Winning Space Race with Data Science

Carlos Moore
8/2/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Tables were constructed via web scraping Wikipedia and API calls to the Space-X API.
  - Data wrangling was employed to deal with null values, inappropriate data types, and missing attributes.
  - Performed EDA in SQL and visualization in MatPlotLib and Seaborn
- Summary of all results
  - Produced several models to predict Falcon 9 stage 1 successful landings
  - Logistic Regression & Support Vector Machines yielded the best F-1 score for out-of-sample testing

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.
- Problems you want to find answers
    - If we can predict successful landings for first stage components, we can better determine the cost of each launch. Armed with this information, competitors could out bid Space-X for future launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Made get requests to Space-X API to collect several attributes of Space-X launches from 2006 to 2022 including: the mass of payloads, orbits achieved, boosters used, and launch sites.
  - Scraped the Wikipedia page on Space-X launches and extracted to Beautiful Soup object to construct a csv for later analysis.
- Perform data wrangling
  - Records were restricted to only Falcon 9 launches.
  - Replaced or removed missing values and established a target.
  - Chose appropriate data types.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models evaluated: Decision Tree, K-Nearest Neighbor, Support Vector Machine, Logistic Regression

Data Collection

Data was collected by first making a get request to the Space-X API to retrieve a JSON for all past launches; then read into a data frame.

Relevant features were chosen and records were restricted to a maximum date of 2020-11-13.

Records of rockets with multiple cores and payloads were also removed.

Using several pre-defined functions, data about individual launches was extracted to multiple lists for future use.

These lists were used to create a new frame for launch data; where all but Falcon 9 boosters were removed.

Finally, null values for payload mass were replaced by the *attribute mean* value.

Data Collection – SpaceX API

- [Space-X Data Collection API Lab](https://github.com/Dichotmy/IBM-Data-Science-Professional/blob/main/10.%20Applied%20Data%20Science%20Capstone/Labs/1.%20SpaceX%20Data%20Collection%20API%20Lab.ipynb)

- https://github.com/Dichotmy/IBM-Data-Science-Professional/blob/main/10.%20Applied%20Data%20Science%20Capstone/Labs/1.%20SpaceX%20Data%20Collection%20API%20Lab.ipynb

API Call: Data Acquisition

- Objective: Obtain data on Space-X launches
- API Endpoint: Utilized Space-X API to fetch data
- Data Acquired: Details of launches and ID keys for rockets, payloads, launch pads, and cores

Normalize Response

- Formatting: JSON response converted to DataFrame
- Data Handling: Data loaded and normalized for analysis

Parse Data

- Feature Selection: *rocket*, *payloads*, *launchpad*, *cores*, *flight_number, date_utc*
- Data Cleaning:
  - Removed rows of launches with multiple cores and payloads
  - Restrict data to launches occurring before November 13th 2020

API Call: Data Acquisition

- Predefined functions use ID keys to make get requests for more data
- Responses are extracted into a new DataFrame for launch data
- Launch data is filtered for only Falcon 9 rocket launches
- Null values for payload mass are replaced with the payload mass average

Data Collection - Scraping

- **Web Scraping Wikipedia for Space-X Launches**

- https://github.com/Dichotmy/IBM-Data-Science-Professional/blob/main/10.%20Applied%20Data%20Science%20Capstone/Labs/2.%20Web%20Scraping%20Wikipedia%20for%20SpaceX%20Launches.ipynb

Static Request

- Send request to Wikipedia page on Falcon 9 rocket launches
- Transfer contents of response to Beautiful Soup object

Extract Data

- Extract all column names from the HTML table header
- Names Discovered: 'Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome'

Parse Tables

- Create empty dictionary using column headers as keys
- Loop through soup object for data to append to dictionary

Normalize

- Formatting: dictionary converted to DataFrame
- Data Handling: Data loaded into csv for storage

Data Wrangling

Data Cleaning and Handling Missing Values → **Data Types and Conversion** → **Feature Exploration** → **Creating New Feature: *Class***

Identified missing values:
- Significant missing values in *LandingPad* column (28.89%)
- No missing values in other columns

Actions taken:
- Analysis considered missing *LandingPad* values as part of the exploration

Overview of data types:

• Integers	3
• Objects	7
• Floats	4
• Boolean	3

Ensured appropriate data types for analysis

Launch Sites:
- Most frequent: CCAFS SLC 40 (55 launches)

Orbits:
- Majority type: LEO (55 occurrences)

Landing Outcomes:
- Categories include successful and unsuccessful landings

Derived feature *Class* based on landing outcomes:
- Positive outcome (1) for successful landings
- Negative outcome (0) for failed or no landings

Calculated success rate: 66.67%

https://github.com/Dichotmy/IBM-Data-Science-Professional/blob/main/10.%20Applied%20Data%20Science%20Capstone/Labs/3.%20Spacex%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

Visualized the relationship between several features of the Falcon 9 launch data set:

- Cat plot of Payload Mass vs Flight Number differentiated by Class
  - Shows greater payloads attempted in latter flights
- Cat plot of Launch Site vs Flight Number differentiated by Class
  - Shows gaps in launch site usage
- Scatter Plot of Launch Site vs Payload Mass differentiated by Class
  - Vandenberg Air Force Base has the least launches
  - Cape Canaveral Air Force Station used mostly for low to mid-range payloads
- Bar Plot showing the success rate of rocket landings by orbit location
  - Orbits for launches achieving a 100% success rate: ES-L1, SSO, HEO, GEO
- Scatter Plot of Orbit vs Flight Number differentiated by Class
- Scatter Plot of Orbit vs Payload Mass
- Line plot revealing the success rate of launches by year
  - Steady increase in successful landings with a drop in 2018

https://github.com/Dichotmy/IBM-Data-Science-Professional/blob/main/10.%20Applied%20Data%20Science%20Capstone/Labs/5.%20EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

Performed data analysis through Jupyter Notebook on SQL-Lite database:

- Loaded launch data set acquired through previous steps to SQL-Lite database
- Displayed the names of distinct Falcon 9 launch sites
- Displayed 5 records where launch sites begin with the string ‘CCA’
- Displayed total mass of payloads carried for customer NASA(CRS)
    - 45596 KG
- Displayed average payload mass carried by booster version F9 v1.1
    - 2928.4 KG
- Date of first successful landing outcome on a ground pad
    - December 22nd 2015
- Show boosters with successful landings on drone ships with a payload mass between 4000 & 6000 KG
- Provide total counts for all mission outcomes
- Provide list of all booster versions that have carried the maximum payload mass
- Show the months, booster versions, and sites of launches with a failure to land on a

