Winning Space Race with Data Science

Rayhaan Pirani
May 7th, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using API
  - Data Collection with web scraping
  - Data Wrangling
  - Exploratory Data Analysis with Visualization
  - Exploratory Data Analysis using SQL
  - Interactive Maps with Folium
  - Interactive Dashboards with Plotly Dash
  - Launch Prediction using Machine Learning
- Summary of all results
  - Results from Exploratory Data Analysis results
  - Results from interactive analytics using maps and dashboards
  - Results from predictions using machine learning algorithms

Introduction

- Project background and context
  - Determine the success of Falcon 9 first stage landings to estimate launch costs
  - SpaceX's Falcon 9 launches cost $62 million compared to competitors' $165 million due to reusability of the first stage
  - Landing prediction helps determine launch cost, aiding bidding against SpaceX
- Problems to find answers for
  - Identify factors of successful rocket landings
  - Examine how various factors interact to impact landing success
  - Determine necessary conditions for a successful landing program

This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

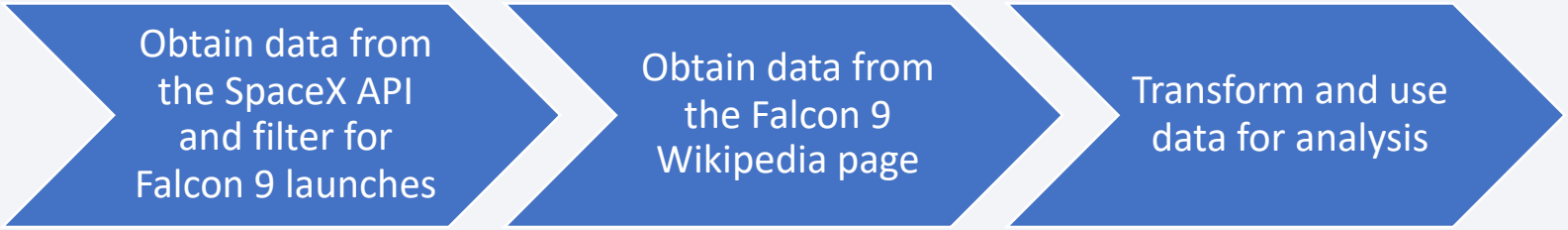
Methodology

Executive Summary

- Data collection methodology:
  - Data collected from the SpaceX API and Falcon 9 Wiki, transformed, and used for analysis
- Perform data wrangling
  - Data analyzed for launch and mission successes, and a class column created to determine success or failure
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models Logistic Regression, SVM, Decision Tree, KNN were used with Grid Search cross-validation to obtain the best hyperparameters and accuracy score computed on test dataset split from the original data into training and test sets

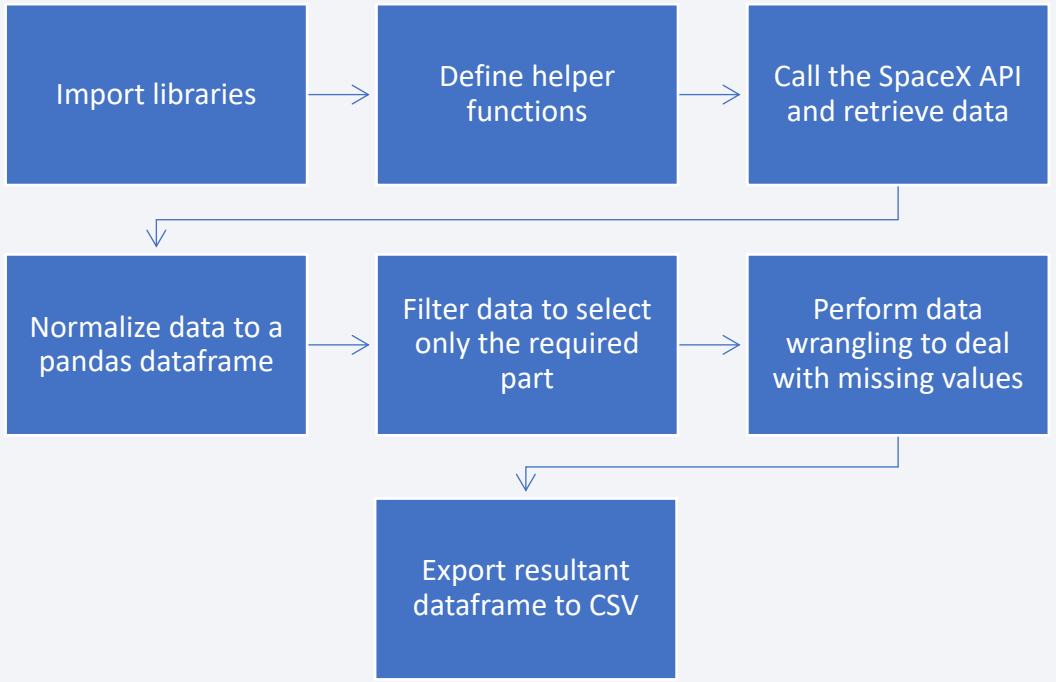
Data Collection

- The primary data sources for this project are
  - SpaceX API
  - Falcon 9 Wiki
- We first obtain the data from the SpaceX API and transform it to a dataframe. We filter the dataset to only include Falcon 9 launches.
- We then obtain the data from the Falcon 9 launch Wiki and transform it.
- We use these two datasets in our analysis.



Data Collection – SpaceX API

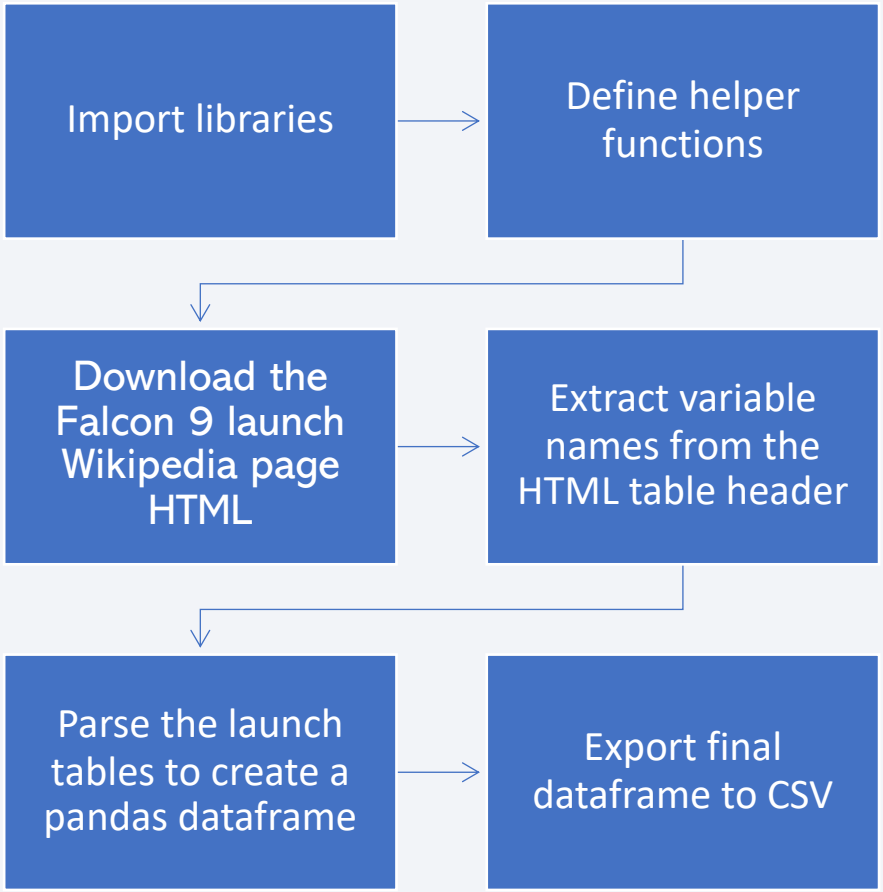
- First, we import the required libraries, such as numpy, pandas, requests, and datetime in our environment.
- Then, we define helper functions for data loading: getBoosterVersion, getLaunchSite, getPayloadData, and getCoreData.
- We then call the SpaceX API and retrieve data in the form of a JSON object.
- We convert this JSON object into a pandas dataframe for easy manipulation and analysis.
- We then filter out irrelevant data, i.e., we only select data relevant to the Falcon 9 booster version.
- We perform data wrangling on the data to deal with missing values.
- Finally, we export the result to a CSV file.



