Winning Space Race with Data Science

<CHE-WEI NIEN>
<8/8/2023>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Exploratory Data Analysis(EDA)
  - Interactive Visual Analytics and Dashboard
  - Predictive Analysis
- Summary of all results
  - EDA results
  - Interactive map and dashboard
  - Prediction results

Introduction

- Project background

  This project is to collect and analyze SpaceX information and build machine learning model to predict Falcon 9 could be reuse at the first stage successfully. As a result, we want to further determine the cost of the launch.

- Problems you want to find answers
  - How to determine the cost of each launch?
  - What kinds of factors are important to the cost of a launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from SpaceX REST API and web scraping related Wiki pages.
- Perform data wrangling
  - the data processing for data wrangling involved using an API to enrich the dataset, filtering out unwanted data, handling NULL values and preparing the data for later steps such as one-hot encoding.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The process includes data preprocessing, feature engineering, data splitting, model selection, modeling training, model evaluation, model evaluation, model comparison and interpretation.

Data Collection

Data were collected from 2 sources below:

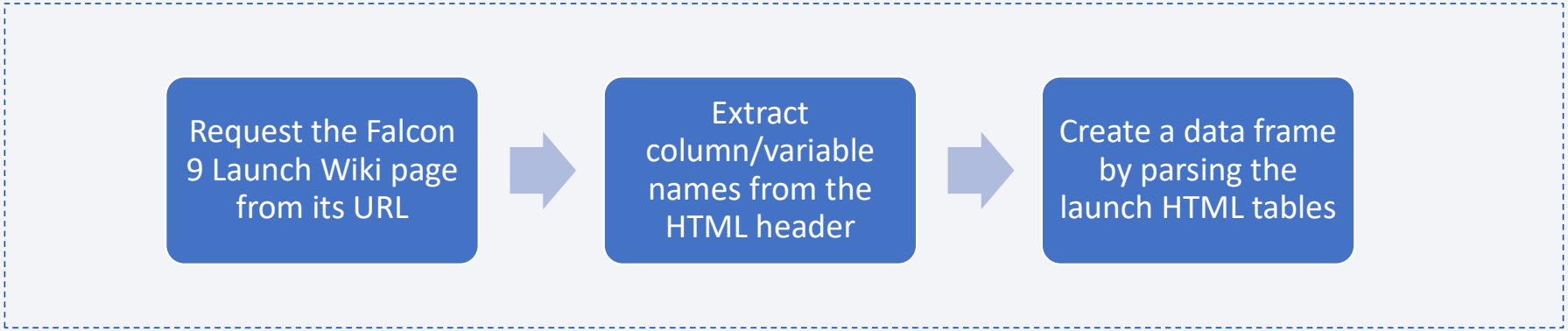
- Space X REST API (https://api.spacexdata.com/v4/rockets/)

- Wiki pages (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

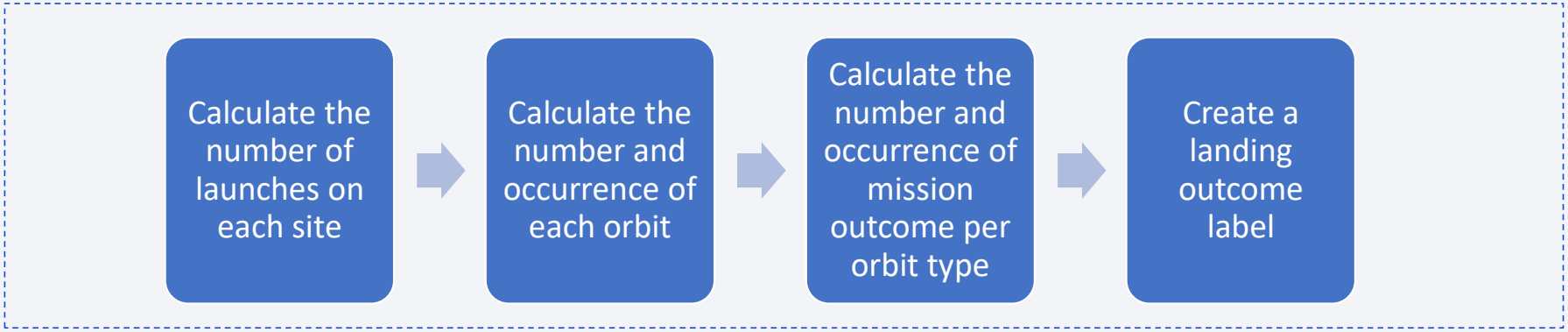
Source code
https://github.com/hanknien/Applied_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

Data Collection - Scraping



Source code
https://github.com/hanknien/Applied_Data_Science_Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling



Source code
https://github.com/hanknien/Applied_Data_Science_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- For data visualization, scatterplots ,barplots, lineplots are used to present the relationship between features in this project.

Plot type	Relationship
Scatterplots	FlightNumber vs. PayloadMass FlightNumber vs. LaunchSite FlightNumber vs. Orbit type PayloadMass vs. LaunchSite PayloadMass vs. Orbit type
Barplots	Success rate of Orbit type
Lineplots	Average launch success yearly trend

Source code
https://github.com/hanknien/Applied_Data_Science_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite%20(1).ipynb

EDA with SQL

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

Source code
https://github.com/hanknien/Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- In this project, we used markers, circles, lines and marker clusters with Folium
- Circles and markers were used to highlight circle area with a text label for launch sites on the site map.
- Marker clusters were a good way to simplify a map containing many markers having the same coordinate.
- Lines were drawn to show the distance between a launch site to its closest city, coastline, highway.

