IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Riccardo Lovato
14-08-2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium

Summary of all results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

Project background and context

- SpaceX aims to place satellites on orbit at the best price in order to meet our customers requirements; Falcon 9 rocket launches can cost 62 million dollars each whereas other contenders announce price between 100 million and 165 millions each launch.
- Space X attractive price mostly relies on a "recoverable" first stage (that is the more expensive part of the rocket system) but success is not guaranteed – there maybe some failures depending on several features.

Problems you want to find answers

- Analyze the success rate of 1st stage landing and how it is correlated to Launch Site, Payload Mass, Booster Version; with that information, pricing of a future launch can be more precisely determined, reducing business risk.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Past launches of Space X are retrieved from SpaceX API: https://api.spacexdata.com/v4
- Multiple API endpoints where used to retrieve various information aspects, and the complete dataset forms a DataFrame
  - https://api.spacexdata.com/v4/launches/past
  - https://api.spacexdata.com/v4/launchpads/
  - https://api.spacexdata.com/v4/payloads/
  - https://api.spacexdata.com/v4/cores/

/launches
- Flight number, Rocket, Payloads, Date
- Launchpad, Payload, Core

/launchpads/
- Launchpad name
- Longitude, Latitude

/payloads/
- Mass
- Orbit

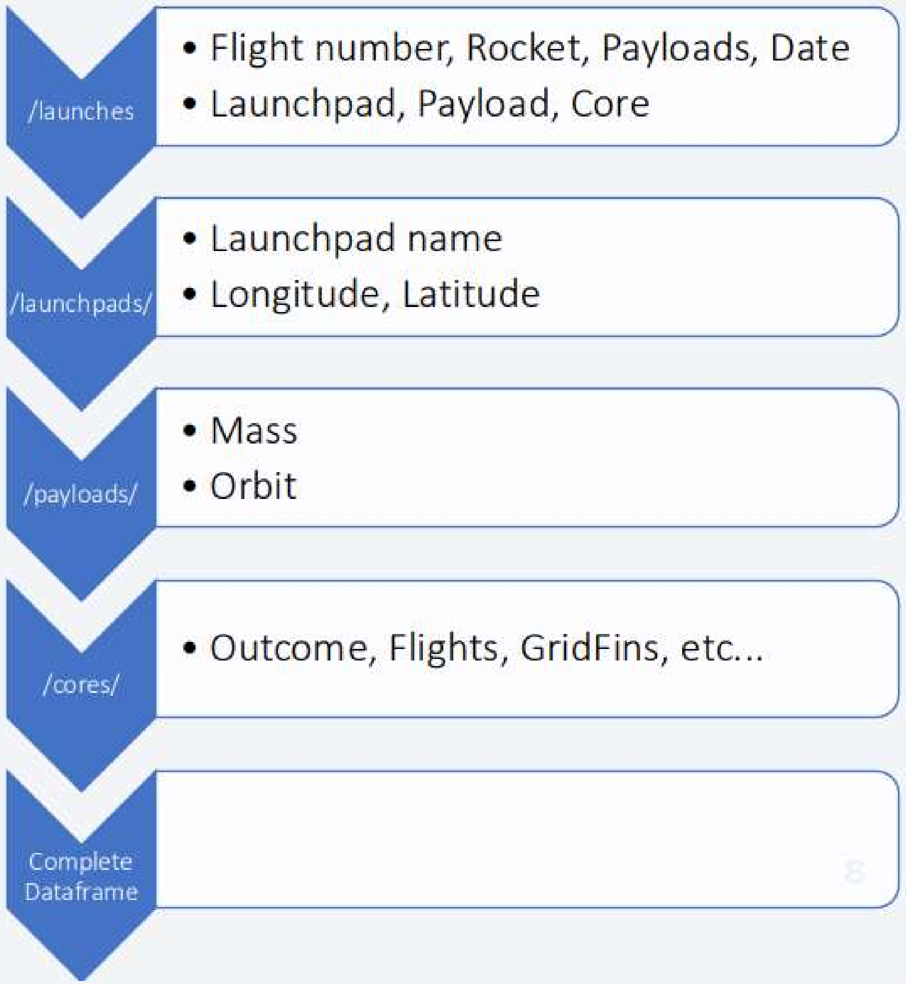
/cores/
- Outcome, Flights, GridFins, etc...

Complete Dataframe

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

- The link to the notebook: https://github.com/852208/IbmDataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

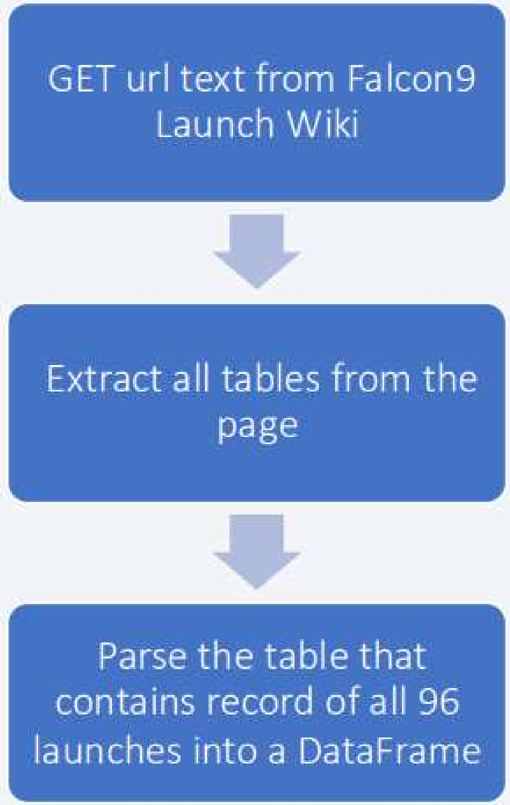


Data Collection - Scraping

- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup

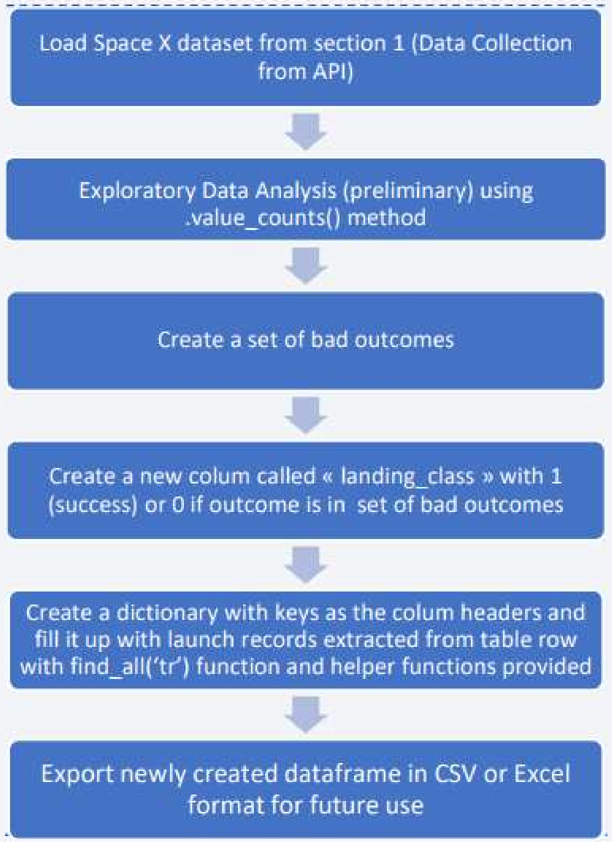
- We parsed the table and converted it into a pandas dataframe.

