IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ashish Upadhyay

20th June 2024

Outline

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

Executive Summary

- This project aims to predict if the first stage of the Falcon 9 Rocket will land successfully.

This will be done by the using Data Collection, Wrangling, EDA, Visualization, and then developing a Machine Learning model to accurately predict the landing outcome.

- The data is based on 4 distinct launch sites, where KSC LC 39A has the highest successful launch counts of all sites.

The Decision Tree Classifier model is found to be the best performing model, with an accuracy of 88% .

Introduction

- In today's Space age, having the lowest cost of launch is vital to any company to attract customers. SpaceX has a cost of $62 Million, far lower than any other company. This low cost is due to them reusing their first stage of the rocket. So, it is important to know if the first stage will land successfully or not.
- In this exercise, we want to know if the first stage will land successfully or not.

  This will help in determining the cost of the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API and Wikipedia.
- Perform data wrangling
  - Training labels were determined and added to the data frame as a column.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Grid Search was used to find the best hyperparameters for various models and then their accuracy was compared to find the best model.

Data Collection

- The datasets were collected from 2 sources-SpaceX API and Wikipedia.
- The data from the SpaceX API was collected with the "requests" library in python. The API gave a JSON file which was then parsed into a dataframe.
- The wikipedia data was collected using web scraping with the help of the python library "Beautiful Soup".

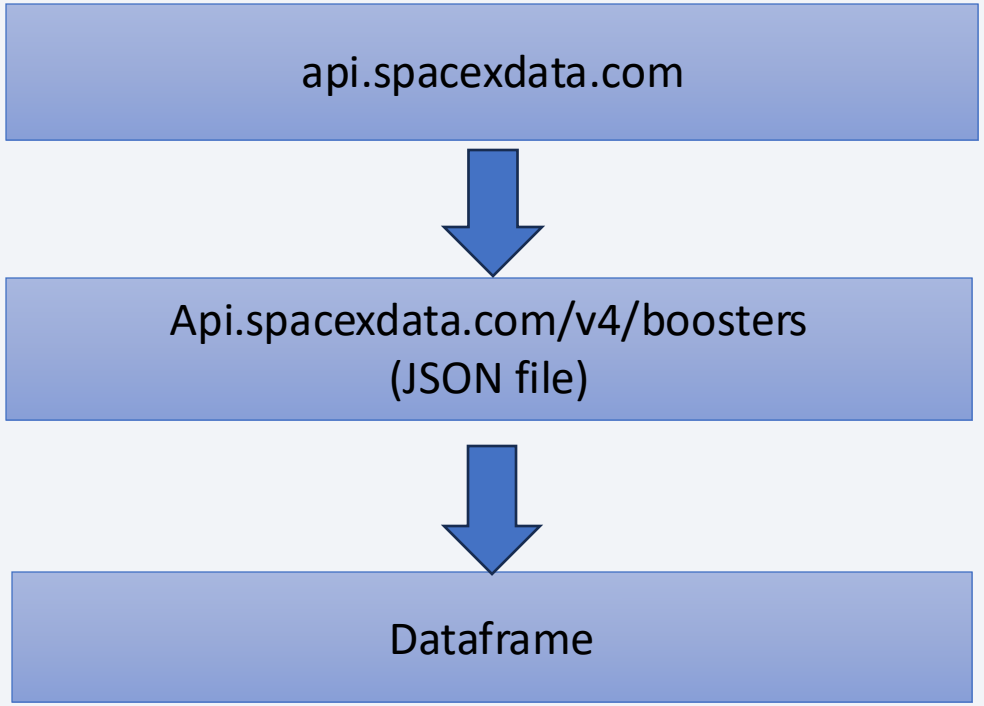
Data Collection – SpaceX API

- Data is collected by creating a response object. The JSON file is then converted to a Data Frame.

  Helper functions are then used to convert Ids to labels.

- Notebook link-

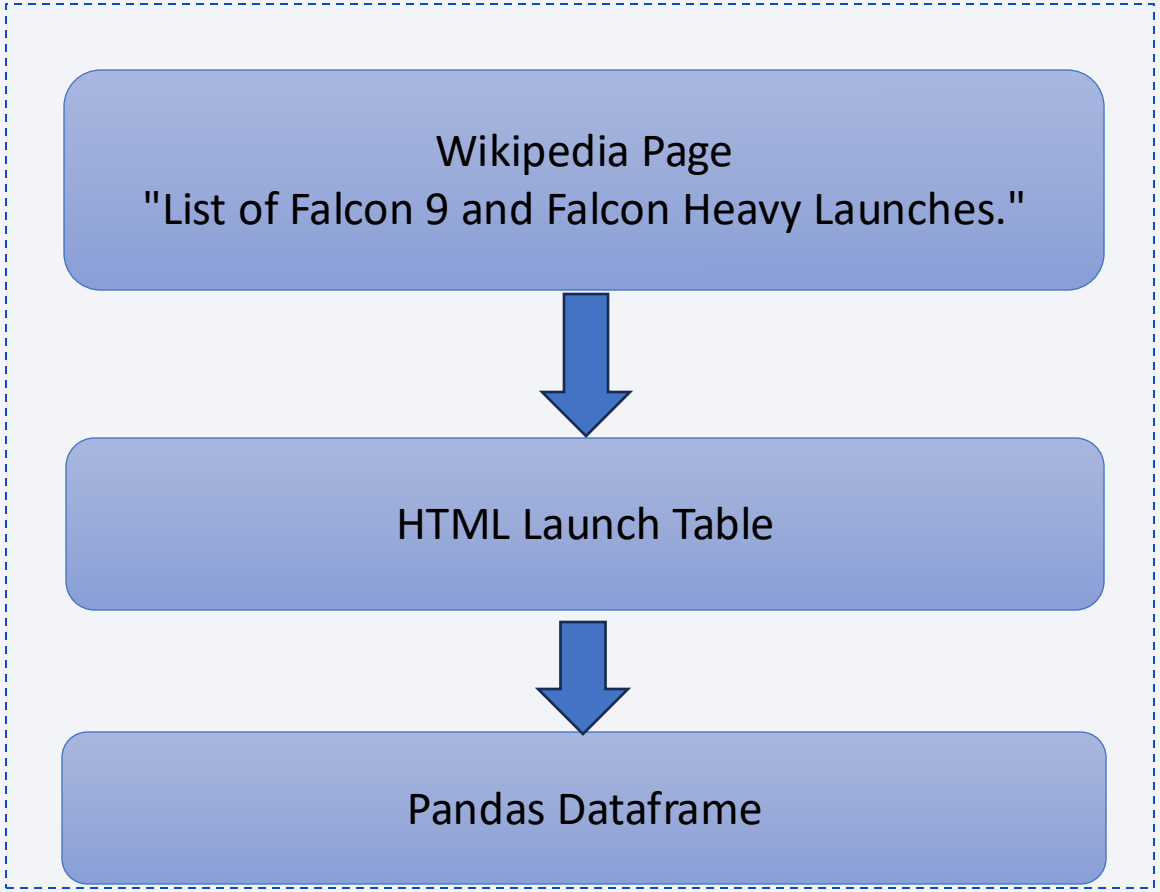
  Data Collection using API



Data Collection - Scraping

- The data is extracted from a table in the wikipedia page "List of Falcon 9 and Falc on Heavy Launches". The table is then parsed and converted to a pandas dataframe.
- Notebook link-

