IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Olugbenga P. Kayode
January 2024

Outline

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

Executive Summary

- This project represents a pioneering endeavor in the realm of data science, specifically tailored to SpaceX's extensive dataset on rocket launches. In an era where space exploration has transitioned into the hands of private entities, SpaceX has emerged as a frontrunner by successfully implementing cost-effective measures, notably through the reuse of initial mission stages.
- The focus is centered on leveraging data science methodologies to ascertain the success factors influencing fresh rocket launches. By employing predictive analytics, the aim is to extract meaningful insights from the rich data SpaceX has accumulated, ultimately contributing to the enhancement of launch success predictions.
- The project commenced with a comprehensive exploration of the publicly available dataset, encompassing launch sites, flight details and mission successes and failures . Through meticulous preprocessing and data cleansing procedures, we curated a robust dataset that served as the foundation for subsequent analyses.
- The core of our analytical approach lies in the application of advanced predictive analytics techniques. By utilizing, logistic regression, support vector machines (SVM), KNN and decision trees classifiers, the task sought to uncover hidden patterns and correlations within the data, offering a predictive framework for evaluating the success of forthcoming launches.
- As our iterative analyses progress, we anticipate not only identifying key success indicators but also contributing to the broader discourse on the future of space exploration. The findings presented in this report encapsulate the culmination of our efforts, providing stakeholders with actionable insights to inform decision-making processes within the dynamic landscape of rocket launches.

Introduction

- In this captivating capstone project, we embark on the thrilling task of predicting the successful landing of the Falcon 9 first stage. The stakes are high, considering SpaceX's strategic cost advantage – their Falcon 9 launches cost a mere 62 million dollars, in stark contrast to competitors charging upwards of 165 million dollars per launch. The key to this cost efficiency lies in SpaceX's groundbreaking ability to reuse the first stage.
- The mission is clear. By accurately predicting the first stage's landing outcome, we unlock the power to determine the overall cost of a launch. This invaluable information holds the potential to reshape the competitive landscape, particularly for companies vying with SpaceX in the rocket launch market. Picture this: armed with insights into landing success, alternate companies can strategically position themselves to bid more effectively against SpaceX.
- In this immersive lab experience, your journey begins with the collection and meticulous formatting of data sourced from a dynamic API. As we delve into the analysis, envision the impact of uncovering patterns that could redefine the space race. Get ready to explore the fascinating interplay of data science and space exploration, where each successful launch is not just a triumph in technology but a financial victory in the fiercely competitive market of rocket launches.

Section 1

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 was sourced from the Spacex's publicly available launch data via API request using python ***requests method*** in JSON format. The data contained massive information about SpaceX launches and specific data items needed to be extracted and normalized using pandas

- A series of helper functions were defined to help extract the needed data items from the mass identification numbers in the launch data

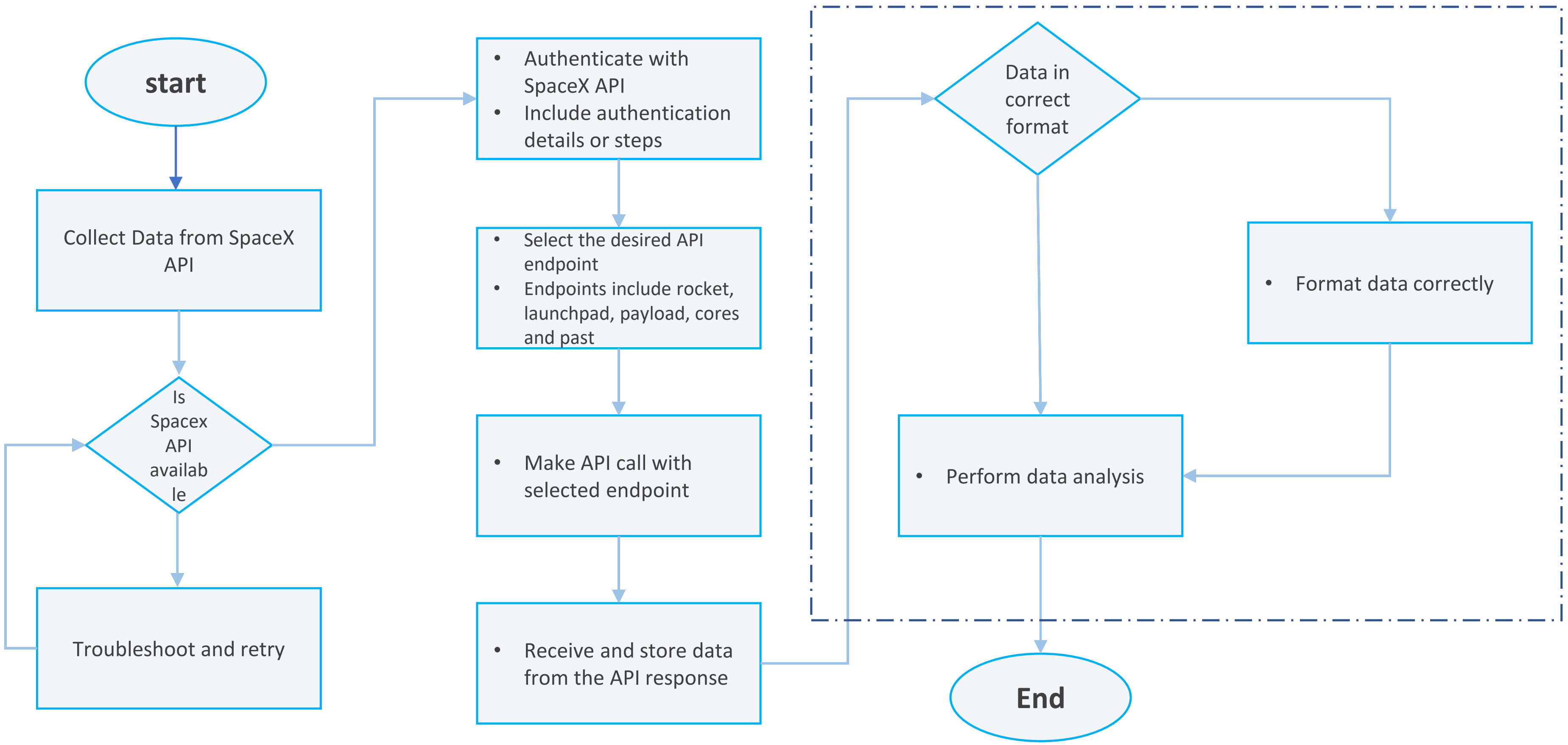
Endpoint	API	Helper Function	Data Obtained
Rocket	https://api.spacexdata.com/v4/rockets/	getBoosterVersion()	BoosterVersion (name)
Launchpad	https://api.spacexdata.com/v4/launchpads/	getLaunchSite()	LaunchSite (name), Longitude, Latitude
Payload	https://api.spacexdata.com/v4/payloads/	getPayloadData()	PayloadMass, Orbit
Cores	https://api.spacexdata.com/v4/cores/	getCoreData()	Block, Reused_Count, Serial, Landing_Success, Landing_Type, Flight, Gridfins, Reused, Legs, Landpad
Past (Rocket Launch)	https://api.spacexdata.com/v4/launches/past		

