Winning Space Race with Data Science

By Aiman Aisamuddin bin Ab Ghapar

21st April 2022

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 and Feature Engineering
  - Launch Sites Location Analysis using Interactive Visual Analytics with Folium and Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive Visual Analytics maps and dashboard
  - Predictive Analytics 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.

Problems to be answered

- What is the factors that contribute to the successful landing of the Falcon 9 first stage?
- The relationships between various features of the datasets to the first stage landing success rate.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- The data was collected through SpaceX REST API and web scraping from Wikipedia using BeautifulSoup

Perform data wrangling

- Landing outcomes are categorized to 1's and 0's to a new column called Class
- Missing values are replaced with their respective mean values
- One-hot encoding is performed on the categorical features

Perform exploratory data analysis (EDA) using visualization and SQL

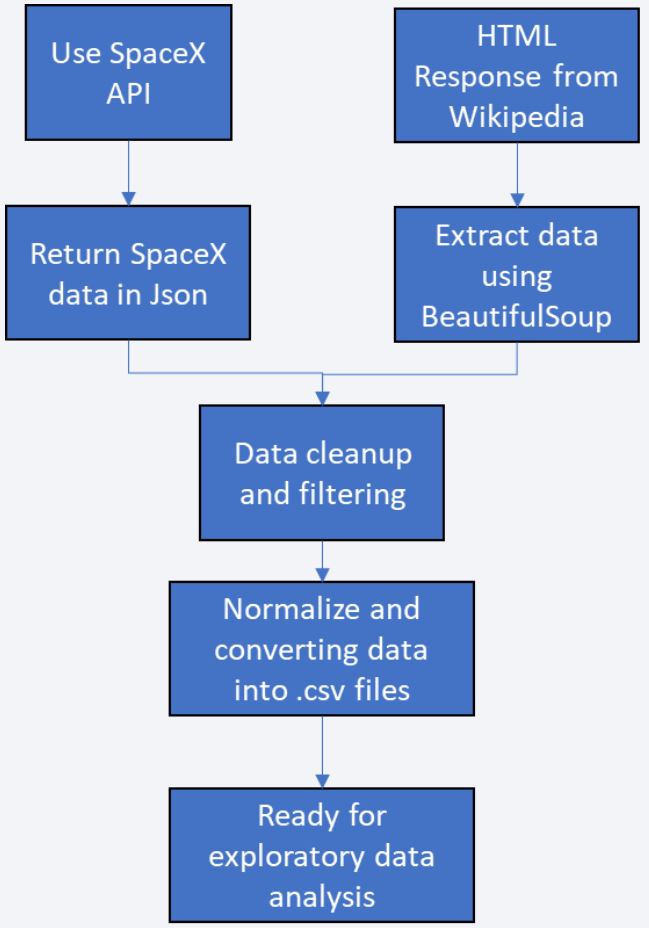
Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- The data are split into train and test datasets
- Classifications models: 1) Logistic Regression, 2) SVM, 3) Decision Tree, and 4) K Nearest Neighbors are used
- GridSearchCV is used on each models to the determine the best parameters

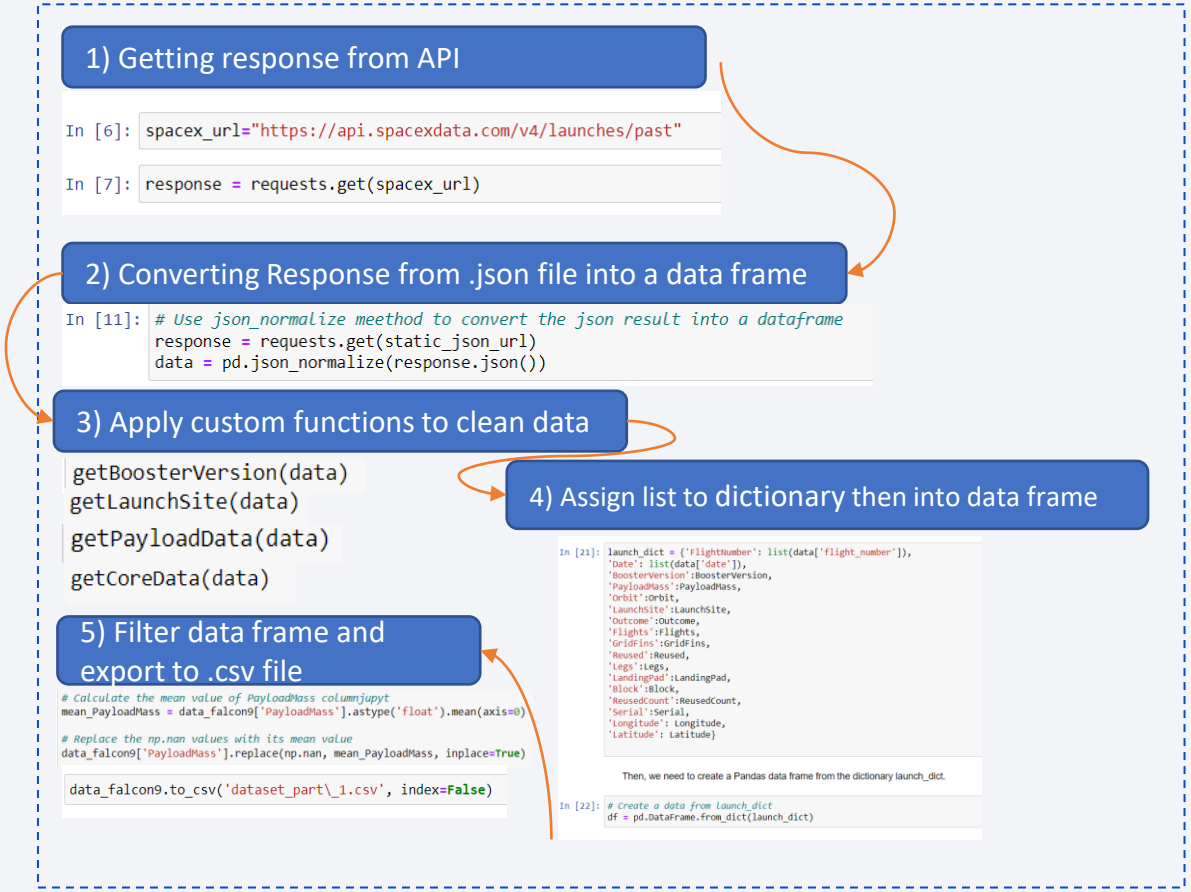
Data Collection

- In this project, data is collected from the SpaceX launch data that is gathered from the SpaceX REST API.
- This API contains launch data including information about the rocket used, booster version, payload mass, launch specifications, landing specifications and landing outcome.
- Launch data of SpaceX Falcon 9 also collected from the Wikipedia using Beautiful Soup



Data Collection – SpaceX API

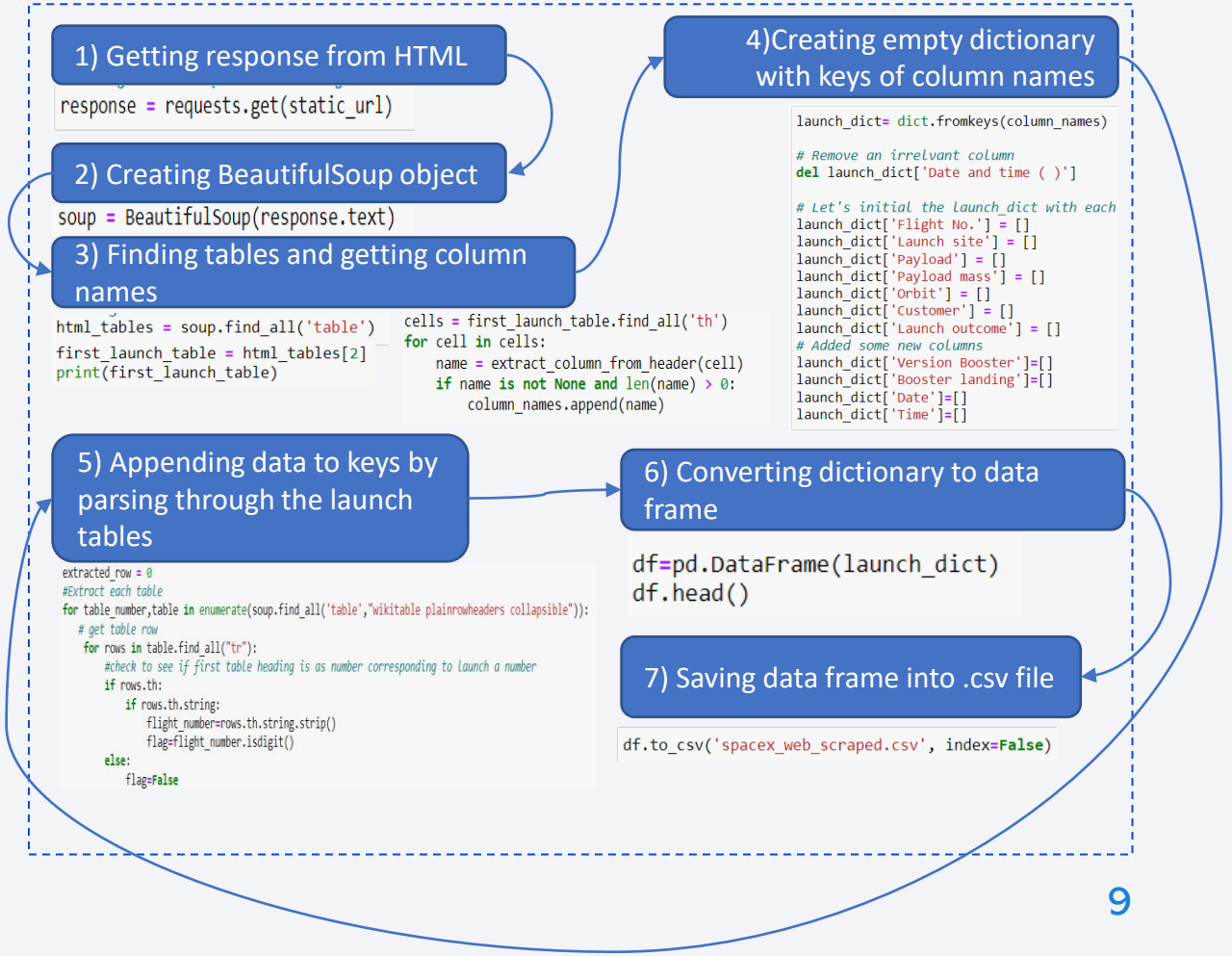
- Data collection with SpaceX REST calls



