IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Umamaheswaran.V

Feb 17,2025

Outline

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

Executive Summary

- SpaceY is a new commercial rocket launch provider who wants to bid against SpaceX
- SpaceX advertises launch services starting at $62 million for missions that allow some fuel to be reserved for landing the 1st stage rocket booster,so that it can be reused
- SpaceX public statements indicate a 1st stage Falcon 9 booster to cost upto $ 15 million to build without including R&D cost recoupment or profit margin
- Given mission parameters such as payload mass and desired orbit,the models produced in this report were able to predict the first stage rocket booster landing successfully with an accuracy level of 84%
- As a result,SpaceY will be able to make more informed bids against SpaceX by using 1st stage landing predictions as proxy for the cost of a launch

Introduction

- This report has been prepared as part of the Applied Data Science Capstone course

- In this course,I take the role of Data Scientist working for a new rocket company SpaceY

- With the help of the data science findings and models in this report,SpaceY will be able to make more informed bids against SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

For this report the data science methodology used as can be outlined as :

- Data collection methodology
- Data wrangling method
- Exploratory data analysis (EDA) :
  - Exploratory data analysis (EDA) using visualization and SQL
- Data Visualization :
  - Interactive visual analytics using Folium and Plotly Dash
- Model Development :
  - Predictive analysis using classification models

Data Collection

- [Data Collection](#) :
- API :
    - Acquired data from open source RestAPI
        - using GET method the SpaceX launch data has been parsed
        - Filtered the dataframe to only include Falcon 9 launches
        - Replaced missing payload mass values from classified missions with mean

Data Collection - Scraping

- [Web Scraping](#) :
    - Acquired historical launch data from ‘https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches’
    - Requested the Falcon 9 Launch Wiki page from its Wikipedia URL
    - Extracted all column/variable names from the HTML table header
    - Parsed the table and converted it into a Pandas dataframe

2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

[hide] Flight No.	Date and time (UTC)	Version, Booster[h]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 △ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 △ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule,[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage – the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 △ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 △ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing the second stage engine twice. The first stage booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.								
82	7 March 2020, 04:50[506]	F9 B5 △ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 △)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
	Last launch of phase 1 of the CRS contract. Carries *Bartolomeo*, an ESA platform for hosting external payloads onto ISS.[508] Originally scheduled to launch on 2 March 2020, the launch date was pushed back due to a second stage engine failure. SpaceX decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, the third flight of the Dragon C112 and the last launch of the cargo Dragon spacecraft.								
83	18 March 2020, 12:16[510]	F9 B5 △ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were reused (Starlink flight in May 2019).[511] Towards the end of the first stage burn, the booster suffered premature shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit.[512] This was the second Starlink launch booster landing failure in a row, later revealed to be caused by residual cleaning fluid trapped inside a sensor.[513]								
84	22 April 2020, 19:30[514]	F9 B5 △ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

Data Wrangling

- Data Wrangling :

- Explored data to determine the label for training supervised models
    - Calculated the number of launches on each site
        - The data contains launch facilities :
            - Cape Canavarel Space Launch Complex 40 VAFB SLC 4E
            - Vandenberg Air Force Base Space Launch Complex 4E (SLC-4E)
            - Kennedy Space Center Launch Complex 39A KSC LC 39A
        - The location of each Launch is placed in LaunchSite
        - Using value_counts() on LaunchSite to determine the number of Launches on each site

```
LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: count, dtype: int64
```

Data wrangling (contd)

- Calculated the number and occurrence of each orbit
- Calculated the number and occurence of mission outcome of the orbits
- Created a landing outcome label from Outcome column
  - Training Label : 'Class'
    - Class = 0; first stage booster did not land successfully
    - None None ; not attempted
    - None ASDS; unable to be attempted due to launch failure
    - False ASDS; drone ship landing failed
    - False Ocean; ocean landing failed
    - False RTLS; ground pad landing failed
  - Class = 1; first stage booster landed successfully
    - True ASDS; drone ship landing succeeded
    - True RTLS; ground pad landing succeeded
    - True Ocean; ocean landing succeeded

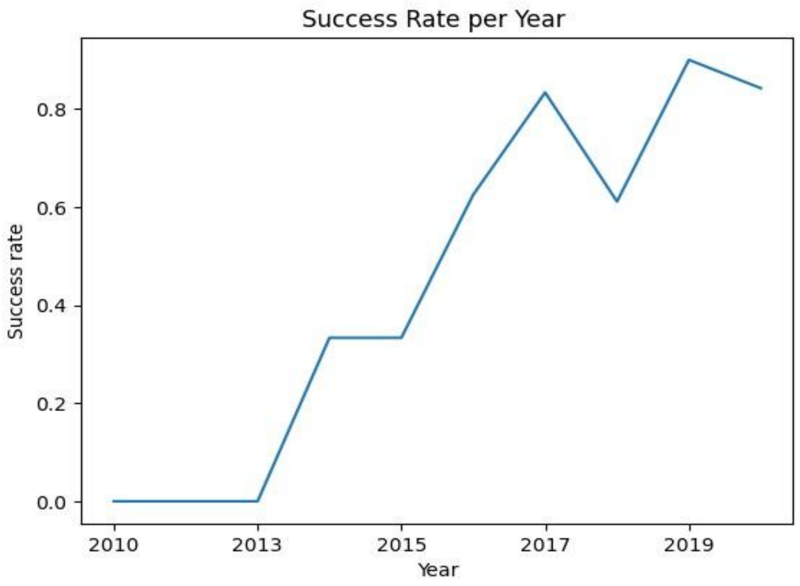
Landing Outcomes
Sample Size = 90
❏ = Class 0
❏ = Class 1

True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
False Ocean	2
None ASDS	2
False RTLS	1

EDA with Data Visualization

- EDA with Visualization :
    - Read the dataset into a Pandas dataframe
        - Plot Libraries : Matplotlib and Seaborn
            - FlightNumber to PayloadMass
            - FlightNumber to LaunchSite
            - Payload to LaunchSite
            - Orbit type to Success rate
            - FlightNumber to Orbit type
            - Payload to Orbit type
            - Year to Success rate

