IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cheng Zhang
02-19-2023

Outline

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

Executive Summary

Summary of methodologies:

Using SpaceX API and web scraping to do data collection
Exploratory Data analysis : data wrangling, data visualization and interactive visual analytics
Prediction using different Machine learning models: Logistic Regression, SVM, KNN, Decision Tree.
Cross Validation technique: GridSearchCV

Summary of results:

Prediction using machine learning models showed the best model to predict which characteristics are important to drive this opportunity by the best way.

Introduction

- Project background and context
  - The goal is to predict if the Falcon 9 first stage will land successfully.
- Problems you want to find answers
  - 1. we can determine the cost of a launch.
  - 2. where is the best place to make launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scraping
- Perform data wrangling
  - Dropping unnecessary columns
  - One hot encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

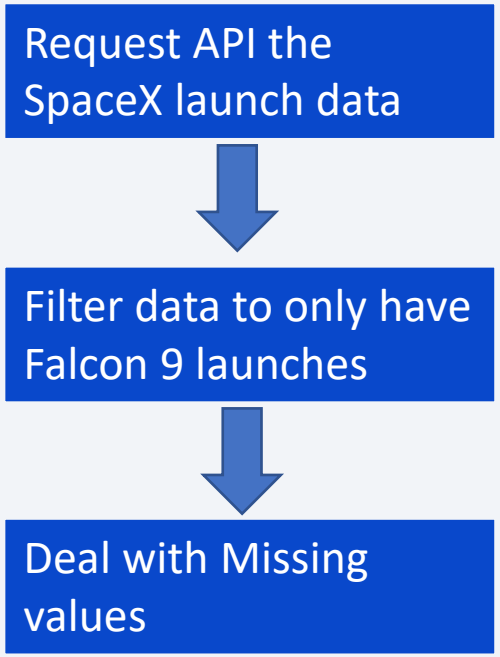
- Datasets are collected from SpaceX API and webscrapping Wikipedia
  - Data sets collected from SpaceX API (https://api.spacexdata.com/v4/launches/past)

  - Data sets collected from Wikipedia by web scrapping (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API

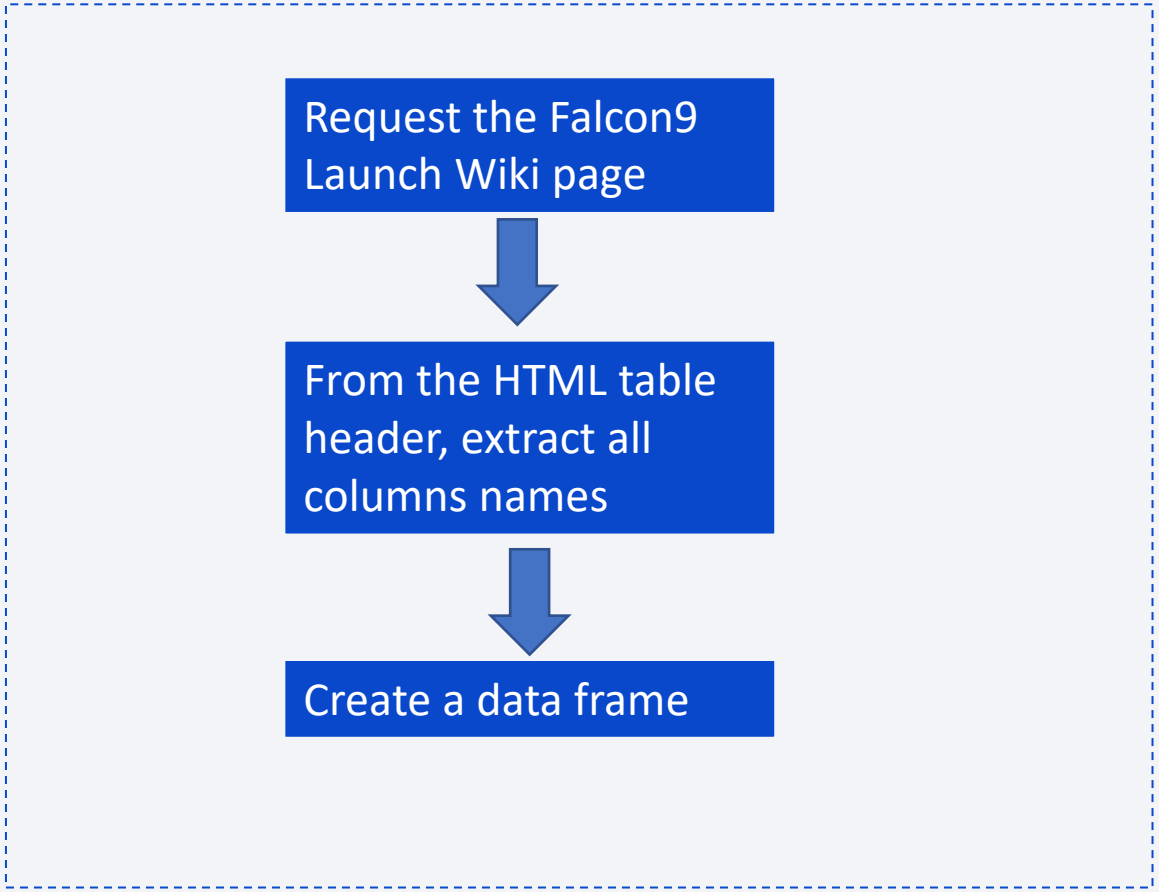
- Data can be obtained from the SpaceX public API

- Source code: https://github.com/Chengz2019/applied-data-science-capstone/blob/6b87cc9da18f0515394f75ba9fafec2e9c26a1d5/spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data sets can be collected from Wikipedia by web scrapping
- Source code: https://github.com/Chengz2019/applied-data-science-capstone/blob/6b87cc9da18f0515394f75ba9fafec2e9c26a1d5/spacex-webscraping.ipynb



Data Wrangling

- Little Exploratory Data Analysis (EDA) was performed on the dataset.
- Summarize launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Create the labeling outcome from the Outcome column.

