SpaceX: Falcon 9 Landing Prediction IBM Data Science Project

Aniekan Ido

Dec, 2022

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Key Insights
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Problem Statement:

SpaceX aims at commercializing space travel at a more affordable cost than competition

SpaceX can launch rockets for a cost of around $60m, in comparism to competition which goes for about $165m.

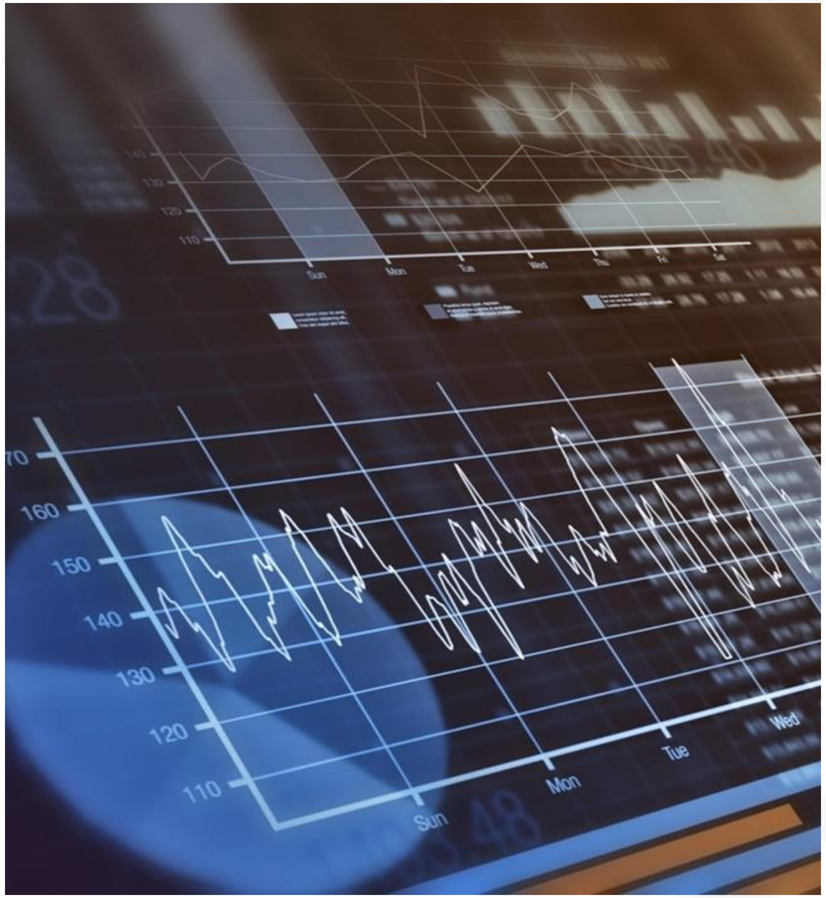
The primary cost saving metrics is the high success rate of stage 1 landing of the Falcon 9 rocket and thus its reusability in future launches

Project Aim:

1. To predict if the new launches will be successful by analyzing historic launch data and recommending what features contributes the most to ensuring a successful launch.

2. Track and Monitor Key KPIs aimed at gauging performance or progress of the rocket launches. Technology used was Python (Plotly dash) and KPIs considered: Total success launches per sites, correlation between Payload and outcome for different Booster versions .

INTRODUCTION



- Summary of Methodology
  - ✓ Data Collection
  - ✓ Data wrangling
  - ✓ EDA with Data visualization
  - ✓ EDA with SQL
  - ✓ Interactive Map with Folium
  - ✓ Interactive Dashboard with Plotly Dash
  - ✓ Machine Learning road Map
- Results Summary
  - ✓ Key Insights for EDA
  - ✓ Predictive Analysis results

METHODOLOGY

- Historical launch data for Falcon 9 was collected using:

    1. SpaceX REST API
    2. Using BeautifulSoup library to scrap data from the Wikipedia page.

- Exploratory Data Analysis and feature Engineering was carried out on Data derived from above and some major key insights were derived.

Tech used : Jupyter notebook, Python (Folium, Plotly Dash, Matplotlib, Pandas), SQL

- Performed predictive analysis using classification models, since we want to predict future outcomes. Each of our 4 model classification algorithm was modelled over the training and test dataset and accuracy was evaluated.

Machine Learning Techniques: Classification { Logistic Regression, Decision tree, SVM and K-nearest Neighbour }

DATA COLLECTION

SpaceX API

Extracted and parse the SpaceX launch data using the GET.request() from "https://api.spacexdata.com/v4/launches/past"	Decoded the response content as a Json using .json()	Converted into a Pandas dataframe using .json_normalize()	API was used again to get information about the launches using the IDs given for each launch and update columns and row	Filtered the database to keep only Falcon 9 launches, Missing value for payload mass was replaced with the mean. Final data coverted to CSV with name: 'dataset_part_1.csv'

IBM-SpaceX-Capstone-Project/jupyter-labs-spacex-data-collection-api.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)

Web Scraping

Extracted and parse the SpaceX launch data using the GET.request	Used BeautifulSoup() to create a BeautifulSoup object from a response text content	Applied find_all() function with `th` element on first_launch_table, Iterate each th element and apply the provided extract_column_from_header() to get a column name	Create a data frame by parsing the launch HTML tables & fill up with launch records extracted from table rows.	Convert the data to CSV with the name 'spacex_web_scraped.csv'

IBM-SpaceX-Capstone-Project/jupyter-labs-webscraping_Falcon_9.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)

DATA WRANGLING

After collecting the data we check the missing data ,and data types and do the following to clean the data :
Replace the missing data with mean .Change data type of the data. Represent categorical data using integer or float dummy numbers - one hot encoding. Perform EDA to find some patterns in the data and determine what would be the label for training supervised models.

Method	Insight	Narrative	Code
.value_counts()	Calculate the number of launches on each site	on the column LaunchSite to determine the number of launches on each site:	df['LaunchSite'].value_counts()
.value_counts()	Calculate the number and occurrence of each orbit	to determine the number and occurrence of each orbit in the column Orbit	df['Orbit'].value_counts()
.value_counts()	Calculate the number and occurrence of mission outcome per orbit type	on the column Outcome to determine the number of landing_outcomes.Then assign it to a variable landing_outcomes. We create a set of outcomes where the second stage did not land successfully from and assign to a variable bad_outcome:	landing_outcomes = df['Outcome'].value_counts()

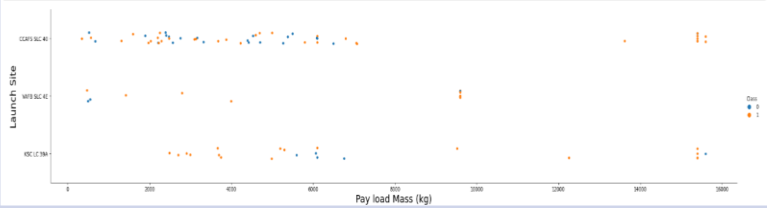
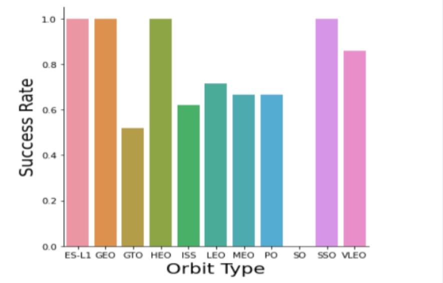
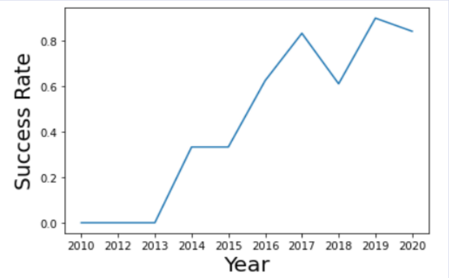
else statement	Create a landing outcome label from Outcome column	Using the Outcome, create a list where the element is zero if the corresponding row in Outcome is in the set bad_outcome; otherwise, it's one. Then assign it to the variable landing_class:	landing_class = []
			for key, value in df['Outcome'].items():
			if value in bad_outcomes:
			landing_class.append(0)
			else:
			landing_class.append(1)

