IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Charlie Soon
15 Oct 2021

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data collection with SpaceX REST API and from Wikipedia
  - Data wrangling
  - Exploratory Data Analysis (EDA) with visualization
  - EDA with SQL
  - Building interactive map with Folium to study launch site geographical location
  - Building dashboard with Plotly Dash to explore data real time
  - Classification predictive analysis with logistic regression, SVM, classification trees and KNN
- Key summary of all results
  - Lower payload mass has a higher success rate
  - SpaceX success rate increase with time
  - KSC LC-39A launch site has the most success rate compared to the rest of the launch sites
  - The classification tree model has the highest accuracy for this dataset

# Introduction

- Background
  - SpaceX advertises Falcon 9 rocket launches on its website to be 62 millions dollars
  - Other providers requires a cost of more than 165 millions dollar
  - The main reason for lower launch cost is because SpaceX can reuse the first stage of the launcher if it land successfully back on Earth
  - By predicting if the first stage will land successfully after the launch, the launch cost can be determined
- Problem
  - To estimate if the first stage will land successfully based on mission parameters in the upcoming launch so that our company can bid against SpaceX

*Successful landing of Falcon 9*

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - SpaceX data is collect using the SpaceX REST API
  - Using Python Beautiful Soup
- Perform data wrangling
  - Describe how data was processed
- 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

- SpaceX data are collected thru two methods:
  - SpaceX REST API endpoints (https://api.spacexdata.com/v4)
  - Web scrapping thru Wiki pages using Python BeautifulSoup

- The next two slides explains how SpaceX REST API and BeautifulSoup is used

# Data Collection – SpaceX API

**SpaceX endpoints**

- url site used: https://api.spacexdata.com/v4/<end point name>
- It contains many end points such as rockets, launchpads, payload, etc..

**Get request**

- request.get() is used to obtain response from the SpaceX url
- Results can be view using the .json() method

**Turn response content into dataframe**

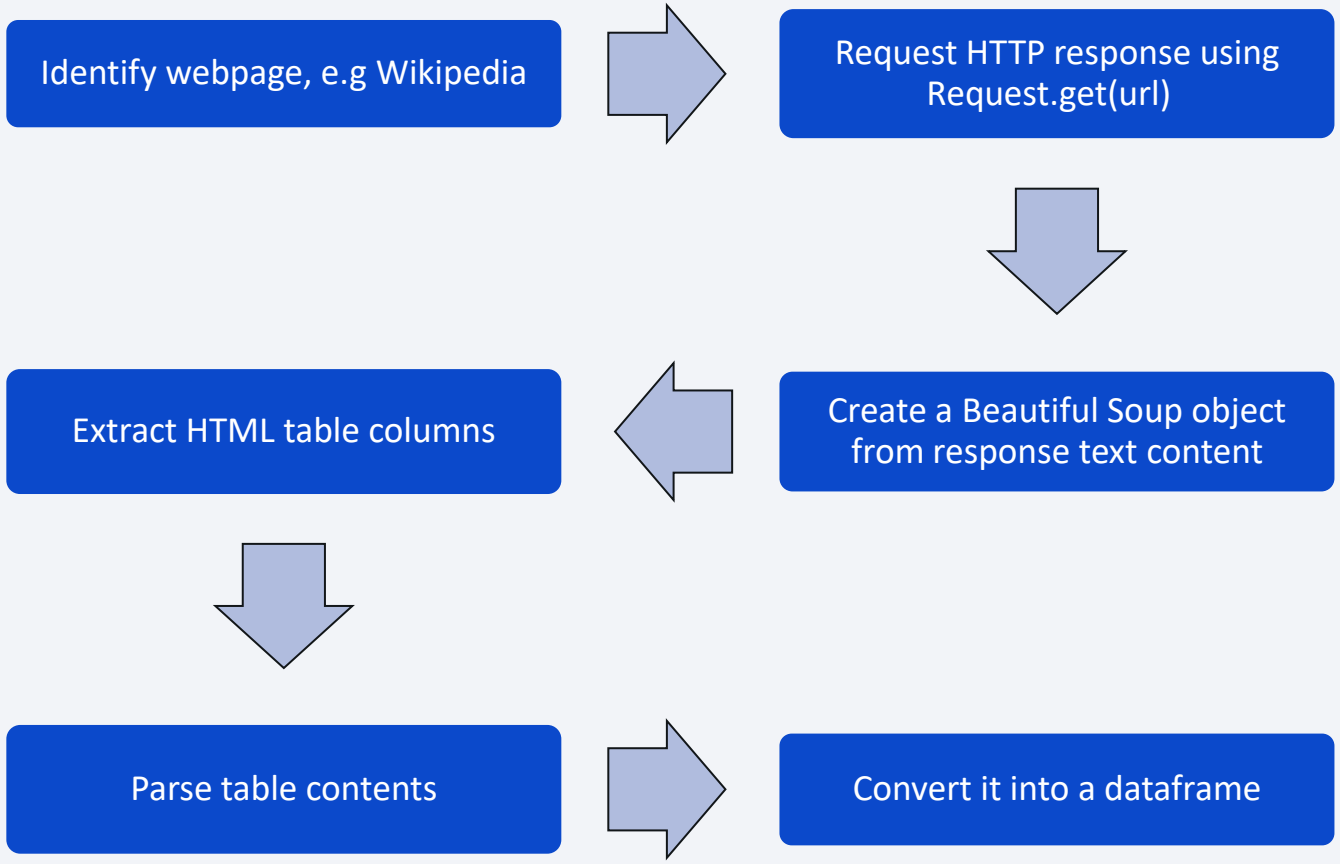
- Using pd.json_normalize(), response can be converted into a flat table as shown on the right

| | static_fire_date_utc | static_fire_date_unix | tbd | net | window | rocket | success | details | crew | ships | capsul |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2006-03-17T00:00:00.000Z | 1.142554e+09 | False | False | 0.0 | 5e9d0d95eda69955f709d1eb | False | Engine failure at 33 seconds and loss of vehicle | [] | [] | [] |
| 1 | None | NaN | False | False | 0.0 | 5e9d0d95eda69955f709d1eb | False | Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage | [] | [] | [] |