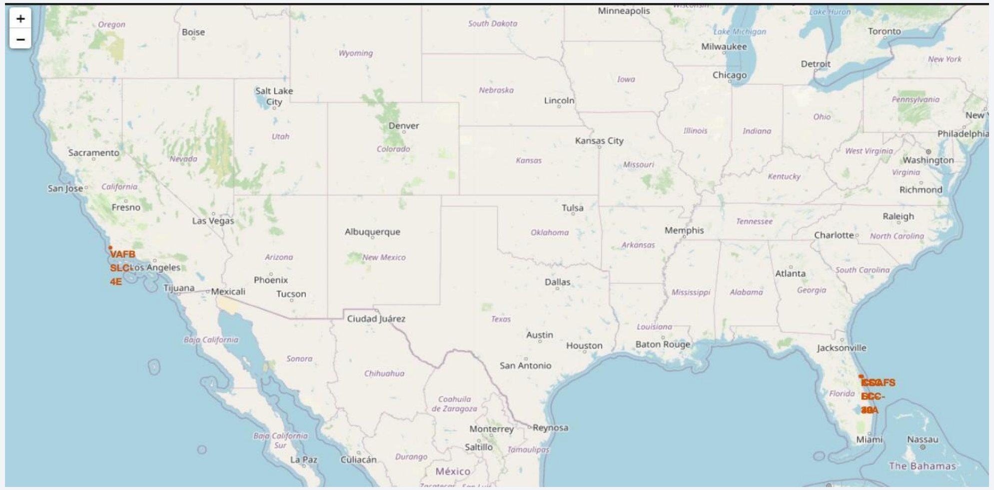
```python
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate

_line = df.groupby("Date")
_line['Class'].mean().plot(kind = 'line',xlabel = 'Year',ylabel = 'Success rate',title = 'Success Rate per Year ')
plt.show()
```



EDA with SQL

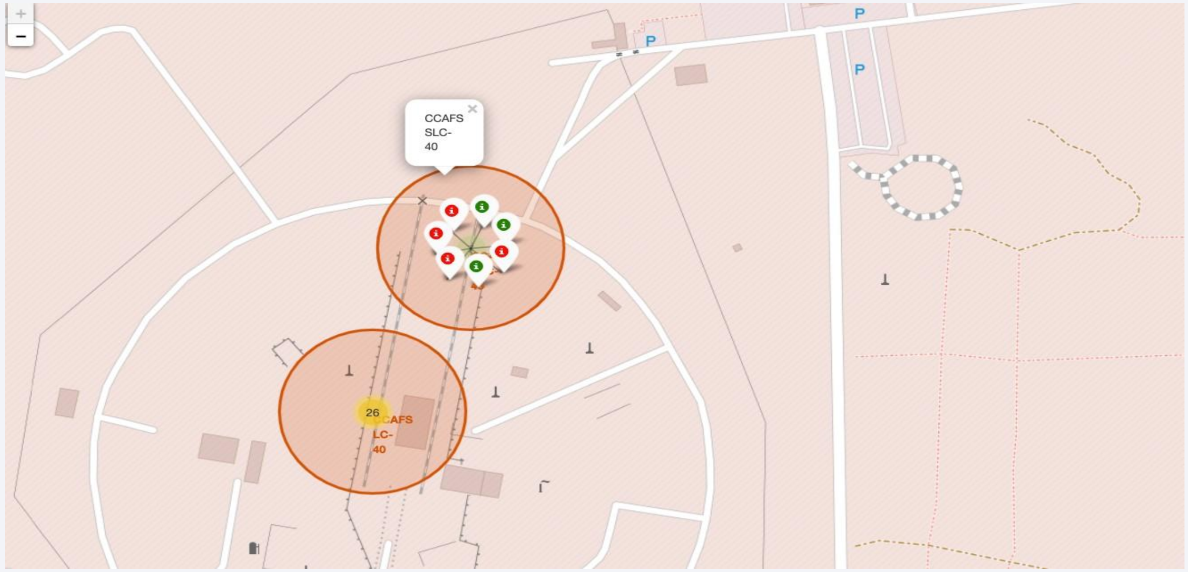
- [EDA with SQL](#) :
    - EDA with SQL
    - Loaded data into an IBM DB2 instance
    - Ran SQL queries to display and list information about
        - Launch sites
        - Payload masses
        - Booster versions
        - Mission outcomes
        - Booster landings

Build an Interactive Map with Folium



[Interactive Visual Analytics with Folium](#) :

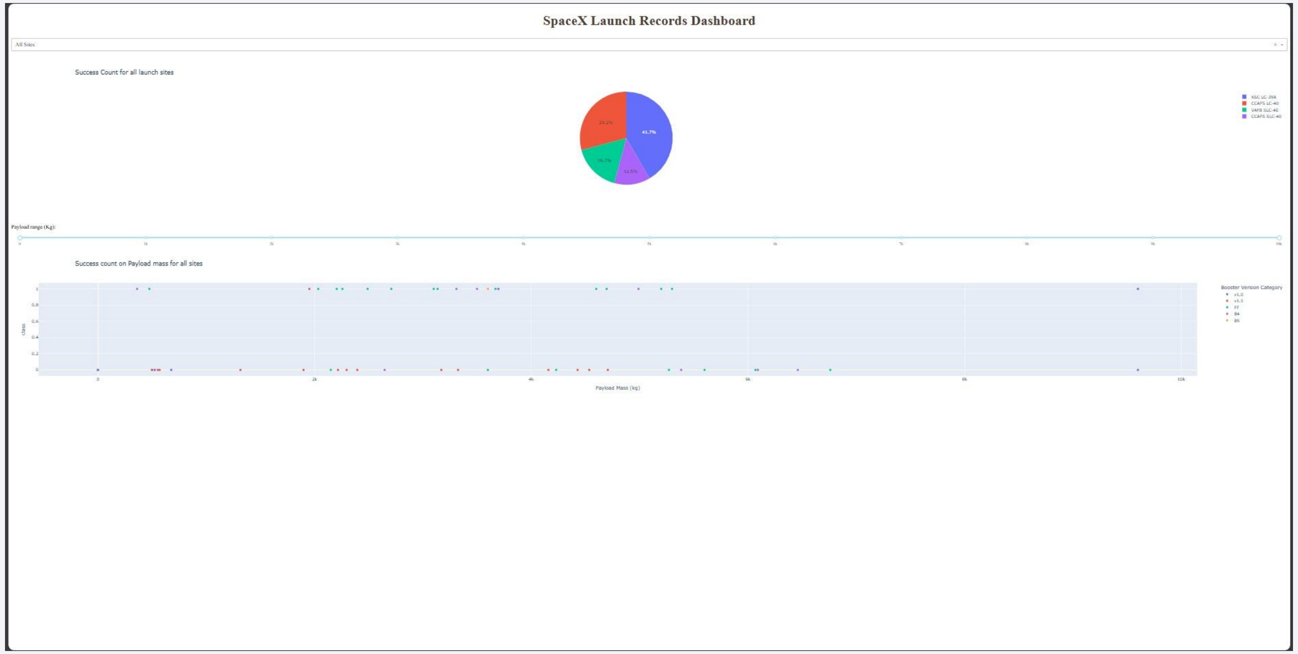
- Marked all launch sites on map
- Marked all success/failure for each site
- Calculated the distance between a launch site to its proximities
  - Railways
  - Highways
  - Coastlines
  - Cities



