IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hoshimov Mirzohamidullo
06/15/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis with SQL
  - Interactive analytics charts
  - Predictive Analytics result

Introduction

- Project background and context
- SpaceX's Falcon 9 launches are much cheaper (around $62 million) than competitors (over $165 million) because they reuse the first stage.. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
  - What are the factors if the rockets will land successfully?
  - The influence of different features that determine the success rate of a successful landing.
  - What conditions are required to ensure a successful landing outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from two sources:
  - Official API of SpaceX: (https://api.spacexdata.com/v4/rockets/)
  - Wikipedia Page Web Scraping : (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - After the summarization and analysis the landing outcome was labeled as True and False based and feature metrics

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - First data was cleaned, performed one-hot encoding and normalization to ensure scaling, divided into test and training sets. Evaluation was based on four classification models (Decission Tree, KNN, Logistic Regression, SVM). The best hyperparameters were selected by using GridSearchCV, which gave the most optimal parameters. According to the accuracy outcome Decision Tree was the best performing model

Data Collection

- Describe how data sets were collected.
- Data was collected from SpaceX API (https://api.spacexdata.com/v4/rockets/)
- As well as from Wikipedia (Web Scraping)
- (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches),

Data Collection - SpaceX API

SpaceX has publicly available API where the data can be accessed

Above you can see the flowchart how API was used

Notebook with code:

https://github.com/Hafizulloevich/IBM-Data-Science-Professional-Certificate/blob/f657cc1a4e03fb0333d9da75f15ad8db70fb2384/Applied%20Data%20Science%20Capstone/Data_Collection_API.ipynb

Data Collection - Scraping

Wikipedia page was used for Web Scraping to obtain the SpaceX data;

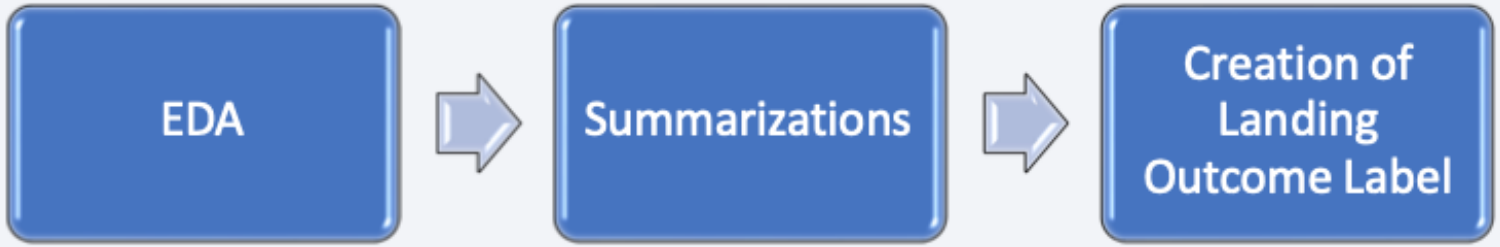
- Data are downloaded from Wikipedia according to the flowchart and then persisted

Notebook with code:

https://github.com/Hafizulloevich/IBM-Data-Science-Professional-Certificate/blob/f657cc1a4e03fb0333d9da75f15ad8db70fb2384/Applied%20Data%20Science%20Capstone/Webscraping.ipynb

Data Wrangling

- Exploratory data analysis were performed with SQL Queries as well as with Visualization techniques.
- The steps are in the below flowchart



https://github.com/Hafizulloevich/IBM-Data-Science-Professional-Certificate/blob/5942be6963663aba6e4d486717ff6731677d1357/Applied%20Data%20Science%20Capstone/Data%20Wrangling.ipynb

EDA with Data Visualization

- The following actions were performed:
- **Visualize the relationship between Flight Number and Launch Site**
- **Visualize the relationship between Payload and Launch Site**
- **Visualize the relationship between success rate of each orbit type**
- **Visualize the relationship between FlightNumber and Orbit type**
- **Visualize the relationship between Payload and Orbit type**
- **Visualize the launch success yearly trend**
- **These graphs were generated to understand the relationships between variables**

https://github.com/Hafizulloevich/IBM-Data-Science-Professional-Certificate/blob/66e4922e384c6c8c08ebe52edebe6ce5cc697a9a/Applied%20Data%20Science%20Capstone/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- SQL queries were performed:
  - Distinct Launchpads for the space mission;
  - The top 5 launch sites with 'CCA' naming
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing out comes, their booster versions, and launch site names for in year 2015;
  - Rank of the count of landing outcomes (such as Failure or Success) between the date 2010-06-04 and 2017-03-20.

https://github.com/Hafizulloevich/IBM-Data-Science-Professional-Certificate/blob/66e4922e384c6c8c08ebe52edebe6ce5cc697a9a/Applied%20Data%20Science%20Capstone/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and cluster markers were created using Folium Maps
- Launch sites were illustrated by Markers;
- Circles were used to indicate areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters highlighted events in each coordinate, like launches sites
- Lines are used to indicate distances between two coordinates.

https://github.com/Hafizulloevich/IBM-Data-Science-Professional-Certificate/blob/56438816da6152cd403f1d3c8918d1584bf6ed8c/Applied%20Data%20Science%20Capstone/Folium_Visualization.ipynb

Build a Dashboard with Plotly Dash

- We generated a dashboard with Plotly Express and Dash
- Pie charts illustrating successful Launches for certain sites were created
- Created scatterplots to show the Relationship of Payload Mass vs Success Rate
- We added a slider to select the payload range

https://github.com/Hafizulloevich/IBM-Data-Science-Professional-Certificate/blob/71542d09ec1846c523ce1272c82db940b4c5f9e4/Applied%20Data%20Science%20Capstone/spacex_dash_visualizations.py

Predictive Analysis (Classification)

•We imported the data with pandas, transformed it with NumPy and Pandas, and divided it into training and testing sets.

•We developed various machine learning models and adjusted their hyperparameters using GridSearchCV.

