Winning Space Race with Data Science

Rui He
6/9/2022

Outline

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

Executive Summary

- Summary of methodologies

This research acquires the data through web scrapping from wiki. Raw data acquired from wiki was first inspected for its structure and then structuralized through transformation. The structured data was used for machine learning to predict the landing success rate based on flight number, payload mass, orbit, launch site, flights, gridfins, resued, legs, landing pad, block, reuse count, and serial. Model used includes support vector machine, decision tree, logic regression, and KNN. The accuracy of model predictions was compared based on test set.

Executive Summary

- Summary of all results
  - Results as follow. The best accuracy towards test set achieved was 0.875 using decision tree with maximum depth of 6.

```
Logic Regression tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
Logic Regression accuracy : 0.8472222222222222

SVM tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
SVM accuracy : 0.8472222222222222

Decision Tree tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto',
'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'random'}
Decision Tree accuracy : 0.875

KNN tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 9, 'p': 1}
KNN accuracy : 0.8472222222222222
```

Introduction

- Project background and context

Sending things to space is expensive, and reusable rockets is a way to reduce cost. Space X is launching reusable rockets and wants to predict the success rate based on data they acquired. Considered the complex relation between features in the data, machine learning is used to perform these experiments.

Introduction

- Problems you want to find answers

Ultimate goal is to predict if the reusable rocket can land successfully after delivering the payload to the orbit.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was acquired from Wiki using webscrapping through request method and structured with beautifulsoup.
- Perform data wrangling
  - Empty data was replaced with mean, and data was converted to structured dataset
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - First standardize features, then use grid search with CV=10 to find best parameters using logic regression, support vector machine, decision tree, and KNN methods.

Data Collection

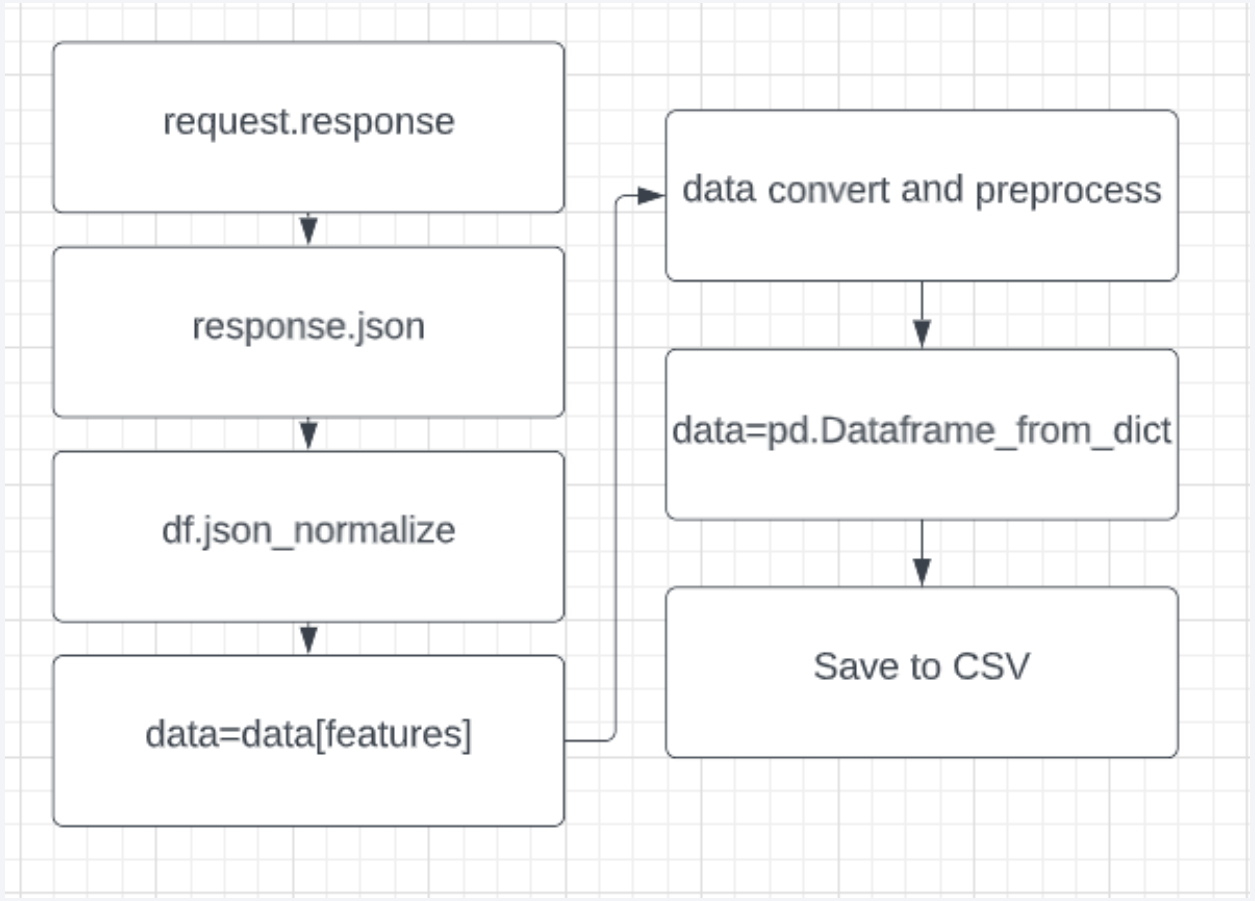
- Data set was acquired from the Wiki using web scrapping.
- First request using response.json to acquire data from 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
- Then format the json to pandas dataframe using pd.json_normalize()
- Data was then selected based on keywords: 'rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc'
- The dataset was selected for launches with one core, one payload, and map the cores and payloads to features.
- The data_utc was converted into dateime

Data Collection – SpaceX API

- Final keys in dataset: FlightNumber, Date,BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longtitude, Latitude

Link:

https://github.com/rhCat/IBMCouse_Capstone/blob/519fcf92e716cbb78b8d68d75b80430f4d6b05bd/Week_1_Data_acquicisium/jupyter-labs-spacex-data-collection-api.ipynb



