Winning Space Race with Data Science

Aleksandra Bal
6 November 2022

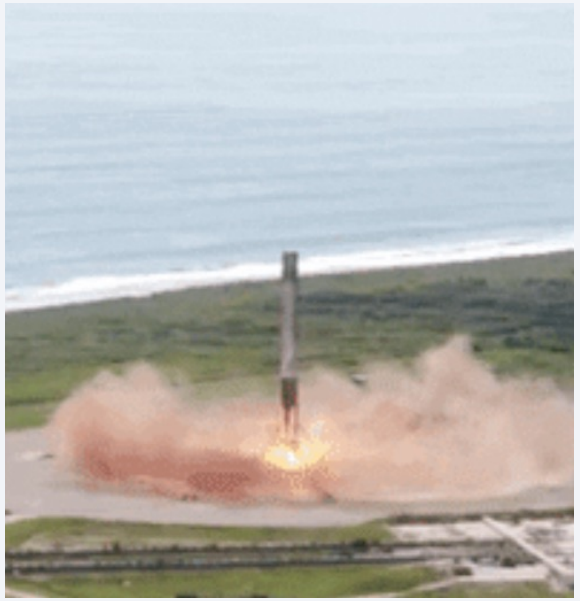
Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

A rocket launch is successful if the first stage of the rocket lands and can be reused. This can be predicted with an SVM model (accuracy: 88%, F1-score: 92%). Other factors that may affect the likelihood of success include:

- Launch site (76.9% of launches at KSC LC-39A were successful)
- Orbit type (ES-L1, GEO, HEO and SSO have 100% success rate)
- Payload mass (most launches with payload > 9000kg were successful but all launches at KSC LC-39A with mass < 6000kg succeeded)



Introduction

Problem: SpaceY wants to provide cost-efficient rocket launches to compete with SpaceX. To do so, it must be able to reuse the first stage of the rocket. The purpose of the project is to identify factors (for example, orbit type, payload mass, launch site) that are strongly correlated with a successful mission where the first stage of the rocket lands.

Question: Can we predict whether the first stage will land based on the available SpaceX data?

Section 1

Methodology

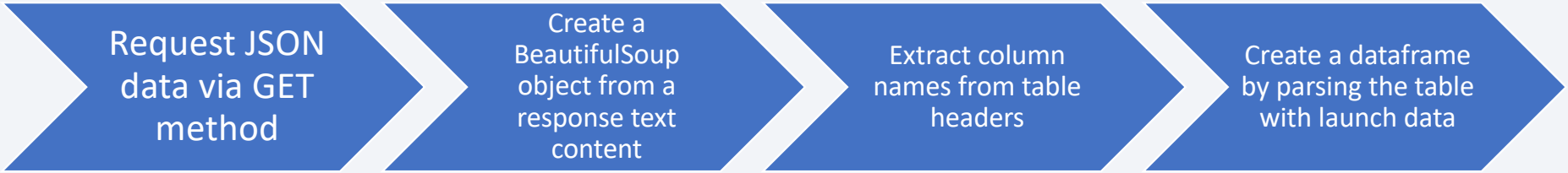
Methodology

- Data collection methodology
  - SpaceX API
  - Wikipedia (web scraping)
- Data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection – SpaceX API

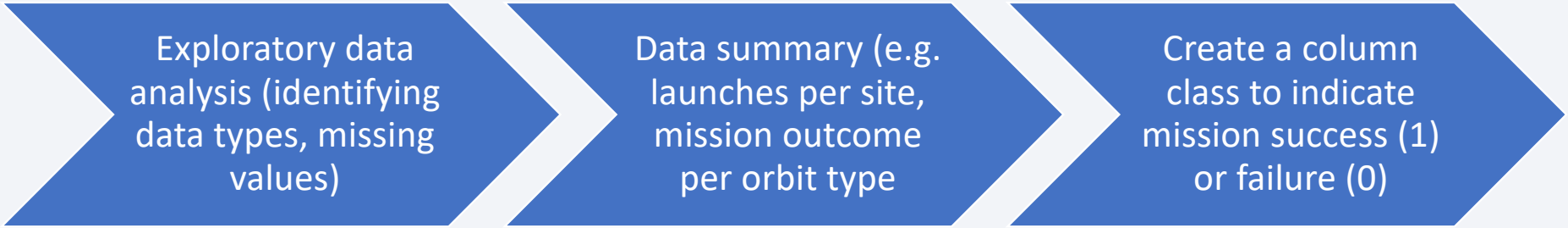
GitHub URL SpaceX API

Data Collection - Scraping



GitHub URL Webscraping

Data Wrangling



GitHub URL Data Wrangling

EDA with Data Visualization

The following charts were created:

- 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
- Success Rate Yearly Trend

GitHub URL Data visualization

EDA with SQL

The following queries were performed:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass.
9. List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
10. Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL SQL

Build an Interactive Map with Folium

All launch sites were marked with a circle and a text label displaying the site name.

