IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Isai Garay
16/05/2025

Outline

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

Executive Summary

- Summary of methodologies

  It has been used the following methodologies: Data Collection via API and Web Scraping, Data Wrangling, Exploratory Data Analysis (EDA) using SQL, Pandas, Matplotlib, Folium and Dash, and Predictive Analysis using Machine Learning Classification Models.

- Summary of all results

  - Success Rate over Time: An improvement in success rate was observed over time.
  - Success Rate by Launch Site: KSC LC-39A shows the highest success rate of all launch sites.
  - Success Rate by Orbit: ES-L1, SSO, HEO, and GEO were observed to have the highest success rates.
  - Payload: Heavier payloads have a high failure rate early on with a significant improvement over time.
  - Predictive Analysis: The DecisionTreeClassifier algorithm has proven highly accurate in predicting landing outcomes.

Introduction

- Project background and context

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

- Problems you want to find answers

  What factors determine if the rocket will land successfully?

  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Retrieval and consolidation from multiple SpaceX API endpoints
  - Web scaping tabular data from Wikipedia
- Perform data wrangling
  - Extracted relevant records
  - Flattened fields and resolved missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Visualize variable relationships
  - Look at the data in aggregate
- Perform interactive visual analytics using Folium and Plotly Dash
  - Mark all launch sites on a map
  - Mark successful and failed launches
  - Calculate distances to proximate locations
  - Provide for interactive exploration of the data
- Perform predictive analysis using classification models
  - Build, evaluate, and compare several predictive classification models

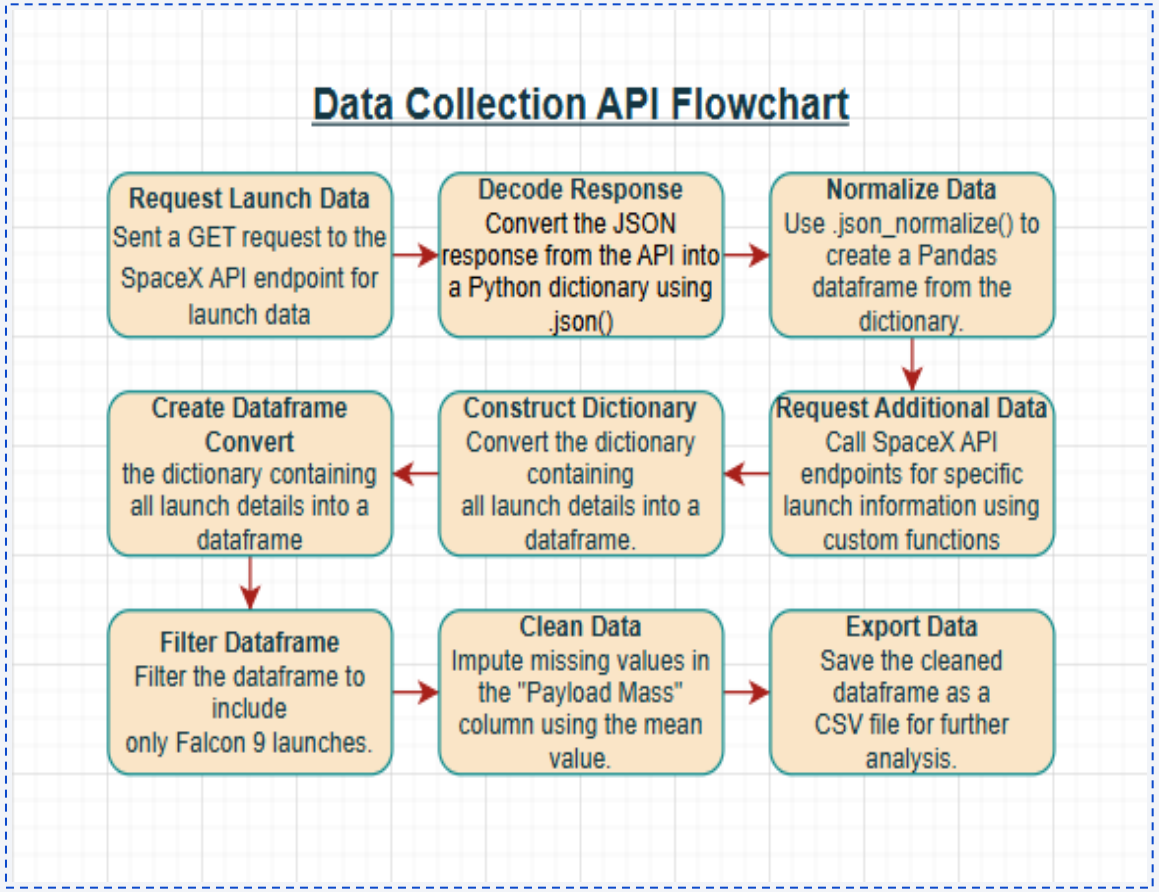
Data Collection

- **Medhods used:**

- Data collection was done using get request to the SpaceX API.
- Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
- Then data was cleaned, checked for missing values and fill in missing values where necessary.
- In addition, web scraping was applied from Wikipedia for Falcon 9 launch records with BeautifulSoup.
- The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

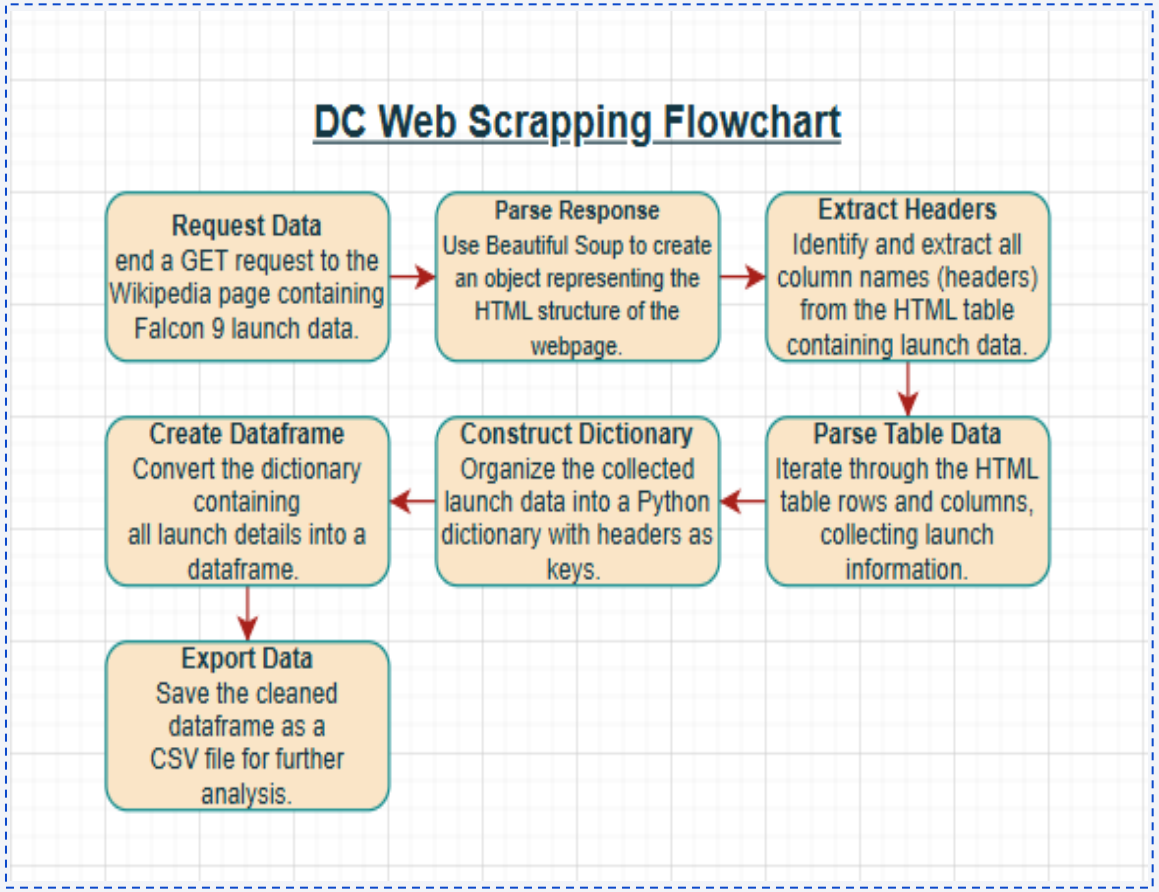
Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/Paul1711/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scrapping was used to extract Falcon 9 launch records with BeautifulSoup
- The table was parsed and converted it into a pandas dataframe.
- GitHub URL of the completed web scraping notebook: https://github.com/Paul1711/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping_Completed.ipynb

