IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Parth Gandhi
December 15, 2022

Outline

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

Executive Summary

- Methodology
  - Data Collection – Using Web Scraping and SpaceX API
  - EDA – Using Data Wrangling and Interactive Data Visualization
  - Predictive Analytics – Using ML Models (with10-fold CV and 20% Test Data): LogReg, SVM, DTs, and KNN Classifiers
- Results Summary
  - EDA Helped Identify Features That Were Best To Predict The Success of Launches
  - A Comparison of Machine Learning Predictions Showed The Most Accurate Model to Predict The Success of Launches Using The Pre-Identified Features Using EDA
  - Although All The Model Had The Same Test Accuracy, DTs Had a Higher Accuracy on Training Data

Introduction

- Background
  - Space Y Needs to Evaluate Its Viability to Compete With Space X
  - This Can Be an Indirect Insight From Whether Space Y Can Launch and Successfully Land First Stage Rockets At a Lower Total Cost
  - For This, Space Y Needs to Know What Factors Correctly Predict Successful Landings, Which Can be Found From The Space X Data Available
  - Using This Information On Predictive Variables of a Successful Stage 1 Launch, Space Y can Optimize Them (For Ex. Launch Location, Payload Mass, etc.) And Reduce The Total Cost Per Launch

- Problems you want to find answers
  - What Factors Affect The Successful Landing of Stage 1 Rocket?
  - Which ML Model/s Can be Used To Predict a Successful Outcome?
  - What Value of the Identified Factor/s Will Optimize The Cost of Launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources
    1. Space X API – https://api.spacexdata.com/v4/rockets/
    2. WebScraping – https://en.Wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- Perform data wrangling
  - Dealing with missing values – imputing with mean
  - Categorical variables converted to numerical using One Hot Encoding
  - Created a landing outcome label based on outcome data after Feature selection process

Methodology

Executive Summary

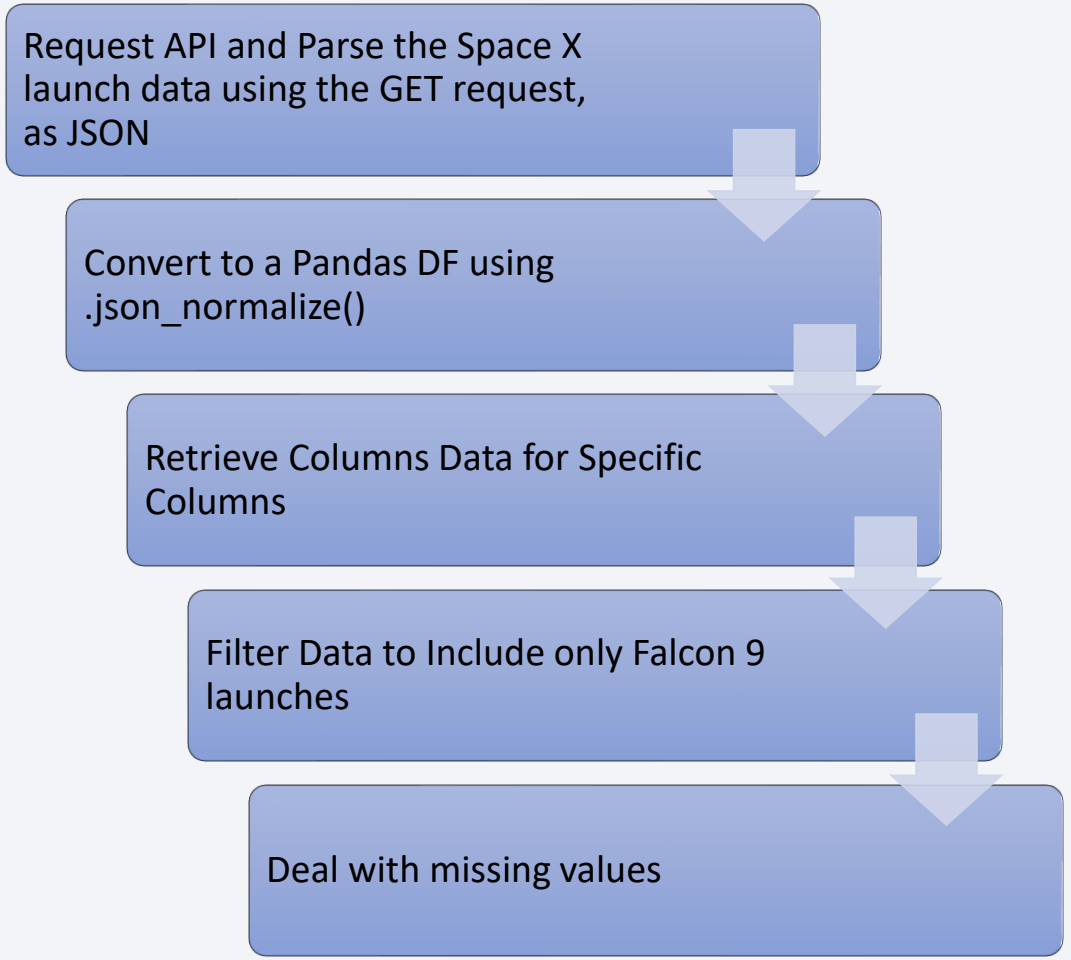
- Performed exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was normalized using standard feature scaling
  - Then the data was split into 20% test data and 80% training data
  - The training data was now used to evaluate the accuracy of 4 classifiers – LogReg, SVM, DT, and KNN along with 10-fold CV and Grid Search to optimize the hyperparameters

Data Collection

- Data from Space X was obtained from 2 sources:
    1. Space X API – https://api.spacexdata.com/v4/rockets/
    2. WebScraping – https://en.Wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

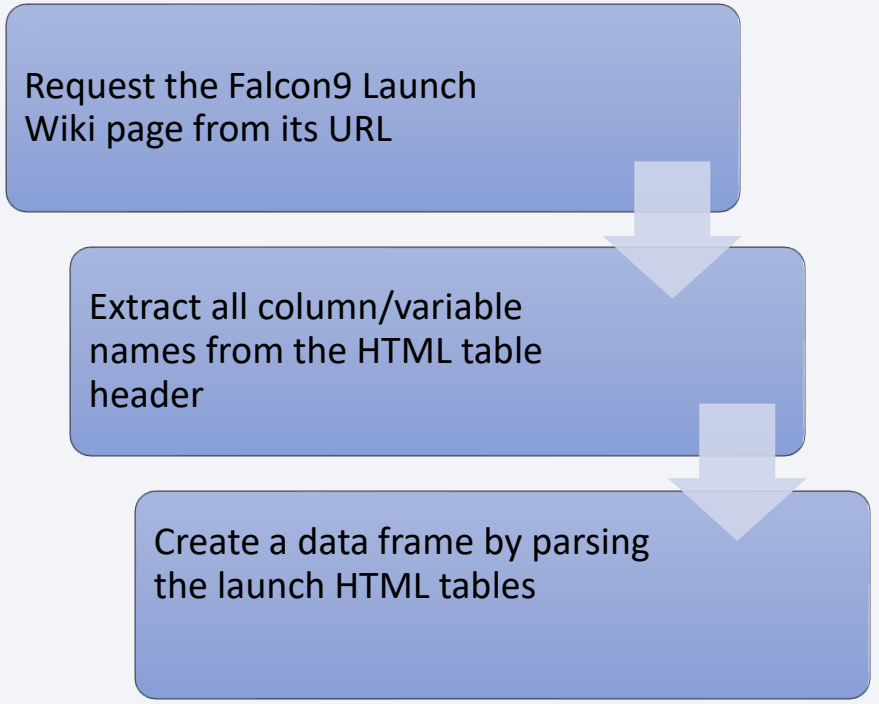
Data Collection – SpaceX API

- SpaceX offers a public API – Data for Falcon 9 Launches can be obtained form here
- See the flowchart for using this API to collect the Data
- GitHub URL – https://github.com/parthgandhi1998/testrepo/blob/master/Module_1.1_spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data for Space X launches can also be downloaded from Wikipedia
- See the flowchart for using Web Scraping to collect the Data
- GitHub URL – https://github.com/parthgandhi1998/testrepo/blob/master/Module_1.2_Data%20Collection%20Using%20Webscraping.ipynb



