IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matthew Dillon
November 3, 2024

Outline

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

Executive Summary

- Methodologies:
  - Collecting Data via API and Web Scraping
  - Exploratory Data Analysis via Python and SQL
  - Data Visualization via Folium and Plotly
  - Predictive Analytics via Machine Learning Models
    - Logistic regression, Support vector machine (SVM), Decision Tree, k-NN
- Summary of all results
  - Various insights gleaned from data exploration and visualization
  - Multiple launch success predictive models with accuracy in excess of 80%

Introduction

- Project Background:
  - SpaceX is a private aerospace manufacturer and space transportation company founded by Elon Musk in 2002 with the goal of reducing space transportation costs to eventually enable the colonization of Mars. Known for achievements like the Falcon 9 reusable rocket and the Starship system, SpaceX has become a leader in launching satellites, crewed missions, and cargo for NASA and private clients.
  - The goal of this project is to utilize data analytics, data science, and machine learning methodologies to predict whether a SpaceX rocket booster landing will be successful after launch.
- Problems you want to find answers
  - What factors determine if a rocket booster will land successfully?
  - What features are closely correlated with booster landing success rate?
  - Can we predict if a rocket launch will be successful given a feature set?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX data collected for this project is public and was obtained in two different ways:

    1) SpaceX API (https://api.spacexdata.com/v4/rockets/)

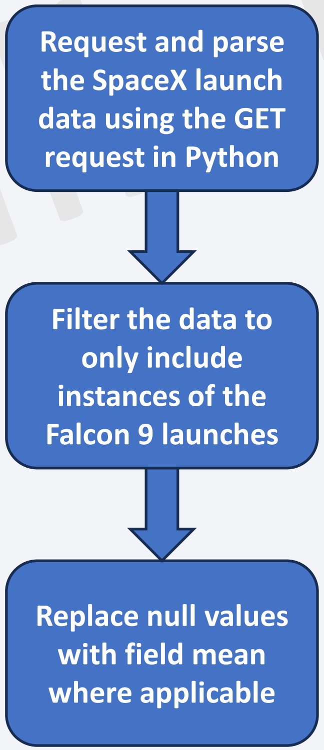
    2) Web Scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Once data was ingested, a binary outcome class feature (dependent variable) was extracted, null values were replaced with the corresponding mean value where applicable, and one-hot-encoding was utilize to convert categorical variables to binary variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained and tested four different machine learning models/methodologies using Grid Search and compared the accuracy score of each to determine the most accurate model

Data Collection

- How was the Data Collected?
  - Data pertaining to individual SpaceX rocket launches was collected from the SpaceX API using the Requests library in Python
  - Features extracted includes: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Latitude, Longitude
  - Additional data was extracted from Wikipedia via Web Scraping with Beautiful Soup in Python
  - Features extracted includes: Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time

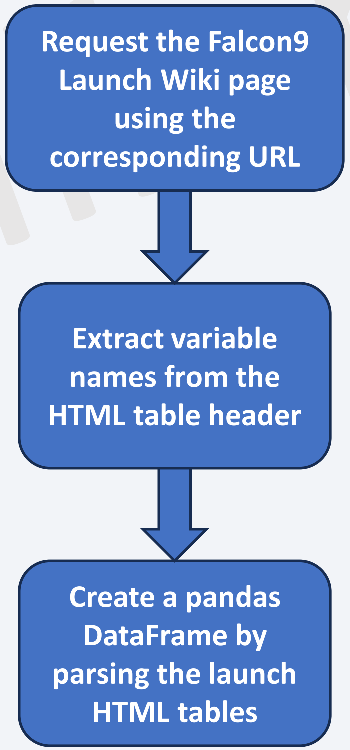
Data Collection – SpaceX API

- SpaceX offers a public API that users can leverage to extract data related to SpaceX rocket launches over time
- The SpaceX API was accessed using the methodology outlined in the flowchart shown to the right
- GitHub URL: https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-1-jupyter-labs-spacex-data-collection-api-v2.ipynb



Data Collection - Scraping

- Wikipedia has a plethora of articles that outline the history and data pertaining to SpaceX's Falcon9 rocket launches
- The tabular data on these Wikipedia pages can be parsed and extracted via Web Scraping with Python's Beautiful Soup library
- The process of extracting and formatting this data is outlined by the flow chart to the right
- GitHub URL: https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-2-jupyter-labs-webscraping.ipynb



Data Wrangling

- Once we collected data from the SpaceX API and the Falcon9 Wikipedia page, we must manipulate it and transform it in a way that allows it to be more easily ingestible and analyzed
- Data exploration, wrangling, and transformation allows us to understand our dataset and its intricacies using Pandas in Python
- The data wrangling steps taken in pursuant to this methodology is outlined in the flow chart to the right
- GitHub URL: https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-3-labs-jupyter-spacex-Data%20wrangling-v2.ipynb

Calculate the number of launches by each site

↓

Calculate the number of occurrences of each orbit type

↓

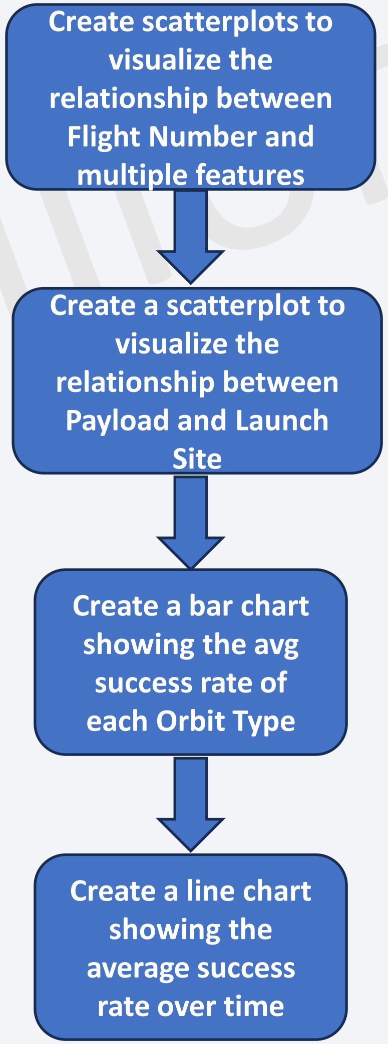
Calculate the number of occurrences of mission outcomes

