IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sharon(Heng) Shi
Dec 13, 2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- SpaceX advertises Falcon 9 rocket launches with much of the savings, becuase SpaceX can reuse the first stage.
- We will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

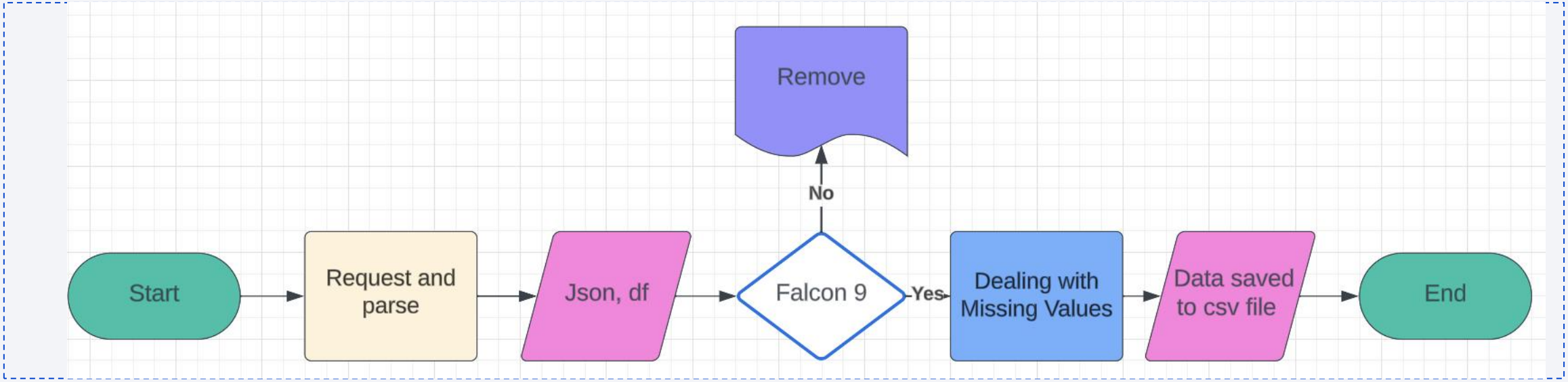
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data is collected in JSON format and using the GET request
- Process of data collection process
  - Request and parse the data using GET request
  - Filter the dataframe to only include Falcon 9 launches
  - Dealing with missing values

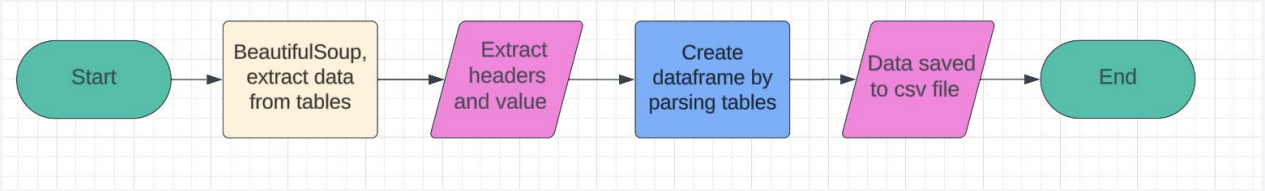
Data Collection – SpaceX API

- https://github.com/Sharon-SHH/IBM-DS-Certificate/blob/main/IBM_10_spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scrap Falcon 9 records with BeautifulSoup; Extract all data from table; Create a data frame by parsing tables.
- https://github.com/Sharon-SHH/IBM-DS-Certificate/blob/main/IBM_10-labs-webscraping.ipynb



Data Wrangling

- Exploratory Data Analysis and determine training labels.
- Determine the number of Launchsite, Orbit and Outcomes.
- https://github.com/Sharon-SHH/IBM-DS-Certificate/blob/main/IBM_10-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Using sns.catplot to visualize the relationship between Flight Number and Launch Site; the relationship between Payload and Launch Site; Using bar chart to visualize the relationship between success rate of each orbit type. catplot can show the relationship of two variables.
- https://github.com/Sharon-SHH/IBM-DS-Certificate/blob/main/IBM_10-dataviz.ipynb

EDA with SQL

- SQL Queries
  - select * from table where Launch_Site
  - like “CCA%”
  - limit(5)
  - sum(Payload_mass_kg)
  - avg(Payload_mass_kg)
  - min(Date), count()
- https://github.com/Sharon-SHH/IBM-DS-Certificate/blob/main/IBM_10-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Using folium.Map to create map objects; Marker can be used to point the specific location and add the makers to a folium map.
- To create a map and create a group of markers on a map.
- https://github.com/Sharon-SHH/IBM-DS-Certificate/blob/main/IBM_10_launchSite_location.ipynb

Build a Dashboard with Plotly Dash

- List plots/graphs and interactions added to a dashboard:
  - Dropdownlist
  - Pie chart
  - Range Slider
  - A callback function (interaction)
- Display all elements on the page clearly and it's user-friendly.
- https://github.com/Sharon-SHH/IBM-DS-Certificate/blob/main/DV_Final_Assignment_Part_2.py