Build a Dashboard with Plotly Dash

Dashboard with Plotly Dash :

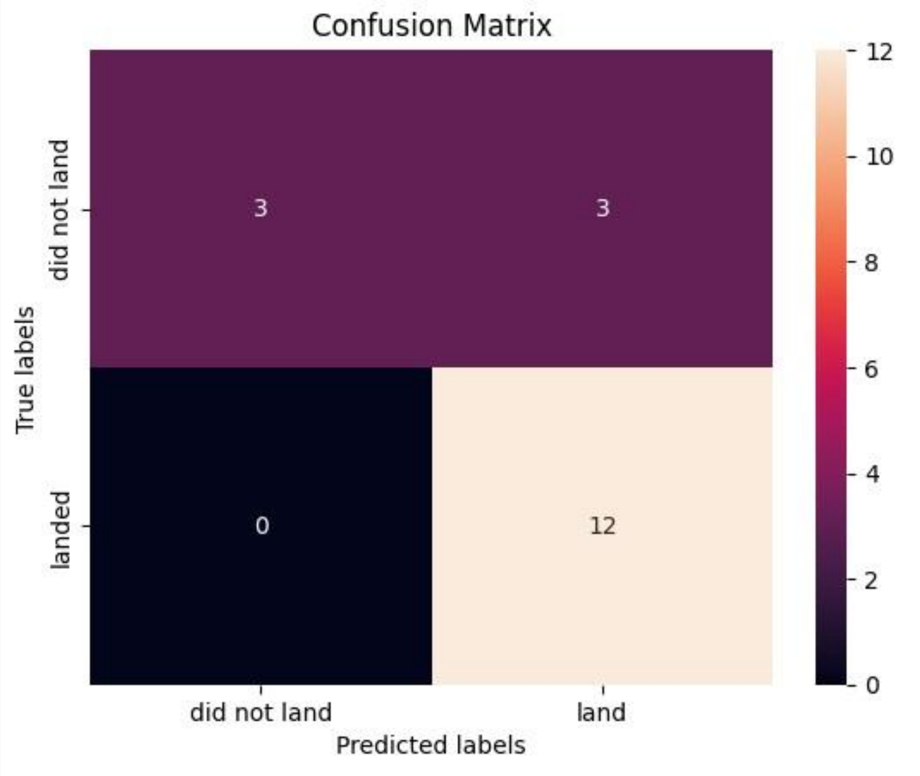
- Added a Launch Site Drop-down Input Component
- Added a callback function to render success-pie-chart based on selected site dropdown
- Added a Range Slider to Select Payload
- Added a callback function to render the success-payload-scatter-chart scatter plot



Predictive Analysis (Classification)

Predictive Analysis (Classification) :

- Loaded the dataframe created during data collection
- Created a column for our training label 'Class' created during data wrangling
- Standardized the data
- Split the data into training data and test data
- Fit the training data to various model types
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors Classifier
- Used a cross validated grid search over a variety of hyperparameters to select the best ones for each model
- Enabled by Scikit learn library function GridSearchCV
- Evaluated accuracy of each model using test data to select the best model



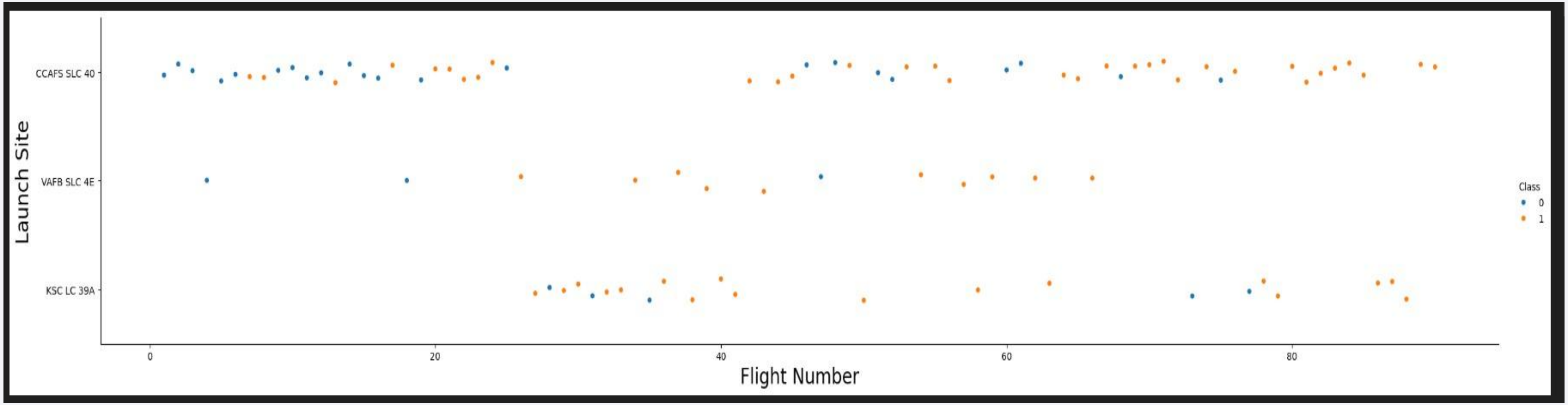
Results

- Exploratory data analysis results
    - Obtained insights about how each important variable would affect the success rate
- Interactive analytics demo in screenshots
    - Obtained the geographical patterns about launch sites using Folium
- Predictive analysis results
    - Obtained best Hyperparameter for SVM, Classification Trees and Logistic Regression and analyzed which method performs best using test data

