IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dominique Blanc
11.08.2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using API & Web Scraping
  - Exploratory Analysis SQL, Pandas & Matplotlib
  - Interactive visual analytics and dashboard
  - Predictive analysis classification
- Summary of all results
  - Total payload mass carried by boosters launched by NASA (CRS) : 45596
  - Average payload mass carried by booster version F9 v1.1 : 2534.66
  - Date of the first successful landing outcome in ground : 2015-12-22
  - Maps from launch sites and distances from facilities modal connexions
  - Machine Learning (ML), best method and score : Support Vector Machine (SVM), 84.8 %

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - Predict if the Falcon 9 first stage will land successfully.
  - Accuracy and launch sites locations

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Import libraries and auxiliary functions, define a series of helper functions, request and parse the data using the GET request. Filter the dataframe to only include Falcon 9 launches
- Perform data wrangling
  - Deal with missing values, calculating and replace the missing values with mean.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize the data, split the data into training and test, create an object using different models, fit the object to find the best parameters, calculate the accuracy.

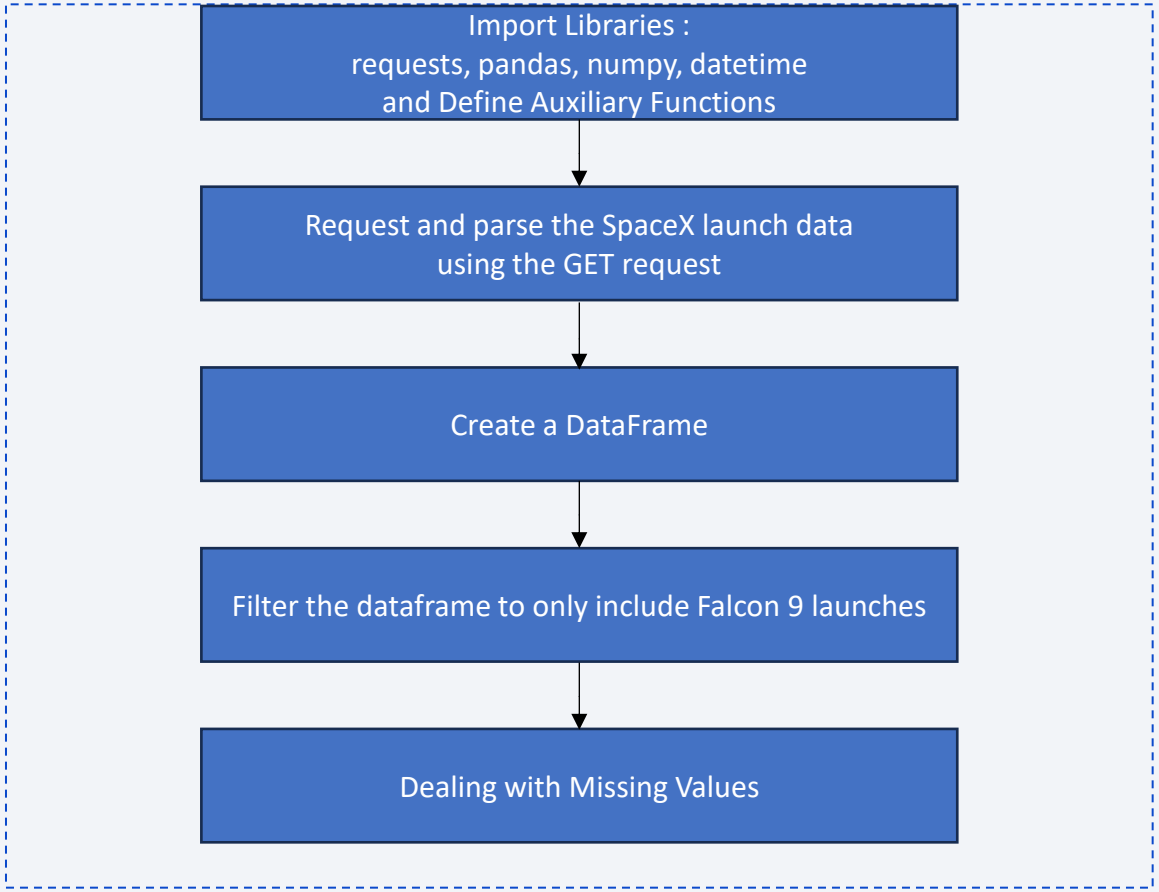
Data Collection

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

Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook :

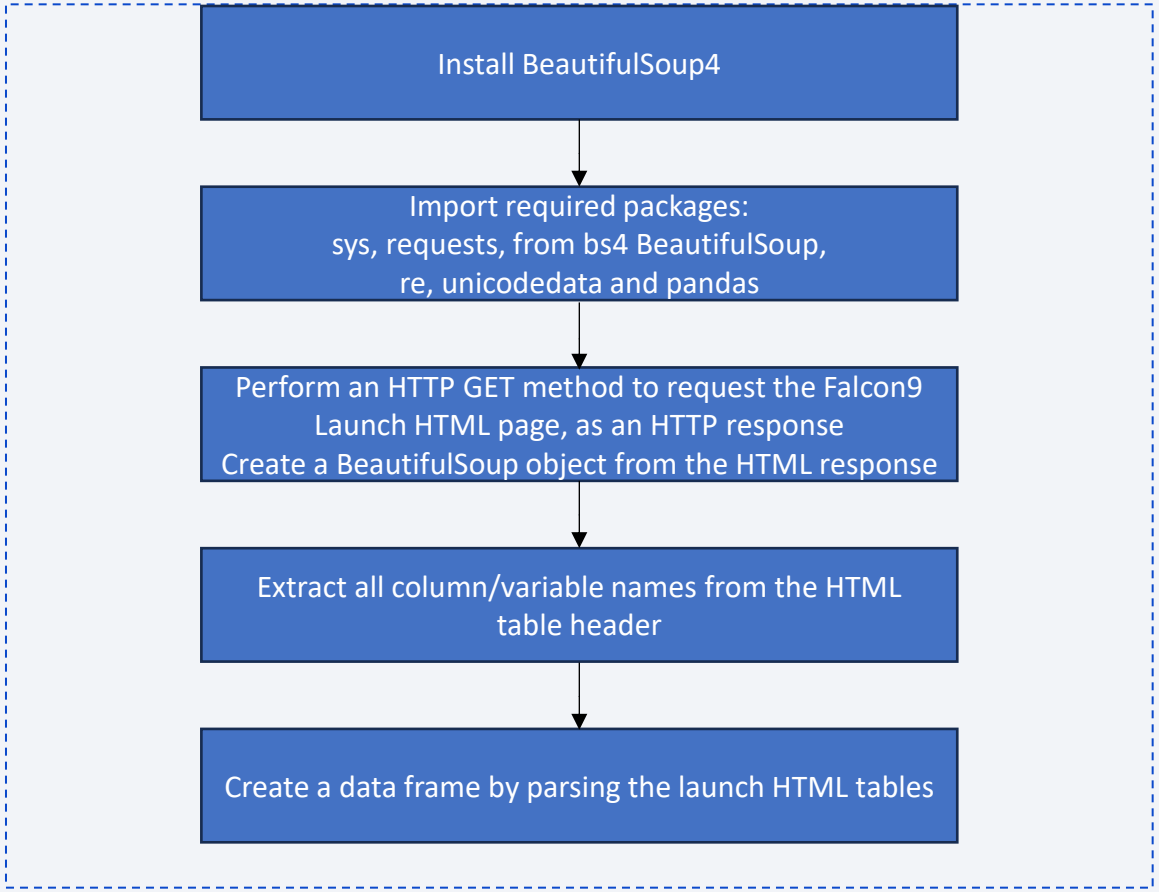
  https://github.com/dominiqueblanc/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection – Scraping

