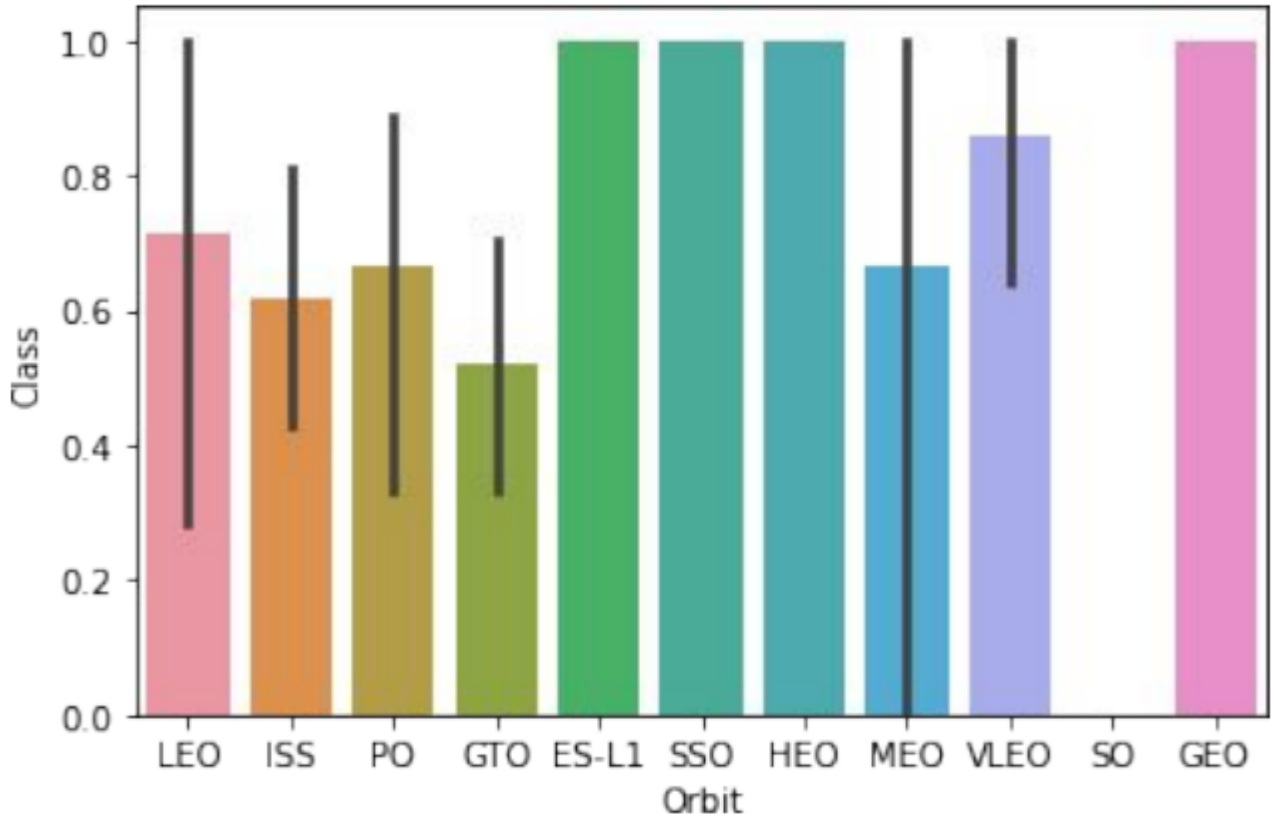
Close to three quarter (74%) of launches the payload mass of less than 7500kg.

For launches with payload mass over 7500kg, the success rate is very high (87%), and all three launch sites have similar success rate.



