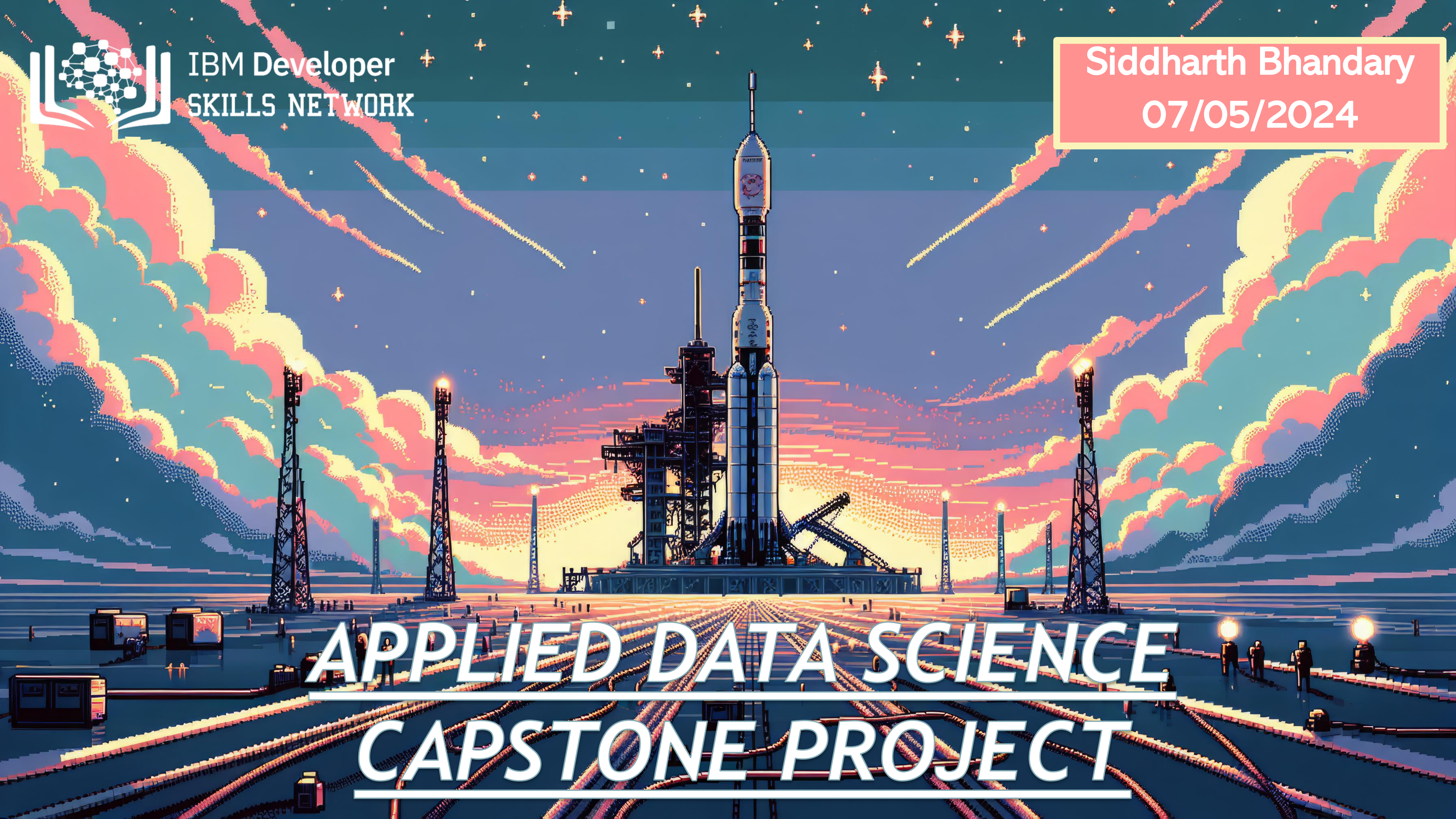


IBM Developer
SKILLS NETWORK

Siddharth Bhandary

07/05/2024

APPLIED DATA SCIENCE CAPSTONE PROJECT

Outline

Executive Summary
Introduction
Methodology
Results
Conclusion
Appendix

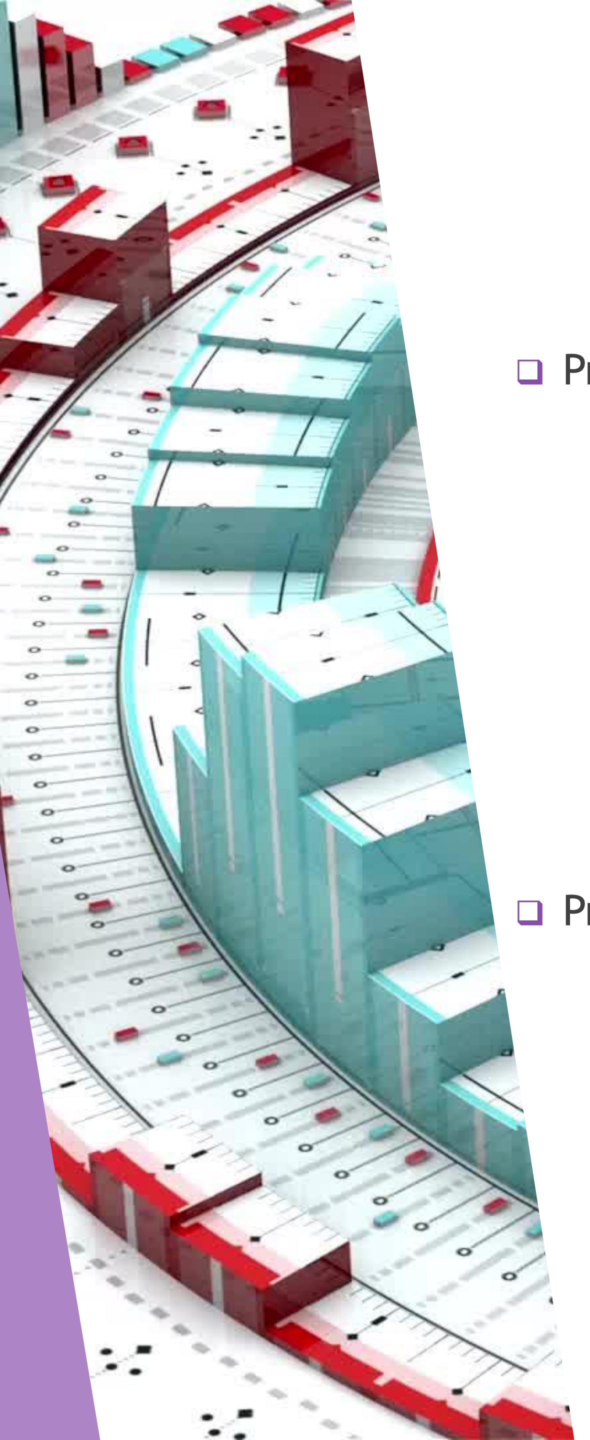
Executive Summary

Summary of methodologies –

- Data Collection
- Data Wrangling
- EDA with Data Visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a dashboard with Plotly
- Predictive analysis (Classification)