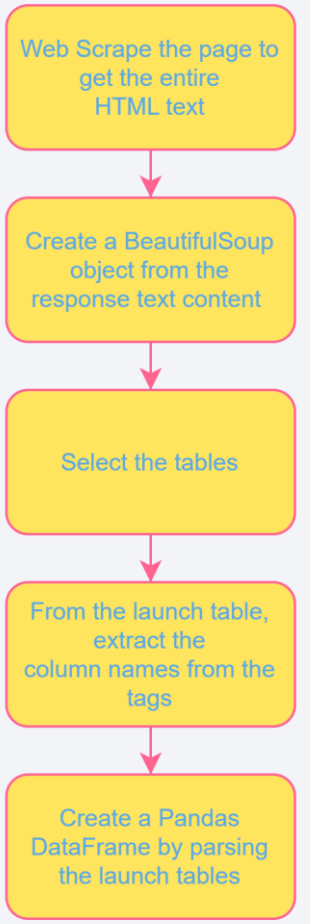


Data Wrangling

- SpaceX launch data was scraped from HTML tables on a permanently-linked copy of the SpaceX Wikipedia webpage (https://en.wikipedia.org/wiki/SpaceX).
- Launch data was extracted from these tables and loaded into a Pandas DataFrame for further analysis.
- GitHub URL (Web Scraping):

  https://github.com/Paul1711/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling_Completed.ipynb

Data Wrangling Flowchart



EDA with Data Visualization

- The following charts were created to look at Launch Site trends
  - Scatterplot to see mission outcome relationship split by Launch Site and Flight Number.
  - Scatterplot to see mission outcome relationship split by Launch Site and Payload.
- The following charts were created to look at Orbit Type trends
  - Bar chart to see mission outcome relationship with Orbit Type.
  - Scatterplot to see mission outcome relationship split by Orbit Type and Flight Number.
  - Scatterplot to see mission outcome relationship split by Orbit Type and Payload.
- The following chart was created to look at trends based on time
  - Line plot to see mission outcome trend by year.
- GitHub URL (EDA with Data Visualization): https://github.com/Paul1711/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-datav.ipynb

EDA with SQL

- SQL queries were written to extract information about:
  - Launch sites
  - Payload masses
  - Dates
  - Booster types
  - Mission outcomes
- GitHub URL (EDA with SQL):
  https://github.com/Paul1711/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite_Completed.ipynb

Build an Interactive Map with Folium

- Map objects were created and added to the Folium map

  - Markers were added for launch sites and for the NASA Johnson Space Center

  - Circles were added for the launch sites.

  - Lines were added to show the distance to the nearby features:

  - Distance from CCAFS LC-40 to the coastline:
    - Distance from CCAFS LC-40 to the rail line
    - Distance from CCAFS LC-40 to the perimeter road

- GitHub URL Folium map:
  https://github.com/Paul1711/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

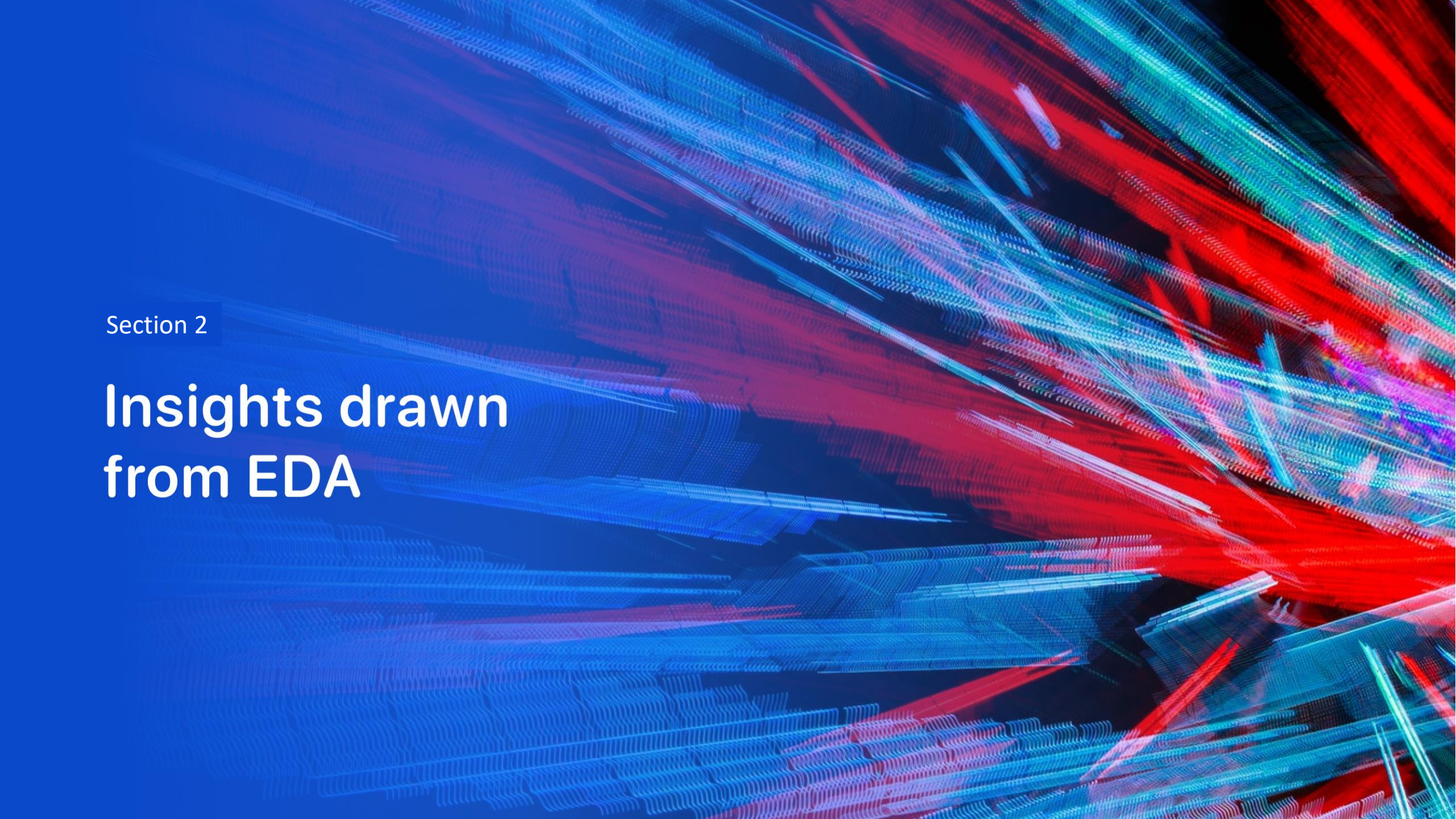
- The Plotly Dash dashboard included a dropdown input to select data from ‘one’ or ‘all’ launch sites to display on the pie chart and scatterplot.
- For ‘one’ launch site, the pie chart displayed the distribution of successful and failed Falcon 9 first stage landings for that site.
- For ‘all’ launch sites, the pie chart displayed the distribution of successful Falcon 9 first stage landings between the sites.
- The input slider is used to filter the payload masses for the scatterplot.
- The scatterplot displayed the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category.
- GitHub URL (Dashboard File): https://github.com/Paul1711/Applied-Data-Science-Capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- The dataset was split into training and testing sets.
- The following machine learning models were trained on
  - the training data set:
  - Logistic Regression
  - SVM (Support Vector Machine)
  - Decision Tree
  - KNN (k-Nearest Neighbors)
- Hyper-parameters were evaluated using GridSearchCV() and the best was selected using the best_params method.
- Using the best hyper-parameters, each of the four models were scored on accuracy by using the testing data set.
- GitHub URL (Machine Learning):
  https://github.com/Paul1711/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

