IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nadira Fawziyya Masnur
19/02/2024

Outline

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

Executive Summary

- Methodologies using **CRISP-DM**
  - Data Understanding
  - Data Preparation
    - **Data Collection** through API and Webscrapping
    - **Data Wrangling**
    - **Explanatory Analysis** with SQL
    - **Explanatory Analysis** with Pandas and Matplotlib
    - **Interactive Visual Analytic** with Folium
  - Modelling (Classification)
  - Evaluation
- Result
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

Project Background and Context

The commercial space industry has experienced remarkable expansion, driven by the efforts of companies such as SpaceX, which have revolutionized space travel, making it more accessible and cost-effective. A pivotal element in SpaceX's triumph lies in its capacity to recycle the first stage of its Falcon 9 rockets, thereby markedly diminishing launch expenses compared to other providers. This project aims to develop a machine learning model (logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbour) tasked with forecasting the outcome of rocket landings, distinguishing between successful and failed attempts, thus offering valuable insights into the dynamics of space exploration.

Problem Identification

- What factors determine whether the rocket will land successfully?
- How does each feature influence the success rate of rocket landing?

Section 1

Methodology

Methodology

Executive Summary (1/2)

- Data collection methodology:
  - Data collected from:
    - Spacexdata https://api.spacexdata.com/v4/ (through **API**) specifically extract Launch Site (Launchpad), Booster Version (Rocket), Payload (Payloads), and Core data.
    - Wikipedia “List of Falcon 9 and Falcon Heavy launches” (through **WebScrapping** using BeautifoulSoup)
- Perform data wrangling
  - Data wrangling performed with create “class” column that contains whether rocket is successfully landed (success=1, fail/bad outcome=0). This class helps to calculate the success rate of rocket landing.

Methodology

Executive Summary (2/2)

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was normalized using standardScaler
  - The classification Model used is Logistic Regression, SVM, decision Tree, and KNN. The parameters selected through gridSearchCV to get the best parameters.
  - The model trained with X_train and Y_train data resulted from splitting the data into 80% data training and 20% data testing.
  - The evaluation performed with score (accuracy score, best score) and confusion matrix.

Data Collection

The data used is extracted through 2 source. Each source was handled with different method, there is using API and webscrapping.

- Spacexapi data (https://api.spacexdata.com/v4/) extractedthrough API. The result data will be look like this

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitud
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.04772

- Wikipedia “List of Falcon 9 and Falcon Heavy launches” (through **WebScrapping** using BeautifoulSoup). The result data will be look like this

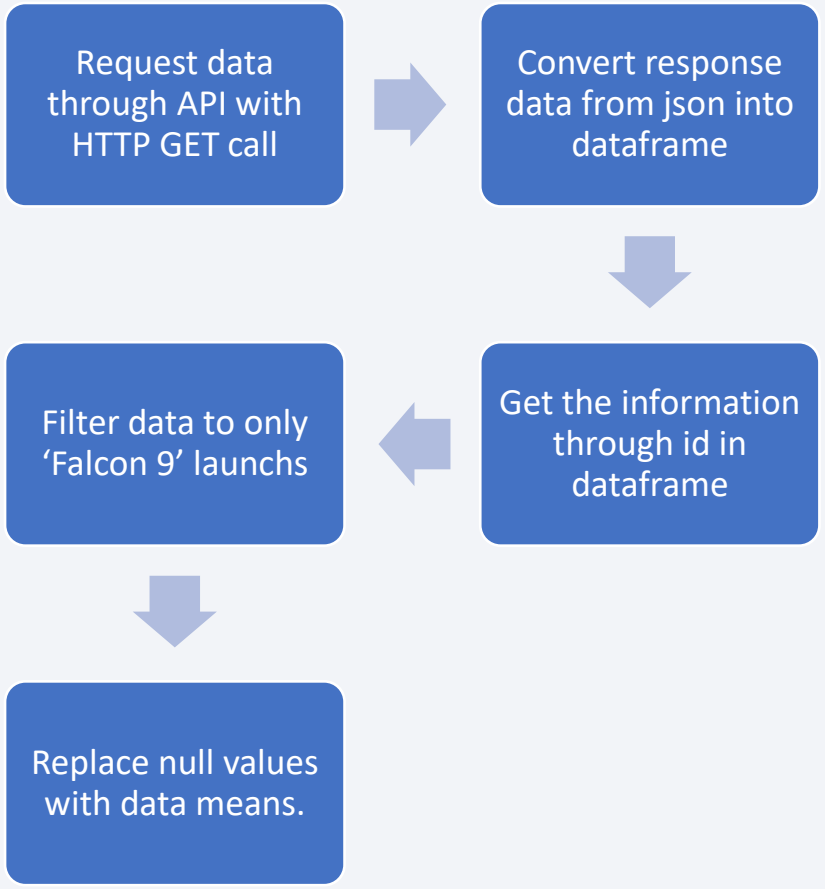
	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45

Data Collection – SpaceX API

This process include collecting the dataframe from api.spacexdata.com through HTTP REST call and data preprocessing as presented in graph beside.

This process could be accessed through this GitHub URL:

https://github.com/458-nadiraF/IBMDataScienceFinalAssignment/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

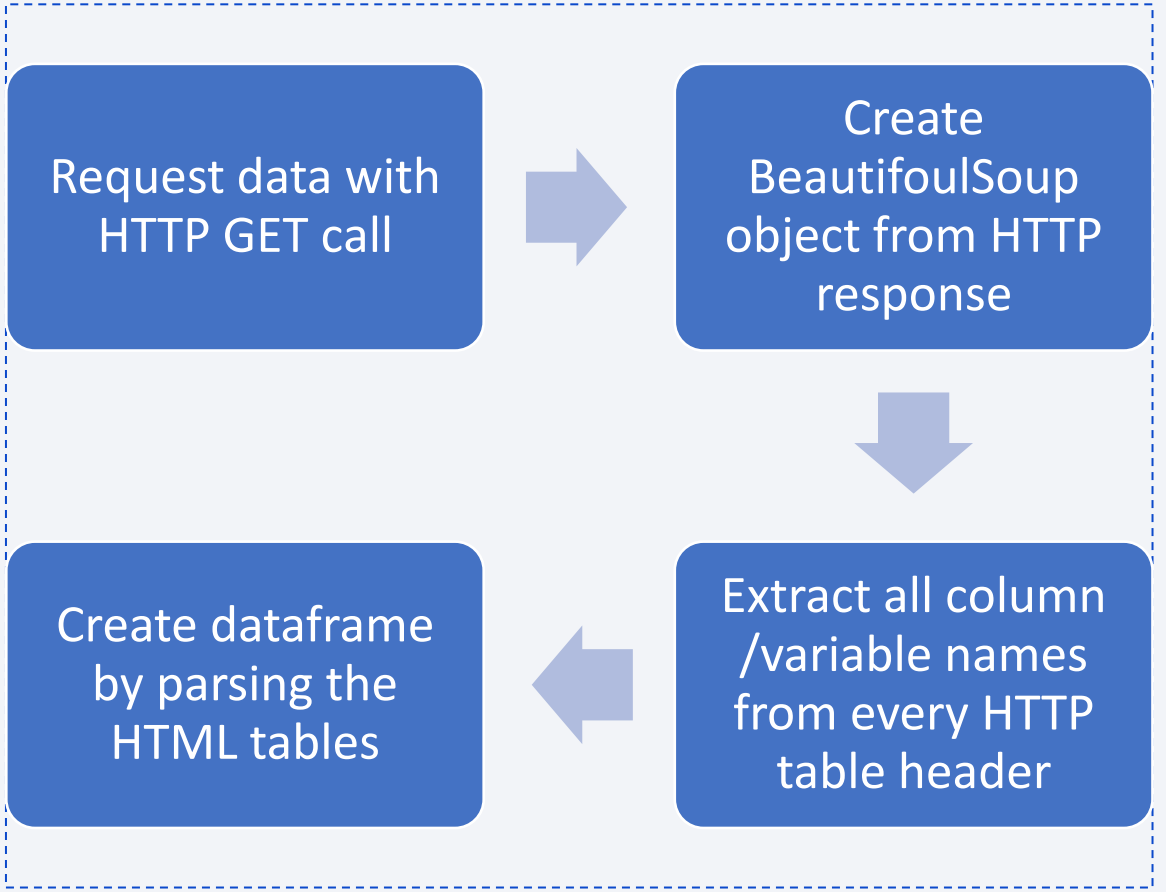


Data Collection - Scraping

This process include collecting the dataframe from Wikipedia “List of Falcon 9 and Falcon Heavy launches” with detailed step presented in graph beside.

