IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nikita Rogozhnikov
21.07.2023

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection using web scrapping and SpaceX API
  - Data Wrangling
  - Exploratory Data Analysis (EDA) using SQL and data visualization
  - Interactive visual analytics and dashboards
  - Predictive analysis (classification) and Grid Search
- Summary of all results:
  - The most common launch sites, orbits, landing spots and payload mass
  - The optimal orbit for a particular payload mass
  - The best parameters for each classification model
  - The model with the highest score is Decision Tree

Introduction

- The objective is to evaluate whether SpaceY can compete with SpaceX and what is needed for a successful launch
- Questions:
  - What is the best location for the launch?
  - What is the best orbit?
  - What is the optimal payload mass?
  - What is the best option for landing?
  - Determine the overall cost for the expected successful launch with the parameters above

Section 1

Methodology

Methodology

Executive Summary

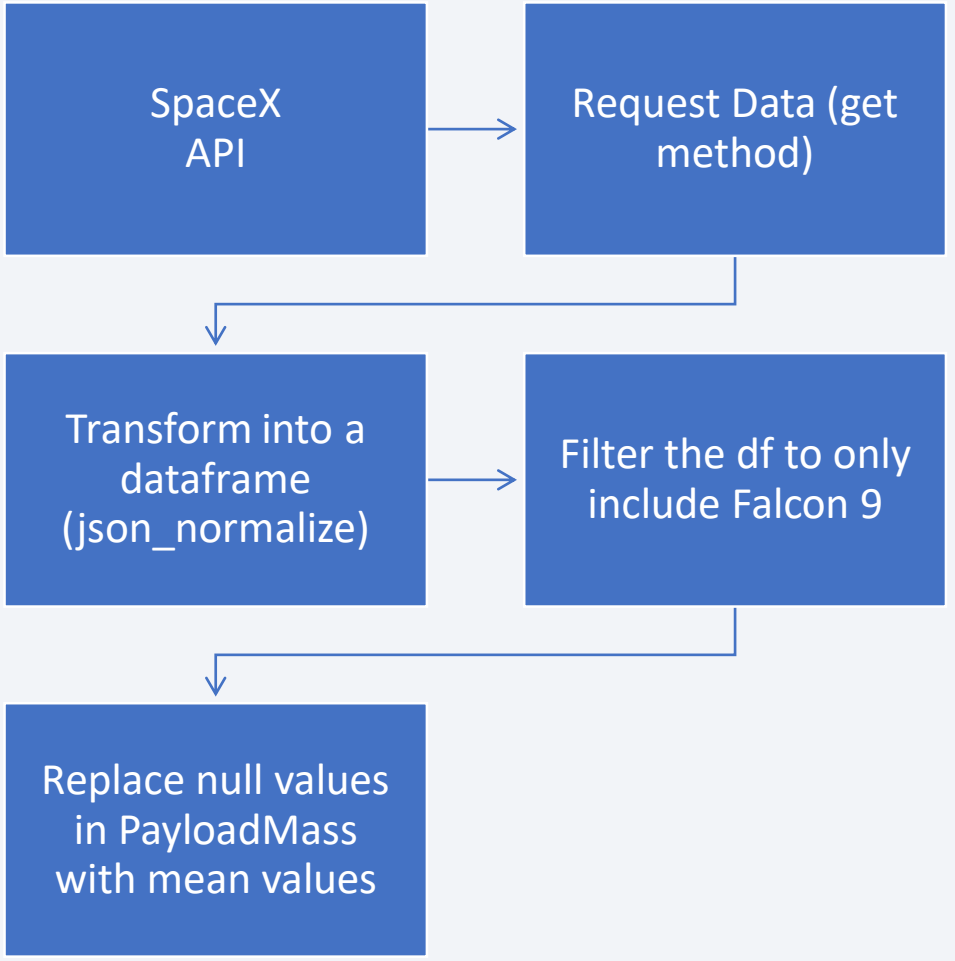
- Data collection methodology:
  - Data was collected from SpaceX API using the get method and from Wiki using the get and beautiful soup
- Perform data wrangling
  - Check the missing values, calculate the number of launches for different orbits, launch sites and mission outcomes. Check the overall successful rate
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create objects for each model, set parameters, use the GridSearch to identify the best parameters, compare the accuracy score

Data Collection

- Data was collected from the SpaceX API using the get method and, consequently, was transformed into a pd dataframe.
- Another source was Wikipedia. The data was collected using the webscrapping method (get request, beautiful soup, html tables). Therefore, a data frame was created by parsing the launch HTML tables.