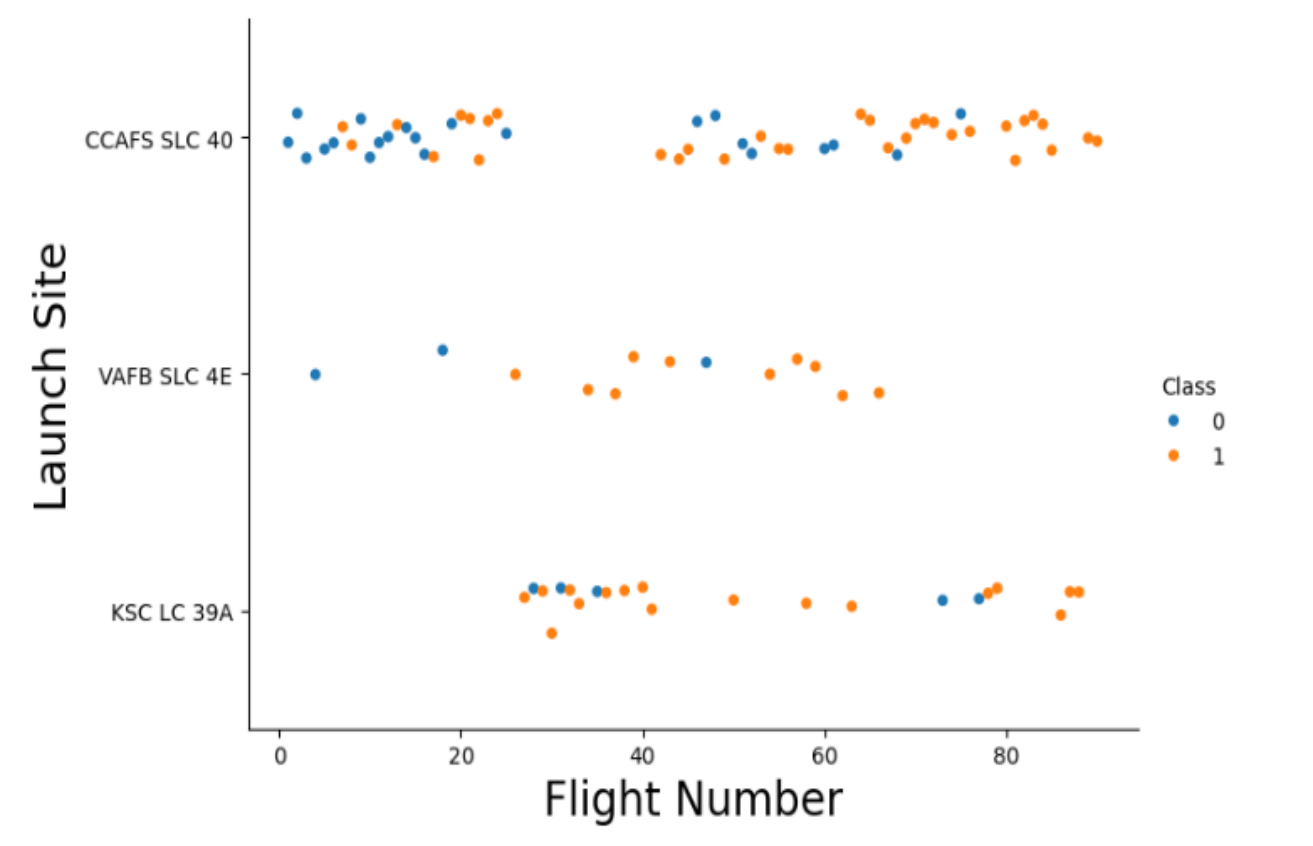
Results

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


Section 2
Insights drawn from EDA

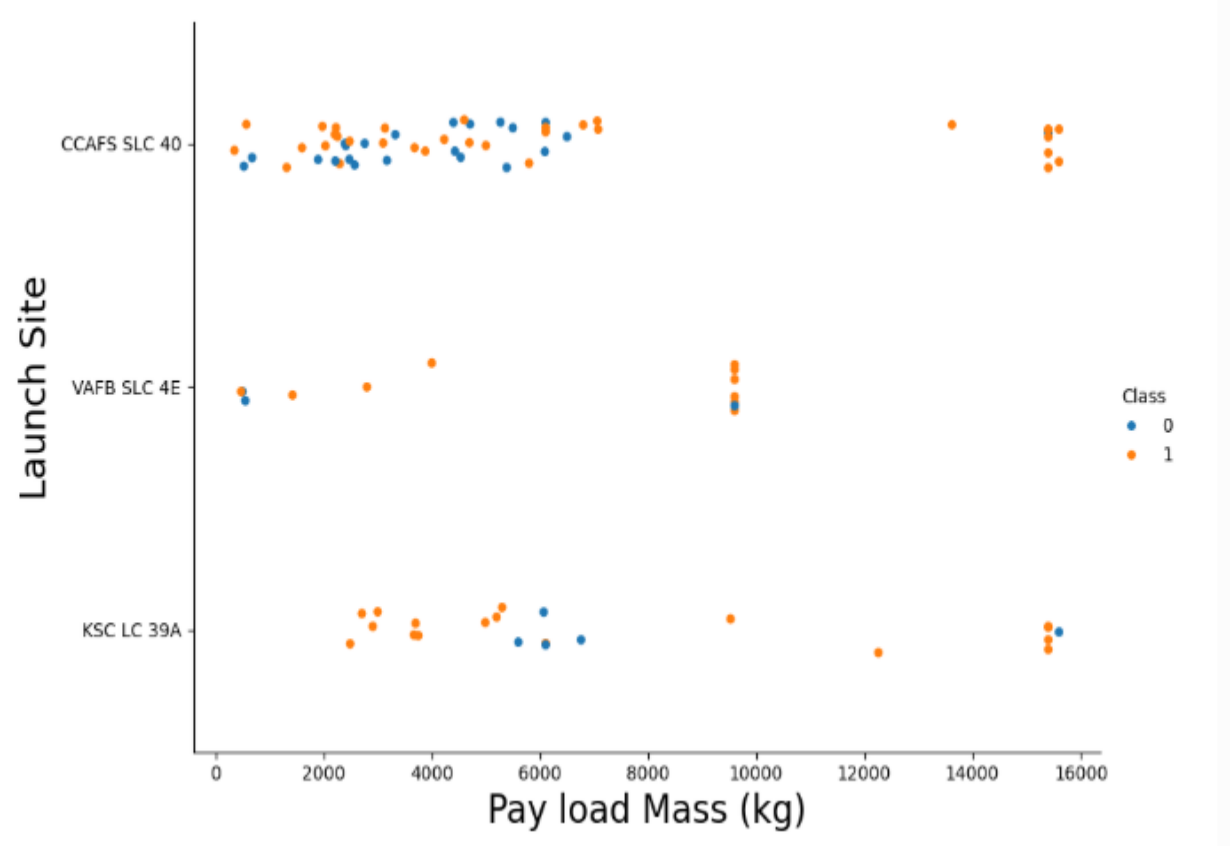
Flight Number vs. Launch Site

- Early launches struggled, recent ones thrived. CCAFS launches most, but VAFB & KSC
- have higher success. This suggests SpaceX is learning and improving over time.