Source code
https://github.com/tflores/applied-data-science-capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

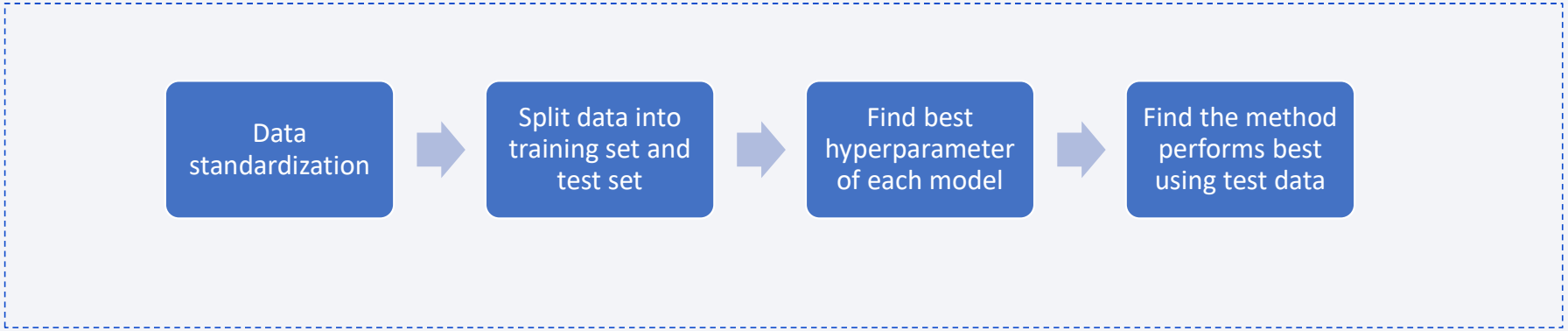
Build a Dashboard with Plotly Dash

- Pie chart and scatter chart were used to show the total successful launches count for all sites and correlation between payload and launch success respectively.

Source code
https://github.com/hanknien/Applied_Data_Science_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We tried to use logistic regression, SVM, decision tree, k nearest neighbors model to predict first stage landing



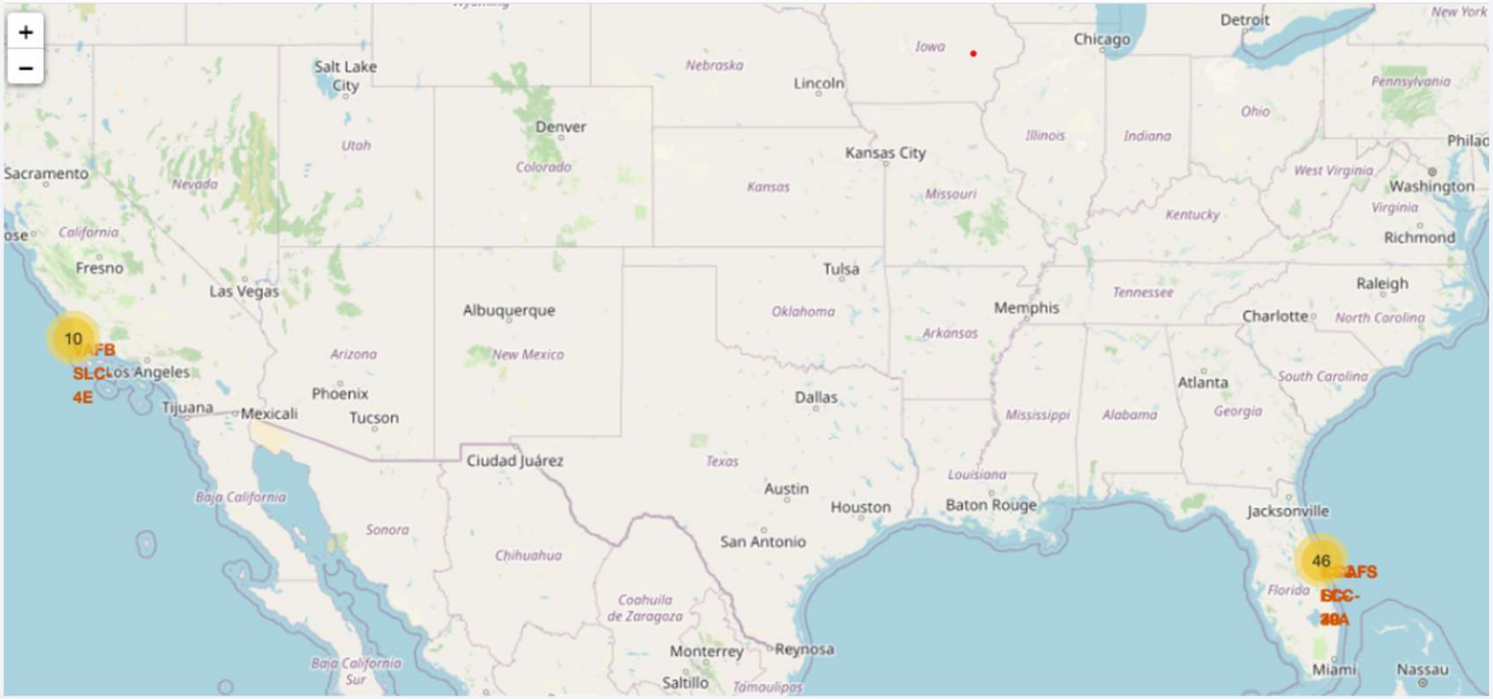
Source code
https://github.com/hanknien/Applied_Data_Science_Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
  - The flight number increases, the first stage is more likely to land successfully.
  - The more massive the payload, the less likely the first stage will return
  - KSC LC-39A and VAFB SLC 4E has a success rate of 77%
  - No rockets launched for heavypayload mass in VAFB SLC
  - ES-L1, GEO, HEO,SSO orbit have high success rate
  - The success rate since 2013 kept increasing till 2020