- The link to the notebook: https://github.com/852208/IbmDataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb



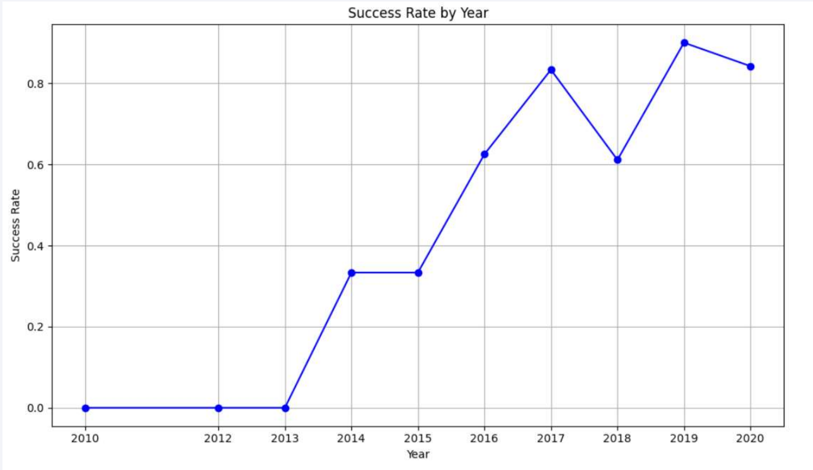
Data Wrangling

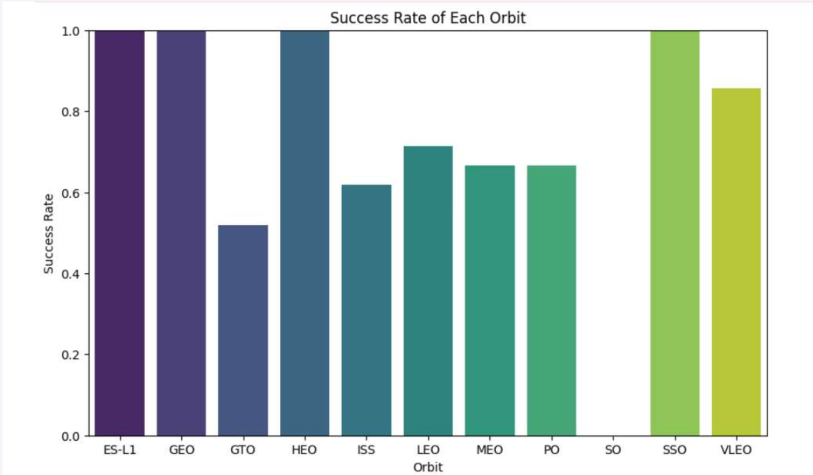
- Objective was to convert Launch outcomes data into Training Labels with 1 means booster successfully landed and 0 when it fails
- We made preliminary EDA to identify:
  - Number of launches on each site;
  - Number and occurrence of each orbit (list of 11)
- We created a new column called "Class" and computed the mean of success: 0,67
- The link to the notebook: https://github.com/852208/IbmDataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Objective of EDA is to predict if the Falcon 9 first stage will land successfully
- We have looked for any correlation/patterns between success and Flight Numbers, Launch Site, Payload Mass, Orbit type, Yearly trend with various plots
- We also have created dummy variables to categorical columns to comply with Machine Learning models (next section)
- The link to the notebook: https://github.com/852208/IbmDataScienceCapstone/blob/main/edadataviz.ipynb



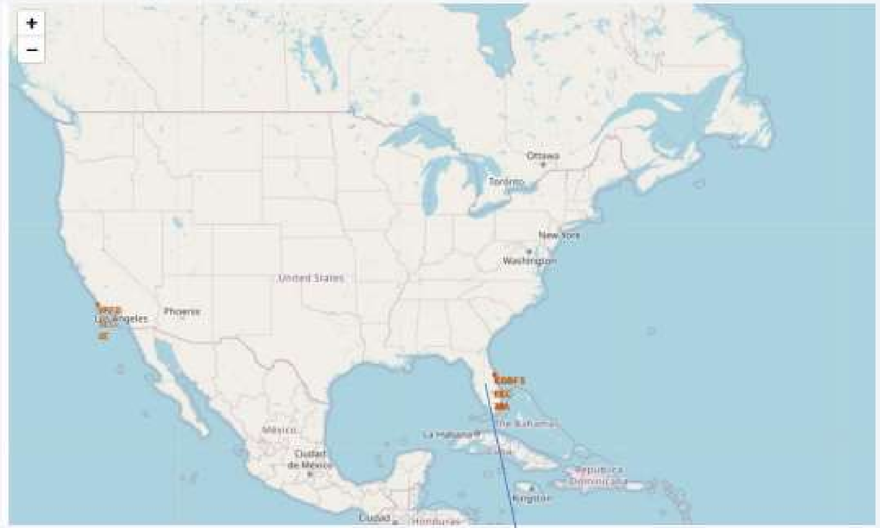


EDA with SQL

- Complementary Exploratory Data Analysis of Space X database using SQL queries
- Got more information on Customers, Booster Version, Payload Mass, type of landing outcomes
- We used sqlite3 package and %sql Magic commands
- We used several type of SQL Commands such as:
  - o Simple command e.g. %sql SELECT * from SPACEXTABLE
  - o Conditional commands e.g. %sql ….. Where {condition}
  - o Aggregate function commands e.g. %sql SELECT sum("column name") from SPACEXTABLE
  - o Grouping and sorting commands e.g. %sql SELECT "col name" from SPACEXTABLE GROUP BY DESC
- The link to the notebook: https://github.com/852208/IbmDataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Objective was to locate Launch Site and their proximities such as railway, highways, city on a map to be used to assess success rate and economic benefit
- We used Space X database and Folium packages and sub-packages
- We have marked the Launch Sites on the map to see their distance from the Equator line using folium.Map function centered on NASA Johnson Space Centers's coordinates
- We used folium.Circle and folium.Marker fucntions to show all 4 Launch Sites on the map with their names
- We have used MarkerCluster objects to show all launch outcomes (success or not) having the same coordinate using folium.Marker and folium.Icon functions and add.child() method to aggregate
- We have calculated the distances between a Launch Site and its proximities and we have drawn a line in between with folium.PolyLine function
- https://github.com/852208/IbmDataScienceCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

