IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thiago Rocha
31st August, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - **Data Collection:** Information was gathered from the SpaceX public API and Wikipedia. The data wrangling process involved extracting launch outcome details to be used as the dependent variable in the Machine Learning Models.
  - **Data Analysis:** SQL queries and various visualizations (including static charts, interactive maps, and a dynamic dashboard) were employed to explore the dataset and address specific questions.
  - **Predictive Modeling:** Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KMN) models were utilized for predictive analysis.
- Summary of all results
  - The launch data encompasses details such as flight number, launch date, payload mass, orbit type, launch site, mission outcome, and other relevant variables.
  - Model performance using Logistic Regression, SVM and KMN models demonstrated comparable performance when applied to this dataset.

Introduction

- **Project background and context**

SpaceX emerged as the leading player in the commercial space era by making space travel more affordable. The company promotes its Falcon 9 rocket launches on its website, priced at $62 million, while other providers charge over $165 million per launch. A significant portion of these savings comes from SpaceX's ability to reuse the first stage of their rockets. By predicting whether the first stage will successfully land, we can estimate the cost of a launch. Using publicly available data and machine learning models, we aim to forecast whether SpaceX will be able to reuse the first stage.

- **Problems you want to find answers**
  - How do factors like payload mass, launch site, number of flights, and orbits influence the success rate of the first stage landing?
  - Has the success rate of landings improved over time?
  - What is the most suitable algorithm for binary classification in this situation?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API and Web Scrapping from Wikipedia were used to collect data.
- Perform data wrangling
  - Data Filtering
  - Working with missing values
  - One Hot Enconding were used to prepare data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Creating, optimizing, and reviewing classification models to ensure the highest quality outcomes.

Data Collection

The data collection process combined API requests from the SpaceX REST API with Web Scrapping from a table on SpaceX's Wikipedia page. Utilizing both methods was necessary to obtain comprehensive information about the launches for a more thorough analysis.

SpaceX REST API data columns acquired: FlightNumber, Date, Booster Version, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

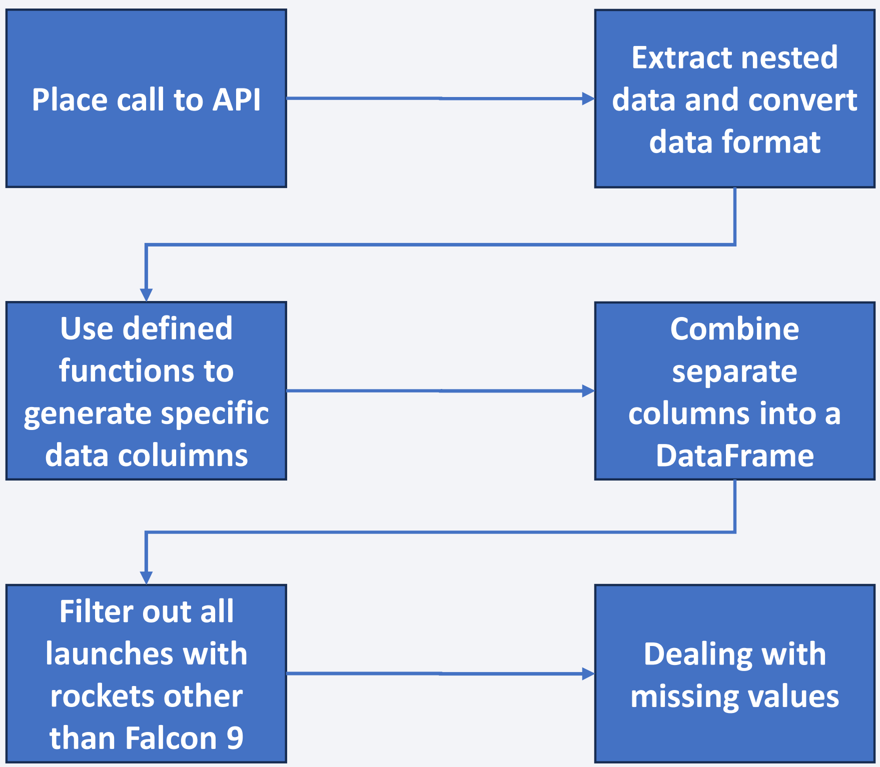
Wikipedia Web Scrapping data columns acquired: FlightNo., Launch site, Payload, PayloadMass, Orbit, Customer, Launchoutcome, Version Booster, Booster landing, Date, Time.

Data Collection – SpaceX API

- The SpaceX API provides data that is publicly available.
- After making a GET request to the SpaceX API and the response is obtained, the data can be imported into a Pandas DataFrame for further analysis.

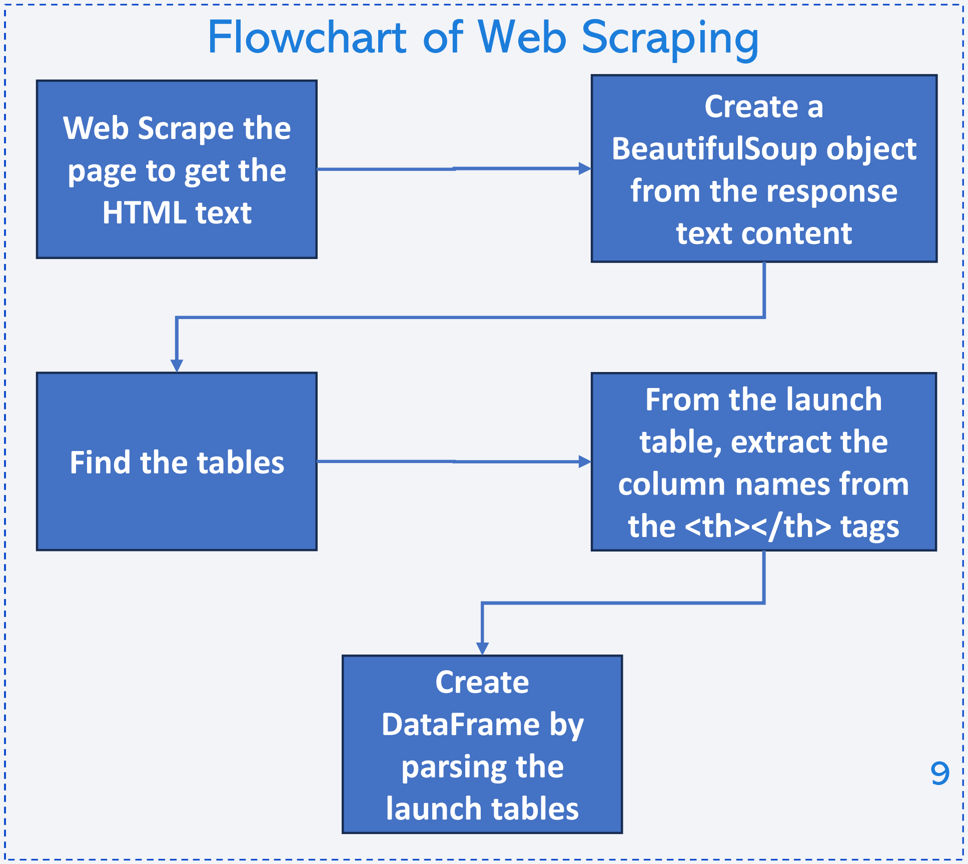
- Data Collection GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Flowchart of SpaceX API Calls



