IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Smit Kalamkar
04/05/2024

Outline

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

Executive Summary

- Summary of methodologies:
  - The data was collected using spacex api and web scrapping using beautiful-soup.
  - EDA was performed which stands for Exploratory Data Analysis.
  - EDA includes Data Wrangling and Data Visualization. Also, Interactive dashboards were created using plotly.
  - Predictions were made using multiple machine learning algorithms.
- Summary of all results:
  - The EDA process provided us with the insights into the data and helped to identify which features are best for the prediction of successful landing.
  - Machine learning algorithms provided us with the statistical prediction using different models and their accuracy. We also plotted the confusion matrix to analyze predictions in detail.

Introduction

- Project background and context:
  - The new company SpaceY wants to compete with SpaceX founded by Allon Musk. We have to evaluate the potentiality of this company to compete with SpaceX.
- Problems you want to find answers:
  - Will the first stage land successfully?
  - Will SpaceX reuse the first stage?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - ○ Data was collected using two sources:
  - ○ 1 – SpaceX API
  - ○ 2 – Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - After analyzing the results, we converted the landing outcome into classes.
    - ▪ Class 0: Bad outcome
    - ▪ Class 1: Good outcome
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was divided into training and test sets. The data was evaluated by different classification models Logistic Regression, SVM, KNN and Decision Tree
  - Accuracy and confusion matrix was plotted for each algorithms for detailed analysis.

Data Collection

- Data was collected from two sources:

  1) SpaceX API: The link containing all the documentation of the API is given below:

     - Link: https://github.com/r-spacex/SpaceX-API/tree/master/docs#rspacex-api-docs

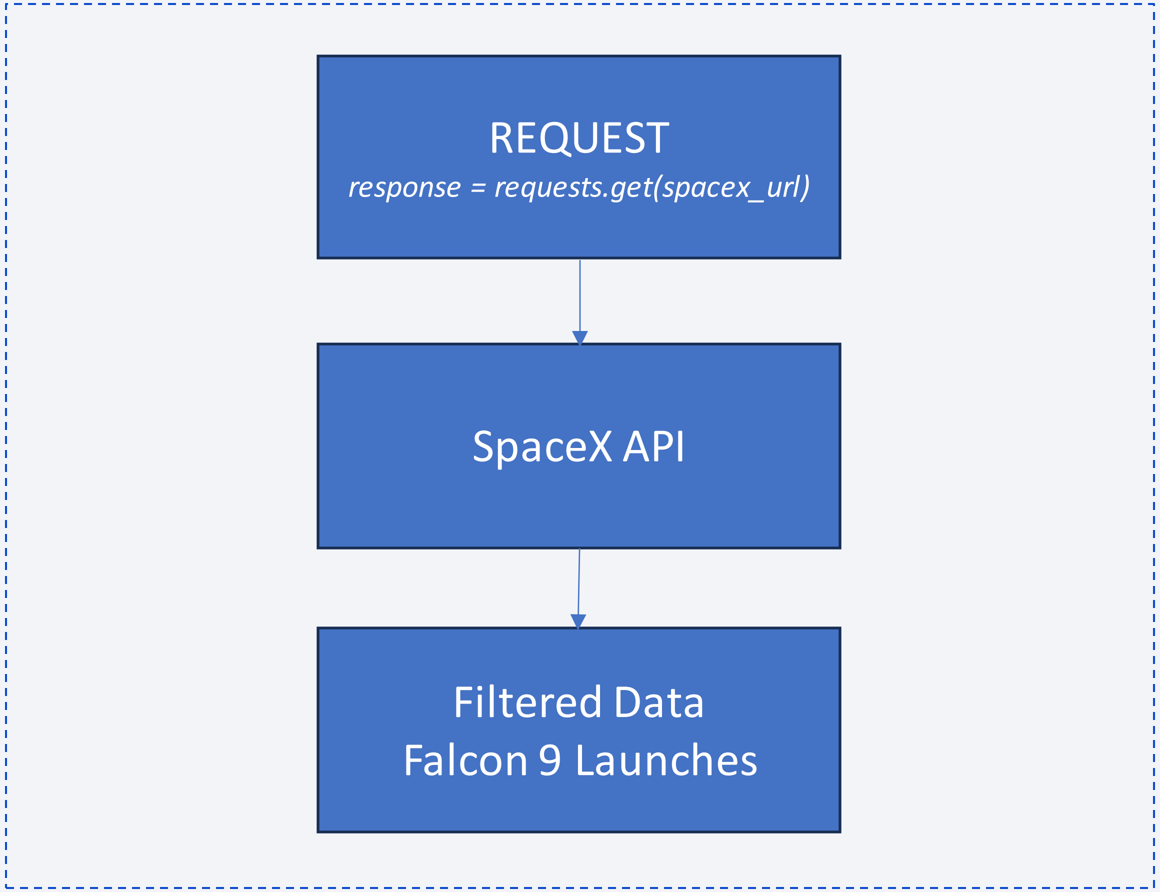
  2) Wikipedia: The data was scrapped from the table in the link given below:

     - Link: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

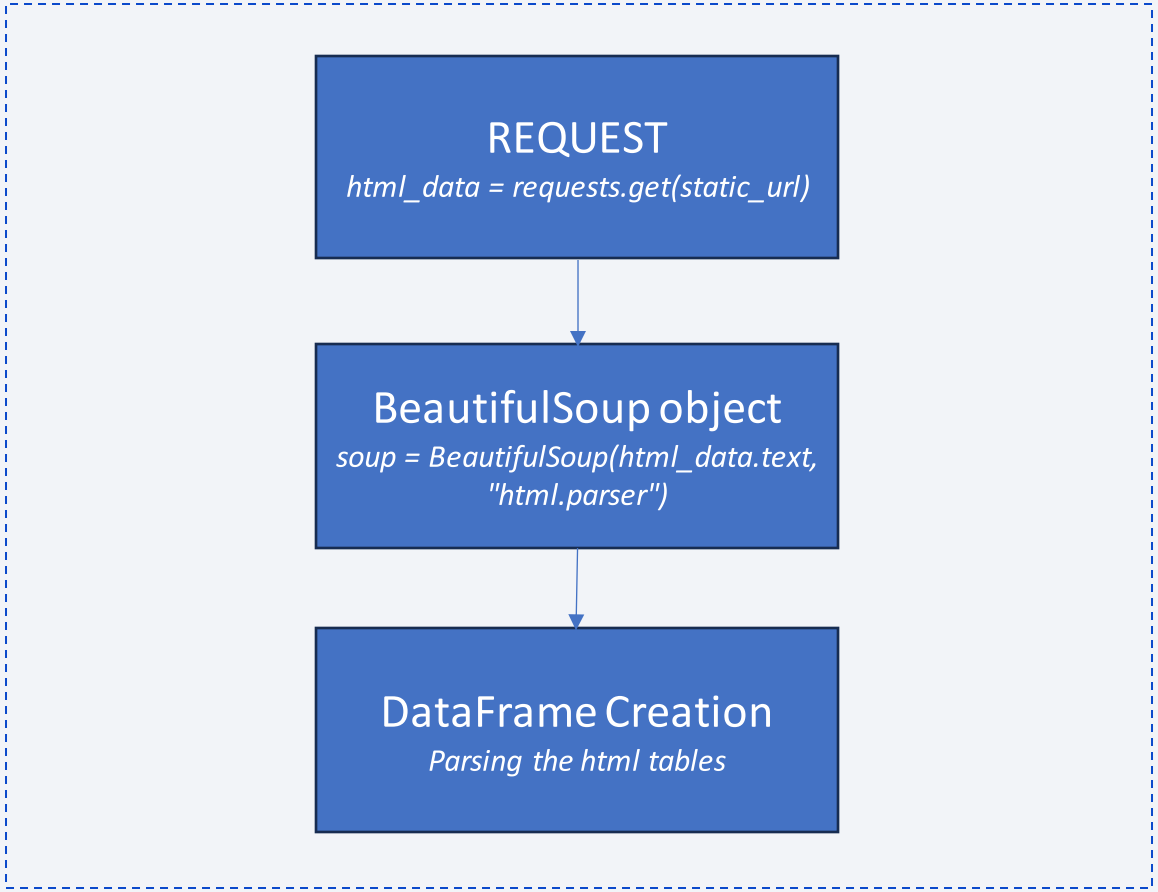
- A data can be collected from the public api provided by SpaceX.
- The steps involved in the process are explained in the flowchart.