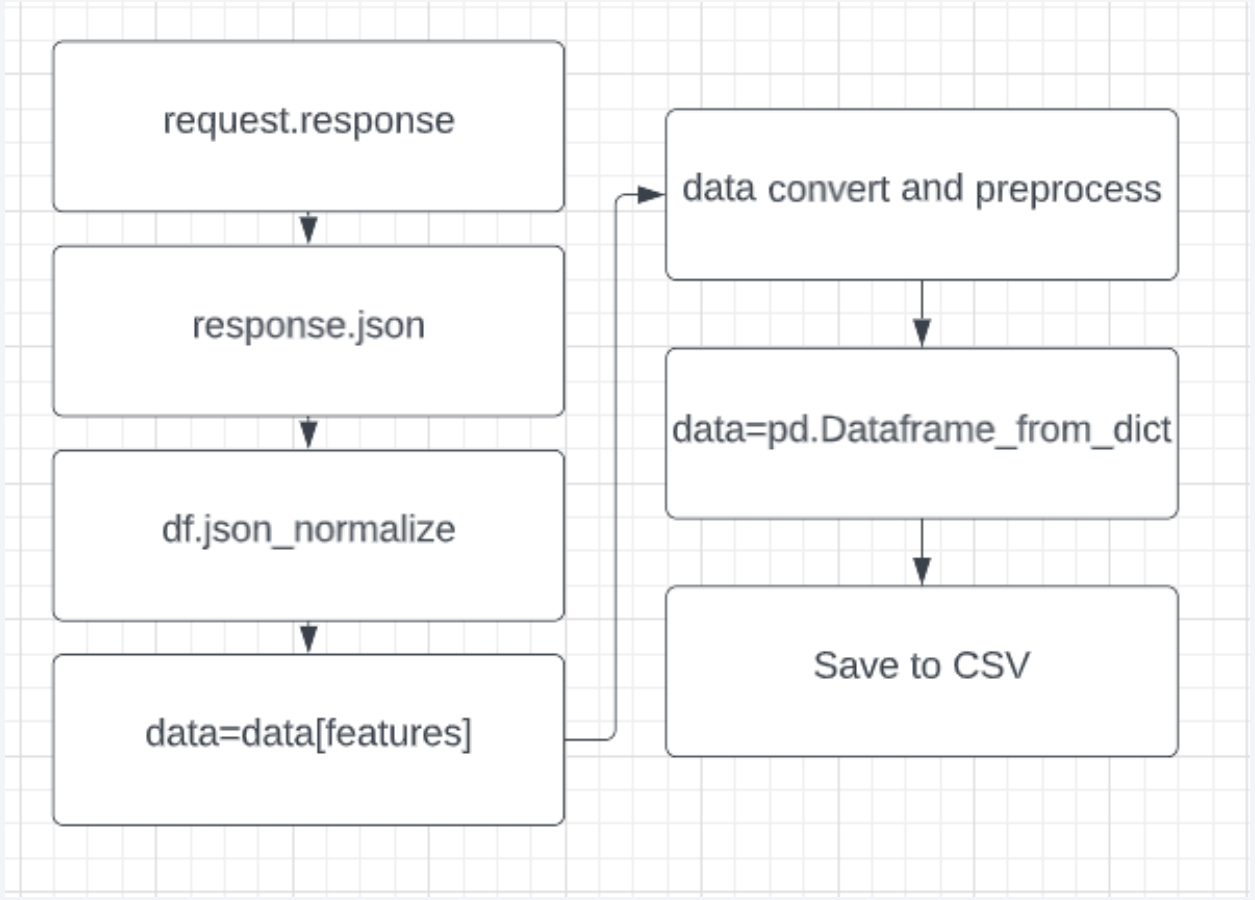
Data Collection - Scraping

- Key processes:

Request, beautifulsoup, process data

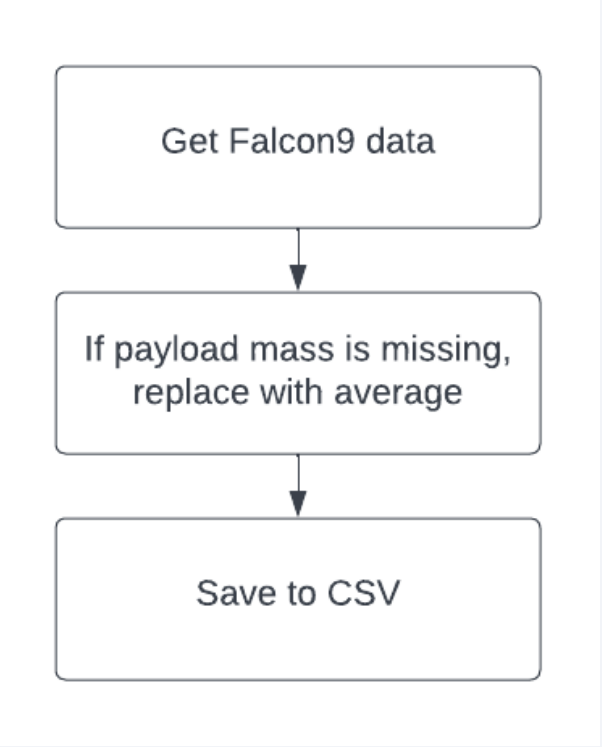
Link:

https://github.com/rhCat/IBMCouse_Capstone/blob/master/Week_1_Data_acquicisium/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data was selected for falcon9 only
- Data with missing values are replaced with average value for payload mass
- Link:
- https://github.com/rhCat/IBMCouse_Capstone/blob/master/Week_1_Data_acquicisium/jupyter-labs-spacex-data-collection-api.ipynb



EDA with Data Visualization

- Charts plotted: Payload mass versus fight number, success rate, Launch site versus flight number, Launchsite versus payload mass, barchart of orbit, scatter of orbit versus flight number, and line chart to be success rate.
- https://github.com/rhCat/IBMCouse_Capstone/blob/master/Week_2_EDA/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Connect to database
- Select everything from database
- 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 launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA with SQL

- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- https://github.com/rhCat/IBMCouse_Capstone/blob/master/Week_2_EDA/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

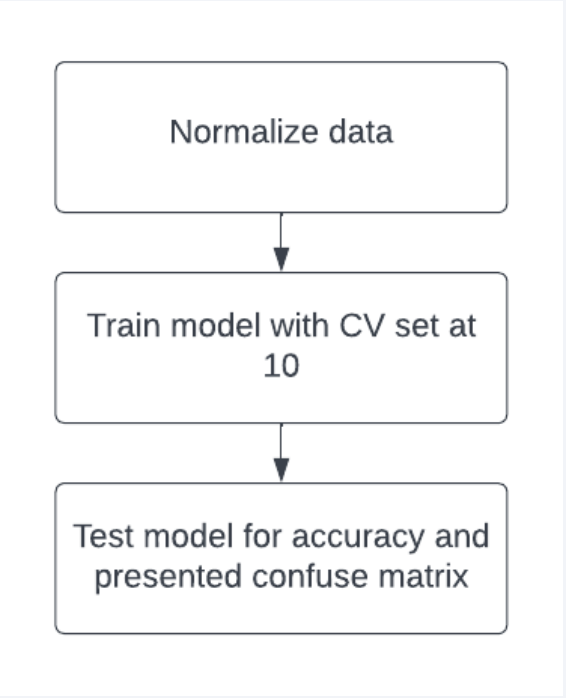
- Mark all launch sites on a mapExplain why you added those objects, Mark the success/failed launches for each site on the map, Calculate the distances between a launch site to its proximities
- Plots and options are added to better view the data
- https://github.com/rhCat/IBMCouse_Capstone/blob/master/Week_3_Visuak_Analytic/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- a Launch Site Drop-down Input ComponentExplain why you added those plots and interactions, a callback function to render success-pie-chart based on selected site dropdown, a Range Slider to Select Payload, a callback function to render the success-payload-scatter-chart scatter plot
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