Data Collection - Scraping

- A Wikipedia page includes tables with information about SpaceX launches.
- By scrapping these tables, launch data can be extracted and imported into a Pandas DataFrame for more in-depth analysis.
- Web Scrapping GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- In the dataset, the are various scenarios where the booster did not land successfully. For instance, a landing attempt might have failed due to an accident. Specifically, “True Ocean” indicates that the booster successfully landed in a specific ocean area, while”False Ocean” means it failed to land in that region. Similarly, “True RTLS” signifies a successful landing on a ground pad, whereas “False RTLS” denotes an unsuccessful landing on a ground pad. “True ASDS” represents a successful landing on a drone ship, while “False ASDS” indicates an unsuccessful landing on a drone ship.
- These outcomes are primarily converted into training labels, with “1” representing a successful landing and “0” representing an unsuccessful one.
- Data Wrangling GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Flowchart of Data Wrangling

Performe Exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

EDA with Data Visualization

- **Plotted Charts:** Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Massvs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend.
- Scatter Plots illustrate the relationship between variables. If a relationship is identified, these plots can be utilized in machine learning models.
- Bar charts are used to compare discrete categories, aiming to display the relationship between specific categories and their corresponding measured values.
- Line charts, on the other hand, reveal trends in data over time (time series).
- EDA with Data Visualization GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/edadataviz.ipynb

EDA with SQL

- **Performed SQL queries:**
  - Displayed the names of unique launch sites used in the space missions.
  - Retrieved 5 records where launch sites start with the string 'CCA'.
  - Calculated the total payload mass carried by boosters launched by NASA (CRS).
  - Computed the average payload mass for booster version F9 v1.1.
  - Listed the date of the first successful landing outcome on a ground pad.
  - Identified the names of boosters that successfully landed on a drone ship with a payload mass greater than 4000 but less than 6000.
  - Counted the total number of successful and failed mission outcomes.
  - Determined the booster versions that carried the maximum payload mass.
  - Found the failed landing outcomes on drone ships, including their booster versions and launch site names, for the months in 2015.
  - Ranked the count of landing outcomes (such a Failure on a drone ship or Successful on a ground pad) between 2010-06-04 and 2017-03-20 in descending order.

- EDA with SQL the GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Markers of All Launch Sites:**
  - Added markers with circles, popub labels, and text labels for NASA Johnson Space Center using its latitude and longitude coordinates as the starting point.
  - Placed markers with circles, popup labels, and text labels for all launch sites, indicating their geographical locations and proximity to the equator and coastlines.
- **Colored Markers for Launch Outcomes at Each Launch Site:**
  - Implemented colored markers to represent successful (green) and failed (red) launches, utilizing marker clusters to highlight launch sites with relatively high success after.
- **Distances from a Launch Site to Nearby Features:**
  - Added colored lines to illustrate distances between the Launch Site KSC LC-39A (as an example) and nearby features such as railways, highways, coastlines, and the nearest city.
- **Interactive Map with Folium GitHub URL:** https://github.com/Thirocha1984/Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

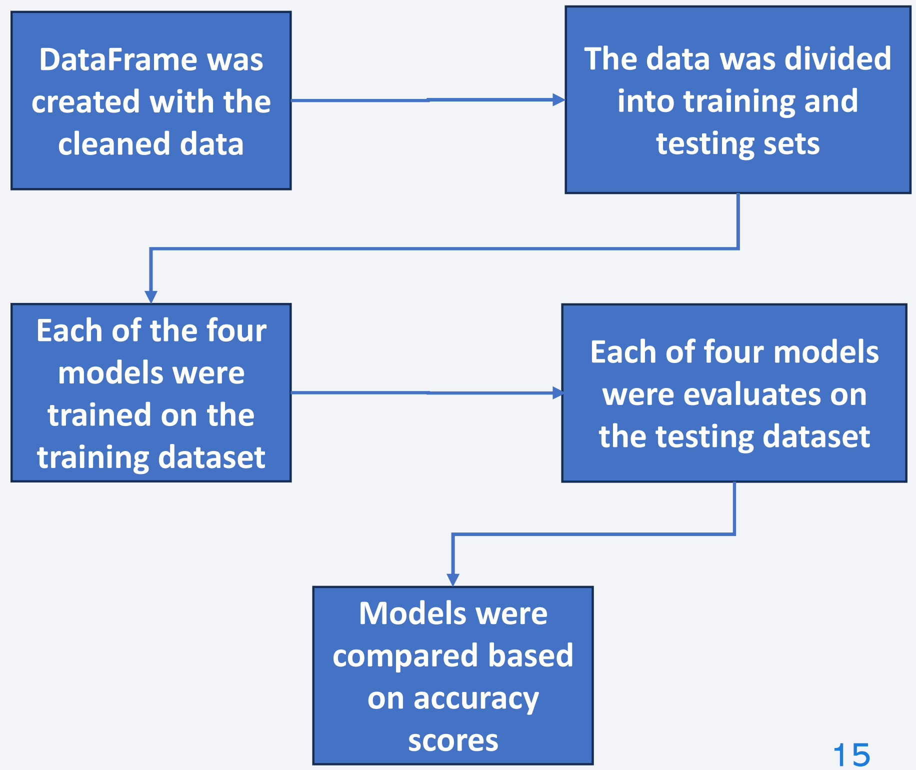
Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List:**
  - Added a dropdown list to select the Launch Site.
- **Pie Chart Showing Successful Launches (All Sites/Specific Site):**
  - Added a pie chart to display the total count of successful launches for all sites and the success vs. failure counts for a specific site if selected.
- **Payload Mass Range Slider:**
  - Added a slider to select the range of payload mass.
- **Scatter Chart of Payload Mass vs. Success Rate for Different Booster Versions:**
  - Added a scatter chart to illustrate the correlation between payload mass and launch success rate across different booster versions.
- **Dashboard with Plotly Dash GitHub URL:** https://github.com/Thirocha1984/Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The dataset was divided into training and testing sets.
- Logistic Regression, SVM (Support Vector Machine), Decision Tree and KNN (k-Nearest Neighbors) machine learning models were trained using the training dataset.
- Hyperparameters were evaluated with GridSearchCV(), and the optimal ones were selected using '.best_params_'.
- Each of the four models was assessed for accuracy using the testing dataset with the best hyperparameters.
- Machine Learning GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Flowchart of Machine Learning