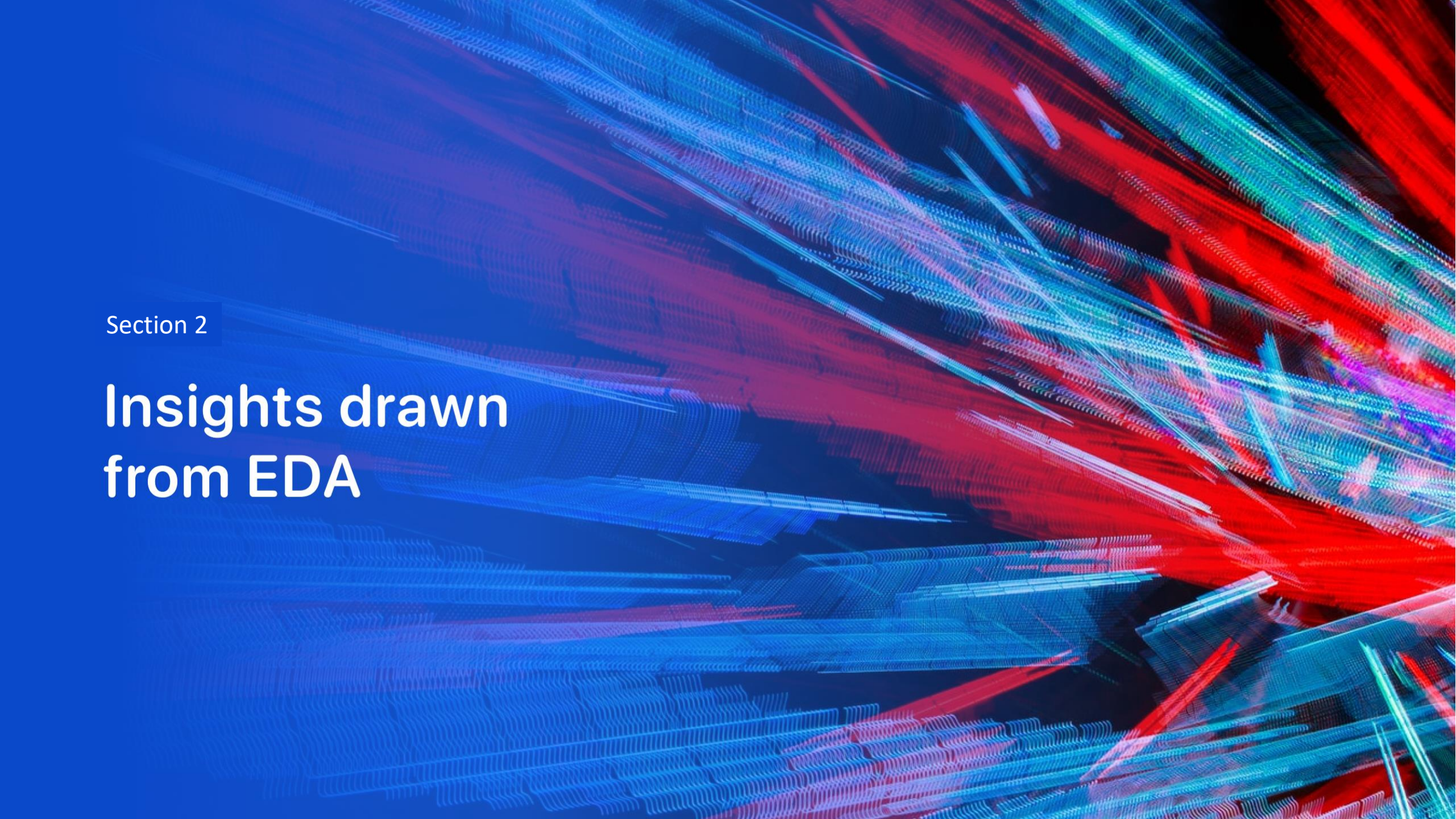
•We evaluated our models using accuracy, enhancing them through feature engineering and algorithm optimization.

•We identified the top-performing classification model

https://github.com/Hafizulloevich/IBM-Data-Science-Professional-Certificate/blob/71542d09ec1846c523ce1272c82db940b4c5f9e4/Applied%20Data%20Science%20Capstone/SpaceX_Machine_Learning.jupyterlite.ipynb

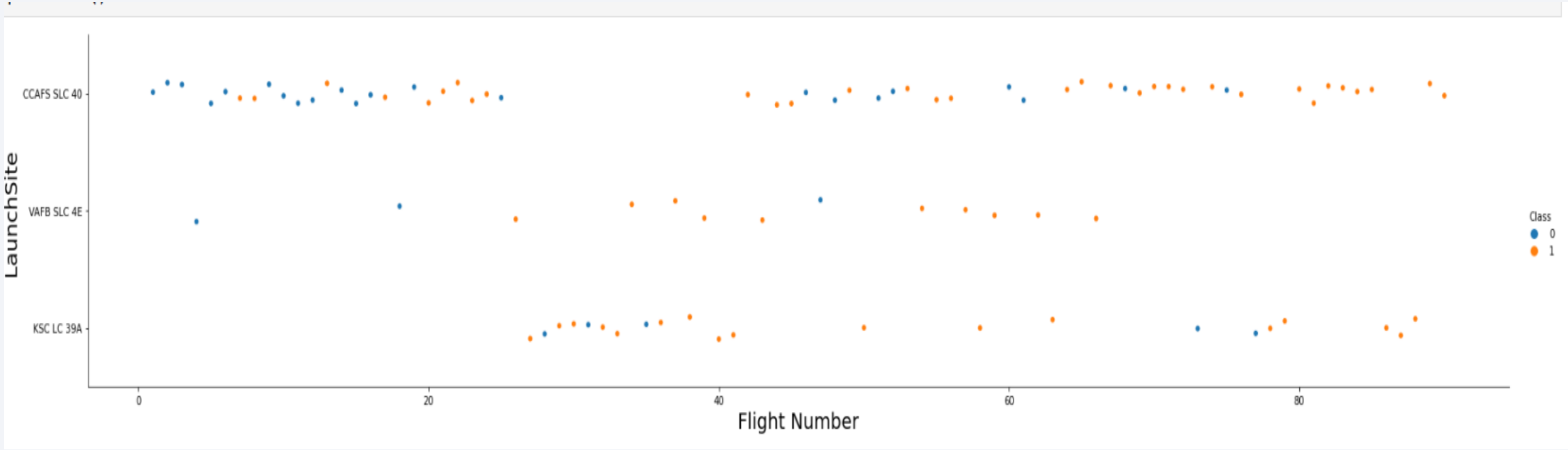
Results

- SpaceX used 4 launch sites in Houston
- First successful landing was performed in 2015
- The average Payload of Falcon 9 is 2928 kg
- Most launches with payload mass under 10,000 kg are from any launch site, but heavier ones happens mainly at CCAFS SLC 40 and KSC LC 39A
- As more time passed as more successful landings increased


