Winning Space Race with Data Science

Ho Nhut Minh
25-Jan-2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Executive Summary

- The following methodologies were used to analyze data:
  - Data Collection using web scraping and SpaceX API;
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
  - Machine Learning Prediction.

- Summary of all results
  - It was possible to collected valuable data from public sources;
  - EDA allowed to identify which features are the best to predict success of launchings;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

The objective is to evaluate the viability of the new company Space Yto compete with Space X.

- Desirable answers:
- The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
- Where is the best place to make launches.

- **Project background and context**

This project will explore successful landing of rocket launch in aero space industry. SpaceX advertises the Falcon 9 rocket on its website at a much lower cost of 62 millions compared to the cost of 165 millions from other providers. Therefore, understanding how SpaceX can achieve this cost efficiency is crucial to break into this market and bid against SpaceX. Indeed, much of the savings is because SpaceX can reuse the first stage. As a result, if we can determine if the first stage will land successfully, we can determine the cost of a overall launch and have appropriate strategic actions to optimize upon.

- **Problems you want to find answers**

We will use data analysis, visualization and relevant statistical model(s) to answer our main problem – i.e. to predict whether the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data collection methodology:

  Data from Space X was obtained from 2 sources:

  Space XAPI(https://api.spacexdata.com/v4/rockets/)

  WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

  The data is initially extracted from SpaceX API of launches in the past, then using predefined helper functions to transform into a data frame containing relevant data attributes. Finally, the data is filtered to only include Falcon 9 launches which is the type of rocket (i.e. booster version) that the analysis will focus on.

- Perform data wrangling

  Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features. Identified that only Payload Mass and Landing Pad attributes have missing information. One of the approaches is to replace missing Payload Mass with the average payload mass value from our dataset. On the other hand, Landing Pad will retain the None value to represent the launches where landing pads were not used and described how data was processed.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Methodology

Executive Summary

- Data collection methodology:

Data collection methodology:

Data from Space X was obtained from 2 sources:

Space XAPI(https://api.spacexdata.com/v4/rockets/)

WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

The data is initially extracted from SpaceX API of launches in the past, then using predefined helper functions to transform into a data frame containing relevant data attributes. Finally, the data is filtered to only include Falcon 9 launches which is the type of rocket (i.e. booster version) that the analysis will focus on.

Data sets were collected from Space XAPI (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

- Perform data wrangling

Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features. Identified that only Payload Mass and Landing Pad attributes have missing information. One of the approaches is to replace missing Payload Mass with the average payload mass value from our dataset. On the other hand, Landing Pad will retain the None value to represent the launches where landing pads were not used and described how data was processed.

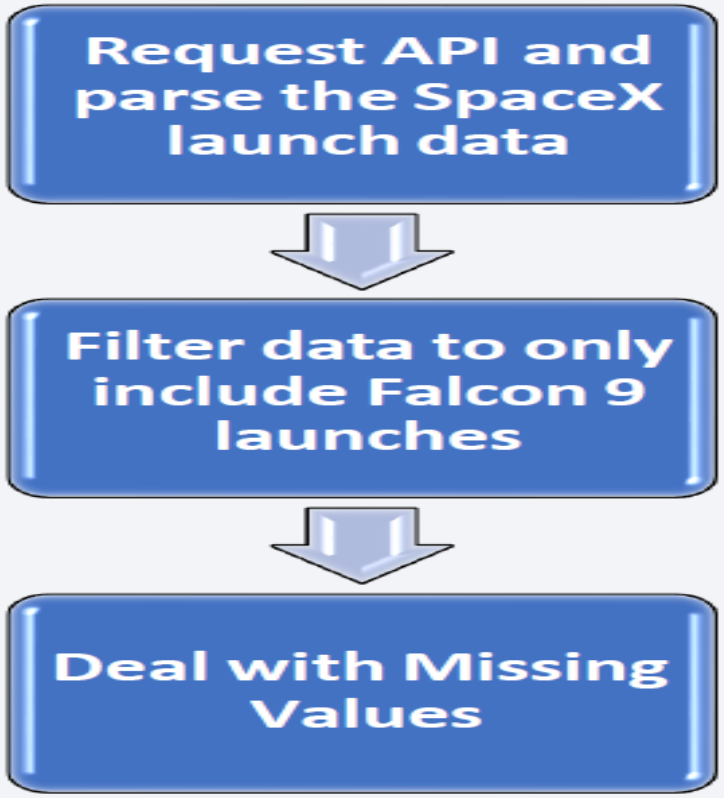
Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

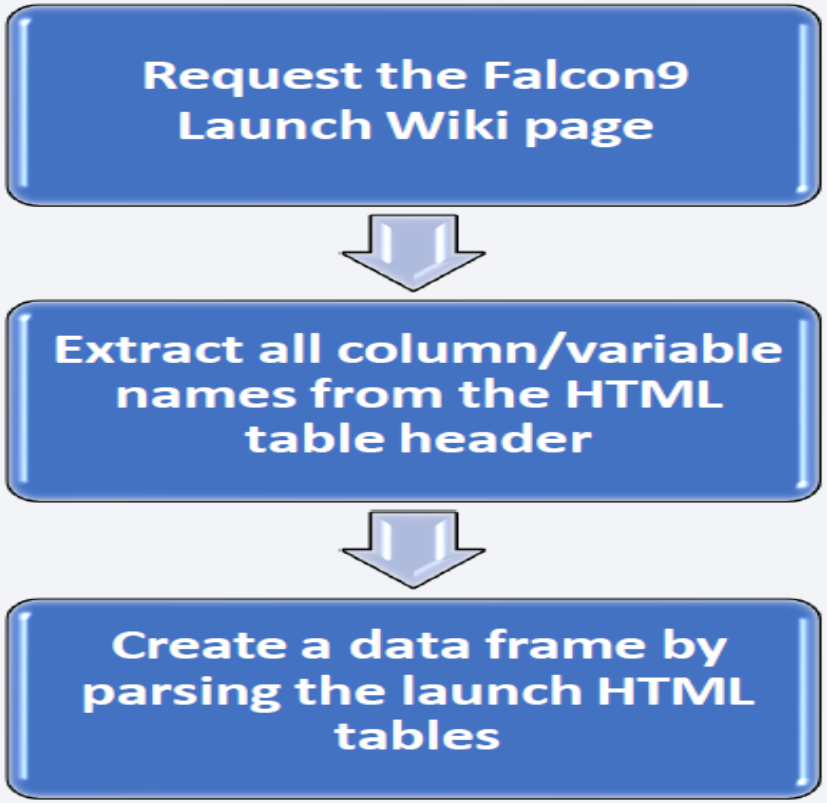
- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
- Source : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects.git



Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.



