Winning Space Race with Data Science

Suvidya Yadav

7.12.23

Outline

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

Executive Summary

The project focused on proficiently applying data science and machine learning techniques to a real-world dataset, specifically, the SpaceX data. Key activities undertaken included data collection, wrangling, and exploratory analysis. This was coupled with data visualization and model development to predict the success of landing for the Falcon 9's first stage.

Various machine learning models were developed for this purpose, including support vector machines, decision tree classifiers, and k-nearest neighbors. These models were evaluated based on their predictive analysis results.

This comprehensive comparison allowed us to understand their respective strengths and weaknesses, thereby identifying the optimal model for our project. The findings and insights derived from this project have been diligently prepared for stakeholder review and consideration. The ultimate goal is to provide a robust machine-learning model that accurately predicts the success of Falcon 9’s first-stage landing, thus contributing to SpaceX's mission success.

Introduction

In this project, expertise in data science and machine learning methodologies was demonstrated by utilizing a real-world data set—the SpaceX Falcon 9 launch data. A comprehensive report was meticulously prepared for the stakeholders, encapsulating the complete process and results of this data science project. The process began with rigorous data collection and data wrangling procedures on the SpaceX data, followed by exploratory data analysis and the development of data visualization models. The primary objective was to predict the successful landing of Falcon 9's first stage. The importance of this prediction stems from the fact that SpaceX offers Falcon 9 rocket launches for 62 million dollars, substantially less than other providers who charge upwards of 165 million dollars. A large portion of these savings can be attributed to SpaceX's ability to reuse the first stage of the rocket. Thus, by predicting the successful landing of the first stage, we can estimate the cost of a launch. This prediction is particularly valuable to other companies considering competing against SpaceX for rocket launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Extracted data using API and Web Scraping techniques
- Perform data wrangling
  - Perform exploratory Data Analysis and determine Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Perform exploratory Data Analysis and determine Training Labels.
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the method performs best using test data

Data Collection

- The data collection was conducted through API requests from Space X's public API and web scraping data from a table in Space X's Wikipedia entry.
- Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,

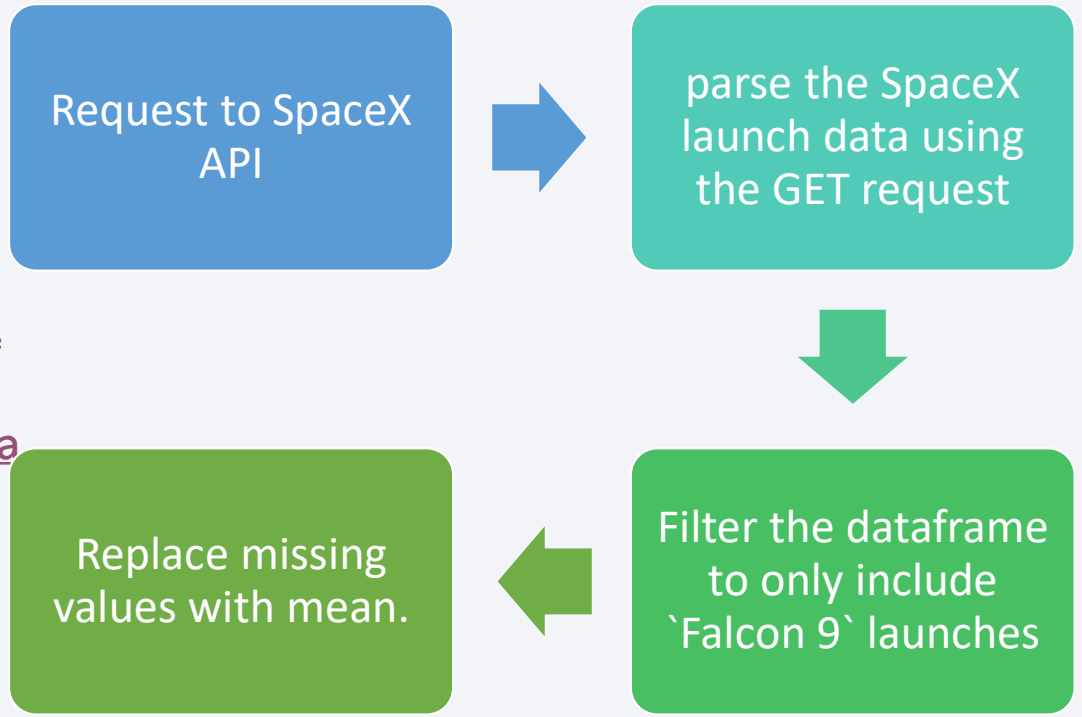
Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- Wikipedia Webscrape Data Columns:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

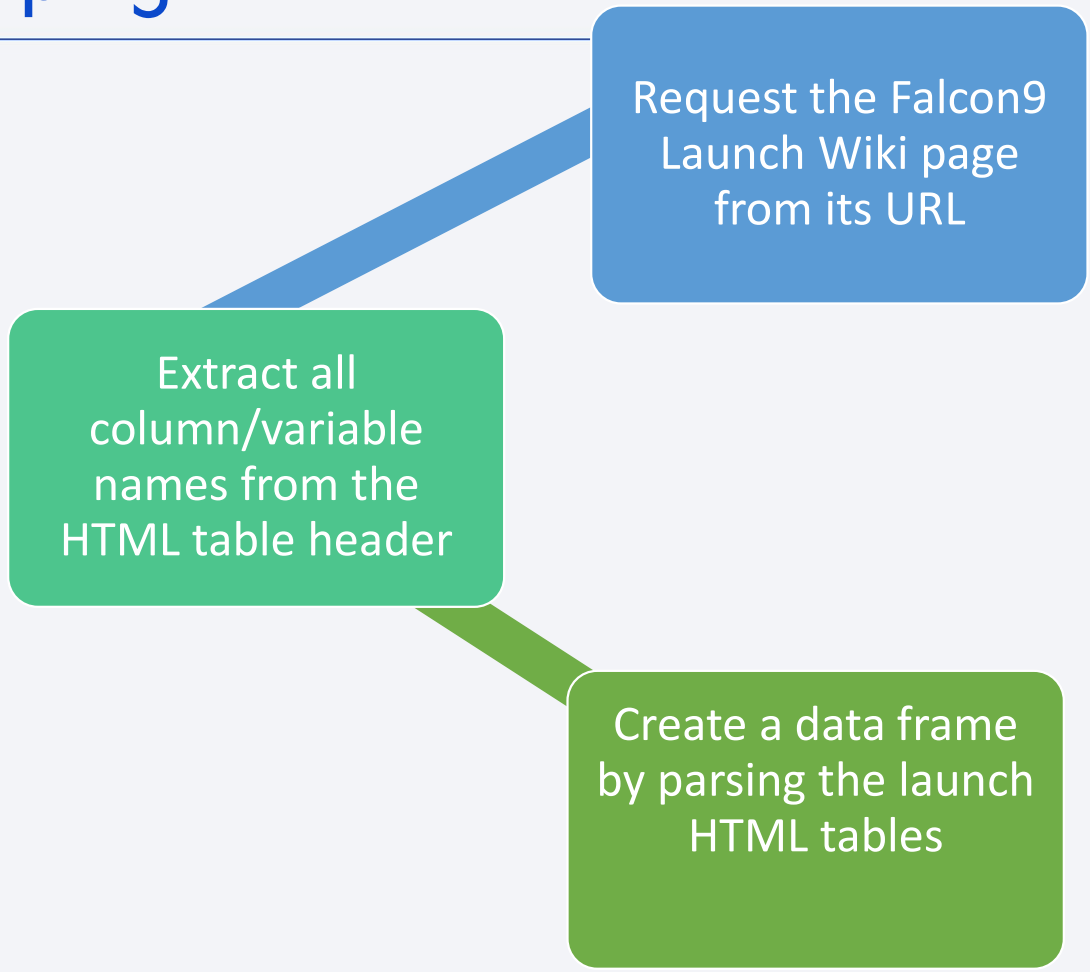
Data Collection – SpaceX API

- GitHub URL - https://github.com/suvidyayadav/appliedcapstoneprojectibm/blob/ce96a172d7478b34480bb300e0bab3229e80885d/Data%20Collection%20API-Lab%201.ipynb