Summary of all results –

- EDA with results
- Interactive analytics demo in screenshots
- Predictive analysis results



Introduction

- Project background and context
  - Space X promotes Falcon 9 rocket flights for a price of 62 million dollars; in comparison, other suppliers charge up to 165 million dollars each launch; a large portion of the cost reductions are attributable to Space X's ability to reuse the first stage.
  - Thus, we can calculate the cost of a launch if we can ascertain if the first stage will land. This data may be utilized should another business choose to submit a bid for a rocket launch against Space X. The project's objective is to build a machine learning pipeline that can forecast whether or not the initial stage will land successfully.
- Problems you want to find answers
  - What factors are able to determine whether their rocket will successfully land?
  - How is the success of landing a rocket correlated to the year it launches?
  - What is the common denominator for failed launch attempts?

SECTION 1

- METHODOLOGY

Methodology

Executive Summary

Data collection methodology:	Using SpaceX Rest API Using Web Scrapping from Wikipedia
Performed data wrangling	Filtering the data
Dealing with missing values	Using One Hot Encoding to prepare the data to a binary classification
Performed exploratory data analysis (EDA) using visualization and SQL Performed interactive visual analytics using Folium and Plotly Dash	
Performed predictive analysis using classification models	Building, tuning and evaluation of classification models to ensure the best results

Data Collection

Data collection was done using get request to the SpaceX API.

Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().

We then cleaned the data, checked for missing values and fill in missing values where necessary.

In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.

The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API : Github Link

Requesting rocket launch data from SpaceX API

Decoding the response content using .json() and turning it into a dataframe using .json_normalize()

Requesting needed information about the launches from SpaceX API by applying custom functions

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Filtering the dataframe to only include Falcon 9 launches

Replacing missing values of Payload Mass column with calculated .mean() for this column

Exporting the data to CSV

Data Collection – Web scraping : Github Link

Requesting Falcon 9 launch data from Wikipedia

Creating a BeautifulSoup object from the HTML response

Extracting all column names from the HTML table header

Collecting the data by parsing HTML tables

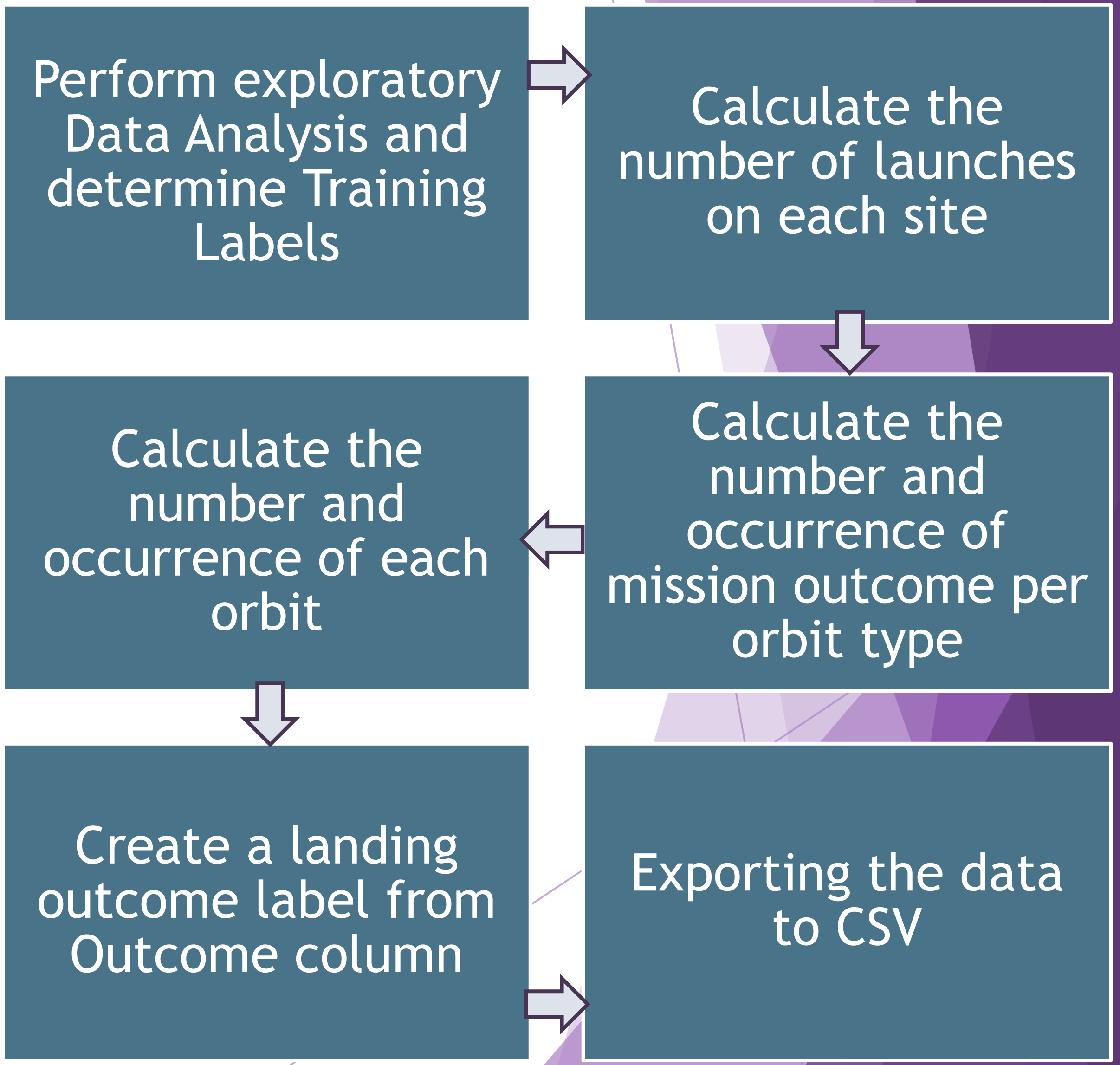
Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Exporting the data to CSV

Data Wrangling : Github Link

- There are other instances in the data set where the booster failed to land properly. Occasionally, an effort at landing is made but the landing is unsuccessful due to an accident; for instance, True Ocean indicates that the mission's outcome was successfully landed in a certain area of the ocean, whereas False Ocean indicates that the mission's outcome was unsuccessfully landed in a particular area of the ocean.
- True RTLS indicates that the mission's result was a successful landing on a ground pad. A mission result of a failed landing on a ground pad is indicated by a false RTLS.A successful landing of the mission outcome on a drone ship is referred to as a true ASDS. False ASDS indicates that the mission's outcome—a drone ship landing—was not accomplished.
- Mostly, we translate those results into Training Labels, where "0" denotes a failed landing and "1" indicates a successful booster landing.

