



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Matthias Herrmann>
<16-April-2025>



Outline

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

Executive Summary

Summary of methodologies

- Data collection via SpaceX API and webscraping
- EDA with SQL and Python
- Visualization with Folium (maps) and Plotly Dash (Dashboard)
- Machine learning with classification models Logistic regression, Decision Tree, KNN, Support Vector machine

Summary of all results

- Success rate of launches is increasing over the years
- As higher the payloadmass as better success rate for the landing
- Orbits with highest success rate are the following SE-L1, GEO, HEO, SSO
- KSC LC-39A is the most successful Launch site
- Payload between 2600 – 4000 seems to be the most successful payload range
- The “FT” Booster Version seems to be the most successful Booster Version

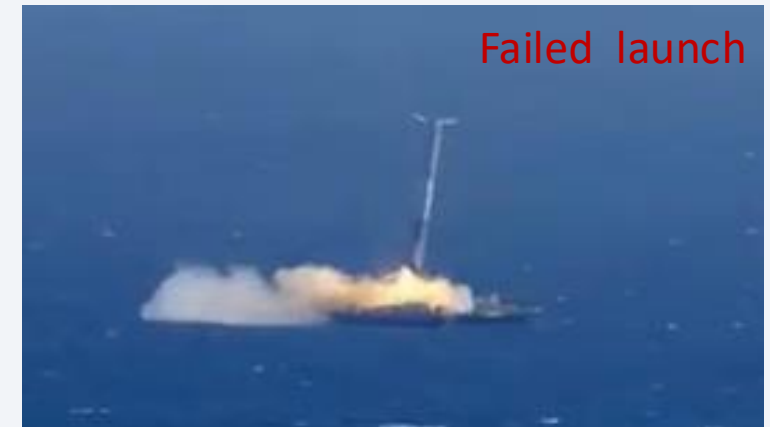
Introduction

Project background and context

- Space X 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 Space X 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 space X for a rocket launch. In this lab, you will create a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.

Goal

- The project goal is to do a EDA with SQL, Python, create a Folium map and a Plotly Dash Dashboard for better understanding, as well as ML model for a “Space X Falcon 9 First Stage Landing Prediction”¶



Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - data via SpaceX API and webscraping
- Perform data wrangling
 - Extract, load and transform/clean the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - 4 classification models have been built (KNN, SVM, Decision Tree, logistic regression) and tuned with gridsearch looking for the best parameters and validated with a confusion matrix and the accuracy score

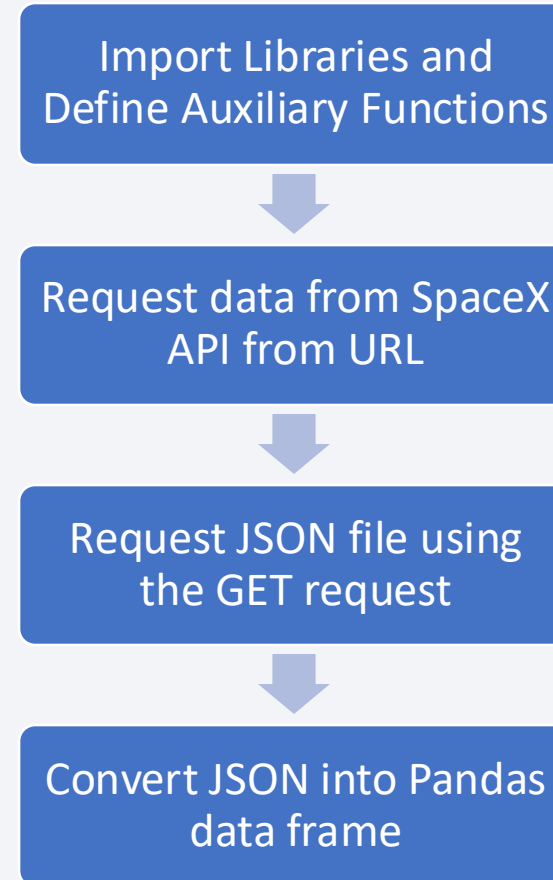
Data Collection

The data has been collected via 2 different ways:

- Request data via SpaceX API
- Extract Falcon 9 launch records HTML table from Wikipedia with BeautifulSoup

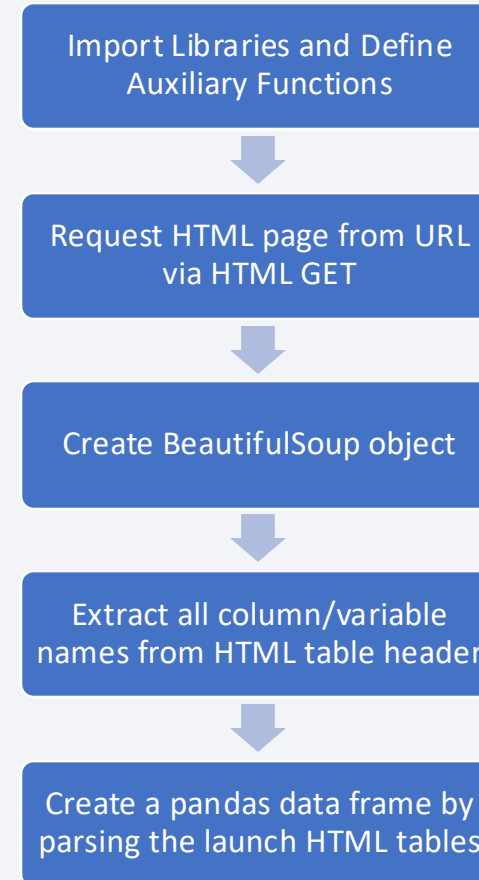
Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
- GitHub URL:
https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_1_Data_collection_api.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- GitHub URL:
https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_2_Web scraping.ipynb



Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- GitHub URL:
https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_3_Data_wrangling.ipynb

