Winning Space Race with Data Science

RAAVI NAGA VENKATA SURYA SAI TANMAI

25th November 2021

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection
  - Data Wrangling
  - EDA with data visualisation
  - EDA with SQL
  - Building interactive map with Folium
  - Dashboard with Plotly Dash
  - Predictive Analysis
- **Summary of all results**
  - Exploratory data analysis
  - Interactive Analytics
  - Predictive analysis result

Introduction

Project background and context

Currently, SpaceX is one of the most successful commercial space travel companies. This project aims to determine if the SpaceX's Falcon 9 first stage will land or not thereby indirectly estimate the total cost for the project. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars.

The information from this data is used by another company called Space Y whose aim is to put a competitive bid to SpaceX for Falcon9 Launches. As a data scientist, the price of each launch should be precisely calculated and create user-friendly dashboards for the team of Space Y.

Introduction

Problems you want to find answers

1. Parameters that influence a successful landing of first stage of Falcon9
2. Interrelationship between rocket launch related variables
3. Average price for a successful launch
4. Ideal conditions that SpaceX need to consider to launch a rocket

Section 1

Methodology

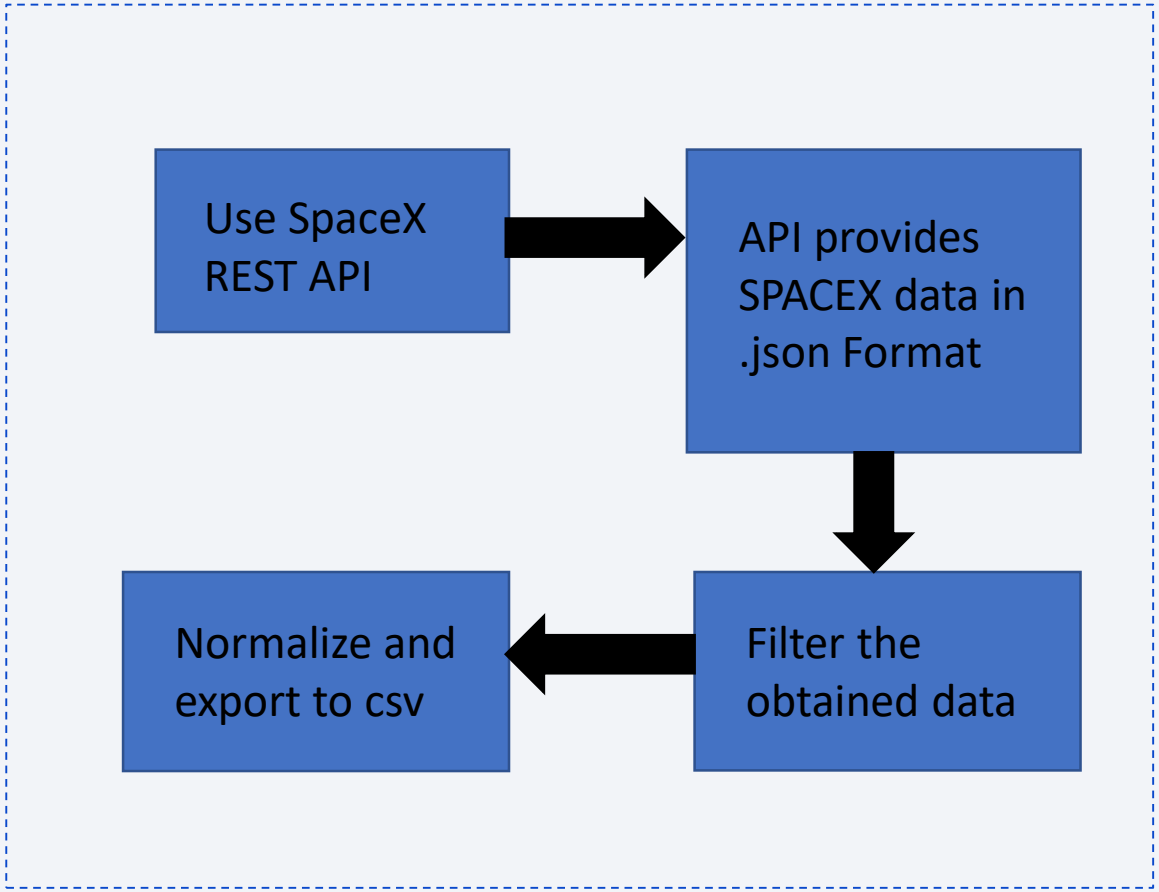
Methodology

Executive Summary

- Data collection methodology:
  - The SpaceX API
  - List of Falcon 9 and Falcon Heavy launches WikiPage
- Performed data wrangling
  - Initially the given data has been cleaned by dropping irrelevant features, removing and replacing NaN values. Then One Hot Encoding is applied to convert categorical data into numerical.
- Performed exploratory data analysis (EDA) using visualization and SQL
  - Used Data Visualization tools like scatter plots, bar graphs, heat maps etc to establish relationships between attributes.
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - The available dataset is split into test and train. Using the train dataset, the Machine learning models are trained on the SpaceX dataset. Then, their accuracy is predicted using various parameters like F-Score, Jaccard Score, etc with the help of test dataset. HyperParameter Tuning is then applied to increase the accuracy of the trained models.