For each site, successful and failed launches were displayed by means of clusters of colour-labelled markers.

For one launch site (KSC LC-39A), the distance to the nearest coastline, railway, highway and city was calculated and displayed by means of lines.

GitHub URL Folium

Build a Dashboard with Plotly Dash

The following plots were created:

- Pie chart showing the percentage of successful launches for all sites and per site (sites are selected via a dropdown list)
- Scatterplot showing the relationship between successful launches and payload mass (a slider allows to select a payload range)

The purpose of the plots is to identify sites with the highest percentage of successful launches and to determine whether there is a relationship between payload mass and launch success.

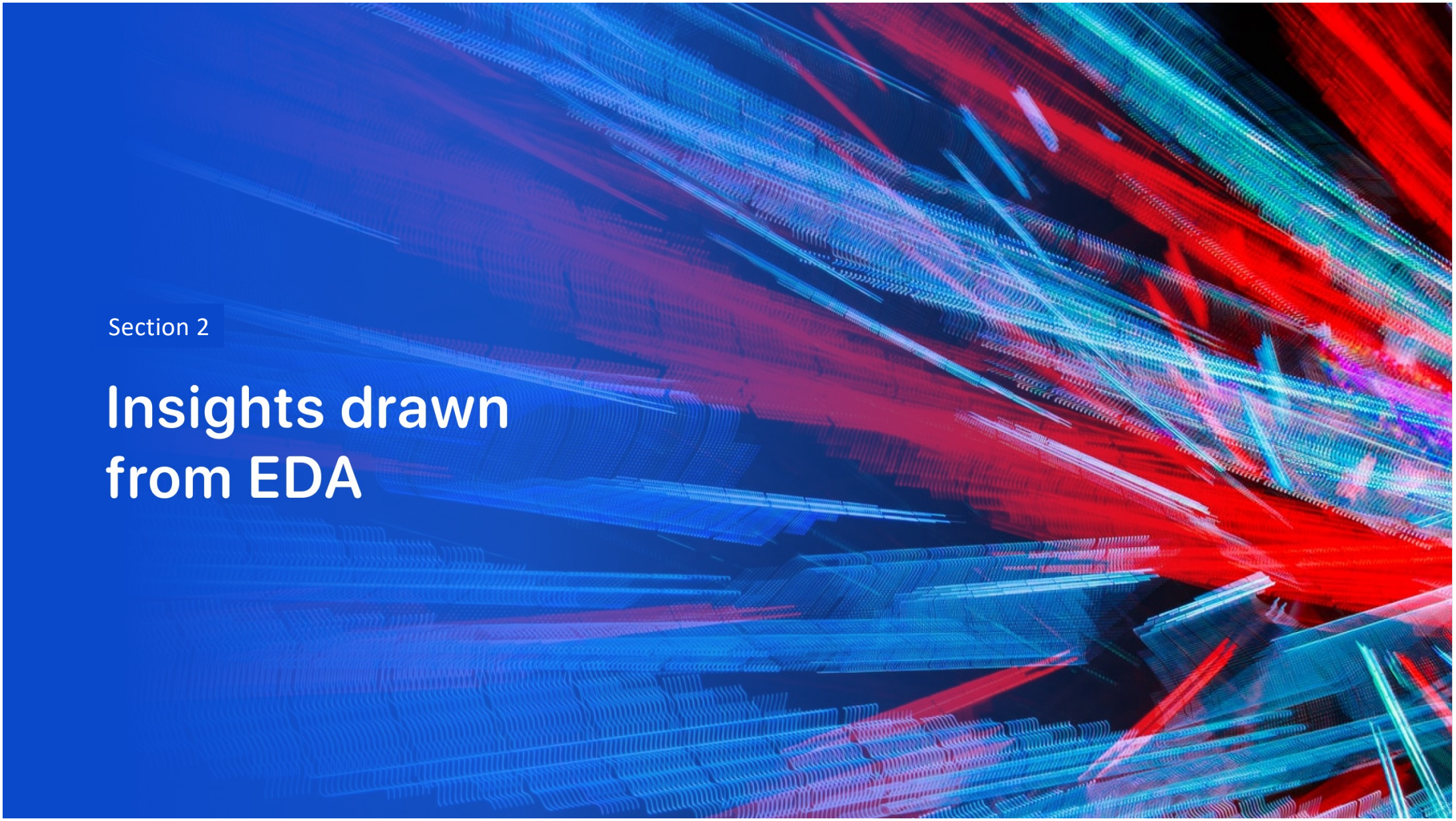
GitHub URL Dashboard

Predictive Analysis (Classification)