- The data obtained were combined into a dictionary and finally converted into a DataFrame.

- Please click here for the full SpaceX Data Collection Jupiter notebook on GITHUB.

[github.com/paulkayode2000/datasciencecoursera/blob/d59876ad54fa203ff2757e7794f8875417a31f39/jupyter_labs_spacex_data_collection_api.ipynb]

Data Collection – SpaceX API Flowchart

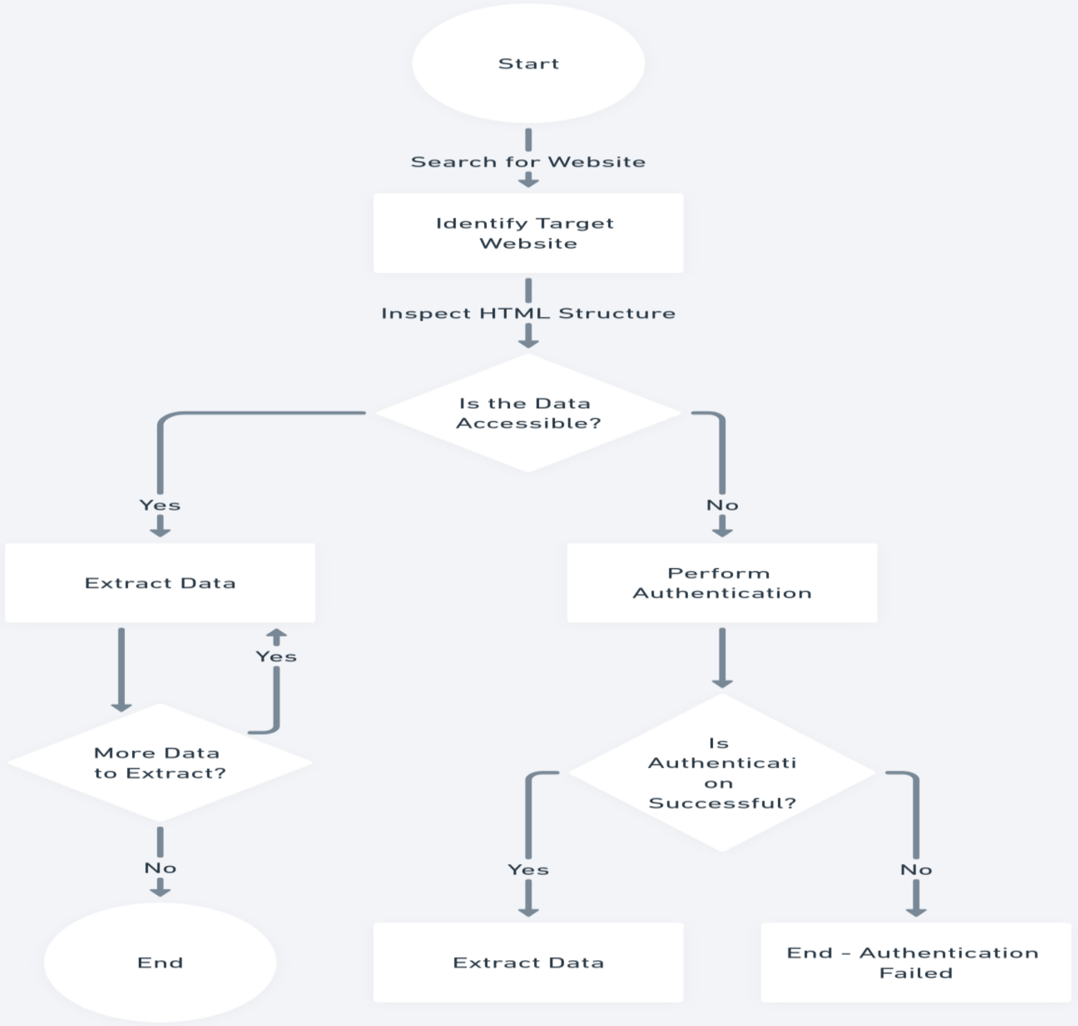


Data Collection - Scraping

- Web scraping was performed using Python's BeautifulSoup libraries to collect Falcon 9 historical launch records from a Wikipedia page titled "List of Falcon9 and Falcon Heavy launches"

