IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Scott Sodoma
July 12, 2024

Outline

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

Executive Summary

- Summary of Methodologies
  - Data Collection through APIs and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Data Visualization
  - Interactive Visual Analysis with Folium and Plotly Dash
  - Machine Learning and Prediction
- Summary of all Results
  - Exploratory Data Analysis results
  - Interactive Analytics screenshots
  - Predictive Analytics results

Introduction

- Project Background and Context
  - SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.
- Questions to be Answered
  - How do variables such as launch site, payload mass, number of flights and orbits affect the success of first stage landing success.
  - Success rate of landings over time
  - Best predictive model for determining successful landings

Section 1

Methodology

Methodology

Executive Summary

- Data collection using SpaceX REST API and Web Scraping from Wikipedia
- Perform data wrangling by using on-hot encoding applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using four different classification models

Data Collection

- Data was collected using two different methods and sources:
  - SpaceX API
    - https://api.spacexdata.com/v4/
  - Wikipedia Page “List of Falcon 9 and Falcon Heavy Launches”
    - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

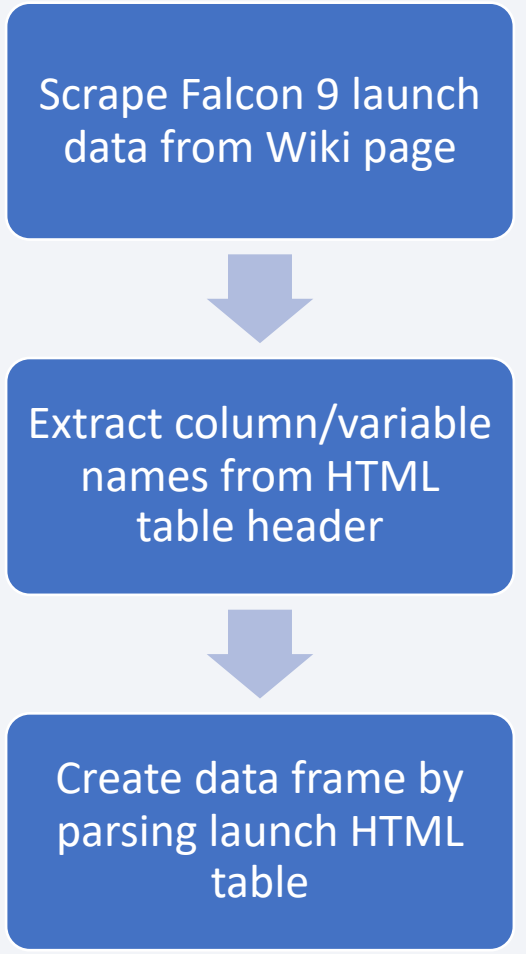
- SpaceX API was used to collect the following data: rocket booster name, launch pad site, payload weight, the landing, the type of the landing, number of flights with that core, whether gridfins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to separate version of cores, the number of times this specific core has been reused, and the serial of the core.

- SpaceX API Location: https://api.spacexdata.com/v4/

- GitHub URL of the SpaceX API notebook: https://github.com/srsodoma/DataScienceClass/blob/main/Mod%201%20Data-Collection-API.ipynb

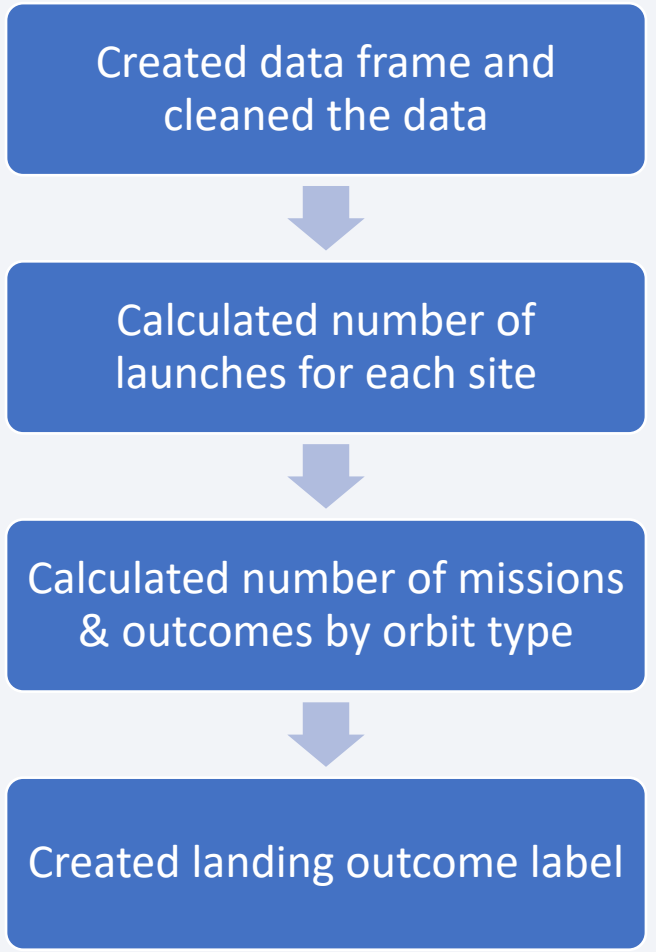
Data Collection – Web Scraping

- Falcon 9 and Falcon Heavy launch data was collected from a Wikipedia page and parsed using BeautifulSoup library functions
- SpaceX Falcon 9 Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- GitHub URL of the Web Scraping notebook: https://github.com/srsodoma/DataScienceClass/blob/main/Mod%201%20Web%20Scraping.ipynb



Data Wrangling

- Performed preliminary Exploratory Data Analysis (EDA) to find patterns in the data and determine what would be the label for training supervised models
- Converted mission outcomes into Training Labels with 1 (booster successfully landed) and 0 (unsuccessful landing)
- GitHub URL of the Data Wrangling notebook: https://github.com/srsodoma/DataScienceClass/blob/main/Mod%201%20Data%20Wrangling%20Lab.ipynb



Exploratory Data Analysis with Data Visualization

- Performed Exploratory Data Analysis and Feature Engineering by looking at the relationship between flight # and launch site, payload mass and launch site, success rate and orbit type, flight # and orbit type, and payload mass and orbit type
- Predicted SpaceX Falcon 9 First Stage Landing success. We also looked at launch success over time
- GitHub URL of the EDA Data Visualization notebook: https://github.com/srsodoma/DataScienceClass/blob/main/Mod%202%20EDA%20Visualization%20Lab.ipynb

Exploratory Data Analysis with SQL

- Summary of SQL queries used on the data:
  - Display names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List date when the first succesful landing outcome in ground pad was achieved
  - List names of boosters which have success in drone ship & payload mass greater than 4,000 but less than 6,000
  - List total number of successful and failure mission outcomes
  - List names of the booster_versions which have carried the maximum payload mass.
  - List records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank the 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.
- GitHub URL of the EDA SQL notebook: https://github.com/srsodoma/DataScienceClass/blob/main/Mod%202%20EDA%20-%20SQL%20Lab.ipynb

Built an Interactive Map with Folium

- Enhanced map by creating data generated markers, circles and lines:
  - Markers used to indicate launch sites
  - Highlighted circles used to add text label on a specific coordinate or site
  - Marker Cluster used to show groups of events at each site (i.e. launch outcomes)
  - Lines were used to show distance between site and nearest coast, city, railway and highway.