Success rate vs. Orbit type

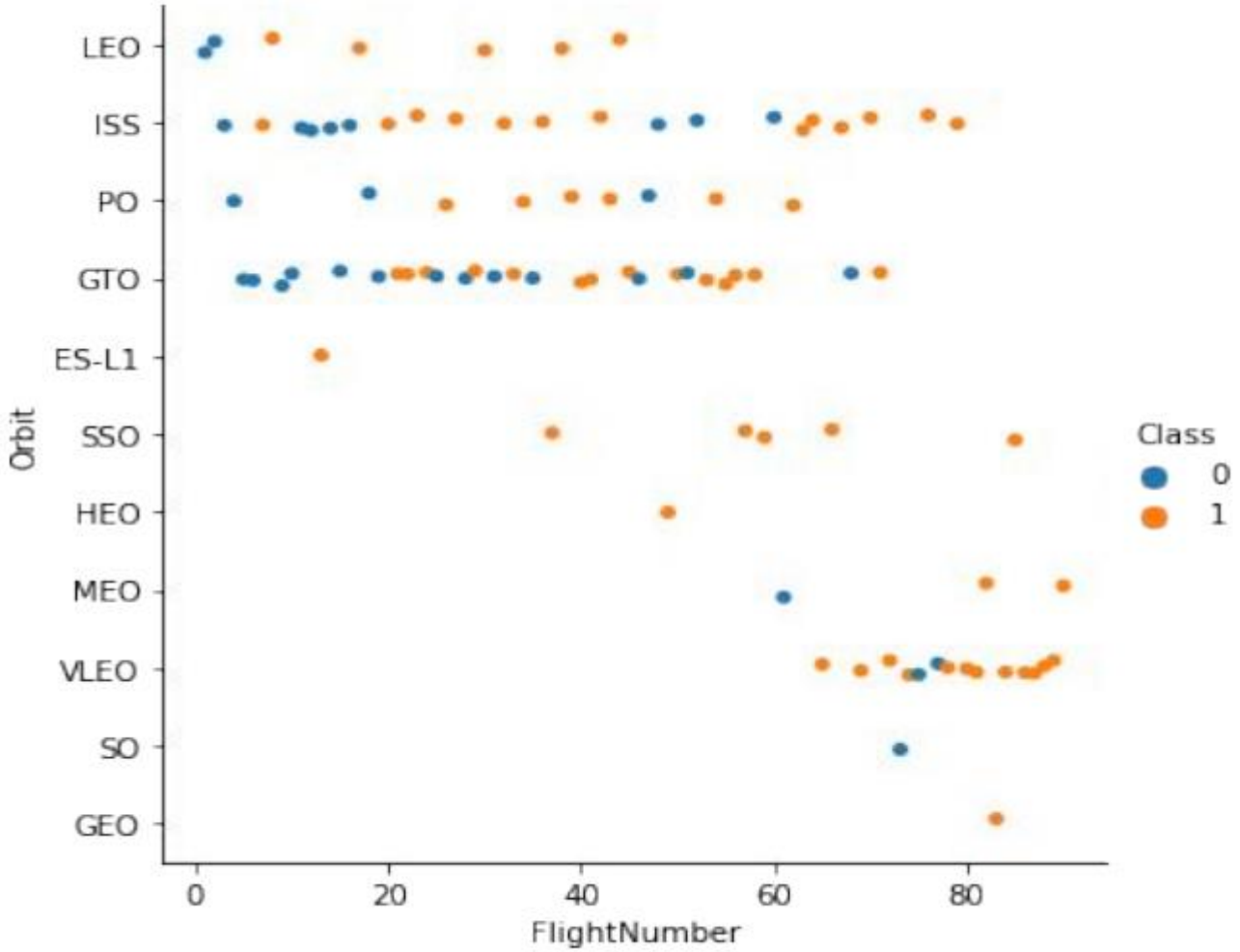
Out of the eleven orbits, four orbits (SSO, HEO, ES-L1, and GEO) have 100% success rate, while one orbit has 0% success rate.



Flight Number vs. Orbit type

Three orbits (GTO, ISS, and VLEO) make up more than 60% of the launches.

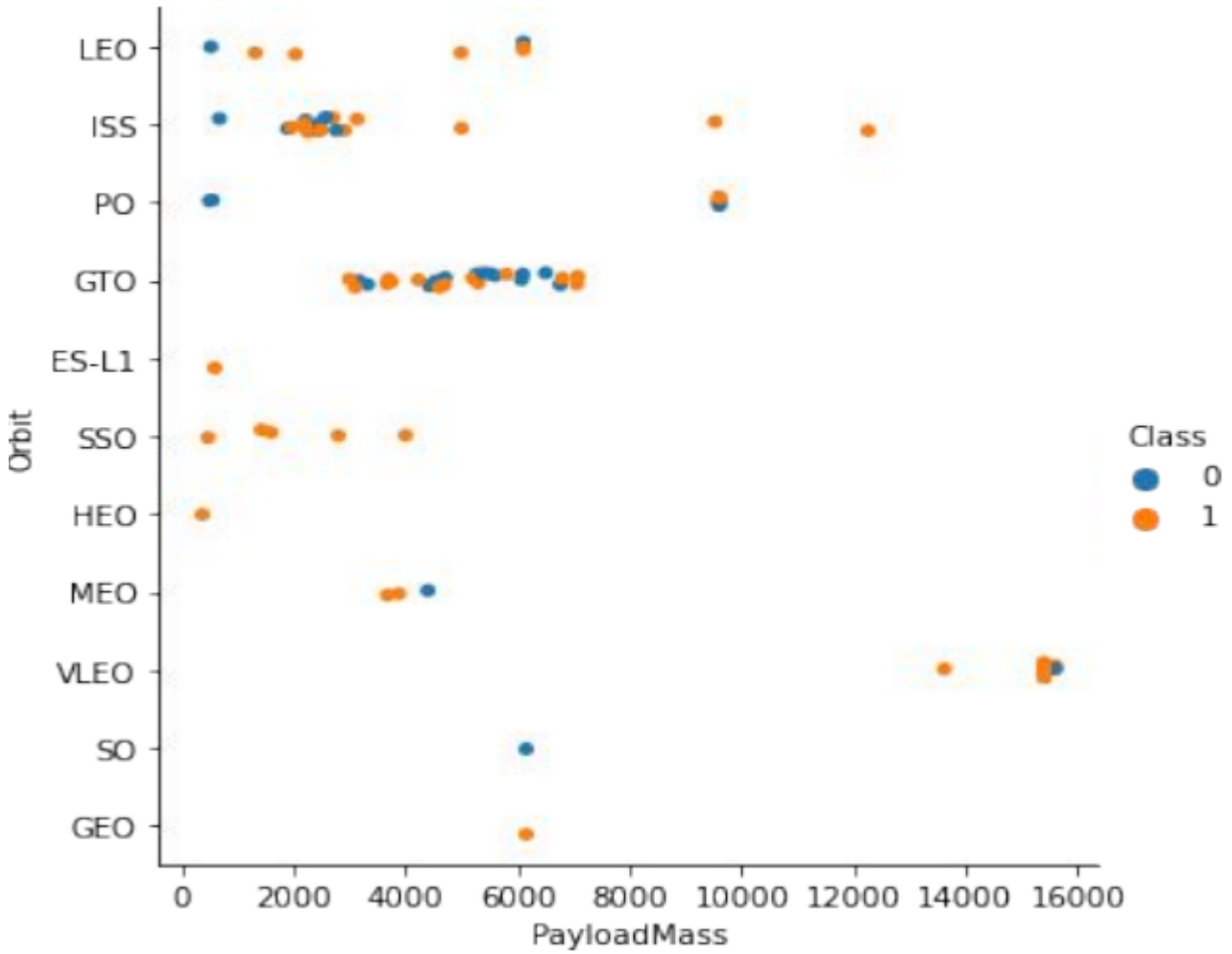
Earlier launches (flight number less than 30) were only for five of eleven orbit, with ES-L1 making up only one launch



