IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jason
October 10, 2022

Outline

- Github Link & Overview
- Executive Summary
- Introduction
- Section 1: Methodology
- Section 2: Insights from EDA
- Section 3: Launch Sites Proximities
- Section 4: Data Dashboard with Plotly
- Section 5: Predictive Analysis
- Conclusion

Github Repository, Notebooks, & Python Files

https://github.com/jasonhooy/DSCapstone

- Collecting the Data:
  1. Data Collection API (REST)
  2. Data Collection with Web Scraping (BeautifulSoup)
- Exploratory Data Analysis (EDA):
  3. With %SQL connected to DB2 data storage
  4. Matplotlib/Seaborn Data Visualizations and Feature Engineering
- Interactive Visual Analytics and Dashboards:
  5. With Folium
  6. With Plotly Dash
- Predictive Analysis (Classification):
  7. Machine Learning Regression Analysis

Executive Summary

- Collected data from SpaceX's website using their APIs and scrapped the SpaceX Wikipedia HTML webpage using Beautiful Soup
- Created a column called "class" to denote successful-failed launches
- Performed EDA using magic SQL (tethered to IBM's DB2), Seaborn, and Matplotlib libraries in Jupyter Notebook
- Created Folium interactive maps with marker clusters and polylines
- Created an interactive data dashboard with a pie chart and scatter plot using an IDE
- Performed predictive analysis using classification models with the train test split method to generate the most accurate model shown in a confusion matrix

Introduction

- The Falcon 9 reusable rocket (depicted to the right) allows SpaceX to underbid competitors by $103 million.
- SpaceX charges $62 million per launch.
- We want to know which launch site, booster version, orbital destination, and payload mass combinations are most successful in order to model an alternative commercial option maximizing the reusable rocket technology.
- Using SpaceX data, we run predictive cost-benefit analysis to determine what parameters would make for a commercially feasible alternative to SpaceX.

Section 1

Methodology

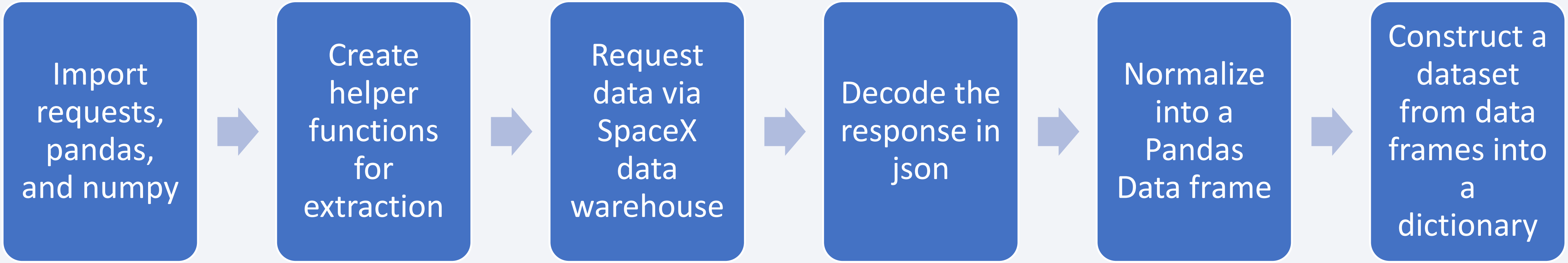
Methodology Outline

- Data Collection & Data Wrangling
  - SpaceX API
  - Web Scraping
- EDA Visualizations and SQL
  - Folium
  - Plotly Dash
- Predictive Analysis Using Regression Analysis
  - Classification Models

Data Collection SpaceX API Overview

- Imported Requests, to make HTTP requests and get data using the APIs. (supported with Pandas and Numpy)
    - Define 4 functions to call and append data via the SpaceX API to one table in data storage:
        - Rocket, Launchpad, Payload, and Cores.
- Create static json request object, request, then parse (normalize) and clean the data.
- Create new Pandas data frames.
- Execute get requests with the 4 predefined functions.
- Create dictionary of data from data get request results.

Data Collection SpaceX API Flowchart



Import requests, pandas, and numpy

Create helper functions for extraction

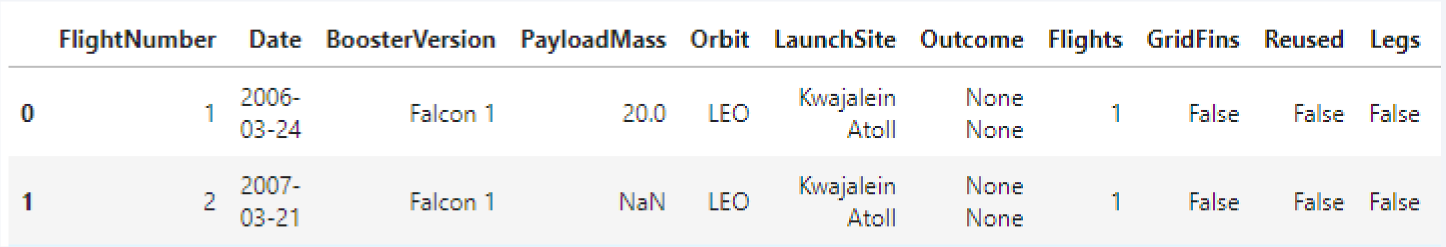
Request data via SpaceX data warehouse

Decode the response in json

Normalize into a Pandas Data frame

Construct a dataset from data frames into a dictionary

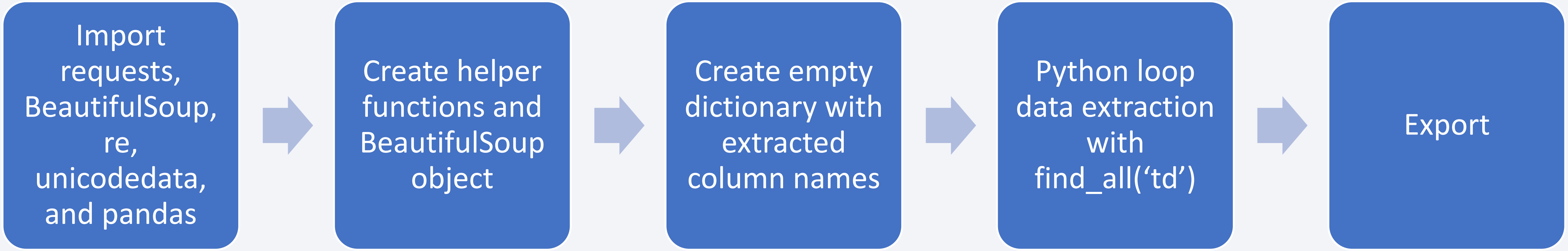
RESULT →



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False

https://github.com/jasonhooy/DSCapstone/blob/main/Getting%20the%20Data.ipynb

Data Collection – Web Scraping


Import requests, BeautifulSoup, re, unicodedata, and pandas

Create helper functions and BeautifulSoup object

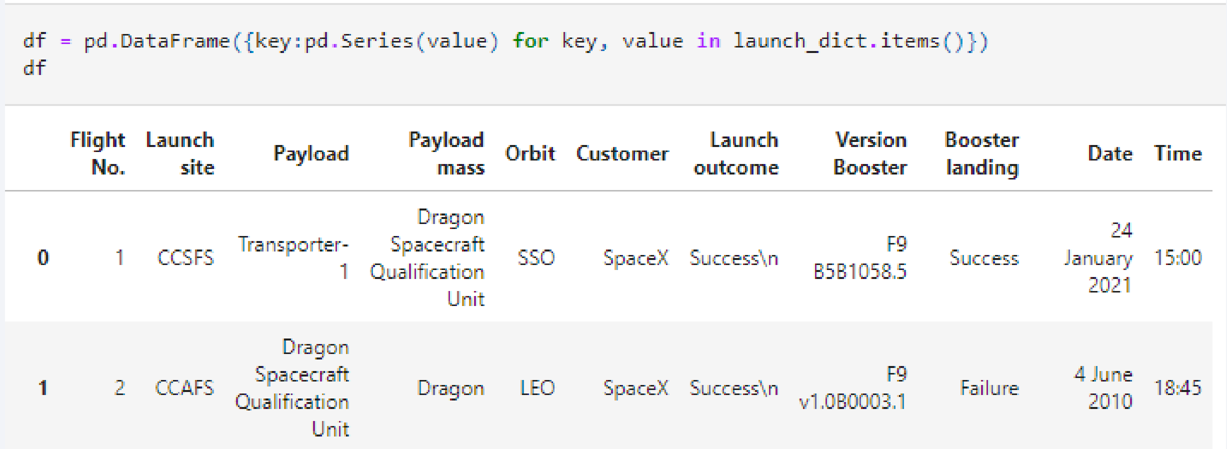
Create empty dictionary with extracted column names

Python loop data extraction with find_all('td')

Export

- Extract HTML Wiki table with BeautifulSoup and find_all
- Parse and convert to Panda data frame

RESULT →



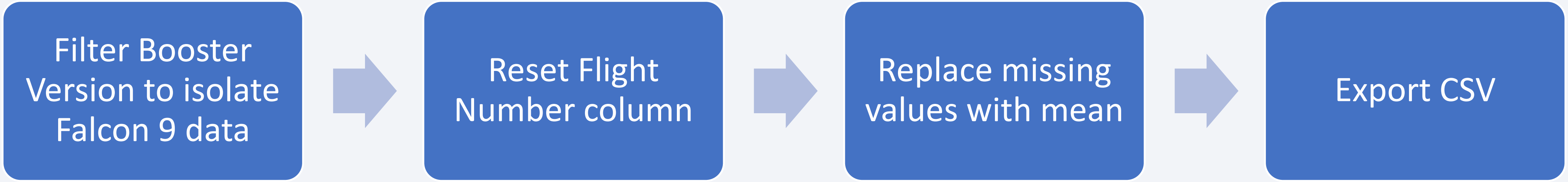
```
df = pd.DataFrame({key:pd.Series(value) for key, value in launch_dict.items()})
df
```

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCSFS	Transporter-1	Dragon Spacecraft Qualification Unit	SSO	SpaceX	Success\n	F9 B5B1058.5	Success	24 January 2021	15:00
1	2	CCAFS	Dragon Spacecraft Qualification Unit	Dragon	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45

https://github.com/jasonhooy/DSCapstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Filter Booster Version column to remove non-Falcon 9 data.
- Reset the data frame Flight Number column to reflect Falcon 9 launches only.
- Identify and replace missing values in the Payload Mass column with the column mean.
- Export CSV for data analysis.



Filter Booster Version to isolate Falcon 9 data

Reset Flight Number column

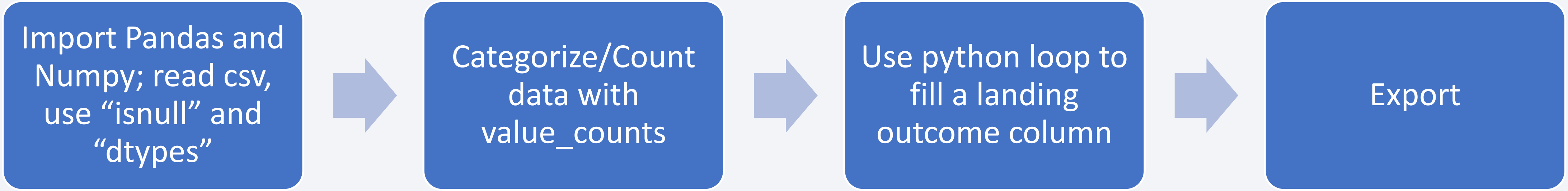
Replace missing values with mean

Export CSV

https://github.com/jasonhooy/DSCapstone/blob/main/Getting%20the%20Data.ipynb

https://github.com/jasonhooy/DSCapstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling Continued



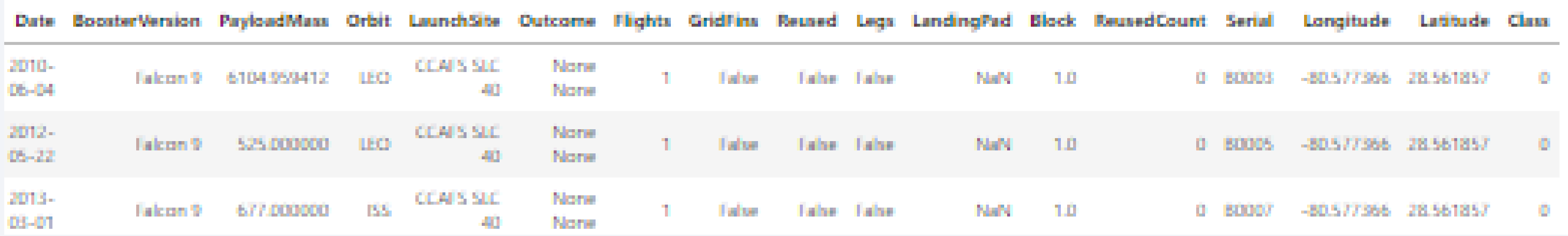
Import Pandas and Numpy; read csv, use “isnull” and “dtypes”

Categorize/Count data with value_counts

Use python loop to fill a landing outcome column

Export

RESULT →



Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0

https://github.com/jasonhooy/DSCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Use seaborn scatterplots to determine the relationships among variables including:
  - Flight number, payload mass, launch site, orbit, and outcome
- Use matplotlib scatter plots and bar charts to assess how frequently launch sites, payload mass, and orbit occur and at what success rates.
- Use seaborn line plot to determine the progression of launch success from 2010 to 2020.
- Scatter, bar, and line charts help discern the nuanced difference among variables impacting first stage rocket success implicating why.