- *CODE: https://github.com/smit-vk/capstone_project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb*



Data Collection - Scraping

- For the web scrapping, we first have to request the html data from url.
- Then we have to create a BeautifulSoup object in order to retrieve the required information from the data.
- *CODE: https://github.com/smit-vk/capstone_project/blob/main/jupyter-labs-webscraping.ipynb*



REQUEST

html_data = requests.get(static_url)

BeautifulSoup object

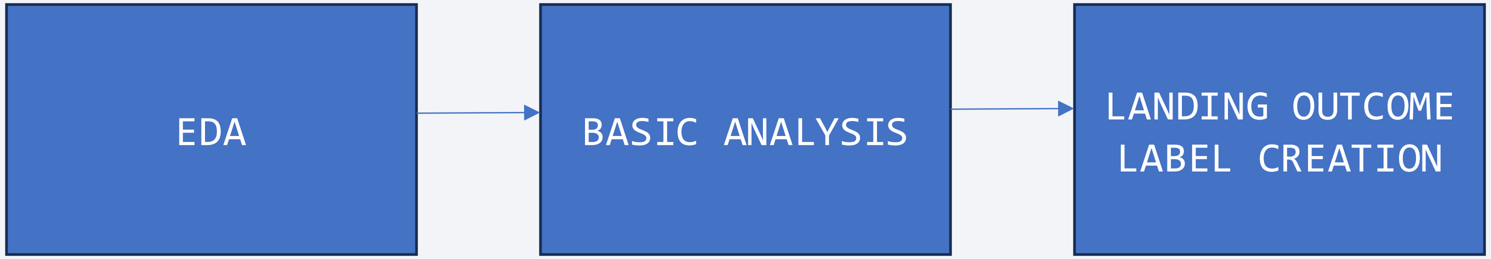
soup = BeautifulSoup(html_data.text, "html.parser")

DataFrame Creation

Parsing the html tables

Data Wrangling

- The EDA – Exploratory Data Analysis was performed on the dataset where null values were identified in order to analyze the data further.
- Basic analysis like number of launches on each site, number of occurence of each orbit, mission outcome per orbit type.
- Landing outcome label was created from outcome column.
- *CODE: https://github.com/smit-vk/capstone_project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb*



EDA

BASIC ANALYSIS

LANDING OUTCOME LABEL CREATION

EDA with Data Visualization

- To analyze the data, some relationships were visualized using bar and scatter plots.
  - Flight No and Payload Mass
  - Flight No and Launch Site
  - Payload and Launch Site
  - Flight No and Orbit Type
  - Payload and Orbit Type
  - Yearly Launch Success Trade

- *CODE: https://github.com/smit-vk/capstone_project/blob/main/edadataviz.ipynb*

EDA with SQL

- The SQL queries performed are listed below:
  - To display the names of the unique launch sites in the space mission
  - To display 5 records where launch sites begin with the string 'CCA'
  - To display the total payload mass carried by boosters launched by NASA (CRS)
  - To display average payload mass carried by booster version F9 v1.1
  - To list the date when the first successful landing outcome in ground pad was achieved.
  - To list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - To list the total number of successful and failure mission outcomes
  - To list the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - To 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
- *CODE: https://github.com/smit-vk/capstone_project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb*

Build an Interactive Map with Folium

- We used Markers, Circles, Line and Marker Cluster to describe the areas and lines on the folium map.
  - Markers indicated the launch sites on the map.
  - Circles were used to indicate areas around specific coordinates to highlight them.
  - Lines are used to indicate distances between two points. For example, to indicate the distance between space station and the nearest coast line.
  - Markers clusters were used to display the group of events. For example, to display the successful and failed launches.

- *CODE: https://github.com/smit-vk/capstone_project/blob/main/lab_jupyter_launch_site_location.ipynb*

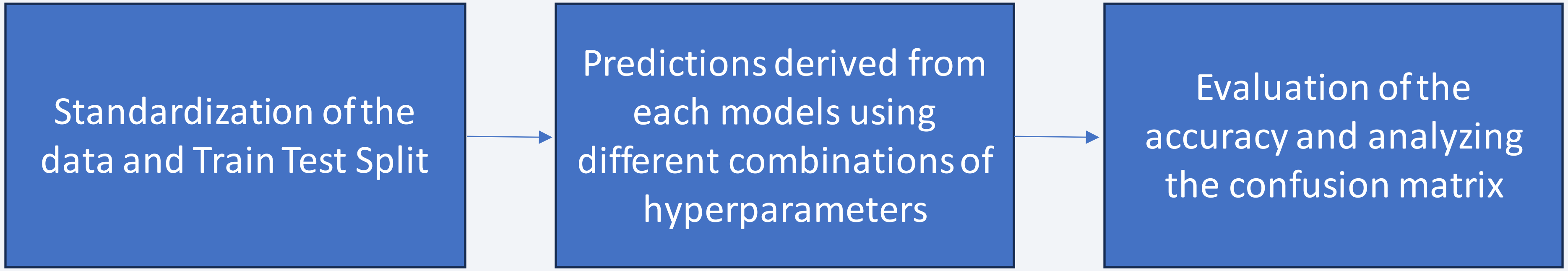
Build a Dashboard with Plotly Dash

- Percentage of total launches by site and Payload range were plotted separately as two different graphs.
- Using this graphs helps in the decision making process to find the best place to launch after analyzing visually.

- *DATA:* *https://github.com/smit-vk/capstone_project/blob/main/spacex_dash_app.py*

Predictive Analysis (Classification)

- We used four classification models for the predictive analysis.
  - Logistic Regression
  - Support Vector Machine (SVM)
  - K-Nearest Neighbour (KNN)
  - Decision Tree



Standardization of the data and Train Test Split

Predictions derived from each models using different combinations of hyperparameters

Evaluation of the accuracy and analyzing the confusion matrix

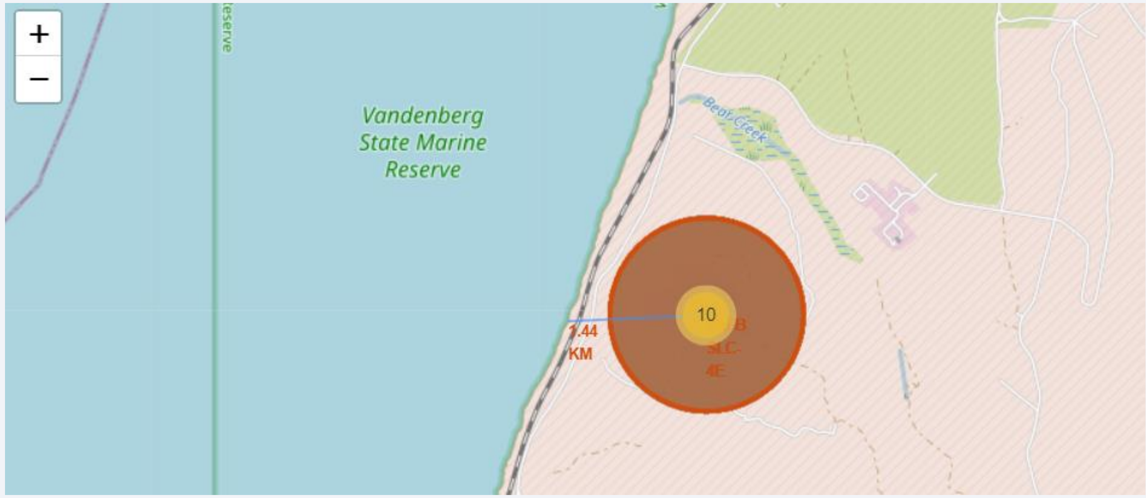
- CODE: *https://github.com/smit-vk/capstone_project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb*

Results

- Exploratory data analysis results
  - We could see that different launch sites has different success rate. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
  - For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass.
  - ES-L1, GEO, HEO, SSO, and VLEO are the Orbits that have high success rate. The SO has the least success rate amongst the orbits.
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
  - The success rate since 2013 kept increasing till 2020. It shows that the landing outcome became successful and better with year.
  - Total payload mass carried by boosters launched by NASA: 45596
  - Average payload mass carried by booster version F9 v1.1: 2928.4
  - Date when the first successful landing outcome in ground pad was achieved: 2015-12-22. It is five years after the first launch.
  - Total number of successful and failure mission outcomes is 100 and 1 respectively. Chances of success is almost100%

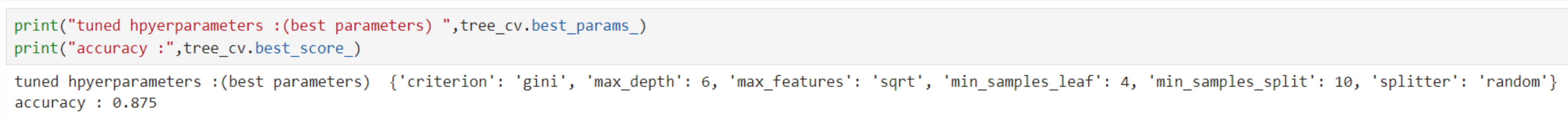
Results

- We used folium for the interactive analysis. As shown in the map, we can easily say that the launch sites are near the coast lines considering the safety. Also there are highways and railways nearby.