Payload vs. Orbit type

For VLEO (very low earth orbit), the payloads for all launches weighted more than 13,000kg.

Expect for VLEO, ISS, and PO, all the launches had a payload mass of less than 7,500kg.



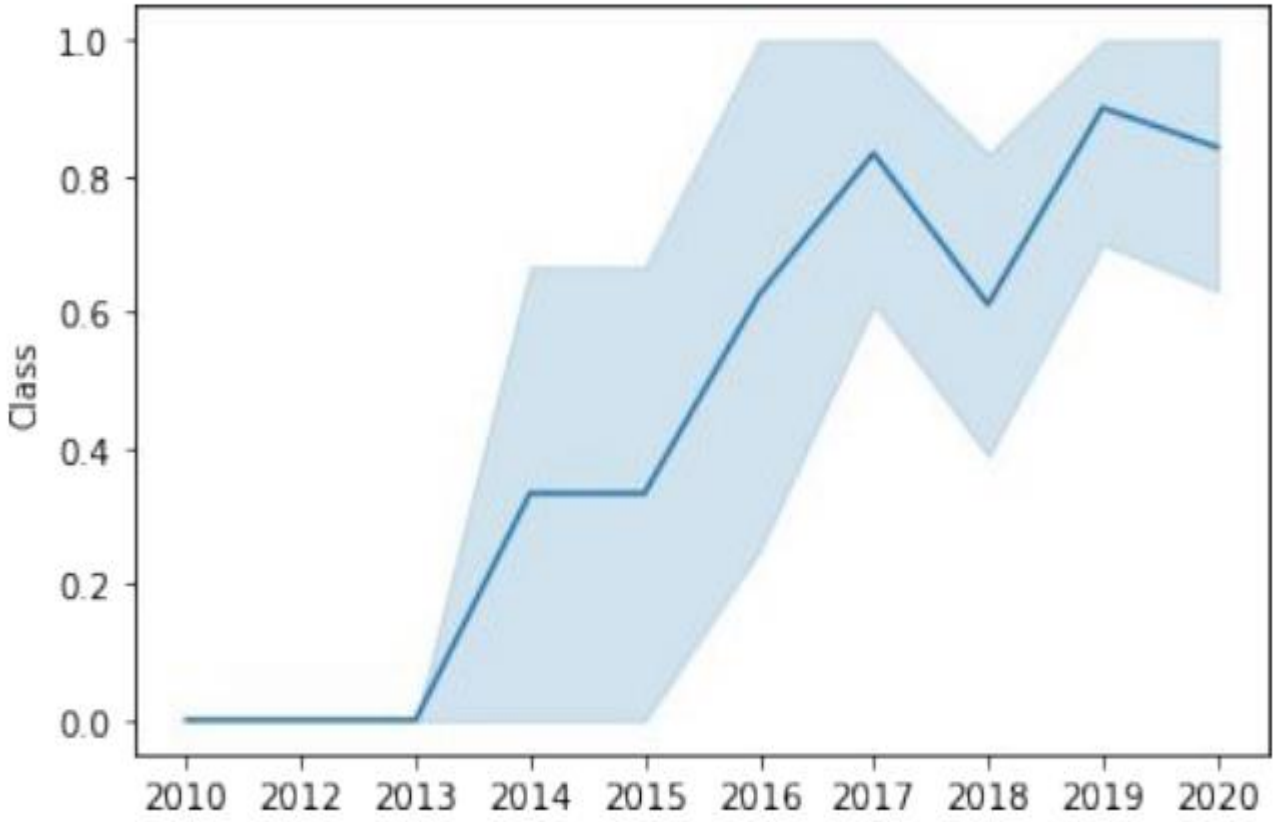
Launch success yearly trend

The year-to-year success rate showed an upward trend from year 2013 to year 2017.

A dip in success rate was observed in year 2018.

The success rate for year 2019 went up after year 2018 and even surpassed the success rate of year 2017.

The success rate in year 202 experienced a slight drop.



EDA with SQL

All launch site names

Query: select distinct(LAUNCH_SITE) from SPACEXTBL

Result: There are four launch sites for SpaceX's dataset

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

CCAFS LC-40 (Previous name for Cape Canaveral Space Launch Complex 40)

CCAFS SLC-40 (Cape Canaveral Space Launch Complex 40)

KSC LC-39A (Kennedy Space Center Launch Complex 39A)

VAFB SLC-4E (Vandenberg Space Launch Complex 4)

Launch site names begin with `CCA`

Query: select * from SPACEXTBL where launch_site like 'CCA%' limit 5
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

The launch records were filtered using LIKE operator and wildcard character % in the WHERE clause.

