

Winning Space Race with Data Science

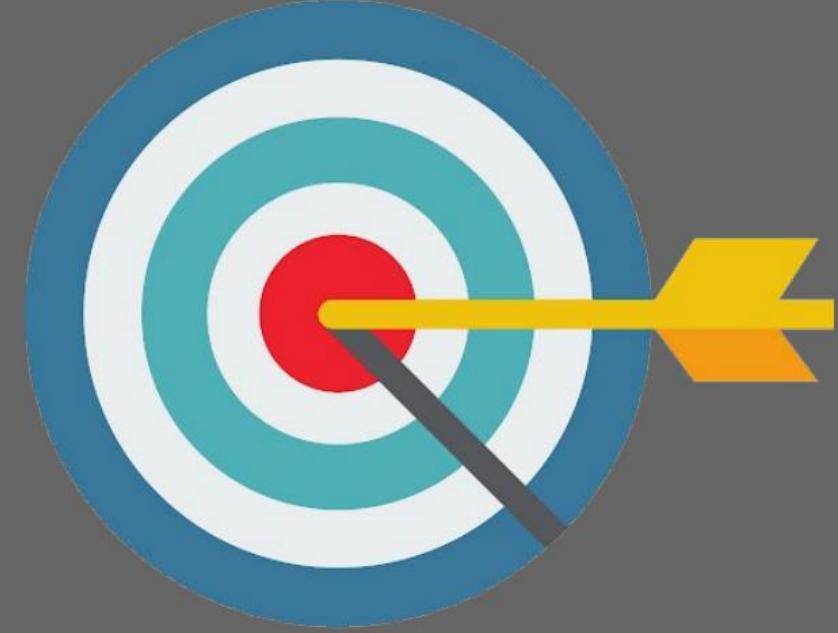
Abhinav Kumar
15 / 07 / 2023



Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

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 Dash
 - Predictive analysis (Classification)
- **Summary of all results**
 - Exploratory data analysis results
 - Interactive analytics demo in screenshots
 - Predictive analysis results

Introduction

- **Project background and context**

We predicted if the Falcon 9 first stage will land successfully. SpaceX 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Common problems that needed solving.**

- What influences if the rocket will land successfully?
- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

- Data collection methodology:
 - SpaceX Rest API
 - (Web Scrapping) from [Wikipedia](#)
- Performed data wrangling (Transforming data for Machine Learning)
 - One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- Performed exploratory data analysis (EDA) using visualization and SQL
 - Plotting : Scatter Graphs, Bar Graphs to show relationships between variables to show patterns of data.
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
 - How to build, tune, evaluate classification models

Data Collection

- The following datasets was collected by
 - We worked with SpaceX launch data that is gathered from the SpaceX REST API.
 - This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
 - Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.
 - The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.

Data Collection – SpaceX API

1 .Getting Response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"  
response = requests.get(spacex_url).json()
```

simplified flow chart

2. Converting Response to a .json file

```
response = requests.get(static_json_url).json()  
data = pd.json_normalize(response)
```

3. Apply custom functions to clean data

```
getLaunchSite(data)  
getPayloadData(data)  
getCoreData(data)
```

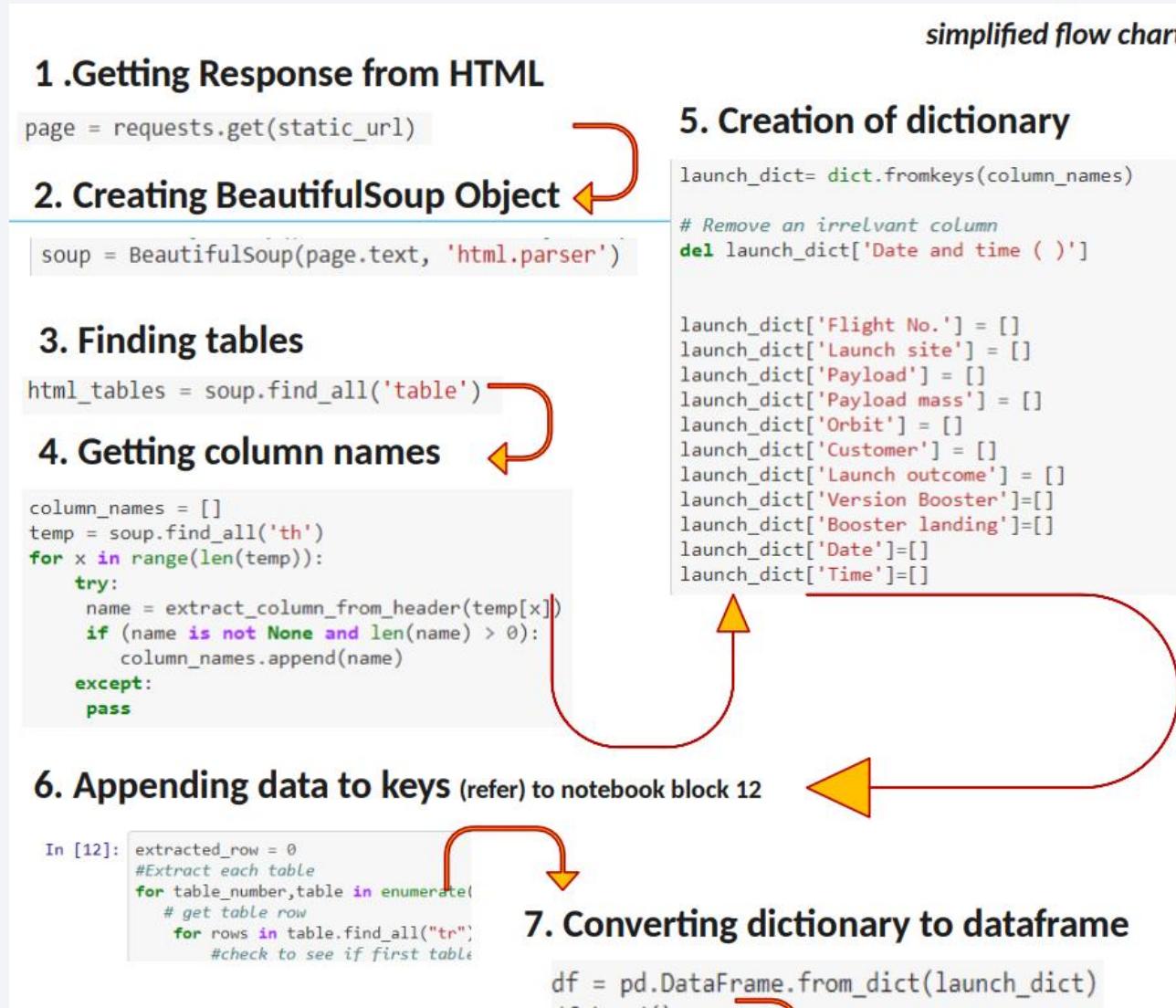
```
getBoosterVersion(data)
```

4. Assign list to dictionary then dataframe

