IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SARTHAK JAIN
16th June 2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of all results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

- Project background and context

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers

- What factors determine if the rocket will land successfully?
- The interaction amongst various features that determine the success rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Filling missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, train, and evaluate models using various algorithms

Data Collection

- Describe how data sets were collected.

We collected data using a combination of SpaceX REST API requests and web scraping from a table in SpaceX's Wikipedia entry. We had to utilize both methods to gather comprehensive information about the launches for a more thorough analysis.

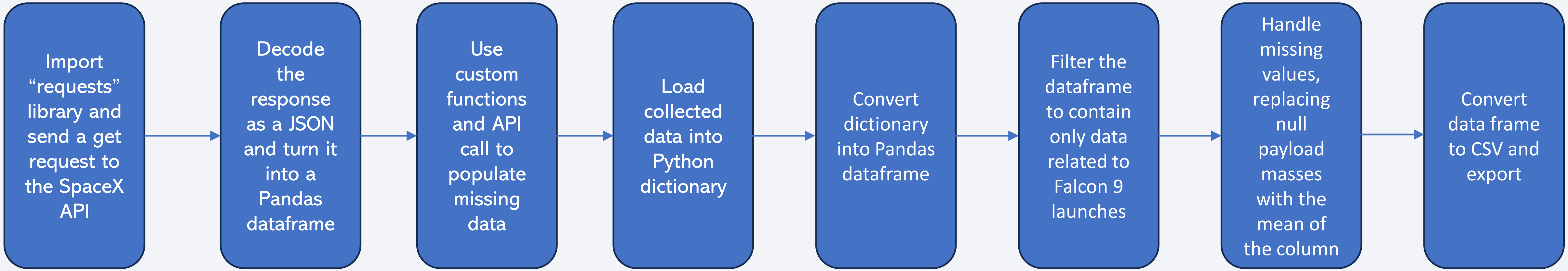
- You need to present your data collection process use key phrases and flowcharts

The following data columns were obtained using SpaceX REST API: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude.

Data Collection – SpaceX API

- Data collected via API followed the below process.
- Launch data collected included date, payload size, rocket booster version, launch site, and launch outcome.

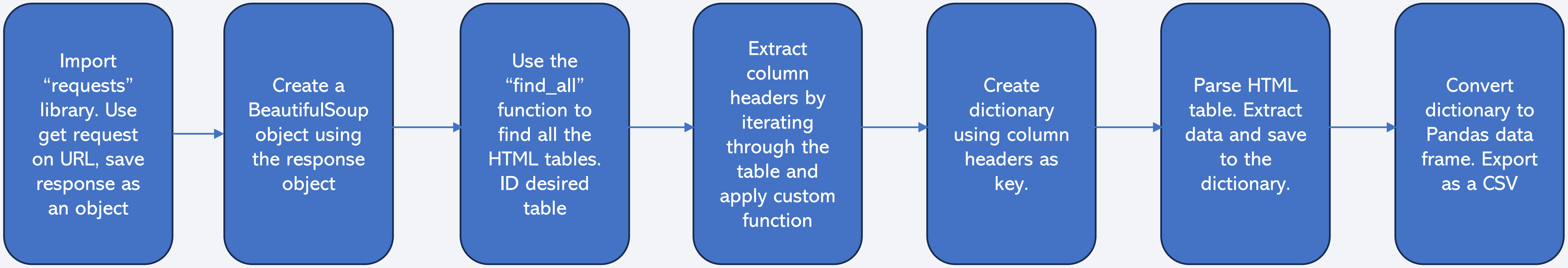
SpaceX REST API process



Data Collection - Scraping

- Data collected via web scraping followed the below process.
- Launch data collected included date, payload size, rocket booster version, launch site, and launch outcome.

SpaceX REST API process



Data Wrangling

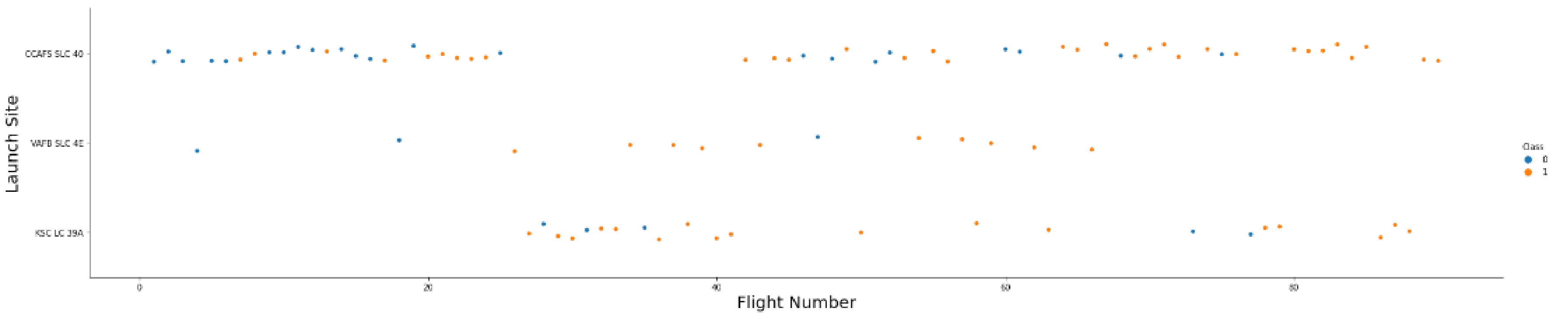
- Conducted data wrangling on the collected data.
- The purpose of data wrangling was to perform initial exploratory data analysis (EDA) and identify potential patterns in the data and define labels for training supervised learning models.
- Tasks in this step included:
  - Calculating the number of launches at each site
  - Calculating the number and occurrence of each launch orbit
  - Calculate the number and occurrence of mission outcome per orbit type.
  - Calculating the number of each landing outcome
  - Creating a binary landing outcome label
- For the landing outcome label, “1” represents the first stage booster successfully landed, and “0” represents the booster was unsuccessful in landing.

