IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akhand Jyoti Tripathi
14 March 2023

Outline

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

Executive Summary

- Data were collected using several methods
- Machine learning models were built
- Data visualizations were created Executive Summary
- Summary of methodologies
- Summary of all results
- The optimal model was acquired
- Visualizations were great for decision making

Introduction

- Introduction

  Project background and context In this project I will work in SpaceX company and try to predict the Falcon 9 first stage. It's important to know if the rockets will land successfully or not because the failure will cost the company much resources.

- Problems that need answers

1. Which factors are behind the failure of landing?
2. Will the rockets land successfully?
3. What the accuracy of a successful landing?

Section 1

Methodology

Methodology

- Method of Data Collection
- With Rest API and Web Scrapping
- Perform data wrangling
- Data were transformed and one hot encoded to be apply later on the Machine Learning models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Discovering new patterns in the data with visualization techniques such as scatter plots
- Perform interactive visual analytics using Folium and Plotly Dash
- Dash and Folium were used to achieve this goal
- Perform predictive analysis using classification models
- Classification machine learning models were built to achieve this goal Methodology

Data Collection

Data sets were collected using the API call from several websites, I collected rocket, launchpad, payloads, and cores data from https://api.spacexdata.com/v4 website. Data Collection

1. Collecting the data with API call

2. Converting to data frame with help of JSON

3. Updating columns and rows (pre-processing)

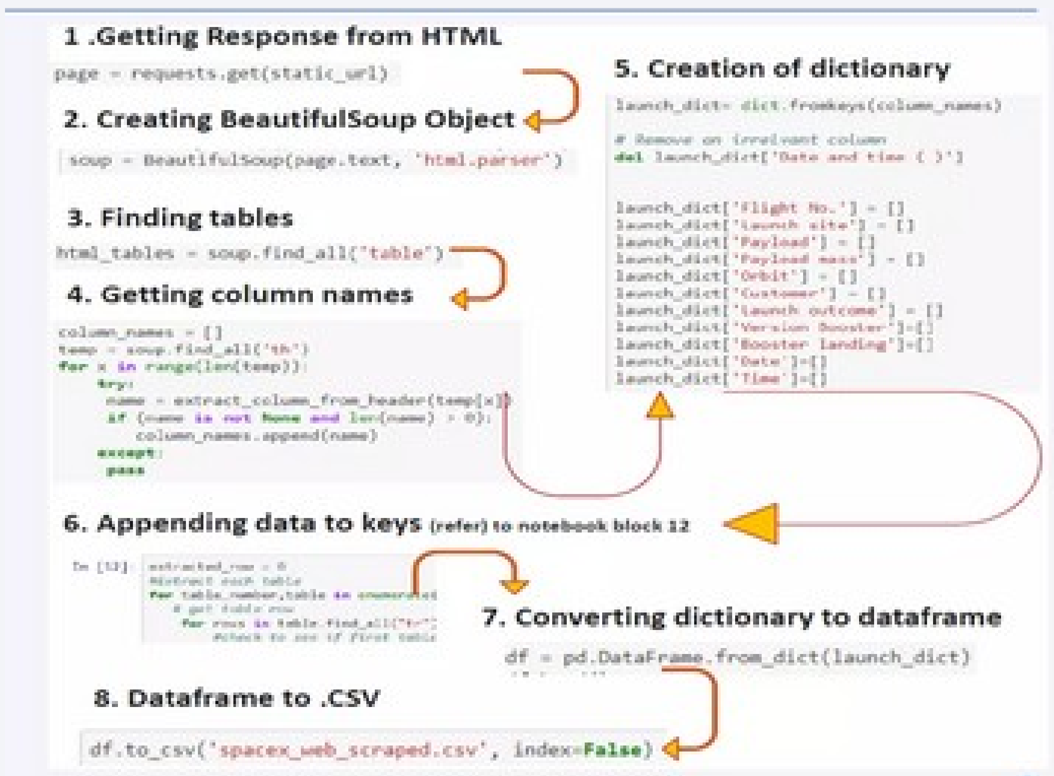
4. Filtering the data to keep only Falcon 9 launches

5. Convert the data to csv file with name 'dataset_part_1.csv'

Data Collection – SpaceX API

- Data Collection – SpaceX API 8

1. Collecting the data with API call
2. Converting to data frame with help of JSON
3. Updating columns and rows (pre-processing)
4. Filtering the data to keep only Falcon 9 launches
5. Convert the data to csv file