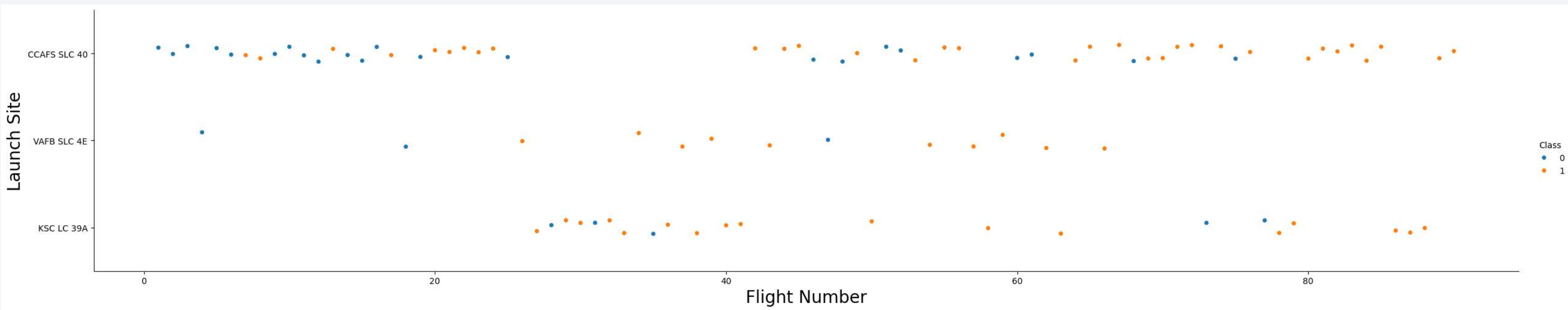
Results

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

Section 2

Insights drawn from EDA

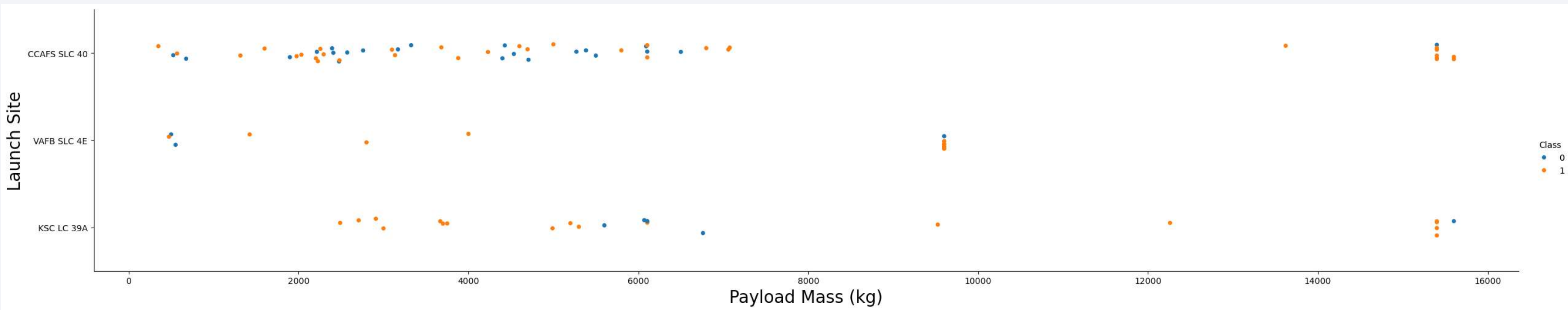
Flight Number vs. Launch Site



INSIGHTS

- The earliest flights all failed while the latest flights all succeeded.
- CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