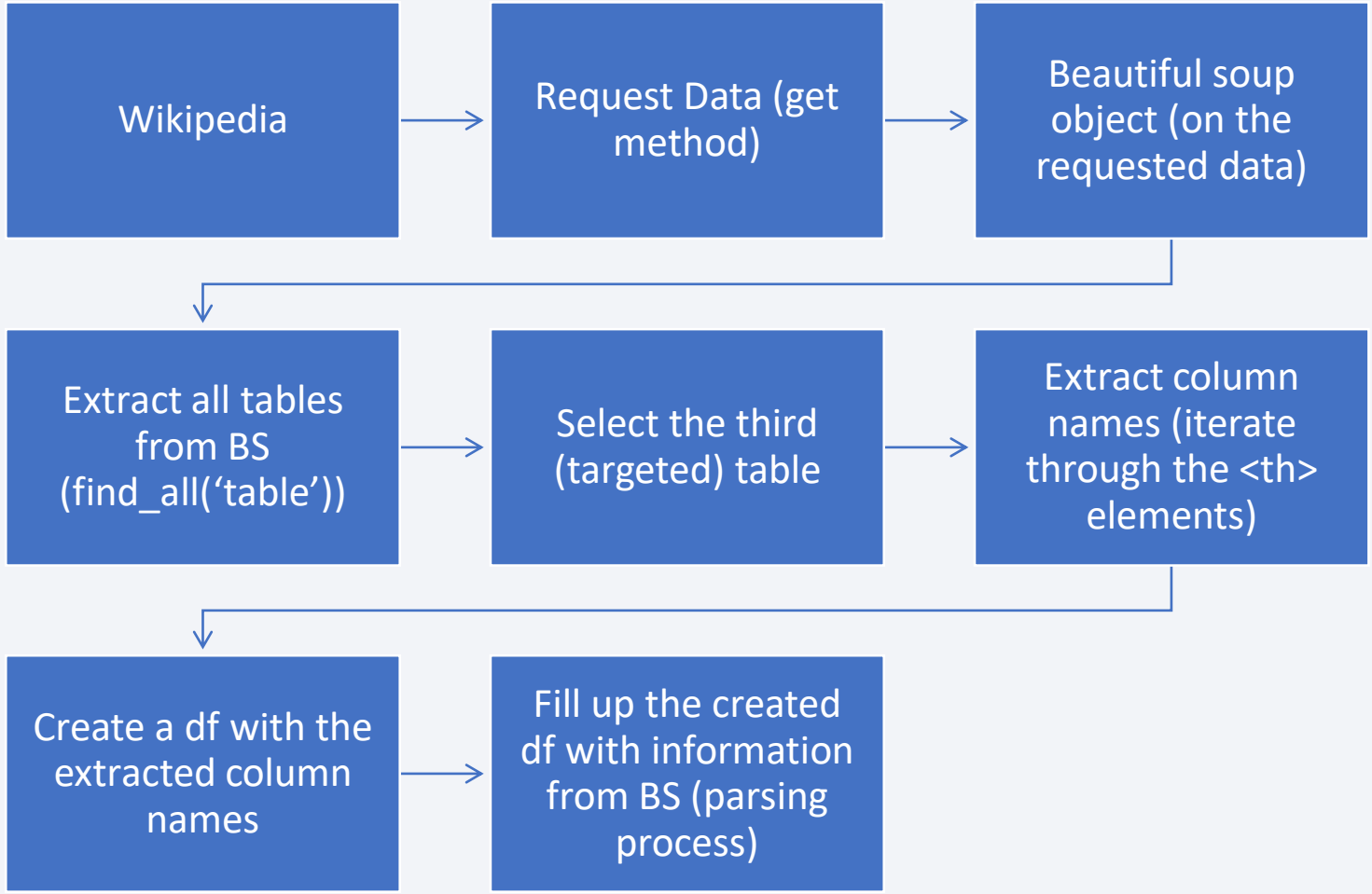
Data Collection – SpaceX API

- Github link



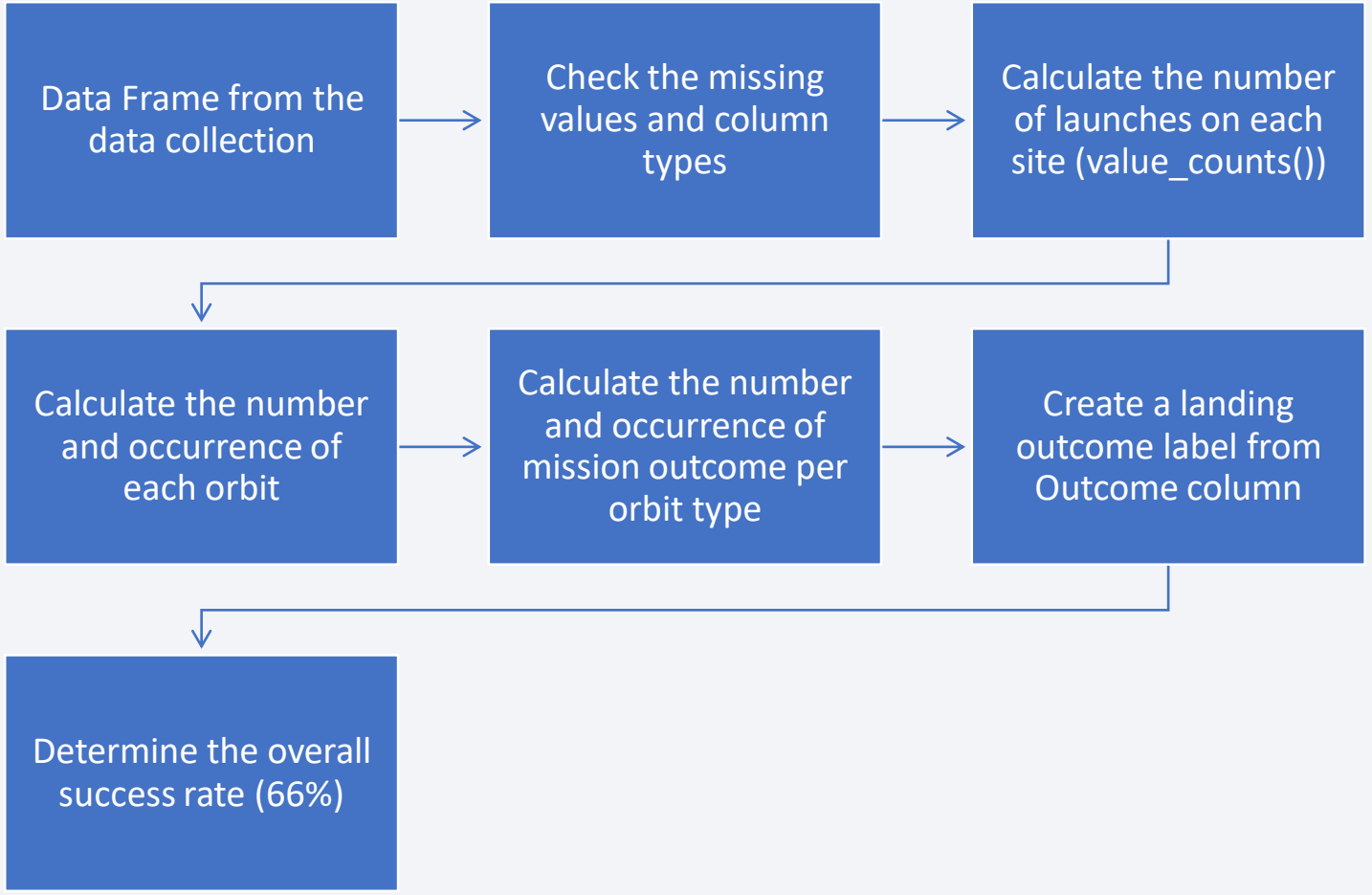
Data Collection - Scraping

- Github link



Data Wrangling

- Github link



EDA with Data Visualization

- Github link

Chart	Why was used
Scatter plot: Payload vs Flight Number	To show the relation of the outcome of the launch on Payload Mass and Flight Number
Scatter plot: Launch Site vs Flight Number	To show the relation of the outcome of the launch on Launch Site and Flight Number
Scatter plot: Launch Site vs Payload	To show the relation of the outcome of the launch on Launch Site and Payload Mass
Bar chart: Success rate vs Orbit type	To show the relation between an orbit type and the success rate
Scatter plot: Orbit type vs Flight number	To show the relation of the outcome of the launch on Orbit type and Flight Number
Scatter plot: Orbit type vs Payload	To show the relation of the outcome of the launch on Orbit type and Payload Mass
Line chart: Success rate vs Year	To show the development of the success rate

EDA with SQL

- Github link
- SQL queries performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - 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.
  - 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.

Build an Interactive Map with Folium

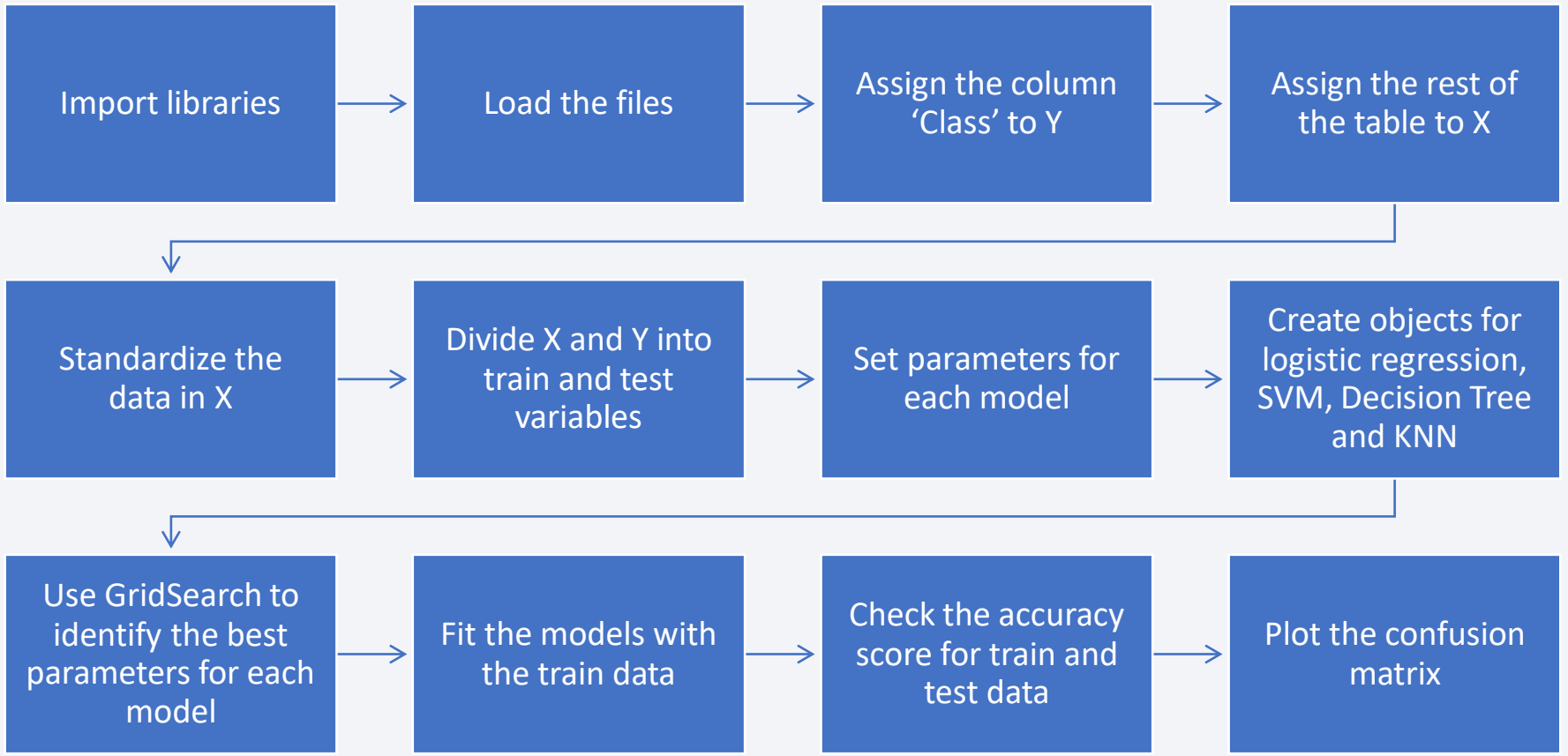
- Github link
- Map objects created:
  - Circles – to see the positions of launch sites
  - Markers – to see the names of launch sites
  - Marker Clusters – to see exact locations of launches in a cluster form (successful / failed)
  - Mouse position (to see coordinates) – to calculate the distance to the nearest coastline / road etc.
  - Lines – to show the distance to the nearest coastline / road / railway / city

Build a Dashboard with Plotly Dash

- Github link
- Plots and interactions added:
  - Dropdown list – to choose a launch site
  - Pie chart – to see either the relation of successful outcomes between different launch sites (for ALL) or the relation of successful and failed outcomes for a particular site
  - Payload range slider – to interactively set up a filter of payload mass
  - Scatter plot – to see the relation of the outcome on the payload mass and booster version category