Total payload mass

Query: select sum(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS_KG from SPACEXTBL where customer = 'NASA (CRS)’ Result: 45596

The total payload mass by NASA (CRS) was obtained by applying the SUM() function to the expression.

The launch records were filtered using WHERE clause. 29 Average pay

Average payload mass by F9 v1.1

Query: select avg(PAYLOAD_MASS__KG_) as AVERAGE_PAYLOAD_MASS_KG from SPACEXTBL where booster_version like 'F9 v1.1%’ Result: 2534

The average payload mass by F9 v1.1 was obtained by applying the AVG() function to the expression.

The displayed records were filtered using LIKE operator and wildcard character % in the WHERE clause.

First successful ground landing date

Query: select date from SPACEXTBL where landing__outcome = 'Success (ground pad)’ order by date limit 1 Result: 2015-12-22

The first successful ground landing date was obtained by sorting using ORDER BY clause.

The launch records were filtered using WHERE clause and LIMIT clause. 31 Successful drone ship landing

Successful drone ship landing with payload between 4000 and 6000

Query: select distinct(booster_version) from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000 Result:

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

All the booster versions with payload mass between 4,000kg and 6,000kg that landed successfully on drone ship were obtained by sorting using DISTINCT() function.

The launch records were filtered using BETWEEN operator and “=” operator in the WHERE clause. The AND operator allows the existence of multiple conditions in the WHERE clause.

Total number of successful and failure mission outcomes

Query: select mission_outcome, count(*) as aggregate from SPACEXTBL group by mission_outcome Result:

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

The total number of successful and failure outcome was obtained by using COUNT() function and GROUP BY clause.

Out of the 101 launches, 99 launches were successful, one launch succeeded in the primary mission but left the secondary payload in a wrong orbit, and one launch ended destroy in flight.

Boosters carried maximum **payload**

Query: select distinct(booster_version) from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL) Result:

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

All the booster versions with payload mass with maximum payload mass were obtained using a subquery.

2015 launch records

Query: select monthname(DATE) as MONTH, landing__outcome, booster_version, launch_site from SPACEXTBL where year(DATE) = 2015 and landing__outcome = 'Failure (drone ship)' Result:

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The launch record with month names, failure landing outcomes in drone ship ,booster versions, and launch site for the months in year 2015 were obtained by filtering using WHERE clause.
- The AND operator allows the existence of multiple conditions in the WHERE clause.
- MONTHNAME() function was applied to the DATE column

Rank success count between 2010-06-04 and 2017-03-20

Query: select landing__outcome, count(landing__outcome) as aggregrate from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' and landing__outcome like '%Success%' group by landing__outcome Result:

landing__outcome	aggregrate
Success (drone ship)	5
Success (ground pad)	3

- The counts of successful landing were obtained by applying COUNT() function to an aggregated data.
- GROUP BY clause was applied to the group the landing outcome. • The launch records were filtered using LIKE operator and wildcard character % in the WHERE clause.
- The AND operator allows the existence of multiple conditions in the WHERE clause.