Source : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects.git

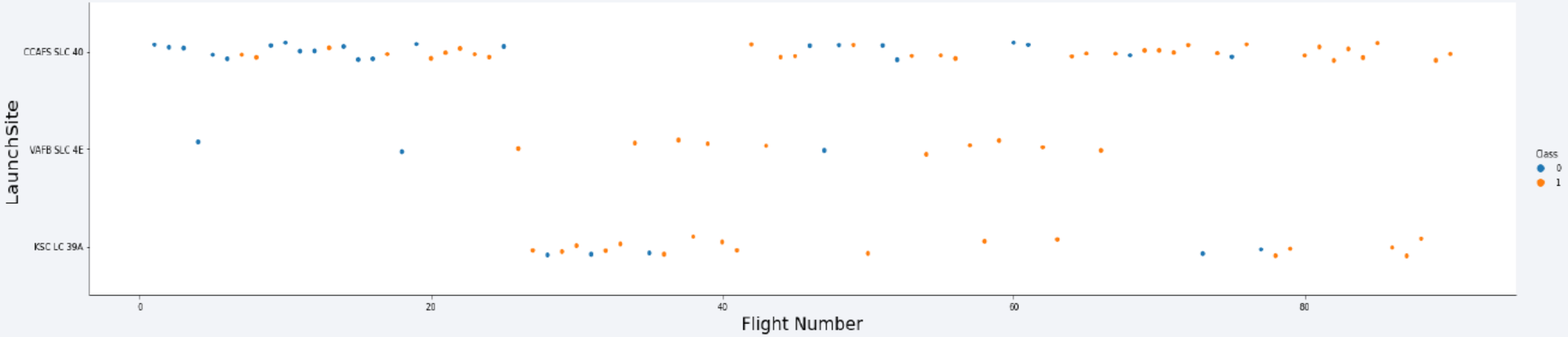
Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.

Source : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects.git

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



Source : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects.git

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;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - 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.

Source : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects.git

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASAJohnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
  - Lines are used to indicate distances between two coordinates.

Source : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects.git

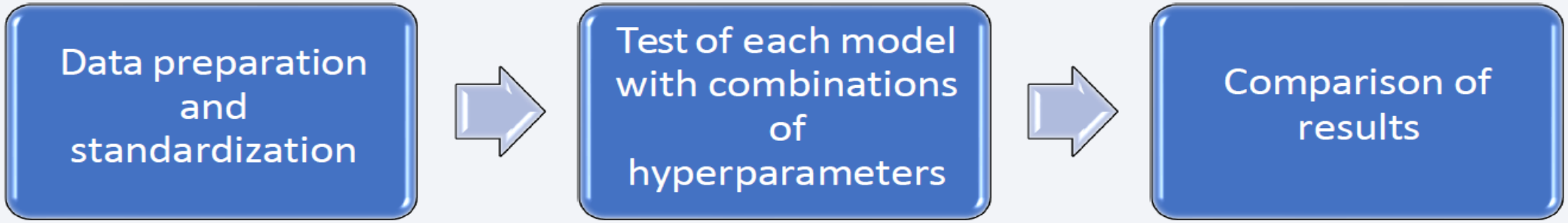
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - 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 : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects.git

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



Source : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects.git

Results

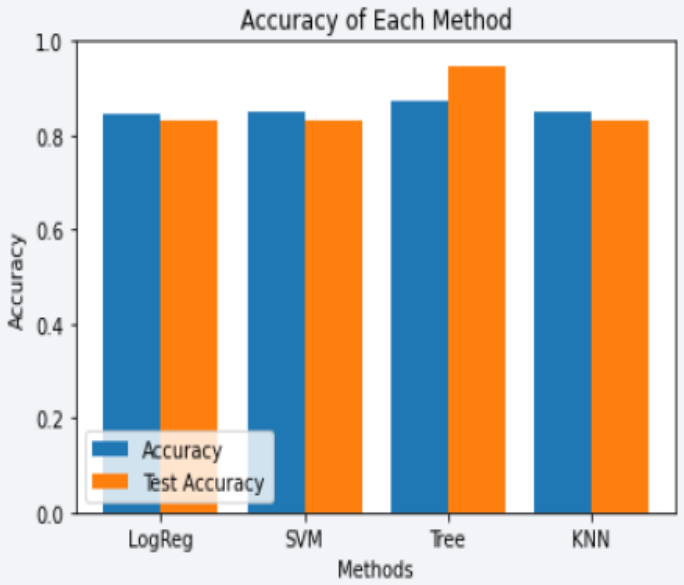
- Exploratory data analysis results:
  - SpaceXuses 4 different launch sites;
  - The first launches were done to SpaceXitself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 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 number of landing outcomes became as better as years passed.

