Winning Space Race with Data Science

Sergio Vázquez Sánchez
07/13/2023

Outline

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

Executive Summary

- **Summary of methodologies**

Machine learning methods allow for the modeling of structures and associations in data in order to derive linkages and make predictions based on previously unknown occurrences. Unsupervised learning (learning from labeled data) and supervised learning (learning from unlabeled data) are two types of machine learning methodologies (discovering hidden patterns in data or extracting features).

- **Summary of all results**

The expected results is the prediction if the launch will be successful an d the rocket can land in the base to reuse it

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. In this lab, you will create a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.

- **Problems you want to find answers**

- What factor determine if the rocket will land successfully?
- The interaction among various features that determine the success rate of landing
- What are the optimal conditions to ensure a successful landing operation?

Section 1

Methodology

Methodology

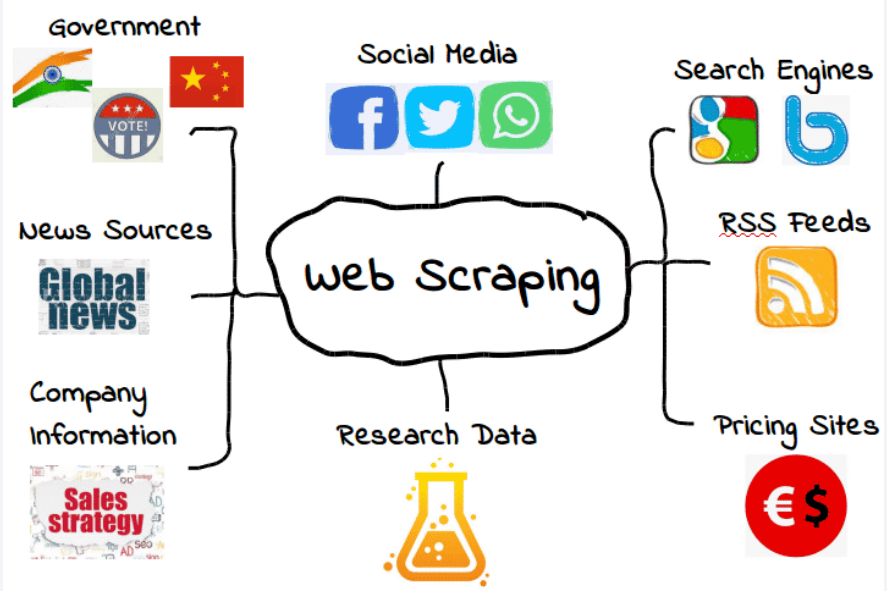
Executive Summary

- Data collection methodology:
  - We use the official data available and provided by Space X. In this data set we can fin some features about the performance and factor involved in the rocket launch like ubication and weather conditions and the total weight of the rocket.
- Perform data wrangling
  - The data was cleaned in Python with pandas Library ensuring of use appropriate data and the most relevant information for the prediction and analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data was collected using Web Scraping with Beautiful Soup Library in wich we can see and retrieve information from tables and text.

Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. There are many different ways to perform web scraping to obtain data from websites. These include using online services, particular API's or even creating your code for web scraping from scratch



Data Collection – SpaceX API

Objectives

In this lab, you will make a get request to the SpaceX API. You will also do some basic data wrangling and formating.

- Request to the SpaceX API
- Clean the requested data

1. Request and parse the SpaceX launch data using the GET request
2. Normalize JSON response into a dataframe
3. Extract only useful columns using auxilary functions
4. Create new pandas dataframe from dictionary
5. Filter dataframe to only include Falcon 9 launches
6. Handle missing values
7. Export to CSV file

- GitHub URL: Data Collection API

Data Collection - Scraping

Objectives

Web scrap Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

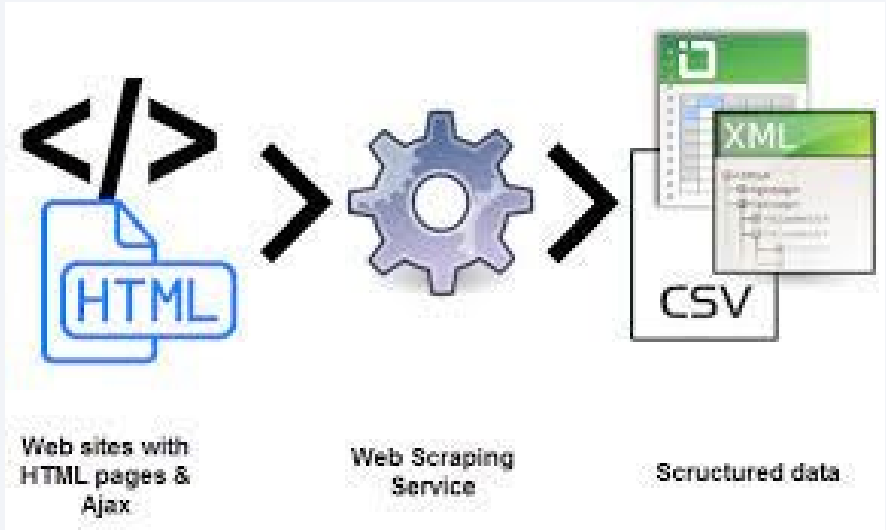
1. Request rocket launch data from its Wikipedia page

2. Extract all column/variable names from the HTML table header

3. Create a data frame by parsing the launch HTML tables

4. Export to CSV file

- GitHub URL: Data Collection Web Scraping



Data Wrangling

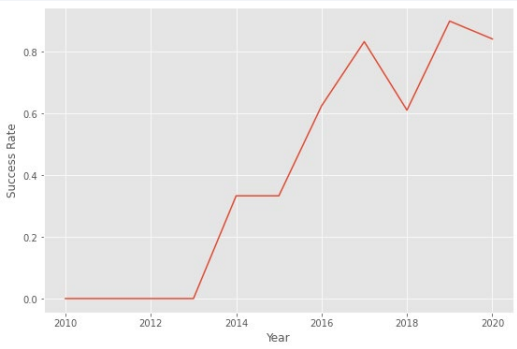
Data cleaning is the process that removes data that does not belong in your dataset. Data transformation is the process of converting data from one format or structure into another.

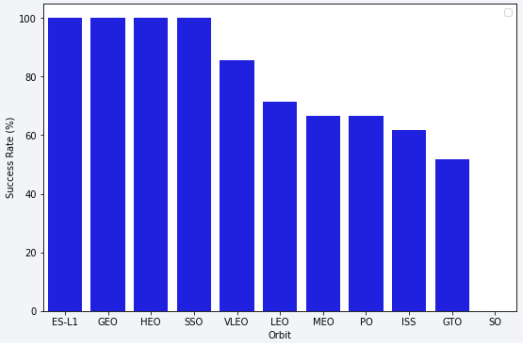
We use this process in order to get good data to feed our **machine learning** model and get high accuracy and avoid bias in the data