Reference notebook: https://github.com/RayhaanPirani/spacex-falcon9-landing/blob/master/SpaceX%20Falcon%209%20first%20stage%20Landing%20Prediction.ipynb

Data Collection – Scraping

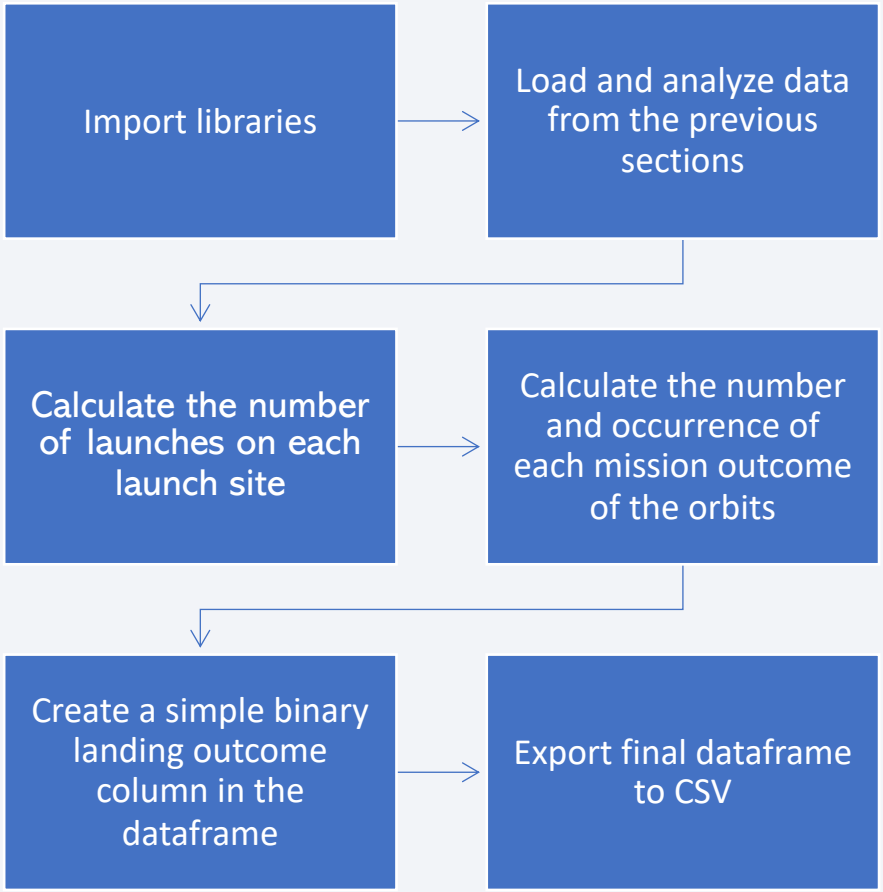
- First, we download and import the required libraries for scraping, such as BeautifulSoup, pandas, and requests.
- Then, we define helper functions to process the scraped HTML.
- We then send a request and download the Falcon 9 launch Wikipedia page in HTML format.
- We extract the variables that we need from the column names of the HTML table header by parsing the page.
- We parse the launch tables in the page and retrieve required data to store it as a pandas dataframe.
- Finally, we export this result to a CSV file as well.



Reference notebook: https://github.com/RayhaanPirani/spacex-falcon9-landing/blob/master/Space%20X%20Falcon%209%20First%20Stage%20Landing%20-%20Scraping.ipynb

Data Wrangling

- First, we download and import the required libraries for data wrangling, such as numpy and pandas.
- Then, we load the data from the previous section into a dataframe and analyze it to check missing values and data types.
- We then calculate the number of launches done from each launch site.
- We calculate the number and occurrence of each mission outcome of the orbits.
- We then make a simple binary landing outcome column in the dataframe called 'Class'.
- Finally, we export this result to a CSV file.



Reference notebook: https://github.com/RayhaanPirani/spacex-falcon9-landing/blob/master/SpaceX%20Falcon%209%20Landing%20Prediction%20-%20Wrangling.ipynb

EDA with Data Visualization

- We first plotted scatter plots between Flight Number and Launch Site, and also between Payload mass and Launch Site to see their correlation.
- We then plotted a bar graph showing the success rate for each Orbit type to find the orbits that had proportionally more successful launches.
- We also plotted the relationships between Flight Number and Orbit type, as well as between Payload mass and Orbit type in the form of scatter plots to find if there is any connection between them.
- The final plot we created was the success rate over the years, to observe the trend of successful launches over time.

Reference notebook: https://github.com/RayhaanPirani/spacex-falcon9-landing/blob/master/Falcon%209%20launch%20prediction%20-%20EDA%20%2B%20Visualization.ipynb

EDA with SQL

- We queried the distinct launch sites in the dataset.
- We looked at the data associated with launch sites which had names starting with 'CCA'.
- We calculated the total payload mass carried by boosters launched by NASA (CRS).
- We determined the average payload mass carried by booster version F9 v1.1.
- We computed the date when the first successful landing in ground pad was made.
- We listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- We found how many missions were a success or a failure.
- We discovered the booster versions involved in carrying the heaviest payload.
- We provided the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- We ranked the count of successful landing_outcomes between the dates 04-06-2010 and 20-03-2017 in descending order.

Reference notebook: https://github.com/RayhaanPirani/spacex-falcon9-landing/blob/master/Falcon%209%20landing%20prediction%20-%20EDA.ipynb

Build an Interactive Map with Folium

- We marked all the launch sites on the map to geographically locate them.
- For each site, we created a marker cluster showing the successful and failed launches in green and red markers respectively to visualize their success rates.
- From a marker, we drew lines between it and its close proximities like the coast line and railroad to demonstrate how close the launch sites are from such places of interest, and how far they are from cities.

Reference notebook: https://github.com/RayhaanPirani/spacex-falcon9-landing/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

