IBM data science capstone project - SpaceX

REYO ELIZABETH JOHN

01-08-2021

Outline

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

EXECUTIVE SUMMARY

- **Summary of methodologies**
- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with folium
- Building a Dashboard with Poltly dash
- Predictive analysis (Classification)
- **Summary of all results**
- Exploratory data analysis results
- Interactive analytics
- Predictive analysis

INTRODUCTION

- Project background and context

We have predicted if the Falcon 9 stage 1 will land successfully.SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.Our job is to determine the price of each launch.

- Problems we want to find answers

What influences if the rocket will land successfully?

The effect each relationship with certain rocket will ipmact in determining the success rate of a successful landing.

What conditions does Spacex have to achieve to get the bbest results and ensure the best rocket scuucess landing rate.

METHODOLOGY

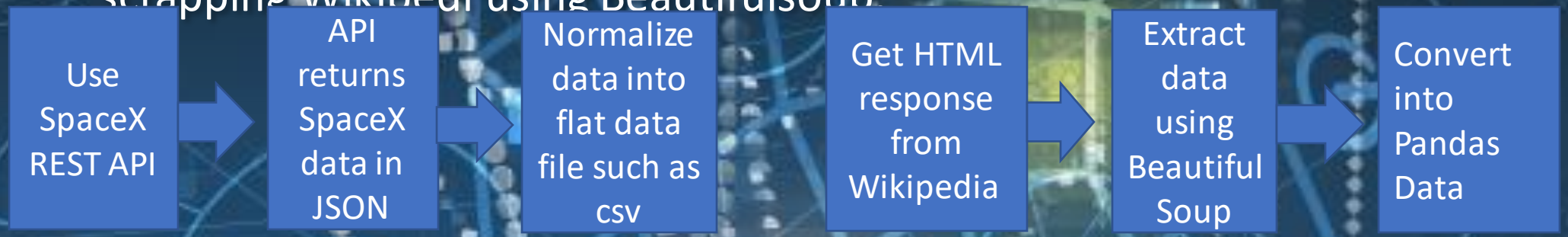
METHODOLOGY

- Data collection methodology:
  - SpaceX Rest API
  - (Web Scrapping) from Wikipedia

- Performed data wrangling(Transforming data for Machine learning)
  - One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- Perfomed expolratory data analysis(EDA) using visualization and SQL
  - Plotting: Scatter Graphs, Bar Graohs to show relationships between variables to show patterns of data.
- Perfomed interactive visual analytics using Folium and Plotly Dash
- Performed predictive anlysis using classification models
  - Models were build using scikit-learn, we used a GridSearch with 10 fold of cross validation, in the end we use the accuracy to determine the best classifier

DATA COLLECTION

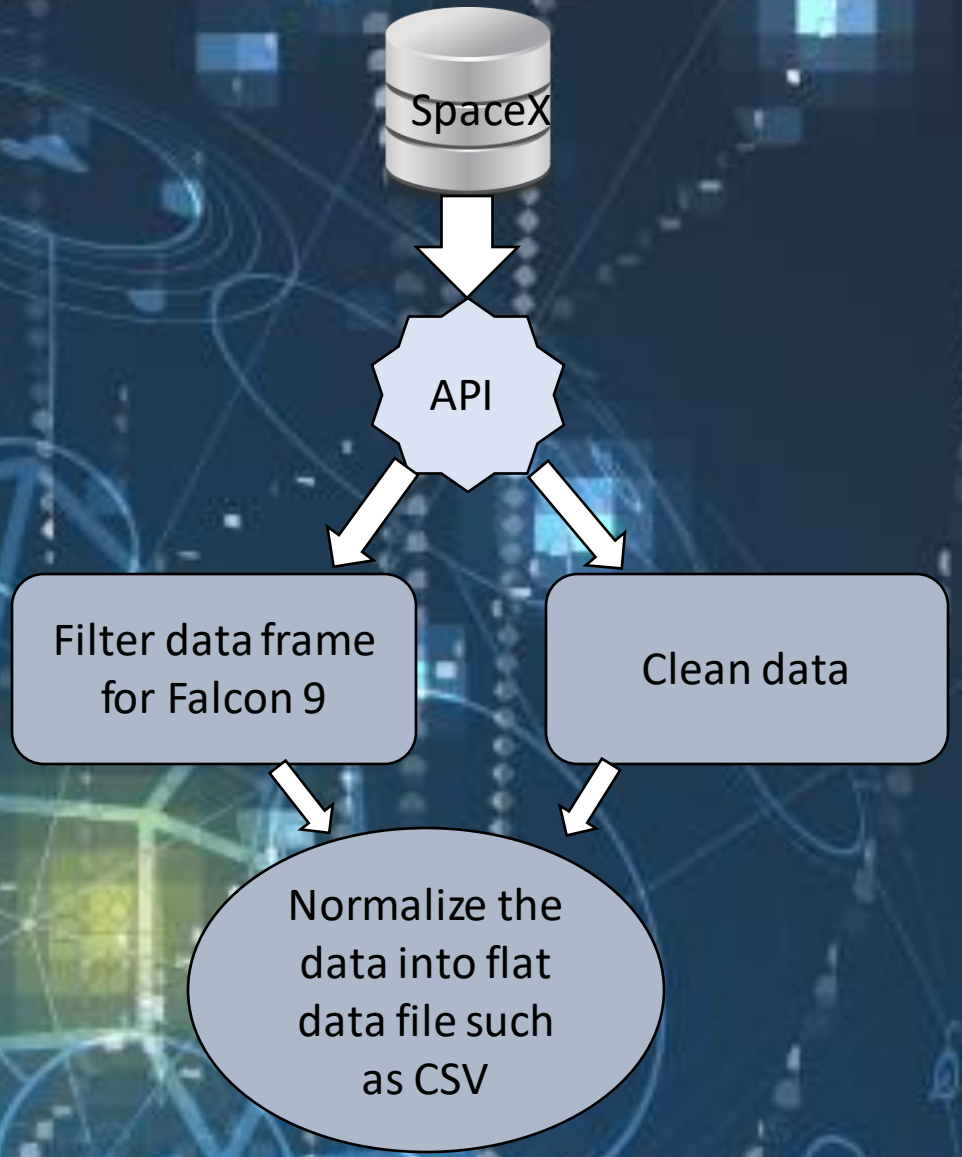
- The folowing datasets was collected:
  - We worked with SpaceX launch data that is gathered from the SpaceX REST API.
  - This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications and landing outcome.
  - The SpaceX REST API endpoints , or URL, starts with api.spacexdata.com/v4/.
  - Another popular data source for obtaining Falcon 9 Launch data is web scrapping Wikipedi using Beautifulsoup



Data Collection –SpaceX API

We collect the data from SaceX REST API , then filter and clean the data set for Falcon 9 and normalize the set .