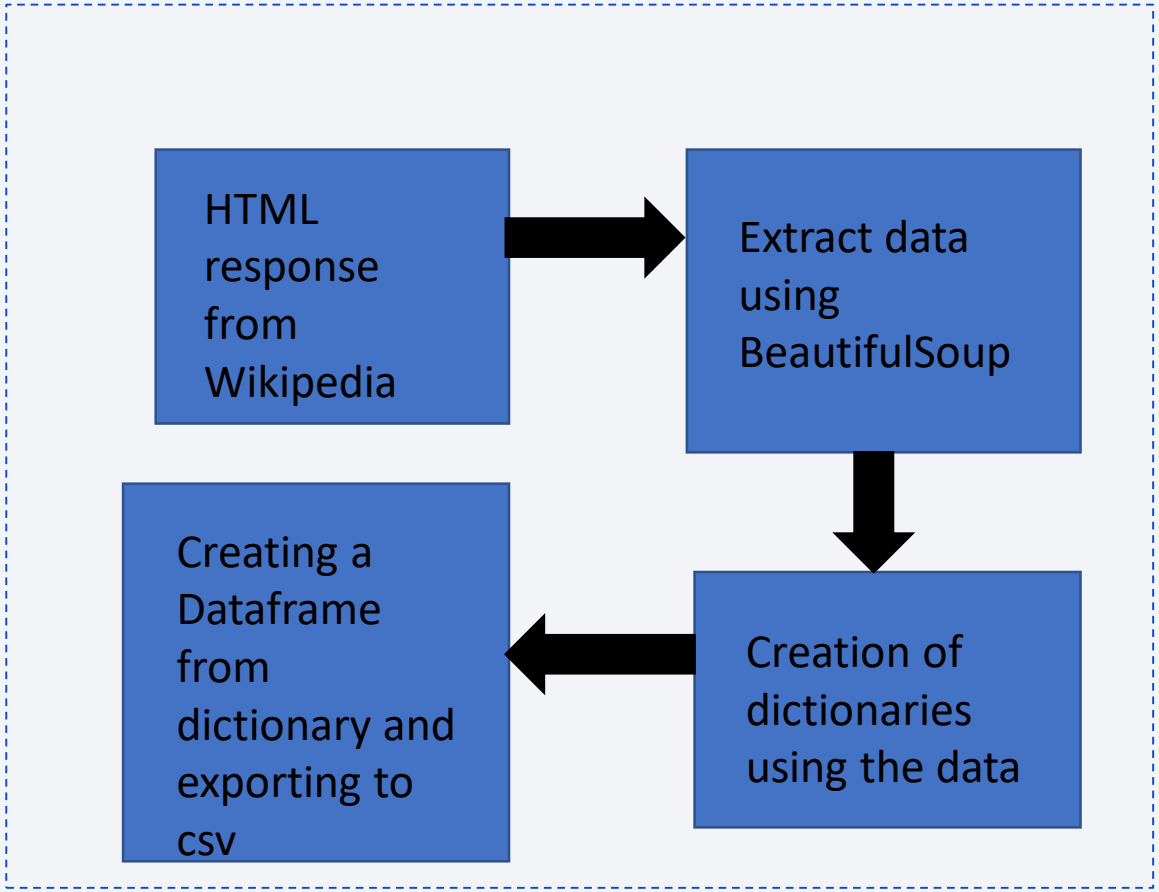
Data Collection – SpaceX API

- Github URL:
- https://github.com/tanmai14/Capstone/blob/main/spacex-data-collection-api.ipynb



Data Collection – Web Scrapping

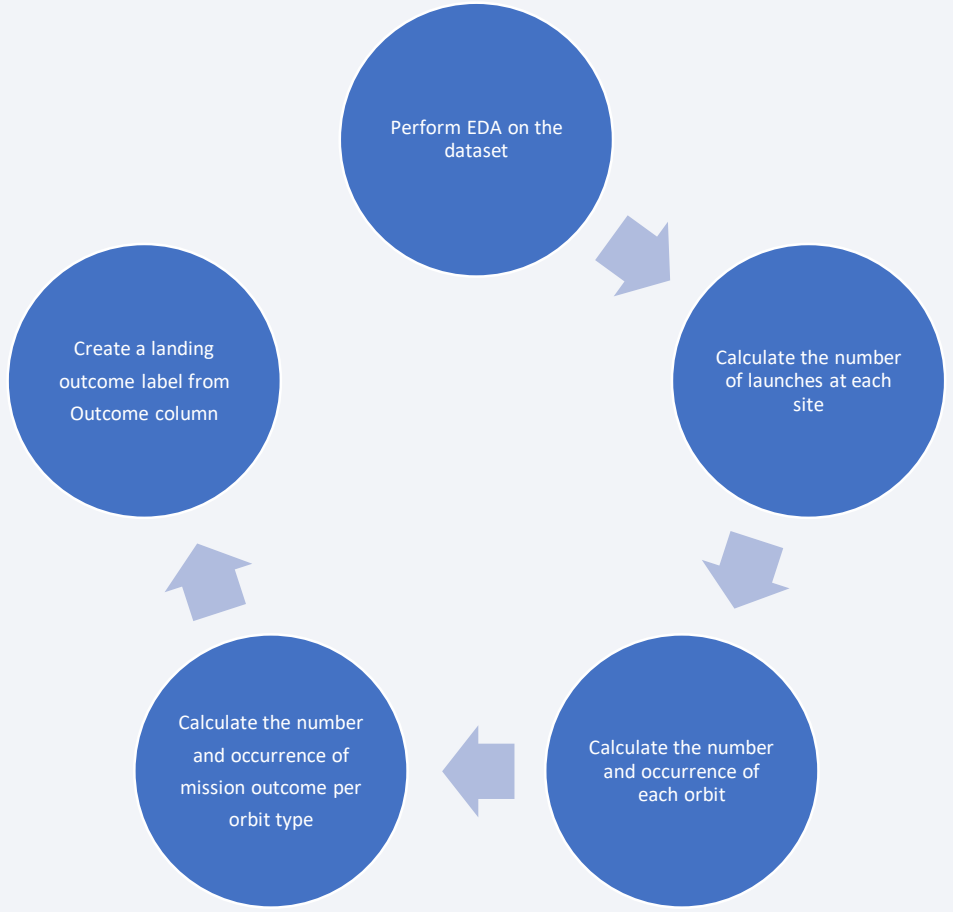
- Github URL:
- https://github.com/tanmai14/Capstone/blob/main/spacex-webscraping.ipynb



Data Wrangling

Github Link

https://github.com/tanmai14/Capstone/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- Charts Plotted:

1. Flight Number VS. Payload Mass
2. Flight Number VS. Launch Site
3. Payload VS. Launch Site
4. Orbit VS. Flight Number
5. Payload VS. Orbit Type
6. Orbit VS. Payload Mass

- Github Link:
- https://github.com/tanmai14/Capstone/blob/main/eda-dataviz.ipynb

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomesin ground pad ,booster versions, launch_sitefor the months in year 2017
- Ranking the count of successful landing_outcomesbetween the date 2010-06-04 and 2017-03-20 in descending order.
- Github Link: https://github.com/tanmai14/Capstone/blob/main/EDA_SQL.ipynb

