Winning Space Race with Data Science

Chae-Hyun Park
19 Nov 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via API and Web Scraping
  - Exploratory Data Analysis (EDA) with Data Visualization
  - EDA with SQL
  - Interactive Map with Folium
  - Dashborads with Plotly Dash
  - Predictive Analysis
- Summary of all results
  - EDA results
  - Interactive Maps and Dashboard
  - Predictive results

Introduction

- Project background and context
  - The aim of this project is to predict if the first stage of the Falcon 9 rocket will successfully land. SpaceX claims that reusing the first stage reduces the cost of rocket launch dramatically down to 62 million dollars, compared to other providers' cost of up to 165 million dollars. By predicting whether the first stage will land, we can determine the cost of a launch, which will be a useful information for competing companies.
- Problems you want to find answers
  - What are the main characteristics of a successful/failed landing?
  - What are the effects of relationships between rocket variables on the success/failure of a landing?
  - What are the conditions for the best landing success rate?

Section 1

Methodology

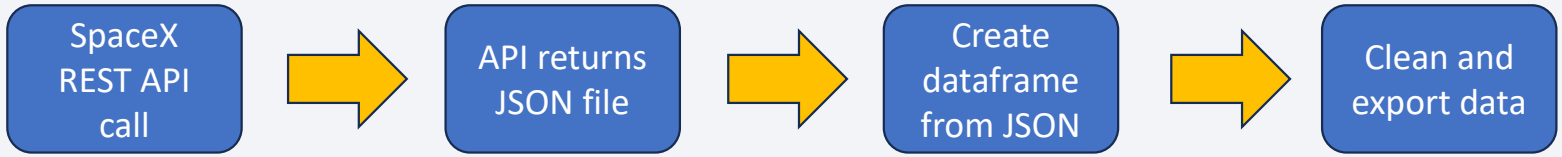
Methodology

Executive Summary

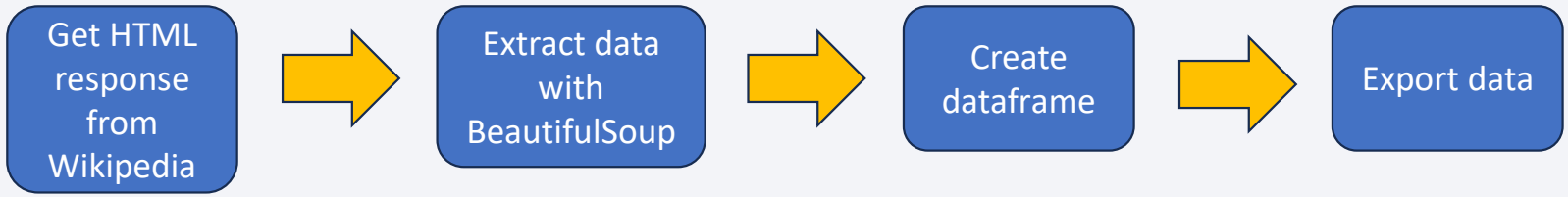
- Data collection methodology:
  - SpaceX REST API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Datasets are collected through SpaceX REST API and Web Scrapping from Wikipedia.
  - SpaceX REST API URL: api.spacexdata.com/v4/



  - Wikipedia URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



Data Collection – SpaceX API

1. Getting response from API
2. Convert response to JSON file
3. Transform data
4. Create dictionary with data
5. Create dataframe
6. Filter dataframe
7. Export to file

- GitHub URL

Data Collection - Scraping

1. Getting Response from HTML
2. Create BeautifulSoup object
3. Find all tables
4. Get column names
5. Create Dictionary
6. Add data to keys
7. Create dataframe from dictionary
8. Export to file

- GitHub URL

Data Wrangling

1. Calculate number of launches for each site
2. Calculate number and occurrence of each orbit
3. Calculate number and occurrence of mission outcome per orbit type
4. Create landing outcome label from Outcome column
5. Export to file

- GitHub URL

EDA with Data Visualization

- Scatter Plot
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Orbit vs. Flight Number
  - Payload vs. Orbit Type
  - Orbit vs. Payload Mass

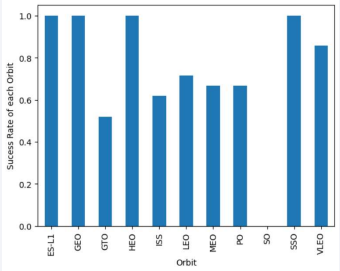
  => shows relationship between two variables (correlation)

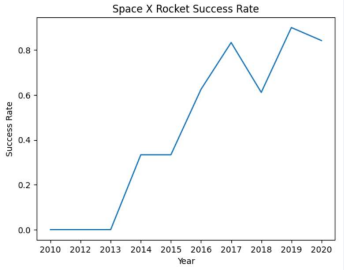
- Bar Plot
  - Success Rate vs. Orbit

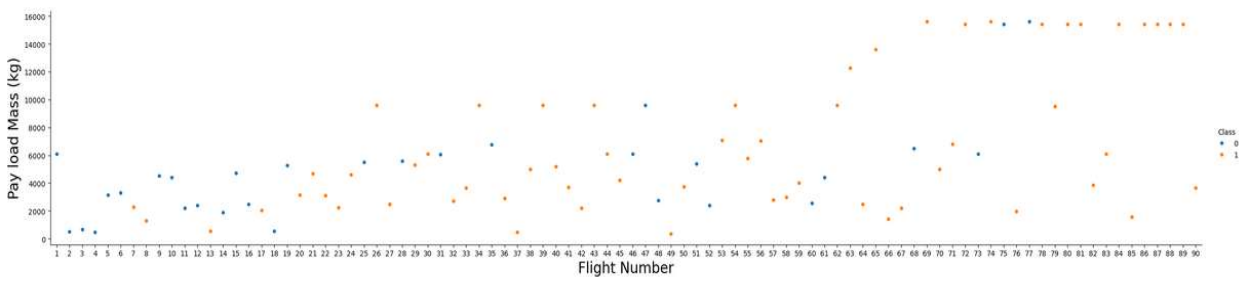
  => shows relationship between numeric and categorical variables

