IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mehrasa Modanlou Jouybari
10.04.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection, data wrangling, exploratory data analysis with data visualization, exploratory data analysis with SQL, building an interactive
    map with Folium, building a dashboard with plotly dash, predictive analysis
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX is a aerospace manufacturer, space transportation services and communications corporation which is a disruptive just like Tesla both founded by Elon Musk. Despite being less than 20 years old SpaceX has managed to reduce the cost by more than 50% compared to other company and said to reduce by 99% when the Starship project will be completed.
  - This is because spaceX has developed Technology to land the first stage booster which is 70% the cost of the rocket. By landing it safely they are able to reuse the booster and the cost of the launches. Using their reused boosters cost 50% less of their their cost to use a new booster and this has made spaceX the company of that dominate the market. SpaceX has gained worldwide attention for a series of historic milestones.
  - It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras 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 SpaceX for a rocket launch.
- Problems you want to find answers
  - How does the success of the first stage landing depend on factors such cargo mass, launch site, number of flights, and orbits?
  - Does the frequency of successful landings rise with time?
  - What is the most effective method that can be applied in this situation for binary classification?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Through SpaceX's API
  - Through scraping Wikipedia
- Perform data wrangling
  - Filtering, handling missing values, using one hot encoding to prepare the data to a binary classification
- 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

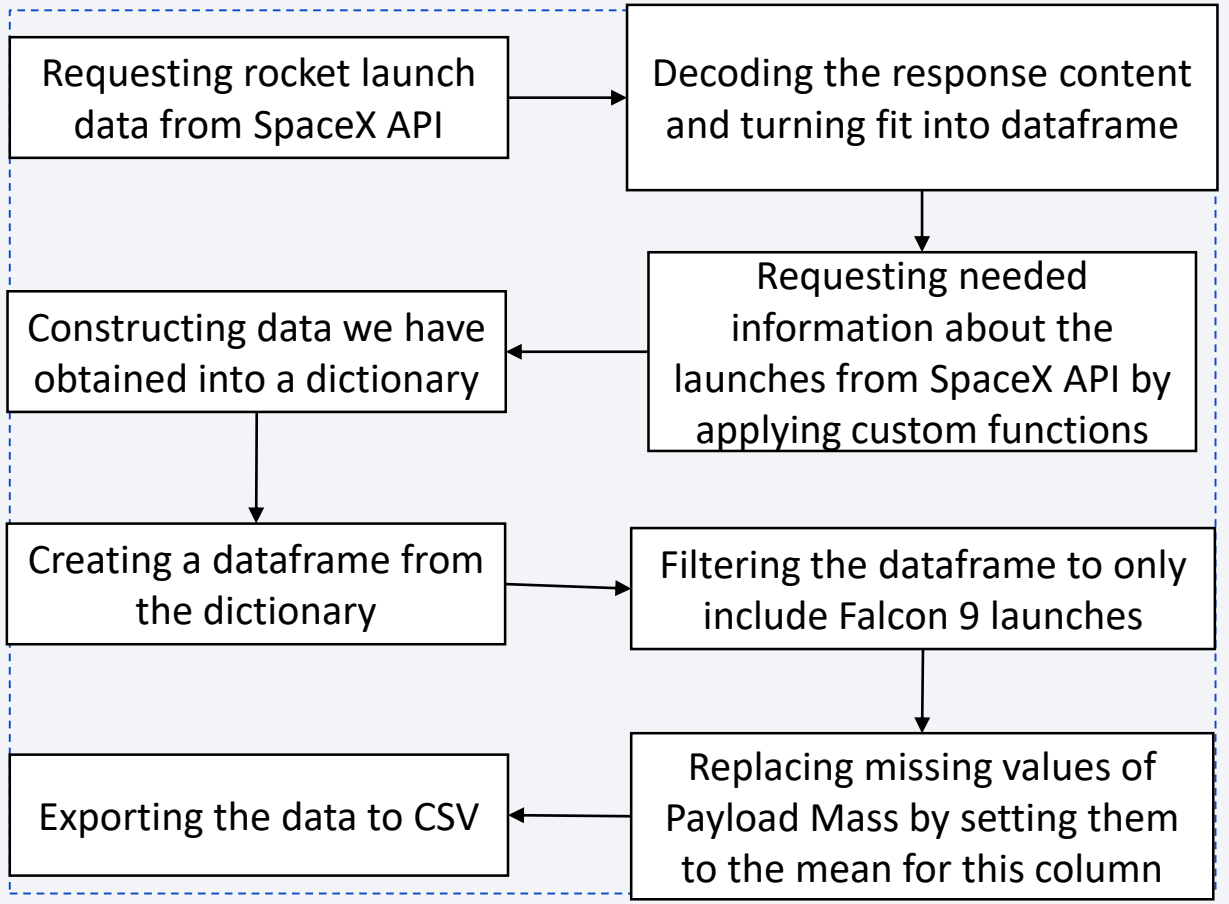
Data Collection

- Data was gathered using a combination of API queries to SpaceX's REST API and web scraping of information from a table in the Wikipedia entry for the company. In order to obtain comprehensive data about the launches for a more in-depth analysis, we had to use both of these data collection techniques.

