IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Federica Zanca
27/05/2024

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Executive Summary

- Summary of methodologies
  - Data collection
    - API, Web scraping
  - Data Wrangling
  - Exploratory data analysis (EDA)
    - SQL, Data visualization (matplotlib, pandas)
  - Interactive visual analytics
    - Folium, Dashboard
  - Predictive analysis (classification)
    - Logistic regression, SVM, Classification tree, KNN

- Summary of results
  - EDA: Launch success improves over time, relationship[s between launch sites, orbit types, payload, flight number
  - Visualisation of where the launch sites are located
  - Predictive analytics shows similar results for different classification models

Introduction

- **Project Background and Context**
- The commercial space age is here, making space travel affordable for everyone.
- Companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX are leading the charge.
- **SpaceX Accomplishments:**
  - Sending spacecraft to the International Space Station.
  - Starlink satellite internet constellation.
  - Manned missions to space.
  - Cost-effective launches with reusable rockets.
- **Problems You Want to Find Answers**
- **Key Question:** Will the first stage of SpaceX's Falcon 9 rocket land successfully?
- **Goal:** Determine the cost of each rocket launch (which depends on whether the first stage will land)
- **Approach:**
  - Gather information about SpaceX's launches.
  - Create dashboards for analysis.
  - Use machine learning to predict the reuse of the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using web scraping and API
- Perform data wrangling
  - Changing categorical features (success or failure specifically) to numbers using one hot encoding.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different machine learning models were used for classification: logistic regression, decision tree, svm and knn. For each, the best parameters to fit the model were found using grid search.

Data Collection

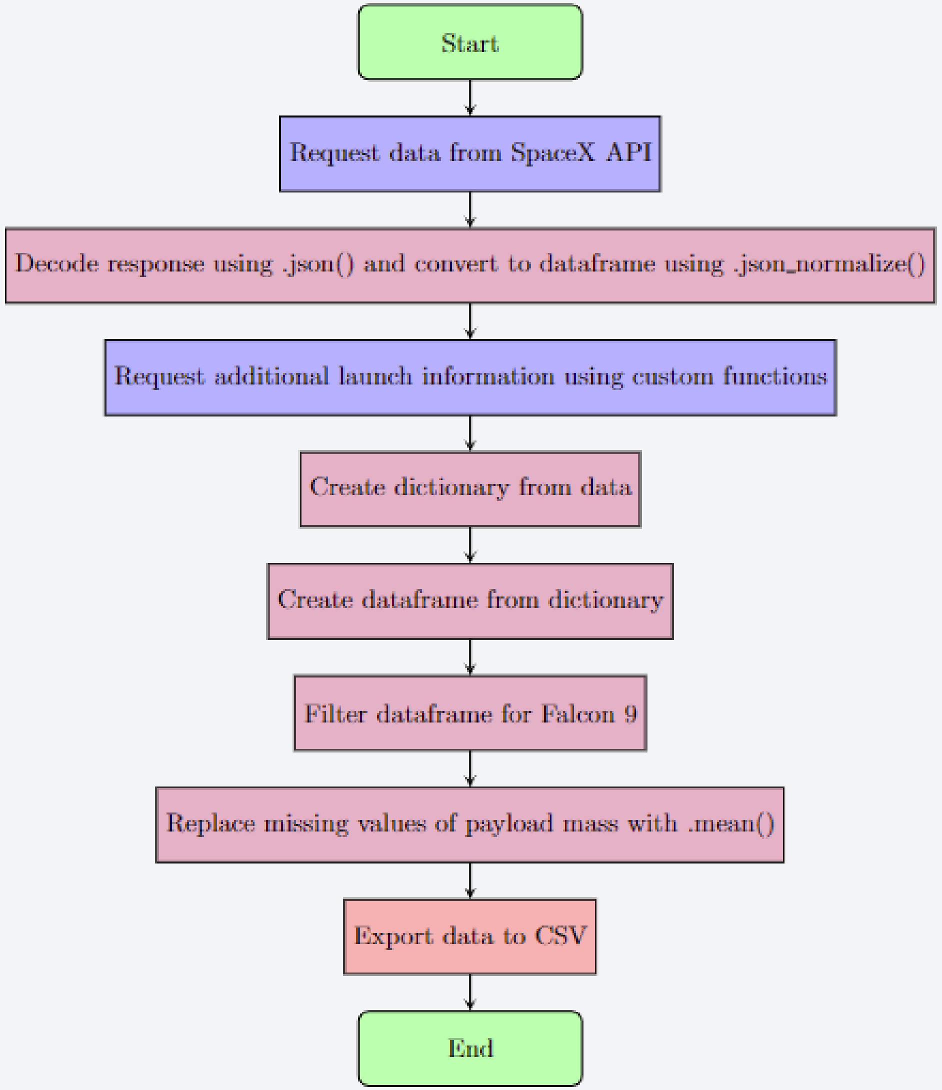
- The data was collected using 2 different methods: API and web scraping

- Get request to the SpaceX API.
- Using .json() function decode the response and read it as a pandas dataframe using .json_normalize().
- Moreover, web scraping from Wikipedia was performed for Falcon 9 launch records with BeautifulSoup.
- Data was cleaned and null values removed for further analysis

Data Collection – SpaceX API

- Data collection with SpaceX REST API

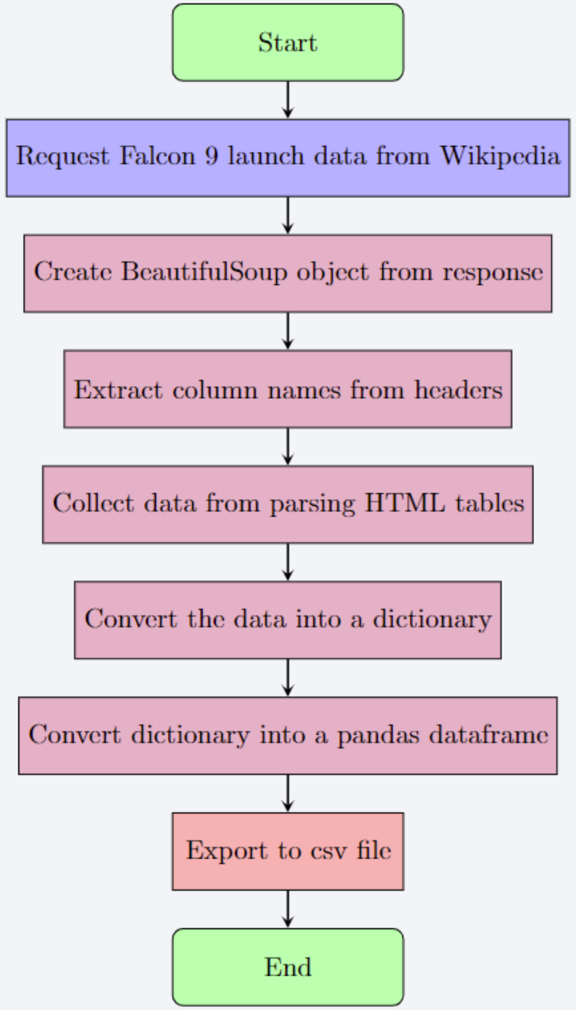
- The code can be found at the GitHub URL: https://github.com/federicazanca/datasc/blob/main/Data%20science%20capstone/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Webscraping Falcon 9 launch records with BeautifulSoup

