IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Riddhi Ladhad
26th July, 2024

Outline

Executive Summary

Summary of methodologies

- **Data Collection** through API & Web Scraping
- **Data Wrangling**
- **Exploratory Data Analysis (EDA)** with SQL & Data Visualization
- **Interactive Visual Analytics** with Folium
- **Interactive Dashboard** with Plotly Dash
- **Predictive Analysis** with Machine Learning

Summary of all results

- Exploratory Data Analysis results
- Interactive Analytics demonstration
- Predictive Analysis results

Introduction

Project background and context

- SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. Instead of using rocket science to determine if the first stage will land successfully, we will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Problems you want to find answers

- Whether or not SpaceX will reuse the first stage?
- Which factors determine the successful landing of Falcon 9?
- How do various features that determine the success rate of landing interact amongst each other?

Section 1

Methodology

Methodology

- Data collection methodology:
  - Collected the Falcon 9 launch data from SpaceX REST API and through web scraping the related wiki pages.
- Perform data wrangling:
  - Converted categorical variables to numerical values by One-hot coding/labelling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split the transformed dataset into training & testing datasets, built ML models using classification models, tuned the hyperparameters & evaluated the models.

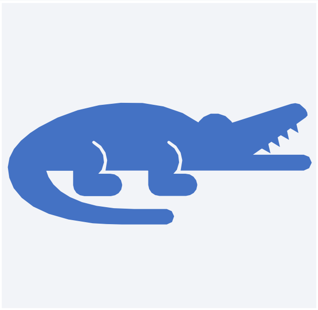
Data Collection

From SpaceX REST API

Used the requests.get() method to request the data from the SpaceX API

Decoded the response content as a Json using .json() method and converted it into a pandas data frame using the .json_normalize() method

Performed initial analysis on the data, checked for missing values and filled in missing values with the average value of that particular feature.



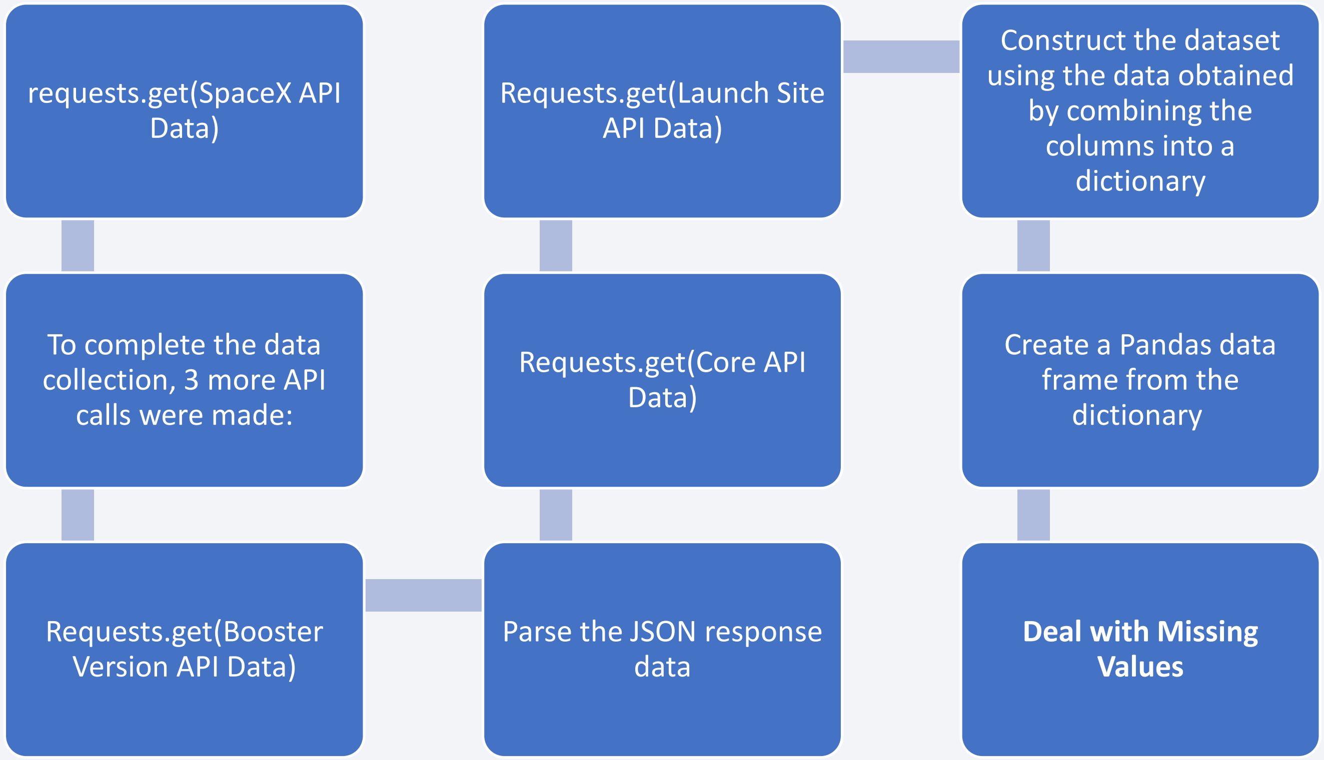
Through Web Scraping

Performed web scraping on Wikipedia for Falcon 9 launch records.

Used the BeautifulSoup to parse the HTML table and convert it to a pandas data frame for further analysis.

Data Collection – SpaceX API

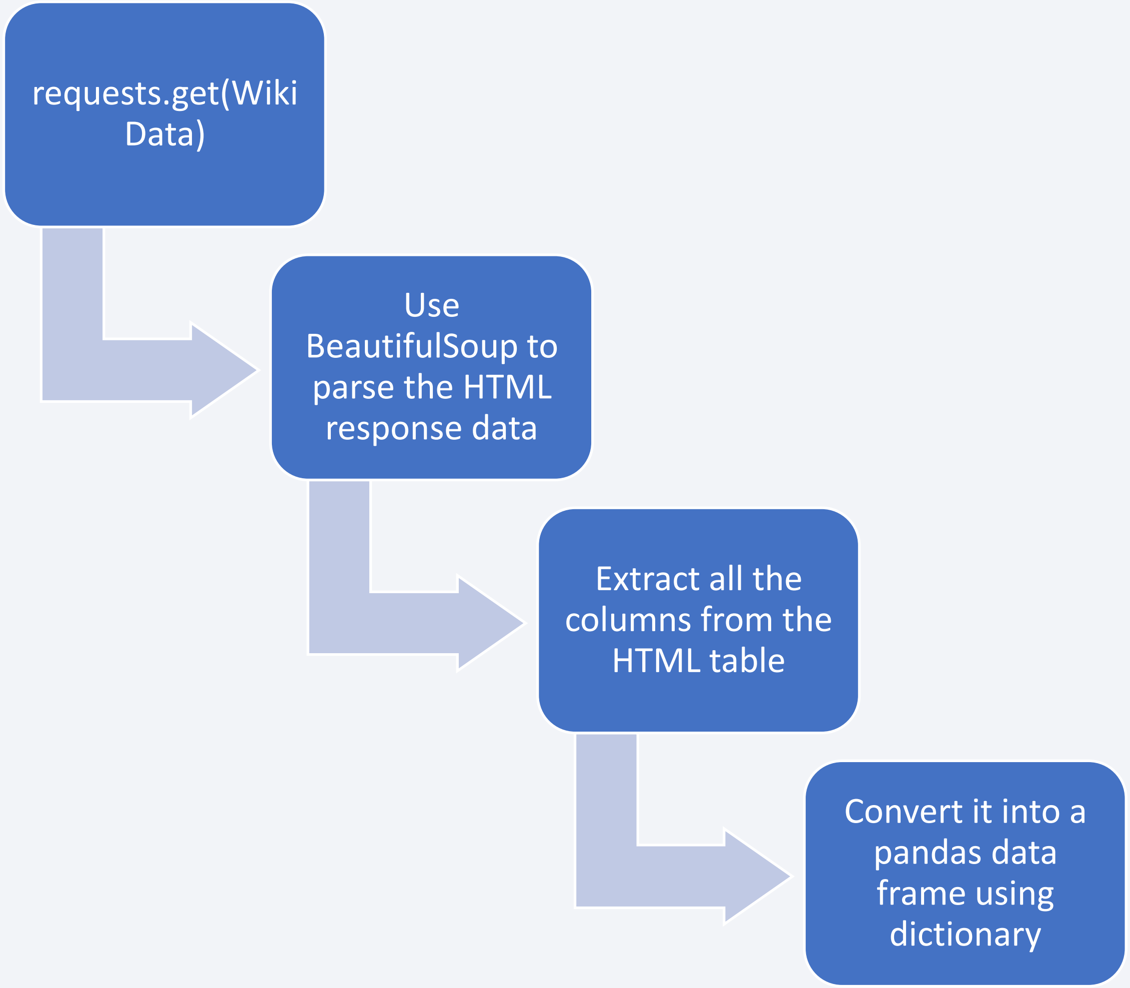
- Used requests.get() to the SpaceX API to collect data, parsed the JSON response and did some basic data wrangling and formatting
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/Riddhi-ladhad/DataScienceCoursera/blob/main/SpaceX/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scraped Falcon 9 launch records HTML table from Wikipedia & parsed the table to convert it into a Pandas data frame

