IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Enrico Findley
16/07/2023

Outline

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

Executive Summary

- The methodologies that used to analyze data are:
    1. Data Collection: collecting data from web scraping and SpaceX API.
    2. Exploratory Data Analysis (EDA): doing data wrangling, data visualization, then interactive visual analysis.
    3. Prediction: doing machine learning prediction
- The results are:
    - There are many valuable data although it is from public data source.
    - EDA process are able to analyze the best feature that can predict launching success.
    - Machine learning prediction is able to give summary about best performing model using all collected data.

Introduction

- Project Space Y are project to make competitor company from Space X
- Main Problem:
  - How to predict if the first stage of landing rocket will land?
  - Where is the best location to launch the rocket?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - API (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - First, the data is loaded then removing the null and missing value.
  - Data exploration is executed to understand the data.
  - Finally, creating the 'Class' column as outcome of each landing whether its success or not.

Methodology

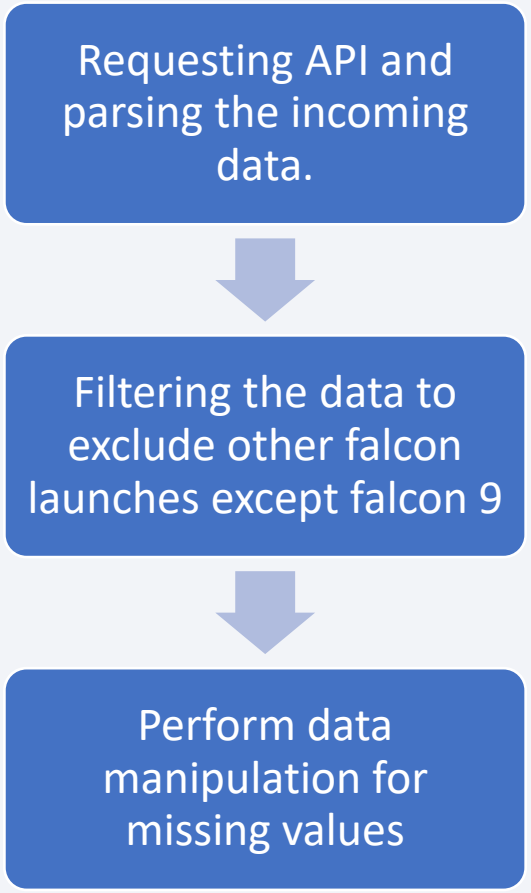
- Perform Exploratory Data Analysis (EDA) using visualization and SQL
  - Visualizing the relationship a number of features.
  - Feature Engineering for converting text feature into number.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Developing machine learning models to predict landing outcome
  - First, data is normalized then divided to train and test datasets, after that the model is developed to trained on training data, and finally the models evaluated by its accuracy which are the best performing models.

Data Collection

- Data sets were collected from:
  - Space X API (https://api.spacexdata.com/v4/rockets/).
  - Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches), using web scraping techniques.

Data Collection – SpaceX API

- Obtained from SpaceX public API, process of collecting data can be seen on flowchart.



- Source code: https://github.com/enricofindley/ibm-ds-certification/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Scraping

- Obtained from wikipedia, process of collecting data can be seen on flowchart.

- Source code: https://github.com/enricofindley/ibm-ds-certification/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/jupyter-labs-webscraping.ipynb

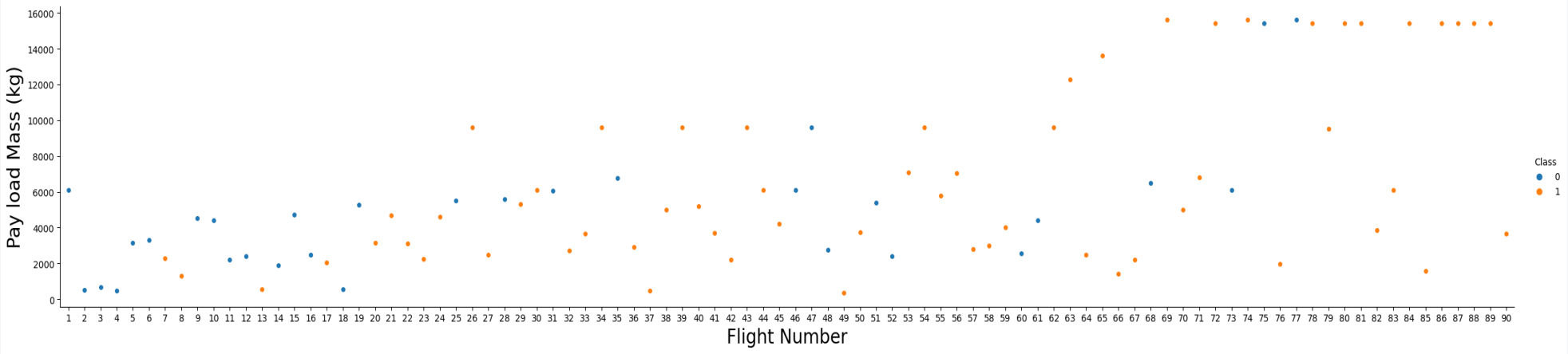
Data Wrangling

- First, the data is loaded then removing the null and missing value.
- Data exploration is executed such as calculate number of launches on each site, occurrence of each orbit, and occurrence of mission outcome in orbit.
- Finally, creating the 'Class' column as outcome of each landing whether its success or not, 1 is representing success and 0 representing failure.

- Source code: https://github.com/enricofindley/ibm-ds-certification/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- 3 type of charts were used for visualizing data relationship between features such as bar plot, scatter plot, and line plot.
- Example:



- Source code: https://github.com/enricofindley/ibm-ds-certification/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission.
  - Top 5 launch sites whose name begin with the string 'CCA’.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by booster version F9 v1.1.
  - Date when the first successful landing outcome in ground pad was achieved.
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg.
  - Total number of successful and failure mission outcomes.
  - Names of the booster versions which have carried the maximum payload mass using subquery.
  - List Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
  - Rank of 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.
- Source code: https://github.com/enricofindley/ibm-ds-certification/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Create Folium maps with following components:
  - Markers as representation for points like launch sites.
  - Circles as representation for highlighted areas around specific coordinates, like NASA Johnson Space Center.
  - Marker clusters as representation for groups of events in each coordinate, like launches in a launch site.
  - Lines as representation for indicate distances between two coordinates.

- Source code: https://github.com/enricofindley/ibm-ds-certification/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Visualizing data using graph and plot for following analysis:
  - Percentage of launches by site.
  - Payload range.
  - This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- Source code: https://github.com/enricofindley/ibm-ds-certification/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/spacex_dash.py

Predictive Analysis (Classification)

- Four classification models were compared its accuracy to find the best performing models: logistic regression, support vector machine, decision tree and k nearest neighbors.

