IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aysegül Tekin
19.12.2024

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

Summary of methodologies

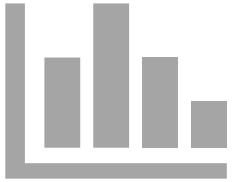
Data Coolection from API and Web Page

Data Wrangling and Processing

Exploratory Data Analysis (EDA) by SQL and Visualization Tools

Data Analytics Using Interactive Maps and Dashboards

Predictive Analysis Using Machine Learning Algortihms



Summary of all results

Results from EDA

Model Evaluation

Determining the Best Model

Introduction

Context

- SpaceX rokets are known for their reusable technology and successful landings.
- If we can predict if the first stage will land, then we can determine the cost of a launch.
- The objective is to predict whether or not the first stage of SpaceX Falcon 9 would be reused.

Problems to find answers to:

- Which factors affect the landing?
- What is the probability of successful landing for a specific launch?

Section 1

Methodology

Methodology

Data collection methodology:

With Rest API and from Wikipedia with Web Scrapping

Perform data wrangling

Data Transformation and One Hot Encoding for ML Models

Perform exploratory data analysis (EDA) using visualization and SQL

Discovering new patterns in the data

Perform interactive visual analytics using Folium and Plotly Dash

Dashboard and Folium

Perform predictive analysis using classification models

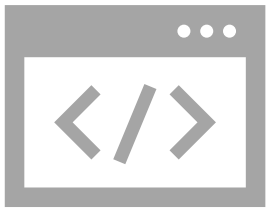
Comparision of Logistic Regression, SVM, Decision Tree, KNN by using GridSearchCV to select best fit model

Data Collection

API

URL="https://api.spacexdata.com/v4/launches/past"

Getting data and cleaning



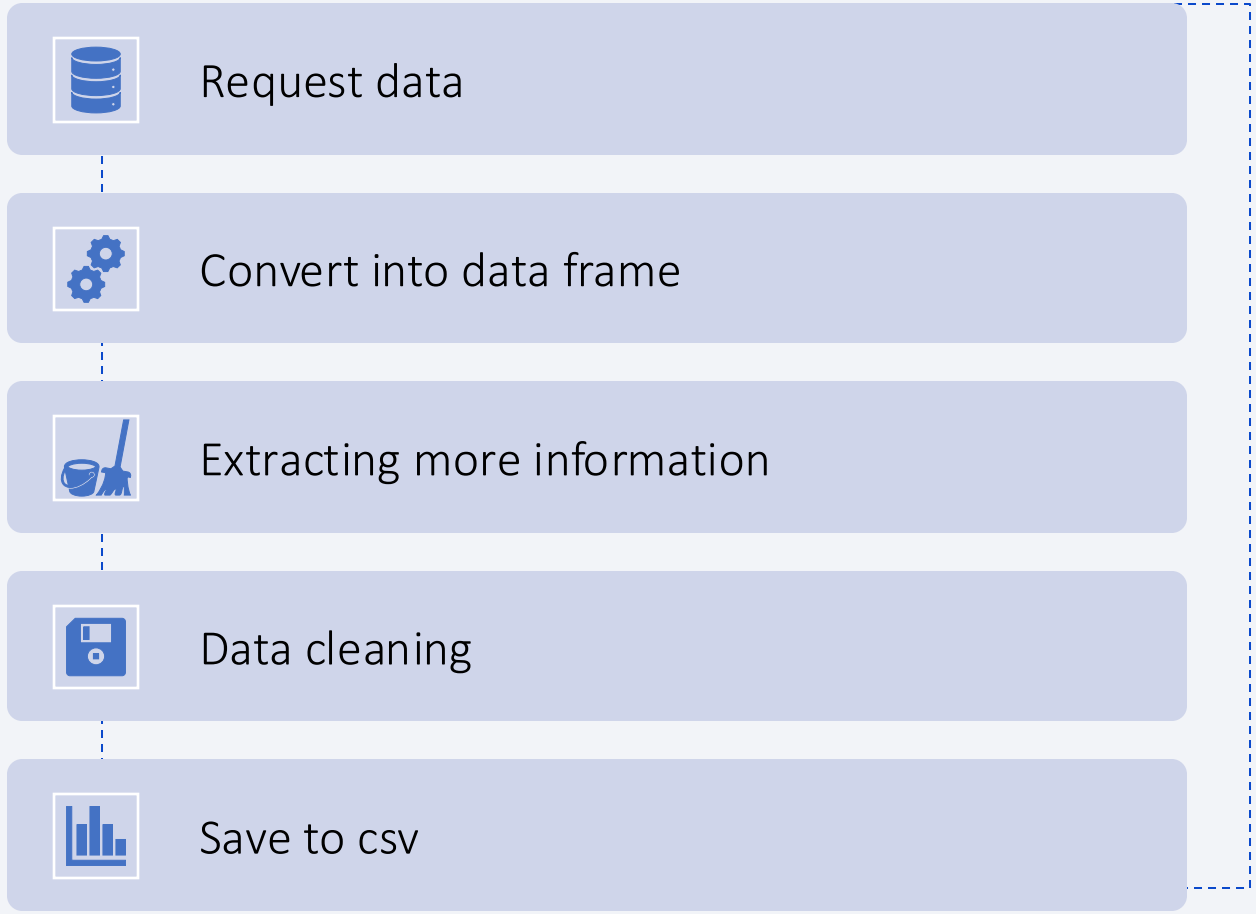
Web Page

URL="https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches"

Extarct the records in HTML format and convert to data frame

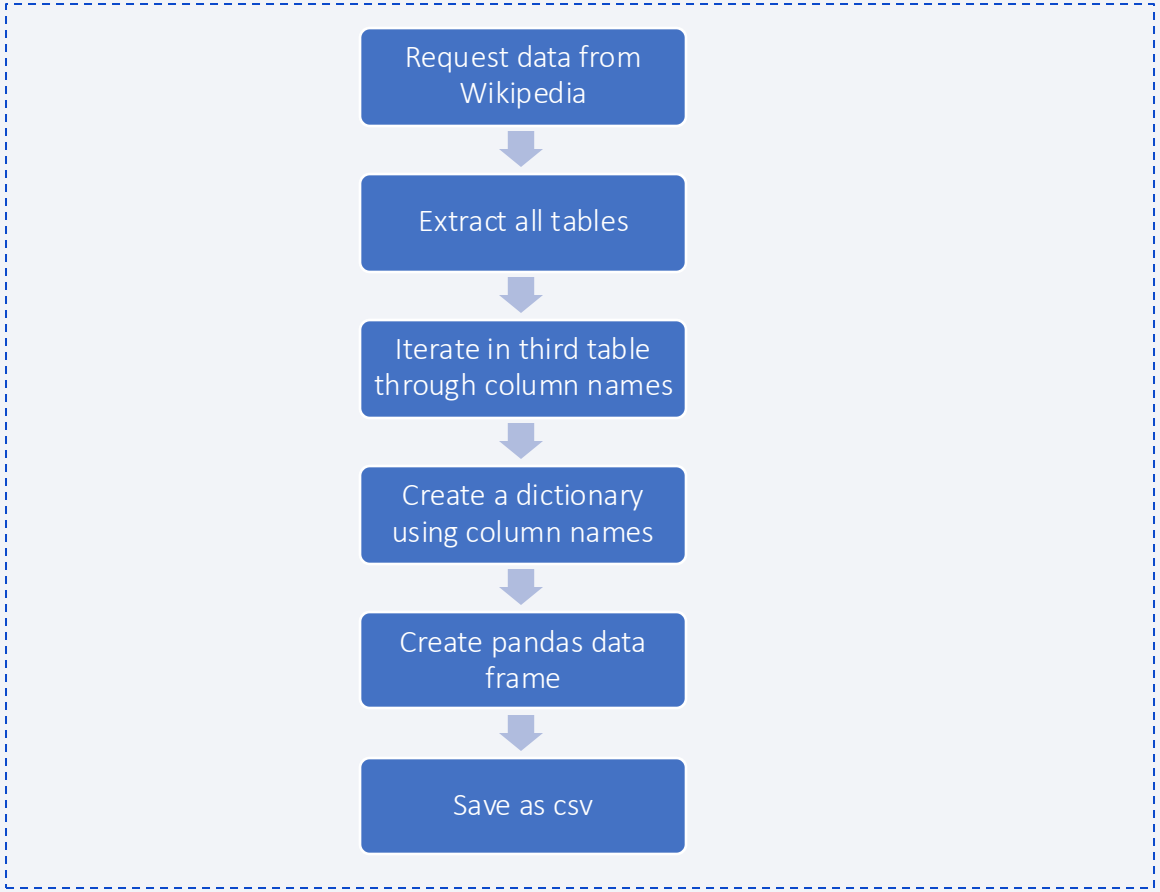
Data Collection – SpaceX API

- Using request.get() method on URL
- Decoding response with .json()
- Converting the data on data frame .json_normalize()
- Extracting more attributes i.e. 'booster name', 'payloads', 'landing outcome'
- Filtering the data frame to only Falcon 9 launches
- Replacing missing values with the mean values
- Export to csv
- GitHub URL : https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

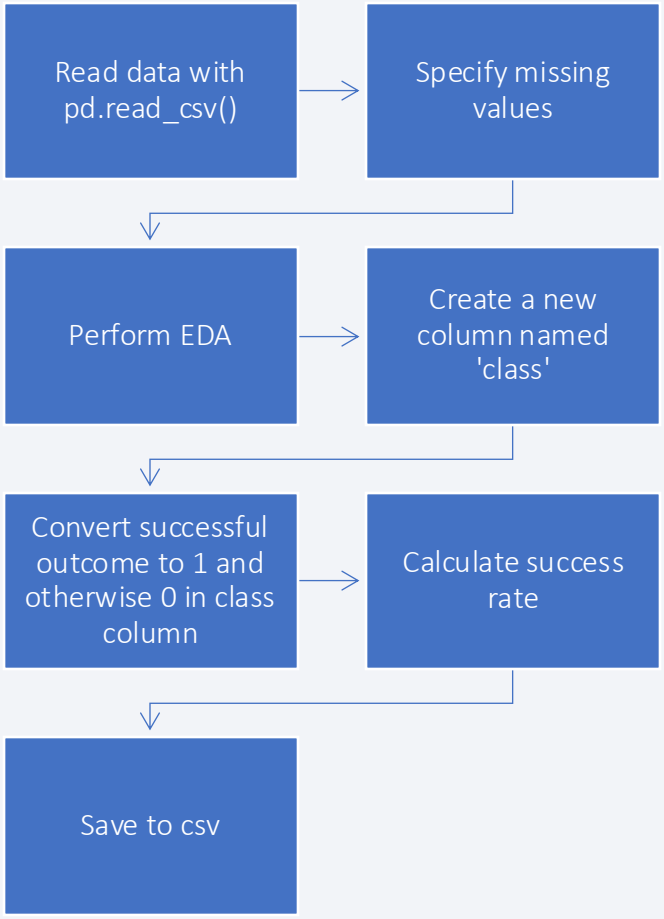


Data Collection - Scraping

- Request data from Wikipedia using requests.get()
- Creating Beatiful Soup objecton on HTML response by using BeatifulSoup() constructor
- Extracting all tables with soup.find_all() method
- In the third table iterating header/column names
- Creating a data frame by parsing launch HTML tables
- GitHubURL: https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling



- Specify the missing values
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
- GitHub URL: https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plot is the best to describe the relation between two variable, so we use scatter plots to visualize how some variables affect the launch outcome e.g
  - Payload mass vs flight number
  - Flight number vs launch site
  - Payload vs launch site
  - Flight number vs orbit type
- Bar plot is the best to compare more than one categorical data, so we use bar plot for success rate of each orbit type
- We use line plot to visualize the yearly trend of the launch success
- GitHub URL:https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/edadataviz%20(1).ipynb

EDA with SQL

- 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 succesful 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 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: https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- To find geographical patterns about launch sites
  - Mark all launch sites on a map with folium.Circle
  - Mark the success/failed launches for each site on the map with folium.map.Marker
  - Create cluster markers with MarkerCluster()
  - Calculate the distances between a launch site to its proximities with MousePosition()
  - Create a marker with distance to a closest city, railway, highway, etc.
  - Draw a line between the marker to the launch site with folium.PolyLine()

- GitHub URL: https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/lab_jupyter_launch_site_location.ipynb

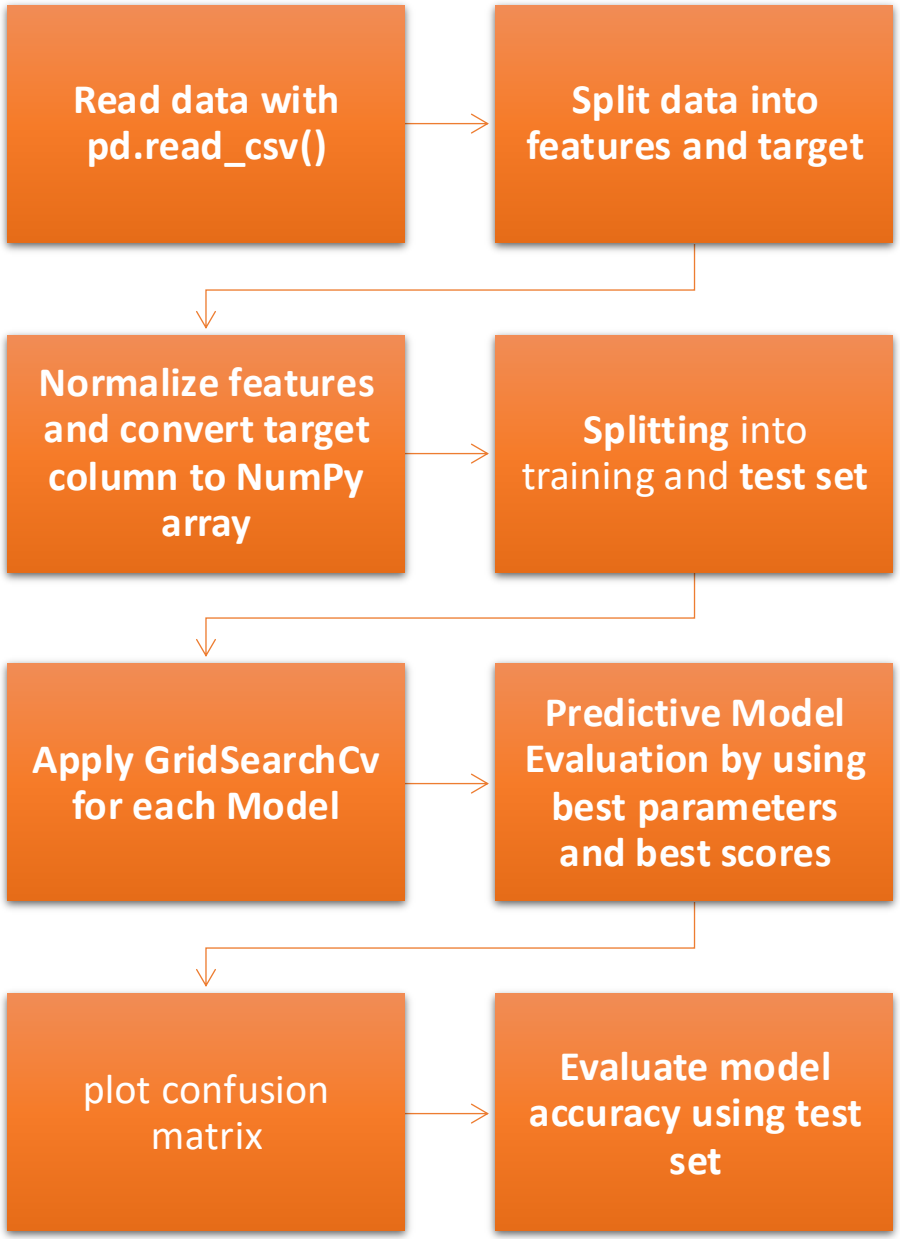
Build a Dashboard with Plotly Dash

- To find a relationship of success rate between launch site and payload
  - Drop-down input containing all the launch sites
  - Rendered pie chart for selected launch site showing the success rate
  - Range slider to specify the range of payload
  - Scatter plot showing the correlation between selected launch site and payload
- GitHub URL : https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/Interactive%20Dash%20Boarding%20Using%20Plotly%20Dash.ipynb

Predictive Analysis (Classification)

- Steps of building and evaluating the classification
  - **Data Collection and Wrangling**: Data was loaded and split into features (X) and target (Y)
  - **Data Standarization**: Features columns were normalized and target column was converted to Numpy array
  - **Splitting**: Data split into training and test set
  - **Predictive Model Evaluation**: GridSearchCV was used to determine best parameters (.best_params_) and best scores (.best_score_) and confusion matrix was plotted
  - **Predictive Model Selection:** Model with highest accuracy is calculated using .score() method
- GitHub URL: https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(2).ipynb

Predictive Analysis (Classification) Flowchart

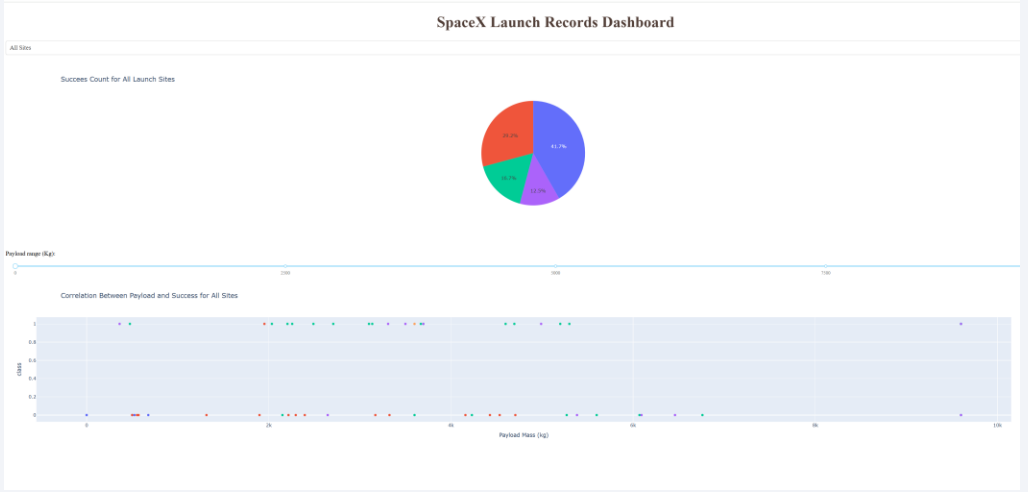


Results

EDA

- KSC LC-39A and VAFB SLC have success rate more then 50%.
- VAFB SLC 4E has no payload more then 10000kg.
- Only in LEO orbit success appears to relate to flight number
- Succes rate is increasing since 2013 till 2020