Section 2
Insights drawn from EDA

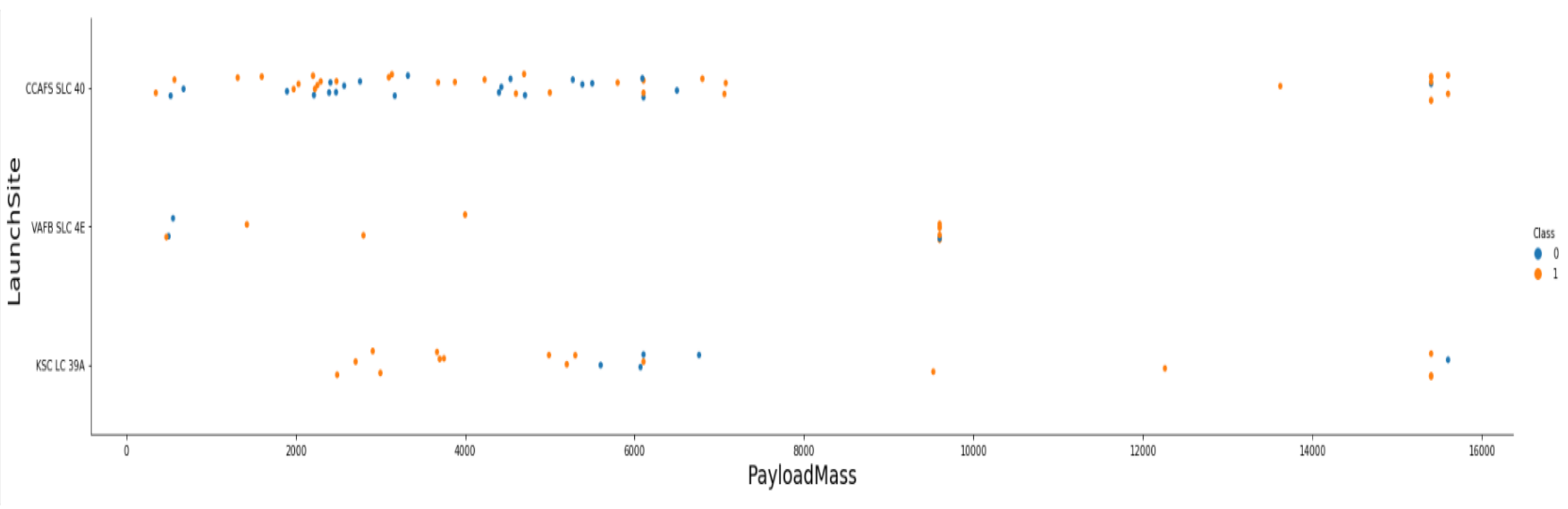
Flight Number vs. Launch Site

- As we can see, the more flights are in the launch site, the more successful landing it has. The most successful launch site is CCAFS SLC 40



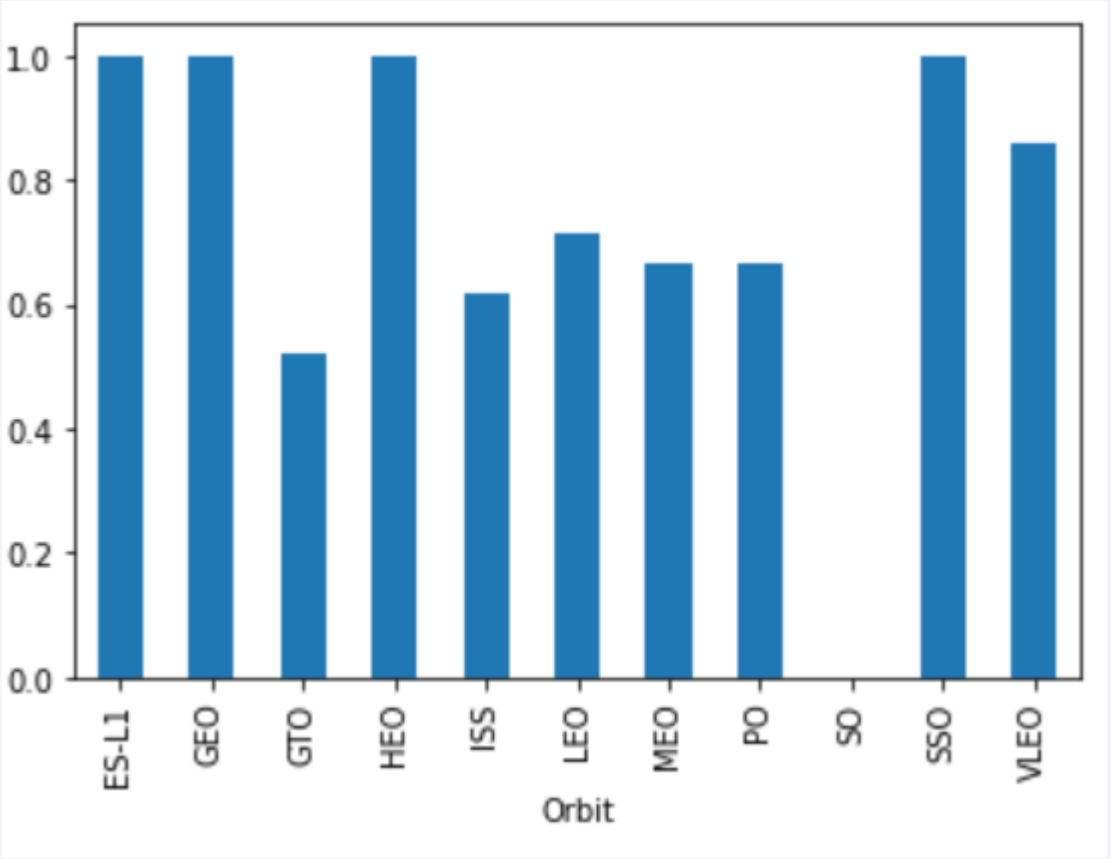
Payload vs. Launch Site

- From the chart we see that launches with Payload Mass higher 8000 have most successful landing outcomes



