IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jerry Tam
23-03-2024

Outline

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

Executive Summary

Summary of methodologies

- First data were collected. Then data wrangling was performed.
- Exploratory data analysis (EDA) was performed using visualization and SQL. Interactive visual analytics was performed using Folium and Plotly Dash.
- Finally, predictive analysis was performed using classification models.

Summary of results

- Decision Tree is the best model to perform landing prediction with following parameters:
- Booster Version, Payload Mass, Orbit, Launch Site, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

Introduction

- SpaceX 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 SpaceX 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 SpaceX for a rocket launch.
- In this research, we will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Data were collected by
  - SpaceX API
  - Web Scraping Records from Wikipedia

Sending request to SpaceX API

Request to the SpaceX API → Clean the requested data

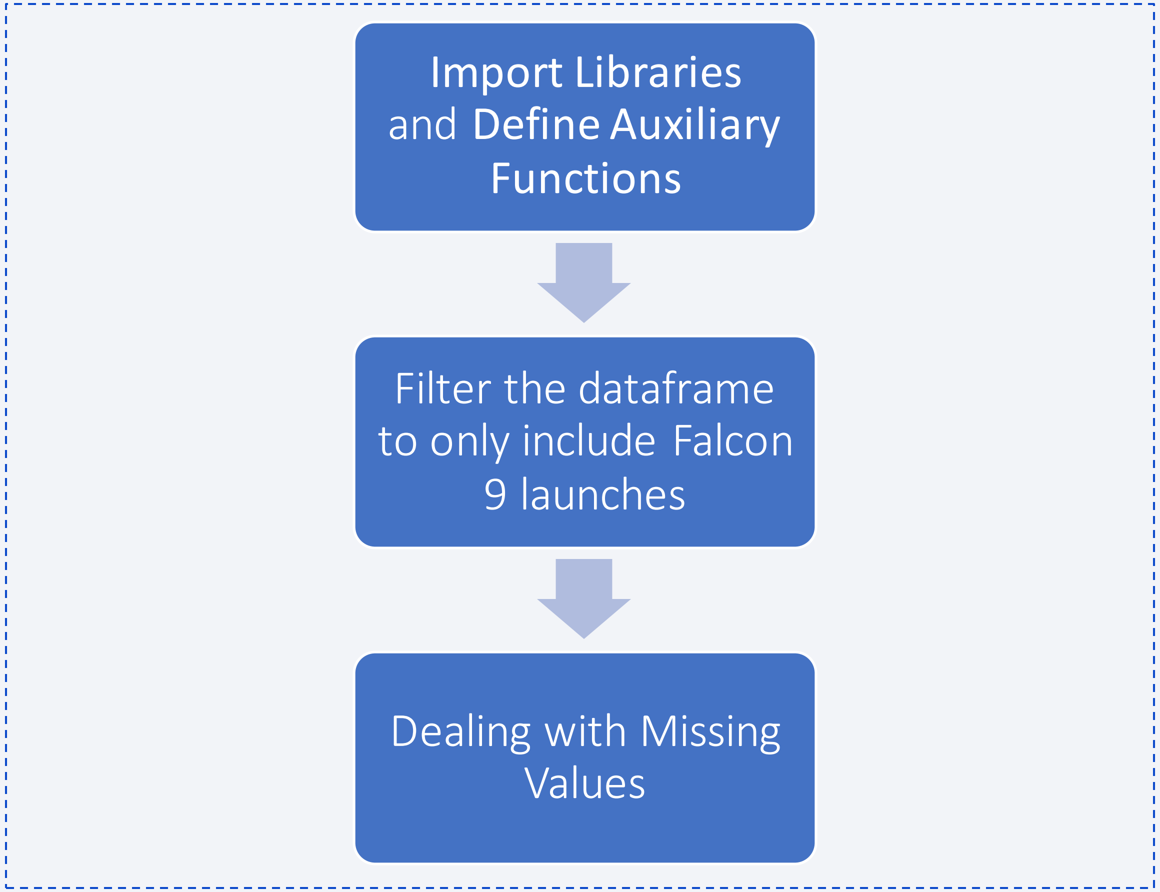
Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia

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

Data Collection – SpaceX API

- A get request was made to the SpaceX API. Some basic data wrangling and formating were done.
  - Request to the SpaceX API
  - Clean the requested data
- Completed SpaceX API calls notebook:
  - https://github.com/zeon2103133/applied_data_science_capstone_project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

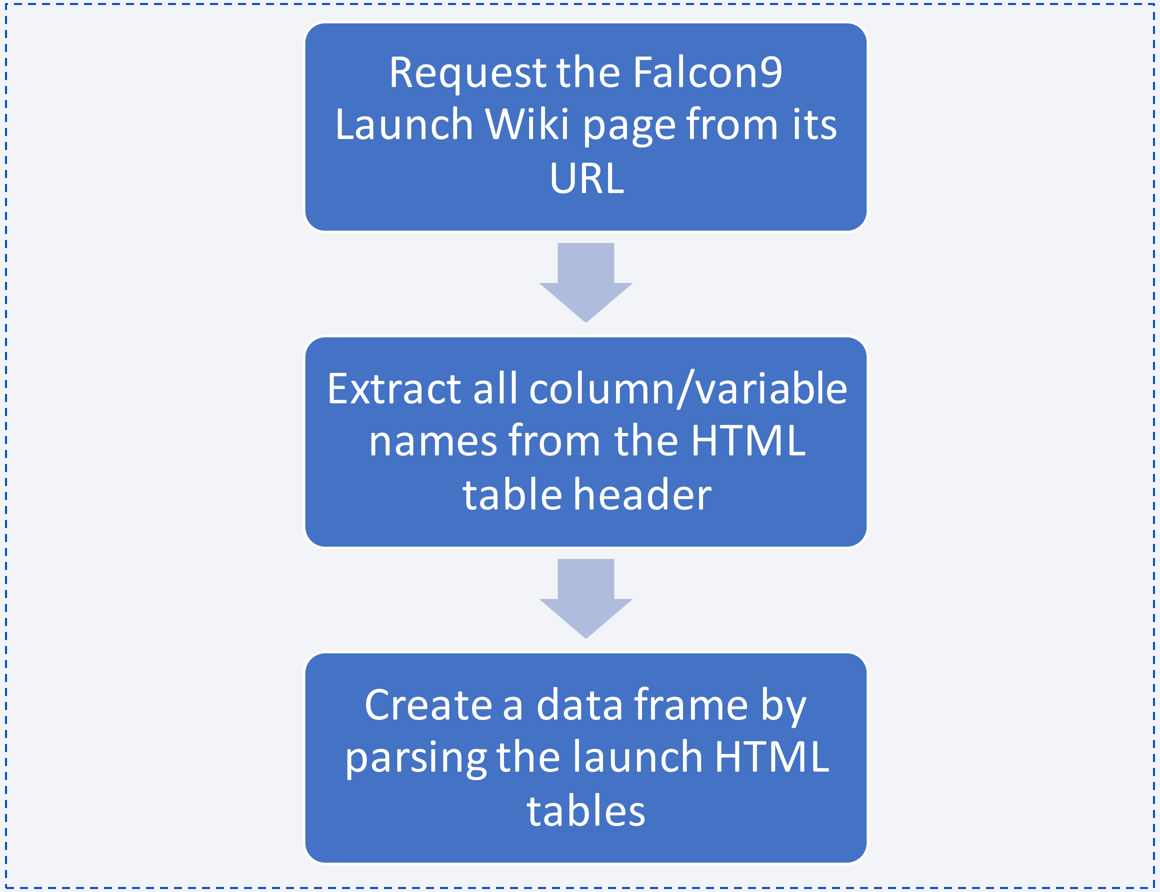
Process



Data Collection - Scraping

- Web scrapping Falcon 9 launch records was done with BeautifulSoup
    - Extract a Falcon 9 launch records HTML table from Wikipedia
    - Parse the table and convert it into a Pandas data frame
- Completed web scraping notebook:
    - https://github.com/zeon2103133/applied_data_science_capstone_project/blob/main/jupyter-labs-webscraping.ipynb