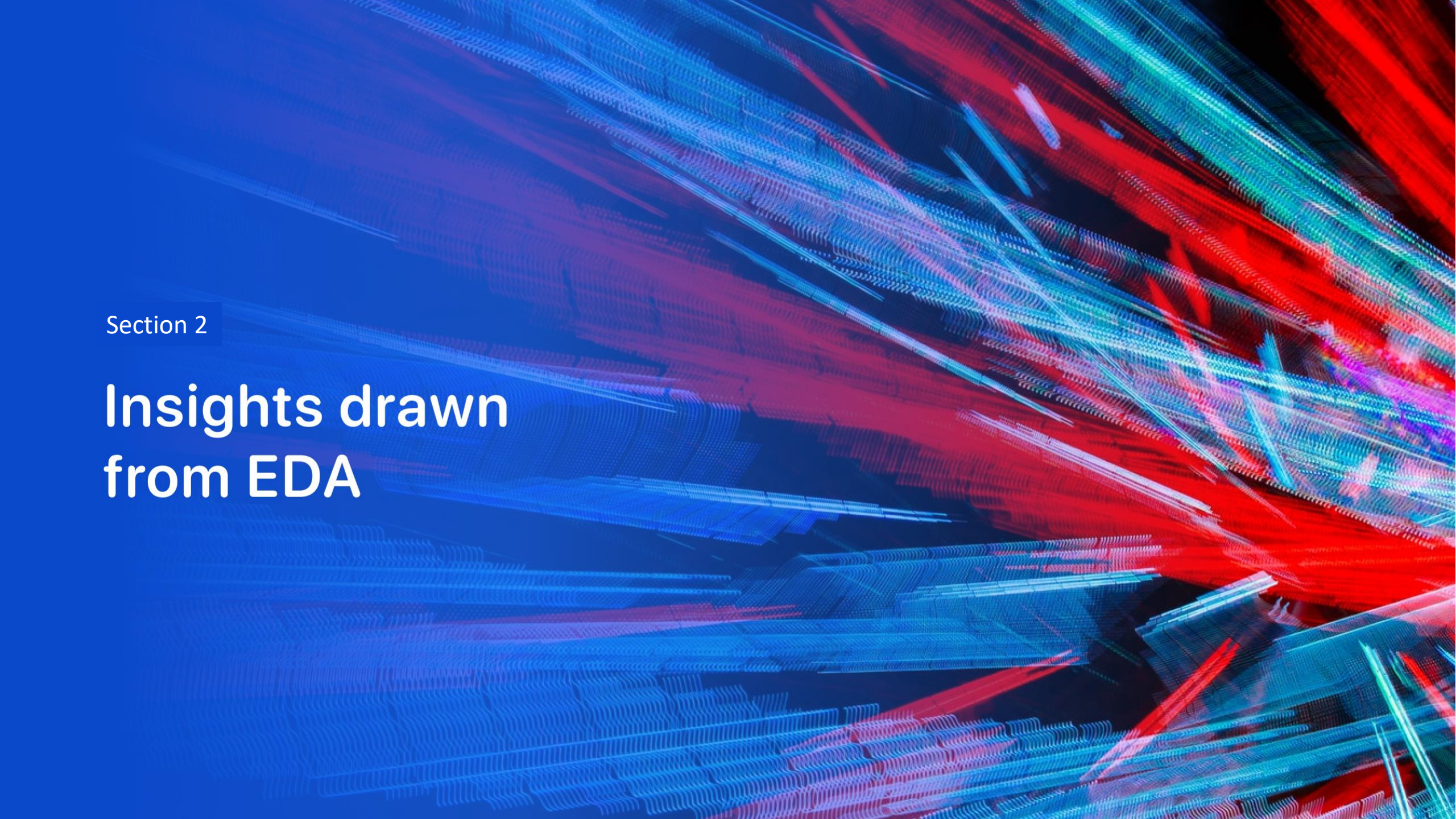
- Logic regression, support vector machine, decision tree, and KNN was used to predict the rocket recover rate by searching the setting with CV =10
- Key: Data normalization, search different parameter combination for each model to find best, and test accuracy
- https://github.com/rhCat/IBMCouse_Capstone/blob/master/Week_4_Predict/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

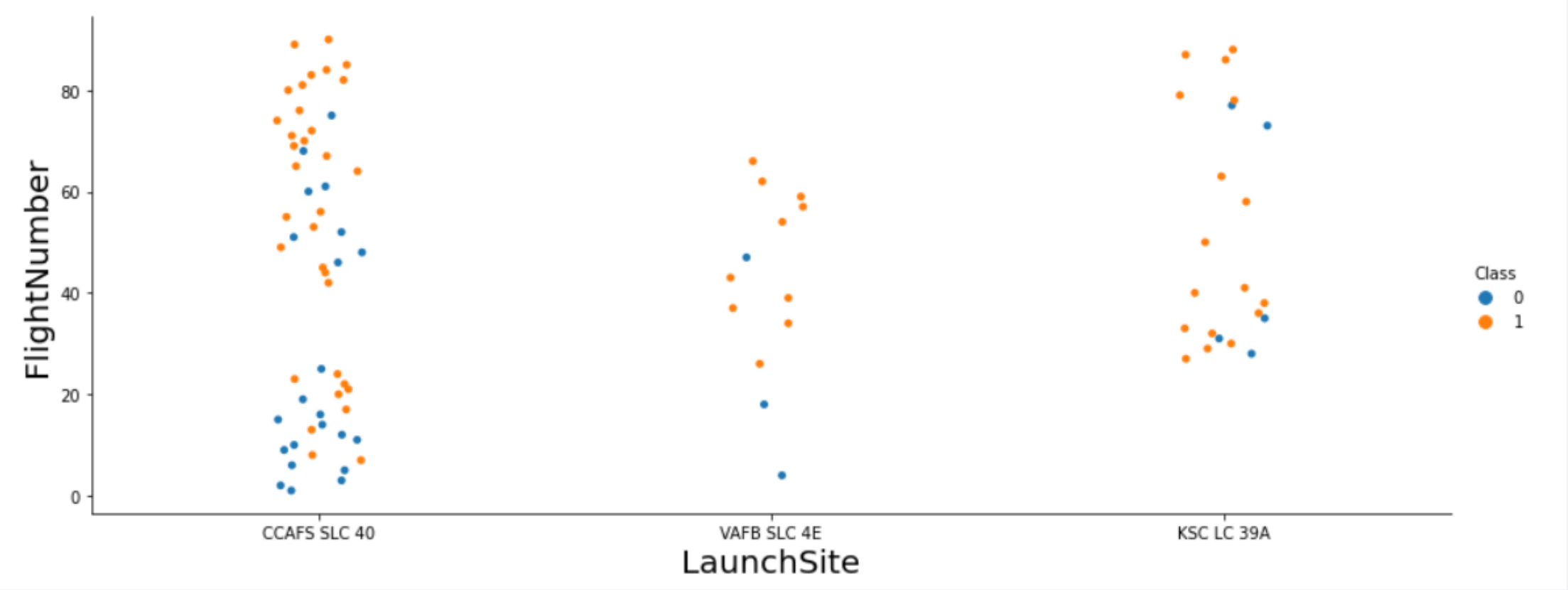
- EDA suggests most relevant features are: FlightNumber, PayloadMass, Orbit, LaunchSite, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial,

```
Logic Regression tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
Logic Regression accuracy : 0.8472222222222222

SVM tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
SVM accuracy : 0.8472222222222222

Decision Tree tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto',
'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'random'}
Decision Tree accuracy : 0.875

KNN tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 9, 'p': 1}
KNN accuracy : 0.8472222222222222
```