Data Wrangling

Create Landing Outcome Label

- Exploratory Data Analysis was performed to for understanding the variables and identifying feature information.
- Calculated Number of Launches on Each site, number and occurrences of each orbit, and the number and occurrence of mission outcome per orbit type
- Created a landing outcome label from the Outcome column

- Data Exploration and Feature Understanding
- Metrics – Site, orbit, and mission outcome
- Using the Outcome label

- GitHub URL – https://github.com/parthgandhi1998/testrepo/blob/master/Module_1.3_Data%20Wrangling.ipynb

EDA with Data Visualization

- For EDA, Scatter plots and Bar plots were used to visualize the relationships between different pairs of features
  - Payload Mass Vs Flight Number
  - Launch Site Vs Flight Number
  - Launch Site Vs Payload Mass
  - Orbit Vs Flight Number
  - Payload Vs Orbit
- GitHub URL - https://github.com/parthgandhi1998/testrepo/blob/master/Module_2.2_EDA_with_Pandas_and_Matplotlib.ipynb

EDA with SQL

- The SQL queries for the following tasks were performed:
    1. **Display the names of the unique launch sites in the space mission**
    2. **Display 5 records where launch sites begin with the string 'CCA'**
    3. **Display the total payload mass carried by boosters launched by NASA (CRS)**
    4. **Display average payload mass carried by booster version F9 v1.1**
    5. **List the date when the first successful landing outcome in ground pad was achieved.**
    6. **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
    7. **List the total number of successful and failure mission outcomes**
    8. **List the names of the booster_versions which have carried the maximum payload mass. Use a subquery**
    9. **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.**
    10. **Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.**
- GitHub URL –

https://github.com/parthgandhi1998/testrepo/blob/master/Module_2.1_EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

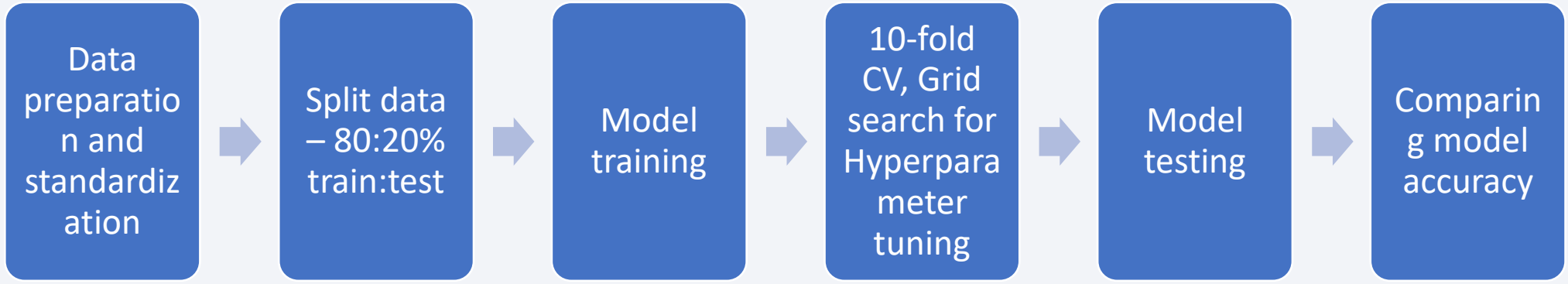
- Map objects such as markers, circles, lines, and marker clusters were created and added to a folium map
- Markers indicate points like launch sites
- Circles indicate highlighted areas around specific coordinates and add a label to it, like NASA Johnson Space Center
- Marker cluster is a collection of multiple marker objects layered into a single marker cluster object to display multiple launch sites
- Lines are for showing distances between two coordinates
- GitHub URL - https://github.com/parthgandhi1998/testrepo/blob/master/Module_3.1_Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize the data:
    1. Percentage of launches by site
    2. Payload range – Range Slider
- This comparison helped identify the relation between payloads and launch sites, and thus decide the best place to launch according to payloads
- GitHub URL - https://github.com/parthgandhi1998/testrepo/tree/master/Module_3.2_Interactive%20Dashboard%20with%20Plotly%20Dash

Predictive Analysis (Classification)

- Predictive Analytics – Using ML Models (with10-fold CV and 20% Test Data): LogReg, SVM, DTs, and KNN Classifiers (Using Grid Search to optimize hyperparameters)



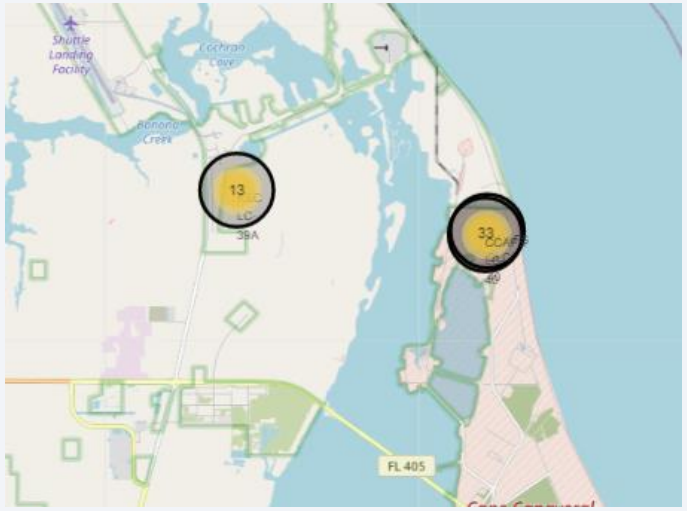
- GitHub URL - https://github.com/parthgandhi1998/testrepo/blob/master/Module_4.1_SpaceX_Machine_Learning_Prediction.ipynb

Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites
  - The first launches were done to Space X itself and NASA
  - The average payload of F9 v1.1 booster is 2,928.4 kg
  - The first success landing outcome happened in 2015, five years after the first launch
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
  - Almost 100% of mission outcomes were successful
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
  - The success rate kept improving since 2013 till 2020