.mean()		determine the success rate:	df["Class"].mean()
.to_csv()		Convert the data into csv file with the name 'dataset_part_2.csv'	df.to_csv("dataset_part_2.csv", index=False)

IBM-SpaceX-Capstone-Project/labs-jupyter-spacex-Data_wrangling.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)

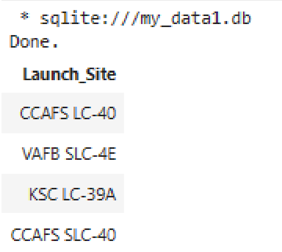
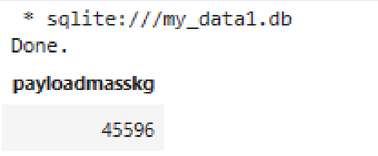
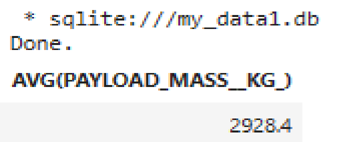
EDA WITH DATA VISUALIZATION

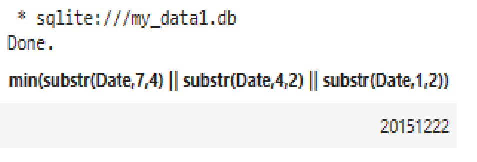
After Data cleaning we then proceeded to Analyzing the data using visualization to get some insights of the launches: 3 types of Visualization were used

Graph Type	Relationship	Picture
Scatter Plot	1. Visualize the relationship between Flight Number and Launch Site 2. Visualize the relationship between Payload and Launch Site 3. Visualize the relationship between FlightNumber and Orbit type 4.Visualize the relationship between Payload and Orbit type	
Bar Chart	1. Visualize the relationship between success rate of each orbit type	
Line chart	1. Visualize the launch success yearly trend	

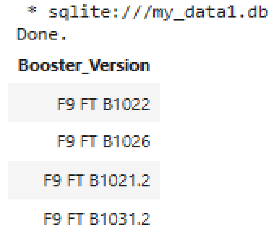
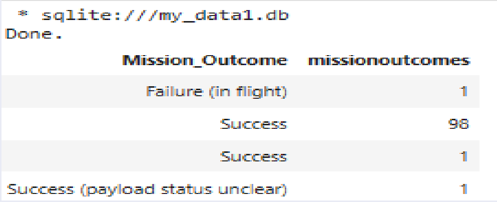
IBM-SpaceX-Capstone-Project/jupyter-labs-eda-dataviz.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)

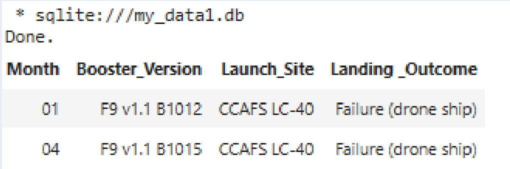
EDA WITH SQLLITE

Question	Code	Output
Display the names of the unique launch sites in the space mission	%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL;	 * sqlite:///my_data1.db Done. **Launch_Site** CCAFS LC-40 VAFB SLC-4E KSC LC-39A CCAFS SLC-40
Display 5 records where launch sites begin with the string 'CCA'	%sql SELECT * from SPACEXTBL where LAUNCH_SITE LIKE ('CCA%') LIMIT 5;	[IBM-SpaceX-Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)]()
Display the total payload mass carried by boosters launched by NASA (CRS)	%sql select sum(PAYLOAD_MASS__KG_) as payloadmasskg from SPACEXTBL Where Customer = 'NASA (CRS)';	 * sqlite:///my_data1.db Done. **payloadmasskg** 45596
Display average payload mass carried by booster version F9 v1.1	%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'	 * sqlite:///my_data1.db Done. **AVG(PAYLOAD_MASS__KG_)** 2928.4

List the date when the
first succesful landing outcome in ground pad was acheived.
	%%sql
	SELECT min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2))
	from SPACEXTBL
	where "Landing _Outcome" ='Success (ground pad)';
		
		* sqlite:///my_data1.db
		Done.
		min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2))
		20151222

EDA WITH SQLLITE

Question	Code	Output
List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000	%%sql select Booster_Version from SPACEXTBL where "Landing _Outcome" ='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999;	 * sqlite:///my_data1.db Done. Booster_Version F9 FT B1022 F9 FT B1026 F9 FT B1021.2 F9 FT B1031.2
List the total number of successful and failure mission outcomes	%sql select MISSION_OUTCOME , count(*) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME	 * sqlite:///my_data1.db Done. Mission_Outcome missionoutcomes Failure (in flight) 1 Success 98 Success 1 Success (payload status unclear) 1
List the names of the booster_versions which have carried the maximum payload mass	%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);	IBM-SpaceX-Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)

