IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gaurav Bansal

09-Dec-2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
  - Summary of all results
- Summary of Results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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 SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Problems you want to find answers

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

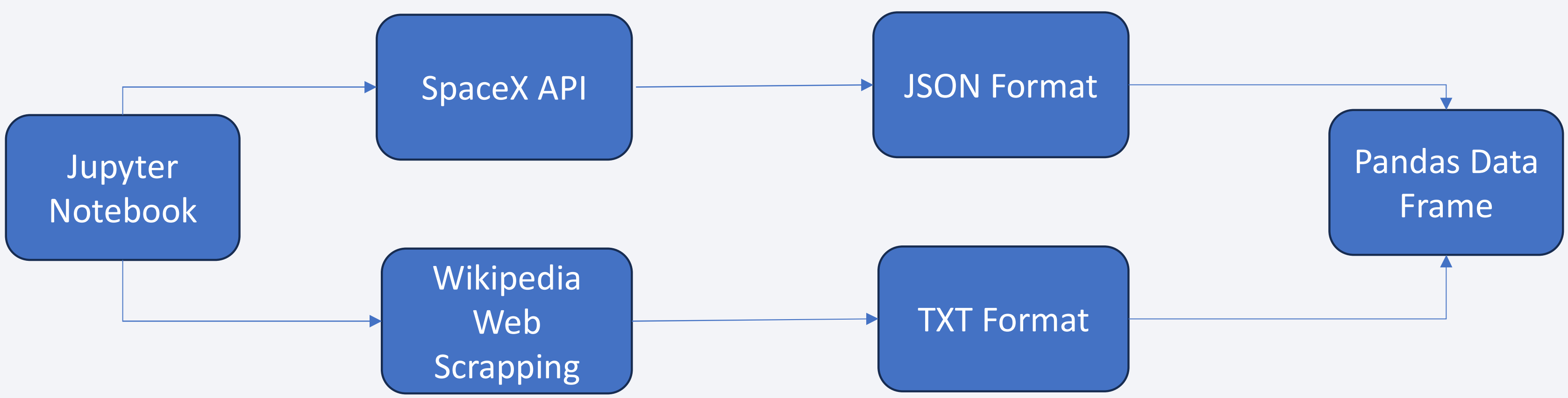
Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardized and transformed data; train/test split data; find best classification algorithm to reach optimal accuracy and ensure best results

Data Collection

- Data collection is the process of gathering data from available sources. This data can be structured, unstructured, or semi-structured. For this project, data was collected via SpaceX API and Web scrapping Wiki pages for relevant launch data
- Data collection process flowcharts



Data Collection – SpaceX API

1.Requesting rocket launch data from SpaceX API

2. Decoding the response content using .json() and turning it into a dataframe using .json_normalize()

3. Requesting needed information about the launches from SpaceX API by applying custom functions

4. Constructing data we have obtained into a dictionary

5. Creating a dataframe from the dictionary

6. Filtering the dataframe to only include Falcon 9 launches

7. Replacing missing values of Payload Mass column with calculated mean() for this column

8. Exporting the data to CSV

GitHub URL of the completed SpaceX API calls notebook:
https://github.com/grb9in/Data-Science/blob/2eeb88360171c3e836a9c66c68606400faa3e00f/Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

- Requesting Falcon 9 launch data from Wikipedia
- Creating a BeautifulSoup object from the HTML response
- Extracting all column names from the HTML table header
- Collecting the data by parsing HTML tables
- Constructing data we have obtained into a dictionary
- Creating a dataframe from the dictionary
- Exporting the data to CSV

GitHub URL of the completed web scraping notebook:
https://github.com/grb9in/Data-Science/blob/c1d18d99a22e06e7f000cb727f32a01e1dcf677a/Data_Collection_with_Web_Scraping.ipynb

Data Wrangling

- In this stage we started by importing pandas and NumPy, loading our collected data in the previous stage to perform our exploratory data analysis which aimed to clean the data and choose the valid features for training a machine learning model.
- data wrangling process flowcharts

1- Loading the collected dataset

2- Identifying and calculating the percentage of the missing values in each attribute

3- Identifying which columns are numerical and categorical

4- Calculating the number of launches on each site

5- Calculating the number and occurrence of each orbit

6- Creating a landing outcome label from Outcome column

7- determining the success rate of returning the first stage of the rocket

GitHub URL of data wrangling related notebook: https://github.com/grb9in/Data-Science/blob/475208cec6de2e991a9b7f0047455e8da5cbfbcc/DataWrangling.ipynb

EDA with Data Visualization

1. Scatter plot: Shows relationship or correlation between two variables making patterns easy to observe. Plotted following charts to visualize:

- Relationship between Flight Number and Launch Site
- Relationship between Payload and Launch Site
- Relationship between Flight Number and Orbit Type
- Relationship between Payload and Orbit Type

2. Bar Chart: Commonly used to compare the values of a variable at a given point in time. Plotted following Bar chart to visualize:

- Relationship between success rate of each orbit type

3. Line Chart: Commonly used to track changes over a period of time. It helps depict trends over time. Plotted following Line chart to observe:

- Average launch success yearly trend

GitHub URL of EDA with data visualization notebook: https://github.com/grb9in/Data-Science/blob/c38295db0f0bc1625ebe3110c408522900814a2c/EDA_VizLab.ipynb

EDA with SQL

- Performed SQL queries:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking 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 EDA with SQL notebook: https://github.com/grb9in/Data-Science/blob/05576877bf6f28586546adac4fdc4e2b33962d8b/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Markers of all Launch Sites: -
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location. - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Colored Markers of the launch outcomes for each Launch Site: -
  - Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities: -
  - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

GitHub URL of interactive map with Folium map: https://github.com/grb9in/Data-Science/blob/7a2c22fb509e6710fb67472446359e7c68b77d11/Interactive_Viz_Folium.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success

GitHub URL of Plotly Dash lab: https://github.com/grb9in/Data-Science/blob/8d1db72b8803b6db2ca7aa4ba259d39b77991fee/Plotly_Dashboard.py

Predictive Analysis (Classification)

