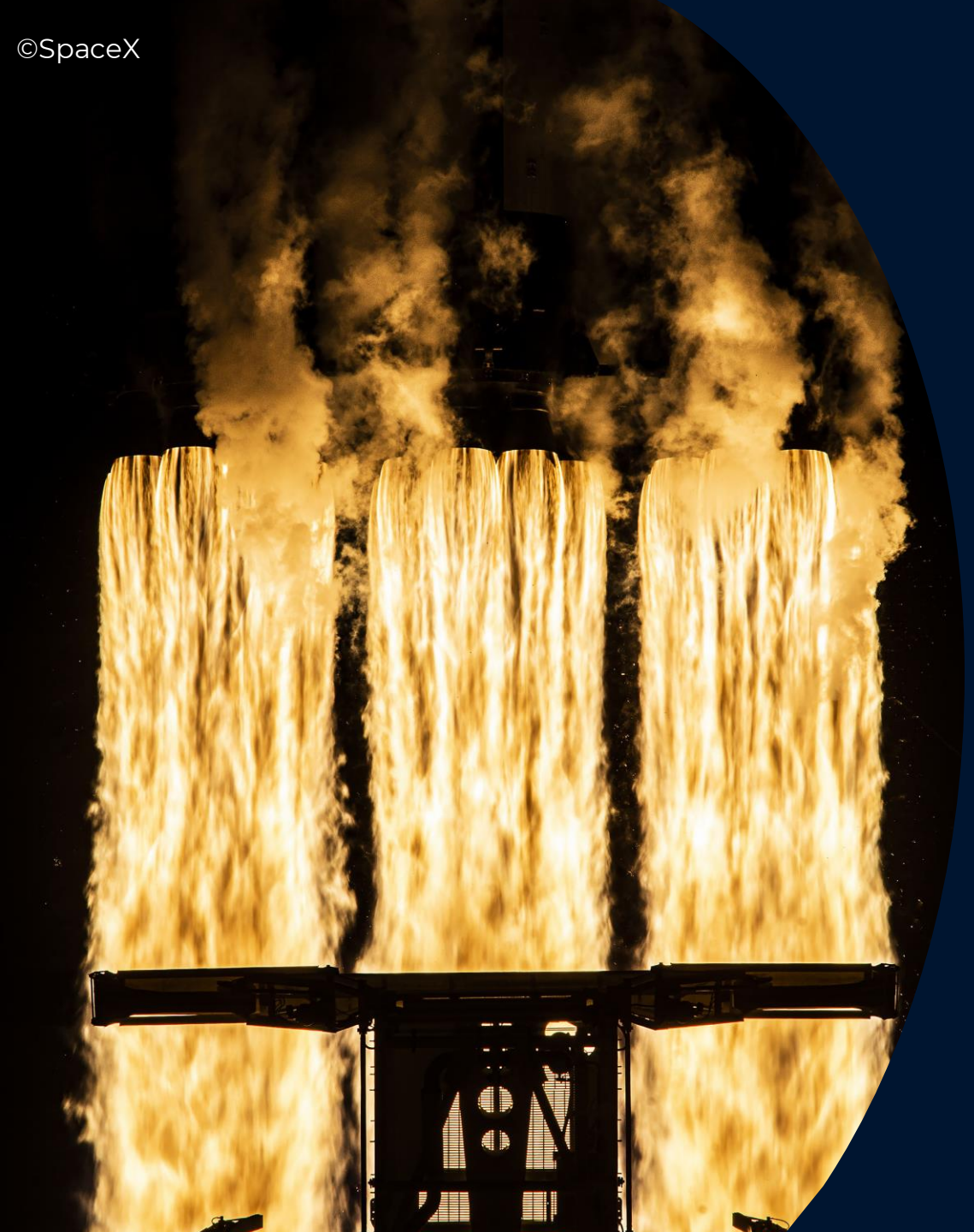
Winning the Space Race with Data Science

Alexander Price
04/10/2023

©SpaceX

Outline

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

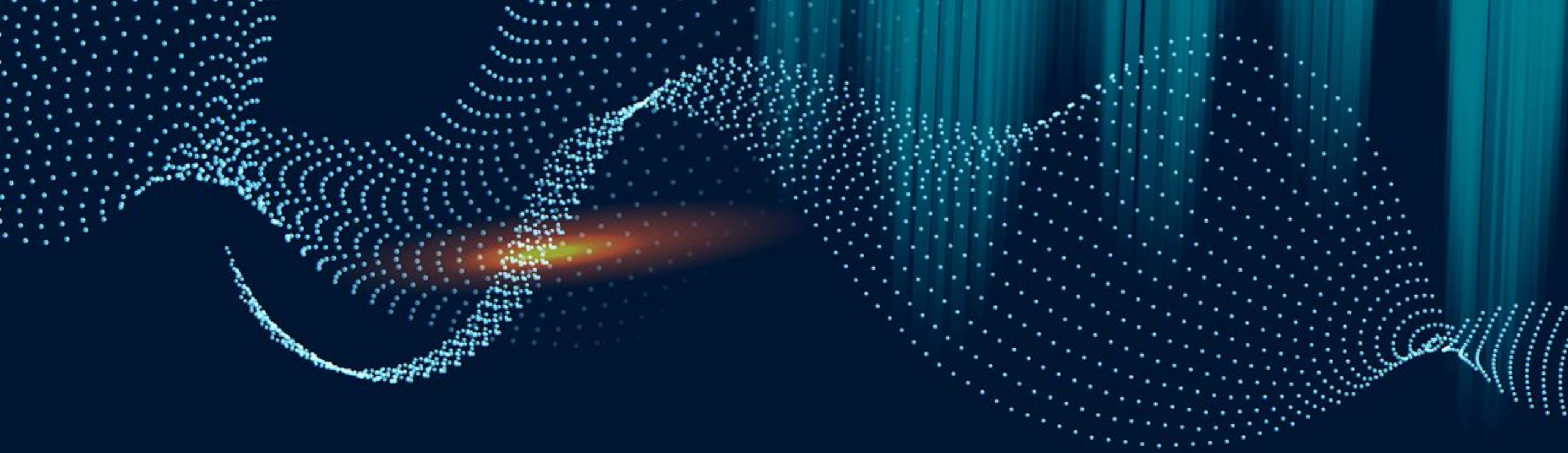


Executive Summary

- Methodologies Used
  - **Data Collecting** from SpaceX Rest API and Web Scraping Wikipedia Table
  - **Data Wrangling** using Exploratory Data Analysis(EDA) to filter and transform the data
  - **Data Analysis** to understand and explore the data using SQL and EDA
  - **Data Visualization** to visualize the data using Folium and Plotly (Dash)
  - **Machine Learning** to model and predict outcomes from the data using classification modeling
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive visualization results using Dash
  - Machine Learning Prediction Result

Introduction

- SpaceX has rapidly emerged as one of the leading space transportation companies in the world, revolutionizing the way we access space. The Falcon 9 rocket is the workhorse of SpaceX's fleet, capable of delivering payloads to low Earth orbit, geostationary transfer orbit, and even beyond. Falcon 9 rocket is one of the first rockets to use boosters that can recover and land themselves. In this report we are trying to predict if a Falcon 9 rocket will land successfully. With each Falcon 9 rocket costing ~75 million dollars needing to know that the rocket launch will be successful would save money in the long run.
- **Problems that need to be answered**
  - Can the data can show what influences, variables, or conditions need to be for a successful or failure of a landing?
  - What can be done to ensure a successful landing in the future?