Data Collection - Scraping

Data Collection - Scraping

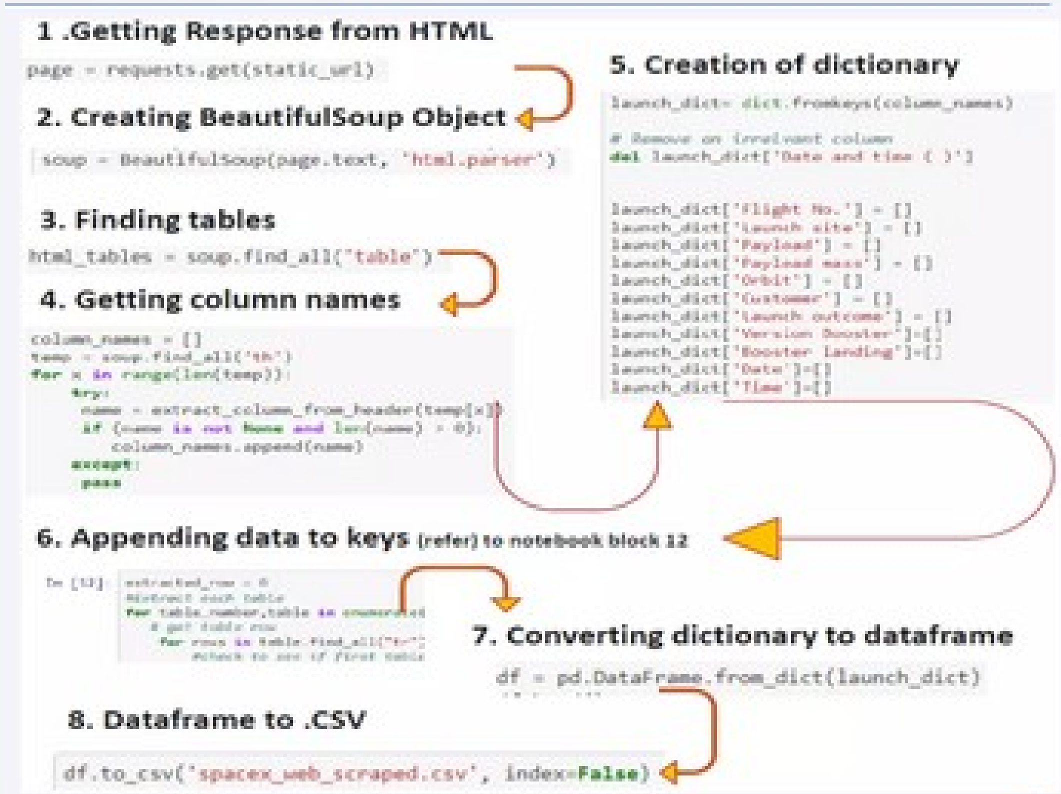
1. Creating the BeautifulSoup object

2. Getting column names

3. Creating the launch_dict

4. Converting to final data frame

5. Convert the data to csv file



Data Wrangling

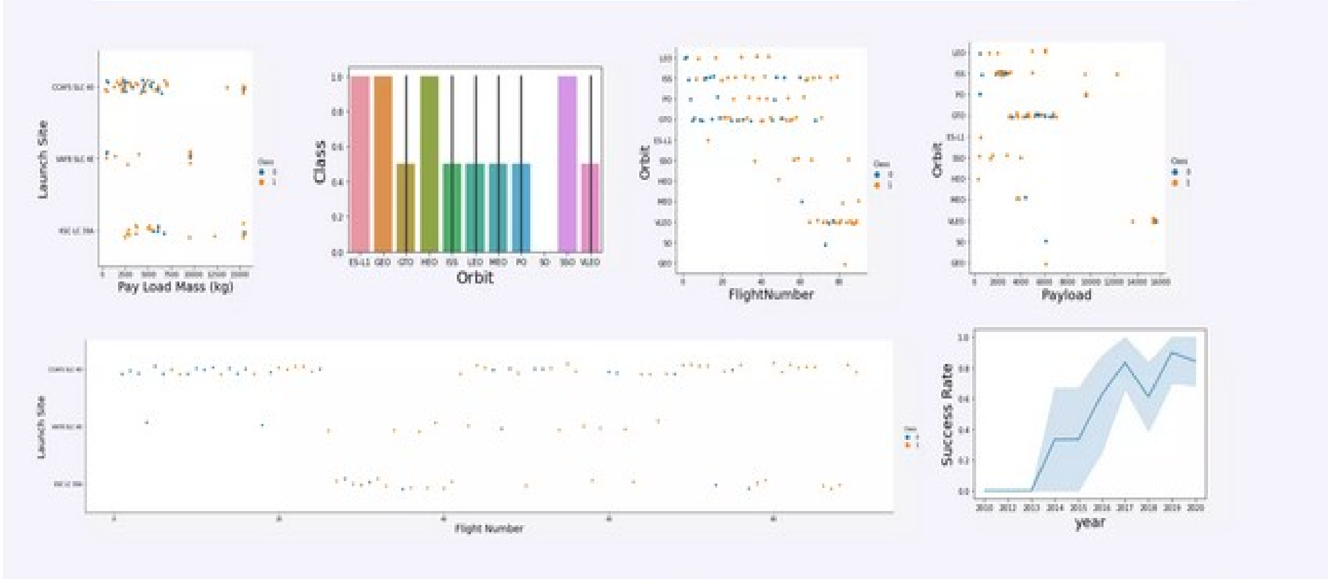
Data Wrangling

1. Loading the data set
2. Creating landing outcomes
3. Finding the bad outcomes
4. Presenting outcomes as 0 and 1
5. Determining the success outcome

EDA with Data Visualization

EDA with Data Visualization

Categorical plot between Flight number and Pay load mass (kg) Bar chart between Orbit and Success rate of each orbit Scatter plot between Orbit and Flight number Line chart between Year and Success rate



EDA with SQL

I used SQL queries to answer the following questions:

- Display the names of the unique launch sites in the space mission
- 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 the first successful landing outcome in-ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 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. Use a subquery
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015
- 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 EDA with SQL

Build an Interactive Map with Folium

folium.Marker() was used to create marks on the maps.

- folium.Circle() was used to create a circles above markers on the map.
- folium.Icon() was used to create an icon on the map.
- folium.PolyLine() was used to create polynomial line between the points.
- folium.plugins.AntPath() was used to create animated line between the points.
- markerCluster() was used to simplify the maps which contain several markers with identical coordination.

Build an Interactive Map with Folium

Build a Dashboard with Plotly Dash

Build a Dashboard with Plotly Dash

- Dash and html components were used as they are the most important thing and almost everything depends on them, such as graphs, tables, dropdowns, etc.
- Pandas was used to simplifying the work by creating dataframe.
- Plotly was used to plot the graphs.
- Pie chart and scatter chart were used to for plotting purposes.
- Rangeslider was used for payload mass range selection.
- Dropdown was used for launch sites.

