Winning Space Race with Data Science

Lee Stetson
February 18, 2022

Executive Summary

- This project uses a wide range of data science techniques to make observations and predictions for SpaceX Falcon 9 landings from the data collection stage all the way through the predictive modelling stage. Data collected via a REST API and web scraping is cleaned for use in the subsequent stages. Data analysis is performed by creating data visualizations using Matplotlib, Seaborn, Folium, and Plotly Dash. Further data analysis is done using SQL to parse datasets. Finally, classification models are developed and evaluated to predict the outcomes of Falcon 9 launches.
- The classification models were highly accurate (accuracy > 0.8) in predicting the landing outcome given parameters payload mass and launch site.

Introduction

- Commercial space travel is a new and rapidly developing field as our desire for space exploration grows. SpaceX is leading the charge with their relatively affordable Falcon 9 rockets.
- Our goal is to use SpaceX data available for Falcon 9 launches to predict whether a launch will successfully land. This prediction will be used to estimate the cost of flight since the first stage of a successful launch can be reused.

Section 1

Methodology

Methodology

- Data collection:
  - REST API
  - Web Scraping
- Data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

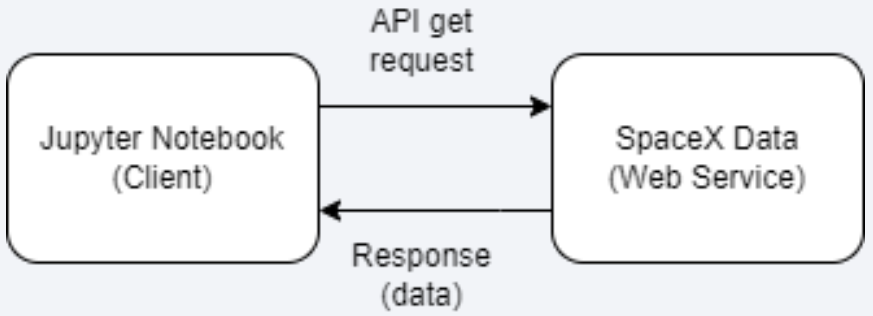
Data Collection

- Data was collected using two methods – a REST API and web scraping.
- A get request to the SpaceX API was used to extract launch data including the booster version, payload mass, orbit, launch site, outcome, longitude and latitude of launch, and other data about each launch.
- The launch data for Falcon 9 and Falcon Heavy launches between June 2010 and June 2021 were extracted from tables on the “List of Falcon 9 and Falcon Heavy launches” Wikipedia page. An HTTP get method and a BeuatifulSoup object were used to parse the webpage.
- For both methods, pandas dataframes were used to tabulate the data.

Data Collection – SpaceX API

- Data was collected using the requests.get() method and a static response object.
- The data was stored in a dictionary with the desired variables, then filtered to only include Falcon 9 launches.
- Finally, missing values for PayloadMass were replaced with the mean of that column.
- Link to notebook containing methods:
  - Data Collection API

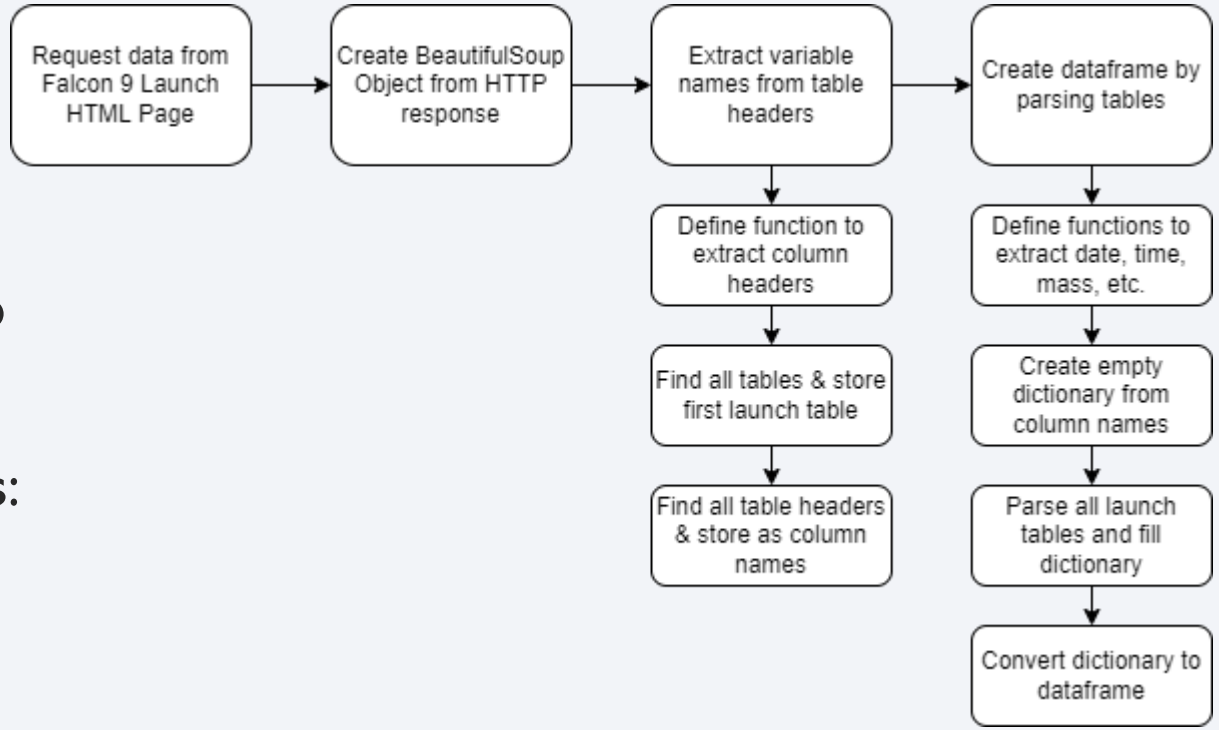
API get request Flowchart



Data Collection – Web Scraping

- Data was collected via a get request from the Wikipedia page.
- A beautiful soup object was used to parse the data and fill a dictionary.
- This dictionary was then converted to a pandas dataframe.
- Link to notebook containing methods:
  - Data Collection Web Scraping

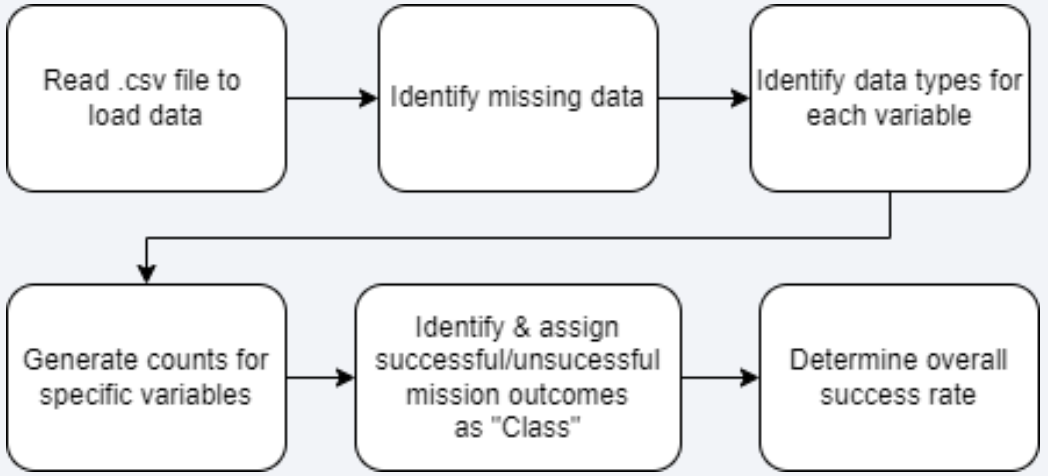
Web Scraping Flowchart



Data Wrangling

- Data from a .cvs file was read into a pandas dataframe
- The initial goal was to create what will ultimately be the target variable, Class, which represents whether a landing was successful
- Link to notebook containing methods:
  - Data Wrangling