Build an Interactive Map with Folium

- To Create an interactive map out of the Launch Data, the Latitude and Longitude Coordinates for each launch location were used to create a Circle Marker with the name of the launch site labelled on it.
- Assigned the dataframe launch_outcomes(failures, successes) to *classes 0 and 1with Green and Red markers on the map in a MarkerCluster()*
- Haversine formula was used to calculate the distance from the launch site to different landmarks in the map
- **Github Url:** https://github.com/tanmai14/Capstone/blob/main/launch_site_locations.ipynb

Build a Dashboard with Plotly Dash

- Used Heroku to host the application
- The dashboard is built with Flask and Dash web framework.
- Graphs used:
  - Pie Charts
  - Scatter Plots

Heroku Link: https://capstone-tanmai.herokuapp.com/

Github Link:

https://github.com/tanmai14/Capstone/blob/main/Dashboard%20SpaceX%20Dataset.ipynb

Predictive Analysis (Classification)

- Training 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
- Estimating the Accuracy
  - Checking the Accuracy
  - Plotting the confusion Matrix
- Tuning the model
  - Feature Engineering
  - Hyper Parameter Tuning
- Selecting the best model
  - Model with the highest accuracy score is chosen as the best model

Github Link:

https://github.com/tanmai14/Capstone/blob/main/machine-learning-prediction-spacex.ipynb

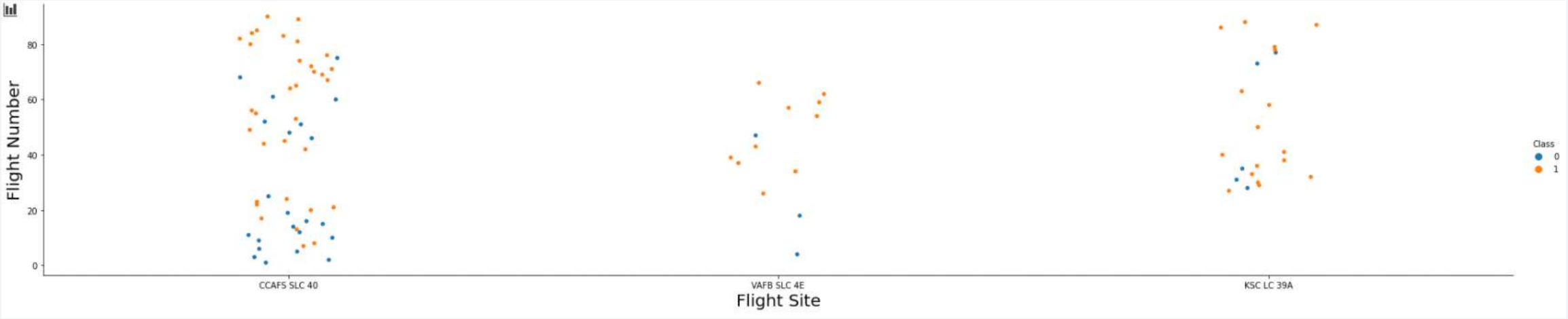
Results

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

Section 2

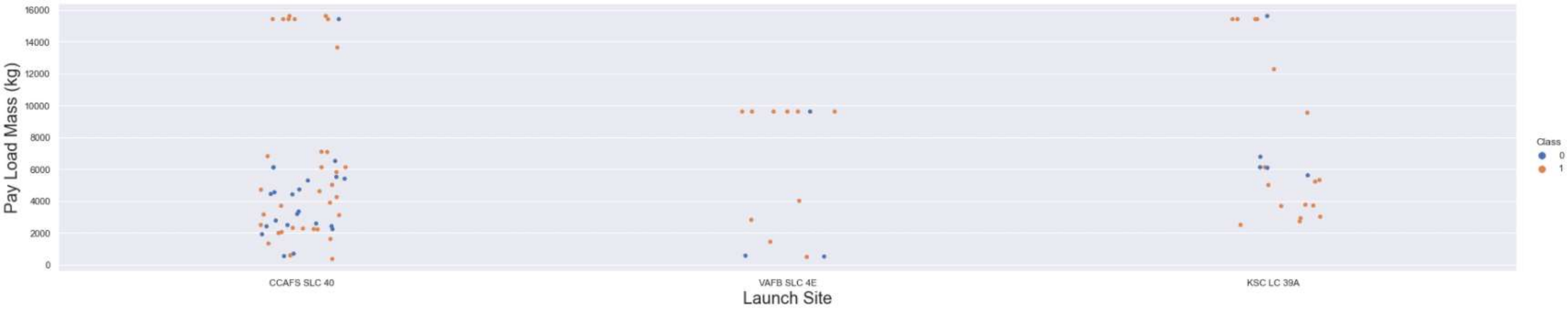
Insights drawn from EDA

Flight Number vs. Launch Site



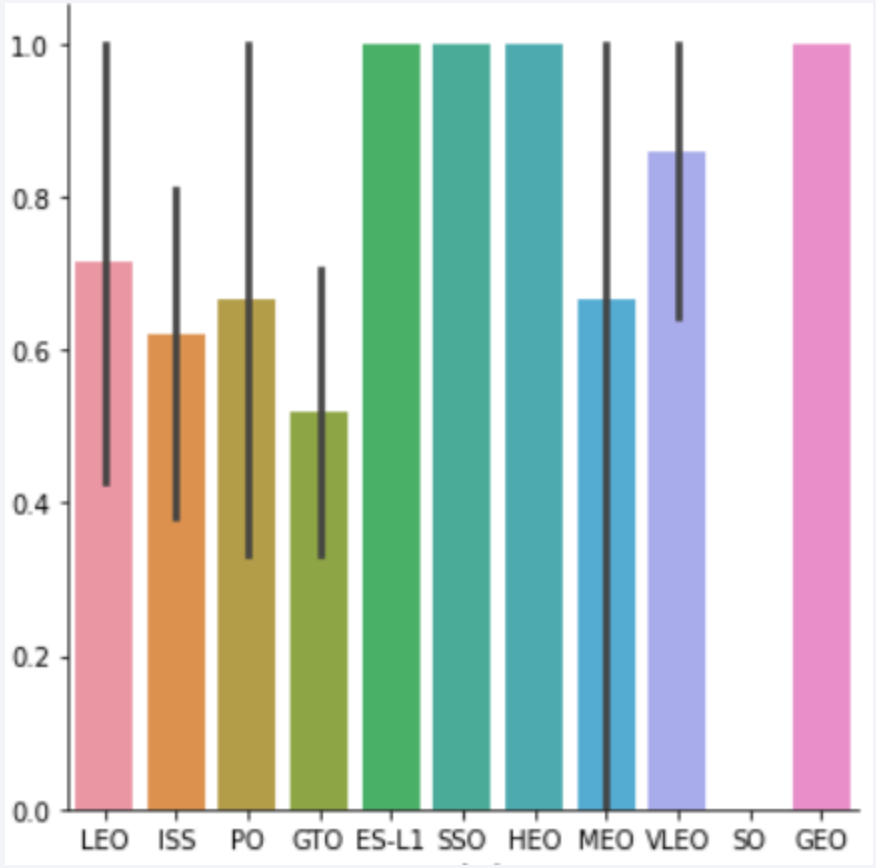
- Greater the number of flights, more is the success rate at a launch site