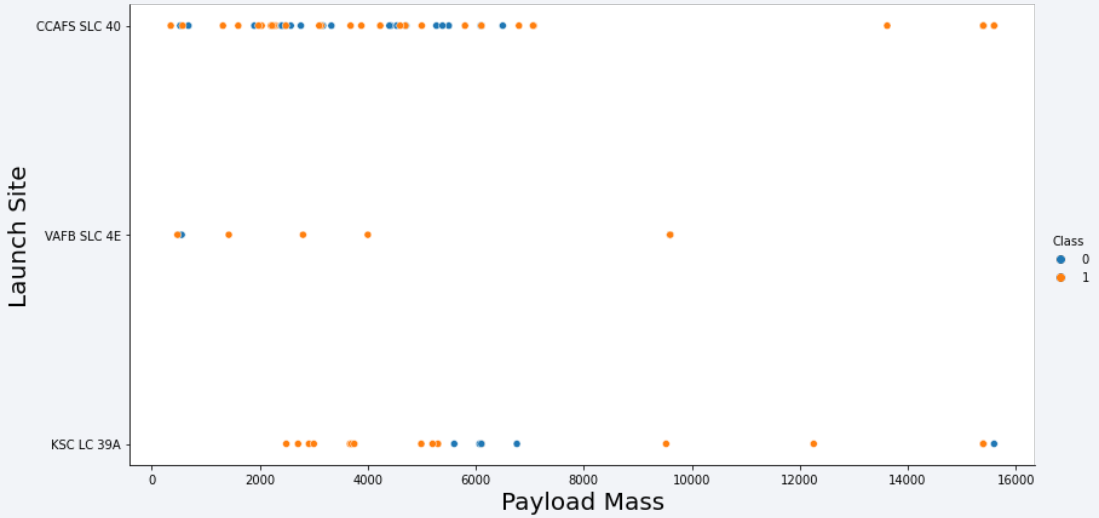
- ✓ Step 1: Remove duplicate or irrelevant observations
- ✓ Step 2: Fix structural errors
- ✓ Step 3: Filter unwanted outliers
- ✓ Step 4: Handle missing data
- ✓ Step 5: Validate and QA

EDA with Data Visualization

Data scientists use exploratory data analysis (EDA) to analyze and investigate data sets and summarize their main features, often using data visualization methods. Helps determine the best way to manage data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or verify assumptions.







- GitHub URL: [EAD Data Visualization]

EDA with SQL

- Understand the Spacex DataSet
- Load the dataset into the corresponding table in a Db2 database
- Execute SQL queries to answer assignment questions

- GitHub URL: SQLite Queries

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

```
%sql select * from SPACEXTBL limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success

Task 9

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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql SELECT date as month ,MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL  where substr(Date,7,4)='2015'
```

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

month	Mission_Outcome	Booster_Version	Launch_Site
01/10/2015	Success	F9 v1.1 B1012	CCAFS LC-40
02/11/2015	Success	F9 v1.1 B1013	CCAFS LC-40
03/02/2015	Success	F9 v1.1 B1014	CCAFS LC-40
14/04/2015	Success	F9 v1.1 B1015	CCAFS LC-40
27/04/2015	Success	F9 v1.1 B1016	CCAFS LC-40
28/06/2015	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
22/12/2015	Success	F9 FT B1019	CCAFS LC-40

Build an Interactive Map with Folium

Objects were created and added to a Folium map. Marker objects were used to show all launch sites on a map as well as the successful/failed launches for each site on the map. Line objects were used to calculate the distances between a launch site to its proximities

- By adding these objects, following geographical patterns about launch sites are found:
  - Are launch sites in close proximity to railways? Yes
  - Are launch sites in close proximity to highways? Yes
  - Are launch sites in close proximity to coastline? Yes
  - Do launch sites keep certain distance away from cities? Yes
- GitHub URL : Interactive Map Analytics with Folium

Build a Dashboard with Plotly Dash

The dashboard application contains two charts:

- A pie chart that shows the successful launch by each site. This chart is useful as you can visualize the distribution of landing outcomes across all launch sites or show the success rate of launches on individual sites.
- A scatter chart that shows the relationship between landing outcomes an the payload mass of different boosters. The dashboard takes two inputs, namely the site(s) and payload mass. This chart is useful as you can visualize how different variables affect the landing outcomes,

Predictive Analysis (Classification)

1. Create column for "Class"

2. Standardizing the data

3. Split ito training and test set

4. Find best Hyperparameter for SVM, Decision Trees, K-Nearest Neighbours and Logistic Regression.

5. Use test data to evaluate models based on their accuracy scores and confusion matrix

- GitHub URL: SpaceX_Machine_Learning_Prediction

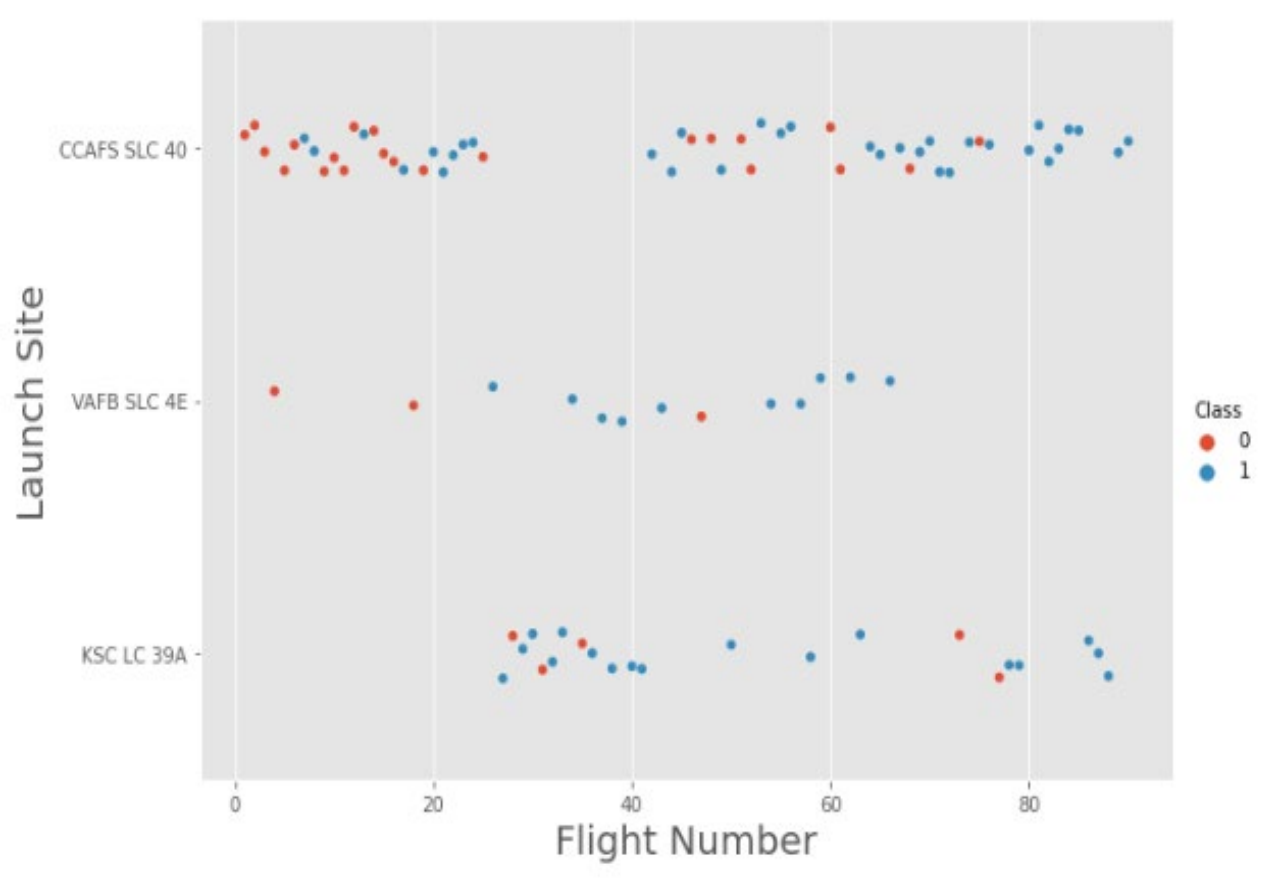
Results

- The results of the exploratory data analysis revealed that the success rate of the Falcon 9 landings was 66.66%
- The predictive analysis results showed that the Decision Tree algorithm was the best classification method with an accuracy of 94%