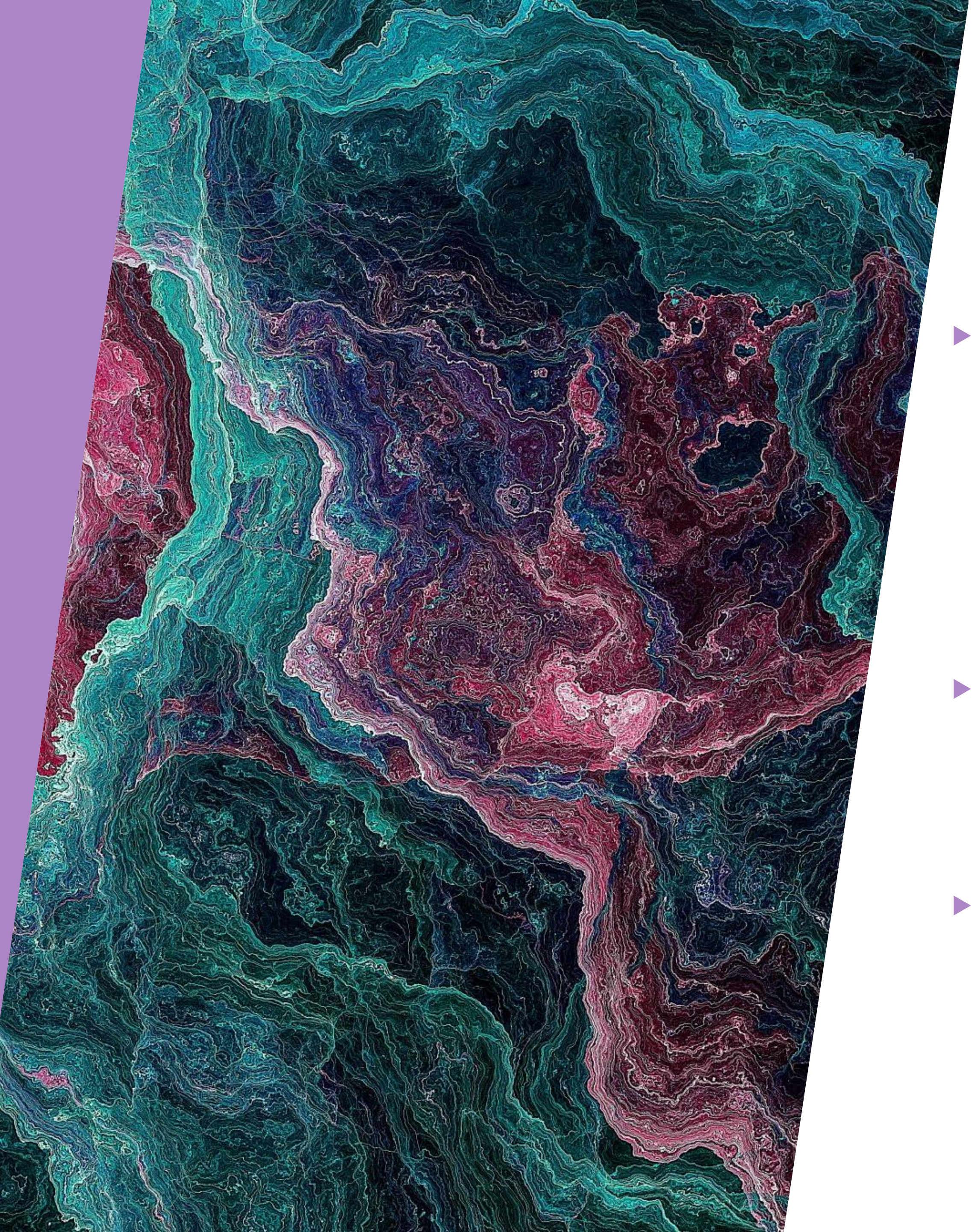


EDA with Data Visualization : [Github Link]

- Charts plotted -
  1. Flight Number vs. Payload Mass
  2. Flight Number vs. Launch Site
  3. Payload Mass vs. Launch Site
  4. Orbit Type vs. Success Rate
  5. Flight Number vs. Orbit Type
  6. Payload Mass vs Orbit
  7. Success Rate Yearly Trend
- Scatter plots show the relationship between variables. If a relationship exists,they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series).

EDA with SQL : Github Link

- Performed SQL queries:
  - 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



Build an Interactive Map with Folium : Github Link

- 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 anexample) and its proximities like Railway, Highway, Coastline and Closest City.

Build a Dashboard with Plotly Dash : Github Link

	Launch Sites Dropdown List:	Added a dropdown list to enable Launch Site selection.
	Pie Chart showing Success Launches (All Sites/Certain Site):	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.
	Slider of Payload Mass Range:	Added a slider to select Payload range.
	Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:	Added a scatter chart to show the correlation between Payload and Launch Success.