Identification of rows with missing values and duplicated rows



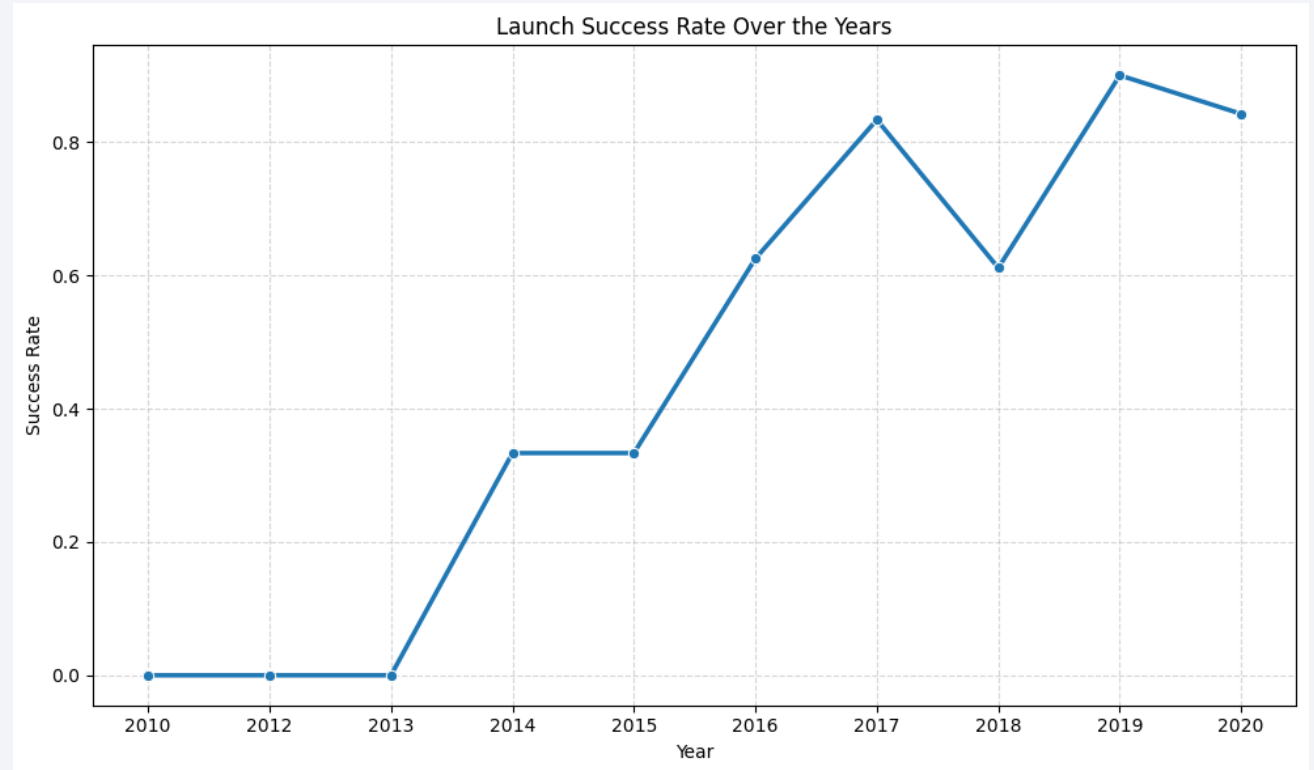
Identify which columns are numerical and categorical



Dealing with missing values (use mean for Payloadmass)

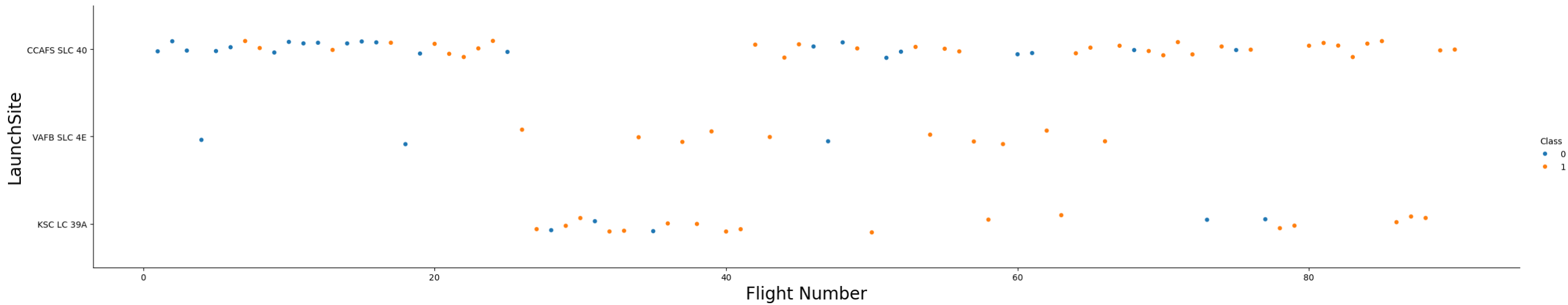
EDA with Data Visualization I

- The line chart “Launch success rate over the years” shows that the success rate since 2013 kept increasing till 2020
- GitHub URL:
https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_5_Eda_data_viz.ipynb



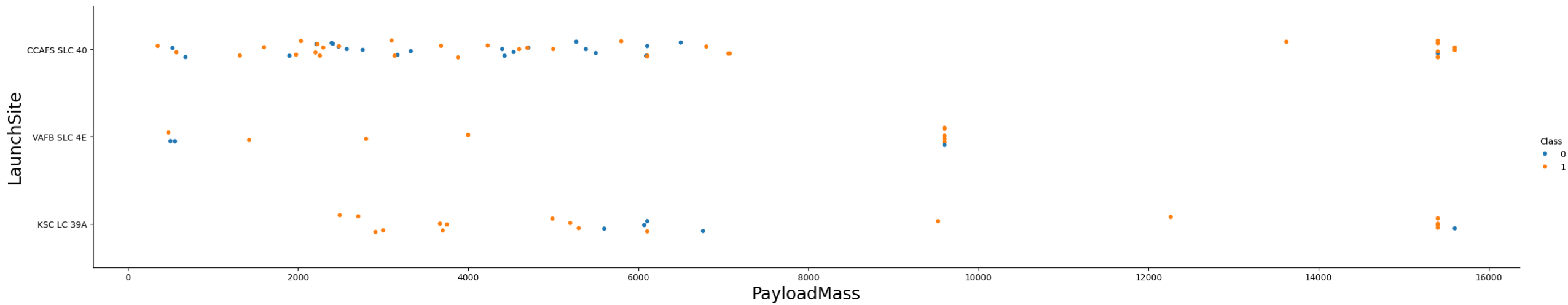
EDA with Data Visualization II

- “Scatter Plot Launch Site versus Flight Number by Class”: It seems like over the time with higher flight number the successful landing increases for all 3 launch sites



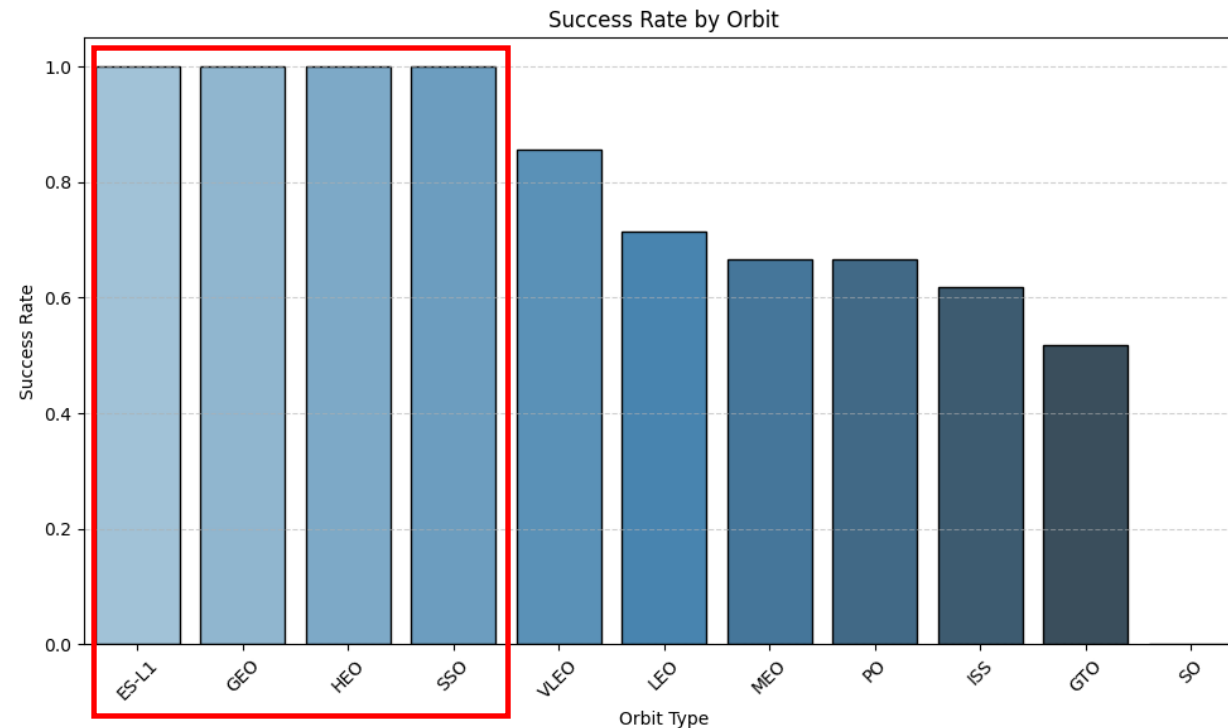
EDA with Data Visualization III

- “Scatter Plot Launch Site versus PayloadMass by Class”: As higher the payloadmass as better success rate for the landing



EDA with Data Visualization IV

- „Bar Chart Success versus Orbit Type“: The chart success rate versus different orbit types shows that the orbits with highest success rate are the following SE-L1, GEO, HEO, SSO



EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- GitHub URL:
https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_4_Eda_sql_coursera_sqlite.ipynb

- Display the names of the unique launch sites
- 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 achieved
- 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 all the booster_versions that 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.