Data Wrangling Flowchart



EDA with Data Visualization

- The following charts were created to observe the Class of each flight with respect to different variables:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Flight Number vs. Orbit
  - Payload Mass vs. Orbit
  - Orbit vs. Success Rate
  - Year vs. Success Rate
- Link to notebook containing methods:
  - Data Visualization

EDA with SQL

- SQL queries were executed to find the following information:
  - Names of unique launch sites
  - 5 records where launch sites begin with ‘CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date of first successful landing on a ground pad
  - Names of boosters with a payload mass between 4000 and 6000 kg which had success landing on a drone ship
  - Total number of successful and failed missions
  - Names of booster versions which carried the maximum payload mass
  - Failed landing outcomes in drone ships in 2015
  - Ranked list of landing outcomes between 2010-06-04 and 2017-03-20
- Link to notebook containing methods:
  - EDA with SQL

Build an Interactive Map with Folium

- Circles with icons were added at each of the four launch sites.
- Marker clusters were added to show the successful and failed launches at each site.
- Lines were added to show the distance from one site to nearby landmarks.

- These objects were added to more effectively visualize the locations of launch sites and their success rates.
- Link to notebook containing methods:
  - Interactive Map with Folium

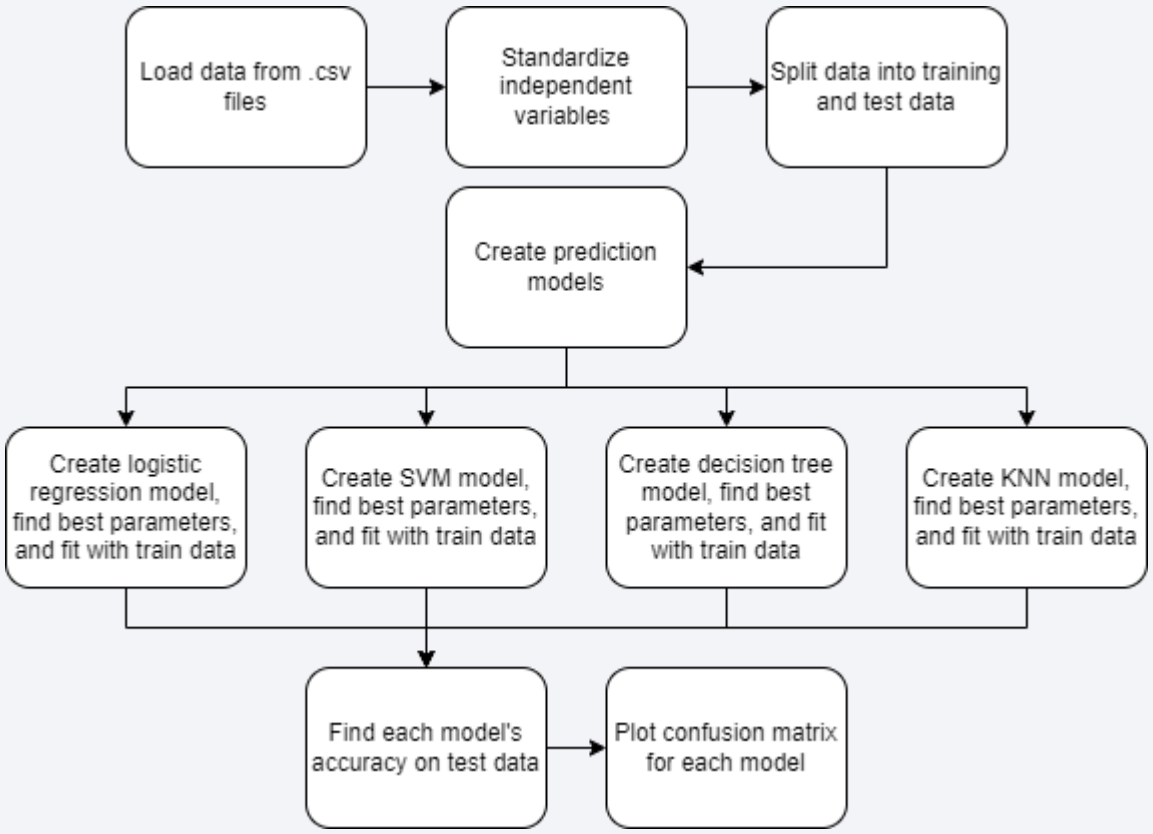
Build a Dashboard with Plotly Dash

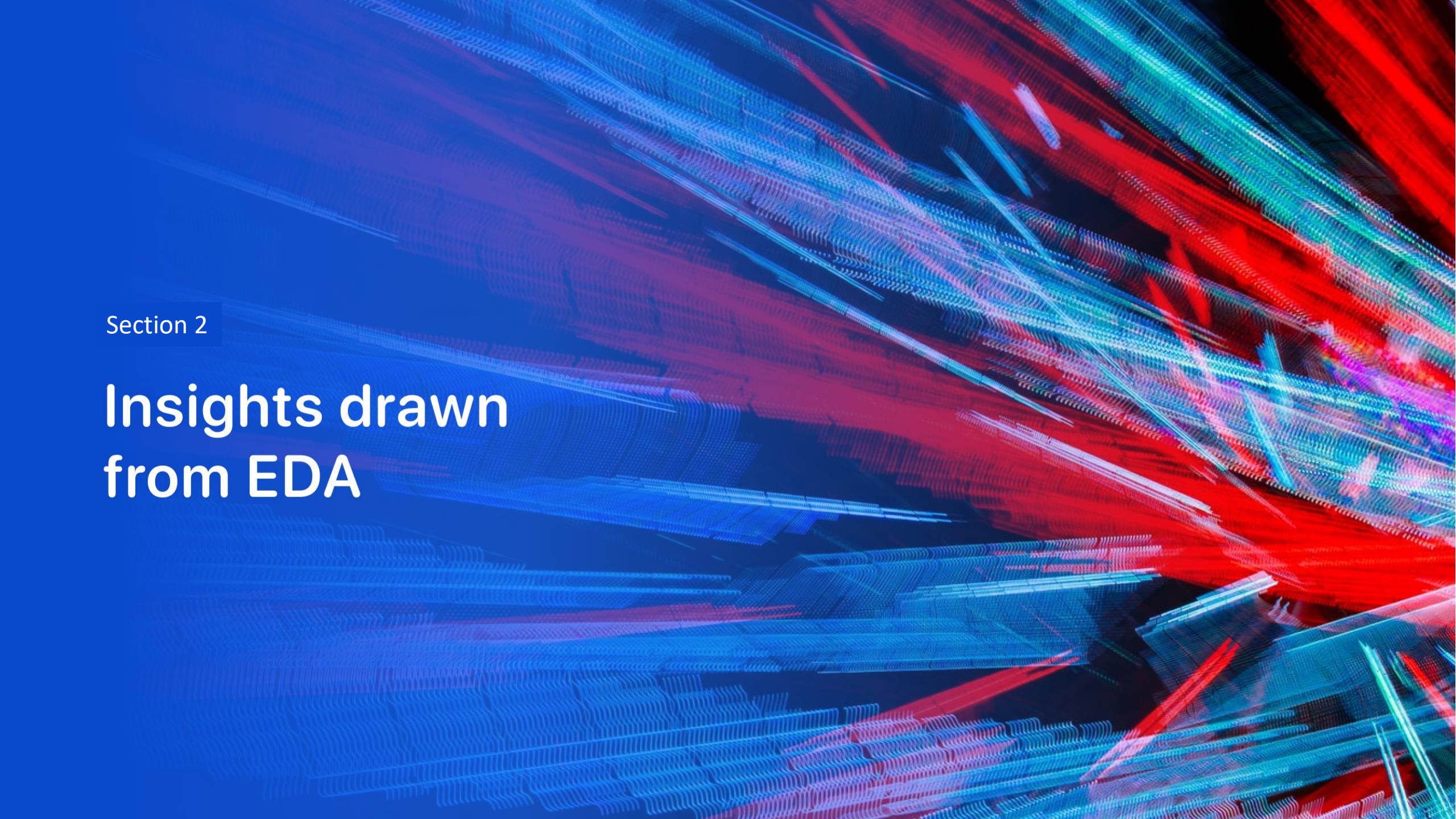
- A dashboard was developed that takes user dropdown input of launch site and slider input of payload mass
- The output is:
  - A pie chart showing the count of successful launches
  - A scatter plot of the successful and failed launches for the given payload mass range
- These charts give a better understanding of how success rate is correlated to payload mass. The interactivity allows for easy comparison between launch sites.
- Link to Python code containing methods:
  - Plotly Dash Dashboard