Predictive Analysis (Classification) : Github Link

Creating a NumPy array from the column “Class” in data

→ Standardizing the data with StandardScaler, then fitting and transforming it

→ Splitting the data into training and testing sets with train_test_split function

→ Creating a GridSearchCV object with cv = 10 to find the best parameters

→ Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

→ Calculating the accuracy on the test data using the method .score() for all models

→ Examining the confusion matrix for all models

→ Finding the method performs best by examining the Jaccard_score and F1_score metrics

Results

Exploratory data analysis results

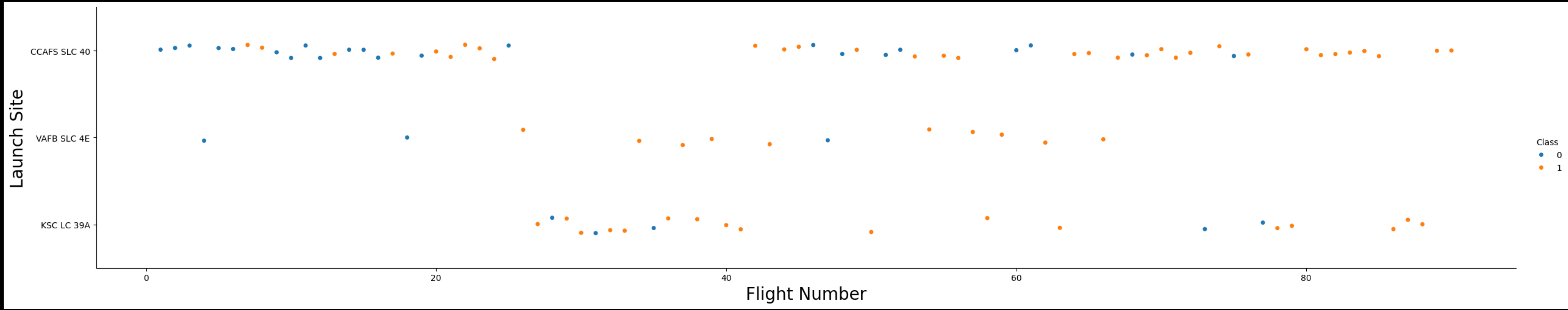
Interactive analytics demo in screenshots

Predictive analysis results

SECTION 2

INSIGHTS DRAWN FROM EDA

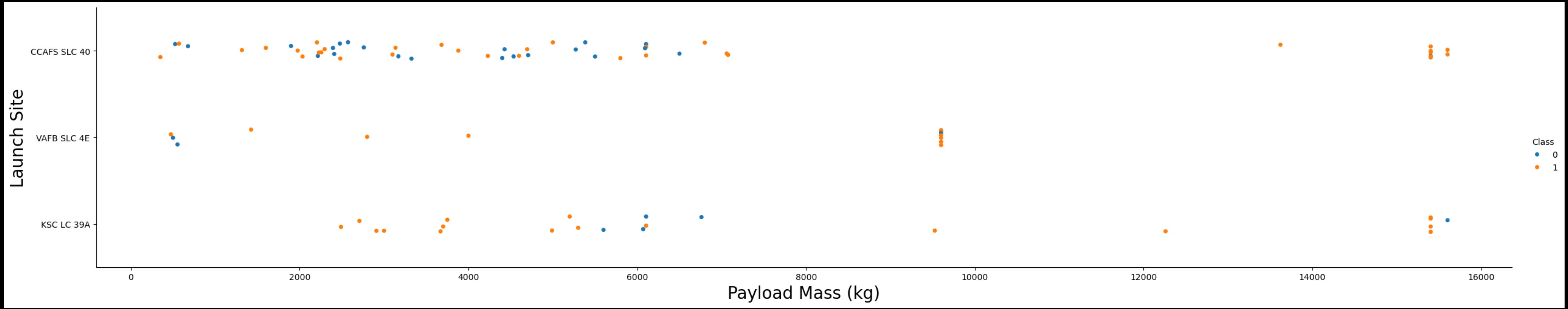
Flight Number vs. Launch Site



Explanation:

- 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



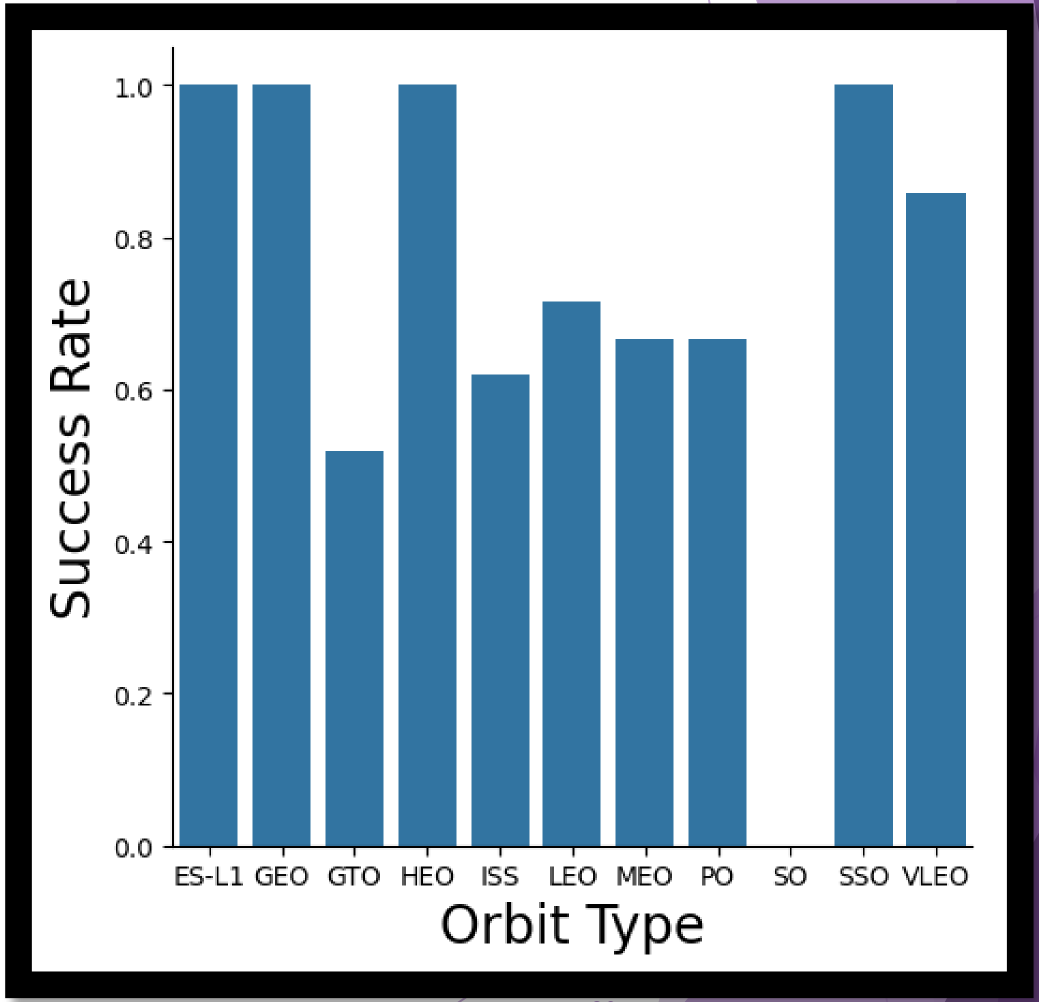
Explanation:

- 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 5500 kg too.

