IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Author : Chimbi Hove

Date: 05 July 2024

Contact info: xyz@gmail.com : (900) 556 4888

Outline

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

Executive Summary

- Summary of methodologies
- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Predictions
- Summary of all results
- EDA results
- Interactive analytics screenshots
- Predictive Analytics result

Introduction

- Project background and context

- The presentation is to show the journey taken in exploring the launches done by SpaceX at different locations and their successes and failures. The data was obtained from a Wikipedia article :

- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- This capstone project's aim is to create a presentation based on the outcomes of all tasks performed throughout the labs. Your presentation will develop into a story of all your data science journey in this project, and it should be compelling and easy to understand.

- Problems we want to solve

- This presentation explores the successes and failures of SpaceX since its initial launch in 2010 all through 2020.

- Presentation also determines factors of successful rocket launches and landings

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: Data was collected using APIs and Web Scraping from Wikipedia
- Performed data wrangling using: one-hot encoding to categorical features
- Performed exploratory data analysis (EDA): using visualization and SQL
- Performed interactive visual analytics using: Folium and Plotly Dash
- Perform predictive analysis using: classification models in Machine Learning which were used to predict accuracy
- Models include: Decision Tree, SVM, Linear Regression, K-Nearest Neighbor

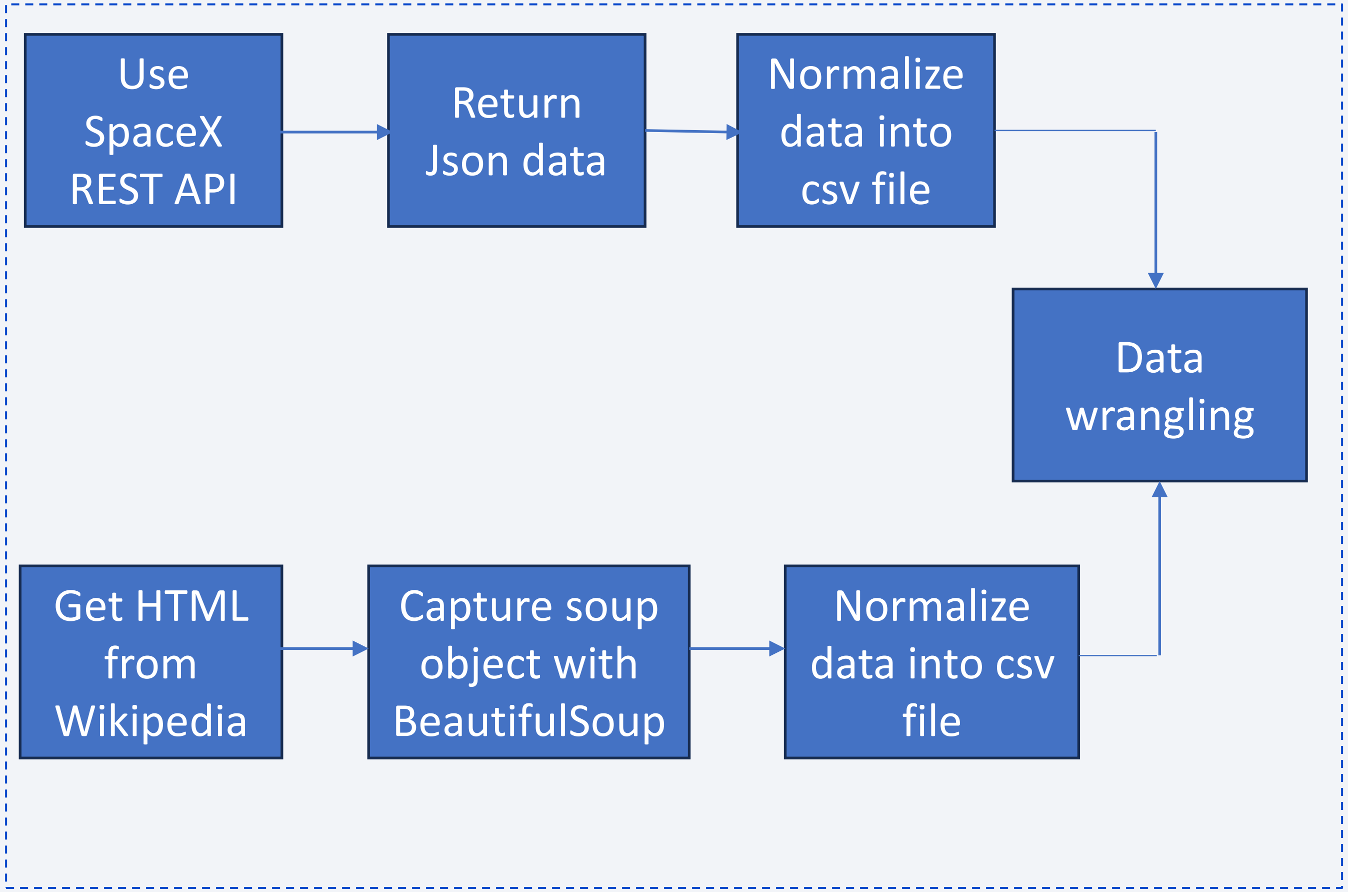
Data Collection

- Describe how data sets were collected.
- Data collection was done using get request to the SpaceX API.
- Decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
- Data was cleaned and checked for missing values and filled in missing values with the mean/avg.
- Then data was web scraping from article with BeautifulSoup.
- Launch records were parsed from html tables and converted to pandas dataframes for analysis.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

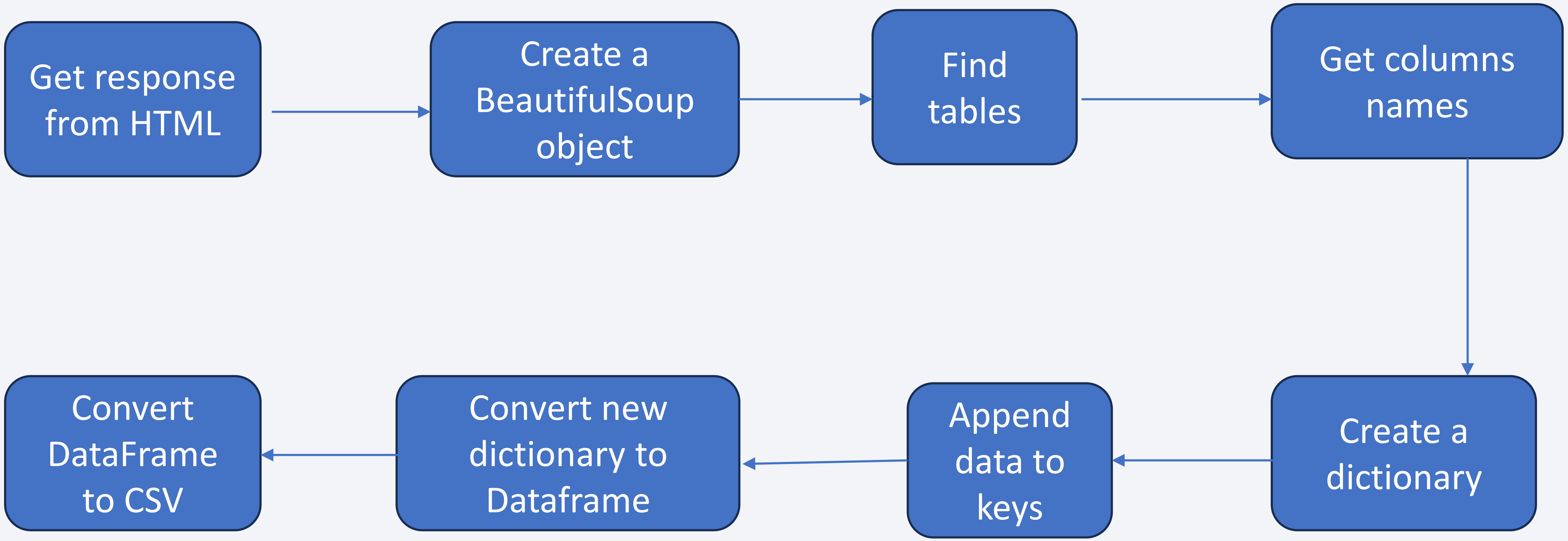
- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