```
launch_dict = {'FlightNumber': list(data['flight_number']),  
               'Date': list(data['date']),  
               'BoosterVersion':BoosterVersion,  
               'PayloadMass':PayloadMass,  
               'Orbit':Orbit,  
               'LaunchSite':LaunchSite,  
               'Outcome':Outcome,  
               'Flights':Flights,  
               'GridFins':GridFins,  
               'Reused':Reused,  
               'Legs':Legs,  
               'LandingPad':LandingPad,  
               'Block':Block,  
               'ReusedCount':ReusedCount,  
               'Serial':Serial,  
               'Longitude': Longitude,  
               'Latitude': Latitude}
```

```
df = pd.DataFrame.from_dict(launch_dict)
```

Data Collection - Scraping



Data Wrangling

Introduction

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.

Process

Perform Exploratory Data Analysis EDA on dataset

Calculate the number of launches at each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Export dataset as .CSV

Create a landing outcome label from Outcome column

Work out success rate for every landing in dataset

EDA with Data Visualization

Scatter Graphs being drawn:

Flight Number VS. Payload Mass

Flight Number VS. Launch Site

Payload VS. Launch Site

Orbit VS. Flight Number Payload

Bar Graph being drawn:

Mean VS. Orbit



A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.

EDA with SQL

Performed SQL queries to gather information about the dataset.

For example of some questions we were asked about the data we needed information about. Which we are using SQL queries to get the answers in the dataset :

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC'
- 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 where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad 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 records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Build an Interactive Map with Folium

To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a *Circle Marker around each launch site with a label of the name of the launch site.*

We assigned the dataframe `launch_outcomes(failures, successes)` to classes *0 and 1* with Green and Red markers on the map in a `MarkerCluster()`

Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks

Build a Dashboard with Plotly Dash

Used Python Anywhere to host the website live 24/7 so you can play around with the data and view the data

- The dashboard is built with Flask and Dash web framework.

Graphs

- Pie Chart showing the total launches by a certain site/all sites
 - display relative proportions of multiple classes of data.
 - size of the circle can be made proportional to the total quantity it represents.

Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions

- It shows the relationship between two variables.
- It is the best method to show you a non-linear pattern.
- The range of data flow, i.e. maximum and minimum value, can be determined.
- Observation and reading are straightforward.

Predictive Analysis (Classification)

BUILDING MODEL

- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.

EVALUATING MODEL

- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix

IMPROVING MODEL

- Feature Engineering
- Algorithm Tuning

FINDING THE BEST PERFORMING CLASSIFICATION MODEL

- The model with the best accuracy score wins the best performing model
- In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

Results

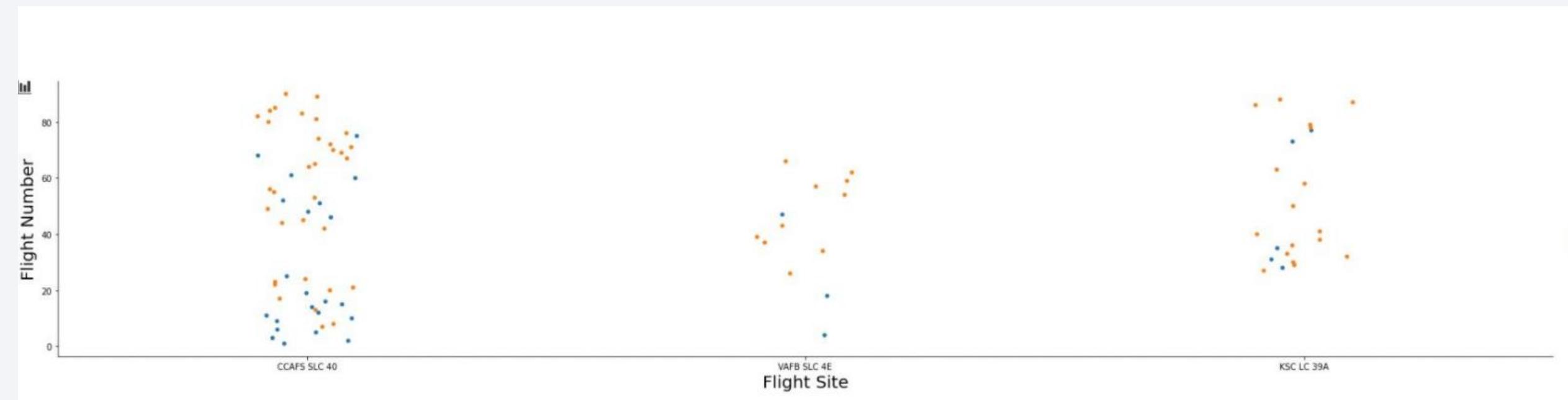
- **Exploratory data analysis results**
- **Interactive analytics demo in screenshots**
- **Predictive analysis results**

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

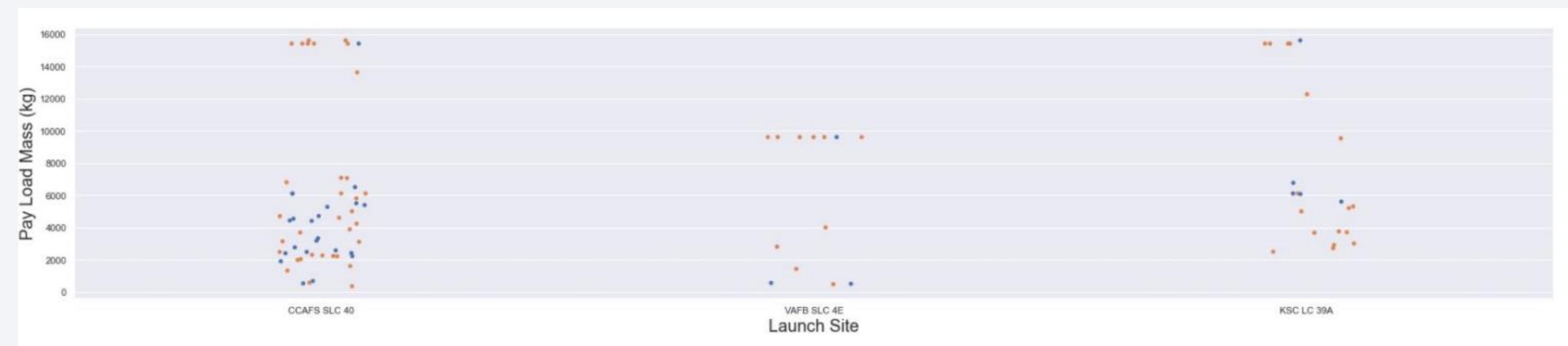
Section 2

Insights drawn from EDA

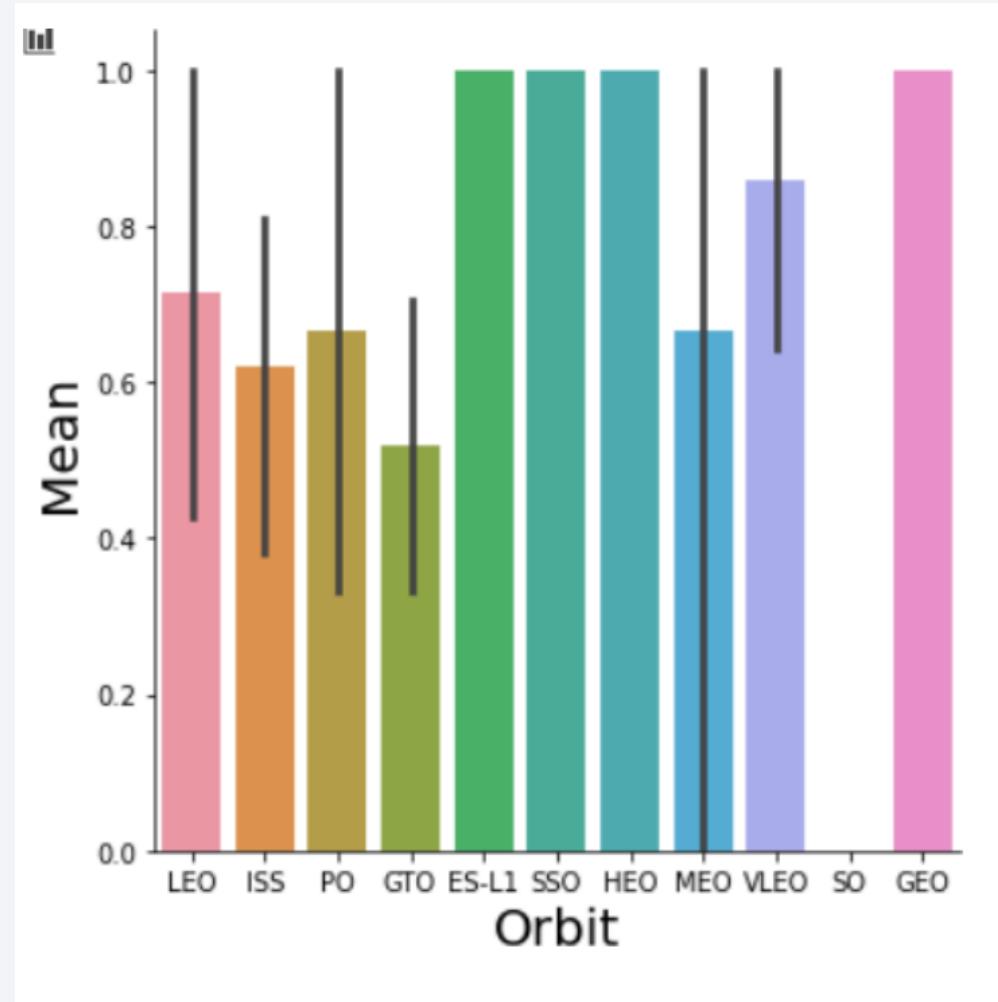
Flight Number vs. Launch Site



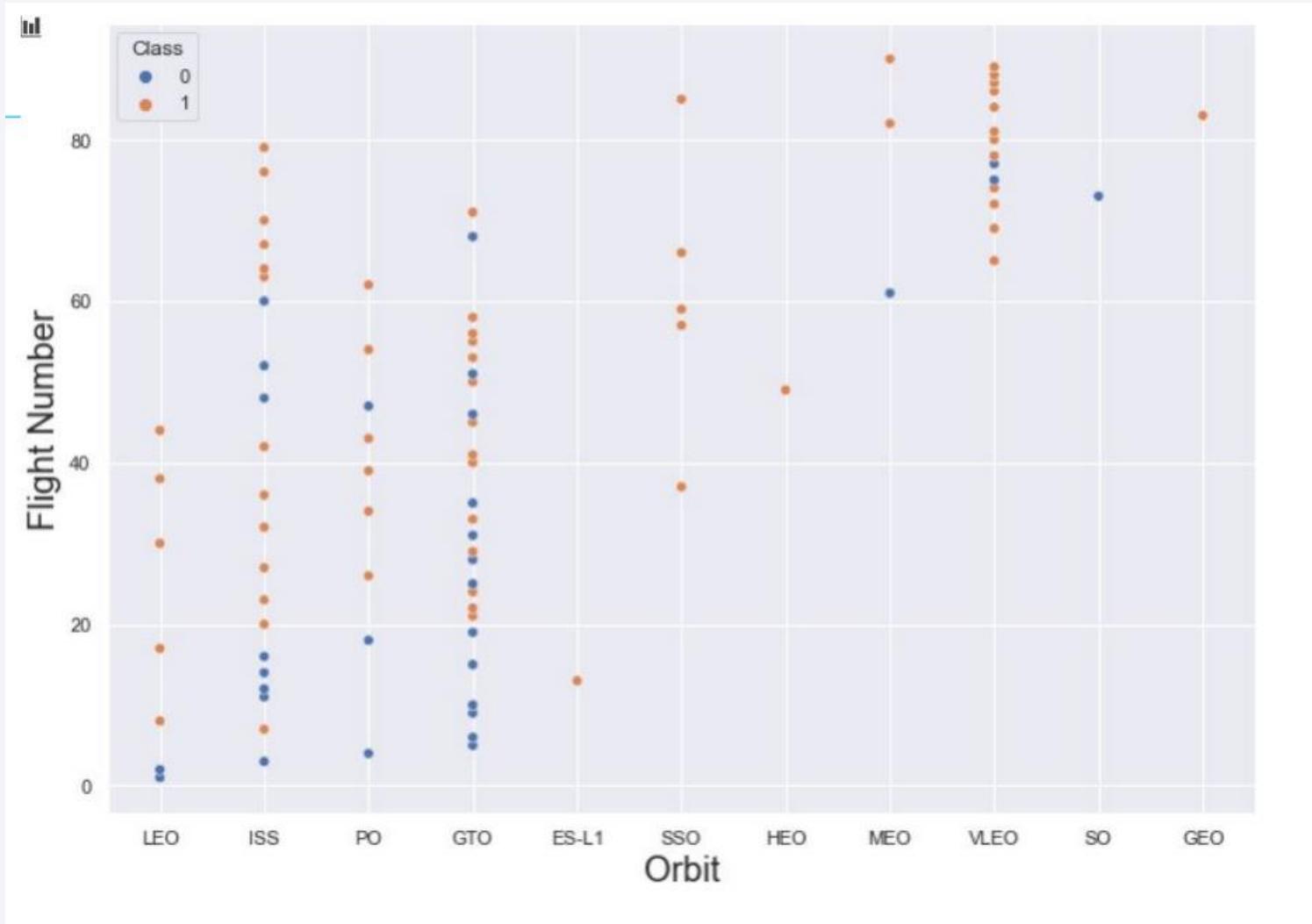
Payload vs. Launch Site



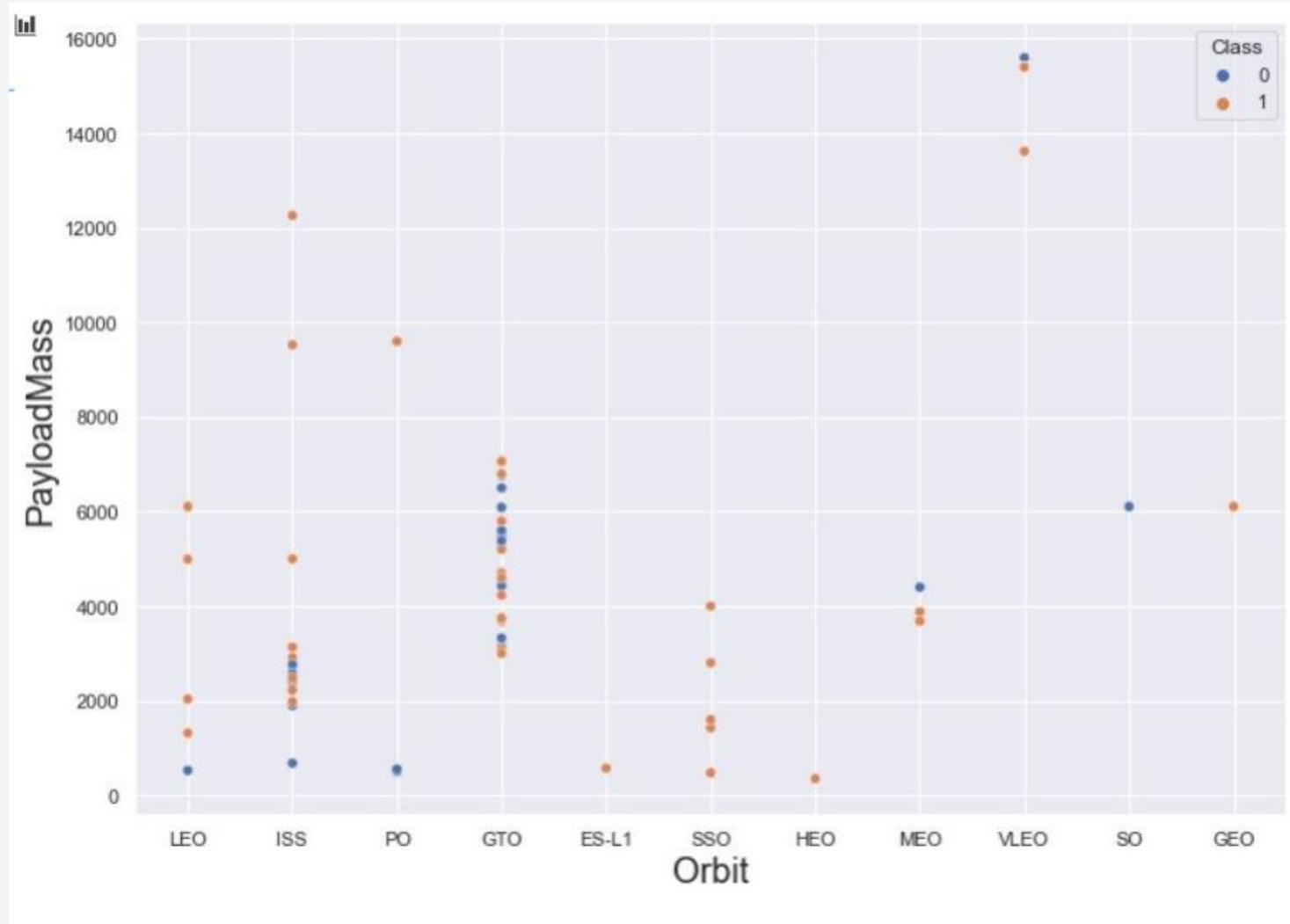
Success Rate vs. Orbit Type



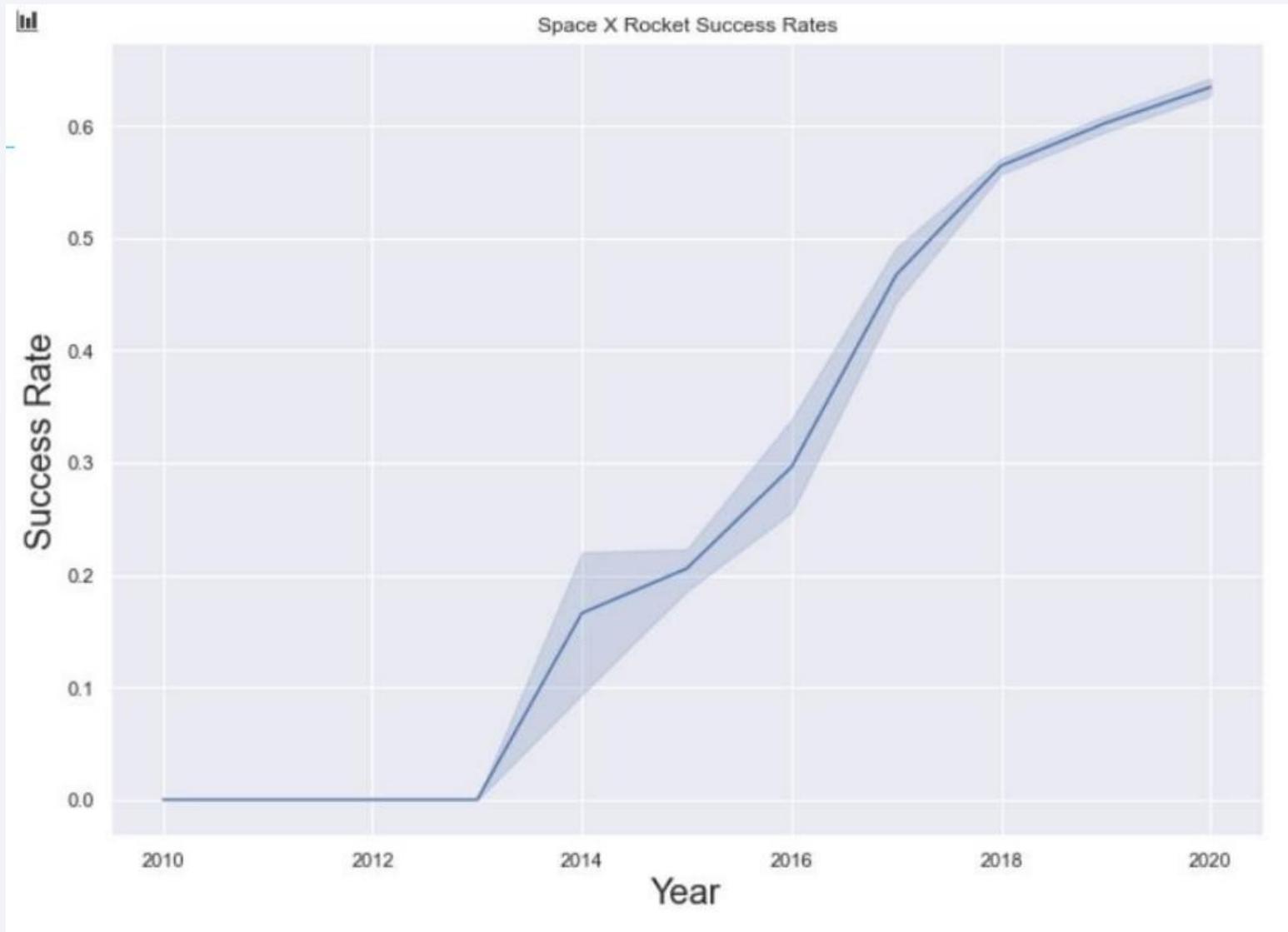
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

SQL QUERY