Section 2

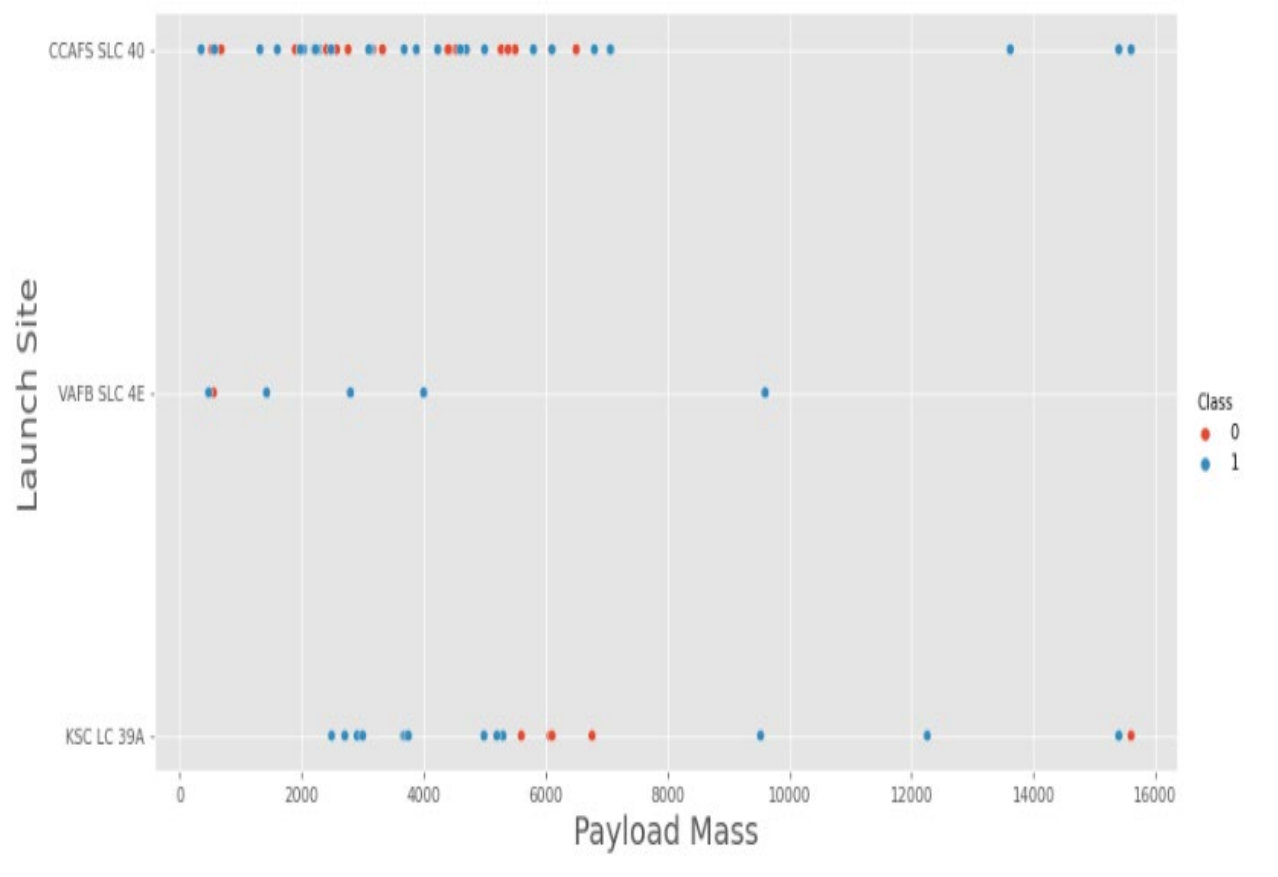
Insights drawn from EDA

Flight Number vs. Launch Site



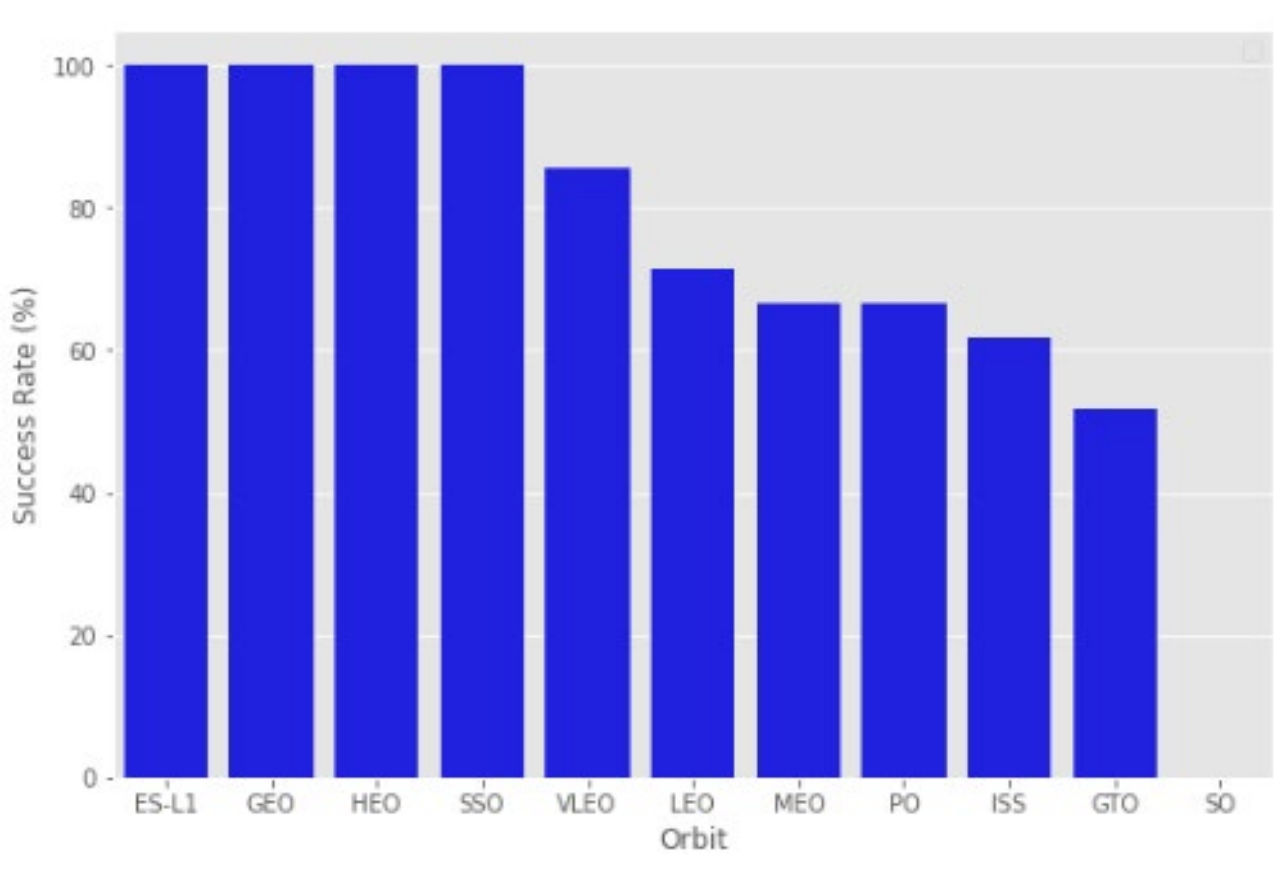
- This figure shows that the success rate increased as the number of flights increased.
- The blue dots represent the successful launches while the red dot represent unsuccessful luanches.
- There seems to be an increase in successful flights after the 40th launch.