↓

Create a binary "Landing Outcome" column (dependent variable)

EDA with Data Visualization

- Once the dataset was transformed, we continue to improve our understanding of the data via visualization in Python
- Both the Matplotlib and Seaborn libraries were used for to build these visualizations
- The data visualizations built pursuant to the exploratory data analysis at this stage is outlined in the flow chart to the right
- GitHub URL: https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-5-jupyter-labs-eda-dataviz-v2.ipynb



EDA with SQL

- The following SQL queries were written to extract additional insights:
  - Extract the distinct Launch Site names
  - Extract 5 records where the Launch Site name begins with “CCA”
  - Calculate the total Payload Mass (kg) for the “NASA (CRS)” Customer
  - Calculate the average Payload Mass (kg) of the “F9 v.1.1” Booster Version
  - Extract the date of the first successful ground pad landing
  - List the distinct Booster names with a successful drone ship Landing Outcome with a Payload Mass between 4000kg – 6000kg
  - List the total number of successful and failure mission outcomes
  - List the Booster Versions that have carried the maximum payload, using a subquery
  - List records from 2015 with a Landing Outcome of failure (drone ship)
  - Rank the count of landing outcomes in descending order between 6/4/2010 and 3/20/2017, inclusive

- GitHub URL: https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Created Folium map Circles to visualize the coordinates of all launch sites
- Utilized Folium map Markers to associate the Site Name with each location
- Added Folium map Icons to represent successful and failed launches by color
- Leveraged Folium lines to calculate distance between various points on the map

- GitHub URL: https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-6-lab-jupyter-launch-site-location-v2.ipynb

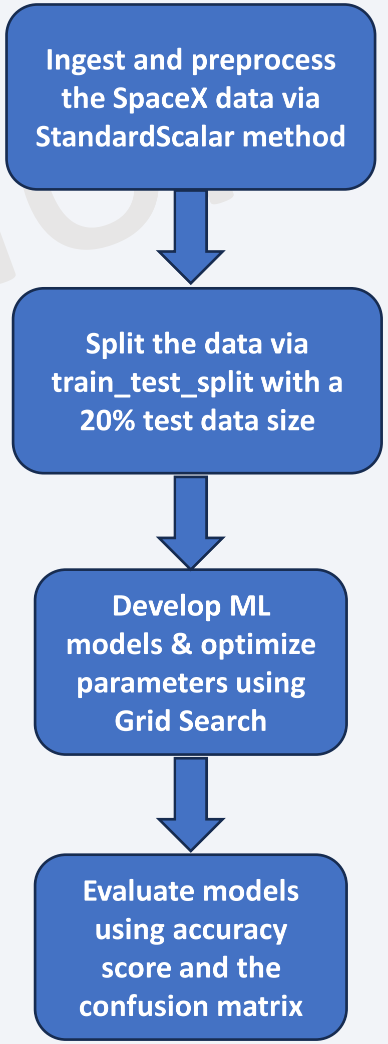
Build a Dashboard with Plotly Dash

- Built an interactive web-hosted dashboard with Python's Plotly Dash library to allow the user to slice the data and glean a multitude of insights
- Added a dropdown selection to filter the data by Launch Site
- Added a pie chart to represent the proportion of successful launches for the subset of data selected
- Added a scatter plot to visualize the correlation between Payload Mass (kg) and launch success, colored by Booster Version for added insights

- GitHub URL: https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-7-Build_a_Dashboard_Application_with_Plotly_Dash.ipynb

Predictive Analysis (Classification)

- Leveraged Python's sklearn library to train, test, ad evaluate four different predictive machine learning classification models to predict the likelihood of launch success given a subset of feature values
- Machine Learning Models developed include Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-NN
- Utilized Grid Search to determine the optimal parameters for each of the aforementioned models resulting in the highest training accuracy
- Evaluated the performance of each model by using the accuracy score as well as the Confusion Matrix results
- GitHub URL: https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-8-SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb



Results

- Exploratory data analysis results
  - Rockets with a higher Payload Mass (kg) seem to have a higher likelihood of success
  - ES-L1, GEO, HEO, and SSO Orbits tend to have high success rates, whereas GTO Orbits have the lowest success rate
  - Launch success rate has generally increased over time, with a significant downtick in 2018 due to multiple launch failures
- Predictive analysis results
  - Logistic Regression out-of-sample accuracy: 83%
  - SVM out-of-sample accuracy: 83%
  - Decision Tree out-of-sample accuracy: 89%
  - kNN out-of-sample accuracy: 83%

```
print('Accuracy for the Logistics Regression model:',    logreg_cv.score(X_test, Y_test))
print( 'Accuracy for the Support Vector Machine model:', svm_cv.score(X_test, Y_test))
print('Accuracy for the Decision tree model:',           tree_cv.score(X_test, Y_test))
print('Accuracy for the kNN model:',                     knn_cv.score(X_test, Y_test))
```