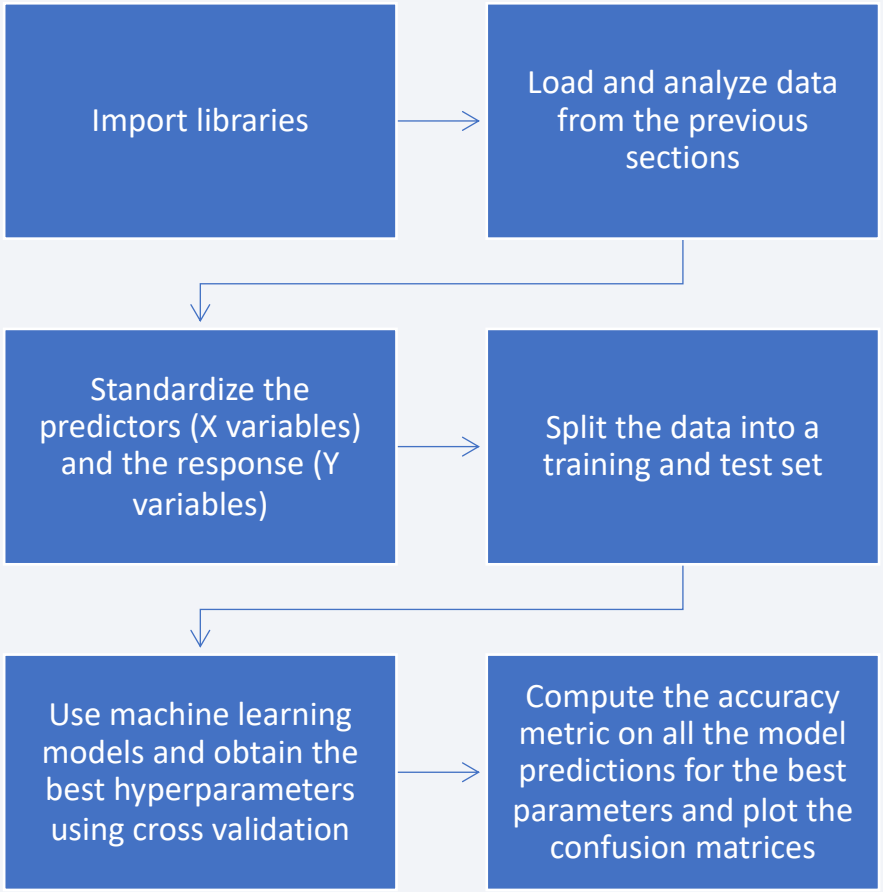
Build a Dashboard with Plotly Dash

- We created a pie chart demonstrating the success rate for each launch site.
  - A generic pie chart would show the proportion of successful launches among all the sites.
  - There is an option to choose a launch site in a drop down in the dashboard. Selecting a launch site from this drop down shows the successful and failed launches in a pie chart for that site.
- We also made a scatter plot that shows the relationship between the success and failure class versus the payload mass for each booster version category.
  - The points in the scatter plot are colored based on the booster version category.
  - A range slider in the dashboard provides the ability for the user to choose their own payload range and focus the chart to display only the selected range.

Reference notebook: https://github.com/RayhaanPirani/spacex-falcon9-landing/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- First, we download and import the required libraries for machine learning and visualization, such as numpy, pandas, scikit-learn, matplotlib and seaborn.
- Then, we load the processed data from the previous sections into a dataframe and analyze it.
- We standardize the response variable Y (in this case, the 'class' denoting the success of a launch) by converting it into a numpy array. We standardize the predictors X using standard scaling.
- We split the data 80%-20% into training and test datasets respectively.
- We use classification machine learning models logistic regression, SVM, decision trees, and K-nearest neighbors. We use grid-search cross-validation for hyperparameter tuning to find the best parameters for each of these models.
- Using the best parameters, we compute the accuracy score for each model on the predictions obtained by them and plot the respective confusion matrices.

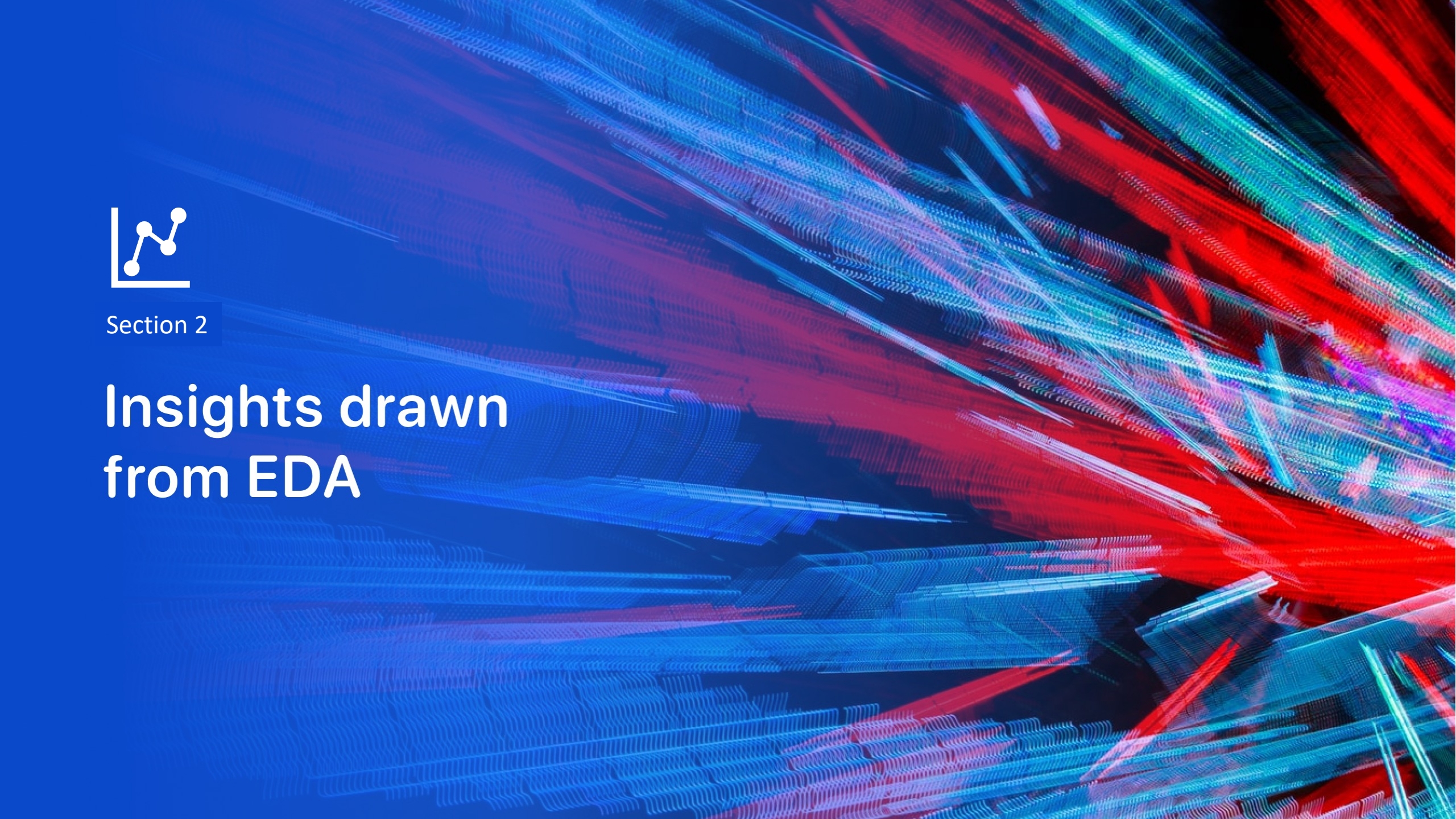


Reference notebook: https://github.com/RayhaanPirani/spacex-falcon9-landing/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

The following section will discuss the results of our analysis. The results that we shall discuss are