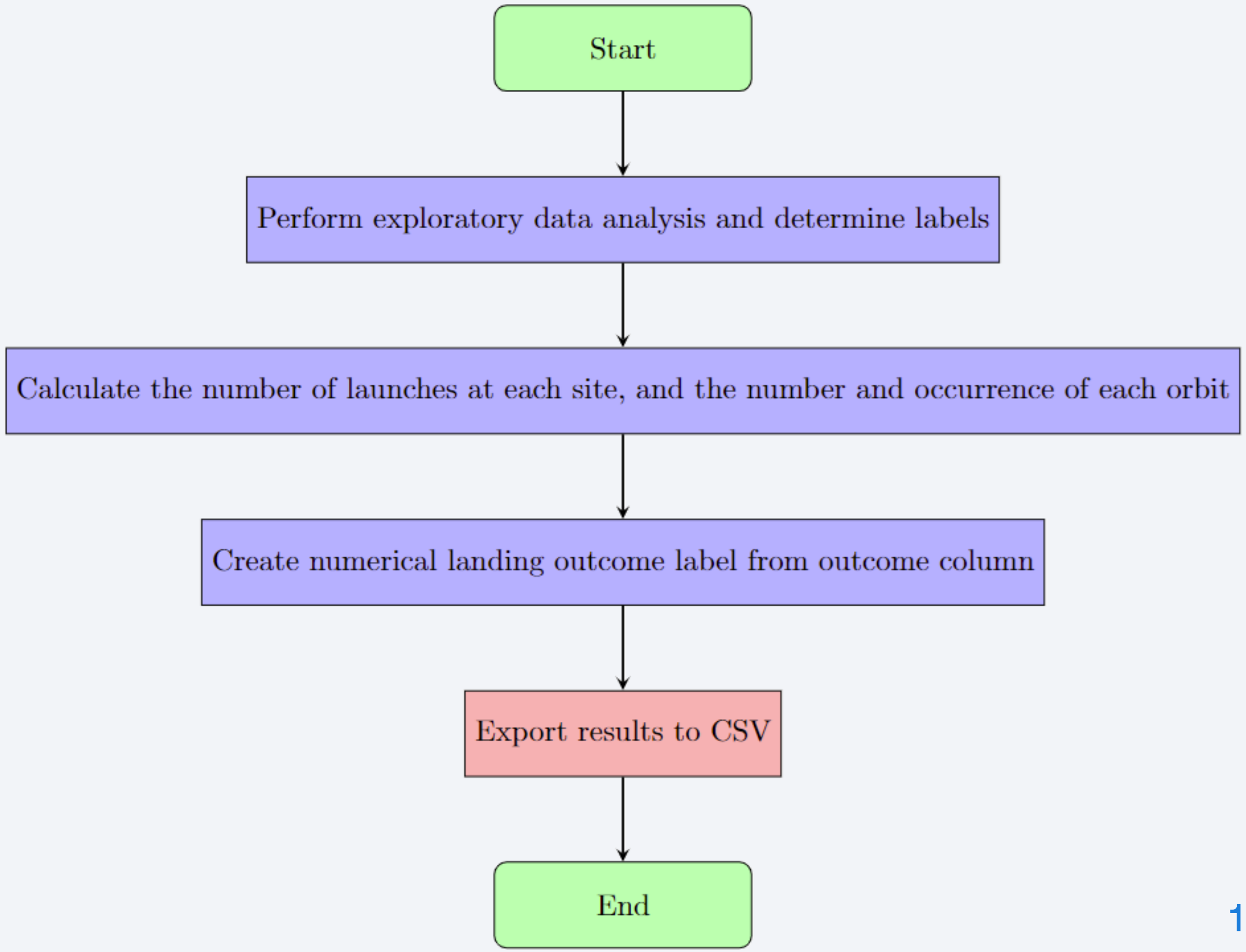
- The code can be found at the GitHub URL: https://github.com/federicazanca/datasc/blob/main/Data%20science%20capstone/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data wrangling: EDA and converting categorical features to numerical ones.

- The code can be found at the GitHub URL: https://github.com/federicazanca/datasc/blob/main/Data%20science%20capstone/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- GitHub URL: https://github.com/federicazanca/datasc/blob/main/Data%20science%20capstone/edadataviz.ipynb

Presented plots (in results section):

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit Type

They are in the form of

Scatter Plots: Explore relationships between variables.

Bar Charts: Compare discrete categories, illustrate relationships among categories and measured values.

EDA with SQL

GitHub URL: https://github.com/federicazanca/datasc/blob/main/Data%20science%20capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

SQL queries to show:

- Names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first succesful landing outcome in ground pad was acheived.
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster_versions which have carried the maximum payload mass. Use a subquery
- Month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- How many landing outcomes (such as Failure (drone ship) or Success (ground pad)) are between the date 2010-06-04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium

Comprehensive Launch Sites and Outcomes Visualization

Markers Indicating Launch Sites

- Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates.

- Added red circles at all other launch sites coordinates with popup labels showing their names using their latitude and longitude coordinates.

Coloured Markers of Launch Outcomes

- Added coloured markers of successful and unsuccessful launches at each launch site to show which launch sites have high success rates.

- Assigned launch outcomes to class 0 for failure and class 1 for success to visually differentiate them on the map.

Distances Between a Launch Site to Proximities

- Added coloured lines to show distances between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city.

- Calculated distances to determine whether launch sites are near railways, highways, and coastlines, and if they maintain a certain distance from cities.

Map with Folium

- Combined the above elements into a comprehensive visualization using Folium, providing a detailed overview of launch site locations, their success rates, and their proximities to important geographical features.

Build a Dashboard with Plotly Dash

GitHub: https://github.com/federicazanca/datasc/blob/main/Data%20science%20capstone/spacex_dash_app.py

Dropdown List with Launch Sites

- Allow users to select all launch sites or a specific launch site for detailed analysis.

Slider of Payload Mass Range

- Allow users to select a payload mass range to filter the data displayed.

Pie Chart Showing Successful Launches

- Provide a pie chart to display successful and unsuccessful launches as a percentage of the total launches.
- Allow users to view total launches by specific sites through interactive pie charts.

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Provide a scatter chart to show the correlation between payload mass and launch success.
- Allow users to see the relationship between payload mass (Kg) and outcome for different booster versions.