Data Collection using Web Scraping



Data Wrangling

- Basis Exploratory Data Analysis(EDA) is performed to find :-
  - -Percentage of null values in each column.
  - -No. of launches from each site.
  - -No. and occurrence of each orbit.
  - -No. and occurrence of mission outcome of each orbit.
- Notebook link-

  Data Wrangling

EDA with SQL

- SQL queries are executed to find (including but not limited to) :-

  -Names of unique launch sites.

  -Total payload mass carried from NASA missions.

  -Total no. Of successful and failed launch outcomes.

- Notebook link -

  EDA with SQL

EDA with Data Visualization

- Data visualization was used to display charts for the following :-

  -relationship between 'flight no.' and 'launch site'.

  -relationship between 'payload mass' and 'launch site'.

  -relationship between success rate of different orbits.

  -relationship between 'flight no.' and 'orbit type'.

  -the launch success yearly trend.

- Notebook link -

  EDA with Data Visualization

Build an Interactive Map with Folium

- The interactive map was created by using folium objects such as Circle, Marker, Marker Cluster, Polyline.
- The folium objects was used to :-

  -mark all launch sites on the map.

  -mark all success/failed launches for each site on the map.

  -calculate the distances between the launch site and its proximities.
- Notebook link -

  Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- The dashboard contains :-

  -a pie chart with a dropdown to show success count for selected/all sites.

  -a plot with range slider to show how success count varies with payload mass.

- The main insight which we are aiming for is how :-

  -how launch outcome varies with launch site.

  -how launch outcome is related to payload mass.

- Notebook link-

  Interactive Dashboard with Plotly Dash

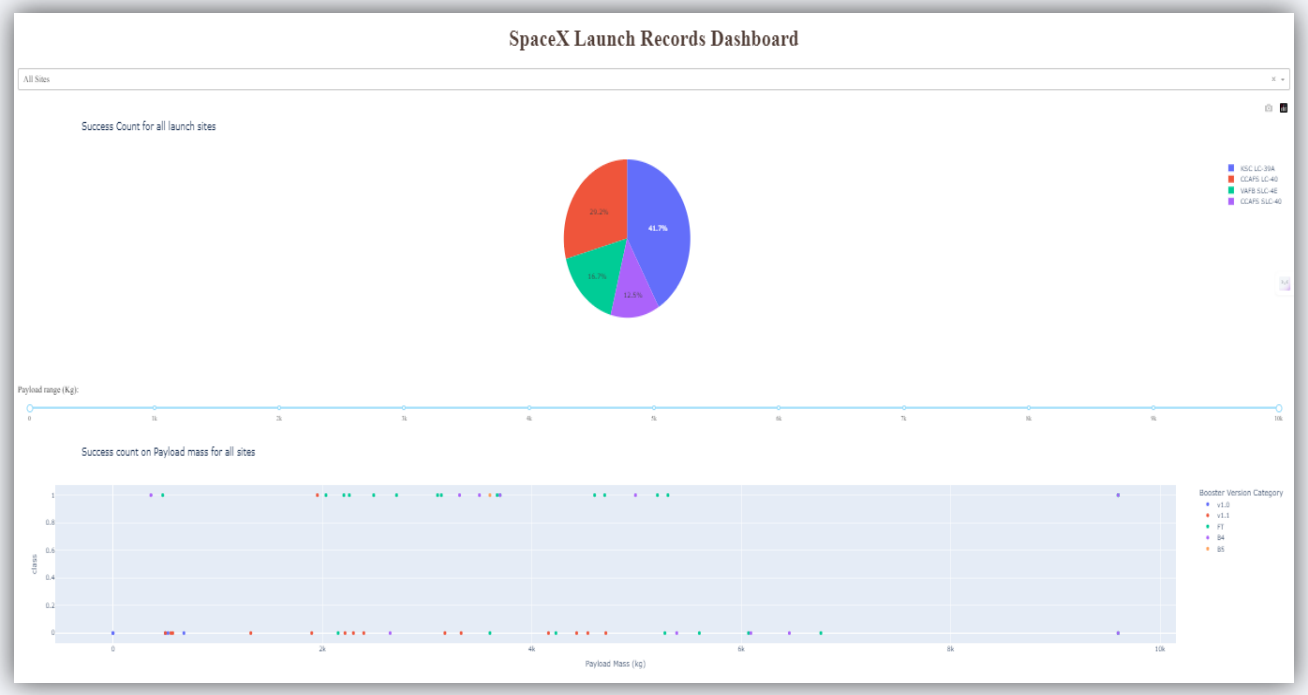
Predictive Analysis (Classification)

