IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tinh Nhat Pham
1/1/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Build an Interactive Map with Folium
  - Build a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

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
- Problems you want to find answers
  - Discover new insights by analyzing Falcon 9 rocket launch data in the first stage.
  - Predict whether or not the first stage will successfully land.

Section 1

Methodology

Methodology

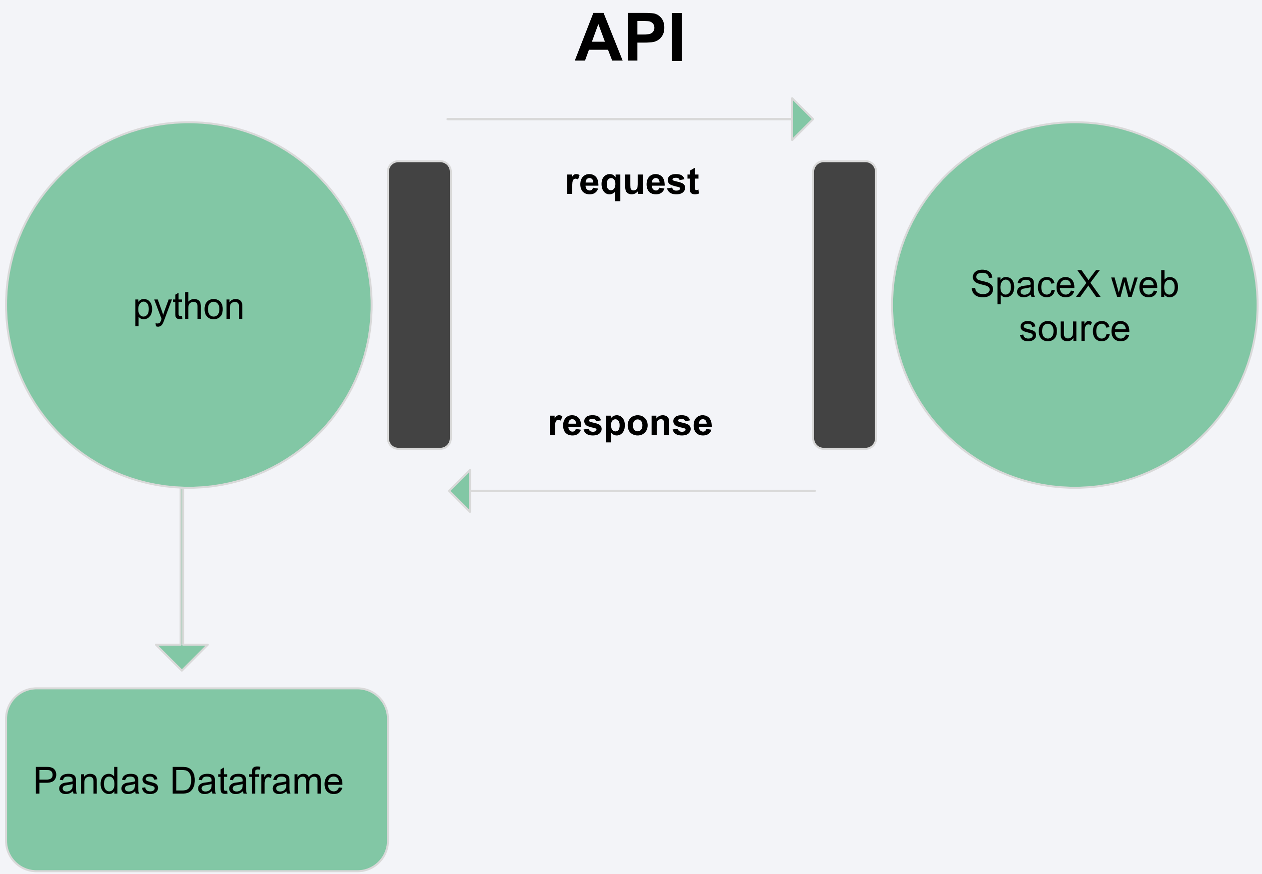
Executive Summary

- 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 – SpaceX API

Key phrases:

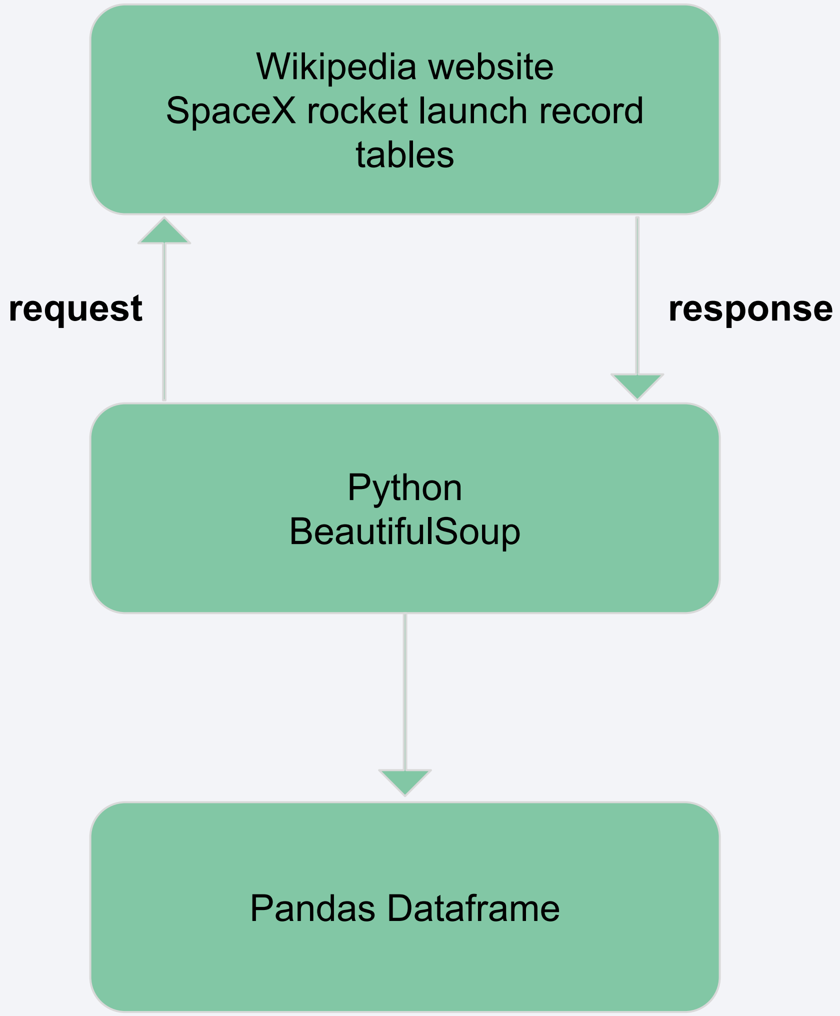
- Send request to get rocket launch data from SpaceX website
- Receive response from the SpaceX website
- Parse response JSON and store in to Pandas Dataframe
- GitHub URL: [Click here]



Data Collection - Scraping

Key phrases:

- Send request to get rocket launch record tables from SpaceX Wikipedia
- Receive response from the SpaceX Wikipedia
- Parse response content using *BeautifulSoup* API
- Store data into Pandas Dataframe
- GitHub URL: [Click here]