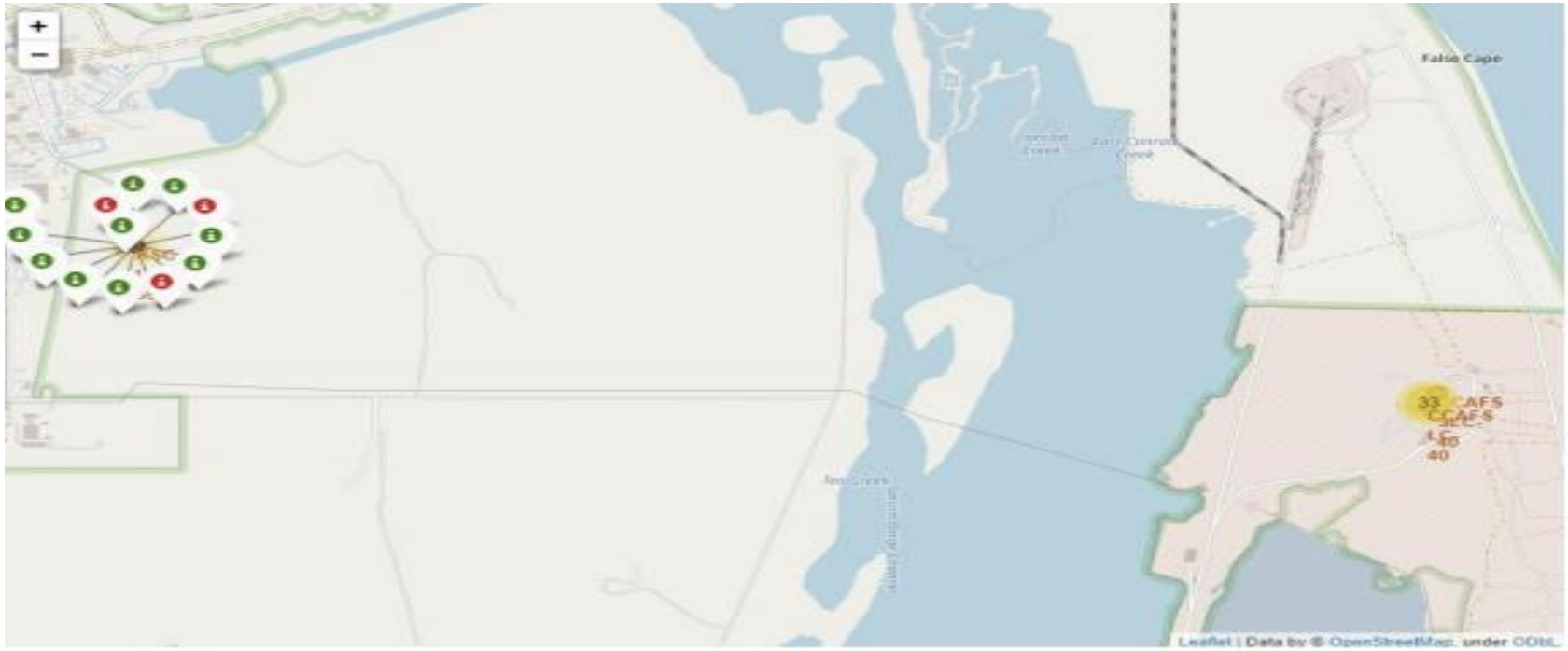
Interactive map with Folium

<Folium map screenshot 1>

There are four launch sites marked on the map. Upon closer inspection, CCAFS LC-40 and CCAFS SLC-40 are the same launch site (albeit slightly different coordinates). CCAFS SLC-40 was previously CCAFS LC-40.

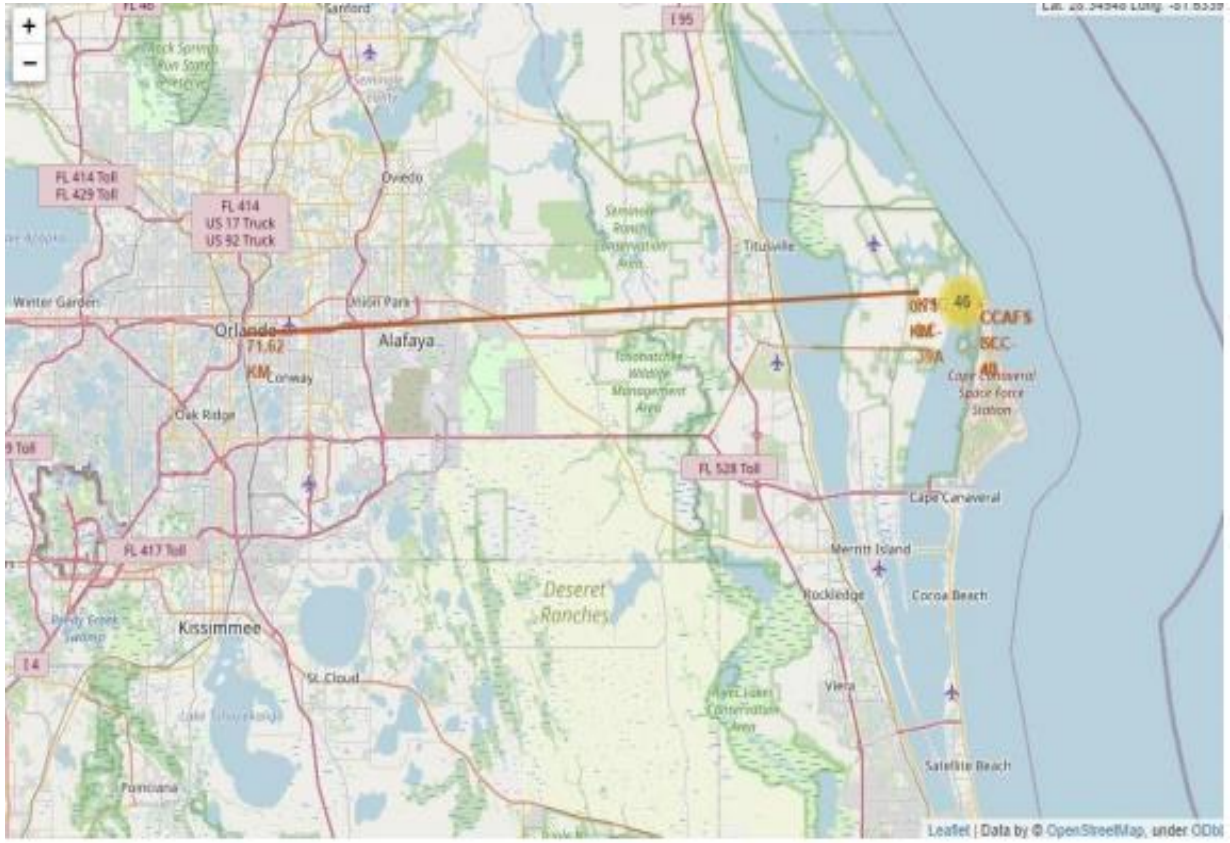
- One launch site was in California while the rest in Florida. All sites are in the southern part of USA.