Section 2
Insights drawn from EDA

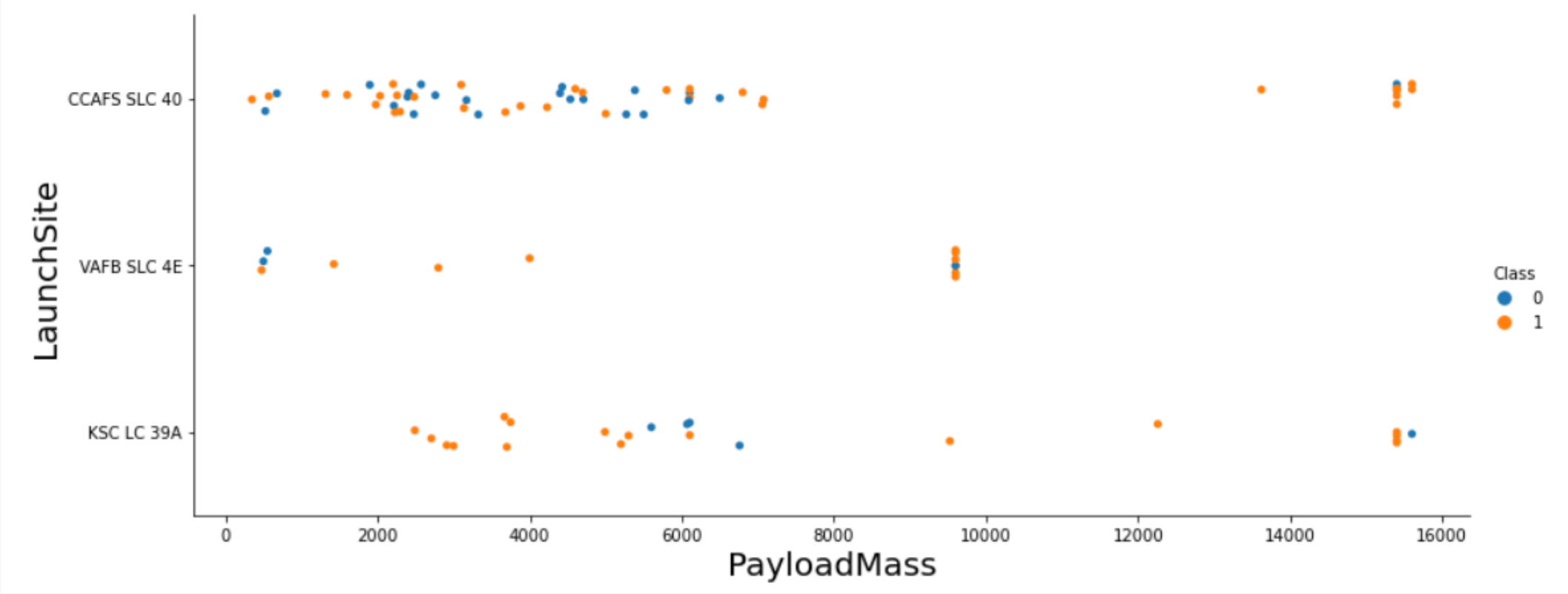
Flight Number vs. Launch Site

- The result shows while most lunches were from CCAFS SLC 40 site, the success rate is lower than the other sites. It is possible that launch sites can impact the success rate.



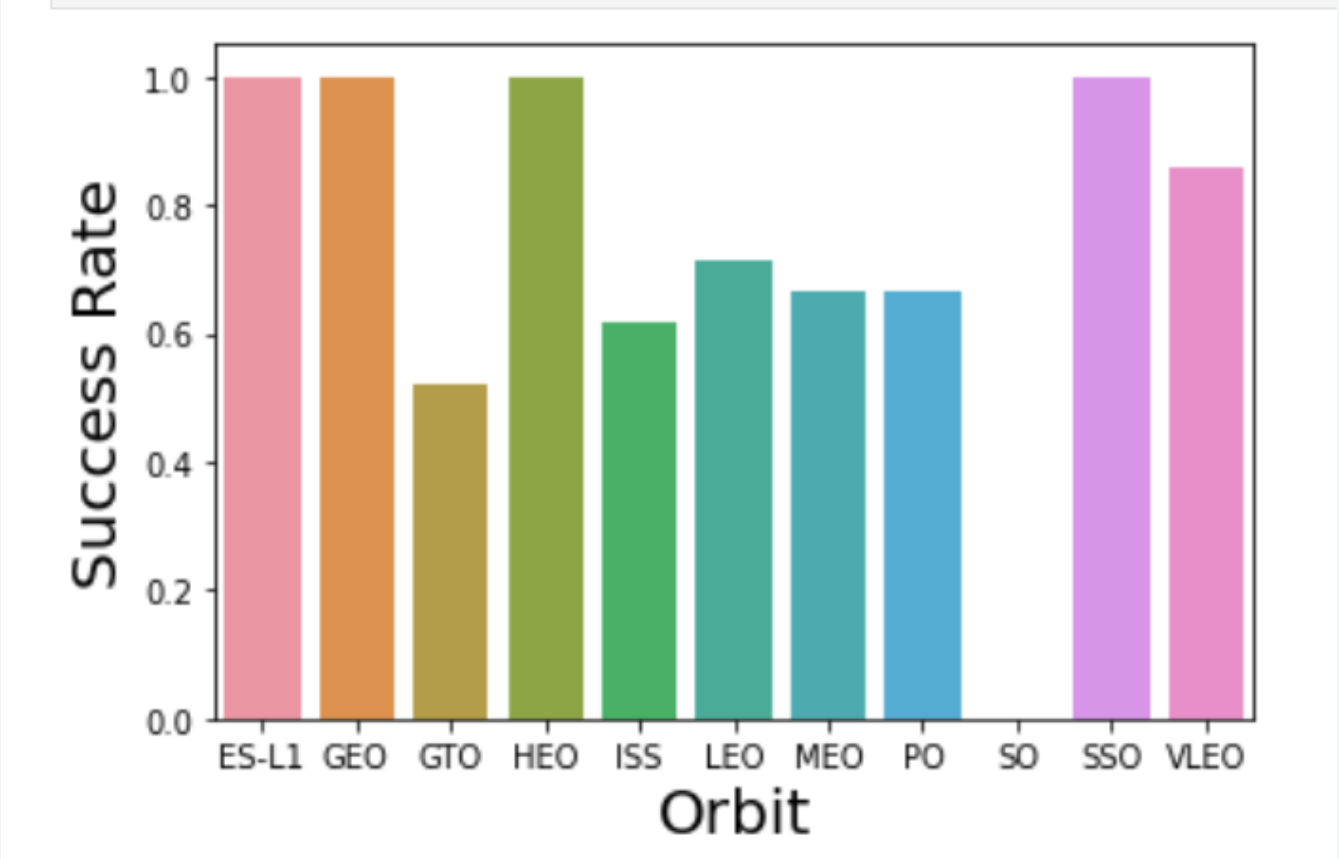
Payload vs. Launch Site

At site CCAFS SLC 40, success rate decrease with payload mass. Fro KSC LC 39A, the failure rate seemingly to be the highest at 8,000. Site VAFB SLC 4E overall have a good success rate.