Data Collection - SpaceX API

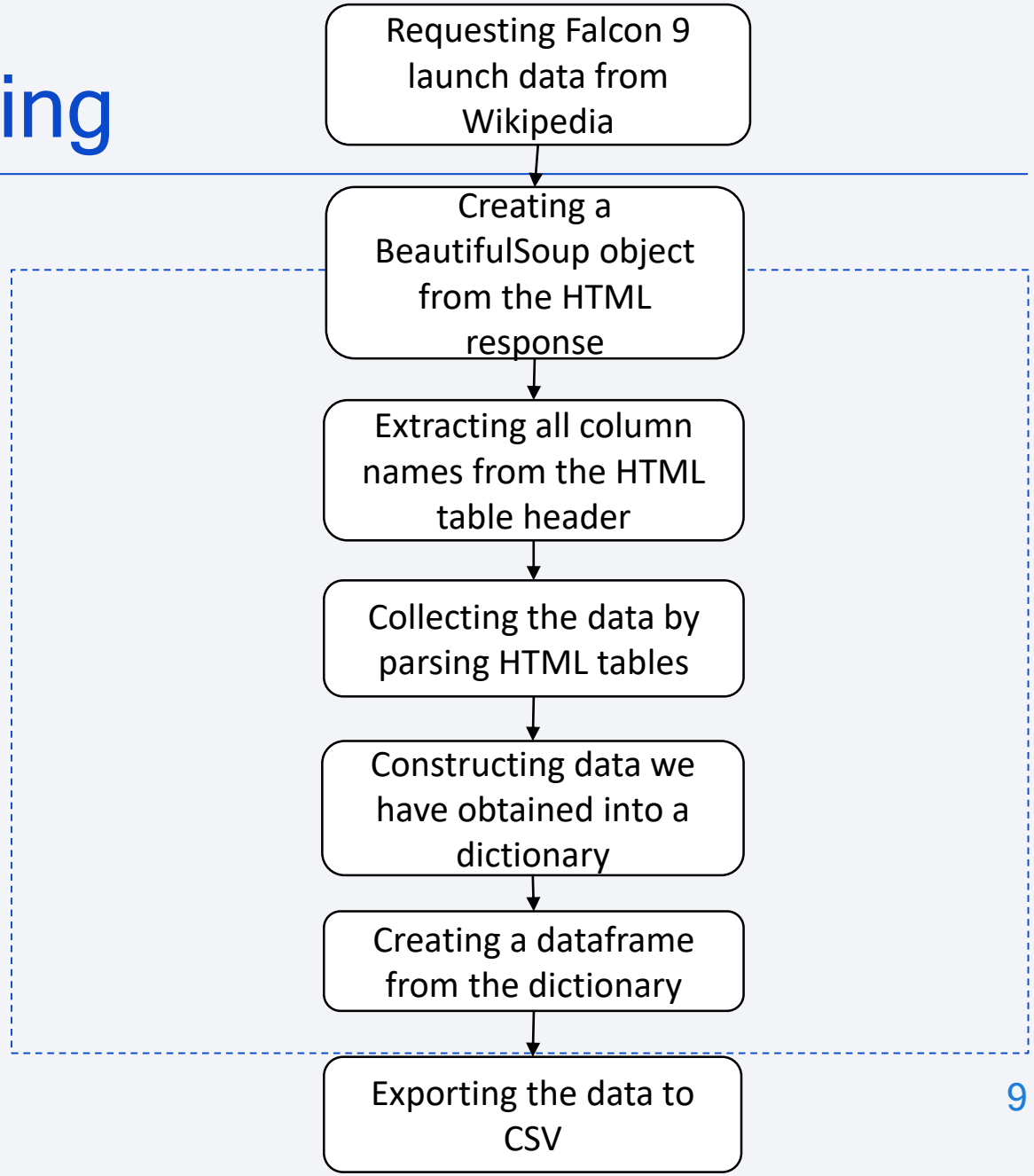
- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/MehrasaModanlou/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb -



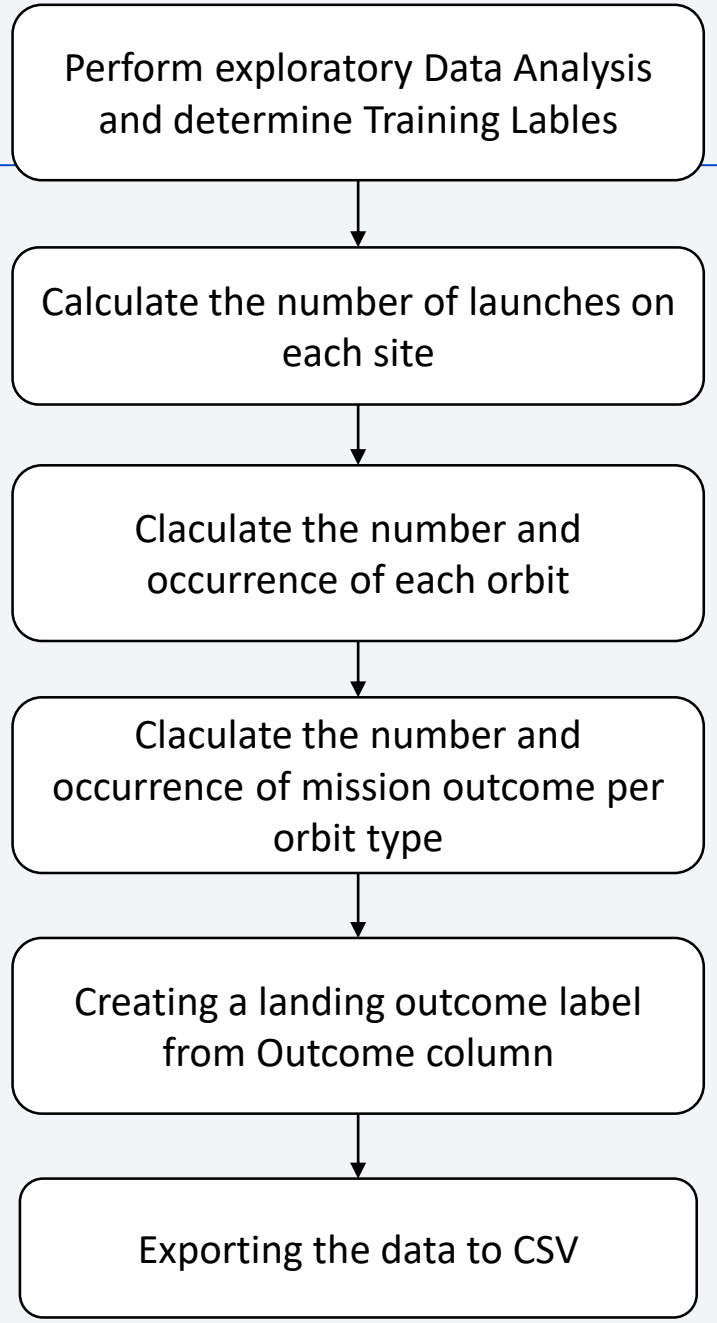
Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- https://github.com/MehrasaModanlou/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb -



Data Wrangling

- After collecting the data we check the missing data and data types, and do one of the following to clean the data:
  - Deal with Null values: Replace the missing data with one-Using mean, median or so.
  - Change data type of the data.
  - Represent categorical data using integer or float dummy numbers – one hot encoding
  - split data into train and test set
  - https://github.com/MehrasaModanlou/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb -



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
  - The relationship between variables is displayed through scatter plots. If a connection can be made, a machine learning model could use it. Bar graphs display comparisons between distinct categories. The objective is to demonstrate the connection between a measured value and the particular categories that are being compared.
    Line diagrams display data trends over time. (time series).
  - https://github.com/MehrasaModanlou/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb -

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string ‘CCA‘
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking 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
- https://github.com/MehrasaModanlou/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb -

Build an Interactive Map with Folium

- Markers of all Launch Sites:
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Coloured Markers of the launch outcomes for each Launch Site:
  - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
  - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.
- https://github.com/MehrasaModanlou/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb -

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection
- Added a pie chart to show the total successful launches count for all sites and the success vs failed counts for the site if a specific launch site was selected
- Added a slider to select payload range
- Added a scatter chart to show the correlation between payload and launch success
- https://github.com/MehrasaModanlou/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py -

Predictive Analysis (Classification)

1. Creating a NymPy array from the column Class in data
2. Standardizing the data with StandardScaler then fitting and transforming it
3. Splitting the data into training and testing sets with train_test_split function
4. Creating a GridSearchCV object with cv=10 to find the best parameters
5. Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
6. Calculating the accuracy on the test data using the method .score() for all models
7. Examining the confusion matrix for all models
8. Finding which method performs best by examining the Jaccard_score and F1_score metrics