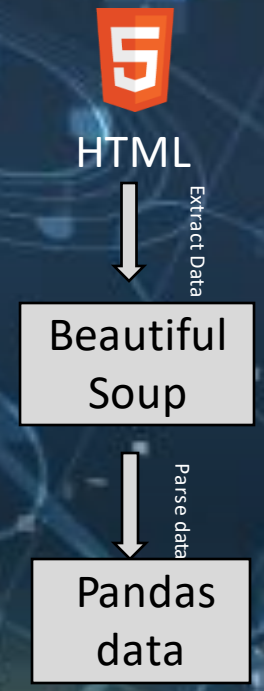
GitHub URL



Data Collection –Web Scraping

We have used BeautifulSoup to web scrap HTML data from Wikipedia

And parse the data into Pandas Dataframe

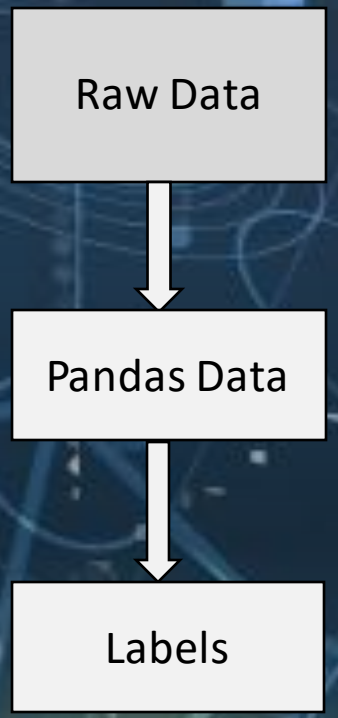


GitHub URL

Data Collection –Wrangling

We have used BeautifulSoup to web scrap HTML data from Wikipedia and parse the data into Pandas Dataframe.

Populate the table with rows collected.



GitHub URL

EDA with Data Visualization

- We used:
- A scatter plot of FlightNumber vs. PayloadMass based on the outcome to see if as the flight number increases,
- The first stage is more likely to land successfully and if there is an impact of PayloadMass
- Then a second scatter plot of LaunchSite vs. FlightNumber based on the outcome to understand if the LaunchSite does not become a problematic element following the number of rockets which are launched there (FlightNumber).
- A third scatter plot to observe if there is any relationship between LaunchSite and their PayloadMass
- A bar chart to visually check if there are any relationship between the *Success Rate* and the *Orbit Type*
- Another scatter plot to visualize the relationship between FlightNumber and *Orbit type*
- A last scatter plot to reveal the relationship between *Payload* and *Orbit type*
- And a line plot to get the average launch success trend

- GitHub URL

EDA with SQL

SQL queries were performed as followed :

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster_versions which have carried the maximum payload mass
- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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
- GitHub URL

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took latitude and longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the data frame launch_outcomes to classes 0 and 1 with Green and Red markers on the map in MarkerCluster().
- Using Haversine's formula we calculated the distance from Launch site to various landmarks to find various trends about what is around the Lauch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks.

GitHub URL

Build a Dashboard with Plotly Dash

- **Pie Chart showing the total launches by a certain site/all**
  - Helps us to understand the success rate on eah launching site
- **Scatter Graph showing the relationship with Outcome and Payload Mass(Kg) for the different Booster Versions**
  - Demonstrates which booster as the higher rate and that it is related to the Payload mass

GitHub URL

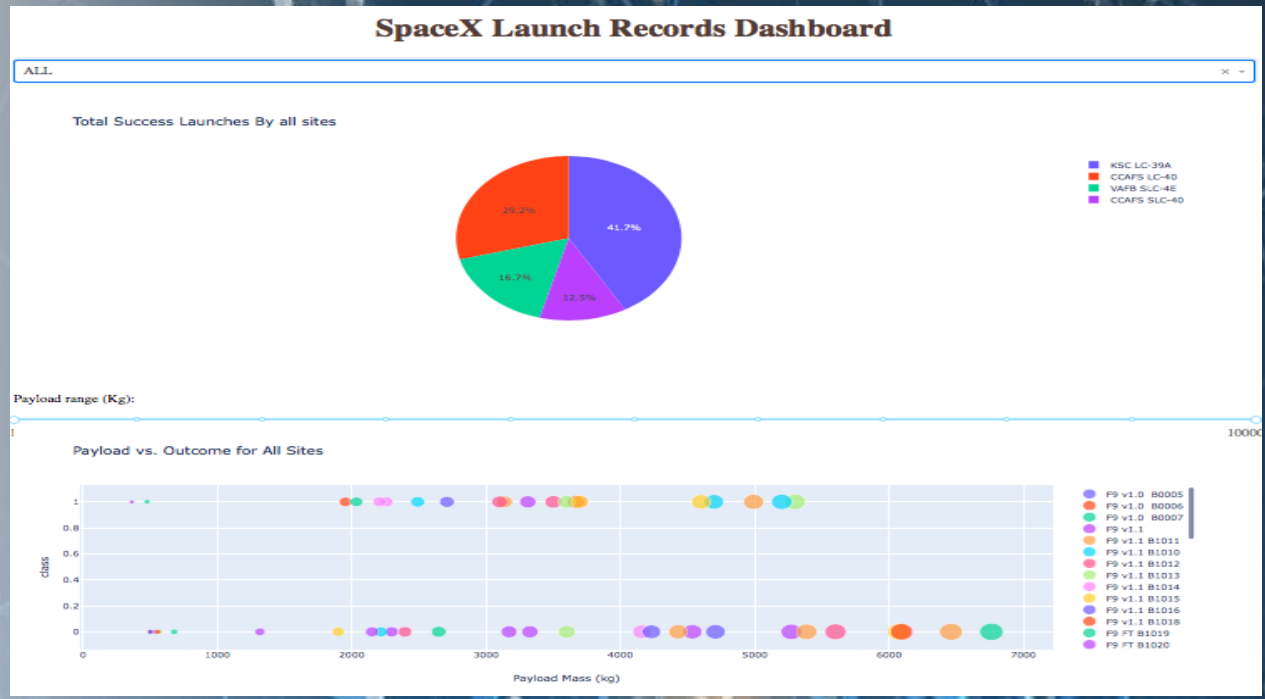
Predictive analysis(Classification)

- **Building Model**
- Load our dataset into Numpy and Pandas
- Transform Data
- Split our data into training and test data sets
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchV
- Fit our datasets into the GridsearchCV
- **Evaluating Model**
- Check accuracy for each model
- Check accuracy of test data
- Plot Confusion Matrix
- **Improving Model**
- Feature Engineering
- **Finding the best performing Classification Model**
- The model with the best accuracy score wins the best performing model

