IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Farzana Zaki
September 14, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - ❑ Data collection using REST API and web scrapping from Wikipedia
  - ❑ Data wrangling
  - ❑ Exploratory data analysis by SQL and data visualization
  - ❑ Interactive visual analytics with Folium and Plotly Dash
  - ❑ Predictive analysis by machine learning models
- Summary of all results
  - ❑ Exploratory data analysis results
  - ❑ Dashboard results
  - ❑ Machine learning prediction 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- The goal of this project is to develop a machine learning pipeline which will predict if the rocket will pass the first stage successfully.

Problem statements

- Major factors/features to determine the landing of first stage of the rocket successfully.
- The operating conditions to ensure a successful landing of the rocket launch.

Section 1

Methodology

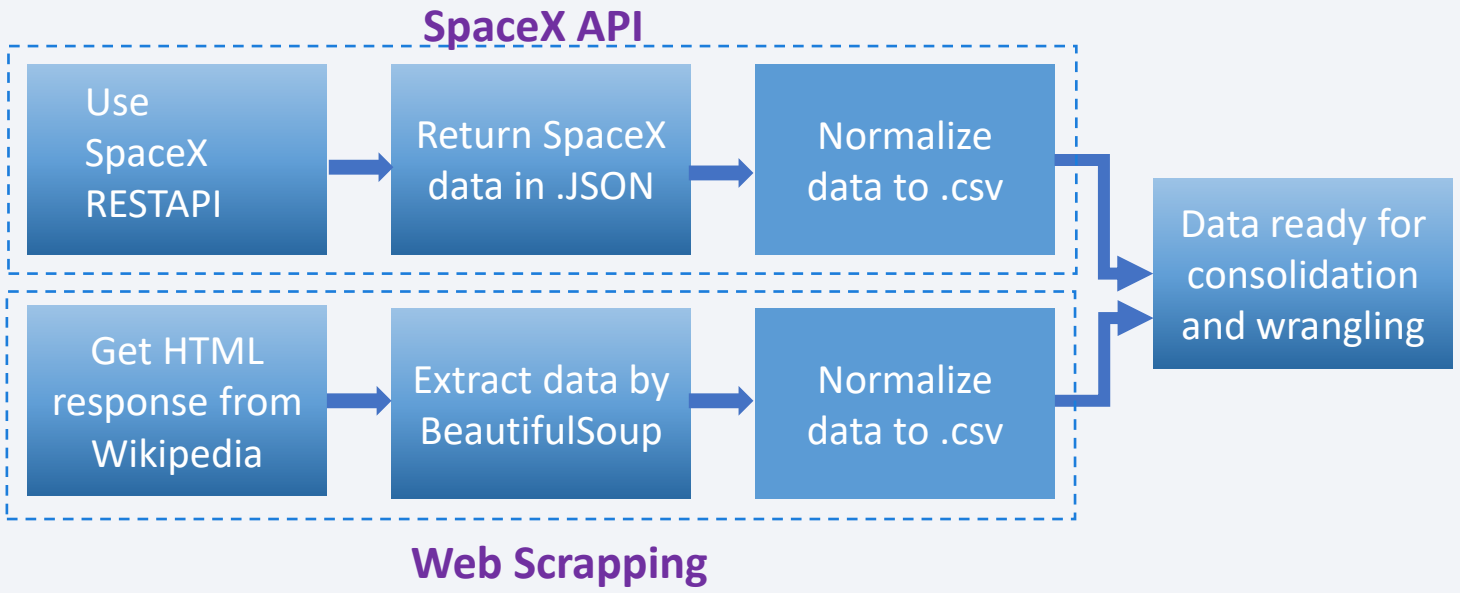
Methodology

Executive Summary

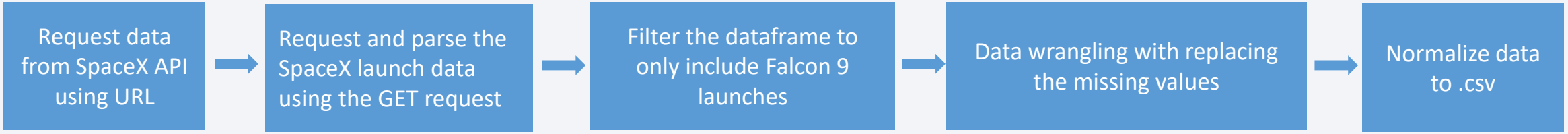
- Data collection methodology:
  - ❑ Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - ❑ One-hot encoding for categorical features of the machine learning, data cleaning of null values and irrelevant columns.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - ❑ EDA using python libraries (Matplotlib, Seaborn) and using SQL, feature engineering.
- Perform interactive visual analytics using Folium and Plotly Dash
  - ❑ Applied Folium to view previously observed correlations interactively.
- Perform predictive analysis using classification models
  - ❑ Classification models (Logistic regression, Support vector machine, K-nearest neighbor and Decision Tree) have been built and evaluated using GridSearch for the best classifier.

Data Collection

- SpaceX launch data is gathered from the SpaceX REST API. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome. The SpaceX REST API has different endpoints. We have used the endpoint api.spacexdata.com/v4/launches/past.
- In addition, web scraping from Wikipedia was performed for Falcon9 launch records using BeautifulSoup.



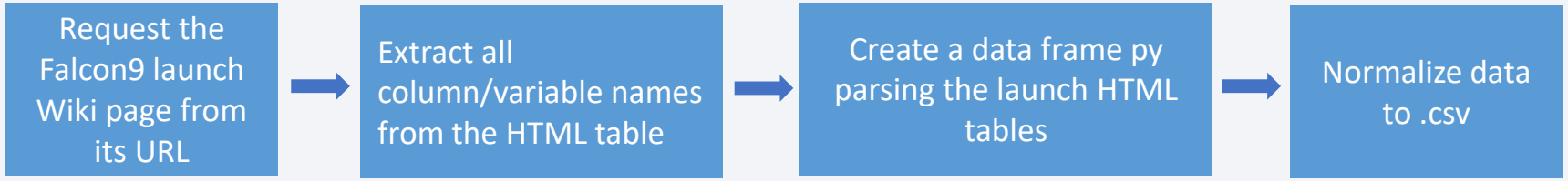
Data Collection – SpaceX API



GitHub URL of SpaceX API:

https://github.com/farzana-zaki/DataScienceProject/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



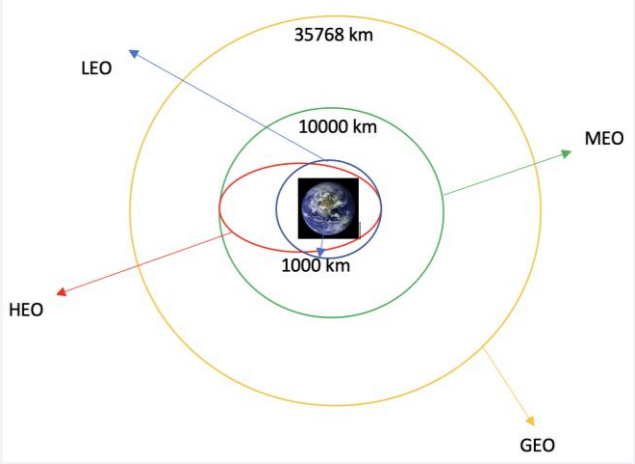
- GitHub URL of web scraping:

https://github.com/farzana-zaki/DataScienceProject/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Performed exploratory data analysis and determined the training labels.
- Calculated the number of launches at each site, and the number and occurrence of each orbits
- Created landing outcome label from outcome column and exported the results to csv.

Check the missing values → Calculate the number of launches on each site → Calculate the number of occurrence of each orbit → Calculate the number and occurrence of mission outcome per orbit type → Create a landing outcome label from Outcome column → Normalize data to .csv



- GitHub URL of data wrangling:

https://github.com/farzana-zaki/DataScienceProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- EDA was performed by visualizing the relationship between the following variables:
  - ❑ Flight number and Launch Site
  - ❑Payload and launch site
  - ❑Success rate of each orbit type
  - ❑Flight number and orbit type
  - ❑Payload and orbit type
  - ❑The launch success yearly trend.