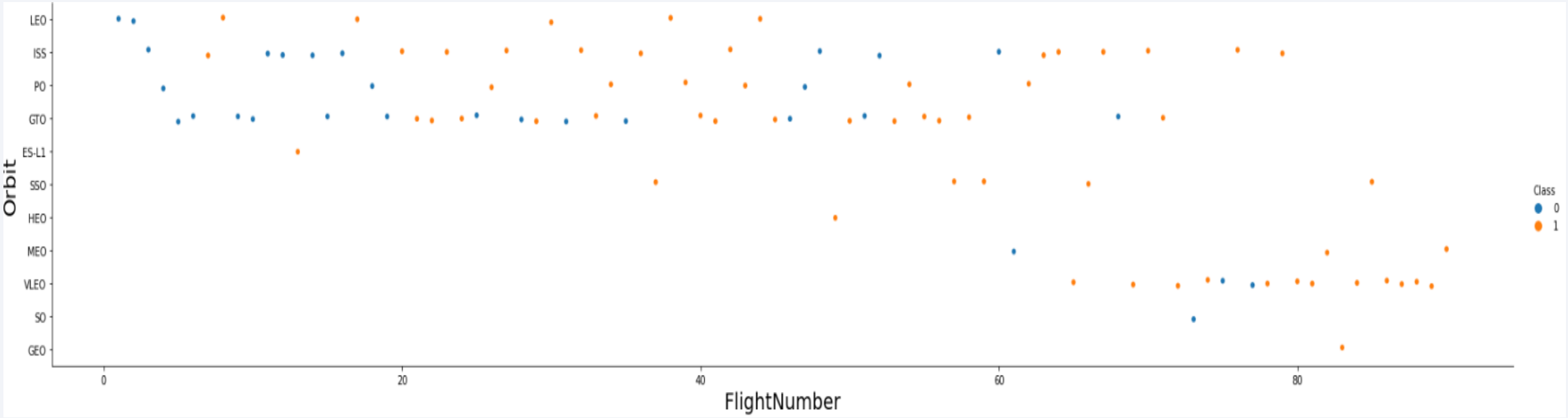
Success Rate vs. Orbit Type

- As illustrated ES-L1, GEO, HEO, SSO have success rate of 100%



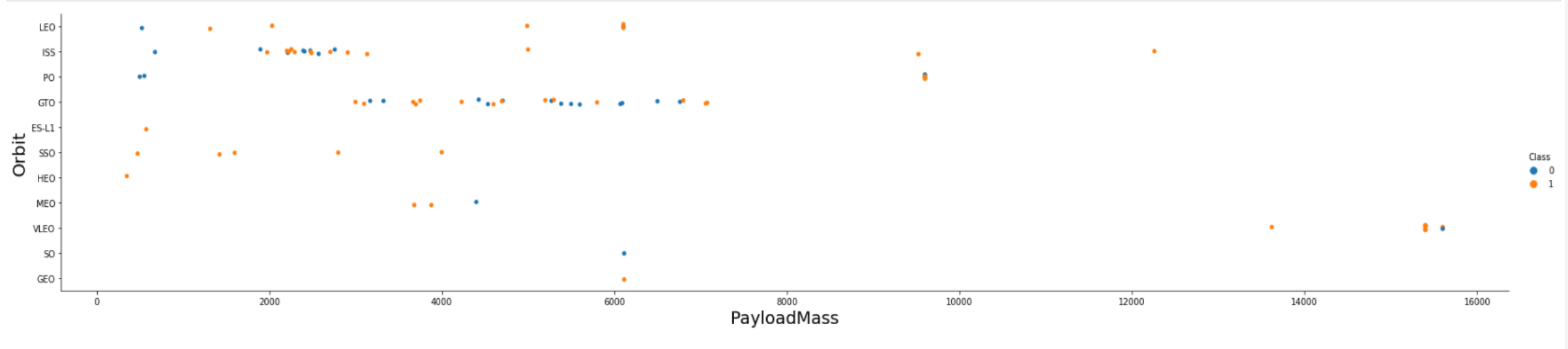
Flight Number vs. Orbit Type

- The flights to ISS VLEO and LEO mostly had successful landing. Moreover, as more flights in the orbit as more successful landing it gets.



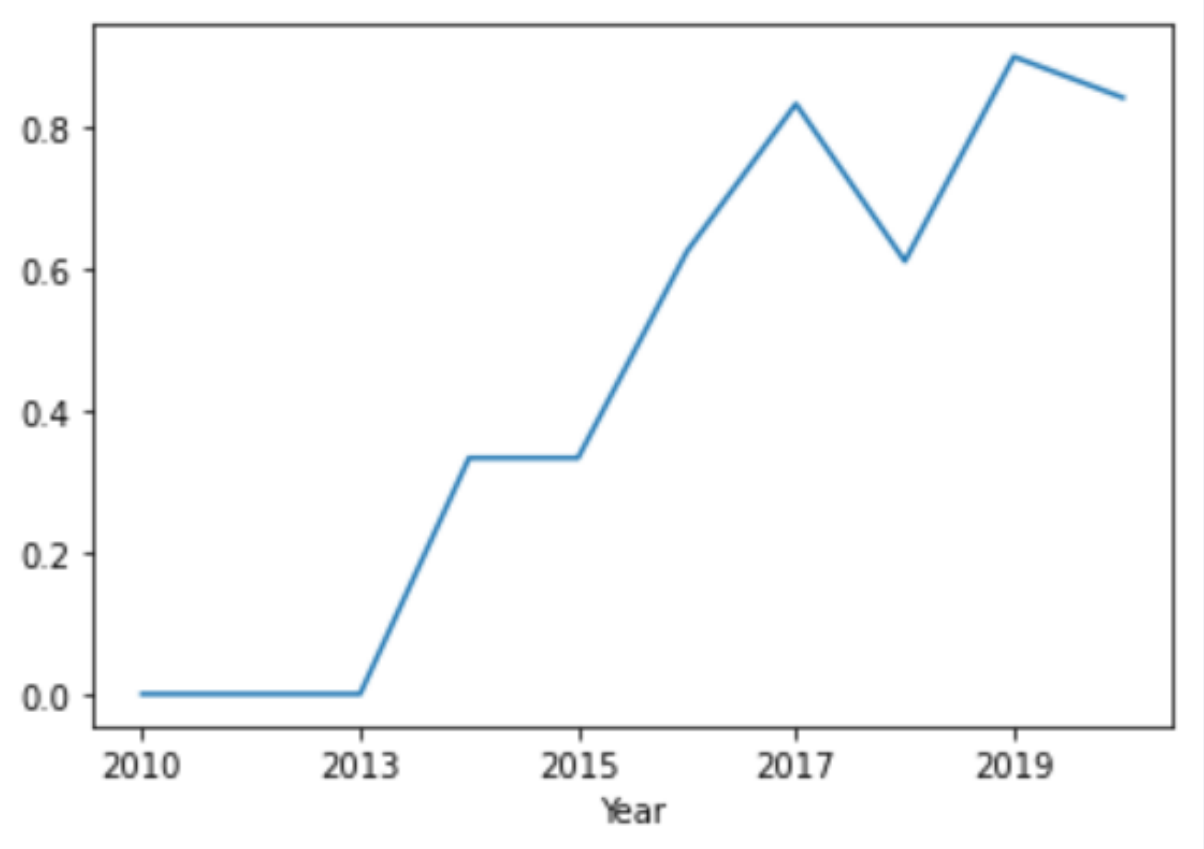
Payload vs. Orbit Type

- Most of the launches to the orbits are below the Payload Mass of 8000. Launches to GTO and ISS mostly had not landed successfully.



