IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jatin Kamboj
30th December, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - The below mentioned methodologies were used to analyze the data:
    1. Data extraction was done using Web Scraping from SpaceX API.
    2. Exploratory Data Analysis, data wrangling, visualization and Interactive visual analytics
    3. Predictive Analysis using Machine Learning

- Summary of all results
  - Exploratory Data Analysis Results
    1. Exploratory Data Analysis allpwed to identify features to predict successful launchings.
    2. Machine Learning Prediction allowed to choose best model for the prediction.

Introduction

- Project background and context
  - SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company capable of returning a spacecraft from low-Earth orbit, and in 2012 our Dragon spacecraft became the first commercial spacecraft to deliver cargo to and from the International Space Station. And in 2020, SpaceX became the first private company to take humans there as well. Based on the data provided by SpaceX API we are going to predict successful launces of rockets for Space Y Company.
- Problems you want to find answers
  - How do various measurable factors like weather, payload mass, launch site affect the probability of a successful rocket launch?
  - To estimate cost and viability of future launches
  - Best locations to make launches for Space Y company

Section 1

Methodology

Methodology

Executive Summary

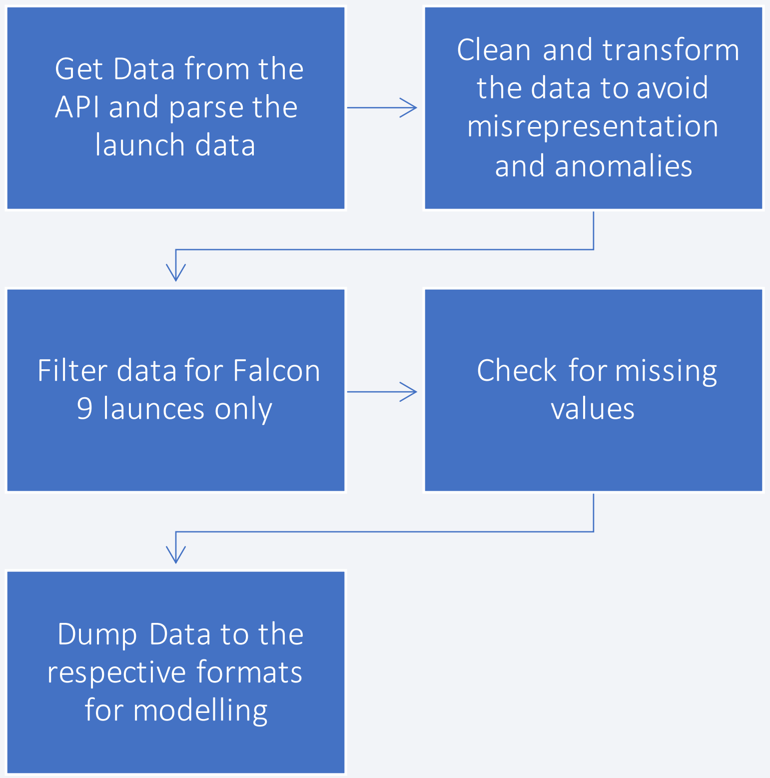
- Data collection methodology:
  - Describe how data was collected
  - Data from SpaceX was obtained from the following sources:
    1. SpaceX API – https://api.spacexdata.com/v4/rockets
    2. Wikipedia - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Describe how data was processed
  - Data was wrangled by creating a landing column based on outcome data after summarizing and analyzing the important features

Data Collection

- Describe how data sets were collected.
  - Data from SpaceX was obtained from the following sources:
    1. SpaceX API – https://api.spacexdata.com/v4/rockets
    2. Wikipedia - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Data collection process flowchart

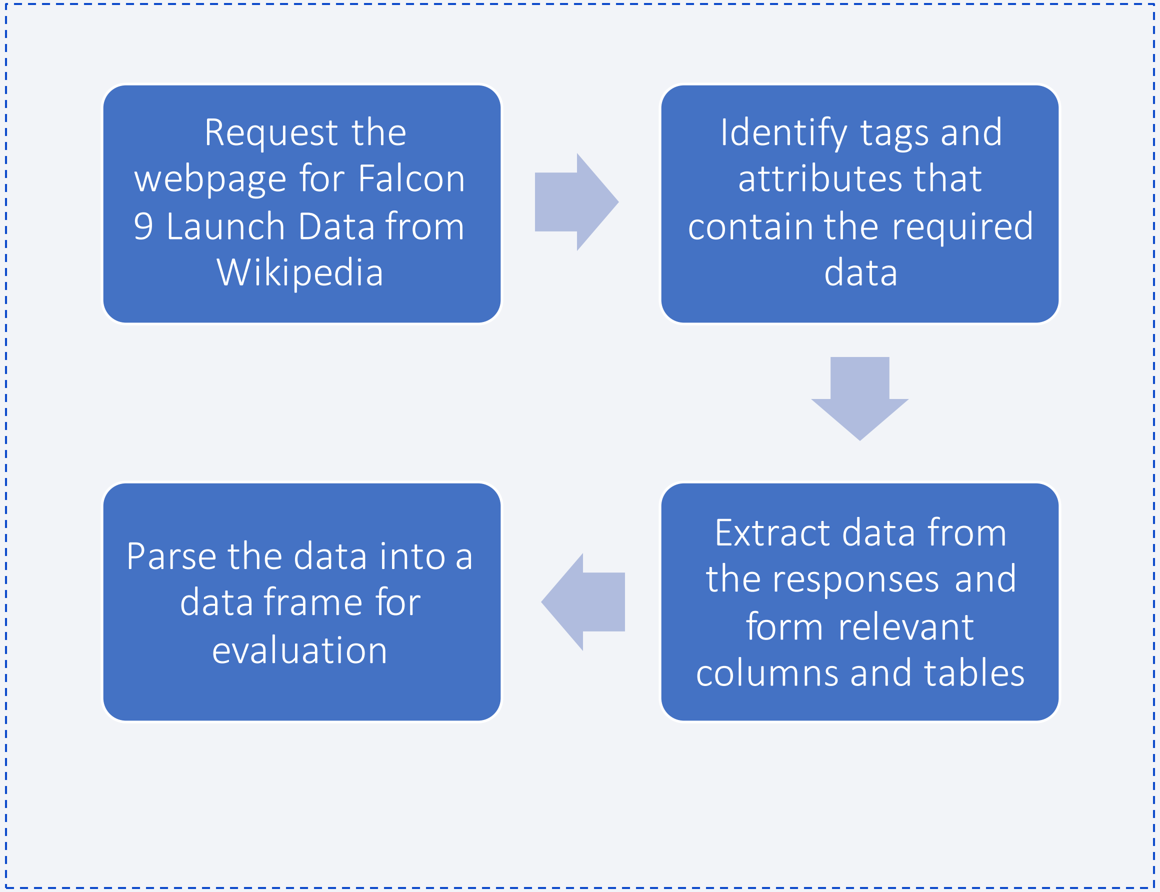
Data Collection – SpaceX API

- SpaceX offers a REST API from where data was obtained and used.
- The API was used according to the flowchart beside and then the data was persisted.
- GitHub URL of the completed SpaceX API calls notebook : https://github.com/MJK618/ibm-data-science-capstone/blob/master/jupyter-labs-spacex-data-collection-api.ipynb



