IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jesse Tsunekawa
07/26/2023

Outline

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

Executive Summary

- Summary of methodologies

  Various python techniques were used to gather, clean, prepare, and analyze Rocket Launch data. This includes:

  -making API connections
  -Web Scraping and Parsing
  -DataFrame Transformation and Manipulation
  -Data Visualization with graphs, maps, and dashboards
  -Machine Learning algorithms

- Summary of all results

  Our algorithms showed consistency in predicting successful landings but no consistency in predicting failed landings. It risks predicting false positives.

Introduction

- ***Background:***
  SpaceX is able to launch its Falcon 9 rocket at a significantly lower cost than other competitors ($62 million vs. $165+ million). Its cost savings can be primarily attributed to its reuse of first stage of the rocket. A wealth of market cap knowledge could be gleaned if we could accurately predict whether the first stage of the rocket will successfully land or not.
- ***Objectives:***
  What factors are related to Successful Landings?
  Can we accurately predict whether a Landing will Fail?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Requesting from API Endpoint (SpaceX)
  - WebScraping (Wikipedia – Falcon 9/Falcon Heavy Launches)
- Perform data wrangling
  - Extract and Standardize data into tables, and Prep for analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create training/test sets, tune training parameters, fit test data, and evaluate model

Data Collection

- Data was collected using two different methodologies:

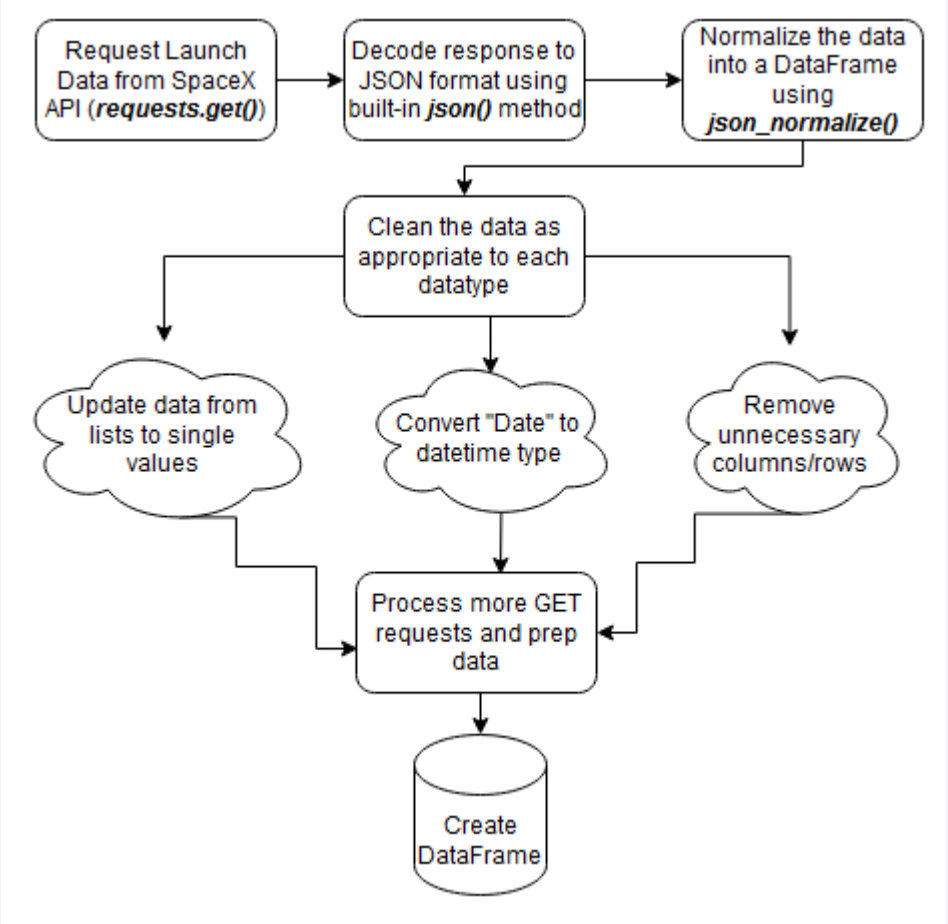
  -Requesting via API calls
  -Webscraping HTML data

- Data had to be prepared in a number of ways to make it useful

  -Parsing through strings to organize data into tables
  -Normalizing/Standardizing the data
  -Identifying data of interest
  -Removing excess/bloat data
  -More precise cleaning of data, such as setting datatype or applying lambda function to ready data in a useful format

Data Collection – SpaceX API

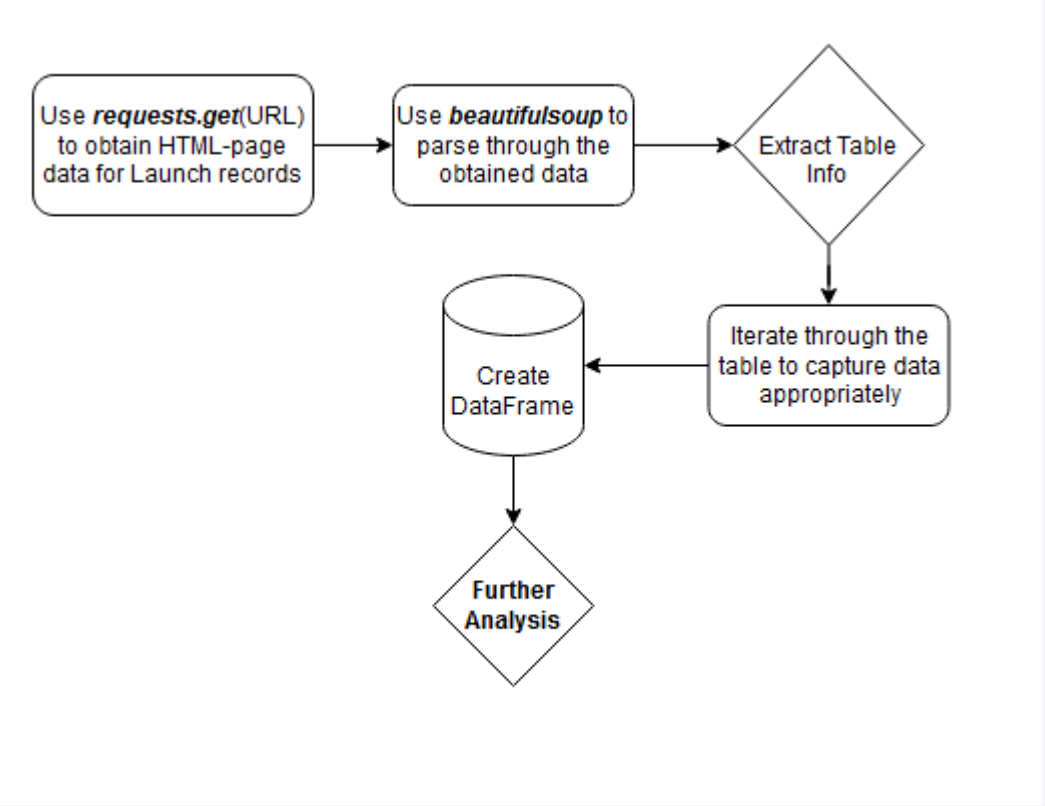
- Create a request to the SpaceX API and clean the received data



- Github: *https://github.com/Tocopherol/Data-Science-Capstone/blob/main/SpaceX-data-collection-api.ipynb*