Source code: https://github.com/Chengz2019/applied-data-science-capstone/blob/6b87cc9da18f0515394f75ba9fafec2e9c26a1d5/EDA-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore the relationship between different pair of variables, scatter plot and bar plots were used to visualize.
  - Flight Number vs Payload Mass, Flight Number vs Launch Site, Launch Site vs Payload Mass, Flight Number vs Orbit Type, Payload vs Orbit Type

- Source code: https://github.com/Chengz2019/applied-data-science-capstone/blob/6b87cc9da18f0515394f75ba9fafec2e9c26a1d5/EDA%20with%20Data%20visualization.ipynb

EDA with SQL

- Used 10 different queries to explore the data:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name beginning with the string ‘CCA’;
  - Total 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;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
  - Rank 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

Source Code: https://github.com/Chengz2019/applied-data-science-capstone/blob/6b87cc9da18f0515394f75ba9fafec2e9c26a1d5/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicates points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space center
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site;
  - Lines are used to indicate distances between two coordinates.

- Source Code: https://github.com/Chengz2019/applied-data-science-capstone/blob/a53a97afc7ce646a92bac99ed1e8af609030c5ca/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- The Percentage of launches by site and payload range were used to visualize data
- This two graphs/plots can be used to quickly analyze the relation between payloads and launch sites, it also helps to identify the best launch location according to payloads.

- Source Code: https://github.com/Chengz2019/applied-data-science-capstone/blob/a53a97afc7ce646a92bac99ed1e8af609030c5ca/spacex_dash_app.py

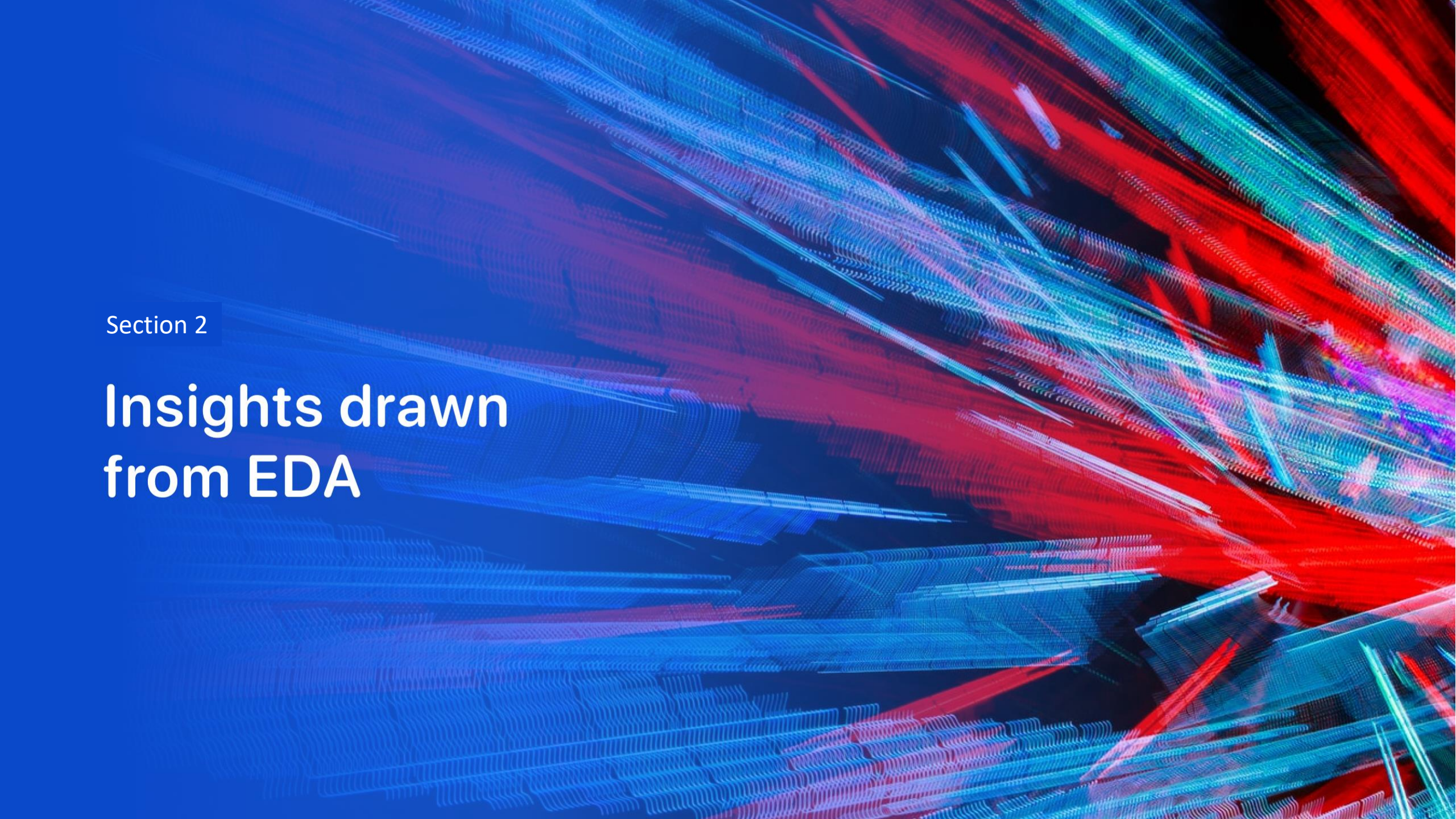
Predictive Analysis (Classification)