GitHub URL

Results

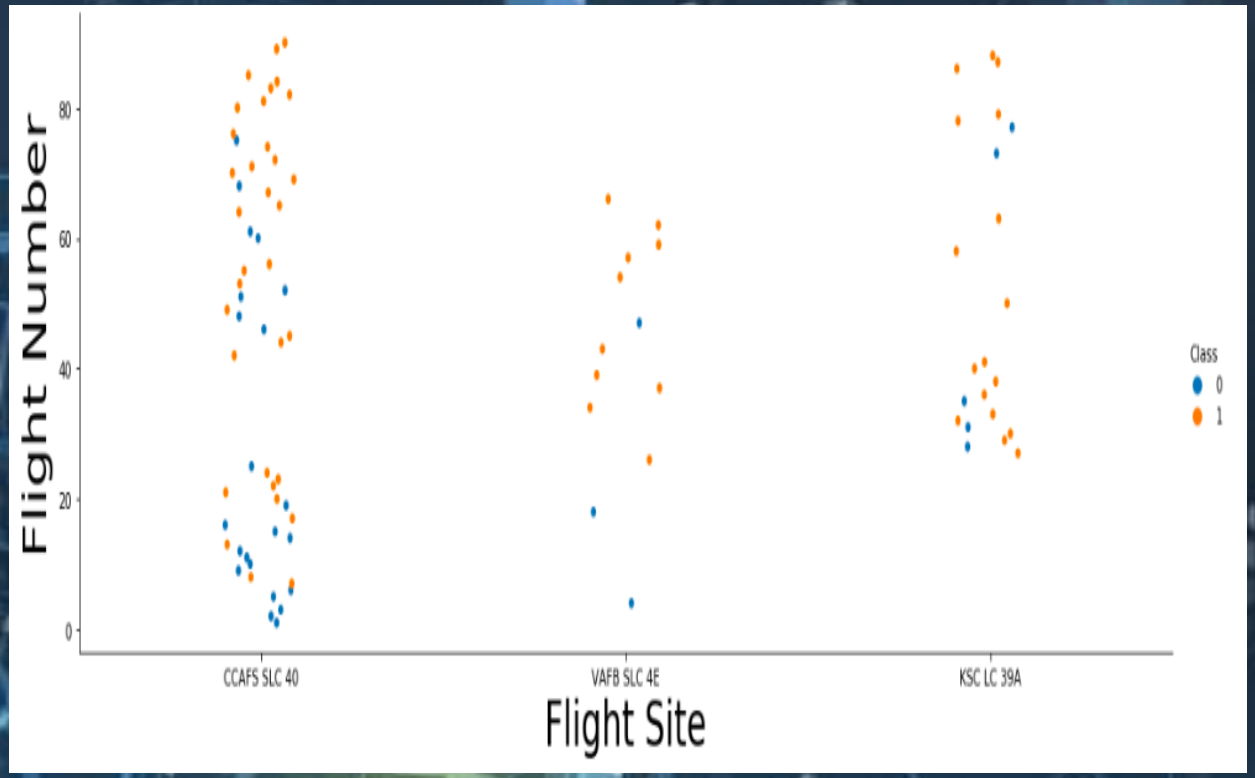
- The EDA demonstrate the importance of the payload mass and also the booster version on the success rate of the launch missions.
- All machine learning models studied were equally performing.



Insights Drawn from EDA

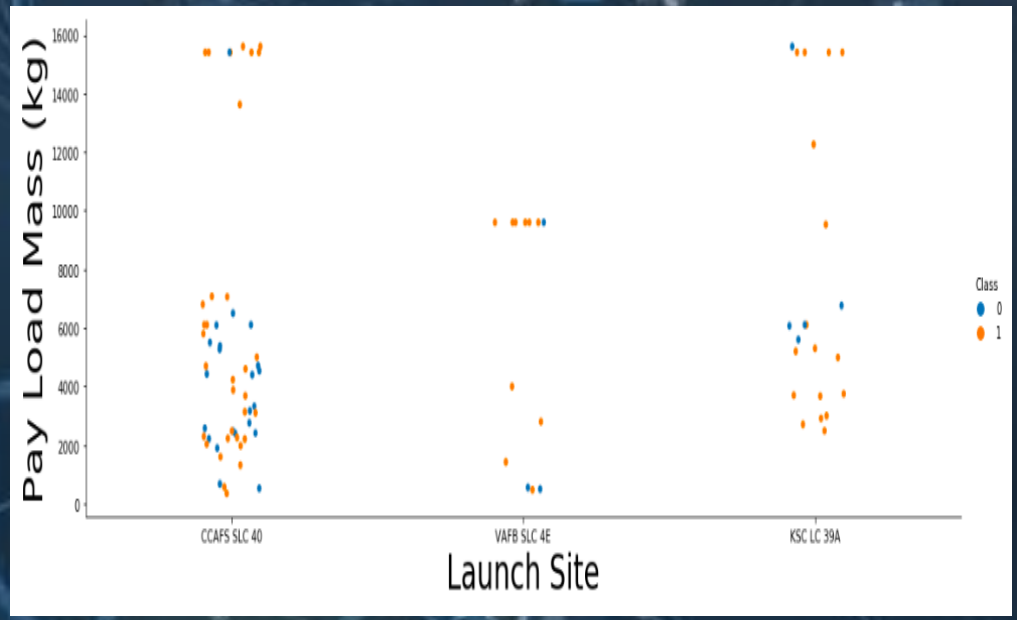
Flight number vs. Launch site

The success rate increases with the number of flights



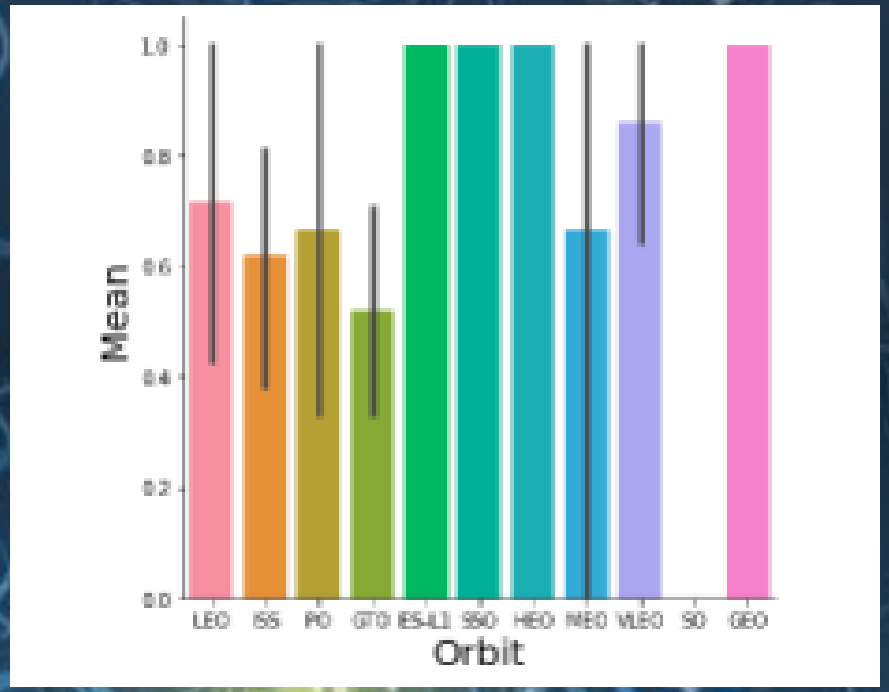
Payload mass vs. Launch site

The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket. There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependant on Pay Load Mass for a success launch.



