IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jens Kugelmann
<Date>

Outline

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

Executive Summary

- The starting point of this project is data collection, gathering as much relevant data as possible. Once the raw data is collected, it needs to be improved through data wrangling. Then exploring the processed data starts by applying SQL queries, statistical analysis and data visualization. Finally, the data is split into groups defined by categorical variables or factors.
- As a results we will have insights about what has in impact on the result of a rocket launch

Introduction

- Sending aircrafts in space is very expensive. With this project I want to find out, what has an impact on the result of a mission. That information can than be used to save money, by having more successful missions
- I want to answer the question:
  - How can rocket launches be more successful?

Section 1

Methodology

Methodology

Executive Summary

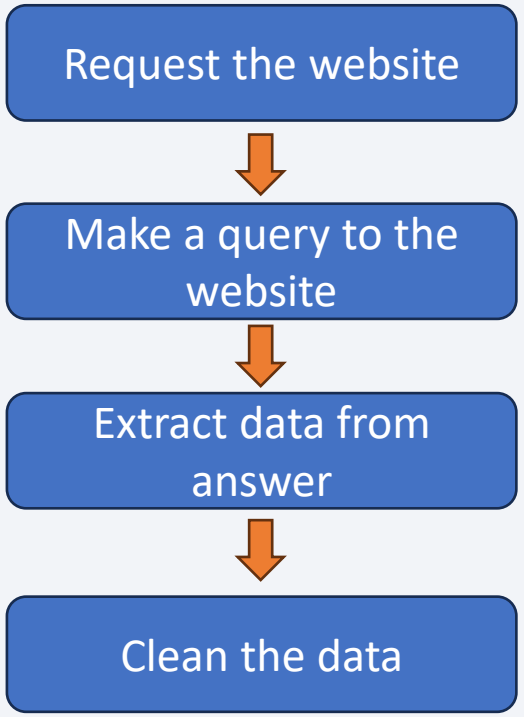
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The necessary data was collected in two ways:
  - First with the API of SpaceX
  - Second by web scraping
- After that the data got processed

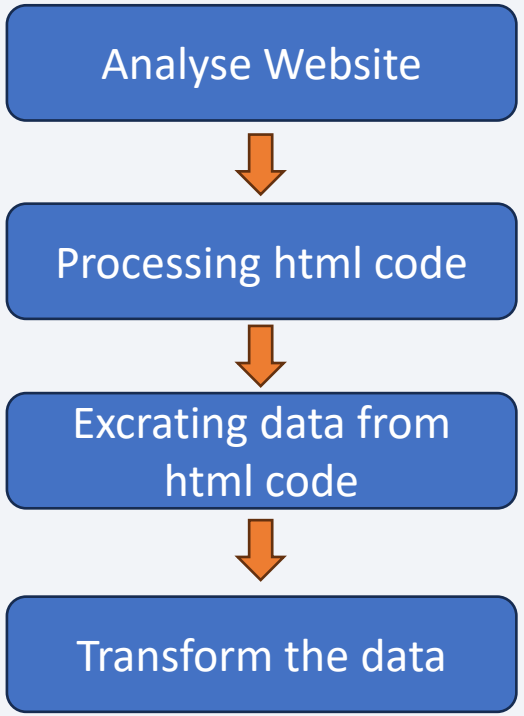
Data Collection – SpaceX API

- The API helps to get the data from SpaceX
- This interfaces answers a request
- The data from the answer needs to be extracted and cleaned
- https://github.com/Strunxkopp/Applied-Data-Science-Capstone/blob/855b7d814d87963b70d55266952eb30fba3c46c1/1.1-spacex-data-collection-API.ipynb



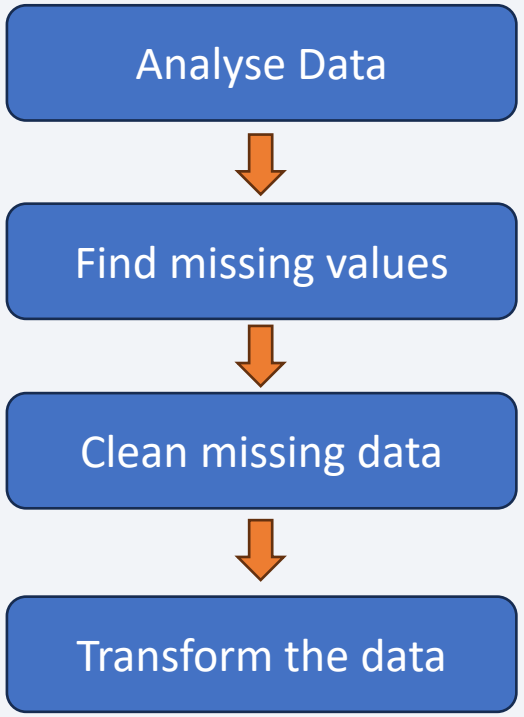
Data Collection - Scraping

- To get the data from the website, an http request was sent
- The answer is a html code with different tables
- The interesting tables need to identified
- The interesting tables get extracted
- The extracted data gets transformed into a DataFrame with useable features
- https://github.com/Strunxkopp/Applied-Data-Science-Capstone/blob/855b7d814d87963b70d55266952eb30fba3c46c1/1.2-spacex-webscraping.ipynb



Data Wrangling

- The collected data needs to be analysed
- Missing values needs to be identified
- The missing values needs to be replaced, or the rows dropped
- Machine readable features need to created from categorical features by creating dummies for example
- https://github.com/Strunxkopp/Applied-Data-Science-Capstone/blob/855b7d814d87963b70d55266952eb30fba3c46c1/1.3-spacex-DataWrangling.ipynb



EDA with Data Visualization

- To visualize the data, scatter plots are plotted to find out if there is a correlation between two selected features.
- The following combination of features are plotted:
  - FlightNumber vs. PayloadMass
  - FlightNumber vs LaunchSite
  - Payload vs Launch Site
  - FlightNumber vs Orbit type
  - Payload vs Orbit
- Bar chart for relationship between success rate and orbit type is plotted
- line chart with success rate over the years is plotted
- https://github.com/Strunxkopp/Applied-Data-Science-Capstone/blob/855b7d814d87963b70d55266952eb30fba3c46c1/2.2-EDA-DataVisualisation.ipynb

EDA with SQL

- I have performed SQL queries that list:
  - the names of the launch sites
  - 5 records where launch sites begin with `CCA`
  - the total payload carried by boosters from NASA
  - the average payload mass carried by booster version F9 v1.1
  - the date of the first successful landing outcome on ground pad
  - the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - the total number of successful and failure mission outcomes
  - names of the booster which have carried the maximum payload mass
  - records from the column 'landing_outcome' that failed landing with a drone ship in year 2015, together with their booster versions, and launch site names
  - Rank the count of landing outcomes
- https://github.com/Strunxkopp/Applied-Data-Science-Capstone/blob/855b7d814d87963b70d55266952eb30fba3c46c1/2.1-EDA-SQL.ipynb

Build an Interactive Map with Folium

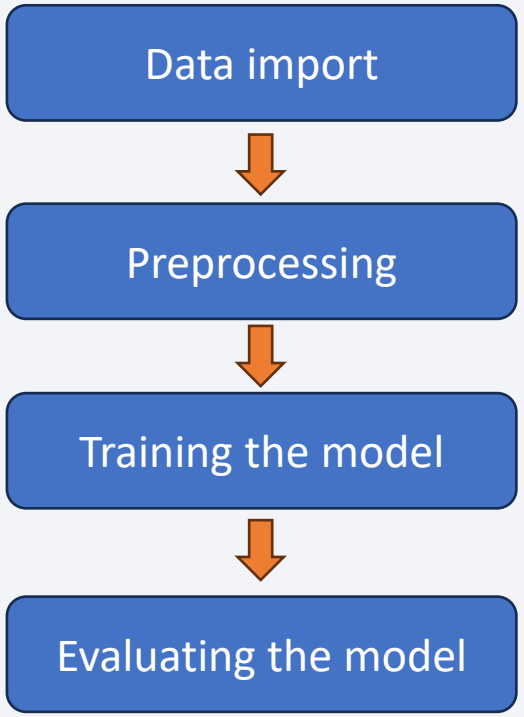
- The Folium library was used to create a map with markers of the launch sites.
- To visualize the success of the launch site, a cluster was created with a marker for each launch and their color corresponding to their success or failure
- Lines from the launch site to things near by were drawn and the distance calculated to see if there are certain things near very successful launch sites, or others further away
- https://github.com/Strunxkopp/Applied-Data-Science-Capstone/blob/855b7d814d87963b70d55266952eb30fba3c46c1/3.1-Visualisation-Folium.ipynb