- Data preparation
  - Load dataset, Normalize data, split data into training and test set.
- Model building:
  - Four different classification model were used – logistic regression, support vector machine, decision tree and k nearest neighbor.
- Model evaluation:
  - Get best hyperparameters for each model
  - Calculate the accuracy for each model with training and test dataset
  - Confusion matrix
- Model comparison
  - Compare each model based on the accuracy

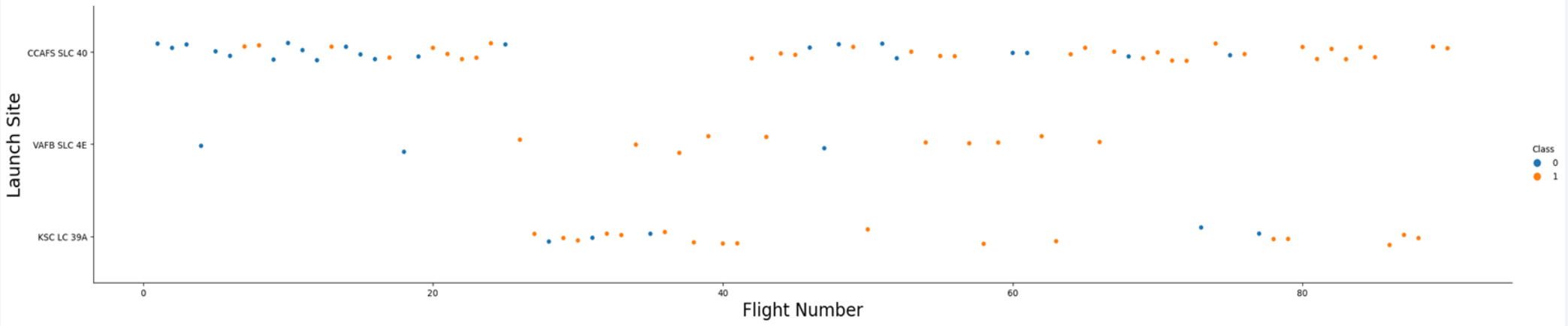
Source Code: https://github.com/Chengz2019/applied-data-science-capstone/blob/a53a97afc7ce646a92bac99ed1e8af609030c5ca/PredictiveAnalysis.ipynb

Results

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

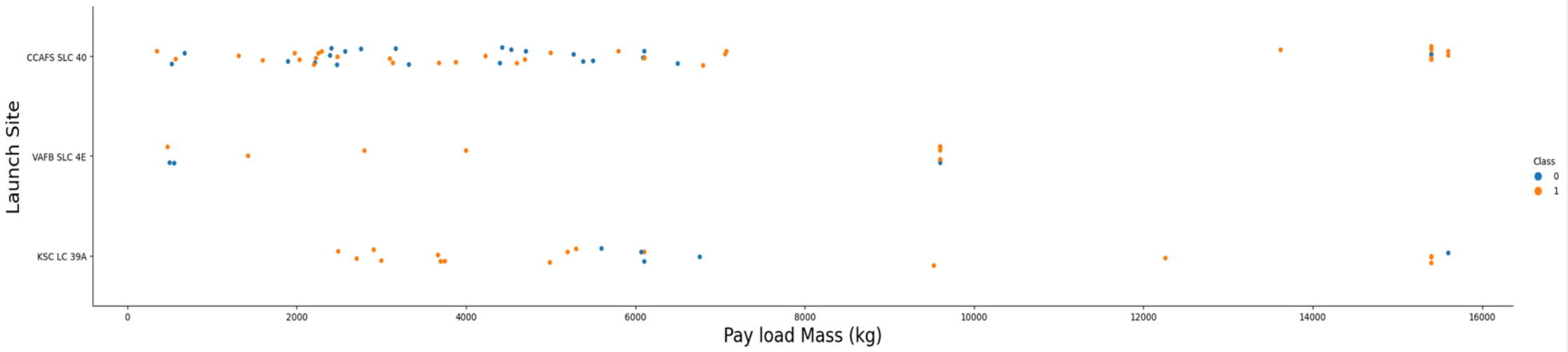

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- From the above plot, we can see the best launch site is CCAF5 SLC 40, where most of recent launches were successful;
- VAFB SLC 4E comes the second place, and KSC LC 39A the thrid.

