IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kuniadi Wandy Huang

31 October 2021

Outline

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

Executive Summary

- SpaceX is able to save cost of their rocket launch by reusing the launch's first stage. For an alternate company to bid against SpaceX for a rocket launch, they would like to predict whether the first stage will land. **If they can determine if the first stage will land, the cost of a launch can be determined**

- Historical launch data are collected from SpaceX API and SpaceX Wikipedia page. Data are mined and stored in cloud database. Exploratory Data Analysis (EDA) and data visualization was conducted using Python libraries.

- Prediction of landing success was conducted using different Machine Learning model: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. **All models show similar accuracy score of 83.3%**

Introduction

- SpaceX cost of its Falcon 9 rocket launch is ~100 million USD cheaper than other rocket provider because they can reuse the first stage.
- For SpaceY to bid against SpaceX for a rocket launch, they need to know the cost of a launch by knowing whether first stage will land
- From historical rocket launch data, SpaceY would like to know whether we can use Machine Learning models to predict the first stage of the rocket launch will be successful or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect data from SpaceX API and Wikipedia
- Perform data wrangling
  - Data obtained from SpaceX and Wikipedia is combined, removed and replaced missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different classification models are tested, evaluated, tuned, and compared

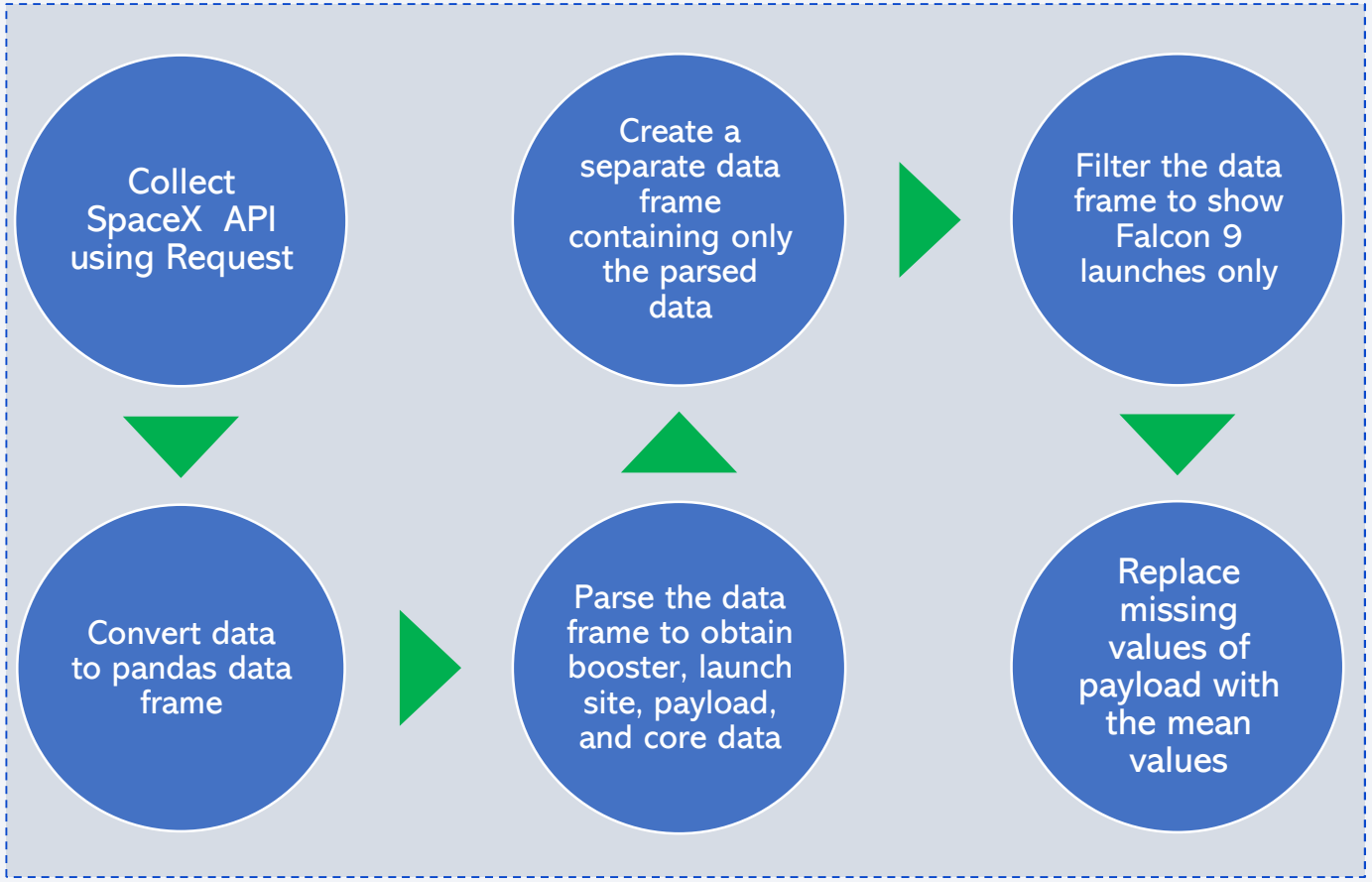
Data Collection – Executive Summary

- SpaceX API data is collected using Requests, and SpaceX Wikipedia data is collected using Beautiful Soup.
- In SpaceX API data, the features / columns are Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcomes, Flights, Grid Fins, Reused, Landing Pad, Block, Reused Count, Serial, Longitude, and Latitude
- In SpaceX Wikipedia page, the features / columns are

Data Collection – SpaceX API

- SpaceX API URL: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json

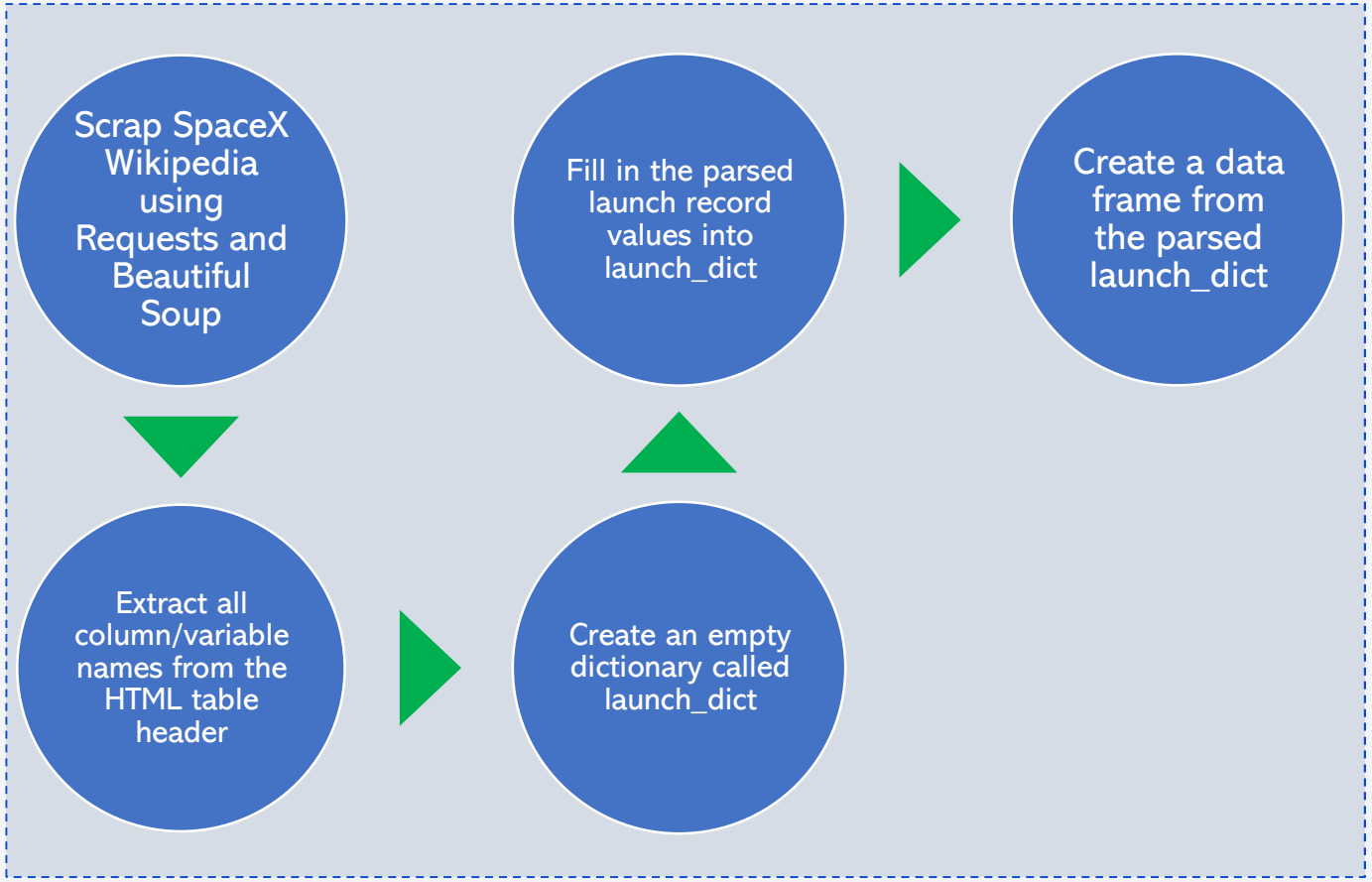
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang/blob/main/Data%20collection.ipynb