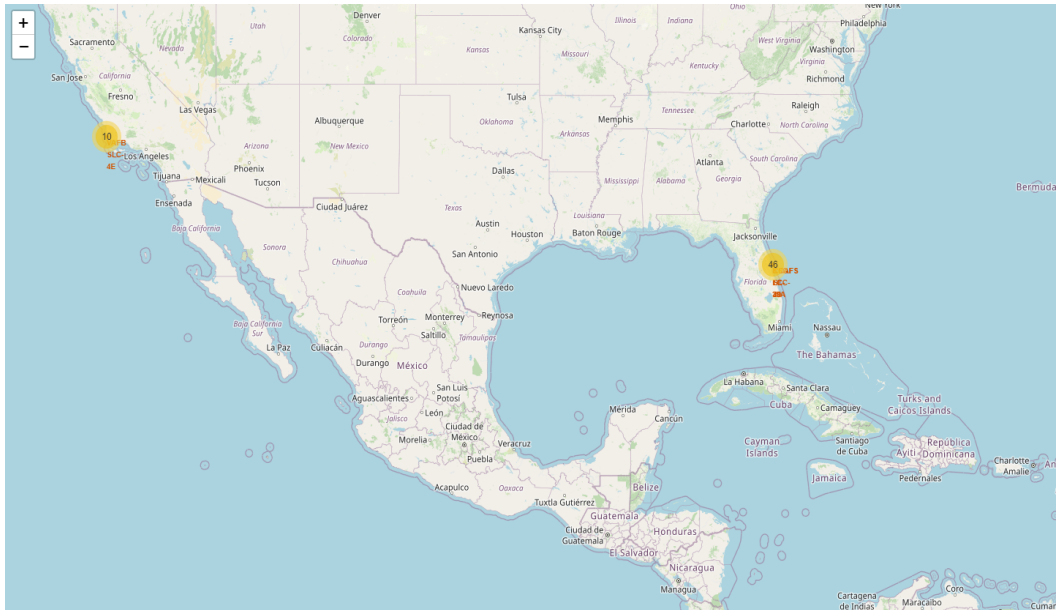
- Source code: https://github.com/enricofindley/ibm-ds-certification/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results:
  - SpaceX uses 4 different launch sites.
  - The first launches were done by SpaceX and NASA.
  - The average payload of F9 v1.1 booster is 2,928 kg.
  - The first successful landing outcome happened in 2015 five year after the first launch.
  - Many Falcon9 booster were successful at landing in drone ships with payload above the average.
  - Almost all of mission outcomes were successful;
  - 2 booster versions failed at landing in drone ships in 2015 such as F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

Results

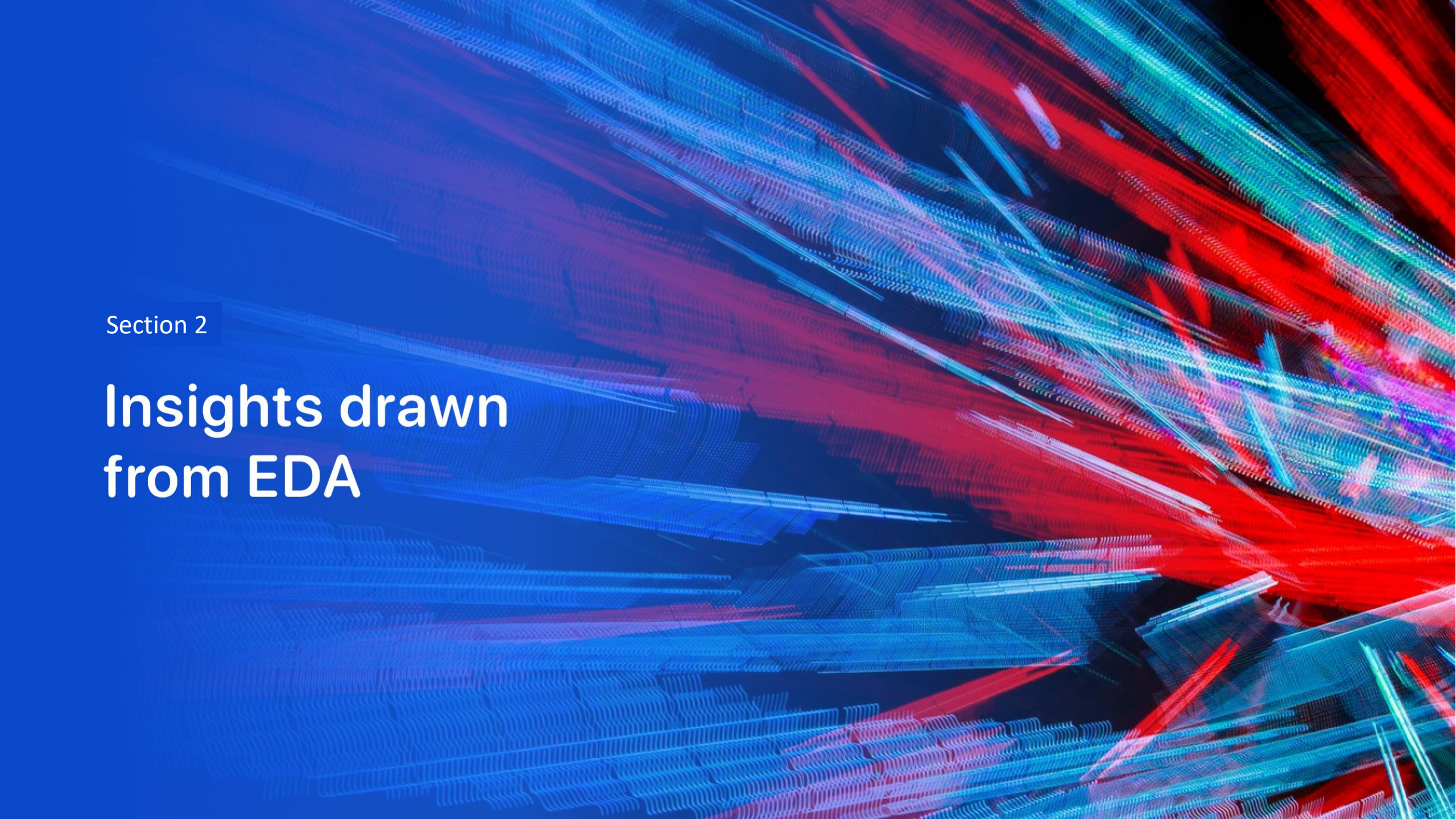
- From interactive analytics launch sites are often to be in safety places in this case near sea, and also have a good logistic infrastructure around such as road and railway.
- Most launches happens at east coast launch sites and the rest launches in west coast.