Process



Data Wrangling

- Exploratory Data Analysis (EDA) was performed and Training Labels were determined
  - Exploratory Data Analysis
  - Determine Training Labels
- You need to present your data wrangling process using key phrases and flowcharts
- Completed data wrangling related notebooks:
  - https://github.com/zeon2103133/applied_data_science_capstone_project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Process

Calculate the number of launches on each site

↓

Calculate the number and occurrence of each orbit

↓

Calculate the number and occurence of mission outcome of the orbits

EDA with Data Visualization

- To visualize the relationship between parameters, following charts were plotted
  - Scatter Plot of Flight Number vs Launch Site
  - Scatter Plot of Payload vs Launch Site
  - Bar Chart of success rates of each orbit type
  - Scatter Plot of Flight Number vs Orbit type
  - Line Chart of launch success rate vs year (launch success yearly trend)
- Completed EDA with data visualization notebook:
  - https://github.com/zeon2103133/applied_data_science_capstone_project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Following SQL queries were performed.
  - select distinct Launch_Site from SPACEXTABLE
  - select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
  - select Customer, sum(PAYLOAD_MASS__KG_) from SPACEXTABLE group by Customer having Customer = 'NASA (CRS)'
  - select Booster_Version, avg(PAYLOAD_MASS__KG_) from SPACEXTABLE group by Booster_Version having Booster_Version = 'F9 v1.1'
  - select min(Date) from (select * from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)')
  - select distinct Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
  - select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome

EDA with SQL

- Following SQL queries were performed. (cont')
  - select distinct booster_version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
  - select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, launch_site from SPACEXTABLE where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
  - select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE group by Landing_Outcome order by count(Landing_Outcome) desc
- Completed EDA with SQL notebook:
  - https://github.com/zeon2103133/applied_data_science_capstone_project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Objects added on the map	Reason of adding
Circles & Markers	To mark all launch sites on a map
Marker Clusters	To mark the success/failed launches for each site on the map
PolyLine	To calculate the distances between a launch site to its proximities

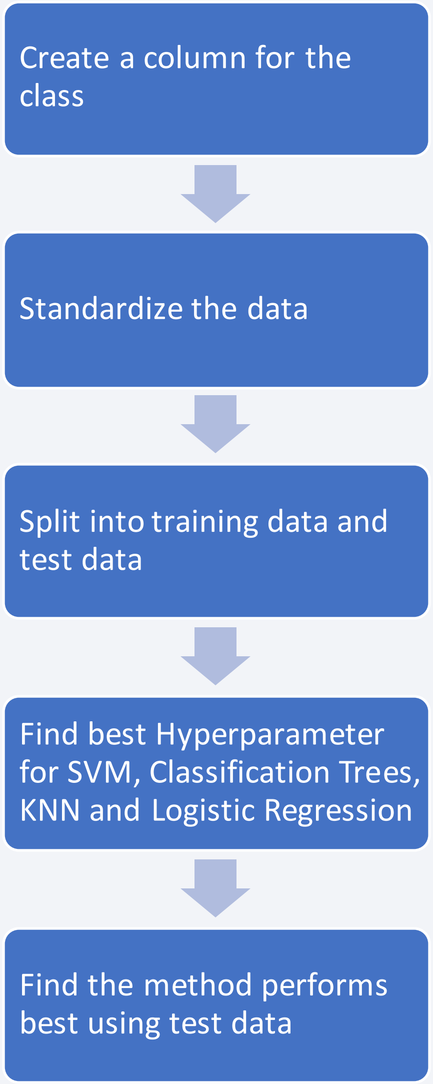
- Completed interactive map with Folium map
  - https://github.com/zeon2103133/applied_data_science_capstone_project/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Dropdown list and a range slider to interact with a pie chart and a scatter point chart, in order to perform interactive visual analytics on SpaceX launch data in real-time.
  - obtain some insights to answer the following five questions:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
- Completed Plotly Dash lab:
  - https://github.com/zeon2103133/applied_data_science_capstone_project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Best performing classification model to predict Falcon 9 First Stage Landing Result was found by following steps.
  - Exploratory Data Analysis was performed and Training Labels were determined.
  - Then, by train data, I found the best hyperparameters for SVM, Classification Trees, k Nearest Neighbors and Logistic Regression.
  - Lastly, test data were used to find the method performs best, by comparing the accuracy among models
- Completed predictive analysis lab:
  - https://github.com/zeon2103133/applied_data_science_capstone_project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



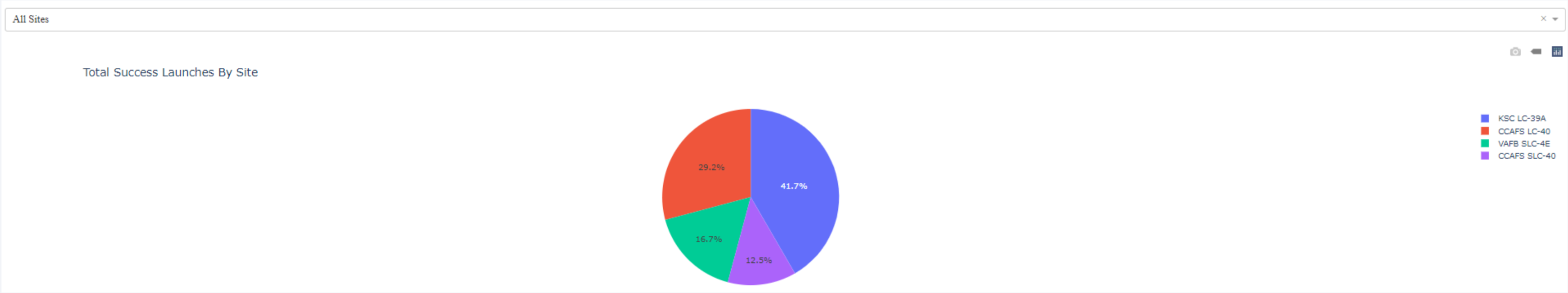
Results

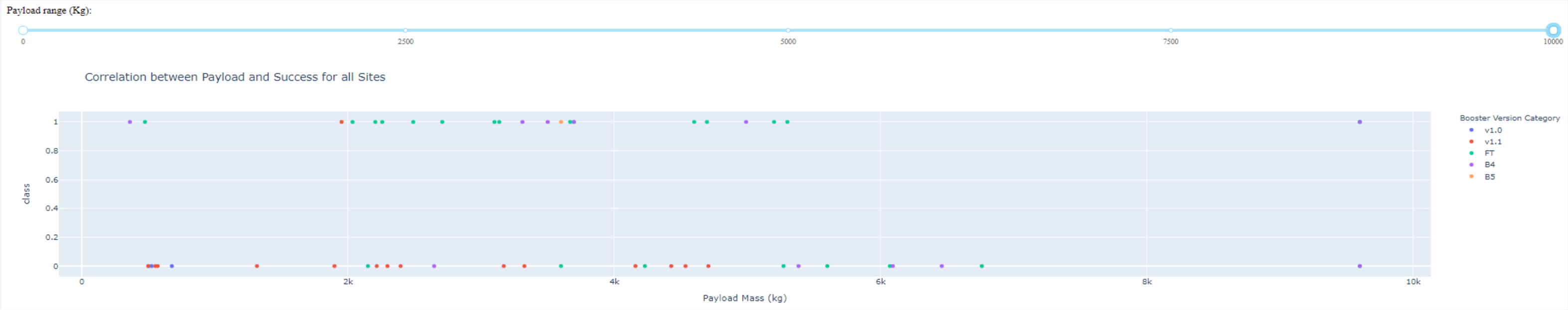
<u>Exploratory data analysis results</u>