Data Collection - Scraping

- SpaceX Wikipedia URL: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

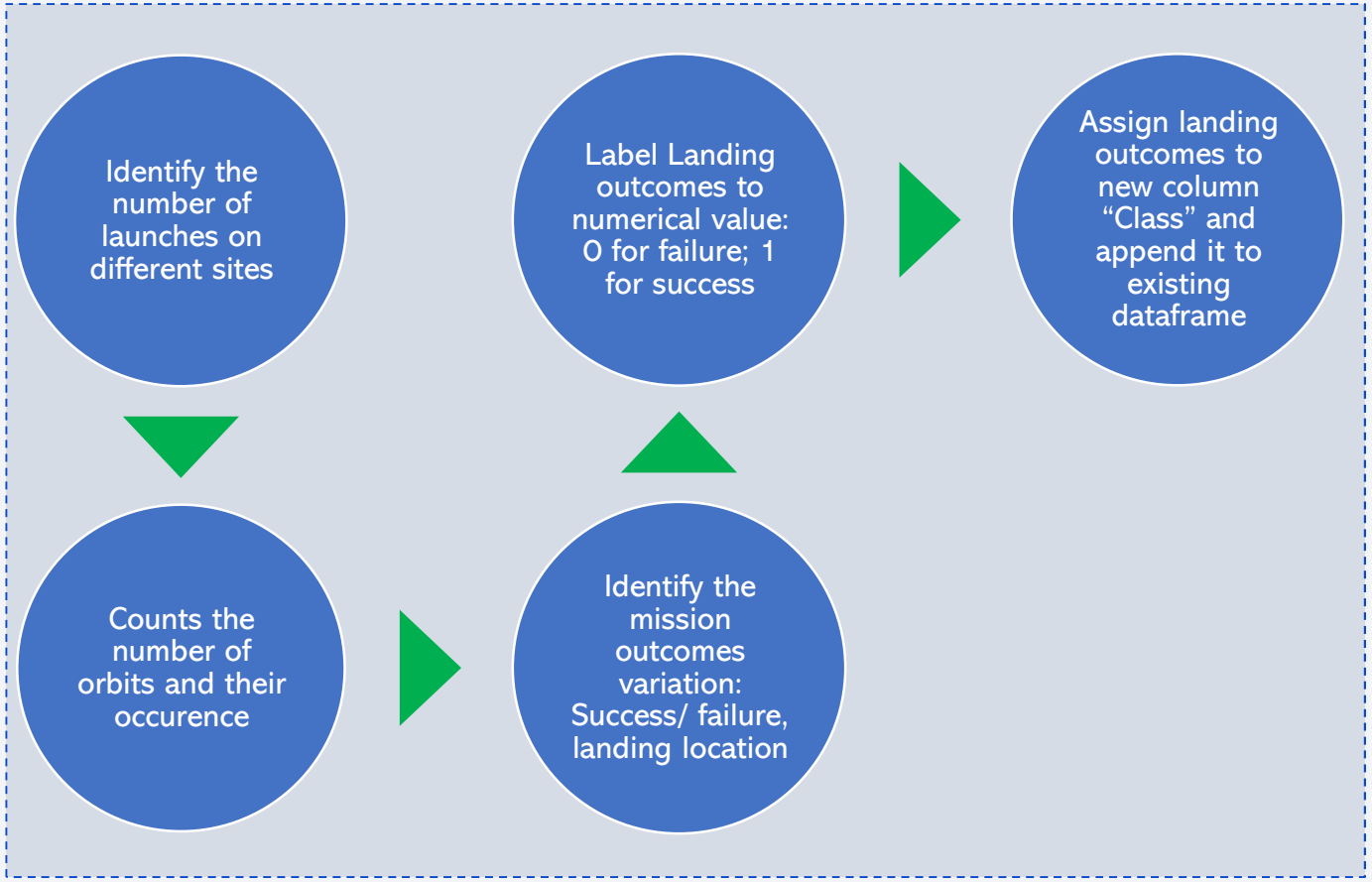
- GitHub URL of the completed SpaceX Wikipedia notebook: https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang/blob/main/data%20collection%20with%20webscraping.ipynb



Data Wrangling

- Some data contains alphabetical value, instead of numerical one. Data wrangling aims to convert this data into usable format before Exploratory Data Analysis (EDA)

- GitHub URL of the completed data wrangling notebook: https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang/blob/main/Data%20wrangling.ipynb



EDA with Data Visualization

- The following charts are used:
  - Scatter Chart: Payload vs Flight number, Launch Site vs Flight number, Launch Site vs Payload mass, Orbit vs Flight number, Orbit vs Payload mass
  - Bar Chart: Class vs Orbit
- We use scatter chart to observe and show relationships between two numeric variables, and bar chart to compare things between different groups
- GitHub URL of the completed EDA notebook: https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang/blob/main/Data%20visualization.ipynb

EDA with SQL

- The following SQL queries was done during EDA:
  - Identify unique launch sites
  - Identify launch sites begin with the string 'CCA'
  - Calculate total and average payload mass carried by boosters launched by NASA (CRS) using SUM and AVERAGE, respectively
  - Identify the date of first successful landing in ground pad
  - Identify boosters used in successful landing at drone ship with 4,000 ≤ payload mass ≤ 6,000
  - Identify the total number and individual number of successful and failure mission outcomes at different landing sites
- GitHub URL of the completed EDA with SQL notebook: https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