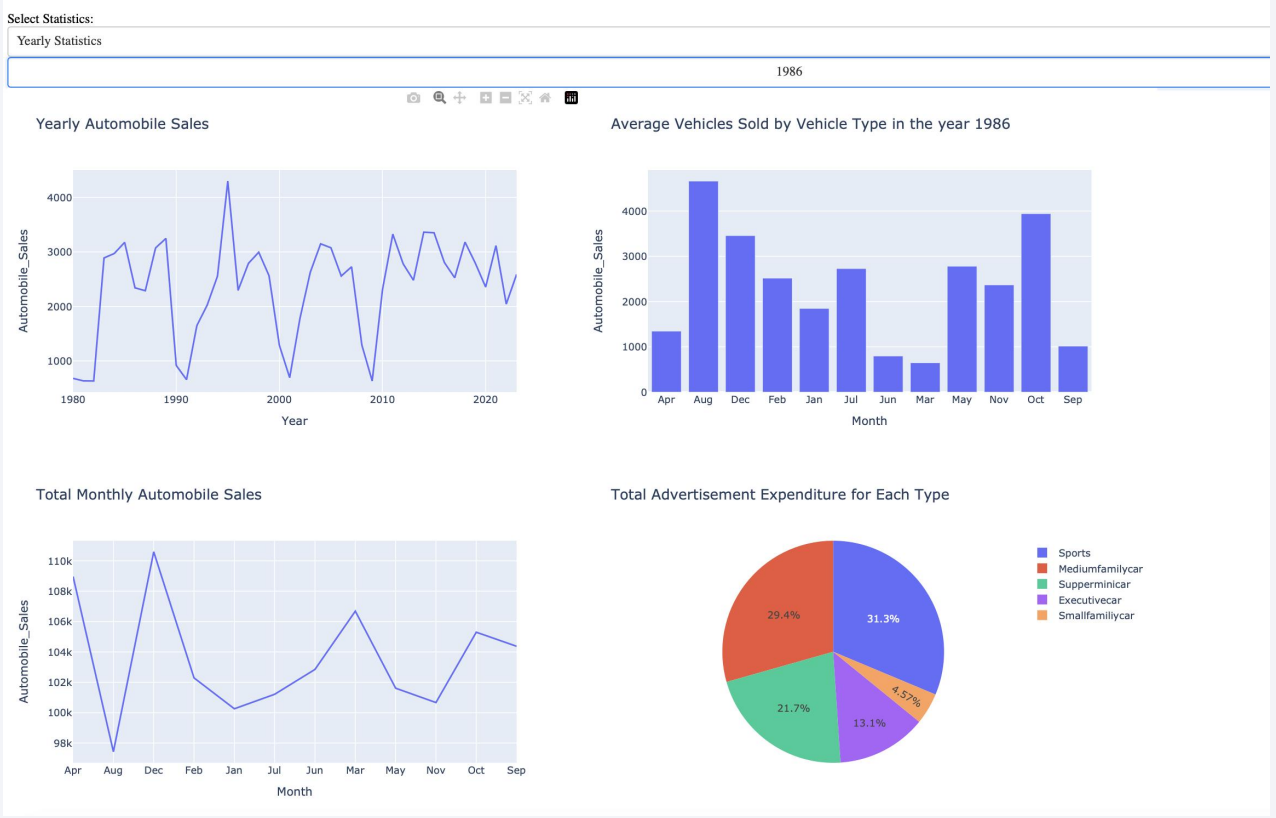
Predictive Analysis (Classification)

- Build a classification model:
  - Load the data and preprocessing: apply StandardScaler to data
  - Use train_test_split to split training data and test data.
  - Create logistic regression model, SVM and decision tree, use training data to train model and get beset parameters.

- Create a model, select significant parameters for it and using training data to train the model and get the best parameters.
- https://github.com/Sharon-SHH/IBM-DS-Certificate/blob/main/IBM_10SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Results

- We find that FlightNumber has a great relationship to Orbit.
- Applied logistic regression, svm, decision tree, knn models, we find that decision tree gives the best performance.

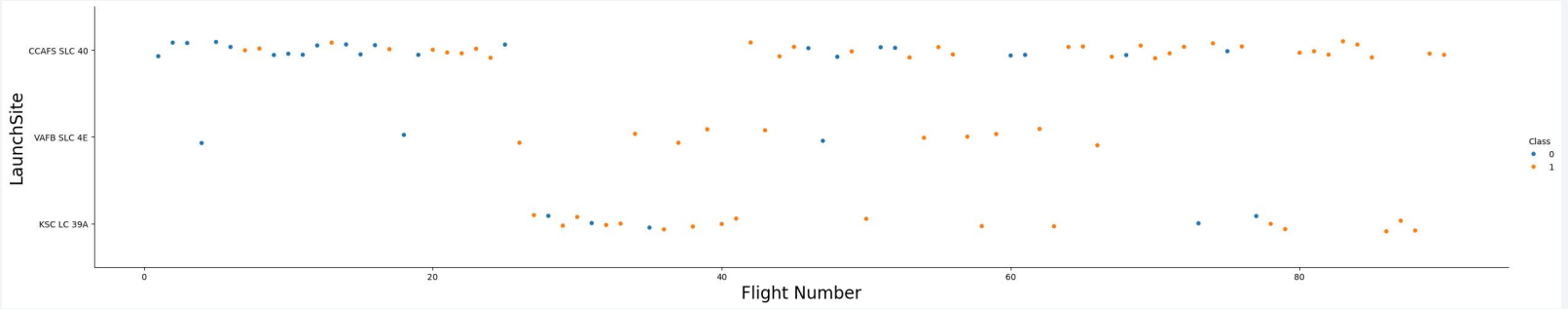


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

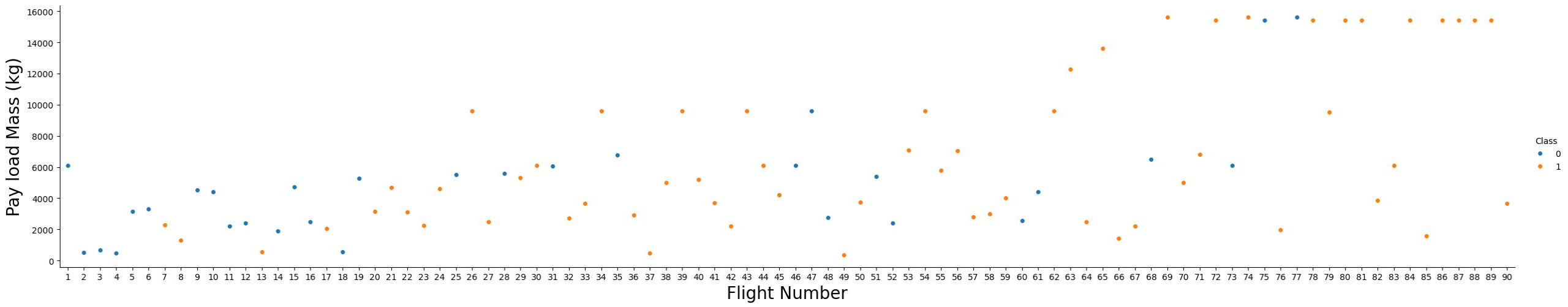
- A scatter plot of Flight Number vs. Launch Site
- X–axis is Flight Number, Y–axis is Launch Site. “CCAFS SLC 40” are mainly selected.