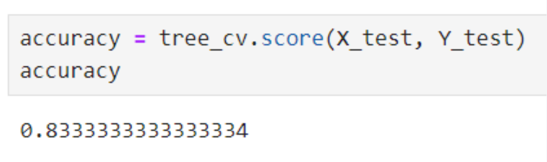
Results

- Predictive analysis results:
  - We used four classification models in our project. Logistic Regression, SVM, KNN and Decision Tree.
  - Analyzing the accuracy and confusion matrix, it is evident that the best model to predict the successful landing is Decision Tree. It has the best accuracy scores.



```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}
accuracy : 0.875
```



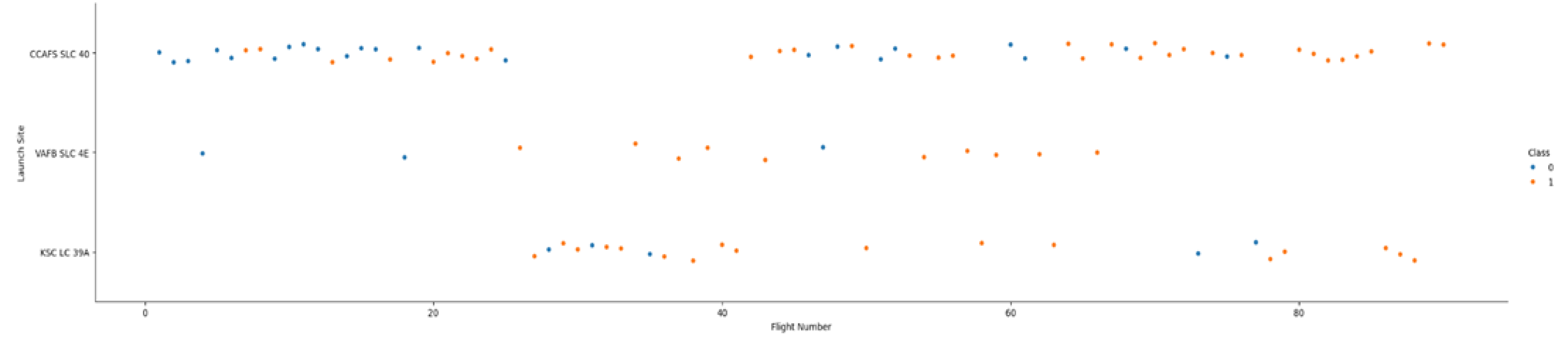
```
accuracy = tree_cv.score(X_test, Y_test)
accuracy
```

```
0.8333333333333334
```

Section 2

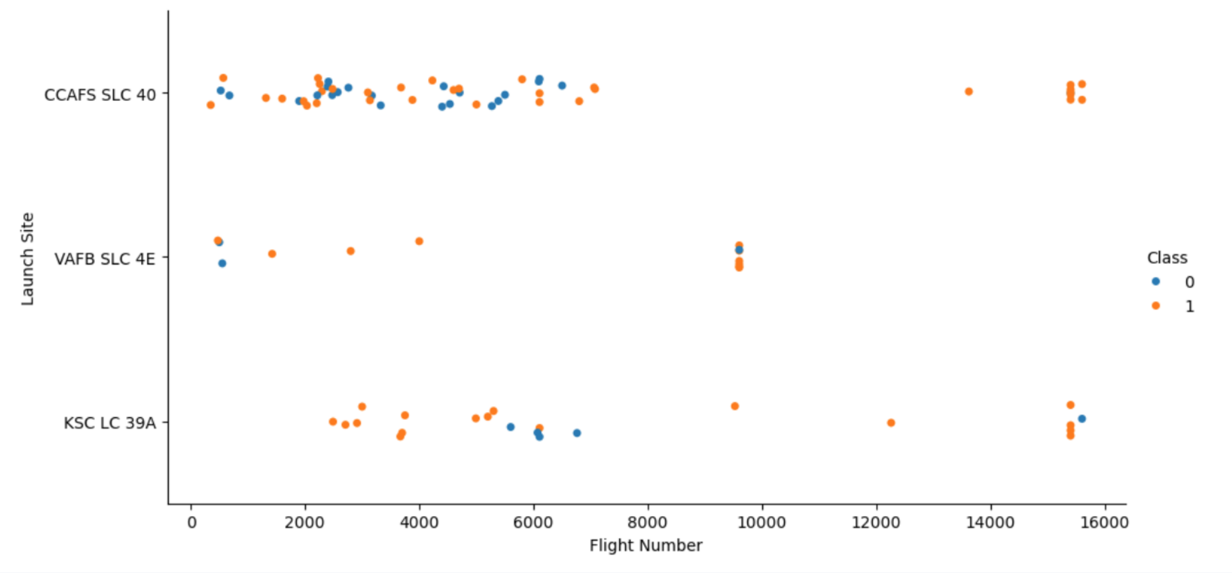
Insights drawn from EDA

Flight Number vs. Launch Site



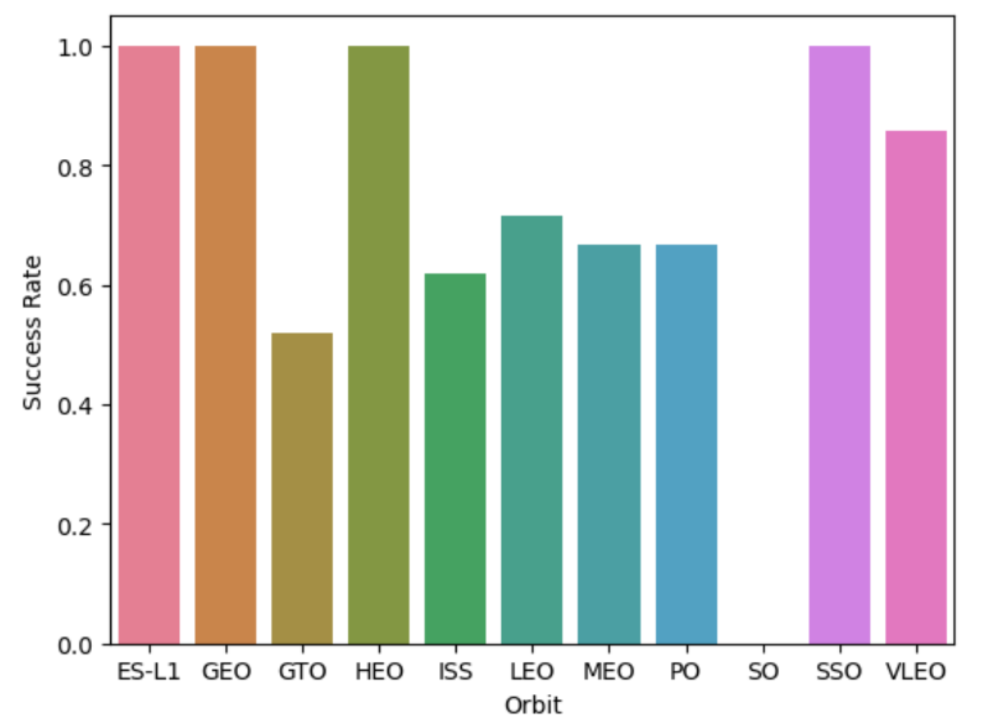
- From the graph, we can say that the most recent launches are happening at CCAF5 SLC 40. Also, it is quite evident that the most recent launches are successful.
- From the graph, one can easily conclude that the most recent launches are successful for all the three launch sites.

Payload vs. Launch Site



- From the plot, we can say that the launches with excessive payload over 12,000 are made from CCAFS SLC 40 and KSC LC 39A. Also, landings seems to have higher success rate when the payloads are over 8000.