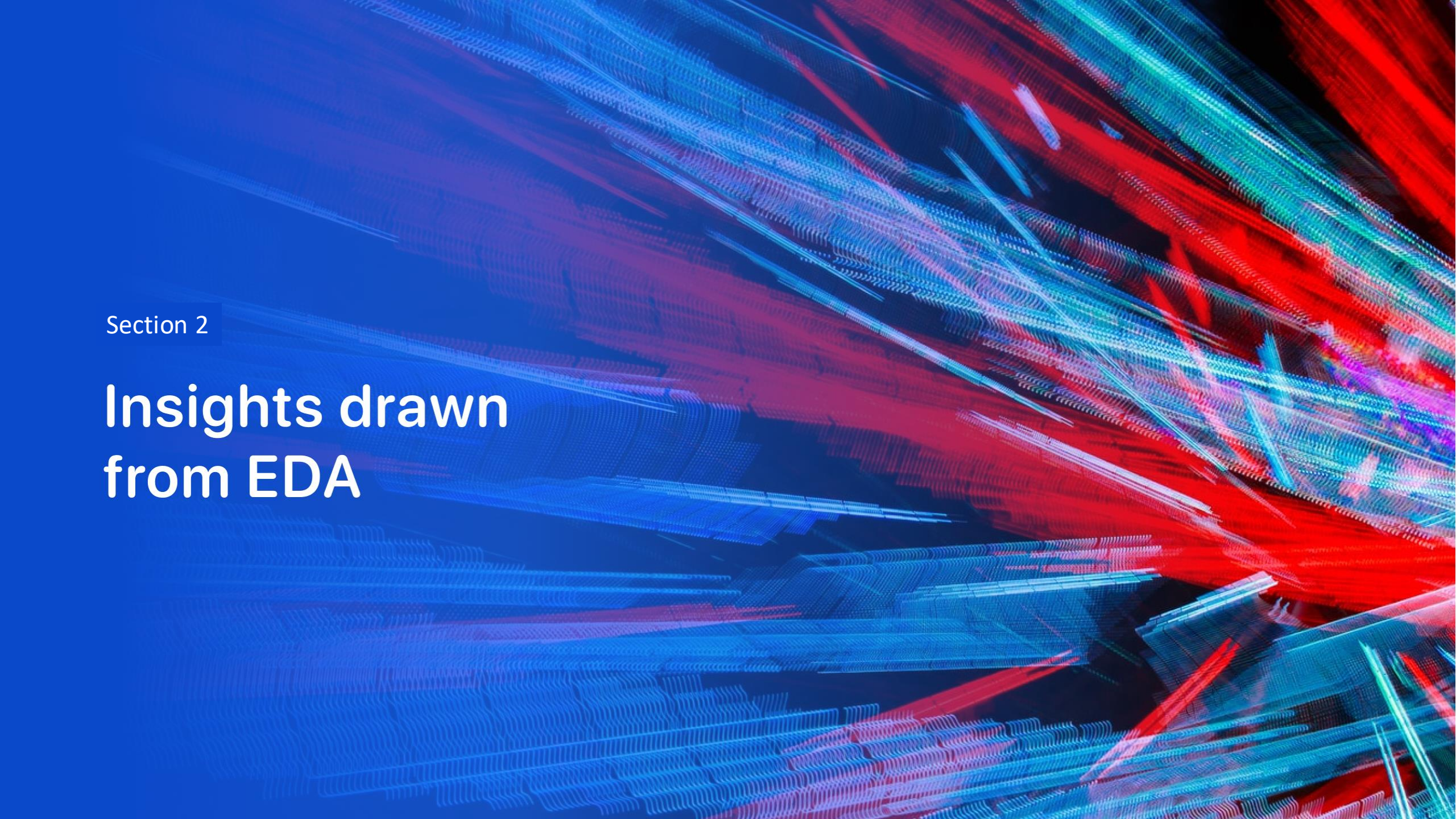
- Grid search method was used to train and finely tune various machine learning models by finding the best hyperparameters for each model.
- We first created an instance of the model, then GridSearchCV method was used to find the best hyperparameters and best accuracy. The accuracy score of all the models was then compared to find the best performing model.
- Notebook link-

  Machine Learning Prediction

Results

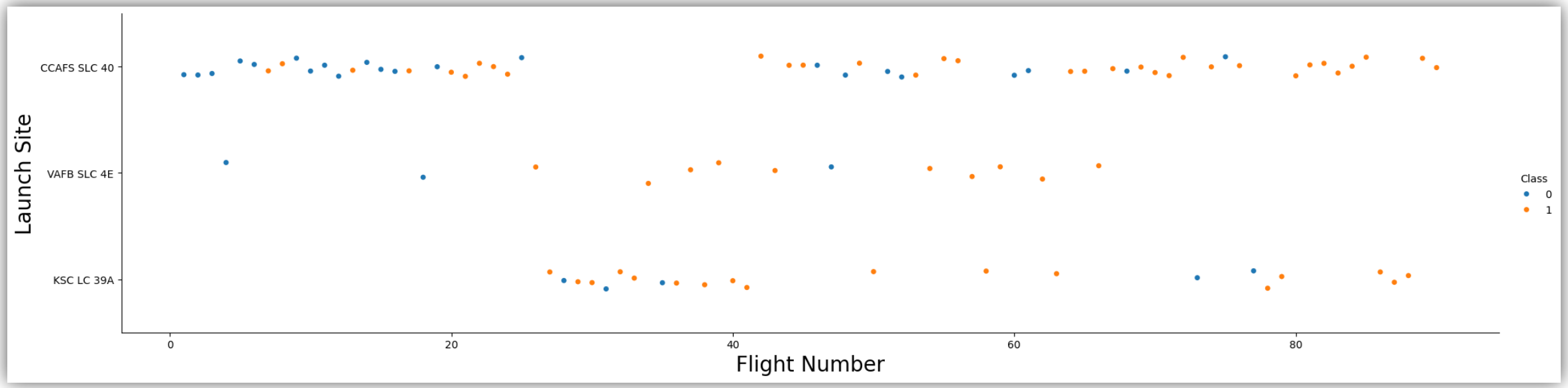
- We find how the launch outcome depends on other features by visualizing various plots to show these relations.
- The adjacent picture shows the resultant dashboard, with the pie chart and the scatter plot.
- The Decision Tree Classifier model is found to be the best performing model, with an accuracy of 88% .




Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- The picture is a scatter plot of Flight Number vs Launch site.
- As the no. of flights increase, 1st stages land successfully increasingly. This is true for all sites.