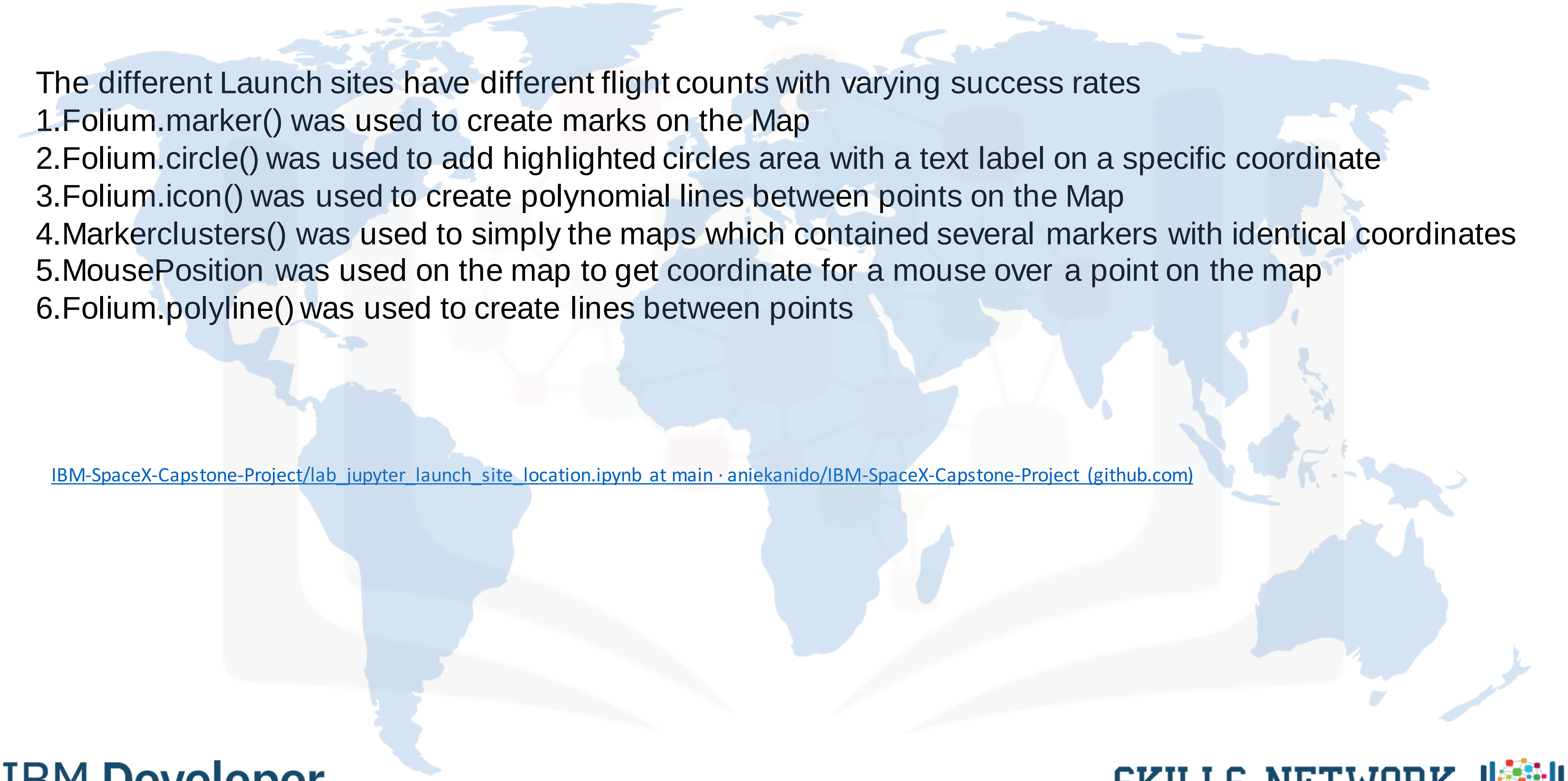
List the records which will display the month names,
failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
	%%sql
	SELECT substr(Date,4,2) as Month,Booster_Version,Launch_Site,"Landing _Outcome"
	FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' and substr(Date,7,4)='2015'
		
		* sqlite:///my_data1.db
		Done.
		Month Booster_Version Launch_Site Landing _Outcome
		01 F9 v1.1 B1012 CCAFS LC-40 Failure (drone ship)
		04 F9 v1.1 B1015 CCAFS LC-40 Failure (drone ship)

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.	%%sql
	SELECT "Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT,DATE
	from SPACEXTBL where substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604' and '20170320'
	group by "Landing _Outcome" order by count("Landing _Outcome") desc
		IBM-SpaceX-Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)

INTERACTIVE MAP WITH FOLIUM

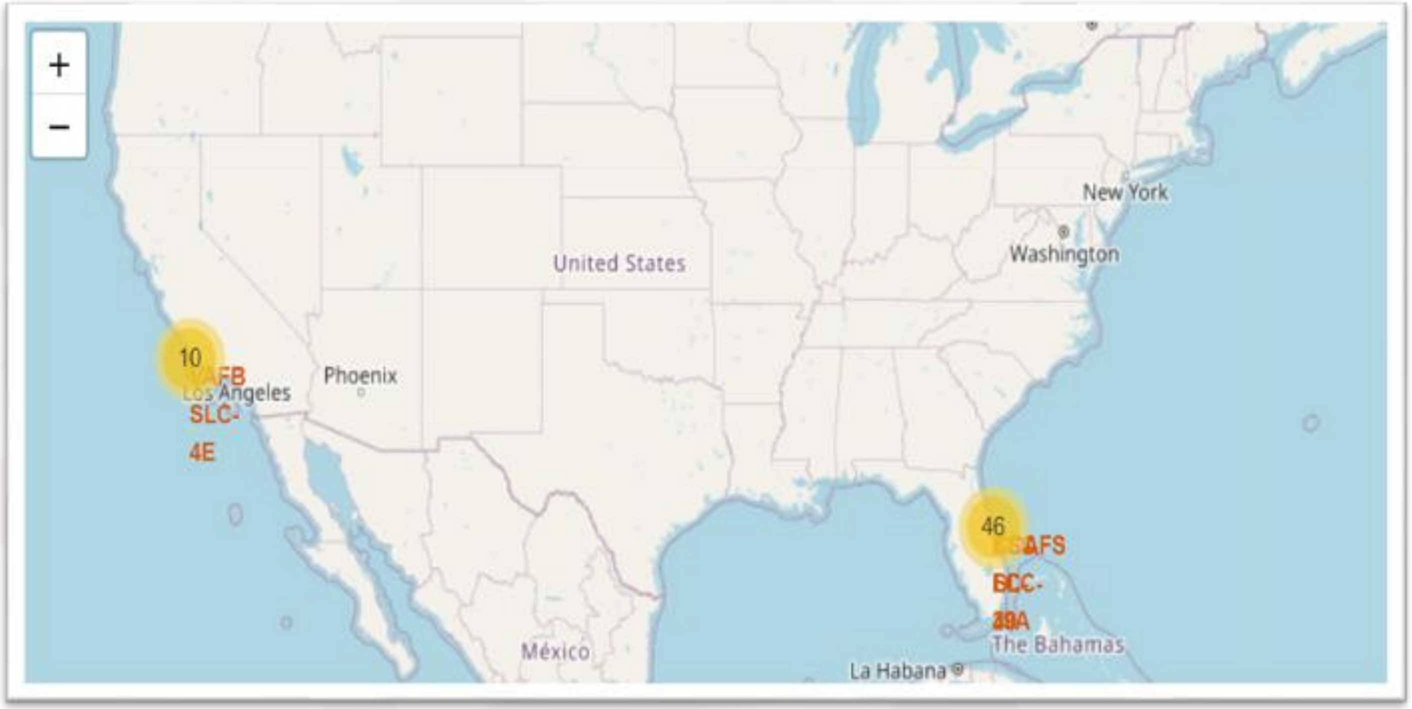
The different Launch sites have different flight counts with varying success rates
1.Folium.marker() was used to create marks on the Map
2.Folium.circle() was used to add highlighted circles area with a text label on a specific coordinate
3.Folium.icon() was used to create polynomial lines between points on the Map
4.Markerclusters() was used to simply the maps which contained several markers with identical coordinates
5.MousePosition was used on the map to get coordinate for a mouse over a point on the map
6.Folium.polyline() was used to create lines between points

[IBM-SpaceX-Capstone-Project/lab_jupyter_launch_site_location.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)](https://github.com)



LAUNCH SITE LOCATION ON MAP

As observed in the map, most of the launch sites are in proximity to the equator line. One reason is that launching a rocket from the coast gives it an additional boost due to the rotational speed of Earth. Another reason is that if something goes wrong during the ascent, the debris will fall into an ocean instead of a densely populated area. Additionally, most launch sites are located near the equator because rockets launched from these sites get an additional natural boost that helps save on fuel and boosters.