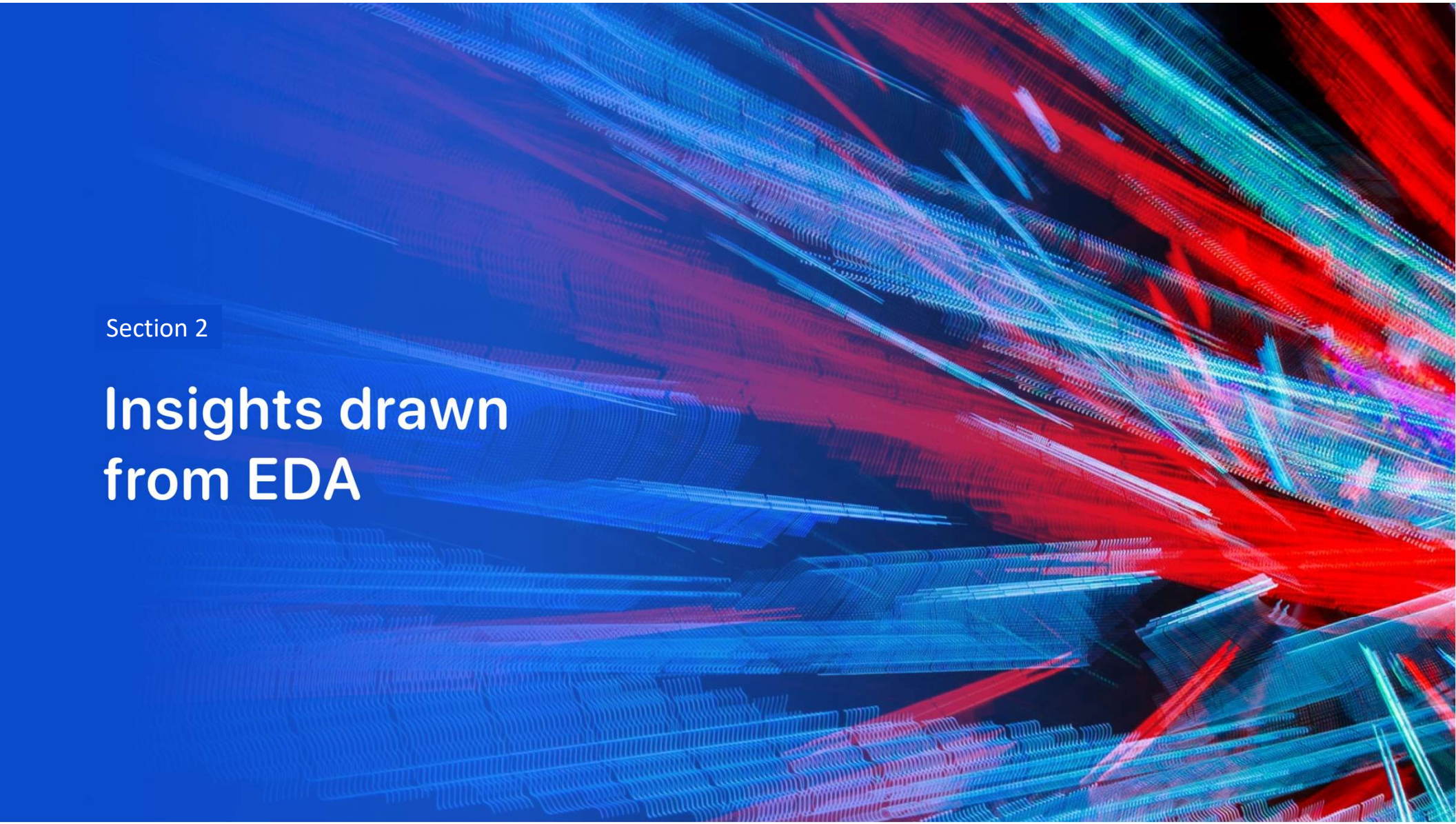
Results

- Interactive analytics demo in screenshots
  - All launch sites are very close proximity to the coast , railway or highway. They also keep certain distance away from cities



Results

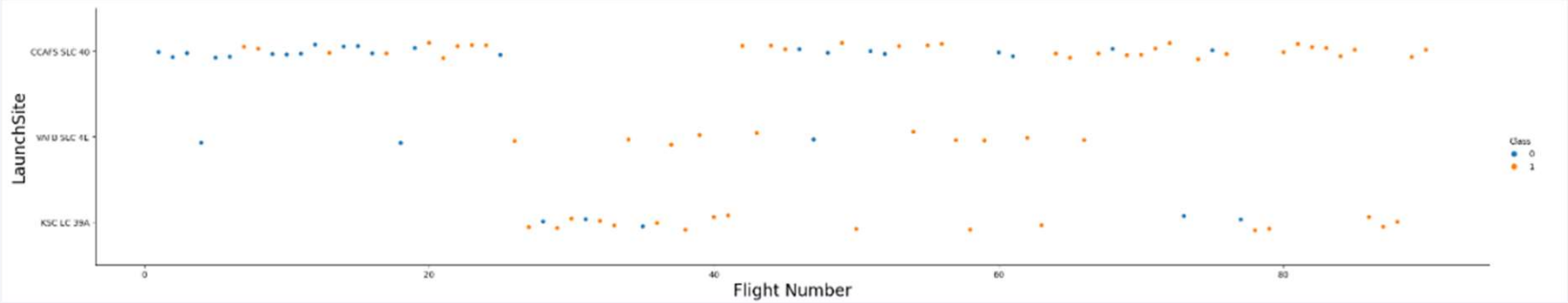
- Predictive analysis results
  - Decision tree model provide the highest accuracy on training set.
  - Four models provide similar accuracy on testing set.