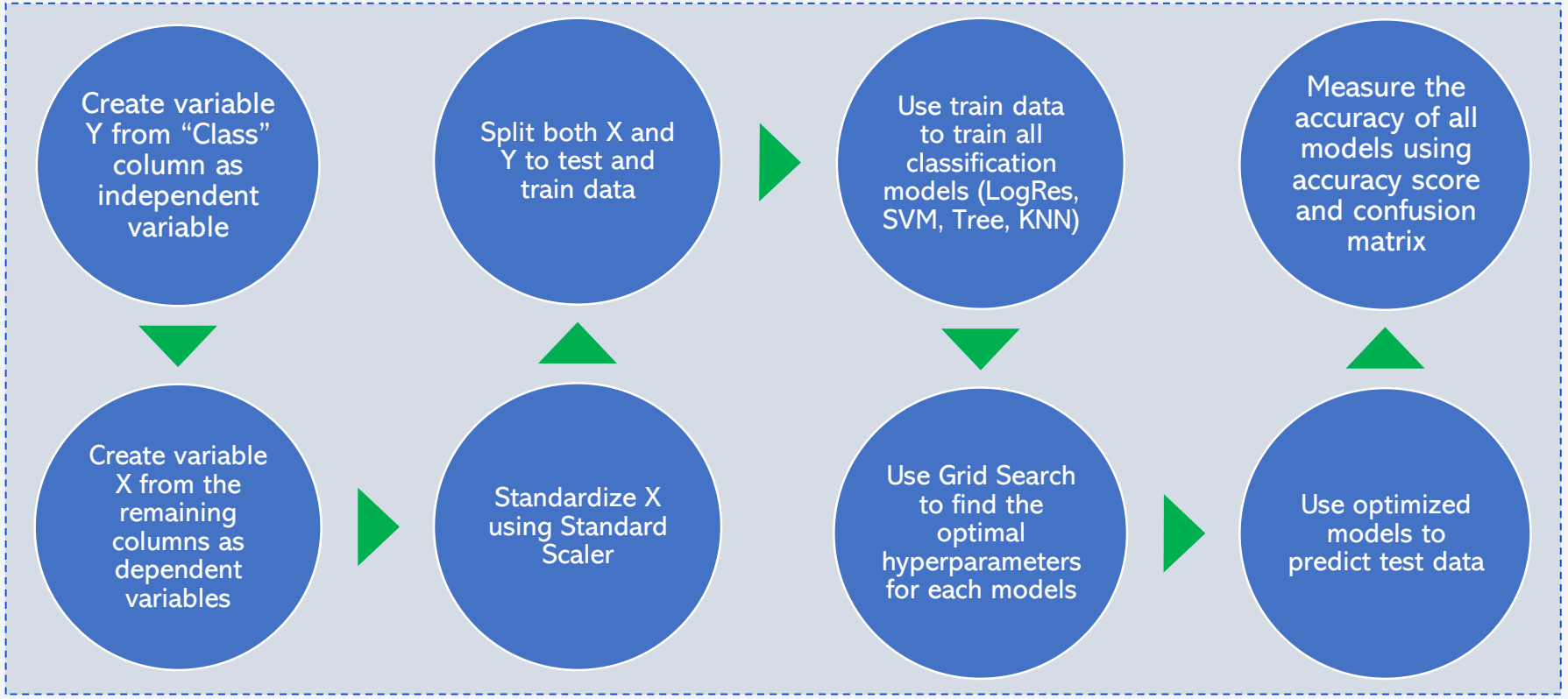
- Folium is used to mark locations of all launch sites, which landing site is a success or a failure, and calculate the distance between launch sites and nearest coastline, railway, highway, and city
- The following marker are used: circle, cluster, and line
- GitHub URL of the completed Folium notebook: https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang/blob/main/Folium%20and%20data%20visualization.ipynb

Build a Dashboard with Plotly Dash

- Two charts were created with the following purpose:
  - Pie chart: To compare the landing success rate of all sites and individual sites
  - Scatter chart: To identify the correlation between payload and the success rate of different sites. Slider were added, so the charts can be adjusted at different payload mass range
- GitHub URL of the completed plotly dash program: https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Different classification model was used: Log Regression, SVM, Decision Tree, and KNN
- Models are optimized and the performance were compared based on its accuracy score
- GitHub URL of the completed predictive analysis notebook: https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang/blob/main/Predictive%20Analysis%20(Classification).ipynb

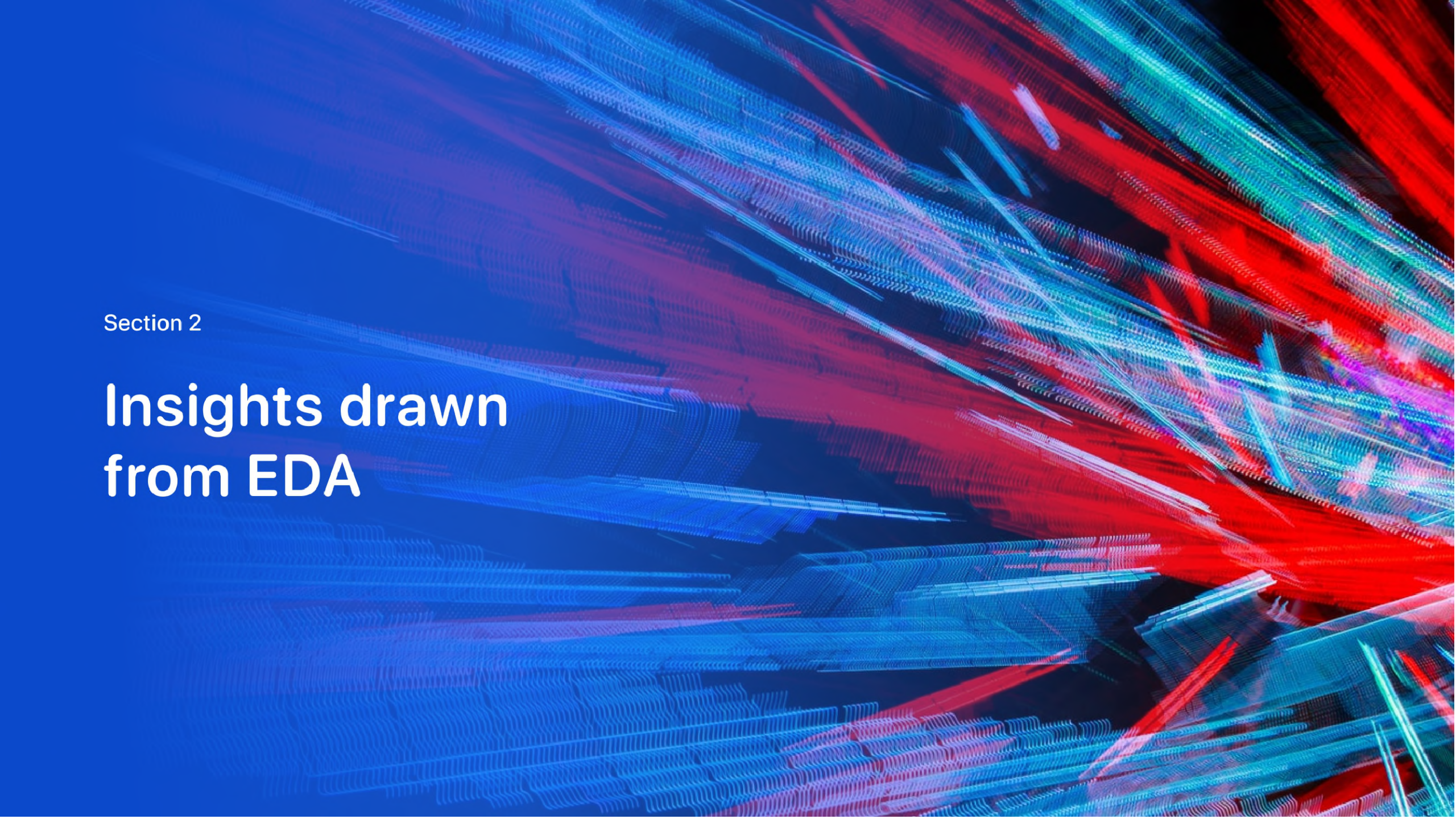


Results

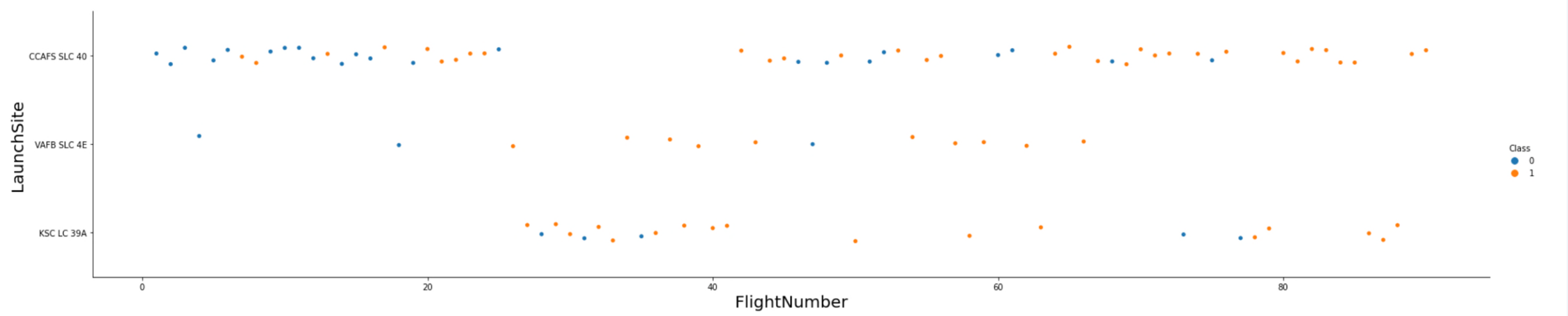
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