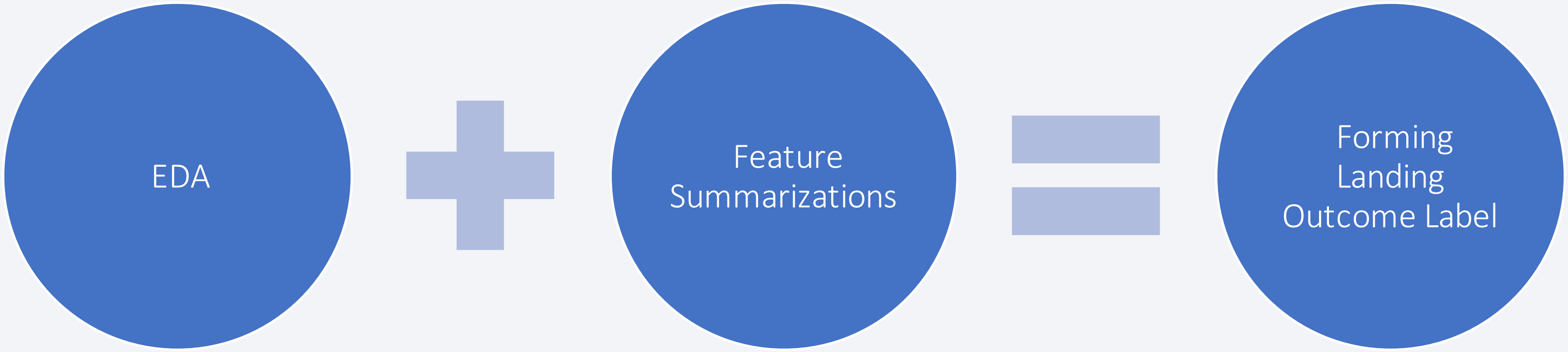
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia webpage via Scraping.
- Data was downloaded from Wikipedia website according to the flowchart and then was pursued for analysis.
- GitHub URL of the completed web scraping notebook : https://github.com/MJK618/ibm-data-science-capstone/blob/master/jupyter-labs-webscraping.ipynb



Data Wrangling

- Describe how data were processed
  - In beginning some Exploratory Data Analysis (EDA) was performed on the dataset.
  - Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
  - Subsequently, the landing outcome label was created from Outcome column.
- You need to present your data wrangling process using key phrases and flowcharts

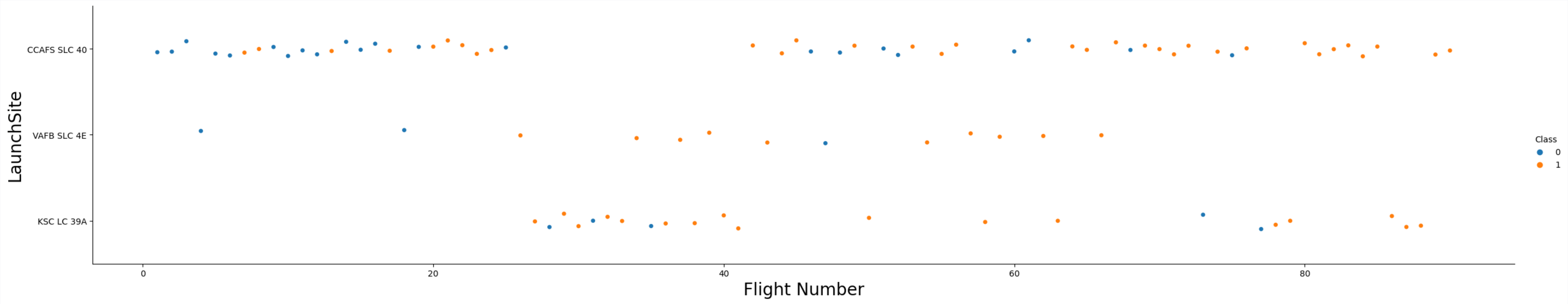


Data Wrangling

- GitHub URL of completed data wrangling related notebooks : https://github.com/MJK618/ibm-data-science-capstone/blob/master/jupyter-labs-data-wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- For exploring the data barplots and scatterplots were used to visualize the relationship between numerous features
- Launch Site X Payload Mass, Payload Mass X Flight Number, Launch Site X Flight Number, Orbit and Flight Number, Payload and Orbit



EDA with Data Visualization

- GitHub URL of completed EDA with data visualization notebook
  : https://github.com/MJK618/ibm-data-science-capstone/blob/master/jupyter-labs-eda-with-data-viz.ipynb

EDA with SQL

- The below SQL queries were performed:
  - Names of the unique launch sites in the space mission,
  - Top 5 launch sites whose name begin with the string 'CCA',
  - Total payload mass carried by boosters launched by NASA (CRS),
  - Average payload mass carried by booster version F9 v1.1,
  - Date when the first successful landing outcome in ground pad was achieved,
  - Total number of successful and failure mission outcomes,
  - Names of the booster versions which have carried the maximum payload mass,
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015, and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or success between dates

- GitHub URL of completed EDA with SQL notebook : https://github.com/MJK618/ibm-data-science-capstone/blob/master/jupyter-

Build an Interactive Map with Folium

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

- GitHub URL of completed interactive map with Folium map
  : https://github.com/MJK618/ibm-data-science-capstone/blob/master/jupyter-labs-interactive-viz-analytics-with-folium-lab.ipynb

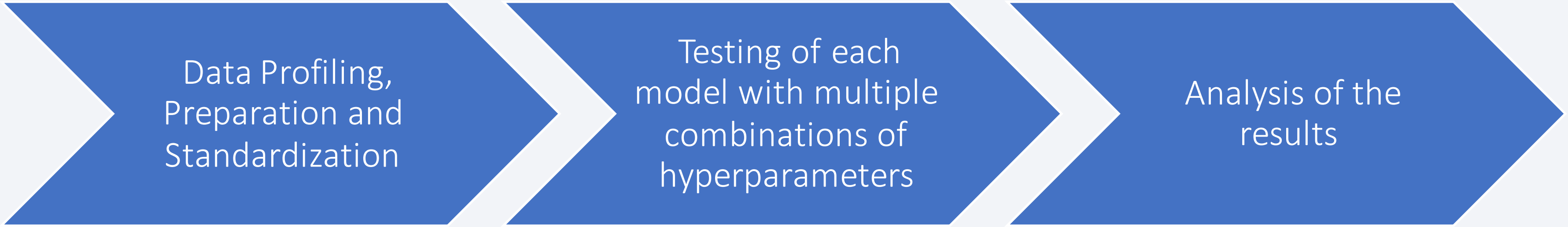
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
    - % of launches by site
    - Payload Range
    - The above combination allowed to briefly analyze the relationship between payloads and launch sites, which helped to identify the best places to launch according to payloads
- GitHub URL of completed Plotly Dash lab : https://github.com/MJK618/ibm-data-science-capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared:

    1. *Logistic Regression,*
    2. *Support Vector Machine,*
    3. *Decision Tree, and*
    4. *K Nearest Neighbors (KNN).*



Predictive Analysis (Classification)

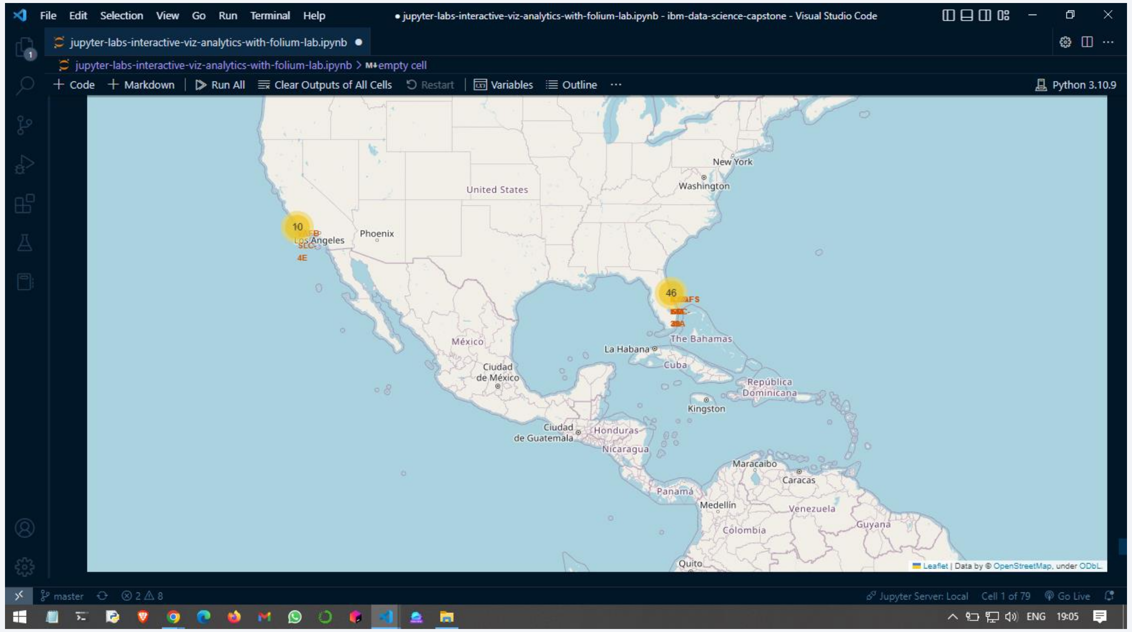
- GitHub URL of completed predictive analysis lab
  : https://github.com/MJK618/ibm-data-science-capstone/blob/master/jupyter-labs-machine-learning-prediction.ipynb