- Creating a NumPy array from the column “Class” in data
- Standardizing the data with StandardScaler, then fitting and transforming it
- Splitting the data into training and testing sets with train_test_split function
- Creating a GridSearchCV object with cv = 10 to find the best parameters
- Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
- Calculating the accuracy on the test data using the method .score() for all models
- Examining the confusion matrix for all models
- Finding the method performs best by examining the Jaccard_score and F1_score metrics

Model development process flowchart:

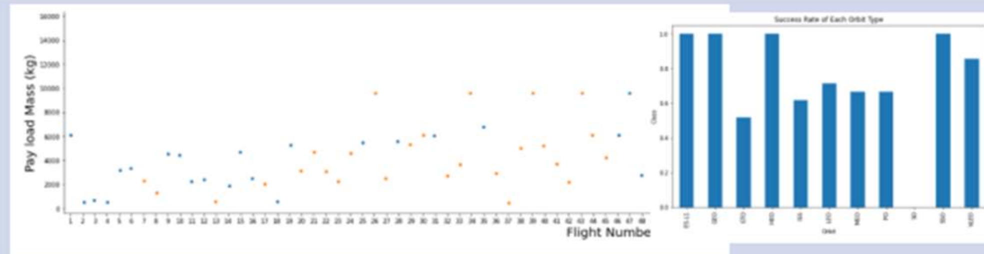
1. Read dataset into Data frame and create a ‘Class’ array
2. Standardize the data
3. Train/Test/Split data in to training and test data sets
4. Create and Refine Models
5. Find the best performing Model

GitHub URL of predictive analysis lab: https://github.com/grb9in/Data-Science/blob/5372a607eff2ba9e32412184be69ba06a62dcc82/PredictiveAnalysis.ipynb

Results

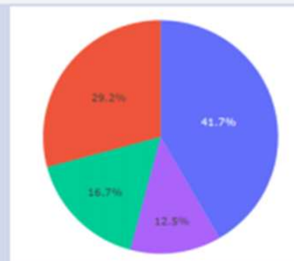
Exploratory data analysis results

- Samples:



Interactive analytics demo in screenshots

- Samples



Predictive analysis results

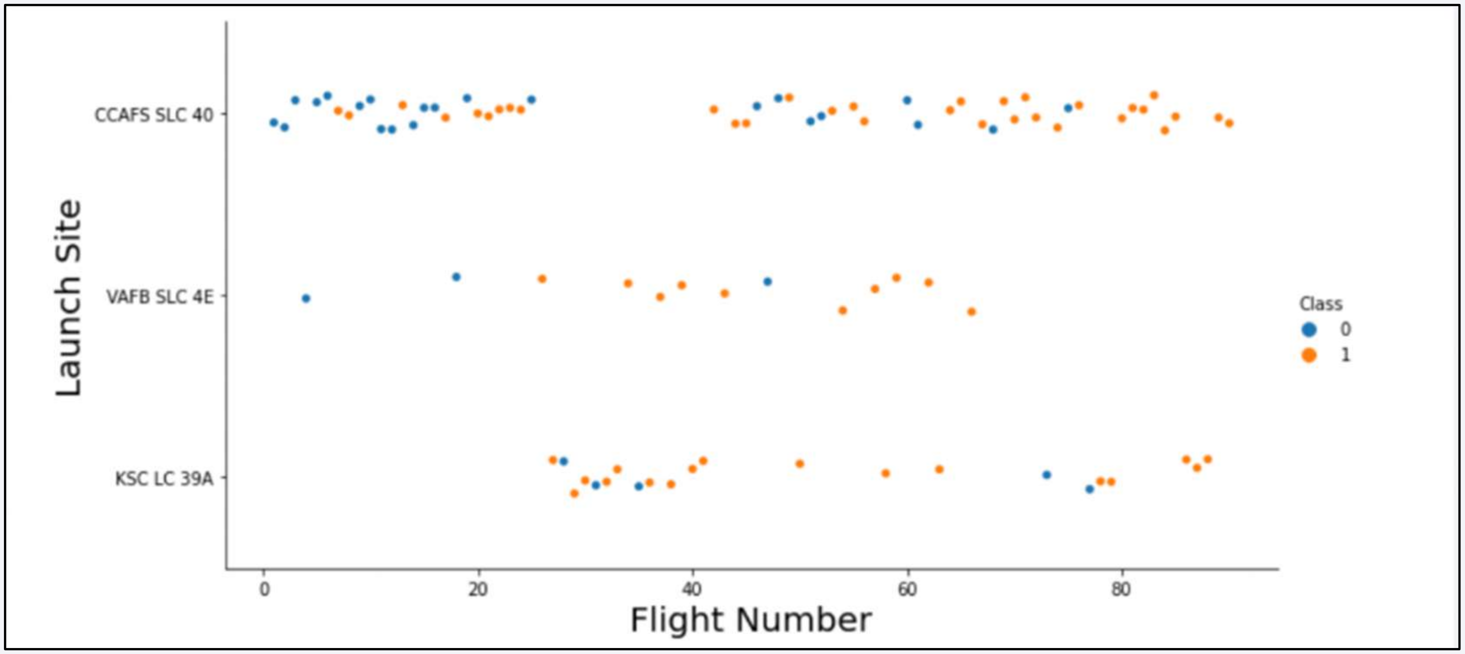
- Samples

	Algo Type	Accuracy Score
2	Decision Tree	0.903571
3	KNN	0.848214
1	SVM	0.848214
0	Logistic Regression	0.846429