Payload vs. Launch Site



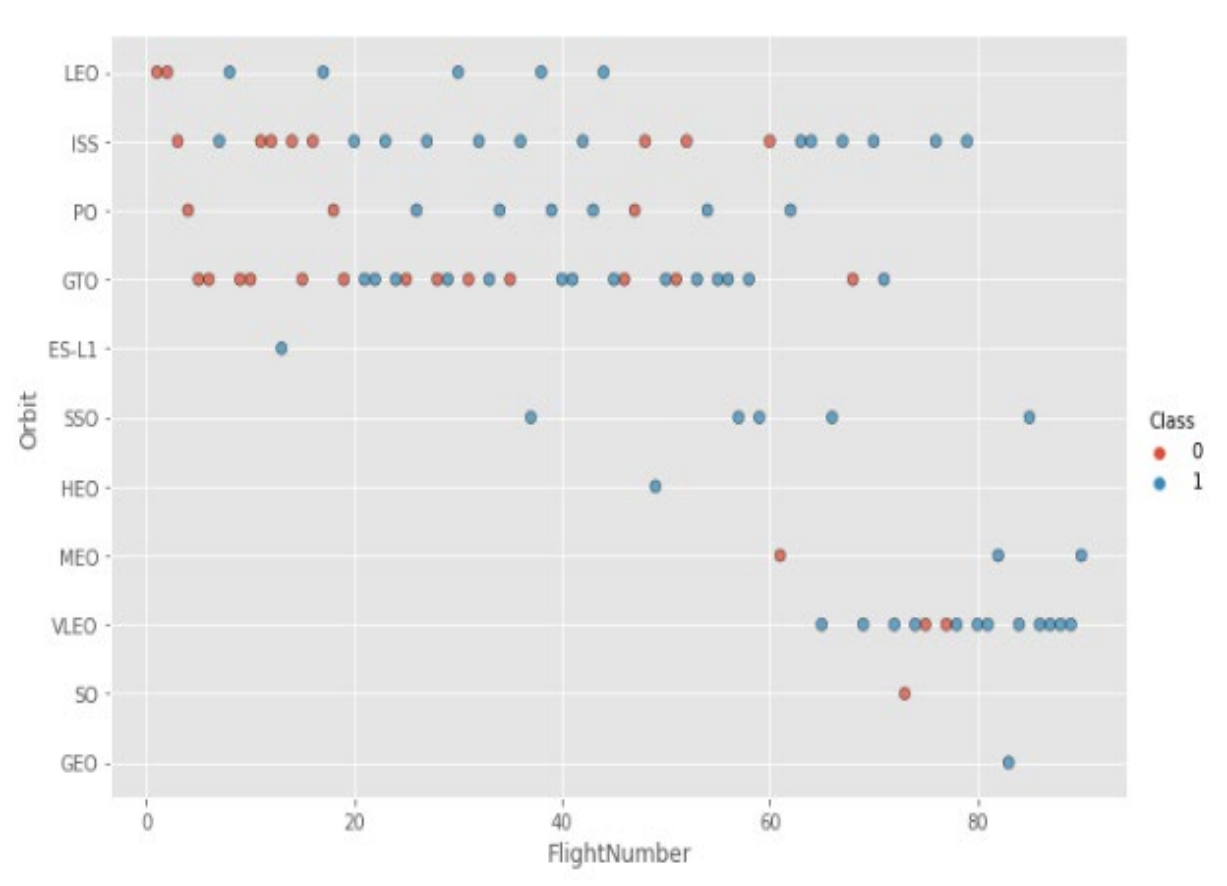
- The blue dots represent the successful launches while the red dots represent unsuccessful launches.
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass
- There seems to be a weak correlation between Payload and Launch Site and therefore decisions cannot be made using this metric.

Success Rate vs. Orbit Type



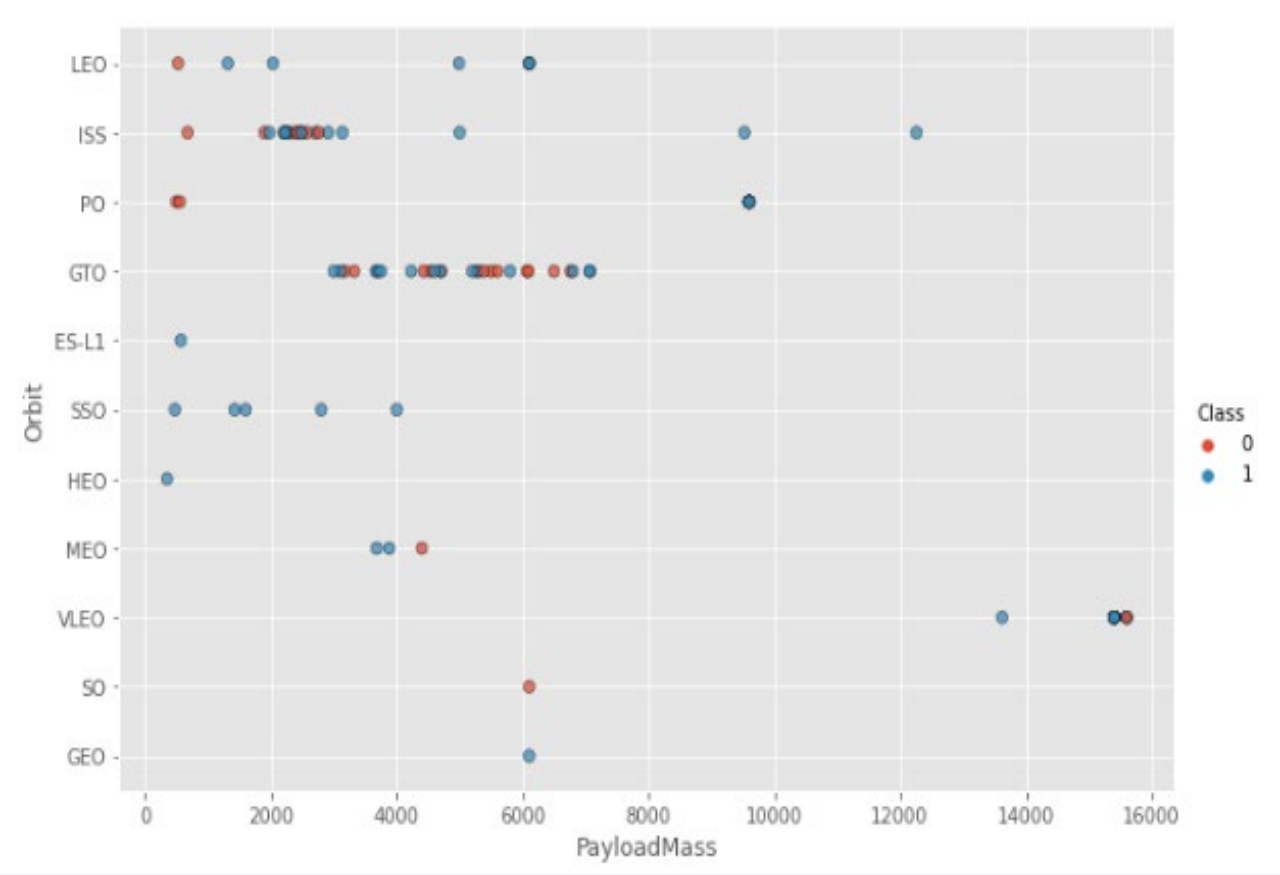
- Orbits SSO, HEO, GEO, and ES-L1 have 100% success rates.
- SO orbit did not have any successful launches with a 0% success rate.