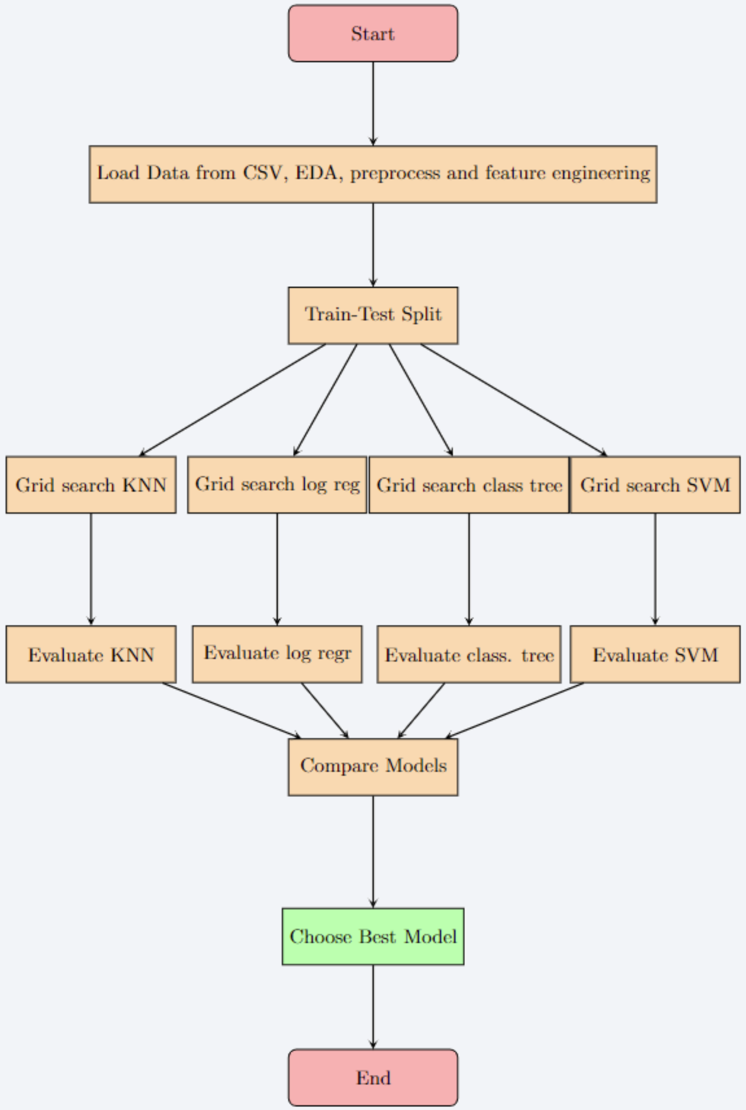
Dashboard with Plotly Dash

- Built an interactive dashboard using Plotly Dash to integrate these features, enabling users to dynamically explore and analyze the launch data.

Predictive Analysis (Classification)

GitHub:
https://github.com/federicazanca/datasc/blob/main/Data%20science%20capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

We found the best hyperparameters for 4 classification models and trained them to get the best predictions



Results

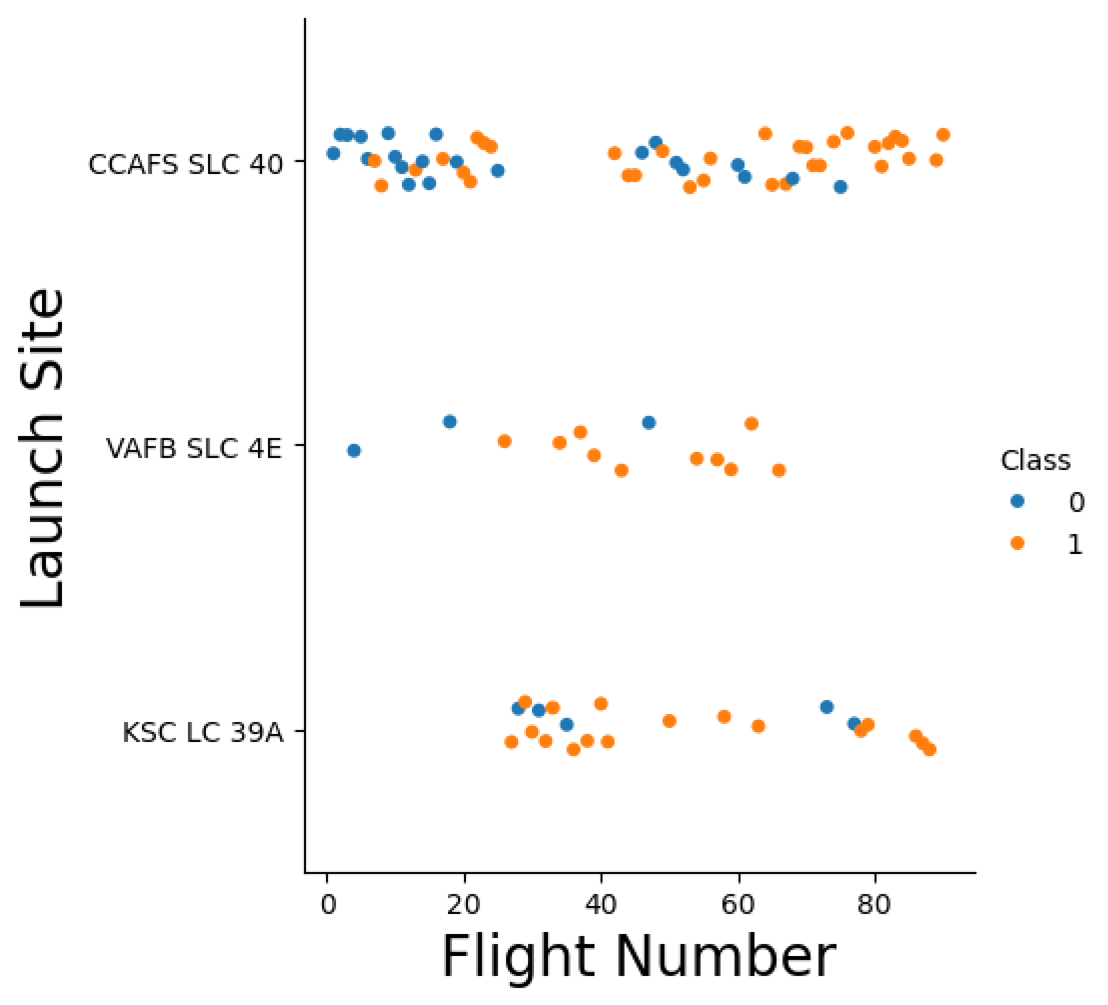
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

