Applied Data Science Capstone

SIDDHANT KISHOR YEOLE

10/09/2022

Outline

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

Executive Summary

- As a part of this project, data on rocket launches was gathered from the public SpaceX API and the SpaceX Wikipedia page. Label creation was used to classify landings as successful or otherwise. The data exploration was followed by analysis using SQL, visualization in the form of folium maps and plotly dashboards. Features were chosen from the gathered data and one-hot encoding was used to convert data into a categorical variable form. In the final step, GridSearchCV was used to find the best parameters for all the machine learning models. The accuracy for all models was also compared as a final step.
- A total of four ML models were created - Logistic Regression, Support Vector Machine, Decision Tree Classifier and K-Nearest Neighbours. The results for each model were similar with an accuracy score of 83.33% for all the models. Minor differences were found in the accuracy score when evaluated on the test data.

Introduction

- The background for this project is the resurgence of the space age and the growing involvement of the global economy in space invention and travel.
- SpaceX has become the flag-bearer of space missions with the best pricing of 62 million USD for a single launch compared to the global average of 162 million USD
- The cost reduction is majorly attributed to the ability of SpaceX to recover the first stage of their rocket and reuse it over a period of time
- As a part of this project, SpaceY is competing with SpaceX and would like to analyze the reasons behind the cost cutting applied by SpaceX
- Problem:

SpaceY will use a machine learning model which will be using a carefully analyzed dataset for predicting the possibility of a successful stage-1 recovery for the rockets.

Section
1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from the SpaceX public API and also from the SpaceX Wikipedia page
- Perform data wrangling
  - To identify the landings - they were categorized as successful or unsuccessful in this step
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different models for predictive analysis were tuned using GridSearchCV

Data Collection

- The data collection was completed using a combination of public API requests to the SpaceX public API along with web scraping applied to a table in the SpaceX Wikipedia.

- The SpaceX API Data was divided into the following columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

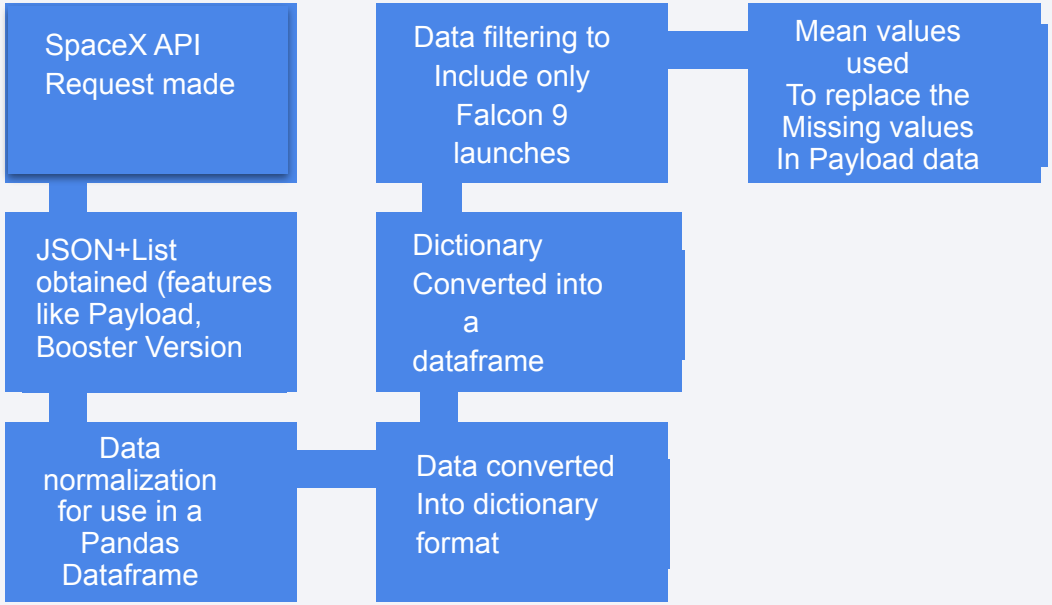
- The web-scraped Wikipedia data contains the following columns:

Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Version, Booster, Booster Landing, Date, Time

Data Collection – SpaceX API

- The flowchart for SpaceX API Calls is as follows:

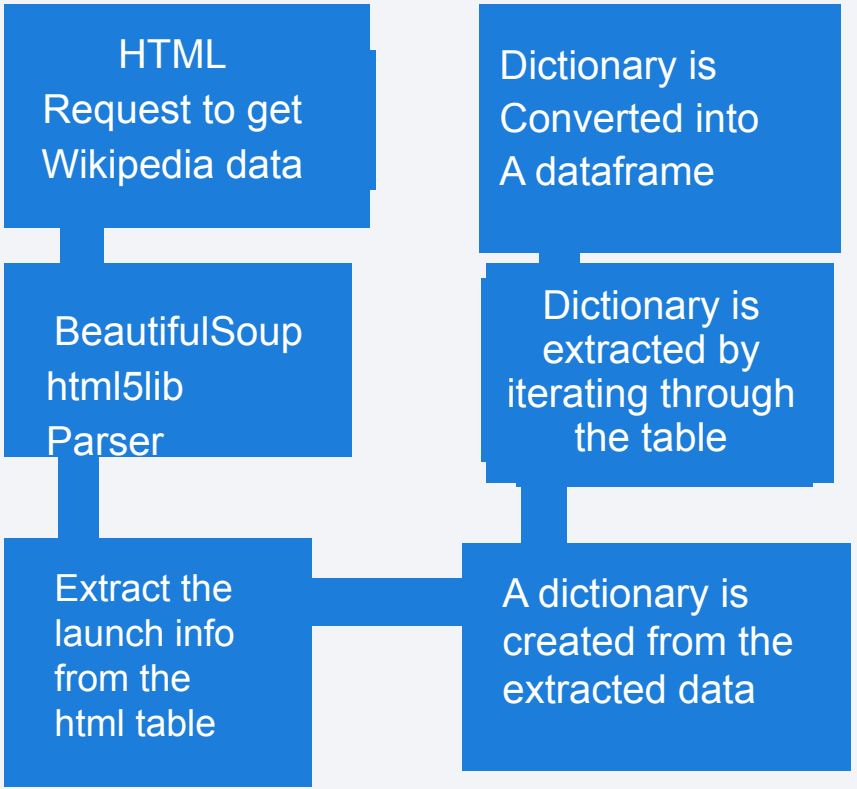
- Github: https://github.com/Daredevil0712/Applied_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- The web-scraping is performed using HTML requests and the BeautifulSoup library