https://github.com/Dichotmy/IBM-Data-Science-

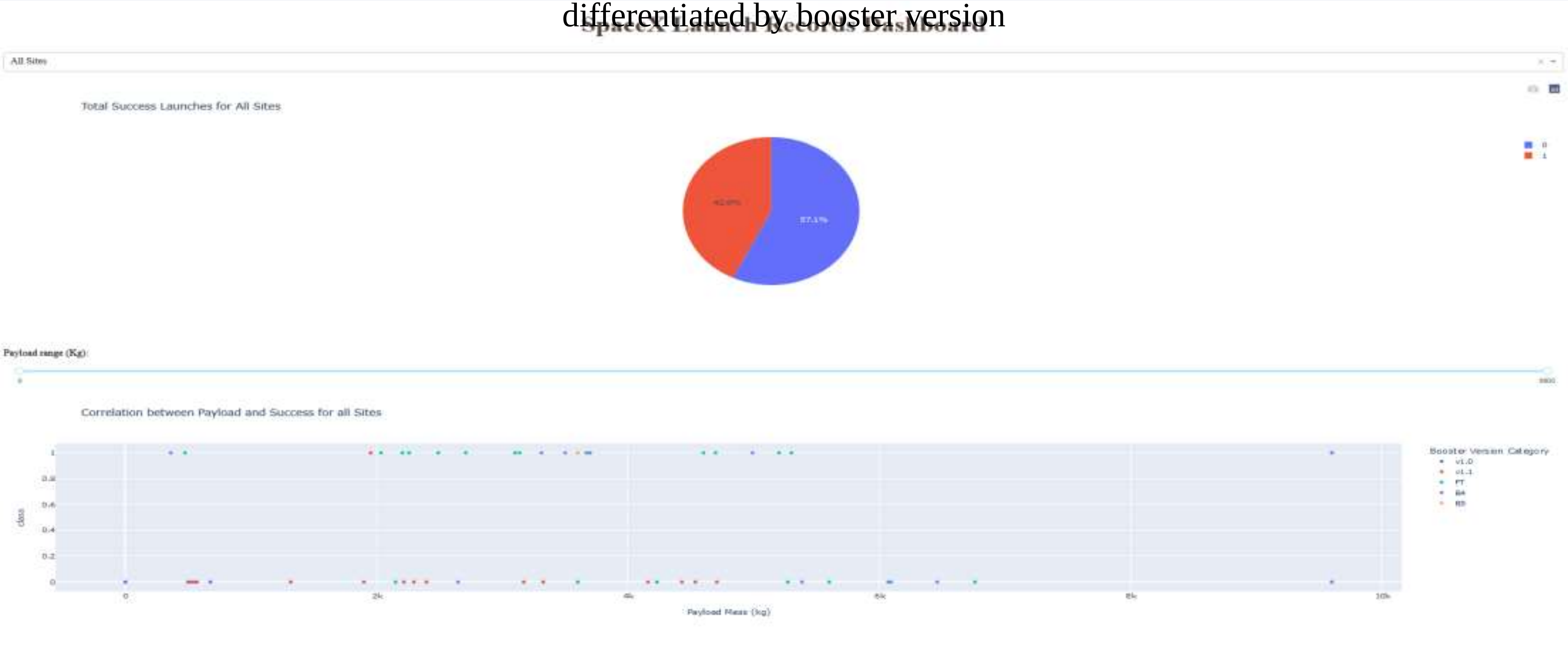
Build an Interactive Map with Folium

Created map markers to display launch site area as well as specific locations

https://github.com/Dichotmy/IBM-Data-Science-Professional/blob/main/10.%20Applied%20Data%20Science%20Capstone/Labs/6.%20Interactive%20Visual%20Analytics%20with%20Folium%20la

Build a Dashboard with Plotly Dash

Plots to show the success ratio of rocket landings by launch site and the correlation between class and outcome differentiated by booster version



https://github.com/Dichotmy/IBM-Data-Science-Professional/blob/main/10.%20Applied%20Data%20Science%20Capstone/Labs/7.%20Build%20a%20Dashboard%20Application%20with%20Plotly%20D

Predictive Analysis (Classification)

Data Collection and Preparation

Load data from CSVs:

- Features: (‘launch data’)
- Target: ('Class')

Target Analysis:

- Assess distribution: value_counts()
- Observation: Class imbalance favoring successful landings by 2 to 1

Data Preprocessing

Convert Target to NumPy Array

Apply StandardScaler() to standardize the features.

Transform Features: fit_transform method used to scale training data.

Data Splitting

Split the data into training and testing sets.

Stratify the data to ensure class proportionality amongst training and testing sets

Parameters used:

- test_size = 0.2
- random_state = 2

Independent Model Tuning

Perform Grid Search for each model independently.

Evaluation Metric: Use accuracy to evaluate model performance

Models Considered:

Decision Tree, K-Nearest Neighbors, Support Vector Machine, Logistic Regression

Model Evaluation

Training & testing accuracy for best model:

LG: 0.85, 0.833 SVM: 0.864, 0.833

DT: 0.918, 0.778 KNN: 0.864, 0.778

Conduct a second concurrent round of Grid Search using F1 score as the evaluation metric for the best models.

Training F1 score:

DT: 0.932 KNN: 0.906

SVM: 0.9 LG: 0.898

Model Validation

Testing: F1, Accuracy, and ROC-AUC:

DT: 0.80, 0.61, 0.60 KNN: 0.83, 0.78, 0.72

SVM: 0.89, 0.83, 0.75 LG:

Evaluation:

Discrepancies in F1 scoring between training and testing for models Decision Tree and KNN suggest over ‘overfitting’.

Selection:

With a good balance between training and testing performance, the **Support Vector Machine** appears to be the best model.

https://github.com/Dichotmy/IBM-Data-Science-

Results

Exploratory Data Analysis reveal a few insights:

- Launch Data shows a steady increase in successful landings with a drop in 2018.
- As confidence in their technology grew Space-X began equipping rockets with heavier payloads.
- All launches achieving the following orbits had a 100% landing success rate and carried light to medium payloads: ES-L1, SSO, HEO, GEO.
- Cape Canaveral Air Force Station saw the highest number of rocket launches.

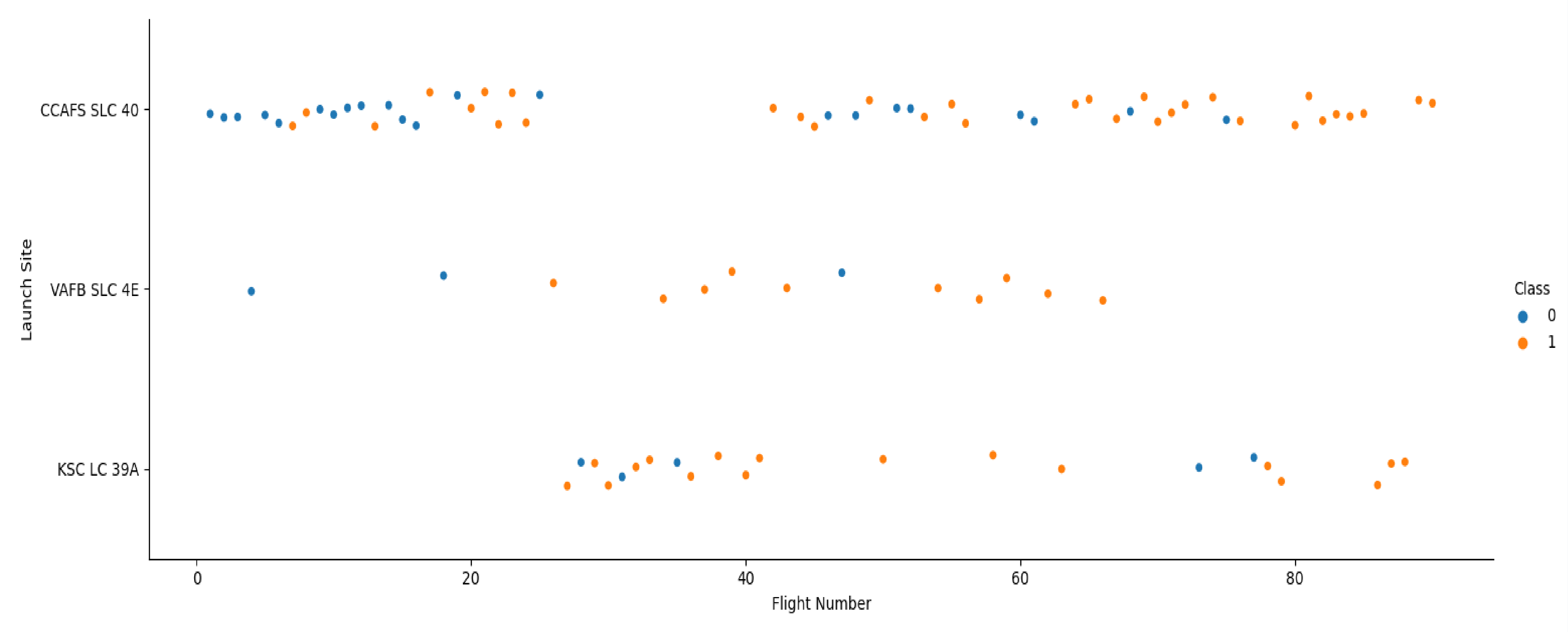
Results of Supervised Machine Learning:

- During independent model evaluation accuracy scores for Logistic Regression and Support Vector Machines showed little difference between training and testing; however Decision Tree and K-Nearest Neighbor dropped by 16.5 & 10.5 percent respectively.
- F1 scoring was used in the training portion of concurrent model evaluation and also for testing along with accuracy and ROC-AUC.
- F1 scoring for SVM and LG models remained virtually identical at approximately 0.9 while DT model performance dropped significantly with a value of 15.3 percent.
- With an F1 score of 0.89, an accuracy of 0.83, and an ROC-AUC of 0.75 the Support Vector Classifier implementing a sigmoid kernel and a gamma of 0.316 was chosen as the best model.

