

# Winning Space Race with Data Science



# Outline

---

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



# Executive Summary

---



- **Summary of methodologies**

This project contains different types of methodologies to carry out a data science project;

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis using SQL
- Constructing an interactive map using Folium Library
- Building an interactive dashboard using Plotly Dash
- Predictive Modeling using different algorithms

- **Summary of all results**

- Exploratory Data Analysis Results along with screenshots and deliverables
- Interactive Analytics Web Application (Dashboard)
- Predictive Analysis Results (Classification)

# Introduction

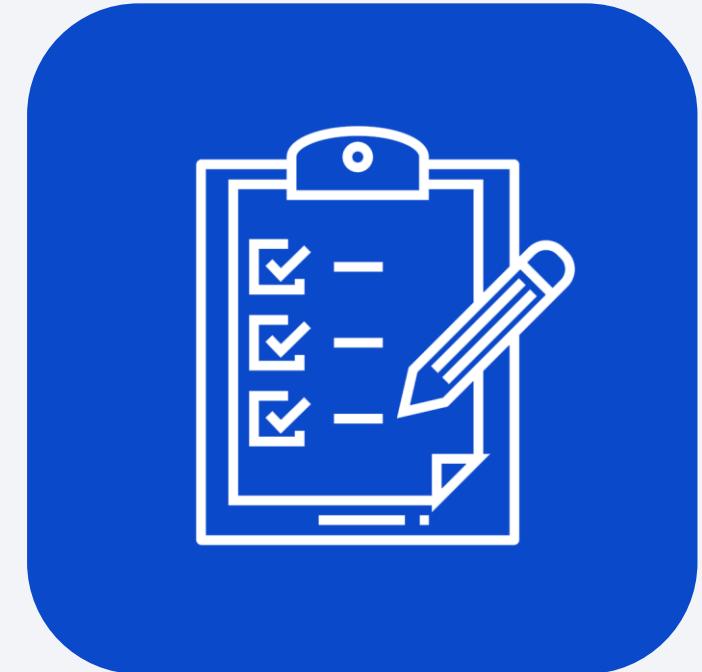
---

- **Project background and context**

SpaceX is the most successful commercial company of space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars. At the heart of SpaceX's spate of successes is the Falcon 9, which has brought down the cost of reaching space and become a springboard. While 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- **Objectives**

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?



Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - Data Collected using SpaceX REST API and web scraped from Wikipedia
- Perform data wrangling
  - Filtered data, dealt with missing values, transformed the data using One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Selected required features, split the data, build models, tuned and performed evaluation to ensure an optimized model

# Data Collection

---

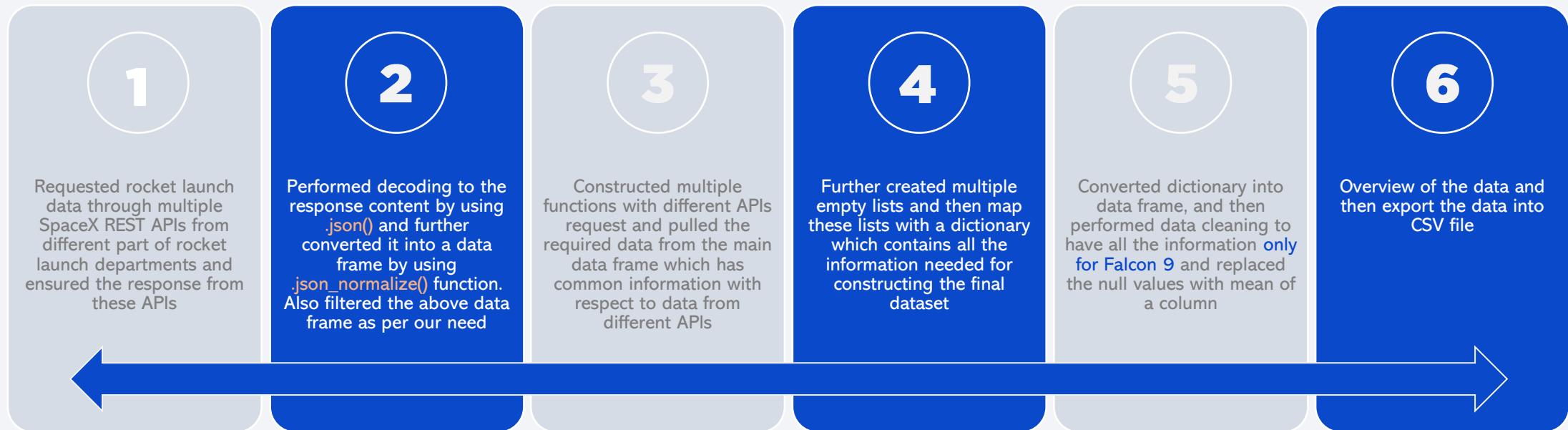


After understanding the context and objectives of project, data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both data collection methods in order to get complete information about the launches for a more detailed analysis.

Multiple columns are obtained by using SpaceX REST API, such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Serial, Longitude, Latitude and more.

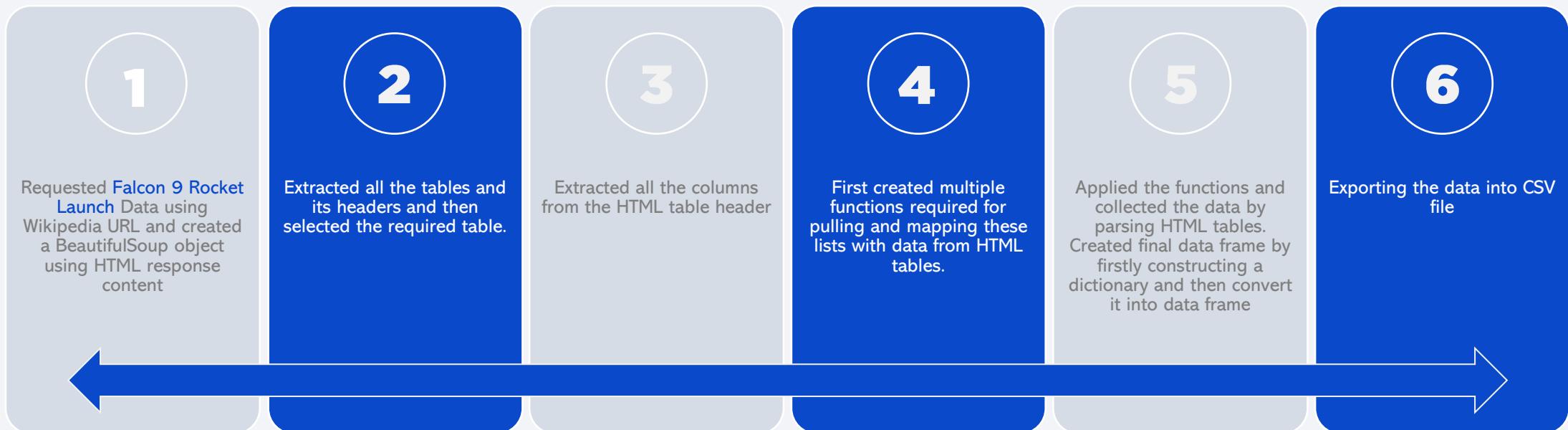
Data Columns are obtained by using Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

# Data Collection – SpaceX API



# Data Collection - Scraping

---



# Data Wrangling

---

Before training the supervised models, some Exploratory Data Analysis (EDA) were conducted to find some patterns in the data and determine what would be the label for training supervised models.