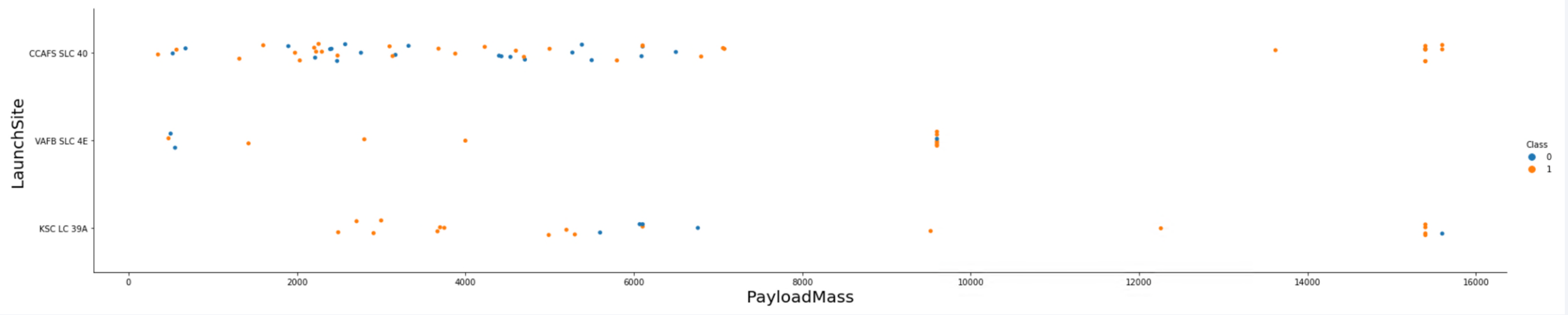


Flight Number vs. Launch Site



- The graph shows that the success rate of the landing increases with flight number
- There is no obvious relation between the location of launch sites with the success rate
- Launching sites CCAFS SLC 40 is the most often used sites

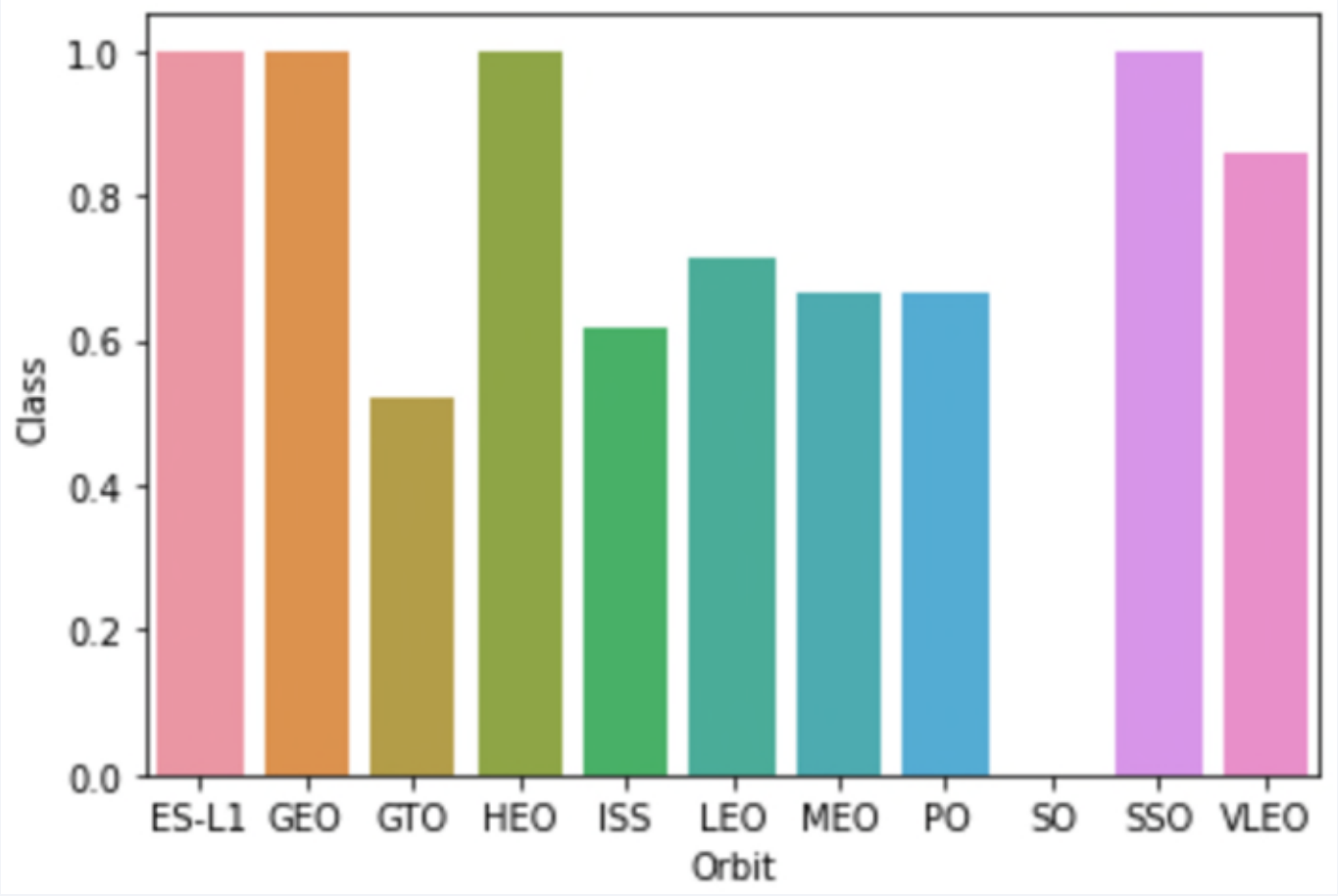
Payload vs. Launch Site



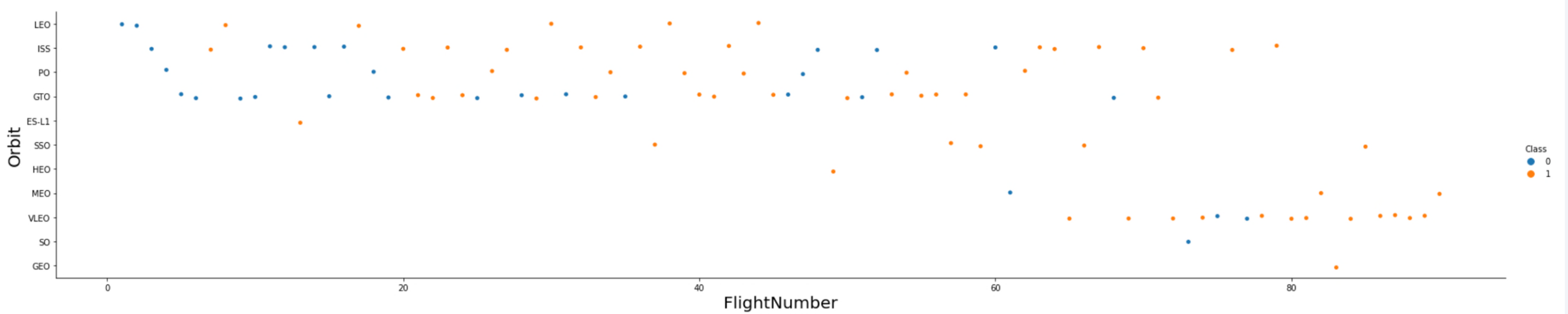
- The graph shows that the success rate of the landing is the highest at payload above 14,000
- Launch site VAFB SLC 4E is not used for rocket with payload mass above 10,000

Success Rate vs. Orbit Type

- Orbit ES-L1, GEO, HEO, and SSO have 100% success rate, followed by VLEO with the second highest at 80% success rate
- Orbit GTO has the lowest success rate of 50%
- Other orbits (ISS, LEO, MEO, and PO) has a success rate ranged between 60 – 70%