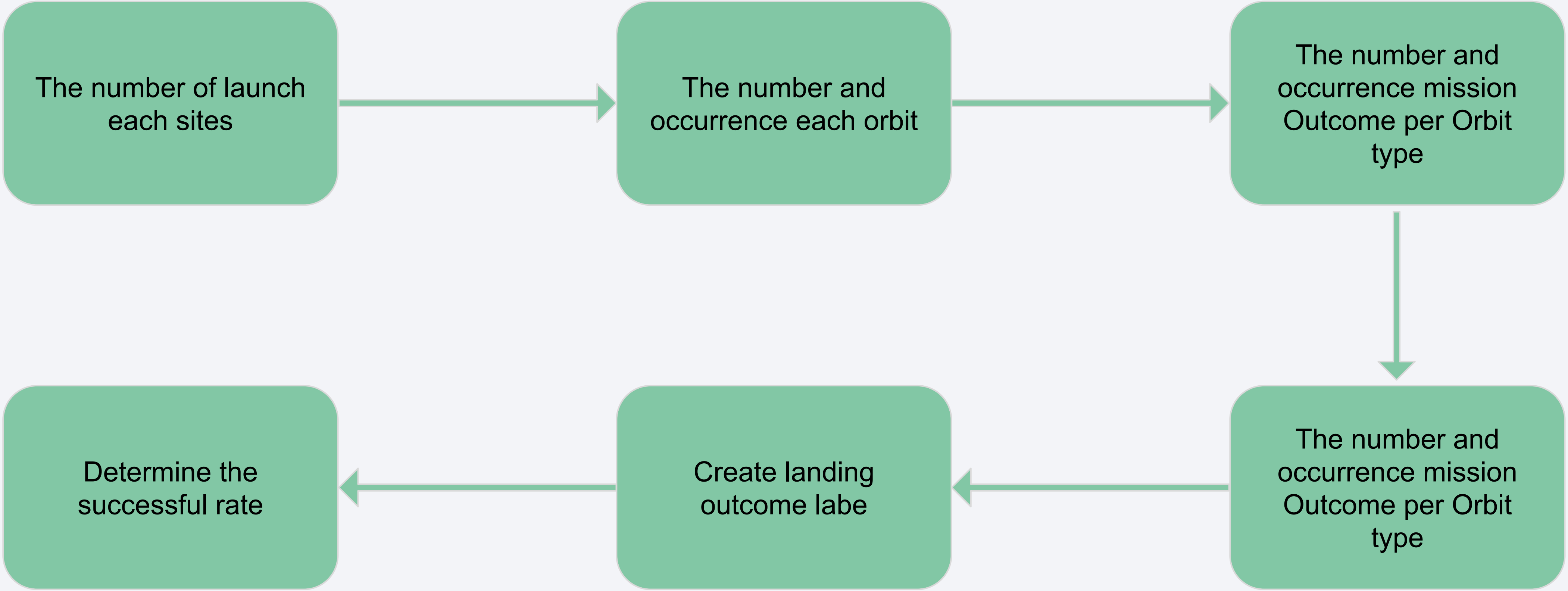


Data Wrangling

Key phrases:

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Determine the successful rate
- GitHub URL: Click here

Data Wrangling (cont.)



EDA with Data Visualization

- Scatter plot: FlightNumber vs PayMassLoad
  - Check the relationship between flight number vs rocket payload mass based on landing outcome
- Scatter plot: FlightNumber vs Launch Site
  - Check the relationship between flight number vs launch sites based on landing outcome
- Scatter plot: PayloadMass vs Launch Site
  - Check the relationship between flight number vs launch sites based on landing outcome

EDA with Data Visualization (cont.)

- Bar plot: Successful rate vs Orbit
  - Visualize the relationship between successful rate of each orbit type
- Scatter plot: FlightNumber vs Orbit type
  - Visualize the relationship between Flight number and Orbit type
- Scatter plot: PayloadMass vs Orbit type
  - Visualize the relationship between Payload mass and Orbit type
- Line plot: Successful rate over the years (2010 - 2020)
  - Visualize the successful rate from 2010 - 2020, which kept increasing since 2013 till 2020
- GitHub URL: Click here

EDA with SQL

- Query the names of the unique launch sites in the space mission
- Query the names of the unique launch sites in the space mission
- Query 5 records where launch sites begin with the string 'CCA'
- Query the total payload mass carried by boosters launched by NASA (CRS)
- Query average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.

EDA with SQL (cont.)

- 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for 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

- GitHub URL: Click here

Build an Interactive Map with Folium

- Create circle object to show the location of NASA center and marker to show the name NASA JSC on the folium map
- Create circle objects to show the locations of Launch Sites and marker objects to show the names of the Launch Sites on the folium map
- Create marker objects to show the success/failed launches for each site on the map, and add to the Cluster marker object
- Calculate the distances between the CCAFS SLC-40 to the nearest coastline, highway, rainway, and city. Then draw the lines of the distances.
- GitHub URL: [Click here]

Build a Dashboard with Plotly Dash

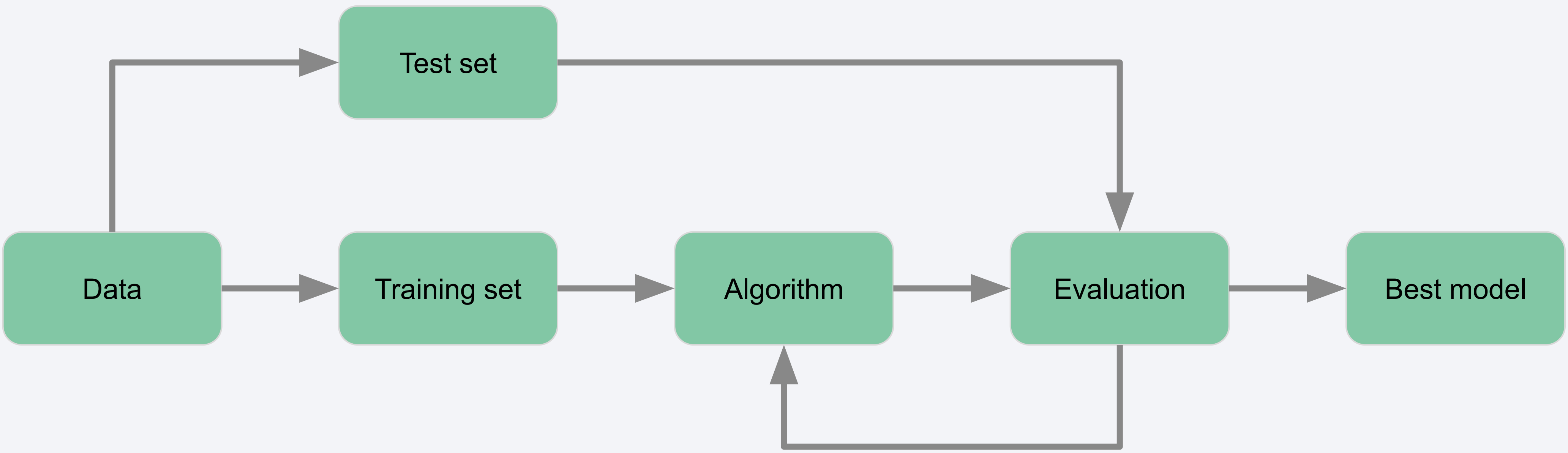
- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add drop down list of Launch sites
- Add a pie chart to show the success/failed launch rate when choosing the site from drop down list
- Add a slider to choose the minimum and maximum payload mass
- Add a scatter plot to show the success/failed launches when choosing the site from drop down list and the min/max payload mass from slider
- GitHub URL: [Click here]

Predictive Analysis (Classification)

- Load the data and store in Pandas Dataframe
- Create features (X) and target (Y)
- Standardize the features X
- Create Train set and Test set by using X and Y
- Using GridSearchCV to find the best parameters for each classifier:
    - Logistic Regression
    - Support Vector Machine
    - Decision Tree
    - K Nearest Neighbors
- Calculate performance score each classifier by predict Test set.
- Evaluate the best performing classification model.

- GitHub URL: Click here

Predictive Analysis (Classification) (cont.)