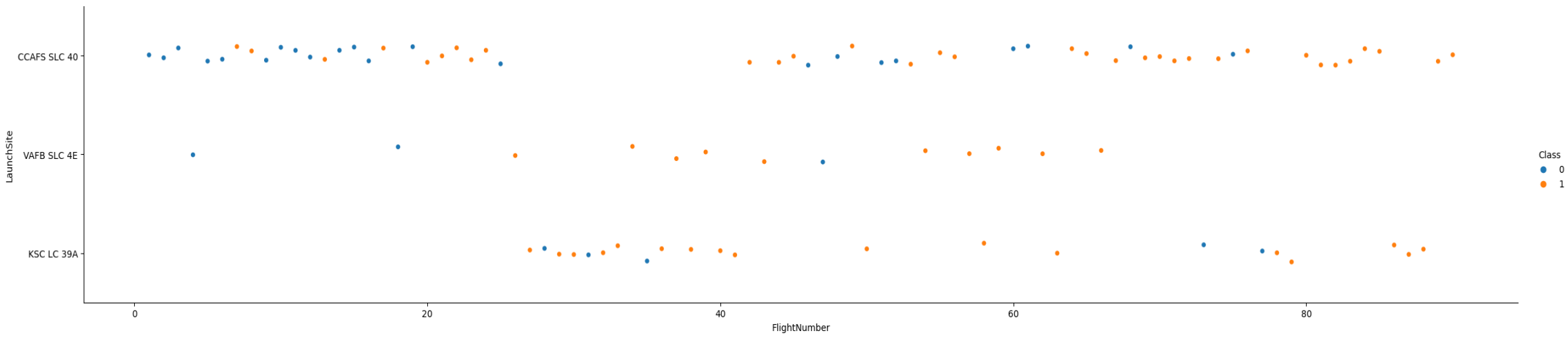
Results

- Predictive models that developed shows that best performing models is decision tree classifier algorithms to predict successful landings, the result can be seen at image below:

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.88929	0.77778
KNN	0.84821	0.83333

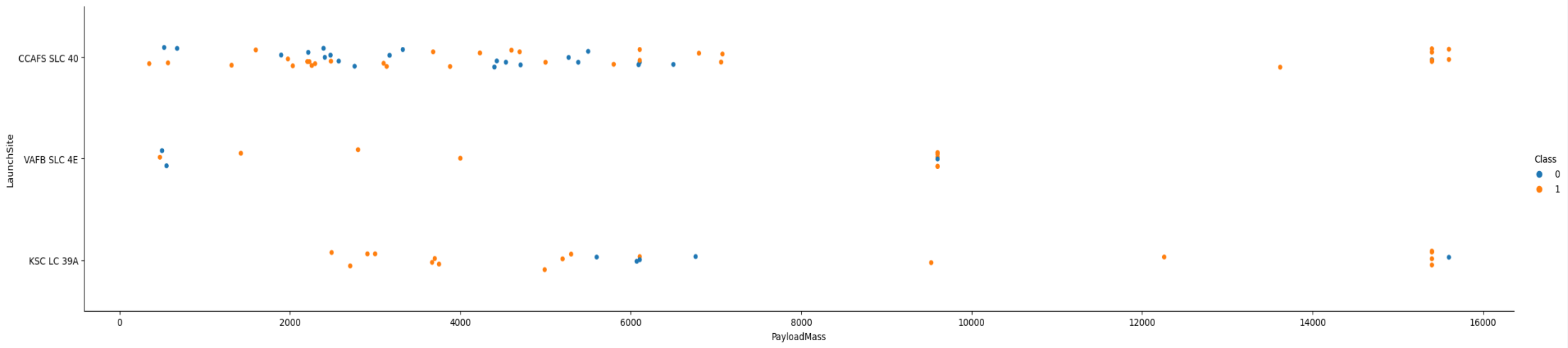

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- From plot above, it's possible to see that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful. In second place VAFB SLC 4E and third place KSC LC 39A.
- It can also be seen that the general success rate improved over time.

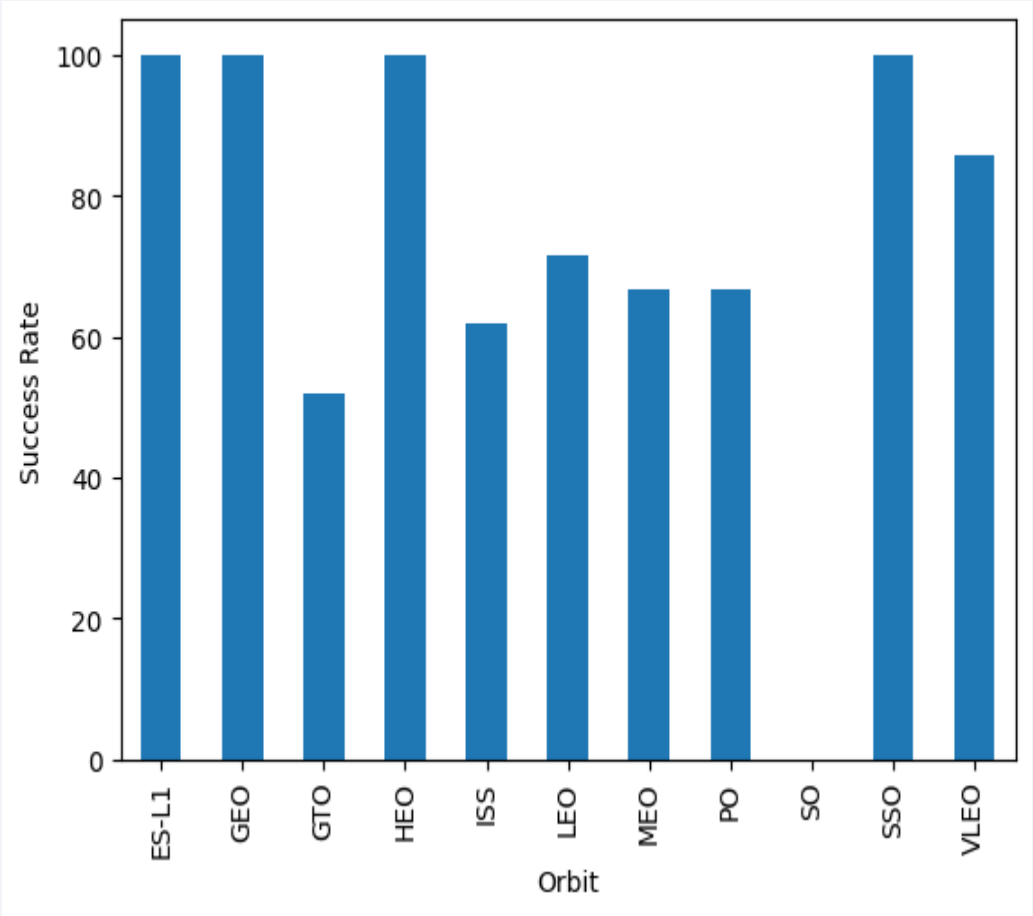
Payload vs. Launch Site



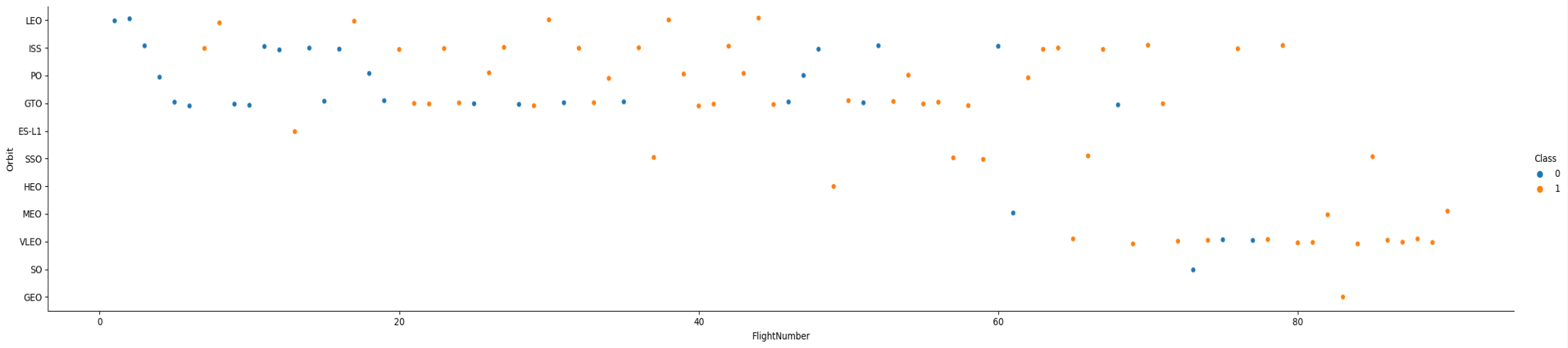
- Payloads over 12,000kg only possible on CCAFS SLC 40 and KSC LC 39A.
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate.

