https://github.com/Antonio-Ngbesu/Applied_Data_Science_Capstone

Winning Space Race with Data Science

Antonio Manzela Ngbesu
21/09/2023

Outline

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

Executive Summary

- Data was collected from public SpaceX API and SpaceX Wikipedia page.
- Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features.
- Changed all categorical variables to binary using one hot encoding.
- Standardized data and used GridSearchCV to find best parameters for machine learning models.
- Visualized accuracy score of all models.

Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.

All ML models produced similar results with accuracy rate of about 83%.
All models over predicted successful landings.

More data is needed for better model determination and accuracy.

Introduction

Background:

- Commercial Space Age is Here
- Space X (Falcon 9) has best pricing ($62 million vs. upwards $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Challenge:

- Space Y asked us to:
  - determine the price of each launch
  - gather public information about Space X and create dashboards for the team
  - train a machine learning model to predict successful Stage 1 recovery

Methodology

OVERVIEW OF DATA COLLECTION, WRANGLING, VISUALIZATION, DASHBOARD, AND MODEL METHODS

Methodology

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection Overview

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

The next slide will show the sequence of processing the data from SpaceX public API and the one after will show sequence of processing the data from webscraping.

Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Webscrape Data Columns:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection SpaceX API

1. Request (Space X APIs)

2. JSON file + Lists(Launch Site, Booster Version, Payload Data)

3. Json_normalize to DataFrame data from JSON

4. Dictionary relevant data

5. Cast dictionary to a DataFrame

6. Filter data to only include Falcon 9 launches

7. Replace missing PayloadMass values with mean

https://github.com/Antonio-Ngbesu/Applied_Data_Science_Capstone/blob/main/Spacex%20Webscraping.ipynb

Data Wrangling

Create a training label with landing outcomes where successful=1 & failure = 0.
Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’
New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.

Value Mapping:

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, False Ocean, False RTLS – set to -> 0.

https://github.com/Antonio-Ngbesu/Applied_Data_Science_Capstone/blob/main/Data%20wrangling.ipynb

EDA with Data Visualization

Exploratory Data Analysis performed on variables flight number, Payload Mass, Launch Site, Orbit, Class and Year

Plot Used:

Flight number vs Payload Mass, Flight number vs Launch site, Payload Mass vs Launch site, Orbit vs Success Rate, Flight number vs Orbit, and Success Yearly Trend.

Scatter plots, line charts, and bar plots were used to compare relationship between variables to decide if a relationship exists so that they could be used in training the machine learning model.

https://github.com/Antonio-Ngbesu/Applied_Data_Science_Capstone/blob/main/EDA%20with%20python.ipynb

EDA with SQL

Loaded dataset into IBM DB2 Database.

Queried using SQL Python integration.

Queries were made to get a better understanding of the dataset.

Queried information about launch site names, mission outcomes,
various pay load sizes of customers and booster versions, and landing outcomes.

https://github.com/Antonio-Ngbesu/Applied_Data_Science_Capstone/blob/main/EDA%20with%20sql.ipynb

Build an Interactive Map with Folium

Launch Sites Locations Analysis with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Cost, and City.

This allows us t understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

https://github.com/Antonio-Ngbesu/Applied_Data_Science_Capstone/blob/main/Launch%20site%20location%20with%20folium.ipynb

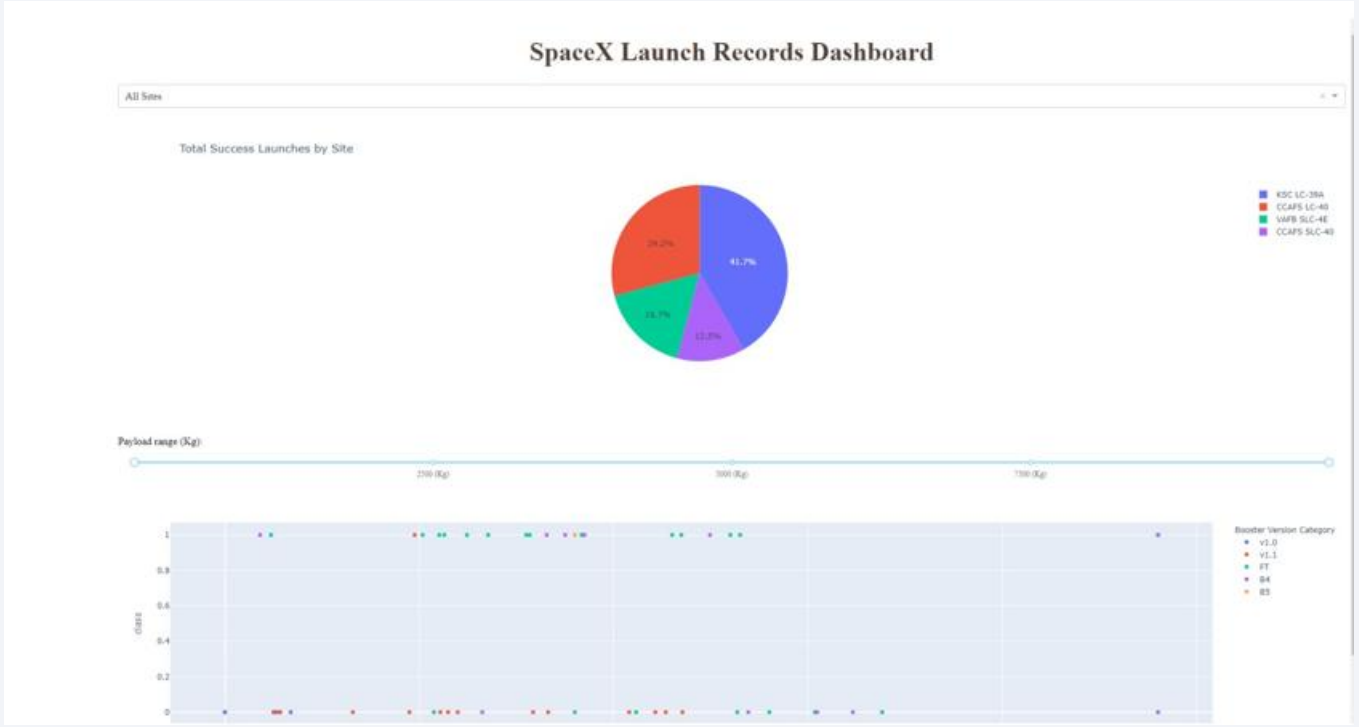
Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