- GitHub URL of the completed web scraping notebook: https://github.com/Riddhi-ladhad/DataScienceCoursera/blob/main/SpaceX/jupyter-labs-webscraping.ipynb

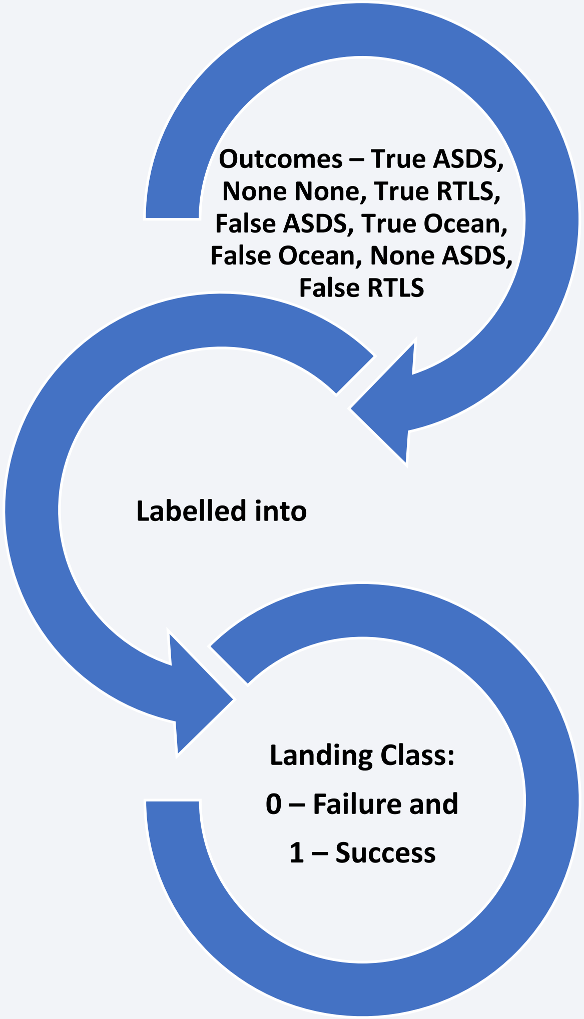


Data Wrangling

- Performed Exploratory Data Analysis & determined training labels.

- Created landing outcome labels from the Outcome column in the data frame.

- GitHub URL of the completed data wrangling notebook: https://github.com/Riddhi-ladhad/DataScienceCoursera/blob/main/SpaceX/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Performed Exploratory data Analysis by plotting various graphs

Used scatter plots to visualize how the following parameters affect the launch success rate:

- Flight Number & Payload Mass
- Flight Number & Launch Sites
- Payload Mass & Launch Sites
- Flight Number & Orbit Types
- Payload Mass & Orbit Types

Used bar chart to visualize the success rate of each orbit

Used the line chart to visualize the yearly launch success trend

GitHub URL of the completed EDA with data visualization notebook: https://github.com/Riddhi-ladhad/DataScienceCoursera/blob/main/SpaceX/edadataviz.ipynb

EDA with SQL

Loaded the SpaceX dataset into a SQLite3 database & queried the DB to get the following insights from the data:

Names of the unique launch sites in the space mission

Total payload mass carried by boosters launched by NASA (CRS)

Average payload mass carried by booster version F9 v1.1

Date of the first successful landing outcome in ground pad

Total number of successful and failure mission outcomes

Booster Versions that have carried the maximum payload mass

Failed landing outcomes in drone ship, their Booster Versions and Launch Sites from the year 2015

Frequency of the Failure (drone ship) & Success (ground pad) landing outcomes

GitHub URL of the completed EDA with SQL notebook:
https://github.com/Riddhi-ladhad/DataScienceCoursera/blob/main/SpaceX/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Marked all the launch sites, and added markers, circles to mark the success or failure of launches for each launch site on the folium map.

Used the color-labeled marker clusters, to identify which launch sites have relatively high success rate (Green for success & Red for failure)

Calculated the distances between a launch site to its proximities & displayed the distance using a line on the map. After plotting the distance lines, it is easy to identify the following:

Are launch sites in close proximity to railways/highways/coastlines?

Do launch sites keep certain distance away from cities?

GitHub URL of the completed interactive map with Folium map: https://github.com/Riddhi-ladhad/DataScienceCoursera/blob/main/SpaceX/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Added a dropdown populated with all the Launch Sites and used a pie chart to visualize the total successful launches for:

all Launch Sites

a specific Launch Site

Added a range slider to select a Payload range & used a scatter plot to visualize how Payload & Mission Outcomes are correlated with different Booster Versions for:

all Launch Sites

a specific Launch Site

GitHub URL of the completed Plotly Dash lab: https://github.com/Riddhi-ladhad/DataScienceCoursera/blob/main/SpaceX/spacex_dash_app.py

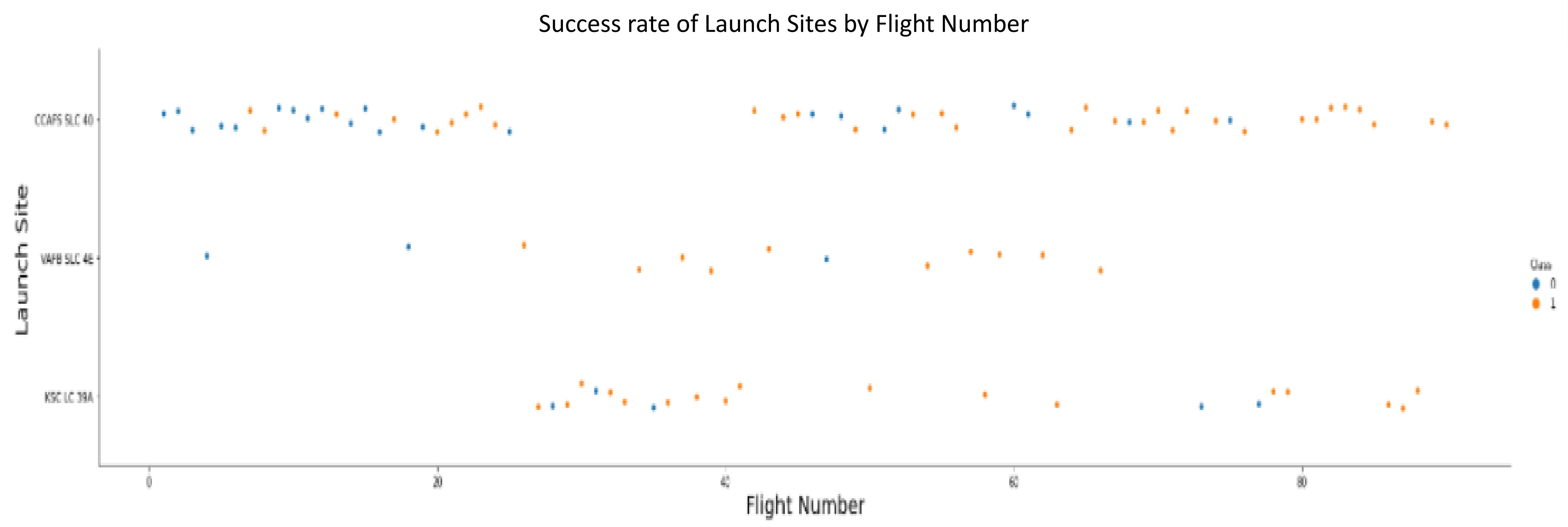
Predictive Analysis (Classification)