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- GitHub URL:
https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_6_Launch_site_location.ipynb

- Markers have been used to show the locations of the launch places, additionally with circles for zooming function
- Clusters have been used to show several starts at one launch location (they all have similar latitude and longitude parameters)
- Line for the equator has been used to identify if the launch locations are near the equator
- PolyLine between a launch site to e.g. a selected coastline point has been implemented showing the calculated distance

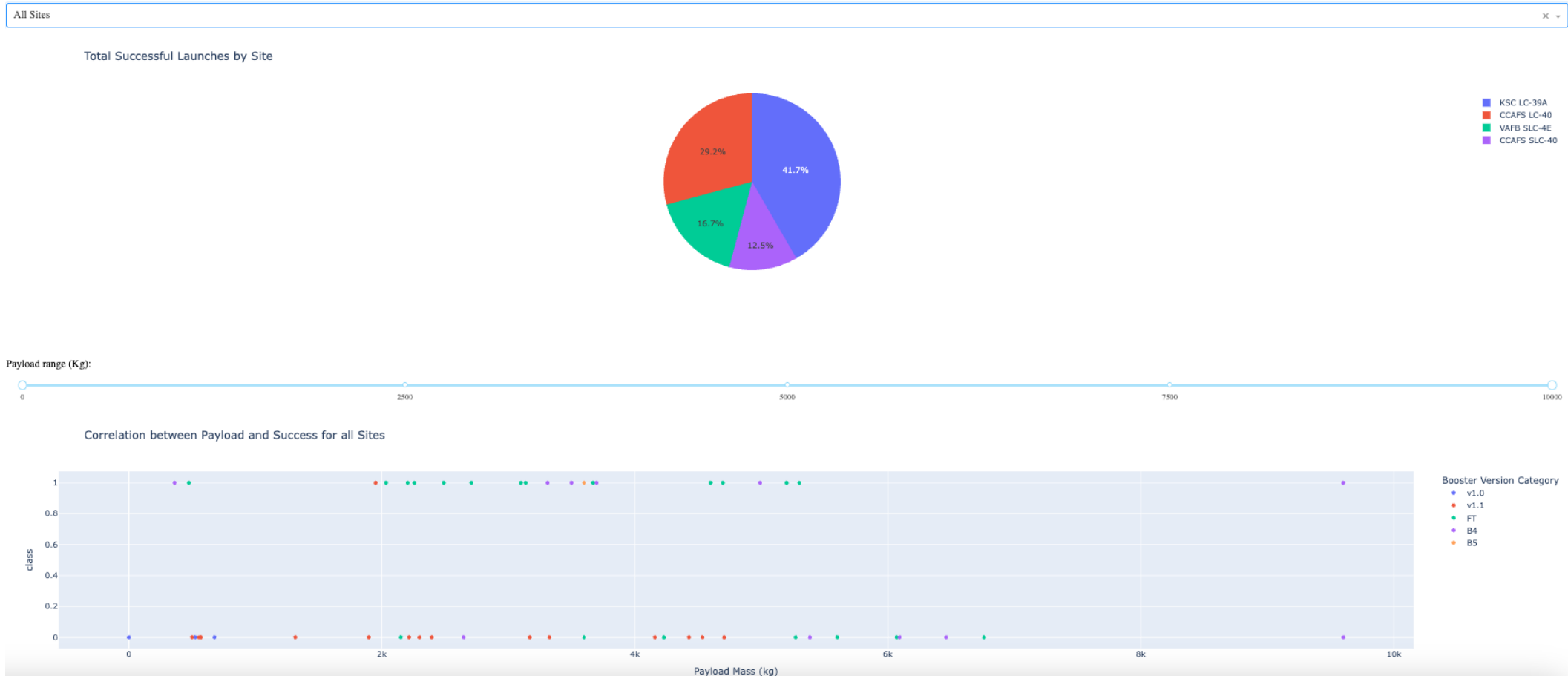
Build a Dashboard with Plotly Dash I

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- GitHub URL:
https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_8_SpaceX_dash_app.py

- One Selection field to select a specific launch location or all
- A pie chart to show success rate of selected launch location/all
- A Range slider to select the payload
- A scatter plot to the correlation between payload and success for specific launch site/all