Section 2

Insights drawn from EDA

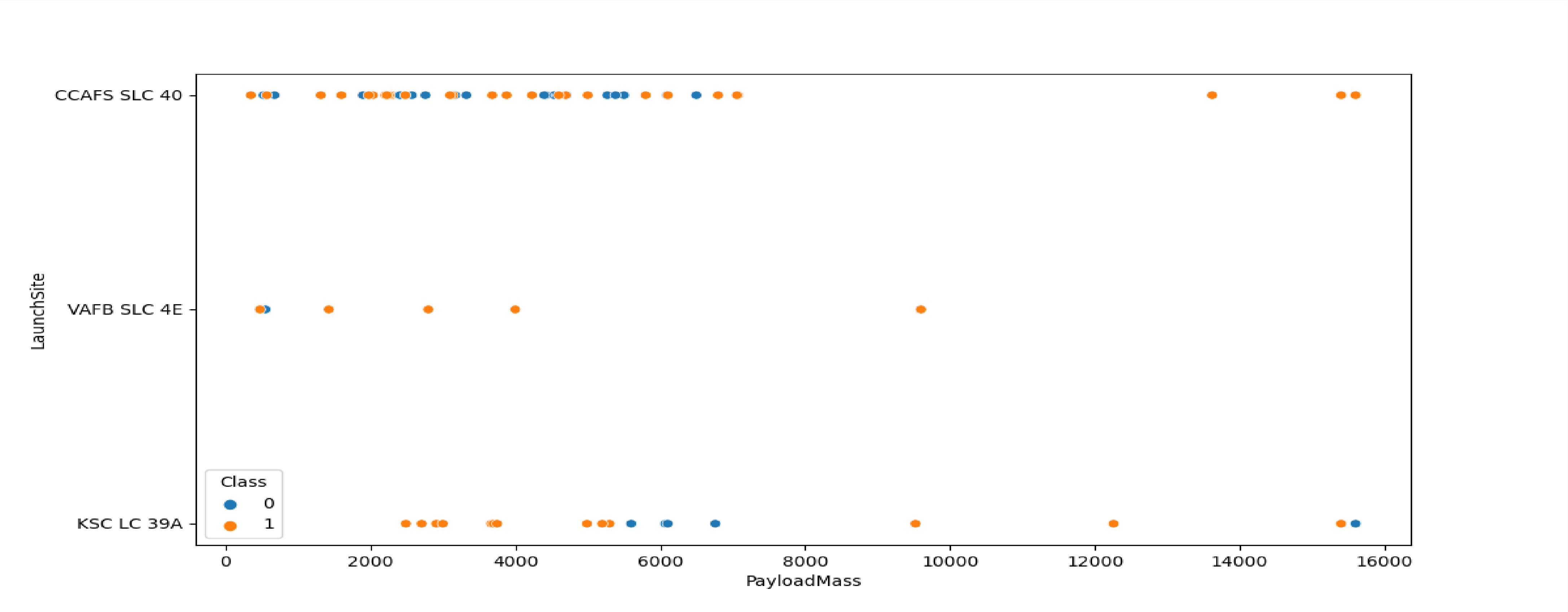
Flight Number vs. Launch Site



Scatter plot segmentation of Launch Site vs Flight Number differentiated by *Class*:

- Shows gaps in usage for all three sites

Payload vs. Launch Site



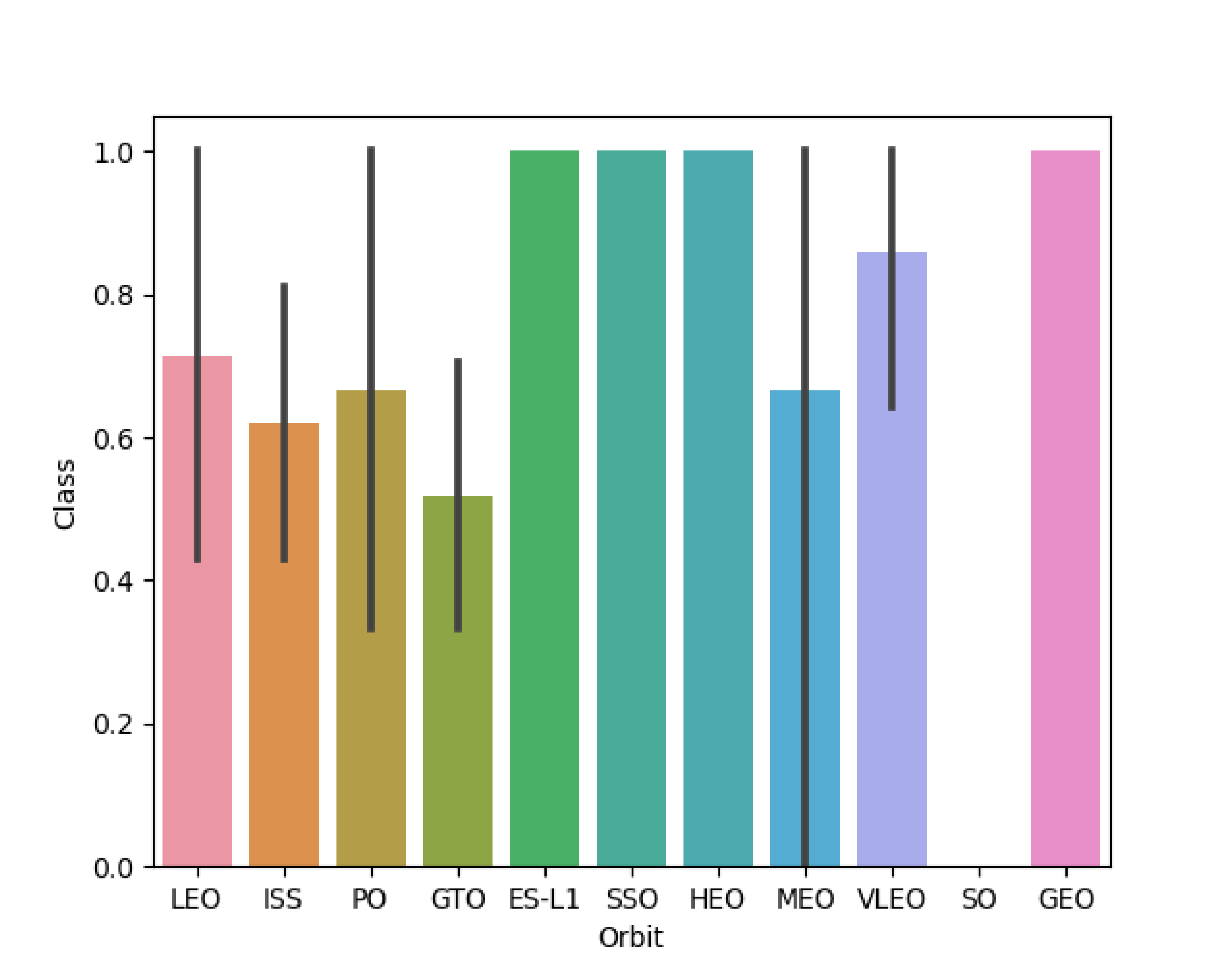
Scatter plot segmentation of Launch Site vs Payload Mass differentiated by *Class*:

- Rockets launched from Kennedy Space Center carry a more even distribution of payload masses
- Vandenberg Air Force Base and Cape Canaveral Space Station launch rockets carrying mostly low to mid-level payload masses