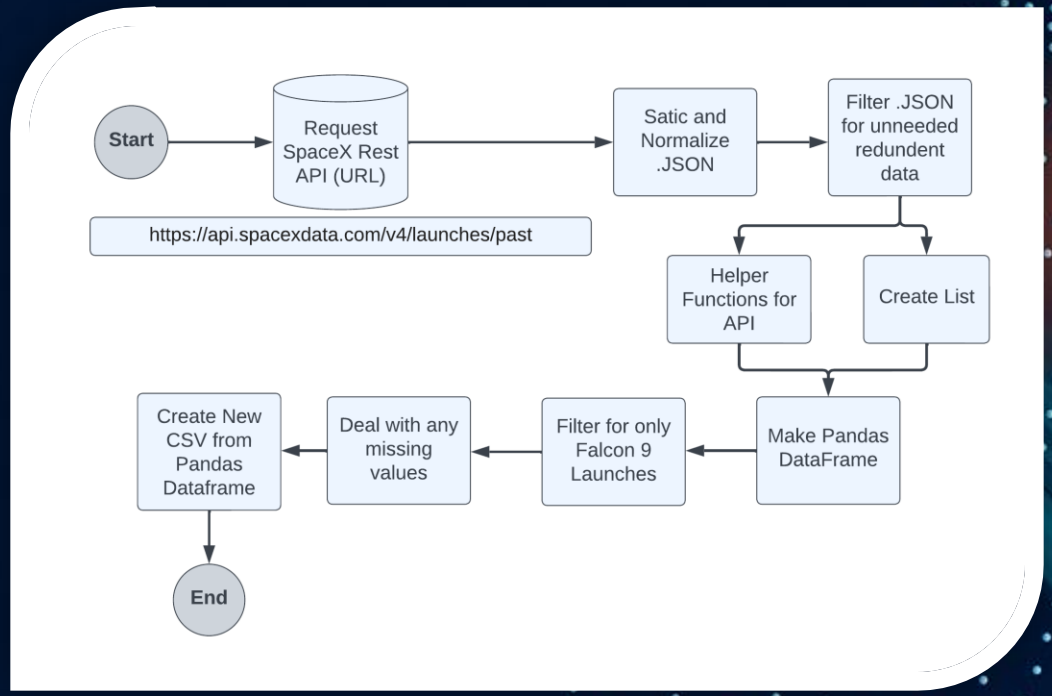
01 Methodolgy

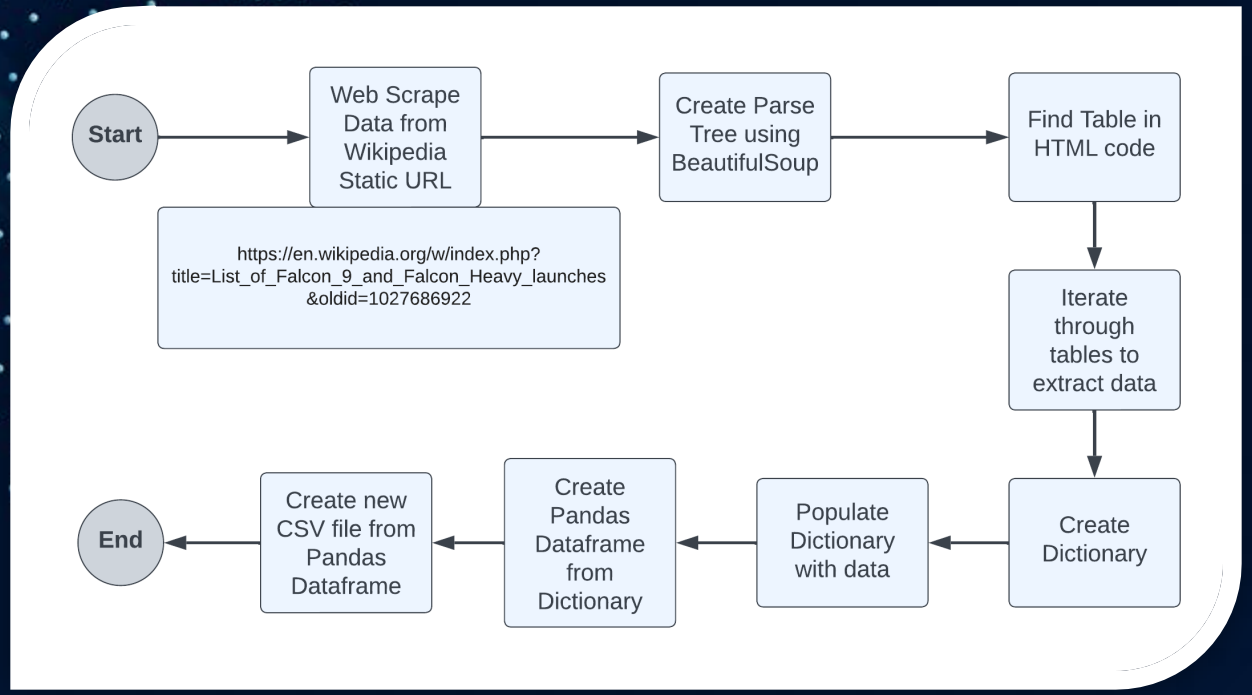
Methodology

- Data collection methodology:
  - **Data Collected from SpaceX Rest API**
  - **Data Web Scraping Space X Chart from Wikipedia**
- Perform data wrangling
  - **Processed using exploratory data analysis (EDA) to transform the data for the machine learning later in the process**
- Perform exploratory data analysis (EDA) using visualization and SQL
  - **Visuals include scatter plots, bar graphs and other to show patterns in the data**
- Perform interactive visual analytics using Folium and Plotly Dash
  - **Visuals include geographic information system(GIS) maps made using Folium**
- Perform predictive analysis using classification models
  - **How to build, tune, and evaluate classification models**

Data Collection

- The two ways that data was collected was through:
  - SpaceX's **Rest API**
  - **Web Scraping** data from a Wikipedia Table on SpaceX rocket launches

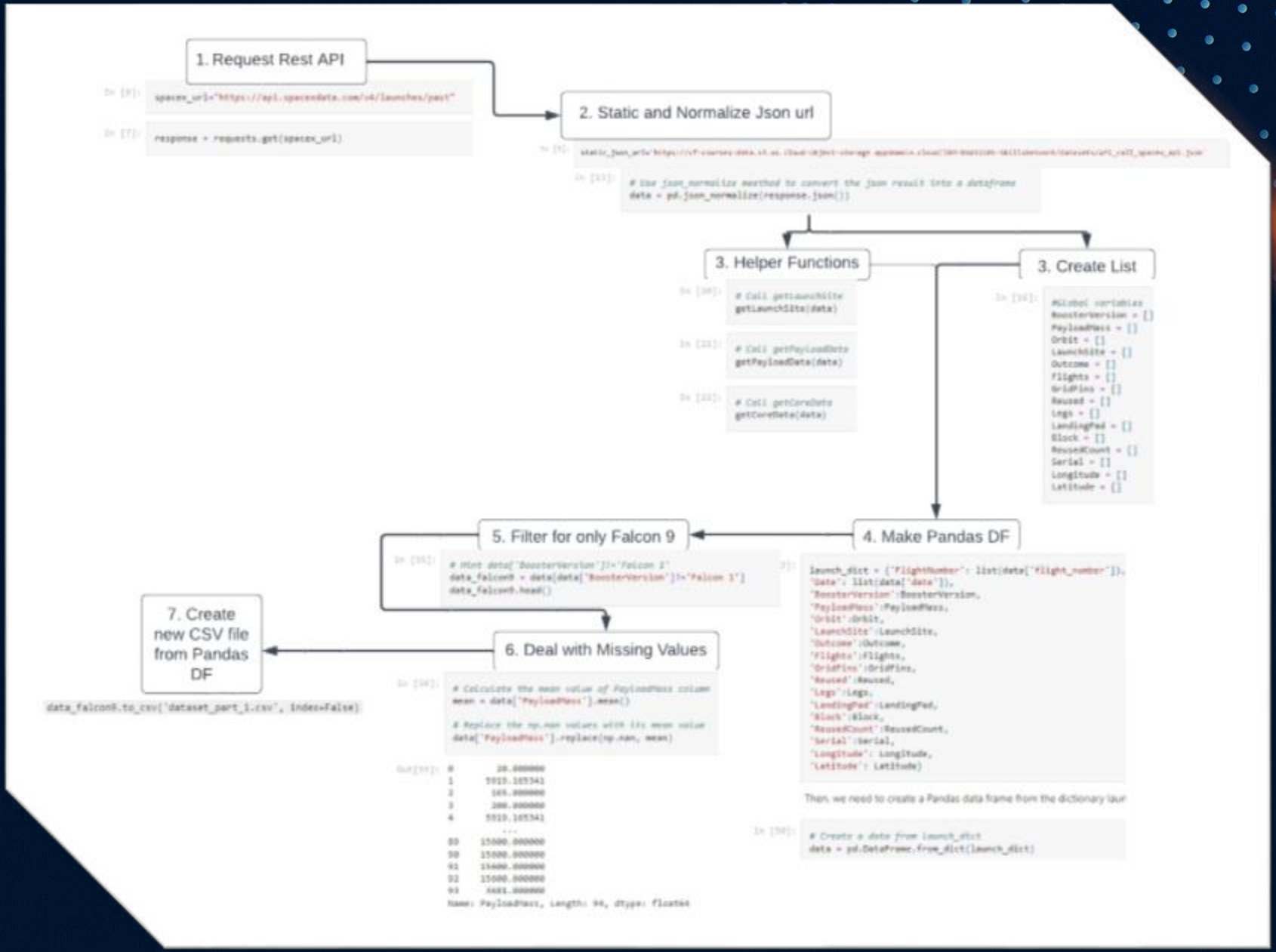




Data Collection – SpaceX API

- This is the process of getting the Rest API from SpaceX and making it into a useable CSV file for exploratory analysis and machine learning later on

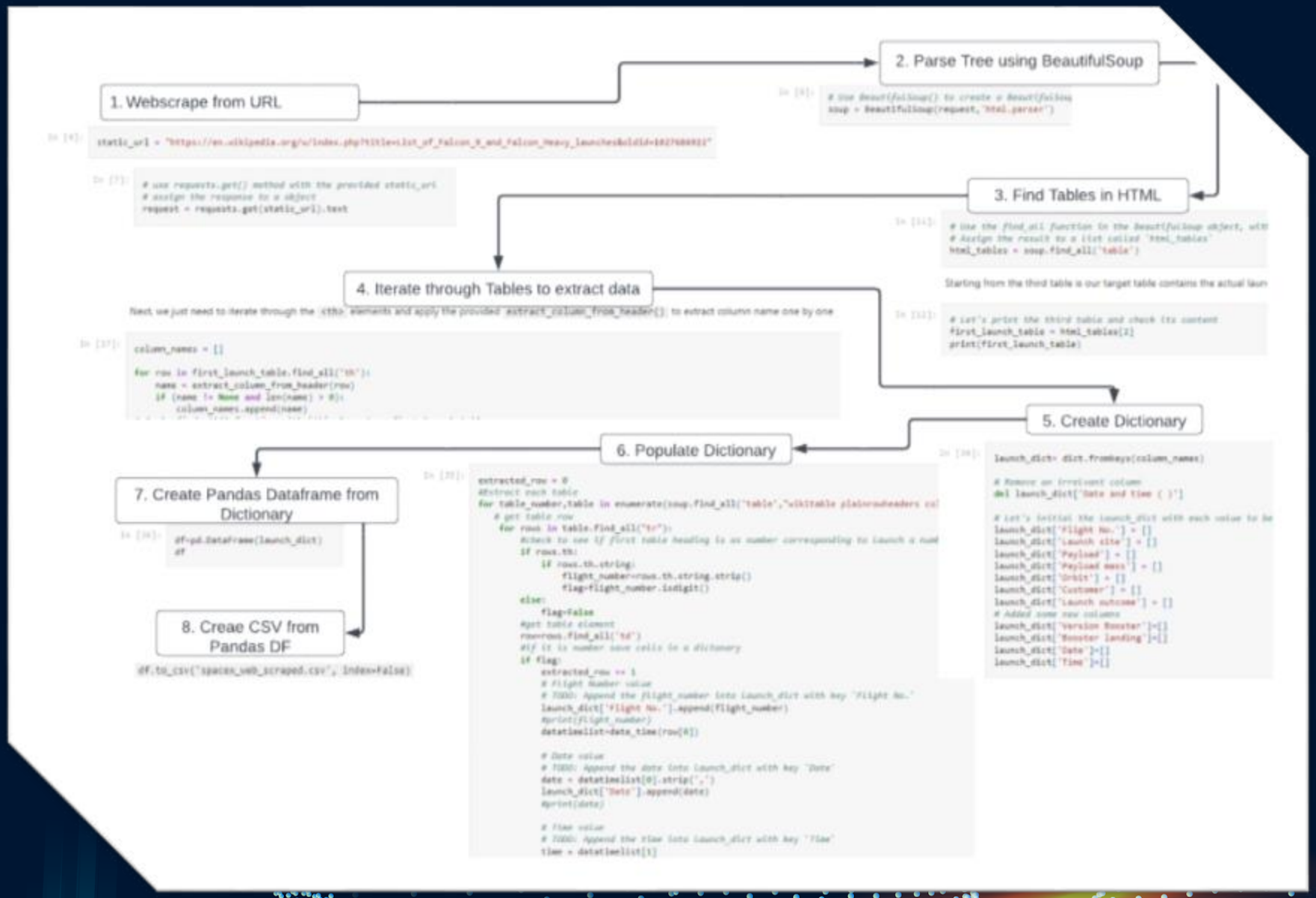
- GitHub Link to Notebook



Data Collection – Web Scrapping

- Web scraping from a Wikipedia page finding the Table in HTML and then turning it into a Pandas Data Frame then make it into a CSV file