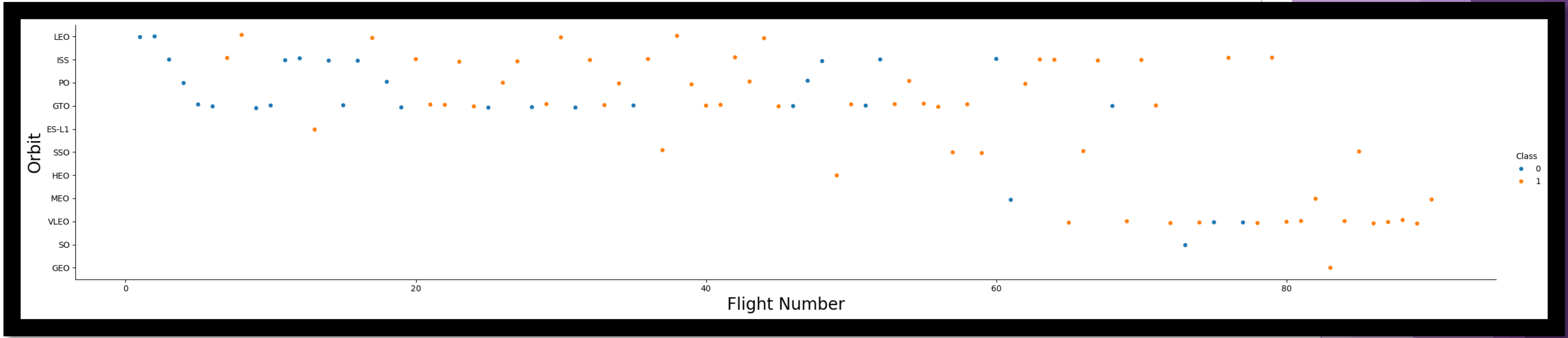
Success Rate vs. Orbit Type

Explanation:

- 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



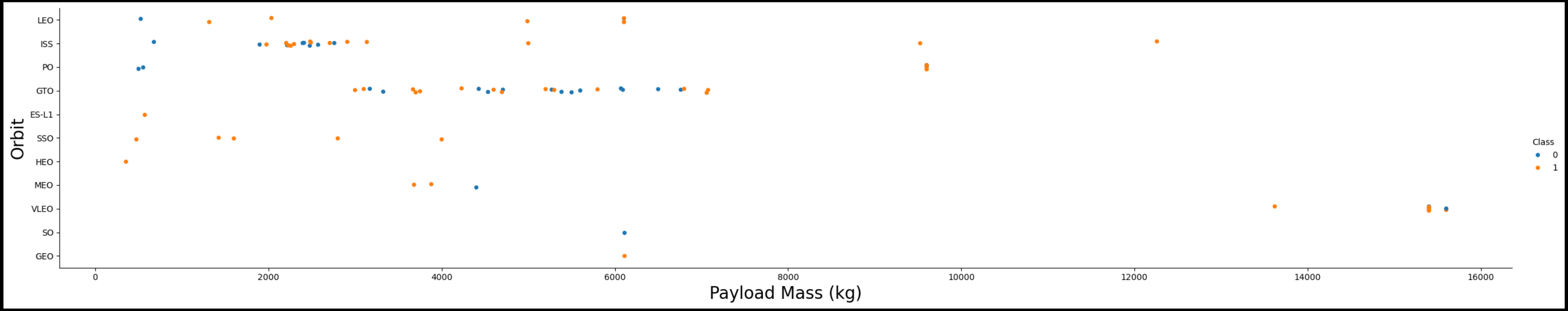
Flight Number vs. Orbit Type



Explanation:

- 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

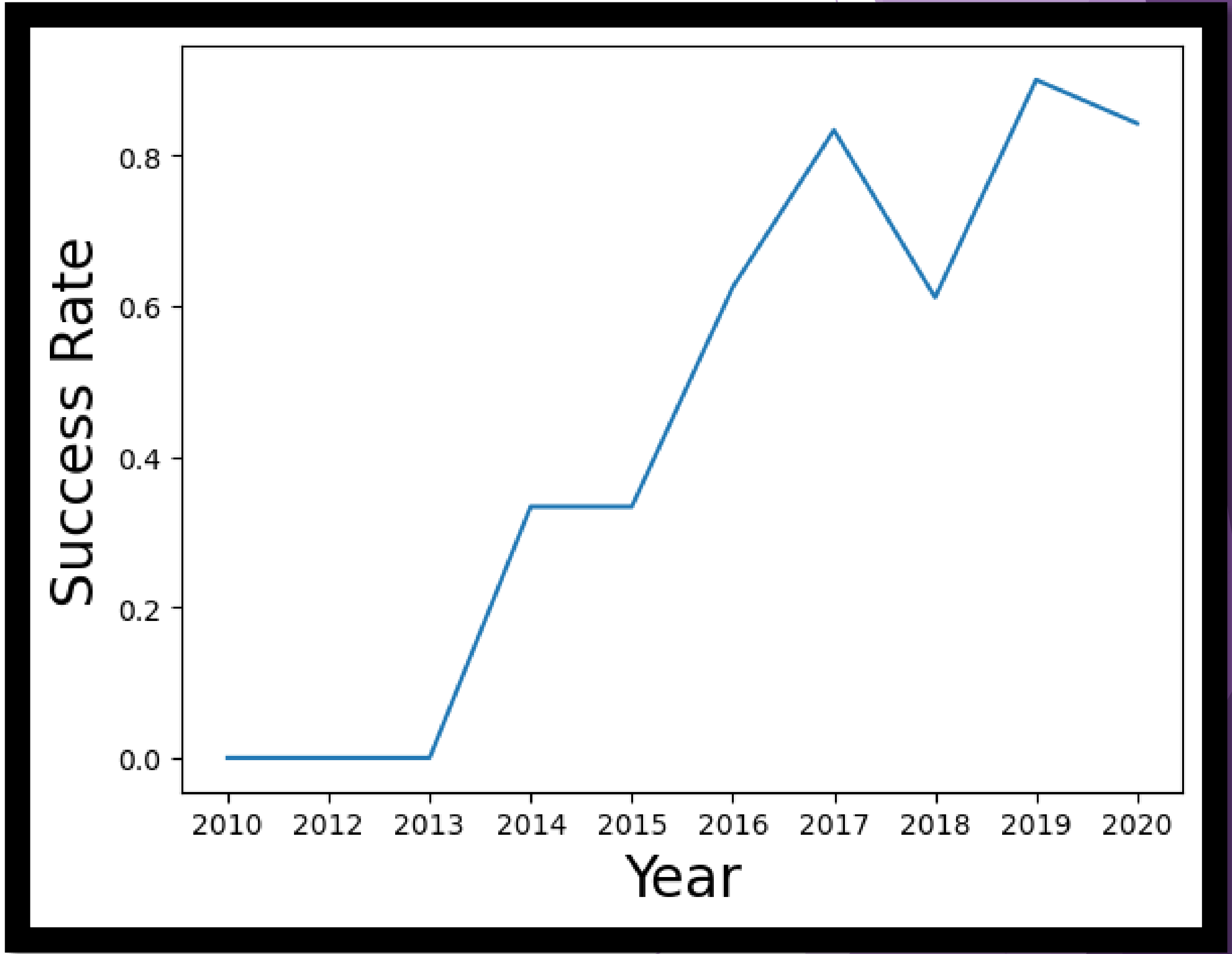


- Explanation:
  - Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

Explanation:

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

- Names of each unique launch site by SpaceX

```
In [16]: %sql select distinct launch_site from SPACEXTABLE;

 * sqlite:///my_data1.db
Done.
Out[16]:
```

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

Launch Site Names Begin with 'CCA'

- 5 records displayed where launch site begins with the string ‘CCA’

In [17]:

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

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

Out[17]:

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

Total Payload Mass

- Total payload mass shown is 45,596 kgs

```
In [18]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';

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

Out[18]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Average payload mass observed by F9 v1.1 is 2534.6666666666665

Display average payload mass carried by booster version F9 v1.1

```
In [19]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';

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

Out[19]:

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- First successful landing observed is 22nd December 2015

```
In [27]: %sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.

Out[27]:
```

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
In [28]: %sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 40

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

Out[28]:

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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes are listed below

```
In [29]: %sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;

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

Out[29]:

Mission_Outcome	total_number
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.

```
In [33]: %sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);

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

Out[33]:

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

- Was unable to get the 2015 launch records. Below is the code snippet and result.

```
In [36]: %%sql select substr(Date, 6,2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
             where landing_outcome = 'Failure (drone ship)' and year(date)=2015;

 * sqlite:///my_data1.db
(sqlite3.OperationalError) wrong number of arguments to function substr()
[SQL: select substr(date) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
      where landing_outcome = 'Failure (drone ship)' and year(date)=2015;]
(Background on this error at: http://sqlalche.me/e/13/e3q8)
```

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.

In [37]:

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

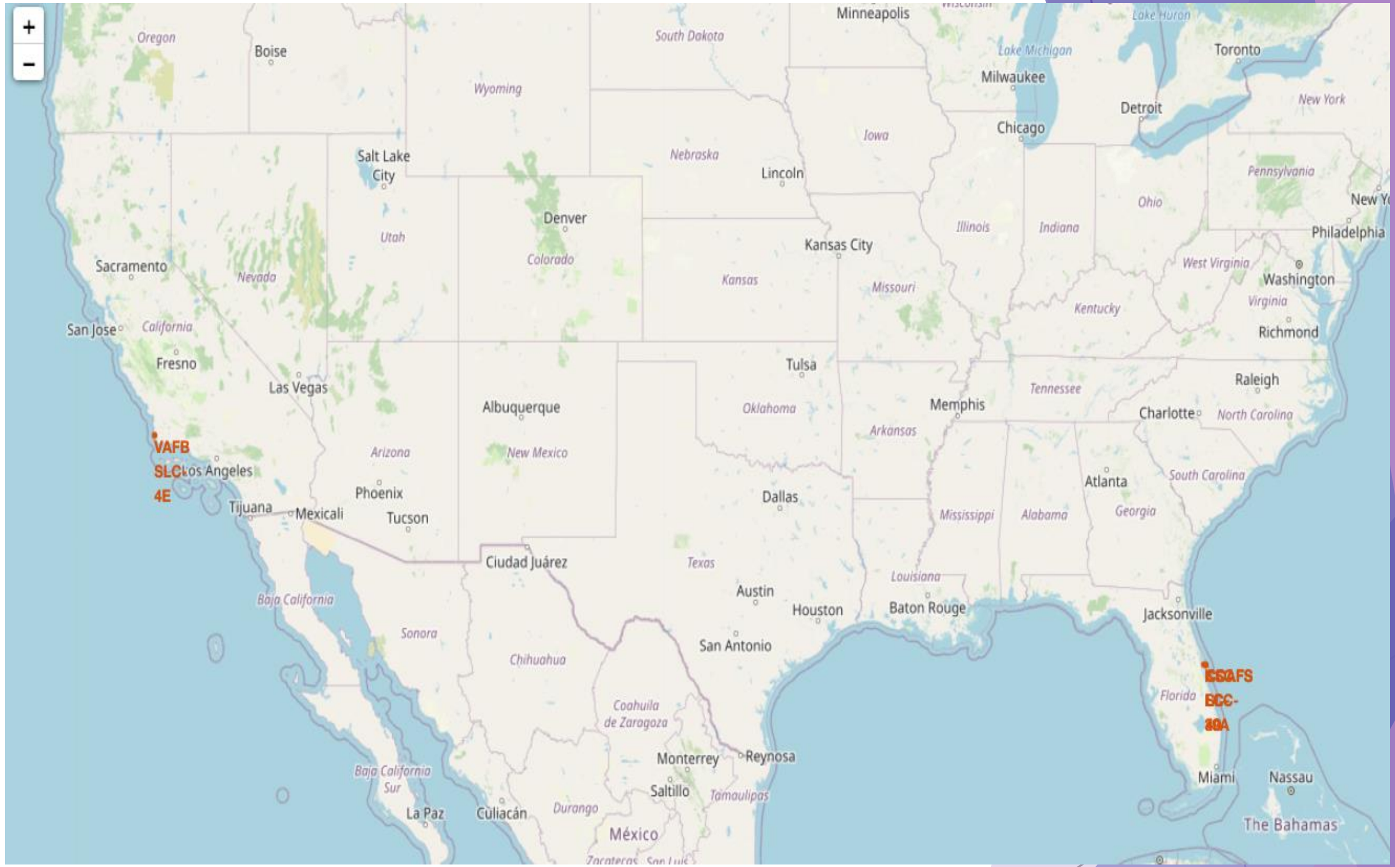
Out[37]:

Landing_Outcome	count_outcomes
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

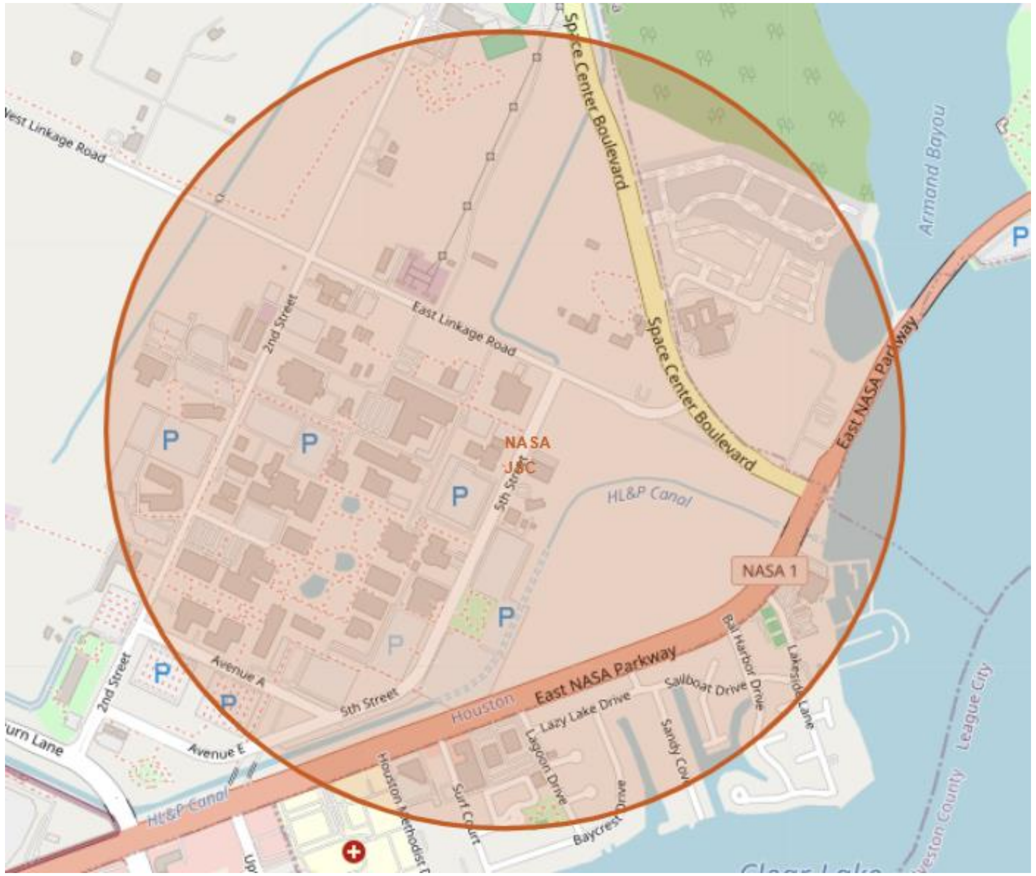
SECTION 3

- LAUNCH SITE PROXIMATE ANALYSIS

All launch sites' location markers on a global map



The majority of launch locations are close to the equator. The equator is the location on Earth's surface where land moves the fastest. At the equator, anything on Earth is already travelling at a speed of 1670 km/h. Launched from the equator, a ship travels upward into space while continuing to orbit the planet at its pre-launch velocity. Inertia is the cause of this. This velocity will assist the spacecraft in maintaining a sufficient speed to remain in orbit.Since all launch sites are located fairly close to the coast, there is little chance that any debris will fall or explode in the vicinity of people when rockets are launched toward the ocean.



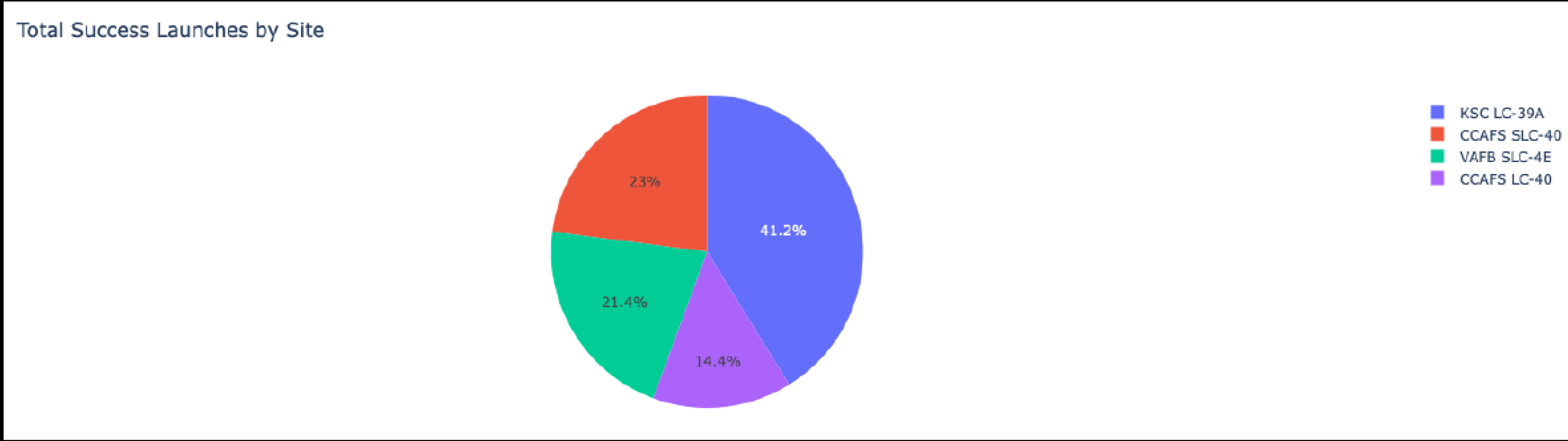
Circled launch area on the map

- This shows the area where the launches have occurred.