- GitHub URL of EDA with data visualization:

https://github.com/farzana-zaki/DataScienceProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- EDA was applied with SQL to get insight from the data. The following queries were calculated:
  - ❑ The names of unique launch sites in the space mission.
  - ❑ 5 records where launch sites begin with “CCA”.
  - ❑ The total payload mass carried by boosters launched by NASA (CRS).
  - ❑ The average payload mass carried by booster version F9 v1.1.
  - ❑ The date when 1st successful landing outcome in ground pad was achieved.
  - ❑ The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - ❑ The total number of successful and failure mission outcomes.
  - ❑ The names of the booster versions which have carried the maximum payload mass.
  - ❑ The failed landing outcomes in drone ship, their booster version and launch site names in 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.
- GitHub URL of EDA with SQL:

https://github.com/farzana-zaki/DataScienceProject/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- All launch sites, and added map objects such as markers to create marks, circles to create a circle above markers, lines to create show the distance between 2 points are done to mark the success or failure of launches for each site on the folium map.
- The feature launch outcomes (failure or success) were assigned to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- Furthermore, the distances between a launch site to its proximities were calculated to figure out the following questions:
  - ❑ Are launch sites near railways, highways and coastlines?
  - ❑ Do launch sites keep certain distance away from cities?
- GitHub URL of interactive map with Folium map:

https://github.com/farzana-zaki/DataScienceProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

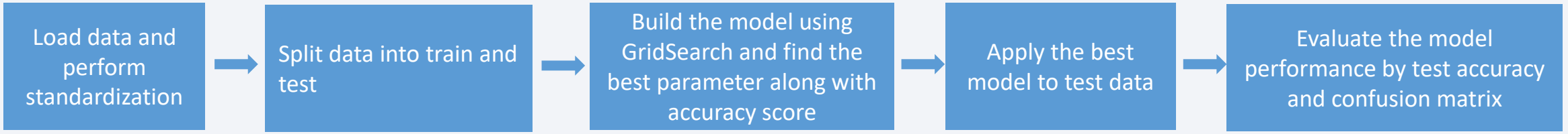
- An interactive dashboard was generated using Plotly dash.
- Pie charts were plotted to show the total launches for 5 different launch sites.
- Scatter graphs are plotted to show the relationship with Outcome and Payload Mass (Kg) for the different booster versions.
- GitHub URL of Plotly Dash lab:

  https://github.com/farzana-zaki/DataScienceProject/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

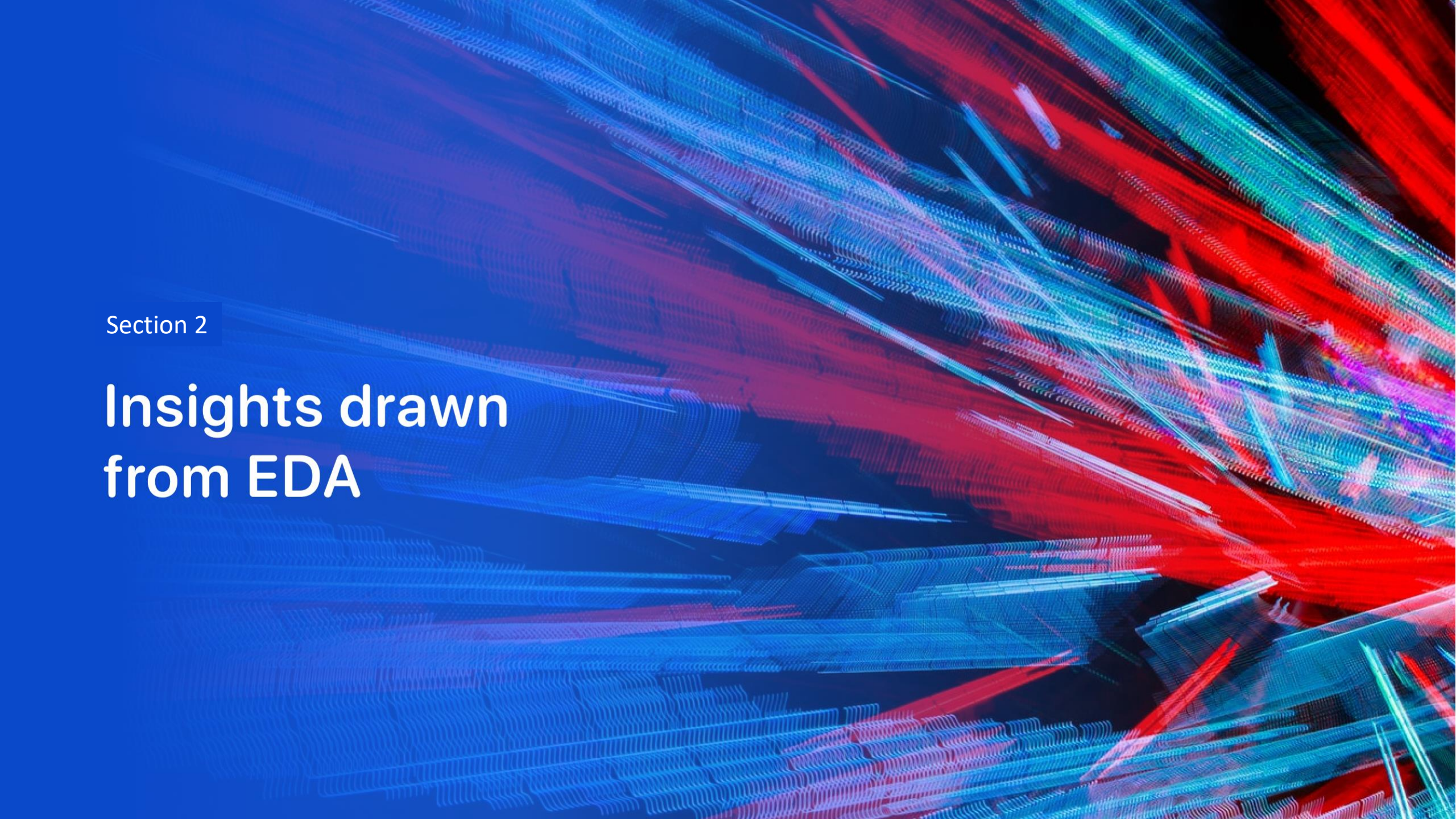
- Data were loaded using numpy and pandas, followed by data transformation and data split into training and testing datasets.
- Various machine learning models (Logistic regression, Decision tree, Support vector machine and KNN) were performed using tunning of different hyperparameters by GridSearchCV.
- Training and test accuracies and confusion matrix were measured as the performance metric for each model.
- Finally, the best performing classification model was found.
- GitHub URL of predictive analysis:

https://github.com/farzana-zaki/DataScienceProject/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