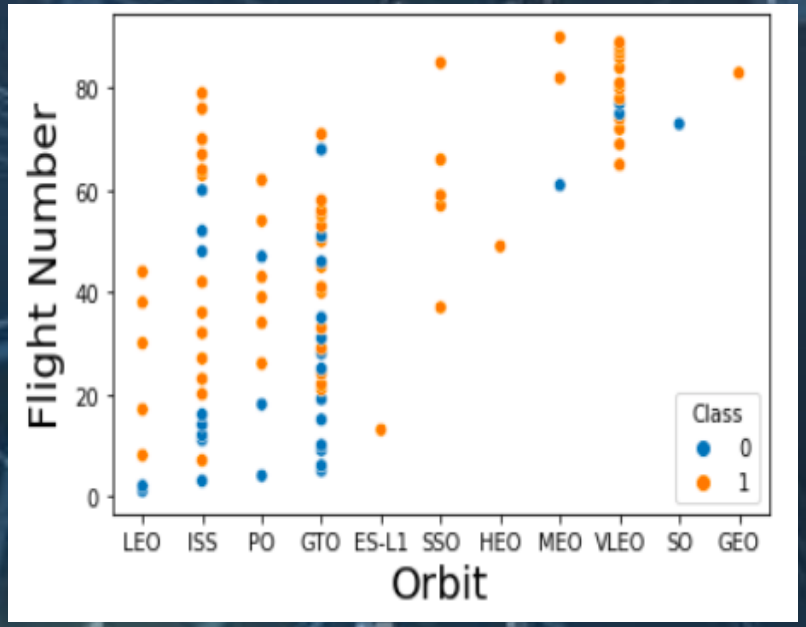
Succcess Rate vs. Orbit Type

Orbits ES-L1,SSO,HEO and GEO has the best success rate



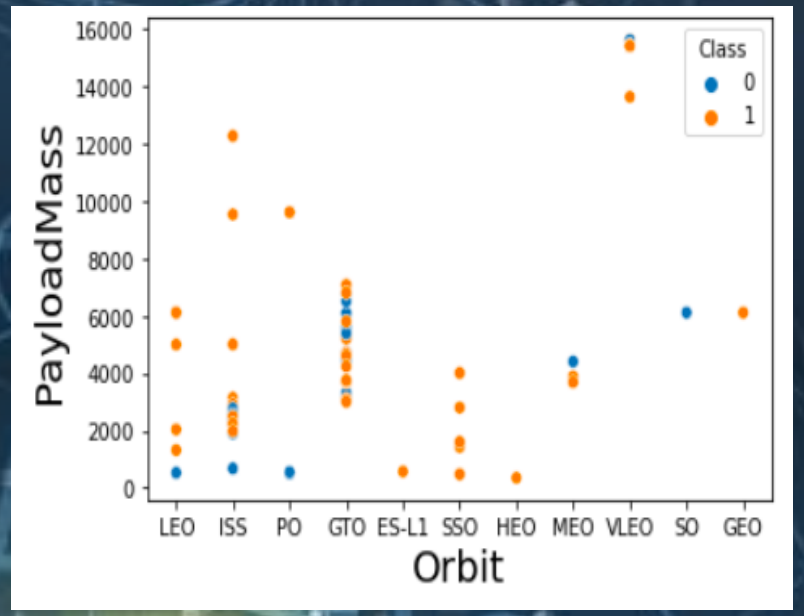
Flight Number vs. Orbit Type

We can 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 mass vs. Orbit Type

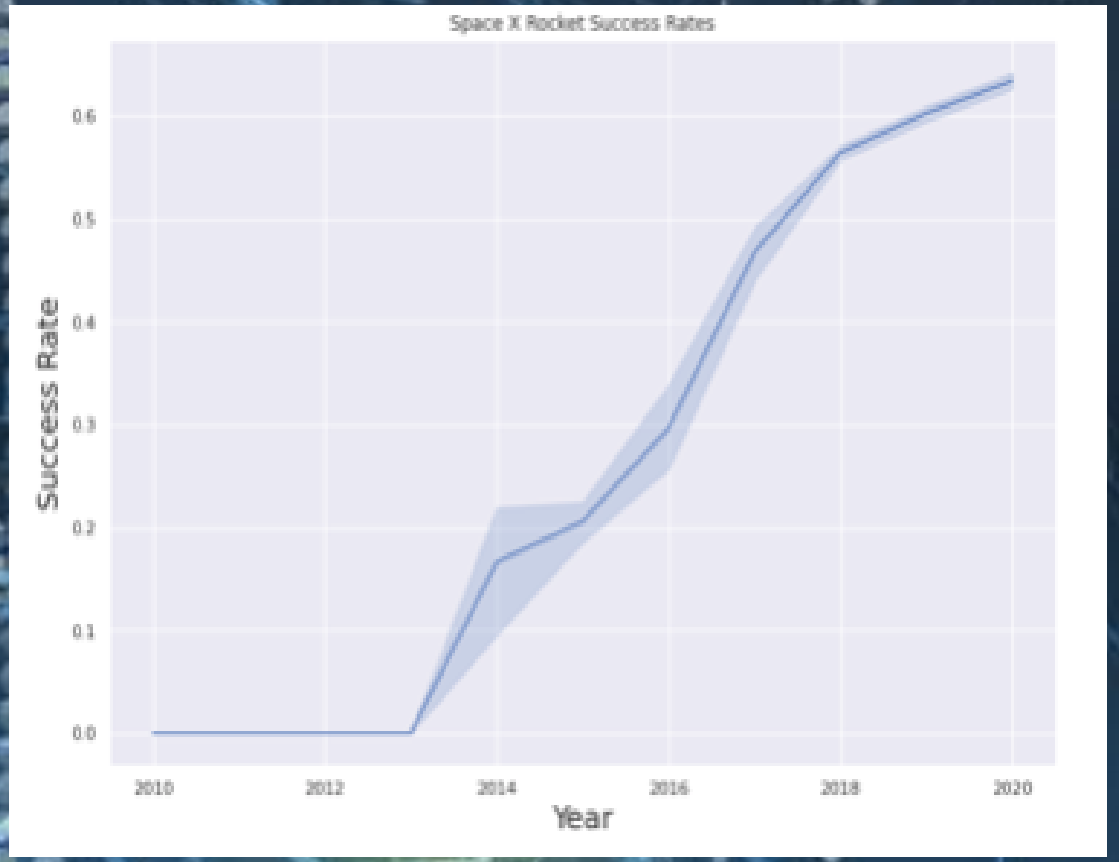
Payload mass have a strong influence on some sites. For example, GTO orbits have a negative influence and for LEO orbits payload mass have posituve influence.