Section 2

Insights drawn from EDA

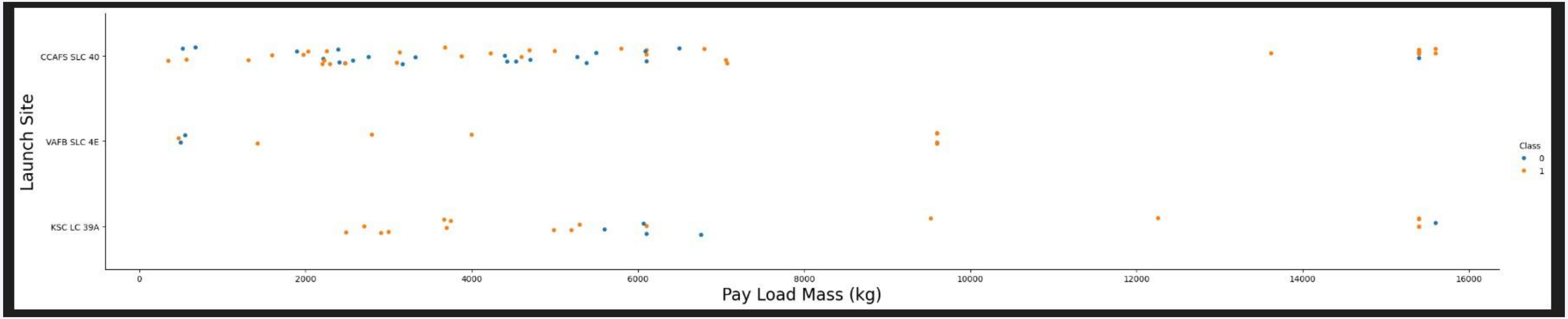
Flight Number vs. Launch Site



- By observation,CCAFS SLC 40 appears to be the most of early 1st stage failure took place

```python
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

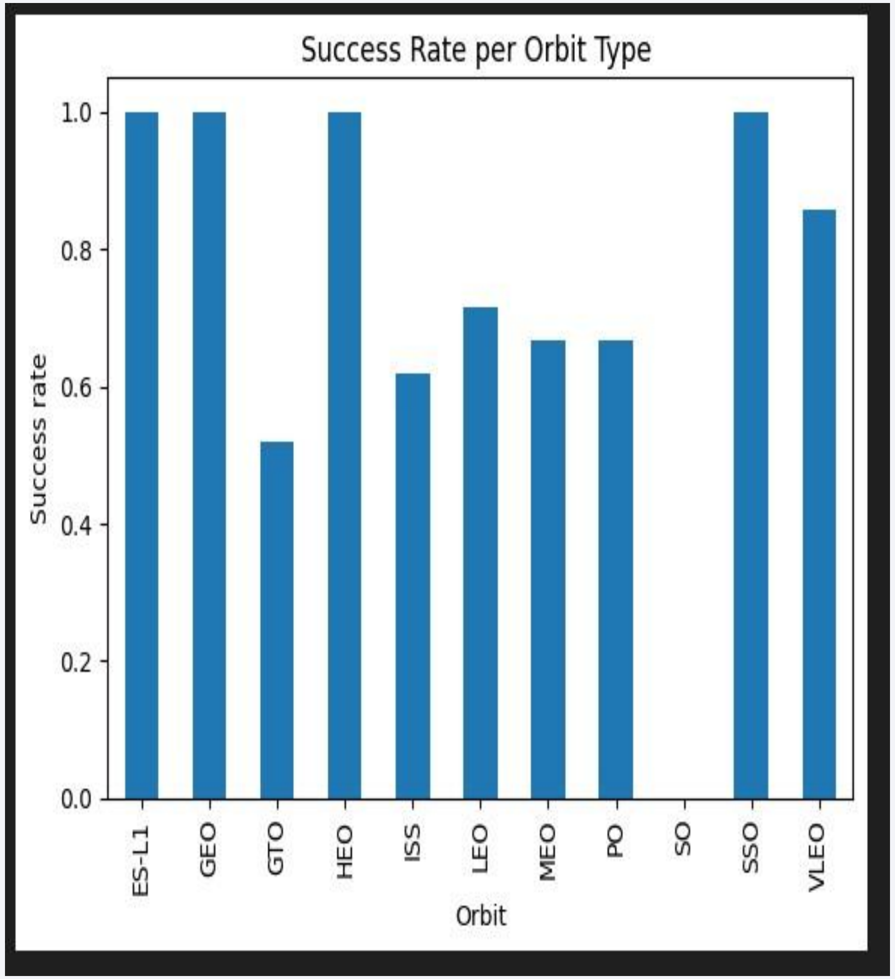
Payload vs. Launch Site



CCAFS SLC 40 and KSC LC 39 A are in favour of heavy payloads

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value

sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

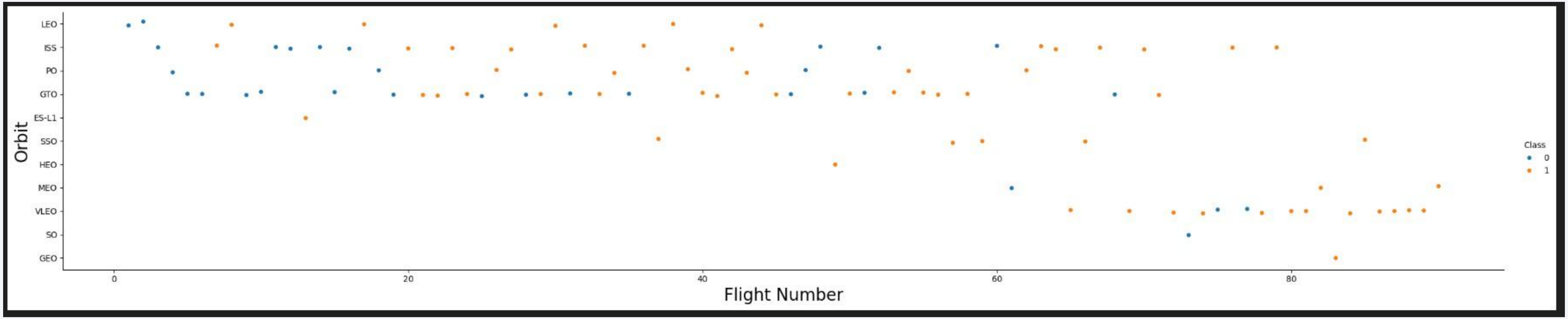
Success Rate vs. Orbit Type




```python
# HINT use groupby method on Orbit column and get the mean of Class column