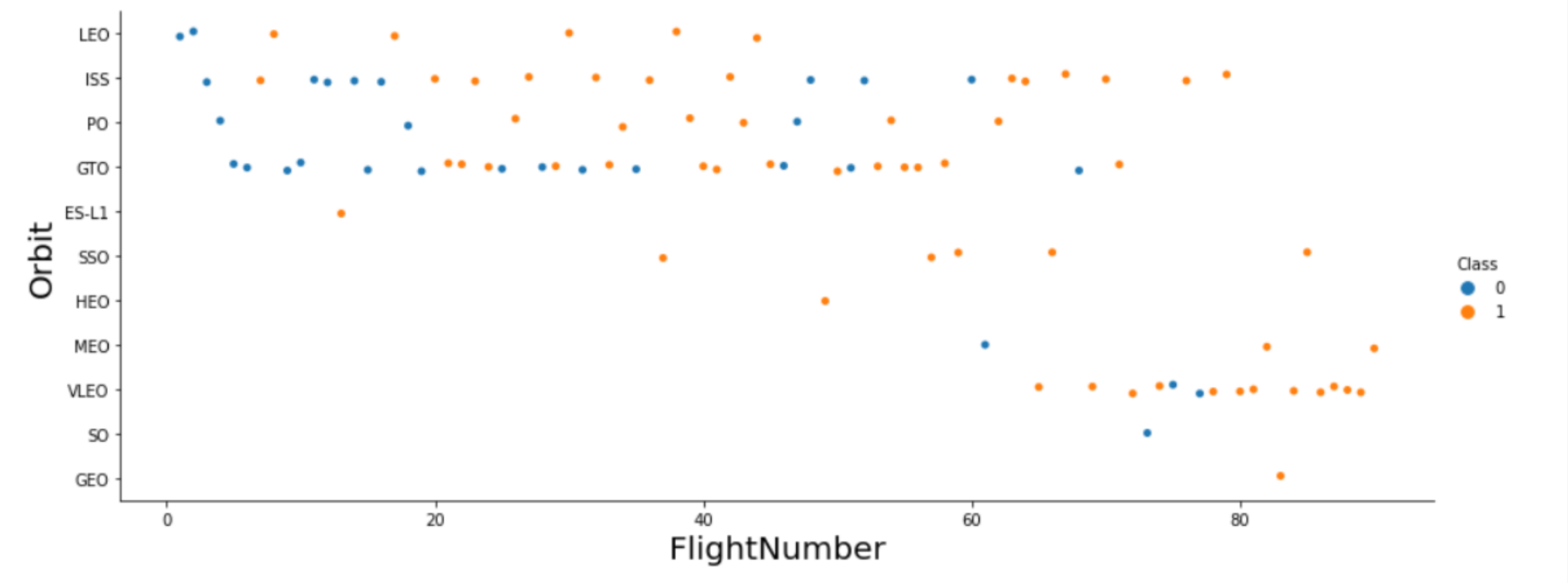
Success Rate vs. Orbit Type

- Results show four types of orbit have perfect success rate while others on average share a 50% chance of success, with SO has a 0 success for 1 flight.



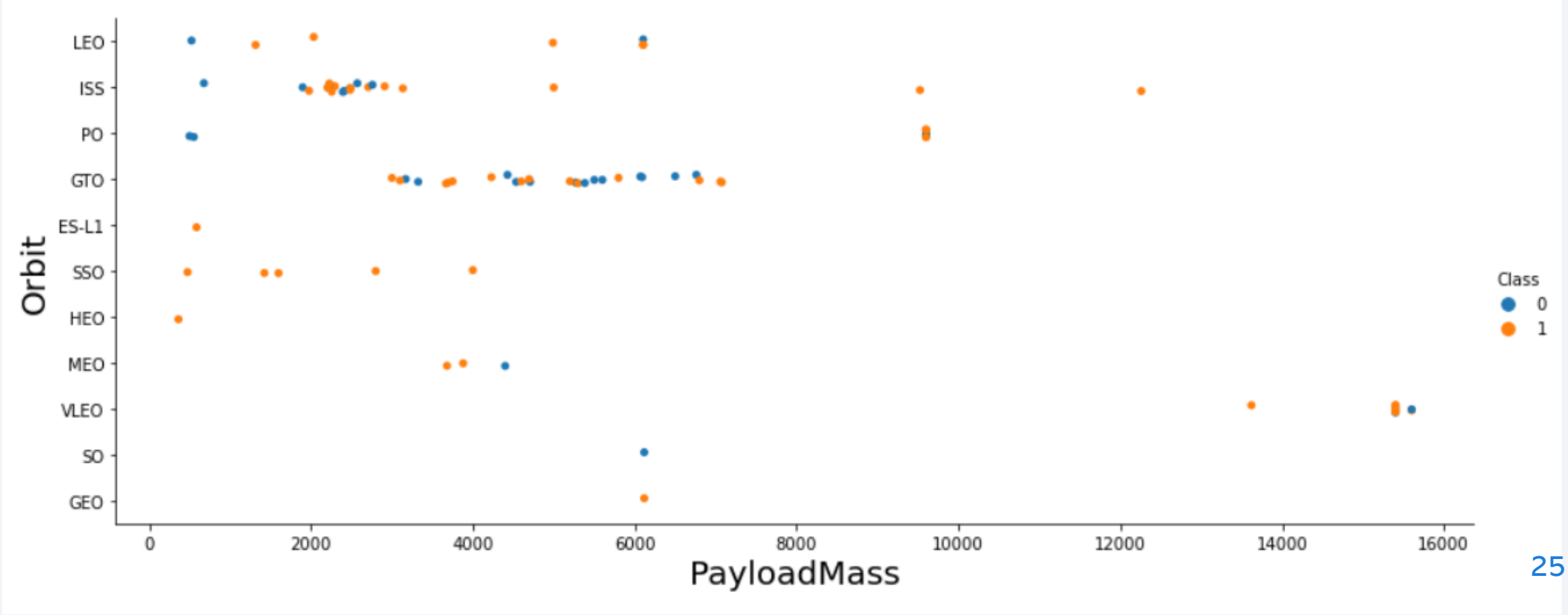
Flight Number vs. Orbit Type

The launch success rate seems to increase with higher (later) flight number on all orbits.