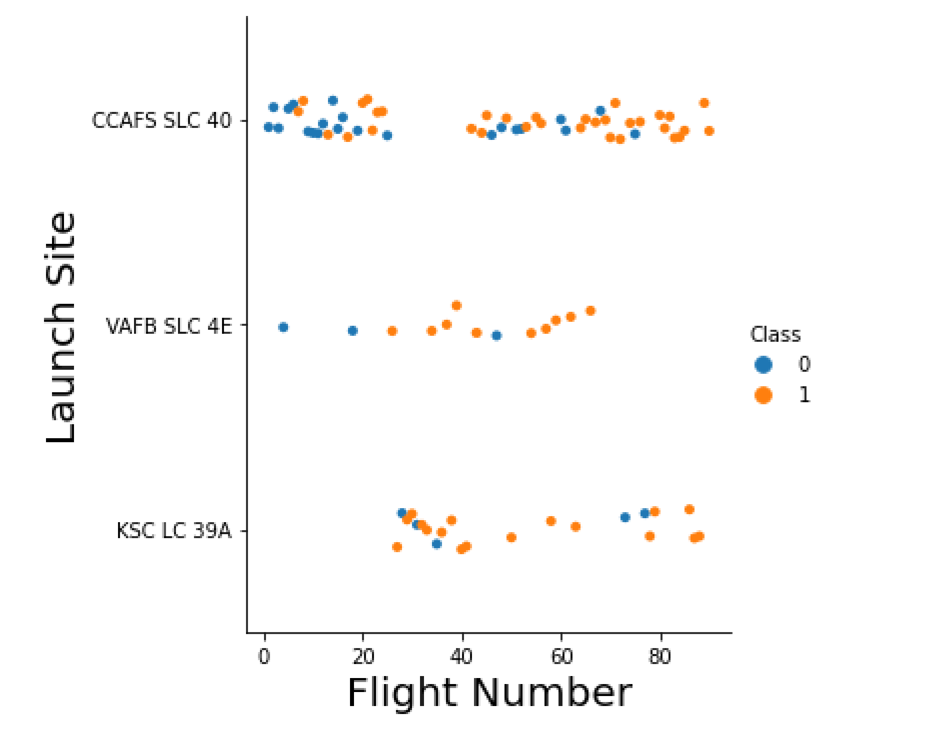
```
Accuracy for the Logistics Regression model: 0.8333333333333334
Accuracy for the Support Vector Machine model: 0.8333333333333334
Accuracy for the Decision tree model: 0.8888888888888888
Accuracy for the kNN model: 0.8333333333333334
```

Section 2

Insights drawn from EDA

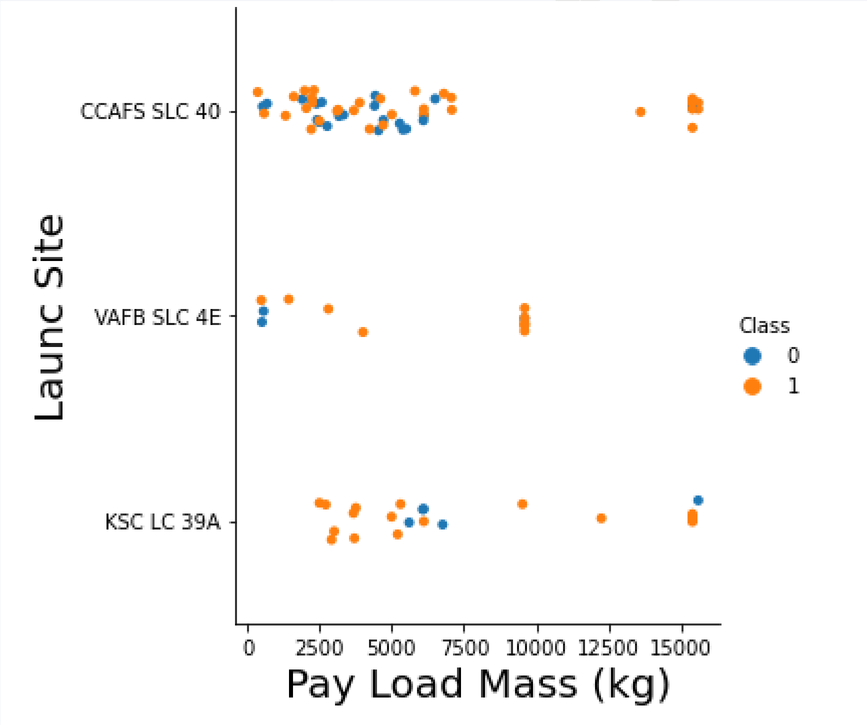
Flight Number vs. Launch Site

- Launches at site KSC LC 39A tend to be quite successful
- Launches at site VAFB SLC 4E tend to be unsuccessful for lower flight numbers and successful for higher flight numbers
- Launches from CCAFS SLC 40 are mixed between successful and unsuccessful



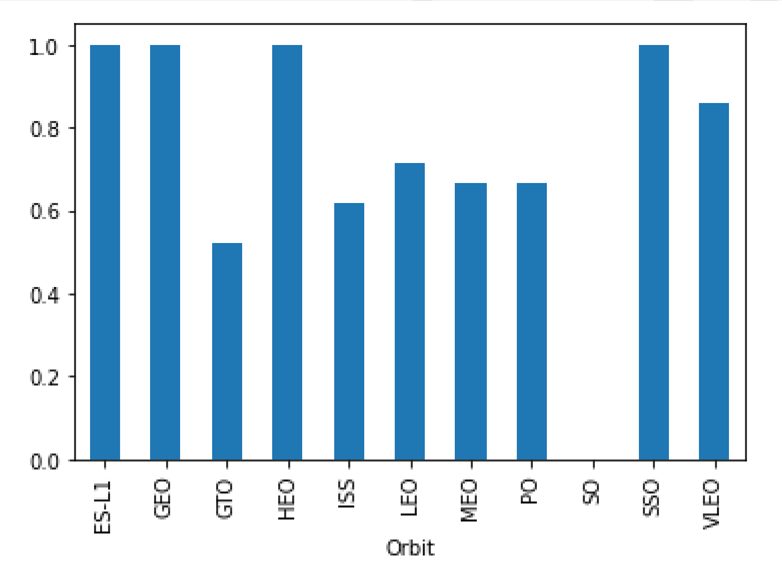
Payload vs. Launch Site

- Launches at site KSC LC 39A tend to be more successful at lower payloads
- Launches at site VAFB SLC 4E tend to be very successful for higher payloads, with the only two failures corresponding to very low payload
- Launches from CCAFS SLC 40 with a higher payload mass are very successful, whereas results for lower masses are mixed



