IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alex T
January 5, 2026

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection using SpaceX API and web scrapping (Wikipedia page with launch data)
  - Data cleaning and wrangling
  - Exploratory data analysis using data visualization and SQL
  - Interactive visual analytics using Folium and Plotly Dash dashboard
  - Predictive analysis using Scikit-learn (classification)
- Summary of all results
  - Heavier payload result in a higher success rate
  - KSC LC-39A has the largest number of successful launches and highest success rate
  - Most predictive models explain 83% of variance in outcomes

Introduction

Background and context

SpaceX is an aerospace and space transportation company. The key part of their approach is to reuse the first stage of a space rocket so that the part can be used for multiple launches thus saving a lot of money and making launch cost much lower. For example, Faclon 9 launch is priced at 62 million dollars which is much lower than other providers' offering of about 165 million dollars.

If an alternate company wants to bid against SpaceX, they might want to predict launch outcomes and costs based on available data to build their business.

- Problems solved in this study
  - Data collection, data analysis and predictive modeling to estimate launch outcomes based on available public data

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using REST API
  - Using Web scraping
- Perform data wrangling
  - Fixed missing data and converted all outcomes to single training label 0 or 1 (Class)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Selected features and used multiple models
  - Cross-validation, hyperparameter grid search and score comparison

Data Collection

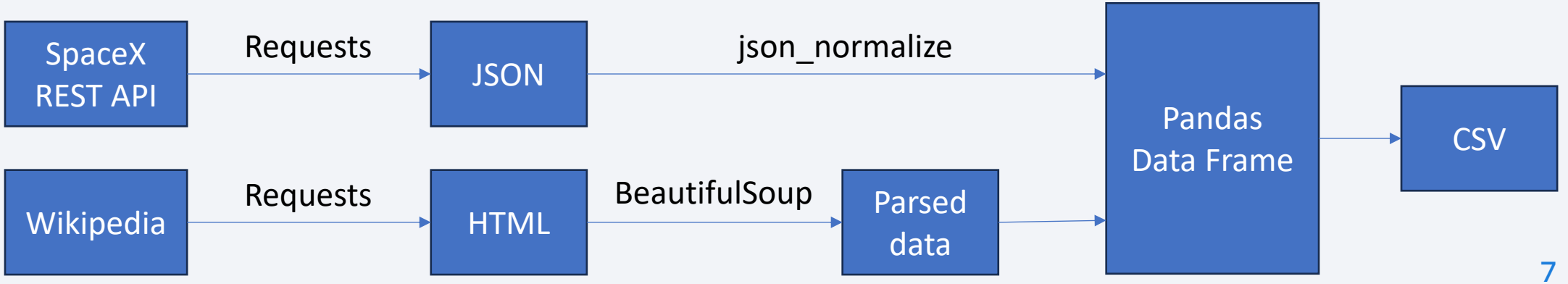
Two data sources:

1. SpaceX REST API using Requests library in JSON format

   https://api.spacexdata.com

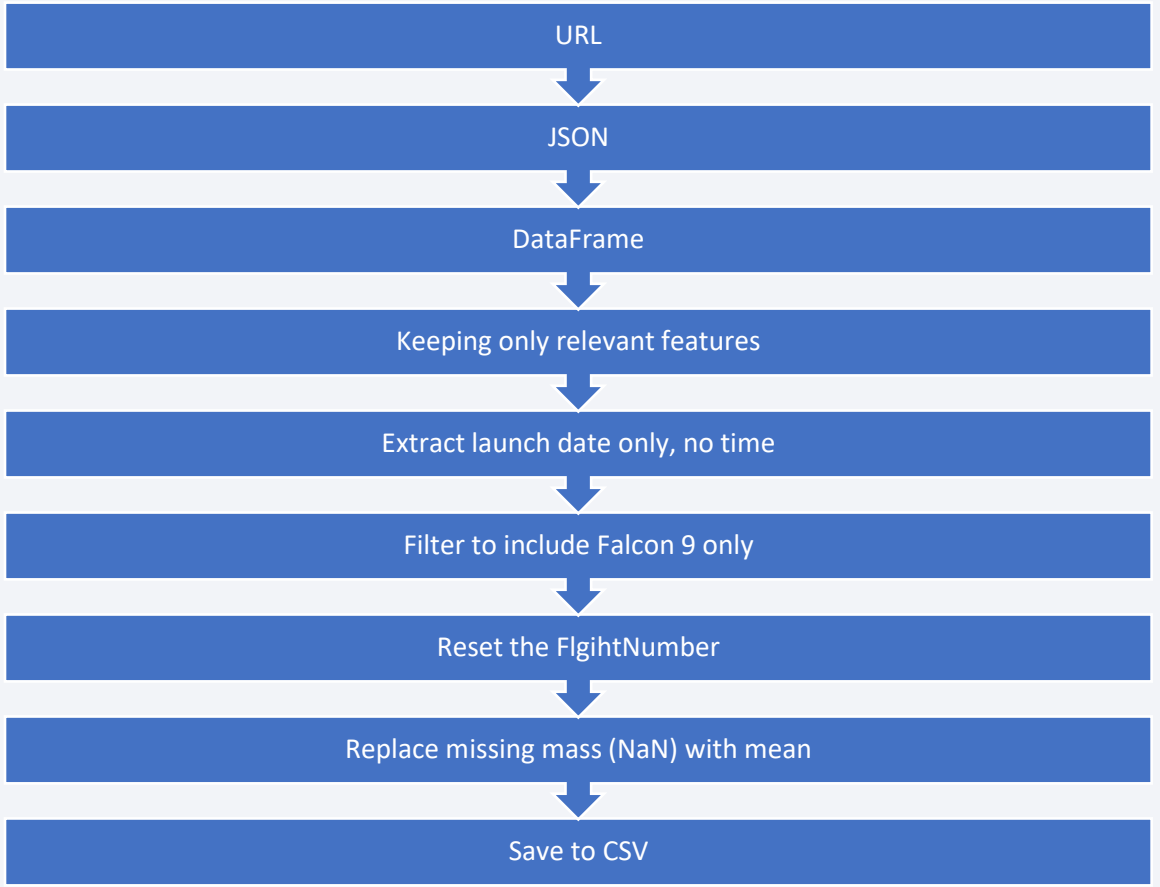
2. Web scraping using Requests and parsing using BeautifulSoup