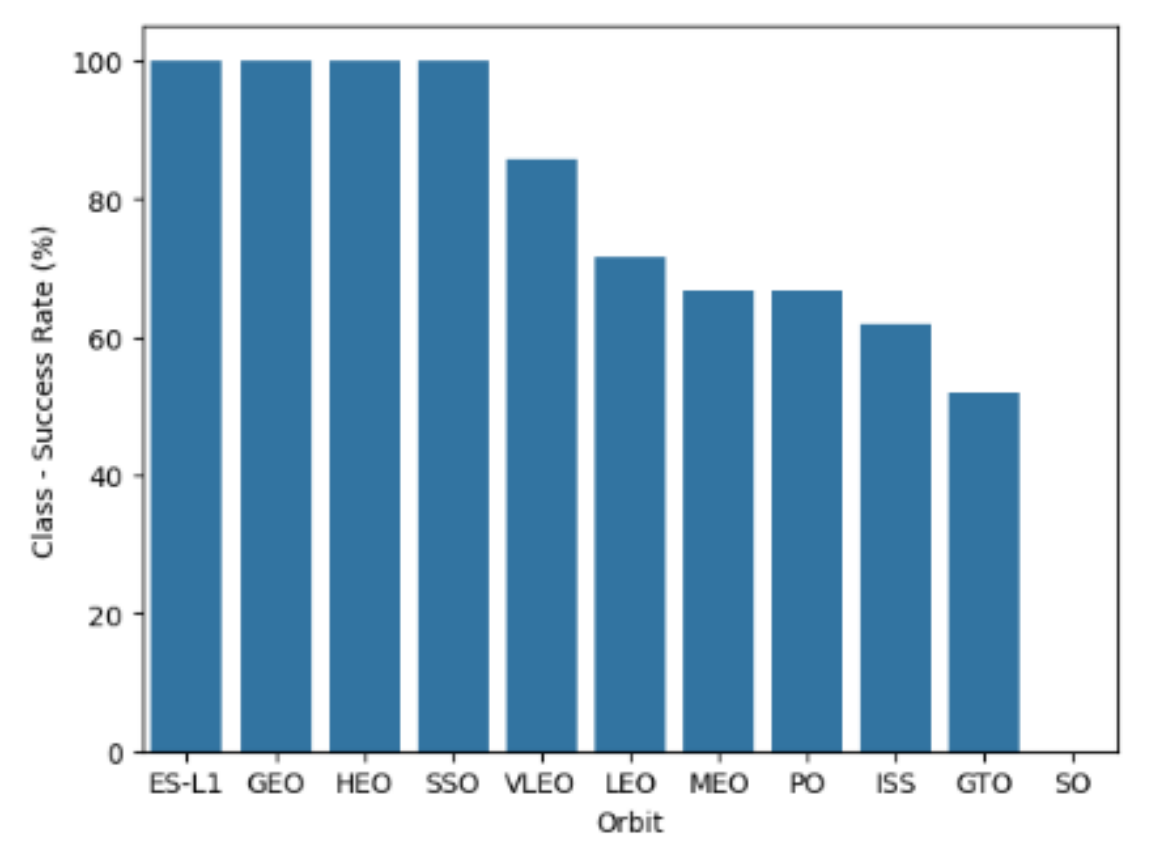
Payload vs. Launch Site

- Early launches & lighter payloads struggled, but SpaceX is learning! Heavier payloads see higher success across sites, with KSC excelling for lighter ones.



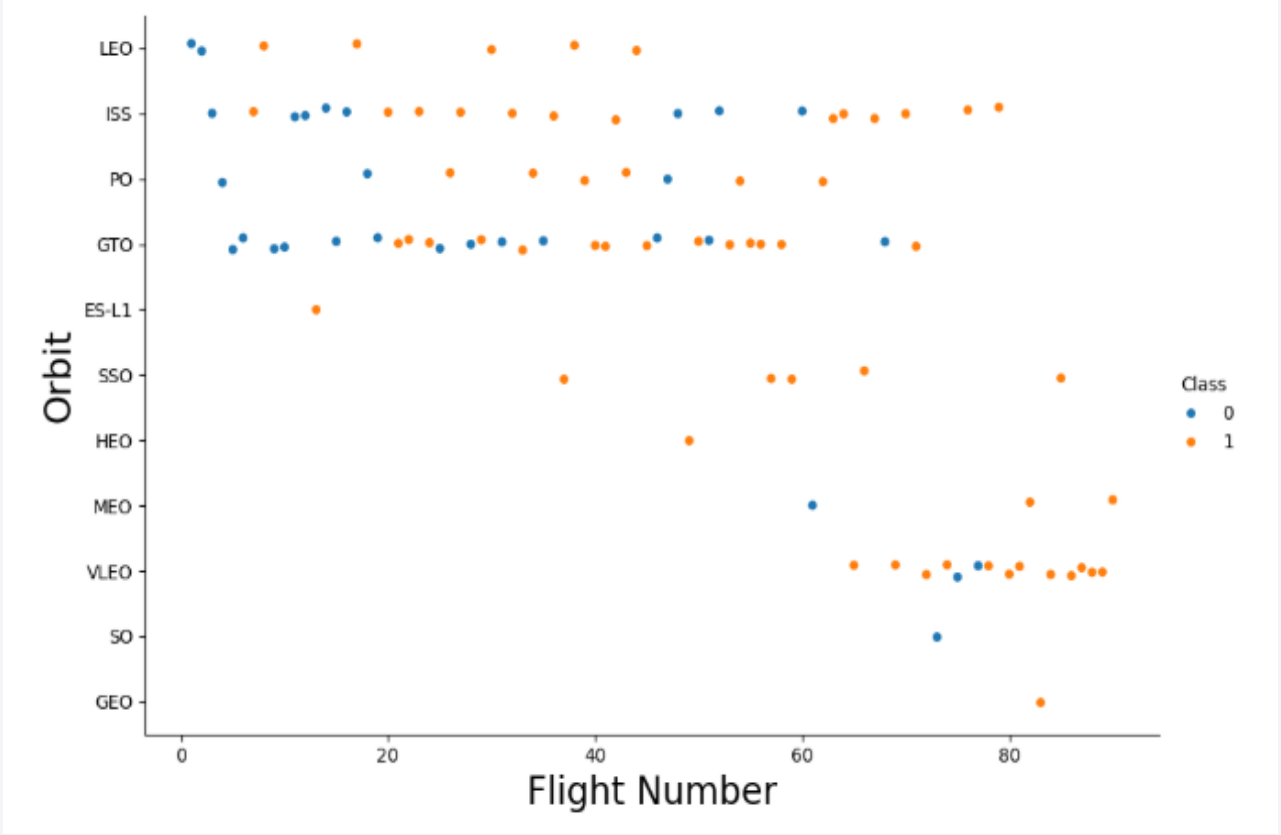
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO and GEO orbits have no failed first stage landings.
- SO orbits have no successful first stage landings.



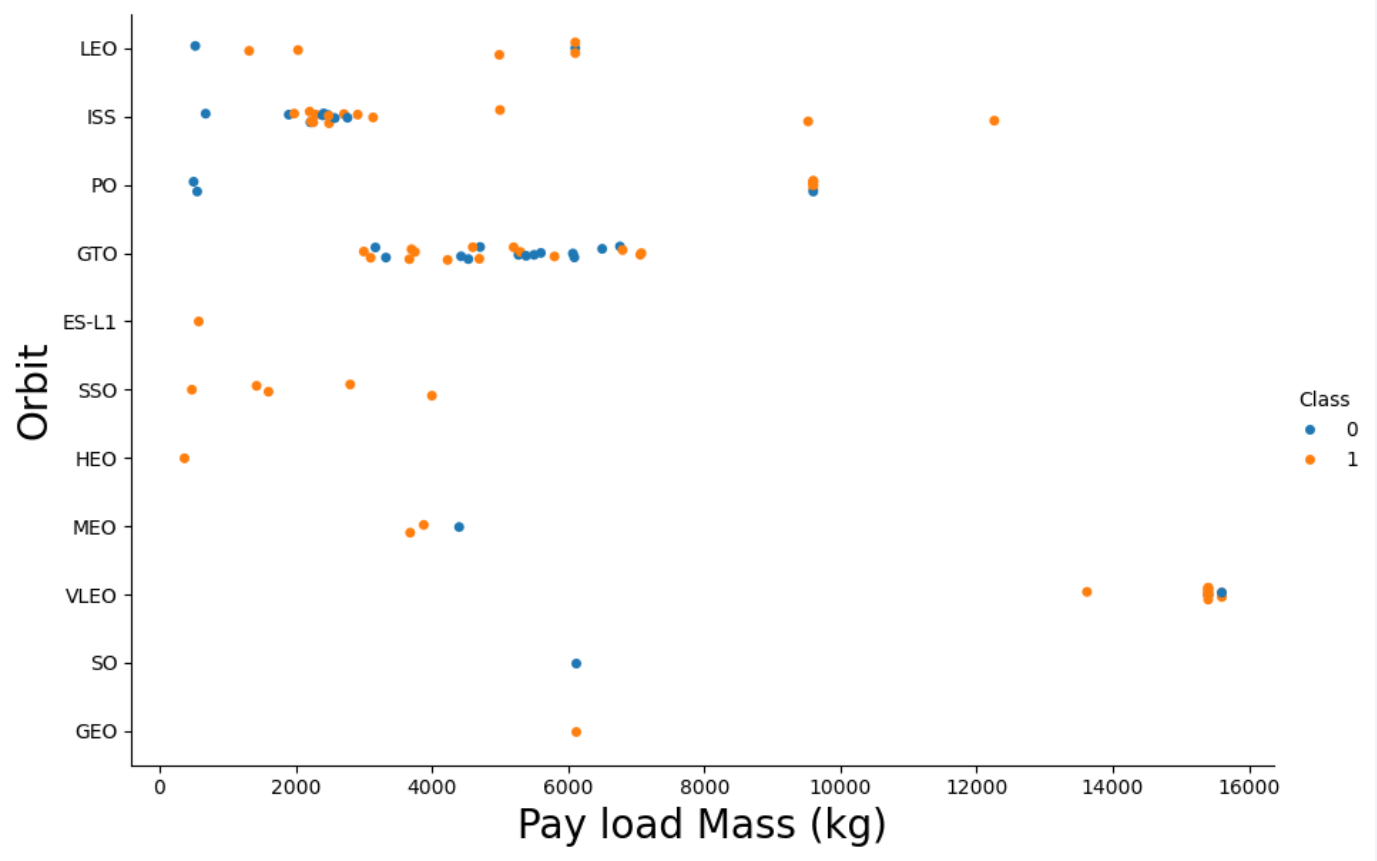
Flight Number vs. Orbit Type

According to the scatter plot in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



