IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AJNF

May 26, 2022

Outline

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

Executive Summary

- ## Summary of methodologies

- Data collection methodology:

    - The data is collected by the use of requests.get() function in python, decoded and turned into a Pandas dataframe with the missing values replaced by the mean. The dataset is filtered from the collection is processed by classifying their numerical outcome labels from their actual outcomes. Various kinds of plots and graphs are used to gain additional insight on the data. SQL queries were done for the querying of the data set. Objects are added to provide insights on the location of the launch sites and how they are distributed in the map. The different models to be considered are the K nearest neighbors method, the Decision Tree method, the Support Vector Machine method, and the Logistics Regression methods

- ## Summary of all results

    - From all the models that are trained and tested, each one can be used according to the model's strength. For this case, since we are predicting the outcome of a successful land, logistic regression will work best when analyzing larger data sets. The case at hand is influenced by different independent variables which could involve more complex models for better predictions. Launch sites are placed at locations where there are railways and highways for material transport, near the coastal areas to access the oceans for possible landings, and away from the cities for public safety. The KSC L-39A station has the highest count of successful landings.

Introduction

- Project background and context
  - SpaceX is known for effectively landing rockets back from lift-off, causing great efficiency in material, time, and costs. To be a viable competitor, SpaceY must know key details on how SpaceX does this remarkable feat through data science.
- Problems you want to find answers
  - Essentially, SpaceY must know what makes the rocket lands fail and most importantly, succeed.
  - SpaceY must consider the different factors that revolve around rocket landing success in order to replicate it.
  - Establish new knowledge on other possible correlations among variables to be considered.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - The data is collected by the use of requests.get() function in python, decoded and turned into a Pandas dataframe with the missing values replaced by the mean.
- Perform data wrangling
    - The dataset filtered from the collection is processed by classifying their numerical outcome labels from their actual outcomes.
- Perform exploratory data analysis (EDA) using visualization and SQL
    - Various kinds of plots and graphs are used to gain additional insight on the data. SQL queries were done for the querying of the data set
- Perform interactive visual analytics using Folium and Plotly Dash
    - Objects are added to provide insights on the location of the launch sites and how they are distributed in the map.
- Perform predictive analysis using classification models
    - The different models to be considered are the K nearest neighbors method, the Decision Tree method, the Support Vector Machine method, and the Logistics Regression method

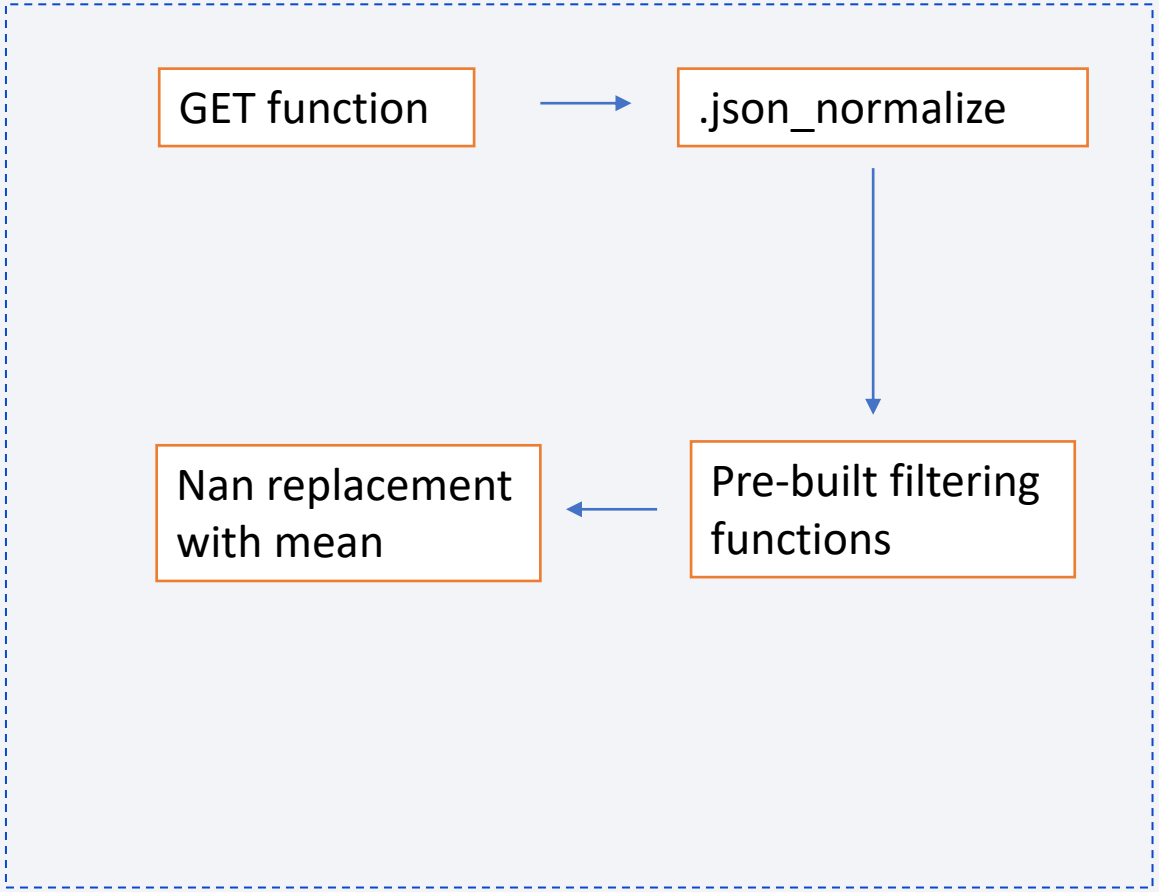
Data Collection

- The data sets are collected using the requests.get() function from python, extracting key information from website of SpaceX and their launches.
- Because the data is in json, it was decoded and turned into a Pandas dataframe using .json_normalize.
- Using the pre-built functions, the features that SpaceY wants to observe are prioritized and extracted from the large dataset, simplifying the data further.
- Missing values are also dealt with by replacing it with the mean values of the column it is under.
- Falcon9 launches are priority data sets, which can be cross referenced and extracted from another website, Wikipedia.

Data Collection – SpaceX API

- Shown in the right is the flowchart done for the data collection process.

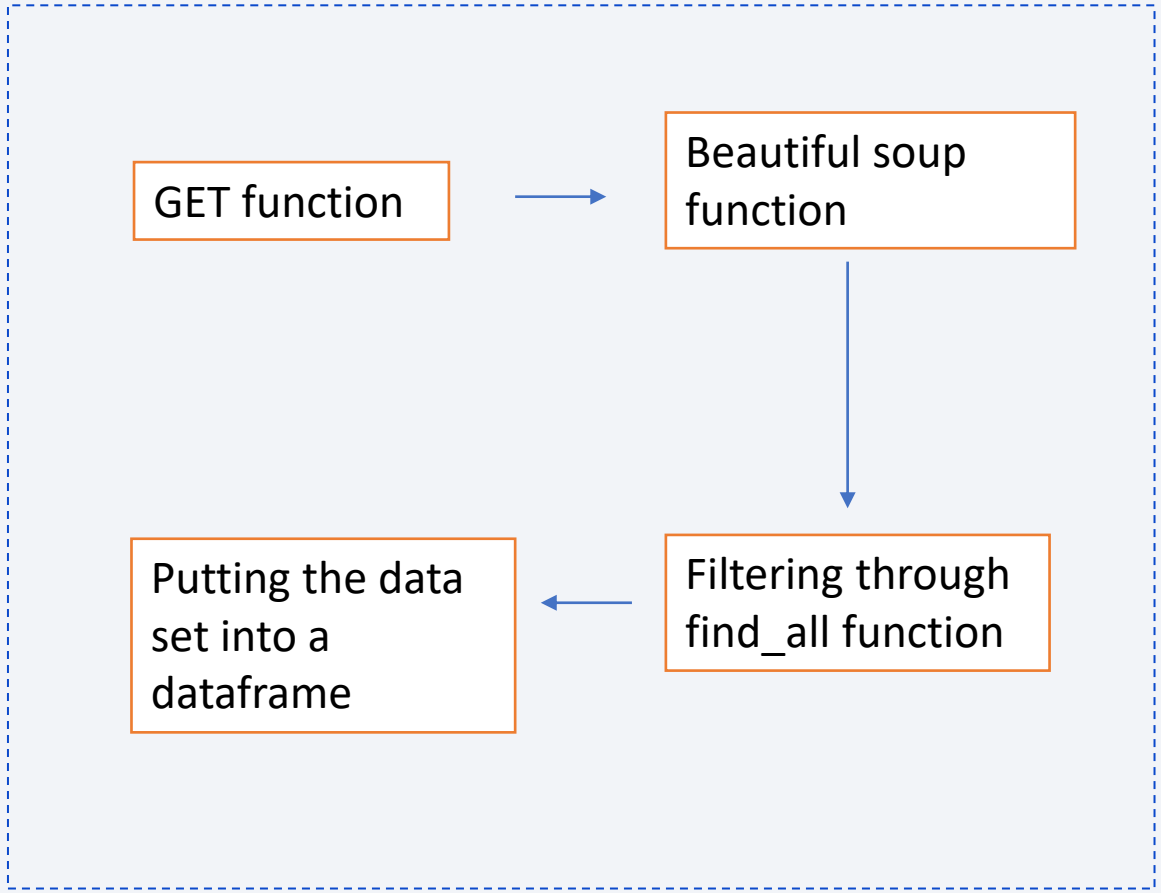
- Github url: https://github.com/Allevium/IBM-Data-Science/blob/main/C10/jupyter-labs-spacex-data-collection-api_1.ipynb



