IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Liviu Petre
23.03.2023

Outline

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

Executive Summary

- Objective
  - To determine the target price of each launch for Space Y by gathering information about Space X competitor and more specifically about their capacity to reuse the first stage
- Summary of methodologies
  - Data collection (using API or Web Scrapping)
  - Data wrangling
  - Exploratory data analysis using SQL
  - Exploratory data analysis using Visualization
  - Interactive visual analytics (Folium)
  - Interactive dashboard (Ploty)
  - Predictive analysis
  - Results presentation (current document)
- Summary of all results – can be seen in the last slide

Introduction

- Project background and context
  - There is a number of new companies activating in the space travel industry
  - One of the most successful is Space X which advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each
  - One of the reasons behind this saving is Space X' capacity of reusing the 1st stage which is the most expensive one. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.
- Problems you want to find answers
  - The main problem is to determine the cost of a launch and in order to do this we must be able to determine if the first stage will land and thus be reused

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected in 2 ways: either through Space X API or through web scrapping a Wikipedia page
- Perform data wrangling
  - Various calculations were made: launches/site, launches/orbit, number of mission outcome/orbit type
  - In the end a landing outcome label was created (1 for success and 0 for fail) – which indicates whether stage 1 was recovered or not

Methodology

Executive Summary contd.

- Perform exploratory data analysis (EDA) using visualization and SQL
  - Space X data in SQL/CSV format was loaded in a DB2 database and various queries were executed (see next slides)
- Perform interactive visual analytics using Folium and Plotly Dash
  - Folium and Plotly were used in order to visualize the data – dashboard can be found in the next slides. This was used in order to get insight into data correlations
- Perform predictive analysis using classification models
  - The data was standardized and split intro train and test data
  - Then various classification models were used: logistic regression, support vector machine, decision tree and k nearest neighbor
  - The classification model with best accuracy was selected and confusion matrix is also presented in the next slides

Data Collection

- Describe how data sets were collected.
  - Data was collected in 2 ways: either through Space X API or through web scrapping a Wikipedia page. Each way workflow is described below and in the next slide.
  - **The SpaceX REST API endpoints**, or URL, starts with api.spacexdata.com/v4/. There are different end points, for example: /capsules and /cores.
    - We used the endpoint api.spacexdata.com/v4/launches/past. to get past launch data.
    - We performed a get request using the requests library to obtain the launch data, which we then used to get the data from the API. The result can be viewed by calling the .json() method. The response is in the form of a JSON, specifically a list of JSON objects which each represent a launch.
    - To convert this JSON to a dataframe, we used the json_normalize function. This function allowed us to "normalize" the structured json data into a flat table.

Data Collection

- Describe how data sets were collected – web scrapping
    - Web scraping related Wiki pages.
        - We used the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records on the Wikipedia page.
        - We parsed the data from those tables and converted them into a Pandas data frame for further visualization and analysis.
        - Some of the columns, like rocket, have an identification number, not actual data. For this we used the API again targeting another endpoint to gather specific data for each ID number for the following: Booster, Launchpad, payload, and core.
        - We also noticed that the launch data we have includes data for the Falcon 1 booster whereas we only want Falcon 9. So Falcon 1 launches were removed.
        - Finally, we replaced the NULL values inside the PayloadMass with the mean. The column LandingPad was left with NULL values, as it is represented when a landing pad is not used

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

Obtain rocket launch data from the following URL:

```
requests.get(spacex_url)
```

Use the API again to get information about the launches using the IDs given for each launch. Specifically we will be using columns rocket, payloads, launchpad, and cores:

```
requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()
requests.get("https://api.spacexdata.com/v4/cores/"+core['core']).json()
```

GitHub URL

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

- perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response
- Create a BeautifulSoup object from the HTML response
- the third table is our target table contains the actual launch records.
- iterate through the <th> elements and apply the provided extract_column_from_header() to extract column name one by one
- fill up the launch_dict with launch records extracted from table rows
- Create the data frame from the dictionary

GitHub URL

Data Wrangling

- Describe how data were processed
  - we converted the outcomes in the table into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful
- You need to present your data wrangling process using key phrases and flowcharts
  - Exploratory Data Analysis
    - the number of launches on each site
    - the number and occurrence of each orbit
    - the number and occurrence of mission outcome per orbit type
  - Determine Training Labels
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose
- GitHub URL

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - We plotted out the FlightNumber vs. PayloadMass and and overlaid the outcome of the launch. We saw that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
  - We then visualized the relationship between Payload and Launch Site. We observed for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000). We also noticed that CSA launch site is more successful on small payloads, while CCA is more successful on heavy payloads
  - We then did a bar chart to Visualize the relationship between success rate of each orbit type. We noticed a high success rate for SSO and VLEO orbits
  - We then visualized the relationship between FlightNumber and Orbit type. We noticed that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
  - Last we visualized Payload vs Orbit Type. We noticed that with heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
- GitHub URL

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - 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 acheived.
  - 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for 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

- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

- GitHub URL. Note: because SQL connection did not work, queries were ran directly in the database.