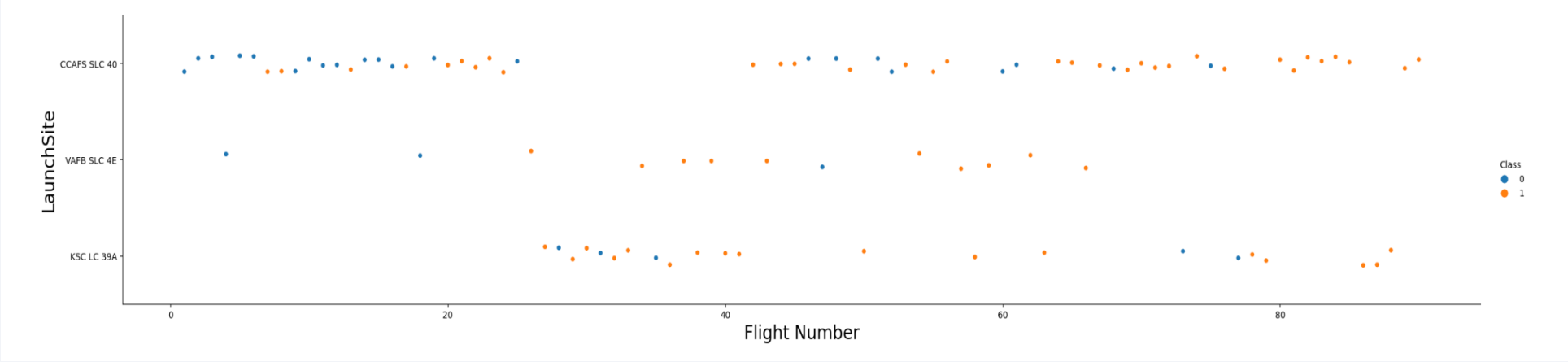
The following predictive models were used to find out which one performs best:

KNN, Decision Tree, SVM and Logistic Regression.

GitHub URL Predictive Analysis


Section 2
Insights drawn
from EDA

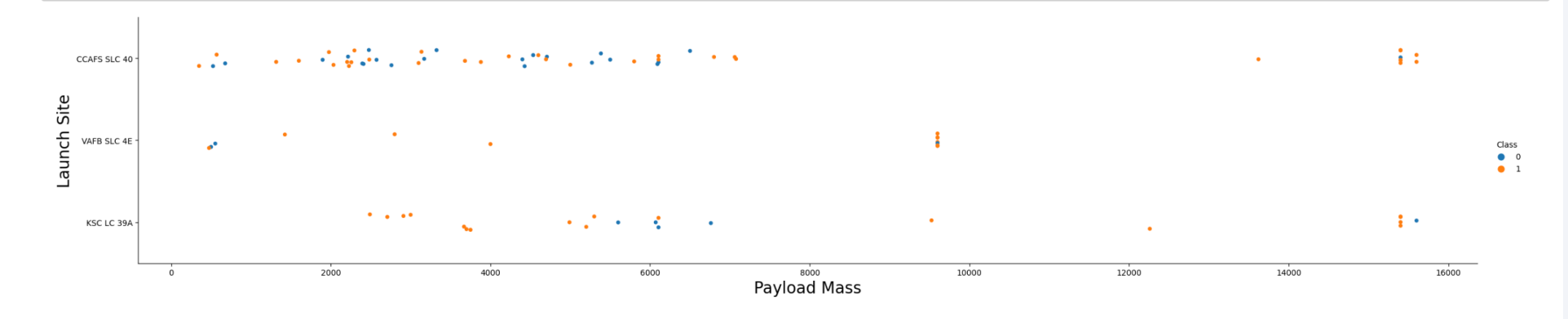
Flight Number vs. Launch Site



As the flight number increases, the first stage is more likely to land successfully. This means that the likelihood of success increases over time.

Most launches took place at CCAFS SLC 40.