Section 2

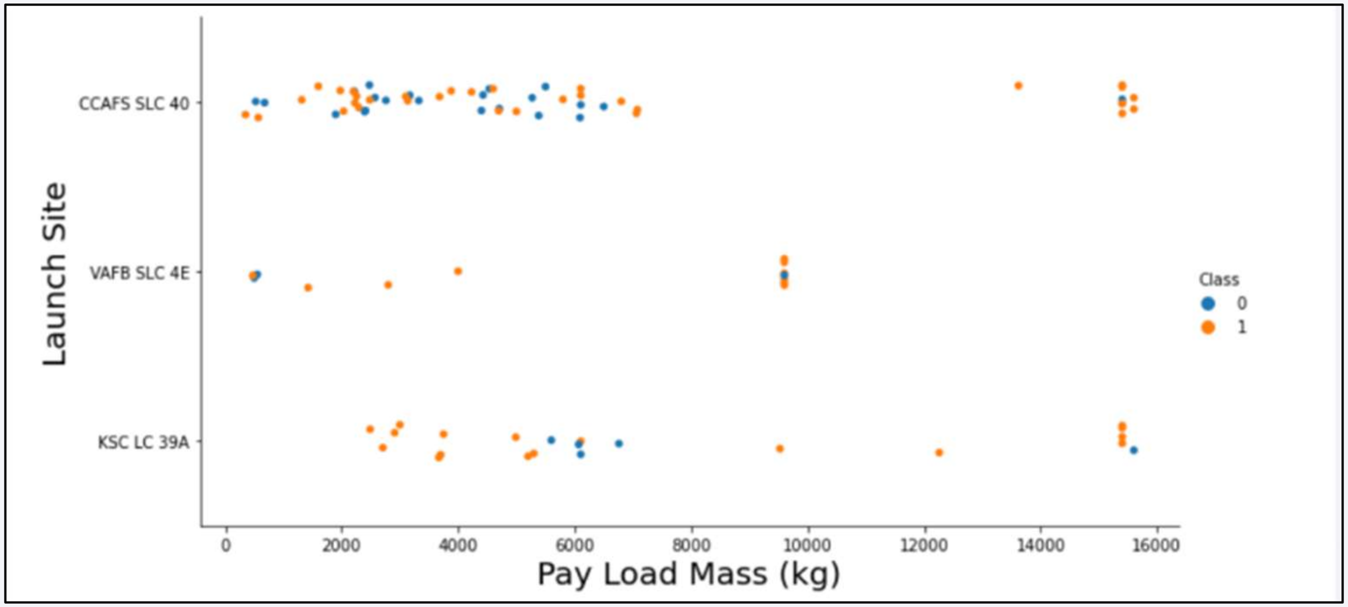
Insights drawn from EDA

Flight Number vs. Launch Site



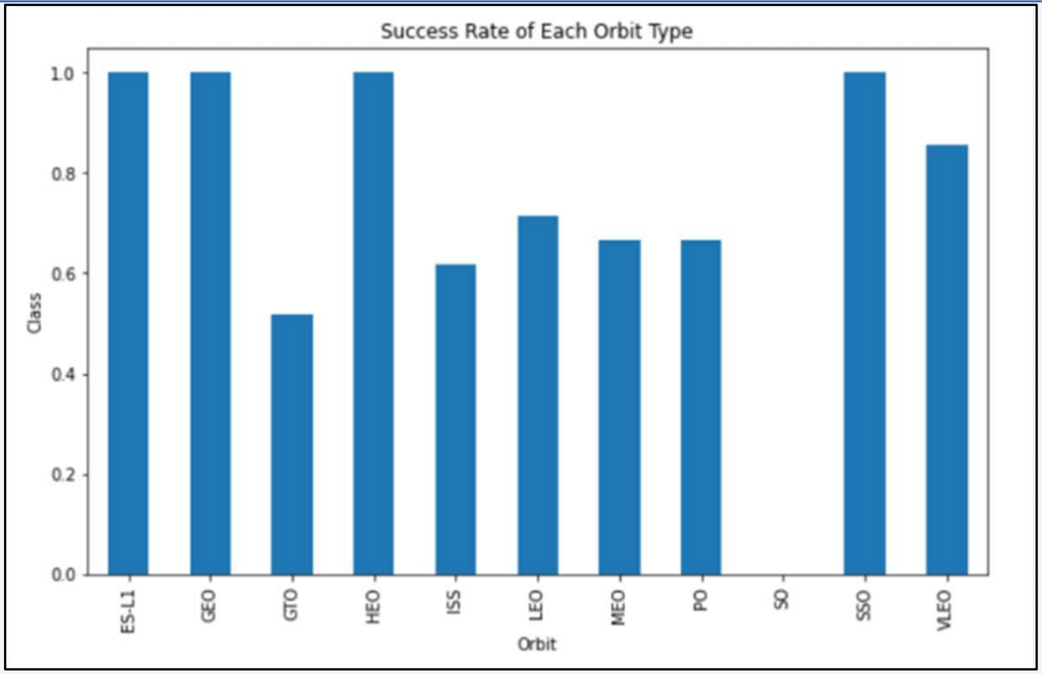
- Success rates (Class=1) increases as the number of flights increase
- For launch site ‘KSC LC 39A’, it takes at least around 25 launches before a first successful launch

Payload vs. Launch Site



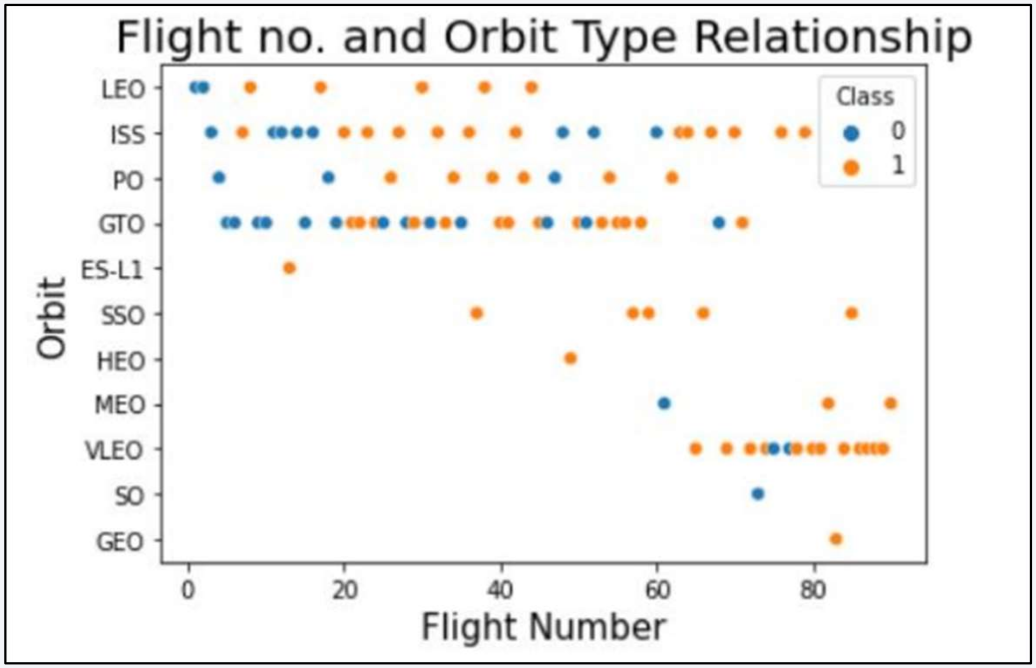
- For launch site ‘VAFB SLC 4E’, there are no rockets launched for payload greater than 10,000 kg
- Percentage of successful launch (Class=1) increases for launch site ‘VAFB SLC 4E’ as the payload mass increases
- There is no clear correlation or pattern between launch site and payload mass