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

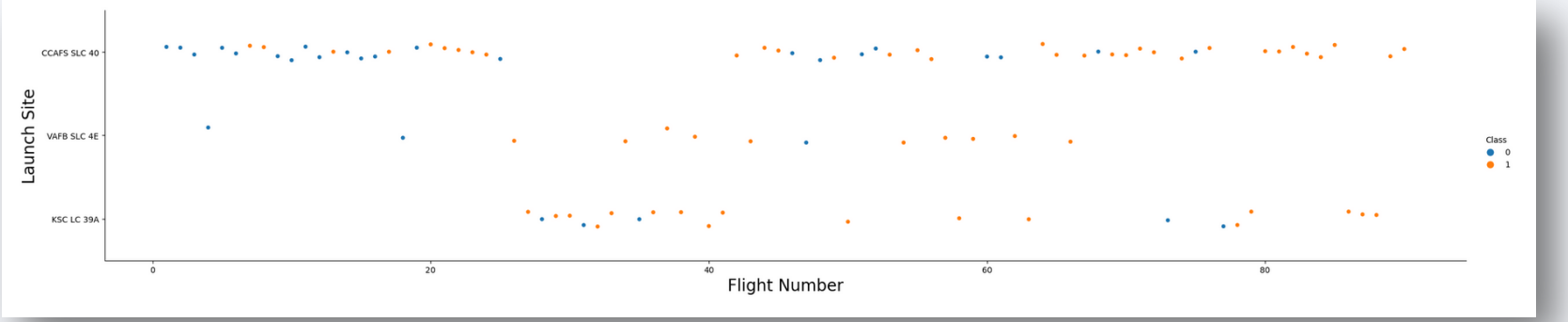


Section 2

Insights drawn from EDA

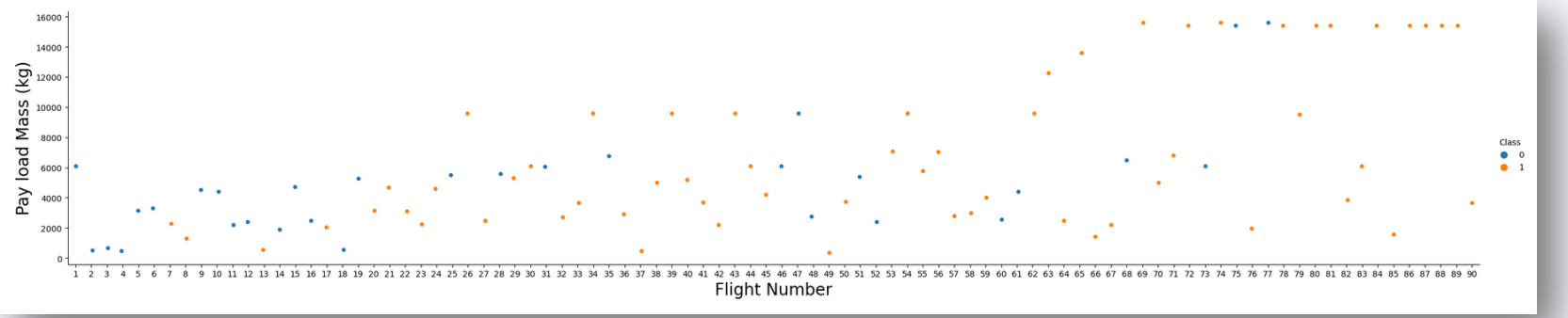
Flight Number vs. Launch Site

- The plot between the Flight Number and the Launch Site shows the launches for each flight number over each launch sites. The points are colored to show successful and unsuccessful launches.
- The following are some observations that we can make from the scatter plot:
  - We observe that in general, higher flight numbers are more successful.
  - 'VAFB SLC 4E' has the least number of launches while 'CCAFS SLC 40' has the most.



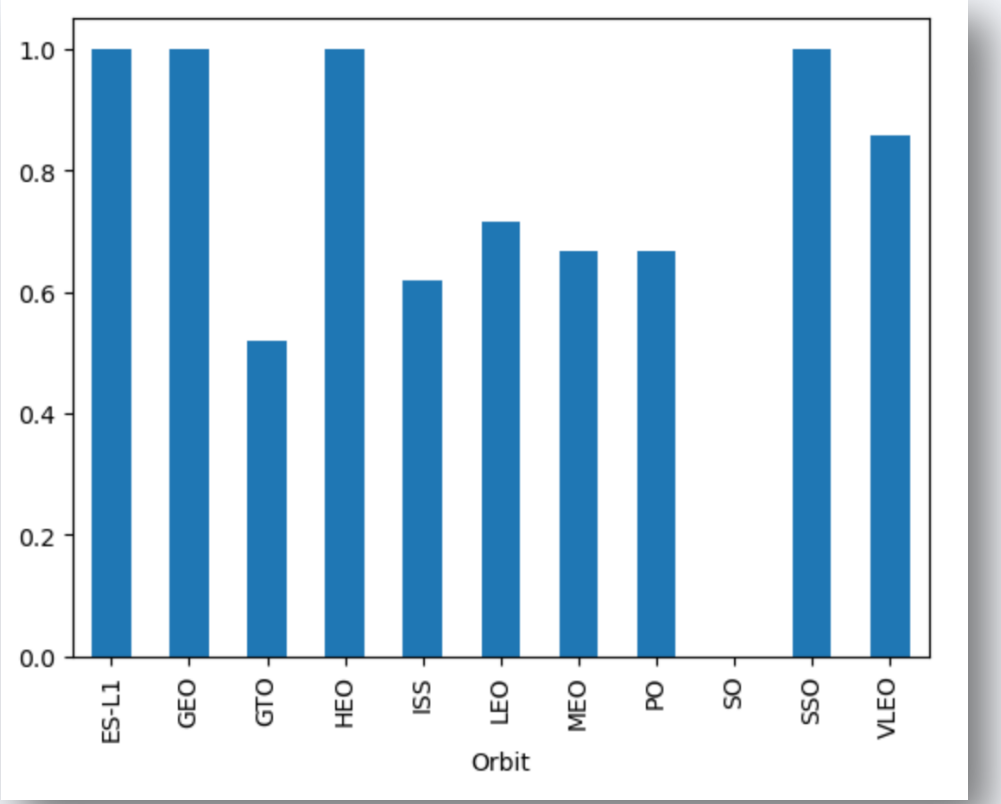
Payload vs. Launch Site

- The plot between the Payload mass and the Launch Site plots the launches showing the payload mass over each launch sites. The points are colored to show successful and unsuccessful launches.
- The following are some observations that we can make from the scatter plot:
  - We observe that in general, higher flight numbers have a heavier payload.
  - Lighter payload mass launches have been more unsuccessful than the heavier ones.



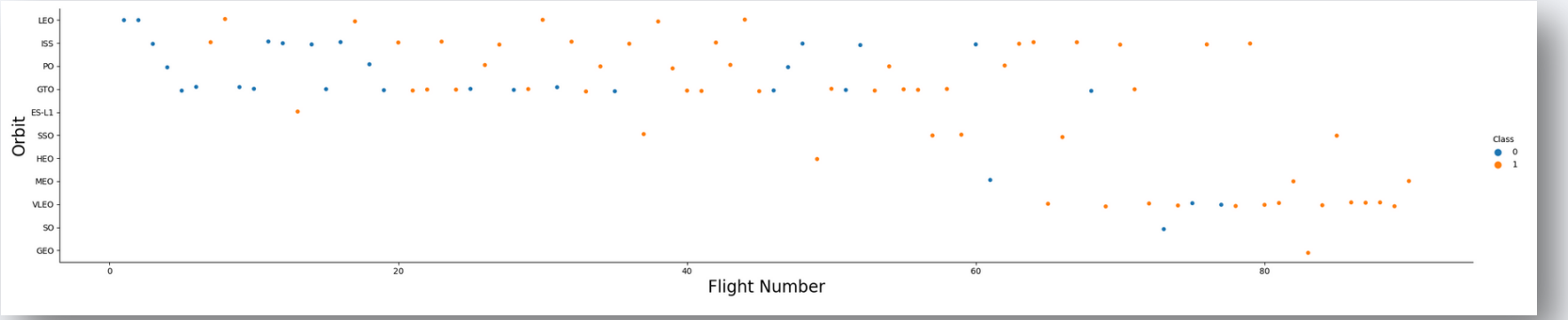
Success Rate vs. Orbit Type

- This bar graph shows the success rate for each orbit type.
- The orbit types ES-L1, GEO, HEO, and SSO have near perfect success rate, while the orbit type SO has never had a successful launch.
- VLEO, LEO, MEO, PO, and ISS orbit types have also been mostly successful, in that order.
- GTO has had around 50% success rate.