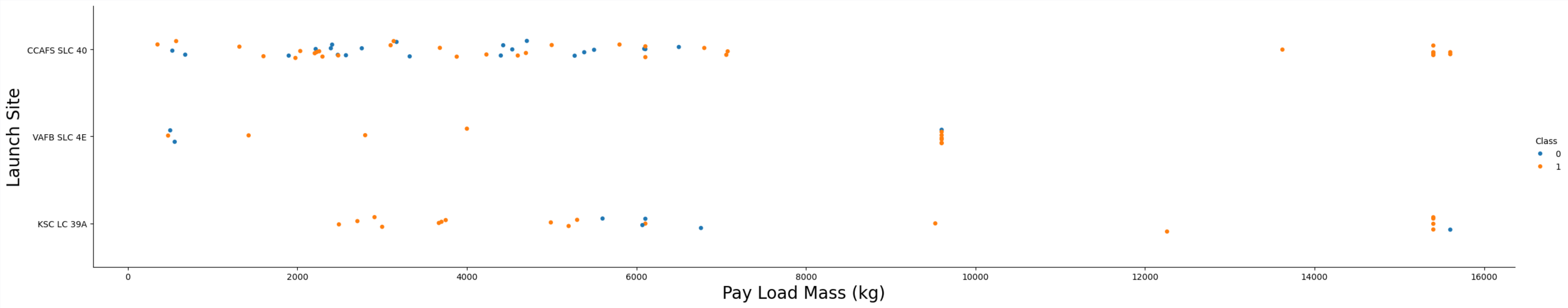
Flight Number vs. Launch Site

- Flight Number vs. Launch Site
- We find that the more flights at a launch site, the greater the success rate at a launch site.