Payload vs. Launch Site

- A scatter plot of Payload vs. Launch Site

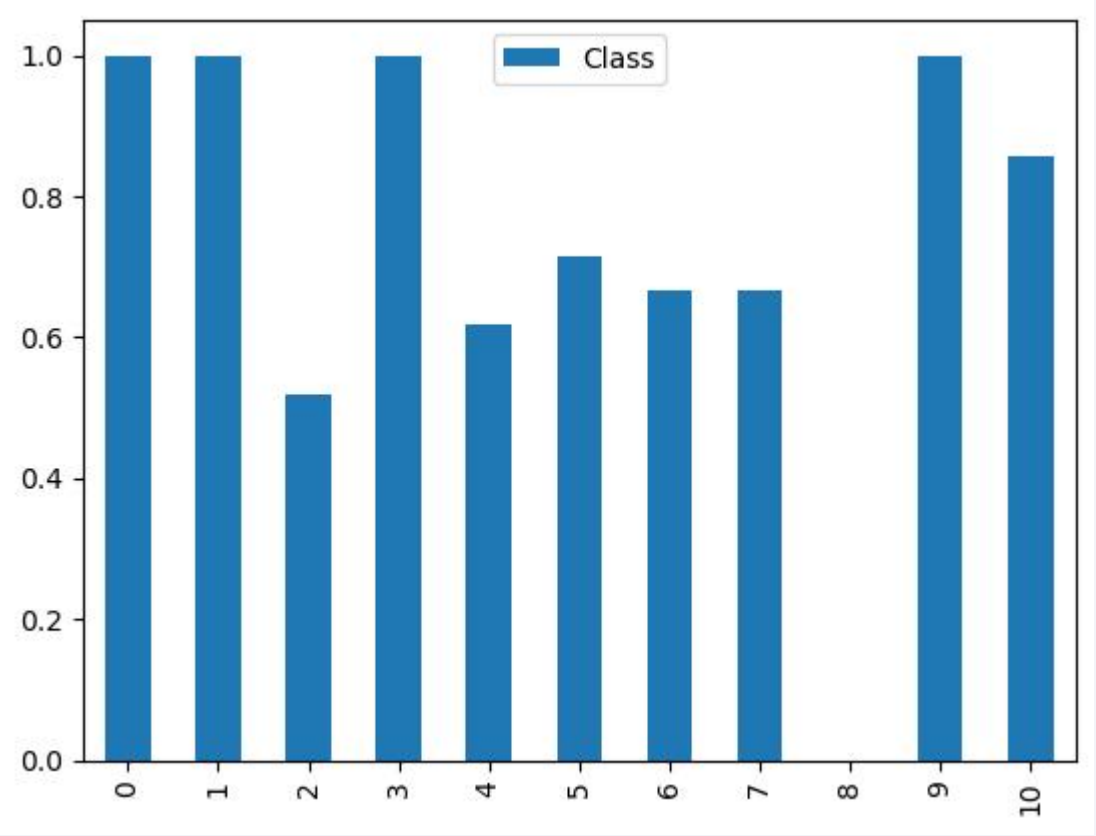
- When Flight Number increases, the first stage is more likely to land successfully. The payload mass is more massive, the first stage will less likely return.



Success Rate vs. Orbit Type

- A bar chart for the success rate of each orbit type

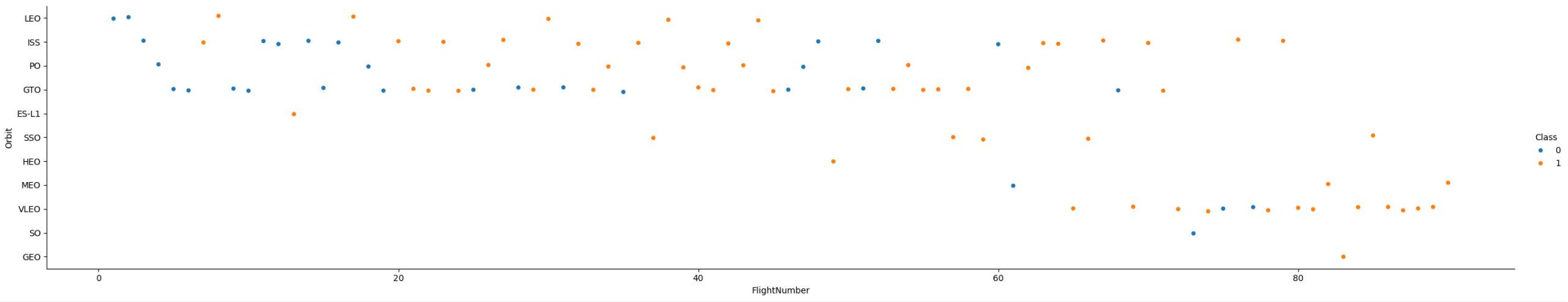
- Most of orbits have high success rate, except 8th–MEO of the orbits.



Flight Number vs. Orbit Type

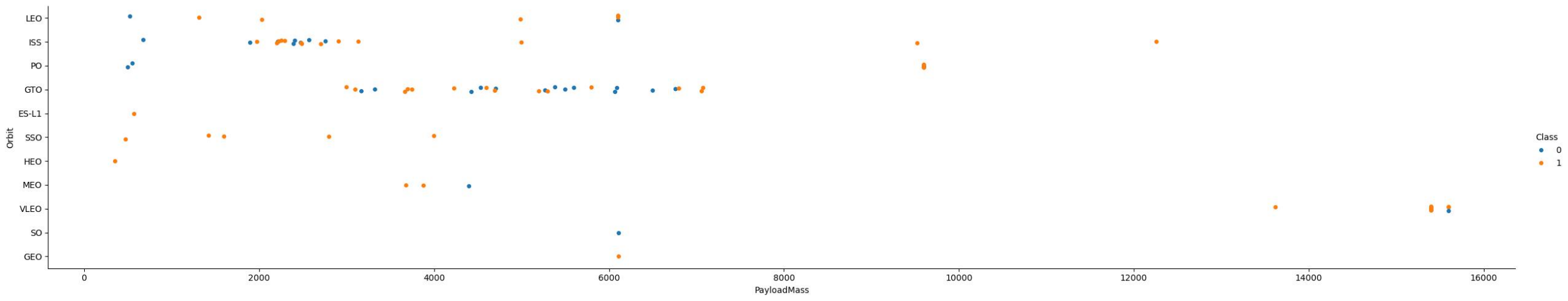
- A scatter point of Flight number vs. Orbit type

- When Flight number is between 20 to 60 and orbits are included in LEO, ISS, PO, GTO, the first stage has a bigger success rate. When Flight number is greater than 60, it also has a bigger success rate.



