Winning Space Race with Data Science

Robin Wu

Nov/12, 2021

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- There are several companies provide space travel such as Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX.
  - The most successful commercial spaceflight is SpaceX.
- 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.
  - If we can determine if the first stage will land, we can determine the cost of a launch.
- Our goal is to use rocket launch data to predict whether SpaceX will attempt to land a rocket successfully or not.
  - This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Based on different circumstances, Could we predict the rocket will attempt to land successfully or not?

Section 1

Methodology

Methodology

Executive Summary (Understanding the scale of the Falcon 9)

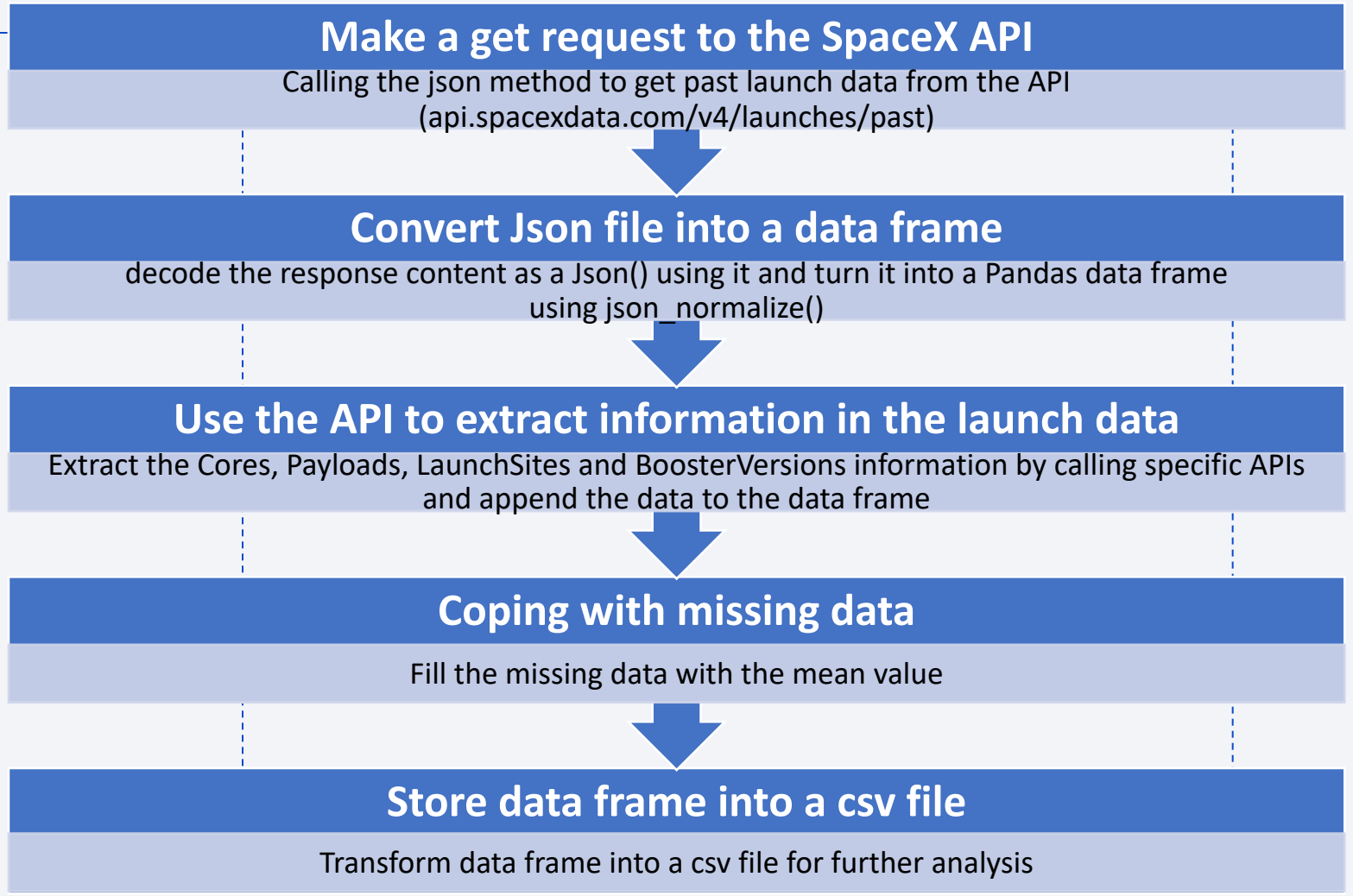
- Data collection methodology:
  - Collecting data from SpaceX REST API and transforming json data into into a flat table form
  - Collecting data from Falcon9 Launch Wikipedia page
- Perform data wrangling
  - Convert all the data in the table into numerical data such like converting landing outcome into 0 and 1
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using Grid Search to train the best hyperparameter of the model and then choose the best model by comparing the prediction of various classification models

Data Collection

- Collecting Falcon9 first stage launches data from SpaceX REST API for further prediction
  - Transforming the collected json data into into a flat table form and parse the data into a reasonable format.
- Collecting Falcon9 Heavy launches data from Wikipedia page
  - Extracting the collected html data into a flat table form and parse the data into a reasonable format.

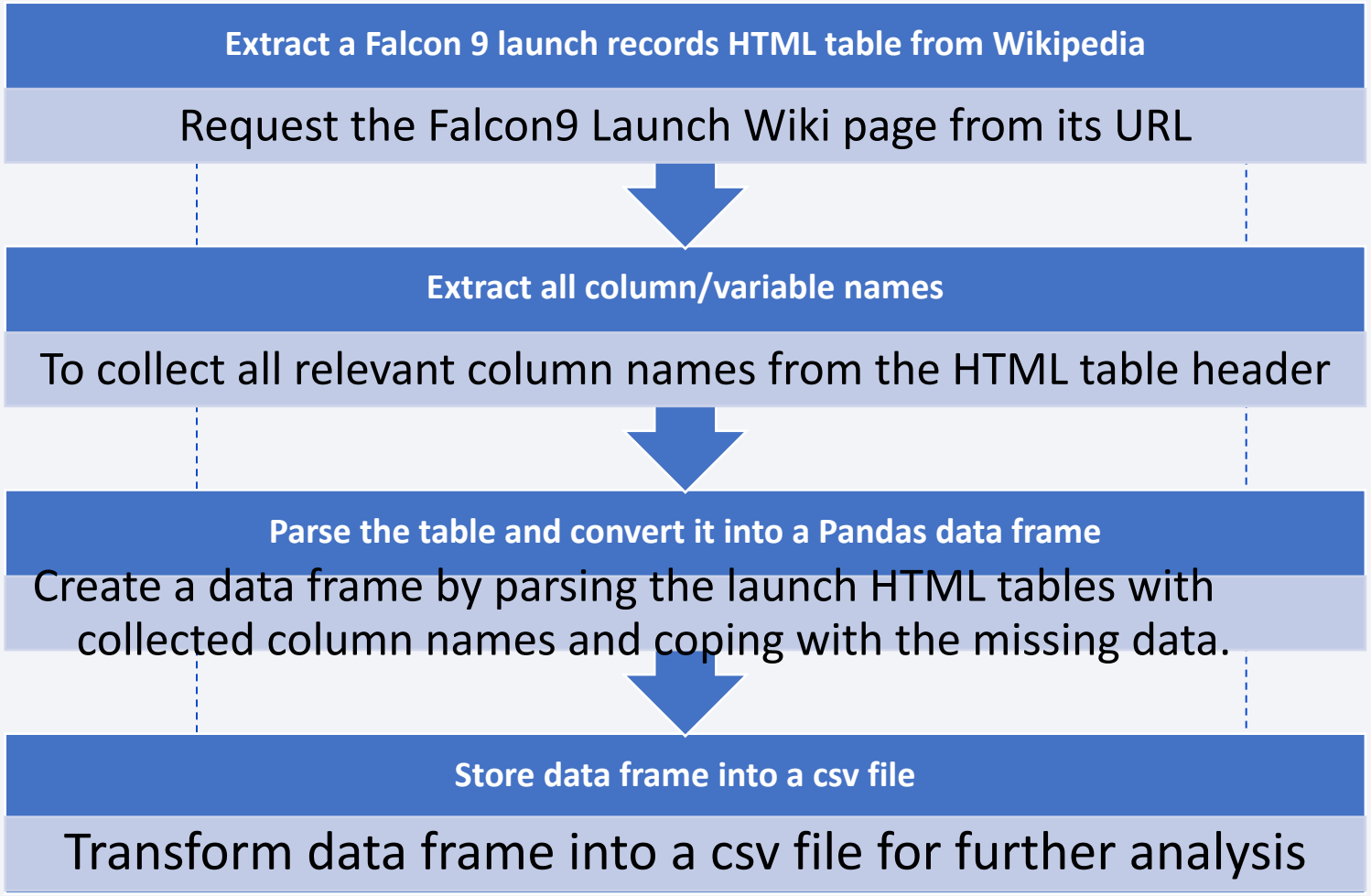
Data Collection – SpaceX API

- Extract a Falcon 9 launch records HTML table from Wikipedia
  - Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.
  - Extract Core, Payload, LaunchSite and BoosterVersion data into dataframe
  - Coping with missing data
- GitHub URL of the completed SpaceX API calls with jupyter notebook (https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Data%20Collection%20with%20Web%20Scraping.ipynb)