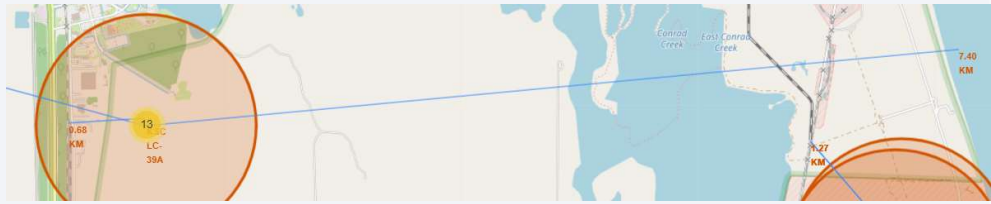
Predictive Analysis (Classification)

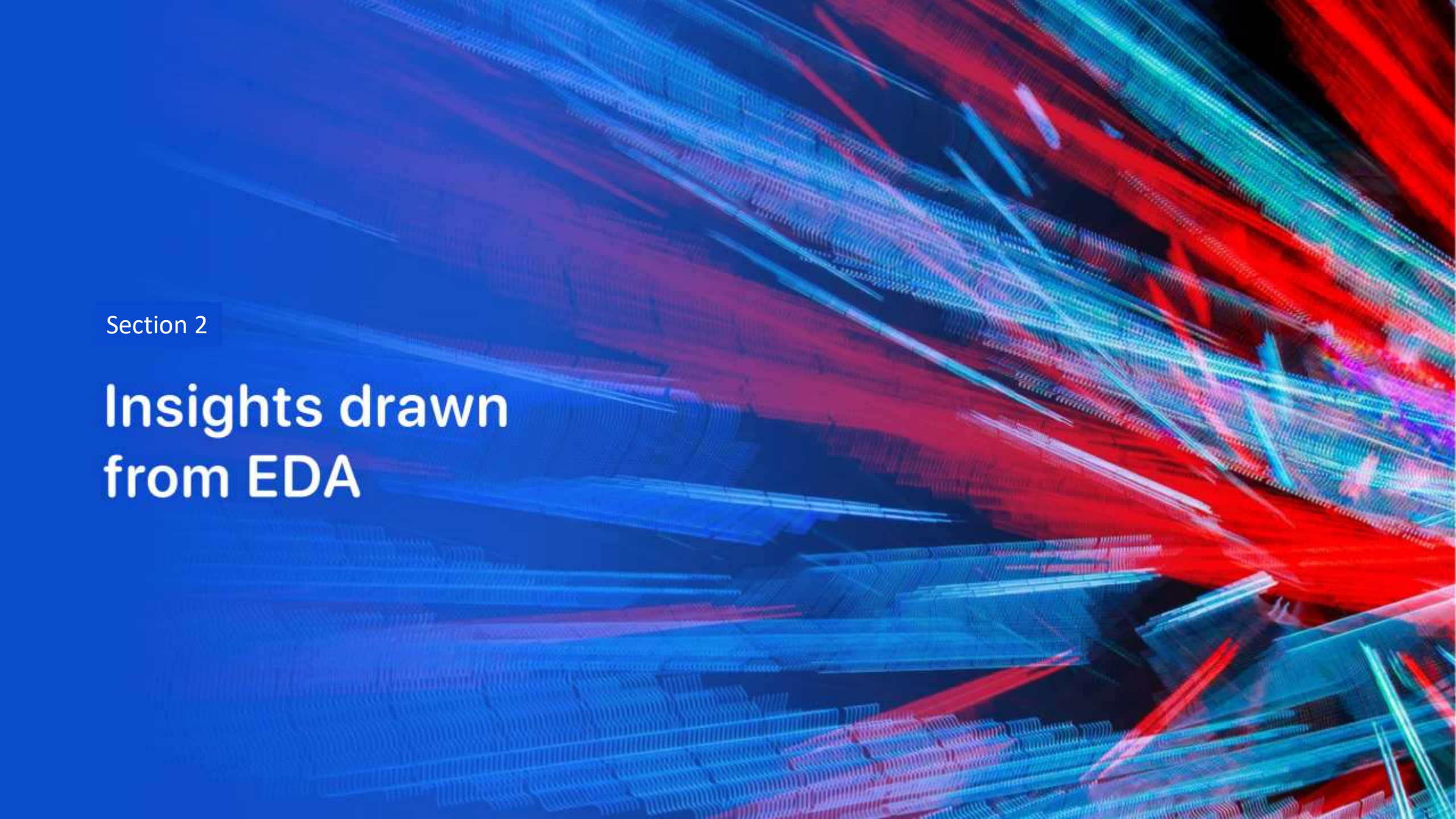
- Github link



Results

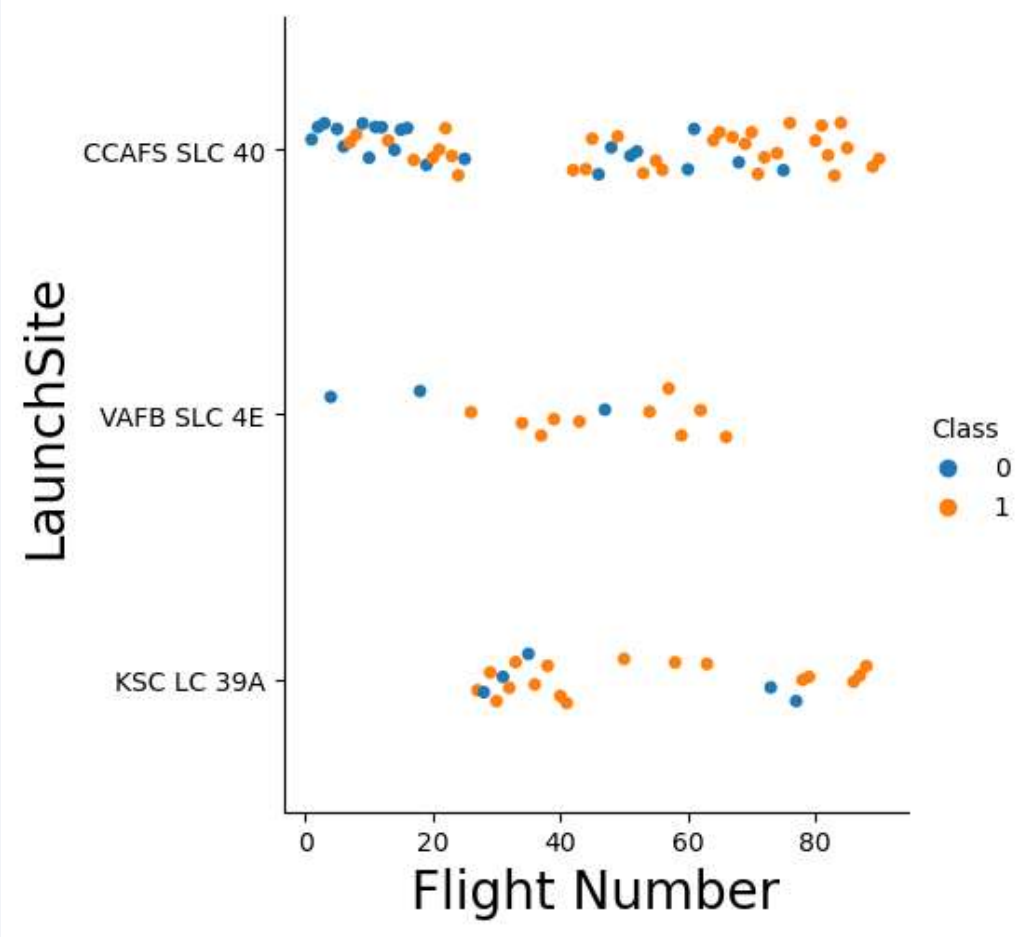
- Exploratory data analysis results
  - SpaceX uses 4 different launch sites
  - The first successful landing was in 2015
  - The success rate was increasing till 2019 and more than doubled compared to the first years
- Interactive analytics demo in screenshots
  - Most of the launches happened in the east cost
  - The launch site are far from cities but close to coastlines, roads and railways
- Predictive analysis results
  - The most successful model is Decision Tree with 94% accuracy scored based on the test data