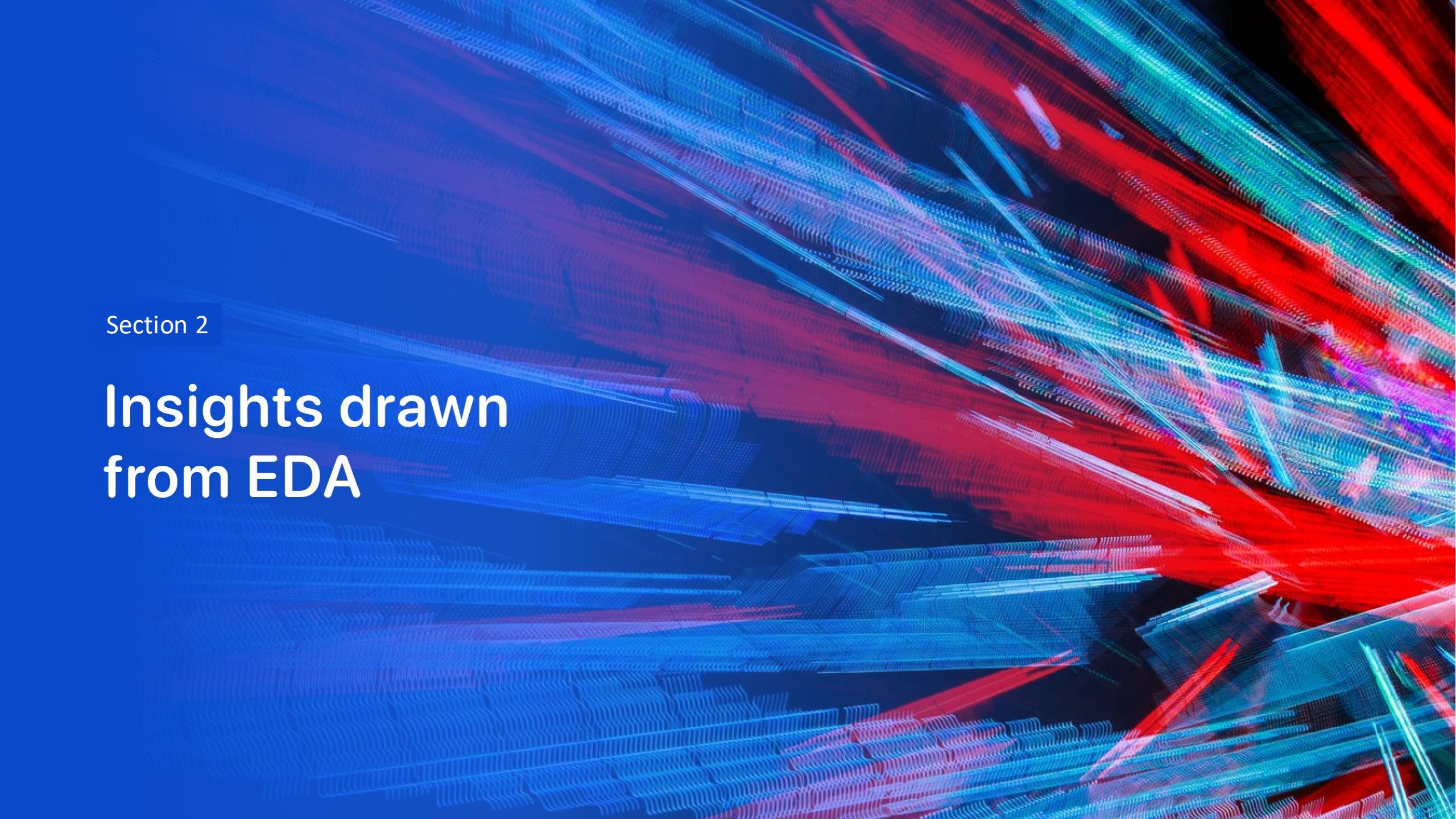
Interactive Analytics



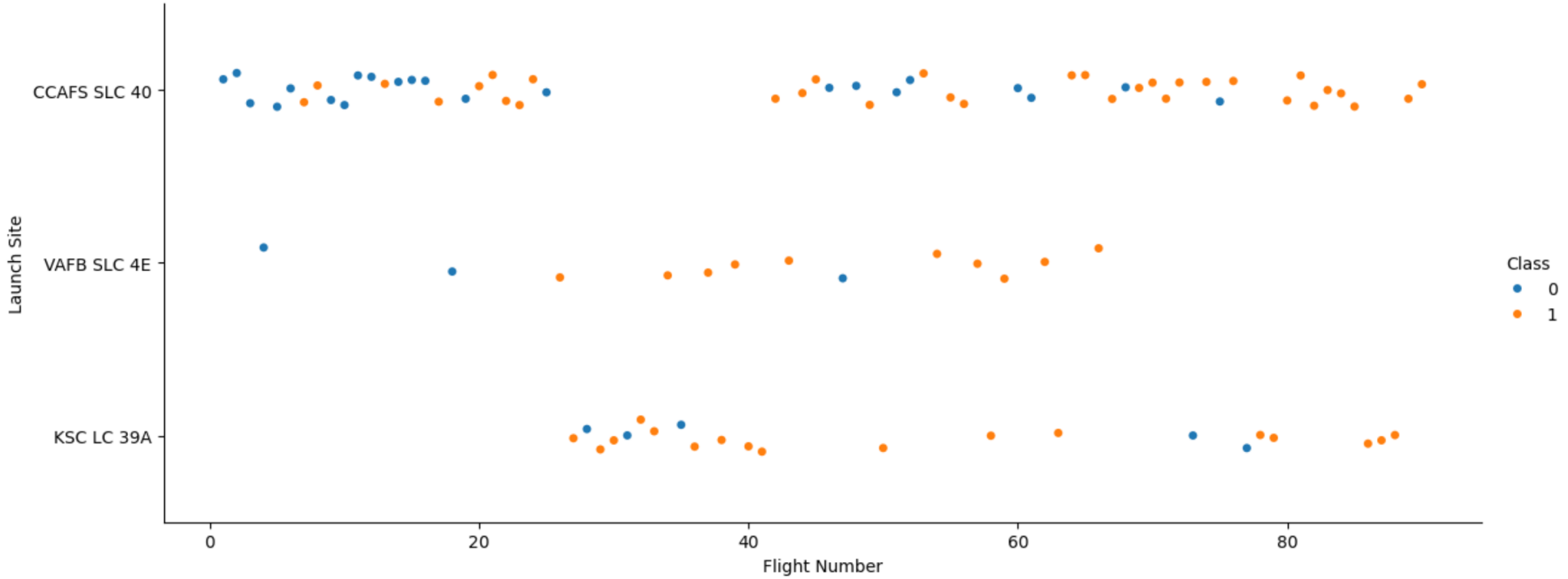
Predictive Analysis

- We compare the models
  - Logistic regression
  - KNN
  - Decision Tree
  - SVM

  Decision Tree has a highest accuracy with 0.87 and is the best model to make a prediction.

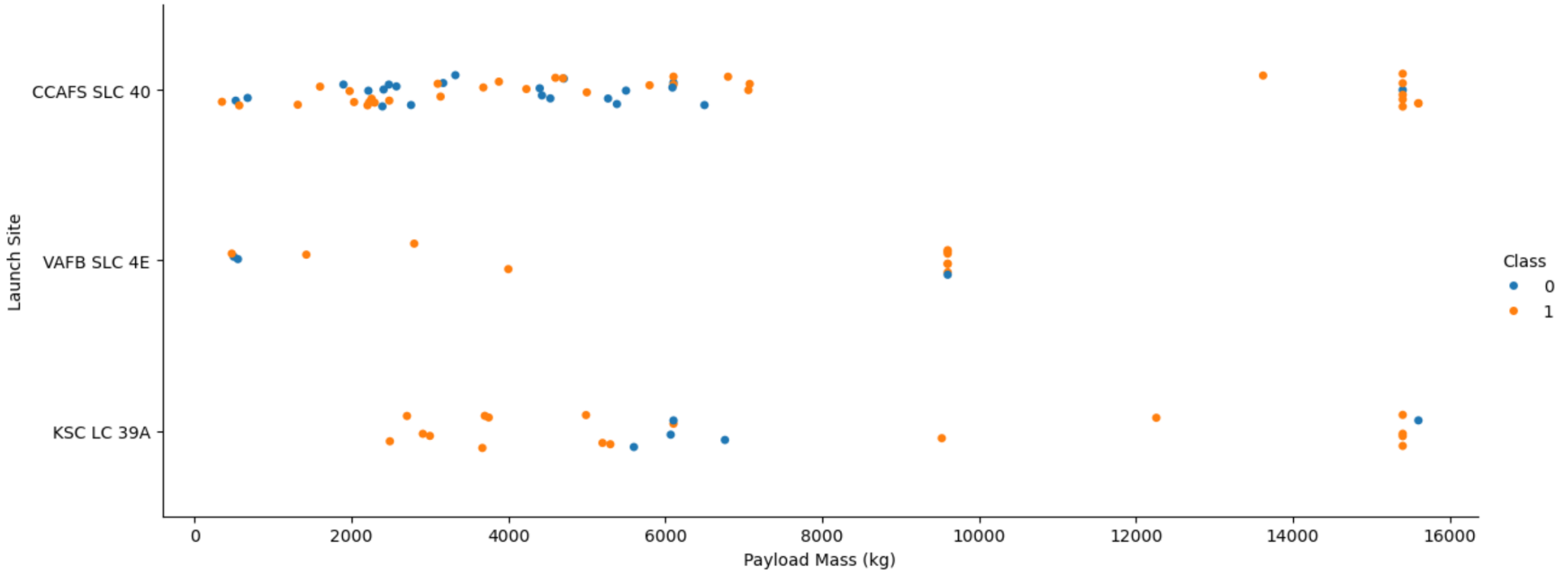

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



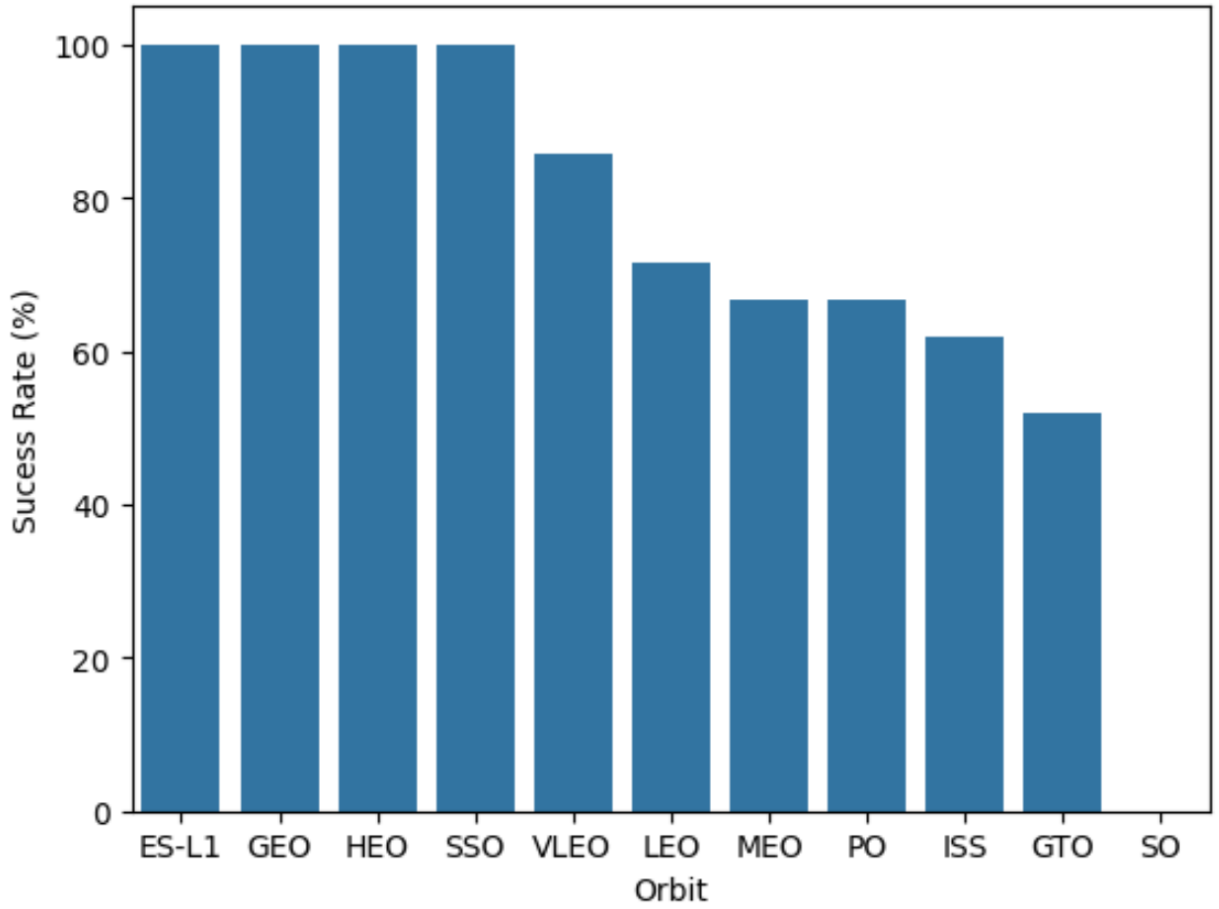
- With the increase of the flight number, the success rate is increasing as well as.
- CCAFS SLC 40 has highest number of rockets compared to the other launches.