- GitHub URL of the Folium-based Interactive Map notebook: https://github.com/srsodoma/DataScienceClass/blob/main/Mod%203%20Visual%20Analytics%20Folium.ipynb

Built a Dashboard with Plotly Dash

- A dashboard application was created that included the following components:
  - Element #1 - Pie Chart
    - Input: a dropdown list of Launch Sites
    - Output: a pie chart showing total successful launches by site
  - Element #2 - Scatter Chart
    - Input: slider bar for selecting Payload Mass
    - Output: scatter chart showing correlation of Payload size to launch success rate for all sites
- Based on these charts, we can determine:
  - Which site has the largest successful launches
  - Which site has the highest launch success rate
  - Which payload range(s) has the highest launch success rate
  - Which payload range(s) has the lowest launch success rate
  - Which F9 Booster version has the highest launch success rate
- GitHub URL of the Dashboard with Plotly Dash notebook: https://github.com/srsodoma/DataScienceClass/blob/main/Mod%203%20Dashboard%20Plotly%20Dash%20spacex_dash_app.py

Predictive Analysis (Classification)

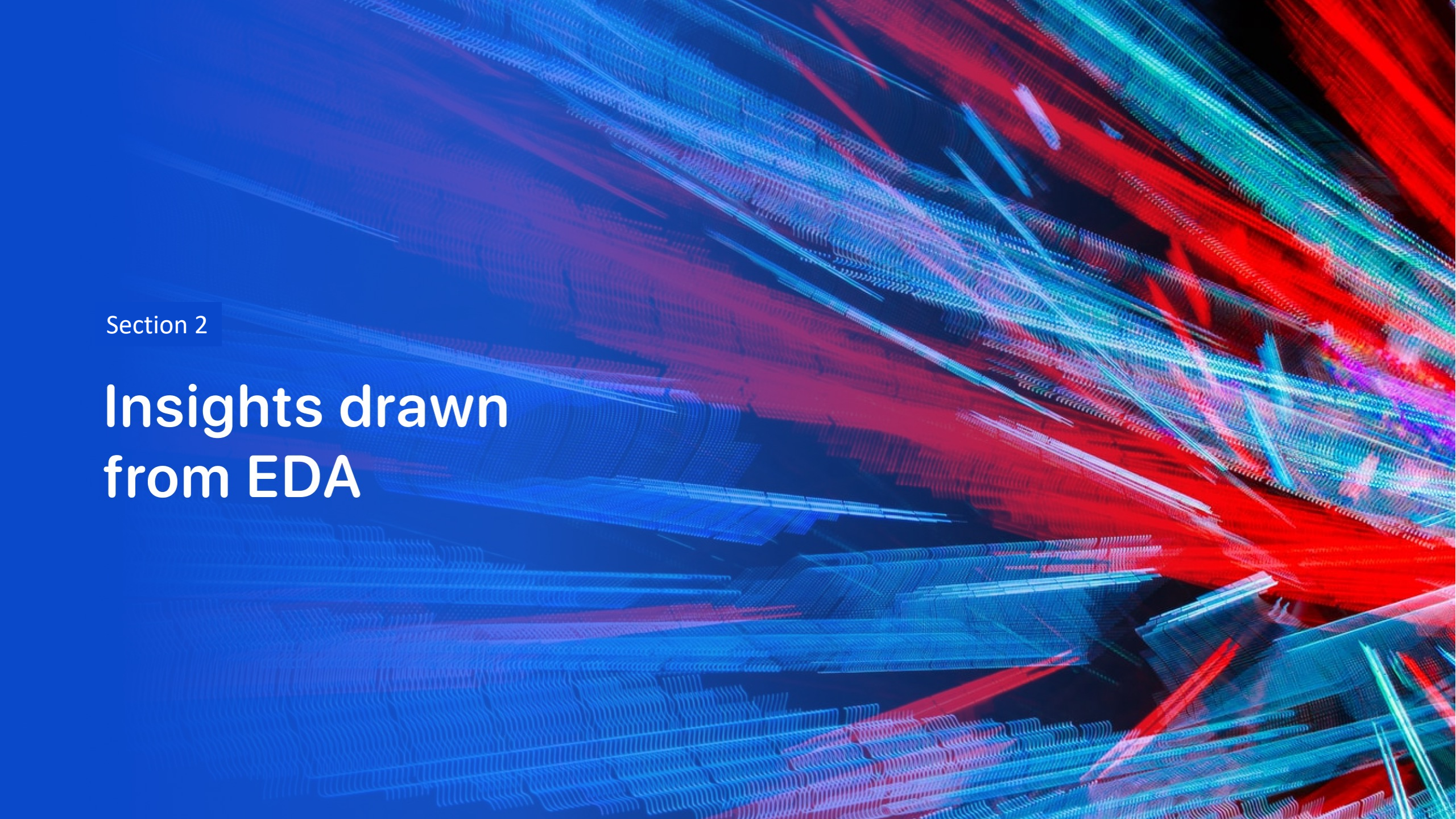
Approach and methodology:

- Performed exploratory Data Analysis and determine Training Labels
- Create a column for the class
- Standardized the data
- Split data into training and test data sets
- Identified best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Identified the best classification method using test data

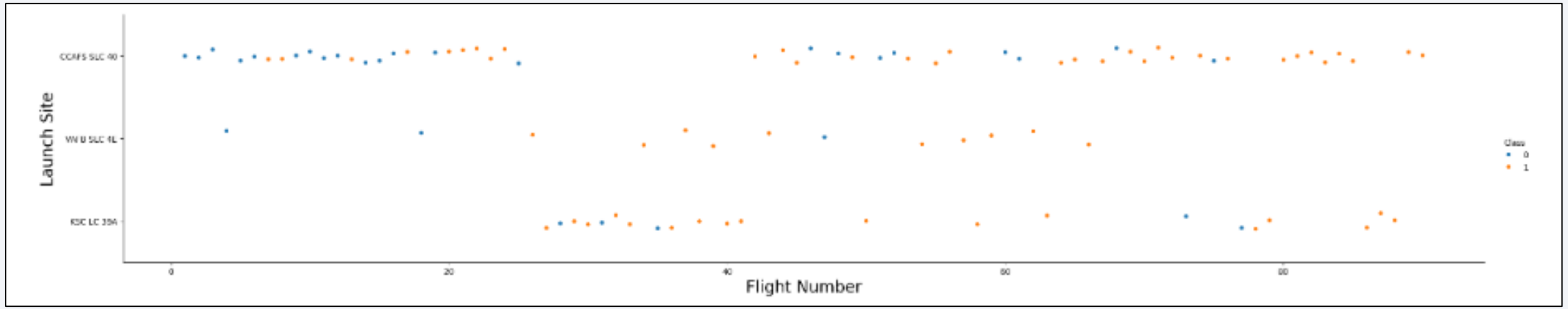
- GitHub URL of the Predictive Analysis notebook: https://github.com/srsodoma/DataScienceClass/blob/main/Mod%204%20SpaceX_Machine%20Learning%20Prediction.ipynb

