

# Winning Space Race with Data Science

Mohamed Ibrahim Abdilahi  
Dec 26, 2022



# 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 (Classification)
- **Summary of all results**
  - Exploratory Data Analysis Results
  - Interactive Analysis demo in screenshots
  - Predictive analysis results

# Introduction

---

- Project background and context
- SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
  - 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?

# Introduction..

---

- **Project Objective**
- The project objective is to evaluate the viability of the new company SpaceY to make prediction on Space X by training and developing models.
- **Desirable answers**
  - The best way to estimate the total cos for launches, by predicting successful landing of the first stage of rockets
  - Where is the best place to make launches.

Section 1

# Methodology

# Methodology

---

- Data collection methodology:
  - Using SpaceX Reset API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with 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

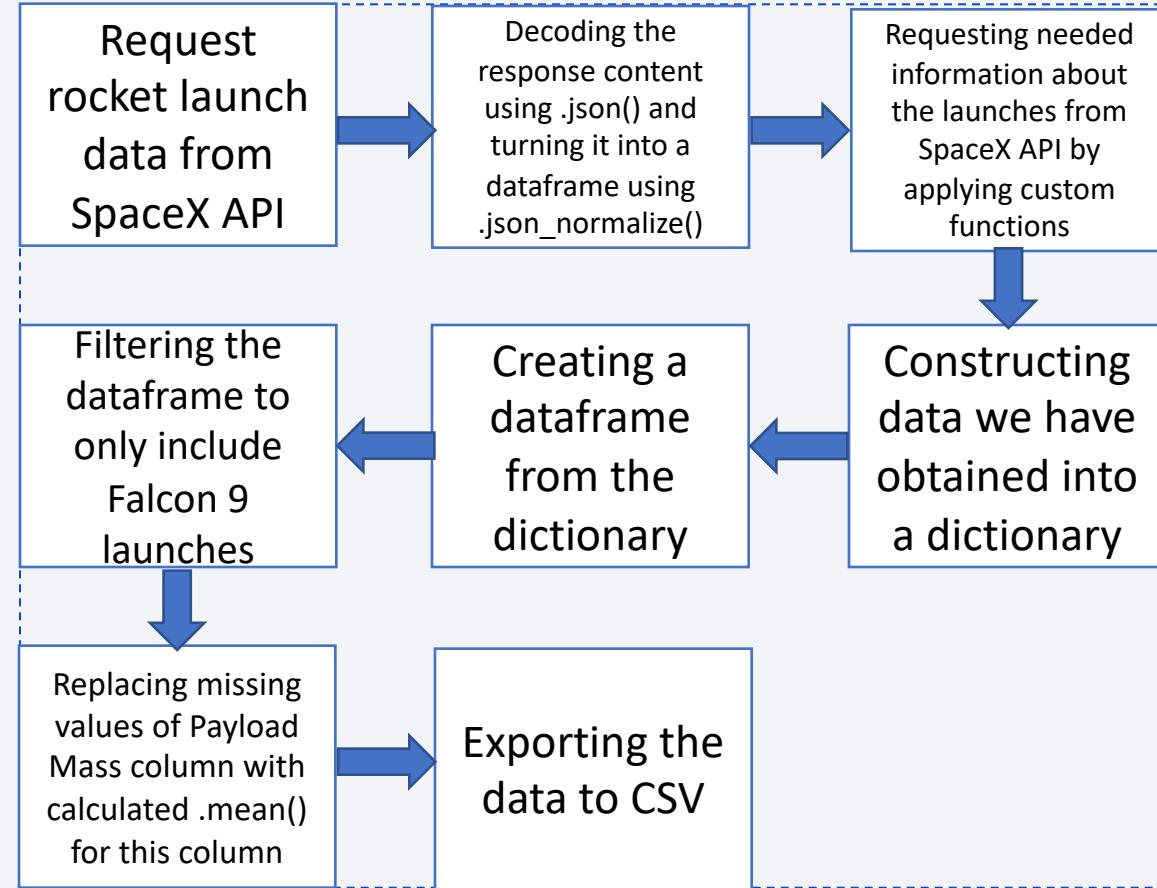
# Data Collection

---

- Data 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 of these data collection methods in order to get complete information about the launches for a more detailed analysis.
- Data Columns are obtained by using SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- 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

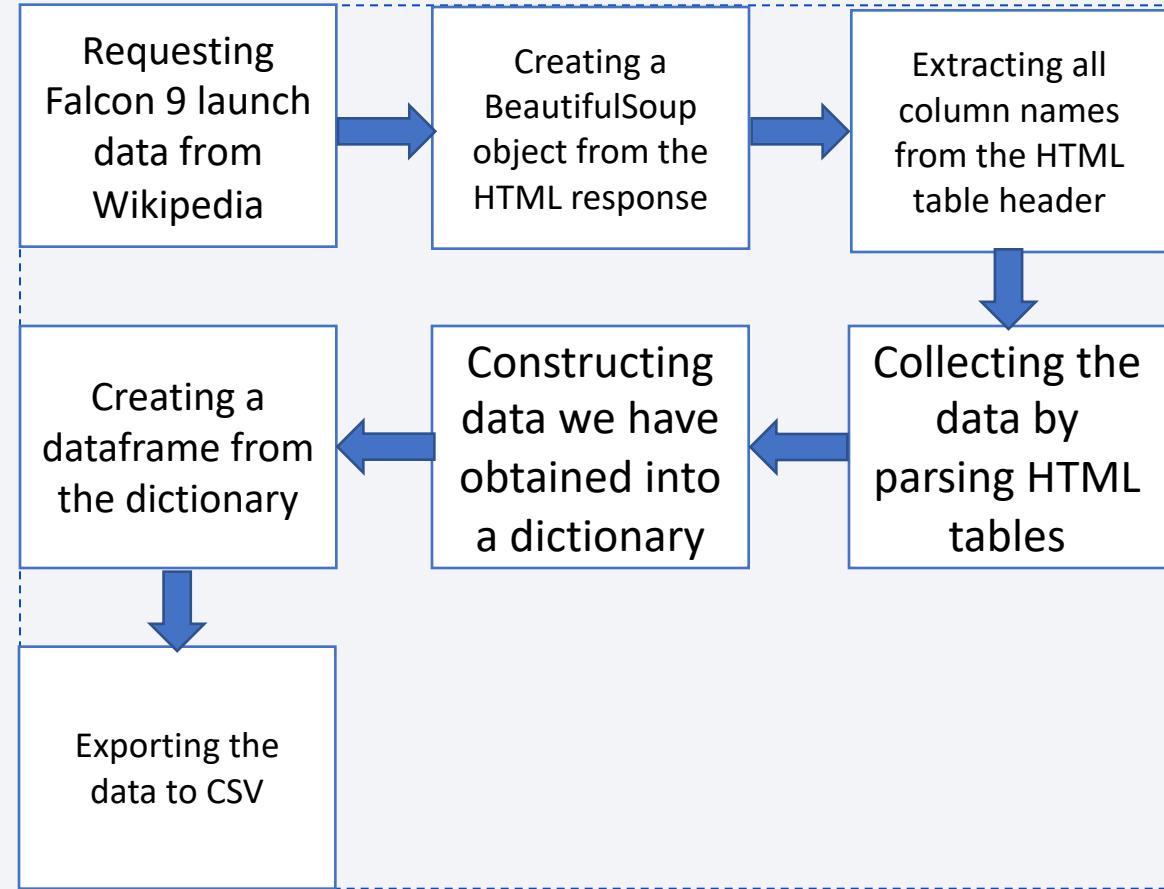
# Data Collection – SpaceX API

- Data collection using SpaceX REST calls using key phrases and flowcharts
- Refer to my GitHub URL of the completed SpaceX API calls notebook:
  - <https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/1-data-collection-api.ipynb>