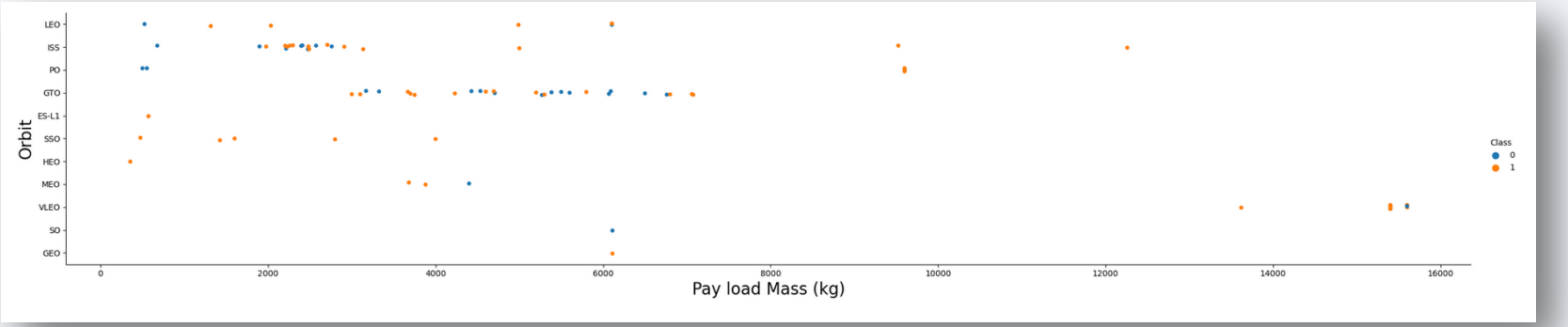
Flight Number vs. Orbit Type

- The plot between the Flight Number and Orbit type shows the launches for each flight number over each orbit type. The points are colored to show successful and unsuccessful launches.
- The following are some observations that we can make from the scatter plot:
  - The orbit VLEO is associated with higher flight numbers.
  - Some orbits like SSO have very few launches. Others like HEO, MEO, SO and GEO have had only one launch in the past.



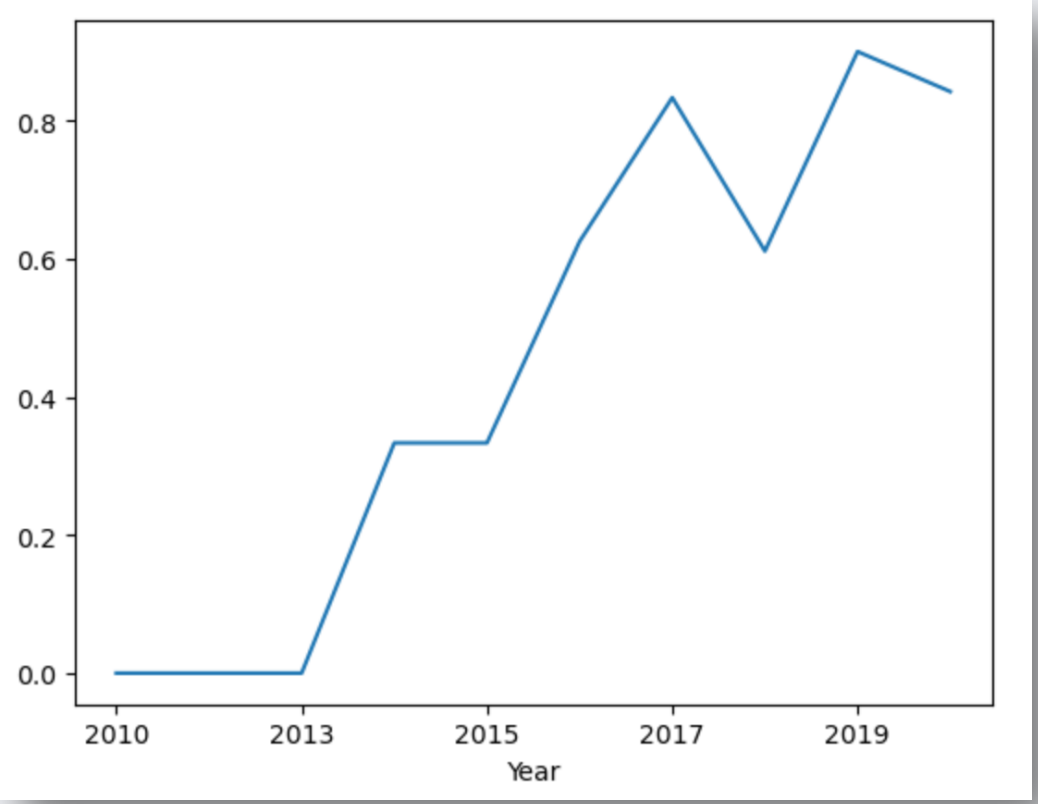
Payload vs. Orbit Type

- The plot between the Payload mass and the Orbit type shows launches with the payload mass over each orbit type. The points are colored to show successful and unsuccessful launches.
- The following are some observations that we can make from the scatter plot:
  - We observe that certain orbit types like ISS, SSO, HEO, MEO are associated with a lower payload mass while others like VLEO have launched only high mass payloads.
  - Some orbit types like SO and GEO have only had one launch in their history.



Launch Success Yearly Trend

- The line chart shows the relationship between time and success rate.
- We observe that there is a general trend of increasing success with time from 2013 to 2020.
- This trend might be because the engineers and scientists involved in the launches learn from previous shortcomings and improve over time.
- We also observe that lately there has been a slight drop in the success rate.



All Launch Site Names

- We select the distinct launch sites from our dataset.
- We find that the dataset has only four launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

- We find 5 records where launch sites begin with ‘CCA’.
- We see that the five records displayed here have the following characteristics.
  - They were launched in the LEO orbit type.
  - They had the booster version F9 v1.0
  - All of them were considered successful missions.
- However, the complete data may show other trends, since we are only viewing the first 5 records.

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- We calculate the sum of the payload column and filter by NASA (CRS) in the customer column.
- We observe that NASA (CRS) was involved in a total of 45,596 kg worth of payload mass launches.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_nasa_crs FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

total_payload_mass_nasa_crs
45596

Average Payload Mass by F9 v1.1

- We calculate the average of the payload column and filter by F9 v1.1 in the booster column.
- We observe that the F9 v1.1 booster was used in launches with an average payload mass of 2,928.4 kg.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Booster_Version" = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- We ordered the date by year, month, and date and selected the minimum date from the ordering to obtain the earliest date.
- We filtered the result by selecting the ‘Success (ground pad) landing outcome.
- The resultant date was ‘2015-12-22’.

```
%sql SELECT MIN(substr(Date,7,4) || '-' || substr(Date,4,2) || '-' || substr(Date,1,2)) AS earliest_ground_pad_success_date \
FROM SPACEXTBL WHERE "Landing _Outcome" == 'Success (ground pad)'
```

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

earliest_ground_pad_success_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.
- We obtain the distinct booster versions and filter the payload mass to be between 4000 and 6000 kg and the landing outcome to be ‘Success (drone ship) to get this result.

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- We select the count of rows for each mission outcome and group the results by mission outcome.
- We observe that almost all missions (100 missions) were considered successful, except for one.

```
%sql SELECT COUNT(*), Mission_Outcome FROM SPACEXTBL GROUP BY Mission_Outcome;
 * sqlite:///my_data1.db
Done.
```

COUNT(*)	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

- We use a subquery to calculate the maximum payload mass.
- We then use the result of that subquery in another query to select the booster version and filter the result by maximum payload mass.
- The resultant booster versions that carried the maximum payload mass are shown in the figure.