Results

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


Section 2
Insights drawn from EDA

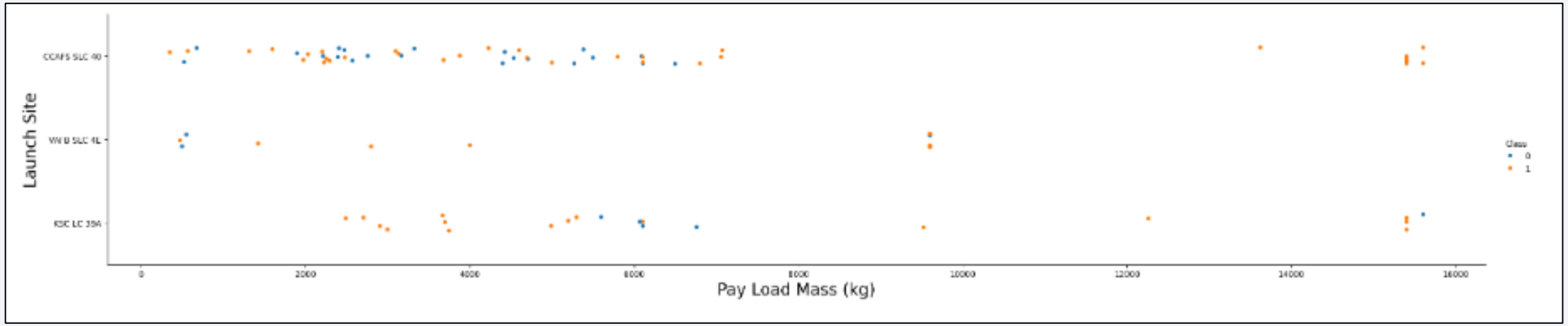
Flight Number vs. Launch Site



Conclusions:

- Rate of successful launches at each site increases as more flights are conducted
- Launch site CCAFS LC-40 has the most launches

Payload Mass (kg) vs. Launch Site



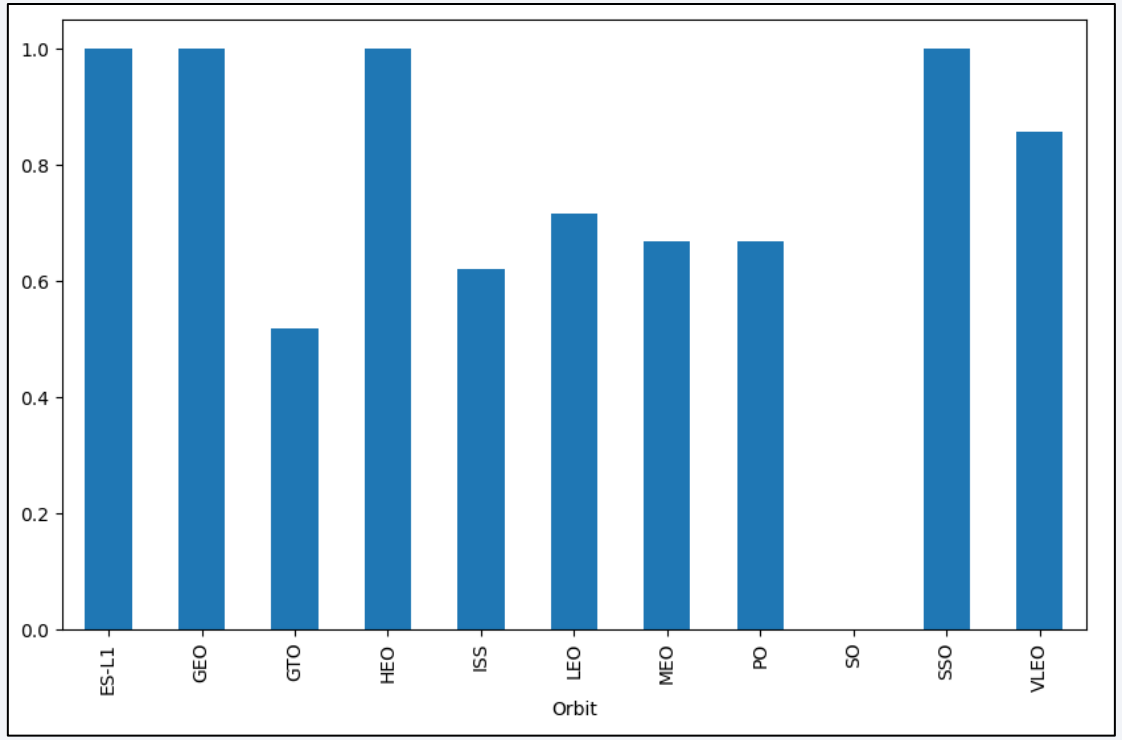
Conclusions:

- The greater the Payload in mass the the higher the success rate
- CCAFS LC-40 is the preferred launch site for small to mid size (0-7,000kg) Payloads
- Launch sites CCAFS LC-40 and KSC LC-39A are preferred for Payloads 10,000kg and greater

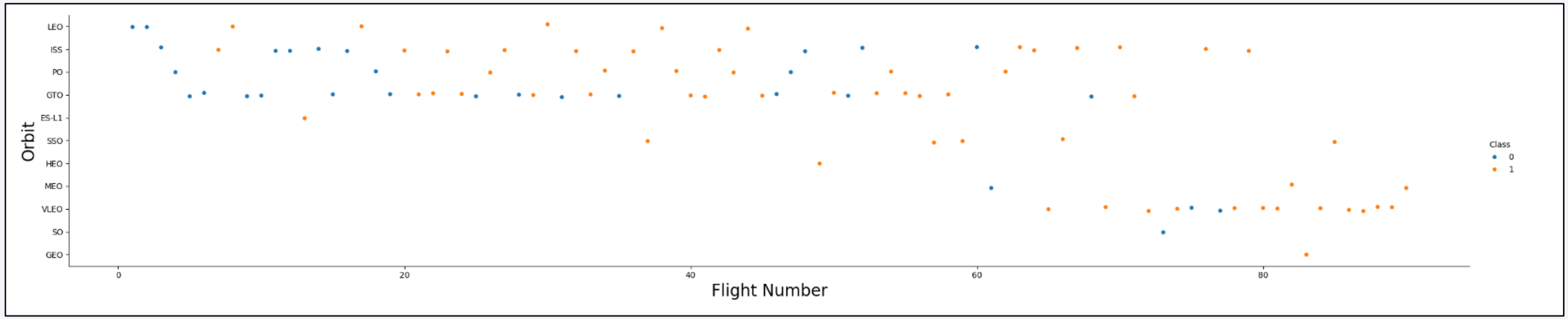
Success Rate vs. Orbit Type

Conclusions:

- Orbits ES-L1, GEO, HEO and SSO have the highest success rate
- Orbit SO had the lowest success rate



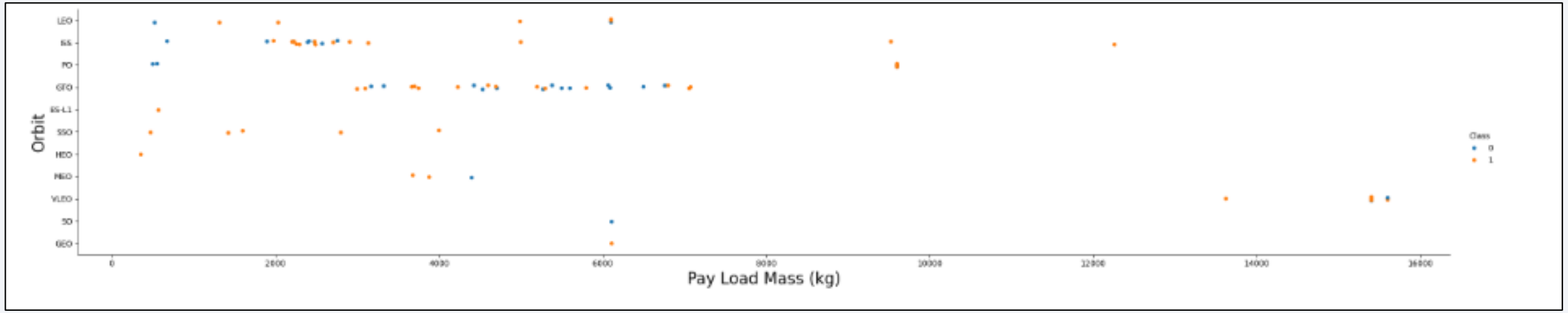
Flight Number vs. Orbit Type



Conclusions:

- Success rate of LEO orbit increases with the number of flights
- There seems to be no relationship between number of flights when in GTO orbit
- VLEO orbit is the preferred or newest orbit of recent launches

Payload vs. Orbit Type



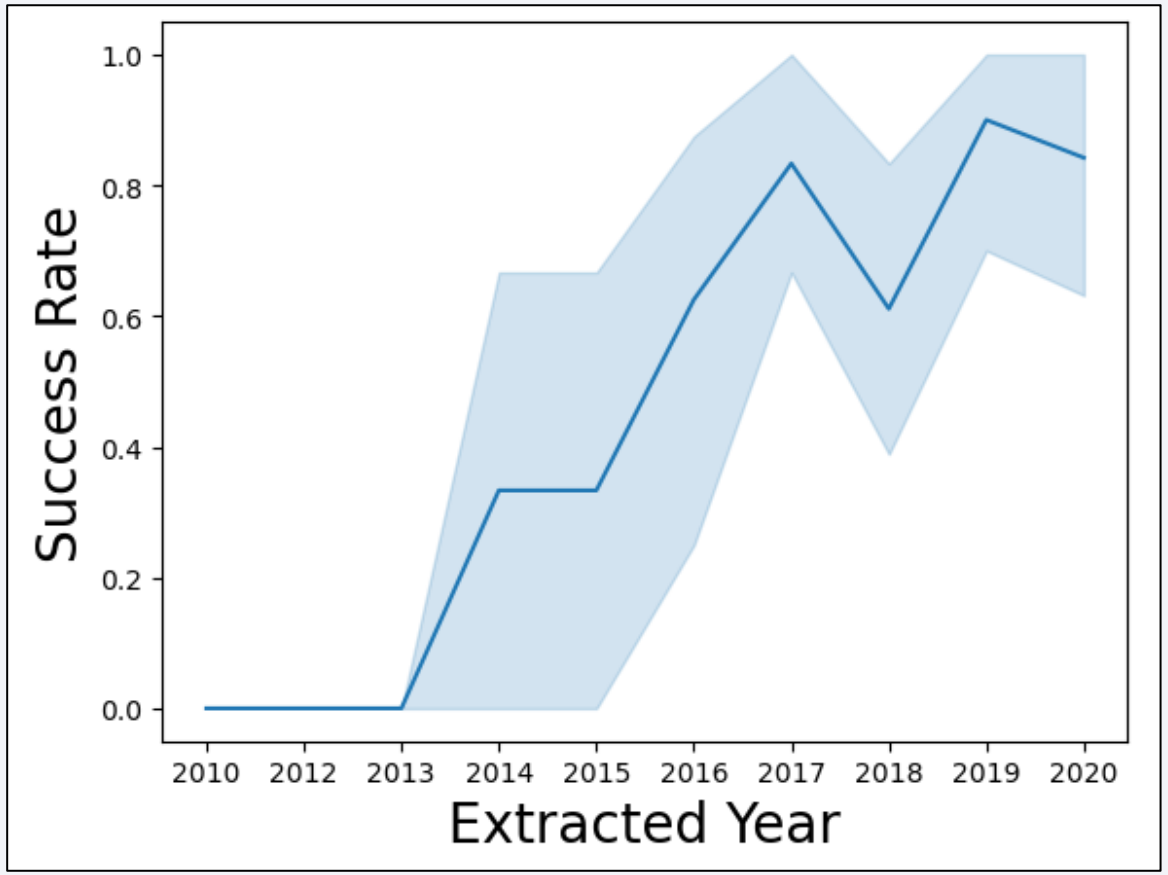
Conclusions:

- With heavy payloads the successful landing or positive landing rate are higher for PO, LEO and ISS orbits.
- For GTO orbit, there doesn't seem to be a relationship between Payload mass and Orbit

Launch Success Yearly Trend

Conclusions:

- Since 2013, the success rate has been increasing until 2020



All Launch Site Names

- Used the DISTINCT keyword to generate a list of unique launch sites from the data
- SQL Query: sql select distinct(launch_site) from SPACEXTABLE

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

5 Records Where Launch Site Names Begin with 'CCA'

- Searched for launch sites that begin with “CCA%” and then used LIMIT keyword to list only 5 records
- SQL Query: sql SELECT * from SPACEXTABLE WHERE launch_site LIKE "CCA%" LIMIT 5

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass By Boosters Launched by NASA (CRS)

- Selected records where NASA (CRS) was the customer and totaled payloads mass for all launches.
- SQL Query: sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE customer = 'NASA (CRS)'

