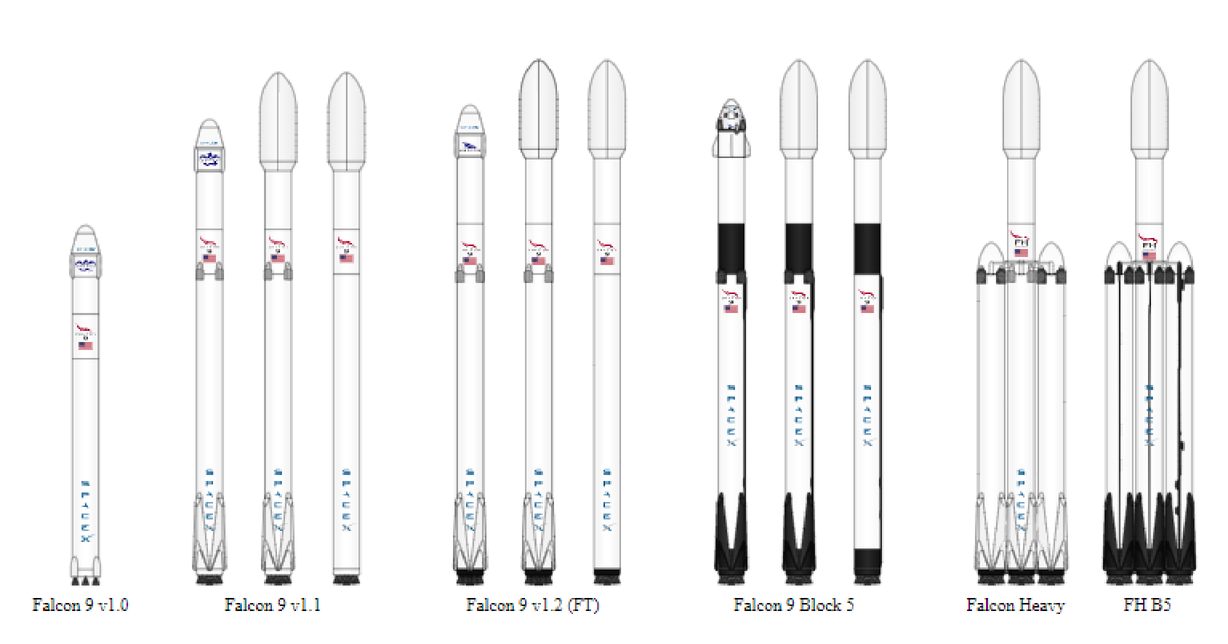
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hariharan Loganathan

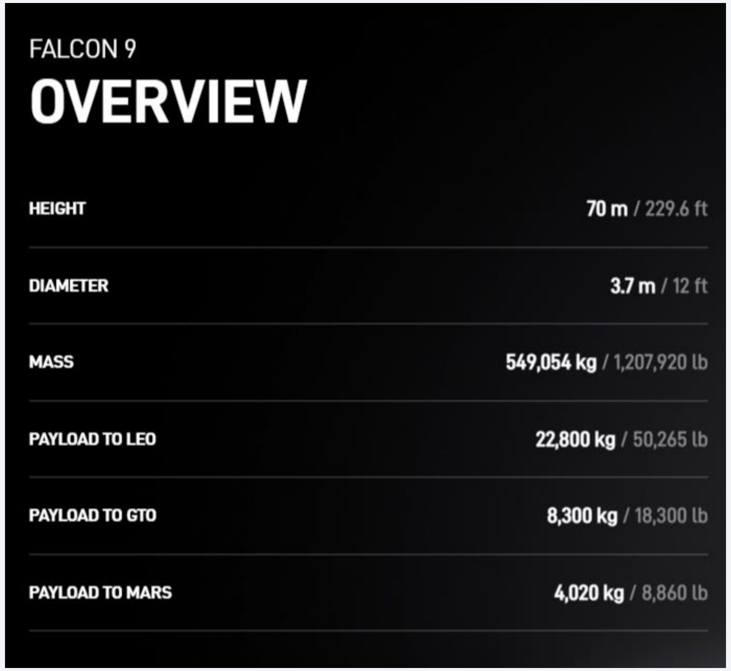
24/04/2023

TABLE OF CONTENTS



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

Executive Summary



FALCON 9

OVERVIEW

HEIGHT	**70 m** / 229.6 ft
DIAMETER	**3.7 m** / 12 ft
MASS	**549,054 kg** / 1,207,920 lb
PAYLOAD TO LEO	**22,800 kg** / 50,265 lb
PAYLOAD TO GTO	**8,300 kg** / 18,300 lb
PAYLOAD TO MARS	**4,020 kg** / 8,860 lb

- SpaceY is a recently established player in the commercial rocket launch industry, aiming to compete with SpaceX by bidding against them for launch contracts.
- SpaceX advertises launch services starting at $67 million, which includes reserving fuel for reusing the first stage rocket booster.
- According to public statements from SpaceX, the cost of building a first stage Falcon 9 rocket booster is estimated to be over $15 million, not including R&D expenses or profit margin.
- The report shows that by considering mission parameters like payload mass and desired orbit, the models were able to predict the first stage rocket booster landing with an accuracy level of 83.3%.
- Using first stage landing predictions as a cost proxy for launches, SpaceY can make more informed bids against SpaceX.

Introduction - **BACKGROUND**

- As part of the Applied Data Science Capstone course, this report has been created to assist SpaceY, a new rocket company.
- As part of the Applied Data Science Capstone course, I am taking the role of a data scientist employed by SpaceY, a newly established company in the rocket industry.
- By utilizing the data science findings and models presented in this report, SpaceY can make better-informed bids against SpaceX for rocket launch contracts.

Introduction – Business Problem

- When the first stage of their rockets is reusable, SpaceX promotes Falcon 9 rocket launches at a cost of 67 million dollars.
- According to public statements by SpaceX, the estimated cost of building the first stage of Falcon 9 rocket is over 15 million dollars, excluding the expenses of R&D cost recovery or profit margin.
- On certain occasions, SpaceX may forgo the reuse of the first stage of the rocket, depending on the mission parameters, such as the payload, orbit requirements, and customer preferences.
- The main objective of this report is to provide an accurate prediction of the probability of a successful landing of the first stage rocket. This prediction can serve as a substitute for estimating the cost of a launch, thereby helping SpaceY to make better-informed bids against SpaceX.

IBM Developer SKILLS NETWORK

Section 1

Methodology

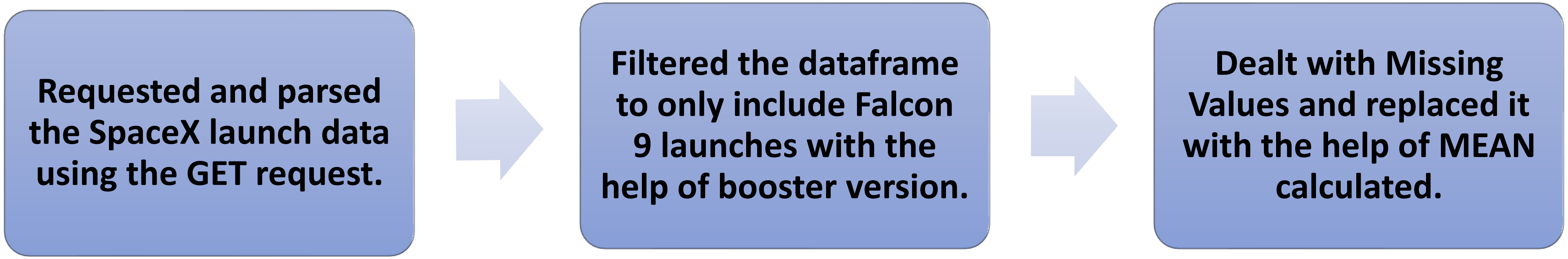
Methodology

In this report, methodology used can be outlined as such:

1. **Data collection**
2. **Data wrangling**
3. **Exploratory data analysis**
4. **Data visualization with SQL and Visualizations**
5. **Building an interactive dashboard using Plotly.**
6. **Model development – Predictive analysis using classification models**
7. **Reporting results to stakeholders.**

Data Collection – SpaceX API

- Acquired historical launch data from Open Source REST API for SpaceX.