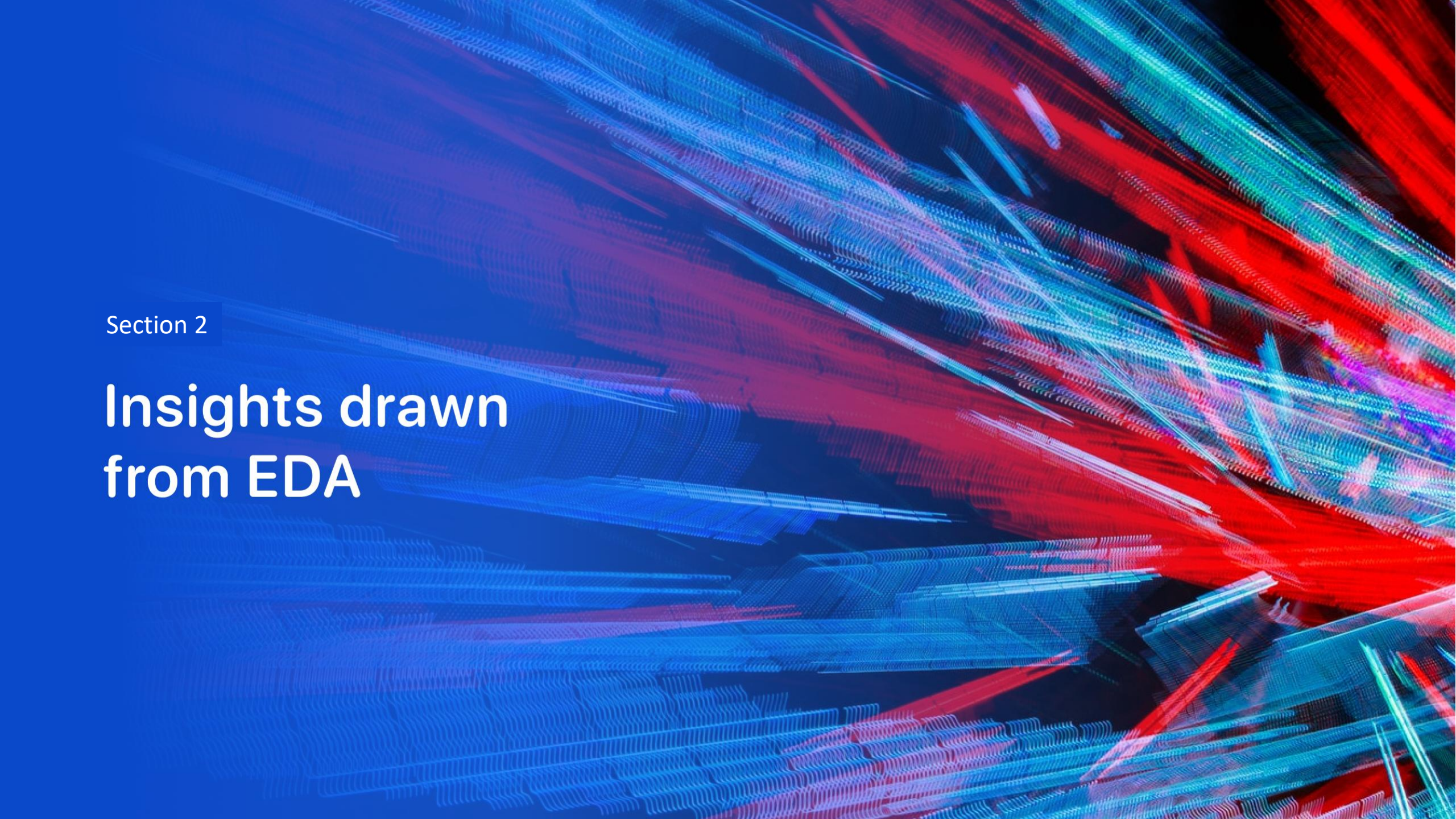
Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results