This process could be accessed through this GitHub URL:

https://github.com/458-nadiraF/IBMDataScienceFinalAssignment/blob/main/jupyter-labs-webscraping.ipynb

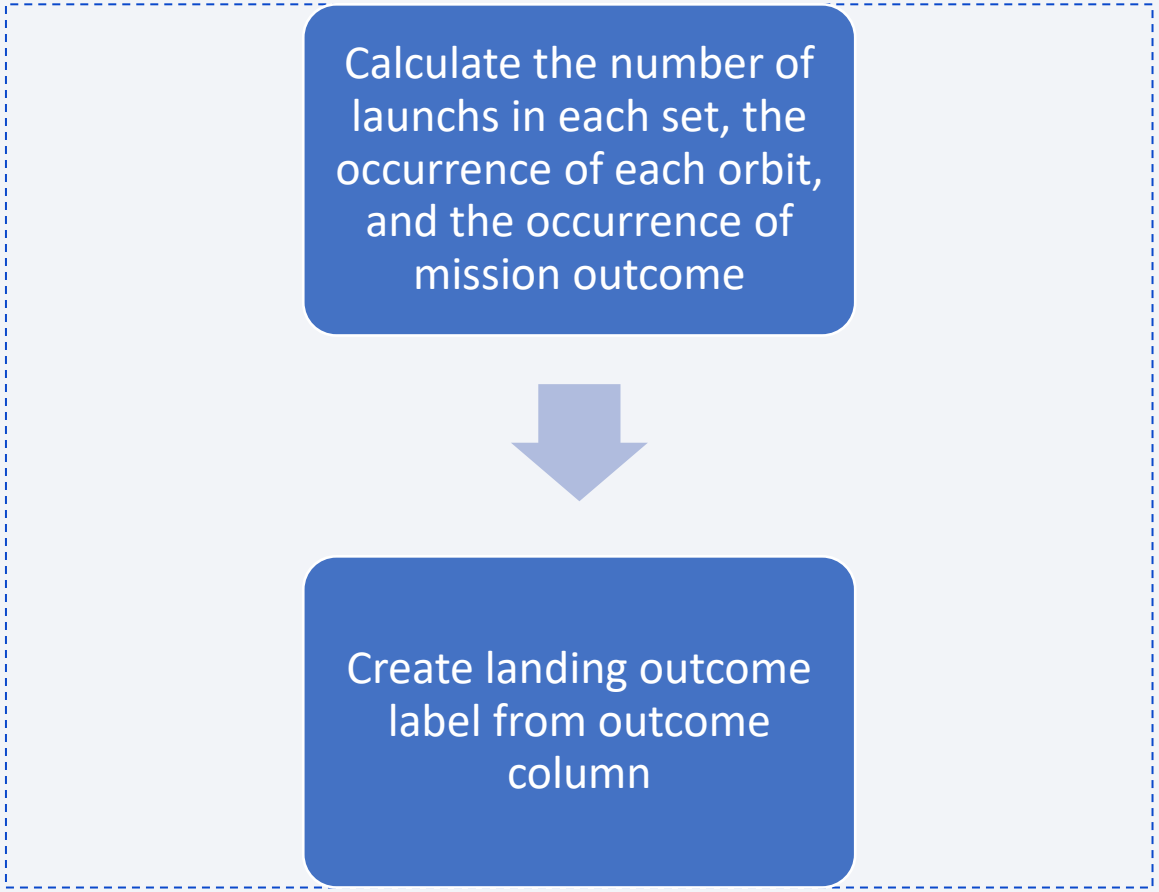


Data Wrangling

This process initiate Explanatory Data Analysis to find some pattern and determine the label for training supervised model. The detailed process shown in graph beside.

This process could be accessed through this GitHub URL:

https://github.com/458-nadiraF/IBMDataScienceFinalAssignment/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

This process performed to see how some features influence to other features or to the success rate. This performance include producing chart to see how this feature interact with each other.

- FlightNumber vs PayloadMass
- FlightNumber vs LaunchSite
- PayloadMass vs LaunchSite
- Orbit vs Class (Success Rate)
- FlightNumber vs Orbit
- PayloadMass vs Orbit
- Year vs Class (Success Rate trends over the year)

This process also include feature engineering (select what features used as data training), applying OneHotEncoder using pd.get_dummies to the column Orbit, LaunchSite, LaunchPad, and Serial, and also include convert type of numeric column to ‘float64’

This process could be accessed through this GitHub URL:

https://github.com/458-nadiraF/IBMDataScienceFinalAssignment/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

This process performed execution of SQL queries as the following:

- 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.

This process could be accessed through this GitHub URL:

https://github.com/458-nadiraF/IBMDataScienceFinalAssignment/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

This process contains following step

- Mark all launch sites on a map

  This step performed by add folium.Circle and folium.map.Marker on each LaunchSite coordinate (Lat, Long). This step make it easier to see the position of the Launch Site.

- Mark the success/failed launches for each site on the map

  This step performed by add folium.Marker (green color for the success launches and red for the failed one) on every records and collect it with marker_cluster. This step make it easier to see the success rate launchs on each Launch Site.

- Calculate the distances between a launch site to its proximities

  This step performed by add folium.PolyLine from LaunchSite to the coordinate of coastline, railway, highway, and city. This step make it easier to see if the success rate launchs on each Launch Site influenced with the proximity of launch site to another object

This process could be accessed through this GitHub URL:

https://github.com/458-nadiraF/IBMDataScienceFinalAssignment/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

This process include building an interactive dashboard with Dash

There is 2 input, Launch Site (or all) and payload mass. With this input, dashboard will shows pie chart of success rate of each launch site and the scatter plot of payload mass vs class and the color of each record will be based on Booster Version Category.

This process could be accessed through this GitHub URL:

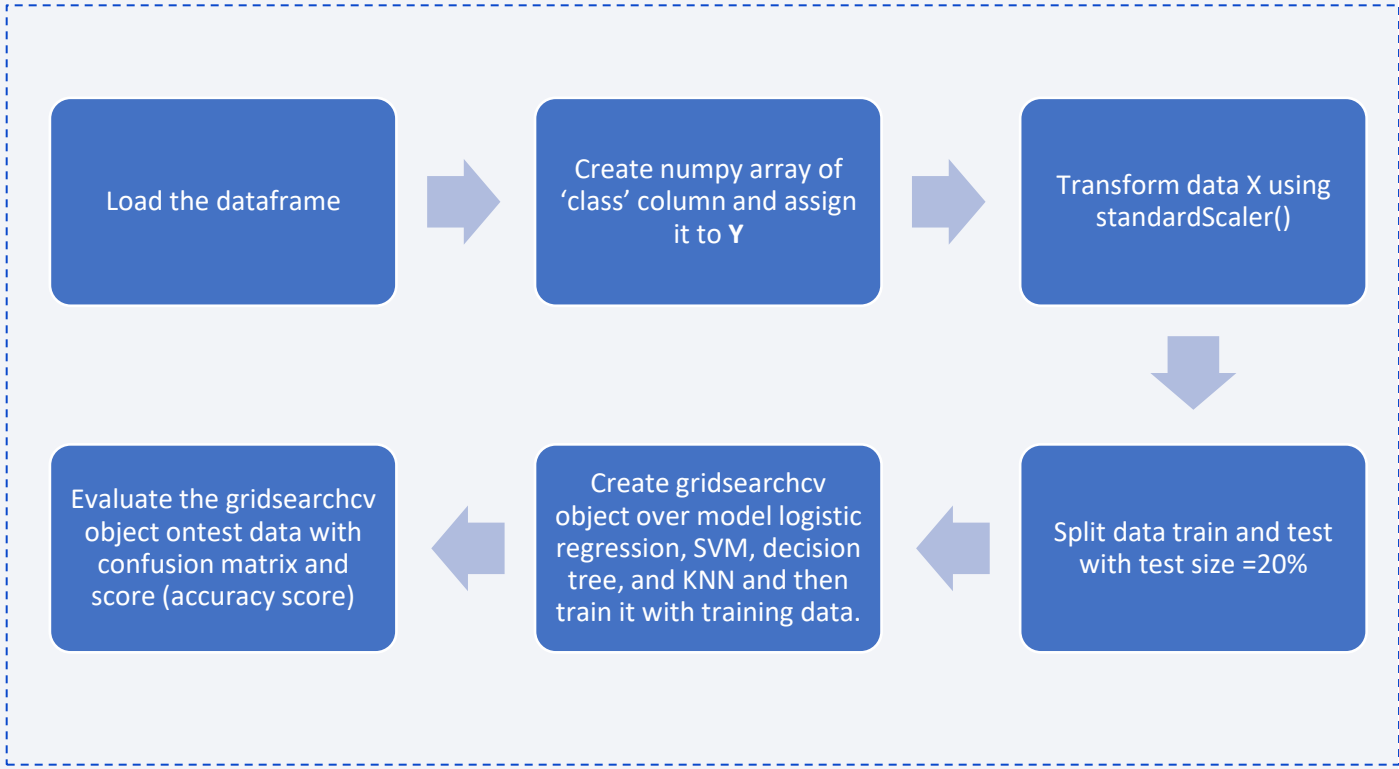
https://github.com/458-nadiraF/IBMDataScienceFinalAssignment/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