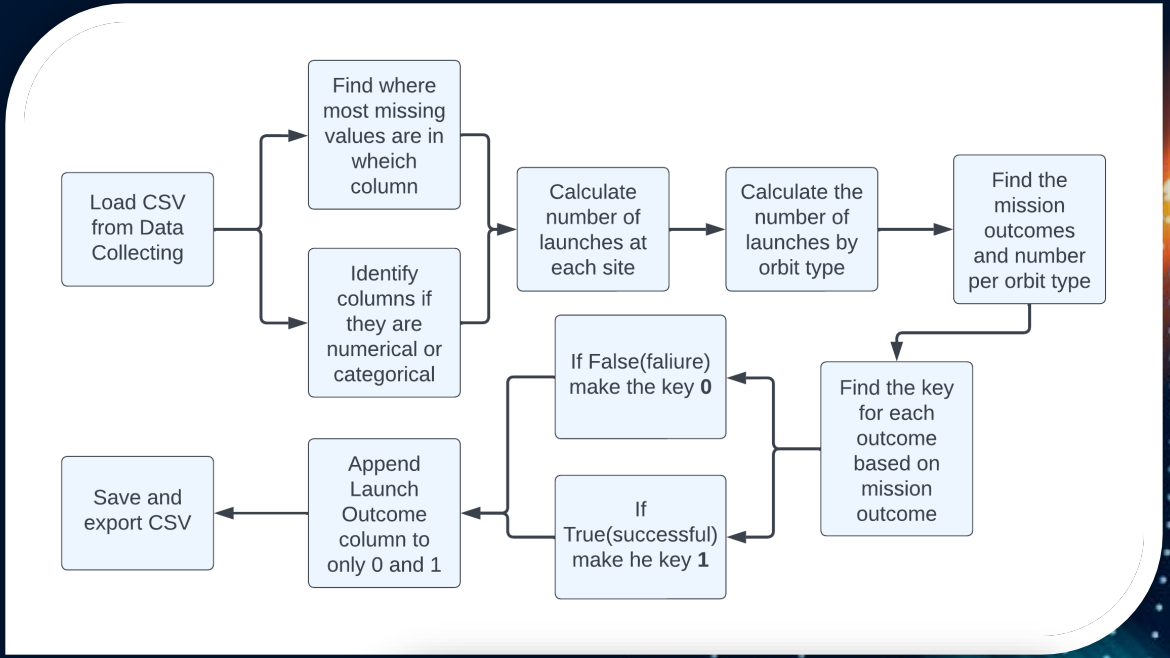
GitHub Link to Notebook

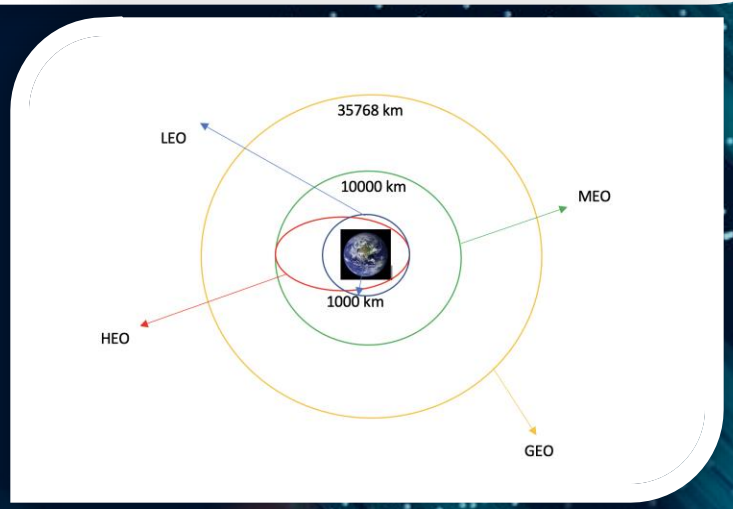


Data Wrangling

- For data wrangling the data was process using Exploratory Data Analysis (EDA). In this data set there were many cases of the booster not landing successfully. Sometimes a landing was attempted but failed due to an accident. For example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad. False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. In this lab we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

- GitHub Link to Notebook





EDA with Data Visualization

- For this Notebook we used Exploratory Data Analysis to make the following visualizations:
  - Scatter Plots – these plots help to see every individual landing to see what or if variables have an influence on success.
    - Flight Number vs Payload Mass (kg)
    - Flight Number vs Launch Site
    - Payload Mass (kg) vs Launch Site
    - Flight Number vs Orbit Type

- Bar Chart – this bar chart is to see if the type of orbit has an influence on landing and which orbit has the best and worst success rates.
  - Success rate vs Orbit Type

- Line Chart – line chart helps us to see if a trend has been happening between to variables in a good trend or a failing trend.
  - Success Rate vs Year

GitHub Link to Notebook

EDA with SQL

- For this Exploratory Data Analysis, we used SQL queries to gather information and understand the dataset.
  - This was loaded onto a Db2 Database on IBM to execute SQL queries on the notebook
- 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. This dataset includes a record for each payload carried during a SpaceX mission into outer space.
- Below is queries searched in the data:
  - Find and display all unique launch sites
  - Find how many launches came from launch sites starting with 'CCA'
  - Find the total payload mass carried by the booster launched by NASA (CRS)
  - Find the average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome on ground pad was achieved
  - List the names of the boosters which have success on the ocean drone ship with a payload mass >4000 kg and <6000 kg
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried maximum payload mass
  - List the failed landing outcomes on ocean drone ships, the booster version, and launch site for the year 2015
  - Then rank the landing outcomes and where the landing was through the dates of 2010-06-04 and 2017-03-20 in descending order

GitHub Link to Notebook

Build an Interactive Map with Folium

- This notebook is visualizing points of interests on a interactive GIS map using folium. The purpose of the maps is to visualize where the launch sites are at and if any factors to contribute to the trajectory and success or failure of launches. The steps for the maps are as follows:
    - Mark all launch sites on map to see if geographic factors contribute to success
    - Mark successful and failed launches for each site
        - Showing all launches that was successful or a failure to see if that a particular site has contributing factors
        - Using data from launch outcomes using 0 and 1 showing up as red(0) and green(1) markers on map
    - Calculate distances between a launch site and its proximities if any proximities to a specific location are contributing factors

- GitHub via nbviewer link to Notebook

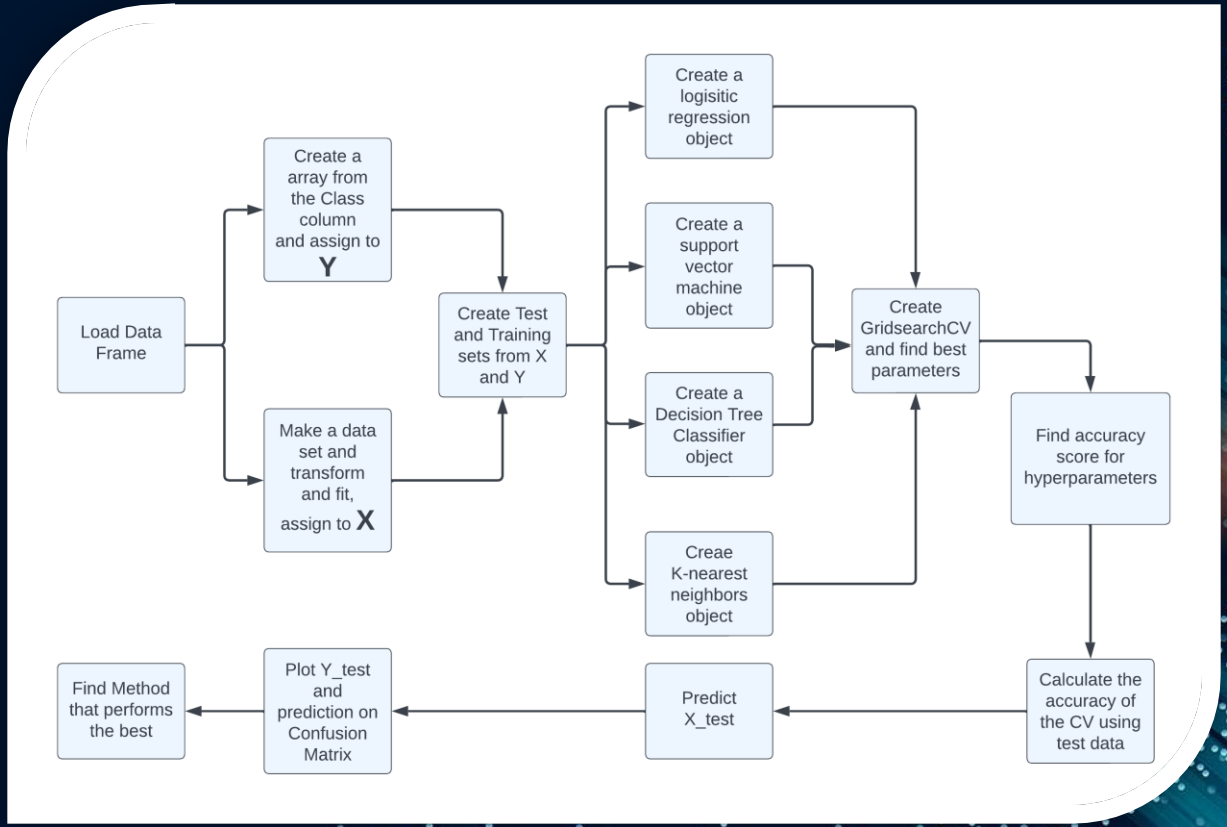
Build a Dashboard with Plotly Dash

- The interactive Dashboard application built in Python using HTML, Pandas, Plotly, and Dash. There are two interactive visualizations used:
  - **Pie Graph**
    - Interactivity includes a dropdown menu to either select individual launch site or seeing all site to see which site has the best success in launches. The pie graph is used because since there are only 4 individual launch sites it shows immediately which site had the most success.
  - **Scatter Plot**
    - The scatter plot is shows Payload Mass (kg) vs if launch was successful. The interactivity is able to select what range the mass is to see the individual points to see if it was successful and if the Payload Mass had any influence on success.

GitHub Link to Code

Predictive Analysis (Classification)

- We are trying to predict if the first stage will land and be reusable to save money over the long run.
- To find the best classification model first need to have a X and Y to create train and test data. Use the training method on the following classification methods:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K-Nearest Neighbor
- Using these methods to then compare the trained data to the test data to find the accuracy and then plot a confusion matrix to visualize the prediction.



GitHub Link to Notebook

Results

- Exploratory data analysis results
- EDA with SQL results
- EDA with Folium Maps
- Interactive analytics demo in screenshots
- Predictive analysis results