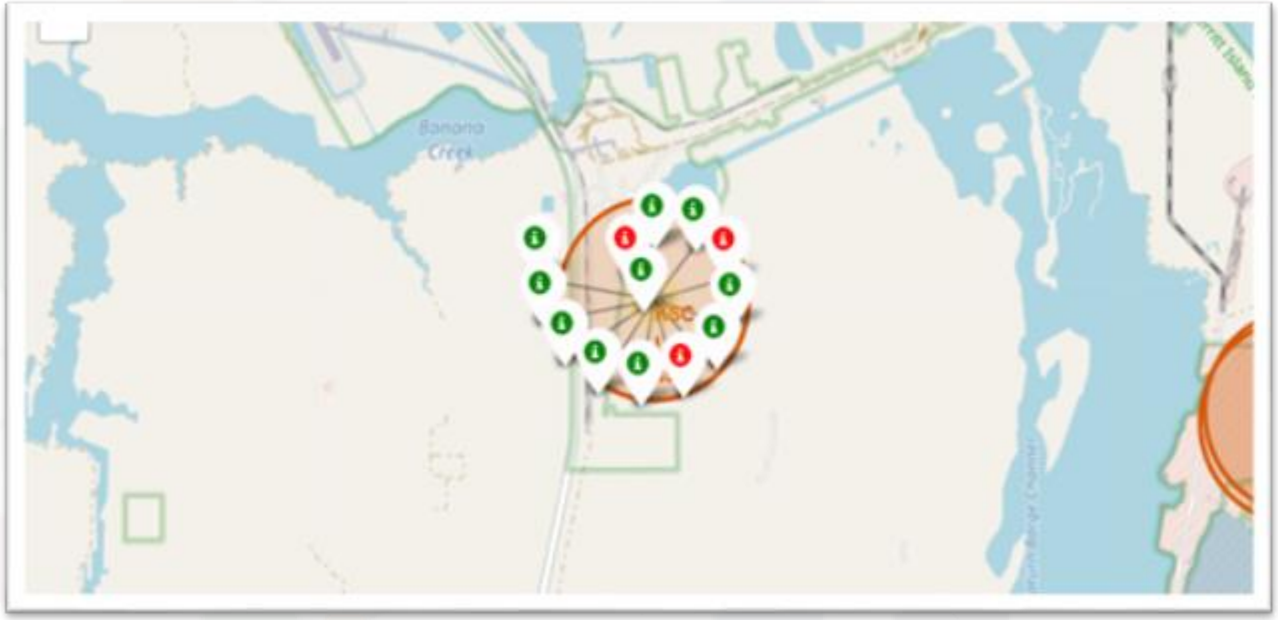


COLOUR-LABELLED LAUNCH OUTCOME ON MAP

This will help the stakeholders to easily identify which launch sites has a relatively high success rate by just looking at the map.

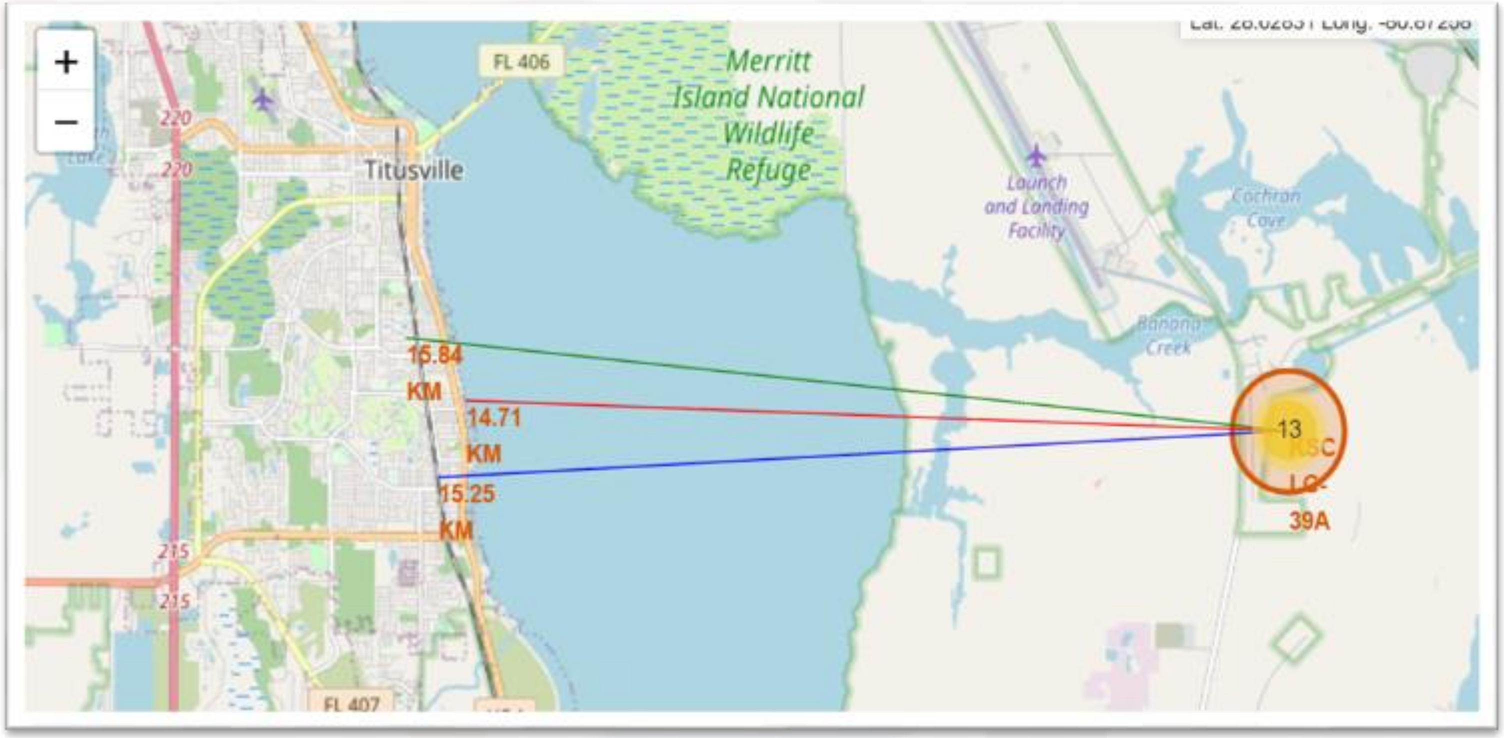
1. Green Marker = Successful launch
2. Red Marker = Failed Launch

This particular Launch site (KSC LC-39A) can be clearly seen to have a very successful launch rate.



LAUNCH SITE TO ITS PROXIMITIES

KSC LC-39A launch site seen in the picture is not too far off from its proximities .