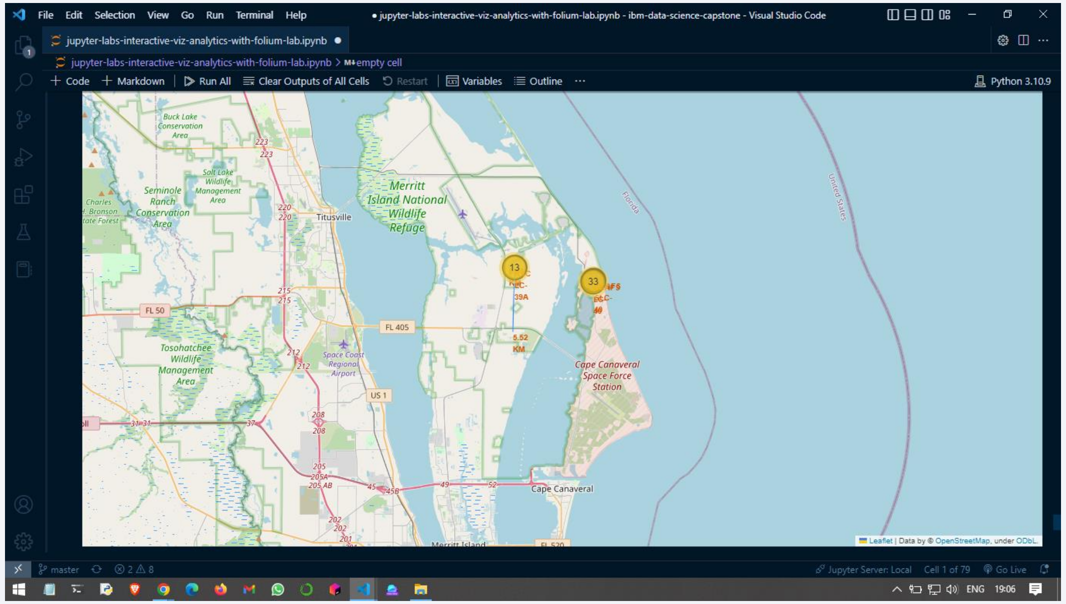
Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.17 Results

Results

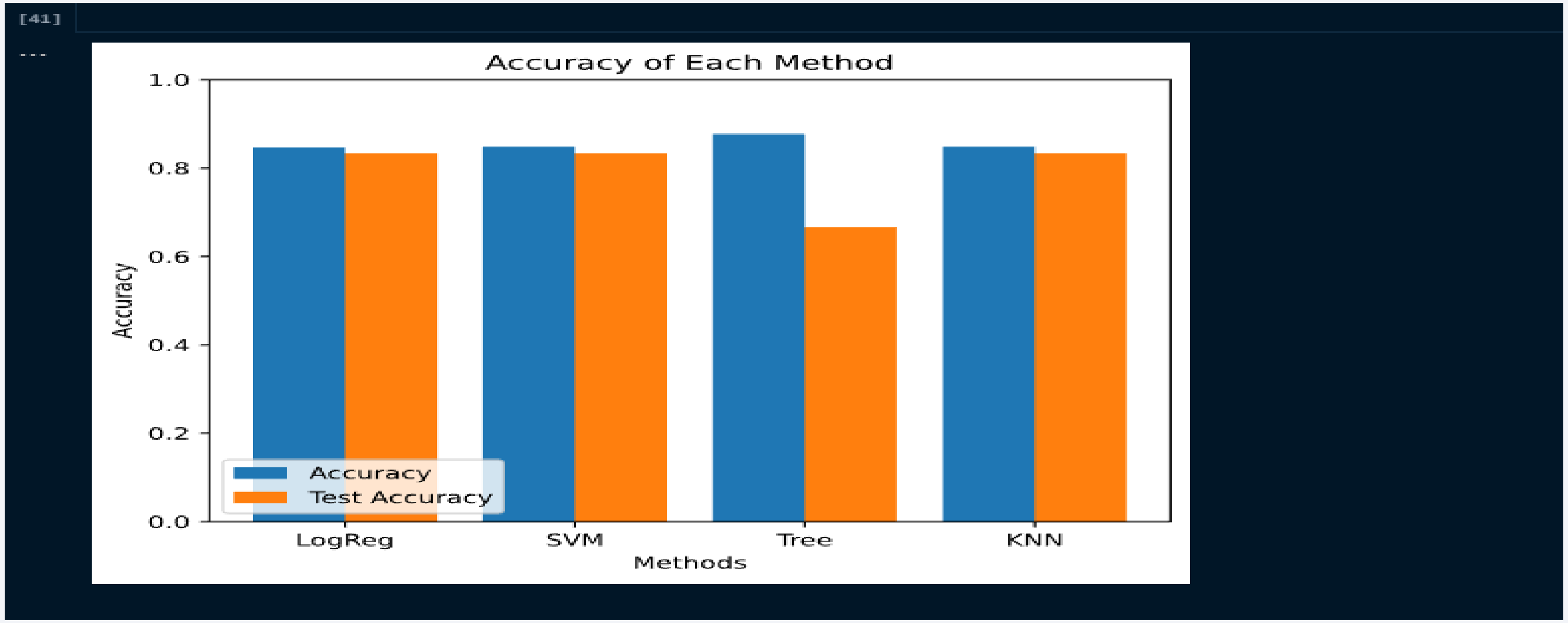
- Using the interactive analysis, it was possible to identify that launch sites use to be in safety places, near sea, for example and having a good logistic infrastructure around
- Most of the successful launches happened on the east coast side.





Results

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87.67% and accuracy for test data over 94%