- Line Plot
  - Success Rate vs. Year

  => shows trends and can be used to make prediction







- GitHub URL

EDA with SQL

- SQL queries to gather and analyze the data from dataset

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA’
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first succesful landing outcome in ground pad was achieved
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. 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
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

- GitHub URL

Build an Interactive Map with Folium

- Folium map object: map centered on NASA Johnson Space Center at Houson, Texas
  - Red circle at NASA Johnson Space Center's coordinate with label showing its name
  - Red circles at each launch site coordinates with label showing launch site name
  - The grouping of points in a cluster to display multiple and different information for the same coordinates
  - Markers to show successful (green) and unsuccessful (red) landings
  - Markers to show distance between launch site to key locations (railway, highway, coastway, city) and to plot a line between them

- GitHub URL

Build a Dashboard with Plotly Dash

- Dropdown
  - allows users to choose the launch site
- Pie Chart
  - shows the percentage of total success/failure for the chosen launch site
- Rangeslider
  - allows users to select a payload mass in a fixed range
- Scatter Plot
  - shows the relationship between launch success and payload mass

- GitHub URL

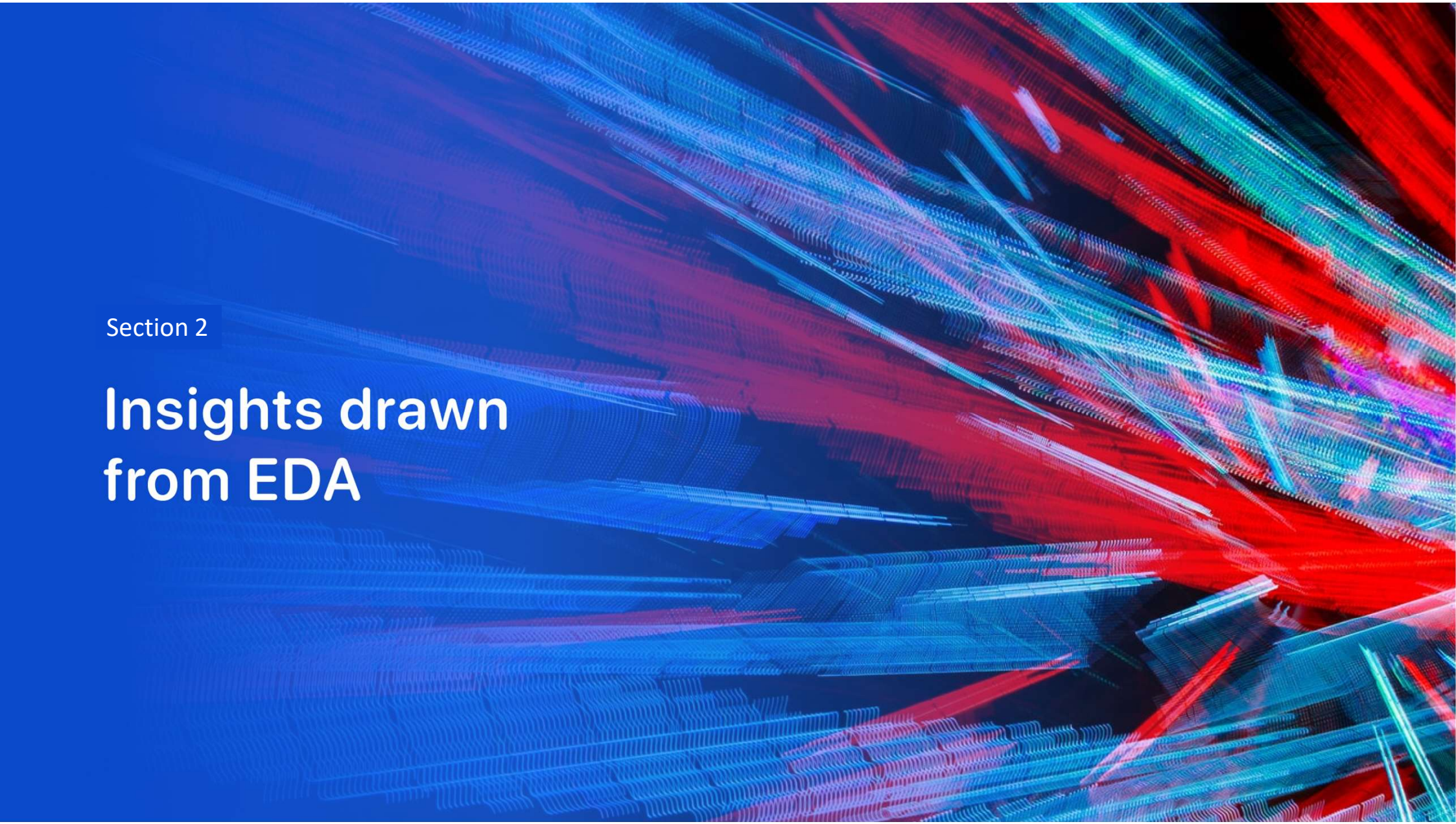
Predictive Analysis (Classification)

- Data Preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets
- Model Preparation
  - Select machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Train GridSearchModel models with training dataset
- Model Evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model Comparison
  - Compare models according to their accuracy
  - Choose the model with the best accuracy