Payload vs. Launch Site



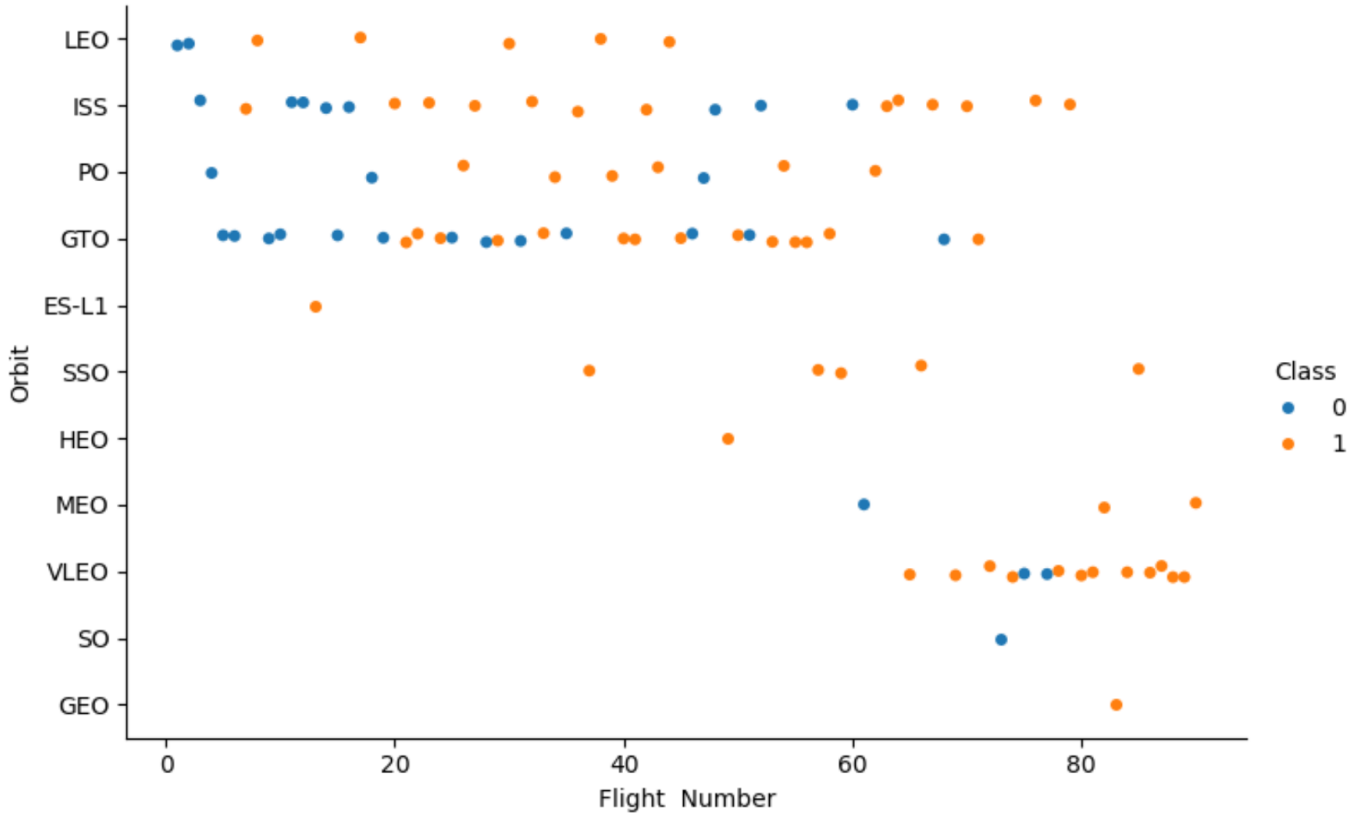
- There is no rockets launches heavy payload mass greater than 1000kg in VAFB SLC 4E.

Success Rate vs. Orbit Type



- ES-L1, GEO, HEO and SSO have the highest success rate.
- SO has the least success rate.

Flight Number vs. Orbit Type

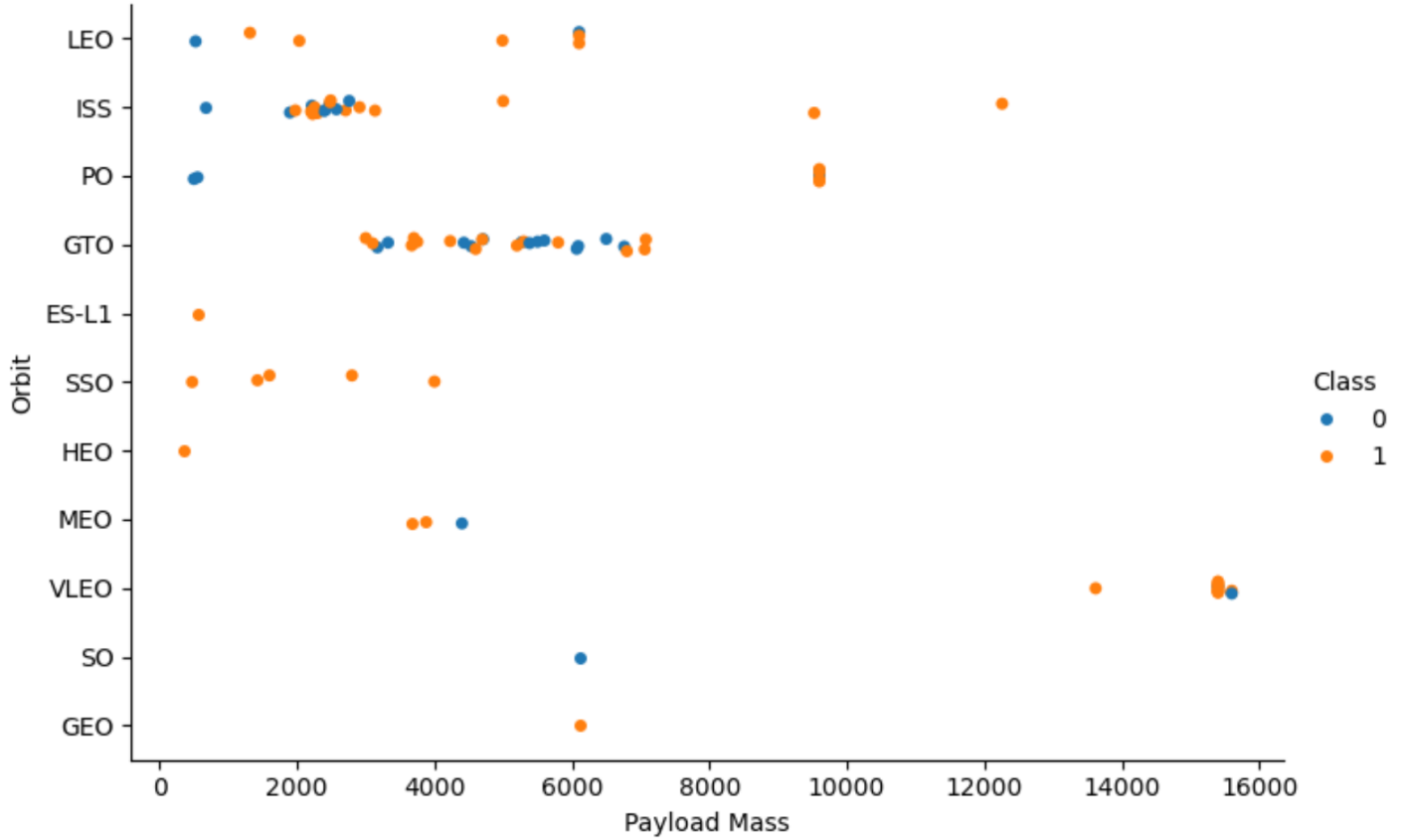


- In the LEO orbit, success seems to be related to the number of flights.
- Conversely, in theother orbits, there appears to be no relationship between flight number and success.