- The complete Jupiter notebook is shared [here](https://github.com/paulkayode2000/datasciencecoursera/blob/d59876ad54fa203ff2757e7794f8875417a31f39/jupyter_labs_webscraping.ipynb)

[https://github.com/paulkayode2000/datasciencecoursera/blob/d59876ad54fa203ff2757e7794f8875417a31f39/jupyter_labs_webscraping.ipynb]



Data Wrangling

- The data was loaded with pandas into a dataframe and the following data wrangling operations were performed:

1. Convert data from JSON to dataframe using pandas
2. Filter data to include only “Falcon 9” records
3. Identify which columns are numerical and categorical
4. Calculate the percentage of missing values in each attribute and remove data with no date as irrelevant
5. Replace missing data for the identified 5 PayloadMass records with the mean value using imputation while the LandingPad column was retained for the 25 records with NULL values to represent when landing pads were not used.
6. Determine the number of launches on each site and the number of occurrence of each orbit and the mission outcome
7. Create a Landing Outcome label from Outcome column into a new column named “Class” and determine the success rate

The GitHub URL of the completed data wrangling notebooks can be found at:

https://github.com/paulkayode2000/datasciencecoursera/blob/d59876ad54fa203ff2757e7794f8875417a31f39/labs_jupyter_spacex_Data_wrangling.ipynb

EDA with Data Visualization

- Charts plotted and the reasons:

	VISUALIZATION	TYPE	OBJECTIVE	REMARK
1	FlightNumber vs. PayloadMass with Outcome (Class) overlay	Scatter Plot	To examine how the FlightNumber and PayloadMass variables would affect the launch outcome.	We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
2	FlightNumber vs Launch_Site with Outcome (Class) overlay	Scatter Plot	To examine the effect of FlightNumber and Launch_Site variables on the launch outcome.	The higher the number of flights undertaken on a site the higher the success rate.
3	Launch_Site vs PayloadMass	Scatter Plot	To examine the effect of PayloadMass and Launch_Site variables on the launch outcome.	There doesn't seem to be a general relationship but for the VAFB-SLC launchsite, there are no rockets launched for heavypayload mass (greater than 10000).
4	Success_Rate vs Orbit	Bar Chat	To visually observe the aggregated success rate of each Orbit.	Rockets launched into 4 orbits (ES-L1, GEO,SSO and VLEO) always had a 100% success rate.
5	Flight_Number vs Orbit type	Scatter Plot	To visually observe the aggregated success rate of each Orbit.	There is no general observation except for LEO orbit where higher flight numbers was associated with successful mission.
6	PayloadMass vs Orbit with outcome overlay	Scatter Plot	To examine the effect of PayloadMass and Orbit variables on the launch outcome.	No reasonable relationship observed except for LEO and ISS where heavy payload seemed to be associated with success.
7	Success rate over the years	Line grapgh	To visually examine the success trend over the years.	An oupward trend of successes was observed over the years.

- Here is the GitHub URL of the completed EDA with data visualization notebook [https://github.com/paulkayode2000/datasciencecoursera/blob/d59876ad54fa203ff2757e7794f8875417a31f39/Module2-jupyter-labs-eda-dataviz.ipynb]

EDA with SQL

The following 10 SQL queries were performed:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass, using a subquery
9. 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
10. 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.

GitHub URL of the completed EDA with SQL notebook can be found here
https://github.com/paulkayode2000/datasciencecoursera/blob/733ab677379ed700691a7d4be2a597cd421e0ade/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Summary of map objects such as markers, circles, lines, etc. created and added to a folium map

- Launch sites in relation to NASA base
- Success/Failed launches for each site
- Distances between a launch site to its proximities (highway, railway, coastline and the equator)

The markers were added to visually reveal each location on the map and the relationship to mission success.

GitHub URL:
https://github.com/paulkayode2000/datasciencecoursera/blob/733ab677379ed700691a7d4be2a597cd421e0ade/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

- The task was to optimize machine learning models, beginning with the partitioning the data into training and test sets, obtain the best scores and training data accuracy to unveil the ideal hyperparameters for

- Support Vector Machines (SVM)
- Classification Trees,
- Logistic Regression, and
- K-Nearest Neighbors Classifier method

This meticulous process sets the stage for uncovering the most powerful predictive method by rigorously evaluating and comparing their performance using the test dataset.