SECTION 4

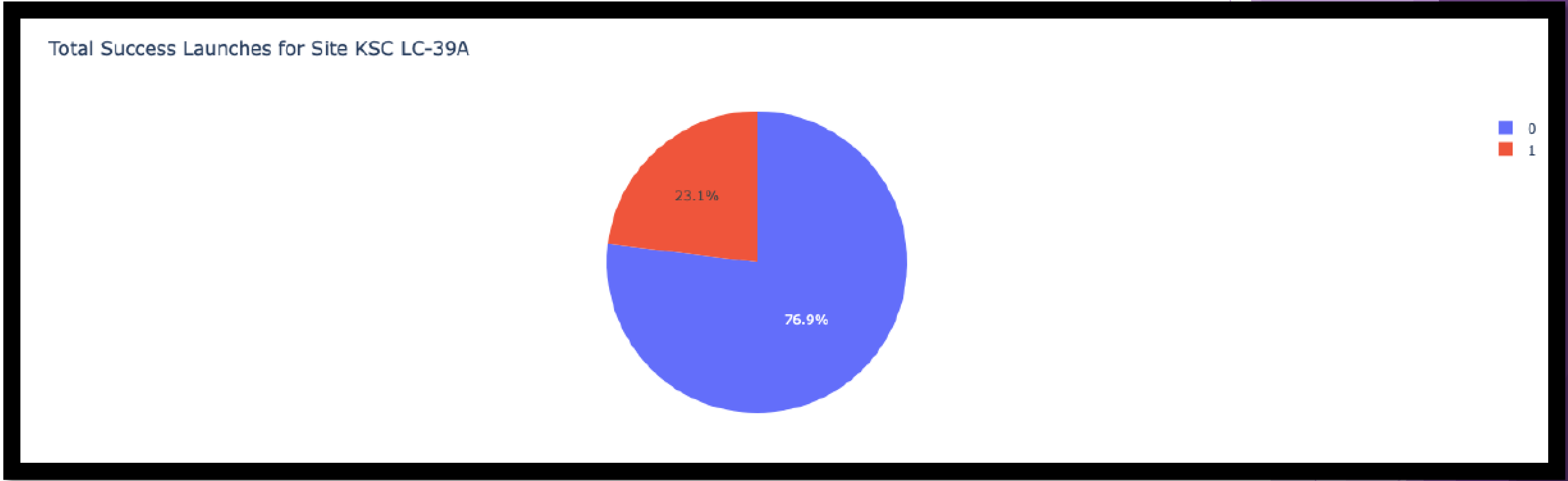
BUILD A DASHBOARD WITH PLOTLY DASH

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.

SECTION 5

- PREDICTIVE ANALYSIS (CLASSIFICATION)

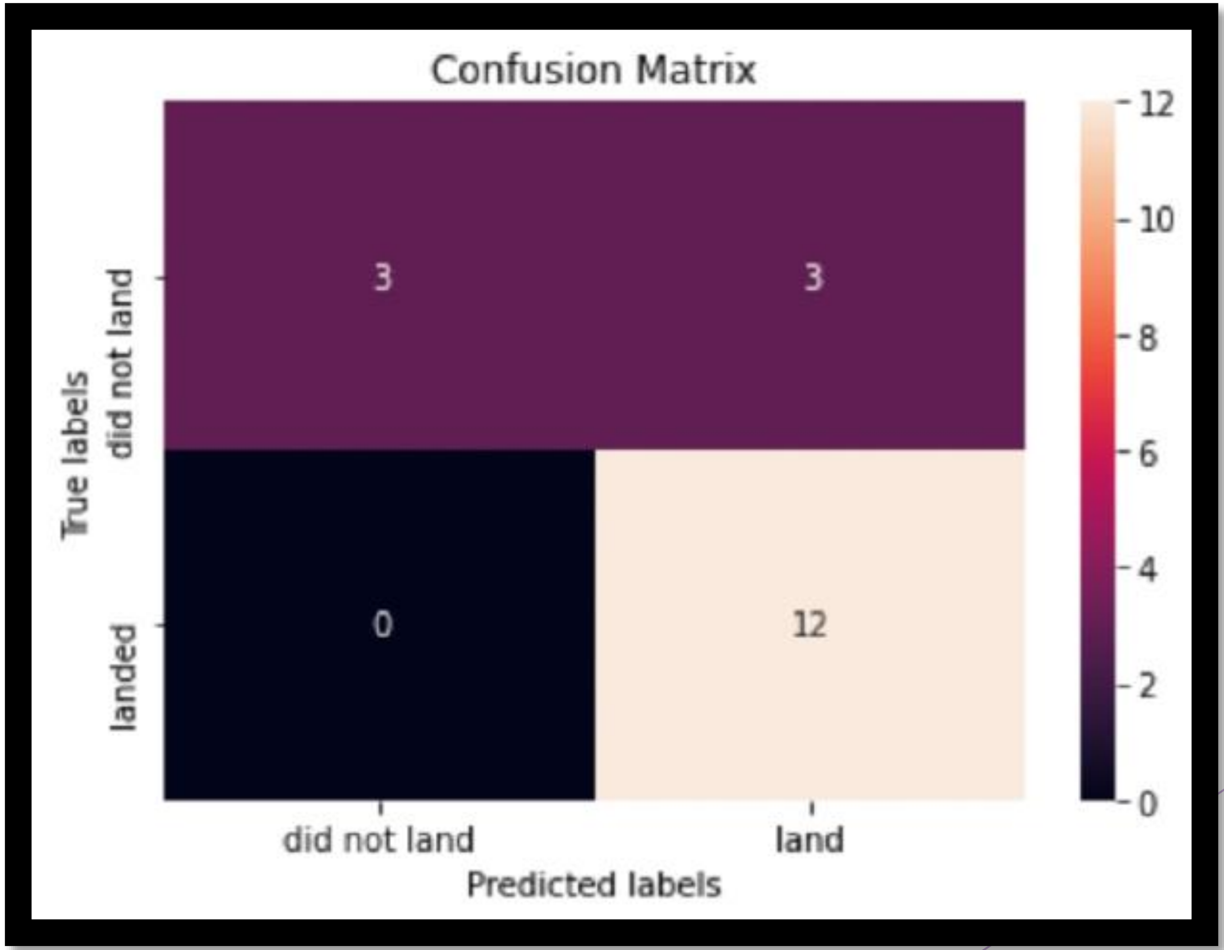
Classification Accuracy

- Unable to verify which strategy works better based on the Test Set scores.
- The short test sample size (18 samples) may be the cause of the same test set scores.
- Consequently, we used the entire dataset to test every approach. • The Decision Tree Model is the best model, as confirmed by the scores of the entire dataset. This model offers the highest accuracy as well as higher scores.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.808219	0.819444
F1_Score	0.909091	0.916031	0.893939	0.900763
Accuracy	0.866667	0.877778	0.844444	0.855556

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

- Decision Tree Model is 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.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- Dataset obtained from – IBM Applied Data Science Capstone Project Course
- Skills taught by – IBM Data Science Professional Certificate Courses


THANK YOU