Success Rate vs. Orbit Type



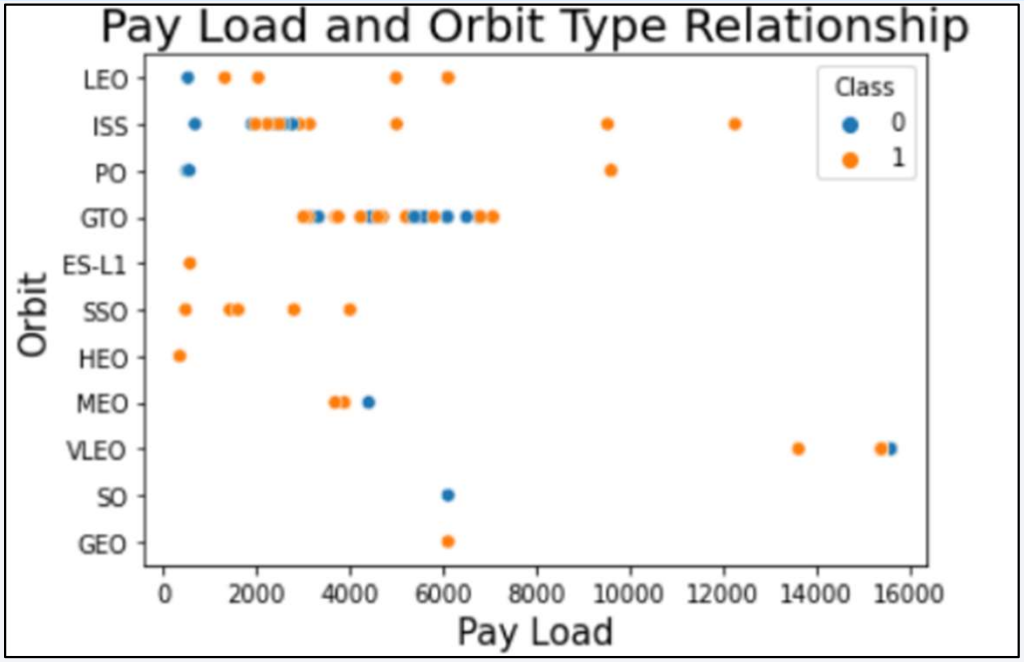
- Orbits ES-LI, GEO, HEO, and SSO have the highest success rates
- GTO orbit has the lowest success rate

Flight Number vs. Orbit Type



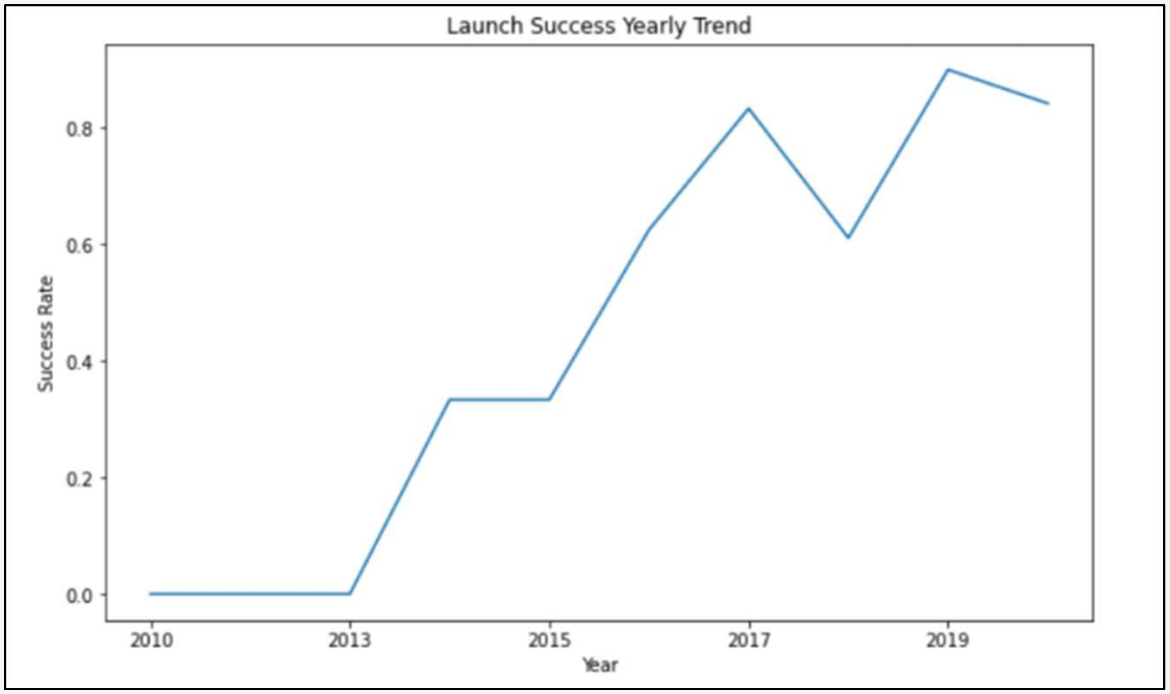
- For orbit VLEO, first successful landing (class=1) doesn't occur until 60+ number of flights
- For most orbits (LEO, ISS, PO, SSO, MEO, VLEO) successful landing rates appear to increase with flight numbers
- There is no relationship between flight number and orbit for GTO

Payload vs. Orbit Type



- Successful landing rates (Class=1) appear to increase with pay load for orbits LEO, ISS, PO, and SSO
- For GEO orbit, there is not clear pattern between payload and orbit for successful or unsuccessful landing

Launch Success Yearly Trend



- Success rate (Class=1) increased by about 80% between 2013 and 2020
- Success rates remained the same between 2010 and 2013 and between 2014 and 2015
- Success rates decreased between 2017 and 2018 and between 2019 and 2020