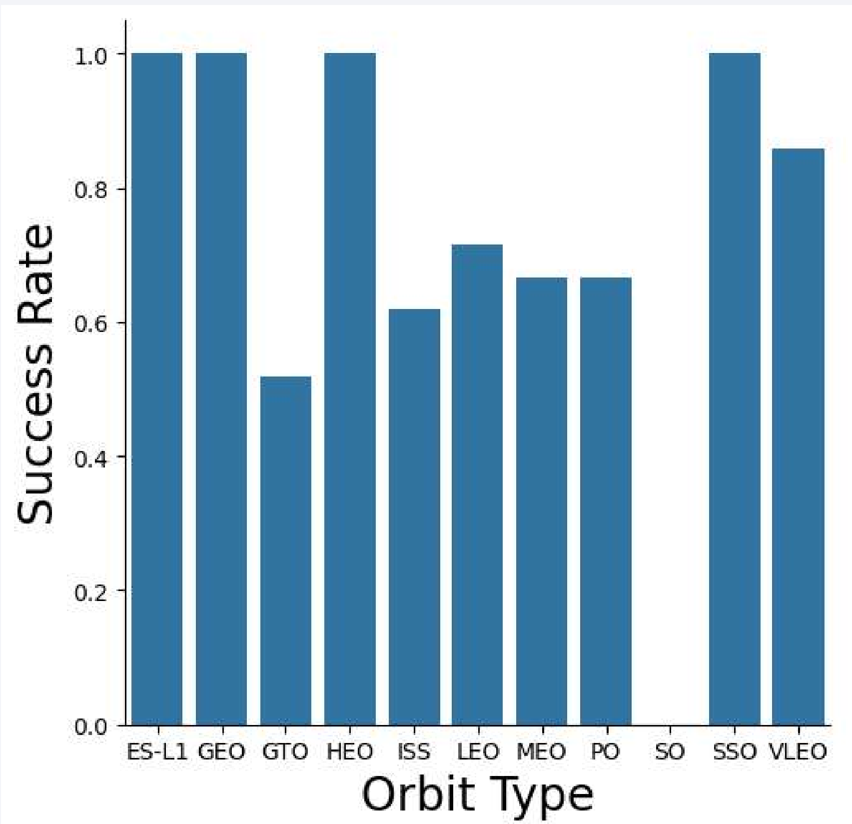
Payload vs. Launch Site



INSIGHTS

- For every launch site the higher the payload mass, the higher success rate.
- Most of the launches with payload mass over 7000kg were success.
- KSC LC 39A has a 100% success rate for payload mass under 5500kg.

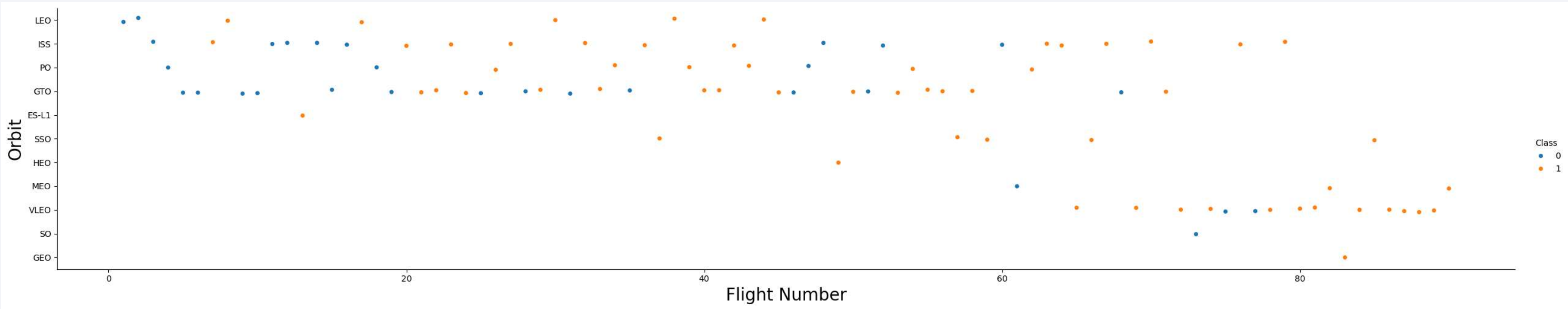
Success Rate vs. Orbit Type



INSIGHTS

- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO.
- Orbits with success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO.
- Orbits with 0% success rate: SO

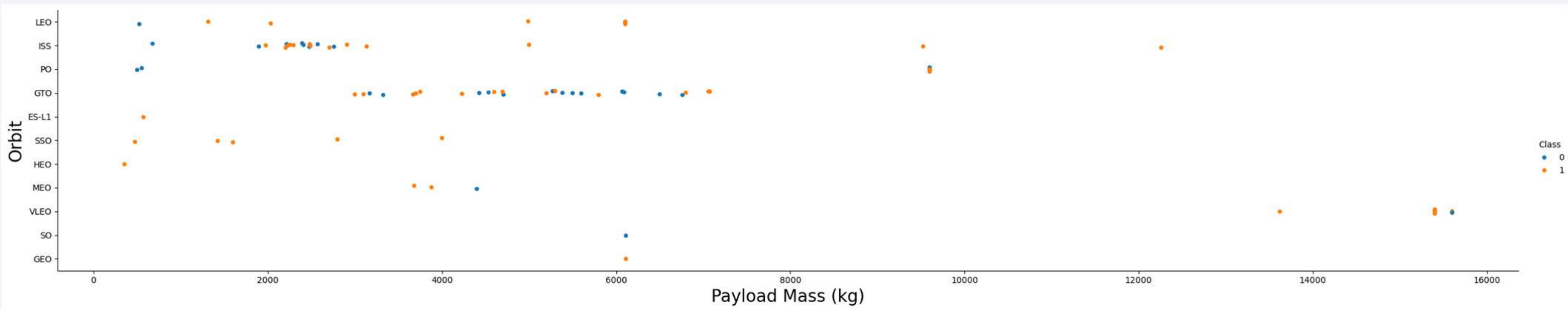
Flight Number vs. Orbit Type



INSIGHTS

- In the LEO Orbit the Success appears related to the number of flights. Conversely, in the GTO Orbit, there appears to be no relationship between flight number and success.

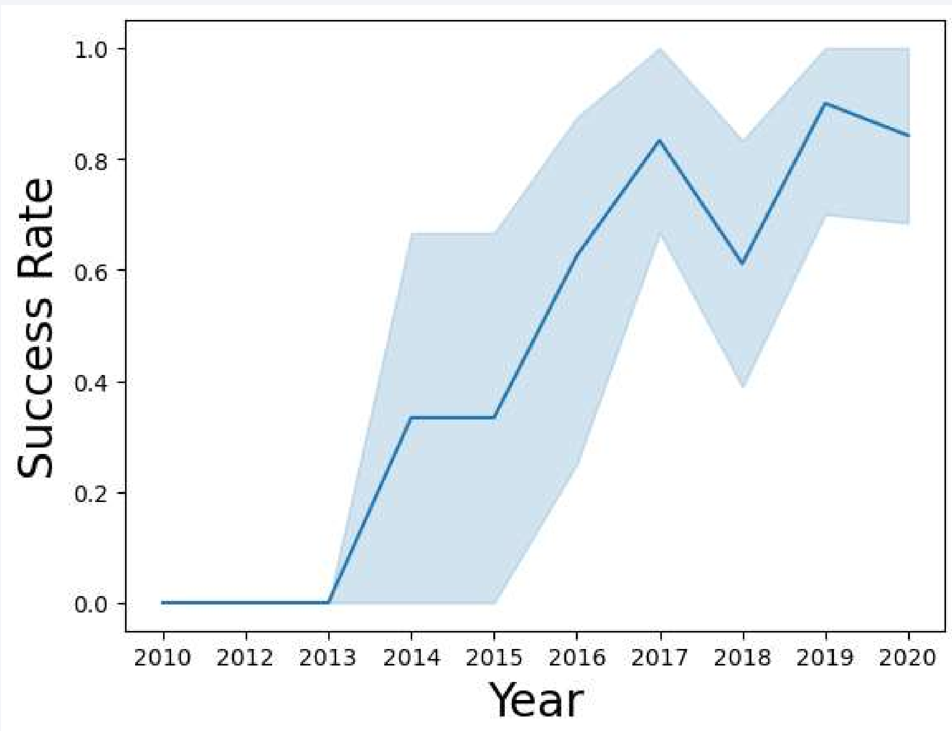
Payload vs. Orbit Type



INSIGHTS

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



INSIGHTS

- We can observe that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

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

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

EXPLANATION

- The are four unique launch sites

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTABLE 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

EXPLANATION

- This is a simple sampling method used to get a overview of the data in the database table.