Section 2

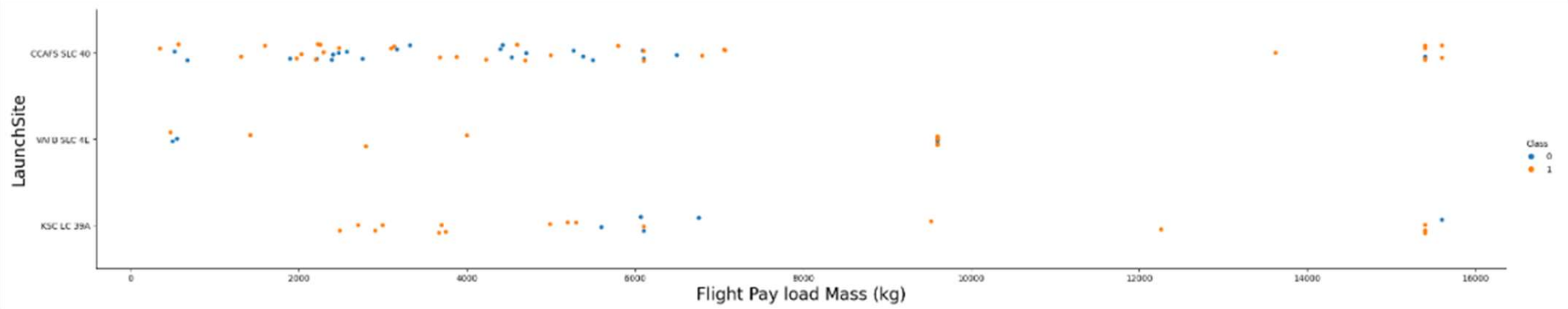
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAF5 SLC 40 site have the most launches. We can see it has consecutively successful launches recently compared to the first several times.
- VAFB SLC 4E has the second most launches and KSC LC 39A has the least.

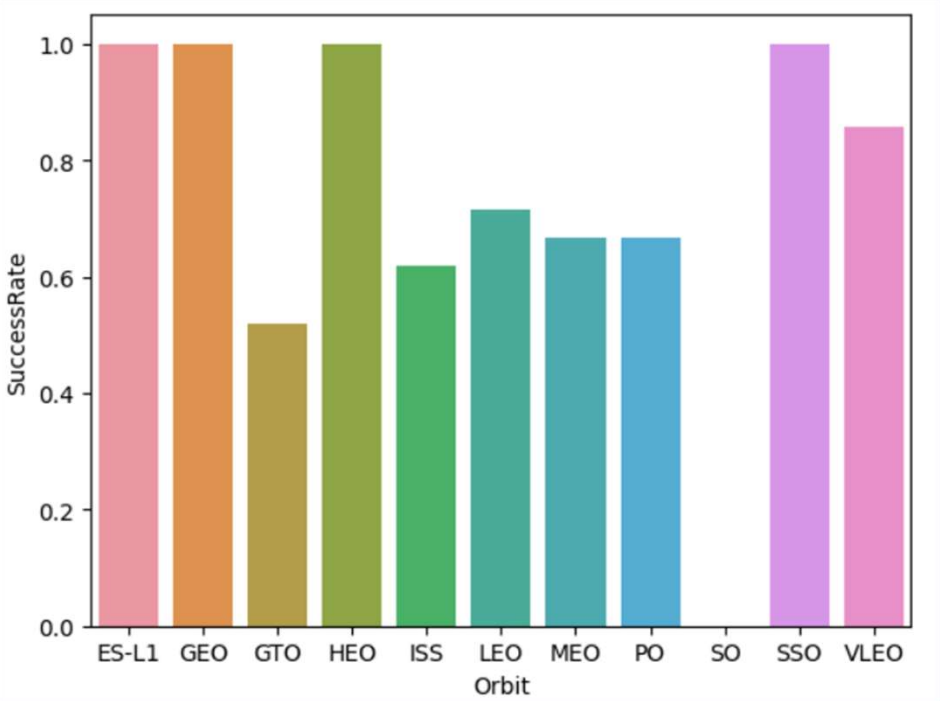
Payload vs. Launch Site



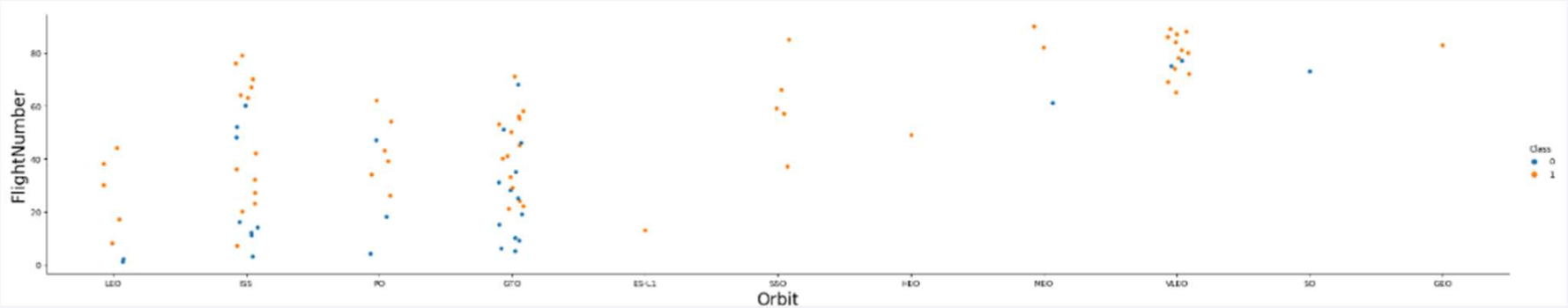
- We can see the high successful rate with flight payload over 8000kg.
- VAFB SLC 4E has no launch mission with flight payload over 10000kg.