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- We select the month extracted from the date, the landing outcome, booster version, and launch site from the dataset and filter the result by landing outcome ‘Failure (drone ship)’, and the year extracted from the date being 2015.
- We obtain the pictured two records as a result.

```
%sql SELECT substr(Date, 4, 2) AS month, "Landing _Outcome", Booster_Version, Launch_Site \
        FROM SPACEXTBL \
        WHERE "Landing _Outcome" = "Failure (drone ship)" AND substr(Date,7,4) = '2015';
```

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

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

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

- We select the landing outcome and count of records, grouped by the landing outcome.
- We filter the Date with the range mentioned.
- We order by the count in descending order.
- We discover that most landing outcomes in the given date range were successful. The second most had no landing attempt.

```
%sql SELECT "Landing _Outcome", COUNT(*) \
        FROM SPACEXTBL \
        WHERE Date BETWEEN '04-06-2010' AND '20-03-2017' \
        GROUP BY "Landing _Outcome" \
        ORDER BY COUNT(*) DESC;
```

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

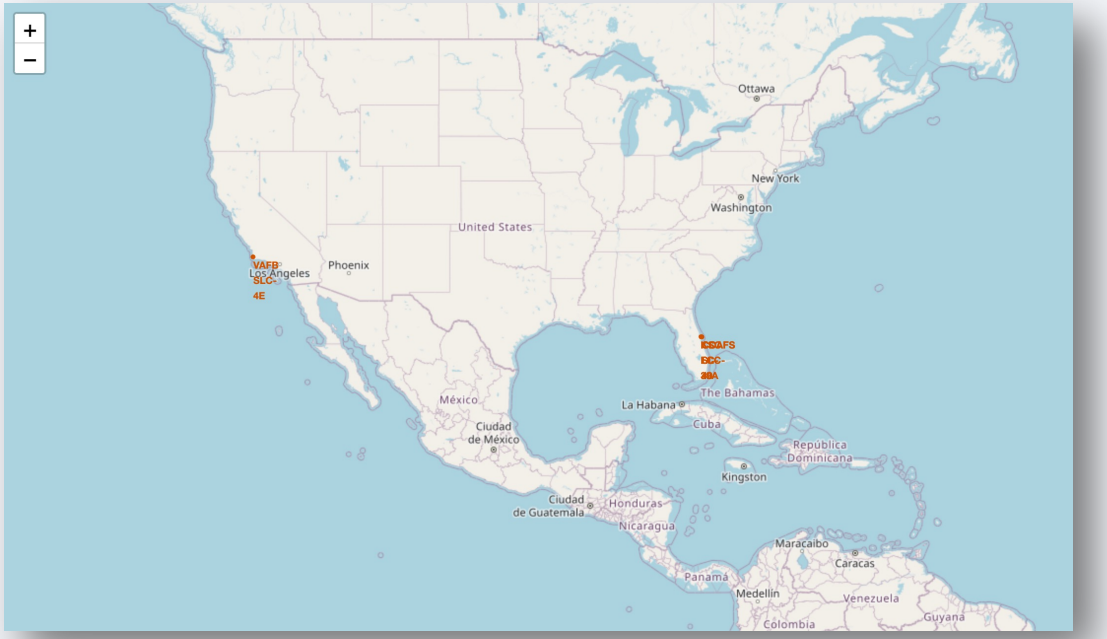
Landing _Outcome	COUNT(*)
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

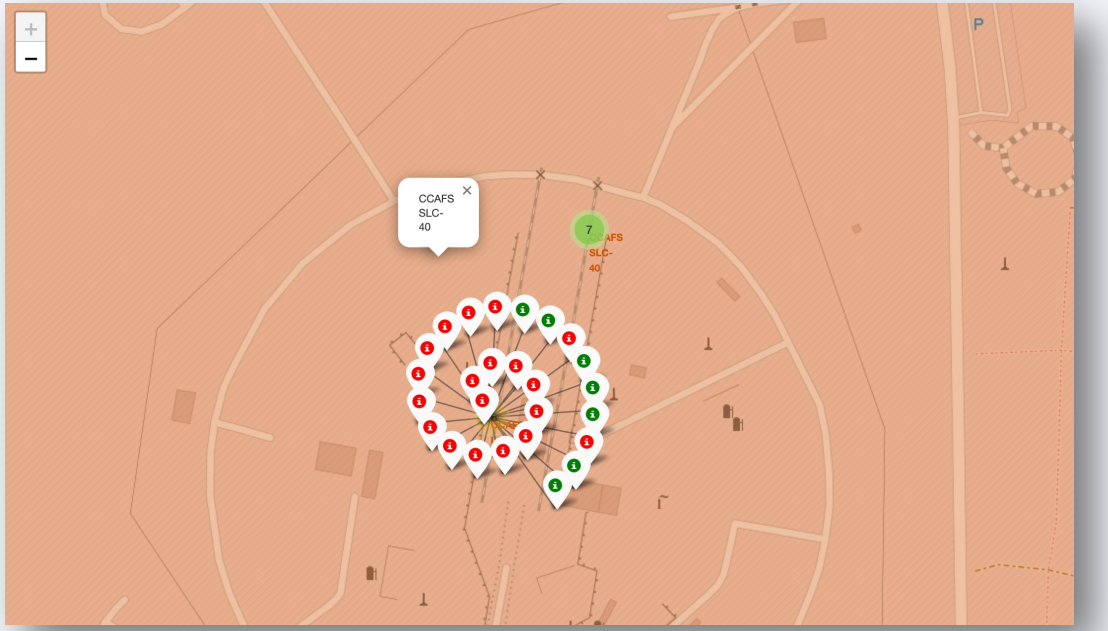
Launch Sites on the World Map

- We observe that all the launch sites are located in the United States and in coastal regions.
- Three launch sites are located in close proximity with each other in the east coast, particularly in Florida.
- The remaining launch site is located in the west coast, in California.