## Descriptive Exploration

Conducted basic data exploration using pandas and identified the data types of all the columns, null values, and more



## Calculation

Performed basic calculations to get familiar with data such as total number of launches for each launch site etc.



## Data Transformation

Transformed landing outcomes into a new column (Class) with binary values where 0 for Failure/None, and 1 for Successful landings



## Export CSV File

Convert Data frame into CSV and export it



[Github URL](#)

# EDA with Data Visualization

---

Charts were plotted:

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

- ✓ 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 plus represents the success rate among Orbit type. Moreover, 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

---

## Tasks completed using 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

---

- **Markers of all Launch Sites:**
  - ✓ First created a map object and further, 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.
- **Colored Markers of the landing outcomes for each Launch Site:**
  - ✓ Added colored 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 colored Lines to show distances between the Launch Site [VAFB SLC-4E](#) (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

# Build a Dashboard with Plotly Dash

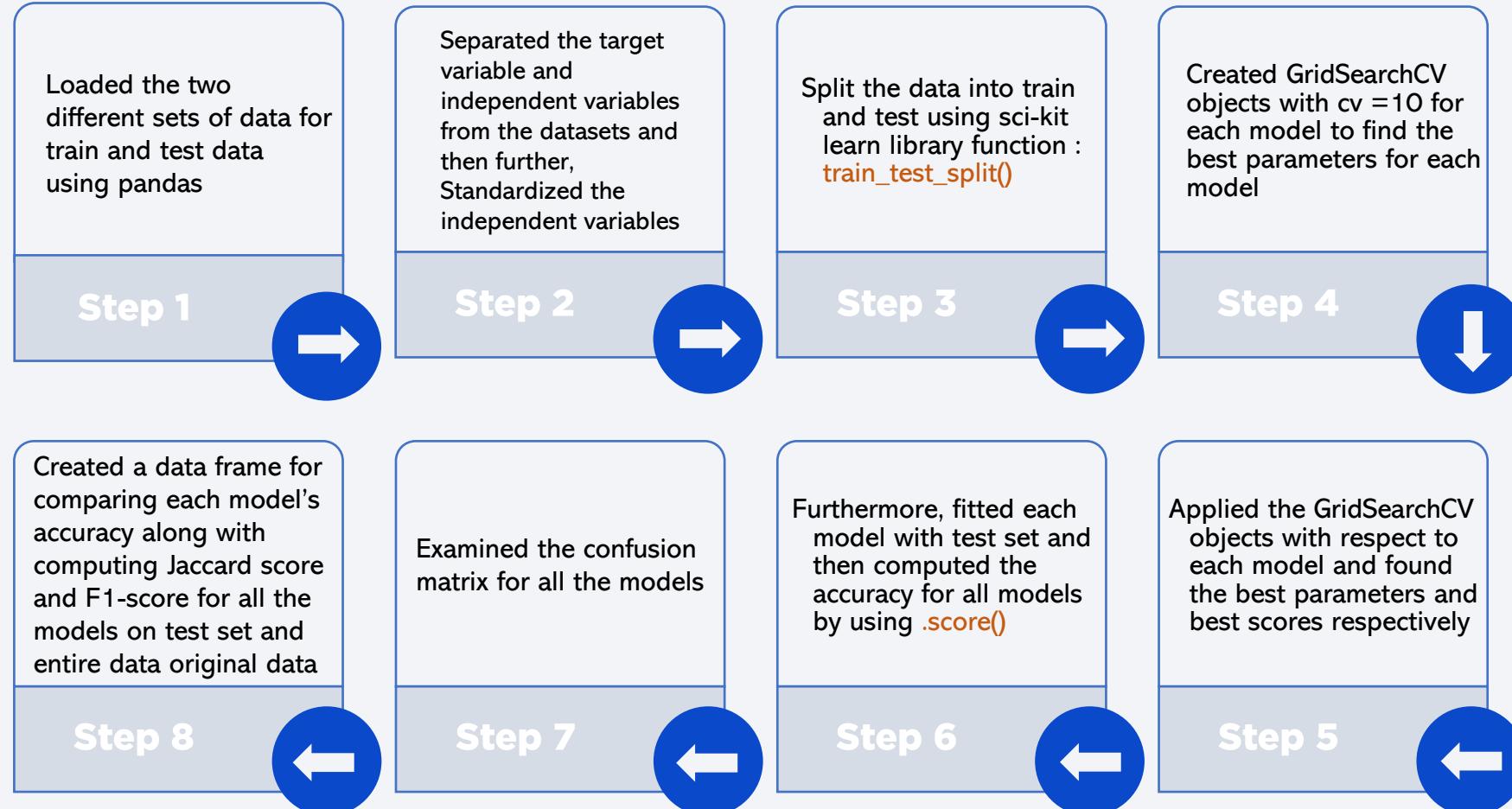
---

- 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 % of successful landings for all sites and % of Success vs. Failed landings for each site, if a specific Launch Site is 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 with respect to booster version.



# Predictive Analysis (Classification)

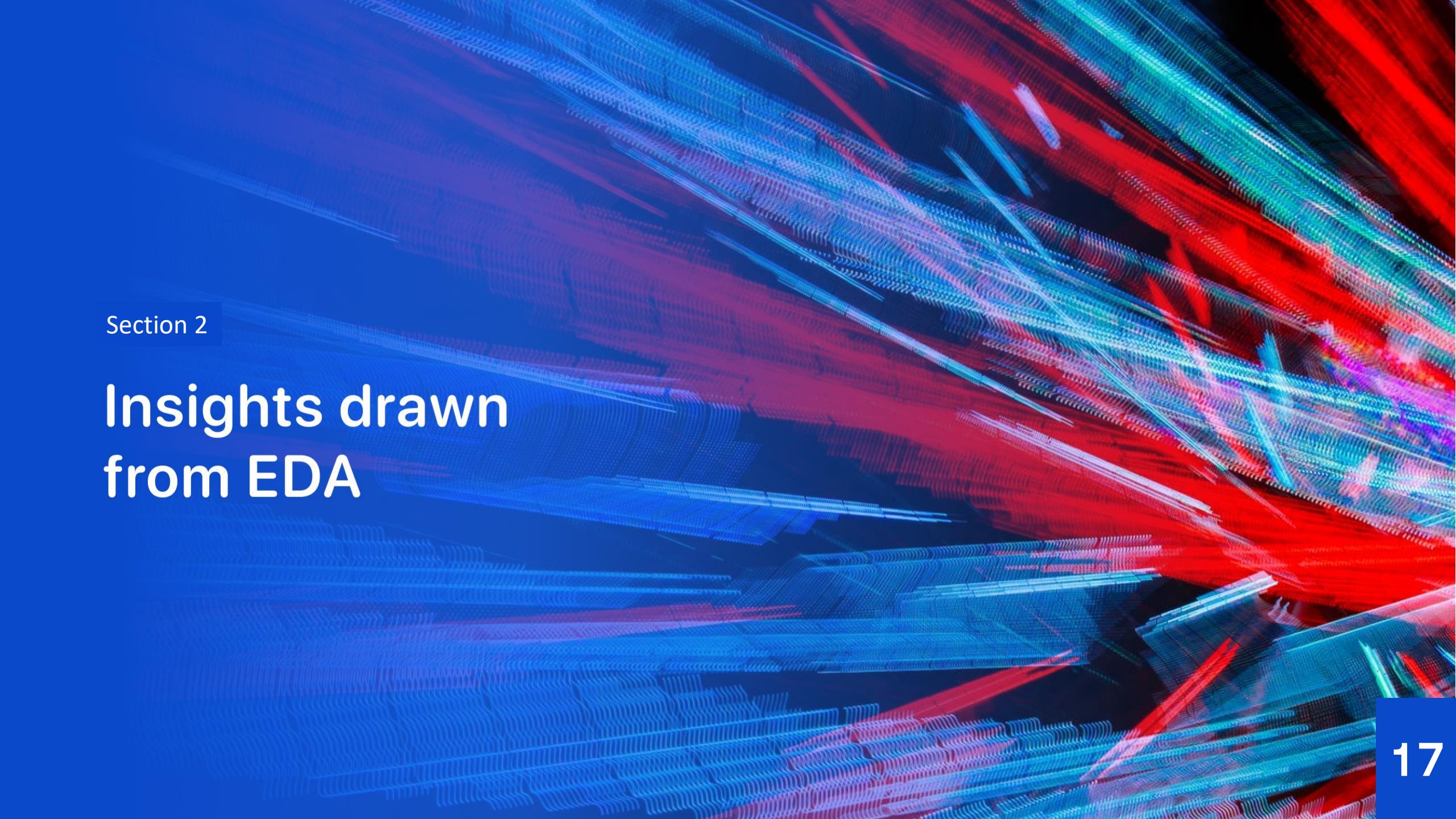
 [Github URL](#)