- GitHub URL of the completed SpaceX API calls notebook https://github.com/chimbih/Applied-Data-Science-Capstone-/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



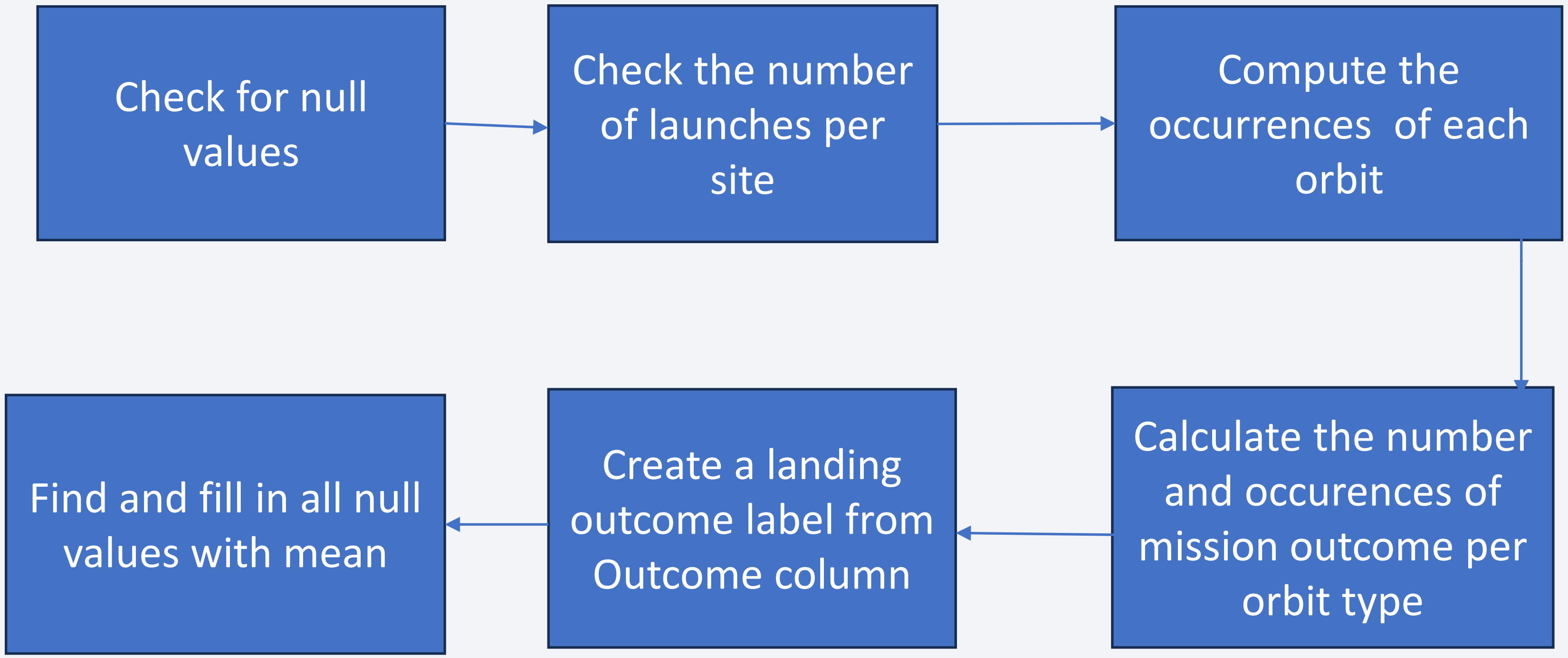
Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts



- https://github.com/chimbih/Web-scraping/blob/main/jupyter-labs-webscraping.ipynb

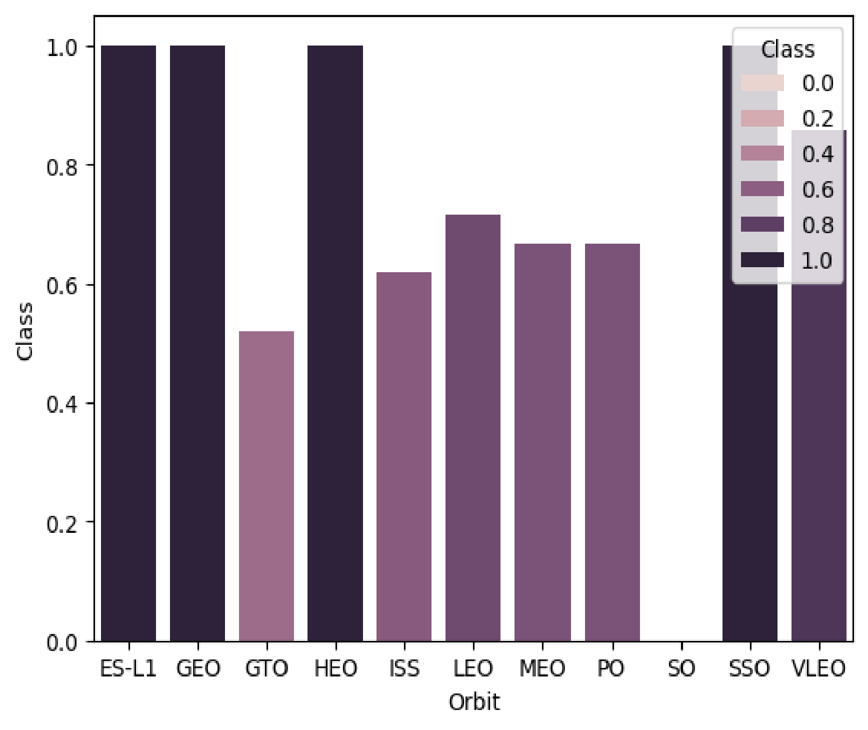
Data Wrangling



- https://github.com/chimbih/Data-wrangling

EDA with Data Visualization

- Data was visualized to link relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.