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Launch sites
  - Successful launches per site
  - Distance between launch site and its proximities
- Explain why you added those objects
  - On the map it is ideal to visualize geographically the locations of the launch sites as well as other information that might be relevant.
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
  - GitHub URL

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

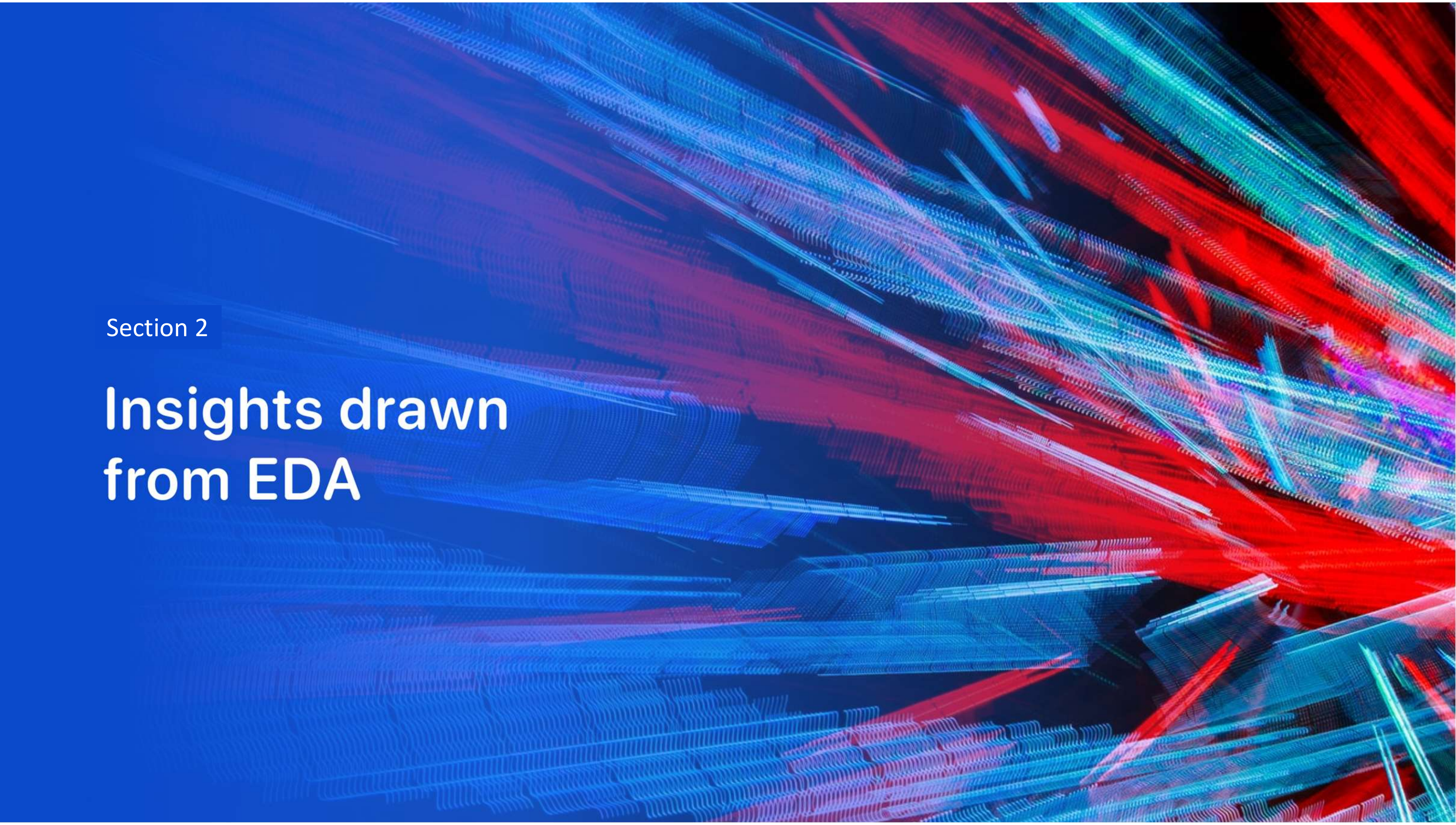
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

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
  - I started from the data frame containing the data
  - The data was standardized and then split into 80% training data and 20% test data
  - 4 models were then created (logistic regression, support vector machine, decision tree and k nearest neighbor) and for each the accuracy was calculated and the confusion matrix was plotted
  - Finally a bar chart was used to see which model has the best accuracy
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- Git hub url