# Data Collection – Scraping

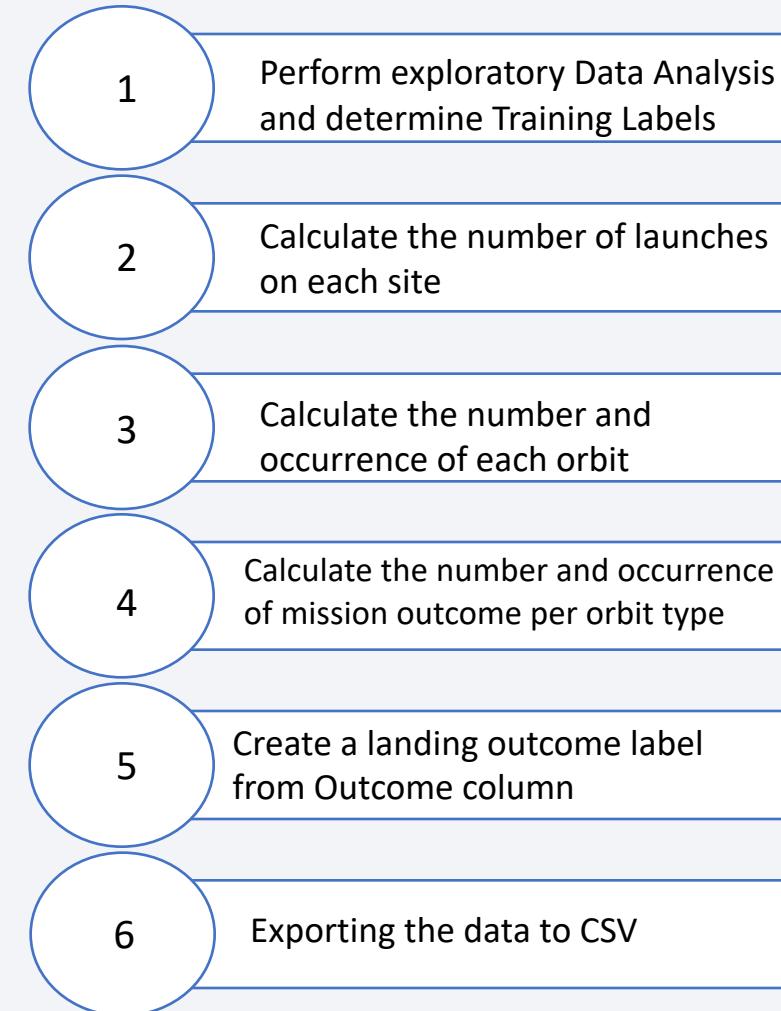
- Data collection was used web scraping process using key phrases and flowcharts.
- Refer to my GitHub URL of the completed SpaceX API calls notebook:
- <https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/2-data-collection-webscraping.ipynb>



# Data Wrangling

---

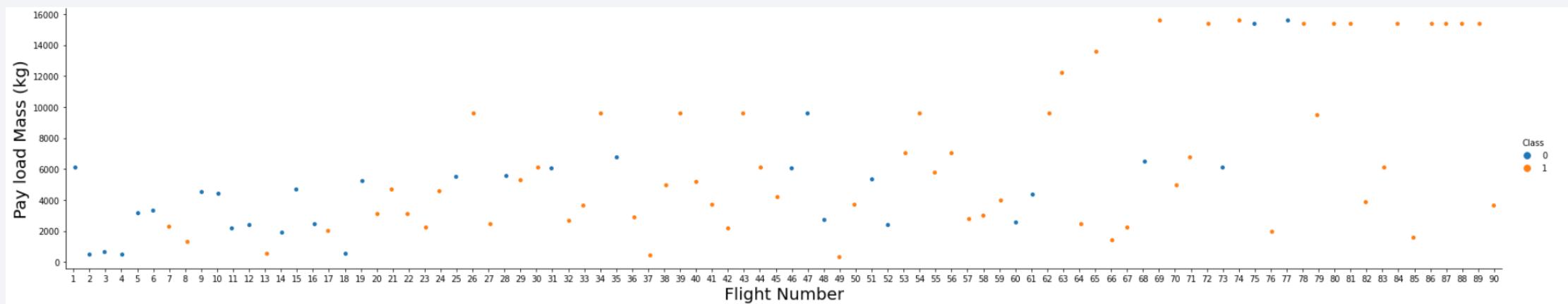
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.  
1
- GitHub URL: <https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/3-Data%20wrangling.ipynb>



# EDA with Data Visualization

---

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



- GitHub URL: <http://localhost:8888/lab/tree/Desktop/IBM/Week%202%20-%20EDA/5.EDA-with-Visaulization.ipynb>

# EDA with SQL

---

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- GitHub URL: <https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/4.%20Complete-EDA-SQL.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
- **GitHub URL:** <https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/5.EDA-with-Visaulization.ipynb>

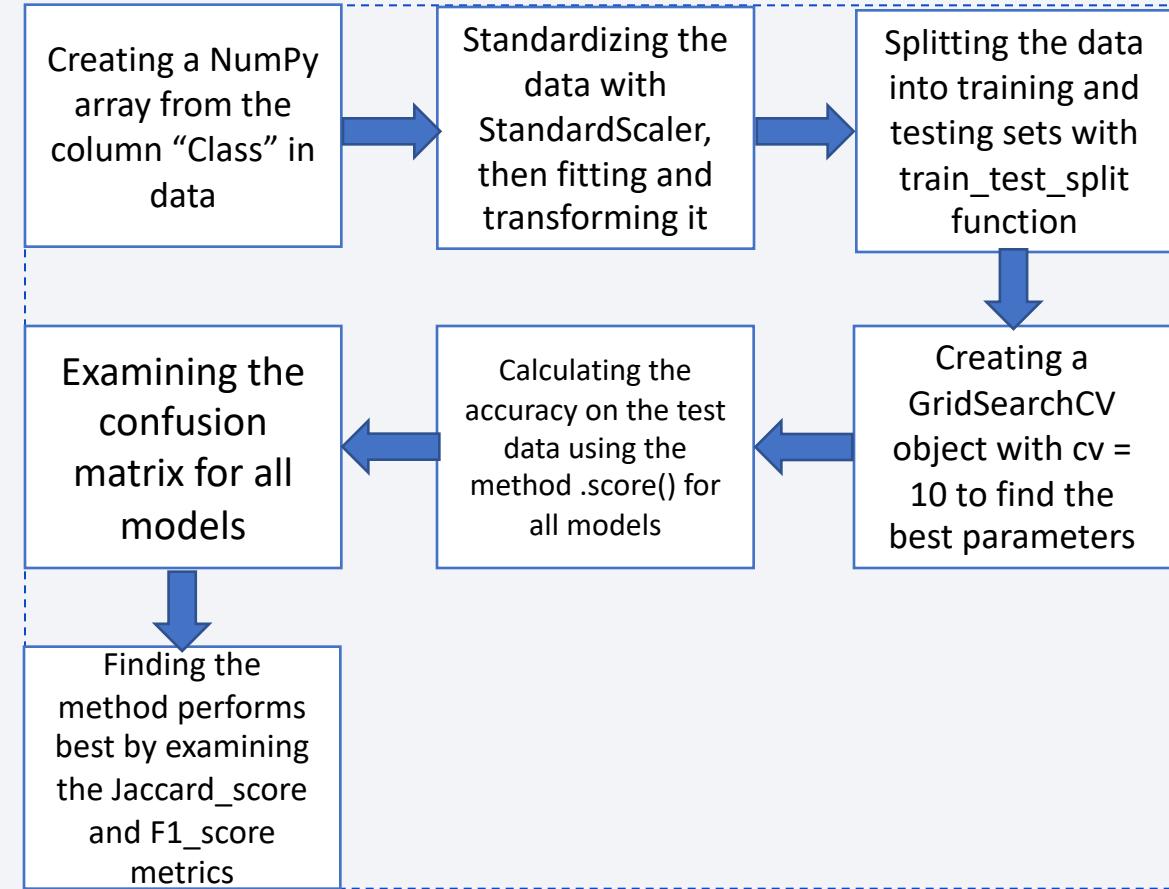
# Build a Dashboard with Plotly Dash

---

- 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 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
- GitHub URL: [https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/7.spacex\\_dash.ipynb](https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/7.spacex_dash.ipynb)

# Predictive Analysis (Classification)

- To make predication we will follow the scientific procedure that we are going to build the models
- Refer to my GitHub URL notebook:
- [https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/8.SpaceX\\_Machine%20Learning%20Prediction\\_Part\\_5.ipynb](https://github.com/Moheid/IBM-SpaceY-DS-Capstone/blob/main/8.SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)