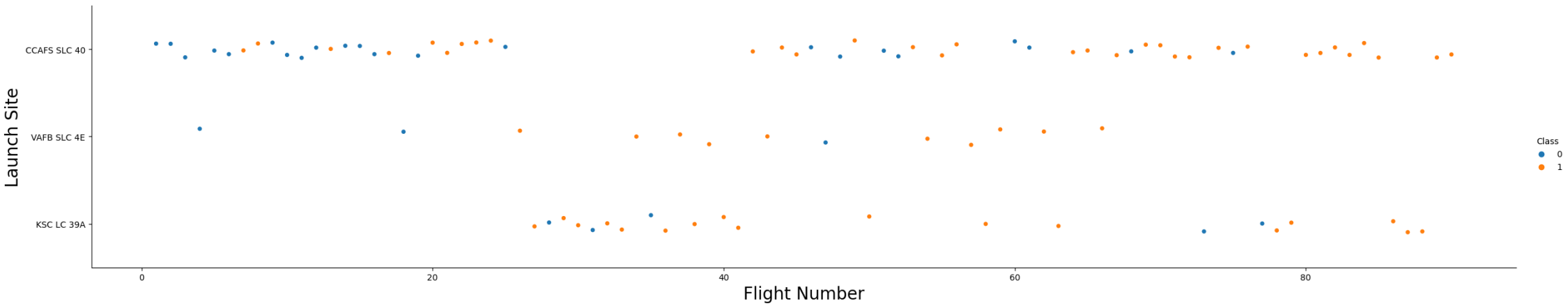


Results

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

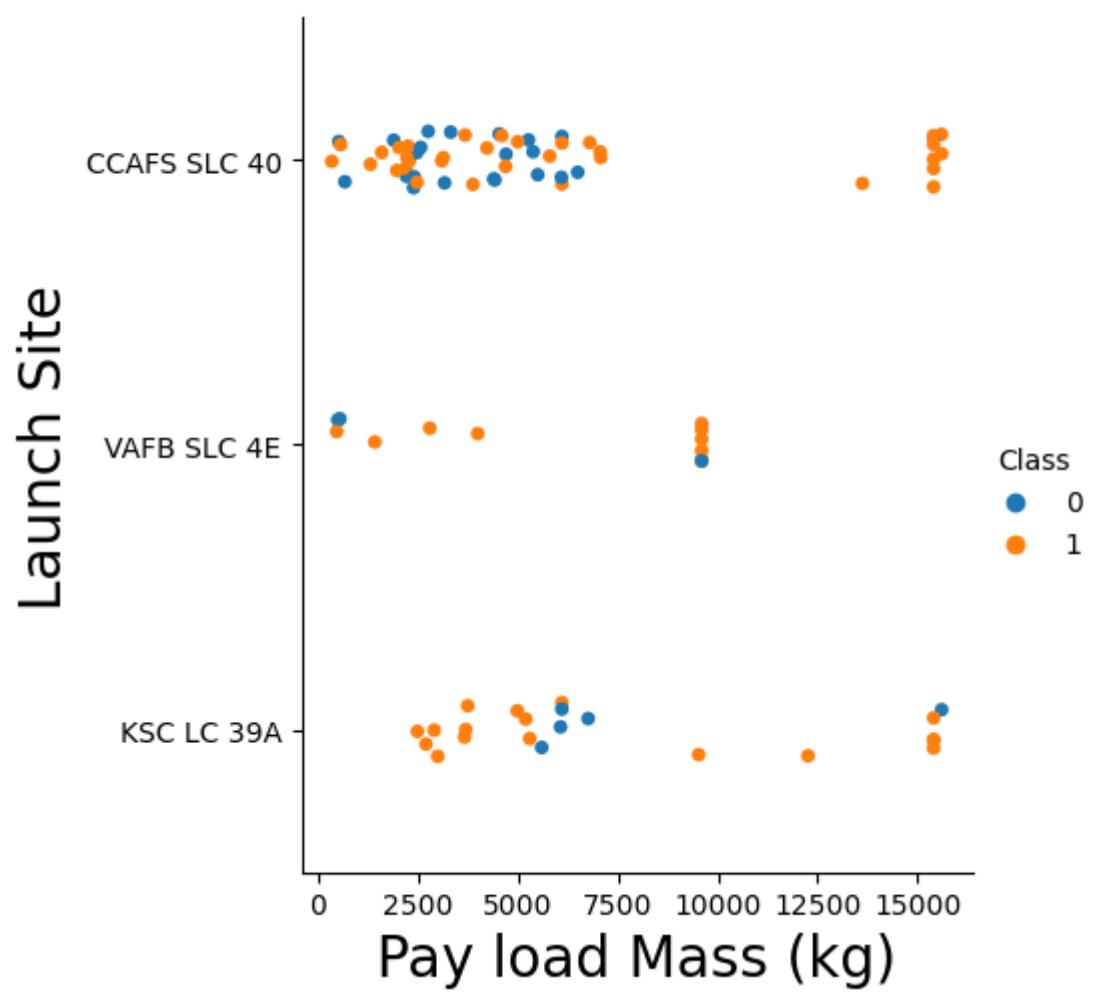

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



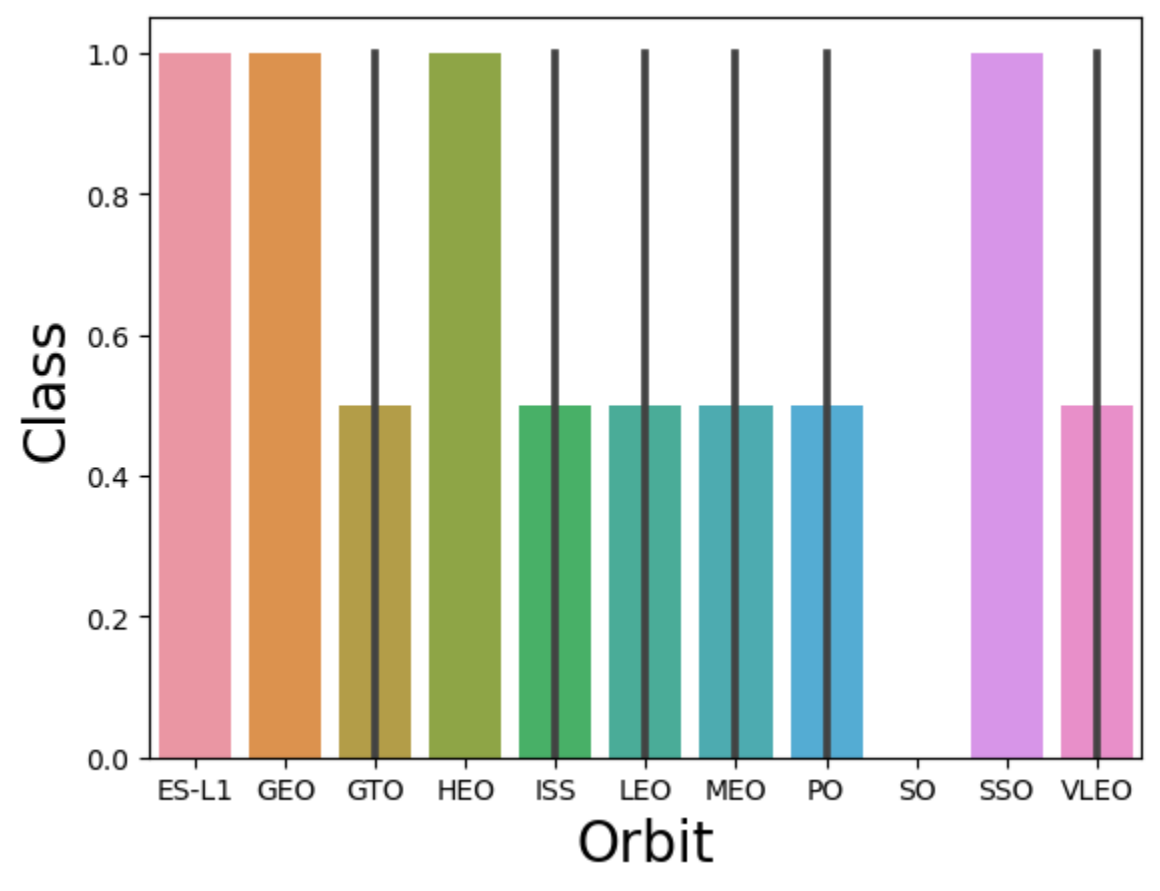
Rockets launch from the launch site CCAFS SLC-40 are significantly higher from the other two launch sites.

Payload vs. Launch Site



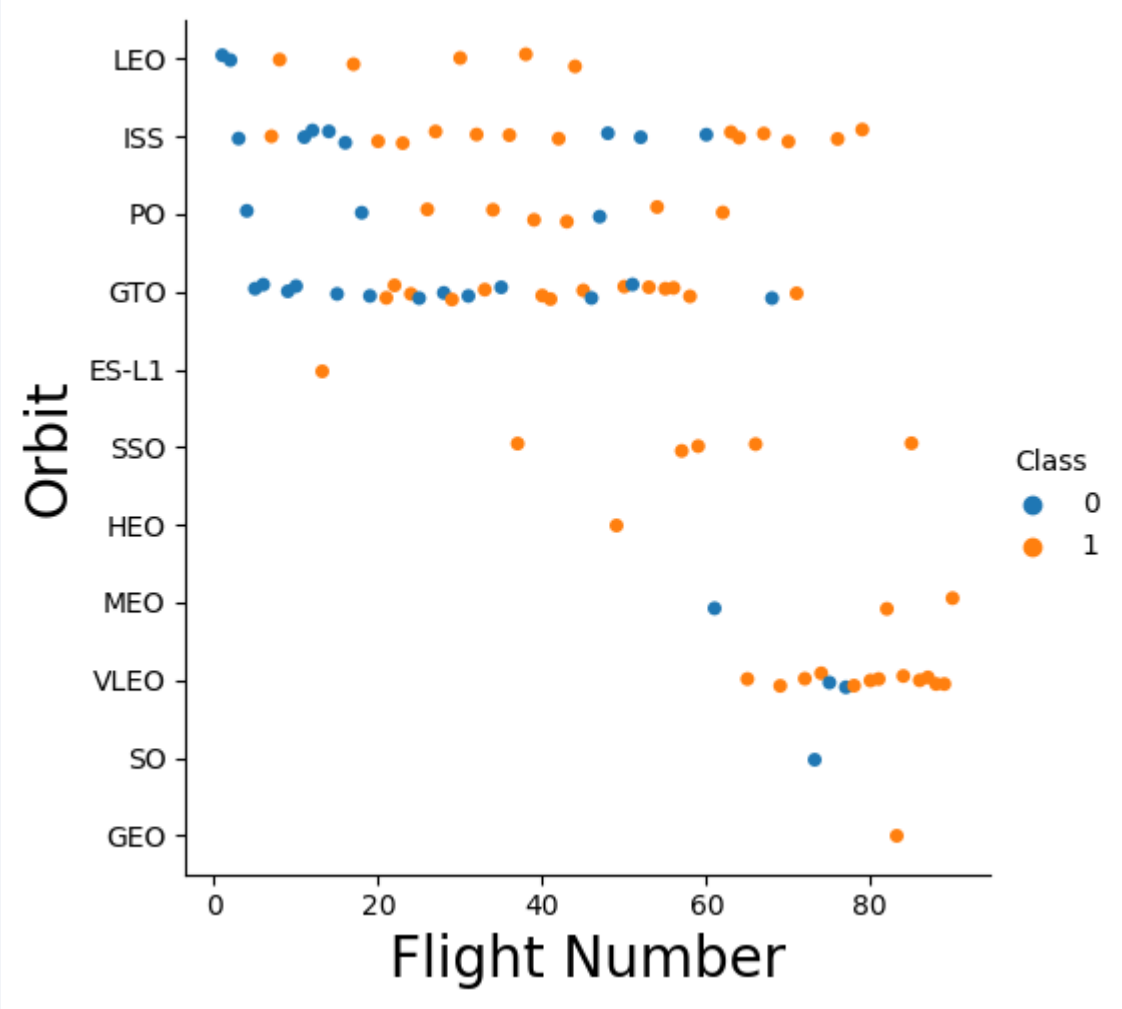
- From launch site CCAFS SLC 40, most of the successful launches occurred for the payload mass in the range up to 6500kg and some successful launches are in the heavy payload mass ranges of 12,500 – 15,000 kg.
- For the VAFB-SLC launch site , there are no rockets launched for heavy payload mass(greater than 10000kg).