Results – see next slides

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

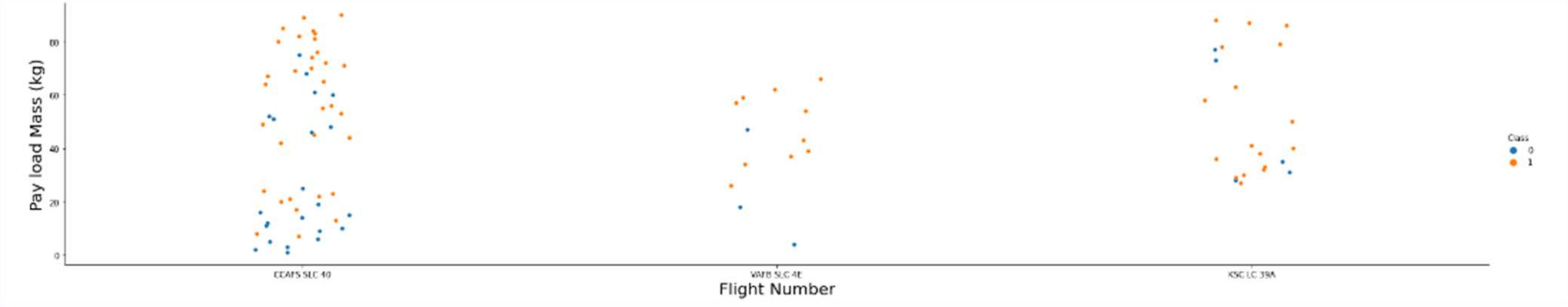


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

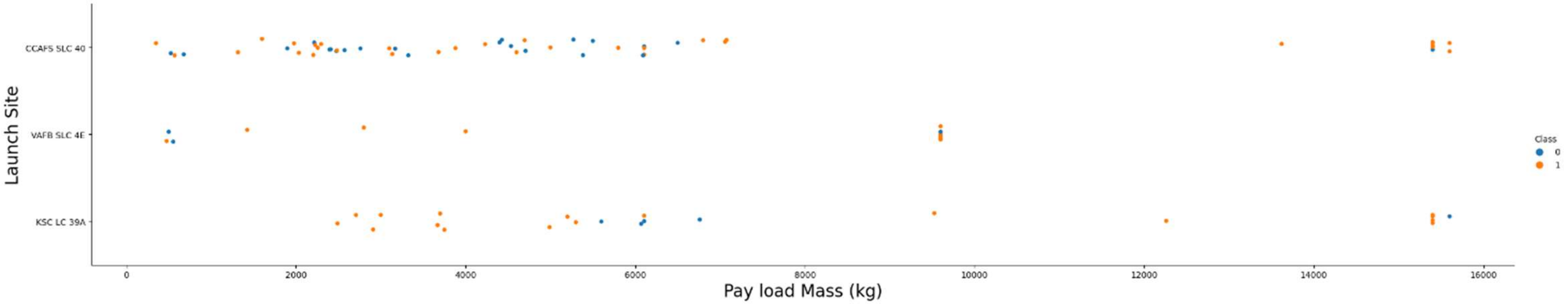
- Show a scatter plot of Flight Number vs. Launch Site.



- We noticed that CCA has most flights, KSC second most and VAFB is 3rd.

Payload vs. Launch Site

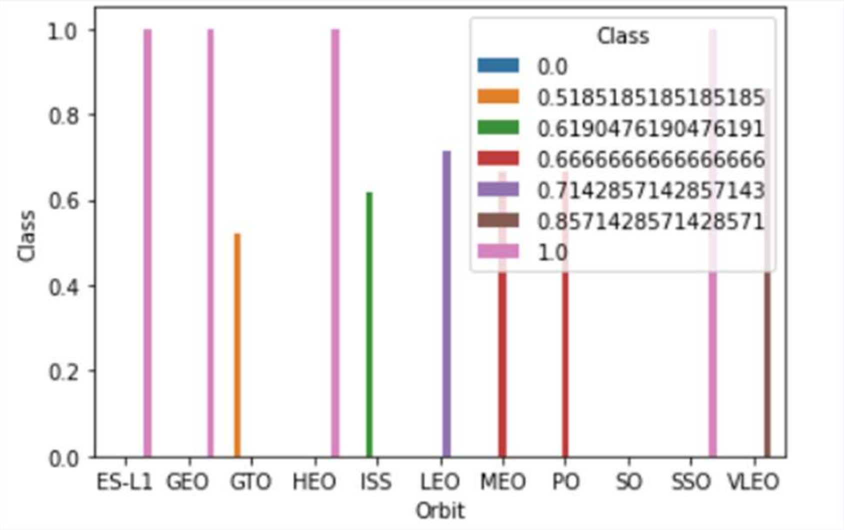
- Show a scatter plot of Payload vs. Launch Site



We observed for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000). We also noticed that CSA launch site is more successful on small payloads, while CCA is more successful on heavy payloads

Success Rate vs. Orbit Type

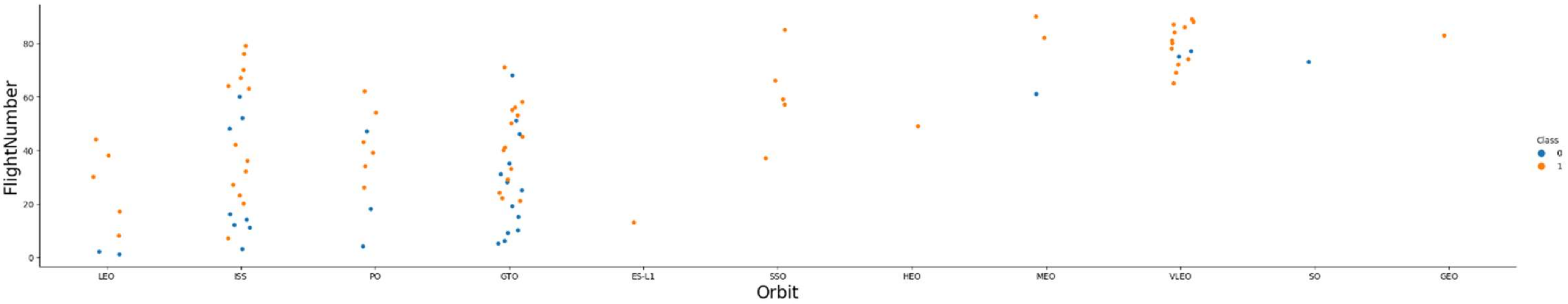
- Show a bar chart for the success rate of each orbit type