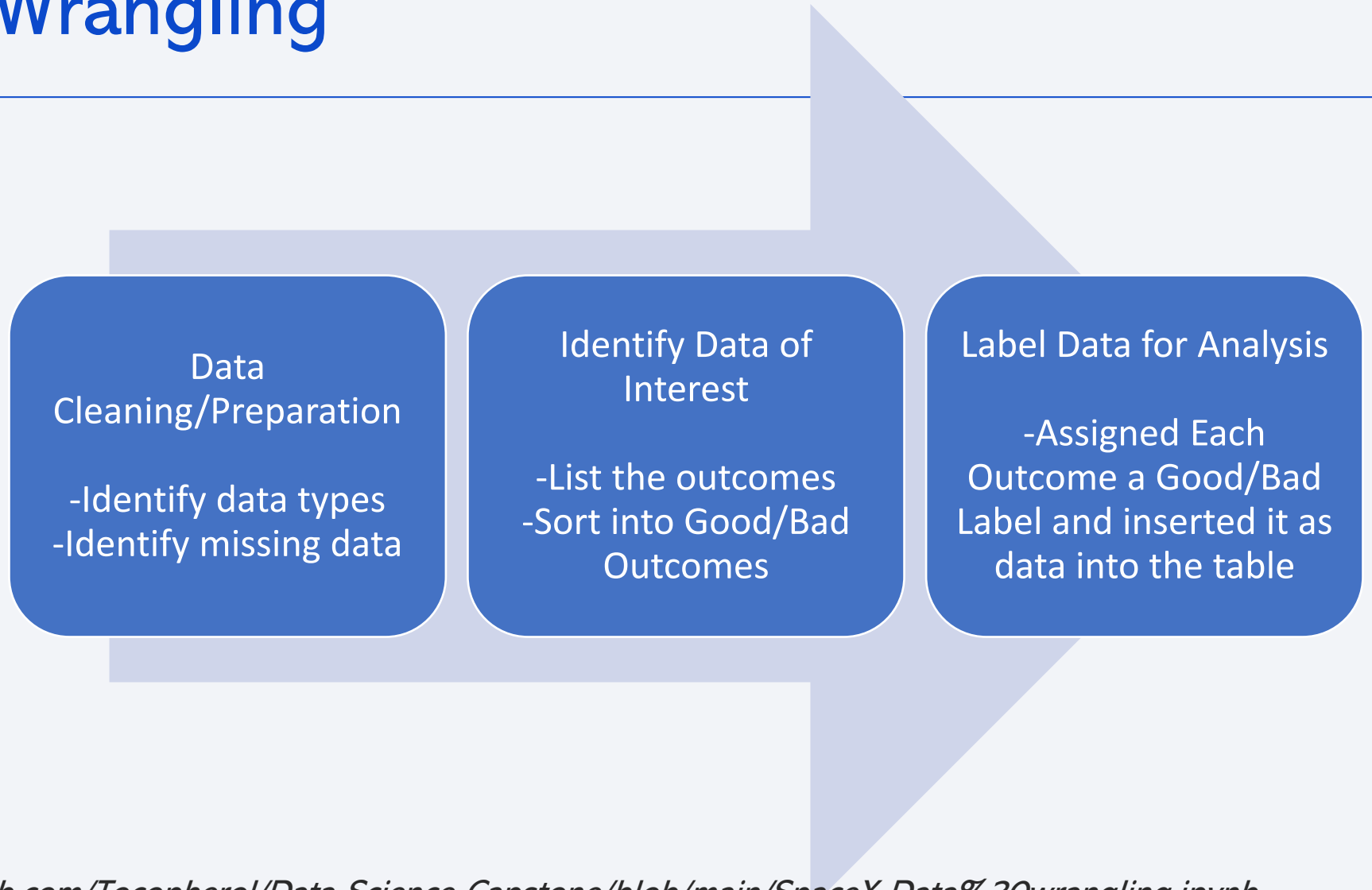
Data Collection - Scraping

- Extract Launch Records from Wikipedia



- Github: *https://github.com/Tocopherol/Data-Science-Capstone/blob/main/SpaceX-webscraping.ipynb*

Data Wrangling



Github:
https://github.com/Tocopherol/Data-Science-Capstone/blob/main/SpaceX-Data%20wrangling.ipynb

EDA with Data Visualization

- A number of charts were plotted in order to try to draw relationships between data. Some comparisons include:
  -Flight Number vs. Launch Site
  -Payload vs. Launch Site
  -Success Rate by Orbit type
  -Flight Number vs. Orbit type
  -Payload vs. Orbit type
- We were able to draw some conclusions, such as Payloads being more successful in certain orbit types

Github:
https://github.com/Tocopherol/Data-Science-Capstone/blob/main/SpaceX-eda-dataviz.ipynb

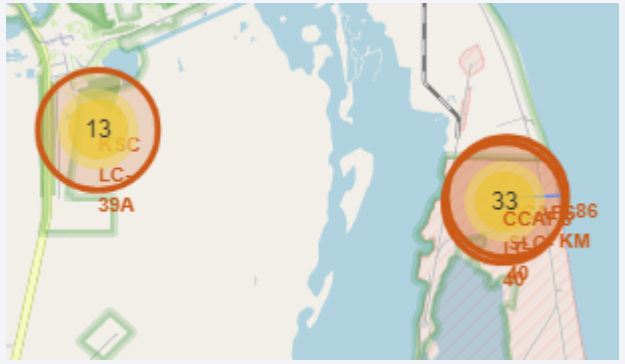
EDA with SQL

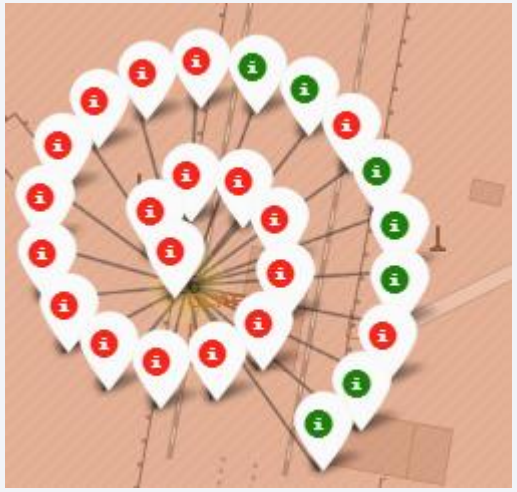
- SQL Queries were performed to answer some basic questions:
  -Create a List of Launch Sites
  -Find records limited to a few CCA Locations
  -Sum the total payload carried by a single Customer
  -Average the payloads carried with a single Booster Version
  -Find the date of the first successful landing on a ground pad
  -Find the Booster Versions used to successfully land payloads of a specific range
  -Count the Total Number of Successful and Failed missions
  -Find the Booster Versions used to carry maximum payload
  -Find particular failure records and the Month of the mission
  -Count the different Landing Outcomes and order by prevalence

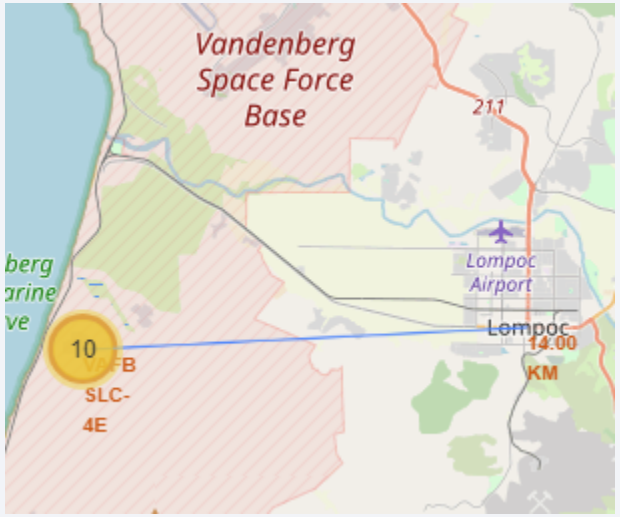
Github:
https://github.com/Tocopherol/Data-Science-Capstone/blob/main/SpaceX-eda-sql-notebook.ipynb

Build an Interactive Map with Folium

- An interactive map was built with Folium to better visualize and understand our data
  -Circles were added to designate launch sites
  -Colored Icon Clusters were added to visualize Launch Outcomes
  -Lines were added to explore distance to points of interests such as coastlines and cities.