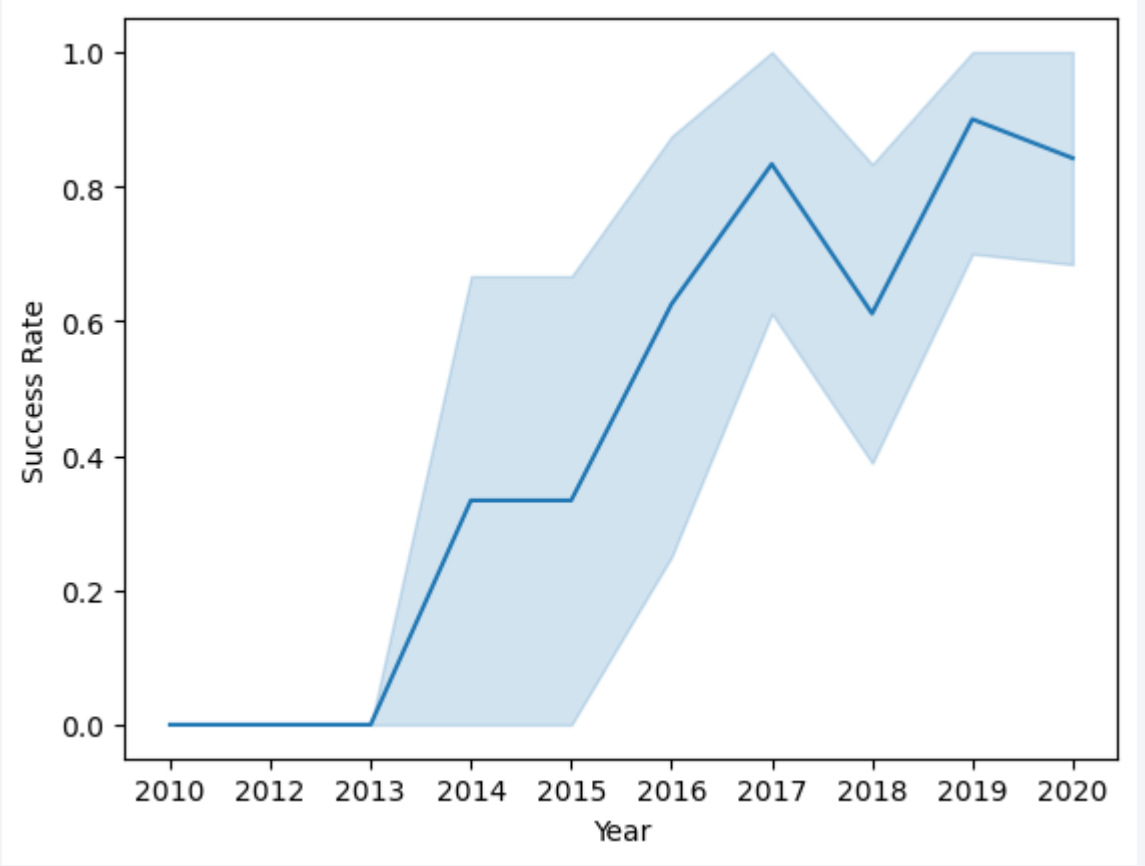
Payload vs. Orbit Type

- Some orbit types showed higher success rates than others.
- Success rate appeared to have no obvious correlation with payload mass.



Launch Success Yearly Trend

- According to the line chart the success rate kept increasing since 2013 to 2020, although there was a decreasing in 2018.



All Launch Site Names

- Distinct method is used to find and show unique Launch Site values from SpaceX data.
- Resulting in four distinct values

```
In [11]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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

Launch Site Names Begin with 'CCA'

- The following code was used to find 5 records where launch sites begin with ‘CCA’

```
In [12]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_site LIKE '%CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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

- Calculating the total payload carried by boosters from NASA
- In order to do that, 'Where' method was used to select only NASA (CRS) in Customer column. The result was 45,596 kg.

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

```
In [13]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

Out[13]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Where and Like method were used to find this number, the result was 2534.6666 kg

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

```
In [14]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE '%F9 v1.1%';
 * sqlite:///my_data1.db
Done.
Out[14]: AVG(PAYLOAD_MASS__KG_)
         2534.6666666666665
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- WHERE and MIN were used to find it, the result was 2015-12-22

```
In [16]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)";
 * sqlite:///my_data1.db
Done.
Out[16]: MIN(Date)
         2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

In [24]:
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%drone ship%' AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

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

Out[24]:

Booster_Version
F9 v1.1 B1012
F9 v1.1 B1015
F9 v1.1 B1018
F9 v1.1 B1017

Total Number of Successful and Failure Mission Outcomes

- Calculating the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

In [20]:
```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS COUNT FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

Out[20]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- A subquery method and MAX command was used to find this list.

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

In [27]:
```
%sql SELECT Booster_Version, (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) AS Max_payload_mass FROM SPACEXTABLE;
```

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

Out[27]:

Booster_Version	Max_payload_mass
F9 v1.0 B0003	15600
F9 v1.0 B0004	15600
F9 v1.0 B0005	15600
F9 v1.0 B0006	15600
F9 v1.0 B0007	15600
F9 v1.1 B1003	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1 B1011	15600
F9 v1.1 B1010	15600
F9 v1.1 B1012	15600

2015 Launch Records

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

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [42]:
```
%sql SELECT SUBSTR(Date, 6, 2), Mission_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE SUBSTR(Date, 0, 5) = '
```

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

Out[42]:

SUBSTR(Date, 6, 2)	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

- 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

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.

In [51]:
```
%sql SELECT Landing_Outcome, COUNT(*) AS count_outcomes FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY count_outcomes DESC;
```

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

Out[51]:

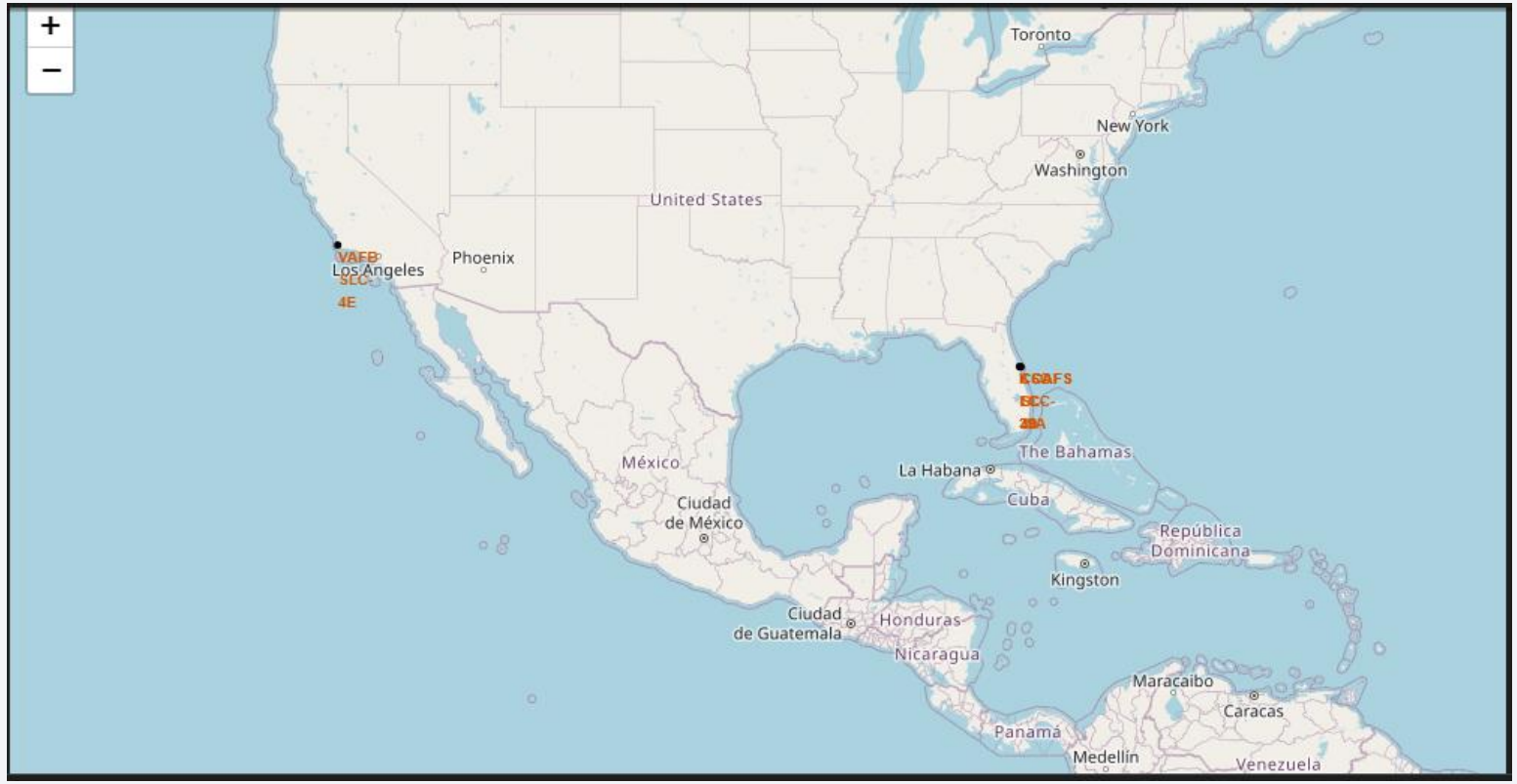
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All Launch Sites' Location Markers On A Global Map