This process include **select** the data (features/X and target/Y), **standardize** the data, split the data (training/testing), building the **model**, build **Gridsearchcv** object, **fitting** the gridsearchcv object over training data, and then **evaluate** model with test data.

This process could be accessed through this GitHub URL:

https://github.com/458-nadiraF/IBMDataScienceFinalAssignment/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



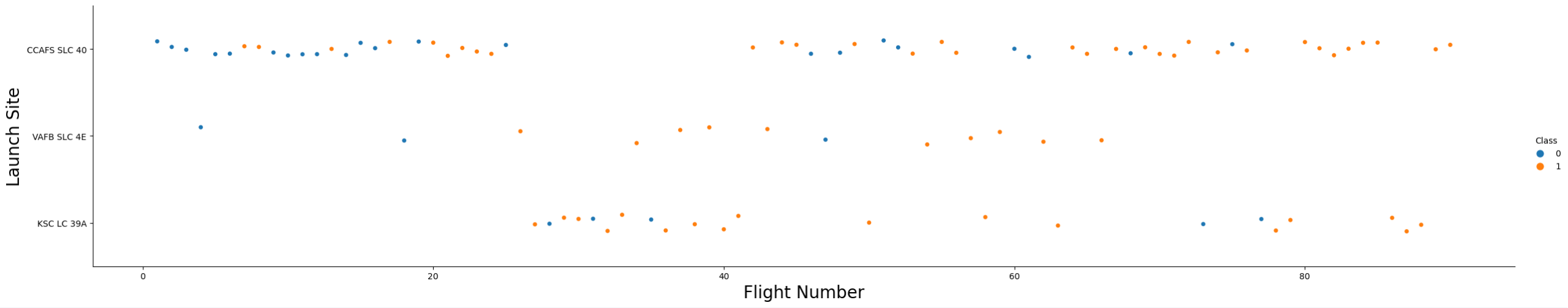
Results

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

Section 2

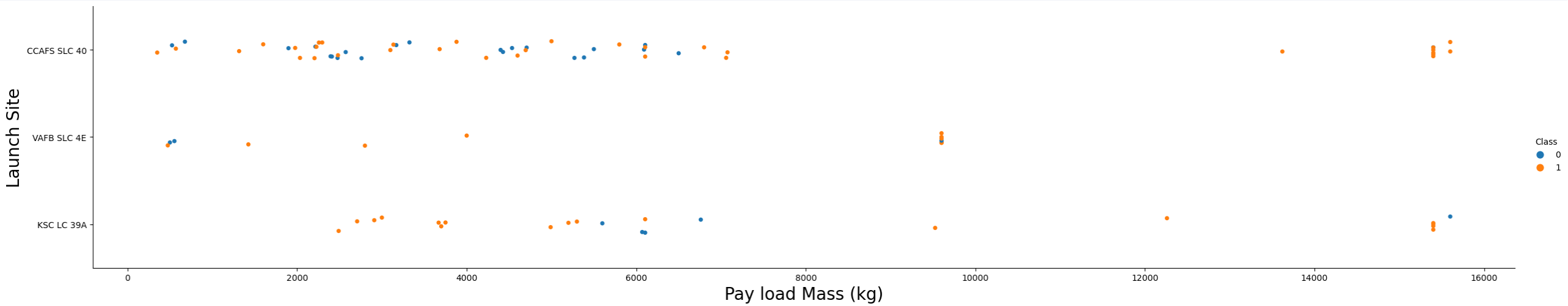
Insights drawn from EDA

Flight Number vs. Launch Site



From the graph we can see that the greater flight number, the greater the success rate of landing.

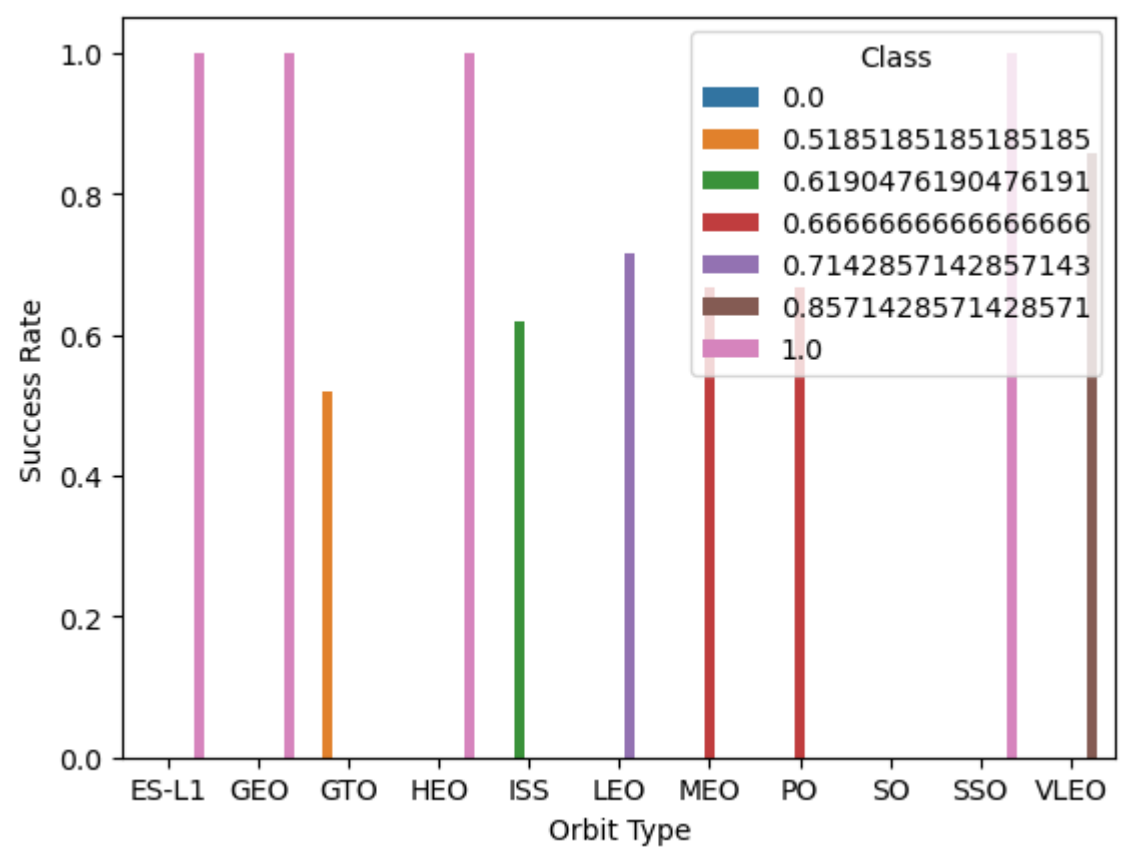
Payload vs. Launch Site



- From the plot, we can see that the greater payload mass, the greater the success rate of landing in launch site ‘CCAFS SLC 40’. For the rest launch site, the payload mass doesn’t seem influence the success rate.