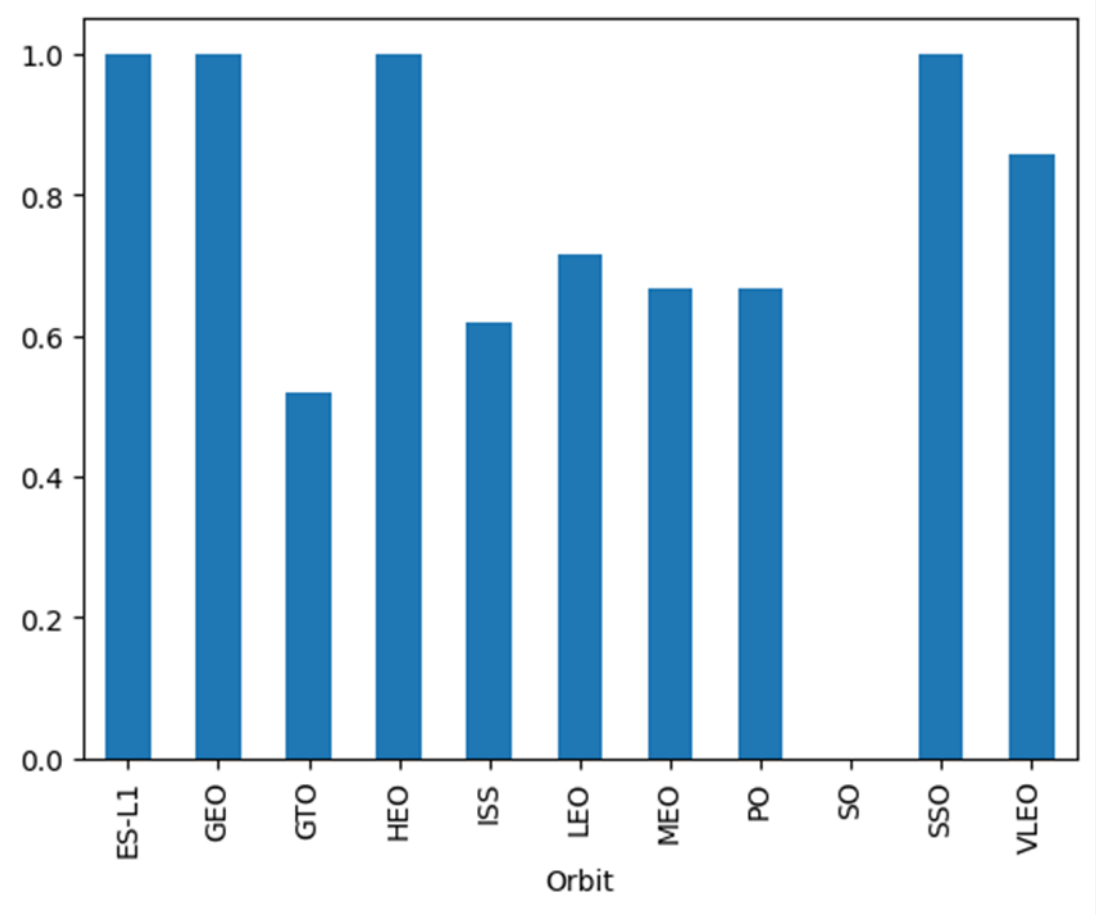
Payload vs. Launch Site



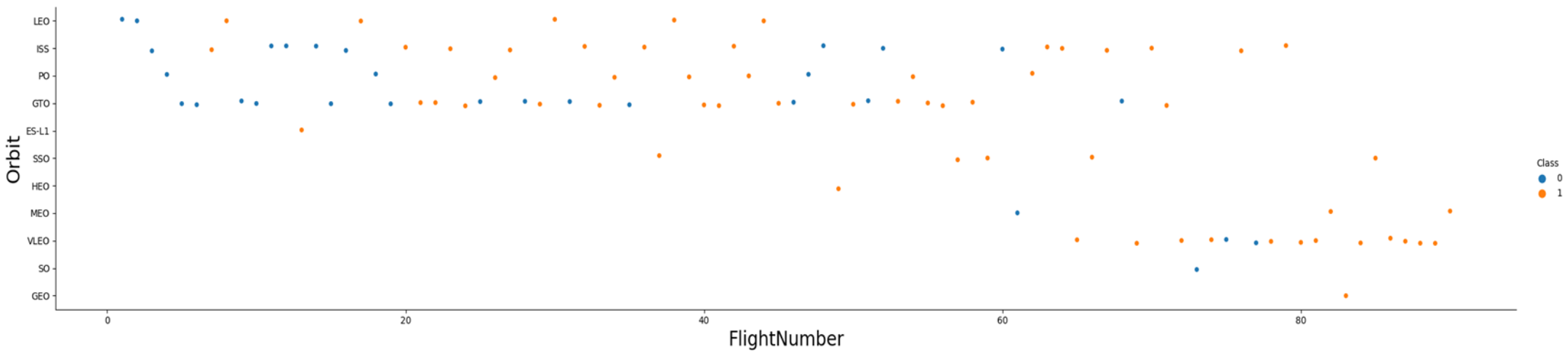
- Payloads over 8000kg have success rate;
- the VAFB-SLC 4E launchsite there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

- The orbits has most success rates are:
  - ESL1
  - GEO
  - HEO
  - SSO

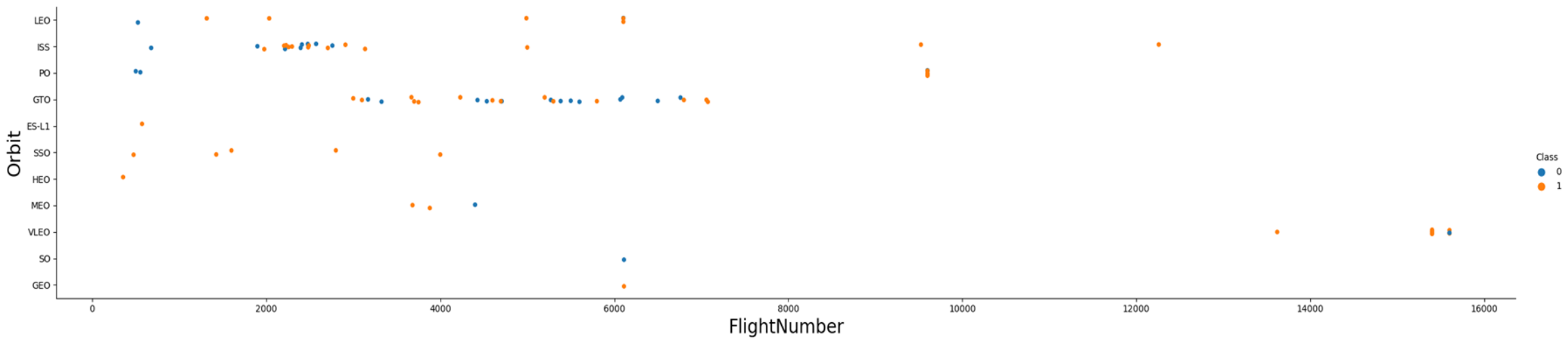


Flight Number vs. Orbit Type



- Success rate improved to all orbits with the increasing flight number.