- SpaceX launch sites are located in United States coasts. Florida and California.

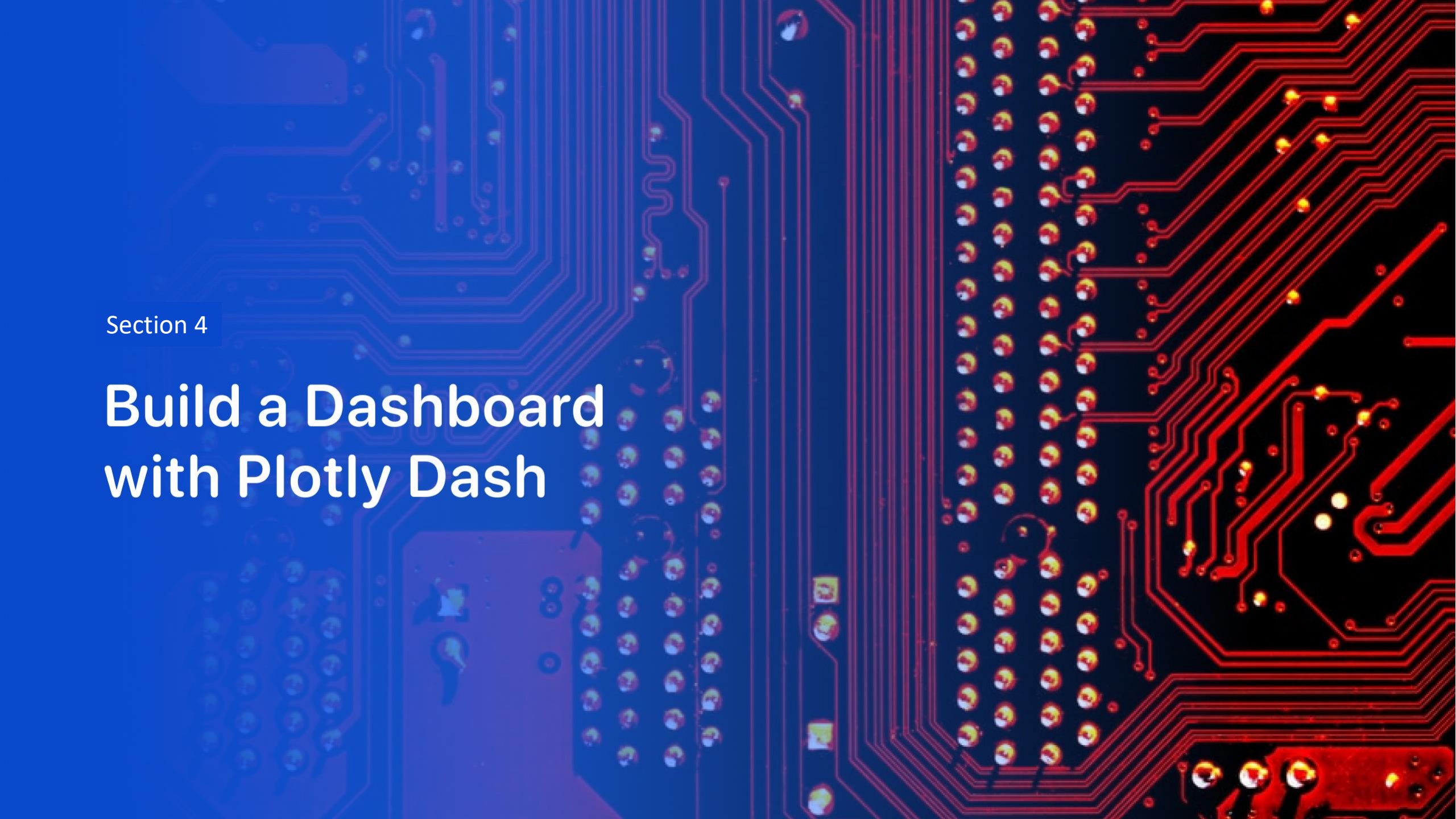


Map Markers of Success/Failed Landings

- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.

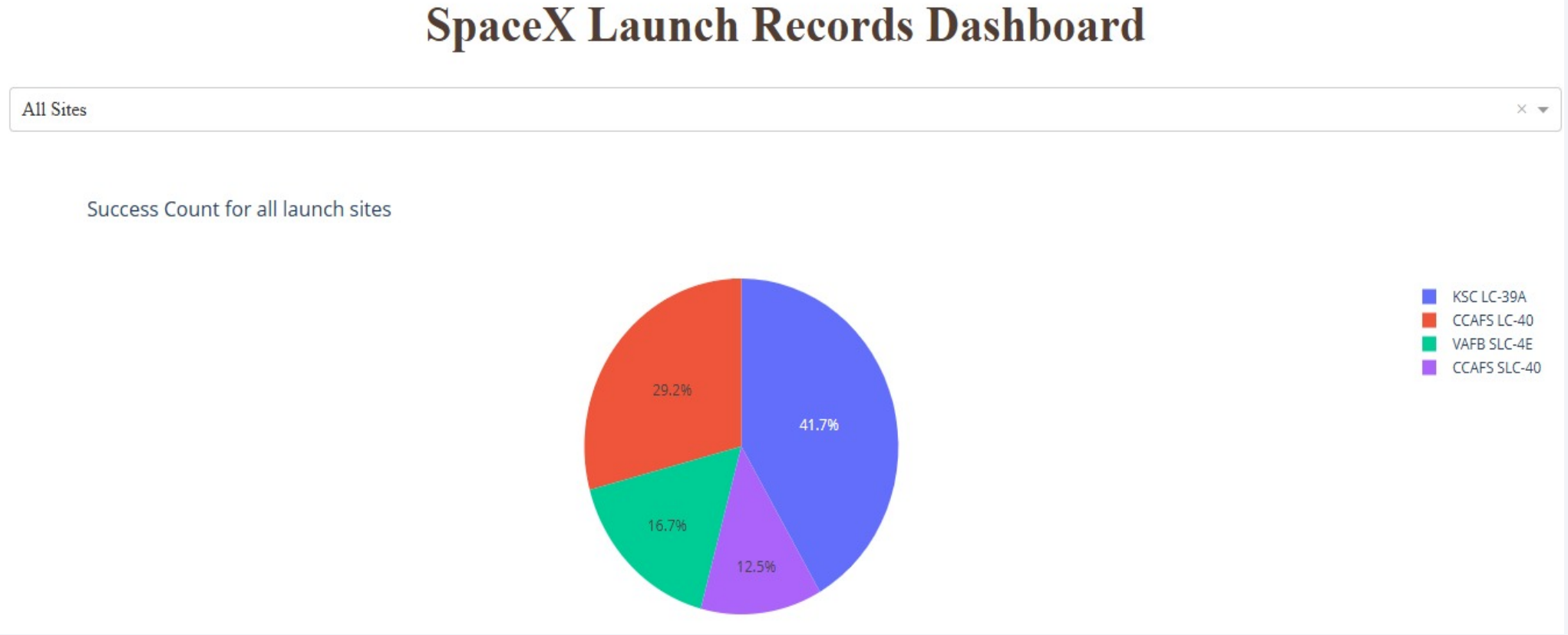
Launch Site distance to landmarks

- A mark with distance is displayed for the nearest city, railway, etc to CCAFS LC-40 which is Samuel C Phillips Railway that is 0.58 km away.