Requested and parsed the SpaceX launch data using the GET request.

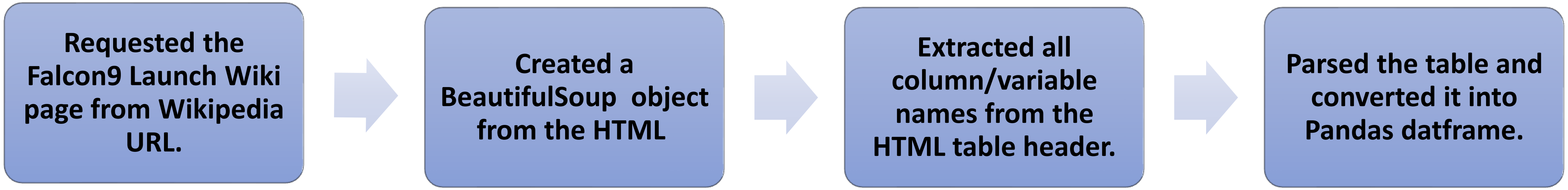
Filtered the dataframe to only include Falcon 9 launches with the help of booster version.

Dealt with Missing Values and replaced it with the help of MEAN calculated.

IBM Developer SKILLS NETWORK

Data Collection – Web Scraping

- Acquired historical launch data from Wikipedia page 'List of Falcon 9 and Falcon Heavy Launches'



Requested the Falcon9 Launch Wiki page from Wikipedia URL.

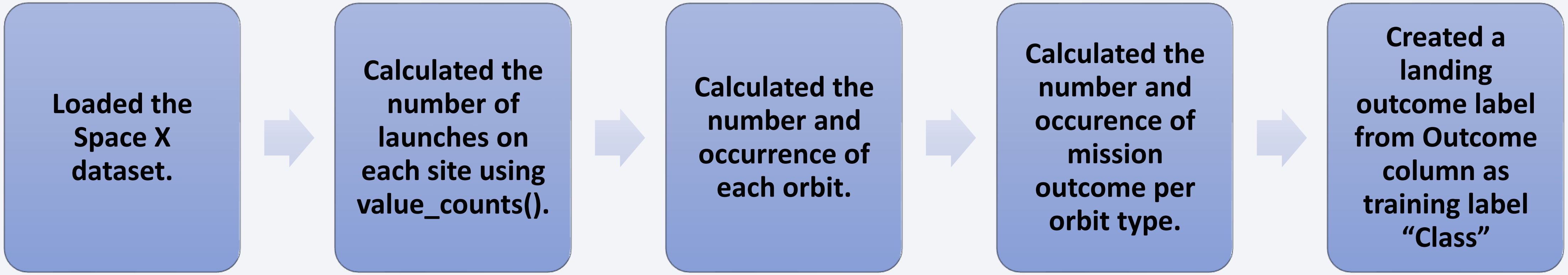
Created a BeautifulSoup object from the HTML

Extracted all column/variable names from the HTML table header.

Parsed the table and converted it into Pandas datframe.

Data Wrangling

- Performed exploratory Data Analysis and determined Training Labels.



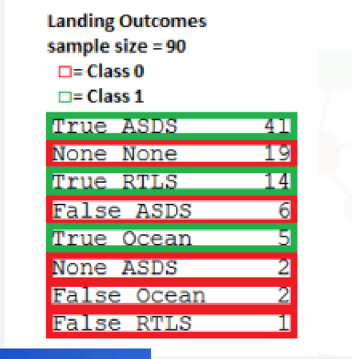
Loaded the Space X dataset.

Calculated the number of launches on each site using value_counts().

Calculated the number and occurrence of each orbit.

Calculated the number and occurence of mission outcome per orbit type.

Created a landing outcome label from Outcome column as training label “Class”



Landing Outcomes
sample size = 90
□= Class 0
□= Class 1

```
True  ASDS     41
None  None     19
True  RTLS     14
False ASDS      6
True  Ocean     5
None  ASDS      2
False Ocean     2
False RTLS      1
```

Class = 0; first stage booster did not land successfully

- None None; not attempted
- None ASDS; unable to be attempted due to launch failure
- False ASDS; drone ship landing failed
- False Ocean; ocean landing failed
- False RTLS; ground pad landing failed

Class = 1; first stage booster landed successfully

- True ASDS; drone ship landing succeeded
- True RTLS; ground pad landing succeeded
- True Ocean; ocean landing succeeded

EDA with Data Visualization

- Read the dataset into a Pandas dataframe.

- Used Matplotlib and Seaborn visualization libraries to plot bar charts and scatter plot to differentiate the following.
  - FlightNumber x PayloadMass †
  - FlightNumber x LaunchSite †
  - Payload x LaunchSite †
  - Orbit type x Success rate
  - FlightNumber x Orbit type †
  - Payload x Orbit type †
  - Year x Success rate †

EDA with SQL

- Loaded data into an IBM DB2 instance
- Ran SQL queries to display and list information about
    1. Launch sites
    2. Payload masses
    3. Booster versions
    4. Mission outcomes
    5. Booster landings

IBM Developer SKILLS NETWORK

Build an Interactive Map with Folium

- Used Python interactive mapping library called Folium and used map objects such as markers, circles, lines, etc.,
- Marked all launch sites on a map
- Marked the successful/failed launches for each site on map
- Calculated the distances between a launch site to its proximities

    1. Railways
    2. Highways
    3. Coastlines
    4. Cities

Build a Dashboard with Plotly Dash

- Used Python interactive dashboarding library called Plotly Dash to enable stakeholders to explore and manipulate data in an interactive and real-time way
- Pie chart showing success rate
- Color coded by launch site
- Scatter chart showing payload mass vs. landing outcome
- Color coded by booster version
- With range slider for limiting payload amount
- Drop-down menu to choose between all sites and individual launch sites

Predictive Analysis (Model development - Classification)

- Imported libraries and defined function to create confusion matrix
- Loaded the data frame created during data collection
- Created a column for our training label 'Class' created during data wrangling
- Standardized the data
- Splitted the data into training data and test data
- Fitted the training data to various model types such as Logistic Regression , Support Vector Machine , Decision Tree Classifier , K Nearest Neighbors Classifier
- Used a cross-validated grid-search over a variety of hyperparameters to select the best ones for each model
- Enabled by Scikit-learn library function GridSearchCV
- Evaluated accuracy of each model using test data to select the best mode

Results

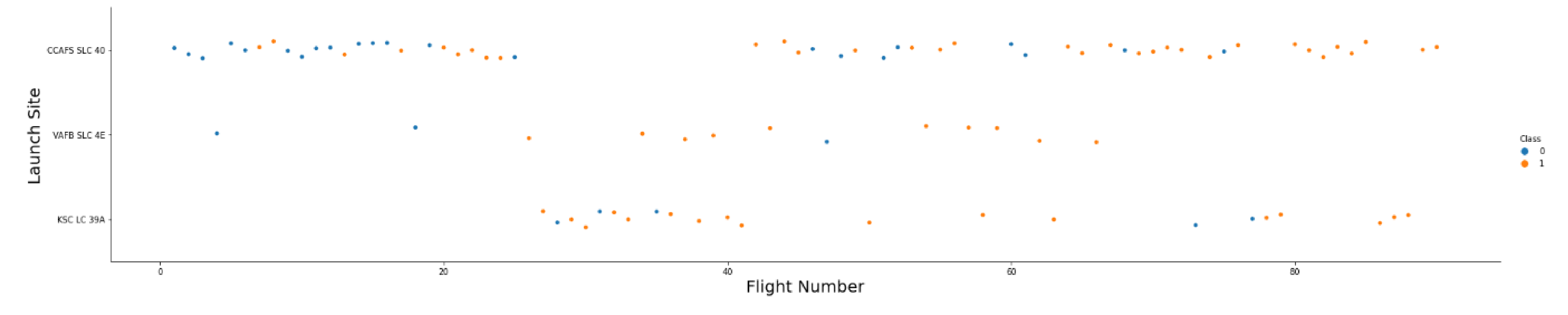
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