Github:
https://github.com/Tocopherol/Data-Science-Capstone/blob/main/SpaceX_launch_sites_with_Folium.ipynb

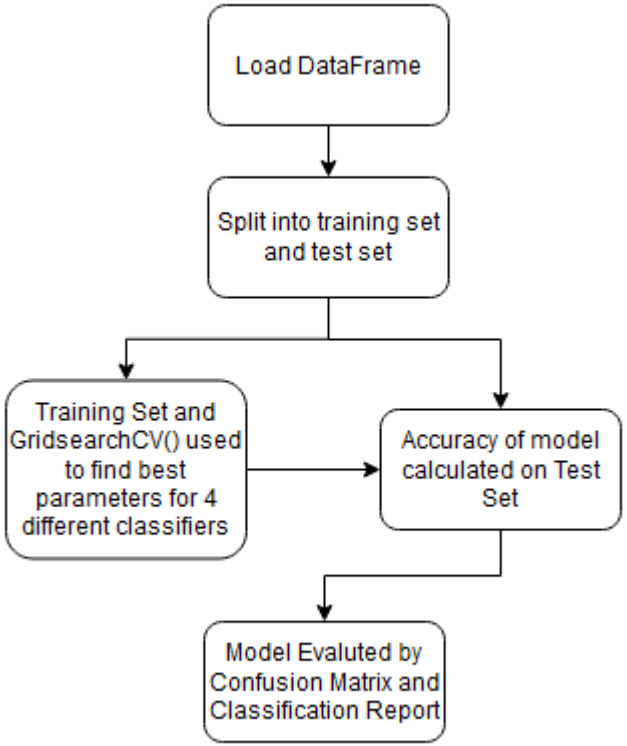
Build a Dashboard with Plotly Dash

- A Pie Chart was added to display success rates by Launch Site
  -This makes it easer to visualize which locations tend to have more successes
  -The chart can display success/failure for a single location, or it can break down all the successful launches by site
- A Scatter Plot was added to display the Success/Failure of payloads by Booster Type.
  -This chart can also be filtered down to show data for a single location
  -A slider was added to be able to filter down our range of interest to specific payloads

Github:
https://github.com/Tocopherol/Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

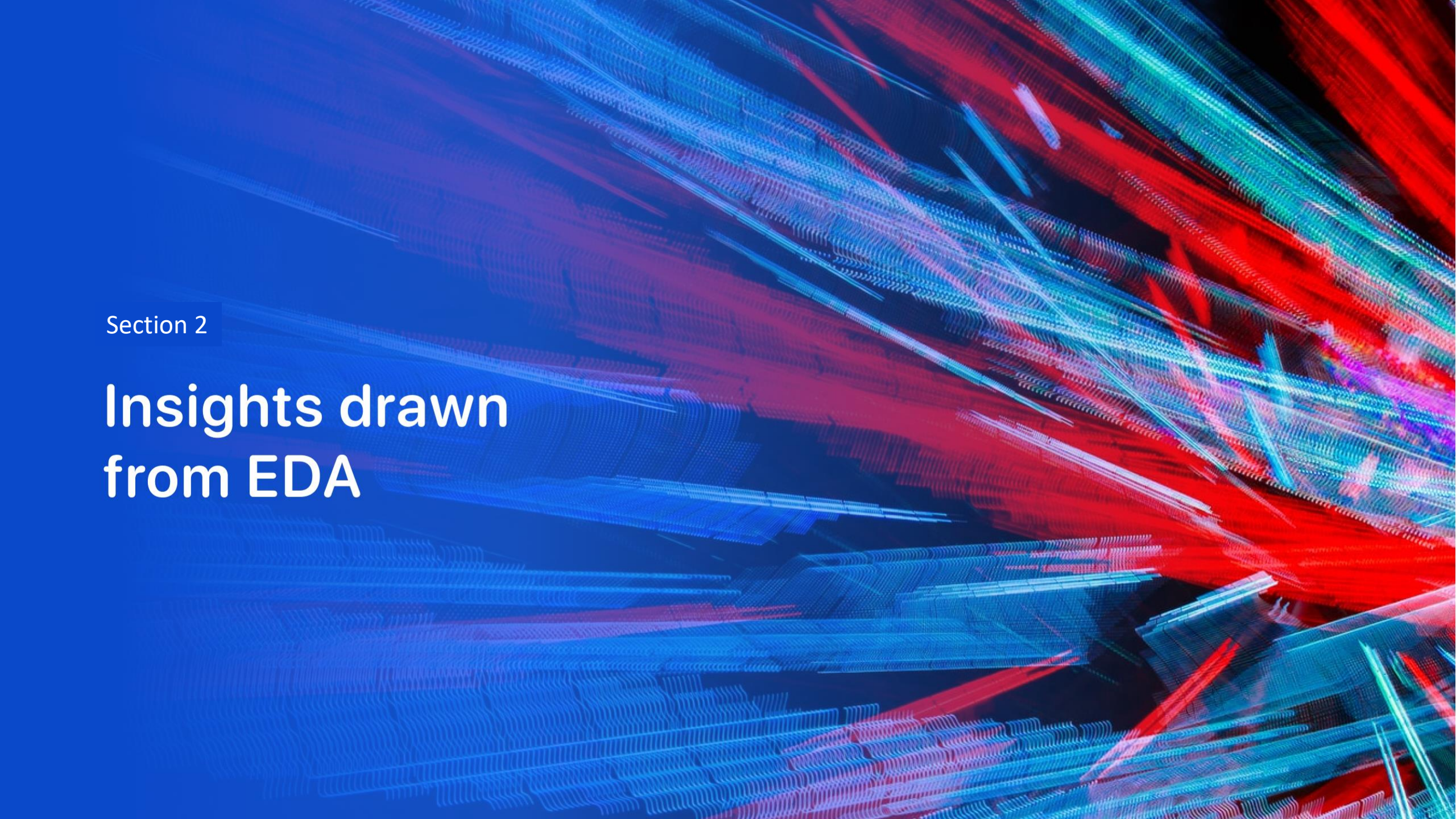
- Our prepared data was split into training and test sets
- Training Data was fit to 4 different classification models, using GridsearchCV() to tune for the best parameters
  -Logistic Regression
  -Support Vector Machine
  -Decision Tree Classifier
  -K-Nearest Neighbor
- Each of the 4 classification models were evaluated with the test data by creating a confusion matrix and generating a classification report



Github:
https://github.com/Tocopherol/Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