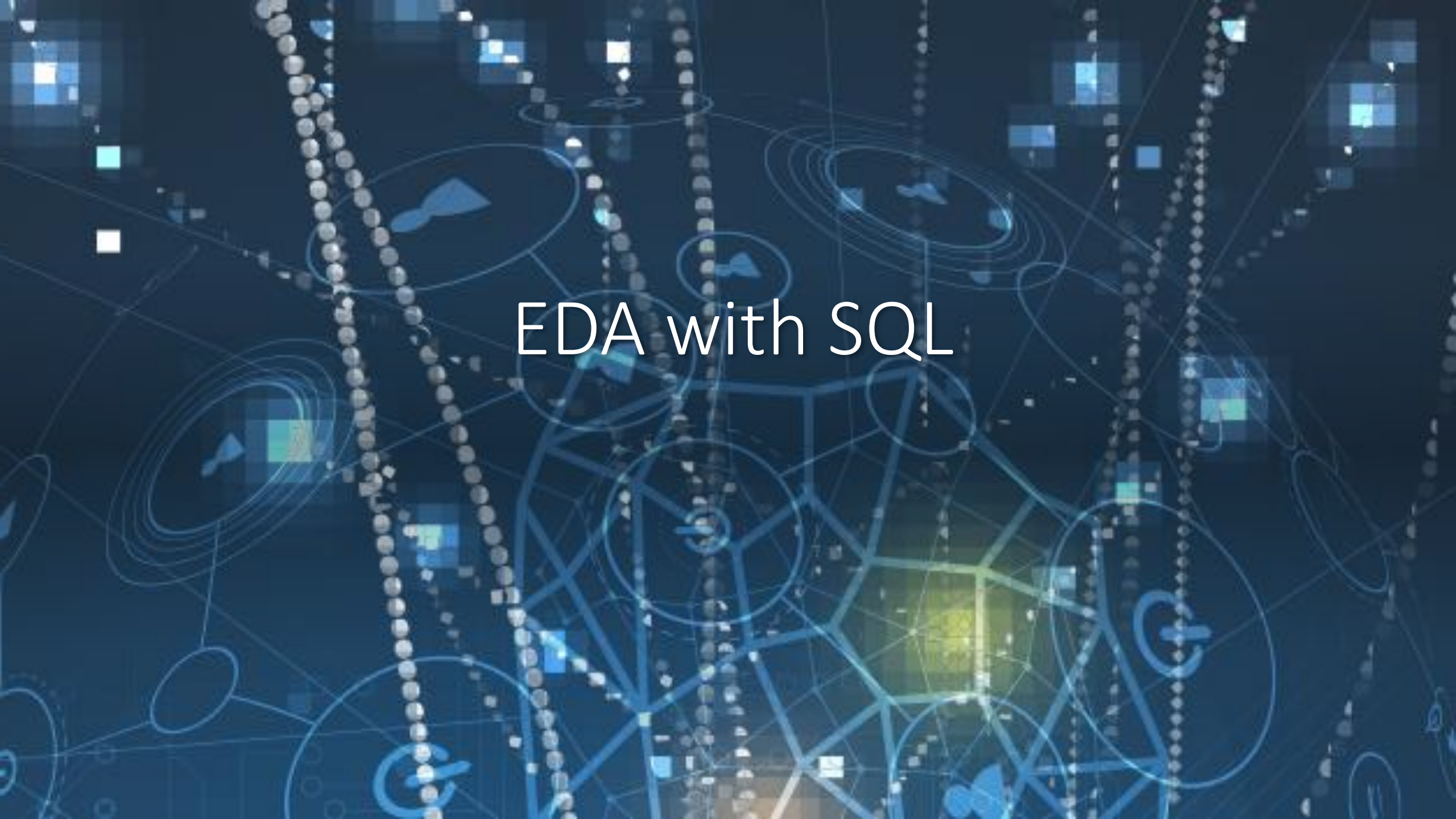


Launch Success Yearly Trend

The success rate increases since

2013




EDA with SQL

All Lauch Site Names

- Launch sites are :
  - CCAFS LC-40
  - CCAFS SLC-40
  - CCAFSSLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- SQL Query:
  - Using the word DISTINCT in query to show the unique values in the Launch_site column

Launch Site Names Begin with 'CCA'

- 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

- SQL Query
- Select all columns from SPACEXTBL where LAUNCH_SITE like 'CCA' and display limited to 5 rows.

Total Payload Mass

- The total payload carried by boosters from NASA is 99980 KG.

- SQL Query
  - Using the function SUM ,we get the total in the column PAYLOAD_MASS__KG_
  - The WHERE clause filters the dataset to only perform calculation on Customer NASA

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is:

  2534 KG

- SQL Query
  - Using the function AVG ,we get the average in the column PAYLOAD_MASS__KG_
  - The WHERE clause filters the dataset to only perform calculation on Booster version F9 v1.1

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was achieved in 2015-12-22

- SQL Query
  - Using the function MIN,we get the minimum date in the column DATE
  - The WHERE clause filters the dataset to only perform calculation on Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are:
  - F9 FT B1032.1
  - F9 B4 B1040.1
  - F9 B4 B1043.1
- SQL Query
  - SELECT only the BOOSTER_VERSION column
  - The WHERE clause filters the dataset to LANDING__OUTCOME like 'Success (ground pad)' **and** PAYLOAD_MASS__KG_ between 4000 **and** 6000

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes :

Mission Outcome	Success or Failure
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- SQL Query
  - SELECT only the BOOSTER_VERSION column
  - The WHERE clause filters the dataset to LANDING__OUTCOME like 'Success (ground pad)' **and** PAYLOAD_MASS__KG_ between 4000 **and** 6000

2015 Launch Records

- List of 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

- SQL Query
  - Select only LANDING__OUTCOME,BOOSTER_VERSION and LAUNCH_SITE
  - The WHERE clause filters the dataset to DATE like '%2015%' and LANDING__OUTCOME like '%Failure'

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

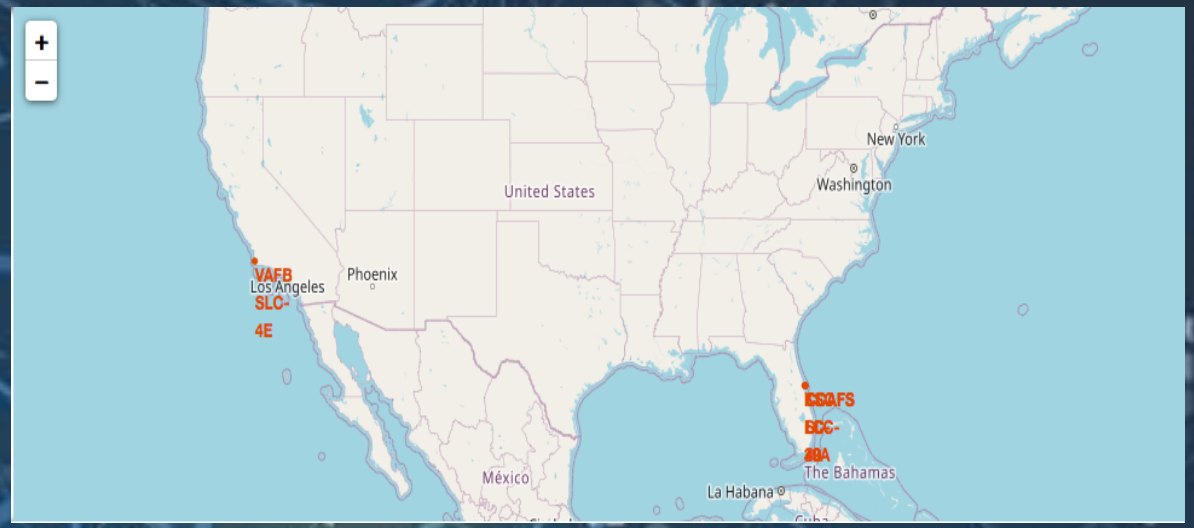
Landing Outcome	No attempt	Failure(drone ship)	Success(drone ship)	Controlled(ocean)	Success(ground pad)	Failure(parachute)	Uncontrolled(ocean)	Precluded(drone ship)
Count	10	5	5	3	3		2	1