Success Rate vs. Orbit Type



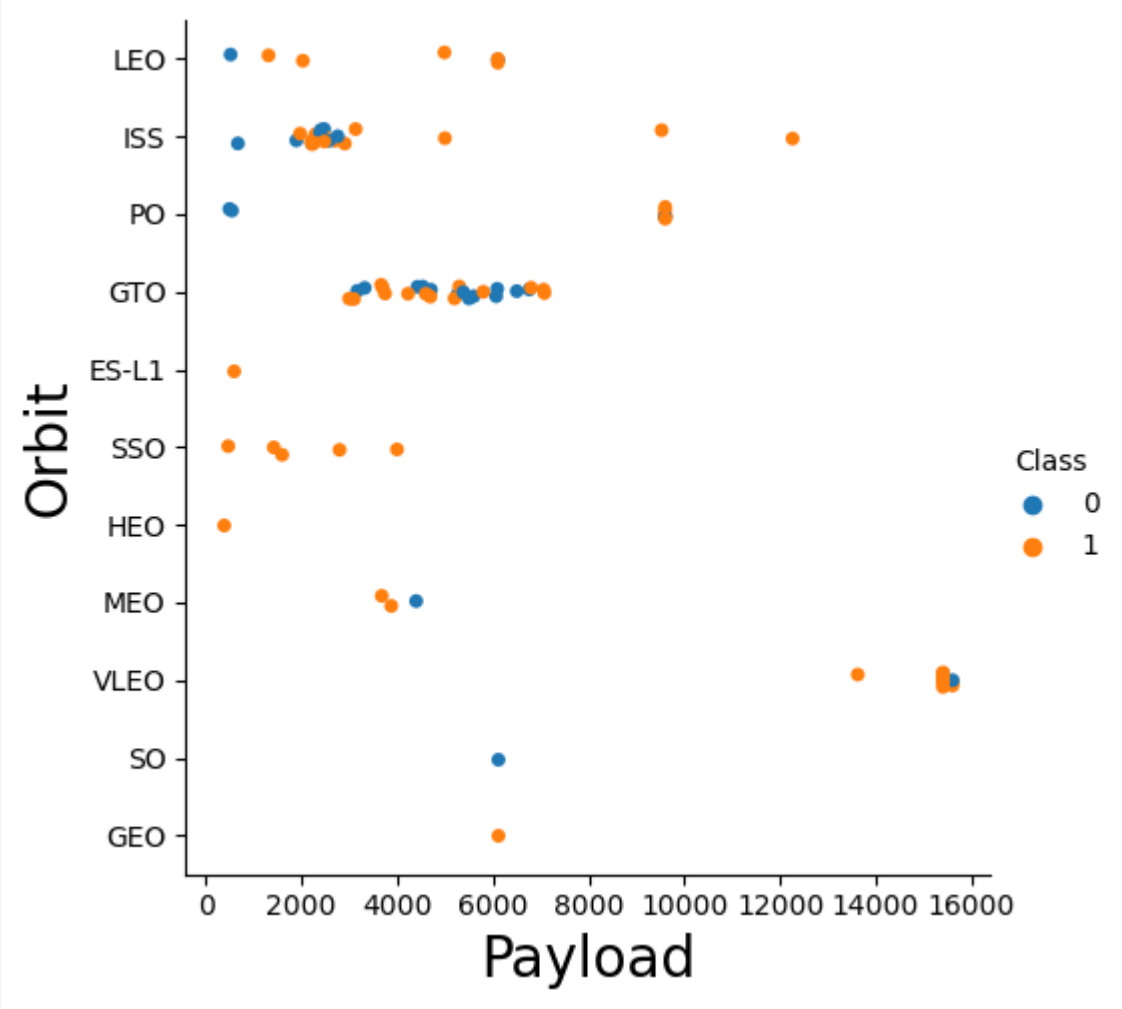
- The orbit types ES-L1, GEO, HEO and SSO have the highest success rate of 100% for the rocket launch.
- Orbit types GTO, ISS, LEO, MEO, PO and VLEO have the success rate of around 50%.

Flight Number vs. Orbit Type



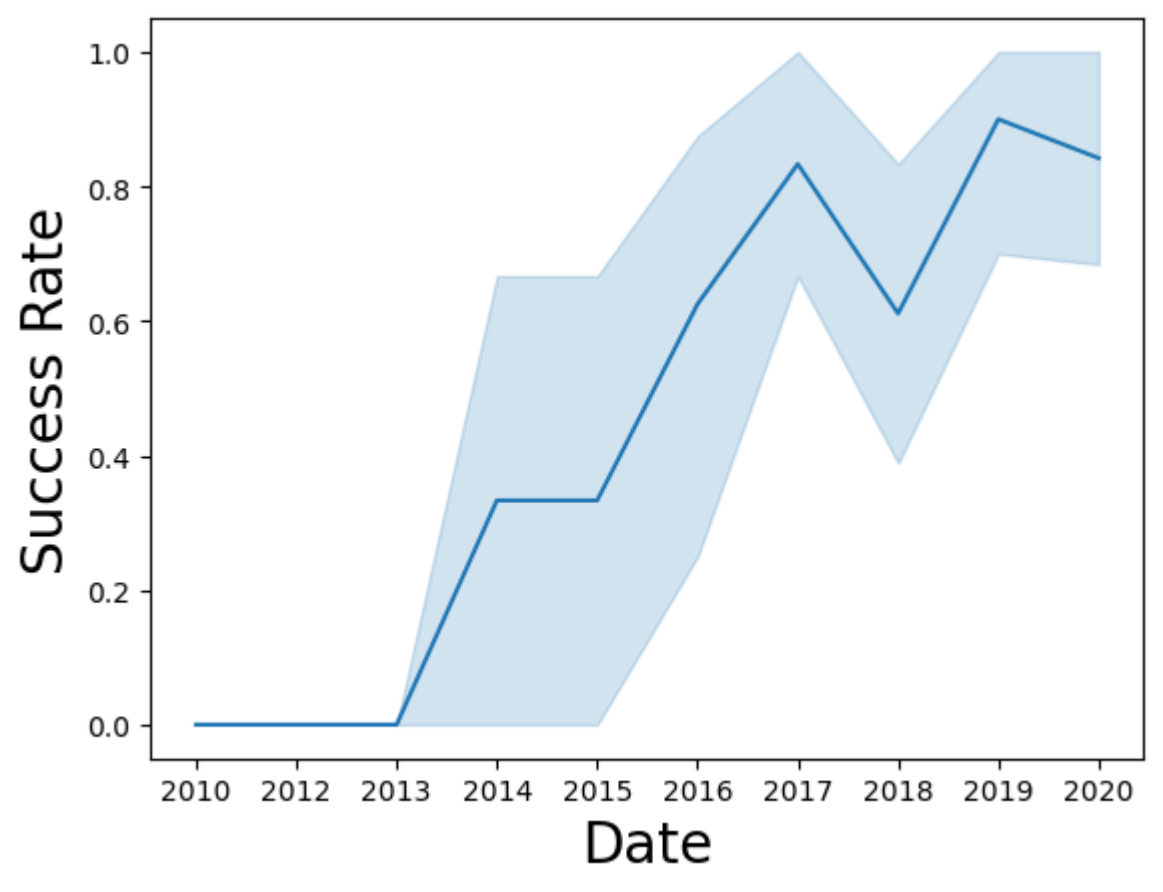
- In the LEO orbit, the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing rate are more for PO, LEO and ISS.
- However, for GTO we cannot distinguish this well as both successful landing rate and unsuccessful landing rate are both present.
- SSO orbit type is most successful for the payload ranges up to 4000kg.
- Heavy payload mass (around 15,000kg) are successful for the VLEO orbit type.

Launch Success Yearly Trend



- The launch success rate is increasing since 2013.
- However, there is a potential drop of success rate in 2018 and then again back to the success rate of around 80% in 2019.