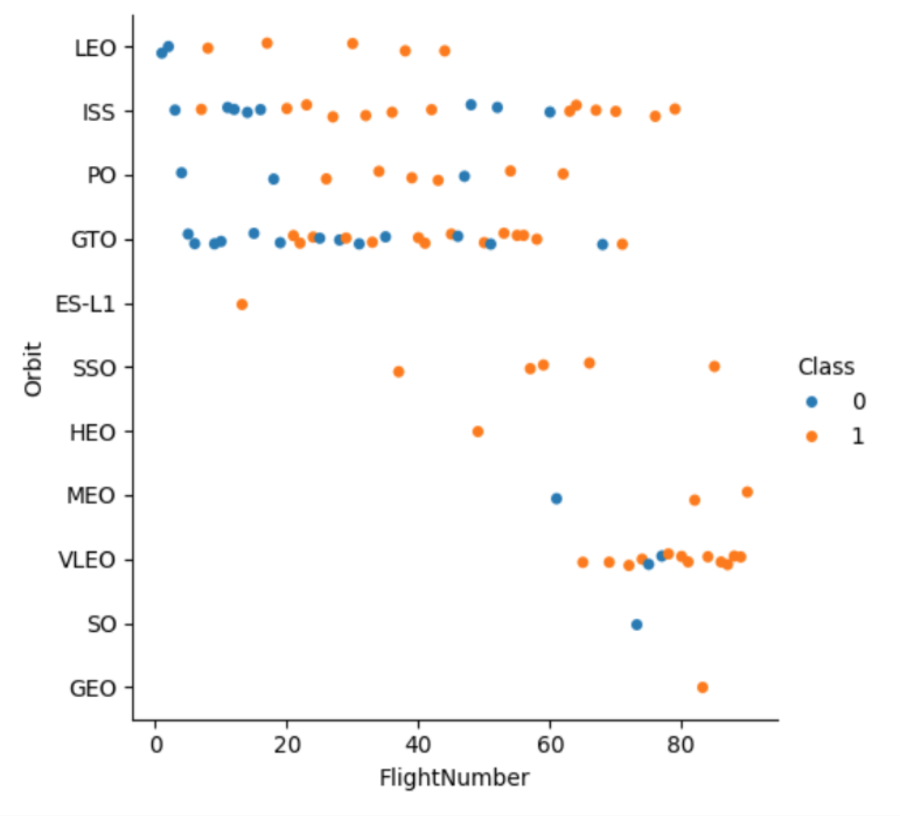
Success Rate vs. Orbit Type



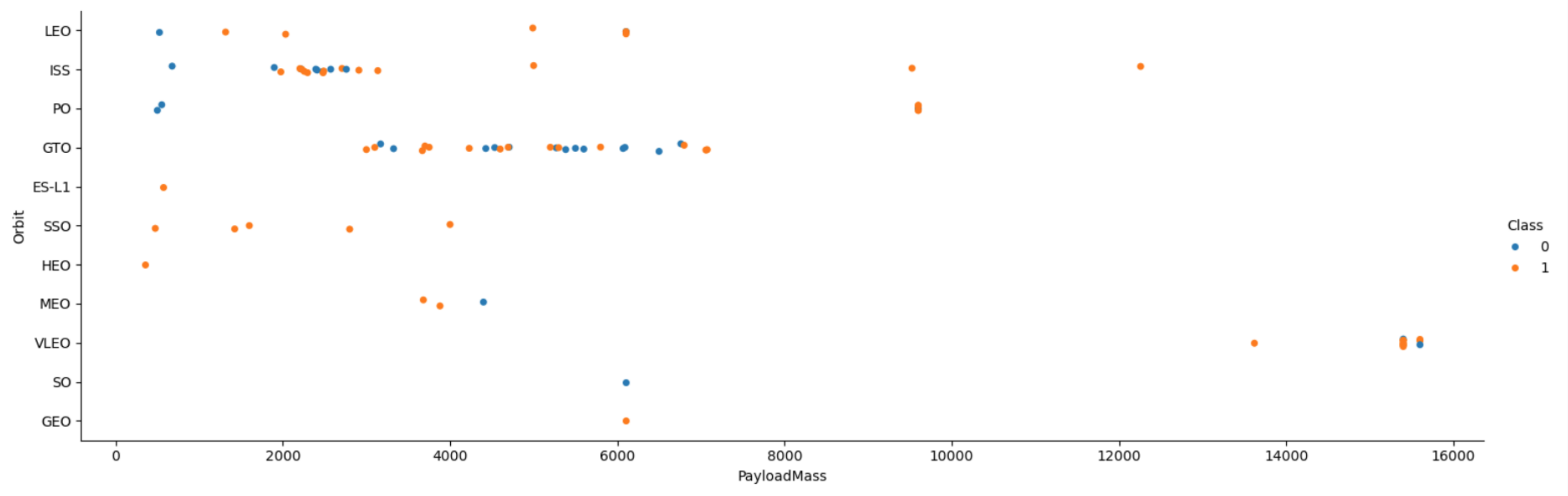
- As we can see in the graph, ES-L1, GEO, HEO, SSO orbits have higher success rate followed by VLEO.

Flight Number vs. Orbit Type

- Most recent launches can be seen at VLEO.
- Also, It is quite evident that the success rate has improved over time.

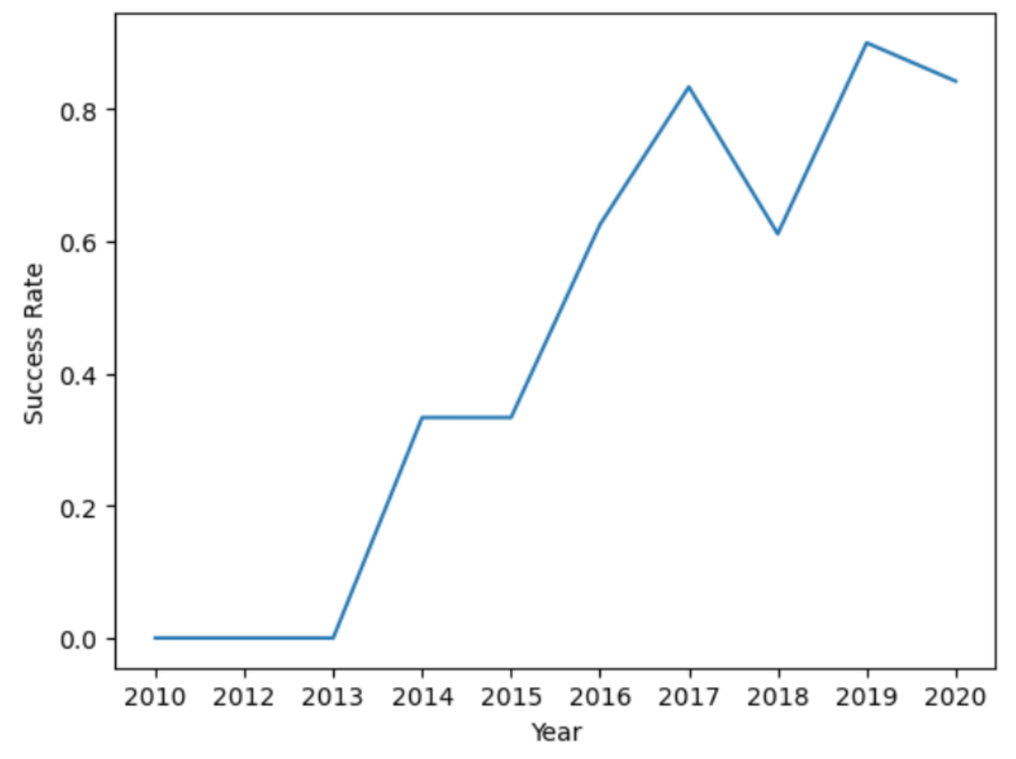


Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- We cannot distinguish the same for orbit GTO as there are both positive and negative landings with the payload ranging from 3000 to 7500.
- With the increased payload mass, the rate of success increases.

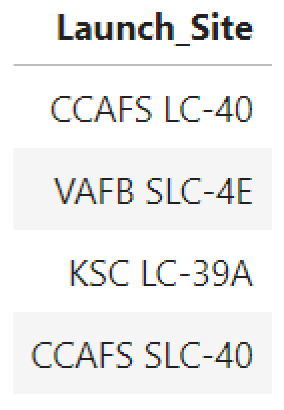
Launch Success Yearly Trend



- As we can see, the success rate since 2013 kept increasing till 2020.
- Success rate was null for the first three years.

All Launch Site Names

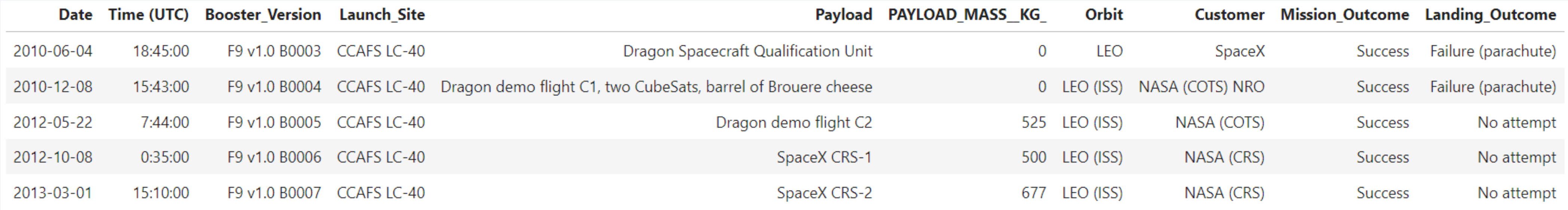
- Here are the names of launch sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- Distinct method was used.



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

Launch Site Names Begin with 'CCA'

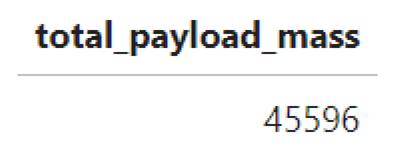
- Given below are the 5 samples of launch sites where name begins with "CCA"
- Here, like statement was used to find the launch sites where name starts with "CCA%"



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

Total Payload Mass

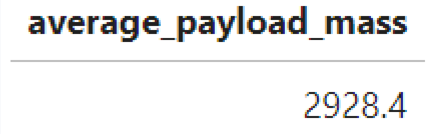
- The total Payload Mass carried by boosters from NASA is 45596.
- Here, the sum method is used to calculate the total payload mass.



total_payload_mass
45596

Average Payload Mass by F9 v1.1

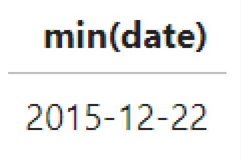
- The average payload mass carried by booster version F9 v1.1 is 2928.4.
- Here, avg method was used to calculate the average payload mass.