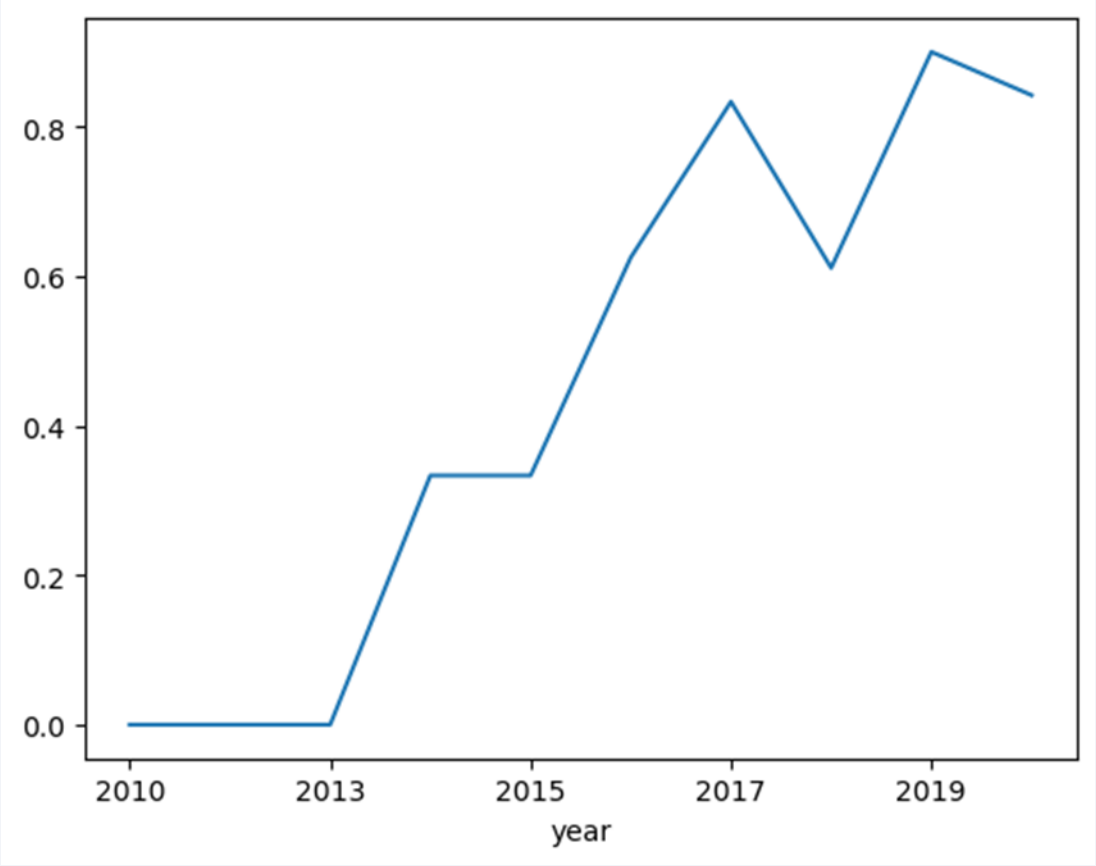
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Since 2013, we can see an increase in the SpaceX Rocket success rate.



All Launch Site Names

- Query:

```
select distinct(launch_site) from spacex;
```

- Result:

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

- Explanation: We can see there are four different launch site from the dataset.

Launch Site Names Begin with 'CCA'

- Query: select * from spacex where launch_site like 'CCA%' limit 5;
- Result:

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)
2012-08-10	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
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

- Explanation: Use the like clause in the where clause to filter the launch sites that begin with ‘CCA’. Limit 5 only shows 5 rows from the filtering.

Total Payload Mass

- Query: select sum(payload_mass__kg_) as total_payload_mass from spacex where customer = 'NASA (CRS)';
- Result:

total_payload_mass
22007

- Explanation: The query gives back the sum of all of payload masses from customer NASA (CRS)

Average Payload Mass by F9 v1.1

- Query:

```
select avg(payload_mass__kg_) as average_payload_mass from spacex where booster_version='F9 v1.1'
```

- Result:

average_payload_mass
3676

- Explanation: This query returns the average of all payload masses where the booster version contains the substring F9 v1.1

First Successful Ground Landing Date

- Query:

```
select min(DATE) from spacex where landing__outcome = 'Success (ground pad)';
```

- Result:

1
2017-01-05

- Explanation: use min function to find the oldest date, use where clause to filter dataset to keep only records which landing was successful.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:

```
select booster_version from spacex where landing__outcome = 'Success (drone ship)' and payload_mass__kg_>4000 and payload_mass__kg_<6000;
```

- Result:

booster_version
F9 FT B1022
F9 FT B1031.2

- Explanation: In this query, the where clause and AND clause to filter the dataset, it will return the booster version where landing was successful and payload mass is between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

- Query:

```
select count(*) as mission_outcome from spacex group by mission_outcome;
```

- Result:

mission_outcome
44
1

- Explanation: Use group by mission_outcome, and count the total number, we can get the total number of successful and failure mission outcome.