Launch Success Yearly Trend

- Each year successful landings increased. In 2019 success rate was almost 100%



All Launch Site Names

- There are 4 launch sites were used to launch the rockets

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

Launch Site Names Begin with 'CCA'

- 5 Launch sites that begin with `CCA`

```
sql SELECT * FROM SPACEXTBL 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

- In the below photo you can see the total payload for NASA(CRS) is 48213

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE Customer  LIKE '%NASA%CRS%'
```

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

TOTAL_PAYLOAD
48213

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

  is 2928 kg.

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- The first successful landing was on 22nd of December in 2015

```
%sql SELECT MIN(Date) AS FIRST_SUCCESS FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%(ground pad)%'
```

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

FIRST_SUCCESS
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster versions which were successful in drone ship landing are F9 FT group

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_
BETWEEN 4000 AND 6000 AND Landing_outcome = 'Success (drone ship)'
```

```
 * 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

- The number of Successful mission outcomes are 100, while failure outcomes are 1

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY Mission_Outcome ORDER BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- Booster Version from F9 B5 type have highest payload of 15600kg

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION
```

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

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

2015 Launch Records

- Failed landing outcomes in drone ship with their booster versions, and launch site names for year 2015

```
%sql SELECT Booster_version, Launch_Site FROM SPACEXTBL
WHERE Landing_outcome = "Failure (drone ship)" AND Date LIKE '%2015%'
```

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

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

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

- Ranking 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(*) AS QTY FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_outcome ORDER BY QTY DESC;
```

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

Landing_Outcome	QTY
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

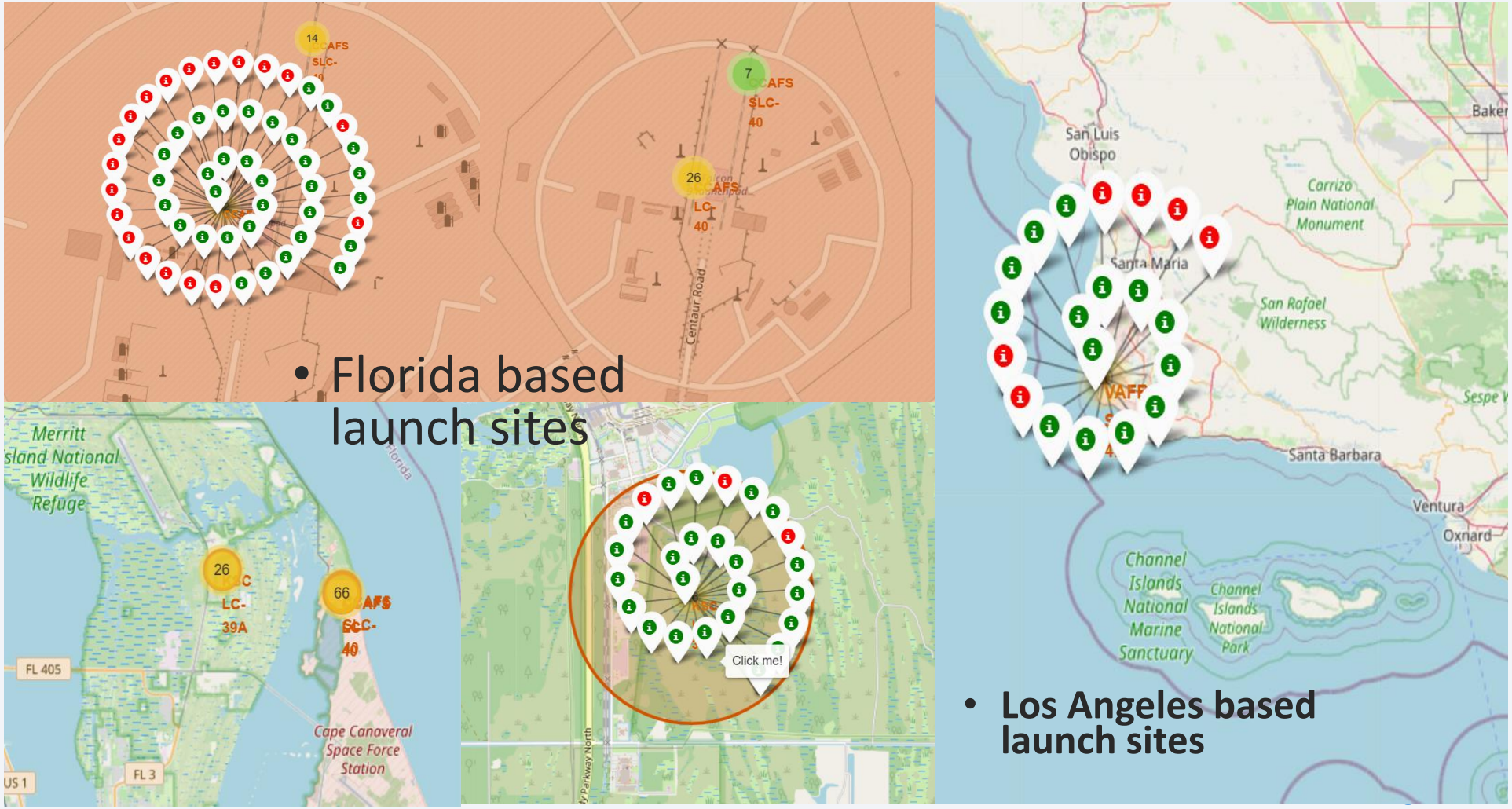
Location of the SpaceX Launch Sites

- SpaceX launch sites are located in USA, near the coastline.