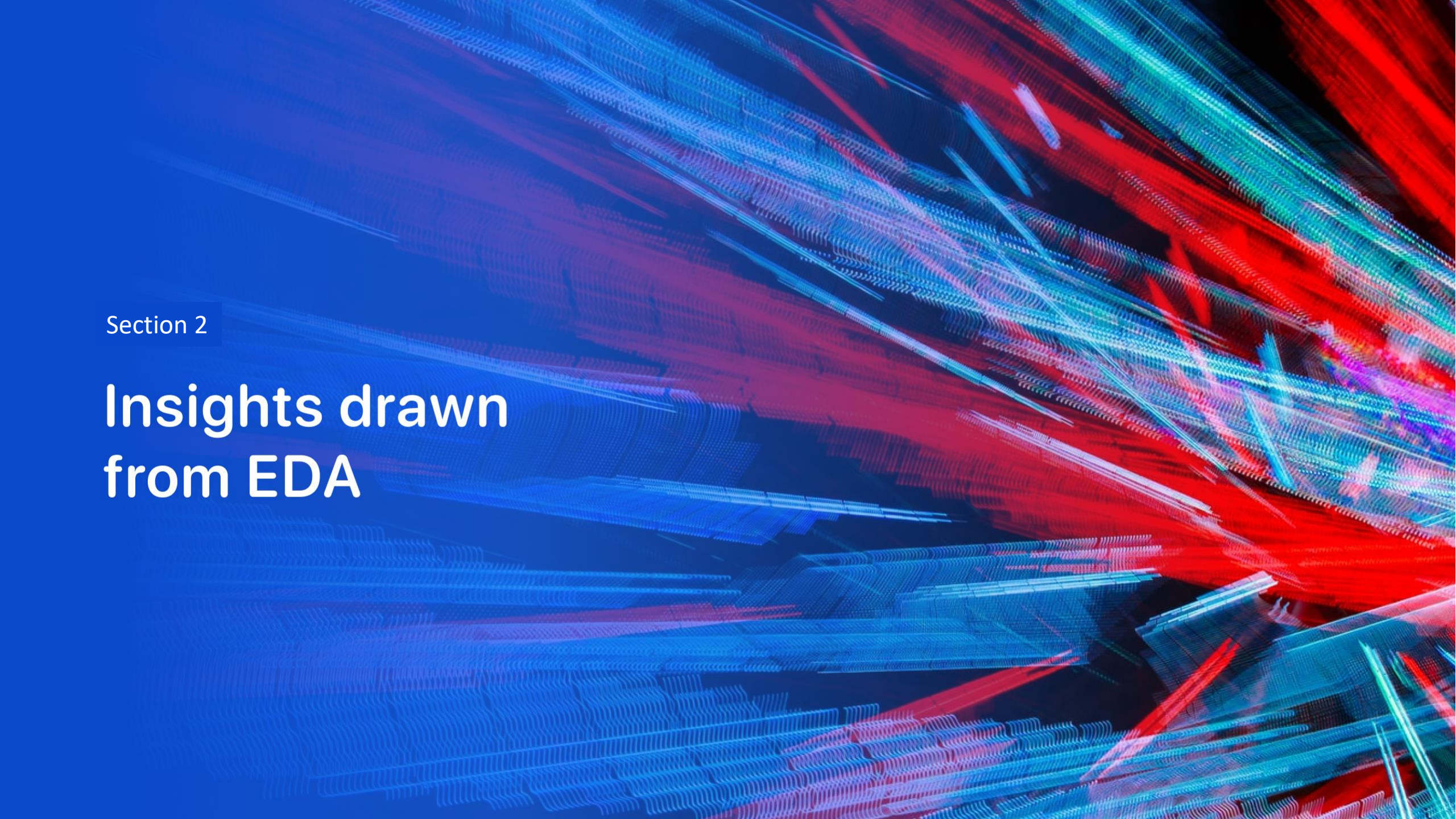
Results

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.

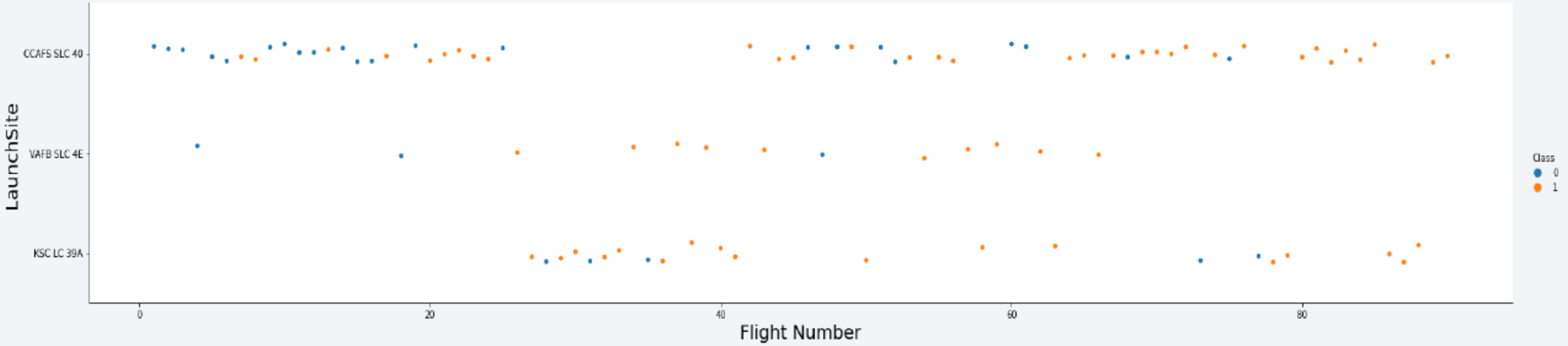
Results

- 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%.



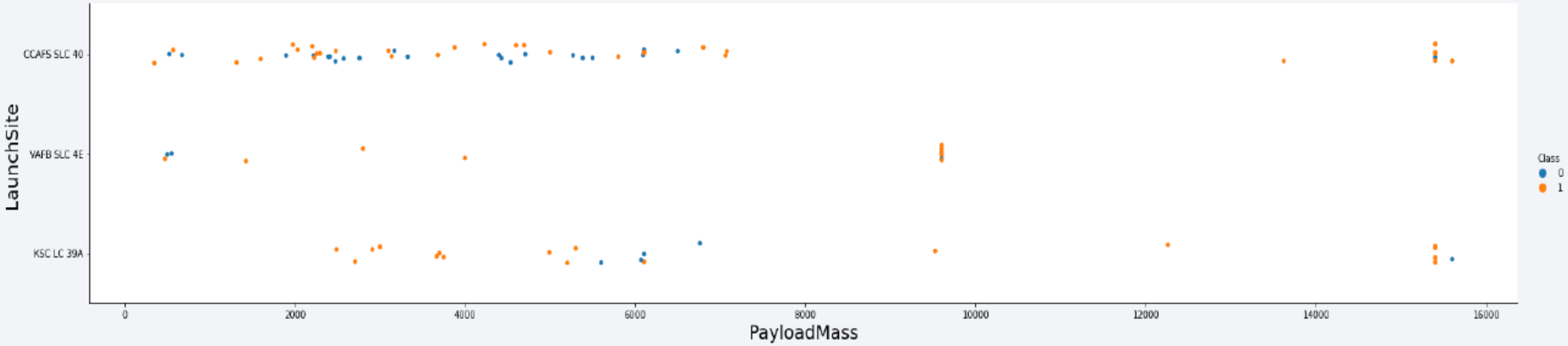

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC40, where most of recent launches were successful;
- In second place VAFB SLC4E and third place KSCLC 39A;
- It's also possible to see that the general success rate improved over time.