Data Collection - Scraping

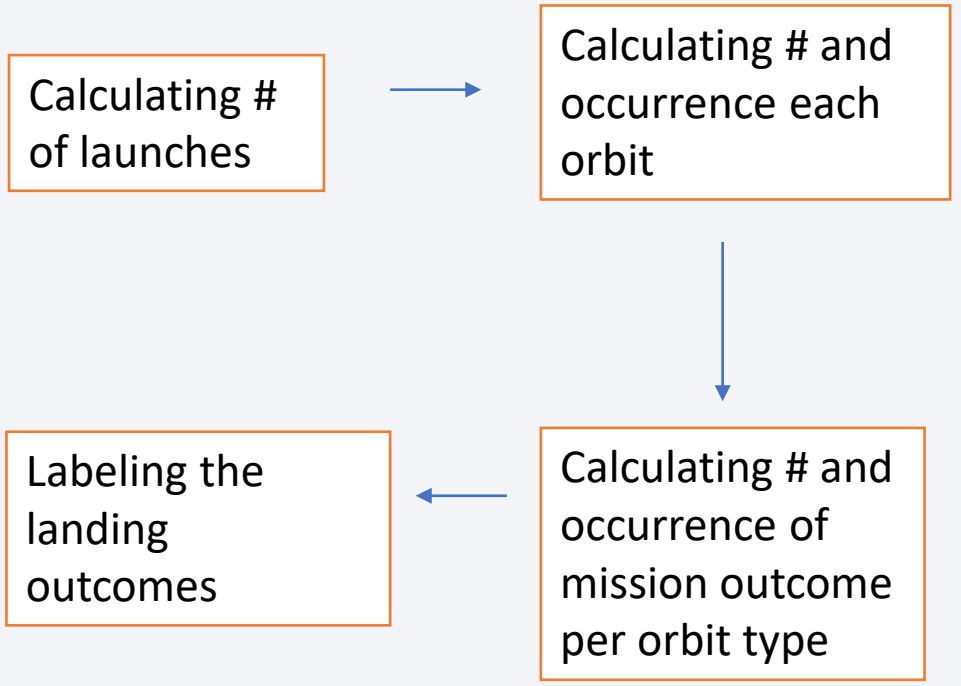
- Shown in the right is the flowchart involving the webscraping process done to extract data from Wikipedia.

- Github url: https://github.com/Allevium/IBM-Data-Science/blob/main/C10/jupyter-labs-webscraping_2.ipynb



Data Wrangling

- The dataset filtered from the collection is processed by classifying their numerical outcome labels from their actual outcomes.
- Actual outcomes are influenced by the data we have on the number of launches, occurrence of each orbit, and the occurrence of mission outcome per orbit type.
- This is done as strings are not recognized by machine learning models to be applied to the data later on.



EDA with Data Visualization

- Scatter plots were generated to provide insight on the possible inferential relationship between the x and varying y variables. A bar graph is also generated to compare the success rates of missions in between orbit types. Scatter plots are utilized again to visualize the relationships varying independent variables and the orbit type. Finally, the yearly trend of launch success is shown in a line graph.
- GitHub URL: https://github.com/Allevium/IBM-Data-Science/blob/main/C10/jupyter-labs-eda-dataviz_5.ipynb

EDA with SQL

- The SQL queries that were done for the querying of the data set are the following:
  - SELECT UNIQUE(launch_site) from SPACEXTBL; (returns unique launch sites)
  - SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5; (returns launch sites which begin with ‘CCA’)
  - SELECT SUM (PAYLOAD_MASS__KG_) AS PAYLOADMASS FROM SPACEXTBL; (returns total payload mass carried by boosters launched by NASA (CRS))
  - SELECT AVG(PAYLOAD_MASS__KG_) AS PAYLOADMASS FROM SPACEXTBL; (returns average payload mass carried by booster version F9 v1.1)
  - SELECT MIN(DATE) FROM SPACEXTBL; (returns when the first successful landing outcome in ground pad was achieved)

EDA with SQL

- The SQL queries that were done for the querying of the data set are the following:
  - SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME='SUCCESS (DRONE SHIP)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000; (returns names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000)
  - SELECT COUNT(MISSION_OUTCOME) AS MISSIONOUTCOMES FROM SPACEXTBL GROUP BY MISSION_OUTCOME; (returns the total number of successful and failure mission outcomes)
  - SELECT BOOSTER_VERSION AS BOOSTERVERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL); (returns names of the booster_versions which have carried the maximum payload mass)
  - SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015'; (returns the list of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC; (returns the 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
- GitHub URL: https://github.com/Allevium/IBM-Data-Science/blob/main/C10/jupyter-labs-eda-sql-coursera_4.ipynb

Build an Interactive Map with Folium

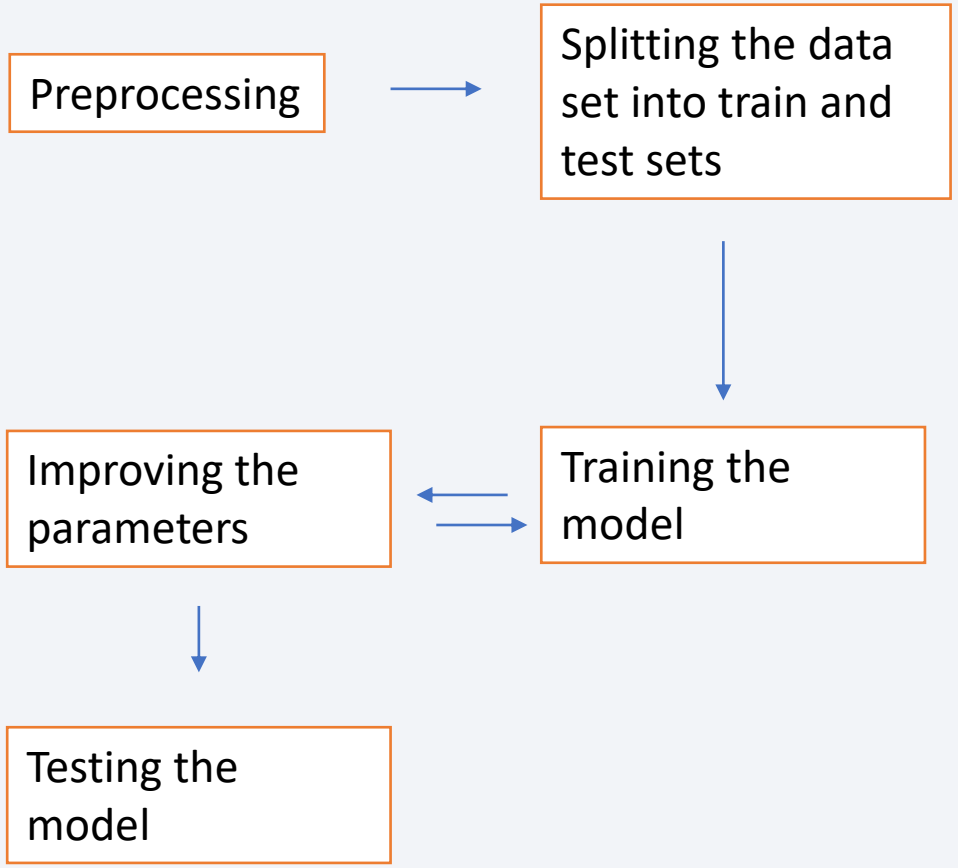
- Circles are used to represent a specific coordinates of launch sites, markers are used to label these circles for quick reading. Marker clusters are utilized to indicate if the launch conducted was a success or not for each launch site. Finally, lines are used to define the proximity of the launch sites to different locations such as railways, highways, and coastlines.
- These objects are added to provide insights on the location of the launch sites and how they are distributed in the map. Additionally, they could also help in choosing launch sites that are away from crowded and busy locations.
- GitHub URL: https://github.com/Allevium/IBM-Data-Science/blob/main/C10/lab_jupyter_launch_site_location_6.ipynb