# Results

---

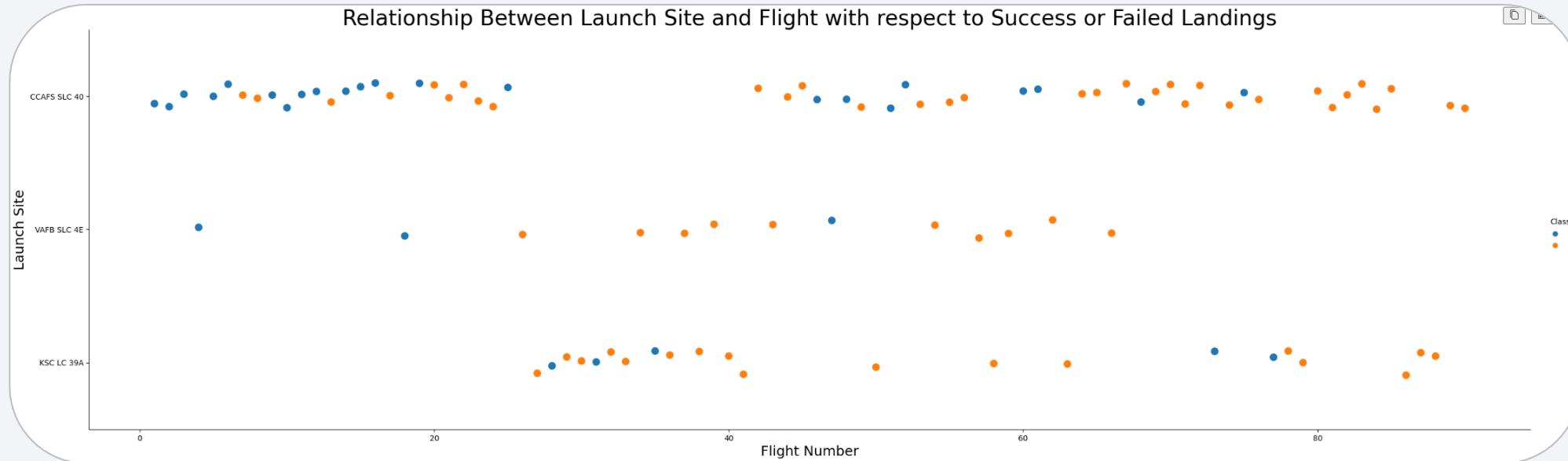
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a 3D wireframe or a network of data points. The overall effect is futuristic and dynamic.

Section 2

## Insights drawn from EDA

# Flight Number vs. Launch Site



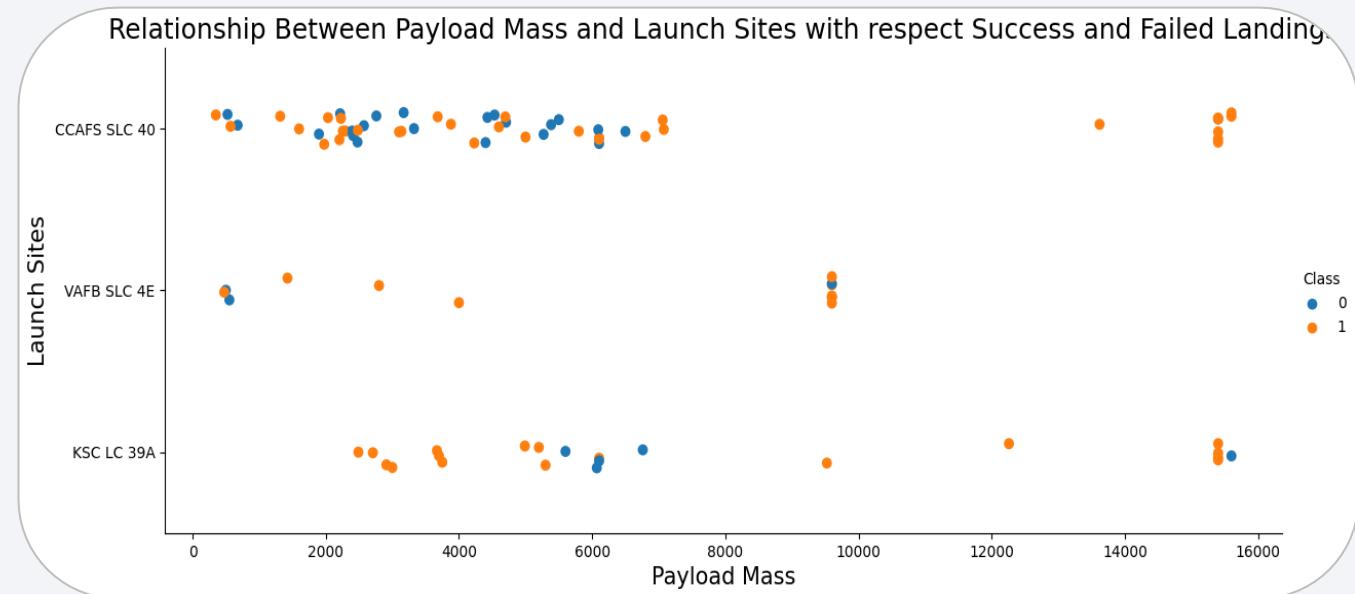
## Key Insights

- ✓ The earliest flights mostly failed while the latest flights mostly succeeded.
- ✓ The CCAFS SLC 40 launch site has performed the most launchings than any other launch sites.
- ✓ 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