Predictive Analysis (Classification)

Predictive Analysis (Classification)

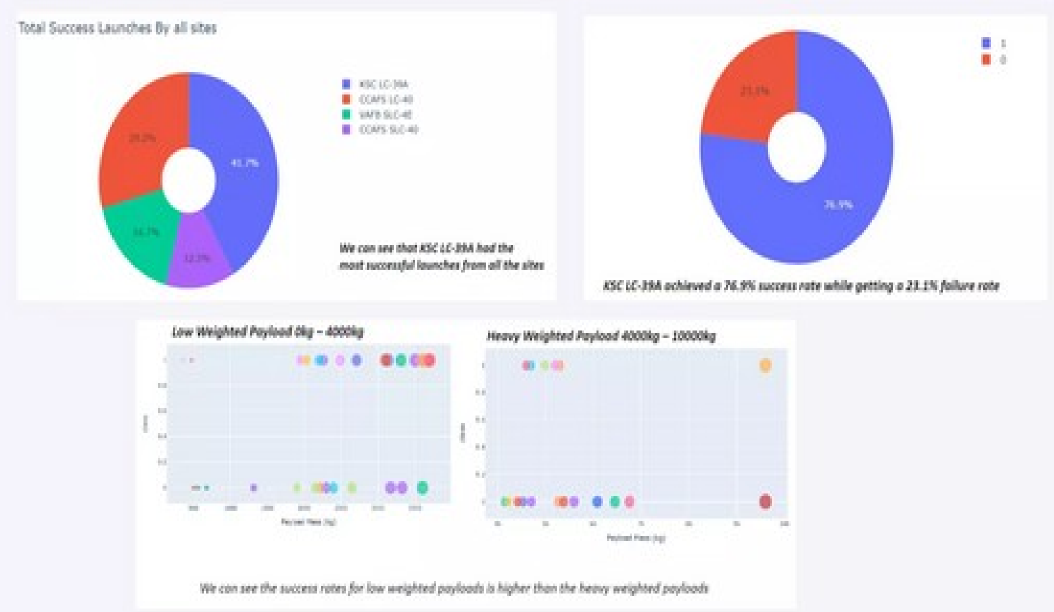
1. Building the model
2. Evaluating the model
3. Finding the optimal model
4. Create column for the class Standardize the data Split the data info train and test sets
5. Build GridSearchCV model and fit the data
6. Find the best hyperparameters for the models
7. Find the best model with highest accuracy
8. Confirm the optimal model
9. Calculating the accuracies
10. Calculating the confusion matrixes
11. Plot the results

Results

Exploratory data analysis results

- Interactive analytics demo in screenshots

- Predictive analysis results



Section 2

Insights drawn from EDA

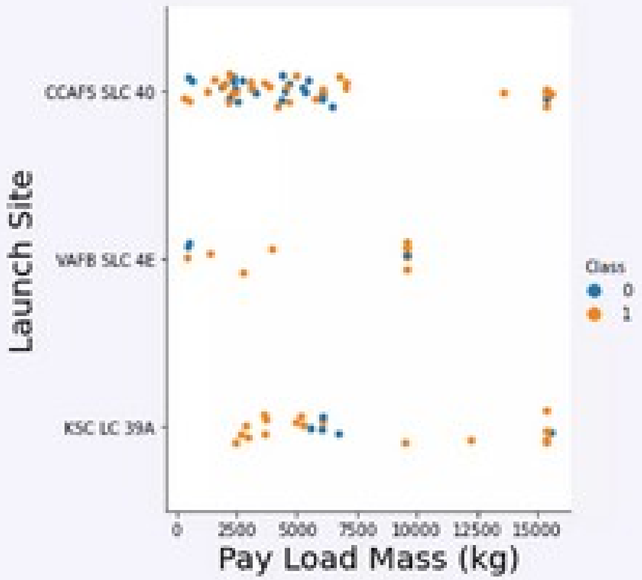
Flight Number vs. Launch Site

Payload vs. Launch Site With the increase of Pay load Mass, the success rate is increasing as well in the launch sites

- Launches from the site of CCAFS SLC 40 are significantly higher than launches form other sites.

Payload vs. Launch Site

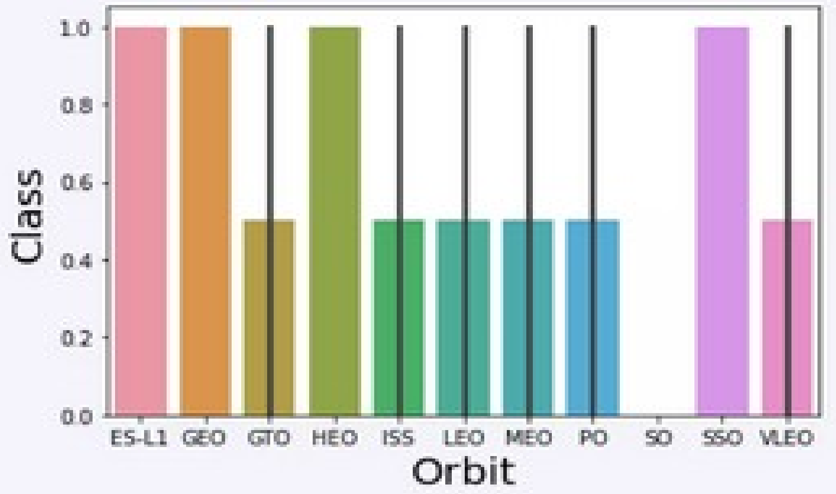
Success Rate vs. Orbit Type ES-L1, GEO, HEO, and SSO have a success rate of 100% SO has a success rate of 0%



- The majority of IPay Loads with lower Mass have been launched from CCAFS SLC 40.