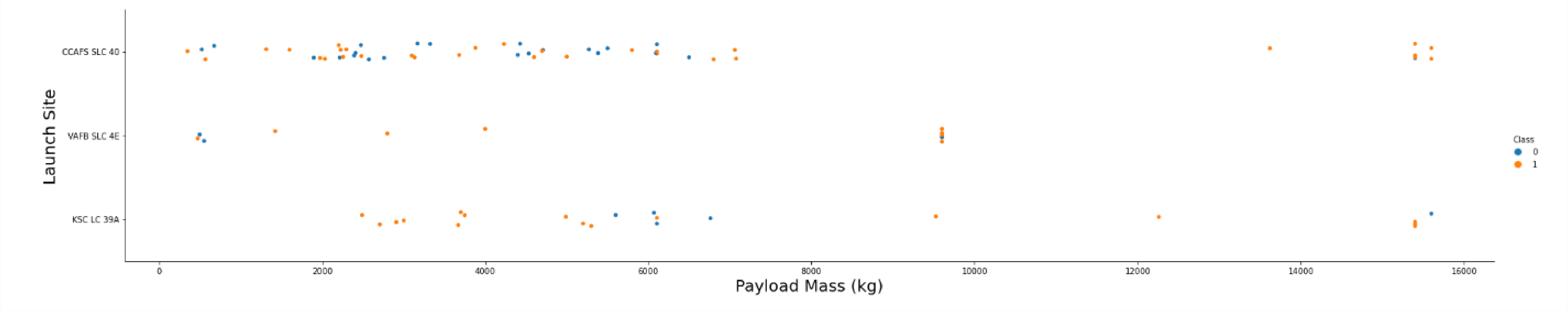
Results : EDA with Data Visualization

Flight Number vs. Launch Site



❑CCAFS SLC 40 appears to have been where most of the early 1st stage landing failures took place.

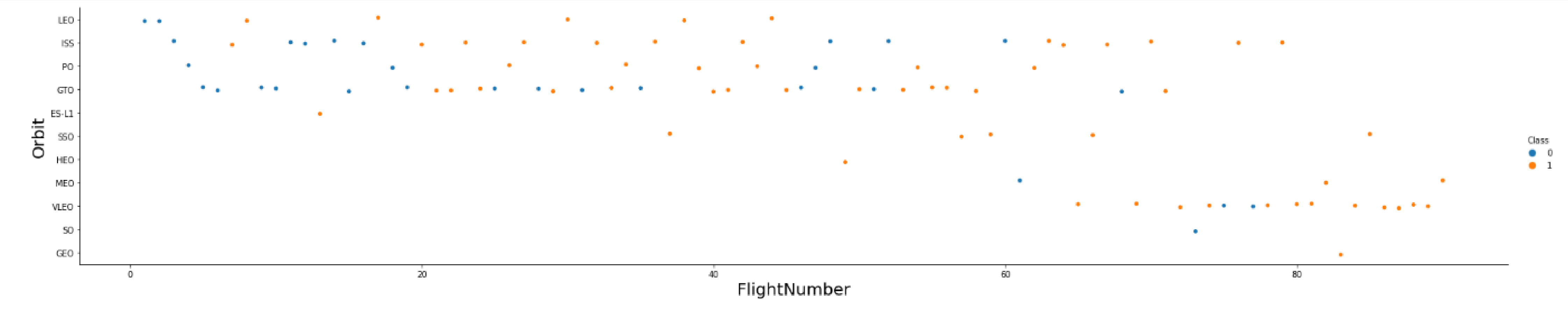
Payload vs. Launch Site



❑CCAFS SLC 40 and KSC LC 39A appear to be favored for heavier payloads.

IBM Developer
SKILLS NETWORK

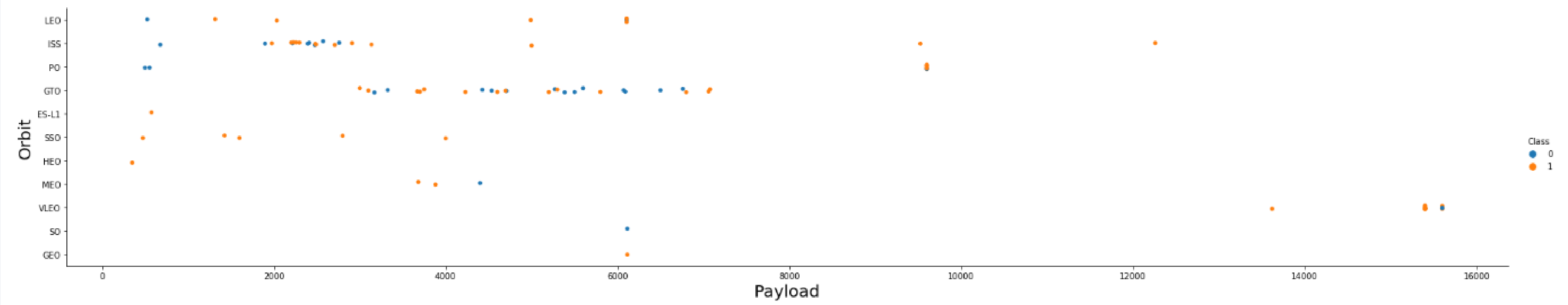
Flight Number vs. Orbit Type



- ❑Flight number positively correlated with 1st stage recovery for all orbit types

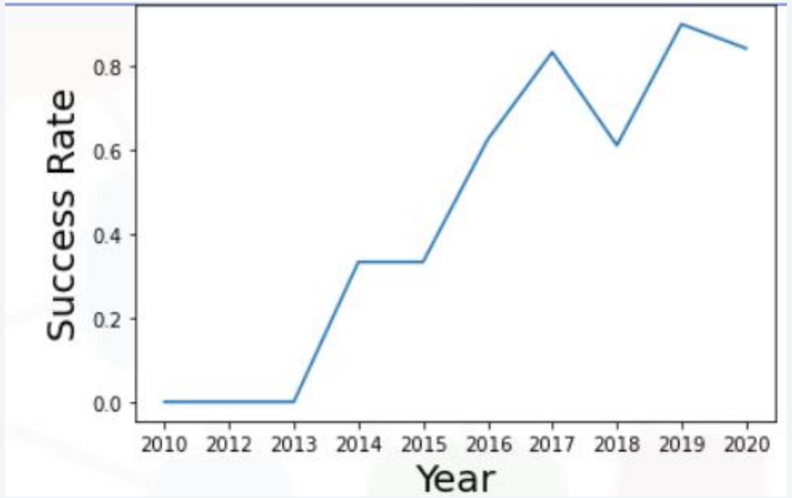
IBM Developer SKILLS NETWORK

Payload vs. Orbit Type

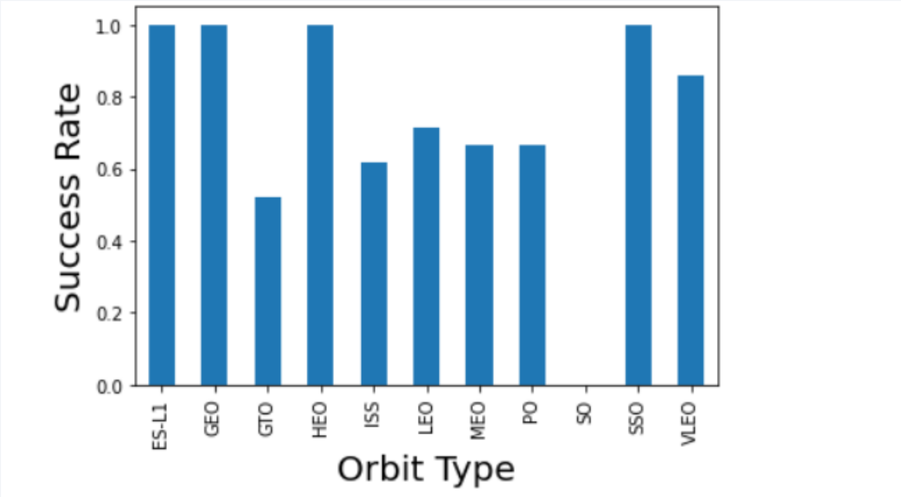


❑Heavier payloads have a negative influence on GTO orbits and positive influence on ISS orbits

Launch Success Yearly Trend and w.r.t Orbit type



❑Success rate trending positively on a yearly basis since 2013.



❑All orbit types except 'SO' have had successful 1st stage landings.