<Folium map screenshot 2>



<Folium map screenshot 3>

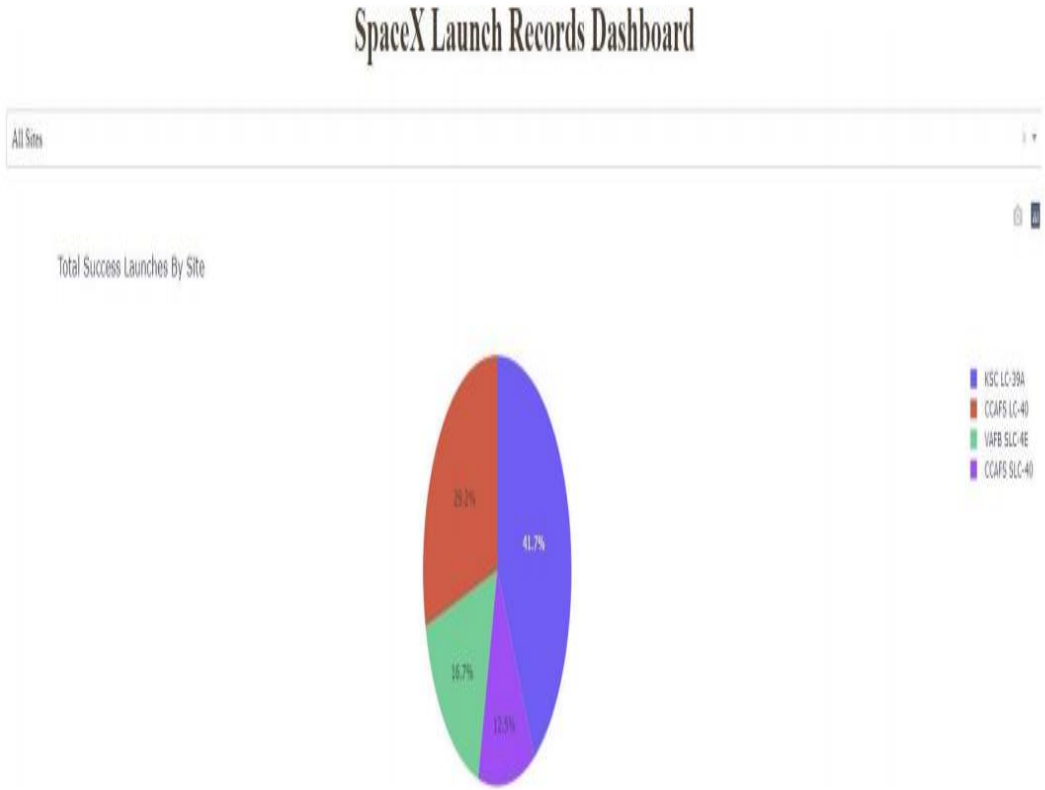
- All launch sites are in close proximity to coastline, highways, and railways.
- Also, all launch site keep a certain distance the major cities.



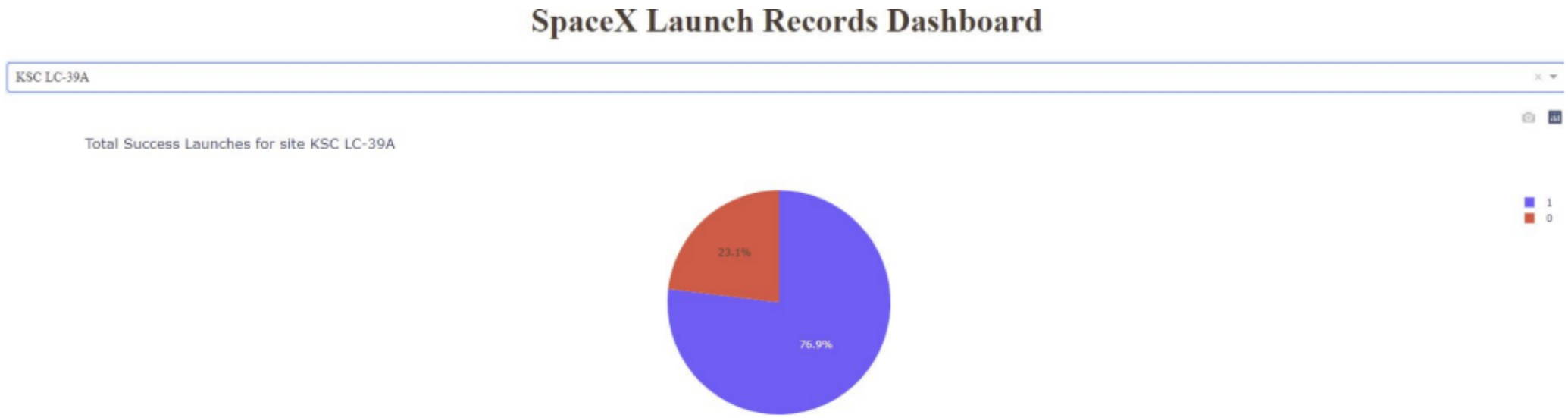
Build a Dashboard with Plotly Dash

<Dashboard screenshot 1>

- Launch site KSC LC-39A produced the highest number of successful landing outcome, while CCAF SLC-40 produced the lowest number of successful landing outcome.
- However, as CCAF SLC-40 is previously known as CCAF LC-40, the lowest number of successful landing outcome is from VAFB SLC-4E.