Build a Dashboard with Plotly Dash

- Pie graphs are added to show the success counts of all launch sites, showing which launch site has the highest count of successful launches and the overall counts. Scatter plots are also added to show the success counts on payload mass for each site and all sites. Dropdowns are also added to enable interaction with the dashboard.
- These plots and interactions are added to make the dashboard user-friendly and interactive, since this showcases the data presentation in real-time, and pushes more users to use it.
- GitHub URL: https://github.com/Allevium/IBM-Data-Science/blob/main/C10/spacex_dash_app_7.py

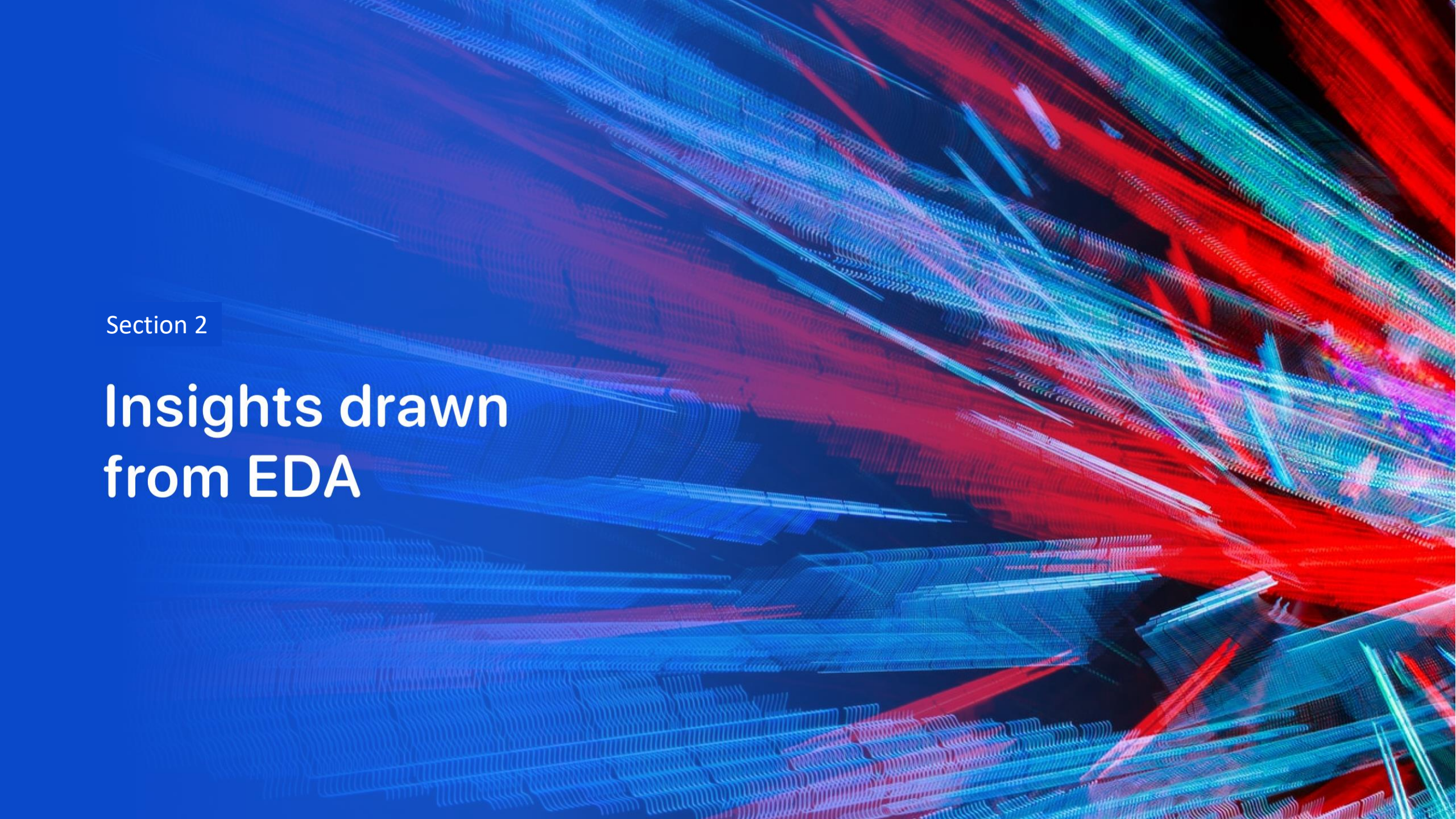
Predictive Analysis (Classification)

- The different models to be considered are the K nearest neighbors method, the Decision Tree method, the Support Vector Machine method, and the Logistics Regression method. Each one is built by splitting the data set into training and testing splits, which are used to train and test the model respectively. Each method has different parameters to consider but was improved by finding the best parameters available with the help of python. The best performing classification model was found to be any of them, since the accuracy tests such as the jaccard scores, f1 scores, and the confusion matrix all generated an accuracy of 0.8334 for all the methods.
- Shown in the right is the summarized flowchart process of the predictive analysis done.
- GitHub URL: https://github.com/Allevium/IBM-Data-Science/blob/main/C10/SpaceX_Machine%20Learning%20Prediction_Part_5_8.ipynb



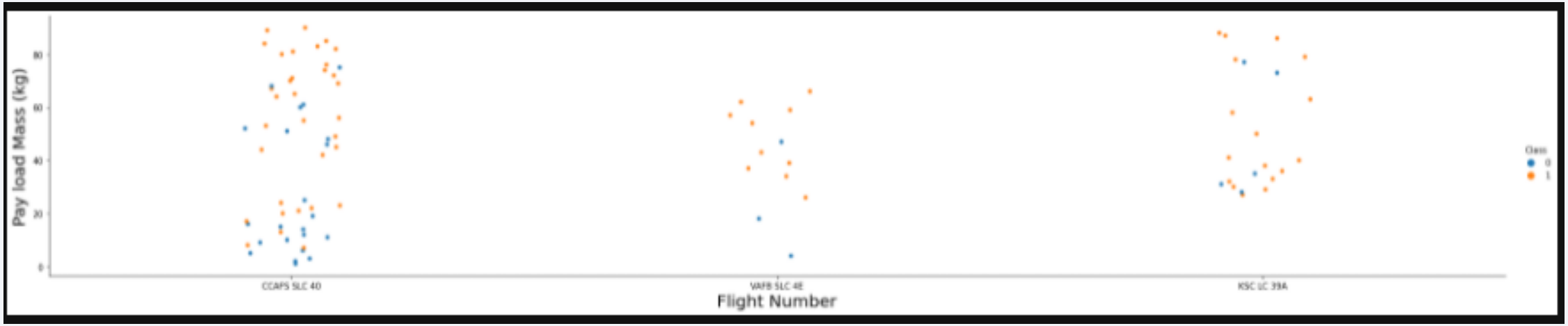
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