Payload vs. Launch Site

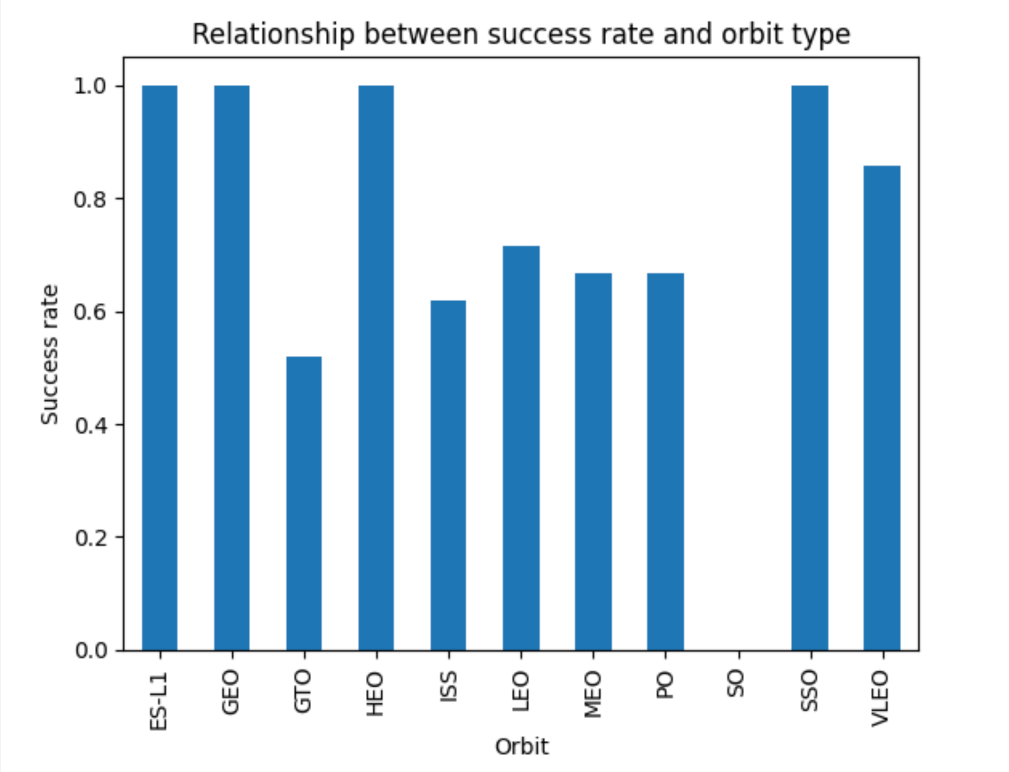


As the payload mass increases, launches tend to be more successful.

Most launches with payload > 9000kg were successful.

All launches at KSC LC 39A with payload mass < 6000kg were successful.

Success Rate vs. Orbit Type

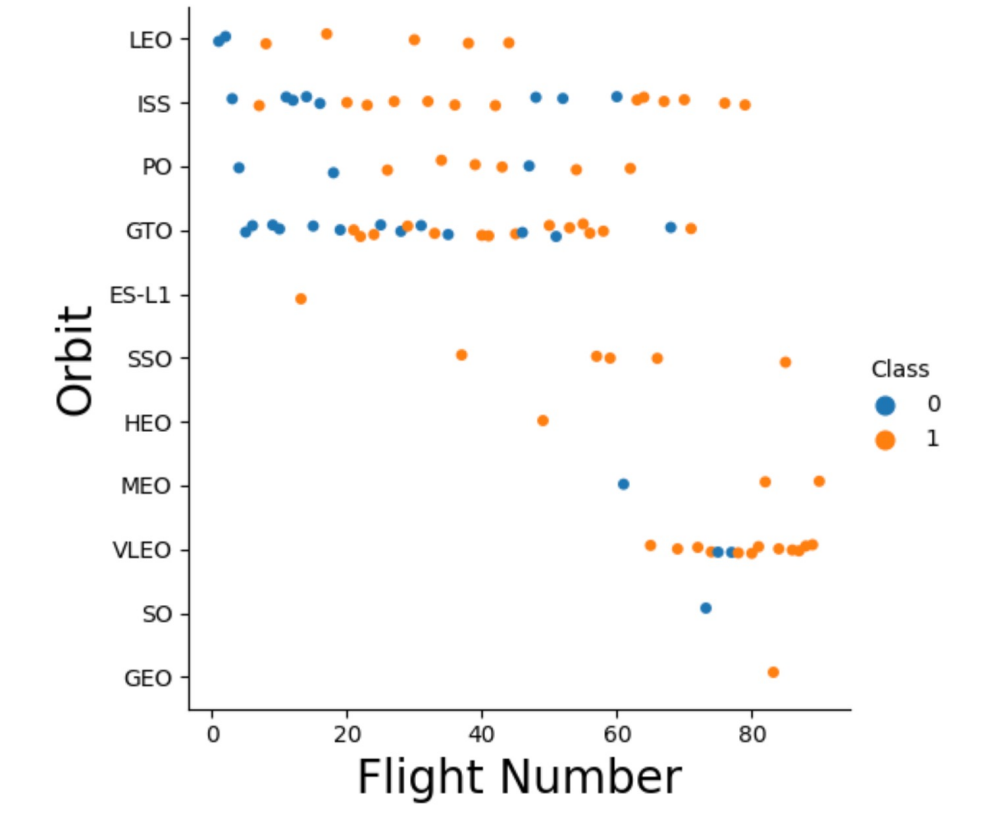


Orbits ES-L1, GEO, HEO and SSO have the highest success rate (100%).