- Github: https://github.com/Daredevil0712/Applied_Data_Science_Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- To categorize data, a training label was created using binary representation. A successful landing is 1 while an unsuccessful landing is denoted as 0.
- The outcomes are split based on - Mission Outcome and Landing location
- A new training label was created - ‘class’ which is set to 1 if ‘Mission Outcome’ is ‘True’
- The value are mapped as follows:
  - True ASDS, True RTLS, & True Ocean – set to - 1
  - None None, False ASDS, None ASDS, False Ocean, False RTLS – set to 0

- Github: https://github.com/Daredevil0712/Applied_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis was performed on the following variables:
- Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Different Plots Used included:
- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
- Scatter plots, line charts, and bar plots were used to compare relationships between variables or to check if a relationship exists so as to use the features in training the machine learning model.

- Github: https://github.com/Daredevil0712/Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- After loading the data set into the IBM db2 database, data was queried using SQL Python integration.
- Queries are written to extract feature-specific data to understand the dataset properly
- Information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes was extracted using the SQL queries in this part of the project.

- Github: https://github.com/Daredevil0712/Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium maps are used to interactively mark Launch Sites, Proximity of the sites to key locations and the nature of landings (successful/ unsuccessful)
- These maps help us understand the reason behind the location of launch sites and related operations in specific areas across the country.
- Github: https://github.com/Daredevil0712/Applied_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The plotly dashboard consists of a pie chart and a scatter plot.
- Pie chart shows the distribution of successful landings across all launch sites. We can also view individual landing statistics across launch sites.
- The scatter plot is used to visualize the relation between launch sites and the payload mass.
- The pie chart is used to visualize the success rate of various launch sites.
- The scatter plot can also help us understand the success rate across launch sites, payload mass and booster versions.
- Github:

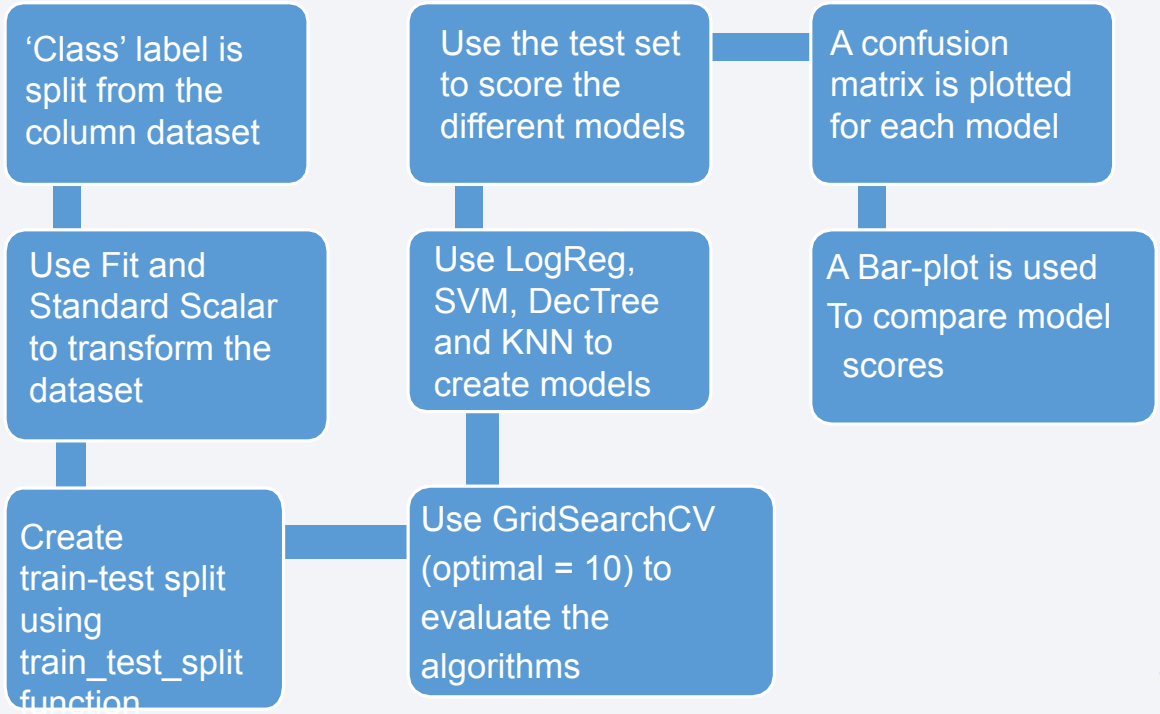
  https://github.com/Daredevil0712/Applied_Data_Science_Capstone/blob/main/Plotly%20Dash%20dashboard_%20spacex_dash_app.py

Predictive Analysis (Classification)