Build a Dashboard with Plotly Dash II

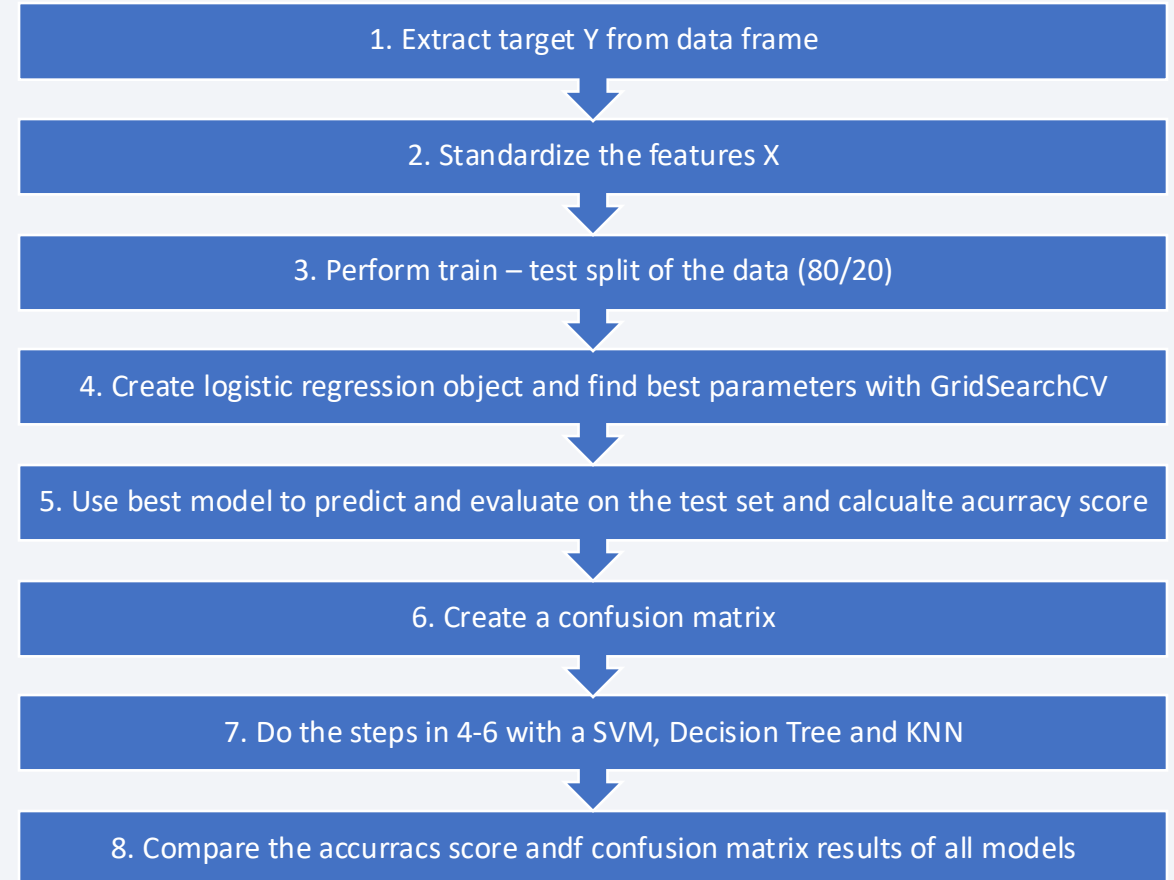
SpaceX Launch Records Dashboard



- GitHub URL: https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_8_Screenshot_Dashboard.png

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- GitHub URL:
https://github.com/NewFishMH/IBM-Data-Science-Professional-Training/blob/432943088a26435ae15643bc6f84d7468529bbdf/Lab_7_Machine_Learning_Prediction.ipynb



Results

Exploratory data analysis results

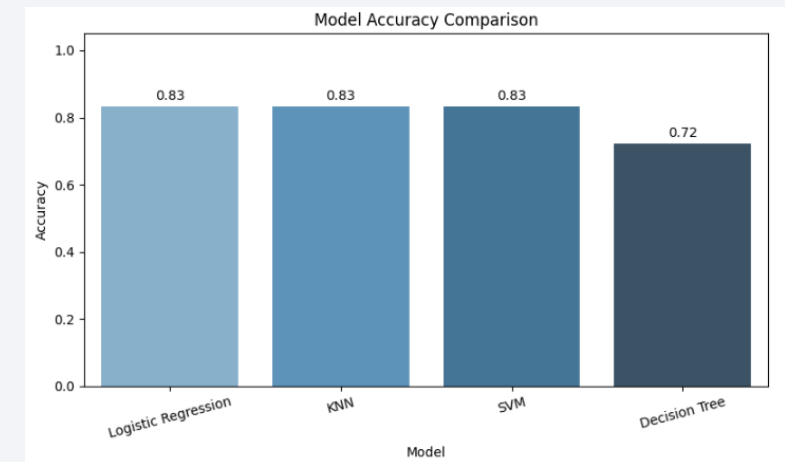
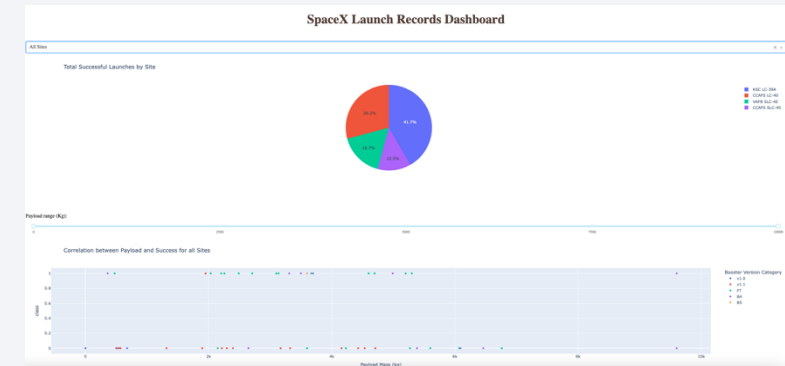
- Success rate of launches is increasing over the years
- As higher the payload mass as better success rate for the landing
- Orbits with highest success rate are the following SE-L1, GEO, HEO, SSO

Interactive analytics demo in screenshots

- See Dashboard screenshot

Predictive analysis results

- See Bar chart screenshot which compares accuracy of different models



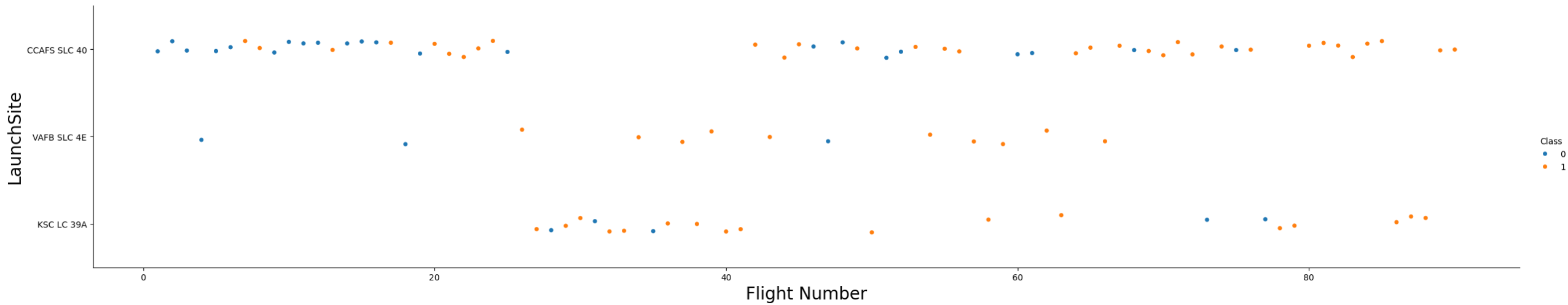
The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is dynamic and technological.

Section 2

Insights drawn from EDA

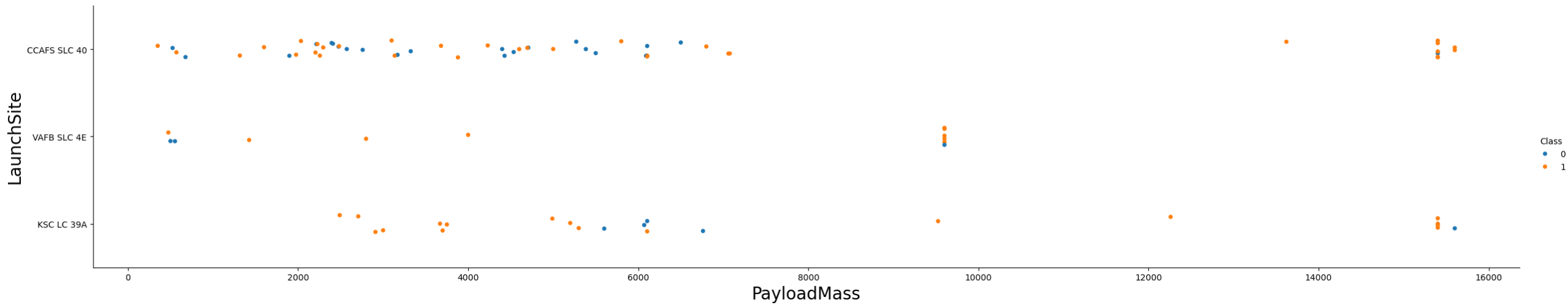
Flight Number vs. Launch Site

- “Scatter Plot Launch Site versus Flight Number by Class”: It seems like over the time with higher flight number the successful landing increases for all 3 launch sites



