IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chad Liu
August 24th, 2023

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

The objective of this study is to thoroughly examine SpaceX Falcon 9 data to forecast the success of the first stage landing using machine learning models. This predictive capability will provide valuable insights for other space agencies deciding to bid against SpaceX.

Summary of Methodologies:

- Gather data through API and Web Scraping
- Perform data transformation
- Employ SQL and data visualizations for exploratory analysis.
- Construct a dashboard for interactive visual analysis
- Conclude by creating a predictive model to determine the likelihood of successful landings.

Summary of All Results:

- The best machine learning classification model to predict the success of landings was the Decision Tree Classifier
- There has been an increase in the likelihood of the first stage successfully landing over time.

Introduction

Project Background and Context

Space X stands out as a foremost contender in the commercial space travel arena, having achieved remarkable success. Prominently featured on its website are Falcon 9 rocket launches, priced at 62 million dollars per launch. This cost sharply contrasts with other providers' fees, which soar to 165 million dollars or more per launch. Notably, Space X achieves substantial savings by reusing the first stage of their rockets. Consequently, accurately predicting the success of the first stage landing could pave the way for precise launch cost estimations. This strategic insight becomes especially pertinent when a competitor aims to challenge Space X's bid in the rocket launch market.

Questions to be answered:

1. What factors influence the achievement of a successful first stage landing?
2. How do these attributes impact the likelihood of a successful landing?
3. How are the success rate, time, and number of flights interrelated?
4. Which model offers the highest accuracy in forecasting successful flight landings?

Section 1

Methodology

Methodology – Executive Summary

Data collection methodology:

- Using SpaceX REST API
- Employing web scraping from Wikipedia

Perform data wrangling:

- Created training labels by converting outcomes into binary format
- Processed data by filtering, handling missing values, and employing One Hot Encoding for binary classification.

Perform exploratory data analysis (EDA) using visualization and SQL

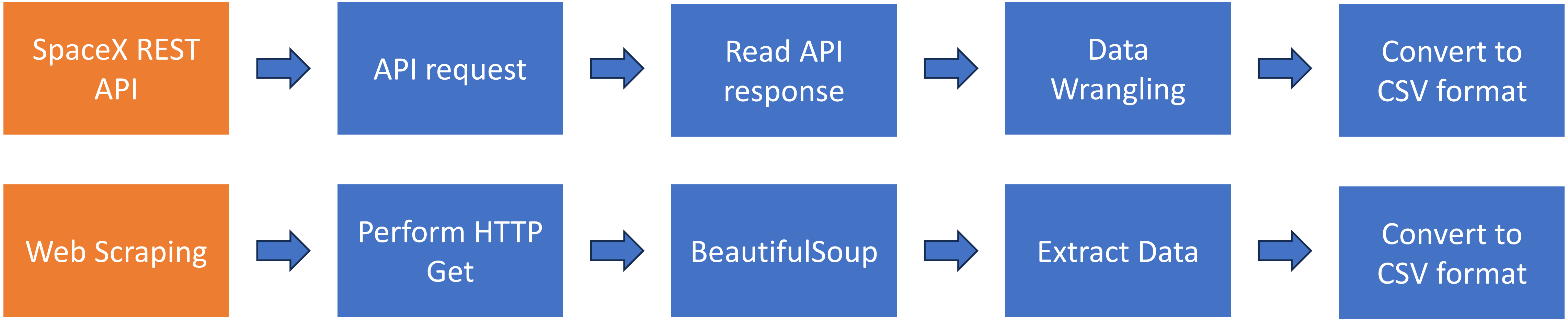
Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models:

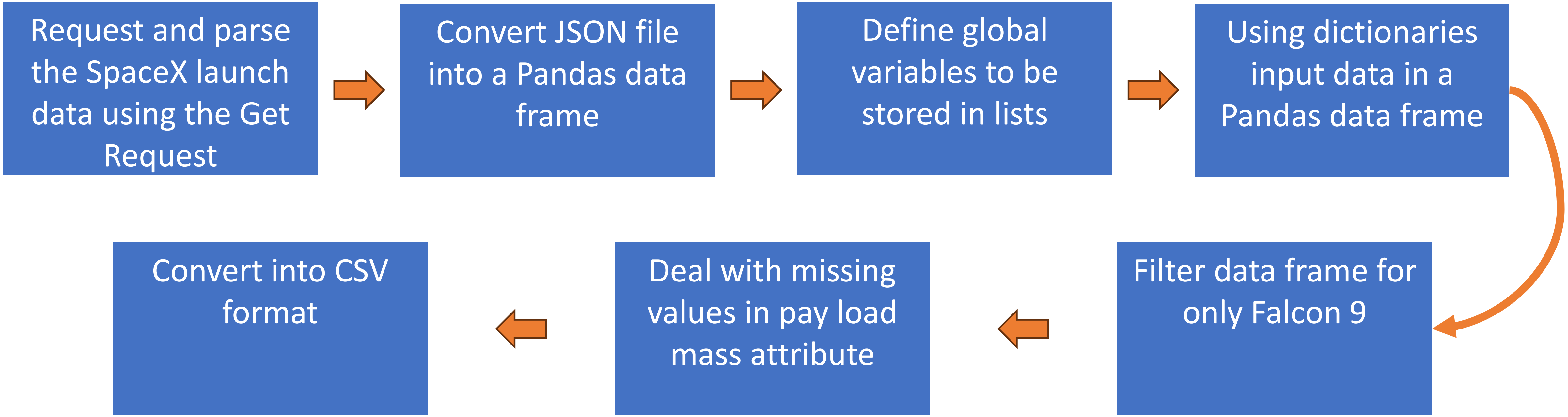
- Conducted data standardization and transformation, partitioned data into training and testing sets, and evaluated and determined the optimal classification algorithm

Data Collection

For the purpose of this project, data gathering involved utilizing the SpaceX API and extracting information from relevant launch data on Wiki pages through web scraping.



Data Collection – SpaceX API



Request and parse the SpaceX launch data using the Get Request

Convert JSON file into a Pandas data frame

Define global variables to be stored in lists

Using dictionaries input data in a Pandas data frame

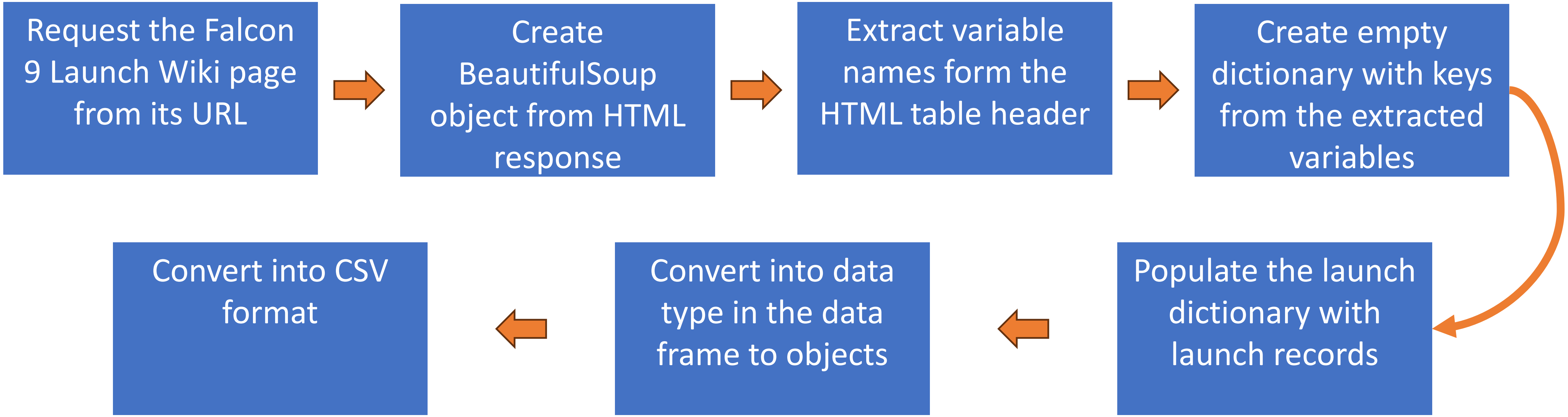
Filter data frame for only Falcon 9

Deal with missing values in pay load mass attribute

Convert into CSV format

References:
Data Collection SpaceX API: https://github.com/OfficialKZYN/IBM-Data-Science-Professional-Certificate/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Data-Collection-SpaceX-Rest-API.ipynb

Data Collection – Web Scrapping



Request the Falcon 9 Launch Wiki page from its URL

Create BeautifulSoup object from HTML response

Extract variable names form the HTML table header

Create empty dictionary with keys from the extracted variables

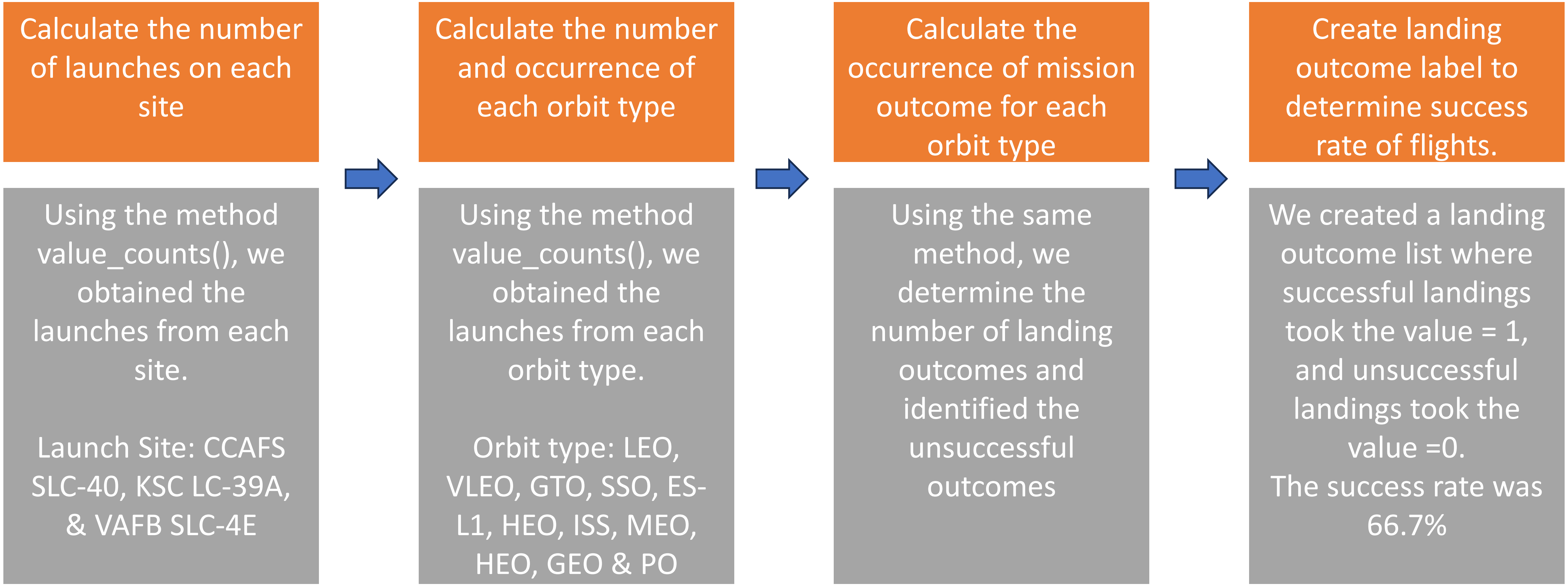
Populate the launch dictionary with launch records

Convert into data type in the data frame to objects

Convert into CSV format

References:
Data Collection Web Scrapping: https://github.com/OfficialKZYN/IBM-Data-Science-Professional-Certificate/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Data-Collection-Web-Scrapping.ipynb

Data Wrangling



Calculate the number of launches on each site

Using the method value_counts(), we obtained the launches from each site.