We noticed a high success rate for SSO and VLEO orbits

Flight Number vs. Orbit Type

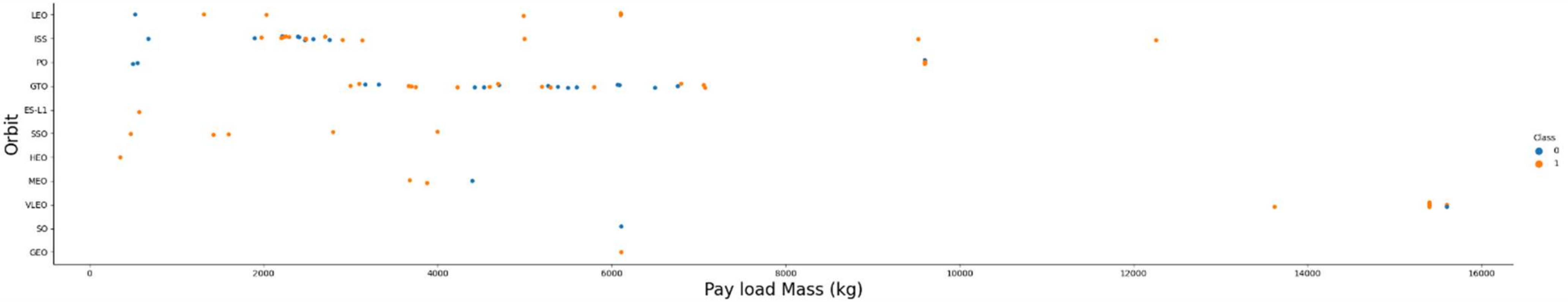
- Show a scatter point of Flight number vs. Orbit type



We noticed that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

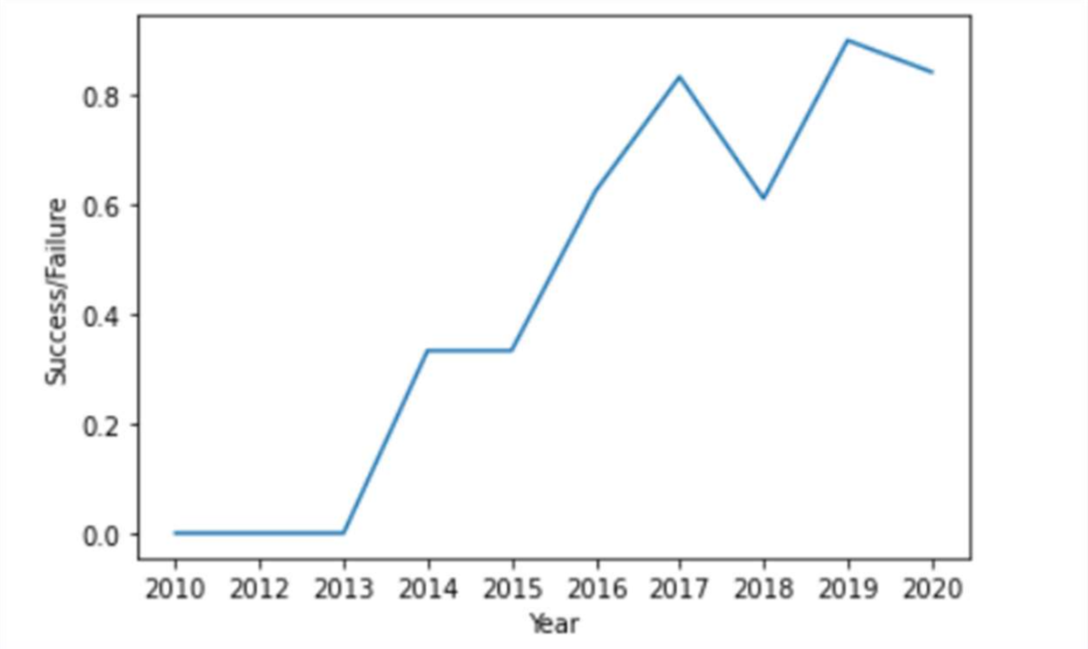
- Show a scatter point of payload vs. orbit type



We noticed that with heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

Launch Success Yearly Trend

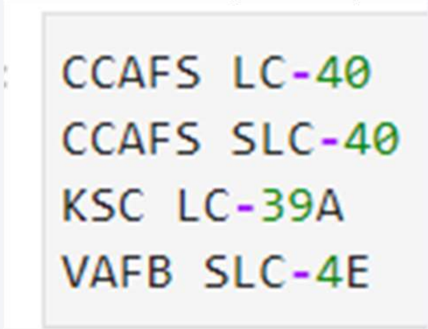
- Show a line chart of yearly average success rate