- https://github.com/chimbih/Dataviz/blob/main/edadataviz%20(1).ipynb

EDA with SQL

- **Display the names of the unique launch sites in the space mission**
- **Display 5 records where launch sites begin with the string 'CCA’**
- **Display the total payload mass carried by boosters launched by NASA (CRS)**
- **Display average payload mass carried by booster version F9 v1.1**
- **List the date when the first succesful landing outcome in ground pad was acheived.**
- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- **List the total number of successful and failure mission outcomes**
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- **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.**
- **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.**

- https://github.com/chimbih/Exploratory-Data-Analysis/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

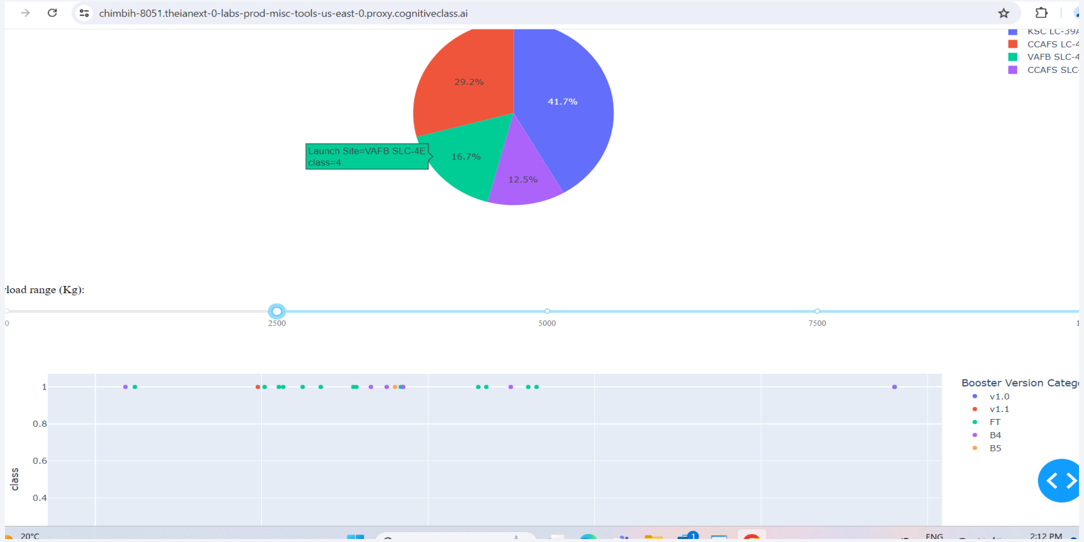
Build an Interactive Map with Folium

- Map objects such as markers, circles, lines were created and added to a folium map
- Markers were added to help identify why launch sites are located where they are for safety reasons, no debris can fall on city residents; if launches fail the debris can fall into the ocean.
- Highways and railway lines are very close to launchsites to help with transportation of equipment and crew for the launches.

- https://github.com/chimbih/Launch-site-analysis-Folium-/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions you have added to a dashboard



- Pie charts of the successes and failures of each launch site as well as
- Explain why you added those plots and interactions
- https://github.com/chimbih/SpaceX-Dashboard/blob/main/spacex_dash_app.py

Calculate the accuracy on the test data using the method score.

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- Load the data
- Create a Numpy array from the column Class in data , by applying to_numpy(), then assigning it to a variable Y
- Standardize the data in X then reassign ot the variable X using transform
- Used the function train_test_split to split the data (X&Y) into training and testing data
- Create a logistic regression object then create a GridSearchCV object, and fit the object to find best parameters
- Output the GridSearh object for logistic regression
- Calculatedt the accuracy on the test data using the score() and built a confusion matrix
- Created an SVM object, then created GridSearch() to calculate accuracy for SVM and confusion matrix.
- Repeated step above to check accuracy and confusion matrices for Decision Tree and KNN

- https://github.com/chimbih/SpaceX-Machine-Learning/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory & Predictive data analysis results
- SVM, KNN, LR models have accuracy in the range of 0.83 to 0.85, while Decision Tree has 0.90 accuracy.
- Low weight payloads have more success than the heavier counterparts
- KSC LC 39A has the most successful launches compared to all other sites
- ES-L1, GEO, HEO & SSO have a high successes

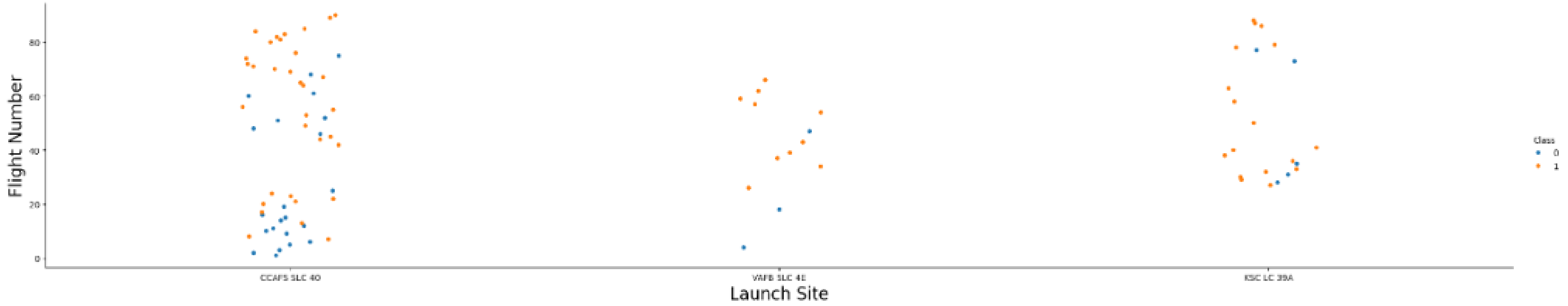
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