```
select DISTINCT  
Launch_Site  
from tb SpaceX
```

QUERY EXPLANATION

Using the word **DISTINCT** in the query means that it will only show Unique values in the **Launch_Site** column from **tb SpaceX**



Unique Launch Sites

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

Launch Site Names Begin with 'CCA'

SQL QUERY

```
select TOP 5 * from tblSpaceX  
WHERE Launch_Site LIKE 'KSC'  
%
```



QUERY EXPLANATION

Using the word **TOP 5** in the query means that it will only show 5 records from **tblSpaceX** and **LIKE** keyword has a wild card with the words '**KSC%**' the percentage in the end suggests that the Launch_Site name must start with KSC.

	Date	Time_UTC	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	19-02-2017	2021-07-02 14:39:00.000000	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
1	16-03-2017	2021-07-02 06:00:00.000000	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2	30-03-2017	2021-07-02 22:27:00.000000	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
3	01-05-2017	2021-07-02 11:15:00.000000	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
4	15-05-2017	2021-07-02 23:21:00.000000	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

```
select SUM(PAYLOAD_MASS_KG) TotalPayloadMass from tblSpaceX  
where Customer = 'NASA (CRS)';TotalPayloadMass
```



Total Payload Mass	
0	45596

QUERY EXPLANATION

Using the function **SUM** summates the total in the column **PAYLOAD_MASS_KG**

The **WHERE** clause filters the dataset to only perform calculations on **Customer NASA (CRS)**

Average Payload Mass by F9 v1.1