# Predictive Analysis (Classification)

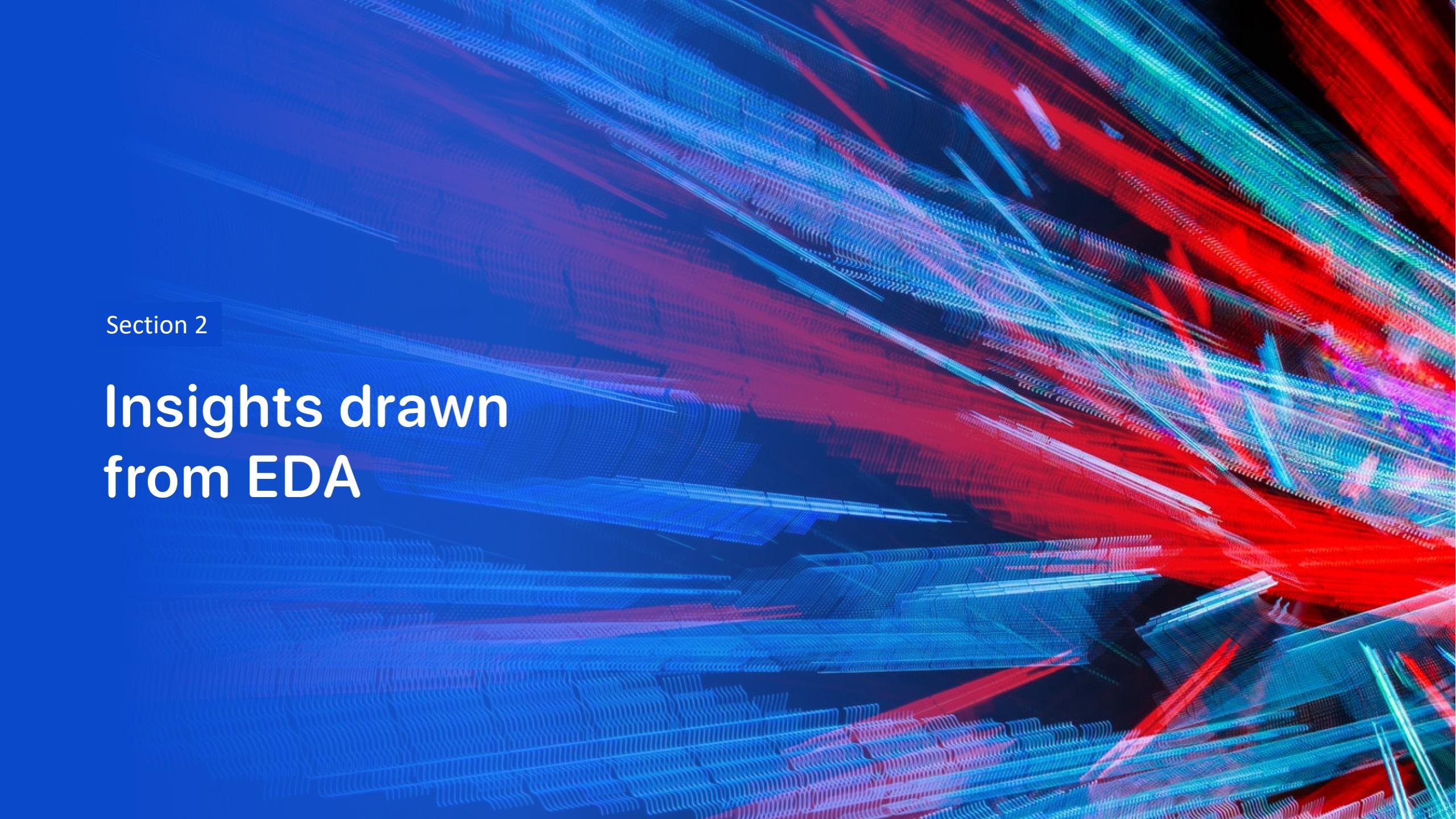
---

- 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
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

# Results

---

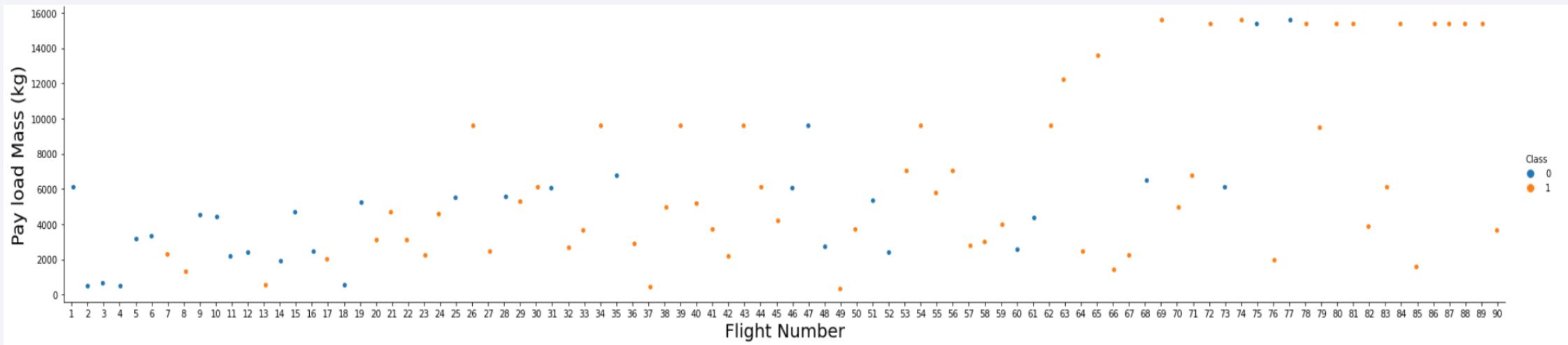
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

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 space or a network of data points. The overall effect is futuristic and dynamic.