Success Rate vs. Orbit Type

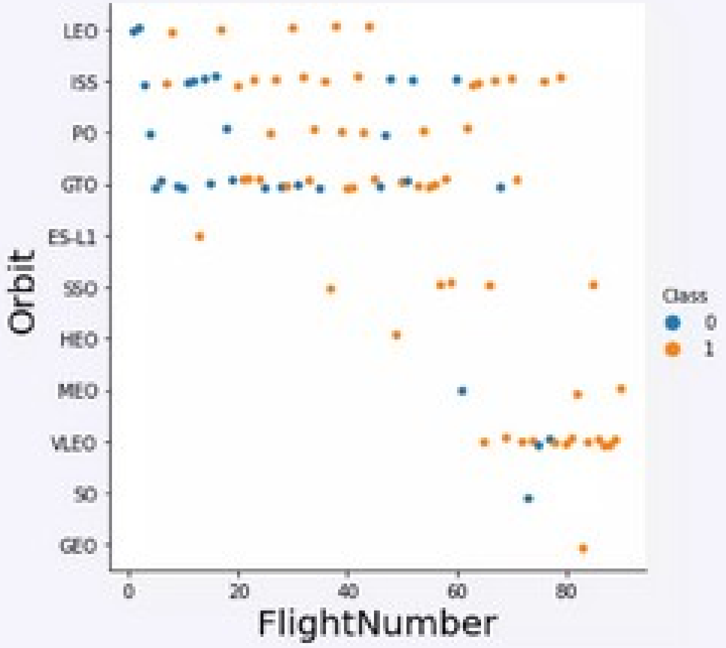
Flight Number vs. Orbit Type It’s hard to tell anything here, but we can say there is no actual relationship between flight number and GTO.



- The orbit types of ES-L1, GEO, HEO, SSO are among the highest success rate.

Flight Number vs. Orbit Type

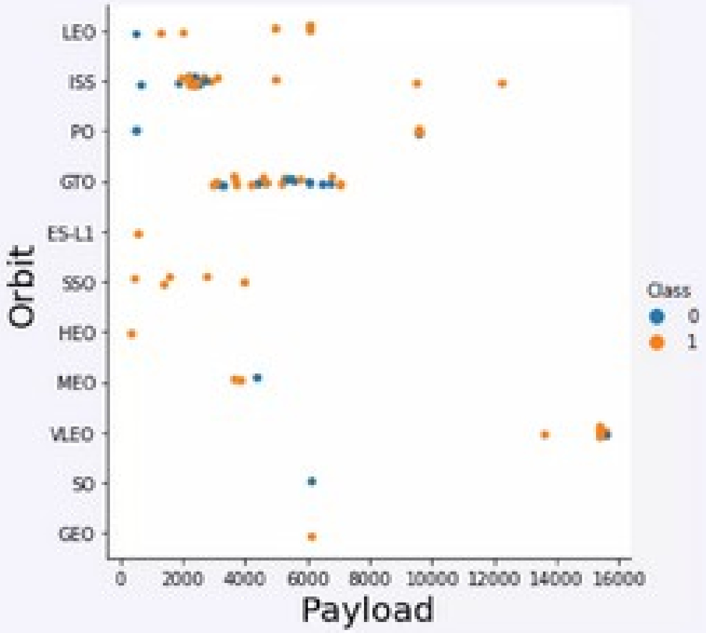
Payload vs. Orbit Type First thing to see is how the Pay load Mass between 2000 and 3000 is affecting ISS. Similarly, Pay load Mass between 3000 and 7000 is affecting GTO.



- A trend can be observed of shifting to VLEO launches in recent years.

Payload vs. Orbit Type

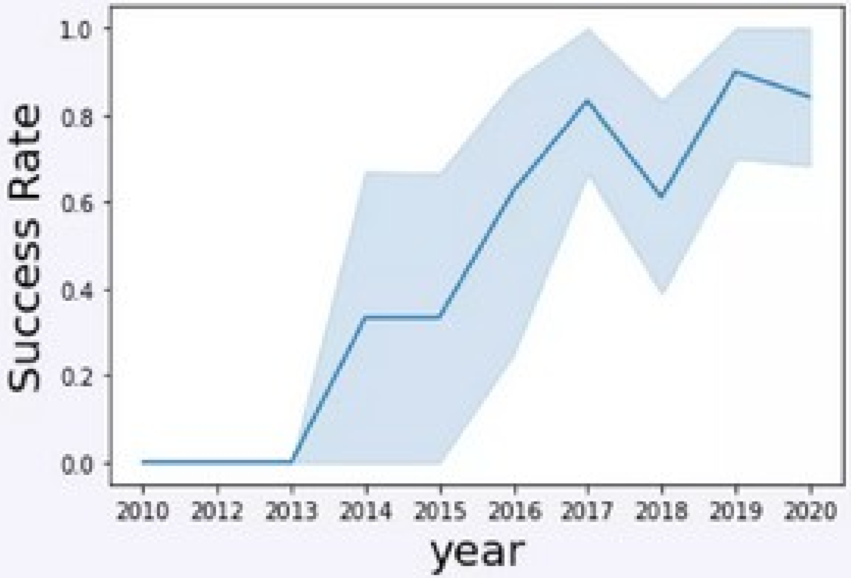
Success Yearly Trend Since the year 2013, there was a massive increase in success rate. However, it dropped little in 2018 but later it got stronger than before.



- There are strong correlation between ISS and Payload at the range around 2000, as well as between GTO and the range of 4000-8000.

Launch Success Yearly Trend

All Launch Site Names We can get the unique values by using “DISTINCT”



- Launch success rate has increased significantly since 2013 and has stablised since 2019, potentially due to advance in technology and lessons learned.

All Launch Site Names

Launch Site Names Begin with 'CCA'
We can get only 5 rows by using "LIMIT"

- %sql select distinct(LAUNCH_SITE) from SPACEXTBL

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

Launch Site Names Begin with 'CCA'

Total Payload Mass We can get the sum of all values by using "SUM"

- %sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5

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

Average Payload Mass by F9 v1.1

Average Payload Mass by F9 v1.1 We can get the average of all values by using “AVG”

- %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'

2928.400000

First Successful Ground Landing Date

First Successful Ground Landing Date We can get the first successful data by using “MIN”, because first date is same with the minimum date

- %sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Successful Drone Ship Landing with Payload between 4000 and 6000 The payload mass data was taken between 4000 and 6000 only, and the landing outcome was determined to be “success drone ship”

- %sql select BOOSTER_VERSION from SPACEXTBL where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000

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

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failure Mission Outcomes We can get the number of all the successful mission by using “COUNT” and LIKE “Success%” We can get the number of all the failure mission by using “COUNT” and LIKE “Failure%”

- %sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'

100

Boosters Carried Maximum Payload

Boosters Carried Maximum Payload We can get the maximum payload masses by using “MAX”