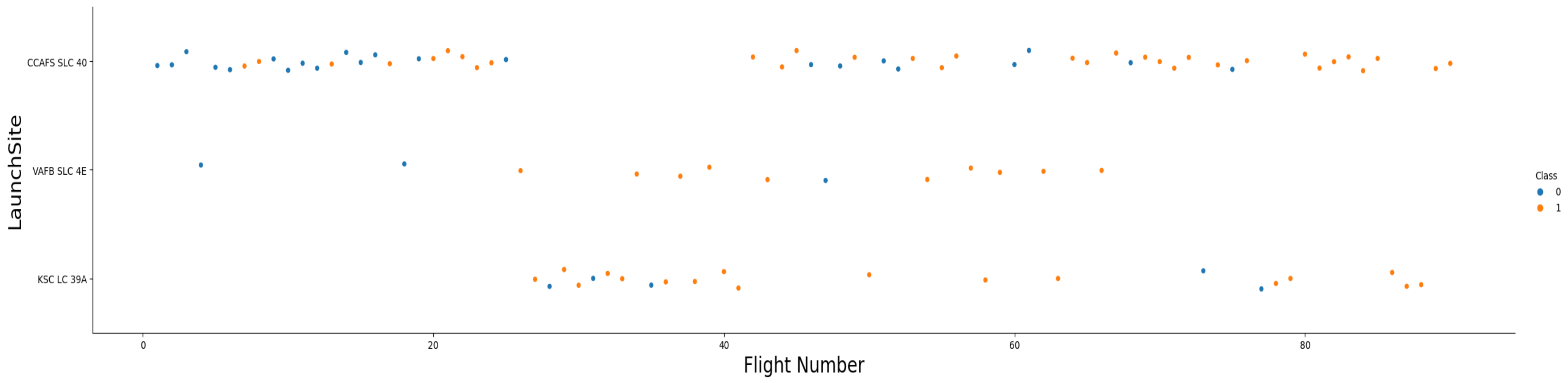


Section 2

Insights drawn from EDA

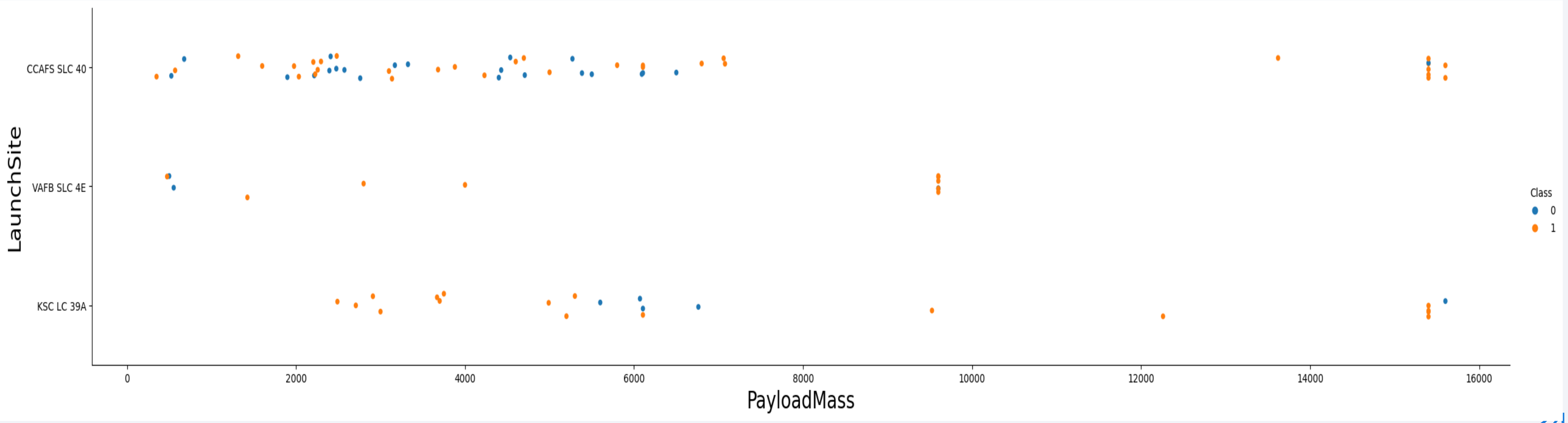
Flight Number vs. Launch Site

- Show the screenshot of the scatter plot with explanations
  - According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful,
  - In second place VAFB SLC 4E and third place KSC LC 39A,
  - It's also possible to see that the general success rate improved over time.



Payload vs. Launch Site

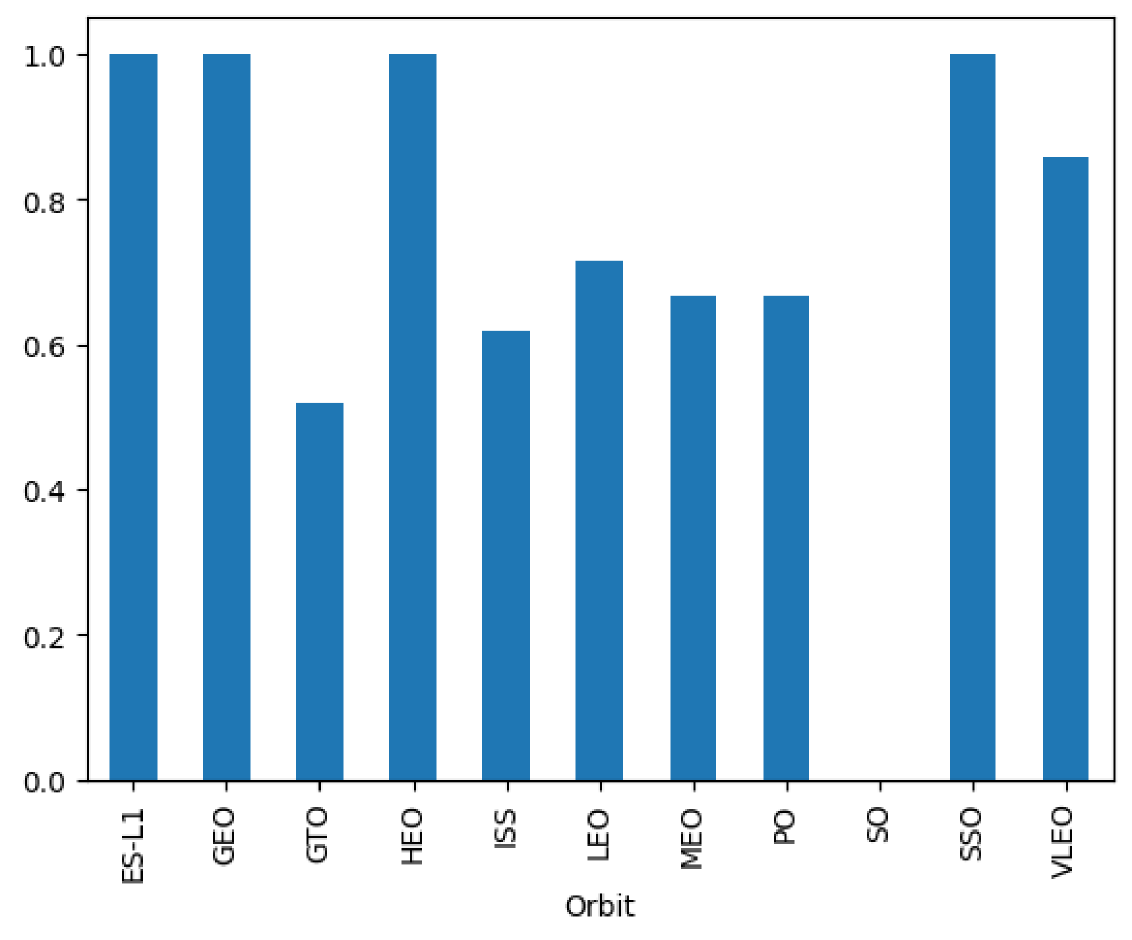
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate,
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.
- Show the screenshot of the scatter plot with explanations



Success Rate vs. Orbit Type

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

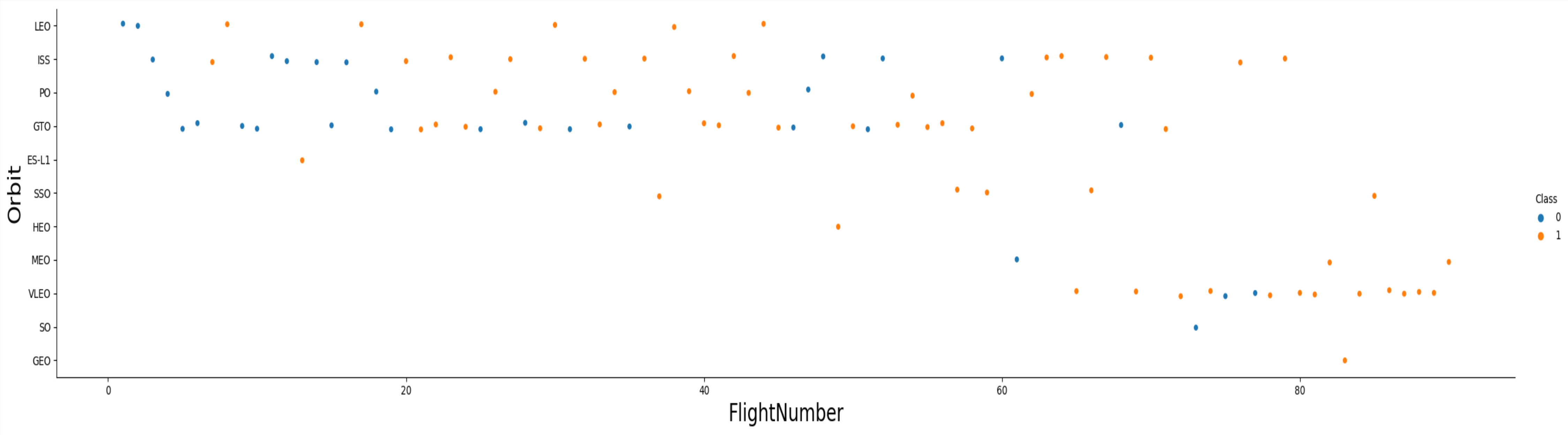
Show the screenshot of the scatter plot with explanations