Build a Dashboard with Plotly Dash

- I have created a Dashboard, where you have a Dropdown menu to select every single launch site, or all launch sites.
- A Pie chart will be created, that displays the success and failure rate of each launch site
- The Dashboard also displays a scatterplot, that shows the successful and failed launches. You can filter the launches with a range sliders by selecting a range of carried payload.
- The plots show the influence, that the payload and the launch site have on the success of the mission
- https://github.com/Strunxkopp/Applied-Data-Science-Capstone/blob/855b7d814d87963b70d55266952eb30fba3c46c1/3.2-Dashboard.ipynb

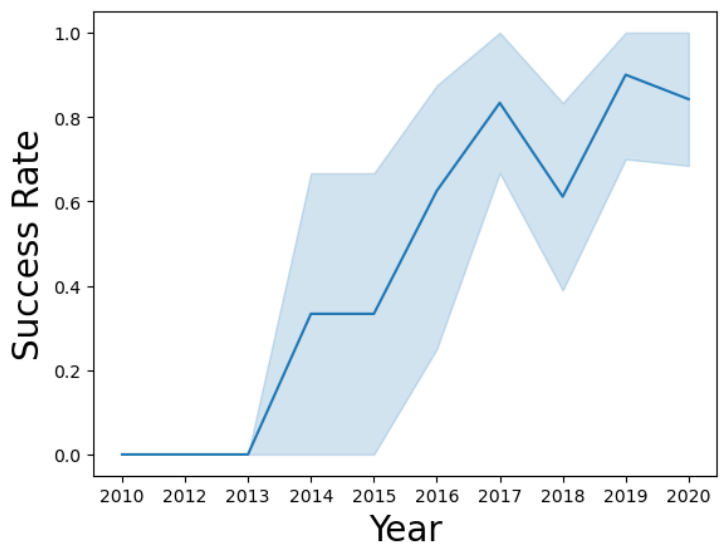
Predictive Analysis (Classification)

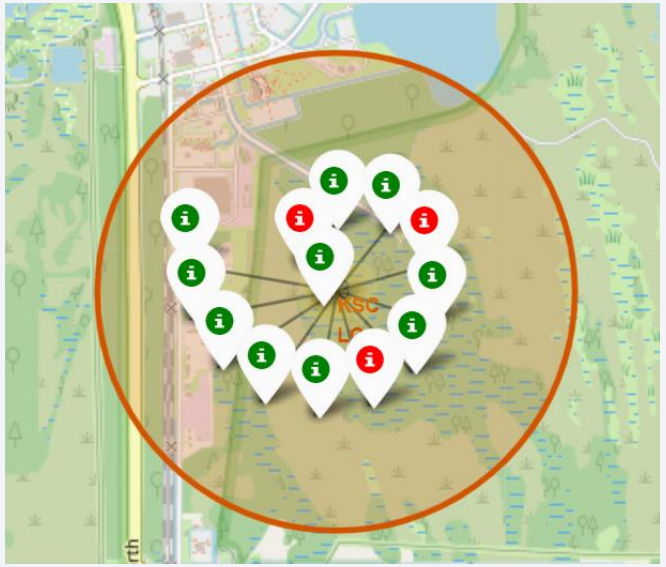
- First the data was read into a DataFrame
- Then the data was preprocessed, to have features to train the model with and a target variable that wants to be predicted. The data gets split into train and test data
- The models get trained with grid search to find the best parameters from a range of parameters for each model
- The model gets evaluated by calculating the accuracy and plotting a confusion matrix
- https://github.com/Strunxkopp/Applied-Data-Science-Capstone/blob/855b7d814d87963b70d55266952eb30fba3c46c1/4.1-SpaceX_Machine%20Learning%20Prediction.ipynb

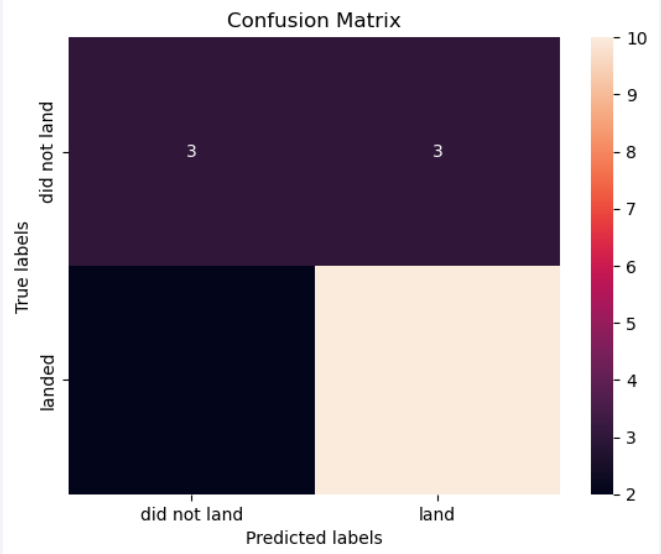


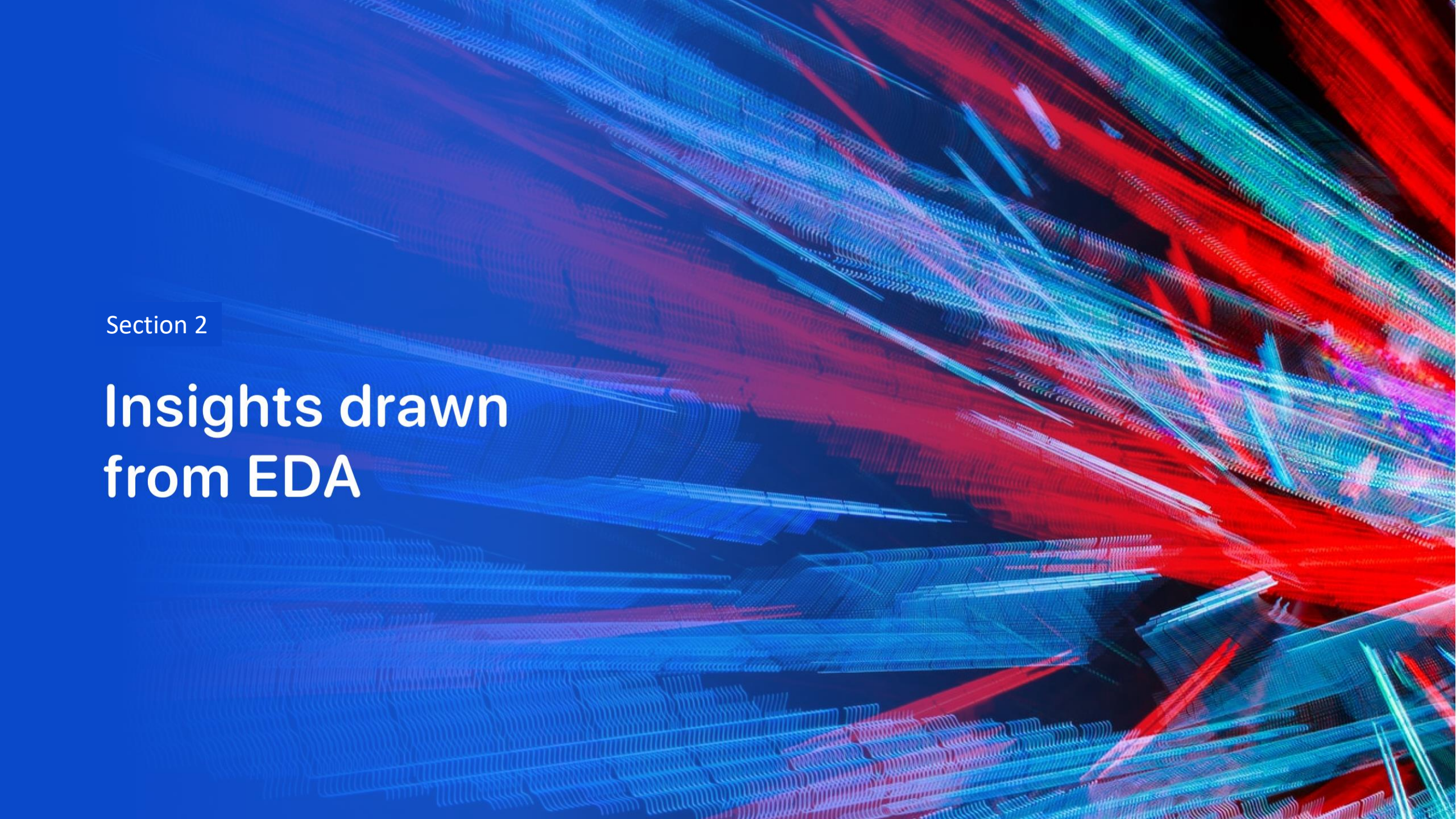
Results

- From the EDA the success rate increased over the years
- Interactive analytics shows that some launch sites are more successful then others
- With the gathered information a model can help predict, if a landing will be successful