Flight Number vs. Orbit Type



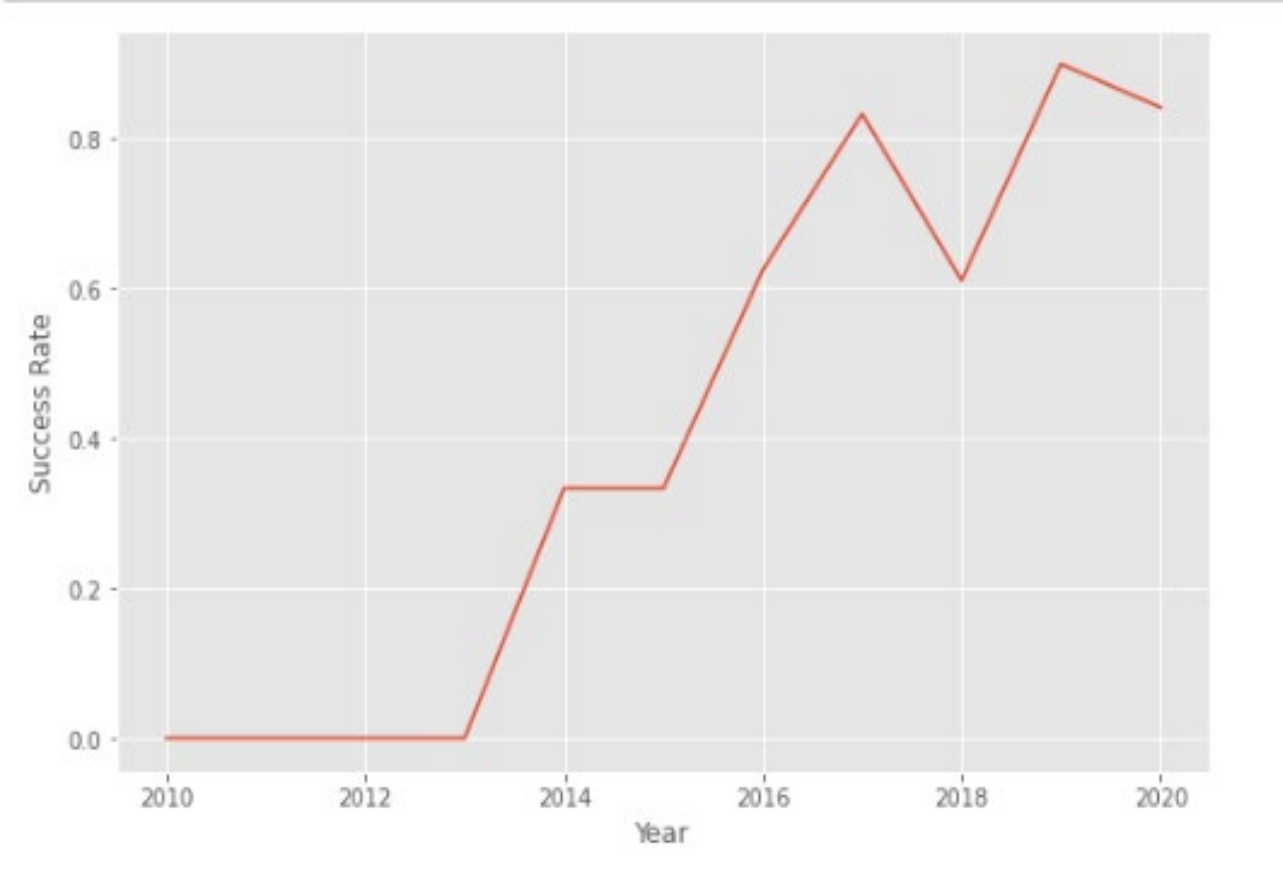
- In the LEO orbit, the success is positively correlated to the the number of flights.
- There seems to be no relationship between flight number in the GTO orbit.
- The SSO orbit has a 100% success rate however with fewer flights than the other orbits
- FLights numbers greater than 40 have a higher success rate than flight numbers between 0-40.

Payload vs. Orbit Type



- As the payloads get heavier, the success rate increases in the PO, SSO, LEO and ISS orbits.
- There seems to be no direct correlation between obrit type and payload mass for GTO orbit as both successful and failed launches are eually present

Launch Success Yearly Trend



- The general trend of the chart shows an increase in landing success rate as the years pass. There is however a dip in 2018 as well as in 2020.

All Launch Site Names

- The DISTINCT clause was used to return only the unique rows from the *launch_site* column.
- The names of the launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E .

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

Launch Site Names Begin with 'CCA'

- The LIMIT and LIKE clauses were used to display only the top five results where the *launch_site* name starts with 'CCA'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- The SUM() function was used to the calculate the total payload carried by boosters from NASA from the *payload_mass__kg* column.

total_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

- The AVG() function was used to the calculate the average payload the average payload mass carried by booster version F9 v1.1
- The WHERE clause was used to filter results so that the calculations were only performed on *booster_versions* only if they were named “F9 v1.1”

avg_payload_mass_kg
2928

First Successful Ground Landing Date

- The MIN(DATE) function was used to find the date of the first successful landing outcome on ground pad
- The WHERE clause ensured that the results were filtered to match only when the *'landing_*outcome' column is 'Success (ground pad)'

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The BETWEEN clause was used to retrieve only those results of payload mass greater than 4000 but less than 6000. The WHERE clause filtered the results to include only boosters which successfully landed on drone ship

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 COUNT() function is used to count the number of occurences of different mission outcomes with the help of the GROUPBY clause applied to the '*mission_outcome*' column. A list of the total number of successful and failure mission outcomes os returned.
- There have been 99 successful mission outcomes out of 101 missions.

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

Boosters Carried Maximum Payload

- The MAX() function was used in a subquery to retrieve a list of boosters which have carried the maximum payload mass

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

2015 Launch Records

- The SELECT statement was used to retrieve multiple columns from the table. The YEAR(DATE) function was used to retrieve only those rows with a 2015 launch date.

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

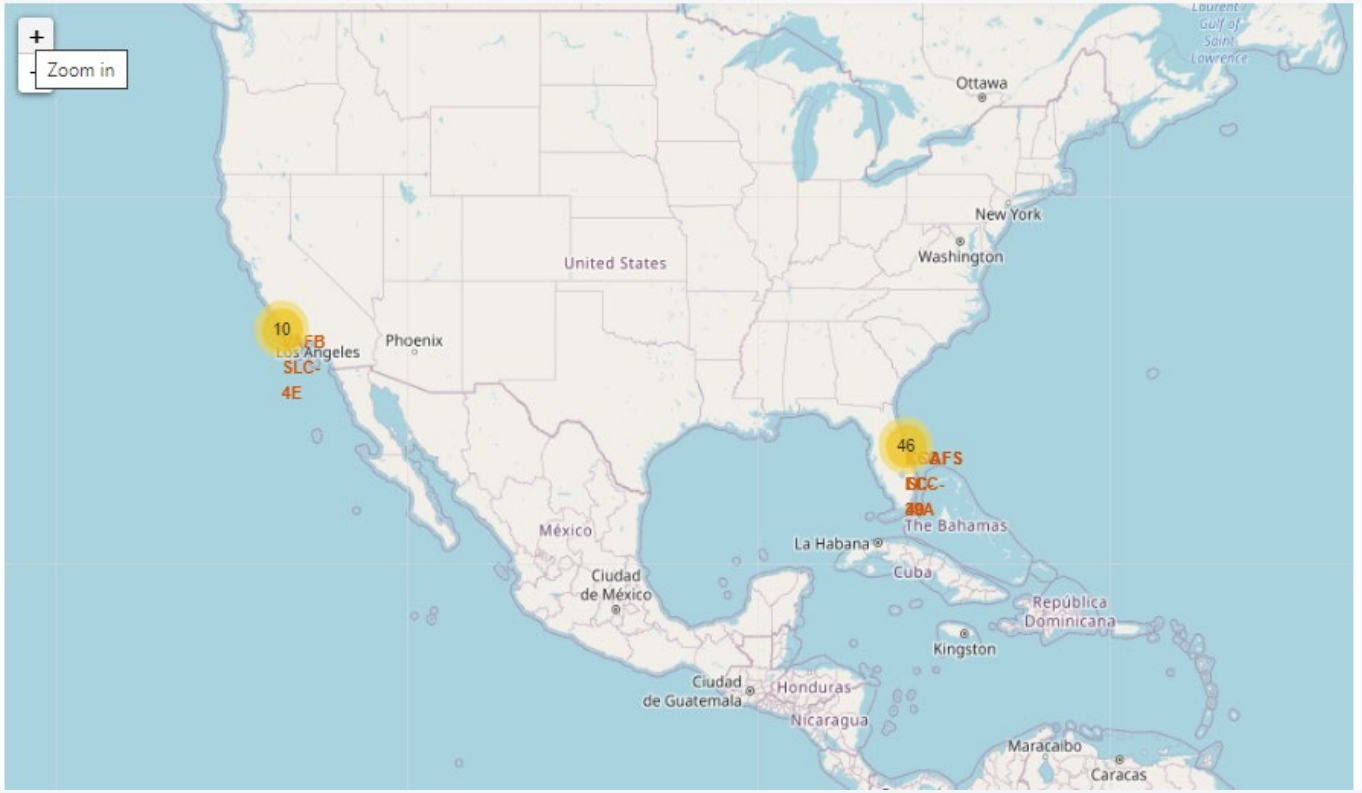
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- COUNT() function was used to count the different *landing outcomes.* The WHERE and BETWEEN clauses filtered the results to only include results between 2010-06-04 and 2017-03-20. The GROUPBY clause ensure that the counts were grouped by their outcome. The ORDERBY and DESC clauses were used to sort the results by descending order.