Payload vs. Launch Site

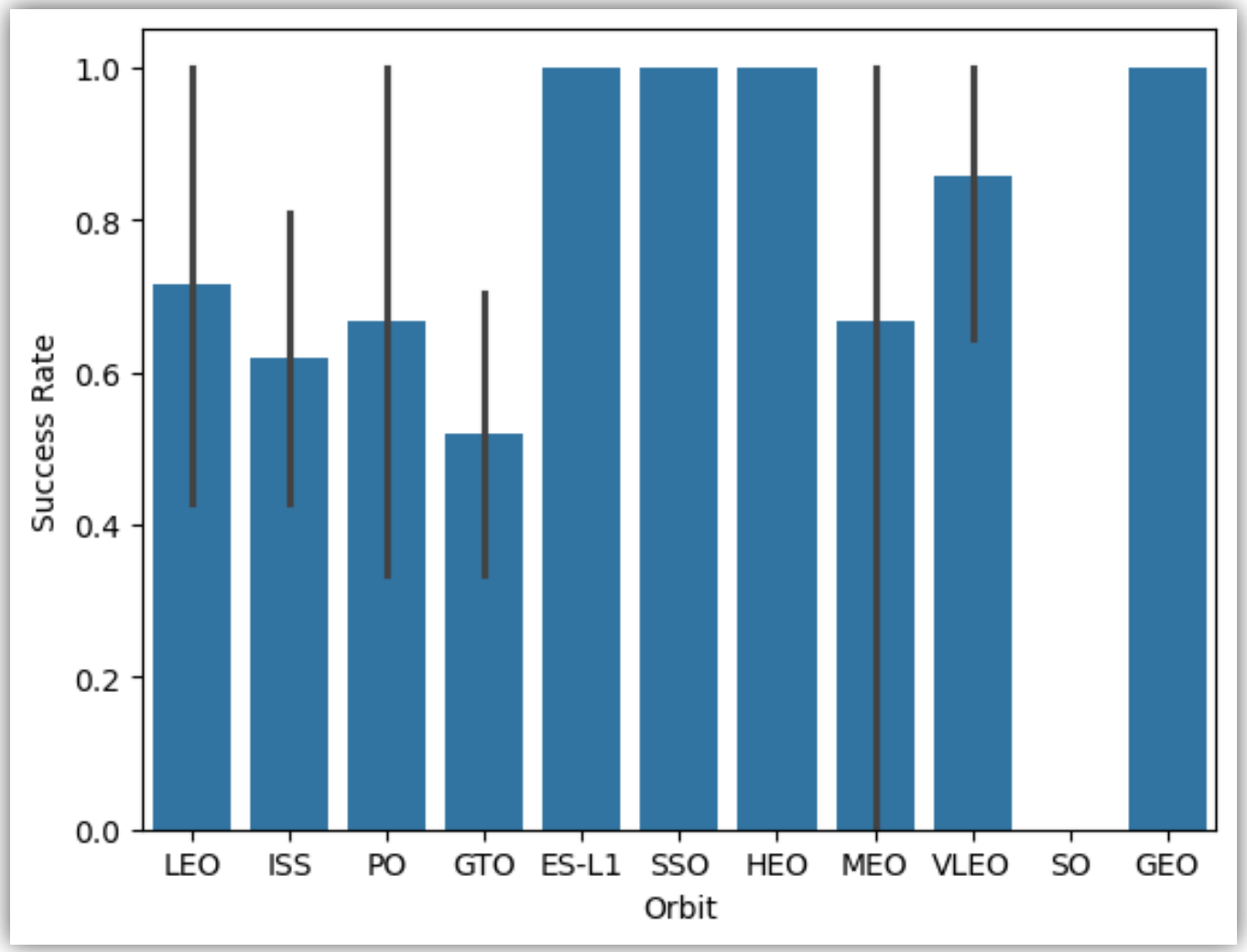
- The picture is a scatter plot of Payload vs. Launch Site.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

- The picture is a bar chart for the success rate of each orbit type.
- The orbits -

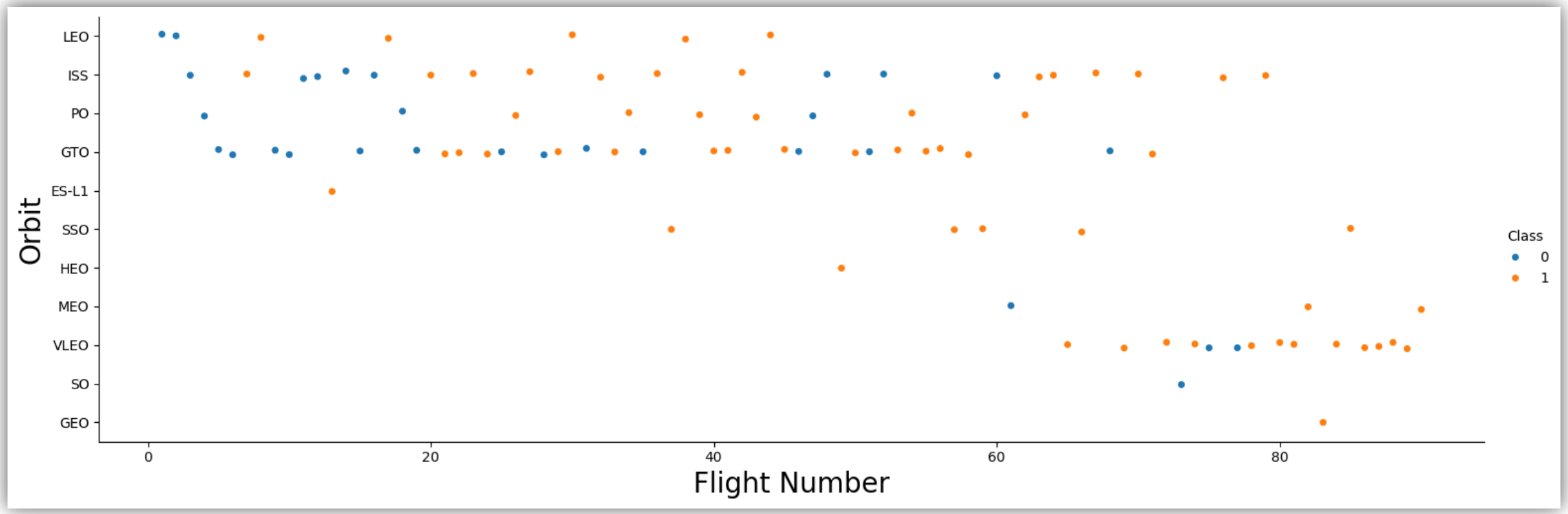
  'ESL1', 'SSO', 'HEO', 'GEO' have a 100% Success Rate.

  'SO' has 0% Success Rate.