Results

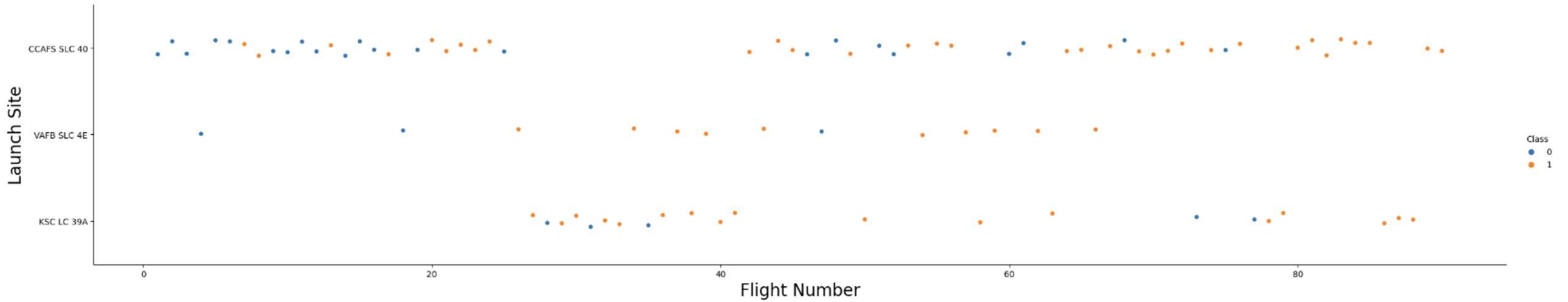
- Exploratory data analysis results
  - Section 3
- Interactive analytics demo in screenshots
  - Sections 4 and 5
- Predictive analysis results
  - Section 6


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

The success rate at the CCAFS Launch Site has improved as the flight number increases

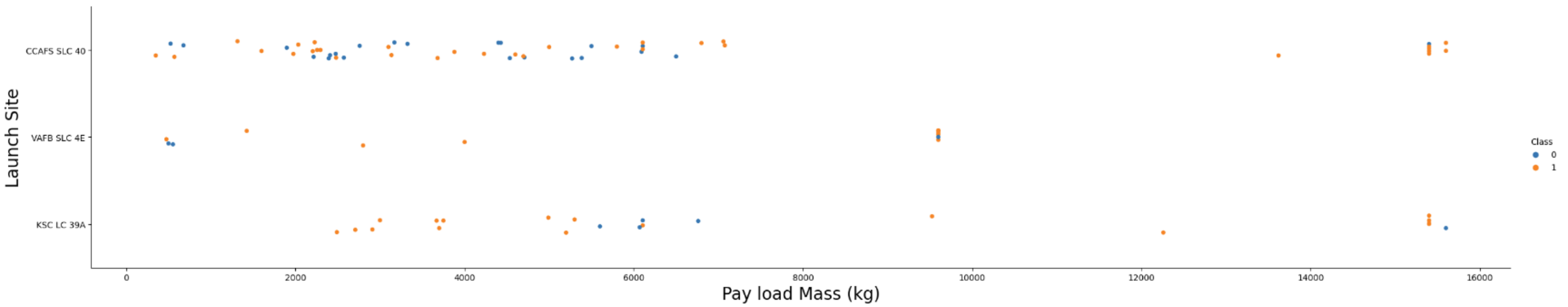
The success rate at the other 2 sites is harder to gauge, but the last 5 flights at each location were successful



Payload vs. Launch Site

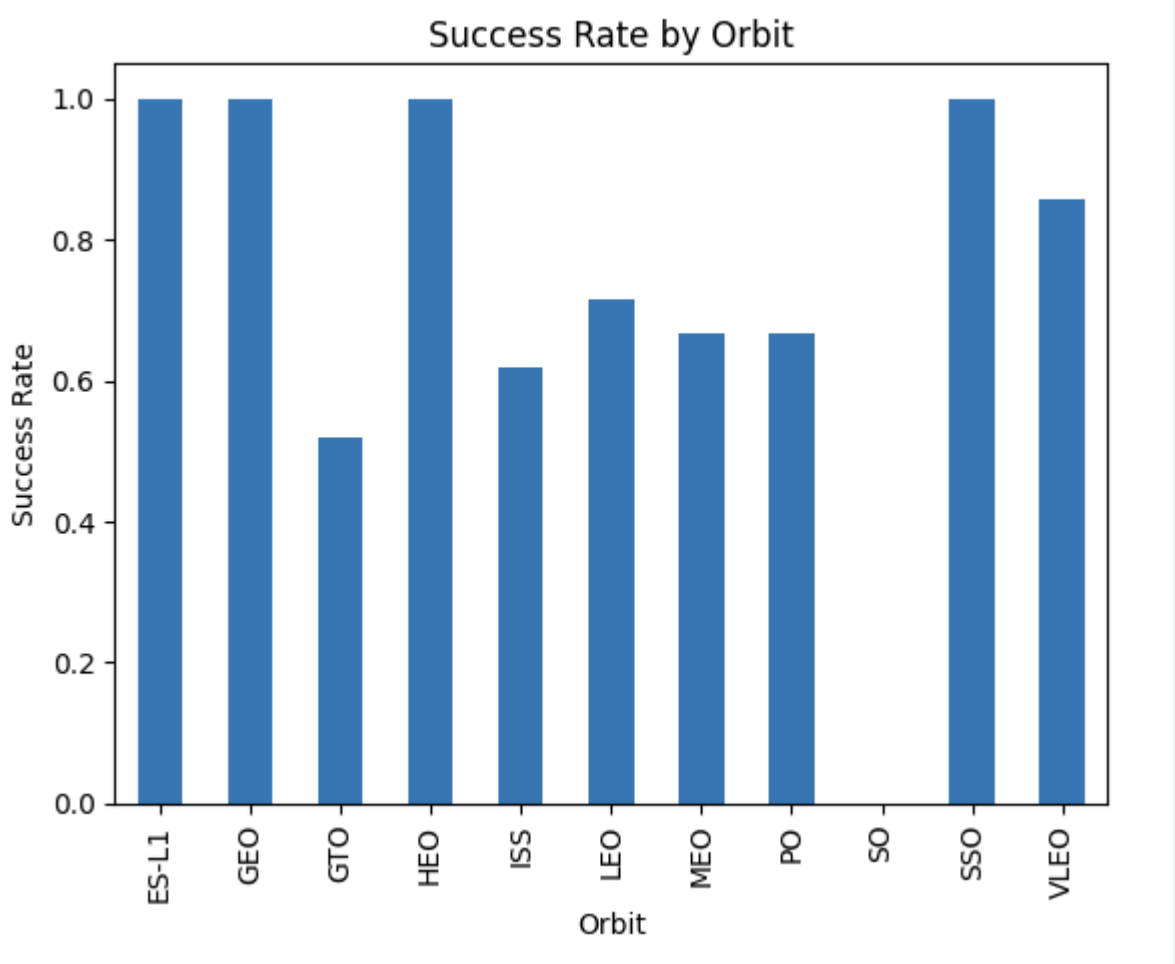
There are no payloads at the VAFB Site greater than 10,000kg