Predictive Analysis (Classification)

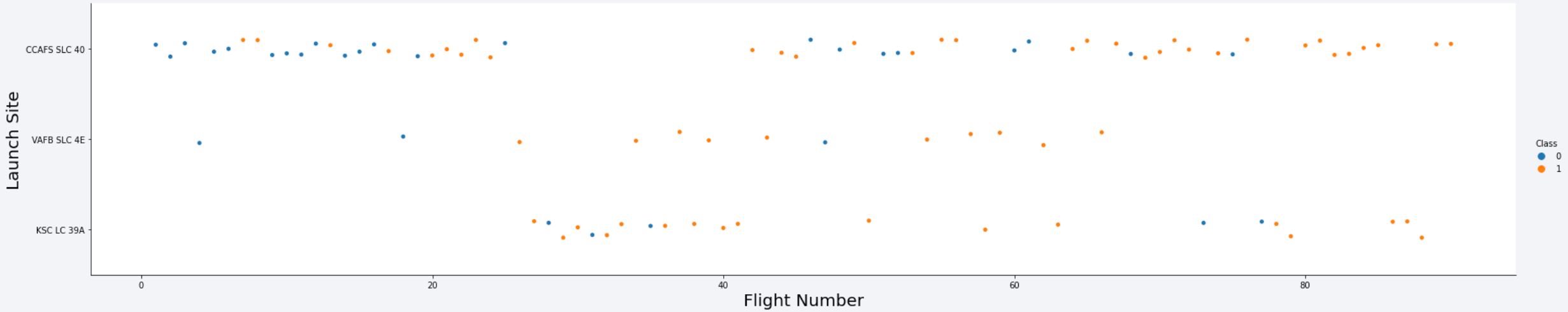
- Four different classification models were developed to predict the success of a given launch. They included:
  - Logistic Regression, Support Vector Machine, Decision Tree, & K-Nearest Neighbors
- Each model's optimal parameters and their accuracies on test data were determined.
- Link to notebook containing methods:
  - Machine Learning Prediction

Predictive Modeling Flowchart



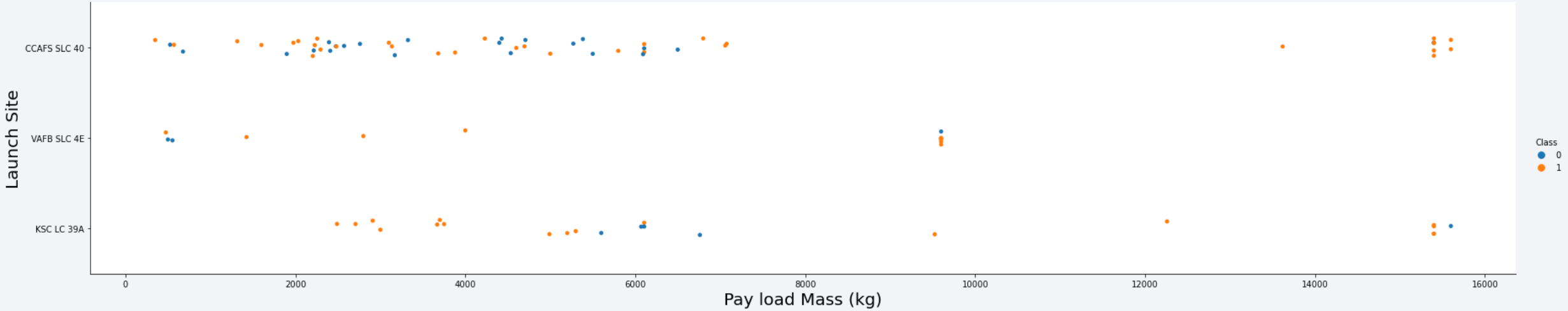

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- This scatter plot shows the location of the launch site for each flight. Most flights were launched from CCAFS SLC 40 while very few were from VAFB SLC 4E. During a period from approximately Flight Numbers 25 to 40, many of the flights were launched from KSC LC 39A.
- Flights denoted class 0 were unsuccessful and class 1 were successful. Not much can be inferred about the impact of launch site on success rate, but as flight number increases, more flights are successful.