landing__outcome	total_number
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

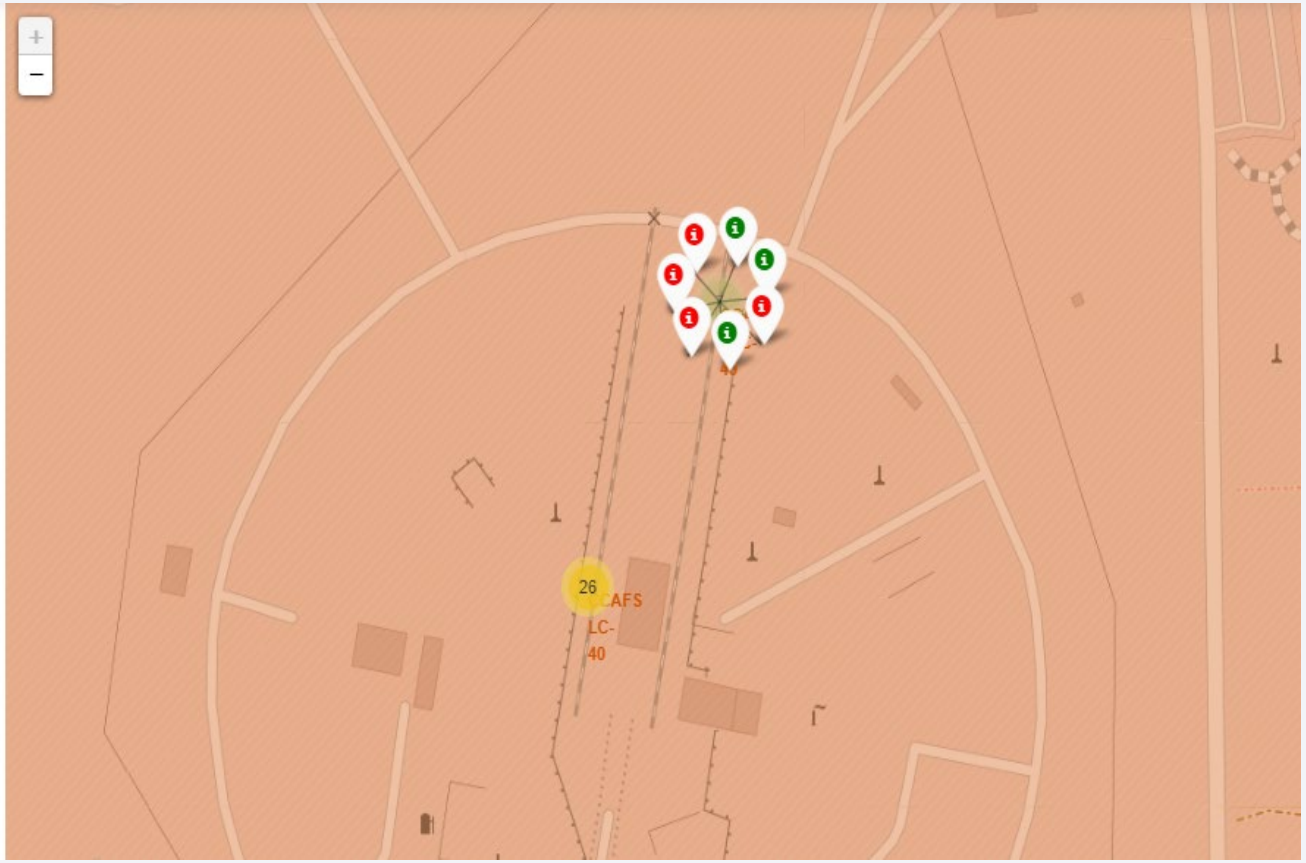

Section 3
Launch Sites
Proximities Analysis

SpaceX Launch Sites Locations



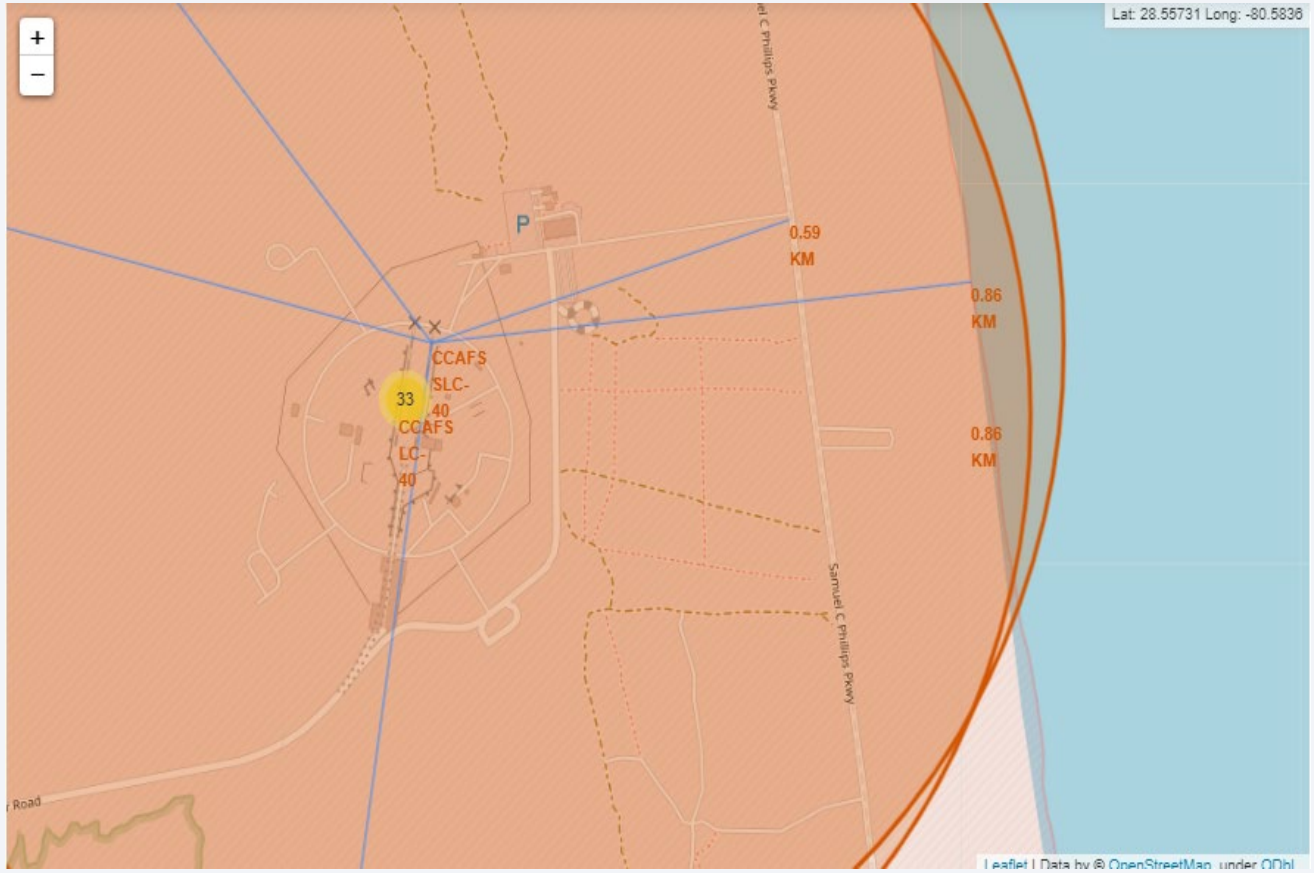
- The yellow markers are indicators of where the locations of all the SpaceX launch sites are situated in the US.
- The launch sites have been strategically placed near the coast

Success or Failure?