IBM Developer SKILLS NETWORK

Results : EDA with SQL

All Launch Site Names

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

Launch Site Names Begin with 'CCA'

- CCAFS LC-40
- CCAFS SLC-40
- Last launch from CCAFS LC-40 was 2016-08-14
- First launch from CCAFS SLC-40 was 2017-12-15

Total Payload Mass carried by boosters launched by NASA (CRS)

- 45,596 Kilogram in total

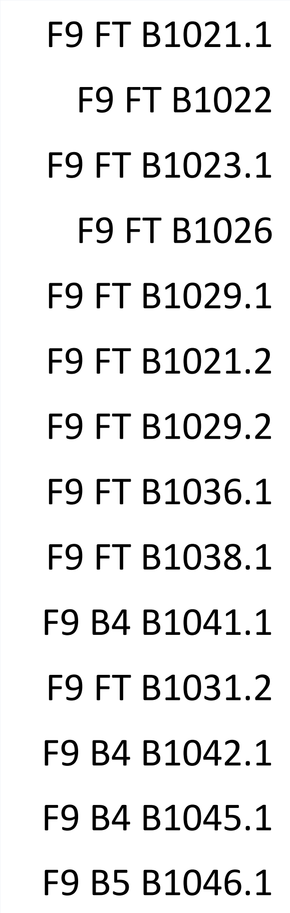
Average payload mass carried by booster version F9 v1.1

- 2534.67 Kilogram

Date when the first successful landing outcome in ground pad was achieved

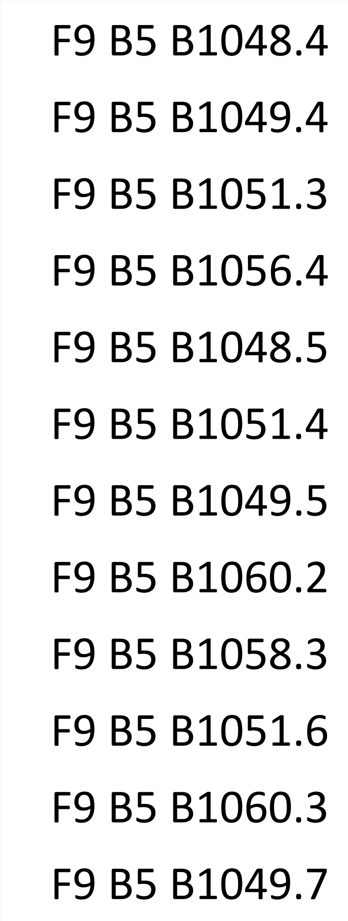
- 01/05/2017 was the first successful landing outcome in ground pad acheived

- Successful Drone Ship Landing with Payload between 4000 and 6000 are listed below.



F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

- Names of the booster versions which have carried the maximum payload mass

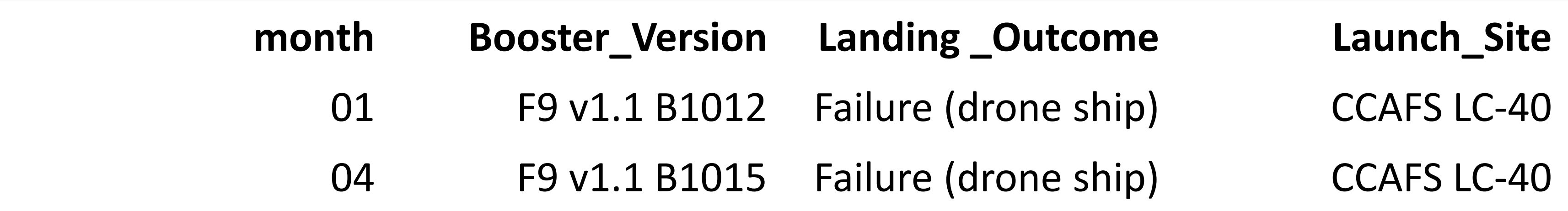


F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	TOTAL_NUMBER
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

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



month	Booster_Version	Landing _Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

The following shows the query of landing outcomes ranked in descending order :

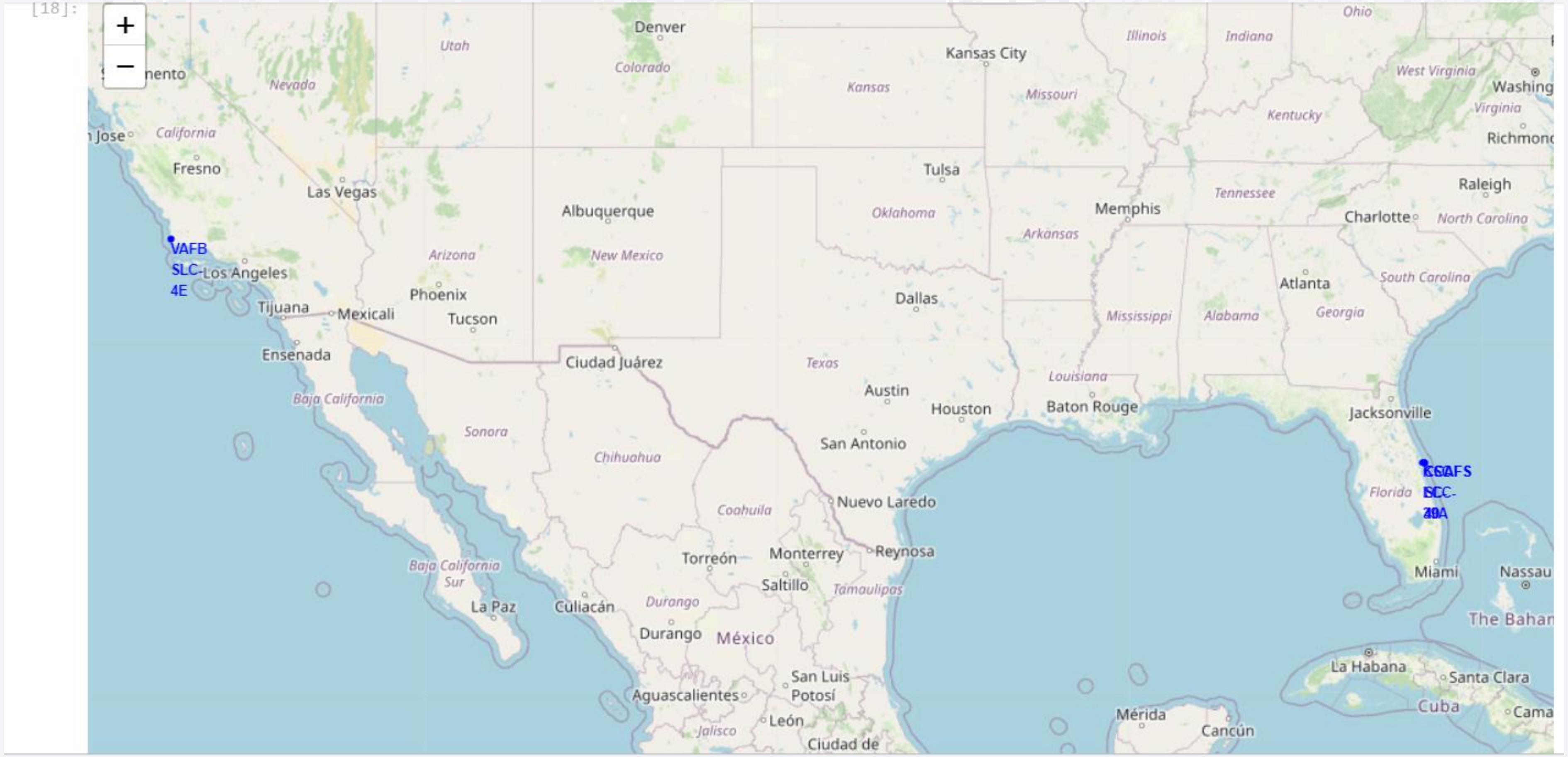
Landing _Outcome	LANDING_OUTCOME_COUNT	Date
No attempt	10	22-05-2012
Success (drone ship)	5	08-04-2016
Failure (drone ship)	5	10-01-2015
Success (ground pad)	3	22-12-2015
Controlled (ocean)	3	18-04-2014
Uncontrolled (ocean)	2	29-09-2013
Failure (parachute)	2	04-06-2010
Precluded (drone ship)	1	28-06-2015