Flight Number vs. Orbit Type

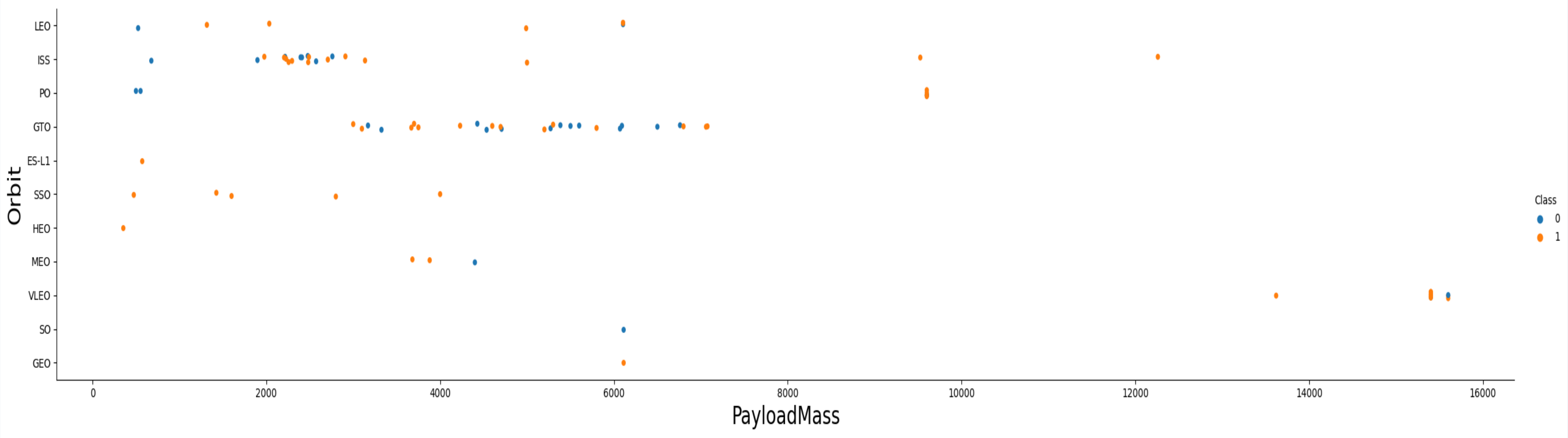
- Apparently, success rate improved over time to all orbits,
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency

- Show the screenshot of the scatter plot with explanations



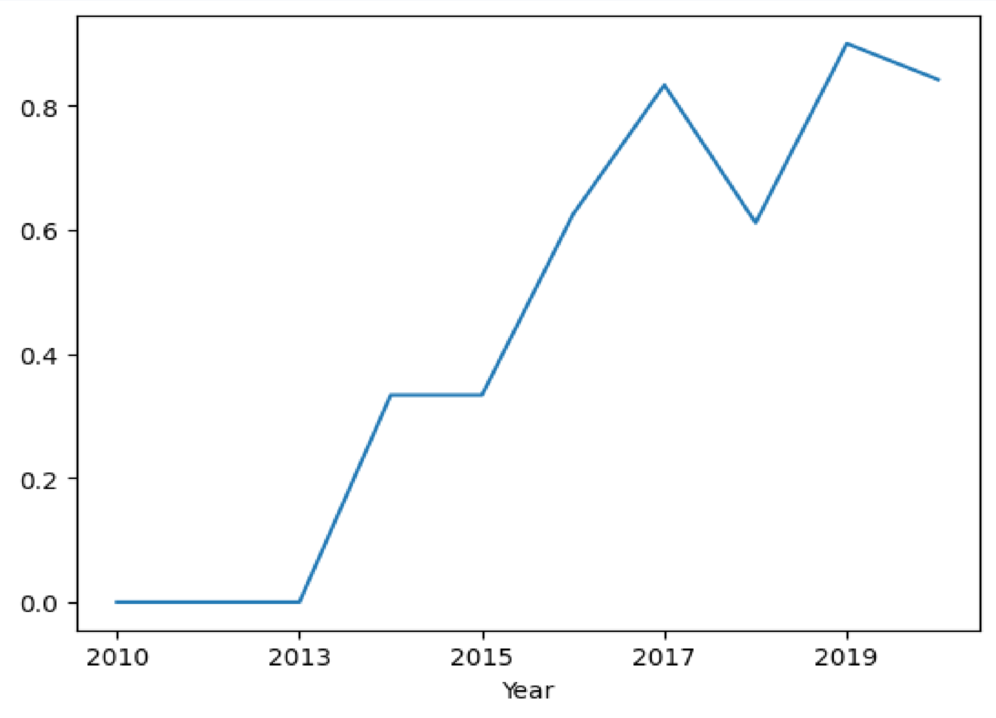
Payload vs. Orbit Type

- Apparently, there is no relation between payload and success rate to orbit GTO,
- ISS orbit has the widest range of payload and a good rate of success,
- There are few launches to the orbits SO and GEO.
- Show the screenshot of the scatter plot with explanations



Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020,
- It seems that the first three years were a period of adjusts and improvement of technology.
- Show the screenshot of the scatter plot with explanations



All Launch Site Names

- According to data, there are four launch sites:
  - They are obtained by selecting unique occurrences of “launch_site” values from the dataset

Task 1

Display the names of the unique launch sites in the space mission

```
[8]
1  sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu01qde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- Here we can see five samples of Cape Canaveral launches.

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

[9]

Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	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) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- The total payload carried by boosters from NASA
- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

[10]

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

total_payload
111268

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1
- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

avg_payload
2928

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad
- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 2015-12-22.

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

first_success_gp
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Selecting distinct booster versions according to the filters above, these 4 are the result.

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUT
```

[14] Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
- Grouping mission outcomes and counting records for each group led us to the summary above

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

[15] Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
- These are the boosters which have carried the maximum payload mass registered in the dataset

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSI
```

[23] Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Only two occurrences are there.

Task 9

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

```
[24]
1  sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;
                                                                                                                                   Python
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

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

Task 10

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

```
[25]
sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY Q
```

Python

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

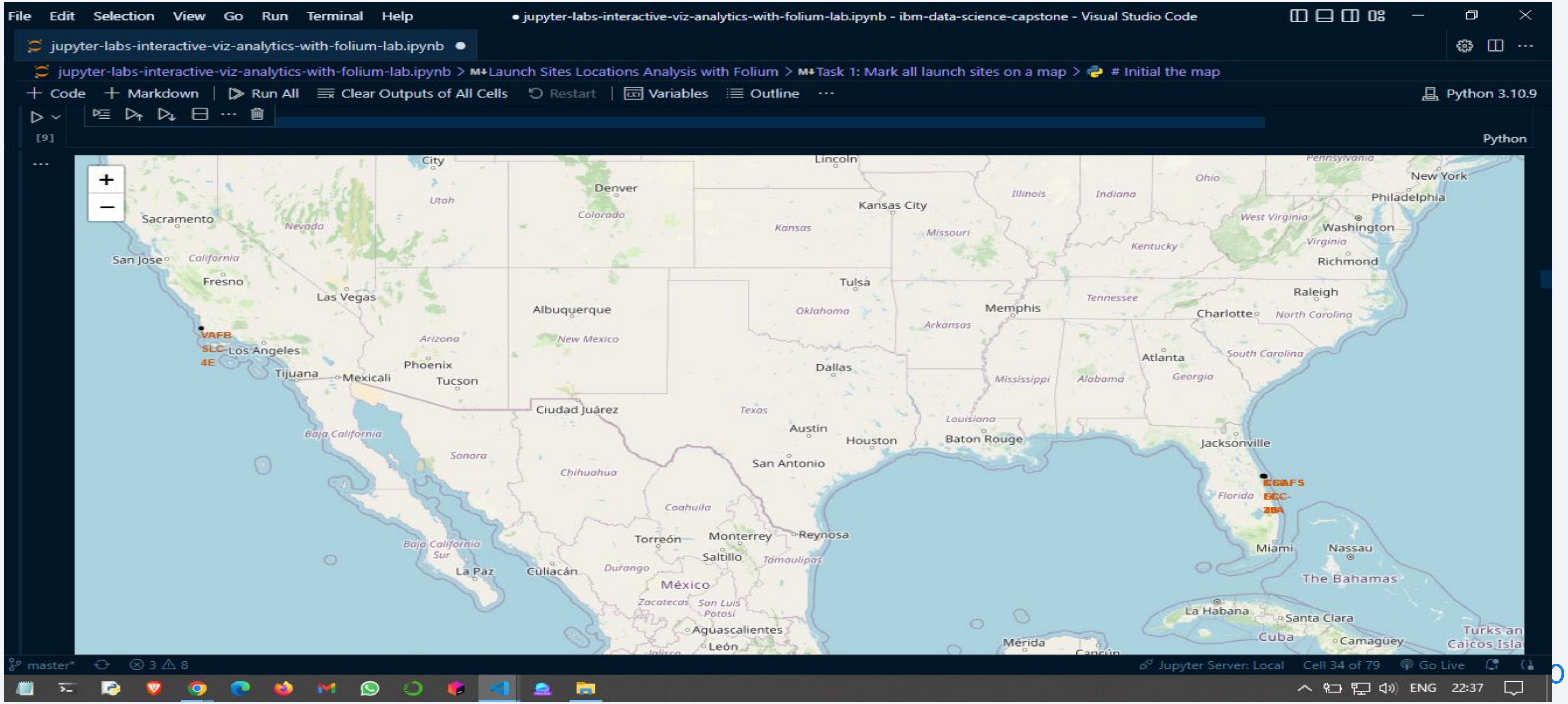
landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