©SpaceX

Section 2

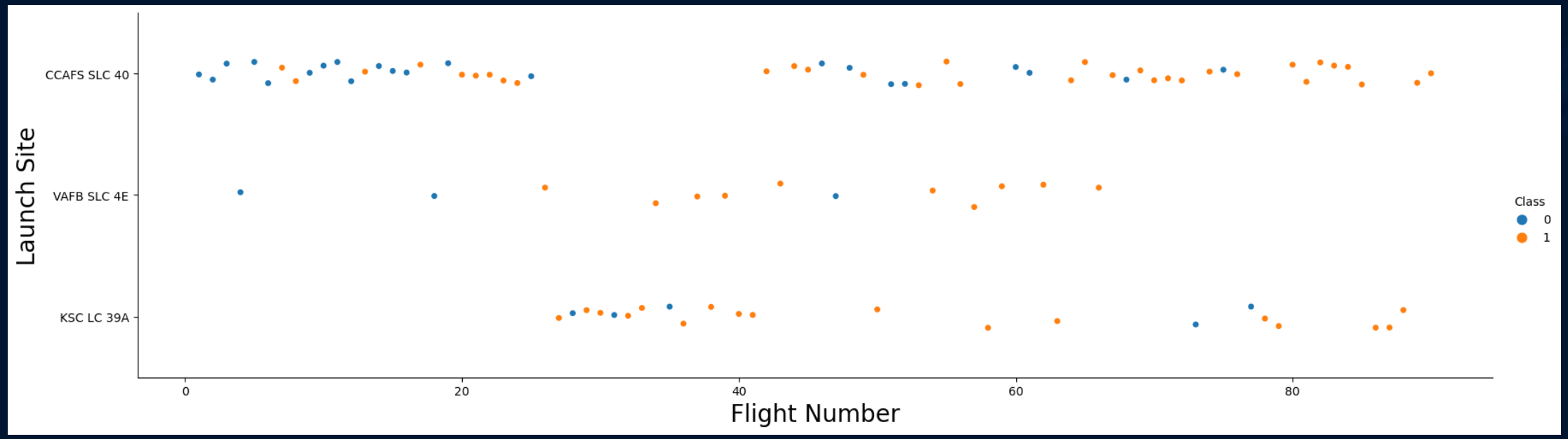
EDA with Visualization

Insights drawn from EDA

©SpaceX

Flight Number vs. Launch Site

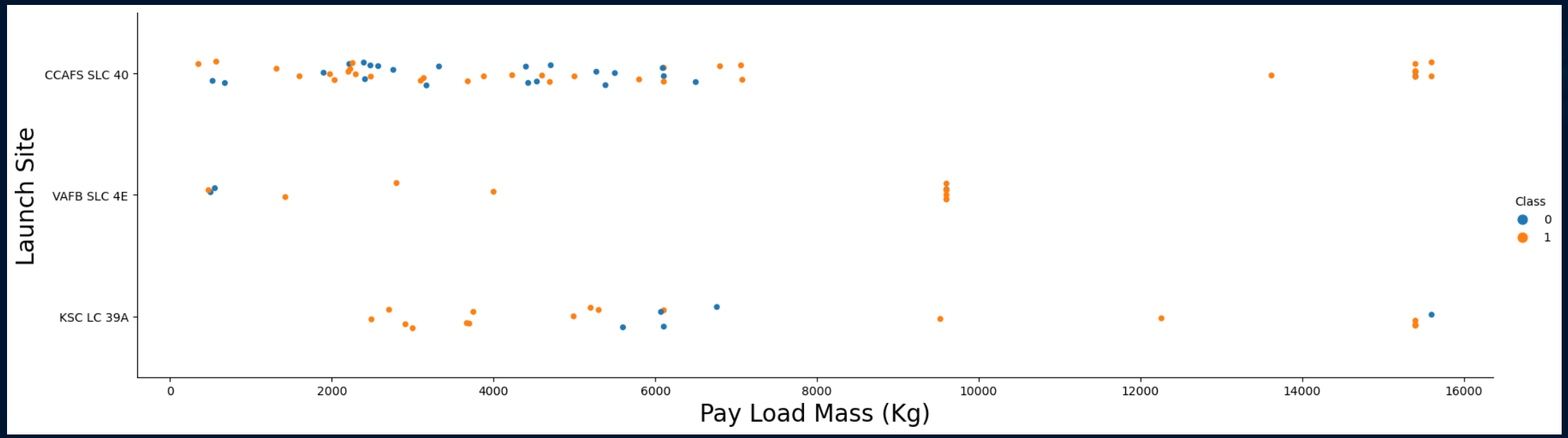
- In this scatter plot, it is showing if the launch site has any variable in the outcome of a launch.
- We can see in the scatter plot KSC LC 39A has a better success rate than the others
- As flight number increases, the chances of success increases because SpaceX is learning from mistakes to have a better success rate on future launches



Blue marker is failure(0) and Orange is Success(1)

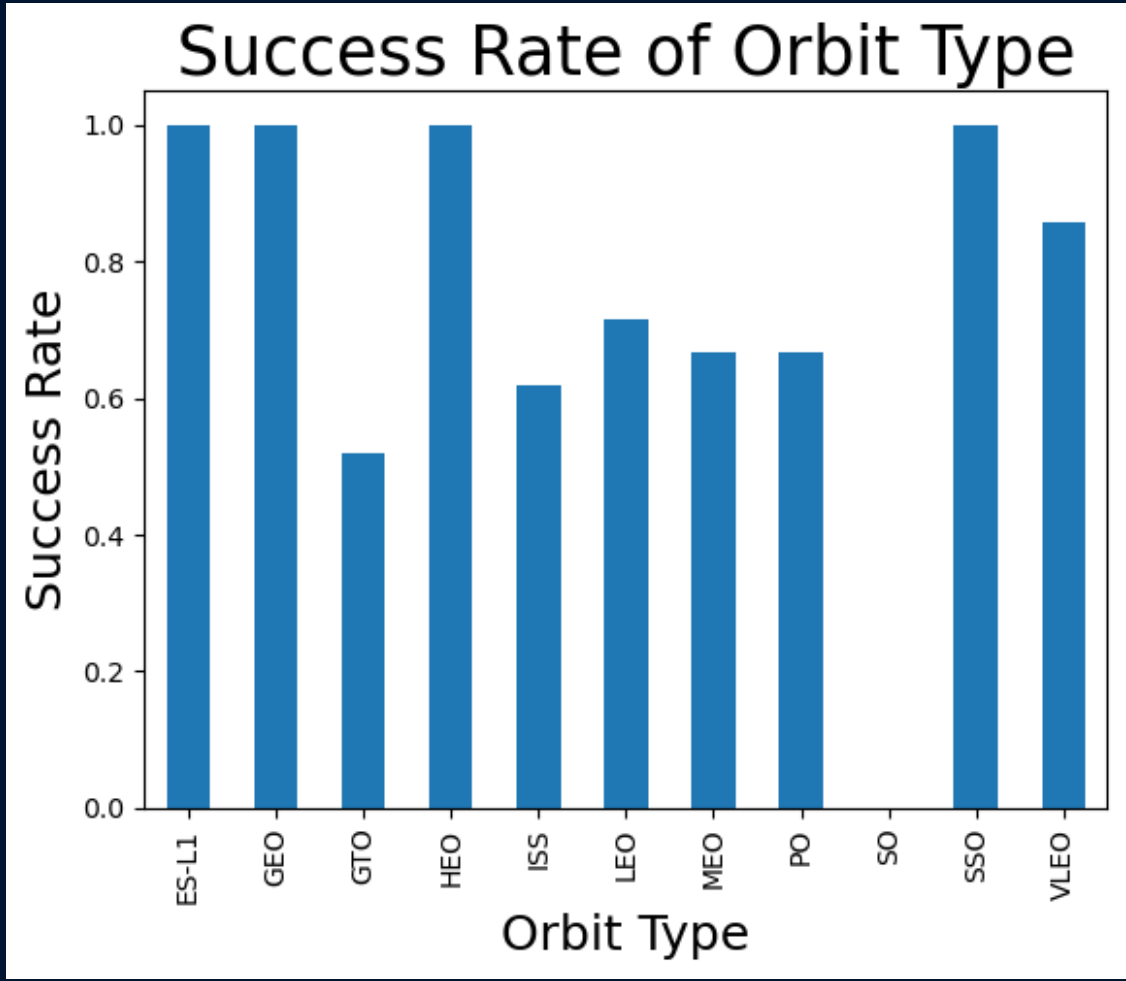
Payload vs. Launch Site

- For this scatter plot we are trying to see if payload mass has a variable at each launch site contributing to a successful outcome.
- In the case the for CCAFS SLC 40 has a great success rate for high mass loads but only between 7000 and 16000 kg.
- Then for launch site KSC LC 39A has a great success rate across the board but seems to have a hard time with payload mass around 6000 kg.
- The results don't have a clear pattern between all the sites to make a prediction that payload mass at a launch site has any effect on the outcome for the launch.



Blue marker is failure(0) and Orange is Success(1)

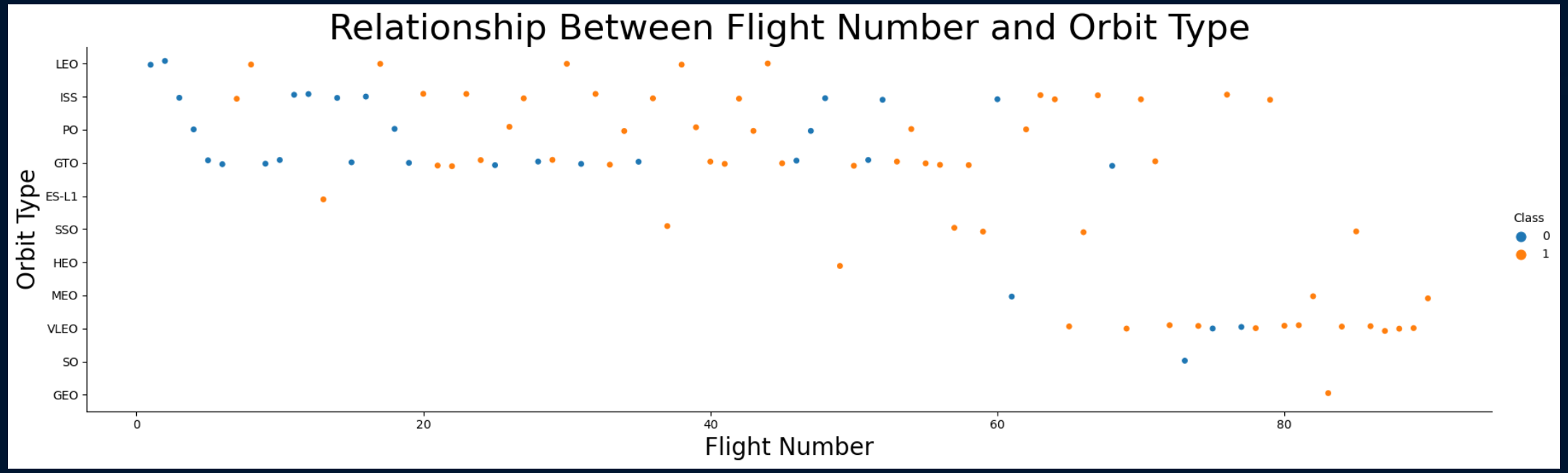
Success Rate vs Orbit Type



- This a bar chart showing the success rate by what type of orbit the launch was going on and to see if orbit type has an influence in launch outcome.
- The chart shows that Orbits ES-L1, GEO, HEO, and SSO have the best success rate
- Orbit types GTO has a 50/50 chance of success.
- Orbit type SO has zero success rate making it the worst success rate used.
- The chart shows there is no clear pattern if orbit type influences the success rate so it should not be used to make a prediction.