- GitHub URL of the completed web scraping notebook :

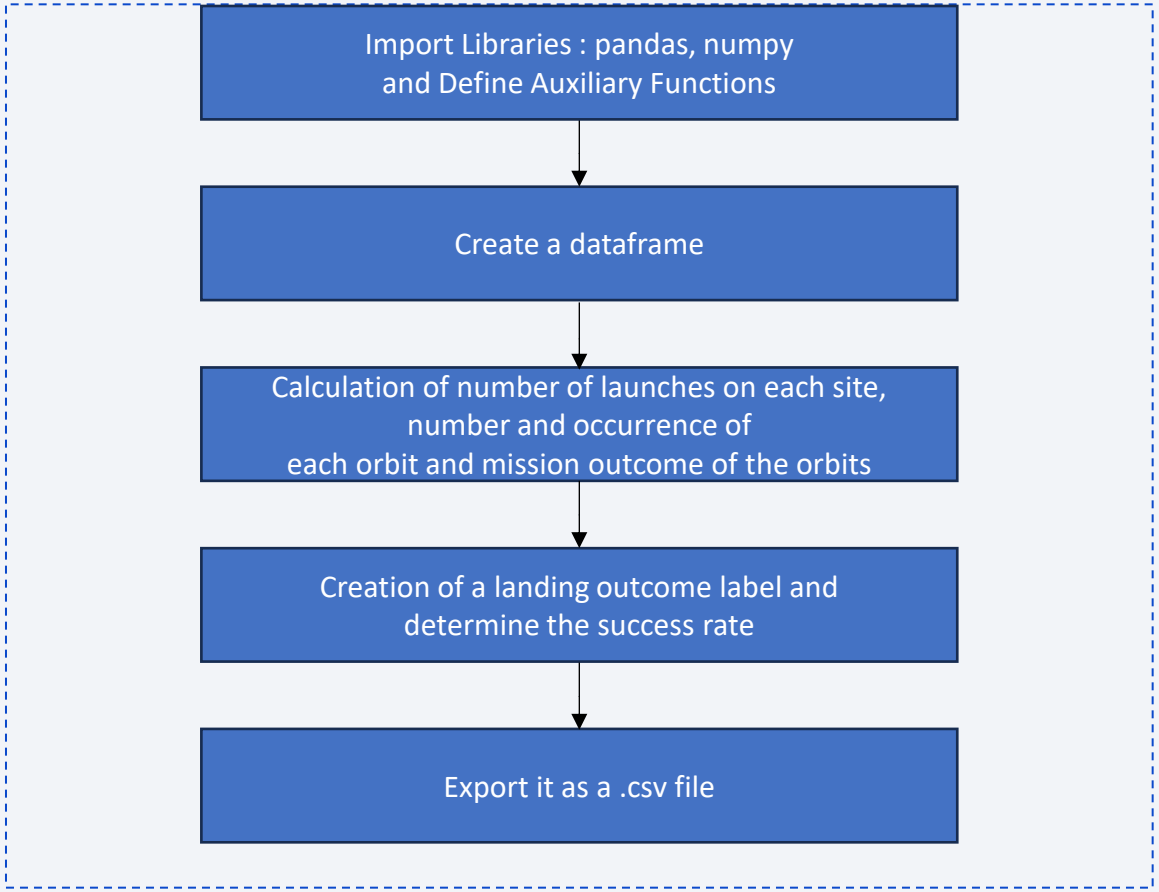
  https://github.com/dominiqueblanc/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- GitHub URL of the completed data wrangling notebook :

  https://github.com/dominiqueblanc/IBM-Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Catplot chart is used to visualize relationships between variables as Flightnumber vs Payloadmass, LaunchSite, OrbitType, etc.
  - Bar chart is used to visualize relationship between success rate and orbit type
  - Line plot is used to visualize the launch success yearly trend
- GitHub URL of completed EDA with data visualization notebook: https://github.com/dominiqueblanc/IBM-Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Create a table SPACEXTABLE
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass. Use a subquery
  - List 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, in descending order.
- GitHub URL of completed EDA with SQL notebook : https://github.com/dominiqueblanc/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Circle
  - Marker clusters
  - Polyline

- Explain why you added those objects
  - Circle to add a highlighted circle area for each launch site
  - Marker clusters to simplify a map containing many markers having the same coordinate, i.e. success/failed launch
  - Polyline to draw a distance between a launch site to its proximities

- GitHub URL of completed interactive map with Folium map : https://github.com/dominiqueblanc/IBM-Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