- Four different machine learning methods were used for predictive analysis

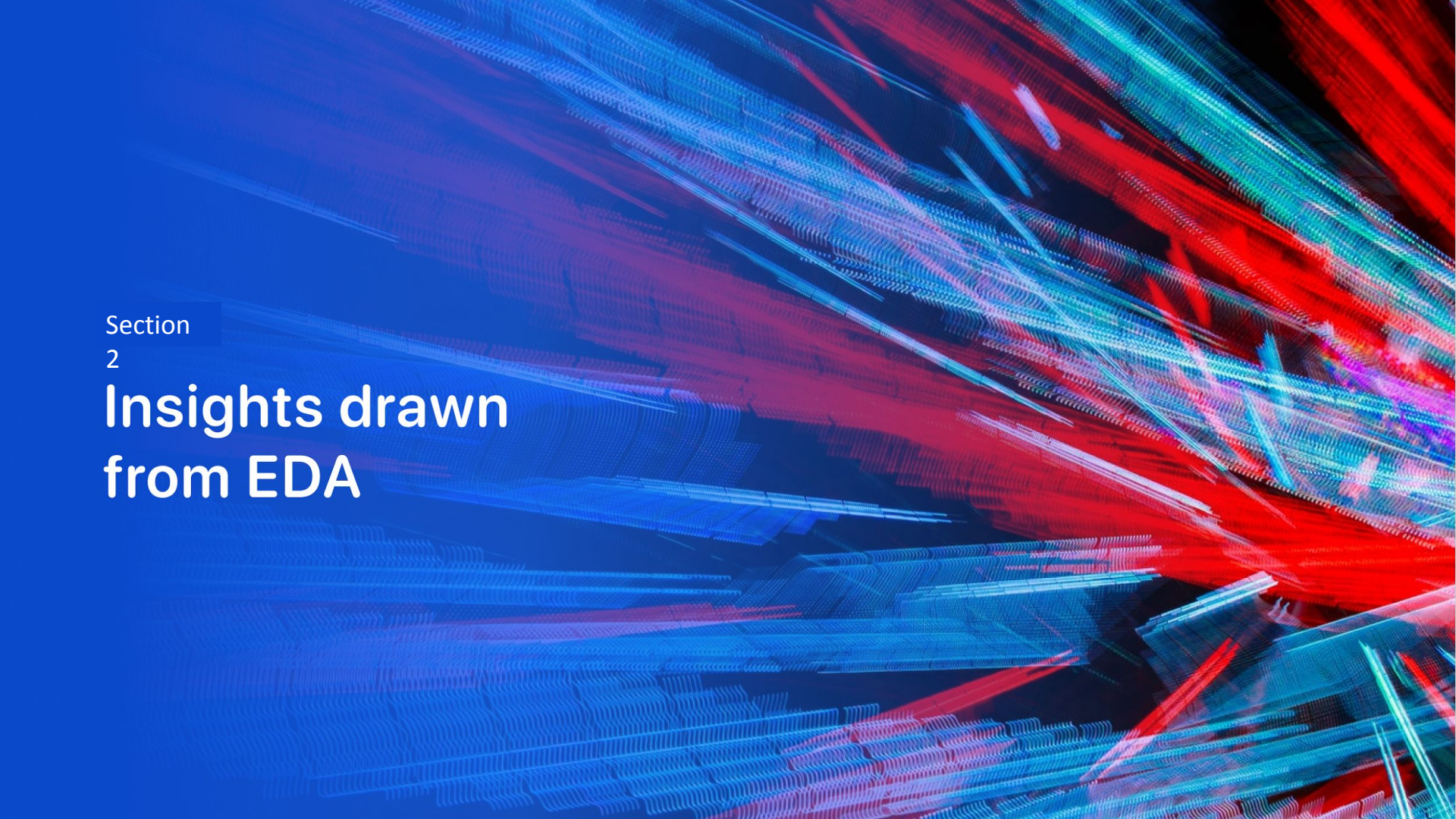
- Github:

  https://github.com/Daredevil0712/Applied_Data_Science_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

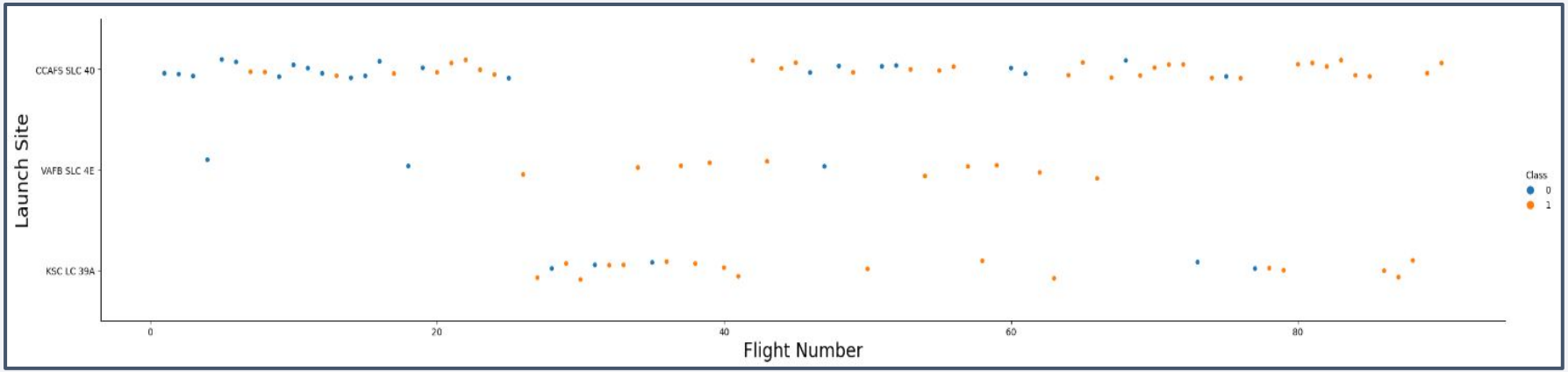
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

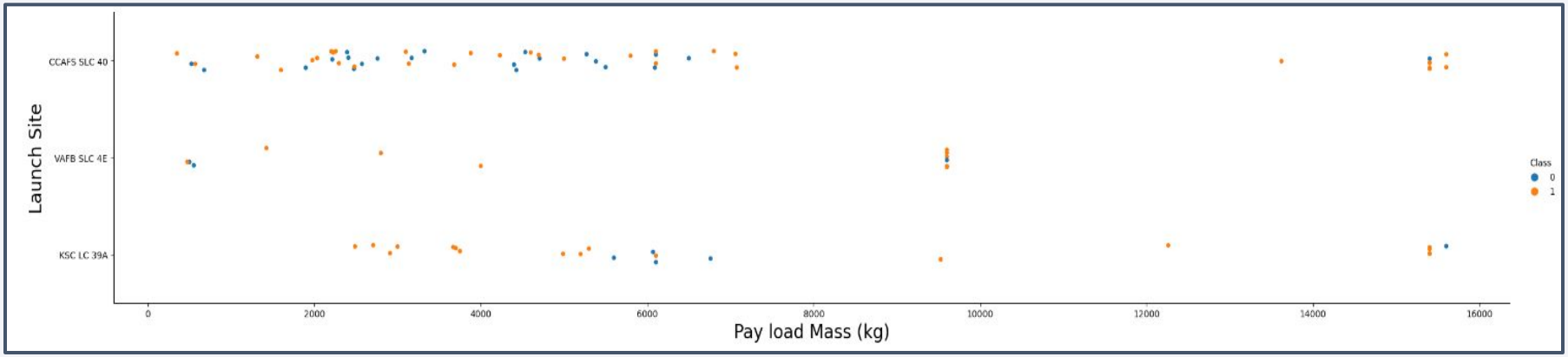
Flight Number vs. Launch Site



The Blue dots refer to unsuccessful launches and yellow signify a successful launch.