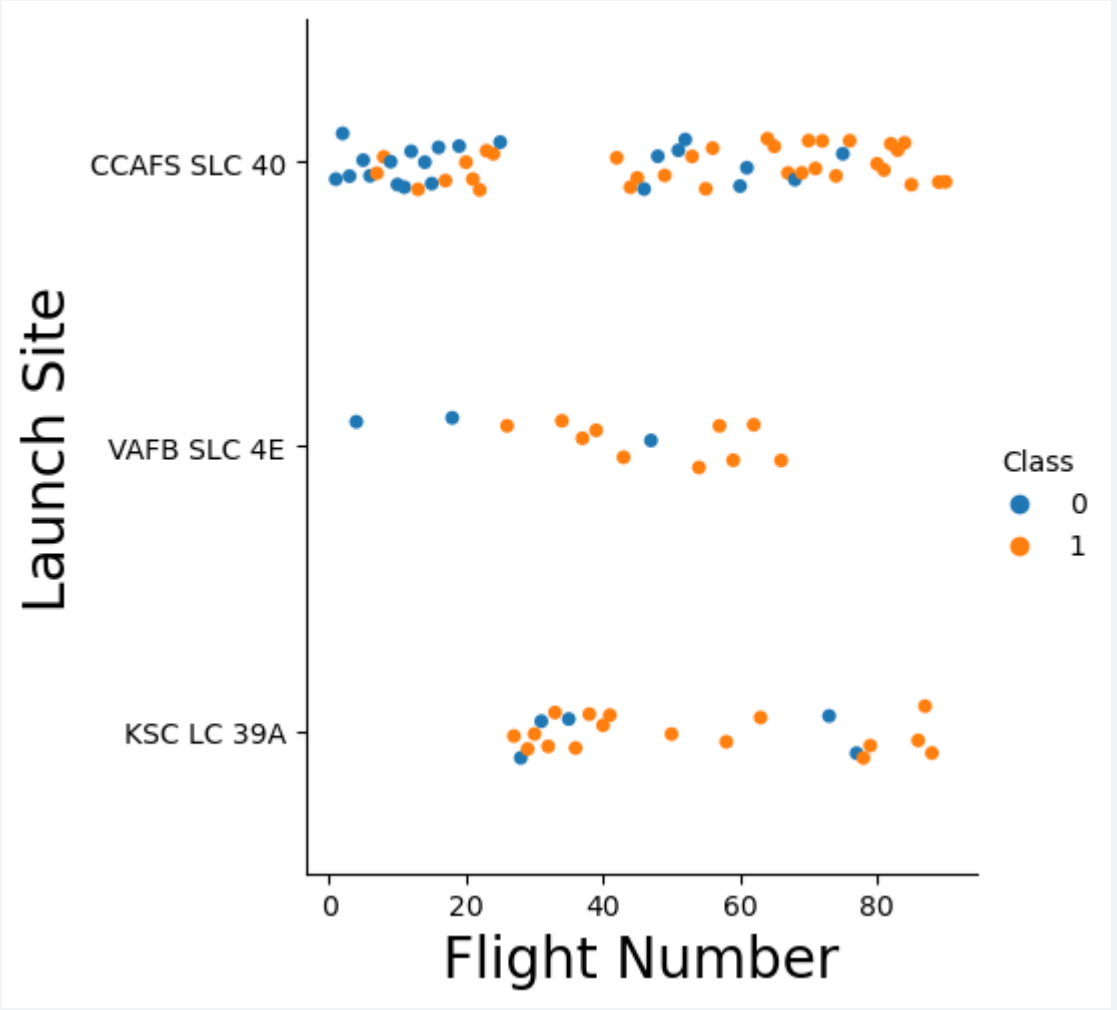





Section 2
Insights drawn from EDA

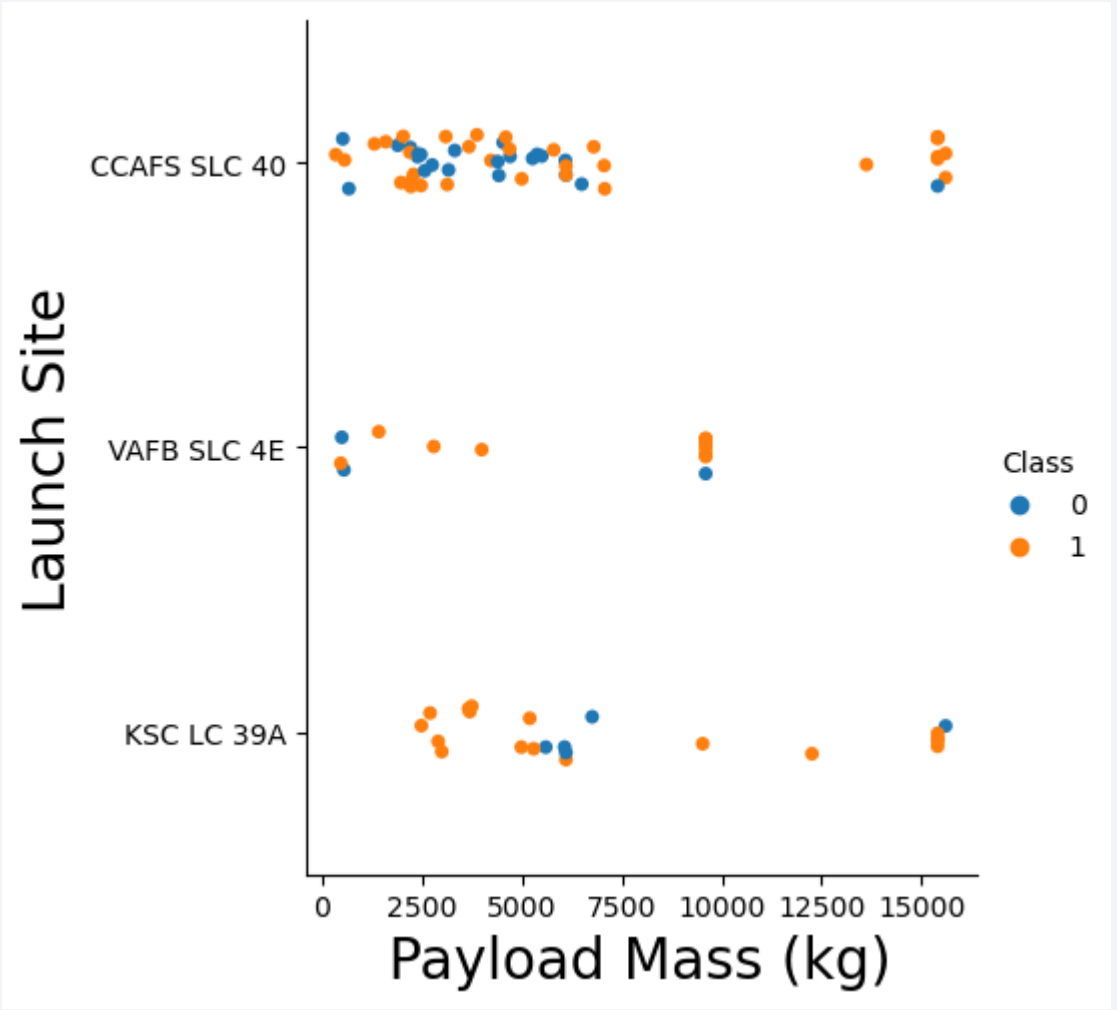
Flight Number vs. Launch Site

- The scatter plot shows the launch site over the flight number
- The first launches on KSC LC 39A were after several launches on CCAFS SLC 40
- With the launches on CCAFS SLC 40 there were no more launches on KSC LC 39A for a while