Results

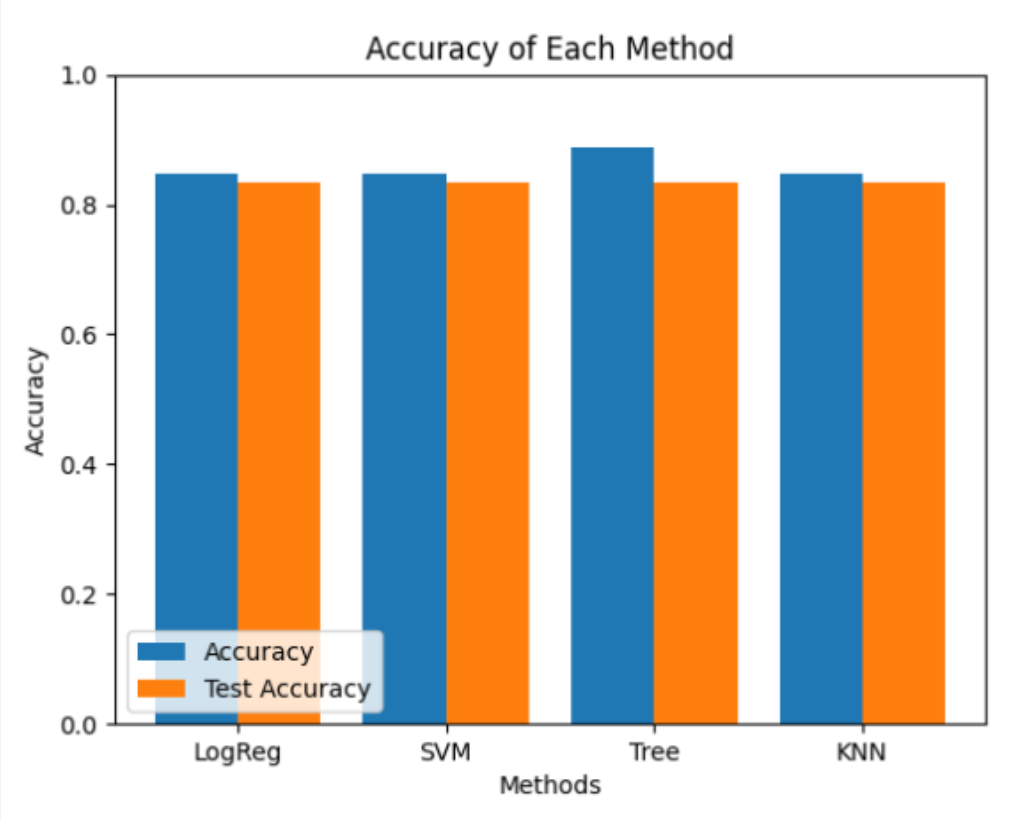
- Interactive analytics showed that launch sites are strategically placed close to the equator and near to safety places – ex. Sea, and have a good logistics infrastructure around them
- Most launches happen at the east coast launch sites (Assuming due to their proximity to the equator)

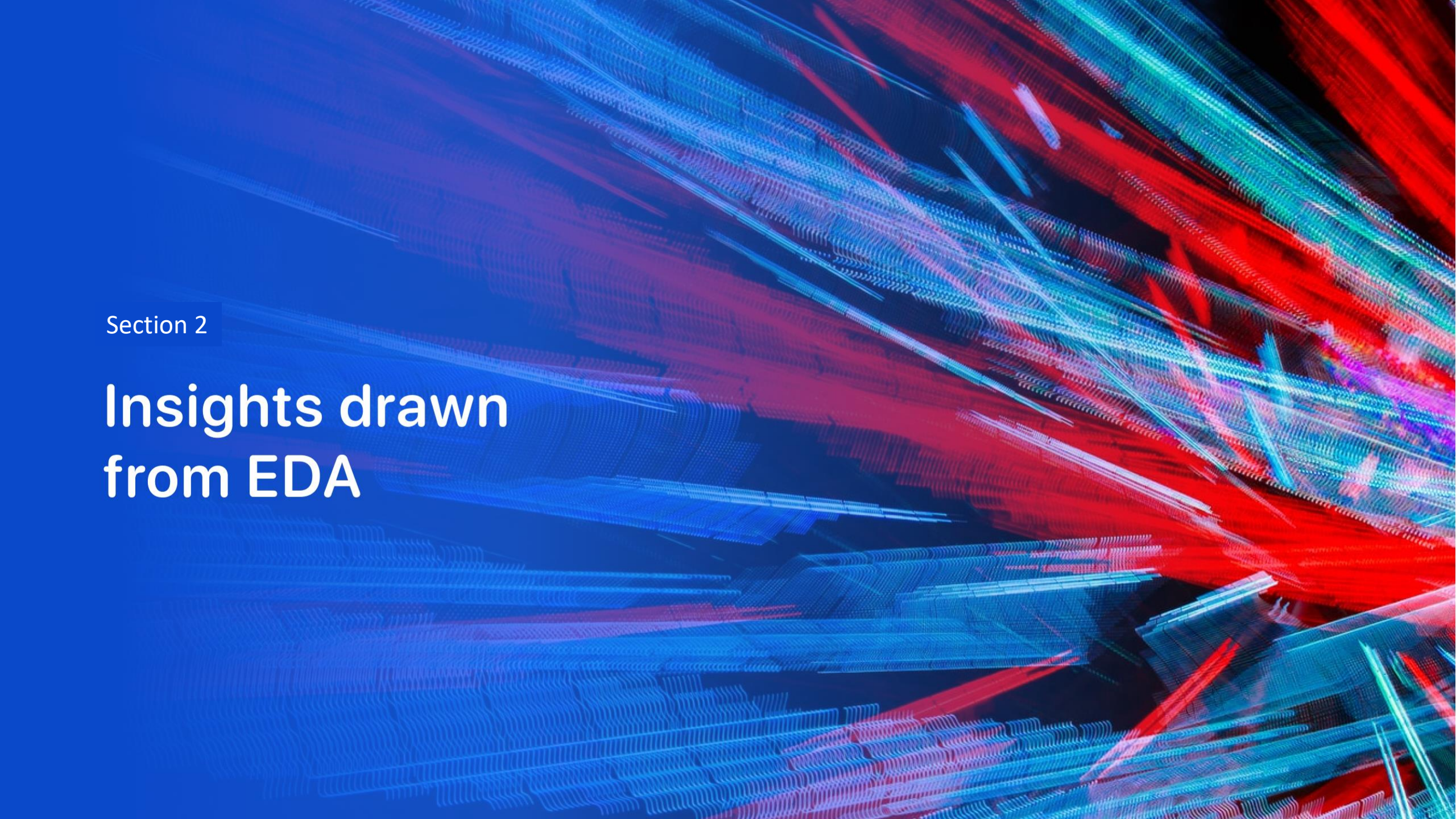


Results

- Predictive analysis results
  - Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.

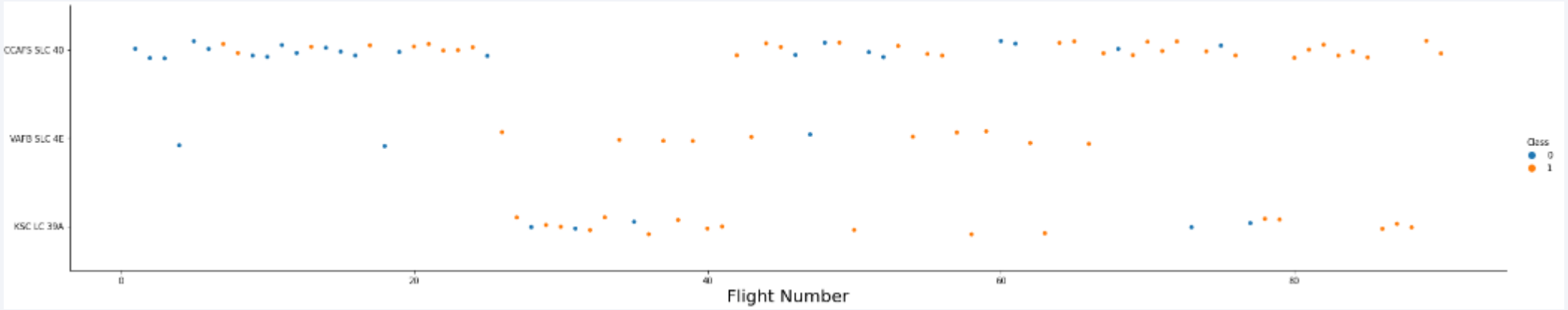
Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.8875	0.83333
KNN	0.84821	0.83333




Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

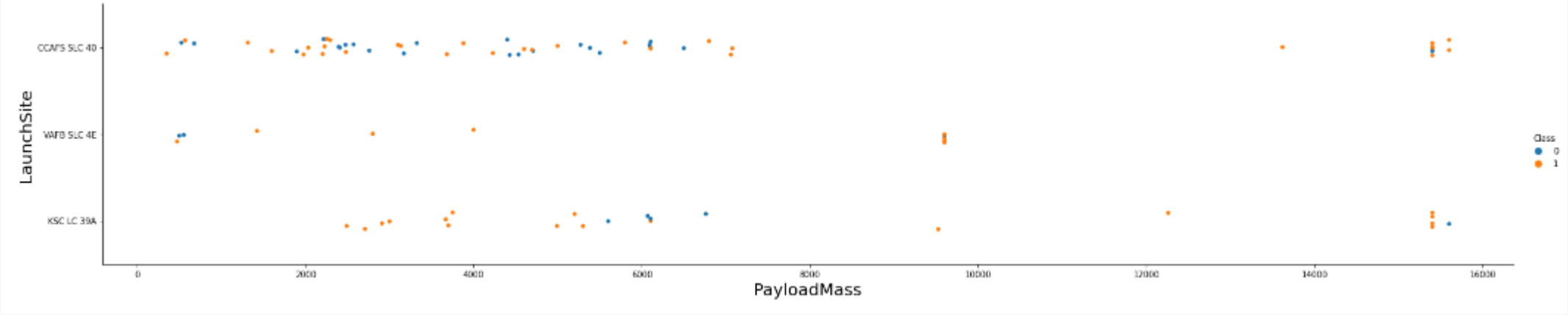
- Scatter plot of Flight Number vs. Launch Site



- CCAF5 SLC 40 seems to be a successful launch site considering more recent years.
- This is followed by VAFB SLC 4E and KSSC LC 39A
- The general success rate has increased with time

Payload vs. Launch Site

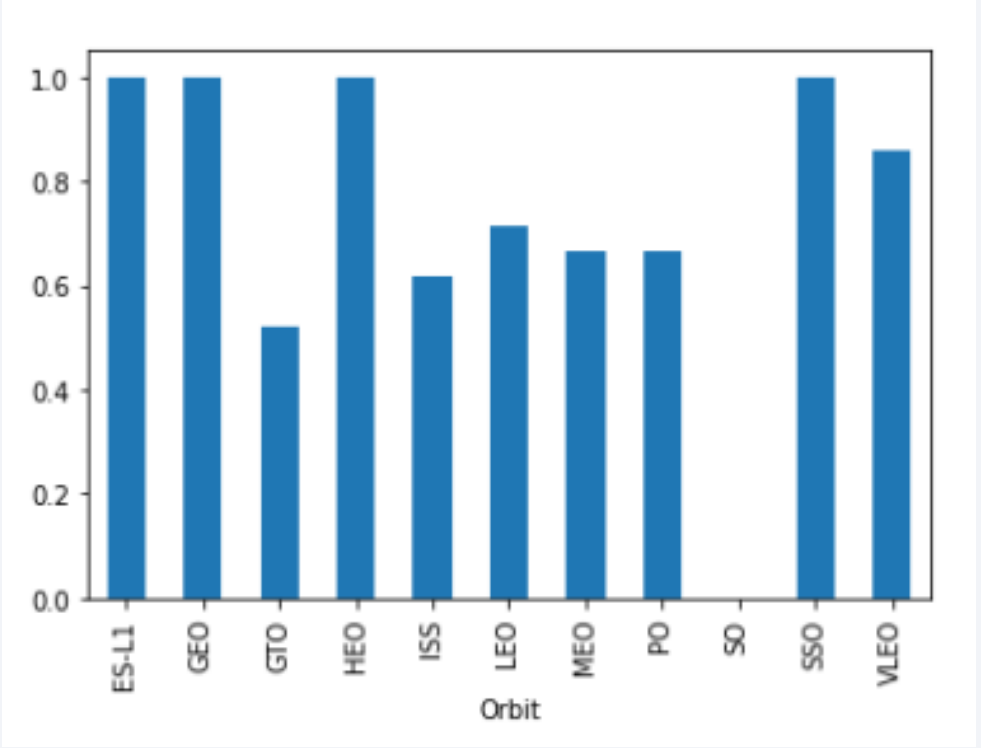
- Scatter plot of Payload Vs Launch Site



- There are fewer payloads above 9000 Kgs, but have a very high success rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.
- VAFB SLC 4E has never launched a payload of more than 10000 Kgs and KSSC LC 39A has never launched a payload less than 2000 Kgs

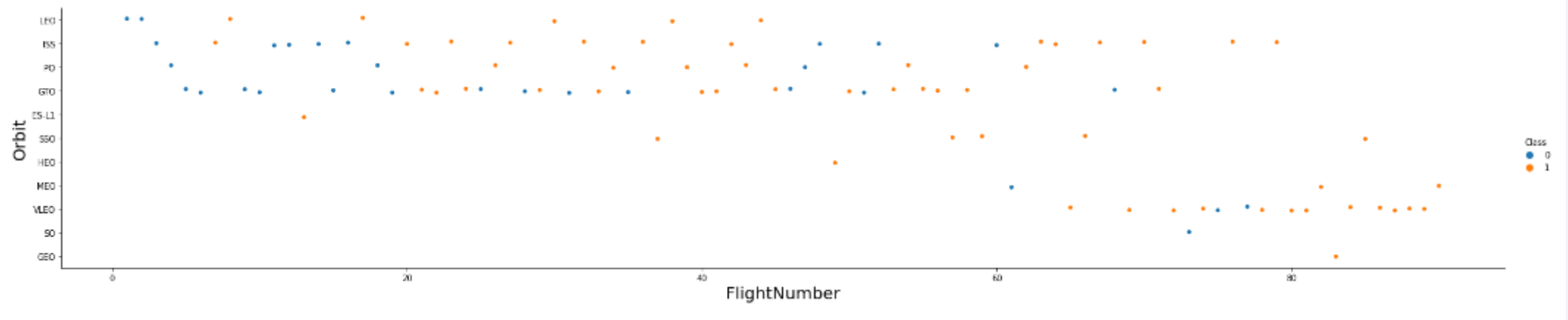
Success Rate vs. Orbit Type

- Highest success rate (=1) are for the orbits-
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by -
  - VLEO (>80%)
  - LFO (>70%)
- SO has no successful launches, and the lowest success rate is about 50% for GTO



Flight Number vs. Orbit Type

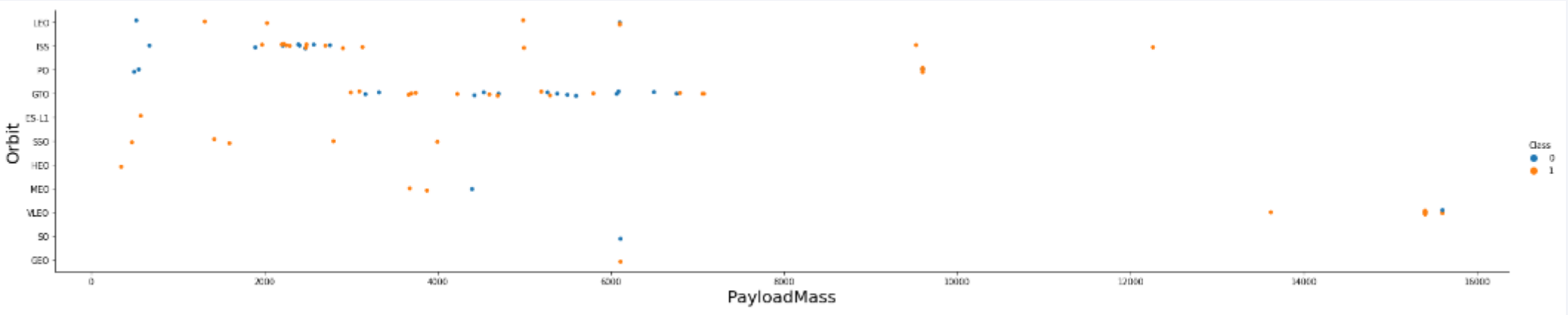
- Scatter plot of Flight number vs. Orbit type



- More recently, the trend has shifted towards VLEO considering the frequency of launches
- The success rate has increased over time in all orbits