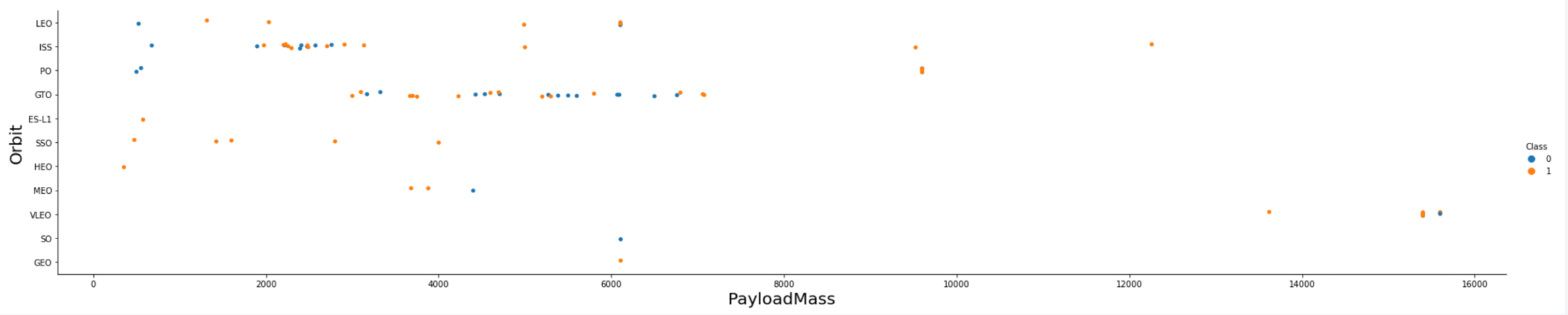


Flight Number vs. Orbit Type



- The graph shows that the success rate of LEO orbit is related with the number of flights
- In the beginning of launch (flight number less than 60) , LEO, ISS, PO, and GTO orbit are often selected
- After flight number higher than 60, the launch started to shift to VLEO orbit

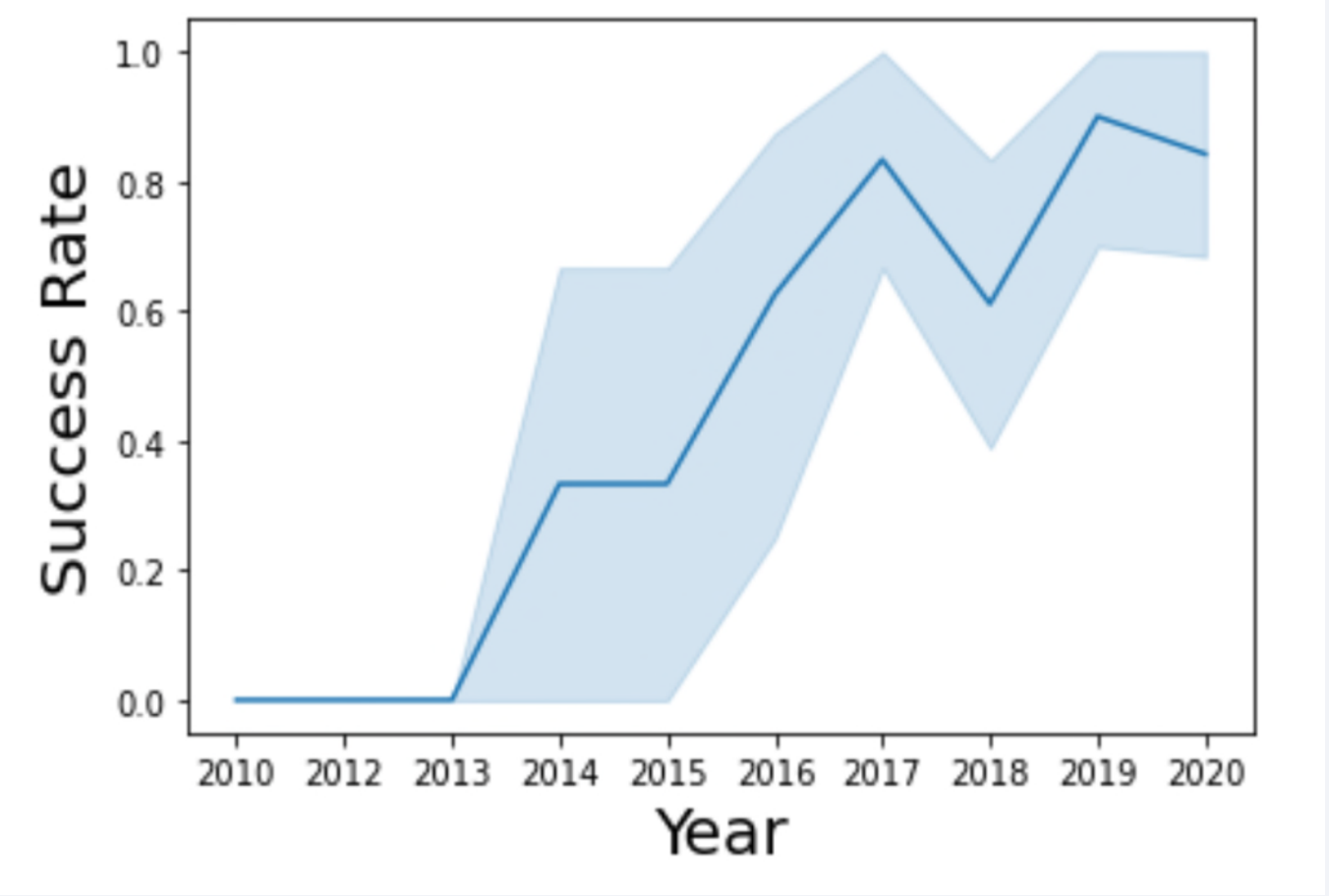
Payload vs. Orbit Type



- For orbit LEO, ISS, and POLAR, success rate increases at higher payload mass
- At payload higher than 13,000, only orbit VLEO is used
- Orbit SSO has 100% success rate, and only tested at payload less than or equal to 4,000

Launch Success Yearly Trend

- Success rate of the launch increases over time, starting from year 2013 to year 2020
- There is setback of success rate in 2018, where it decreases from around 80% to 60%. However, it recovers back to 90% in 2019



All Launch Site Names

- There are 4 launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

```
In [7]: %sql SELECT DISTINCT launch_site FROM DB2
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4
Done.
```

Out[7]:

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

Launch Site Names Begin with 'CCA'

```
In [8]: %sql SELECT * FROM DB2 WHERE (launch_site LIKE 'CCA%') LIMIT 5;
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[8]:

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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	0:35:00	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

- The earliest record where launch sites names begin with ‘CCA’ is in 2010
- All earliest record of site names begin with ‘CCA’ are either a failure or no attempt

Total Payload Mass

```
In [9]: %sql SELECT SUM(payload_mass__kg_) FROM DB2 WHERE (customer = 'NASA (CRS)')
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgt
Done.
```

Out[9]:

1
45596

- The total payload mass carried by boosters from NASA is 45,596 kg

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [10]:
```
%sql SELECT AVG(payload_mass__kg_) FROM DB2 WHERE (booster_version LIKE 'F9 v1.1%')
```

```
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00
Done.
```

Out[10]:

1
2534

- The average payload mass carried by boosters from NASA is 2,534 kg

First Successful Ground Landing Date

```
In [12]: %sql SELECT MIN(DATE) FROM DB2 WHERE landing_outcome='Success (ground pad)'
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu
Done.
```

Out[12]:

1
2015-12-22

- The first successful landing on ground pad is in 2015, 5 years since the first launch testing

Successful Drone Ship Landing with Payload between 4000 and 6000

In [16]:
```
%sql SELECT booster_version FROM DB2 WHERE (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing_outcome = 'Success (drone ship)')
```

```
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[16]:

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

- There are 4 types of boosters with 4000 ≤ payload mass ≤ 6000 had been used during successful drone ship landing:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

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

```
In [22]: %sql SELECT launch_site,booster_version FROM DB2 WHERE landing_outcome = 'Failure (drone ship)'
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.a
Done.
```