- Success rate of landing increase against the number of flights.
- The VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- Orbits ES-L1, GEO, HEO and SSO have the highest success rate (100%).
- The LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
- The success rate since 2013 kept increasing till 2020

Results

Interactive analytics demo





Predictive analysis results

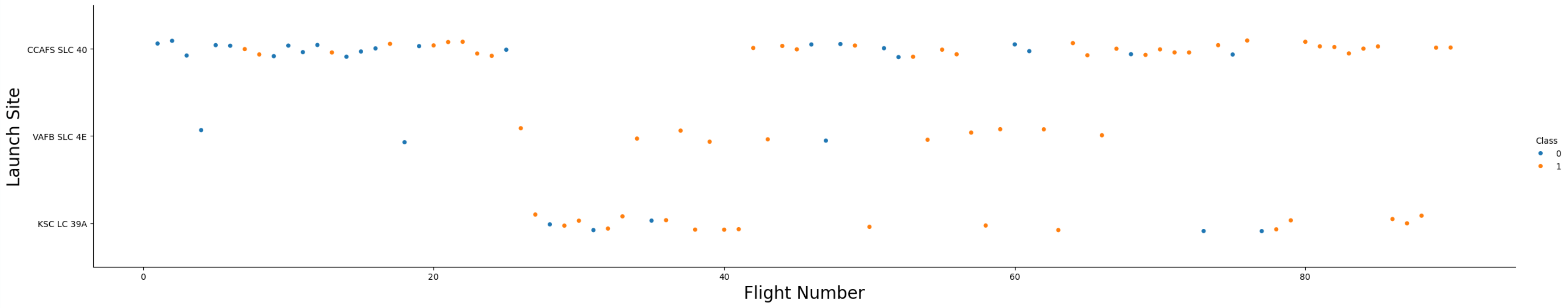
Classification Trees performed the best since it attained the highest accuracy score.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

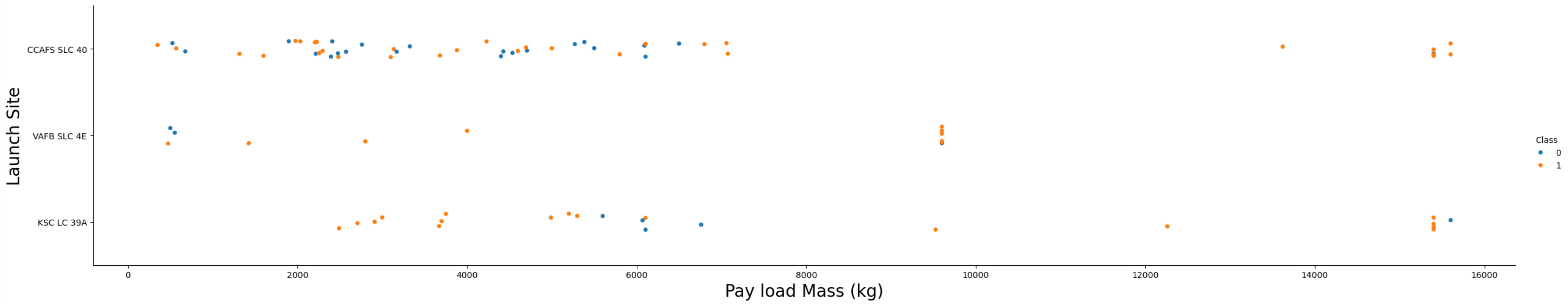
Scatter plot of Flight Number vs. Launch Site



In early time, launches mainly took place in site "CCAFS LC-40" and landing success rate is above 50%. After above 25th flights, launches took place more in KSC LC-39A and VAFB SLC 4E as well and success rate increased.