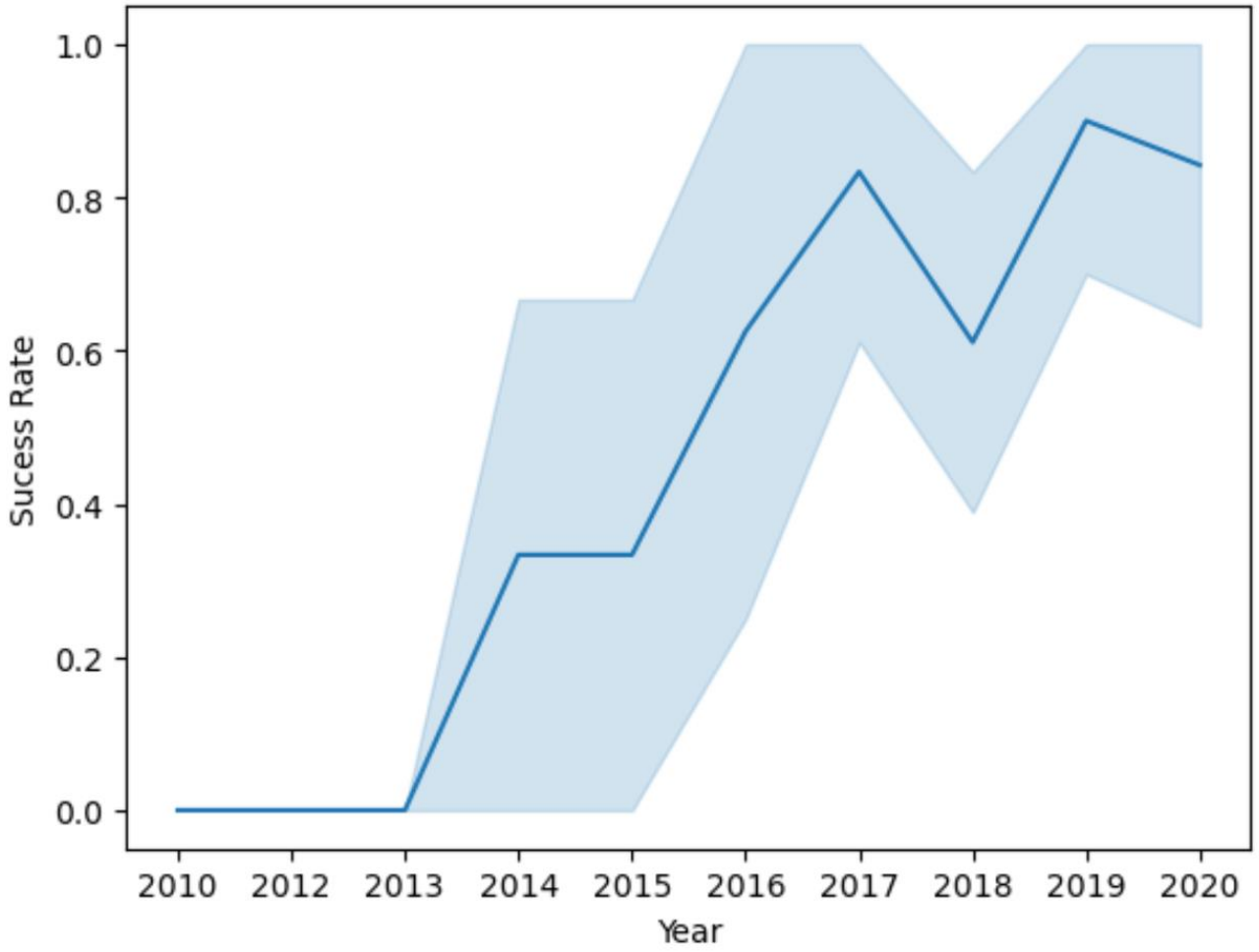
Payload vs. Orbit Type

- In PO, LEO and ISS successful landings are more with heavy payloads.

- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings.



Launch Success Yearly Trend



- It is observed that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

- There are 4 different launches names

```
%sql select distinct Launch_Site as Launch_Sites from spacextable
```

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

Launch Site Names Begin with 'CCA'

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

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	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

- Keyword "LIKE" is used to get names beginning with 'CCA'
- By using LIMIT 5, we can see only 5 results

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) as Payload_Mass_Sum, customer from SPACEXTABLE where customer='NASA (CRS)'
```

- "SUM" function is used to calculate the total payload mass for customers with name 'NASA (CRS)'

Payload_Mass_Sum	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as payload_average, booster_version from spacextable where booster_version ='F9 v1.1'
```

payload_average	Booster_Version
2928.4	F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is calculated by using the "AVG" function.

First Successful Ground Landing Date

```
%sql select min(date) from spacextable where landing_outcome='Success (ground pad)'
```

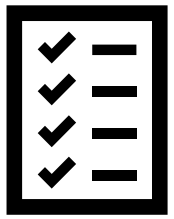
- To find the first successful landing, "MIN" function is used for 'DATE' column.

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from spacextable where landing_outcome='Success (drone ship)' and payload_mass__kg_>4000 and payload_mass__kg_<6000
```

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



- Using "AND" keyword, the desired data is eliminated.
- We listed the successful drone ship landings with payload between 4000 and 6000.
- The result is 4 rockets.

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(mission_outcome) as total_number from spacextable group by "mission_outcome"
```

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

- "COUNT" function is used to find the total number of successful missions and failed missions.
- As a result, there are 99 successful mission, 1 failed and unclear mission.

Boosters Carried Maximum Payload

```
%sql select booster_version from spacextable where payload_mass__kg_=(select max(payload_mass__kg_) from spacextable)
```

- Sub-query and "MAX" function are used to get boosters that carried the maximum payload.
- As a result, there are 12 booster version.

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 select substr(Date, 0, 5) as Year, substr(Date, 6,2) as Month, booster_version, launch_site, landing_outcome from spacextable where substr(Date, 0,
```

Year	Month	Booster_Version	Launch_Site	Landing_Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Substr() is used to get the month and year from the date column.
- As a result, there are two failed launch in 2015 by booster version B1012 and B1015.

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