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Load CSV to IBM Cloud DB2 data warehouse
- Connect data to Jupyter Notebook and load %sql extension
- Import Pandas to establish data frames/table
- Altered Table to unify naming conventions
- Conducted analysis with the following queries determining:
  - Sum of Booster Version (CRS) payload mass
  - Average payload mass carried by booster version F9 v1.1
  - Listed Booster Versions that carried the max payload mass
  - Listed time-specific failed launches by drone ship landing launch site
  - Listed successful and failed launch outcomes by landing type in a specified time

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Used site_map.add_child(circle/marker) to plot the following launch sites on an interactive map:
- CCAFS_LC_40, CCAFS_SLC_40, KSC_LC_49A, VAFB_SLC_4E.
- Created a column correlating color to class values 0 & 1 as red/green.
- Added folium.marker and marker cluster to display the colored plots in the vicinity of each launch site.
- Used folium.polyline to add lines (with distance) from the launch sites to the nearst: rail, highway, and city.
- These markers, circles, and lines provide insight into the success of each launch site and what civil amenities (infrastructure) are needed to successfully operate a launch site.

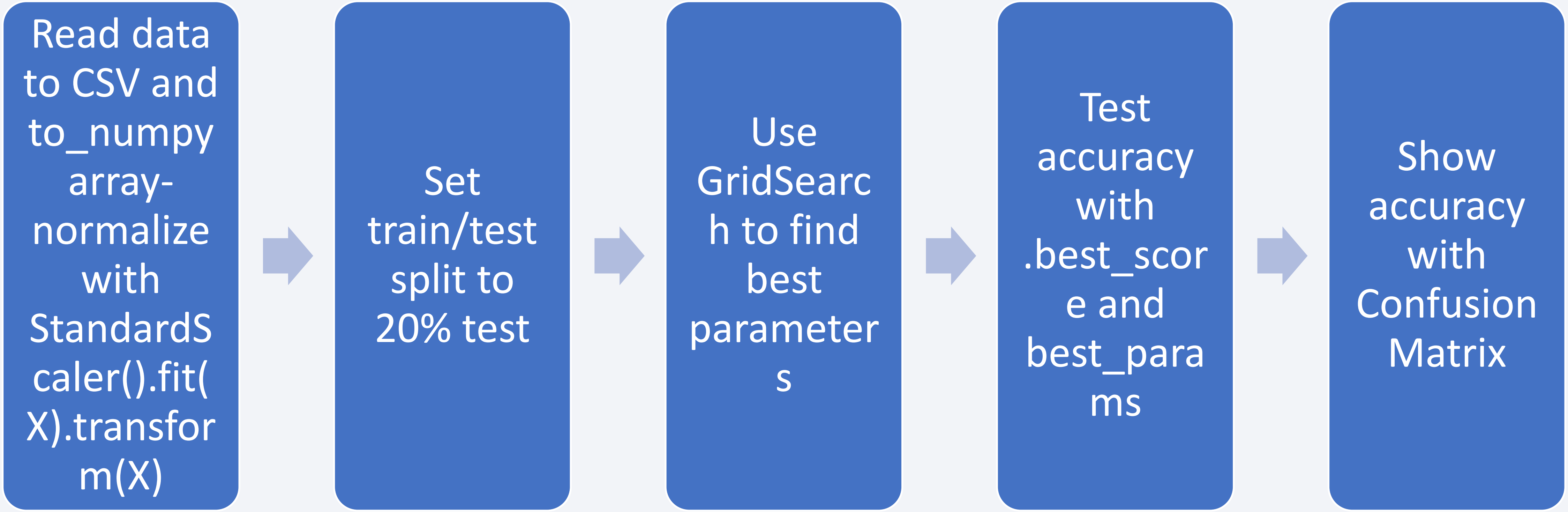
https://github.com/jasonhooy/DSCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Used dcc.Dropdown to create interactive list of each launch site.
- Used dcc.RangeSlider to create an interactive payload range to change data displayed by weight.
- Created a pie chart with a call back function aligned with the dropdown.
- Created a scatter plot to show the relationship between payload mass and success.

https://github.com/jasonhooy/DSCapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



Read data to CSV and to_numpy array-normalize with StandardScaler().fit(X).transform(X)

Set train/test split to 20% test

Use GridSearch to find best parameters

Test accuracy with .best_score and best_params

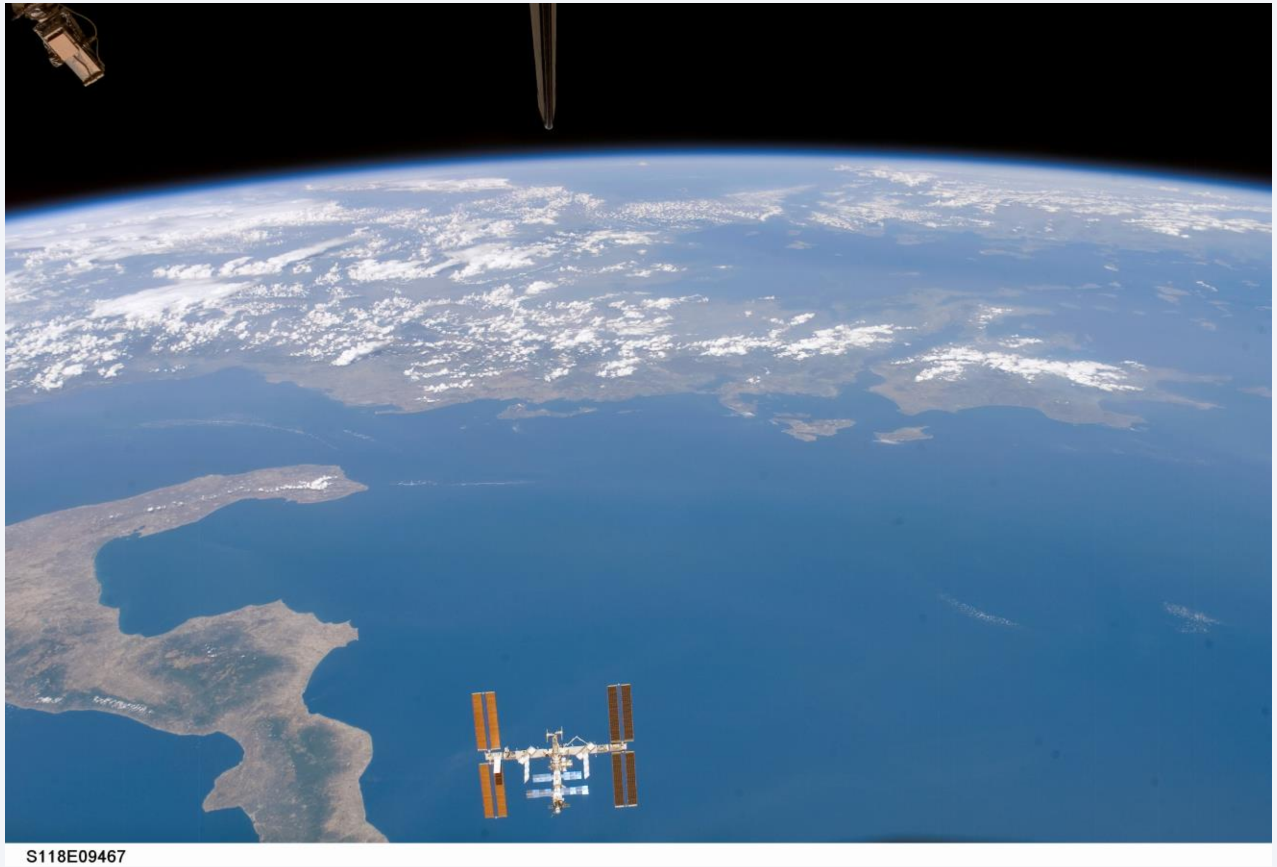
Show accuracy with Confusion Matrix

- Built, evaluated and improved a classification model cylcing the process on: logistic regression, SVM, decision tree, and K Nearest Neighbor.
- The decision tree was the most accurate model with 94% accuracy.

https://github.com/jasonhooy/DSCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(2).ipynb

Results Outline

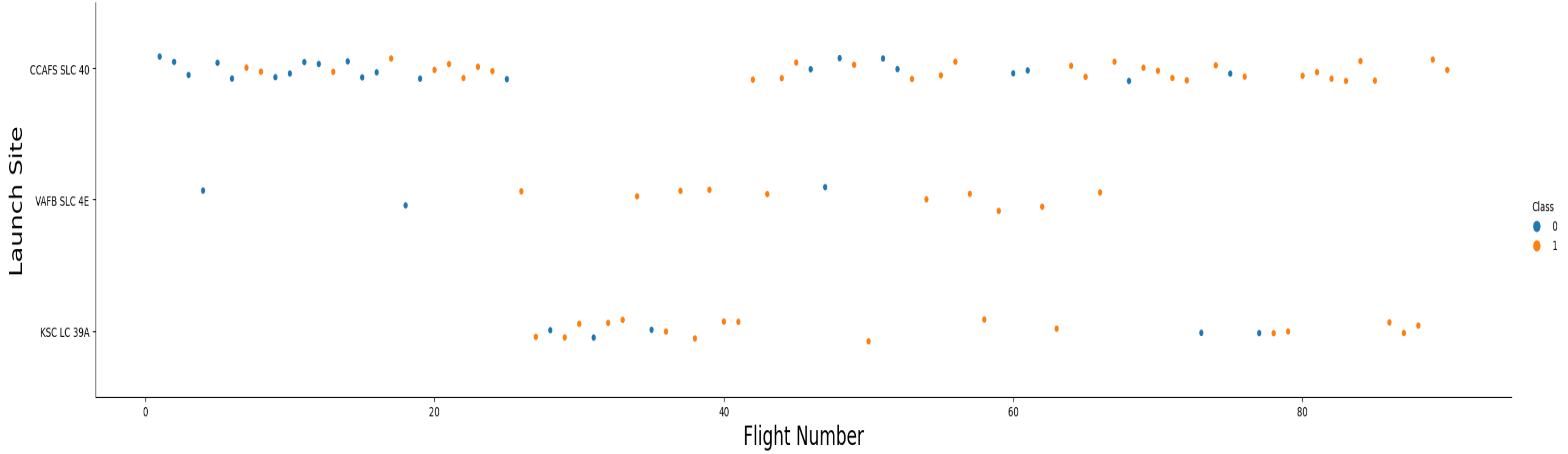
- Exploratory data analysis results
- Interactive analytics demonstration in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