Payload vs. Orbit Type

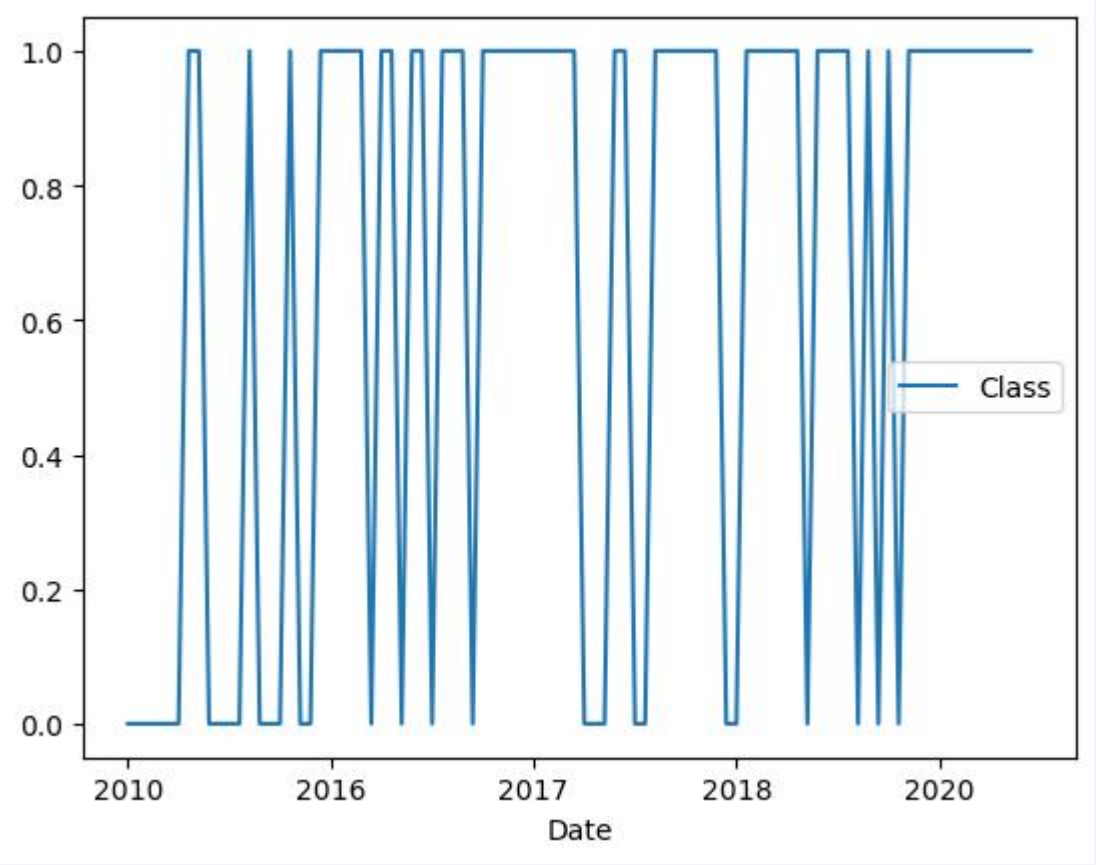
- A scatter point of payload vs. orbit type
- With heavy payloads the successful landing or positive landing rate are more for PO (polar), LEO and ISS.



Launch Success Yearly Trend

- A line chart of yearly average success rate

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

- The unique launch sites
- distinct to get unique elements

```
%sql select distinct `Launch_Site` from SPACEXTABLE
```

```
* sqlite:///my_data1.db
one.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Display 5 records where launch sites begin with `CCA`

```
%sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit(5)
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Calculate the total payload carried by boosters from NASA
- Use sum() to compute the total payload carried from NASA

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

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Use avg() to compute the average payload mass.

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version="F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- To get the first successful landing outcome, use min() to get the earliest date.

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

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql select Booster_Version from SPACEXTABLE where Landing_Outcome="Success (drone ship)" and PAYLOAD_MASS__KG_ > 4000
        and PAYLOAD_MASS__KG_ < 6000
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- To get both successful and failure mission outcomes, use count() and or in the where condition.

```
%sql select count() from SPACEXTABLE where Mission_Outcome = "Success" or Mission_Outcome = "failure"
```

```
 * sqlite:///my_data1.db
Done.
```

count()
98

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Use a subquery to get the maximum of payload mass, the retrieve the names of the boosters.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql select Booster_Version from SPACEXTABLE
        where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

- Use substr() to get Month and Year; there're two conditions in this query.

```
%%sql select substr(`Date`, 6, 2) as Month, Booster_Version,Launch_Site from SPACEXTABLE
        where substr(`Date`, 0, 5) = '2015' and Landing_Outcome like "Failure%"
```

```
 * sqlite:///my_data1.db
Done.
```