Section 3

Launch Sites Proximities Analysis

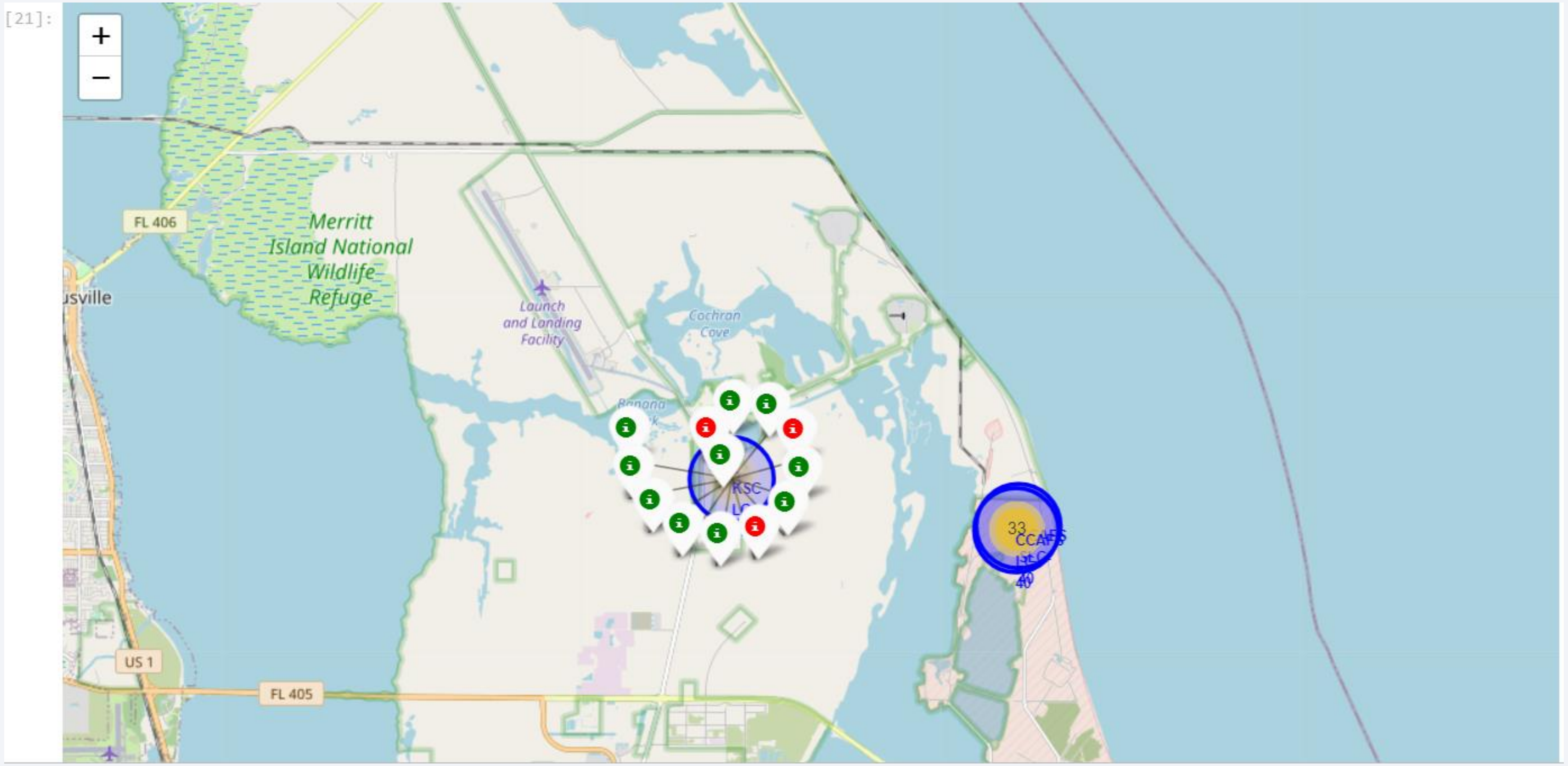
Results : Launch Site Location Analysis

✓ Visualizing the launch sites on a map highlights the importance of launch site proximity to the coast and equator using folium

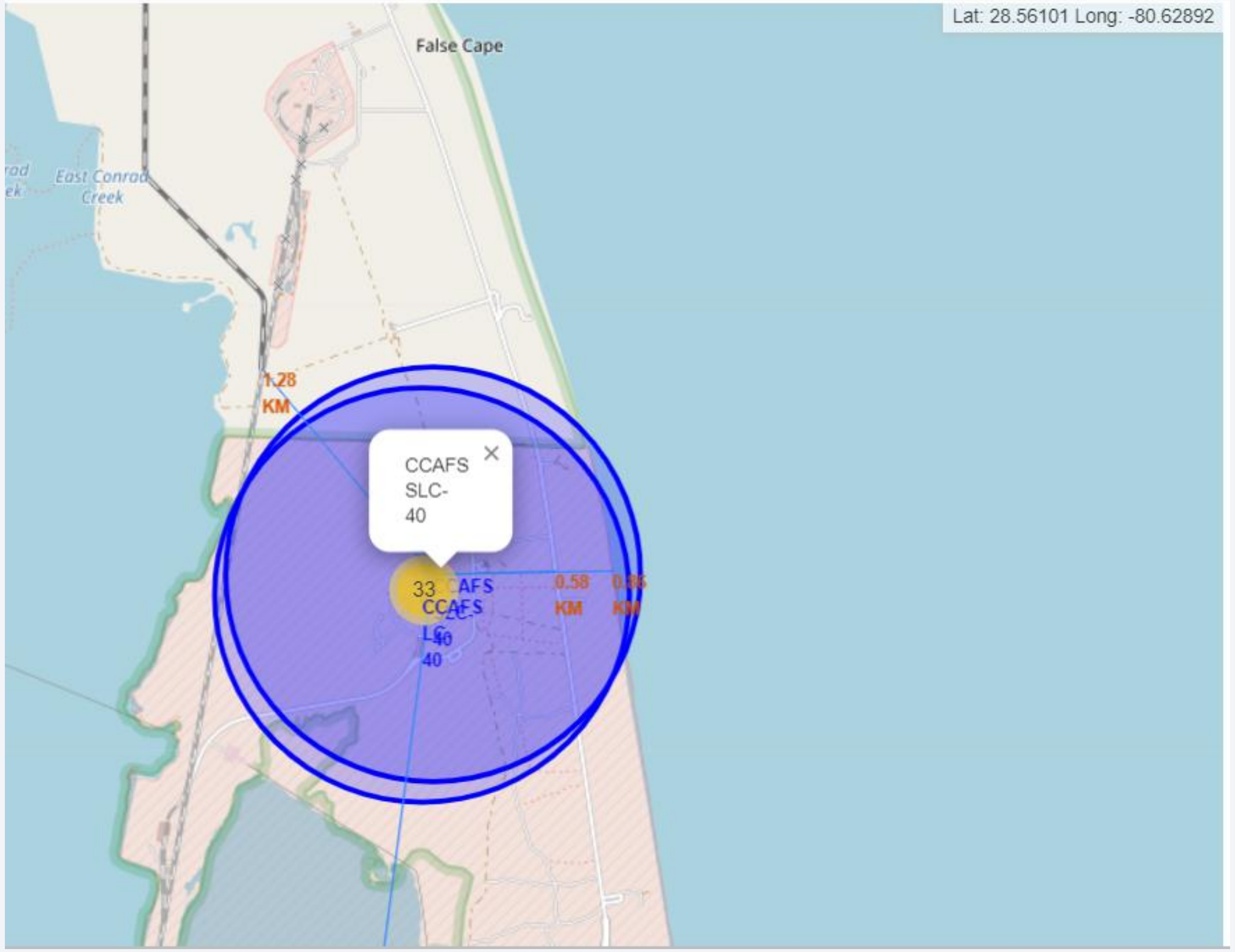


Note : In Github maps cannot be used although I have applied trusted option in notebook. It will work if you download it run it on your platform

- ✓ Visualizing the booster landing outcomes for each launch site highlights which launch sites have relatively high success rates, namely KSC LC39A


[21]:
+
−
FL 406
Merritt Island National Wildlife Refuge
usville
Launch and Landing Facility
Cochran Cove
KSC LC
33
CCAFS
US 1
FL 405