_bar = df.groupby("Orbit")
_bar['Class'].mean().plot(kind = 'bar',xlabel = 'Orbit',ylabel = 'Success rate',title = 'Success Rate per Orbit Type ')
plt.show()
```

- Orbits otherthan SO have successful 1st stage landing

Flight Number vs. Orbit Type



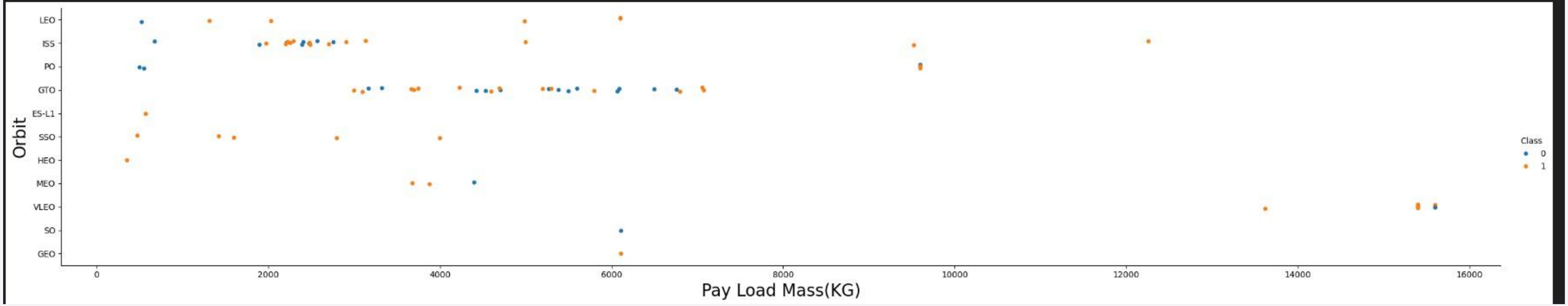

```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value

sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

Orbit between LEO and GTO are showing results proportional with the flight number

Payload vs. Orbit Type

```python
# Plot a scatter point chart with x axis to be Payload Mass and y axis to be the Orbit, and hue to be the class value

sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay Load Mass(KG)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

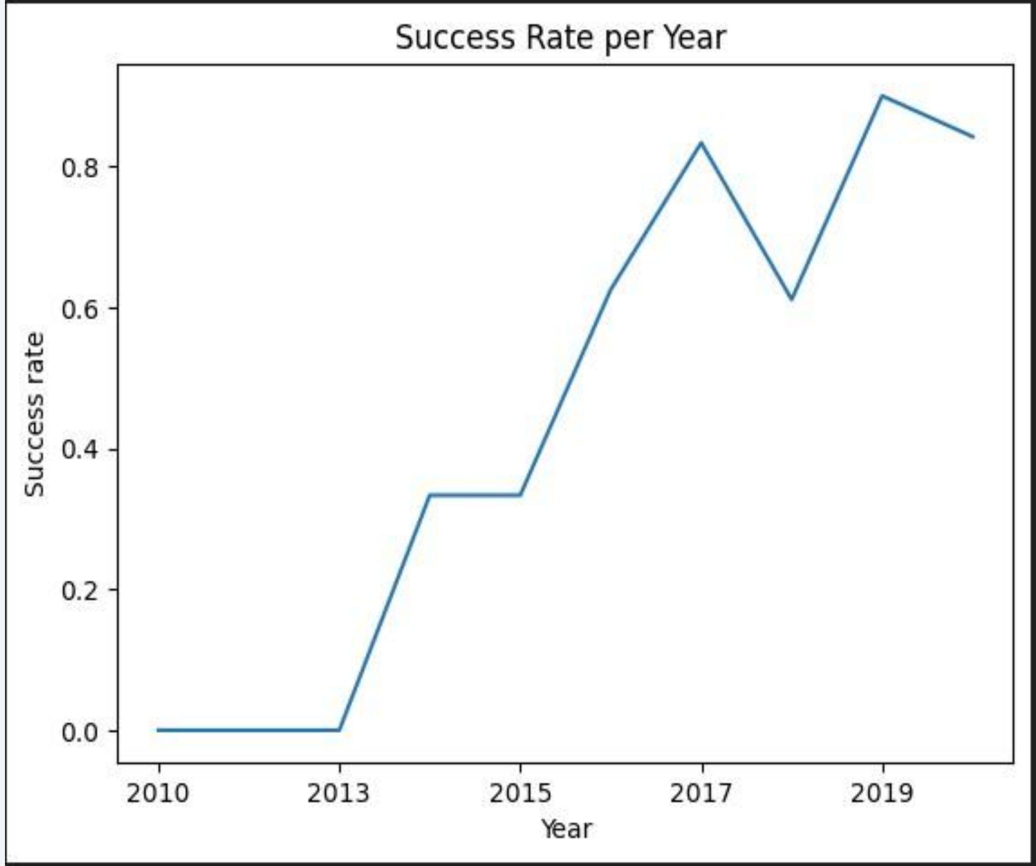
Python



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

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate

_line = df.groupby("Date")
_line['Class'].mean().plot(kind = 'line',xlabel = 'Year',ylabel = 'Success rate',title = 'Success Rate per Year ')
plt.show()
```



The sucess rate since 2013 kept increasing till 2020

All Launch Site Names

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

Using DISTINCT(LAUNCH_SITE) from SPACEX table obtained the all unique site names

select DISTINCT(LAUNCH_SITE) from SPACEXTBL

Launch Site Names Begin 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	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

Using the ‘like CCA%’ obtained the the name of launch site begin with CCa and using ‘limit’ keyword to filter the result to 5 values

select * from SPACEXTBL where LAUNCH_SITE like "CCA%" limit 5

Total Payload Mass

Total PAYLOAD mass carried by Boosters launched by NASA (CRS) is 45596 kg

Using the sum function to obtain the sum of payloadmass with the Customer is NASA (CRS)

select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER='NASA (CRS)'

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928.4 kg

Applied the avg function to pay load mass by filtering the booster value to F9V1.1 to extract the desired result

select avg(payload_mass__kg_) from spacextbl where Booster_Version = 'F9 v1.1'

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad is 2015-12-22

By applying min() function to date whose Landing outcome value is Success(ground pad)

select min(date) from spacextbl where Landing_Outcome = 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

The list of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are

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

Using the logical operator AND to apply multiple filter to extract the result

select booster_version from spacextbl where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ >4000 and payload_mass__kg_<6000

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes :

- 1 Failure (in flight)
- 99 Success
- 1 Success (payload status unclear)

Update SPACEXTBL set Mission_Outcome = 'Success' where Mission_Outcome = 'Success '

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass

- 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

select Booster_Version from spacextbl where Payload_Mass__kg_=(select max(payload_mass__kg_) from spacextbl)

