Winning Space Race with Data Science

Fatemeh Eslaminasab

Summer 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **Summary of methodologies**
- SpaceX data collection using SpaceX API & Web Scraping
- SpaceX data wrangling
- SapceX expletory data analysis
- SpaceX EDA data visualization using matplotlib & seaborn
- SpaceX launch site analysis with plotly Dash
- SpaceX Machine Learning predictions (Logistic- Svm- Decision tree- KNN)
- **Summary of all results**
- EDA results
- Interactive Visual Analytics and Dashboards
- Predictive Analysis(Classification)

Introduction

- Project background

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.

- Problems you want to find answers

In this project, the goal is to predict if the Falcon 9 first stage will land successfully using data from Falcon 9 rocket launches advertised on its website.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

Via API

- I made a get request to the SpaceX API. A series of helper functions was defined that will help us use the API to extract information using identification numbers in the launch data.
- Finally to make the requested JSON results more consistent, the SpaceX launch data was requested and parsed using the GET request and then decoded the response content as a Json result which was then converted into a Pandas data frame.

Via web scrapping

- Performed web scraping to collect Falcon 9 historical launch records from a Wikipedia. Using BeautifulSoup and request Libraries, I extract the Falcon 9 launch HTML table records, Parsed the table and converted it into a Pandas dataframe.

Data Collection – SpaceX API

- Data collected using SpaceX API then parsed the SpaceX launch data as a Json which was then converted into a Pandas dataframe

- Here is the project GitHub link: the first section is related to data collection SpaceX-Falcon-9-link

1.1. loading data

```
[ ]  spacex_url="https://api.spacexdata.com/v4/launches/past"
     response = requests.get(spacex_url)
     print(response.content)
```

```
[ ]  json_result = response.json()
     data = pd.json_normalize(json_result)
```

```
[ ]  data.head()
```

Data Wrangling

- Filter the dataframe to only include Falcon 9 launches
- Dealing with missing values: replacing them with mean of column:

For the PayloadMass, missing data values were replaced using mean value of column.

```
data_falcon9 = launch_df[launch_df['BoosterVersion'] == 'Falcon 9']
data_falcon9.head()
```

```
payload_mass_mean = data_falcon9['PayloadMass'].mean()
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].replace(np.nan, payload_mass_mean)
```

- Here is the project GitHub link: the second section is related to data wrangling SpaceX-Falcon-9-link

EDA with Data Visualization

- Also performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

- Exploratory Data Analysis using data visualization (Matplotlib, seaborn, etc)

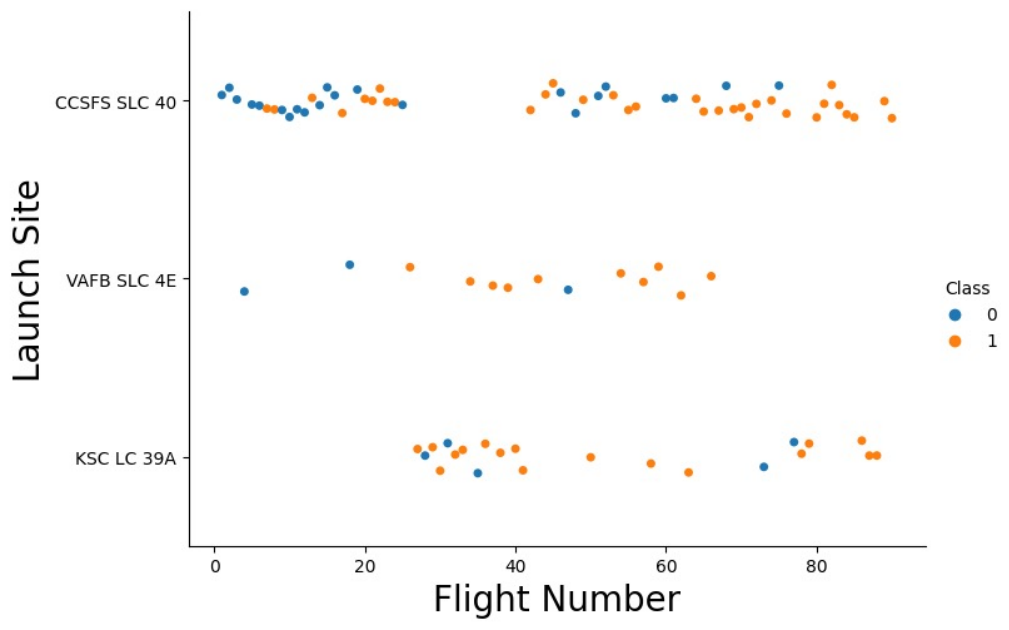
- Preparing Data Feature Engineering

- Here is the project GitHub link: the third and fourth sections is related to Exploratory analysis and Feature engineering SpaceX-Falcon-9-link

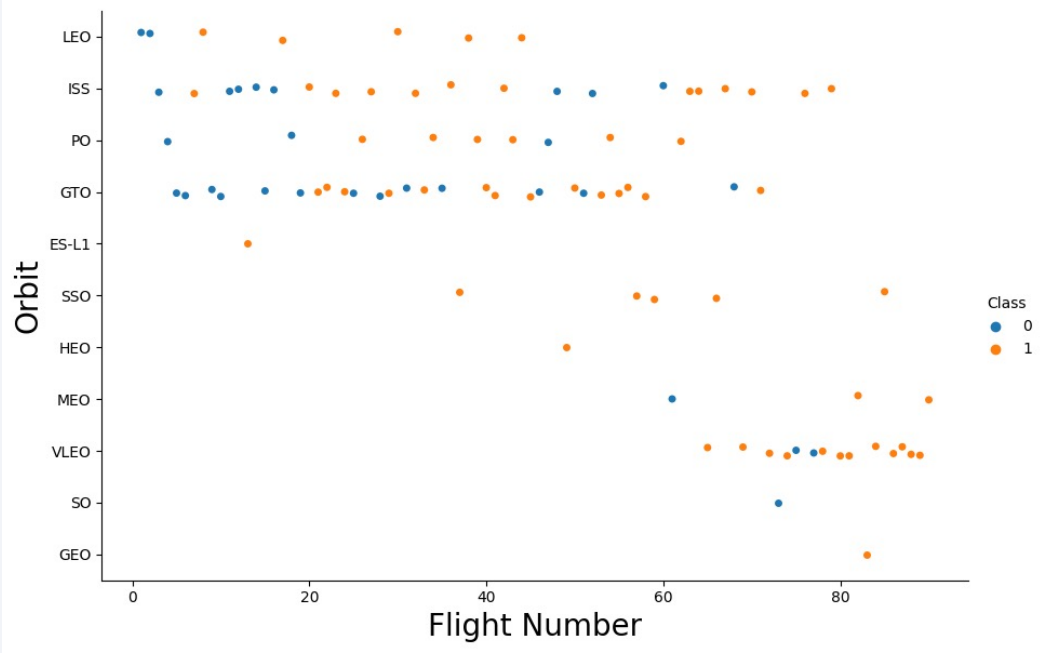
- A few samples of data visualization are shown in following slides

EDA with Data Visualization

- Used scatter plots to Visualize the relationship between Flight Number and Launch Site.