An example of a normalize json

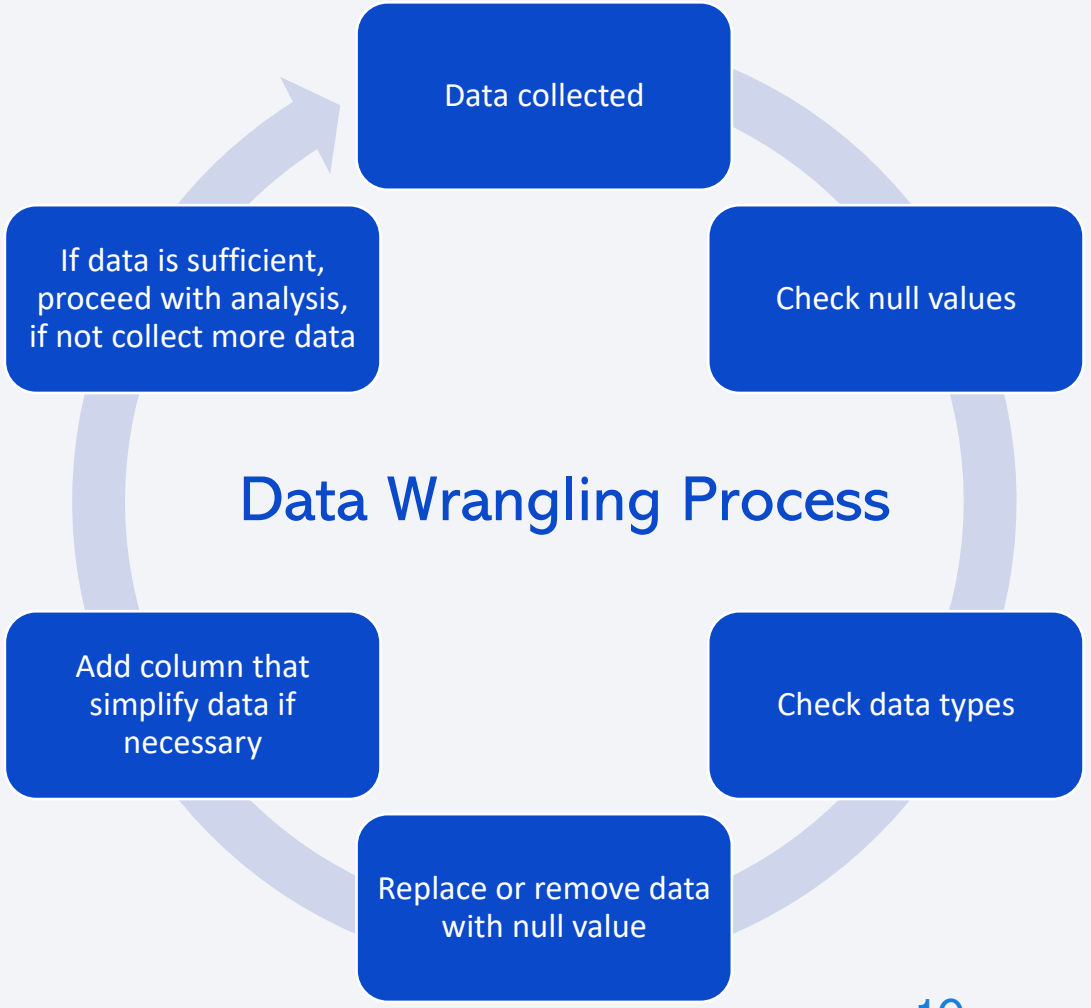
# Data Collection - Scraping

- Web scrapping is performed using Beautiful Soup
- Beautiful Soup is a Python library for pulling data out of HTML and XML files
- The flowcharts describe the steps when using Beautiful Soup



# Data Wrangling

- After data is collected, it has to be cleaned up before it can be used for analysis
- The data set for this study consists of BoosterVersion, LaunchSite, PayloadData, CoreData, etc.
- Since the key interest is on Falcon 9 launcher, only data related to Falcon 9 launch is used
- Null values in Payloads are replaced with the mean Payload Mass data
- Based on landing_outcomes, a column 'class' is created to simply the information
  - '0' for not successful
  - '1' for successful



# EDA with Data Visualization

- In this study, for exploratory data analysis and visualization, the following plots are used:
  - Scatter plot
  - Bar chart
  - Line chart
- The following parameters are plotted to study its relationships with success rate (class) shown in the table

| X axis | Y axis | |
|---|---|---|
| Flight Number | Payload Mass | If payload mass would affect launch outcome |
| Flight Number | Launch Site | If a particular launch site has higher landing success rate |
| Payload Mass | Launch Site | If there is any relationship between launch sites and their payload mass |
| Orbit | Flight Number | If there is relationship between success rate and type of orbit |
| Flight Number | Orbit | If there is relationship between flight number and orbit type |
| Payload Mass | Orbit | If there is relationship between payload mass and orbit type |
| Year | Class | If the success rate has relationship with year |

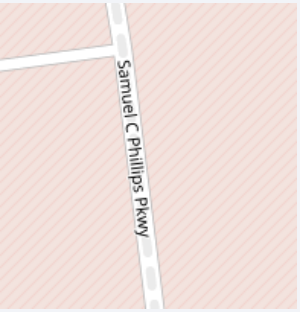
# EDA with SQL

- Using Python DB2 magic, the following SQL queries is used to explore the data

| SQL Query | EDA |
|---|---|
| SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL | Display unique launch site |
| SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'KSC%' LIMIT 5 | Display 5 records where launch sites being with string ‘KSC’ |
| SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER LIKE 'NASA (CRS)%' | Display the total payload mass carried by boosters launched by NASA (CRS) |
| SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1' | Display average payload mass carried by booster version F9 v1.1 |
| SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE 'Success (drone ship)' | List the date where the succesful landing outcome in drone ship was acheived. |
| SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE 'Success (ground pad)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 | List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000 |
| SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS COUNT FROM SPACEXTBL GROUP BY MISSION_OUTCOME | List the total number of successful and failure mission outcomes |
| SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) | List the names of the booster_versions which have carried the maximum payload mass. Use a subquery |
| SELECT MONTH(DATE) AS Month, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE YEAR(DATE) = 2017 AND LANDING__OUTCOME LIKE 'Success (ground pad)' | List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017 |
| SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20’ AND LANDING__OUTCOME LIKE ‘success%’ GROUP BY LANDING__OUTCOME ORDER BY COUNT DESC | Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order. |

# Build an Interactive Map with Folium

- To study if location of launch site and landmarks has relationship with success rate, an interactive map is build using Folium to answer the following questions:
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?
- The following steps are taken:
  - Launch sites are marked on the map using it's latitude and longitude with a circle area
  - Success and failed launches for each site is added
  - The distance between the launch site to symbols such as railway, highway and cities are calculated using coordinate labelled with a poly line object


# Build a Dashboard with Plotly Dash

- To help with visual analytics on launch data in real time, a dashboard is created using Plotly Dash to answer the following questions:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
- The following dashboard are created
  - Pie chart to display launch success rate vs launch site
  - Scatter plot to display payload mass vs success rate for each booster version by launch site selection

# Predictive Analysis (Classification)

- To select the best classification model, three mains steps has to be performed (see image on right)
  - Standardize the data – independent variables are transform standard normally distributed data using the *sklearn.preprocessing,standardscaler* library
  - Split data into training and test data – Data is separate randomly into data used for training the model and testing the model using the *train_test_split function*
  - Finding the best hyperparameters for models – The best parameters for different classification models are found using the *GridSearchCV* function by feeding in the estimator, the parameter range to be optimized and training data
- In this study, the SVM, classification tree and logistic regression model are explored to find the most suitable model for prediction