GitHub URL:https://github.com/aiman-aisa/Applied-Data-Science-Capstone-Project/blob/a94c3c99b833606d99bfa5a4502e58d1175c3988/Data%20Collection%20API.ipynb

Data Collection - Scraping

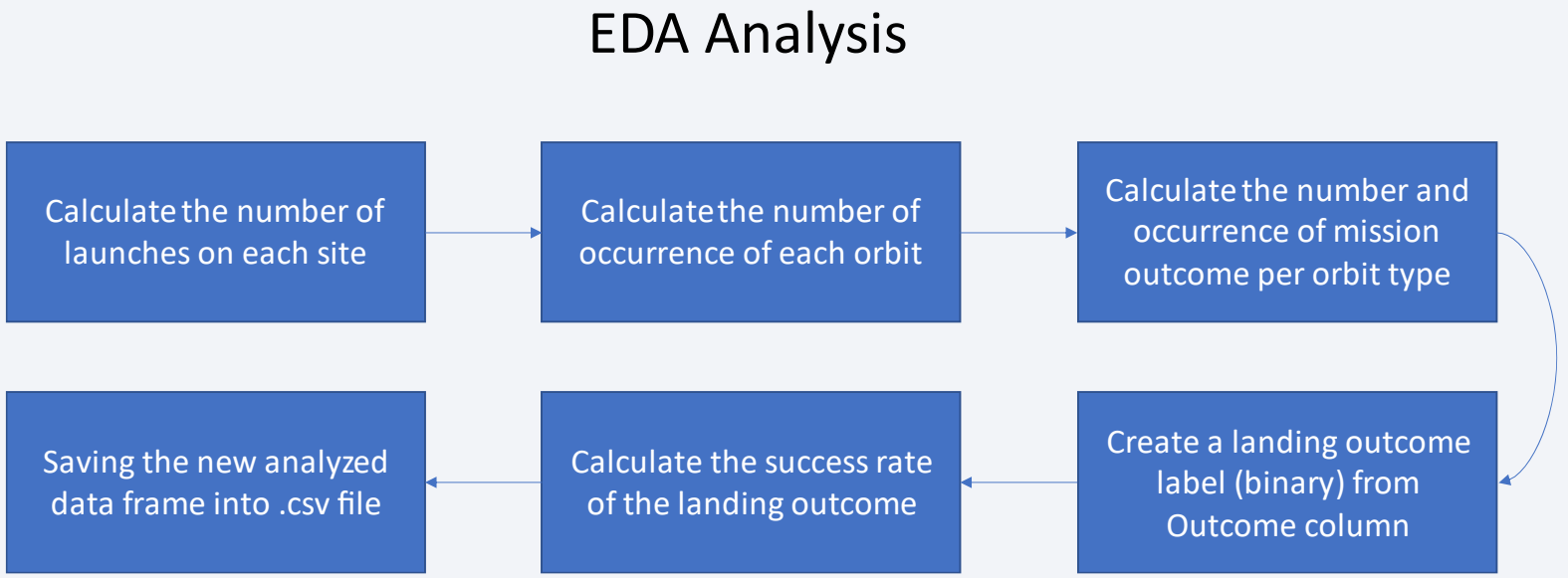
- Web scrapping from Wikipedia using Beautiful Soup



GitHUB URL: https://github.com/aiman-aisa/Applied-Data-Science-Capstone-Project/blob/a94c3c99b833606d99bfa5a4502e58d1175c3988/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

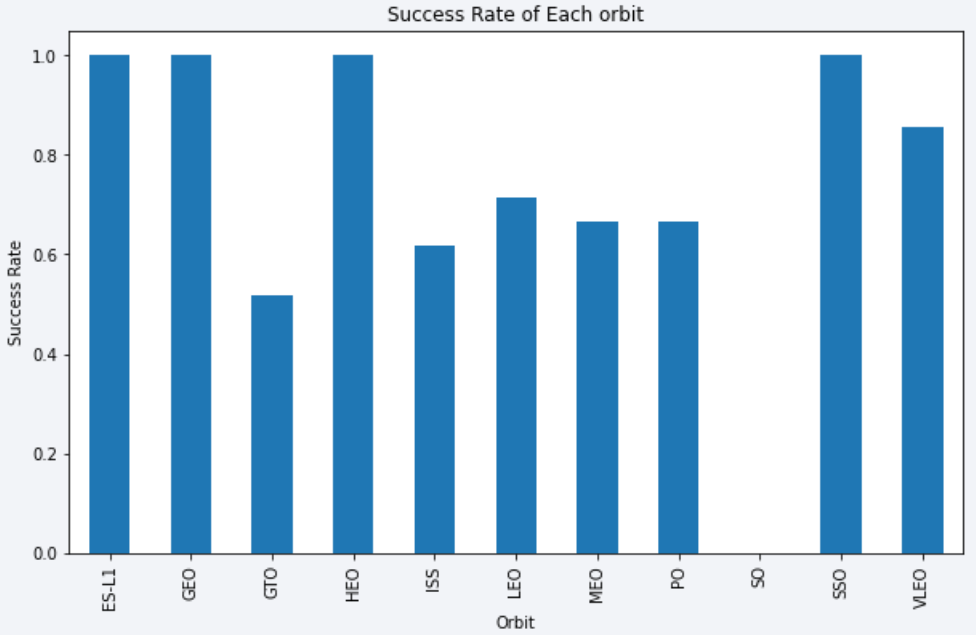
- The null data are replaced by the mean of the respective data. Then EDA analysis is initially performed as below:

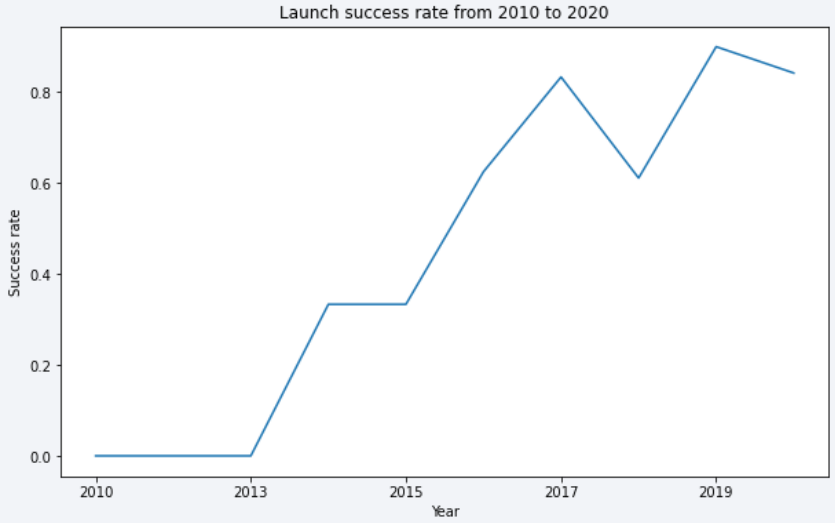


Link to the notebook:https://github.com/aiman-aisa/Applied-Data-Science-Capstone-Project/blob/a94c3c99b833606d99bfa5a4502e58d1175c3988/EDA.ipynb

EDA with Data Visualization

- Scatter plot is used to visualize the relationship between various parameters with the landing outcomes of the launch
- Line plot and bar charts are also used to visualize the success rate of the landing of each orbit type and the success rate yearly trend





Link to the notebook: https://github.com/aiman-aisa/Applied-Data-Science-Capstone-Project/blob/a94c3c99b833606d99bfa5a4502e58d1175c3988/EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- SQL queries you performed includes:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launch by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass using subquery
  - List the failed landing outcomes in drone ship, their booster versions and launch site names for in year 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
- Link to the notebook: https://github.com/aiman-aisa/Applied-Data-Science-Capstone-Project/blob/a94c3c99b833606d99bfa5a4502e58d1175c3988/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Using Folium, each launch sites are marked using markers, circles, lines and info signs.
- Circles are used to marked the coordinates of the launch sites and markers are used to write the launch sites names.
- We also add marker clusters on each sites consisting of the outcomes of the launch, where green as being successful launch and red otherwise.
- Finally, lines and distance markers are made to indicates the distance of the launch sites to the important features in its proximity such as railways, highways, cities and coastlines.