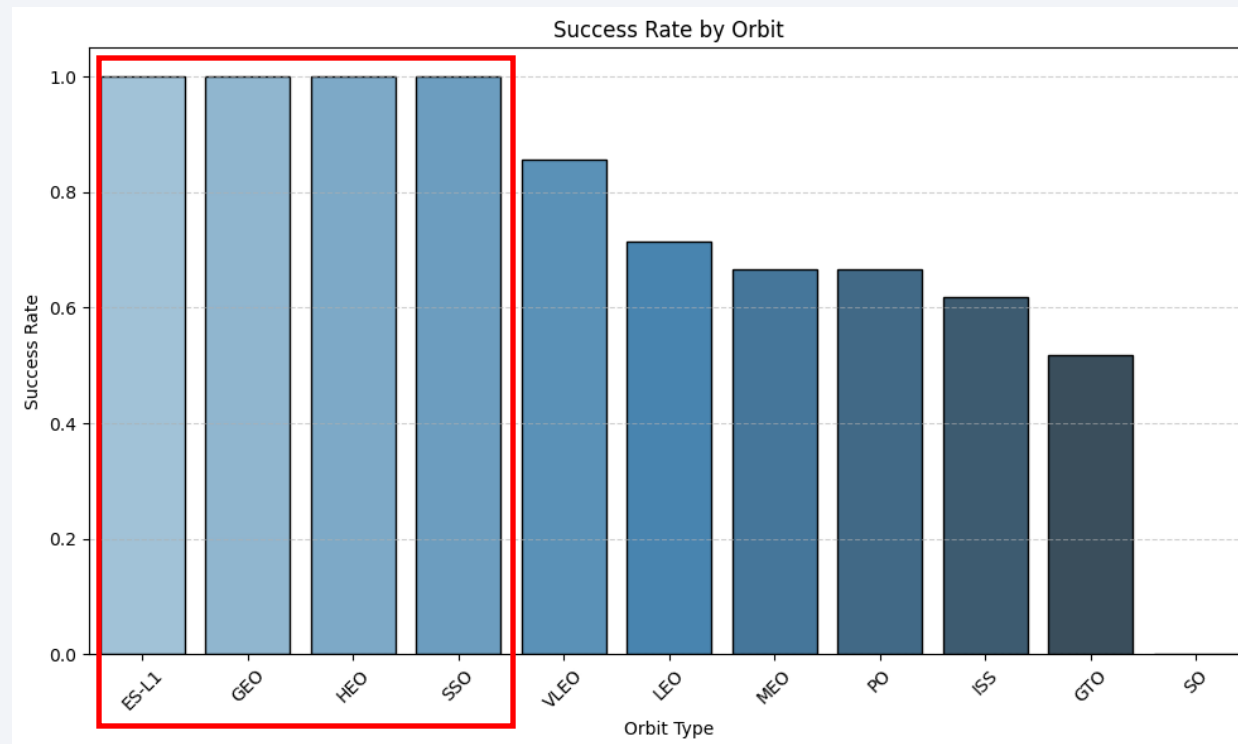
Payload vs. Launch Site

- “Scatter Plot Launch Site versus PayloadMass by Classs”: As higher the payloadmass as better success rate for the landing



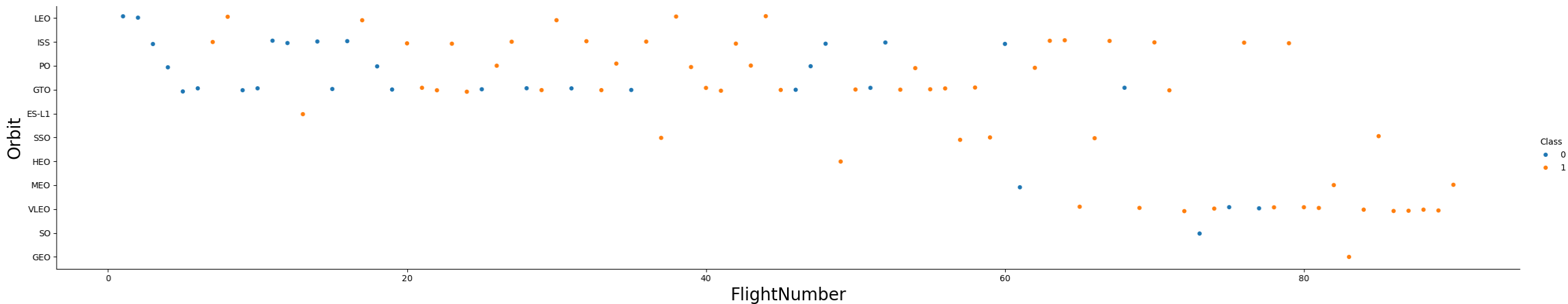
Success Rate vs. Orbit Type

- „Bar Chart Success versus Orbit Type“: The chart success rate versus different orbit types shows that the orbits with highest success rate are the following SE-L1, GEO, HEO, SSO



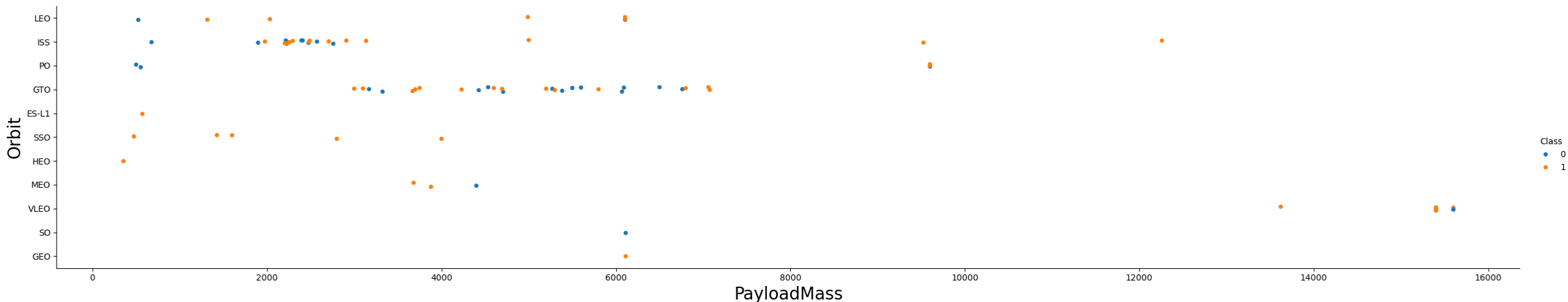
Flight Number vs. Orbit Type

- “Scatter Plot Flight number versus Orbit Type by Class”: You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