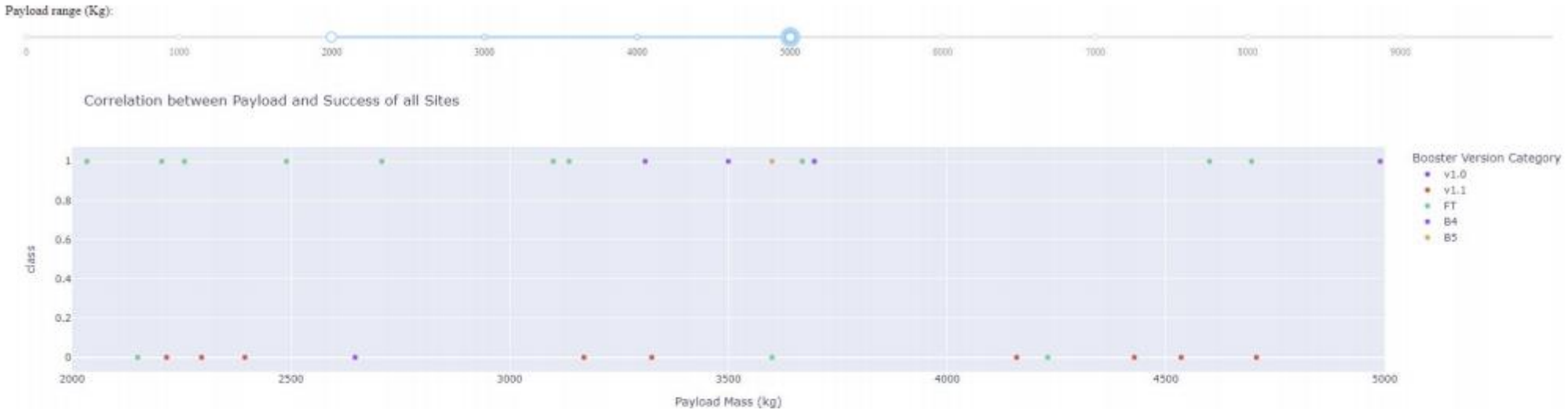


<Dashboard screenshot 2>



- KSC LC-39A has the highest success rate. Out of the thirteen launches, ten have successfully landed.

<Dashboard screenshot 3>

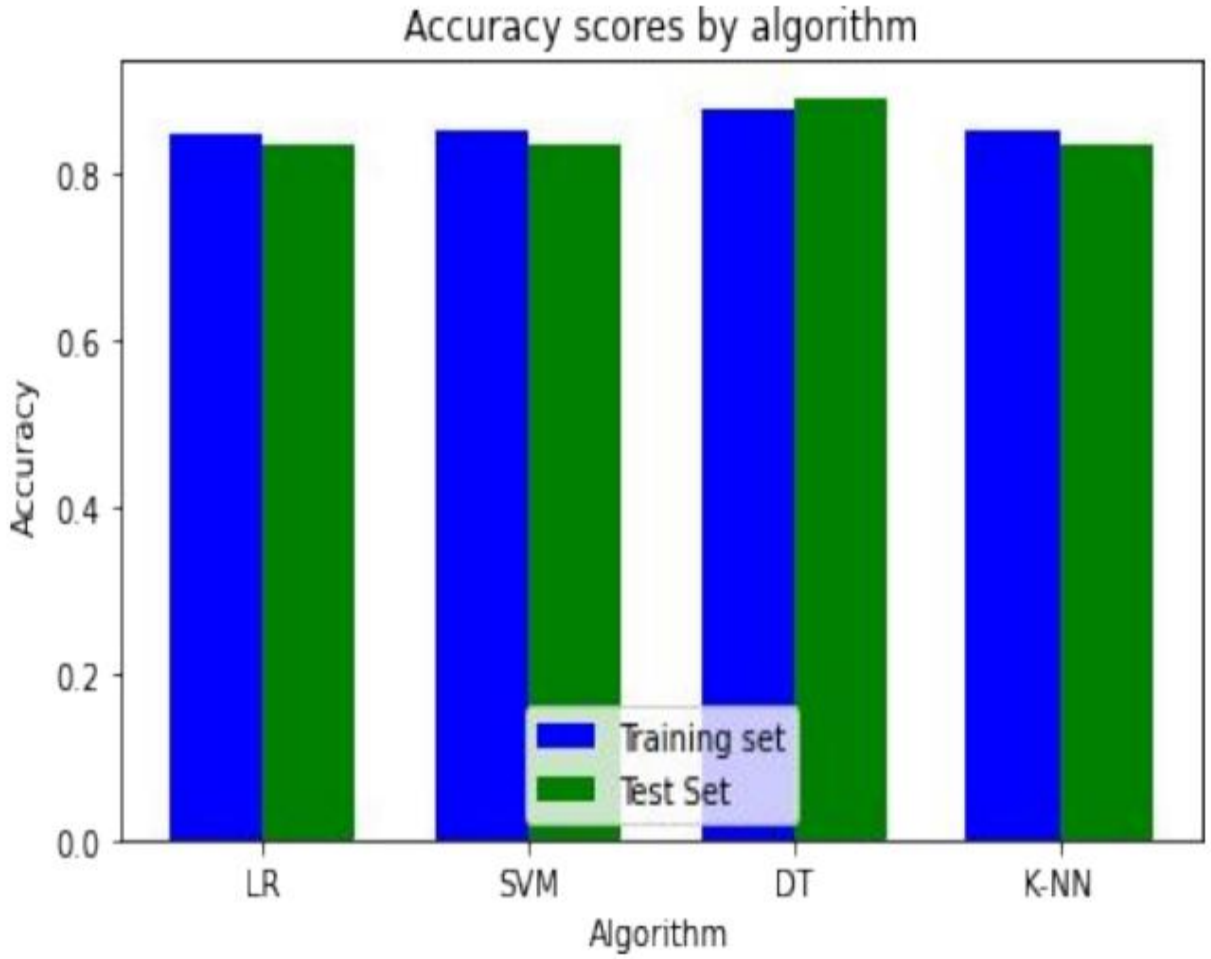


Slightly more than half of the launches had the payload mass of between 2,000kg to 5,000kg. Payload mass of this range had 16 out 29 successful landing outcome.