- https://github.com/MehrasaModanlou/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb -

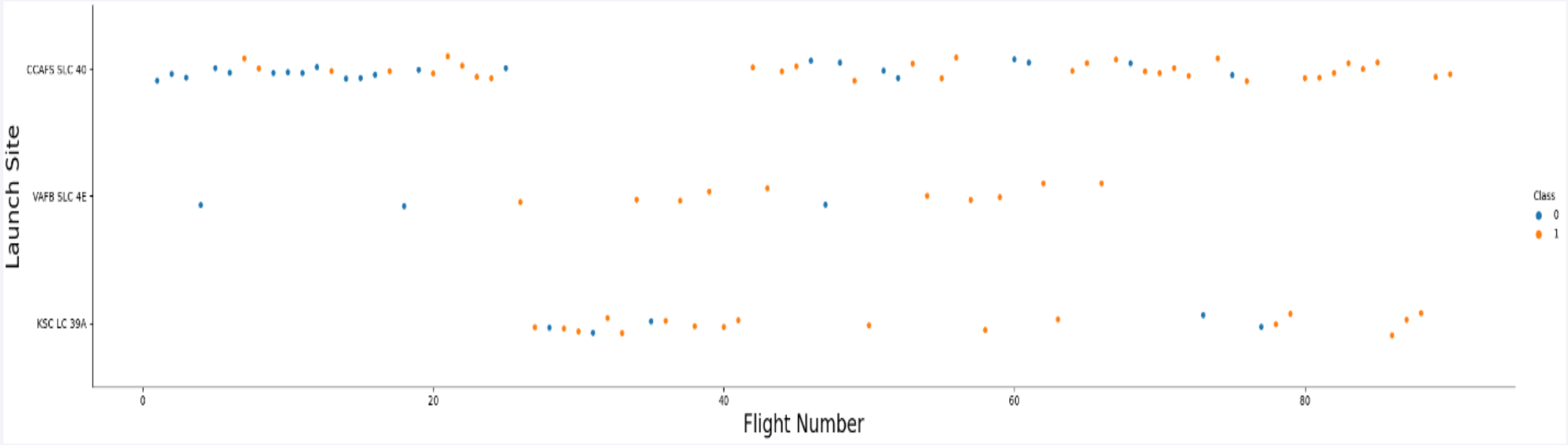
Results

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

Section 2

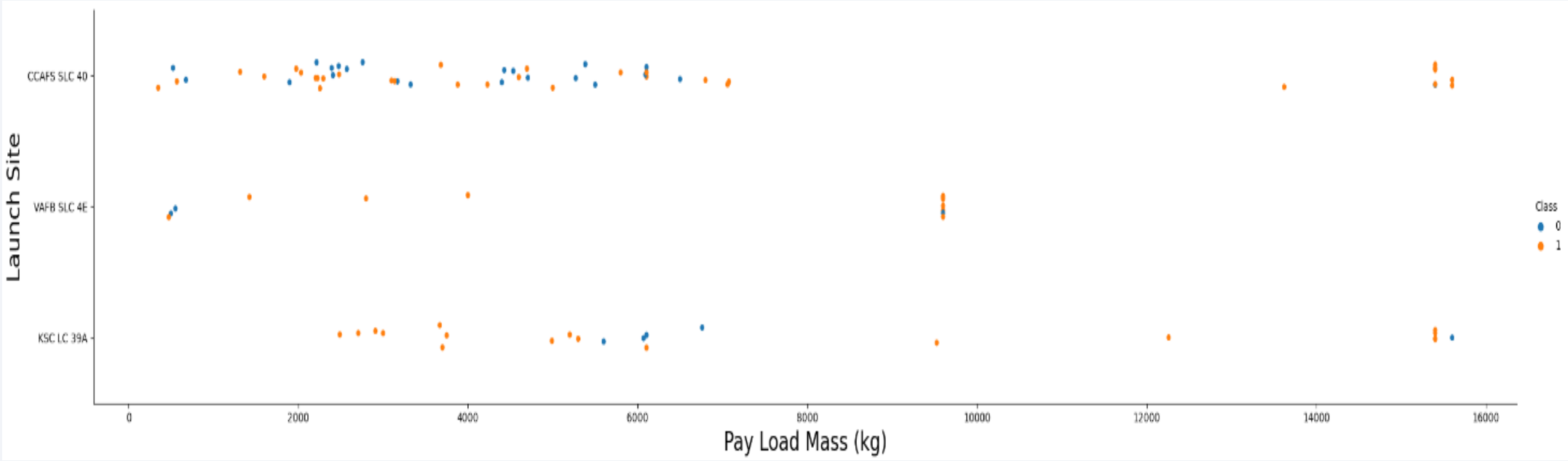
Insights drawn from EDA

Flight Number vs. Launch Site



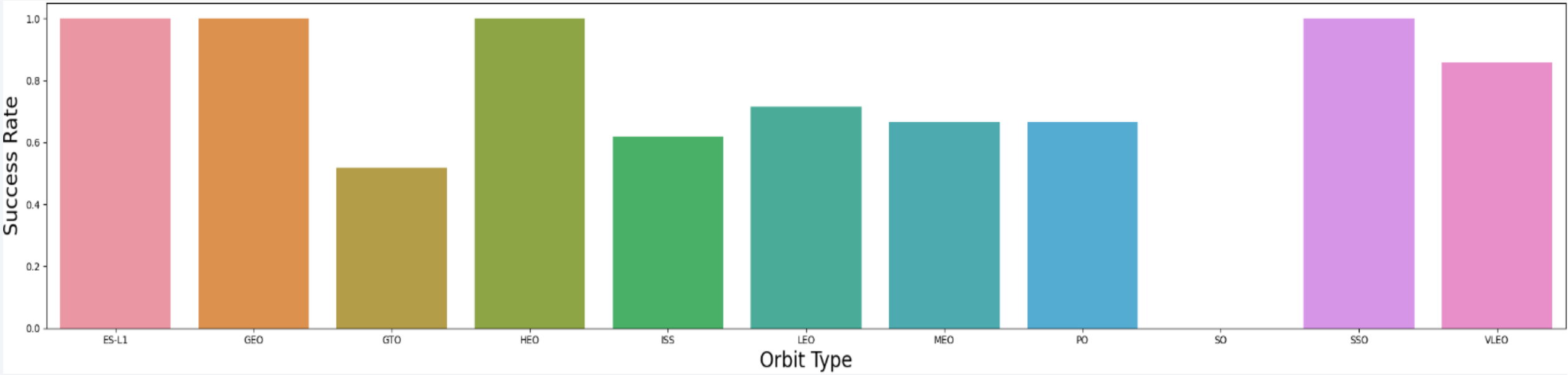
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success