2015 Launch Records

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

- 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

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

Landing_Outcome	QTY
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

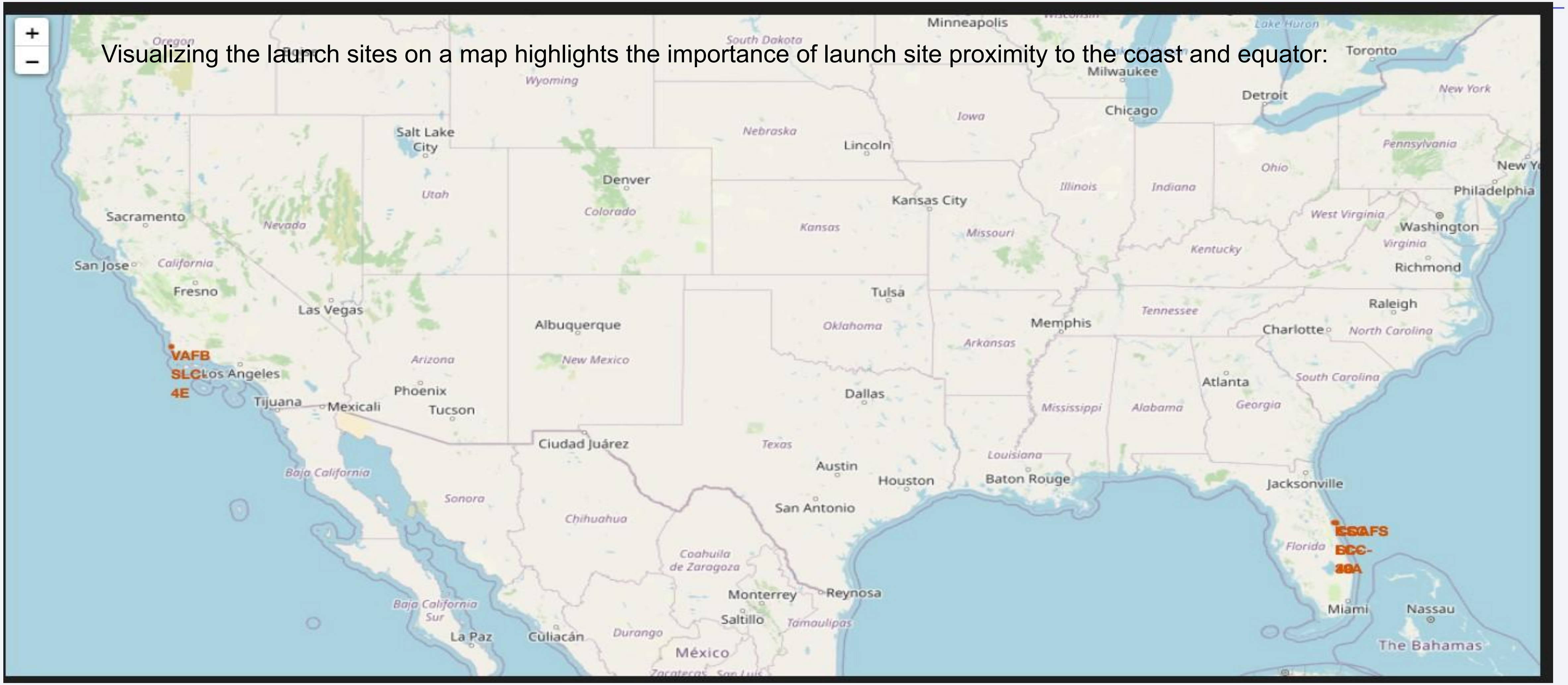
```
SELECT Landing_Outcome, COUNT(*) AS
QTY FROM SPACEXTBL WHERE Date
BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome ORDER BY
QTY DESC
```

Section 3

Launch Sites Proximities Analysis

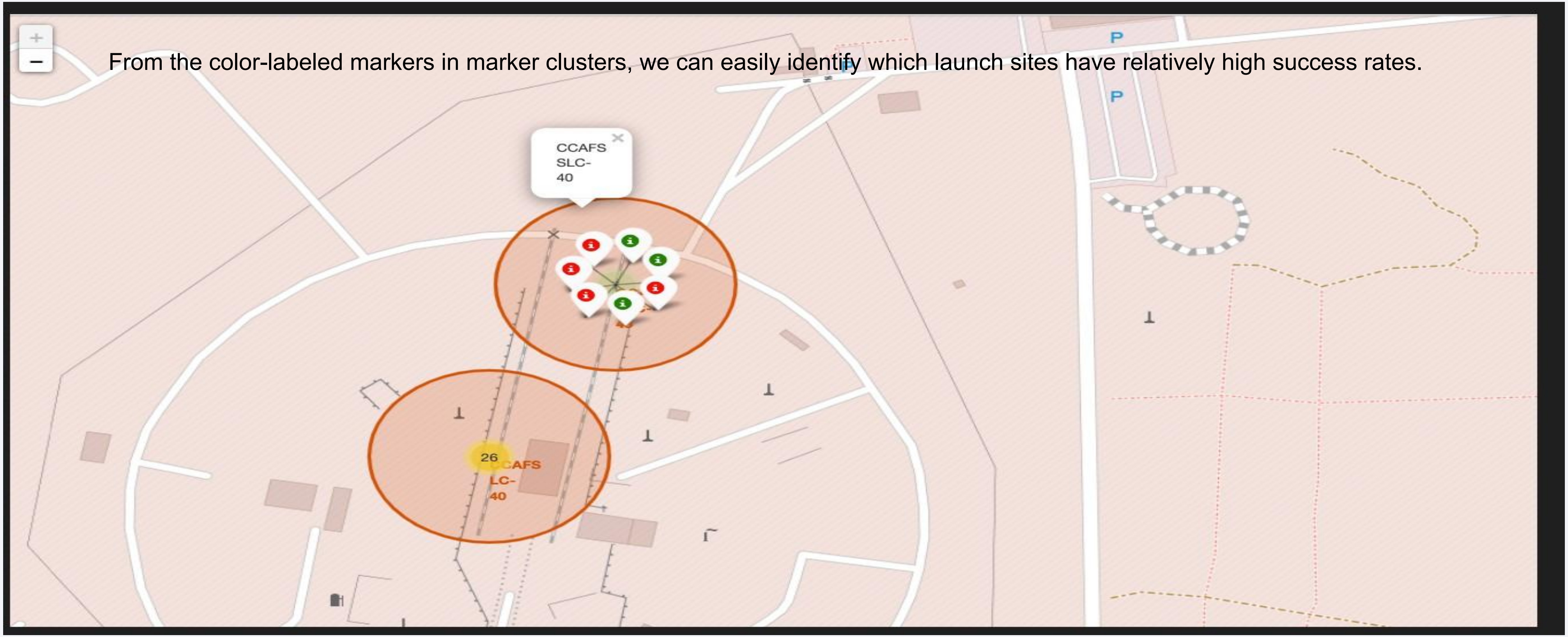
Launch Site Location Analysis :