- GitHub URL

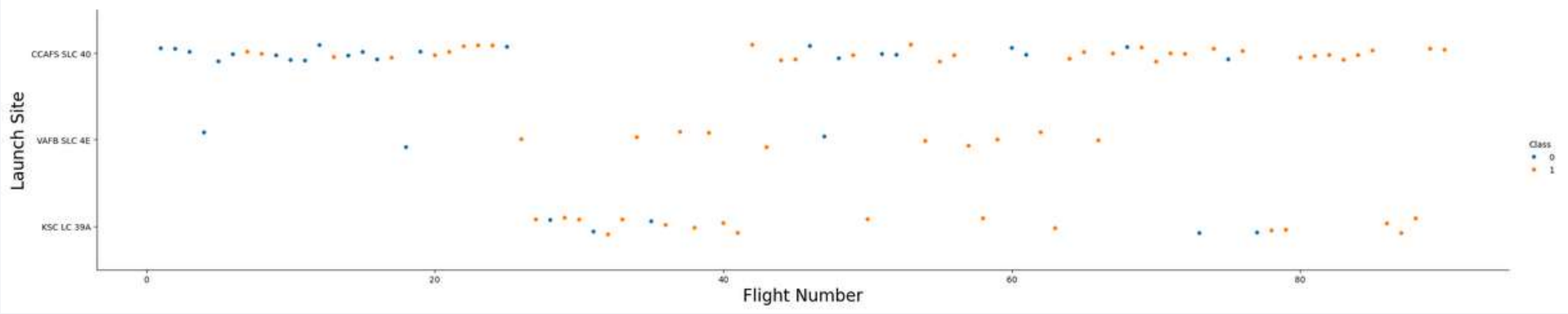
Results

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


Section 2
Insights drawn from EDA

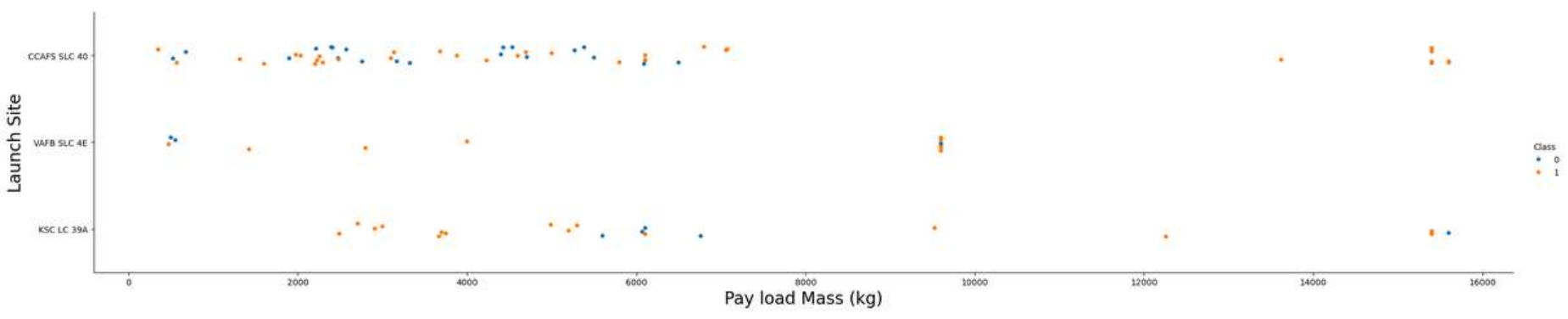
Flight Number vs. Launch Site

- The plot shows that there is no discernable dependency of the success rate on the flight number for all three launch sites



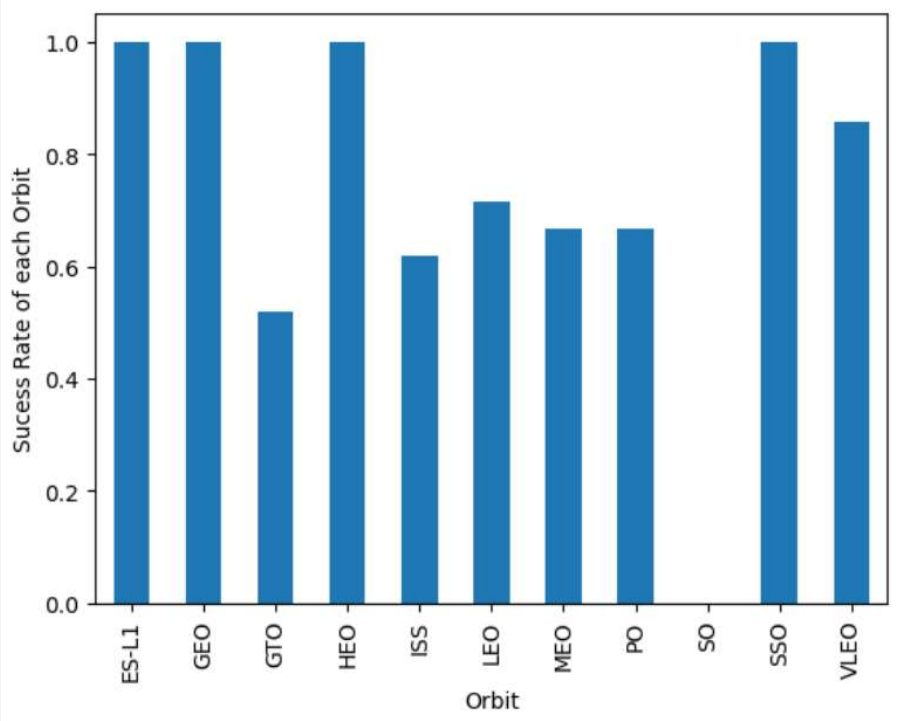
Payload vs. Launch Site

- For VAFB SLC 4E, a heavy payload of up to 10000 kg may be appropriate for successful landing.
- For KSC LC 39A and CCAFS SLC 40, a heavier payload may increase the success rate, but too high pay load mass might also cause failure of landing.