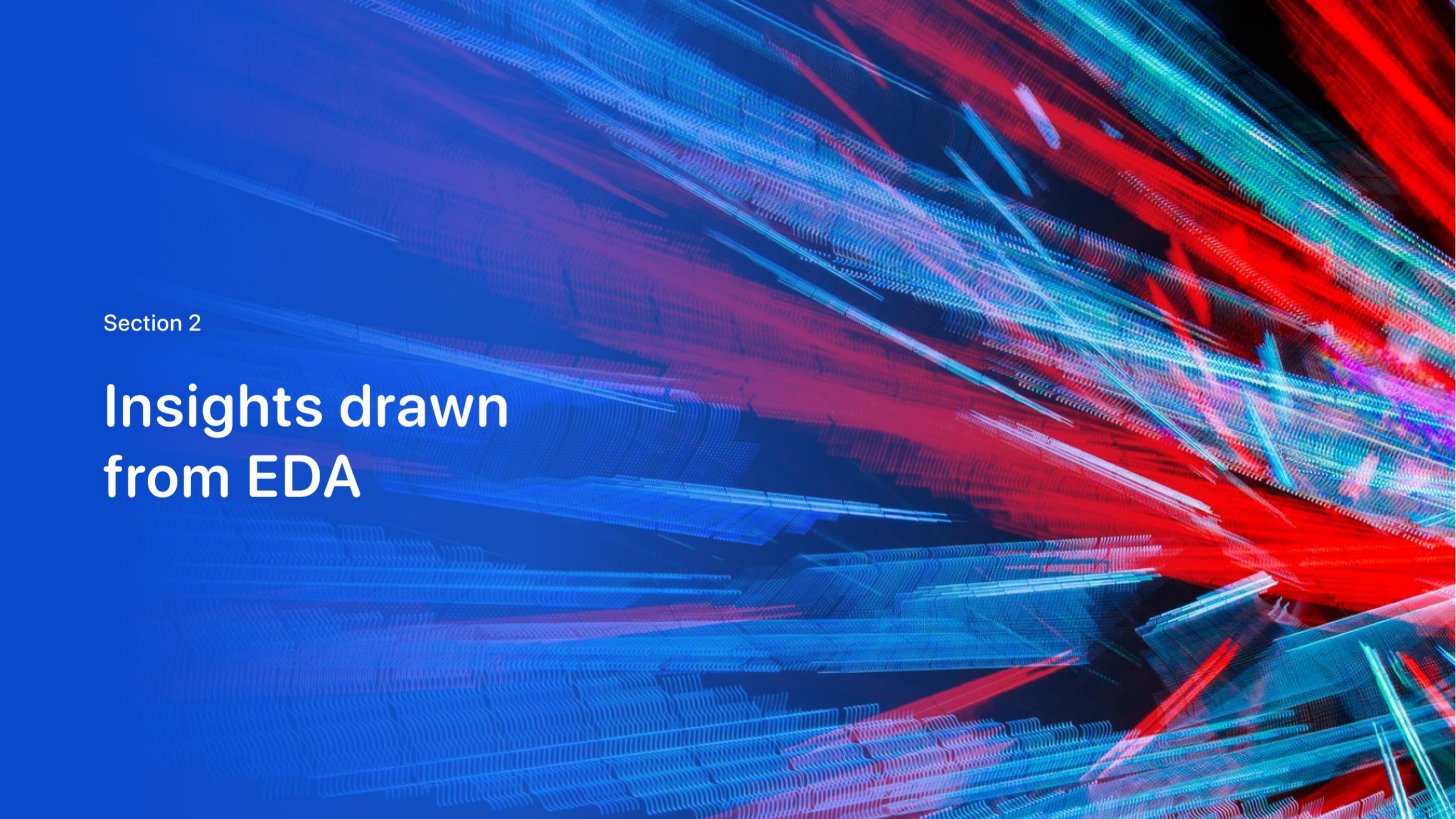

Perform data predicting using test data with model to calculate the score to select best model

# Results

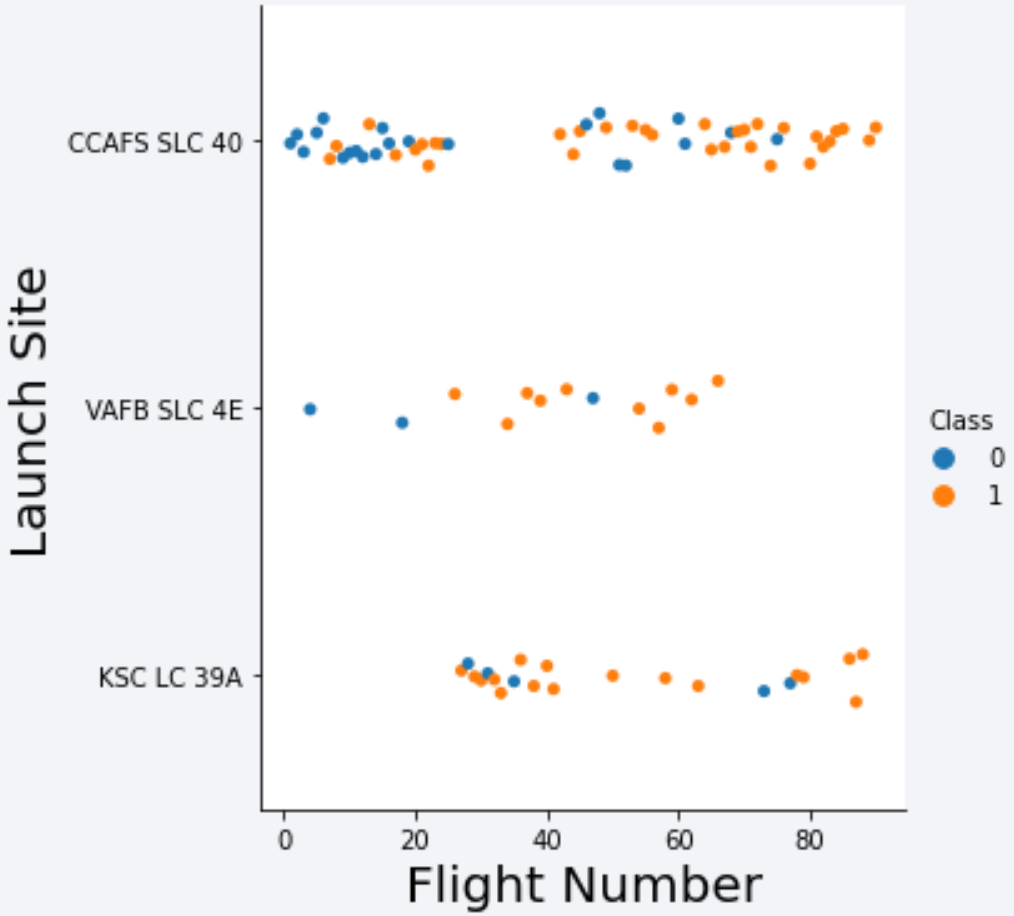
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