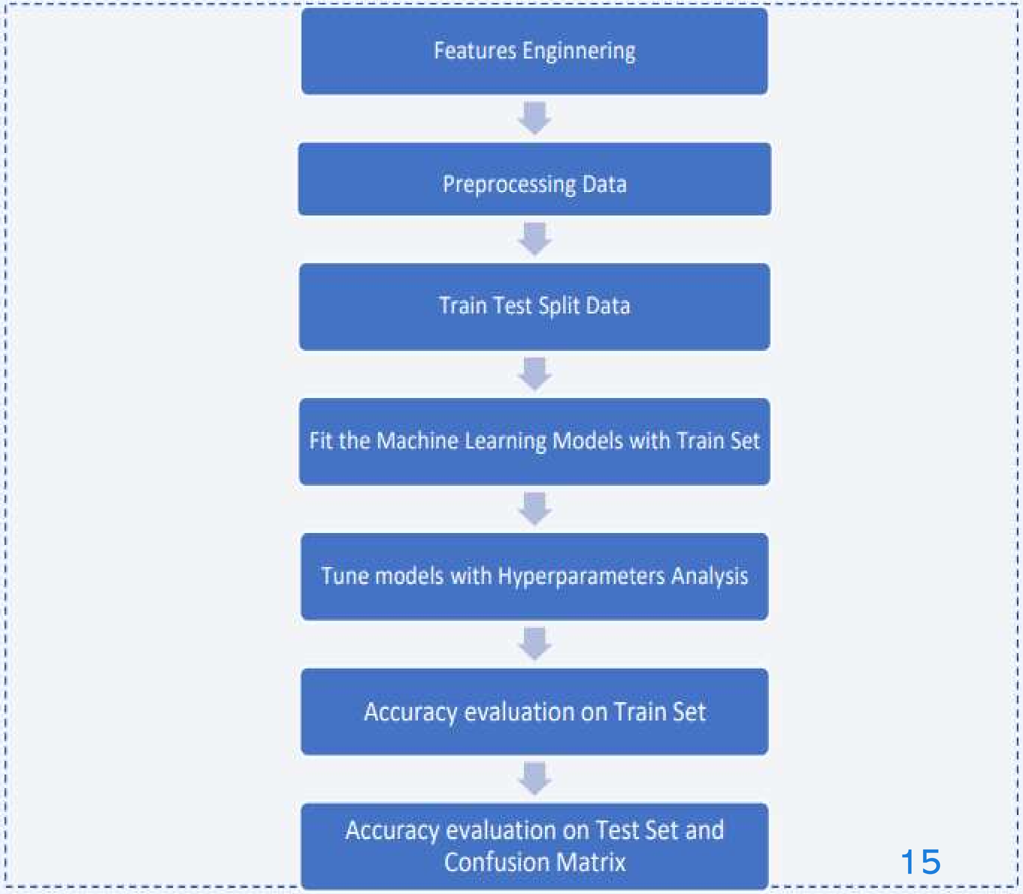


Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.
- The link to the notebook: https://github.com/852208/IbmDataScienceCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

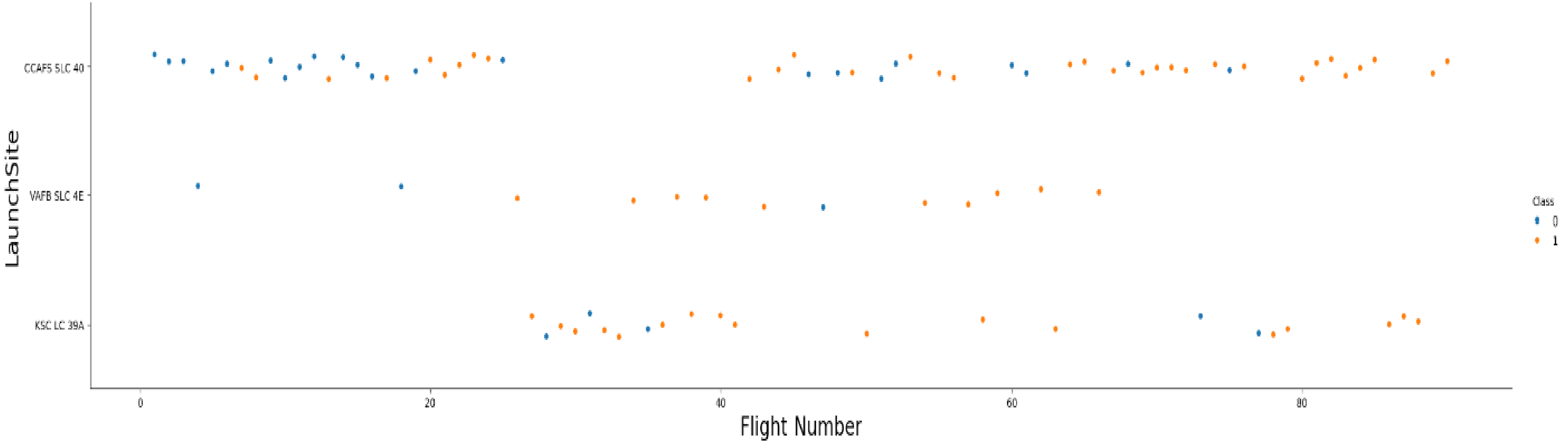
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 3

Insights drawn from EDA

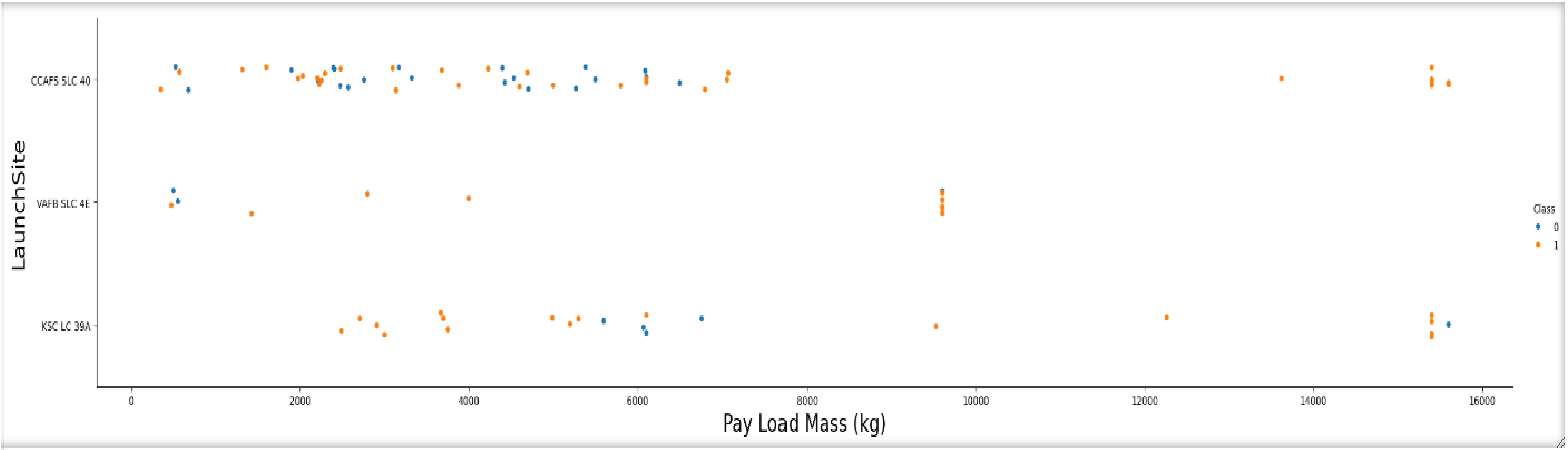
Flight Number vs. Launch Site

- Most of flight numbers started in the CCAFS SLC 40 Launch Site with many failures (class 1 blue mark). Starting FN23 Space X decided for some reason (unavailability of the Launch Site ?) to use the other two launch Sites
- They came back to the original Launch Site after Flight Number 41
- Starting from Flight Number 60, the outcome drastically improved (most are class 1).