Success Rate vs. Orbit Type

- Highest success rate are with orbits ES-L1, GEO, HEO, and SSO.

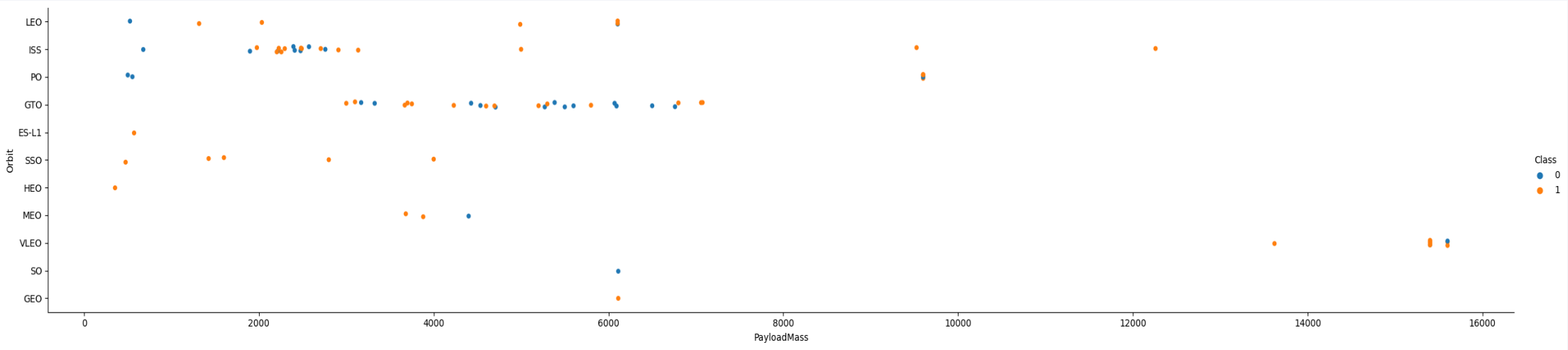


Flight Number vs. Orbit Type



- Success rate constantly improving over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

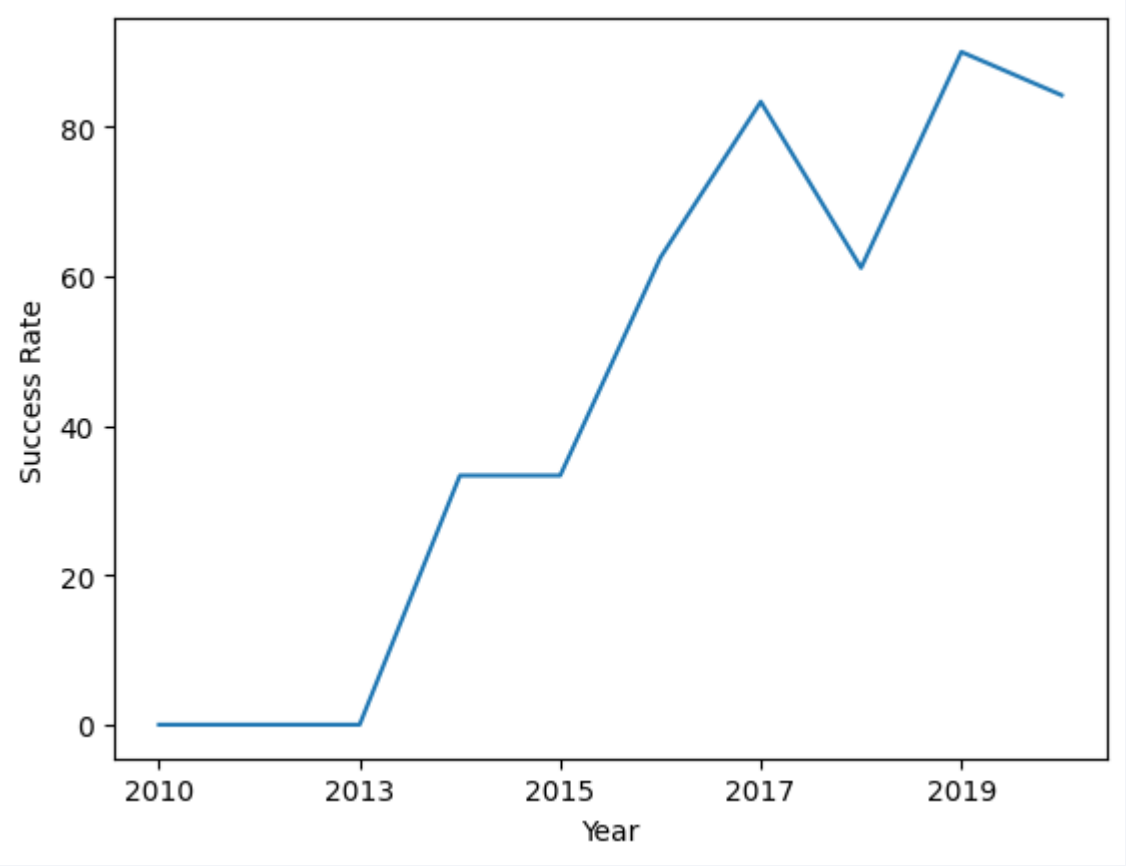
Payload vs. Orbit Type



- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- The line chart show success rate show uptrend from 2013 until 2020.
- The first 3 year remain flat because its still not giving best performance yet.



All Launch Site Names

- Below table are all launch site name:

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

- That result obtained from selecting unique value from launch_site column in SPACEXTBL

Launch Site Names Begin with 'CCA'

- Find 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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- Above are 5 sample of Cape Canaveral launches.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA:

SUM(PAYLOAD_MASS__KG_)
45596.0

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

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1:

- Data filtered by the booster version and calculating the average payload mass.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad:

MIN(DATE)
01/08/2018

- Selecting lowest date value and filtering to find only ground pad landing success giving result in above table.

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:

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

- The result is earned by executing query selecting unique booster version according to filters above.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes:

COUNT(MISSION_OUTCOME)
98

- Result above is successful launch with 1 other result with payload unclear and 1 failure launch.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:

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

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

2015 Launch Records

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

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

- Only 2 failed landing outcome occurrence in 2015.