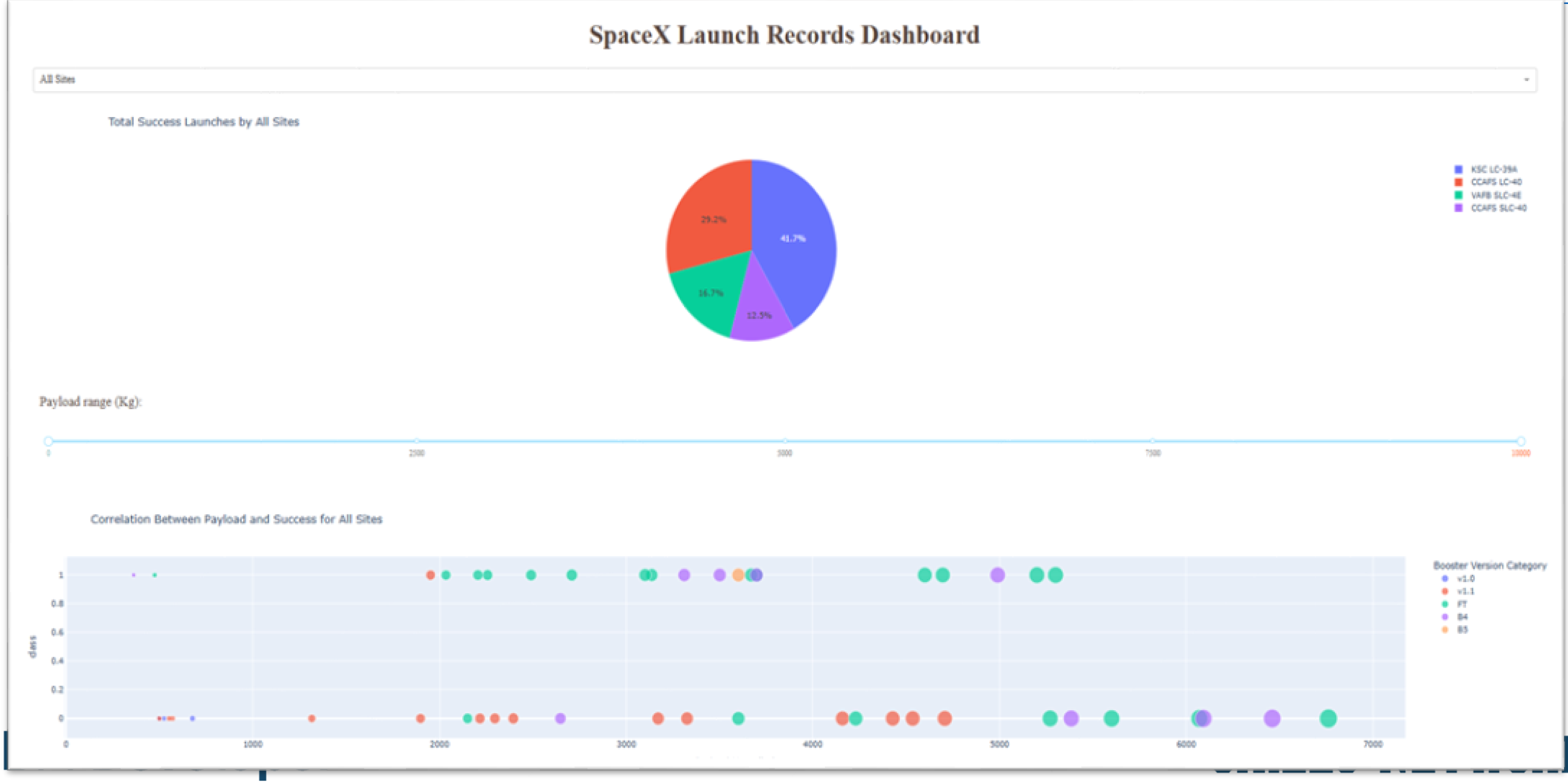


BUILD A DASHBOARD WITH PLOTLY DASH

IBM-SpaceX-Capstone-Project/Interactive Dashboard with Ploty Dash_IBM Capstone Project.ipynb at main · aniekanido/IBM-SpaceX-Capstone-Project (github.com)

Visualization of the launches from the sites

1. Success launch % per Site
2. Visualize impact of payload Mass of launch outcome for the different Booster category version



SpaceX Launch Records Dashboard

All Sites

Total Success Launches by All Sites

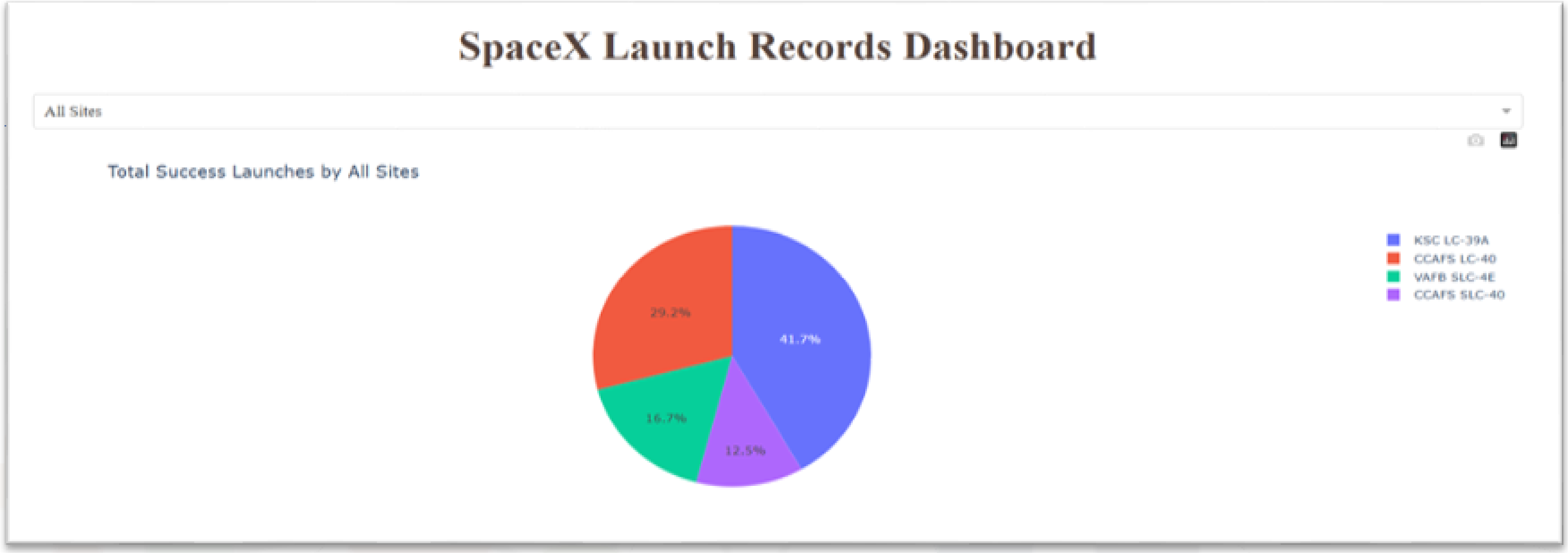
Payload range (Kg):

Correlation Between Payload and Success for All Sites

DASHBOARD

KSC LC 39A has the most count of successful launch (41%)