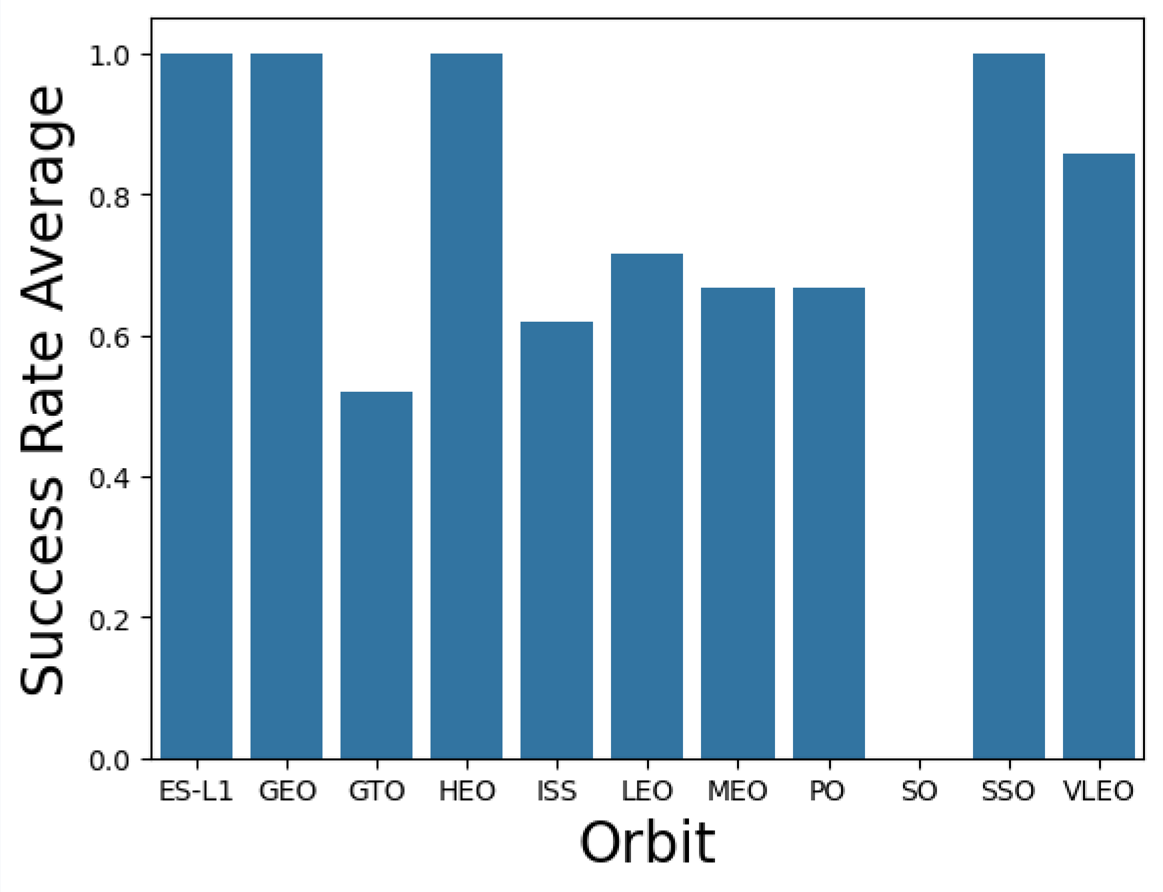
Payload vs. Launch Site

- A big payload mass seems correlated to success



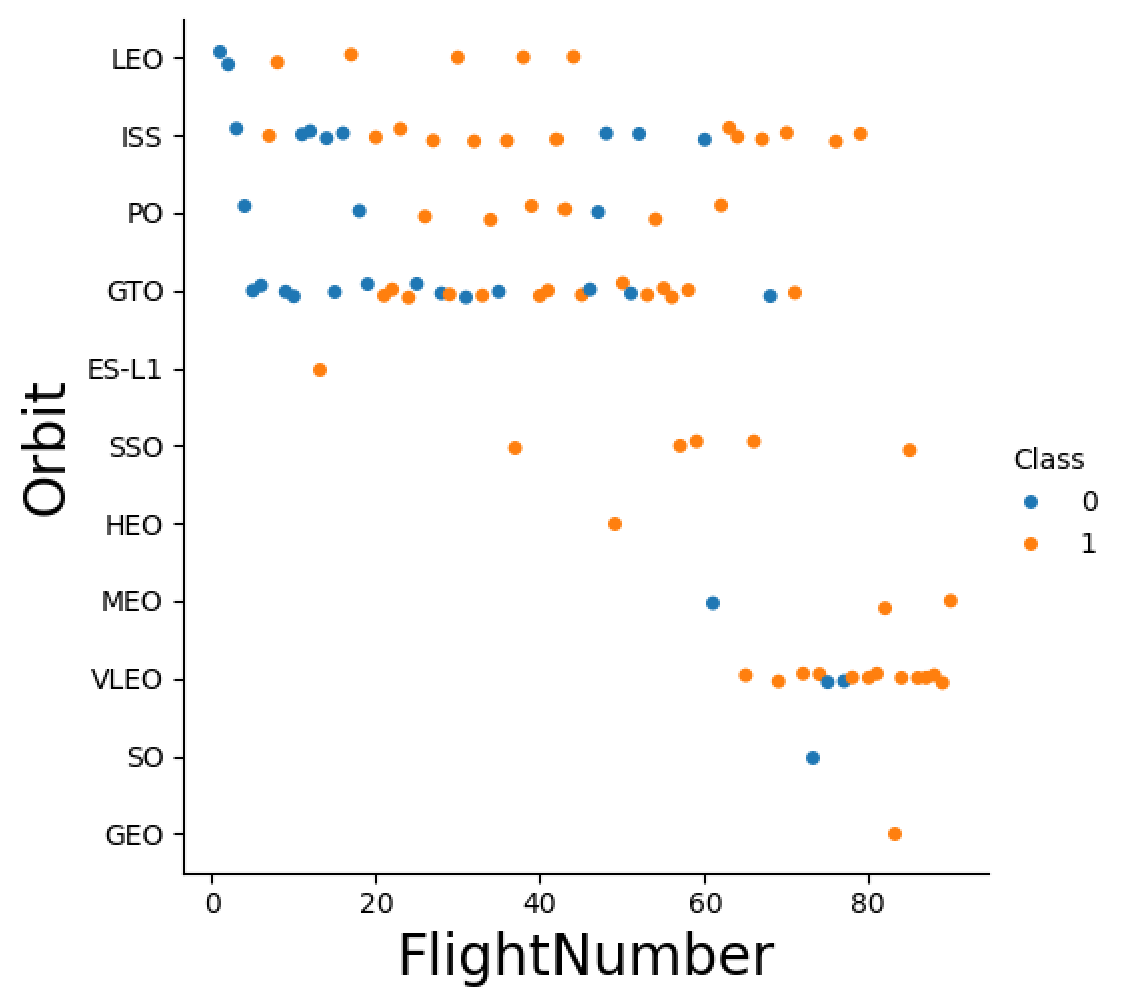
Success Rate vs. Orbit Type

GTO seems to be the orbit with less success rate while SSO, HEO, GEO, ES-L1 perform very well



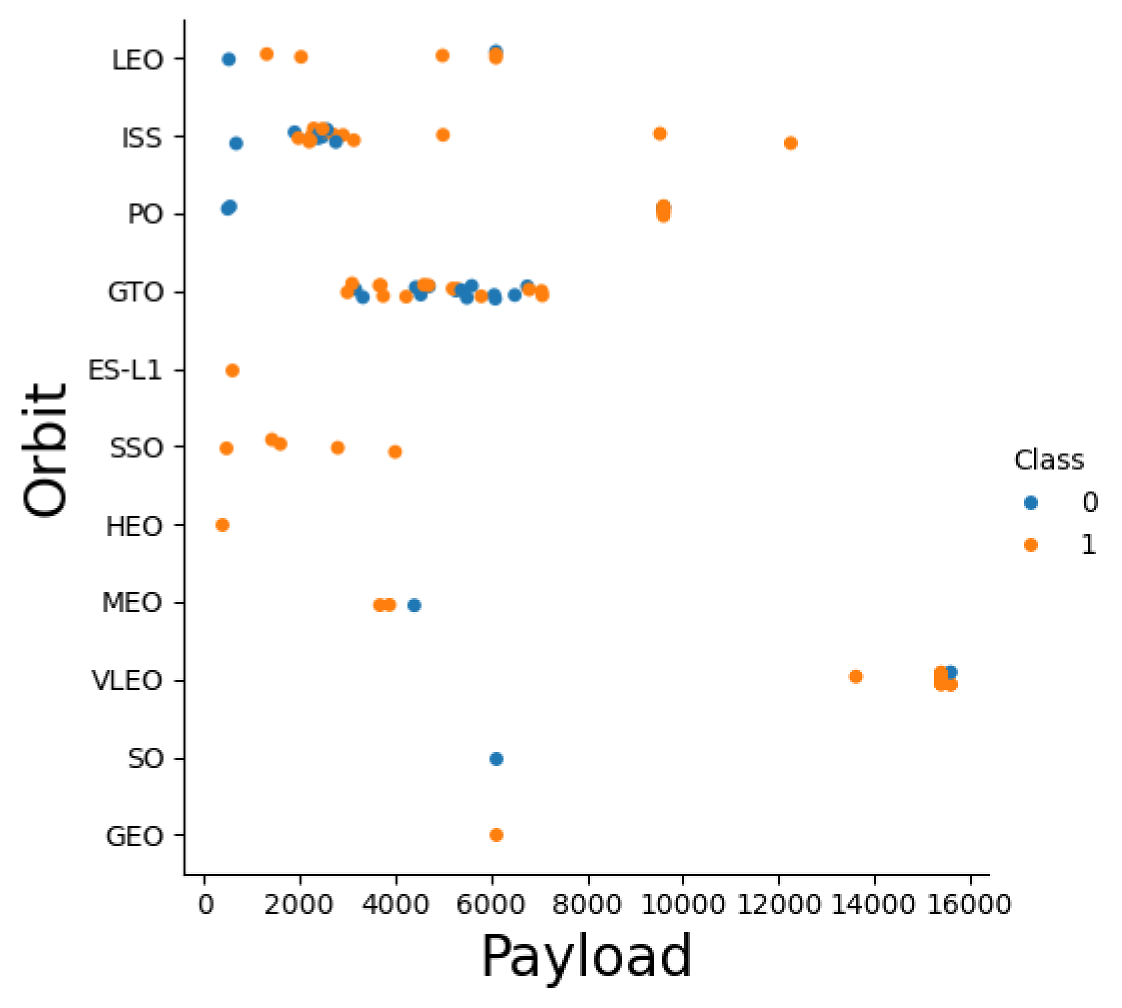
Flight Number vs. Orbit Type

- In some orbits (eg LEO), success is related to the number of flights whereas in others (like GTO), there is no relationship between flight number and the orbit.