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:

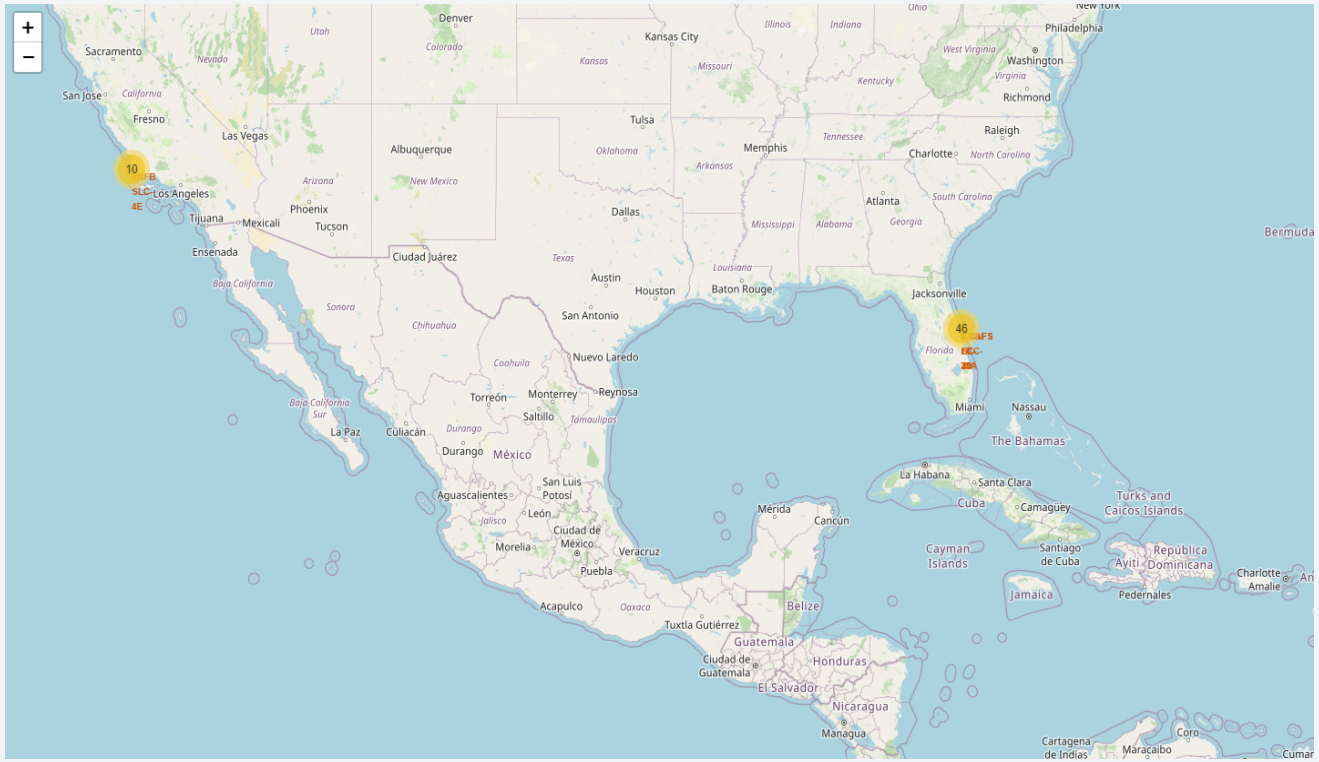
- Query show no findings.

Section 3

Launch Sites Proximities Analysis

Launch Sites General Location

- All launch sites nearing the sea and also near the roads and railroads.