Visualizing the railway, highway, coastline, and city proximities for each launch site allows us to see how close each are present

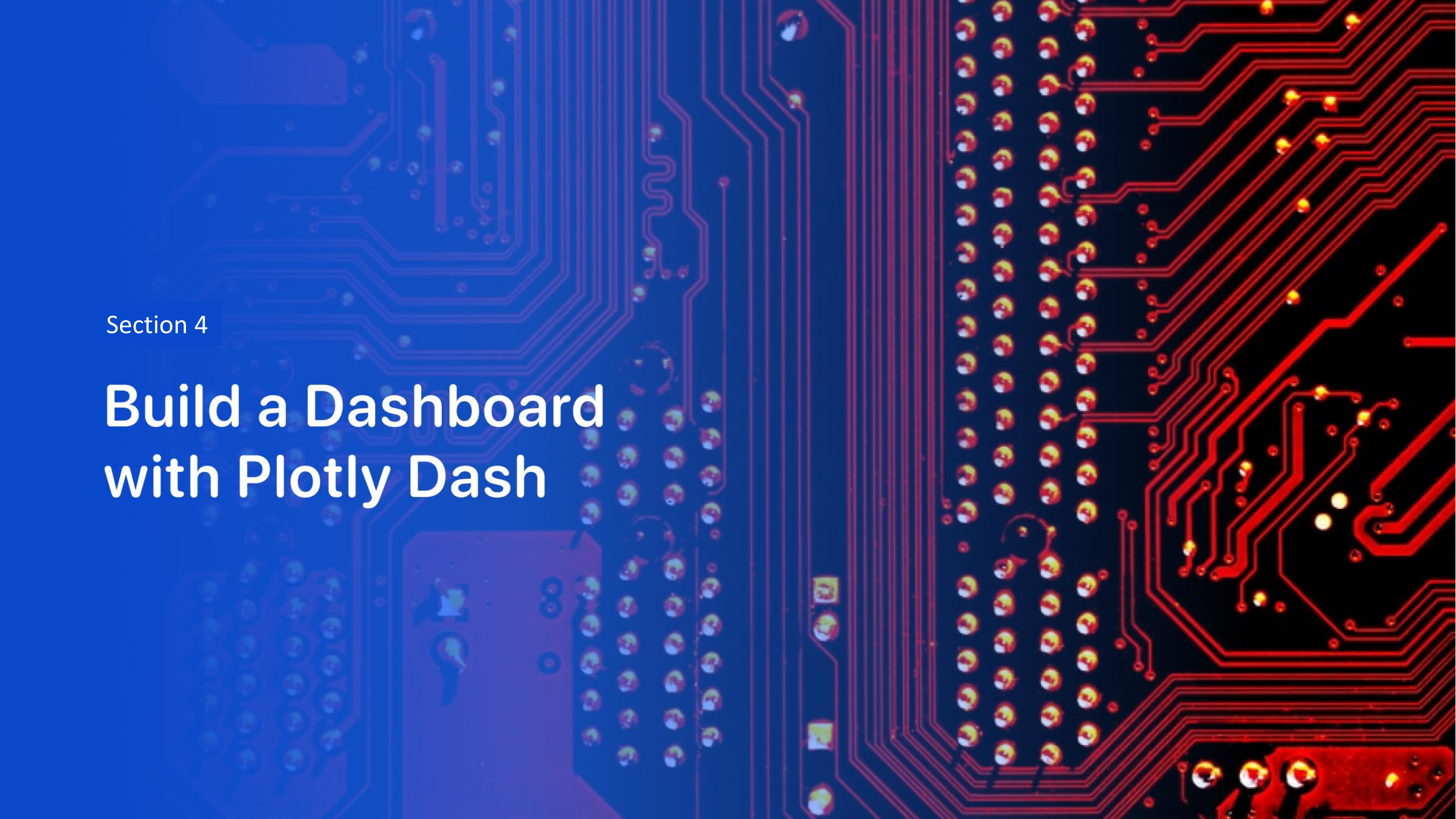


Proximities for CCAFS SLC-40:

- ➢ Railway: 1.28 km
  - transporting heavy cargo
- ➢ Highway: 0.58 km
  - transporting personal and equipment
- ➢ Coastline: 0.86 km
  - optionality to abort launch and attempt water landing
  - minimizing risk from falling debris
- ➢ City: 51.43 km
  - minimizing danger to population dense areas

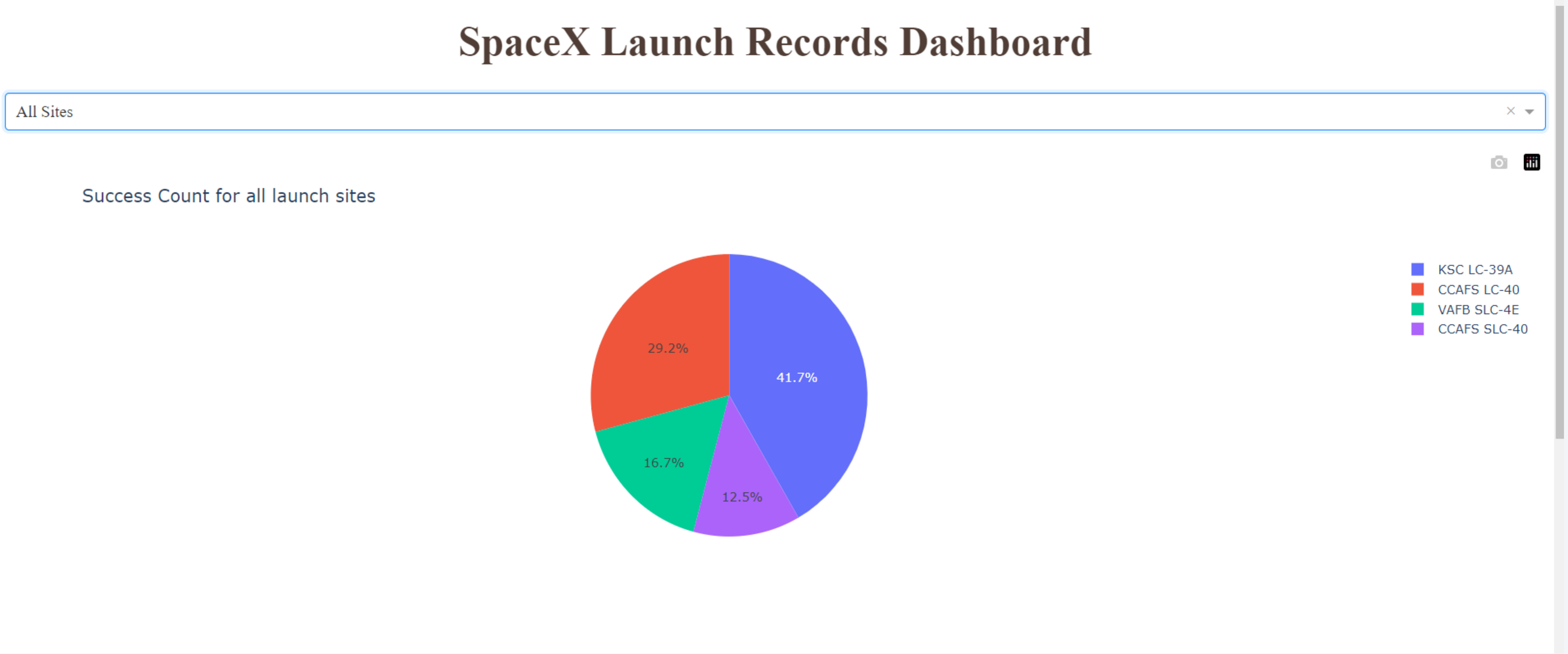
Section 4

Build a Dashboard with Plotly Dash



Results : Launch Records Dashboard

- ❖ Drop-down menu to choose between all sites and individual launch sites.
- ❖ Color coded by launch sites.
- ❖ Pie chart showing booster landing success rate