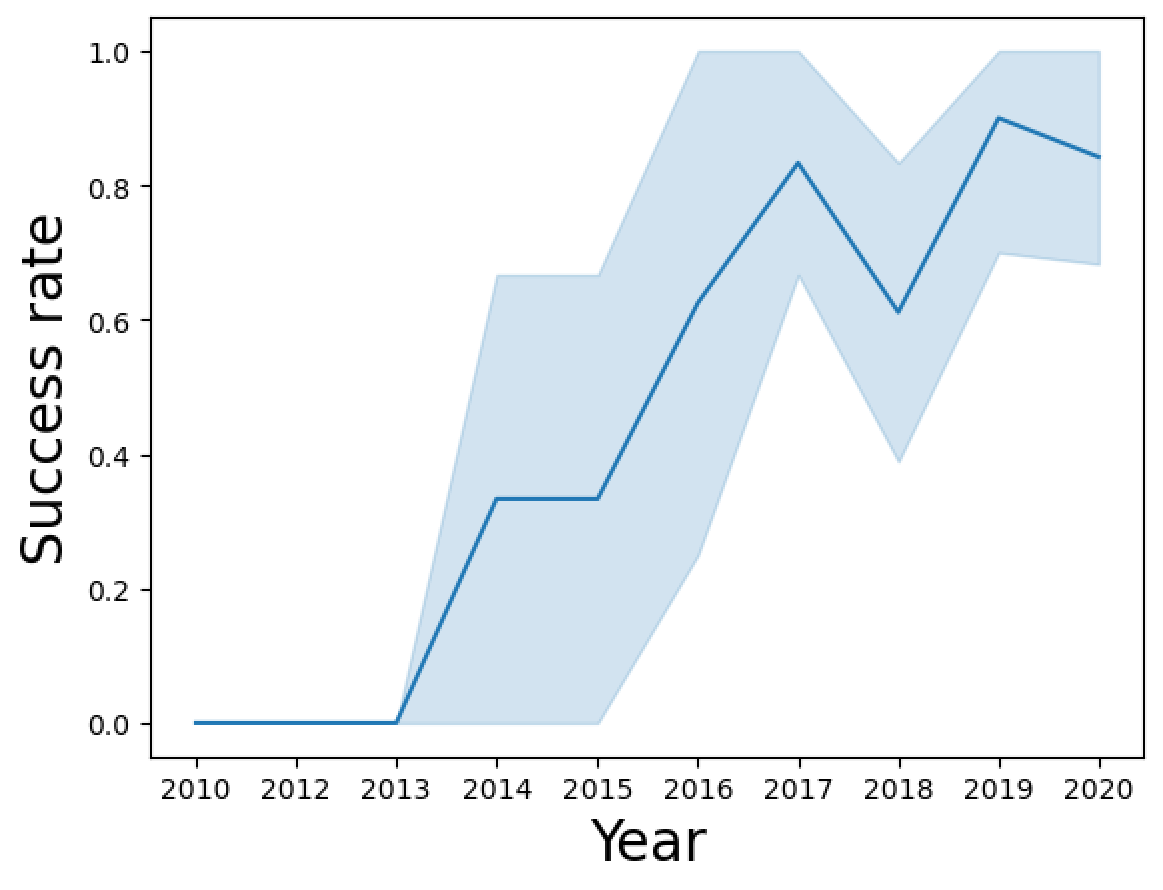
Payload vs. Orbit Type

- With heavy payloads, PO, LEO and ISS orbits have more successful landings.



Launch Success Yearly Trend

- Success rate increases throughout the years



All Launch Site Names

- We use SQL and sqlite in jupyter to query the database

- The launch site names are queried using select

```
%sql select distinct Launch_Site from SPACEXTBL
```

```
 * 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'

- Find 5 records where launch sites begin with `CCA`
- We use the keyword like to define that the launch site should contain CCA (it should actually be CCA% if we want it to begin with CCA but it is the same in this case)
- We use limit to get 5 results

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
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 (para
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No a
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No a
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No a

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Sum keyword used to calculate the sum of some values

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer is "NASA (CRS)"
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Avg keyword used to calculate the average of some values

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version is "F9 v1.1"
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Using min to find the first successful landing

```
%sql select min(Date),Landing_Outcome from SPACEXTBL where Landing_Outcome is "Success (ground pad)"
```

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

min(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

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
- Using keyword between to define a range for a selection

```
%sql select Booster_version from SPACEXTBL where Landing_Outcome is "Success (drone ship)" and  PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Keyword group by is used to group results with the same mission outcome

```
%%sql
SELECT Mission_Outcome, count(Mission_Outcome)
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

Mission_Outcome	count(Mission_Outcome)
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
- Using a subquery to select specifically boosters with maximum payload mass

```
%%sql
select Booster_version, PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ in
(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Here we use substring for the dates as SQLLite does not support monthnames