Success Rate vs. Orbit Type

- Four orbits with 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO

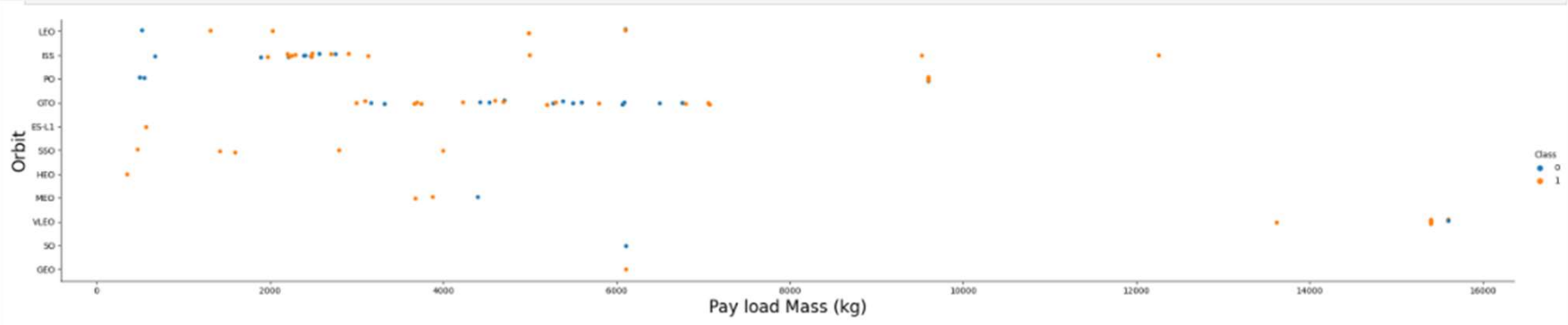


Flight Number vs. Orbit Type



- ISS and GTO have the most flight numbers in the beginning 60 flights
- VLEO replaces the place after the 60 flights
- No matter which orbits the success rate is getting higher in recent flights

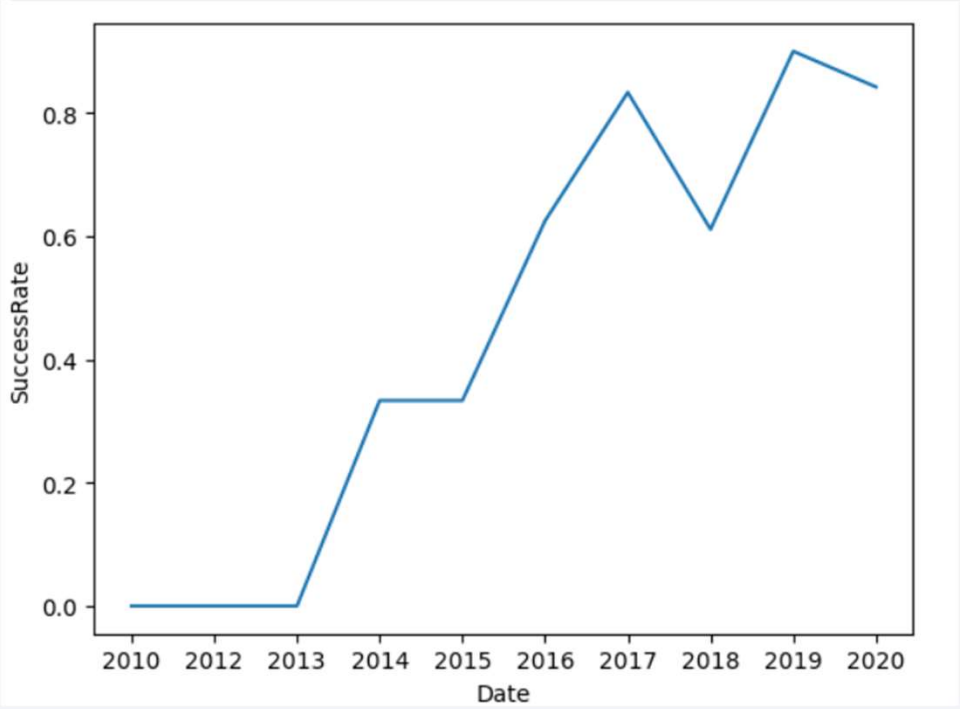
Payload vs. Orbit Type



- Among different payload setting, success rate seems to have no relationship with it.
- ISS and GTO orbit have the widest range of payload

Launch Success Yearly Trend

- The success rate has a sharp increase since 2013



All Launch Site Names

- There are 4 launch sites whose name are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
- We select distinct values of “launch_site” from data.

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

Launch Site Names 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

Total Payload Mass

- Total payload carried by boosters from NASA are 45596 kg
- We sum up the payload which carried by “customer” equal to “NASA (CRS)” from data.

TOTAL_PAYLOAD_MASS
45596.0

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4
- We get average of payload which carried by “Booster_Version” equal to “F9 v1.1”

AVERAGE_PAYLOAD_MASS
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 01/08/2018
- We get the minimum value of “Date” with “Landing_Outcome” equal to “Success (ground pad) from data.

min(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are as below.

- We get the booster version with 2 filtering conditions which are "Landing_Outcome" equal to "Success (drone ship)" and "PAYLOAD_MASS_KG" greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are as below
- We count the occurrences of all possible “Mission_Outcome” using “group by” function in SQL.