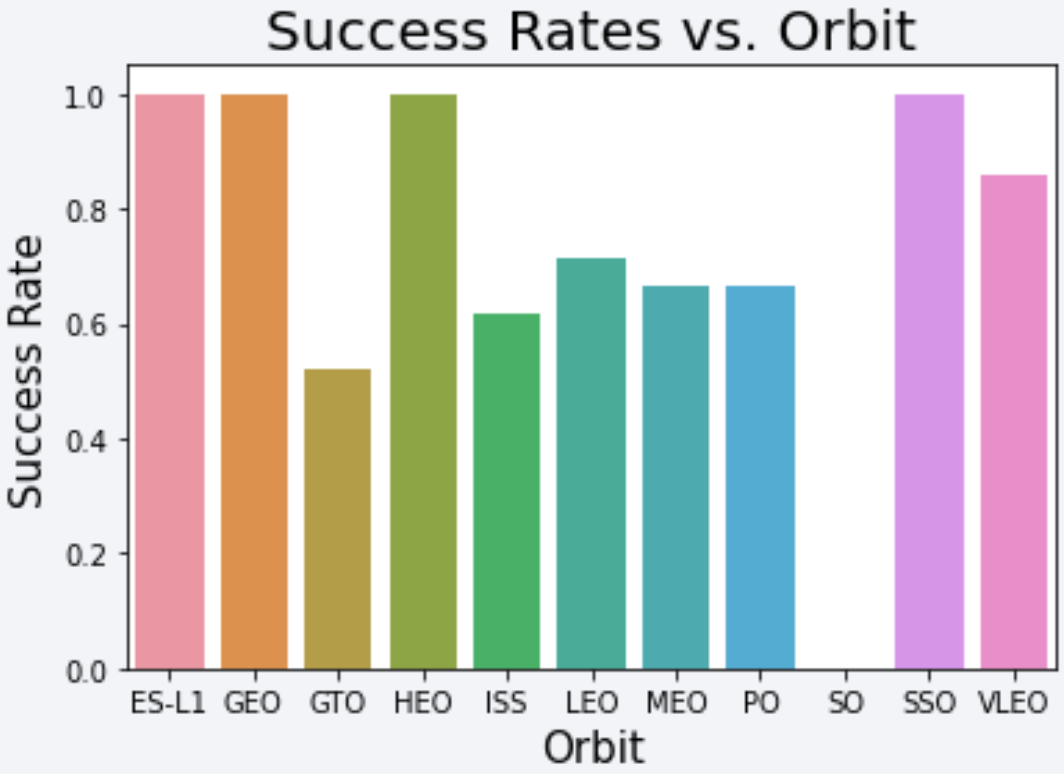
Payload vs. Launch Site



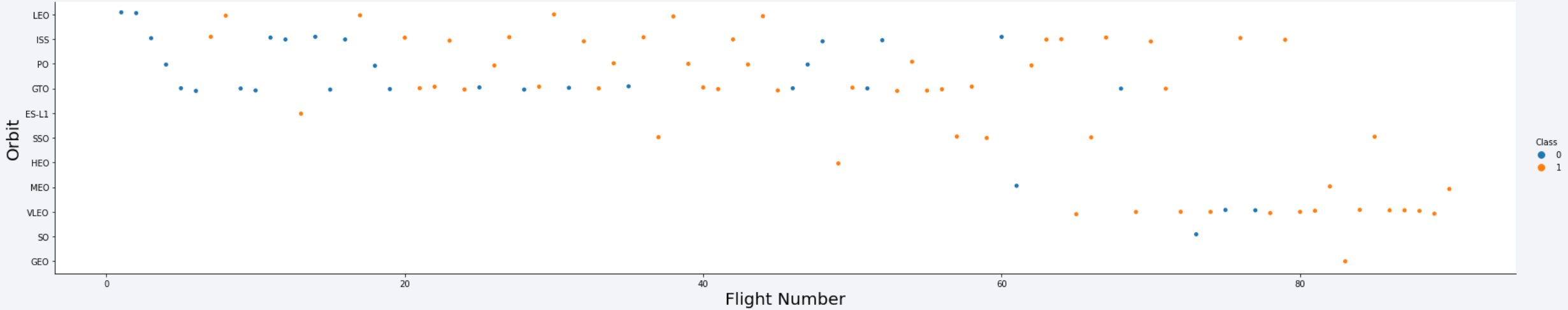
- The payload and launch site of each flight are shown. The majority of flights had payload masses of less than 8,000 kg and only CCAFS SLC 40 and KSC LC 39A launched payloads greater than 10,000 kg.
- It should be noted that larger payloads (> 8,000 kg) had a higher success rate than those with smaller payloads.

Success Rate vs. Orbit Type

- The success rate of each of the 11 types of orbits is shown with 1.0 representing a 100% success rate.
- ES-L1, GEO, HEO, and SSO orbits had a 100% success rate while the SO orbit type had a 0% success rate.
- All orbits except the SO orbit had a success rate over 50%.

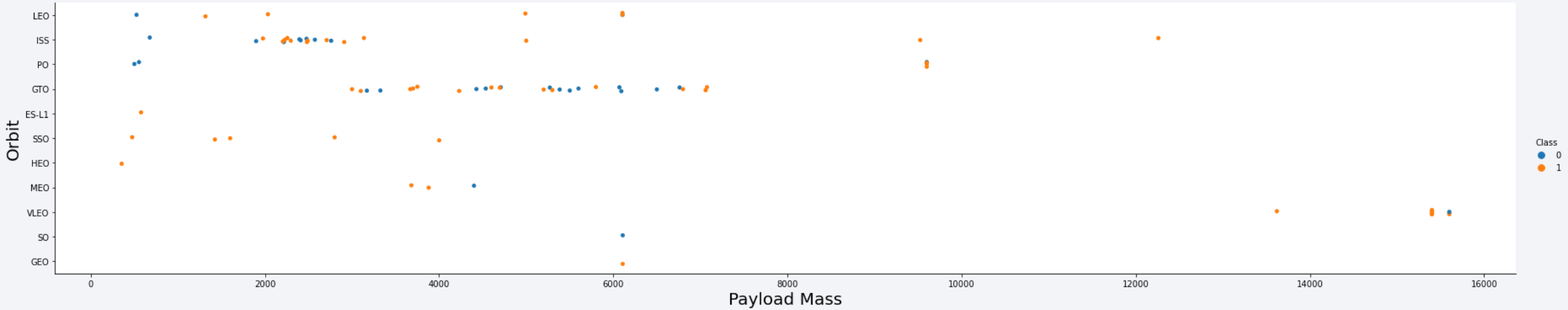


Flight Number vs. Orbit Type



- The orbit type for each flight number is shown. Most of the early flights were of the LEO, ISS, PO, and GTO types and most of the later flights were the VLEO orbit type.
- There is a scarcity of flights in the ES-L1, HEO, MEO, SO, and GEO orbits, so their success rates may not be indicative of the impact of their orbit type on success.