Section 2
Insights drawn
from EDA

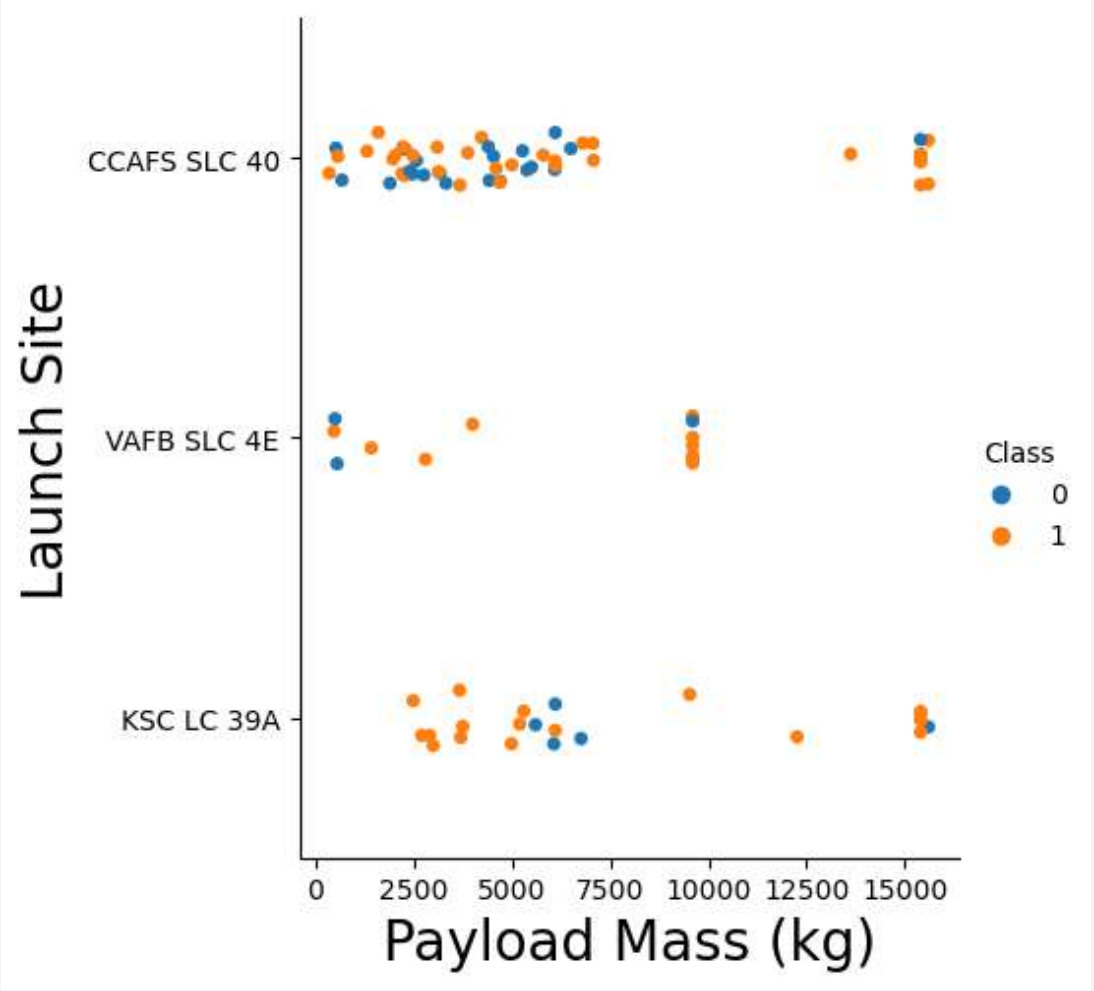
Flight Number vs. Launch Site

- Explanation: the first launches were performed in CCAFS SLC 40 and had poor successful rates. Afterwards, it was moved to other sites (VAFB and KSC) where the successful rate improved. However, the latest launches were performed again on CCAFS with the higher successful rate.
- Conclusion: Launch site seems to have almost no correlation with the outcome, even though the flight number does what could be caused by the improved technology and eliminated problems during the first launches



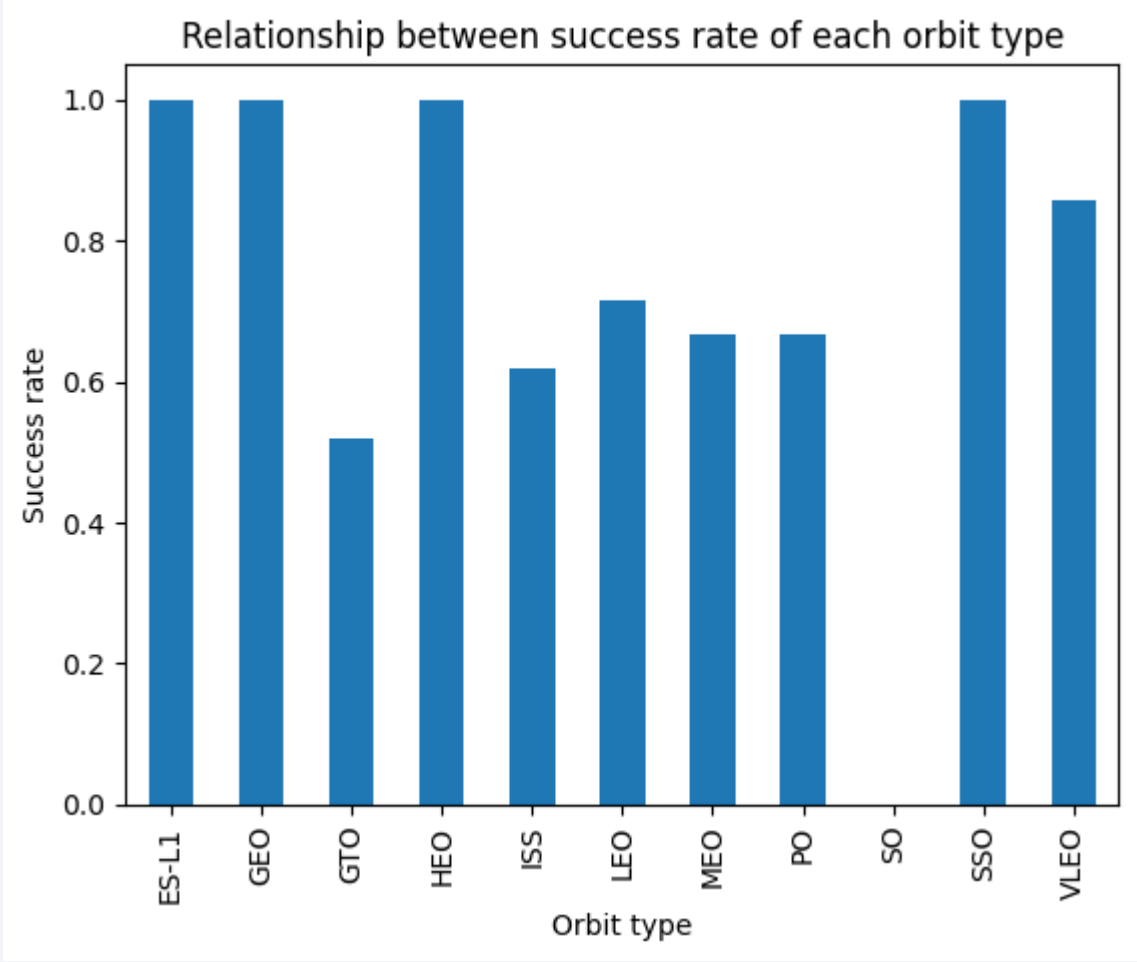
Payload vs. Launch Site

- There are no rockets launched for heavypayload mass(greater than 10000) from the VAFB-SLC launchsite