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by Booster F9 v1.1

- Selected records where Booster Version was F9 v1.1 and averaged payloads mass for those launches.
- SQL Query: sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version like "F9 v1.1%"

avg(PAYLOAD_MASS__KG_)
2534.666666666665

First Successful Ground Landing Date

- Selected records where there was a successful landing on a ground pad and then used MIN to select the oldest (or first) successful ground landing
- SQL Query: sql SELECT min(date) FROM SPACEXTABLE WHERE landing_outcome="Success (ground pad)"

min(date)
2015-12-22

Booster Names with Successful Drone Ship Landing and Payloads between 4000 and 6000

- Selected Booster names (versions) where there was a successful drone ship landing and payload was between 4,000kg – 6,000kg
- SQL Query: sql SELECT booster_version FROM SPACEXTABLE WHERE landing_outcome="Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 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

- Counted records my mission outcome
- SQL Query: sql SELECT mission_outcome, COUNT(*) AS TOTAL FROM SPACEXTABLE GROUP BY mission_outcome

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

List of Boosters that have Carried Maximum Payload

- Determine maximum payload size and then select the unique Booster names (versions) that have carried that payload
- SQL Query: sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

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

List of Failed Drone Ship Landings in 2015

- Select records from 2015 that had failed drone ship landing outcomes
- SQL Query: sql select substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE \ where substr(Date,0,5)='2015' and Landing_Outcome is 'Failure (drone ship)'

month	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

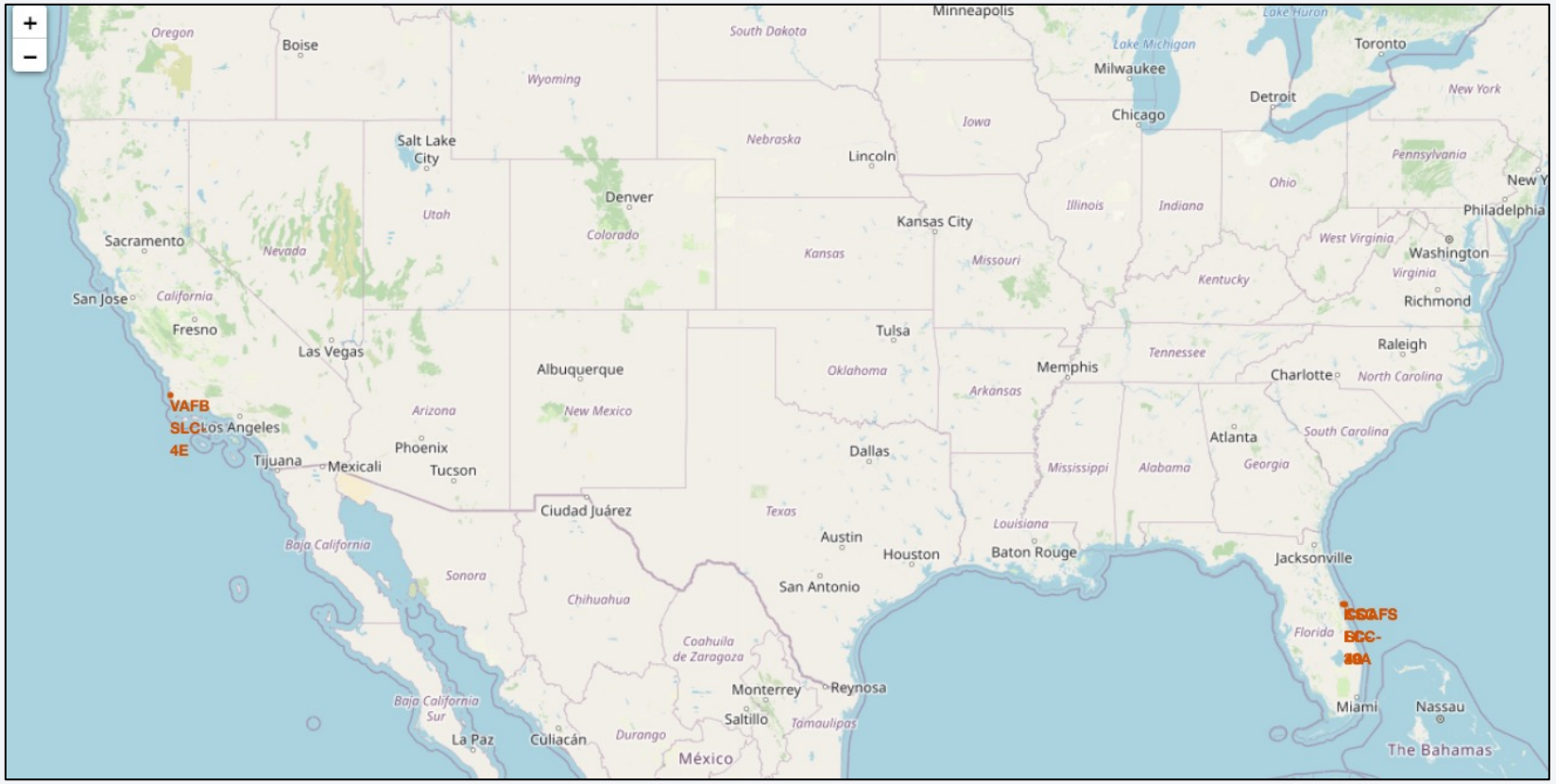
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Selected landings from 2015 that had failed drone ship landing outcomes
- SQL Query: sql SELECT "Landing_Outcome", count(*) as 'Quantity' FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-02' group by "Landing_Outcome" order by "Quantity" descql

Landing_Outcome	Quantity
No attempt	9
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