# Insights drawn from EDA

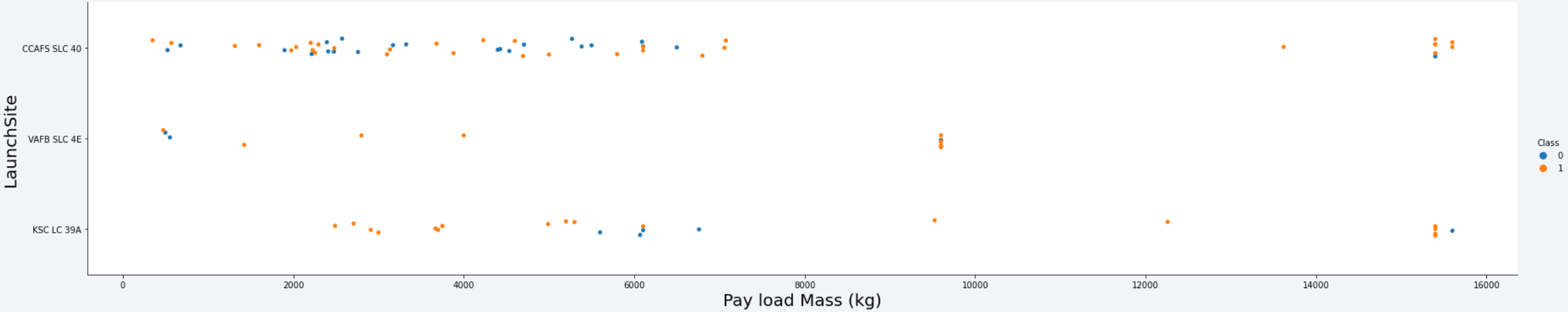


# Flight Number vs. Launch Site



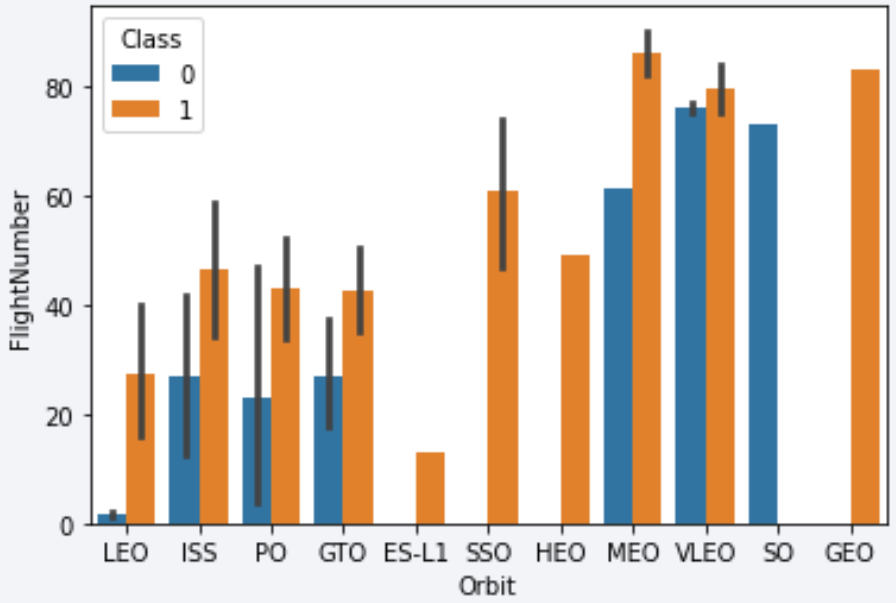

- The scatter plot shows the flight number vs. launch site
- From the plot, it is observed that
  - CCAFS SLC-40 has the highest number of launches and the success rate increases with number of flights
  - Between flight number 20 to 40, CCAFS SLC-40 site has no launches and most likely launches are done in other launch sites
  - VAFB-SLC-4E site has the least number of launch

# Payload vs. Launch Site



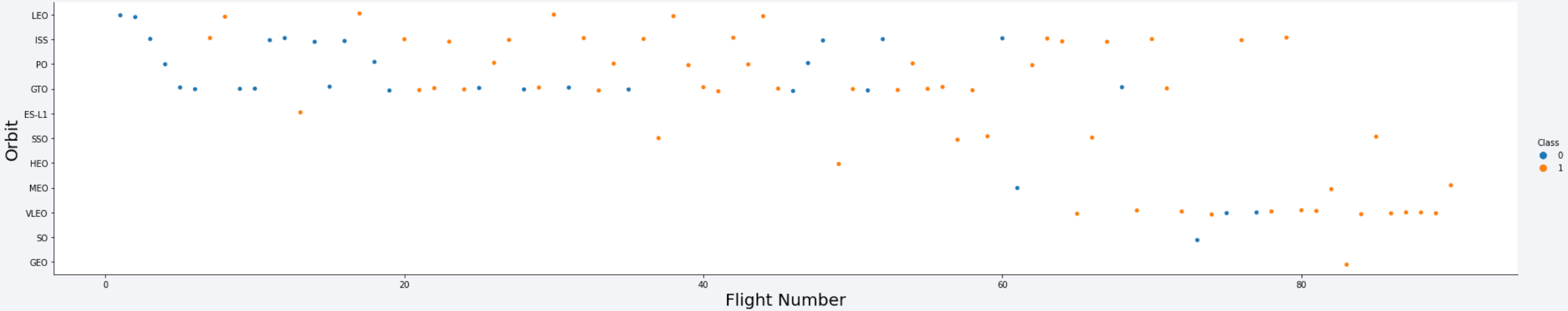

- The plot shows the scatter plot of payload vs. launch site
- It is observed that
  - Most of the payload mass launches for all three sites are between 0 to 8000kg
  - There are fewer launches with payload > 15000kg
  - CCAFS-SLC-40 site has the most launches for payload mass of 0 to 8000kg

# Success Rate vs. Orbit Type



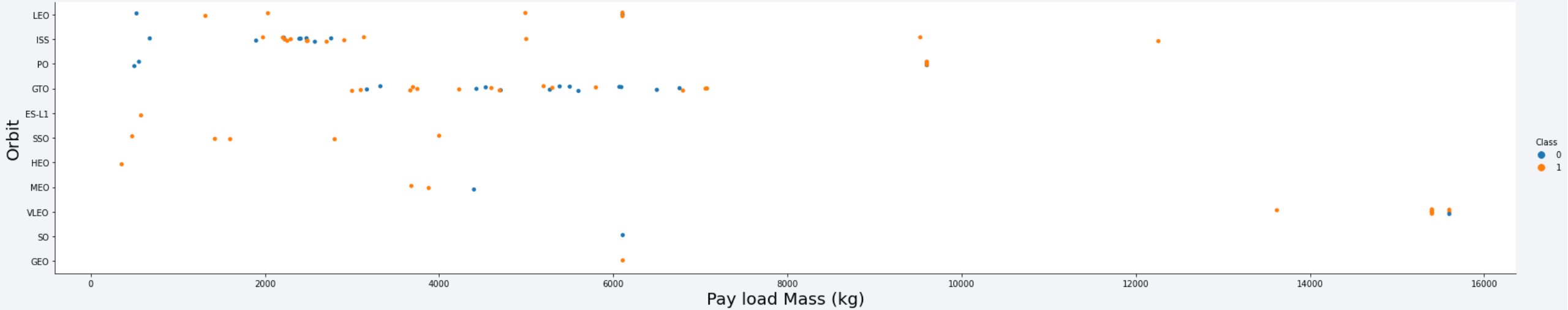

- The plot shows the bar chart of success rate vs orbit type
- It is observed that
  - ES-L1, SSO, HEO and GEO orbit has 100% success rate
  - SO has success rate is 0%