Payloads at the CCAFS Site jumped from <8,000kg to around 14,000+ kg



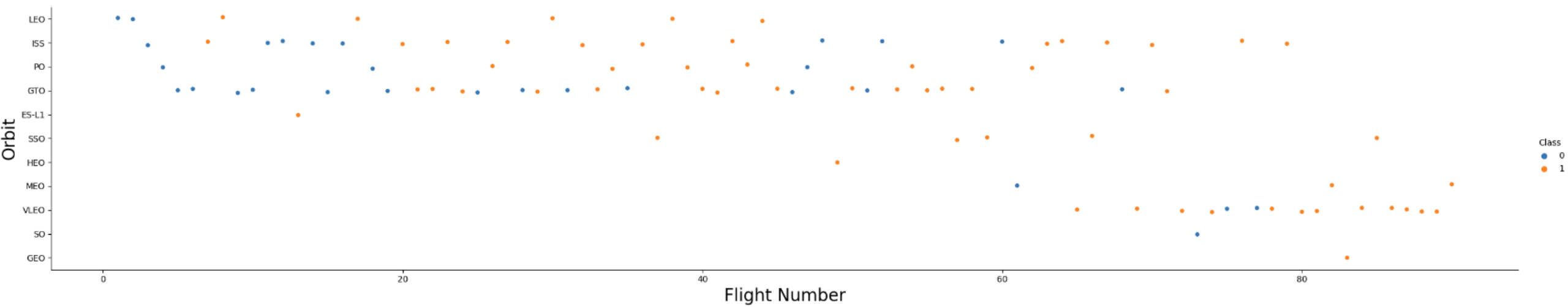
Success Rate vs. Orbit Type

- All flights into ES-L1, GEO, HEO, and SSO orbits were successful
- There were no successful flights into SO orbit



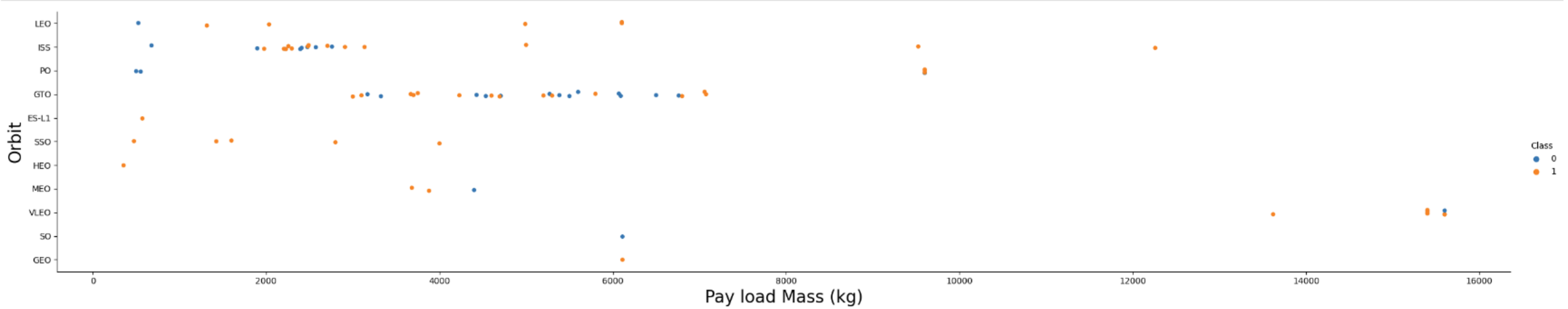
Flight Number vs. Orbit Type

- Launches into LEO orbit improved with Flight Number
- There does not appear to be a relationship between flight number and success for GTO orbit



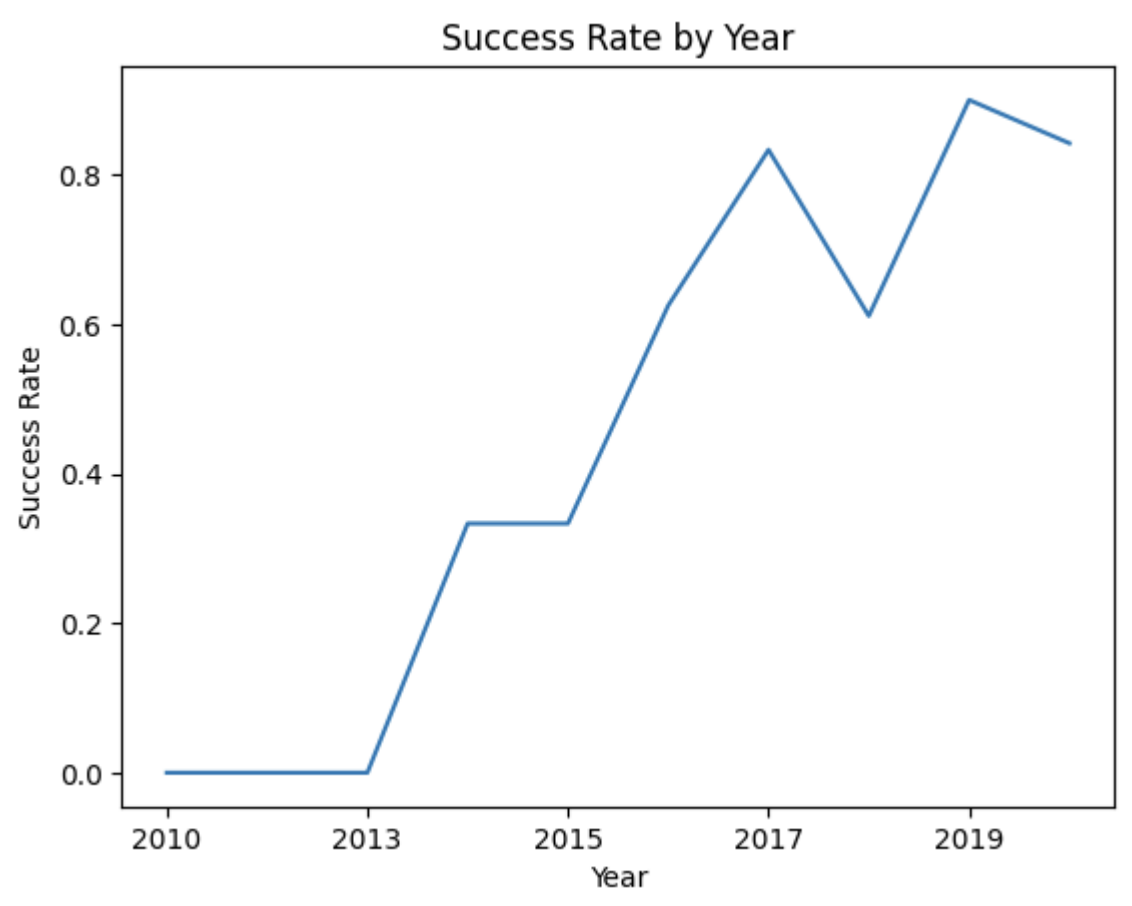
Payload vs. Orbit Type

- Heavy Payloads were more successful in LEO, ISS and PO orbits
- Success Rate in GTO does not appear to be related to Payload Mass



Launch Success Yearly Trend

- Overall, success rates have been trending up since 2013
- Average success rate of the last 4 years of data is over 75%



All Launch Site Names

- The Distinct Launch Site Names are:

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

SQL QUERY:

select distinct "Launch_Site" from SPACEXTBL

Launch Site Names Begin with 'CCA'

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- SQL QUERY:
  *select * from SPACEXTBL where "Launch_Site" like 'CCA%' LIMIT 5*

Total Payload Mass

- The Total Payload Mass carried by all flights from NASA (CRS) was found to be 45,596.0 kg