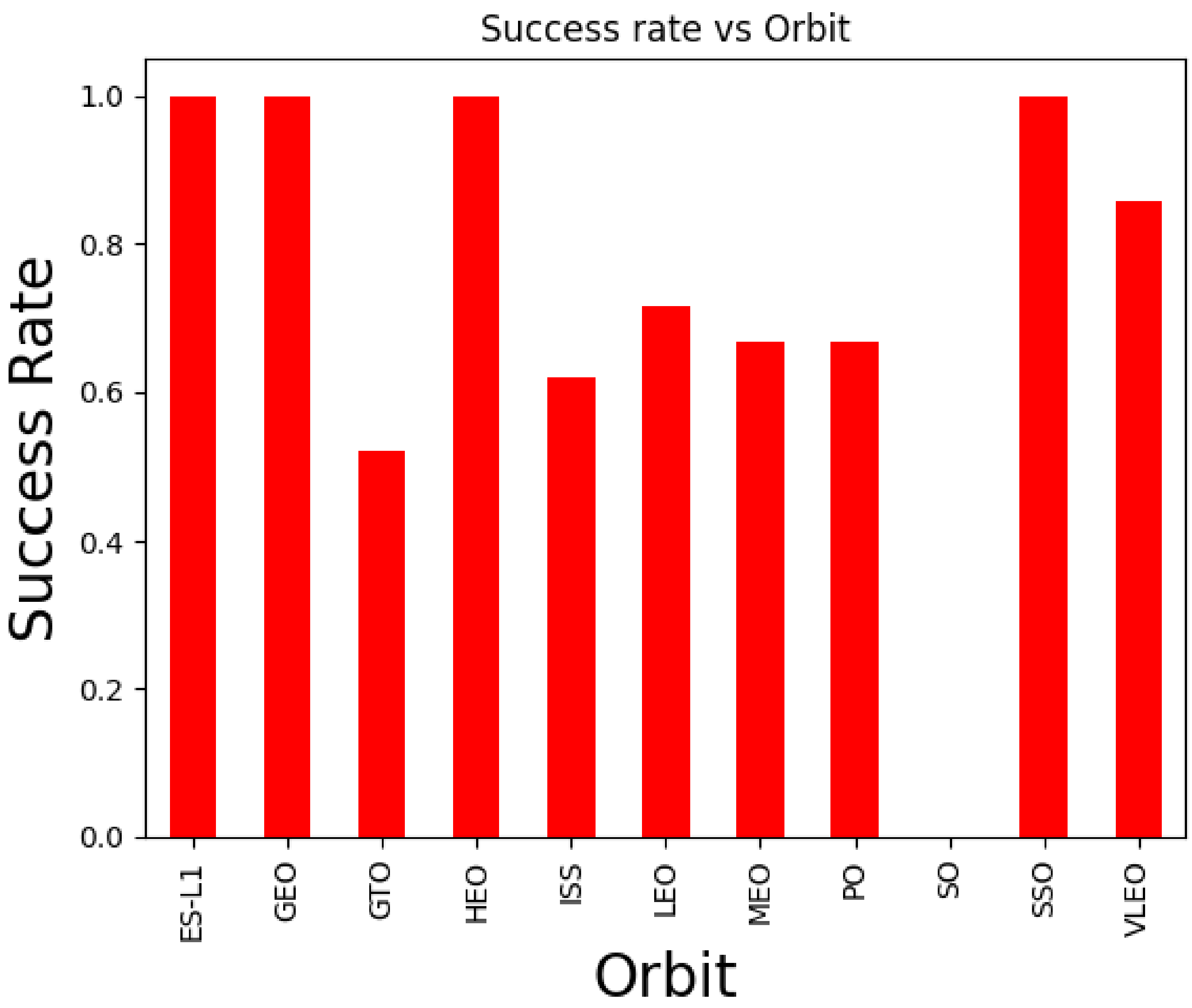
Payload vs. Launch Site

- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)
- It seems that payload range 4000-6000 kg is critical since there are many failures



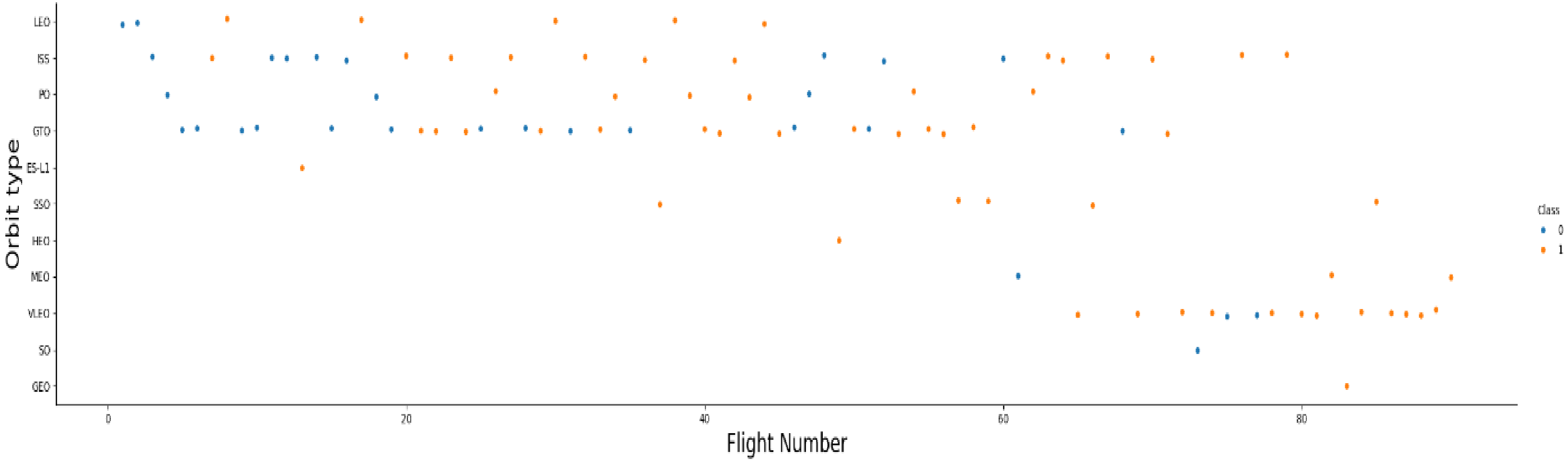
Success Rate vs. Orbit Type

- Orbit with 100% success are ES-L1, GEO, HEO, SSO however there has been only a few launches (see next slide)
- LEO and VLEO orbits had relatively good success because they are low altitude
- The success rate is smaller for high altitude orbit such as GTO