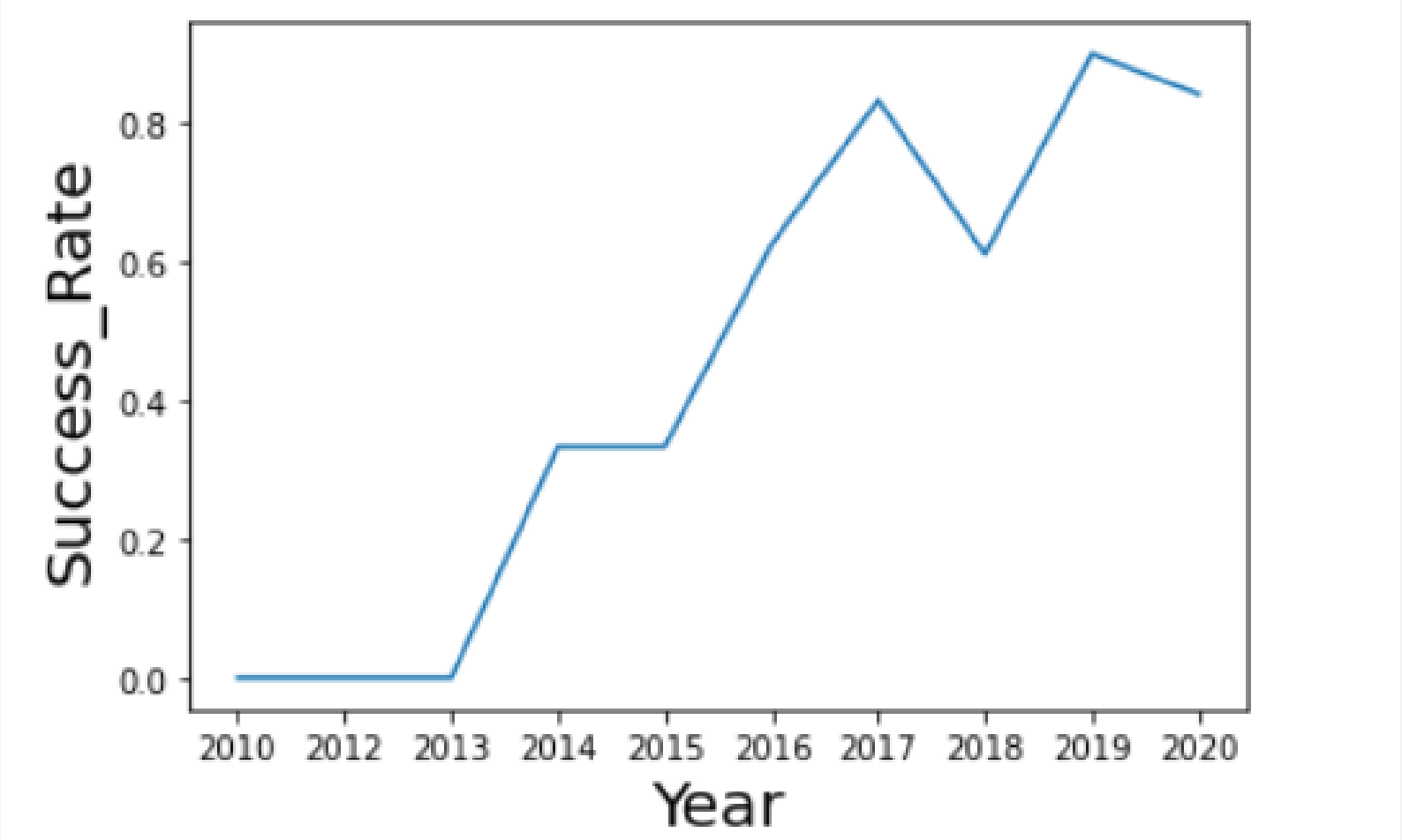
Payload vs. Orbit Type

- The increase of payload mass does not seem to have an impact on success rate for different orbit type



Launch Success Yearly Trend

- Success rate in general increases each year.



All Launch Site Names

- Select distinct can be used to get unique values

13]:
```
%%sql
    Select distinct Launch_Site
    from SPACEXTBL
```

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

13]:

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

Launch Site Names Begin with 'CCA'

- Where __ like '%string_to_find%' is the way to do it

```
%%sql
    Select *
    from SPACEXTBL
    where Launch_Site like '%CCA%'
    limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Task 3

Total Payload Mass

- Select sum() + where is the method to do it.

17]:
```
%%sql
    select sum(PAYLOAD_MASS__KG_)
    from SPACEXTBL
    where Customer like '%NASA (CRS)%'
```

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

17]:

sum(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

- Select avg() + where is the method to do it.

In [18]:
```
%%sql
    select avg(PAYLOAD_MASS__KG_)
    from SPACEXTBL
    where Booster_Version like '%F9 v1.1%'
```

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

Out[18]:

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

Task 5

First Successful Ground Landing Date

- While min() method does the job, I use sort to inspect the dataset

In [50]:

```
%%sql
    select *
    from SPACEXTBL
    where "Landing _Outcome" like '%(ground pad)%'
    order by "Date" desc
    limit 1
```

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

Out[50]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Use between to get the values

In [54]:

```
%%sql
    select Booster_Version
    from SPACEXTBL
    where "Landing _Outcome" like '%Success (drone ship)%'
    and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Out[54]:

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

Task 7

Total Number of Successful and Failure Mission Outcomes

- Select count() does the job

In [58]:
```
%%sql
    select count(Mission_Outcome)
    from SPACEXTBL
    where Mission_Outcome like '%Success%';
```

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

Out[58]:

count(Mission_Outcome)
100

Boosters Carried Maximum Payload

- Subquery with select Max can do it. In this case, it is the max payload each launch

In [87]:

```
%%sql
    select Booster_Version
    from SPACEXTBL
    where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Out[87]:

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

- Use and to add additional requests

In [107…

```
%%sql
    select substr(Date, 4, 2) as Month, "Landing _Outcome", Booster_Version, Launch_Site
    from SPACEXTBL
    where substr(Date,7,4)='2015'
    AND "Landing _Outcome" like "%(drone ship)%"
    and "Landing _Outcome" not like "%Success%"
```

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

Out[107…

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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

- Use group and order to categorize and sum up results

In [111…