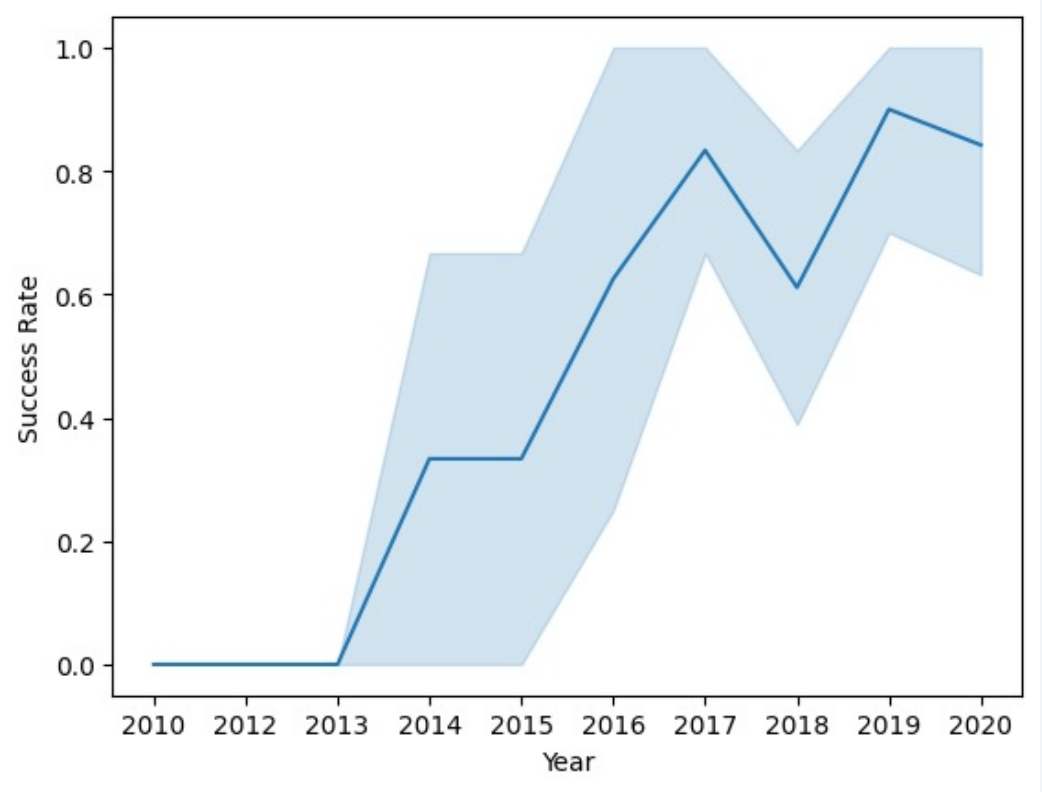


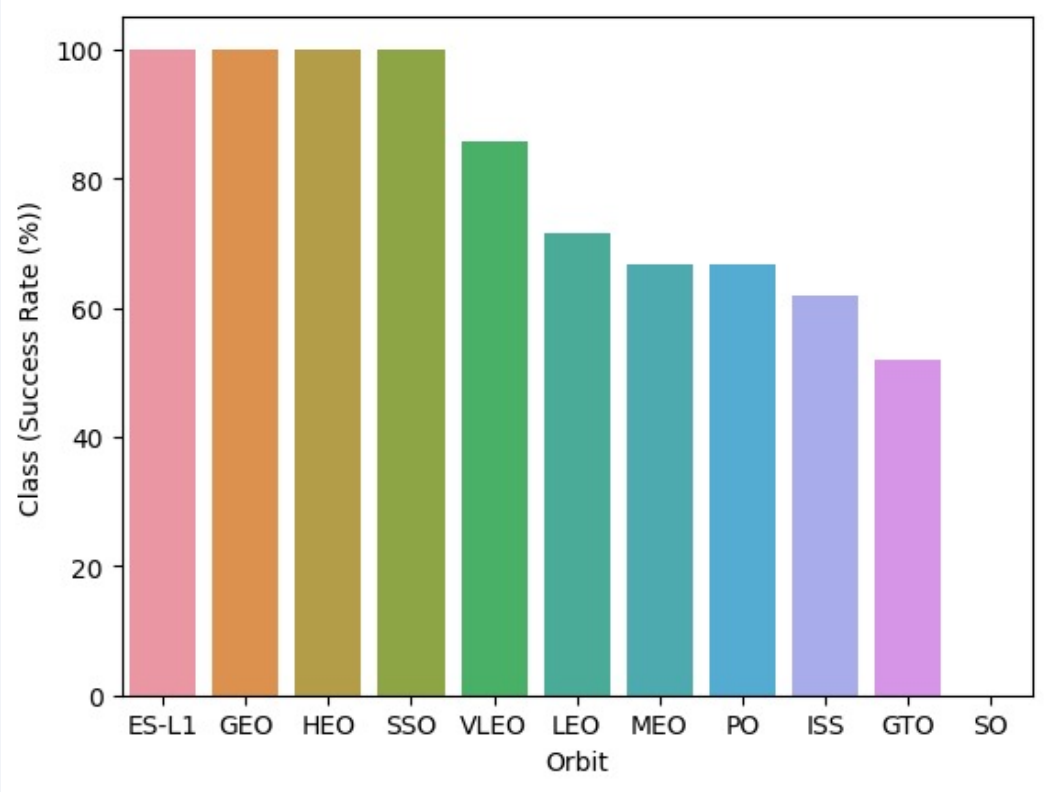
- Used scatter plots FlightNumber and Orbit type



EDA with Data Visualization

- Line plot to Visualize the launch success yearly trend.
- Used Bar chart to Visualize the relationship between success rate of each orbit type





EDA with SQL

- A connection to SQLlite database is created and the SpaceX dataset is then copied in a table in the database
- These are samples of SQL queries were performed for EDA

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

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

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

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

- The total number of successful and failure mission outcomes

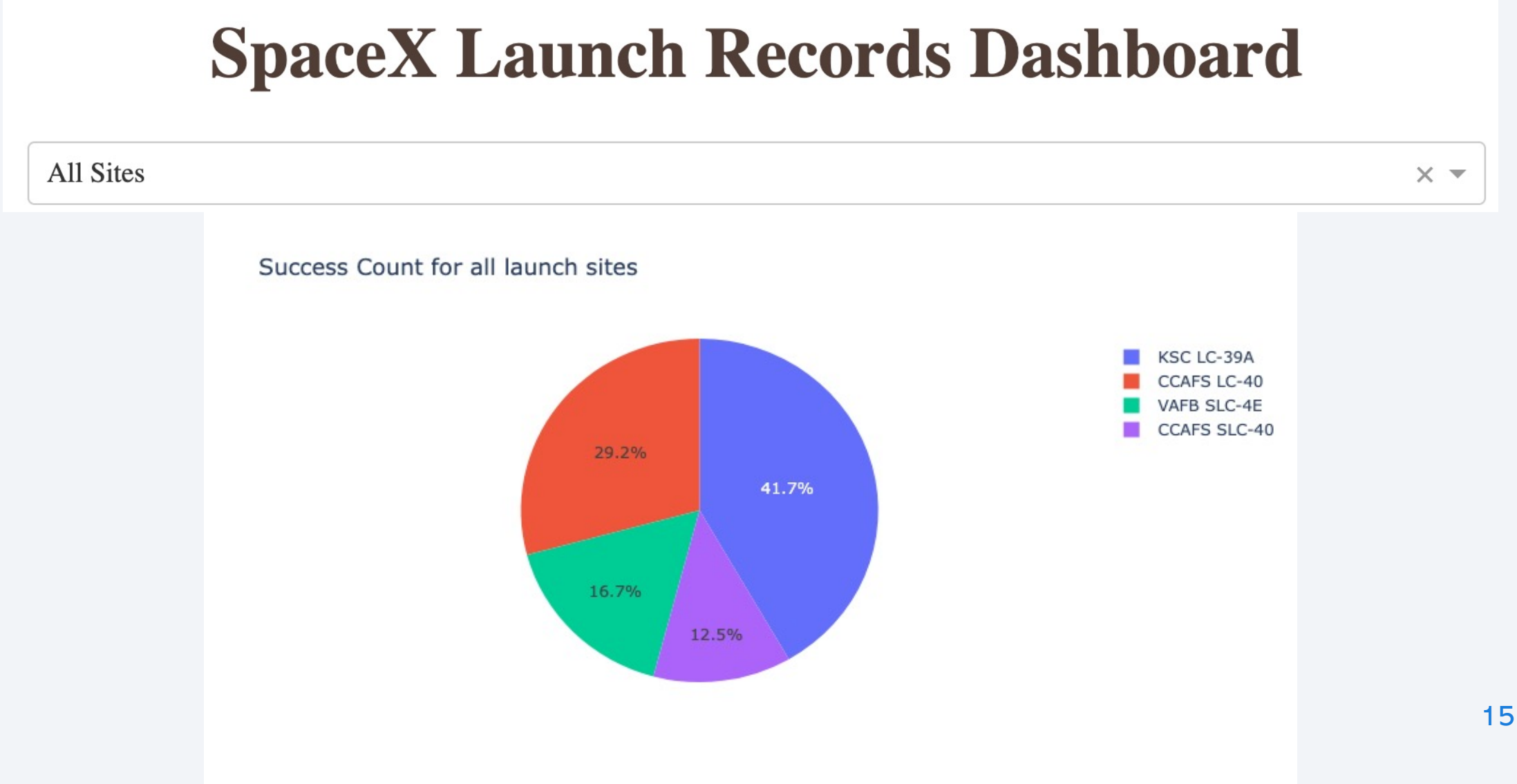
```
%sql SELECT "Mission_Outcome", COUNT(*) AS total_count FROM SPACEXTABLE  \
GROUP BY "Mission_Outcome";
```