Payload vs. Orbit Type

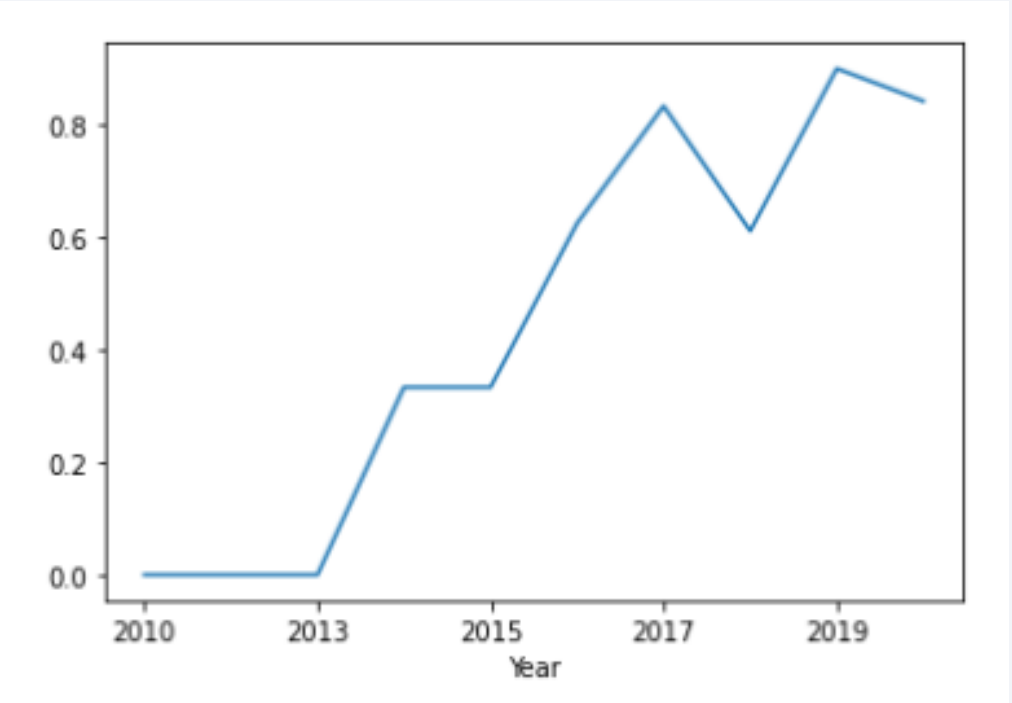
- Scatter plot of payload vs. orbit type



- There has been only one launch each to SO and GEO orbits
- Majority of the payloads have been launched to GTO and ISS
- ISS has the widest range of payloads and a good success rate as well

Launch Success Yearly Trend

- The success rate kept increasing from 2013 till 2020
- The initial 3 years showed no growth however the improvement took place in a step-wise fashion bi-yearly, seeing a dip in 2018 before climbing up again in 2019-2020



All Launch Site Names

- There are 4 unique launch sites

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

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- These are the 5 launches from Cape Canaveral

Total Payload Mass

- Total payload carried by boosters from NASA

TOTAL_PAYLOAD
111268

- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Avg payload mass carried by booster version F9 v1.1

AVG_PAYLOAD
2928.4

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928.4 kg.

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad

FIRST_SUCCESS_GP
01-05-2017

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

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

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

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

- Grouping mission outcomes and counting records for each group led us to the summary above

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster_Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- These were the 2 failed landings for the year 2015

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

- Ranking 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

- The classification of outcomes shows the types other than failure and success to consider, like No attempt and Controlled (ocean)

Landing_Outcome	QTY
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

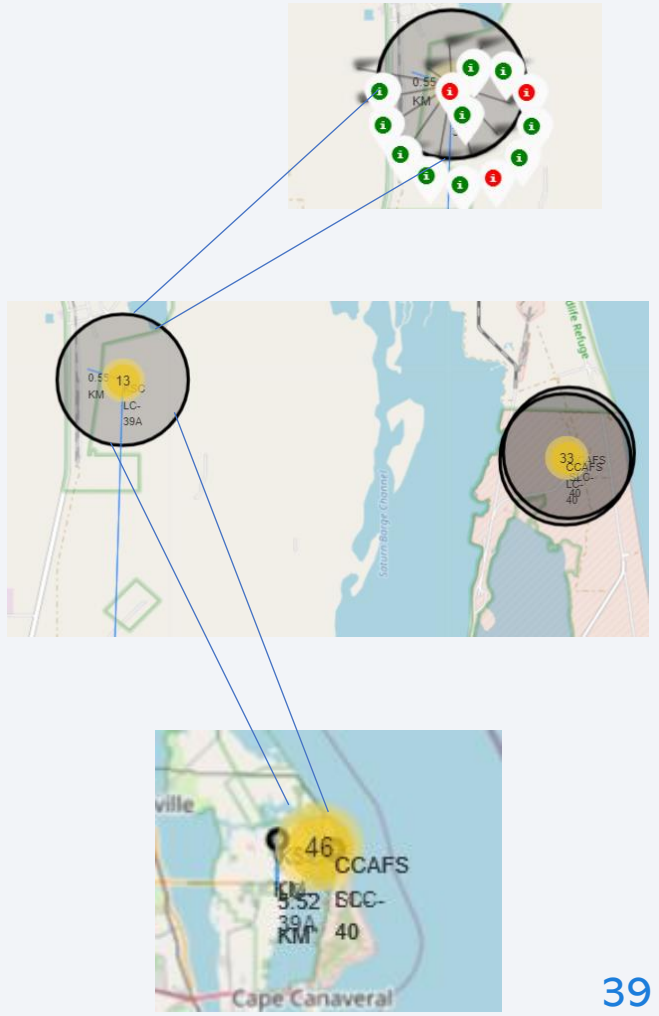
Launch Sites Proximities Analysis

All Launch Sites

- Launch sites are near to the sea and equator largely.
- They are not too far from roads and railroad for logistic conveniences.