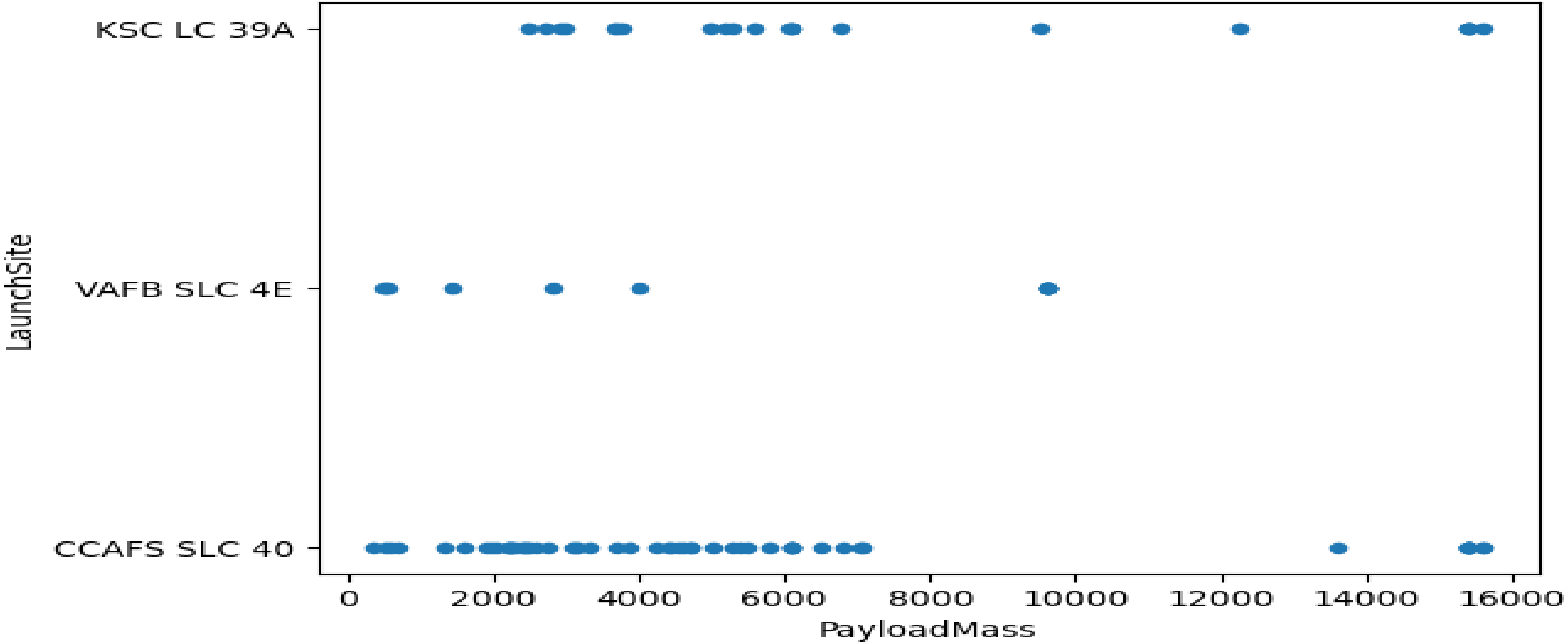
Flight Number vs. Launch Site



- This scatter plot shows CCAF SLC 40 is the preferred launch site with increasing success with every new flight

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

Payload vs. Launch Site

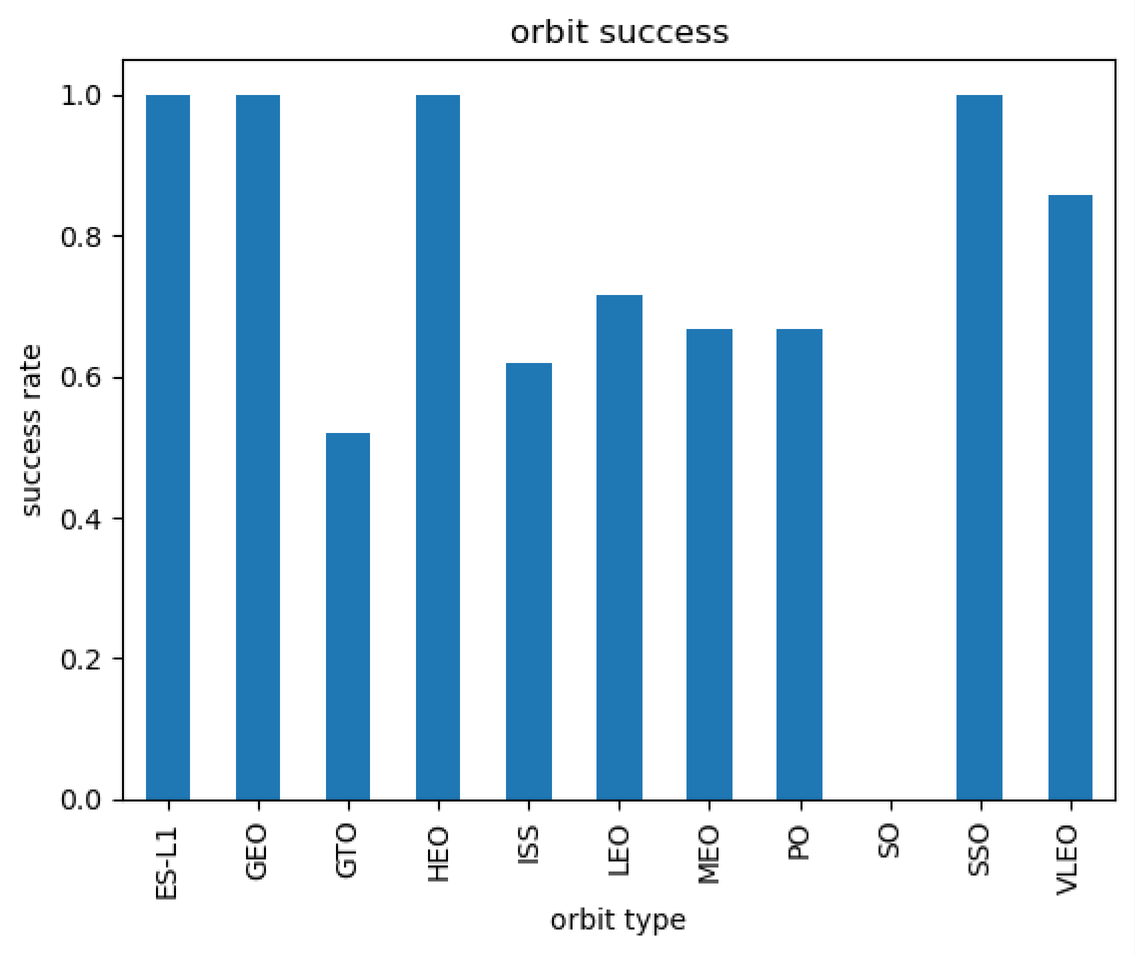


- CCAFS SLC 40 is primarily used for low weight payload launches below 7500 Kgs
- KSC LC 39A is used for launches of many weight varieties

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

Success Rate vs. Orbit Type

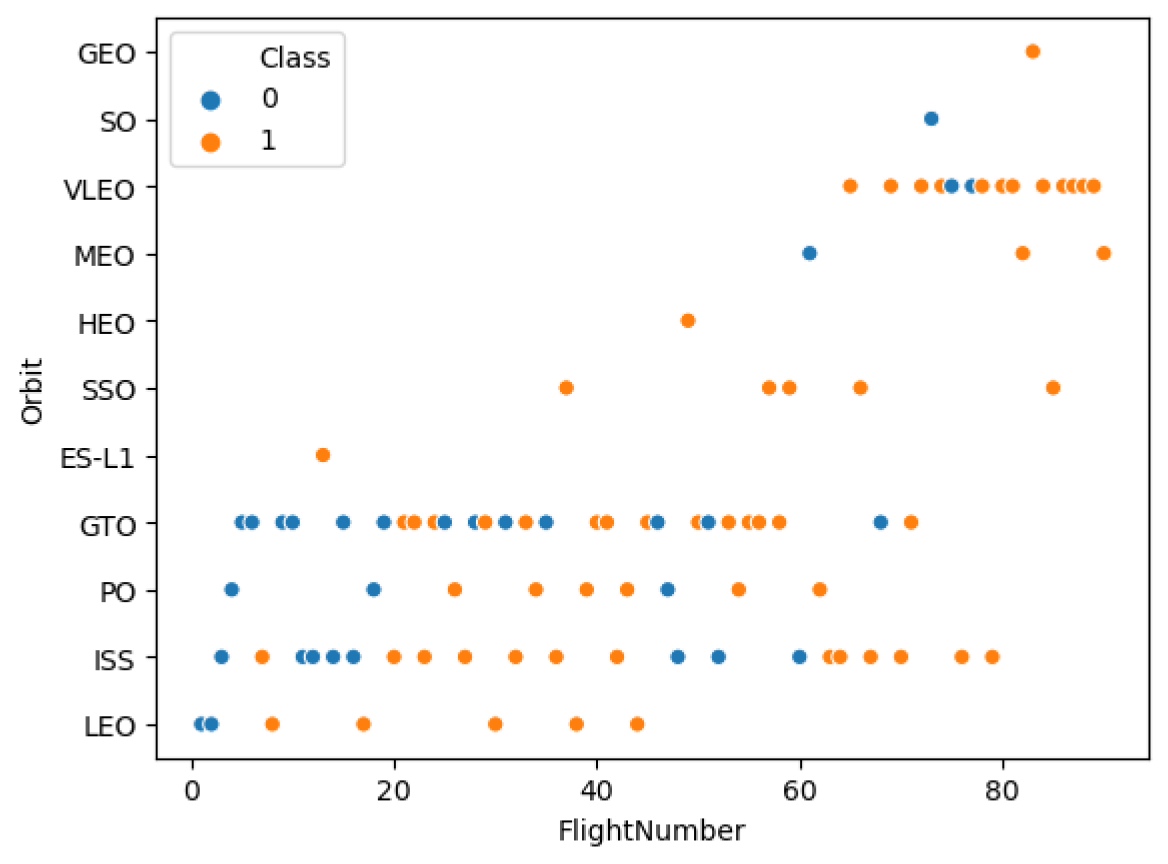
- 100% successful launches to:
  - ES-L1, GEO, HEO, and SSO
- High probability of success in the above orbital destinations are the best orbits to shuttle our rockets.
- We want to avoid orbits with <80% success:
  - GTO, ISS, MEO, PO, and SO