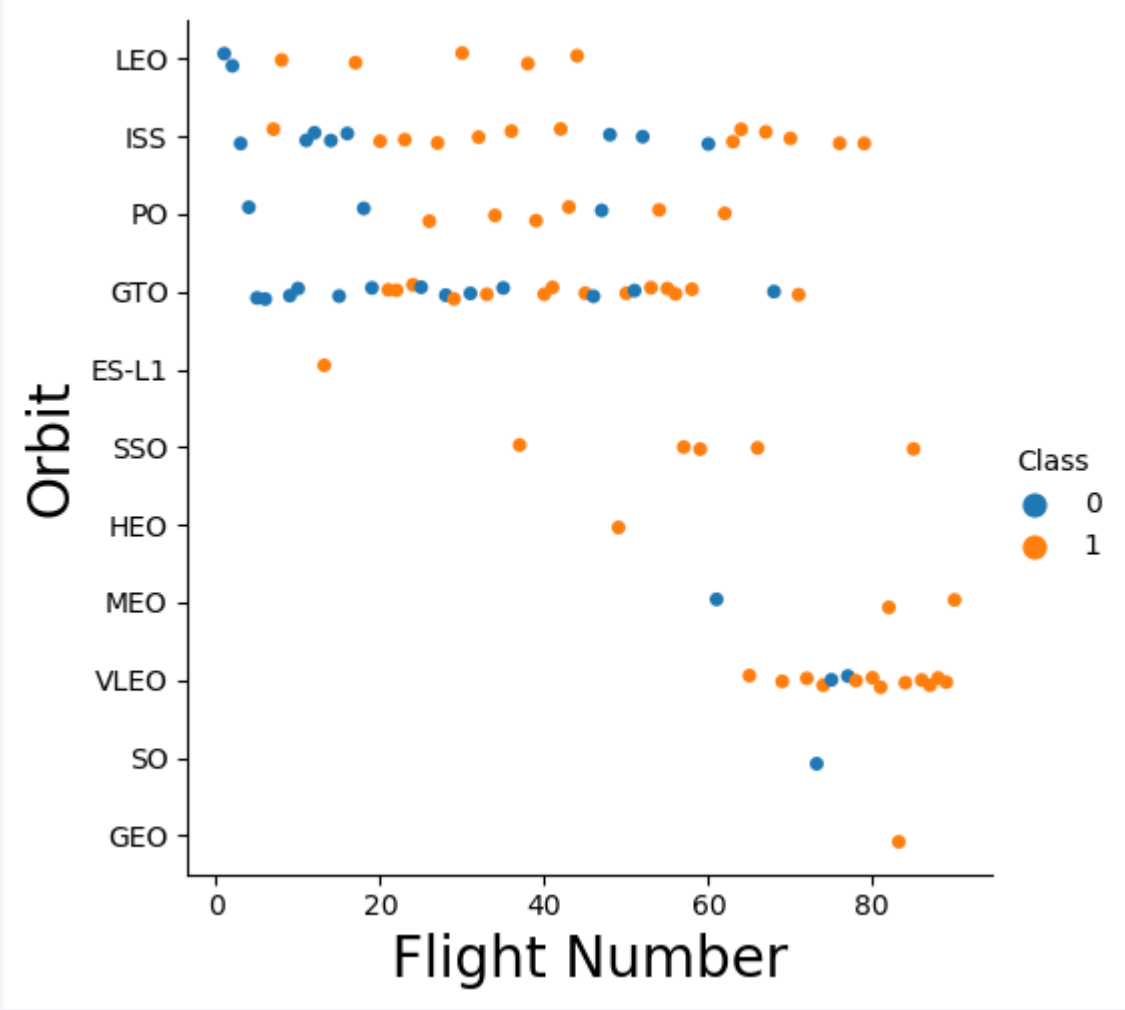
Success Rate vs. Orbit Type

From the chart, we can see that there is 4 orbit type that have success rate 1.0, there is: ES-L1, GEO, HEO, and SSO.

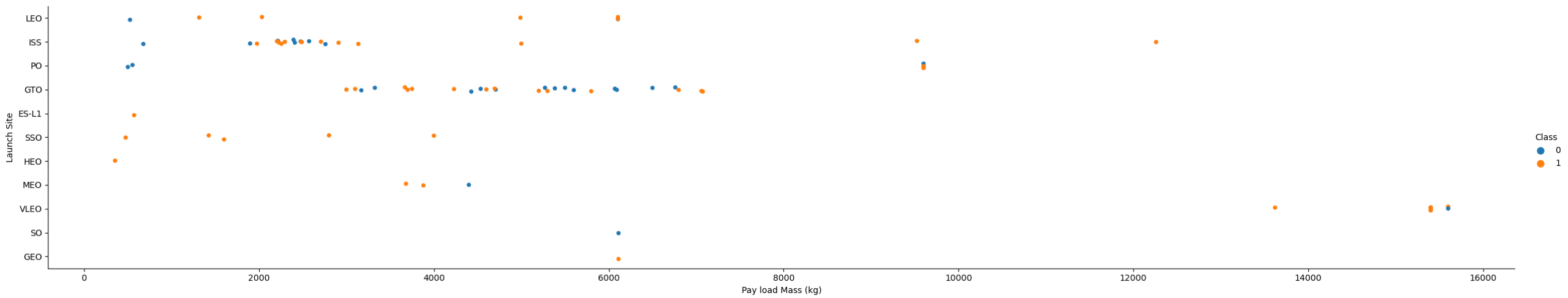


Flight Number vs. Orbit Type

From the plot, we can see that the greater flight number, the greater the success rate of landing on orbit type ‘LEO’ and ‘MEO’. The rest of orbit type success rate doesn’t seem influenced by flight number