- Determined training labels for landing outcomes, standardized & transformed the dataset
- Split the dataset into training and test datasets
- Trained the dataset using Logistic Regression, Support Vector Machine (SVM), Decision Tree and k-Nearest Neighbors classification models
- Fit the models to find the best parameters using GridSearchCV
- Calculated accuracy of all the models on training & testing datasets and plotted the confusion matrix for all the models
- Found the best performing classification model
- GitHub URL of the completed predictive analysis lab: https://github.com/Riddhi-ladhad/DataScienceCoursera/blob/main/SpaceX/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

Section 2

Insights drawn from EDA

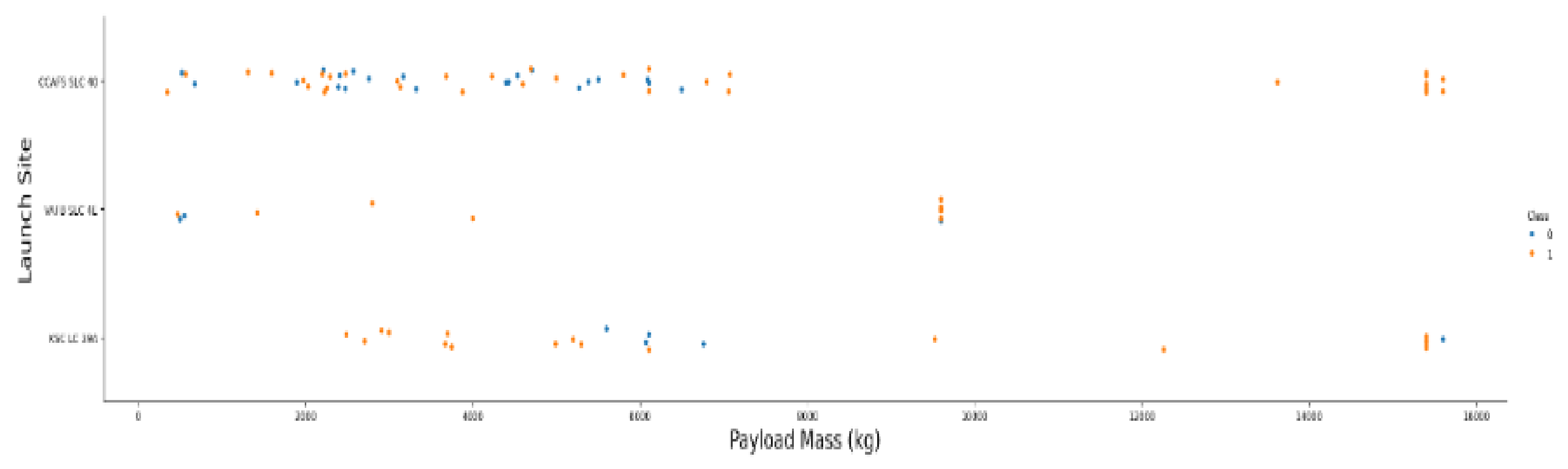
Flight Number vs. Launch Site



- Highest rockets were launched from the CCAFS SLC 40 Launch Site
- VAFB SLC 4E & KSC LC 39A Launch Sites have higher launching success rates
- Earlier flights have lower success rates than the later flights