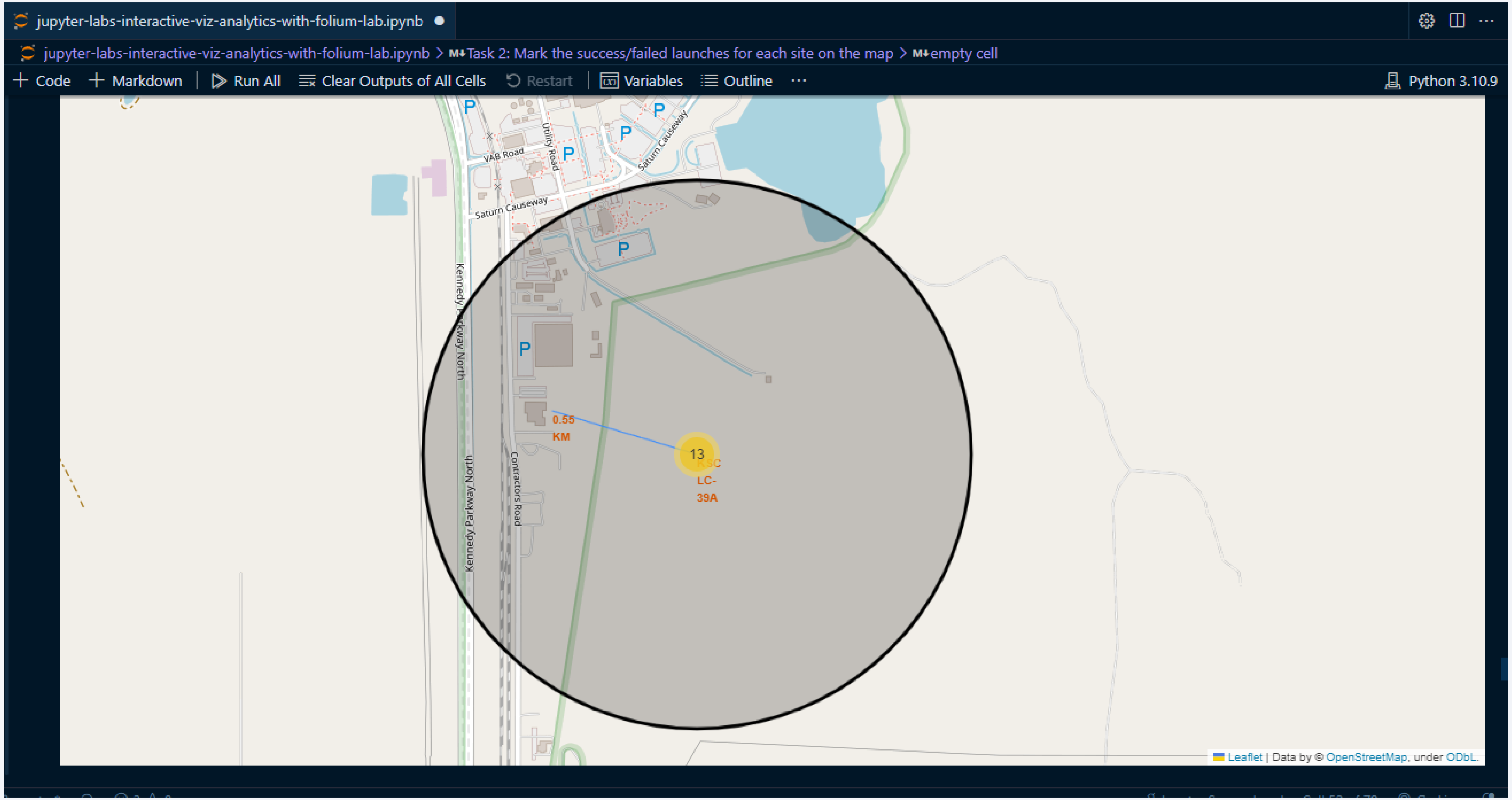
All of the Launch Sites on Map

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



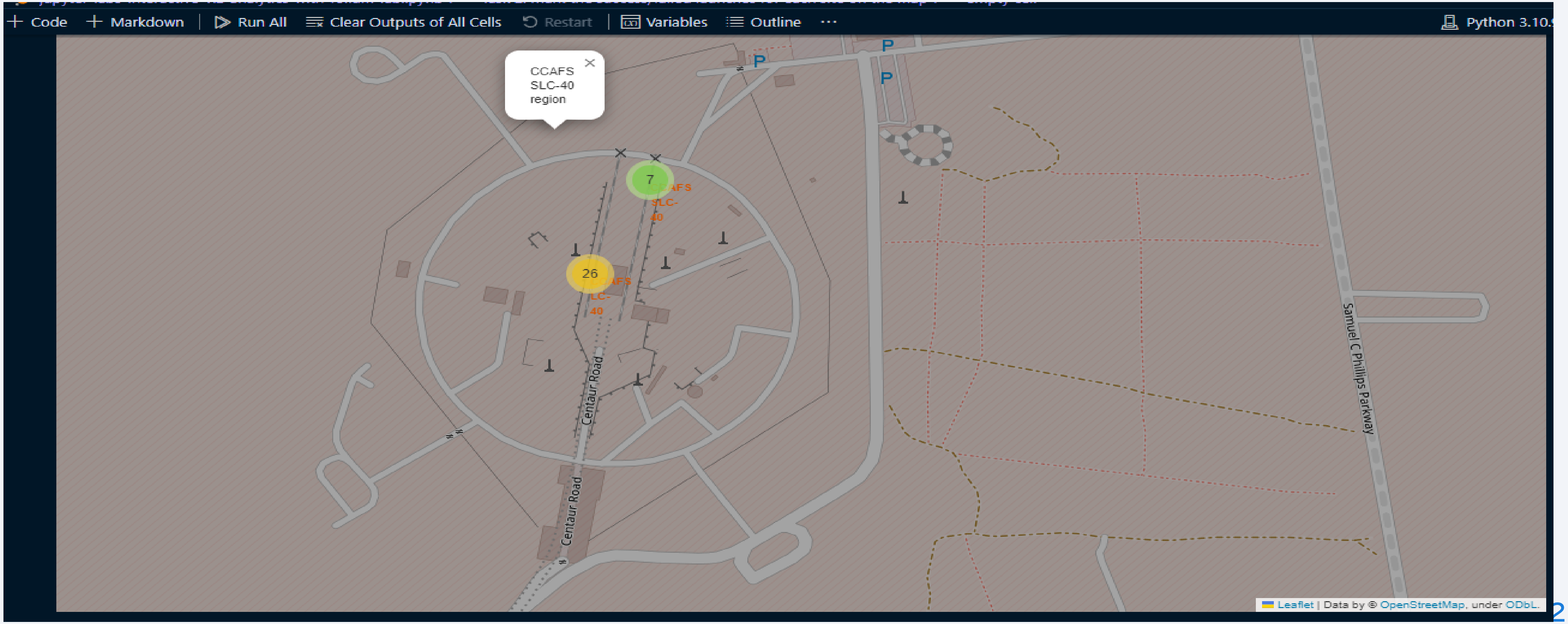
Launch Outcomes by Site

- Example site launch outcomes



Safety and Logistics

- Generated folium map and the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