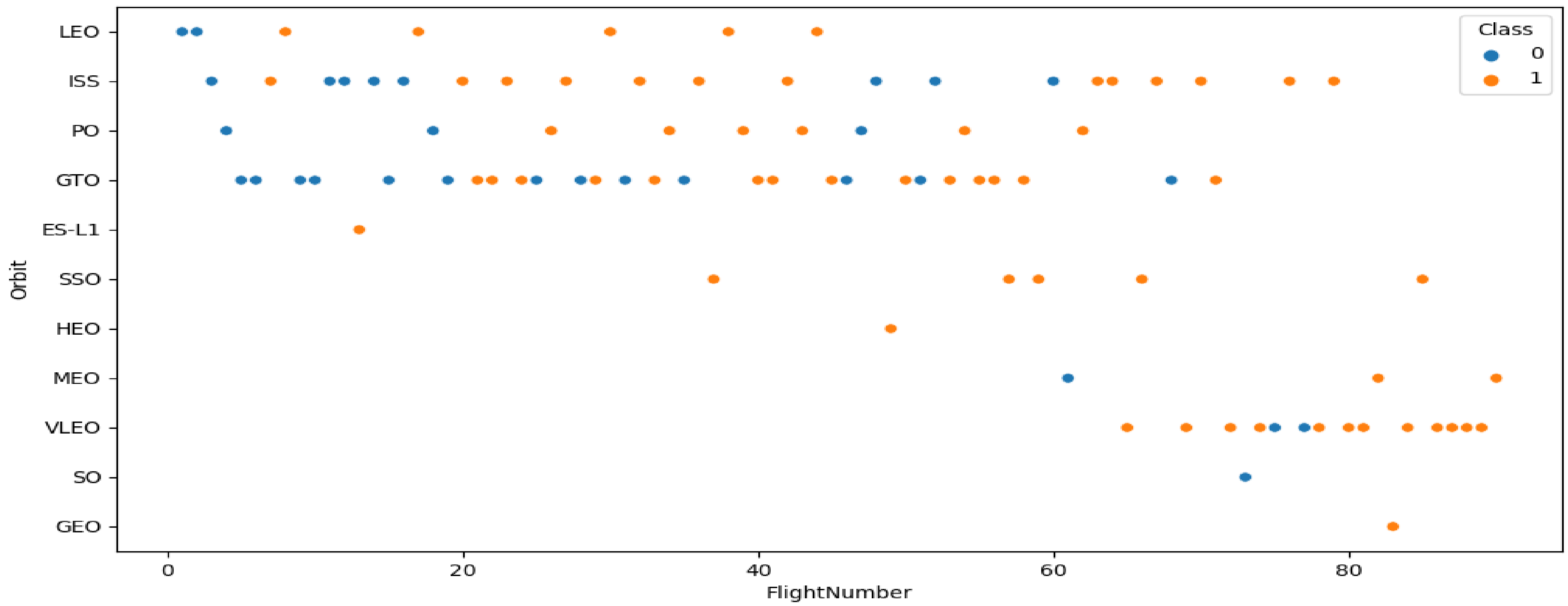
Success Rate vs. Orbit Type

Bar Chart measuring the success rate of rocket landings by type of orbit achieved during launch

- Orbits with 100 % Success:
  - Earth-Sun Lagrange Point 1
  - Sun-Synchronous Orbit
  - Highly Elliptical Orbit
  - Geostationary Earth Orbit



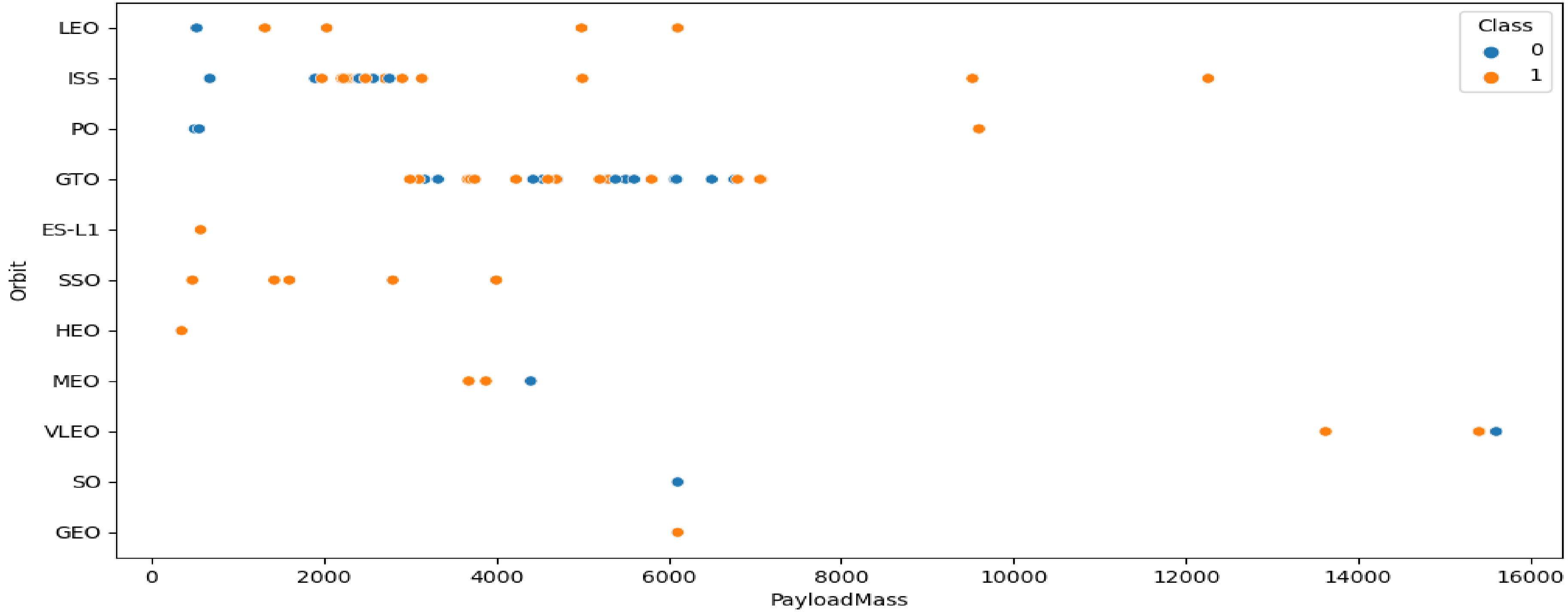
Flight Number vs. Orbit Type



Scatter plot of Orbit types achieved after launch vs Flight Number:

- Orbits ES-L1, HEO, and GEO only have a single launch
- Launches targeting the orbits of MEO, VLEO, SO, and GEO occur later in the flight record

Payload vs. Orbit Type



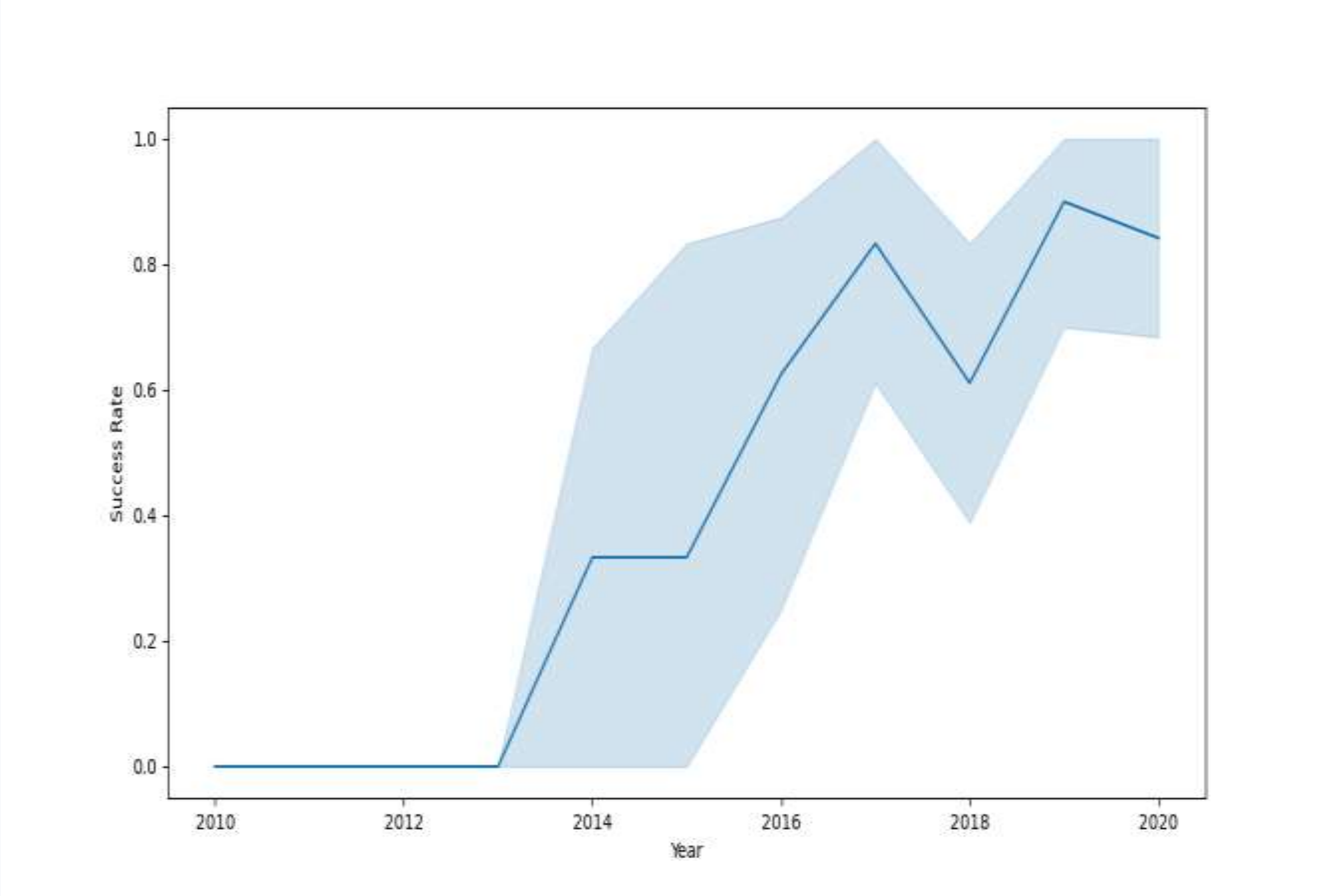
Scatter plot of Orbit types achieved after launch vs Payload Mass:

- **Most orbits were achieved by rockets carrying low to medium payload masses**
- **Rockets for the VLEO orbit carried large payloads**

Launch Success Yearly Trend

Line plot of successful landing rate vs Year:

- Plot shows no successful landings from 2010 to 2013
- There is a steady step-wise increase in successful landings until 2017
- 2018 shows a drop in the success rate leading to another increase for 2019



All Launch Site Names

Unique Launch Sites:

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

Queried the SQL-Lite database with the following command:

- %sql select distinct Launch_Site as `Launch Sites` from SPACEXTABLE

Launch Site Names Begin with 'CCA'

The first 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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon…	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo…	0	LEO (ISS)	NASA (COTS)…	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo…	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

Queried the SQL-Lite database with the following command:

- %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' Limit 5

Total Payload Mass

Total payload mass carried by boosters from NASA (CRS):

- 45596 KG

Queried the SQL-Lite database with the following command:

- %sql select sum(PAYLOAD_MASS__KG_) as `Total Payload Mass for NASA (CRS)` from SPACEXTABLE where Customer == 'NASA (CRS)'

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1:

- 2928.4 KG

Queried the SQL-Lite database with the following command:

- %sql select avg(PAYLOAD_MASS__KG_) as `AVG_MASS (F9 v1.1)` from SPACEXTABLE where Booster_Version == 'F9 v1.1'

First Successful Ground Landing Date

Dates of the first successful landing outcome on ground pad:

- 2015-12-22

Queried the SQL-Lite database with the following command:

- %%sqlselect Date from SPACEXTABLE where lower(Landing_Outcome) like '%success%ground%‘ order by Datelimit 1

Successful Drone Ship Landing with Payload between 4000 and 6000 KG

Names of boosters which have successfully landed on a drone ship and had a payload mass between 4000 & 6000 KG:

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

Queried the SQL-Lite database with the following command:

- %%sqlselect Booster_Versionfrom SPACEXTABLE where lower(Landing_Outcome) like '%success%drone%' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000

Total Number of Successful and Failure Mission Outcomes

Queried the SQL-Lite database with the following command:

- %sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome

Returned:

- **Mission_Outcome** **count(*)**
- Failure (in flight) 1
- Success 98
- Success 1
- Success (payload status unclear) 1

Updated table with the following query to amend error:

- %%sql UPDATE SPACEXTABLE SET Mission_Outcome = 'Success‘ WHERE Mission_Outcome == 'Success '

Boosters Carried Maximum Payload

Names of boosters 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

Queried the SQL-Lite database with the following command:

- %%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

2015 Launch Records

List of months, booster versions, and sites of launches with a failure to land on a drone ship in 2015:

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

Queried the SQL-Lite database with the following command:

- %%sqlSELECT CASE strftime('%m', Date) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March' WHEN '04' THEN 'April' WHEN '05' THEN 'May' WHEN '06' THEN 'June' WHEN '07' THEN 'July' WHEN '08' THEN 'August' WHEN '09' THEN 'September' WHEN '10' THEN 'October' WHEN '11' THEN 'November' WHEN '12' THEN 'December' END AS Month_Name, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE strftime('%Y', Date) = '2015' AND Landing_Outcome = 'Failure (drone ship)';

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

Count of all landing outcomes between 2010-06-04 and 2017-03-20from greatest to least:

Landing_Outcome	Total
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

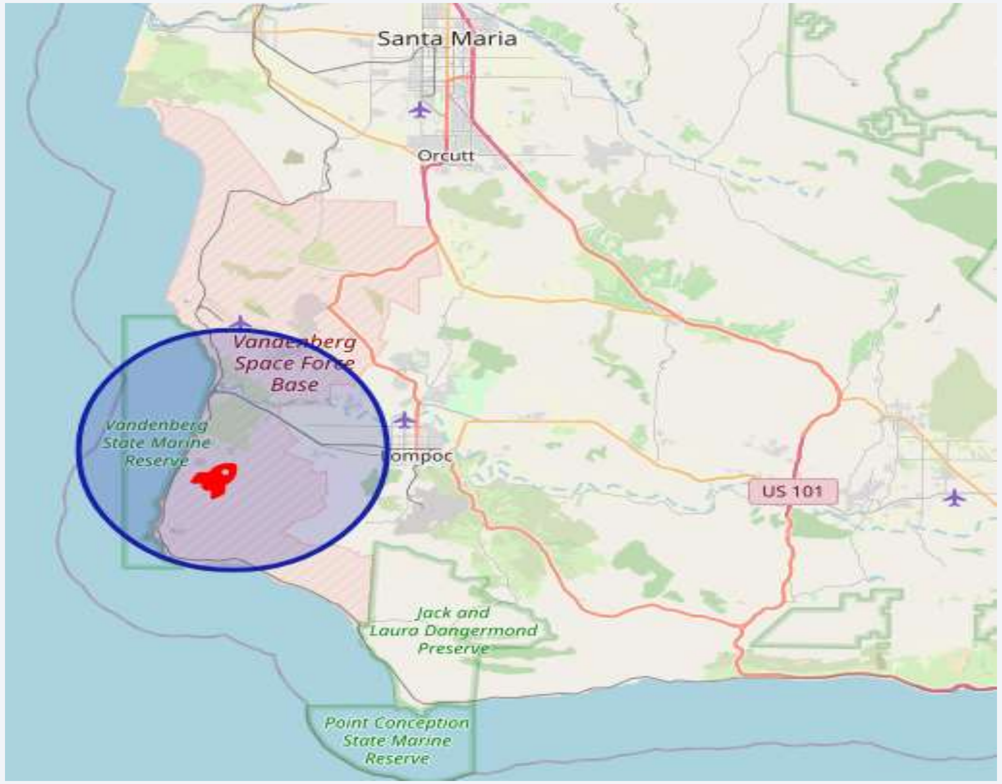
Queried the SQL-Lite database with the following command:

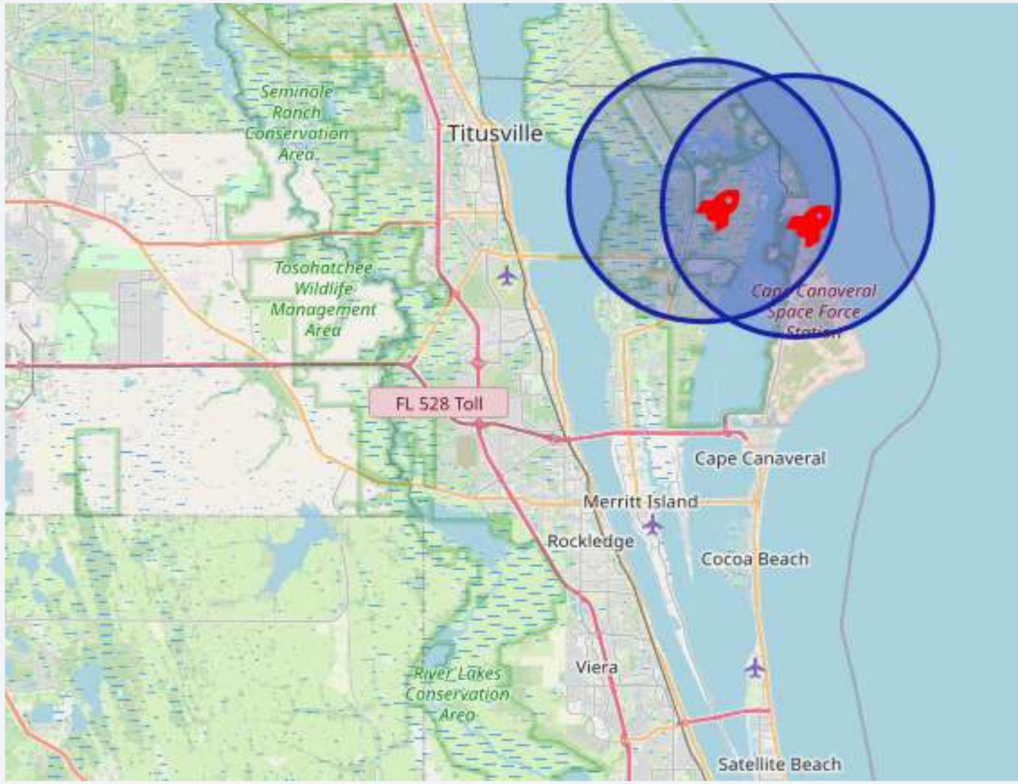
- %%sql SELECT Landing_Outcome, COUNT(*) AS Total FROM SPACEXTABLEWHERE Date BETWEEN '2010-06-04' AND '2017-03-20‘ GROUP BY Landing_Outcome ORDER BY Total DESC

Section 3

Launch Sites Proximities Analysis

Map of Space-X Launch Sites





- Left is a map of the VAFB SLC-4E launch site in Vandenberg Space Force Base
- Right is a map of the CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A launch sites in Cape Canaveral Space Force Station and Kennedy Space Center

Map 2

Marker locations for California and Florida launch sites

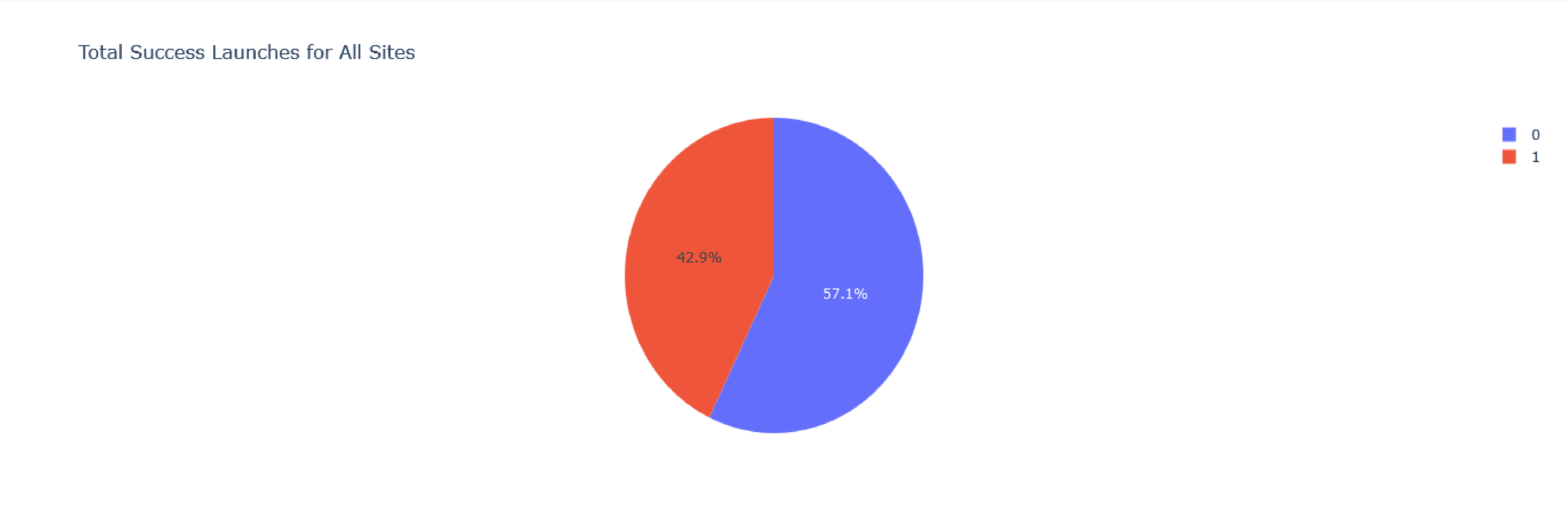
Map 3


Salt Lake City
Nebraska
Lincoln
Ohio
Denver
Utah
Illinois
Indiana
Kansas City
Sacramento
Colorado
West Virginia
Nevada
Kansas
Missouri
Kentucky
San Jose
California
Fresno
Tulsa
Las Vegas
Tennessee
Albuquerque
Oklahoma
Memphis
Charlotte
North
Arkansas
New Mexico
Arizona
Los Angeles
Atlanta
South Carolina
Phoenix
Dallas
Tijuana
Mexicali
Tucson
Mississippi
Alabama
Georgia
Ciudad Juárez
Texas
Louisiana
Austin
Baja California
Houston
Baton Rouge
Jacksonville
Sonora
San Antonio
Chihuahua
Florida
Nuevo Laredo
Coahuila
Reynosa
Torreón
Monterrey
Baja California Sur
Saltillo
Miami
Tamaulipas
La Paz
Culiacán
Durango

Section 4

Build a Dashboard with Plotly Dash

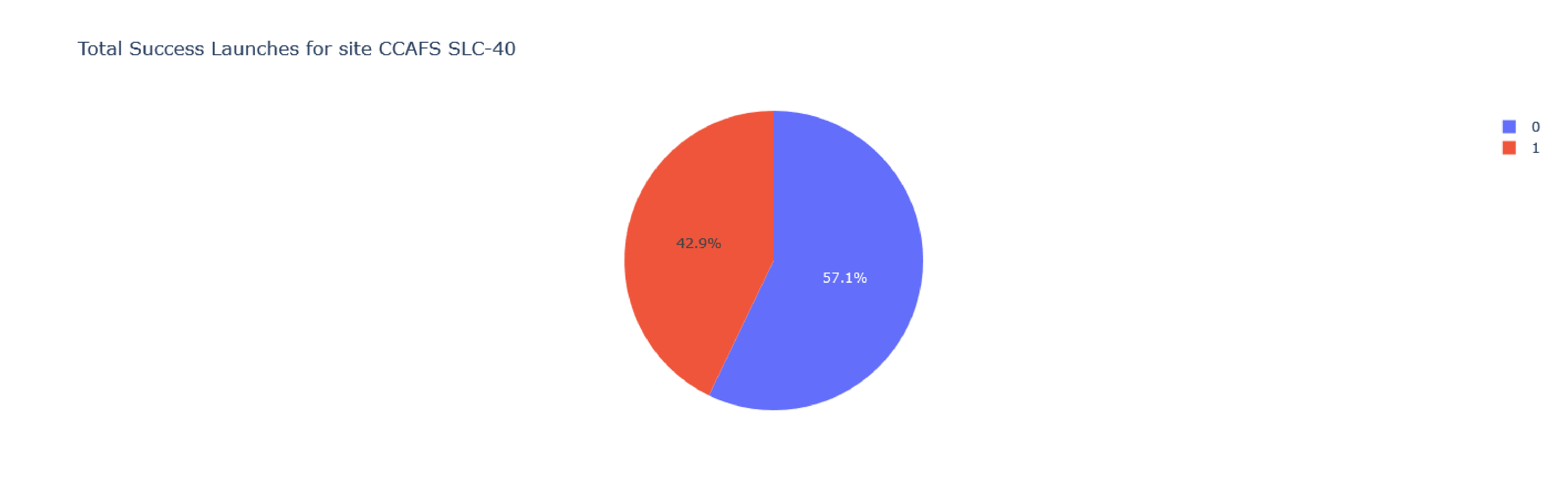
Overall Successful Landing Rate



The above pie chart shows the landing rate of stage 1 rockets for all launch sites.