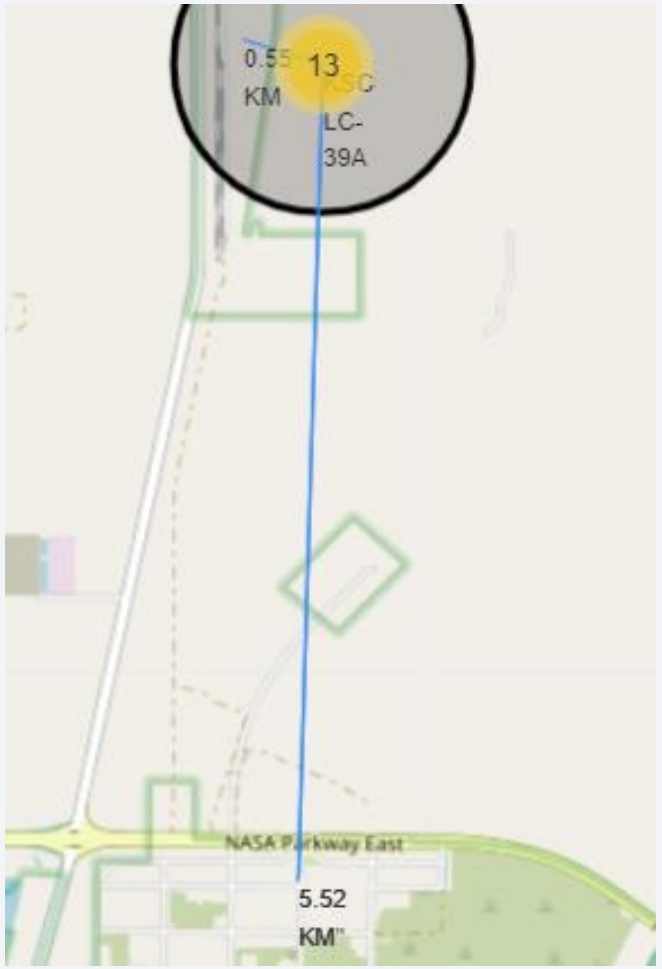
Launch Outcome by Site

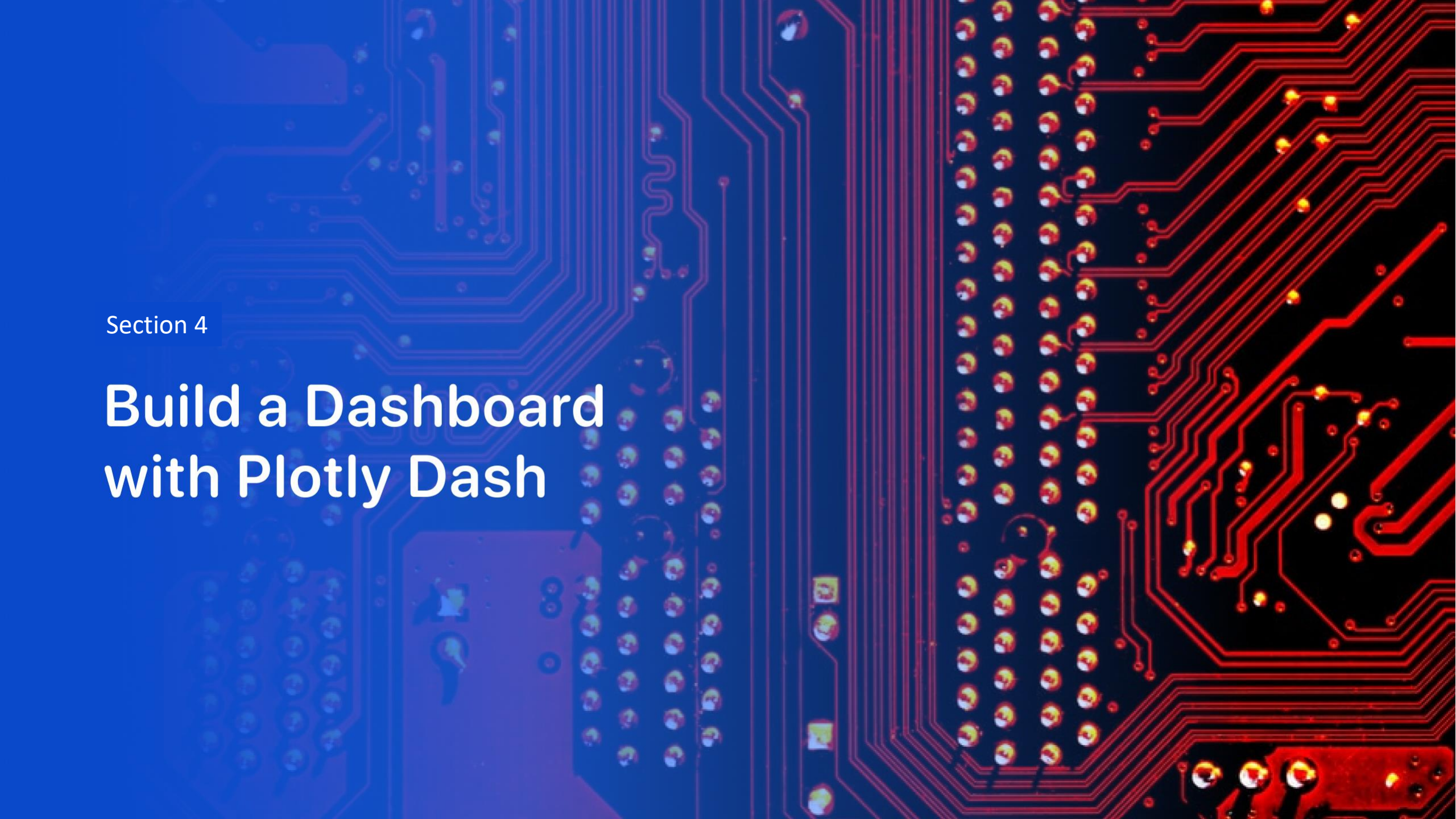
- Example of KSC LC-39A launch site launch outcomes



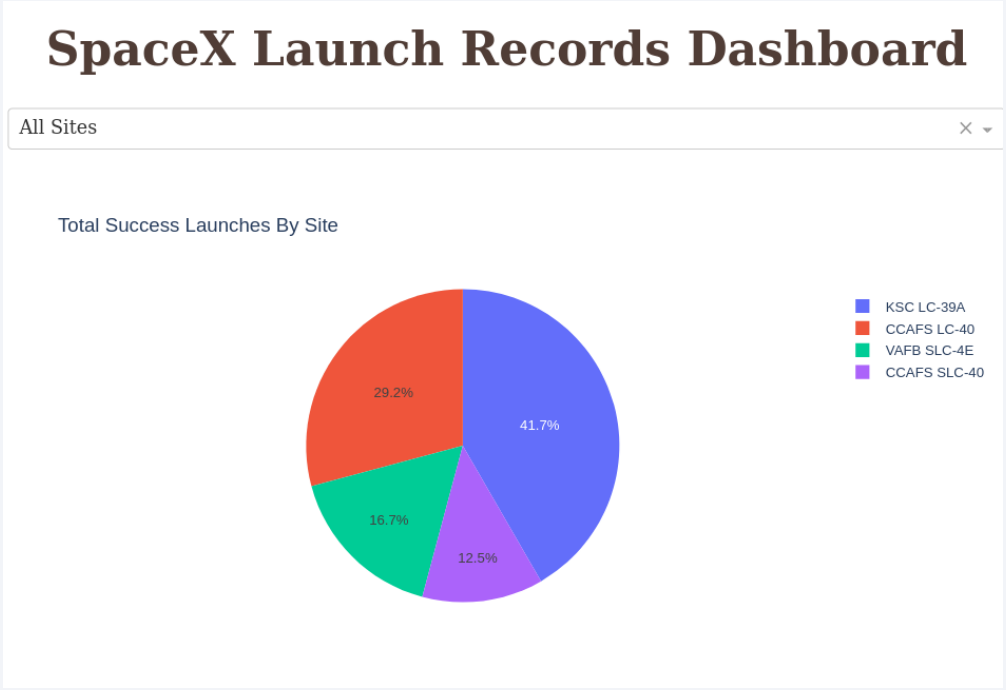
Logistics and Safety

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.