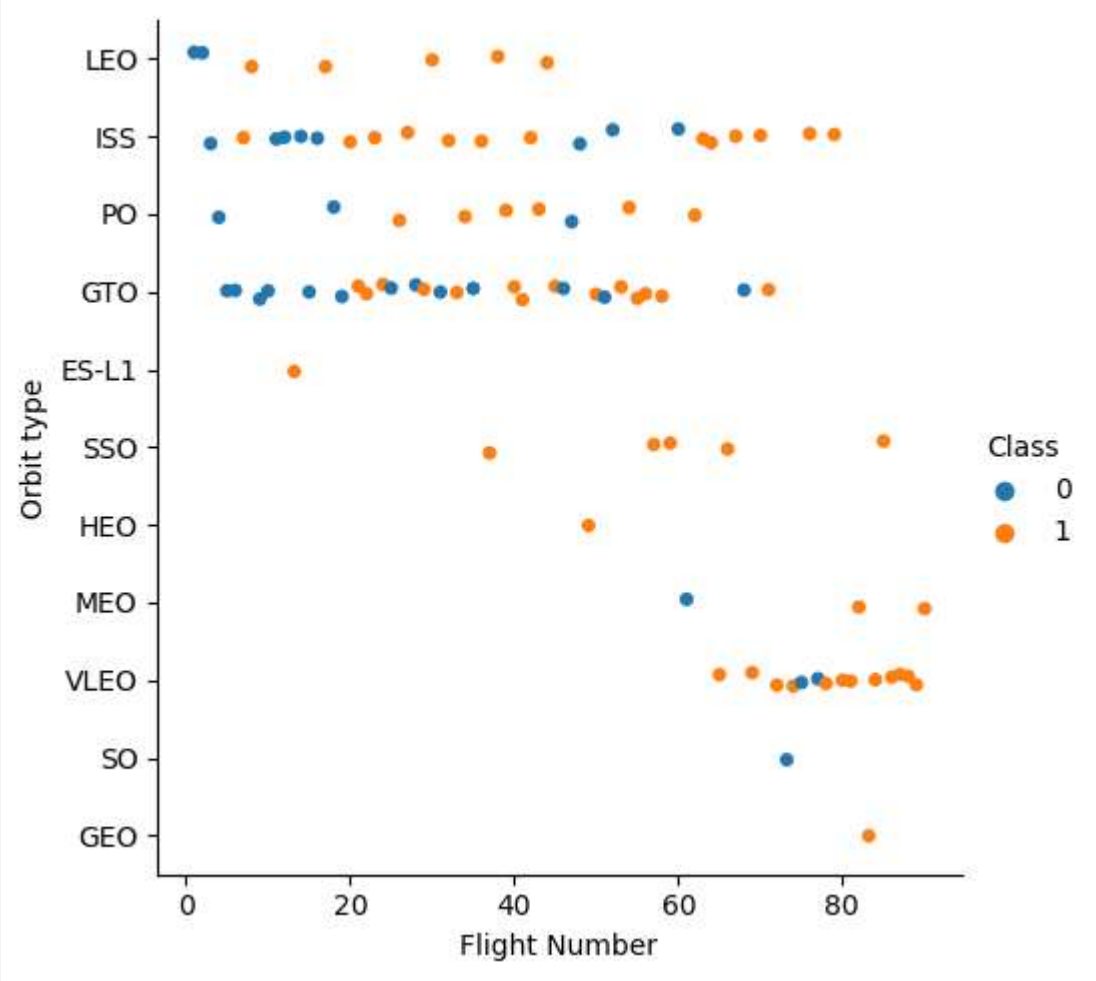
Success Rate vs. Orbit Type

- It seems that ES-L1, GEO, HEO and SSO have the max success rate. However, there was only 1 launch for each orbit. If we set a filter to see at least 2 launches for an orbit, VLEO has the max success rate (80%)



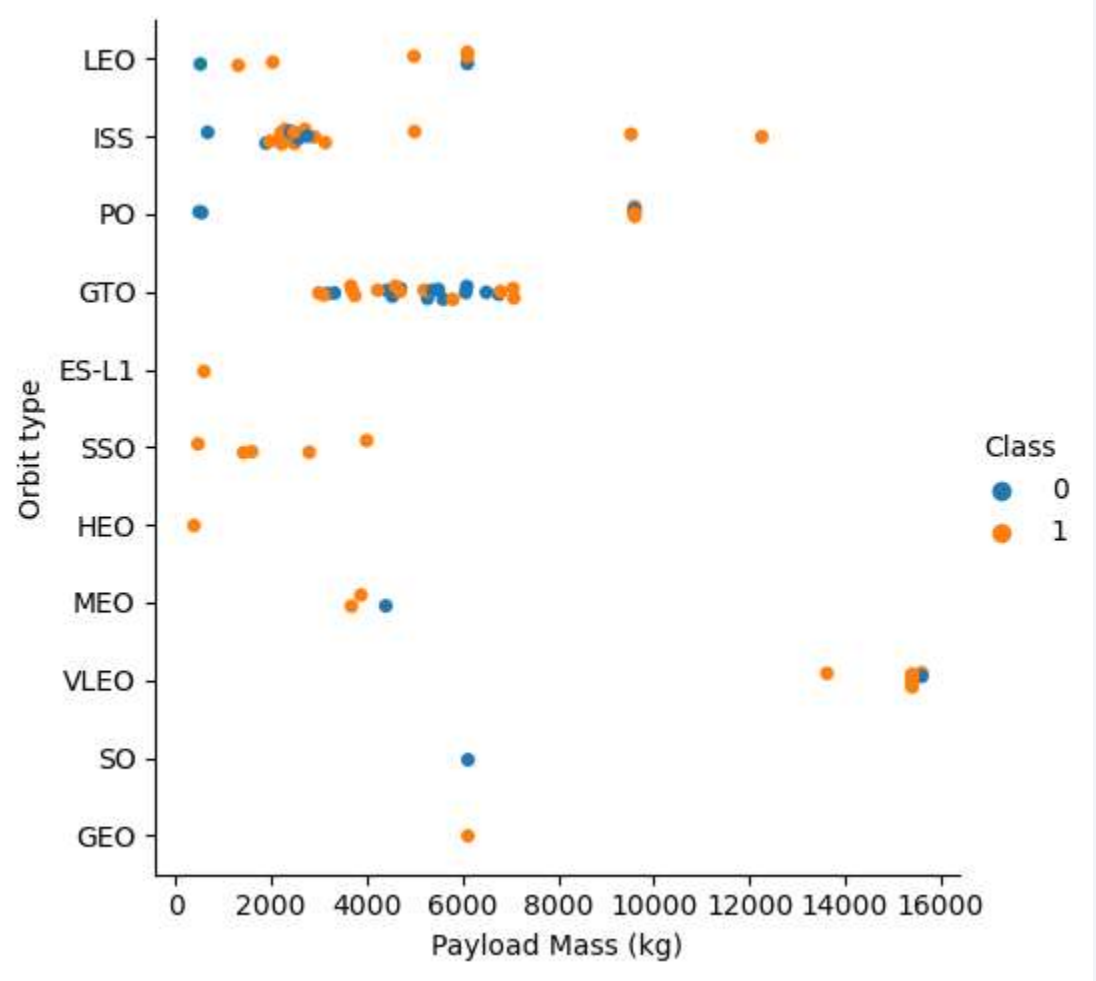
Flight Number vs. Orbit Type

- During the first launches different kinds of orbits were used. Nevertheless, VLEO is used more often in the latest flights.



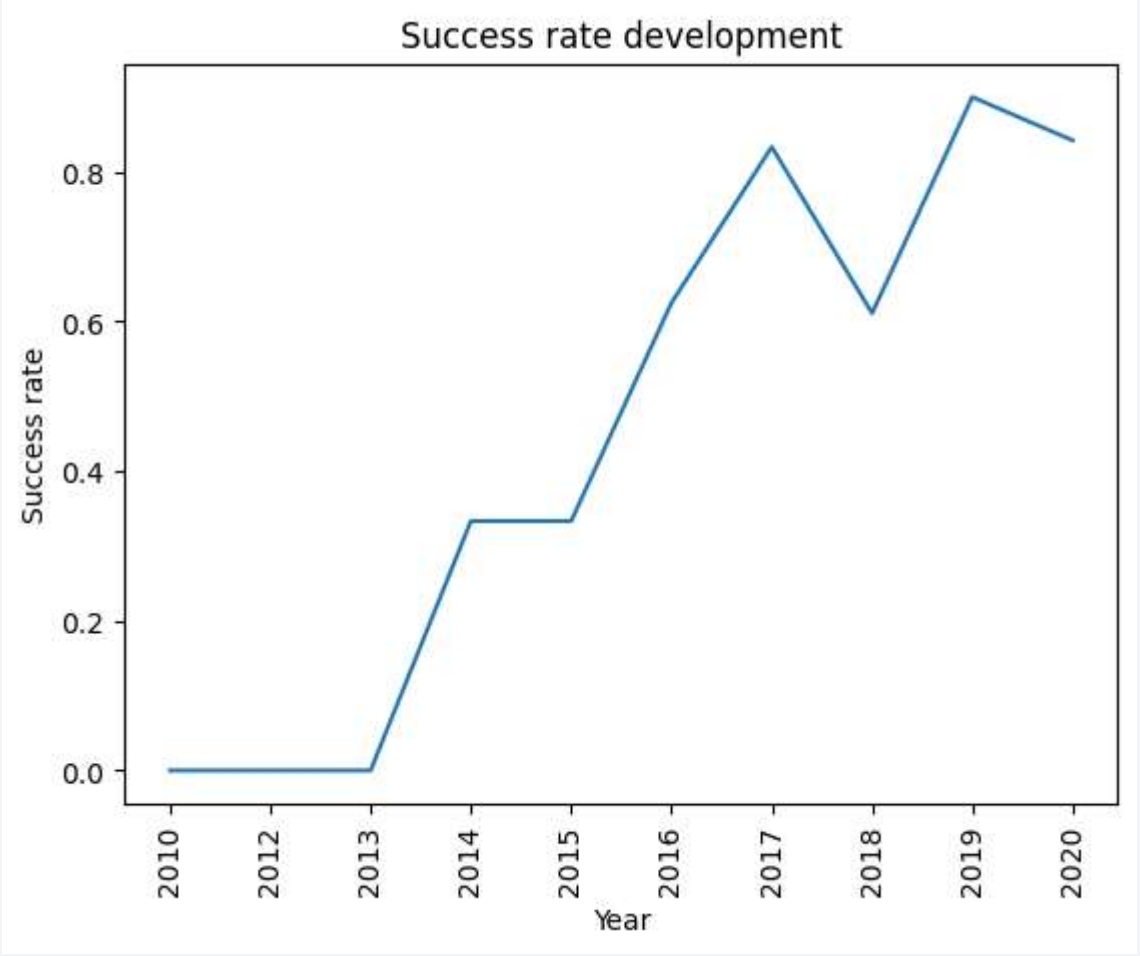
Payload vs. Orbit Type

- Different kinds of orbits are used for low payload mass, even though VLEO is used more for heavy launches



Launch Success Yearly Trend

- It is visible that between 2014 and 2017 the success rate doubled



All Launch Site Names

- There are 4 different launch sites ->

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	L
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