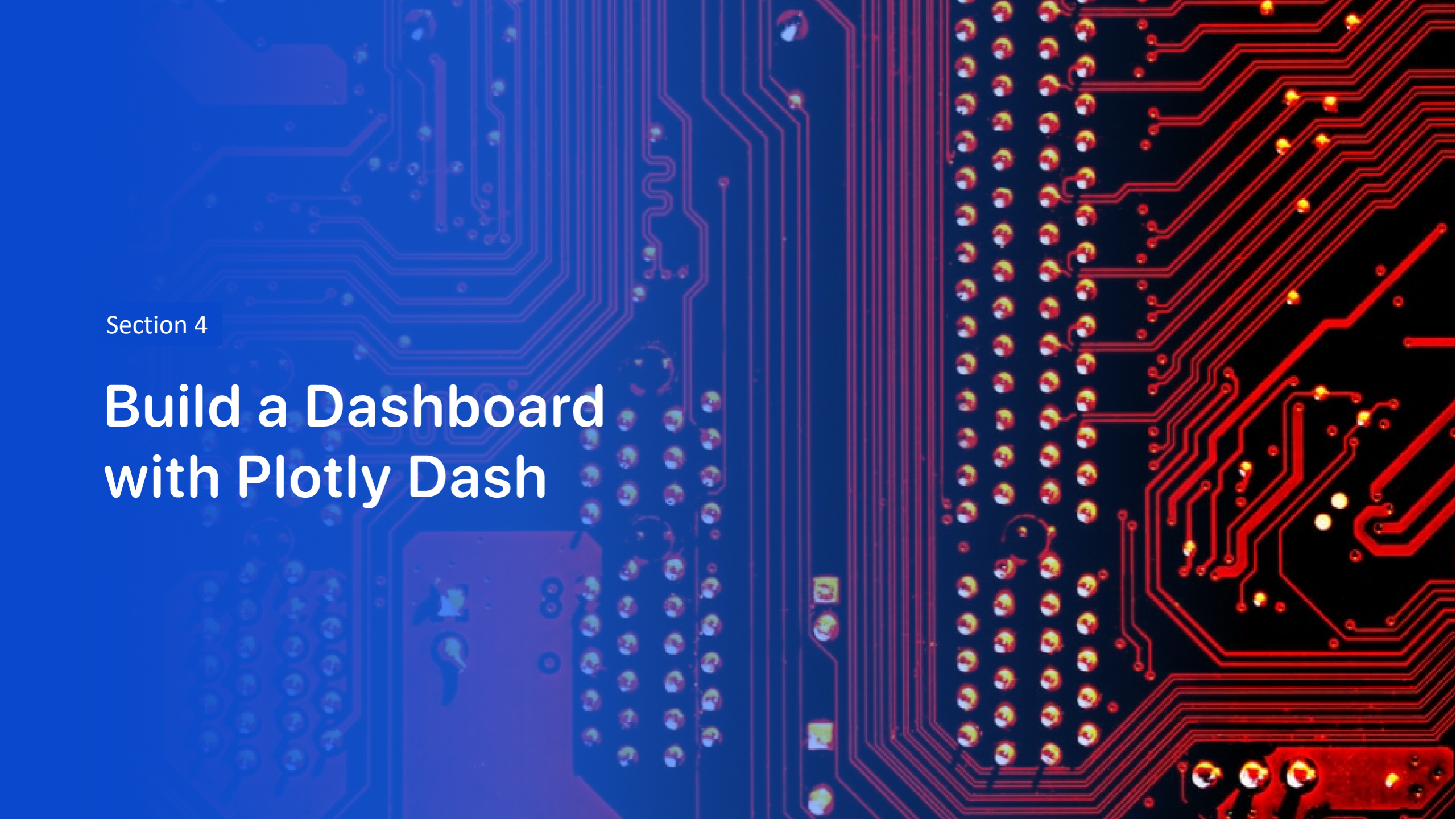


- When we zoom in on a launch site, we can click on the launch site which will display marker clusters of successful landings (green) or failed landing (red).

Launch Site Proximities

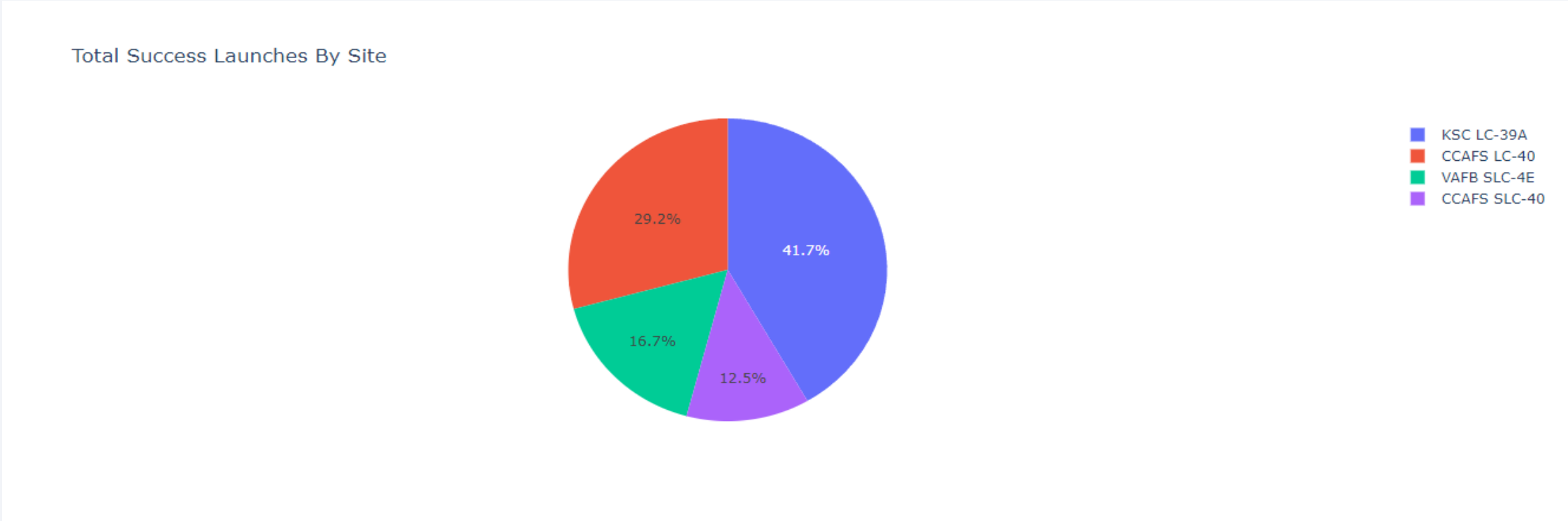


- The generated map shows that the selected launch site is close to a highway for transportation of personnel and equipment. The launch site is also close to the coastlines for launch failure testing.
- The launch sites also maintain a certain distance from the cities. (Can be viewed in notebook).