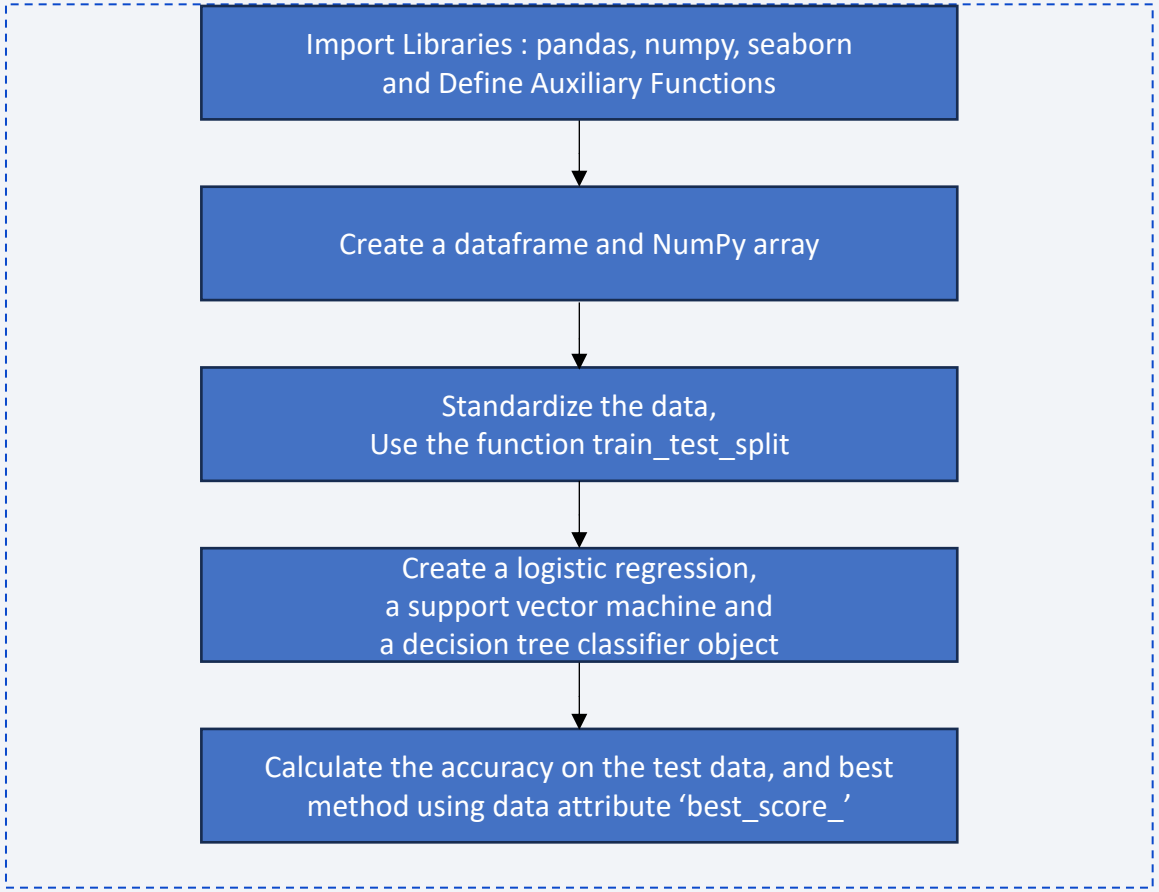
Build a Dashboard with Plotly Dash

- 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
- GitHub URL of completed predictive analysis lab :

  https://github.com/dominiqueblanc/IBM-Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