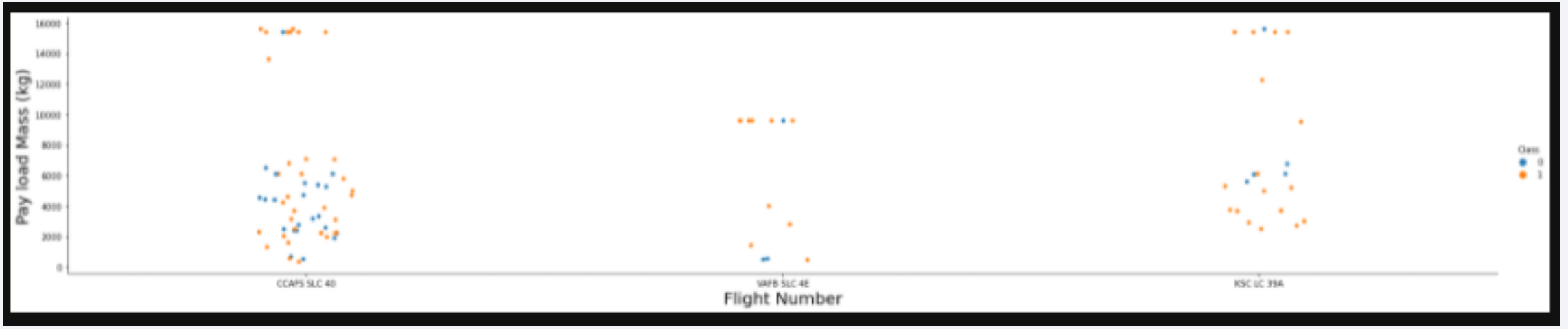
We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Flight Number vs. Launch Site



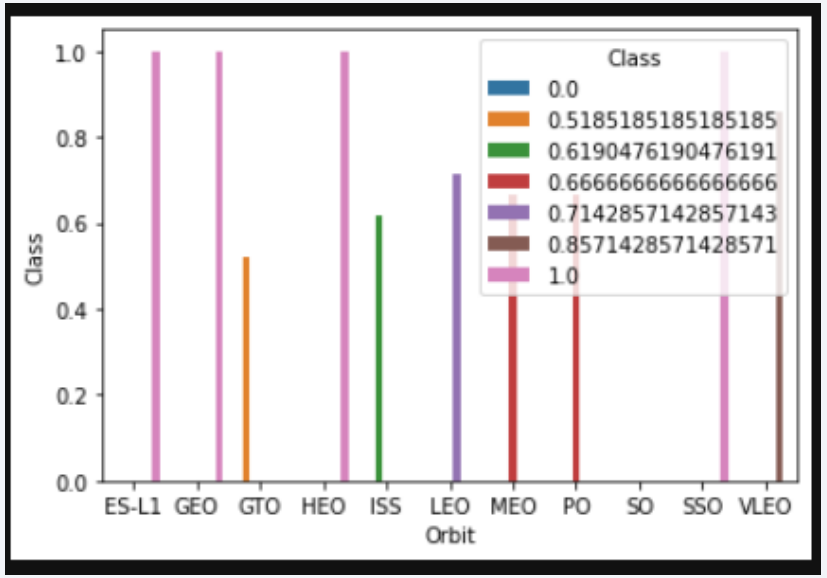
- From the plot, it can be observed that KSC LC-39A and VAFB SLC 4E both have relatively higher counts of orange dots, pertaining to successful outcomes. It seems that the the chance of successful outcomes increases, if it has an increasing flight number and a pay load mass greater than 20 kg. Generally, there is a high amount of pay load masses on lower flight numbers.

Payload vs. Launch Site



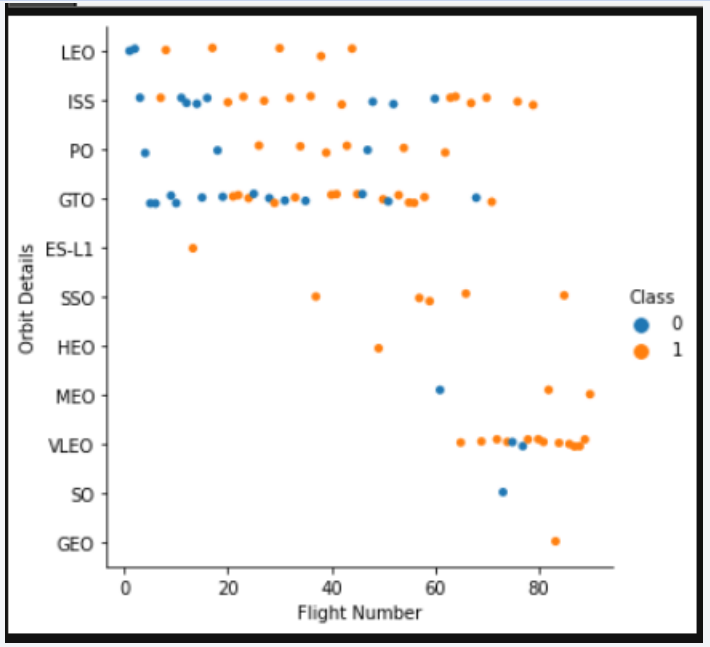
- From the plot, it can be observed that each lunch site has a higher success rate at their higher pay load masses. Differentiating the launch sites, the launch site CCAFS LC-40 has the greatest number of launches, followed by the KSC LC-39A, and lastly, VAFB SLC 4E. The maximum pay load masses of CCAFS LC-40 and VAFB SLC 4E are the same.

Success Rate vs. Orbit Type



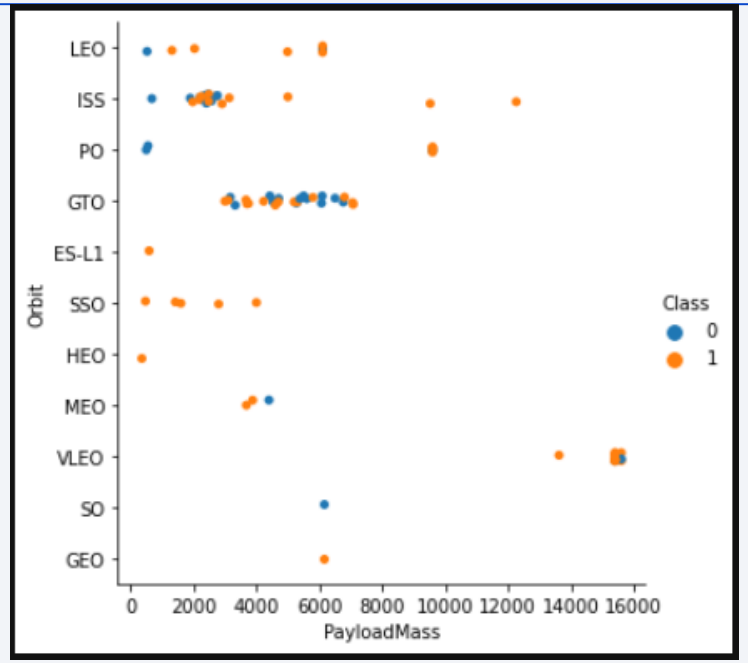
- From the bar graph, it can be observed that the pink-colored bars are the launch sites that have the highest success rates which are the ES-L1, GEO, HEO, and SSO . Meanwhile, the lowest is the SO launch site with flat 0 success rate.