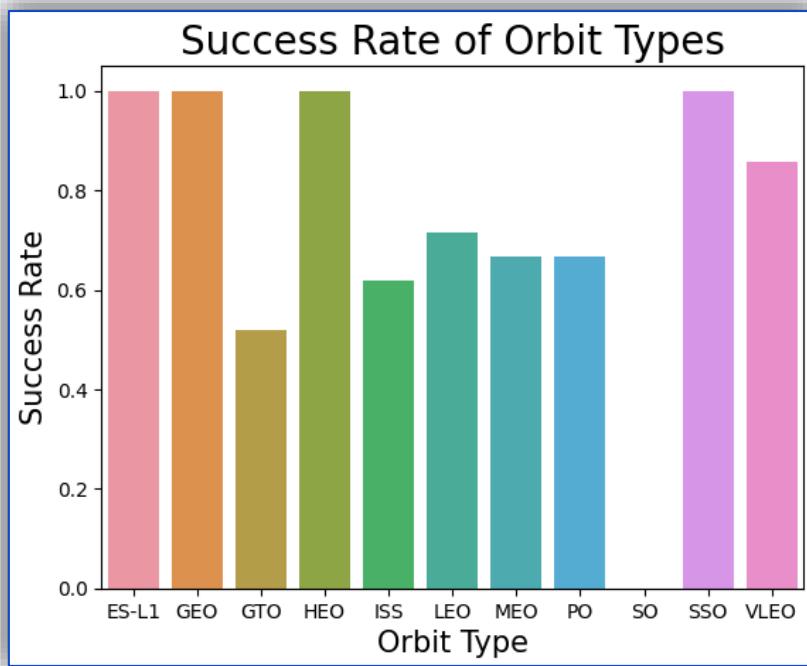
## Key Insights:

- ✓ As mentioned above, for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- ✓ 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.
- ✓ For Payload mass between 7100 and 15000, launches for any site didn't take place much as compared to other payload mass.



# Success Rate vs. Orbit Type

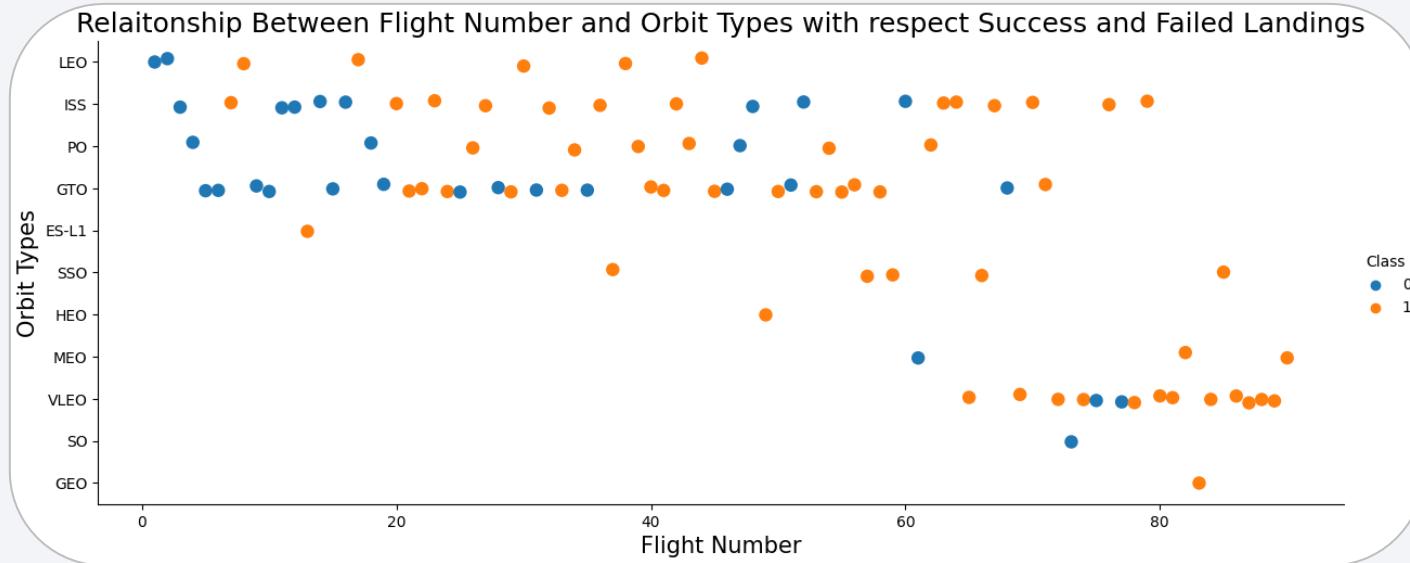
---



## Key Insights:

- ✓ The success rate of Orbit Types; ES-L1, GEO, SSO, and HEO are 100%
- ✓ GTO orbit has the lowest success rate among other orbits
- ✓ Orbit with 0% success rate is SO
- ✓ Orbits with above 60% success rate are ISS, LEO, MEO, PO, and VLEO

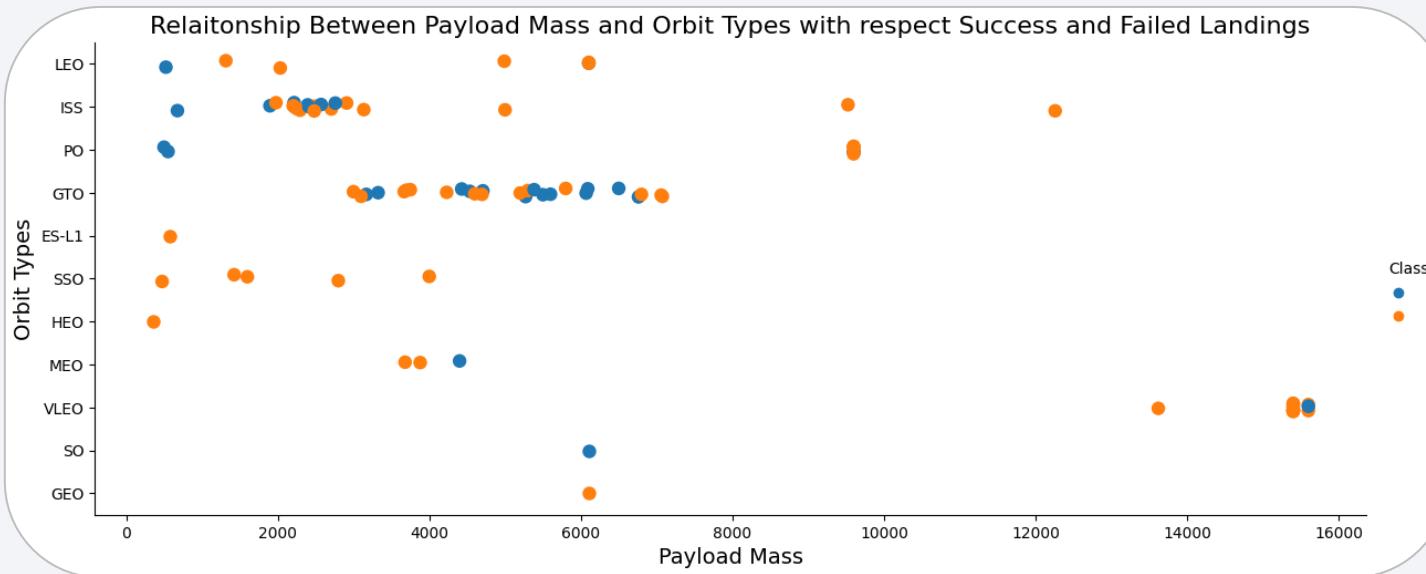
# Flight Number vs. Orbit Type



## Key Insights:

- ✓ As we can that earliest flights between 0-20 are less successful comparatively.
- ✓ Whereas the flights after 55, the success rate seems be more effective across all orbits.
- ✓ VLEO and ISS appear to be having more successful flights.