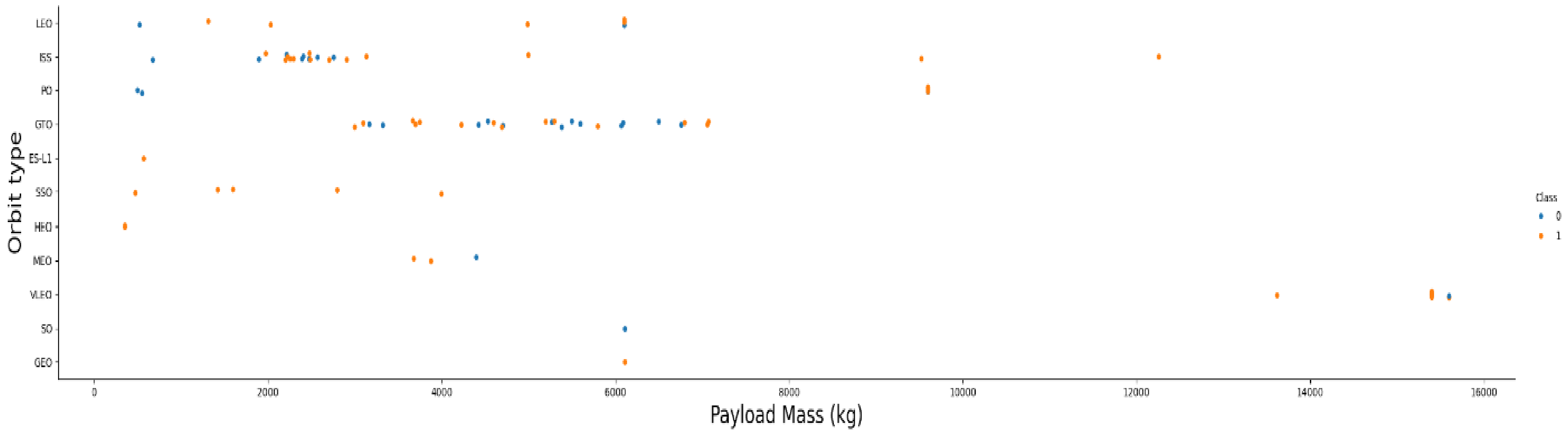
Flight Number vs. Orbit Type

- GTO, ISS,VLEO are the most used orbits especially VLEO since Flight Number 60
- Success rate has improved regularly



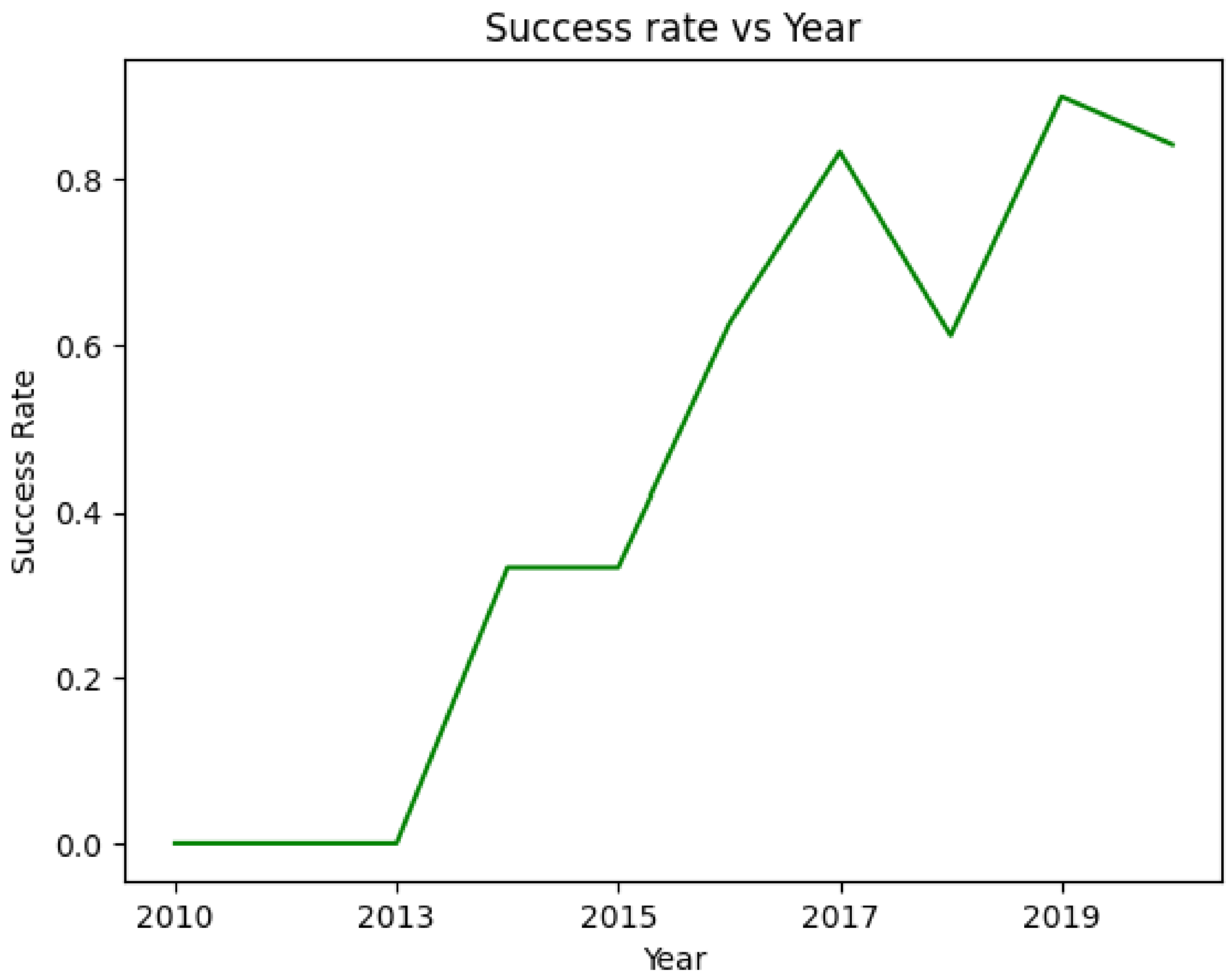
Payload vs. Orbit Type

- 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

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



All Launch Site Names

- 4 Unique launch sites obtained with SQL queries:

- VAFB SLC-4E is situated on the West Coast in California (Vandenberg Space Force Base, owner US Space Force)

- The three other ones are situated in Florida at Cape Canaveral and own either to the US Air Force or to NASA (KSC LC-39A)

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

Launch Site Names Begin with 'CCA'

- Here are 5 records where launch sites begin with `CCA`
- First test of Falcon 9 started in 2010 from CCAFS LC-40 Launch site then from VAFB SLC-4E in 2013
- Small Payload mass and short orbit (LEO) to start with
- “landing outcome” is about landing the first stage which started in 2012

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

Total Payload Mass

- NASA is the main Customer of Space X
- Total payload carried by boosters from NASA ≈ 100 tons,
- About half (45 tons) is for NASA (Commercial Resupply services) with the aim of supplying ISS

- We used the following query : %sql select SUM(PAYLOAD_MASS__KG_) as PayLoad_Total_Mass_All_NASA from SPACEXTABLE where Customer like 'NASA%'

PayLoad_Total_Mass_All_NASA
99980

Customer	PayLoad_Total_Mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2534 kg
- It varies between 500 kg and 4700 kg

Average_Payload_Mass
2534.6666666666665

- We used the following SQL command : %sql SELECT "Booster_Version", AVG("PAYLOAD_MASS_KG") as Payload_Mass_Avg from SPACEXTABLE GROUP BY "Booster_Version“

First Successful Ground Landing Date

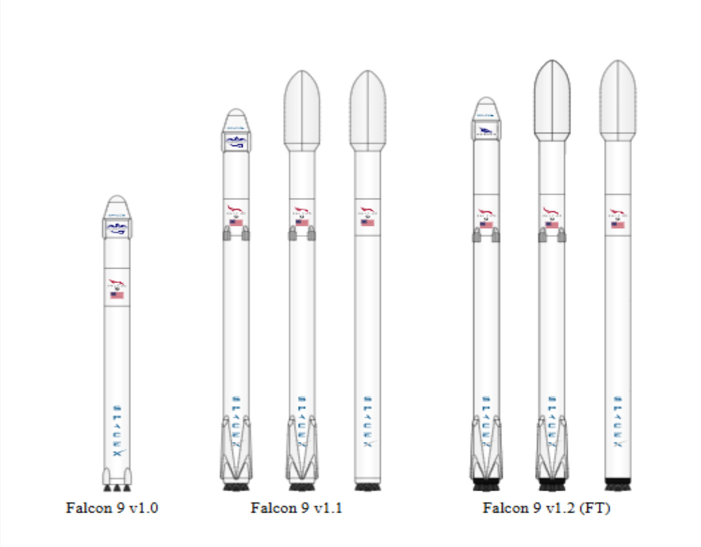
- Table on the right shows the earlier dates of Landing Outcome types
- It can be seen from the table that the first successful landing outcome in ground was on **22 December 2015**