average_payload_mass
2928.4

First Successful Ground Landing Date

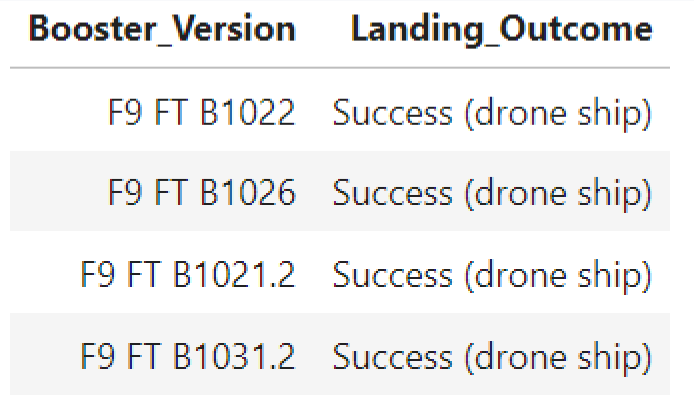
- The date of the first successful landing outcome is 22nd Dec, 2015.
- Here, min method was used with the condition.



min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

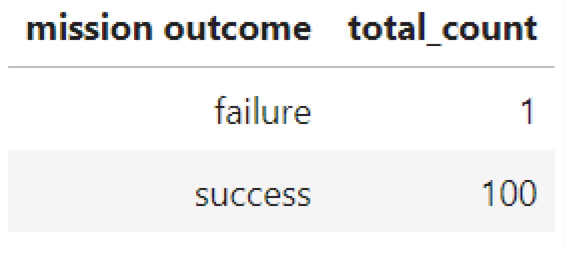
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 can be seen in the table.



Booster_Version	Landing_Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

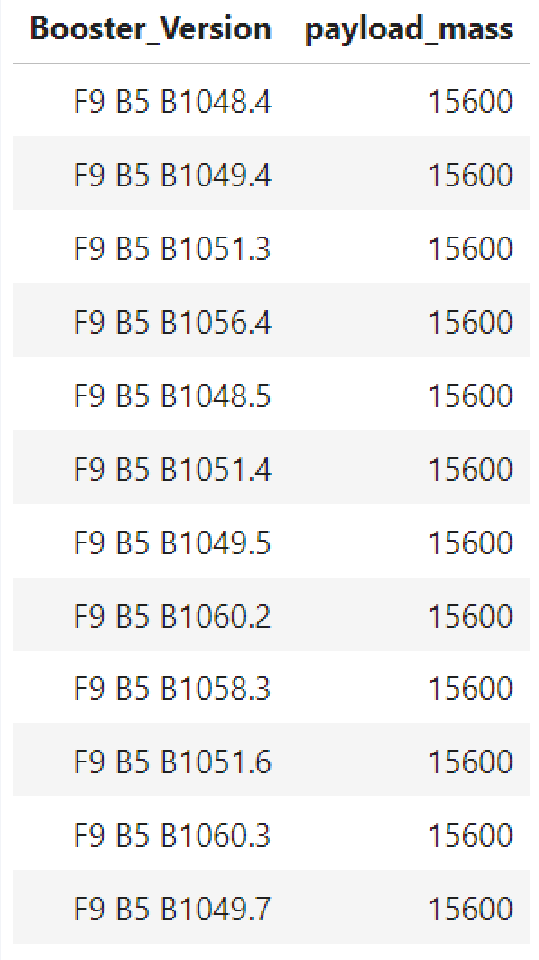
- Total number of successful and failure mission outcomes can be seen in the table given below.
- Here, mission outcomes containing "success" and "failure" were grouped separately using the case method and count method was applied afterwards to achieve this summary.



mission outcome	total_count
failure	1
success	100

Boosters Carried Maximum Payload

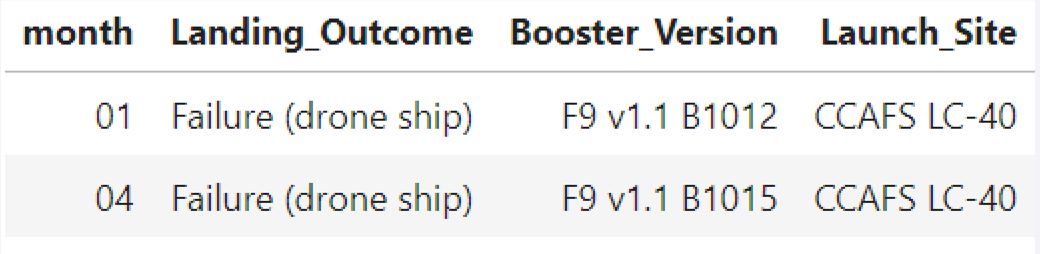
- The names of the booster which have carried the maximum payload mass can be seen in the figure of the list attached.
- Subquery was used to achieve this summary.



Booster_Version	payload_mass
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

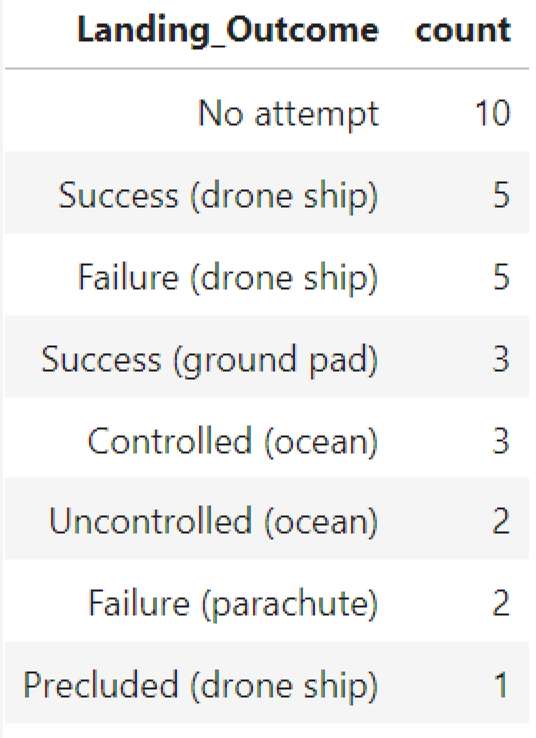
- substr method was used to filter results based on the year.



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

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

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order can be seen in the figure of list attached.



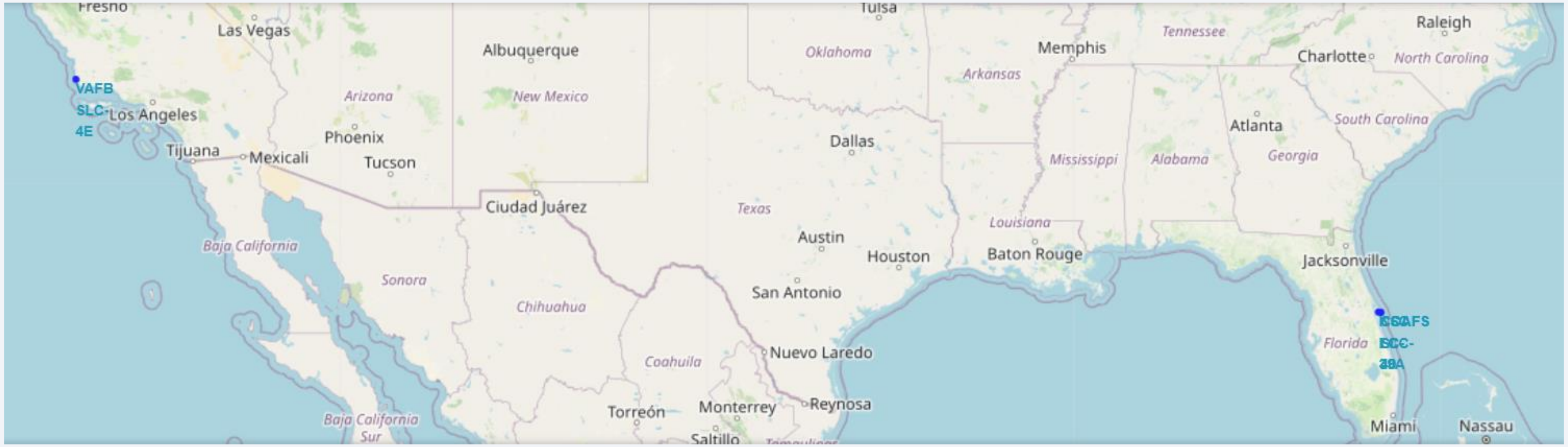
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

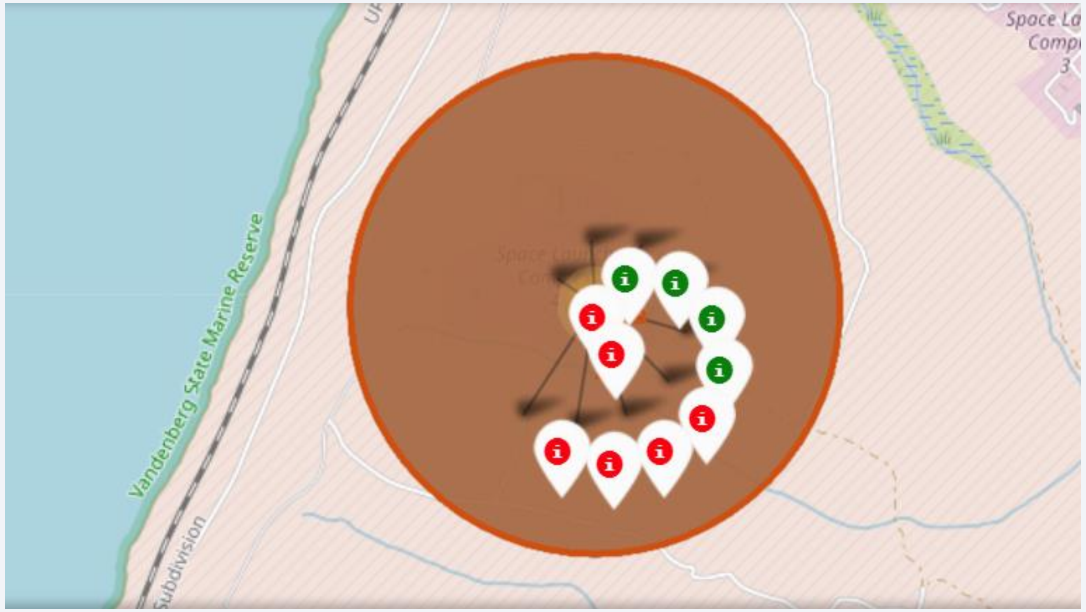
All Launch Sites

- All the launch sites are near coast line as we can determine from the map.