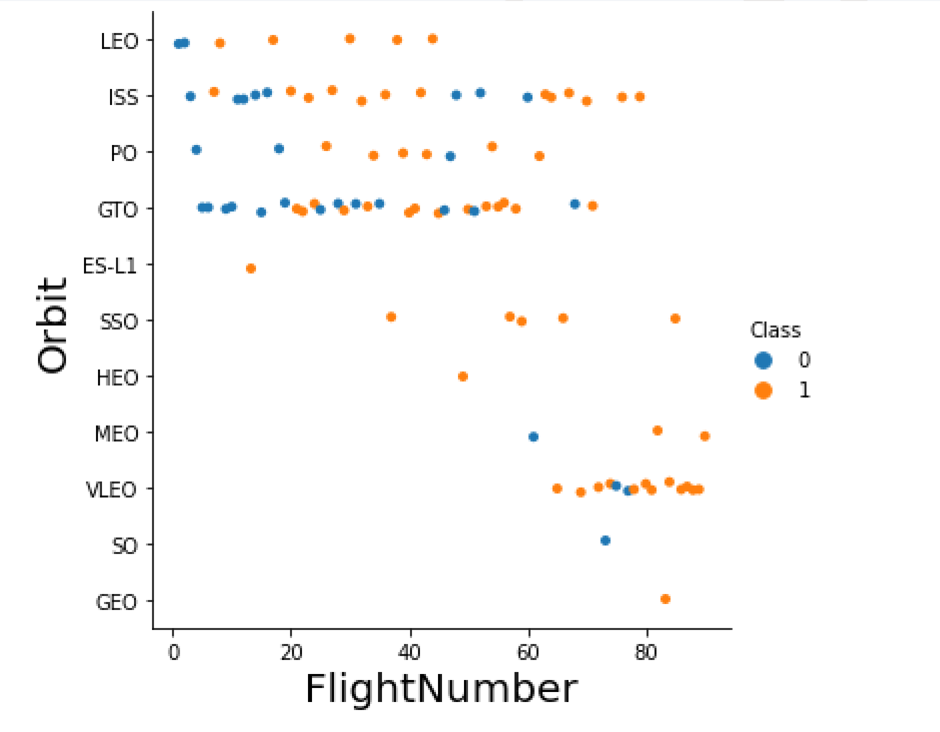
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO Orbits all have a very high success rate
- GTO, ISS, MEO, and PO Orbits have the lowest success rates respectively



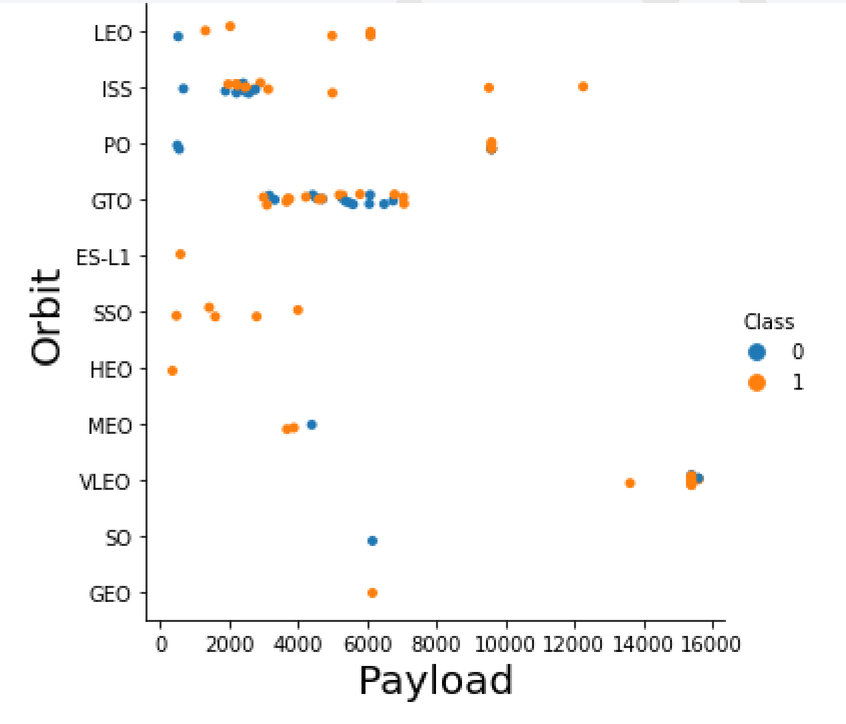
Flight Number vs. Orbit Type

- Lower flight numbers tend to have unsuccessful launches across all orbits
- The VLEO orbit has an impressively high success rate when considering the high number of launches from this site
- The GTO launches have the most failed launches and are therefore the hardest to predict



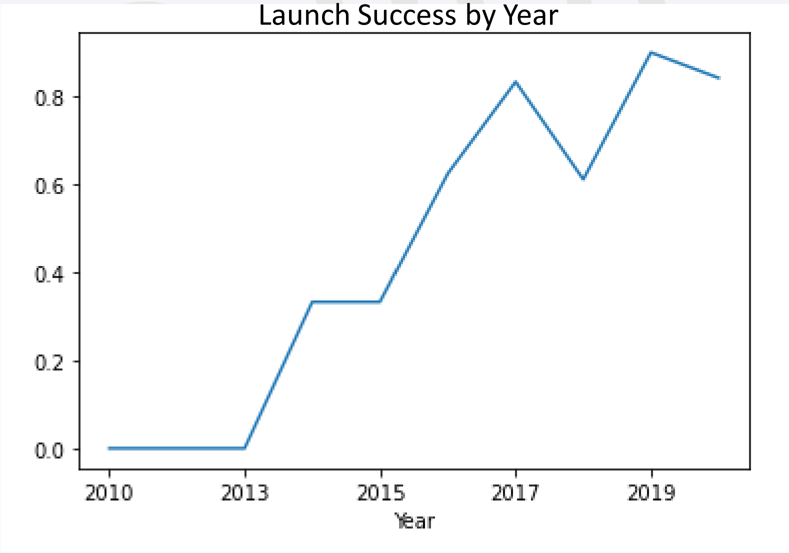
Payload vs. Orbit Type

- Lower payloads for LEO, ISS, and PO orbits tend to be highly correlated with failure
- Lower payloads for ES-L1, SSO, HEO, and MEO tend to be correlated with successful launches
- Higher payloads generally seem to have higher success rates