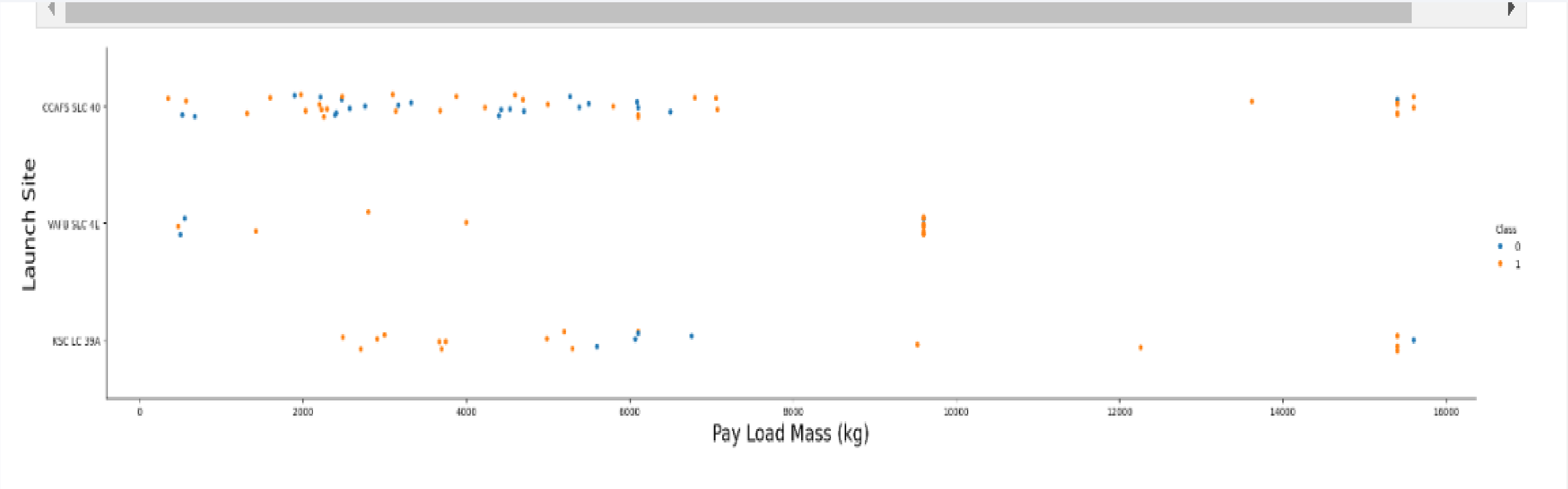
```
In [5]:
### TASK 1: Visualize the relationship between Flight Number and Launch Site
sns.catplot(y="FlightNumber", x="LaunchSite", hue="Class", data=df, aspect = 5)
plt.xlabel("Launch Site",fontsize=20)
plt.ylabel("Flight Number",fontsize=20)
plt.show()
```



CCAFS SLC -40 has the most flights while VAFB SCL-4E has the least amount of flights.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

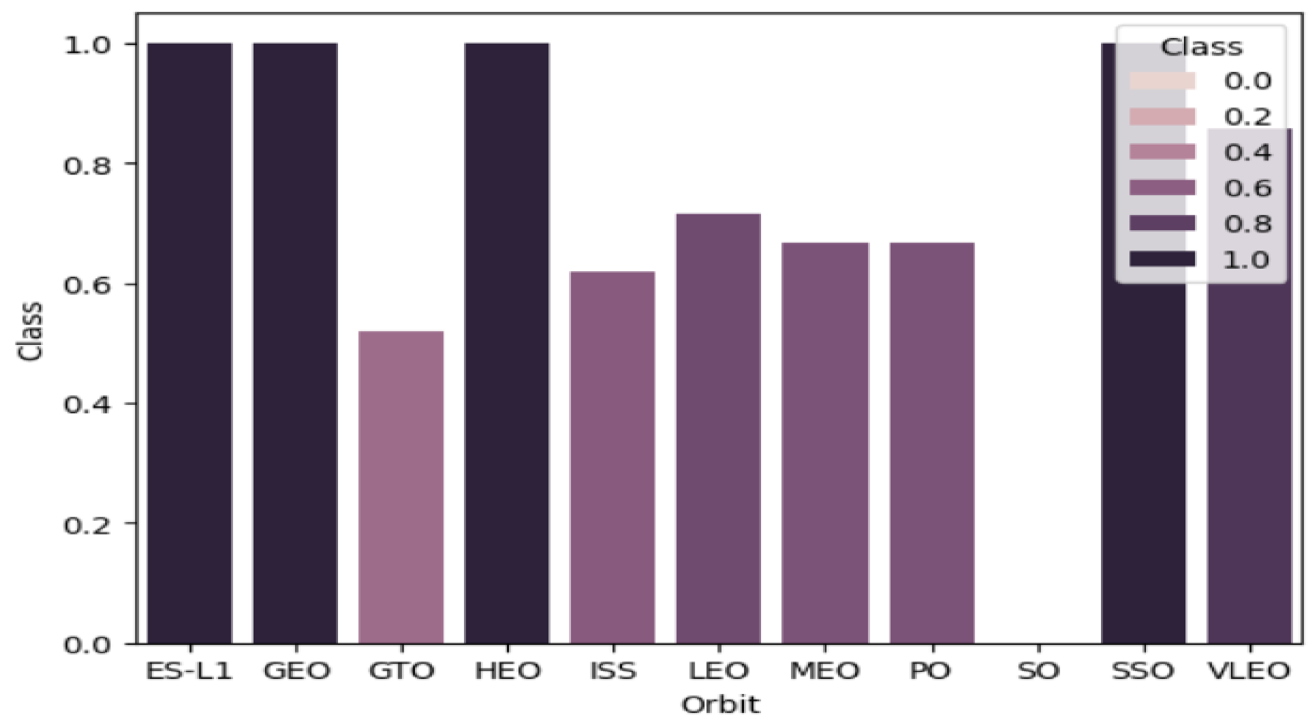


- Show the screenshot of the scatter plot with explanations

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

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

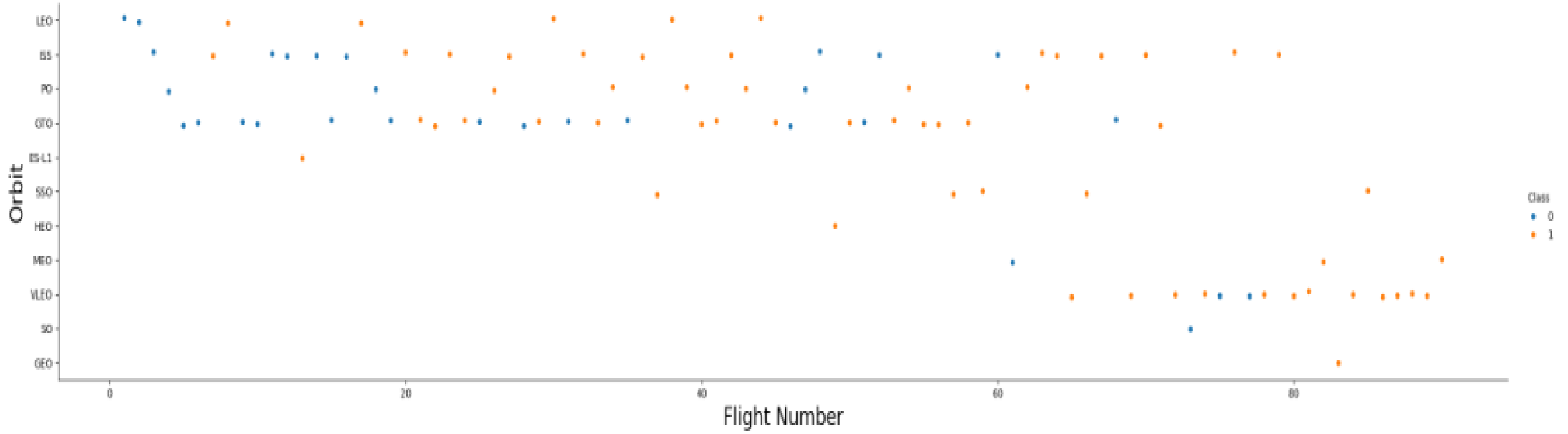


Analyze the ploted bar chart try to find which orbits have high sucess rate.