We can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

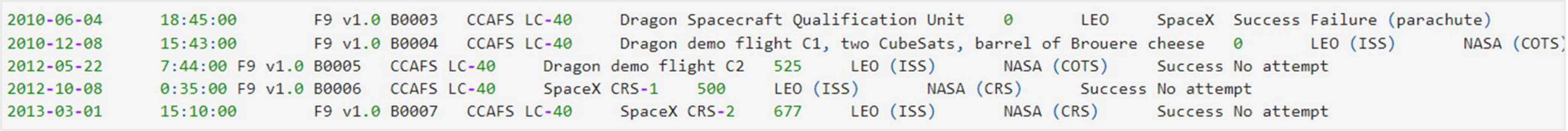
- Find the names of the unique launch sites.
- Present your query result with a short explanation here



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

Launch Site Names Begin with 'CCA'

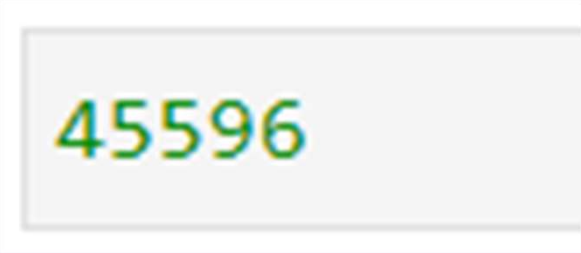
- Find 5 records where launch sites begin with `CCA`.
- Present your query result with a short explanation here



```
2010-06-04      18:45:00        F9 v1.0 B0003   CCAFS LC-40     Dragon Spacecraft Qualification Unit    0       LEO     SpaceX  Success Failure (parachute)
2010-12-08      15:43:00        F9 v1.0 B0004   CCAFS LC-40     Dragon demo flight C1, two CubeSats, barrel of Brouere cheese   0       LEO (ISS)       NASA (COTS)
2012-05-22      7:44:00 F9 v1.0 B0005   CCAFS LC-40     Dragon demo flight C2   525     LEO (ISS)       NASA (COTS)     Success No attempt
2012-10-08      0:35:00 F9 v1.0 B0006   CCAFS LC-40     SpaceX CRS-1    500     LEO (ISS)       NASA (CRS)      Success No attempt
2013-03-01      15:10:00        F9 v1.0 B0007   CCAFS LC-40     SpaceX CRS-2    677     LEO (ISS)       NASA (CRS)      Success No attempt
```

Total Payload Mass

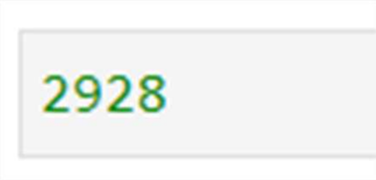
- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here



```
45596
```

Average Payload Mass by F9 v1.1

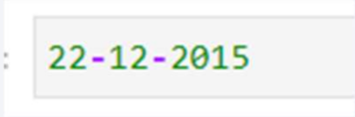
- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here



```
2928
```

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

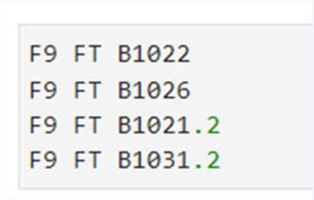


```
22-12-2015
```

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

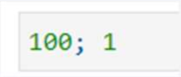
- Present your query result with a short explanation here



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

Total Number of Successful and Failure Mission Outcomes

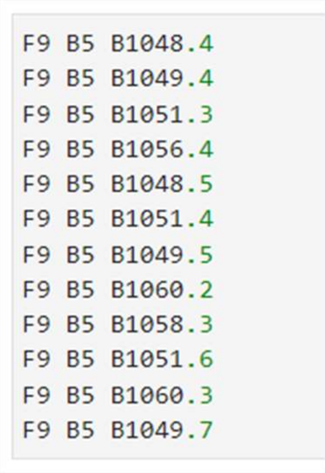
- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here



```
100; 1
```

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

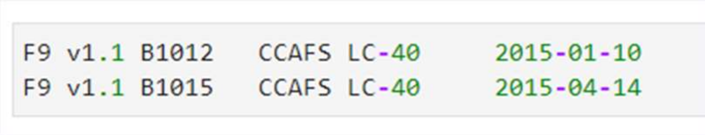


```
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7
```

2015 Launch Records

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

- Present your query result with a short explanation here

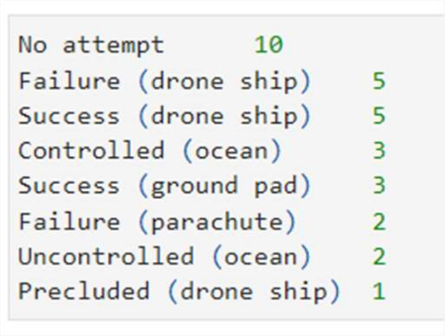


```
F9 v1.1 B1012   CCAFS LC-40     2015-01-10
F9 v1.1 B1015   CCAFS LC-40     2015-04-14
```