Launch Site: CCAFS SLC-40, KSC LC-39A, & VAFB SLC-4E

Calculate the number and occurrence of each orbit type

Using the method value_counts(), we obtained the launches from each orbit type.

Orbit type: LEO, VLEO, GTO, SSO, ES-L1, HEO, ISS, MEO, HEO, GEO & PO

Calculate the occurrence of mission outcome for each orbit type

Using the same method, we determine the number of landing outcomes and identified the unsuccessful outcomes

Create landing outcome label to determine success rate of flights.

We created a landing outcome list where successful landings took the value = 1, and unsuccessful landings took the value =0.
The success rate was 66.7%

References:
Data Wrangling: https://github.com/OfficialKZYN/IBM-Data-Science-Professional-Certificate/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Data-Wrangling.ipynb

EDA with Data Visualization

Summary of Charts:

- Payload Mass vs Flight Number Scatter Plot - this plot depicts the success rate of landings with varying payload mass and the number of flights, enabling us to determine the relationship between Payload mass and flight number on landing success.
- Flight Number vs Launch Site Scatter Plot – This plot enables us to visualize the success rate and number of flights from each Launch site, evaluating whether the launch site has an impact on landing success rate.
- Flight Number vs. Orbit Type Scatter Plot – This plot determines whether the high success rate of orbits is due to high number of flights.
- Orbit Type vs. Payload Mass Scatter Plot– This plot which orbits can carry high payload and which orbits can carry low payload mass.
- Landing Success vs. Orbit Type Bar Plot – This plot helps find which orbits have the highest success rate
- Average Success Rate vs. Year Line Plot – This plot depicts whether the number of successful landings has increased over the years, indicating improvement in design and implementation of rocket.

References:
EDA with Data Visualization: https://github.com/OfficialKZYN/IBM-Data-Science-Professional-Certificate/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/EDA-with-Data-Visualisation.ipynb

EDA with SQL

SQL Queries Performed:

Display the names of the launch sites in the space mission.

Display 5 records where launch sites begin with string ‘CCA’

Display the total payload mass by boosters launched by NASA (CRS)

Display average payload mass carried by booster version F9 v1.1

List the date when the first successful landing outcome for ground pad

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

List the total number of successful and failure mission outcomes

List the names of the booster versions which have carried the highest payload mass.

List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

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

References:
EDA with SQL: https://github.com/OfficialKZYN/IBM-Data-Science-Professional-Certificate/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/EDA-with-SQL.ipynb

Build an Interactive Map with Folium

Summary of Map Objects:

1. Marked all launch sites - this allows us to quickly see the locations of the 4 unique launch sites identified in the data set.
2. Marked launch sites with red circles – this allows us to see the launch sites area without having to zoom to each city to see.
3. Marked the success / failed launches for each site – this enables us to quickly identify the success rate of each launch site through the map.
4. Marked the distance between the launch site and nearest coastal line- this allows us to understand the distance the rocket are tested for their landings.
5. Marked the distance between the launch site and nearest railway – Ensures safe distance for testing
6. Marked the distance between the launch site and nearest highway – Enables us to understand how rural the area these rockets are being tested are.
7. Marked the distance between the launch site and nearest city – Ensures safe distance from major cities in case of unsuccessful landings

References:
Interactive Map with Folium: https://github.com/OfficialKZYN/IBM-Data-Science-Professional-Certificate/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Build-an-interactive-map-with-Folium.ipynb

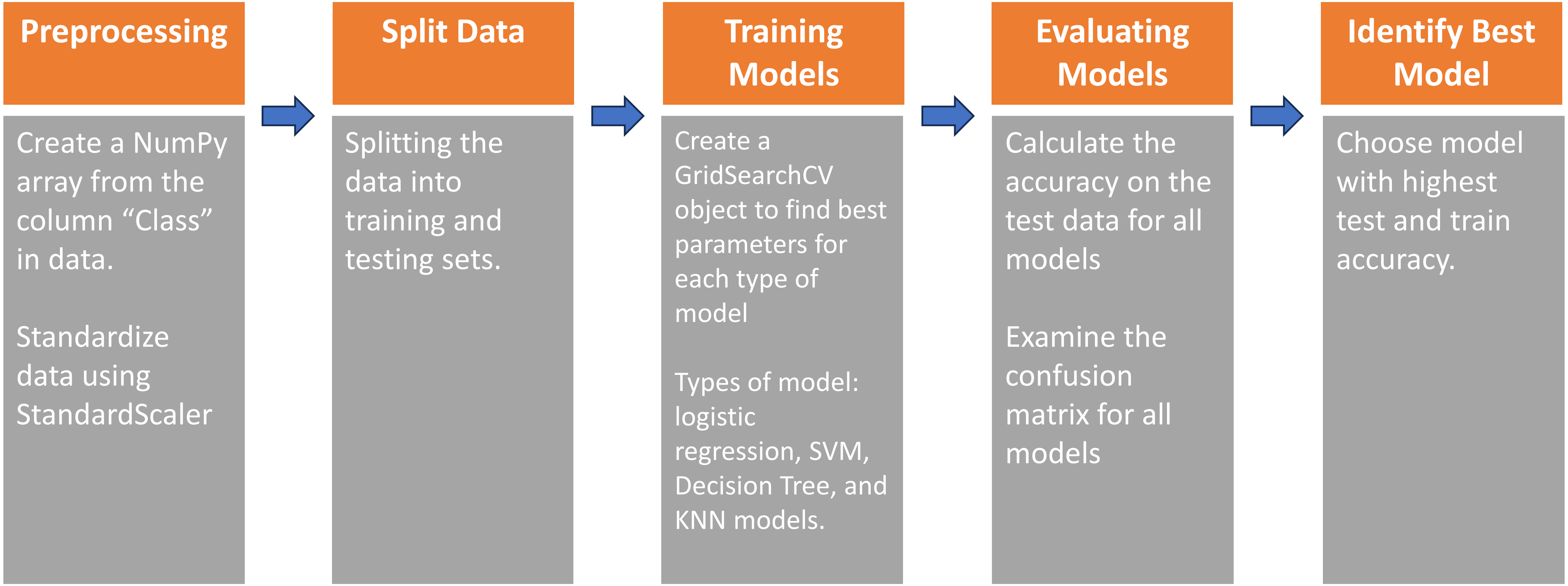
Build a Dashboard with Plotly Dash

Summary of Dashboard:

1. Added a dropdown list to enable selection of different Launch sites
2. Added a call back function to display a pie chart on selected site through the drop down, enabling stakeholders to access the total successful launches count for all sites and the success rate of launches on selected sites.
3. Added a payload mass range slider to easily access the payload range
   - Able to access landing outcomes for a range of payload mass with different types of booster versions
4. Added a scatter graph to display the relationship between the payload mass and success rate of landings for the different types of booster versions.
   - This enables users to quickly see the relationship between the range of payload with different types of booster versions and success rate.

References:
Dashboard with Plotly Dash: https://github.com/OfficialKZYN/IBM-Data-Science-Professional-Certificate/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Dashboard-with-Plotly-Dash.py

Predictive Analysis (Classification)



Preprocessing

Create a NumPy array from the column “Class” in data.

Standardize data using StandardScaler

Split Data

Splitting the data into training and testing sets.

Training Models

Create a GridSearchCV object to find best parameters for each type of model

Types of model: logistic regression, SVM, Decision Tree, and KNN models.

Evaluating Models

Calculate the accuracy on the test data for all models

Examine the confusion matrix for all models

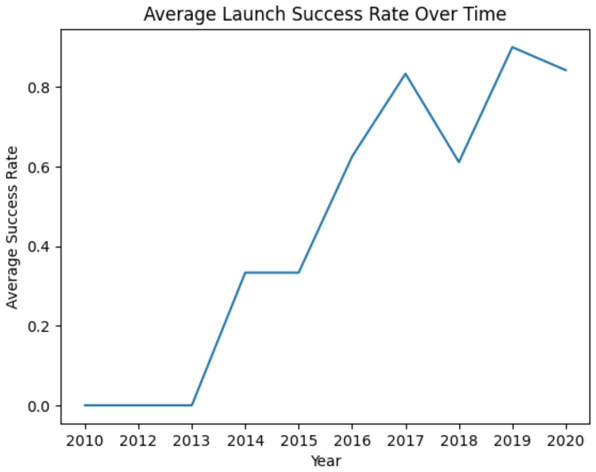
Identify Best Model

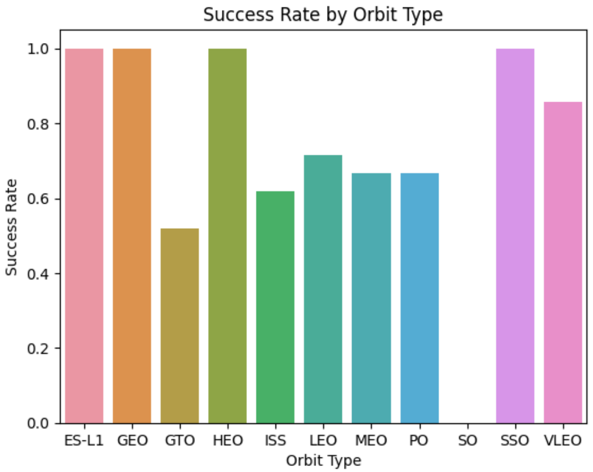
Choose model with highest test and train accuracy.