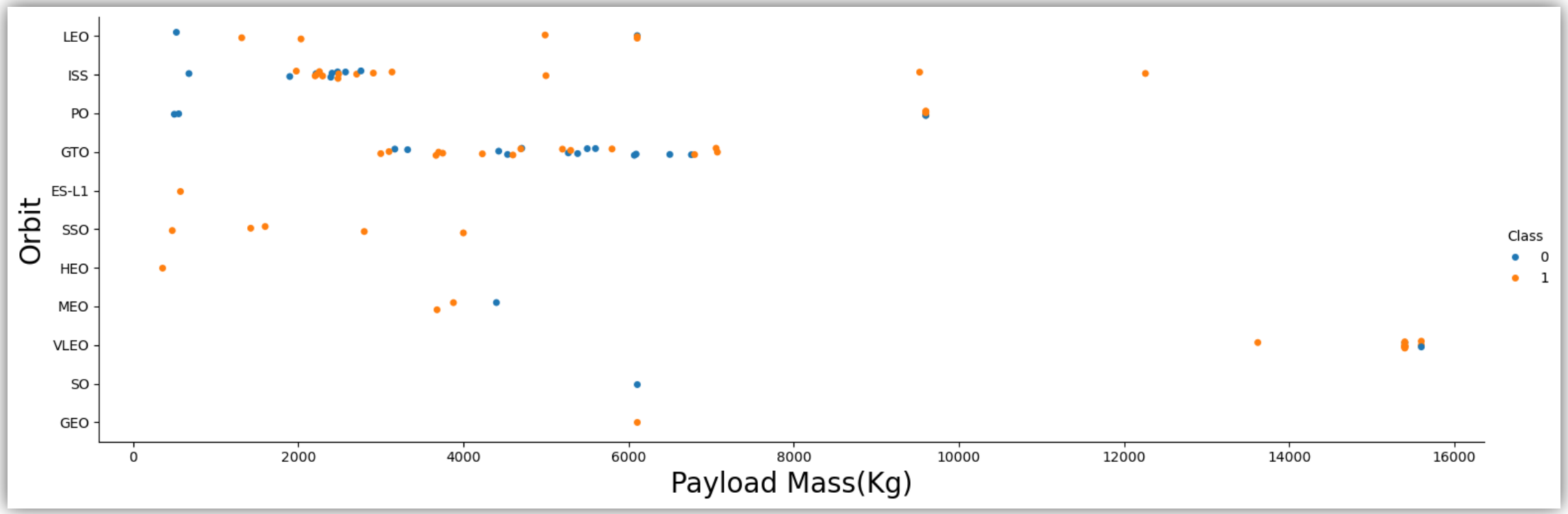
Flight Number vs. Orbit Type

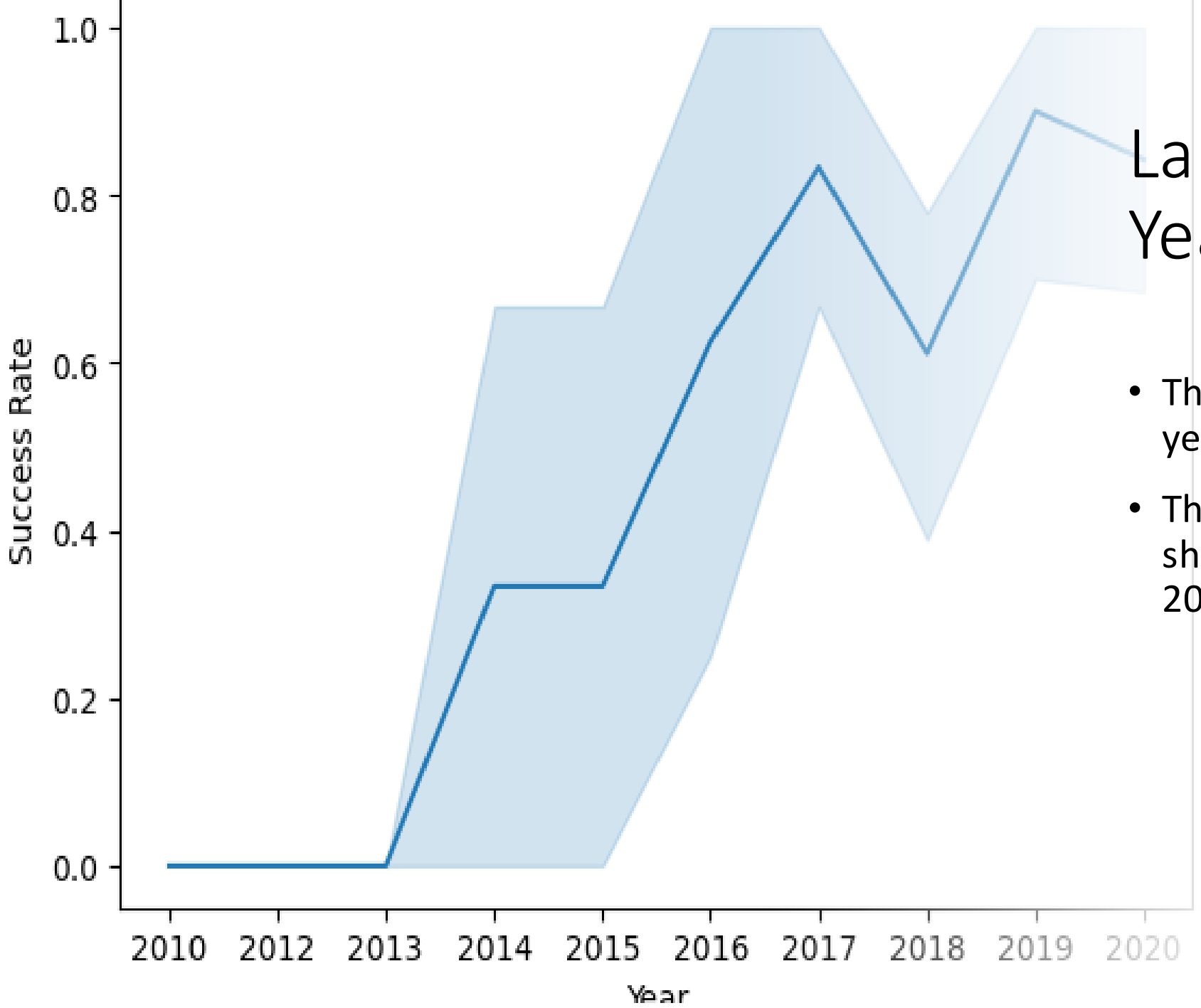
- The picture shows a scatter point of Flight number vs. Orbit type.
- In LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

- The picture shows a scatter point of payload vs. orbit type.
- With heavy payloads the successful landing rate is more for Polar, LEO and ISS orbits.





Launch Success Yearly Trend

- The picture is a line chart of yearly average success rate.
- The yearly success rate has sharply increased after 2013.

All Launch Site Names

- A query is run to find the names of the unique launch sites.
- The results show that there are 4 unique launch sites.
- The resultant table is shown below.

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

Launch Site Names Begin with 'CCA'

- A query is run to find 5 records where launch sites begin with `CCA`.
- The resulting table is shown below.

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

- A query is run to calculate the total payload carried by boosters from NASA.
- The results show that the total payload mass from NASA was 45596 Kgs.

sum("PAYLOAD_MASS__KG_"
45596

Average Payload Mass by F9 v1.1

- A query is run to calculate the average payload mass carried by booster version F9 v1.1
- The results show that the average payload mass carried by F9 v1.1 is 2928.4 Kgs.