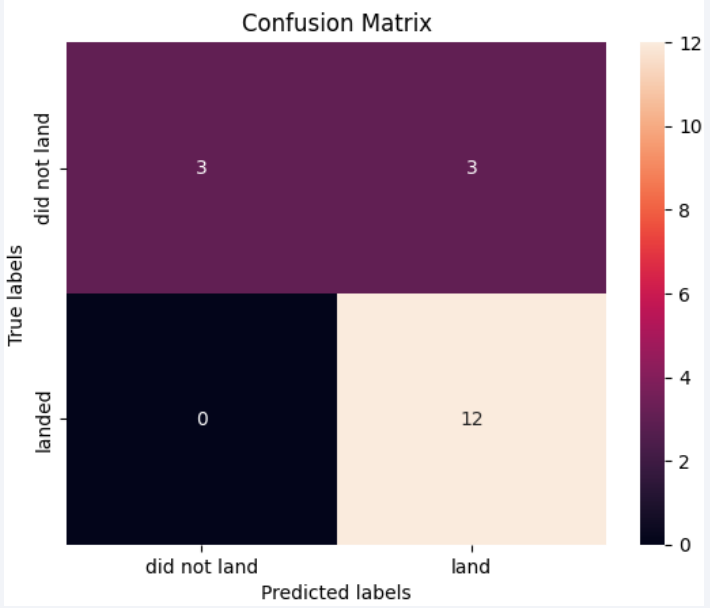


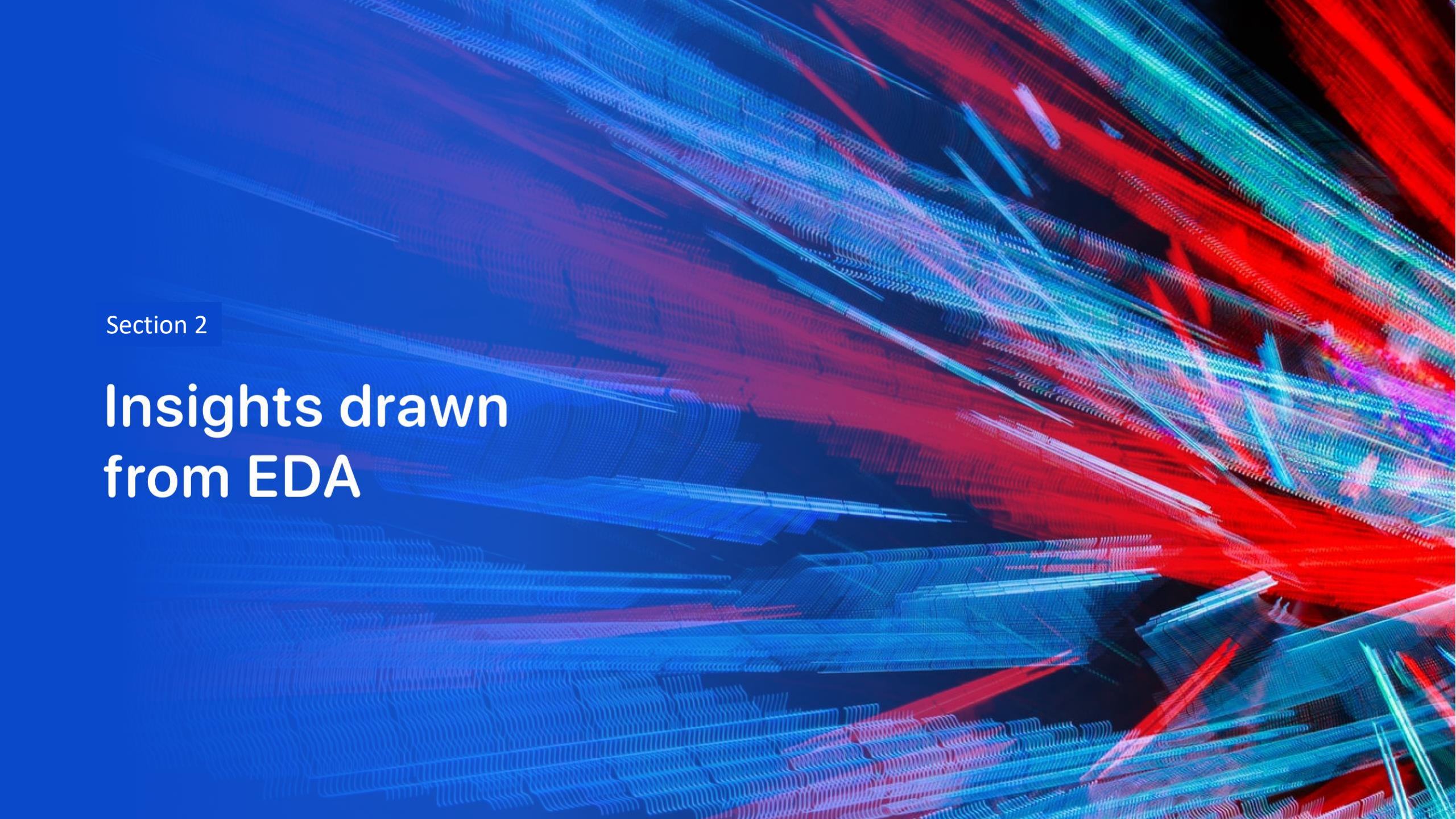
Results

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

Method :
K Nearest Neighbours (KNN)

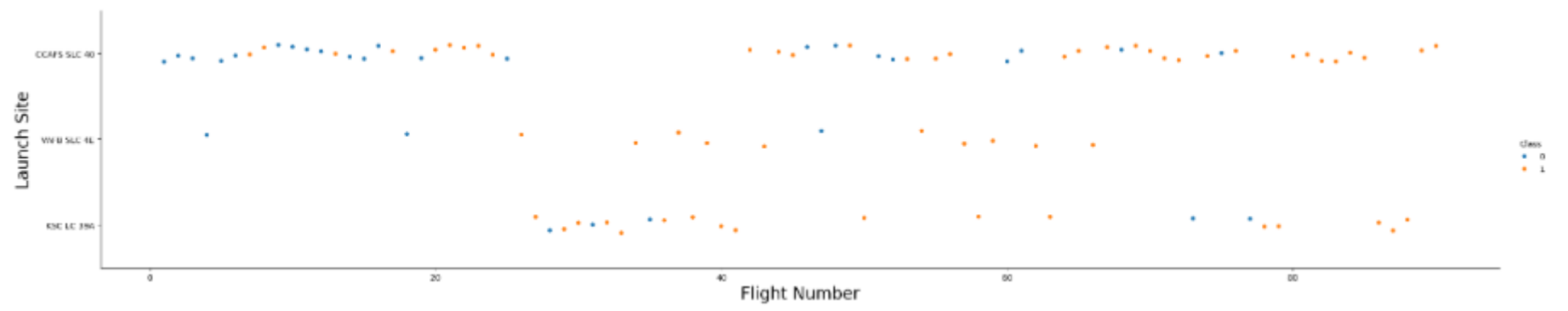
Accuracy :
84.8 %




Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

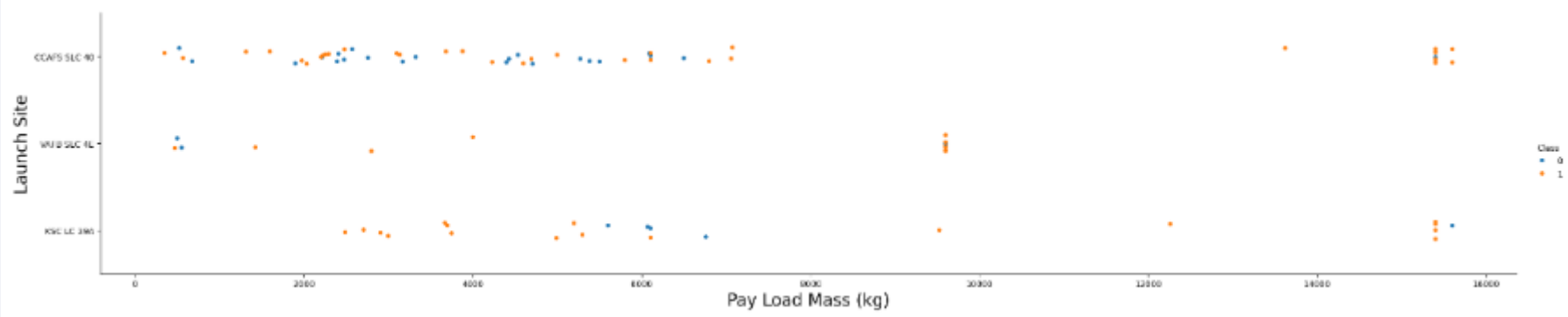
- Show a scatter plot of Flight Number vs. Launch Site



- Show the screenshot of the scatter plot with explanations
  - Most of launches has been done at CCAFS-SLC-40 launch site
  - Early launches were a success while, by time, many failures occurred.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