- %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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

2015 Launch Records We can get the months by using month(DATE) and in the WHERE function we assigned the year value to “2015”

- %sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between ‘2015-01-01' and ‘2015-12-31') order by date desc

time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
[illegible]	[illegible]	CCAFS LC-	[illegible]	[illegible]	[illegible]	SKY Perfect JSAT	[illegible]	[illegible]

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

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20 By using “ORDER” we can order the values in descending order, and with “COUNT” we can count all numbers as we did previously

- %sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc

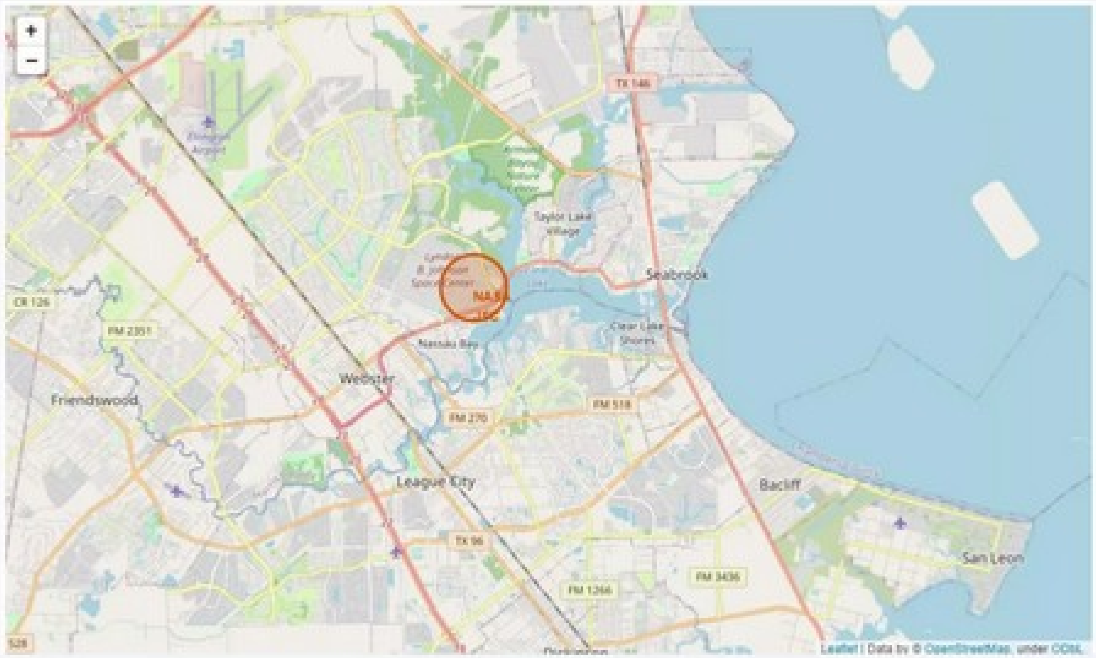
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