https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

Flight Number vs. Orbit Type

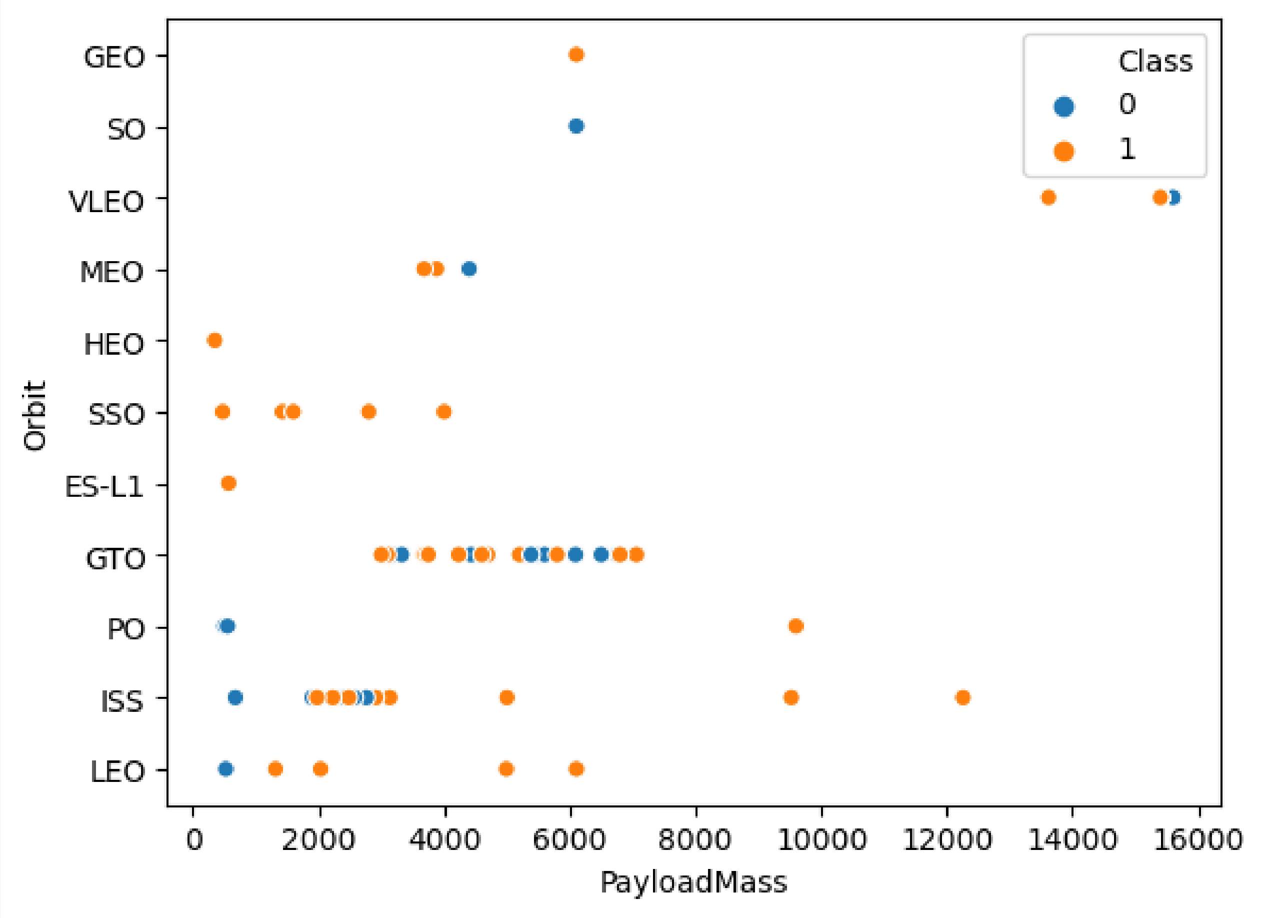
- LEO Launches were phased out
- But, VLEO flights are occurring more regularly
- 1:4 of the last 40 launches went to VLOE
- The last 6 launches to the ISS were successful



https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

Payload vs. Orbit Type

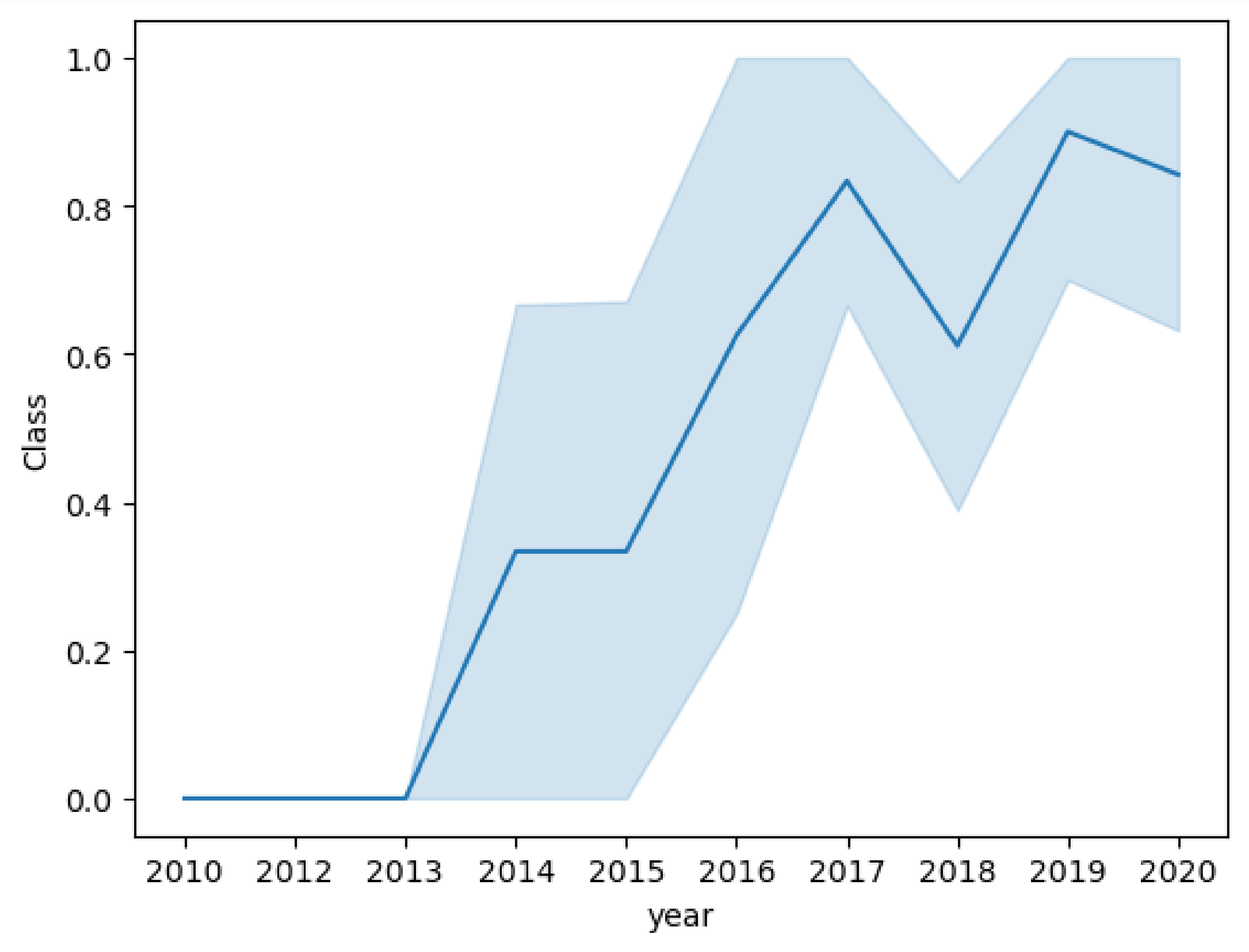
- The heaviest payloads are sent to VLEO
- Successful launches to GTO falls precipitously with a payload greater than 5000 Kgs.



https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

Launch Success Yearly Trend

- Launch success increased every year except 2018
- Success doubled between 2014 and 2017



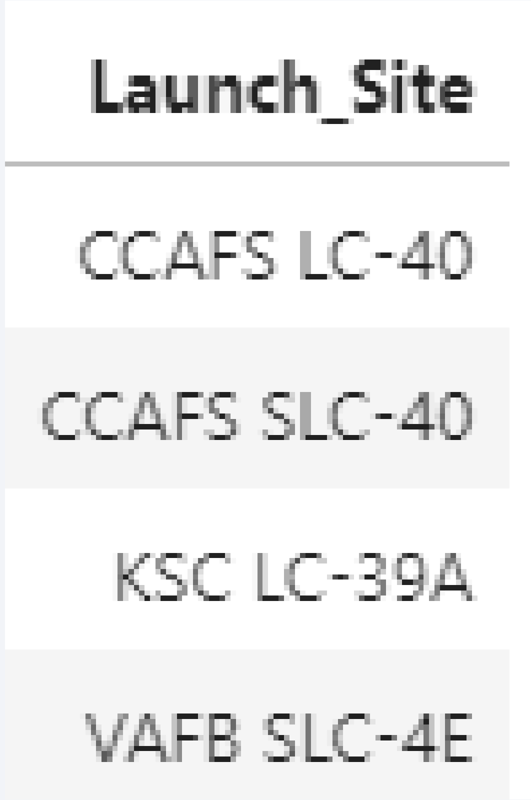
https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

All Launch Site Names

Display the names of the unique launch sites in the space mission

Query:

%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL ORDER BY "Launch_Site“



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

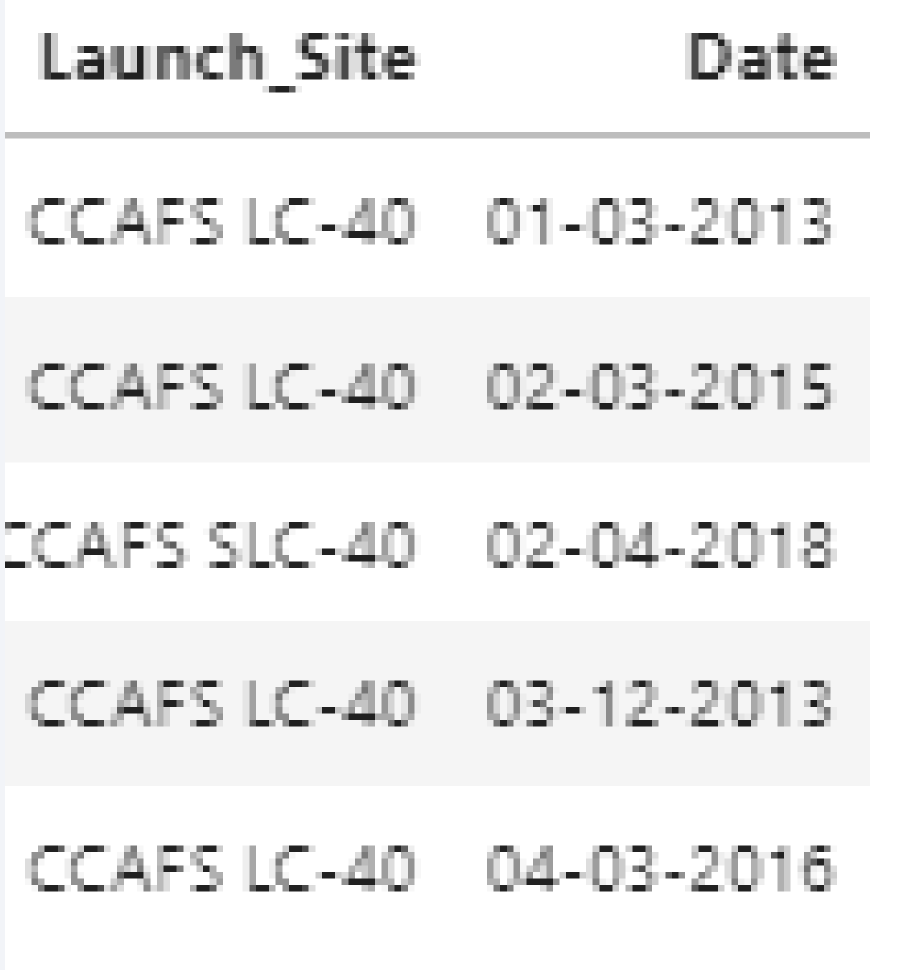
https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA’

Query:

%sql SELECT Launch_Site, DATE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE '%CCA%' ORDER BY DATE LIMIT 5

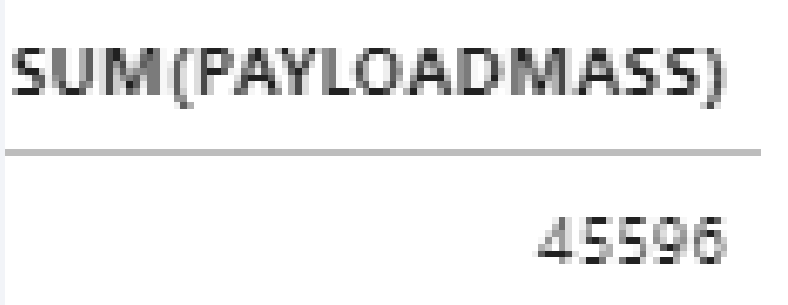


Launch_Site	Date
CCAFS LC-40	01-03-2013
CCAFS LC-40	02-03-2015
CCAFS SLC-40	02-04-2018
CCAFS LC-40	03-12-2013
CCAFS LC-40	04-03-2016

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS)



SUM(PAYLOADMASS)
45596

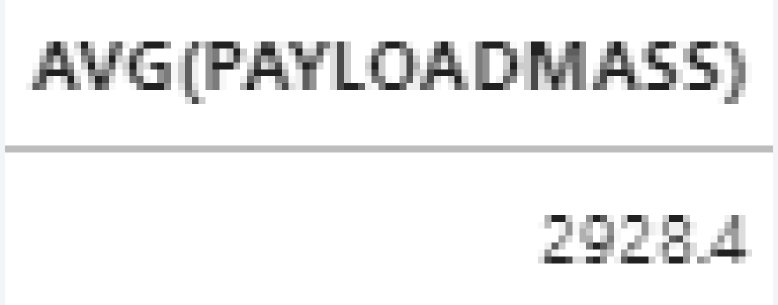
Query:

%sql SELECT SUM(PAYLOADMASS)
FROM SPACEXTBL WHERE
CUSTOMER = "NASA (CRS)“

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Average Payload Mass by F9 v1.1

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



AVG(PAYLOADMASS)
2928.4

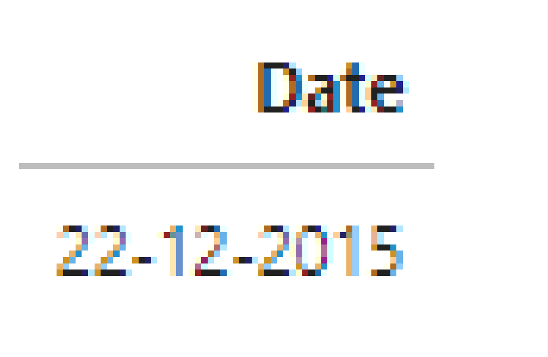
Query:

```
%sql SELECT AVG(PAYLOADMASS)
FROM SPACEXTBL WHERE
BOOSTER_VERSION = "F9 v1.1"
```

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was achieved.