Predictive analysis (Classification)

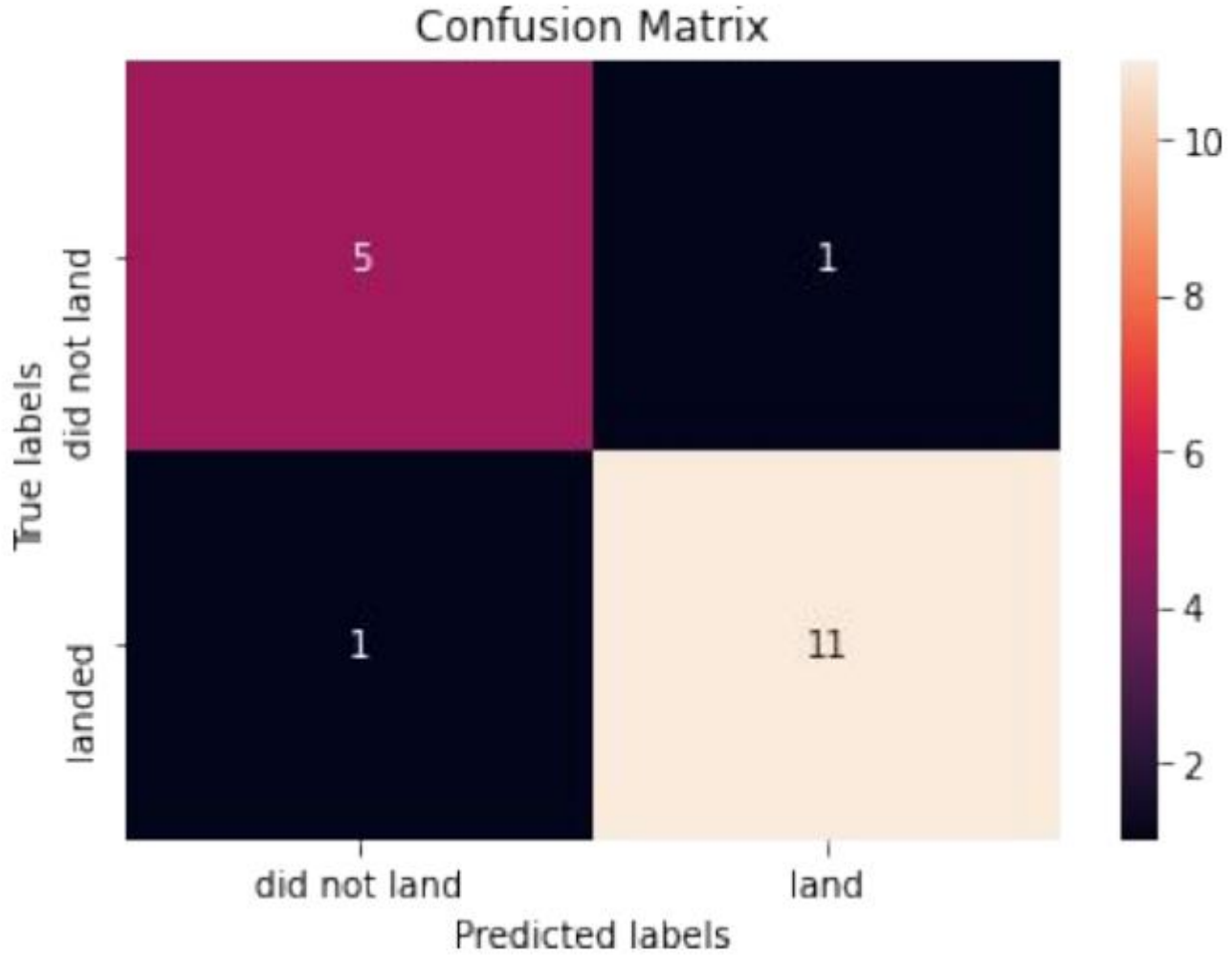
Classification Accuracy

All four classification algorithms have accuracy of more than 80% and Decision Tree algorithm has the highest prediction accuracy in both training set and test set.



Confusion Matrix

Out of the 18 test set data, the trained Decision Tree model has correctly predicted 16 outcome, i.e., 88.9%. Taking success landing as positive, the precision and recall of the model are both 0.917. With identical precision and recall, the F1 score is also 0.917. 47



CONCLUSION

- Data of SpaceX launches were obtained through SpaceX REST API and scraping Wikipedia pages.
- SQL queries and Python libraries (Seaborn and pyplot) were used to gain further insights through some basic statistical analysis and data visualization.
- Folium was used to better understand geospatial aspect of the data, whilst Dash was used to build a web application.
- By training machine learning models (without rocket science knowledge), we can predict the success of the landing of the first stage of Falcon 9 with accuracy of greater than 80%. In this project, Decision Tree had the best accuracy in both test set and training set.

APPENDIX

- API URLs:
  https://api.spacexdata.com/v4/launches/past
  https://api.spacexdata.com/v4/rockets/
  https://api.spacexdata.com/v4/launchpads/
  https://api.spacexdata.com/v4/payloads/
  https://api.spacexdata.com/v4/cores/
- Wikipedia page used for web scraping:
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Dataset for Machine Learning model training:
https://cf-courses-data.s3.us.cloud-objectstorage.appdomain.cloud/IBMDS0321EN-SkillsN etwork/datasets/dataset_part_2.csv