Model Development Process

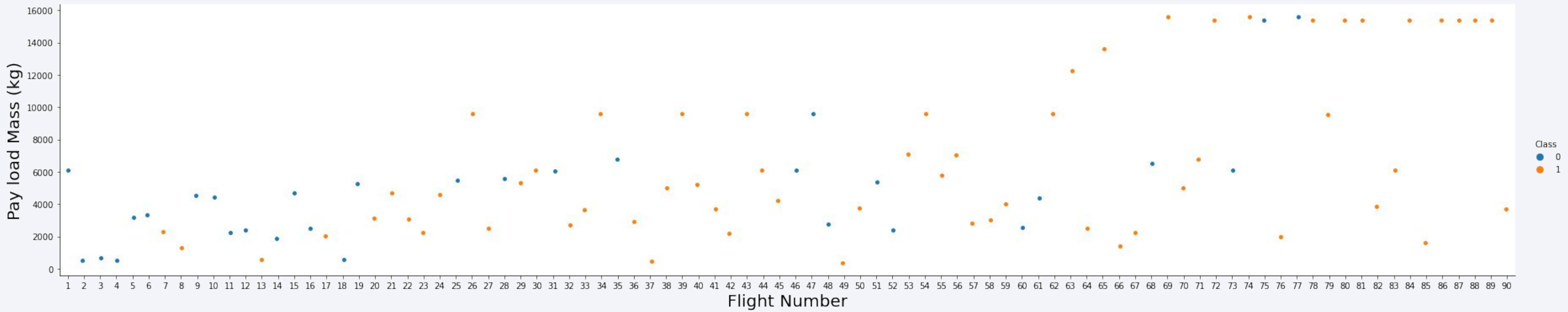
Results

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

Section 2

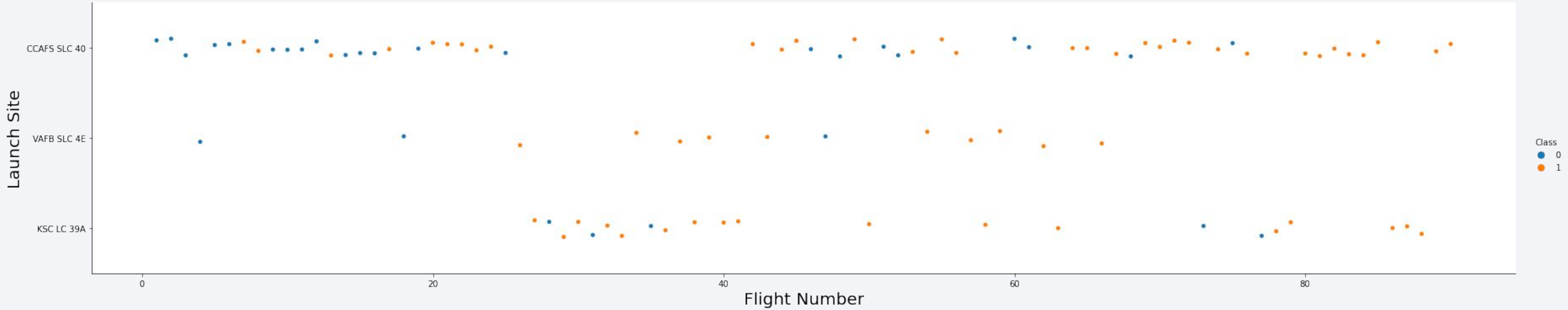
Insights drawn from EDA

Flight Number vs. Payload Mass



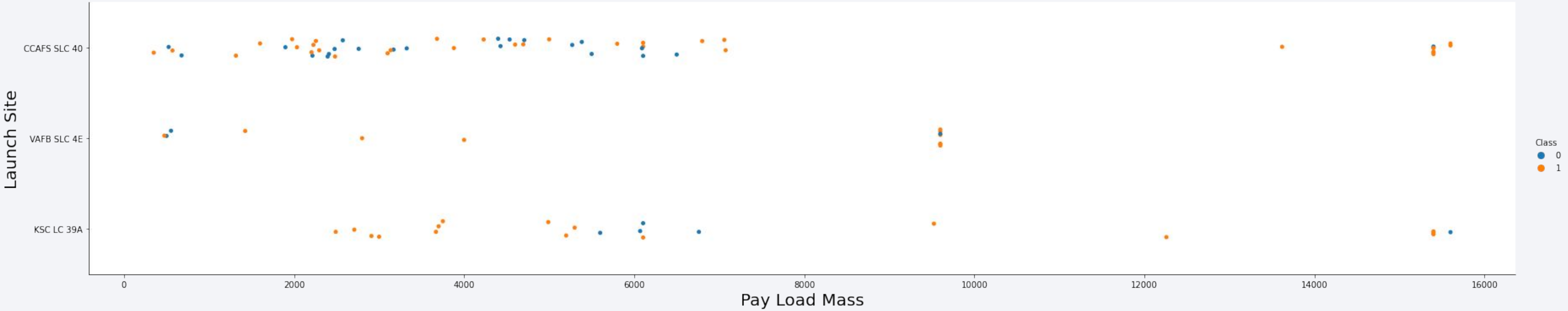
- The plot shows that as the number of flights increases, so does the likelihood of a successful landing.
- The greater the payload mass, the better the landing outcome.

Flight Number vs. Launch Site



- The higher flight number the more success.
- CCAFS SLC 40 launch site has the lowest successful rate (60%)
- VAFB SLC 4E and KSC LC 39A has the most successful rate (77%)

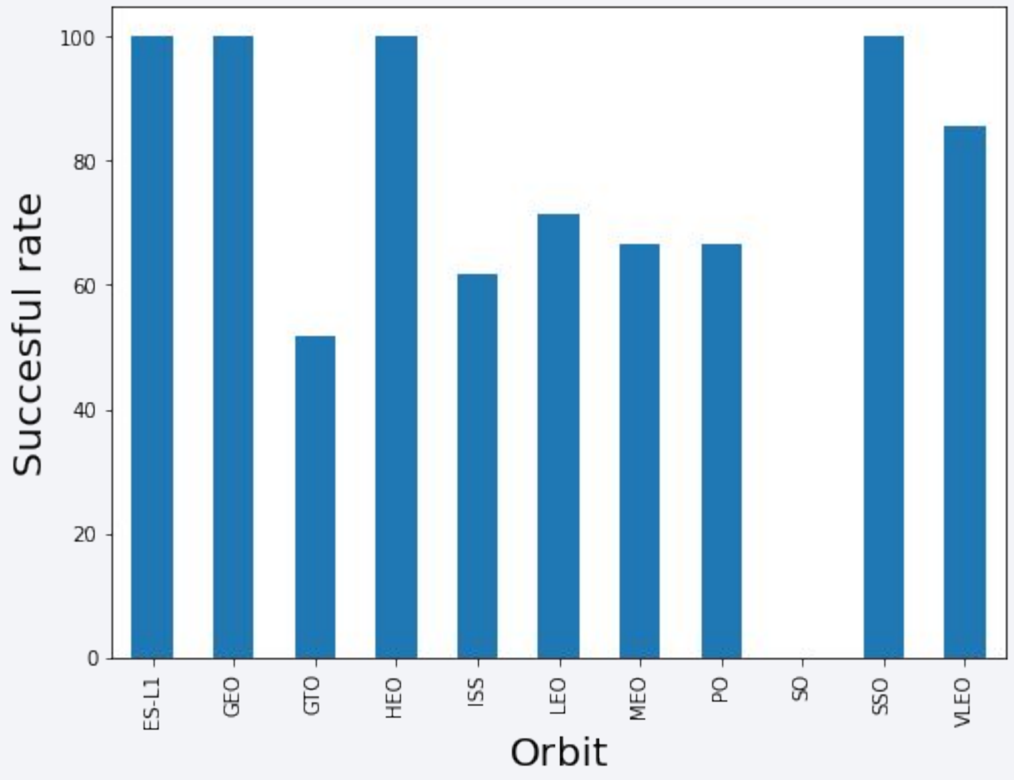
Payload vs. Launch Site



- The VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000)
- The CCAFS SLC 40 launch site has the most rockets launched, but mostly below 8000
- The KSC LC 39A has the most successful rate compare to other sites

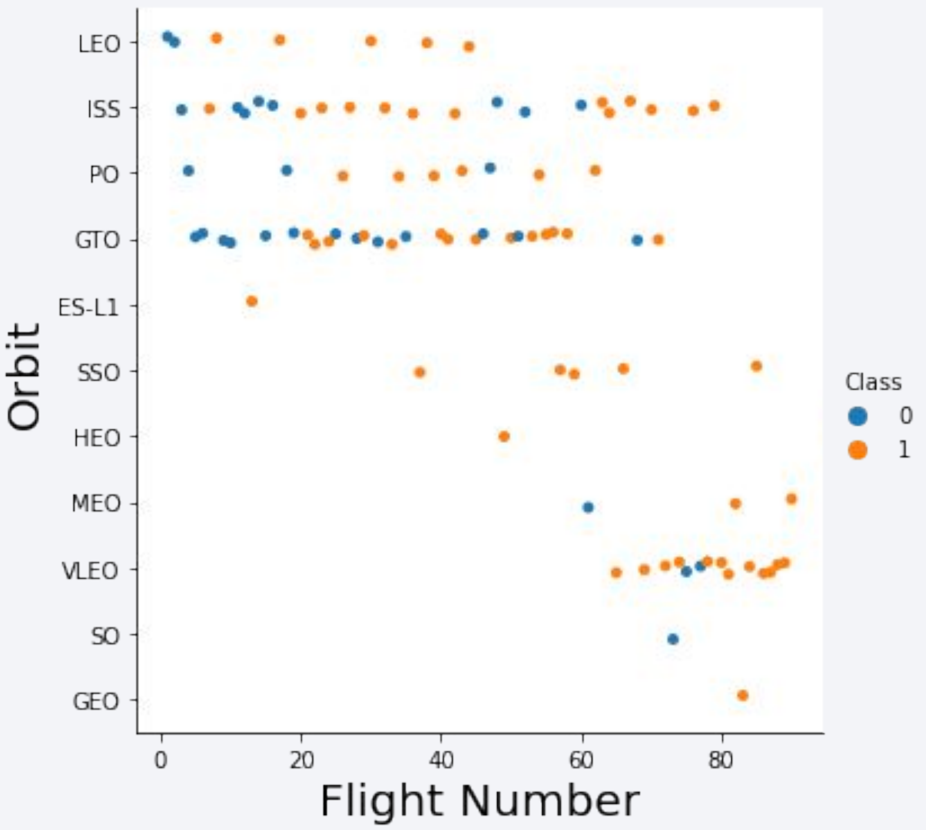
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO are the orbits that have 100% success rate.
- GTO is the orbit that has lowest success rate (about 55%)
- ISS, LEO, MEO, PO orbits have similar success rate (60%-70%)