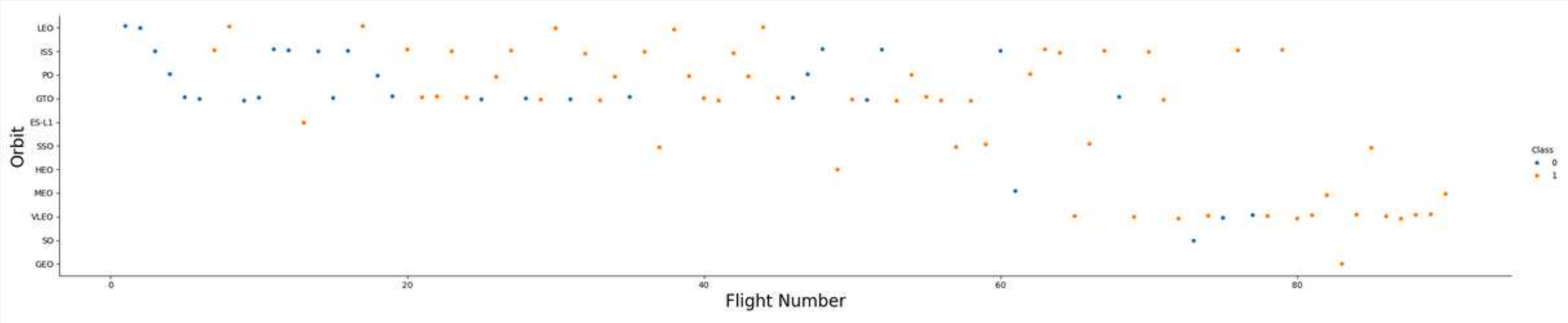
Success Rate vs. Orbit Type

- The bar chart shows that the success rate is most high for ES-L1, GEO, HEO and SSO.



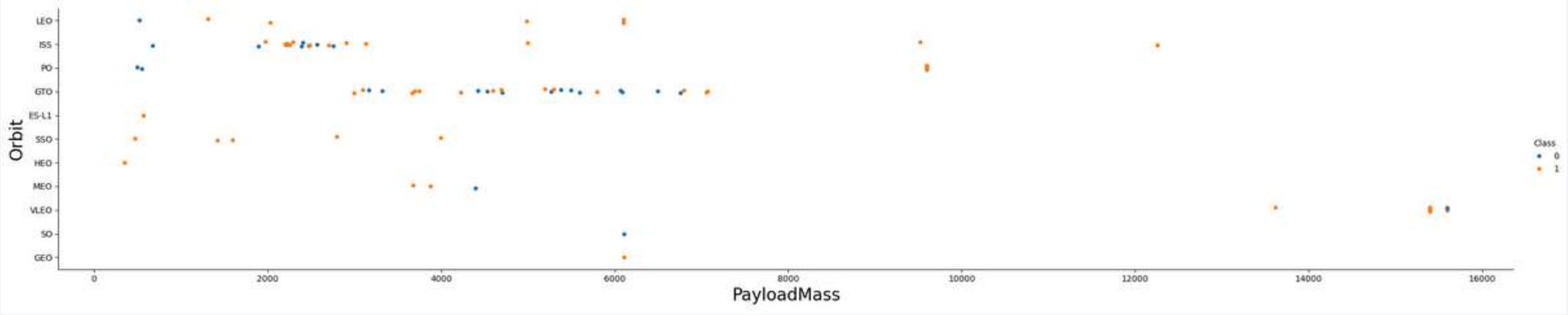
Flight Number vs. Orbit Type

- The scatter plot shows that the success rate appears to be higher for higher flight number in the LEO orbit.
- There seems to be no relationship between flight number when in GTO orbit.



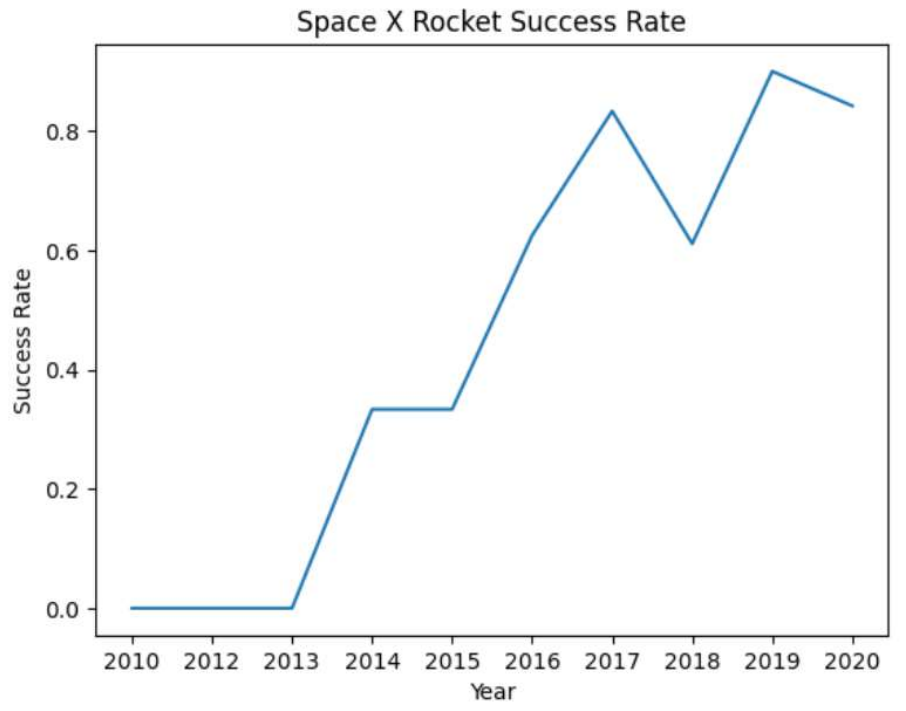
Payload vs. Orbit Type

- With heavy payloads the success rate is higher for Polar, LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.



Launch Success Yearly Trend

- The line chart shows that the success rate continuously increased from 2013 until 2017.
- After a downfall between 2017 and 2018, the success rate increased again till 2019.