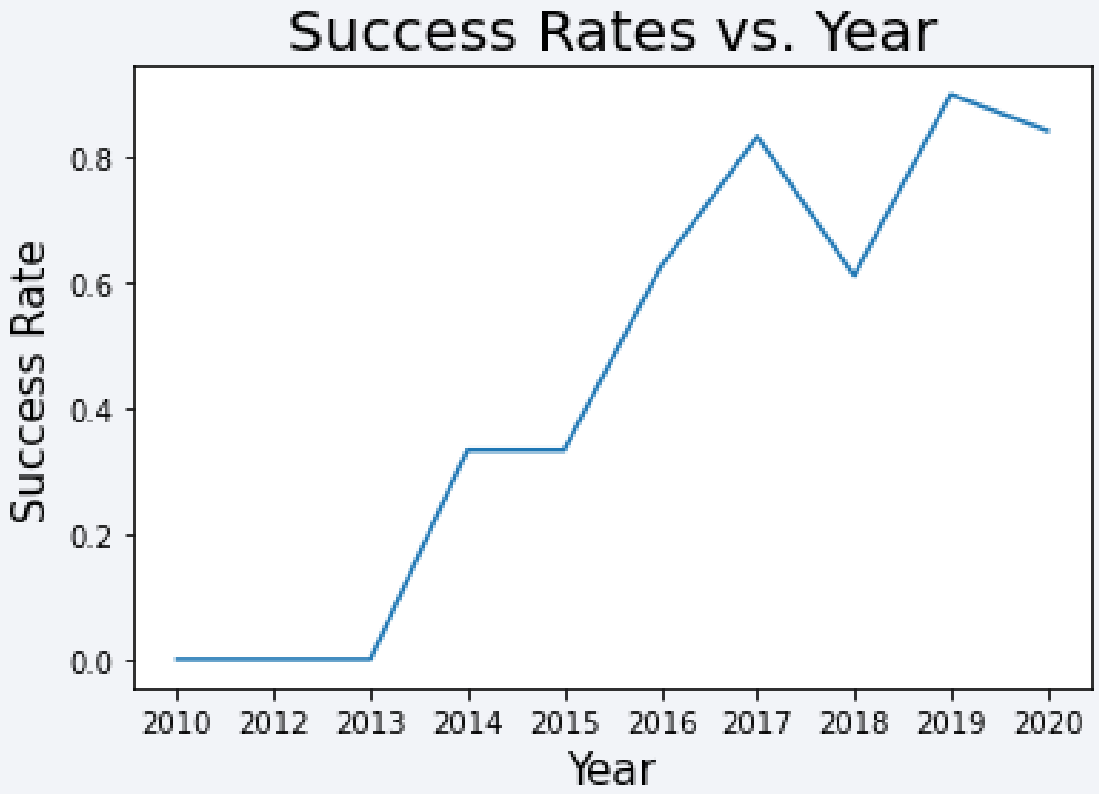
Payload vs. Orbit Type



- The orbit type and payload mass for each flight are shown. Most orbits contained flights with similar payload masses, however there are outliers.

Launch Success Yearly Trend

- The success rates for flights in each year are shown
- Success rates show an upward trend with time from 0.0 in 2013 up to above 0.8 by 2019.
- The increase in success rate is not linear and seems to flatten at about 0.8.



All Launch Site Names

- Unique Launch Sites:
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E
- SQL Query: SELECT DISTINCT Launch_Site FROM SPACEXTBL
- This query searches the SpaceX table located in an IBM database for distinct Launch Sites.

Launch Site Names Begin with 'CCA'

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	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	00:35:00	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

- The first 5 records where the launch site begins with ‘CCA’ are shown above.
- SQL Query: SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Total Payload Mass

- Total payload carried by boosters from NASA (CRS): **45,596 kg**
- SQL Query:

```
SELECT SUM(PAYLOAD_MASS__KG_) AS NASA_CRS_SUM_OF_PAYLOAD_MASS
    FROM SPACEXTBL
    WHERE Customer = 'NASA (CRS)'
```

- This query sums the payload masses with the sum() function using the column alias “NASA_CRS_SUM_OF_PAYLOAD_MASS “

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: **2,928 kg**
- SQL Query:

```
SELECT AVG(PAYLOAD_MASS__KG_)
        AS F9_V1_1_AVERAGE_PAYLOAD_MASS
        FROM SPACEXTBL
        WHERE Booster_Version = 'F9 v1.1'
```

- This query averages the payload masses for booster version F9 v1.1 with the avg() function using the column alias “F9_V1_1_AVERAGE_PAYLOAD_MASS “

First Successful Ground Landing Date

- The first successful landing on a ground pad was **2015-12-22**.
- SQL Query:

```
SELECT MIN(Date) FROM SPACEXTBL
    WHERE Landing__Outcome = 'Success (ground pad)'
```

- This query displays the earliest date that the landing outcome was 'Success (ground pad)' using the min() function.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
  - F9 FT B1021.2
  - F9 FT B1031.2
  - F9 FT B1022
  - F9 FT B1026
- SQL Query:

```
SELECT DISTINCT Booster_Version FROM SPACEXTBL
    WHERE Landing__Outcome = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

- This query displays the results described above by filtering with two where conditions.

Total Number of Successful and Failure Mission Outcomes

- Count of mission outcomes:

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

- SQL Query:

```
SELECT Mission_Outcome, COUNT(*) AS COUNT FROM SPACEXTBL
    GROUP BY Mission_Outcome
```

- This query returns the count for failures and successes of flights by grouping by mission outcome.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass:
  - 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:

```
SELECT DISTINCT Booster_Version FROM SPACEXTBL
    WHERE PAYLOAD_MASS__KG_ =
    (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

- This query returns the boosters that have the max payload mass using a nested query.

2015 Launch Records

- Failed landings in drone ship in 2015:

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- SQL Query:

```
SELECT Booster_Version, Launch_Site, Landing__Outcome
    FROM SPACEXTBL
    WHERE Landing__Outcome = 'Failure (drone ship)'
    AND Date LIKE '2015%'
```

- This query returns the booster version and launch site for drone ship failures in 2015.

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

- Count of landing outcomes between the date 2010-06-04 and 2017-03-20:
- SQL Query:

```
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
```

- This query returns a sorted list of the count for each landing outcome in a date range.