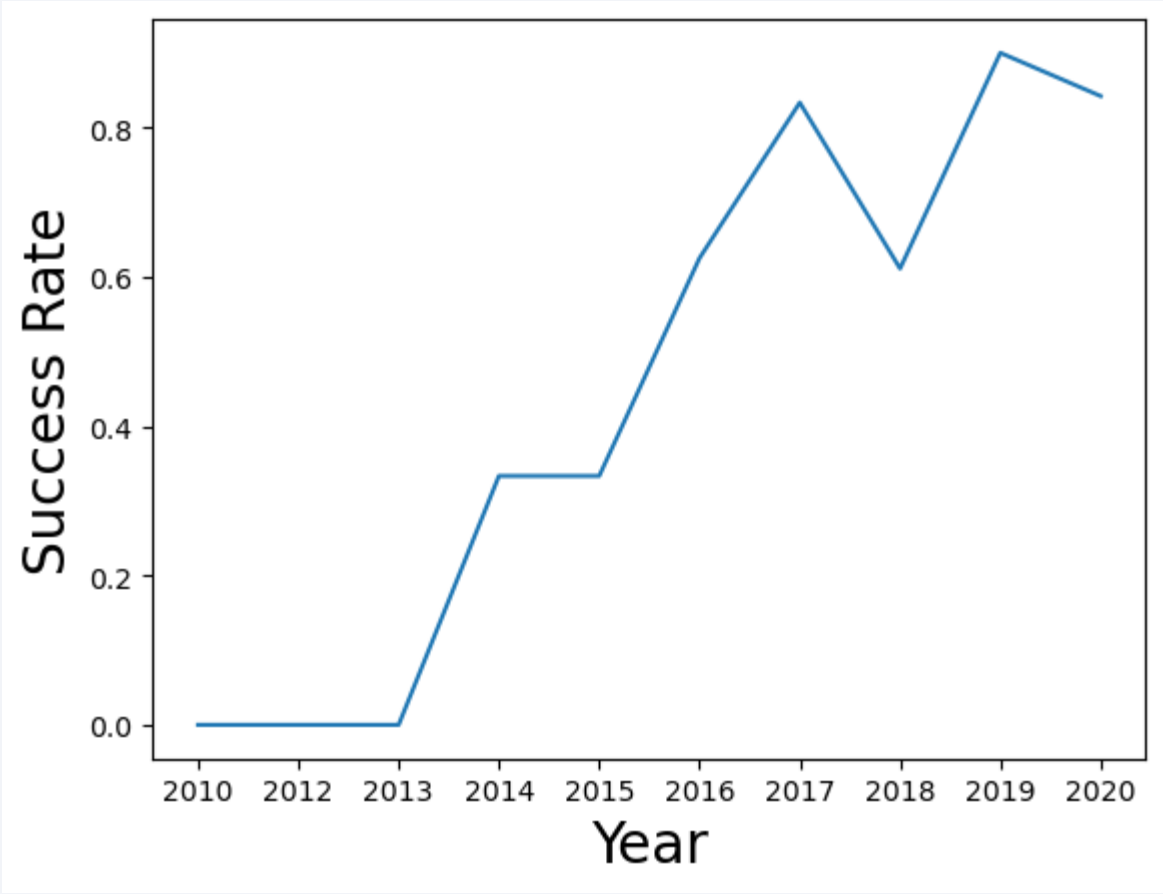
Payload vs. Orbit Type



From the plot, we can see that the greater flight number, the greater the success rate of landing on orbit type ‘LEO’ and ‘ISS’. The rest of orbit type success rate doesn’t seem influenced by flight number

Launch Success Yearly Trend

From the chart, we can see the success rate have positive trend line from 2010 until 2017 and decrease in 2018.



All Launch Site Names

This query using ‘DISTINCT’ to return unique value of ‘Launch_Site’ column

```
%sql select distinct "Launch_Site" FROM SPACEXTABLE
```

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

Launch Site Names Begin with 'CCA'

This query using LIKE ‘CCA%’ phrase to return any record with Launch Site name started with ‘CCA’ and LIMIT 5 to just return 5 records.

```
%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

Total Payload Mass

This query using SUM() to return the total ‘PAYLOAD_MASS__KG_” column and where “Customer” = “NASA (CRS)” to filter the records.

```
%sql select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer" ="NASA (CRS)"
```

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

This query using AVG() to return the average ‘PAYLOAD_MASS__KG_” column and where “Booster_Version” LIKE “F9 v1.1%” to filter the records that starts with F9 v1.1.

```
%sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" like 'F9 v1.1%'
```

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

This query using **min(date)** to return the minimum of date column and **where** “Landing_Outcome”= “Success (ground pad)” to filter the records

```
%sql select min(date) from SPACEXTABLE where "Landing_Outcome"="Success (ground pad)"
```

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

This query using **SELECT** “Booster_Version” to return the ‘Booster Version” record, **WHERE** “Landing_Outcome”=“Success (drone ship)” to filter the records, and **BETWEEN** clause to filter the “PAYLOAD_MASS__KG_” value to 4000-6000.

```
%sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome"="Success (drone ship)" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

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

Total Number of Successful and Failure Mission Outcomes

This query using **COUNT()** clause to count the records with the same value.

```
%sql select "Mission_Outcome", count("Mission_Outcome") from SPACEXTABLE
GROUP BY "Mission_Outcome"
```

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

Boosters Carried Maximum Payload

This query using subquery to get the max of payload mass and use it to filter the first select clause

```
%sql select "Booster_Version" FROM SPACEXTABLE where
"PAYLOAD_MASS__KG_"=(select MAX("PAYLOAD_MASS__KG_") FROM
SPACEXTABLE)
```

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

This query use **SUBSTRING()** to change the month number extracted with **substr()** and then also filter the records to just in 2015 using **substr**.

```
%sql select SUBSTRING('JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC ',
(substr(Date,6,2)*4)-3,3) AS Month_Name, Landing_Outcome, Booster_Version,
Launch_Site from SPACEXTABLE where "Landing_Outcome"="Failure (drone ship)" AND
substr(Date,0,5)='2015'
```

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
JAN	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APR	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

This query using **BETWEEN** clause to filter Date just in between 2010-06-04 and 2017-03-20, **GROUP BY** Landing_Outcome and **ORDER BY VAL** to ordering the result based on VAL or **COUNT(**Landing_Outcome**).**

```
%sql select Landing_Outcome, count(Landing_Outcome) AS
VAL from SPACEXTABLE WHERE Date Between '2010-06-04'
and '2017-03-20' group by Landing_Outcome Order by VAL
DESC
```

Landing_Outcome	VAL
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

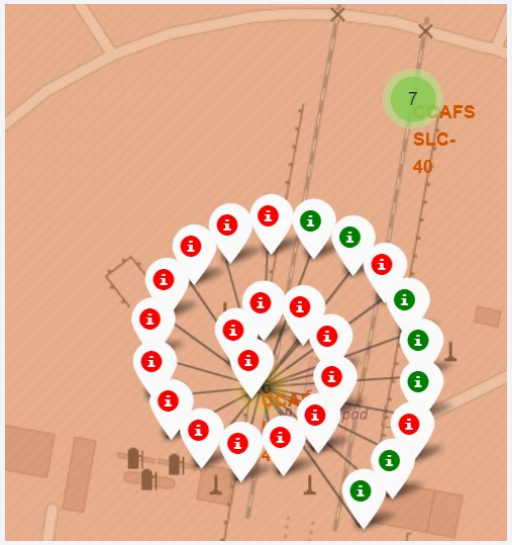
Launch Sites Proximities Analysis

Mark the LaunchSite with Circle

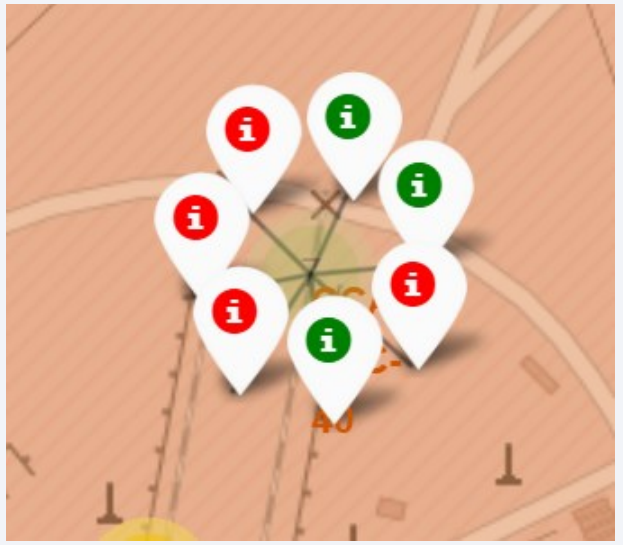
From the map, we can see that the launch site in the data basically just in 2, in Florida and California