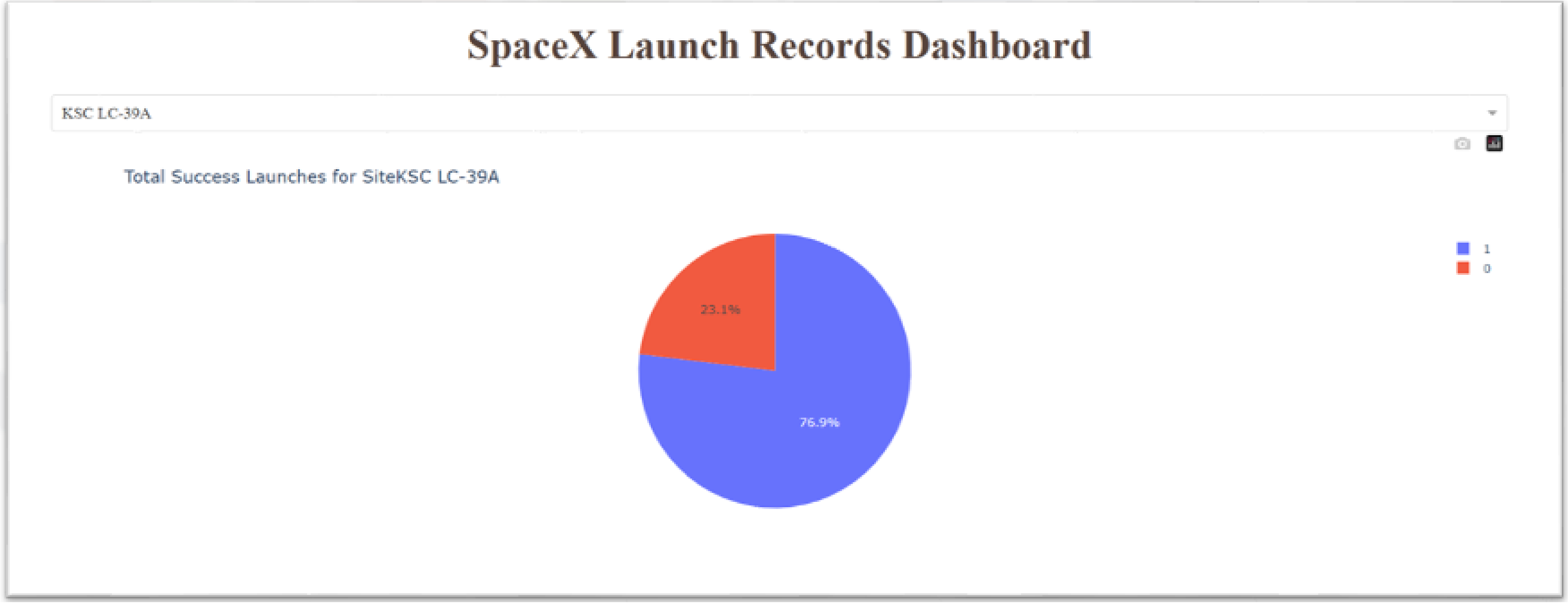
77% of her launch was successful



SpaceX Launch Records Dashboard

All Sites

Total Success Launches by All Sites



SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for SiteKSC LC-39A

MACHINE LEARNING ROADMAP

Data Pre-processing ●

- *Exploratory data Analysis*
- *Handling Missing values*
- *Data transformation .*
- *Feature engineering .*

Model Selection & Improvement ●

- *Train/Test Split*
- *Algorithm Selection*
- *Hyperparameter tuning*
  - A. *Grid Search cross validation*

Model Evaluation ●

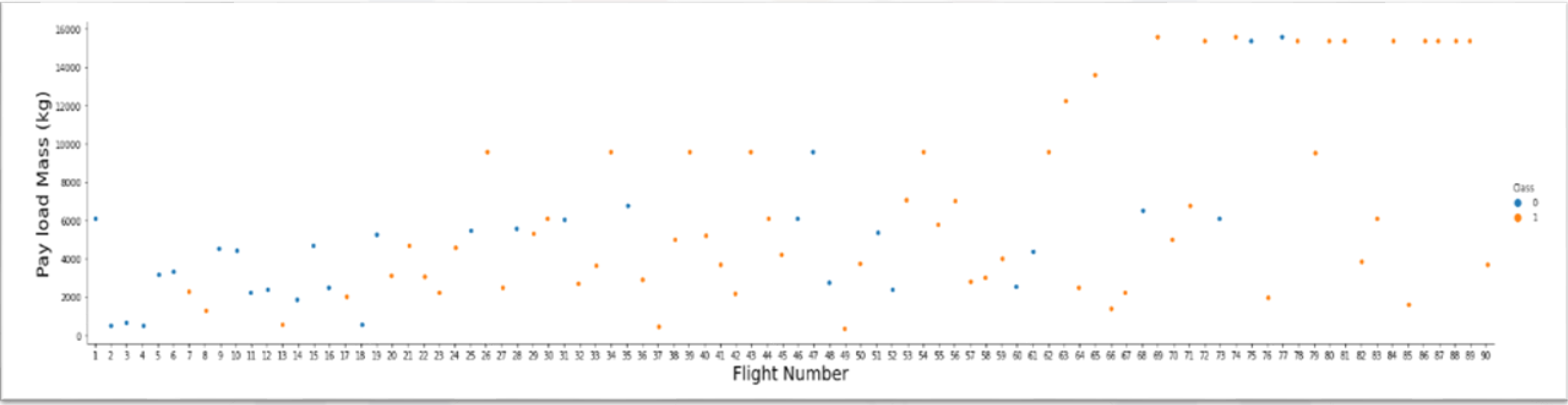
- *Classification Report*
- *Confusion Matrix .*

Selecting optimal Model ●

- *Model Scoring .*

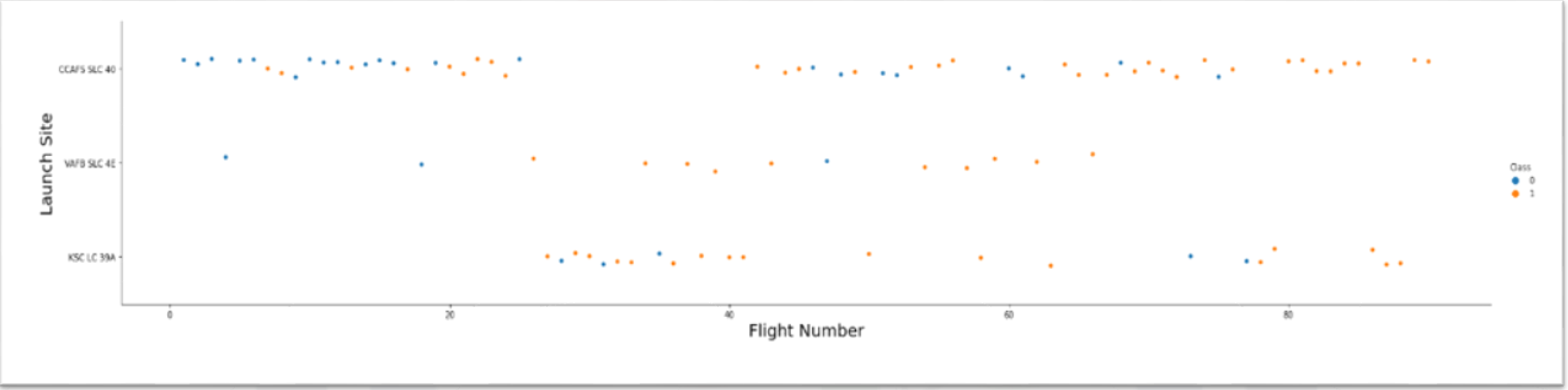
KEY INSIGHTS FROM DATA EXPLORATION

RESULTS – FlightNumber Vs PayloadMass



We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

RESULTS – FlightNumber Vs LaunchSite



The different Launch sites have different flight counts with varying success rates