Flight Number vs. Orbit Type



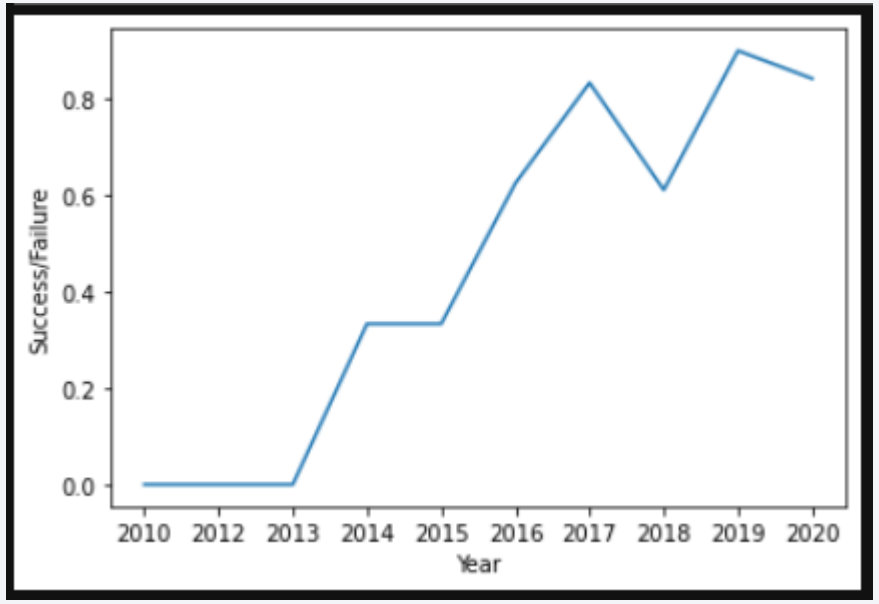
- From the plot, it can be observed that there seems to be an increasing trend in success rate as flight number increases in the general sense except for the GTO orbit. The GTO and ISS orbits have the highest counts of failure dots in lower flight numbers.

Payload vs. Orbit Type



- From the plot, it can be observed that there seems to be an increasing trend in success rate as the payload mass increases in the general sense except for the GTO orbit again.

Launch Success Yearly Trend



- From the line graph, it can be observed there seems to be an increasing trend in success rate from 2010 to an all time high of 0.8 on 2017, followed by a sharp decrease to 0.6 2 years afterwards. Another all time high of a success rate of approximately 0.9 during 2019 is recorded followed by a gentle decrease towards 2020.

All Launch Site Names

- Find the names of the unique launch sites
  - SELECT UNIQUE(launch_site) from SPACEXTBL;
- Present your query result with a short explanation here
  - The query above returns unique launch sites.

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

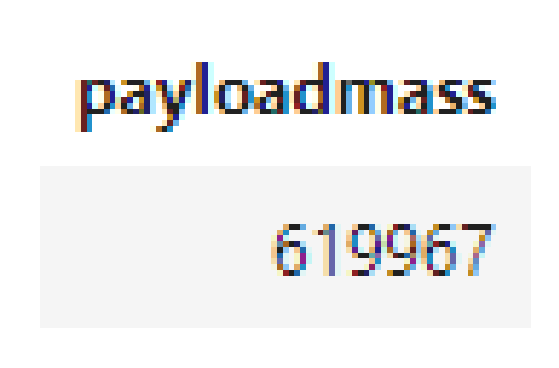
Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
  - SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5; (returns launch sites which begin with ‘CCA’)
- Present your query result with a short explanation here
  - The query above returns launch sites which begin with ‘CCA’

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - SELECT SUM (PAYLOAD_MASS__KG_) AS PAYLOADMASS FROM SPACEXTBL;
- Present your query result with a short explanation here
  - The query above returns total payload mass carried by boosters launched by NASA (CRS)



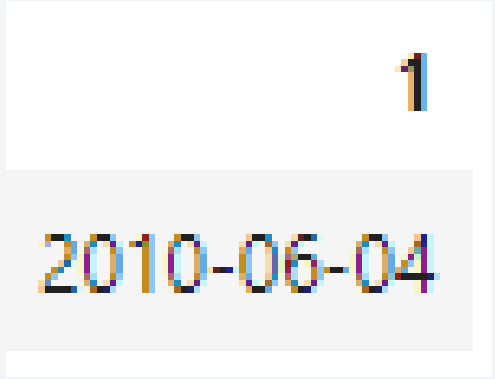
Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - SELECT AVG(PAYLOAD_MASS__KG_) AS PAYLOADMASS FROM SPACEXTBL;
- Present your query result with a short explanation here
  - The query above returns average payload mass carried by booster version F9 v1.1

payloadmass
6138

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - SELECT MIN(DATE) FROM SPACEXTBL;
- Present your query result with a short explanation here
  - The query above returns the basically the minimum date from the data set.



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
  - SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME='SUCCESS (DRONE SHIP)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
- Present your query result with a short explanation here
  - The query above returns 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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - SELECT COUNT(MISSION_OUTCOME) AS MISSIONOUTCOMES FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
- Present your query result with a short explanation here
  - The query above returns total number of successful and failure mission outcomes

missionoutcomes
1
99
1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - SELECT BOOSTER_VERSION AS BOOSTERVERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
- Present your query result with a short explanation here
  - The query above returns the names of the booster which have carried the maximum payload mass

boosterversion
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

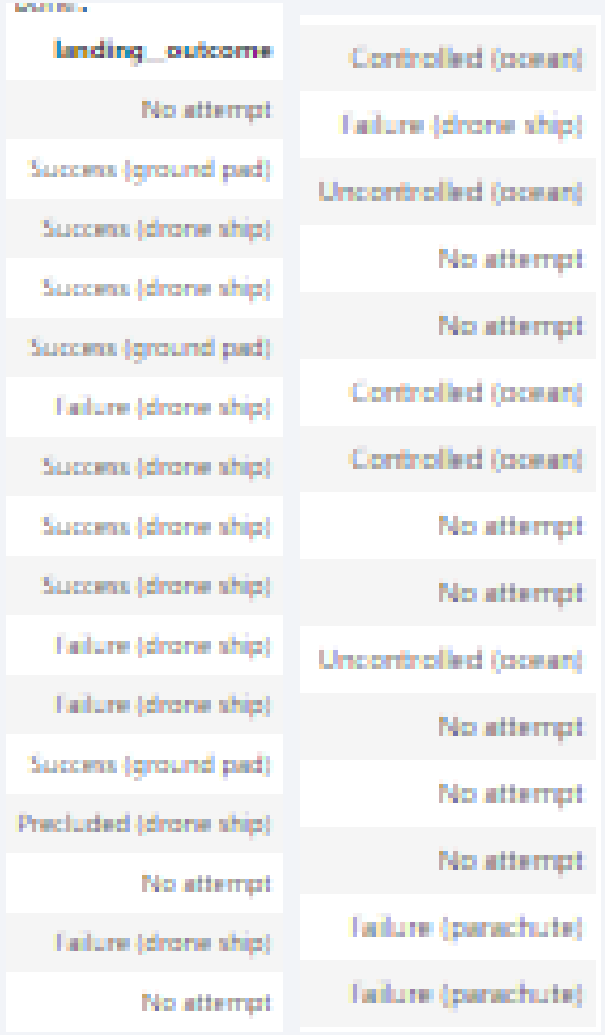
2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015';
- Present your query result with a short explanation here
  - The query above returns failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	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
  - SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
- Present your query result with a short explanation here
  - The query above returns the 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
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt
Failure (drone ship)
No attempt