# Payload vs. Orbit Type



## Key Insights:

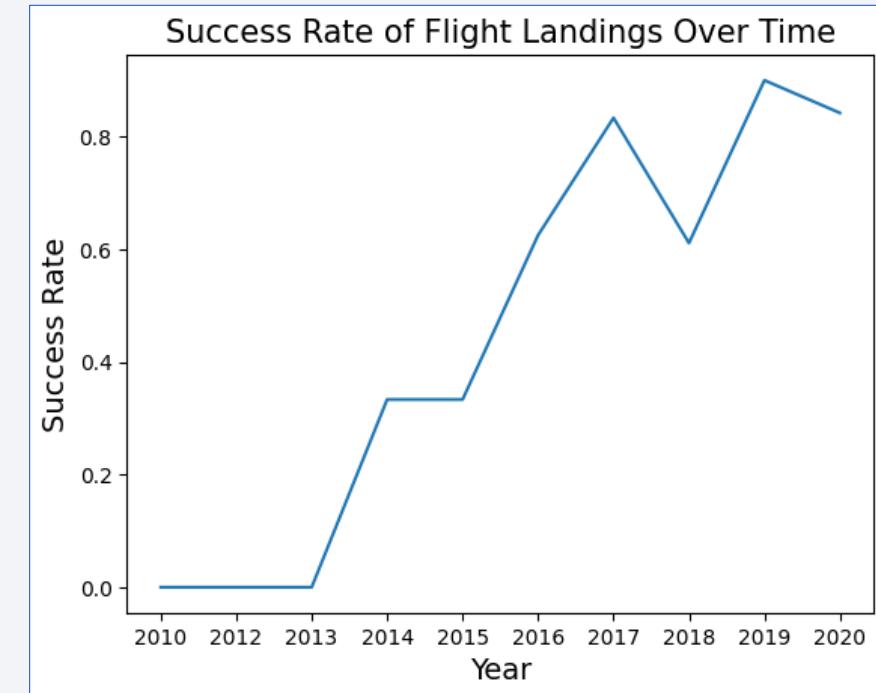
- ✓ Successful landings, for ES-L1, SSO, HEO, and MEO, appear to be more effective for the payload mass between 100 - 4000.
- ✓ With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

# Launch Success Yearly Trend

---

## Key Insights:

- ✓ We can observe that the success rate since 2013 kept increasing till 2020
- ✓ Success rate is highest for year 2019.
- ✓ While success rate is lowest in the initial years (2010-2013).
- ✓ We can also observe there is a decline in success rate in year 2018.



# EDA Using SQL : All Launch Site Names

---

The result shows the unique name of all the launch sites of SpaceX

```
%%sql  
SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE;
```

\* [sqlite:///my\\_data1.db](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'

Find 5 records where launch sites begin with `CCA`

Display 5 records where launch sites begin with the string 'CCA'

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

# Total Payload Mass

Calculate the total payload carried by boosters from NASA

```
%%sql
SELECT
    Customer,
    sum("PAYLOAD_MASS__KG_") AS total_payload_mass
FROM SPACEXTABLE
WHERE Customer = "NASA (CRS)"
GROUP BY Customer
ORDER BY total_payload_mass DESC;
```

```
* sqlite:///my\_data1.db
Done.
```

| Customer   | total_payload_mass |
|------------|--------------------|
| NASA (CRS) | 45596              |

# Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

```
%%sql  
  
SELECT  
    "Booster_Version",  
    AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass  
FROM SPACEXTABLE  
WHERE "Booster_Version" = 'F9 v1.1'  
GROUP BY "Booster_Version"
```

```
* sqlite:///my\_data1.db  
Done.
```

| Booster_Version | avg_payload_mass |
|-----------------|------------------|
| F9 v1.1         | 2928.4           |

# First Successful Ground Landing Date

---

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

```
%%sql
SELECT
    MIN("Date") as first_time_successful_landing_ground
FROM SPACEXTABLE
WHERE "Landing_Outcome" = "Success (ground pad)";

* sqlite:///my\_data1.db
Done.
```

**first\_time\_successful\_landing\_ground**

2015-12-22

## Successful Drone Ship Landing with Payload between 4000 and 6000

---

```
%%sql

SELECT
    "Booster_Version",
    "Landing_Outcome",
    "PAYLOAD_MASS_KG_" AS payload_mass
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS_KG_" BETWEEN 4000 AND 6000
AND "Landing_Outcome" = "Success (drone ship)"

* sqlite:///my\_data1.db
Done.



| Booster_Version | Landing_Outcome      | payload_mass |
|-----------------|----------------------|--------------|
| 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         |


```

List of names of boosters which have successfully landed on drone ship and had payload mass 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(*) AS "total_mission_outcomes"
FROM SPACEXTABLE
GROUP BY "Mission_Outcome"

* sqlite:///my\_data1.db
Done.



| Mission_Outcome                  | total_mission_outcomes |
|----------------------------------|------------------------|
| 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

```
%%sql
SELECT
    "Booster_Version",
    "PAYLOAD_MASS__KG_" max_payload
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)

* sqlite:///my\_data1.db
Done.



| Booster_Version | max_payload |
|-----------------|-------------|
| 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

---

List the failed landing\_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT
    "Date",
    substr("Date", 6, 2) AS "Month",
    strftime('%Y', "Date") AS "Year", -- we can also use this method to extract year or month or day etc.
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = "Failure (drone ship)" AND "Year" = "2015"
```

\* [sqlite:///my\\_data1.db](sqlite:///my_data1.db)

Done.

| Date       | Month | Year | Landing_Outcome      | Booster_Version | Launch_Site |
|------------|-------|------|----------------------|-----------------|-------------|
| 2015-10-01 | 10    | 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