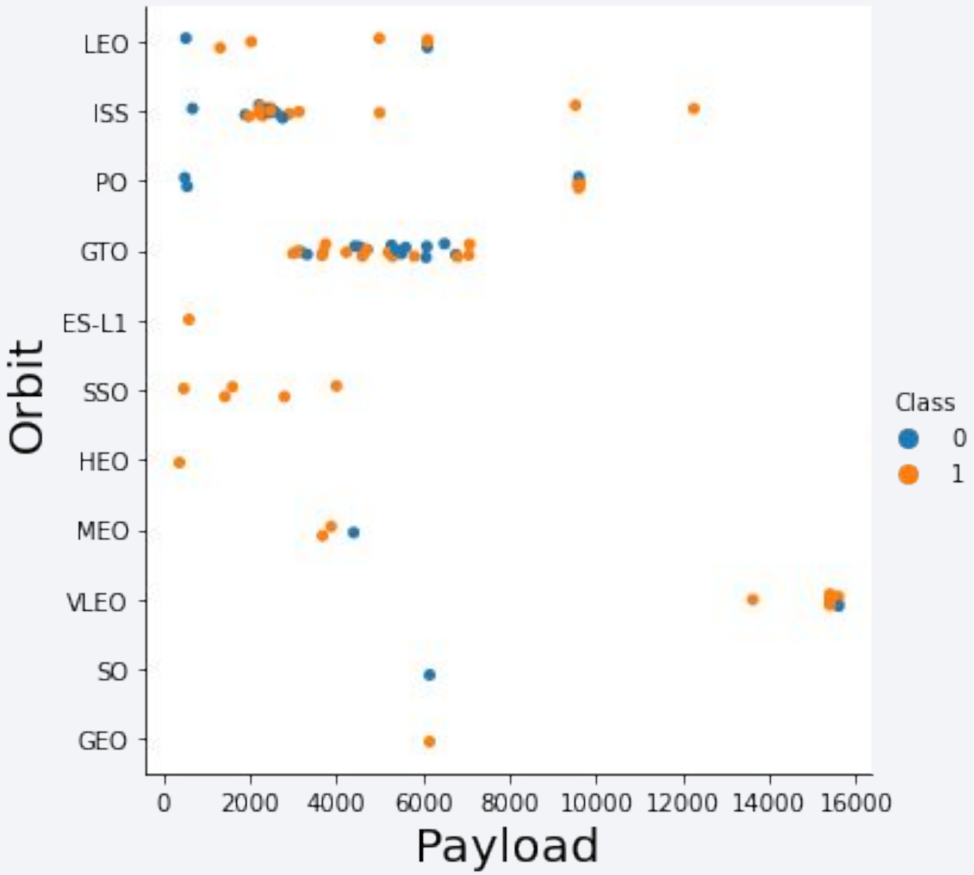
Flight Number vs. Orbit Type

- LEO is the only orbit where the success rate is related to the amount of flights.
- In other orbits, there is no relationship between success rate and flight number.



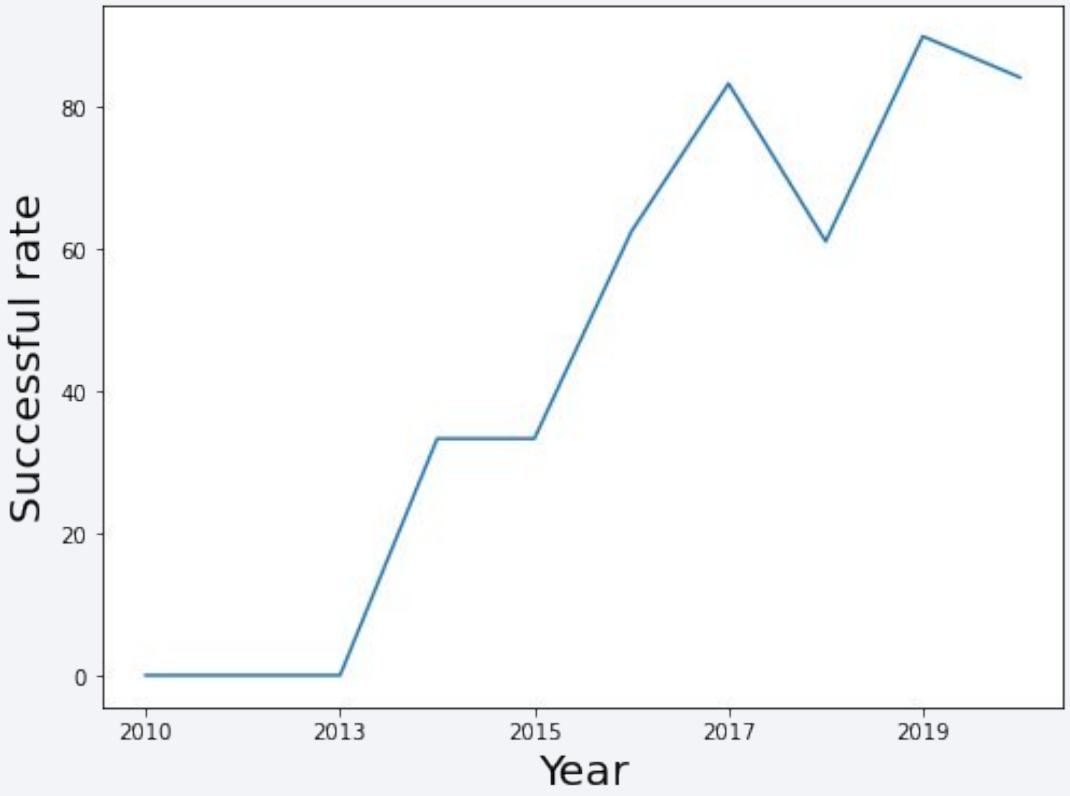
Payload vs. Orbit Type

- We can see that LEO, ISS, and PO have a higher success rate when the payload mass is large.
- There is no relationship between success rate and payload in other orbits.



Launch Success Yearly Trend

- The success rate kept increasing since 2013.
- The reasons could be the improvement after each failure and the advancement of technology over time.



All Launch Site Names

- Find the names of the unique launch sites

```
%%sql
select distinct LAUNCH_SITE
from SPACEXDATASET
```

- Use the keyword ***distinct*** to select the unique launch sites

Launch Site Names Begin with 'CCA'

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

```
%%sql
select distinct LAUNCH_SITE
from SPACEXDATASET
where launch_site like 'CCA%'
```

- Use ***distinct*** and ***where*** to select the launch sites with the name starts with **CCA**

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%%sql
select sum(PAYLOAD_MASS__KG_) as total_payload_mass
from SPACEXDATASET
where CUSTOMER = 'NASA (CRS)'
```

- Use ***sum*** function to summary the total payload mass and ***where*** statement to apply condition (customer is NASA)

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 average_payload_mass
from SPACEXDATASET
where BOOSTER_VERSION = 'F9 v1.1'
```

- Use ***avg*** function to calculate the average of payload mass and ***where*** statement to apply condition (BOOSTER_VERSION is F9 v1.1)

First Successful Ground Landing Date

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

```
%%sql
select min(DATE)
from SPACEXDATASET
where LANDING__OUTCOME = 'Success (ground pad)'
```

- Use ***min*** function to get the first date (minimum) of the first successful landing outcome on ground pad.
- Use ***where*** statement to apply condition (LANDING__OUTCOME is 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

```
%%sql
select BOOSTER_VERSION
from SPACEXDATASET
where LANDING__OUTCOME = 'Success (drone ship)' and
(PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)
```

- Use the comparison operators **>** and **<** to select the PAYLOAD_MASS__KG_ that greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql
select MISSION_OUTCOME, count(MISSION_OUTCOME) as total
from SPACEXDATASET
group by MISSION_OUTCOME
```

- Use ***count*** function to count the number of mission outcomes, and using ***group by*** statement to group the value by the mission outcomes.

Boosters Carried Maximum Payload

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

```
%%sql
select BOOSTER_VERSION
from SPACEXDATASET
where PAYLOAD_MASS__KG_ = (select
max(PAYLOAD_MASS__KG_) from SPACEXDATASET)
```

- Use the **subquery** to return the maximum payload mass(***max*** function) and get the booster version that has the payload equal to the maximum payload mass

2015 Launch Records

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