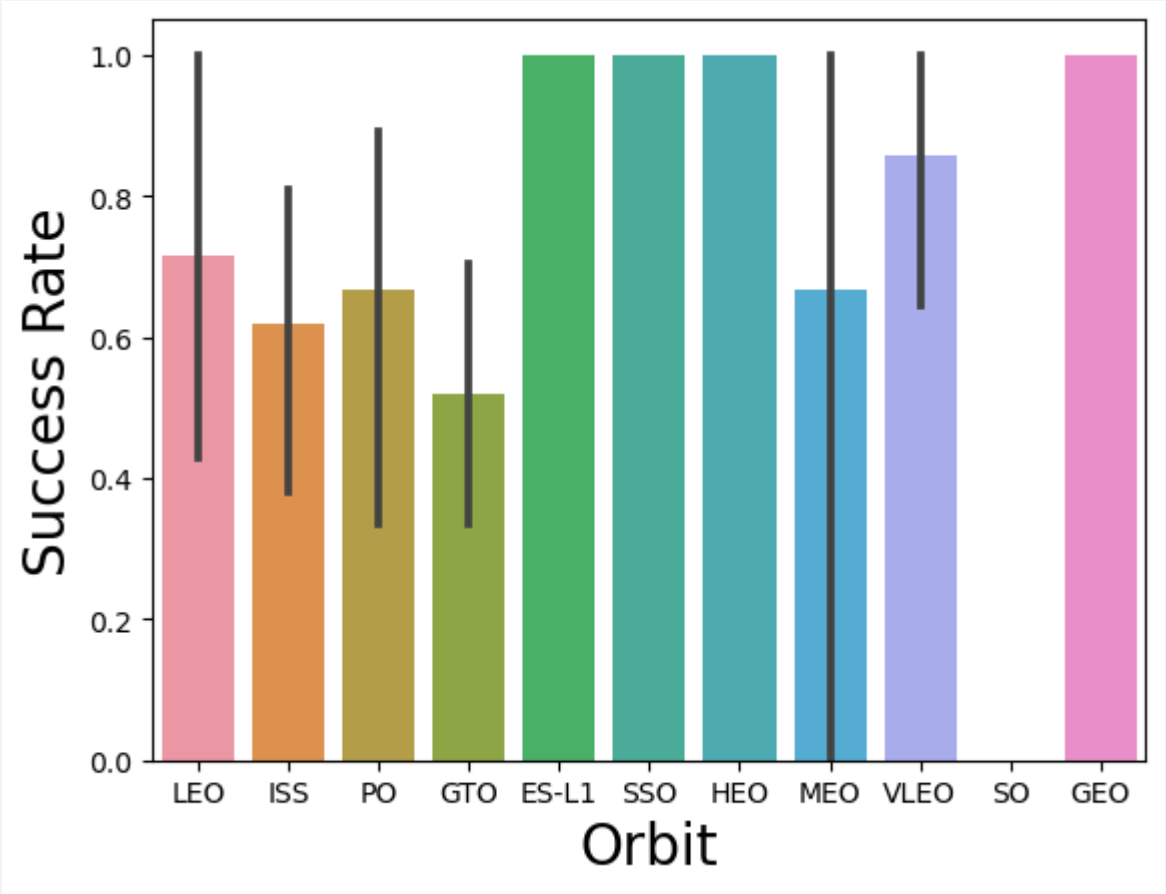
Payload vs. Launch Site

- From this scatter plot it seems like some launch sites are preferred for certain payload masses



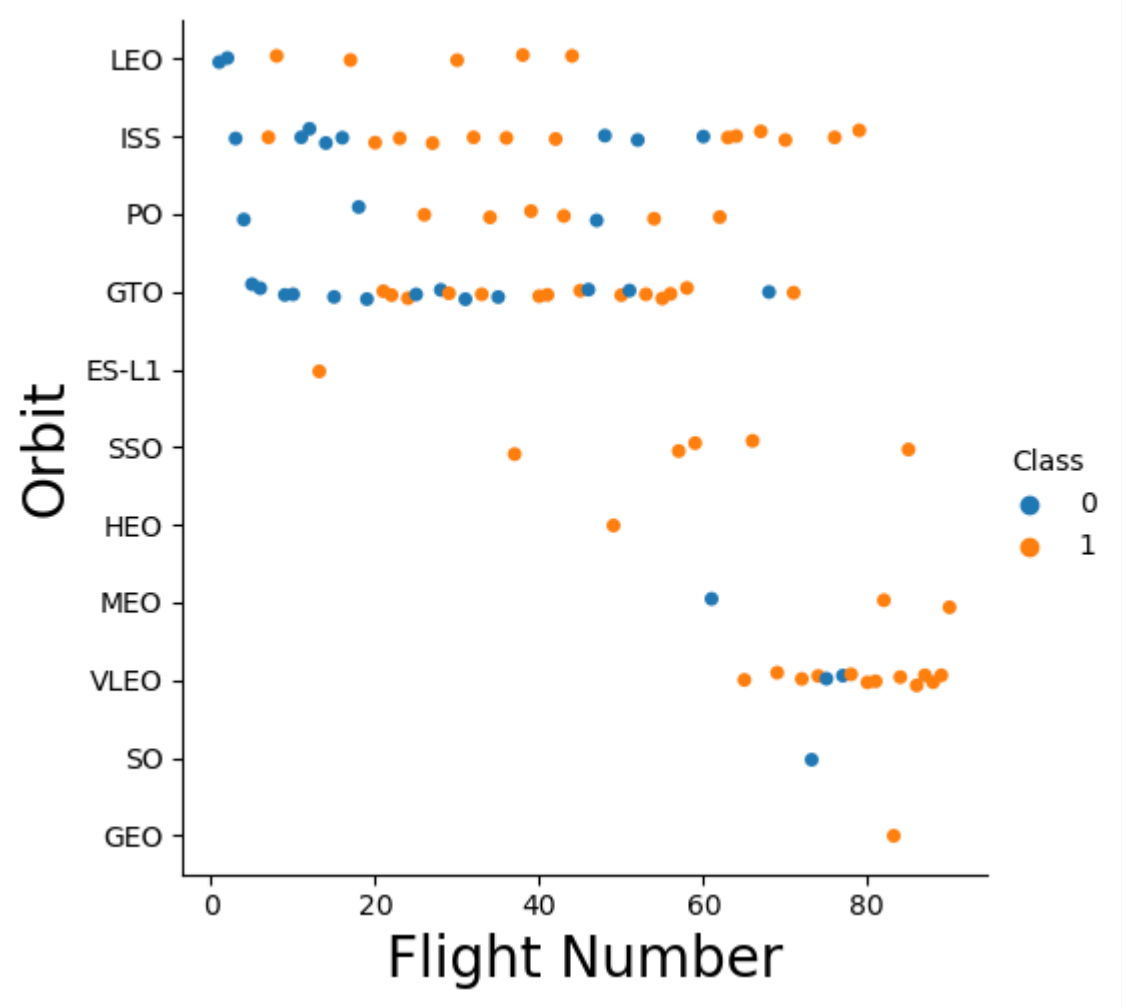
Success Rate vs. Orbit Type

- The success rate varies a lot
- These orbits have a success rate of 100%:
  - ES-L1
  - SSO
  - HEO
  - GEO