All Launch Site Names

- SQL Query: SELECT DISTINCT “LaunchSite” FROM SPACEXTBL

  => The SELECT DISTINCT statement removes redundant LaunchSite

- Result: CCAFS LC-40, VAFB SLC-4E, KC LC-39A, CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- SQL Query: SELECT * FROM SPACEXTBL WHERE “LaunchSite” LIKE ‘%CCA%’ LIMIT 5

  => The WHERE clause followed by LIKE clause filters launch sites that contain the substring CCA. LIMIT 5 shows 5 records from filtering

- Result:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

- SQL Query: SELECT SUM("PAYLOAD_MASS_KG_") FROM SPACEXTBL WHERE "CUSTOMER" = "NASA (CRS)"

  => This query returns the sum of all payload masses where the customer is NASA (CRS)

- Result: 45596

Average Payload Mass by F9 v1.1

- SQL Query: SELECT AVG("PAYLOAD_MASS_KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'

  => This query returns the average of all payload masses where the booster version contains the substring F9 v1.1

- Result: 2534.6666666666665

First Successful Ground Landing Date

- SQL Query: SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Success%'

  => With this query, we select the oldest successful landing.

  => The WHERE clause filters dataset to keep only records where landing was successful.

  => With the MIN function, we select the record with the oldest date.

- Result: 2017-05-01

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL Query: %sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING_OUTCOME" = 'Success (drone ship)' AND "PAYLOAD_MASS_KG_" > 4000 AND "PAYLOAD_MASS_KG_" < 6000;

  => This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg.

  => The WHERE and AND clauses filter the dataset.

- Result:

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

Total Number of Successful and Failure Mission Outcomes

- SQL Query:

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

  => With the first SELECT, we show the subqueries that return results.

  => The first subquery counts the successful mission.

  => The second subquery counts the unsuccessful mission.

  => The WHERE clause followed by LIKE clause filters mission outcome.

  => The COUNT function counts records filtered.

- Result: 100 Success, 1 Failure

Boosters Carried Maximum Payload

- SQL Query:

```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

  => Subquery is used to filter data by returning only the heaviest payload mass with MAX function.

  => The main query uses subquery results and returns unique booster version (SELECT DISTINCT) with the heaviest payload mass.

- Result:

Booster_Version
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

- SQL Query:

```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

  => This query returns month, booster version, launch site where landing was unsuccessful and landing date took place in 2015.

  => Substr function processes date in order to take month or year.

  => Substr(DATE, 4, 2) shows month.

  => Substr(DATE,7, 4) shows year.

- Result:

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- SQL Query:

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

=> This query returns landing outcomes and their count where mission was successful and date is between 04/06/2010 and 20/03/2017.

=> The GROUP BY clause groups results by landing outcome and ORDER BY COUNT DESC shows results in decreasing order.

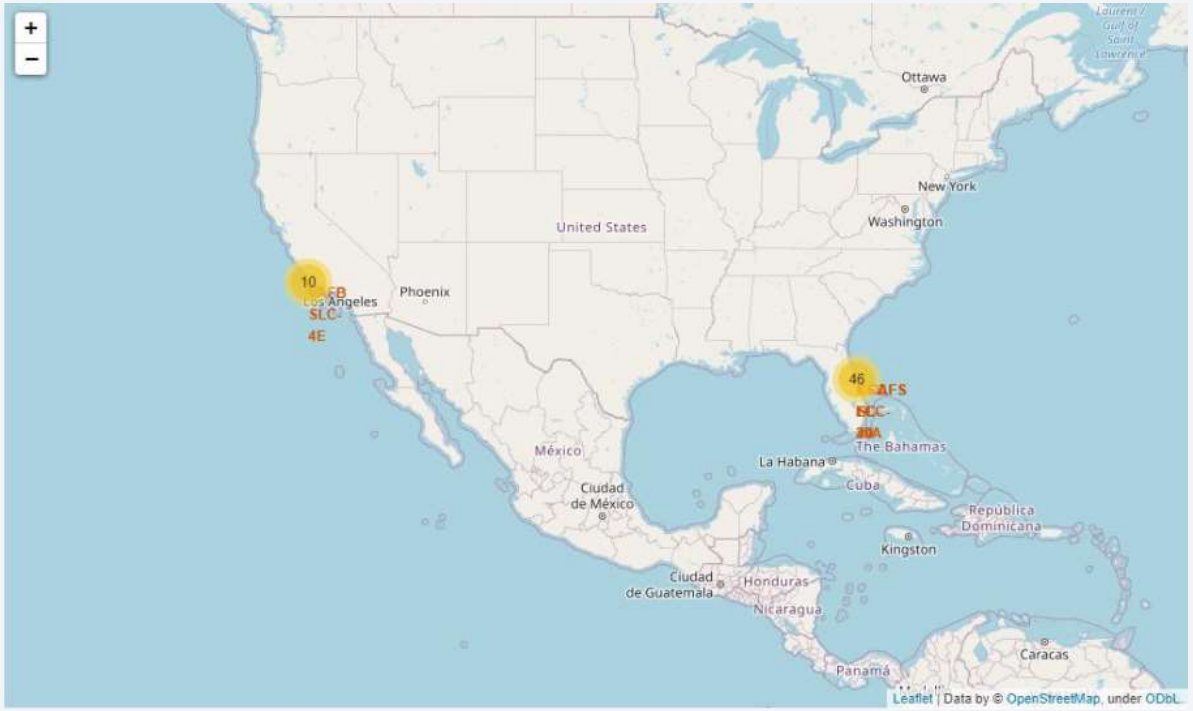
- Result:

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6


Section 3
Launch Sites
Proximities Analysis

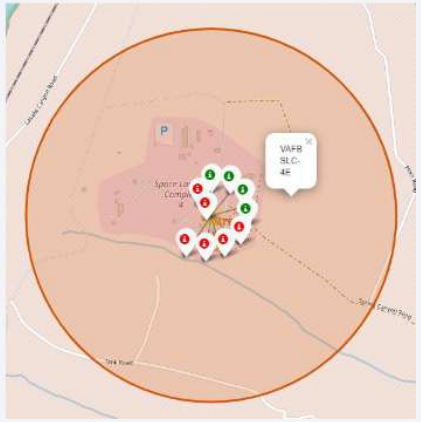
<Folium Map Screenshot 1>

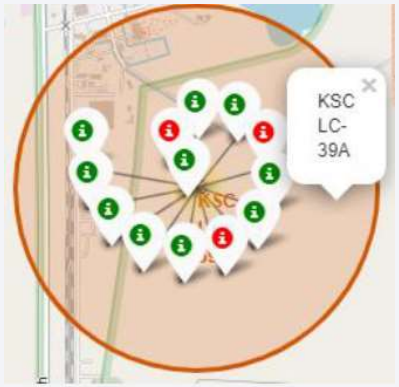
- The Folium Map shows that the Space X launch sites are located on the coast of the United States

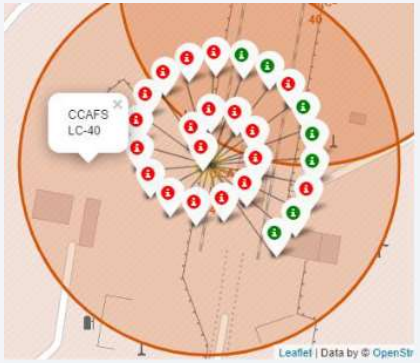


<Folium Map Screenshot 2>

- Green marker represents successful launches.
- Red marker represents unsuccessful launches.
- KSC LC-39A has the highest launch success rate.