Date_Minimum	Landing_Outcome
2014-04-18	Controlled (ocean)
2018-12-05	Failure
2015-01-10	Failure (drone ship)
2010-06-04	Failure (parachute)
2012-05-22	No attempt
2019-08-06	No attempt
2015-06-28	Precluded (drone ship)
2018-07-22	Success
2016-04-08	Success (drone ship)
2015-12-22	Success (ground pad)
2013-09-29	Uncontrolled (ocean)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version	Landing_Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)



- We used the following Query : %sql SELECT "Booster_Version","Landing_Outcome“ from SPACEXTABLE WHERE "Landing_Outcome"='Success (drone ship)' AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes
- Mission outcome was very good 99%

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

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass (15,6 tons) : see table on the right
- %sql SELECT "Booster_Version","PAYLOAD_MASS__KG_" from SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") from SPACEXTABLE); (we used a subquery for this complex query)

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

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

Month_Name	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

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

- Present your query result with a short explanation here

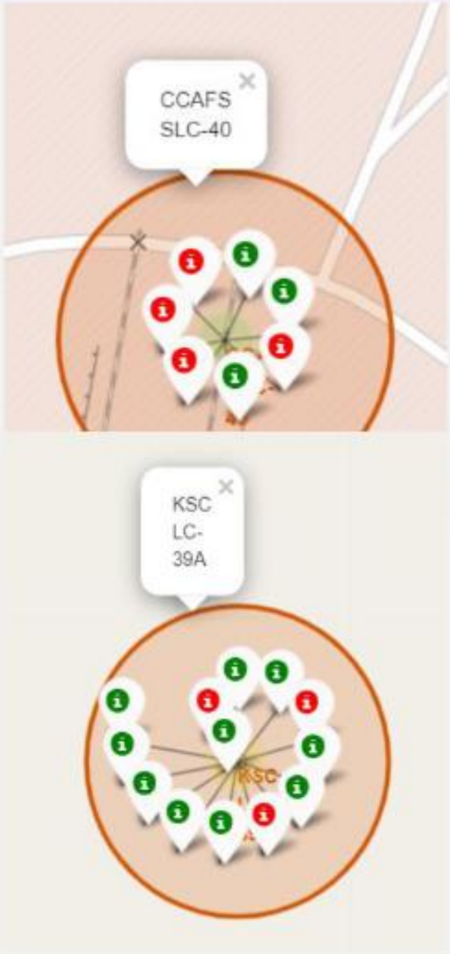
Landing_Outcome	Landing_Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