GitHub URL:

https://github.com/paulkayode2000/datasciencecoursera/blob/b300c2bd20e657e90c1186991a8145cc03783ec3/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

Exploratory Data Analysis

	PARAMETERS	RESULT
1	Launch Sites	There were 4 distinct launch sites in the data analyzed, with 2 each on the east and west coast
2	FlightNumber vs. PayloadMass with Outcome (Class) overlay	We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
3	FlightNumber vs Launch_Site with Outcome (Class) overlay	The higher the number of flights undertaken on a site the higher the success rate.
4	Success_Rate vs Orbit	Rockets launched into 4 orbits (ES-L1, GEO,SSO and VLEO) always had a 100% success rate.
5	Success rate over the years	An oupward trend of successes was observed over the years between 2013 and 2020

Results

Interactive Analytics

	PARAMETERS	RESULT
1	Location	There were 4 distinct launch sites in the data analyzed, with 2 each on the east and west coast
2	Proximity to Equator	None
3	Proximity to the coast	All launch sites were situated close to the sea
4	Proximity to highway	None
5	Proximity to Railway	None

Results

Machine Learning Predictive Models

	ML Algorithm	Accuracy on test data	Remark
1	Logistic Regression	0.833333333	Next best prediction model after DTC
2	SVM	N/A	Could not be run successfully due to high computing resource demand
3	Decision Tree Classifier	0.871428571	Best ML model for predicting the success of a rocket launch
4	K Neighbor Classifier	0.664285714	Worst of the 3 ML model for the project's Objective

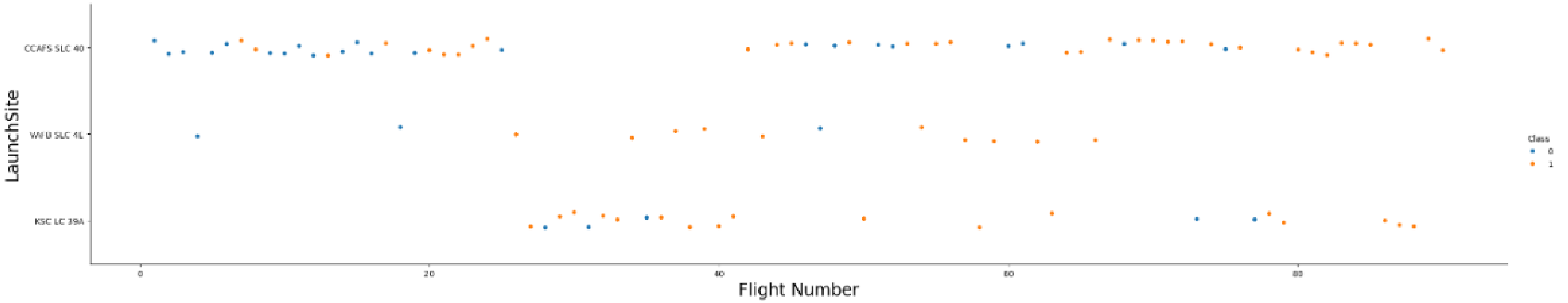
The Decision Tree Classifier ML model was best for predicting the success of a Falcon 9 rocket launch

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

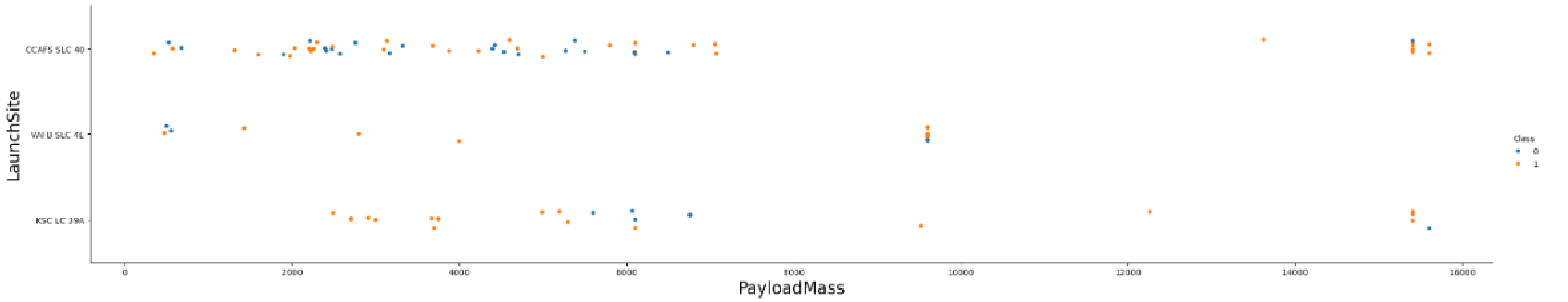
```
### TASK 1: Visualize the relationship between Flight Number and Launch Site
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



INSIGHT: The higher the number of flights undertaken on a site the higher the success rate.

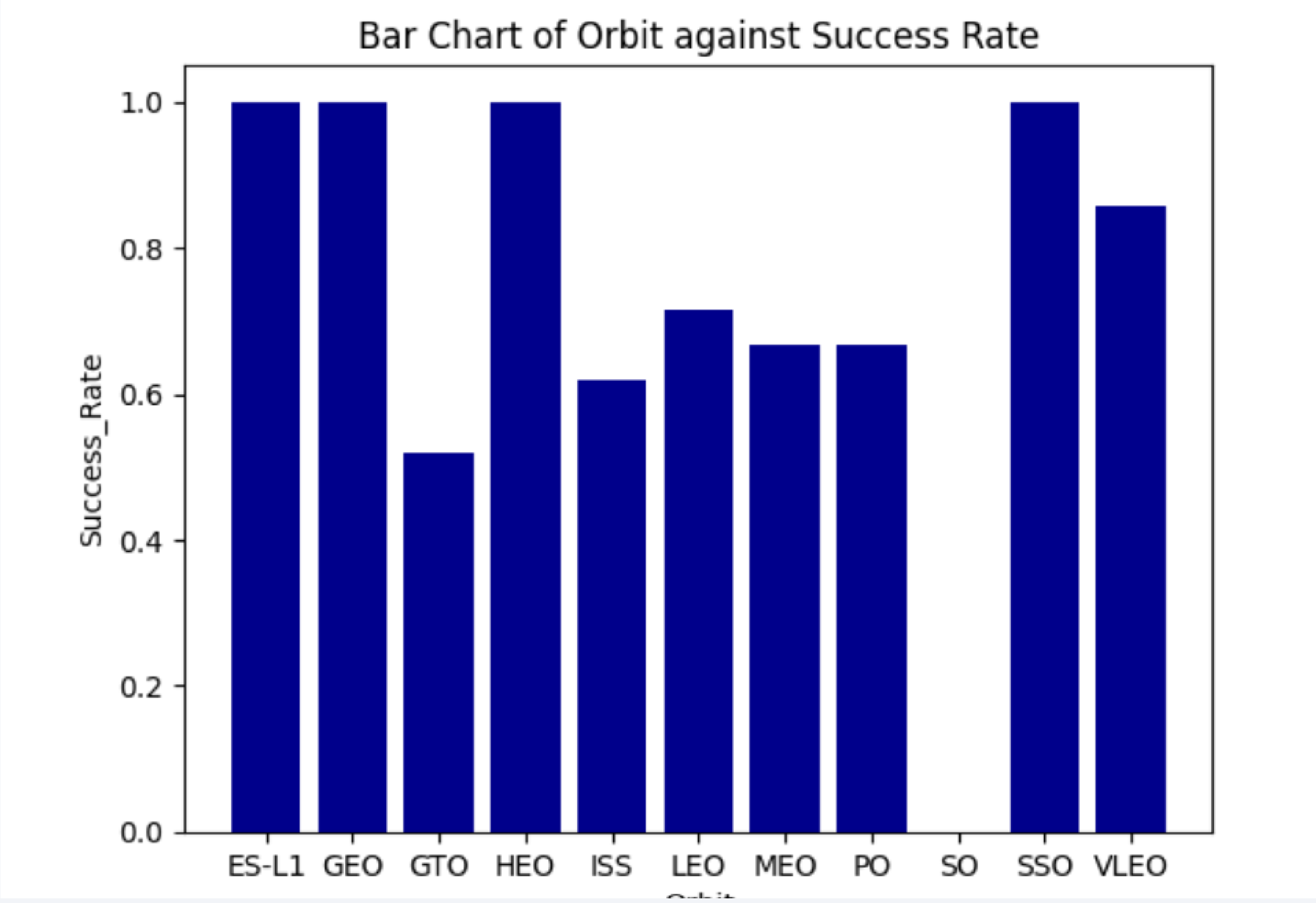
Payload vs. Launch Site