- Over time, the launches observe more success as the number of flights increase
- CCAFS appears to be the site with most launches and also has the highest number of successful launches.
- Overall, the trend shifts towards successful launches after the initial 20 flights.

Payload vs. Launch Site

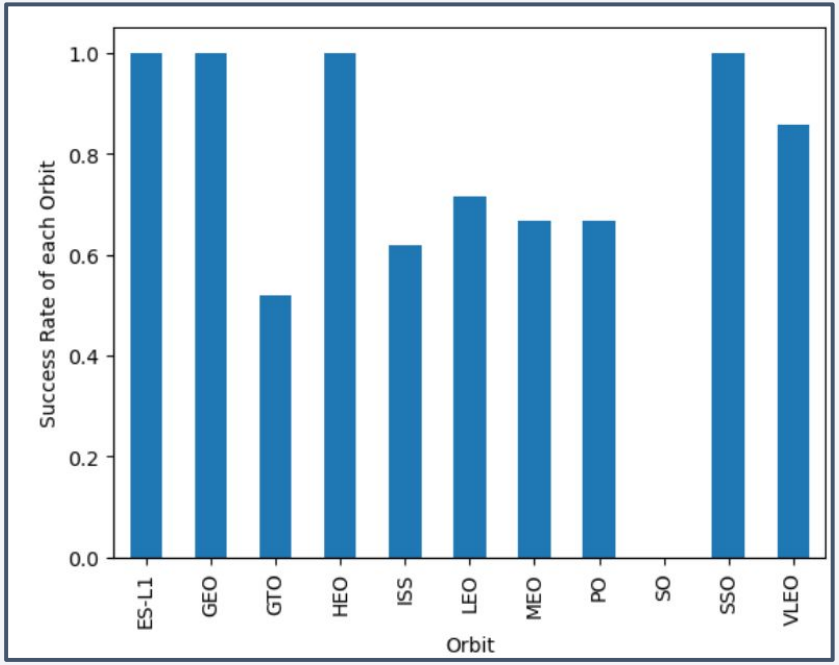


The Blue dots refer to unsuccessful launches and yellow signify a successful launch.

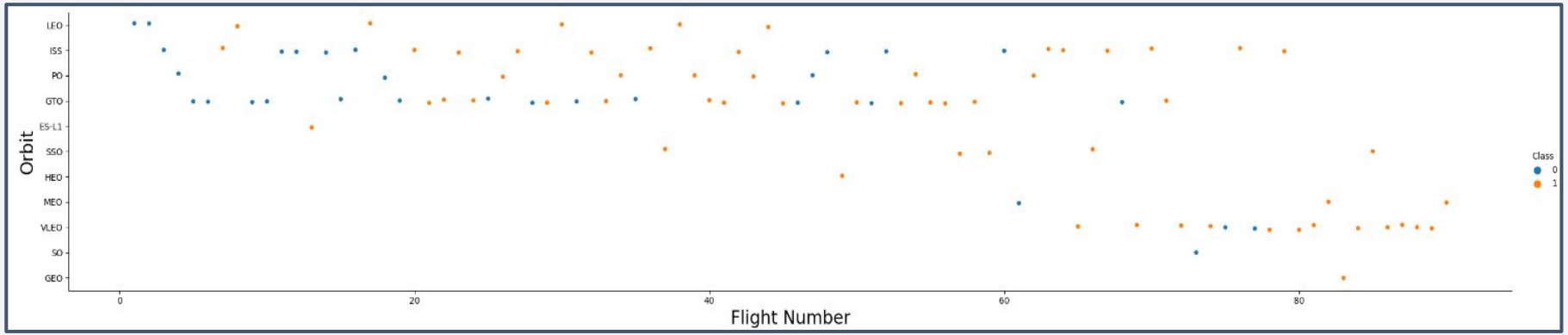
- The most frequent payload observed at launch sites is concentrated in the range of 0-6000 kg.
- The VAFB site appears to be favored for heavier payloads beyond 8000 kg.

Success Rate vs. Orbit Type

- 4 Orbits have the maximum success rate of 100% - ELS-1, GEO, HEO, SSO
- VLEO has an appreciably high success rate of around 80%
- The SO orbit does not see any success in the launch
- While the GTO has a success rate around 50%, the number of flights launches in this orbit are relatively higher compared to the other cases.



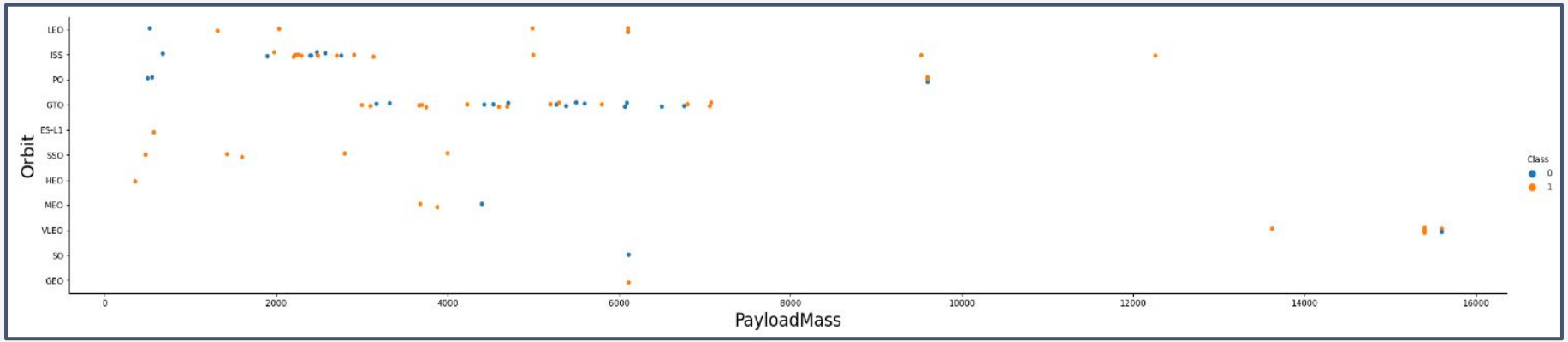
Flight Number vs. Orbit Type



The Blue dots refer to unsuccessful launches and yellow signify a successful launch.