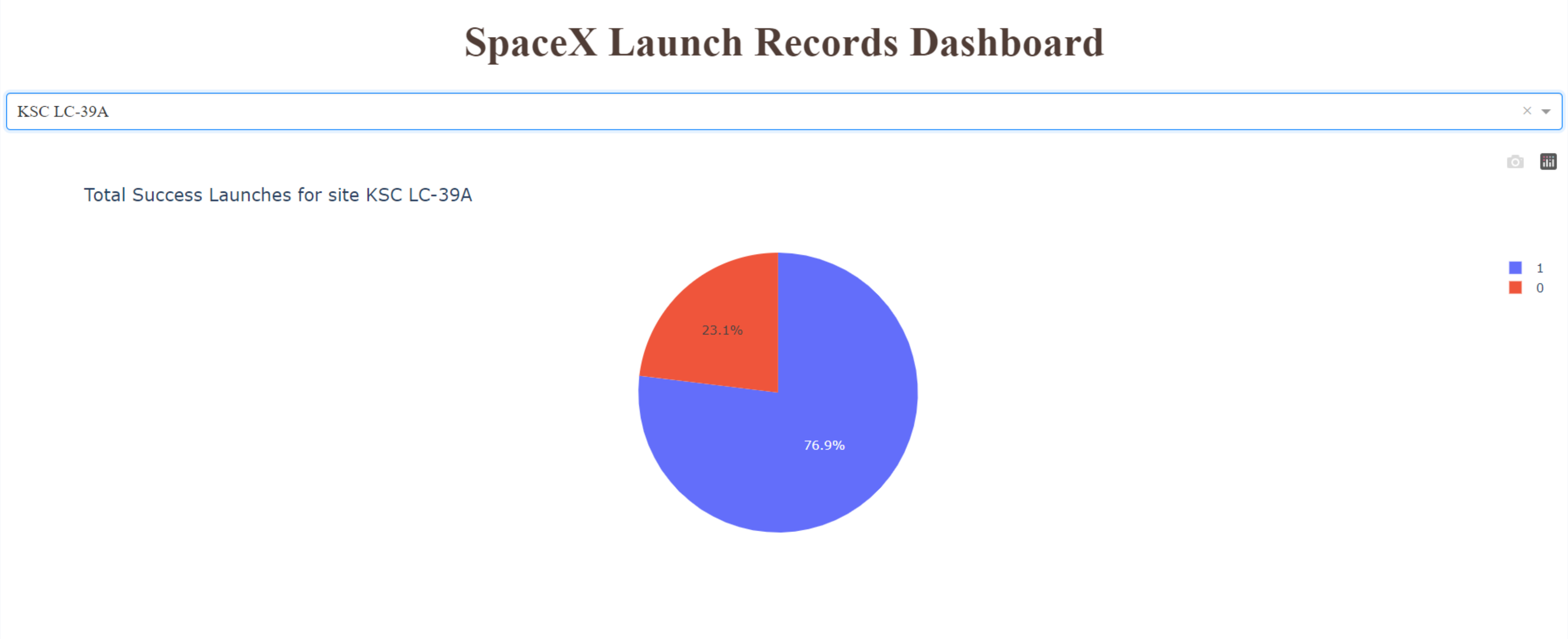
SpaceX Launch Records Dashboard

All Sites

Success Count for all launch sites

Dashboard Observations

- Enabling stakeholders to explore and manipulate the data in an interactive and real-time way
- KSC LC-39A has the highest booster landing success rate.