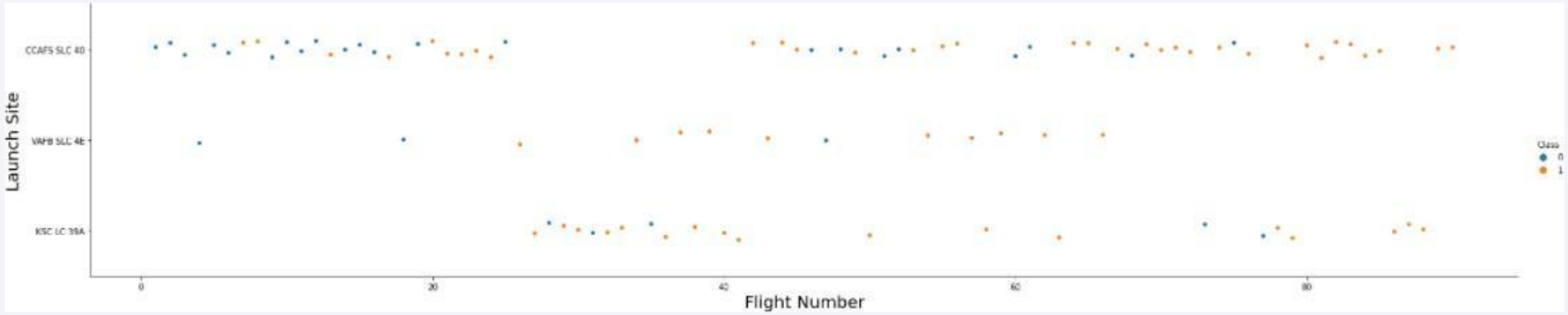


This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Number vs. Launch Site



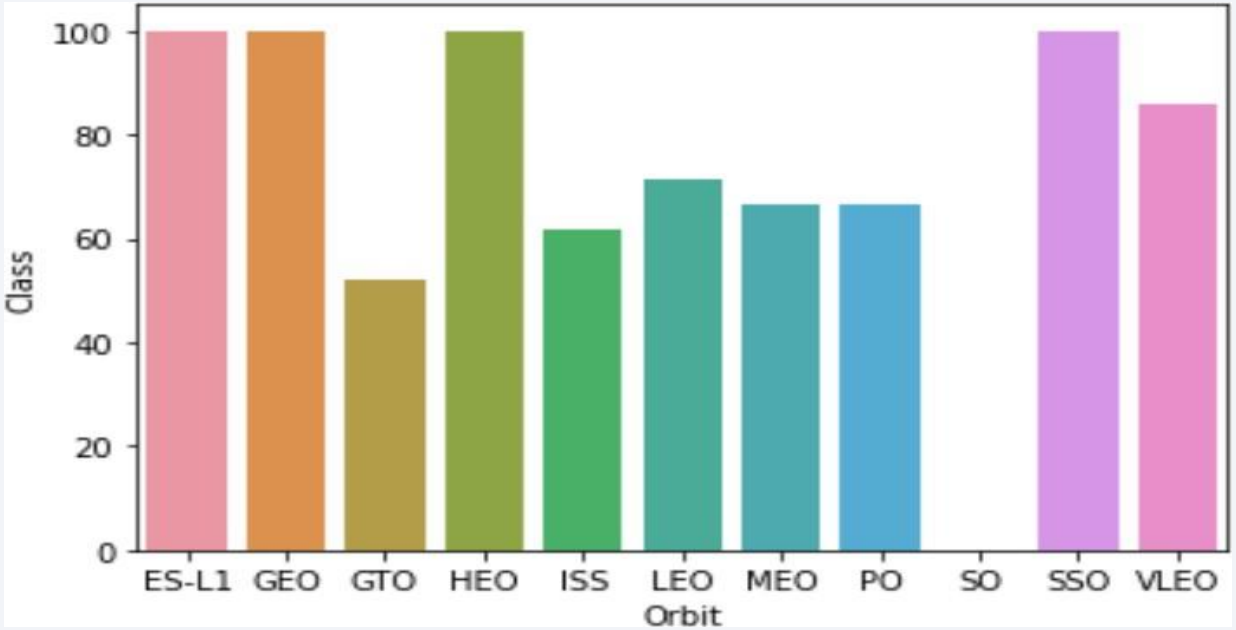
Orange indicates: Successful Launch
Blue indicates: Unsuccessful Launch