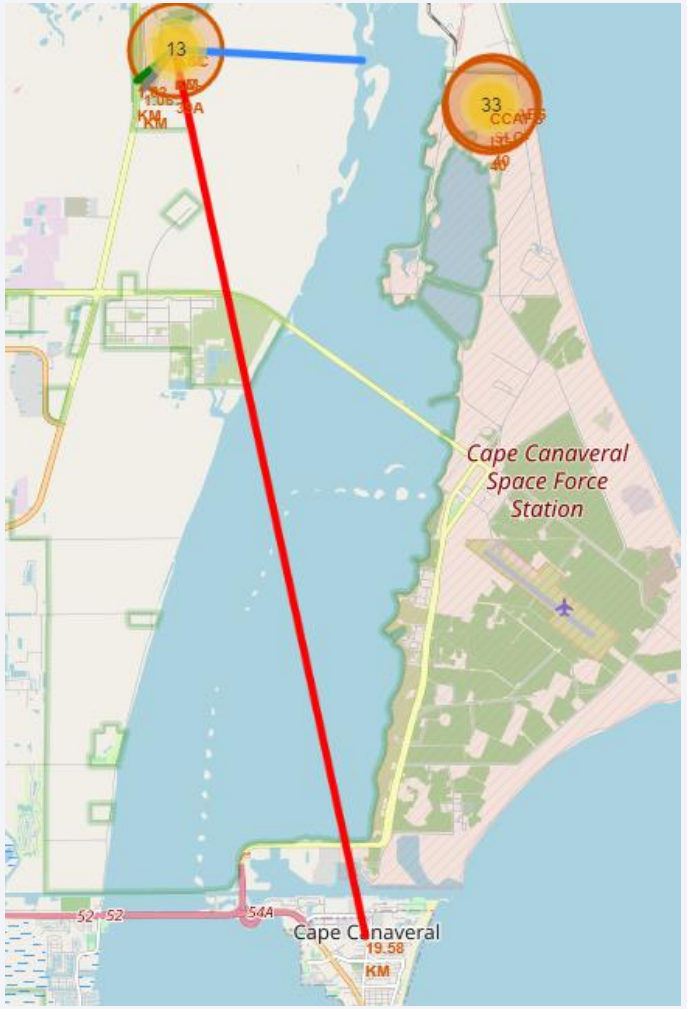


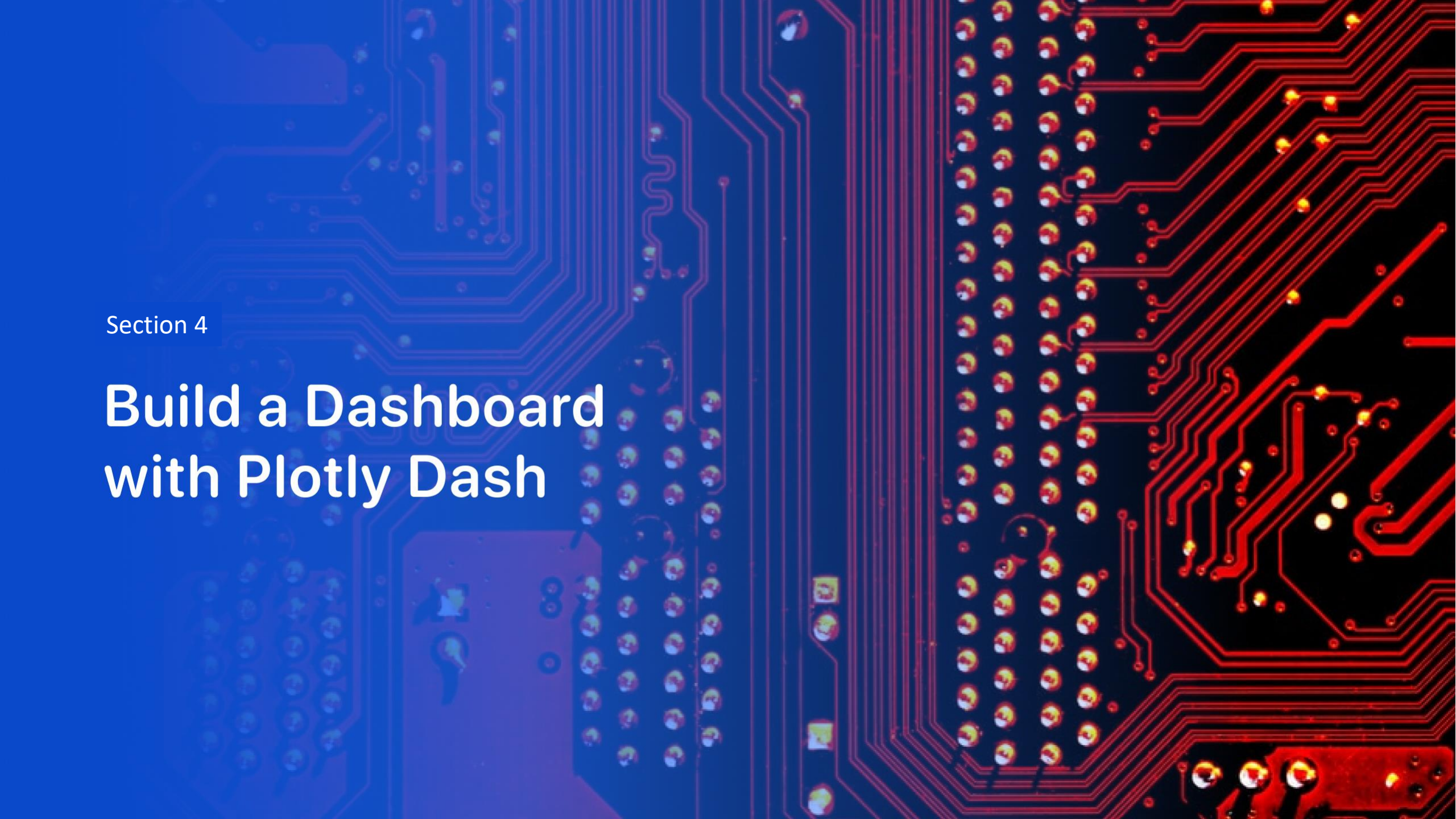
Launch Outcomes by Site

- Launch outcomes grouped by sites, below can be seen on image that the green marker indicate successful ones and the red ones indicate failure.

Strategic Location

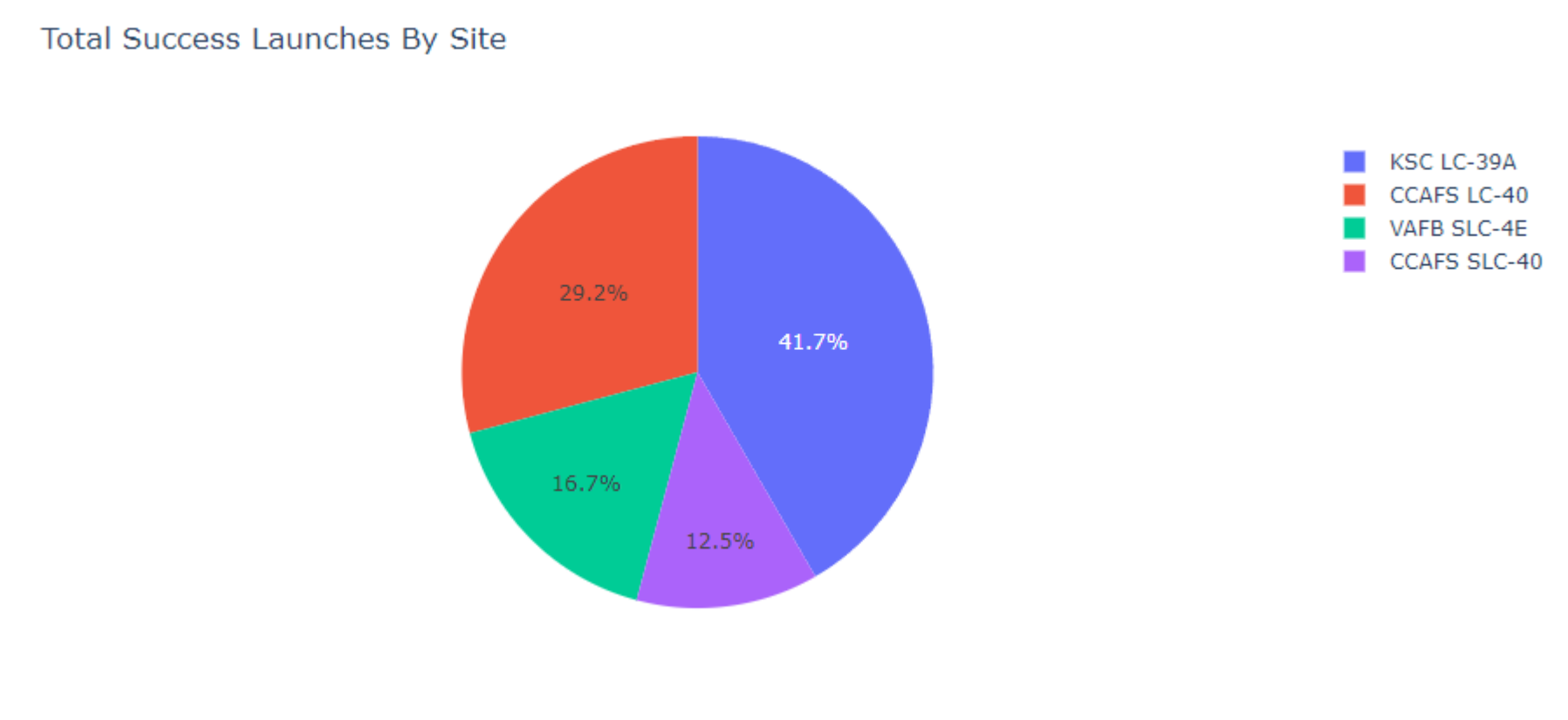
- Launch sites as can be seen on image on the right the location is near the sea, roads, and railroads but far from city.