Section 2

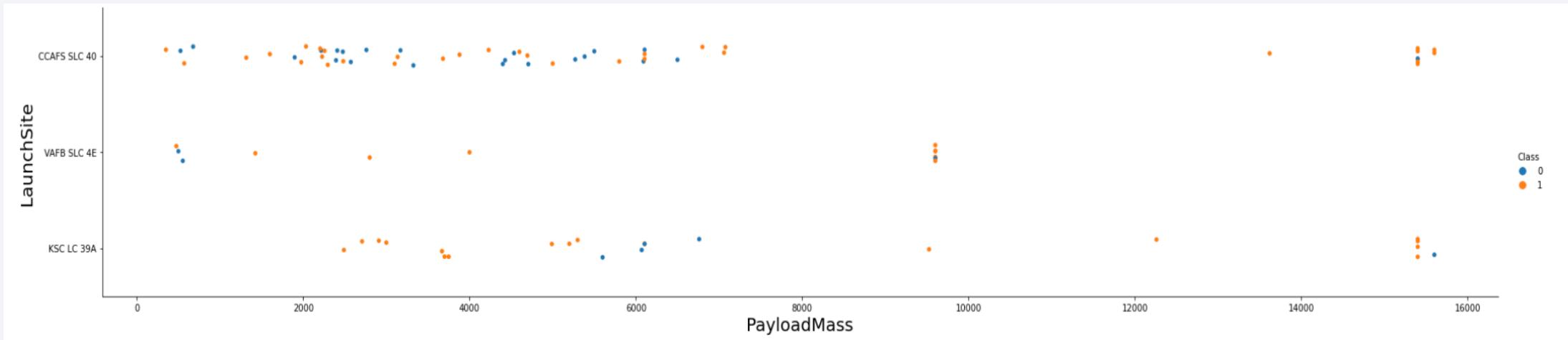
## Insights drawn from EDA

# 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

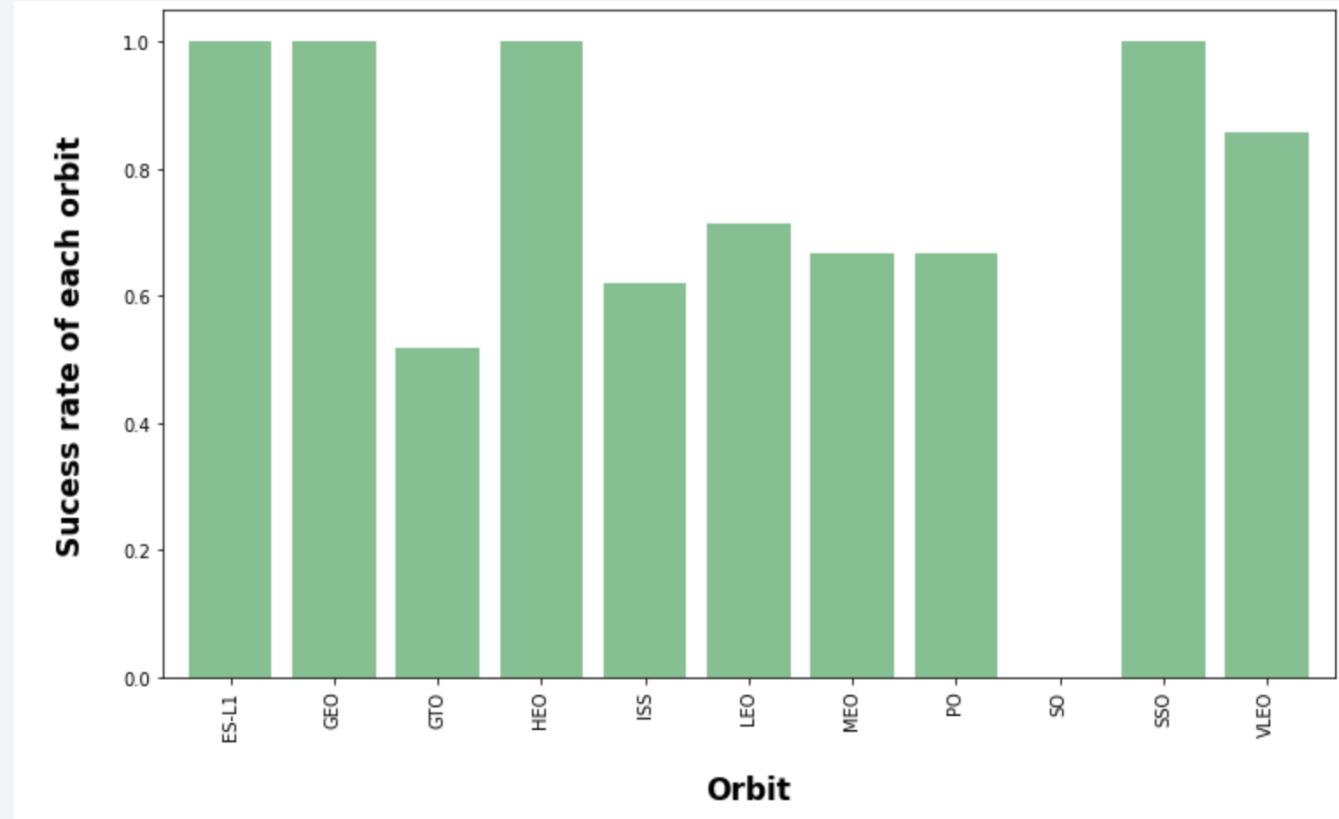


- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

# Success Rate vs. Orbit Type

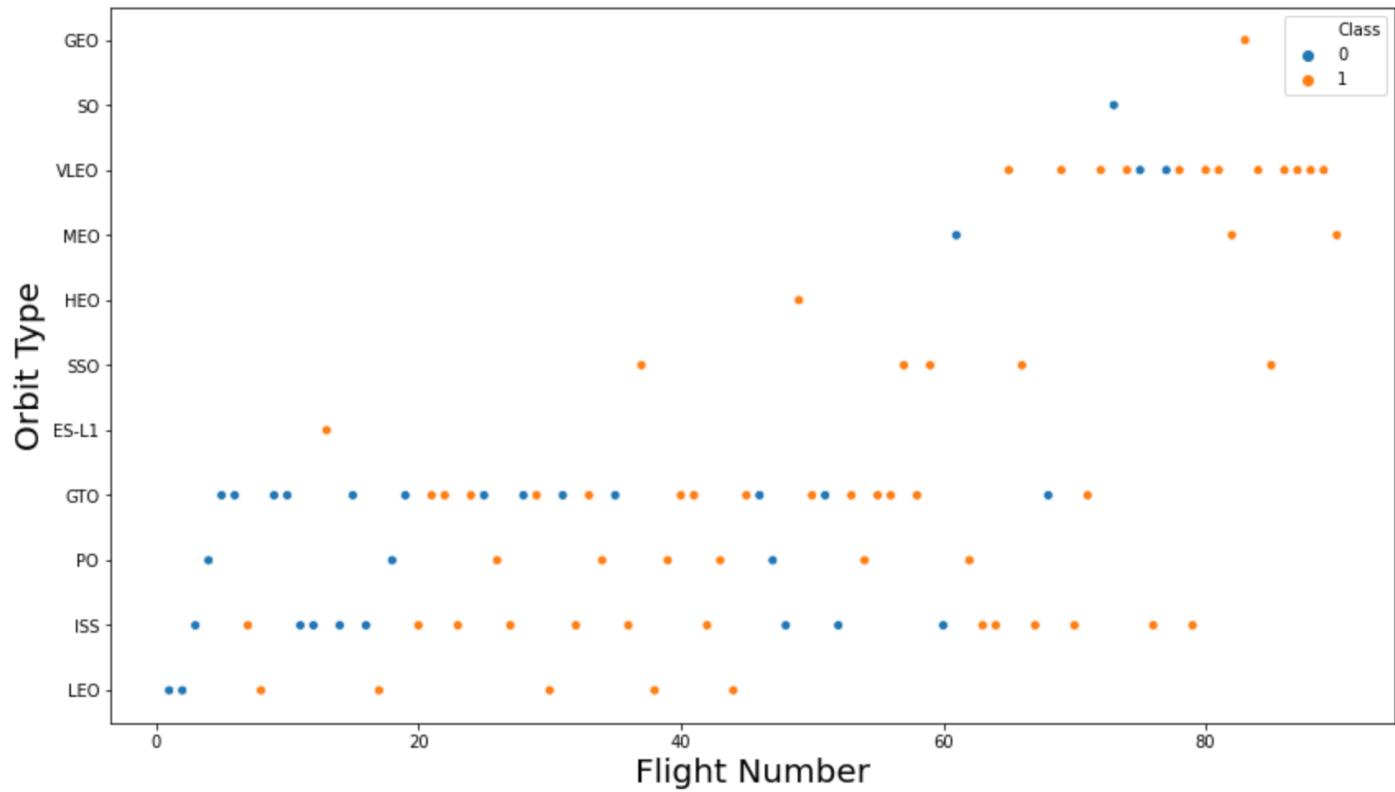
---

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



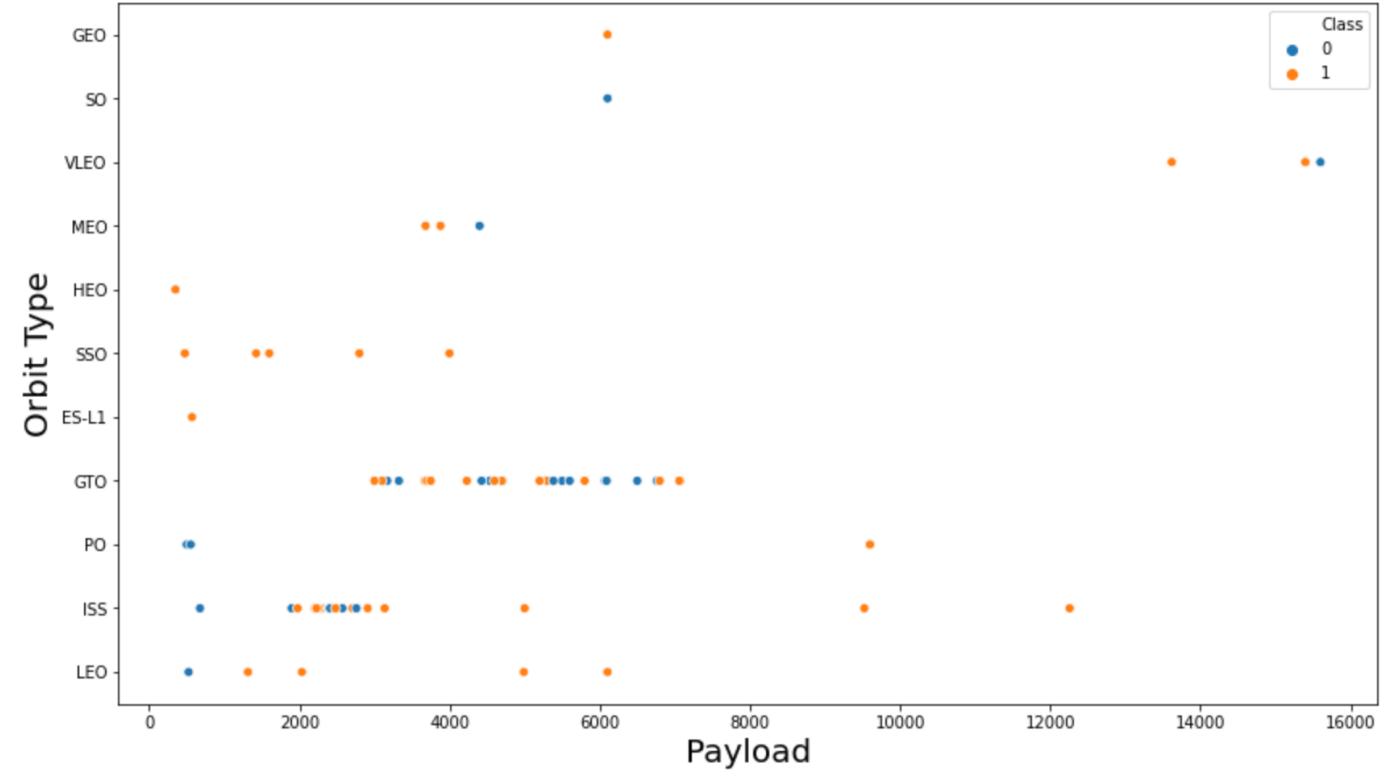
# Flight Number vs. Orbit Type

- 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

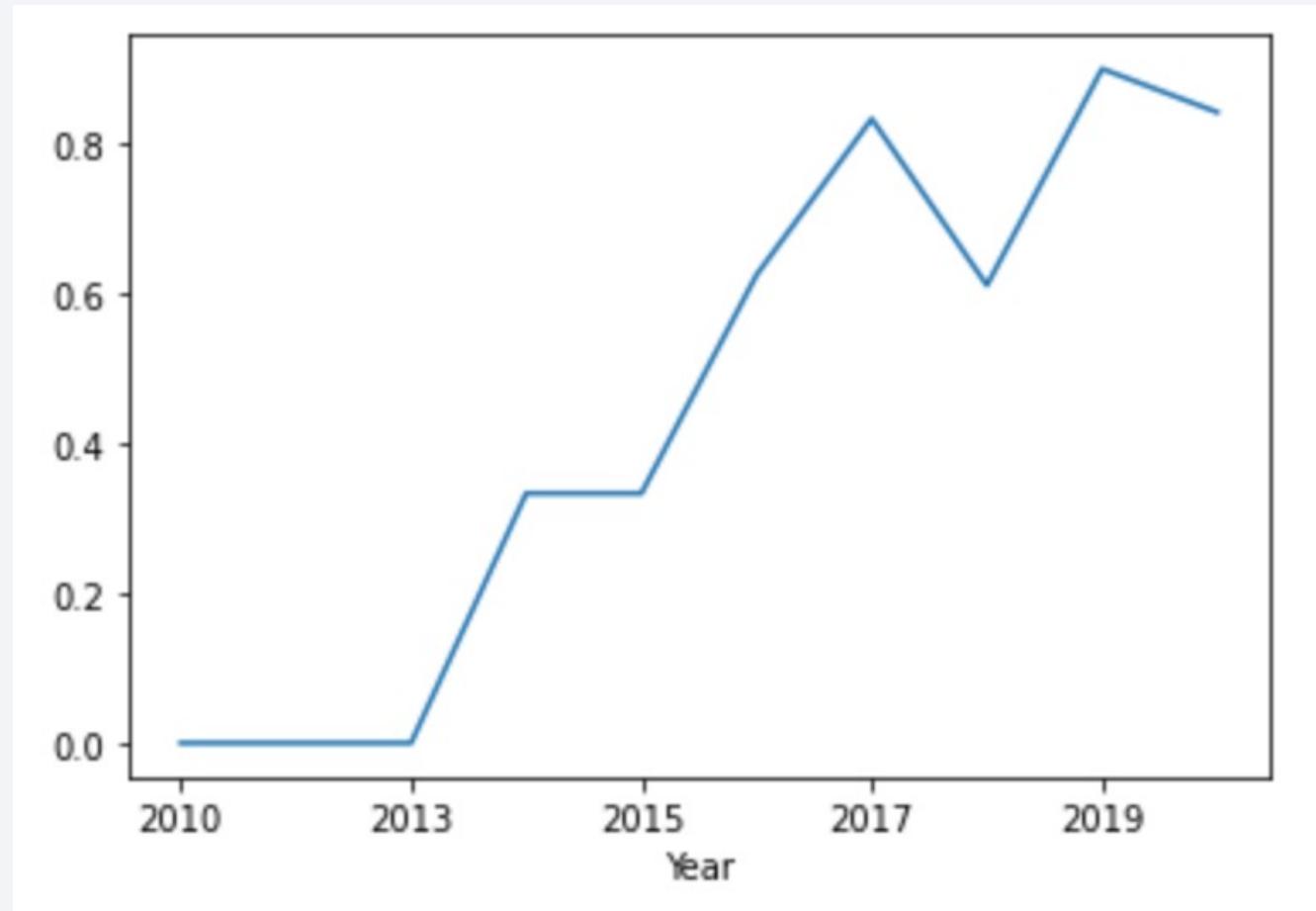
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



# Launch Success Yearly Trend

---

- you can observe that the success rate since 2013 kept increasing till 2020



# All Launch Site Names

```
# Select relevant sub-columns: `Launch Site`, `Lat(Latitude)`, `Long(Longitude)`, `class`  
spacex_df = spacex_df[['Launch Site', 'Lat', 'Long', 'class']]  
launch_sites_df = spacex_df.groupby(['Launch Site'], as_index=False).first()  
launch_sites_df = launch_sites_df[['Launch Site', 'Lat', 'Long']]  
launch_sites_df
```

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

- Above coordinates are just plain numbers that can not give you any intuitive insights about where are those launch sites. If you are very good at geography, you can interpret those numbers directly in your mind. If not, that's fine too. Let's visualize those locations by pinning them on a map.

# Launch Site Names Begin with 'CCA'

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bludb  
Done.
```

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

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

# Total Payload Mass

---

```
In [6]: %sql select sum(payload_mass_kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

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

# Average Payload Mass by F9 v1.1

---

```
In [6]: %sql select sum(payload_mass_kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[6]: total_payload_mass
        45596
```

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

# First Successful Ground Landing Date

---

```
In [7]: %sql select avg(payload_mass_kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.  
Out[7]:

| average_payload_mass |
|----------------------|
| 2534                 |


```

- Listing the date when the first successful landing outcome in ground pad was achieved

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing_outcome = 'Success (drone ship)' and payload_mass_kg_ between 4000 and 6000;  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

Out[9]:

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

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

# Total Number of Successful and Failure Mission Outcomes

---

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Listing the total number of successful and failure mission outcomes.

# Boosters Carried Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass_kg_ = (select max(payload_mass_kg_) from SPACEXDATASET);
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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