- Initially, SpaceX favored a lot of Lower Earth Orbits while eventually switching to Medium and VLEO.
- As expected at the start, the number of unsuccessful launches is higher.
- The best performance is observed in sun-synchronous orbits.

Payload vs. Orbit Type

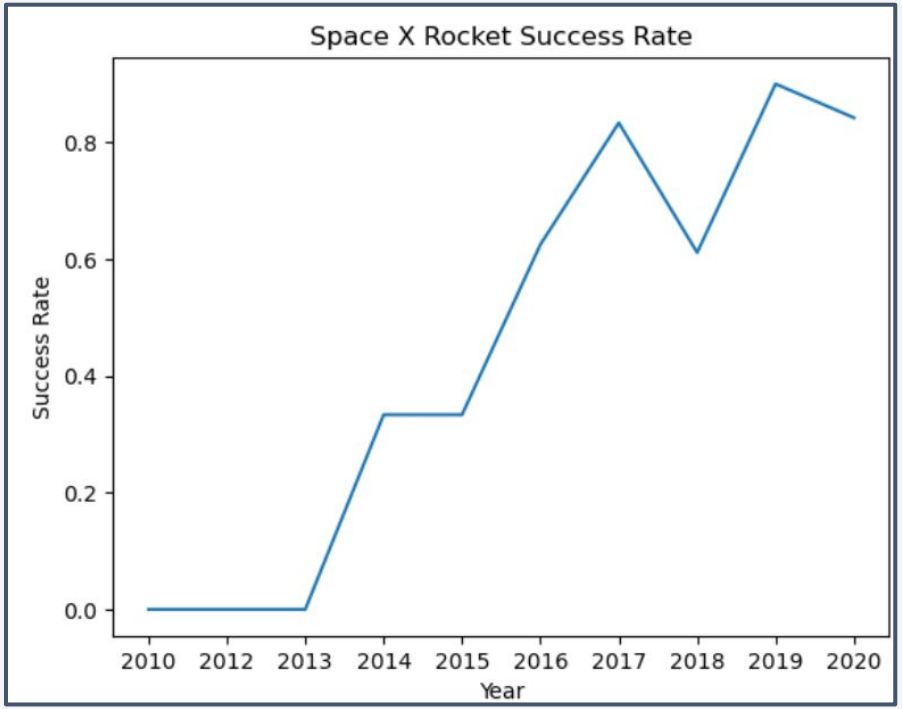


The Blue dots refer to unsuccessful launches and yellow signify a successful launch.

- The payload is majorly restricted in the range of 0-6000kg.
- The highest payload is sent in VLEO.
- LEO has mostly low payload mass

Launch Success Yearly Trend

- The rate of success for the first few years is very low.
- From 2015, the success rate observes a significant jump by almost double (40% to 80%)
- A slight reduction was observed in 2018
- The recent trend has seen the success rate hover around 80% for the rocket launches.



All Launch Site Names

```
In [ ]: %sql SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL

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

Out[ ]:

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

Query to obtain Unique Launch Site Names

- A total of 4 unique launch site names are obtained from this query

- It looks like the CCAFS SLC-40 is a new launch site and CCAFS LC-40 is the older location.

Launch Site Names Begin with 'CCA'

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

In [10]:
```
%sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

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

Out[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The query selects the launch sites with name starting with ‘CCA’ using a delimiter for obtaining 5 records from the entire database

Total Payload Mass

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

In [13]:
```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
```

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

Out[13]:

SUM("PAYLOAD_MASS__KG_")
45596

- The total Payload Mass carried by boosters launched by NASA (CRS) is 45596 kg.
- CRS are supplies sent to the International Space Station (ISS)

Average Payload Mass by F9 v1.1

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

In [14]:
```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

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

Out[14]:

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

- The average payload mass by F9 v1.1 is 2534.66 Kgs

- The value is on the lower end of our general payload mass for the launches.

First Successful Ground Landing Date

```
In [23]: %sql SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'

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

Out[23]:

MIN("DATE")
01-05-2017

- The first successful ground landing was observed on 01 May 2017.

- The first three years from 2012-2014 did not have any successful landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [25]:
```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

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

Out[25]:

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

- 4 drone ship landings were completed successfully with a payload between 4000-6000 kgs.

Total Number of Successful and Failure Mission Outcomes

In [26]:
```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

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

Out[26]:

SUCCESS	FAILURE
100	1

- The success to failure ratio for missions is extremely high upto 99%. Thus, it is clear that most of the failures are intended and well-managed.

Boosters Carried Maximum Payload

In [28]:
```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

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

Out[28]:

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

- A total of 12 boosters carried the maximum payload mass
- All the rockets belong to the same series F9 B5 B10xx.x

2015 Launch Records

In [31]:
```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

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

Out[31]:

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- Two failed drone ship landings were seen in 2015 - one in the month of January and the second in the month of April

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

In [34]:

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

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

Out[34]:

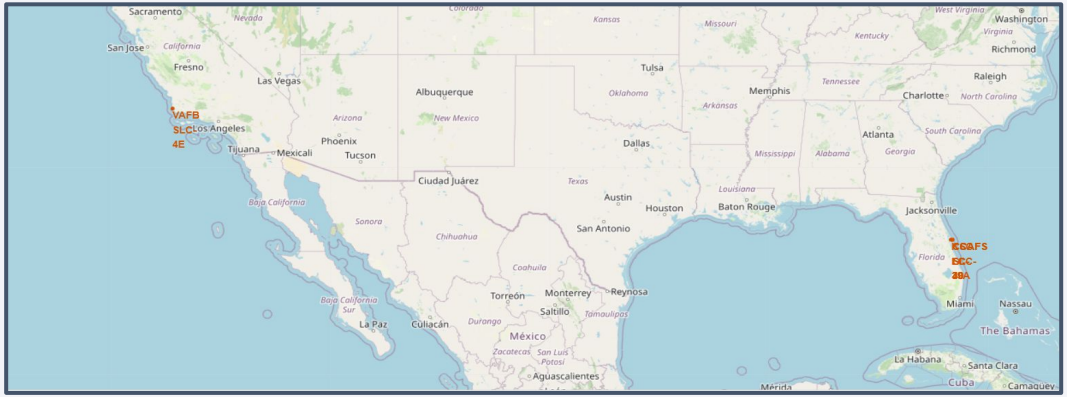
Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