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

- Present your query result with a short explanation here



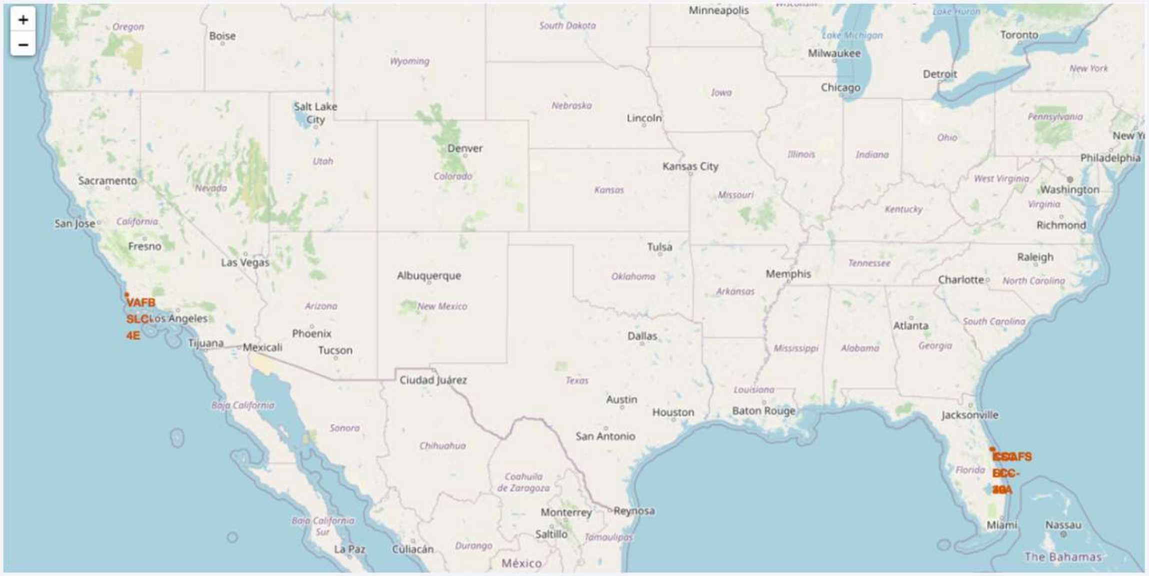
```
No attempt      10
Failure (drone ship)    5
Success (drone ship)    5
Controlled (ocean)      3
Success (ground pad)    3
Failure (parachute)     2
Uncontrolled (ocean)    2
Precluded (drone ship)  1
```

Section 3

Launch Sites Proximities Analysis

Launch sites

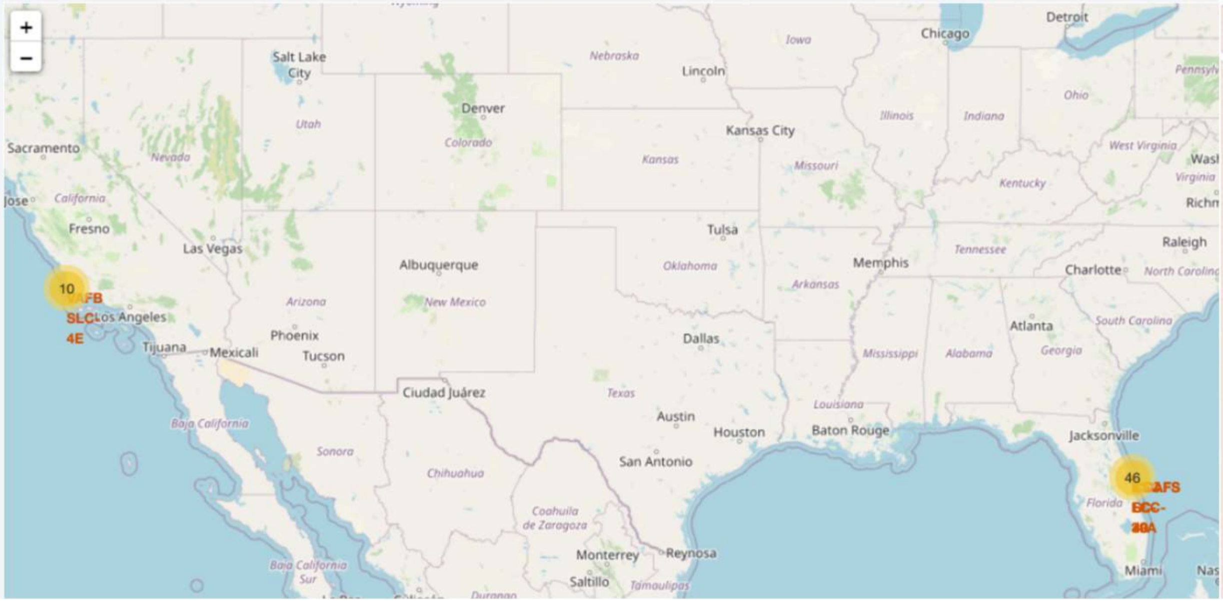
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot.