Total Payload Mass

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

In [20]:
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

Out[20]:

total_payload_mass
45596

EXPLANATION

- The total payload mass carried by boosters from NASA (CRS) is 45,596kg.

Average Payload Mass by F9 v1.1

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

```
In [21]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';

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

Out[21]:

average_payload_mass
2534.6666666666665

EXPLANATION

- The average payload mass carried by booster F9 v1.1 is 2534,67kg.

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [22]:
```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

Out[22]:

first_successful_landing
2015-12-22

EXPLANATION

- The first successful landing outcome on ground pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
[23]: %sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

[23]:

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

EXPLANATION

- The four booster versions that have successfully landed on a drone ship with a payload mass exceeding 4000kg but less than 6000kg are listed above.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

[25]:
```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

[25]:

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

EXPLANATION

- The total number of success and failure mission outcomes are listed above.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[26]: %sql select booster_version from SPACEXTABLE WHERE\
payload_mass__kg_ = (select max (payload_mass__kg_) from SPACEXTABLE);

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

[26]:

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

EXPLANATION

- The largest payload mass recorded in this dataset is 15600kg. The twelve Falcon 9 boosters carried this amount of payload mass are listed beside.

2015 Launch Records

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.

```
%sql select substr(date, 6,2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE\
     where landing_outcome = 'Failure (drone ship)' and substr(date, 0,5)='2015';
```

```
 * 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)

EXPLANATION

- In 2015, there are two unsuccessful landings on drone ship, both originating from CCAFS LC-40. One incident took place in January, and the other in April.

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.

[43]:
```
%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.
```

[43]:

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

EXPLANATION

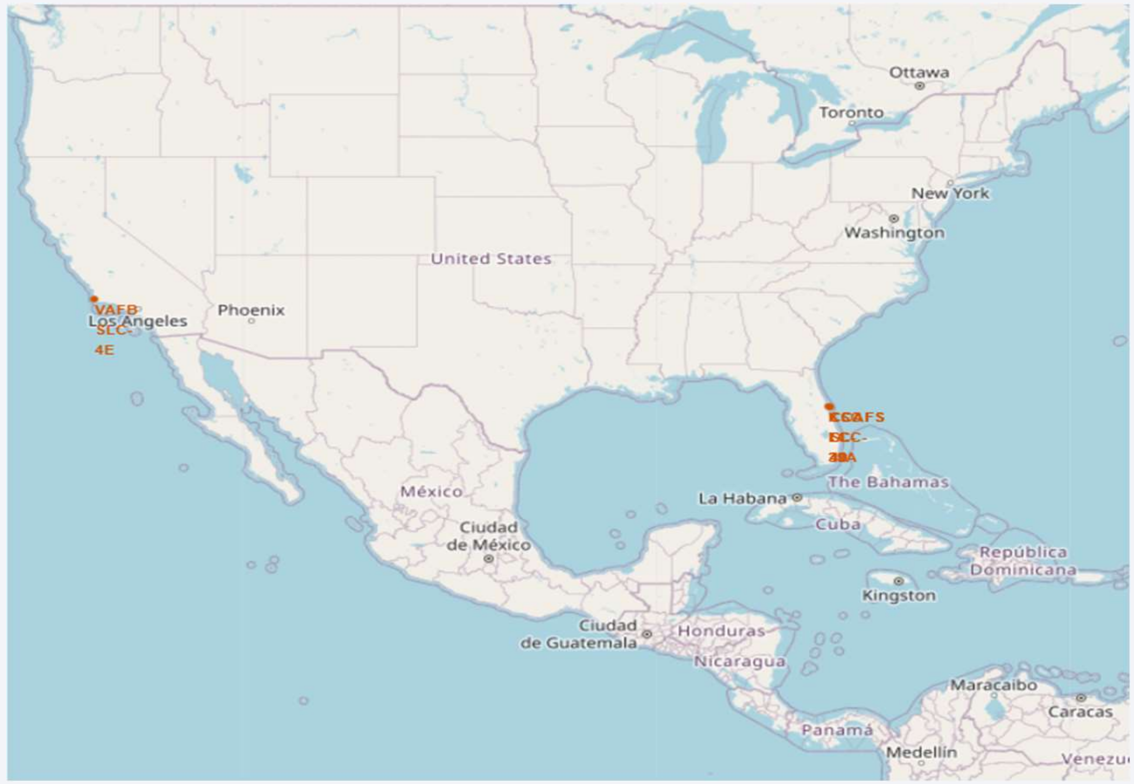
- ‘Not attempt’ is the most common landing outcome.