Customer	TOTAL_MASS
NASA (CRS)	45596.0

SQL QUERY:
SELECT "Customer", SUM("PAYLOAD_MASS__KG_") AS 'TOTAL_MASS' from SPACEXTBL WHERE "Customer"='NASA (CRS)' group by "Customer"

Average Payload Mass by F9 v1.1

- The Average Payload Mass of all flights using the F9 v1.1 Booster was found to be 2,928.4 kg

Booster_Version	AVERAGE_MASS
F9 v1.1	2928.4

SQL QUERY:

SELECT "Booster_Version", AVG("PAYLOAD_MASS__KG_") AS 'AVERAGE_MASS' from SPACEXTBL WHERE "Booster_Version"='F9 v1.1' group by "Booster_Version"

First Successful Ground Landing Date

- The first successful Ground Pad Landing occurred on Jan 8, 2018

Date
01/08/2018

SQL QUERY:

SELECT MIN("Date") as 'Date' from SPACEXTBL where "Landing_Outcome"='Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000 was found to be the following:

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

SQL QUERY:

SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_MASS__KG_">4000 and "PAYLOAD_MASS__KG_"<6000

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is shown below

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

--An anomaly likely attributed to an extra space/hidden character seems to have grouped one of the “Success” outcomes by itself

SQL QUERY:
Select "Mission_Outcome", Count("Mission_Outcome") as 'Number' from SPACEXTBL Group by "Mission_Outcome" ORDER BY Count("Mission_Outcome") DESC

Boosters Carried Maximum Payload

- A number of boosters carried the maximum payload mass

-The Maximum Payload mass was found to be 15,600 kg

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

SQL QUERY:
select distinct "Booster_Version" from SPACEXTBL where
"PAYLOAD_MASS__KG_" =(select max("PAYLOAD_MASS__KG_") from SPACEXTBL)

2015 Launch Records

- The failed Drone Ship Landings in 2015 are shown below:

Month	Booster_Version	Launch_Site	Landing_Outcome
October	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

SQL QUERY:
select case when substr("Date",4,2)='10' then 'October' when substr("Date",4,2)='04' then 'April' else substr("Date",4,2) end 'Month', "Booster_Version", "Launch_Site", "Landing_Outcome" from SPACEXTBL where substr("Date",7,4)='2015' and "Landing_Outcome"='Failure (drone ship)'

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

- The Landing Outcomes between 2010-06-04 and 2017-03-20 are shown here

Landing_Outcome	Number
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

SQL QUERY:

sql Select "Landing_Outcome", Count("Landing_Outcome") as 'Number' from SPACEXTBL WHERE "2010-06-04"<"Date"<"2017-03-20" Group by "Landing_Outcome" Order by Count("Landing_Outcome") DESC

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

- Here we see that Launch Sites are located near coastlines
- We might also infer that they are located as close to the Equator as possible, given their locations in the more Southern parts of the coastline while still being in the U.S.

Launch Site Outcomes

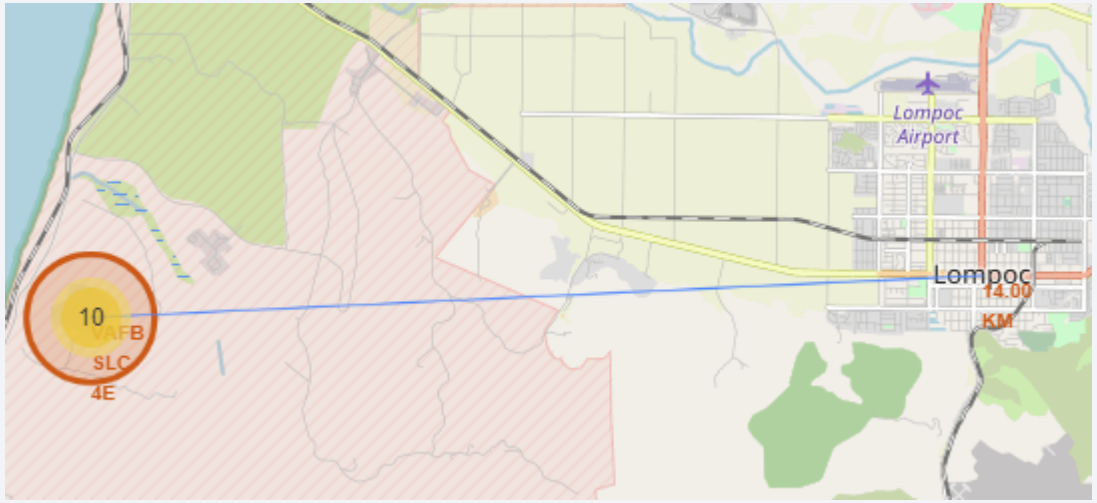
- Each Launch is marked on the appropriate Launch Site as either Failed (red icon) or Succeeded (green icon)

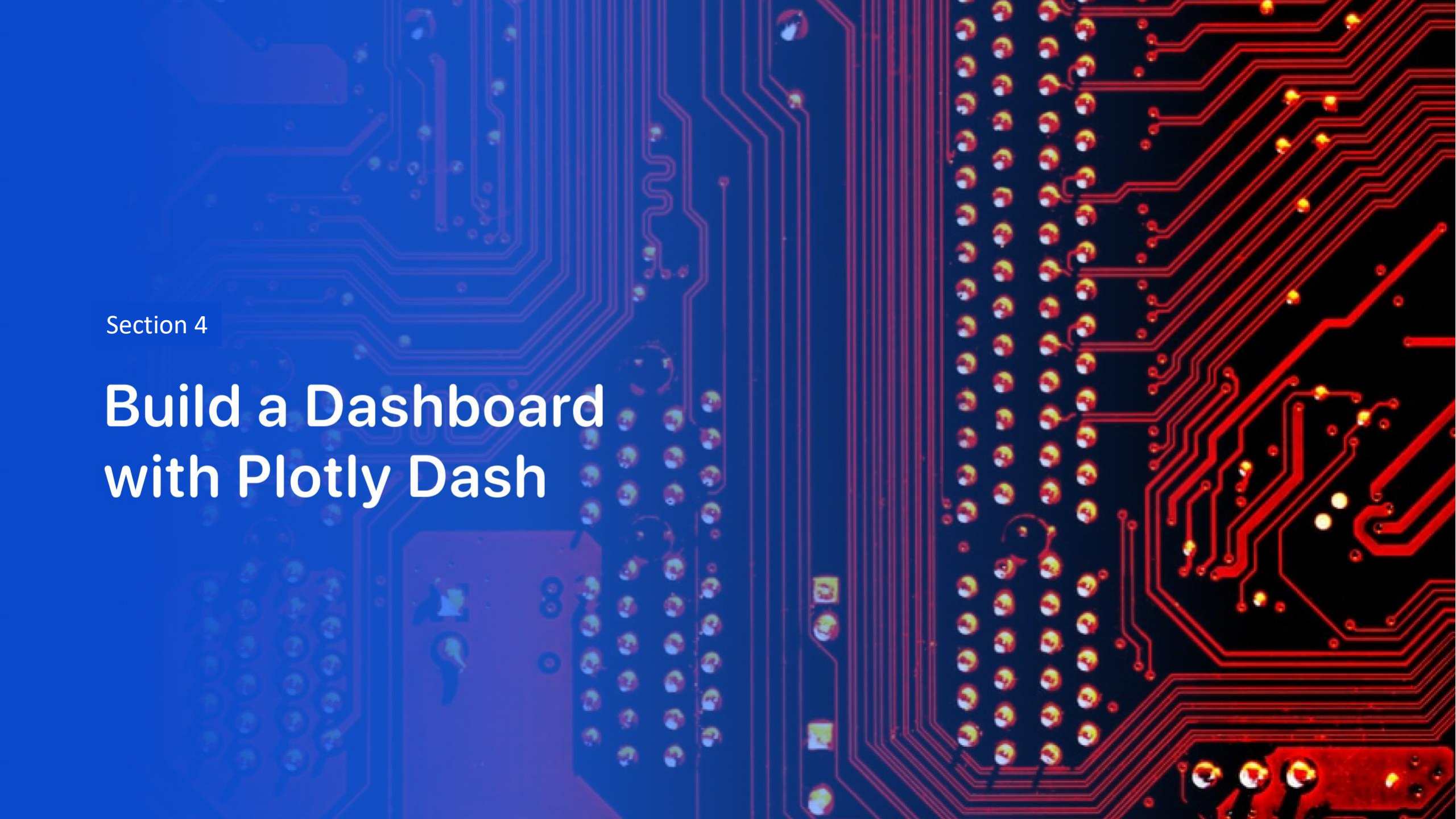
- We can easily visualize the performance at each launch site this way.
- At the example launch site on the right, we see a large number of failed launches