Out[22]:

launch_site	booster_version
CCAFS LC-40	F9 v1.1 B1012
CCAFS LC-40	F9 v1.1 B1015
VAFB SLC-4E	F9 v1.1 B1017
CCAFS LC-40	F9 FT B1020
CCAFS LC-40	F9 FT B1024

- The total number of successful mission is close to 100%, regardless whether the landing of stage 1 is successful or not
- There is only 1 flight resulting in a failure

Boosters Carried Maximum Payload

```
In [20]: %sql SELECT booster_version FROM DB2 WHERE payload_mass__kg_=(SELECT MAX(payload_mass__kg_) FROM DB2)
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdom
Done.
```

Out[20]:

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

- There are total of 12 booster versions carries the highest payload, at 15,600 kg

2015 Launch Records

```
In [22]: %sql SELECT launch_site,booster_version FROM DB2 WHERE landing_outcome = 'Failure (drone ship)'
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.a
Done.
```

Out[22]:

launch_site	booster_version
CCAFS LC-40	F9 v1.1 B1012
CCAFS LC-40	F9 v1.1 B1015
VAFB SLC-4E	F9 v1.1 B1017
CCAFS LC-40	F9 FT B1020
CCAFS LC-40	F9 FT B1024

- There are 2 launch sites where landing outcomes at drone ship is a failure: CCAFS LC-40 and VAFB SLC-4E
- There are 5 types of booster versions had been used during failure landing at drone ship