EDA with Data Visualization

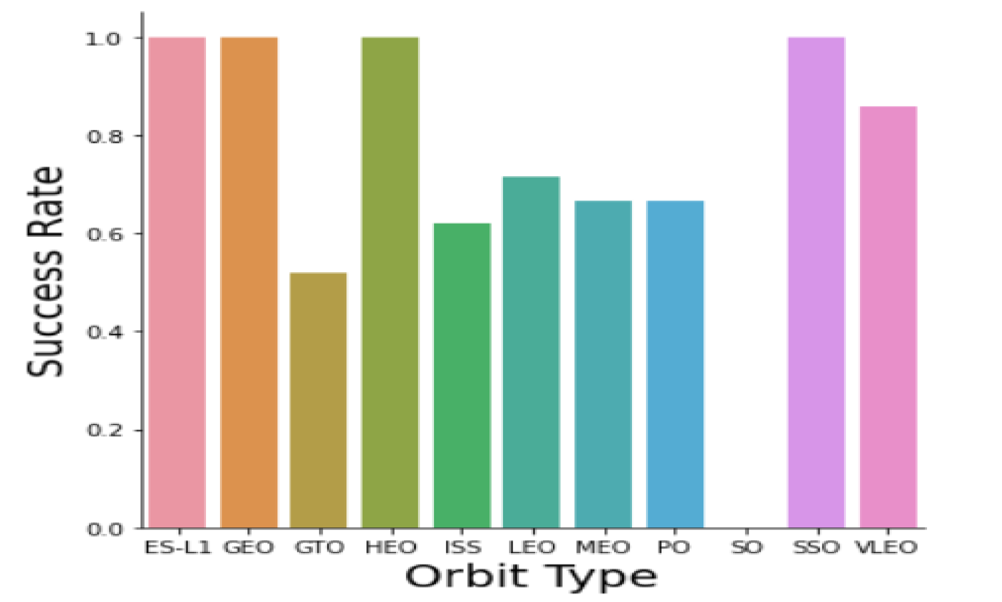
- As part of the EDA process, several plots were created to examine trends in the data.
- Scatter Plot: Show the relationship/correlation between two variables. Used to identify patterns. The following scatter plots were created:
  - Flight Number vs Payload Mass, with color indicating launch outcome
  - Flight Number vs Launch Site Location, with color indicating launch outcome
  - Payload Mass vs Launch Site Location, with color indicating launch outcome
  - Flight Number vs Orbit Type, with color indicating launch outcome
  - Payload Mass vs Orbit Type, with color indicating launch outcome
- Bar Chart: Used to compare values among discrete categories. The bar chart created for this analysis illustrated success rate for each launch orbit type.
- Line Chart: Typically used to show time series trends. The line chart created for this analysis illustrated annual success rate over time (from 2010-2020)

EDA with Data Visualization - Continued

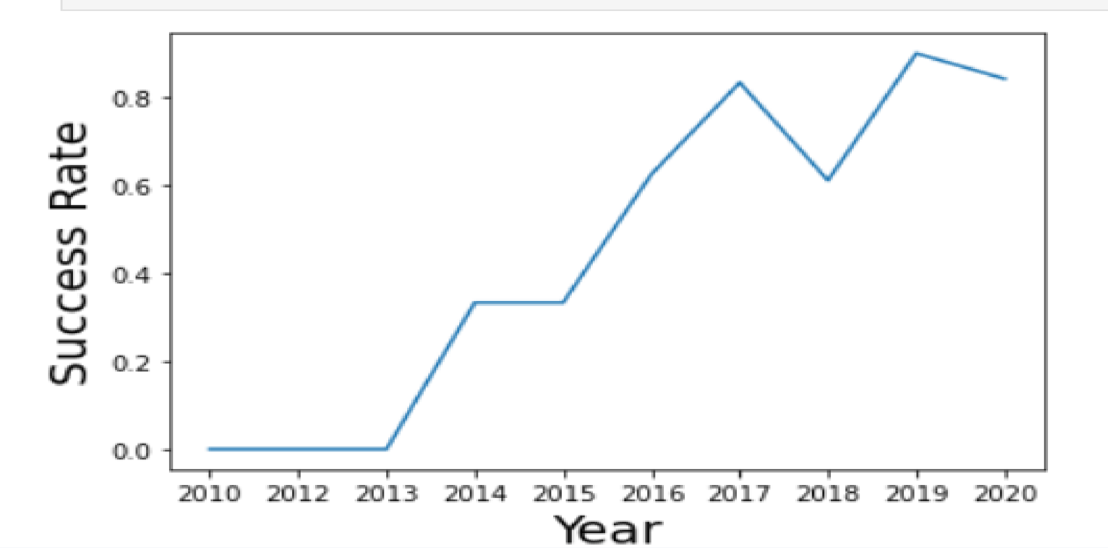
- Scatter Plot:



- Bar Chart:



- Line Chart:



EDA with SQL

- Used SQL to conduct additional EDA on launch data, performing various queries to better understand the data and identify any trends or patterns.
- The following queries were performed:
  - Display the names of the unique launch sites.
  - Display 5 records where launch sites begin with the string “CCA”.
  - Display the total payload mass caried by boosters launched for NASA (CRS).
  - Display the average payload mass carried by F9 v1.1 boosters.
  - List the date when the first successful landing outcome on a ground pad was achieved.
  - List the names of the boosters which landed successfully on a drone ship and have a payload mass between 4000 kg and 6000 kg.
  - List the total number of successful and unsuccessful mission outcomes.
  - List the names of the booster versions which carried the maximum payload mass
  - List records that failed landings on drone ships in 2015
  - Rank the count of landing outcomes between 06/04/2010 and 03/20/2017 in descending order

Build an Interactive Map with Folium

- Built an interactive map using the Folium library to illustrate geospatial data related to the launches.
- First, created to a map to show all launch sites.
  - Added circles to denote the location of each launch site, with a popup label displaying the site name.
  - Added markers to display the name of the launch site by each circle.
- Next, indicated the result of the launches at each site.
  - Added markers for each launch and added color to indicate success (green) or failure (red).
  - Created marker clusters at each site to improve readability.
- Last, calculated distance from each launch site to nearby points of interest (highway, railroad, airport, etc.).
  - Added MousePosition to determine coordinates and wrote function to calculate distances between coordinates.
  - Added a PolyLine between site CCAFS SLC-40 and the coastline, with distance as the label.
  - Added a PolyLine with distance between site VAFB SLC-4E and the nearest railroad.
  - Added a PloyLine with distance between site VAFB SLC-4E and the nearest city.

Build a Dashboard with Plotly Dash

- Used Plotly Dash to build an interactive dashboard, allowing users to adjust parameters and see updated charts in real time.
- Created a Pie Chart with a dropdown menu listing the launch sites.
    - When all launch sites selected, pie chart displays the percent of successful launches at each site.
    - When a single launch site selected, pie chart displays number of successes and failures at that site.
    - This is a useful visualization for identifying which site experienced the most successful launches.
- Created a Scatter Chart of Payload Mass vs. Launch Outcomes for each Booster version.
    - Displays any correlation between payload mass and success rates.
    - Coloring points by Booster version provides additional information which Boosters have the highest success rates.
    - Created range slider for Payload Mass, allowing the user to set a range for the x-axis on the chart.

Predictive Analysis (Classification)

Create a NumPy array from the "Class" data (value want to predict)

→ Standardize , fit, and transform the data using StandardScaler. Save to "X"

→ Split the data into training and test data sets. (80% training, 20% test)

→ Train a logistic regression model, using GridSearchCV to find the best parameters.

→ Follow the same process to train SVM, Decision Tree, and KNN Models.

→ Calculate accuracy using .score() and create confusion matrix for each model.

→ Compare scores to determine the best performing model.