Boosters Carried Maximum Payload

- Query:

```
select booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
```

- Result:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

- Explanation: Use the subquery to find the heaviest payload mass with max function. The query returns the booster version with heaviest payload mass.

2015 Launch Records

- Query:

```
select landing__outcome,booster_version,launch_site, date from spacex where YEAR(DATE) = 2015 and landing__outcome='Failure (drone ship)';
```

- Result:

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

- Explanation: the query returns the booster version and launch site in 2015 which launch is failure.

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

- Query:

```
select landing__outcome, count(landing__outcome) as landing_outcome_rank from spacex where Date between '2010-06-04' and '2017-03-20' group by landing__outcome ord
```

- Result:

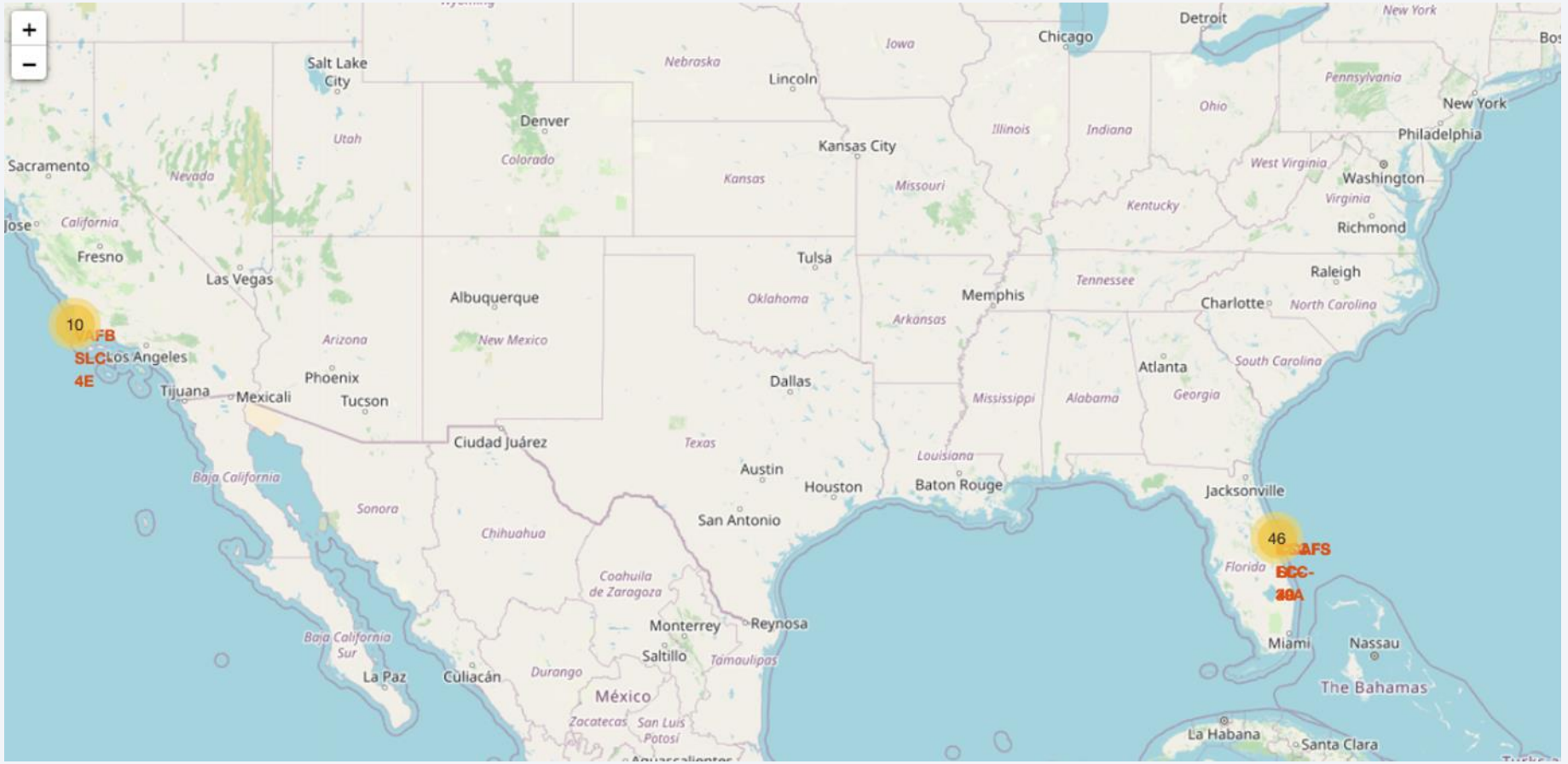
landing__outcome	landing_outcome_rank
No attempt	7
Failure (drone ship)	2
Success (drone ship)	2
Success (ground pad)	2
Controlled (ocean)	1
Failure (parachute)	1

Explanation: The query returns the landing outcomes, use group by clause to group result by landing outcome.