```
%sql select landing_outcome, count(*) as total from spacextable where Date between '2010-06-04' and '2017-03-20' group by landing_outcome order by total
```

Toggle output scrolling

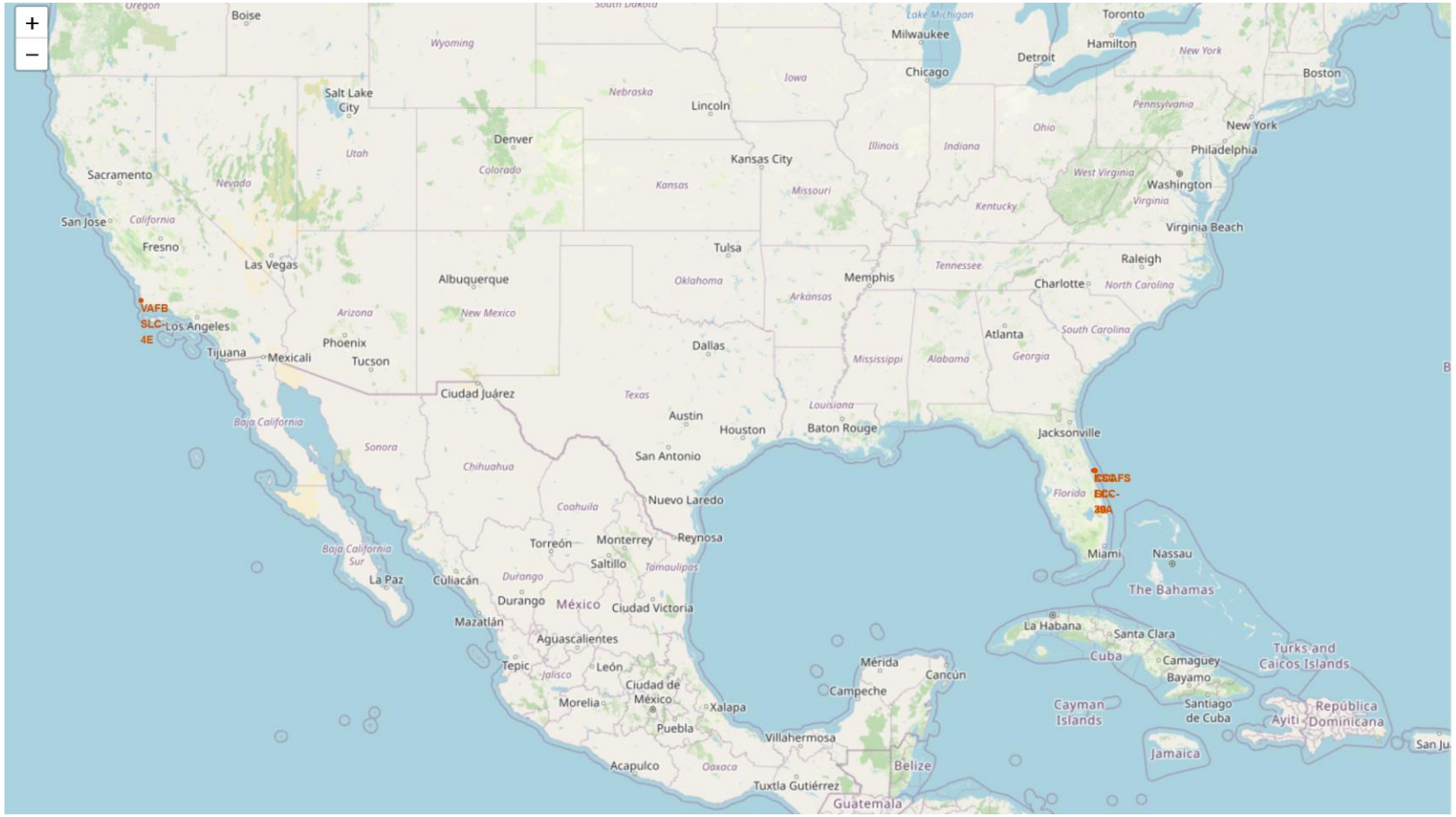
Landing_Outcome	total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- "GROUP BY", "ORDER BY", "DESC", "BETWEEN" and "COUNT" keywords are used to rank landing outcomes between 2010-06-04 and 2017-03-20.

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

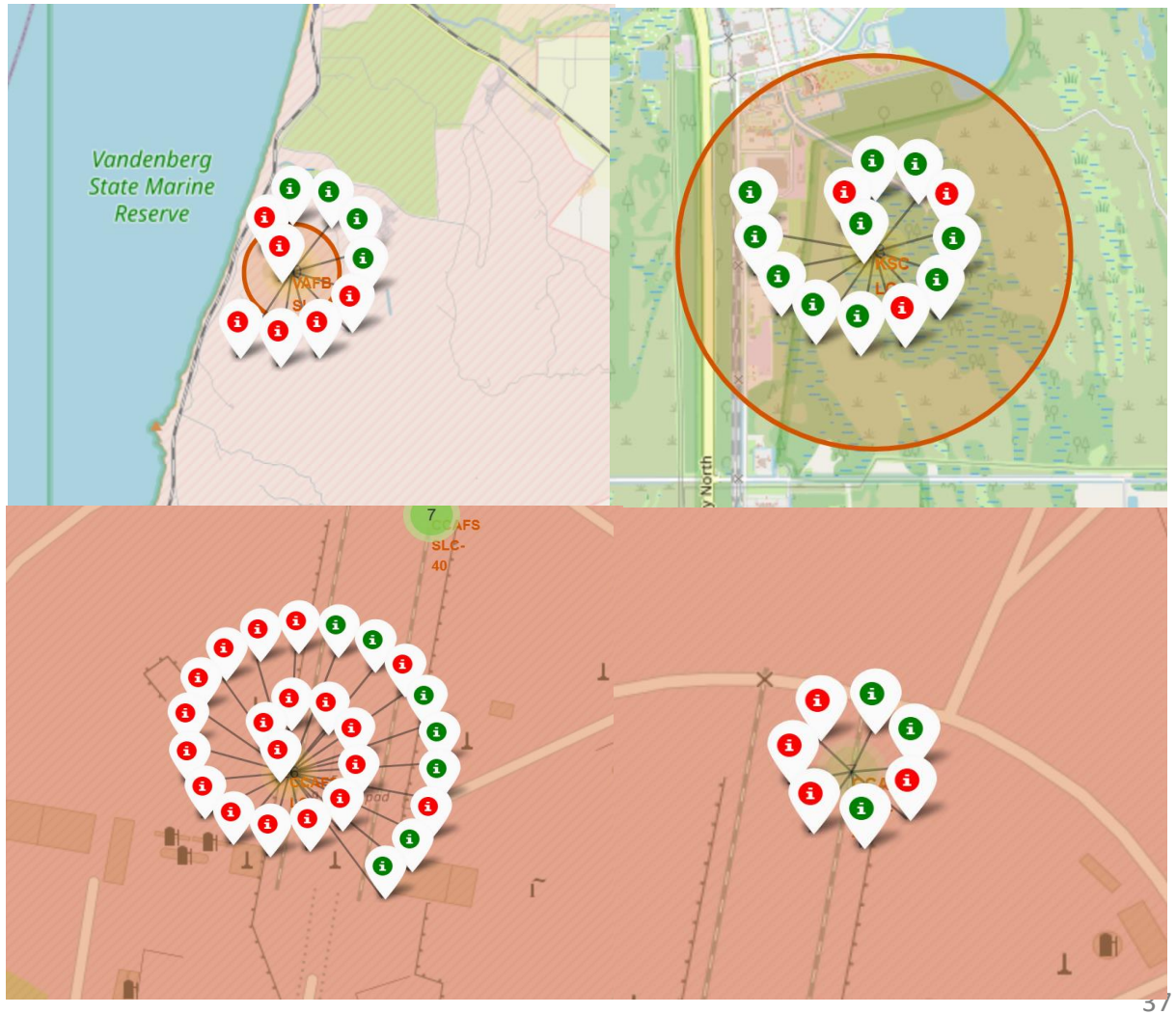


- All launch sites are very close to coast and Equator line.

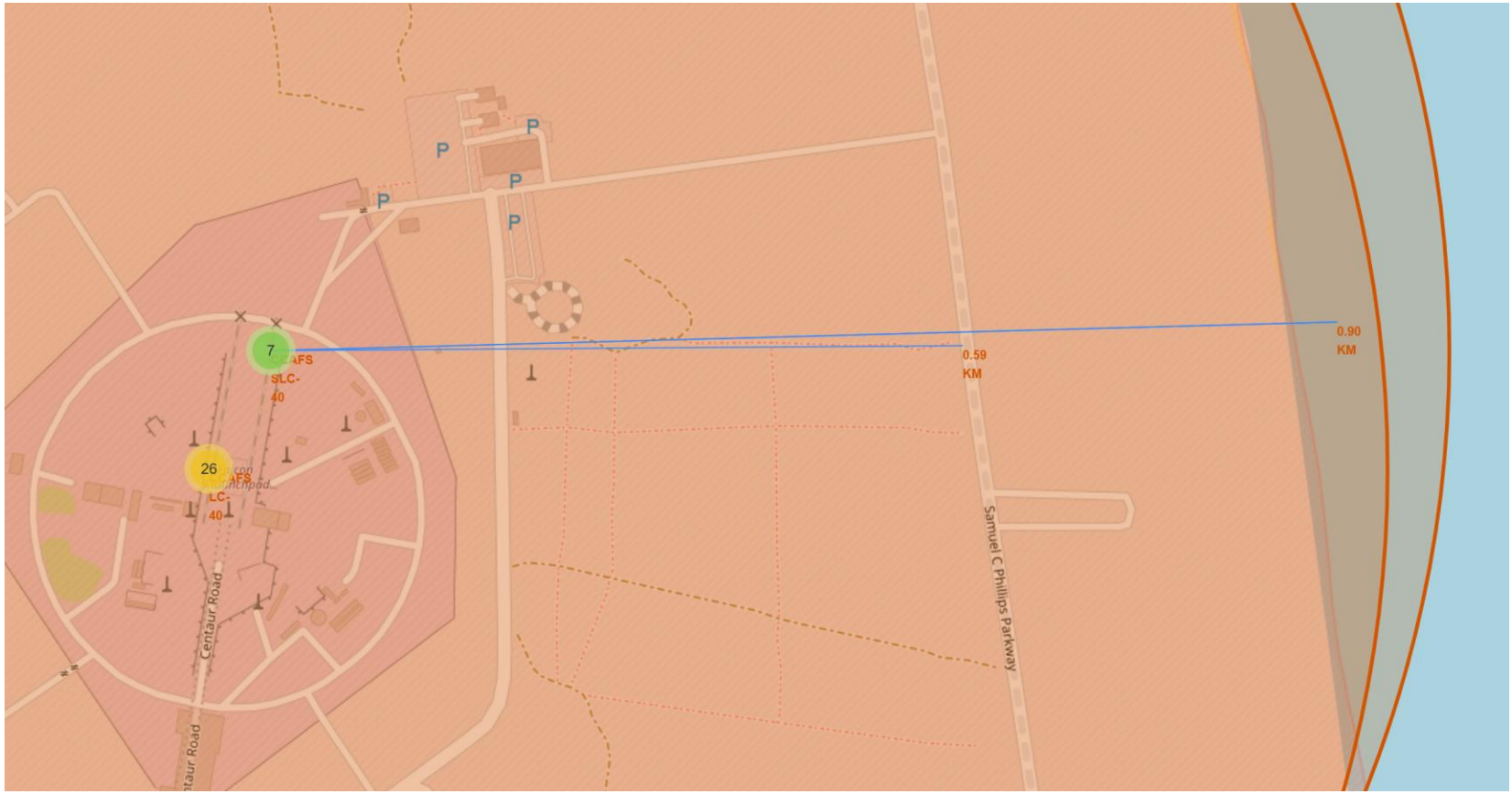
Launch Outcomes

- The map shows the launch outcomes for different :
  - Top left: VAFB SLC-4E
  - Top right: KSC LC-39A
  - Bottom left: CCAFS LC-40
  - Bottom right: CCAFS SLC-40

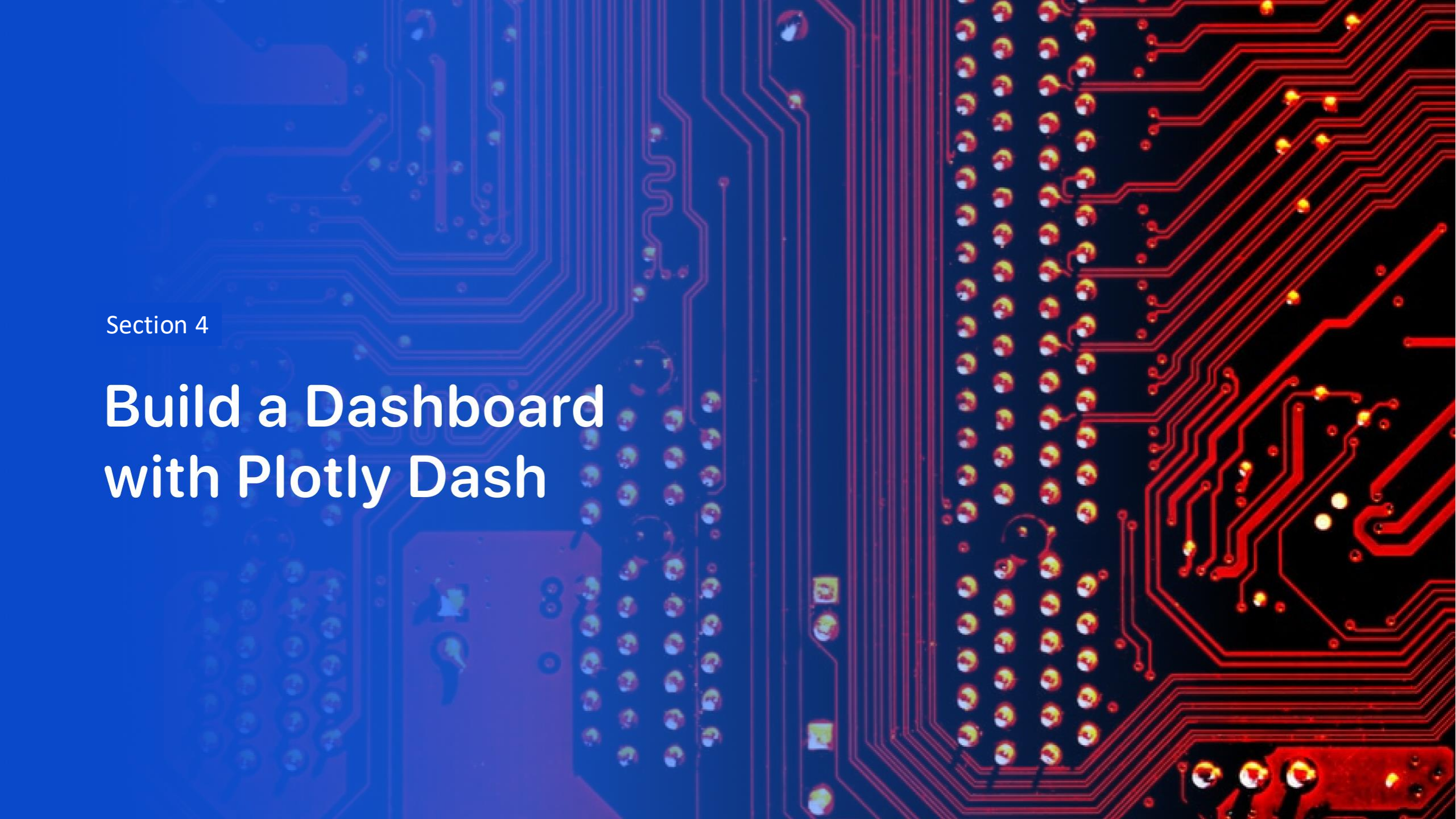
- Red and green icons represent the failed and successful outcomes respectively.



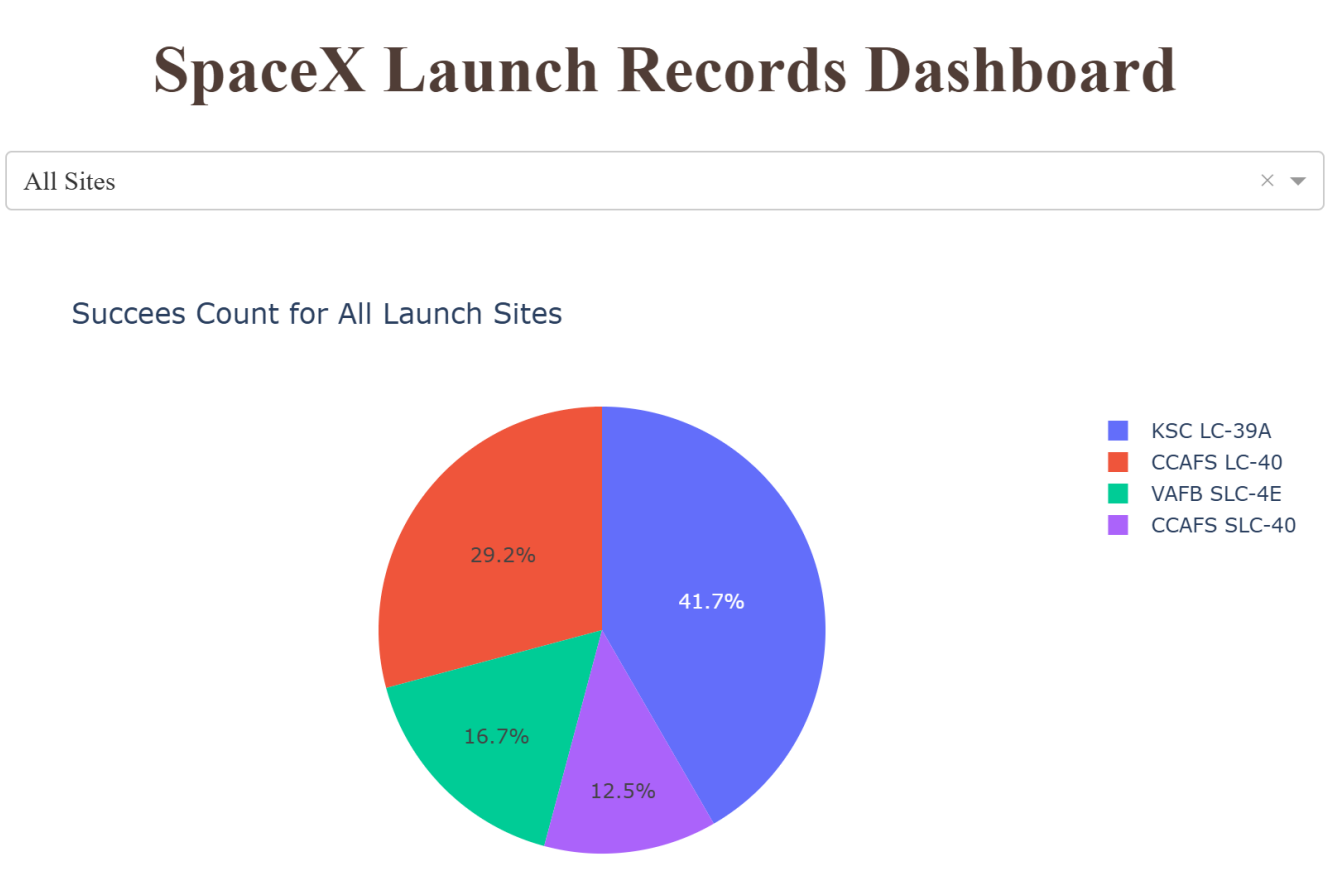
Launch Site Distance



- CCAFS SLC-40 is 0.9 km from the coast and 0.59km from the highway (Samuel C Philips Parkway).
- It is very close to coast and highway.

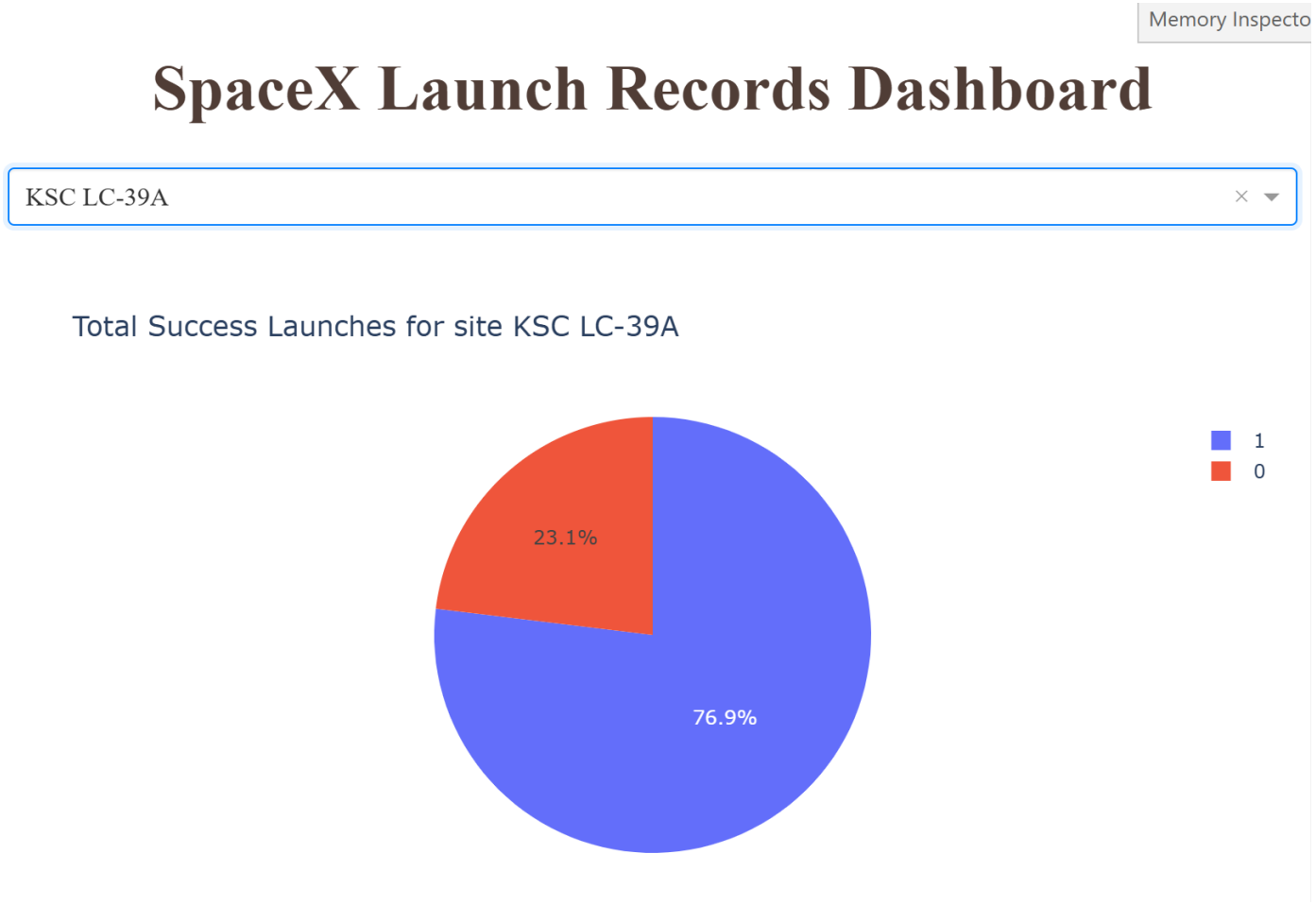

Section 4