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- A query is run to find the dates of the first successful landing outcome on ground pad
- The results show that the first successful ground landing date was 22-12-2015.

min("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- A query is run to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- The results below show the 4 versions that fulfill the requirements.

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

Total Number of Successful and Failure Mission Outcomes

- A query is run to calculate the total number of successful and failure mission outcomes
- The resulting table is shown below.

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

Boosters Carried Maximum Payload

- A query is run to list the names of the booster which have carried the maximum payload mass.
- The resulting table is shown on the right.

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- A query is run to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- The resulting table is shown below.

Month	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

- A query is run to 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.
- The resulting table is shown below.

LANDING_OUTCOME	TOTAL	DATE
No attempt	21	2012-05-22
Success (drone ship)	14	2016-04-08
Success (ground pad)	9	2015-12-22
Failure (drone ship)	5	2015-01-10
Controlled (ocean)	5	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

Section 3

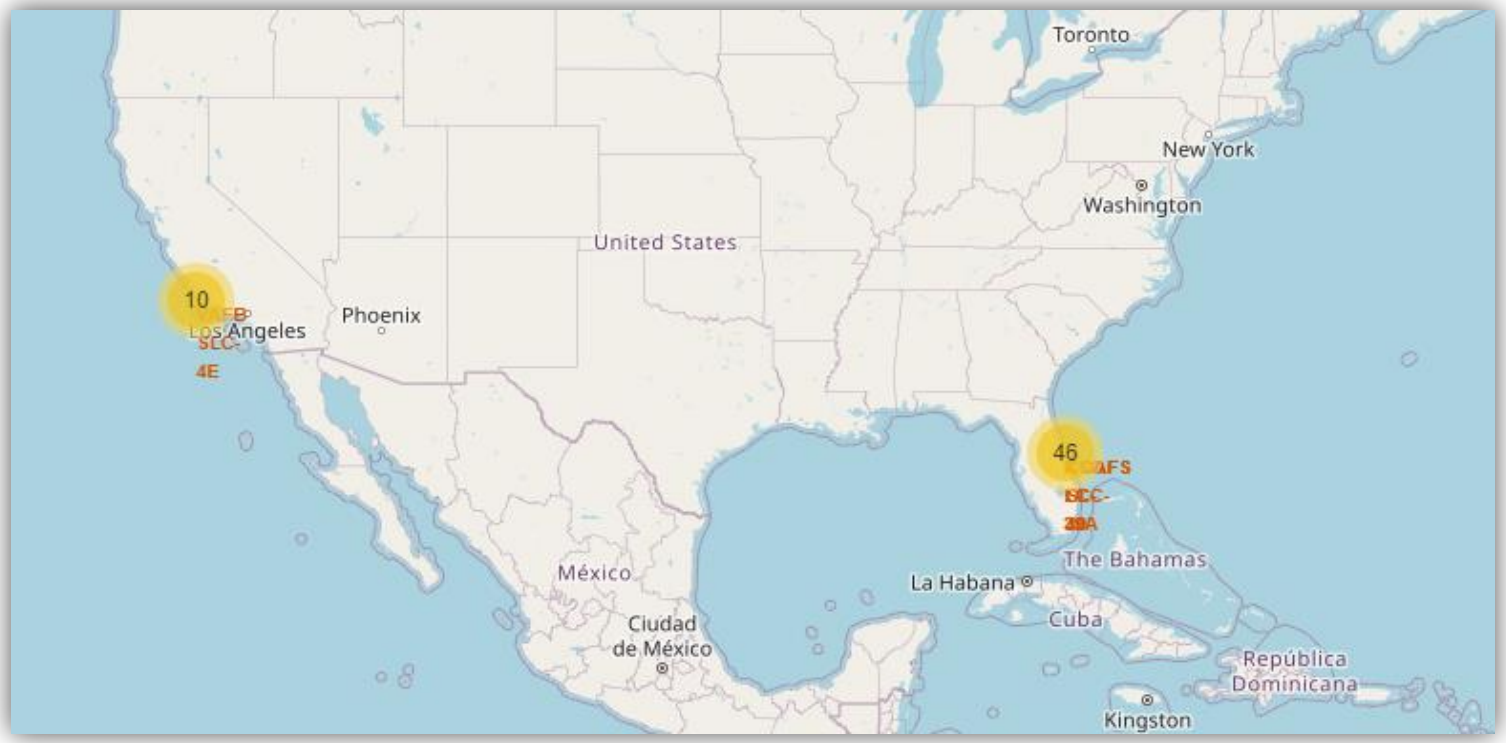
Launch Sites Proximities Analysis

Launch Site Locations

- The 2 main launch locations are marked on the map using folium markers.
- The output map, which is shown below, shows us the locations of the 2 launch sites.