   https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



Data Collection – SpaceX API

- spacex_url=https://api.spacexdata.com/v4/launches/past
- response = requests.get(spacex_url)
- data = pd.json_normalize(response.json())
- data_falcon9 = launch_data[launch_data['BoosterVersion']!='Falcon 1']
- …loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
- https://github.com/tepl/coursera-applied-data-science-capstone/blob/main/module-1-introduction/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- static_url = https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- response = requests.get(static_url, headers=headers)
- soup = BeautifulSoup(response.content, 'html.parser')
- html_tables = soup.findAll('table')
- https://github.com/tepl/coursera-applied-data-science-capstone/blob/main/module-1-introduction/jupyter-labs-webscraping.ipynb

Data Wrangling

- There are multiple different case where the booster did not land successfully. For example, ocean, ground pad, drone ship etc. True or False for success.
- Convert all those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- All steps:
  - Calculated the number of launches on each site
  - Calculated the number and occurrence of each orbit
  - Calculated the number and occurence of mission outcome of the orbits
  - Created a landing outcome label **Class** from Outcome column
- https://github.com/tepl/coursera-applied-data-science-capstone/blob/main/module-1-introduction/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts plotted:

  0. Scatter plot of FlightNumber vs. PayloadMass with color-coded Class (hue,1 success, 0 failure). Helps to see how flight number and payload would affect the launch outcome.

  1. Categorical plot FlightNumber vs LaunchSite with color-coded Class. Helps to see how launch site affect the outcome.

  2. Categorical plot Payload Mass Vs. Launch Site with color-coded Class. Helps to see which sites used which payloads and correlation with outcomes as well.

  3. Bar plot Class vs Orbit type. Helps to check if there are any relationship between success rate and orbit type.

  4. Categorical plot FlightNumber and Orbit type. Helps to see if there is a correlation between flight number and success rate for different orbit types.

  5. Categorical plot Payload Mass and Orbit type. Helps to see if there is a correlation between payload and success rate for different orbit types.

  6. Line plot Date (Year) vs Class. Help to see yearly trend in the success rate.

- Variables identified: 'FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial'

- Dummy variables: 'Orbit', 'LaunchSite', 'LandingPad', 'Serial' using one-hot encoding

- https://github.com/tepl/coursera-applied-data-science-capstone/blob/main/module-2-eda/edadataviz.ipynb

EDA with SQL

1. Names of the unique launch sites in the space mission
2. 5 records where launch sites begin with the string ‘CCA’
3. Total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. Date when the first successful landing outcome in ground pad was achieved
6. Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. Total number of successful and failure mission outcomes
8. Booster versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function
9. Records with the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
10. Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

https://github.com/tepl/coursera-applied-data-science-capstone/blob/main/module-2-eda/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Summary of used map objects added to the map using Folium

1. Marked all launch sites on a map using Circles and Markers to see locations
2. Marked the success/failed launches for each site using MarkerCluster to see how successful each site is
3. Marked proximities using PolyLine and calculated the distances to those proximities

Used MousePosition to find coordinates

- https://github.com/tepl/coursera-applied-data-science-capstone/blob/main/module-3-visualization-dashboard/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Summary of plots/graphs in the dashboard

1. Pie chart of launch success count for all sites to find the most successful site
2. Pie chart of launch success ratio of the best site to see what the actual ratio is
3. Scatter plot of payload vs. launch outcome to identify the most successful payload range and booster version.

- https://github.com/tepl/coursera-applied-data-science-capstone/blob/main/module-3-visualization-dashboard/spacex-dash-app.py