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 3

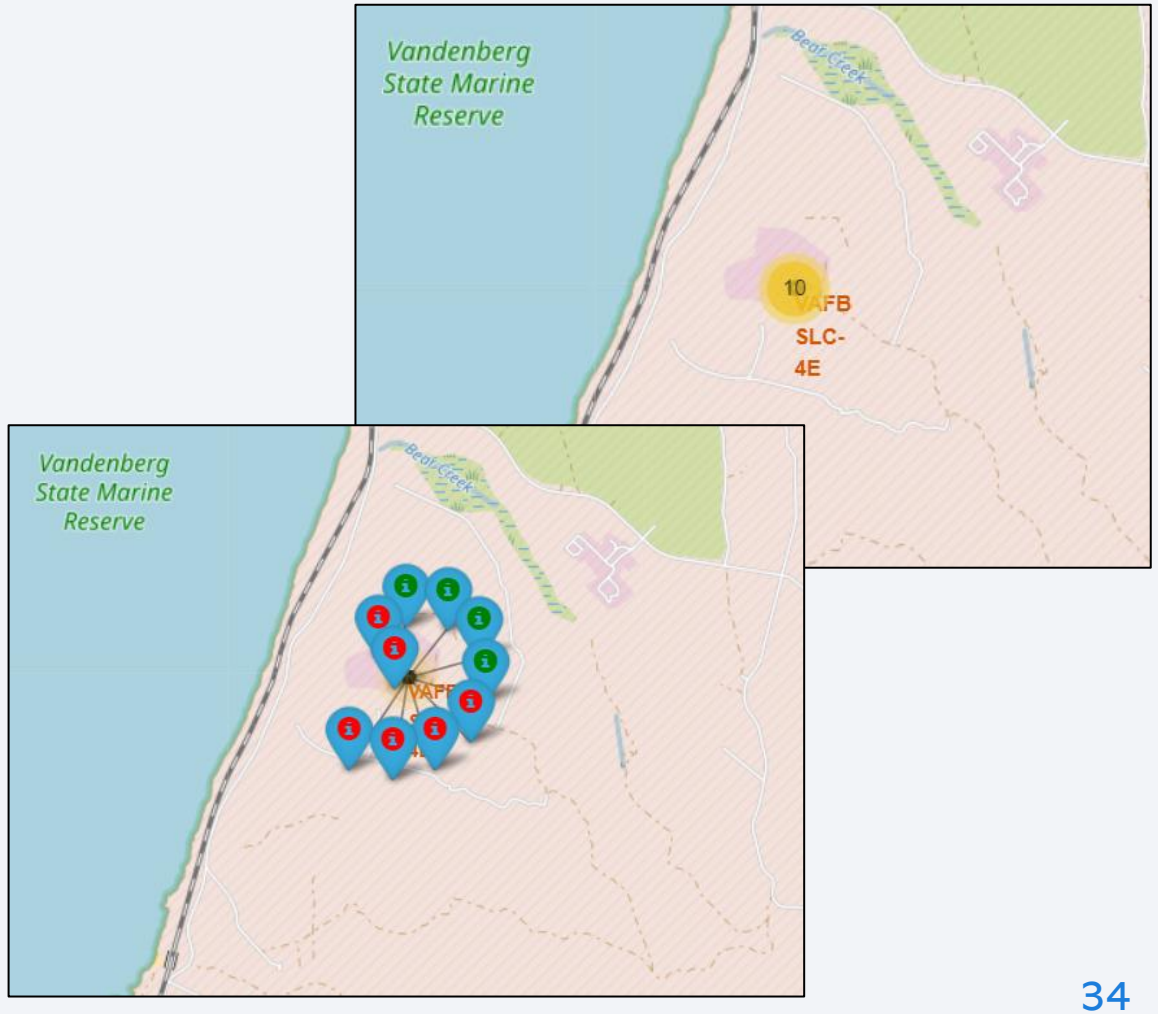
Launch Sites Proximities Analysis

Folium Map – Launch Site Locations

- Locations of CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, & VAFB SLC-4E launch sites marked with circles

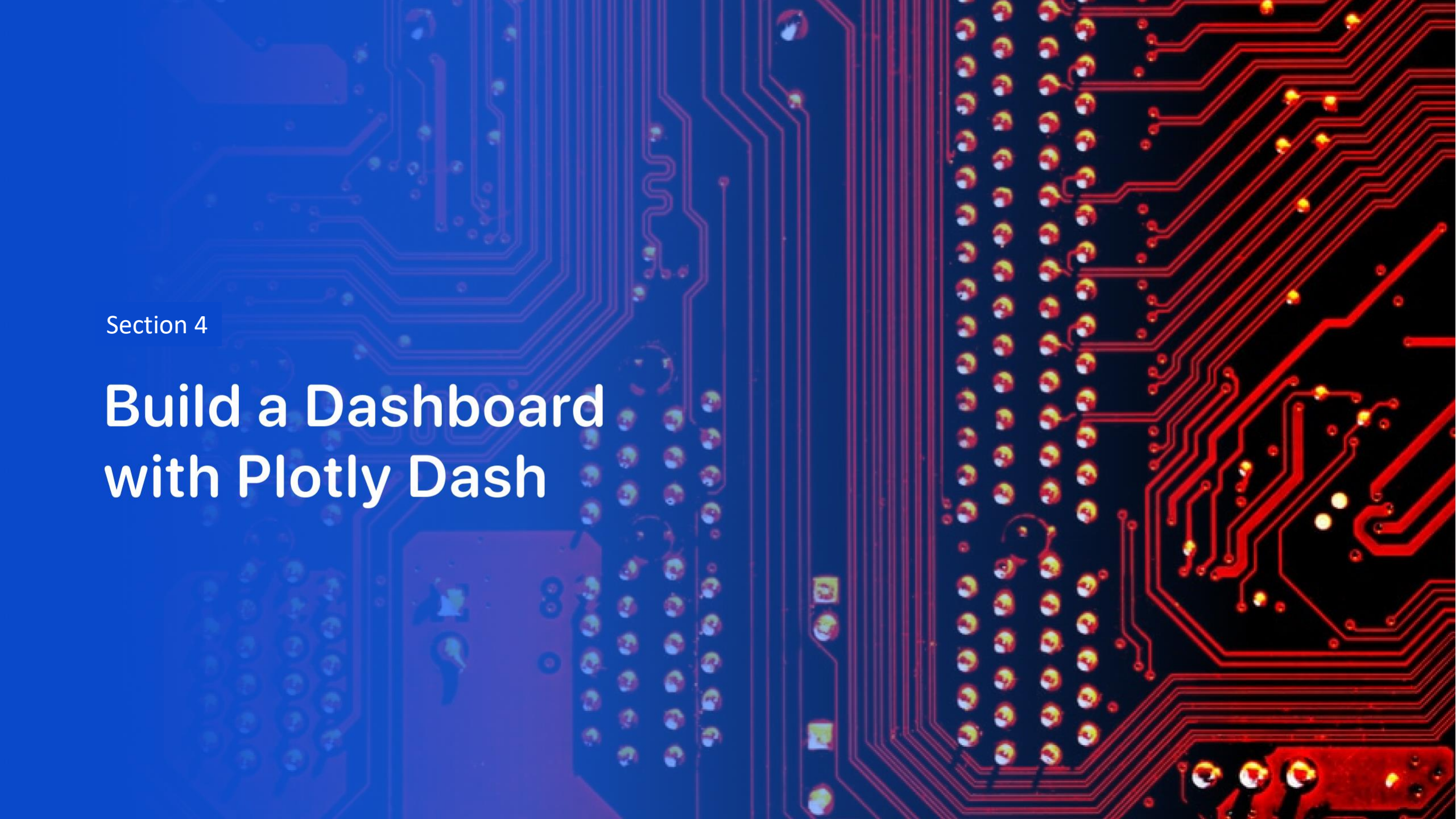
Folium Map – Launch Outcomes

- Launch outcomes for the VAFB SLC-4E launch site are shown with a marker cluster that can be exploded to see the individual outcomes.
- The cluster count makes it easy to pinpoint which sites had the most launches and the color-coded icons aid in class identification.



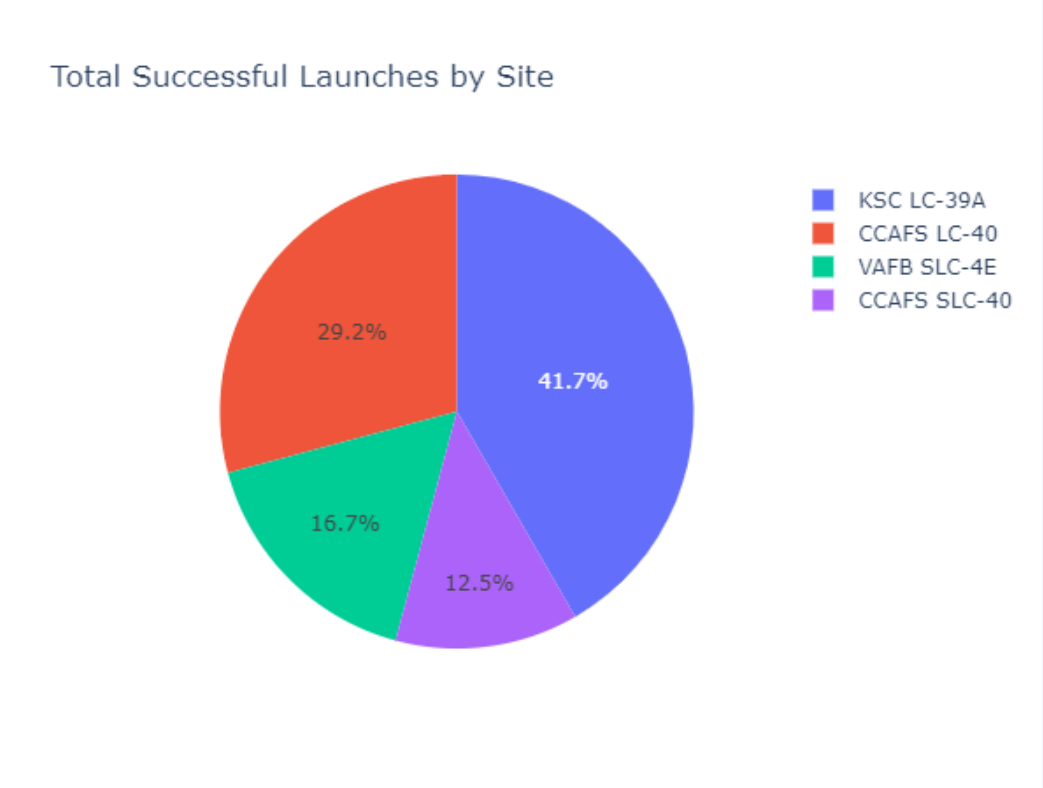
Folium Map – Launch Site Landmark Proximity