Data Collection - Scraping

- GitHub URL- https://github.com/suvidyayadav/appliedcapstoneprojectibm/blob/ce96a172d7478b34480bb300e0bab3229e80885d/Webscraping-Lab%202.ipynb



Data Wrangling

- Performed some Exploratory Data Analysis (EDA) to find patterns in the data and determine what would be the label for training supervised models. The data contains several Space X launch facilities: Cape Canaveral Space Launch Complex 40 VAFB SLC 4E, Vandenberg Air Force Base Space Launch Complex 4E (SLC-4E), and Kennedy Space Center Launch Complex 39A KSC LC 39A.The location of each Launch Is placed in the column LaunchSite.
- Calculated the number of launches for each site.
- Calculated the number and occurrence of each orbit
- Calculated the number and occurrence of mission outcome of the orbits
- Created a landing outcome label from the Outcome column. Converted those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- GitHub URL- https://github.com/suvidyayadav/appliedcapstoneprojectibm/blob/ce96a172d7478b34480bb300e0bab3229e80885d/Data%20Wrangling-Lab%203.ipynb

EDA with Data Visualization

Charts plotted:

- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

Scatter plots, Cat plots, line charts, and bar plots were used to compare relationships between variables to determine if a relationship exists so that they could potentially be used in training the machine learning model

GitHub URL-

https://github.com/suvidyayadav/appliedcapstoneprojectibm/blob/ce96a172d7478b34480bb300e0bab3229e80885d/EDA%20Data%20Visualisation-Lab%205.ipynb

EDA with SQL

SQL queries performed:

- Displayed the names of the unique launch sites in the space mission.
- Displayed 5 records where launch sites begin with the string 'CCA'.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed average payload mass carried by booster version F9 v1.1.
- Listed the date when the first successful landing outcome in the ground pad was achieved.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listed the total number of successful and failure mission outcomes.
- Listed the names of the booster_versions which have carried the maximum payload mass using a subquery.
- Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranked 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-
https://github.com/suvidyayadav/appliedcapstoneprojectibm/blob/ce96a172d7478b34480bb300e0bab3229e80885d/EDA%20SQL-Lab%204.ipynb

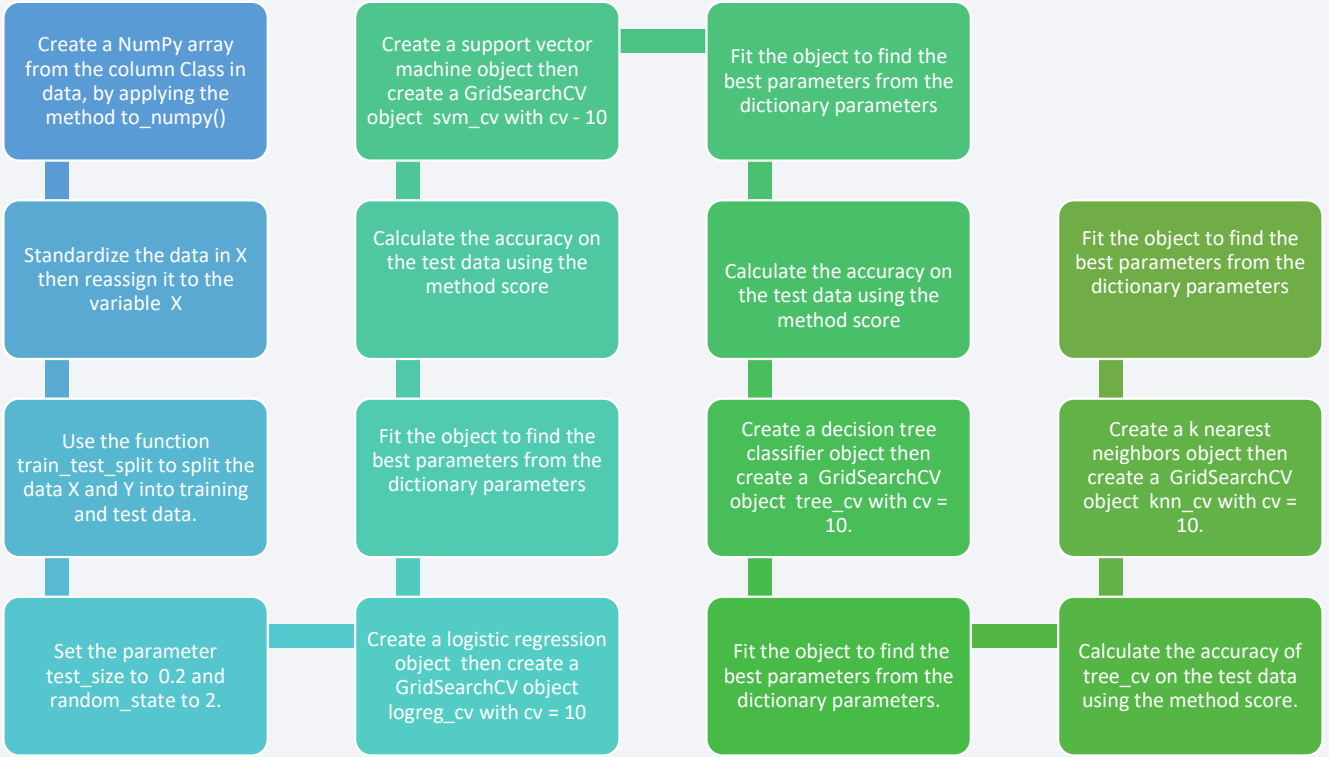
Build an Interactive Map with Folium

- Marked all launch sites on a map using Folium map objects, marker, Circle.
- Marked the success/failed launches for each site on the map,If a launch was successful `(class=1)`, then used a green marker and if a launch was failed, used a red marker `(class=0)`
- Calculated the distances between a launch site to its proximities.
- GitHub URL- https://github.com/suvidyayadav/appliedcapstoneprojectibm/blob/ce96a172d7478b34480bb300e0bab3229e80885d/Folium-Lab%206.ipynb

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List: - Added a dropdown list to enable Launch Site selection.
- Pie Chart displaying Success Launches: - Added a pie chart to display the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Slider for Payload Mass Range: - Added a slider to select Payload range.
- Plot of Payload Mass vs. Success Rate for the different Booster Versions: Visualized the correlation between Payload and Launch Success using Scatter Plot.
- GitHub URL- https://github.com/suvidyayadav/appliedcapstoneprojectibm/blob/092bdc4bf26a051a48447f6fe5393957aa13309e/spacex_dash_app.py