Section 4
Build a Dashboard with Plotly Dash

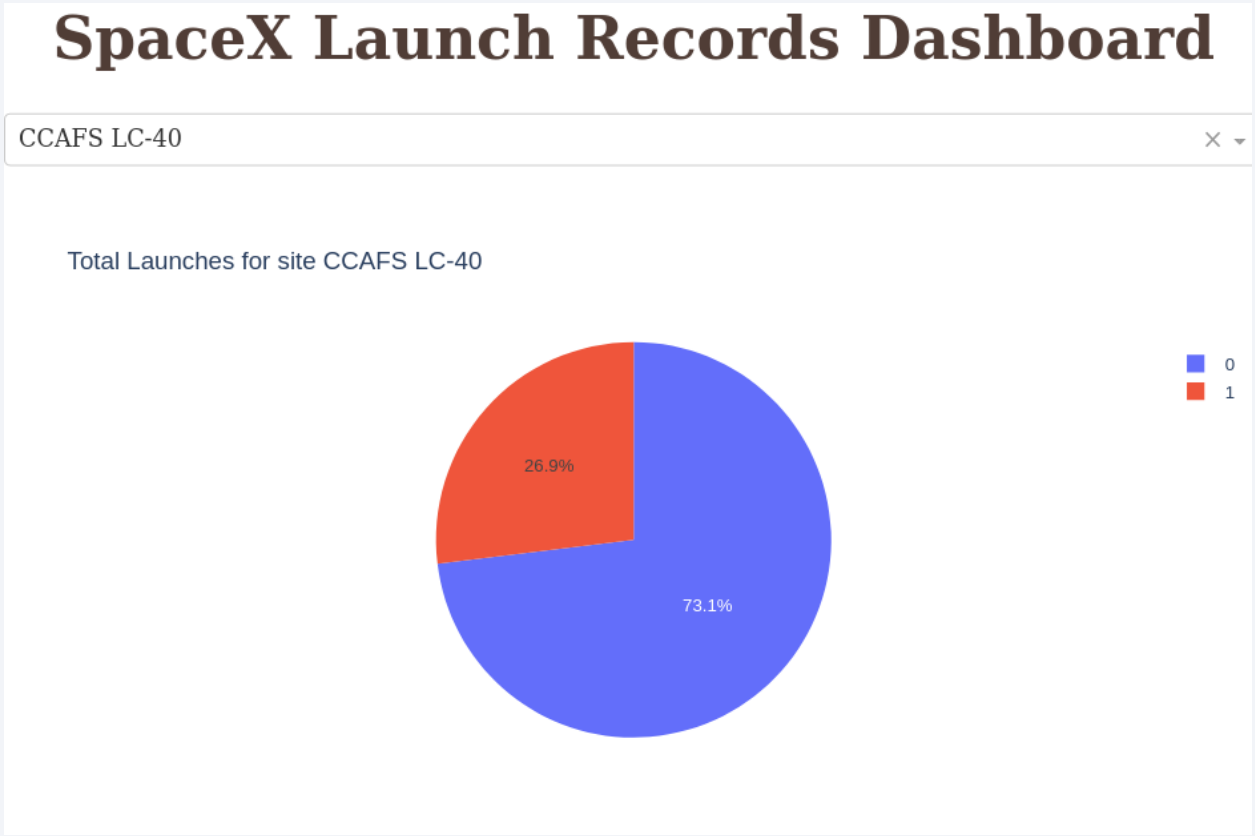
Successful Launches by Site



The launch site greatly influences the success of missions, with KSC LC-39A having the highest share of successful launches i.e. 41.7%

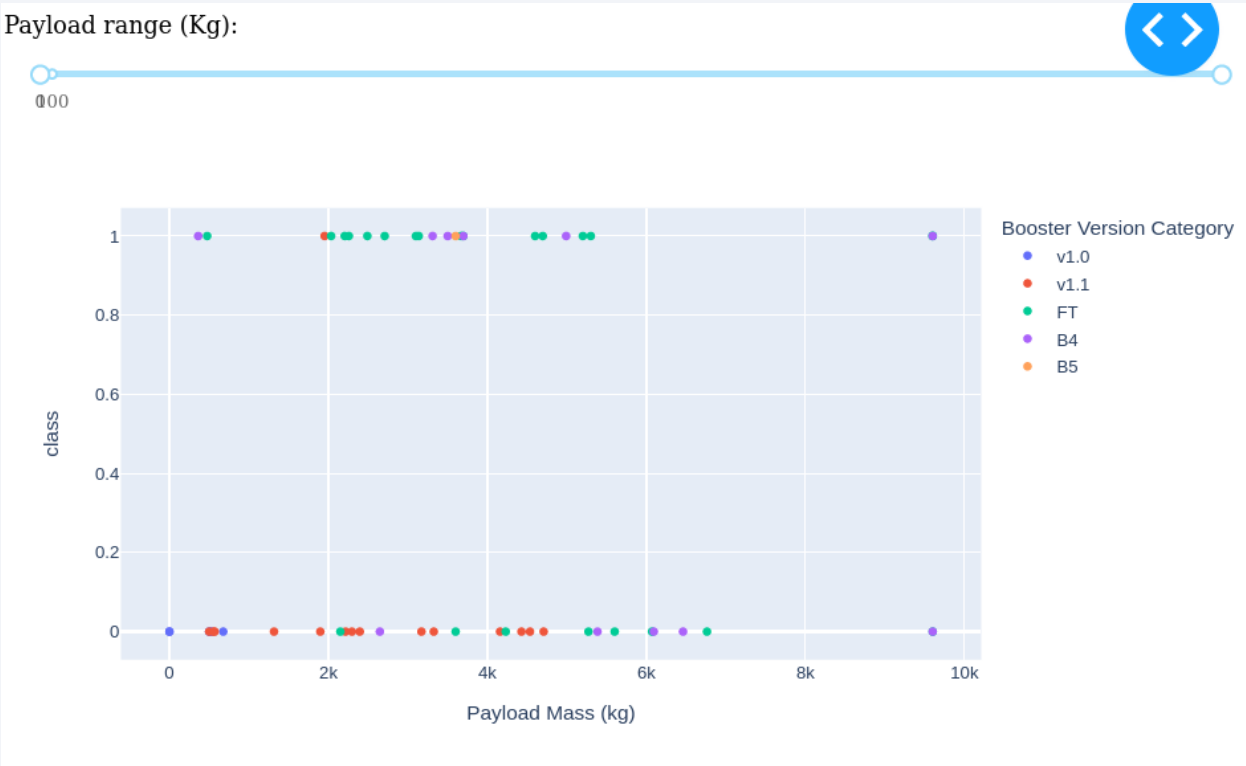
Launch Success Ratio for KSC LC-39A

- 76.9% of launches are successful on this site



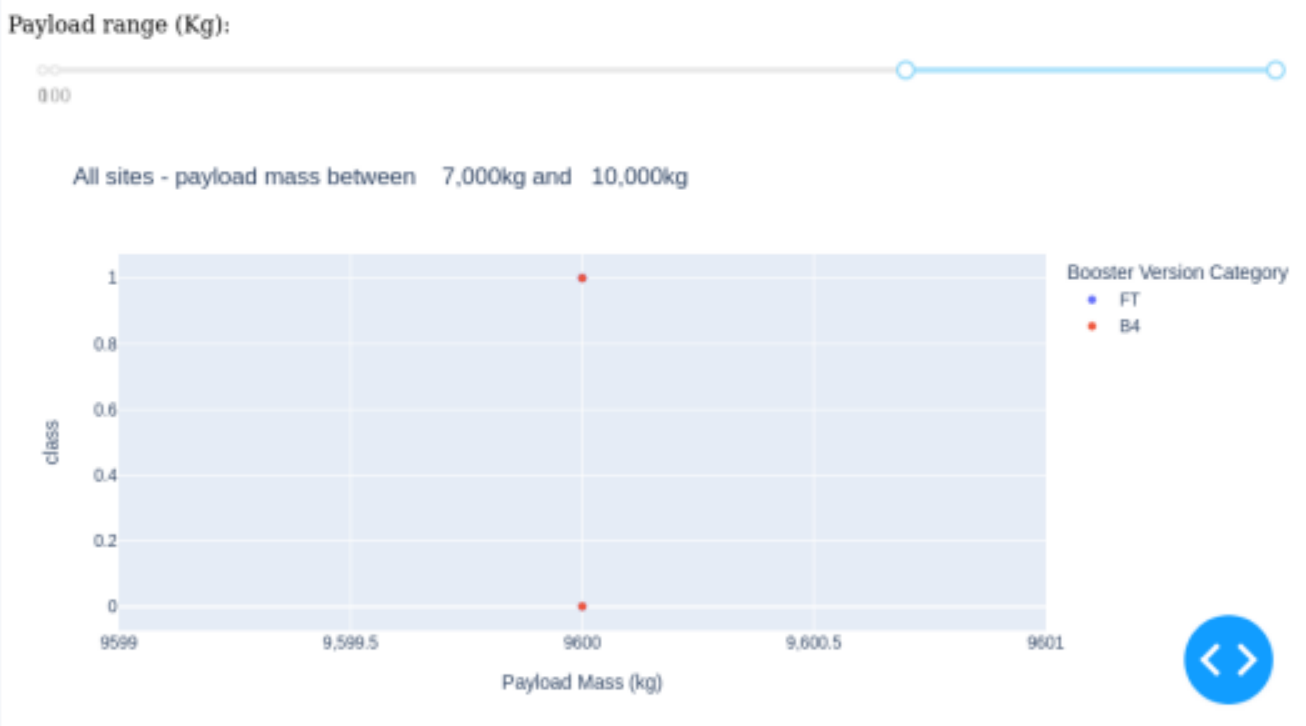
Payload vs. Launch Outcome

- Payloads under 6,000kg and FT boosters are the most successful combination.