Section 3

Launch Sites Proximities Analysis

Launch Site Locations – Falcon 9

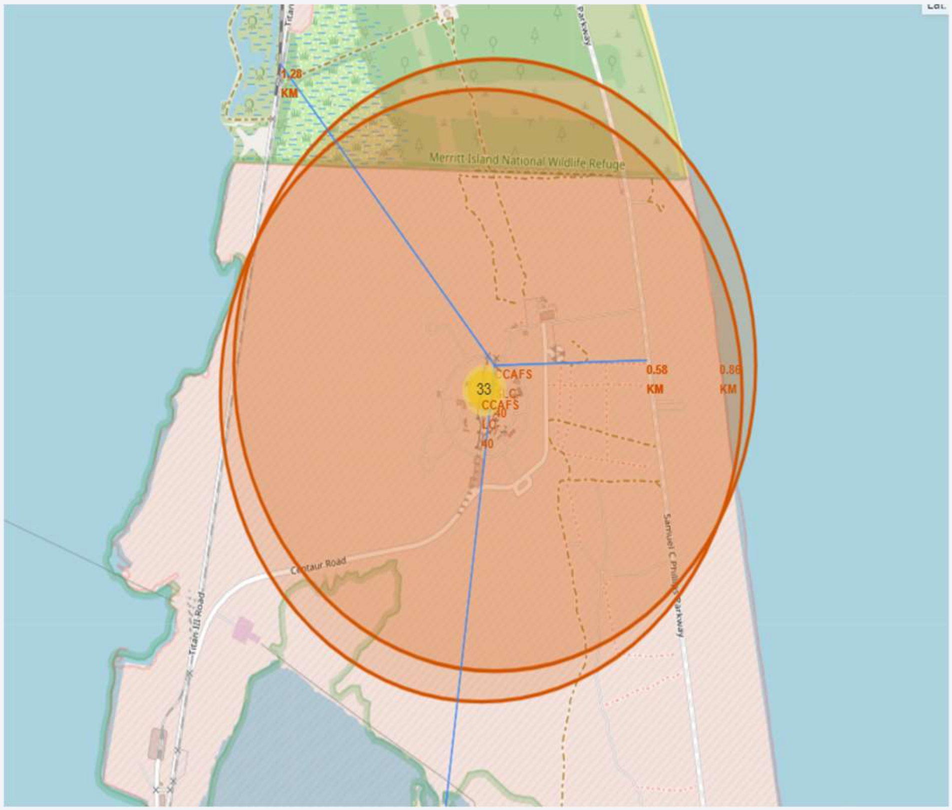
- VAFB SLC-4E (California, USA)
  - Vanderberg Air Force Base Space Launch Complex 4E
- KSC LC-39A (Florida, USA)
  - Kennedy Space Center Launch Complex 39A
- CCAFS LC-40 (Florida, USA)
  - Cape Canaveral Air Force Station Launch Complex 40
- CCAFS SLC-40 (Florida, USA)
  - ´Cape Canaveral Air Force Station Space Launch Complex 40



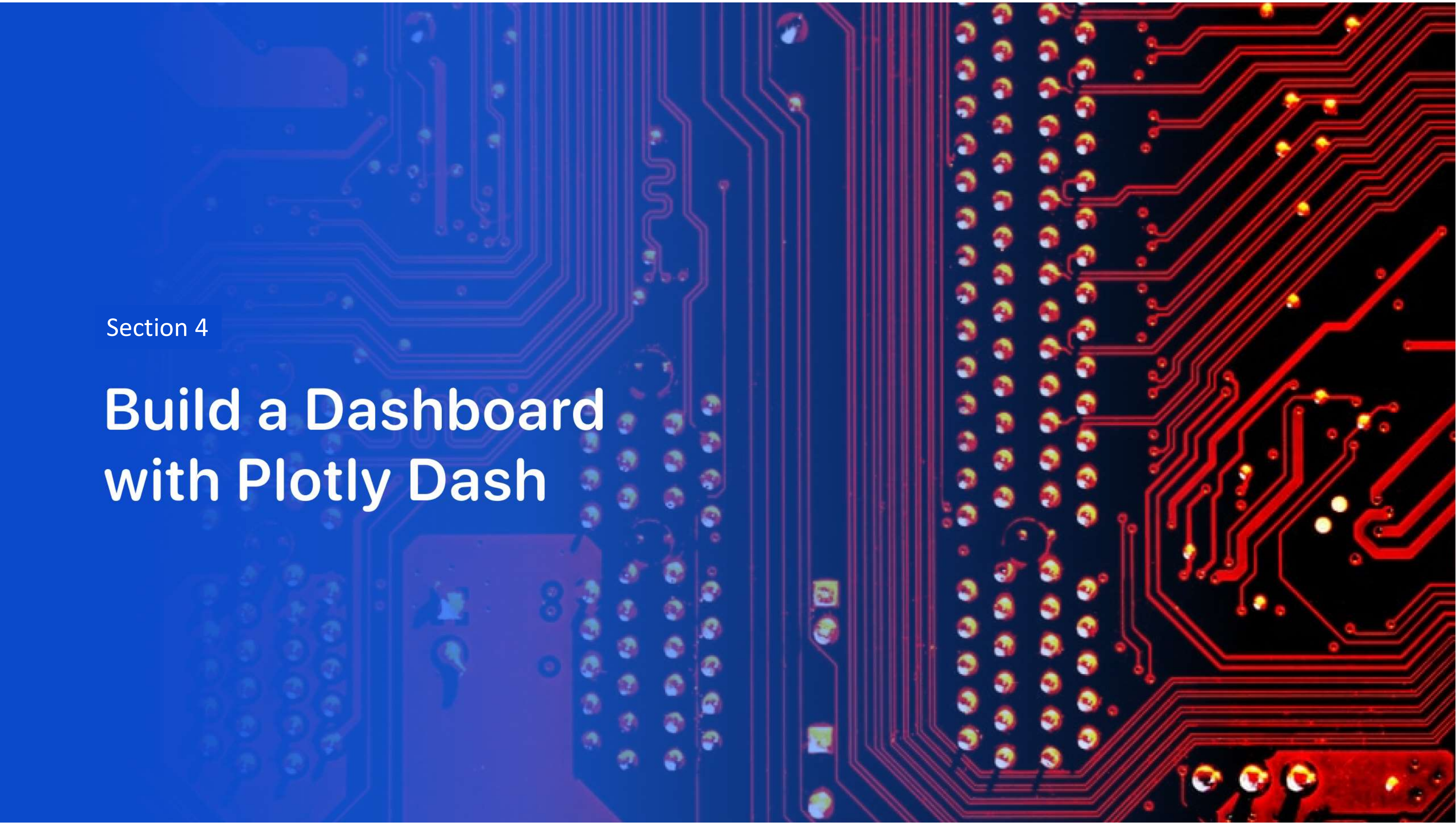
Success/Failed Landings – Map Markers

- By examining the color-coded markers, we can easily identify which launch sites have relatively high success rates:
  - Green Marker = Successful Launch
  - Red Marker = Unsuccessful Launch
- Launch Site KSC LC-39A exhibits a very high success rate.