Payload vs. Launch Site



- Though there is no clear insight from this plot, we can see that that for CCAFS SLC 40, greater the payload more confidently we can estimate its success

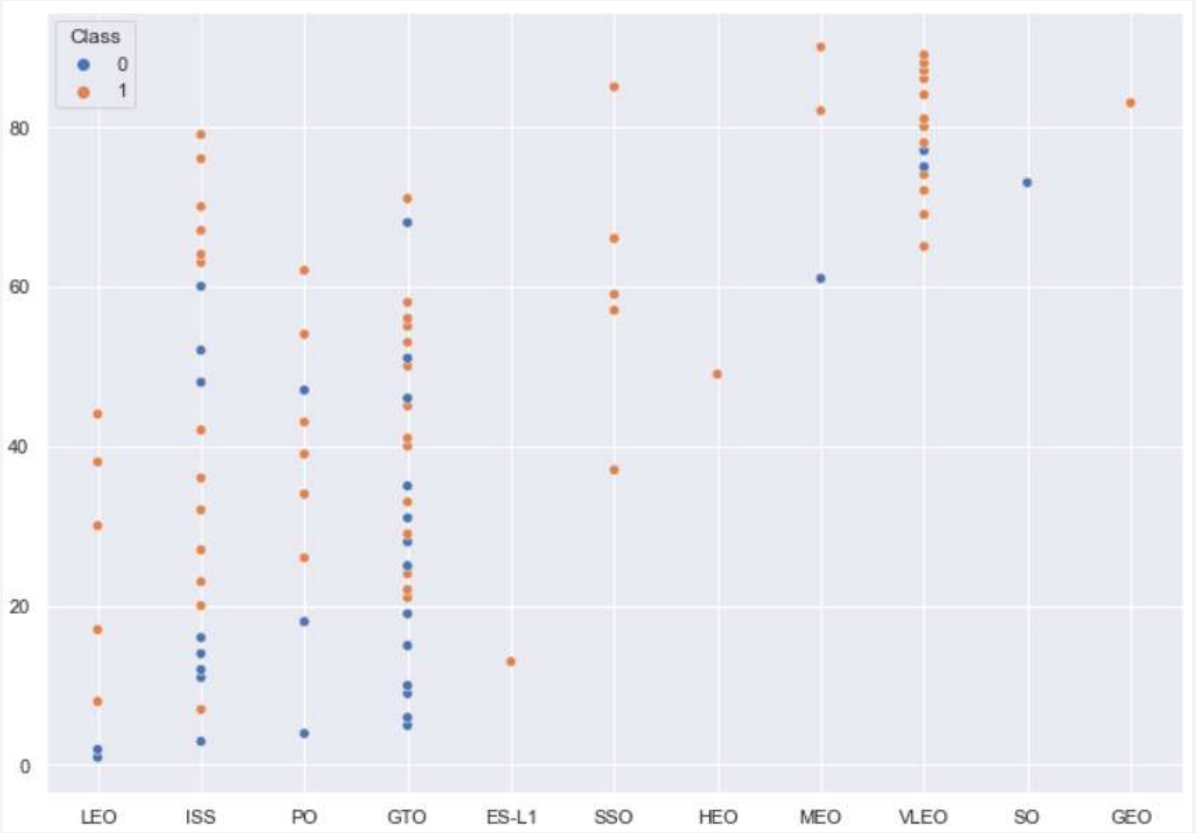
Success Rate vs. Orbit Type



- Orbits ES-L1, SSO, HEO and GEO have the highest success rate

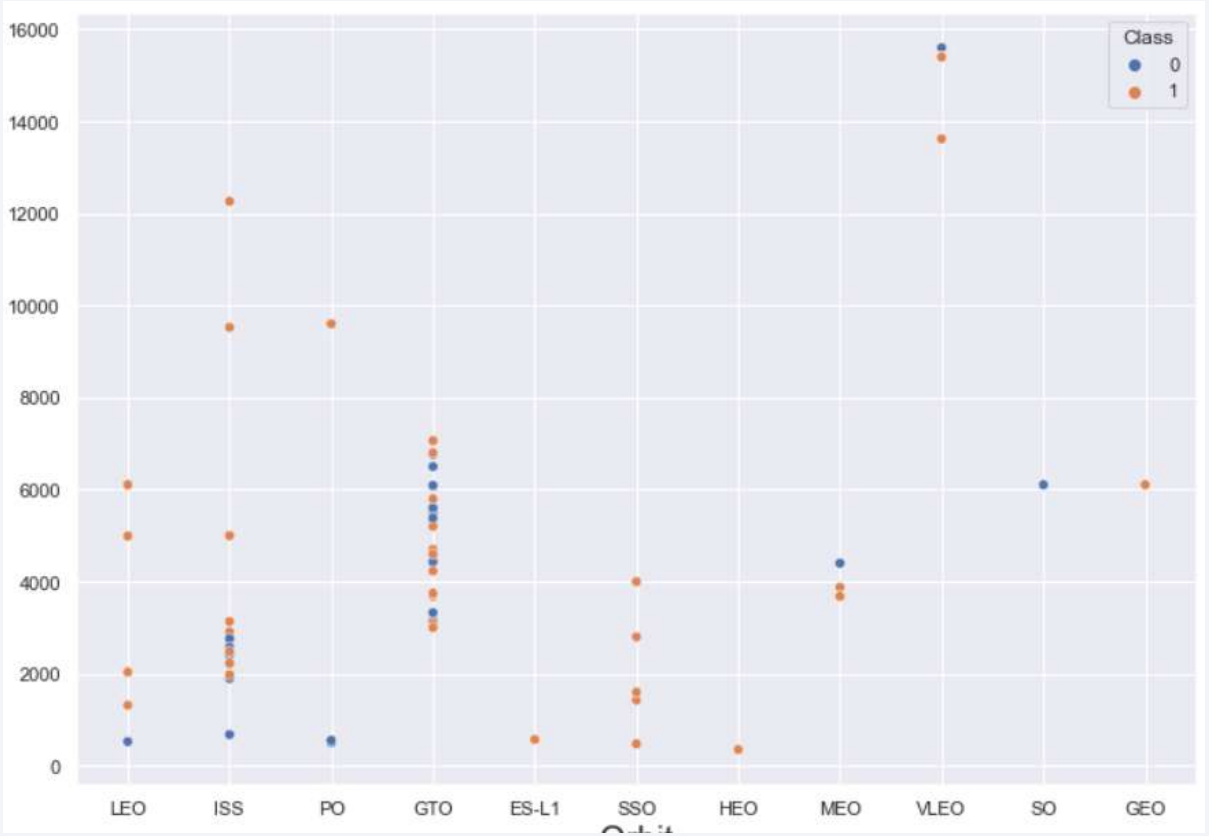
Flight Number vs. Orbit Type

- Only the LEO Orbit type clearly changes with a change in Flight Number. Rest all remain ambigous



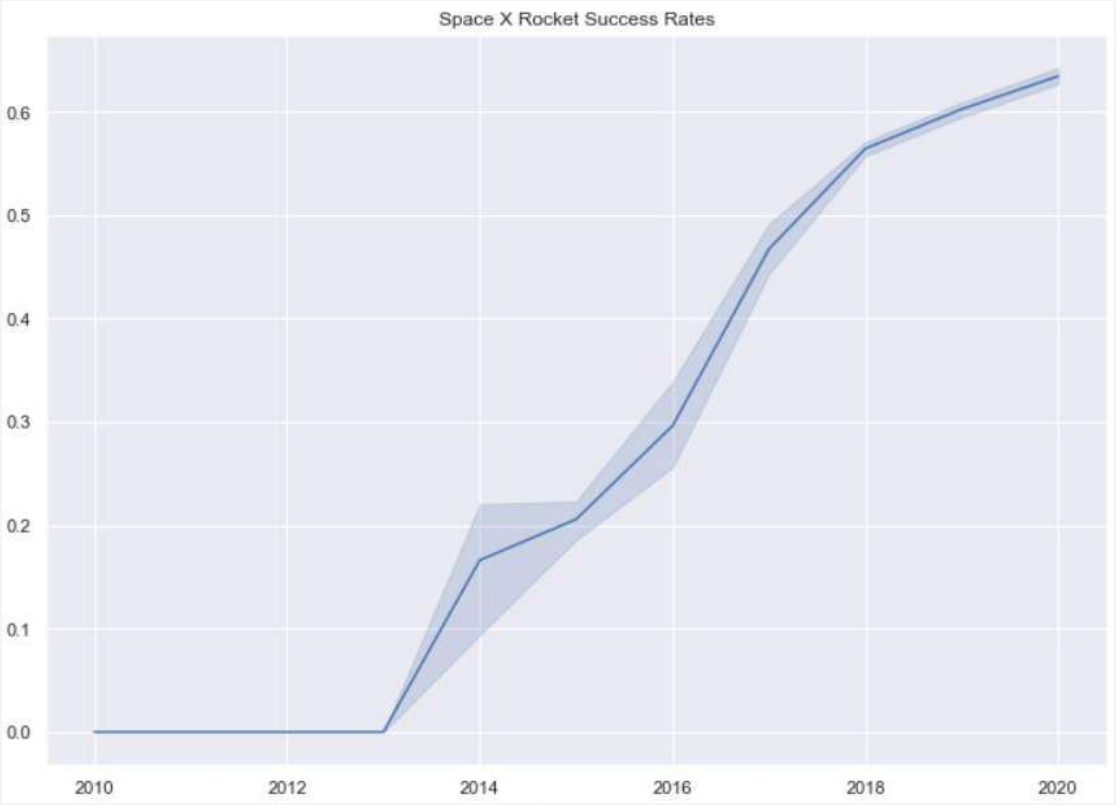
Payload vs. Orbit Type

- ISS orbit is highly successful with increasing payload and SSO orbit is successful when payload is below the range of 4000.
- The GTO orbit has nearly equal odds for success or failure at all given payload weights



Launch Success Yearly Trend

- The rate of suceess has been constantly accelerating since the year 2013



All Launch Site Names

- **Query**

Select DISTINCT Launch_site from tblSpcaeX

- **Output**
- CCAFS LC-40