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

# 2015 Launch Records

---

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET  
where landing__outcome = 'Failure (drone ship)' and year(date)=2015;  
  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.  
Out[12]:
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

```
In [13]: %%sql select landing_outcome, count(*) as count_outcomes from SPACEXDATASET  
    where date between '2010-06-04' and '2017-03-20'  
    group by landing_outcome  
    order by count_outcomes desc;  
  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

Out[13]:

landing_outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- 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.

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against a dark blue-black 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 right, the green and yellow glow of the aurora borealis is visible. The atmosphere of the Earth is thin and hazy, appearing as a light blue band near the horizon.

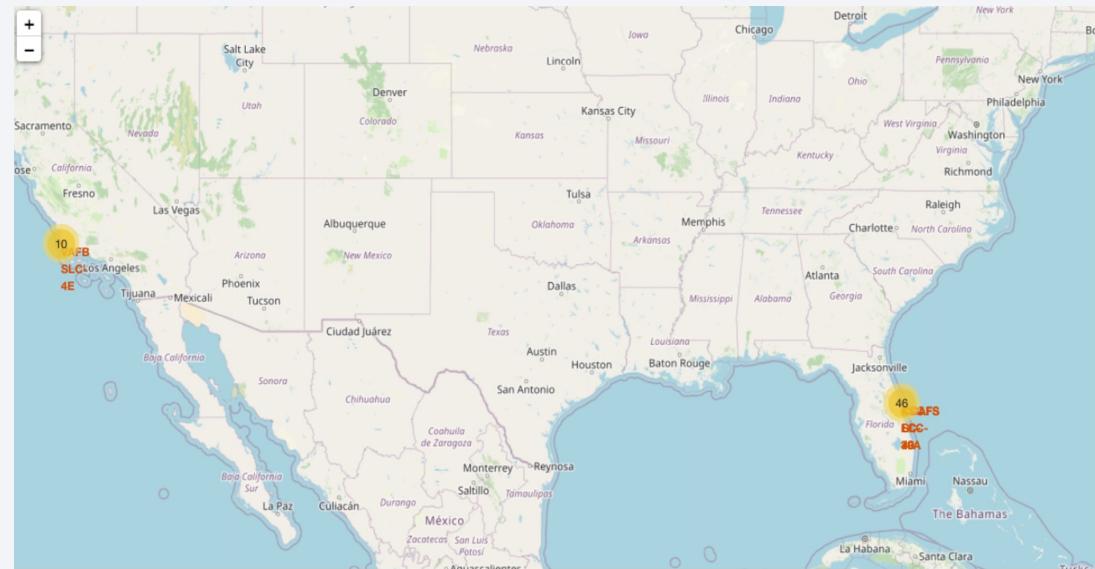
Section 3

# Launch Sites Proximities Analysis

# All Launch Sites' Location Markers on a Global Map

---

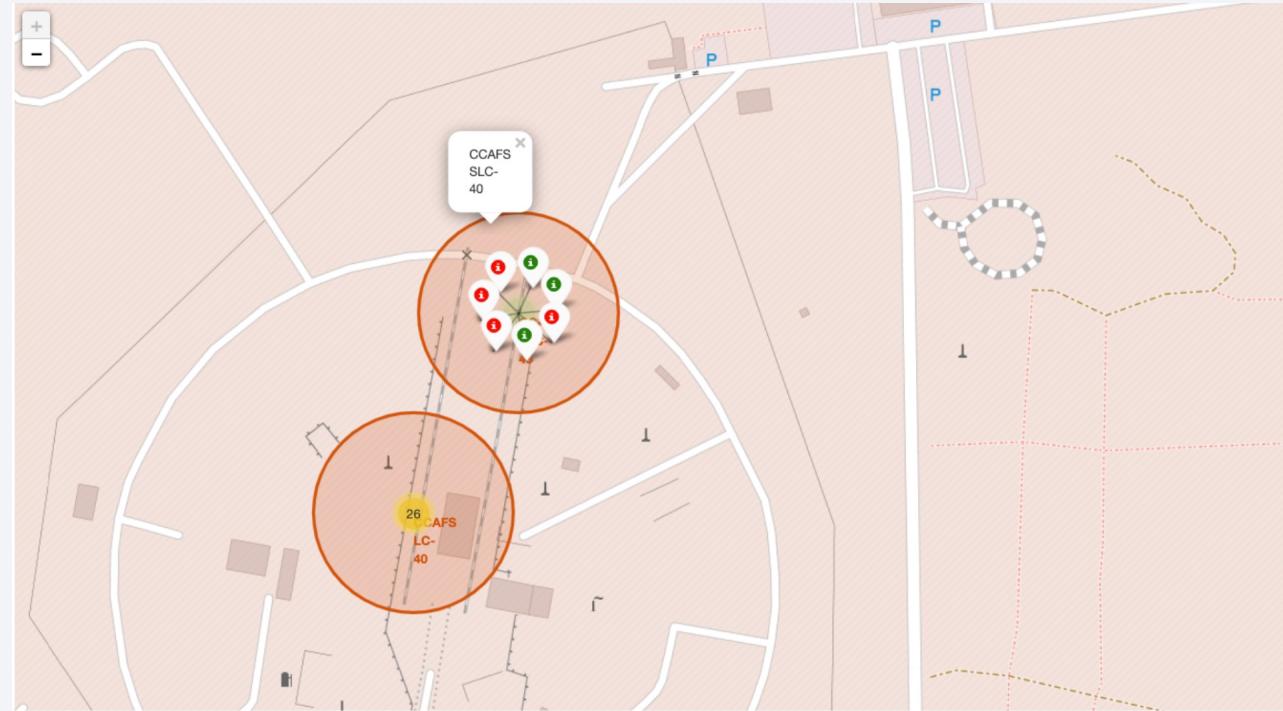
- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, 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

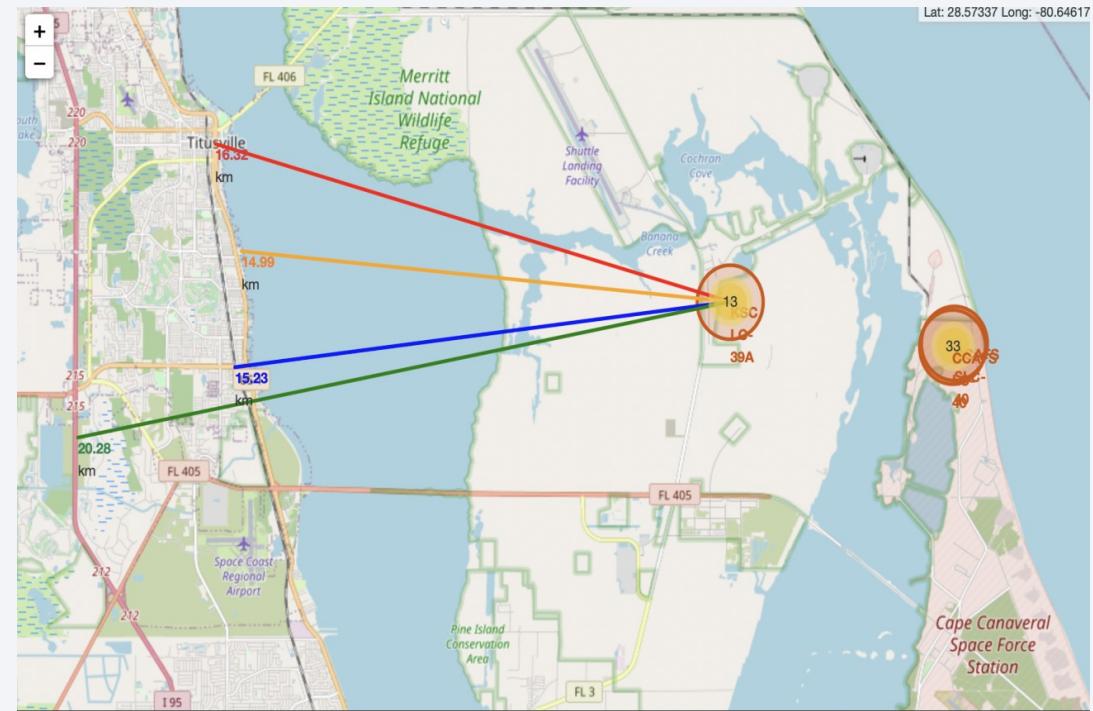
---

- 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.



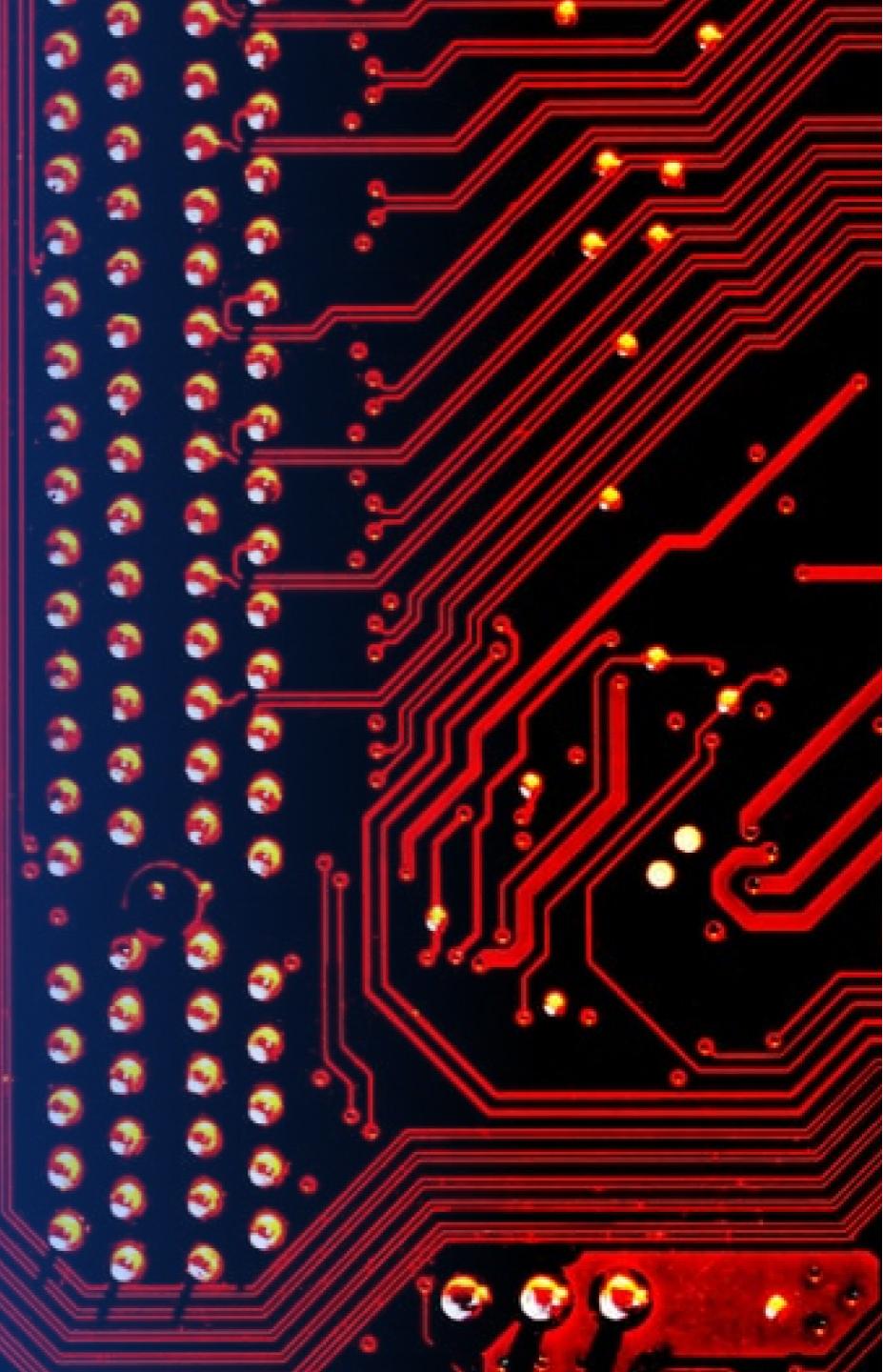
# Distance from the launch site KSC LC-39A to its proximities

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is: - relative close to railway (15.23 km) - relative close to highway (20.28 km) - relative close to coastline (14.99 km).
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

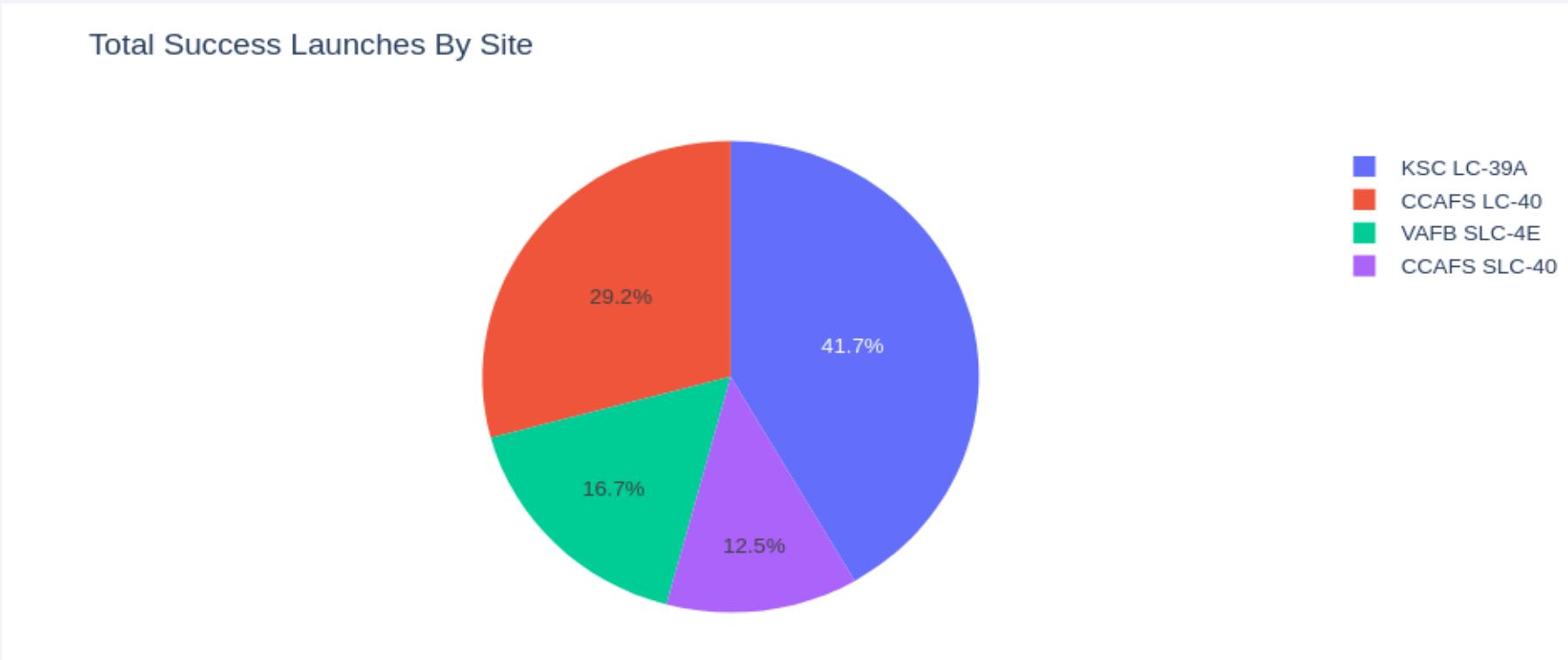


Section 4

# Build a Dashboard with Plotly Dash

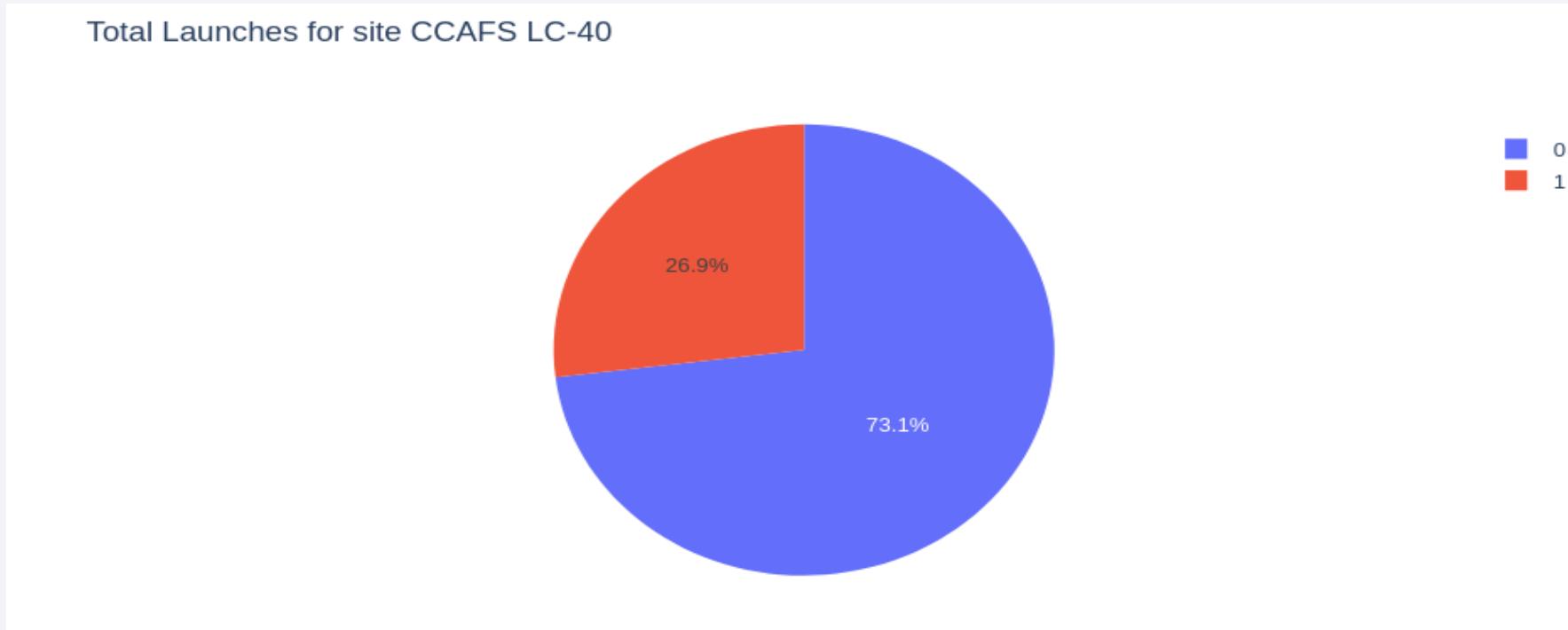