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are as shown below

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

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are shown below
- We get booster versions and launch site name with 2 filtering conditions which are “Landing_Outcome” equal to “Failure (drone ship)” and launch date’s year equal to 2015

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

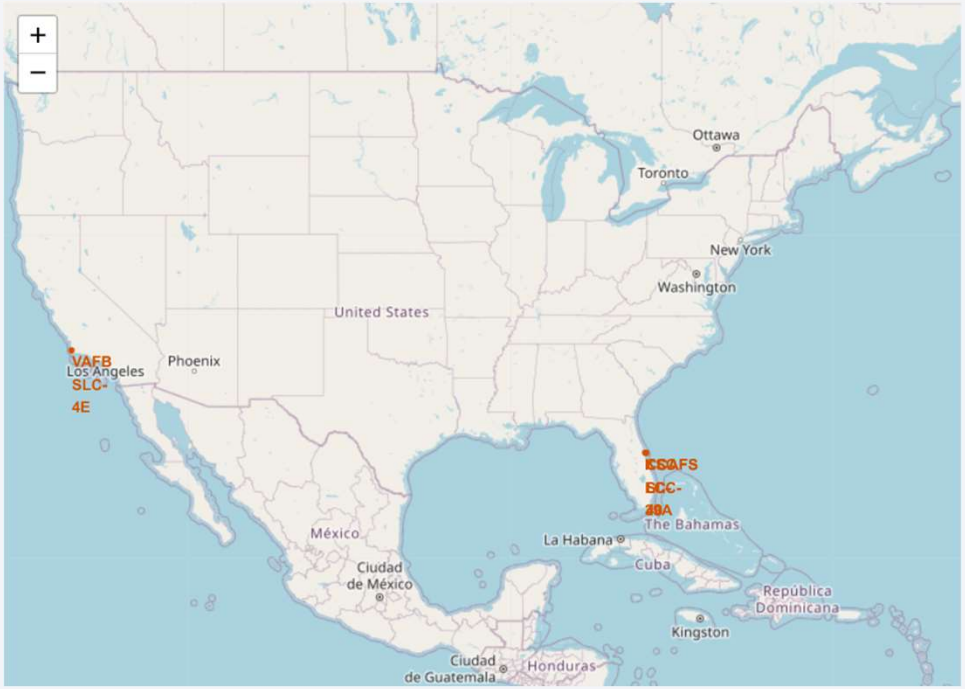
Landing_Outcome	count(Landing_Outcome)
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2

- We count the occurrences of possible “Landing Outcome” with launch date between 2010-06-04 and 2017-03-20 and show the result by the numbers of occurrences in descending order.