Payload vs. Launch Site



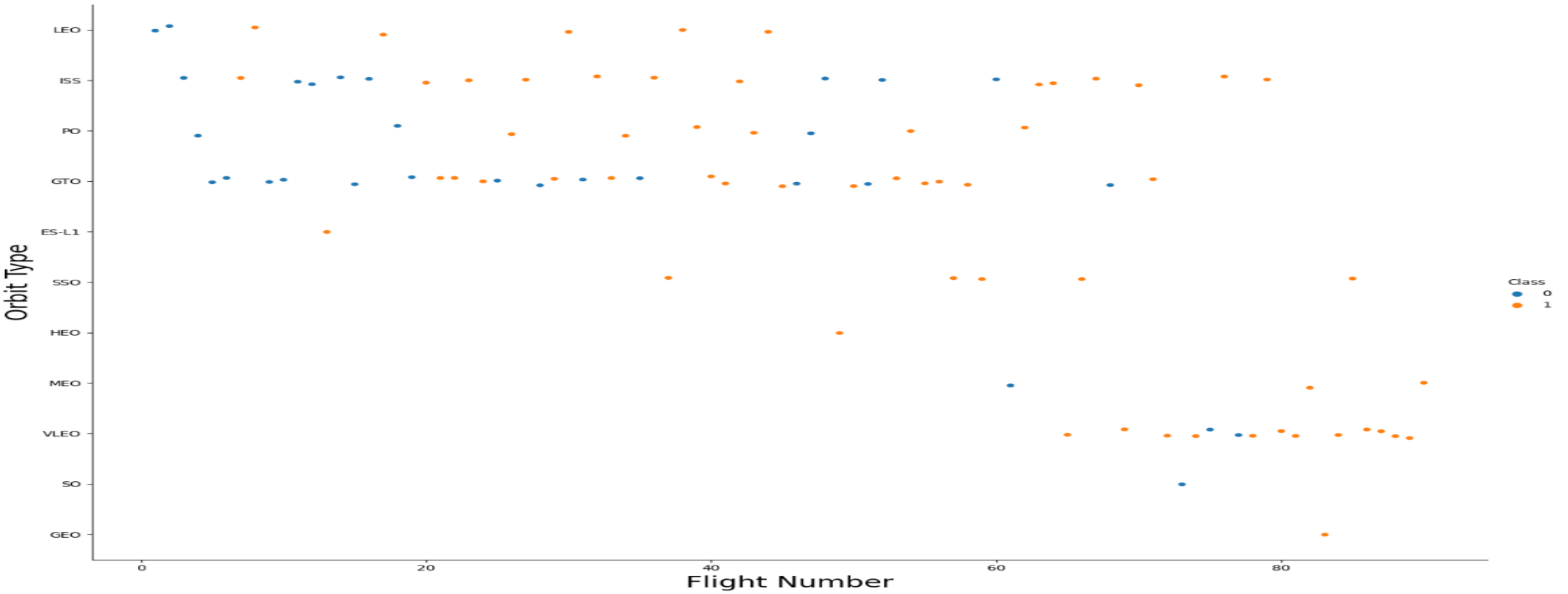
- For every launch site the higher the payload mass, the higher the success rate
- Most of the launches with payload mass over 7000 kg were successful
- KSC LC 39A has a 100% success rate for payload mass under 5500kg too

Success Rate vs. Orbit Type



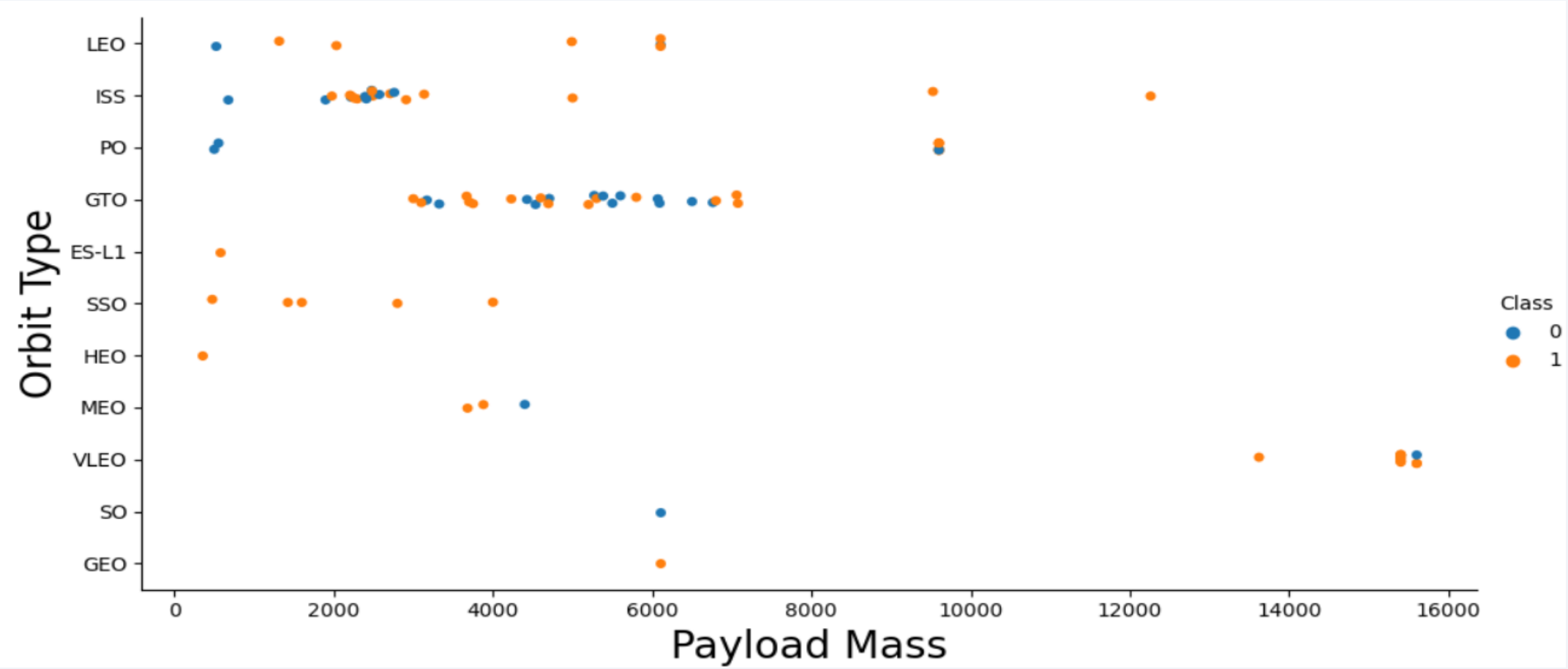
- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: SO
- Orbits with success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO

Flight Number vs. Orbit Type



- In the LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

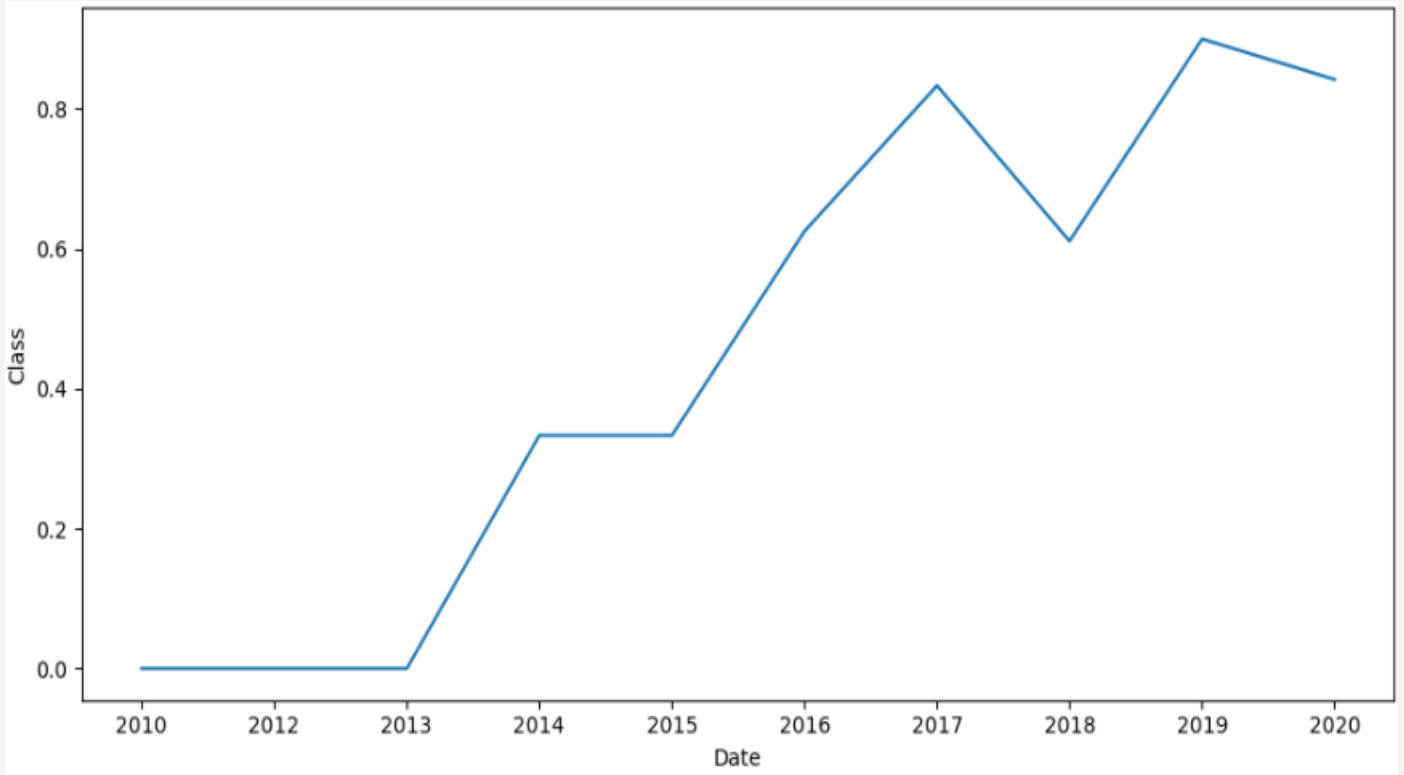
Payload vs. Orbit Type



- Heavy payloads have a negative influence on GTO orbits and positive influence o Polar LEO (ISS) orbits

Launch Success Yearly Trend

- the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Displaying the names of the unique launch sites in the space mission

```
%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'

- Displaying 5 records where launch sites begin with the string 'CCA'.

```
%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 _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Displaying the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS 'TOTAL_PAYLOAD_MASS'
FROM SPACEXTBL
WHERE "Customer" = "NASA (CRS)"
```

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

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

- Displaying average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS 'AVG_PM_F9_v1.1'
FROM SPACEXTBL
WHERE "Booster_Version" = "F9 v1.1"
```

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

AVG_PM_F9_v1.1
2928.4

First Successful Ground Landing Date

- Listing the date when the first succesful landing outcome in ground pad was acheived.

```
q = pd.read_sql("select min(Date) as First_Success_Date from spacexdata where Landing__Outcome='Success (ground pad)'", conn)
q
```

	First_Success_Date
0	2015-12-22 00:00:00

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql SELECT DISTINCT("Booster_Version"), "Landing _Outcome", PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Booster_Version	Landing _Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Listing the total number of successful and failure mission outcomes

```
q = pd.read_sql("select substr(Mission_Outcome,1,7) as Mission_Outcome, count(*) from spacexdata  group by 1", conn)
q
```

	Mission_Outcome	count(*)
0	Failure	1
1	Success	100

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass

```
%%sql
SELECT DISTINCT("Booster_Version"), PAYLOAD_MASS__KG_ AS MAX_PAYLOAD_MASS
FROM SPACEXTBL
WHERE MAX_PAYLOAD_MASS = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Booster_Version	MAX_PAYLOAD_MASS
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
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

```
%%sql
SELECT substr(Date,7,4) AS Year,
substr(Date, 4, 2) AS Month,
"Date",
"Landing _Outcome",
"Booster_Version",
"Launch_Site"
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE "%Failure (drone ship)%" AND Year='2015'
```

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

Year	Month	Date	Landing _Outcome	Booster_Version	Launch_Site
2015	01	10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	14-04-2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Ranking 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