Build a Dashboard with Plotly Dash

Success Count for All Launch Sites



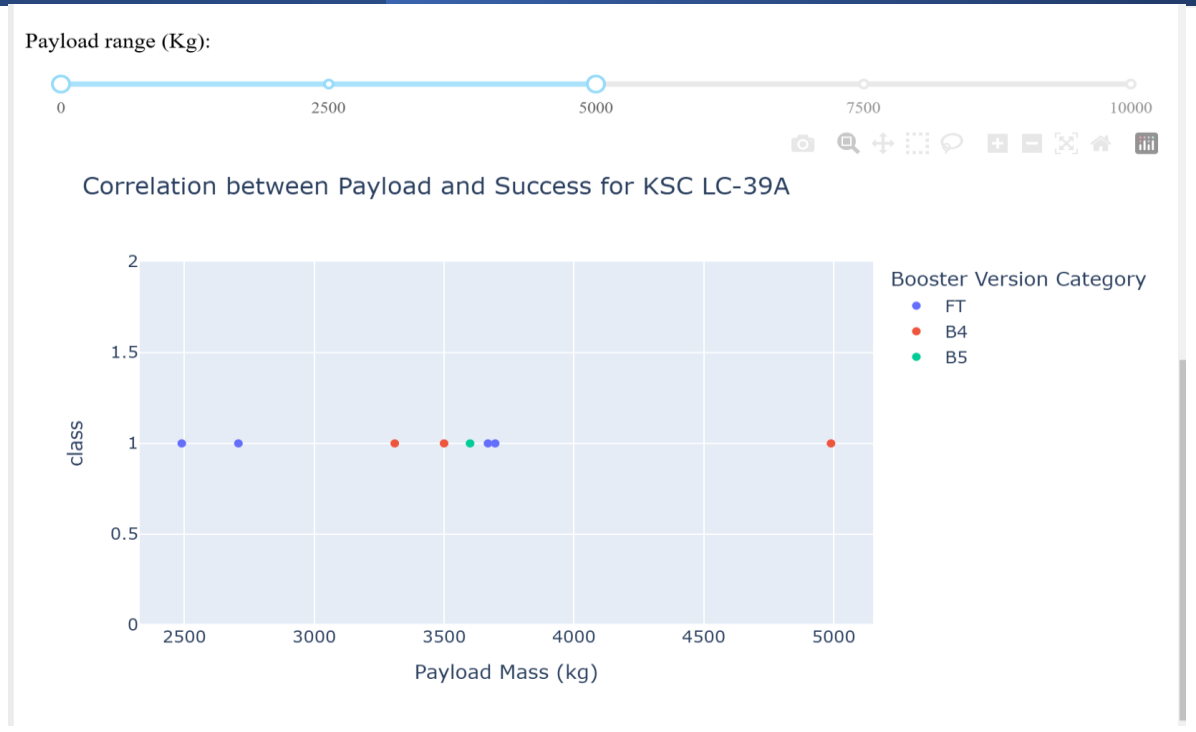
- As it is shown in the pie chart, KSC LC-39A has the largest success rate with 41.7%.
- CCAFS SLC-40 has the smallest succes rate with 12.5%.

Launch Site with Highest Success Ratio



- KSC LC-39A has a highest success rate with 76.9% .
- Only 23.1% of rockets are failed in this launch site.

Payload vs. Launch Outcome



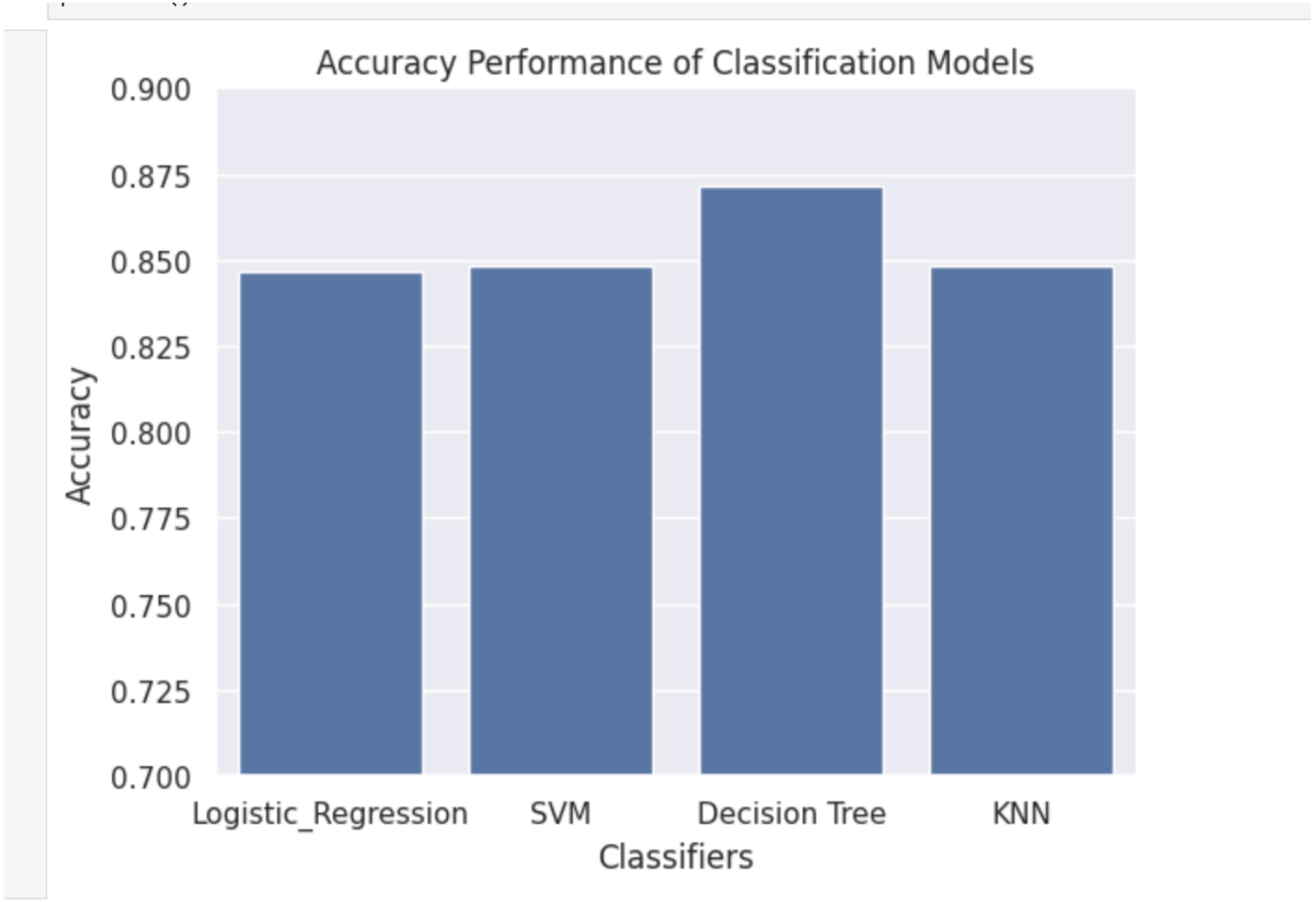
- Above scatter plot shows correlation between payload and success rate for all sites, for the payload range 0 to 10000kg.

- Second chart is for launch KSC LC-39A, which has a highest success rate.
- We can see that, with payload mass between 0 to 10000 kg, there are only 3 booster version is used and all are successful.

Section 5

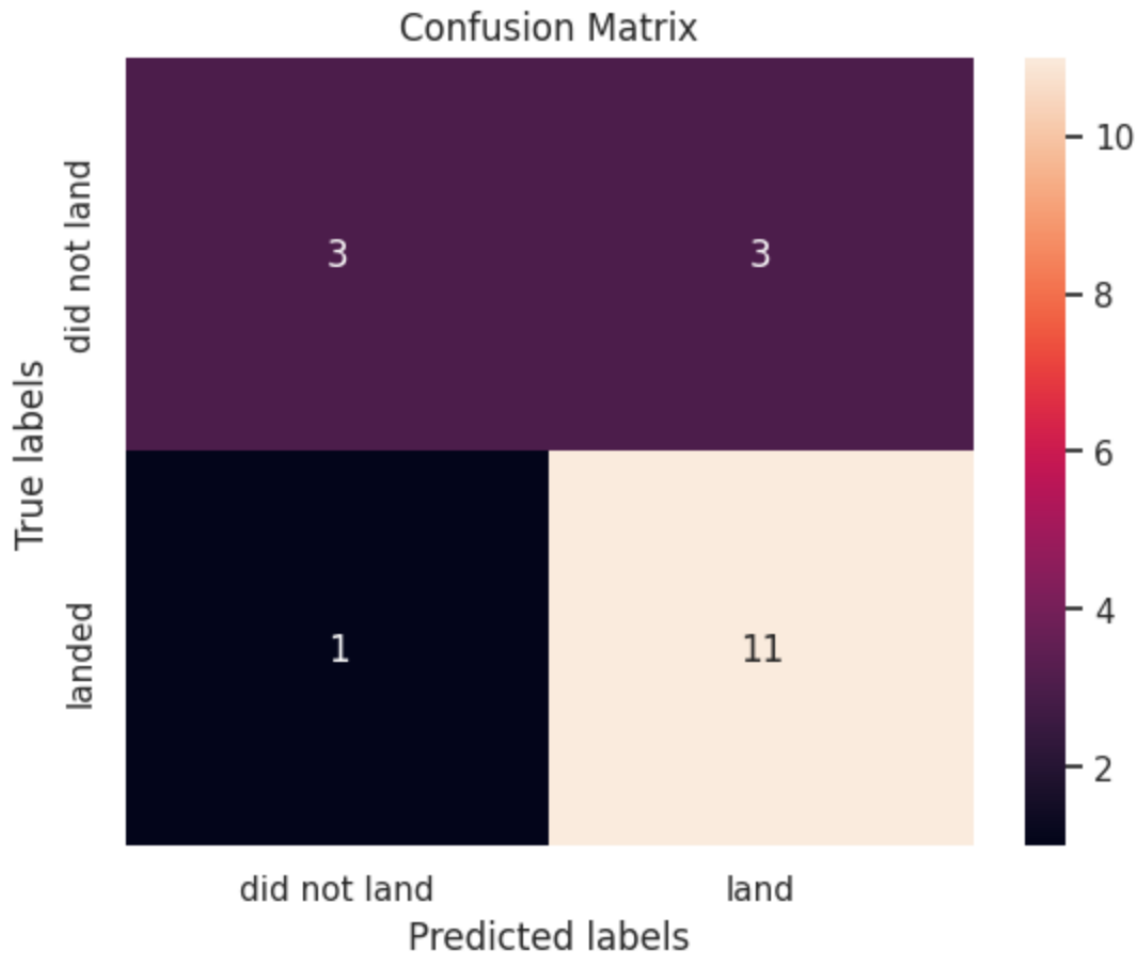
Predictive Analysis (Classification)

Classification Accuracy



- From bar chart, we can see that decision tree classifier has the best accuracy score with approximately 87.5%.
- Other classification models have the almost same accuracy score which is app. 84%.

Confusion **Matrix**



- From the test set, decision tree classifier was able to predict 11 of 12 landed observations correctly. This called true positives. Only 1 of 12 is wrongly estimated. (false negative)
- Also 3 of 6 failed observations are correctly predicted. This means 3 'did not land' estimation is true negative.
- There are also 3 incorrectly 'land' predicted observation. This is false positive.

Conclusions

- Launch site KSC LC-39A has a highest success rate.
- There is correlation between payload mass and success rate for some launch sites. More massive the payload, less likely to successful landing.
- Orbit type also should be considered. VLEO, ES-L 1, HEO and SSO have a larger success rate, whereas SO has the lowest success rate.
- There has been an increase in the success rate since 2013 till 2020.
- Lauch sites very close to coast but enough far away from the crowded places.
- Compared to other classification models, decision tree classification model has the highest accuracy. It can be used landing outcome prediction.

Appendix

- GitHub Repo: https://github.com/AysegulTkn/SpaceX-Falcon-9-first-stage-Landing-Prediction
- For more graphs, you can also see : https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Thank you!