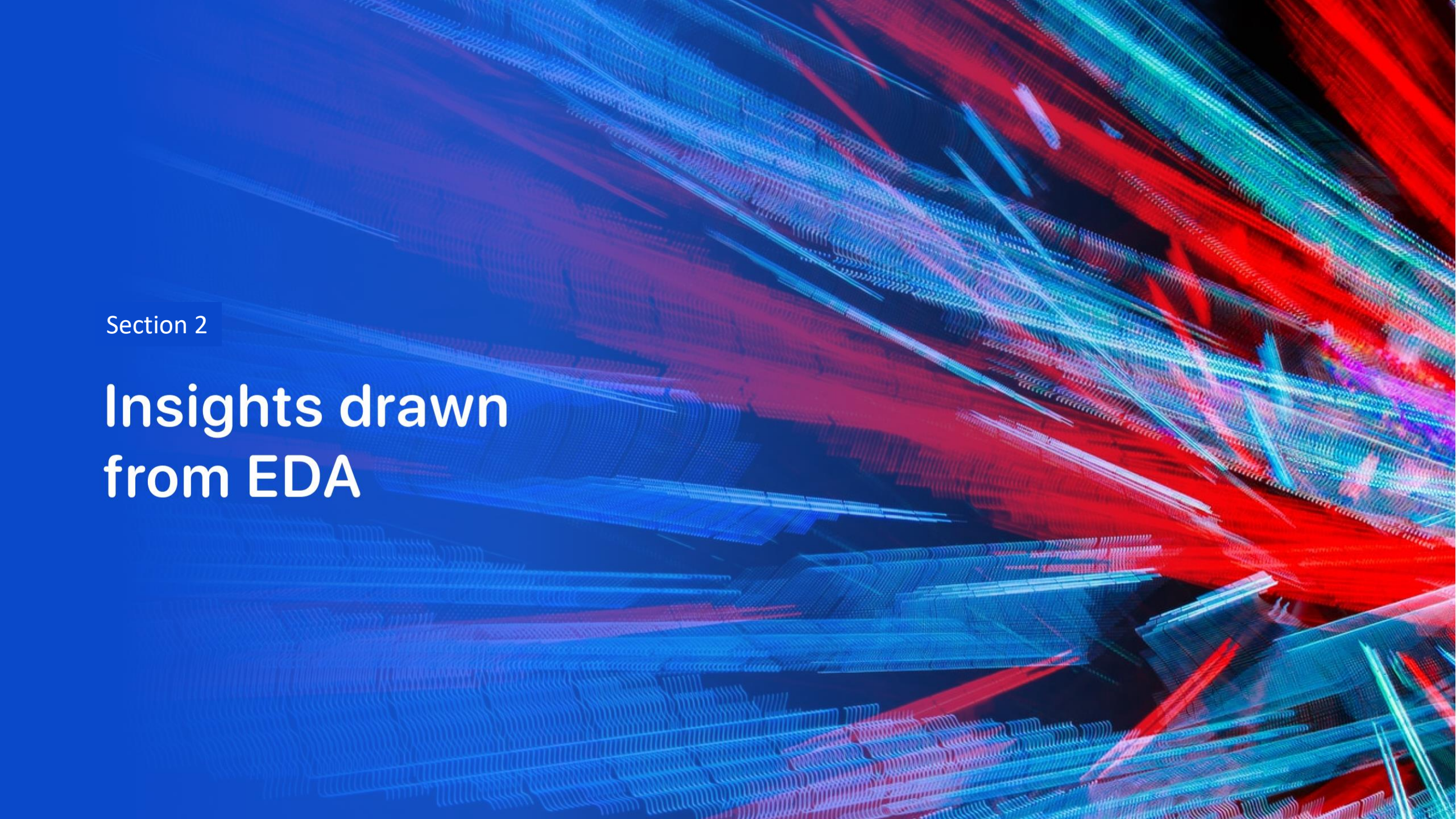
Predictive Analysis (Classification)

Summary of analysis:

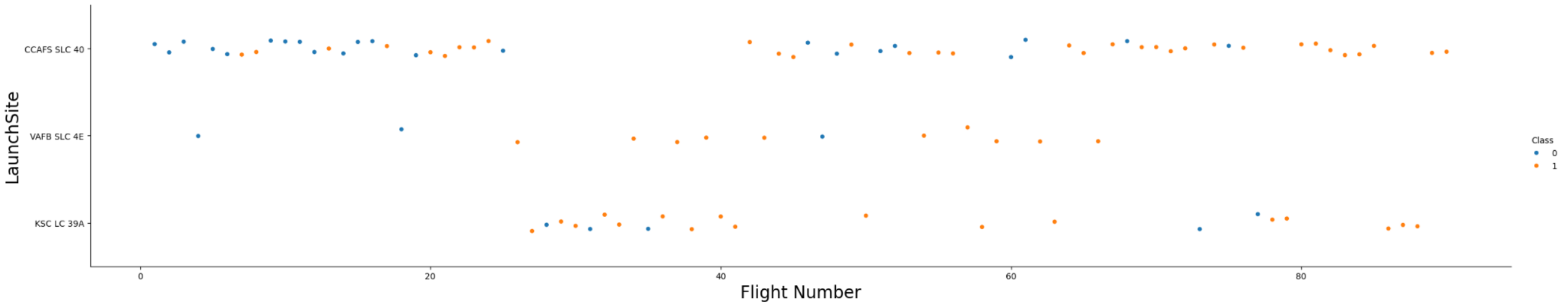
- Preprocessing by standardizing the data
- Train-set split by 20/80, 20% for test data
- Fit multiple models: logistic regression, support vector machine, decision tree classifier and k-nearest neighbors
- Using GridSearchCV to find best hyperparameters with cross-validation cv = 10
- Prediction and scoring on test data and making a bar chart
- Plotting confusion matrix to see how well classification works
- https://github.com/tepl/coursera-applied-data-science-capstone/blob/main/module-4-predictive-analysis/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