```
%%sql
select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
from SPACEXDATASET
where year(DATE) = 2015 and LANDING__OUTCOME = 'Failure (drone
ship)'
```

- Use ***year*** function to extract the year from DATE to select the year in 2015

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

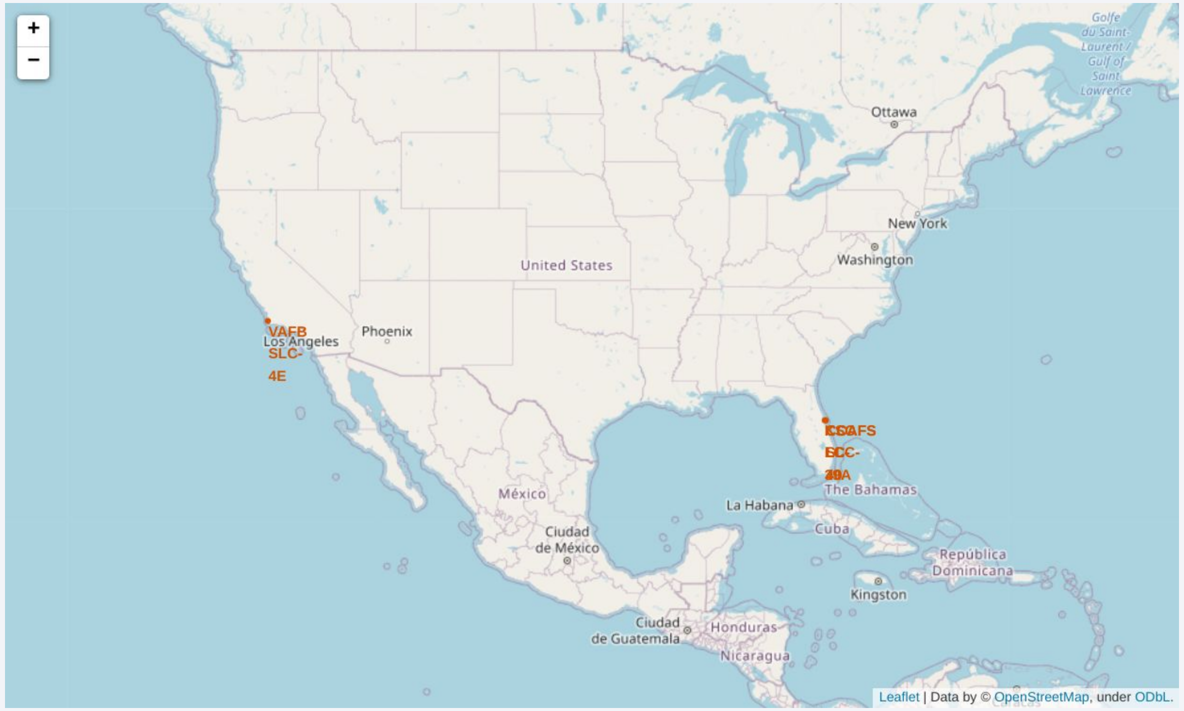
```
%%sql
select LANDING__OUTCOME, count(LANDING__OUTCOME) as total
from SPACEXDATASET
where DATE between '2010-06-04' and '2017-03-20'
group by LANDING__OUTCOME
order by total desc
```

- Use ***between*** operator to get the date from 2010-06-04 and 2017-03-20.
- Use ***group by*** statement to group result by landing outcome
- Use ***order by*** statement with ***desc*** to sort the result in descending total order.

Section 4

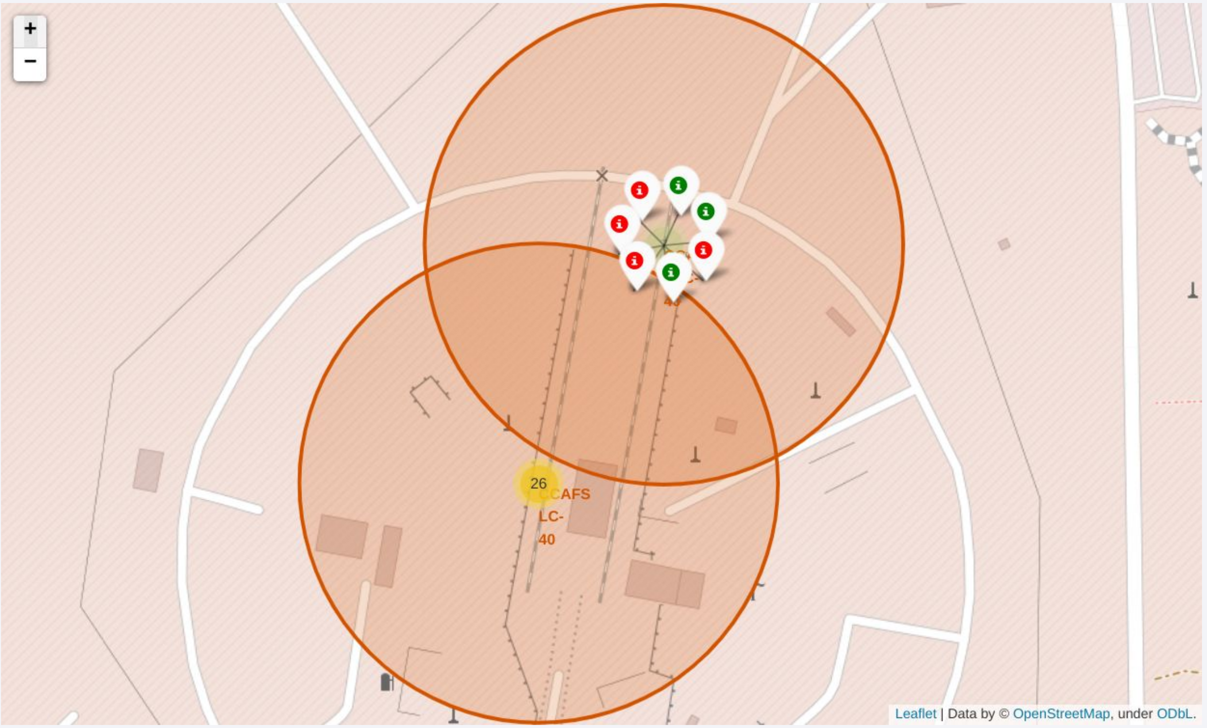
Launch Sites Proximities Analysis

Launch site locations



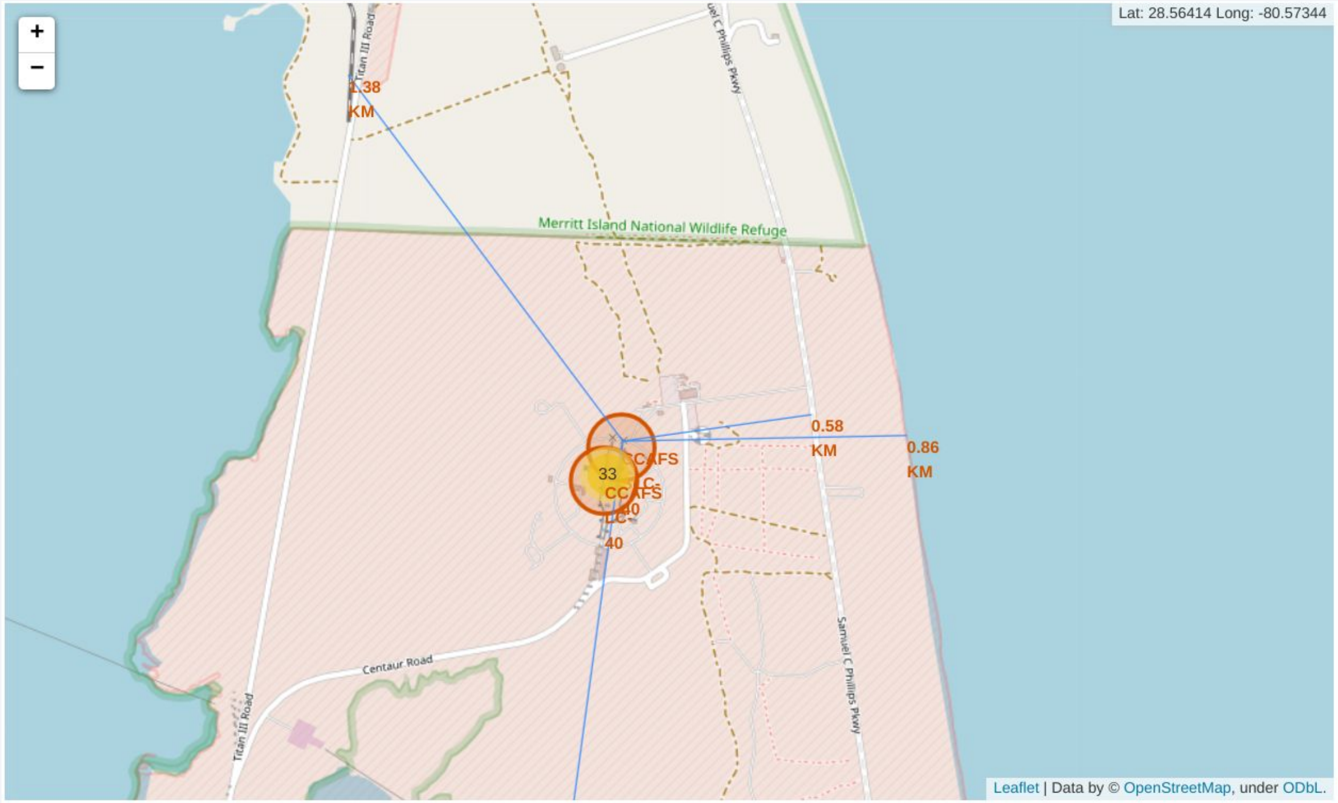
- All the launch sites in proximity to the Equator line, it's because of the rotational speed of Earth that helps the launch.
- All the launch sites is located very close to the coast to reduce the risks over populated areas.

Landing outcome on each landing site