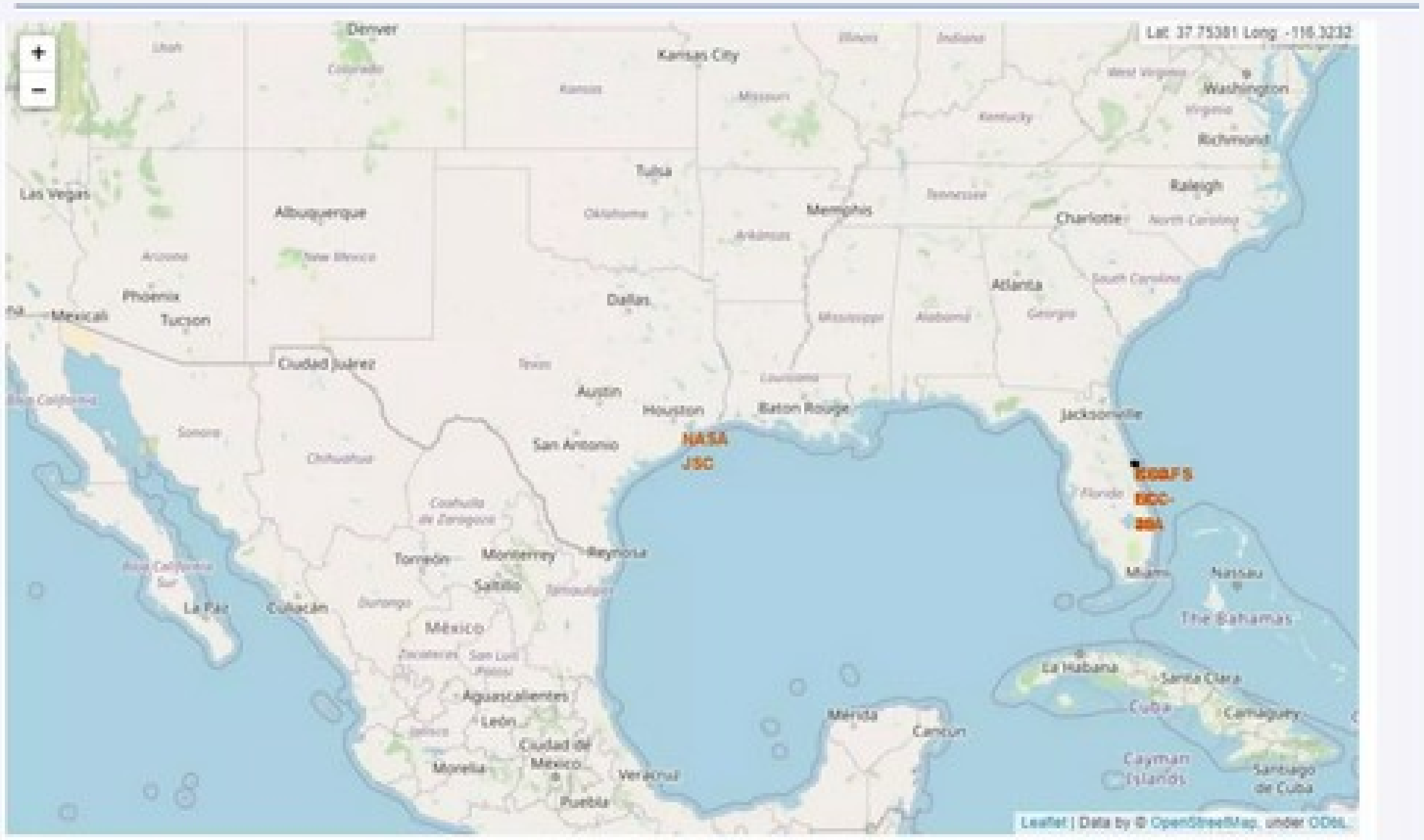
<Folium Map Screenshot 1>

Launch Sites to its Proximities All distances from launch sites to its proximities, they weren’t far from railway tracks.