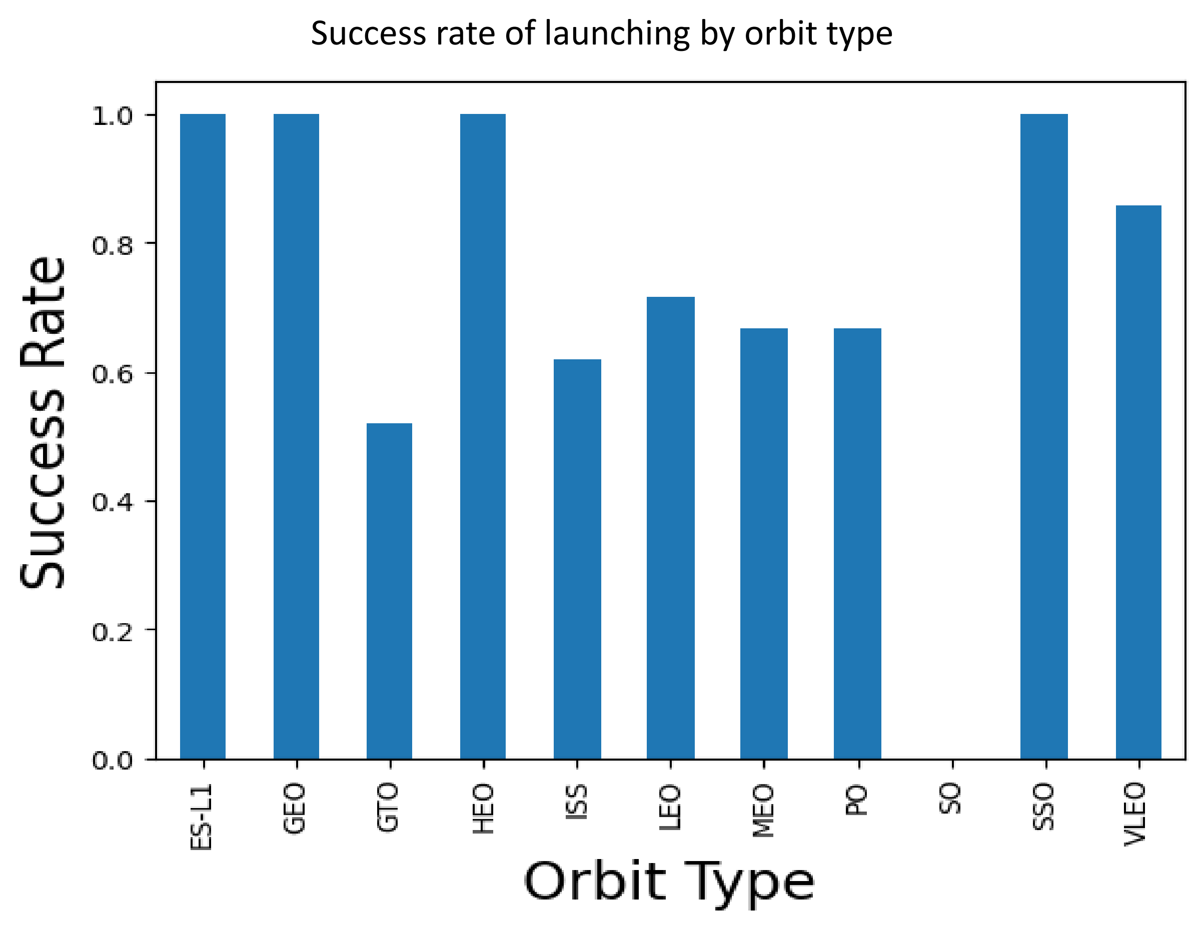
Payload vs. Launch Site

Success rate of Launch Sites by Payload Mass



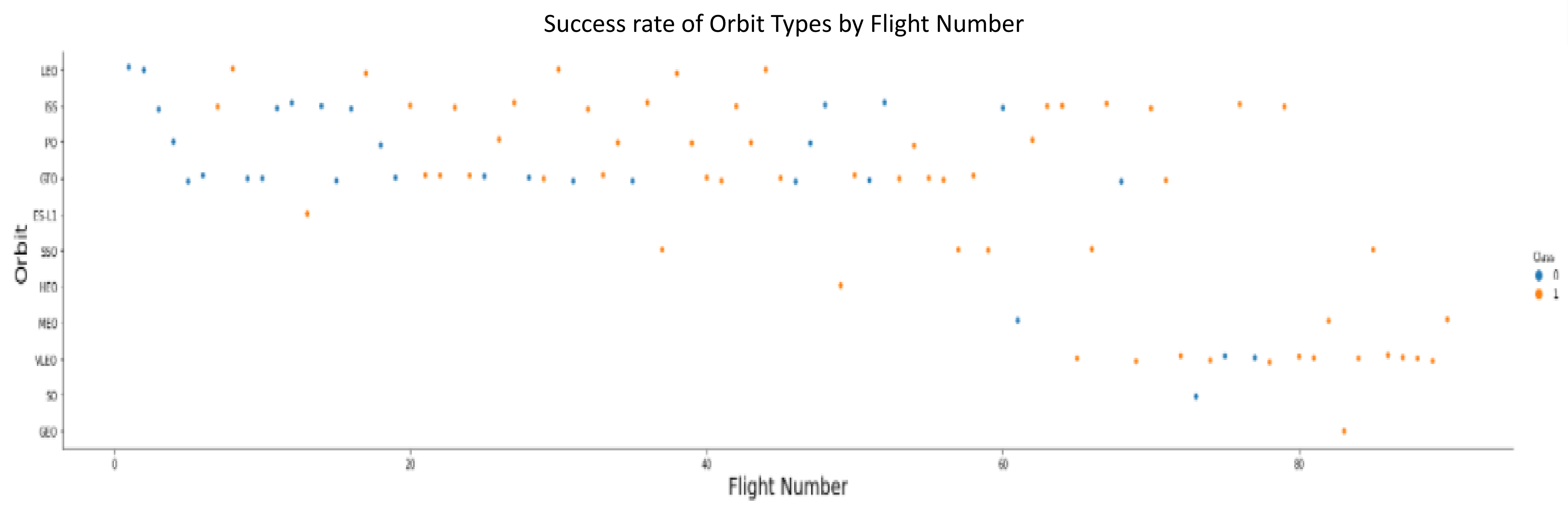
- For the VAFB SLC 4E Launch Site, there are no rockets launched for heavy Payload Mass (greater than 10,000 kg)
- Rockets with Payload Mass greater than 6,000 kg have higher success rates
- KSC LC 39A Launch Site has failed launches for Payload Mass between 5,500 kg & 7,500 kg

Success Rate vs. Orbit Type



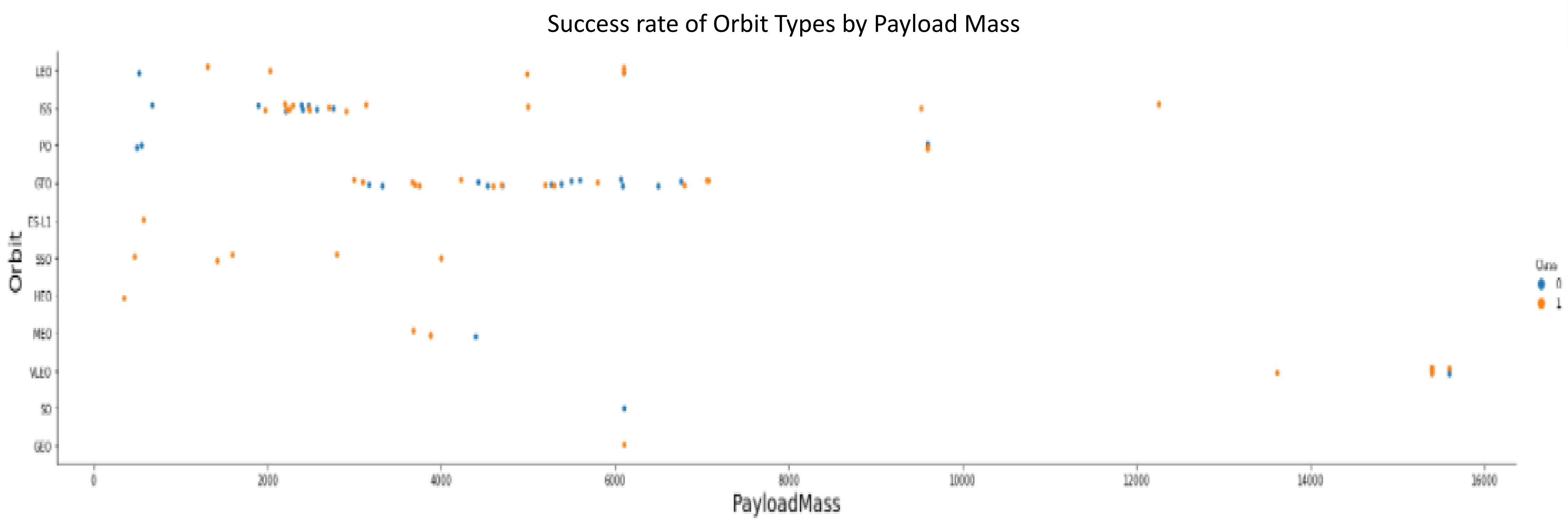
- Launch Sites ES-L1, GEO, HEO and SSO have 100% success rate
- SO Launch Site has 0% success rate
- Rest of the Launch Sites – GTO, ISS, LEO, MEO, PO and VLEO have 50% – 80% success rate