Section 4
Build a Dashboard with Plotly Dash

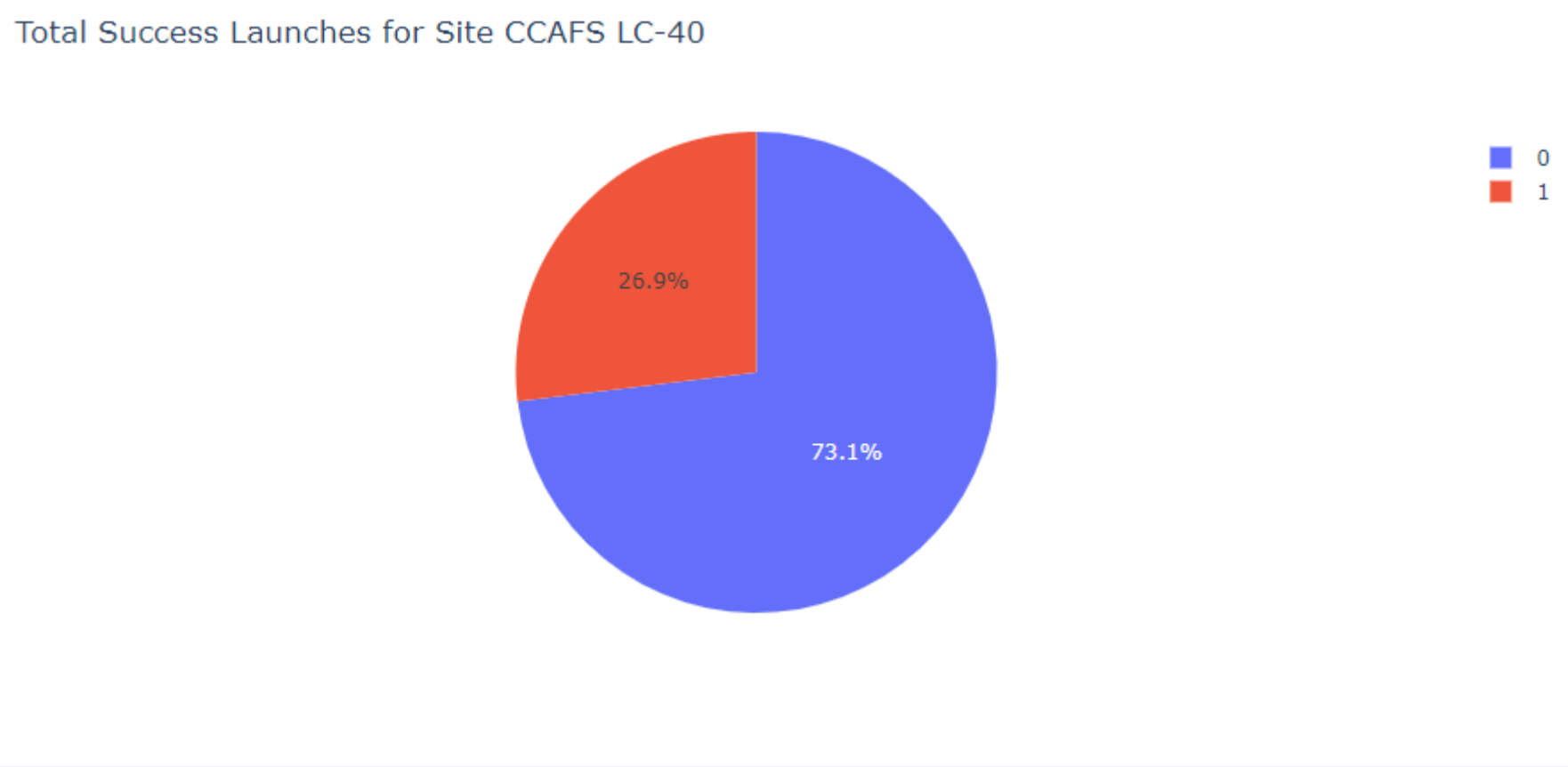
Successful Launches by Site

- Location is significant factor of successful launch.



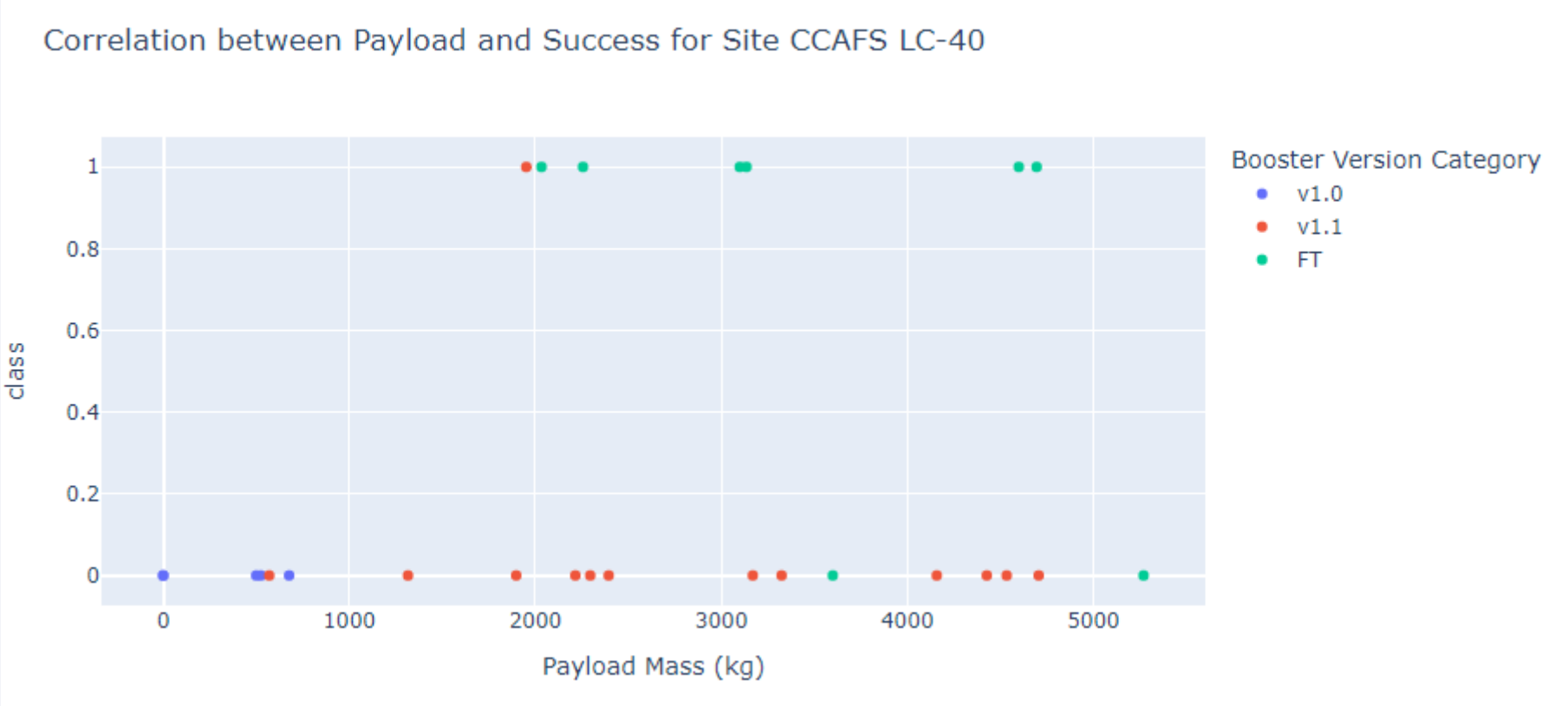
Launch Success Ratio for CCAFS LC-40

- 73,1% launches are successful in this site.