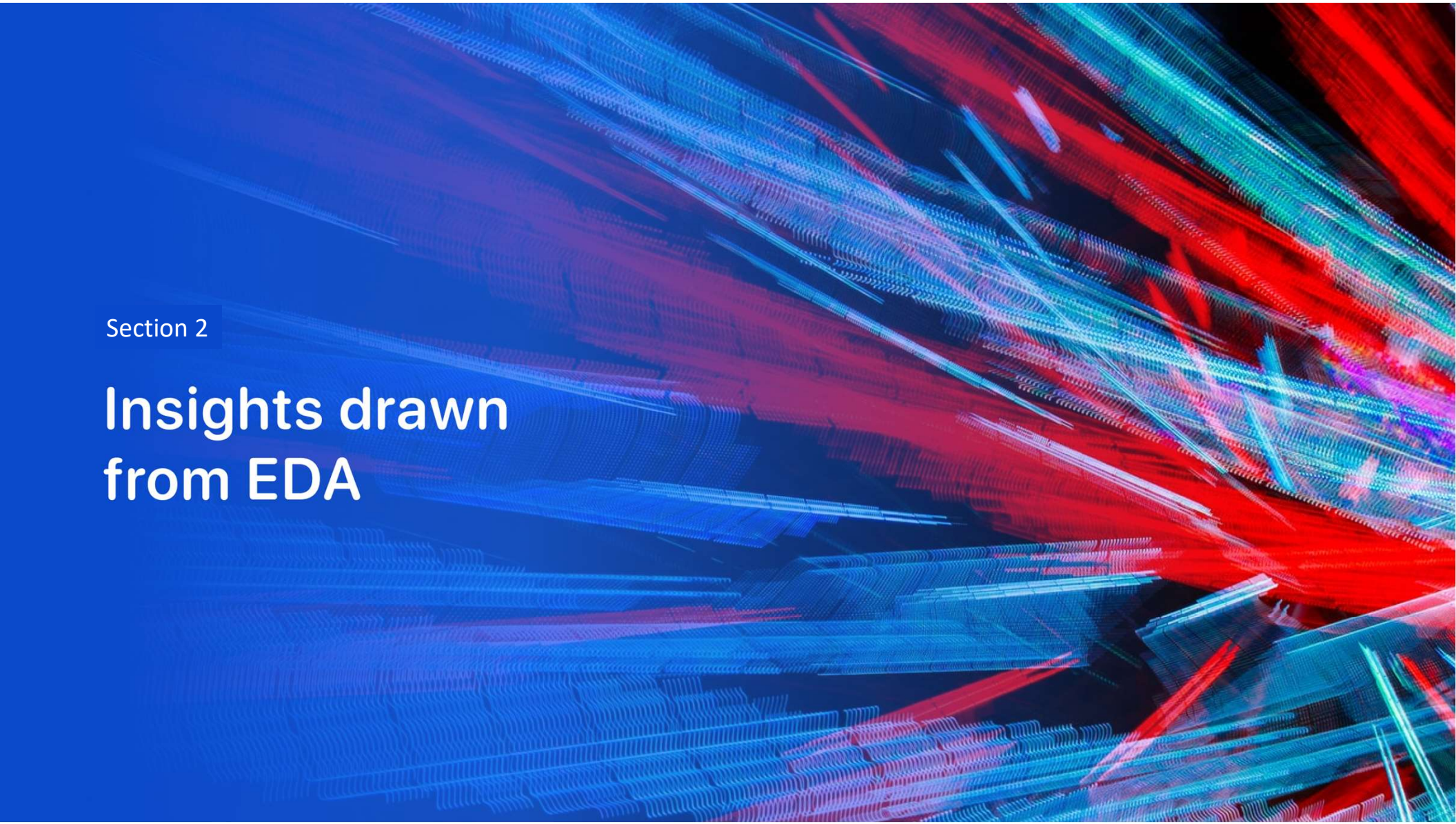
Predictive Analysis (Classification)



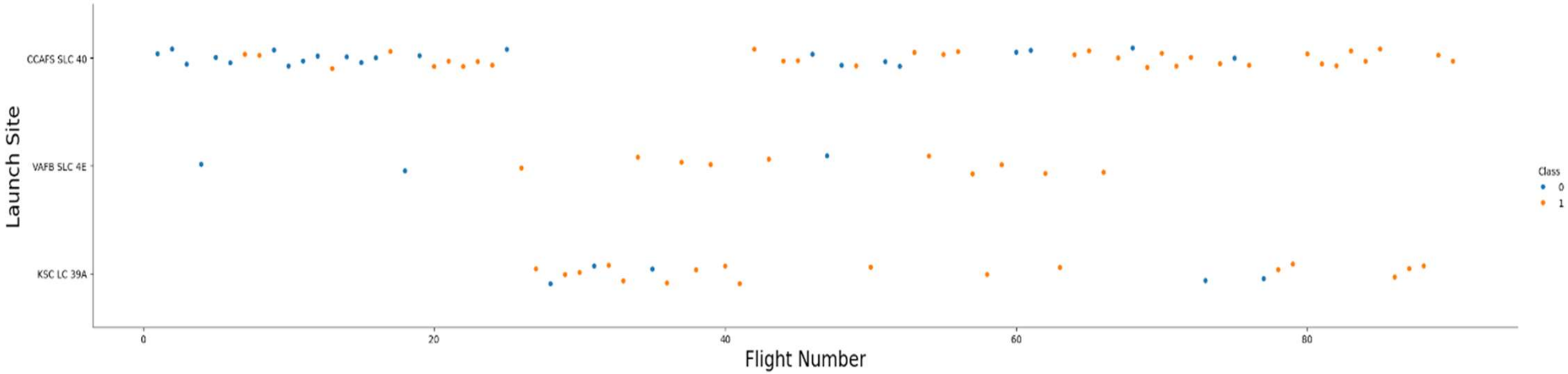
GitHub URL-
https://github.com/suvidyayadav/appliedcapstoneprojectibm/blob/092bdc4bf26a051a48447f6fe5393957aa13309e/Machine%20Learning-Lab%207.ipynb

Results

- Exploratory data analysis results
- Interactive analytics results
- Predictive analysis results

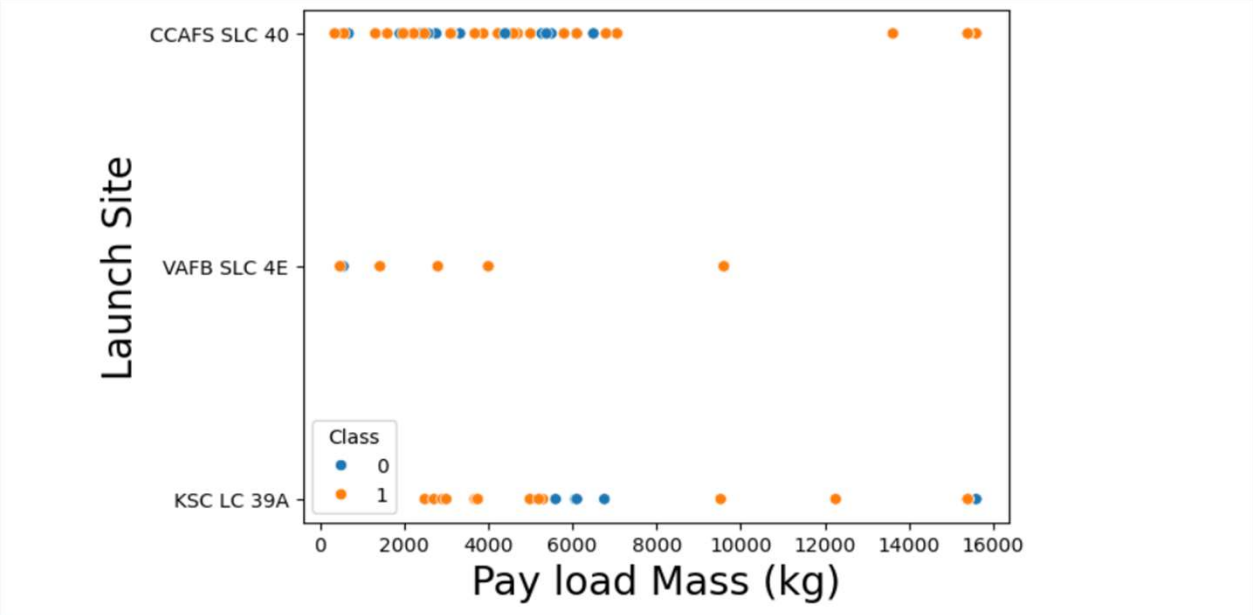

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 launch site has the maximum number of launches.
- The recent flights have an increased success rate in comparison to the earlier flights.