# Flight Number vs. Orbit Type



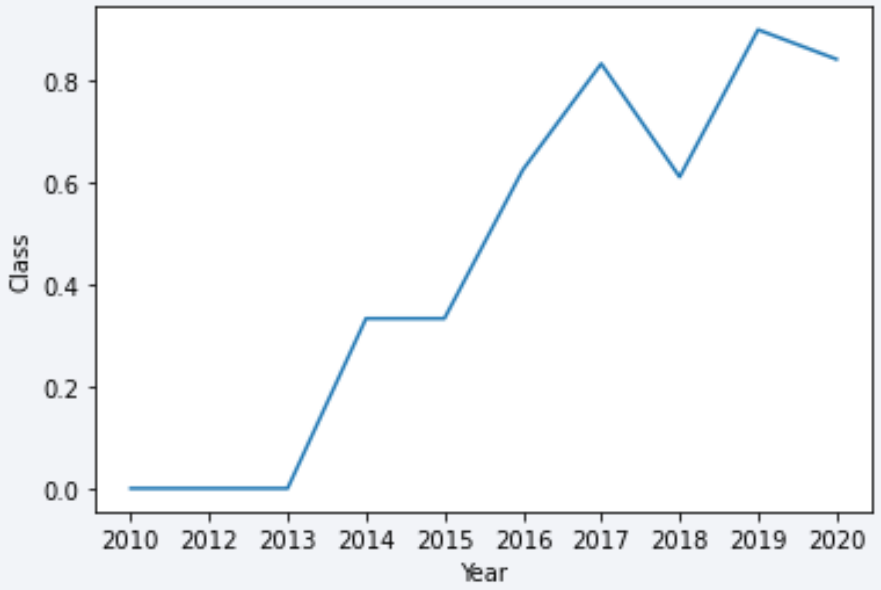

- The plot shows the scatter plot of flight number vs. orbit type
- It is observed that
  - For LEO, ISS and VLEO orbit, the success rate increase with flight number
- From the previously slide, it is seen that ES-L1, SSO, HEO and GEO orbit has 100% success rate
  - From the scatter plot, note that for these orbits, the number of flight number is less than 5

# Payload Mass vs. Orbit Type




- The plot shows the scatter plot of payload mass vs. orbit type
- It is observed that
  - Heavier payload affects the success rate of launch to GTO orbit
  - Increase in payload mass up has positive effects to ISS and LEO orbit

# Launch Success Yearly Trend




- The plot shows the line plot of success rate yearly trend
- It is observed that after 2013, the success rate shows an upward trend with time showing that SpaceX is learning from failure and improving over the years

# All Launch Site Names

- Find the names of the unique launch sites
  - There are a total of 4 unique launch sites used for Falcon 9 launch

```
In [4]: %sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL
 * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[4]:

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

# Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
  - The records for launch site beginning with CCA is presented. The records of both CCAFS SLC-40 and CCAFS LC-40 will be obtained

```
In [20]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[20]:

| 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 |

# Total Payload Mass

- Calculate the total payload carried by boosters from NASA (CRS)
  - The total payload mass carried by NASA (CRS) is 48213kg

```
In [19]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_MASS FROM SPACEXTBL \
             WHERE CUSTOMER LIKE 'NASA (CRS)%'
 * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[19]:

| total_mass |
|---|
| 48213 |

# Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - The total average payload mass carried by F9 v1.1 booster is 2928kg

```
In [22]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_MASS FROM SPACEXTBL \
             WHERE BOOSTER_VERSION LIKE 'F9 v1.1'
 * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
Out[22]:
```

| average_mass |
|---|
| 2928 |

# First Successful Ground Landing Date

- Find the date of the first successful landing outcome on ground pad
  - The first success ground pad launch is as early as 2015-15-22

```
In [24]: %sql SELECT MIN(DATE) FROM SPACEXTBL \
             WHERE LANDING__OUTCOME LIKE 'Success (ground pad)'
          * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
         Done.
```

Out[24]:

| 1 |
|---|
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - The booster version are listed in the query output below

In [26]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
        WHERE LANDING__OUTCOME LIKE 'Success (drone ship)' \
        AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[26]:

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

# Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - The total number of mission is 101 with 99 success mission and 1 failure mission. 1 of the mission is success, however the status of the payload is unclear

```
In [10]: %sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS COUNT FROM SPACEXTBL \
         GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[10]:

| mission_outcome | COUNT |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - The maximum payload is 15600kg and the booster which have carried the maximum payload mass is shown in the query output below

```
In [11]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
             WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
 * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[11]:

| 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 |

# 2017 Launch Records

- List the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
  - MONTH represents the number of the month. E.g. 2 is February
  - In 2017, there are a total of 6 successful landing on ground pad from F9 FT and B5 boosters in launch site KSC-LC-39A and CCAFS-SLC-40

```
In [12]: %sql SELECT MONTH(DATE) AS Month, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL \
         WHERE YEAR(DATE) = 2017 AND LANDING__OUTCOME LIKE 'Success (ground pad)'
```

```
 * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[12]:

| MONTH | landing__outcome | booster_version | launch_site |
|---|---|---|---|
| 2 | Success (ground pad) | F9 FT B1031.1 | KSC LC-39A |
| 5 | Success (ground pad) | F9 FT B1032.1 | KSC LC-39A |
| 6 | Success (ground pad) | F9 FT B1035.1 | KSC LC-39A |
| 8 | Success (ground pad) | F9 B4 B1039.1 | KSC LC-39A |
| 9 | Success (ground pad) | F9 B4 B1040.1 | KSC LC-39A |
| 12 | Success (ground pad) | F9 FT B1035.2 | CCAFS SLC-40 |

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

- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order
  - The total of successful launching outcome between 2010-06-04 and 2017-03-20 is 8
  - 5 of them are successful landing on drone ship and 3 are successful landing on ground pad

```
In [32]: %sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT FROM SPACEXTBL \
             WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND LANDING__OUTCOME LIKE 'Success%' GROUP BY LANDING__OUTCOME \
             ORDER BY COUNT DESC

          * ibm_db_sa://tjr72306:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
        Done.
```

Out[32]:

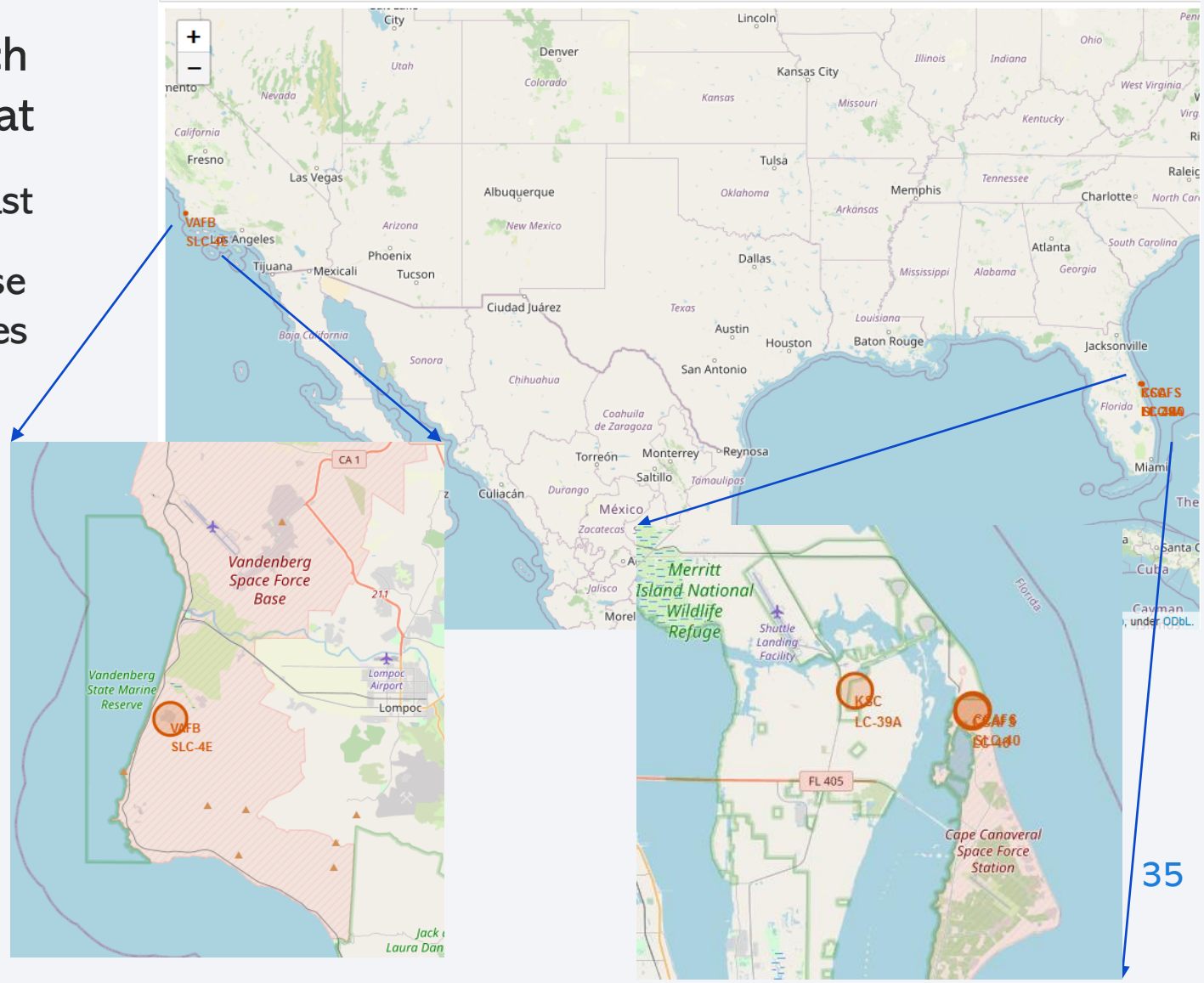
| landing__outcome | COUNT |
|---|---|
| Success (drone ship) | 5 |
| Success (ground pad) | 3 |

Section 4

# Launch Sites Proximities Analysis

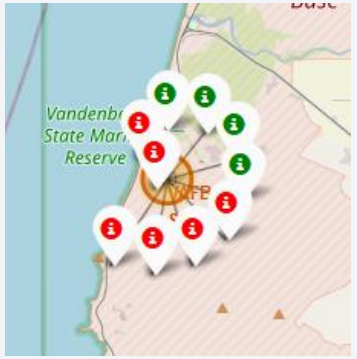
# Visualizing Launch Site Locations

- From the visualization of launch sites on map, it is observed that
  - Launch sites are close to the coast
  - They are located away from dense populated area, mainly in airbases



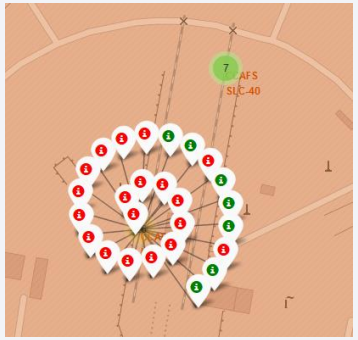
# Launch Site Success Rate

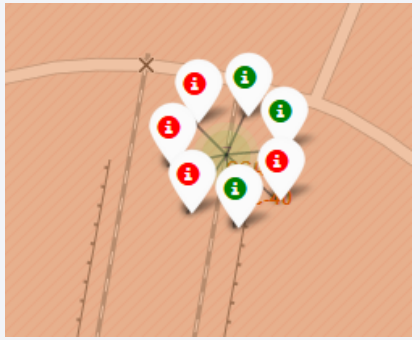
- The success rate for each launch site is displayed in the figures
  - Red marker shows failure
  - Green marker shows success
- From the figures, it is observed that
  - CCAFS LC-40 has the most launches
  - KSC LC-39A has the highest success rate