Link to notebook: https://github.com/aiman-aisa/Applied-Data-Science-Capstone-Project/blob/a94c3c99b833606d99bfa5a4502e58d1175c3988/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

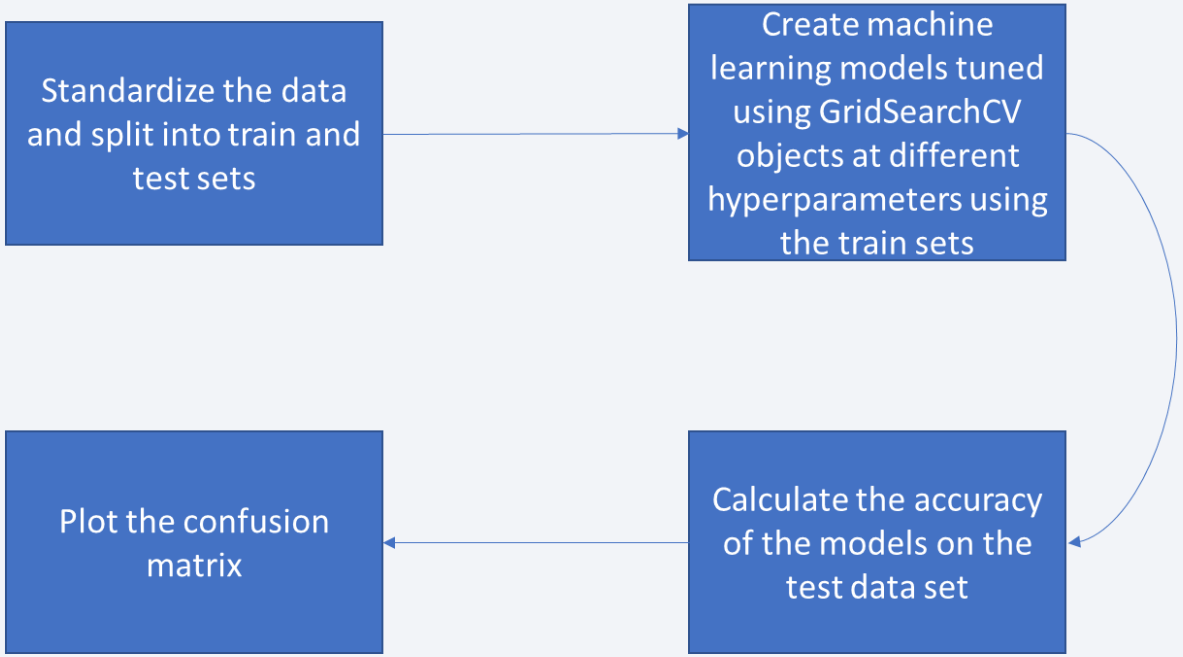
Build a Dashboard with Plotly Dash

- Using Plotly Dash, an interactive visualization board are created to visualize the findings of the launch sites.
- Using pie charts, success rate of the SpaceX launch sites are visualized.
- Scatter graphs are created to visualize the relationship between the success rates of the launch with the payload mass for different booster version.

The following is the link to the Python code to create the Plotly Dash: https://github.com/aiman-aisa/Applied-Data-Science-Capstone-Project/blob/a94c3c99b833606d99bfa5a4502e58d1175c3988/spacex_dash_app.py

Predictive Analysis (Classification)

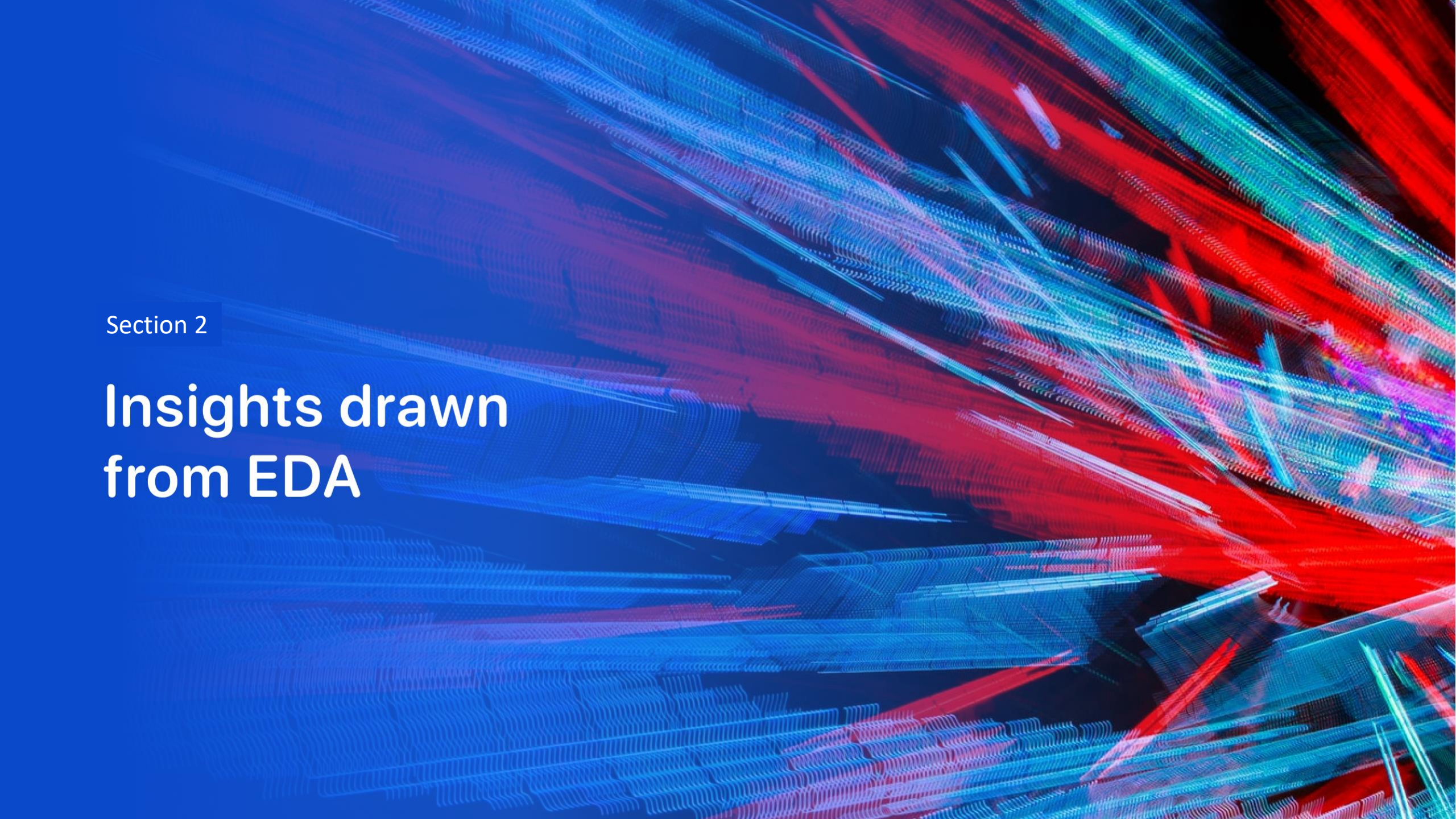
- The data are first loaded using numpy and pandas. Then they are standardize and splitted into test and train sets.
- In training the train sets, GridsearchCV objects are created for each machine learning models to tune the models at different hyperparameters.
- The accuracy of each models are computed using the test sets and compared.

