Winning Space Race with Data Science

Winstan Onyango Otieno
21 September 2022

Outline

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

Executive Summary

- Summary of methodologies

Data was collected using web scrapping HTML tables that contain valuable Falcon 9 launch records. Data wrangling was done, exploratory data analysis (EDA) using visualization and SQL. Interactive visual analytics performed using Folium and Plotly Dash and predictive analysis done using classification models.

- Summary of all results

Most flights launched at CCAFS SLC 40. Fewer rockets launched for payloadmass above 6000. When flights are few success rate is low. Most orbits have payload below 6,000. Launch Site KSC LC-39A has highest success rate. Decision Tree model had highest classification accuracy.

Introduction

- Project background and context

In this project, we predicted if the Falcon 9 first stage will land successfully. 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. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

- Problems you want to find answers

Space Y that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk. Project is to determine the price of each launch. Also to determine if the first stage will land successfully. This will be done by gathering information about Space X and creating dashboards.

Section 1

Methodology

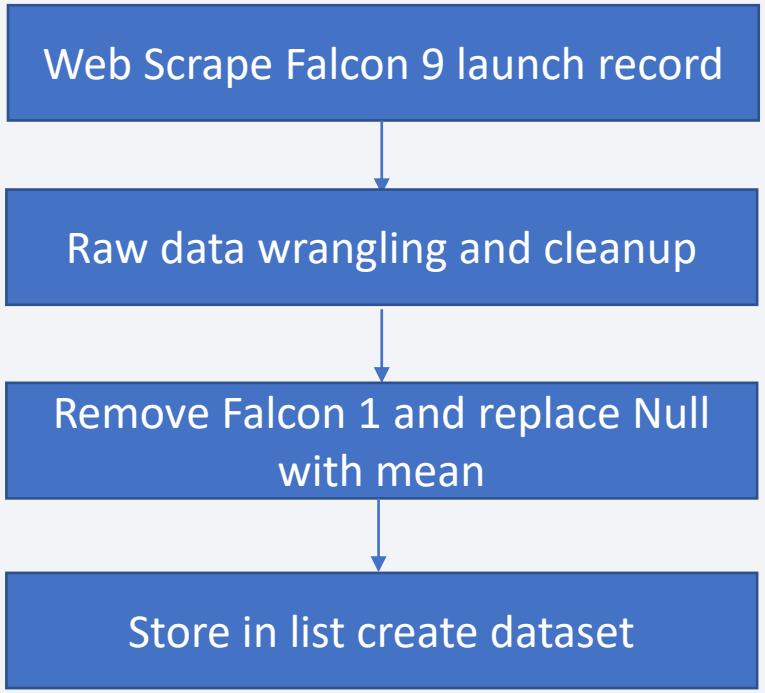
Methodology

Executive Summary

- Data collection methodology:
  - Using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.
- Perform data wrangling
  - Parse the data from web scrapped tables and convert them into a Pandas data frame for further visualization and analysis. Raw data is wrangled into a clean dataset which provides meaningful data. Null values are replaced by mean
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - split data into training data and test data to find the best Hyperparameter for SVM, Classification Trees, and Logistic Regression. Then find the method that performs best using test data.

Data Collection

Data sets were collected using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.