Payload vs. Launch Site



- The higher the payload mass, the higher the success rate.
- Most of the payload mass values lie under the range of 0-7000kg.

Success Rate vs. Orbit Type

• Orbits with 100% success rate:

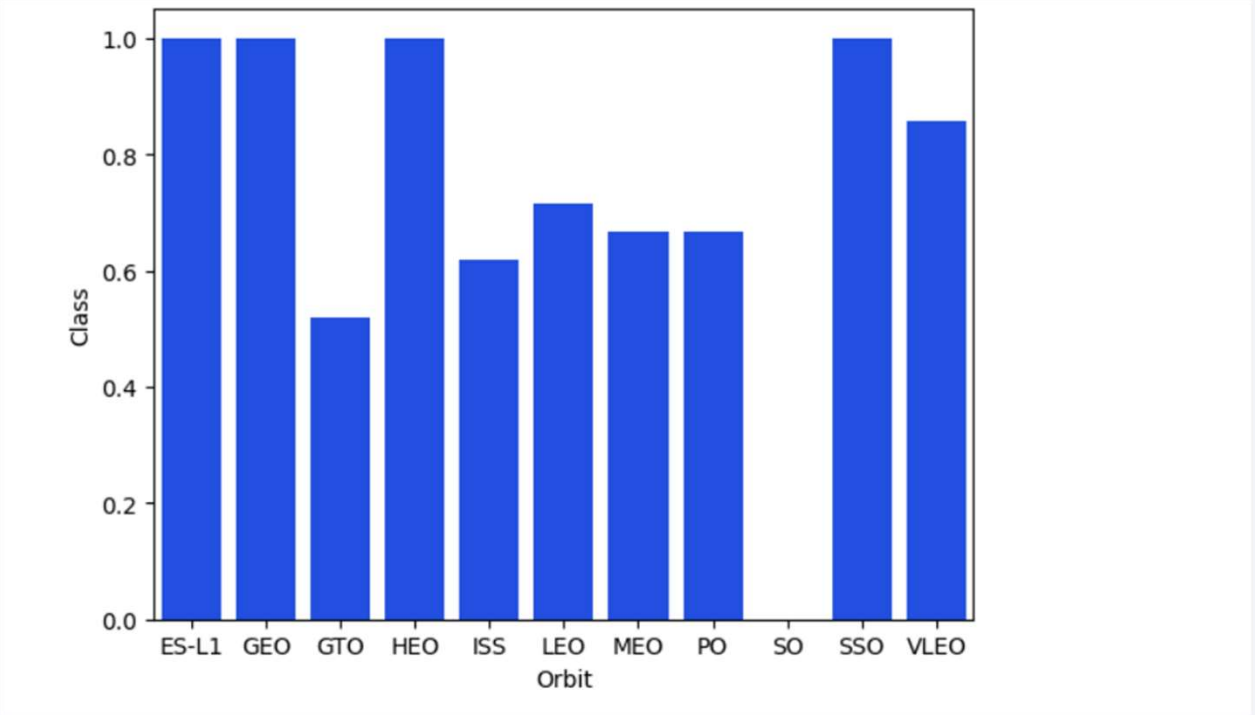
- ES-L1, GEO, HEO, SSO

• Orbits with 0% success rate:

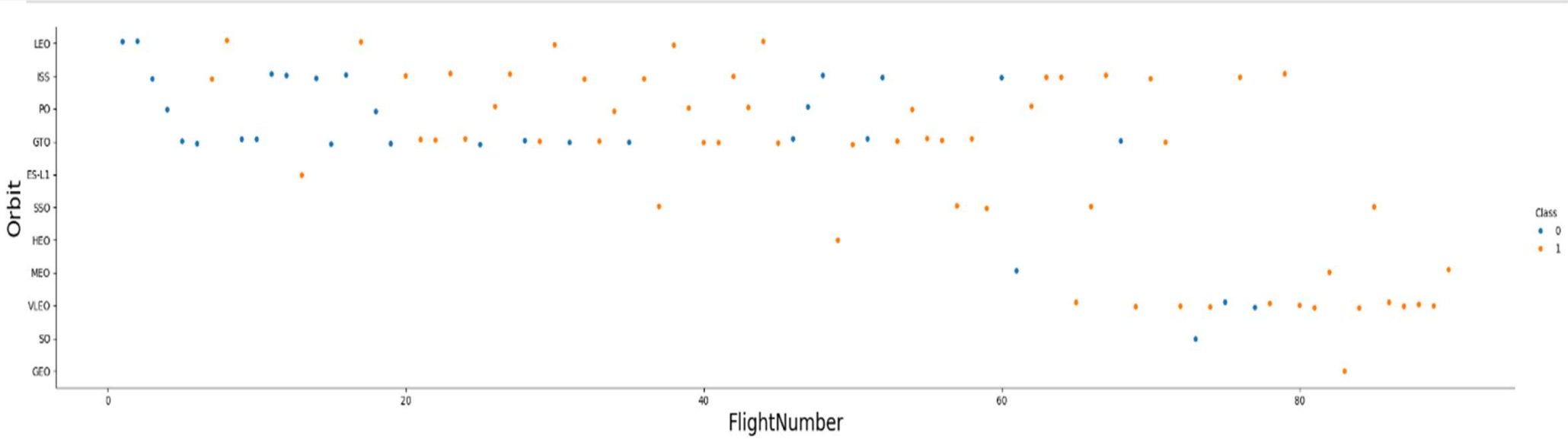
- SO

• Orbits with the success rate between 0% and 100%:

- GTO, ISS, LEO, MEO, PO



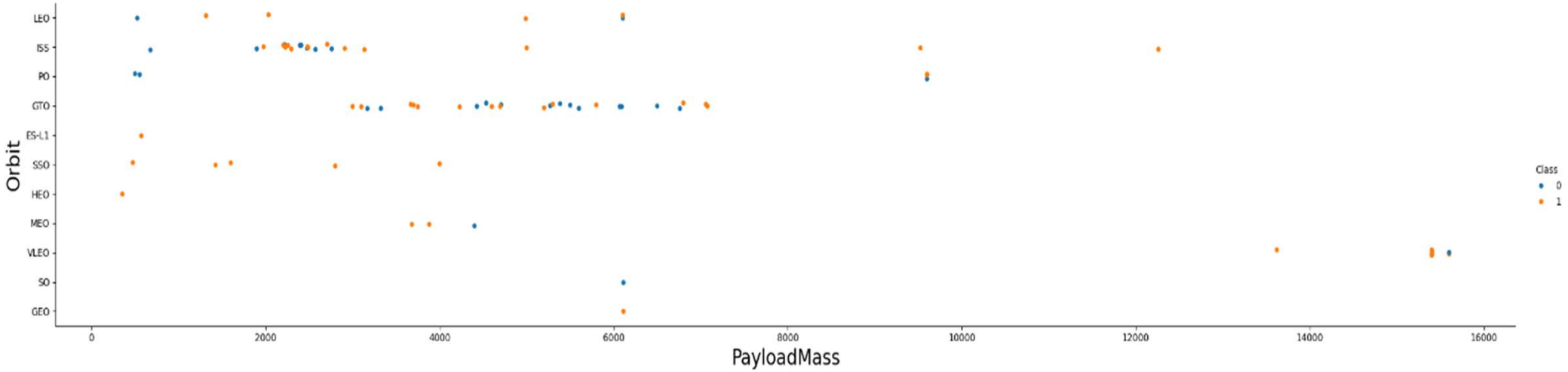
Flight Number vs. Orbit Type



Success appears to have a relationship with the number of flights in the LEO orbit.

There seems to be no relationship between flight numbers and success in GTO orbit.

Payload vs. Orbit Type



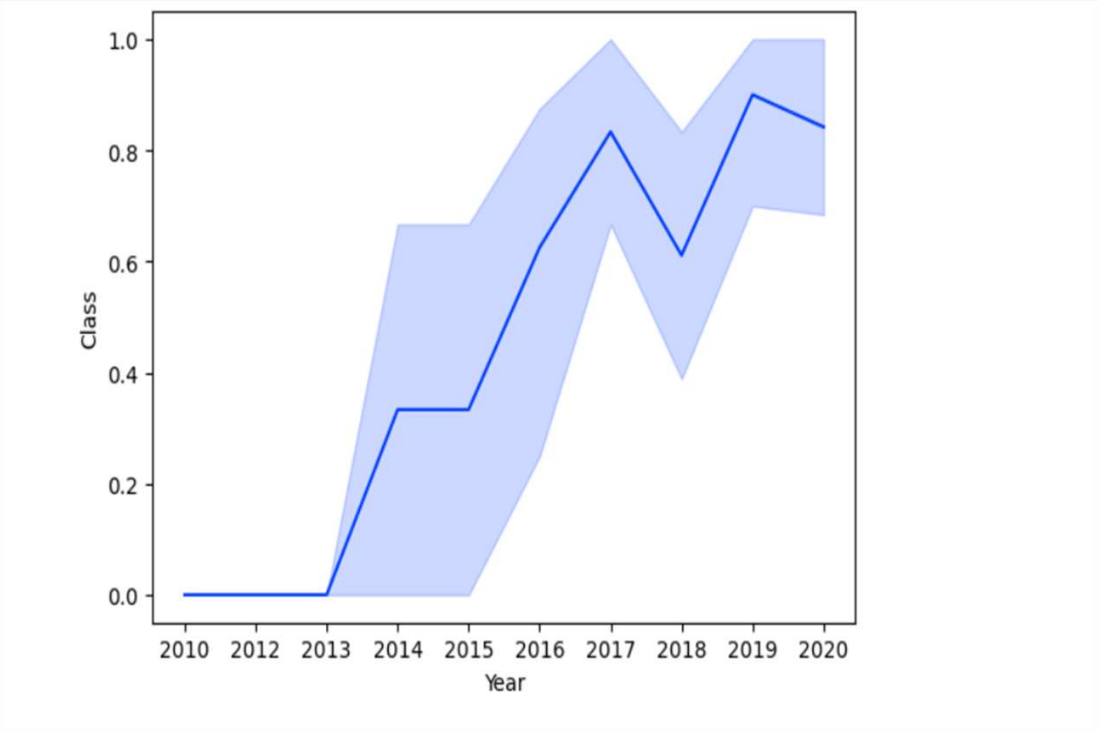
Payload mass correlates with orbit:

- LEO and SSO seem to have relatively low payload mass

- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

- The success rate is observed to be increasing over time.
- There is a decline observed in the year 2018.



All Launch Site Names

- The names of the unique launch sites:

```
In [17]: %sql select distinct Launch_Site from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[17]:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

```
In [18]: %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

Out[18]:

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

- The total payload carried by boosters from NASA:

```
In [30]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like 'NASA (CRS)%'
 * sqlite:///my_data1.db
Done.
Out[30]: sum(PAYLOAD_MASS__KG_)
         48213
```

Average Payload Mass by F9 v1.1

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

In [31]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

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

Out[31]:

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad:

In [37]:
```
%sql select Date from SPACEXTBL where Landing_outcome like '%ground%' order by Date limit 1
```

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

Out[37]:

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
In [38]: %sql select * from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
 * sqlite:///my_data1.db
Done.
```

Out[38]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

In [60]:
```
%sql select substr(Mission_Outcome,1,7) as Mission_Outcome, count(*) from SPACEXTBL  group by 1
```

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

Out[60]:

Mission_Outcome	count(*)
Failure	1
Success	100

Boosters Carried Maximum Payload

- List of the names of the booster which have carried the maximum payload mass

```
In [58]: %sql select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

Out[58]:

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

In [63]:
```
%sql select distinct Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome='Failure (drone shi
```

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

Out[63]:

Landing_Outcome	Booster_Version	Launch_Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
Failure (drone ship)	F9 FT B1020	CCAFS LC-40
Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

- Ranking of 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.

```
In [64]: %sql select Landing_Outcome, count(*) from SPACEXTBL where Date between '2011-06-04' and '2017-03-20' group by Landing_Outcc
 * sqlite:///my_data1.db
Done.
```

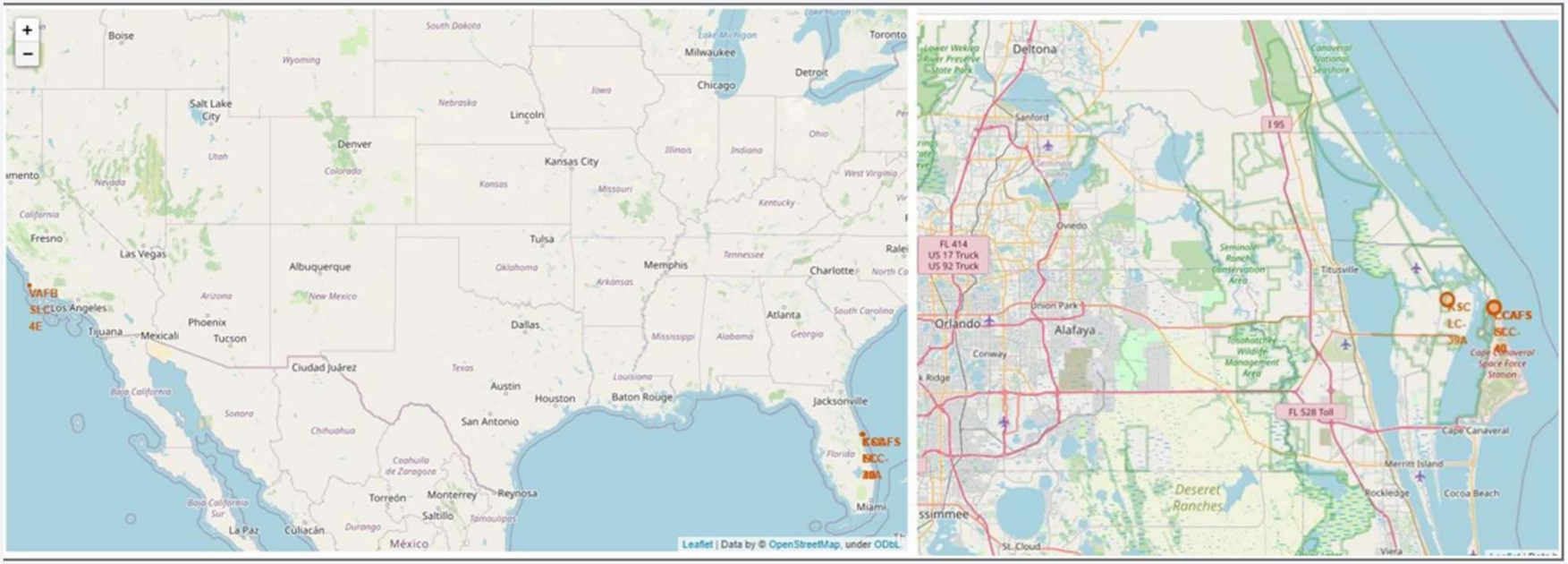
Out[64]:

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

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

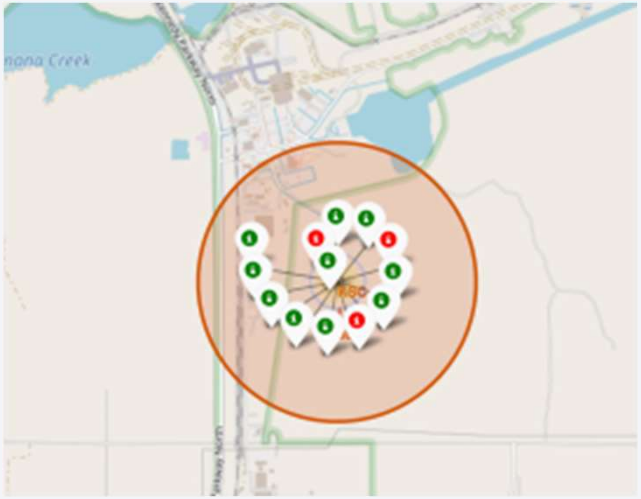


Launch Site Locations in US and a Closer view of Florida Launch site locations.

Most of the launch site locations are close to water.

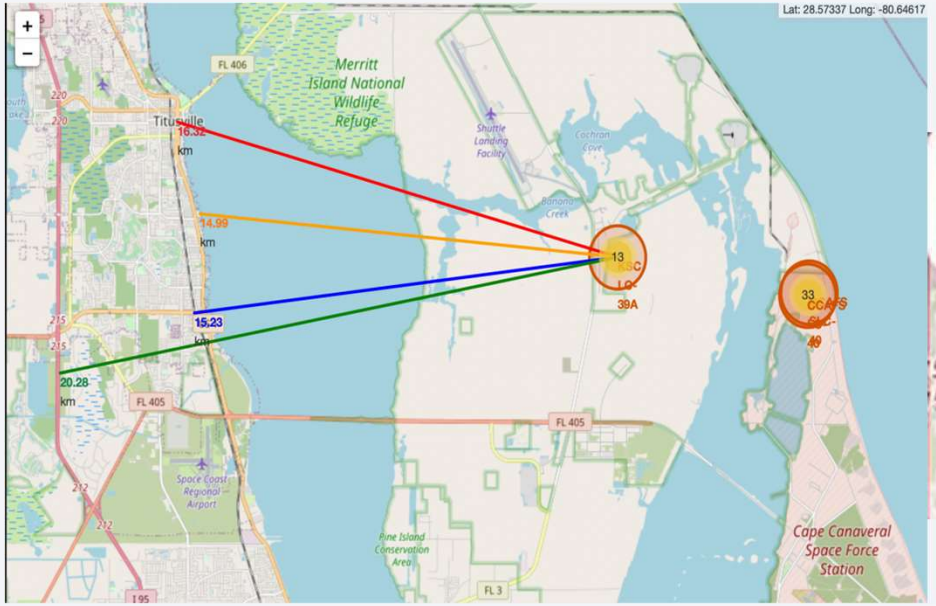
Launch Markers- Color indicated

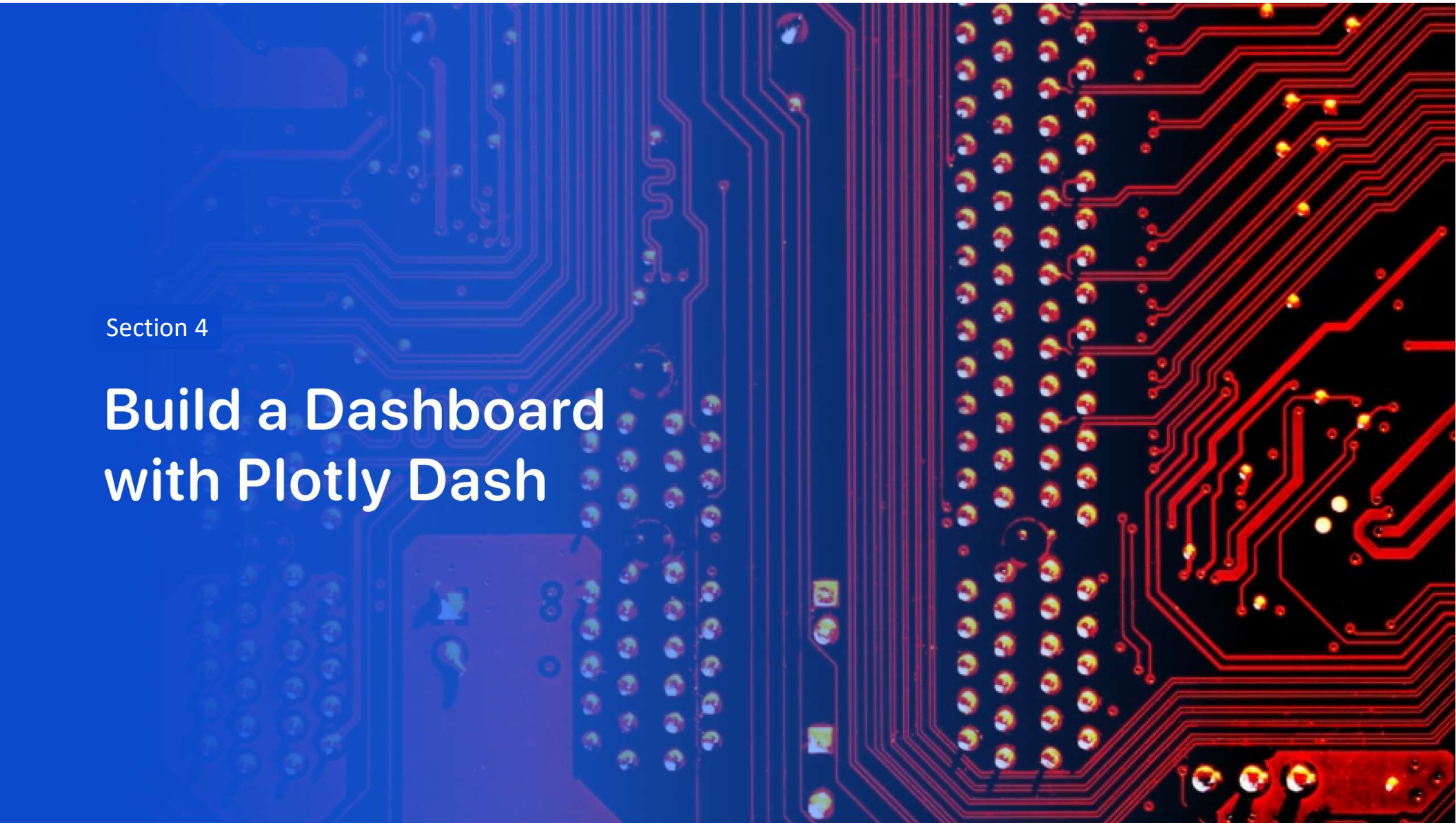
- Color-coded launch markers, green indicating successful Launch while red colored markers indicate a failed attempt at launch.