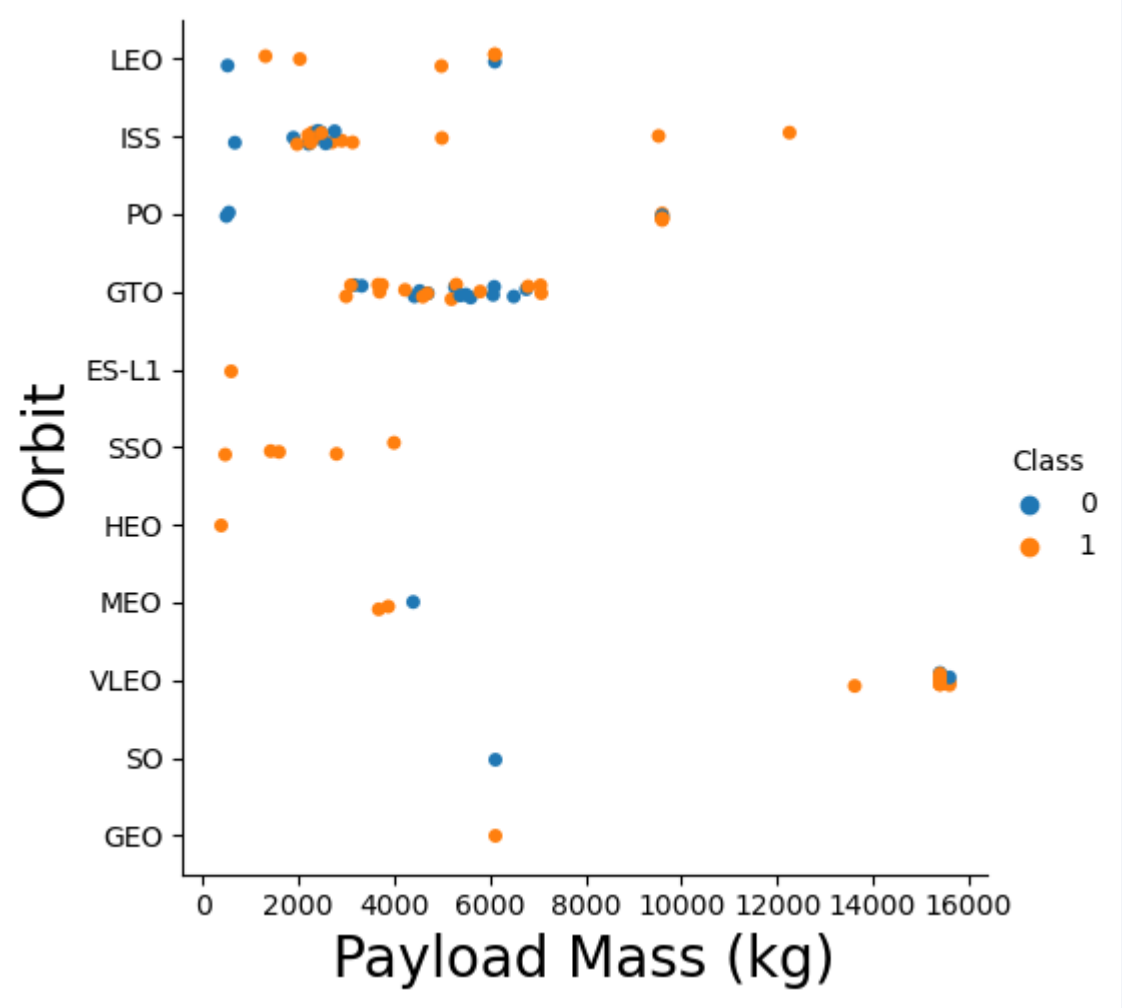
Flight Number vs. Orbit Type

- The scatter plot shows that the first launches went to lower orbits
- Later there came launches to higher orbits



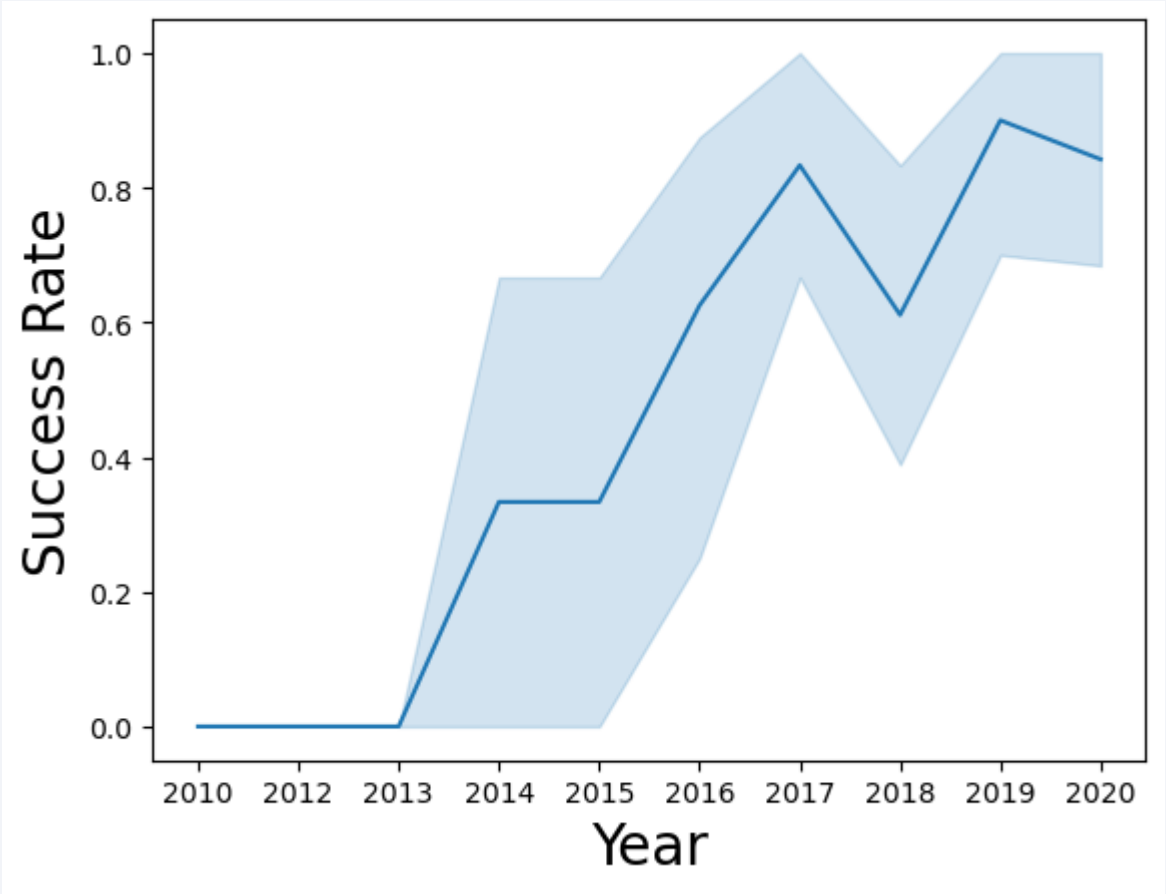
Payload vs. Orbit Type

- The scatter plot shows the payload mass over the orbit
- It seems like that certain orbits have a specific range for the payload mass



Launch Success Yearly Trend

- The Line plot shows the succes rate over the years
- It indicates, that the success increased over the years



All Launch Site Names

- The query outputs the the names of the launch sites
- Only the coloumn ‘Launch_Site’ is shown and every duplicate records are removed

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

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

Launch Site Names Begin with 'CCA'

- The query outputs 5 records where launch sites begin with `CCA`
- The coloumn ‘Launch_Site’ gets filtered for records that starts with ‘CCA’ and the number of displayed records are limited to 5

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

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

Total Payload Mass

- The query outputs the total payload carried by boosters from NASA
- In the query all records are summed up, where the Cusomer is NASA
- NASA's boosters carried 107010kg payload in total

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload from SPACEXTBL where Customer like '%NASA%';
```

total_payload
107010

Average Payload Mass by F9 v1.1

- The query outputs the average payload mass carried by booster version F9 v1.1
- The average of the coloumn ‘PAYLOAD_MASS__KG_’ carried by booster version F9 v1.1 is 2928.4kg

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1';
```

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The query outputs the date of the first successful landing outcome on ground pad
- The first successful landing on ground pad was on 2015-12-22

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query outputs the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The filtered records have to fulfill the condition ‘Success (drone ship)’ and needs to within the givin payload mass range
- There are 4 Booster Version that fulfill these criteria

```
%sql select distinct Booster_Version from SPACEXTBL where Landing_Outcome like 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

- The query outputs the total number of successful and failure mission outcomes
- 100 mission were successful and only 1 mission failed

```
%sql select count(Mission_Outcome) as Mission_Success from SPACEXTBL where Mission_Outcome like 'Success%';
```

Mission_Success
100

```
%sql select count(Mission_Outcome) as Mission_Failure from SPACEXTBL where Mission_Outcome not like 'Success%';
```

Mission_Failure
1

Boosters Carried Maximum Payload

- To the right you see a list with names of the booster which have carried the maximum payload mass
- The subquery is nested in the where condition to check if the record has the maximum payload mass

```
%sql select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ == (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

- The query lists the records from the column ‘landing_outcome’ that failed landing with a drone ship in year 2015, together with their booster versions, and launch site names
- The result are the two records in the table on the right

```
%%sql
SELECT
    CASE
        WHEN substr(Date, 6, 2) = '01' THEN 'January'
        WHEN substr(Date, 6, 2) = '02' THEN 'February'
        WHEN substr(Date, 6, 2) = '03' THEN 'March'
        WHEN substr(Date, 6, 2) = '04' THEN 'April'
        WHEN substr(Date, 6, 2) = '05' THEN 'May'
        WHEN substr(Date, 6, 2) = '06' THEN 'June'
        WHEN substr(Date, 6, 2) = '07' THEN 'July'
        WHEN substr(Date, 6, 2) = '08' THEN 'August'
        WHEN substr(Date, 6, 2) = '09' THEN 'September'
        WHEN substr(Date, 6, 2) = '10' THEN 'October'
        WHEN substr(Date, 6, 2) = '11' THEN 'November'
        WHEN substr(Date, 6, 2) = '12' THEN 'December'
    END AS month,
    landing_outcome,
    booster_version,
    launch_site
FROM SPACEXTBL
WHERE substr(Date, 0, 5) = '2015'
  AND landing_outcome  LIKE 'Failure (drone ship)';
```

month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	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
- Present your query result with a short explanation here

```
%%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;
```