Launch Site Proximities

- Launch sites have close proximities to Coastlines and Railways, likely to aid in Disaster Prevention and Construction/Material Supply respectively.

- Launch sites are kept at a distance from major population centers, likely to reduce impact of launches on the general public



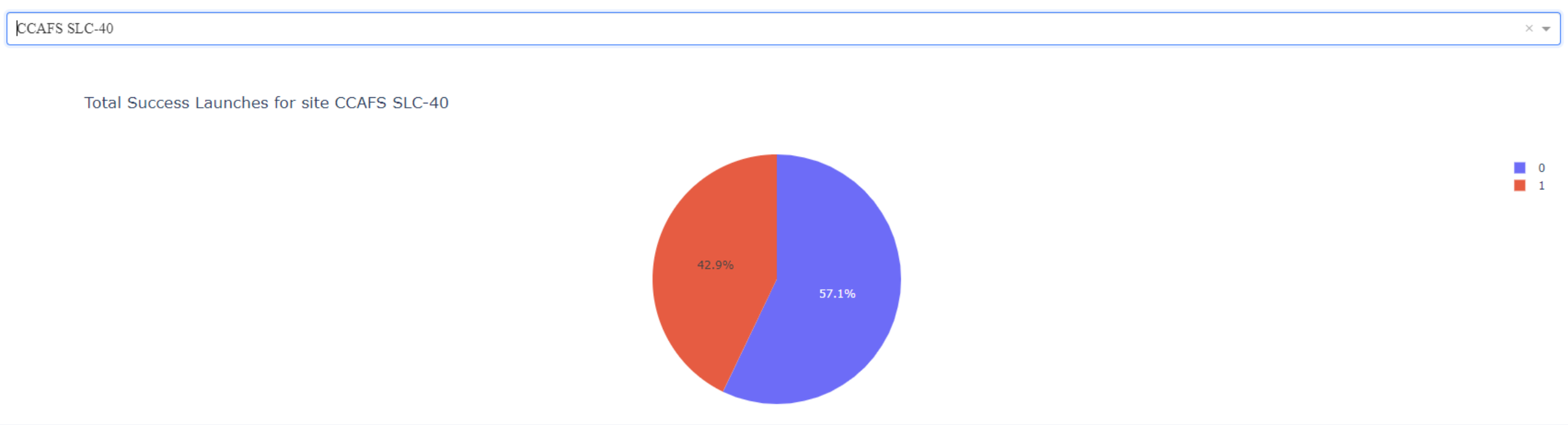

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches By Site

- Launch Site KSC LC-39A had the greatest number of successful launches

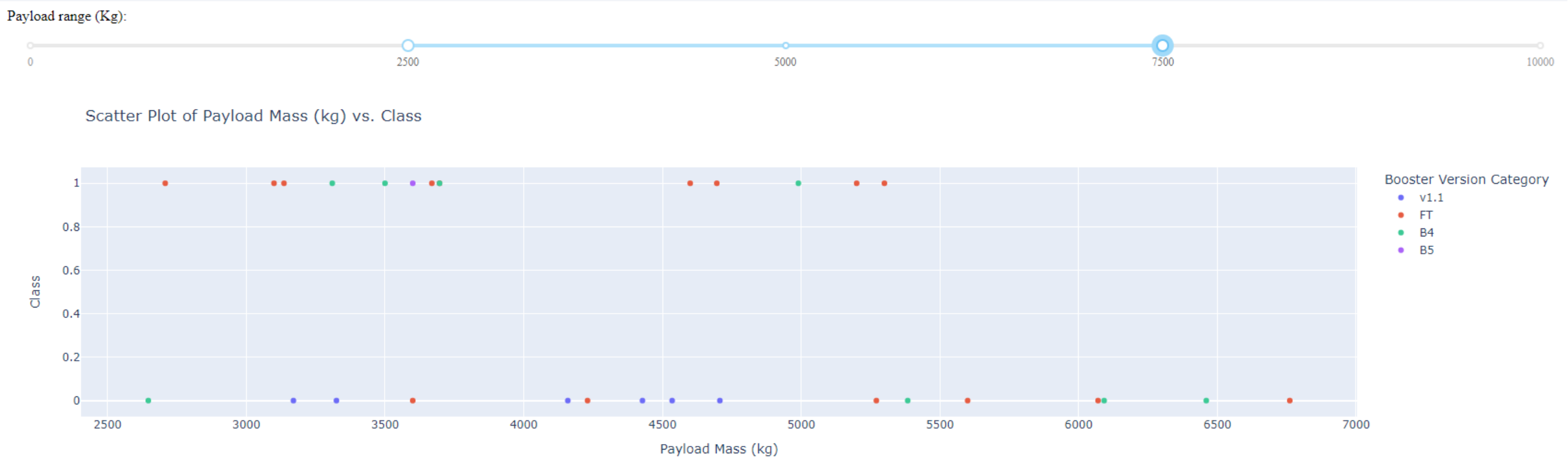
Highest Success Rate

- Launch Site CCAFS SLC-40 had the highest success rate at 42.9% of launches



Launch Outcomes by Payload Mass

- There were no successful Flights in the mid-range for booster v1.1 and no successful flights for any booster in the ~5500-7500 Payload range