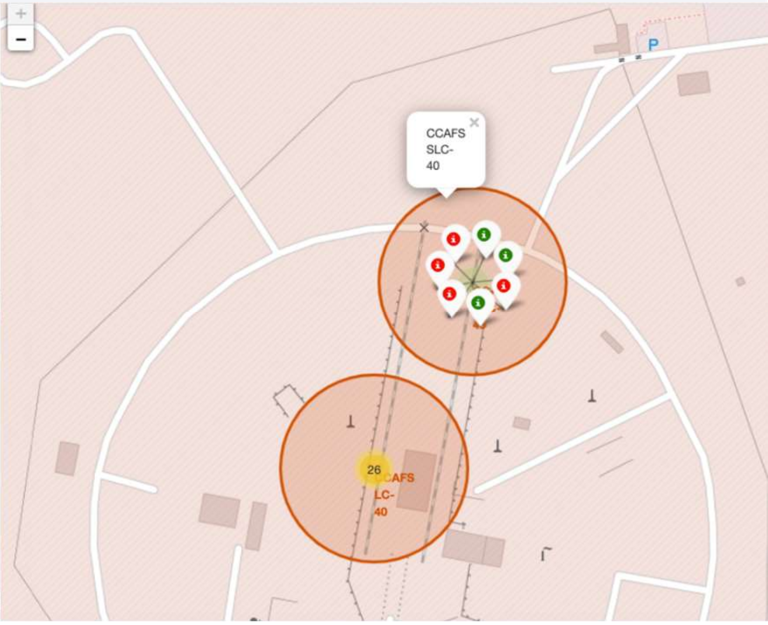


- All launch sites are on the coast to make it easier to land in the ocean and avoid areas with population
- The launch sites are on different coasts for logistic reasons.

Successful launches per site

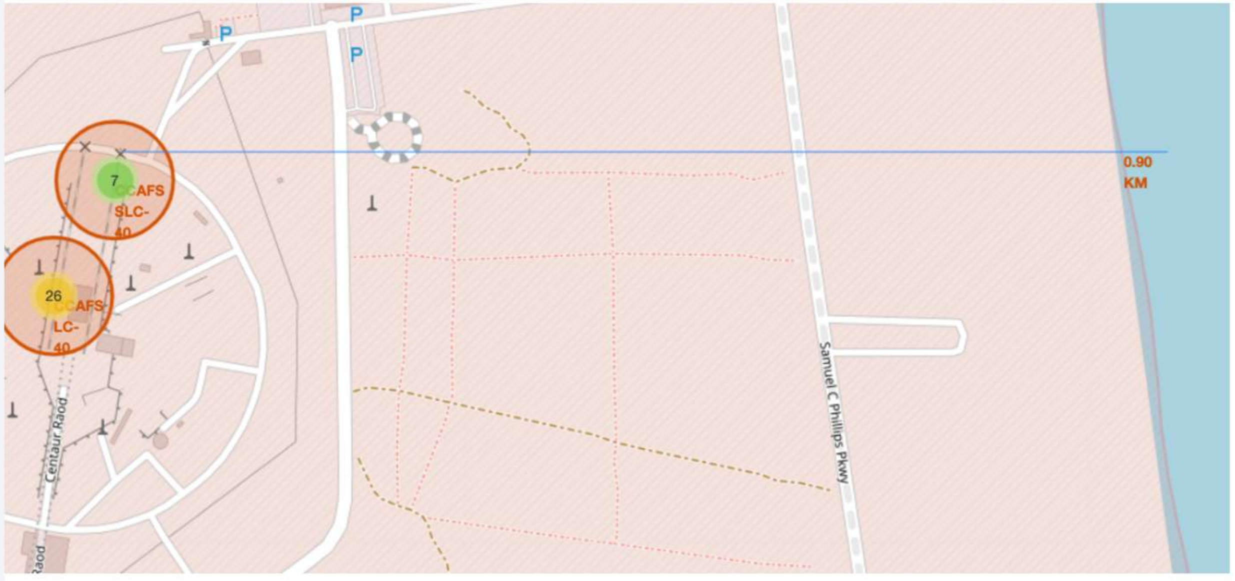
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map