References:
Predictive Analysis: https://github.com/OfficialKZYN/IBM-Data-Science-Professional-Certificate/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Prediction-Analysis-Classification.ipynb

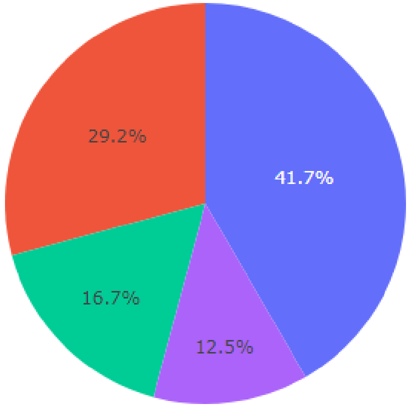
Results

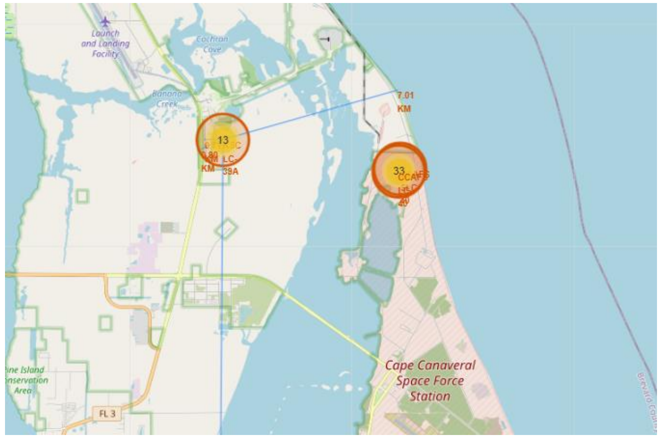
Exploratory Analysis Results:





Interactive Analytics Result




Launch and Landing Facility
Cochran Cove
Banana Creek
7.01 KM
13
33
Cape Canaveral Space Force Station
FL 3

Predictive Analysis Results:

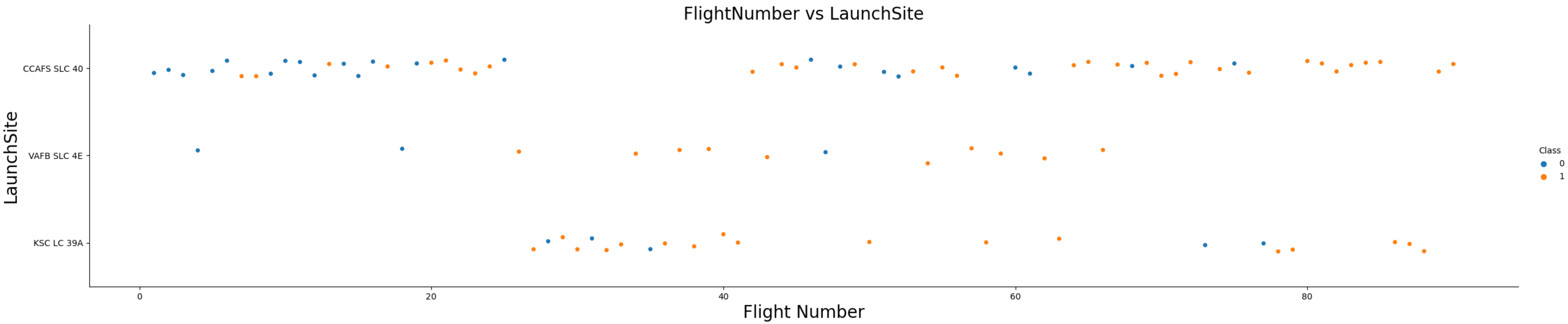
Model Type	Accuracy Score	Test Data Accuracy Score
Decision Tree	0.876786	0.833333
KNN	0.875000	0.833333
SVM	0.848214	0.833333
Logistic Regression	0.846429	0.833333

- From the predictive analysis, it shows that the Decision Tree Classifier is the best model to predict Falcon 9 successful landings, having an accuracy of 87% and test data accuracy of 83%.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



Insights:

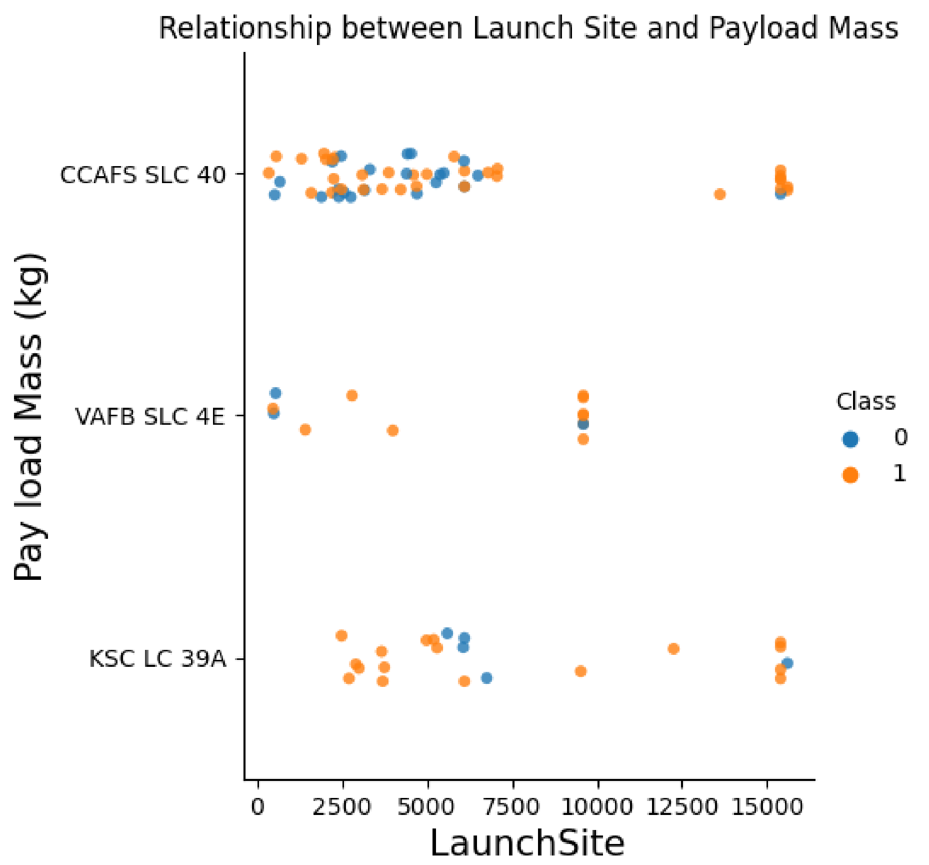
1. From the chart, we can see that there are 3 unique launch sites and CCAFS SLC-40 has the highest number of flights.

2. As the number of flights increases, the number of successful landings has increased as well, showing an improvement in the design and implementation of the Falcon 9 rocket.

3. Most launches recently are conducted on the CCAFS SLC-40 site.

References:

Payload vs. Launch Site

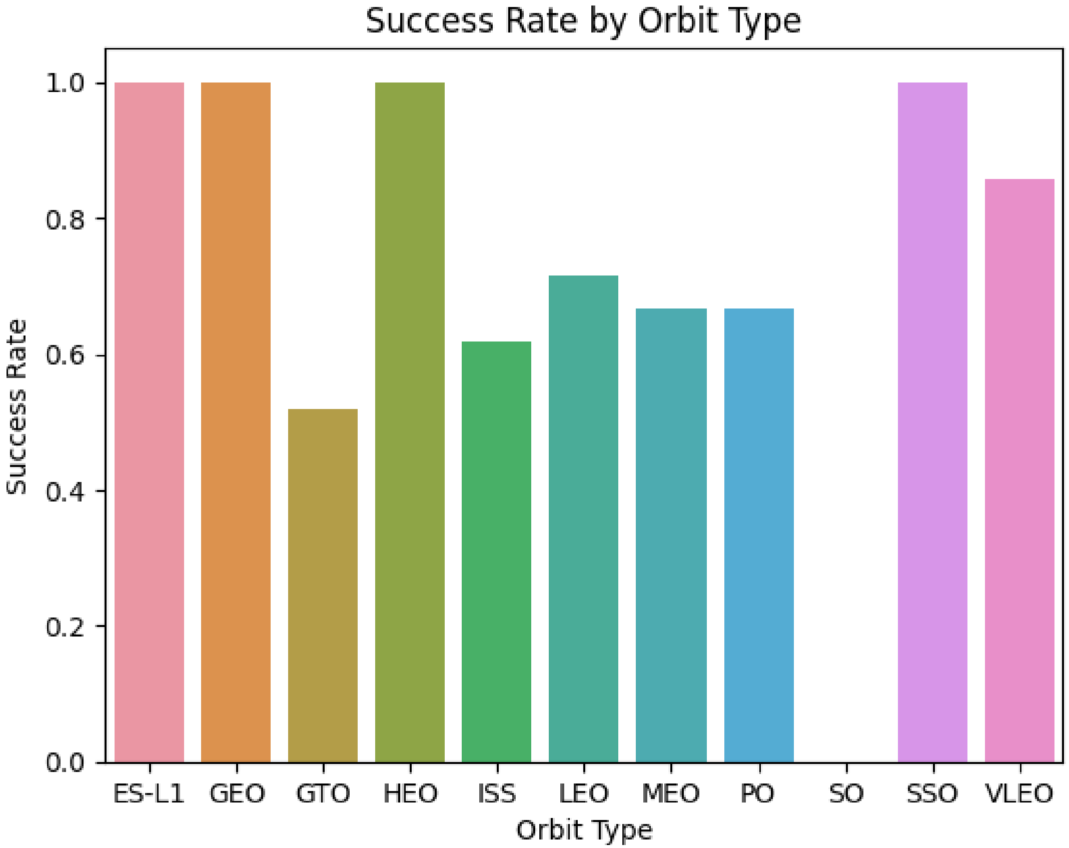


References:

Insights:

1. This chart displays the relationship between Launch Site, the payload mass, and landing outcomes for each launch.
2. Surprisingly, payload mass of larger than 10,000 kg have an excellent success rate
3. VAFB SLC launch site has no rockets that launched for heavy payload mass (greater than 10,000kg)

Success Rate vs. Orbit Type



Insights:

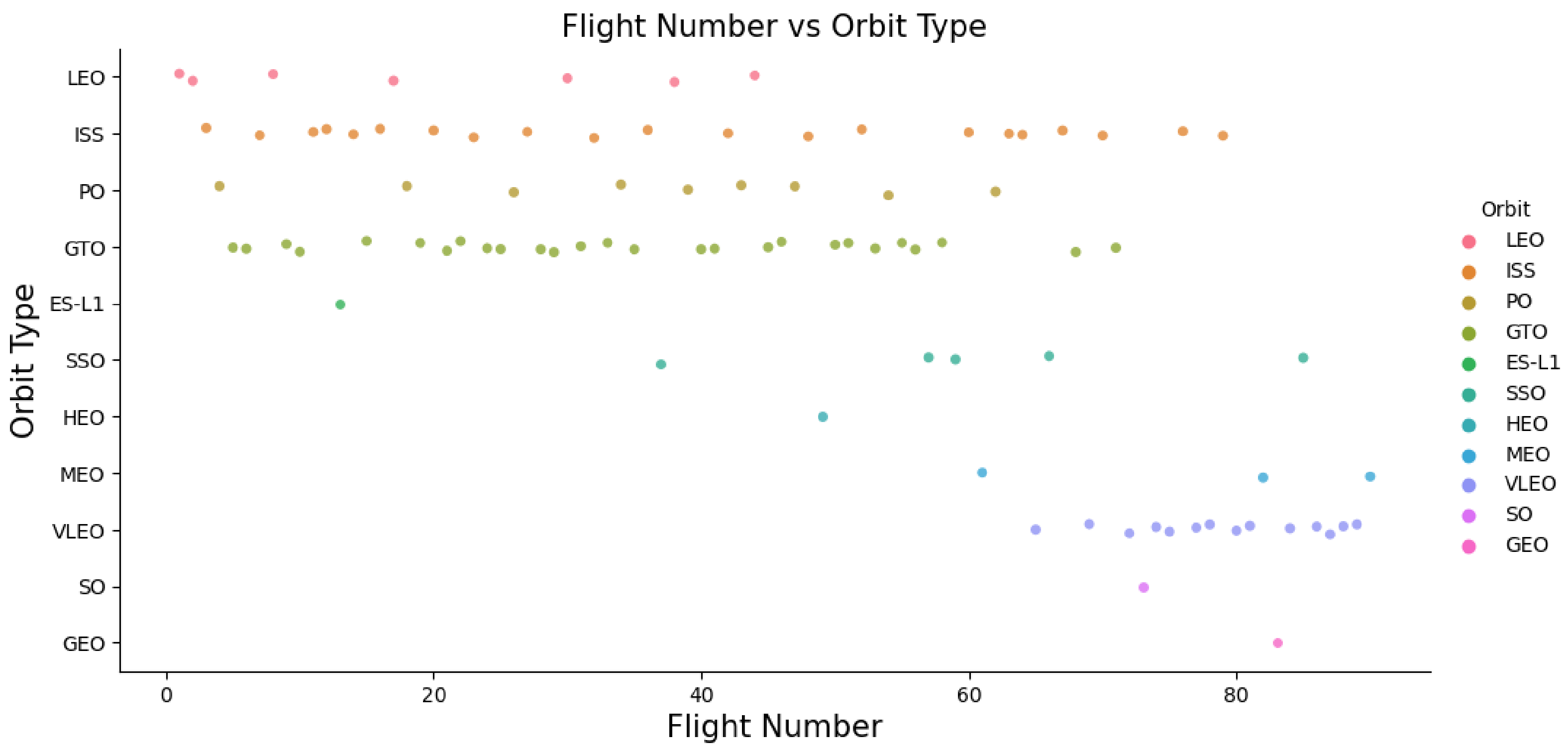
1. This chart displays the relationship between the success rate of landings with the orbit type the rocket was intended

2. ES-L1, GEO, HEO, & SSO are orbit types that has the highest success rate of successful landings

3. We have yet to seen a successful landing launch for the orbit type SO.

References:

Flight Number vs. Orbit Type



Insights:

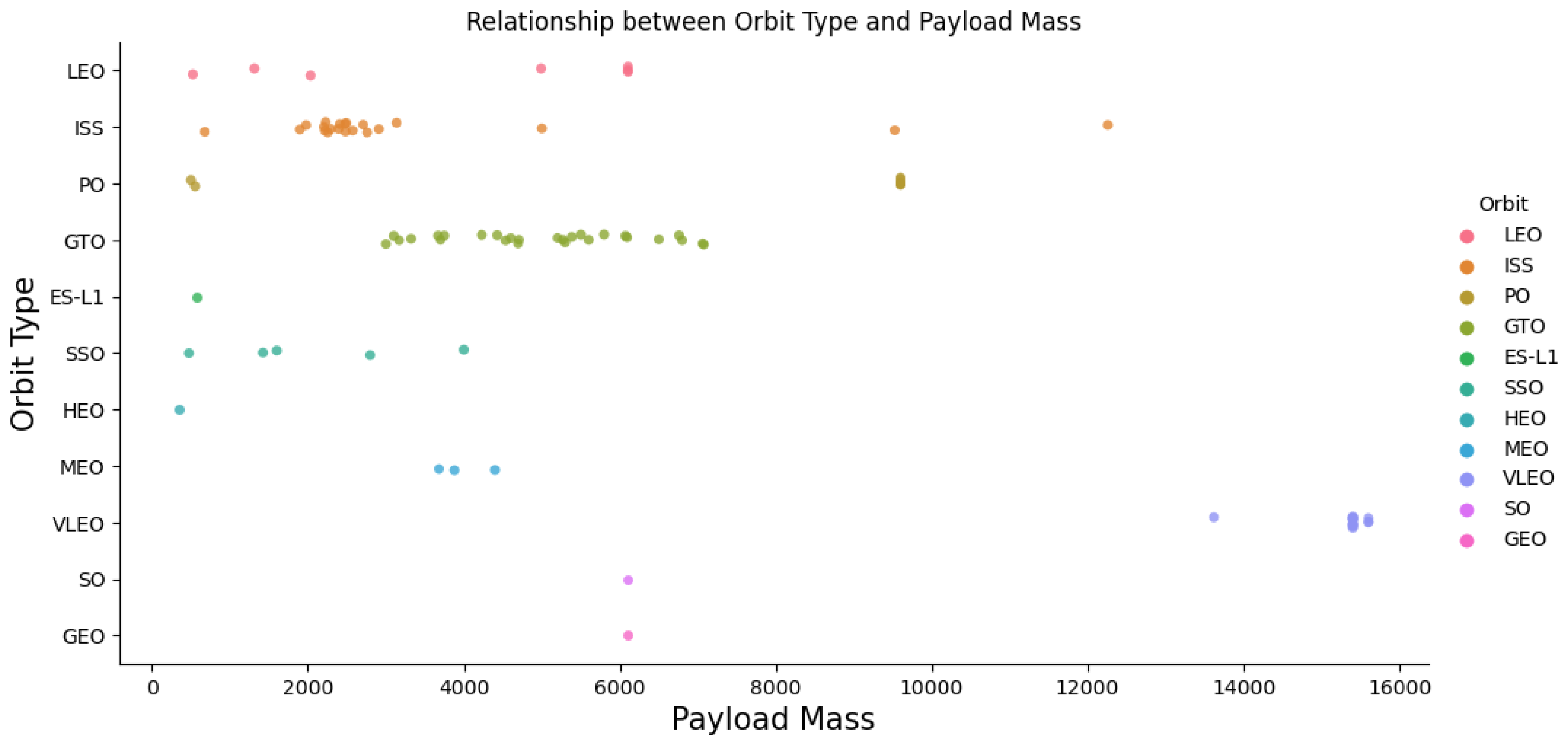
1. GTO, ISS, and VLEO has the highest number of flights.

2. Despite SO having no successful landings could be due to its small flight number of 1.

3. Most recent flights have been focusing on the VLEO orbit type, where previously Falcon9 launches have been focusing on LEO, ISS, PO, and GTO.

References:

Payload vs. Orbit Type



Insights:

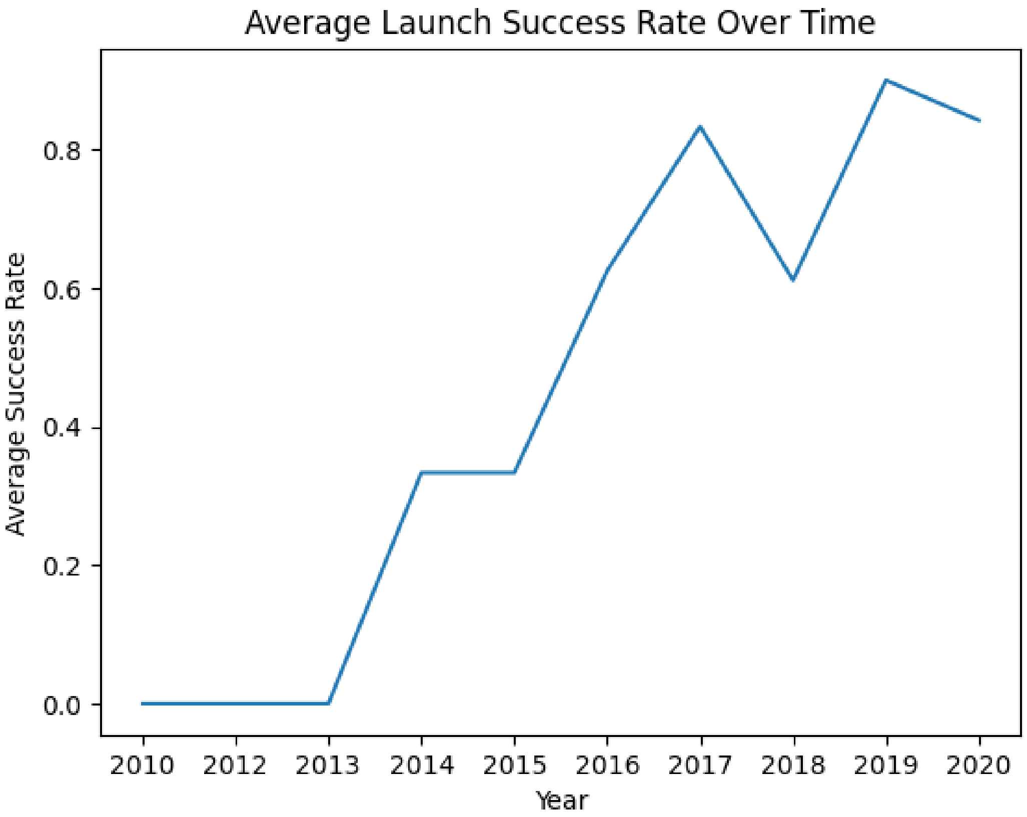
1. Most orbit types have been tested with a payload mass of less than 8,000kg.

2. The increase in the number of flights of VLEO orbit type could be due to its capability of holding a heavier payload mass compared to the rest ranging from 12,000kg – 16,000kg

3. HEO orbit type has the lowest payload mass launch approximately 500kg.

References:

Launch Success Yearly Trend



Insights:

1. This chart displays the relationship between the success rate of landings over time
2. The average success rate of landings have increased over time.
3. We have seen an improvement from 0% in 2013 to 85% in 2019.