- SQL Query
  - Select the distinct LANDING__OUTCOME and count it.
  - The WHERE clause filters the dataset to DATE between '2010-06-04' and '2017-03-20' and groupby LANDING__OUTCOME in descending order using ORDER BY.

Launch Sites Proximities Analysis

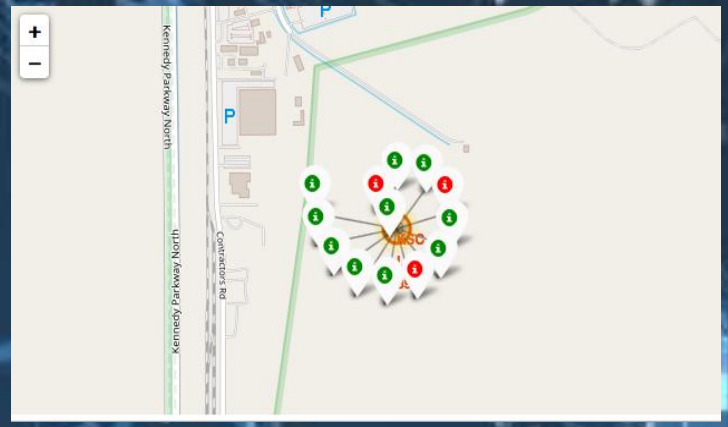
Launch sites in the US

We can see that the SpaceX launch sites are in the United States of Ameica's coasts, Florida and California



Success or Failure Marker

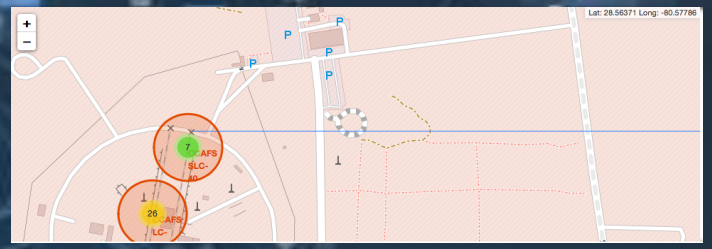
Green Marker shows successful Launches and Red Marker shows failures



Launch sites proximities

We can see the proximity of the launching site and the railway station and proximities with the coastline,city and closest Highway