```python
### TASK 2: Visualize the relationship between Payload and Launch Site
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



There doesn't seem to be a general relationship but for the VAFB-SLC launchsite, there are no rockets launched for heavypayload mass (greater than 10000).

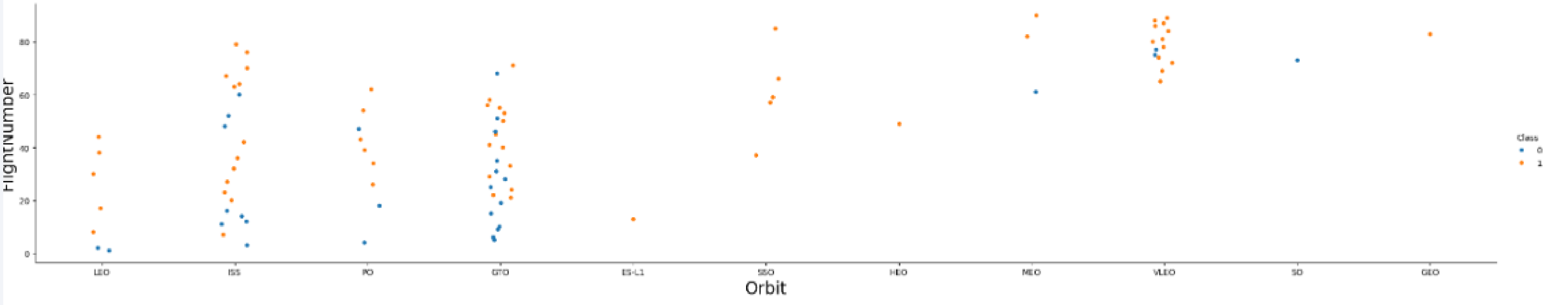
Success Rate vs. Orbit Type



Rockets launched into 4 orbits (ES-L1, GEO,SSO and VLEO) always had a 100% success rate