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

In [37]:
```
%sql SELECT landing_outcome, COUNT(landing_outcome) FROM DB2 WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY COUNT(landing_outcome) DESC
```

```
 * ibm_db_sa://gyd94708:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

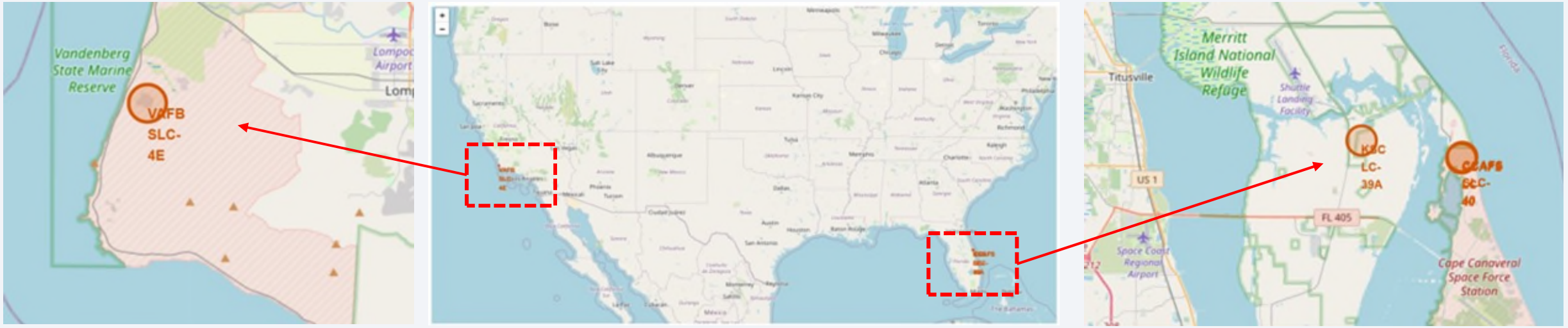
Out[37]:

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

- The top three highest rank of landing outcomes between 2010 and 2017 are ‘No Attempt’, followed by ‘Failure at drone ship’ and Success at drone ship’

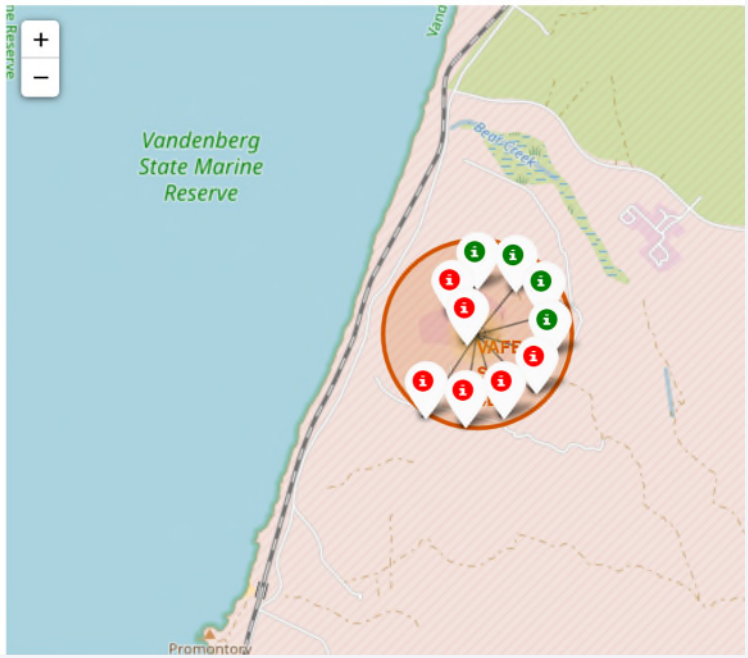

Section 4
Launch Sites
Proximities Analysis

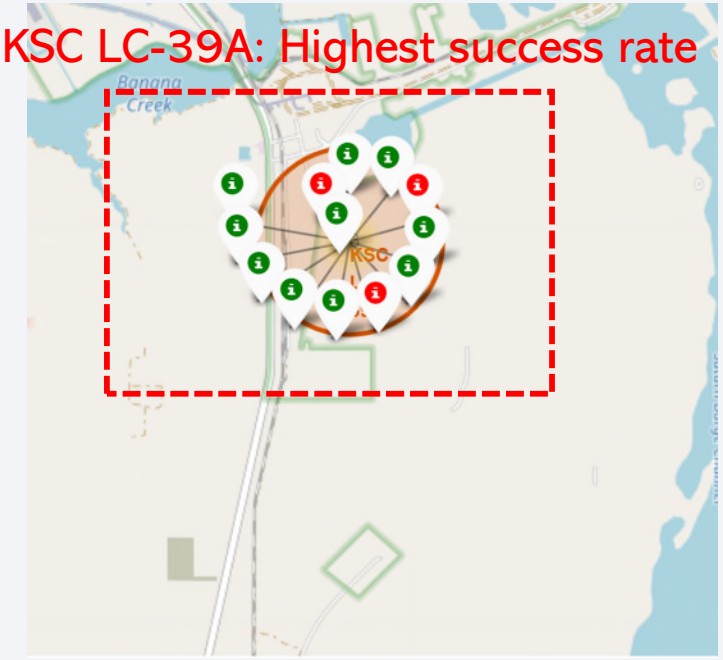
Locations of All Launch Sites on a Map

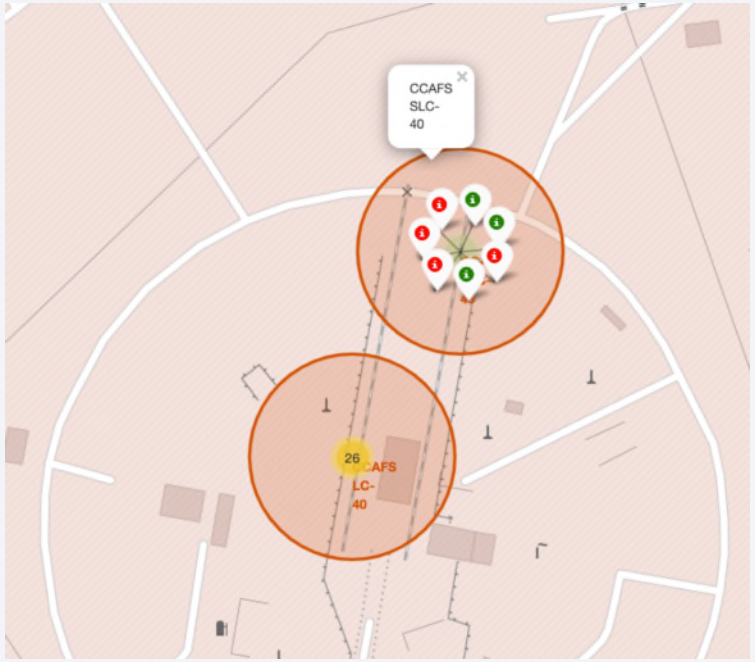


- There are total of 4 launch locations in the US: One is in the west of USA, and three are in the east of USA
- VAFB SLC-4E is located in the west USA, while CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A are located in the east of USA

Success/Failed Launch Markers for Every Sites







- Between all launch sites, KSC LC-39A has the highest success rate of launching sites, with only 3 failures out of 13 launches (see middle figure)

Distances between a launch site to its proximities

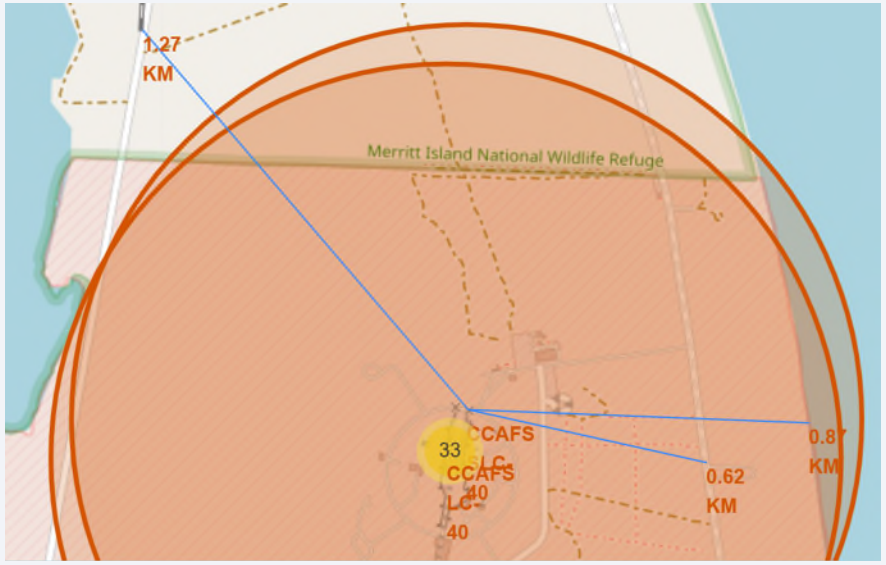


Fig.1 Distance between launching site CAFS LC-40 and coast, railway, highway

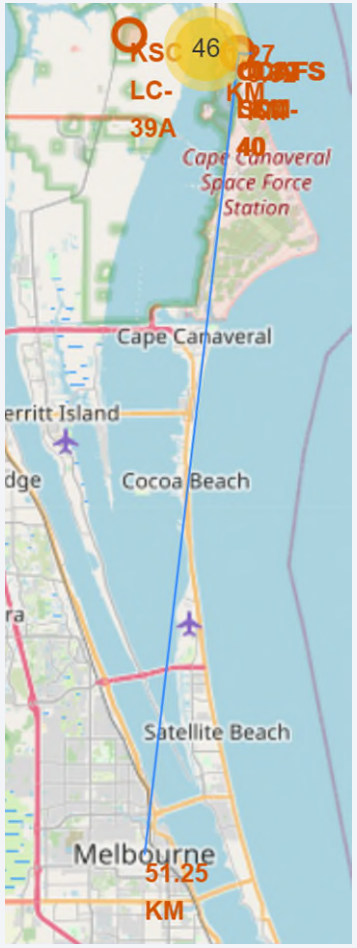
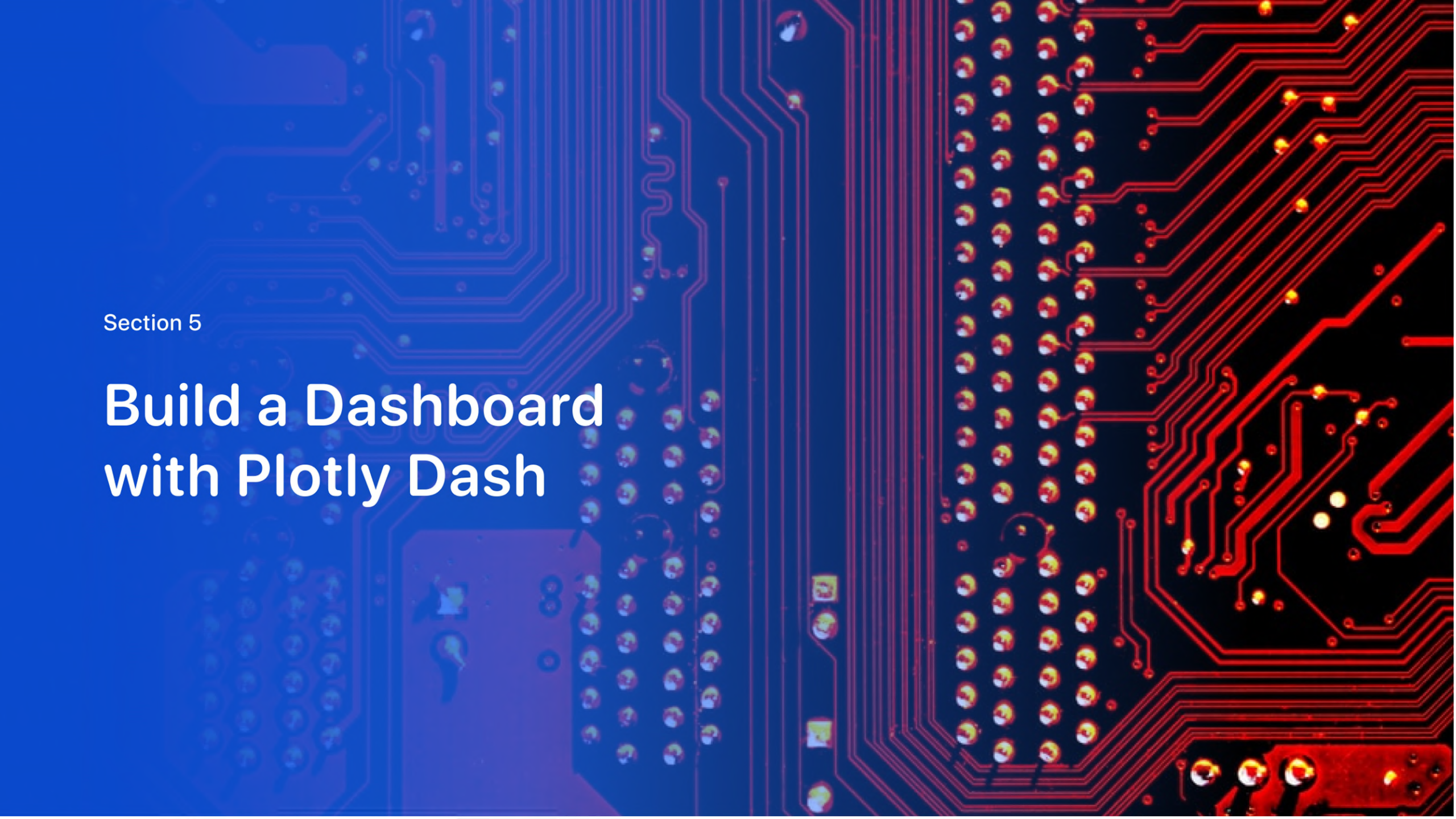


Fig.2 Distance between launching site CAFS LC-40 to Melbourne city

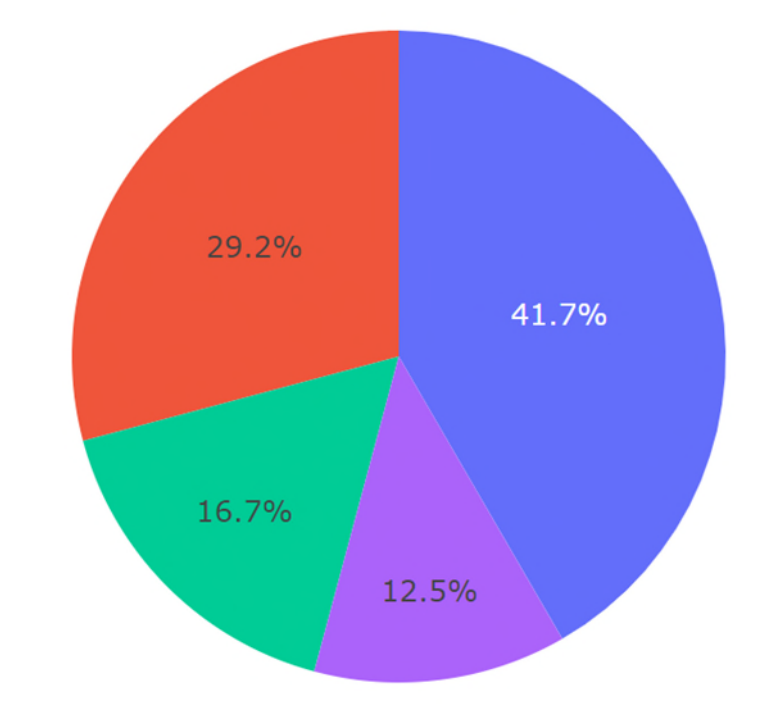
- The distance between launching site CAFS LC-40 to nearest coast is 0.87 km
- The distance between launching site CAFS LC-40 to nearest railway is 1.27 km
- The distance between launching site CAFS LC-40 to nearest highway is 0.62 km
- The distance between launching site CAFS LC-40 to Melbourne city is 51.25 km



Section 5

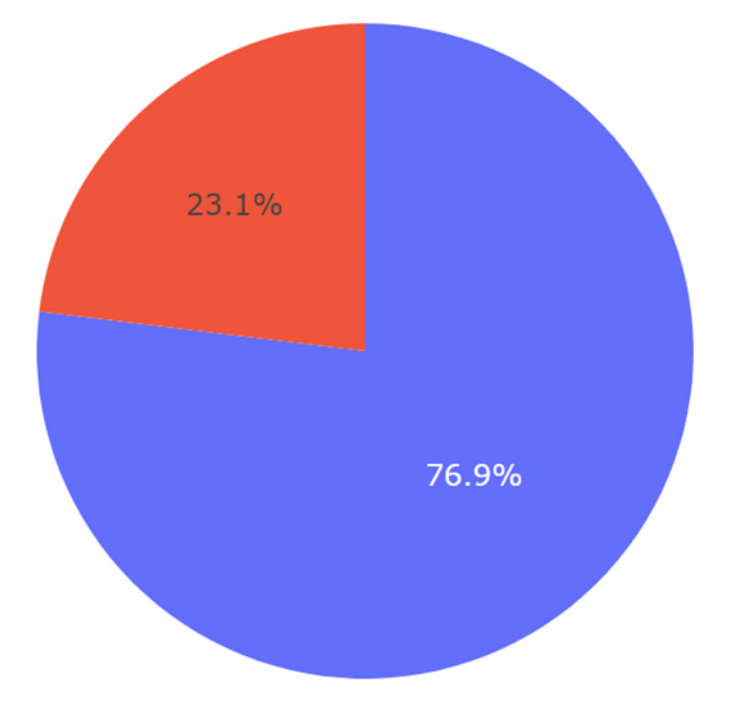
Build a Dashboard with Plotly Dash

Total Success Rate of Launches of All Sites



- KSC LC-39A has the highest success rate between all launching sites, at 41.7%