- Here are 5 records of launches with a launch site started with CCA

Total Payload Mass

- Total payload of boosters by NASA is 45 596 kg

Total_payload_NASA_CRS
45596.0

Average Payload Mass by F9 v1.1

- Average payload of a booster F9 v1.1 is 2 928,4 kg

Average_payload_F9_v1dot1
2928.4

First Successful Ground Landing Date

- The first successful ground landing data is 01.08.2018

MIN("Date")
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- Here are 3 boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- Here is a list of different mission outcomes
- Even though the values of failed missions are not that significant, here is a major representation of ‘No attempt’

Landing_Outcome	COUNT(*)
None	898
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- The list of boosters with the max payload mass
- The max payload mass is, therefore, 15 600 kg

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- There are only 2 launch records with these predefined parameters – one in April and the other one in October

MONTH	Landing_Outcome	Booster_Version	Launch_Site
10	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

- The greatest count has the value ‘No attempt’ – 10
- It is followed by Success (ground pad), Success (drone ship) and Failure (drone ship) with 5 counts each

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

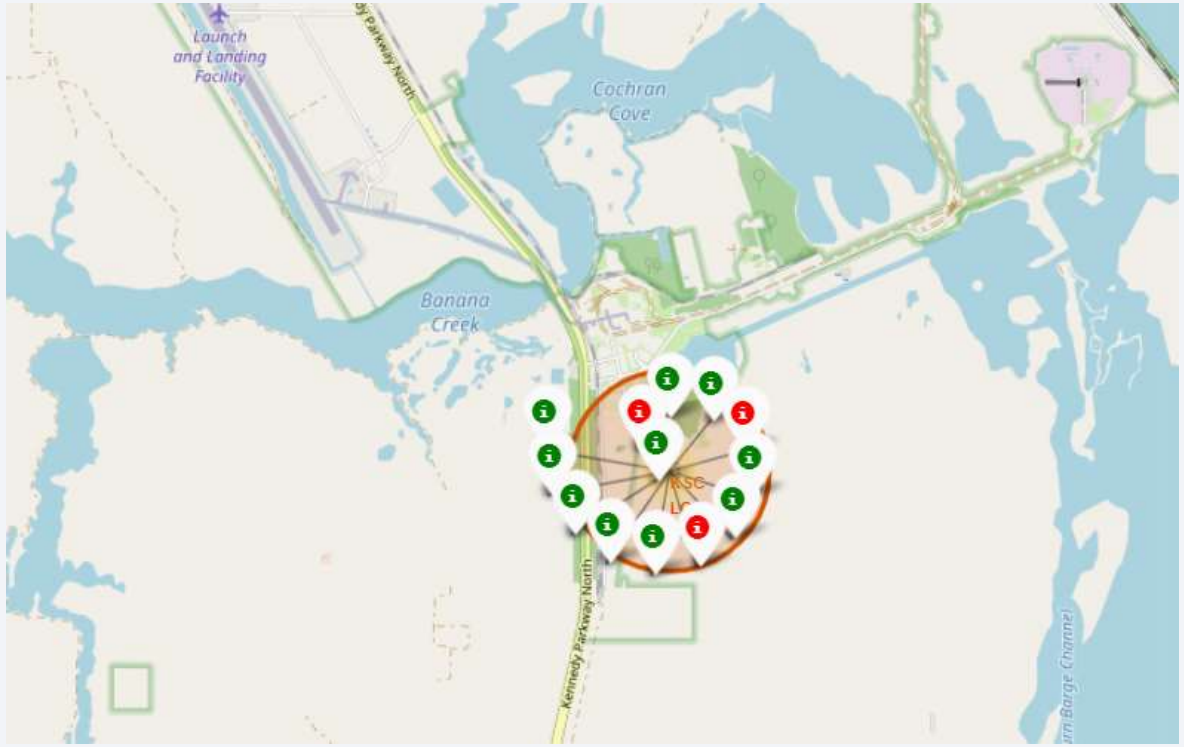
Section 3

Launch Sites Proximities Analysis

Global map: launch sites locations

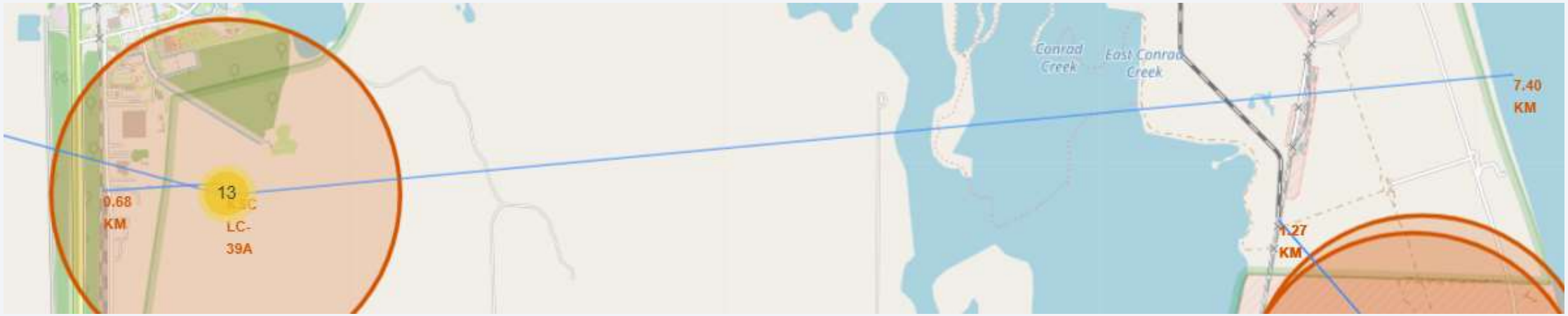
- 3 Launch sites are located on the East coast
- 1 Launch site is located on the West coast

Outcomes in KSC LC-39A



- Here is an example of outcomes in KSC LC-39A launch site
- It is visible that this launch site has a high success rate

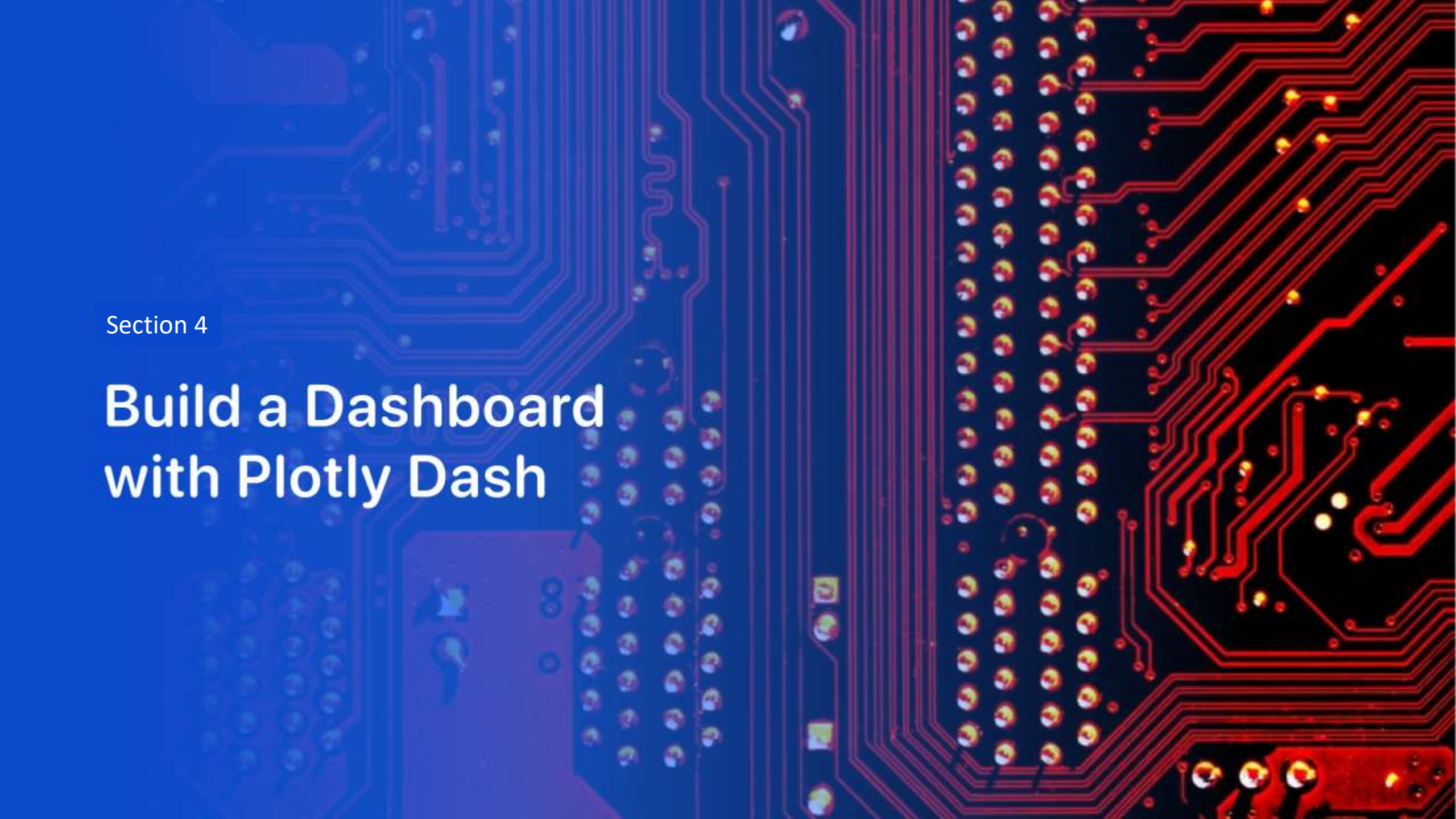
Distances between KSC LC-39A and roads, railway etc.