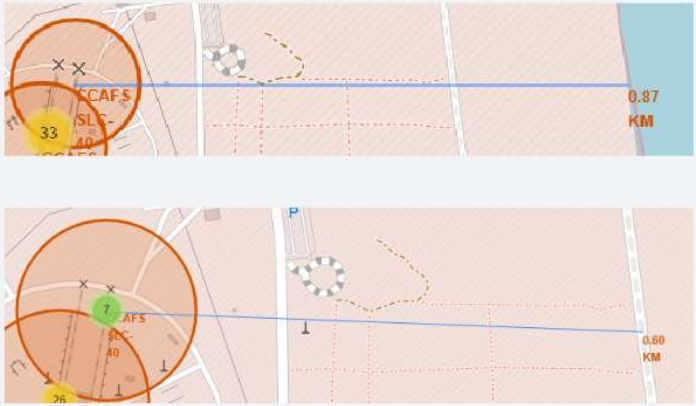


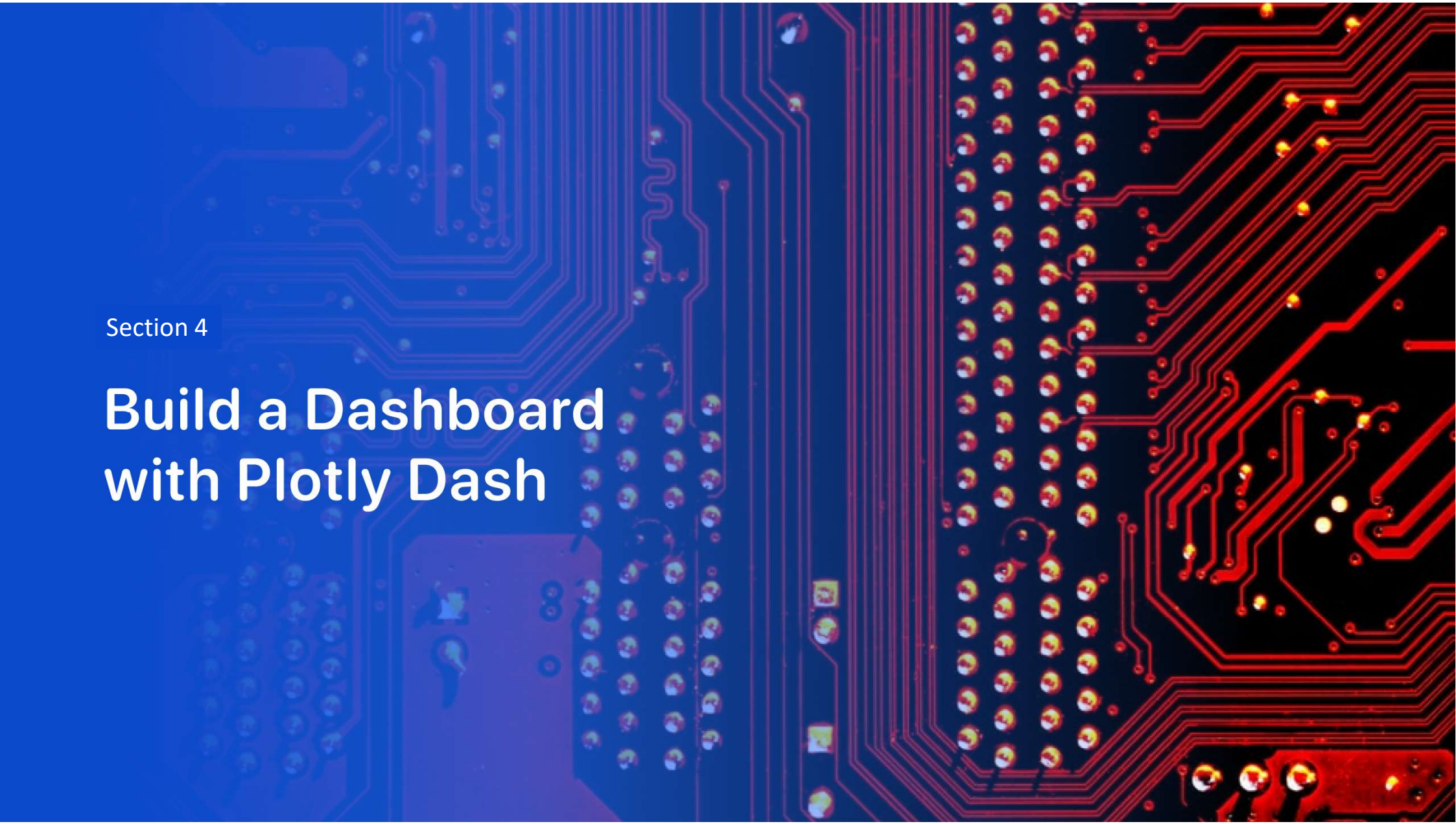




<Folium Map Screenshot 3>

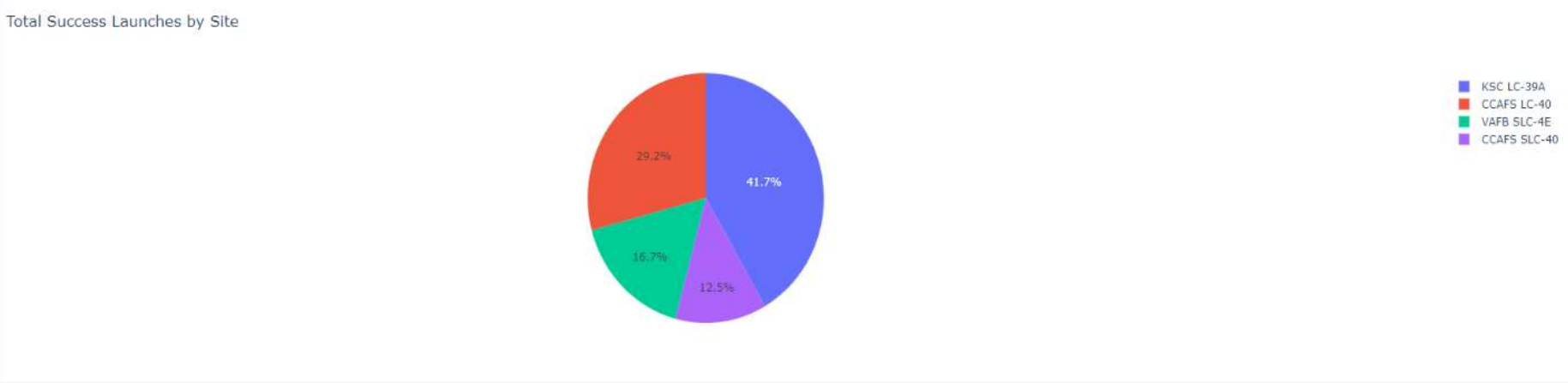
- Is CCAFS SLC-40 in close proximity to railways ? Yes
- Is CCAFS SLC-40 in close proximity to highways ? Yes
- Is CCAFS SLC-40 in close proximity to coastline ? Yes
- Do CCAFS SLC-40 keeps certain distance away from cities ? No




Section 4
Build a Dashboard with Plotly Dash

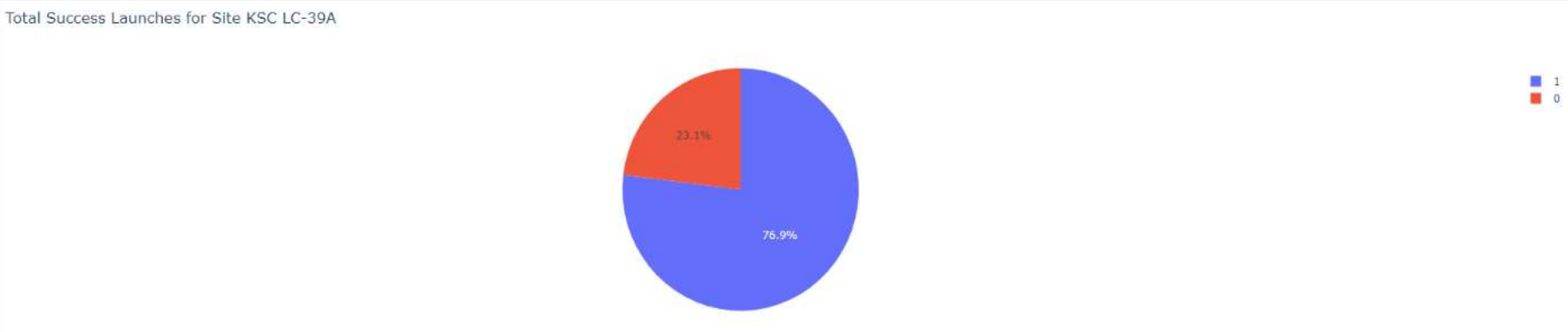
<Dashboard Screenshot 1>

- KSC LC-39A has the best success rate of launches.