Section 3

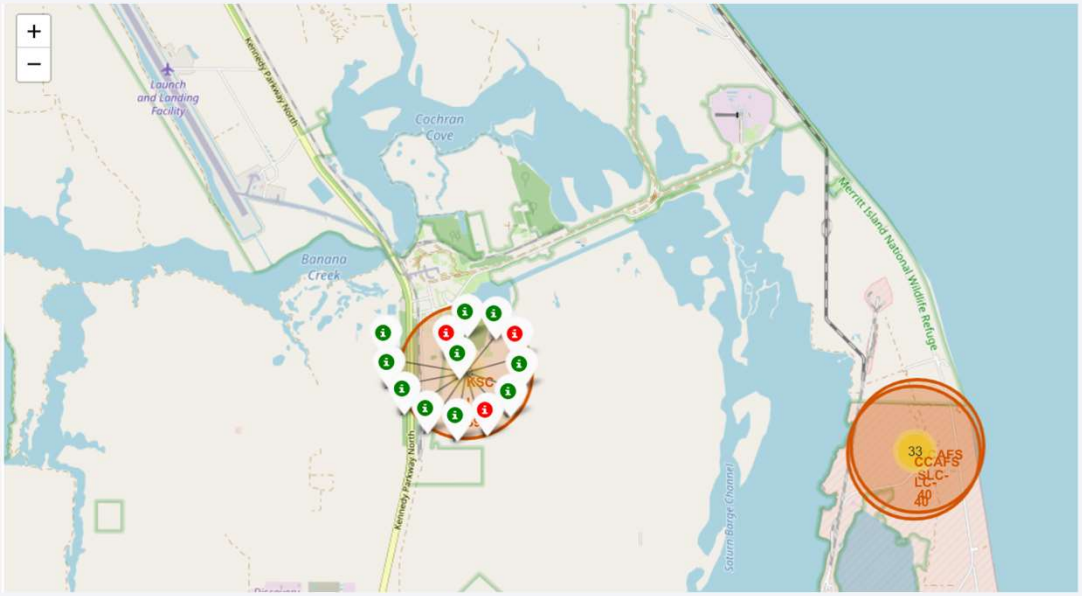
Launch Sites Proximities Analysis

All launch sites



- From above we can see the sites are very close to sea or railroads

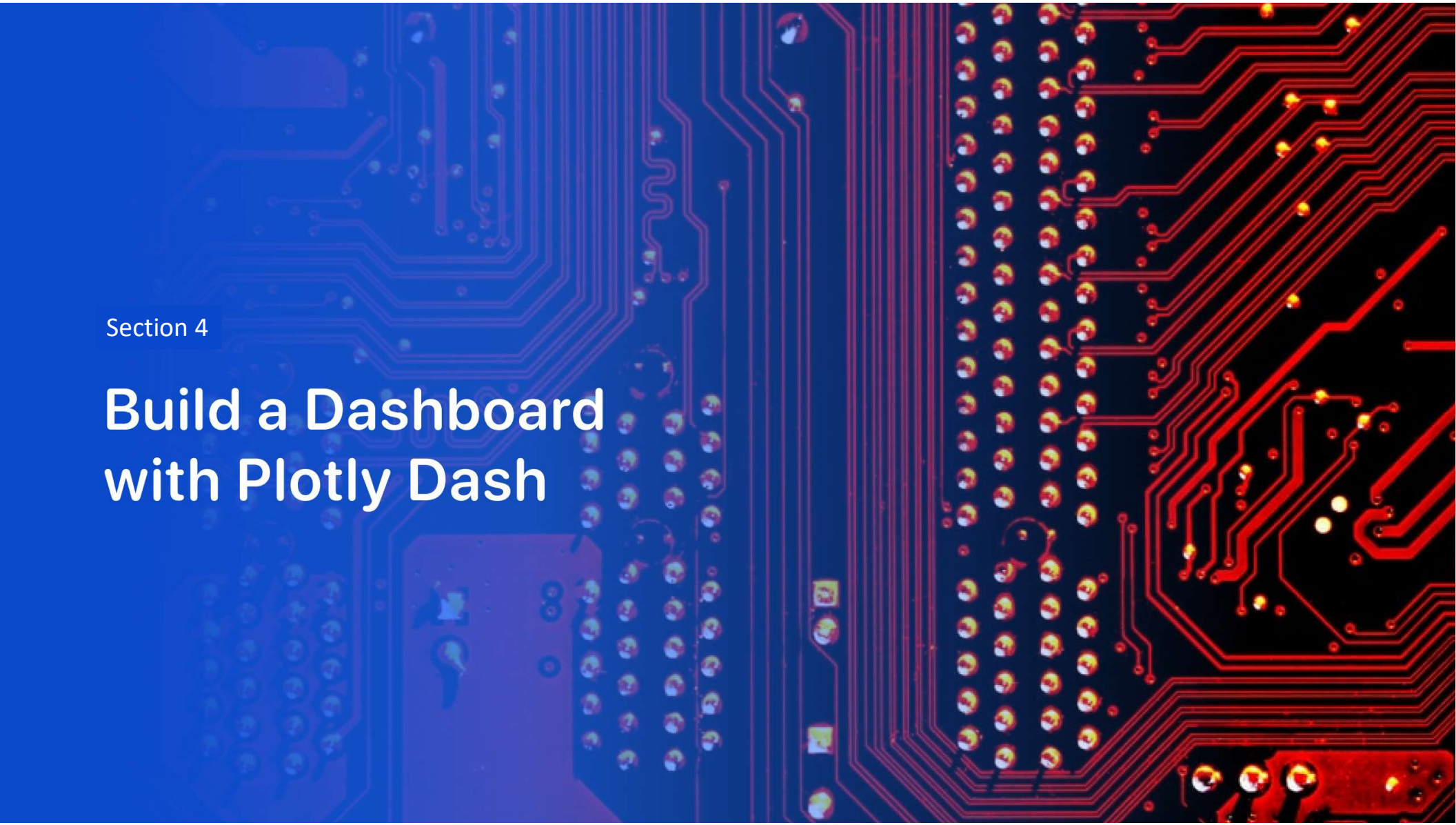
Launch outcomes of KSC LC-39A



- We can see 10 successful outcomes with green markers and 3 failure outcomes with red markers.

Proximities to CCAFS SLC-40

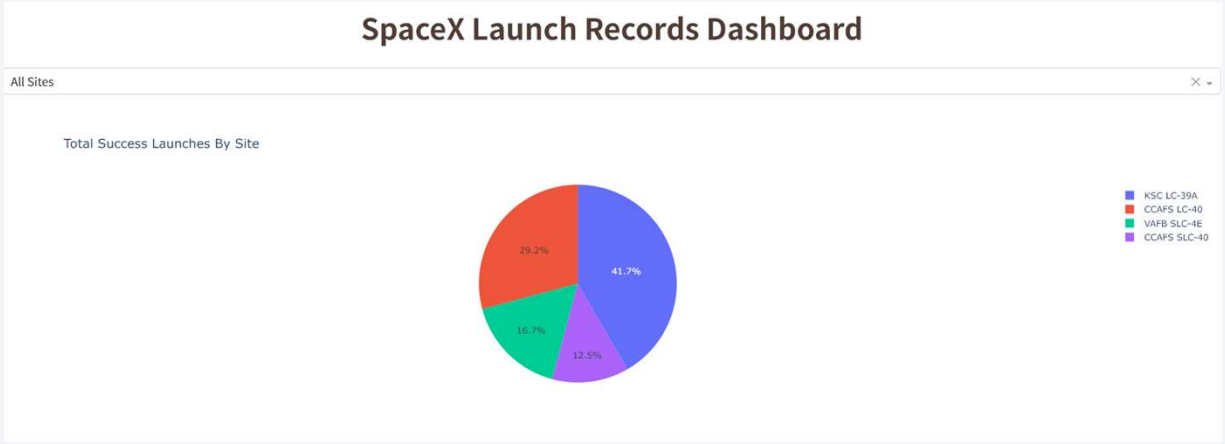
- CCAFS SLC-40 is very close to coastline and far from the big cities for safety.