Mark the success/failed launches for each site on the map

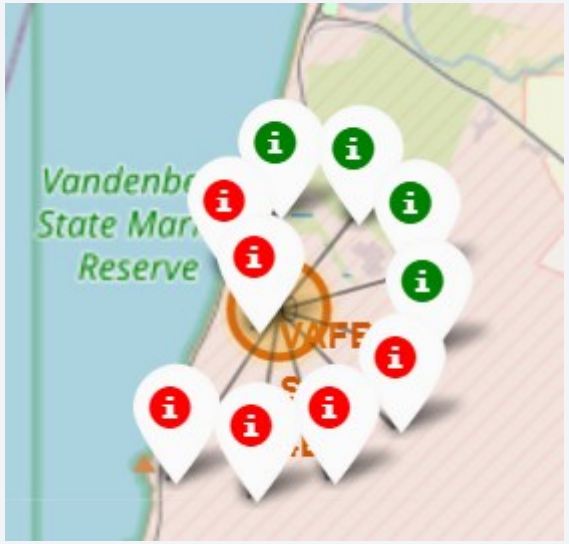
KSC LC-39A



CCAFS LC-40



CCAFS SLC-40



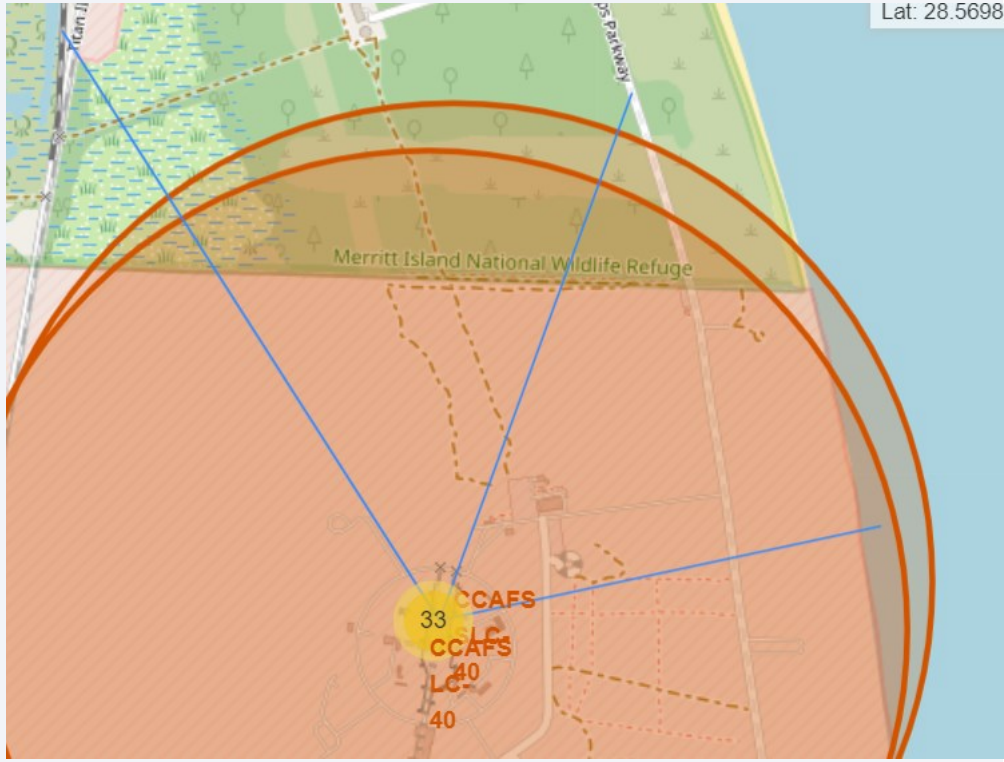
VAFB SLC 4E

From this picture, we can see that the launch site KSC LC-39A have the highest rate among others.

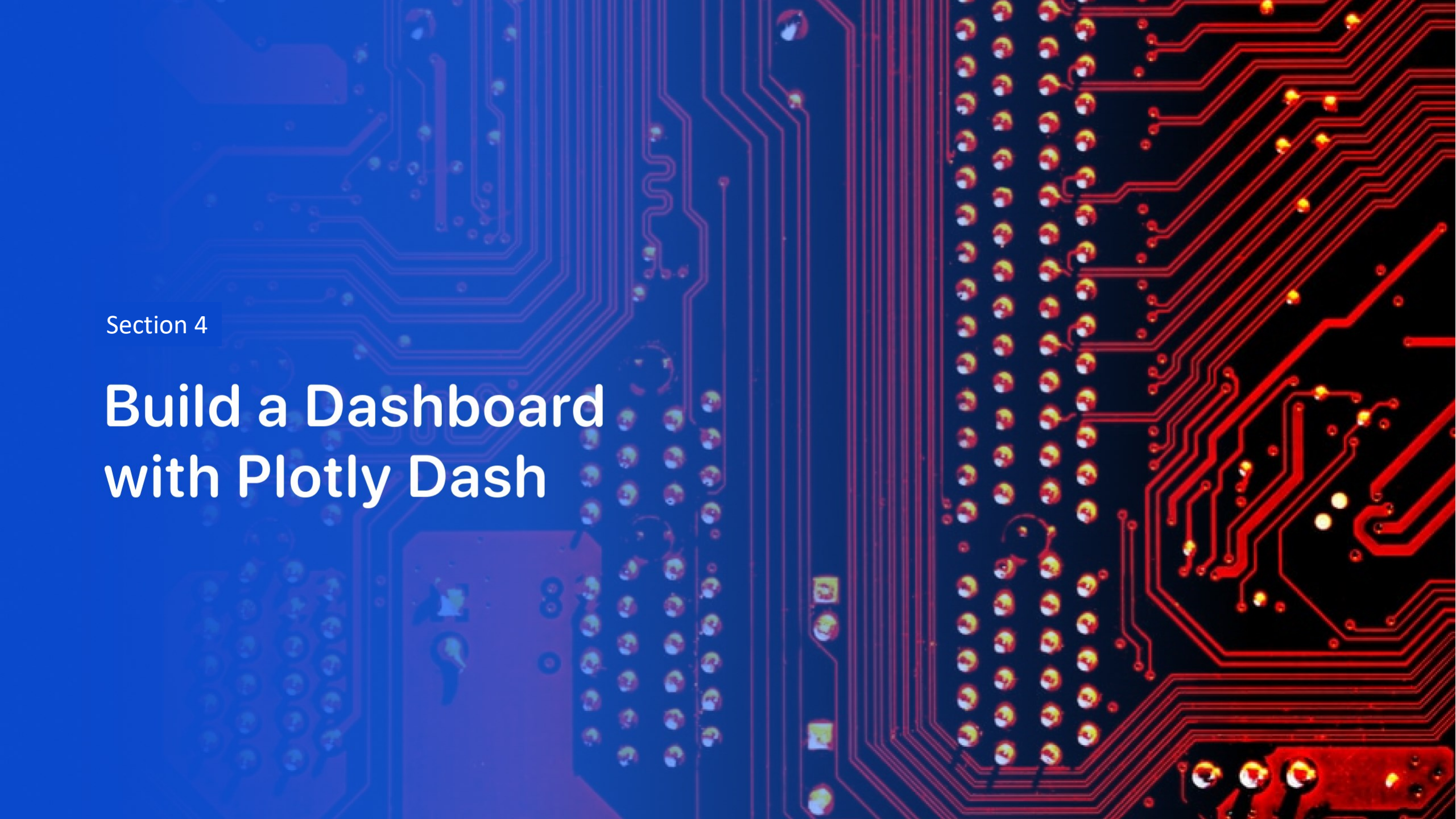
Draw PolyLine

Line proximity to closest city

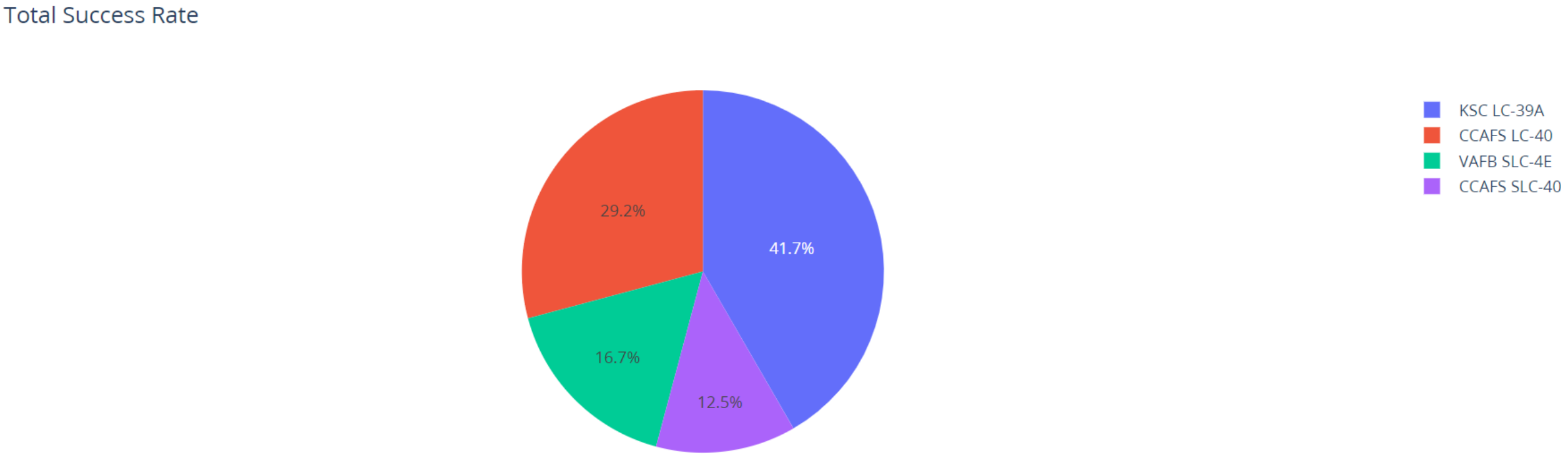
- Are launch sites in close proximity to railways? yes
- Are launch sites in close proximity to highways? yes
- Are launch sites in close proximity to coastline? yes
- Do launch sites keep certain distance away from cities? no