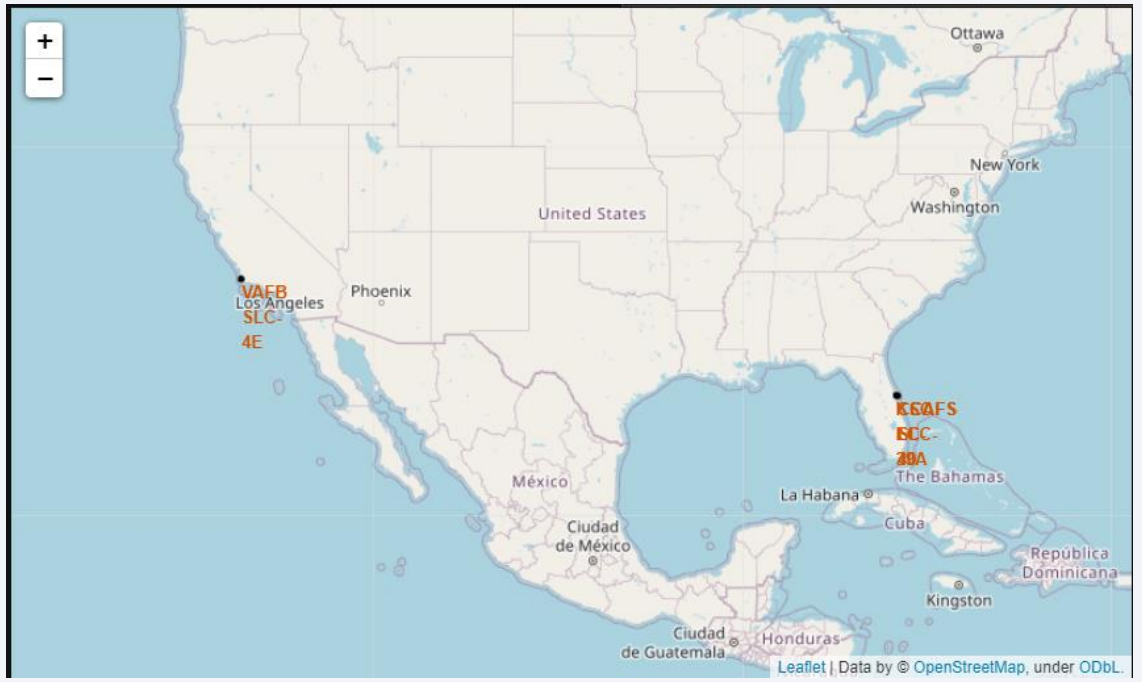
Controlled (ocean)
Failure (drone ship)
Uncontrolled (ocean)
No attempt
No attempt
Controlled (ocean)
Controlled (ocean)
No attempt
No attempt
Uncontrolled (ocean)
No attempt
No attempt
No attempt
Failure (parachute)
Failure (parachute)

Section 3

Launch Sites Proximities Analysis

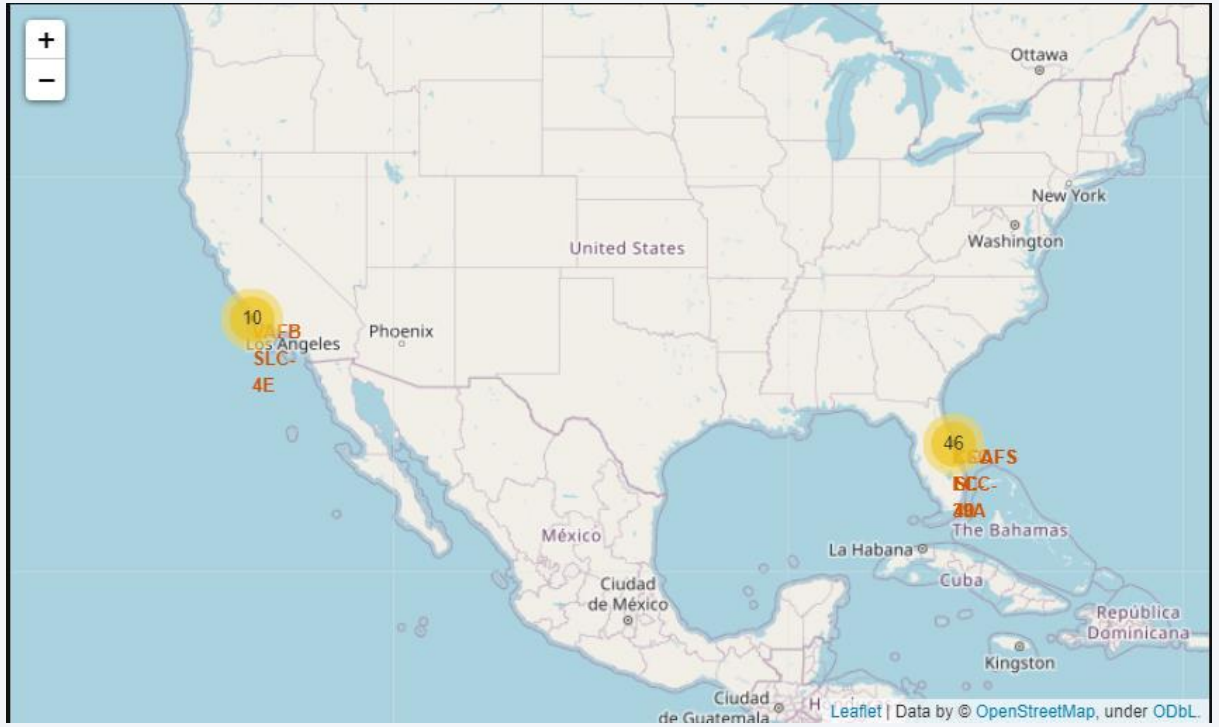
Launch Site Map v.1

- There seems to be two overlapping launch sites due to the zoom, but are separable once zoomed in.
- All of the launch sites seem to be at the coastal areas.