1.Earlier flights launch were from CCAFS-SLC-40 site ,Followed by KSC-LC-39A

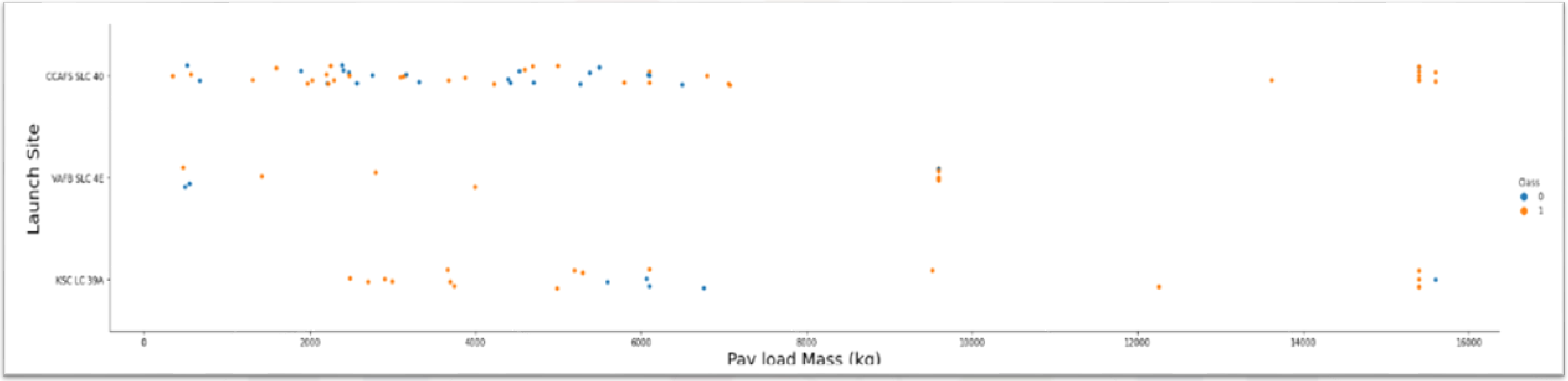
2.CCAFS SLC40, with a total of 55 flights and 33 successful (60%)

3.VAFB SLC 4E, with a total of 13 flights and 10 successful (77%)

4.KSC LC 39A, with a total of 22 flights and 17 successful (77%)

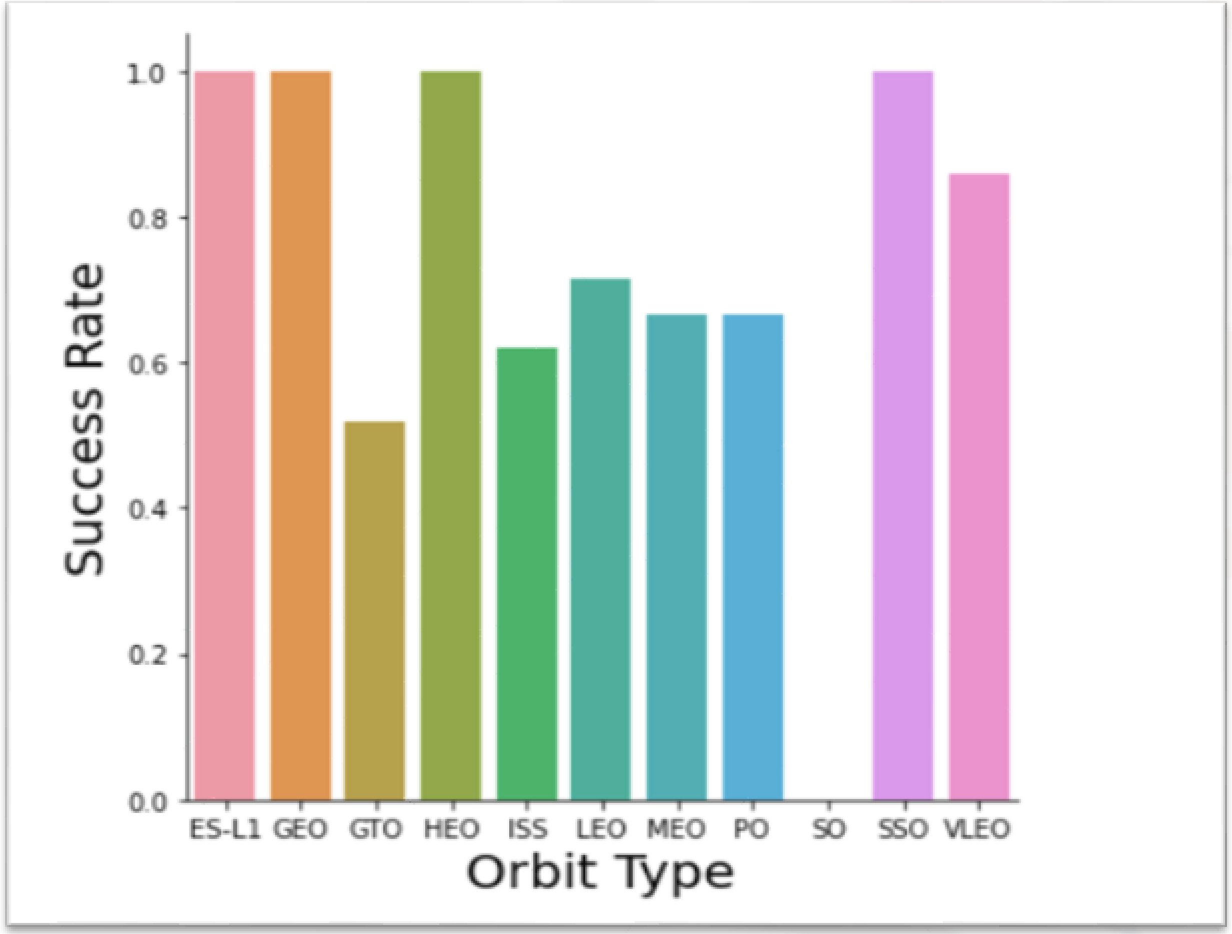
5.Most Launches are Launched from CCAFS-SLC-40

RESULTS – Payload Vs LaunchSite



1. CCAFS SLC 40 has more Higher Payload Launches and Low Payload Lauches .
2. KSC LC 39A has a 100% success rate for payload mass under 5800kg
3. VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