- Here is the complete EDA with SQL GitHub link: SQL_spaceX_link

Build a Dashboard with Plotly Dash

- Built an interactive dashboard application with Plotlydash by:
- Adding a Launch Site Drop-down Input Component
- Adding a callback function to render success-pie-chart based on selected site dropdown
- Adding a Range Slider to Select Payload
- Adding a callback function to render the success-payload-scatter-chart scatter plot
- Here is the GitHub URL of your completed PlotlyDash lab: SpaceX_launch_dashboard_link

Build a Dashboard with Plotly Dash



Predictive Analysis (Classification)

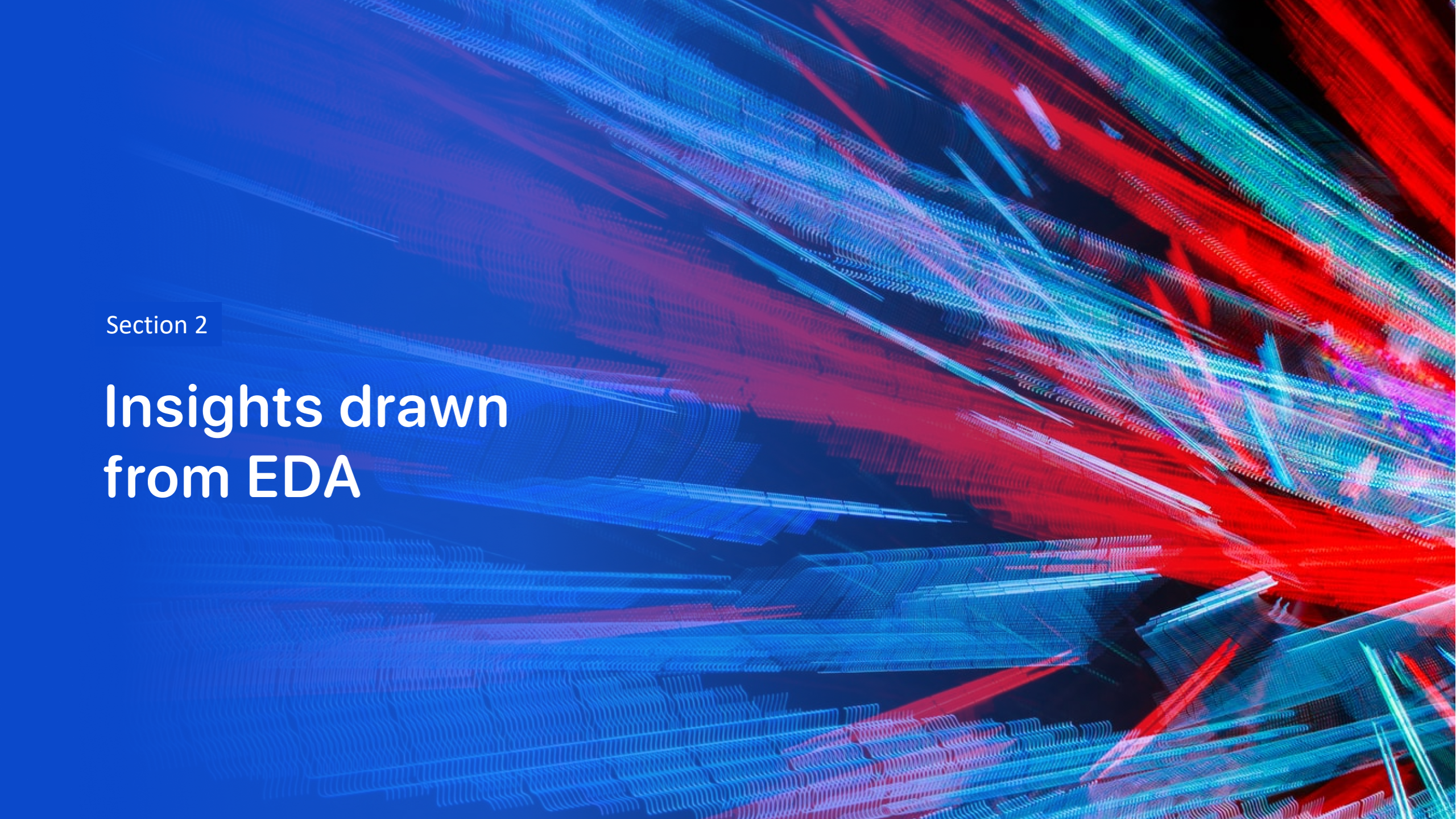
- The goal is to predict if the first stage will land, so we can determine the cost of a launch
- Predictive analysis was done using 4 different method: Logistic regression, Support vector machine, Decision Tree, K nearest neighbors

Steps:

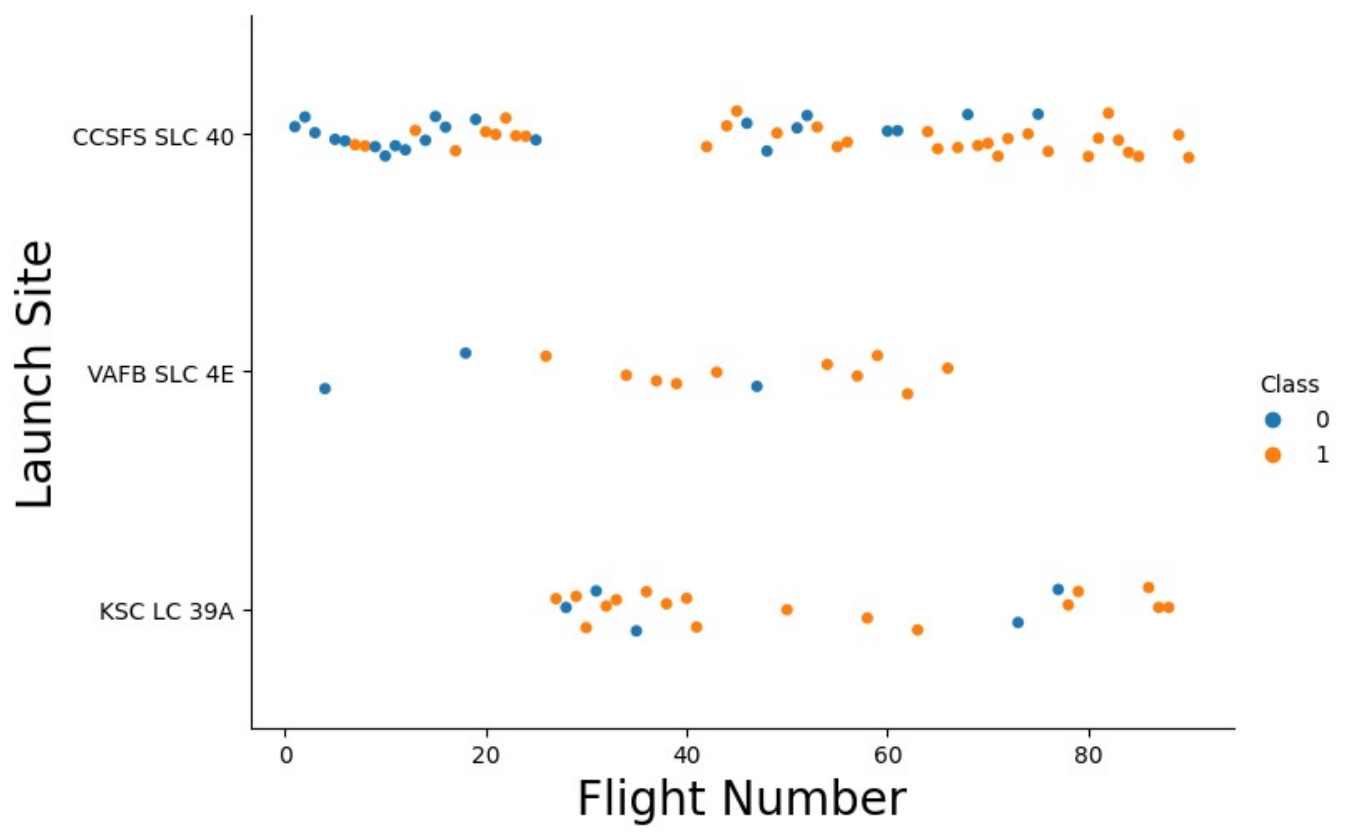
- Perform exploratory Data Analysis and determine Training Labels
- create a column for the class
- Standardize the data
- Split into training data and test data
- Find best Hyperparameter using GridSearch for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data
- Here is the project GitHub link: the last section is related Predictive analysis SpaceX-Falcon-9-link