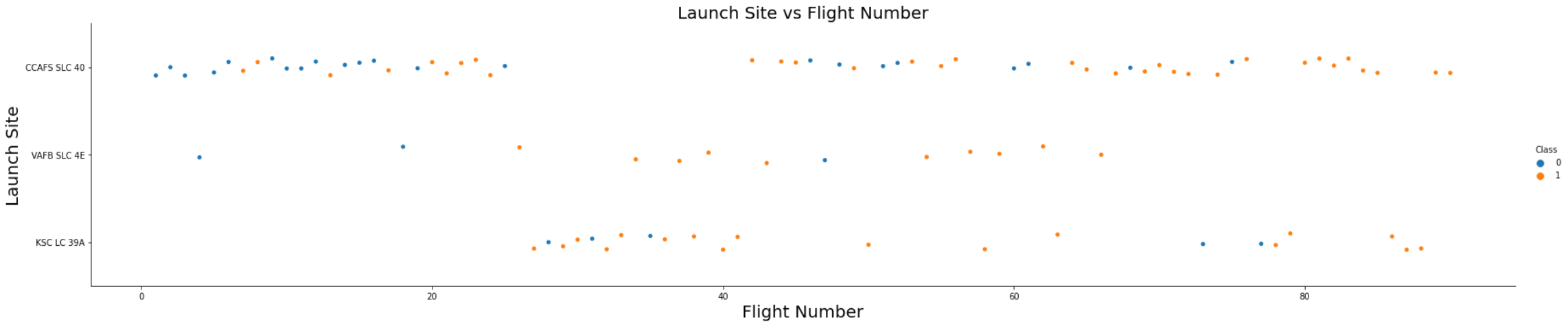


Link to Notebook: https://github.com/aiman-aisa/Applied-Data-Science-Capstone-Project/blob/a94c3c99b833606d99bfa5a4502e58d1175c3988/Machine%20Learning%20Prediction.ipynb

Results

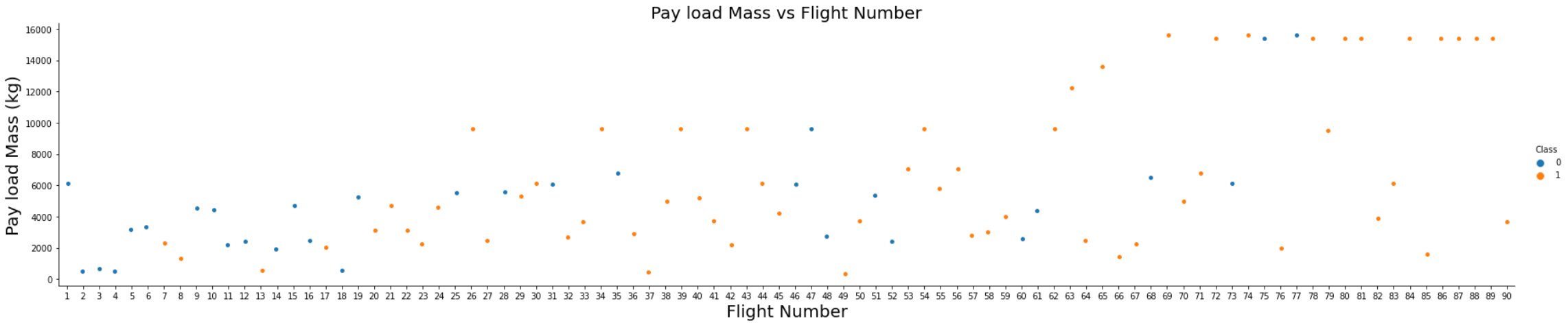
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA



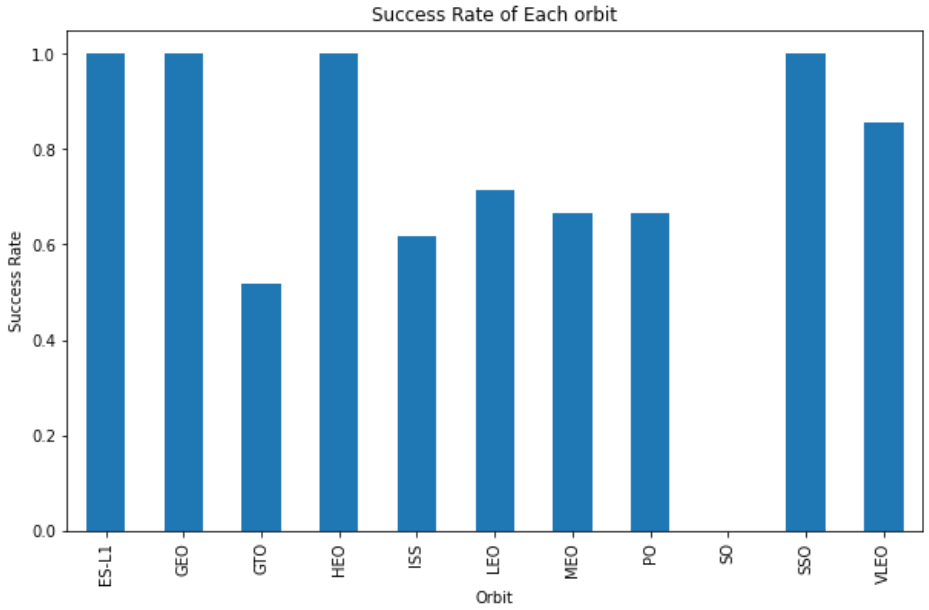
Flight Number vs. Launch Site

- The launch sites have different success rates. CCAFS LC-40 has a success rate of 60% while KSC LC-39A and VAFB SLC 4E has a success rate of about 77%



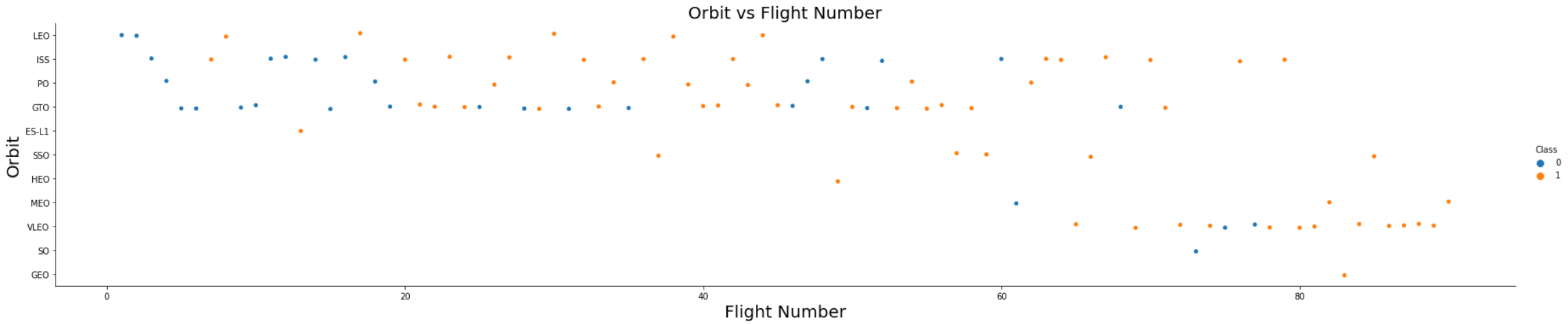
Payload vs. Launch Site

- As the flight number increases, the first stage is more likely to land successfully
- Plus, it seems that as the payload mass increases, the less likely the first stage will return



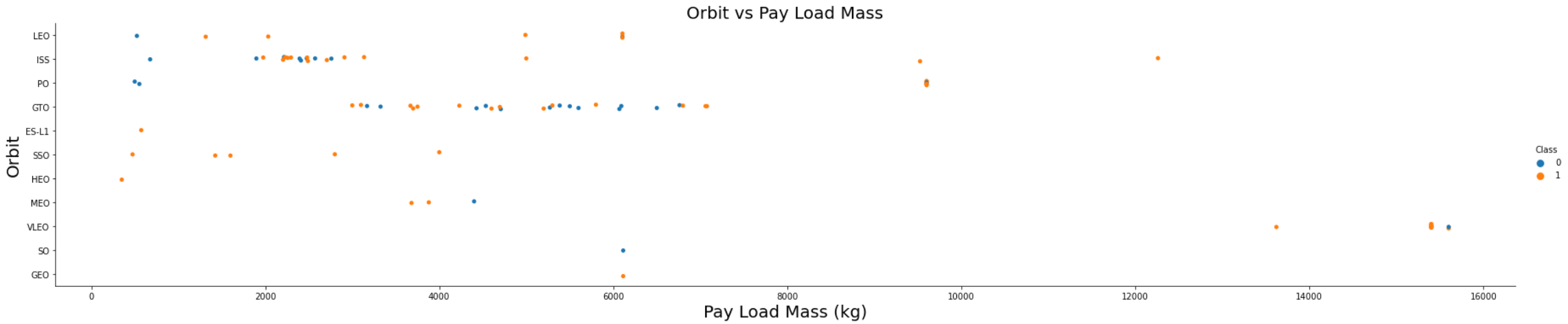
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO all have 100% launch success rate.
- SO orbit has 0% launch success rate



Flight Number vs. Orbit Type

- The launch success rate of at the LEO orbit seems to show a relation with the number of flights.
- There seems to be no relationship between flight number when in GTO orbit