Distance to nearest city



Distance to railway station

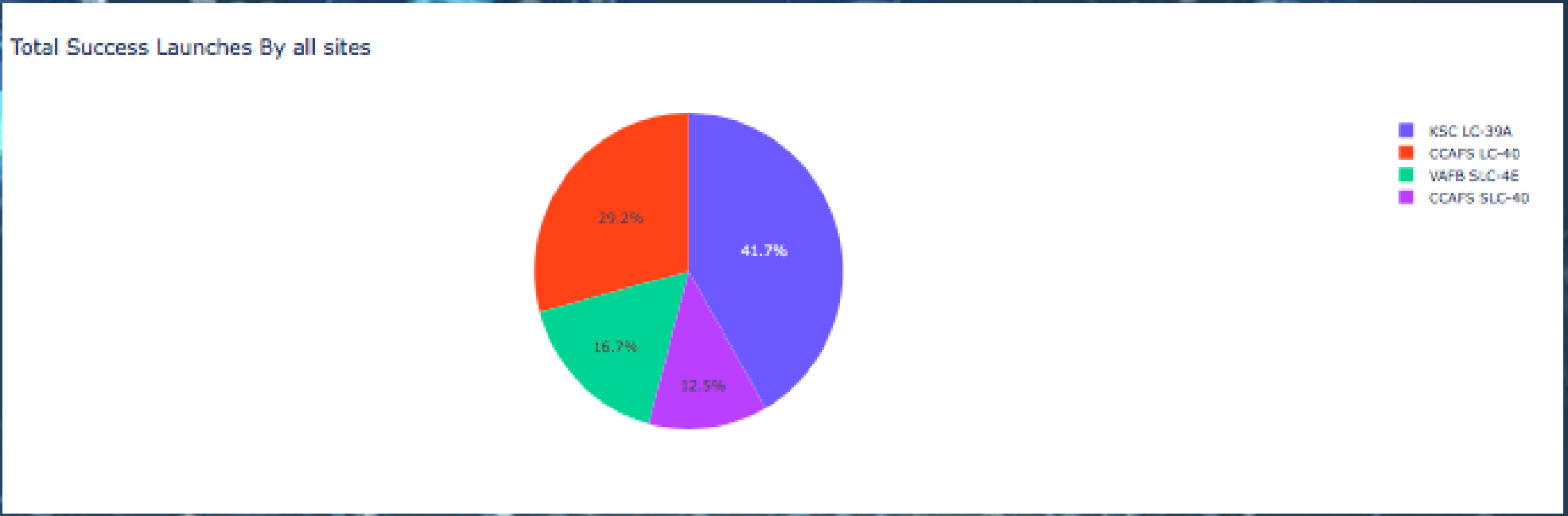
Distance to coastline

Distance to nearest highway


Building a Dashboard with
Plotly Dash

Success rate of all sites

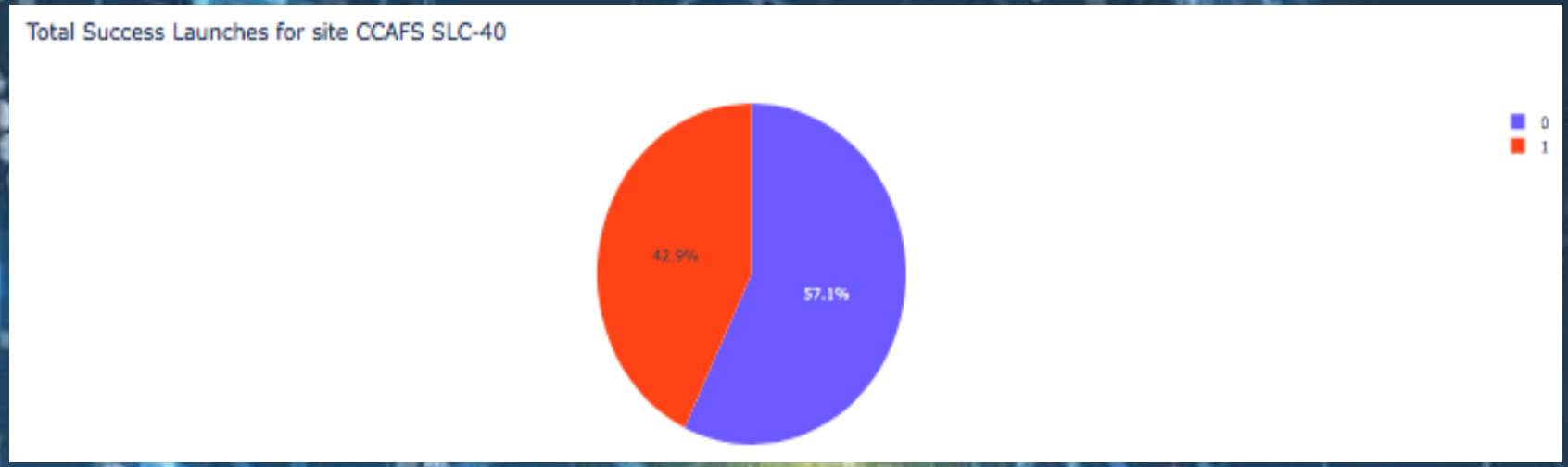
We can see that KSC LC –39A has the most successful launches among all the sites.



Success rate of CCAFA SLC-40 best ratio

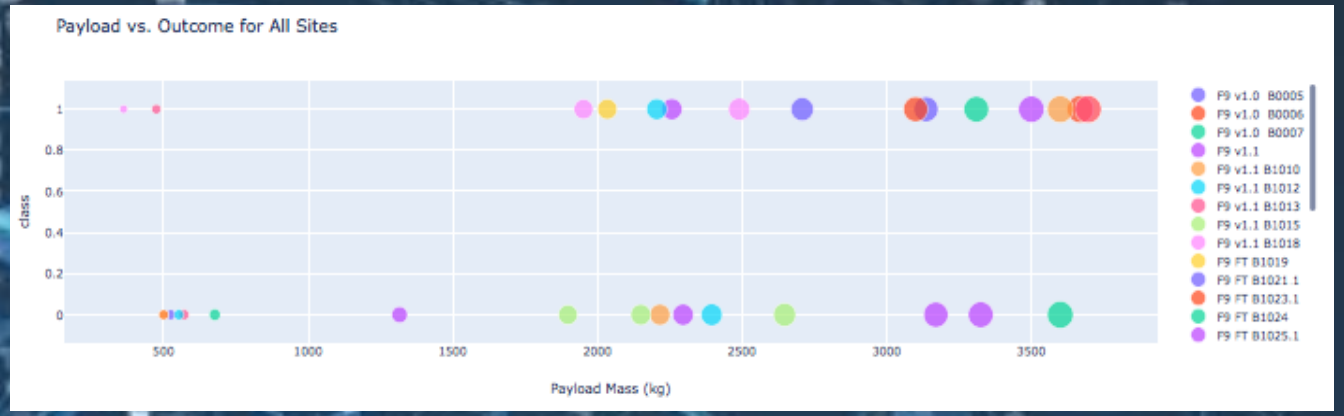
We can see that CCAFA SLC-40

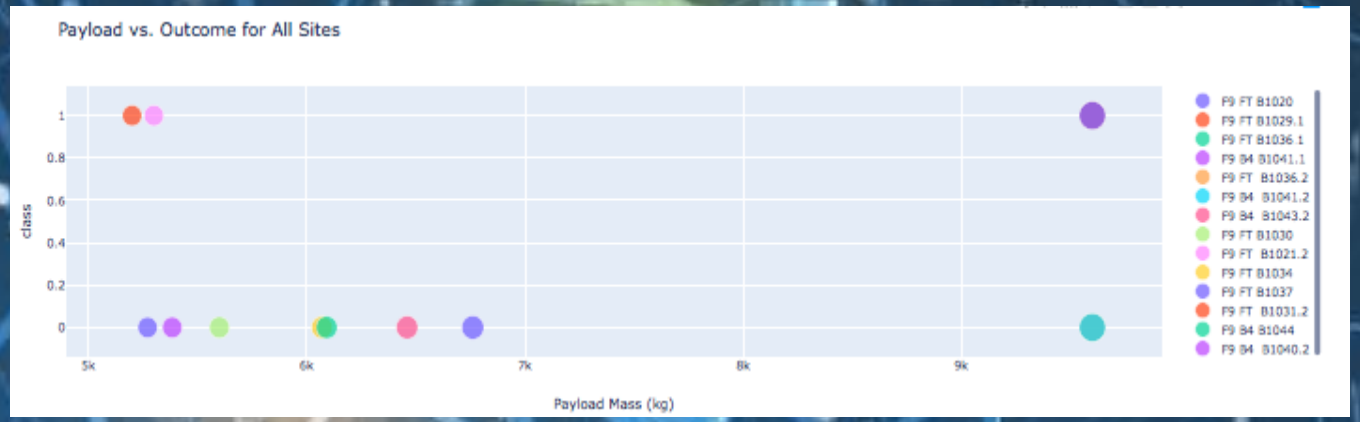
has 42.9% successful launches



Payload vs. Launch Outcome scatter plot for all sites

It is clear from the scatter plots that success rate is high for low weighted pay loads than heavy payloads