Results

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

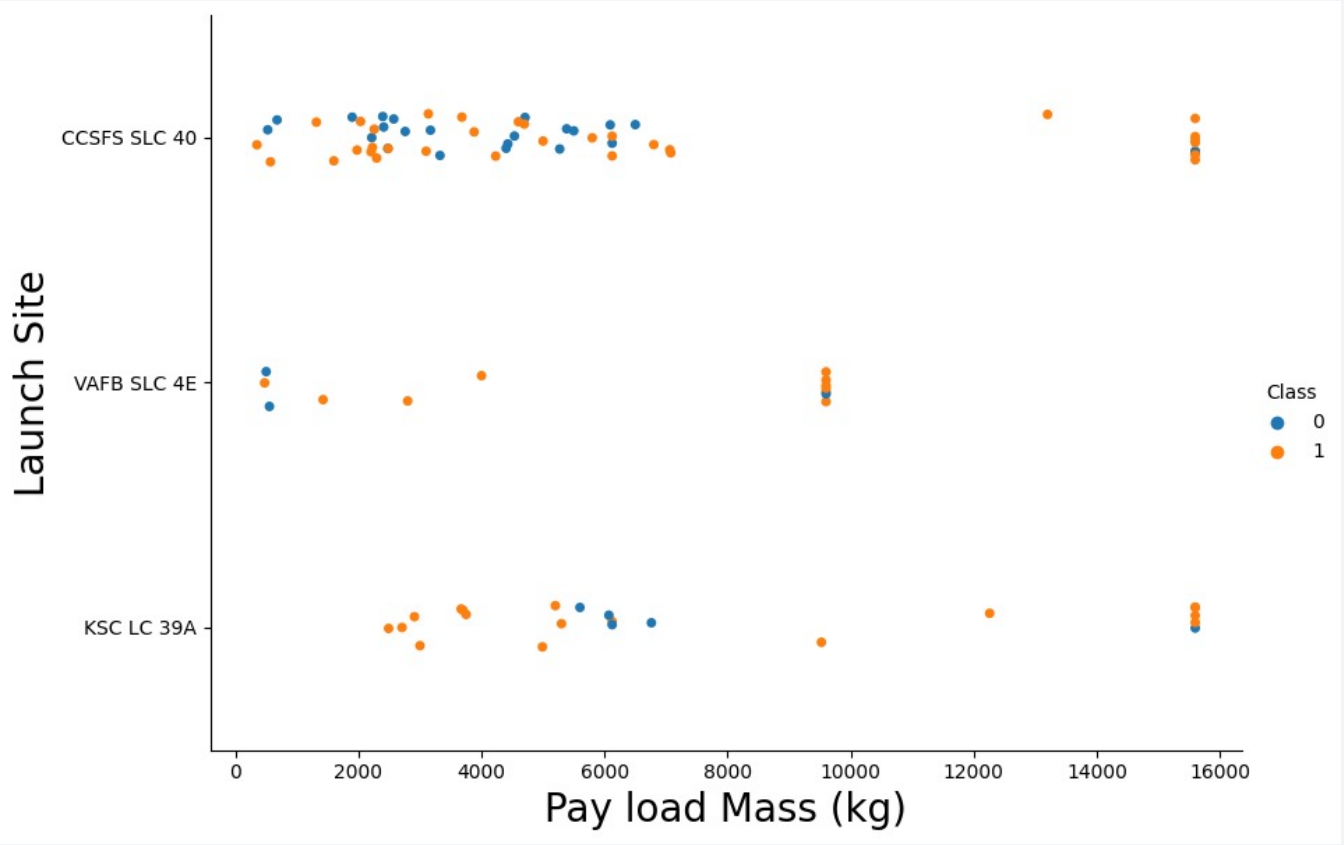

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



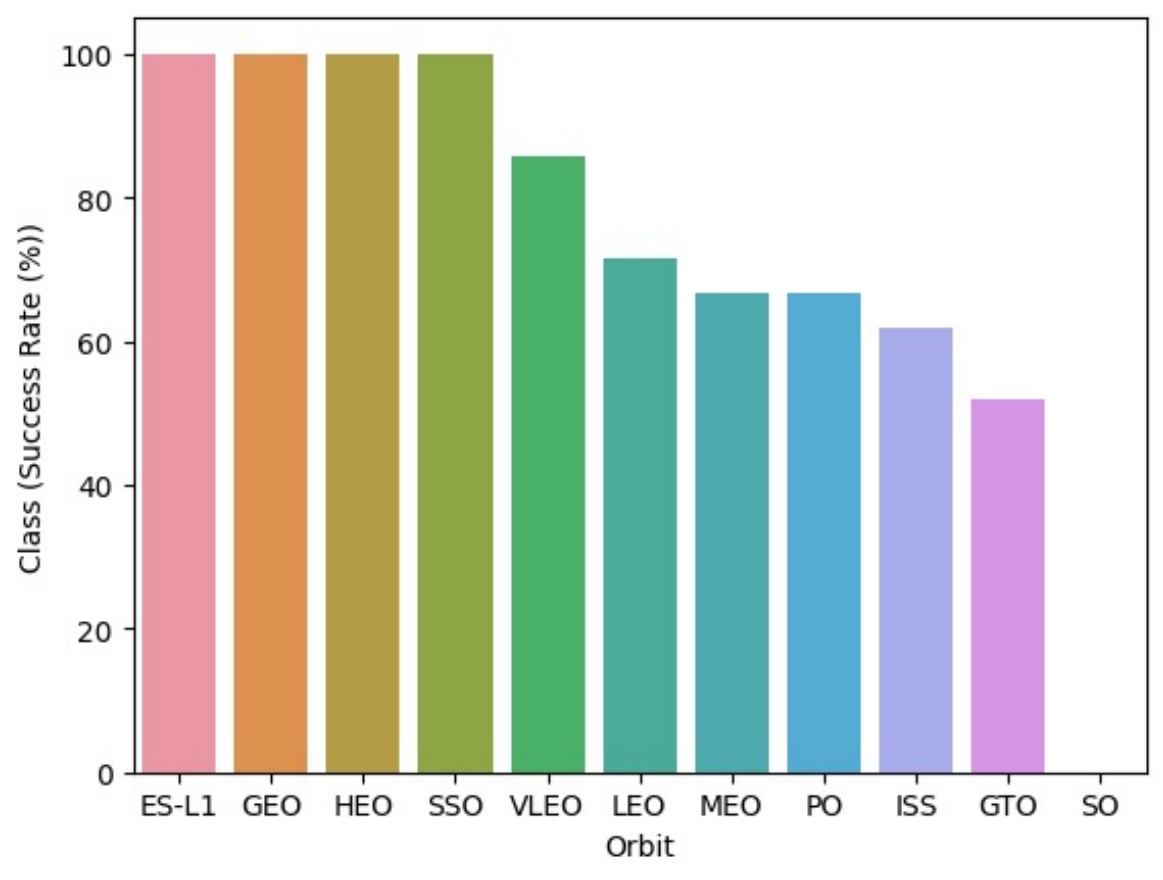
- It can be concluded, as the flight number increases in 3 types of launch sites, so does the success rate. For instance, the success rate for the VAFB SLC 4E launch site is 100% after the Flight number 50. Both KSC LC 39A and CCAFS SLC 40 have a 100% success rates after 80th flight.