Date
22-12-2015

Query:

```
%sql SELECT DATE FROM SPACEXTBL
WHERE LANDINGLOC = "Success
(ground pad)" ORDER BY DATE DESC
LIMIT 1
```

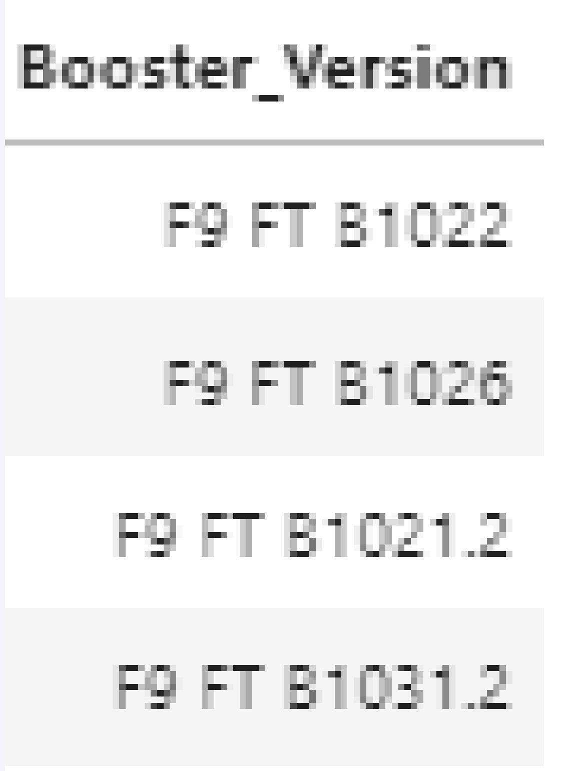
https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Query:

%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDINGLOC = "Success (drone ship)" AND PAYLOADMASS BETWEEN 4001 AND 5999



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

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Number of Successful and Failure Mission Outcomes

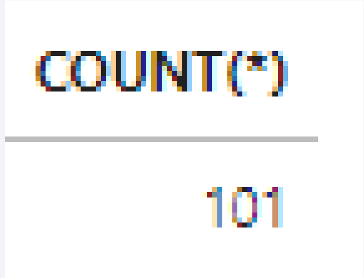
List the total number of successful and failure mission outcomes

Queries:

1. %sql SELECT COUNT(*) FROM SPACEXTBL

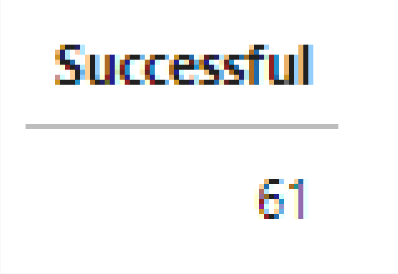
2. %sql SELECT COUNT(*) as Successful FROM SPACEXTBL WHERE MISSION_OUTCOME like '%Success%' AND LANDINGLOC like '%Success%’

3. #therefore, total failures = 101 (total) - 61 (successful) = 50 (failure)



COUNT(*)
101

-



Successful
61

=

Failure

50

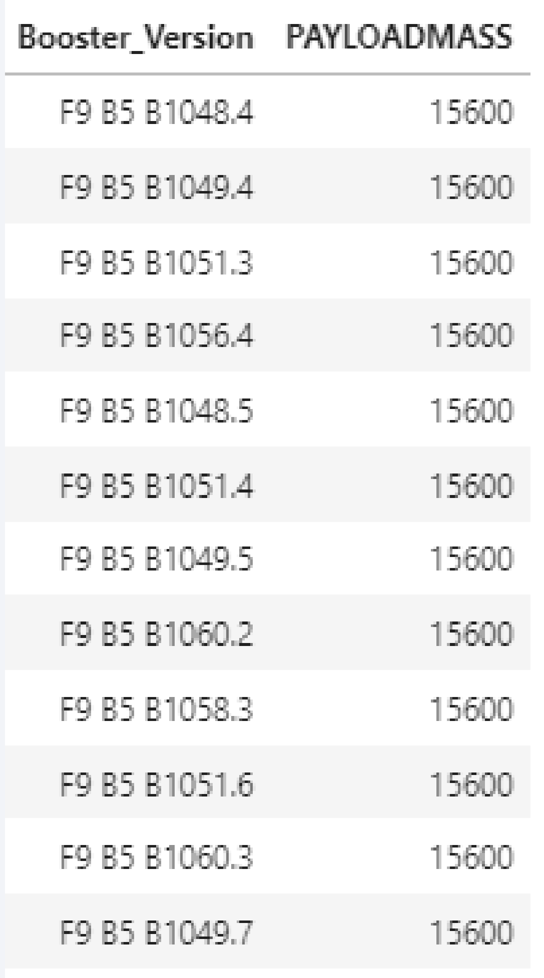
https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Boosters Carried Maximum Payload

List the names of the booster versions which have carried the maximum payload mass.

Query (using subquery):

```
%sql SELECT DISTINCT BOOSTER_VERSION,
PAYLOADMASS FROM SPACEXTBL WHERE
PAYLOADMASS = (SELECT MAX(PAYLOADMASS)
FROM SPACEXTBL)
```



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

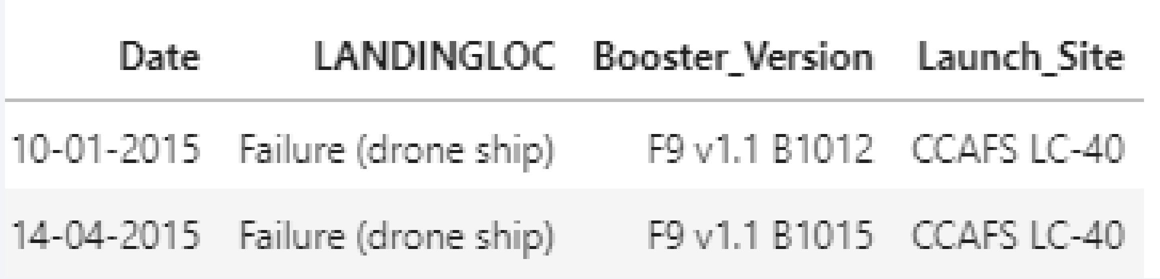
https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

2015 Launch Records

List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.

Query:

```
%sql SELECT DATE, LANDINGLOC,
BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL WHERE DATE like
'%2015' AND LANDINGLOC =
"Failure (drone ship)"
```



Date	LANDINGLOC	Booster_Version	Launch_Site
10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
14-04-2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

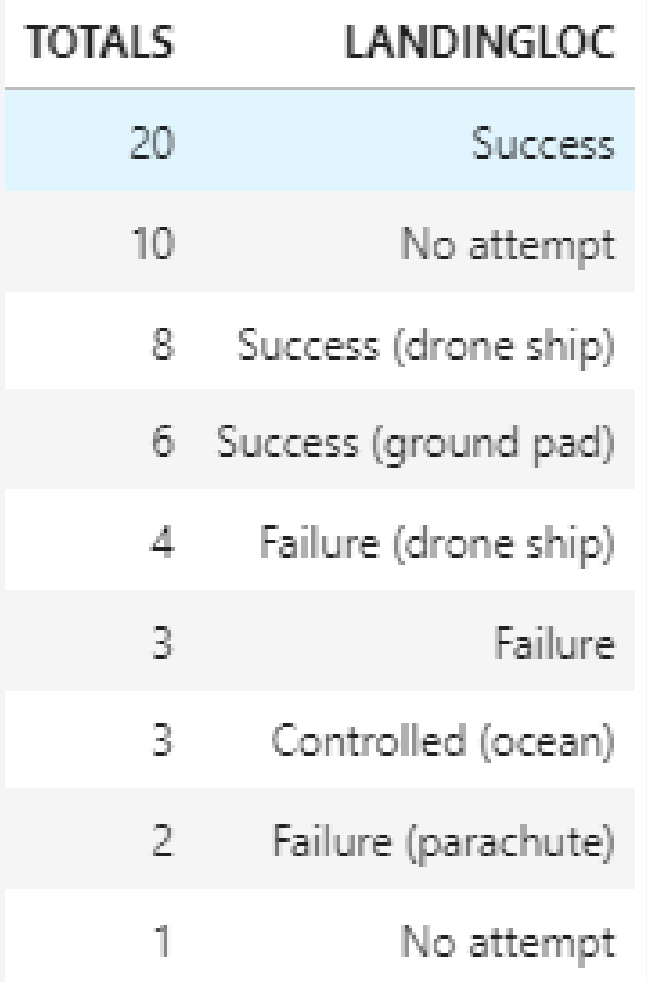
https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

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

Rank the count of successful and Failed landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Query:

%sql SELECT COUNT(LANDINGLOC) AS TOTALS, LANDINGLOC FROM SPACEXTBL WHERE DATE BETWEEN "04-06-2010" AND "20-03-2017" GROUP BY LANDINGLOC ORDER BY COUNT(LANDINGLOC) DESC



TOTALS	LANDINGLOC
20	Success
10	No attempt
8	Success (drone ship)
6	Success (ground pad)
4	Failure (drone ship)
3	Failure
3	Controlled (ocean)
2	Failure (parachute)
1	No attempt

https://github.com/jasonhooy/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

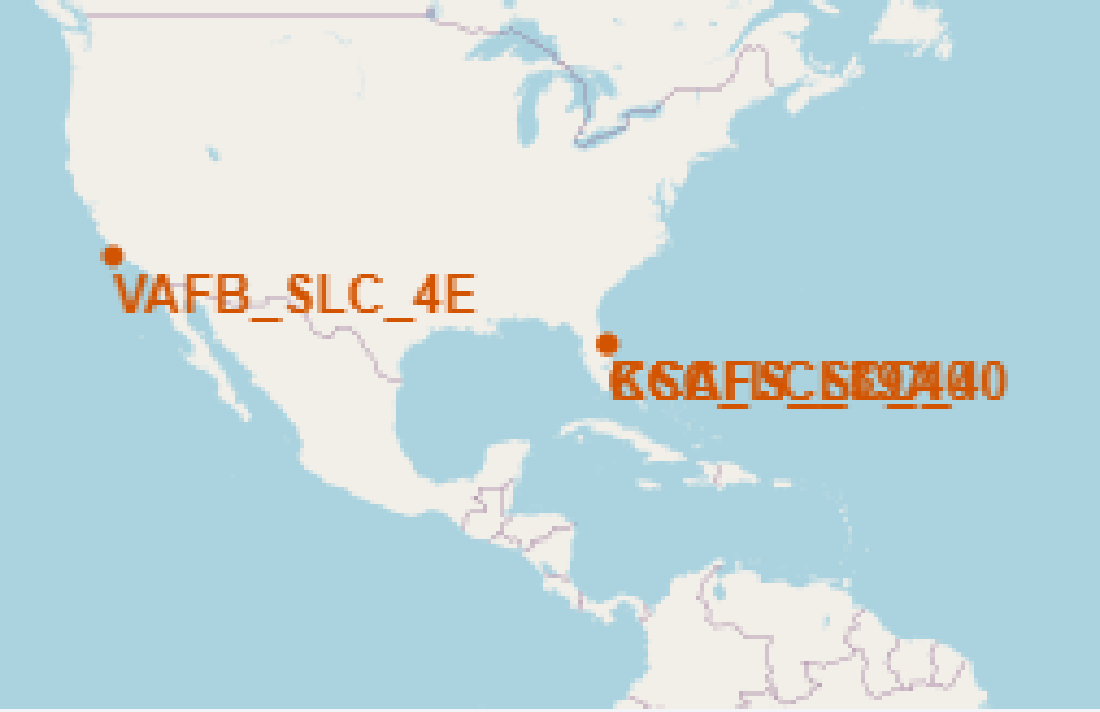
Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites

Locations depicted:

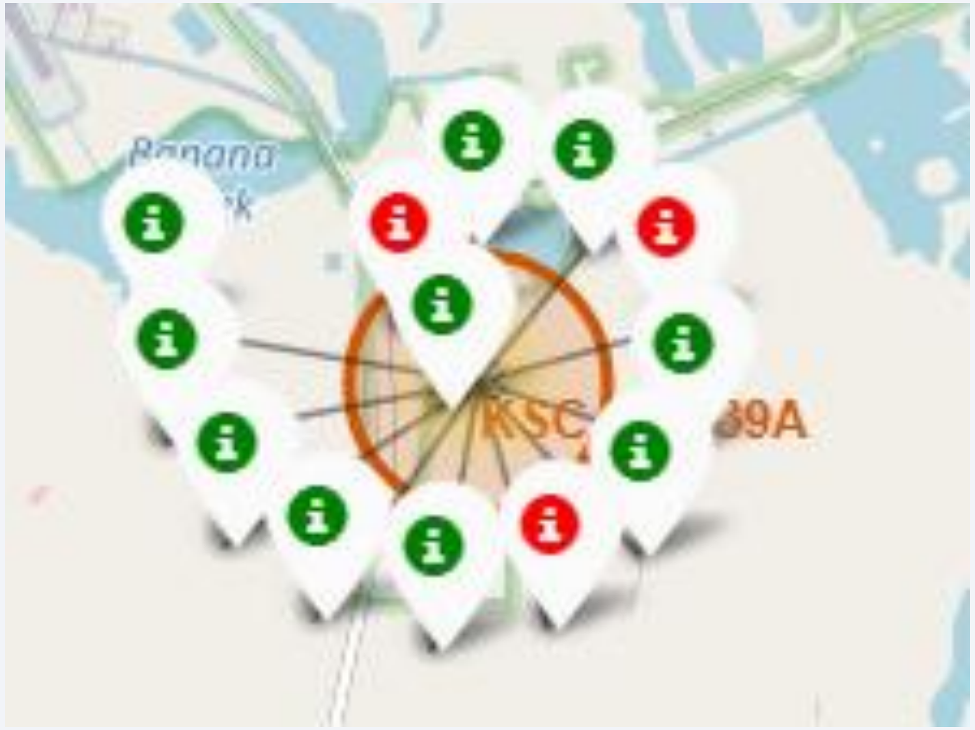
- CCAFS_LC_40, CCAFS_SLC_40, KSC_LC_39A, VAFB_SLC_4E
- Each site is located near the ocean
- Each site is roughly the same latitude and approximate distance from the equator.



https://github.com/jasonhooy/DSCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

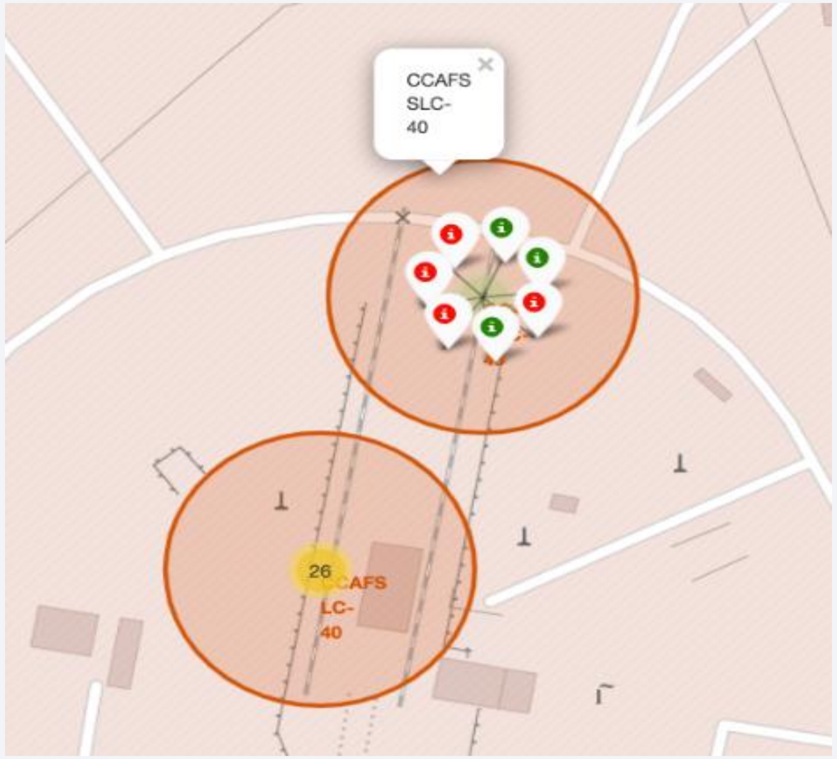
KSC LC-39A Cluster Map

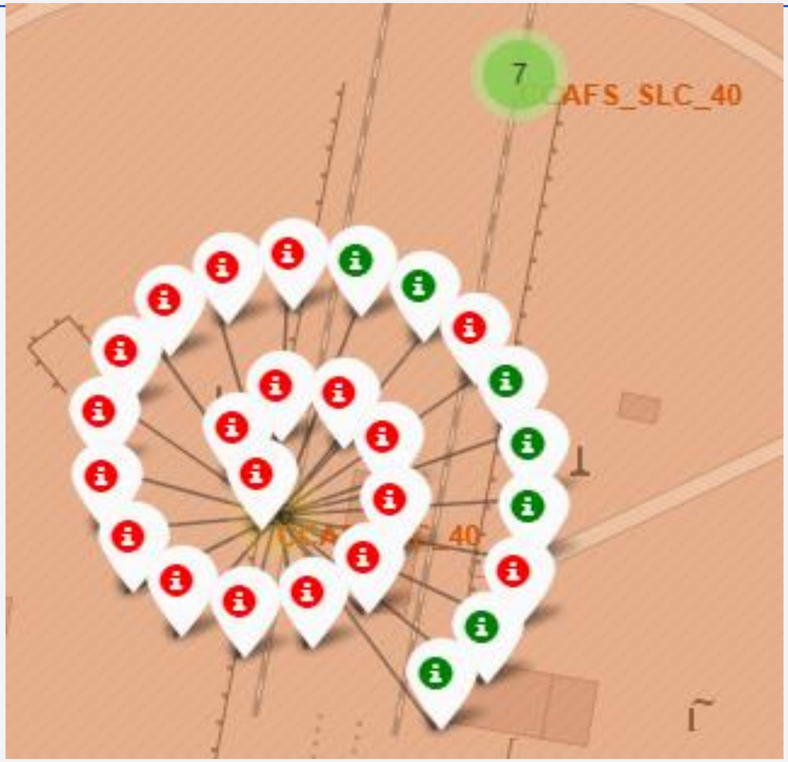
- This cluster marker interactively displays the number of successful/failed launches at launch site KSC LC-39A.
- Successful launches are colored green
- Failed launches are colored red
- Of the 13 launches, 3 failed
- **KSC LC 39A is the best performing launch site**



https://github.com/jasonhooy/DSCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

AFS SLC 40 & AFS LC 40 Cluster Maps



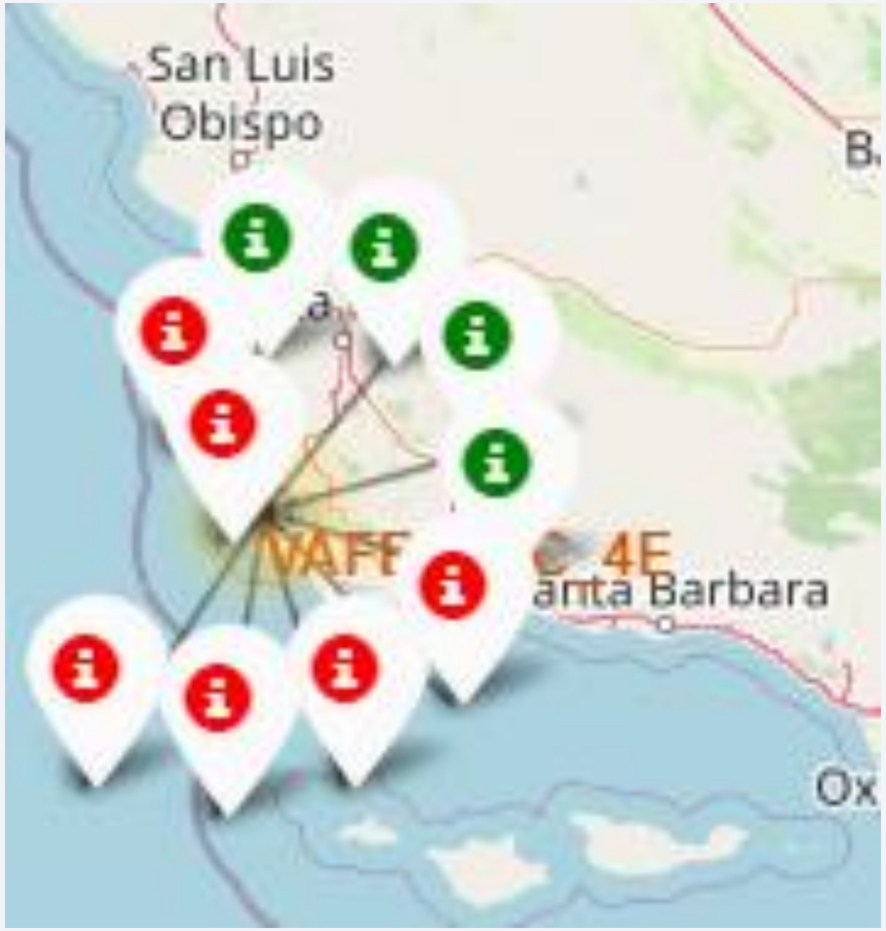


- Left- CCAFS SLC 40 with 7 launches, 4 of which failed
- Right- AFS LC 40 with 26 launches, 19 of which failed

https://github.com/jasonhooy/DSCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