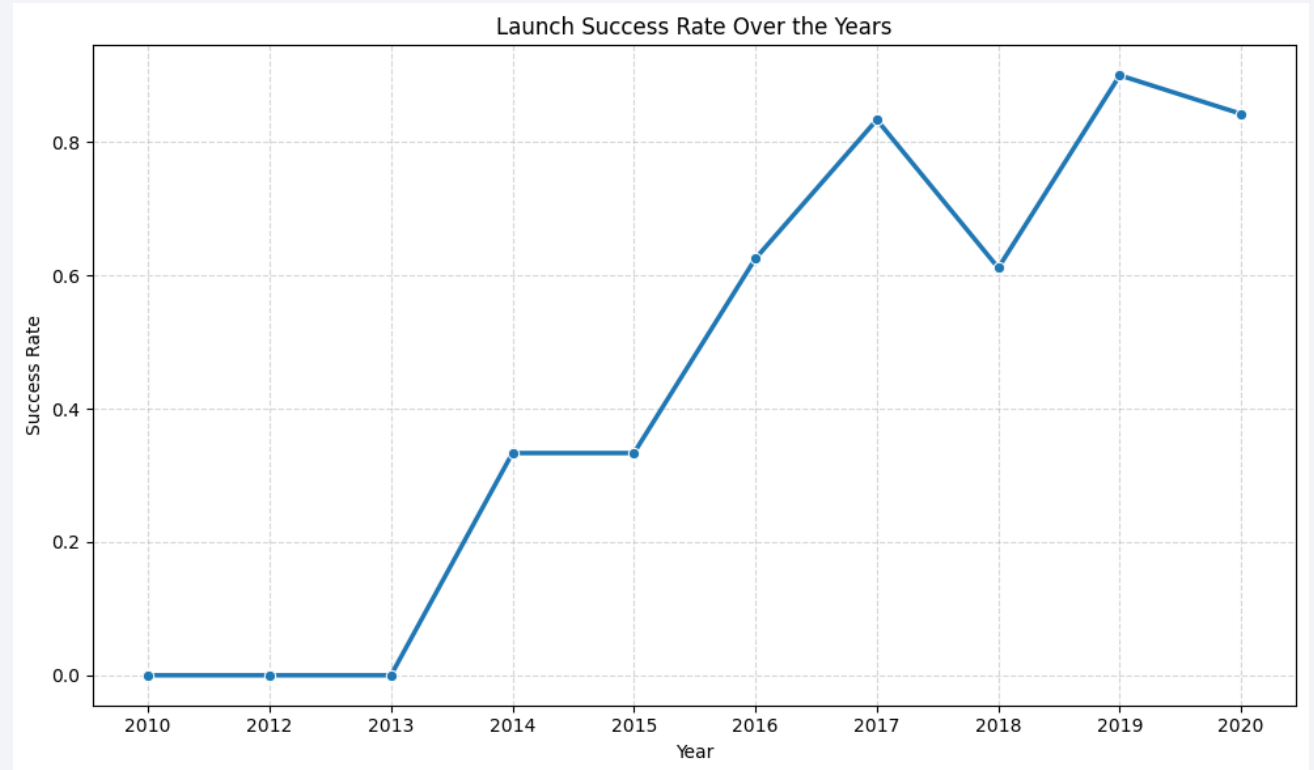
Payload vs. Orbit Type

- “Scatter Plot Payload versus Orbit Type by Class”: With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- The line chart “Launch success rate over the years” shows that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites
- Following SQL Code has been used
- See result in the screenshot on the right

%%sql

```
SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

```
Out [55]:
```

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'
- Following SQL Code has been used
- See result in the screenshot on the right

%%sql

```
SELECT * FROM SPACEXTBL
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

Out[58]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (g
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 (g
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	h
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	h
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	h

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Following SQL Code has been used
- See result in the screenshot on the right

%%sql

```
SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass  
FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```



```
Out[64]: TotalPayloadMass  
45596
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Following SQL Code has been used
- See result in the screenshot on the right

```
%%sql SELECT AVG(PAYLOAD__MASS__KG_) AS  
AveragePayloadMass FROM SPACEXTBL WHERE  
Booster_Version = 'F9 v1.1';
```

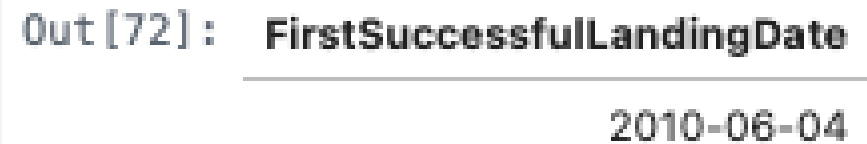


Out[70]: AveragePayloadMass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Following SQL Code has been used
- See result in the screenshot on the right

```
%%sql SELECT MIN(Date) AS  
FirstSuccessfulLandingDate FROM SPACEXTBL WHERE  
Mission_Outcome = "Success"
```



```
Out [72]:  FirstSuccessfulLandingDate  
          2010-06-04
```

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
- Following SQL Code has been used
- See result in the screenshot on the right

```
%%sql SELECT Booster_Version FROM SPACEXTBL  
WHERE Mission_Outcome = "Success" AND  
PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND  
Landing_Outcome = "Success (drone ship)"
```

Out [76] :	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
- Following SQL Code has been used
- See result in the screenshot on the right (2nd: more detailed)

```
%%sql SELECT CASE WHEN Mission_Outcome LIKE  
'Success%' THEN 'Successful' WHEN  
Mission_Outcome LIKE 'Failure%' THEN  
'Unsuccessful' ELSE 'Other' END AS Mission_Status,  
COUNT(*) AS Total_Missions FROM SPACEXTBL  
GROUP BY Mission_Status;
```

Out [87]:

Mission_Status	Total_Missions
Successful	100
Unsuccessful	1

Out [86]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Following SQL Code has been used
- See result in the screenshot on the right

```
%%sql SELECT Booster_Version FROM SPACEXTBL  
WHERE PAYLOAD_MASS__KG_ = (SELECT  
MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

Out [90]: **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 the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Following SQL Code has been used
- See result in the screenshot on the right

```
%%sql SELECT SUBSTR(Date, 6, 2) AS Month,  
Landing_Outcome, Booster_Version, Launch_Site  
FROM SPACEXTBL WHERE -- Landing_Outcome  
LIKE '%Failure%' AND Landing_Outcome LIKE  
'%drone ship%' AND SUBSTR(Date, 1, 4) = '2015';
```

Out [98]:	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
	06	Precluded (drone ship)	F9 v1.1 B1018	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
- Following SQL Code has been used
- See result in the screenshot on the right