All Launch Site Names

- Query:
  - select distinct Launch_Site from spacextbl
- Explanations and Results:
  - ‘distinct’ returns only unique values from the queries column (Launch_Site)
  - There are 4 unique launch sites

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

Launch Site Names Begin with 'CCA'

- Query:
  - select * from spacextbl where Launch_Site LIKE 'CCA%' limit 5;
- Explanations and Results:
  - Using keyword ‘Like’ and format ‘CCA%’, returns records where ‘Launch_Site’ column starts with “CCA”.
  - Limit 5, limits the number of returned records to 5

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	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_) from spacextbl where Customer = 'NASA (CRS)’
- Explanations and Results:
  - ‘sum’ adds column ‘PAYLOAD_MASS_KG’ and returns total payload mass for customers named ‘NASA (CRS)’

45596

Average Payload Mass by F9 v1.1

- Query:
  - select avg(PAYLOAD_MASS__KG_) from spacextbl where Booster_Version LIKE 'F9 v1.1';
- Explanations and Results:
  - ‘avg’ keyword returns the average of payload mass in ‘PAYLOAD_MASS_KG’ column where booster version is ‘F9 v1.1’

2928

First Successful Ground Landing Date

- Query:
  - select min(Date) as min_date from spacextbl where Landing__Outcome = 'Success (ground pad)';
- Explanations and Results:
  - 'min(Date)' selects the first or the oldest date from the 'Date' column where first successful landing on group pad was achieved
  - Where clause defines the criteria to return date for scenarios where 'Landing_Outcome' value is equal to 'Success (ground pad)'

min_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:
  - select Booster_Version from spacextbl where (PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_ < 6000) and (Landing__Outcome = 'Success (drone ship)');
- Explanations and Results:
  - The query finds the booster version where payload mass is greater than 4000 but less than 6000 and the landing outcome is success in drone ship
  - The ‘and’ operator in the where clause returns booster versions where both conditions in the where clause are true

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(Mission_Outcome) as counts from spacextbl group by Mission_Outcome;
- Explanations and Results:
  - The ‘group by’ keyword arranges identical data in a column in to group
  - In this case, number of mission outcomes by types of outcomes are grouped in column ‘counts’

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

Boosters Carried Maximum Payload

- Query:
  - select Booster_Version, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl);
- Explanations and Results:
  - The sub query returns the maximum payload mass by using keywork 'max' on the pay load mass column
  - The main query returns booster versions and respective payload mass where payload mass is maximum with value of 15600

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Query:
  - select Landing__Outcome, Booster_Version, Launch_Site from spacextbl where Landing__Outcome = 'Failure (drone ship)' and year(Date) = '2015';
- Explanations and Results:
  - The query lists landing outcome, booster version, and the launch site where landing outcome is failed in drone ship and the year is 2015
  - The ‘and’ operator in the where clause returns booster versions where both conditions in the where clause are true. Also, ‘year’ keywork extracts the year from column ‘Date
  - The results identify launch site as ‘CCAFS LC-40’ and booster version as F9 v1.1 B1012 and B1015 that had failed landing outcomes in drop ship in the year 2015

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

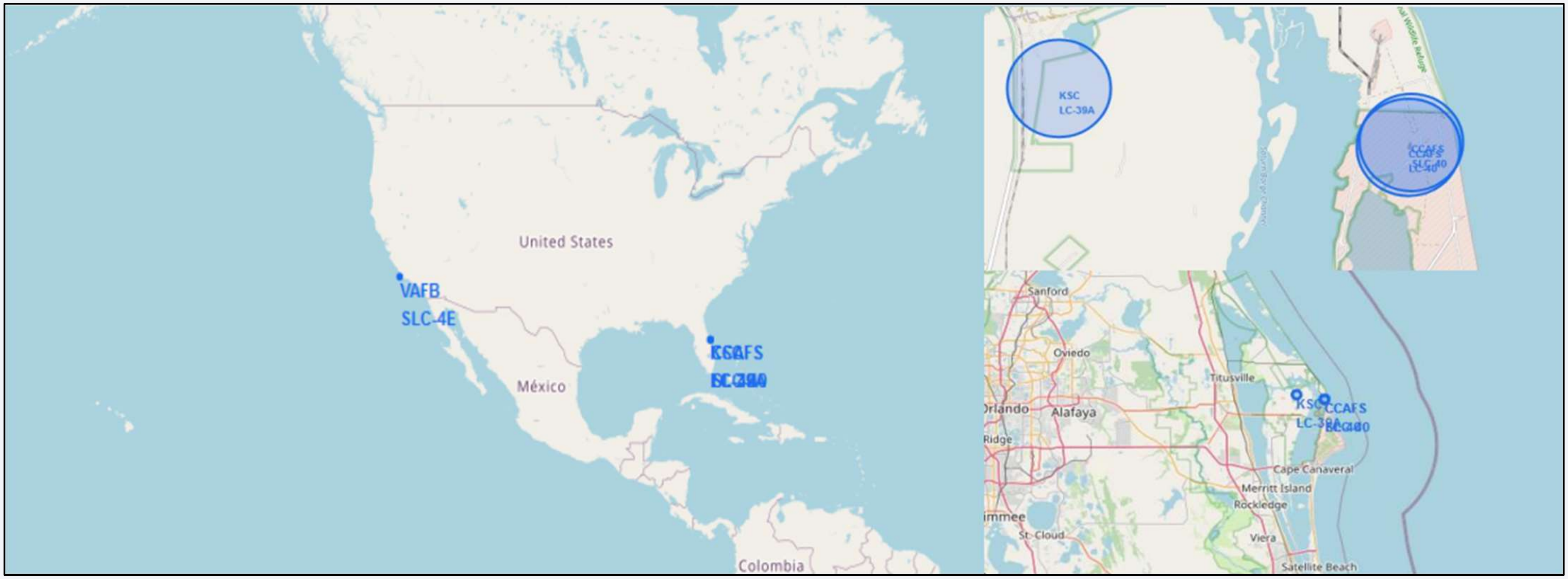
- Query:
  - select Landing__Outcome, count(*) as LandingCounts from spacextbl where Date between '2010-06-04' and '2017-03-20' group by Landing__Outcome order by count(*) desc;
- Explanations and Results:
  - The 'group by' key word arranges data in column 'Landing__Outcome' into groups
  - The 'between' and 'and' keywords return data that is between 2010-06-04 and 2017-03-20
  - The 'order by' keyword arranges the counts column in descending order
  - The result of the query is a ranked list of landing outcome counts per the specified date range