All Launch Site Names

%sql select distinct (Launch_Site) from SPACEXTBL

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

- 4 launch sites are available in the dataset.

Launch Site Names Begin with 'CCA'

%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

- ‘sum’ aggregated function was used to get the total payload mass for the NASA (CRS).

Average Payload Mass by F9 v1.1

```
In [16]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
Out[16]: avg(PAYLOAD_MASS__KG_)
         2928.4
```

- ‘avg’ aggregated function was used to get the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

```
In [17]: %sql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
Out[17]:
```

min(Date)
2015-12-22

- ‘min’ aggregated function was used to get the first successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

- ‘where’ function was used to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as Count from SPACEXTBL
where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
group by MISSION_OUTCOME
```

Mission_Outcome	Count
Failure (in flight)	1
Success	98

- ‘count’ and ‘group by’ functions were used to calculate the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
In [24]: %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

Out[24]:

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

- A ‘sub query’ was used to list the names of the booster which have carried the maximum payload mass.

2015 Launch Records

```
%sql select substr(Date,6,2) as Month, Landing_Outcome, BOOSTER_VERSION,LAUNCH_SITE
from SPACEXTBL where Landing_Outcome = 'Failure (drone ship)' and substr(Date,1,4) ='2015'
```

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

- ‘substr’ was used to get the month and year name from the data set.
- 2 lists are found for the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

%sql select * from SPACEXTBL where Landing_Outcome = 'Failure (drone ship)' or Landing_Outcome = 'Success (ground pad)' and (Date between '2010-06-04' and '2017-03-20') order by Date desc

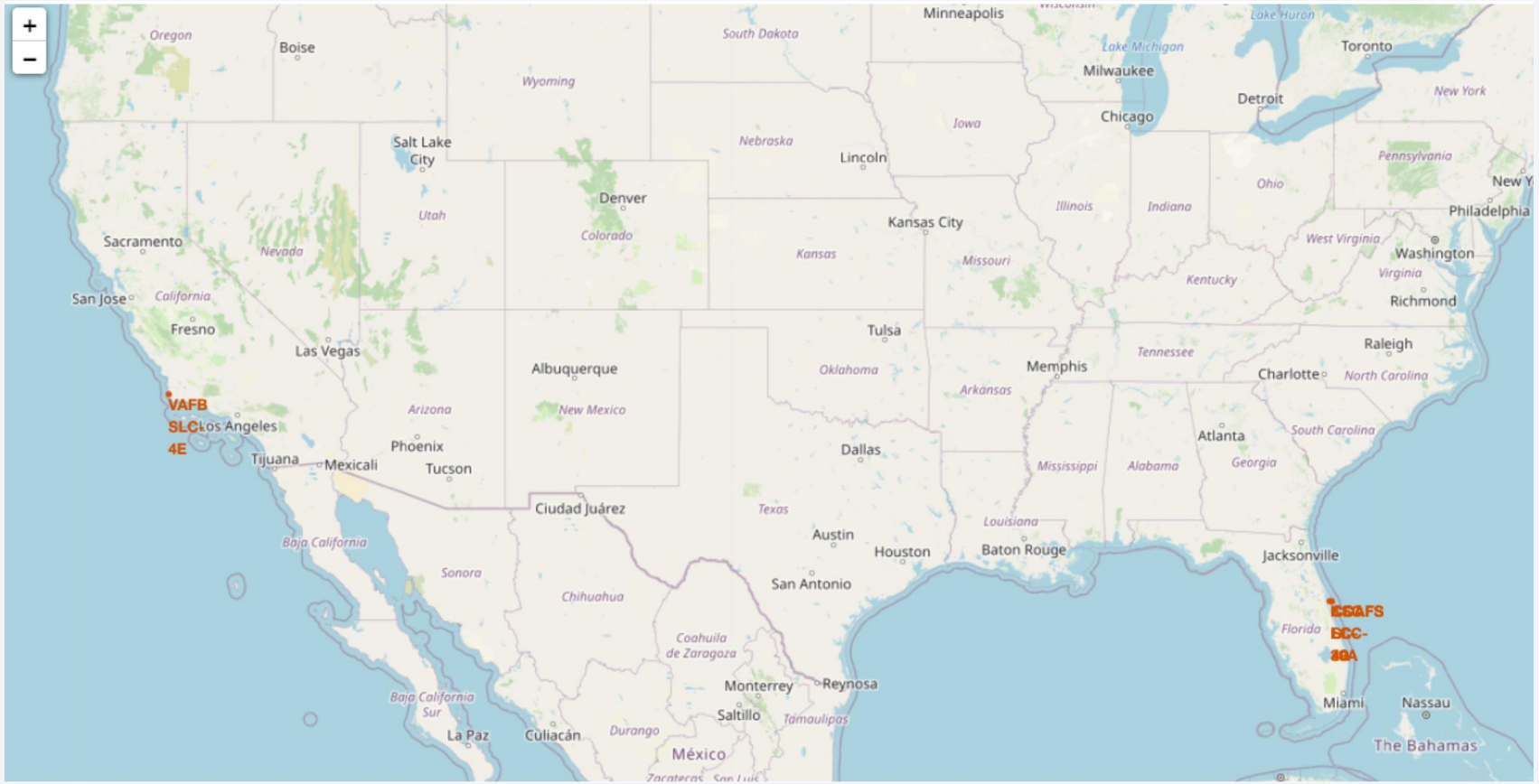
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-03-06	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	ABS-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	Success	Failure (drone ship)
2016-04-03	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success	Failure (drone ship)
2016-01-17	18:42:00	F9 v1.1 B1017	VAFB SLC-4E	Jason-3	553	LEO	NASA (LSP) NOAA CNES	Success	Failure (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
2015-10-01	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

- ‘where’, ‘and/or’, ‘order by’, and ‘desc’ functions were used 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.

Section 3

Launch Sites Proximities Analysis

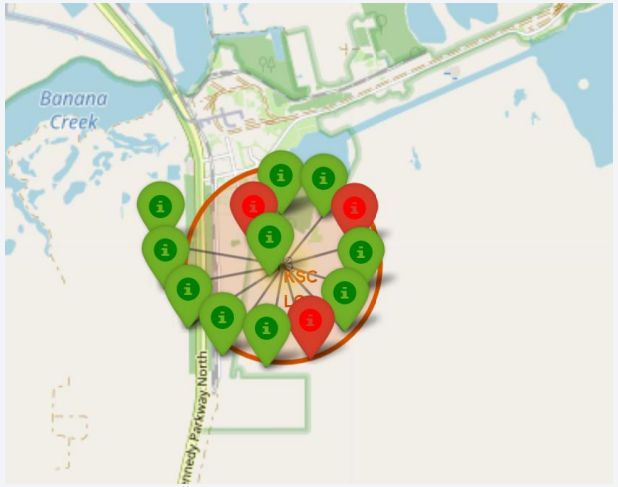
All Launch Sites Location Marker



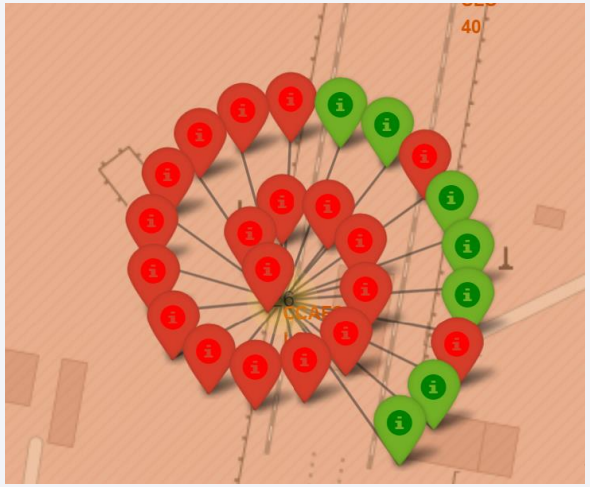
- All the launch sites in the USA (near California and Florida).

Color-labelled Launch Outcomes on Launch Sites