Data Collection - Scraping

- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
- Parse the table and convert it into a Pandas data frame
  - Extract all column/variable names from the HTML table header
  - Create a data frame by parsing the launch HTML tables
- GitHub URL of the completed Data Scraping with Jupyter notebook (https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Data%20Collection%20with%20Web%20Scraping.ipynb)



Data Wrangling

- In the Data Wrangling, I perform Exploratory Data Analysis (EDA) to find some patterns in the data from SpaceX API determine what would be the label for training supervised models.
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type
  - Create a landing outcome label from Outcome column

- GitHub URL of the completed Data Wrangling with Jupyter Notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/EDA%20lab.ipynb)

EDA with Data Visualization

- Visualize the relationship between two features in Falcon9 launches data from the plot and see how these features related to launching successfully or not
  - Does a higher Flight Number on the specific Launch Site will launch successfully?
  - Does the heavy Payload mass affect the launch success on the specific Launch Site?
  - To observe which orbit type has the higher success rate.
  - Does a higher Flight Number on the specific Orbit type will launch successfully?
  - Does a heavy Payload mass affect the launch success on the specific Orbit?
  - Does the launch success increase yearly?
- Create dummy variables to categorical columns
- Cast all numeric columns to float64 for further prediction
- GitHub URL of the completed Data Visualization with Jupyter Notebook (https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/EDA%20with%20Visualization%20lab.ipynb)

EDA with SQL

- Load the Falcon9 heavy launches dataset into the corresponding table in an IBM Db2 database to observe
  - The names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
- GitHub URL of the completed Data in SQL with Jupyter Notebook (https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/EDA%20with%20SQL%20lab.ipynb)

Build an Interactive Map with Folium

- Launch Sites Locations Analysis with Folium by using Falcon9 heavy launches data
- Mark all launch sites on a map
  - Does most of the launch sites located in similar places?
- Mark the success/failed launches for each site on the map
  - Does specific location relate to the launch success.
- Calculate the distances between a launch site to its proximities
  - Does each of the launch site very close proximity to the coast?
- GitHub URL of the completed interactive map with Folium map by Jupyter Notebook (https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb)

Build a Dashboard with Plotly Dash

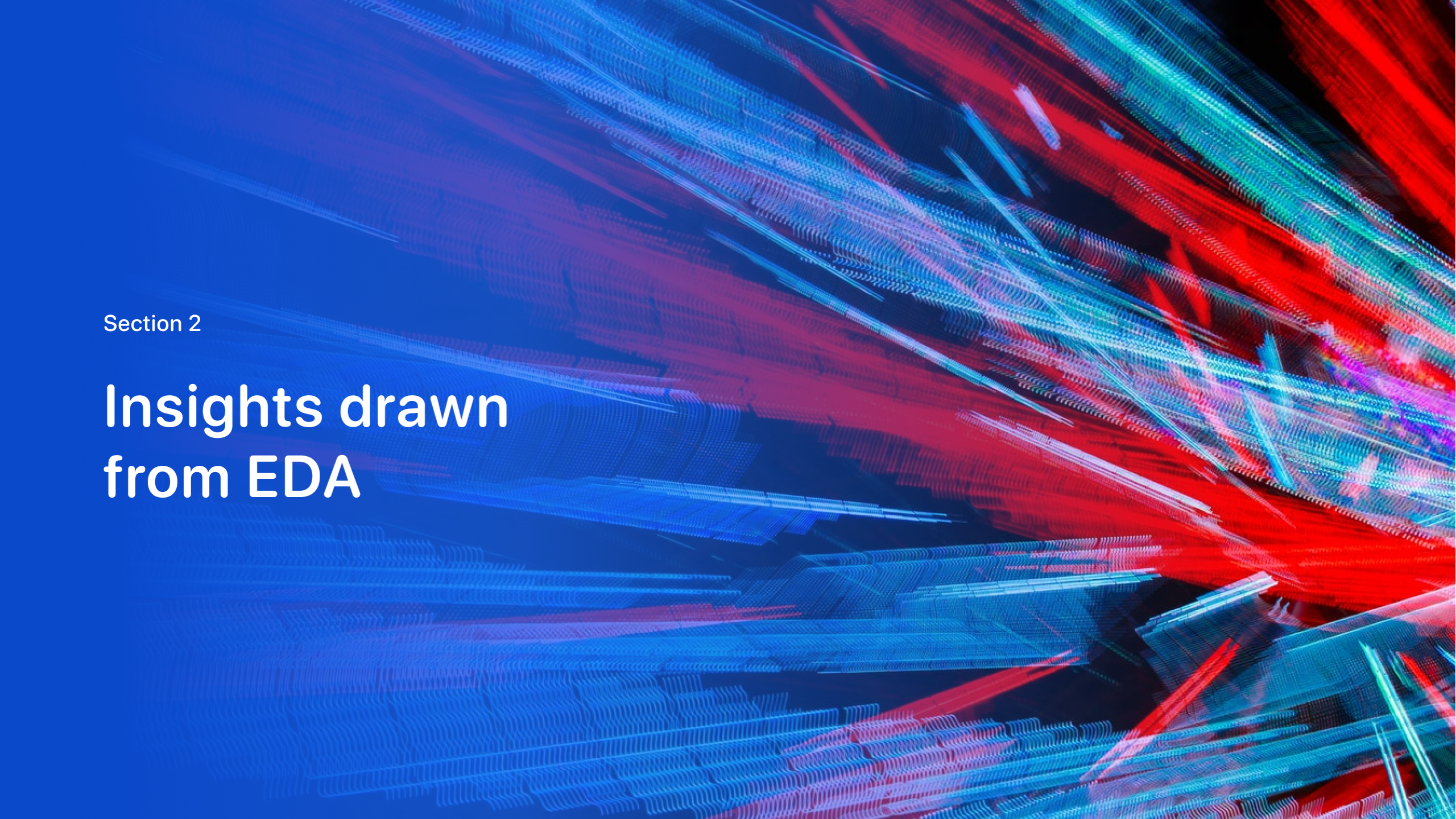
- Add a Launch Site Drop-down Input Component
  - To enable Launch Site selection
- Add a callback function to render success-pie-chart based on selected site dropdown
  - A pie chart to show the total successful launches count for all sites
- Add a Range Slider to Select Payload
  - The plot will simply change based on the selected range of the payload on the slider
- Add a callback function to render the success-payload-scatter-chart scatter plot
  - A scatter chart to show the correlation between payload and launch success
- GitHub URL of the completed Plotly Dash lab by Jupyter Notebook
  (https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.ipynb)

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels
  - Create a column for the class (launch success or not)
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data
- GitHub URL of the completed of the completed predictive analysis lab by Jupyter Notebook
  (https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Machine%20Learning%20Prediction.ipynb)