---

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(*) AS "Count"
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY "Count" DESC
```

\* [sqlite:///my\\_data1.db](sqlite:///my_data1.db)  
Done.

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

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against the dark void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States appears. In the upper left quadrant, the green and blue glow of the aurora borealis is visible in the upper atmosphere.

Section 3

# Launch Sites Proximities Analysis

# SpaceX Launch Sites Location

## Explanation:

- ✓ Most of Launch sites considered in this project are in proximity to the Equator line. Launch sites are made at the closest point possible to Equator line, because anything on the surface of the Earth at the equator is already moving at the maximum speed (1670 kilometers per hour). For example; launching from the equator makes the spacecraft move almost 500 km/hour faster once it is launched compared halfway to north pole.
- ✓ All launch sites considered in this project are in very close proximity to the coast. Before deploying the rockets in the sky, SpaceX makes sure to minimize the risk of having any debris dropping or exploding near people.



# Color-labeled Launch Outcomes

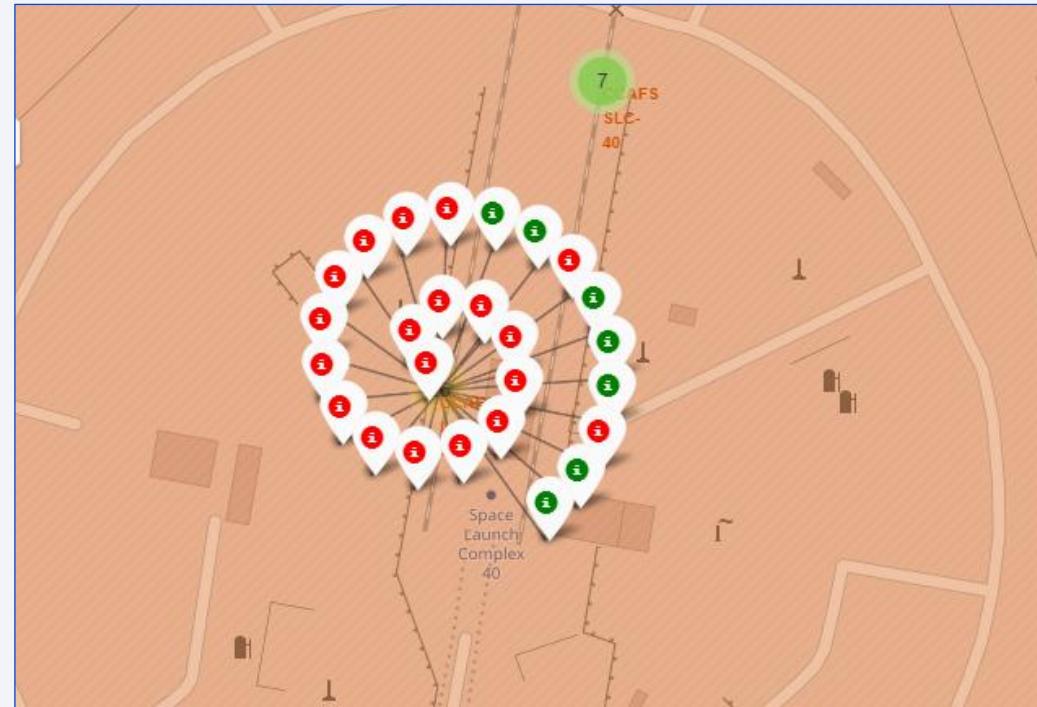
---

## Explanation:

From the color-labeled markers, we can easily identify which launch sites have relatively high success rates or failure rate.

- **Green Marker** = Successful Launch
- **Red Marker** = Failed Launch

Launch Site KSC LC-39A has a very high Success Rate.

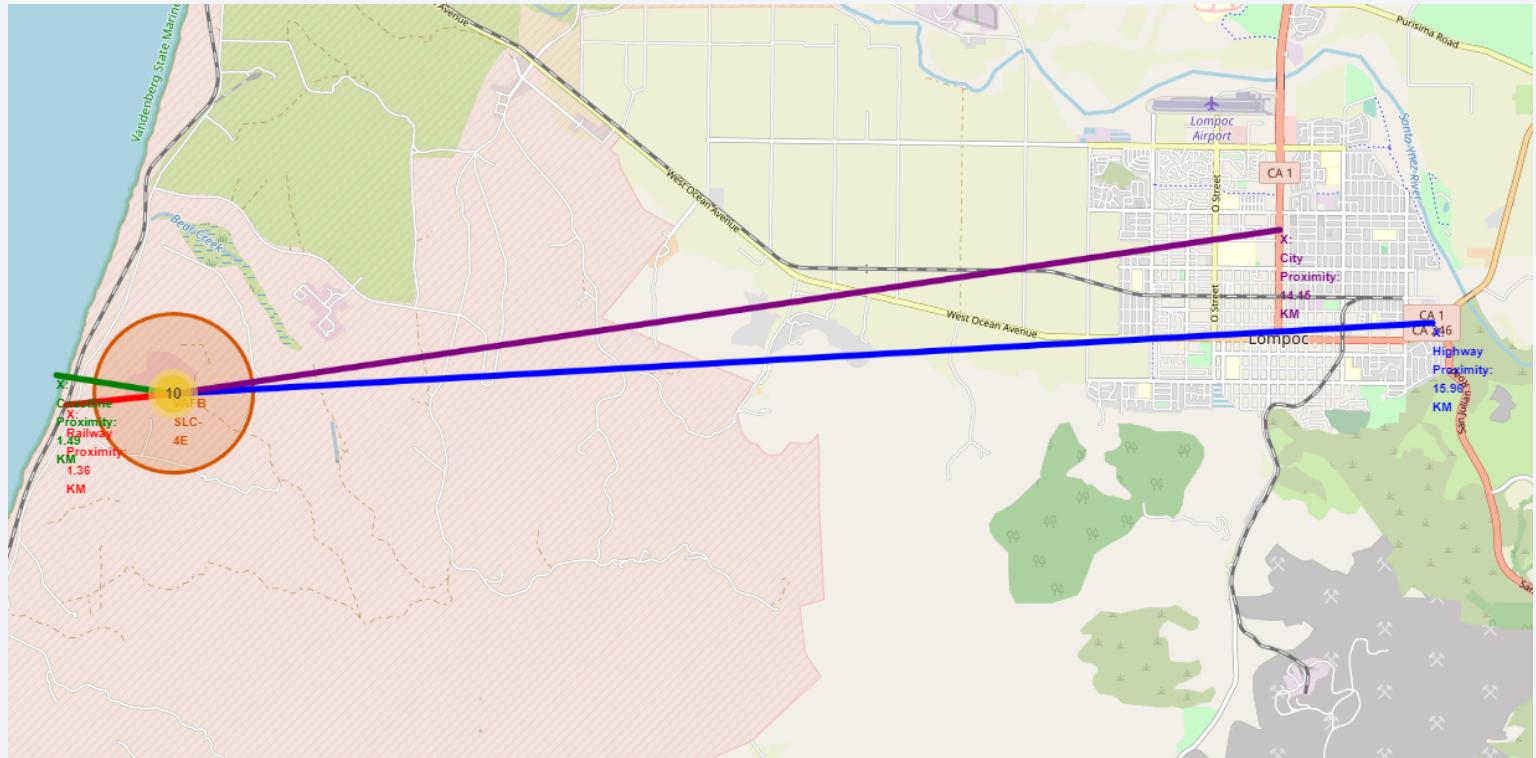


# Launch Site and its proximities

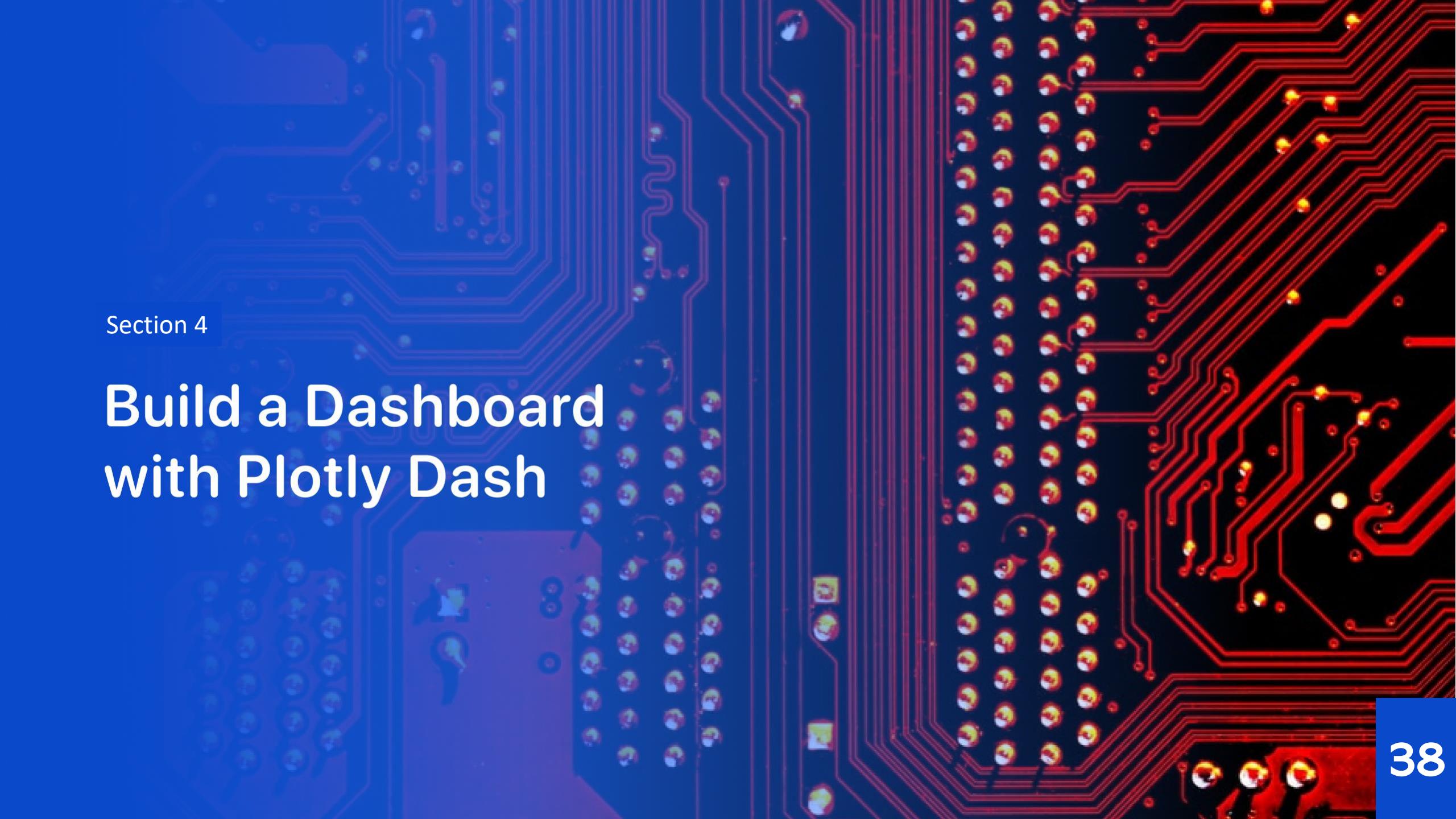
## Explanation:

From the visual analysis of the launch site VAFB SLC-4E we can clearly see that it is:

- relatively close to railway proximity (1.36 KM)
- relatively close to highway (15.96 km)
- relatively close to coastline (1.49 km)