Section 4
Build a Dashboard with Plotly Dash

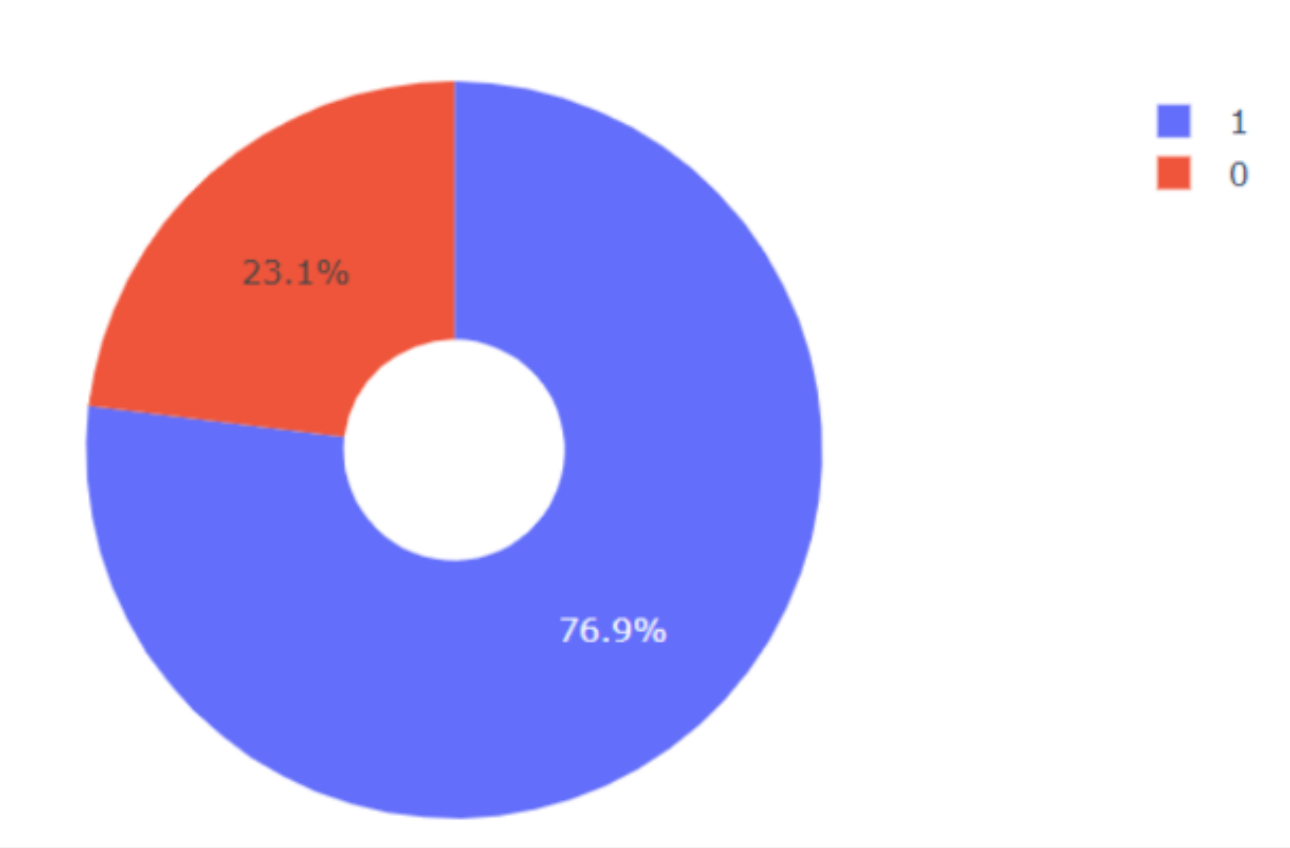
Total Success Launches By Site

- The dropdown menu allowed the selection of one or all launch sites.
- With all launch sites selected, the pie chart displayed the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites.
- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A



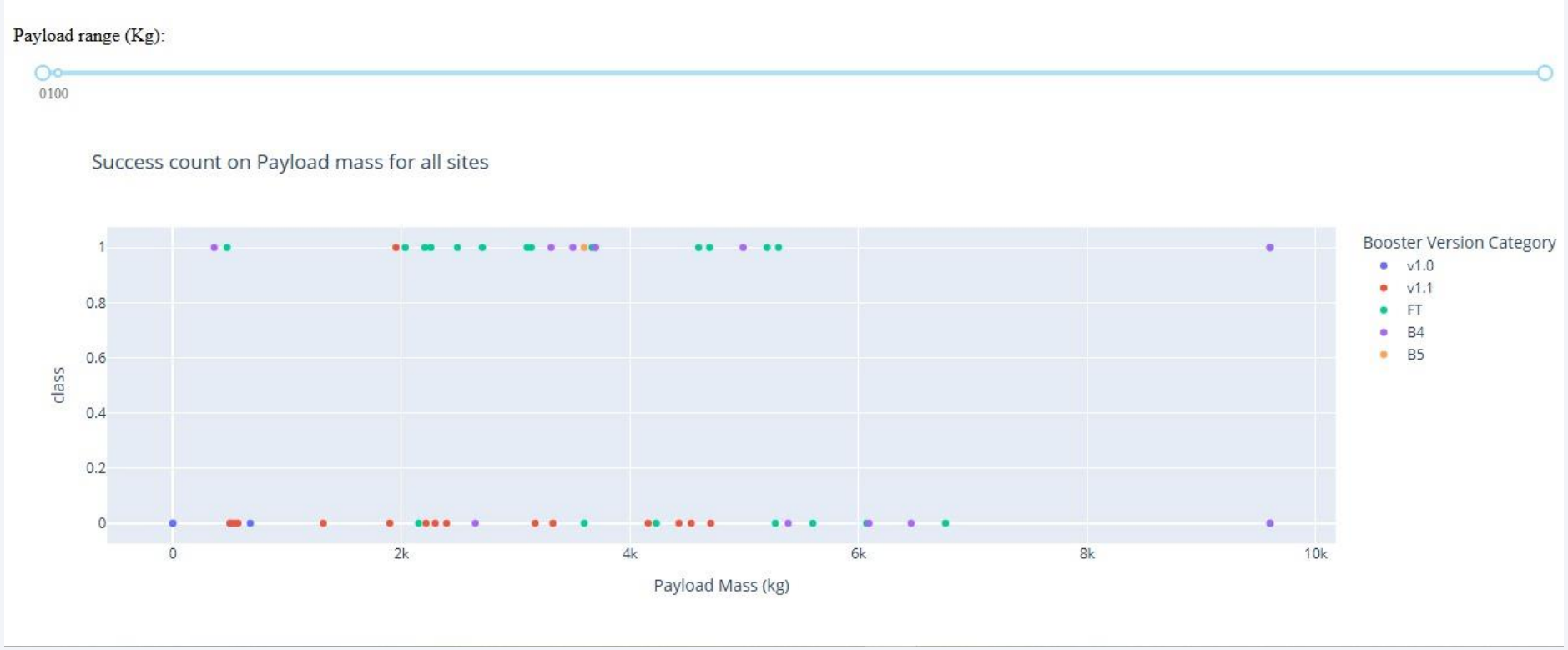
Launch Site with Highest Launch Success Ratio

- Falcon 9 first stage failed landings are indicated by the ‘0’ Class (blue) and successful landings by the ‘1’ Class (red).
- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%)



Payload vs. Launch Outcome For Different Payloads

- The scatter plot represents launch outcome for payload mass between 0 to 10000 kg for various Booster Version Category.