ES_L1, GEO, HEO & SSO have high success rate

Flight Number vs. Orbit Type

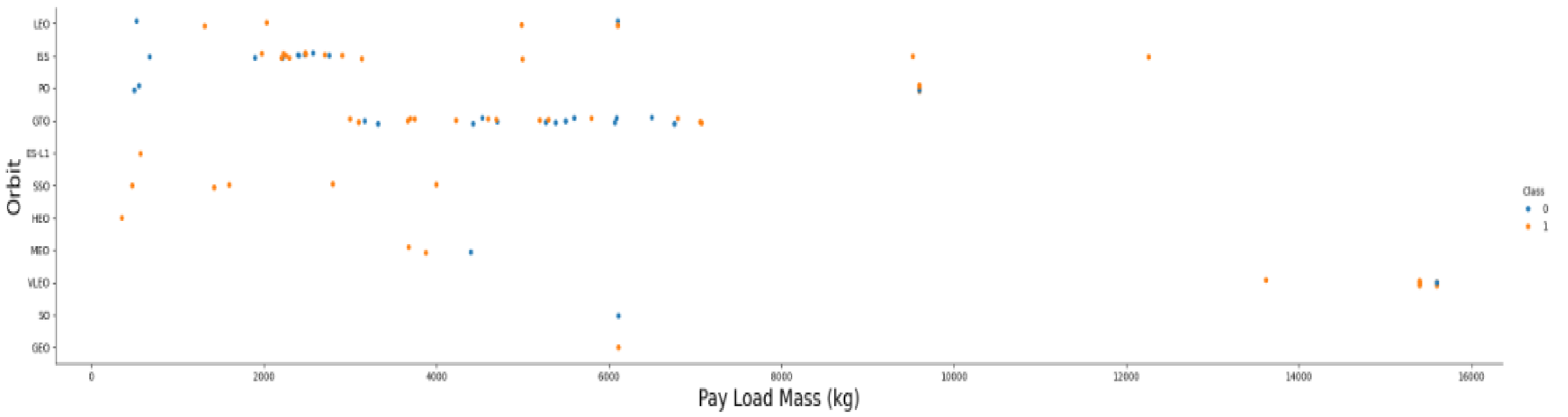
- Show a scatter point of Flight number vs. Orbit type



You should see that 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

- Show a scatter point of payload vs. orbit type

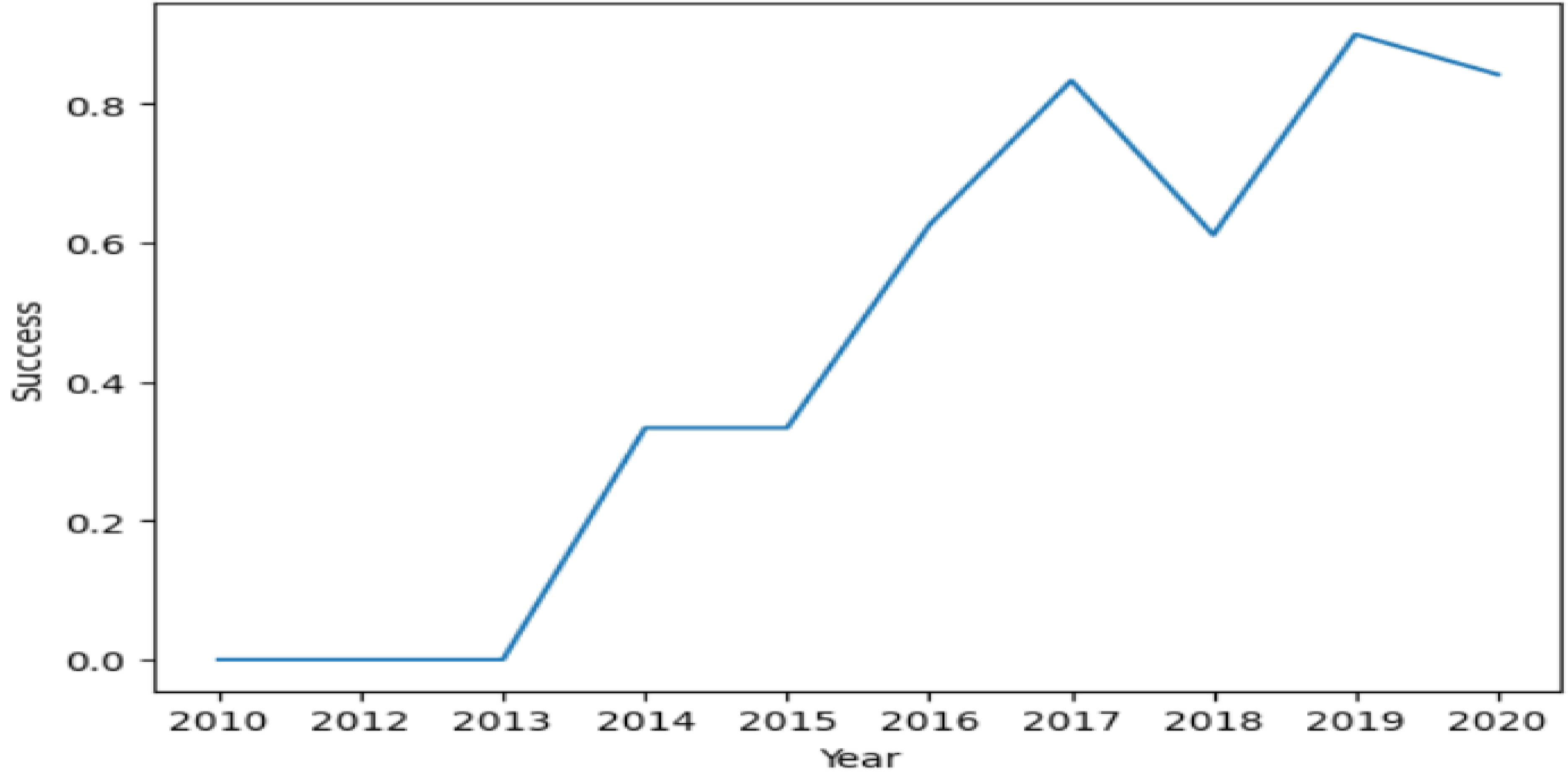


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

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

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites

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

- Present your query result with a short explanation here

```
%sql select DISTINCT ("LAUNCH_SITE") from SPACEXTBL;
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%sql select LAUNCH_SITE from SPACEXTBL WHERE (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

- Present your query result with a short explanation here

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

There is only one launch site beginning with ‘CCA’

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL WHERE "Customer" like 'NASA (CRS)%';
```