Flight Number vs Orbit Type

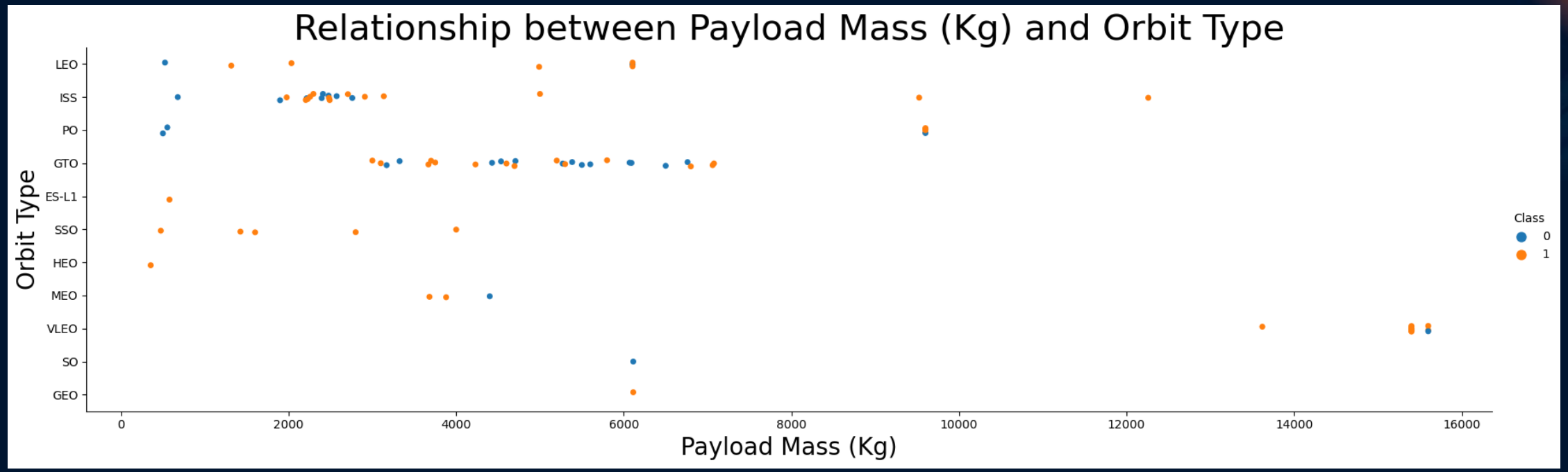
- This scatter plot shows if Flight Number and Orbit Type have an influence on the success rate.
- It shows as Flight Number increases more orbit types are being used and it also shows the increase in success rate.
- Looking at the plot, you can see one pattern for LEO orbit type and an increase in flight number shows a little influence. With others there is no influence with orbit type on flight number with the outcome of the launch.



Blue marker is failure(0) and Orange is Success(1)

Payload vs. Orbit Type

- This Scatter Plot shows if the payload mass has any influence with orbit type on successful launch.
- It seems that if the payload is heavier than 8000 kg there is a higher success rate in orbits ISS, PO, and VLEO
- For payloads less than 8000 kg Orbits LEO and SSO have a better chance at success



Blue marker is failure(0) and Orange is Success(1)

Launch Success Yearly Trend

- The line plot shows an average continued success rate of launches from 2013 to 2020.

Section 3

EDA with SQL

SQL Queries

©NASA

All Launch Site Names

- SQL Query
  - **%sql** select DISTINCT(LAUNCH_SITE) AS "Unique Launch Sites" from SPACEX
- This Query shows all the Launch Sites used in the Data Set and the exact names for them
- Florida
  - CCAFS LC-40
    - Cape Canaveral Launch Complex 40
  - CCAFS SLC-40
    - Cape Canaveral Space Launch Complex 40
  - KSC LC-39A
    - Kennedy Space Center Launch Complex 39A
- California
  - VAFB SLC-4E
    - Vandenberg Space Launch Complex 4E

Unique Launch Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

DATE	time_utc	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site Names Begin with 'CCA'

- SQL Query
  - %sql select * from SPACEX where LAUNCH_SITE like 'CCA%' limit 5;
- This query shows only launches that took place at CCA(Cape Canaveral) limiting 5 to search.
- The search includes date/time, what booster, payload carrying, payload mass, what orbit is chosen, the customer, mission outcome, and if landing was success/failure.

Photo: Theresa Cross / Spaceflight Insider

TOTAL PAYLOAD MASS (KG)
45596

Total Payload Mass

Launched by NASA

- SQL Query
  - %sql select SUM(PAYLOAD_MASS_KG_) AS "TOTAL PAYLOAD MASS (KG)" from SPACEX where CUSTOMER = 'NASA (CRS)'
- This query shows the total payload mass for all the launches payload mass (Kg) where the customer for the rocket was NASA
  - 45,596 Kg

AVG Payload MASS (KG)
2928

Average Payload Mass by F9 v1.1 Rocket

- SQL Query
  - %sql SELECT AVG(PAYLOAD_MASS_KG_) as "AVG Payload MASS (KG)" FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1'
- This query shows the average payload mass (kg) for Falcon 9 version 1.1 rockets.
  - 2928 Kg

First Successful Ground Landing Date

- SQL Query
  - %sql SELECT MIN(DATE) AS "First Successful Landing Outcome" FROM SPACEX WHERE LANDING_OUTCOME = 'Success (ground pad)'
- Below is the First Successful Landing Data for the Falcon 9 booster. This marked the first time in history that a booster landed itself instead of being tossed in space or the ocean.
- By landing, Space X can reuse boosters and save money for future missions.

First Successful Landing Outcome
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL Query
  - %%sql

```
SELECT BOOSTER_VERSION as "Booster Version",
PAYLOAD_MASS_KG_  as "Payload Mass(Kg)"

FROM SPACEX

WHERE LANDING_OUTCOME  = 'Success (drone ship)'

AND PAYLOAD_MASS_KG_  BETWEEN 4000 AND 6000;
```

- This query shows all the successful landings on the drone ship for the falcon 9 boosters to land on in the middle of the ocean. The search is limited to the rockets that were carrying a payload mass between 4000 kg and 6000 kg.

Booster Version	Payload Mass(Kg)
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

Mission Outcome	Total Number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- SQL Query
  - %sql SELECT MISSION_OUTCOME as "Mission Outcome", COUNT(MISSION_OUTCOME) AS "Total Number" FROM SPACEX GROUP BY MISSION_OUTCOME
- The table query displays the total number of mission outcome by success or failure by the given data .
- Notable things that are shown are that there is a great success rate with two exceptions; it that showed a complete failure and a successful mission, but payload information is unknown.
  - Failure (in flight)
  - Success (payload status unclear)

Boosters Carried Maximum Payload

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

- SQL Query with subquery
  - %%sql

```
SELECT DISTINCT BOOSTER_VERSION as "Booster Version"
FROM SPACEX
WHERE PAYLOAD_MASS_KG_  = (
SELECT MAX(PAYLOAD_MASS_KG_)
FROM SPACEX);
```

- This is a list of all the booster versions that carried the maximum payload (kg) when launching.

Failed Landing Outcomes for 2015

Landing on Drone Ship

- SQL Query
  - %%sql

```
SELECT LANDING_OUTCOME as "Landing
Outcome", BOOSTER_VERSION as "Booster Version",
LAUNCH_SITE as "Launch Site"
FROM SPACEX
WHERE LANDING_OUTCOME = 'Failure (drone ship)'
        AND YEAR(DATE) = 2015;
```

- For the year 2015 there was two failed landing outcomes when trying to land on a drone ship in the ocean. Both launches were from Cape Canaveral, FL.

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

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

- SQL Query
  - %%sql

```
SELECT LANDING_OUTCOME as "Landing Outcome", COUNT(LANDING_OUTCOME) as Total
FROM SPACEX
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY TOTAL DESC;
```

Landing Outcome	total
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

- Shown is the ranking of all the landing outcomes for the dates between 06/04/2010 and 03/20/2017
- The query shows that there were ten attempts to land made during the time frame
- The successful and failure is mixed with different reasons for other landing outcomes

Section 4

Launch Sites

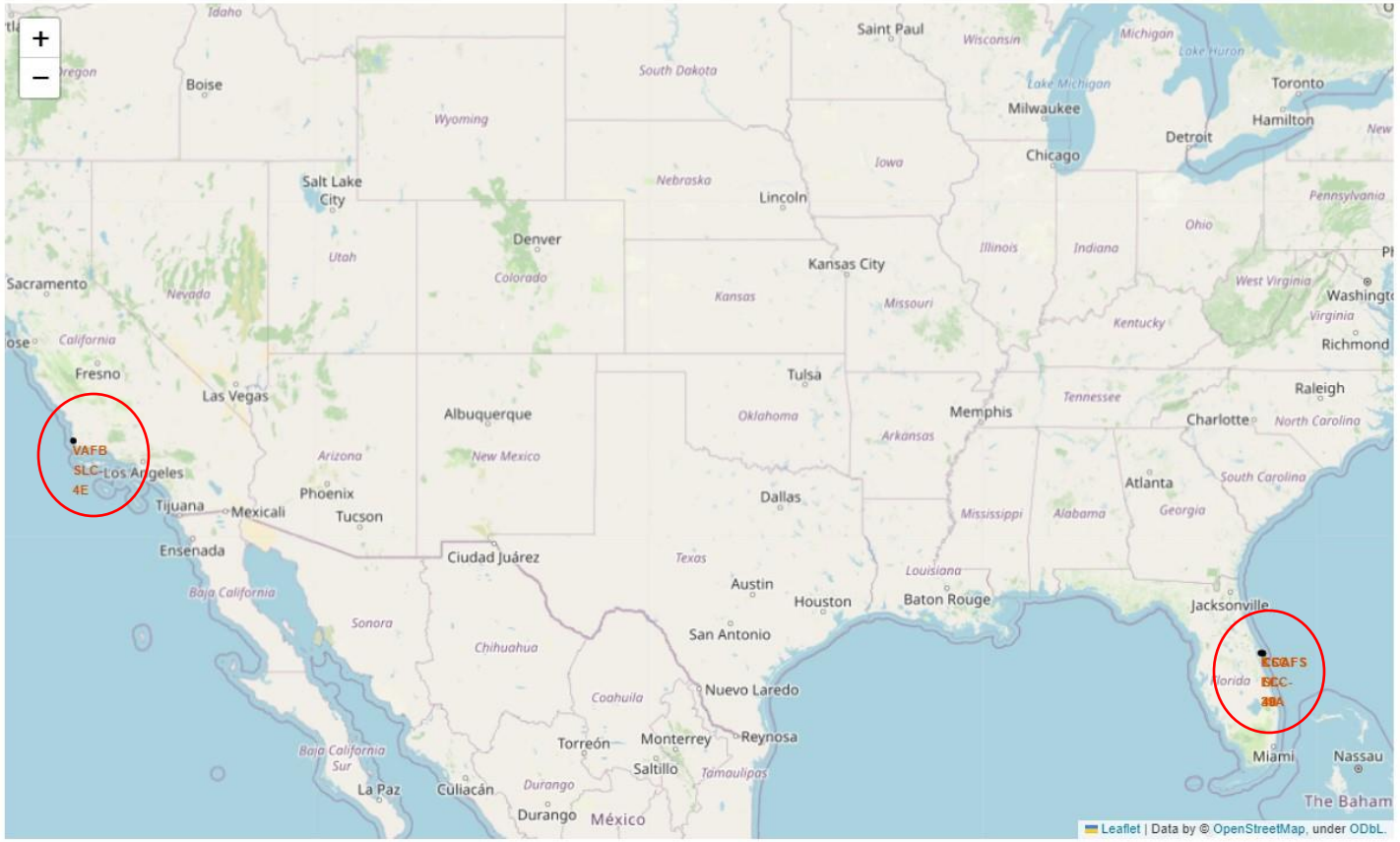
Proximities Analysis

©NASA

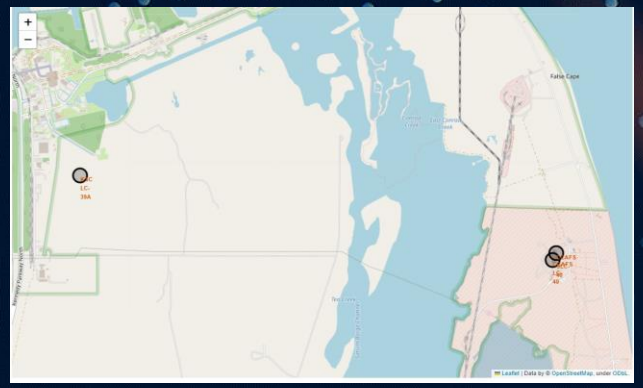
Locations of Launch Sites

- The United States map shows the locations of all the launch sites that SpaceX has used in the given data
- Three in Florida and one in California
  - All at launch sites used by NASA in the past
- All are near or on the coast