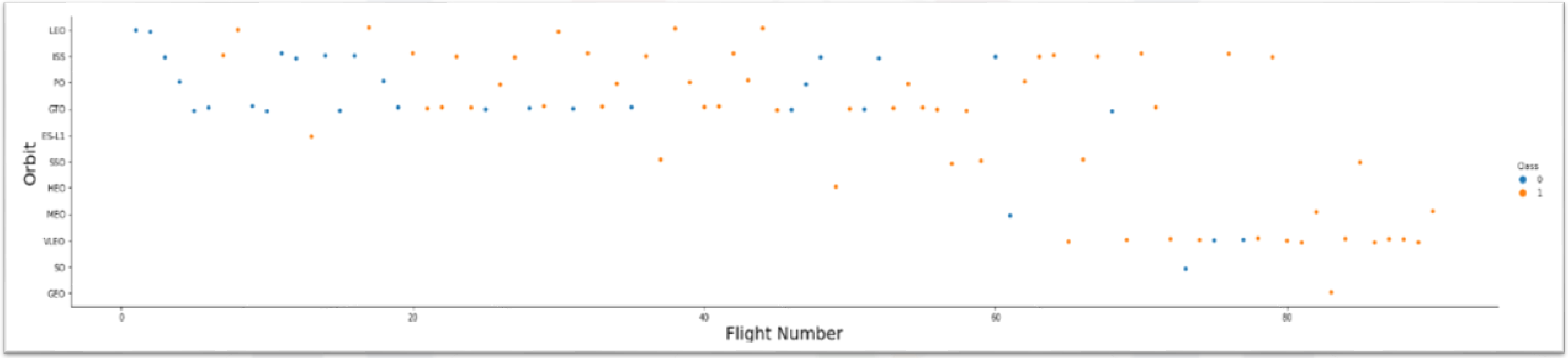
RESULTS – Success rate Vs Orbit Type



1. Orbits ES-L1, GEO, SSO, HEO have the highest success rate at 100%

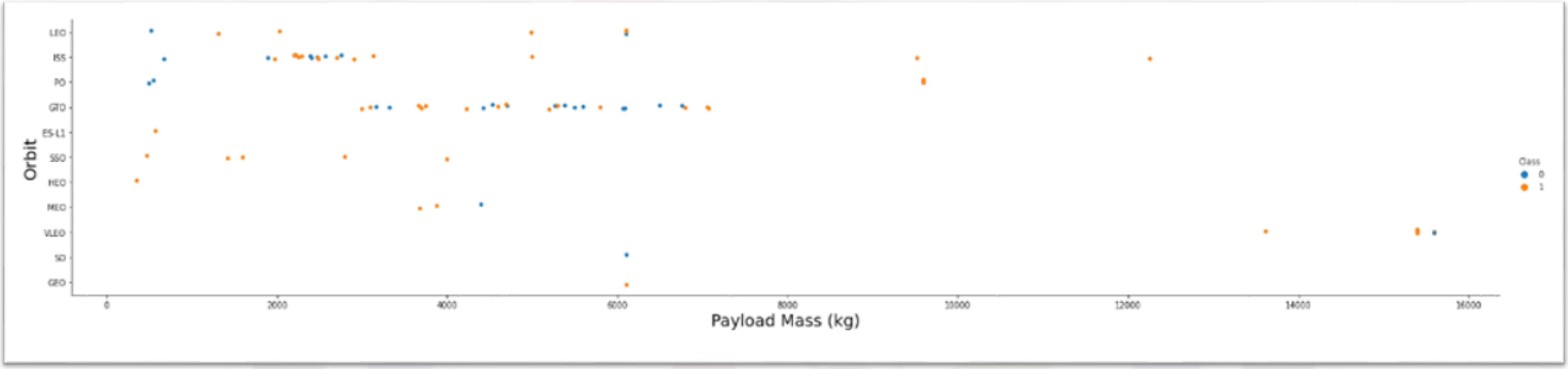
2. SO Orbit has 0% success rate

RESULTS - FlightNumber Vs Orbit Type



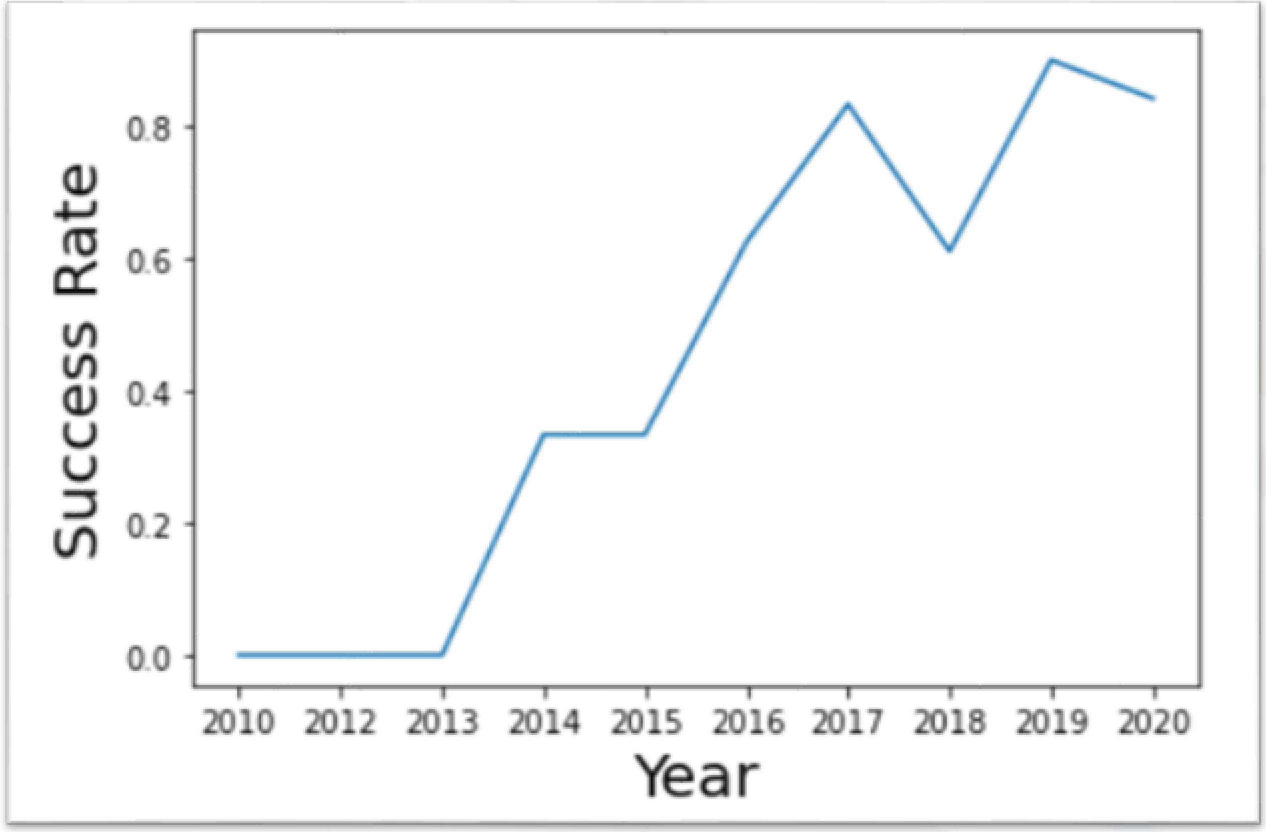
We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

RESULTS – Payload Mass 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.

RESULTS



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

PREDICTIVE ANALYSIS (CLASSIFICATION)

MODEL SELECTION AND EVALUATION

Each of our 4 selected model classification algorithm was modelled over the training and test dataset and evaluated . 4 metrics were used to evaluate each of the Models

Model	Accuracy	Precision	Recall	F1 Score	
Logistic Regression	83.33%	87%	83%	81%	✓
Support Vector Machine	83.33%	87%	83%	81%	✓
Decision Tree	83.33%	87%	83%	81%	✓
K-nearest Neighbour	77.78%	83%	78%	74%	

3 out of the 4 Models performed similarly, hence any can be used to make predictions

CONCLUSIONS

- The payload mass is an important feature, the more massive the payload, the less likely the first stage will return.
- An increase in flights number increases the chances of a successful launch
- Outcome of launches are getting increasing successful over time
- Most launches were from KSC PAD 39A since most of them were to VLEO,GEO or ISS which makes it a good site to launch from
- Winning models for our prediction @ 83.33% accuracy and F1 score of 81% are : Logistic Regression , Support Vector Machine and Decision Tree

THANK YOU