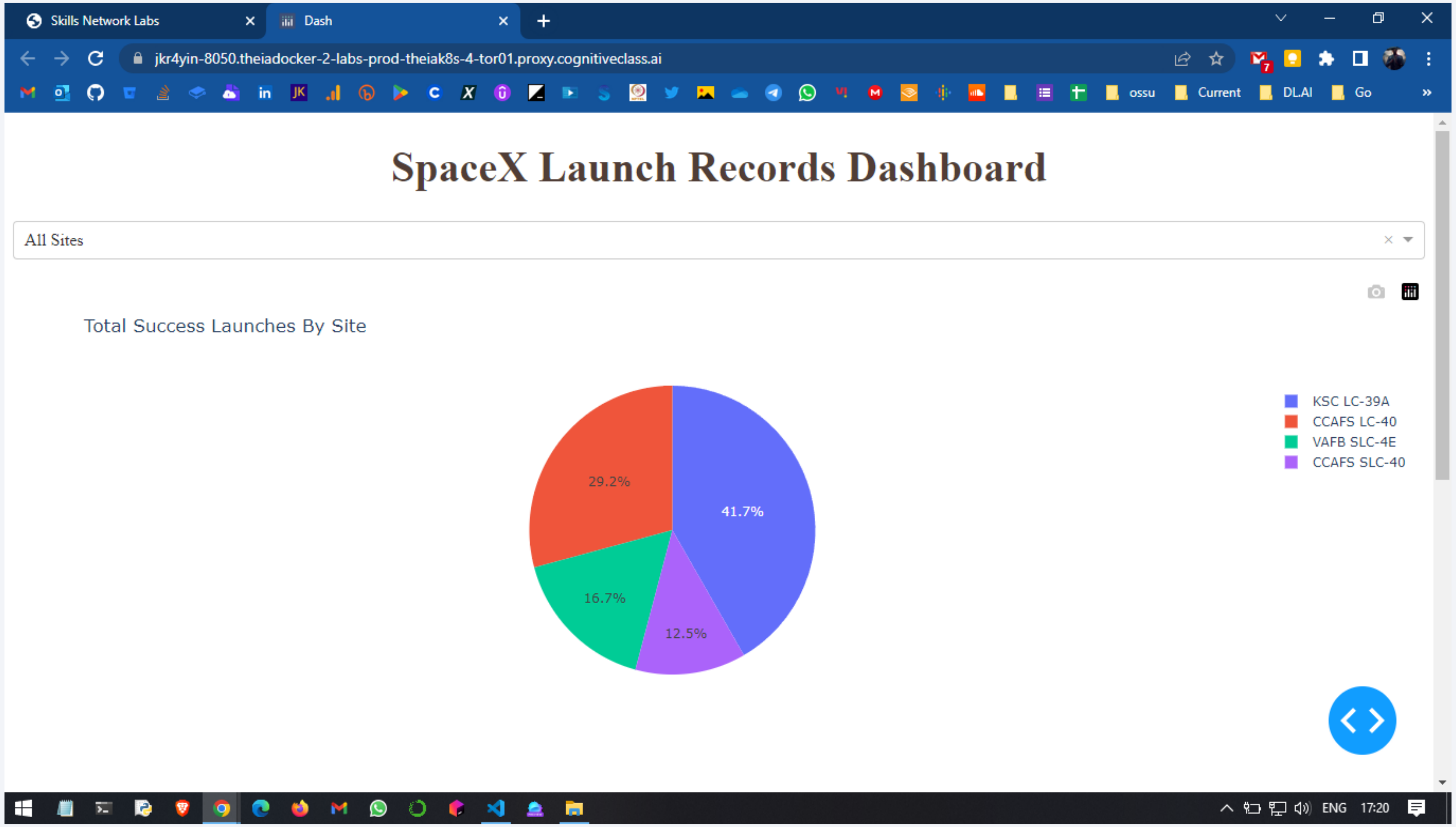


Section 4

Build a Dashboard with Plotly Dash

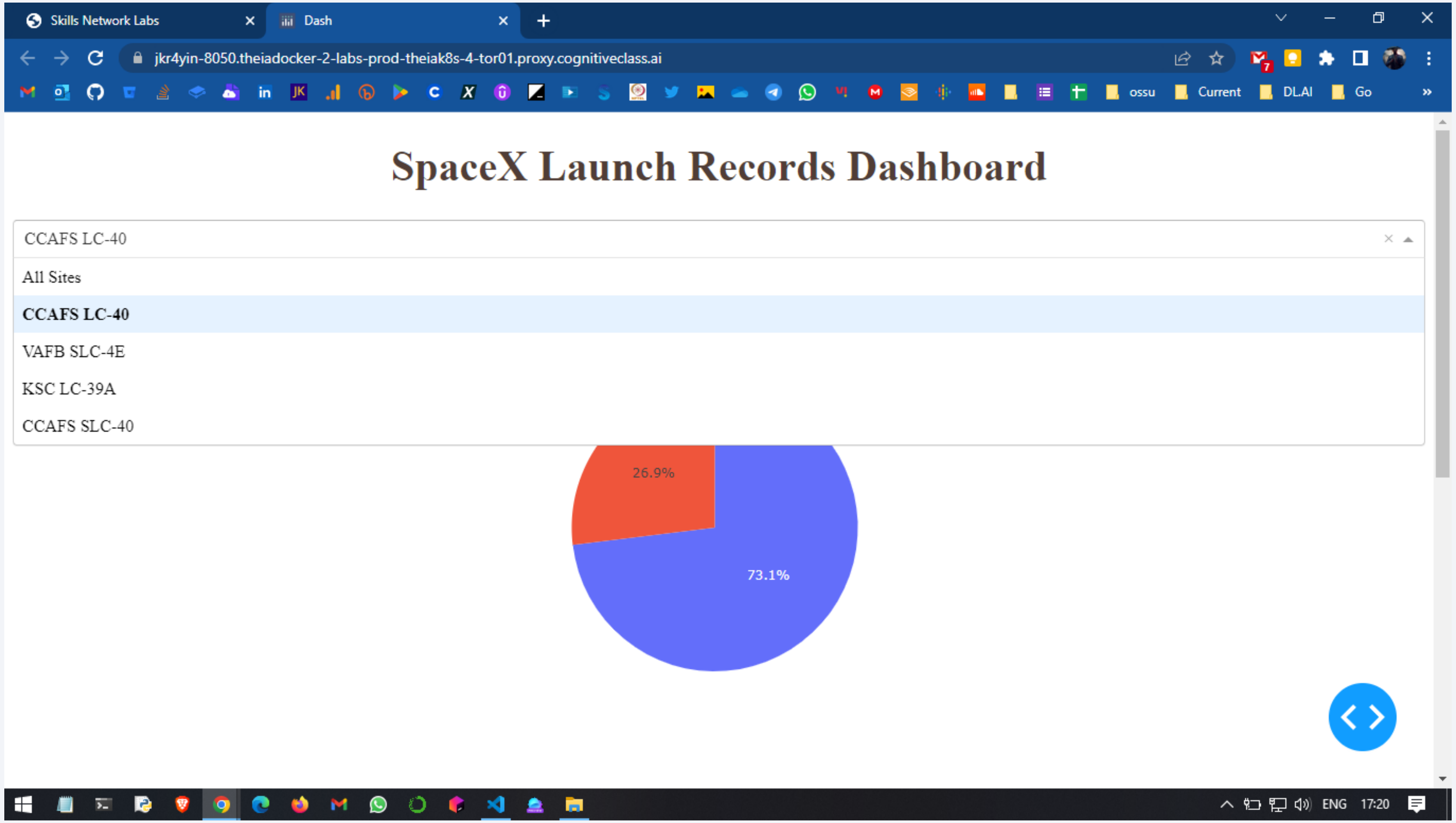
Successful Launches by Site

The place from where launches are done seems to be a very important factor of success of missions