Flight Number vs. Orbit Type



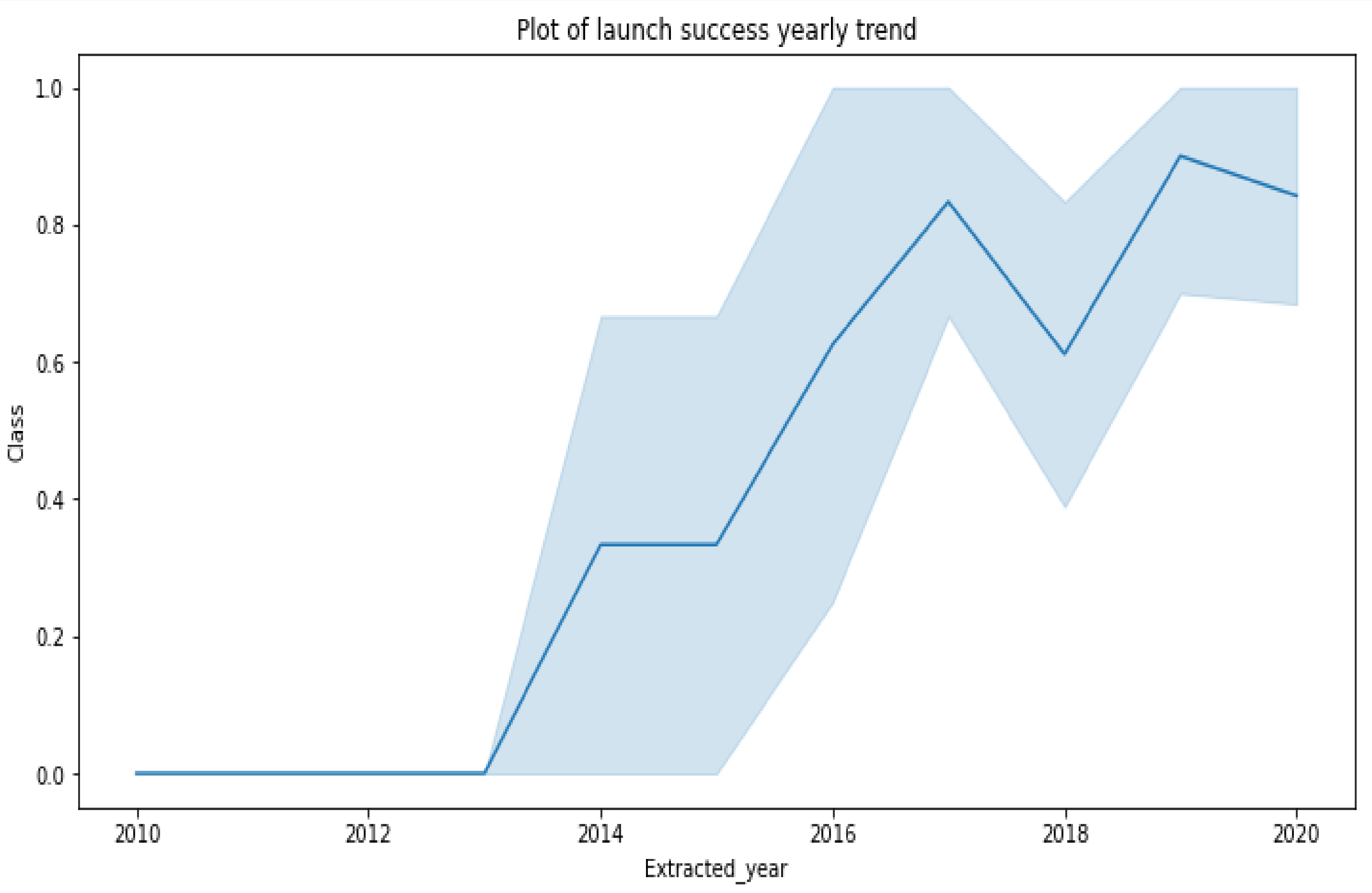
- Success rate increases with the no. of flights for each orbit
- However, the same is not true for the GTO orbit

Payload vs. Orbit Type



- Orbits LEO, SSO, HEO, MEO, SO & GEO have not launched rockets with Payload Mass greater than 7,000 kg
- SSO, HEO, GEO and ES-L1 orbits have 100% success rate
- Orbits ISS & VLEO have higher success rate for higher Payload Mass (greater than 8,000 kg)

Launch Success Yearly Trend



- Success rate increased from 2013-2017 and then from 2018-2019
- 2019 being the year during which rockets were launched most successfully
- Overall success rate improved since 2013

Launch Site Names

Names of all the Launch Sites

```
results = pd.read_sql('select distinct "Launch_Site" from spacextable;', con)
results
```

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Names of all the Launch Sites that begin with 'CCA'

```
results = pd.read_sql('select * from spacextable where "Launch_Site" like "CCA%" limit 5;', con)
results
```

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcor
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachu
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachu
2	2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
3	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
4	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Payload Mass

Total Payload Mass carried by NASA (CRS)

```
results = pd.read_sql('select sum(payload_mass__kg_) from spacextable where customer = "NASA (CRS)";', con)
results
```

	sum(payload_mass__kg_)
0	45596

Average Payload Mass carried by booster Version F9 v1.1

```
results = pd.read_sql('select avg(payload_mass__kg_) from spacextable where booster_version like "F9 v1.1%";', con)
results
```

	avg(payload_mass__kg_)
0	2534.666667

Successful Landing

First Successful Ground Landing Date

```
results = pd.read_sql('select min(date) from spacextable where landing_outcome = "Success (ground pad)";', con)
results
```

	min(date)
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
results = pd.read_sql('select booster_version from spacextable where landing_outcome = "Success (drone ship)" and payload_ma
results
```

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failure Mission Outcomes

```
results = pd.read_sql('select mission_outcome, count(mission_outcome) from spacextable group by mission_outcome;', con)
results
```

	Mission_Outcome	count(mission_outcome)
0	Failure (in flight)	1
1	Success	98
2	Success	1
3	Success (payload status unclear)	1

Failed Landing Outcomes in drone ship, their Booster Versions and Launch Site names in 2015

```
results = pd.read_sql('select substr(Date, 6,2) as month, landing_outcome, booster_version, launch_site from spacextable whe
results
```

	month	Landing_Outcome	Booster_Version	Launch_Site
0	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Boosters Carried Maximum Payload

```
results = pd.read_sql('select booster_version from spacextable where payload_mass__kg_ = (select max(payload_mass__kg_) fro
results
```

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

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