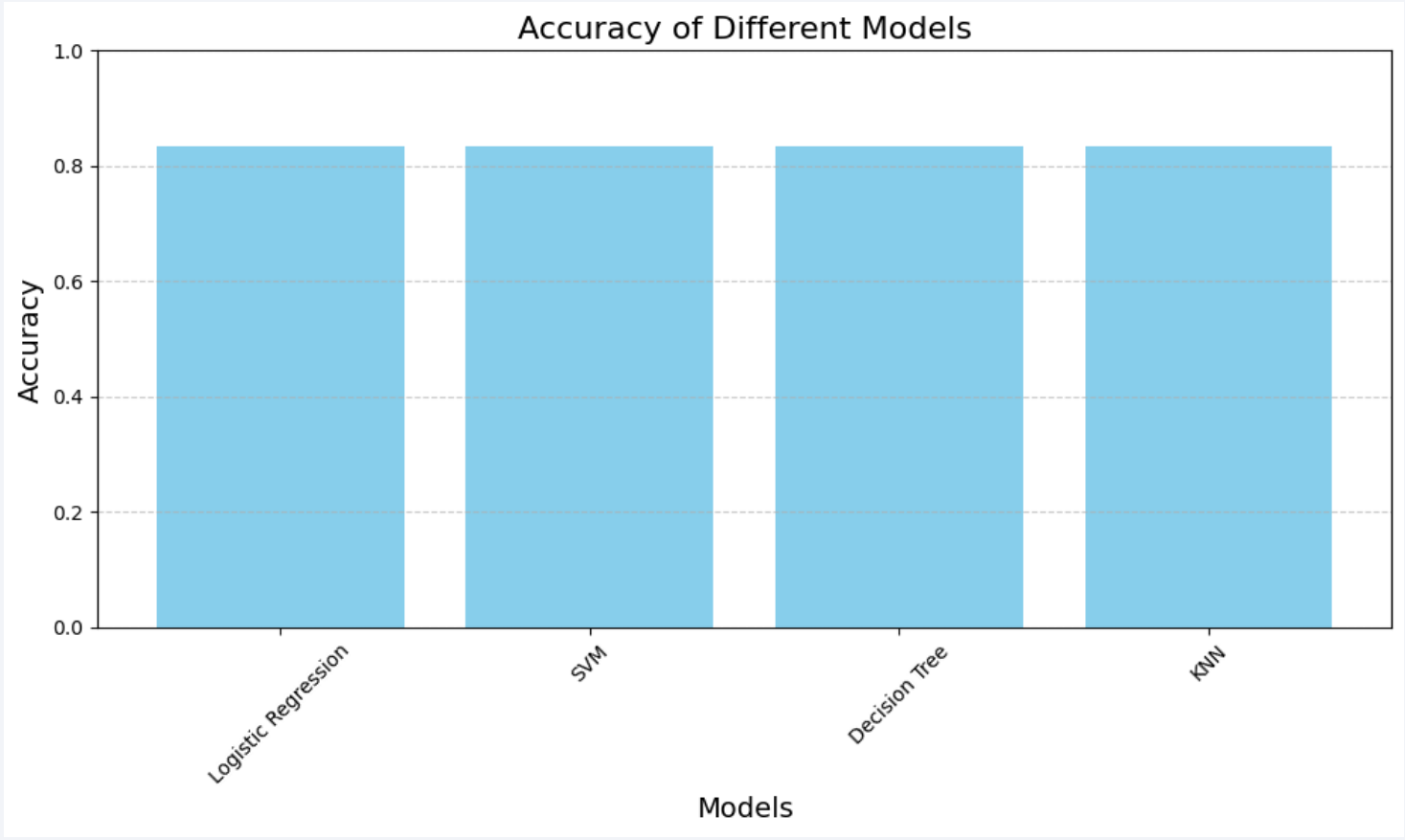
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four models where built (Logistic Regression, SVM, Decision Tree and KNN). Looking at their accuracy rate all of them have equal values, meaning any of them can be use for this prediction analysis.

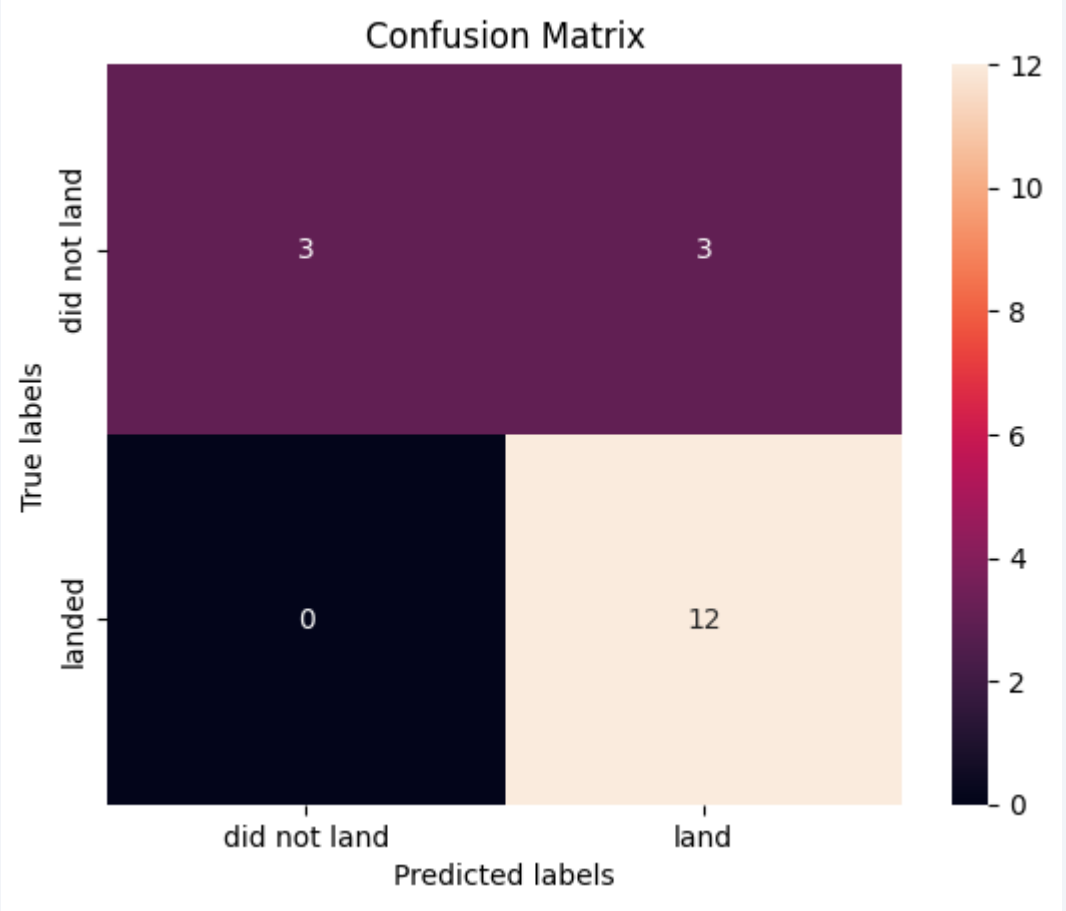
```
In [33]: print('LR Accuracy:', '{:.2%}'.format(logreg_accuracy))
         print( 'SVM Accuracy:', '{:.2%}'.format(svm_accuracy))
         print('Decision Tree Accuracy:', '{:.2%}'.format(tree_accuracy))
         print('KNN Accuracy:', '{:.2%}'.format(knn_accuracy))

LR Accuracy: 83.33%
SVM Accuracy: 83.33%
Decision Tree Accuracy: 83.33%
KNN Accuracy: 83.33%
```



Confusion Matrix

- Similarly to the accuracy rate, The confusion matrix have the same distribution for all four models.

- Prediction Breakdown:
  - 12 True Positives and 3 True Negatives
  - 3 False Positives and 0 False Negatives



Conclusions

- SpaceX's record for Falcon 9 first stage landing outcomes has improved.
- The trend is toward to a better performance and greater success as more launches are made.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- Any of all four machine learning models (LR, SVM, DT and KNN) can be used to predict future SpaceX Falcon 9 first stage landing outcomes as their accuracy rate and matrix confusion are equal.

Appendix

- Sources:

  - SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
  - Wikipedia (Webpage): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - SpaceX (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321ENSkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=0000026UJ&utm_term=10006555&utm_id=NA-SkillsNetworkChannel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
  - Launch Geo (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
  - Launch Dash (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

- Jupyter Notebooks and Plotly Dashboard File:

  - https://github.com/Paul1711/Applied-Data-Science-Capstone/tree/main

Thank you!