Payload vs. Launch Site



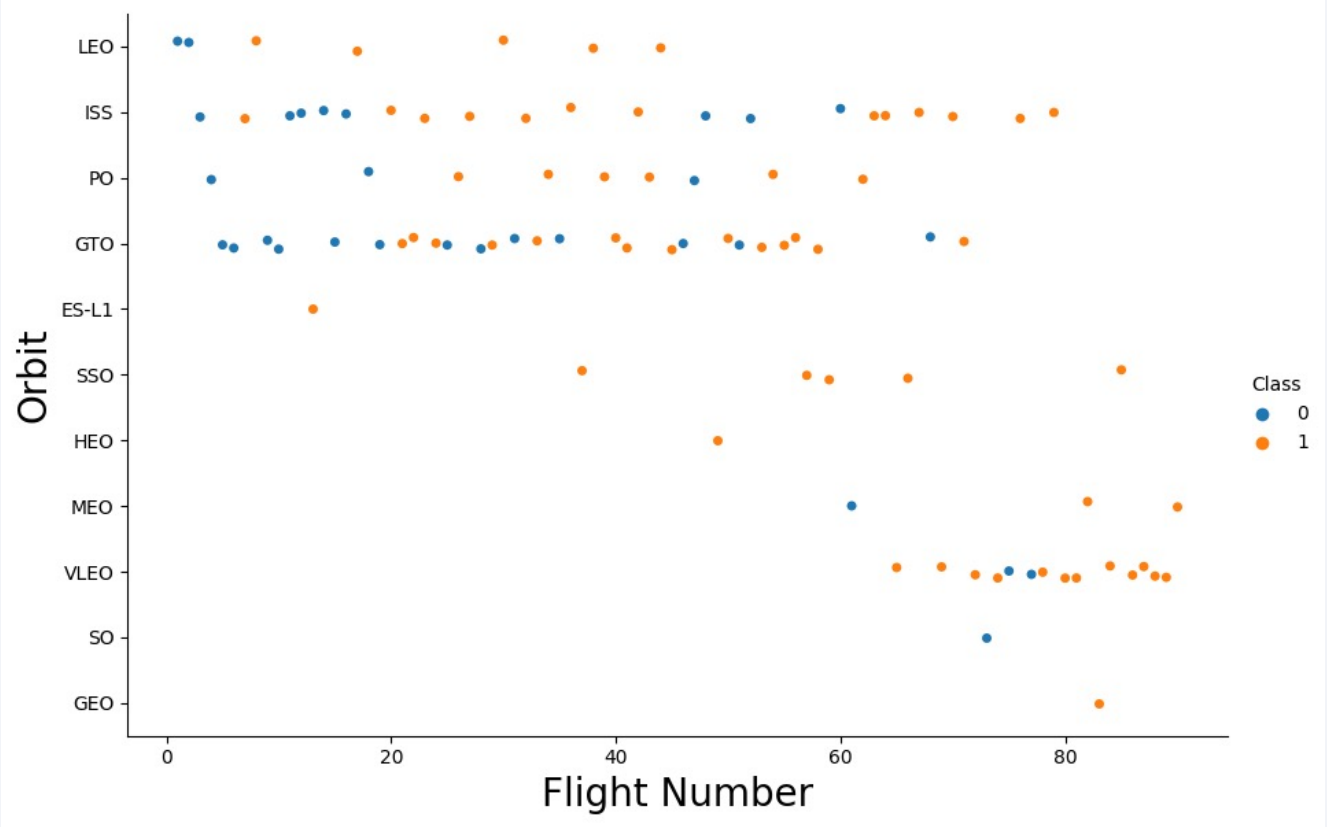
- Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



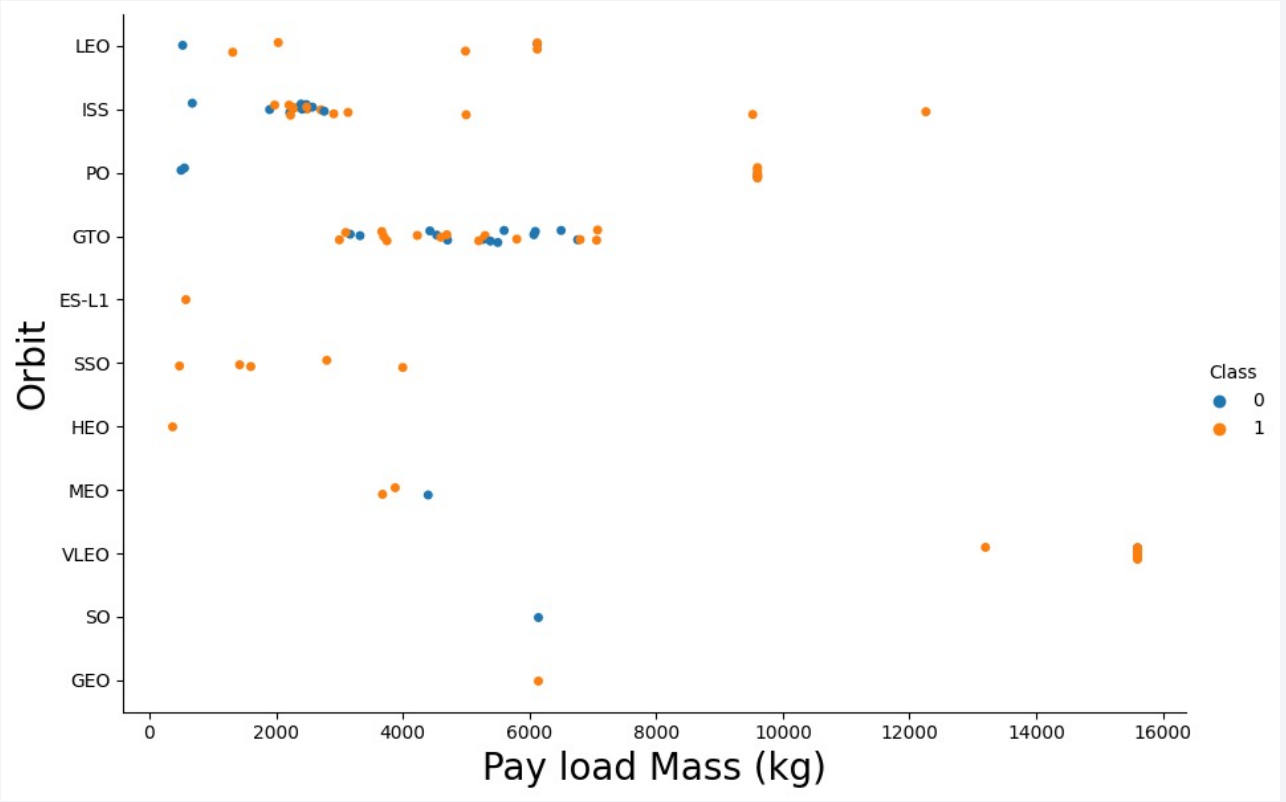
- Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate.