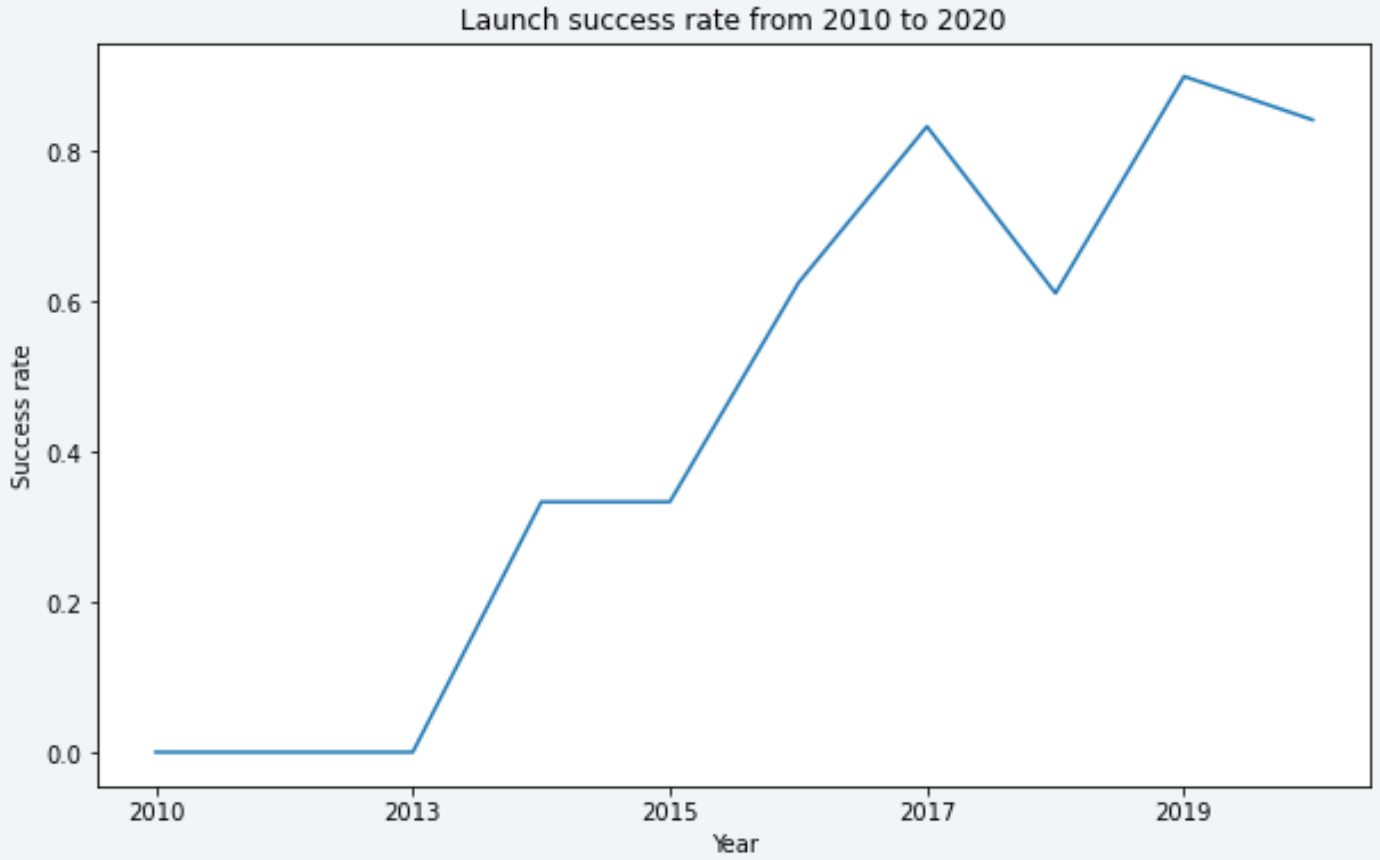


Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS orbits
- The GTO orbit does not show any relation between success rate and the payload mass

Launch Success Yearly Trend

- The launch success rate shows and increasing trend from 2013 until 2020



All Launch Site Names

- SELECT DISTINCT statement is used to get all the launch site names

```
In [20]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXDATASET;
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0fbb
Done.
```

Out[20]:

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

In [21]:
```
%sql SELECT * FROM SPACEXDATASET \
WHERE LAUNCH_SITE LIKE 'CCA%' \
LIMIT 5;
```

```
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

Out[21]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- LIKE operator with % is used to select launch site names begin with ‘CCA’

Launch Site Names Begin with 'CCA'

```
In [31]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXDATASET \
         WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0fbb
Done.
```

Out[31]:

1
45596

- SUM() function is used to calculate the total payload mass

Total Payload Mass

Average Payload Mass by F9 v1.1

- AVG() functions is used to determine the average payload mass carried by booster version F9 v1.1.

```
In [30]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXDATASET \
         WHERE BOOSTER_VERSION = 'F9 v1.1';
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0fbb7
Done.
```

Out[30]:

1
2928

In [32]:
```
%sql SELECT MIN(DATE) FROM SPACEXDATASET \
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0
Done.
```

Out[32]:

1
2015-12-22

- MIN() functions is used on the DATE column to determine the dates of the first successful landing outcome on ground pad.

First Successful Ground Landing Date

Successful Drone Ship Landing with Payload between 4000 and 6000

- WHERE and BETWEEN operator is 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.

In [33]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXDATASET \
WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0
Done.
```

Out[33]:

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

In [41]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_OUTCOME \
FROM SPACEXDATASET \
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0
Done.
```

Out[41]:

mission_outcome	total_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- COUNT() function is used to calculate the total number of successful and failure mission outcomes.
- Group by operator is used to present the result grouped by the mission outcome

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

- Subquery is used to list the names of the booster which have carried the maximum payload mass.

In [43]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXDATASET \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0l
Done.
```

Out[43]:

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

In [45]:
```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEXDATASET \
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND \
DATE LIKE '2015%'
```

```
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.data
Done.
```

Out[45]:

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

2015 Launch Records

- LIKE and WHERE operators are used to list 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

- COUNT() functions is used to count the landing outcomes
- WHERE and BETWEEN operator is used to get the result between the specific dates
- GROUP by operator is used to group the result based on the landing outcome
- ORDER BY operator is used to order the result in descending

In [68]:
```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT FROM SPACEXDATASET \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC;
```

```
 * ibm_db_sa://zdq79269:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.da
Done.
```

Out[68]:

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

Section 3

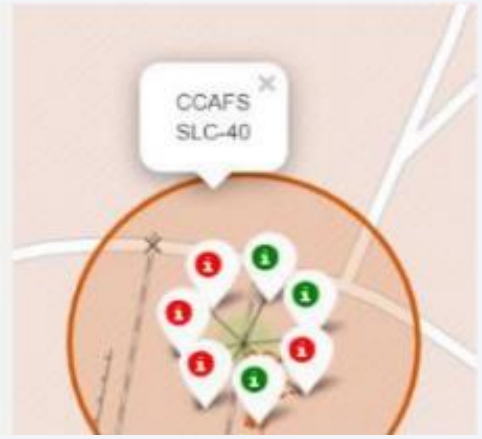
Launch Sites Proximities Analysis

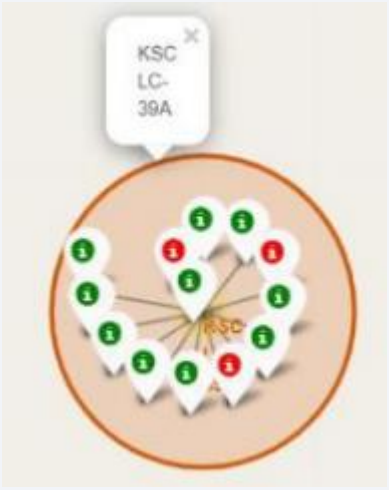
SpaceX Launch Sites Map Markers

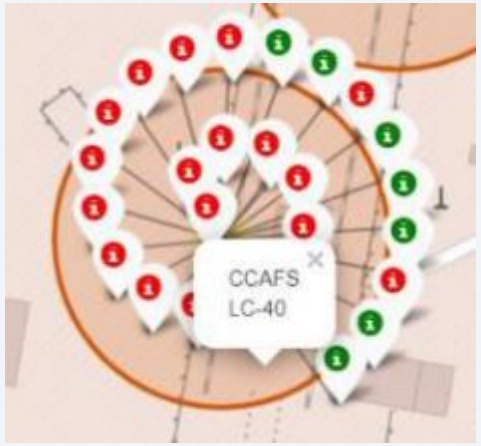
We can see that the three of the launch sites are in Florida and the other one is California