Distance of Key location to its proximities

- The distance of the key location from its proximities can be easily calculated by hovering or navigating through the key location.

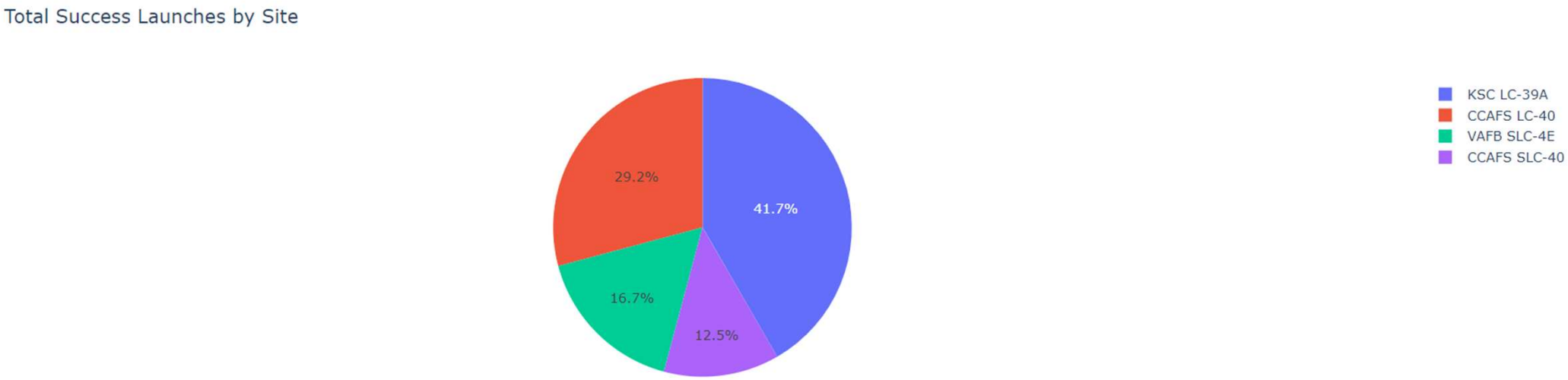




Section 4

Build a Dashboard with Plotly Dash

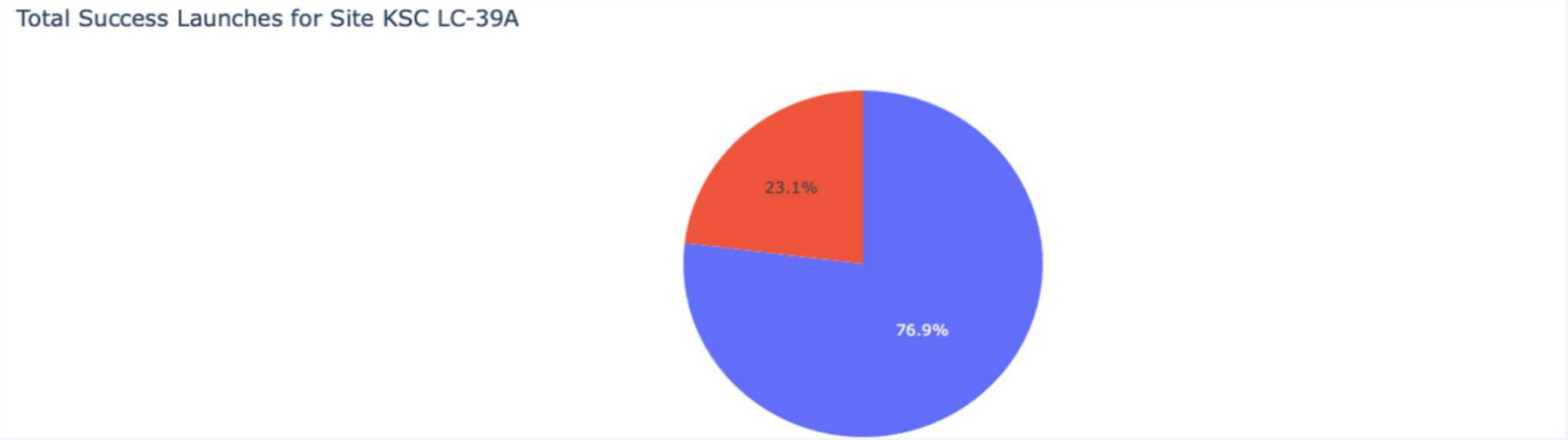
Successful Launches across Launch sites



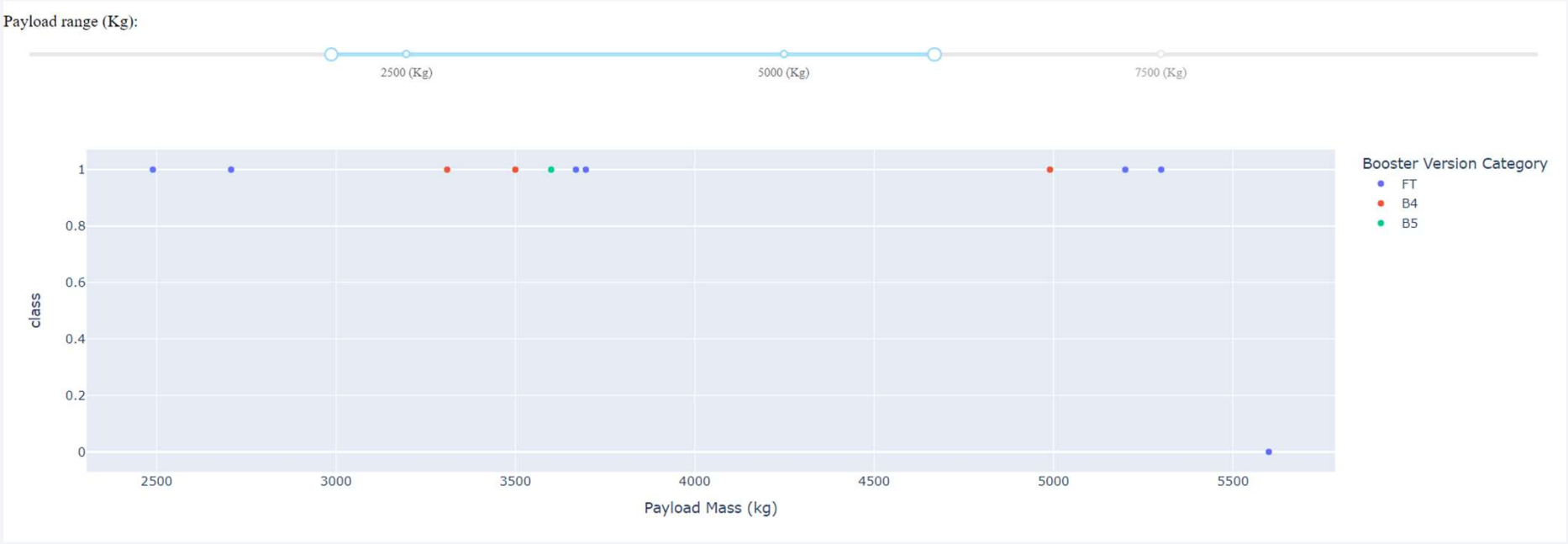
Ration of total no. of Successful launches across different launch sites can be observed in the pie chart.