Flight Number vs. Orbit Type

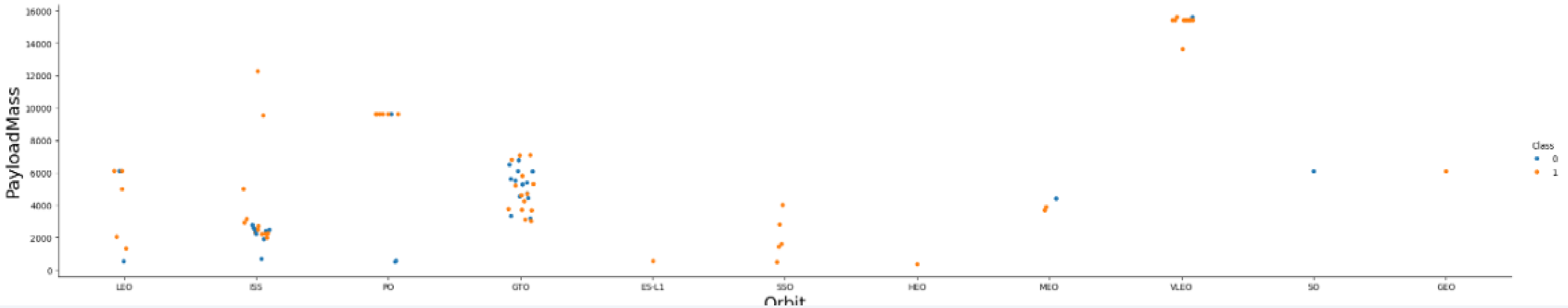
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be
sns.catplot(y="FlightNumber", x="Orbit", hue="Class", data=df, aspect = 5)
plt.xlabel("Orbit",fontsize=20)
plt.ylabel("FlightNumber",fontsize=20)
plt.show()
```



There is no general observation except for LEO orbit where higher flight numbers was associated with successful mission.

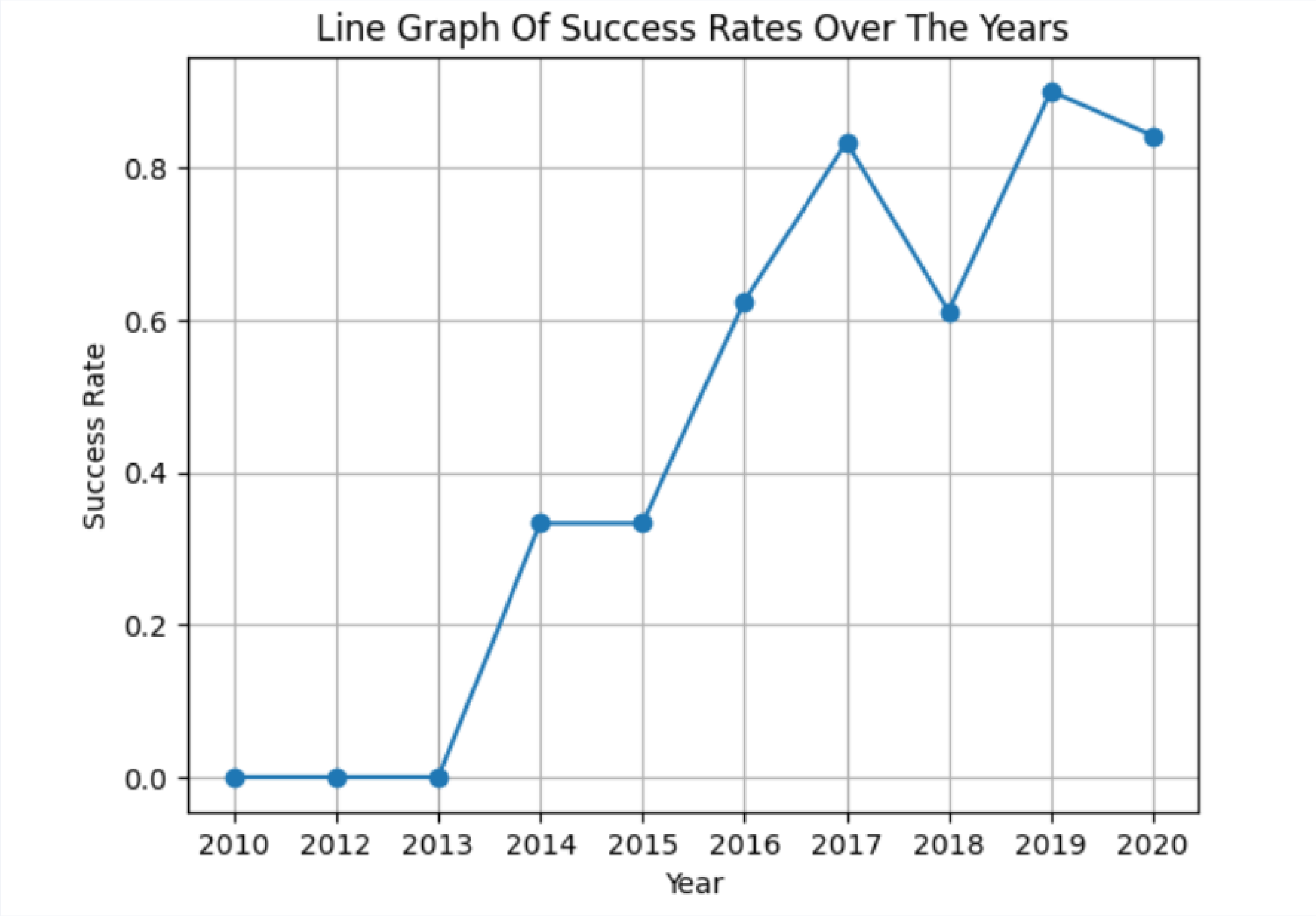
Payload vs. Orbit Type

```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the
sns.catplot(y="PayloadMass", x="Orbit", hue="Class", data=df, aspect = 5)
plt.xlabel("Orbit",fontsize=20)
plt.ylabel("PayloadMass",fontsize=20)
plt.show()
```



No reasonable relationship observed except for LEO and ISS where heavy payload seemed to be associated with success.

Launch Success Yearly Trend



An upward trend of successes was observed over the years between 2013 and 2020

All Launch Site Names

Using SQL, 4 distinct launch sites were discovered from the data as shown below:

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