Landing Outcome Indication with markers

- Green marker indicates successful landing, while red marker indicates the opposite

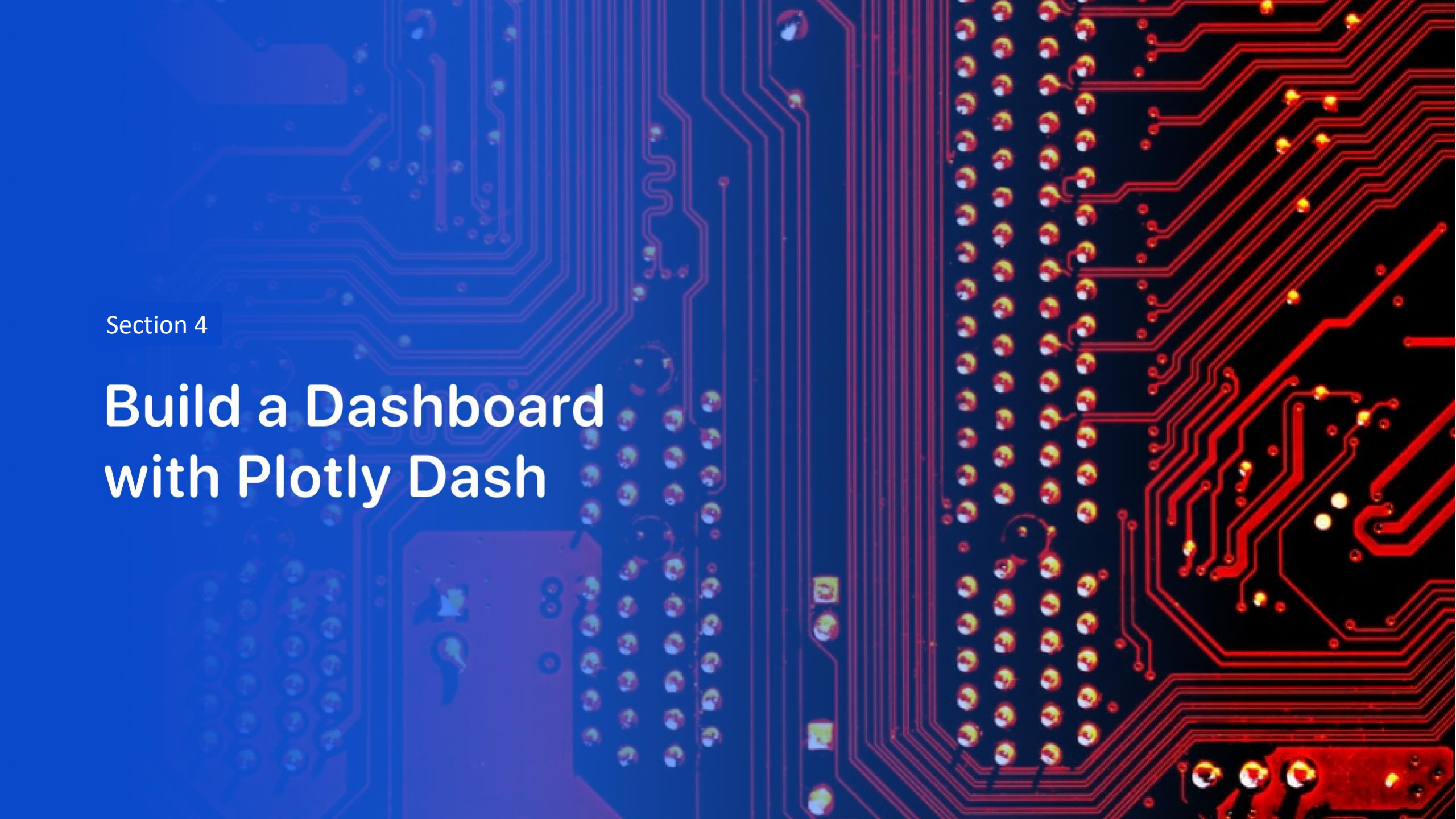
- There are 92 launches in Florida and 20 in Los Angeles



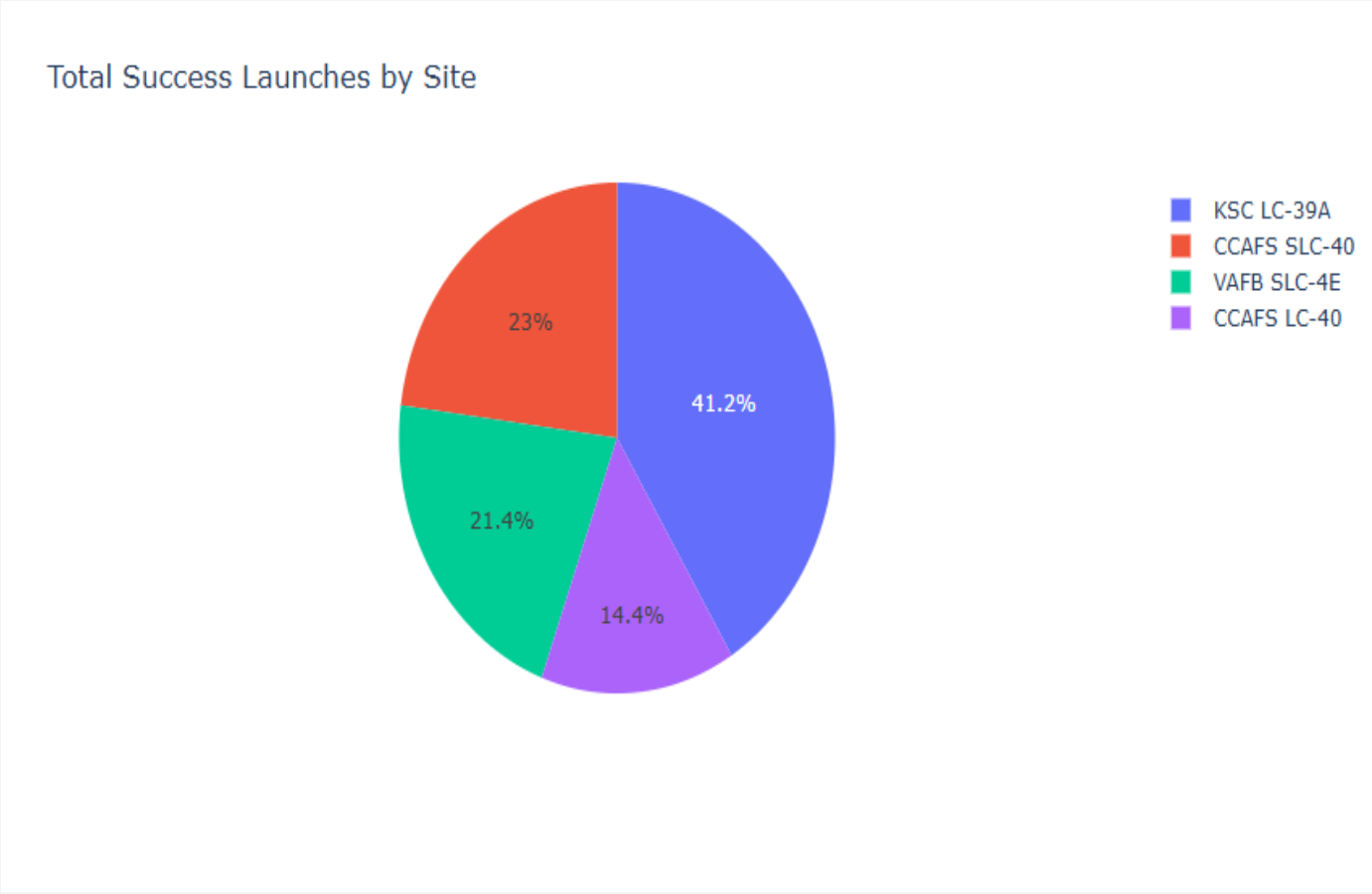
The distance between Coastline and Rail Road