AFS SLC 40 & AFS LC 40 Cluster Maps

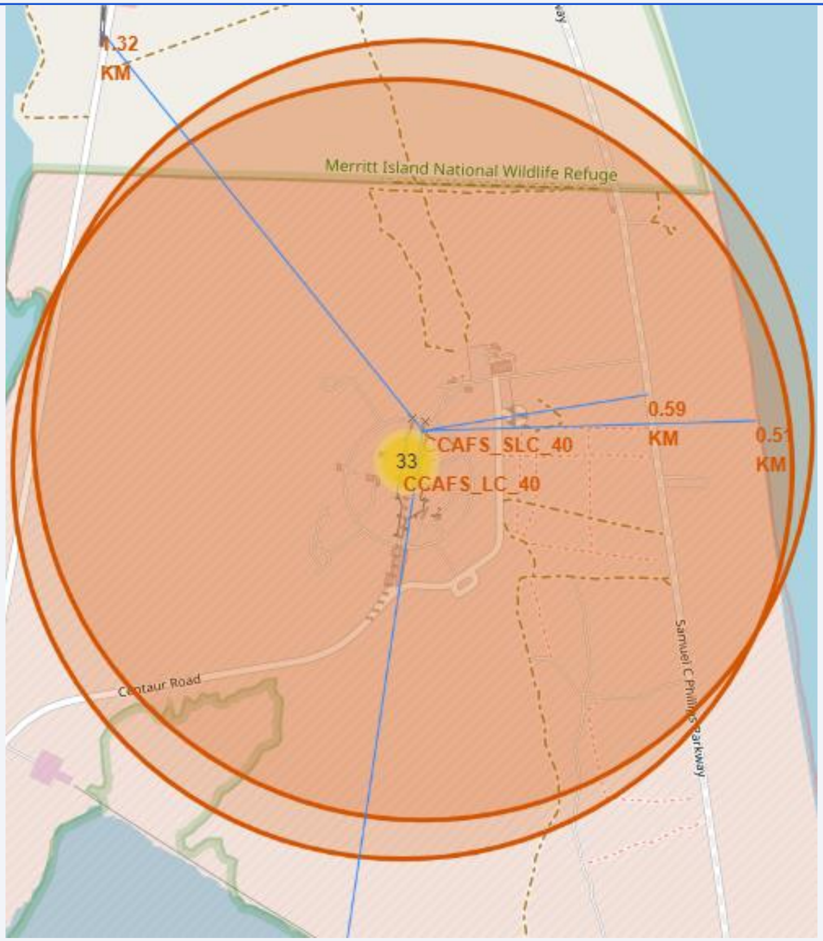
- The VAFB SLC 4E cluster map interactively shows 10 launches
- Of the 10 launces, 4 succeeded and 6 failed



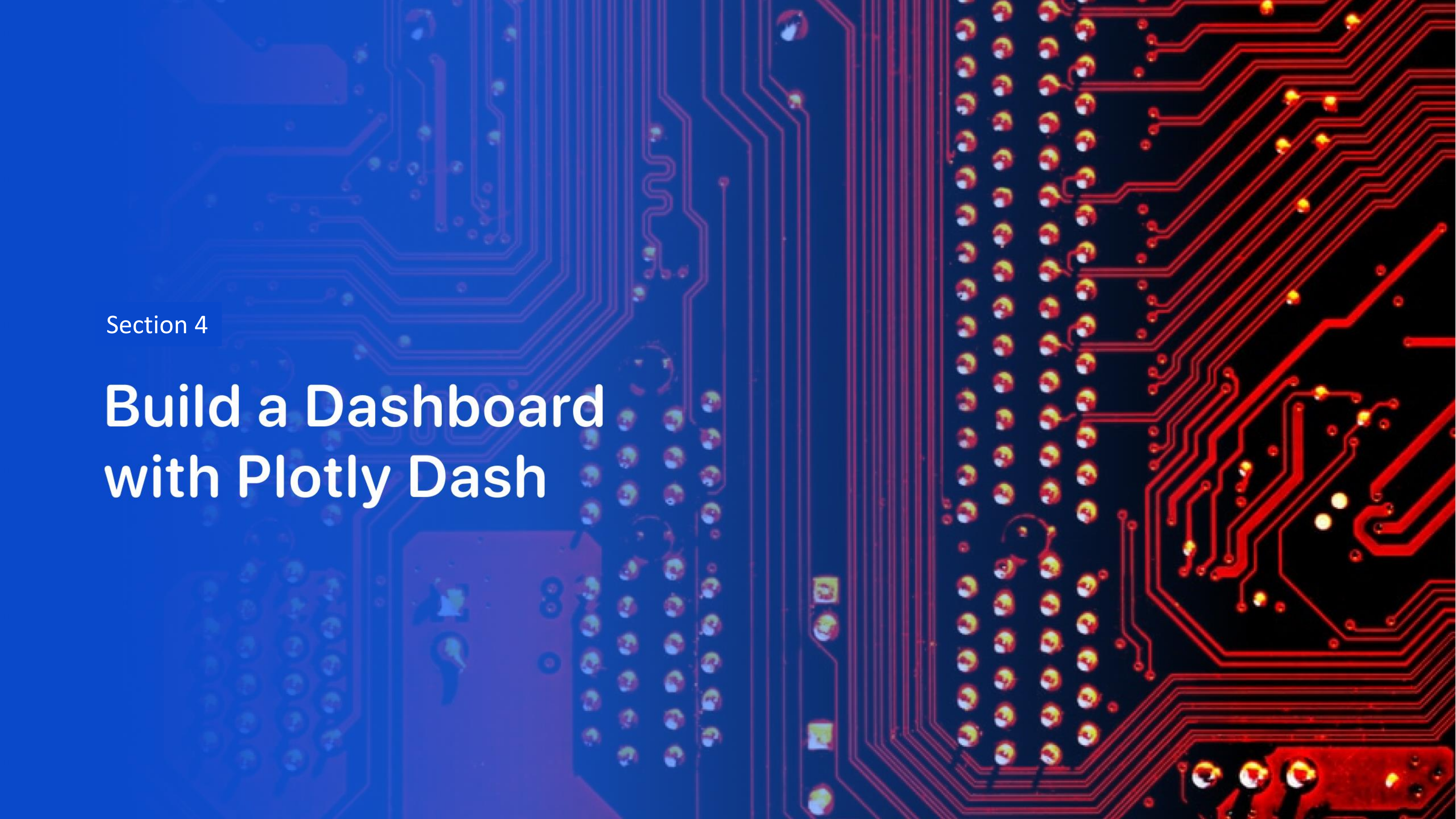
https://github.com/jasonhooy/DSCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Launch Site Proximity to Civil Infrastructure

- The interactive polyline map shows the CAFS SLC 40 launch site is:
- 1.32 Km from a rail line
- 0.58 Km from a highway
- 0.51 Km from the ocean
- 13.36 Km from the nearest city
- Each site is roughly the same distance from civil and geographic features listed above


1.32 KM
Merritt Island National Wildlife Refuge
0.59 KM
0.51 KM
33
CCAFS_SLC_40
CCAFS_LC_40
Centaur Road
Samuel C Phillips Parkway

https://github.com/jasonhooy/DSCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

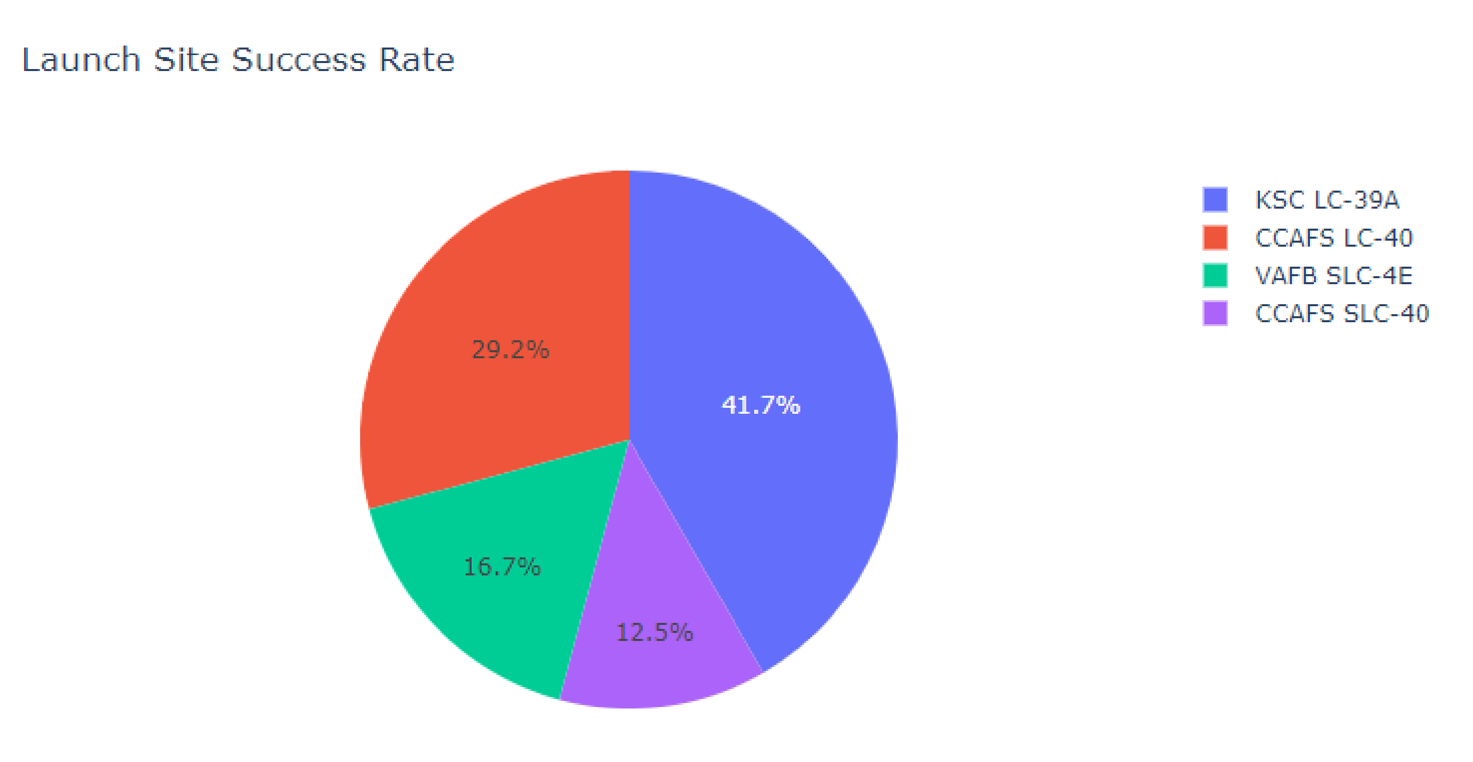


Section 4

Build a Dashboard with Plotly Dash

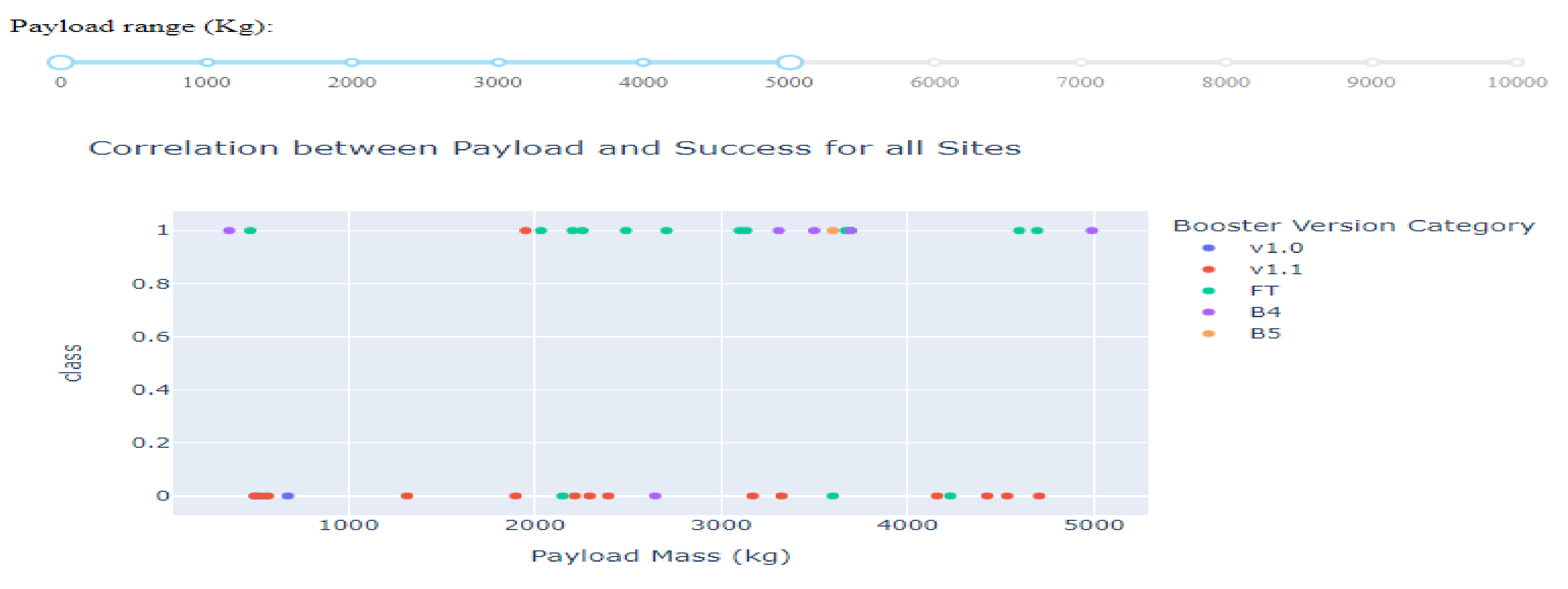
SpaceX Launch Records Dashboard

KSC LC 39 is not only the most successful launch site, but as a function of the whole it makes up the plurality of all success



https://github.com/jasonhooy/DSCapstone/commit/361ea327f4df67c989d421c4d4c5ccaa968d1cd3