- CCAFS SLC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC

Launch Site Names Begin with 'CCA'

- **Query**

Select TOP 5 * from tblSpaceX where Launch_site LIKE ‘KSC%’

- **Output**

	Date	Time_UTC	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	19-02-2017	2021-07-02 14:39:00.0000000	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
1	16-03-2017	2021-07-02 06:00:00.0000000	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2	30-03-2017	2021-07-02 22:27:00.0000000	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
3	01-05-2017	2021-07-02 11:15:00.0000000	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
4	15-05-2017	2021-07-02 23:21:00.0000000	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

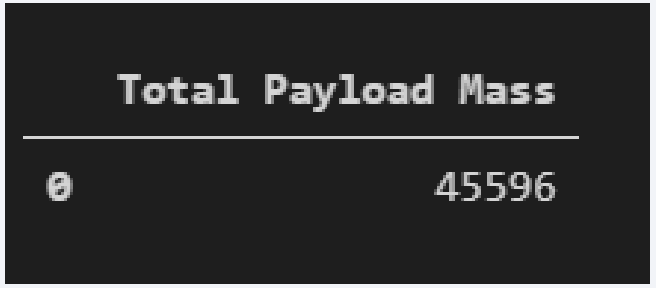
- The keyword Top 5 in the query shows only the first 5 records after executing the given query

Total Payload Mass

- **Query**

Select SUM(PAYLOAD_MASS_KG_) TotalPayloadMass from tblSpaceX where Customer='NASA(CRS)', 'TotalPayloadMass'