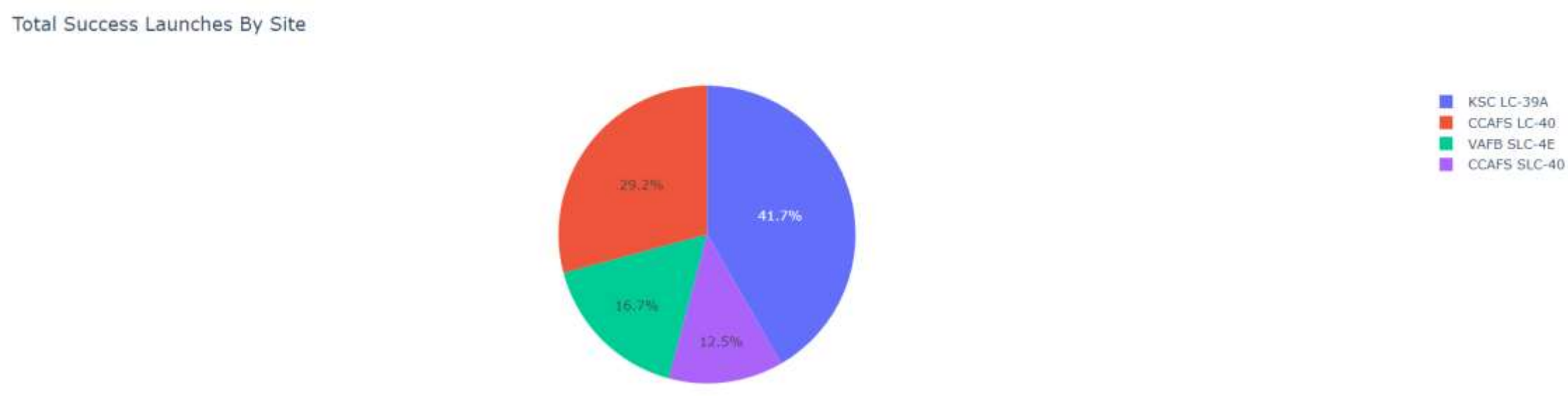
- Here is an example of KSC LC-39A
- It is approx. 0.7 km to the nearest road and railway
- It has 7.4 km to the coastline
- Even though it is not visible on the screenshot, it has approx. 16 km to the nearest city (Titusville)

Section 4

Build a Dashboard with Plotly Dash

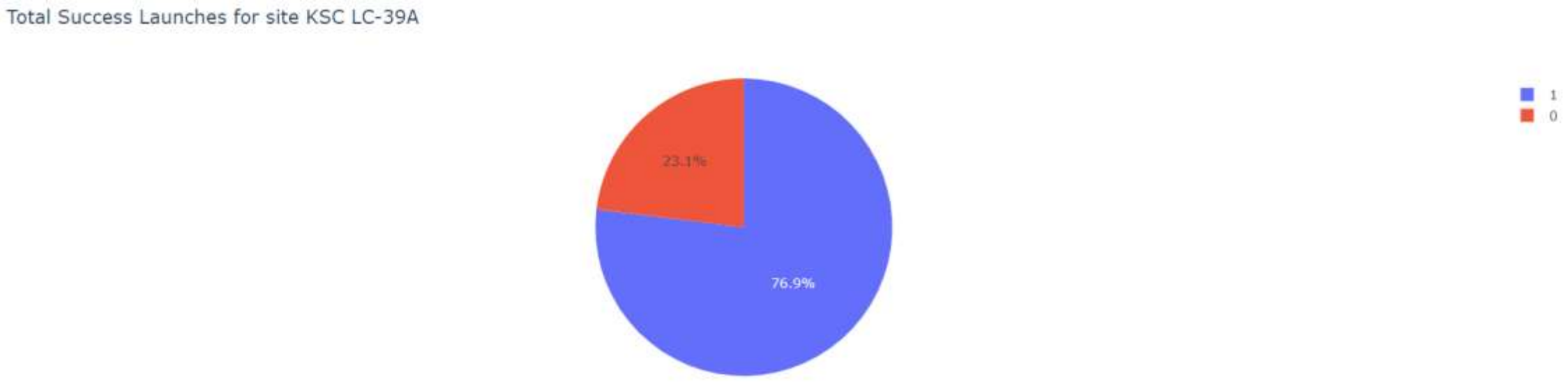


Total Success Launches By Site



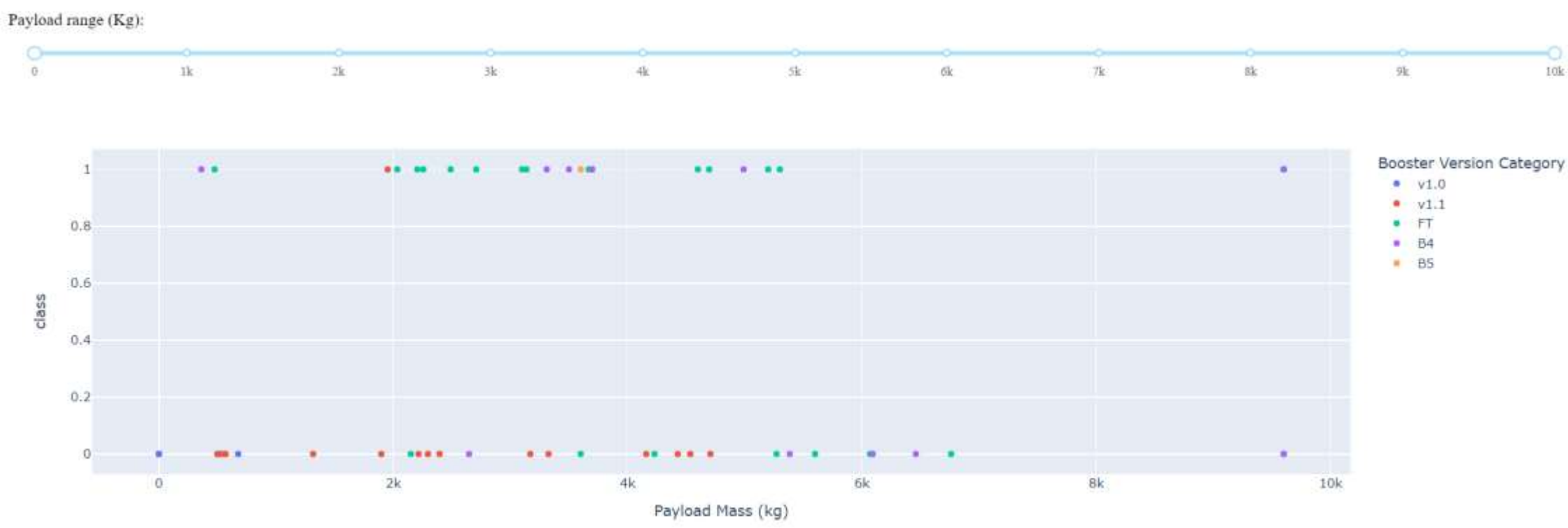
- The most successful launch site is KSC LC-39A which has almost 42% out of the total successful outcomes within all launch sites