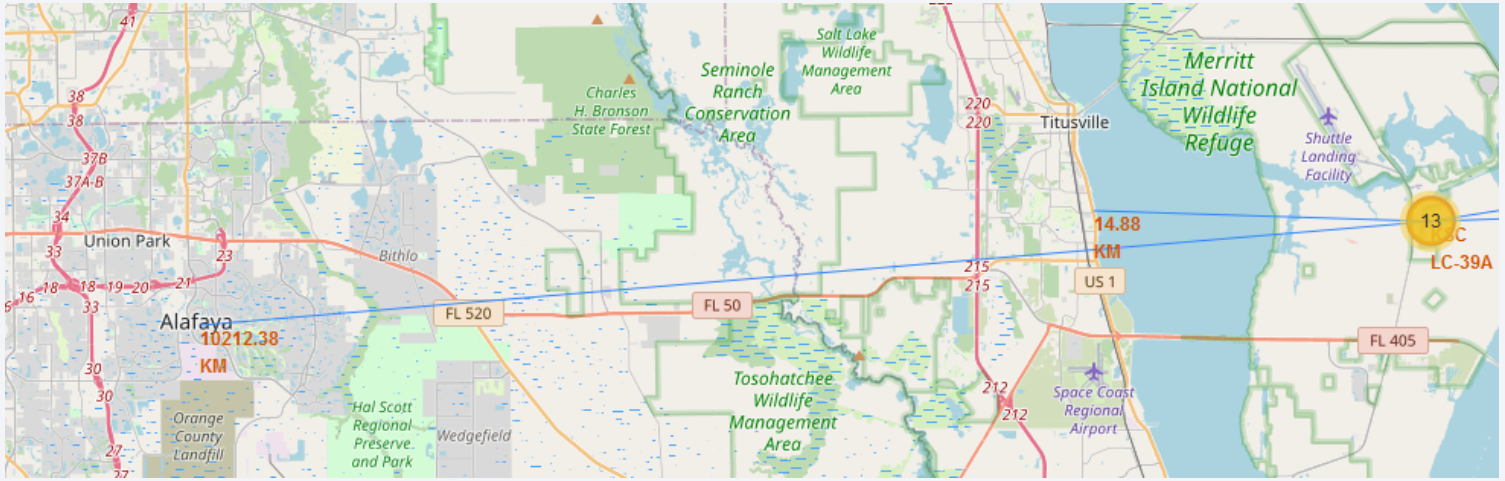
VAFE SLC-4E

KSC LC-39A


CCAFS LC-40


CCAFS SLC-40

# Proximity of KSC LC-39A Launch Site

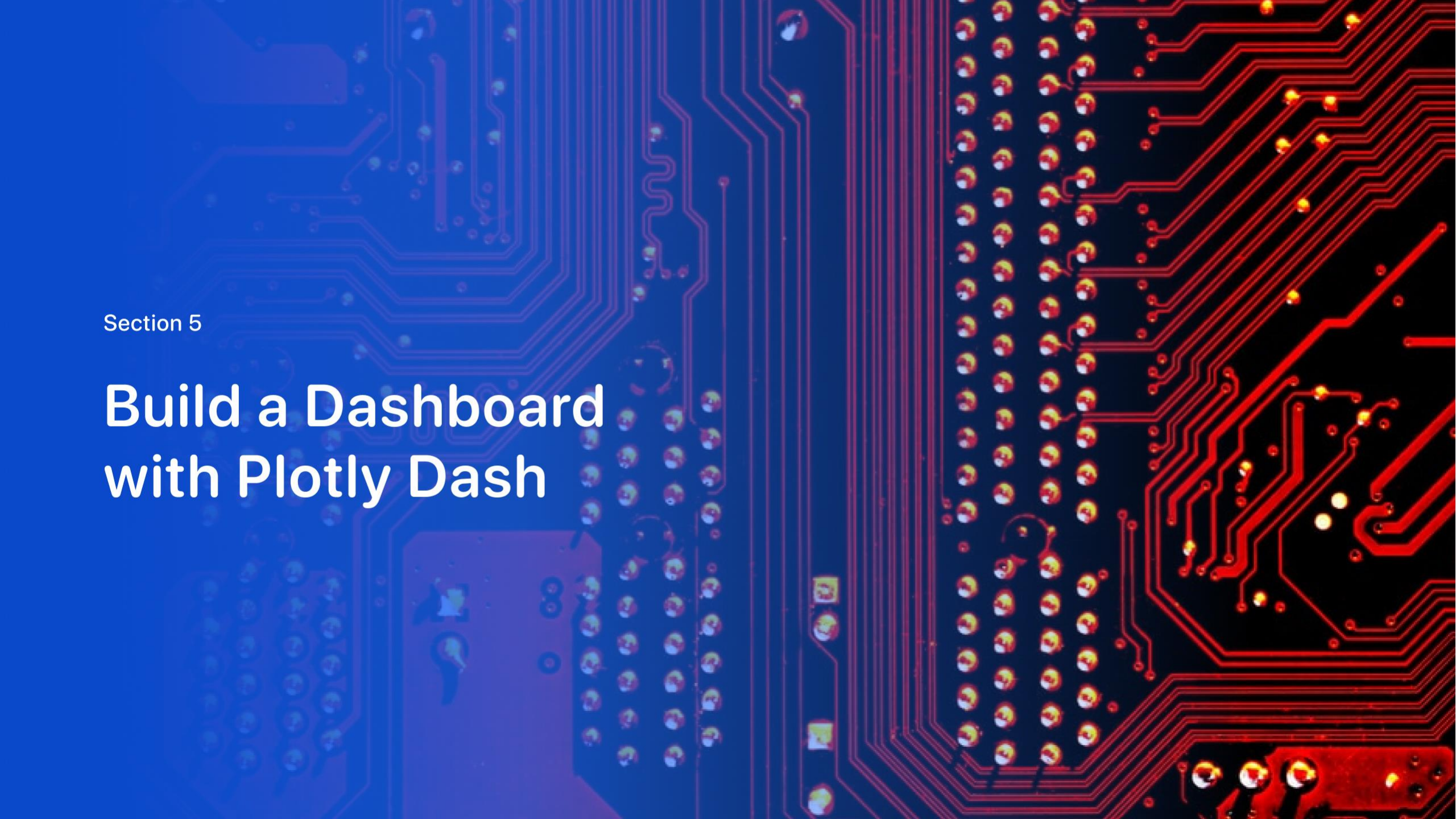




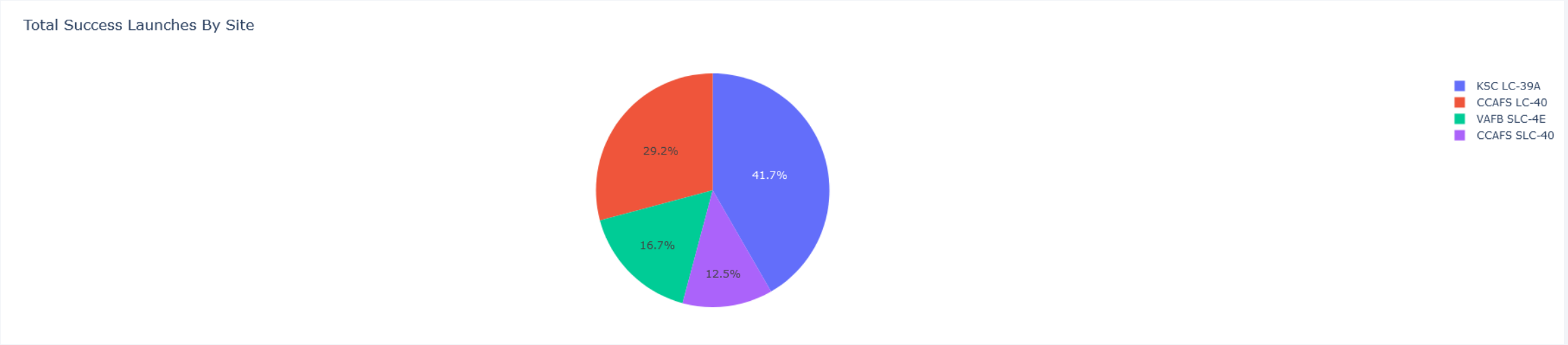
- From previous slides, KSC LC-39A launch site has the highest success rate, we will be studying if the launch site to its proximities has any relationships
- It is observed that
  - KSC LC-39A is near to the coastline
  - KSC LC-39A is near to a railway which is used to travel to the landing facility
  - KSC LC-39A is away from highways and cities