References:

All Launch Site Names

According to the dataset, there are 4 launch sites:

Launch Site Namesc
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

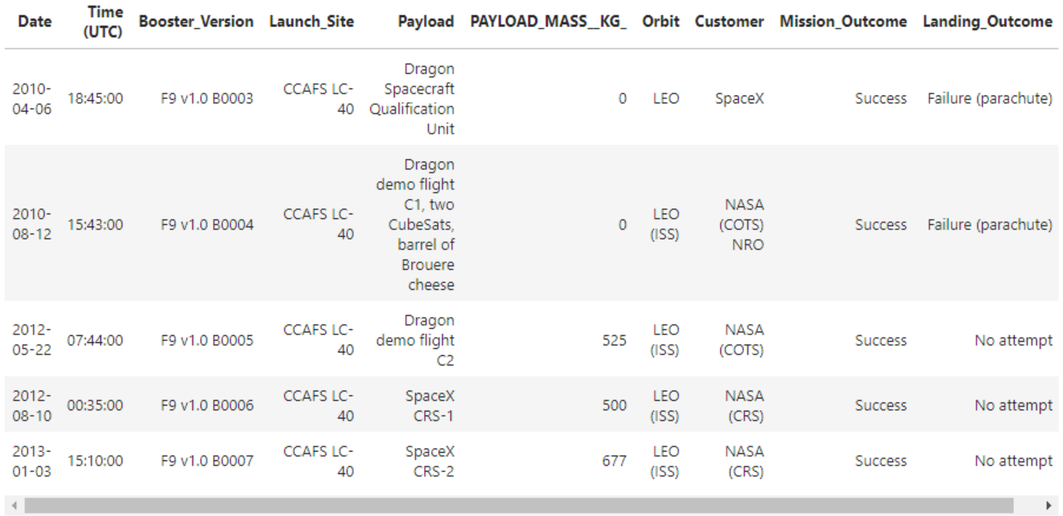
They are acquired by choosing distinct instances of "launch_site" values from the dataset.

SQL Query: Select DISTINCT "Launch_Site" from SPACEXTBL

References:

Launch Site Names Begin with 'CCA'

According to the dataset, the 5 records where launch sites begin with the string ‘CCA’:



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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	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

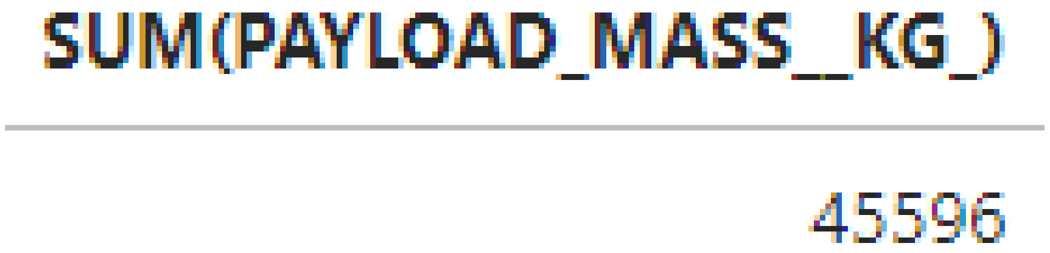
They are acquired by choosing distinct instances of "launch_site" values from the dataset that begin with the string ‘CCA’.

SQL Query: SELECT * from SPACEXTBL where "Launch_Site" like "CCA%" LIMIT 5

References:

Total Payload Mass

According to the dataset, the total payload mass carried by boosters launched by NASA (CRS):



SUM(PAYLOAD_MASS__KG_)
45596

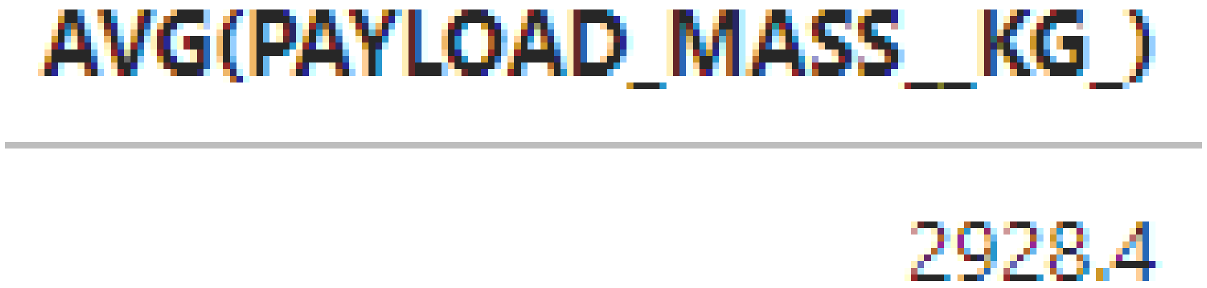
They are acquired by selecting the sum of payload mass where the boosters were launched by NASA (CRS)

SQL Query: Select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE "Customer" == "NASA (CRS)"

References:

Average Payload Mass by F9 v1.1

According to the dataset, the average payload mass carried by booster version F9 v1.1:



AVG(PAYLOAD_MASS__KG_)
2928.4

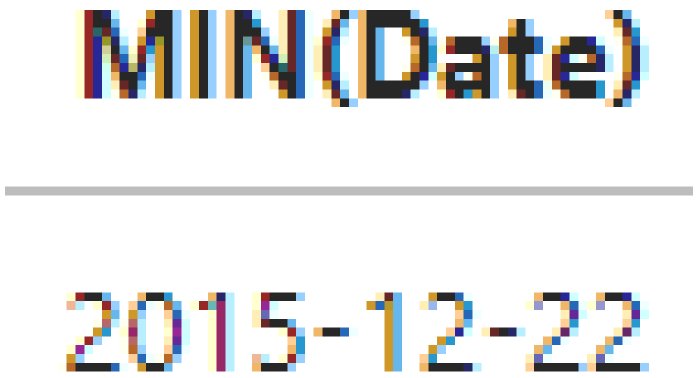
They are acquired by selecting the average payload mass where the booster version was F9 v1.1

SQL Query: Select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE "Booster_Version" == "F9 v1.1"

References:

First Successful Ground Landing Date

According to the dataset, the first successful landing outcome in ground pad was:



MIN(Date)
2015-12-22

They are acquired by selecting the minimum date value from the table where the landing outcome is successful for ground pad

SQL Query: Select MIN(Date) from SPACEXTBL WHERE "Landing_Outcome" LIKE "Success (g%"

References:

Successful Drone Ship Landing with Payload between 4000 and 6000

According to the dataset, boosters with success in drone ship and have payload mass greater than 4000 but less than 6000:

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

They are acquired by selecting the distinct booster version where the landing outcome is successful in drone ship and a payload mass greater than 4000 but less than 6000

SQL Query: SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" LIKE "Success (d%" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 and 6000

References:

Total Number of Successful and Failure Mission Outcomes

According to the dataset, the total number of successful and failure mission outcomes:

Mission Outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

They are acquired by selecting the mission outcome and counting the mission outcomes

SQL Query: Select Mission_Outcome, COUNT(MISSION_OUTCOME) from SPACEXTBL GROUP BY "MISSION_OUTCOME"

References:

Boosters Carried Maximum Payload

According to the dataset, the booster versions which have carried the maximum payload mass:

Booster Versions		
F9 B5 B1048.4	F9 B5 B1048.5	F9 B5 B1058.3
F9 B5 B1049.4	F9 B5 B1051.4	F9 B5 B1051.6
F9 B5 B1051.3	F9 B5 B1049.5	F9 B5 B1060.3
F9 B5 B1056.4	F9 B5 B1060.2	F9 B5 B1049.7

They are acquired by selecting distinct booster version where payload mass is the highest payload mass from the dataset.

SQL Query: SELECT DISTINCT "Booster_Version" from SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

References:

2015 Launch Records

According to the dataset, the failed launch records for drone ship in 2015:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

They are acquired by selecting the months of 2015 where the landing outcome is a failure for drone ship in 2015

SQL Query: SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure (d%' AND substr(Date, 1, 4) = '2015’

References:

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

According to the dataset, the landing outcomes between 2010-06-04 and 2017-03-20:

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

They are acquired by selecting the landing outcome and count of the landing outcome where the launch date was between 2010-06-04 and 2017-03-20.

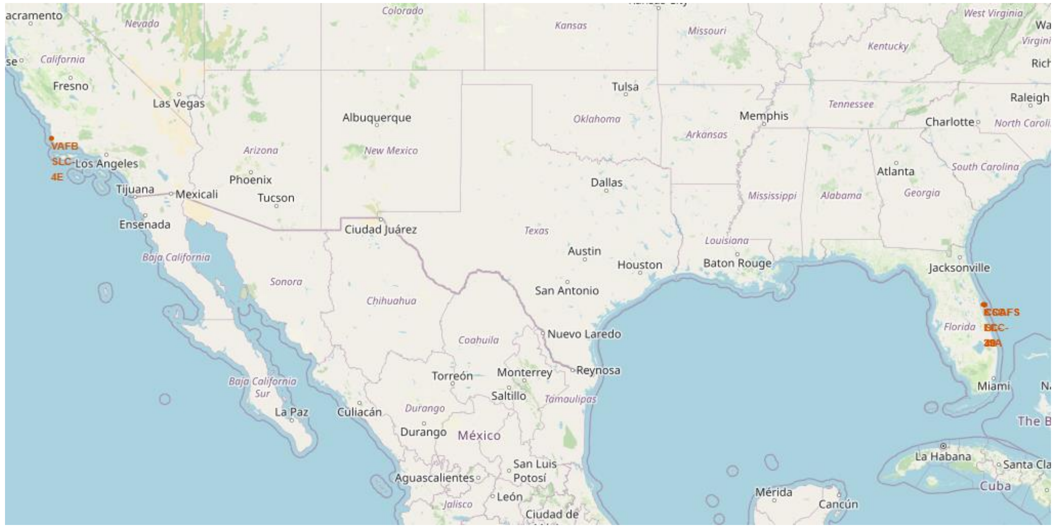
SQL Query: SELECT Landing_Outcome, COUNT(*) as Count FROM SPACEXTBL Where Date between '2010-06-04' and '2017-03-20' Group by Landing_Outcome Order by Co

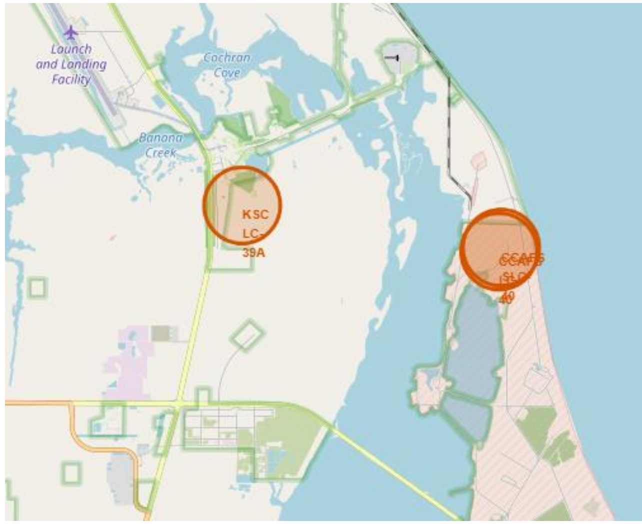
References:

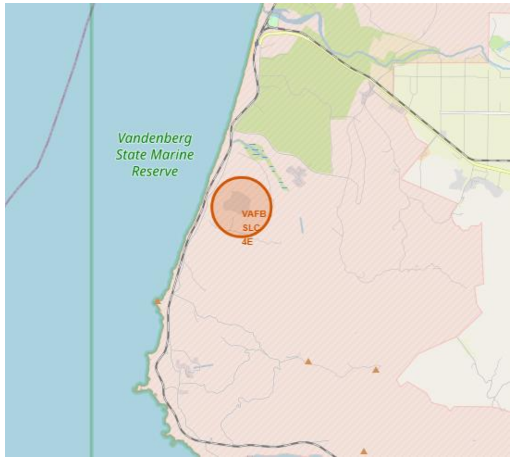
Section 3

Launch Sites Proximities Analysis

Space X Falcon9 – All Launch Sites

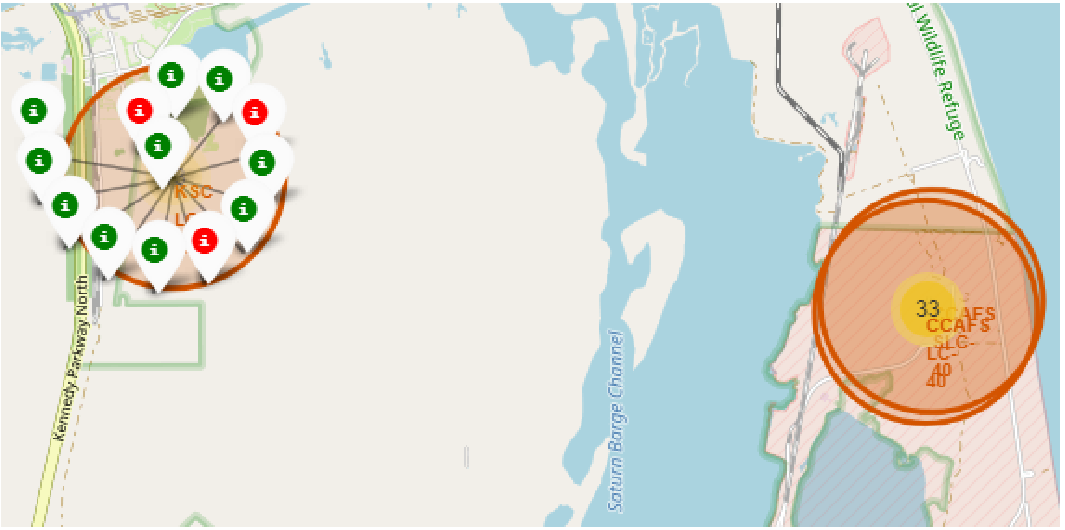


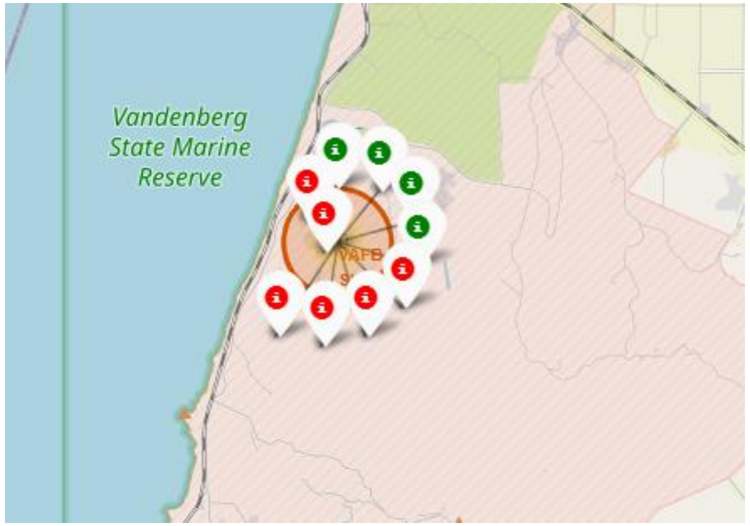




- In the top left figure, it clearly depicts the unique launch sites
- We can see that all launch sites are close to the ocean for less travel for the rocket, but instead testing for landing outcomes.
- The red circles depict the zone of the launch site

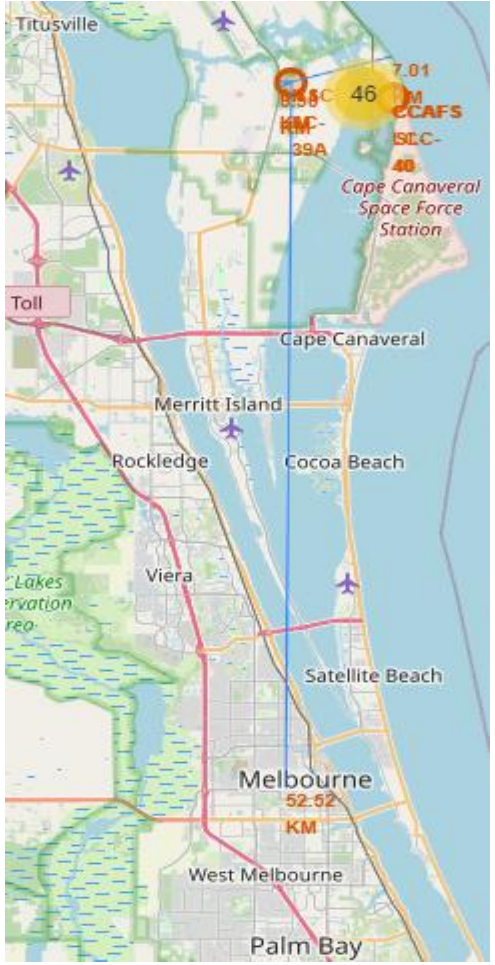
Launch Site Landing Outcomes - Map

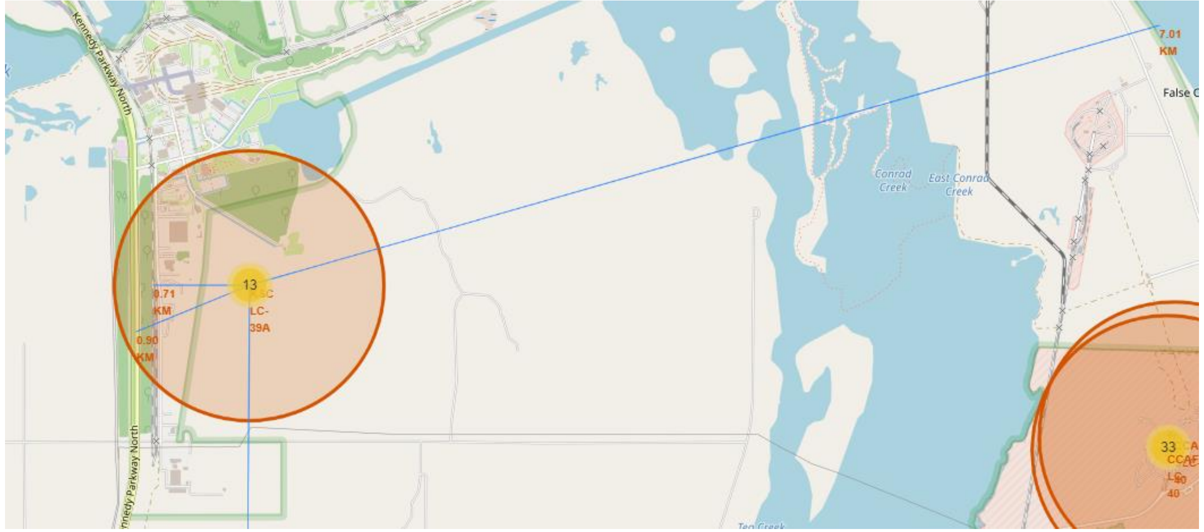




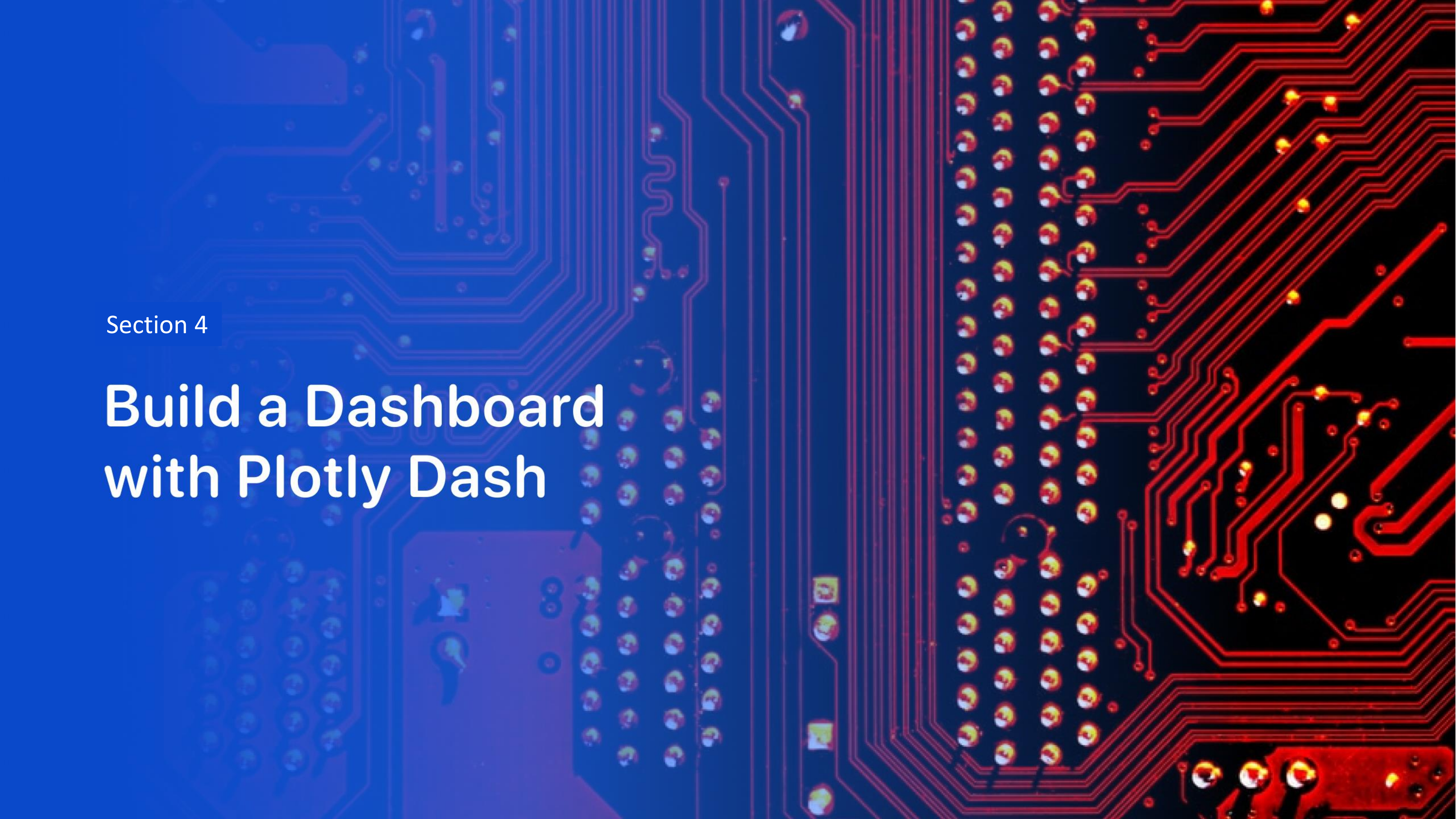
- The figure clearly depicts the landing outcomes from each site once selected.
- The green marker indicates that the landing was successful.
- The red marker indicates the landing was unsuccessful.
- We can see that KSC LC-39A launch site had more successful landing outcomes compared to VAFB SLC-4E launch site.