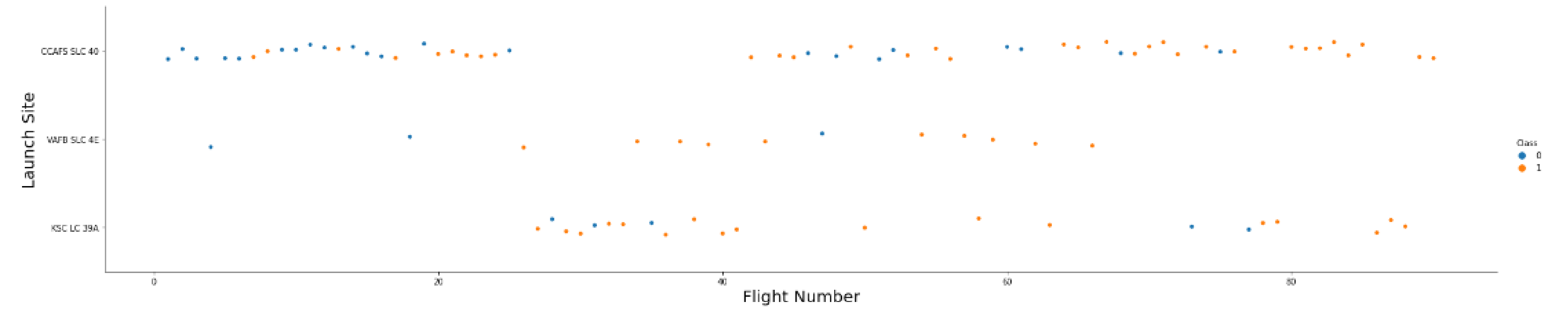
Results

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

Section 2

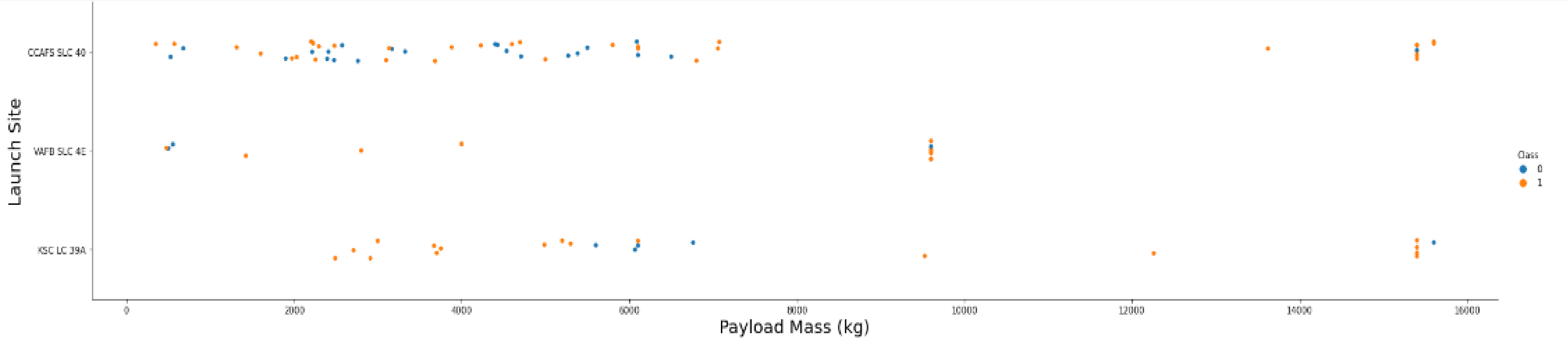
Insights drawn from EDA

Flight Number vs. Launch Site



- Flight numbers are on the x-axis, launch sites are on the y-axis, with blue data points indicating mission failure and orange data points indicating mission success.
- Site CCAFS SLC 40 had the highest number of launches, including 18 of the first 20 launches.
- Success rate improved over time, with early launches having a high failure rate, and later launches experiencing higher success rates.