Results

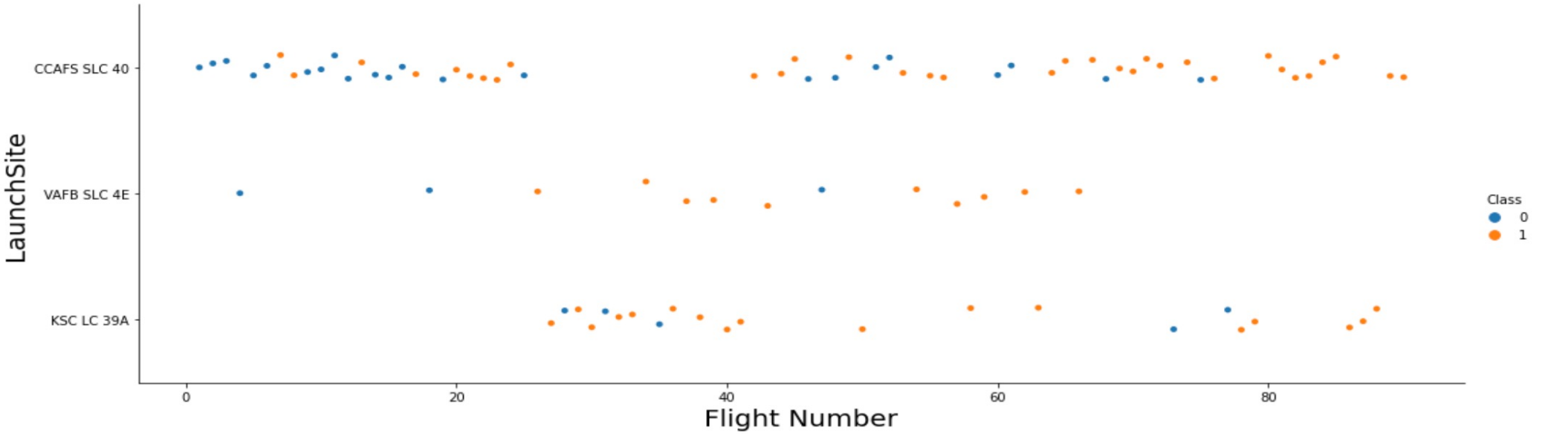
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

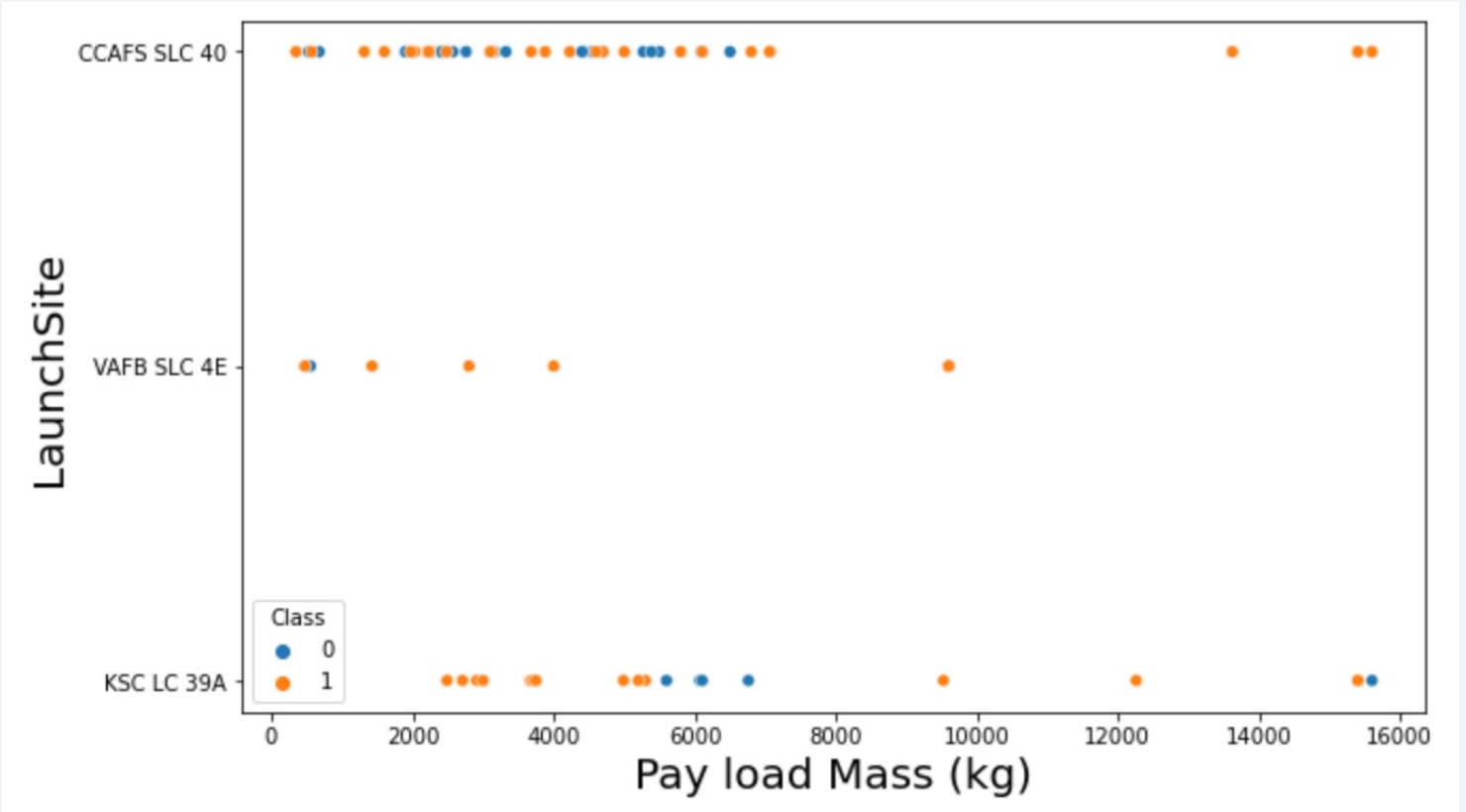
Flight Number vs. Launch Site



- The plot shows that
  - The launch site VAFB SLC 4E has 100% launch success when the flight number is larger than 50.
  - The launch site CCAFS SLC 40 has 100% launch success when the flight number is larger than 75.
  - The launch site KSC LC 39A has 100% launch success when the flight number is larger than 78.

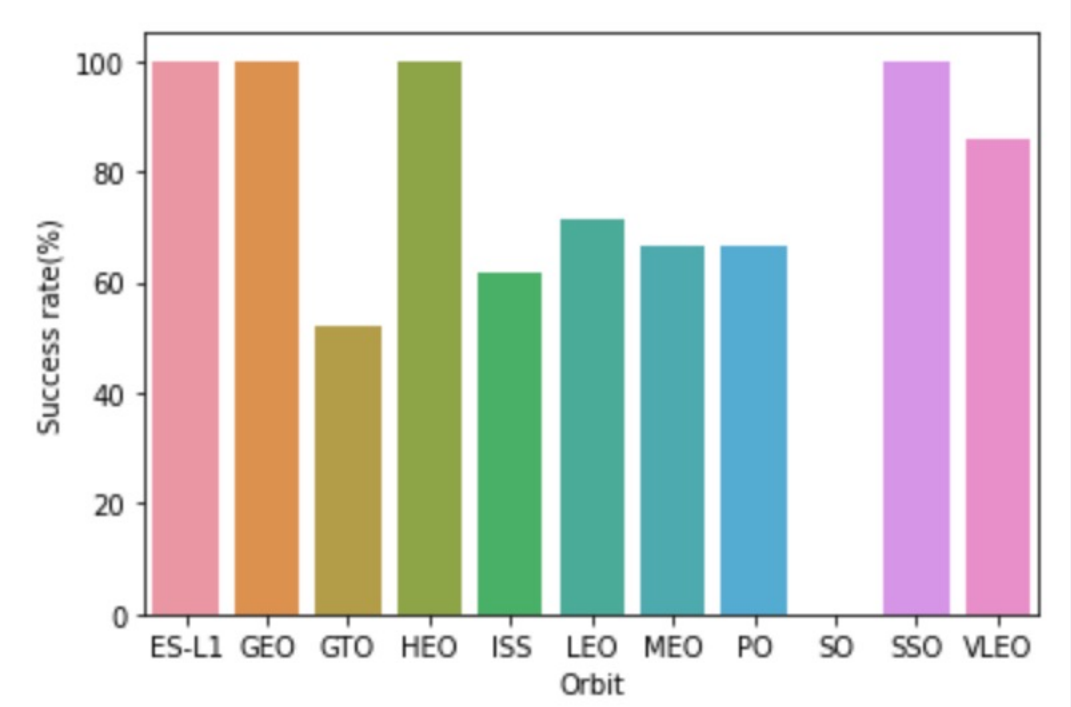
Payload vs. Launch Site

- The plot shows the launch site CCAFS SLC 40 has a 100% launch success rate when the payload mass is heavier (heavy than 14000kg)