KSC LC-39A proximities to nearest city, railway, highway, & coastline





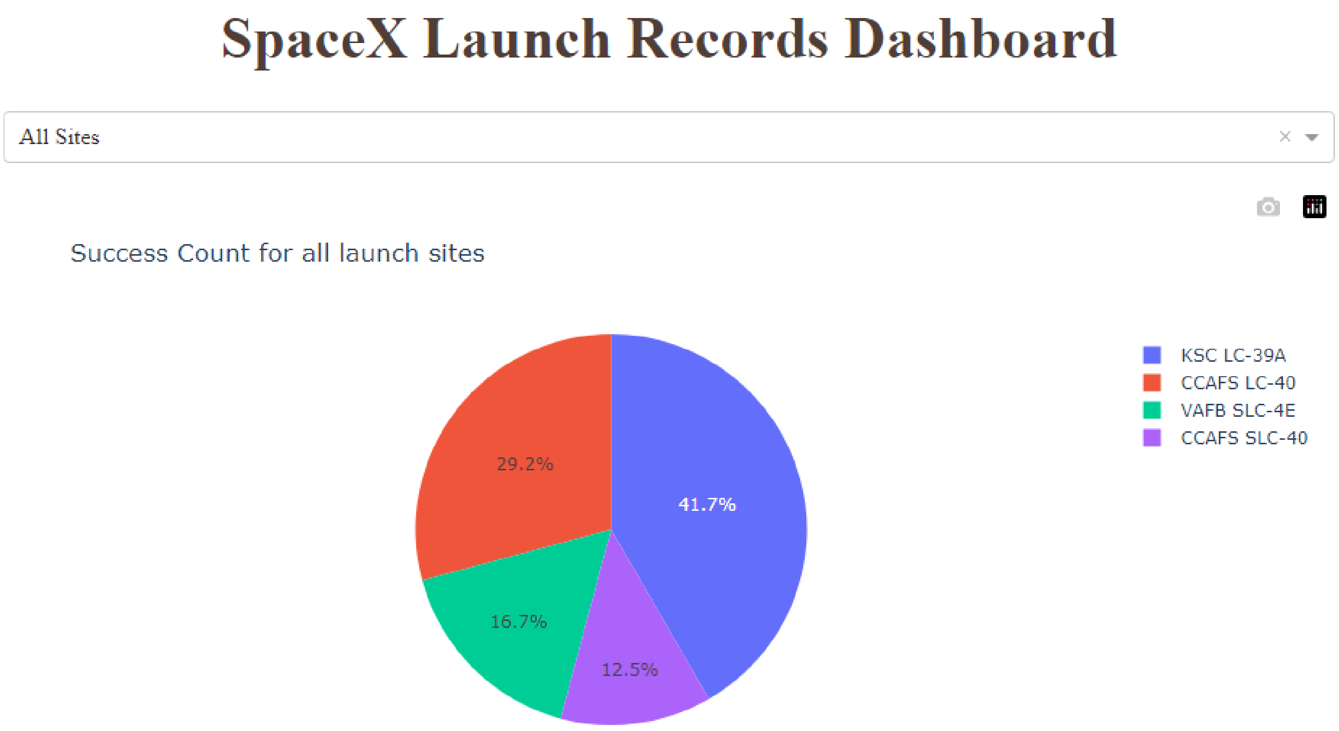
- The figure clearly depicts the distance of the launch site to the nearest coastline, railway, highway & city.
- It is evident that the launch site is close to the coastline to test successful landings.
- The launch site are away from the city, which could be due to safety issues, but still close to railway and highways.



Section 4

Build a Dashboard with Plotly Dash

Dashboard Launch Success for All Sites

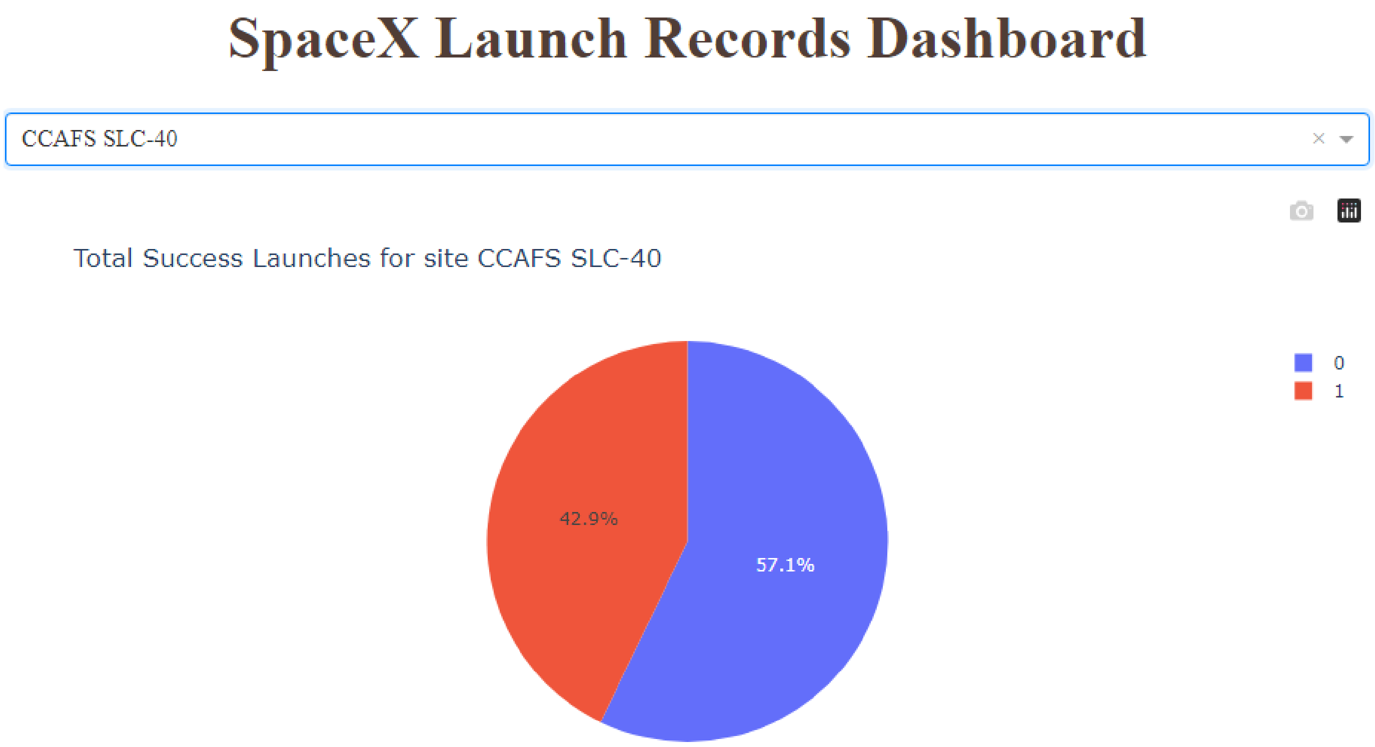


SpaceX Launch Records Dashboard

All Sites

- This figure portrays the SpaceX launch records for the Falcon9 for all launch sites.
- We can see that KSC LC-39A has the highest successful landings and CCAFS SLC-40 has the lowest successful landing outcomes.
- The place where launches occur seem to have an impact on the success of landing outcomes

Launch Site with Highest Launch Success Ratio



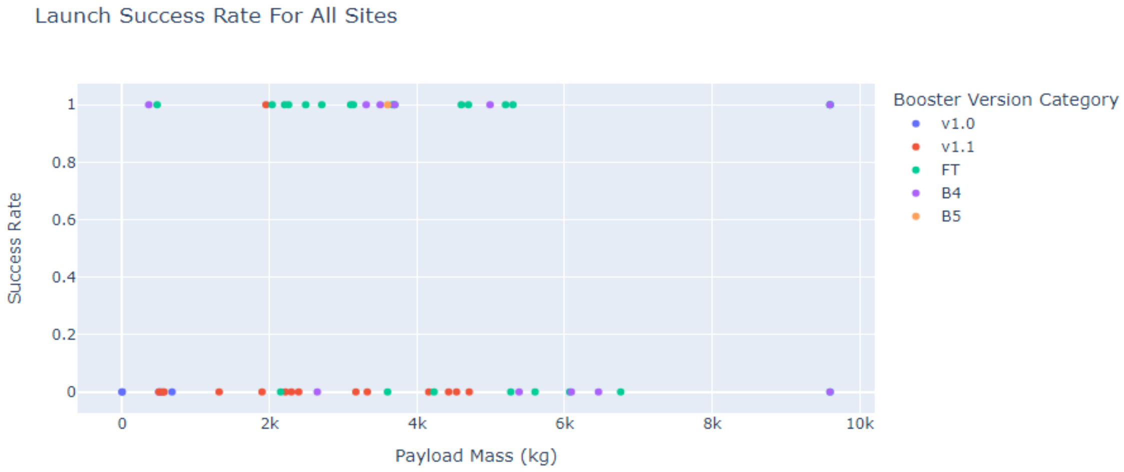
SpaceX Launch Records Dashboard

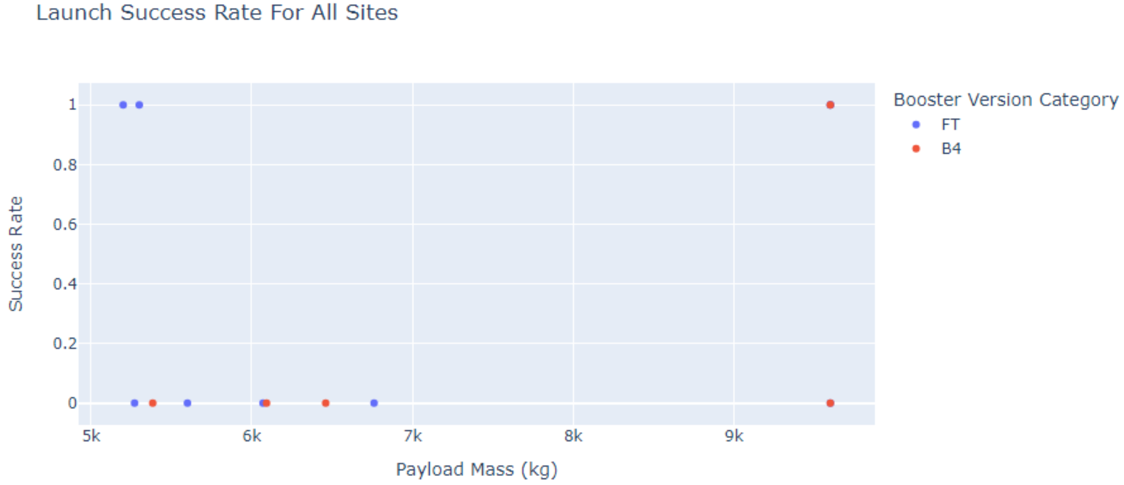
CCAFS SLC-40

Total Success Launches for site CCAFS SLC-40

- Despite having the lowest number of successful launches, CCAFS SLC-40 has the highest launch success ratio with a success rate of 42.9%.
- It is likely that in the future SpaceX will launch more rockets form this launch site as it has the highest success rate compared to other launch sites