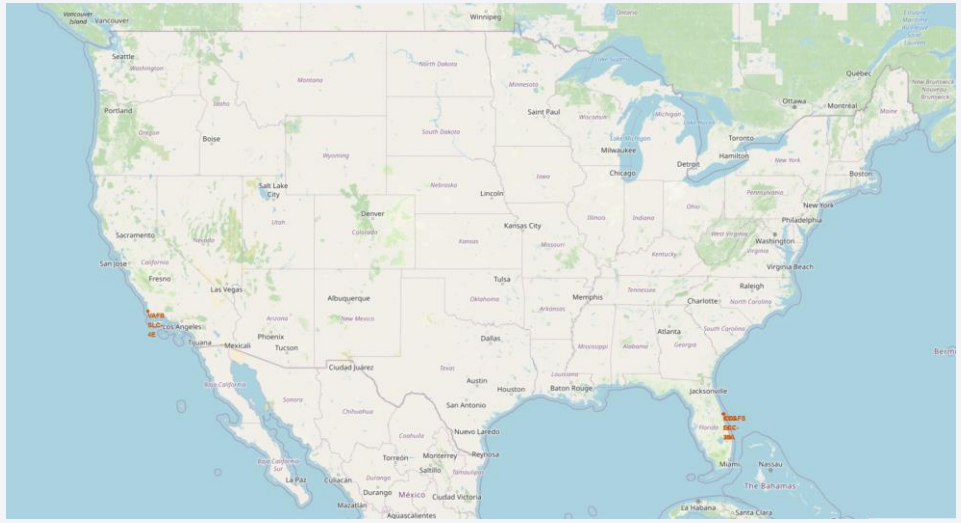
Landing_Outcome	count(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

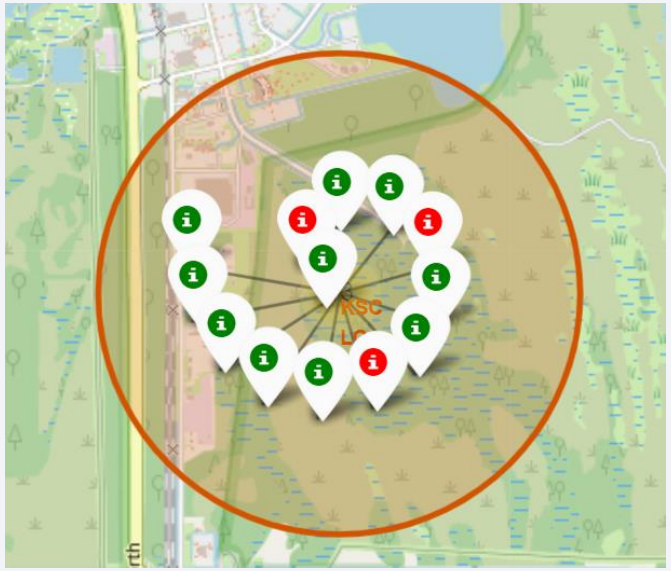
World map with launch sites

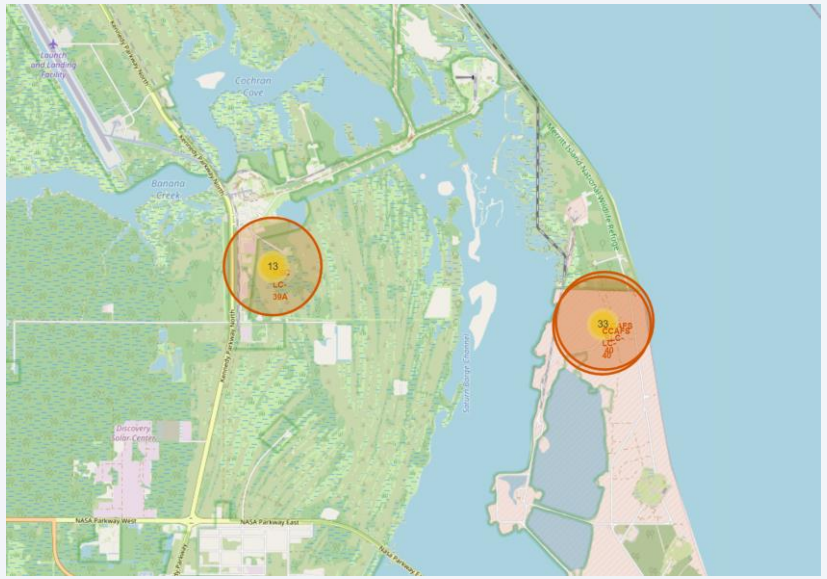
- The launch sites are one in California and two in Florida, USA
- All three launch sites are close to the coast
- All launch sites are far away from the Equator

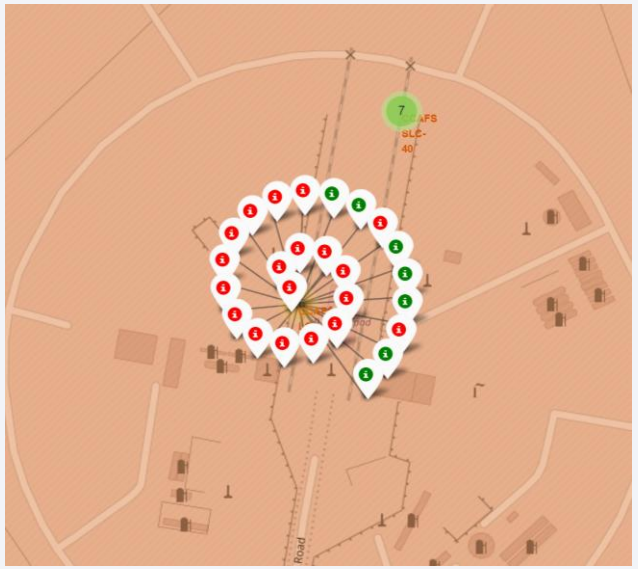


Map with success/failed launches for each site

- The screenshot in the middle shows the launch sites in FLorida
- The launch site KSC LC 39A has a higher (left) success rate than CCAFS LC-40 (right)