Payload vs. Launch Outcome < 5000 Kgs

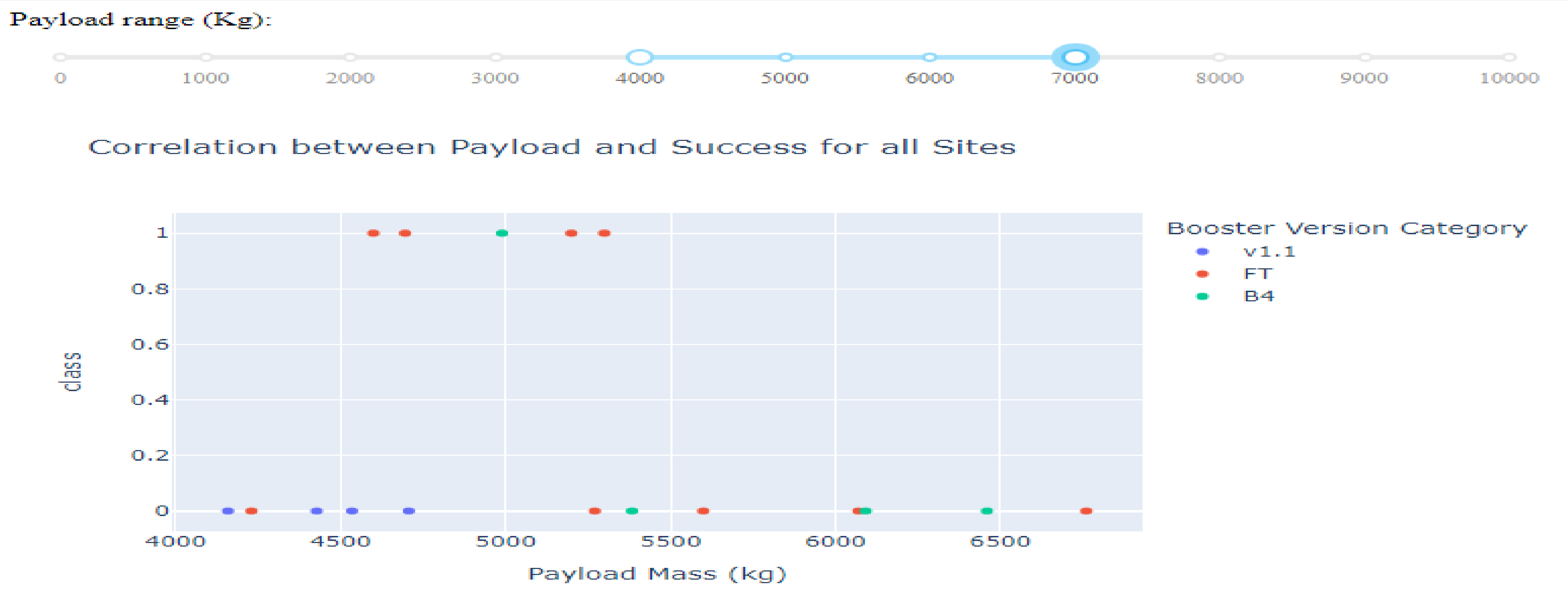


Payload range (Kg):

The FT Booster Version with a payload between 2000 and 3500 Kgs succeeds 87.5% of the time.

https://github.com/jasonhooy/DSCapstone/commit/361ea327f4df67c989d421c4d4c5ccaa968d1cd3

Payload vs. Launch Outcome 4000 < X < 7500 Kgs



The FT Booster Version with a payload greater than 5500 Kgs fails 100% of the time.

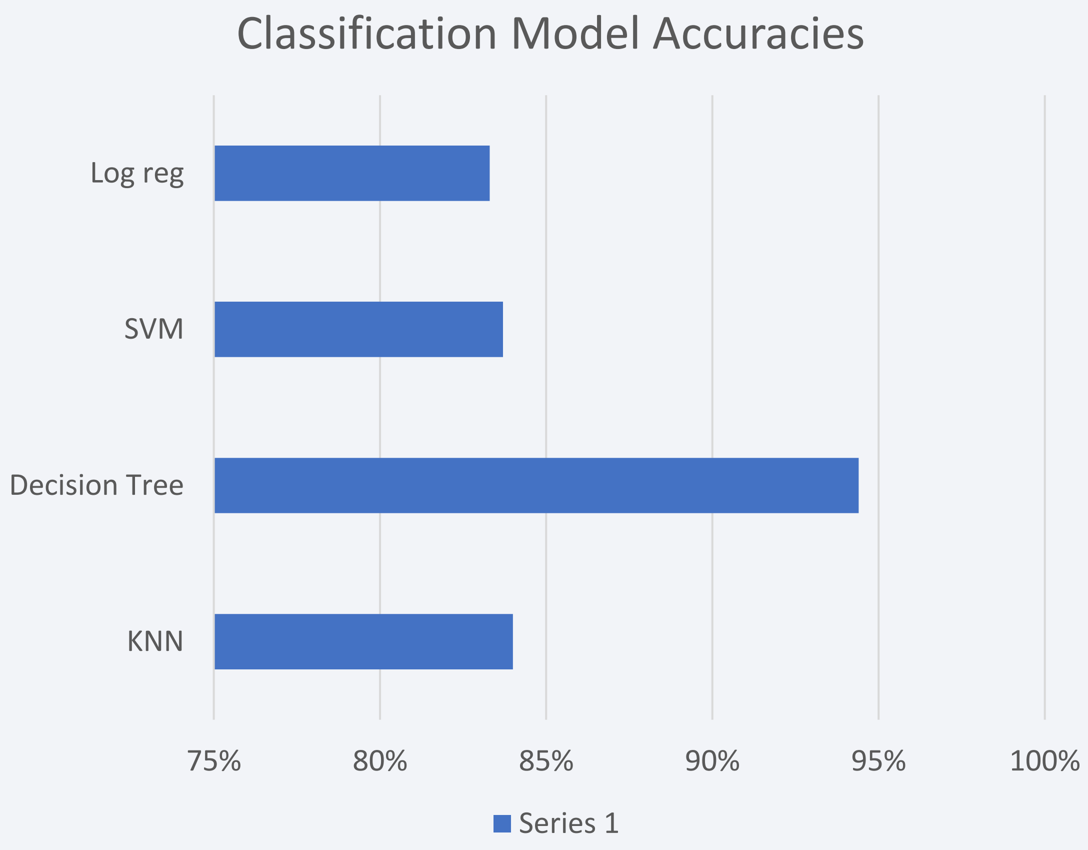
https://github.com/jasonhooy/DSCapstone/commit/361ea327f4df67c989d421c4d4c5ccaa968d1cd3

Section 5

Predictive Analysis (Classification)

Classification Accuracy

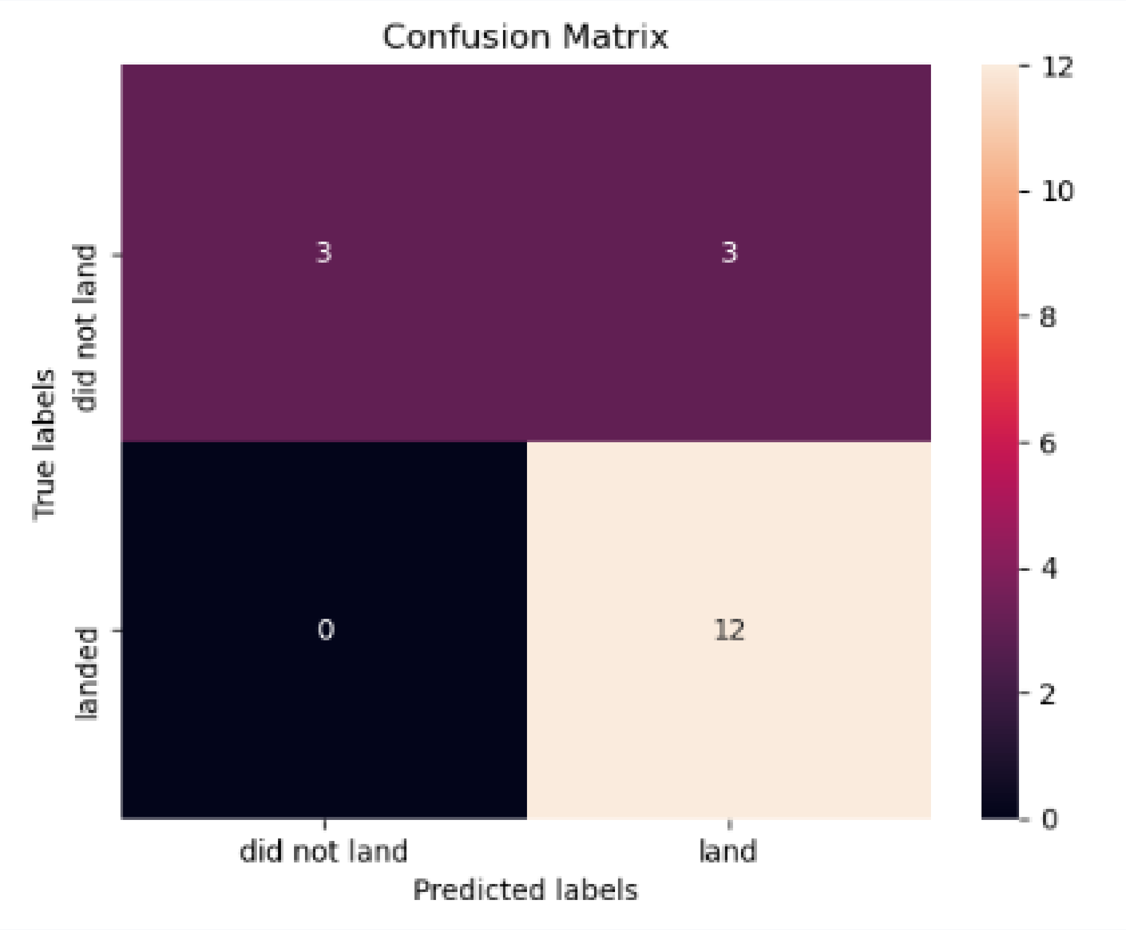
- The Decision Tree classification model performed 10% better than the next best model, KNN.
- All Log Reg, SVM, and KNN held similar accuracies around 83-85%.



https://github.com/jasonhooy/DSCapstone

Confusion Matrix

- While a np.int64 error derailed some analysis, the decision tree showed the most promising confusion matrix with an accuracy of 94%.
- 15 of 18 predictions were correct.



https://github.com/jasonhooy/DSCapstone

Conclusions

- Launches to orbital space ES-L1, GEO, HEO, and SSO are 100% successful
- 1:4 of SpaceX's last 40 launches went to VLOE
- KSC LC 39A is the most successful launch site holding the widest variety of launches in terms of payload mass
- Booster Version F9 B5 B104(4, 5, & 6) are used for maximum payload missions
- Booster Version F9 B4 enjoys high success carrying a payload mass between 2,000-3,000 Kgs.

=

To maximize the success of an alternative to SpaceX, said company needs:

1. A launch site emulating KSC LC 39A,
2. To specialize in launches to ES-L1, GEO, HEO, and SSO,
3. Build rockets with Booster Version F9 B4
4. Carry launches with a payload mass between 2,000 & 3,000 Kgs.

Thank you!