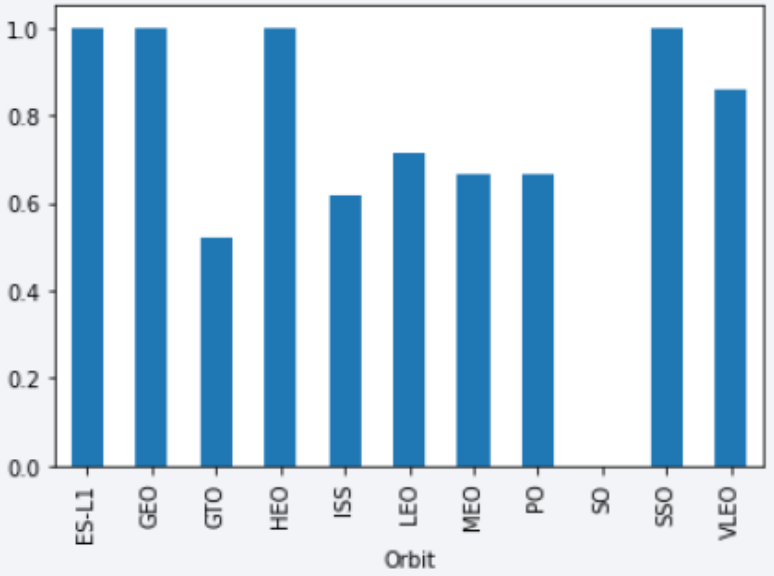
Payload vs. Launch Site



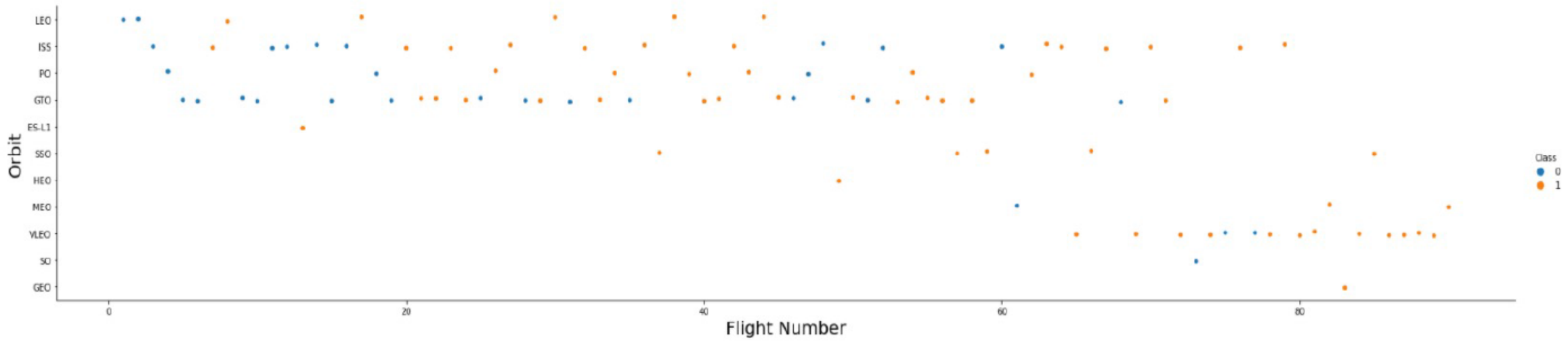
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFSSLC40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).



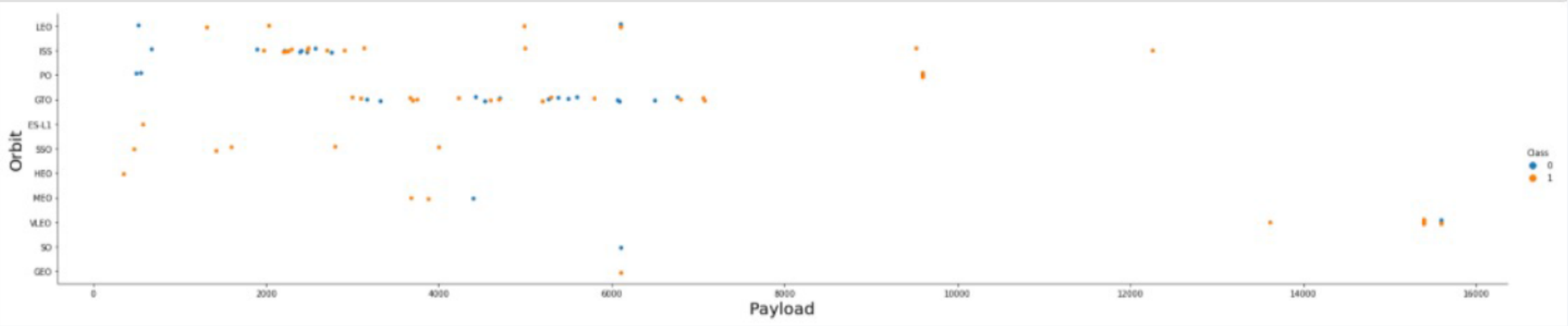
Flight Number vs. Orbit Type



Plot of Orbit vs Flight Number across the outcomes of launch

The relationship seems to vary based on specific orbit, e.g. in the LEO orbit, the Success rate appears to be related to the number of flights. Whereas there seems to be no relationship between flight number when in the GTO orbit