- The distances from the CCAFS LC-40 to the coastline (0.94 km) and to Titusville (23.10 km) are shown with lines and distance labels.
- The lines help with visualization of distances to nearby locations.


Section 4
Build a Dashboard with Plotly Dash

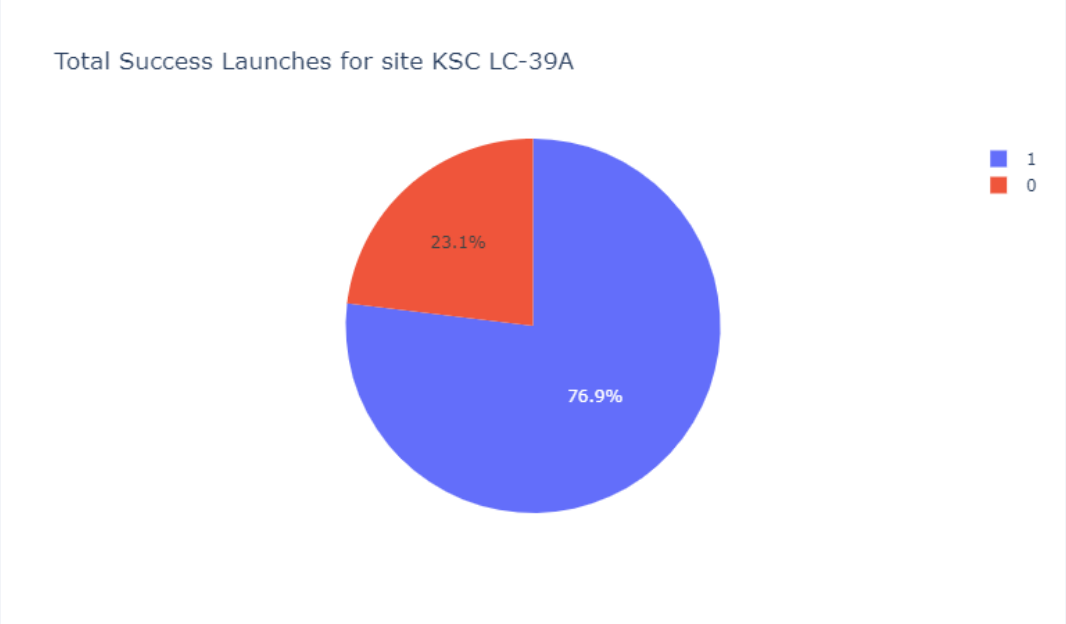
Dashboard – Launch Success (All Sites)

- The percentage of successful flights launched from each site is shown. KSC LC-39A launched the most successful flights (10 out of 24 total)
- While the CCAFS LC-40 site accounted for 46.4% of all flights, it only launched 29.2% of the total successful flights.



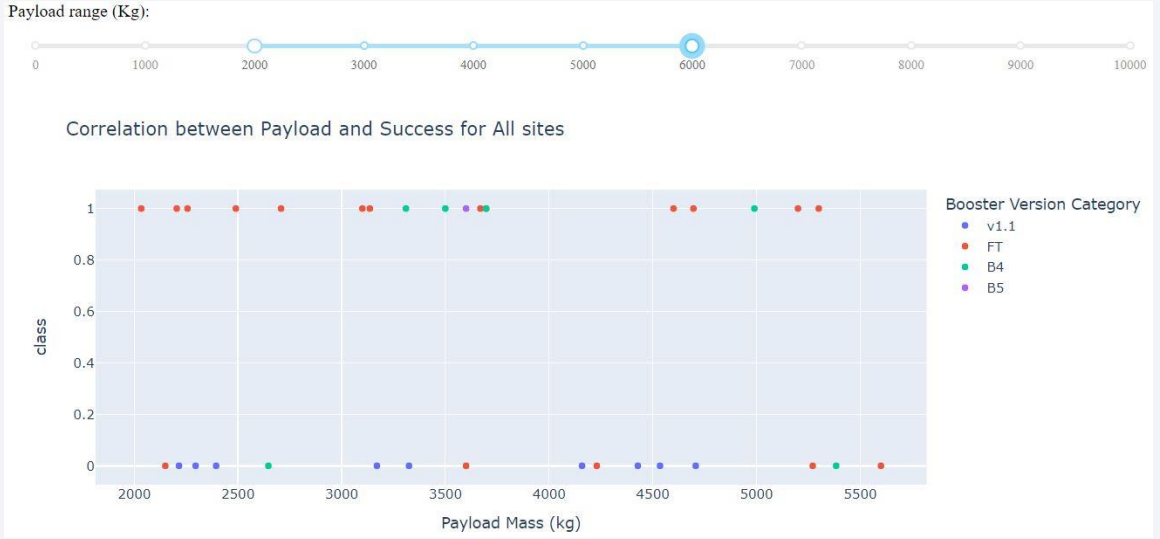
Dashboard – Success of Launches from KSC LC-39A

- The KSC LC-39A launch site had the highest success rate at 76.9% (10 out of 13 flights successful).
- This site accounts for 41.7% of all successful flights.



Dashboard – Success by Payload Mass

- Flight success for two payload mass ranges (0-10,000 kg and 2,000-6,000 kg) are shown.
- The truncated range where most flights fall into shows that payload mass alone may not give a clear indication of success rate.