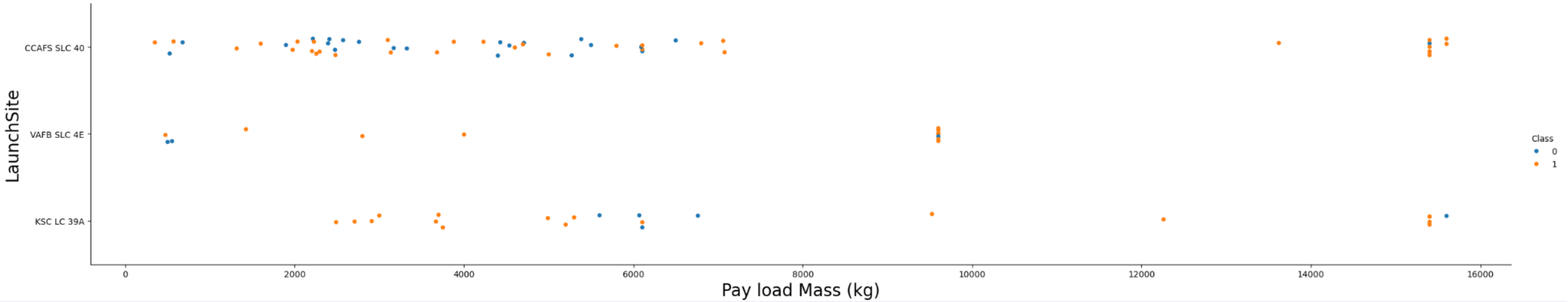

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



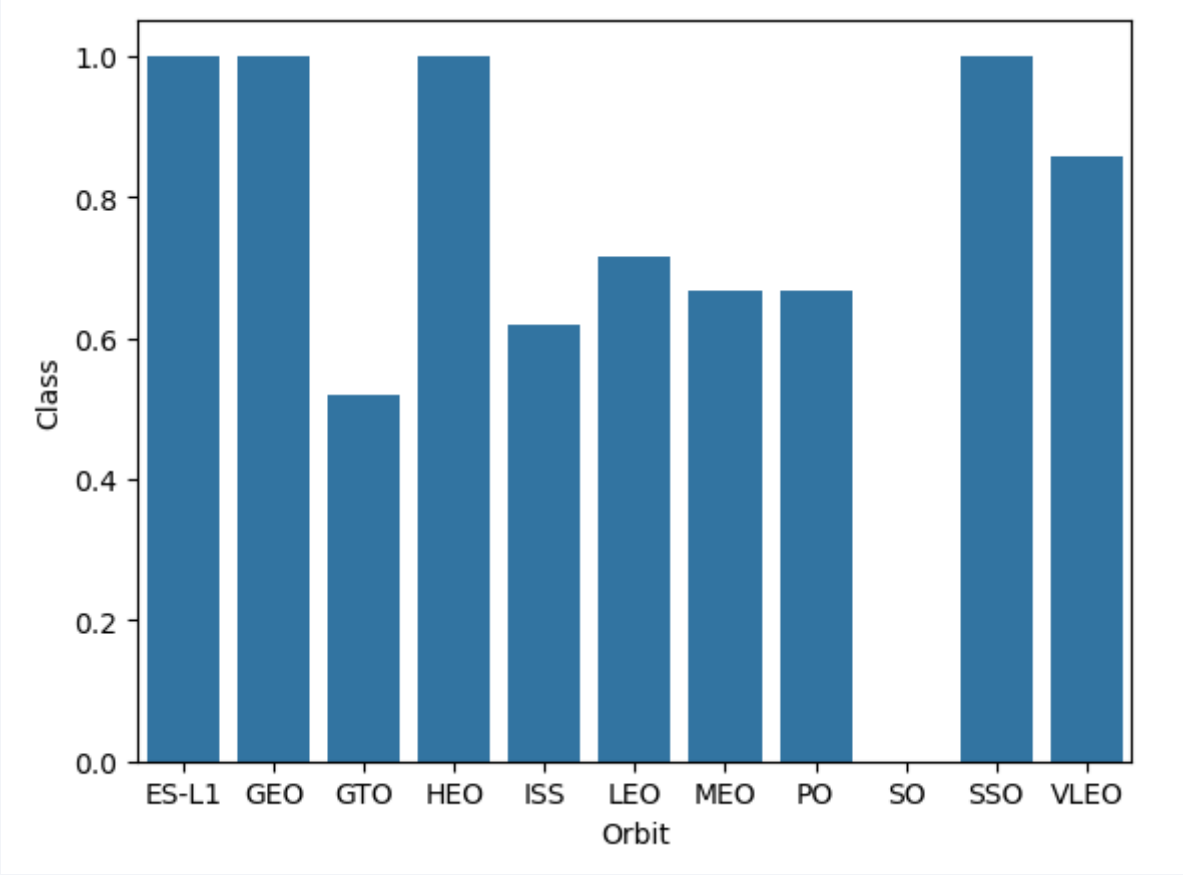
- CCAFS SLC 40 has many failures especially for low flight numbers
- KSC LC 39A does not have low flight numbers.

Payload vs. Launch Site



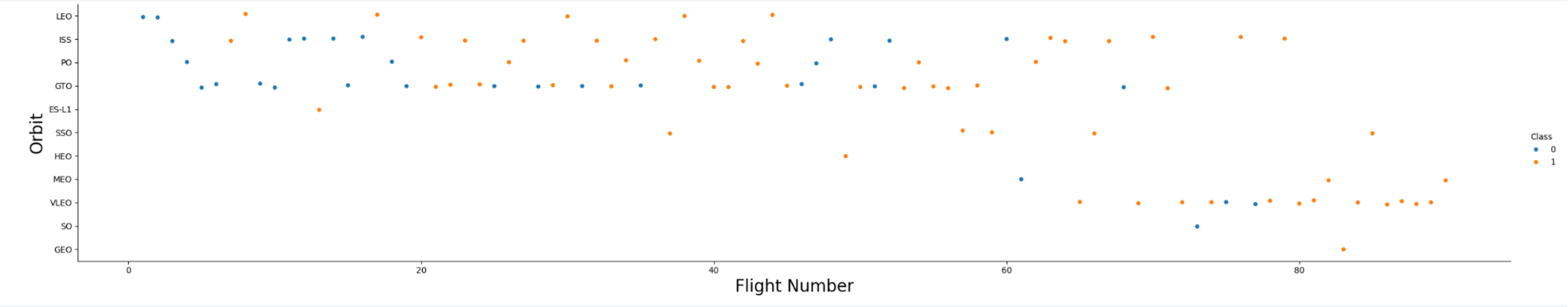
- For VAFB-SLC there are no rockets launched for heavy payload mass.
- Heavy payloads are more successful.