Payload vs. Launch Site

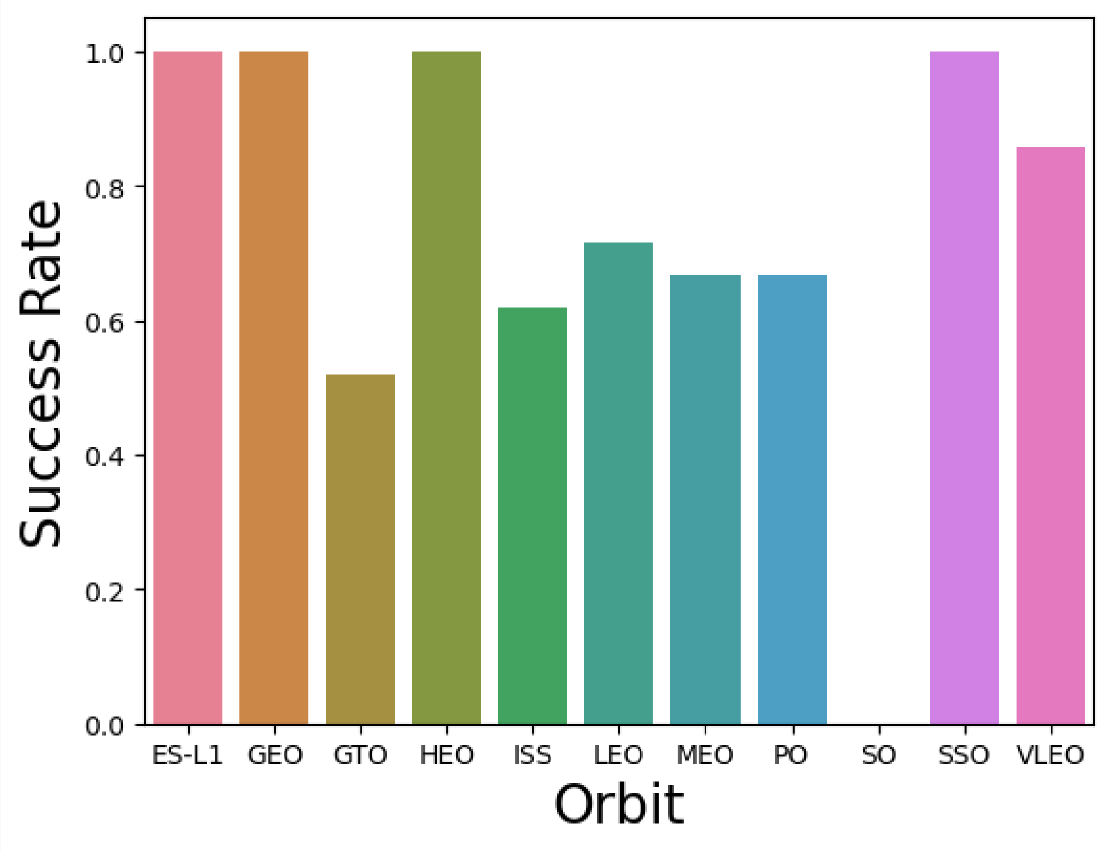
Scatter plot of Payload vs. Launch Site



For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

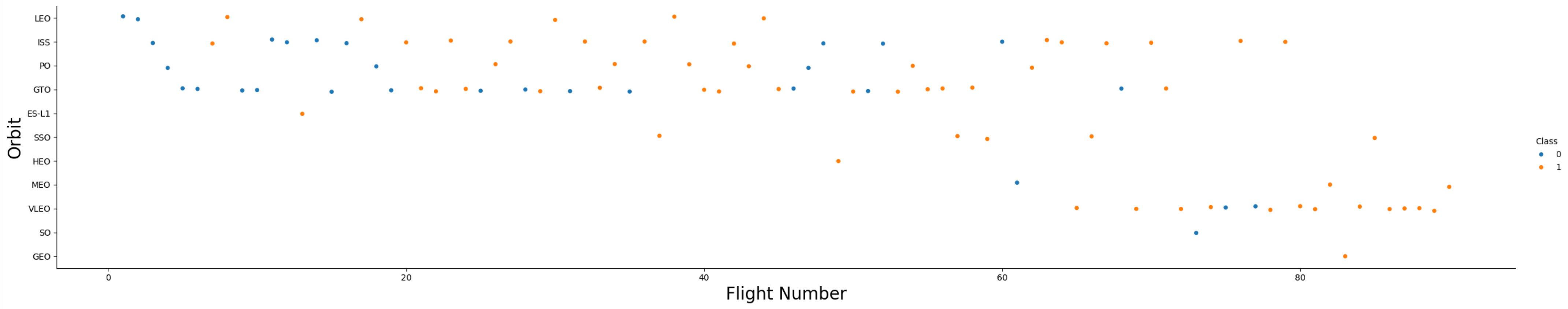
Bar chart for the success rate of each orbit type



Orbits ES-L1, GEO, HEO and SSO have the highest success rate (100%).

Flight Number vs. Orbit Type

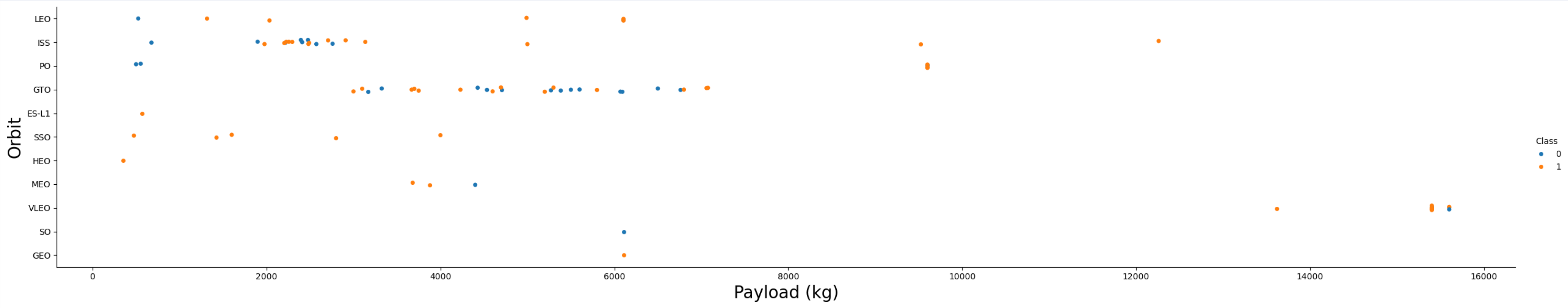
Scatter point of Flight number vs. Orbit type



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

Payload vs. Orbit Type

Scatter point of Flight number vs. Orbit type

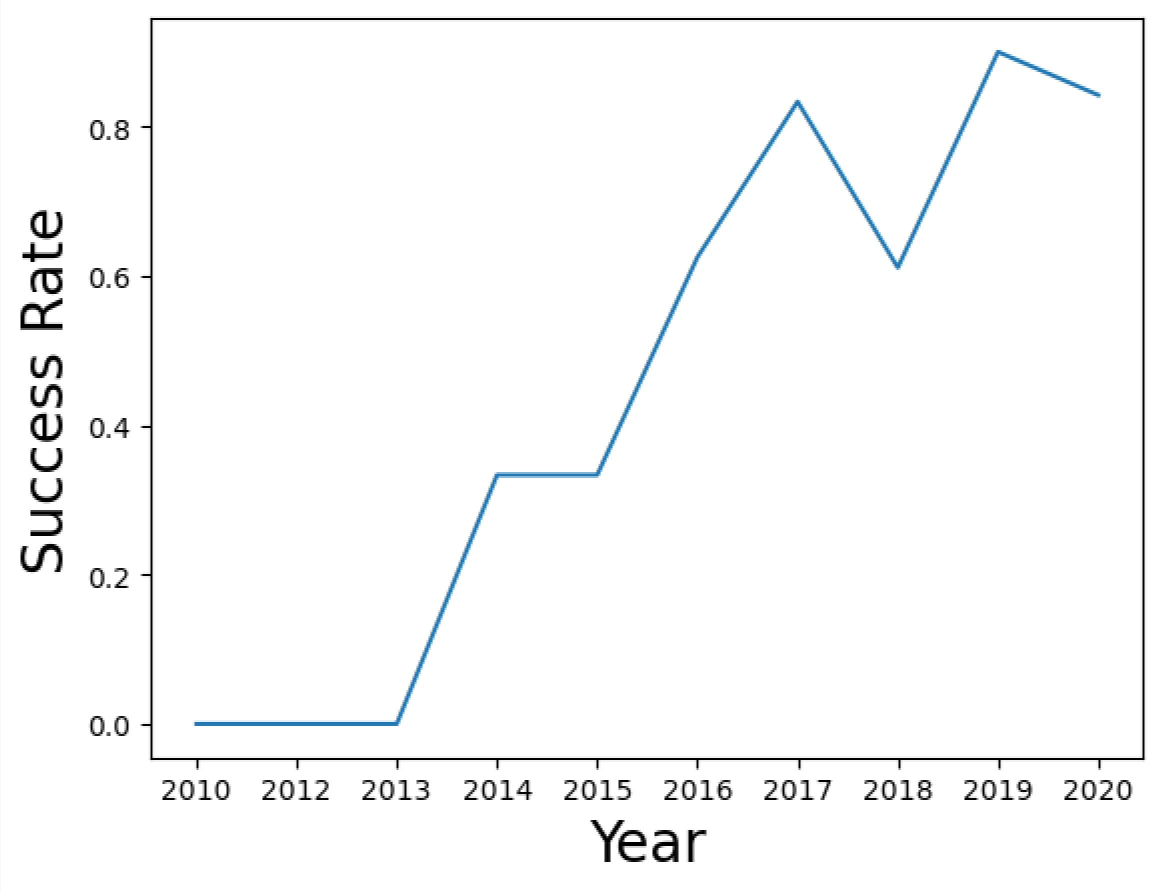


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there.