Launch Sites are away from cities.

The distance between launch sites and coastline as well as Rail Road are within 1km

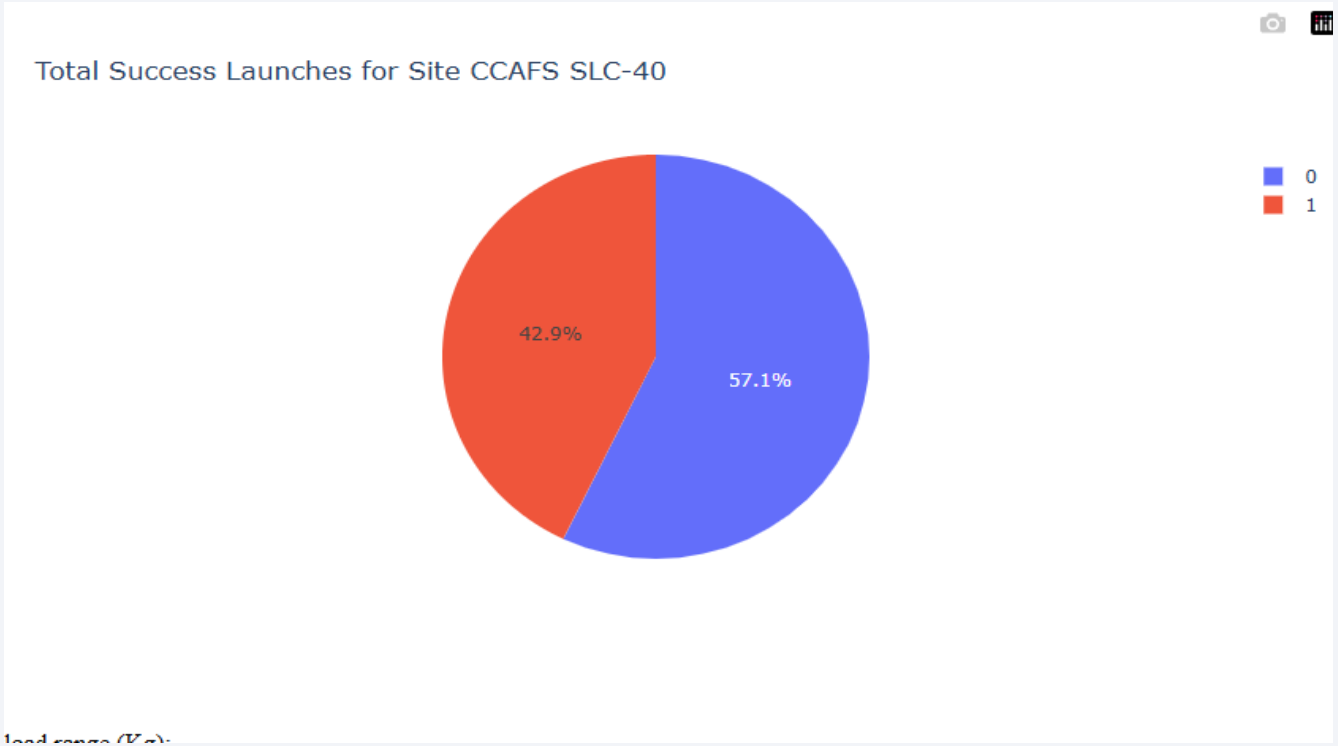

Section 4
Build a Dashboard with Plotly Dash

Successful landing percentage of total by each Launch Site



As illustrated in the pie chart, KSC LC-39A had
the largest proportion (41%) of successful landing among
other launch sites.

Pie Chart of the largest success to failure ratio launch Site



Launch Site CCAFS SLC-40 had the more Successful landing results compare to unsuccessful
of 57%