```
select AVG(PAYLOAD_MASS_KG) AveragePayloadMass from tblSpaceX  
where Booster_Version = 'F9 v1.1'
```



Average Payload Mass	
0	2928

QUERY EXPLANATION

Using the function **AVG** works out the average in the column **PAYLOAD_MASS_KG**

The **WHERE** clause filters the dataset to only perform calculations on **Booster_version F9 v1.1**

First Successful Ground Landing Date

```
select MIN(Date) SLO from tbSpaceX where Landing_Outcome = "Success (drone ship)"
```



QUERY EXPLANATION

Using the function **MIN** works out the minimum date in the column **Date**

The **WHERE** clause filters the dataset to only perform calculations on **Landing_Outcome Success (drone ship)**

Date which first Successful landing outcome in drone ship was achieved.	
0	06-05-2016

Successful Drone Ship Landing with Payload between 4000 and 6000

```
select Booster_Version from tbSpaceX where Landing_Outcome = 'Success (ground pad)'  
AND Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000
```



Date which first Successful landing outcome in drone ship was achieved.	
0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

QUERY EXPLANATION

Selecting only *Booster_Version*

The **WHERE** clause filters the dataset to *Landing_Outcome = Success (drone ship)*

The **AND** clause specifies additional filter conditions
Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

Total Number of Successful and Failure Mission Outcomes

```
SELECT(SELECT Count(Mission_Outcome) from tbISpaceX where Mission_Outcome  
LIKE '%Success%') as Successful_Mission_Outcomes,  
(SELECT Count(Mission_Outcome) from tbISpaceX where Mission_Outcome LIKE  
'%Failure%') as Failure_Mission_Coutcomes
```



Successful_Mission_Outcomes	Failure_Mission_Outcomes
0	100

QUERY EXPLANATION

a much harder query I must say, we used subqueries here to produce the results. The **LIKE '%foo%'** wildcard shows that in the record the **foo** phrase is in any part of the string in the records for example.

Boosters Carried Maximum Payload

SQL QUERY

```
SELECT DISTINCT Booster_Version,  
MAX(PAYLOAD_MASS  
_KG_) AS [Maximum Payload Mass]  
FROM tbISpaceX GROUP BY Booster_Version ORDER  
BY [Maximum Payload Mass] DESC
```

QUERY EXPLANATION

Using the word **DISTINCT** in the query means that it will only show Unique values in the **Booster_Version** column from **tblSpaceX**.
GROUP BY puts the list in order set to a certain condition.
DESC means its arranging the dataset into descending order



	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0
97 rows x 2 columns		

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