All launch sites global map markers

Markers showing launch sites with color labels



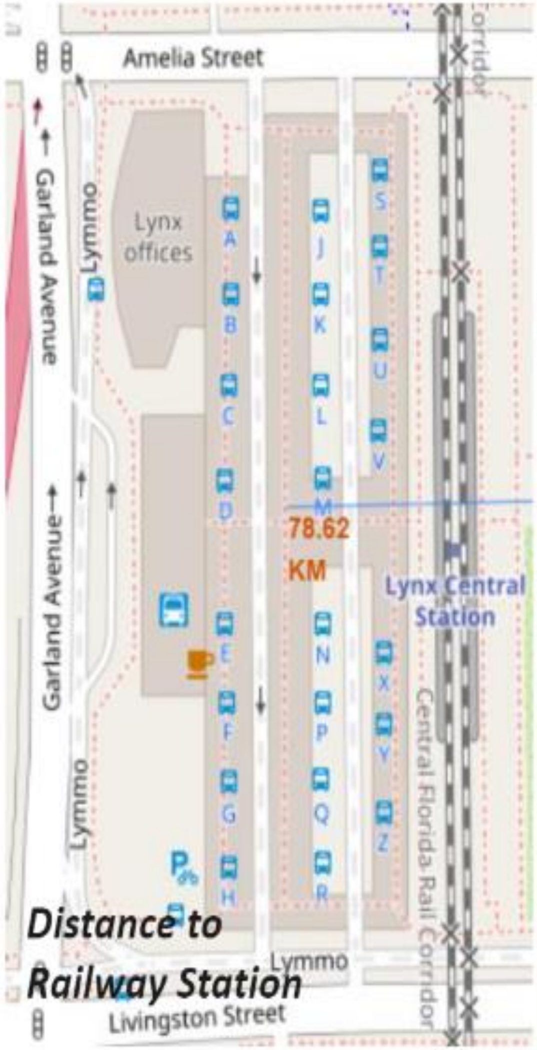
Florida Launch Sites

Green Marker shows successful Launches and Red Marker shows Failures

California Launch Site

37

Launch Site distance to landmarks

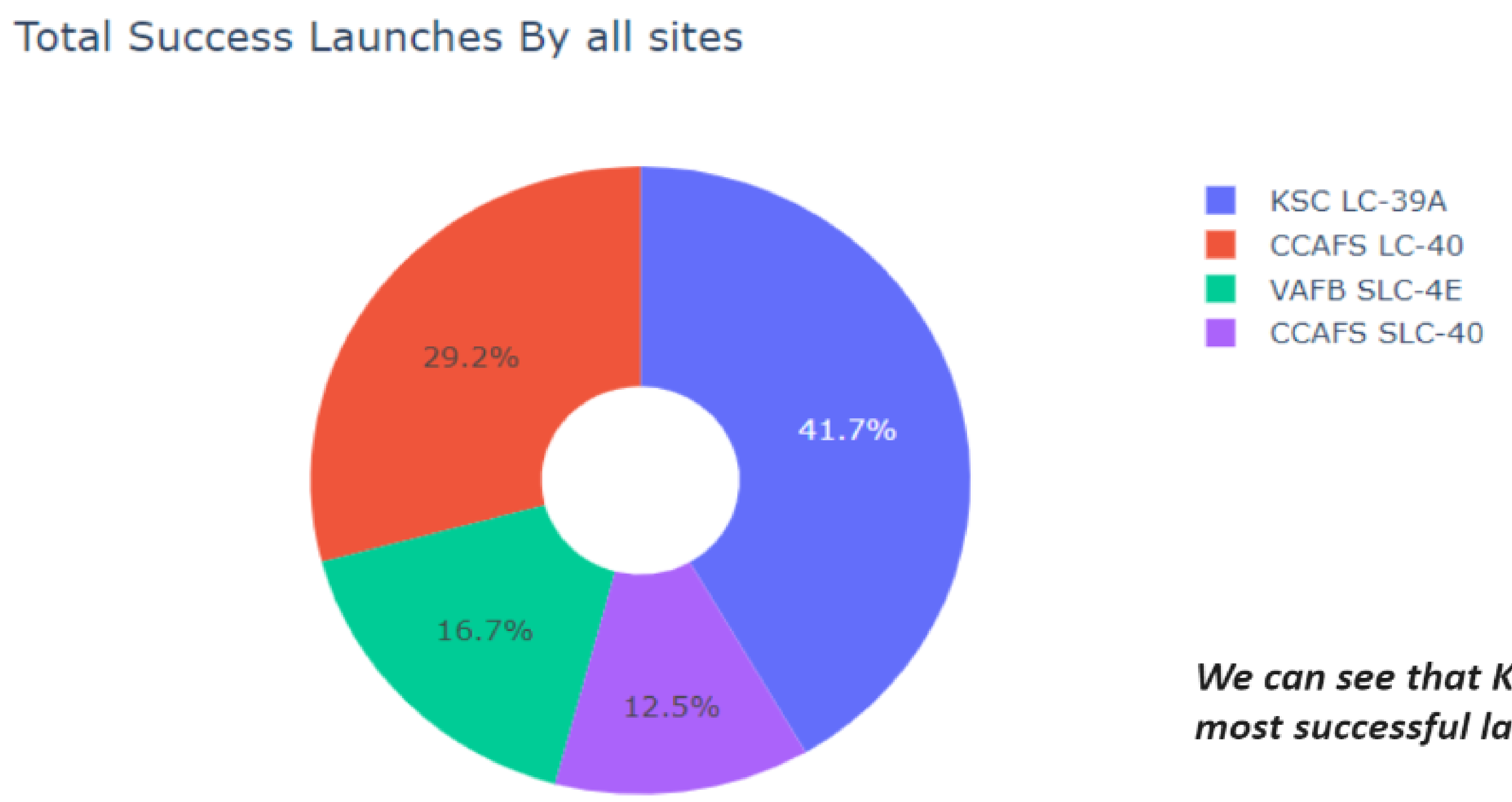


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 5

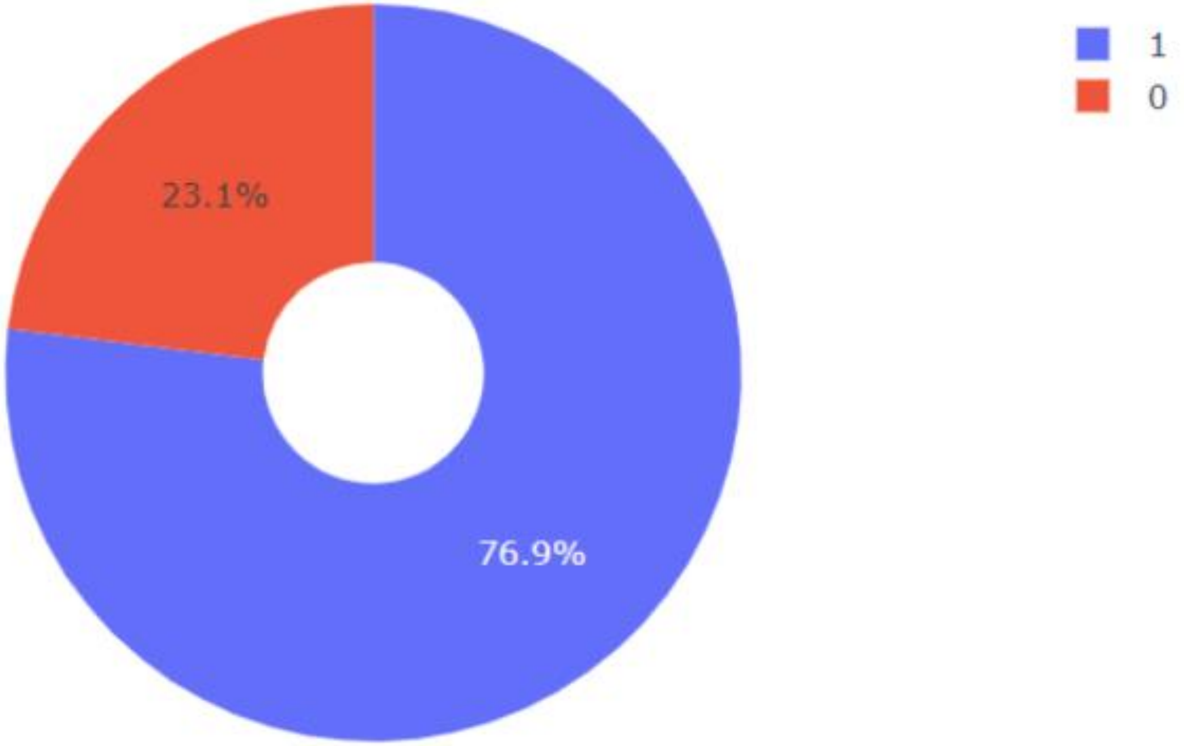
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



We can see that KSC LC-39A had the most successful launches from all the sites

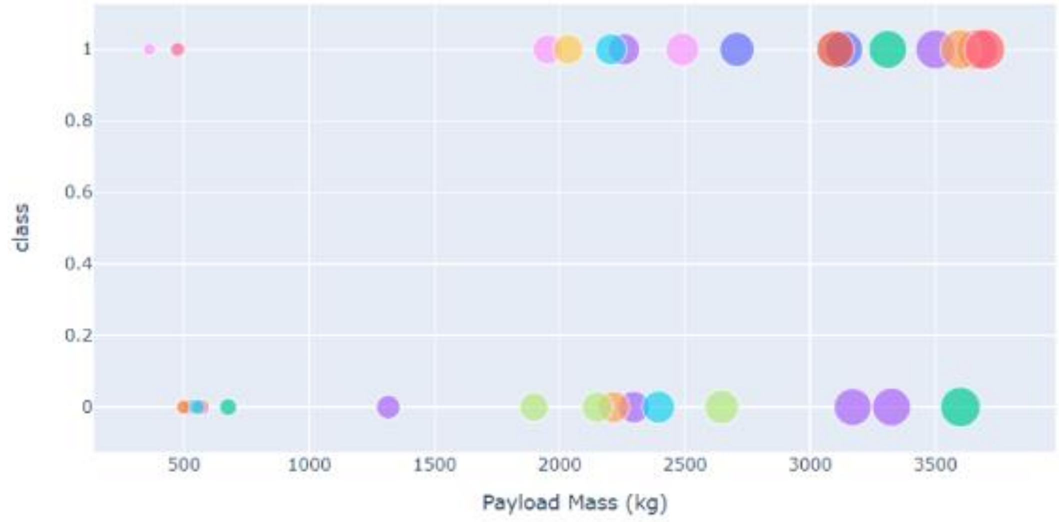
Pie chart showing the Launch site with the highest launch success ratio