Flight Number vs. Orbit Type



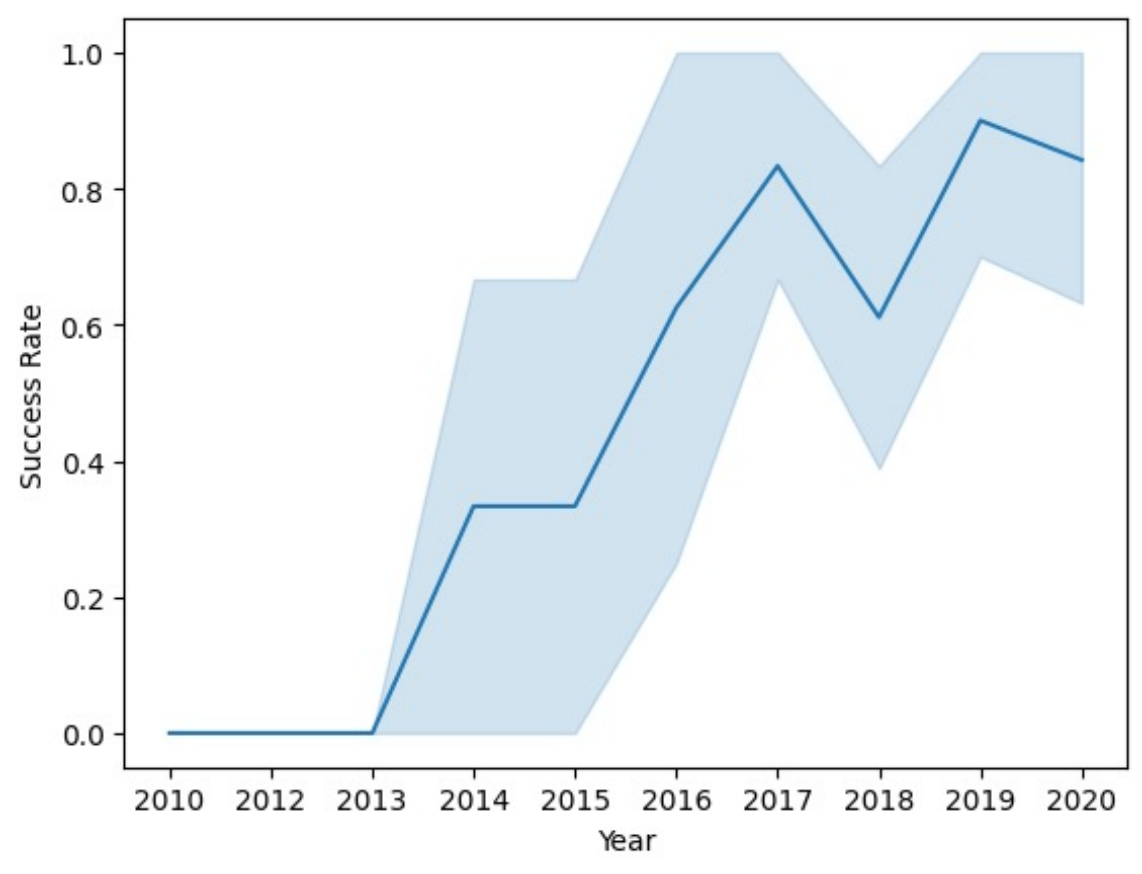
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

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

- Unique launch sites
- `%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;`

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- `%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

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 GROUP BY "Customer" HAVING "Customer" = 'NASA (CRS)';`

```
 * sqlite:///spaceX_data1.db
Done.
```

Customer	total_payload_mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1
- `%sql SELECT "Booster_Version", AVG("PAYLOAD_MASS__KG_") AS average_payload_mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';`

```
 * sqlite:///spaceX_data1.db
Done.
```

Booster_Version	average_payload_mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad
-
```
%sql SELECT MIN("Date") AS first_successful_landing_date \
FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

```
 * sqlite:///spaceX_data1.db
Done.
```

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT "Booster_Version","PAYLOAD_MASS__KG_" FROM \
  SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND
"PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

```
 * sqlite:///spaceX_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
- `%sql SELECT "Mission_Outcome", COUNT(*) AS total_count FROM SPACEXTABLE \`

```
GROUP BY "Mission_Outcome";
```

```
 * sqlite:///spaceX_data1.db
Done.
```

Mission_Outcome	total_count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
- `%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE \`

```
"PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_")FROM SPACEXTABLE);
```

```
 * sqlite:///spaceX_data1.db
Done.
```

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

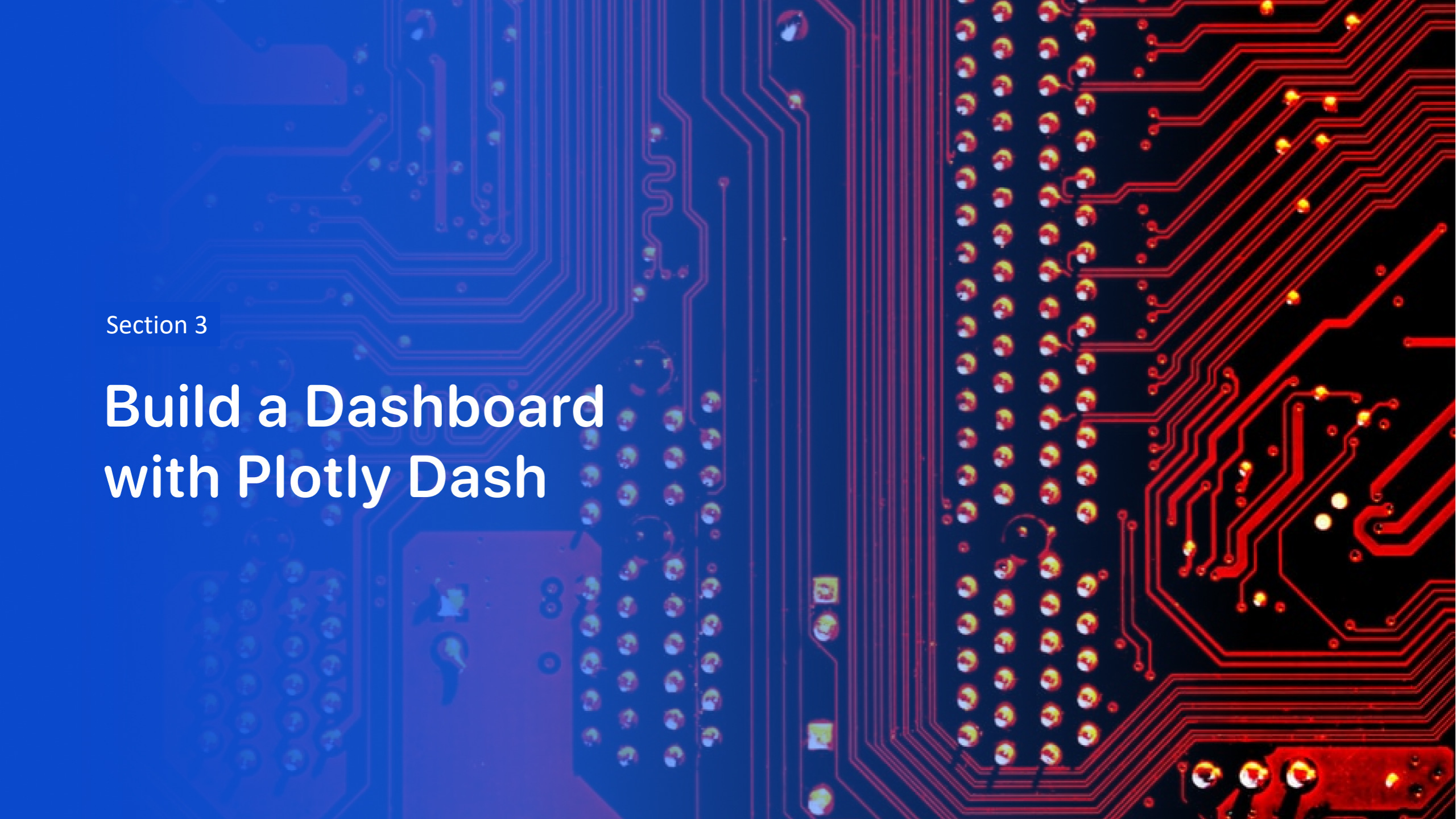
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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