Section 5

# Build a Dashboard with Plotly Dash

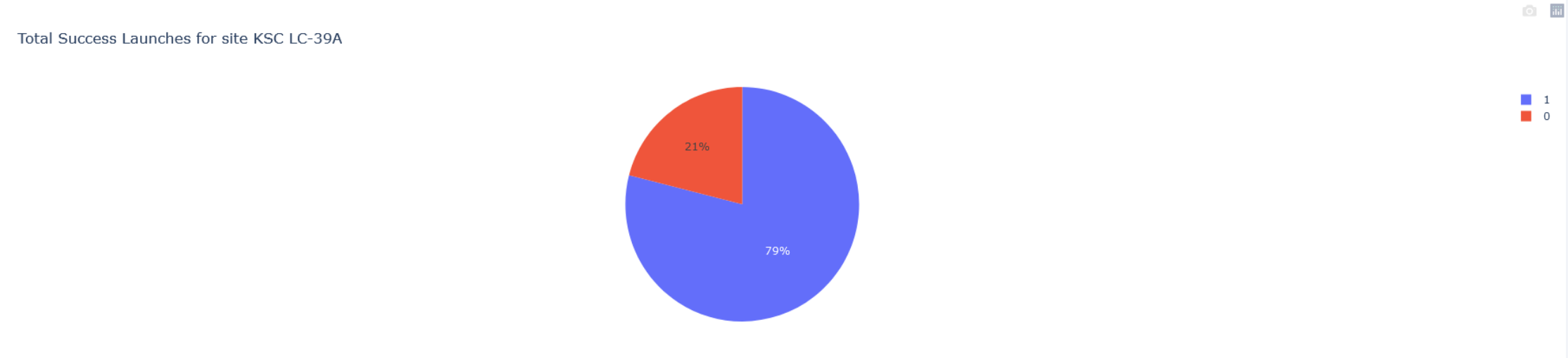


# Success Rate by Launch Site




- The pie chart shows the success launches by launch site
- KSC LC-39A launch site has the highest success rate
- CCAFS SLC-40 launch site has the lowest success rate

# Success Rate for KSC LC-39A




- From the previous slides, it is shown that KSC LC-39A has the highest success rate
  - 79% of them are success rate

# Success Rate for Payload between 0 to 7500kg


- Most of the launches payload mass range from 0 to 7500kg
- Higher success rate for payload mass between 2000 to 5500kg
- The highest success rate booster version for 0 to 7500kg is FT
- The highest failure rate booster version is v1.1 for 0 to 7500kg

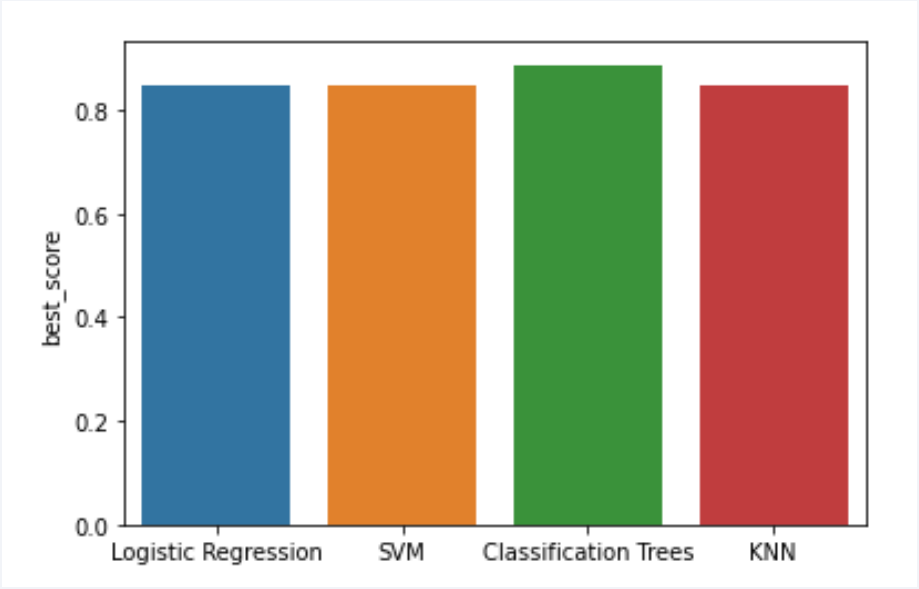
Section 6

# Predictive Analysis (Classification)

# Classification Accuracy

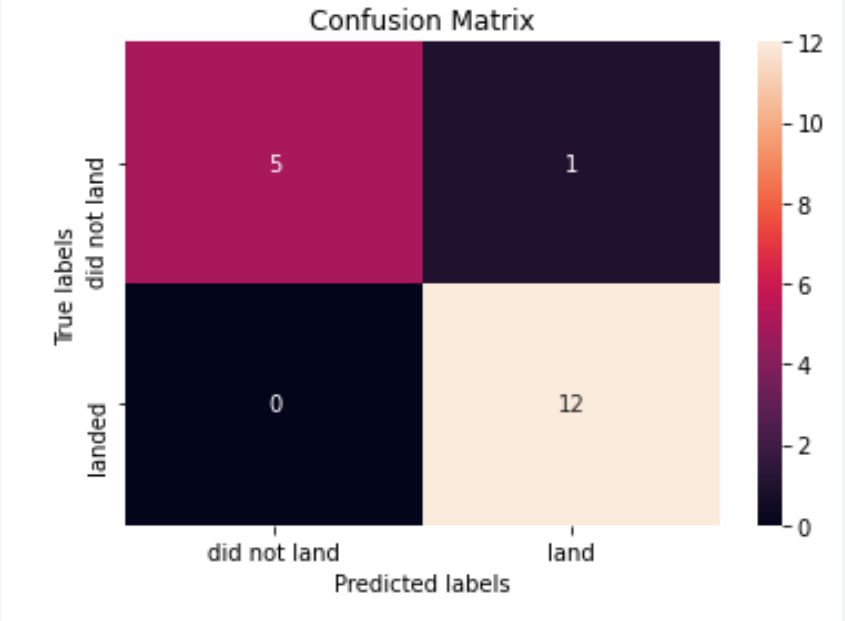
- Four models are trained with the same set of data to find the best hyperparameters
  - From the barplot, classification trees has the highest accuracy
  - The model best score is 0.8875000000000002, with the following parameters:

*{'criterion': 'entropy', 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}*



# Confusion Matrix – Classification Tree

- From the confusion matrix
  - The classification tree model is accurate in predicting the landed (12 out of 12 correct)
  - For true did not land, the accuracy is 83.33% (1 out of 6 incorrect)



# Conclusions

- CCAFS SLC-40 has the highest number of launches
- Most of the launches have payload mass < 8000kg
- For LEO, ISS and VLEO orbit, the success rate increase with flight number
- Success rate of SpaceX increase with time
- The launch sites are close to the coast and away from populations
- KSC LC-39A launch site has the highest success rate
- Higher success rate for payload mass between 2000 to 5500kg
- Highest failure rate booster version is v1.1
- Classification Tree has shown to have higher accuracy for this dataset

# Appendix – Github link

https://github.com/charliesoon/IBM-Data_Science_Professional_Cert_Program

# Appendix – Calculate distance between latitude and longitude

## Haversine formula

*Haversine formula:*

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

*where* $\varphi$ *is latitude,* $\lambda$ *is longitude,* R *is earth's radius (mean radius = 6,371km); note that angles need to be in radians to pass to trig functions!*

Reference: https://www.movable-type.co.uk/scripts/latlong.html

# Thank you!