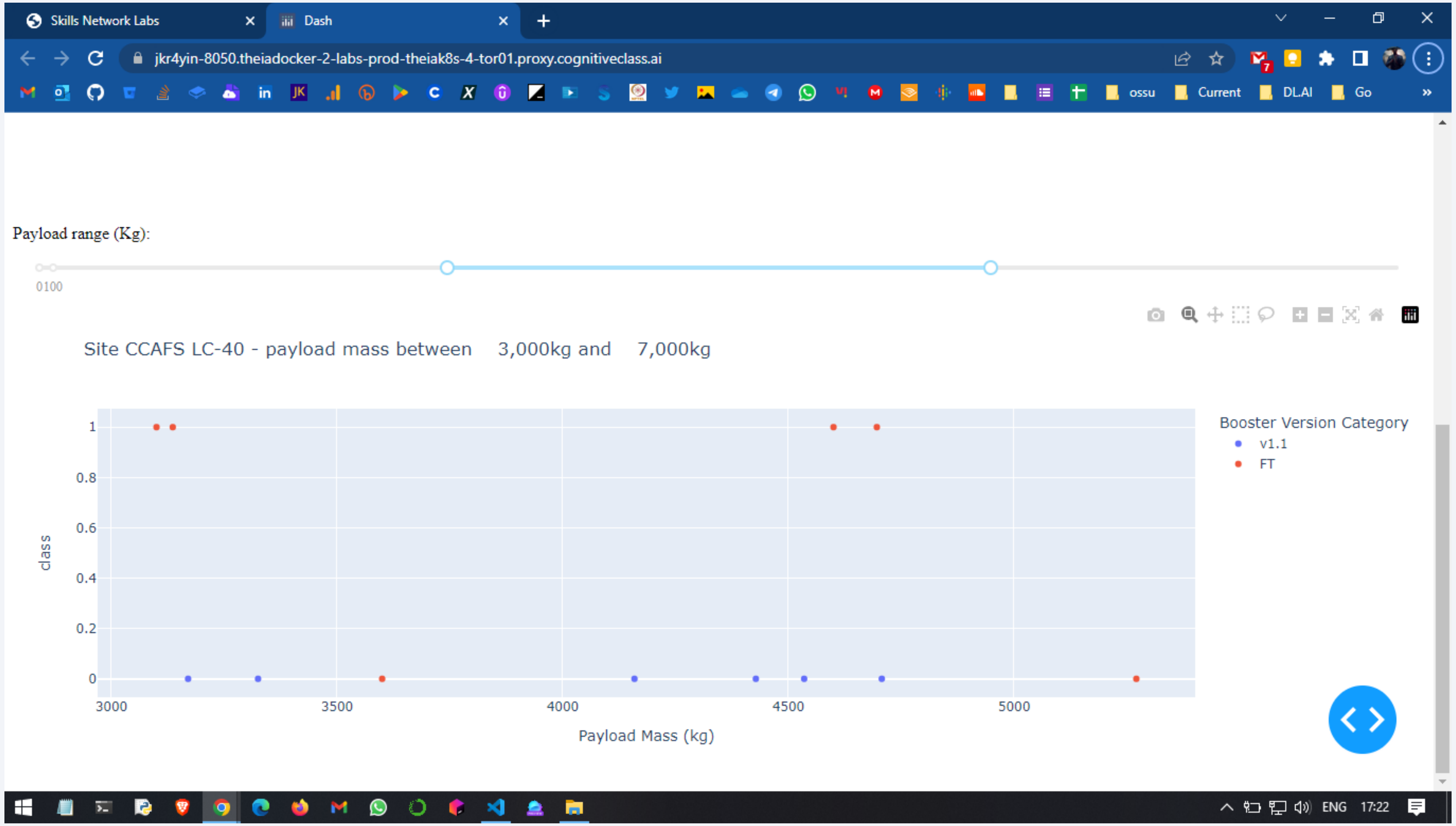
Success Ratio for a Launch

- 73.1% of launches are successful in this site.



Payload vs Launch Outcome

- Payloads under 6,000kg and FT boosters are the most successful combination.

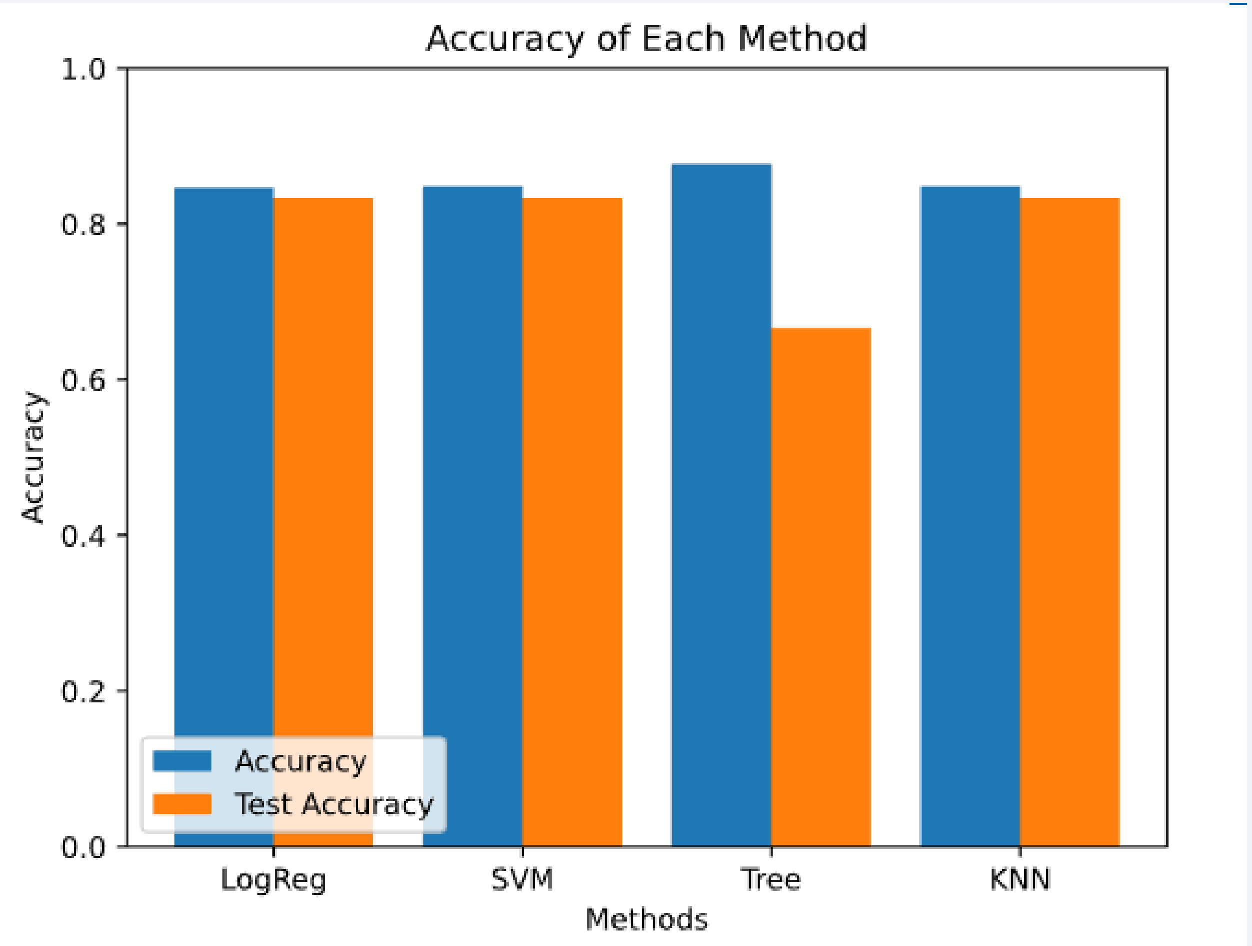


Section 5

Predictive Analysis (Classification)

Classification Accuracy

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



Confusion Matrix of Decision Tree Classifier

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

Conclusions

- Two different data sources were analyzed to refine the conclusions along the process,
- The most suitable launch site is KSC LC-39A in terms of over all success rate.
- Launches that have payload above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets,
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!