```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

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

```
%sql select * from SPACEXTABLE where substr(Launch_Site, 1, 3) = '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

Total Payload Mass

Total payload carried by boosters from NASA

```
# Display the total payload mass carried by boosters launched by NASA (CRS)
%sql select sum([PAYLOAD_MASS__KG_]) from SPACEXTABLE
```

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

sum([PAYLOAD_MASS__KG_])
619967

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

```
%sql select AVG([PAYLOAD_MASS__KG_]) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

AVG([PAYLOAD_MASS__KG_])
2928.4

First Successful Ground Landing Date

Date of the first successful landing outcome on ground pad

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

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 SPACEXTABLE where Mission_Outcome = 'Success' and [PAYLOAD_MASS__KG_] between 4000 and 600
```

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

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

```
%sql select count(*) from SPACEXTABLE where Mission_Outcome = 'Success'
```

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

count(*)
98

```
%sql select count(*) from SPACEXTABLE where Mission_Outcome != 'Success'
```

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

count(*)
3

There were 98 successful and only 3 failed missions observed using SQL

Boosters Carried Maximum Payload

- Names of the booster versions which have carried the maximum payload mass

- There were 12 Booster Versions identified

```
%sql select Booster_Version from SPACEXTABLE where [PAYLOAD_MASS__KG_] = (select max([PAYLOAD_MASS__KG_]) from SPACEXTABLE)
```

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

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

2015 Launch Records

- List of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 2 records were found as shown here

```
%sql select substr(Date, 6,2) as Month_In_2015, Launch_Site, Booster_Version, Landing_Outcome from SPACEXTABLE where substr(
```

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

Month_In_2015	Launch_Site	Booster_Version	Landing_Outcome
01	CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
04	CCAFS LC-40	F9 v1.1 B1015	Failure (drone ship)

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

```
%sql select Landing_Outcome, count(*) as Ranking from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcom order by Ranking desc;
```

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