```
results = pd.read_sql('select landing_outcome, count(landing_outcome) from spacextable where date between "2010-06-04" and '
results
```

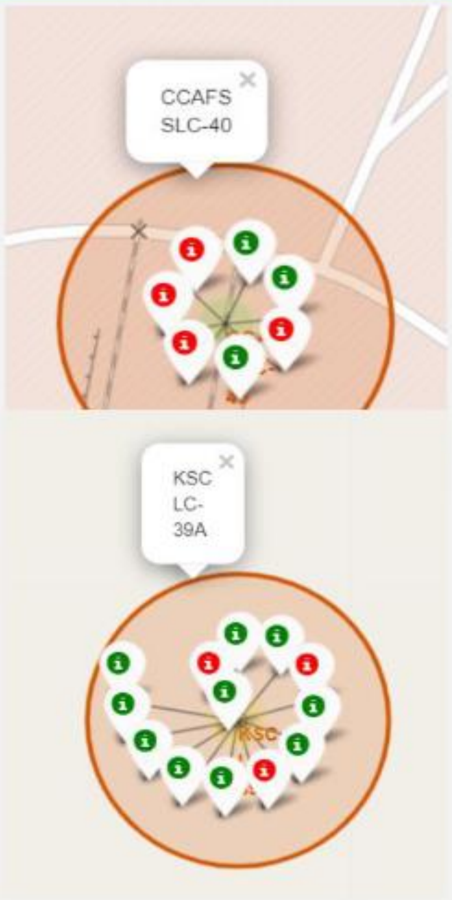
	Landing_Outcome	count(landing_outcome)
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	2
7	Precluded (drone ship)	1

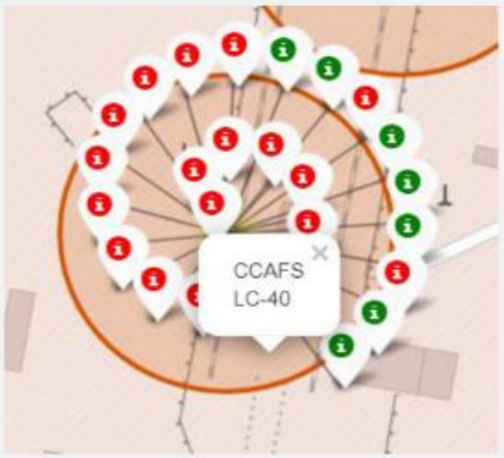
Section 3

Launch Sites Proximities Analysis

All Launch Sites' location markers on a global map

Color-labeled Launching Outcomes on the map





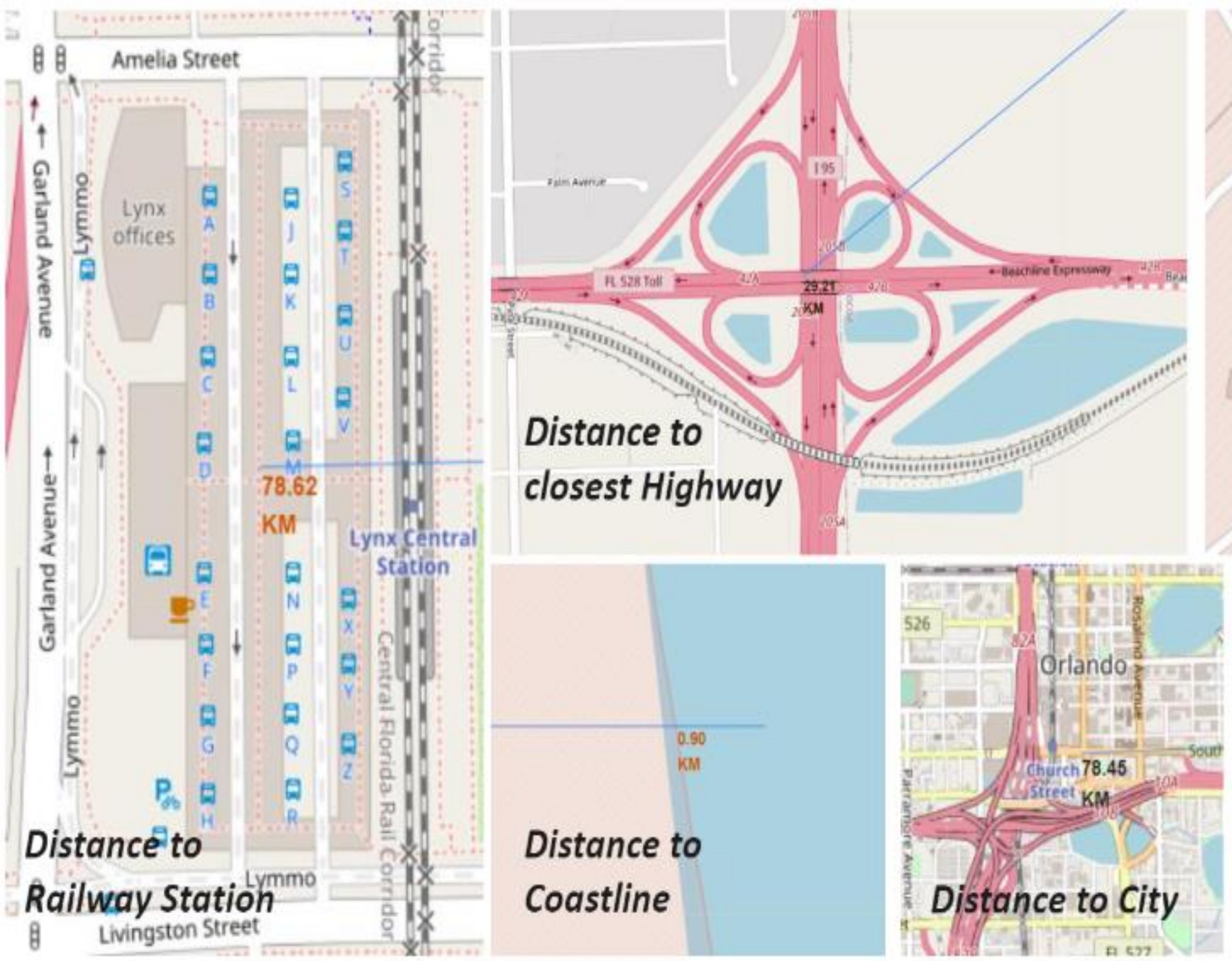
Florida Launch Sites

Green Marker shows successful Launches and Red Marker shows Failures

California Launch Site

37

Launch Site distance to its proximities

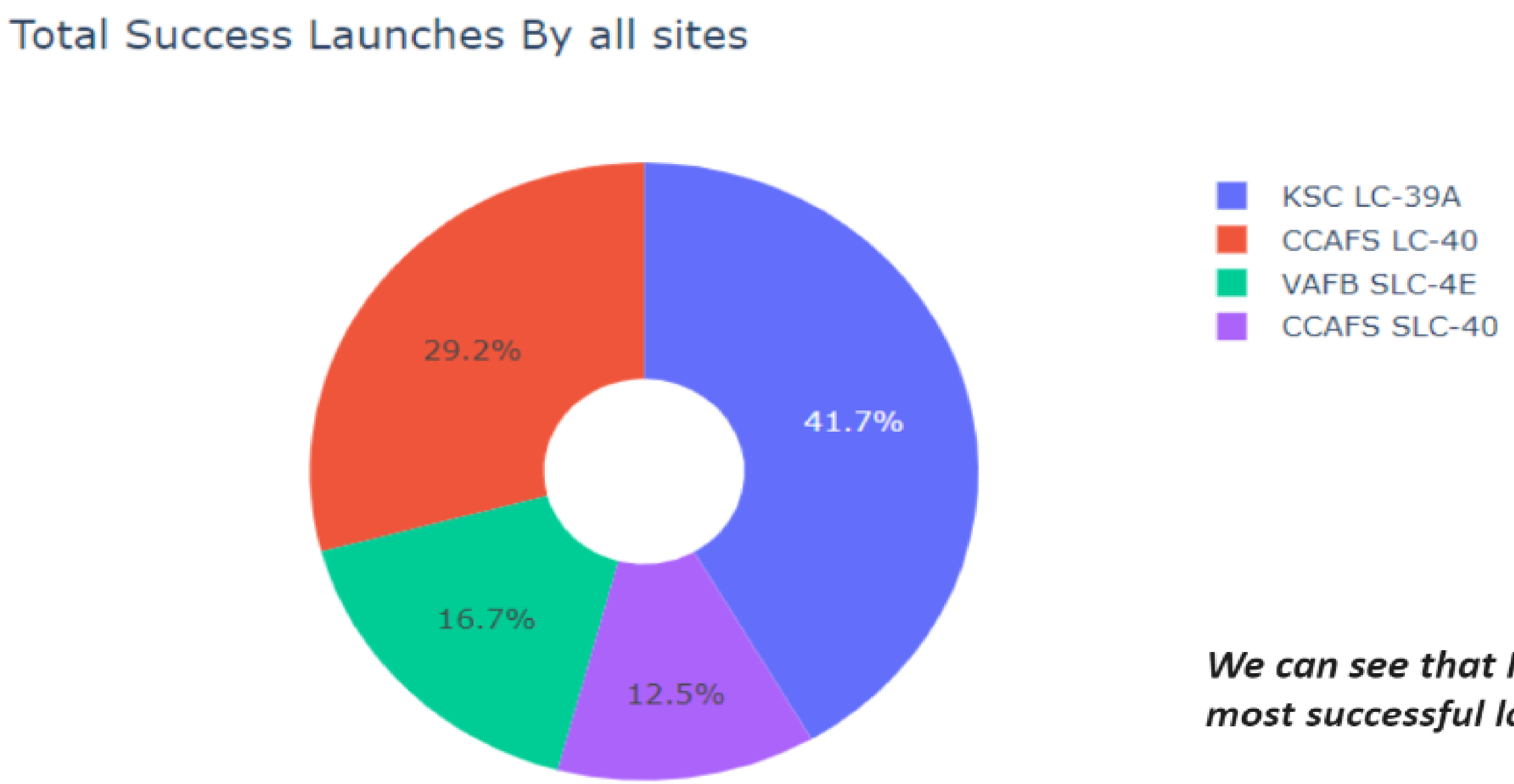


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 4