Launch Success Yearly Trend

Line chart of yearly average success rate



The success 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

Following sql magic command was called to extract the desired data:

'%sql select distinct Launch_Site from SPACEXTABLE'

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

Following sql magic command was called to extract the desired data:

'%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5'

Total Payload Mass

Customer	sum(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

Following sql magic command was called to extract the desired data:

' %sql select Customer, sum(PAYLOAD_MASS__KG_) from SPACEXTABLE group by Customer having Customer = "NASA (CRS)" '

Average Payload Mass by F9 v1.1

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

Following sql magic command was called to extract the desired data:

' %sql select Booster_Version, avg(PAYLOAD_MASS__KG_) from SPACEXTABLE group by Booster_Version having Booster_Version = "F9 v1.1" '

First Successful Ground Landing Date

min(Date)
2015-12-22

Following sql magic command was called to extract the desired data:

' %sql select min(Date) from (select * from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)') '

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Following sql magic command was called to extract the desired data:

' %%sql select distinct Booster_Version from SPACEXTABLE

where Landing_Outcome = 'Success (drone ship)' and

PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000 '

Total Number of Successful and Failure Mission Outcomes

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

Following sql magic command was called to extract the desired data:

' %sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome '

Boosters Carried Maximum Payload

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

Following sql magic command was called to extract the desired data:

' %%sql select distinct booster_version from SPACEXTABLE where PAYLOAD_MASS__KG_ =

(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE) '

2015 Launch Records

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Following sql magic command was called to extract the desired data:

' %%sql select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, launch_site from SPACEXTABLE where Landing_Outcome = "Failure (drone ship)" and substr(Date,0,5)="2015" '

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

Landing_Outcome	count(Landing_Outcome)
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

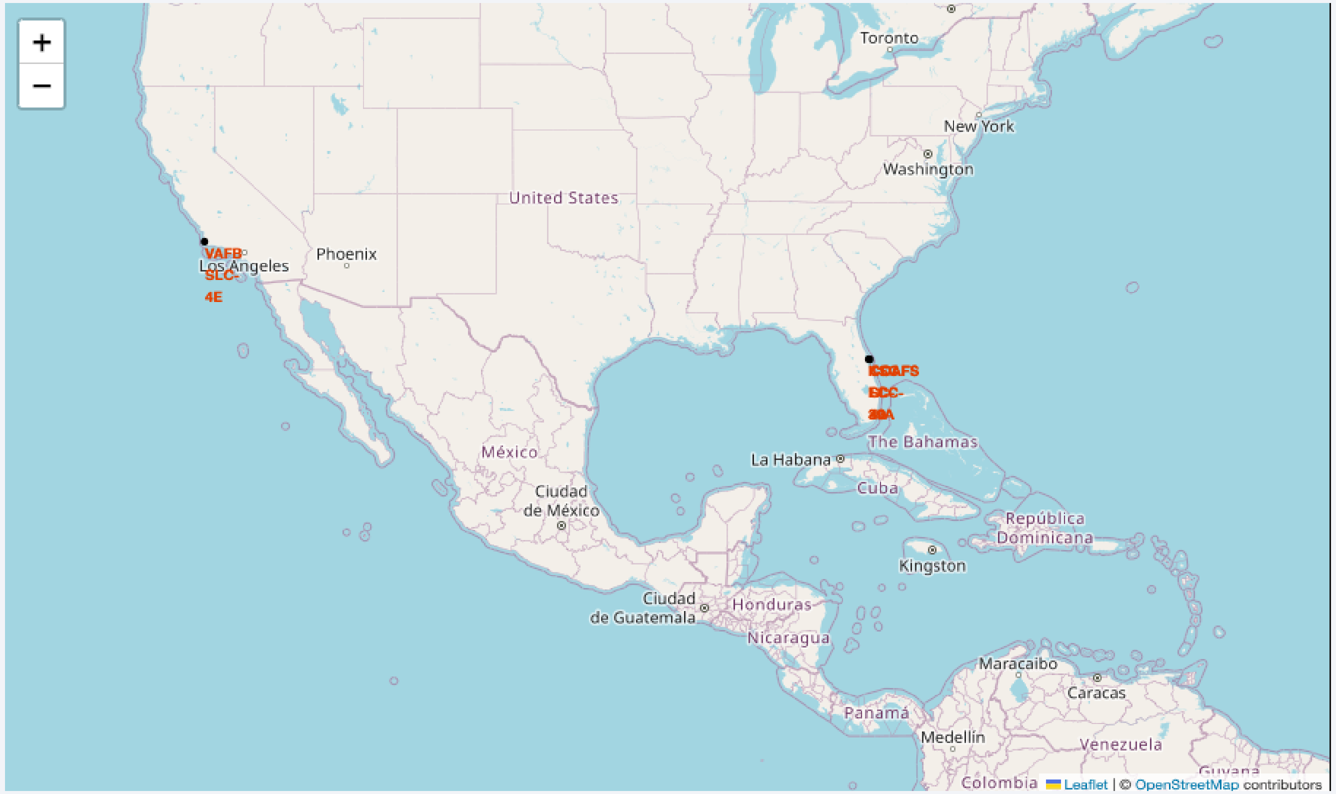
Following sql magic command was called to extract the desired data:

' %%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE group by Landing_Outcome order by count(Landing_Outcome) desc '

Section 3

Launch Sites Proximities Analysis

Locations of all launch sites

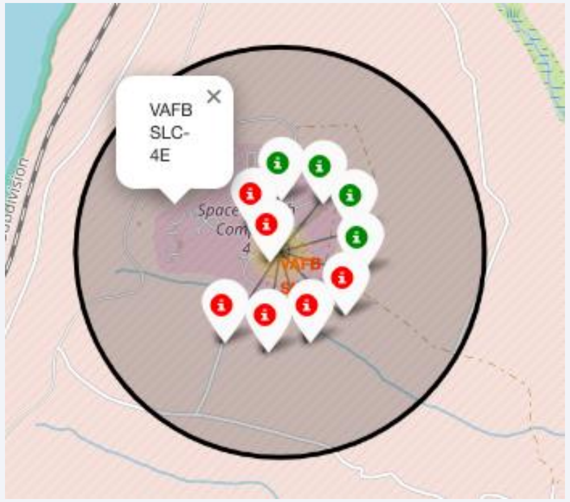


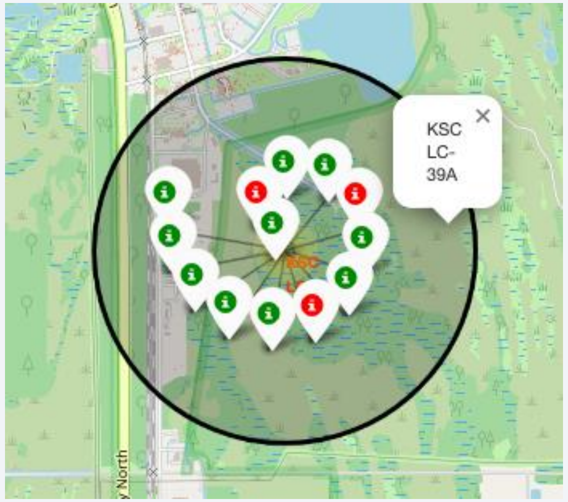
Findings:

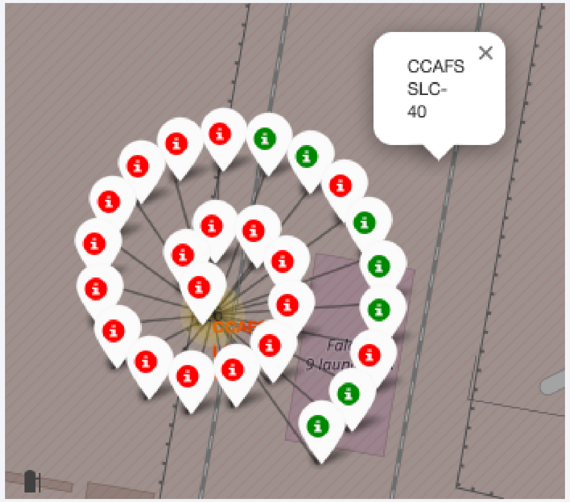
•All launch sites are in proximity to the Equator line.