Success Rate vs. Orbit Type



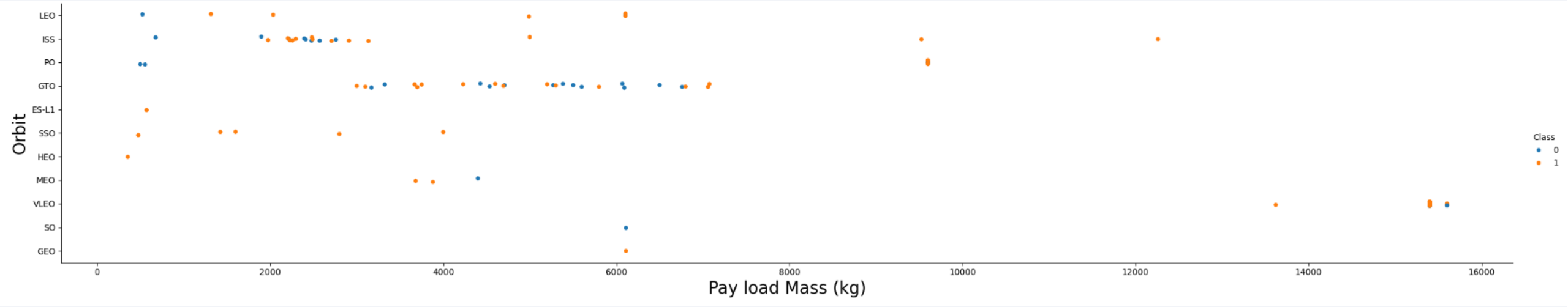
- Orbits with highest success rate are ES-L1, GEO , HEO and SSO.

Flight Number vs. Orbit Type



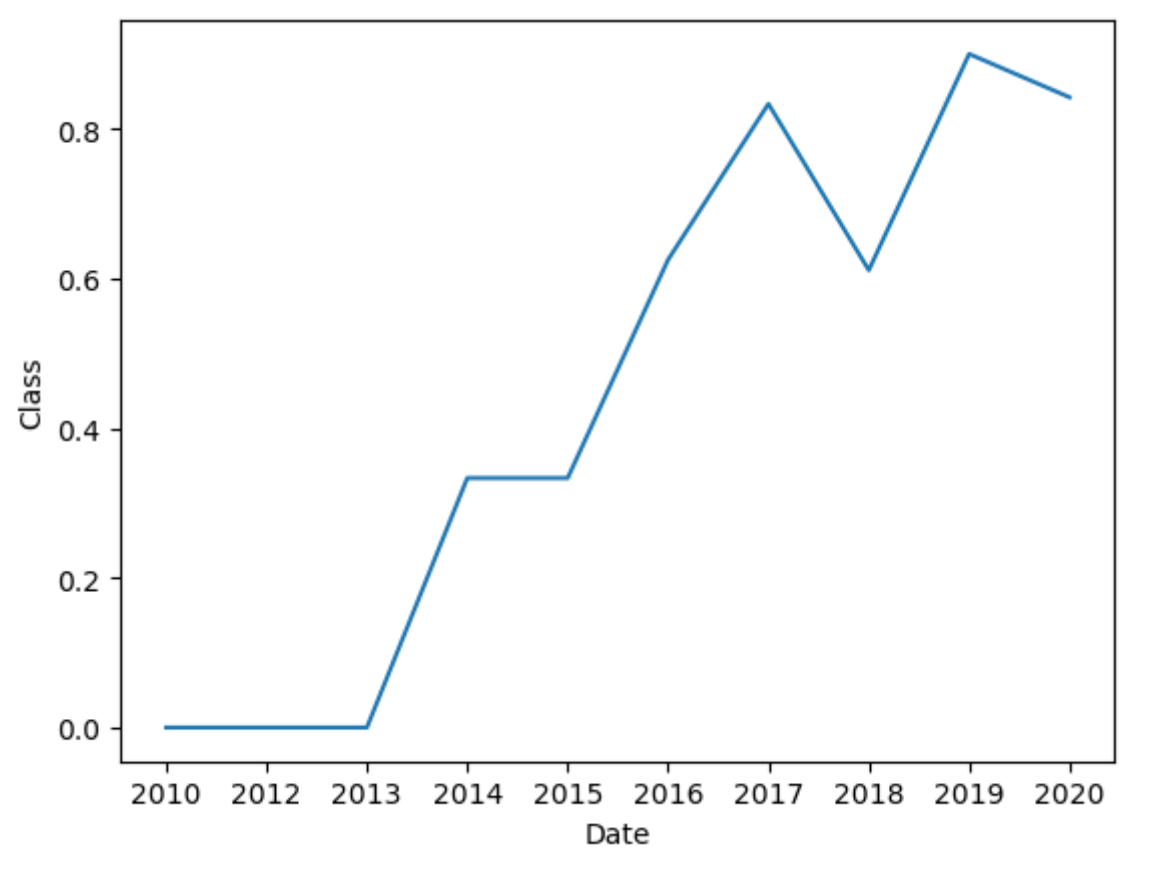
- For LEO orbit the success rate correlates with the flight number.
- For GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



- With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
- For GTO, both outcomes are present.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

- SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE;
- There are 4 unique names

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

Launch Site Names Begin with 'CCA'

- SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%';
- Top 5 CCA launches

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

Total Payload Mass

- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Customer" = "NASA (CRS)";
- Total payload mass by NASA

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1";
- Average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
- Dates of the first successful landing outcome on ground pad

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

- SELECT COUNT(*) FROM SPACEXTABLE WHERE "Landing_Outcome" Like 'Success%' OR "Landing_Outcome" Like 'Failure%';
- Total number of successful and failure mission outcomes

COUNT(*)
71

Boosters Carried Maximum Payload

- SELECT "Booster_Version" FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
- Names of the booster which have carried the maximum payload mass

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

2015 Launch Records

- SELECT SUBSTR(Date, 6, 2), "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND SUBSTR(Date,0,5) = '2015';
- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