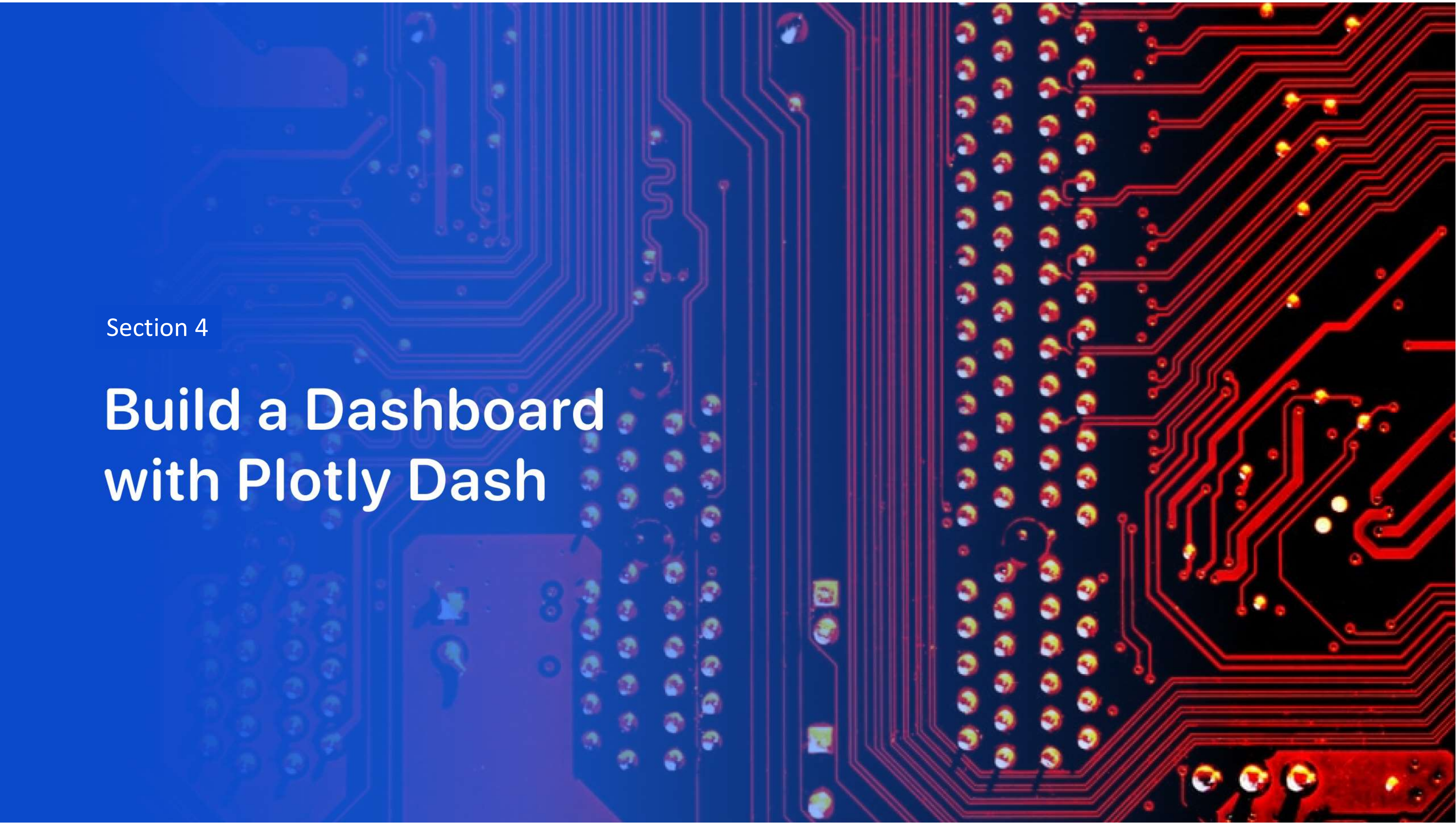


Distance between launch site and its proximities

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot



- The site is close to coast line to be able to transport various equipment
- It is outside the city and close to railway



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)

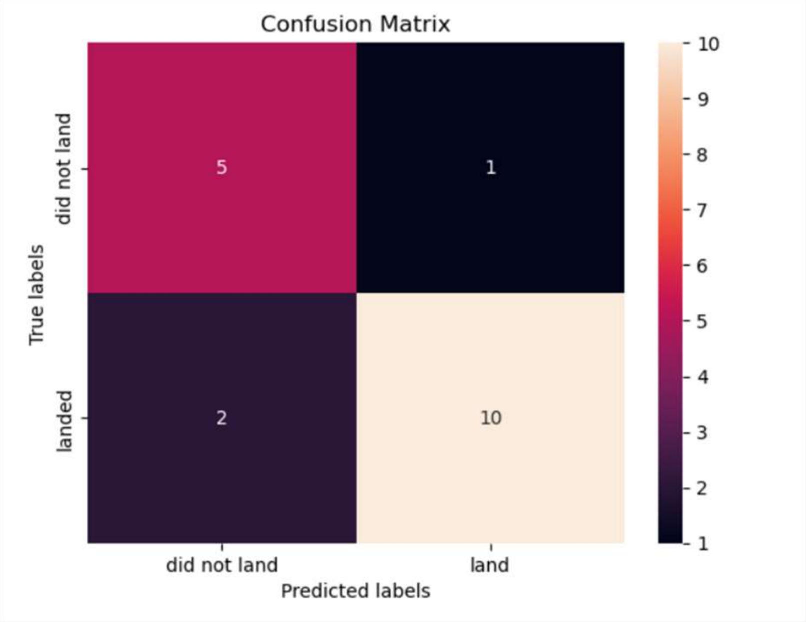
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy

- Best model is DecisionTree with tuned hpyerparameters :(best parameters) {'criterion': 'entropy', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'random’}
- And accuracy : 0.8875

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



Conclusions

- The more flights at a launch site, the greater the success rate. Launch success rate increased since 2013.
- Orbits ES-L1, GEO, HEO, SSO, VLEO are more successful.
- KSC LC-39A is the most successful launch site
- The Decision tree classifier is the best machine learning algorithm for this task. It can be used to predict the outcome.

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

Thank you!