Payload vs. Launch Site

Payload vs. Launch Site

Orange indicates: Successful Launch
Blue indicates: Unsuccessful Launch

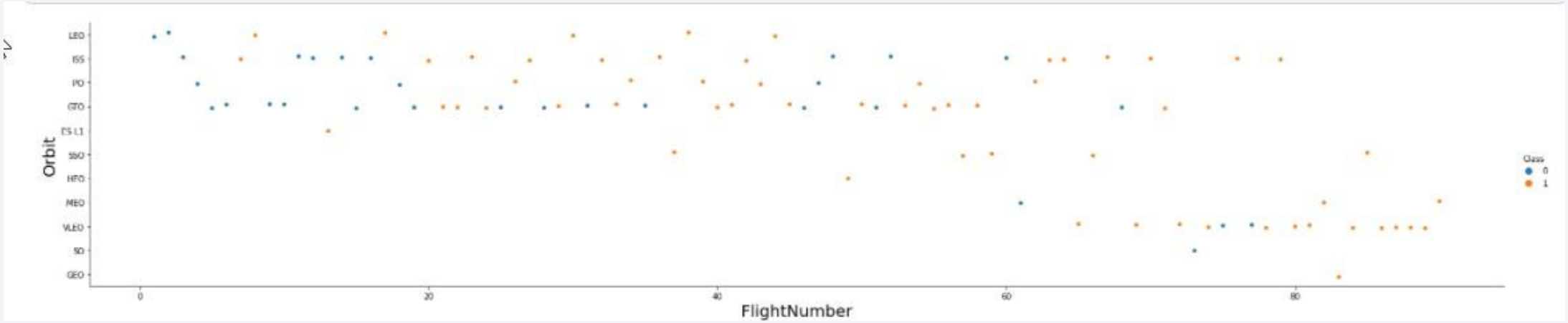
Success Rate vs. Orbit Type



Success Rate Scale with %

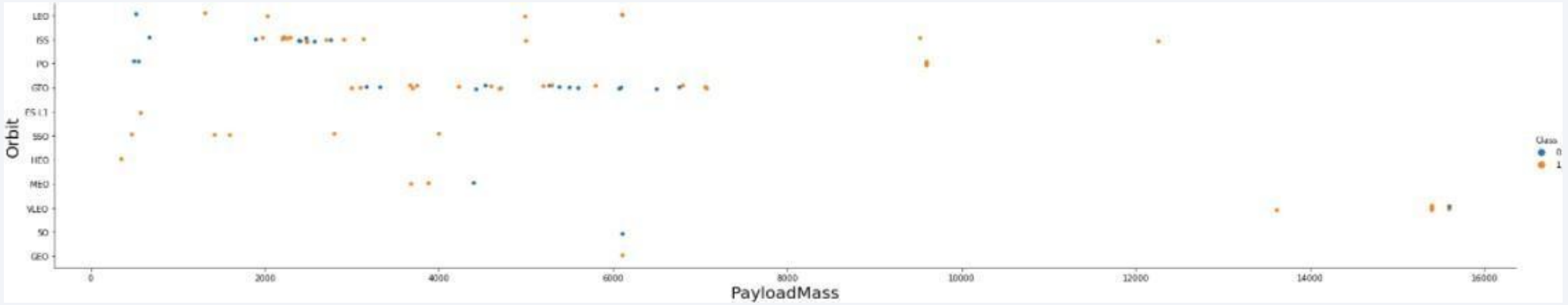
Flight Number vs. Orbit Type

Flight number vs. Orbit type



Orange indicates: Successful Launch
Purple indicates: Unsuccessful Launch

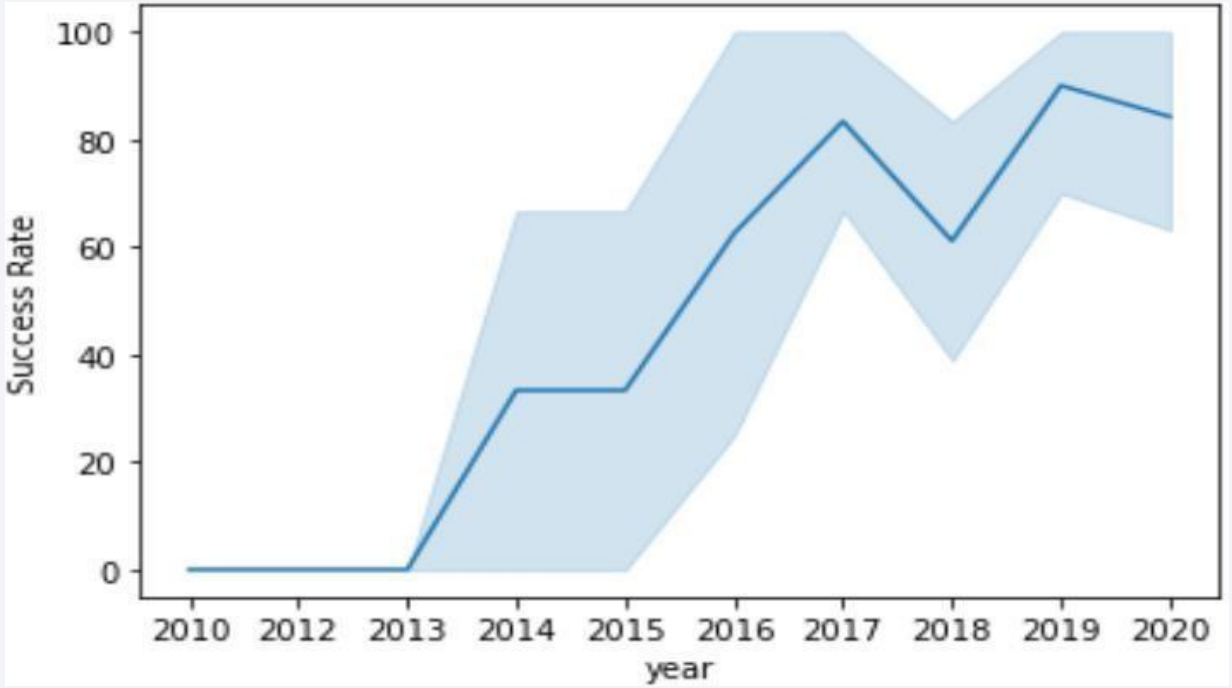
Payload vs. Orbit Type



Orange indicates: Successful Launch

Purple indicates: Unsuccessful Launch

Launch Success Yearly Trend



95% confidence interval
(light blue shading)

All Launch Site Names

Task 1

Display the names of the unique launch sites in the spac

```
n [10]: %sql select DISTINCT LAUNCH_SITE from SPACEXTBL
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8
d:32733/BLUDB
Done.
```

ut[10]:

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

- Query unique launch site names from database.
- CCAFS SLC -40 and CCAFSSLC -40 likely all represent the same
- launch site with data entry errors.
- CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC-40,