Payload vs. Launch Outcomes for All Sites



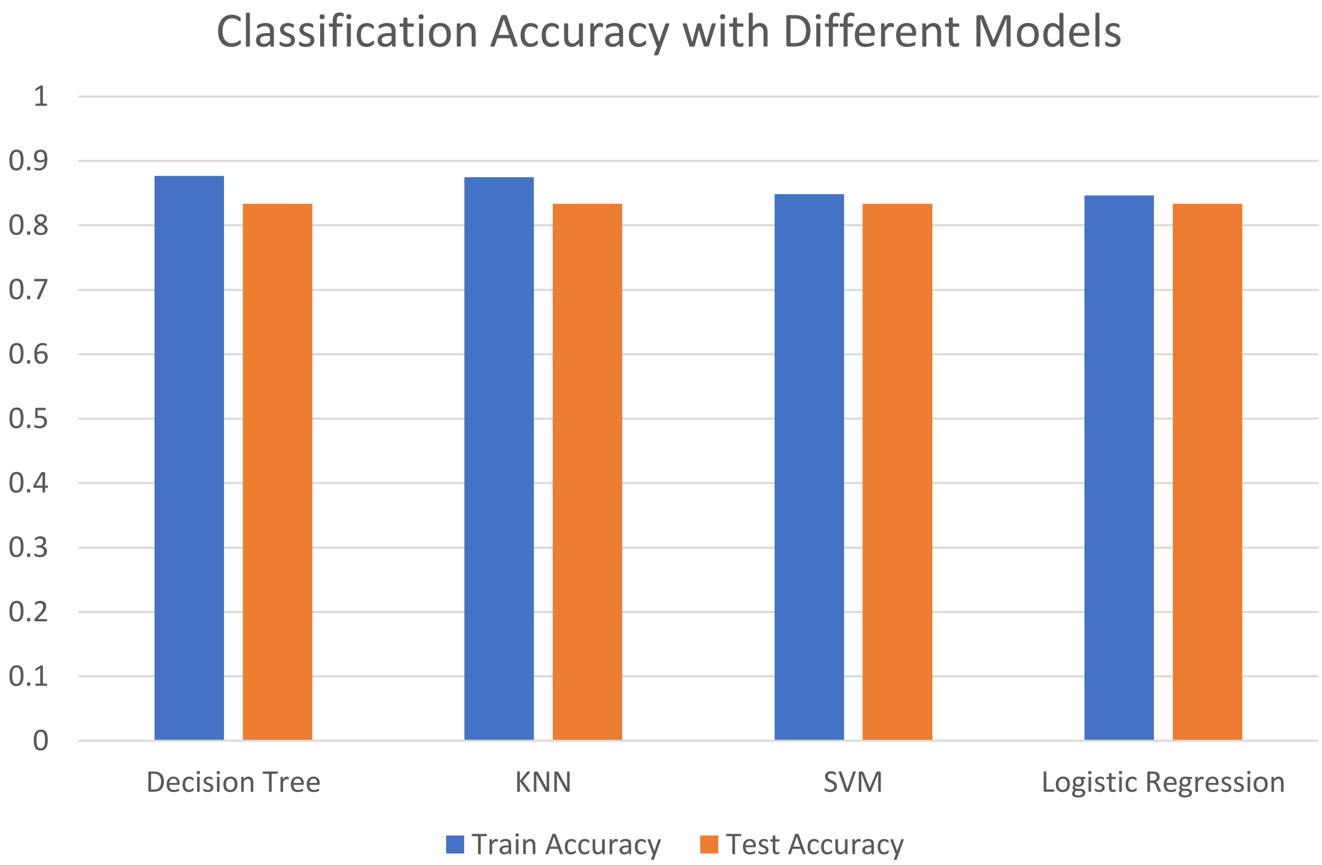


- The top figure depicts the launch success rate for different booster version carrying payload mass from 0kg – 10,000kg.
- The bottom figure depicts the launch success rate for different booster version carrying payload mass from 5,000kg – 10,000kg.
- We can see that a significant portion of the successful launches are for lighter payload mass (less than 5,000kg)
- The FT and B5 have successful landing outcomes with heavier payload mass.

Section 5

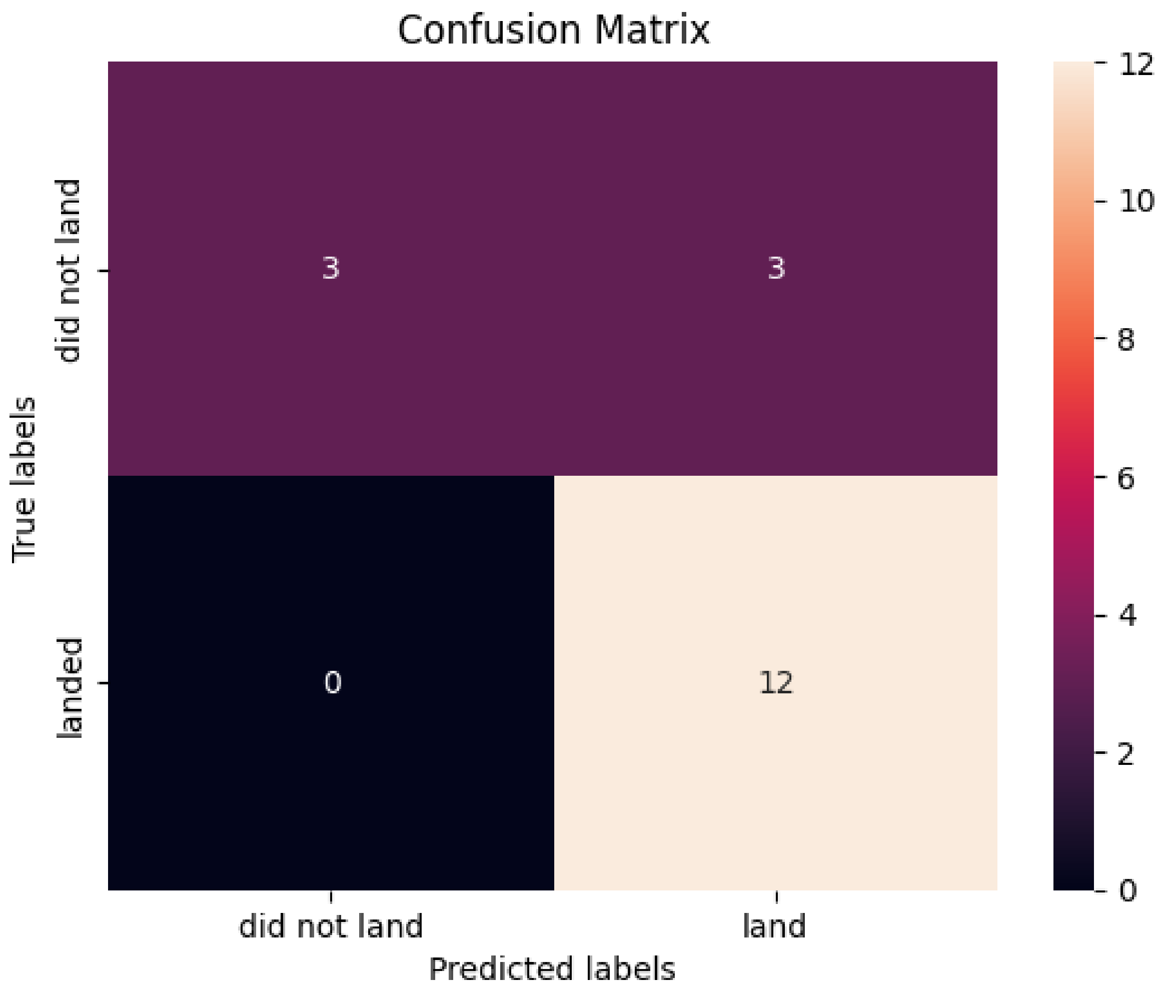
Predictive Analysis (Classification)

Classification Accuracy



- The figure depicts the classification accuracy with Decision Tree, KNN, SVM, & Logistic Regression Model
- The testing accuracy score on the test data is the same for all classification algorithms of 83.3%.
- This test accuracy could be due to the small test sample of only 18 launches, we may need a broader data set to fine tune the models
- Based on the training accuracy scores, the model with best score to predict the landing outcomes is the Decision Tree Classifier with a training accuracy of 87.7%

Decision Tree Classifier Confusion Matrix



- From the confusion matrix, we can see that the model made 18 predictions.
- 12 instances were forecasted to result in successful landings, and indeed, they did culminate in successful landings (True positive)
- In 3 instances (top left), predictions indicated failed landing, and these predictions aligned with the absence of successful landings (True negative).
- In 3 cases (top right), predictions suggested successful landings, yet these landings did not come to fruition as anticipated (False positive).

Conclusions

- With the progression of SpaceX's flight testing over time, there was a corresponding rise in the likelihood of the first stage successfully landing.
- The highest number of successful landings came from KSC LC-39A, but the highest proportion of successful landings was at CCAFS SLC-40.
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rate.
- The Decision Tree Classifier seems to be the best classification model to predict first stage landing outcomes. However, more data may be required to fine tune the model as the test accuracy was the same for all models.

Appendix

- To avoid a constant warning message that may crash the notebook in predictive analysis for decision tree classifier, it is crucial to run the below:
  - def warn(*args, **kwargs):
  -     pass
  - import warnings
  - warnings.warn = warn
- To avoid data type conflicts, it is crucial to check all data types in data collection. Specifically, the “customer” section in the for loop required altering.

Thank you!