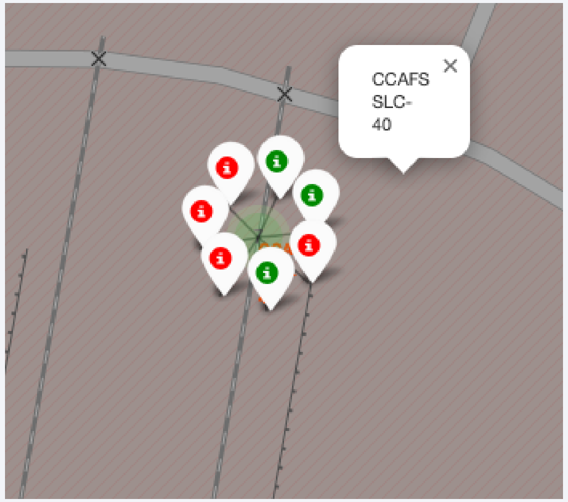
•All launch sites are in very close proximity to the coast.

Marks of Success/Failed Launches for Each Site





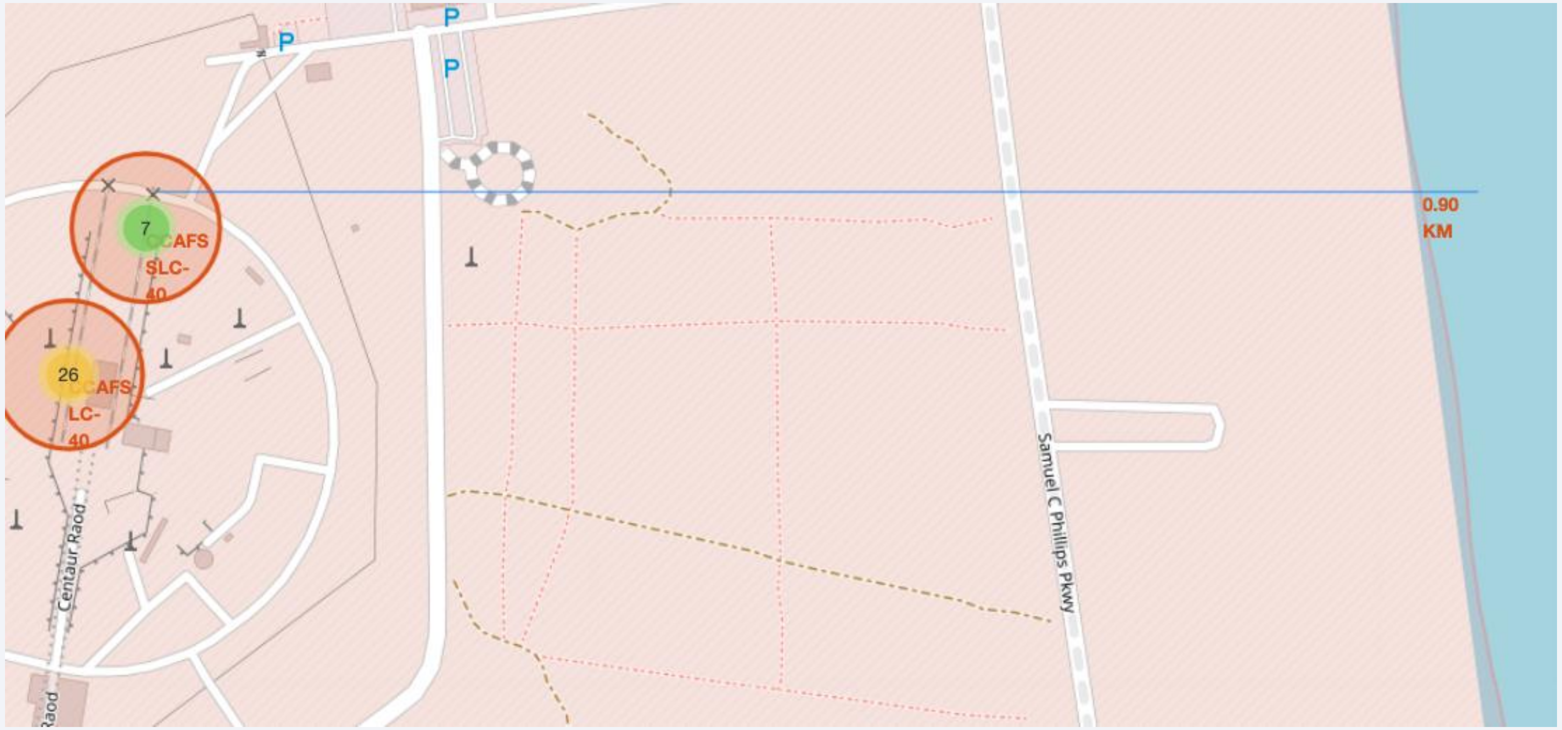




Findings:

•Site "KSC LC-39A" has highest success launch rate.

Distances between a launch site to its proximities



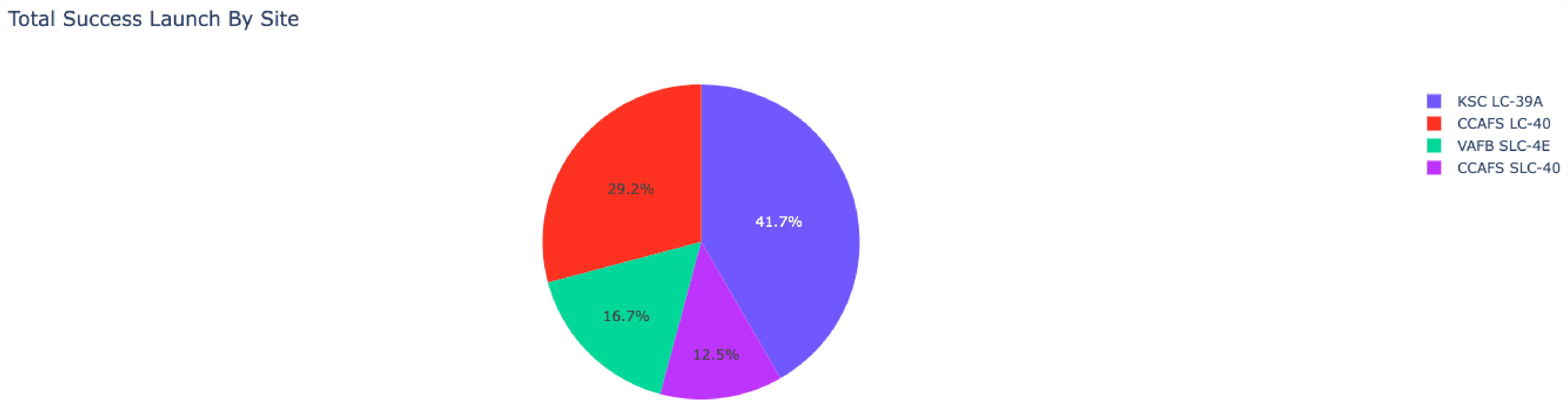
Findings:

- Site "CCAFS LC-40" is close to the coastline.