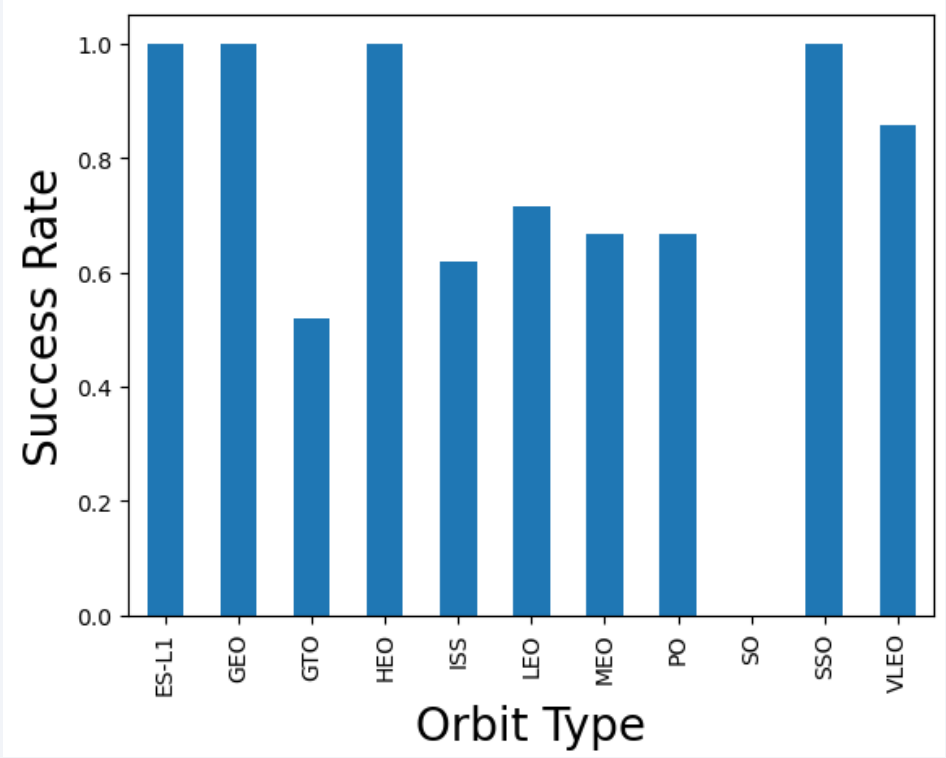


- Show the screenshot of the scatter plot with explanations
  - At VAFB-SLC launch site, there are no rockets launched for heavy payload mass (> 10000).
  - At CCAFS-SLC, there are no rockets launched between 7000 and 150000 (excepting one)

Success Rate vs. Orbit Type

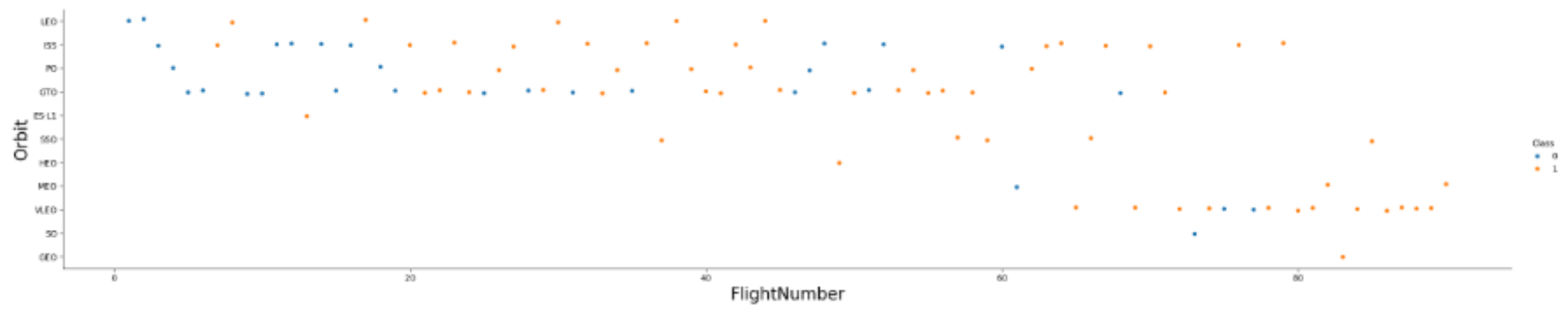
- Show a bar chart for the success rate of each orbit type

- Show the screenshot of the scatter plot with explanations
  - ES-L1, GEO, HEO and SSO has the best success rate while SO has none.



Flight Number vs. Orbit Type

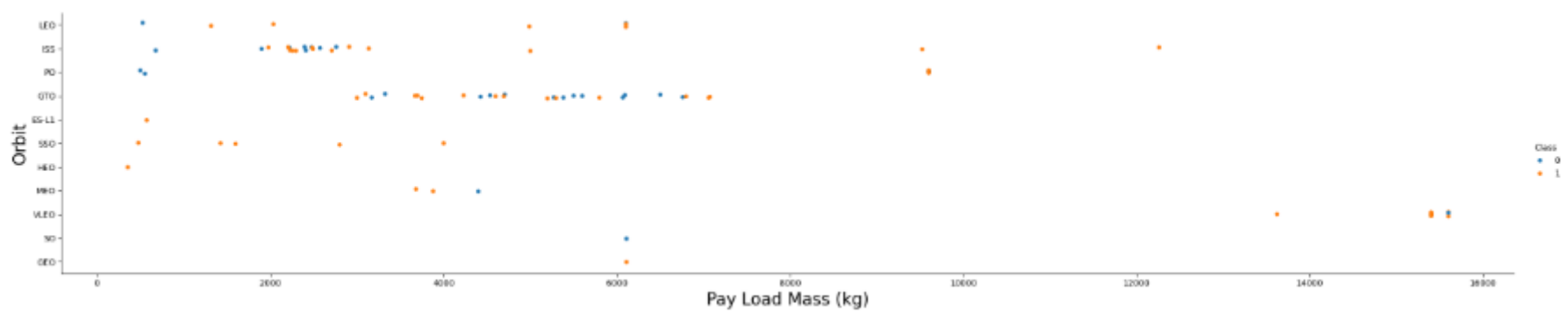
- Show a scatter point of Flight number vs. Orbit type



- Show the screenshot of the scatter plot with explanations
  - In the LEO orbit, success seems to be related to the number of flights. On the other hand, in the GTO orbit, no relationship appears between flight number and success