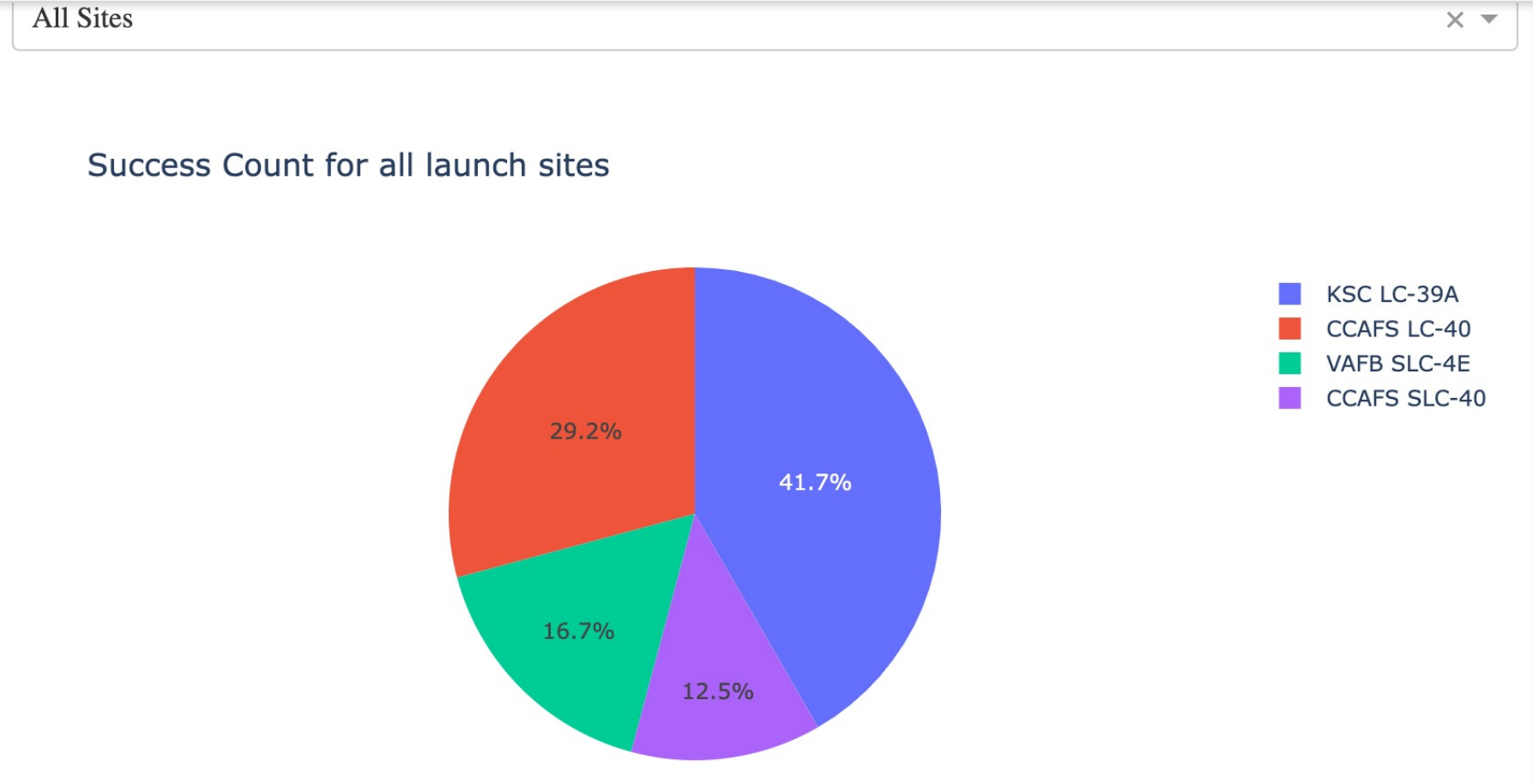
```
%sql SELECT "Landing_Outcome", COUNT(*) AS outcome_count FROM SPACEXTABLE \
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY "Landing_Outcome" \
ORDER BY outcome_count DESC;
```

```
 * sqlite:///spaceX_data1.db
Done.
```