VAFB SLC-4E
Near California



KSC LC-39A



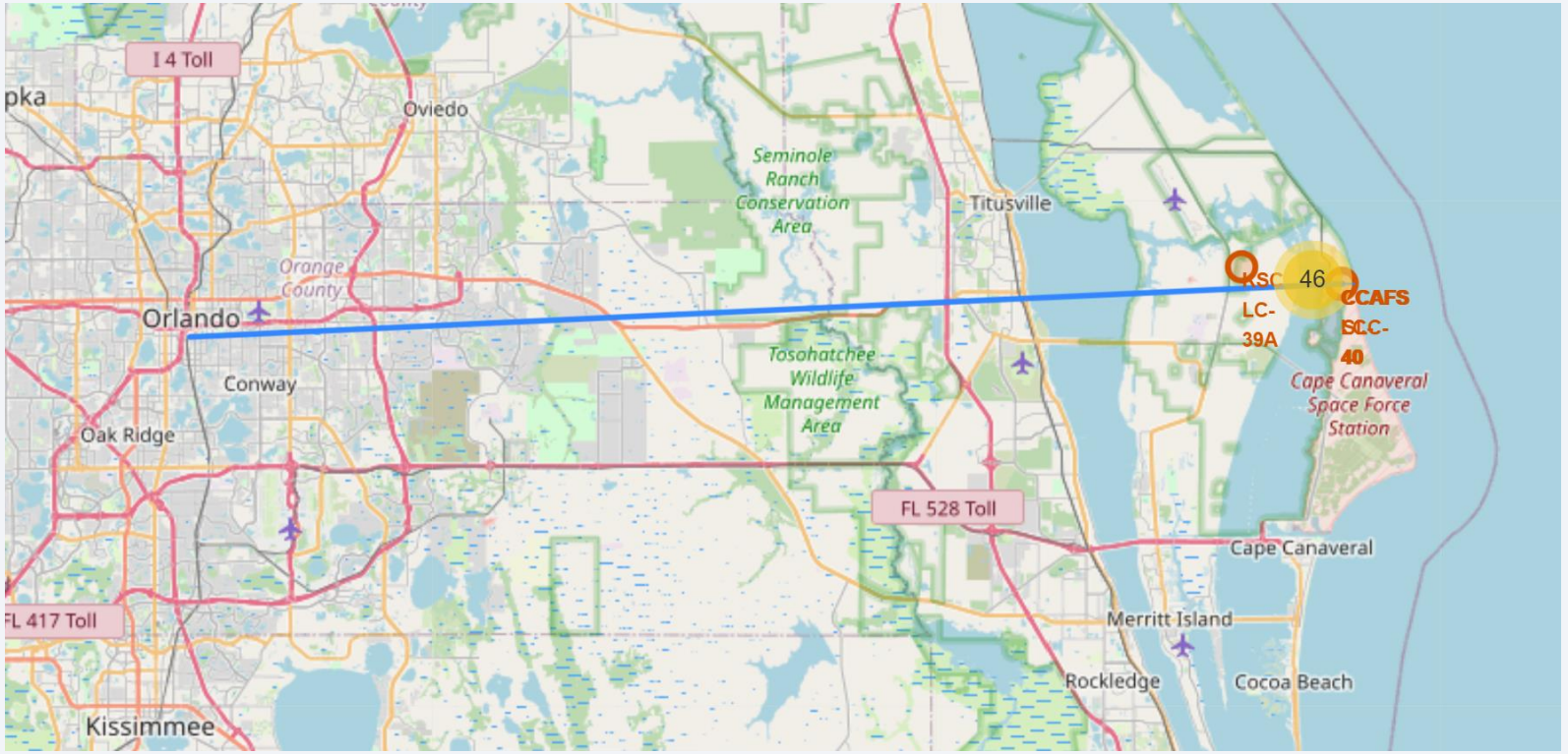
CCAFS LC-40

CCAFS SLC-40

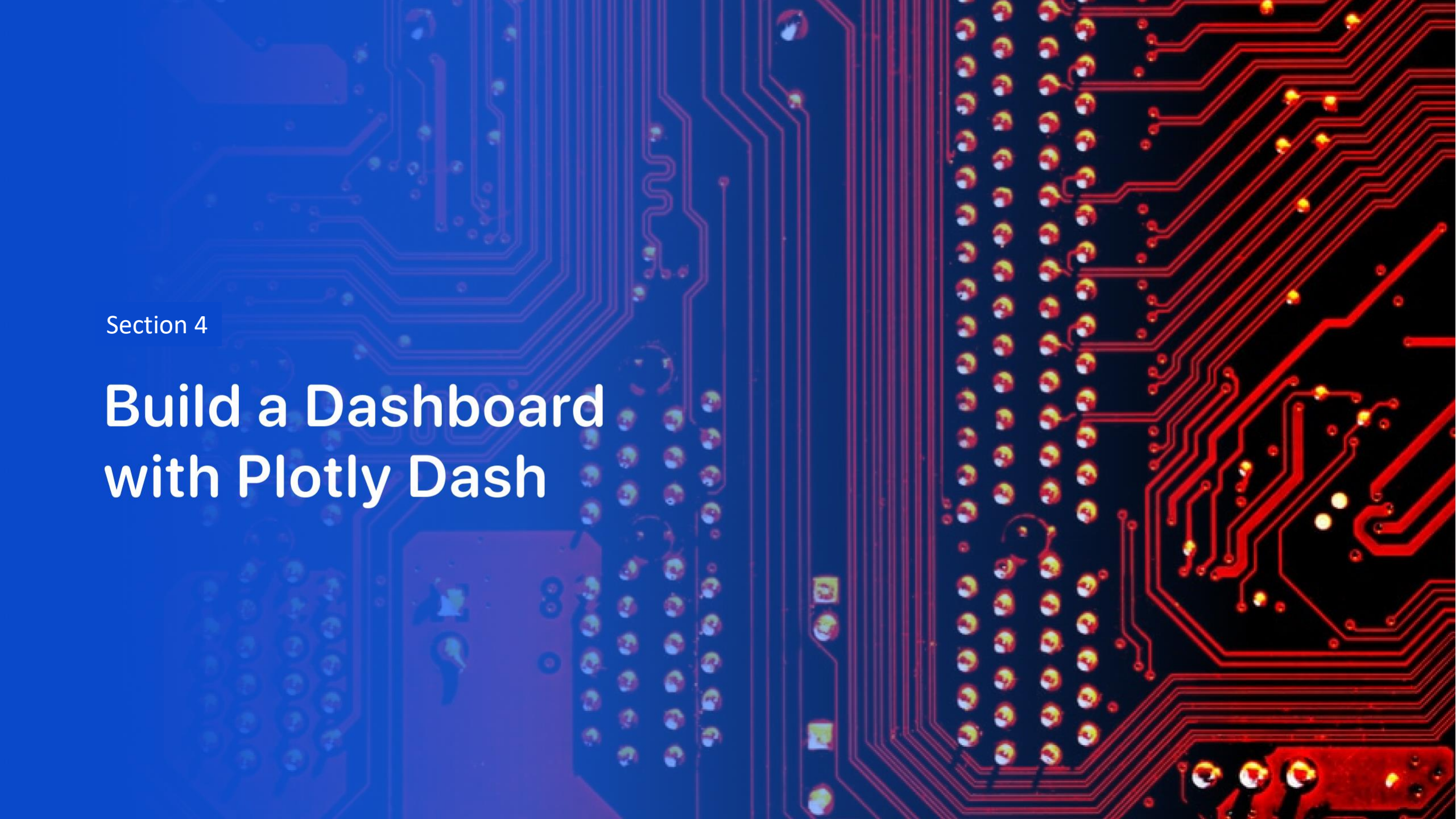
Near Florida

- One site is found near California (VAFB SLC-4E) and 3 launch sites are found near Florida (KSC LC-39A, CCAFS LC-40 and CCAFS SLC-40).
- Green marker indicates successful launch and the red marker indicates the unsuccessful launch.

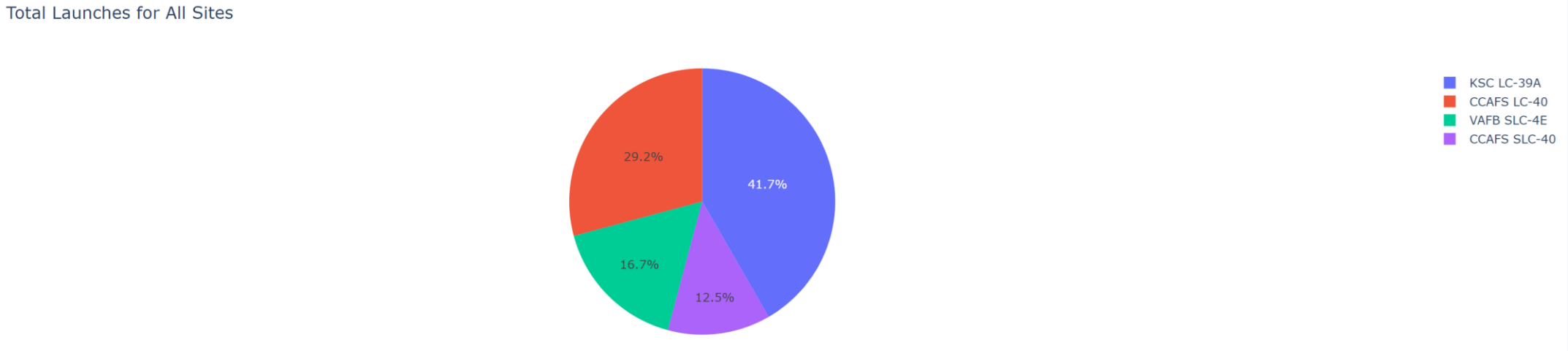
Launch Site to its Proximities



- All the Florida launch sites are nearby and they are not far from the railway tracks.

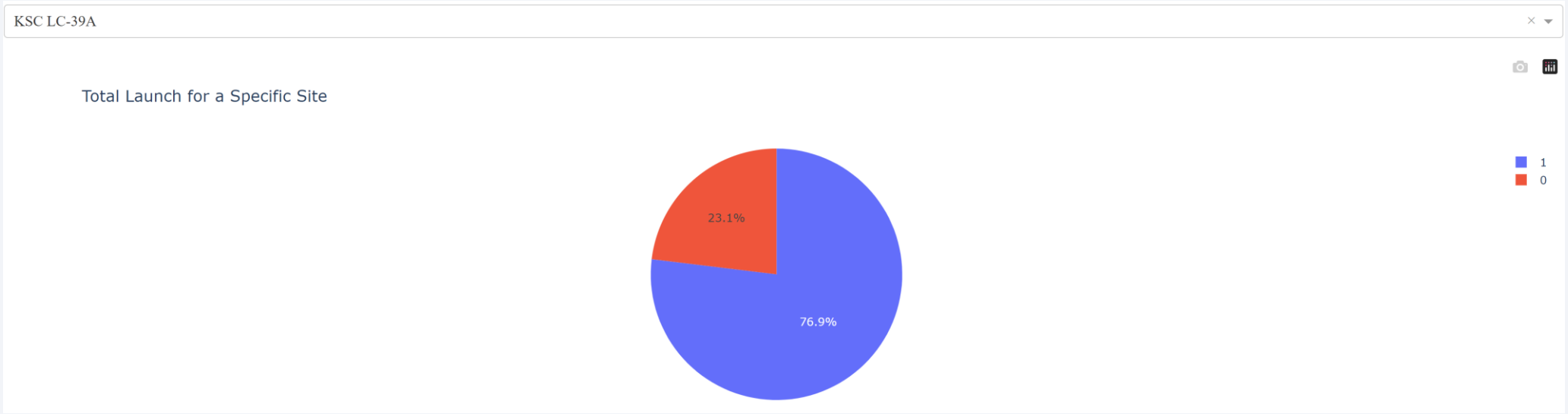

Section 4
Build a Dashboard with Plotly Dash

Launch success rate (Total success launches for all sites)



- KSC LC-39A had the most successful launch rate (41.7%), followed by CCAFS LC-40 with successful launch rate of 29.2% for all sites.
- On the other hand, VAFB SLC-4E had the success rate of 16.7% and CCAFS SLC-40 has the successful launch rate of 12.5% for all sites.

Launch Site (KSC LC-39A) with highest score



- KSC LC-39A launch site has 76.9% successful launching rate.

Payload, Booster version vs. Launch Outcome

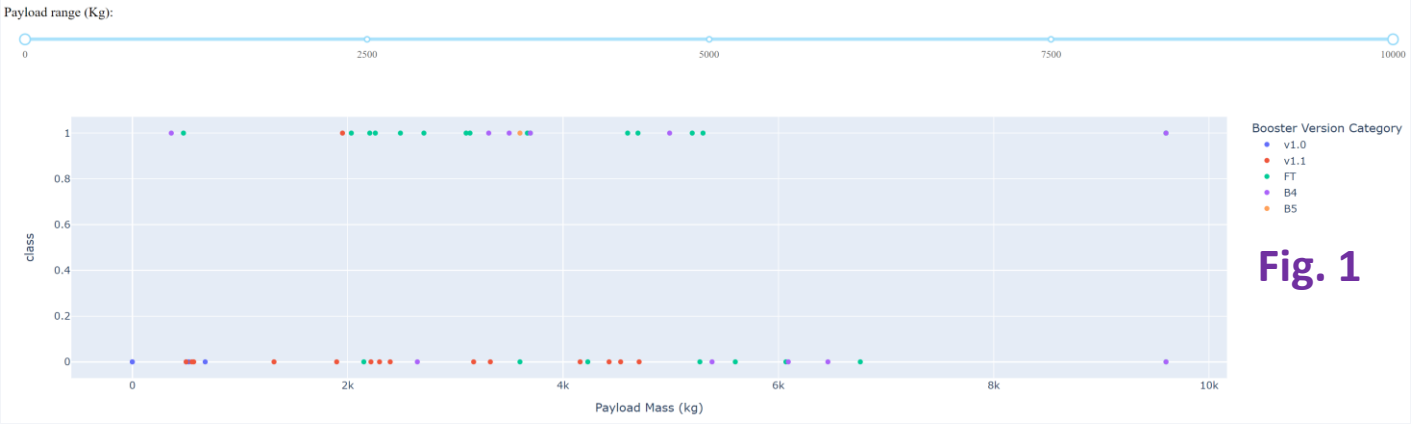


Fig. 1

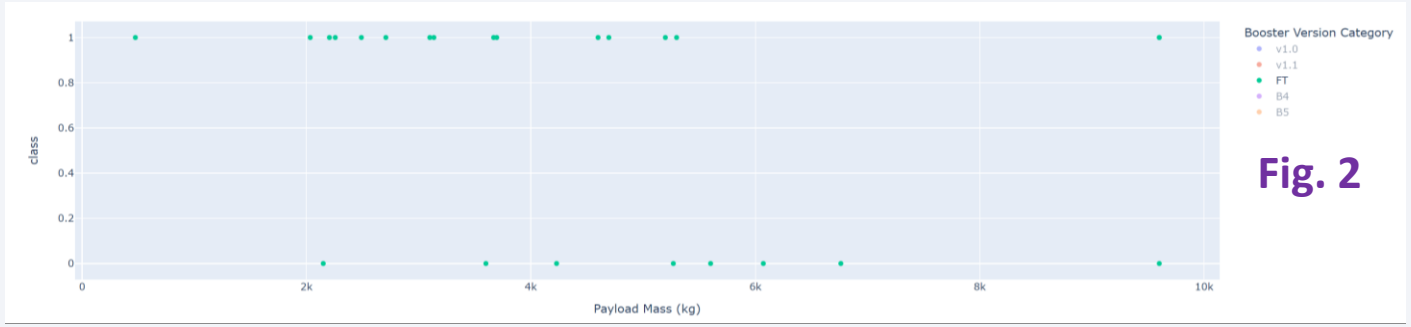


Fig. 2

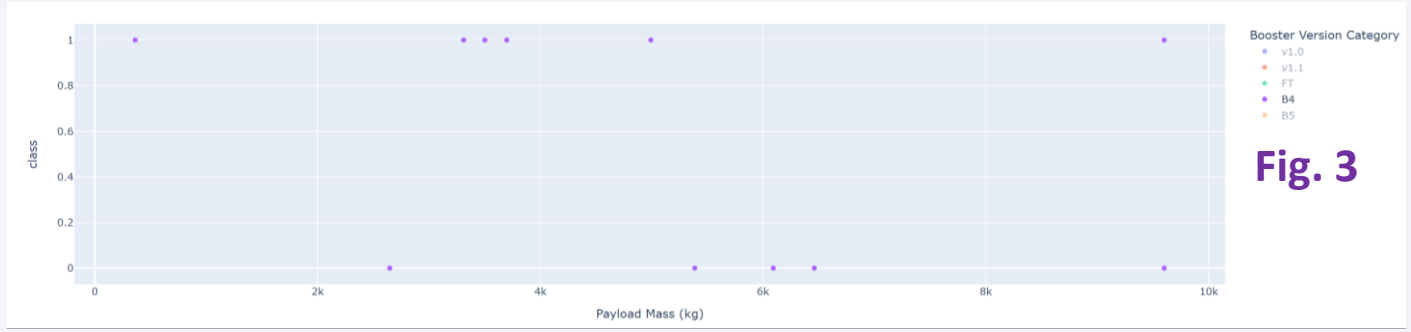


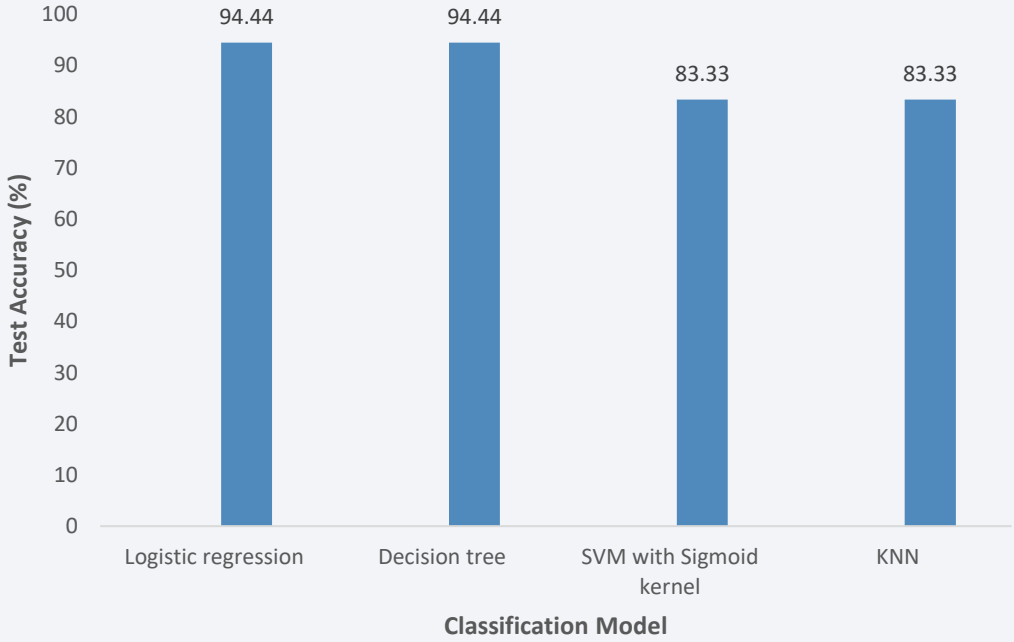
Fig. 3

- Payload Range vs Launch Outcome: Most successful launch outcomes are observed in the payload range of 2000 – 5300 kg (Fig. 1).