```
%%sql SELECT Landing_Outcome, COUNT(*) AS  
Outcome_Count FROM SPACEXTBL WHERE Date BETWEEN  
'2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome  
ORDER BY Outcome_Count DESC;
```

Out [101...]	Landing_Outcome	Outcome_Count
	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

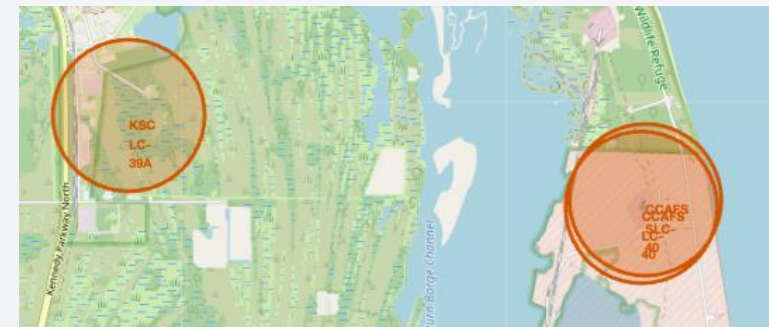
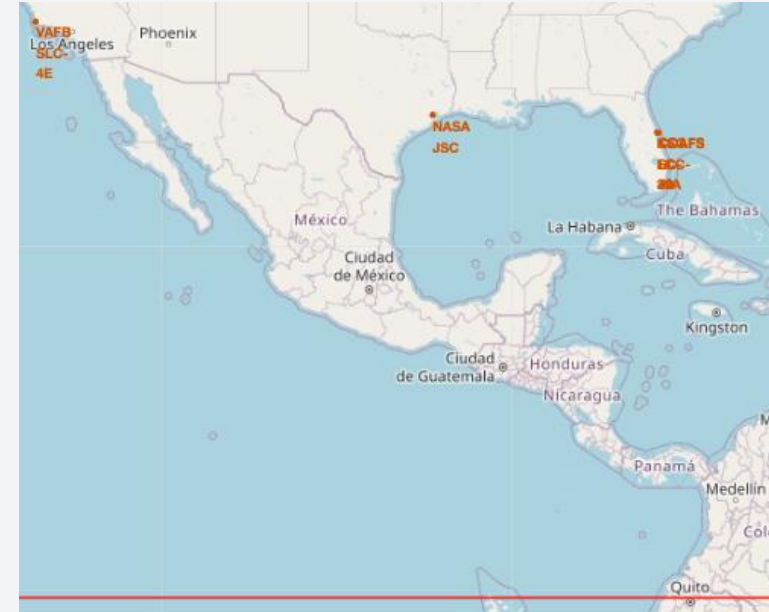
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

Launch Sites Proximities Analysis

Launch site locations

- On the upper map all launch sites and the NASA Johnson space center are shown. The one near Los Angeles called VAFB SLC-4E
- The lower map shows 3 remaining launch sites zoomed-in since they very close from each other and hardly to see on the upper map (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A)



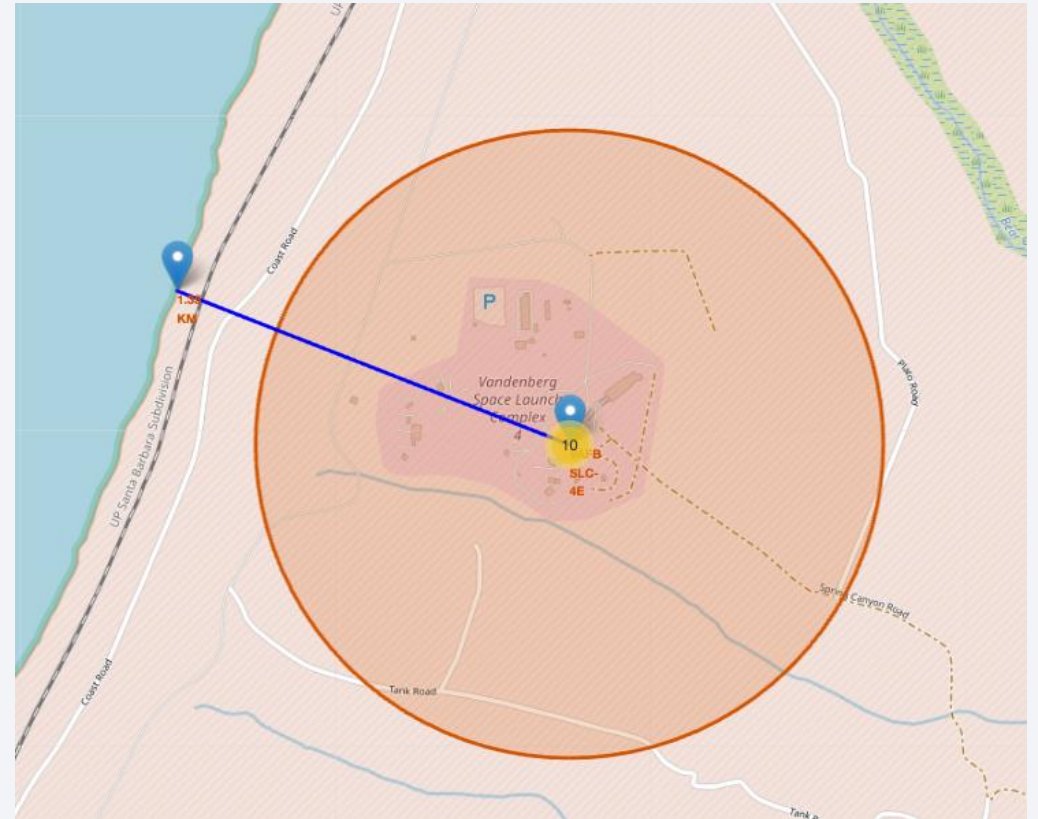
Launch outcomes

- On the map you see the number of launches for each launch site
- When you click on the number you see in red the failed and in green the successful launches



Distance between launch site and coastline

- The chart shows the VAFB SLC-4e launch site and the closed distance to the coastline.
- The orange circle shows the radius around the launch site and the yellow circle with "10" shows the number of launches
- The polyline shows the distance from Marker 1 (launch site) to Marker 2 (coastline)
- The calculated distance using longitude and latitude parameters is 1.35km





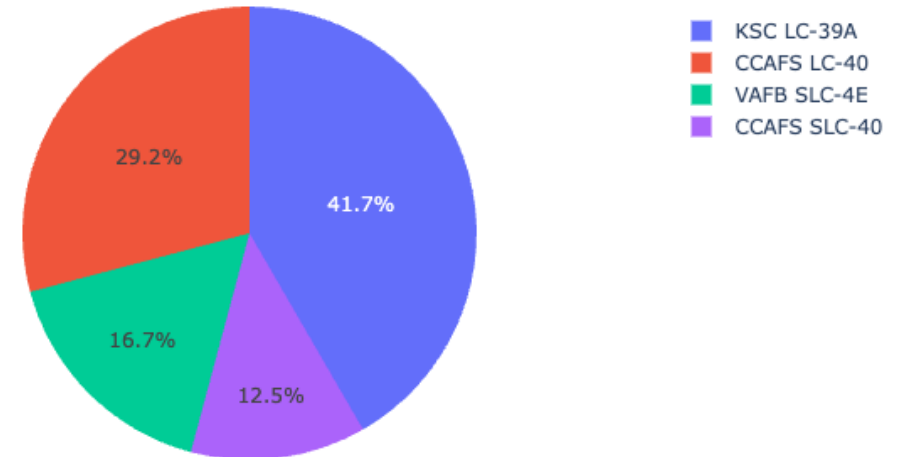
Section 4

Build a Dashboard with Plotly Dash

Launch success count for all sites

- The pie chart shows the successful launches by site
- The KSC LC-39A seems to be the most successful launch site with 41.7% of all sites
- CCAFS SLC-40 seems to be the worst launch site

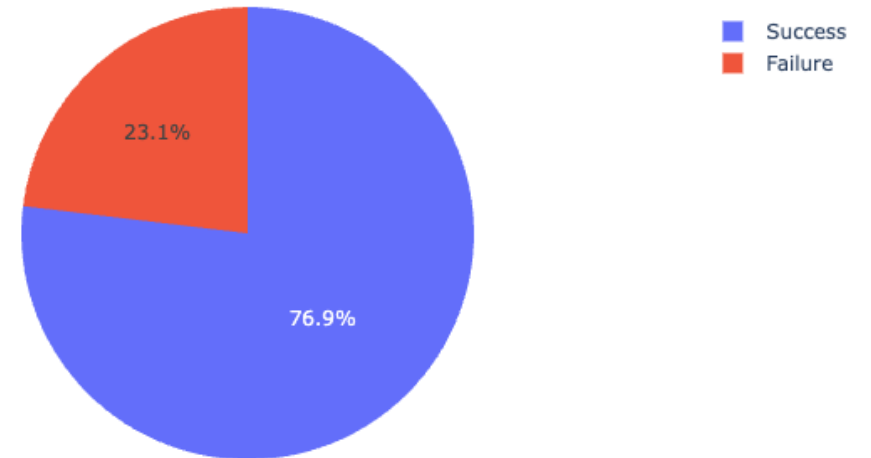
Total Successful Launches by Site



Launch site with highest launch success ratio

- The pie chart shows the KSC LC-39A launch site which has the highest launch success ratio
- There are 76.9% successful launches so far