Landing_Outcome	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

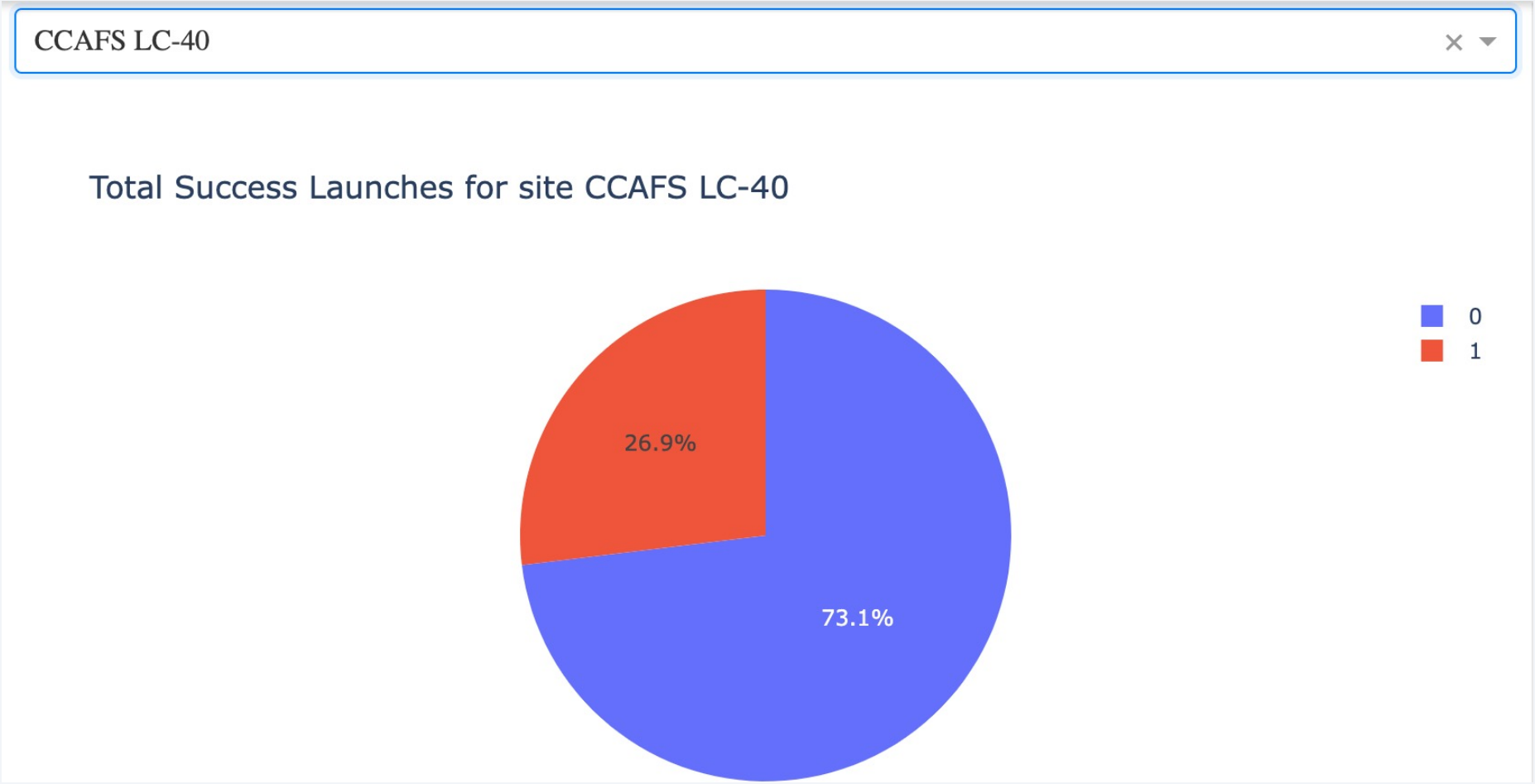

Section 3
Build a Dashboard with Plotly Dash

Pie-Chart for launch success count for all sites



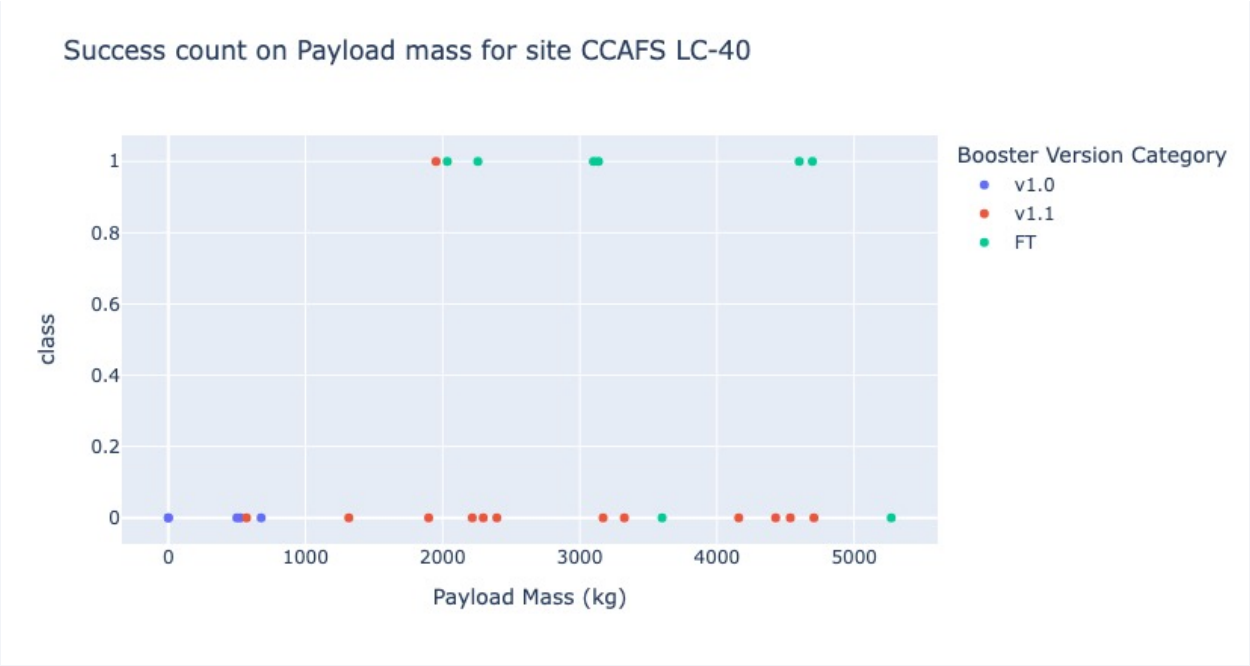
- Launch site KSC LC-39A has the highest launch success rate at 42% followed by CCAFS LC-40 at 29%, VAFB SLC-4E at 17% and lastly launch site CCAFS SLC-40 with a success rate of 13%

Pie chart for the launch site with 2ndhighest launch success ratio



- Launch site CCAFS LC-40 had the 2ndhighest success ratio of 73% success against 27% failed launches

Payload vs. Launch Outcome scatter plot for all sites



- For Launch site CCAFS LC-40 the booster version FT has the largest success rate from a payload mass of >2000kg

Section 4

Predictive Analysis (Classification)

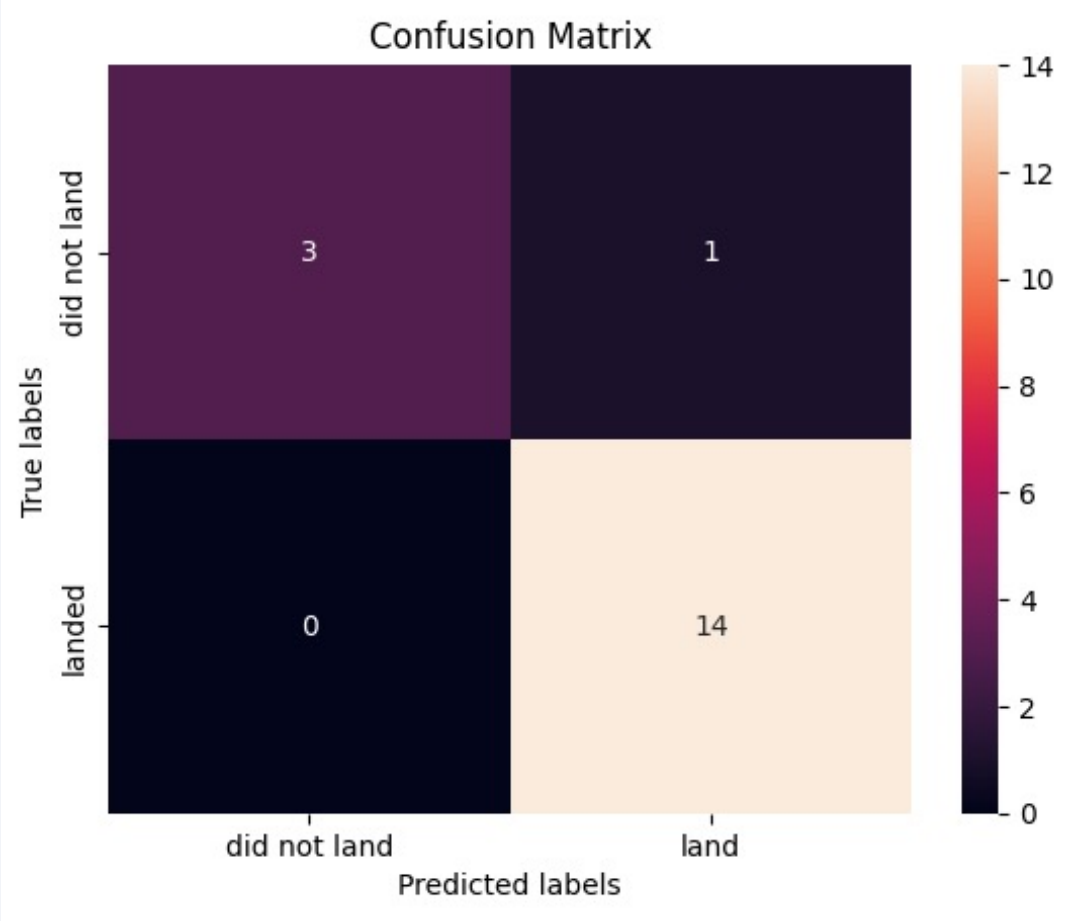
Classification Accuracy

- The built models accuracy on test data

index	
Method	Test Data Accuracy
Logistic_Reg	0.9444444444444444
SVM	0.8888888888888888
Decision Tree	0.6111111111111112
KNN	0.8888888888888888

Confusion Matrix

- The best performing model was **Logistics Regression** and this the confusion matrix related to that



Conclusions

- Different launch sites have different success rates. **CCAFS LC-40**, has a success rate of **60 %**, while **KSC LC-39A** and **VAFB SLC 4E** has a success rate of **77%**.
- We can deduce that, as the flight number increases in each of the 3 launch sites, so does the success rate. For instance, the success rate for the VAFB SLC 4E launch site is 100% after the Flight number 50. Both KSC LC 39A and CCAFS SLC 40 have a 100% success rates after 80th flight
- If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- Orbits **ES-L1, GEO, HEO & SSO have the highest success rates at 100%**, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate.
- 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
- 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
- Finally the success rate since 2013 kept increasing till 2020.

Thank you!