<Folium Map Screenshot 2>

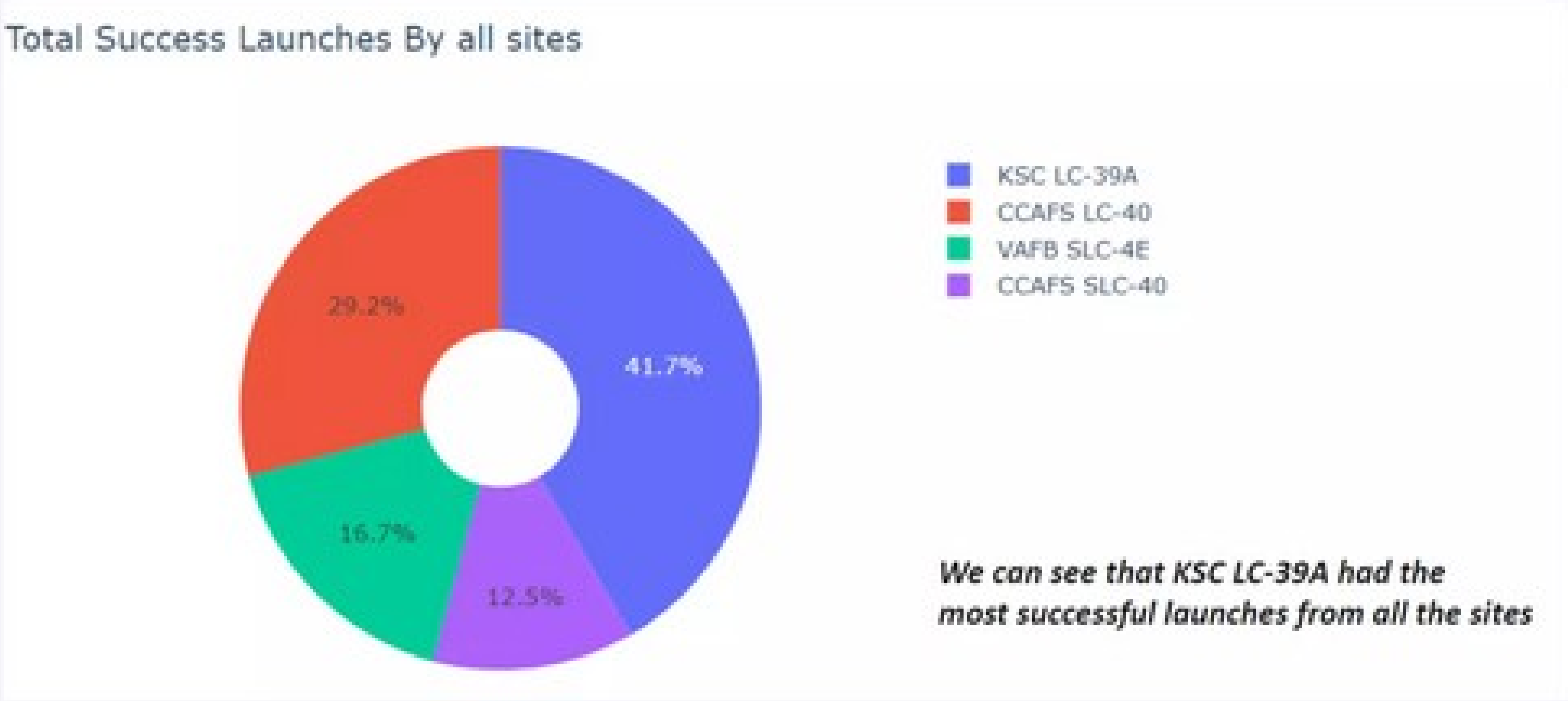
<Folium Map Screenshot 3>



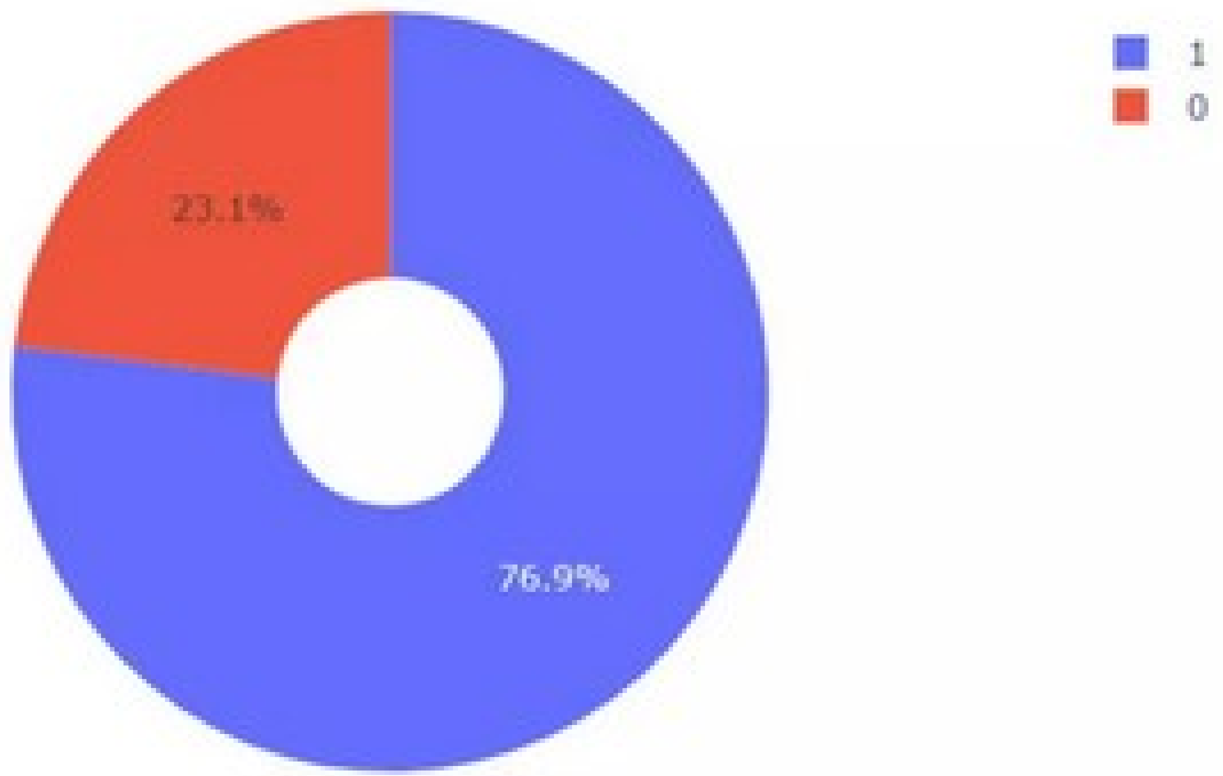
Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

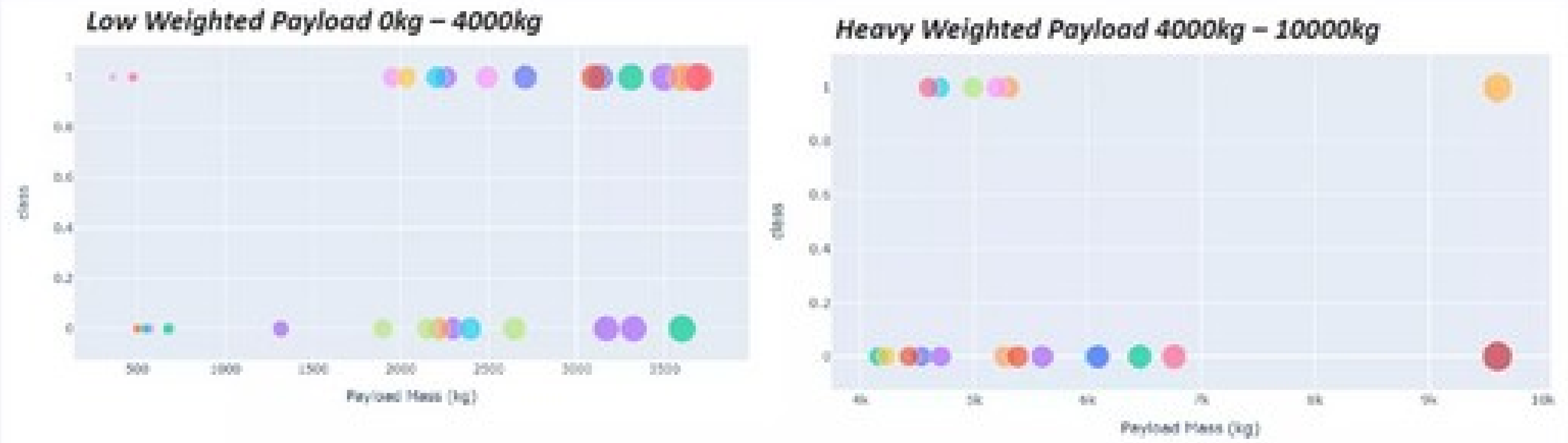


<Dashboard Screenshot 2>



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

<Dashboard Screenshot 3>



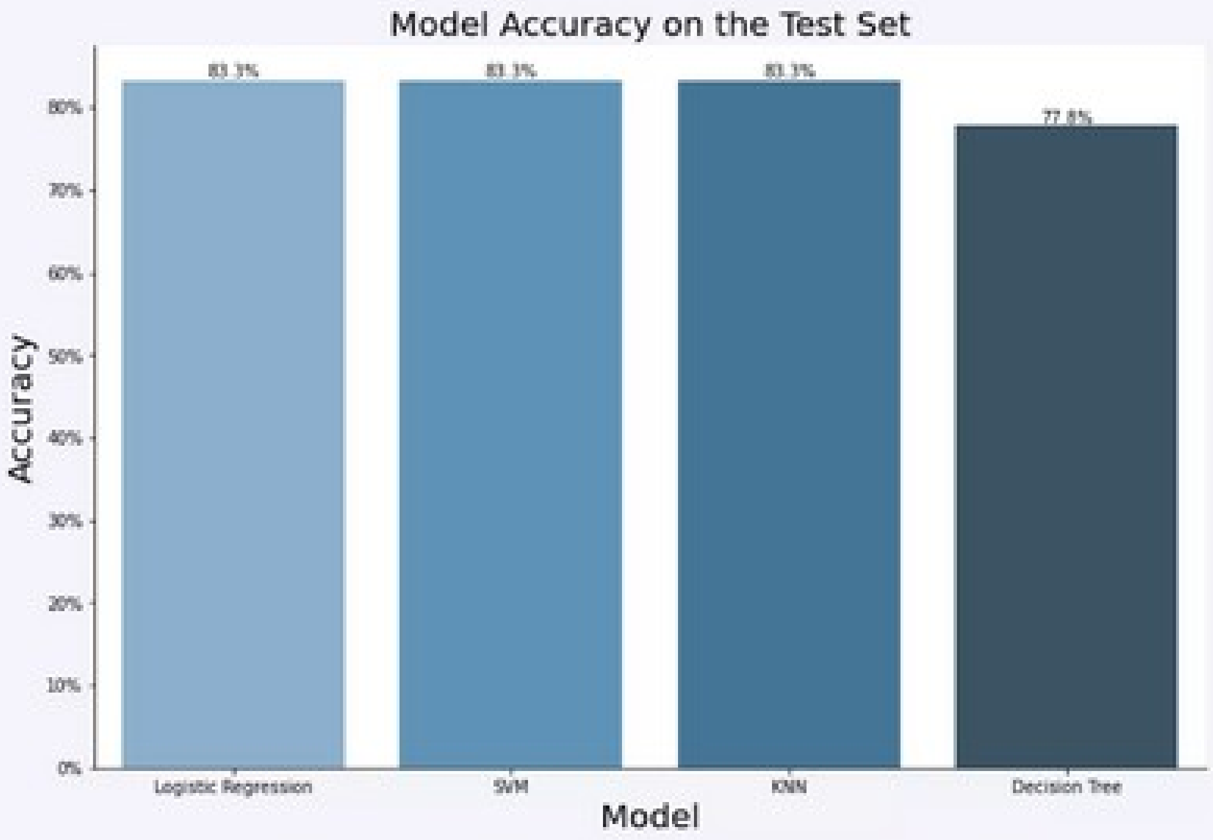
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