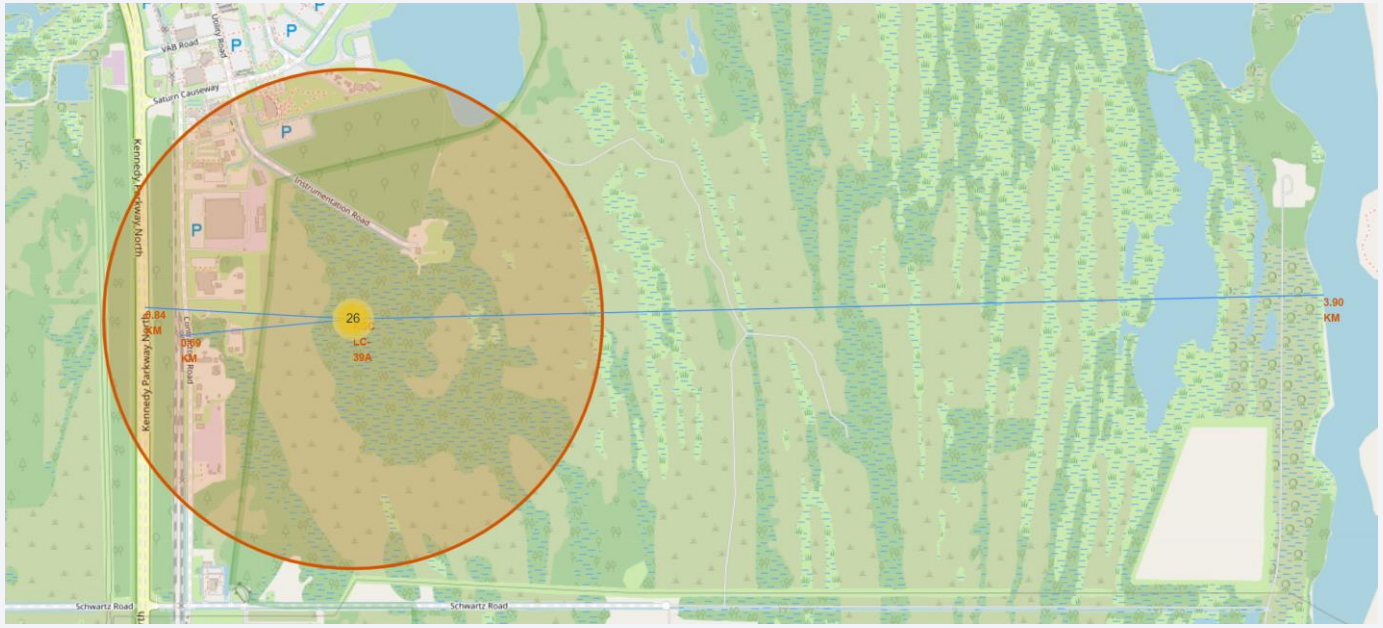


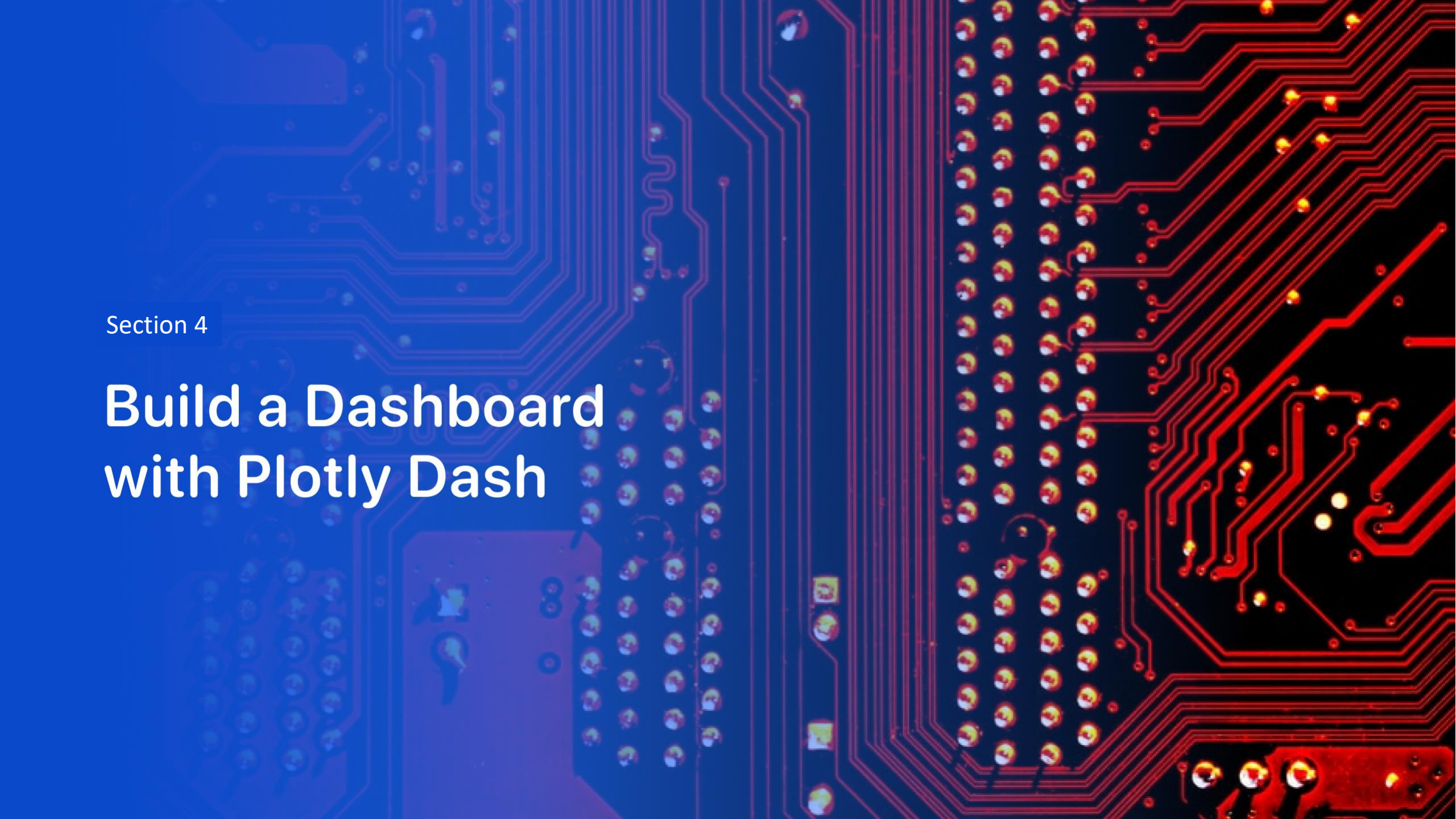




Map with calculated proximities

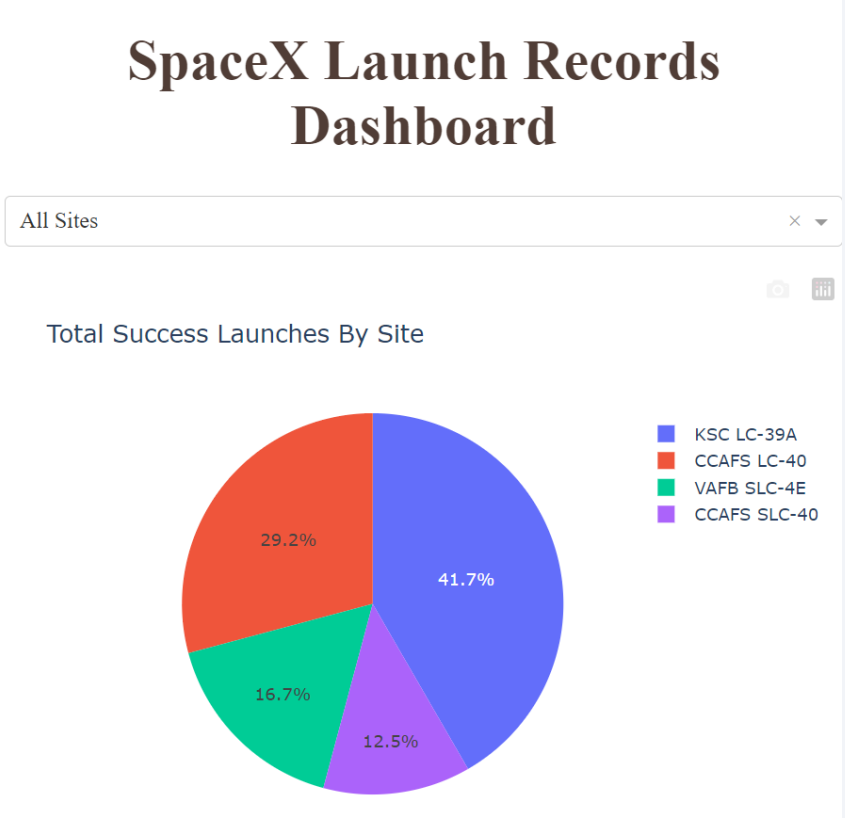
- The launch sites are close to railways, highways and coastlines
- Cities are further away from the launch sites
- The launch site keeps distance to cities to avoid accidents if there is a failure




Section 4
Build a Dashboard with Plotly Dash

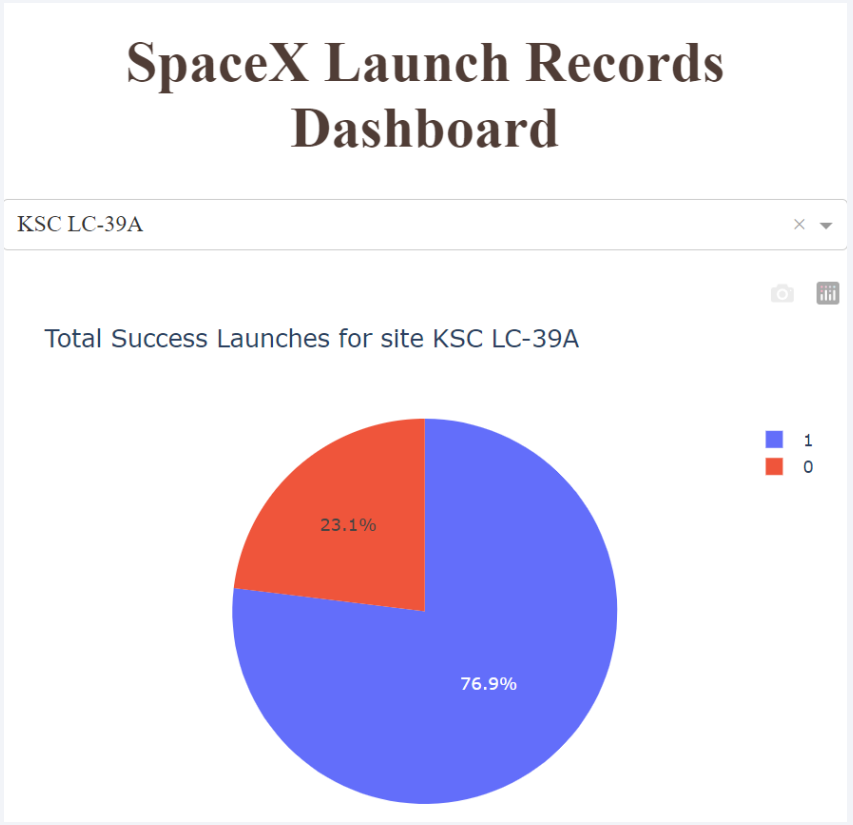
Pie chart with total success launches by site