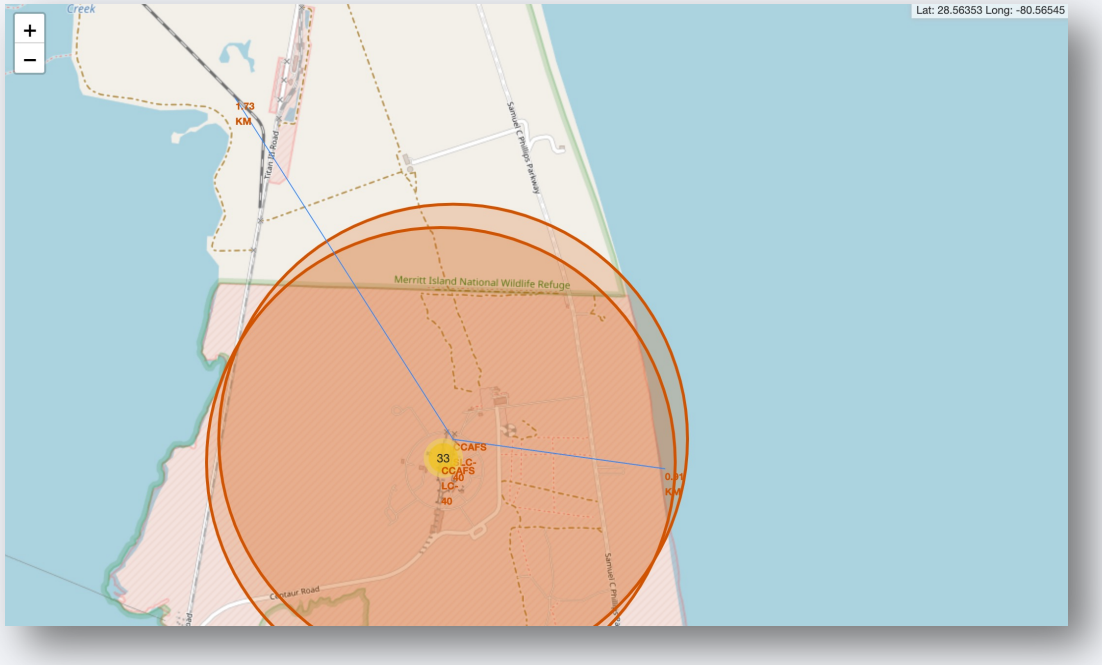
Site Success/Failed Launches Marking

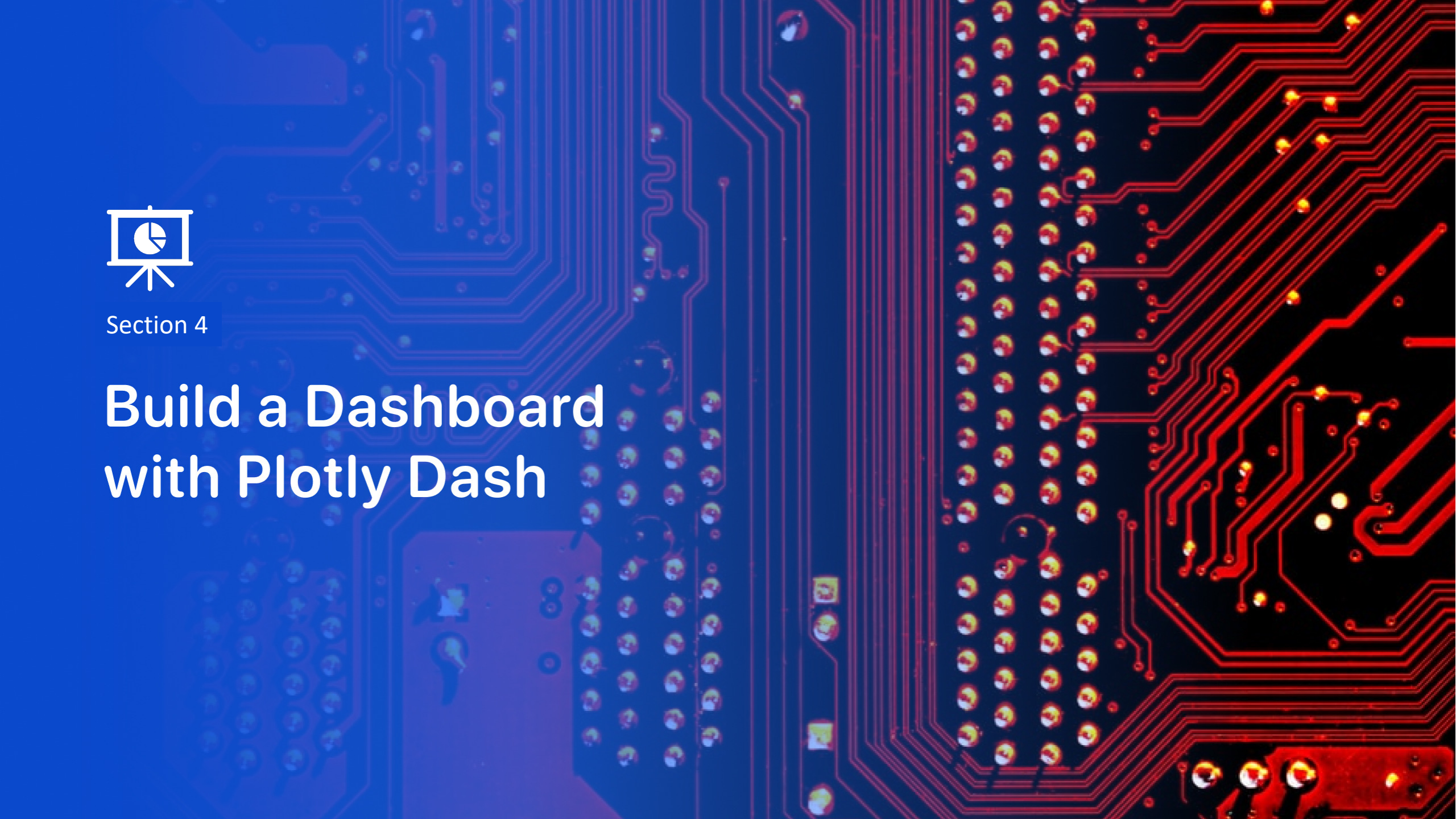
- The zoomed-in portion of the map shows the number of successful and unsuccessful launches with its own marker.
- The markings shown are for the site CCAFS LC-40. The same markings were also made for all the other sites.
- We can see that there were 19 failed launches and 7 successful ones at this site.



Distance between launch site to proximities

- The zoomed-in portion of the map shows the distance between the launch site and its proximities.
- The markings shown are for the site CCAFS SLC-40. The proximities shown are the nearest coastline and railway line.
- We can see that this site is 0.91 km away from the coastline and 1.73 km away from the NASA railroad.



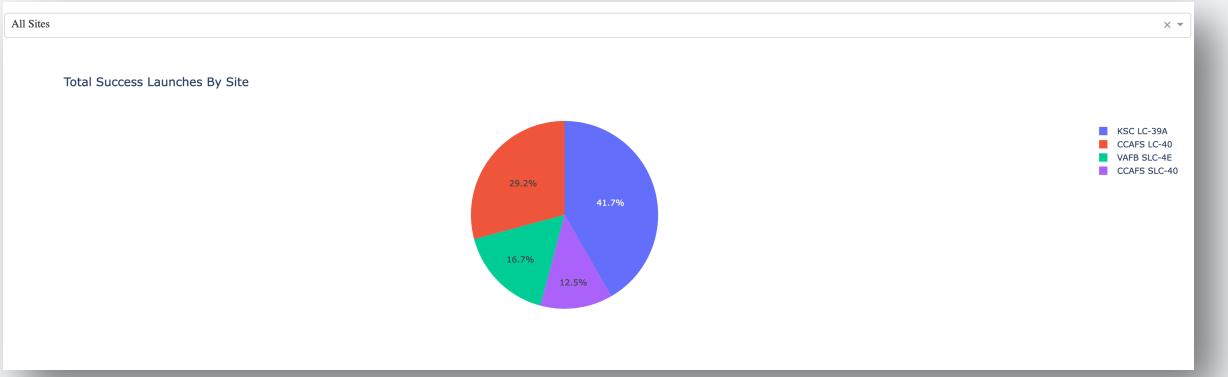


Section 4

Build a Dashboard with Plotly Dash

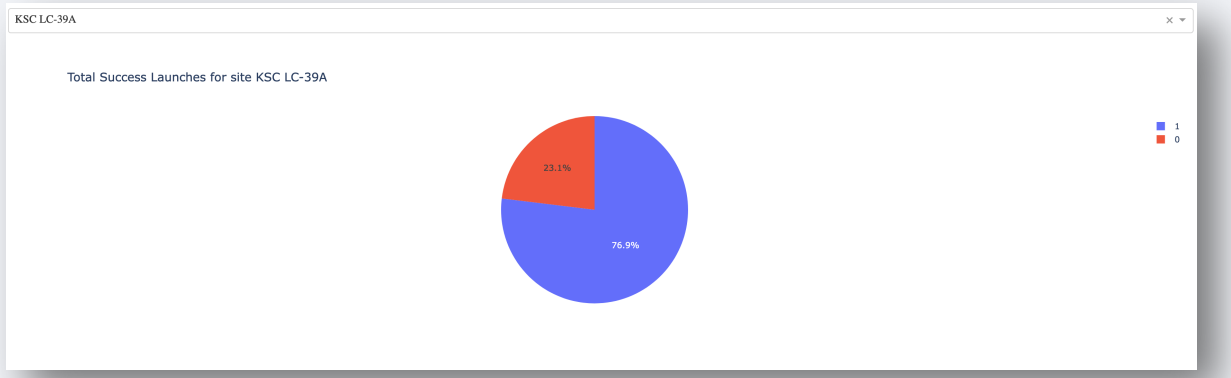
Share of Successful Launches by Site

- The pie chart shows the proportion of successful launches for every launch site.
- The launch site ‘KSC LC-39A’ has the highest proportion of successful launches with almost half of the successful launches being from this site.
- ‘CCAFS LC-40’ has the next highest successful launches with over a quarter launches followed by the sites ‘VAFB SLC-4E’ and ‘CCAFS SLC-40’.