- Based on the color-labeled markers in marker clusters, we can see the KSC LC-39A has the most successful landing outcome (10 success out of 13 landings)
- CCAFS LC-40 has the lowest success rate, with 19 failures out of 26 landings.

Launch site vs its proximities

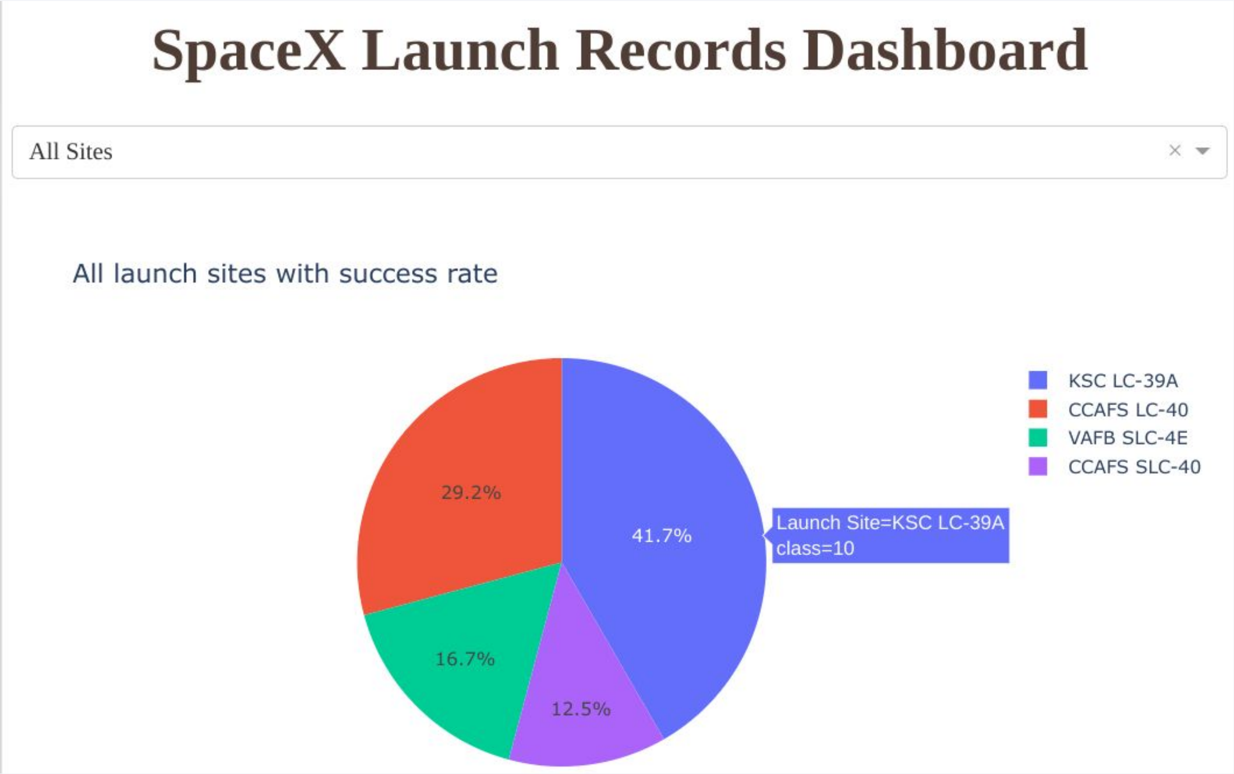


- The launch site (CCAFS SLC-40) in close proximity to railway: 1.38 Km
- The launch site (CCAFS SLC-40) in close proximity to highway: 0.58 Km
- The launch site (CCAFS SLC-40) in close proximity to coastline: 0.86 Km
- The launch site (CCAFS SLC-40) keeps 49.86 Km away from Melbourne city

Section 5

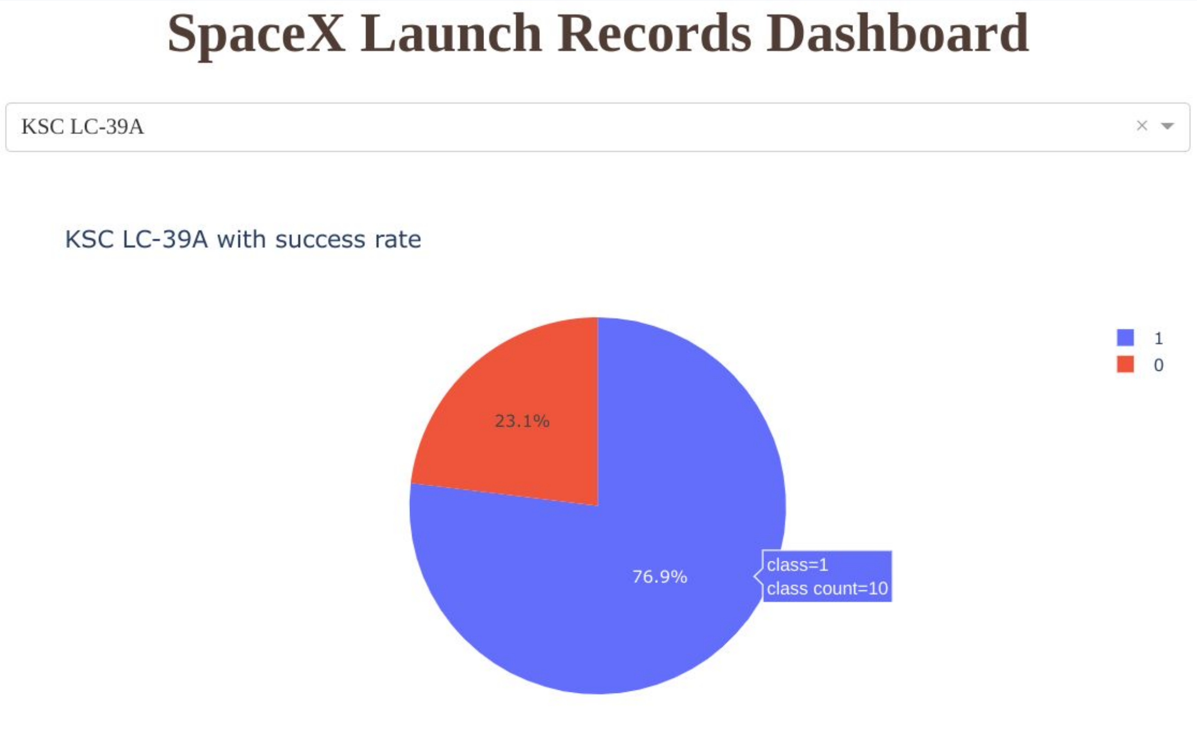
Build a Dashboard with Plotly Dash

Launch sites success rate



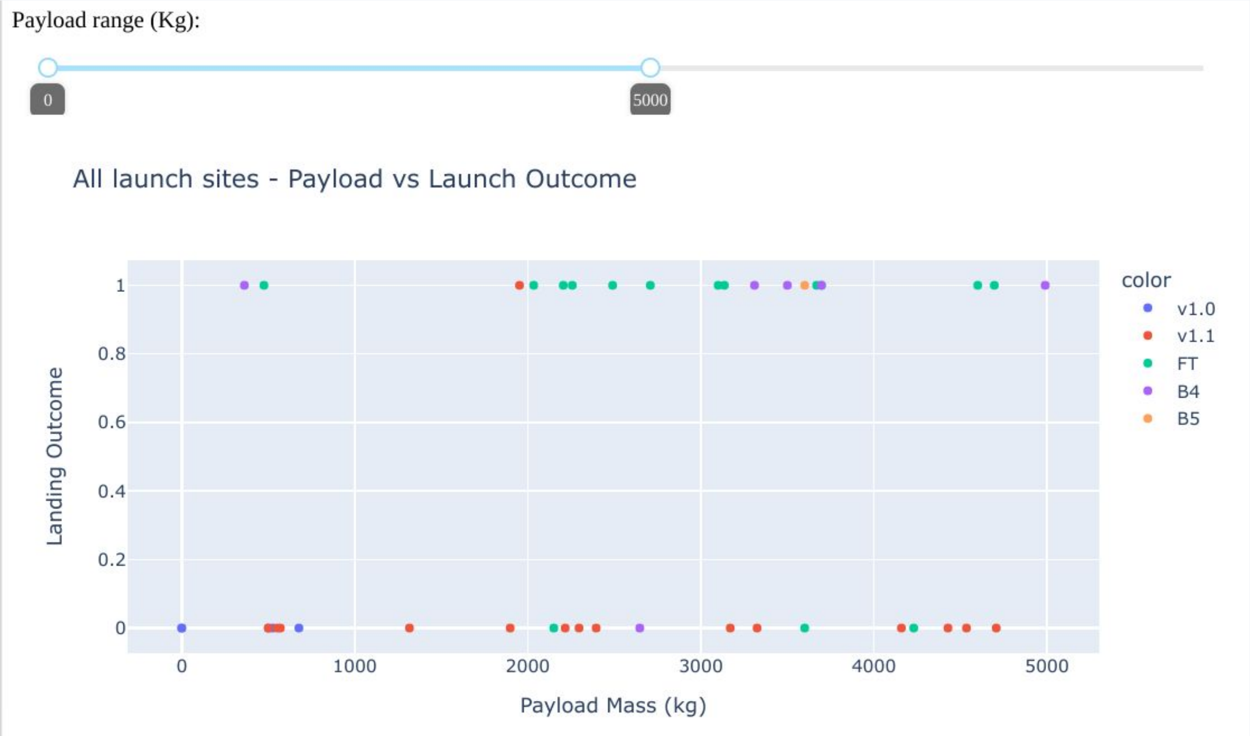
- The KSC LC-39A has the highest landing success rate (41.7 percent of total rate).
- CCFS SLC-40 has the lowest landing success rate (12.5%)

Launch site - success ratio



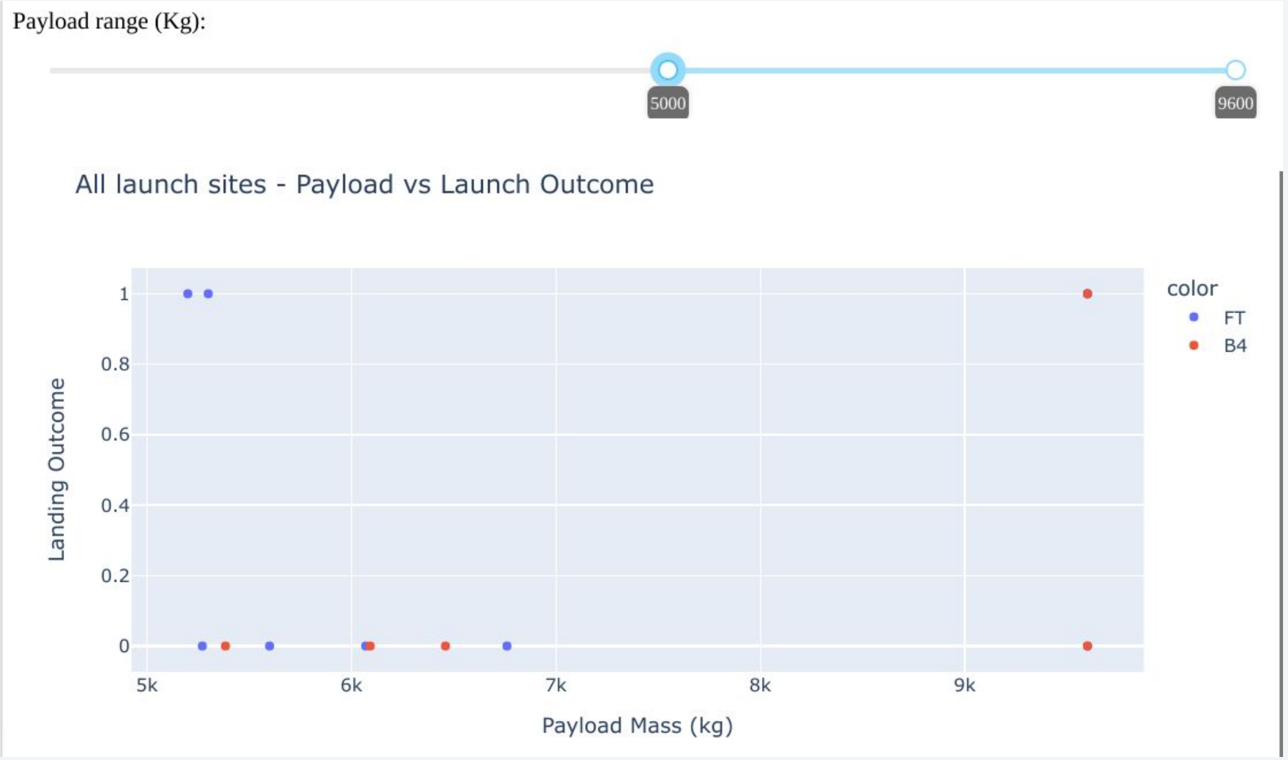
- KSC LC-39A has the highest launch success ration (76.9% success vs 23.1% fail).

Success rate - Payload vs Landing outcome



- Payload range 0 - 5000: FT version have the largest success rate vs V1.1 version has the most failure rate.

Success rate - Payload vs Landing outcome

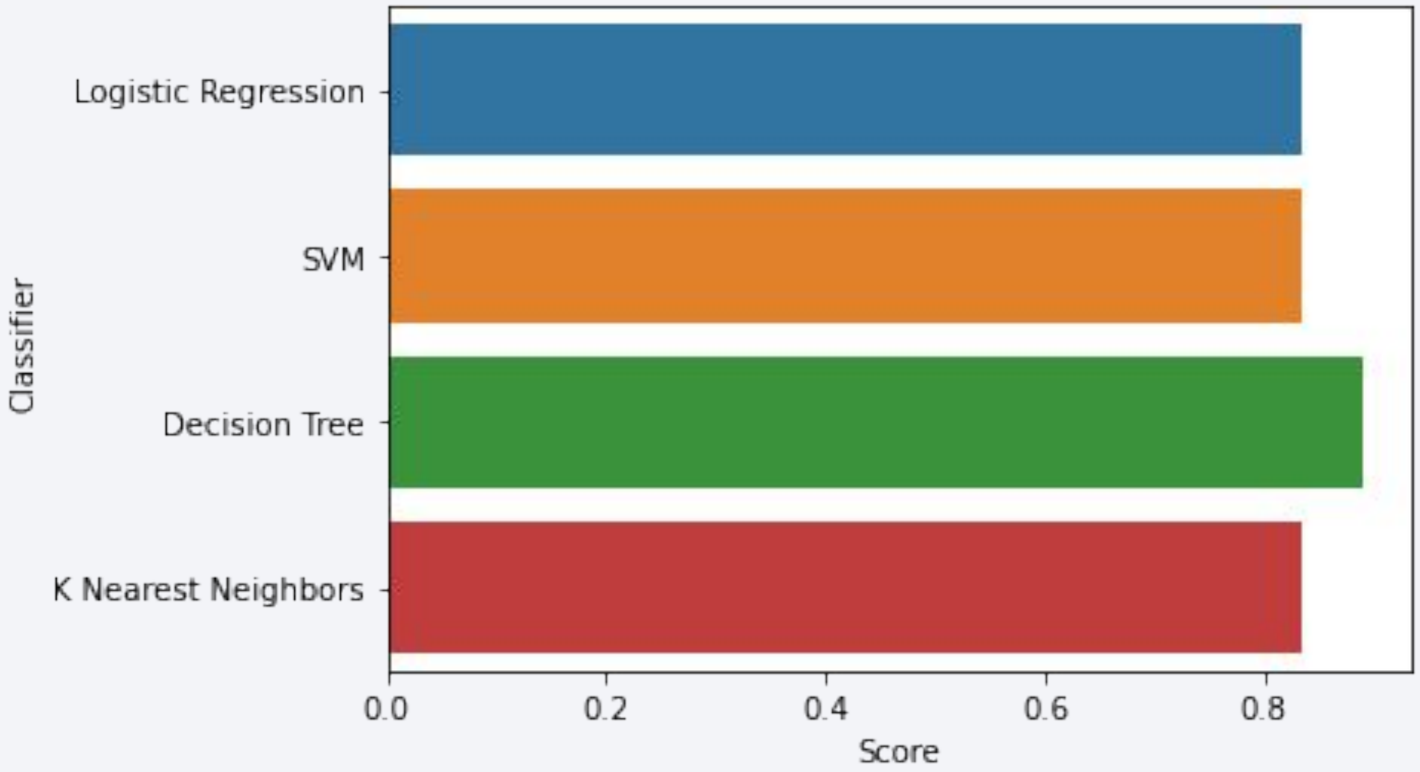