```
SELECT COUNT(Landing_Outcome)  
FROM    tblSpaceX  
WHERE (Landing_Outcome LIKE '%Success%')  
AND (Date > '04-06-2010') AND (Date < '20-  
03-2017')
```

QUERY EXPLANATION

Function **COUNT** counts records in column
WHERE filters data

LIKE (wildcard) AND (conditions) AND (conditions)



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

0

34

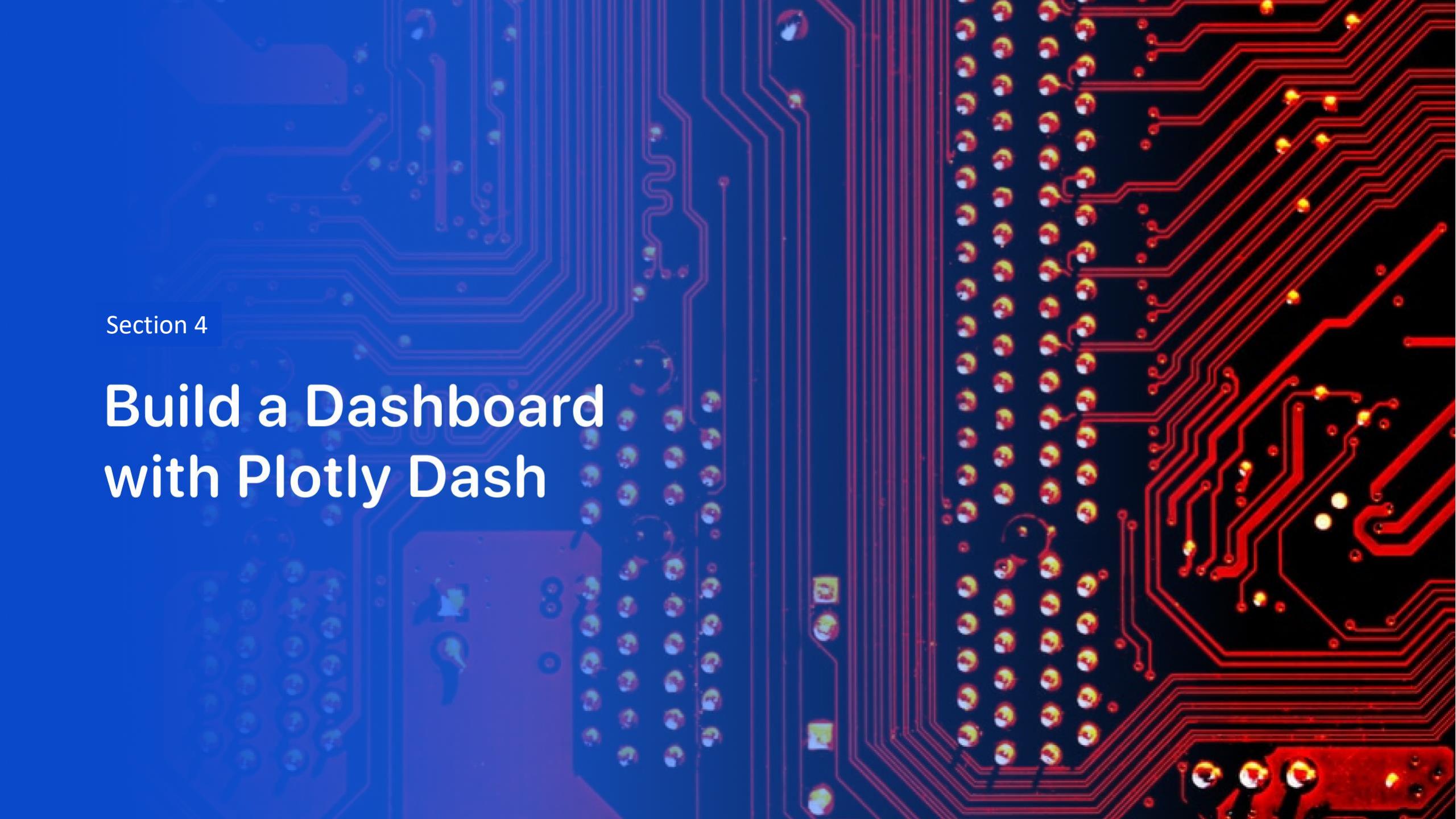
The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth's horizon against a dark blue sky. Numerous glowing yellow and white points represent city lights, concentrated in coastal and urban areas. In the upper right quadrant, there are bright green and yellow bands of light, likely the Aurora Borealis or Australis. The overall atmosphere is dark and mysterious.

Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on a global map



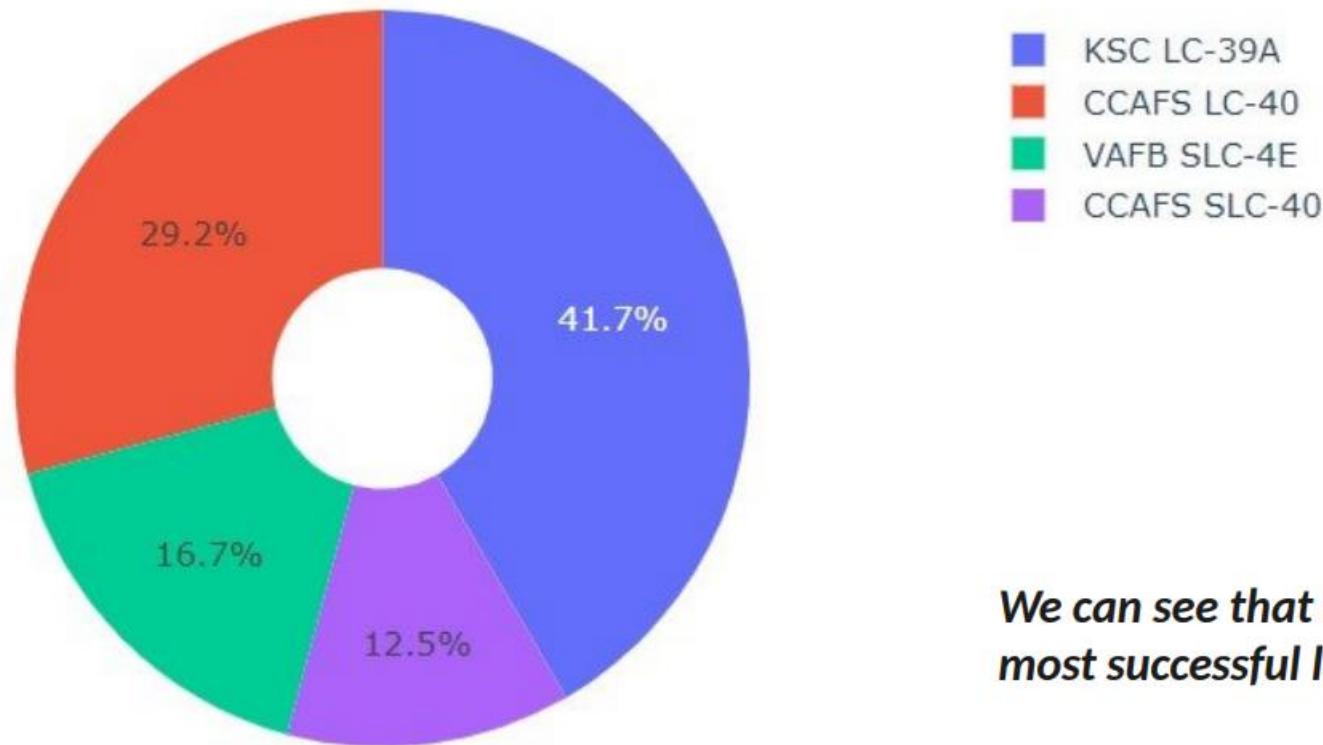


Section 4

Build a Dashboard with Plotly Dash