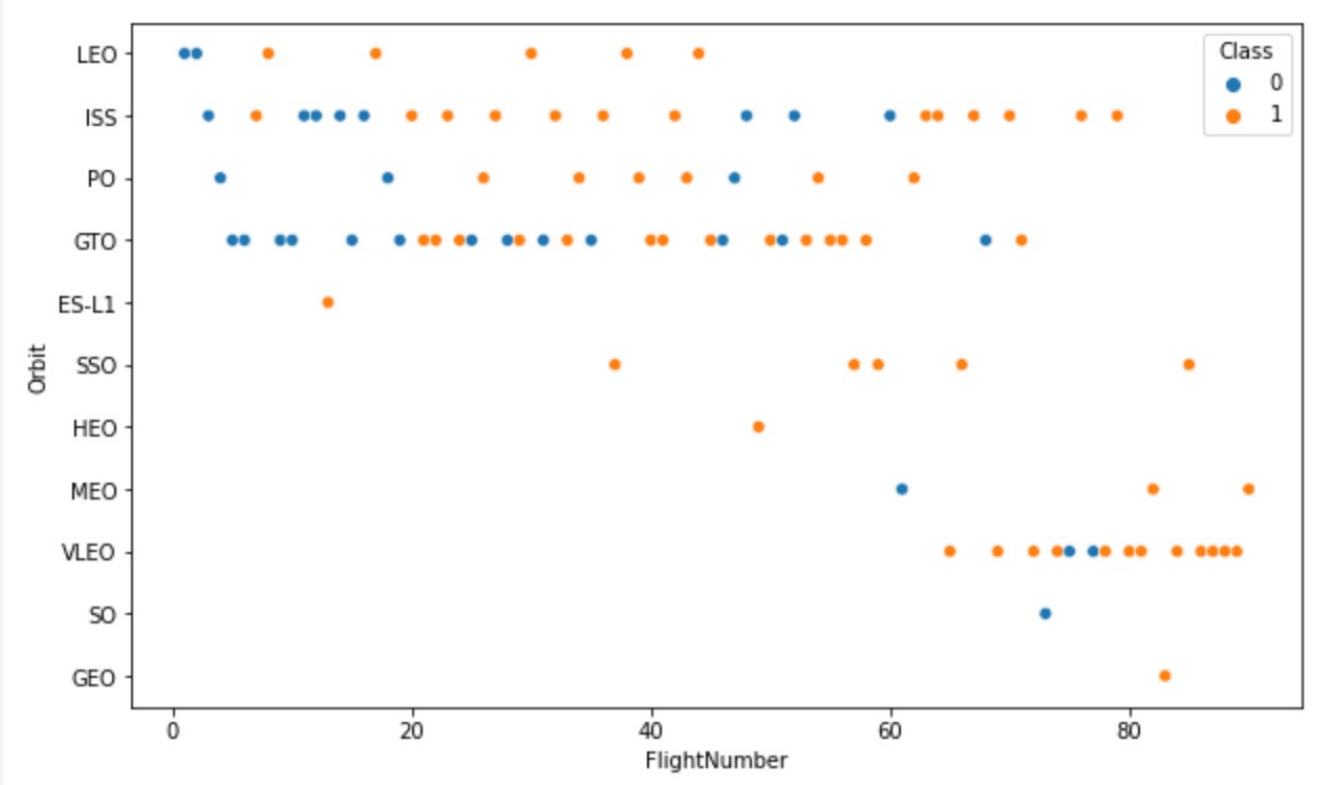
Success Rate vs. Orbit Type

- The bar chart shows the orbit type of ES-L1, GEO, HEO, SSO perform the highest success rate



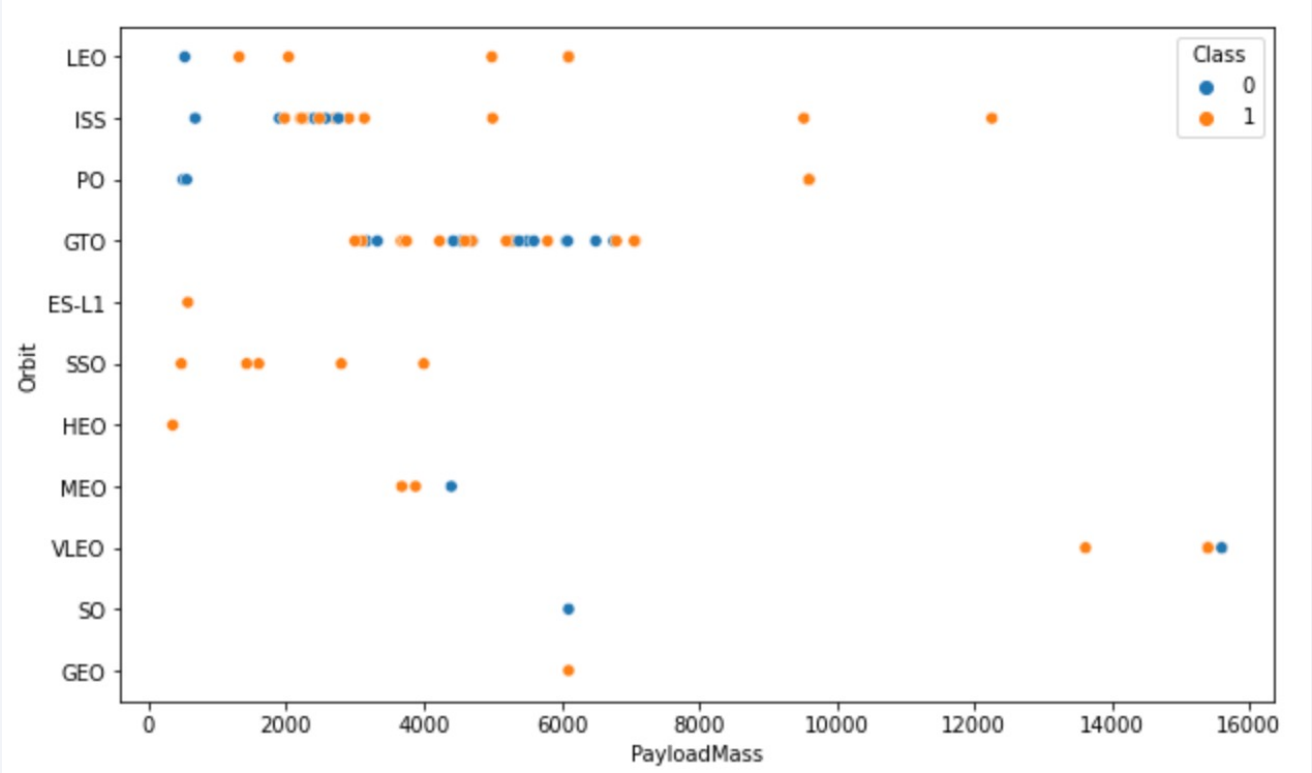
Flight Number vs. Orbit Type

- The plot shows that in the LEO orbit the Success appears related to the number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