Distance from Launch Site to Proximities



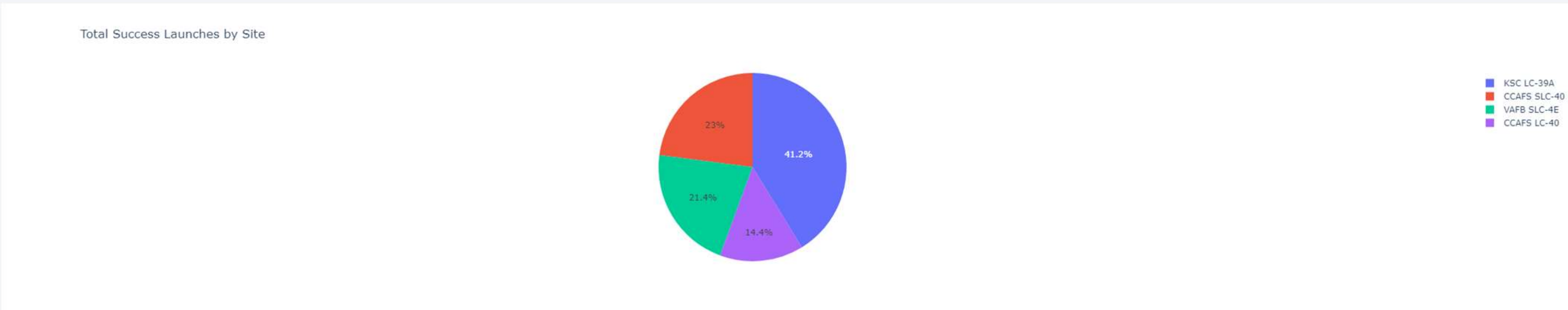
- The CCAFS SLC-40 and CCAFS LC-40 launch sites have coordinates that are nearly identical but not exactly the same.
- Proximities for CCAFS SLC-40:
  - Railway: 1.28km
    - Useful transporting heavy cargo
  - Highway: 0.58km
    - Useful transporting for equipment and employees
  - Coastline: 0.86km
    - Possibility to abort launch and attempt a water landing
    - Minimizing risk from falling debris
  - City: 51.43km
    - Reducing danger to population dense areas



Section 4

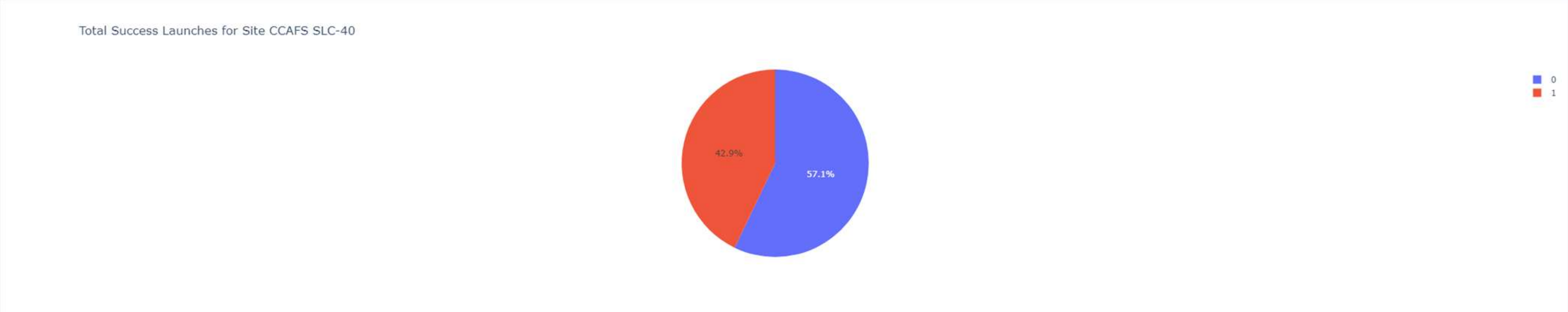
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



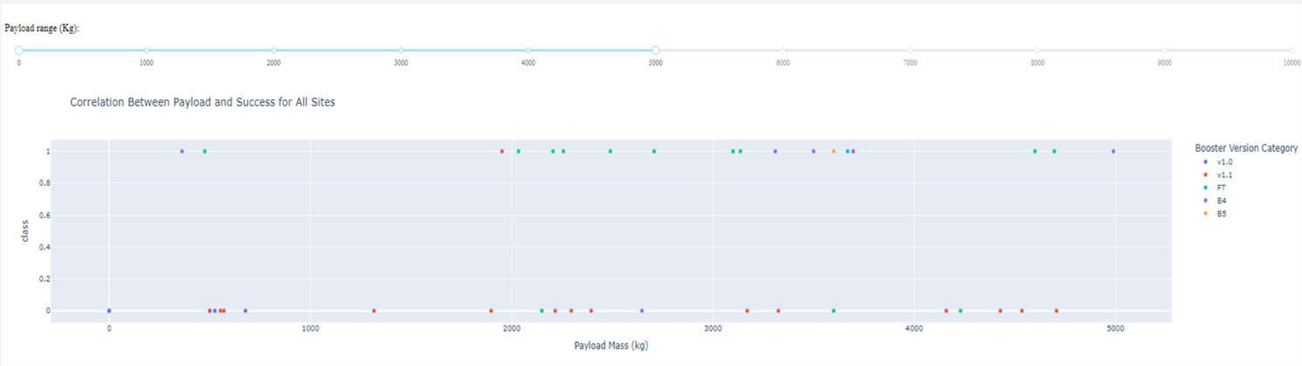
- Through the dropdown menu was made a selection for success launches for all sites. The pie chart displays the distribution of successful Falcon 9 first stage outcomes between the different launch sites where can be observed that KSC LC-39A launch site has the most successful launches with 41,2% of the total.