Output



	Total Payload Mass
0	45596

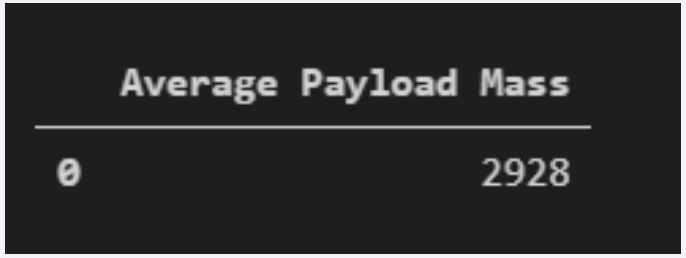
The function SUM() returns the total sum of the specified column and where clause filters the data as required

Average Payload Mass by F9 v1.1

Query:

Select AVG(PAYLOAD_MASS_KG_) AveragePayloadMass from tblSpaceX
where Booster_Version='F9v1.1‘

Output:



	Average Payload Mass
0	2928

The function AVG() returns the Average value of the specified column and where clause filters the data as required

First Successful Ground Landing Date

- Select MIN(Date) SLO from tblSpaceX where Landing_Outcome= ”Success (droneship)”

	Date which first Successful landing outcome in drone ship was acheived.
0	06-05-2016

- The function MIN() returns the Minimum value of the specified column and where clause filters the data as required

Successful Drone Ship Landing with Payload between 4000 and 6000

- Select Booster_Version from tblSpaceX where Landing_Outcome=‘ Success(groundpad)’ AND Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ <6000

	Date which first Successful landing outcome in drone ship was acheived.
0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

The AND Clause specifies filter conditions and where clause filters the data as required

Total Number of Successful and Failure Mission Outcomes

- SELECT (SELECT Count (Mission_Outcome) from tblSpaceX where Mission_Outcome LIKE '%Success%') as Successful_Mission_Outcomes, (SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome LIKE '%Failure%') as Failure_Mission_Coutcomes

	Successful_Mission_Outcomes	Failure_Mission_Outcomes
0	100	1

- Here, the concept of nested queries has been used to generate the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

- SELECT DISTINCT Booster_Version, MAX(PAYLOAD_MASS KG) AS [Maximum Payload Mass] FROM tblSpaceX GROUP BY Booster Version ORDER BY [Maximum Payload Mass] DESC

	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
...	...	...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0

97 rows x 2 columns

Group By: Arrange the data according to certain column
DESC: Descending Order
Distinct: Unique Values

2015 Launch Records

- SELECT DATENAME(month, DATEADD(month,MONTH(CONVERT(date, Date, 105)), 0) - 1) AS Month, Booster_Version, Launch_Site, Landing Outcome FROM tblSpaceX WHERE (Landing_Outcome LIKE N'%Success%') AND (YEAR(CONVERT(date, Date, 105)) = '2015')

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
March	F9 FT B1021.2	KSC LC-39A	Success (drone ship)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1029.2	KSC LC-39A	Success (drone ship)
June	F9 FT B1036.1	VAFB SLC-4E	Success (drone ship)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
August	F9 FT B1038.1	VAFB SLC-4E	Success (drone ship)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
October	F9 B4 B1041.1	VAFB SLC-4E	Success (drone ship)
October	F9 FT B1031.2	KSC LC-39A	Success (drone ship)
October	F9 B4 B1042.1	KSC LC-39A	Success (drone ship)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

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

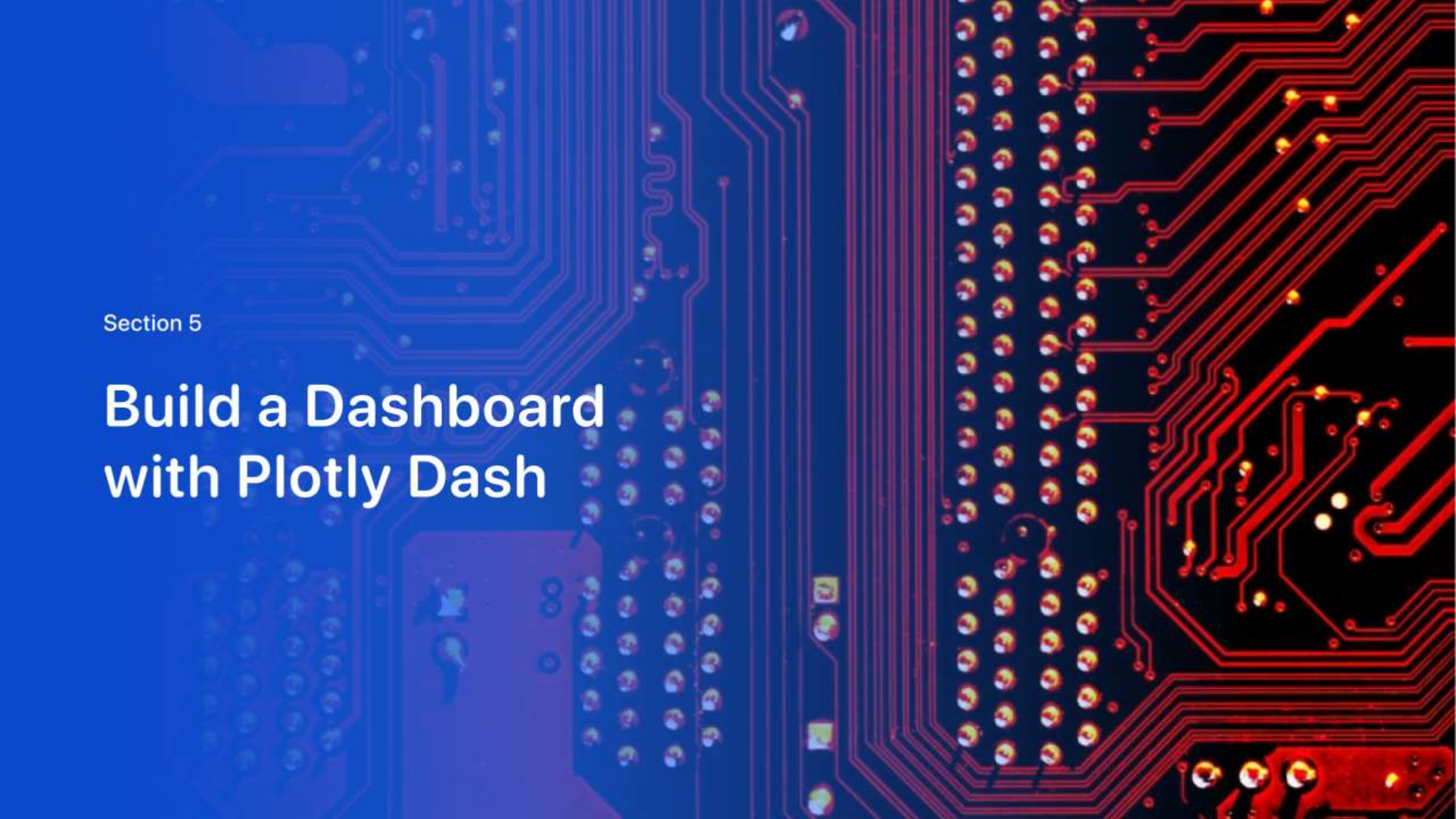
- SELECT COUNT(Landing_Outcome)FROM tblSpaceX WHERE (Landing_Outcome LIKE '%Success%') AND (Date > '04-06-2010')AND (Date < '20-03-2017')