```
q = pd.read_sql("select Landing__Outcome, Date, count(*) as Count from spacexdata where Landing__Outcome like '%Success%' and Date between '2010-06-04
q
```

	Landing__Outcome	Date	Count
0	Success (drone ship)	2016-04-08 00:00:00	5
1	Success (ground pad)	2015-12-22 00:00:00	3

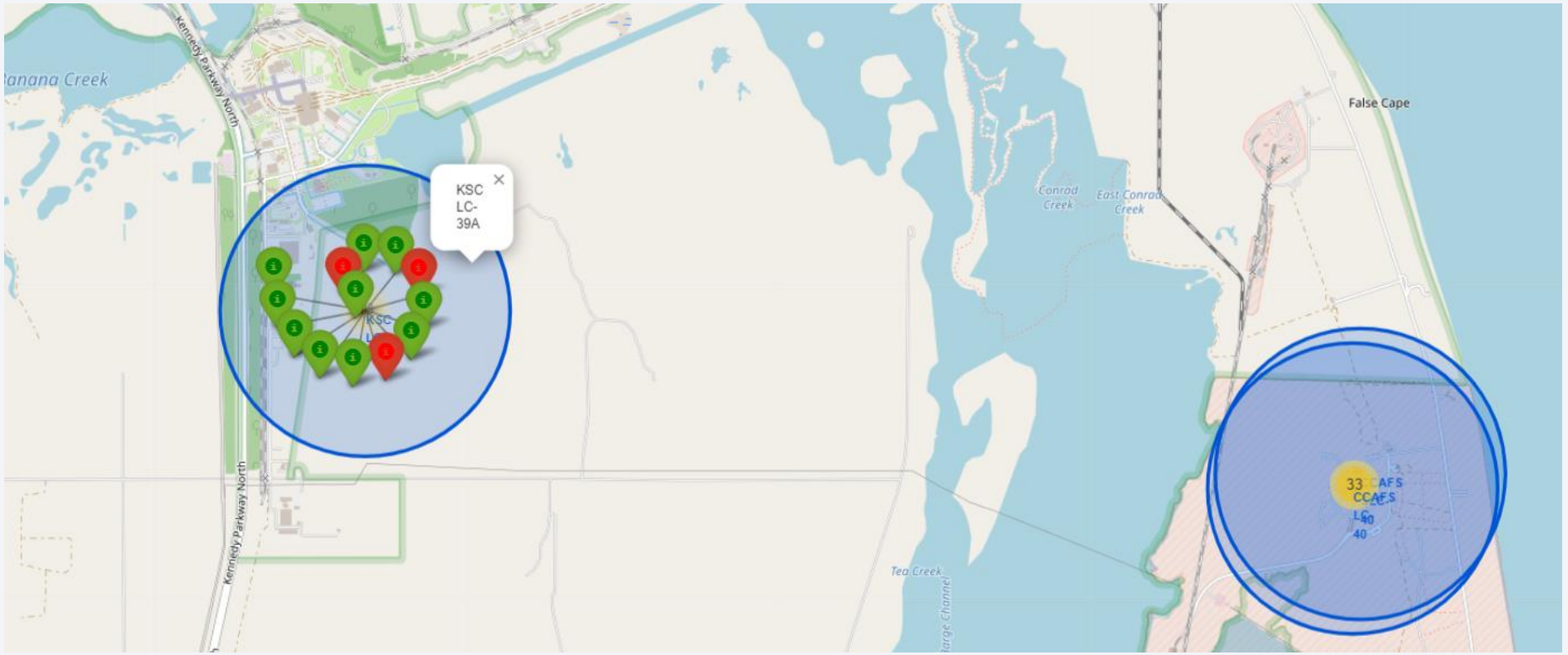

Section 3
Launch Sites
Proximities Analysis

All launch sites' location markers on a global map

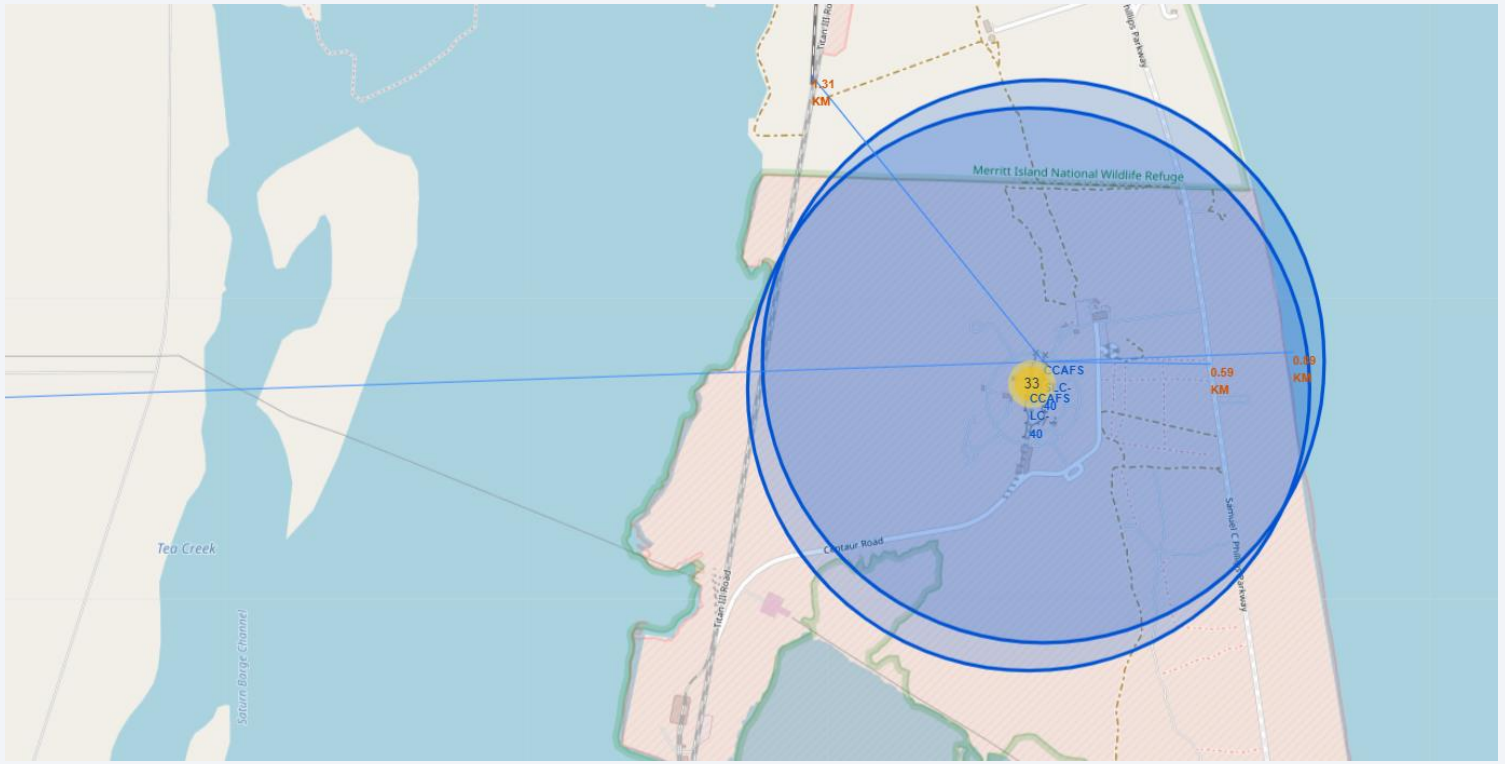
- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.

Colour-labeled launch records on the map

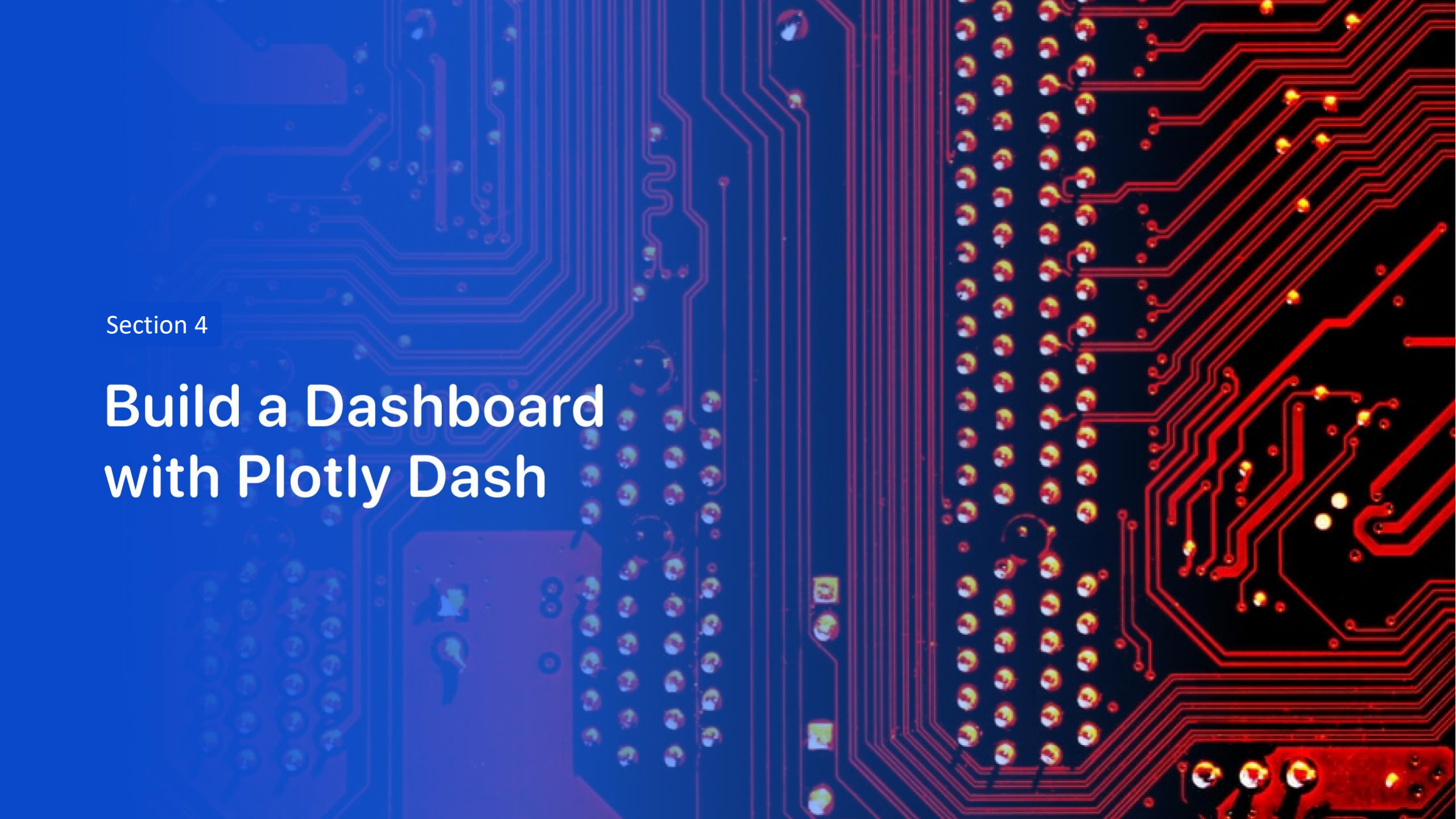
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates. Green Marker = Successful Launch, Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.



Distance from the launch site CCAFS SLC-40 to its proximities

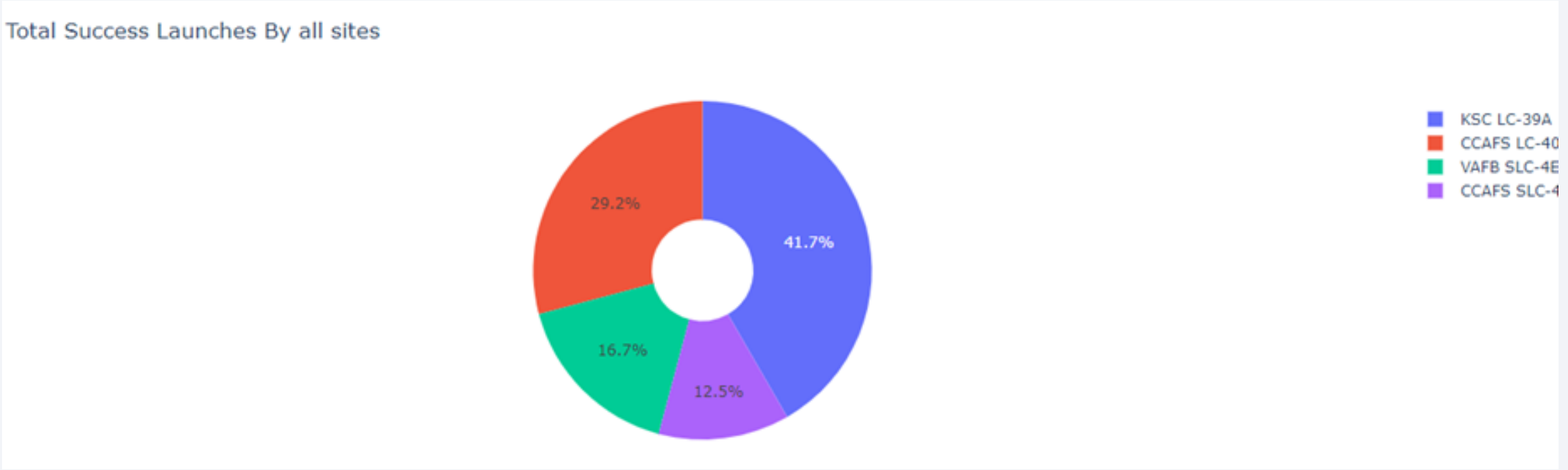


- From the visual analysis of the launch site CCAFS SLC-40 we can clearly see that it is: relative close to railway (1.31 KM), relative close to highway (0.59 KM), relative close to coastline (0.89 KM).