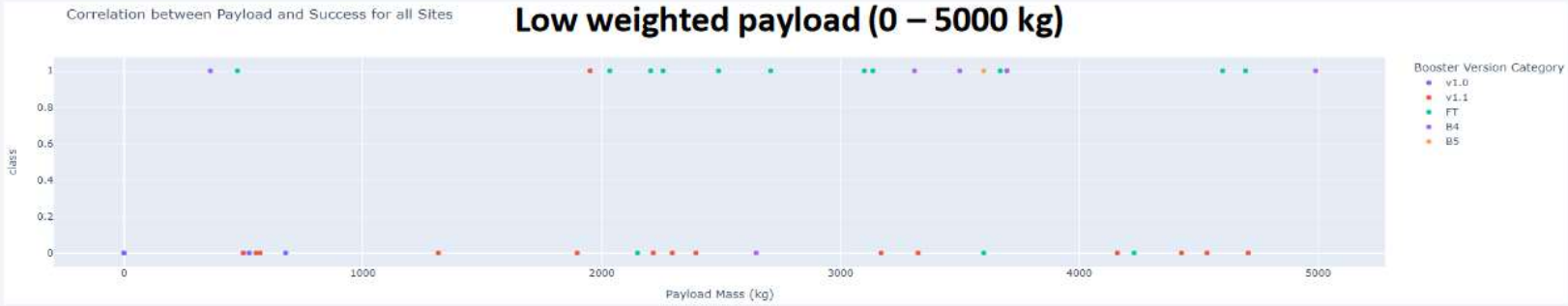
<Dashboard Screenshot 2>

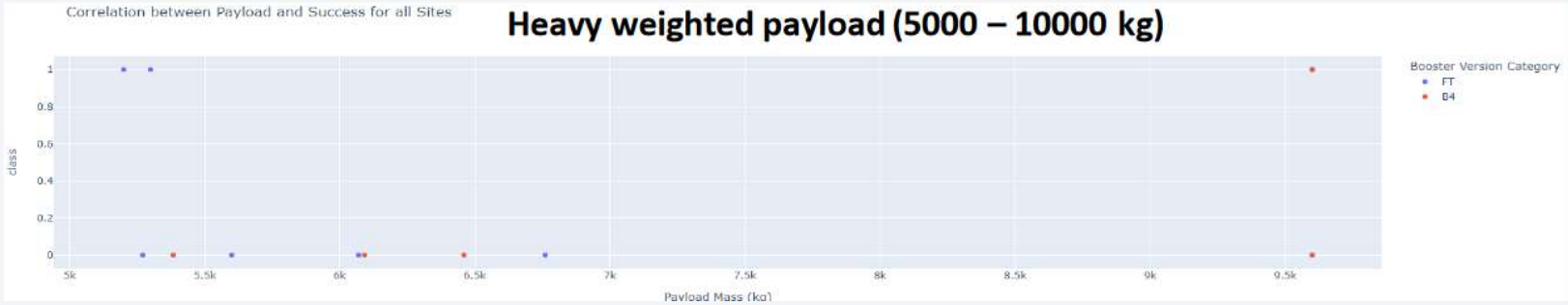
- KSC LC-39A has a 76.9% success rate and a 23.1% failure rate.



<Dashboard Screenshot 3>

- Low weighted payloads have a better success rate than the heavy weighted payloads





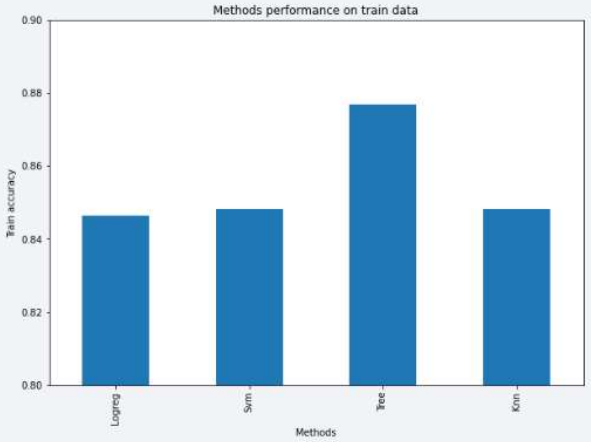
Section 5

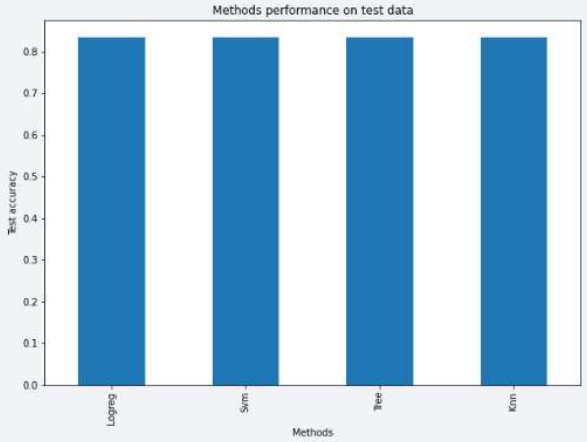
Predictive Analysis (Classification)

Classification Accuracy

- All methods showed similar performances in the accuracy test. More test data would be needed to determine the best method. For now, the decision tree seems to be a best choice.

	Accuracy Train	Accuracy Test
Tree	0.876786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333



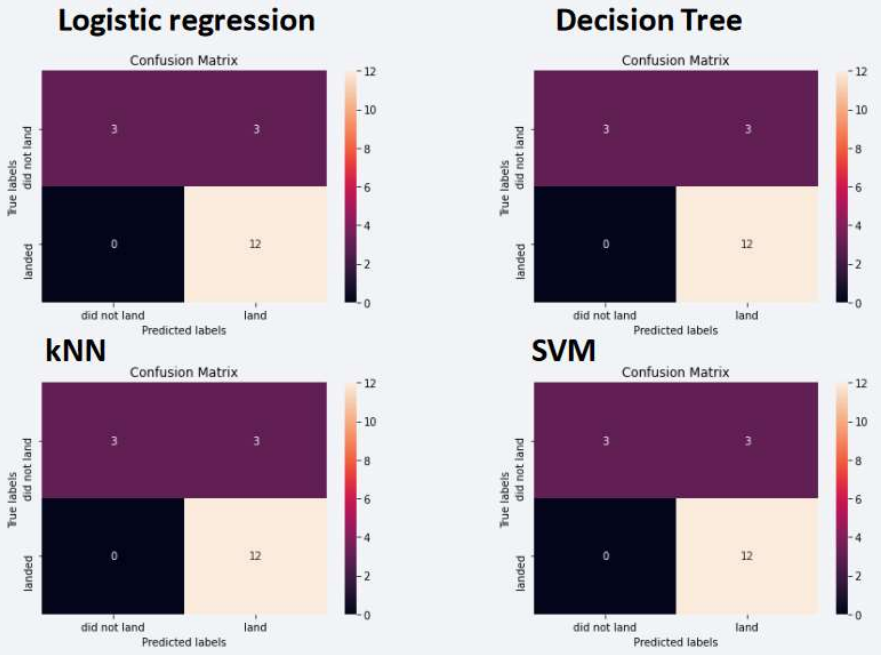


Decision tree best parameters

```
tuned hyperparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf':
4, 'min_samples_split': 2, 'splitter': 'random'}
```

Confusion Matrix

- As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models are false positives.



Predicted values	Actual values: 1	Actual values: 0
1	TP	FP
0	FN	TN

Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates: GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

Thank you!