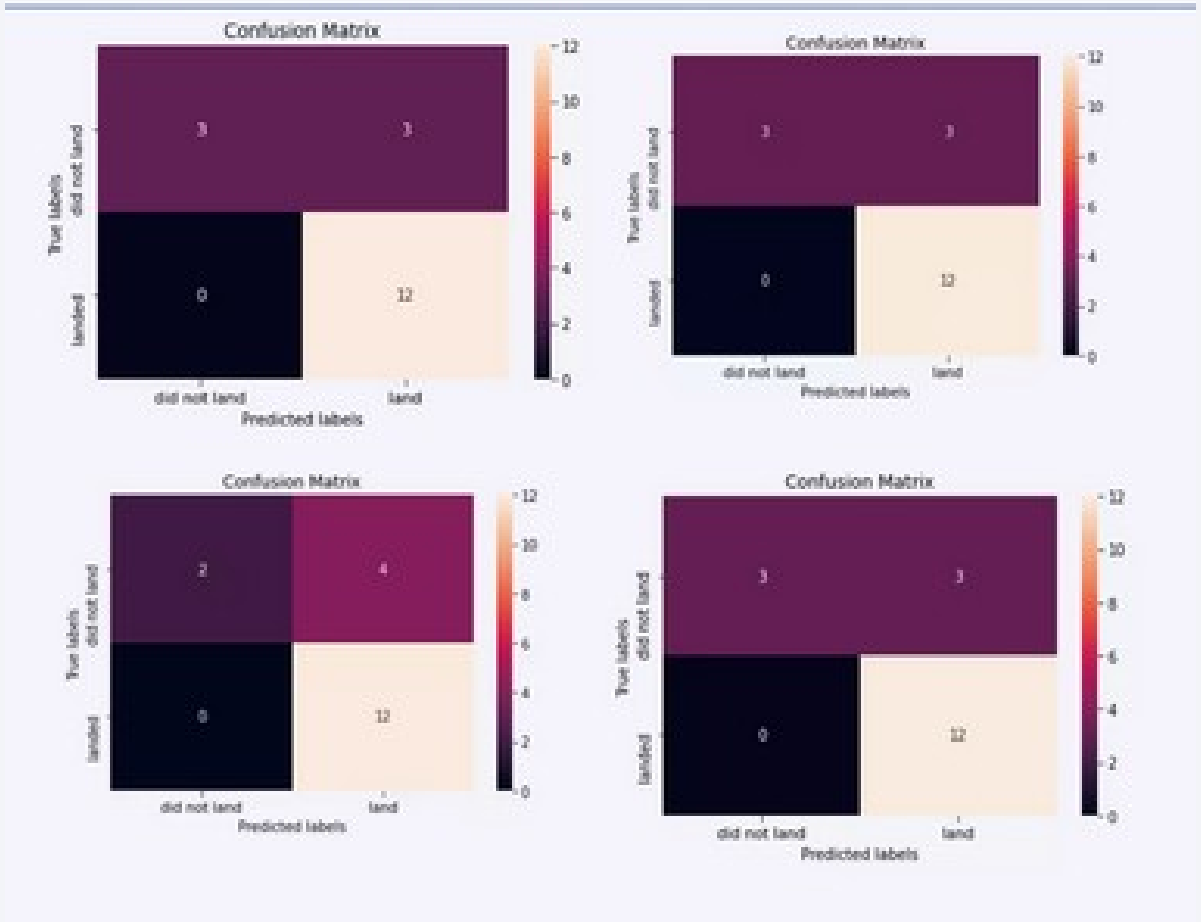
Classification Accuracy

Classification Accuracy Decision Tree has the highest accuracy with almost 0.89, then comes the remaining models with almost same accuracy of 0.84



Confusion Matrix

Confusion Matrix Sensitivity = 1.00, formula: TPR = TP / (TP + FN) Specificity = 0.50, formula: SPC = TN / (FP + TN) Precision = 0.80, formula: PPV = TP / (TP + FP) Accuracy = 0.83, formula: ACC = (TP + TN) / (P + N) F1 Score = 0.89, formula: F1 = 2TP / (2TP + FP + FN) False Positive Rate = 0.50, formula: FPR = FP / (FP + TN) False Discovery Rate = 0.20, formula: FDR = FP / (FP + TP) True Positive (TP) False Positive (FP) False Negative (FN) True Negative (TN)



Conclusions

The site with highest score is KSC LC-39A

The payload of 0 kg to 5000 kg was more diverse than 6000 kg to 10000 kg

Decision Tree was the optimal model with accuracy of almost 0.89

We calculated the launch sites distance to its proximities Conclusions

Thank you!