Line proximity to coastline, highways, and railways

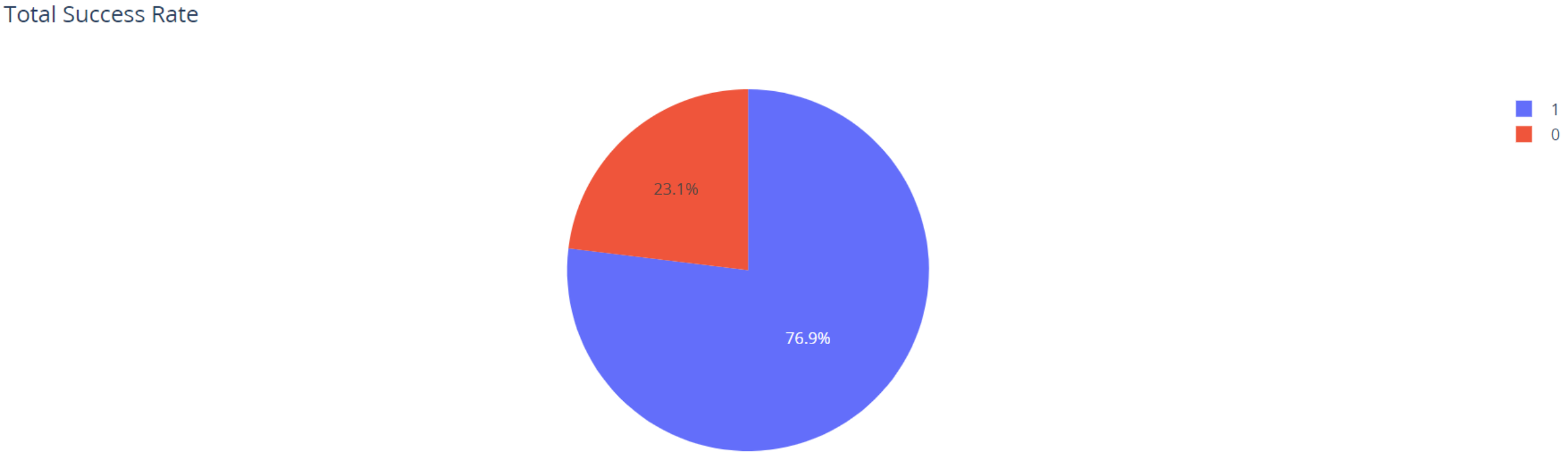

Section 4
Build a Dashboard with Plotly Dash

Pie chart of success rate in all Launch Site



From this pie chart, we can see KSC LC-39A Launch Site have the highest success rate among others with success rate is 41.7% from the total success landing records. The second highest is CCAFS LC-40 with success rate is 29.2% from the total success landing records.

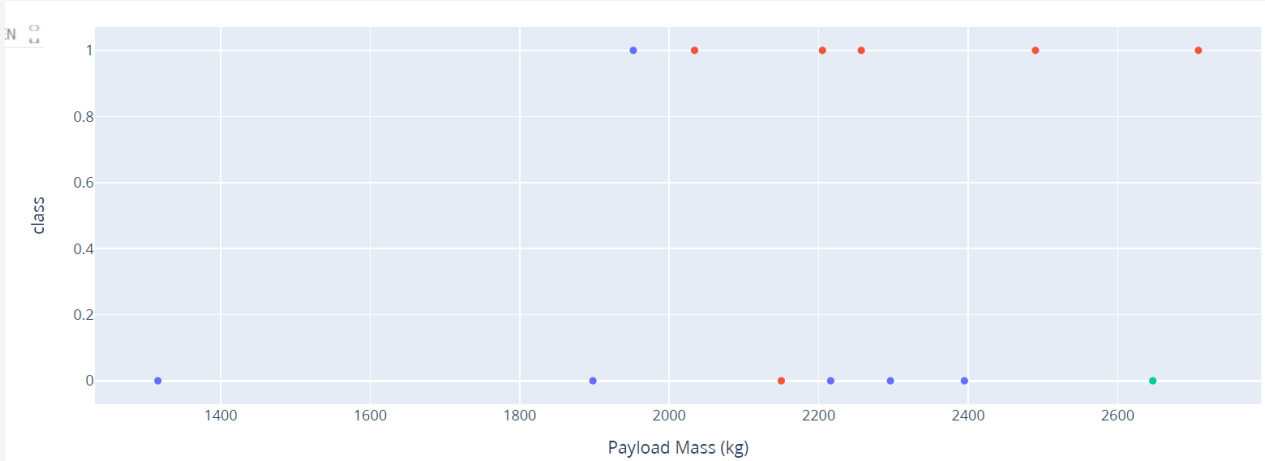
Pie Chart in Launch Site with Highest Success Rate

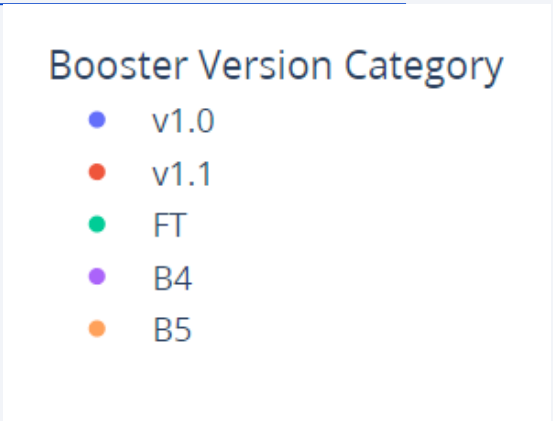


From this pie chart, we can see the success rate of KSC LC-39A is 76.9 % which is also the highest among other launch site. This result also already affirmed in previous section that KSC LC-39A have most successful landing compared to the others.

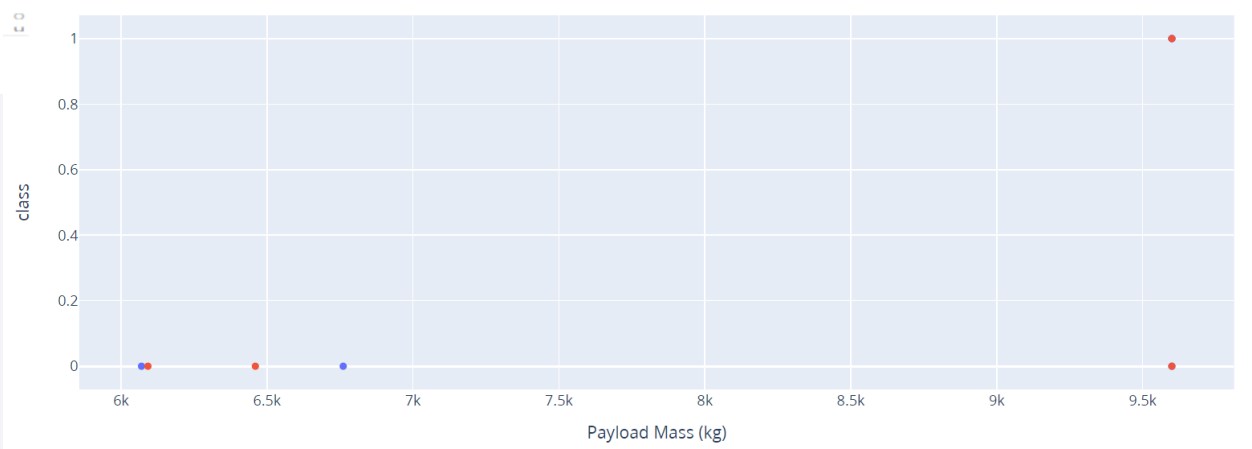
Scatter Plot for all Launch Site with certain Payload