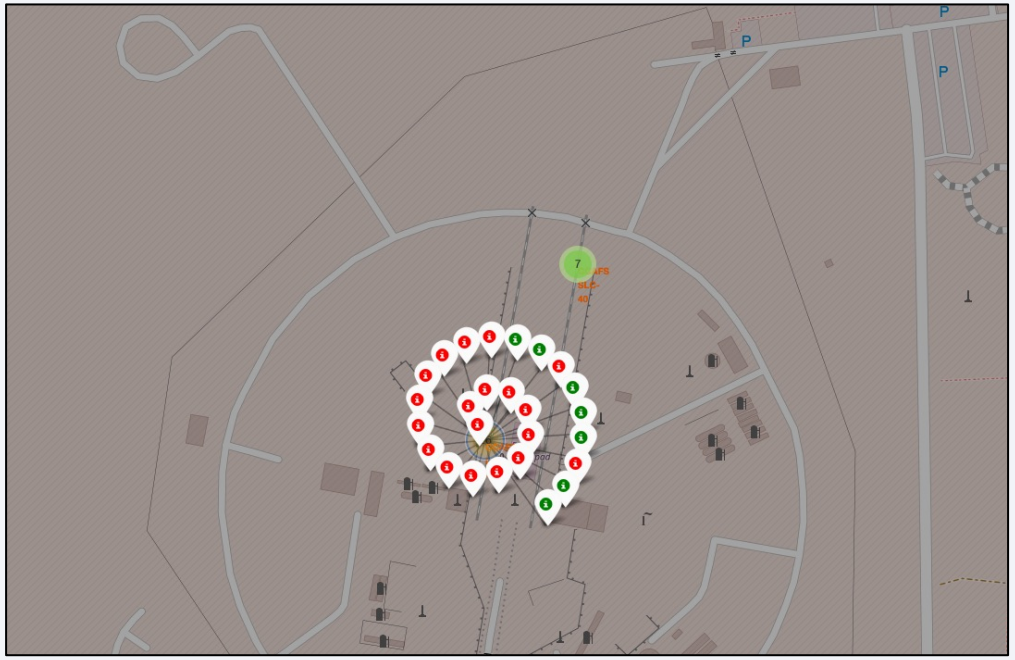

Section 3
Launch Sites
Proximities Analysis

Folium Map of US Launch Sites



- Launch sites typically located near West or East coastline

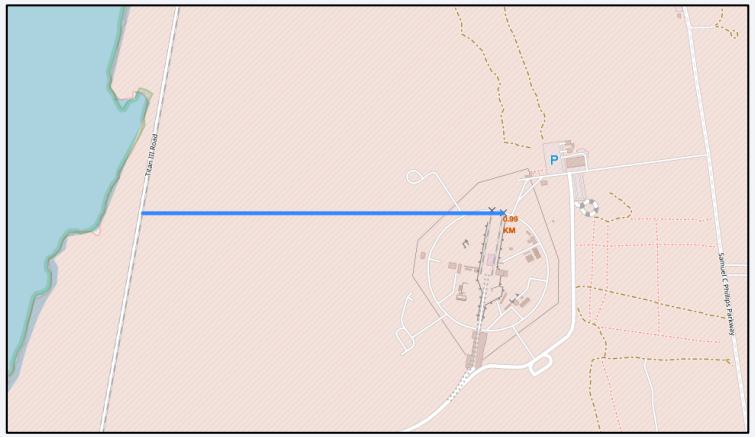
Launch Outcomes by Site



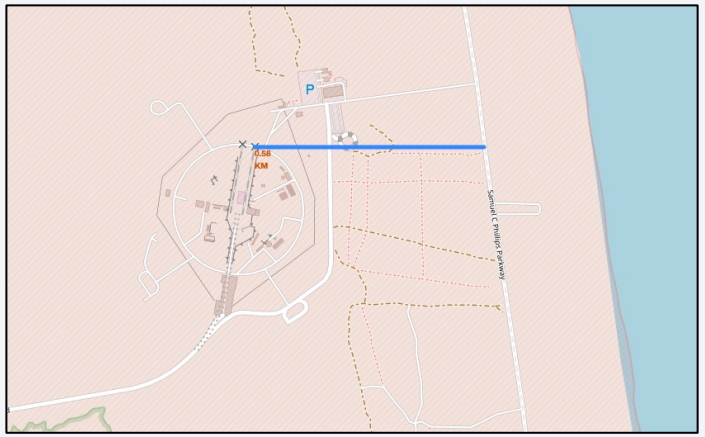
- Colored marker represent launch result; Green markers represent successful launches; red indicate failed launches