Roughly 43% of all Falcon 9 rockets landed successfully.

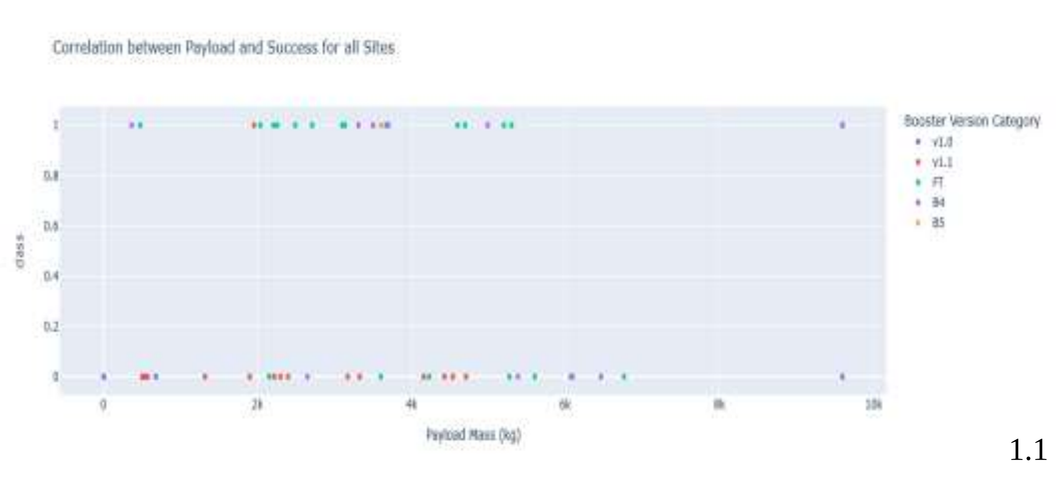
Highest Successful Landings



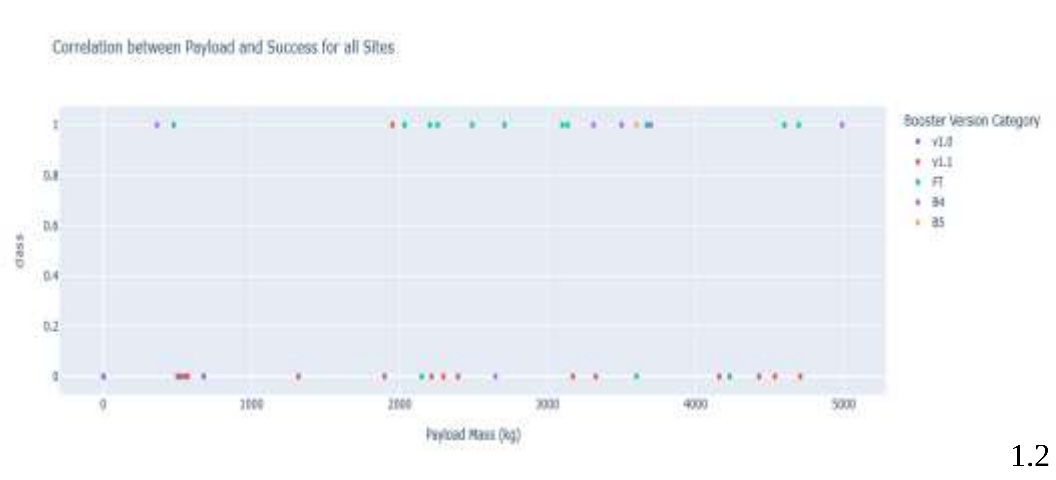
Pie chart for the success rate of all rockets launching from the Cape Canaveral 'Slick 40' SLC-40 launch site.

This site produces the highest rate of successful landings.

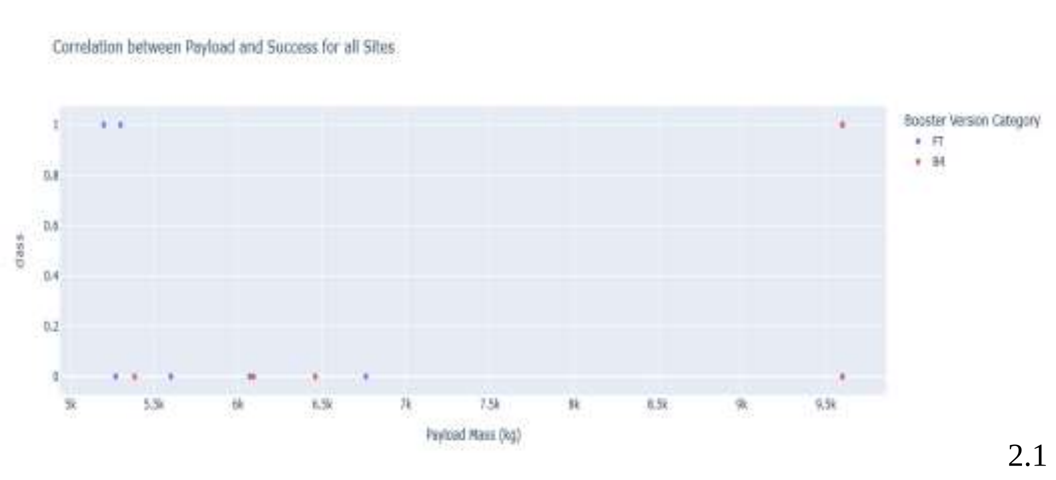
Launch Outcome vs Payload Mass



1.1



1.2



2.1

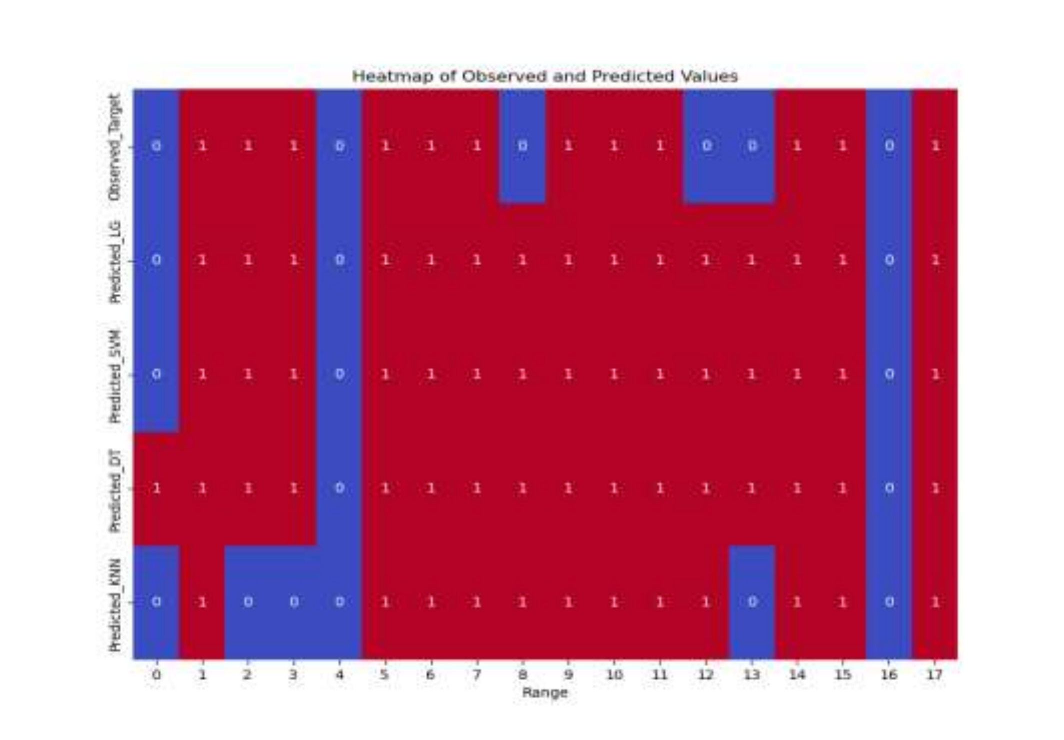
‘Class’ outcomes of launches differentiated by booster version:

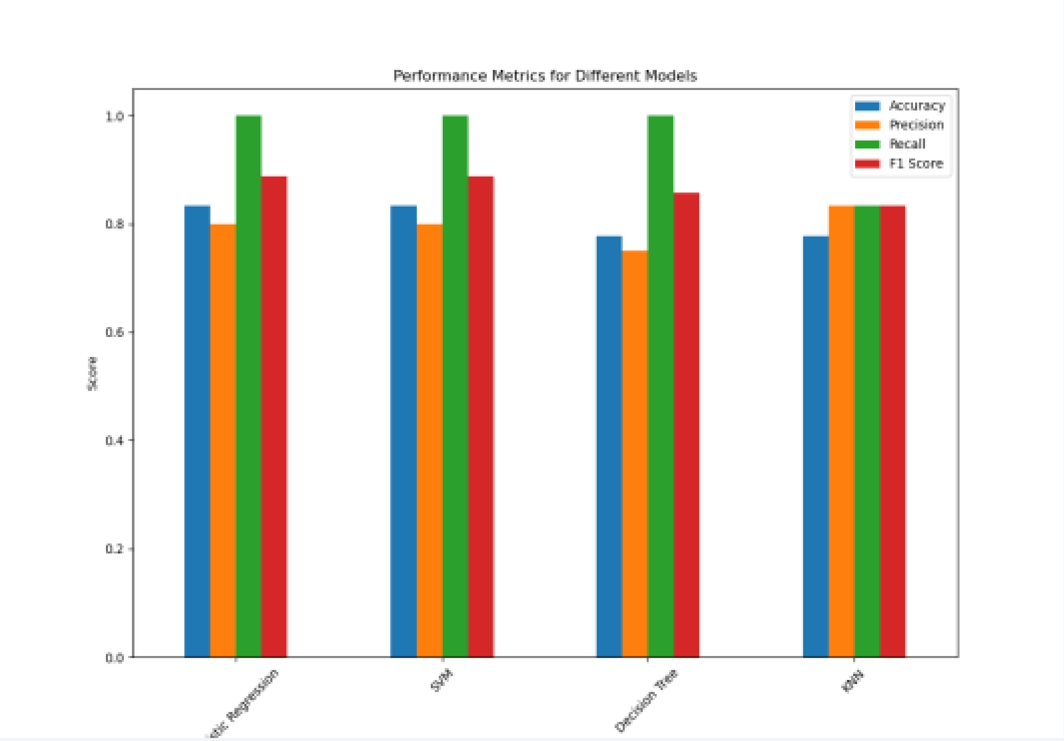
- Plot 1.1: The entire range of payload masses
  - Five unique booster for payloads ranging from 0 to 10000 KG
- Plot 1.2: ‘Payload Mass’ range from 0 to 5000 KG
  - Booster v1.0 used for very light loads
  - Booster v1.1 used for loads between 500 and 5000 KG
- Plot 2.1: ‘Payload Mass’ range from 5000 to 10000 KG
  - Only two boosters used for mid to heavy loads
  - B4 is the only booster used above 7000 KG

Section 5

Predictive Analysis (Classification)

Classification Accuracy





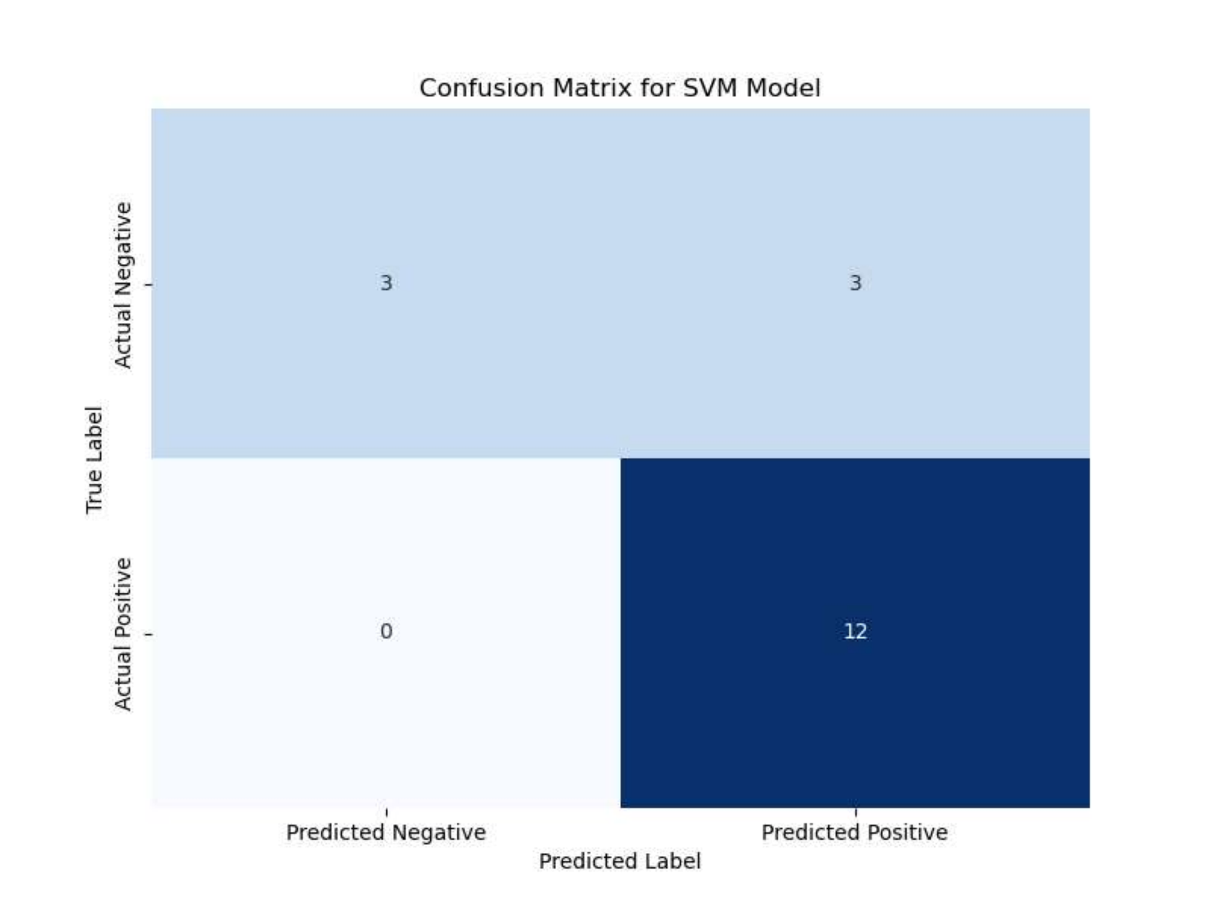
- On the top left is a Heat map showing the observed and predicted values for the set of testing data used to evaluate each model
  - The testing set contained a total of 18 records
- On the top right is a bar chart displaying the metrics Accuracy, Precision, Recall, and F1 Score for each model
  - Logistic Regression, SVM, and Decision Tree produce perfect Recall and has high values for all other metrics

Confusion Matrix

Support Vector Machine Confusion Matrix

Support Vector Classifier chosen as best predictive model

- Best Parameters:
  - 'C': 1.0, 'degree': 1, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid‘
- Classifier Metrics:
  - F-1 Test Score: 0.8889
  - Accuracy Test Score: 0.8333
  - ROC-AUC Score: 0.7500
- Summary
  - Excellent at predicting negative classes when detected
  - Fails to detect half of all negative classes
  - Predicts 80% of all positive classes: 3 False Positives
  - Has perfect recall for positive class: 0 False Negatives
  - Overall the model performs quite well at identifying positive cases but is not so good at detecting negative cases



Conclusions

Successful rocket landings increased as Space-X gained more experience.

Launches occurred most frequently at Cape Canaveral Air Force Station with rockets launched from site SLC “Slick” 40 having the highest rate of successful landings.

Booster v1.1 has the overall lowest success rate

The Decision Tree Classifier always performs the best in training but the worst in testing which suggests that the model is overfitting the training data.

All models produce identical confusion matrices save KNN which has a 78% accuracy and does a better job at detecting and predicting its positive classes.

Results for Logistic Regression were very similar to SVM in both training and testing but was ultimately disqualified for having a lower ROC-AUC score.

Appendix

All support data for this presentation can be found at the following link:

- https://github.com/Dichotmy/IBM-Data-Science-Professional/tree/main/10.%20Applied%20Data%20Science%20Capstone/Labs

Thank you!