Launch Success Yearly Trend

- There were no successful launches from 2010-2013
- Launch success has generally increased strongly from 2010 to 2020
- There is a downtick in launch success in 2018 due to a high proportion of failures



All Launch Site Names

- List of distinct launch site names

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE "CCA%" LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Total payload carried by boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_)  FROM SPACEXTBL WHERE CUSTOMER="NASA (CRS)";

 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE "F9 V1.1%";
```

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME="Success (ground pad)";
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload 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

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000) AND (Landing_Outcome="Success (drone s
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome='F
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, ranked in descending order

```
%sql SELECT Landing_Outcome, COUNT(*) FROM SPACEXTBL WHERE DATE BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY Landing_Outcome HA
```

```
 * sqlite:///my_data1.db
Done.
```

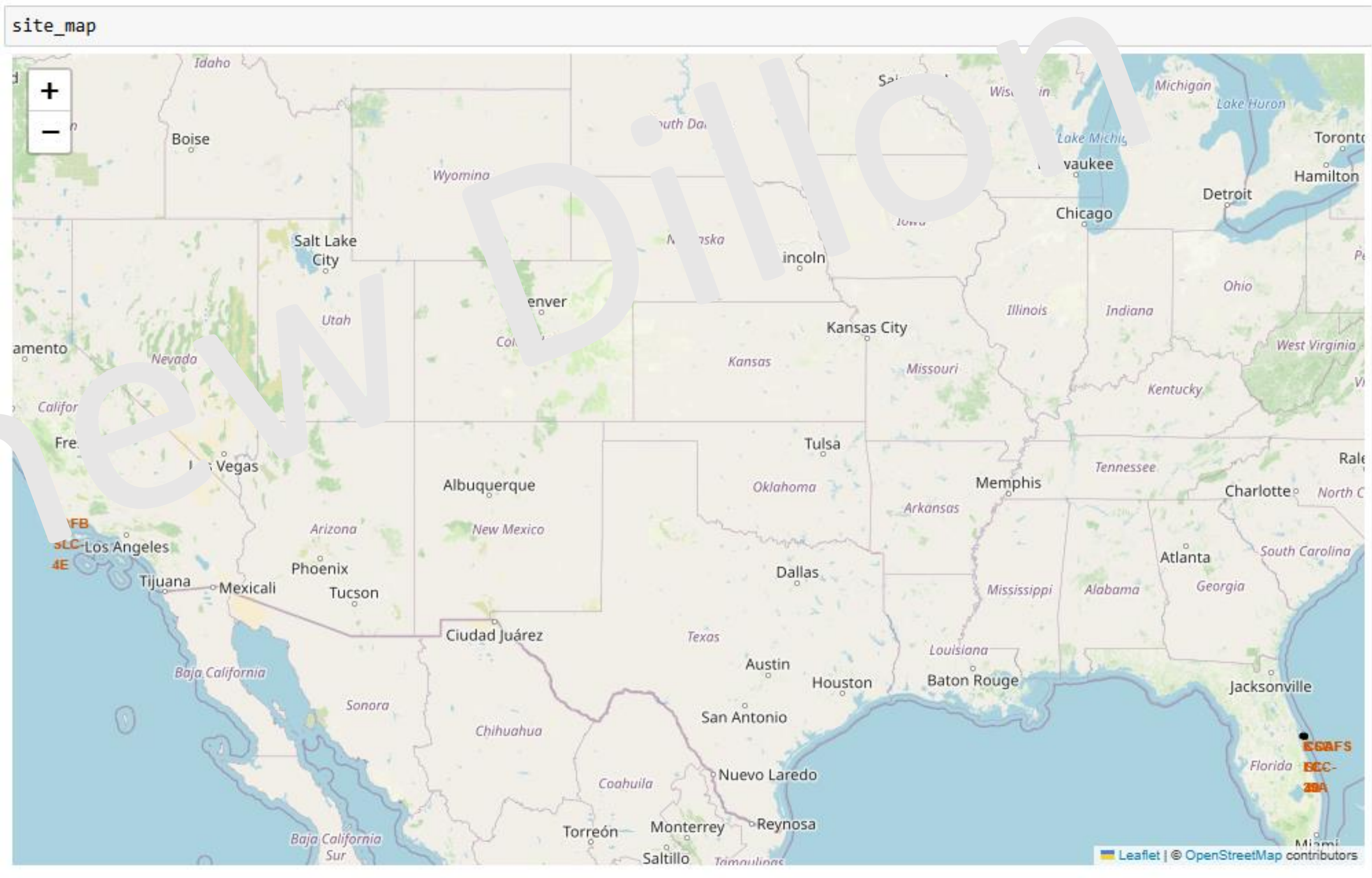
Landing_Outcome	COUNT(*)
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Folium Map – All SpaceX Falcon9 Launch Sites

- Once we plotted the SpaceX Falcon9 rocket launch sites on a map, it becomes evident that SpaceX targets coastal regions of the United States for their launches, specifically in the following two locations:

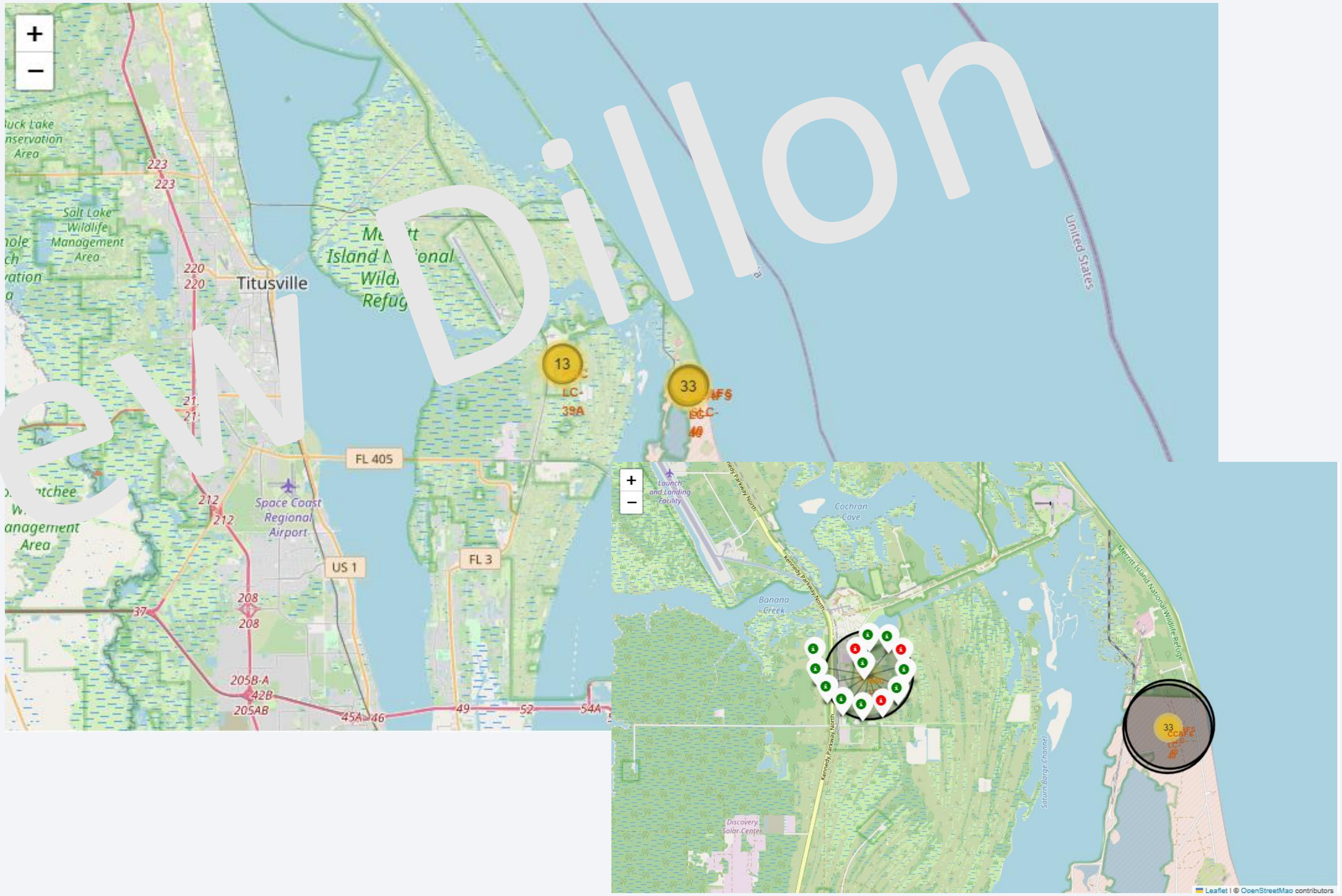


1) Southwestern California

2) Eastern Florida

Launch Success Rate by Launch Site

- By adding color-coordinated markers to our Launch Site location map, we can approximate the overall success of each launch site by identifying the number of green and red markers within each launch site
- The second screenshot shows a more zoomed-in view of the first, both focusing on Florida



Calculating Distance between Landmarks and Launch Sites

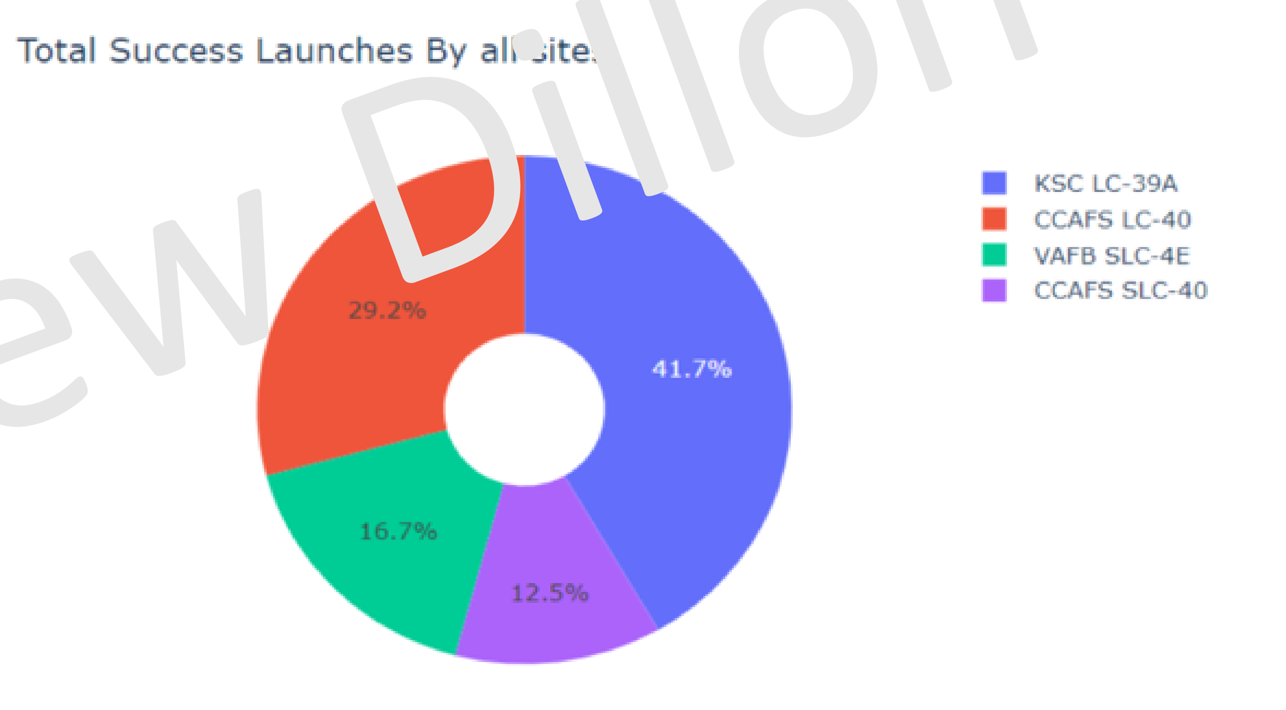
- By leveraging a simple Euclidean distance formula, we can use latitude and longitude coordinates to calculate how far the launch site is from a landmark
- We see in the screenshot to the right that this launch site is a mere 0.87km from the coastline, identified with red text

Section 4

Build a Dashboard with Plotly Dash

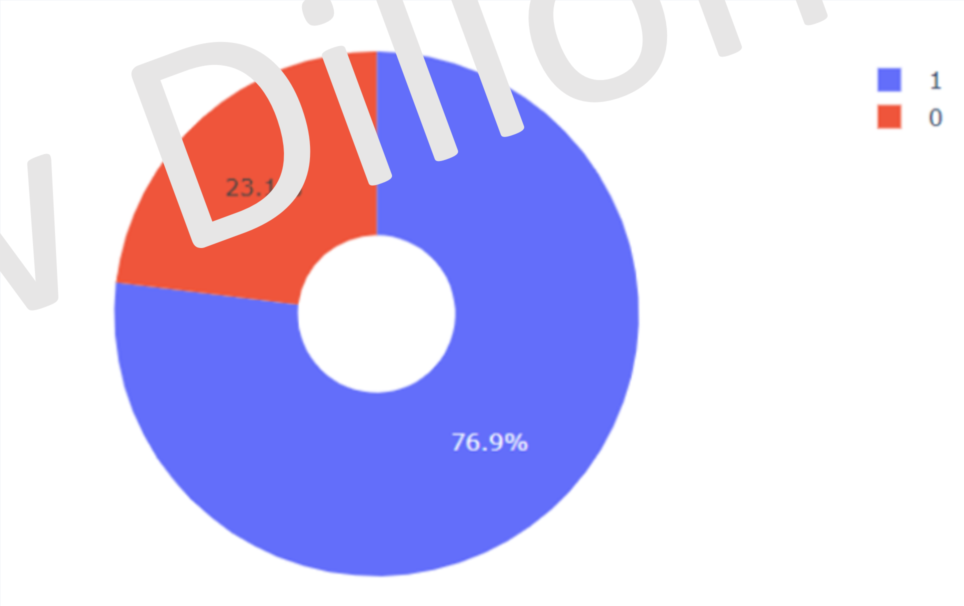
Percentage of Successful Launches by Site

- With a pie chart, we can easily identify which launch sites have the highest and lowest number of successful launches in our dataset
- We see that KSC LC-39A has the highest number of successful launches at nearly 42% of all successes, whereas CCAFS SLC-40 has the lowest at 12.5% of all successful launches



Success Rate of the Most Successful Launch Site

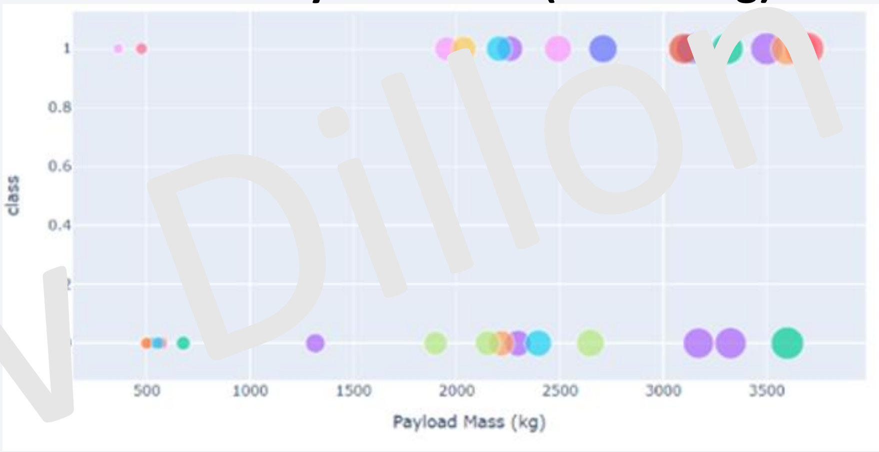
- With a pie chart, we can easily identify the success rate of a specific launch site
- Launch site KSC LC-39A has the highest launch success rate of all launch sites at nearly 77% success, whereas 23% of the launches from this site were classified as a failure