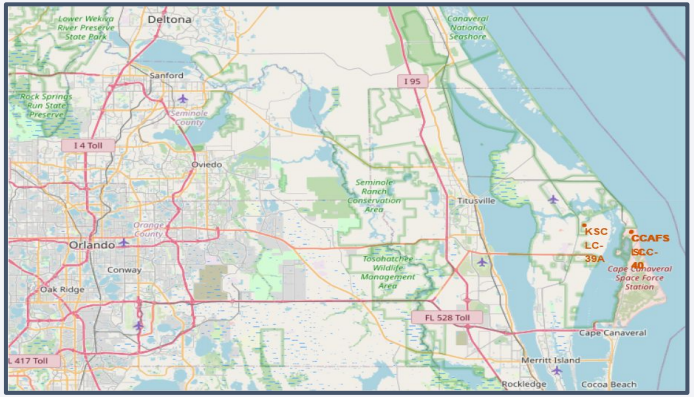
- A total of 20 successful landings were observed in the specified period over a course of 7 years.
- The landings were almost equally successful on drone ships as well as ground pads.

Section 3

Launch Sites Proximities Analysis

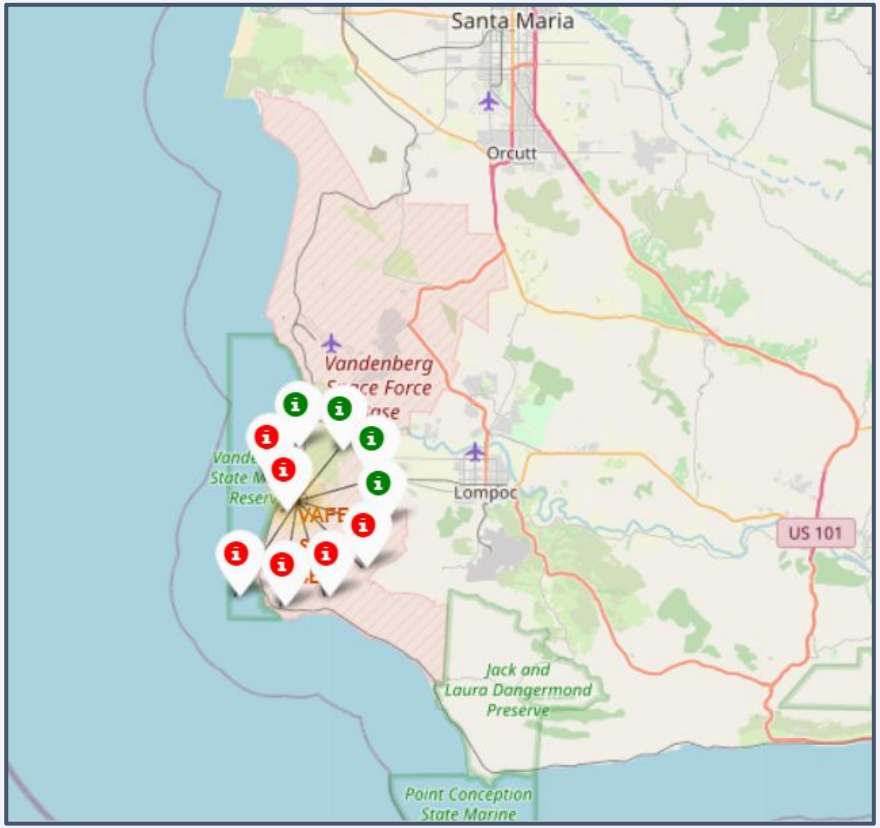
Launch Site Locations





- Of the two maps, the left one shows the launch site locations across USA. On the right, we see a picture of the multiple launch sites located in Florida within a short distance of each other.