SUBSTR(Date, 6, 2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

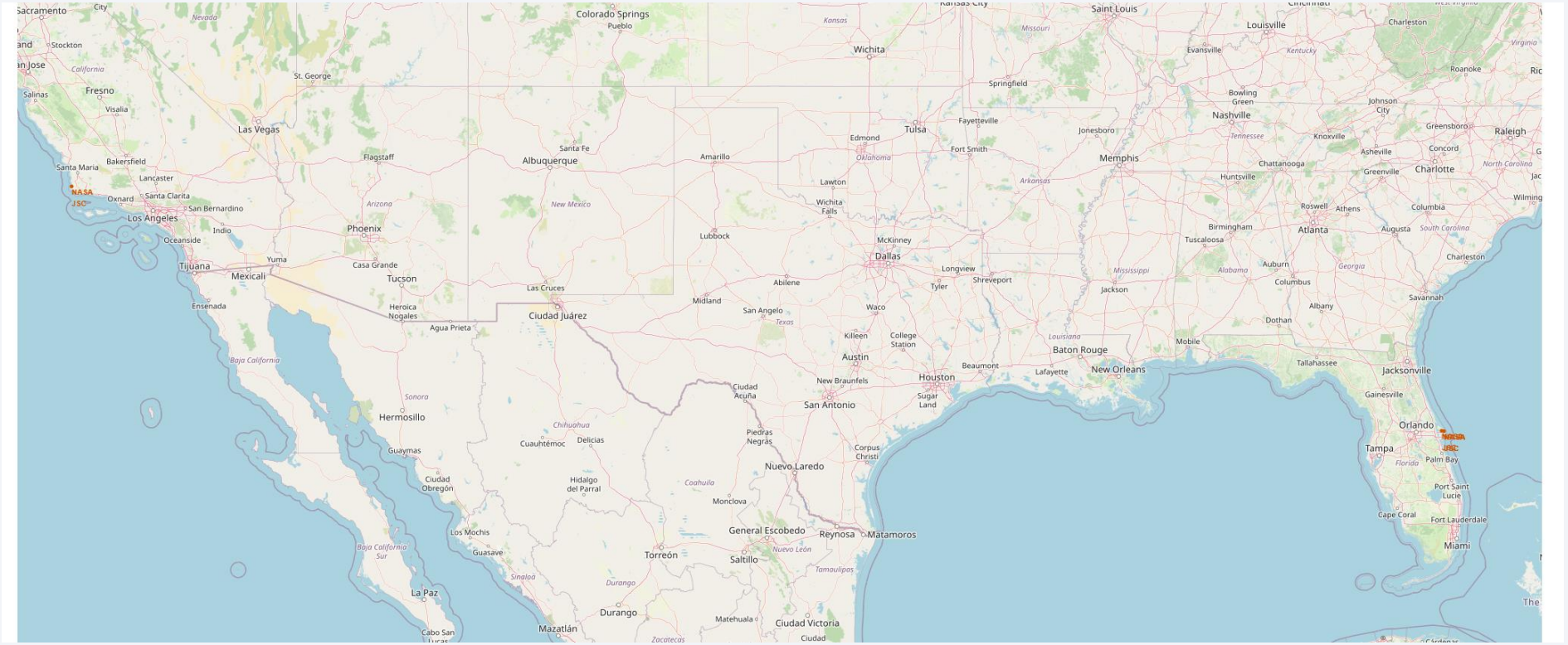
- SELECT "Landing_Outcome", COUNT(*) AS OUTCOMES_COUNT FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY OUTCOMES_COUNT DESC;
- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing_Outcome	OUTCOMES_COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All launch location



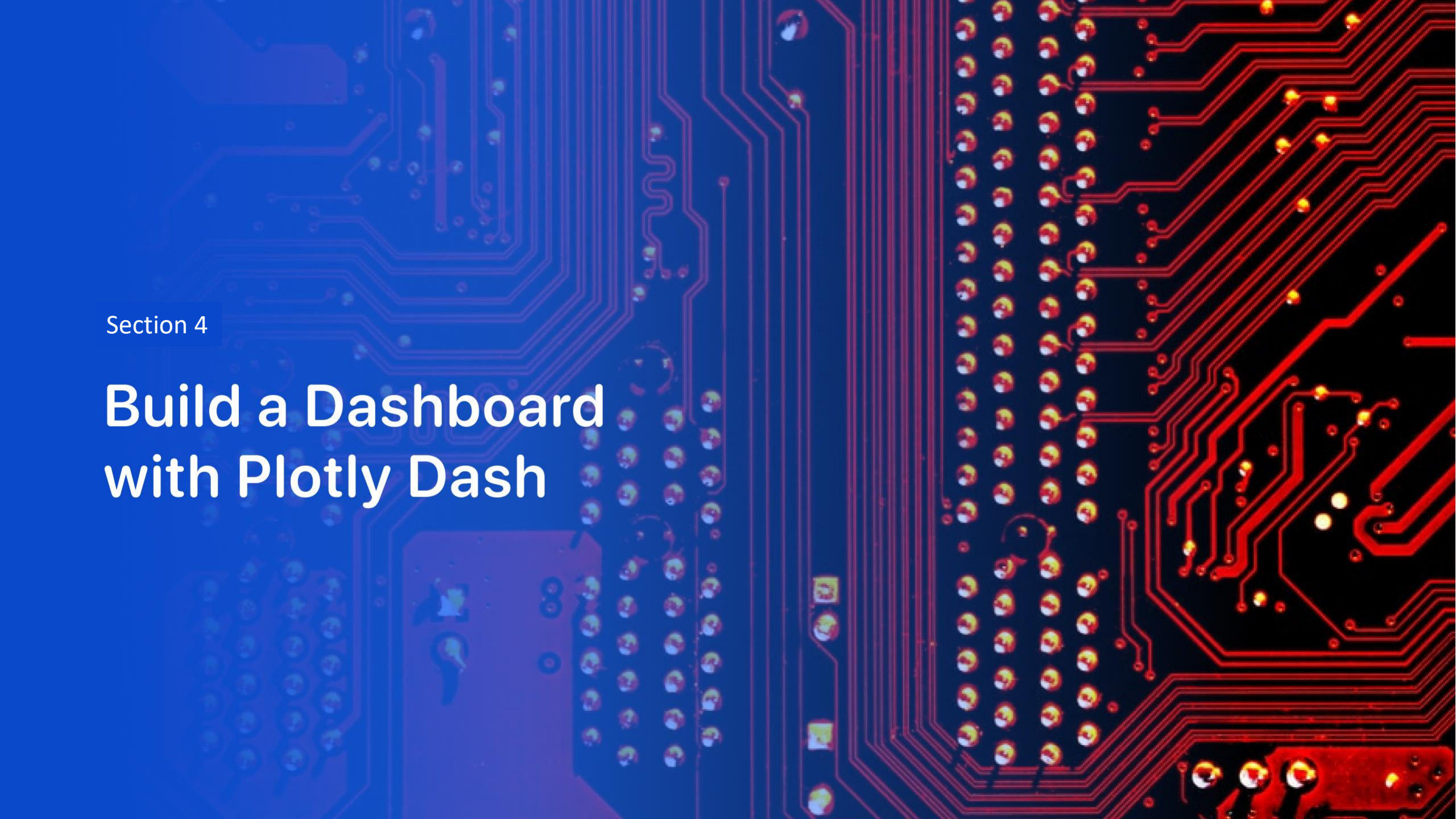
- There are 4 launch locations
- California: VAFB SLC-4E
- Florida: CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A