	Successful Landing Outcomes Between 2010-06-04 and 2017-03-20
0	34

Section 4

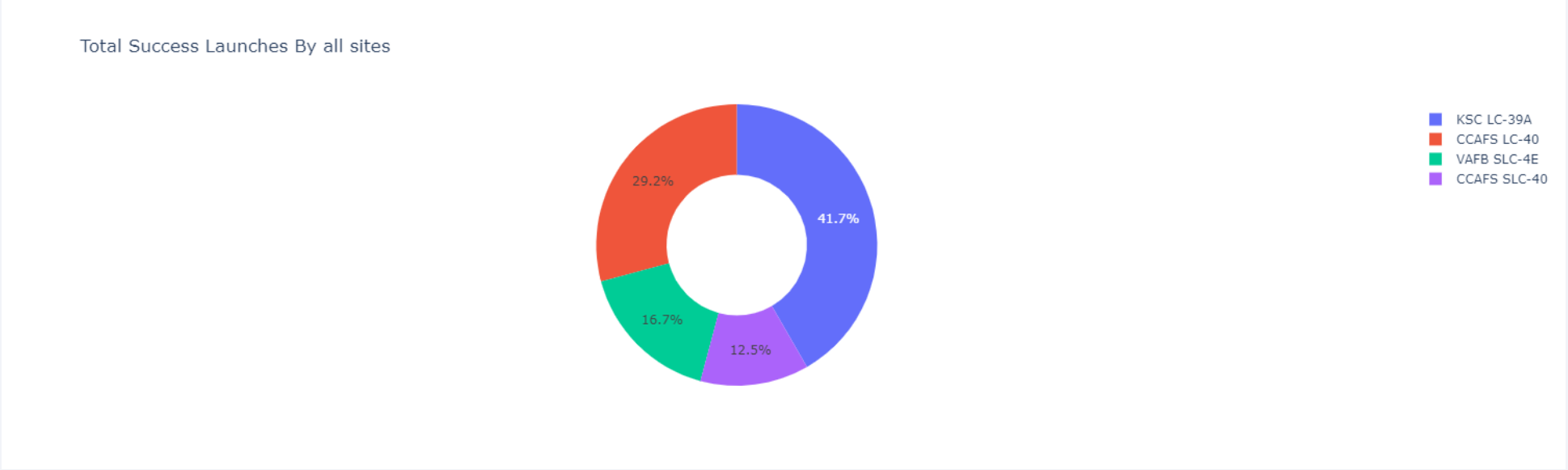
Launch Sites Proximities Analysis

Global Map of Launch Sites


Section 5
Build a Dashboard with Plotly Dash

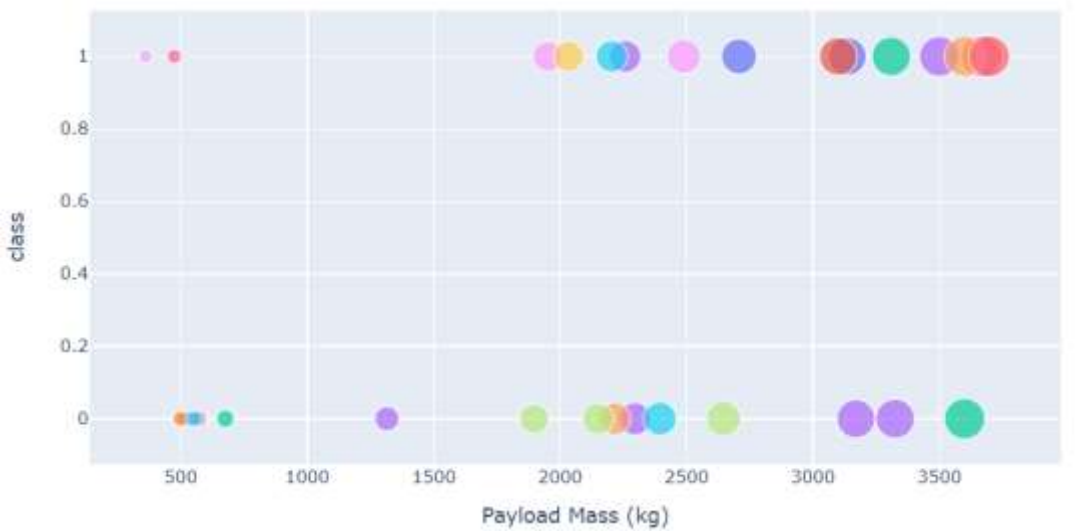
Pie chart for the launch site with the highest launch success ratio