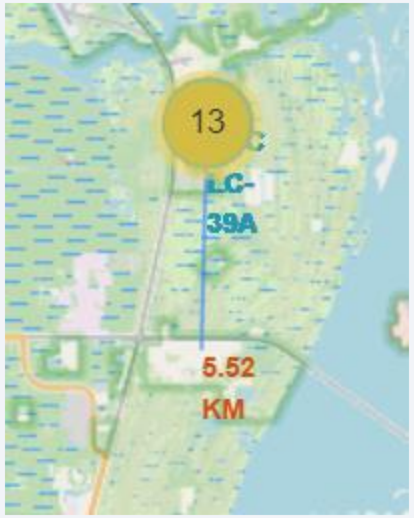
Successful/Failed Launches for each site

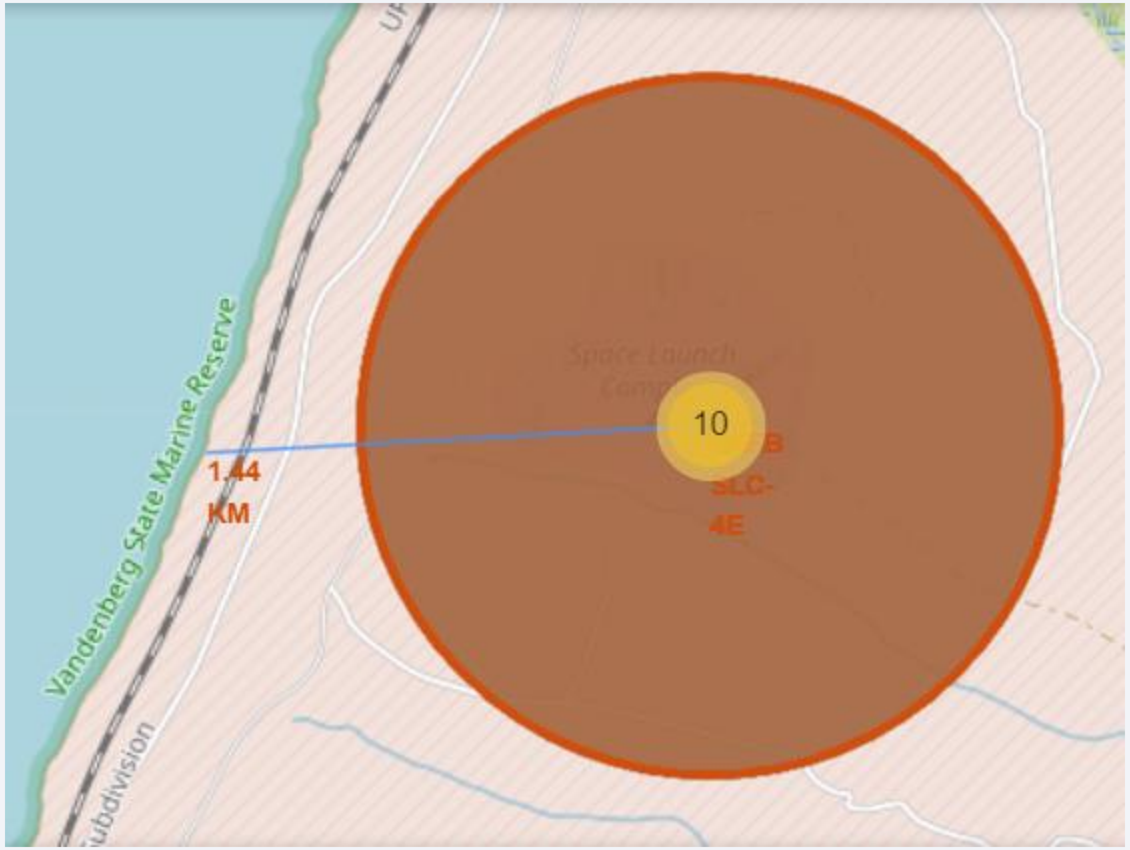
- Green markings indicate the successful launches and the red ones indicate the failed ones.



Distances between a launch site to its proximities

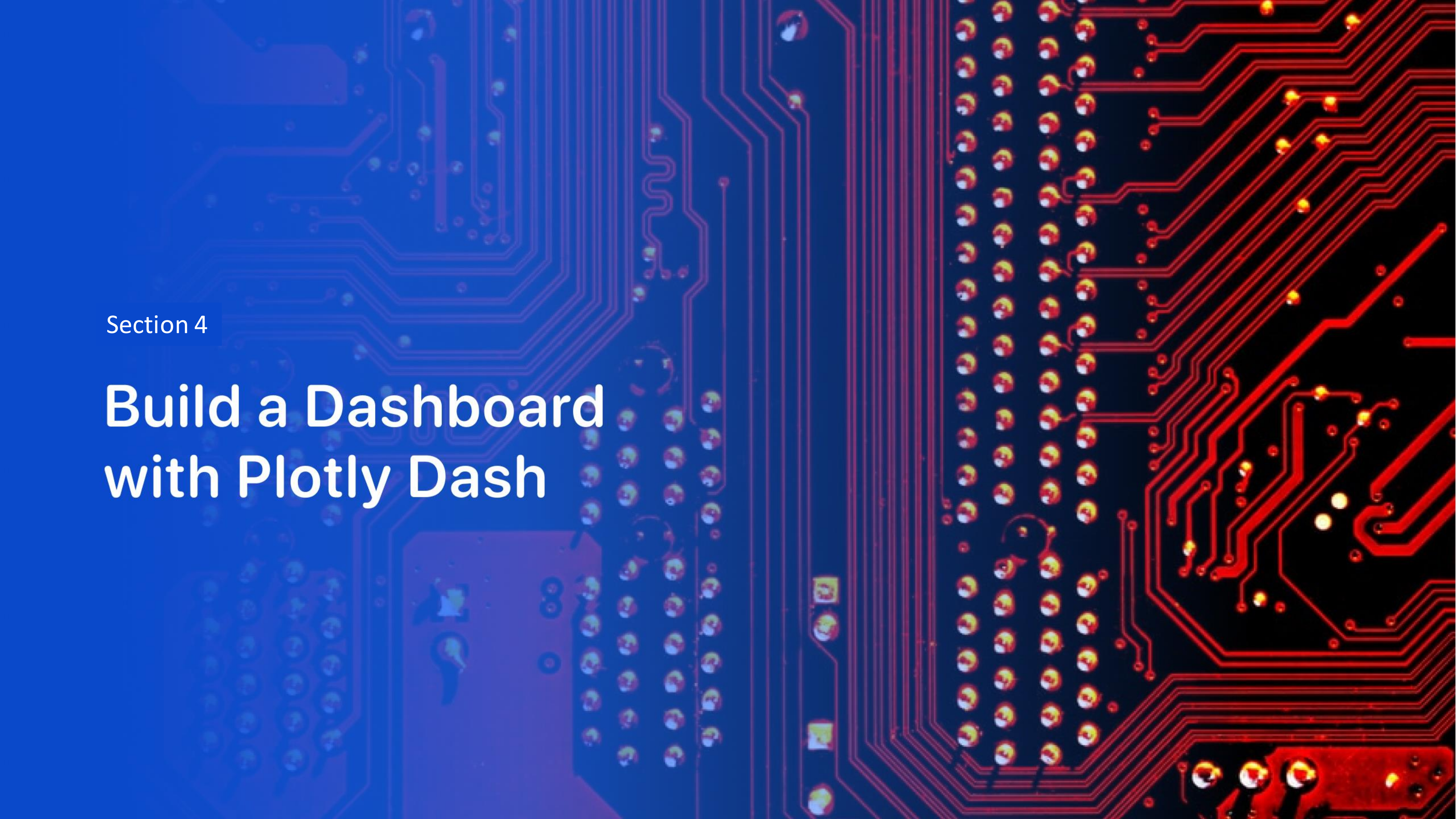
- As we can see, the distance between launch sites and their proximities is described in the graph.