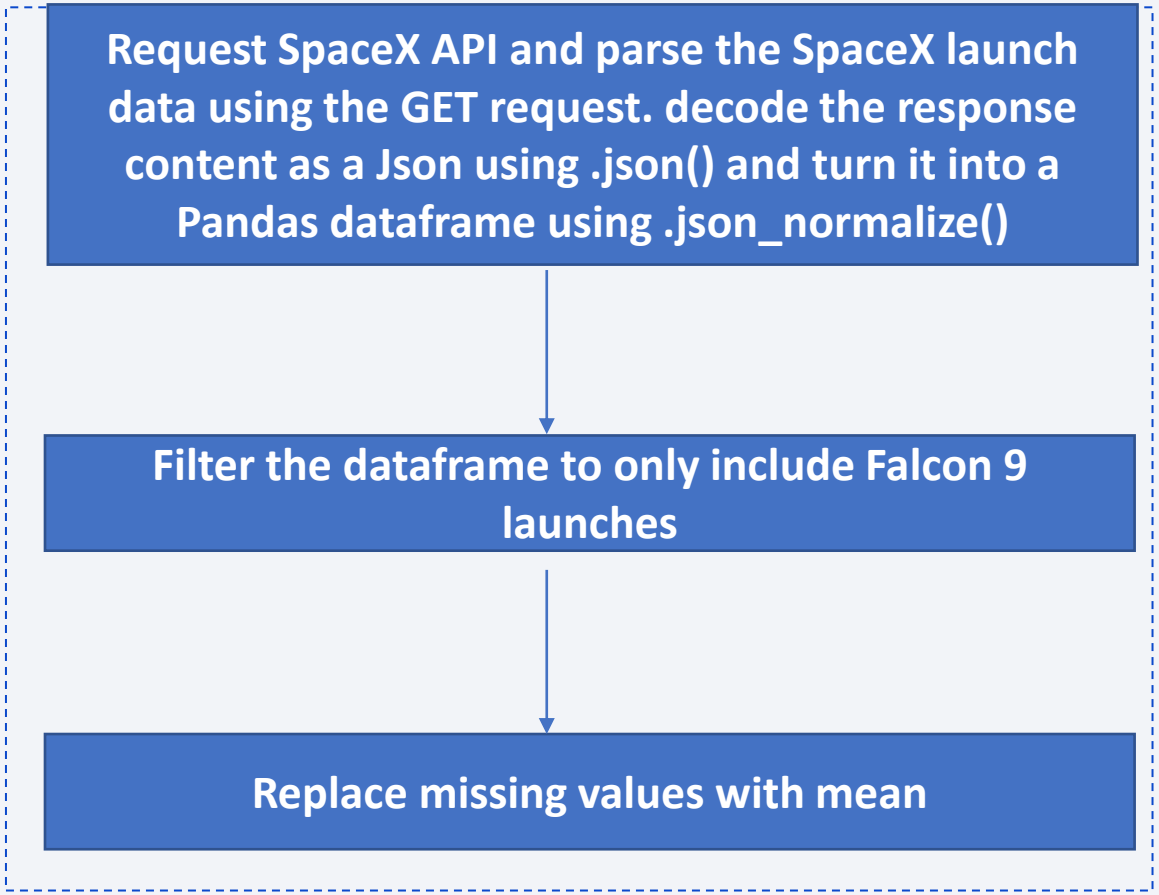


Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

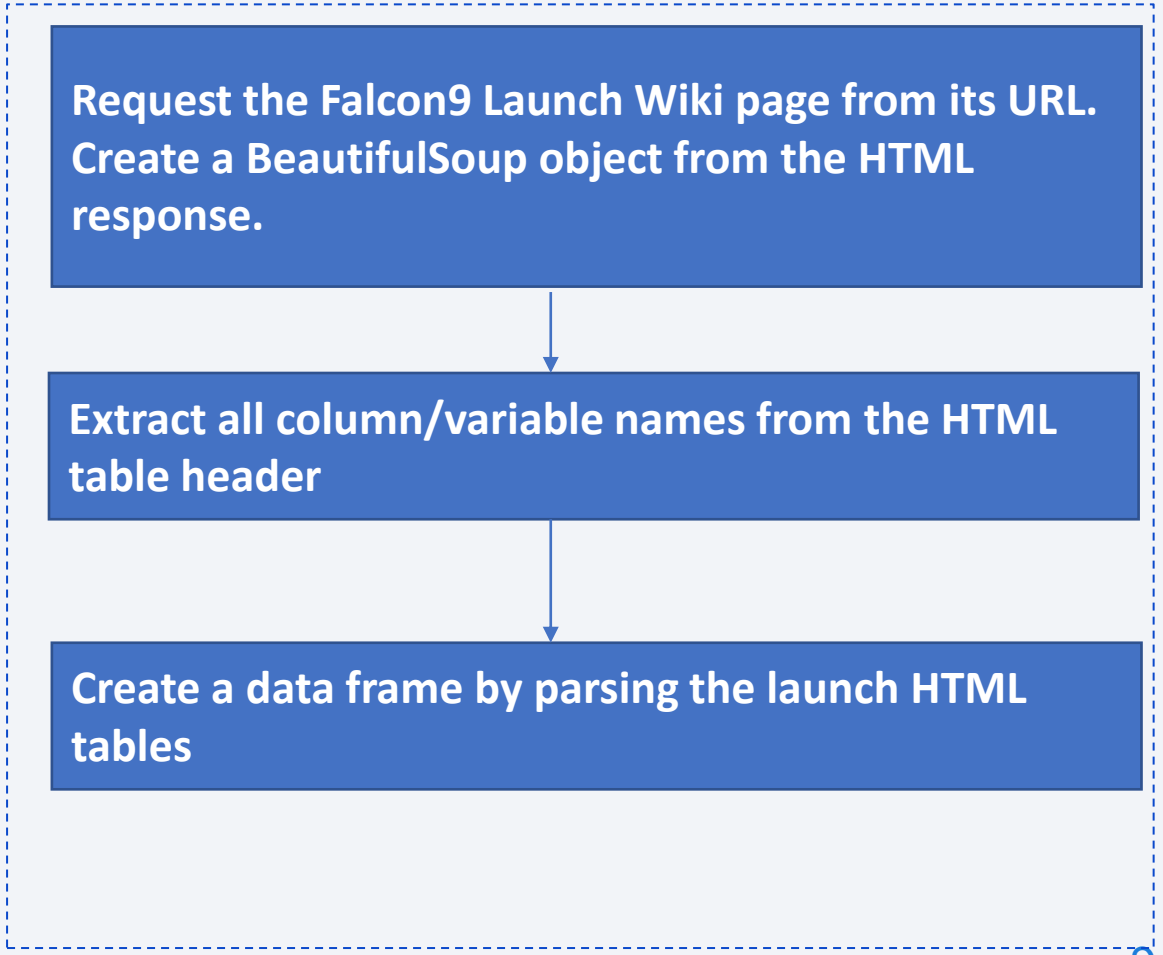
- GitHub URL of the completed SpaceX API calls



Data Collection - Scraping

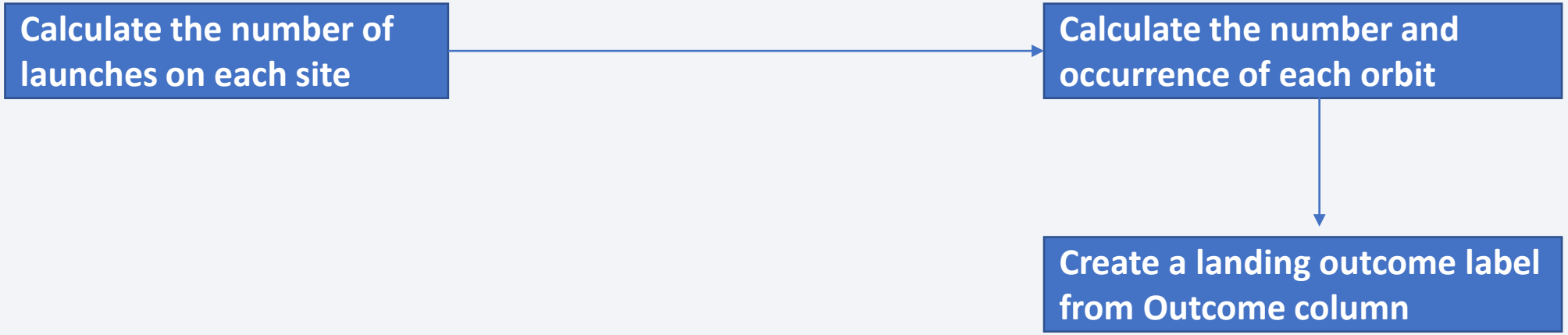
- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



Data Wrangling

- Describe how data were processed: perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- You need to present your data wrangling process using key phrases and flowcharts



- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

A dashboard to analyze launch records interactively with Plotly Dash. Interactive map to analyze the launch site proximity with Folium.

- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

- launch_site_location
- Build a Dashboard Application with Plotly Dash

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

- SELECT DISTINCT Launch_Site FROM chicago.spacextbl;
- SELECT * FROM chicago.spacextbl WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- SELECT SUM(PAYLOAD_MASS__KG_) FROM chicago.spacextbl WHERE customer ='NASA (CRS)';
- SELECT avg(PAYLOAD_MASS__KG_) FROM chicago.spacextbl WHERE Booster_Version ='F9 v1.1';
- SELECT min(Date) FROM chicago.spacextbl WHERE `spacextbl`.`Landing _Outcome` ='Success (ground pad)';
- SELECT Booster_Version FROM chicago.spacextbl WHERE `spacextbl`.`Landing _Outcome` = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
- SELECT COUNT(Mission_Outcome) Success FROM chicago.spacextbl WHERE Mission_Outcome LIKE '%Success%';
- SELECT COUNT(Mission_Outcome)Fail FROM chicago.spacextbl WHERE Mission_Outcome LIKE '%Fail%';
- SELECT Booster_Version FROM chicago.spacextbl WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM chicago.spacextbl);;
- SELECT `Landing _Outcome`,Booster_Version, Launch_Site FROM chicago.spacextbl WHERE `spacextbl`.`Landing _Outcome` ='Failure (drone ship)' AND YEAR(DATE) = 2015;
- SELECT `Landing _Outcome`,COUNT(`Landing _Outcome`) tt FROM chicago.spacextbl WHERE DATE Between '2010-06-04' and '2017-03-20' group by `Landing _Outcome` ORDER BY tt desc;

- GitHub URL

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- folium.Circle, folium.Marker, MarkerCluster , MousePosition

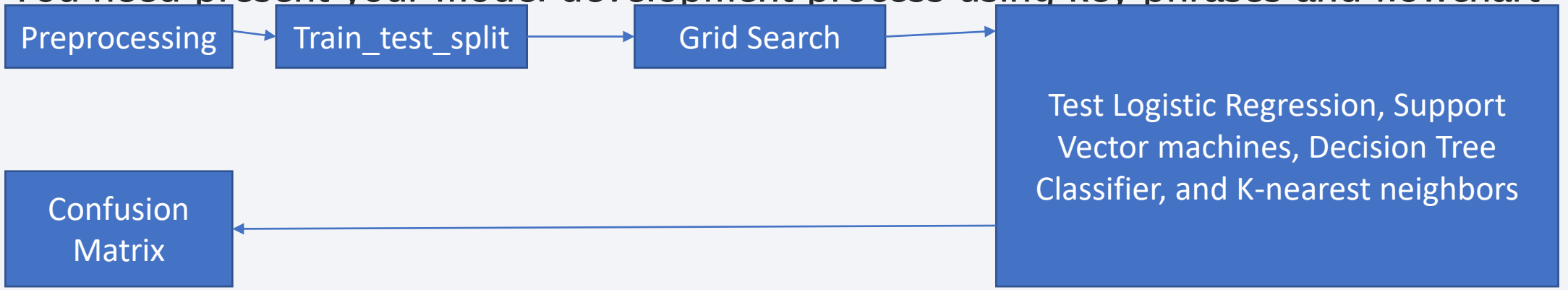
- Explain why you added those objects
- To Mark the success/failed launches for each site on the map
- To get coordinate for a mouse over a point on the map
- To calculate the distance between the coastline, point and the launch site
- GitHub URL

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Launch Site Drop-down Input Component
- success-pie-chart
- Range Slider to Select Payload
- Callback function to render the success-payload-scatter-chart scatter plot
- Explain why you added those plots and interactions
- For users to perform interactive visual analytics on SpaceX launch data in real-time.
- GitHub URL

Predictive Analysis (Classification)

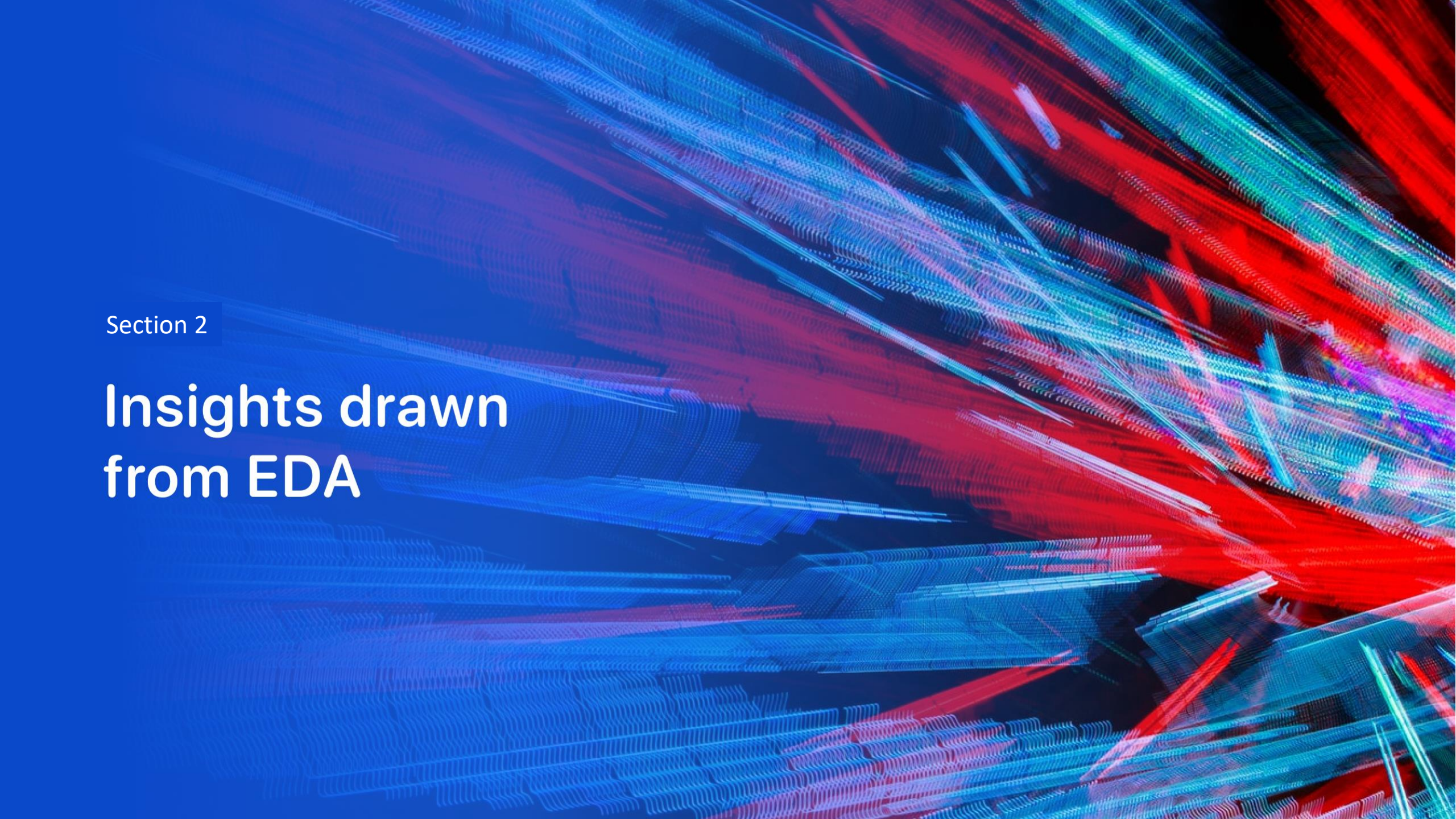
- Summarize how you built, evaluated, improved, and found the best performing classification model
- machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully
- Test Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors and output the confusion matrix
- You need present your model development process using key phrases and flowchart