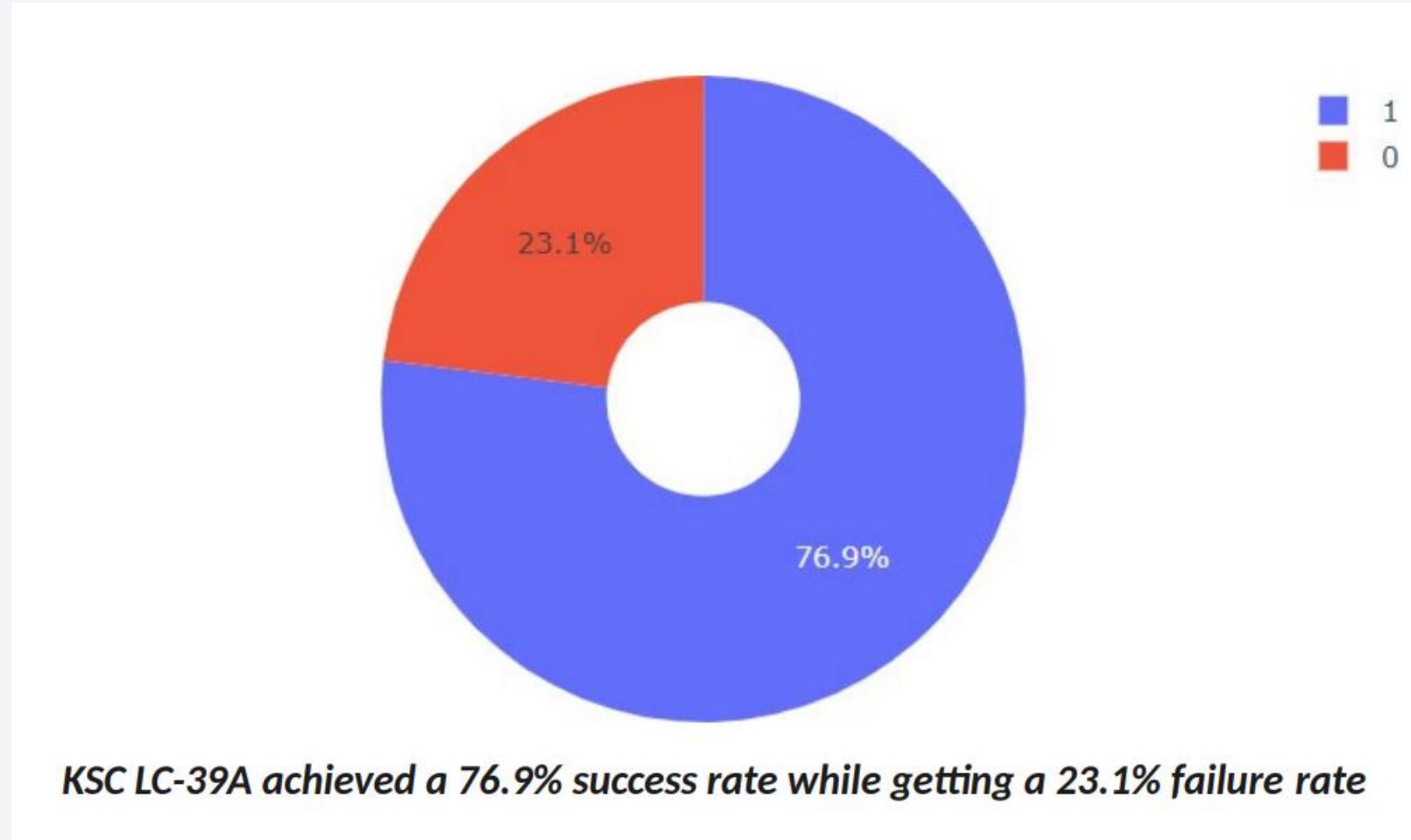
Launch success count

Total Success Launches By all sites



We can see that KSC LC-39A had the most successful launches from all the sites

Piechart for the launch site



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 landscape. The overall effect is modern and professional.

Section 5

Predictive Analysis (Classification)

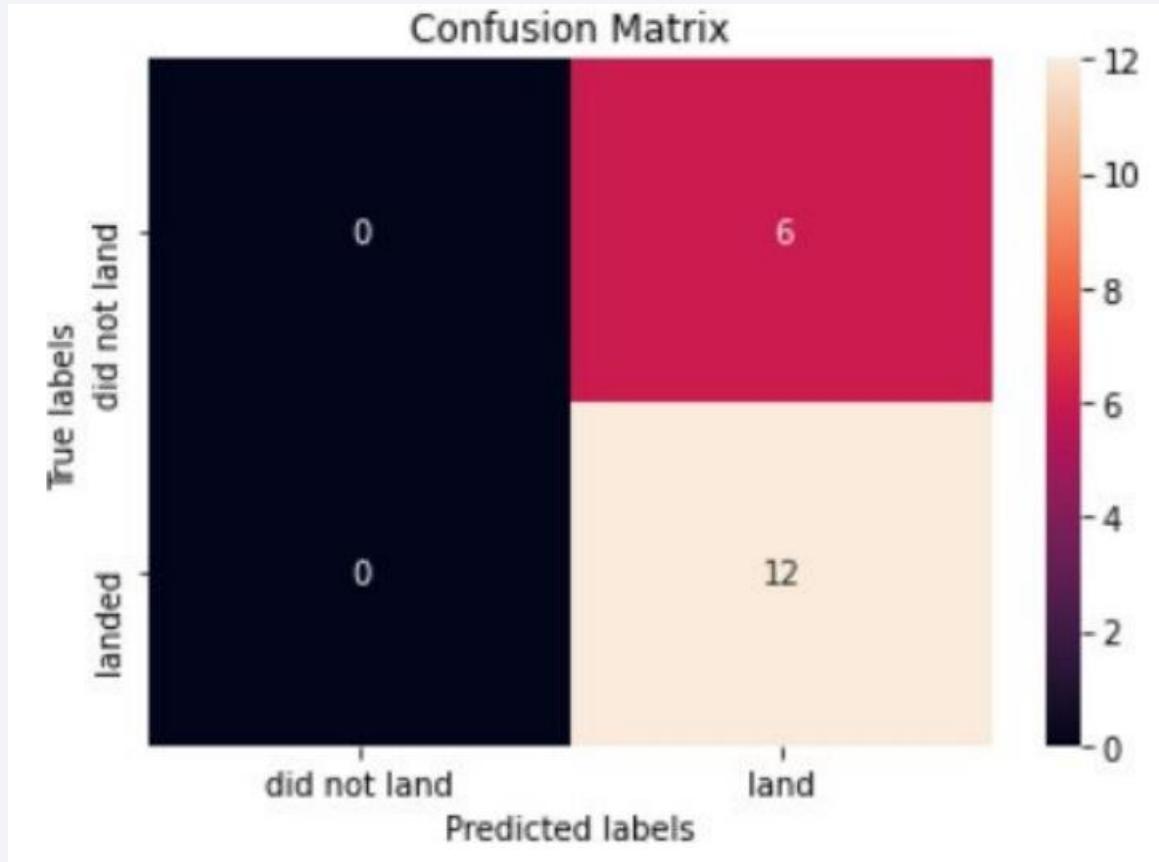
Classification Accuracy

As you can see our accuracy is extremely close but we do have a winner its down to decimal places! using this function

```
bestalgorithm = max(algorithms, key=algorithms.get)
```

	Accuracy	Algorithm
0	0.653571	KNN
1	0.667857	Tree
2	0.667857	LogisticRegression

Confusion Matrix



Conclusions

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- We can see that KSC LC-39A had the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Thank you!