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

Section 3

Launch Sites Proximities Analysis

Folium Map: Launch Sites



All site locations are near the coast and Equator line, SpaceX focuses on locations that are close to water and the zeroth latitude for the purpose of avoiding any undesired accidents. The launch sites are distributed in two states California and Florida

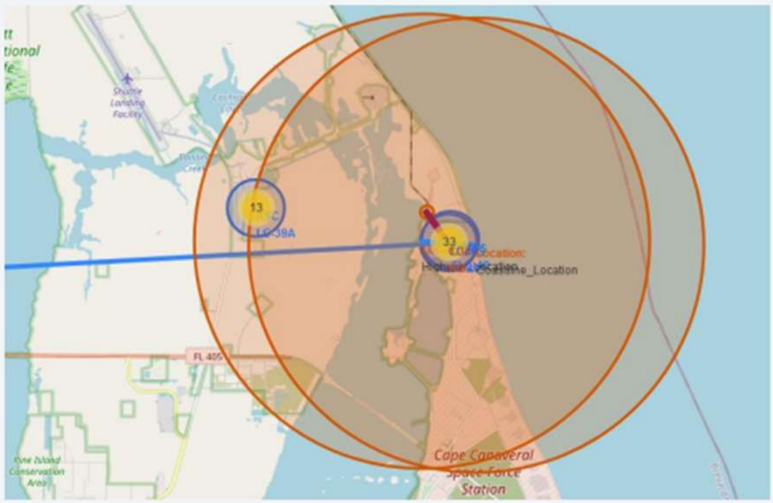
Folium Map: Success rate for each launch location

From the color-labeled markers in marker clusters, we can easily identify which launch sites have relatively high success rates.

Green Marker = Successful Return

Red Marker = Failed Return

Folium Map: Closest Proximities to CCAFS LC-40



Proximities Cordinates

	Location	Lat	Long
0	Orlando_Location	28.52300	-81.38260
1	Coastline_Location	28.56146	-80.56746
2	Highway_Location	28.56270	-80.58703

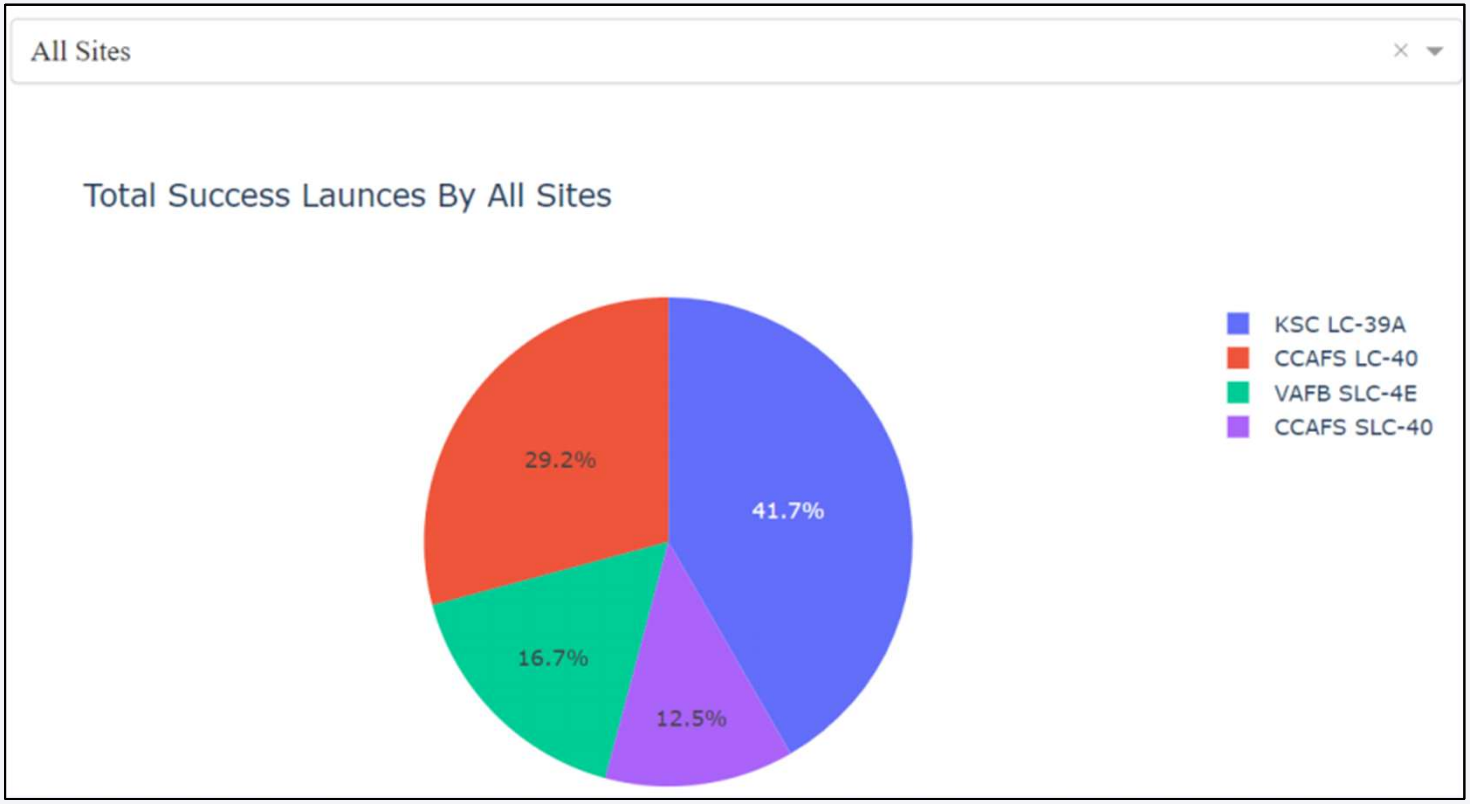
we calculated the distances between the launch site (CCAFS LC-40) to its proximities