Color-labeled launch outcomes on the map

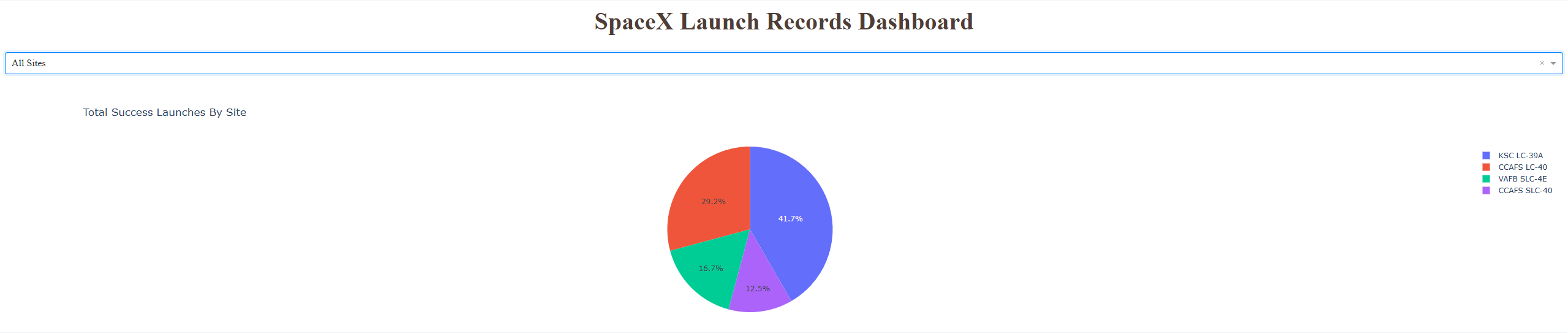
- KSC LC-39A has the highest success ratio (a lot of green markers, only a few are red)

Launch site to its proximities

- CCAFS LC-40 proximity examples
- To coast line 0.92 KM
- To road 0.67 KM

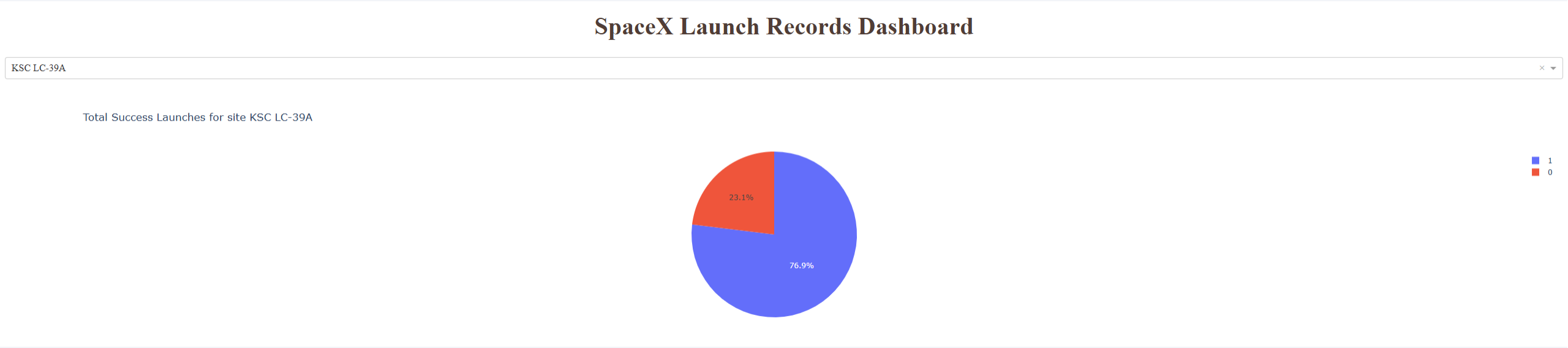

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



- KSC LC-39A has the highest number of successful launches

Launch site with highest launch success ratio



- KSC LC-39A launch ratio is very high 3:1
- A lot of flights were successful, only 23% have failed