Total Success Rate of Launches of Site KSC LC-39A



- KSC LC-39A has the highest success rate of launching sites, with only 3 failures out of 13 launches. The total success rate of this site is 76.9%

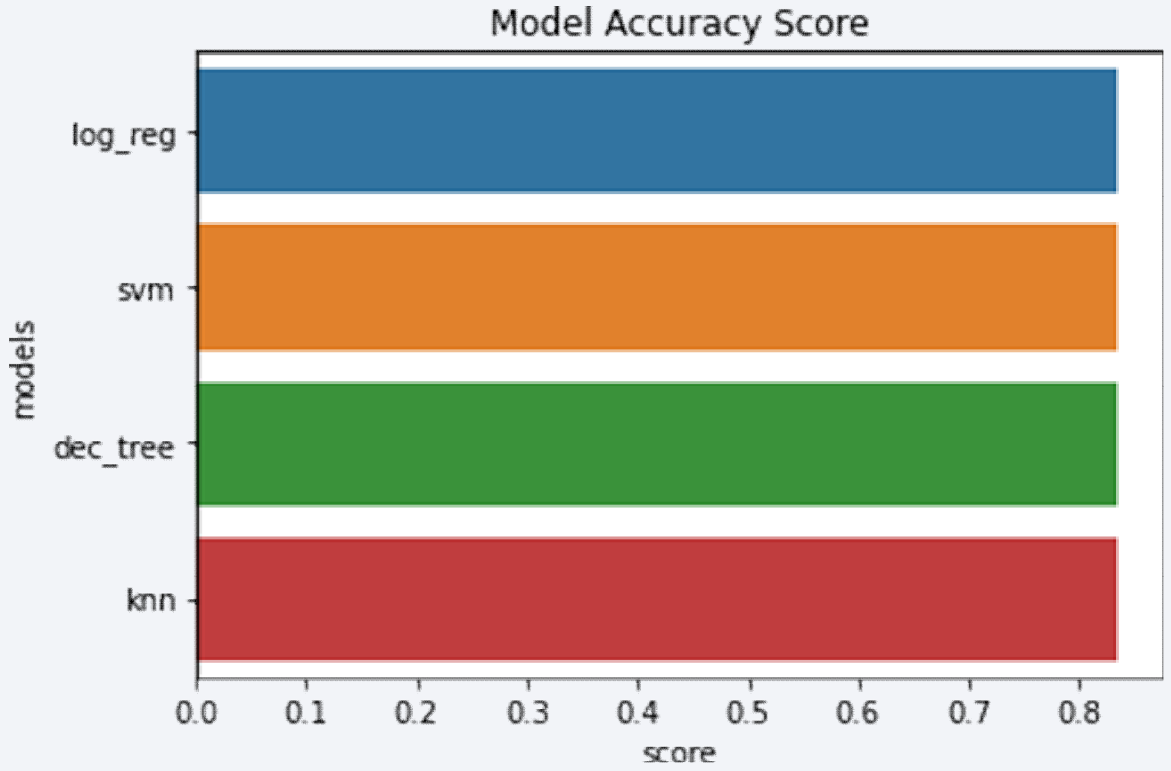
Correlation Between Payload & Success Rate of KSC LC-39A

- Booster FT can used at wide range of payload, which can be as low as 2,500 kg and as high as 5,500 kg
- At payload below 5,500 kg, booster B4 and B5 has a 100% success rate, although the number of samples are limited
- Booster B5 is only used once at payload below 5,000 kg

Section 6

Predictive Analysis (Classification)

Classification Accuracy



- All 4 models (lor regression, SVM, decision tree, and KNN) showed similar accuracy score, at 83.3%
- The number of data is rather small, at 90 samples. At test size of 20%, the test data used is only 18 samples
- More data is recommended to determine the best classification model for this problem

Confusion Matrix

- Because all models showing same accuracy, the confusion of matrix of all models showed the same results
- There are 12 True Positive and 3 True Negative results
- There are 3 False Positive and 0 False Negative results from the model, which contribute to the error in accuracy

Conclusions

- We managed to collect the SpaceX API and Wikipedia using Requests and Beautiful Soup libraries
- The data is wrangled, cleaned, and stored in Db2 cloud database for easier data management
- From EDA, we managed to identify that there is relation between Launch Site vs Flight Number, between Landing Success Rate vs Payload Mass, and between Landing Success Rate vs Orbit
- By creating Pie chart, we managed to identify that KSC LC-39A has the highest success rate between all of launching sites, with the success rate of 76.9%
- We developed machine learning models to predict whether a launch will have a successful Stage , with the accuracy of 83.3%
- SpaceY will be able to predict the success rate of landing site with a good accuracy, to compete with SpaceX for the next bidding

Appendix

My Github Repository URL:

https://github.com/setzerace/IBM-Data-Science-Capstone-Project-Huang

Thank you!