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

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

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

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(Landing_Outcome)
from SPACEXTBL
where date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(Landing_Outcome) desc
```

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

Landing_Outcome	count(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2

Section 3

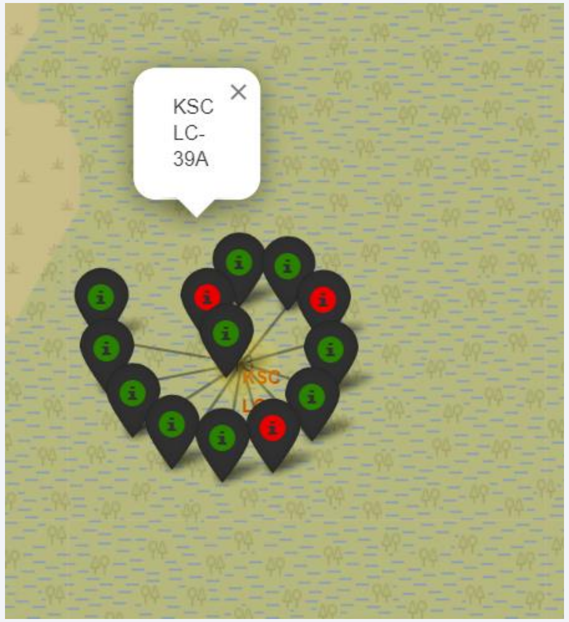
Launch Sites Proximities Analysis

All launch sites

- Launch sites are in the US, in Florida and California

Markers indicating success and failure

- The red markers indicate the failure of a launch and the green one success.
- The number on the launch sites zoomed out represents the total number of launches



Distances between launch sites and coast

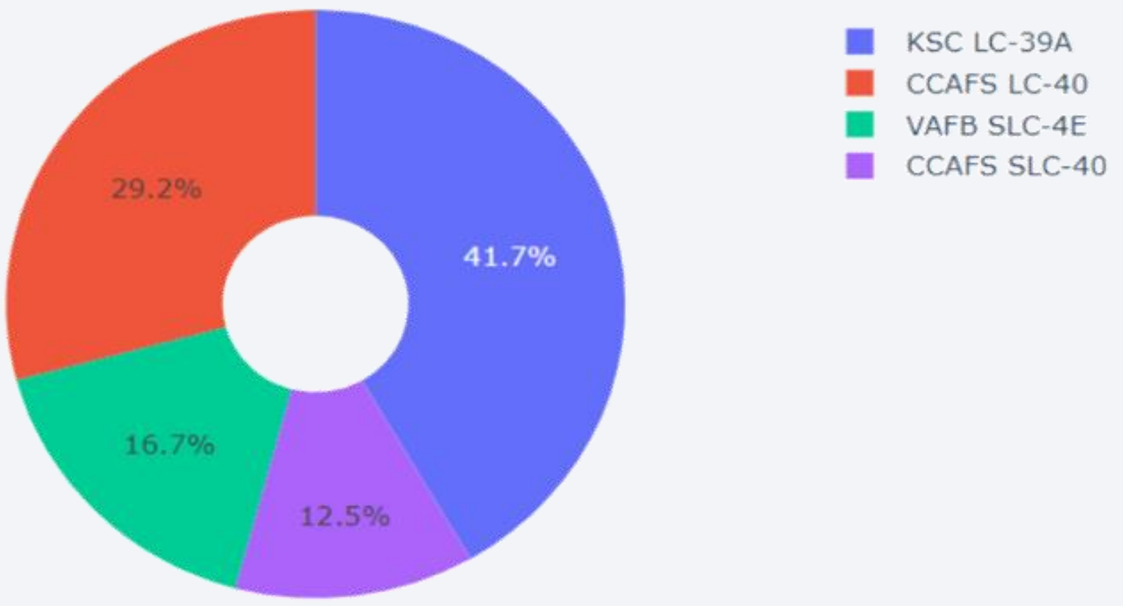
- Folium can also be used to calculate distances between parts of the map. The figure shows the distance between one of the sites and the coast.
- We calculated that launch sites are in close proximity to the coastline and cities, but not to highways and railways

Section 4

Build a Dashboard with Plotly Dash

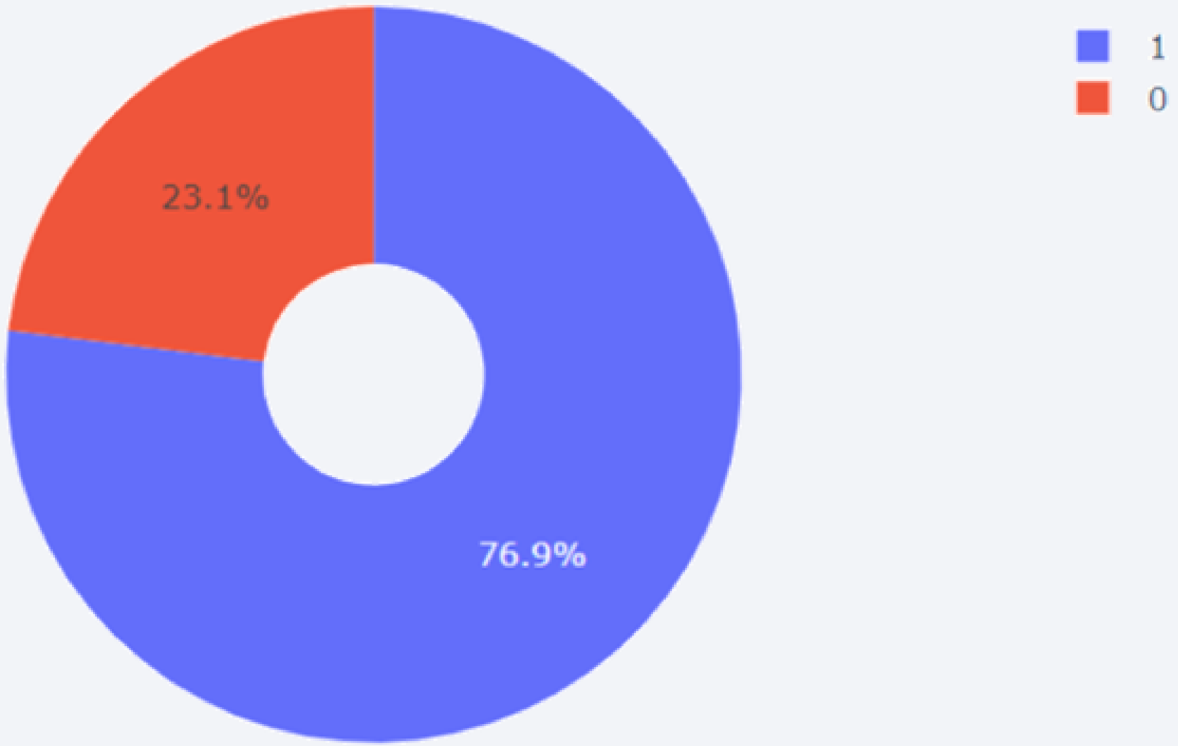
Pie chart of success count for al sites

- The pie chart shows that KSC LC39A has more successful launches than the others



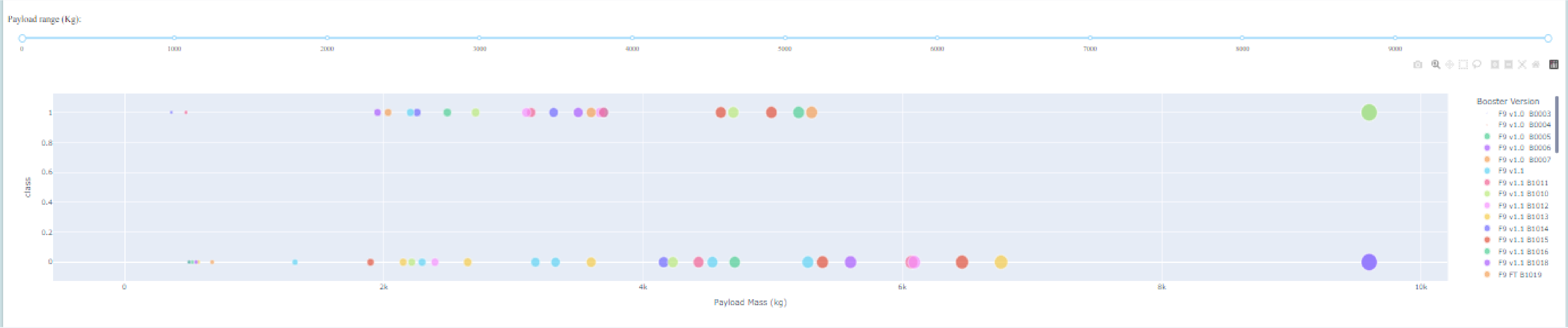
Pie chart for KSC LC39A

- The site KSC LC39A has 76.9 % of success and 23.1 of failure for its launches