Launch Markers

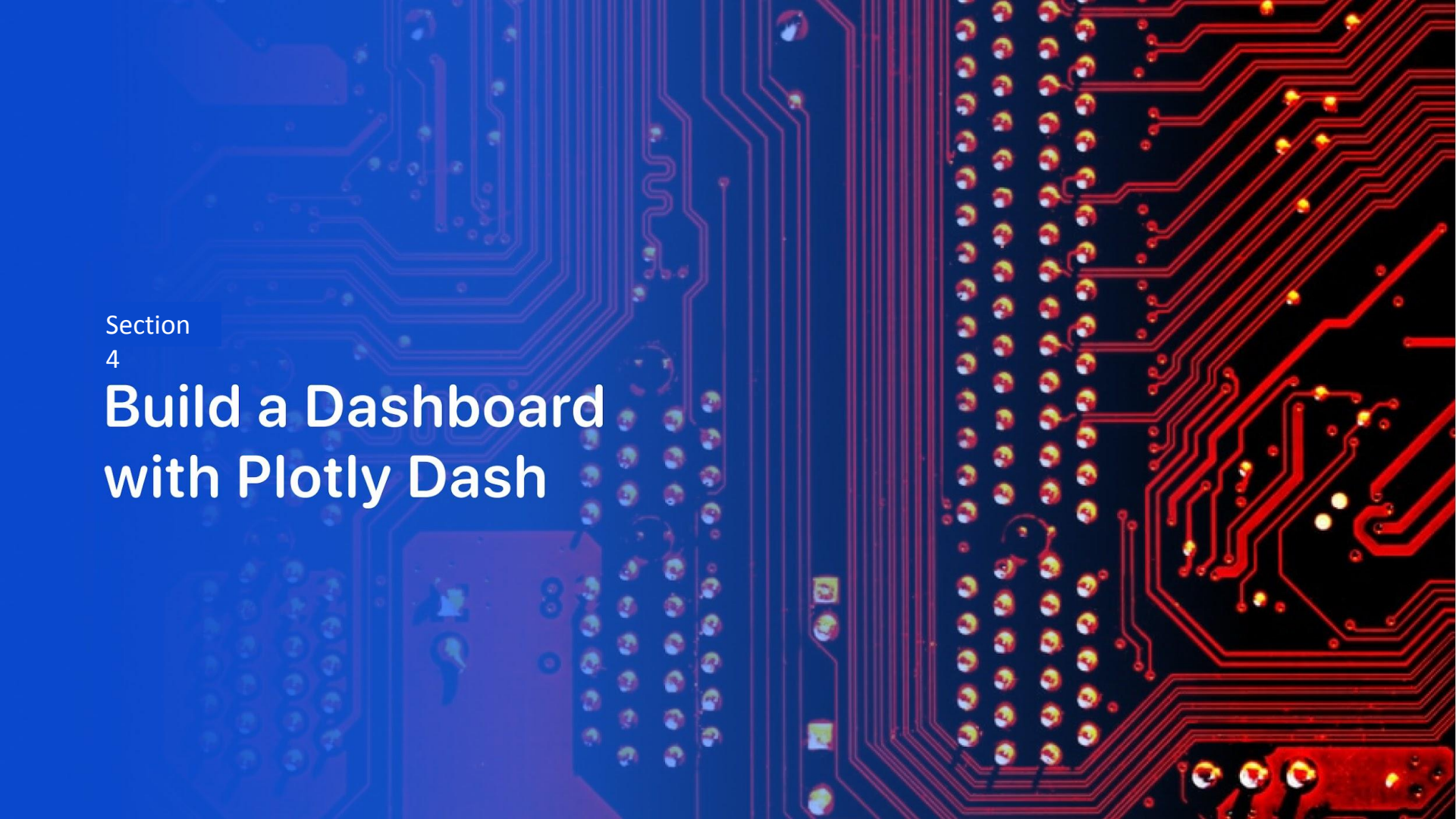


- Every launch site can be viewed in the form of color-coded markers where red represents a failed landing and green markers indicate a successful attempt.

- The screenshot is of the Vandenberg Space force base on the West Coast of USA in California.

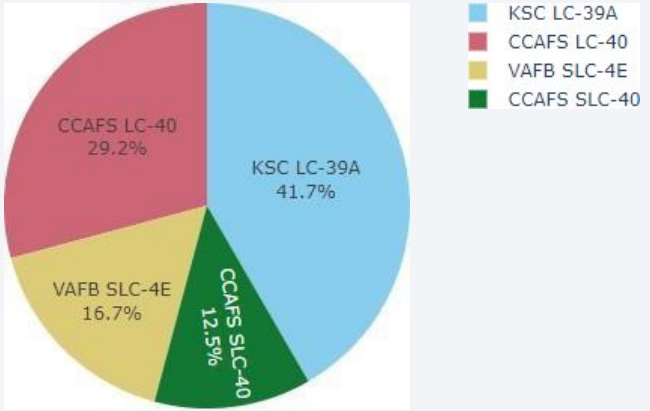
Key Location Distance Plots

- As we can see the launch sites are located in close proximity to public transport and the coast to allow easy availability and transport of resources.
- In the screenshot, the road is located at 600m while the coast is at 900m from the CCAFS-SLC site on the east coast in Florida.


Section
4
Build a Dashboard
with Plotly Dash

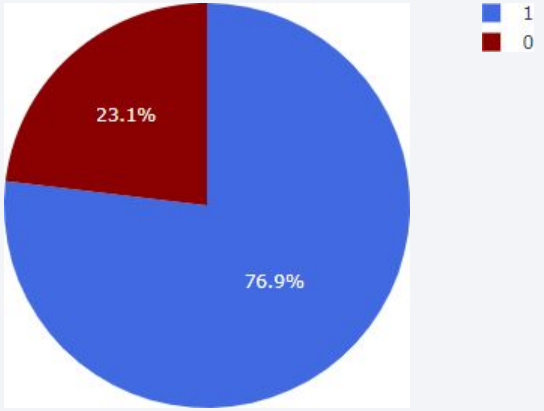
Successful Launches at different Sites

- The pie chart shows the distribution of successful launches across various launch sites.
- CCAFS LC-40 is the old name of CCAFS SLC-40.
- A majority of the successful landings are done prior to the renaming of the site.
- VAFB has the smallest share of successful landings. The reasons include smaller sample size and the location close to the coast.



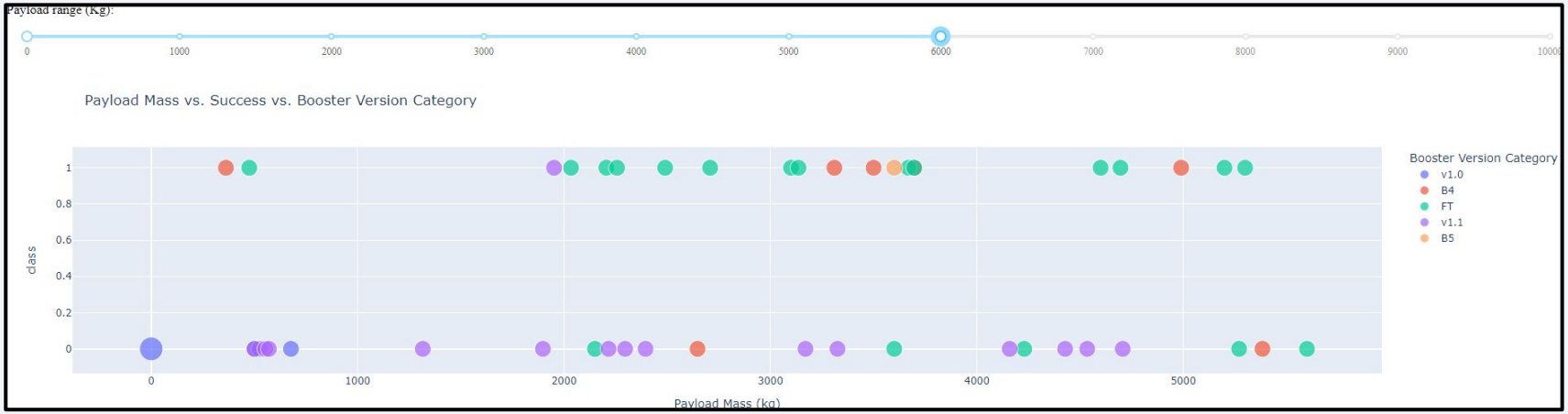
Highest Success Rate

- The site KSC LC 39-A has the highest success rate indicated by the binary 1 based on our encoding among all other launch sites.
- 76.9% of the landings are successful at this location