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

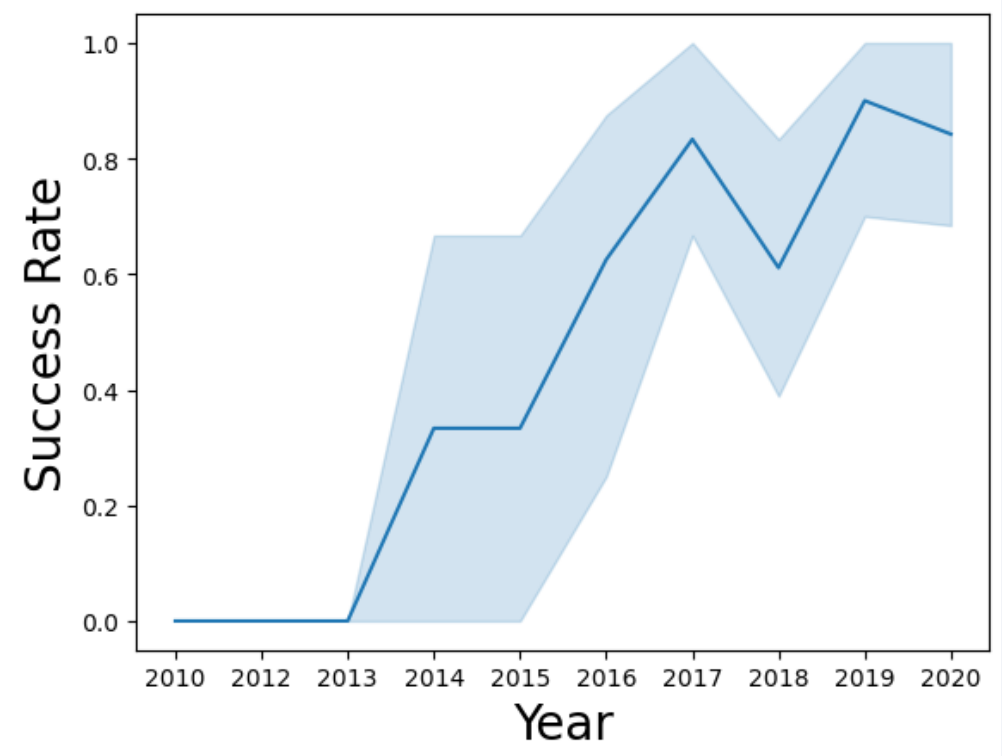


- Show the screenshot of the scatter plot with explanations
  - Polar, LEO and ISS have the most successful/positive landing rate with heavy payloads.
  - For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present

Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations
  - Success rate increased non-stop since 2013 until 2017.
  - However, it decreased in 2018 around 20%, until it increased again to its best in 2019, showing a strong corrective action capacity



All Launch Site Names

- Find the names of the unique launch sites

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

- Present your query result with a short explanation here

```
%sql select Launch_Site from SPACEXTBL group by Launch_Site
```

  - By grouping by Launch_Site, the result will display the unique name of each.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

  - Using ‘like’ and a joker ‘%’ express the occurrences that begins with the string defined between semi-quotes or parentheses (i.e. ‘CCA%’).
  - A limit 5, restrict the display to the number of records.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

sum(PAYLOAD_MASS__KG_)
45596

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer=='NASA (CRS)'
```

  - Using the function ‘sum’ of the attribute ‘Payload’ calculates the total.
  - Using ‘where’ to be specific, with a condition like (sum where customer is NASA)

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

- Using the ‘avg’ function to the attribute ‘Payload’ calculates the average
- Using ‘where’ specifies the booster version (average of a certain booster version)
- Using ‘like’ will consider as ‘true’ what looks like, even if not rigorously exact in a string
- Using a joker ‘%’ display all occurrences of a string that begins with what’s on its left : (i.e. ’F9 v1.1’)

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

min(date)
2015-12-22

```
%sql select min(date) from SPACEXTBL where Landing_Outcome='Success (ground pad)'
```

- Using the ‘min’ function will display the lowest result. As a date, the lowest means the 1st.
- The condition is ‘where’ the landing was successful on ground pad, according the records syntax ‘Success (ground pad)’

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

```
%sql select * from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

  - I always prefer select all the fields with a '*' so I can check the content and then be more specific if needed.
  - Conditions are : 1) successful landing on drone ship and 2) payload between two values
  - A 'where' is placed to display all 'Success (drone ship)' of Landings, plus an 'and' for the 2nd condition on the payload, using the 'between' function that replaces '>' and '<' signs.

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

sion	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	count(Mission_outcome)
)003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)	100
1018	CCAFS LC-40	SpaceX CRS-7	1952	LEO (ISS)	NASA (CRS)	Failure (in flight)	Precluded (drone ship)	1

```
%sql select *, count(Mission_outcome) from SPACEXTBL group by Mission_Outcome like 'Fail%'
```

- Using the function ‘count’ will sum the number of missions
- Using both ‘group by’ and ‘like’ will split in two categories : failures and not failures (=success)

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

```
%sql select * from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl)
```

  - Using ‘where’ determine the condition : all that have carried the max payload mass
  - Using the function ‘max’ applied to the ‘Payload’ display the highest mass
  - Using an intricated condition ‘where’ a selection of Payload (the max carried) is called a subquery. It looks like ‘where’ something equal a non-value but a selection of values i.e. all results where the value equal the maximum value of the column of the table.

2015 Launch Records

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

```
#%sql select *, substr(Date, 6,2) as month, substr(Date,0,5)='2015' as year from SPACEXTBL where PAYLOAD_MASS__KG_ == max(P
%sql select *, substr(Date, 6,2) as month, substr(Date,0,5)='2015' as year from SPACEXTBL where Landing_Outcome like 'Fail%
```

- Present your query result with a short explanation here
  - Using ‘where’ as a condition, to display the failed landing in drone
  - Using ‘substr’ function to split the date string as month and year : Date for the column, 6 to specify that starts from the 6th character from left-to-right, 2 to specify the number of characters = month from 01 to 12)

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

```
%sql select *, count(Landing_Outcome) from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by Landing_Outc
```

```
from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(Landing_Outcome) desc
```

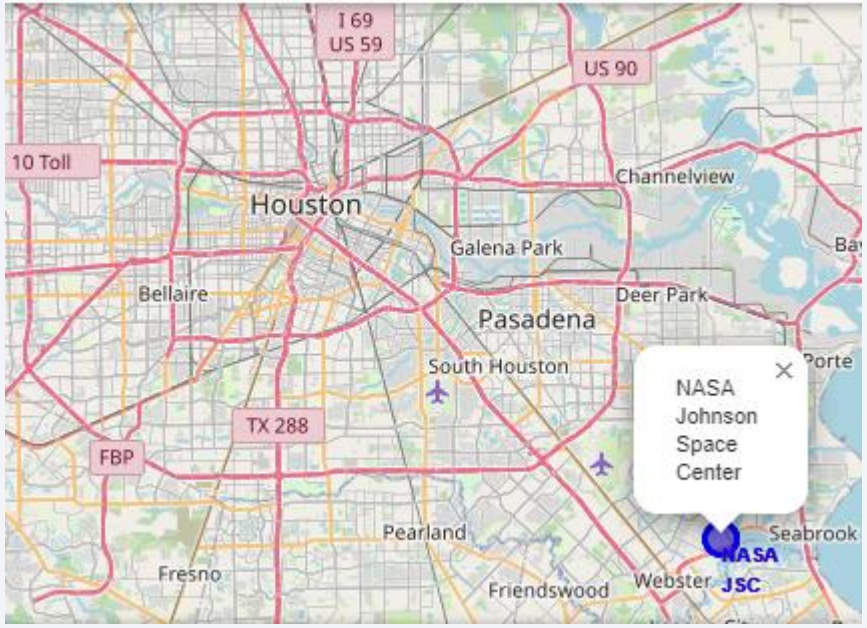
- Present your query result with a short explanation here
  - Using ‘order by’ sorts the results, which lead to get the ranking
  - Using a ‘count’ function make the attribute unique (as a group by)
  - Using ‘where’ added to ‘between’ are combined conditions that applies in this case.