Section 4

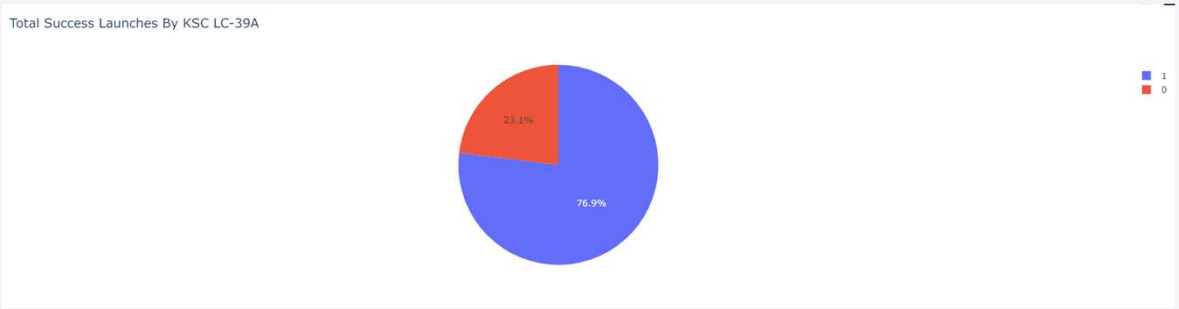
Build a Dashboard with Plotly Dash

SpaceX Launch Records



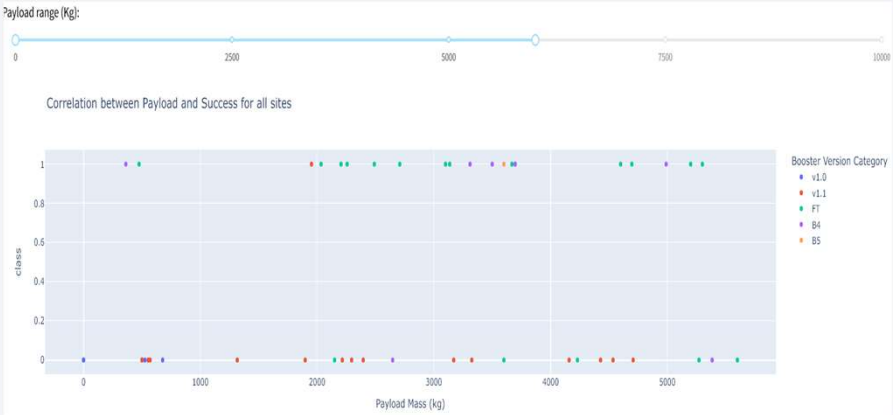
- KSC LC-39A and CCAFS LC-40 covers over 70% of successful launches.

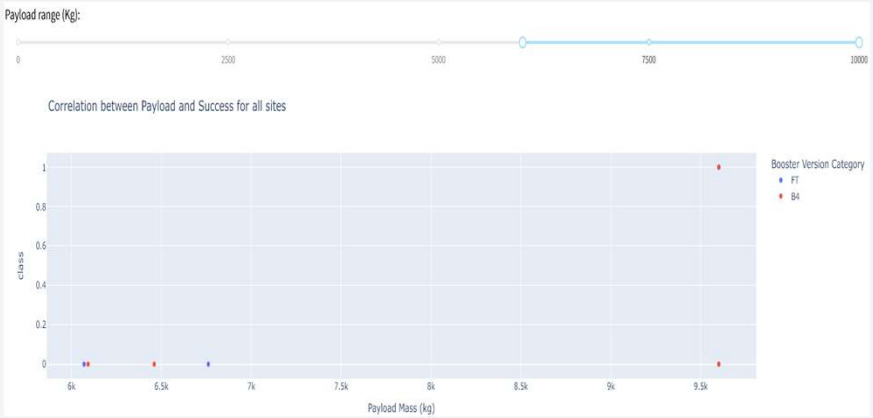
Success rate of KSC LC-39A



- We can see KSC LC-39A has high success rate up to 76.9%

Correlation between payload and success rate





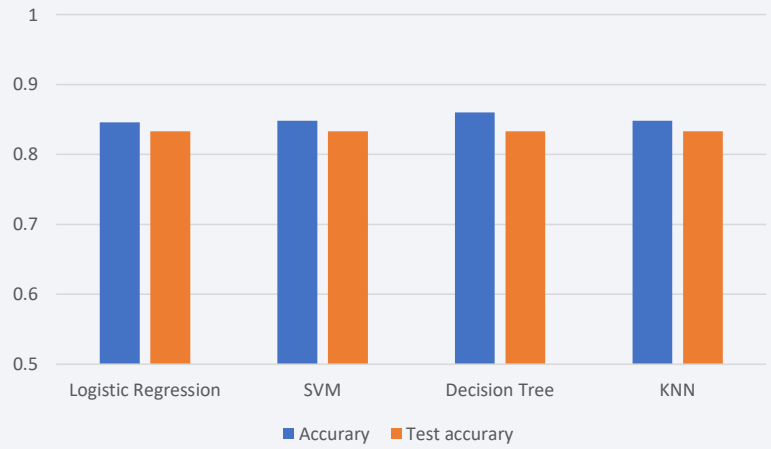
- From above scatter plots, we can easily identify FT booster version provides the most successful launches with payload under 6,000kg
- With payload above 6,000kg, only one launch was successful. It's hard to get conclusive result to infer which booster version is better.

Section 5

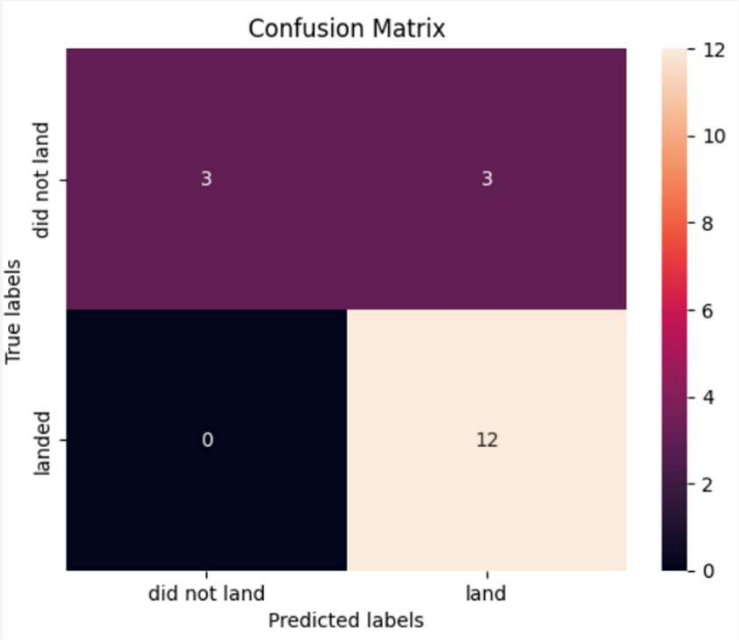
Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree model provides the highest accuracy of 86%.
- Four models get similar test accuracy of 83%



Confusion Matrix



- Decision tree model can predict the launch landed with high precision.

Conclusions

- The successful landing outcomes become more and more since 2013
- KSC LC-39A is the best launch site because of its highest success rate
- The flights with payload over 8,000kg have high successful rate
- Decision Tree model is good classifier to predict if the landing is successful

Thank you!