- Add the GitHub URL

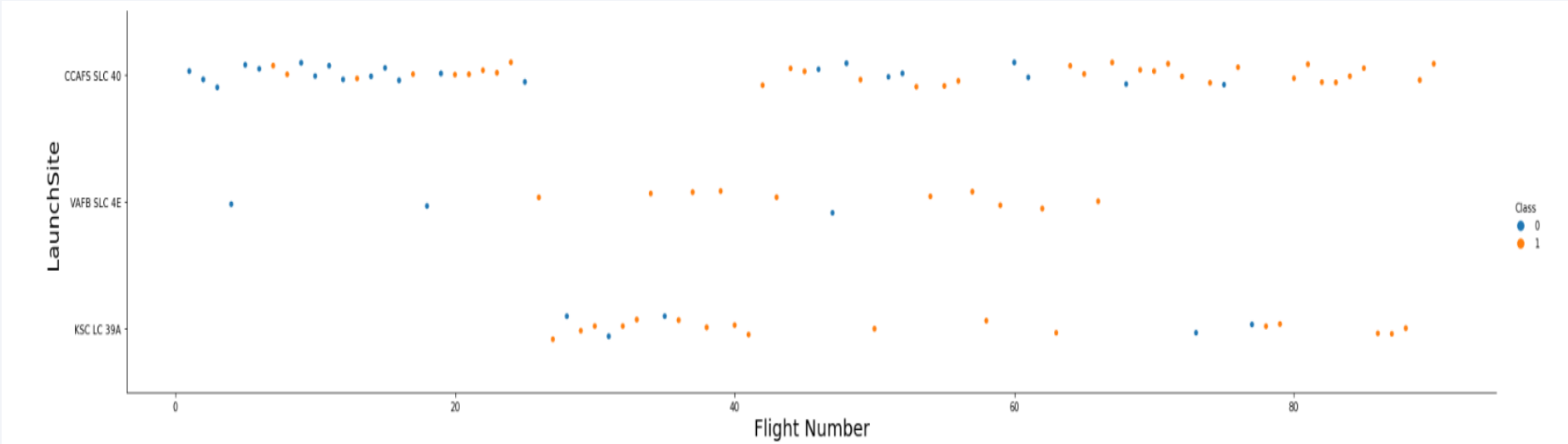
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

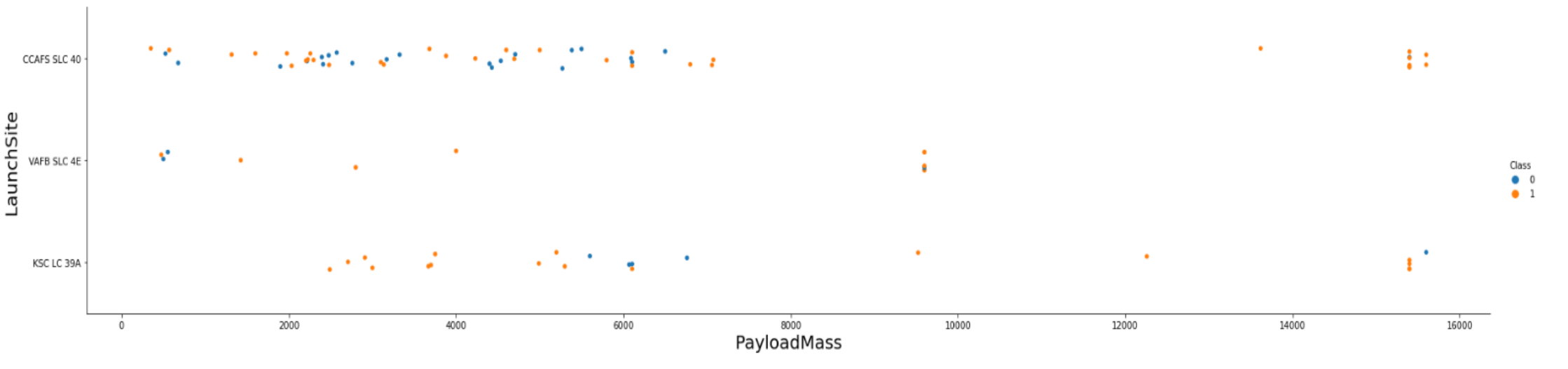
- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations



- Most flights launched at CCAFS SLC 40. Few flights launched at VAFB SLC 4E

Payload vs. Launch Site

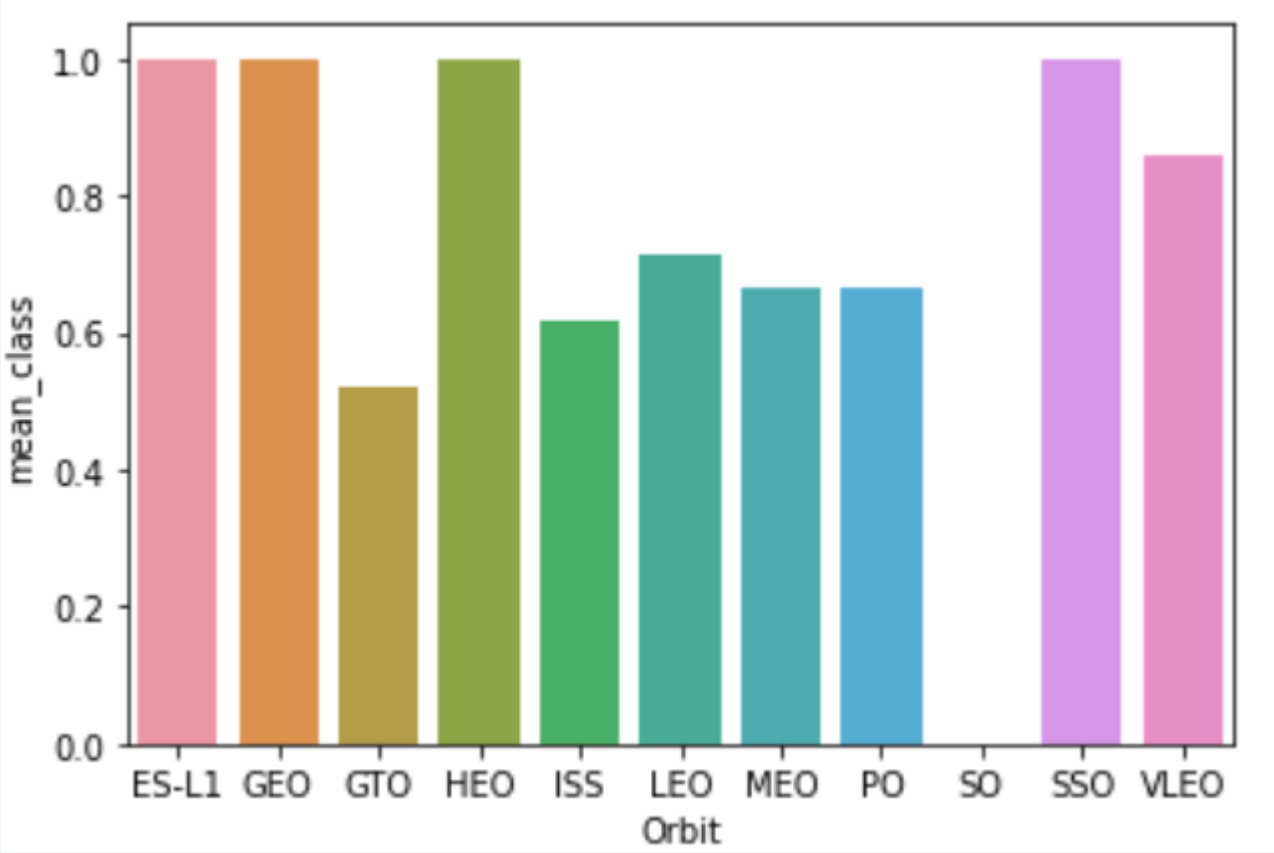
- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations



- VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- Fewer rockets launched for payloadmass above 6000

Success Rate vs. Orbit Type

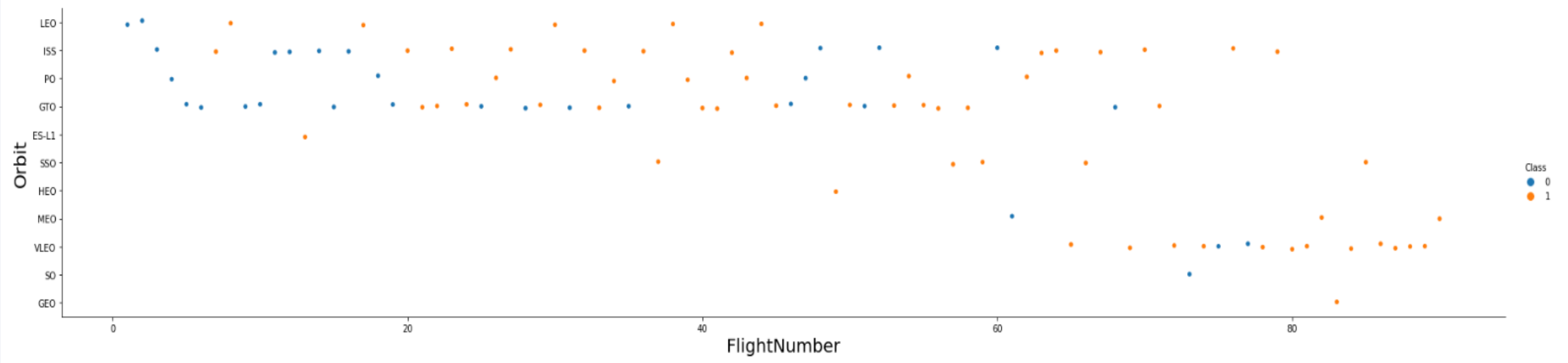
- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations



Orbit ES-L1, GEO, HEO, SSO has highest success rate at 1.0 followed by VLEO at 0.8, then majority range from 0.5 – 0.7. SO has 0.0 success rate

Flight Number vs. Orbit Type

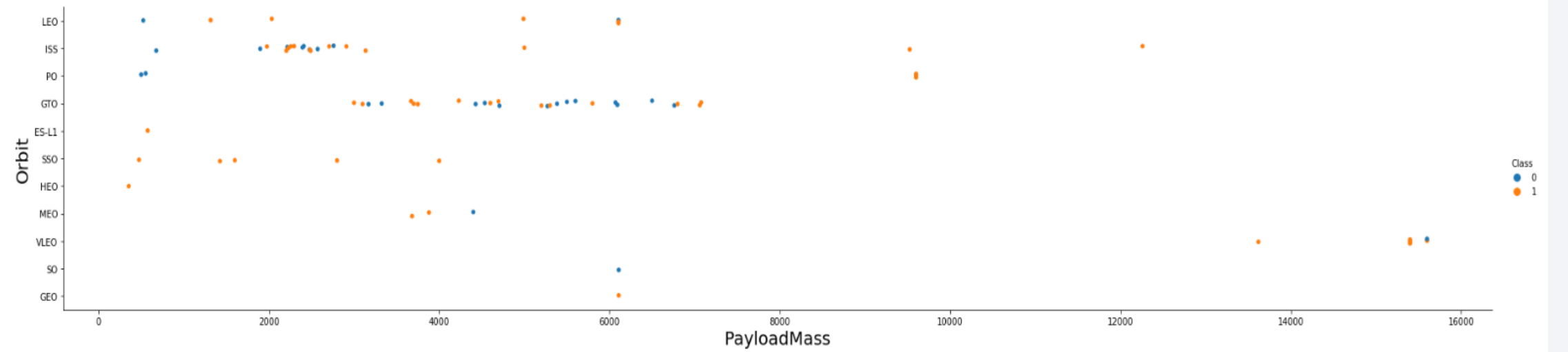
- Show a scatter point of Flight number vs. Orbit type
- Show the screenshot of the scatter plot with explanations



- Orbits LEO, ISS, PO, GTO and ES L1 have both low and high number of flights, rest are reached as number of flights increase. When flights are few success rate is low.