Section 3

Launch Sites Proximities Analysis

NASA Johnson Space Center at Houston, Texas

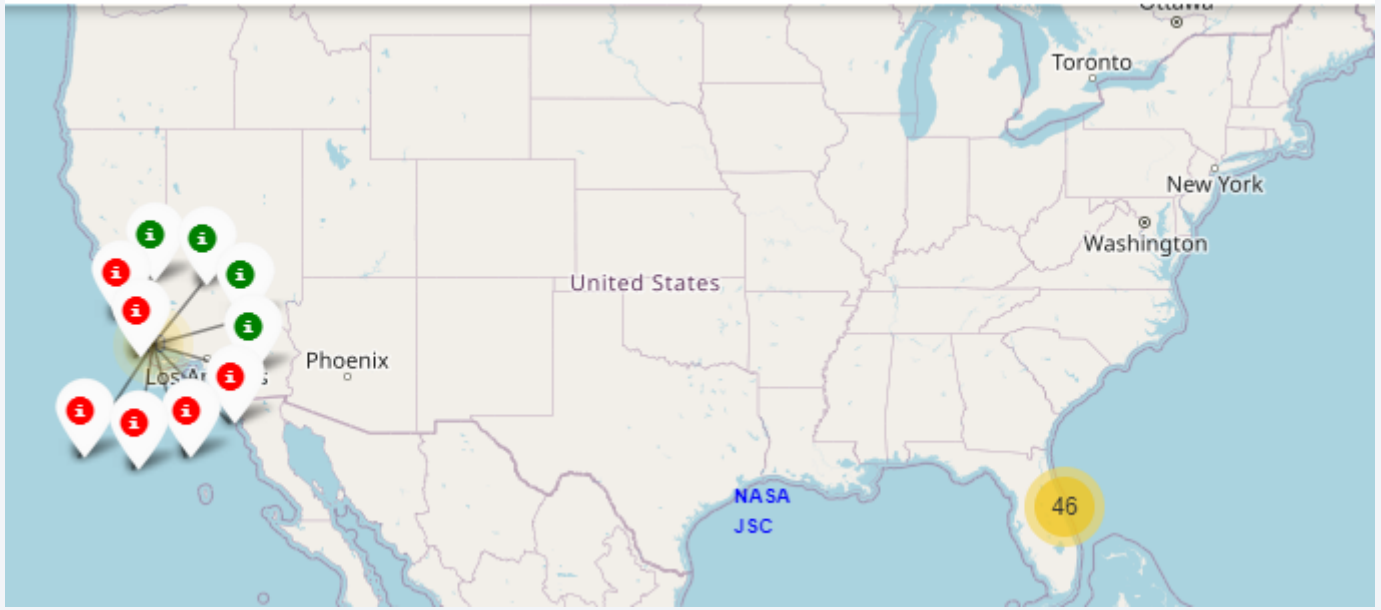
- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot

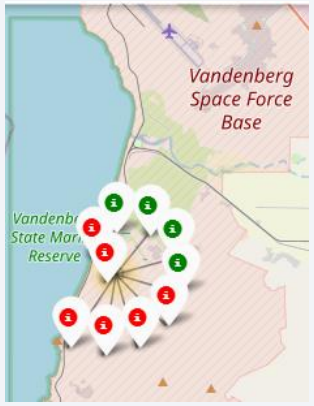


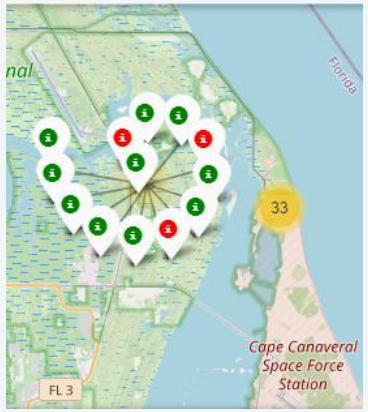
US Launch Sites

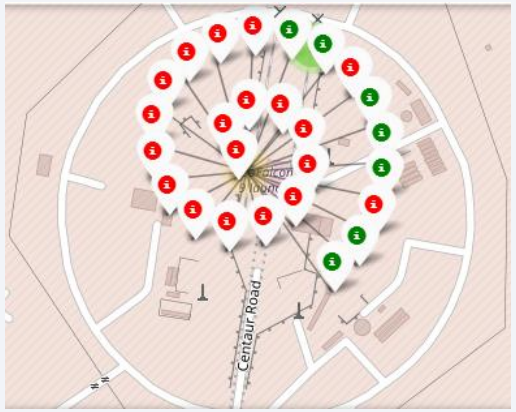
- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