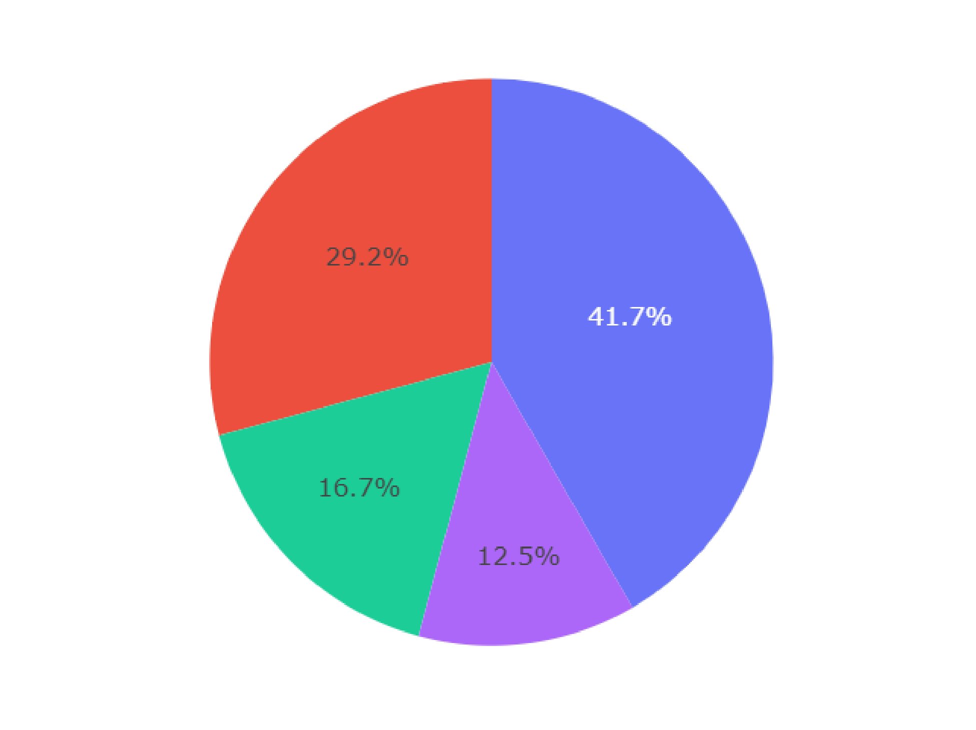
Section 4

Build a Dashboard with Plotly Dash



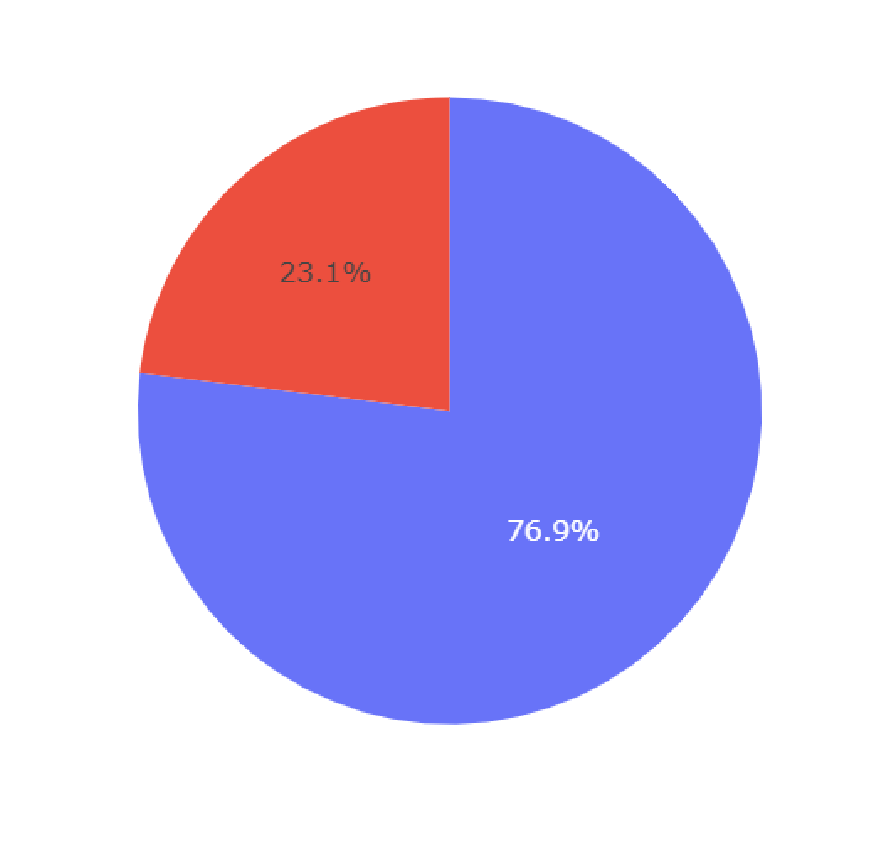
Total Successful Launches by Site

- KSC LC-39A is on the top when it comes to total successful launches. Color blue represents the same.



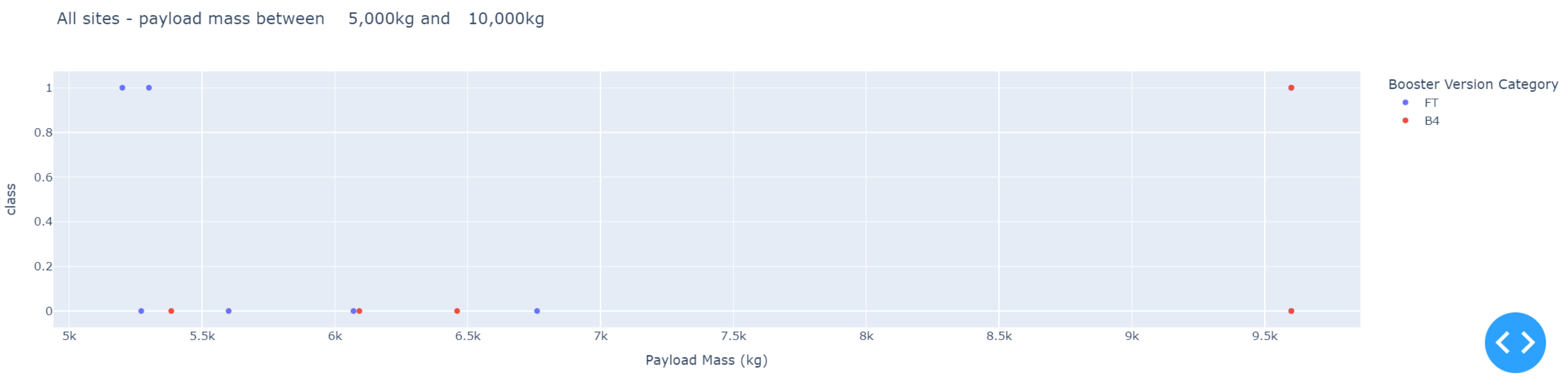
Launch site with the highest success ratio

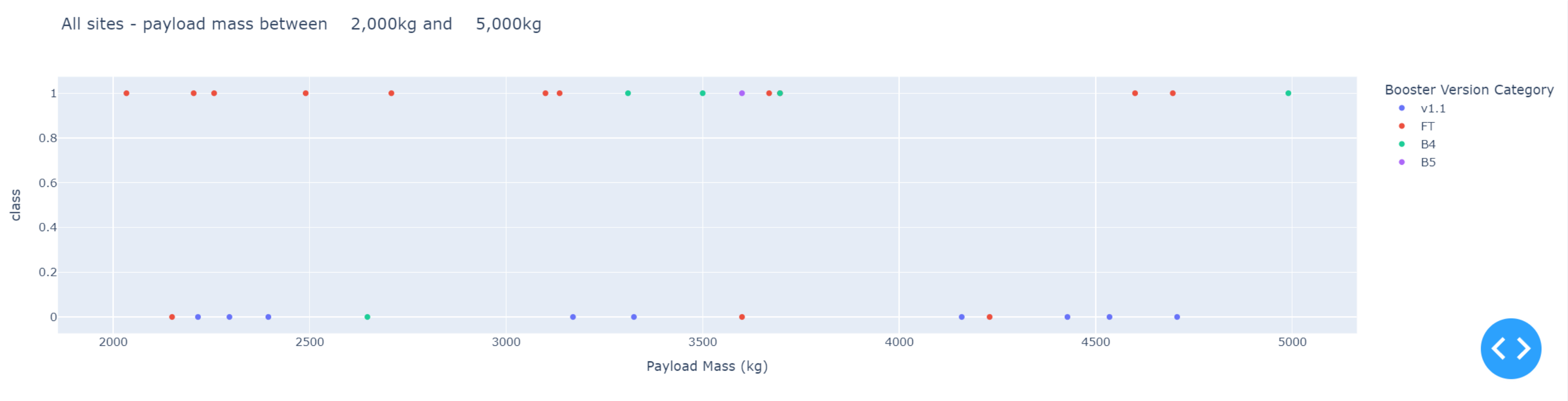
- Attached is the pie chart of the success ratio for KSC LC-39A.
- Here, we can see that the 76.9% of the launches are successful.