Scatter plot of Payload vs Launch Outcome for all sites

The dashboard visualiser allows to use the slider to visualise success rate of different booster versions in different payload ranges

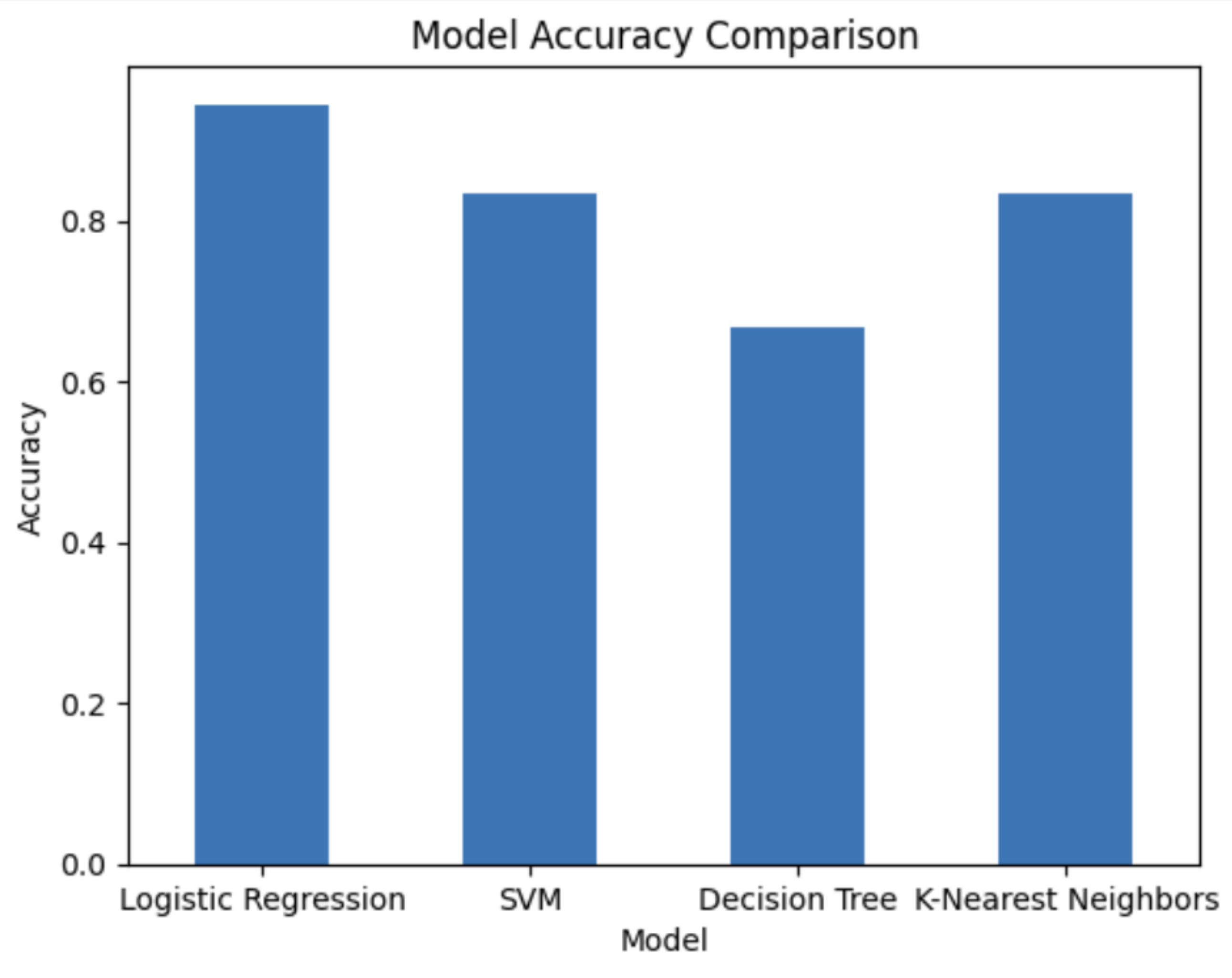


Section 5

Predictive Analysis (Classification)

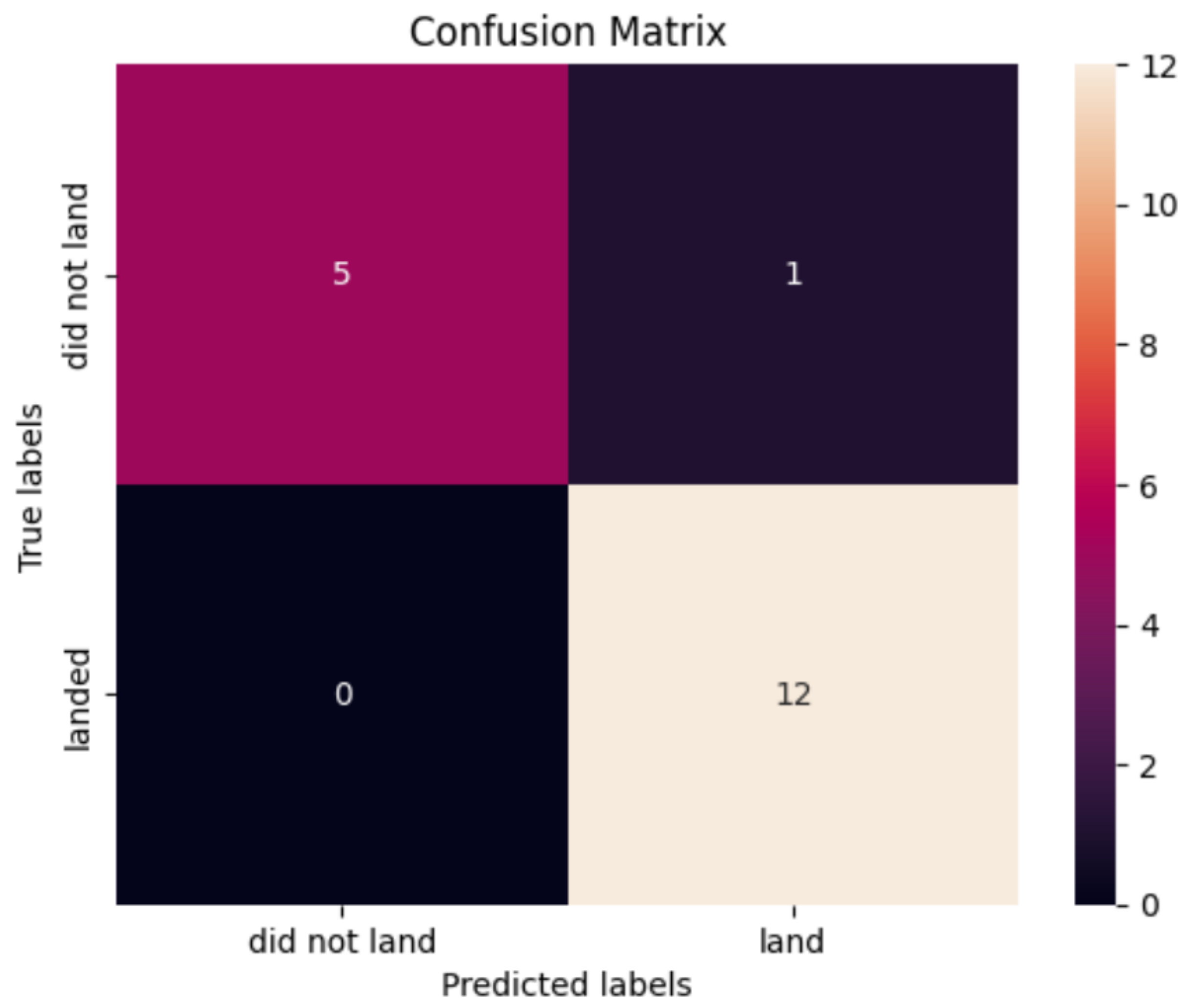
Classification Accuracy

- The logistic regression model used for classification has the best accuracy in this case



Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

Launch Frequency and Success Rate: The larger the flight amount at a launch site, the greater the success rate. Launch success rate has been increasing steadily from 2013 to 2020.

Optimal Launch Sites: KSC LC-39A stands out with the highest success rate among all launch sites. Notably, it boasts a 100% success rate for launches carrying payloads less than 5,500 kg.

Geographical Advantage: Most launch sites are located near the equator and close to the coast, providing a natural boost due to the Earth's rotational speed and reducing additional fuel and booster requirements.

Orbital Success: Orbits such as ES-L1, GEO, HEO, SSO, and VLEO demonstrate the highest success rates.

Payload Mass Correlation: Across all launch sites, a higher payload mass (kg) is associated with a higher success rate.

Model Performance: The logistic regression classifier is identified as the best machine learning algorithm for this task, slightly outperforming other models on the test set.

Thank you!