Section 5

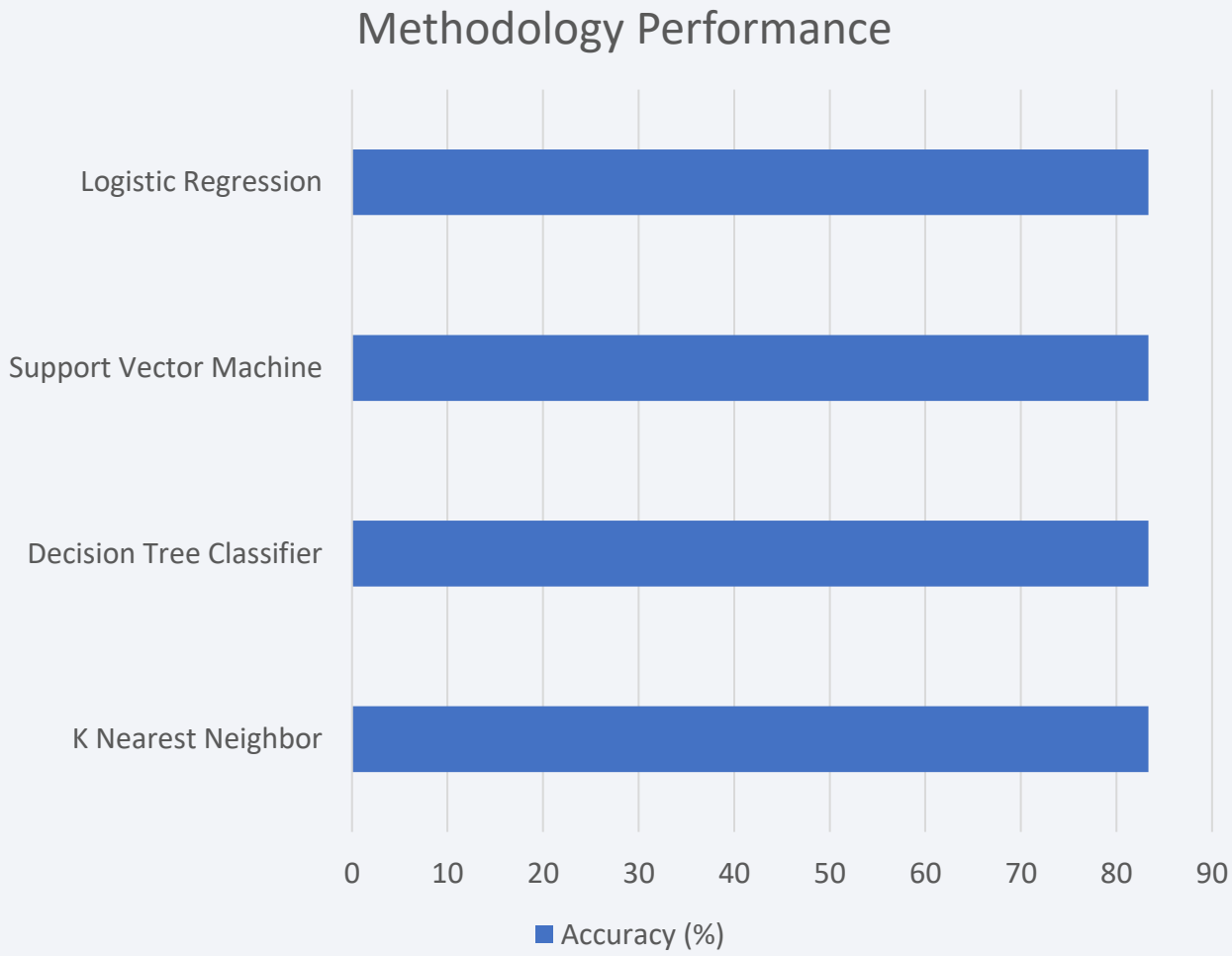
Predictive Analysis (Classification)

Classification Accuracy

- Comparison of 4 different methodologies
- All methodologies showed equal performance when predicting test data
  -Classification Reports and R2_score showed equal competency amongst all options

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

LR_R2: 0.25
SVM_R2: 0.25
Tree_R2: 0.25
KNN_R2: 0.25



Confusion Matrix

- All methodologies yielded accuracy results of 83.3% with identical Confusion Matrices

K Nearest Neighbor

Params: (algo: auto, n_neighbors: 5, p:1)

Tree Classification

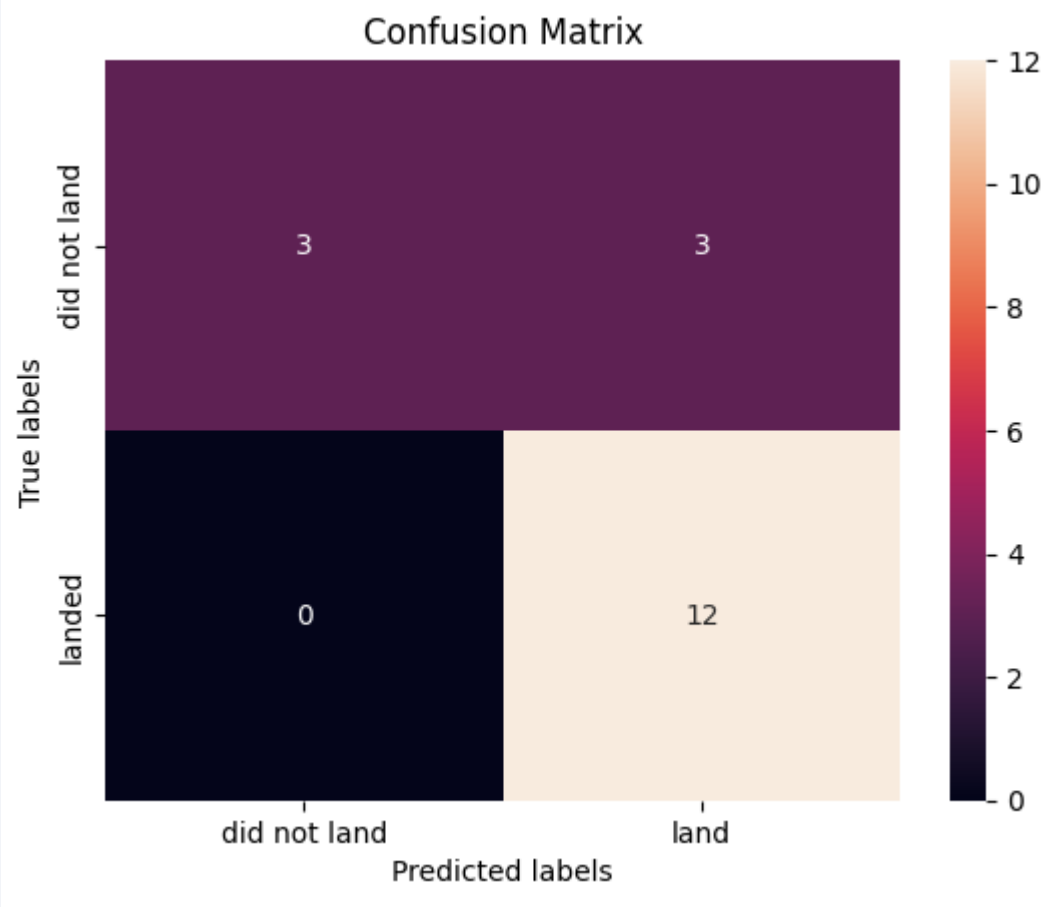
Params: (criterion: entropy, max_depth: 4)

Support Vector Machine

Params: (C: 1.0, gamma: .03162, kernel: sigmoid)

Logistic Regression

Params: (C: .01, penalty: l2, solver: lbfgs)



Conclusions

- Data must be reliably gathered, cleaned, and otherwise prepared in order for it to be accurately compared and analysed
- The data was explored and analyzed in multiple ways to try and identify where comparisons might be made to answer our problem
- Ultimately, the prediction methodologies accurately predicted successful landings but failed to predict failed landings 50% of the time
- The data *may* be able to be improved by increasing the size of the training set or adjusting parameters, but as always, more data would be preferred, especially around failed landings.

Appendix

- The entire repository for this analysis project can be found here: *https://github.com/Tocopherol/Data-Science-Capstone/tree/main*

Thank you!