```
%%sql
    select "Landing _Outcome", count("Landing _Outcome") as Success_count
    from SPACEXTBL
    where date between "04-06-2010" and "20-03-2017"
    And "Landing _Outcome" like "%Success%"
    group by "Landing _Outcome"
    order by Success_count desc
```

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

Out[111…

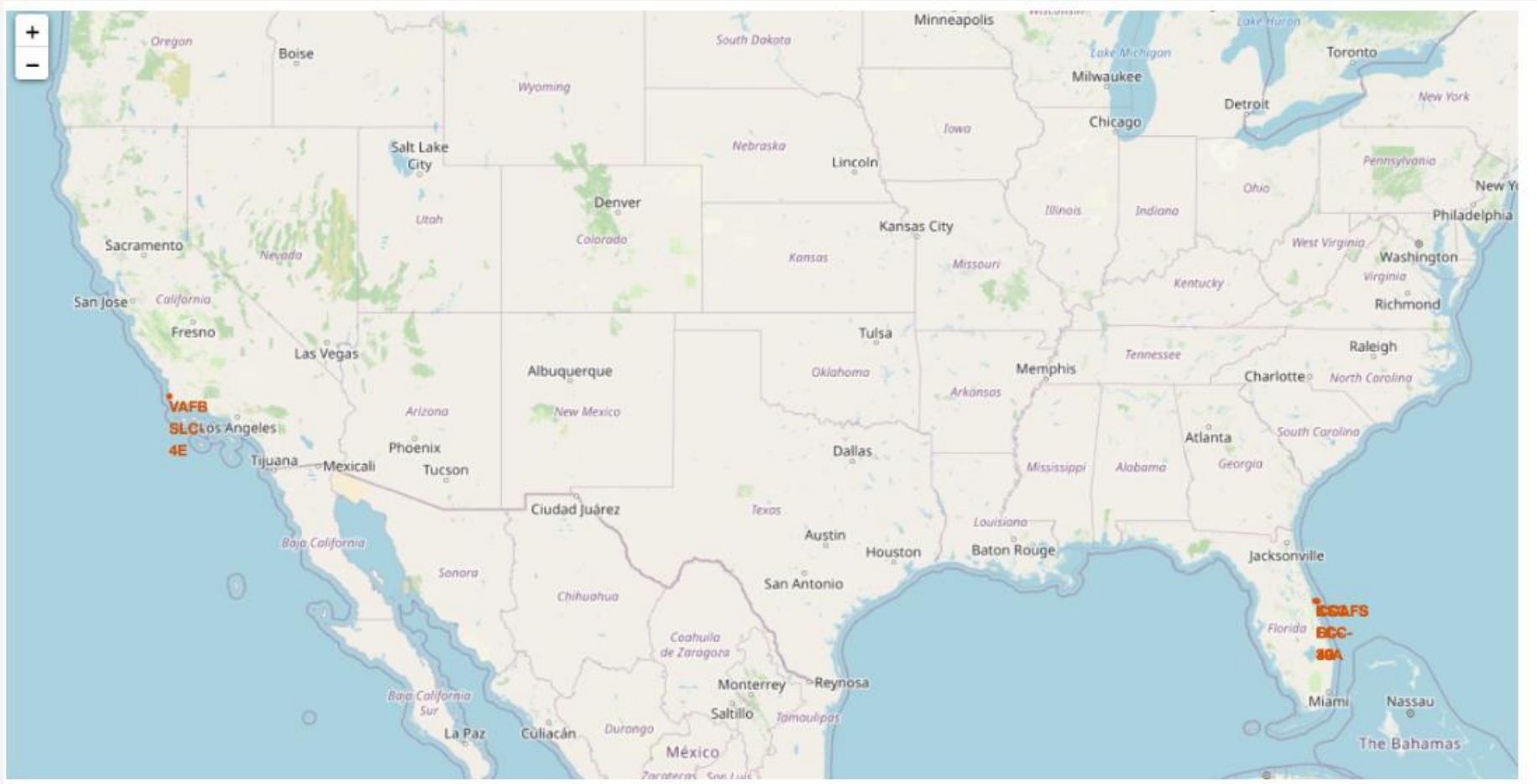
Landing _Outcome	Success_count
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

- One site at west cost, others all in Florida

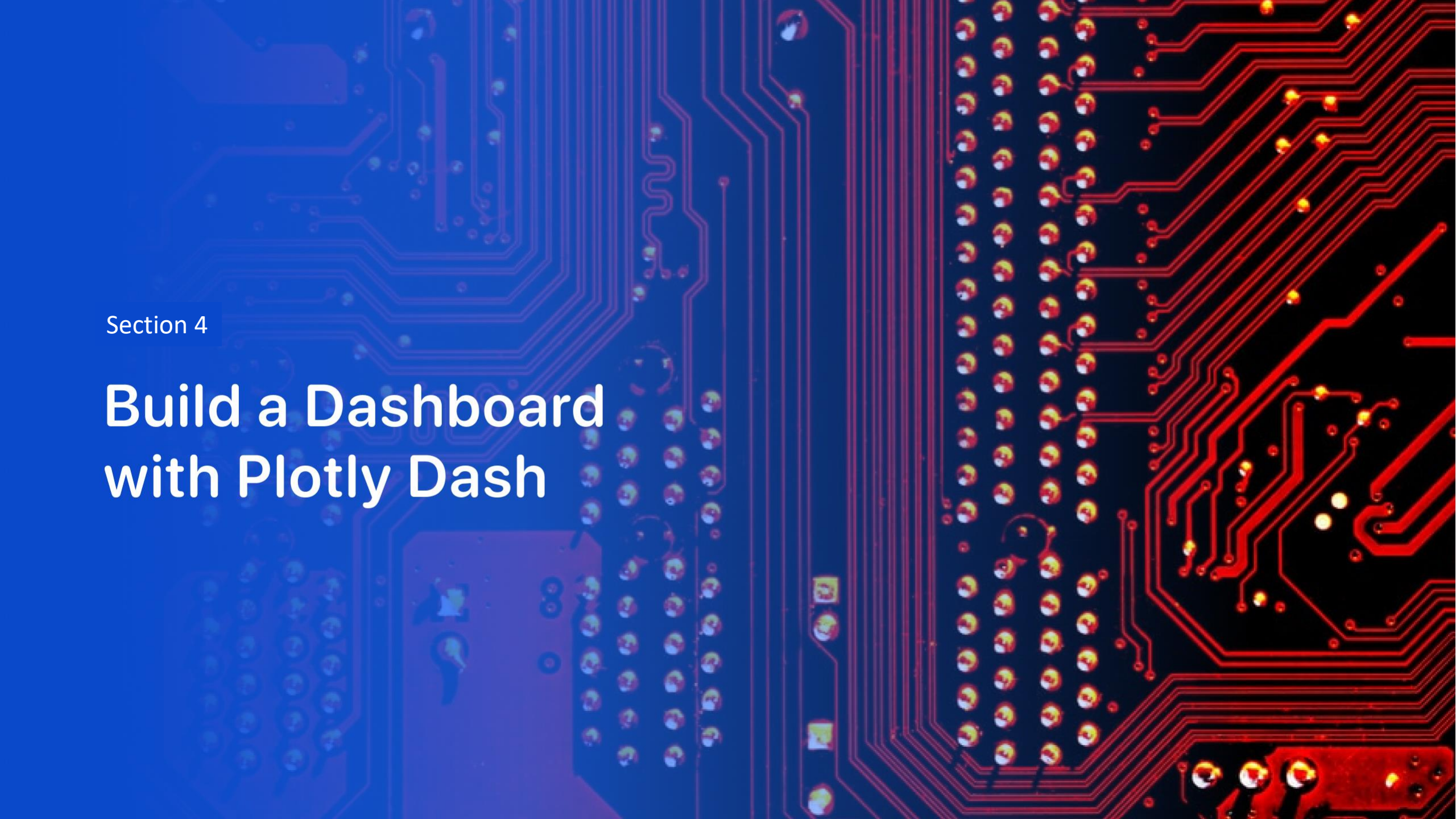


<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<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>

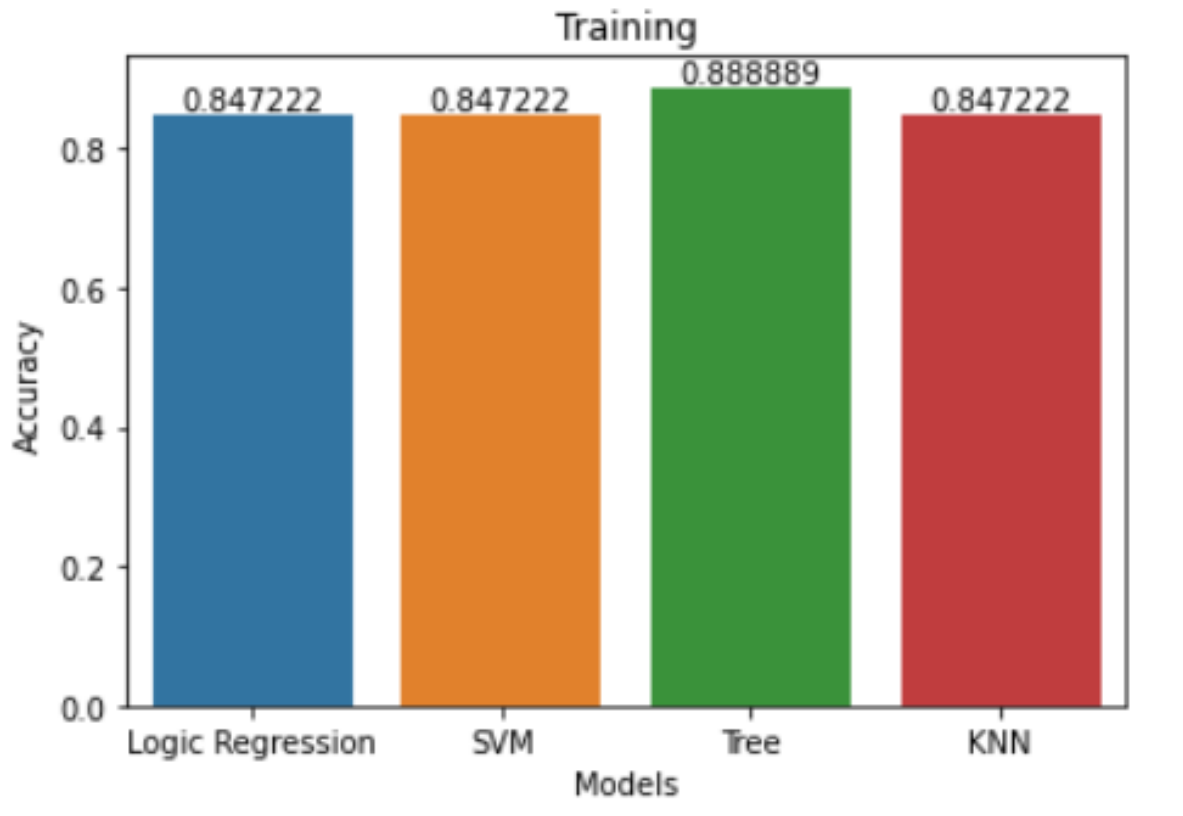
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version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

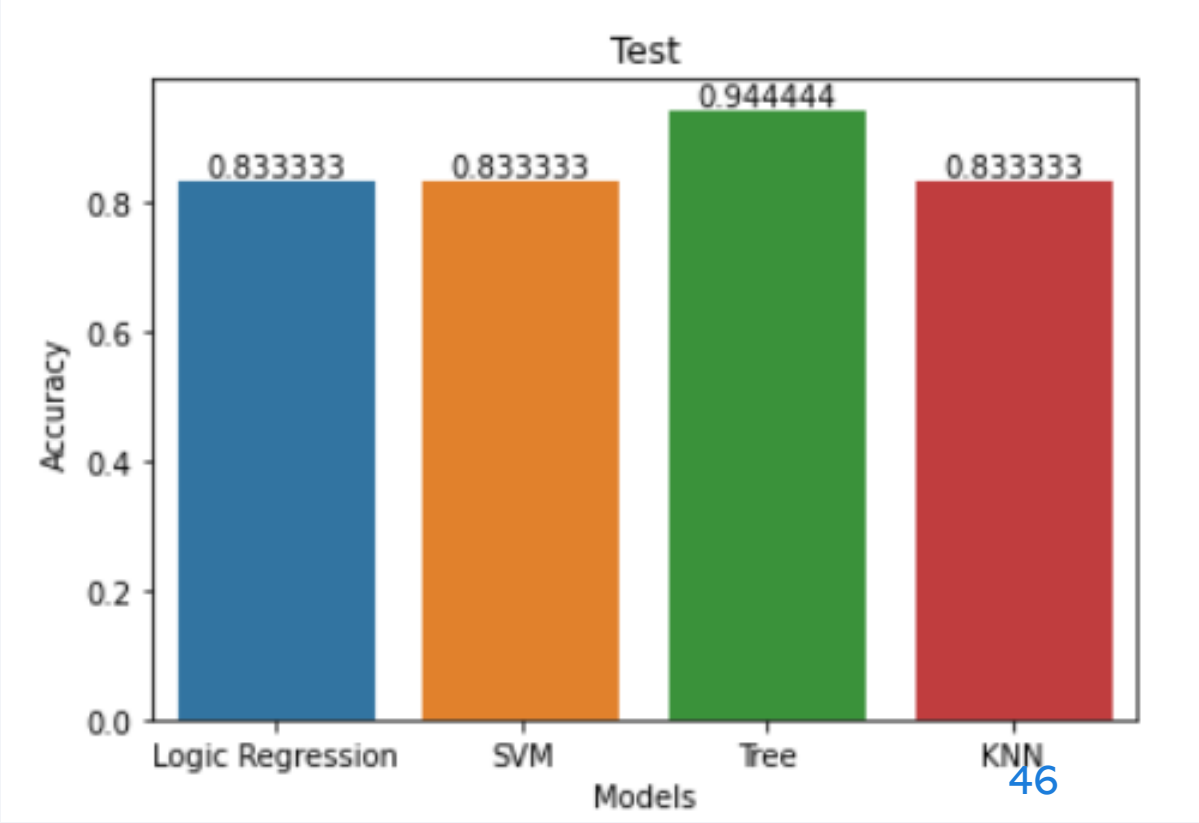