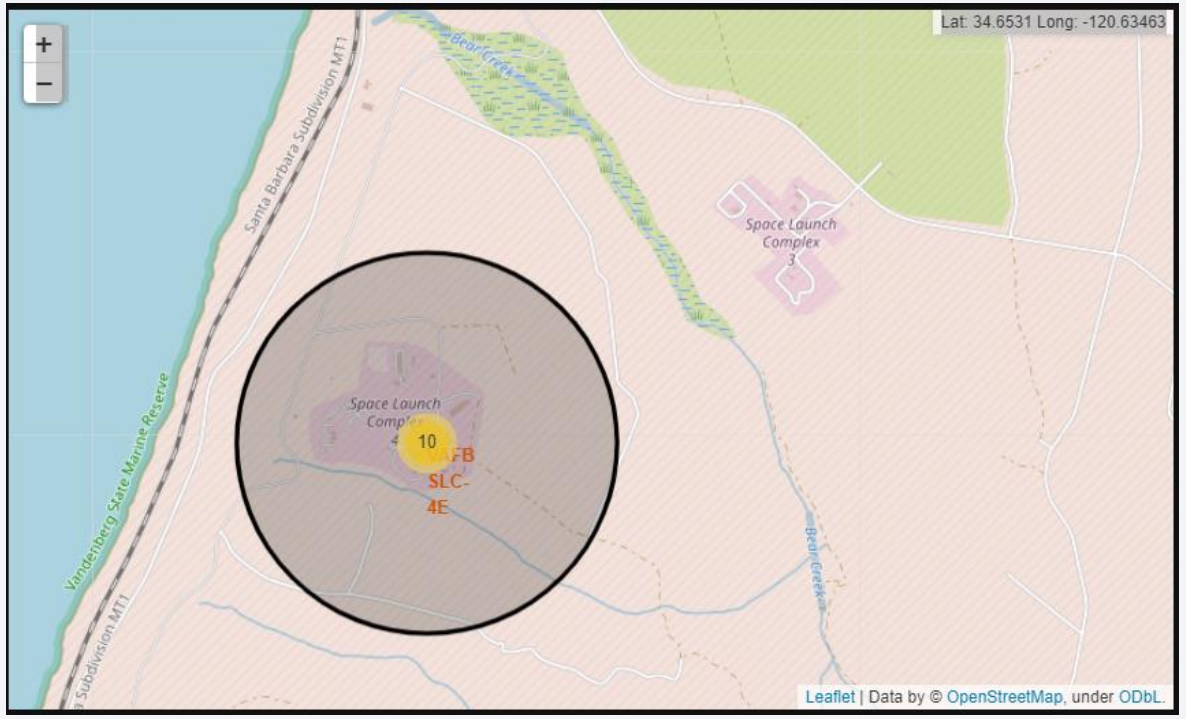
Launch Site Map with launch outcomes

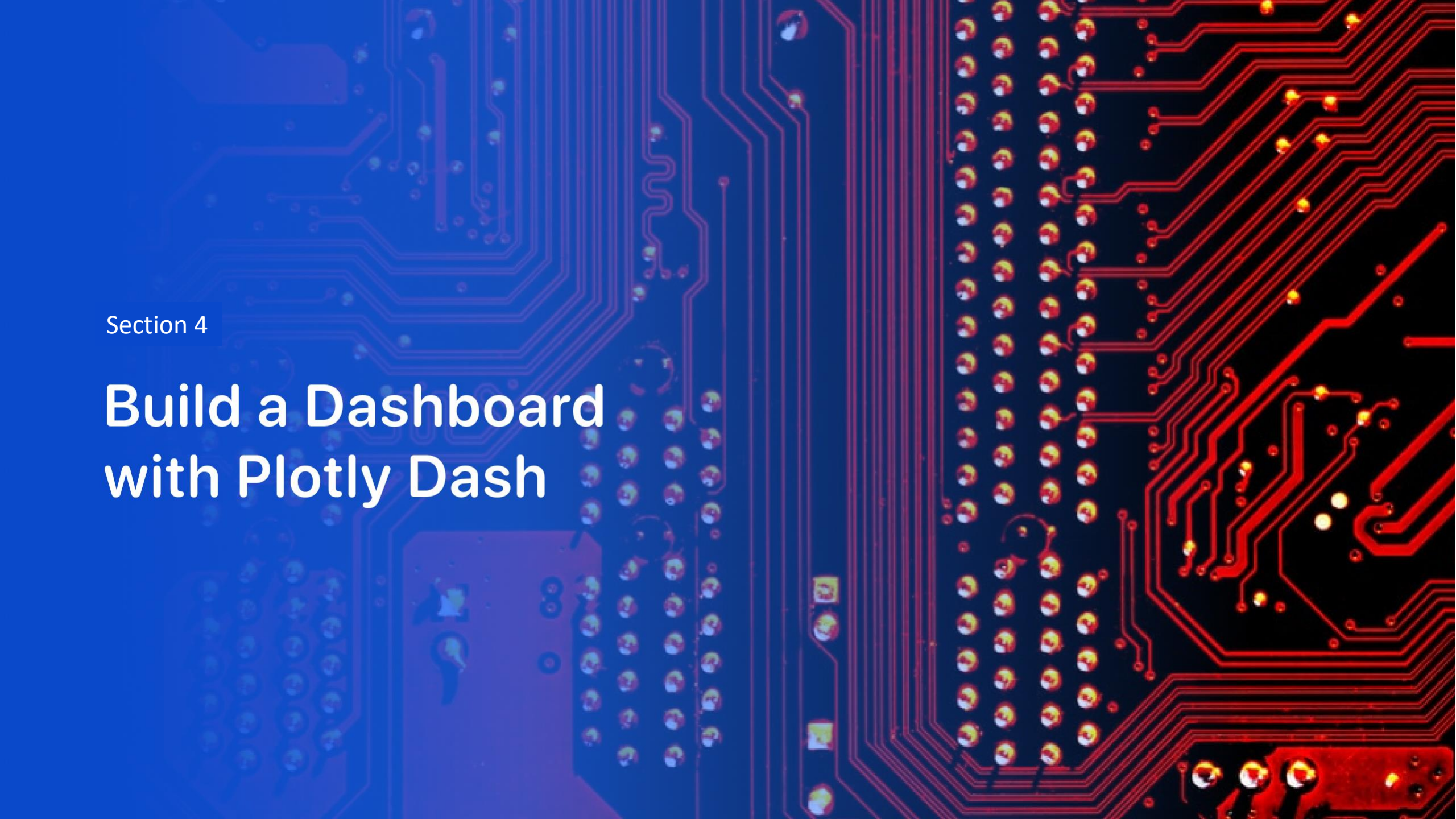
- The outcomes of the two close stations are aggregated, but taking the average of the success counts, it is still higher than the site situated on the left side of the map.
- Cluster markers are an effective tool for aggregating records for labeling.



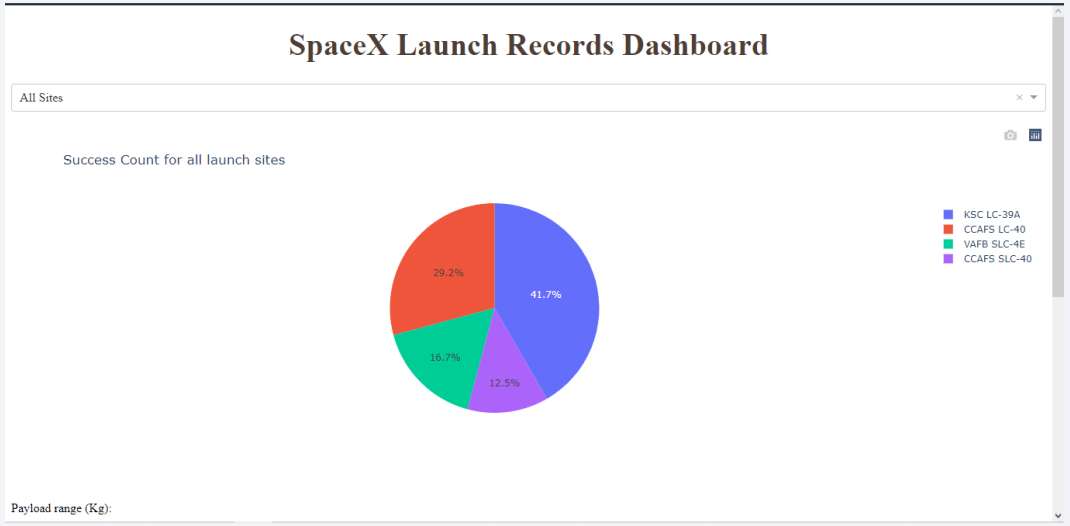
Launch Site Proximity Map

- Generally, the launch sites are in close proximity to coastlines, allowing access to the sea for possible landings there.
- The sites are also close to railways and highways, so that material transport is smooth.
- The sites are far from busy and crowded areas such as cities.



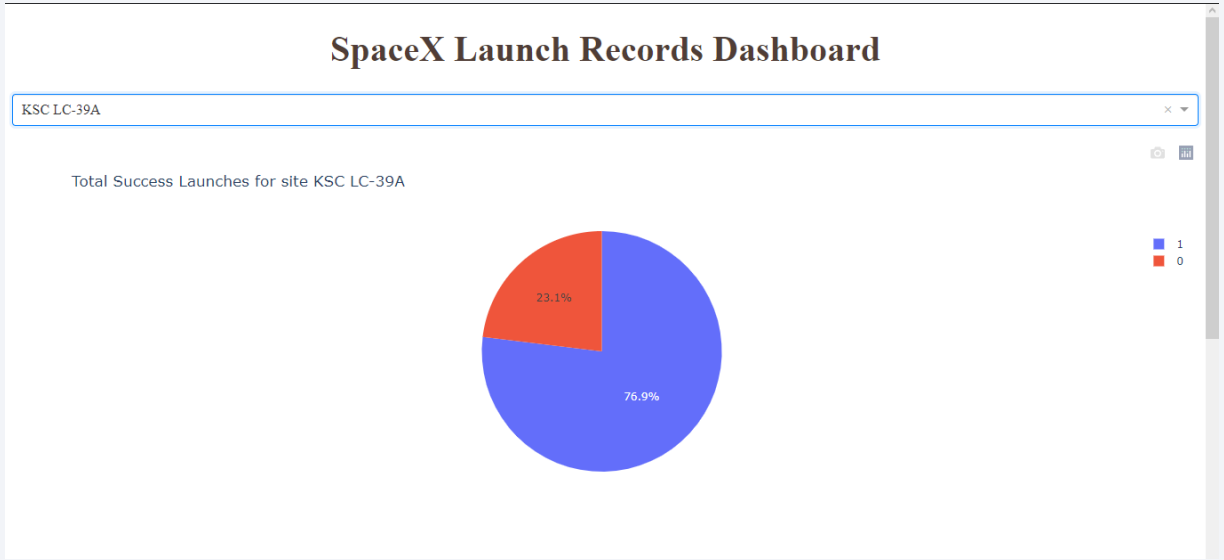

Section 4
Build a Dashboard with Plotly Dash

Launch Overall Success Count (Pie)