Out[8]:



FLORIDA

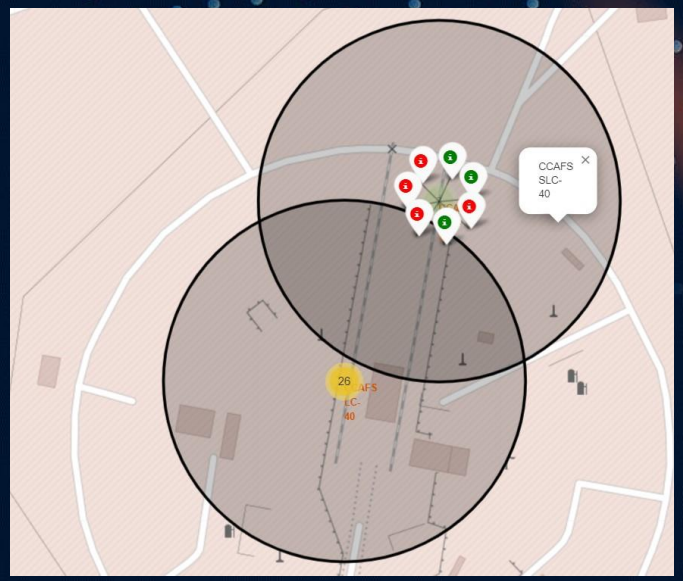


CALIFORNIA

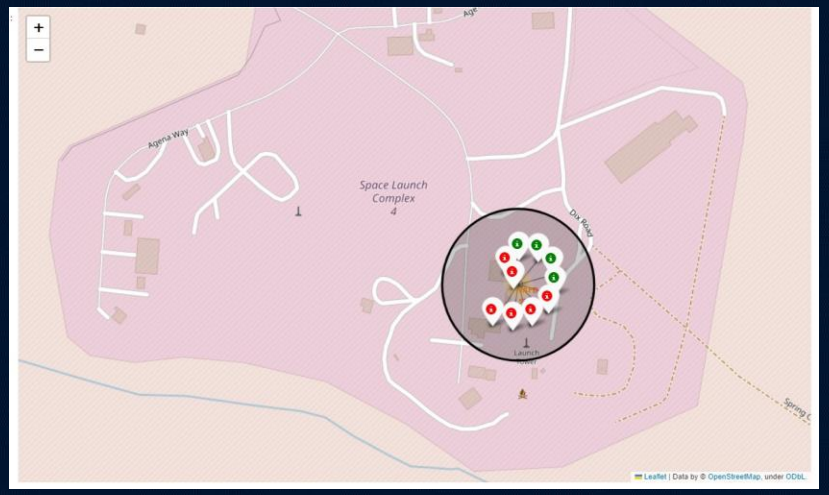
Success or Failure Markers at Launch Sites

- Shown are maps that show markers for each site, when clicked on, pop up with red showing failure and green showing success.
- For each launch site, you can see where more failures or more success in launches have happened visually per site.

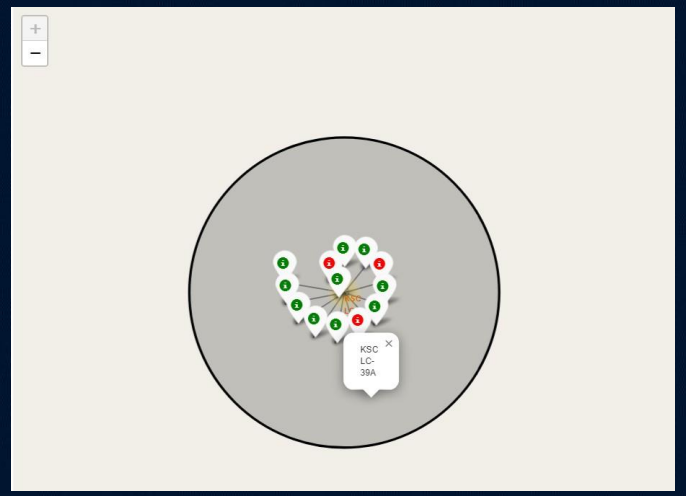
CCAFS SLC-40



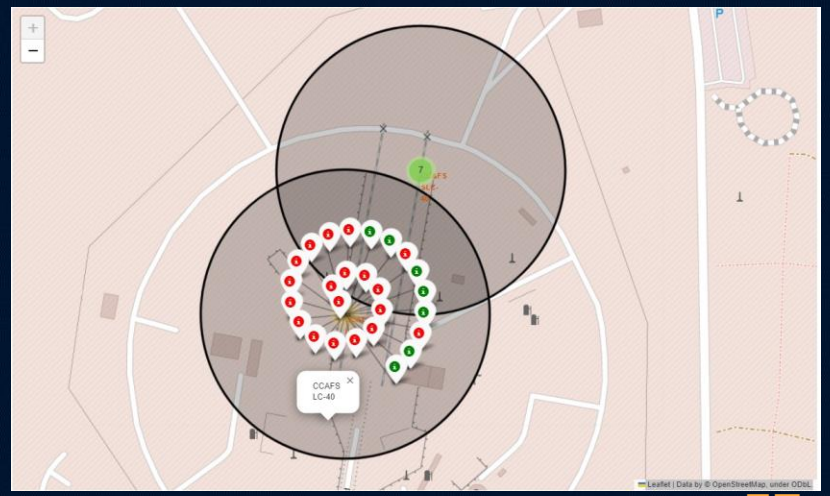
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40



Proximities to CCAFS SLC-40

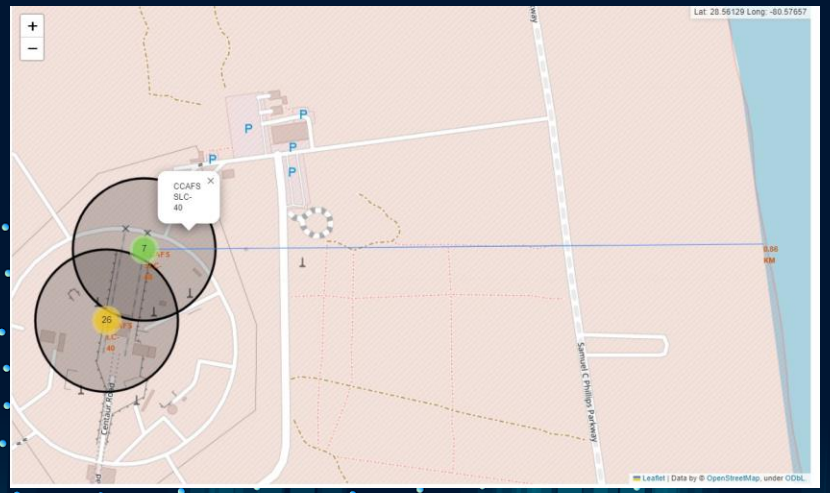
Distances in km

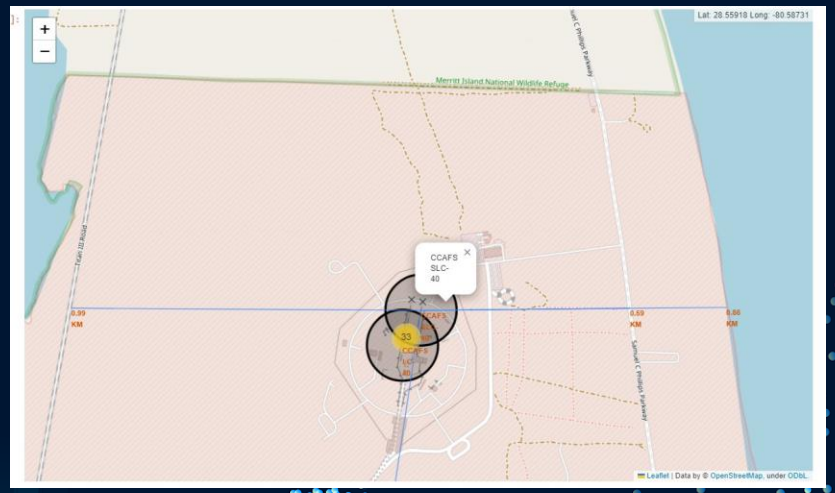
```
distance_highway = 0.5910563196201084  km
distance_railroad = 0.9896321330874676  km
distance_city = 18.13251322366318  km
```

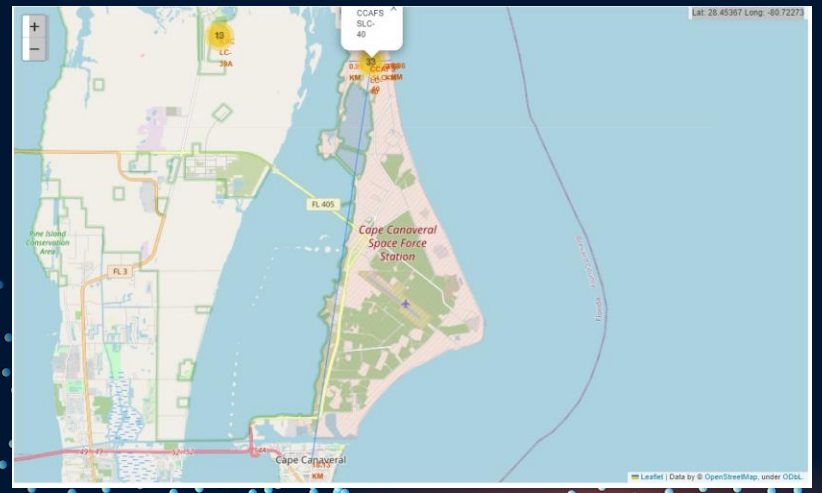
Distance to Coast

```
0.8646127425056958  km
```

- The maps shown illustrate how far the closest coastline, highway, railroad, and city for Launch Site CCAFS SLC-40 at Cape Canaveral in Florida.
- The calculations on the right side show the exact distance to the specified point
  - The highway, railroad, and the coastline are all equal to or less than a kilometer.
  - 1 km = 0.621371 mile