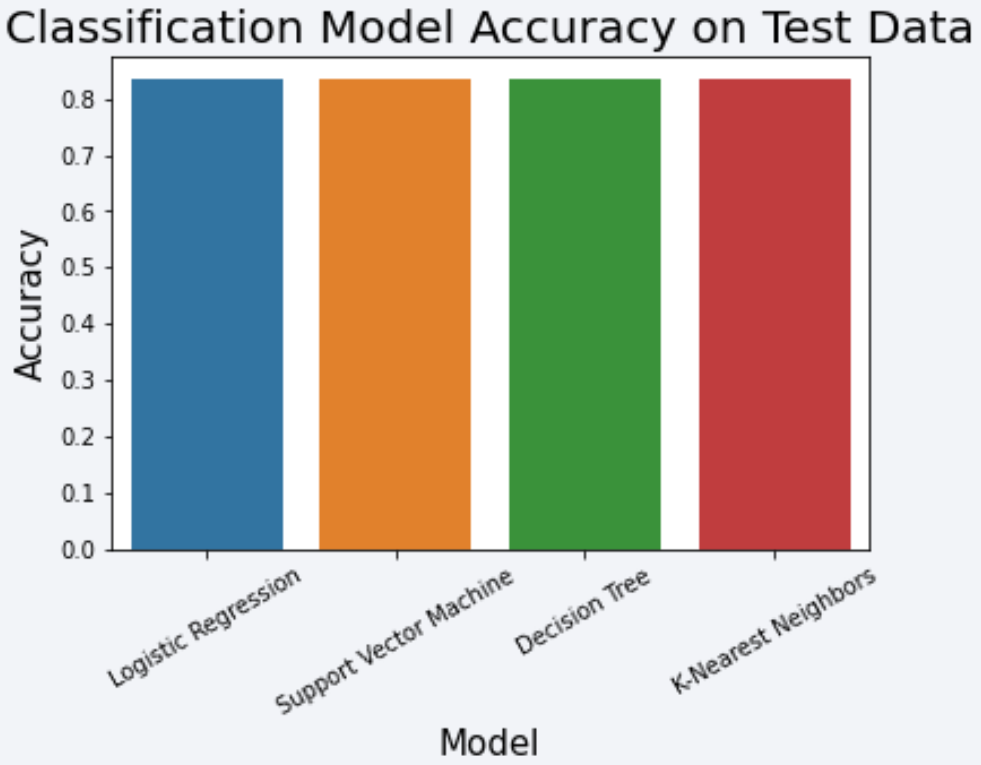


Section 5

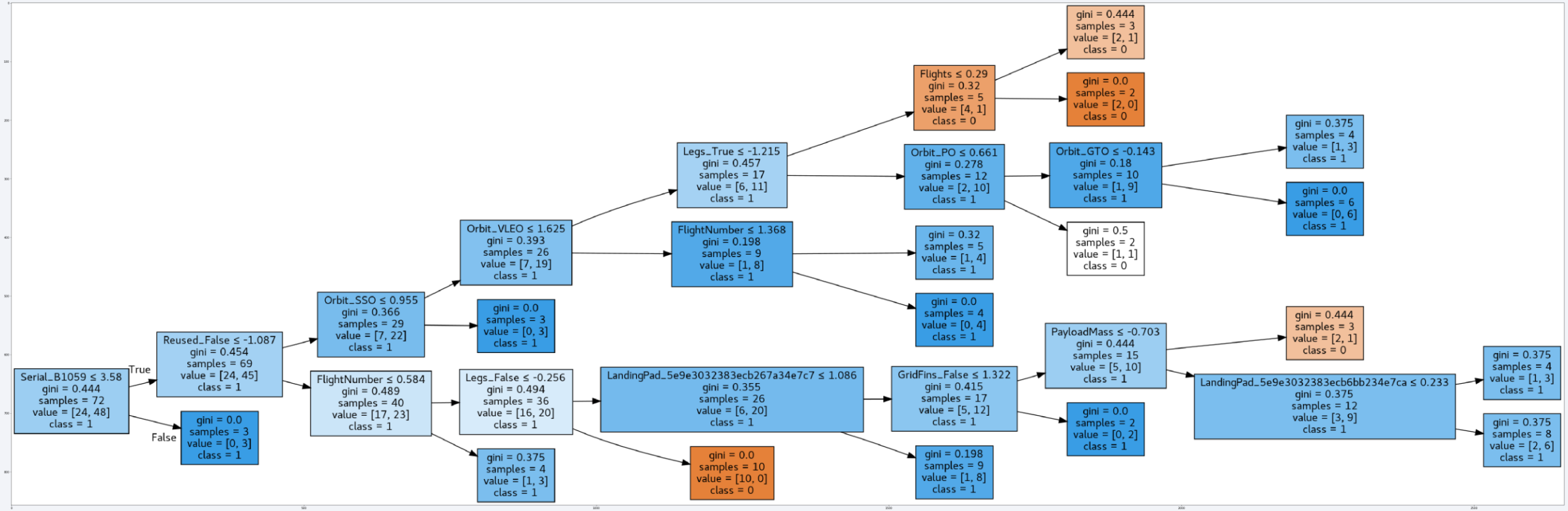
Predictive Analysis (Classification)

Classification Accuracy

- All the models had an 0.8333 accuracy on the test data. This is because of the small size of the test set.
- The decision tree had the highest accuracy on training data at 0.8875.



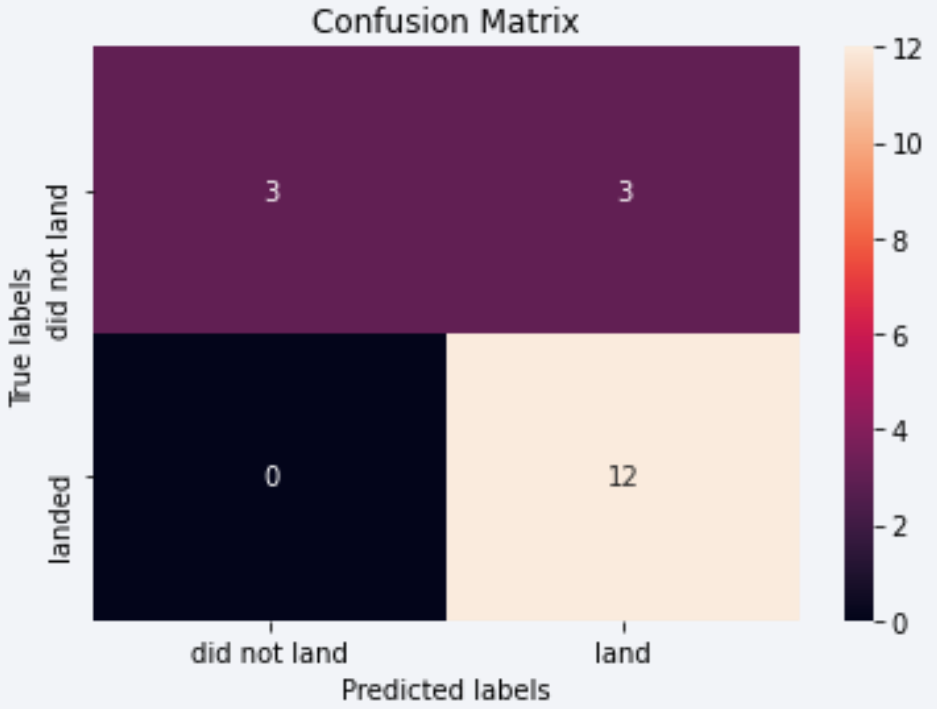
Decision Tree Visualization



* Note the tree feature values are standardized.

Confusion Matrix

- The confusion matric for all models were identical as shown on the right.
- The models were effective in prediction launch outcome (class), but there were a significant number of false positives (predicted successful landing when actual landing failed).



Conclusions

- No clear trends were initially apparent by looking at the charts of a single independent variable and the class of the launch.
- The success rate increased as time went on, so more recent flights (and upcoming flights) are more likely to succeed. This means older flights may be less useful in prediction.
- All classification models had the same accuracy on the small test data set, but the decision tree was slightly more effective on the training data.
- A more accurate and effective model may be produced by adding relevant independent variables such as weather conditions and landing location.

Appendix

- Github repository with all notebooks and code:
  - https://github.com/lstetson30/Data-Science-Capstone

Thank you!