Color labeled launch outcomes

Florida Launch Sites



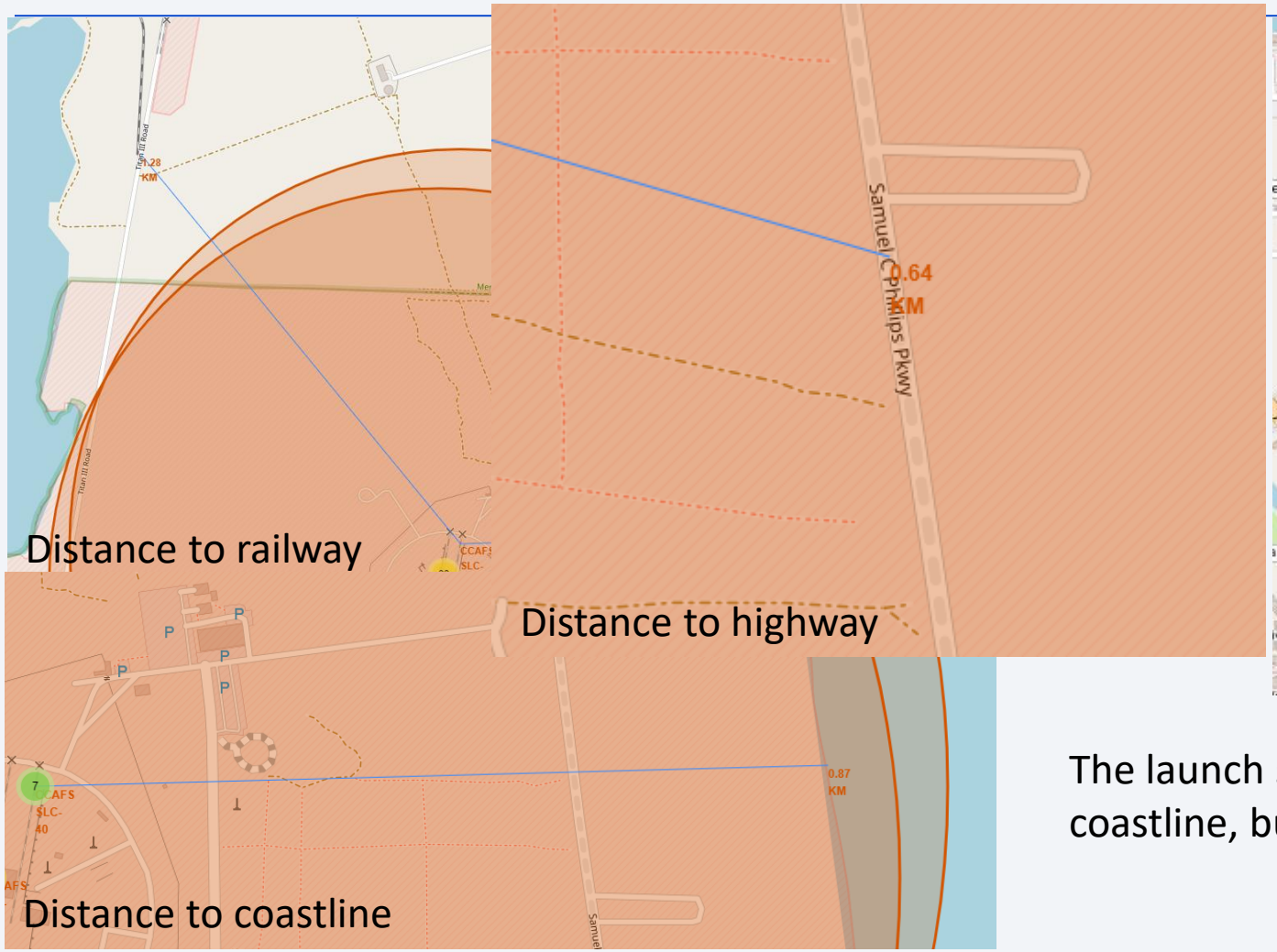


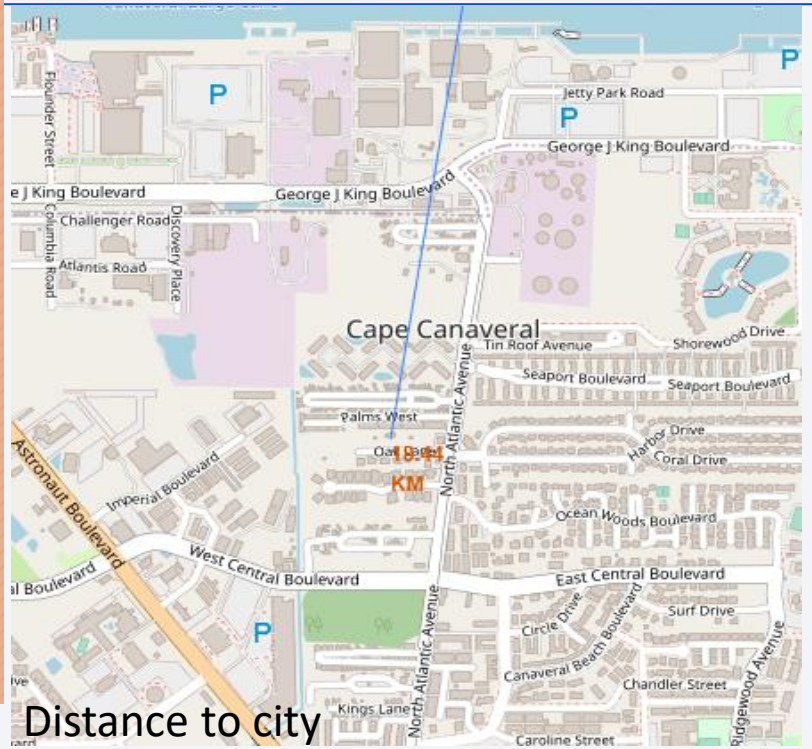


California Launch Sites

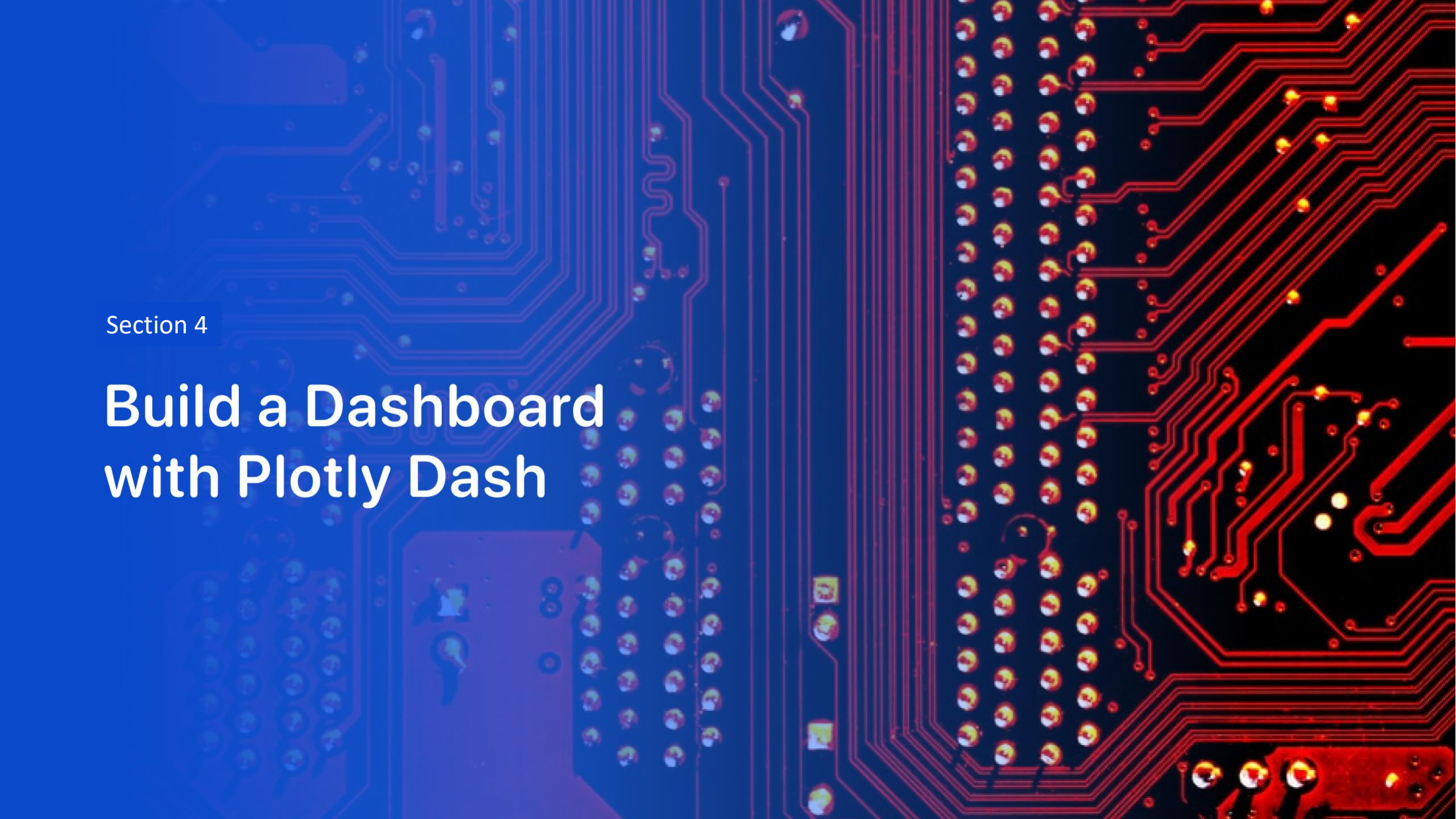
Green marker shows successful launches and Red marker shows failed launches