- Payload range 5000 - 9600: FT version have the most success rate vs B4 version has the most failure rate.
- In heavy payload mass, only FT and B4 boosters are used (greater than 5000 kg).

Section 6

Predictive Analysis (Classification)

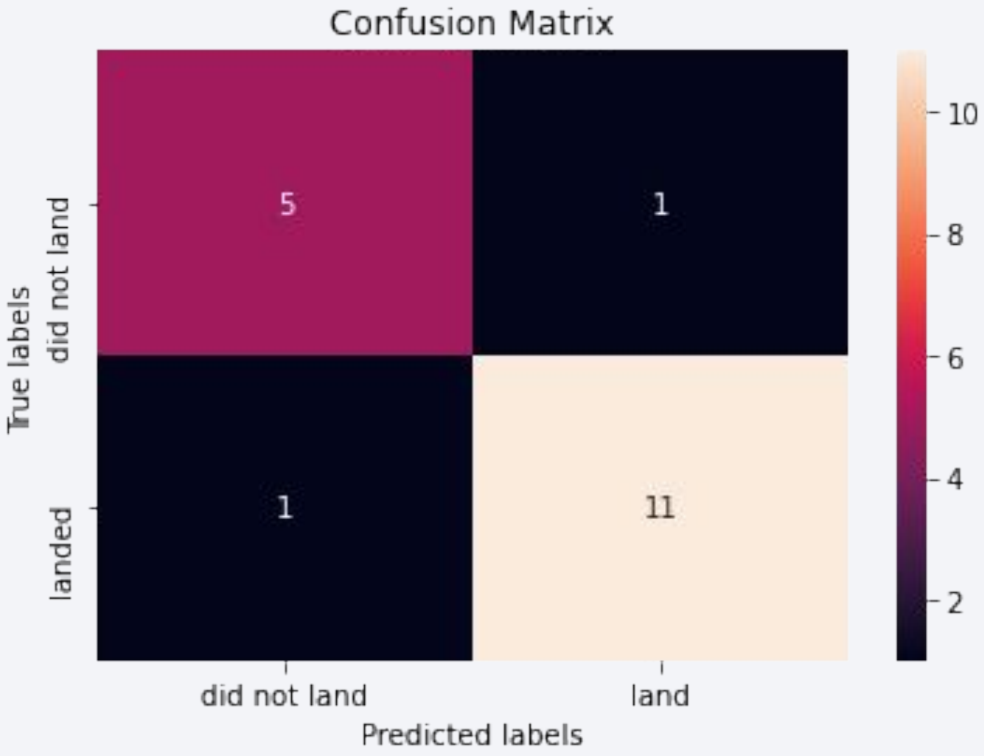
Classification Accuracy



- As the chart shows, the Decision Tree classifier has the highest score (88% accuracy)

Confusion Matrix

- The confusion matrix plot shows the accuracy of the model predictions:
  - 11 True Positives vs 1 False Negative
  - 5 True Negatives vs 1 False Positive



Conclusions

- The greater payload mass, the better landing outcome.
- KSC LC-39A has the most success rate compare to others.
- ES-L1, GEO, HEO, SSO are the orbits that have 100% success rate.
- The reasons could be the improvement after each failure and the advancement of technology over time.
- All the launch sites in proximity to the Equator line
- All the launch sites is located very close to the coast to reduce the risks over populated areas.
- The Decision Tree Classifier is the best model.

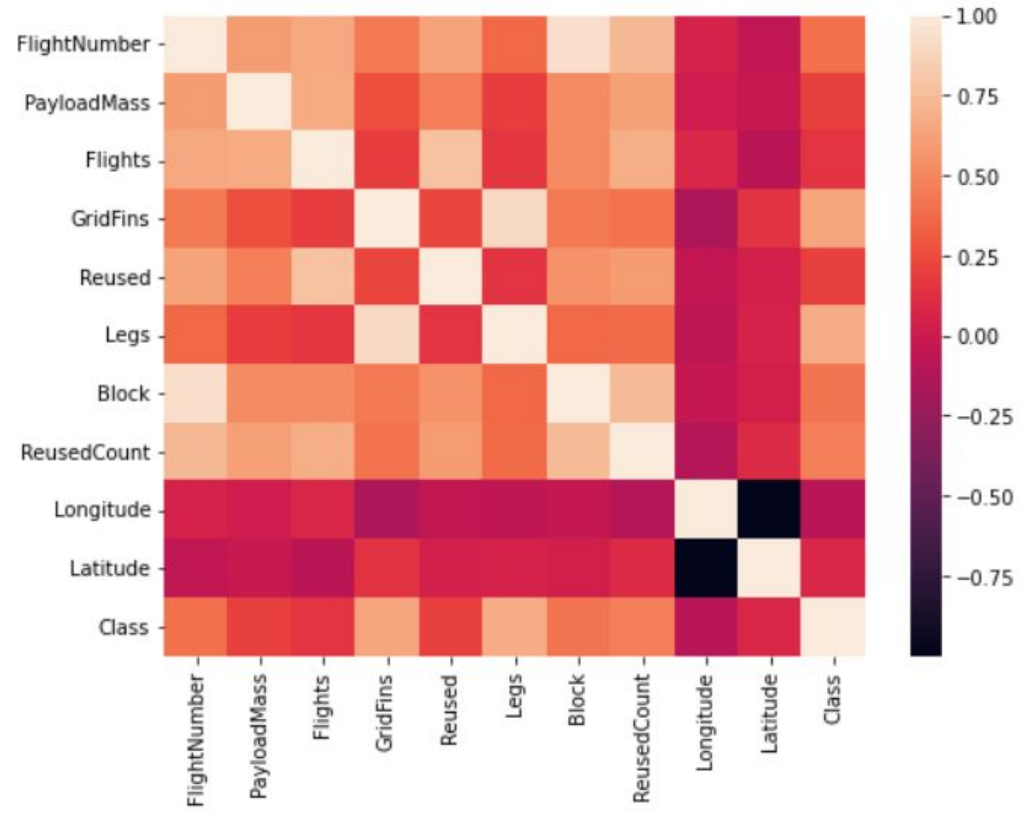
Appendix

- Use seaborn heatmap plot to show the correlation in the dataset.
- Show the flight number over the year.

Appendix