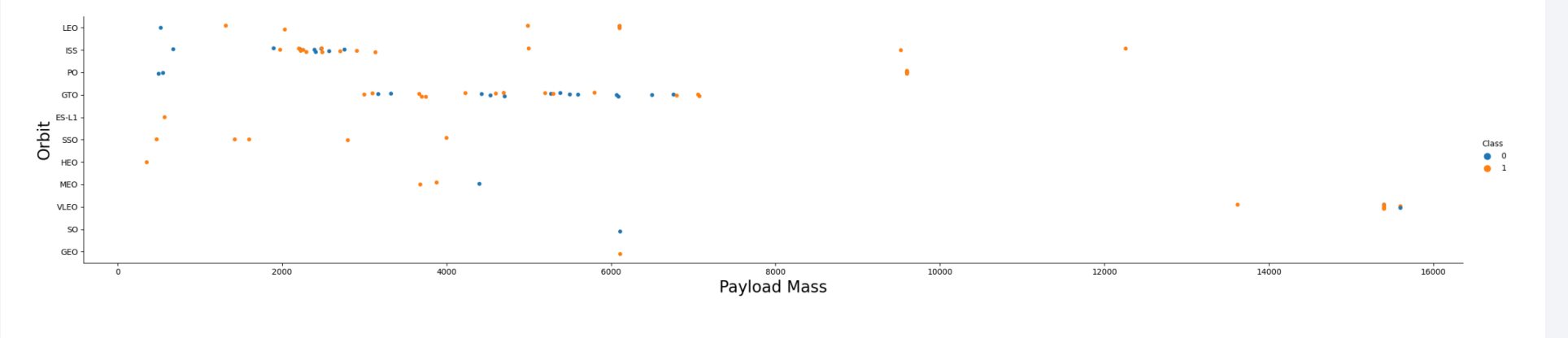
Orbit SO has the lowest success rate (0%).

Flight Number vs. Orbit Type

There is no clear relationship between flight number and orbit type.



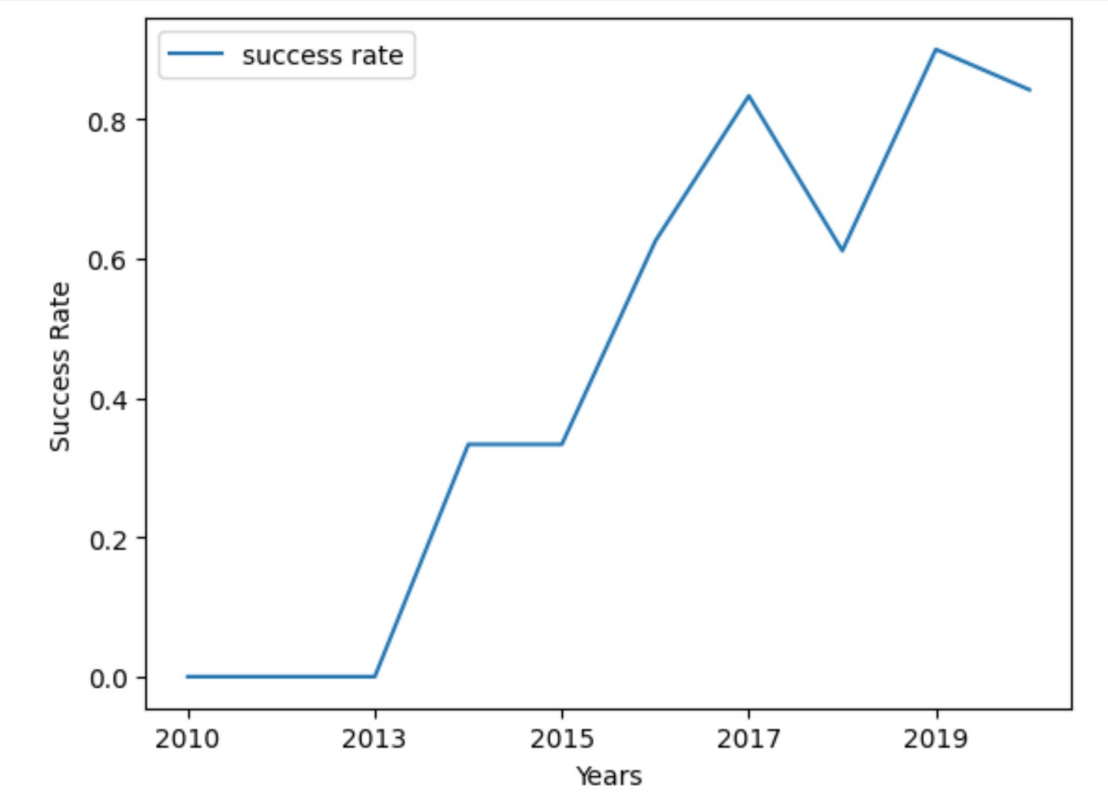
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

Launch Success Yearly Trend

The success rate shows an upward trend. It kept increasing since 2013 with a temporary dip in 2018.