Scatter Plot: Payload and Outcome

- Here are two scatter plots displaying results for the payload mass more than 5,000 and the payload mass between 2000 and 5000.





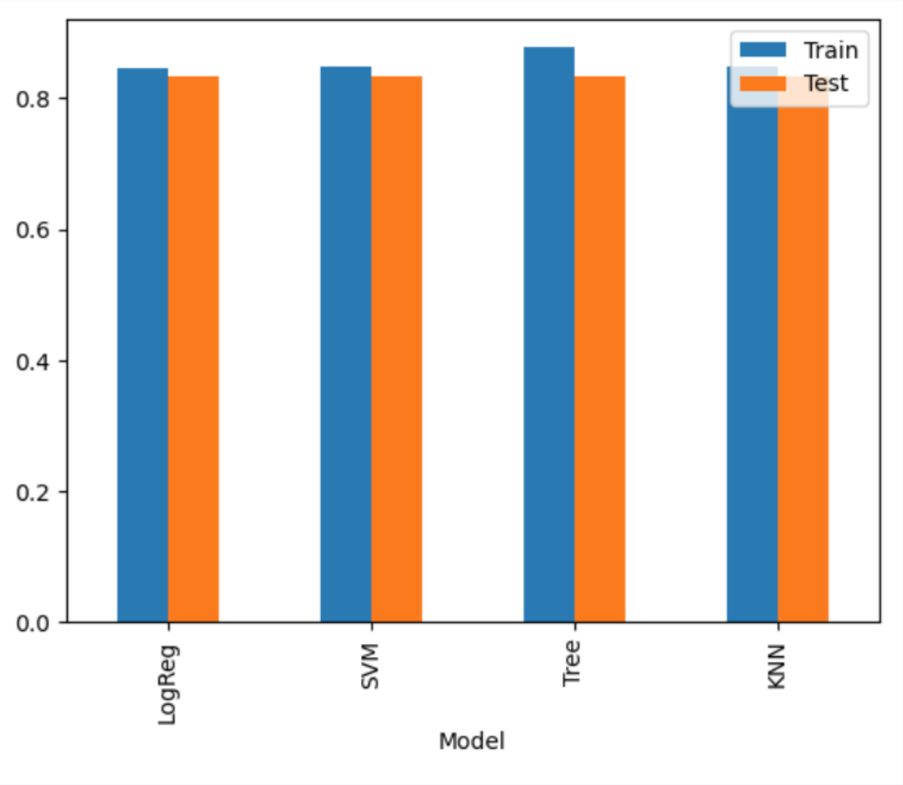
- For payload mass more than 5000, only FT and B4 booster versions can be noticed.

Section 5

Predictive Analysis (Classification)

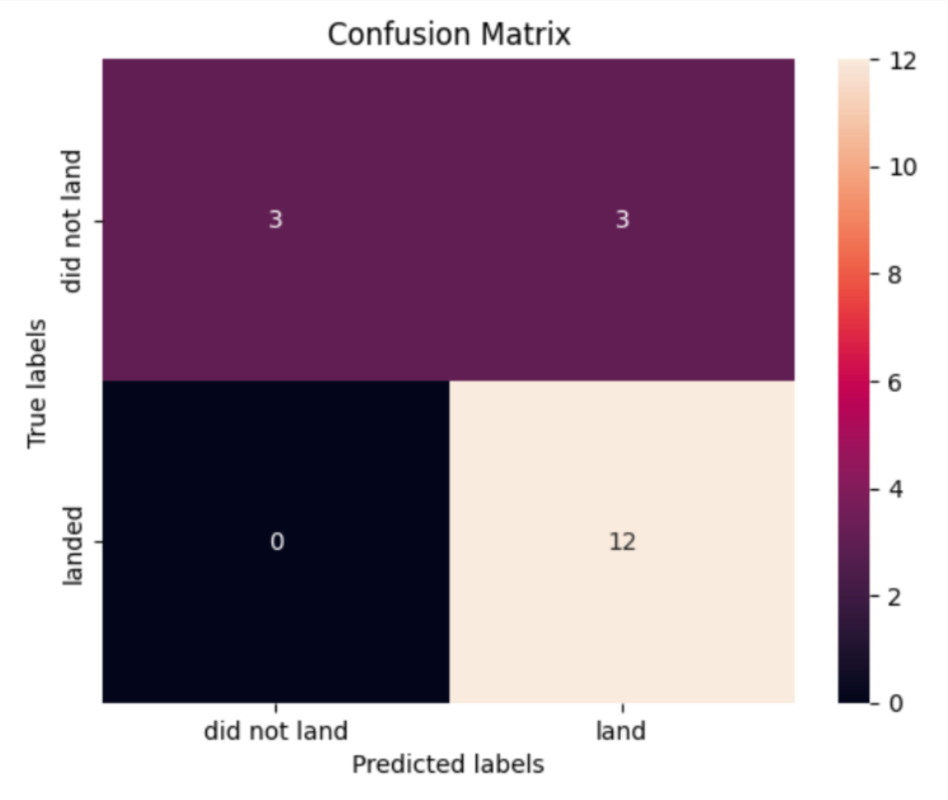
Classification Accuracy

- From the bar plot, it is quite evident that the Decision Tree as best overall accuracy comparted to other models.



Confusion Matrix

- Here's the confusion matrix of decision tree classifier.
- As we can see, there are only 3 False Positives. Also, there are 0 False Negatives. This indicates that this model performs really well.

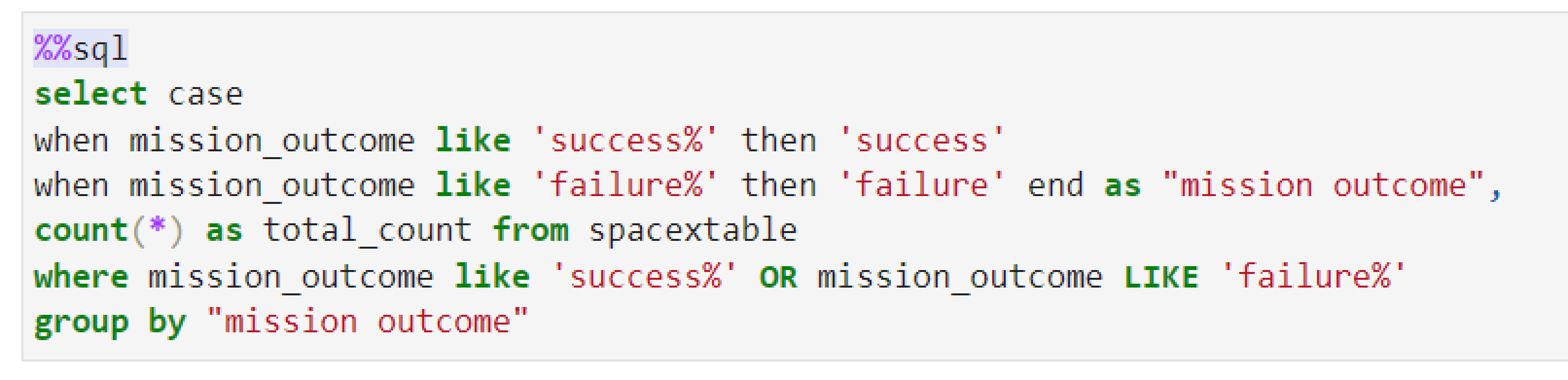


Conclusions

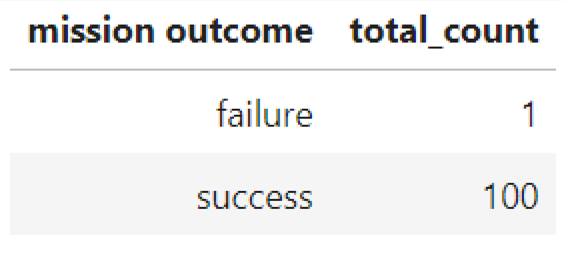
- ES-L1, GEO, HEO, SSO, and VLEO are the Orbits that have high success rate. The SO has the least success rate amongst the orbits.
- The success rate since 2013 kept increasing till 2020. It shows that the landing outcome became successful and better with year.
- Date when the first successful landing outcome in ground pad was achieved: 2015-12-22. It is five years after the first launch.
- The most recent launches are happening at CCAF5 SLC 40. Also, it is quite evident that the most recent launches are successful.
- The launches with excessive payload over 12,000 are made from CCAFS SLC 40 and KSC LC 39A. Also, landings seems to have higher success rate when the payloads are over 8000.

Appendix

- I used the following SQL code in order to summarize total successful and failed outcomes.



```
%%sql
select case
when mission_outcome like 'success%' then 'success'
when mission_outcome like 'failure%' then 'failure' end as "mission outcome",
count(*) as total_count from spacextable
where mission_outcome like 'success%' OR mission_outcome LIKE 'failure%'
group by "mission outcome"
```



mission outcome	total_count
failure	1
success	100

Thank you!