Visualizing the launch sites on a map highlights the importance of launch site proximity to the coast and equator:

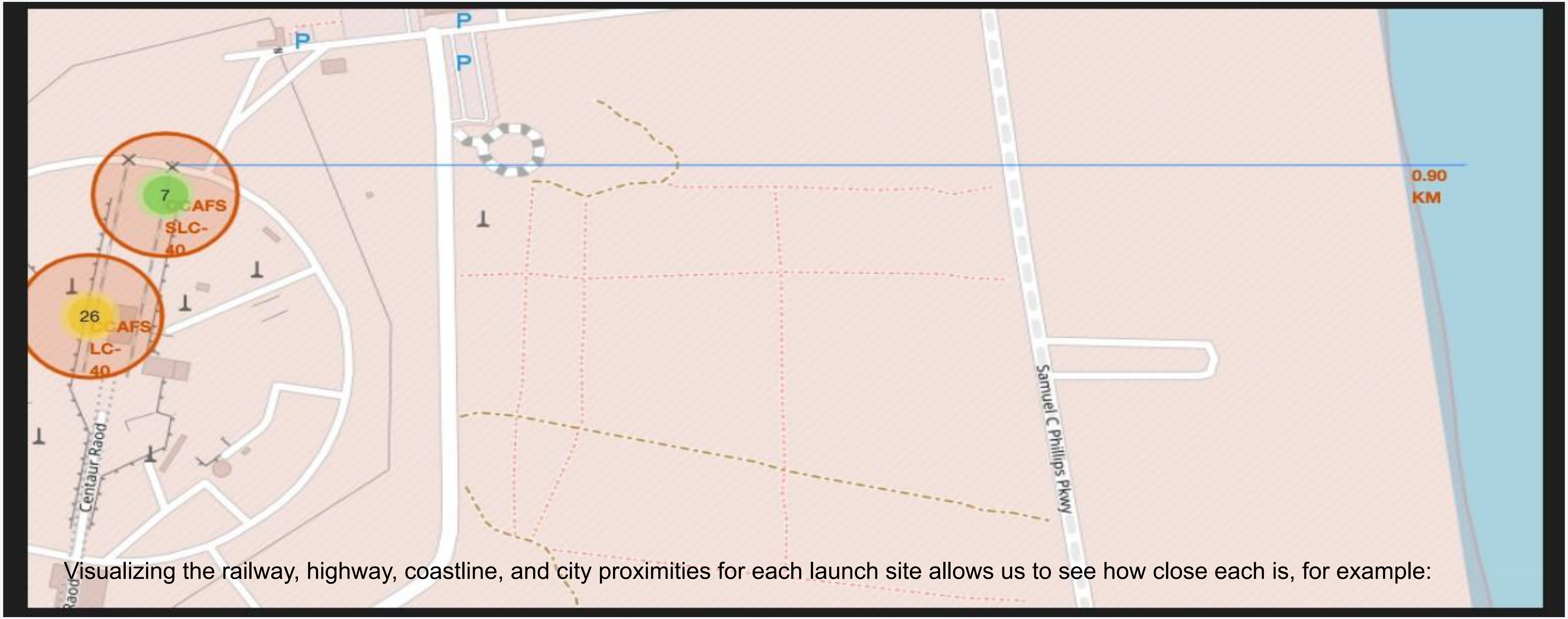


Launch Site Location Analysis (Contd)

From the color-labeled markers in marker clusters, we can easily identify which launch sites have relatively high success rates.



Launch Site Location analysis(Contd)

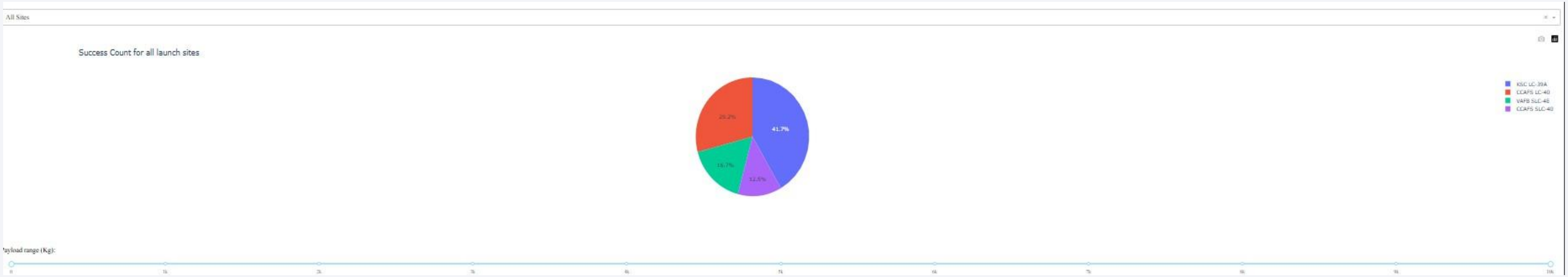


Visualizing the railway, highway, coastline, and city proximities for each launch site allows us to see how close each is, for example:

Section 4

Build a Dashboard with Plotly Dash

Dashboard



- Pie chart showing booster landing success rate
- Drop down menu to choose between all sites and individual launch sites
- Color coded by launch site
  - KSC LC-39A
  - CCAFS LC-40
  - VAFB SLC-4E
  - CCAFS SLC-40