KSC LC-39A Success Rate (blue=success)

Payload Mass vs. Success vs. Booster Version



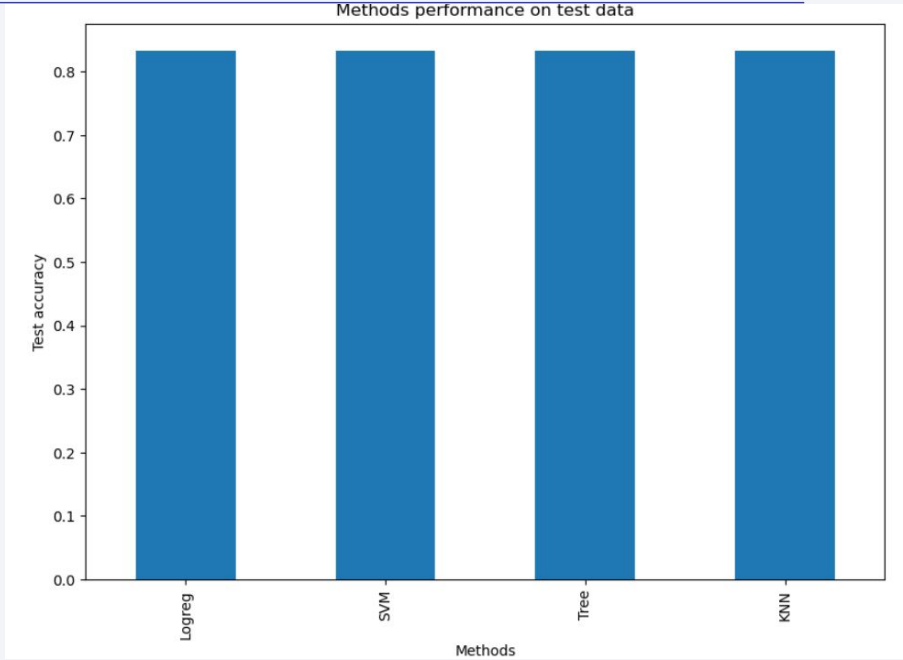
- Plotly dashboard has a Payload range selector. The range is limited to 10000 kgs instead of the max payload value of 15600 kgs.
- Class 1 and 0 indicate successful and failed landings respectively.
- The scatter plot represents booster version and number of launches in the form of the color and point size of the plot points respectively.

Section
5

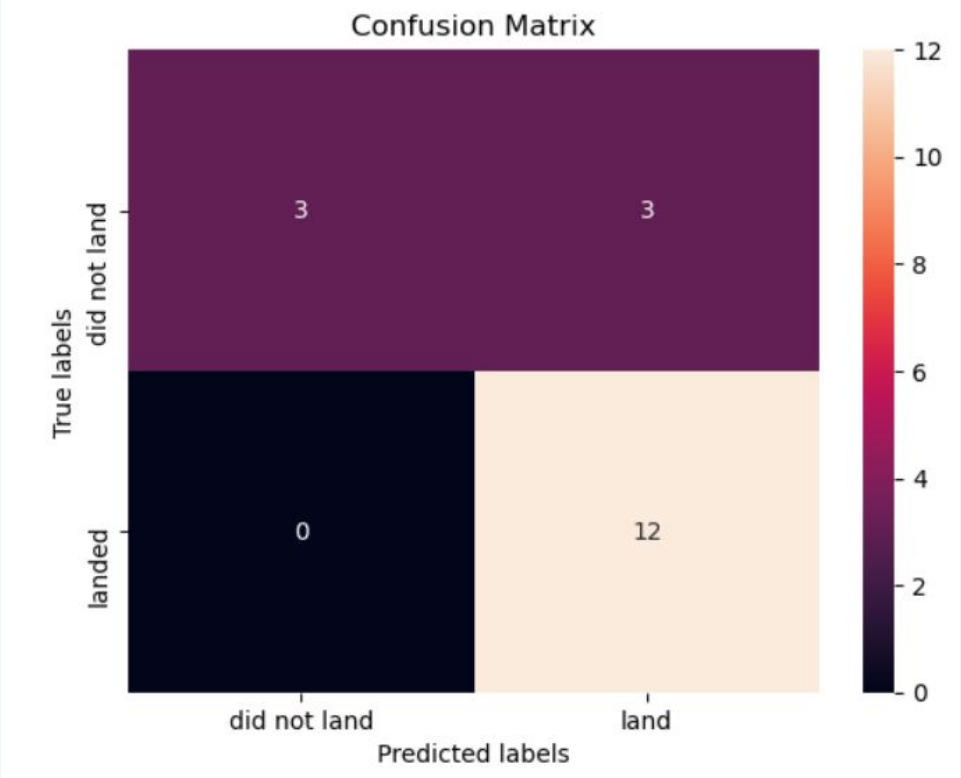
Predictive Analysis (Classification)

Classification Accuracy

- The accuracy score for all four methods - Logistic Regression, Support Vector Machine (SVM), Decision Tree and K-nearest Neighbours is similar with a value of 83.33%
- One of the reasons for this result could be the small sample of the test set, with more data, a more definitive conclusion could be obtained.



Confusion Matrix



- As models perform similarly on the test data, a single confusion matrix can be used for all the machine learning models used in prediction.

Conclusions

- The goal of this project was to create an effective plan based on a machine learning model for a company SpaceY to bid against SpaceX while saving 100 million USD on space launches.

- The data required for this exercise was obtained using the SpaceX API and wikipedia.
- The data was cleaned, sorted and then stored in a tabular database.
- Visualization was completed with tools such as maps, plots and dashboards to understand the characteristics of this data.
- In the final step, a machine learning model was built with 4 diverse algorithms and performed at an accuracy of 83%

- Based on these findings, SpaceY could determine if a successful Stage-1 landing is possible before the launch to help the company save a lot of resources and cost, thus allowing it to compete with SpaceX and bid against them in the journey to space exploration.

Appendix

- Github Repository URL

  https://github.com/Daredevil0712/Applied_Data_Science_Capstone

Thank you!