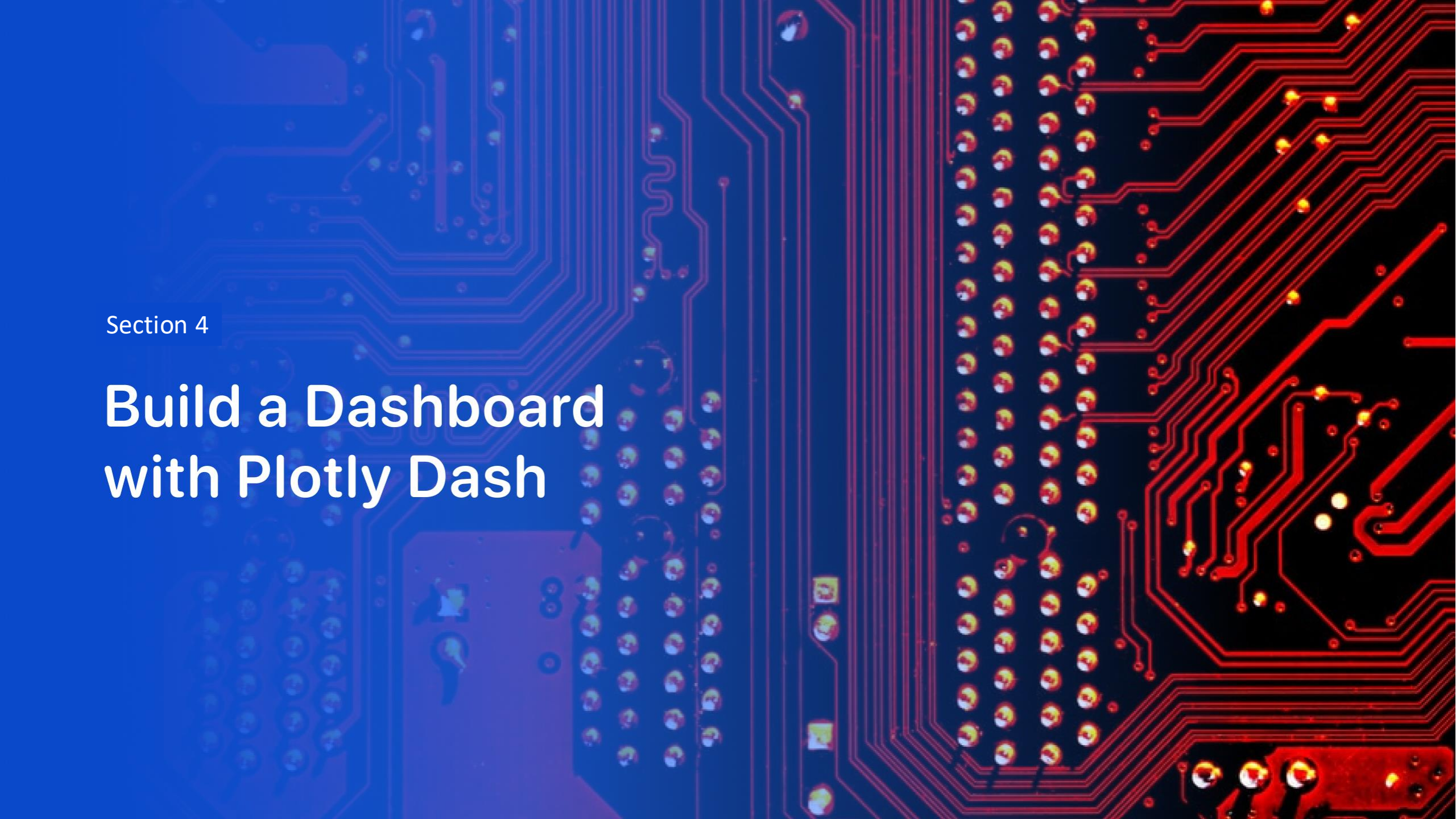
Launch Outcomes

- Each launch site is color coded by the number of successful launches.
- It was done by using folium 'marker cluster'.
- The image on the right shows the resulting map.



Launch Site Proximities

- The distances form the launch site to the closest coastline, highway and city are c alculated.
- Folium polyline objects are then created to draw lines from launch site to its proximities.
- The picture on the right shows the lines.


Section 4
Build a Dashboard with Plotly Dash

Total Success Count

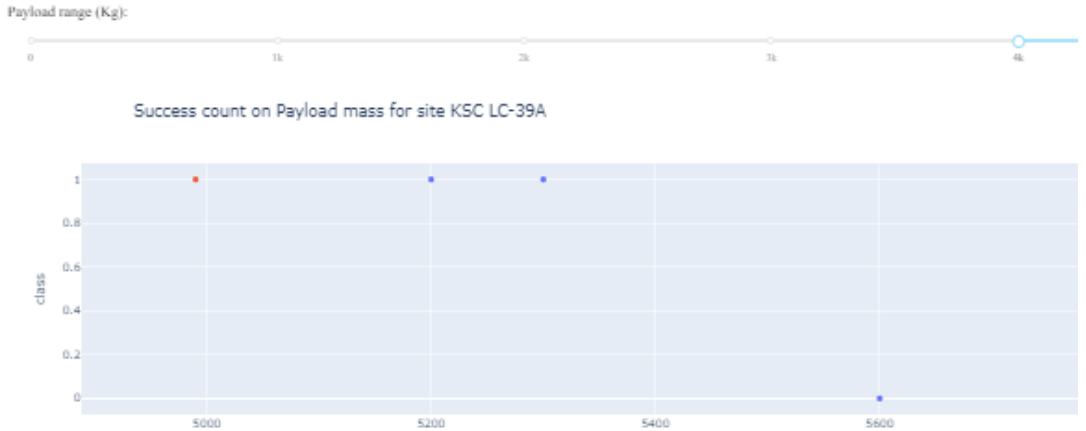
- The picture shows a pie chart of launch success count for all sites.
- It shows that site 'KSC LC-39A' has the highest number of successful launch counts of all sites.

Success Count for 'KSC LC-39A'

- The picture below shows the pie chart for success ratio of the launch site 'KSC LC-39A.
- The chart shows that about 77 % of the launches were successful (Blue) and 23% were unsuccessful (Orange).

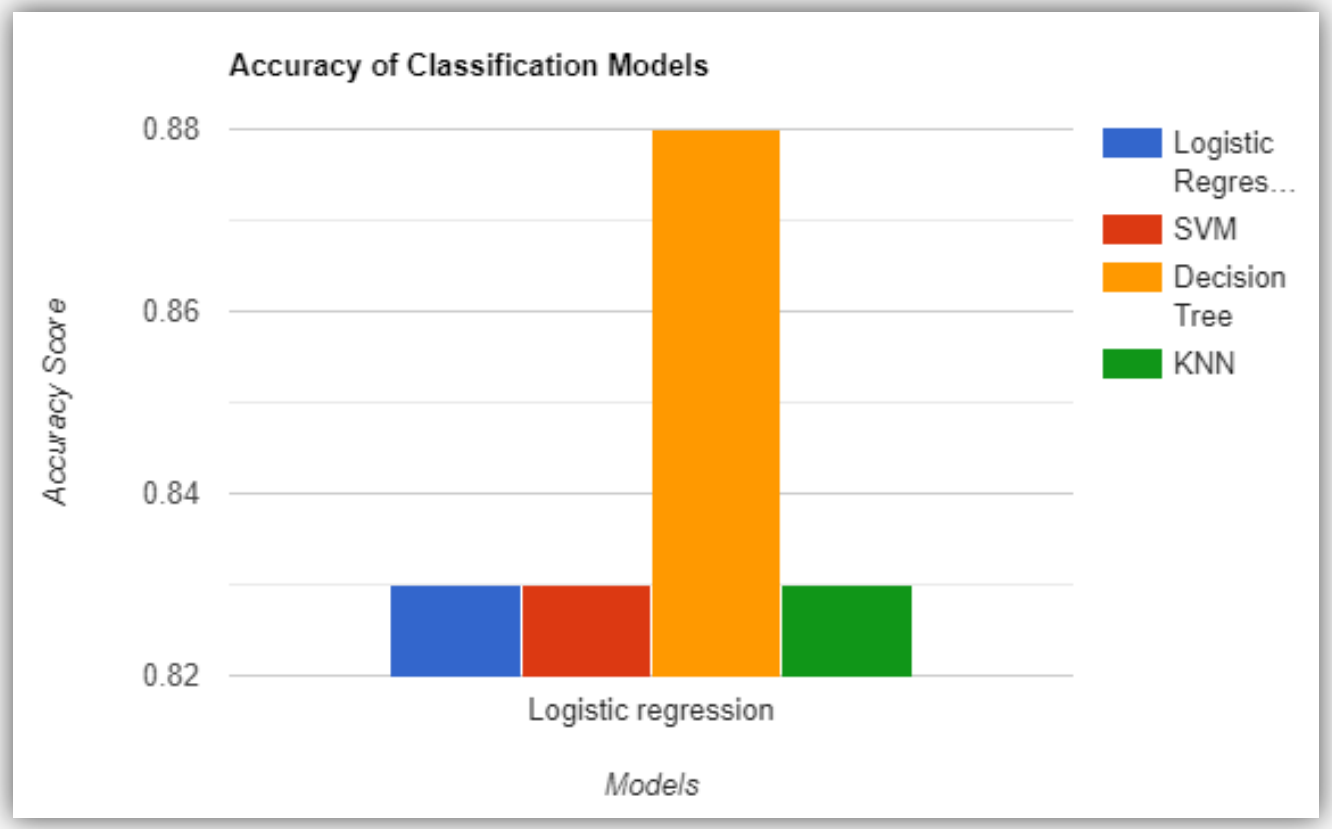
Relation between Payload and Mission Outcome