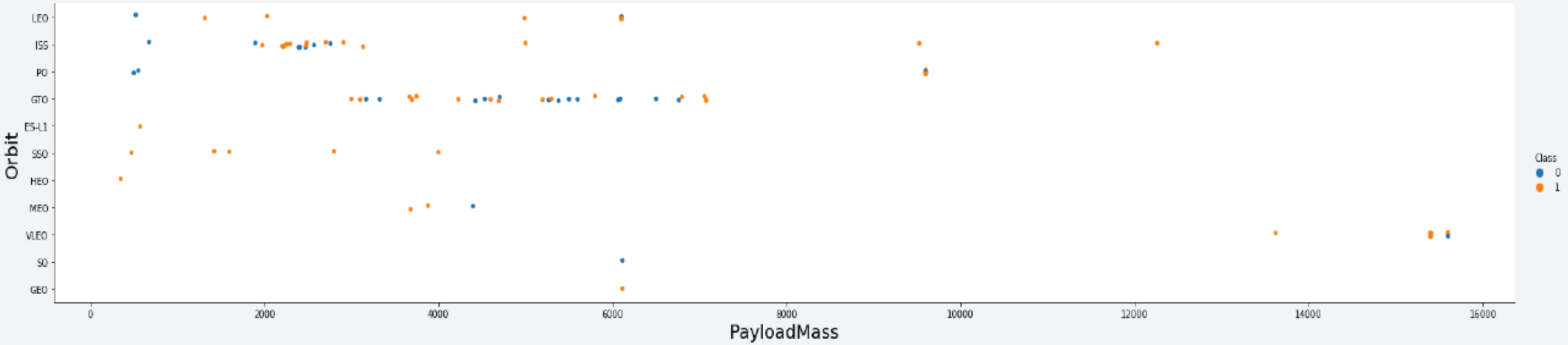
Payload vs. Orbit Type



Plot of Orbit vs Payload across the outcomes of launch

The relationship seems to vary based on specific orbit again here. With heavy payloads the successful landing are more for Polar, LEO and ISS. On the other hand, there is no clear distinction when looking at GTO

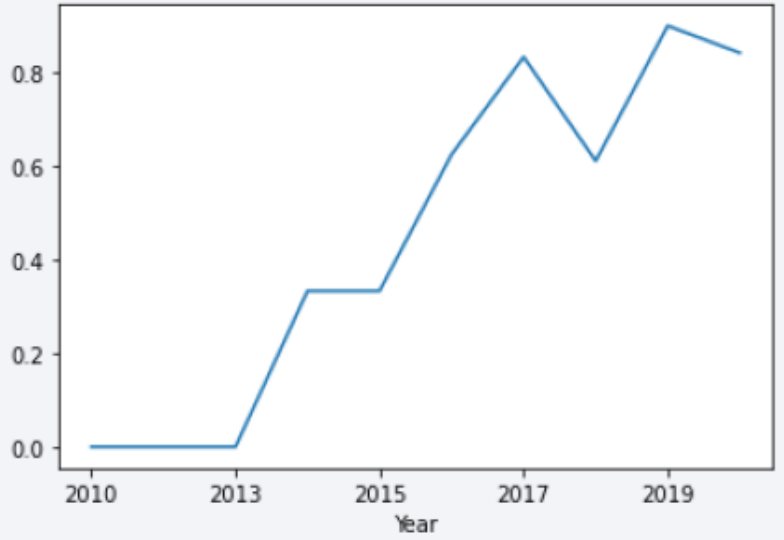
Payload vs. Orbit Type



- Apparently, there is no relation between payload and success rate to orbit GTO;
- ISSorbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SOand GEO.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are four 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
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- The total payload mass carried by boosters launched by NASA (CRS) is 45,596 kgs

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

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Avg Payload (kg)
2.928

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- 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

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

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

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

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

Boosters Carried Maximum Payload

- Boosters 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

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

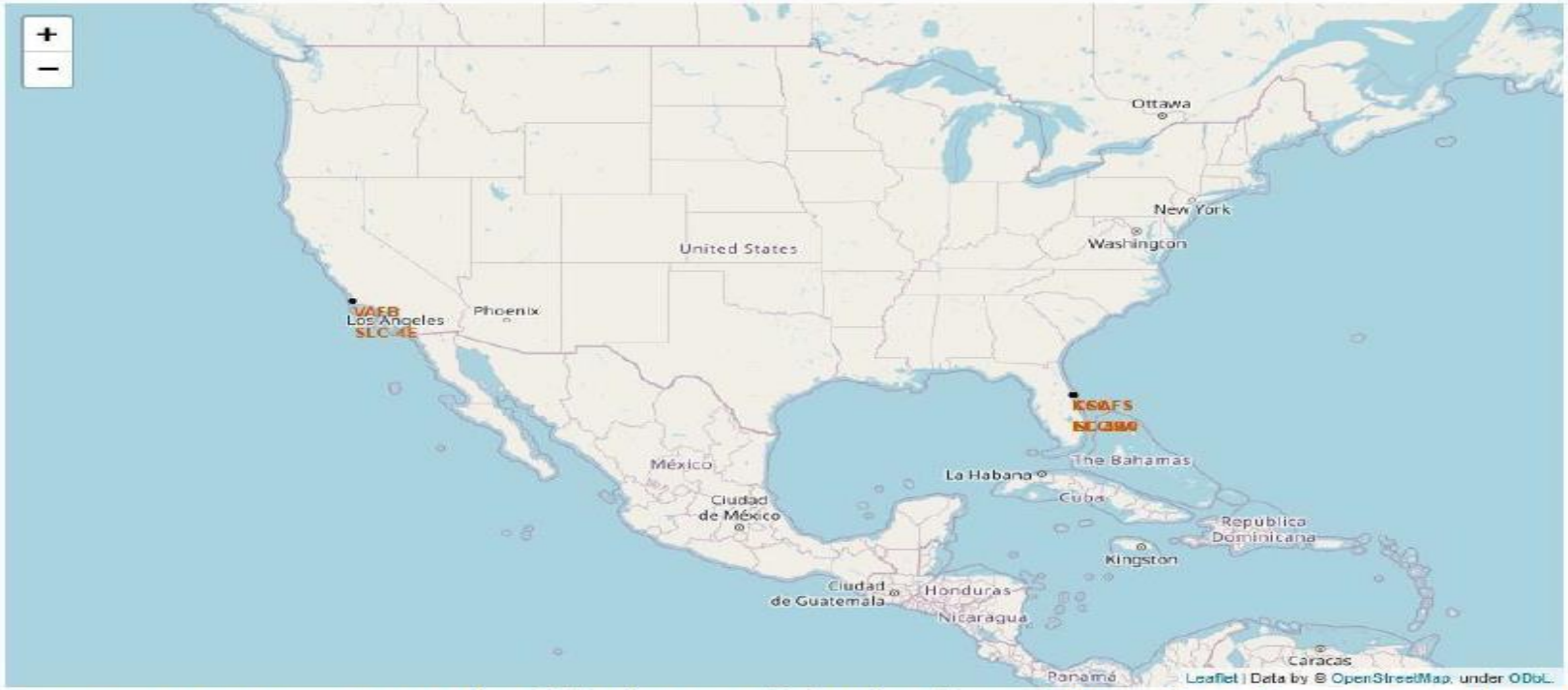
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- This view of data alerts us that “No attempt” must be taken in account.