Proximity of Railway/Highway/Coastline to Launch Sites

Distance to Coastline

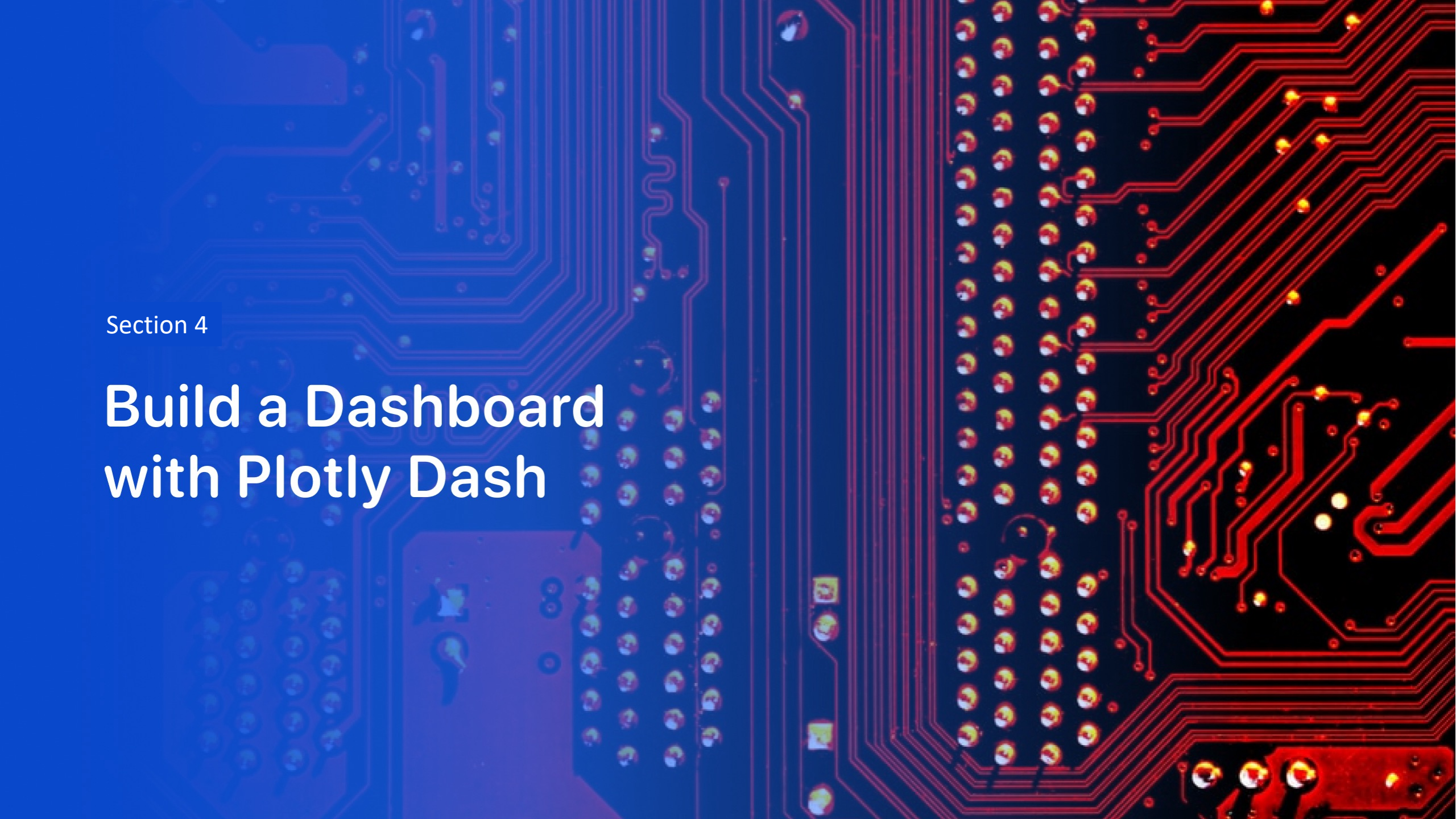


Distance to Railway

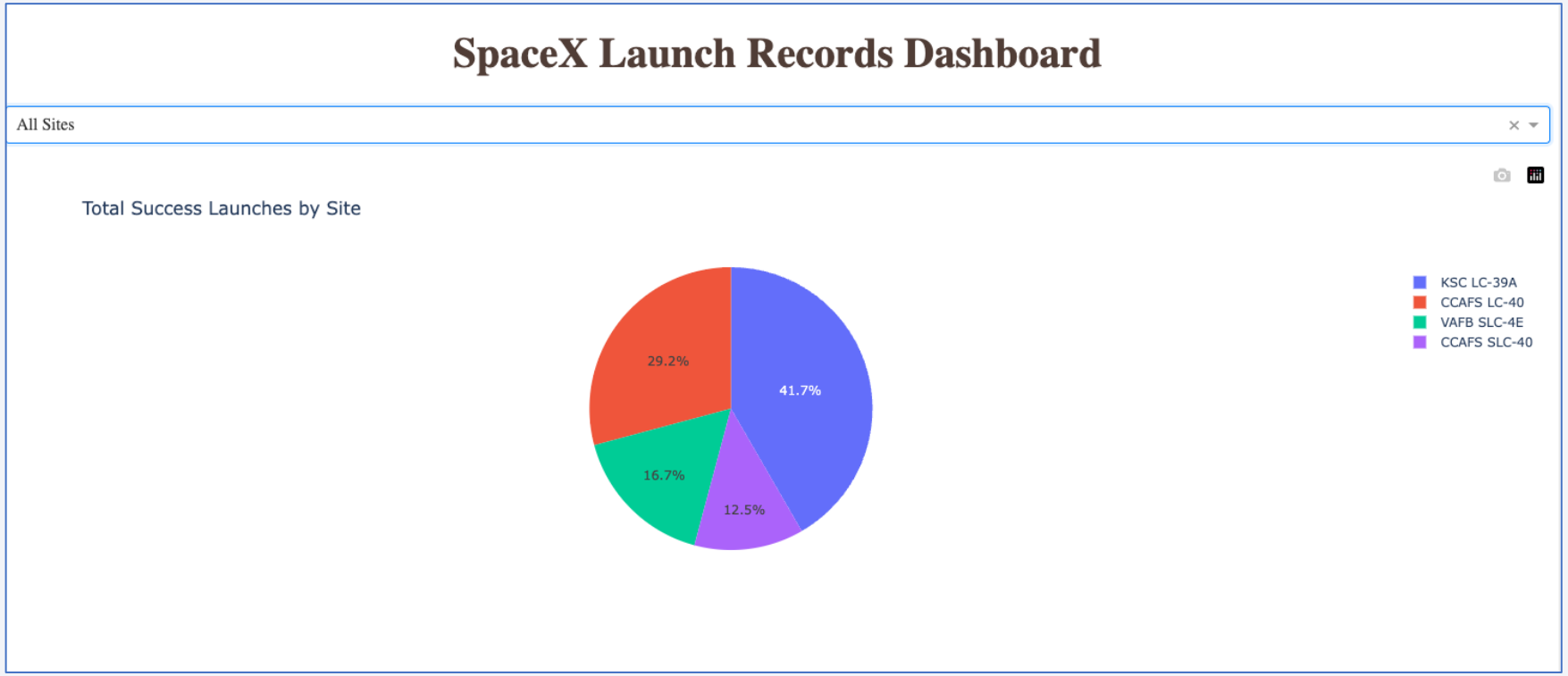


Distance to Highway

- Calculating distance of coastline, highway and railways to launch site is critical to determining launch setup and recovery aspects