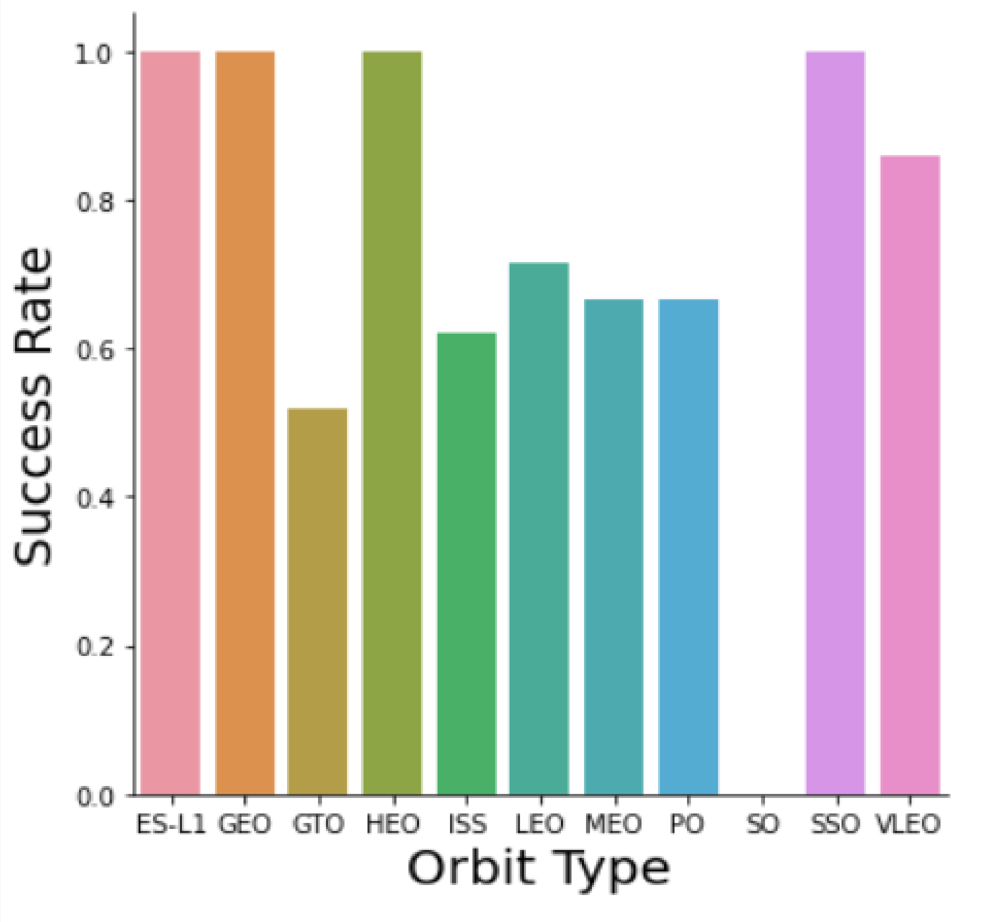
Payload vs. Launch Site



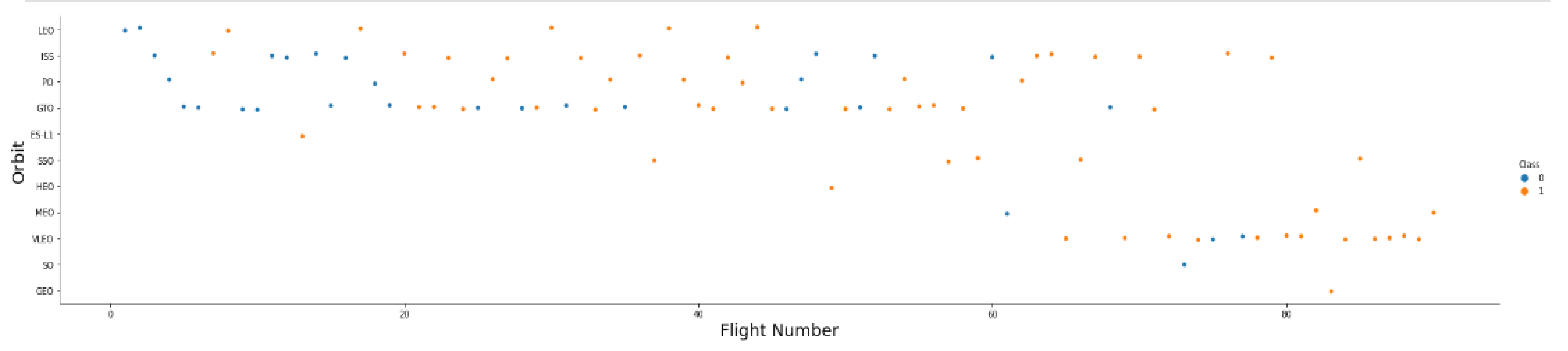
- Payload Mass (in kg) is on the x-axis, Launch Site is on the y-axis, with blue data points indicating failure, and orange data points representing success.
- The majority of the launches carried payloads less than 7,000 kg.
- Site VAFB SLC 4E did not launch a rocket with a payload greater than 10,000 kg.
- High payload launches (greater than 8,000 kg) experienced a high success rate.

Success Rate vs. Orbit Type

- Orbit type is the x-axis, success rate is on the y-axis.
- ES-L1, GEO, HEO, and SSO had the highest success rates at 100%.
- SO had the lowest success rate, at 0%.
- GTO, ISS, LEO, MEO, and PO all had success rates between 50% and 80%.



Flight Number vs. Orbit Type

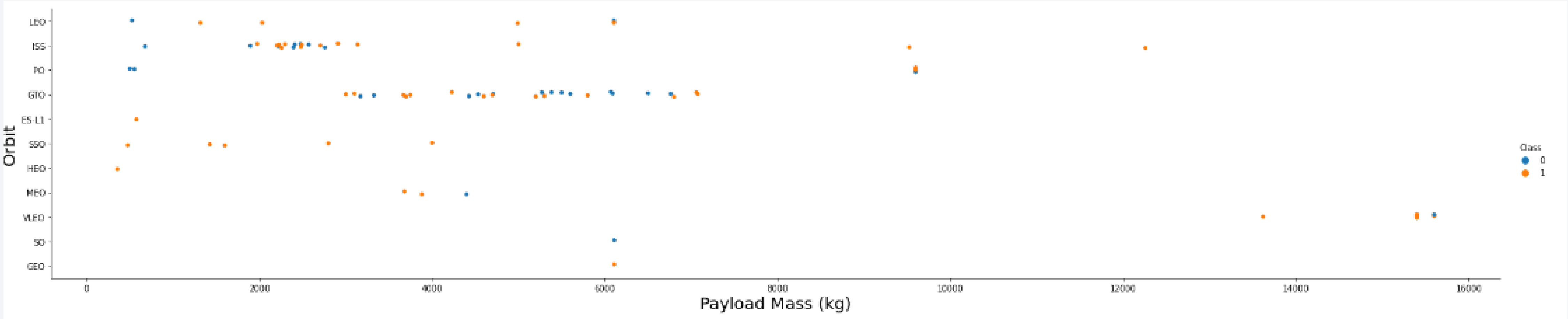


- Flight number is on the x-axis, orbit type is on the y-axis, with blue data points indicating mission failure and orange data points indicating mission success.

- Majority of launches up to flight 55 had orbits of LEO, ISS, PO, or GTO.

- For LEO, success rate appears to improve over the launches, while GTO does not demonstrate a clear relationship.