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

- 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

```
%%sql select Landing_Outcome, count(*) from
        (select * from SPACEXTABLE where `Date` >= '2010-06-04' and  `Date` <= '2017-03-20'
        order by `Date` Desc)
        group by Landing_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

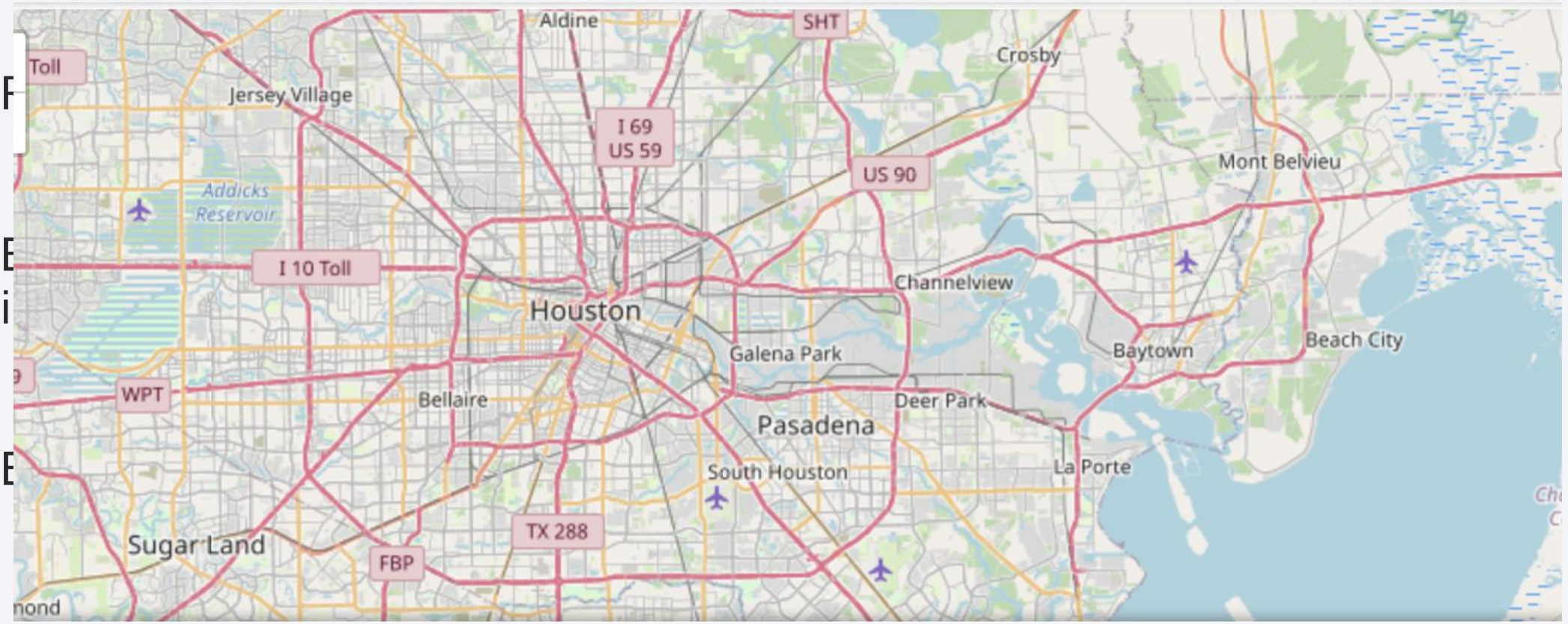

Section 3
Launch Sites
Proximities Analysis

Launch Sites on a global map

- We can only see two points on the map, but there are four points on it. Because two points are much closer.