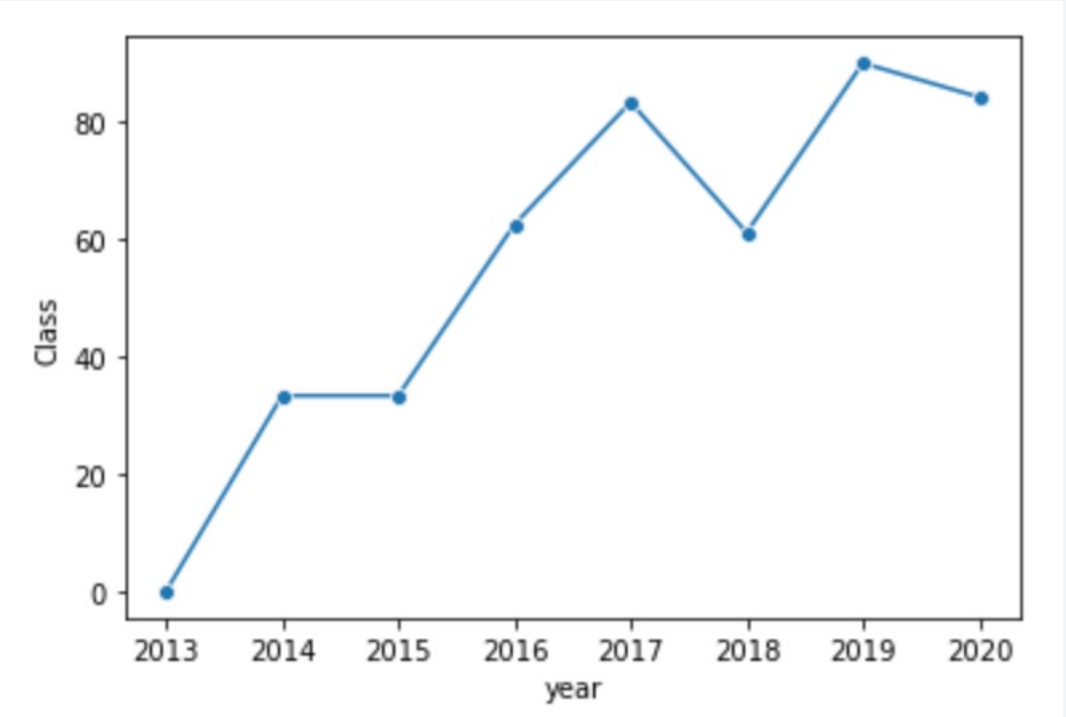
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here



Launch Success Yearly Trend

- The Launch success keep increasing yearly from 2013, but it decreases in 2018 and 2020.

	year	Class
0	2010	0.000000
1	2012	0.000000
2	2013	0.000000
3	2014	0.333333
4	2015	0.333333
5	2016	0.625000
6	2017	0.833333
7	2018	0.611111
8	2019	0.900000
9	2020	0.842105



All Launch Site Names

- Find the names of the unique launch sites
- By the result of SQL query , there are four sites in the past launch data.

In [4]:
```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

```
 * ibm_db_sa://nrt78679:***@5
atabases.appdomain.cloud:3273
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` from the SpaceX past launch data
- Which shows the launch site CCAFS start loading payload from 2012.

In [7]:
```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

```
 * ibm_db_sa://nrt78679:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1
atabases.appdomain.cloud:32733/BLUDB
Done.
```

Out[7]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677

Total Payload Mass

- The total payload carried by boosters from NASA is 45595

In [17]:
```
%%sql
SELECT customer, SUM(payload_mass__kg_) as total_payload
FROM SPACEXTBL
GROUP BY customer
HAVING customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://nrt78679:***@54a2f15b-5c0f-46df-8954-7e38e61
atabases.appdomain.cloud:32733/BLUDB
Done.
```

Out[17]:

customer	total_payload
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is extremely different and is hard to find some trends.

```
In [16]: %%sql
         SELECT booster_version, AVG(payload_mass__kg_) as average_payload
         FROM SPACEXTBL
         GROUP BY booster_version
         HAVING booster_version LIKE 'F9 v1.1%'
```

booster_version	average_payload
F9 v1.1	2928
F9 v1.1 B1003	500
F9 v1.1 B1010	2216
F9 v1.1 B1011	4428
F9 v1.1 B1012	2395
F9 v1.1 B1013	570
F9 v1.1 B1014	4159
F9 v1.1 B1015	1898
F9 v1.1 B1016	4707
F9 v1.1 B1017	553
F9 v1.1 B1018	1952

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is on Dec 22, 2015
- There are 10 different landing outcomes in the records.

```
%%sql
SELECT MIN(DATE) as DATE
FROM SPACEXTBL
WHERE landing__outcome ='Success (ground pad)'
```

```
 * ibm_db_sa://nrt78679:***@54a2f15b-5c0f-46df-
atabases.appdomain.cloud:32733/BLUDB
Done.
```

DATE
2015-12-22

```
%%sql
SELECT DISTINCT landing__outcome
FROM SPACEXTBL
```

```
 * ibm_db_sa://nrt78679:***@54a2f
atabases.appdomain.cloud:32733/BI
Done.
```

landing__outcome
Controlled (ocean)
Failure
Failure (drone ship)
Failure (parachute)
No attempt
Precluded (drone ship)
Success
Success (drone ship)
Success (ground pad)
Uncontrolled (ocean)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT customer, booster_version, landing__outcome, payload_mass__kg_
FROM SPACEXTBL
WHERE landing__outcome ='Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 and 6
000
```

- There are two different names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

customer	booster_version	landing__outcome	payload_mass__kg_
SKY Perfect JSAT Group	F9 FT B1022	Success (drone ship)	4696
SKY Perfect JSAT Group	F9 FT B1026	Success (drone ship)	4600
SES	F9 FT B1021.2	Success (drone ship)	5300
SES EchoStar	F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes shows that there's only one failure outcome and almost 99% of the outcomes are successful.

```
%%sql
SELECT mission_outcome, COUNT(*) as total_number
FROM SPACEXTBL
GROUP BY mission_outcome
```

```
 * ibm_db_sa://nrt78679:***@54a2f15b-5c0f-46df-89
atabases.appdomain.cloud:32733/BLUDB
Done.
```

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

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, payload_mass__kg_
FROM SPACEXTBL
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
ORDER BY booster_version
```

- The names of the booster which have carried the maximum payload mass (15600kg) are with similar booster name.

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- There are only two failed landing outcomes in drone ship, their launch site names for in year 2015 are the same.

```
%%sql
SELECT DATE, booster_version, launch_site, landing__outcome
FROM SPACEXTBL
WHERE landing__outcome ='Failure (drone ship)' AND YEAR(DATE) = 2015
```

```
 * ibm_db_sa://nrt78679:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1og
atabases.appdomain.cloud:32733/BLUDB
Done.
```

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)

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20
  - Overall, Not attempt has the most counts of landing outcome.
  - The successful rate is nearly 25% during 2010 to 2017

```
%%sql
SELECT LANDING__OUTCOME, COUNT(*) as COUNT
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY COUNT DESC
```