Payload vs. Orbit Type



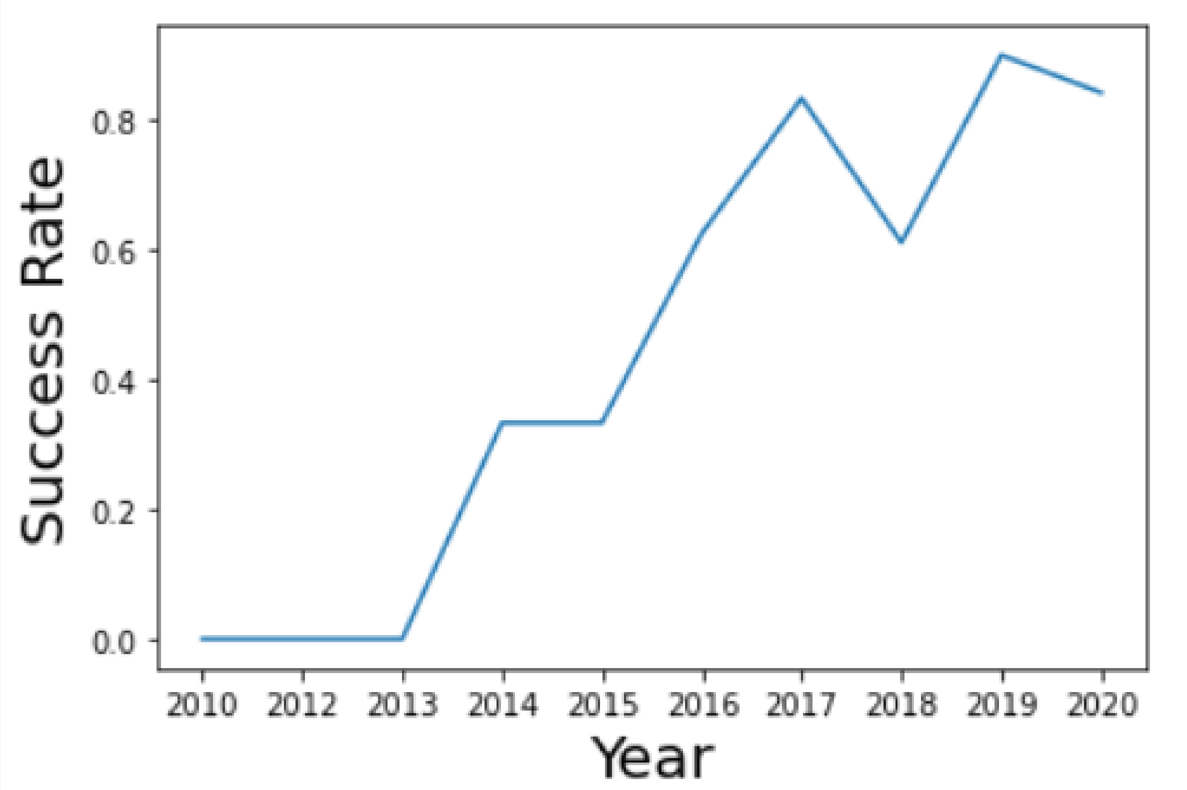
- Payload Mass (in kg) is the x-axis, orbit type is the y-axis, with blue data points indicating mission failure and orange data points indicating success.

- Success rates for PO, ISS, and LEO increase as payload mass increases.

- GTO does not display any clear correlation between success and payload mass.

Launch Success Yearly Trend

- Year is the x-axis, success rate is the yaxis.
- Launches from 2010-2013 had a 0% success rate.
- Success rate improved between 2013-2020.
- There is drastically improvement between 2015-2017.
- There is a drop in success rate between 2017-2018



All Launch Site Names

- Query : select distinct launch_site from SPACEXDATASET
- Result :

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

Launch Site Names Begin with 'CCA'

- Query : select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
- Result :

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

Total Payload Mass

- Query : select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)’;
- Result :

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Query : select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
- Result :

average_payload_mass
2534

First Successful Ground Landing Date

- Query : select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
- Result :

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query : select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
- Result :

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

Total Number of Successful and Failure Mission Outcomes

- **Query** : select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
- **Result** :

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

Boosters Carried Maximum Payload

- **Query** : select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
- **Result** :

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

- Query : **select** monthname() **as month**, , booster_version, launch_site, landing__outcome **from** SPACEXDATASET **where**landing__outcome **=** 'Failure (drone ship)' **and year**()**=**2015;
- Result :

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Query : **select** landing__outcome, **count**(*) **as** count_outcomes **from** SPACEXDATASET **where** **between** '2010-06-04' **and** '2017-03-20' **group by** landing__outcome **order by** count_outcomes **desc**;
- Result :

landing__outcome	count_outcomes
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