- KSC LC-39A is the most successful
- CCAFS SLC-40 is the least successful

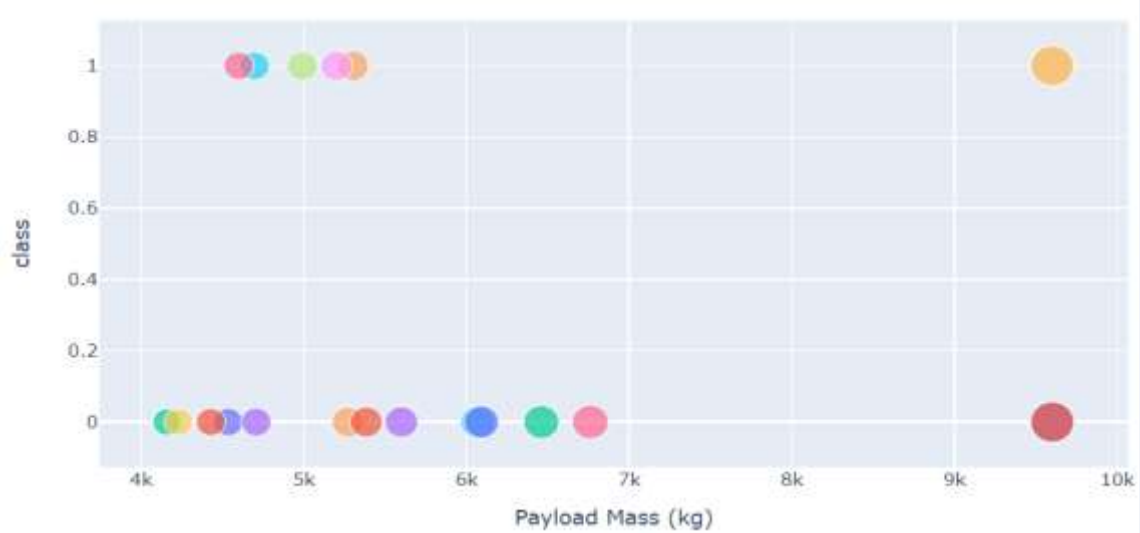


Payload vs Launch outcome scatter plot

- Low Payload



- High Payload



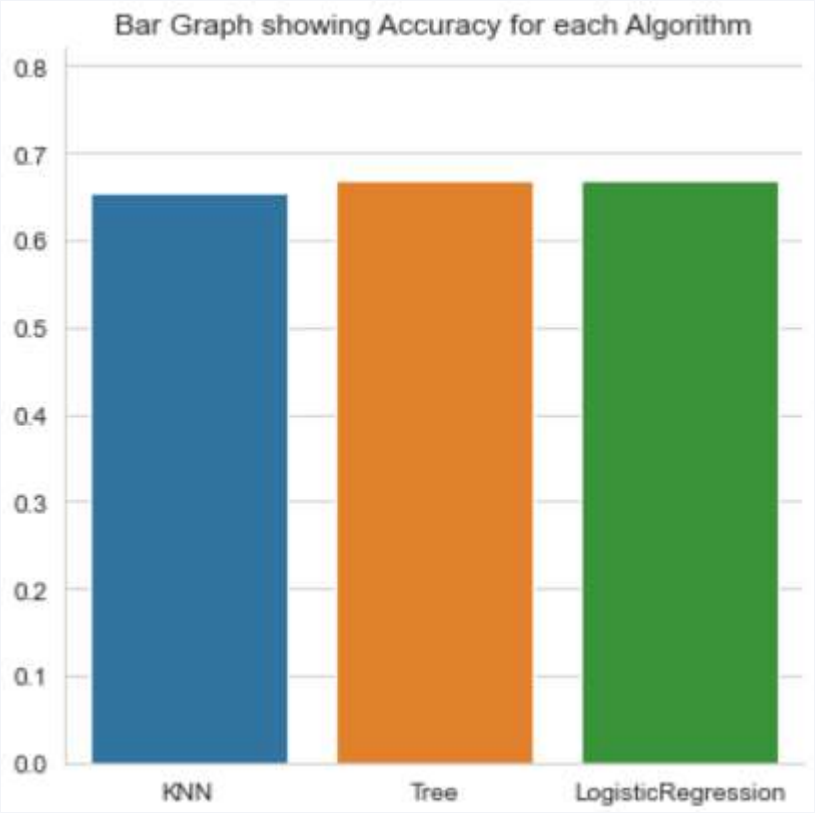
Lower the payload, higher is the probability for a successful launch

Section 6

Predictive Analysis (Classification)

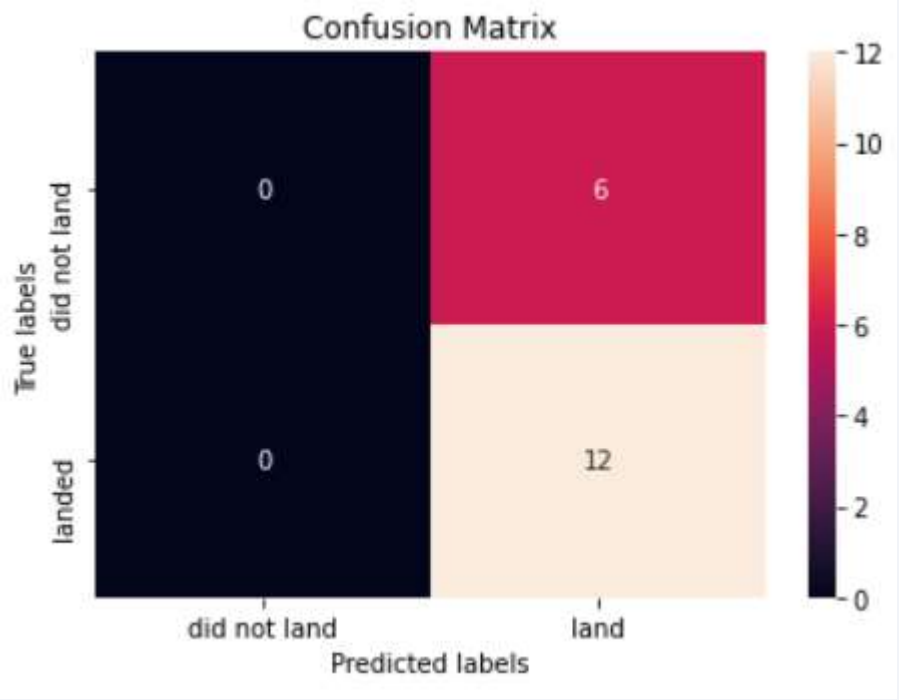
Classification Accuracy

- Though all the models are nearly similar in terms of accuracy, the bar graph suggests that the Decision Tree algorithm is the most accurate compared to its counterparts.



Confusion Matrix

- Given is the confusion matrix for the prediction the trained model has done.
- It can be clearly seen that the a minor proportion of the predictions are False Positives which may bias the model while predicting an output



Conclusions

- The success rate of the launches is being drastically improved over the years
- Less Payload implies more chance of success
- Decision Tree Algorithm has the highest accuracy for this case study
- KSC LC-39A is the most successful launch site
- ESL1, SSO, HEO and GEO are the best orbits for successful launch

Thank you!