Section 3

Launch Sites Proximities Analysis

All launch sites



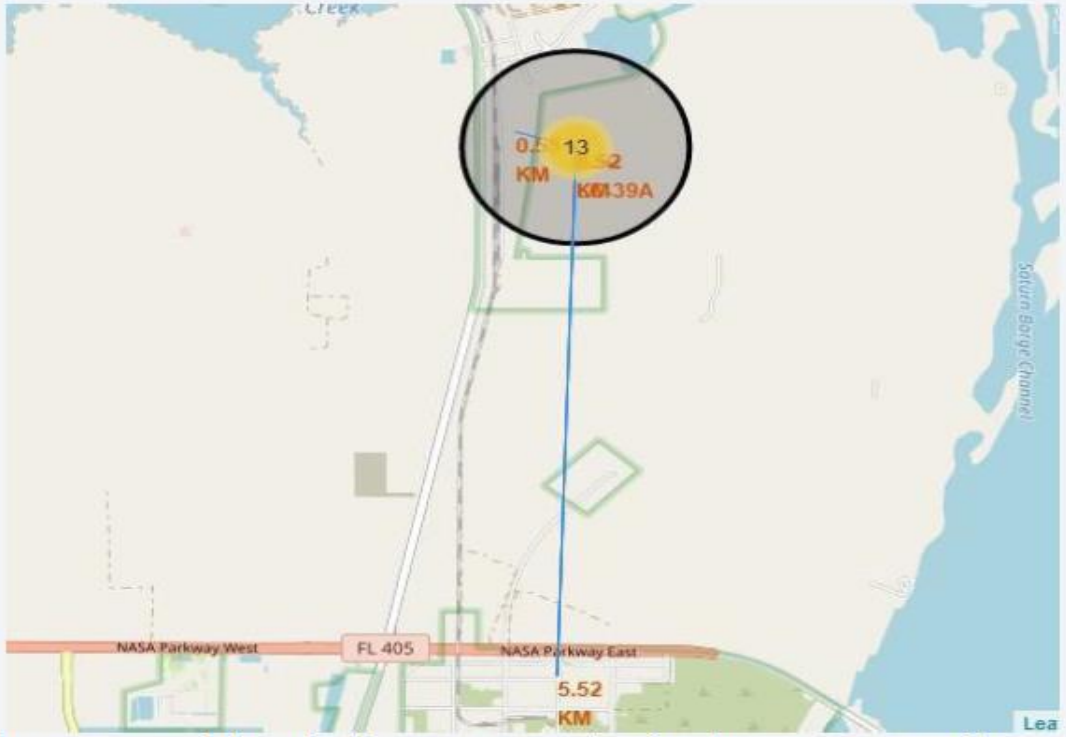
- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

Launch Outcomes by Site

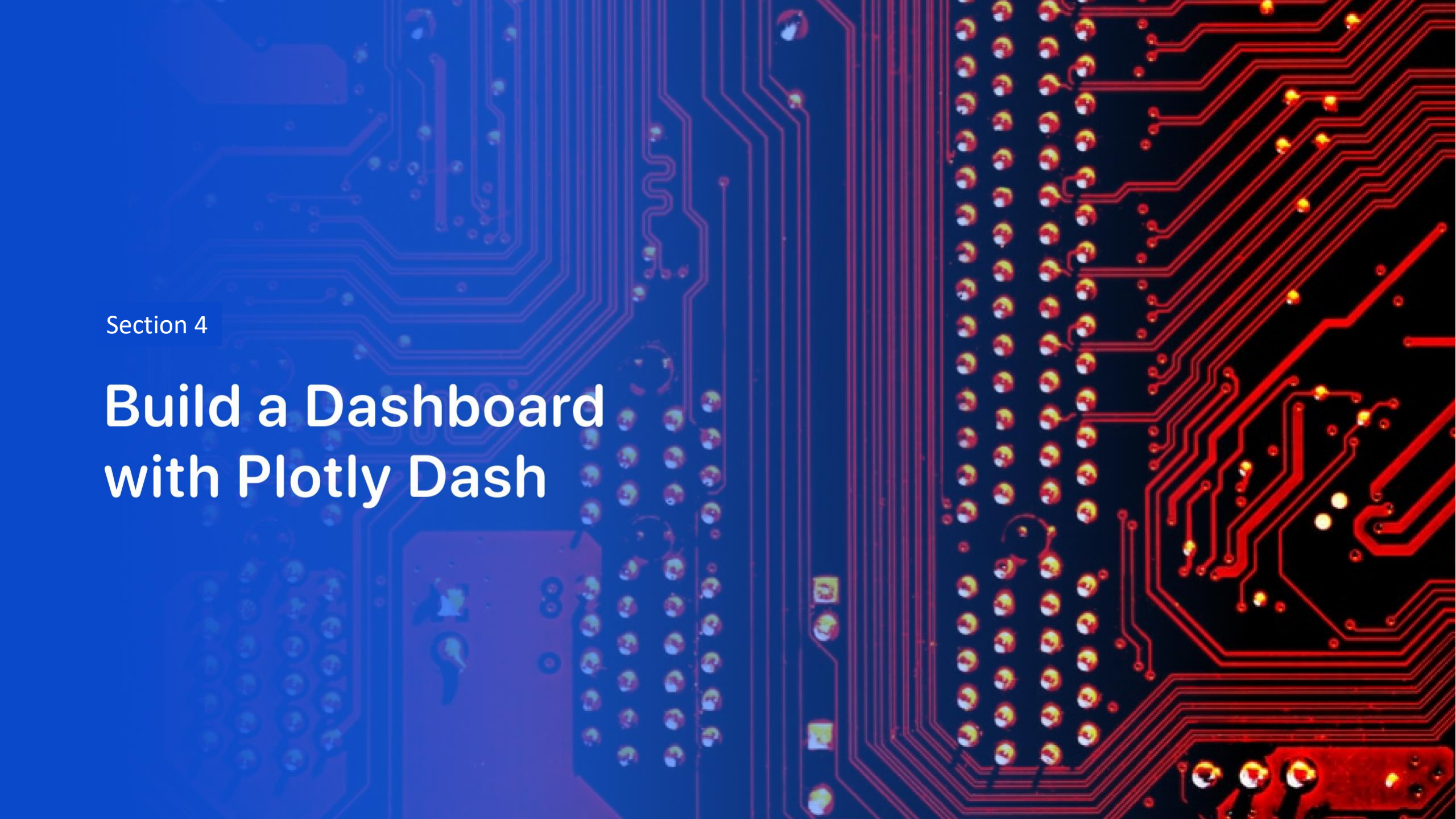
- Example of KSCLC-39A launch site launch outcomes

- Green markers indicate successful and red ones indicate failure.