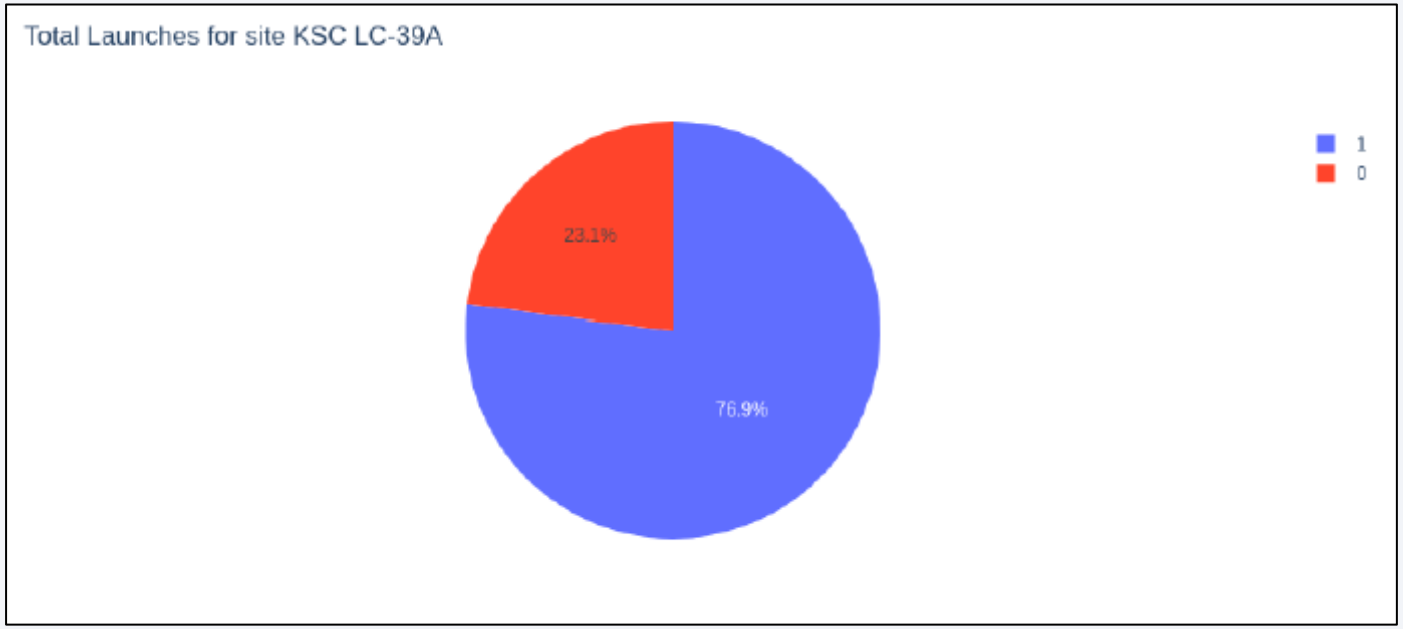

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



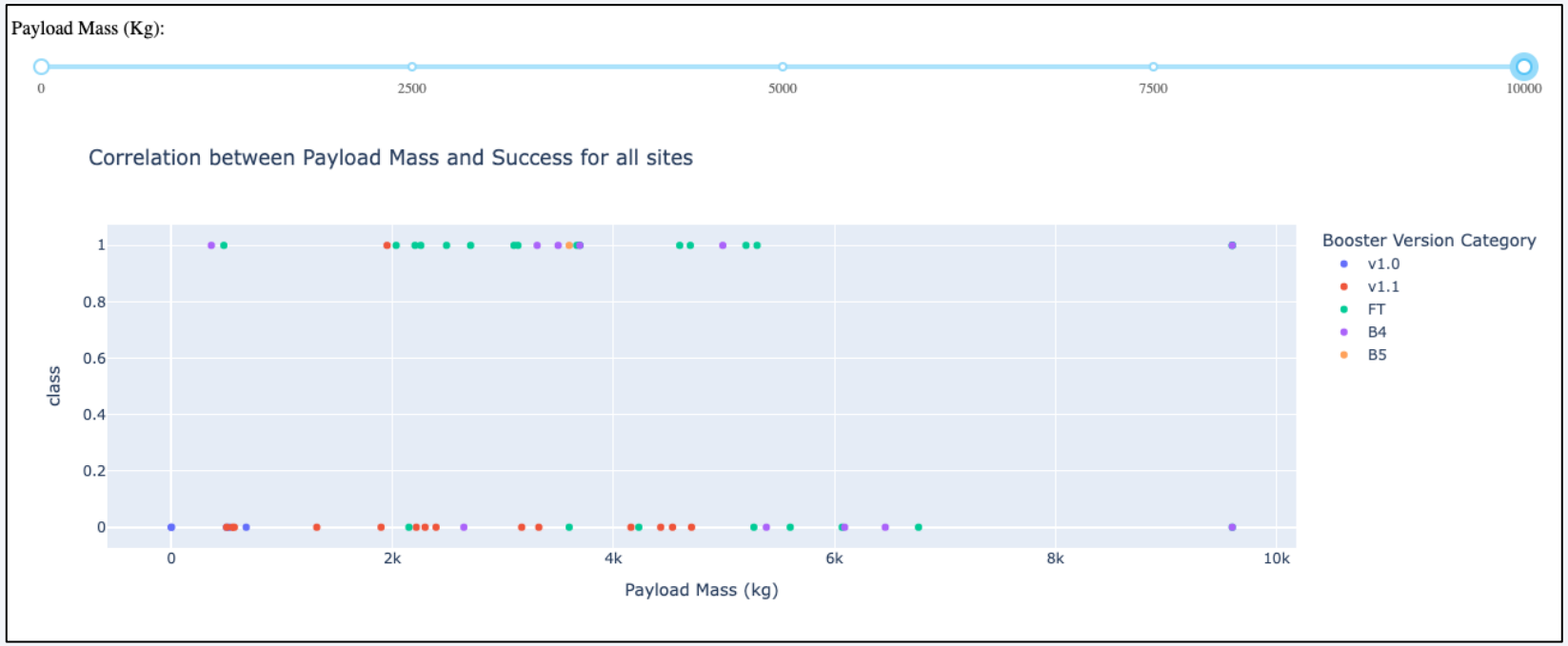
- Pie chart provides a quick way to assess which site has the most successful launches and provides actual percentage breakdown for each site
- Based on the chart, launch site KSC LC-39A has had the most successful launches

Highest Launch Success Site = KSC LC-39A



- Based on the chart, KSC LC-39A site had the most successful launches with 76.9%

Payload vs Launch Outcome



Conclusions:

- Payloads between 2K – 6K have the highest success rate
- Payloads between 6K – 8K have the lowest success rate
- Booster version FT has the highest launch success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

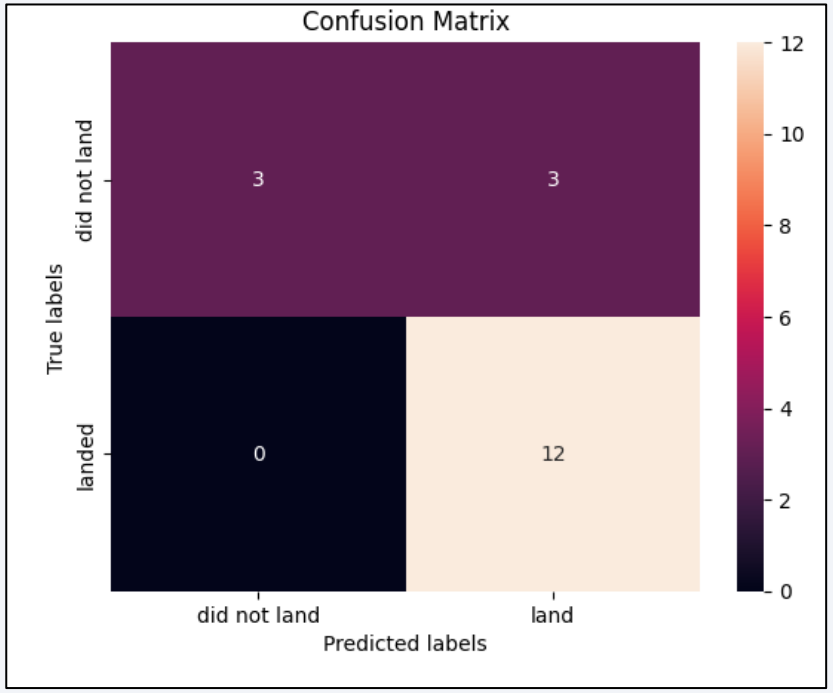
- Four classification models were tested: LogReg, SVM, Decision Tree and KNN.
- Decision Tree classification model was the most accurate at 87%

Find the method performs best:

```
print('Logistic Regression Score', logreg_cv.best_score_)
print('SVM Score', svm_cv.best_score_)
print('Tree Score', tree_cv.best_score_)
print('KNN Score', knn_cv.best_score_)
```

```
Logistic Regression Score 0.8464285714285713
SVM Score 0.8482142857142856
Tree Score 0.8732142857142856
KNN Score 0.8482142857142858
```

Confusion Matrix of Decision Tree Classification Model



- Decision Tree classification model was the best model as it demonstrated the largest number of true positive and true negative outcomes

Conclusions

- The success rate of launches increases the more launches conducted per site
- The best launch site is KSC LC-39A with the highest success rate
- Launch Payloads greater thank 7,00ks are more successful
- Orbits ES-L1, GEO, HEO, SSO and VLEO have the highest success rate
- Decision Tree classifier is the best ML model for this particular use case

Thank you!