Launch Site with maximum success ratio

The Launch Site KSC LC-39A has maximum success ratio with success percentage of 76.9%



Payload Mass vs Launch Outcomes



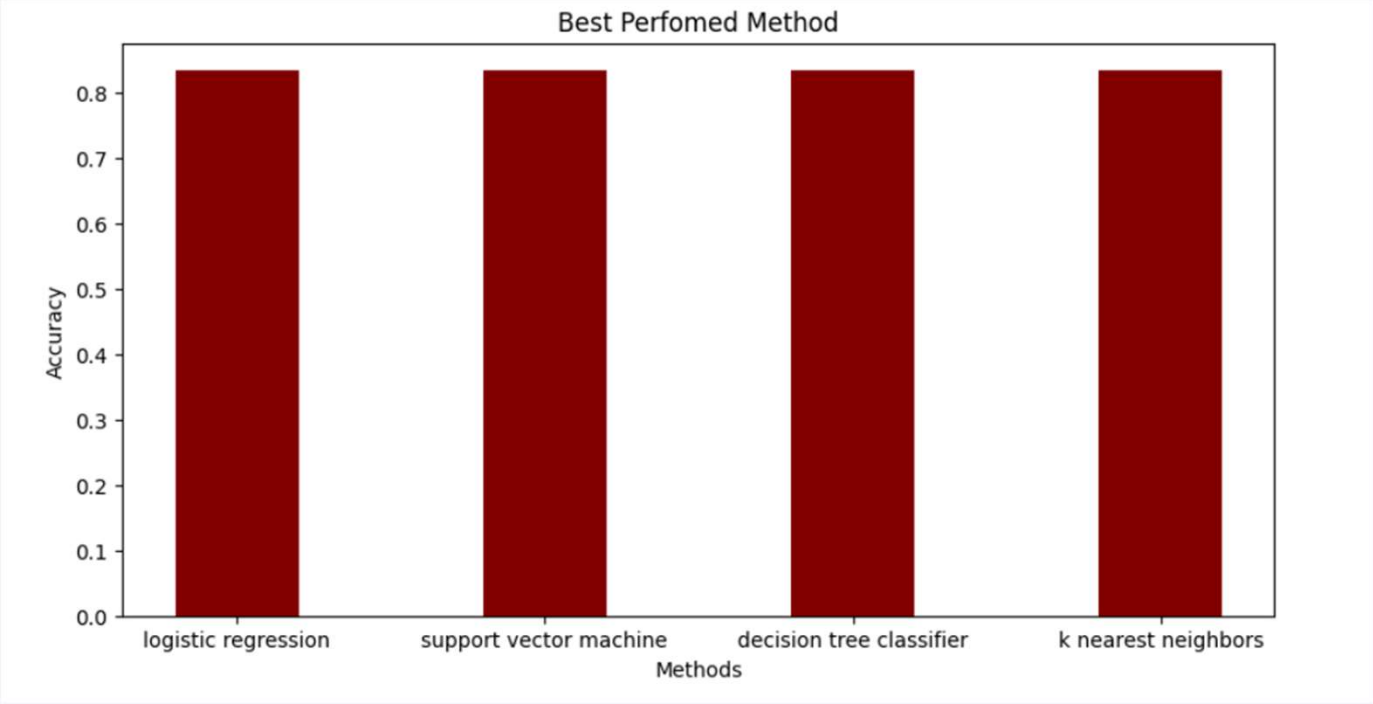
- The charts show the relation between payload mass and launch outcomes, most of the highest values lie in the range of 2500-5500 kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

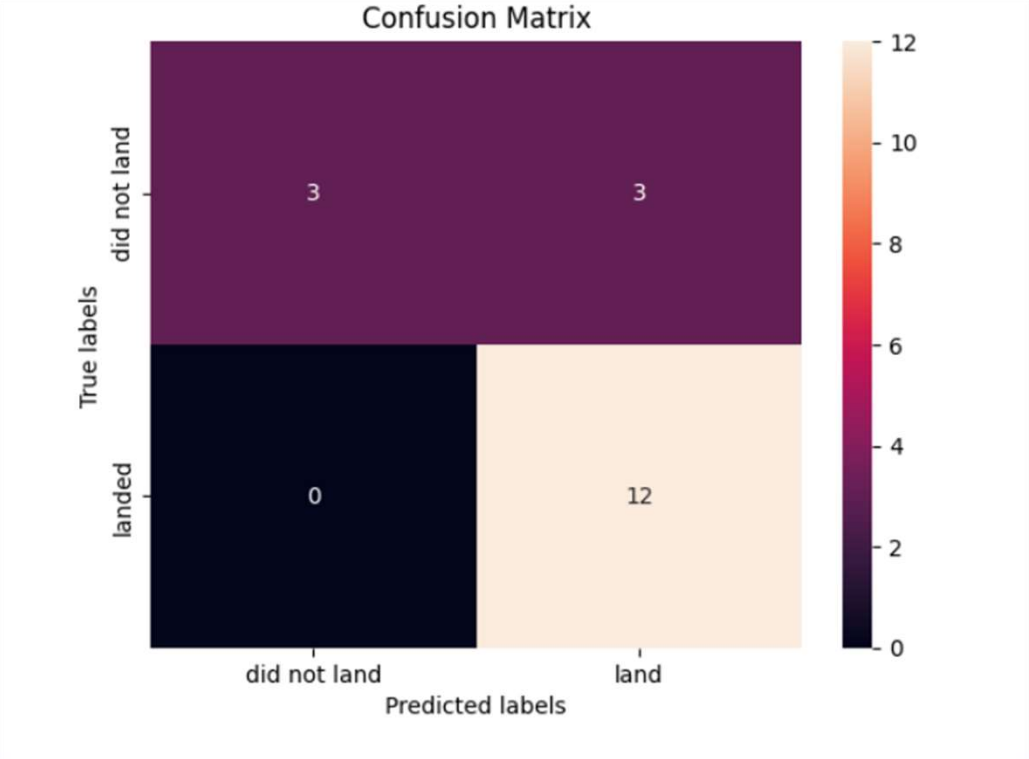
- All models were observed to have had virtually the same accuracy on the test set at 83.33% accuracy. The test size of the data is small at only sample size of 18.



Confusion Matrix

All models had almost similar accuracy for the test set so, the confusion matrix is the same across all models.

- The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).



Conclusions

- Demonstrated expertise in data science and machine learning methodologies using a real-world data set, culminating in a comprehensive report for stakeholders.
- Successfully performed various tasks such as data collection, data wrangling, exploratory data analysis, data visualization, model development, and model evaluation using the SpaceX data.
- Focused on predicting the successful landing of the Falcon 9's first stage using developed machine learning models.
- Constructed different machine learning models, including support vector machines, decision tree classifiers, and k-nearest neighbors.
- Thoroughly evaluated the model results for predictive analysis.
- Compared the strengths and weaknesses of each model, leading to the identification of the optimal model for the task.

Appendix

- Credits and Acknowledgments
- Special thanks to Instructors

Thank you!