# Launch success count for all sites



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

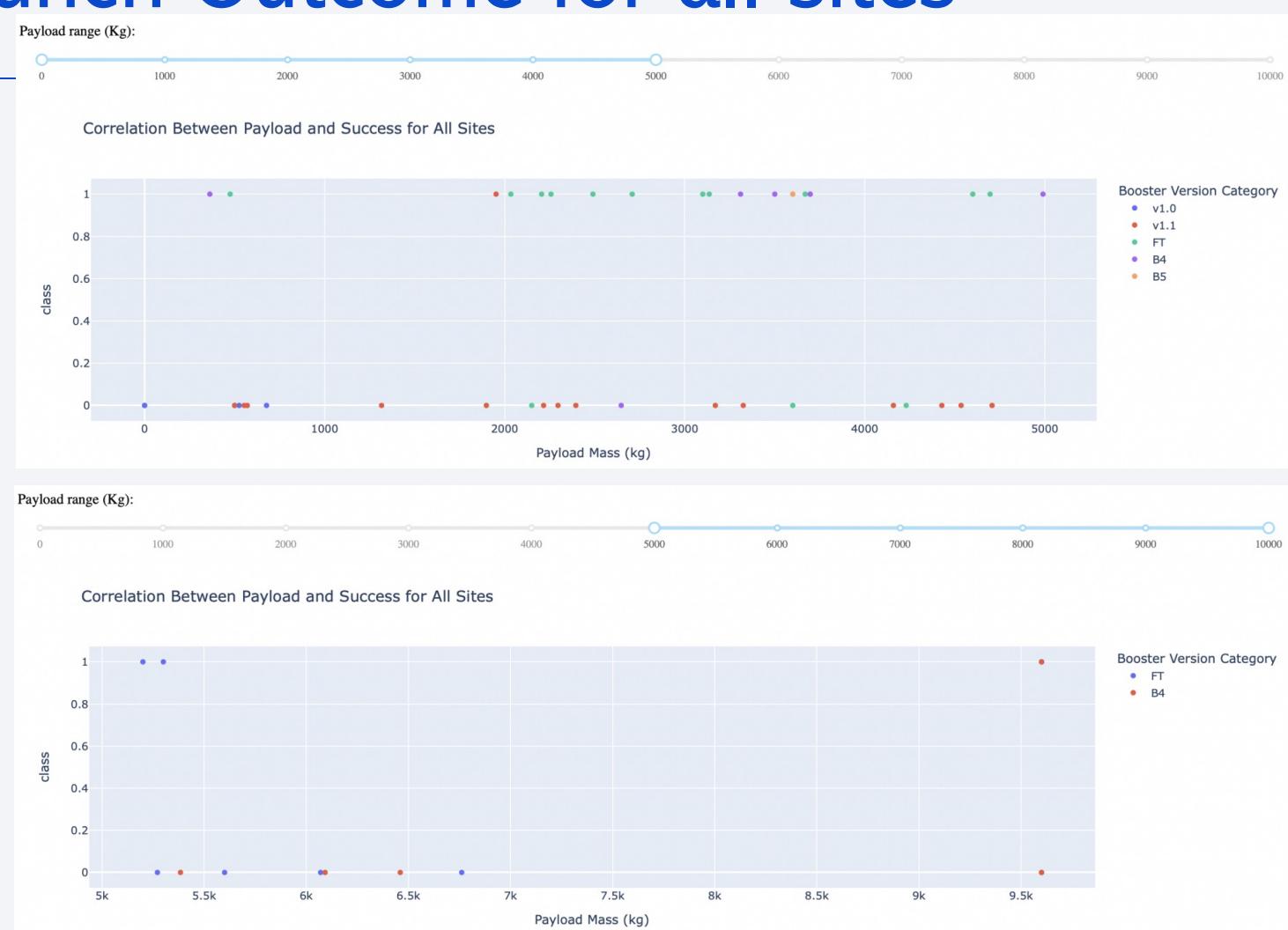
# Launch site with highest launch success ratio



- KSC LC-39A has the highest launch success rate (73.1%) 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



The background of the slide features a dynamic, abstract design. It consists of several thick, curved lines that transition from a bright yellow 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 road. The overall effect is modern and professional.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

---

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

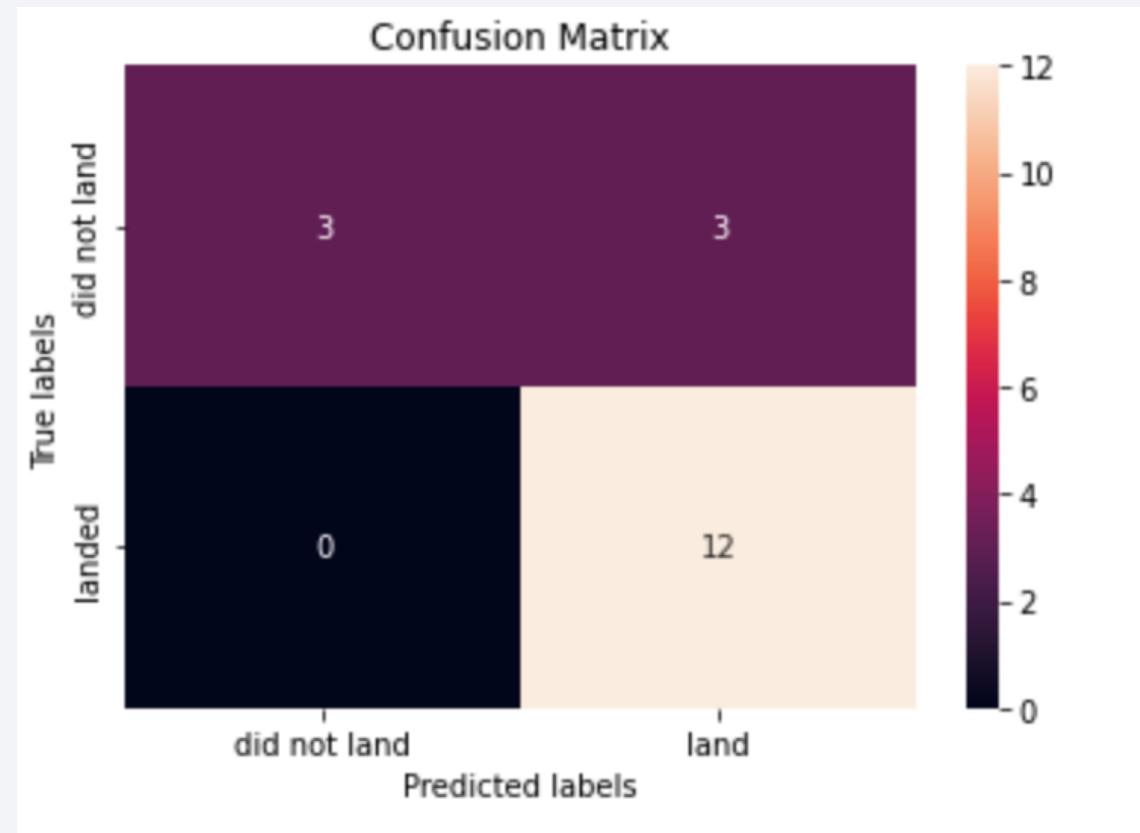
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

# 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.
- 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

---

- 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!