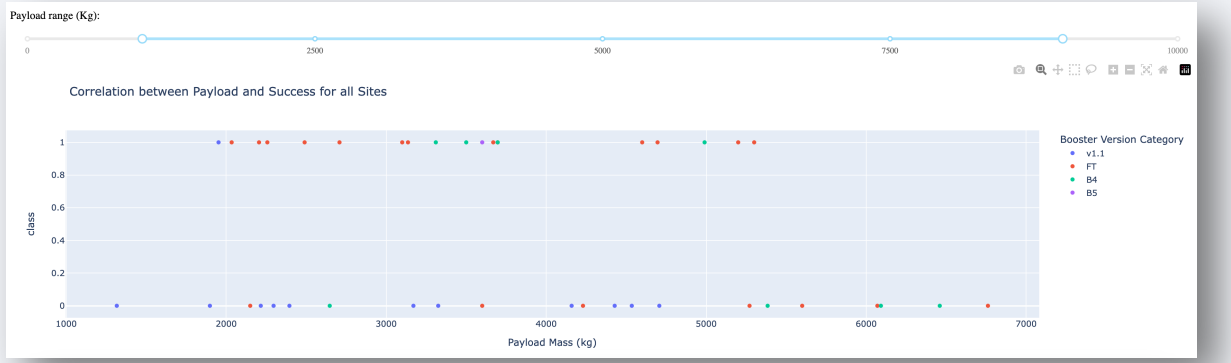
Successful and Failed Launches for a single site

- The pie chart shows the proportion of successful launches for a single launch site.
- This chart shows the proportion of successful and failed launches for the site ‘KSC LC-39A’.
- Over three quarters of the launches for this site were successful.



Correlation between payload mass & success

- The scatter plot shows the correlation between the success rate and payload mass.
- The colored points depict the booster version category.
- We observe that for the range of 1000 kg to 7000 kg of payload mass, lower payload mass, particularly 2000 to 4000 kg, tend to have a higher success rate.
- Booster version category FT was the most successful, while the v1.1 booster version was the least.

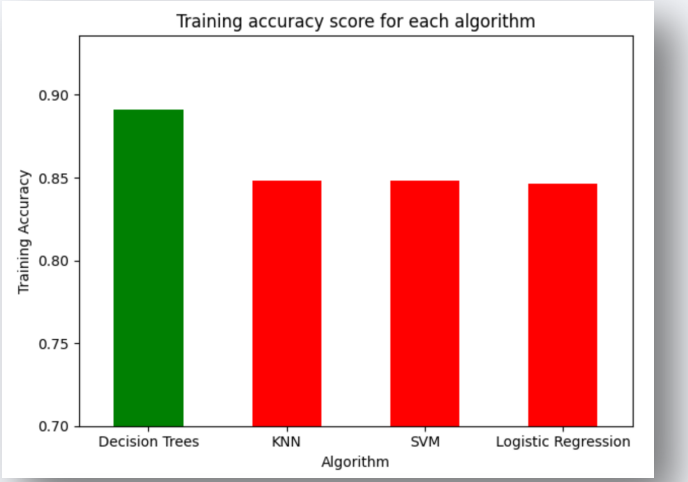


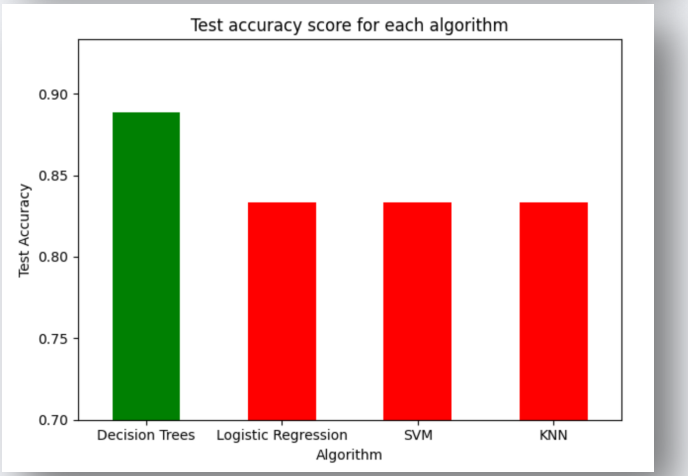
Section 5

Predictive Analysis (Classification)

Classification Accuracy

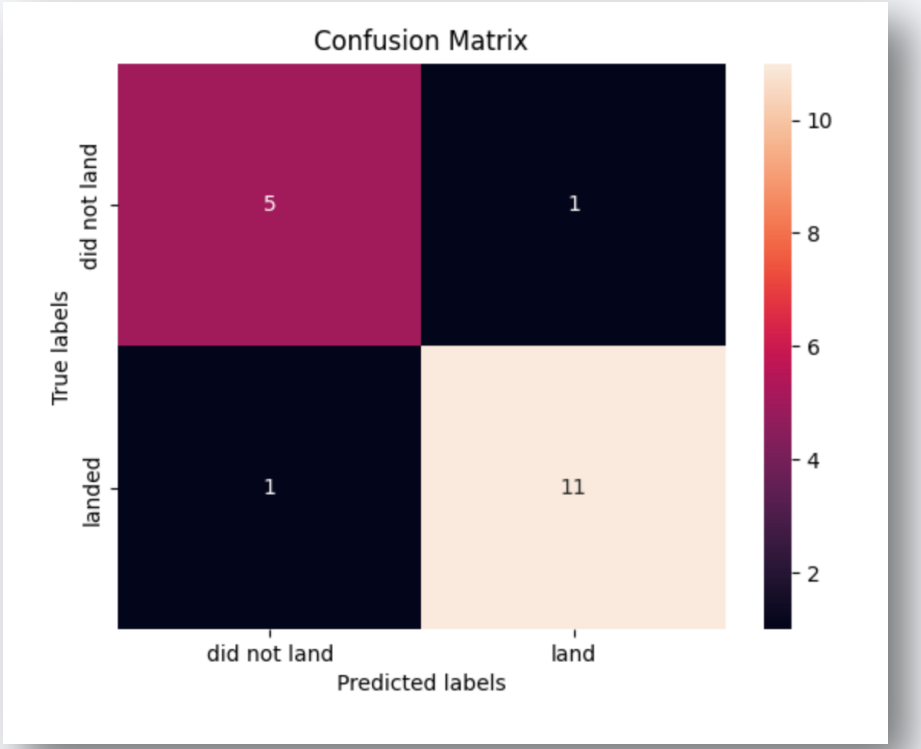
- Bar charts visualizing the training and testing accuracy score for all the models are shown in the figures.
- They demonstrate that the Decision Tree classifier had the highest accuracy (almost 90%).
- Other models viz., logistic regression, SVM, and KNN demonstrated similar and equivalent performance.





Confusion Matrix

- The confusion matrix for the best performing model (Decision Trees) is shown in the figure.
- The matrix compares the number of labels predicted by the model with the actual labels.
- We observe that the model correctly predicted 5 instances where the landing was a failure and 11 instances where it was a success.
- However, the model also made two incorrect predictions, 1 false-positive and 1 false-negative.



Conclusions

- ‘CCAFS SLC 40’ launch site had the largest number of launches.
- In general, higher flight numbers are more successfully launched.
- The orbit types ES-L1, GEO, HEO, and SSO have near perfect success rate.
- We observe that there is a general trend of increasing success in launches with time from 2013 to 2020.
- Falcon 9 has been launched only from four launch sites; one in California and three in Florida.
- Falcon 9’s first successful ground pad launch was in late 2015, over five years from its initial launch.
- Most mission outcomes were regarded as successful.
- Launch sites are close to the coastline and have railways built near them, but they are away from other proximities like cities and highways.
- Booster version FT was the most successful, while the v1.1 booster version was the least.
- The launch site ‘KSC LC-39A’ has the highest proportion of successful launches.
- The Decision Tree classifier is the best suited to predict if a launch would be a success or not.

Appendix

- SpaceX API URL: https://api.spacexdata.com/v4/launches/past
- Falcon 9 Wiki URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches
- GitHub Project URL: https://github.com/RayhaanPirani/spacex-falcon9-landing/
- Tools used: IBM Skills Network Lab, IBM Watson, Python, SQLite, Folium, Plotly Dash, scikit-learn

Thank you!