```
 * ibm_db_sa://nrt78679:***@54a2f15b-5c0f-46df-89
atabases.appdomain.cloud:32733/BLUDB
Done.
```

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 4

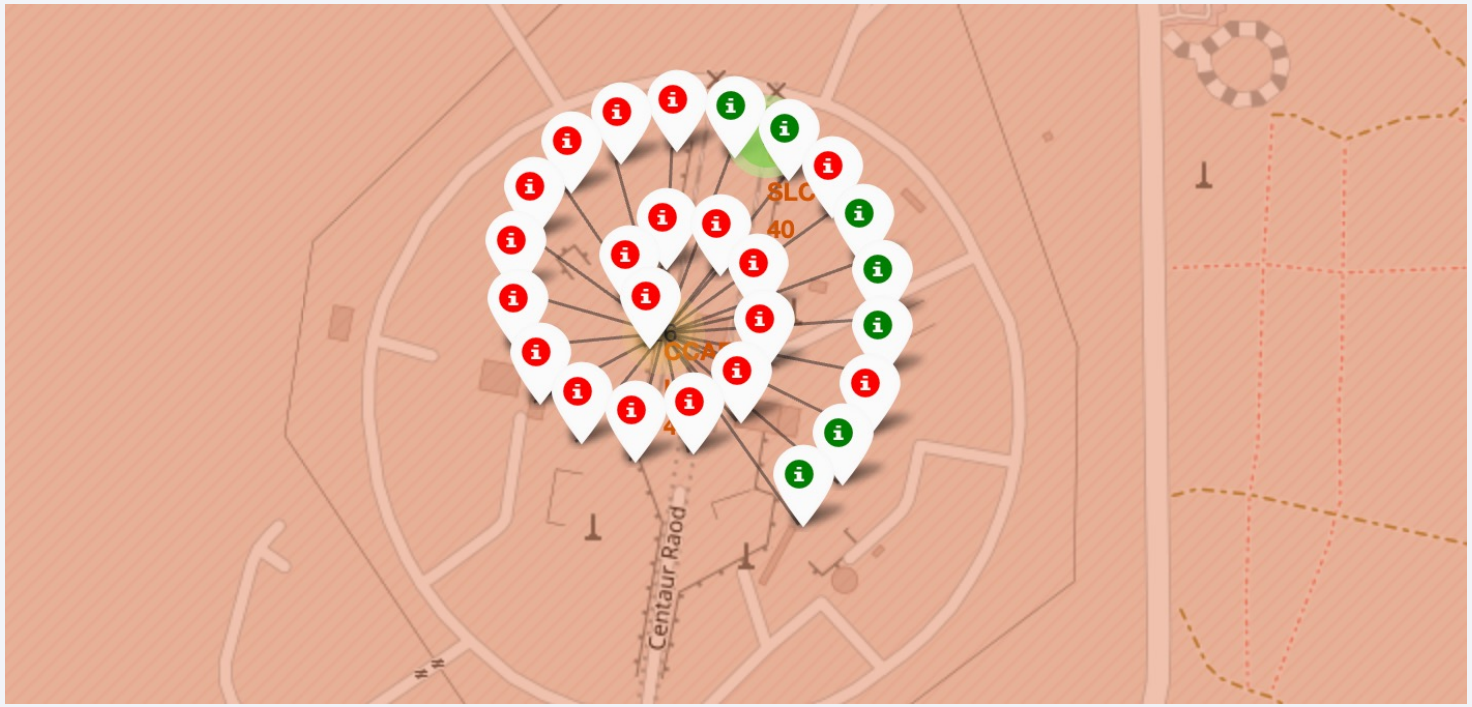
Launch Sites Proximities Analysis

Mark all launch sites on a map

- Find each launch site's location on a map using site's latitude and longitude coordinates.
- Create and
  add Circle and Marker fo
  r each launch site on the site map.

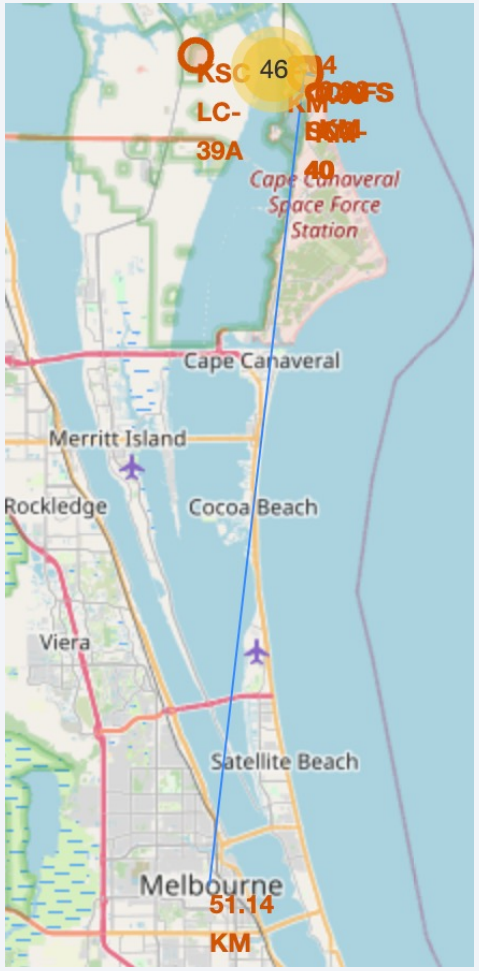
Mark the success/failed launches for each site on the map

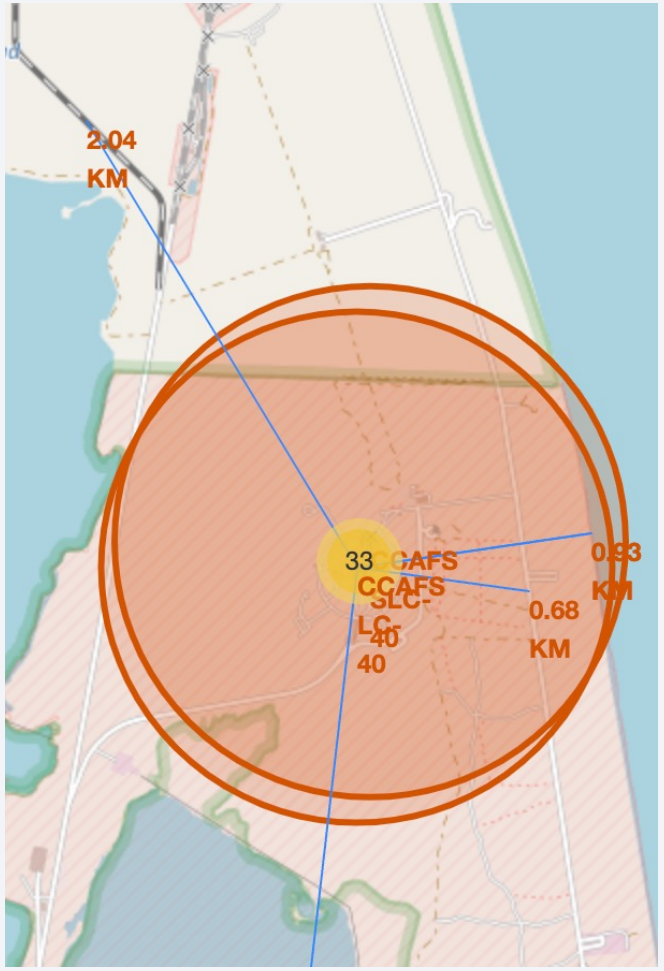
- Enhance the map by adding the launch outcomes for each site
- Create markers for all launch records. If a launch was successful (class=1), then use a green marker and if a launch was failed, use a red marker (class=0)

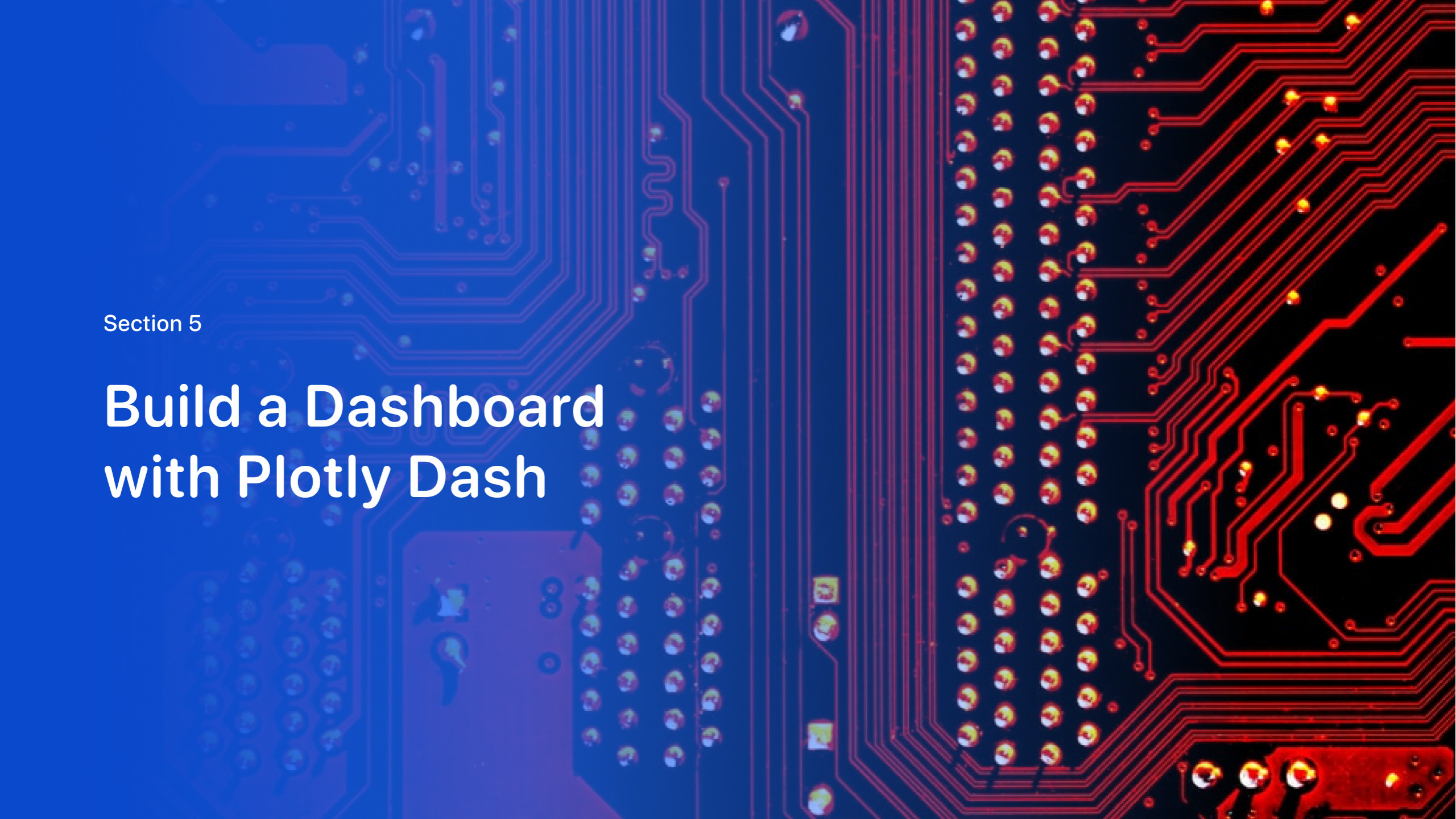


Calculate the distances between a launch site to its proximities

- To explore and analyze the proximities of launch sites.
- Add a MousePosition on the map to get coordinate of closest city, railway and highway on the map.
- Then, calculate and draw the distance between launch site and closest city, railway and highway.





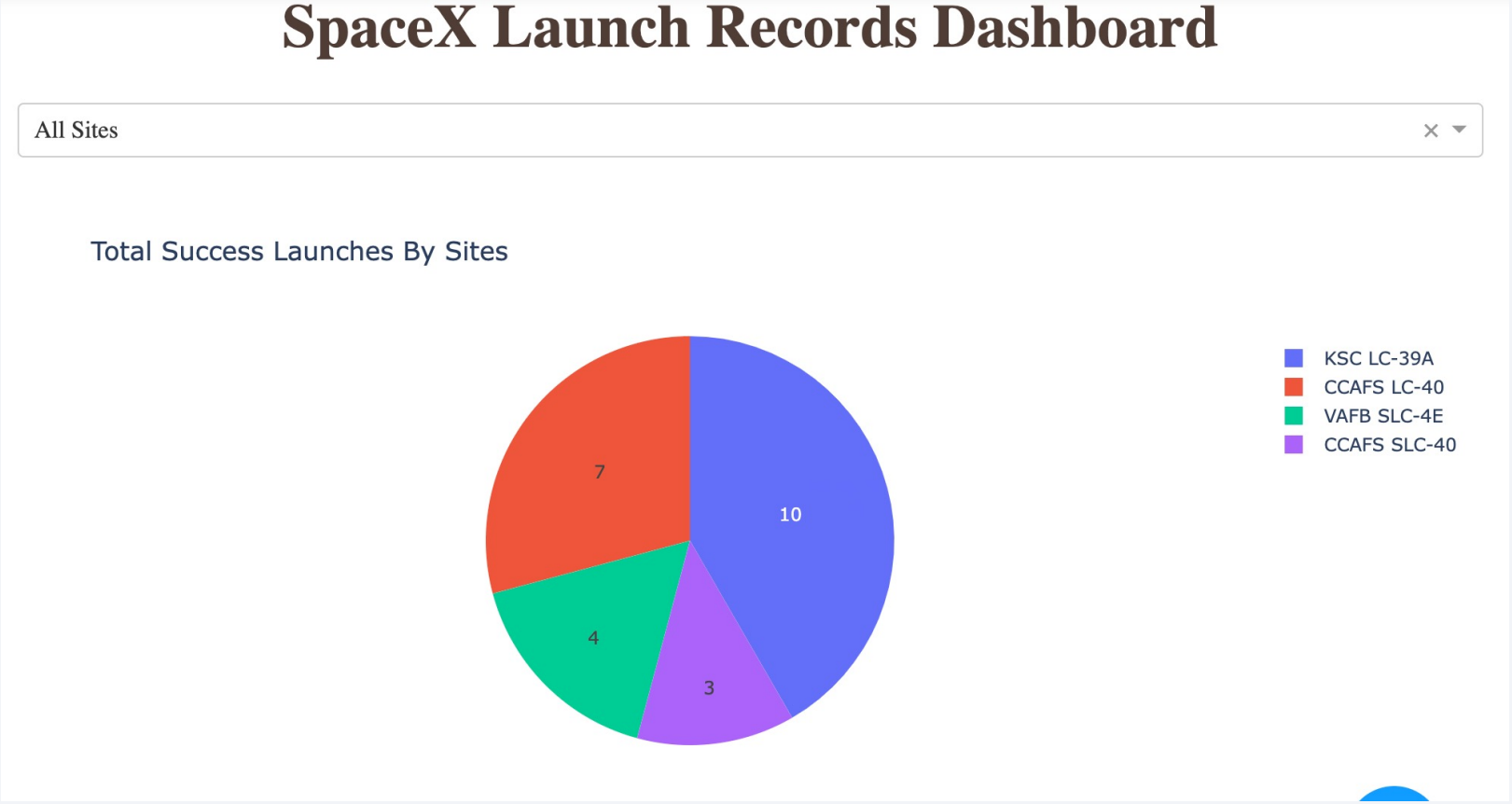


Section 5

Build a Dashboard with Plotly Dash

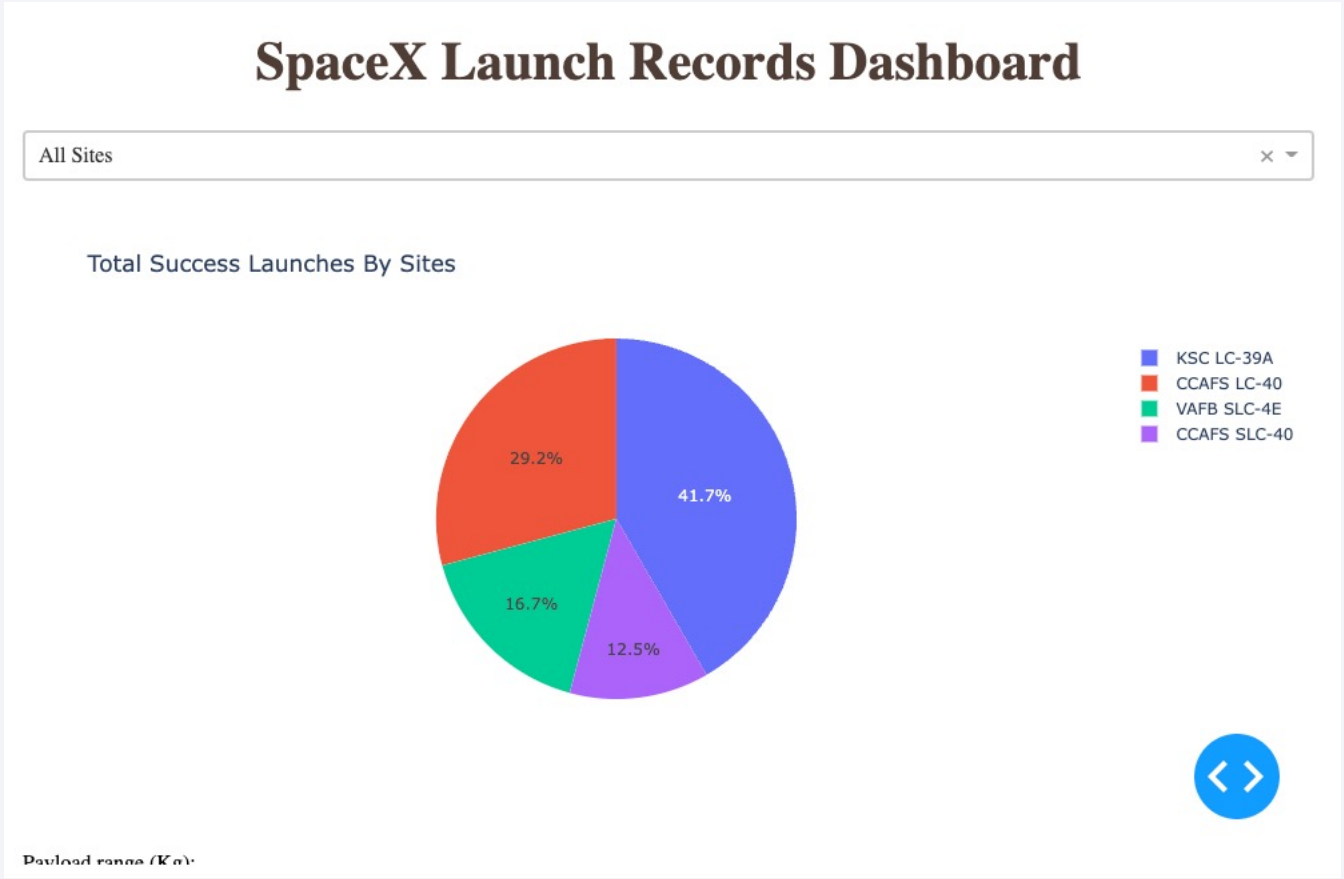
Pie chart for the launch success count on Dashboard

- Plot a pie chart for the launch site with launch success count on Dashboard