Total Success Launches for site KSC LC-39A



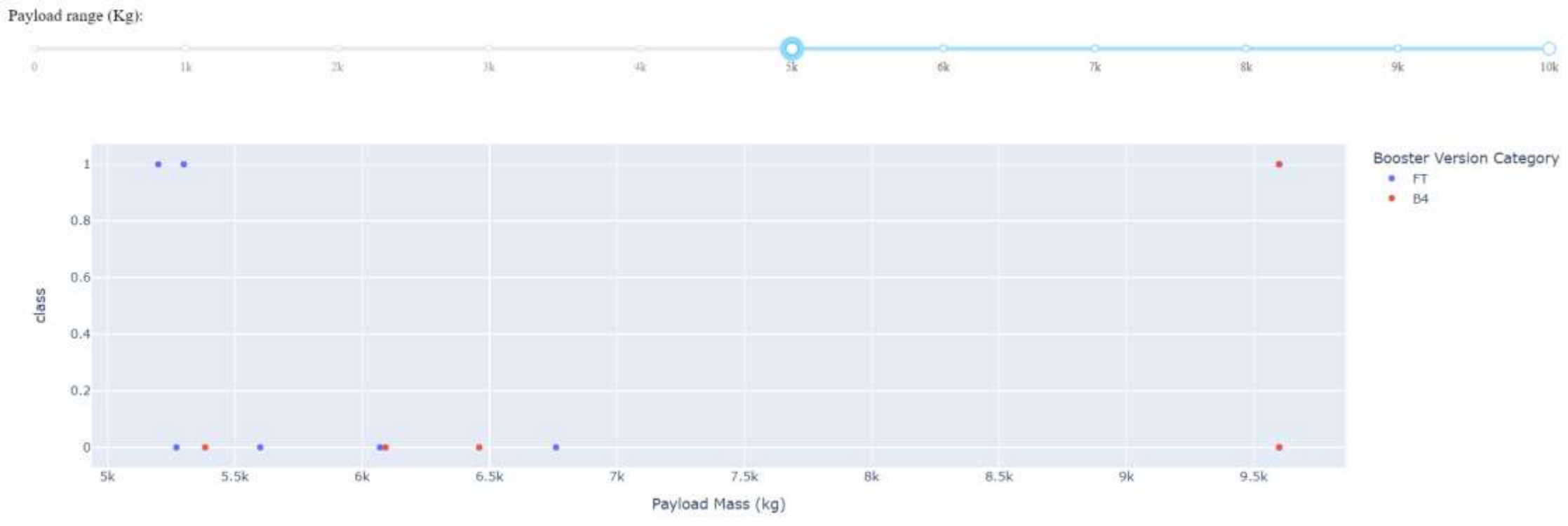
- The launch site mentioned on the previous slide is shown here
- It has the highest successful rates – nearly 77%, what means that approx. 3 out of 4 launches land successfully

Payload vs Outcome vs Booster category



- No filter on payload is set here
- It is visible that booster version FT has the highest successful rate, even though v1.0 and v1.1 have the lowest

Payload vs Outcome vs Booster category



- Here is a filter on payload >5k kg
- It is visible that the launches with a massive payload tend to fail as there were only 3 successful landings

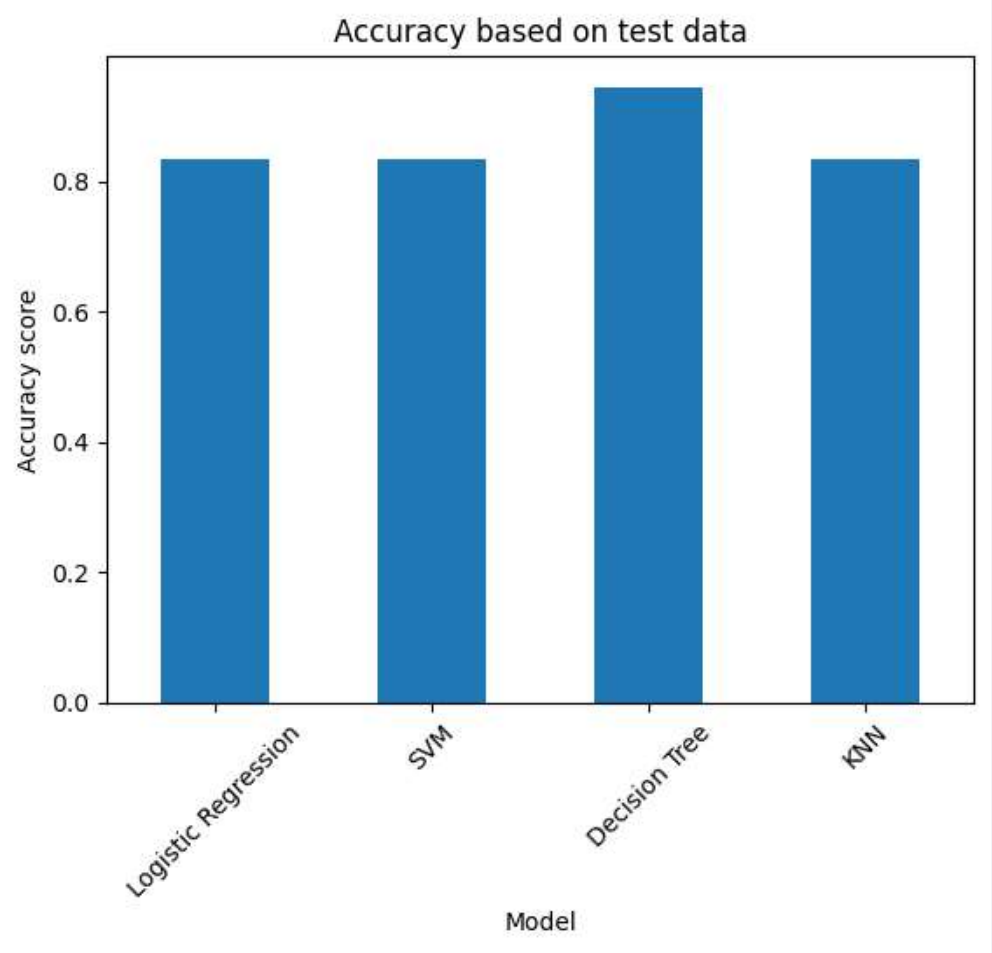
Section 5

Predictive Analysis (Classification)

Classification Accuracy

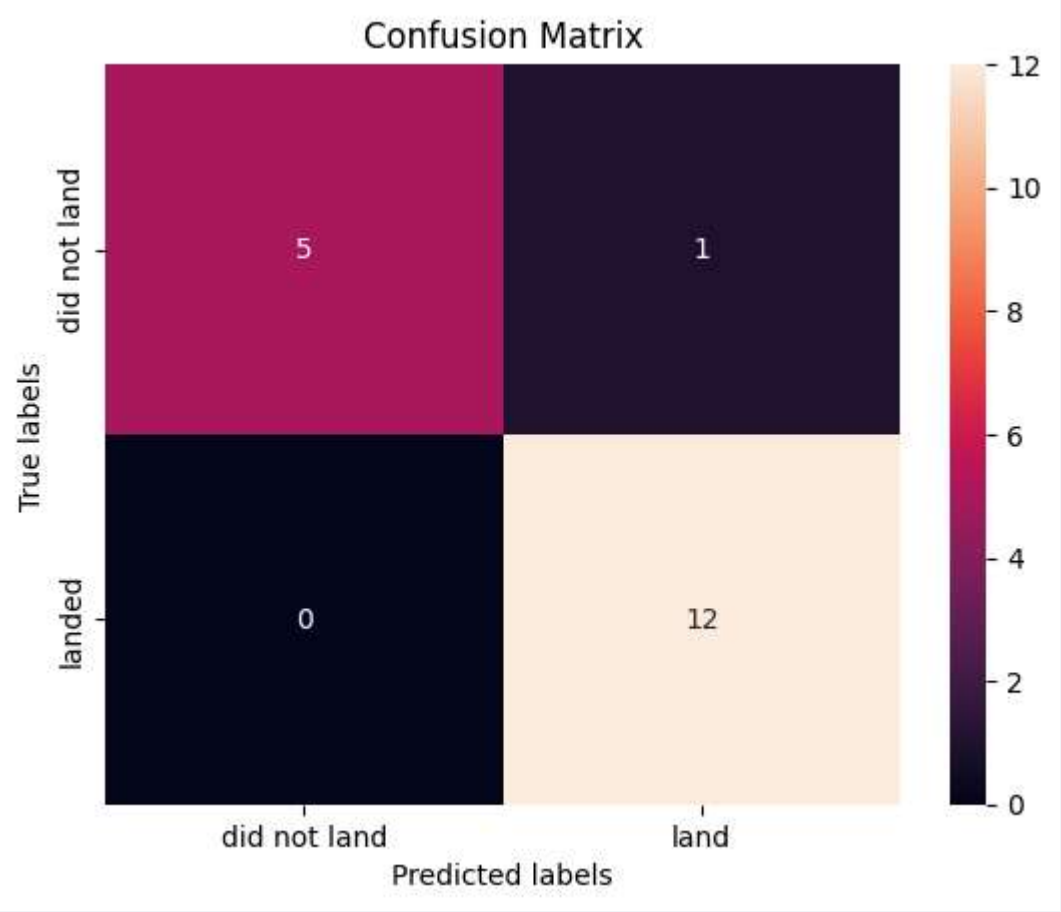
- The model with the highest accuracy score based on test data is Decision Tree (94,4%)
- It was achieved with the following parameters:

- 'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'



Confusion Matrix

- Here is the confusion matrix of Decision Tree
- It correctly predicted all the successful landings (0 error)
- However, there is one false prediction for the failed landings



Conclusions

- The best performed model is Decision Tree
- The rest of the models have the same accuracy score on the test data

Method	Accuracy/Train	Accuracy/Test
Logistic Regression	0.846429	0.833333
SVM	0.848214	0.833333
Decision Tree	0.891071	0.944444
KNN	0.848214	0.833333

Thank you!