Section 5

Predictive Analysis (Classification)

Classification Accuracy

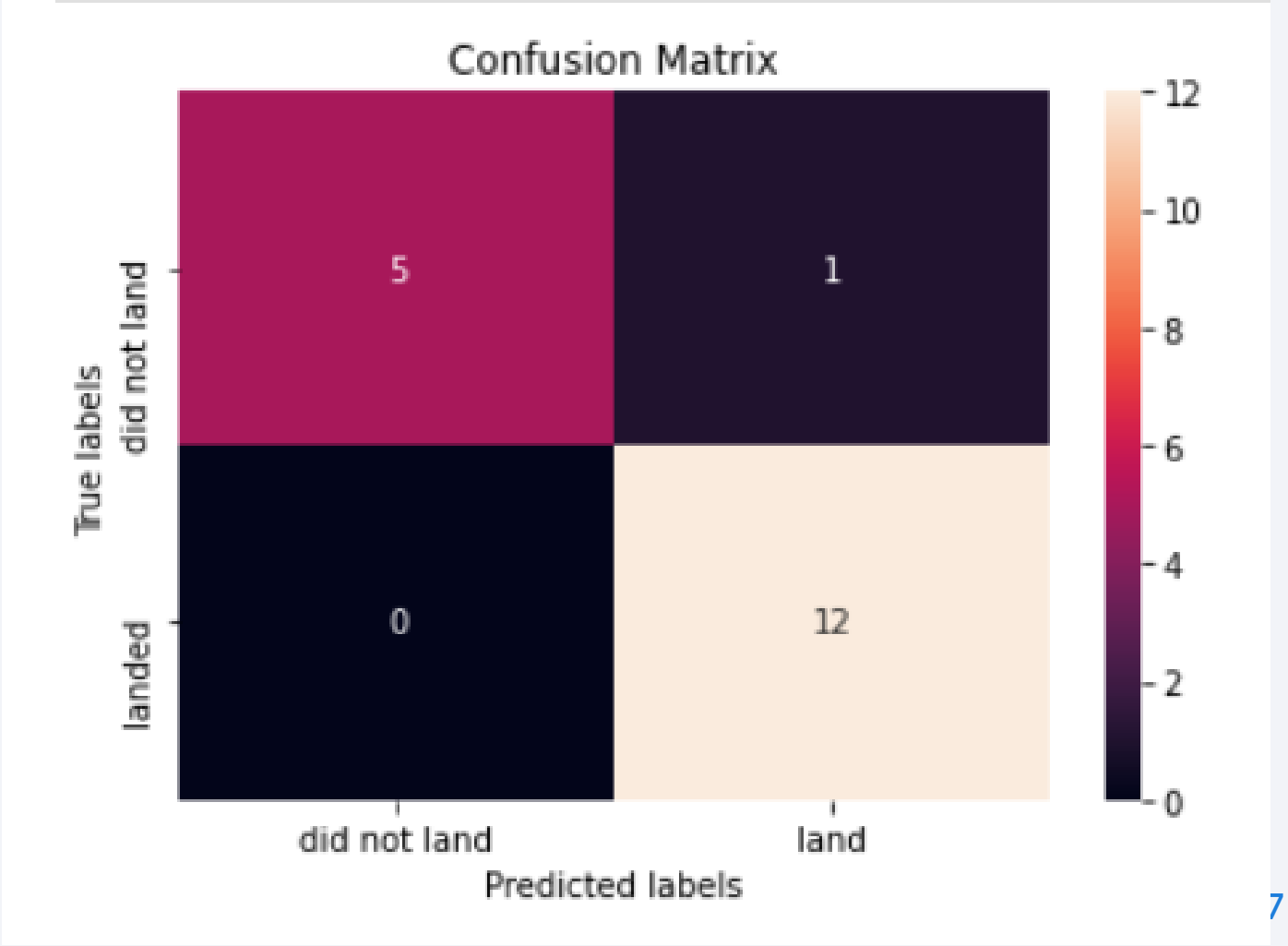
- The best result was from decision tree with training accuracy of 88.9% and test accuracy of 94.4%





Confusion Matrix

- The precision (true positive/all positive prediction)=12/13, the recall was (true positive / (true positive + false negative)) =12/12
- F1 score was 2*(12/13*12/12)/(12/13 +12/12)=0.96



Conclusions

- Success rate increase with time
- Success rate show dependency on complex variables
- However, dataset need to be expanded to better present all cases
- The decision tree best predict the success rate
- Although the prediction can be used as an indicator for predicting results, it depends on precious knowledge
- Selection of launching criteria still need to be depend on project requirements

Thank you!