Also, the launch site VAFB SLC-4E is relatively close to its closest city Lompoc (14.45 km).

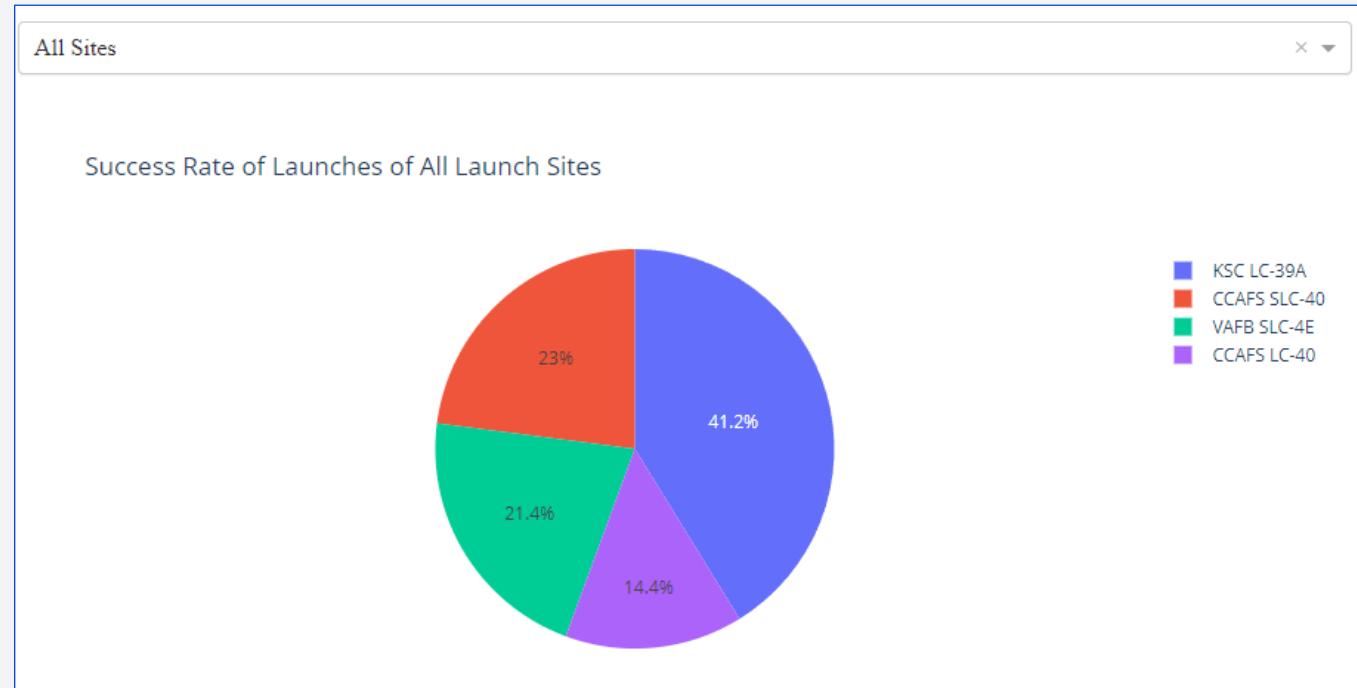


Section 4

# Build a Dashboard with Plotly Dash

# SpaceX Launch Site Successful Launches (%)

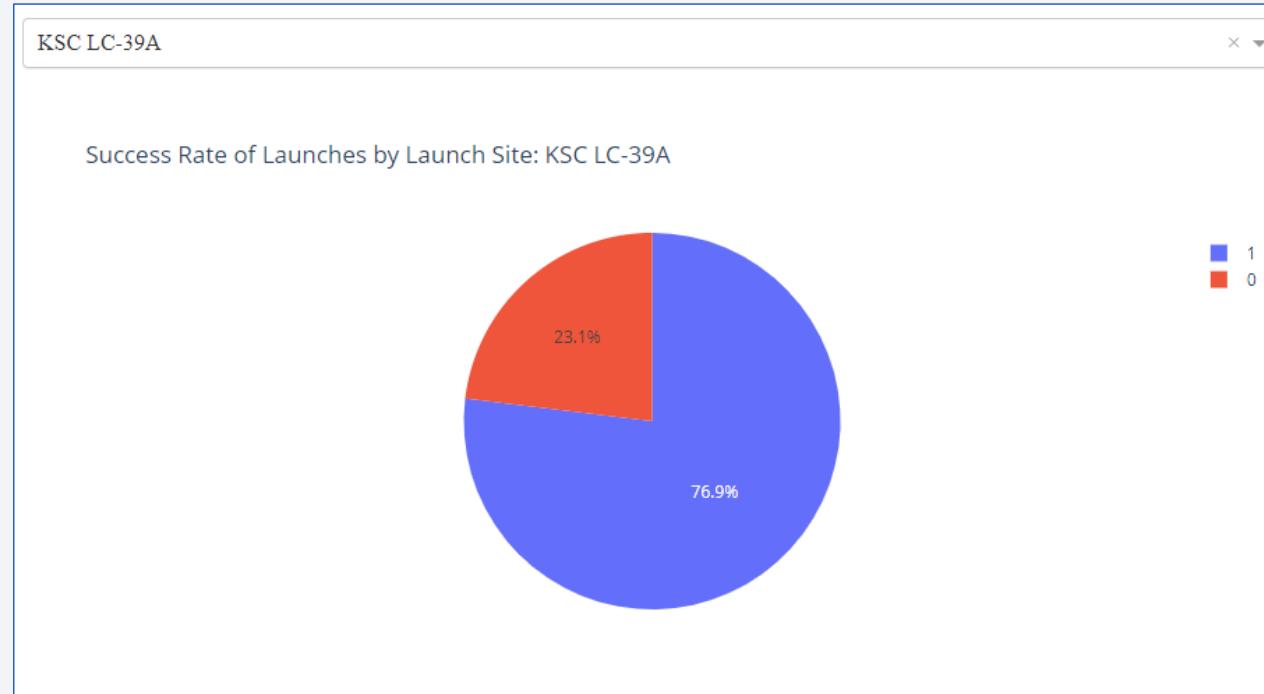
It clearly shows that the KSC LC-39A has the highest rate of success launches among others despite the fact how many launches have been taken placed at each launch site.



# KSC LC-39A Success Rate

---

As we can see that KSC LC-39A has the highest rate (~77%) of the success launches with 10 successful landings and 3 failed landings.



# Payload Mass vs. Launch Outcomes

Between 0 – 7000 of payload mass, the most successful booster is FT whereas booster v1.1 has most failed launch outcomes.



The background of the slide features a dynamic, abstract design. It consists of several thick, curved lines that transition from a bright yellow-green at the top right to a deep blue at the bottom left. These lines create a sense of motion and depth, resembling a tunnel or a stylized landscape. The overall effect is modern and professional.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

---

- From these screenshots, we can observe that, based on the test sets, the KKN model has the highest scores along with accuracy among other models.
- While based on the original data performance, we can't confirm which model is the best. However, it apparently appears that KNN model has the best accuracy among others despite the scores taken into consideration.

Accuracy and Scores on Test sets

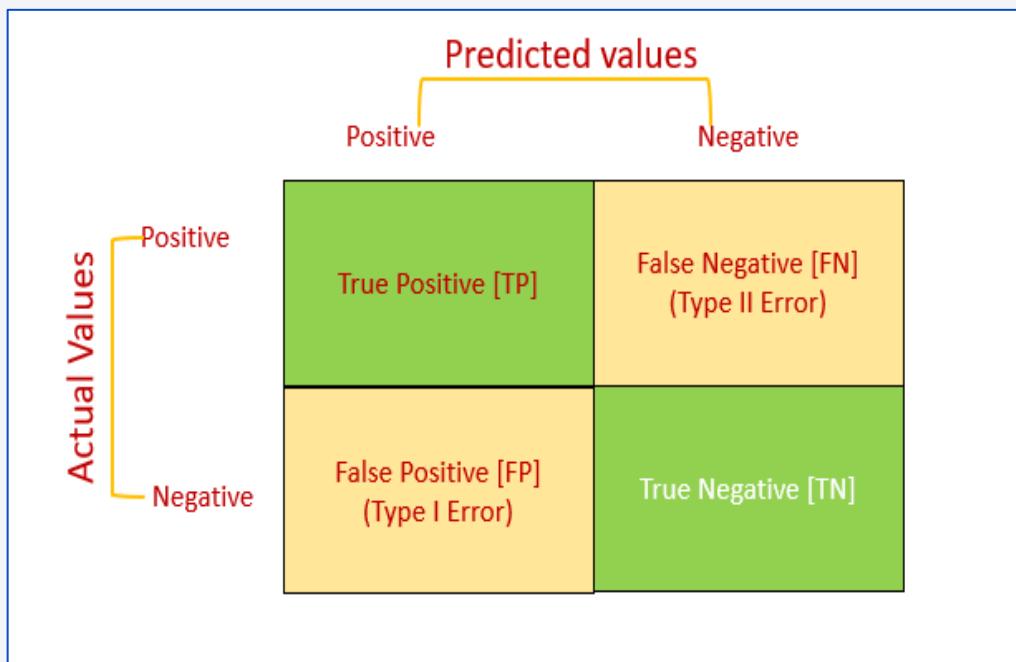
|   | Model Name             | Jaccard Score | F1-Score | Accuracy |
|---|------------------------|---------------|----------|----------|
| 0 | Logistic Regression    | 0.80          | 0.888889 | 0.833333 |
| 1 | Support Vector Machine | 0.80          | 0.888889 | 0.833333 |
| 2 | Decision Tree          | 0.75          | 0.857143 | 0.777778 |