Orlando City Distance ≈ 78.8 Km,
Coastline Distance ≈ 0.97 Km,
Highway Distance ≈ 0.95Km

Section 4

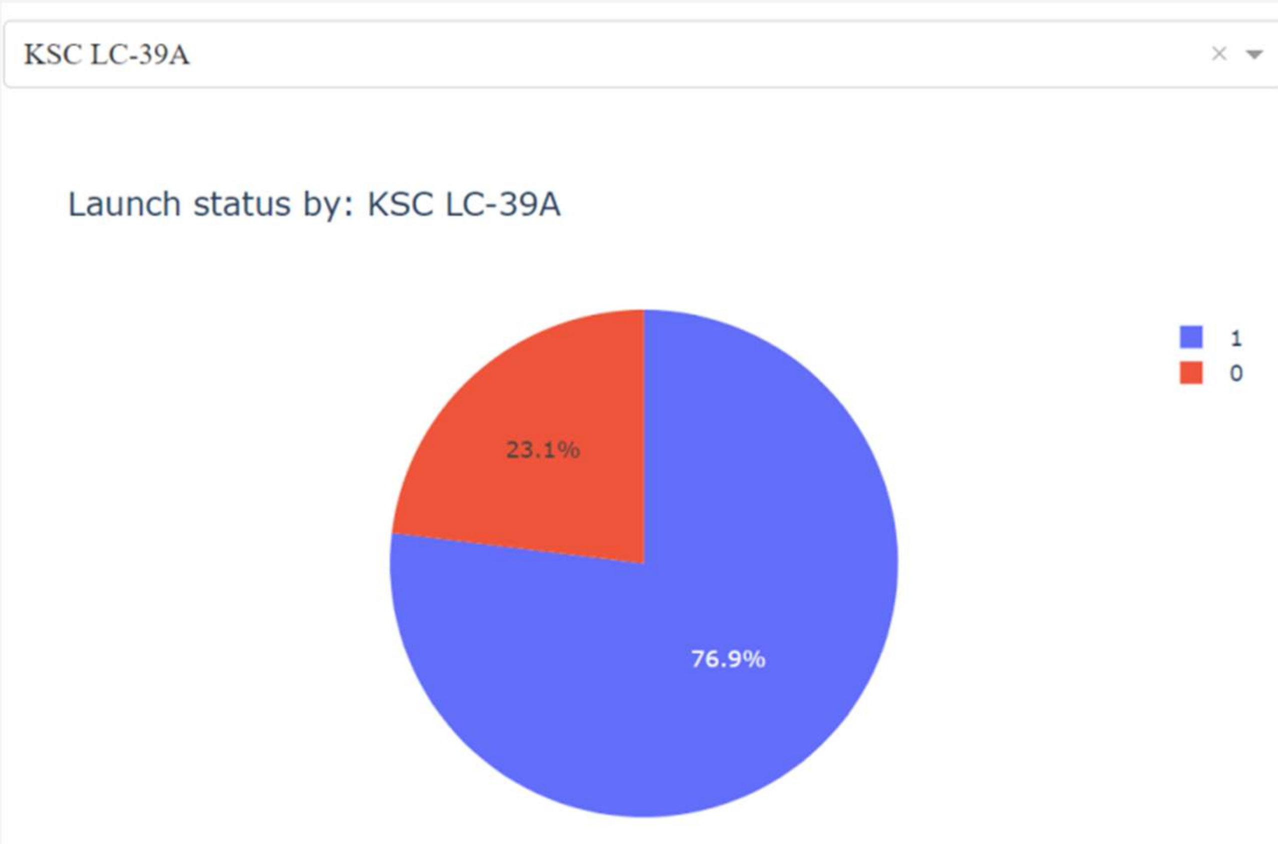
Build a Dashboard with Plotly Dash

Dashboard: Launch success count for all sites



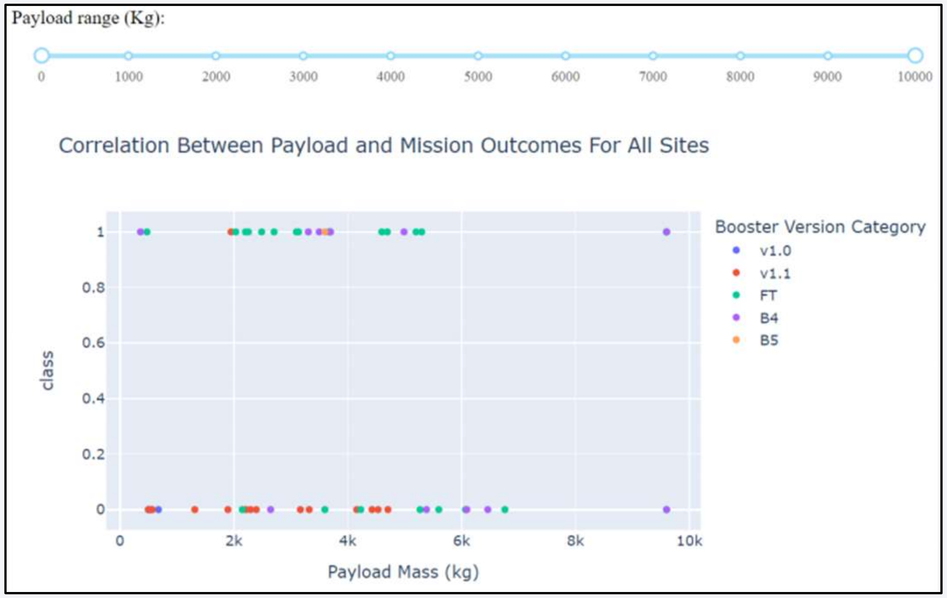
- Launch Site ‘KSC LC-39A’ has the highest launch success rate of with 41.7%
- Launch Site ‘CCAFS SLC-40’ has the lowest launch success rate of only12.5%

Dashboard: Launch success for KSC LC 39A



- KSC LC-39A Launch Site has the highest launch success rate with 10 successful and only 3 failed landings.
- Launch success rate is 76.9%
- Launch success failure rate is 23.1%