- The highest launch success count is 10 by the launch site KSC LC-39A



Total Success Launches By Sites

- Plot a success-pie-chart based on selected site dropdown
- The highest launch success ratio is 41.7% with the launch site KSC LC-39A
- Add a Launch Site Drop-down Input Component
- Add a callback function to render success-pie-chart based on selected site dropdown



Correlation between Payload and Success for Sites

- Plot a success-payload-scatter-chart based on selected slider range of payload input
- Add a Range Slider to Select Payload
- Add a callback function to render the success-payload-scatter-chart scatter plot

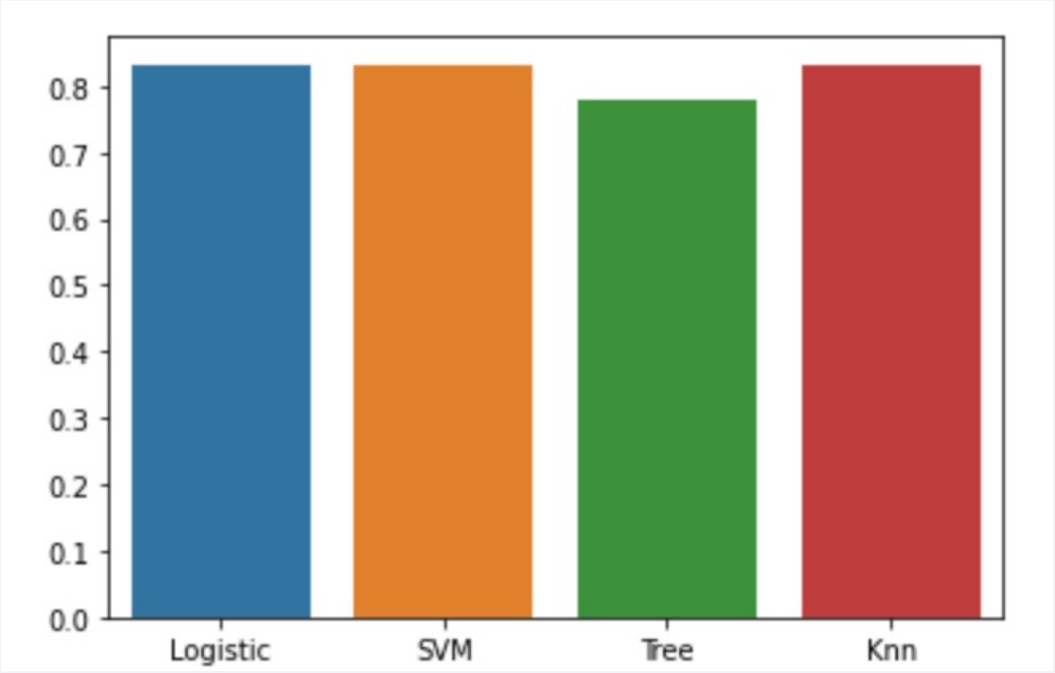
Section 6

Predictive Analysis (Classification)

Classification Accuracy

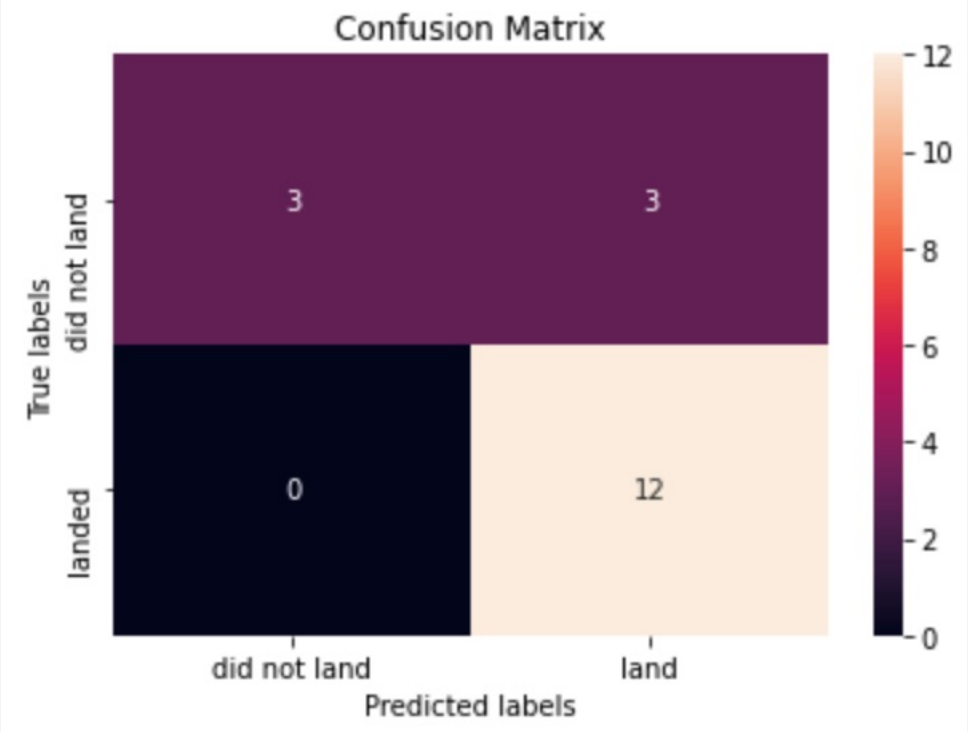
- Visualize the built model accuracy for all built classification models, in a bar chart

- Logistic, SVM and KNN are the models have the highest classification accuracy of 83%



Confusion Matrix

- The confusion matrix of Logistic, SVM and KNN, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- We analyzed the data for finding relationship between two features
  - The scatter plot shows the launch site CCAFS SLC 40 has a 100% launch success rate when the payload mass is heavier (heavy than 14000kg)
- We built a Dashboard which plot the pie chart by selecting dropdown as input and shows the scatter plot by sliding the input range slider
- Since our goal is to use rocket launch data to predict whether SpaceX will attempt to land a rocket successfully or not
  - I built a machine learning model which perform 83% accuracy to predict the launch success ratio.

Appendix

- GitHub URL of the completed SpaceX API calls with jupyter notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Data%20Collection%20with%20Web%20Scraping.ipynb)
- GitHub URL of the completed Data Scraping with jupyter notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Data%20Collection%20with%20Web%20Scraping.ipynb)
- GitHub URL of the completed Data Wrangling with Jupyter Notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/EDA%20lab.ipynb)
- GitHub URL of the completed Data Visualization with Jupyter Notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/EDA%20with%20Visualization%20lab.ipynb)
- GitHub URL of the completed Data in SQL with Jupyter Notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/EDA%20with%20SQL%20lab.ipynb)
- GitHub URL of the completed interactive map with Folium map by Jupyter Notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb)
- GitHub URL of the completed Plotly Dash lab by Jupyter Notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.ipynb)
- GitHub URL of the completed of the completed predictive analysis lab by Jupyter Notebook
(https://github.com/coolbosses/coursera_capstone_project/blob/f2f64f8a33c9a21953077564bfaacc960c3b92f7/Machine%20Learning%20Prediction.ipynb)

Thank you!