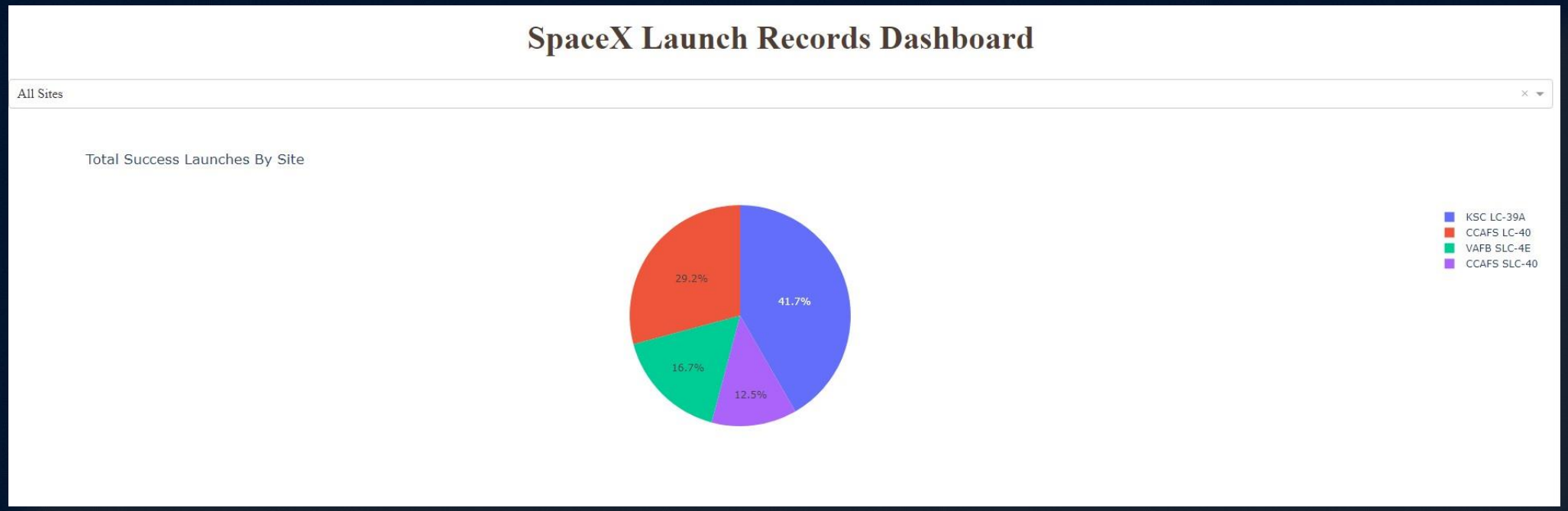
Section 5

Build a Dashboard with Plotly Dash

©SpaceX

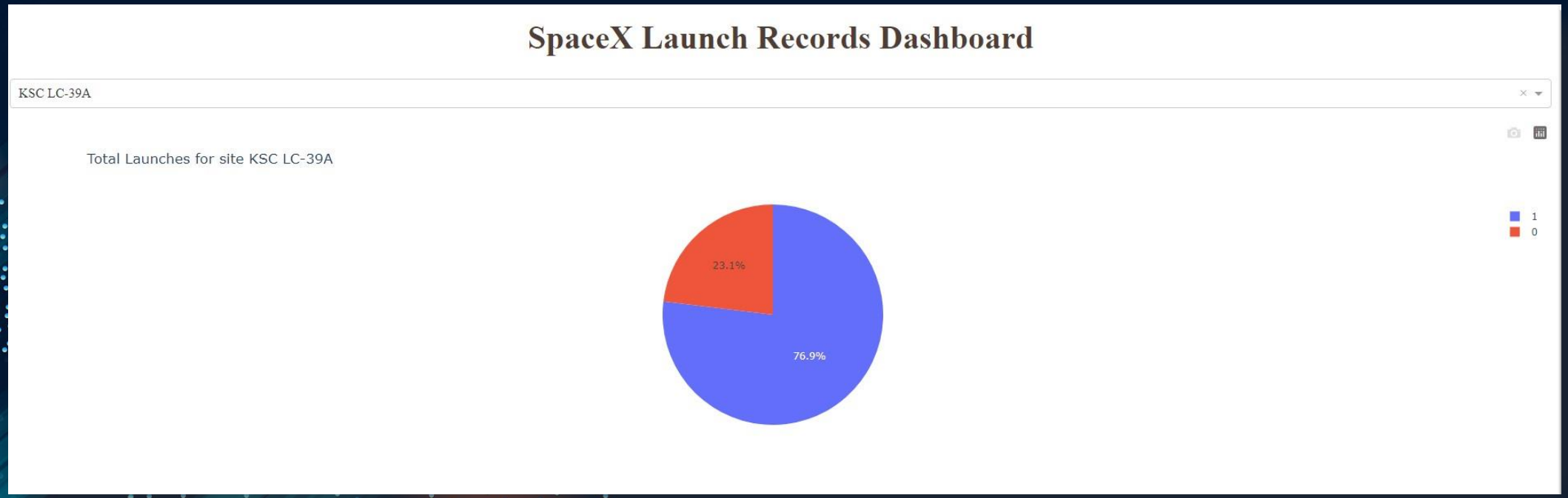
Total Success for all Launch Sites

- For this section, a Dashboard was built using Plotly Dash which is interactive.
- Below you can select between sites or all sites and see the Total Success ratio.
- The Pie Chart shows what site has the most success. It shows that KSC LC-39A is the most successful with 41.7% success rate overall.

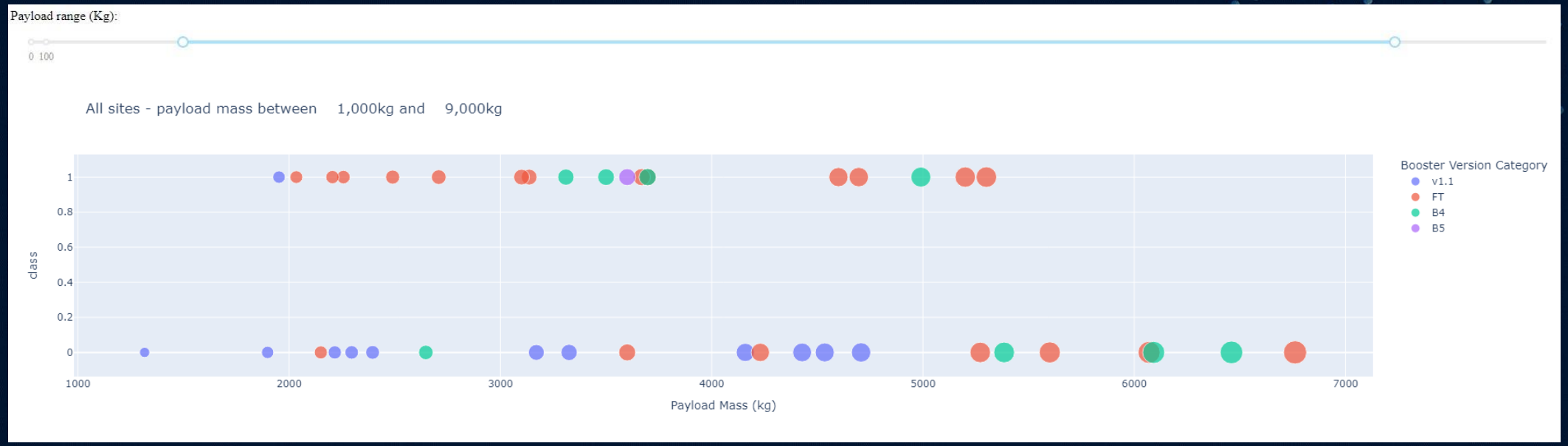


Launch Site with Highest Success Ratio

- Shown below is the most successful site is at Kennedy Space Center in Florida, LC-39A
- LC-39A has had a 76.9% success and a 23.1% failure from all the launches that was launched by SpaceX on this launch pad.
- Amazingly, this is also the launch site for Saturn V rockets and the Space Shuttle.



Success Rate by Payload Mass



- Below the pie graph on the Dashboard there is a scatter plot with a slider to select what Payload Range (Kg) you would want to look at.
- The Scatter Plot itself shows Success vs Payload Mass (Kg) with success on top of the plot and failure on bottom. With the colors showing what booster version it was.
- The plot shows a interesting pattern in that the Booster version FT has a higher success rate between 2000 kg and ~7000 kg. Also, Booster version B5 has a lower success rate, showing that it needs improvement or to be replaced based on data.
- Another pattern shows that lighter payloads have a higher chance of success than heavier.

Section 6

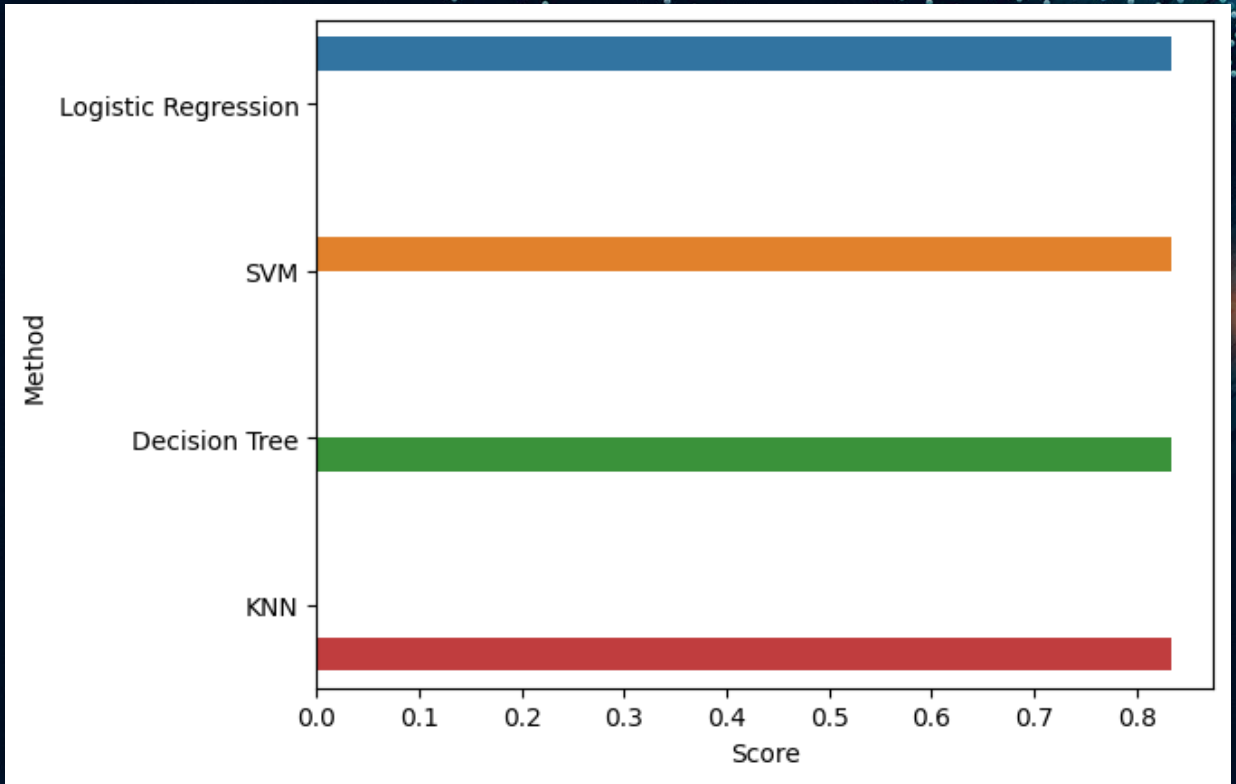
Predictive Analysis (Classification)