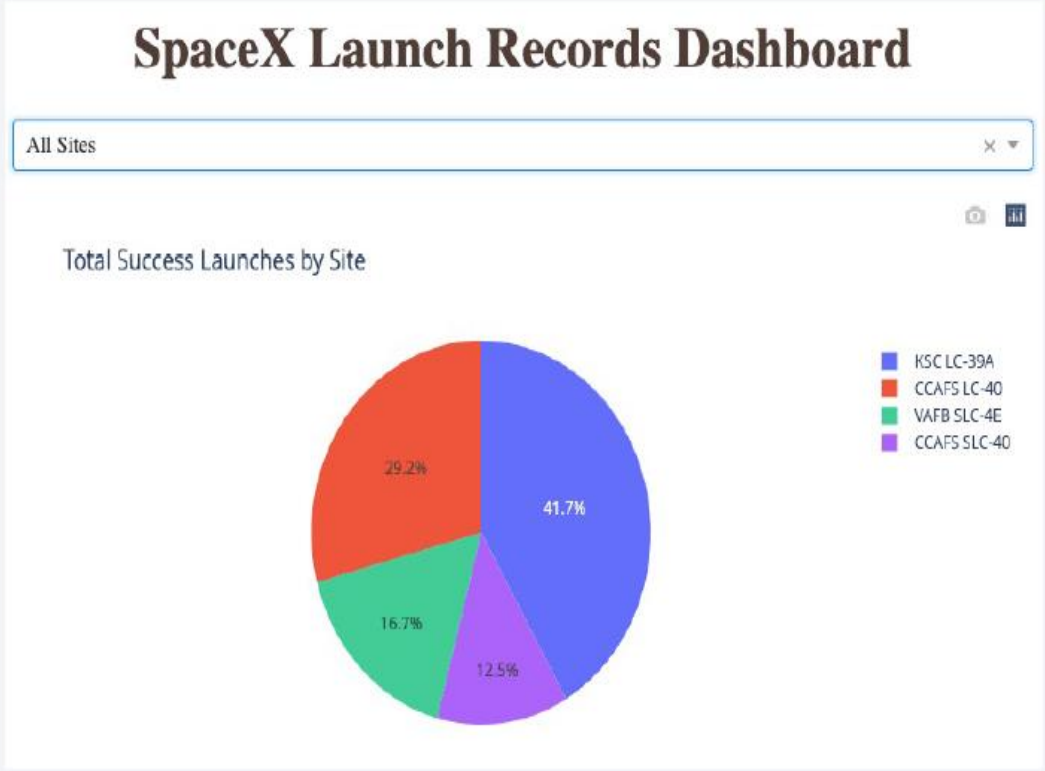
Logistics and Safety



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


Section 4
Build a Dashboard with Plotly Dash

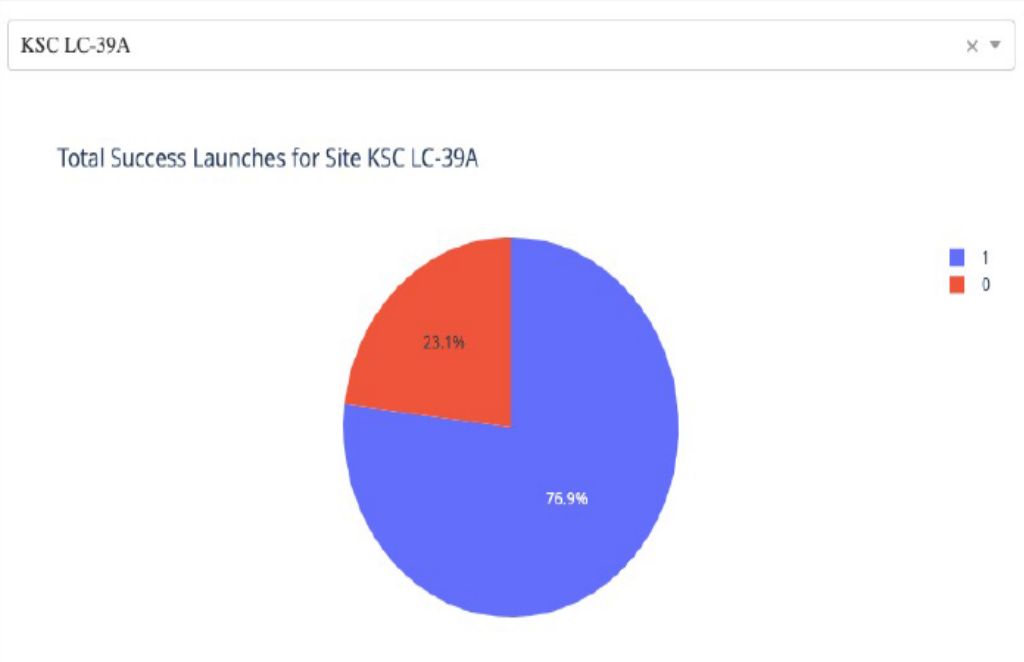
Total Success Launches by Site



The KSC LC-39A launch site contributes the highest percentage to total success launches across all sites (41.7%).

On the other hand, CCAFS SLC-40 controls the smallest amount to total success launches across all sites (12.5%).