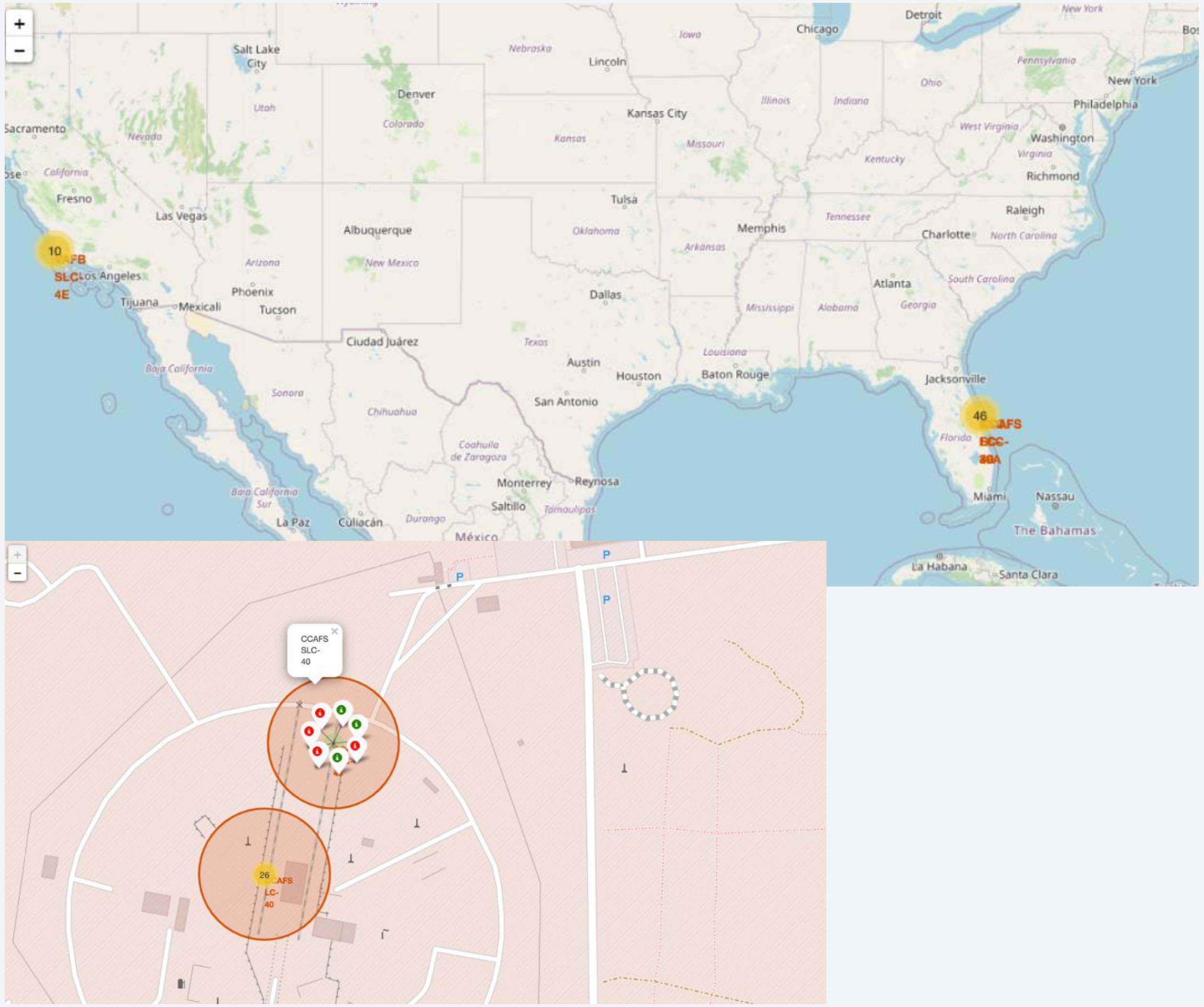
Launch Sites Proximities Analysis

Map Of All SpaceX Falcon 9 Launch Site

- This map shows the location of the two launch sites.

- All launch sites are in the southern portion of the United States and are close to the coast.

Launch Outcome By Site

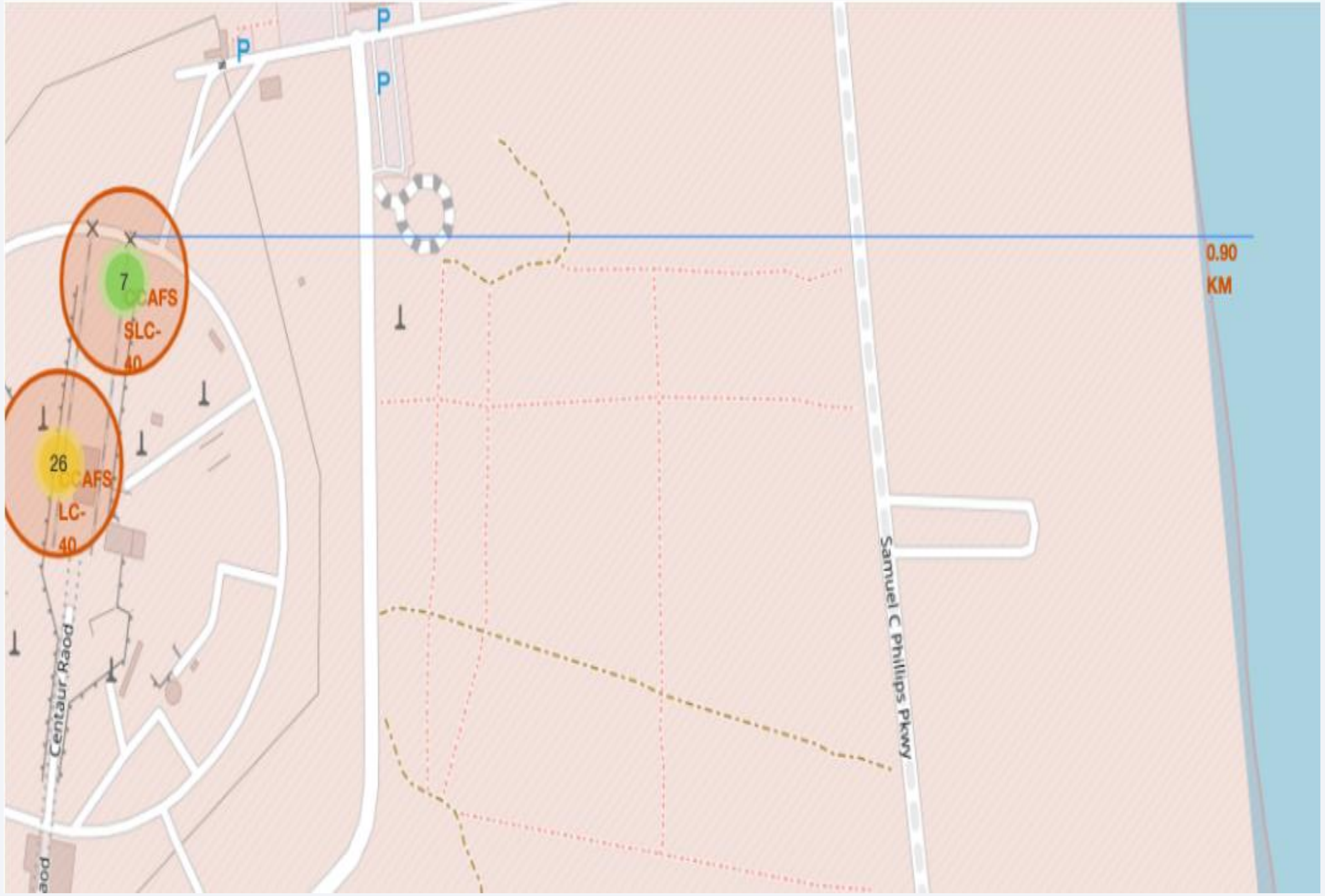


- Added Marker Clusters to each launch site to indicate the number of launches at each site.