©SpaceX

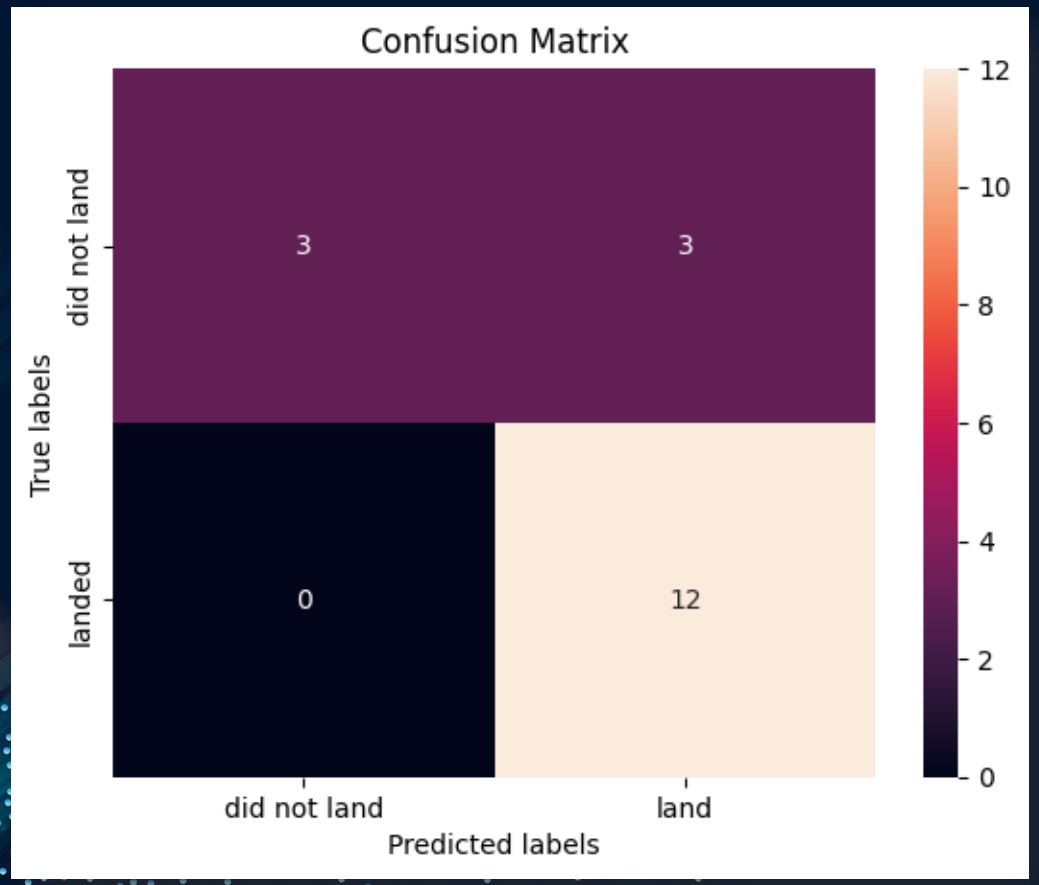
Classification Accuracy

- Using different training and testing data, but with a small data set, the graph and the scores are the same for all the test done.
  - It has an 83% accuracy
- The Tests used are as follows:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbor (KNN)

```
Logistic Regression Test Accuracy Score : 0.8333333333333334
SVM Test Accuracy Score : 0.8333333333333334
Decision Tree Test Accuracy Score : 0.8333333333333334
KNN Test Accuracy Score : 0.8333333333333334
```



Confusion Matrix



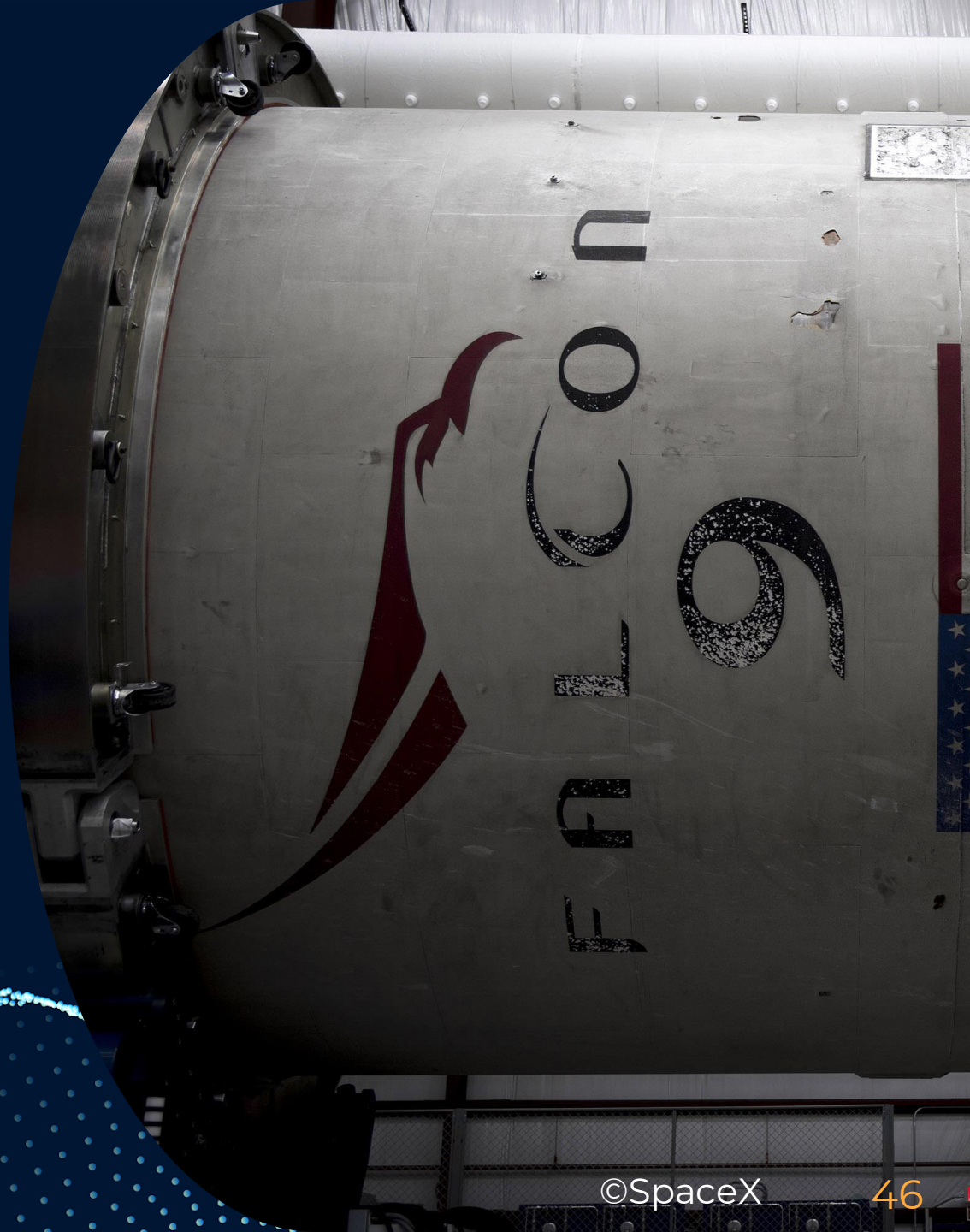
- The confusion matrix shown is the same for all the tests done giving the same result.
- This shows that the predication that twelve will land and three do not land.
- There three false positives which gives a error in the train and test set.

Radecic, D.

Predicted Values	Actual Values: Negative	Actual Values: Positive
Negative	True Negative (TN)	False Positive (FP)
Positive	False Negative (FN)	True Positive (TP)

Conclusions

- SpaceX will continue to have more successful launches as years progress
- KSC LC-39A is should be used more for successful launches, data showing that it is the most successful
- For booster with a payload heavier than 8000 kg there is a higher success rate in orbits ISS, PO, and VLEO
- For payloads less than 8000 kg Orbits LEO and SSO have a better chance at success
- Orbits ES-L1, GEO, HEO, and SSO have the best success rate
- All machine learning test conclude a prediction of 83% success for landing.



©SpaceX

Appendix

- For The Classification Accuracy slide for the bar chart, here is the code with the data frame used for the plot.

```
In [32]:
log_score = logreg_cv.score(X_test, Y_test)
svm_score = svm_cv.score(X_test, Y_test)
tree_score = tree_cv.score(X_test, Y_test)
KNN_score = knn_cv.score(X_test, Y_test)
data_accuracy = {'Method':['Logistic Regression', 'SVM', 'Decision Tree', 'KNN'],'Score':[log_score, svm_score,tree_score, KNN_score]}


Accuracy_Table = pd.DataFrame(data=data_accuracy)
Accuracy_Table
```

```
sns.barplot(x='Score', y='Method', hue='Method', ci=30, data=Accuracy_Table)
plt.legend([], [], frameon=False)
plt.show
```

- Used Jupyter nbviewer.com for the Folium Notebook
- https://nbviewer.org/

[Link to the Full GitHub]

Made flowcharts through Lucidchart

Lucidchart

lucidchart.com

References

- *NASA*. (n.d.). NASA. Retrieved April 19, 2023, from http://nasa.gov
- *NASA Images | NASA*. (n.d.). NASA. Retrieved April 19, 2023, from https://www.nasa.gov/multimedia/imagegallery/index.html
- *NASA Invites Media to Next SpaceX Space Station Cargo Launch | NASA*. (n.d.). NASA. Retrieved April 19, 2023, from https://www.nasa.gov/press-release/nasa-invites-media-to-next-spacex-space-station-cargo-launch
- Neufeld, M. (2020, May 26). *Launch Complex 39: From Saturn to Shuttle to SpaceX and SLS | Smithsonian Voices | National Air and Space Museum Smithsonian Magazine*. History, Travel, Arts, Science, People, Places Smithsonian Magazine; Smithsonian Magazine. https://www.smithsonianmag.com/blogs/air-space-museum/2020/05/26/launch-complex-39-saturn-shuttle-spacex-and-sls/
- *Official SpaceX Photos | Flickr*. (n.d.). Flickr. Retrieved April 19, 2023, from https://www.flickr.com/photos/spacex
- Radecic, D. (2018, September 28). *A Non-Confusing Guide to Confusion Matrix*. Towards Data Science. https://www.towardsdatascience.com/a-non-confusing-guide-to-confusion-matrix-7071d2c2204f
- Sheetz, M. (2022, March 23). *SpaceX raises prices for launches and Starlink due to inflation*. CNBC; CNBC. https://www.cnbc.com/2022/03/23/spacex-raises-prices-for-launches-and-starlink-due-to-inflation.html
- *SpaceX*. (n.d.). SpaceX. Retrieved April 19, 2023, from http://spacex.com
- *SpaceX launch cadence sees two Falcon 9 boosters in port for first time - SpaceFlight Insider*. (2021, March 20). SpaceFlight Insider; https://www.facebook.com/pages/The-Spaceflight-Group/254774621252230. https://www.spaceflightinsider.com/organizations/space-exploration-technologies/spacex-launch-cadence-sees-two-falcon-9-boosters-in-port-for-first-time/

Thank You

©NASA