Section 4

Build a Dashboard with Plotly Dash

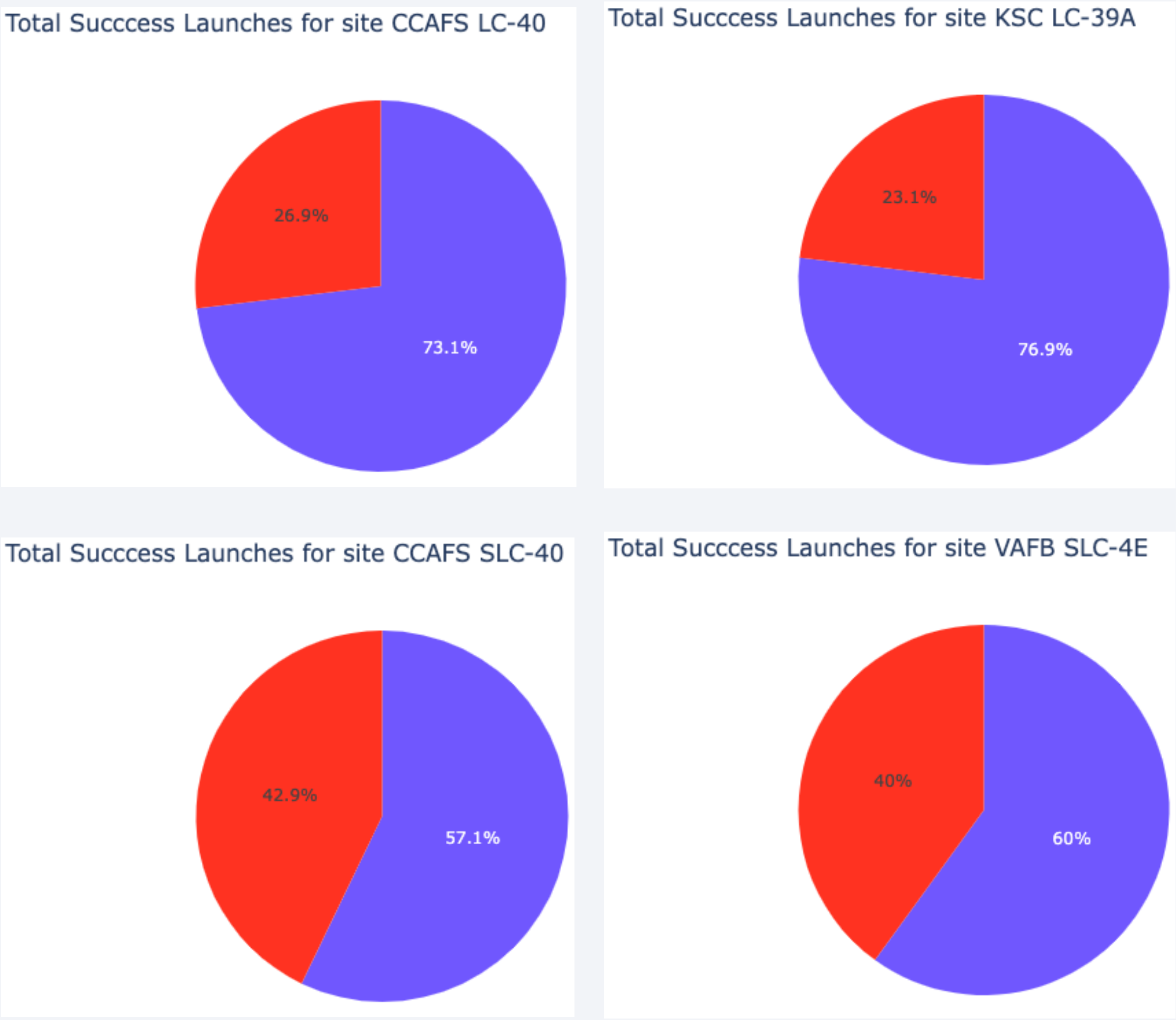
Numbers of Successful Launches of Each Site

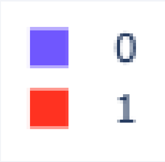


Findings:

- Site "KSC LC-39A" has the largest successful launches

Success Ratio of Lauches per Site

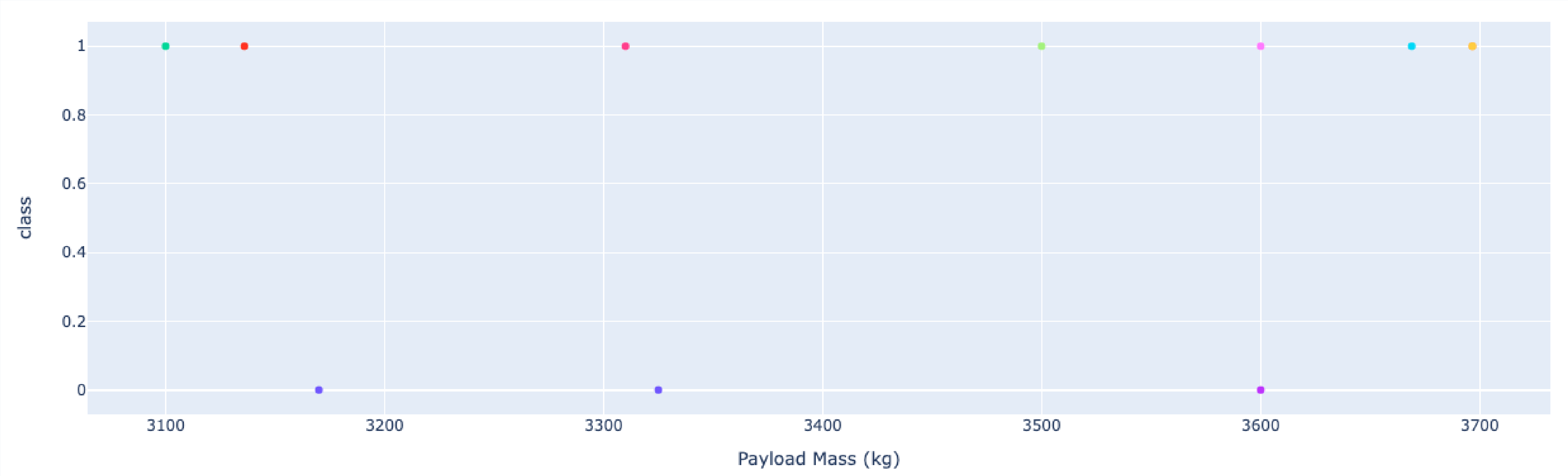




Findings:

- Site "CCAFS SLC-40" has the highest launch success ratio

Scatter Plot of Payload vs. Launch



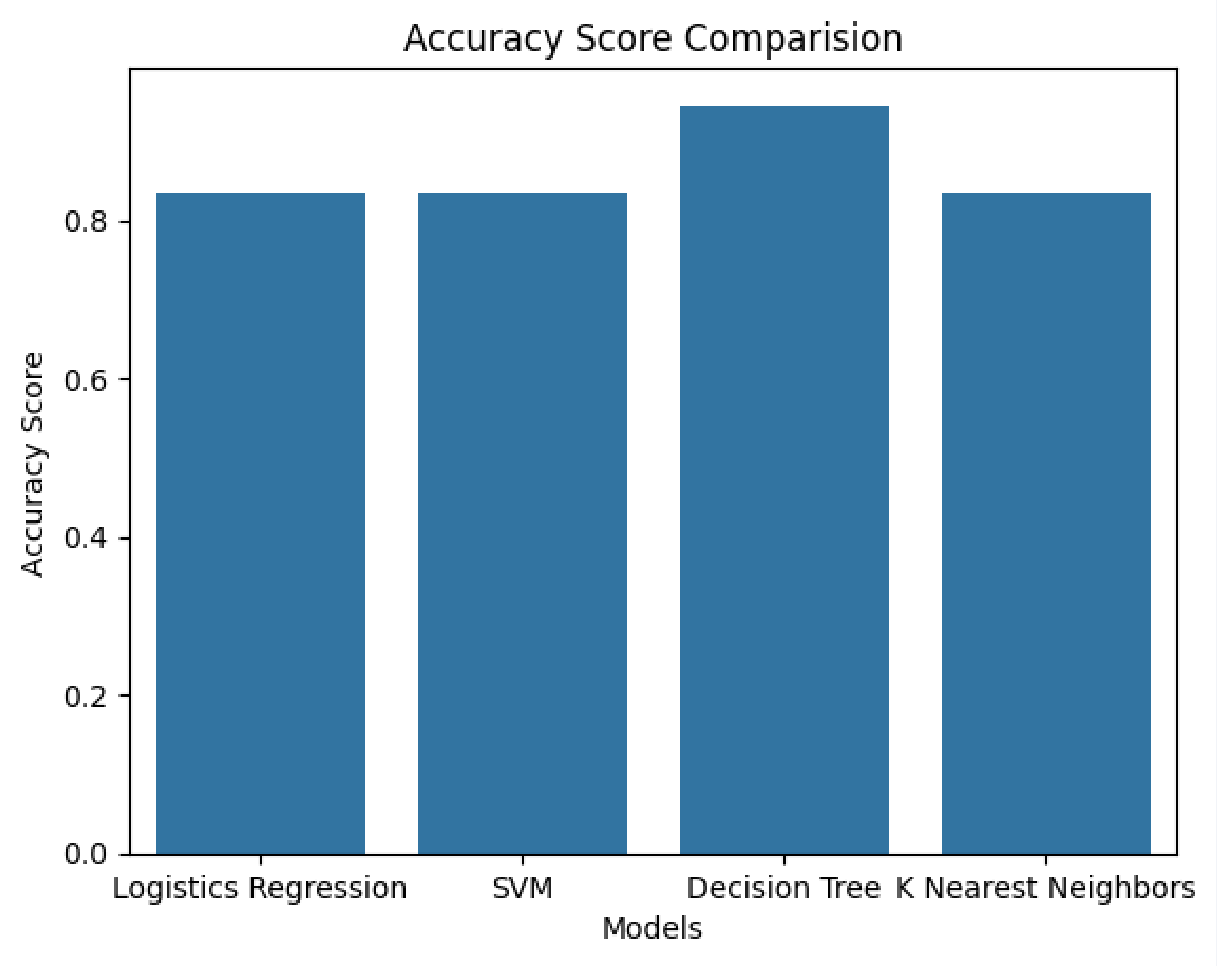
Findings:

- Payload range 3,000 ~ 4,000 kg has the largest success rate.

Section 5

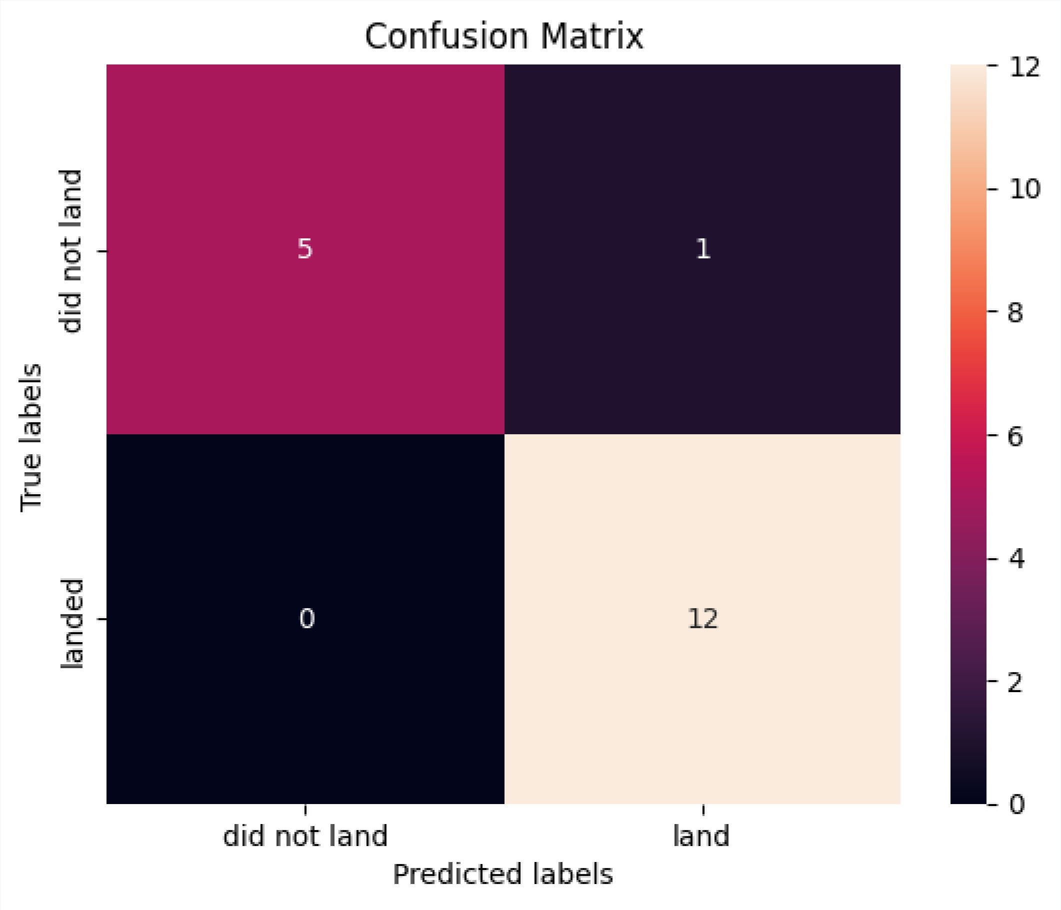
Predictive Analysis (Classification)

Classification Accuracy



Decision Tree Classification performed the best since it attained the highest accuracy score.

Confusion Matrix



Confusion matrix of the best performing model, the Decision Tree, shows largest number of true negative (5), while the other 3 models have only (3).

Conclusions

- The success rate since 2013 kept increasing till 2020.
- In early time, launches mainly took place in site "CCAFS LC-40" and landing success rate is above 50%. After above 25th flights, launches took place more in KSC LC-39A and VAFB SLC 4E as well and success rate increased.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- Site "KSC LC-39A" has highest success launch rate and the largest successful launches as well.
- Orbits ES-L1, GEO, HEO and SSO have the highest success launch rate (100%).
- In the LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there.
- All launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast.
- Decision Tree Classification performed the best on predicting landing result.

Appendix

- GitHub Project Repository:
  - https://github.com/zeon2103133/applied_data_science_capstone_project/tree/main

Thank you!