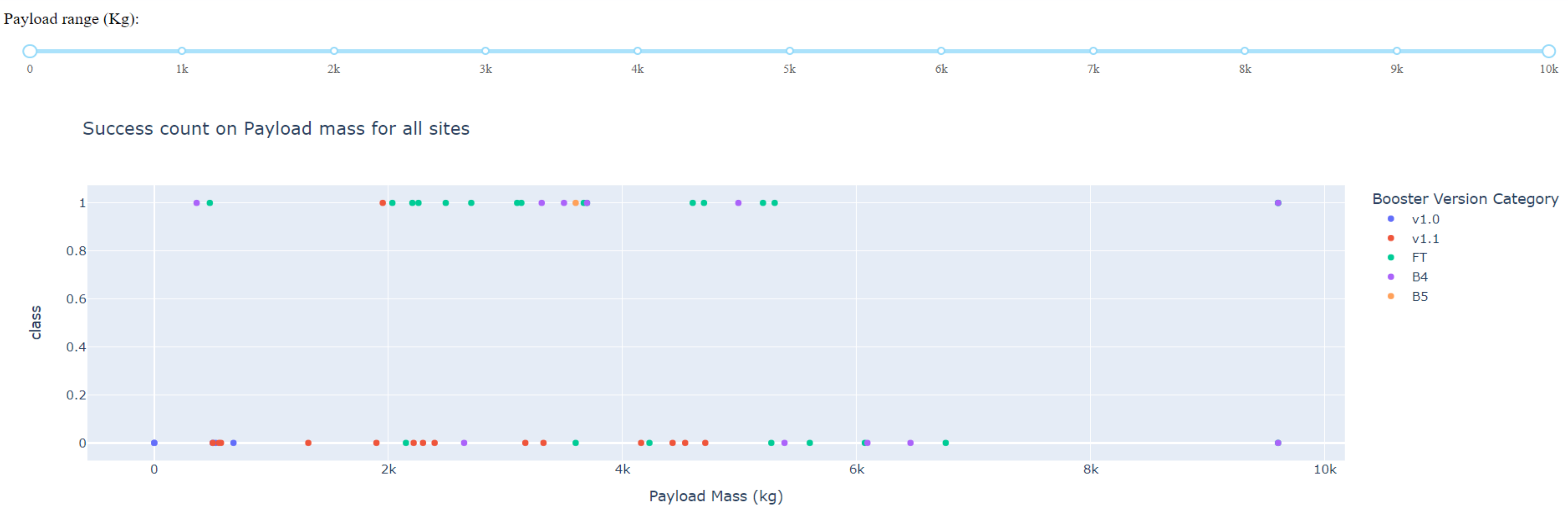


SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

Dashboard Observations

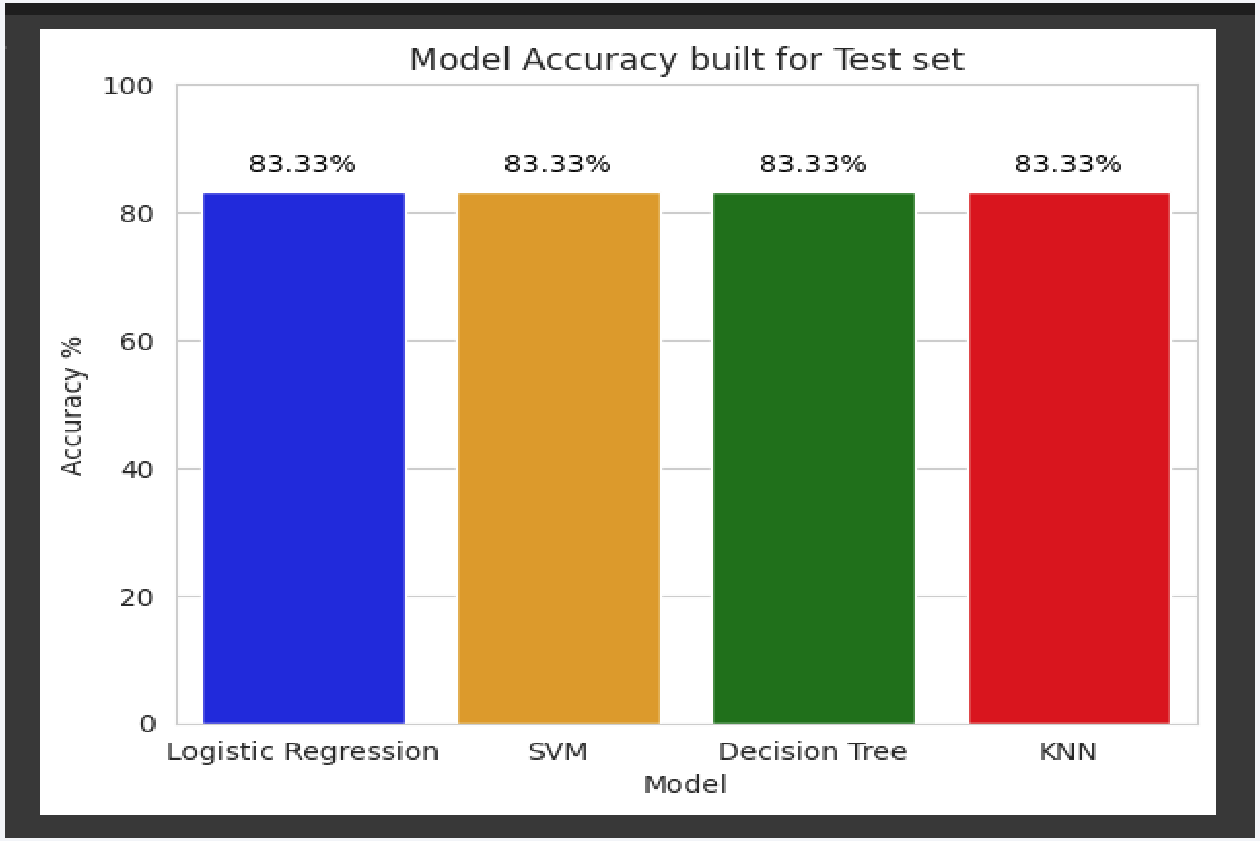


- ✓ Range slider for limiting payload amount.
- ✓ Scatter chart showing payload mass vs. landing outcome .
- ✓ Color coded by booster version.
- ✓ Payloads < 5,300 kg had the highest booster landing success rate .
- ✓ Payloads > 5,300 kg had the lowest booster landing success rate.

Section 5

Predictive Analysis (Classification)

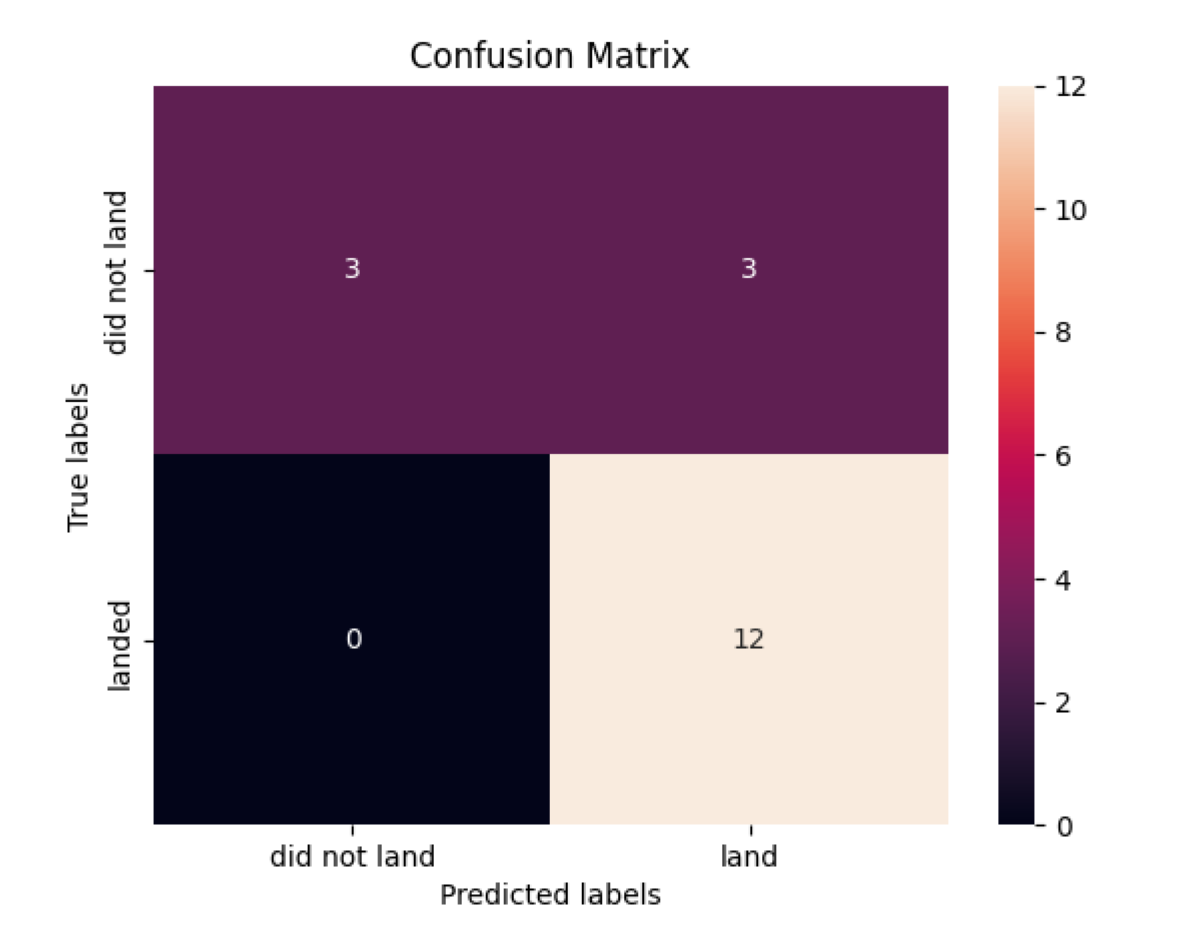
Results : Classification Accuracy



❑Visualized the built model accuracy for all built classification models, in a bar chart

❑All the models built on the test set has the accuracy score of 83.33%.

Confusion Matrix



- By analyzing the confusion matrix, we can identify the strengths and weaknesses of the model and make adjustments to improve its performance.

- In this case, there are 12 true positive predictions (TP), meaning that the model correctly predicted that 12 observations landed. There are 3 true negative predictions (TN), meaning that the model correctly predicted that 3 observations did not land. There are 3 false positive predictions (FP), meaning that the model predicted that 3 observations landed, but they actually did not. Finally, there are 0 false negative predictions (FN), meaning that the model correctly predicted that all 12 observations landed.

- Overall, this model has a high accuracy and recall, indicating that it is good at correctly classifying observations that land. However, its precision is somewhat lower, meaning that it has a tendency to predict that some observations land even when they do not.

Conclusions

- This report provides models that SpaceY can use to make predictions about the successful landing of the 1st stage booster by SpaceX. These models have an accuracy of 83.3%, meaning that they can predict the outcome with a high level of confidence.

- SpaceY will be able to make more informed bids against SpaceX. This is because SpaceY will have a more accurate idea of when they can expect the SpaceX bid to include the cost of a sacrificed 1st stage booster, allowing them to better plan and allocate their resources.

- Future study for improving the project :

  - ✓ Improving the feature selection and feature engineering.
  - ✓ Use of different algorithm or model
  - ✓ Add more data**:** Adding more data can sometimes help to reduce false positives by providing more examples for the model to learn from. However, this may not always be feasible or practical.
  - ✓ Obtaining the most effective combination of model and hyperparameters, the top-performing combination has been selected and frozen. The next step is to fit this combination using the complete dataset instead of just the training data.
  - ✓ Incorporate additional launch data to the dataset and model as it becomes available.

Appendix

➢ Python code snippets, SQL queries, charts, Notebook outputs, data sets analysis and models:

- ❑ https://github.com/Harilm10/IBM---Applied-Data-Science-Capstone-Project/blob/master/Data%20Collection%20API.ipynb
- ❑ https://github.com/Harilm10/IBM---Applied-Data-Science-Capstone-Project/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb
- ❑ https://github.com/Harilm10/IBM---Applied-Data-Science-Capstone-Project/blob/master/Exploratory%20Data%20Analysis%20(Data%20Wrangling).ipynb
- ❑ https://github.com/Harilm10/IBM---Applied-Data-Science-Capstone-Project/blob/master/EDA%20with-sql-coursera_sqllite.ipynb
- ❑ https://github.com/Harilm10/IBM---Applied-Data-Science-Capstone-Project/blob/master/EDA%20with-Data%20Visualization.ipynb
- ❑ https://github.com/Harilm10/IBM---Applied-Data-Science-Capstone-Project/blob/master/Lauchsites%20interaction%20with%20Folium%20map.ipynb
- ❑ https://github.com/Harilm10/IBM---Applied-Data-Science-Capstone-Project/blob/master/spacex_dash_app.py
- ❑ https://github.com/Harilm10/IBM---Applied-Data-Science-Capstone-Project/blob/master/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

➢ Acknowledgment :

I would like to thank IBM for providing this course and to work with real world dataset problem and also all the instructors of IBM who helped and guided me towards the project.

Thank you!