- The dashboard makes it easy to select the most successful launchsite
- Here KSC LC-39A has with 41.7% the most success



Pie chart with total success launches for KSC LC-39A

- Now the launch site KSC LC-39A is selected
- The pie chart shows the percentage of success with here 76.9%



Scatterplots with different payloads and their outcome

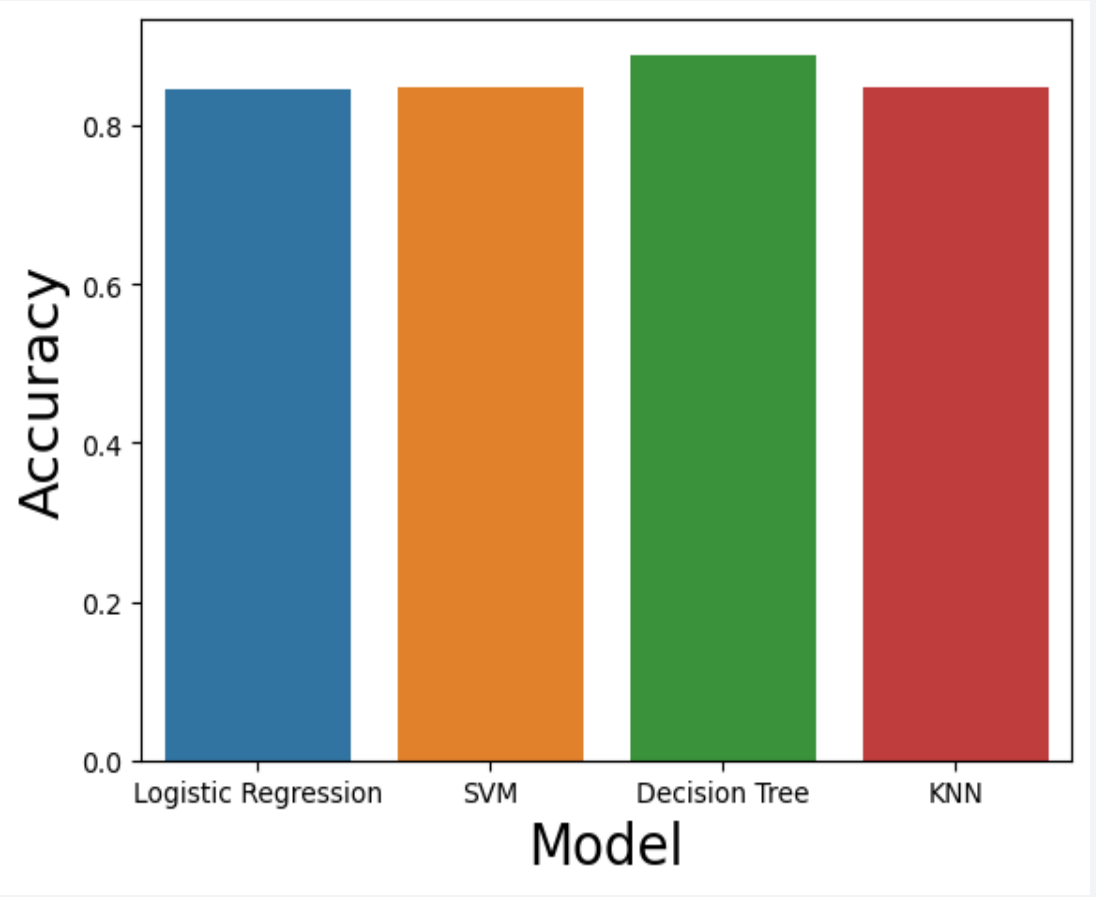
- On this dashboard, the payload can be adjusted, to get a scatterplot with the outcome with the selected range of payload
- It seems like the missions with higher payloads have more success

Section 5

Predictive Analysis (Classification)

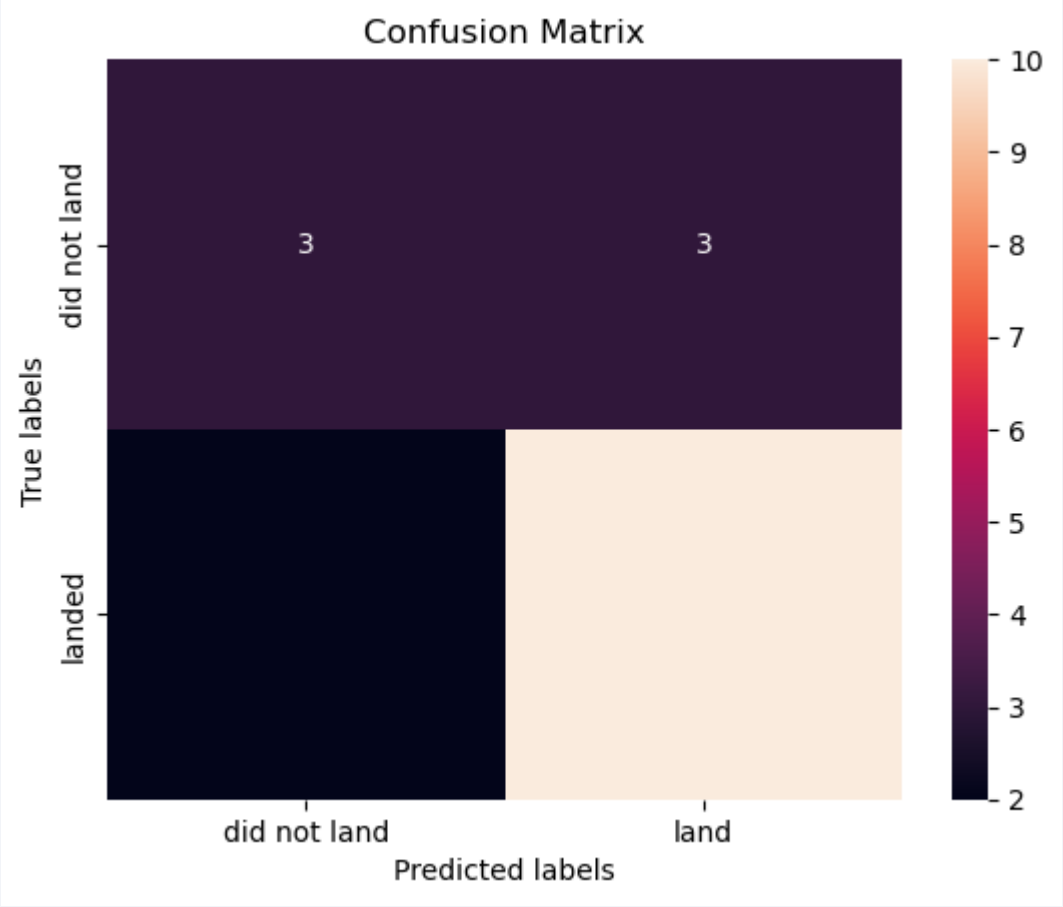
Classification Accuracy

- The bar chart shows the Accuracy for each model that was trained with the collected data.
- The highest accuracy has the Decision Tree Model.



Confusion Matrix

- To the right is the confusion matrix for the decision tree model.
- The model is not very accurate on boosters that did not land, because there is some amount of false negative predictions like for false positive.



Conclusions

- Point 1: The success rate increased over the year.
- Point 2: The success depends on the Launch site.
- Point 3: A classification model can help predicting, if the landing will be successful

Thank you!