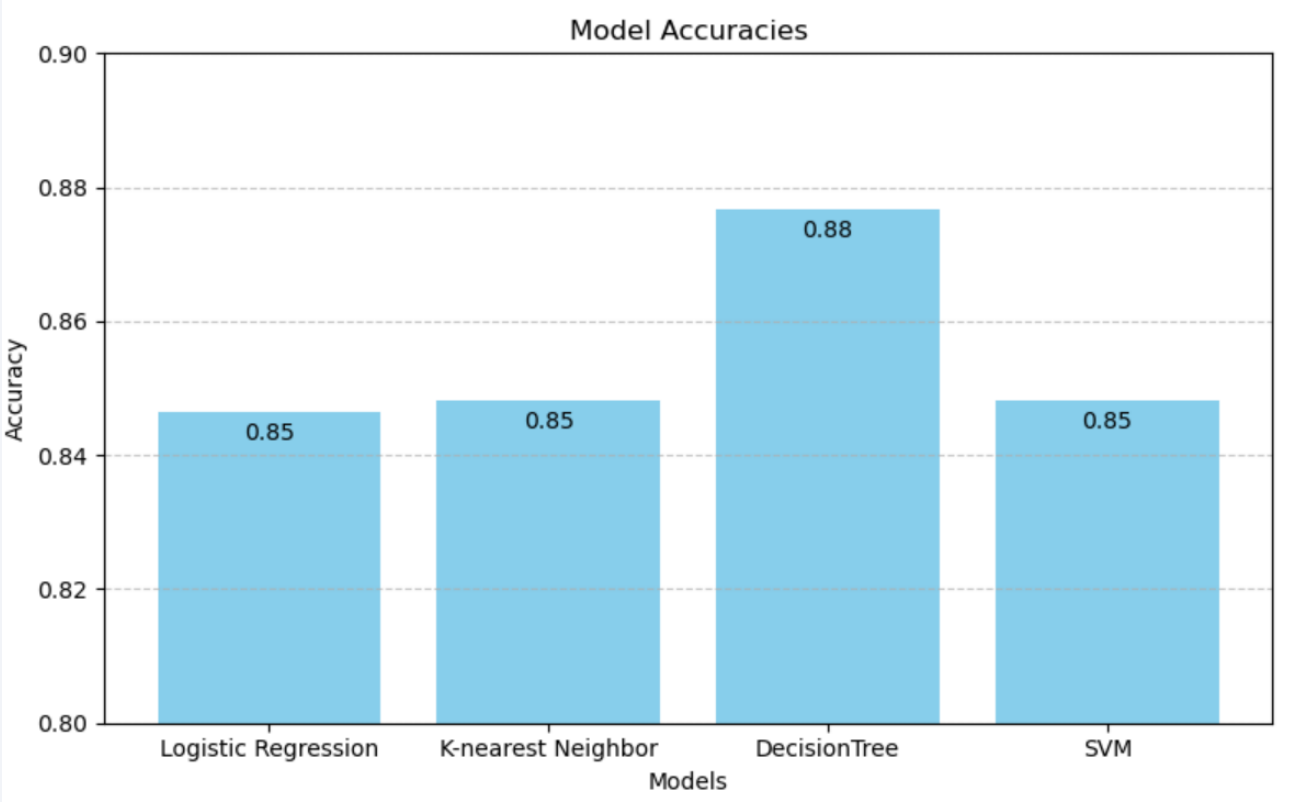
Relationships between Payload vs Landing Outcome

Section 5

Predictive Analysis (Classification)

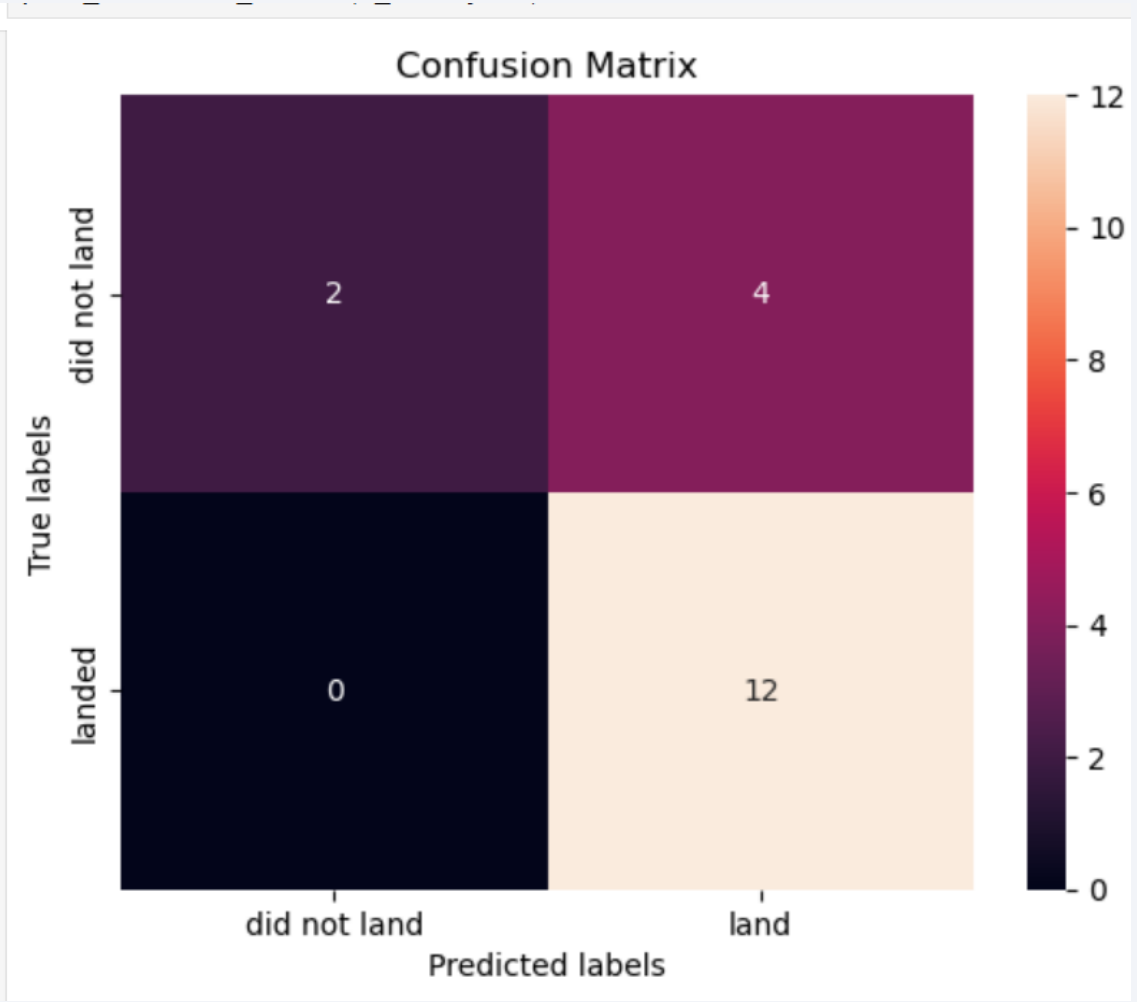
Classification Accuracy

- The chart illustrates that Decision Tree classifier has the highest accuracy of 88% and performs best among all of the models



Confusion Matrix

- This illustration is the confusion matrix of Decision Tree Classifier. The problem is the False Positive. It means model outputted 4 launches as landed while in reality it did not



Conclusions

In conclusion we can say:

The more flights at a launch site, the greater the success rate at that launch site.

Launch success rate started to increase in 2013 till 2020.

Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

KSC LC-39A had the most successful launches of any sites.

Payload mass heavier than 8000 guarantees successful landing

The Decision Tree Classifier is the most optimal model for classifying into classes (Success & Fail)

Thank you!