Launch Site Distance to landmarks

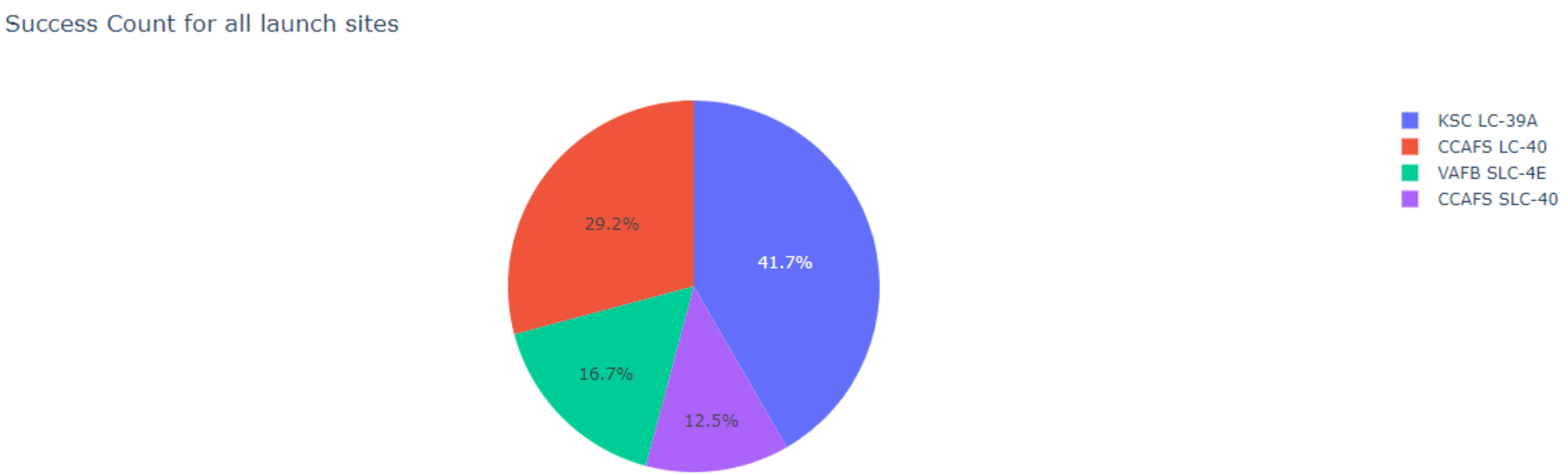




The launch site are near the road, railway and coastline, but keep a safe distance to the city

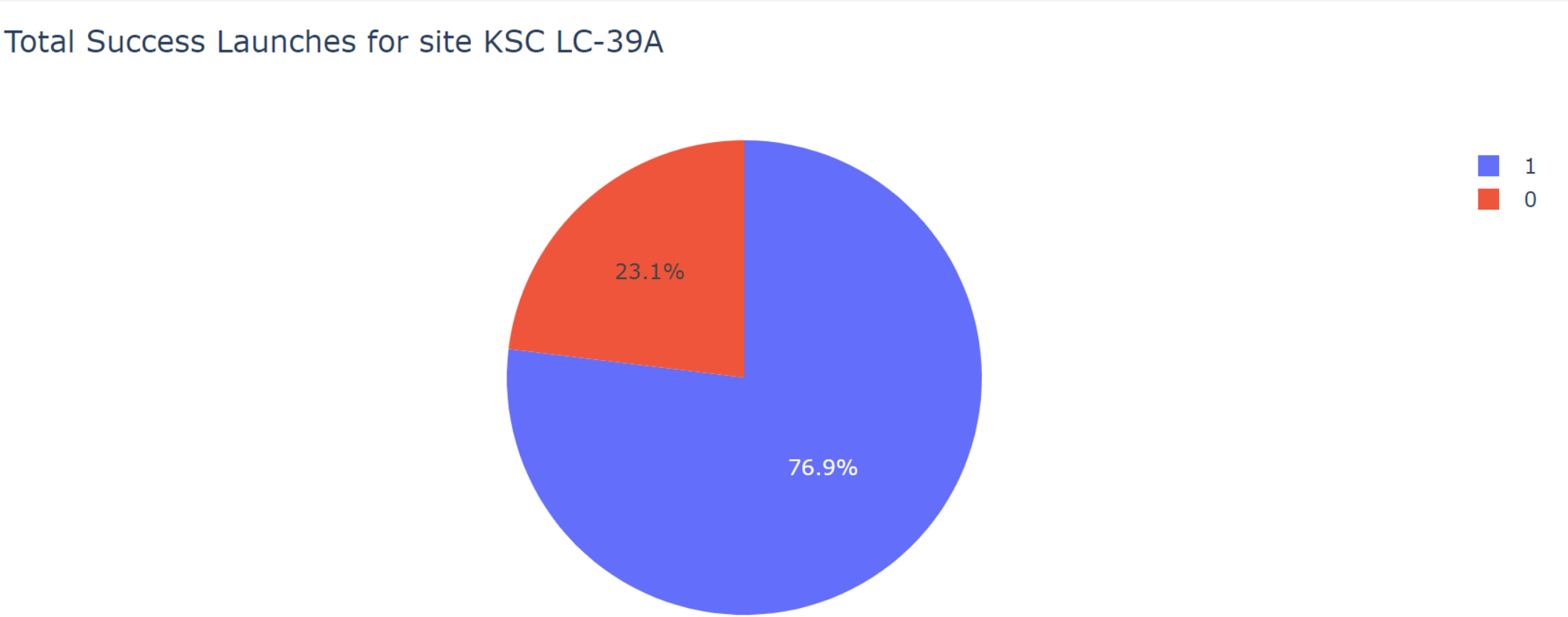

Section 4
Build a Dashboard with Plotly Dash

Launch success rate for all launch sites



KSC LC-39A has the highest launch success rate

Pie chart at the highest success rate launch site

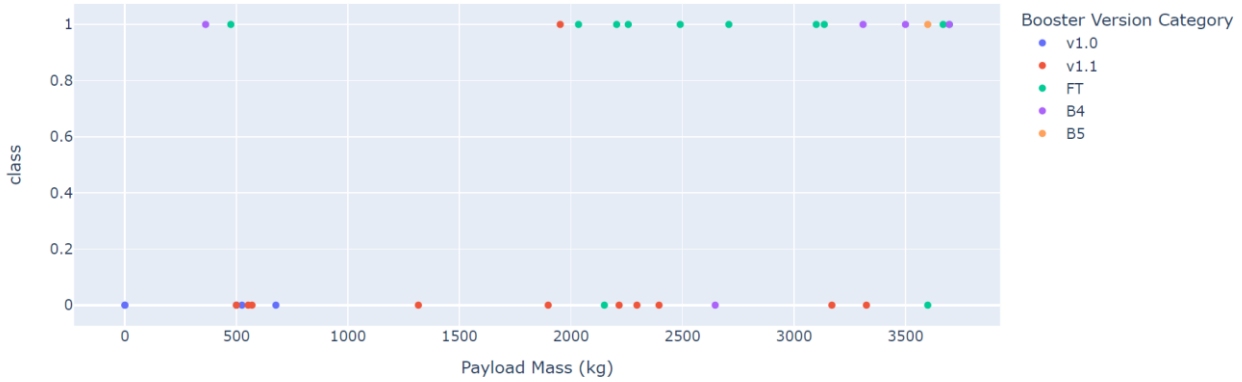


KSC LC-39A has 76.9% successful landing and 23.1% failures

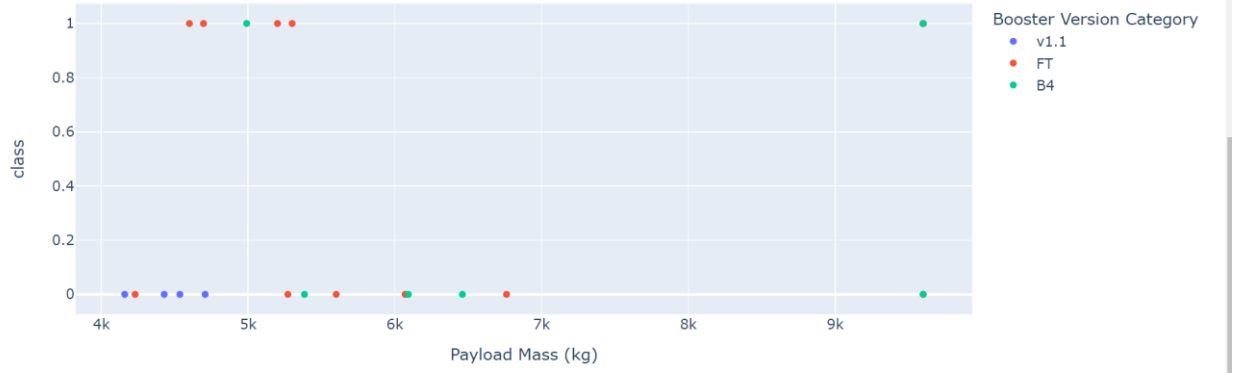
Payload vs Launch Outcome Scatter Plot

- The success rate for low weighted payload is higher than the heavy weighted payload
- Indirectly, we found that, the booster version of v1.0 and B5 are not used in heavy weighted payload launches