```
_Outcome, count(*) as Ranking from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Ranking desc;
```

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

Landing_Outcome	Ranking
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

There were 8 Landing Outcomes with the highest ranked being “No attempt”

Section 3

Launch Sites Proximities Analysis

Lauch Site Locations Analysis

There were 2 launch sites each on the east and west coasts as shown above.

- Are launch sites in close proximity to railways? No

- Are launch sites in close proximity to highways? No

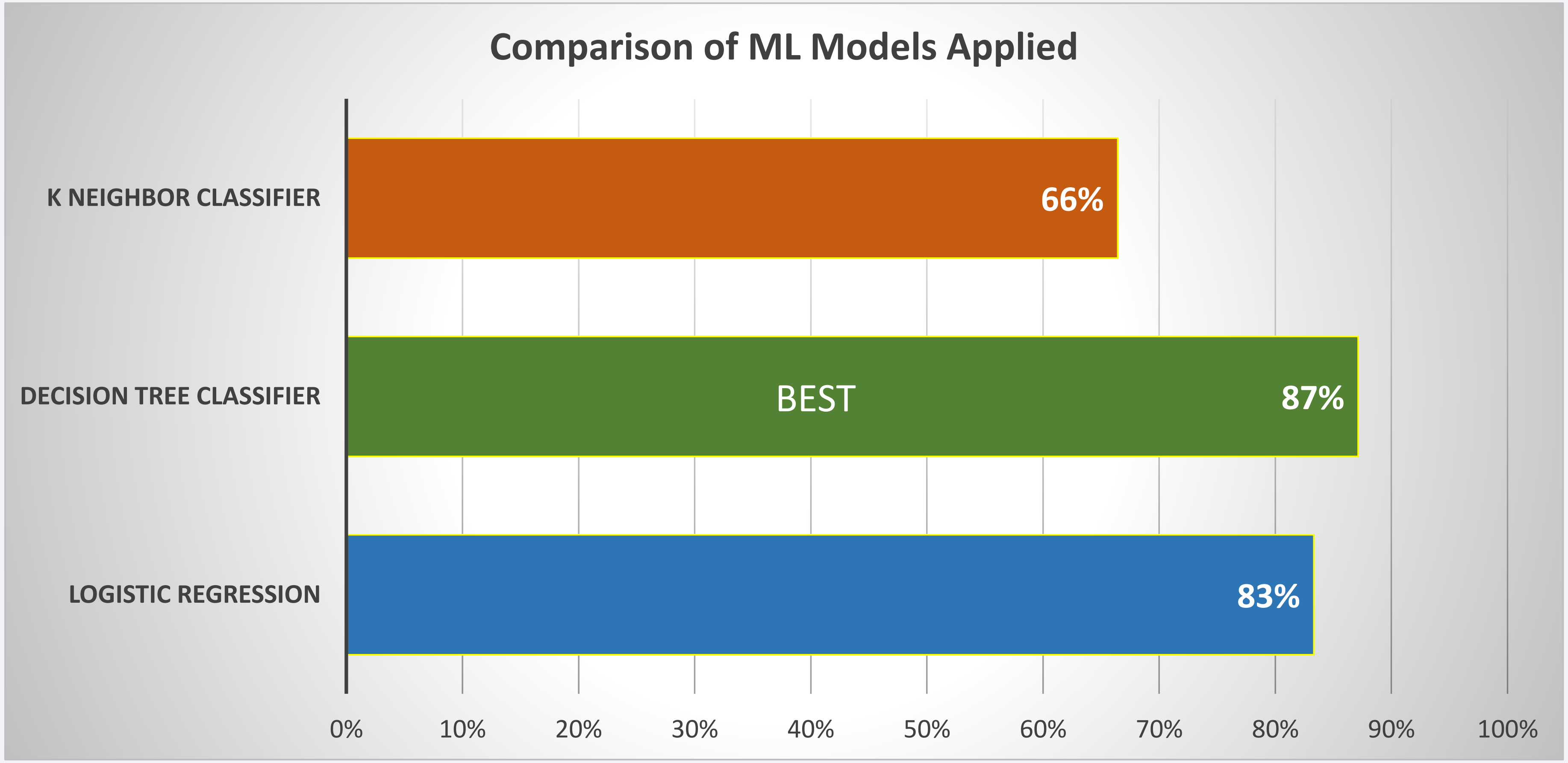
- Are launch sites in close proximity to coastline? Yes

- Do launch sites keep certain distance away from cities? Yes

Section 5

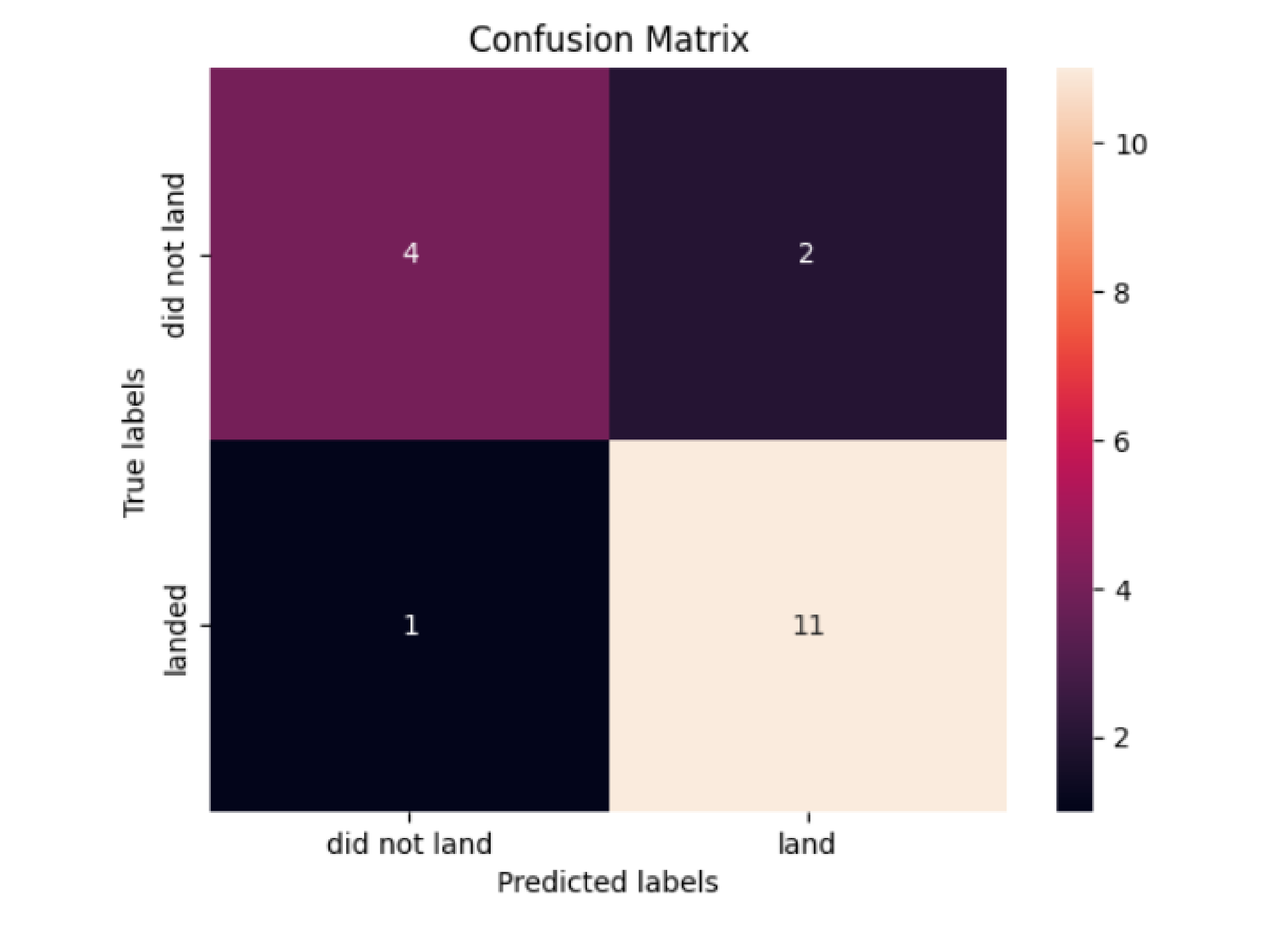
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

Confusion matrix of the best performing Decision Tree Classifier Model



Conclusions

- Rockets launched into 4 orbits (ES-L1, GEO,SSO and VLEO) always had a 100% success rate
- There were 4 distinct launch sites of which 2 each are located on the east and west coasts of the country but away from highways and railways. The sites are also far away from the earths equator.
- An upward trend in successful launches were recorded between 2013 and 2020
- The higher the number of flights undertaken on a site the higher the success rate; the more massive the payload, the less likely the first stage can be reused.
- The Decision Tree Classifier ML model was best for predicting the success of a Falcon 9 rocket launch with 87% prediction accuracy

Thank you!