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
In [16]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.clou
d:32733/BLUDB
Done.
```

Out[16]:

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

First 5 entries in database with Launch Site name beginning with CCA.

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL
where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

SUM
22007

This query sums the total payload mass in Kg where NASA was the customer.

CRS stands for Commercial Resupply Service which indicates that these payloads were sent to the international Space station (ISS).

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

2015 Launch Records

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

- Present your query result with a short explanation here

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

- Present your query result with a short explanation here

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

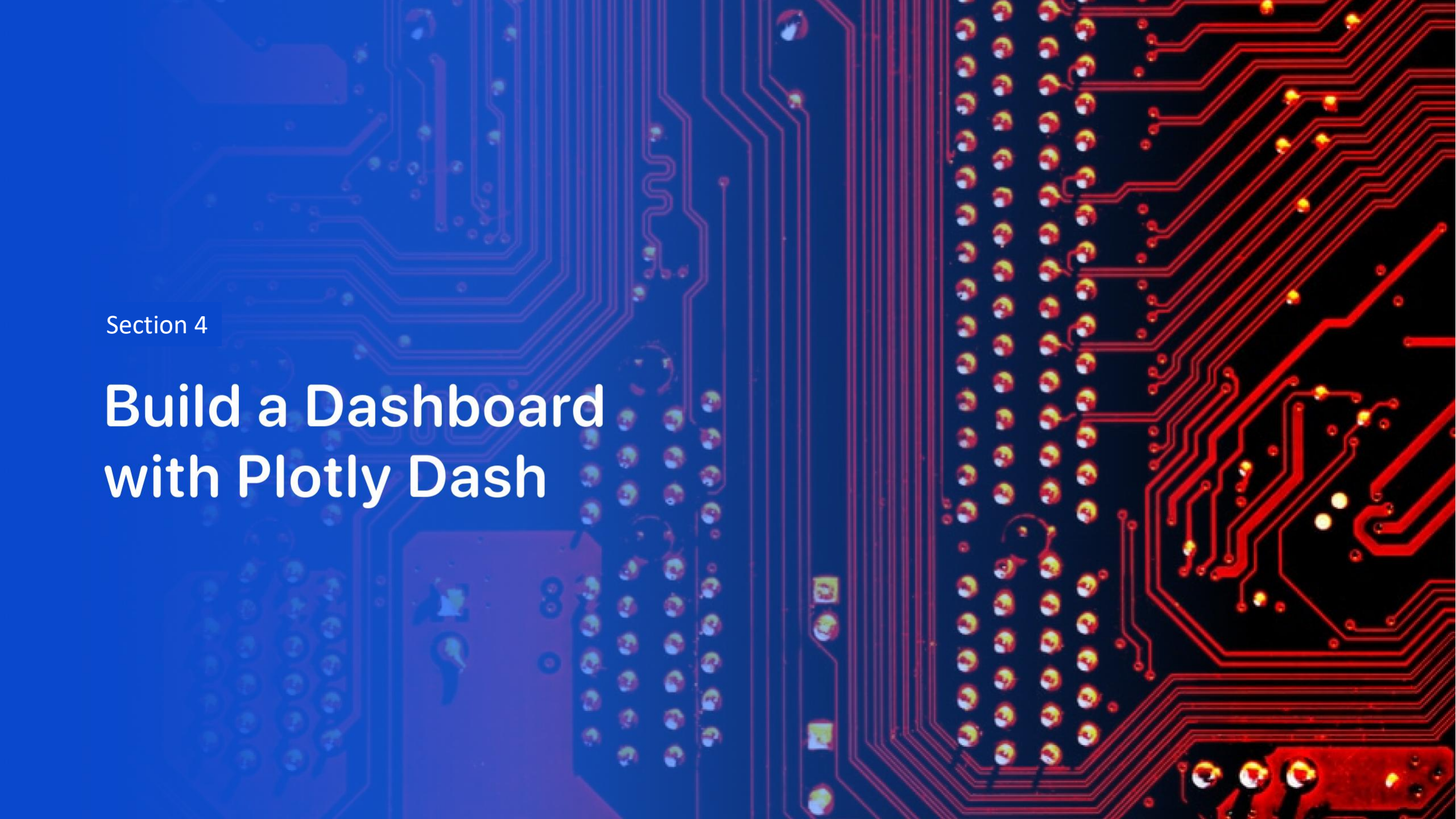
- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 2>

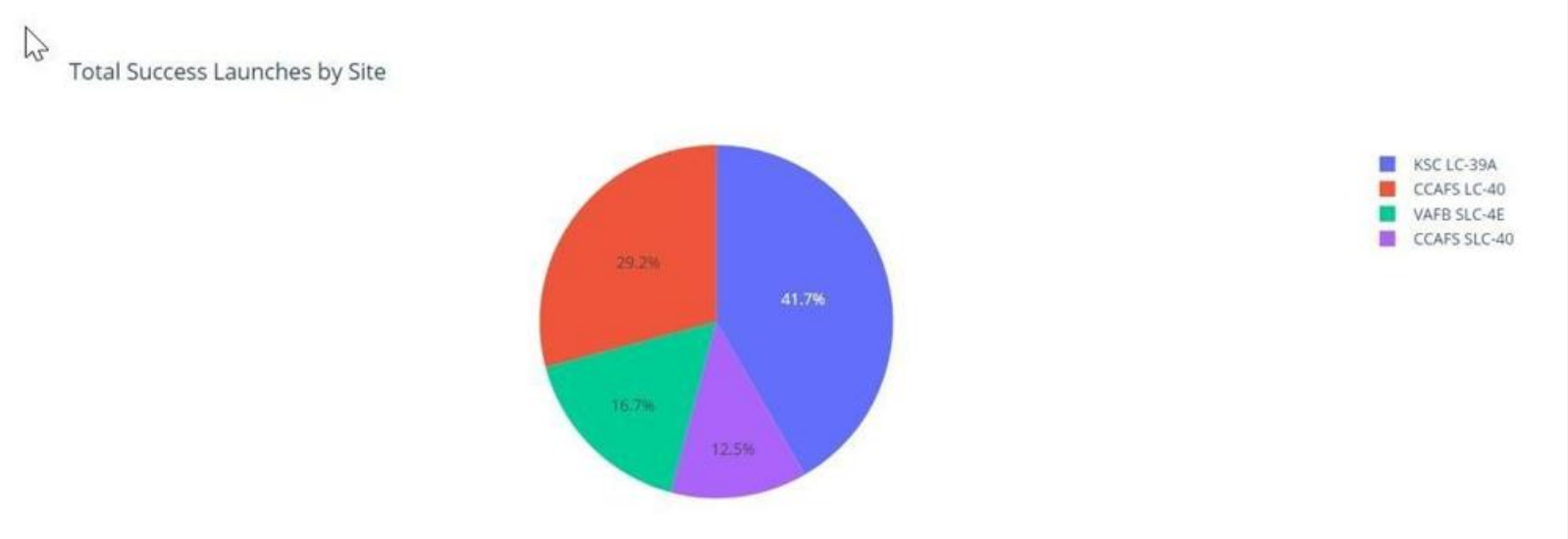
- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

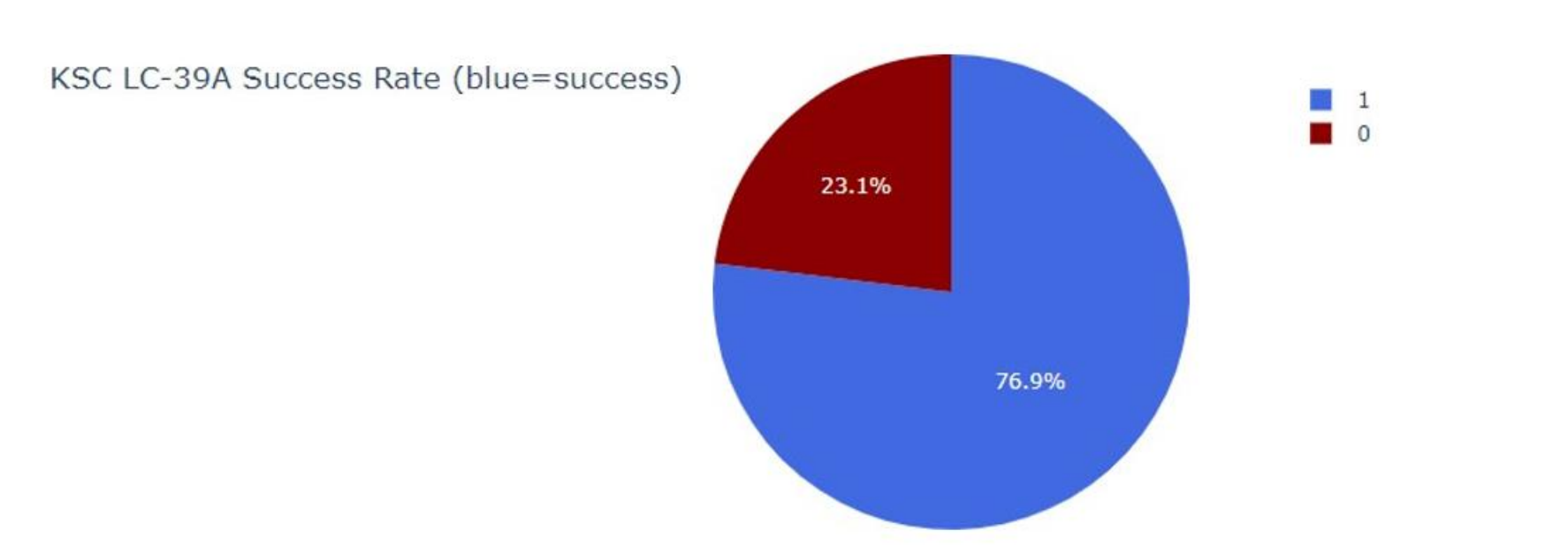
- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot


Section 4
Build a Dashboard with Plotly Dash

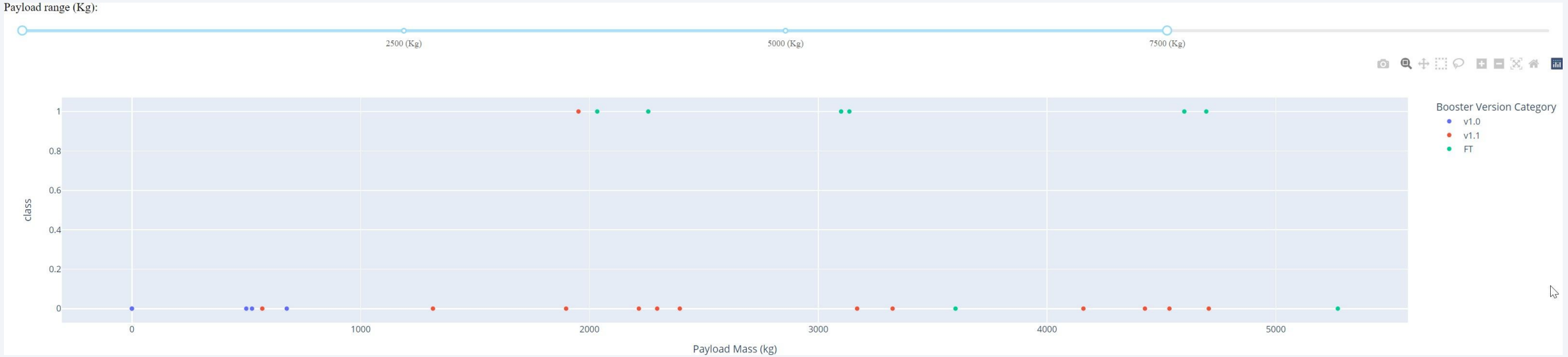
<Dashboard Screenshot 1>



<Dashboard Screenshot 2>



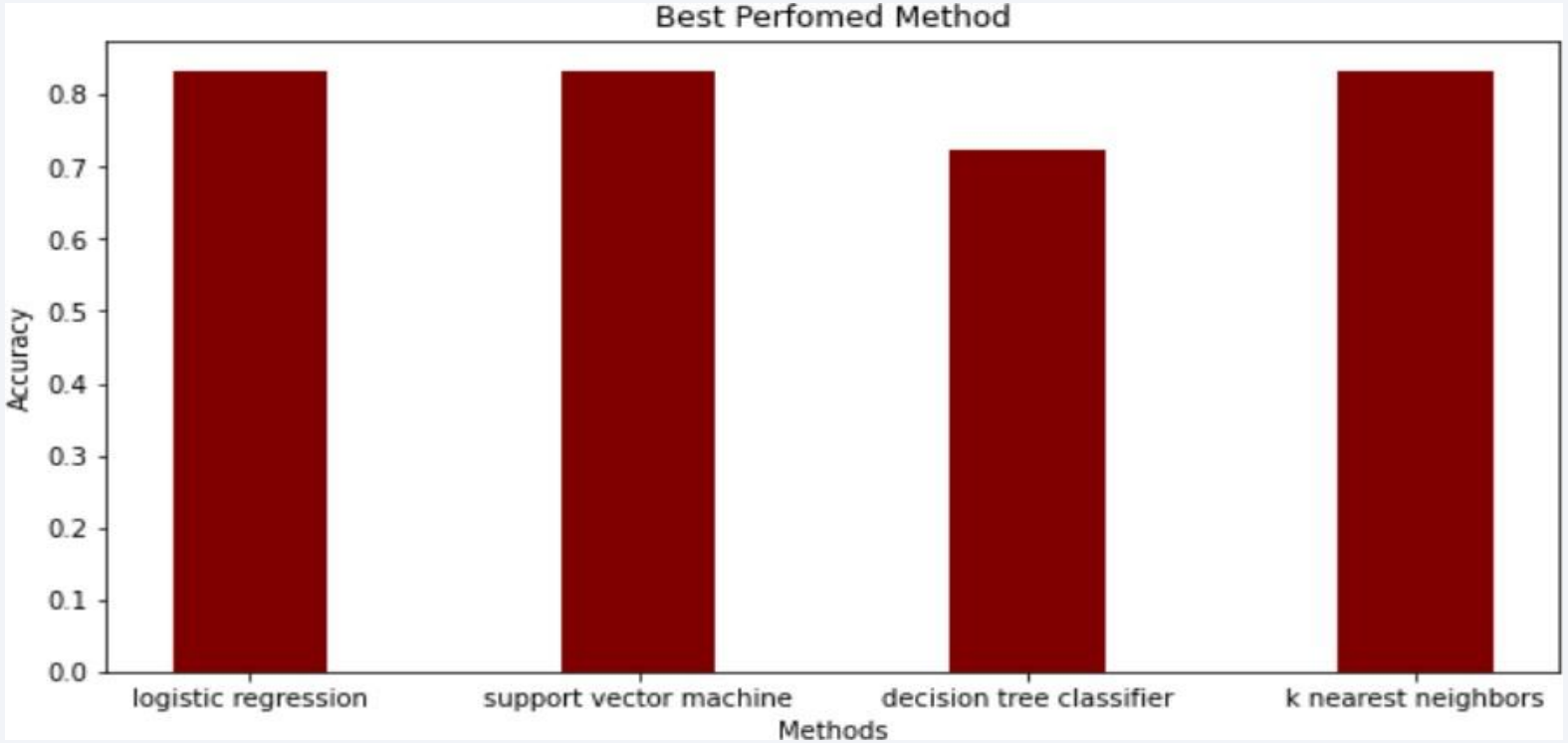