- And there are 23.1% failed launches recorded

Success vs. Failure for site: KSC LC-39A



Payload versus Launch Outcome

- The screenshots showing “Payload vs. Launch Outcome scatter plots” for all sites, with different payload selected in the range slider
- 1. picture shows that there a lot of failed launches e.g. with a payload between 3000 - 7000
- 2. picture shows that the payload between 2600 – 4000 seems to be the most successful payload range
- 3. picture shows that the “FT” Booster Version seems to be the most successful Booster Version



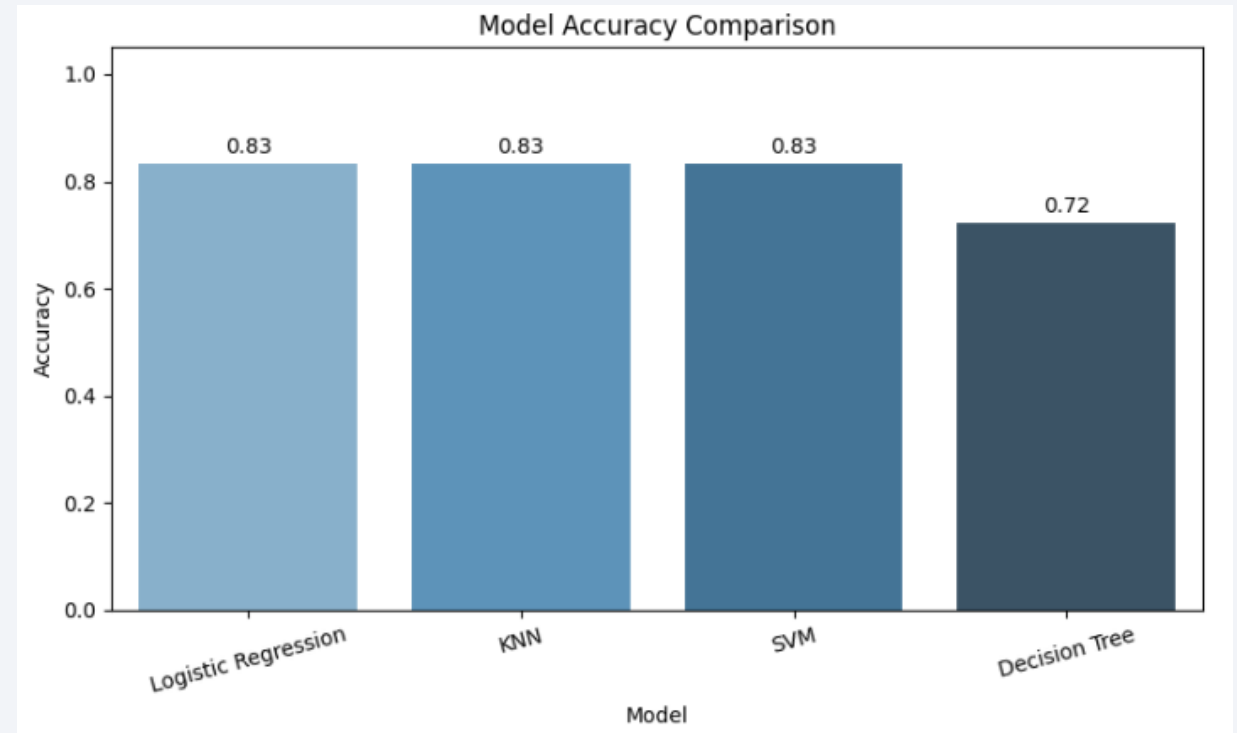


Section 5

Predictive Analysis (Classification)

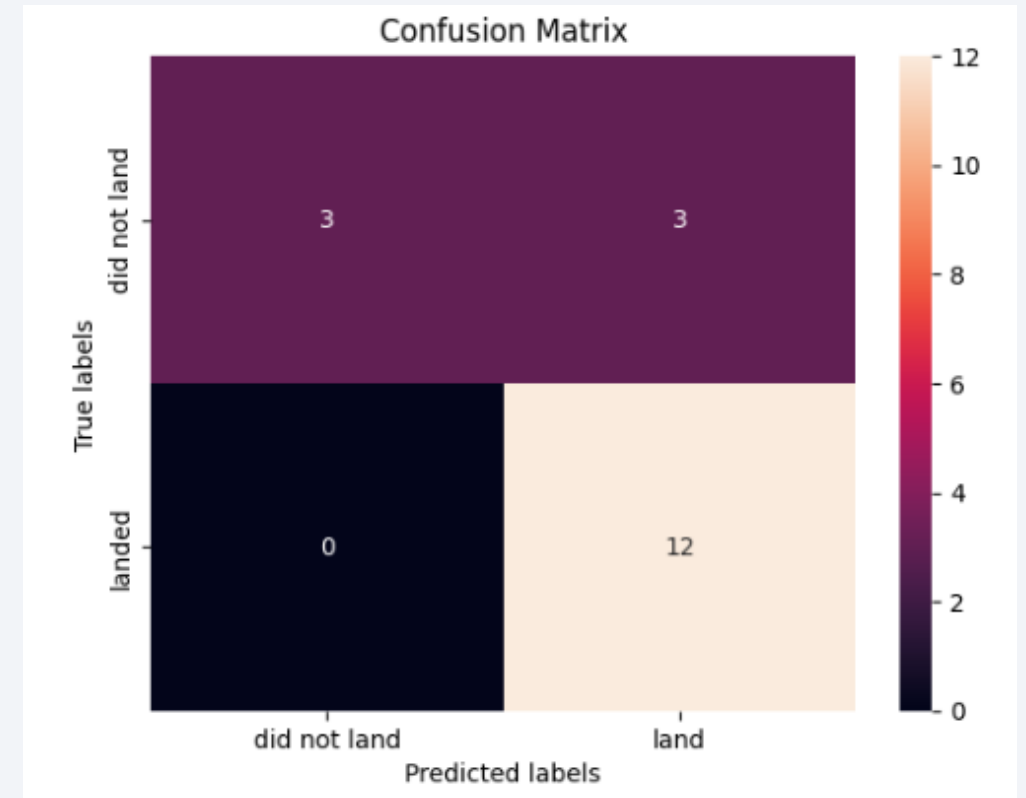
Classification Accuracy

- The Bar chart shows the accuracy for all built classification models
- It seems like that logistic regression ,KNN, SVM are the winner and have all the same value
- The decision tree has lowest accuracy



Confusion Matrix

- That is the confusion matrix of the best performing model. Its equal for KNN, SVM and logistic regression
- It shows 12 True positives and 3 false negatives














Conclusions

- Success rate of launches is increasing over the years
- As higher the payload mass as better success rate for the landing
- Orbits with highest success rate are the following SE-L1, GEO, HEO, SSO
- KSC LC-39A is the most successful Launch site
- Payload between 2600 – 4000 seems to be the most successful payload range
- The “FT” Booster Version seems to be the most successful Booster Version

Appendix – Github Infos

- That are the uploaded files on the Github repository for this capstone project
- Github URL:
<https://github.com/NewFishMH/BM-Data-Science-Professional-Training>

 DS_Capstone_MH.pdf
 Lab_1_Data_collection_api.ipynb
 Lab_2_Web scraping.ipynb
 Lab_3_Data_wrangling.ipynb
 Lab_4_Eda_sql_coursera_sqlite.ipynb
 Lab_5_Eda_data_viz.ipynb
 Lab_6_Launch_site_location.ipynb
 Lab_7_Machine_Learning_Prediction.ipynb
 Lab_8_Screenshot_Dashboard.png
 Lab_8_SpaceX_dash_app.py
 README.md

Thank you!