Section 3

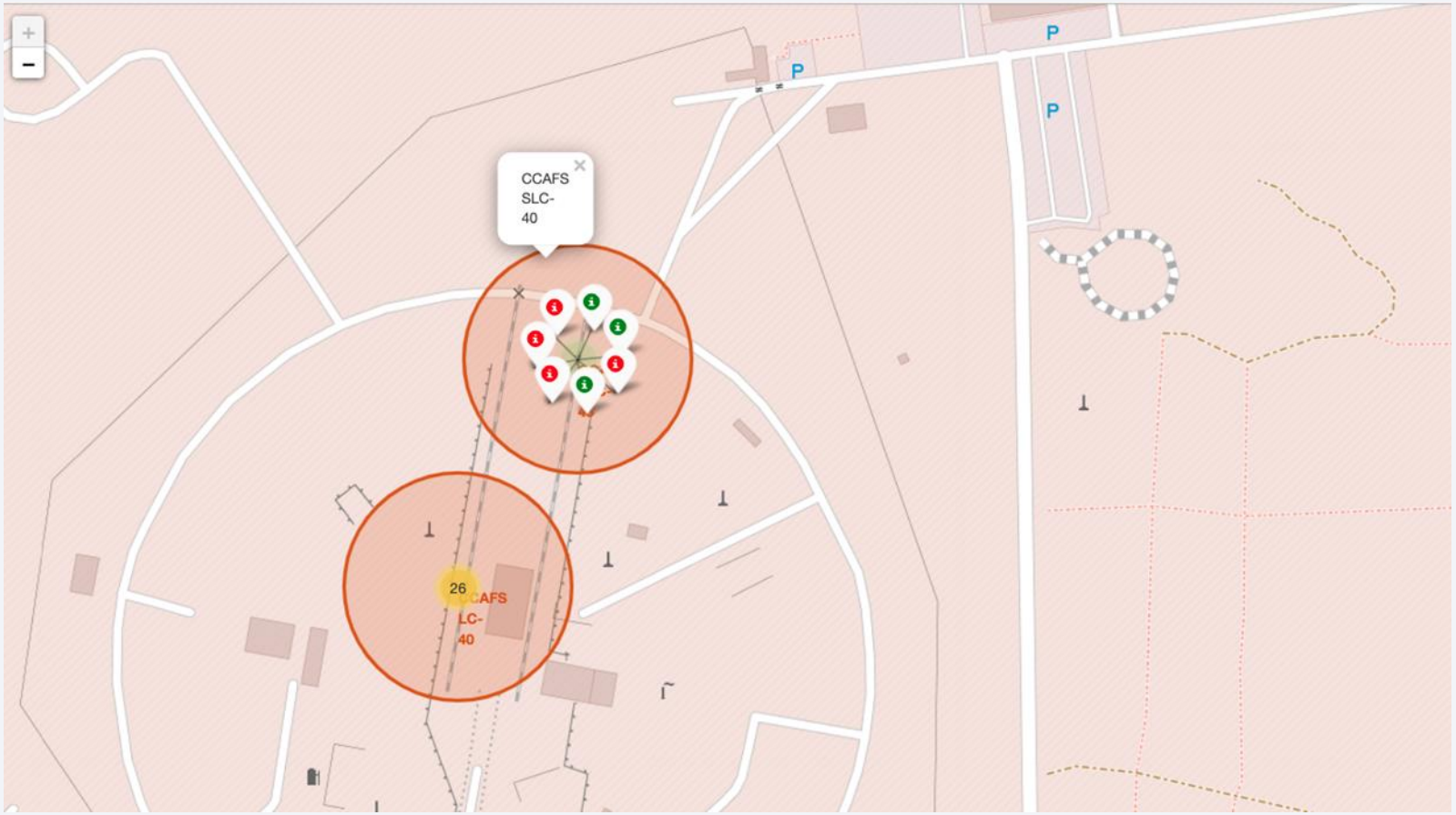
Launch Sites Proximities Analysis

Folium Map – All launch sites



- From the map, we can see all launch sites are located near sea, but not far away from roads and railroads.