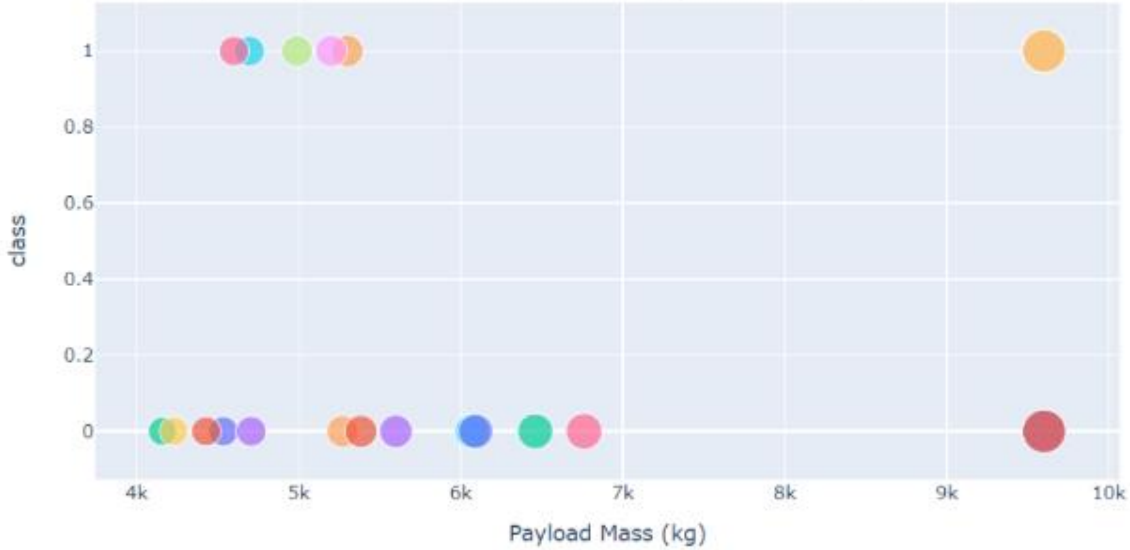
KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

Low Weighted Payload 0kg – 4000kg



Heavy Weighted Payload 4000kg – 10000kg



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

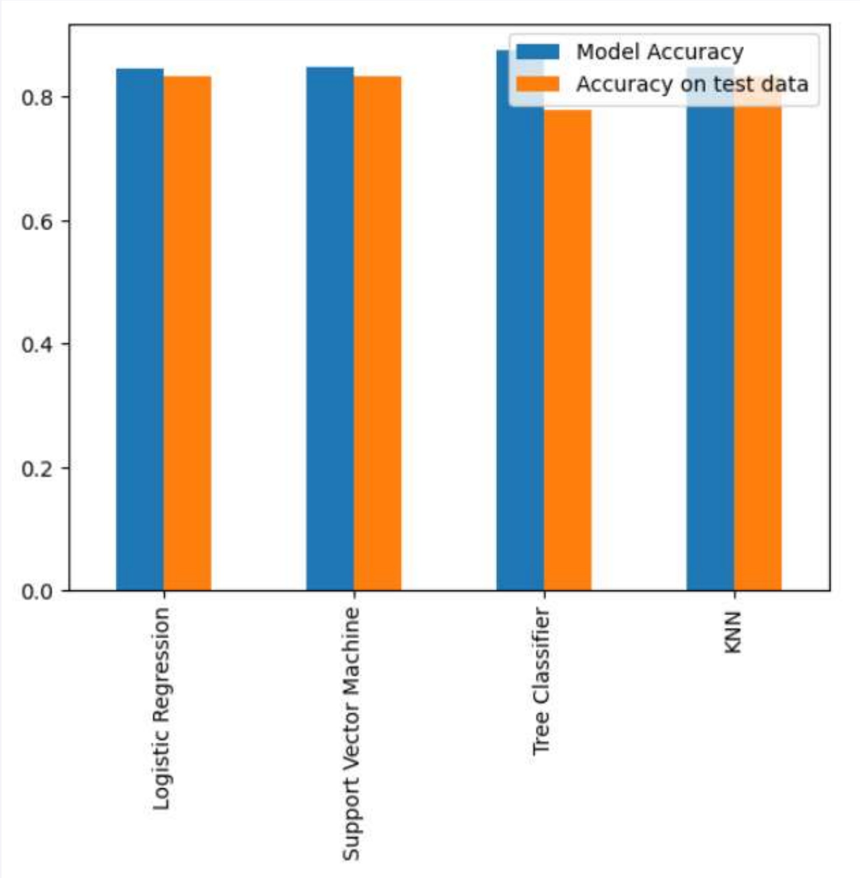
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Model accuracy and accuracy on test data is shown on the bar graph for each Model that has been tested

- Decision Tree Classifier is the best model with Model accuracy of 0,875 and Accuracy on Test data of 0,78

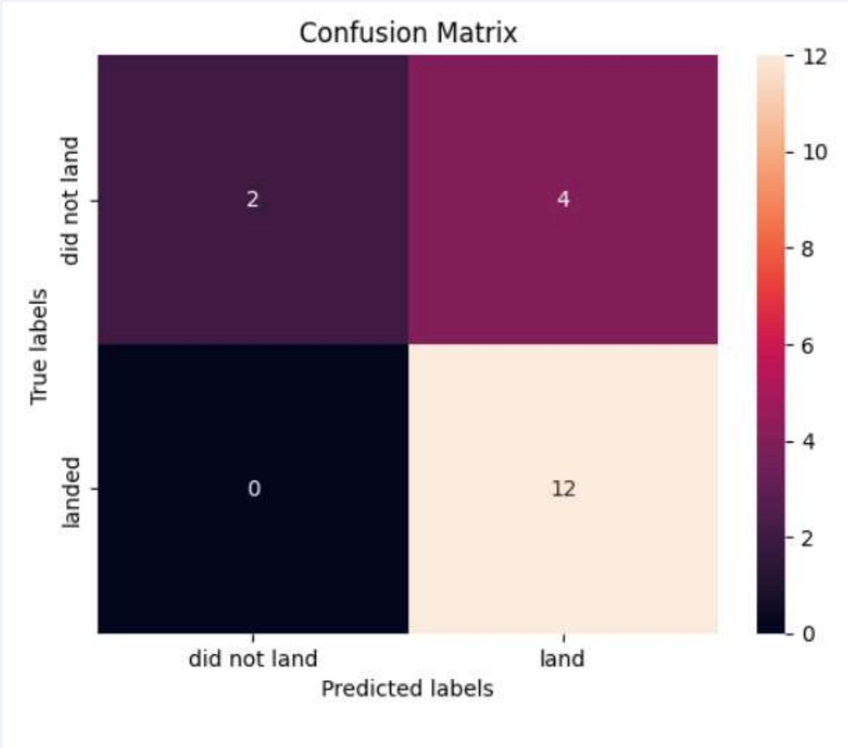


Confusion Matrix

- Confusion matrix of the best performing model Decision Tree Classifier
- Parameters of the Best Classifier are:

```
Best estimator found: DecisionTreeClassifier(criterion='entropy', max_depth=6, max_features='sqrt',
                       min_samples_split=10, splitter='random')
```

- It can distinguish between different classes quite well. False Positive and Fase Negative are 4 and 0, so it is a good predictor for the 1st stage landing success



Conclusions

• Space X uses 4 Launch sites. The 3 ones in Florida are closer to Equator Line and have better proximities that the one in Califormia

• CCAFS SLC 40 was the first Launch test at the beginning of the Space X history so the success rate is rather small. VABF SLC-4E in California does not operate payload above 9t because it is farther from the Equator line. KSC LC-39A is the best Launch site with 77% of 1st stage landing success.

• High altitude orbit (e.g. GTO) and payload in the range 4000-6000 kg have not so good success rate

• V1.0 and v1.1 booster version did not perform well enough whereas FT booster and B4 booster show better performance especially starting from 2017/Flight Number 60. Booster B5 is reserved for heavy payload (15 tons) and performs well.

• With all the available data, we have built a Decision Tree Classifier that predicts the landing outcome with good confidence. It can be used by Space Y to decide on a bid against SpaceX.


Thank you!