- Also the launch site CCAFS SLC-40 is relative far from its closest city Orlando (78.63 KM).


Section 4
Build a Dashboard with Plotly Dash

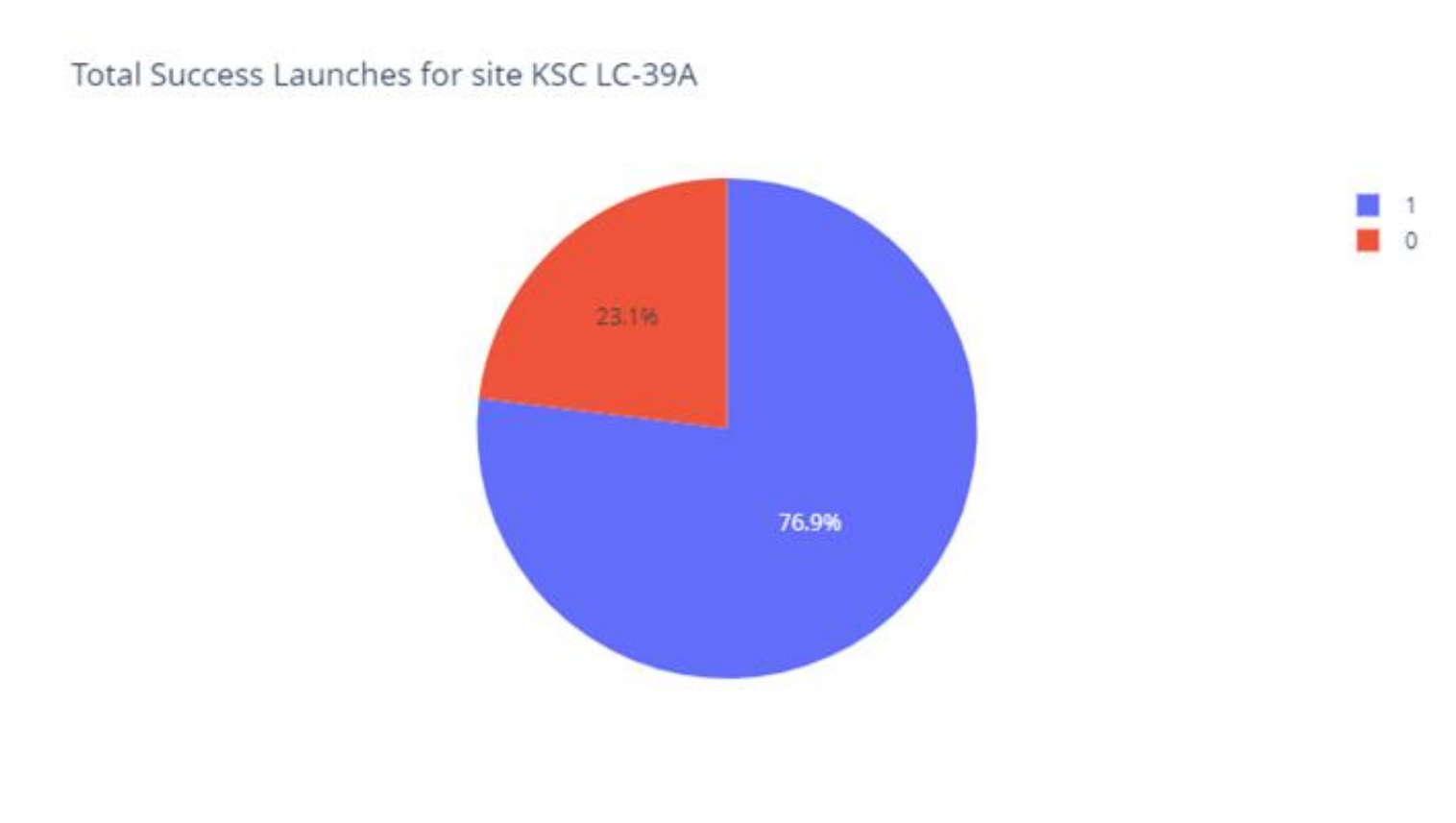
Launch success count for all sites

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.



Launch site with highest launch success ratio

- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



Payload Mass vs. Launch Outcome for all sites

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.
- Among F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.), FT has the highest launch success rate with 15 successes and 8 failures.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- KNN, Logistic Regression and SVM obtained superior results against Decision Tree.
- The best accuracy score was about 0.94%

Find the method performs best:

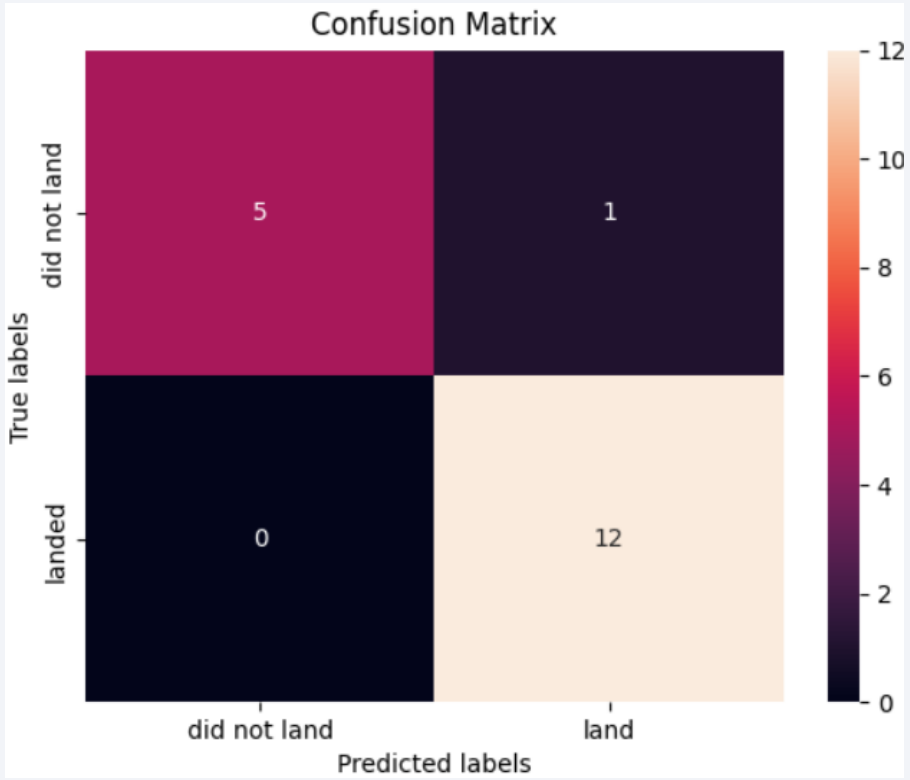
```
# We can dermine the best method/classifier by the highest performance measure results on test(unseen)data

print("score on GridSearchCv with Logistic Regression: ", logreg_cv.score(x_test, y_test))
print("score on GridSearchCv with Support Vector Machine: ", svm_cv.score(x_test, y_test))
print("score on GridSearchCv with  Decision Tree Classifier: ", tree_cv.score(x_test, y_test))
print("score on GridSearchCv with K-Nearest Neighbors: ", knn_cv.score(x_test, y_test))
```

```
score on GridSearchCv with Logistic Regression:  0.9444444444444444
score on GridSearchCv with Support Vector Machine:  0.9444444444444444
score on GridSearchCv with  Decision Tree Classifier:  0.8333333333333334
score on GridSearchCv with K-Nearest Neighbors:  0.9444444444444444
```

Confusion Matrix

- Examining the confusion matrix, we see that Logistic Regression, Support Vector Machine, K-Nearest Neighbors can distinguish between the different classes. We see that the major problem is false positive.



Conclusions

- Logistic Regression, Support Vector Machine, K-Nearest Neighbors Models are the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!