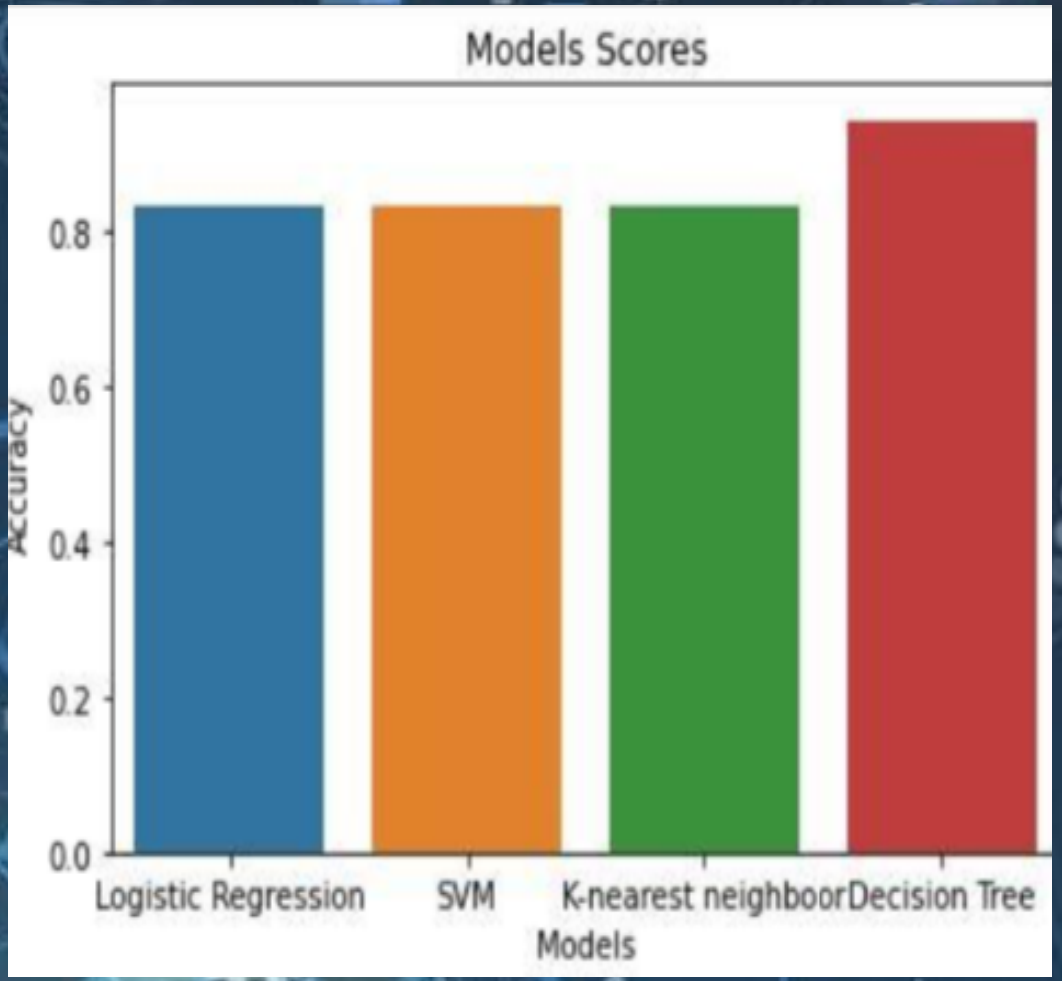





Predictive Analysis
(Classification)

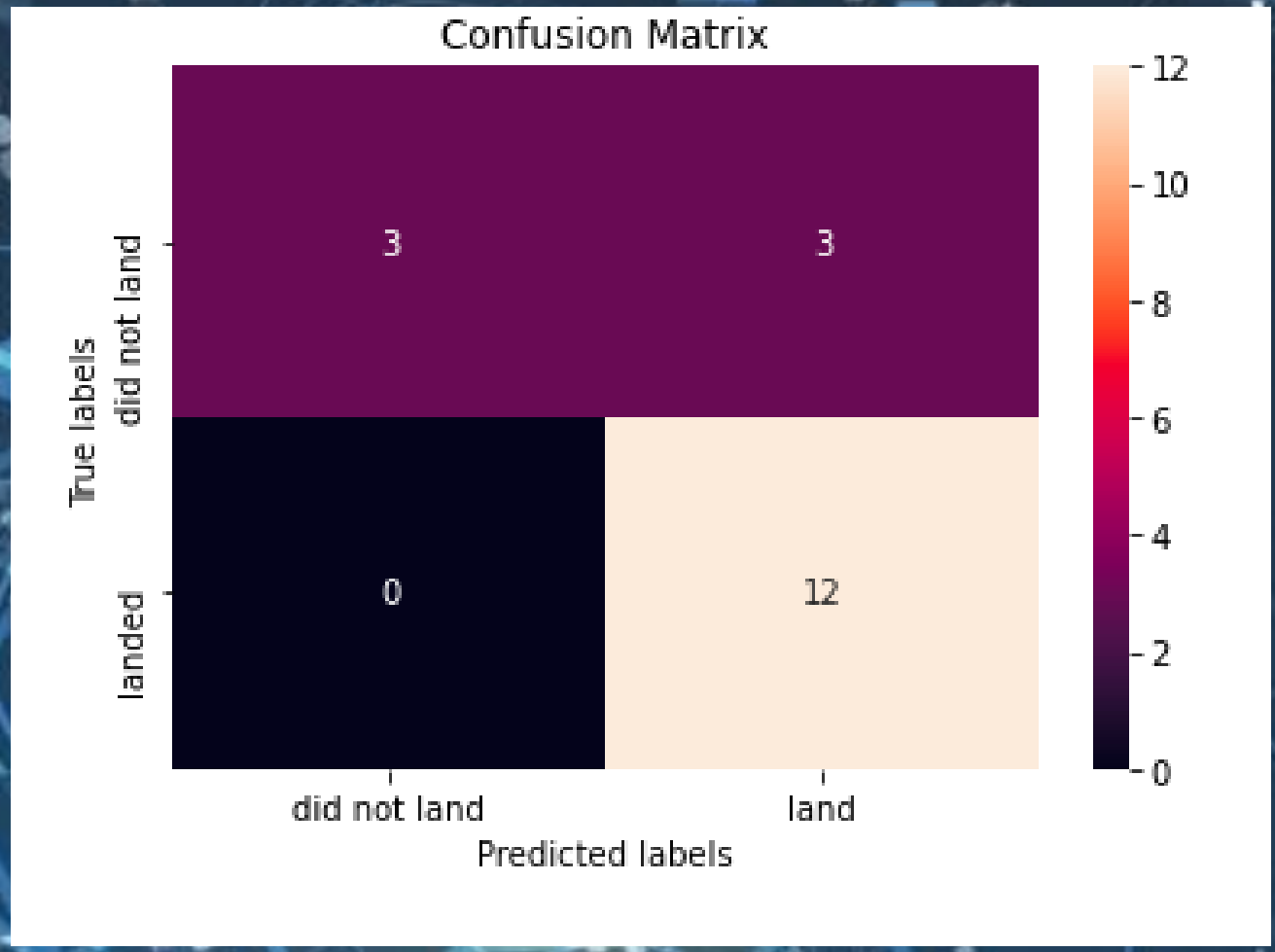
Classification Accuracy

The best model based on accuracy is a Decision Tree Classifier with a score of 0.9166



Confusion Matrix

This summarizes the performance of classification algorithm. Examining the confusion matrix, we can distinguish between different classes. We see that thr major problem is false positive.



Conclusion

- All algorithms performs equally, since the data set is small.
- Payloads with low weight performs better than that with heavy weights.
- The success rate increses each year since 2013 .
- KSC LC –39A has the most successful launches from all the sites.
- Orbit ES-L1,SSO,HEO and GEO have more success rate than others

Appendix

- Haversine formula

Haversine formula

Haversine formula is used to calculate the distance between two GPS coordinates with Python

Formula:

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi 1)\cdot\cos(\phi 2)\cdot\sin^2\left(\frac{\Delta\lambda}{2}\right)$$

With:

$$c = 2\cdot \arctan 2\left(\sqrt{a}, \sqrt{1-a}\right)$$

and

$$d = R\,.\,c$$

Where:

ϕ = latitude

λ = longitude

R = 6371 (Earth's mean radius in km)


Thank You!