- The pictures on the right show scatter plots of Payload mass vs Class ,grouped by booster version.
- It is seen that when Payload mass goes above 3000 Kgs, the number of successful mission outcomes increase more and more.



Section 5

Predictive Analysis (Classification)

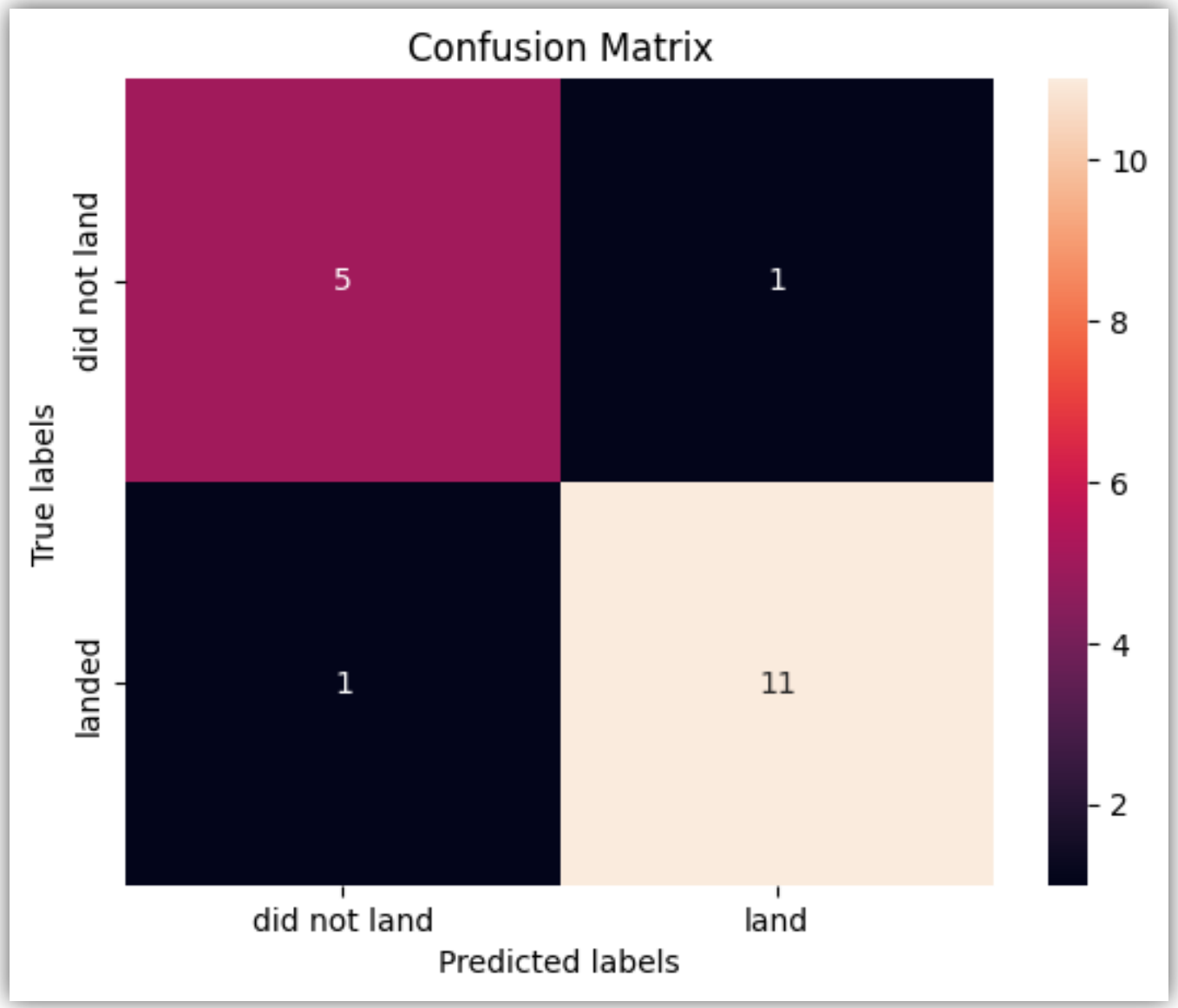


Classification Accuracy

- The image depicts a bar chart comparing accuracy of all 4 classification models th at were trained.
- The 'Decision tree classifier' model has an accuracy of 0.88 , which is the highest of all models.

Confusion Matrix

- The image shows the confusion matrix for the Decision tree classifier model.
- There were only 2 wrongly classified cases , 1 each for successful and unsuccessful o utcomes.



Conclusions

- After fitting and evaluating various models, it is decided that 'Decision tree classifier' model is the best for predicting the launch outcomes.
- The model is 88% accurate in predicting the launch outcome successfully.
- The best hyperparameters for the model are 'entropy' criterion and a maximum depth of 4.

Appendix

- Logistic Regression
- Support vector machines (SVM)
- Decision tree classifiers
- K-nearest neighbors (KNN)
- Confusion Matrix

Thank you!