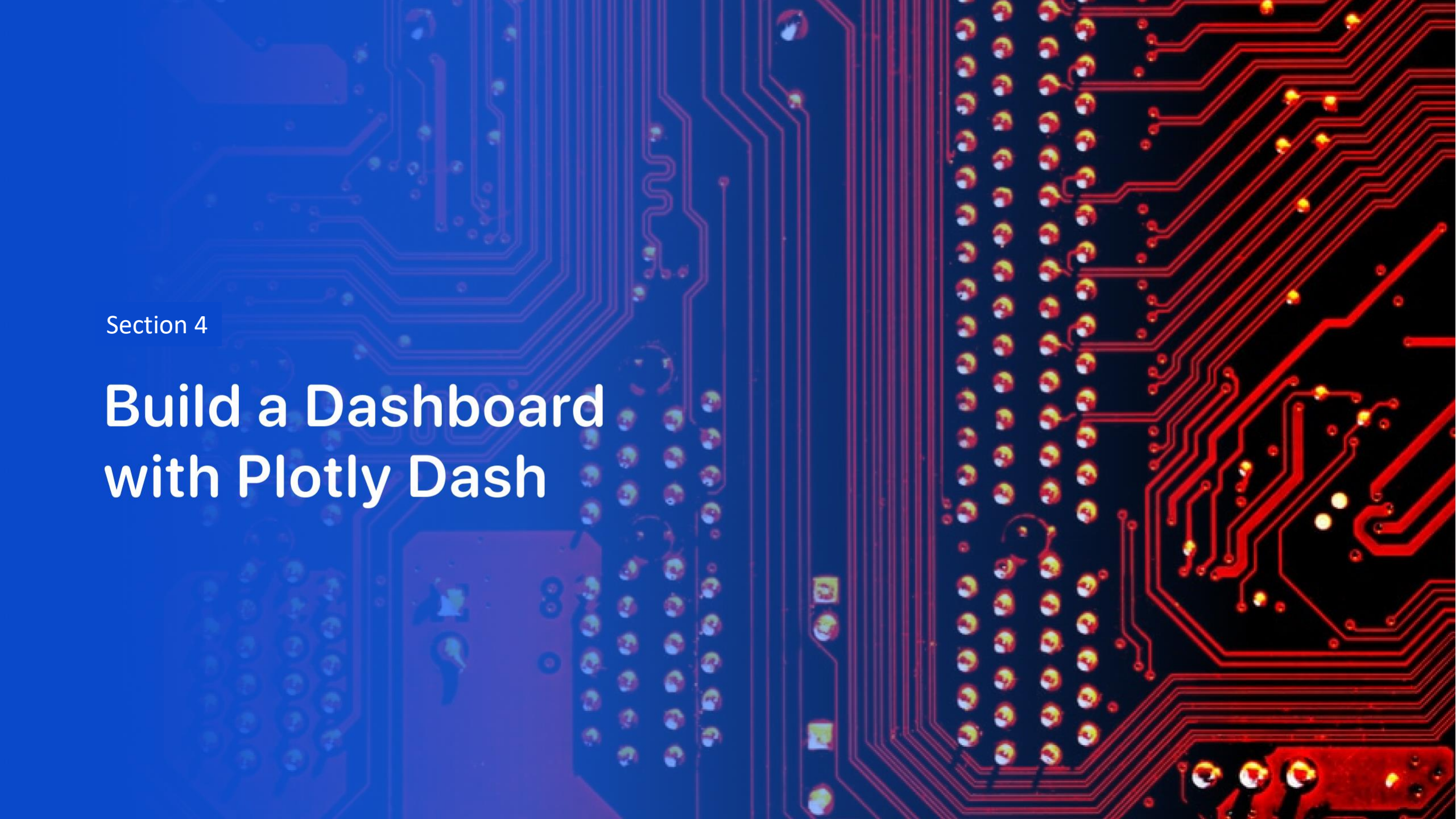
Launch Outcomes by Site



- Green markers represent successful and read ones represent failure.

Distance

- Replace <Folium map screenshot 3> title with an appropriate title


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

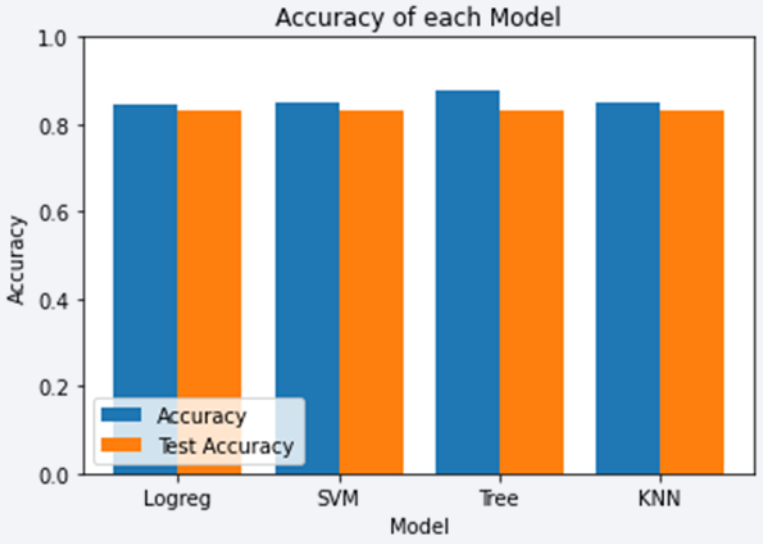
Section 5

Predictive Analysis (Classification)

Classification Accuracy

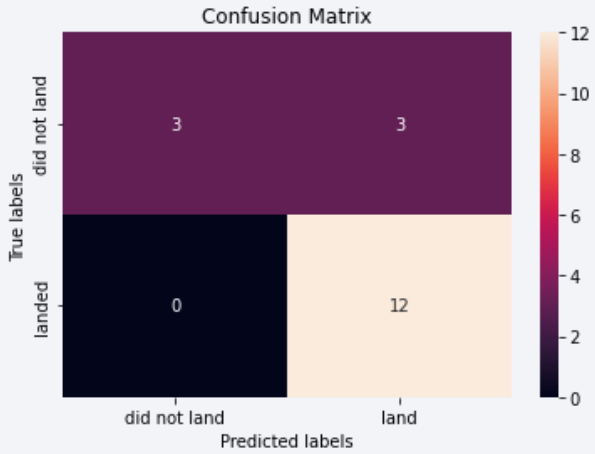
- Accuracies are tested for four different models, which is plotted on the right.

- All of the four models have similar accuracy on the test dataset, 83.33%. While decision tree has higher accuracy than other three models, 87%.



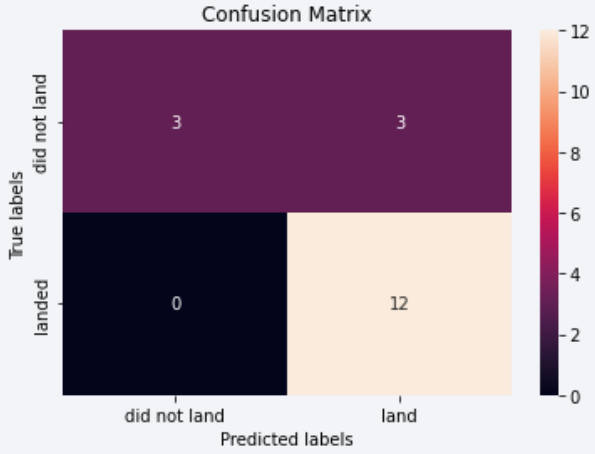
Confusion Matrix

Logistic regression



SVM

Decision Tree



KNN

Since the test accuracy are the same, the confusion matrices are also the same.

Conclusions

- The success of a mission can be affected by a few factors, such as the orbit, the launch site and the number of the previous launches.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Successful landing outcomes seem to improve over time, according to the evolution of processes and rockets;
- For this dataset, Desion tree models has better accuracy.

Thank you!