- Present your query result with a short explanation here

payloadmass
48213

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL WHERE Booster_Version = ‘F9 V1.12’;
```

- Present your query result with a short explanation here

payloadmass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql select min(Date) from SPACEXTBL WHERE Landing_Outcome='Success (ground pad)';
```

- Present your query result with a short explanation here

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

%sql select Booster_Version from SPACEXTBL WHERE LANDING_OUTCOME = 'Success(drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;

- Present your query result with a short explanation here

Booster_Version

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select MISSION_OUTCOME, count(*) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

- Present your query result with a short explanation here

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_)from SPACEXTBL);

boosterversion
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

2015 Launch Records

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

```
%sql select CASE SUBSTR("Date", 6, 2) WHEN '01' THEN 'JANUARY' WHEN '02' THEN 'FEBRUARY' WHEN '03' THEN 'MARCH' WHEN '04' THEN 'APRIL'\
WHEN '05' THEN 'MAY' WHEN '06' THEN 'JUNE' WHEN '07' THEN 'JULY' WHEN '08' THEN 'AUGUST' WHEN '09' THEN 'SEPT' WHEN '10' THEN 'OCTOBER'\
WHEN '11' THEN 'NOVEMBER' WHEN '12' THEN 'DECEMBER' ELSE 'Unknown' END AS MONTH, "Landing_Outcome" = 'Failure(drone ship)', "Booster_Version", \
"Launch_Site" from SPACEXTBL WHERE SUBSTR("Date", 0, 5)='2015';
```

MONTH	"Landing_Outcome" = 'Failure(drone ship)'	Booster_Version	Launch_Site
JANUARY	0	F9 v1.1 B1012	CCAFS LC-40
FEBRUARY	0	F9 v1.1 B1013	CCAFS LC-40
MARCH	0	F9 v1.1 B1014	CCAFS LC-40
APRIL	0	F9 v1.1 B1015	CCAFS LC-40
APRIL	0	F9 v1.1 B1016	CCAFS LC-40
JUNE	0	F9 v1.1 B1018	CCAFS LC-40
DECEMBER	0	F9 FT B1019	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

```
%sql select LANDING_OUTCOME, count(*) as count_outcomes from SPACEXTBL\ where DATE between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count_outcomes DESC;
```

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

Section 3

Launch Sites Proximities Analysis

All launch sites global map markers

Now, you can explore the map by zoom-in/out the marked areas , and try to answer the following questions:

- Are all launch sites in proximity to the Equator line?
- Are all launch sites in very close proximity to the coast?

Also please try to explain your findings.

Explanation

The launchsites are all close to the equator and parallel to the equator. The sites are also on coastal lands.

Mark the success/failed launches for each site on the map

From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

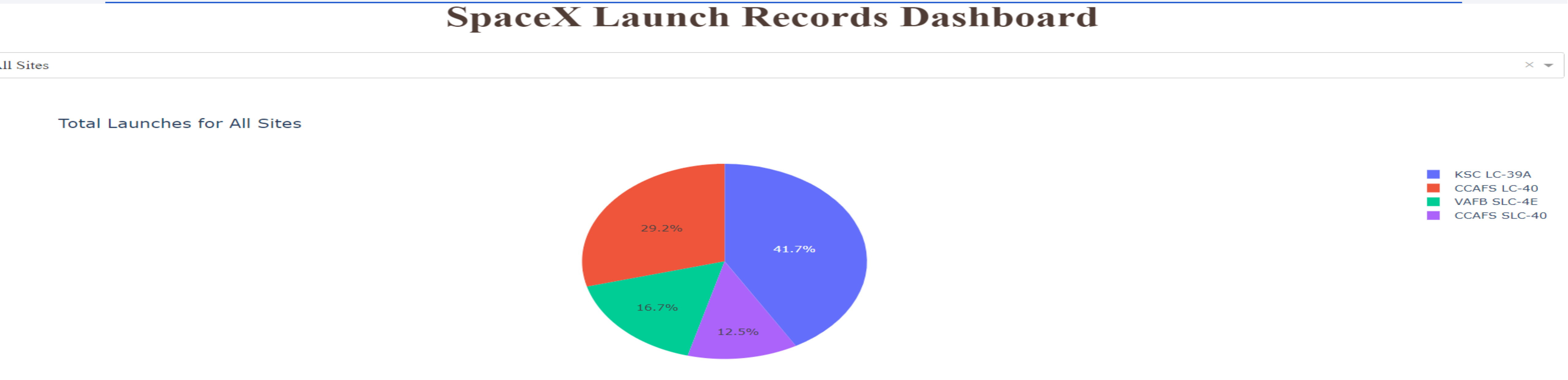
Calculate the distances between a launch site to its proximities

- From the plots railway lines and highways are about 1.28km from the launch site, to allow easy transportation of equipment and crew.
- The coastline is only 0.86km from the launch site, so rockets can fly over the ocean during launch also allowing crew to abort launches and have quick water landings. Also debris from the launch can fall into the ocean and not damage property.
- Launch sites are far from cities to avoid damage to property and people, in the case above Orlando city is 51.43km from launch site CCAFS SLC-40.

Section 4

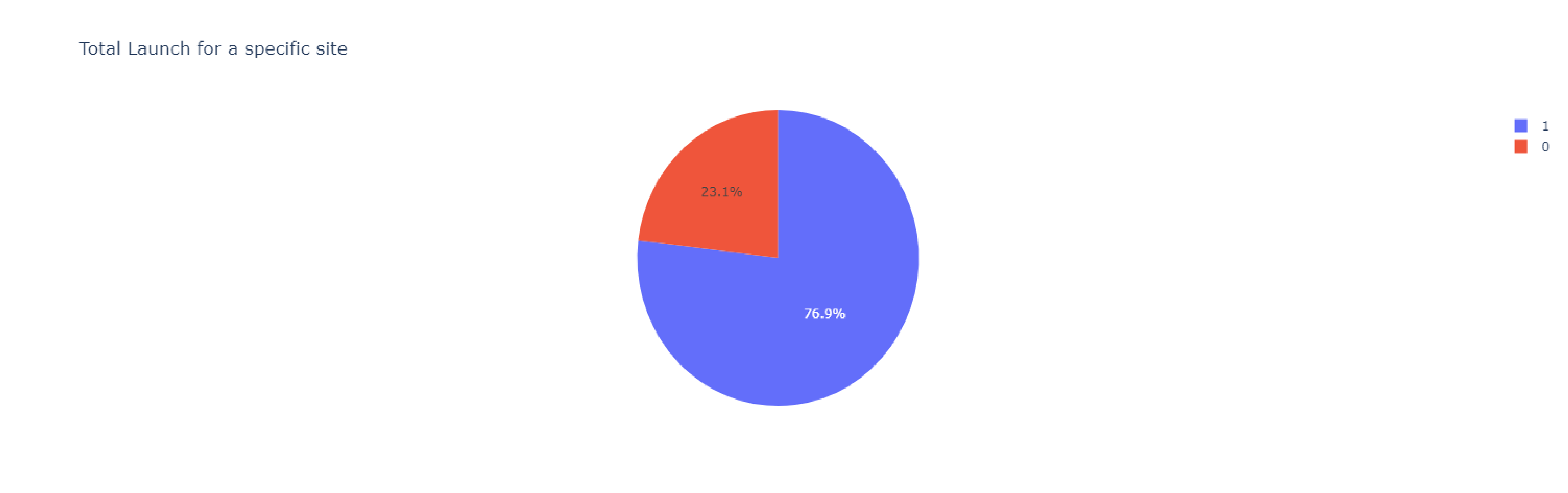
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