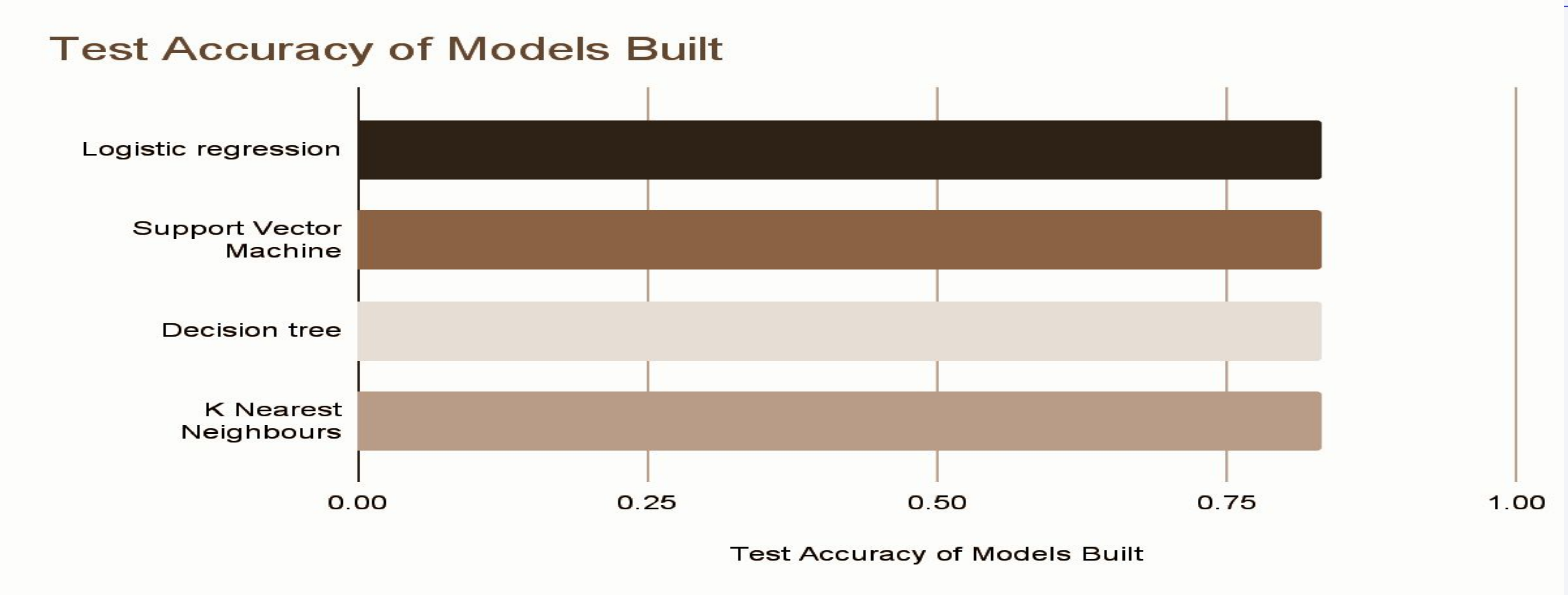
Dashboard(cONTD)

- Range slider for limiting payload amount
- Scatter chart showing payload mass vs. landing outcome
- Color coded by booster version
  - v1.0
  - v1.1
  - FT
  - B4
  - B5

Section 5

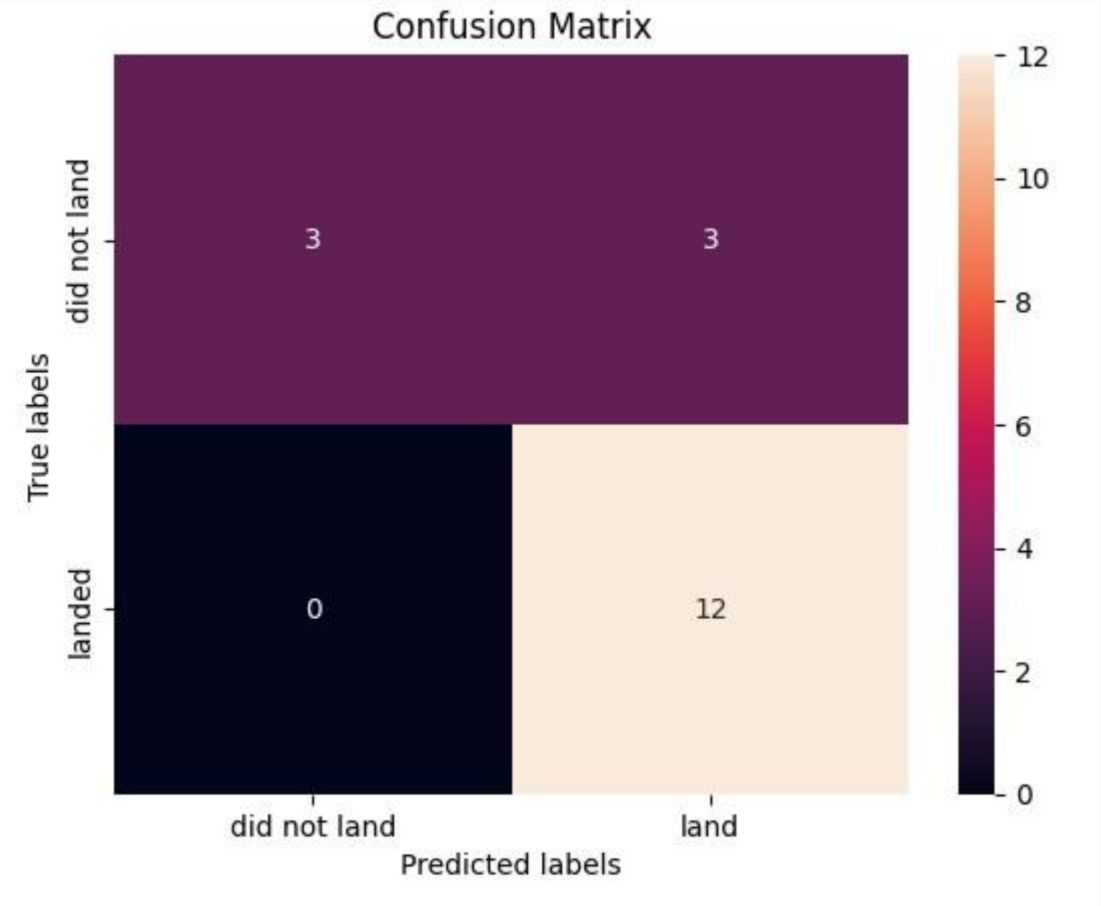
Predictive Analysis (Classification)

Classification Accuracy



Each of the four models built came back with the same accuracy score, 83.33%

Confusion Matrix



- The confusion matrix is same for all four models
- Models predicted slightly false values for test data and training data

Conclusions

- Using the models from this report SpaceY can predict the land of 1st booster with an accuracy of 83.331%
- Less cost price by predicting the failure of 1st stage boosters will enable SpaceY to bid more cost wise against SpaceX

Appendix

Reference Links :

- https://github.com/umamaheswaranv/datascience_capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
- https://github.com/umamaheswaranv/datascience_capstone/blob/main/jupyter-labs-webscraping.ipynb
- https://github.com/umamaheswaranv/datascience_capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
- https://github.com/umamaheswaranv/datascience_capstone/blob/main/edadataviz.ipynb
- https://github.com/umamaheswaranv/datascience_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- https://github.com/umamaheswaranv/datascience_capstone/blob/main/lab_jupyter_launch_site_location.ipynb
- https://github.com/umamaheswaranv/datascience_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb
- https://github.com/umamaheswaranv/datascience_capstone/blob/main/spacex_dash_app%20(1).py

Thank you!