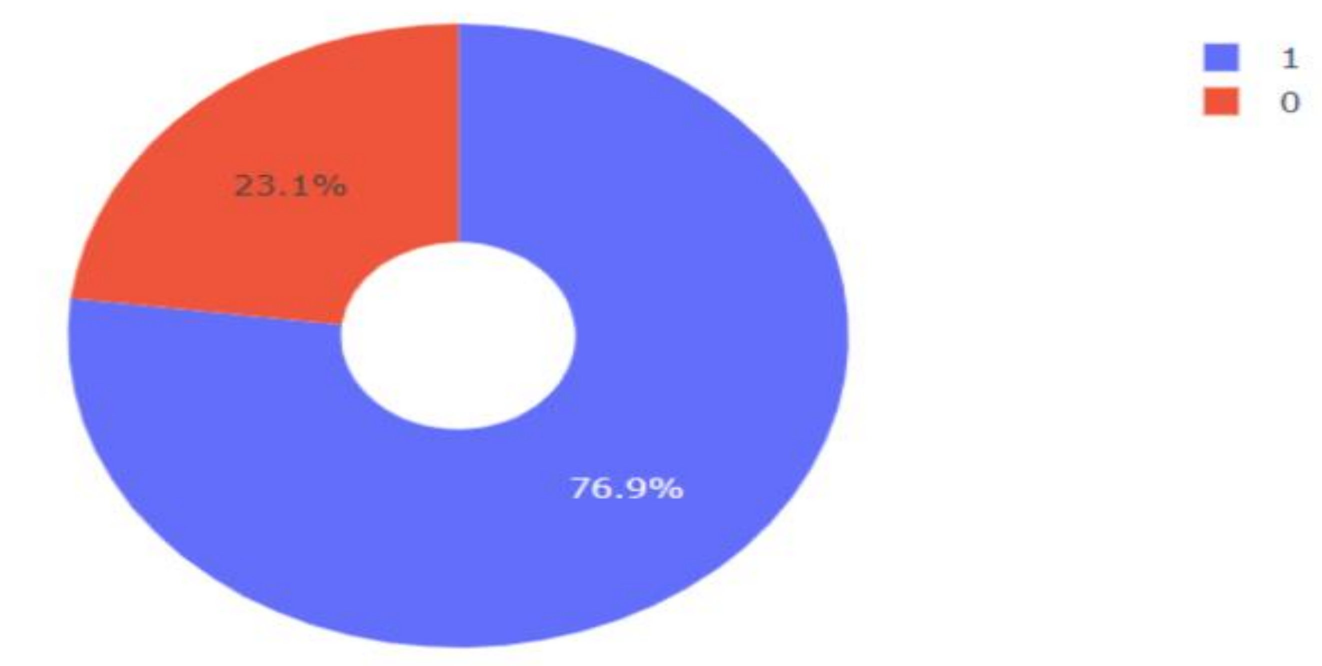
Build a Dashboard with Plotly Dash

Launch success by Sites



We can see that KSC LC-39A had the most successful launches from all the sites

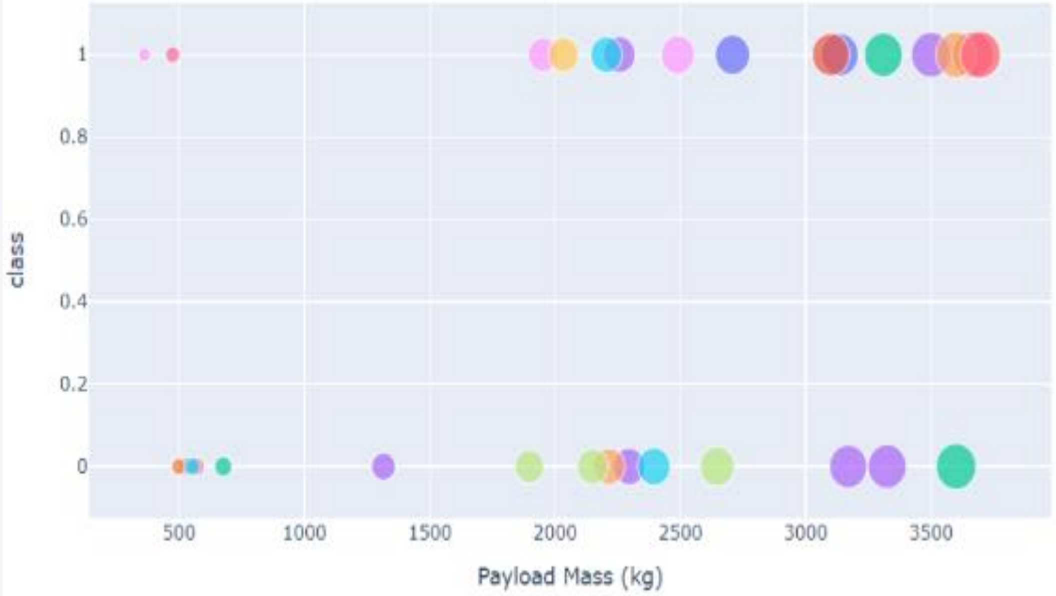
Success rate of KSC LC 39A Launch Site



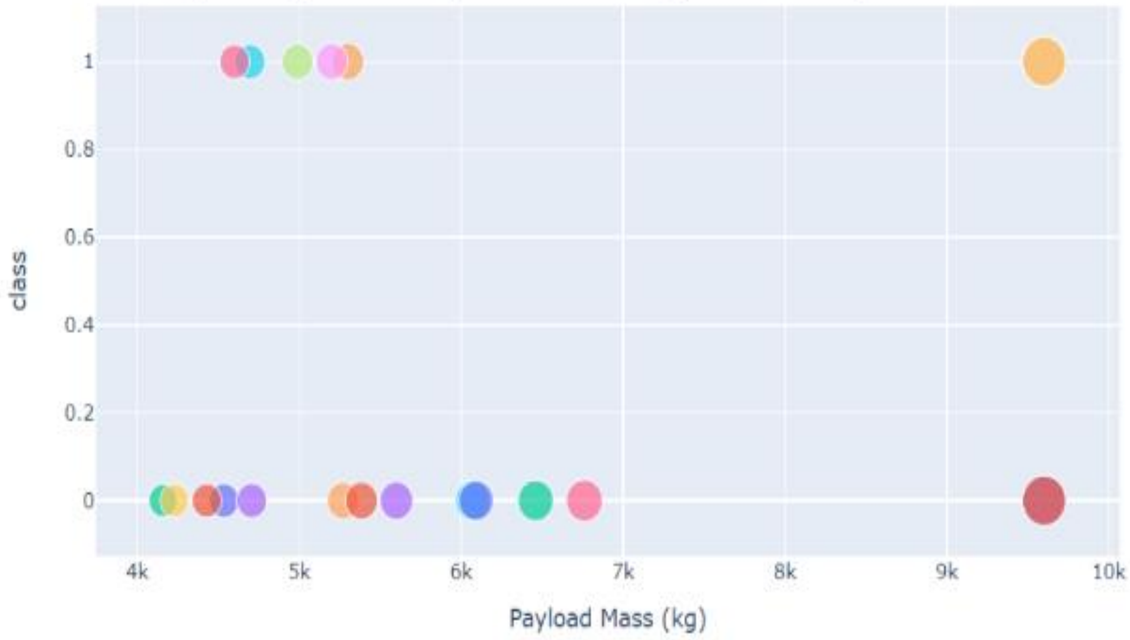
KSC LC 39A has had 10 successful launches & 3 failed launches

Payload Mass & Success

Low Weighted Payload 0kg – 4000kg



Heavy Weighted Payload 4000kg – 10000kg

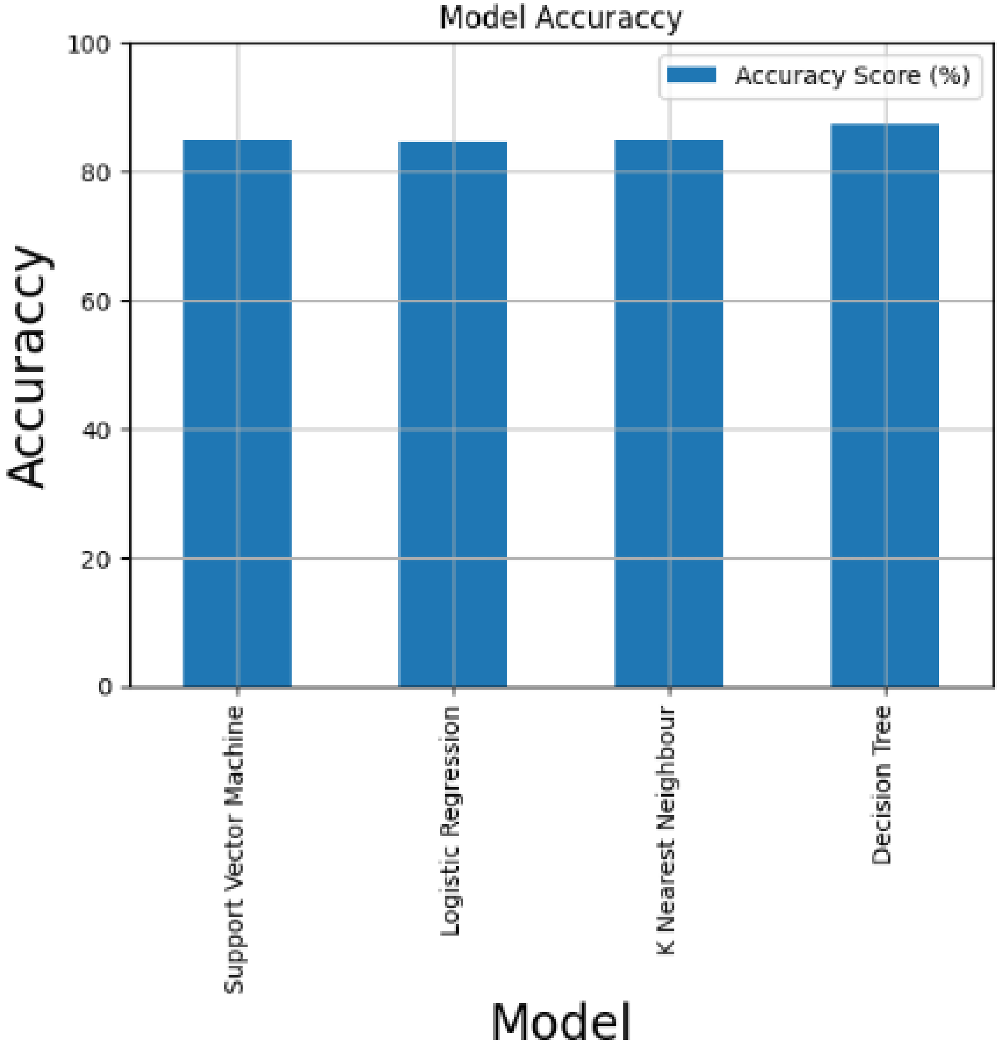


We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

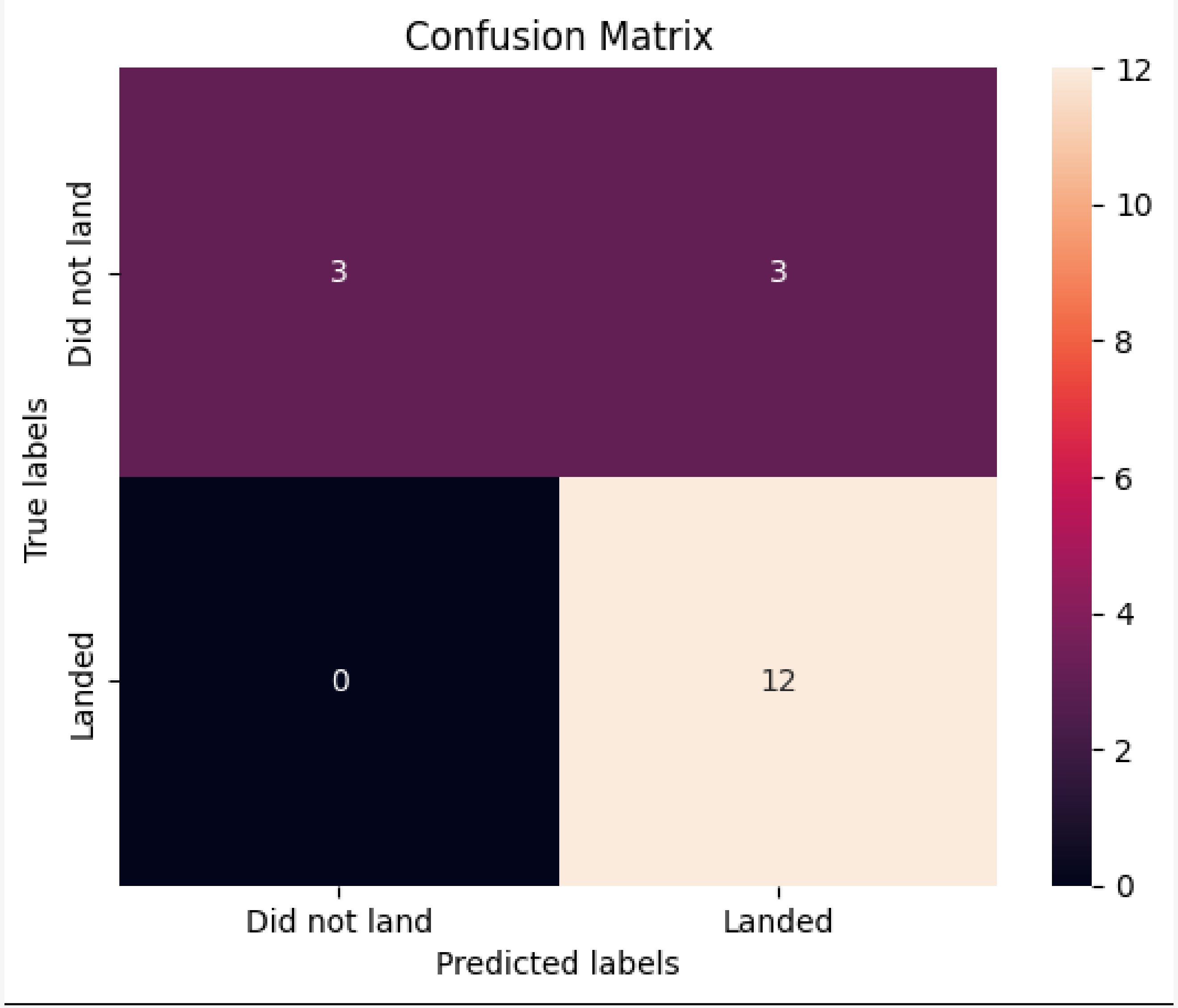
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree Classification model is the best as it has the highest accuracy

	Accuracy Score (%)
Support Vector Machine	84.82
Logistic Regression	84.64
K Nearest Neighbour	84.82
Decision Tree	87.32



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes
- The major problem is the false positives .i.e., failed landing marked as successful landing by the classifier

Conclusions

- All the models performed similarly on the training dataset with Decision Tree model having slightly higher accuracy than the other models
- Launch success rate increases overtime since 2013
- Launch Site KSC LC 39A has the has the highest success rate. It also has 100% success rate for launches with Payload Mass less than 5,500 kg
- Orbits ES-L1, GEO, HEO & SSO have 100% launch success rate
- The higher the Payload Mass, the higher the launch success rate

Thank you!