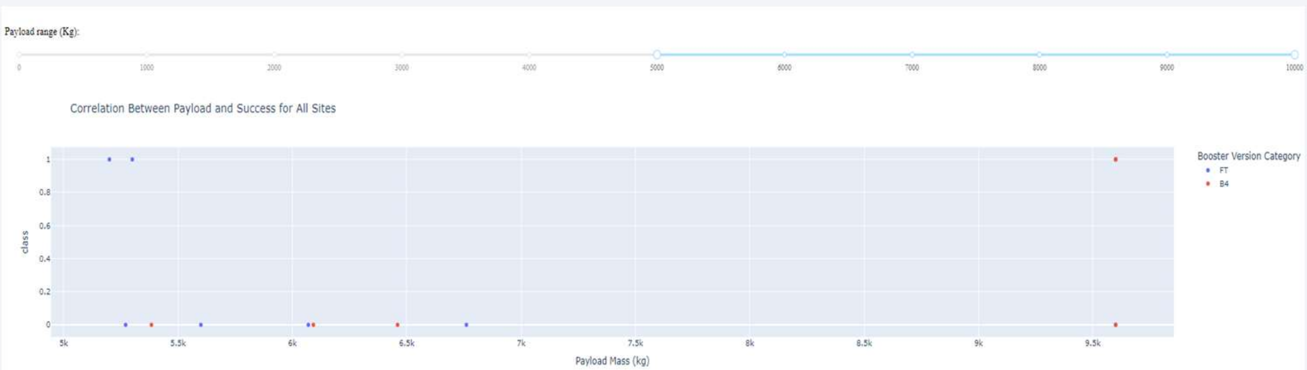
Launch Site with Highest Launch Successful Ratio



- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42,9%).
- Falcon 9 first stage successful landings are indicated by the '1' Class (red wedge in the pie chart) and failed landings by the '0' Class (blue wedge in the pie chart).

Payload Mass vs. Launch Outcome





- Payload Mass vs. Launch Outcome scatter plots for All Sites, with different payload selected in the range slider.
- The payload range from about 2000kg to 5000kg has the largest success rate.
- The 'FT' booster version category has the largest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Entire DataSet

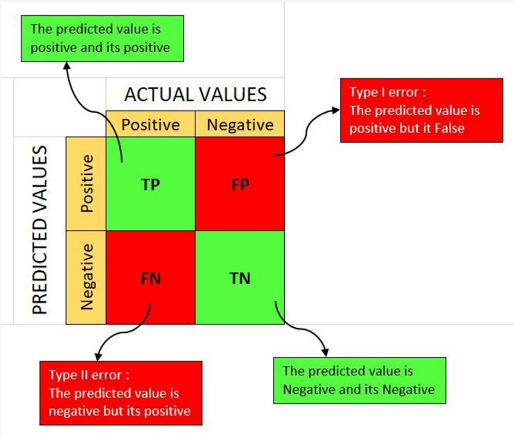
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.826087	0.819444
F1_Score	0.909091	0.916031	0.904762	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

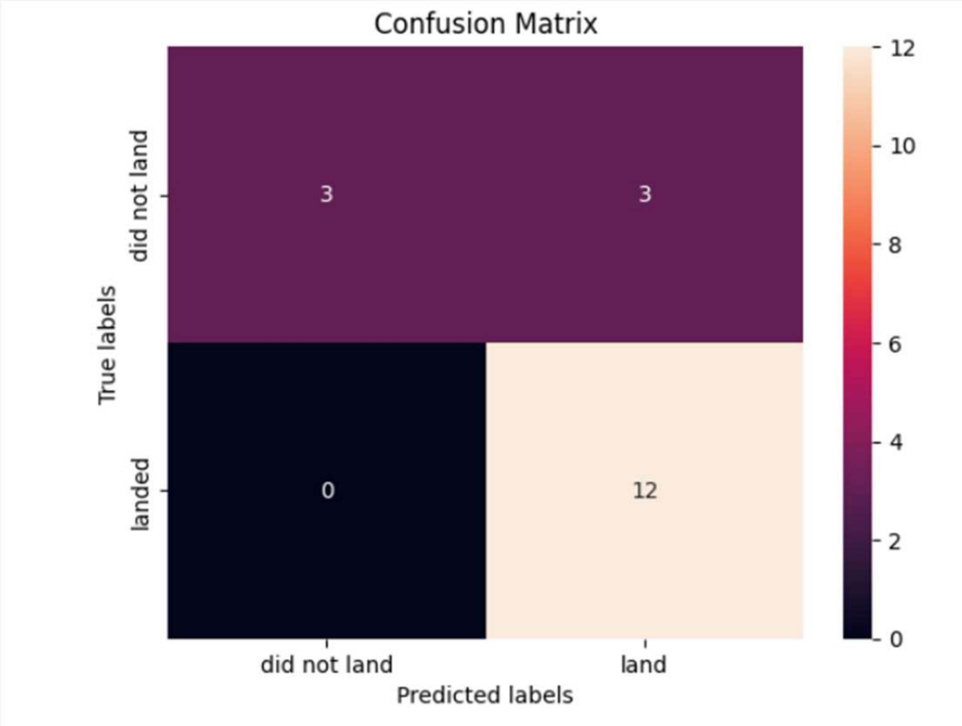
EXPLANATION

- Based on the scores from the Test Set, we cannot determine which method performs best.
- The similar Test Set scores may be attributed to the small sample size (18 samples). Therefore, we evaluatedall methods using the entire Dataset.
- The results from the entire Dataset confirm that the best model is the SVM (Support Vector Machine) model. This model not only has the higher scores but also the highest accuracy.

Confusion Matrix

- Here is the confusion matrix for the Logistic Regression Model.
- Prediction Breakdown:
  - 12 True Positives and 3 True Negatives
  - 3 False Positives and 0 False Negatives
- The figure below shows how Confusion Matrix should be read





Conclusions

- The success rate of launches has improved over the years.
- KSC LC-39A has the highest launch success rate among the sites.
- Most launch sites are located near the Equator Line, and all are very close to the coast.
- Orbits ES-L1, GEO, HEO, and SSO all have 100% success rate.
- Launches with lower payload mass yield better results compared to those with higher payload mass
- The Machine Learning Models can be used to predict future SpaceX Falcon 9 first stage landing outcomes.

Appendix

- Datasets utilized in this project:
  - SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
  - Wikipedia (Web Page): 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-DS0321EN-SkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-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 DashBoard Python File:
  - Data Collection GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
  - Web Scrapping GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb
  - Data Wrangling GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
  - EDA with Data Visualization GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/edadataviz.ipynb
  - EDA with SQL GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
  - Folium Maps GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb
  - Dashboard File GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/spacex_dash_app.py
  - Machine Learning GitHub URL: https://github.com/Thirocha1984/Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!