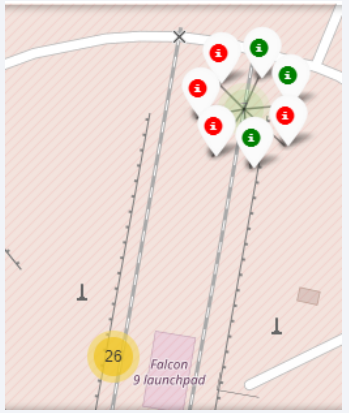
US Success/Failed launches





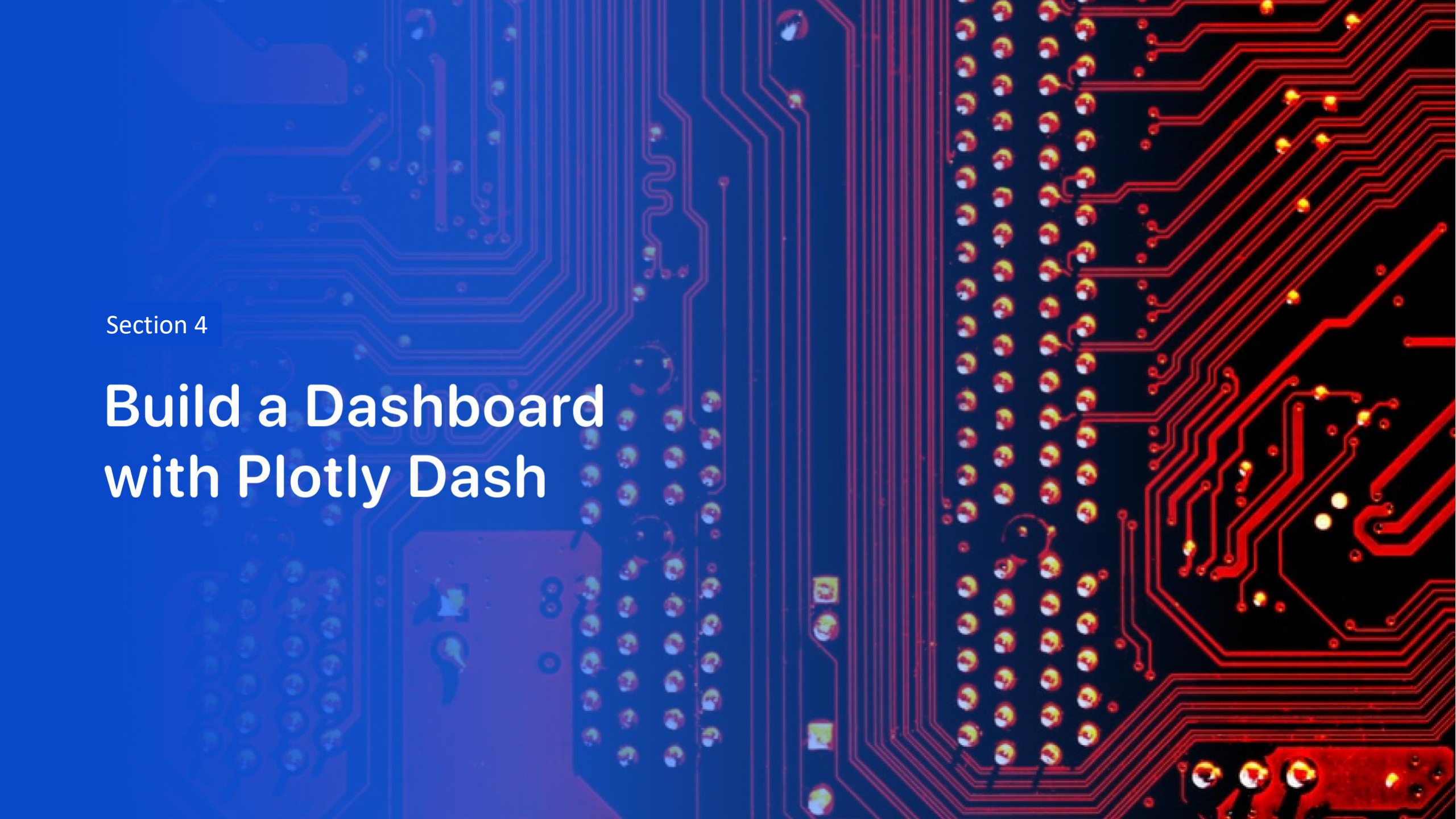






US Launch Sites and proximities findings

- Launch sites are in close proximity as well as to coastline, highway and railroad, while it keep certain distance away from cities.
- Facilities are very next to the launch sites, which makes it easy to join for staff, visitors, goods and supplies.
- Railroad next to the launch sites are specific due to the extra-size of the elements in the space field, however the railroad is linked to the national network, passengers, staff and visitors can easy access by train.
- The proximity of the coastline as well as the distance away from the cities makes the launches safer for the residents. If an accident may always occur, it minimizes the impacts of crashes and falling debris.


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

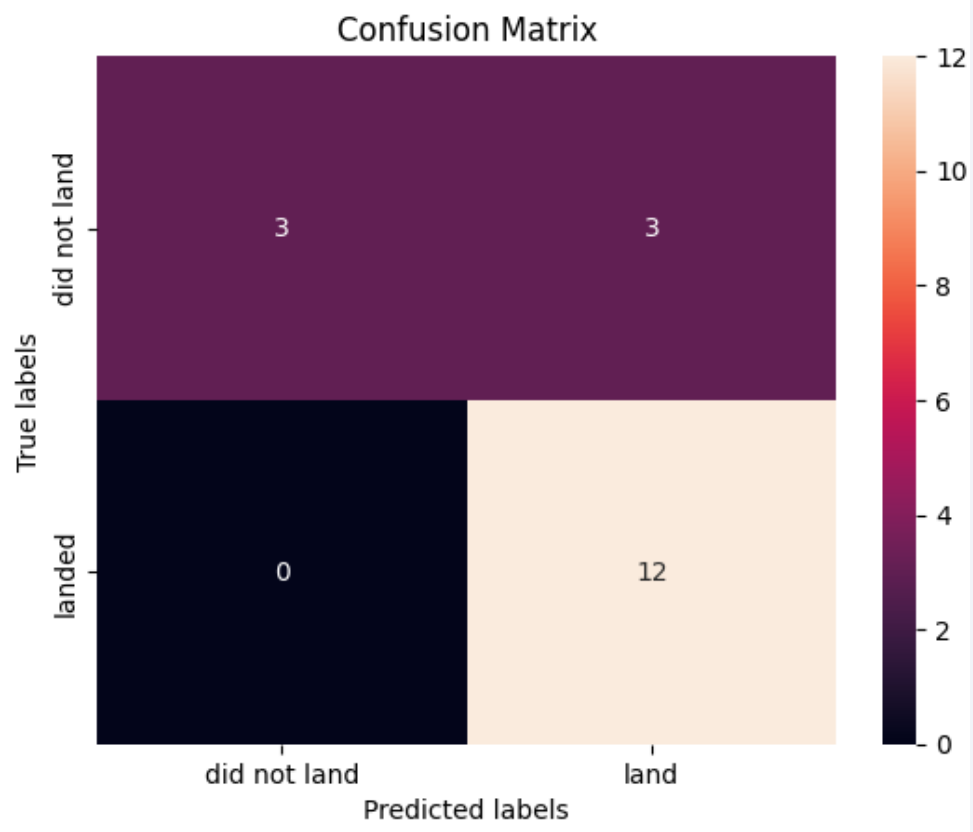
Classification Accuracy

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

- Find which model has the highest classification accuracy

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation
  - K Nearest Neighbors (KNN)
  - Accuracy : 84.8 %
    - 12 True Positives
    - 3 False Positives



Conclusions

- Point 1
  - With a 62 million dollars cost and an accuracy of 84.8%, we can estimate a cost of 73.2 million dollars each (62/84.8*100).
- Point 2
  - The next target would be to increase the accuracy to reduce the waste due to landing failures, which is currently around 15.2 % = 9.5 million dollars.
- Point 3
  - Some boosters and/or launchers landing success is related to the payload mass.
    We need to identify further the root causes

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- Notebooks : https://github.com/dominiqueblanc/IBM-Applied-Data-Science-Capstone

Thank you!