Payload vs. Launch Outcome

- Not enough data to estimate the risk of launches over 7000 Kgs



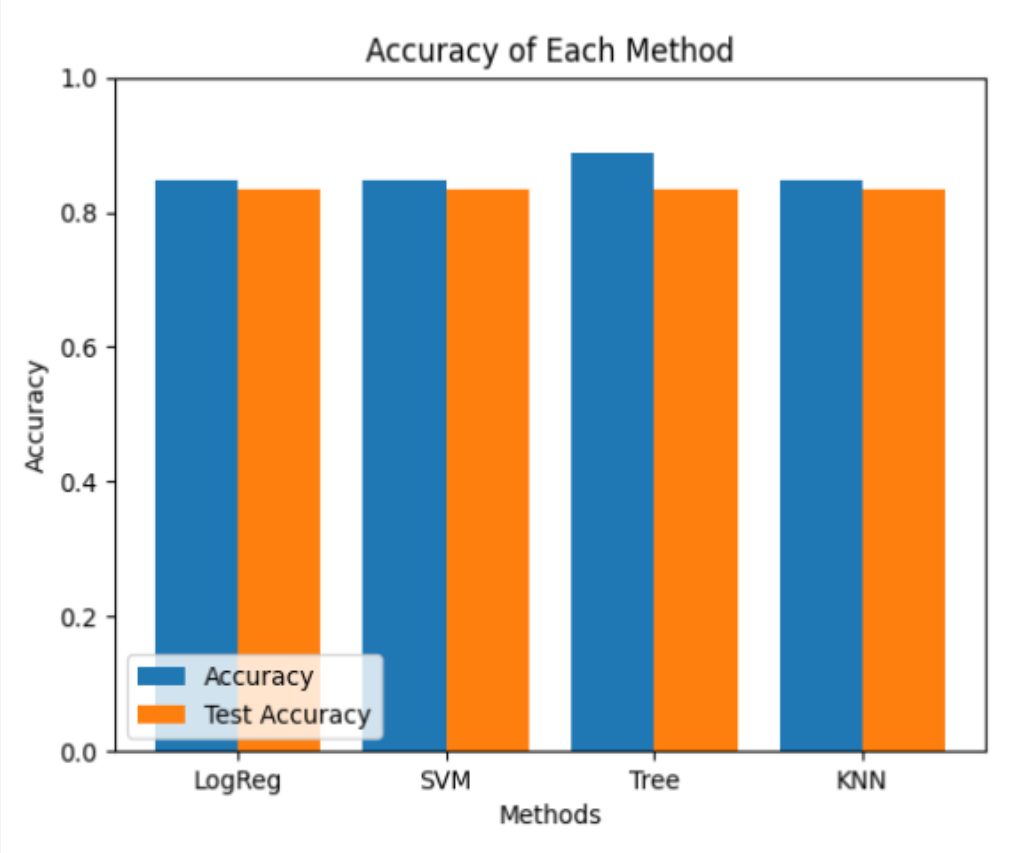
Section 5

Predictive Analysis (Classification)

Classification Accuracy

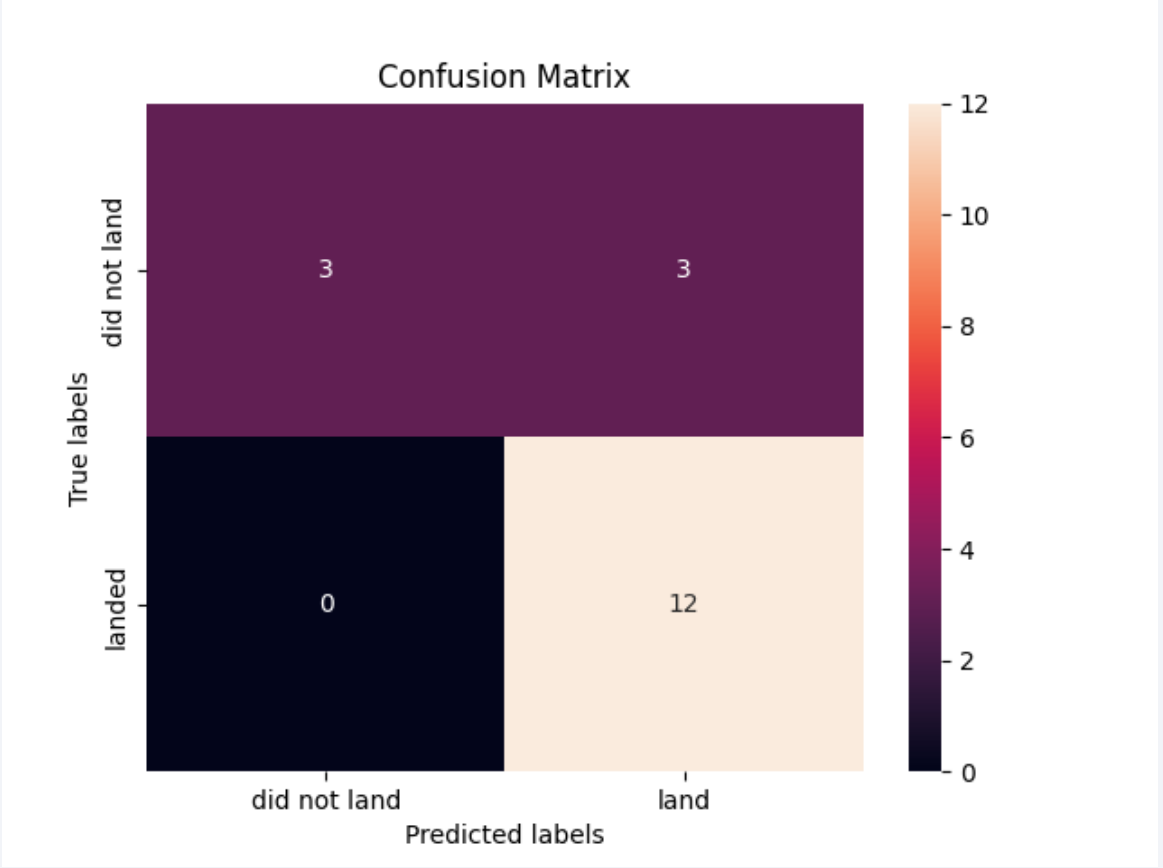
- 4 ML models were tested, and their accuracies are plotted in the bar graph-
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.8875	0.83333
KNN	0.84821	0.83333



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing a high value for true positive and true negative compared to the false ones.



Conclusions

- The launch sites are located near the coast and equator, and are close to rail/road/air ways for logistic conveniences
- The best launch site according to success rates is KSC LC-39A. However, CCAF5 SLC 40 seems to be more successful launch site considering more recent years.
- Payloads under 6,000kg and FT boosters are the most successful combination.
- Most mission outcomes are successful, and the success rate of launches has kept increasing over time – which explains the technological advancements (better rockets and processes) and a better understanding of factors affecting the launch outcomes
- Decision Trees Classifier can be used to predict successful landings and increase profits by cutting on the cost of stage 1 rockets.

Appendix

- np.random.seed **must be used while using ML models for reproducibility.**
- **Folium maps can be viewed by opening the notebook in ‘github.dev’ option**

Thank you!