Payload vs. Orbit Type

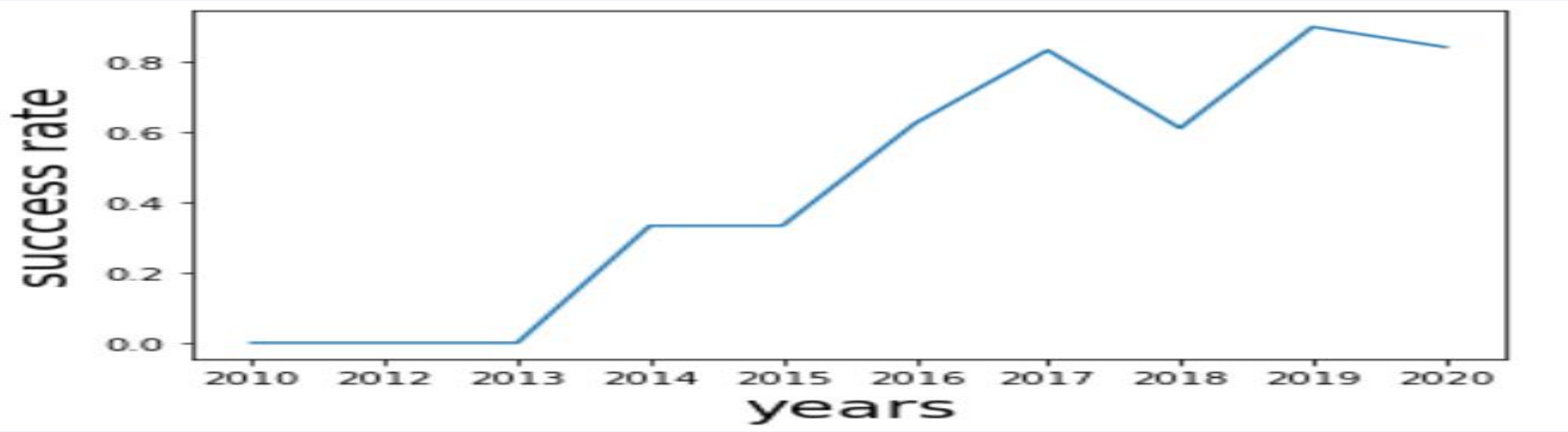
- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations



- Most orbits have payload below 6,000

Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations



- Success rates increased over the years from 2013 to 2017. A slight decrease in 2018 then continued increase to 2020

All Launch Site Names

- Find the names of the unique launch sites

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

- Present your query result with a short explanation here

```
SELECT DISTINCT Launch_Site FROM chicago.spacextbl;
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

- Present your query result with a short explanation here
- SELECT * FROM chicago.spacextbl WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

- Present your query result with a short explanation here

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM chicago.spacextbl WHERE
customer ='NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4000

- Present your query result with a short explanation here

```
SELECT avg(PAYLOAD_MASS__KG_) FROM chicago.spacextbl  WHERE
Booster_Version ='F9 v1.1';
```

- Uses the avg function on the payload column and filter booster version F9 v1.1.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

min(Date)
2015-12-22 00:00:00

- Present your query result with a short explanation here

```
SELECT min(Date) FROM chicago.spacextbl WHERE  `spacextbl`.`Landing
_Outcome` ='Success (ground pad)';
```

- The first successful date is minimum date using min() function and filter 'Success (ground pad)';

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

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

- Present your query result with a short explanation here

```
SELECT Booster_Version FROM chicago.spacextbl WHERE  `spacextbl`.`Landing_Outcome` = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000  AND PAYLOAD_MASS__KG_ < 6000;
```

Filter has landing outcome of 'Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Success	Fail
100	1

- Present your query result with a short explanation here

```
SELECT COUNT(Mission_Outcome) Success FROM chicago.spacextbl WHERE Mission_Outcome
LIKE '%Success%';

SELECT COUNT(Mission_Outcome) Fail FROM chicago.spacextbl

WHERE Mission_Outcome LIKE '%Fail%';
```

- All mission outcome with success are counted also with fail

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

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

```
SELECT Booster_Version FROM
chicago.spacextbl
WHERE   PAYLOAD_MASS__KG_ = (SELECT
MAX(PAYLOAD_MASS__KG_) FROM
chicago.spacextbl);
```

Using subquery the maximum payload is filtered

2015 Launch Records

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

Landing _Outcome	Booster_Version	Launch_Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Present your query result with a short explanation here

```
SELECT `Landing _Outcome`,Booster_Version, Launch_Site FROM chicago.spacextbl WHERE `spacextbl`.`Landing _Outcome` ='Failure (drone ship)' AND YEAR(DATE) = 2015
```

Columns Landing _Outcome, Booster_Version, Launch_Site with Landing _Outcome` ='Failure (drone ship)'

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing _Outcome	tt
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