  - ❑ Overall Success rate = ~ 54.55% by all 5 Booster version categories.
- Booster Version vs Launch Outcome: Most successful launch outcomes are observed by FT booster version (Fig. 2).
  - ❑ Overall Success rate = 63.64% for 500 – 9500 Kg payload range.
  - ❑ Success rate = 70.59% for 2000 – 5500 Kg payload range.
- B4 had success rate of 66.67% in the payload range of 2000 – 5500 Kg. (Fig. 3).
- On the other hand, v1.0 and v1.1 were unsuccessful for any payload ranges.

Section 5

Predictive Analysis (Classification)

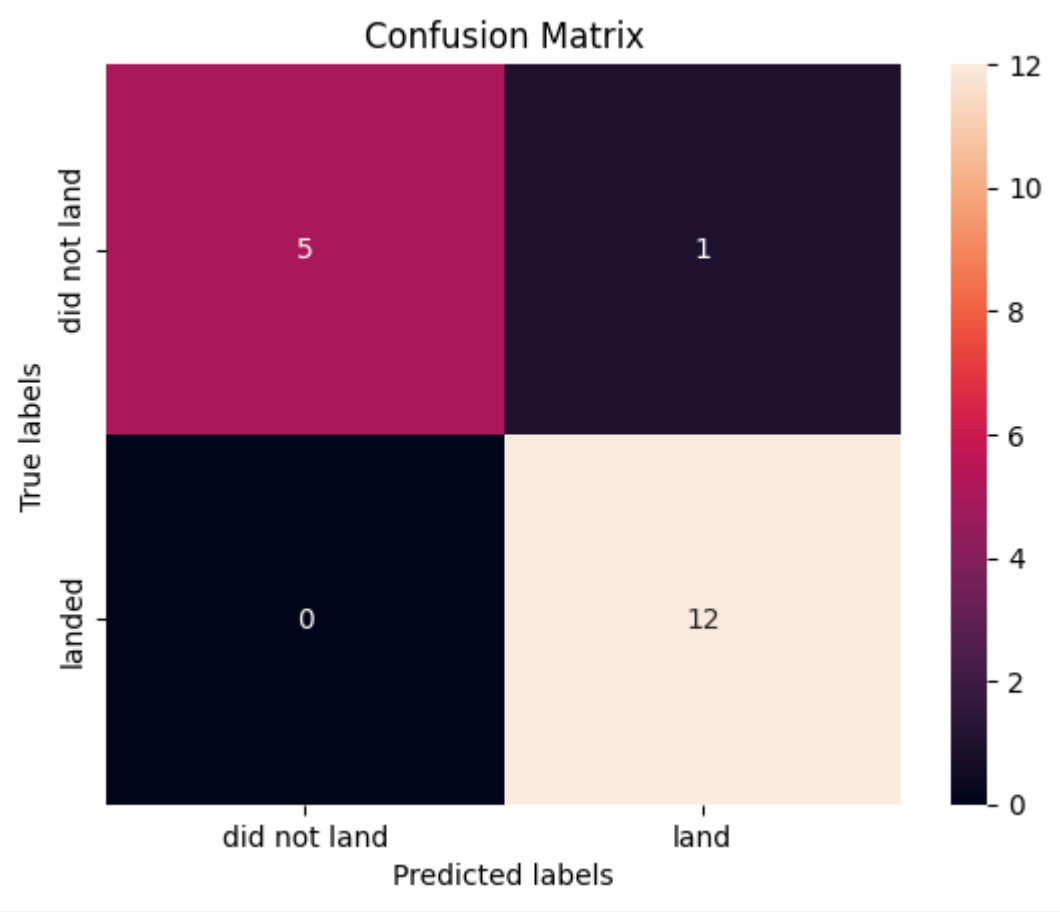
Classification Accuracy



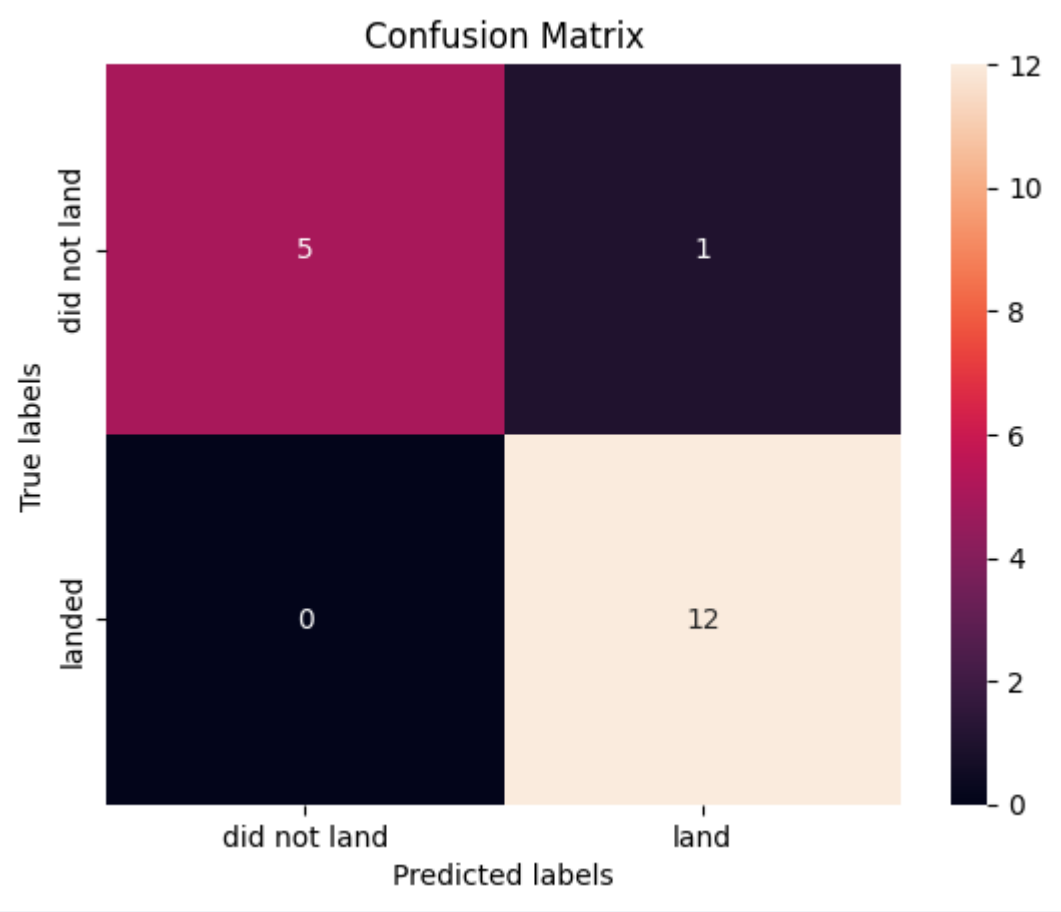
- Both Logistic regression and decision tree models had the best test accuracies of 94.44%.

Confusion Matrix

Logistic regression



Decision Tree



Conclusions

- KSC LC-39A had the most successful launches of any sites.
- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Most successful launch outcomes are observed in the payload range of 2000 – 5300 kg.
- FT Booster version had the most successful launch outcomes with success rate of around 71% in the 2000 – 5300 kg payload range and around 64% success rate in the 500 – 9500 kg payload range.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- Both Logistic Regression and Decision tree classifiers had shown the optimal machine learning algorithms for the provided dataset with test accuracy of 94.44%.

Appendix

All the codes can be found on my GitHub:

https://github.com/farzana-zaki/DataScienceProject

Thank you!