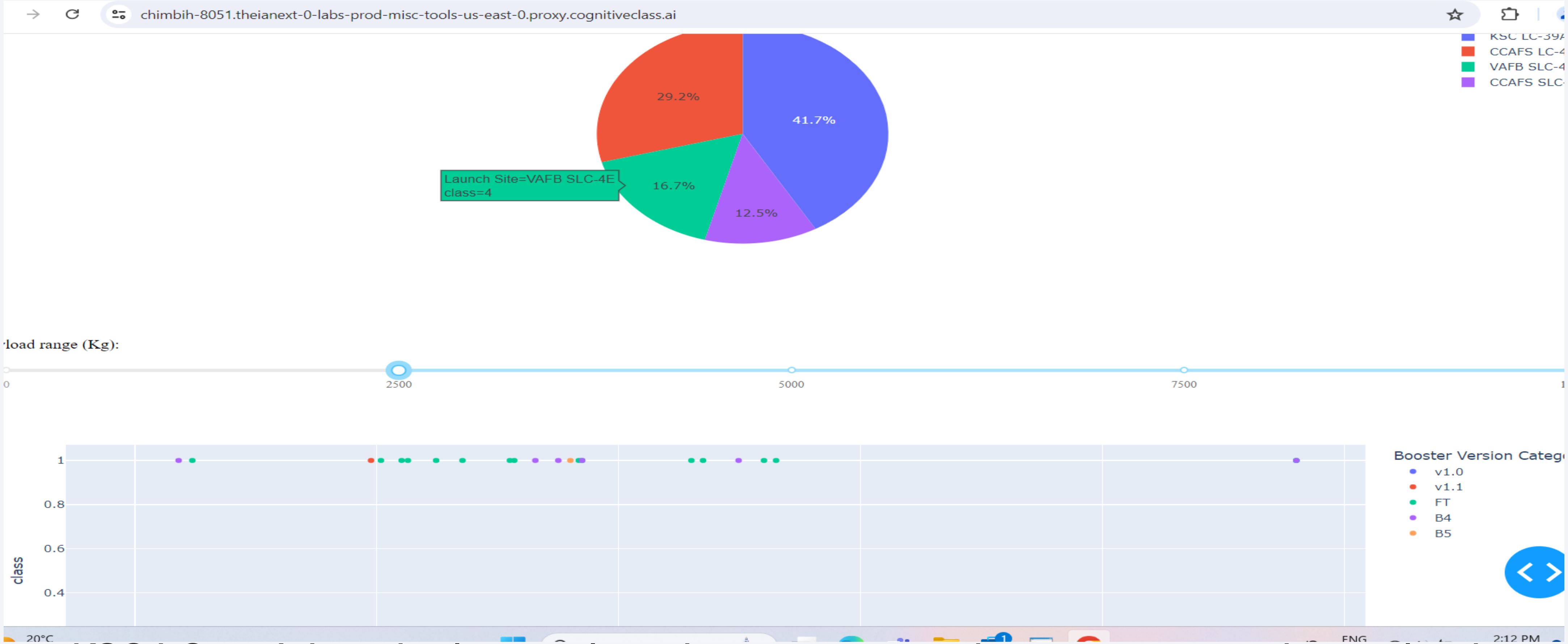
- Explain the important elements and findings on the screenshot
- KSC SC-39A has highest success with the largest share of 41.7% and CCAFS SLC-40 has the least success with 12.5%