```
plt.figure(figsize=(8, 6))
sns.heatmap(data=df.corr())
```

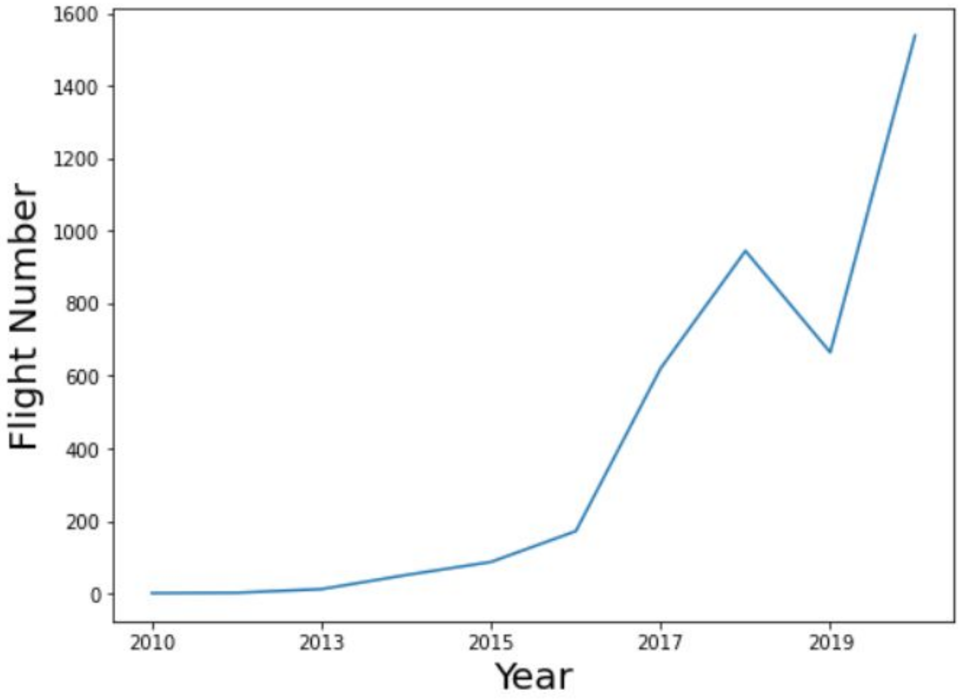
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fec25a77490>
```



Appendix

```
flight_by_year_df = df.groupby('Year')['FlightNumber'].sum()
flight_by_year_df.plot(figsize=(8, 6))
plt.xlabel("Year", fontsize=20)
plt.ylabel("Flight Number", fontsize=20)
```

```
Text(0, 0.5, 'Flight Number')
```



Thank you!