Dashboard: Payload vs. Launch Outcome Plot for All Sites

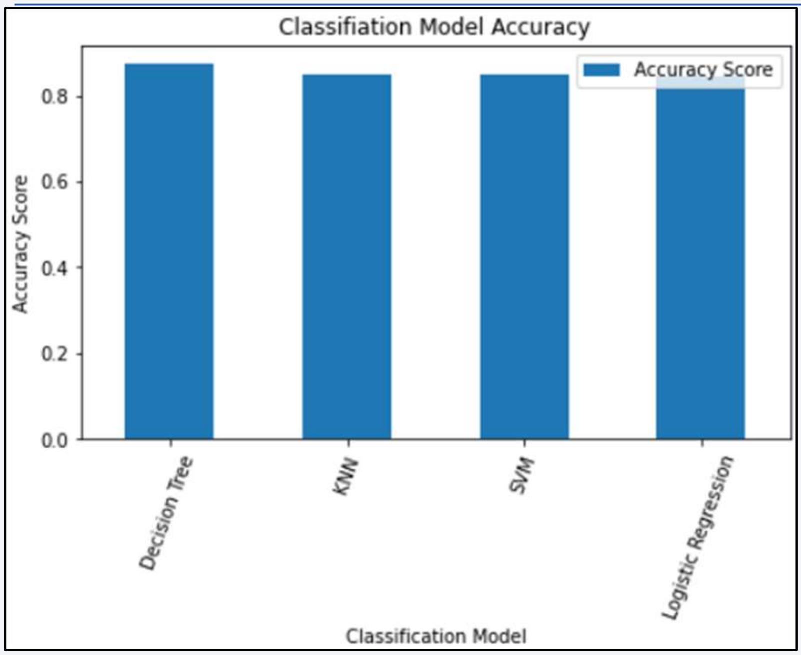


- Most successful launches are in the payload range from 2000 to about 5500
- Booster version category ‘FT’ has the most successful launches
- Only booster with a success launch when payload is greater than 6k is ‘B4’

Section 5

Predictive Analysis (Classification)

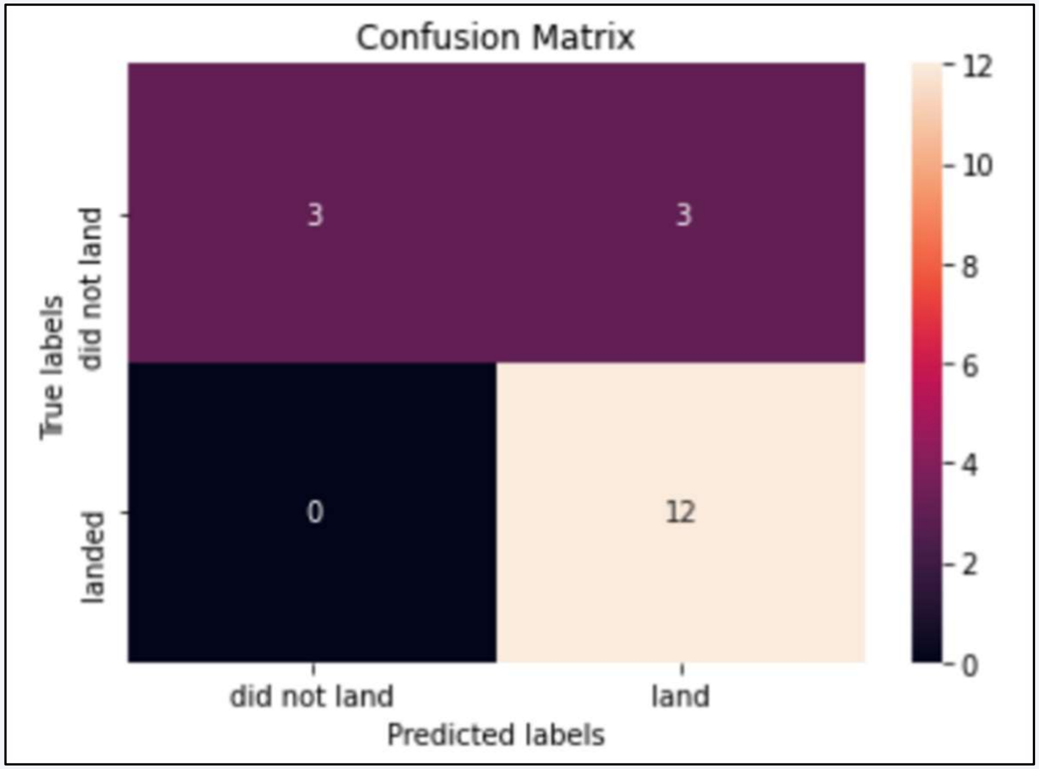
Classification Accuracy



	Algo Type	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.875000	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

- Based on the Accuracy scores and as also evident from the bar chart, Decision Tree algorithm has the highest classification score with a value of .8750
- Accuracy Score on the test data is the same for all the classification algorithms based on the data set with a value of .8333
- Given that the Accuracy scores for Classification algorithms are very close and the test scores are the same, we may need a broader data set to further tune the models

Confusion Matrix



- The confusion matrix is same for all the models (LR, SVM, Decision Tree, KNN)
- Per the confusion matrix, the classifier made 18 predictions
- 12 scenarios were predicted Yes for landing, and they did land successfully (True positive)
- 3 scenarios (top left) were predicted No for landing, and they did not land (True negative)
- 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive)
- Overall, the classifier is correct about 83% of the time ((TP + TN) / Total) with a misclassification or error rate ((FP + FN) / Total) of about 16.5%

Conclusions

- For the given dataset, best performing Machine Learning Classification Model is the Decision Tree with an accuracy of about 87.5%. When the models were scored on the test data, the accuracy score was about 83% for all models. More data may be needed to further tune the models and find a potential better fit.
- As the numbers of flights increase, the first stage is more likely to land successfully
- Success rates appear go up as Payload increases but there is no clear correlation between Payload mass and success rates
- Launch success rate increased by about 80% from 2013 to 2020
- Launch Site 'KSC LC-39A' has the highest launch success rate and Launch Site 'CCAFS SLC-40' has the lowest launch success rate
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest
- Lunch sites are located strategically away from the cities and closer to coastline, railroads, and highways

Appendix

- Special Thanks to: Instructors: IBM Data Science Professional Certificate | Coursera

- My GitHub Link that includes work done as part of this course learnings: grb9in/Data-Science


Thank you!