- The top map illustrates the small scale view. Yellow circles represent the clusters, the number showing the number of launches.

- The bottom map shows a zoomed in view of the VAFB SLC 4E launch site. Markers in the cluster are assigned a color:
  - Red – Failed landing
  - Green – Successful landing

Launch Site Distance To Landmark



- All launch sites are near the coast to launch rockets over the water and are near a major transportation route (highway/railroad)

Section 5

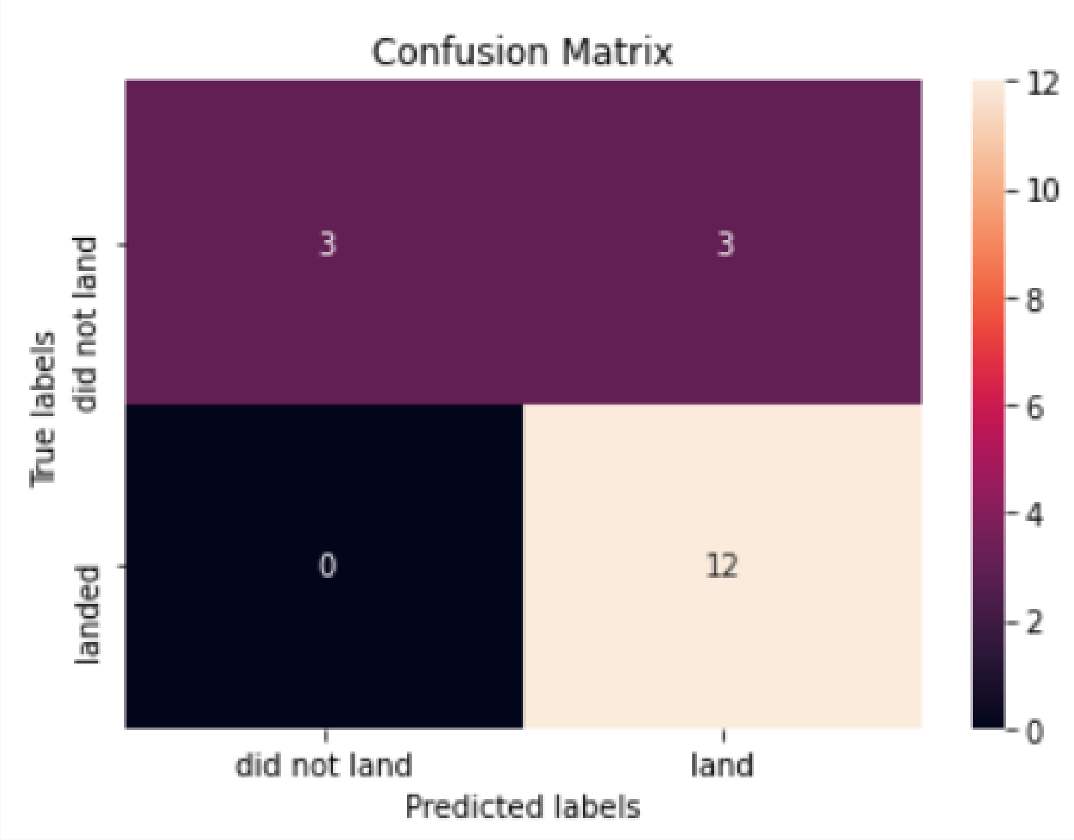
Predictive Analysis (Classification)

Classification Accuracy

- The Decision Tree Classification Model scored the best of the four models.
- All four models have similar accuracy scores.
    - Highest = Decision Tree (0.889)
    - Lowest = Logistic Regression (0.846)
- All models have the same accuracy score on the test data set (0.833).
- As new data becomes available for training, one model may appear as the definitive best.

Confusion Matrix

- Models predicted the outcome of 18 launches.
  - Accurately predicted 15 of 18 outcomes. (83.3%)
  - 3 of the predicted successes failed. (16.7%)



Conclusions

- Findings from Exploratory Data Analysis (EDA):
  - As more rockets are launched, success rate improves (flight number and success rate positively correlated).
  - ES-L1, GEO, HEO, and SSO orbits had the highest success rates (100%).
  - Success rates improved from 2013-2020, from 0% to ~80%.
- Findings from Proximities Analysis:
  - Launch sites are in the southern United States, as near the equator as practical.
  - Launch sites are near the coast and a major highway or railroad.
- From Predictive Analysis:
  - Decision Tree Classification scored the best, but all four models performed similarly well.
  - All models experienced Type I errors, which is the less desirable error and can result in underestimate costs.
  - As new data is available, using it to train/test the data should improve results.

Thank you!