<Dashboard Screenshot 3>



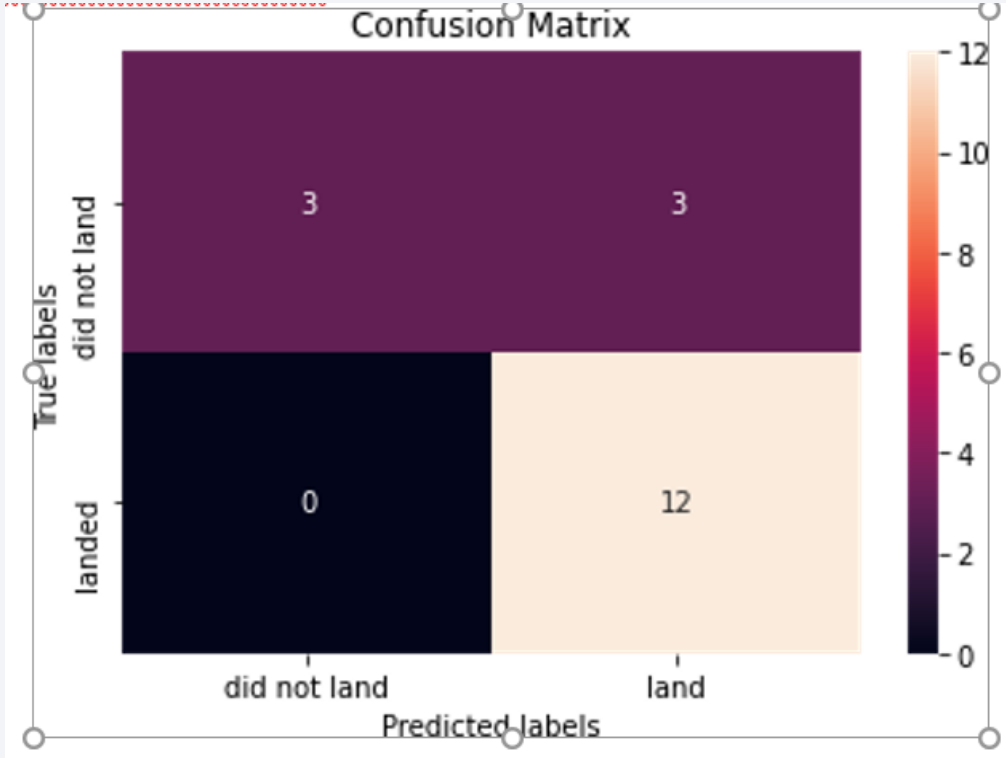
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Correct predictions are on a diagonal from top left to bottomright.

Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- More data should be collected to better determine the best machine learning model and improve the accuracy

Appendix

GitHub Repo

https://github.com/Antonio-Ngbesu/Applied_Data_Science_Capstone

- Space X Data

- Wikipedia

Thanks to everyone

Thank you!