Payload vs Launch Outcome

- Payloads under 6,000kg and FT boosters are the most successful combination for sites CCAFS LC-40

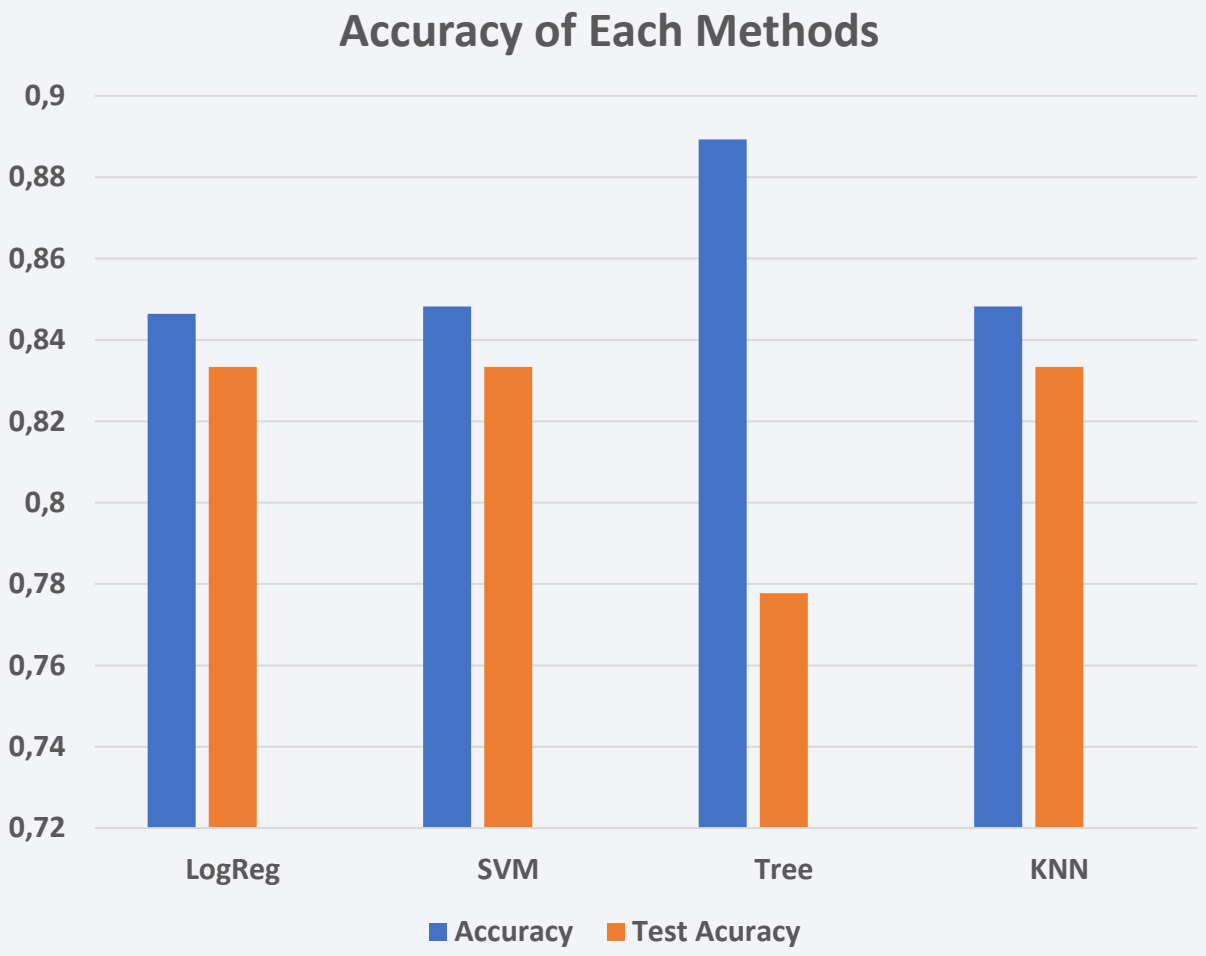


Section 5

Predictive Analysis (Classification)

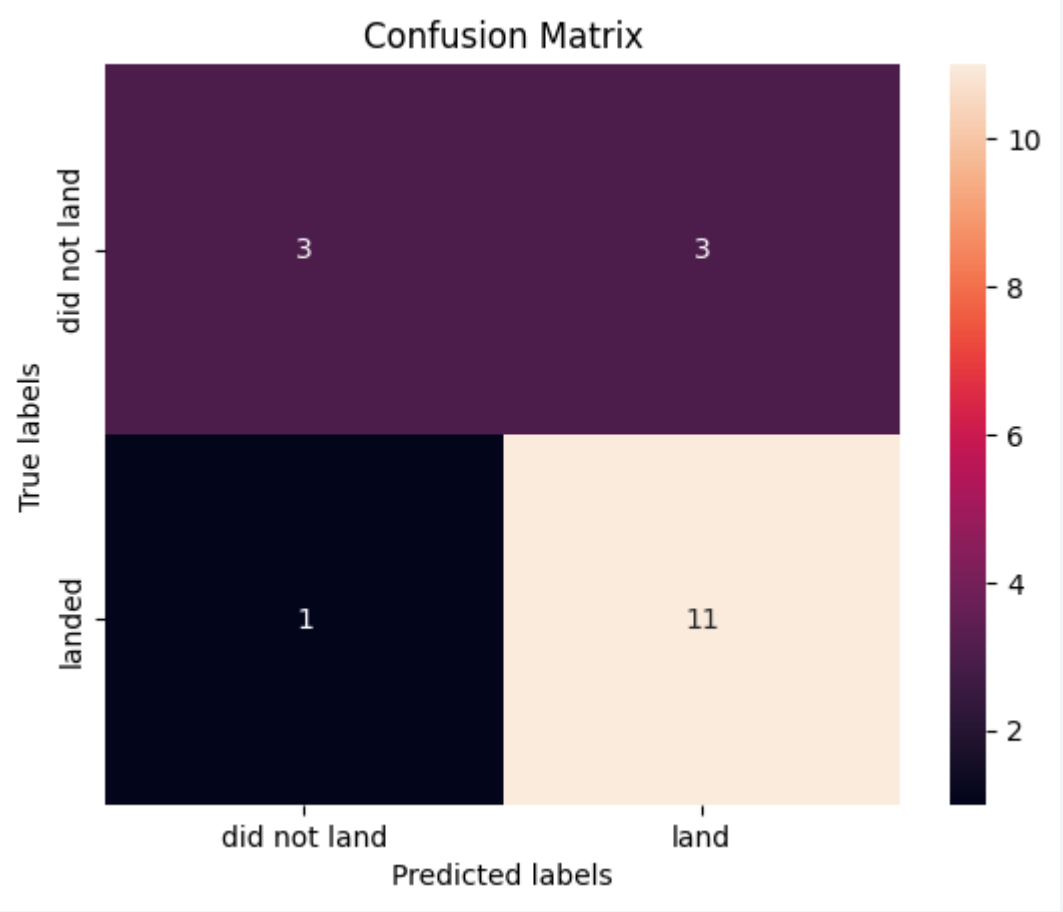
Classification Accuracy

- Beside are all classification model plotted to find best models.
- Decision tree have the highest accuracy.



Confusion Matrix

- Decision tree classifier show high number of true positive and true negative



Conclusions

- The best launch site is KSC LC-39A.
- Launches above 7,000kg are less risky.
- Successful landing outcomes seem to improve over time, according the evolution of technologies.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- Folium didn’t show maps on Github, so I took screenshots.

Thank you!