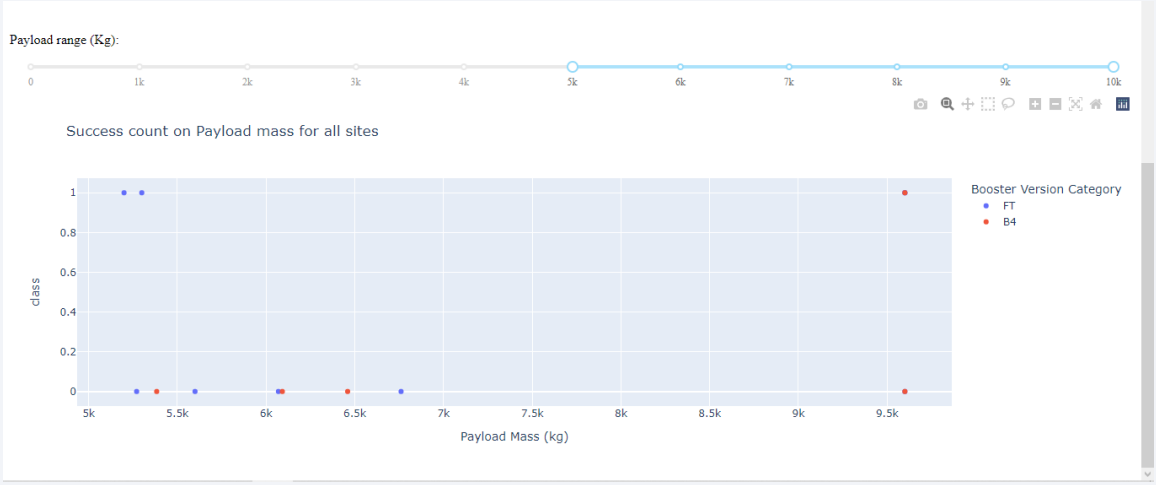
- The KSC LC-39A has the highest overall launch success count according to the pie chart, followed by the CCAFS LC-40, with the lowest being CCAFS SLC-40.
- The 2 lowest success count stations almost has the same number of counts.

KSC LC-39A Success Count (Pie)



- Out of 13 launches, KSC L-39A has succeeded 10 times, and failed only 3 times.
- The success counts of this launch station is more than double than its failure counts.

Success count on (5k-10k kg) payload mass



- In the given range of 5000 to 10000 kg, only two booster version categories are only present.
- The FT and B4 booster version categories both have relatively the same number of counts in both 0 and 1, implying that there seems to be no relationship between the variables in the given range.

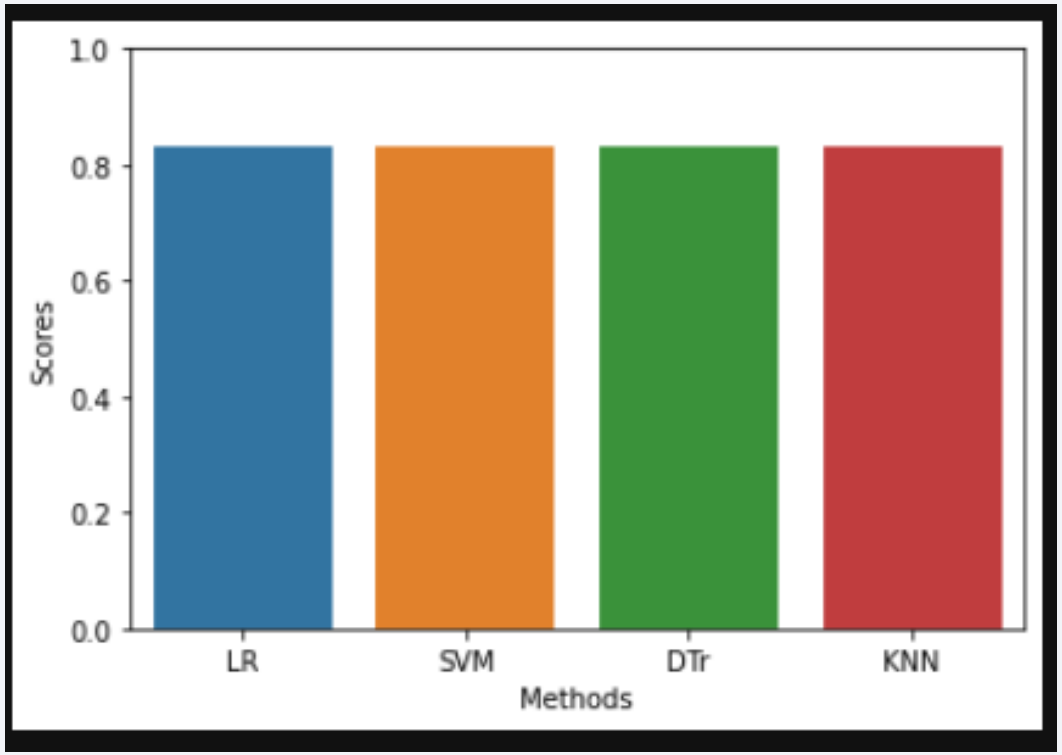
Section 5

Predictive Analysis (Classification)

Classification Accuracy

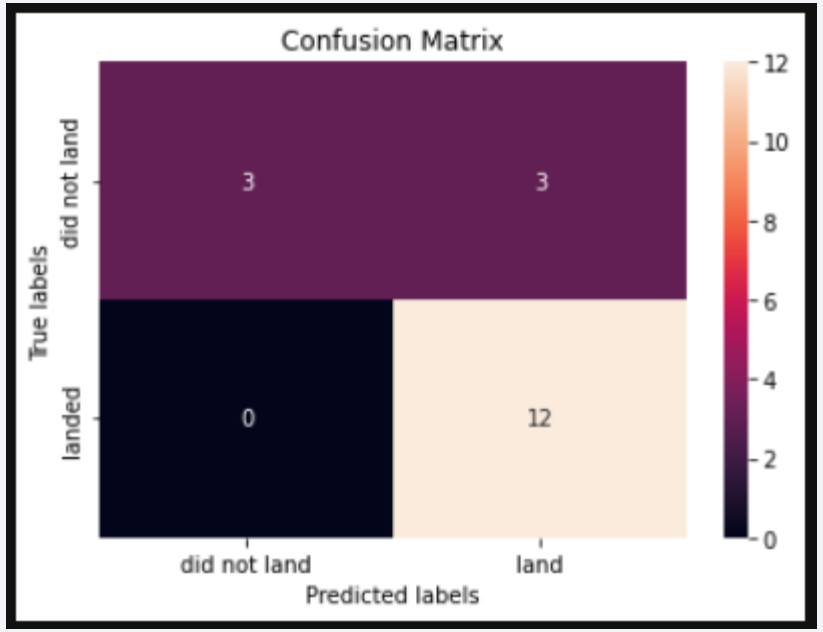
- Visualize the built model accuracy for all built classification models, in a bar chart

- There is no model which has the highest classification accuracy since their accuracies are equal at 0.83.



Confusion Matrix

- All of the models share the same appearance of confusion matrices, as they are of equal accuracies.
- In the first row, the model correctly predicted that 3 outcomes will not land, but wrongly predicted that 3 will land.
- In the 2nd row, the model correctly predicted that 12 outcomes will land.



Conclusions

- From all the models that are trained and tested, each one can be used according to the model's strength. For this case, since we are predicting the outcome of a successful land, logistic regression will work best when analyzing larger data sets.
- The case at hand is influenced by different independent variables which could involve more complex models for better predictions.
- Launch sites are placed at locations where there are railways and highways for material transport, near the coastal areas to access the oceans for possible landings, and away from the cities for public safety.
- The KSC L-39A station has the highest count of successful landings.

Appendix

- All in https://github.com/Allevium/IBM-Data-Science

Thank you!