- In the pie chart legend, 1 refers to a success whereas 0 refers to a failure



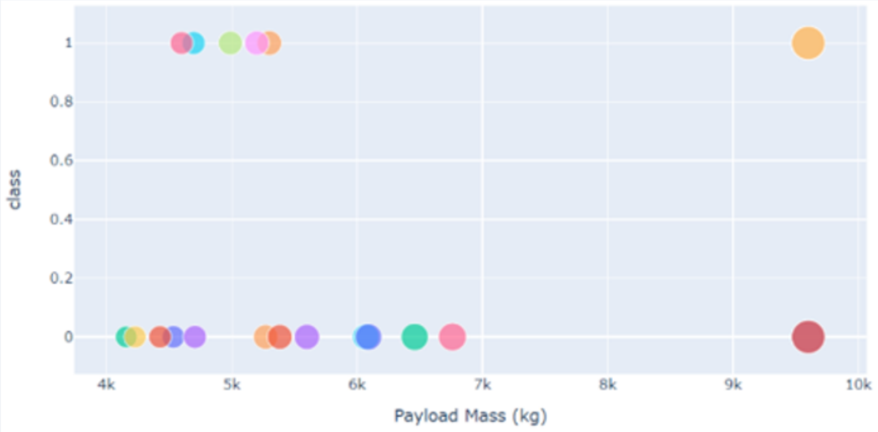
Launch Success vs. Payload Side Scatter Plots

- Using a scatterplot, we can easily identify correlations between Payload Mass and Launch Success
- For lower payload masses, there does not appear to be a strong correlation between payload mass and launch success
- For higher payload masses, the majority are failures, while interestingly payload masses around 5,000kg are highly successful

Lower Payload Mass (<= 4000kg)



Higher Payload Mass (> 4000kg)

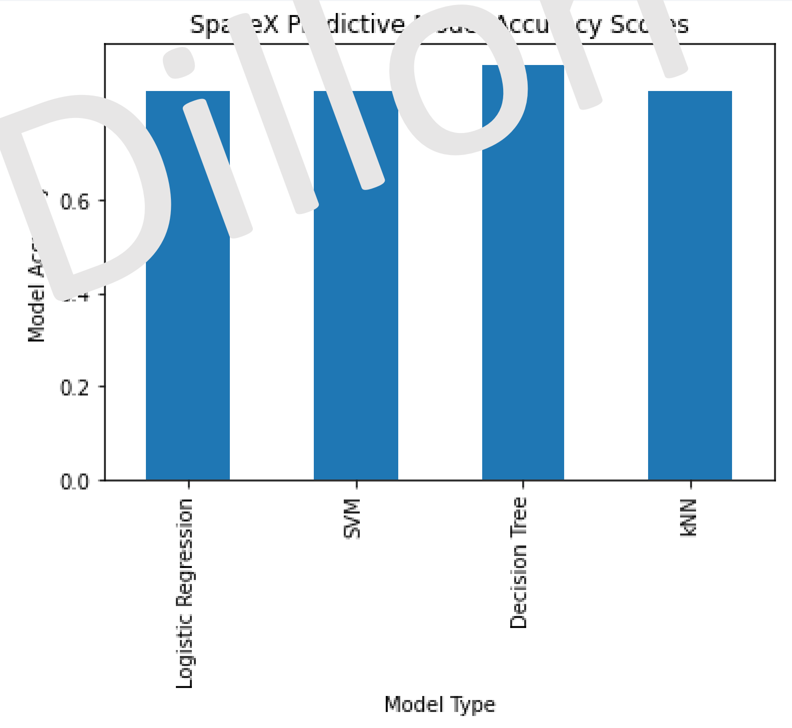


Section 5

Predictive Analysis (Classification)

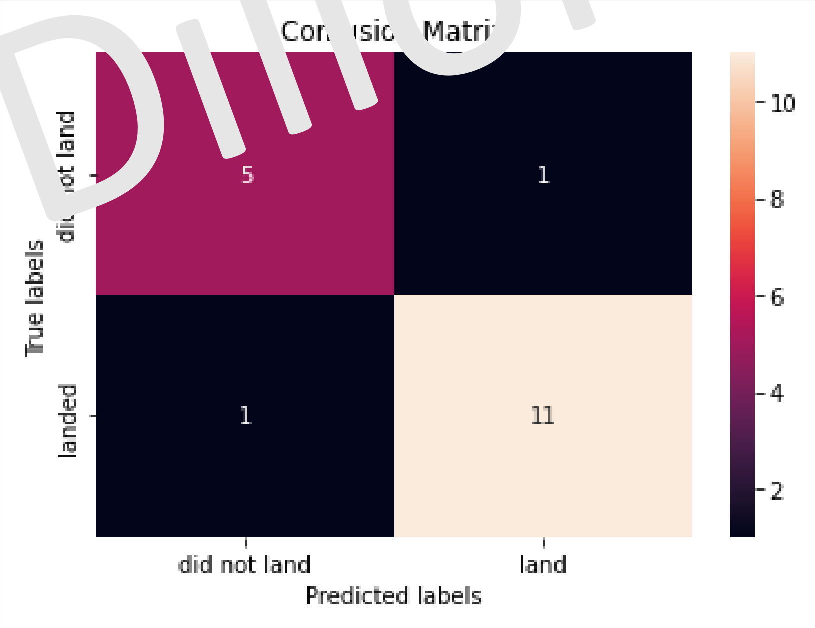
Classification Accuracy

- The out-of-sample accuracy of our four predictive machine learning models can be seen in the matplotlib bar chart to the right
- While all models perform quite well, the Decision Tree model has the highest accuracy at roughly 89%



Confusion Matrix

- The Confusion Matrix for our Decision Tree model is shown to the right
- The model correctly identified 11 launches that landed successfully and 5 launches that failed
- The model incorrectly predicted one failed launch was a success (false positive) and one successful launch was a failure (false negative)
- 16/18 launches were predicted accurately



Conclusions

- Based on our analysis and exploration, the key findings are as follows:

1. Launch sites with a higher number of launches over time tend to have a higher overall launch success rate
2. SpaceX's launch success rate has steadily increased from 2013 until 2020, with one downturn in 2018 which has since recovered and has been surpassed
3. The KSC LC-39A launch site is the most successful of any launch site based on a variety of metrics
4. We have developed a Decision Tree predictive classification model that will be able to predict launch success with roughly 90% accuracy on unseen data
5. We will need to tailor our machine learning approach as new data continues to become available for our investigation and analysis

Appendix

GitHub URLs for Python Scripts Referenced:

- https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-1-jupyter-labs-spacex-data-collection-api-v2.ipynb
- https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-2-jupyter-labs-webscraping.ipynb
- https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-3-labs-jupyter-spacex-Data%20wrangling-v2.ipynb
- https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-4-jupyter-labs-eda-sql-coursera_sqllite.ipynb
- https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-5-jupyter-labs-eda-dataviz-v2.ipynb
- https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-6-lab-jupyter-launch-site-location-v2.ipynb
- https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-7-Build_a_Dashboard_Application_with_Plotly_Dash.ipynb
- https://github.com/matthewdillon1/Applied_Data_Science_Capstone_Final_Project/blob/main/MDillon-8-SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Thank you!