Total Launch for a specific site



KSC LC-39A has 76.9% success rate and a 23.1% failure rate

Payload vs Launch outcome scatter plot for all sites

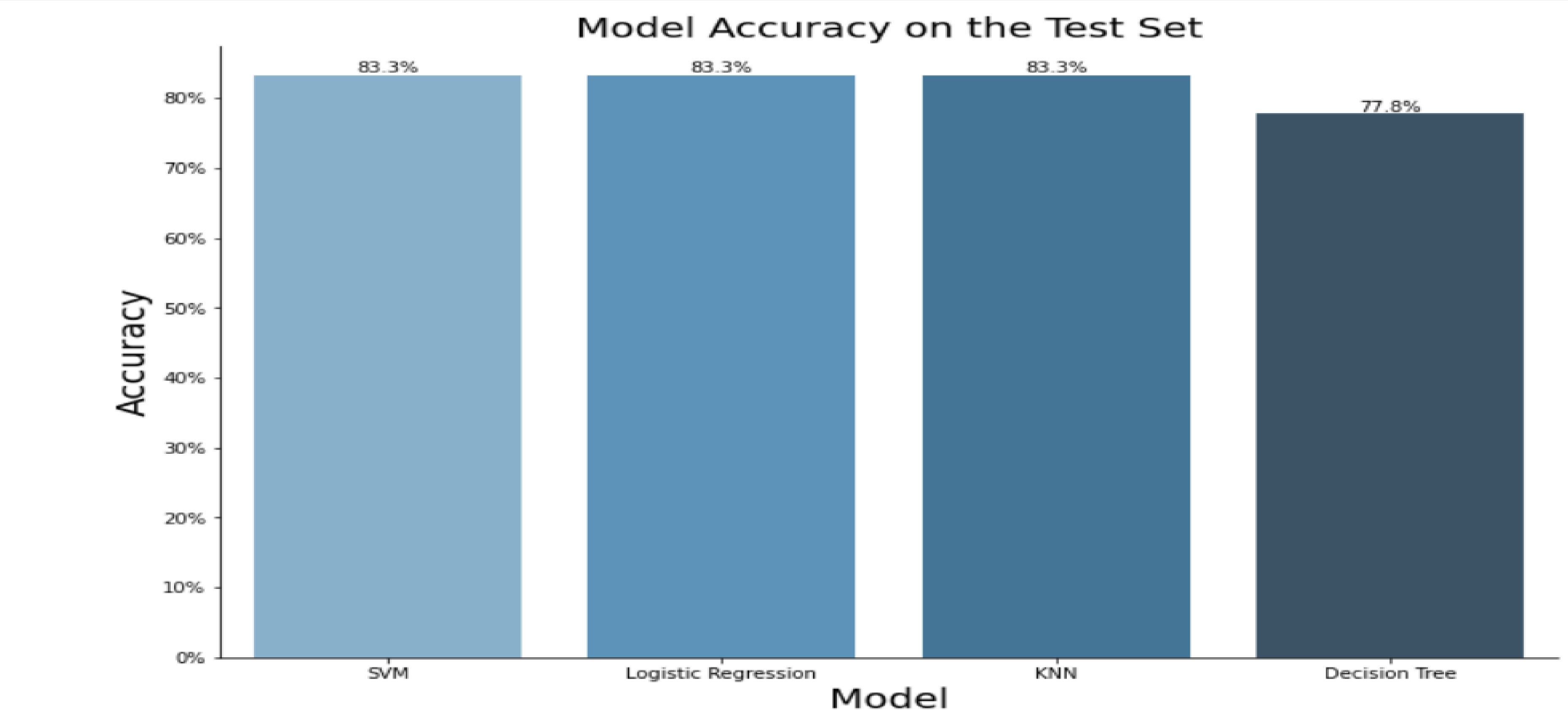


KSC LC-39A has the largest launch success rate, the 4000 – 6000kg payload mass has the highest success rate and the FT has the the highest launch success rate.

Section 5

Predictive Analysis (Classification)

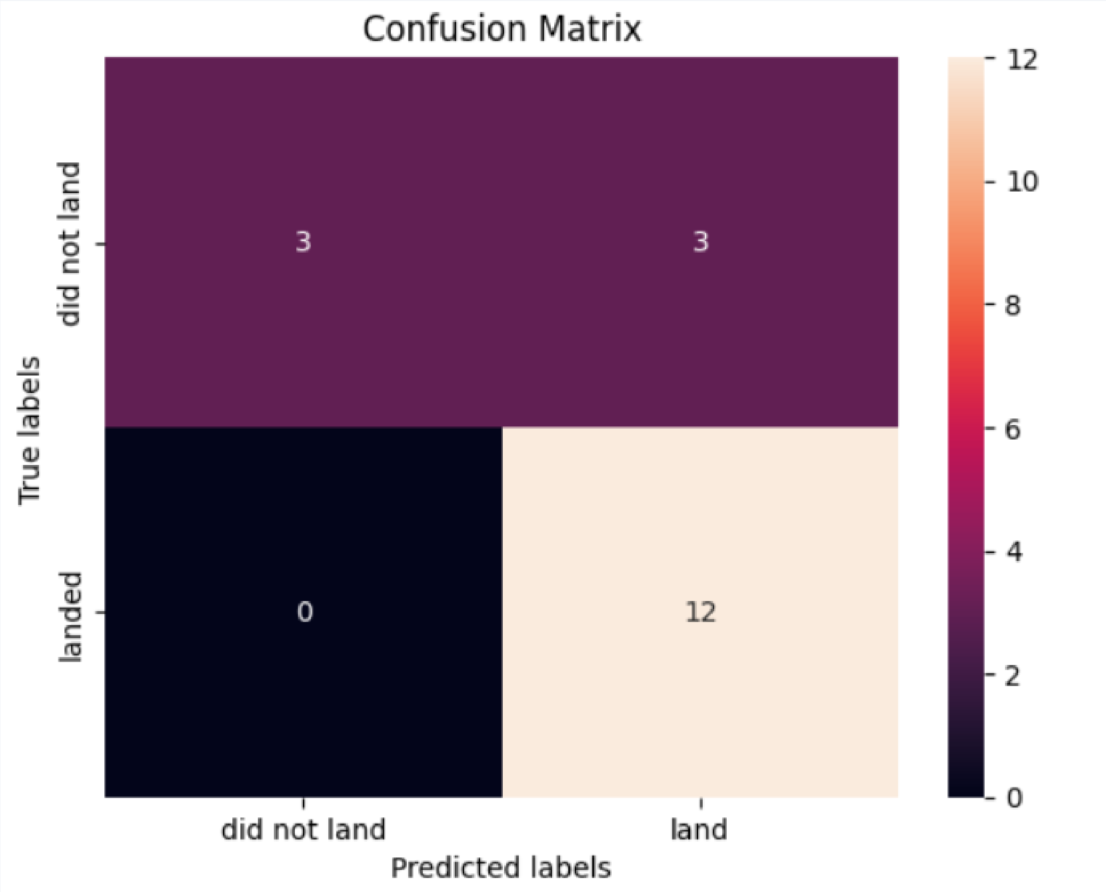
Classification Accuracy



The accuracy is basically the same around 0.83 for KNN, LR, SVM, although Decision Tree has a slightly lower accuracy it is not too far off.

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation

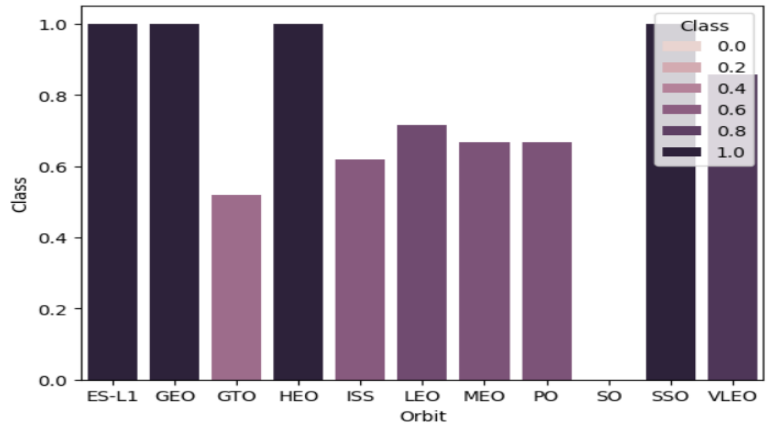


Conclusions

- Spacex success rate is steadly increasing over time
- There is a higher success rate with payloads in the 2500kg to 6000kg range than heavier weights
- KSC LC-39A is the most successful launch site
- Orbits: ES-L1, GEO, HEO & SSO have the highest success rates

Appendix

- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- **%sql** select CASE SUBSTR("Date", 6, 2) WHEN '01' THEN 'JANUARY' WHEN '02' THEN 'FEBRUARY' WHEN '03' THEN 'MARCH' WHEN '04' THEN 'APRIL'\ WHEN '05' THEN 'MAY' WHEN '06' THEN 'JUNE' WHEN '07' THEN 'JULY' WHEN '08' THEN 'AUGUST' WHEN '09' THEN 'SEPT' WHEN '10' THEN 'OCTOBER'\ WHEN '11' THEN 'NOVEMBER' WHEN '12' THEN 'DECEMBER' ELSE 'Unknown' END AS MONTH, "Landing_Outcome" = 'Failure(drone ship)', "Booster_Version", \ "Launch_Site" **from** SPACEXTBL WHERE SUBSTR("Date", 0, 5)='2015';



Analyze the ploted bar chart try to find which orbits have high sucess rate.

ES_L1, GEO, HEO & SSO have high success rate

Thank you!