Section 4
Build a Dashboard with Plotly Dash

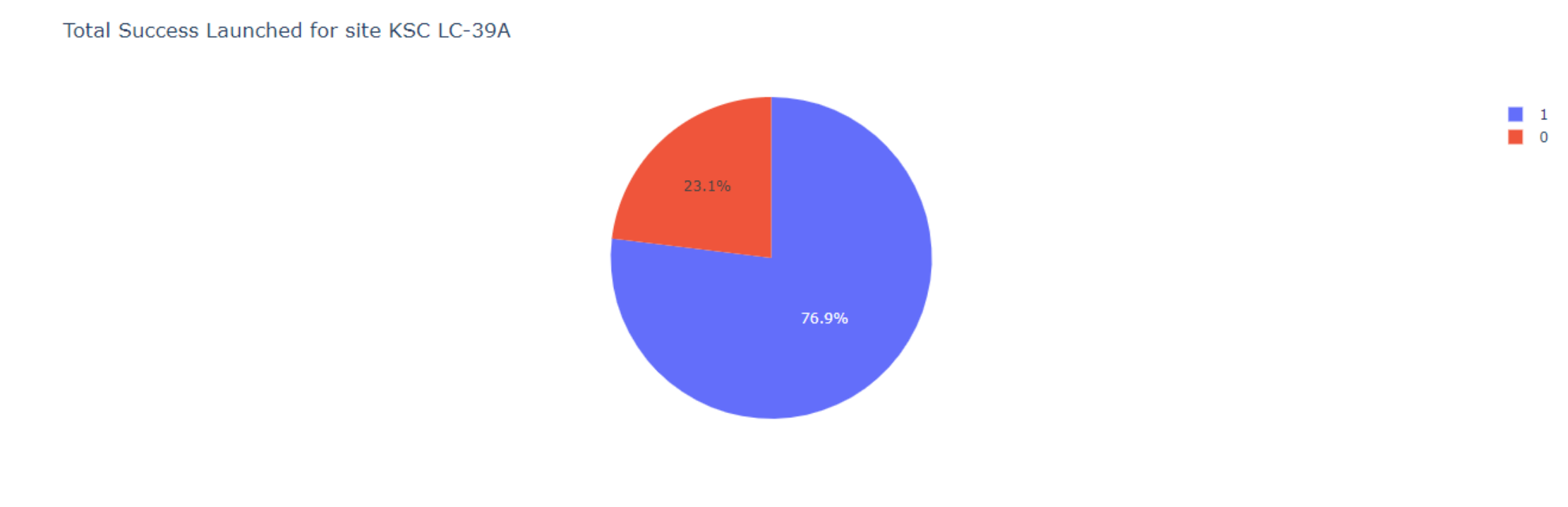
Total Successful Launches By Site

- The KSC LC-39A Launch site has the most successful launches with 10 in total.



Launch Site With Highest Success Ratio

- The KSLC-39A has the highest success rate with 76.9%.



Payloads vs Launch Outcome

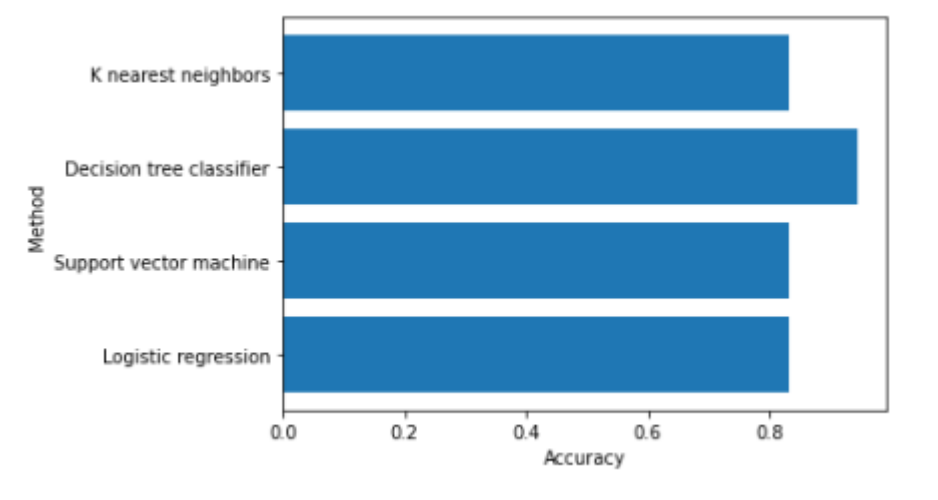
- The launch success rate for payloads 0-2500 kg is slightly lower than that of payloads 2500-5000 kg. There is in fact not much difference between the two.
- The booster version that has the largest success rate, in both weight ranges is the *v1.1*.

Section 5

Predictive Analysis (Classification)

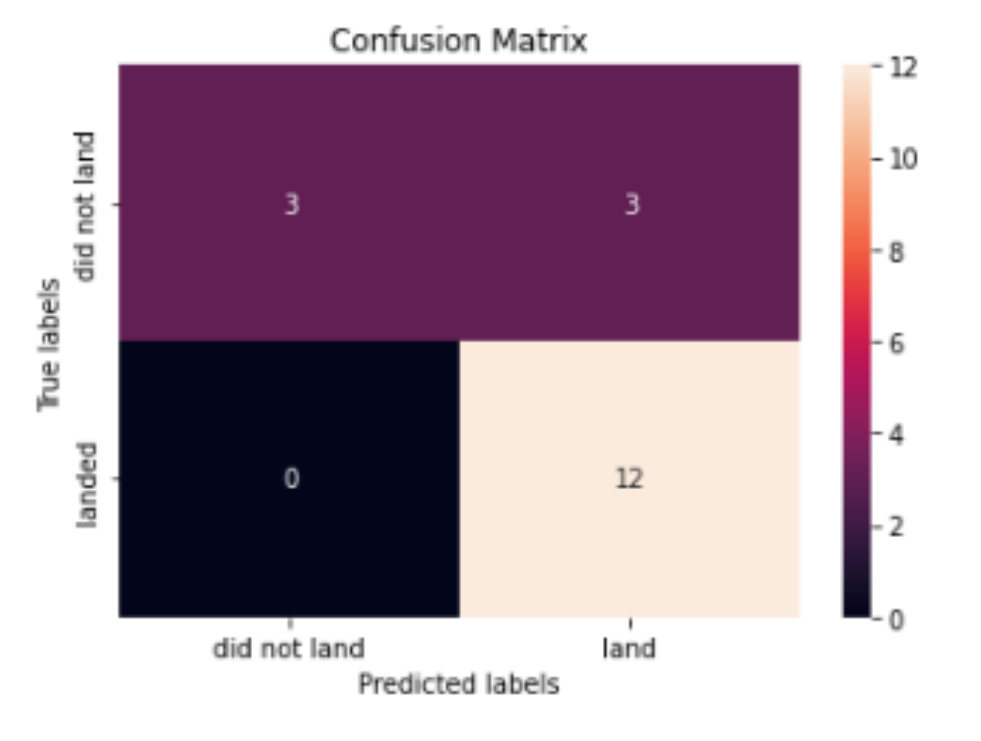
Classification Accuracy

- The Decision Tree classifier had the best accuracy at 94%.



	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333

Confusion Matrix



- The model predicted 12 successful landings when the True label was successful (True Positive) and 3 unsuccessful landings when the True label was failure (True Negative).
- The model also predicted 3 successful landings when the True label was unsuccessful landing (False Positive).
- The model generally predicted successful landings.

Conclusions

- The analysis showed that there is a positive correlation between number of flights and success rate as the success rate has improved over the years.
- There are certain orbits like SSO, HEO, GEO, and ES-L1 where launches were the most successful.
- Success rate can be linked to payload mass as the lighter payloads generally proved to be more successful than the heavier payloads.
- The launch sites are strategically located near highways and railways for transportation of personel and cargo, but also far away from cities for safety.
- The best predictive model to use for this dataset is the Decision Tree Classifier as it had the highest accuracy with 94%.

Appendix

- Coursera Project Link: https://www.coursera.org/learn/applied-data-science-capstone/home/welcome
- GitHub Repository: https://github.com/SergioDAtaAnalyst/Space-X_Landing_Prediction

Thank you!