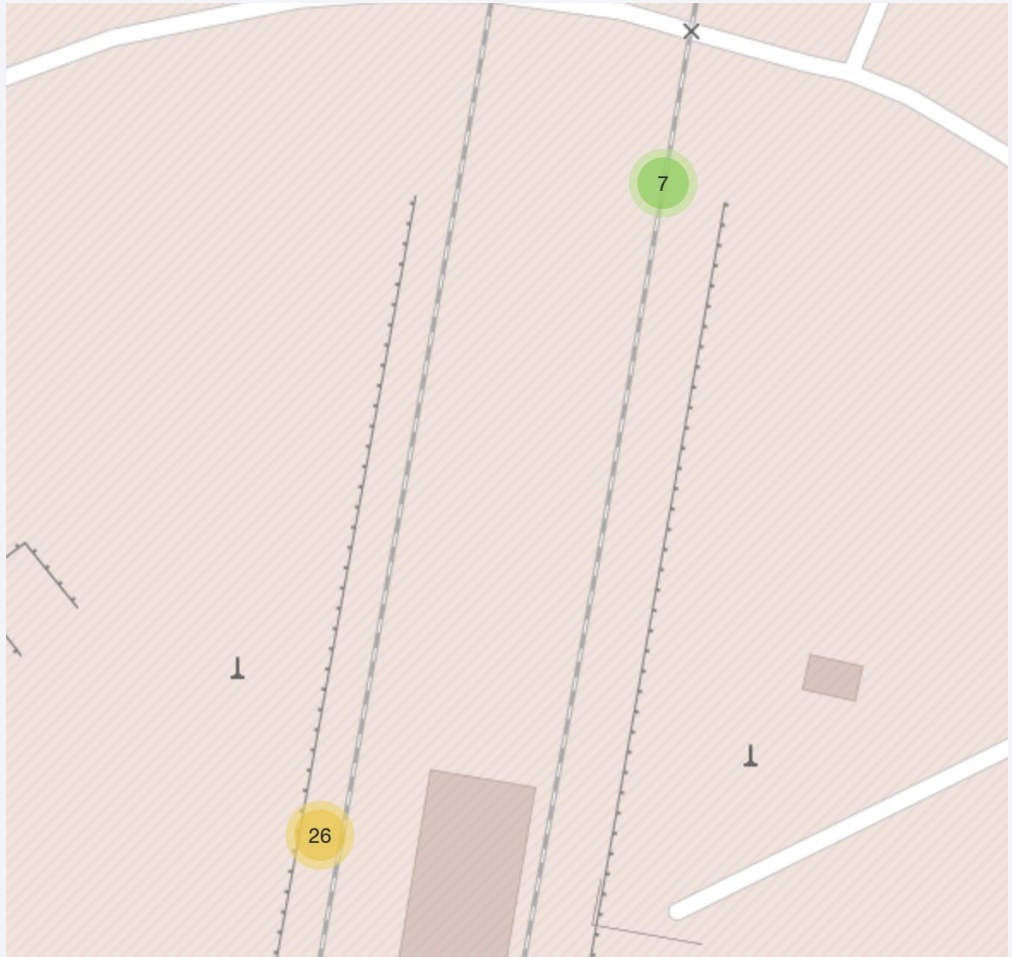
Locat to NASA Johnson Space Center at Houston, Texas

- F
- E
  i
- E



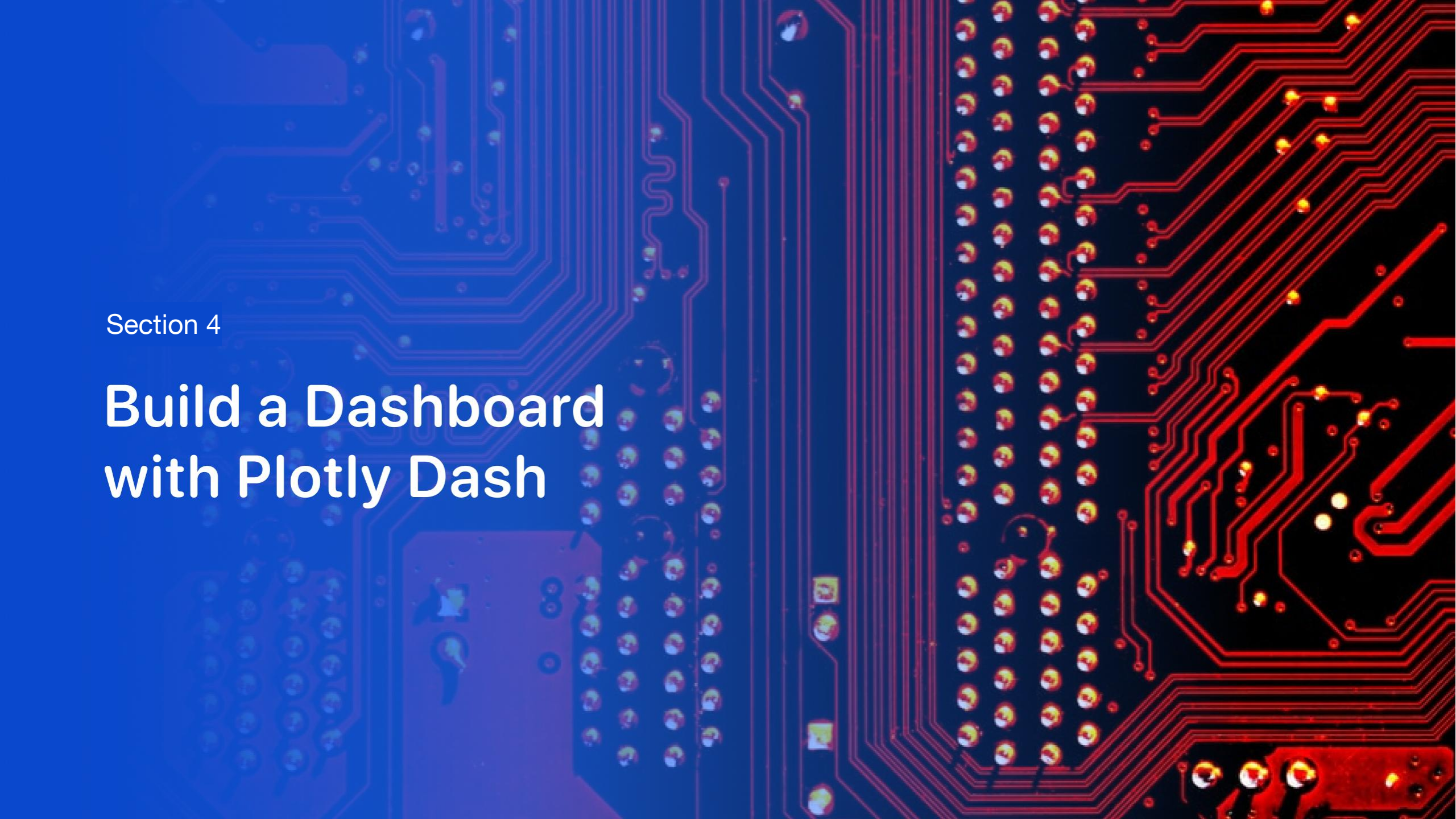
Label success/failure of the launch

- If a launch is successful, “class=1” and label is in green marker.
- If a launch is failed, “class=0” and label is in red marker.



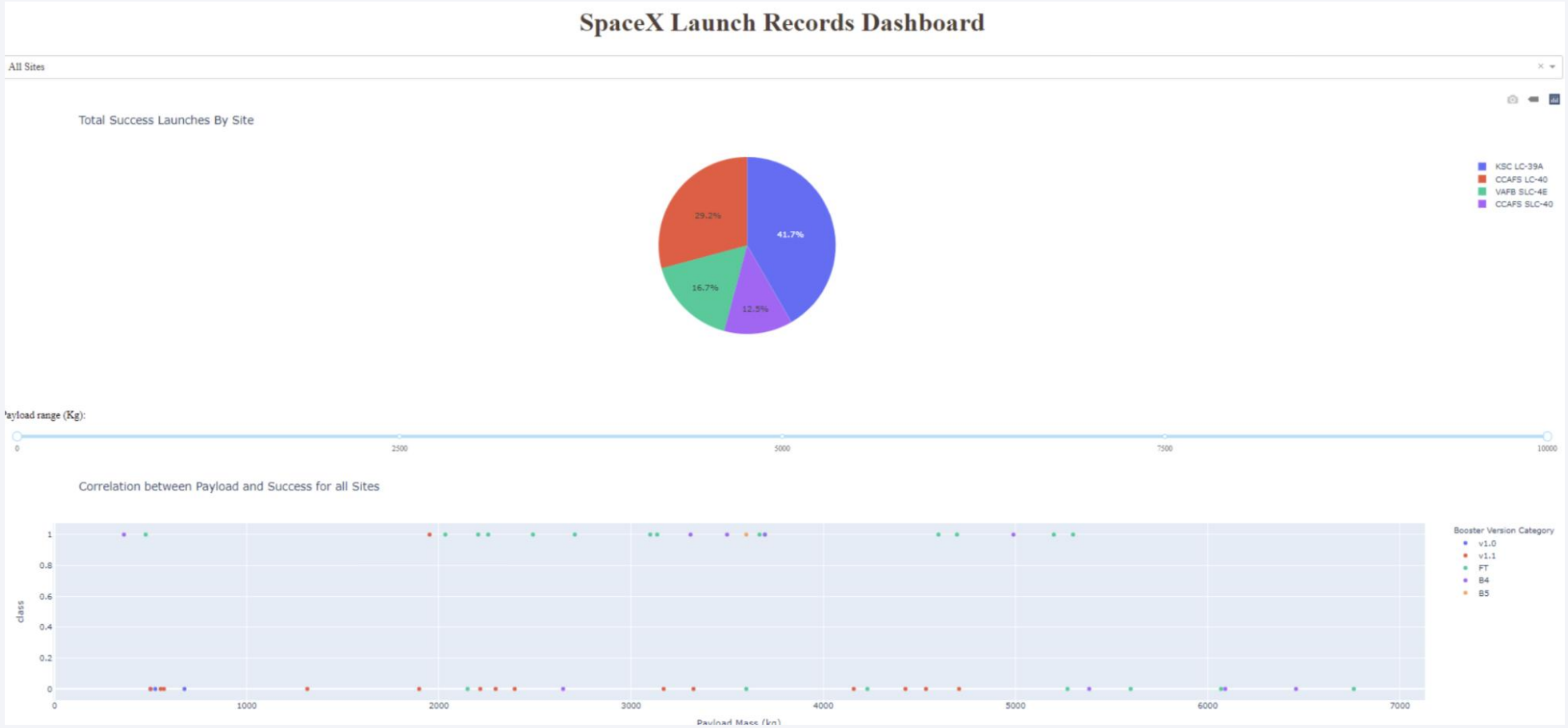
Draw the distances between a launch site to the closest coast

- Find coordinate of the closest coastline and use PolyLine to draw the distance between a launch site to the closest coast.