Low Weighted Payload (0 – 4000 kg)



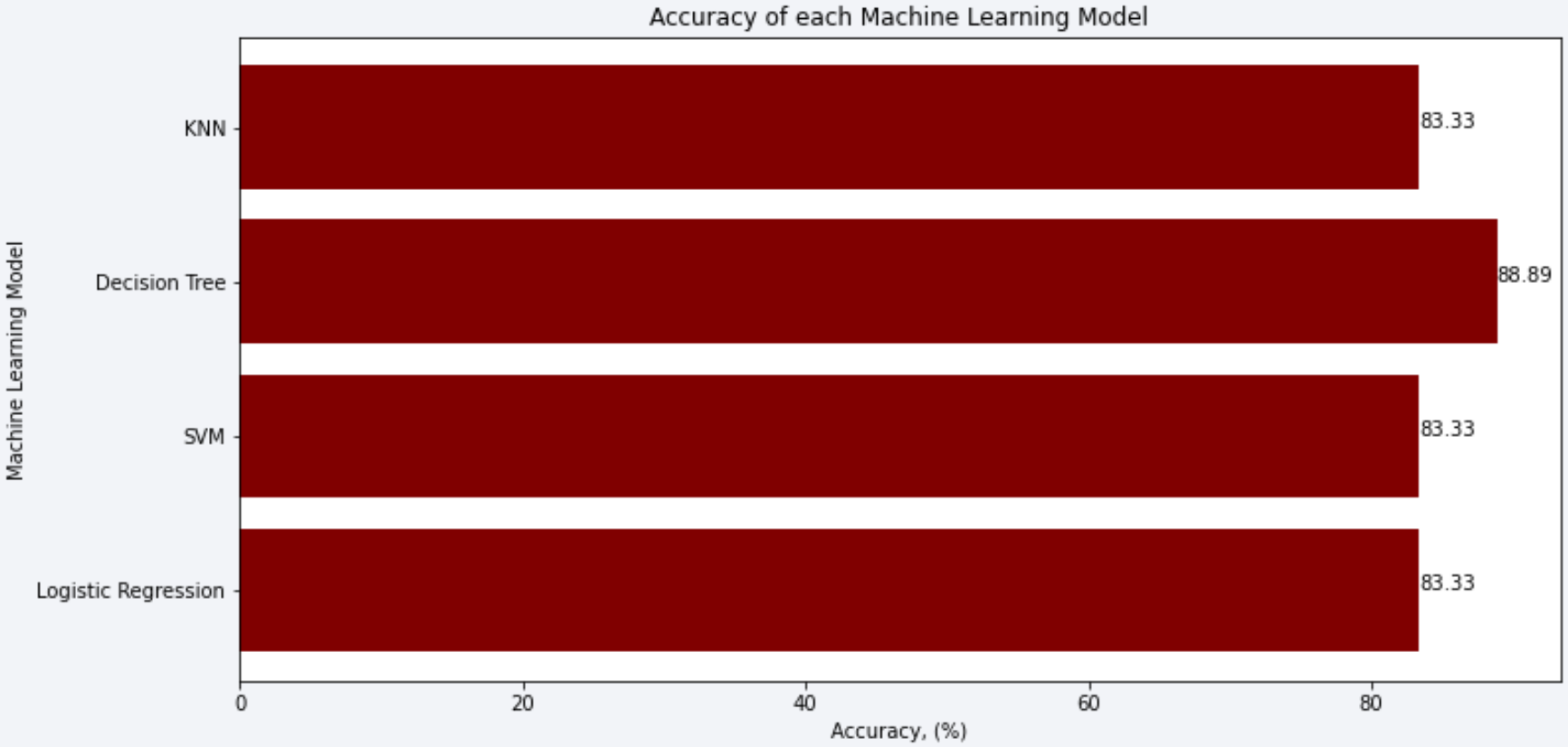
Heavy Weighted Payload (4000 kg – 10000kg)



Section 5

Predictive Analysis (Classification)

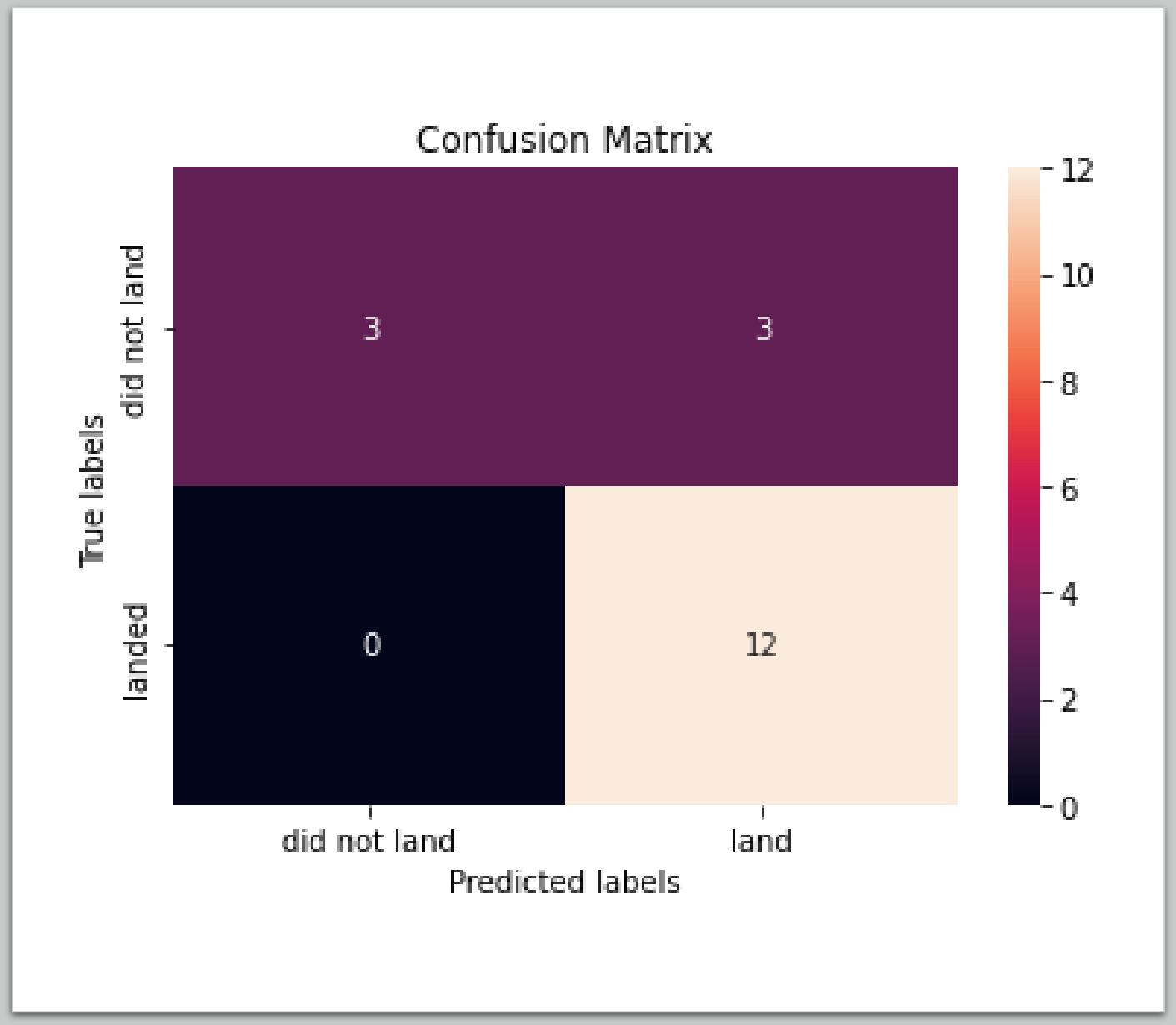
Classification Accuracy



Decision tree classification model has the highest accuracy in predicting the landing outcomes

Confusion Matrix

- The figure shows the confusion matrix of the best classifier for this project which is the decision tree.
- The decision tree classifier successfully predict correctly the successful landing outcome
- However, the major problem is the false positives where the classifier marked the unsuccessful landing as successful here



Conclusions

- As the flight number increases, the first stage is more likely to land successfully.
- As the payload mass increases, the less likely the first stage will return.
- The launch success rate shows and increasing trend from 2013 until 2020.
- The launch site are near the road, railway and coastline, but keep a safe distance to the city.
- The success rate for low weighted payload is higher than the heavy weighted payload.
- The decision tree classifier is the best machine learning model to be used in this predicting the landing outcome of the launches.

Thank you!