```
SELECT `Landing
_Outcome`,COUNT(`Landing
_Outcome`) tt  FROM
chicago.spacextbl
WHERE DATE Between '2010-
06-04' and '2017-03-20'
group by `Landing _Outcome`
ORDER BY tt desc
```

- Present your query result with a short explanation here

Landing outcome counts are ranked by desc.

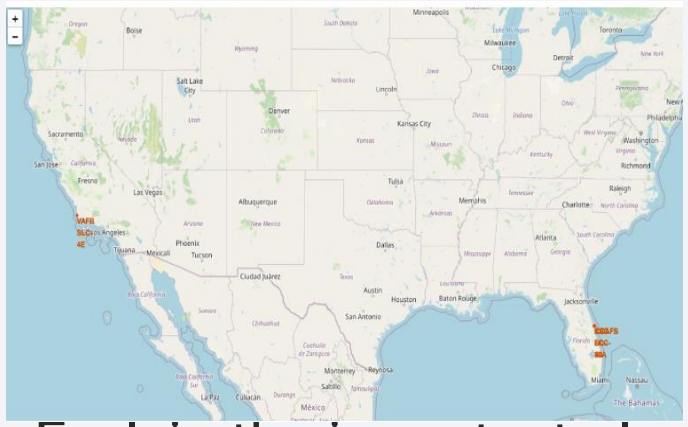
Section 3

Launch Sites Proximities Analysis

Mark all launch sites on a map

- Replace <Folium map screenshot 1> title with an appropriate title

- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

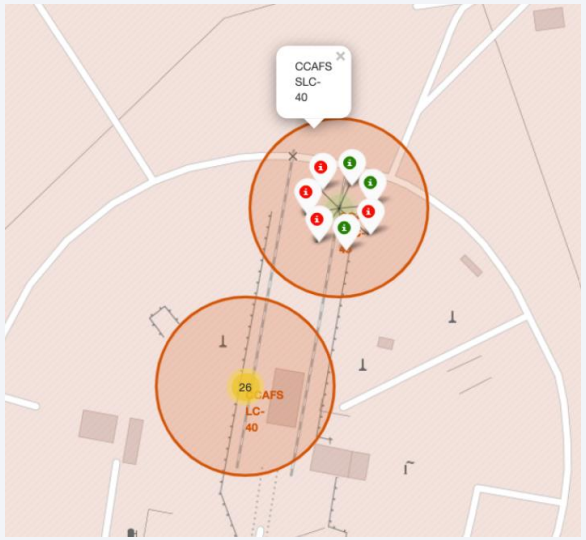


- Explain the important elements and findings on the screenshot

- In Florida and Los Angeles areas

The success/failed launches for each site on the map

- 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

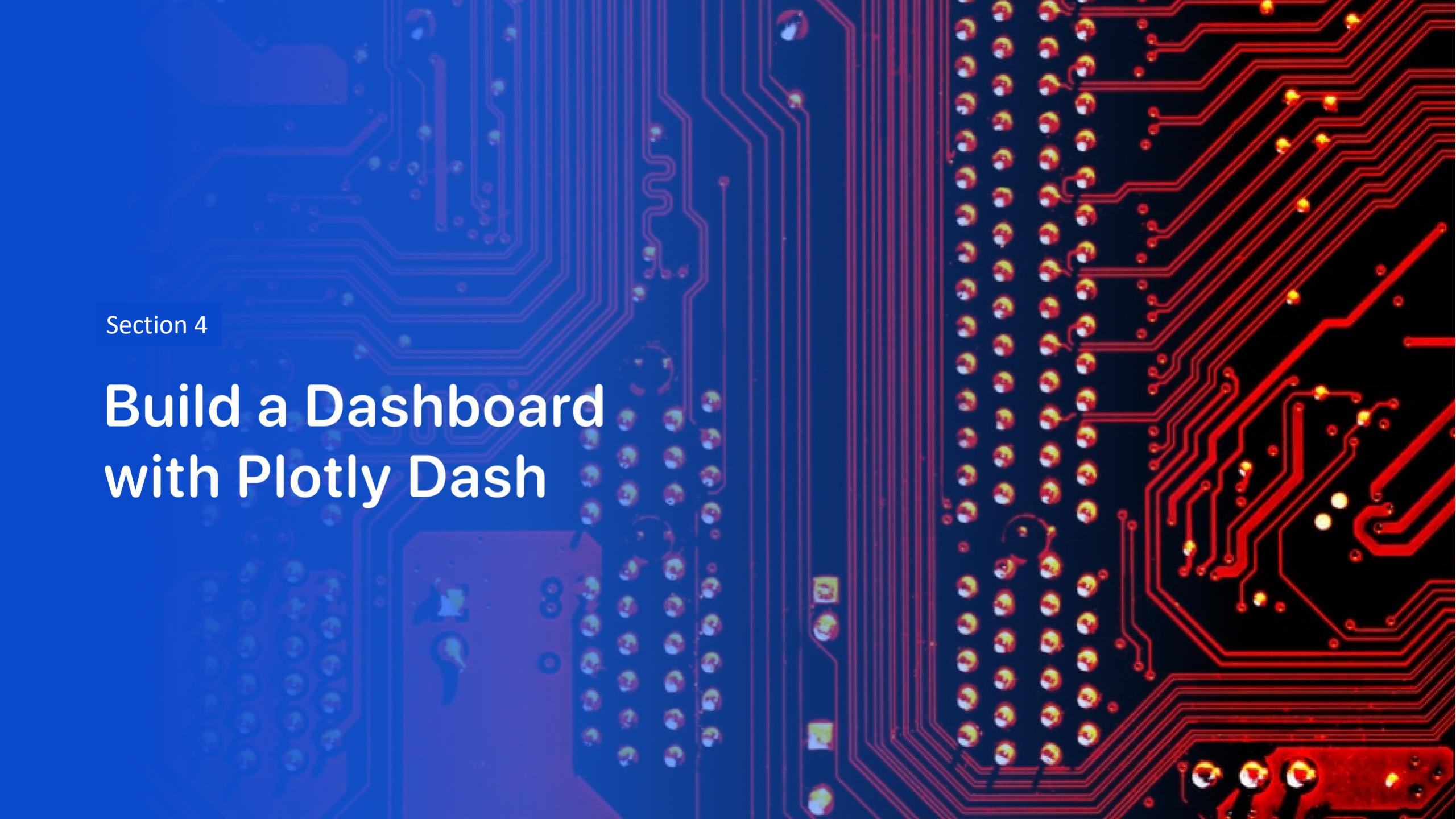


- Explain the important elements and findings on the screenshot

high success rates are green, and failure are red

The distances between a launch site to its proximities

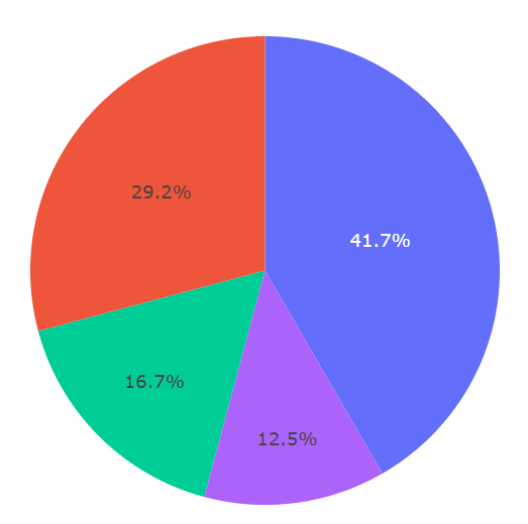
- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the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


Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites

- Replace <Dashboard screenshot 1> title with an appropriate title

- Show the screenshot of launch success count for all sites, in a piechart

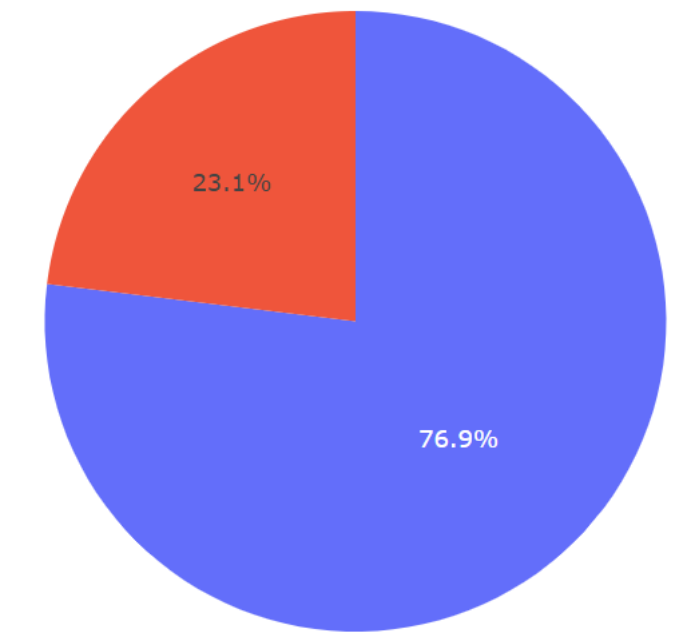


KSC LC-39A has highest success rate at 41.7% followed by CCAFS LC-40 at 29.2% then VAFB SLC-4E at 16.7% and KSC LC-39A at 12.5 %

- Explain the important elements and findings on the screenshot

Launch site with highest launch success ratio

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio

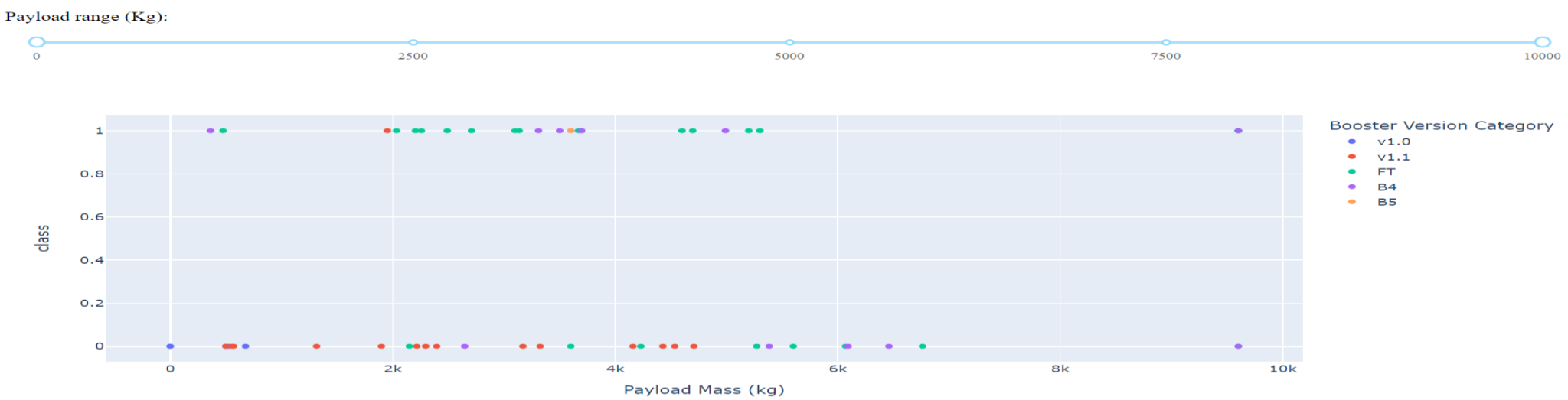


The launch site with highest success ratio is KSC LC-39A with success 76.9% and fail at 23.1%

- Explain the important elements and findings on the screenshot

Payload vs. Launch Outcome scatter plot for all sites

- 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



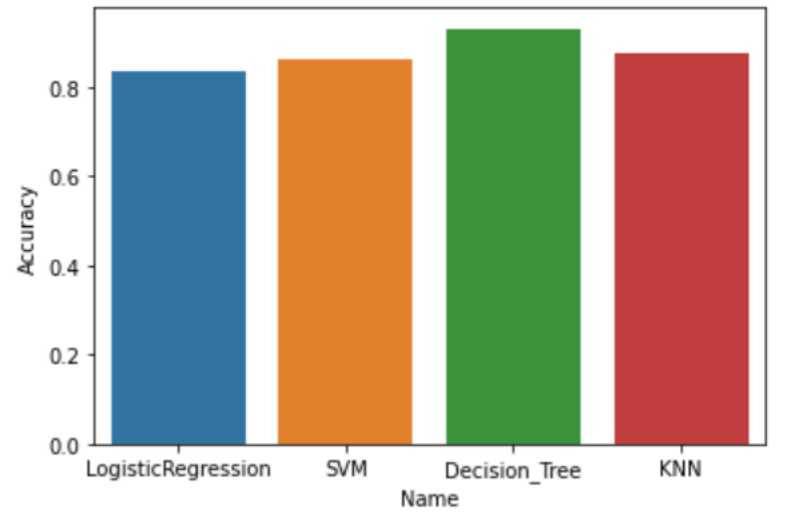
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.
- Booster version with largest success rate is FT and B4.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

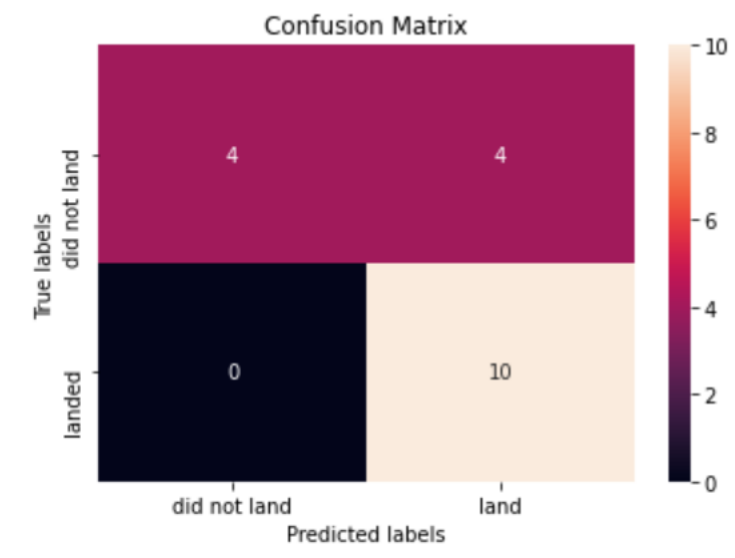
- Visualize the built model accuracy for all built classification models, in a bar chart



- Find which model has the highest classification accuracy. Decision Tree have highest accuracy on built classification on training data

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



High prediction landing at 10 which landed. Confirmed by 0 prediction that did not land confirmed by 0 landing

Conclusions

- Decision tree model have highest accuracy on built classification on training data
- Both KNN and SVM performs best using test data with accuracy of 0.7777777777777778
- High prediction landing at 10 which landed. Confirmed by 0 prediction that did not land confirmed by 0 landing
- Most flights launched at CCAFS SLC 40

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- https://github.com/Win1otieno2/IBM-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

```
methods = {'Name':['LogisticRegression','SVM','Decision_Tree','KNN'],'Accuracy':[0.8357142857142857,0.8625,0.9303571428571429,0.8767857142857143]}

methods

methods_df  = pd.DataFrame.from_dict(methods)

methods_df

sns.barplot(data=methods_df, x="Name", y="Accuracy")
```

Thank you!