Section 4
Build a Dashboard with Plotly Dash

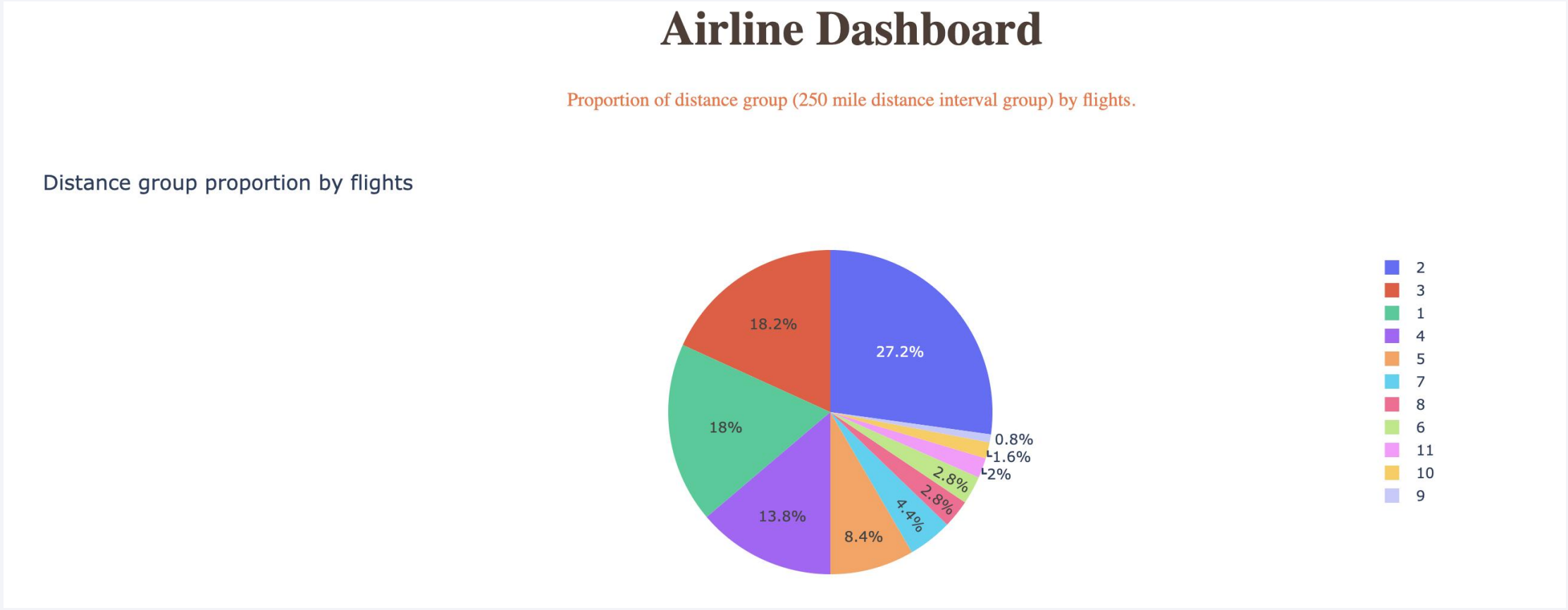
SpaceX Launch Records Dashboard

- Total success Launches by site



Airline Dashboard

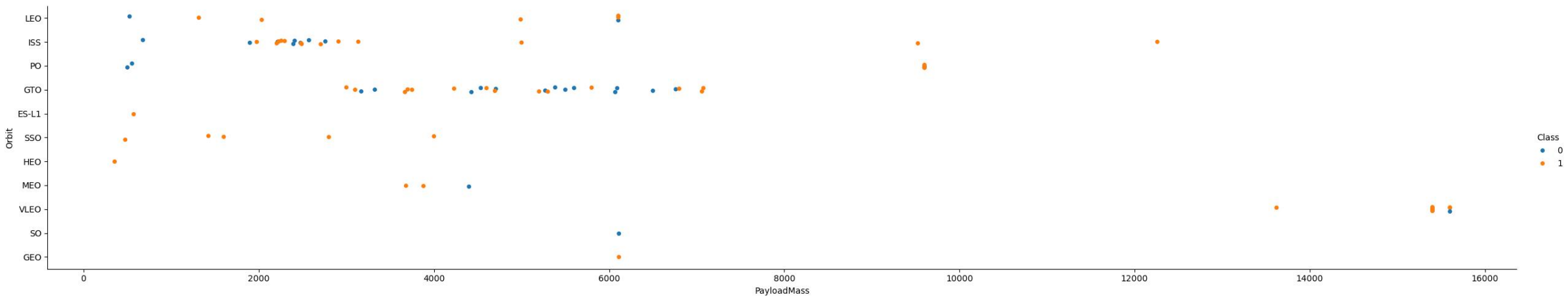
- Distance group proportion by flights



Range Slider to Select Payload

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

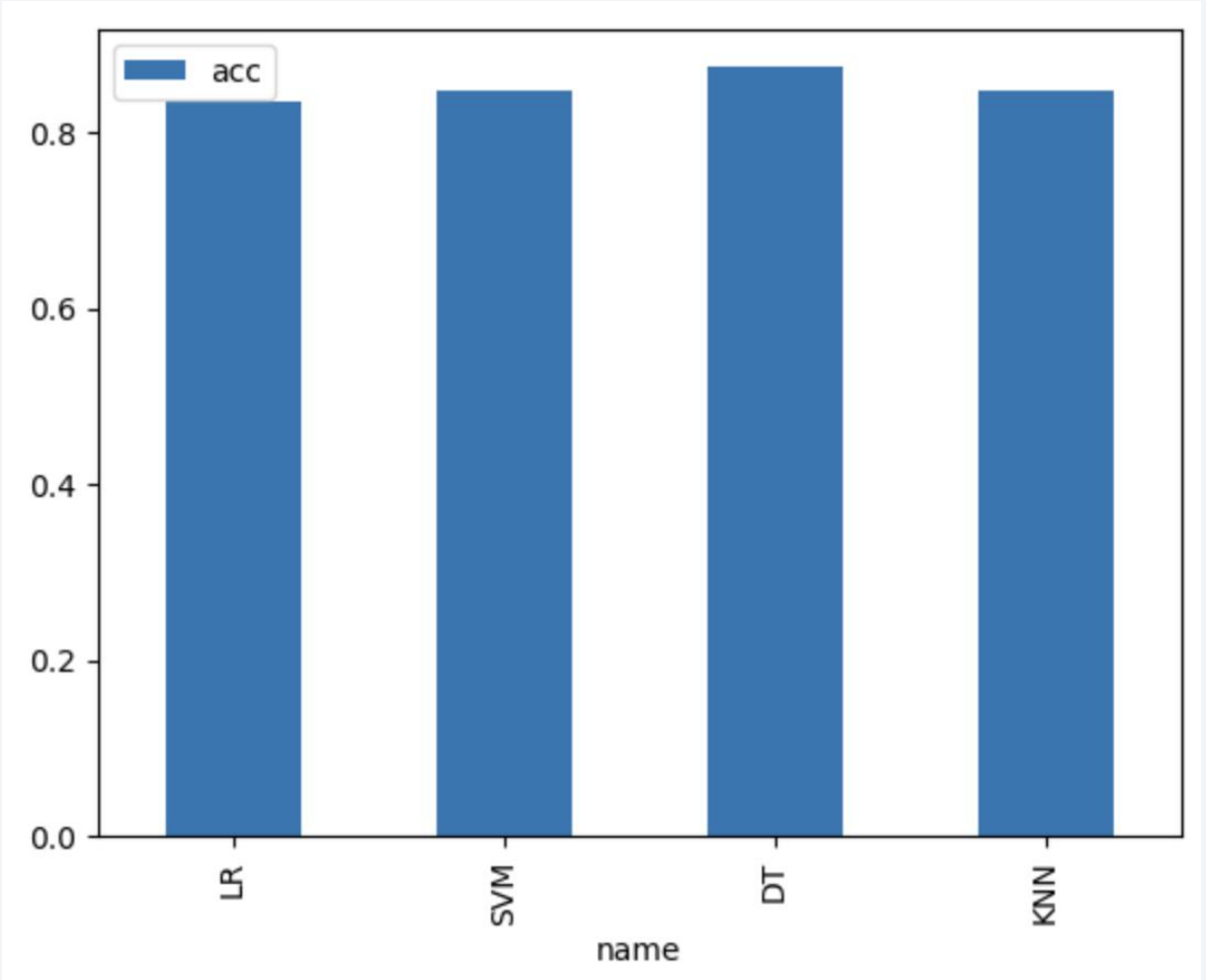


Section 5

Predictive Analysis (Classification)

Classification Accuracy

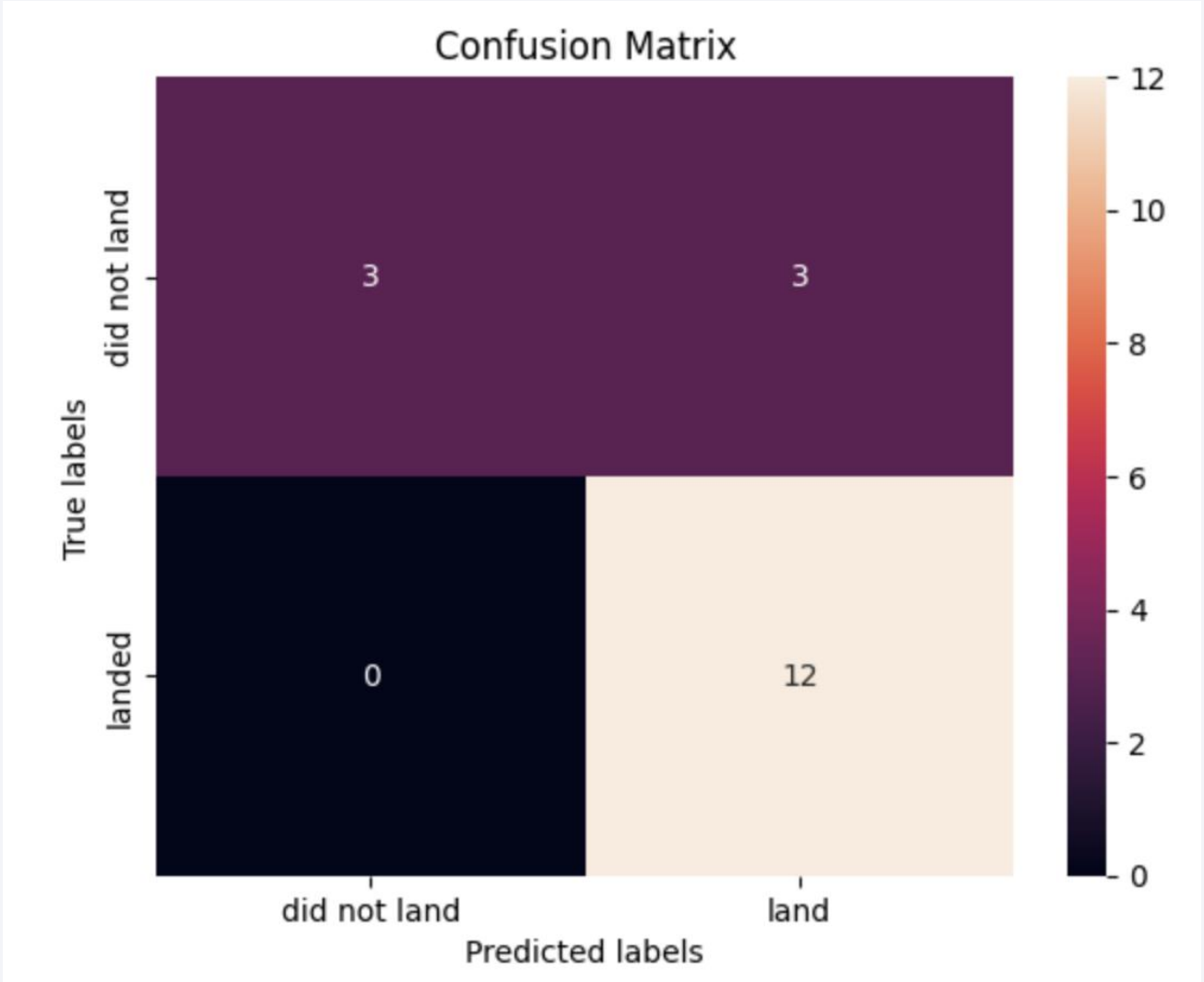
- For models as shows in the bar chart, Decision Tree gives the best classification accuracy.



Confusion Matrix

- TP=3, TN=12, FN=3, FP=0
- acc = (TP + TN)/(TP+TN+FN+FP)
- precision = TP/(TP+FP) = 1
- recall = TP/(TP+FN) = 0.5

It is concluded that the precision is excellent, but the recall is satisfactory.



Conclusions

- SpaceX Falcon 9 works well in first stage landing.
- Flight number, Payload, and Orbit Type impact the success of the launches.
- Decision Tree yields optimal performance when predicting the success of a landing.

Appendix

- https://github.com/Sharon-SHH/IBM-DS-Certificate/tree/main

Thank you!