| 3 | K-Nearest Neighbours   | 1.00          | 1.000000 | 1.000000 |

Accuracy and Scores on entire data

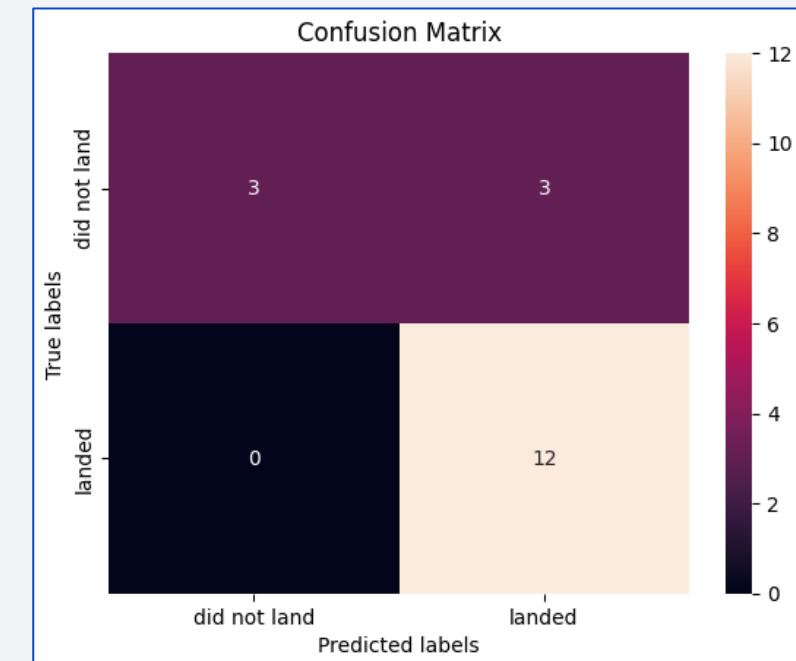
|   | Model Name             | Jaccard Score | F1-Score | Accuracy |
|---|------------------------|---------------|----------|----------|
| 0 | Logistic Regression    | 0.833333      | 0.909091 | 0.833333 |
| 1 | Support Vector Machine | 0.845070      | 0.916031 | 0.833333 |
| 2 | Decision Tree          | 0.753247      | 0.859259 | 0.777778 |
| 3 | K-Nearest Neighbours   | 0.732394      | 0.845528 | 1.000000 |

# Confusion Matrix

A confusion matrix presents a table layout of the different outcomes of the prediction and results of a classification problem and helps visualize its outcomes. It plots a table of all the predicted and actual values of a classifier.



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

---

- KNN 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 successful booster is FT for payload of under 7000.
- 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.
- Orbit ES-L1, GEO, HEO and SSO have 100% success rate.



# Appendix

Hello! I am Pavlo - self-motivated, go-getter, and self-taught individual

With a proven knack of providing data-driven actionable insights along with recommending informed-decision making to the stakeholders, I always thrive on learning about in-demand tools and technologies and adding them into my expertise, is what keeps me motivated and inspired on a daily basis. I started this project with a broader picture which includes – to advance my skills in Data Science and gain bigger picture on different kind of machine learning models. Not only I was exposed to many aspects of Data Science but also, I have embarked on many new tools and techniques throughout this course, which, I believe, will lead me to bring a unique blend of technical skills on the table.

---

Many Many Thanks To:

Instructors

Coursera

IBM

Thank you!