Total Success Launches for Site KSC LC-39A

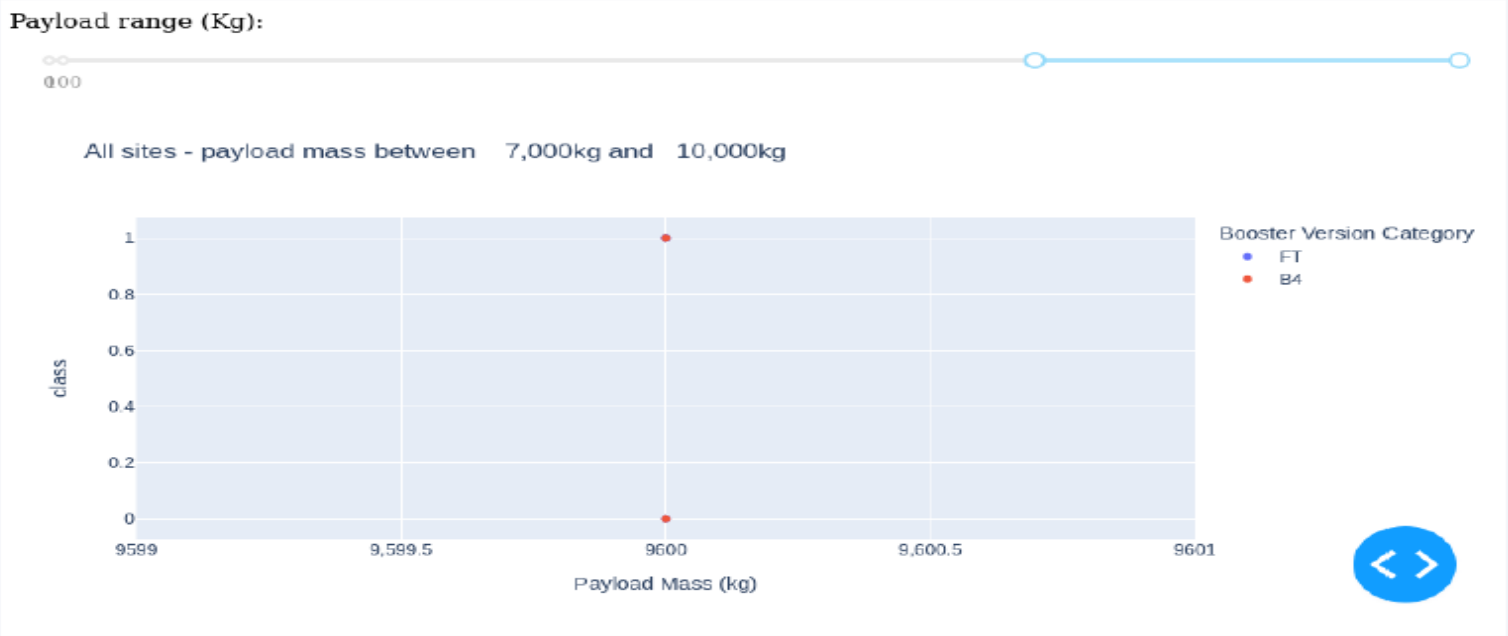


The KSC LC-39A launch site has almost 80% success launch rate for launches taking place at its site, whereas only 20% failed.

Launch outcomes between 2000-4000 kgs across all sites

When the payload mass slider is adjusted to between 2000 – 4000 kgs, the success outcomes increases compared to the failed outcomes. The performance seems to also be best for FT booster version based on colour indicated in the graph.

Payload vs. Launch Outcome



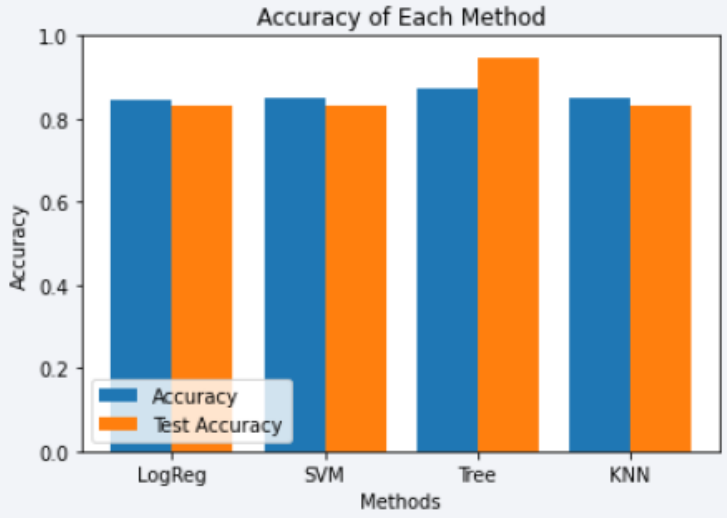
- There’s not enough data to estimate risk of launches over 7,000kg

Section 5

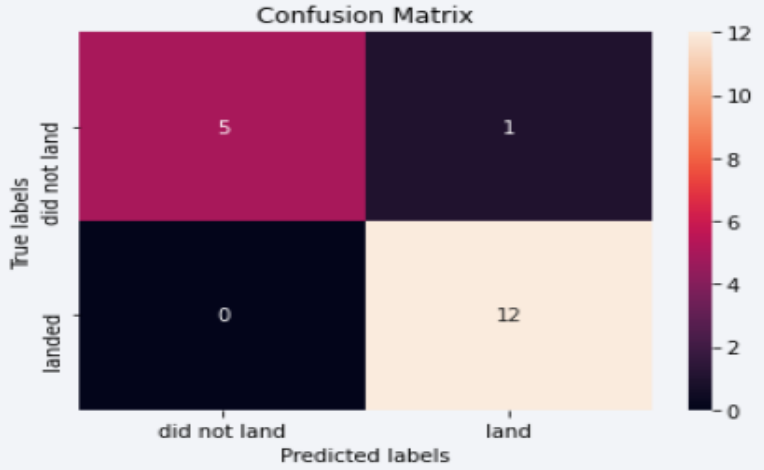
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix of Decision Tree Classifier



- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSCLC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- Include any relevant assets like
  - Python code snippets:

    https://builtin.com/data-science/python-code-snippets
  - SQL queries:

    https://towardsdatascience.com/sql-queries-in-python-51ef85b92c1e
  - SQL Cheatsheets

    https://learnsql.com/blog/sql-basics-cheat-sheet/
  - Charts:
  - https://plotly.com/python/basic-charts/
  - Notebook outputs:
  - https://jupyterbook.org/en/stable/content/code-outputs.html
  - data sets that you may have created during this project : https://github.com/ikollenchery/Applied-Data-Science-Capstone-Projects/blob/97db29128d1801aa0e6142e91fdd60f0cd83d7b9/Spacex.csv

Thank you!