Payload Mass (1000-3000)





Payload Mass (6000-9000)



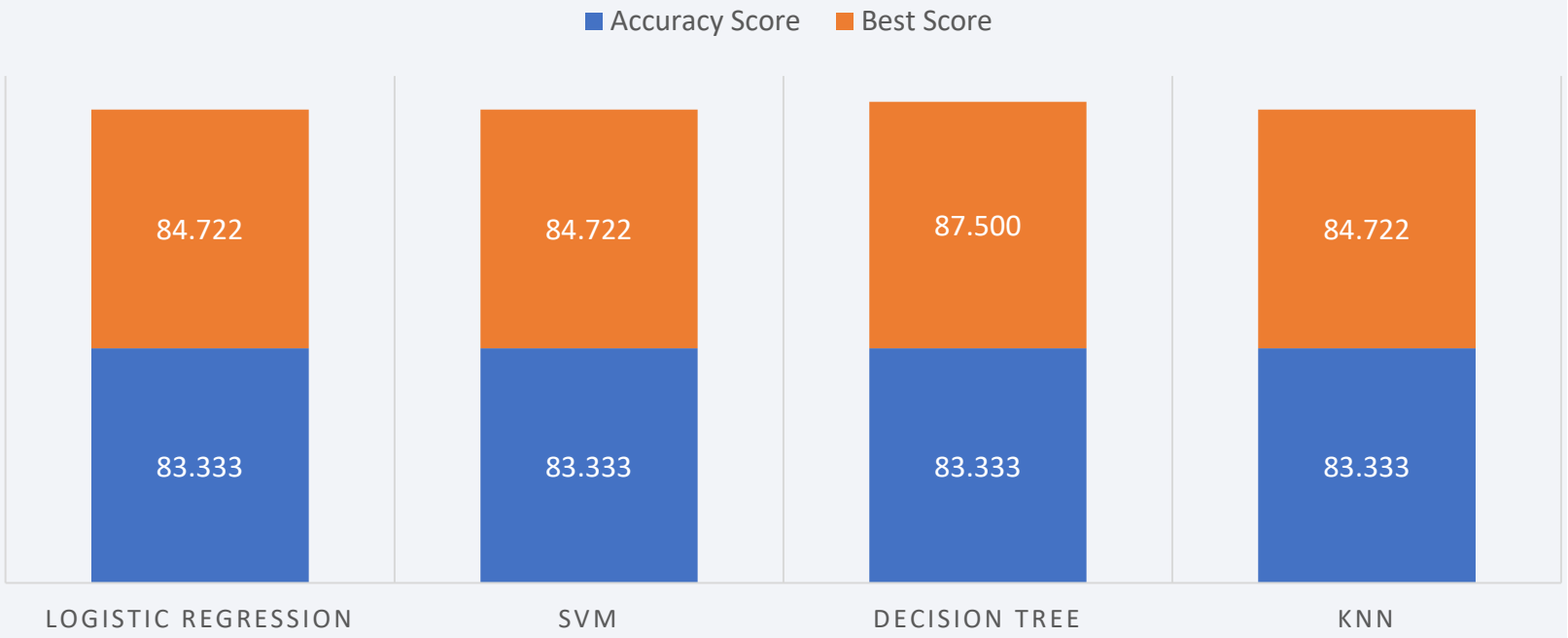
From the plots, we can see the success rate of lower payload mass (1000-3000) is greater than the one with more heavy payload mass (6000-9000).

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Based on the graph, the classification accuracy results for all four models (logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbour) shows the same score, which is 83.333%. This implies that each model performs equally well in predicting the success or failure of rocket landings, highlighting the robustness and consistency of their predictive capabilities.

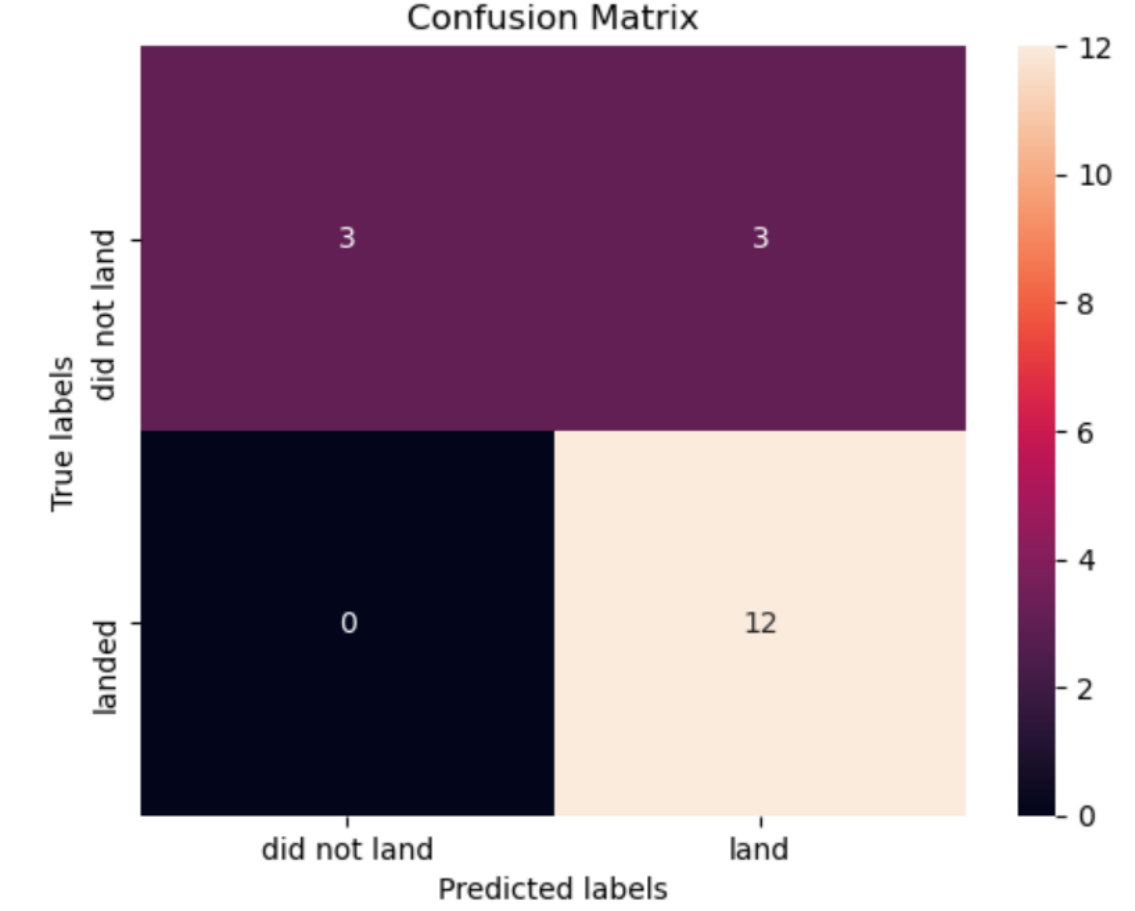


However, the best score results from the best parameters found using grid search indicate that the **Decision tree model** has the highest best score value of 87.5%.

Confusion Matrix

The graph beside show confusion matrix from all model (have the same result) which can be seen there is 12 test sample as True Positive, 3 test sample as True Negative, and 3 test sample as False Negative.

It means the model is still have mistake of predicting the record with did not land label and the model proficient in predict the data with landed label



Conclusions

- The greater the Flight Number, the higher the Success Rate at certain Launch Sites
- The Launch Success rate have positive trend line from 2010 until 2020 although there is a decrease in 2018
- Orbit type that have 100.0% success rate is: ES-L1, GEO, HEO, and SSO. All the orbit type have >=51.8% success rate
- Launch site KSC LC-39A have the highest success rate among others, it is 76.9%
- The Decision tree have the best 'best score' among other models, it is 87.50%
- The test accuracy score of the four model have the same exact score, it is 83.33%

Thank you!