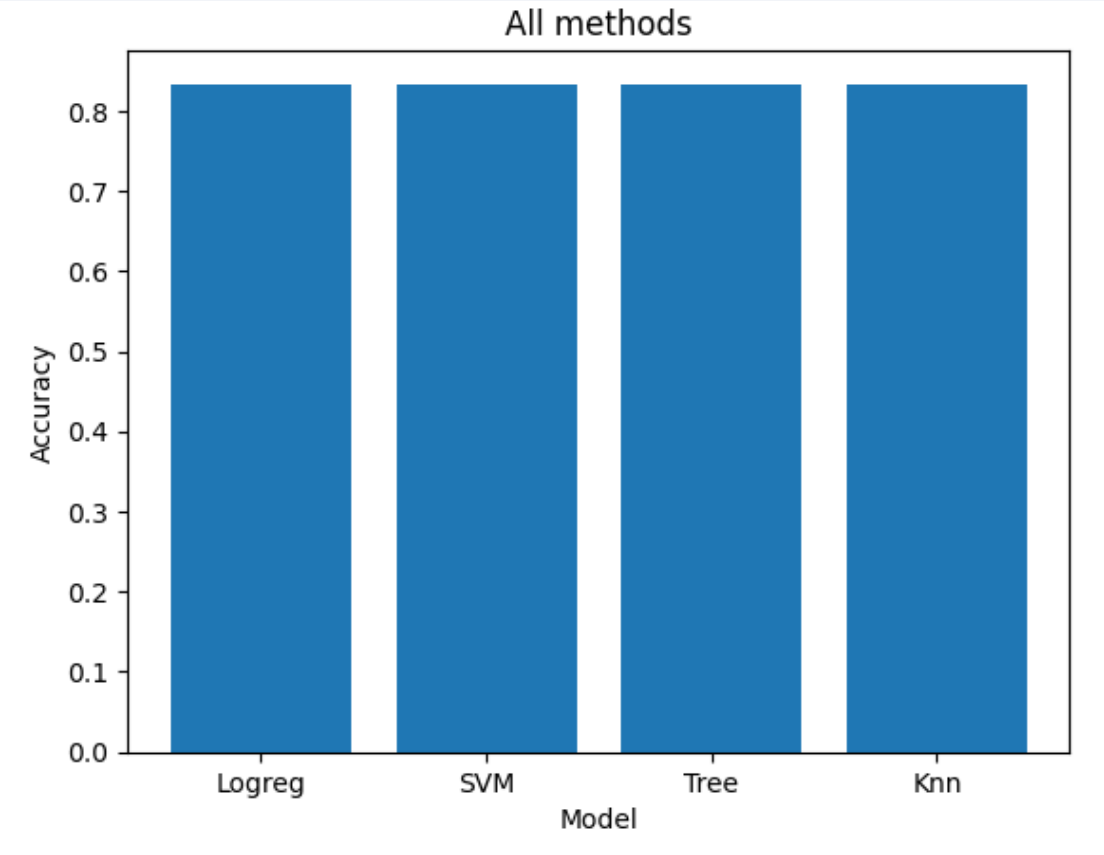
Payload vs. Launch Outcome

- In the payload range 2000 – 6000 kg booster version FT has the highest success rate

Section 5

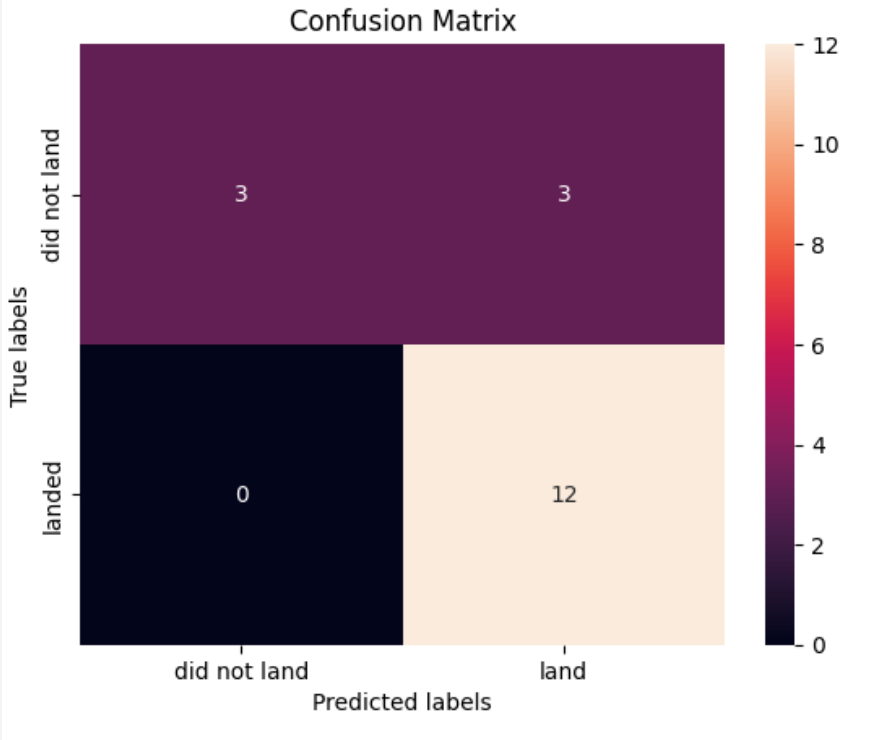
Predictive Analysis (Classification)

Classification Accuracy



- All perform the same on the test data with accuracy of 0.833.
- Decision tree classifier seems to do better on the training data then other methods.

Confusion Matrix



- Diagonal elements show how many predicted labels match true labels
  - [0, 0] True Positives – count 3 (True label is not landed, Predicted label is also not landed)
  - [1, 1] True Negatives – count 12 (True label is landed, Predicted label is also landed)
- Off-diagonal elements how many were predicted incorrectly
  - [0, 1] False Negatives – count 3 (True label is not landed, Predicted label is landed)
  - [1, 0] False Positives – count 0 (True label is landed, Predicted label is not landed)

Conclusions

- Heavier payload result in a higher success rate
- KSC LC-39A has the largest number of successful launches and highest success rate
- Most predictive models explain 83% of variance in outcomes

Appendix

- All relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, and data sets used in this project can be found on the project Github.

Additional insight

Task8 in Prediction lab has a typo. It says to use ‘auto’ for max_features however DecisionTreeClassifier() does not have auto option and the calculation shows error.
To fix: 'max_features': ['log2', 'sqrt']

Thank you!