All Launch Site Names

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

```
In [28]: %sql Select distinct "Launch_Site" from SPACEXTBL
```

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

Out[28]:

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

There are four launch sites. However, based on Wikipedia, CCAFS LC-40 is a former name of CCAFS SLC-40. Therefore, there are actually three unique launch sites.

Launch Site Names Begin with 'CCA'

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

In [32]:
```
%sql Select * from SPACEXTBL where "Launch_Site" like "CCA%" limit 5
```

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

Out[32]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

In [34]:
```
%sql Select sum("PAYLOAD_MASS__KG_") as "Total payload mass" from SPACEXTBL where "Customer" = "NASA (CRS)"
```

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

Out[34]:

Total payload mass
45596

Average Payload Mass by F9 v1.1

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

In [35]:
```
%sql Select avg("PAYLOAD_MASS__KG_") as "avg payload F9v1.1" from SPACEXTBL where "Booster_Version" like "F9 v1.1%"
```

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

Out[35]:

avg payload F9v1.1
2534.6666666666665

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [37]:
```
%sql Select min("Date") from SPACEXTBL where "Landing _Outcome" = "Success (ground pad)"
```

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

Out[37]:

min("Date")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [41]:
```
%sql Select "Booster_Version" from SPACEXTBL where "Landing _Outcome"= "Success (drone ship)"
and "PAYLOAD_MASS__KG_" >400 and "PAYLOAD_MASS__KG_" <6000
```

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

Out[41]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1038.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [44]:
```
%sql Select count(*) as "Successful mission" from SPACEXTBL where "Mission_Outcome" Like "%Success%"
```

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

Out[44]:

Successful mission
100

In [45]:
```
%sql Select count(*) as "Unuccessful mission" from SPACEXTBL where "Mission_Outcome" Not like "%Success%"
```

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

Out[45]:

Unuccessful mission
1

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [47]:
```
%sql Select "Booster_Version" from SPACEXTBL where "PAYLOAD_MASS__KG_" =
(Select max("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

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

Out[47]:

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [50]:
```
%sql Select substr("Date", 4, 2), "Landing _Outcome", "Booster_Version", "Launch_Site"
from SPACEXTBL where substr("Date", 7, 4) = "2015" and "Landing _Outcome" = "Failure (drone ship)"
```

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

Out[50]:

substr("Date", 4, 2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

In [52]:
```
%sql Select "Landing _Outcome", count("Landing _Outcome") from SPACEXTBL where "Date" between "04-06-2010" and "20-03-2
group by "Landing _Outcome" having "Landing _Outcome" like "s%" order by count("Landing _Outcome") desc
```

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

Out[52]:

Landing _Outcome	count("Landing _Outcome")
Success	20
Success (drone ship)	8
Success (ground pad)	6


Section 3
Launch Sites
Proximities Analysis

Launch site locations

All launch sites are located near the coastline and close the the equator line.

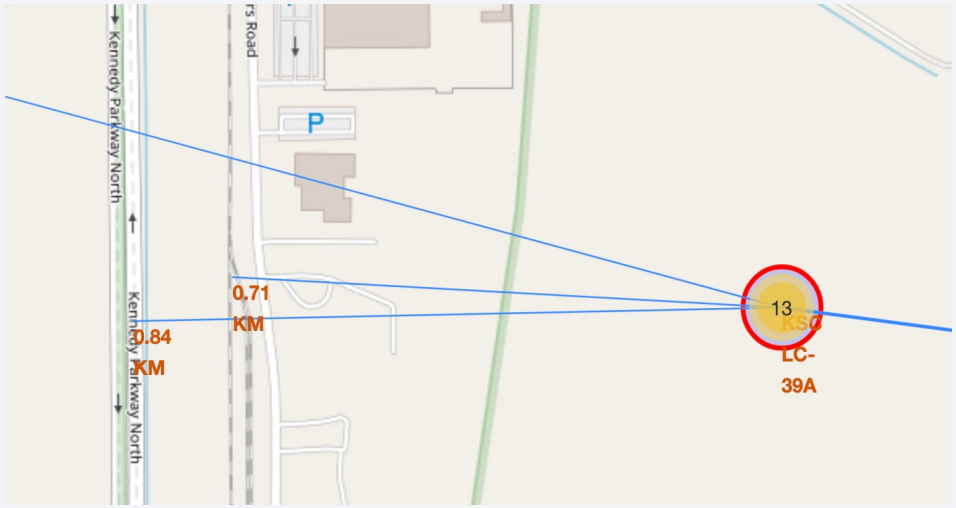
This minimizes the risk that a rocket explosion would affect areas where people live.

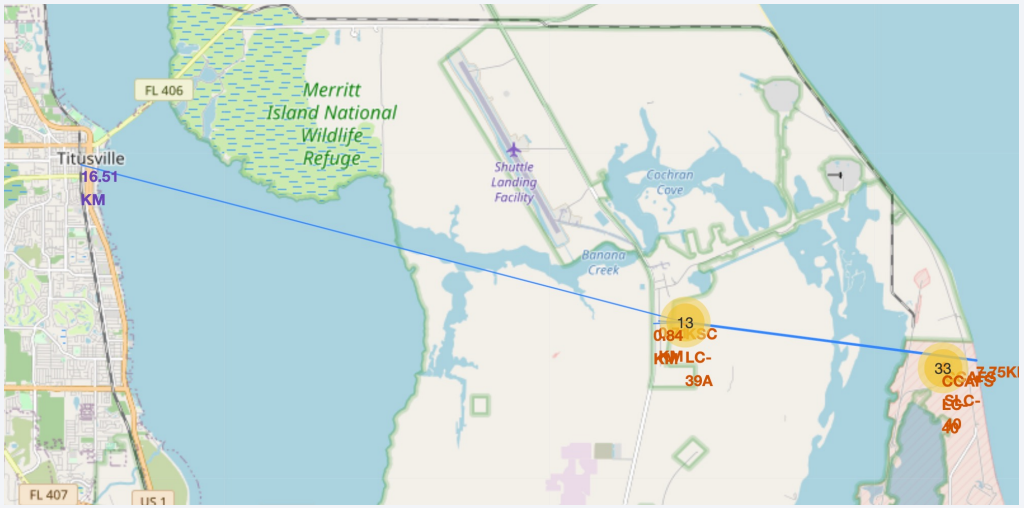
Earth rotates faster at the Equator than it does at any other place. If a rocket is launched from a site near Earth's equator, it can take optimum advantage of the Earth's substantial rotational speed.

Color-labelled launch outcomes

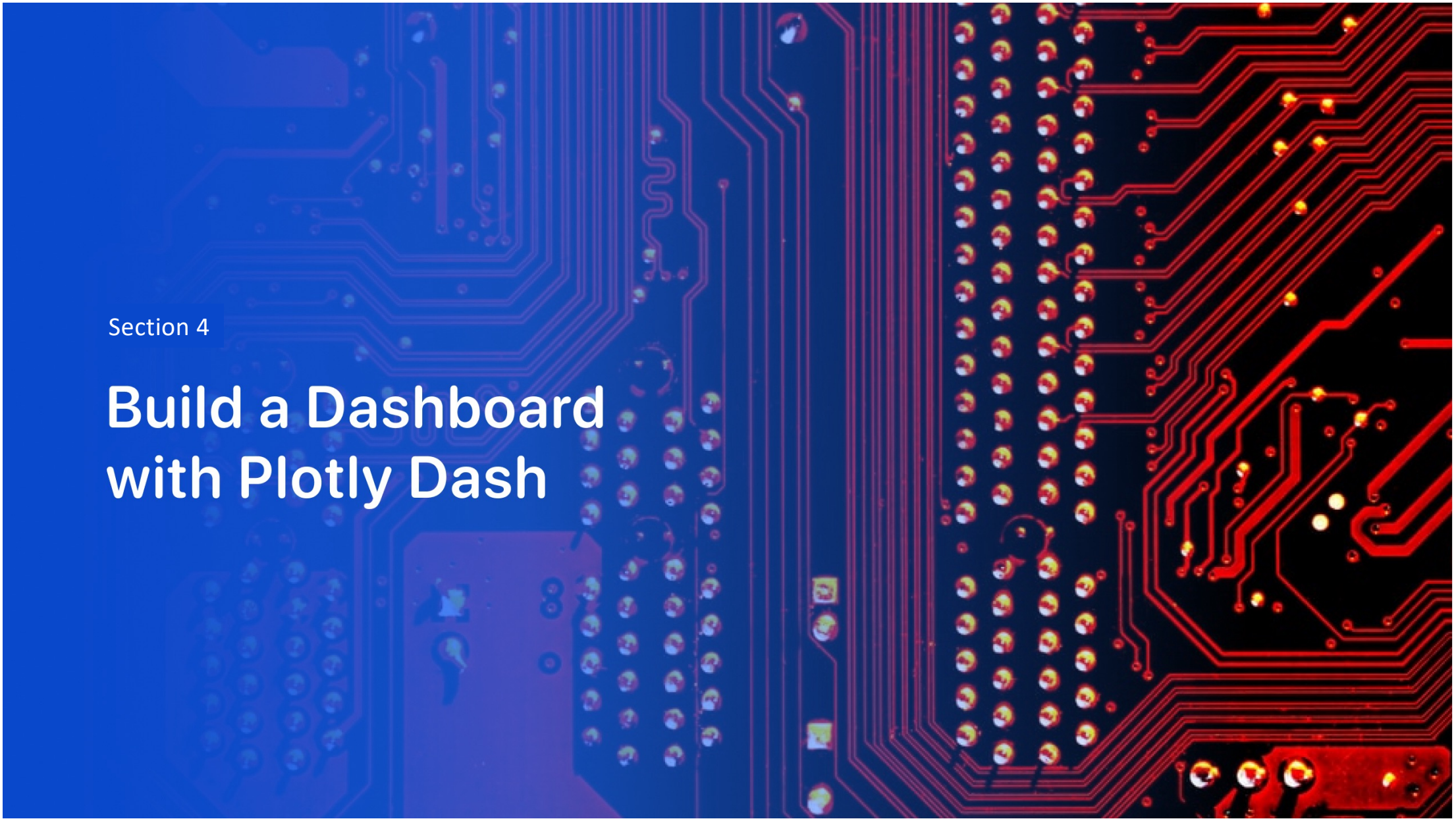
Launch site KSC LC-39A has a high percentage of successful outcomes (indicated by green markers).

Launch site KSC LC-39A

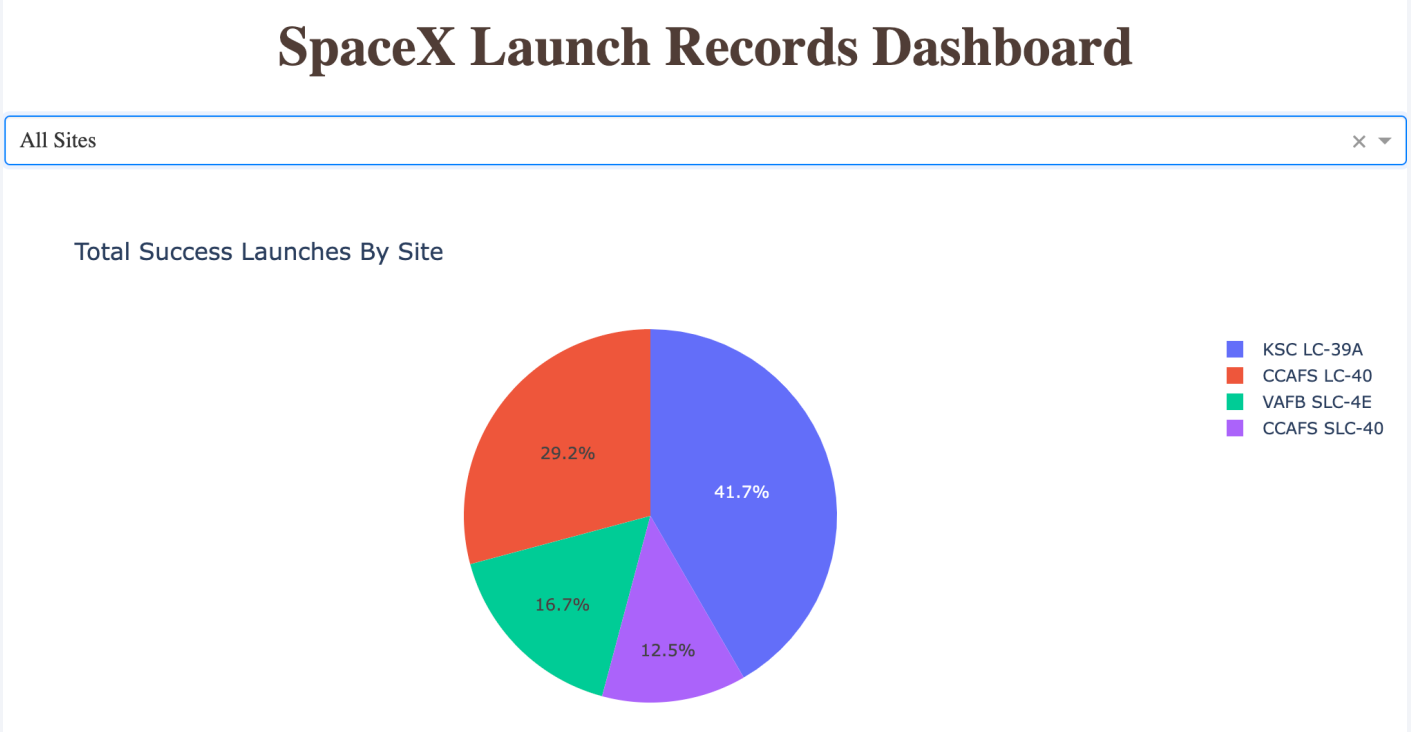





```
Distance to the nearest city:  16.51 km
Distance to the nearest railway:  0.71 km
Distance to the nearest highway:  0.84 km
Distance to the nearest coastline:  7.75 km
```

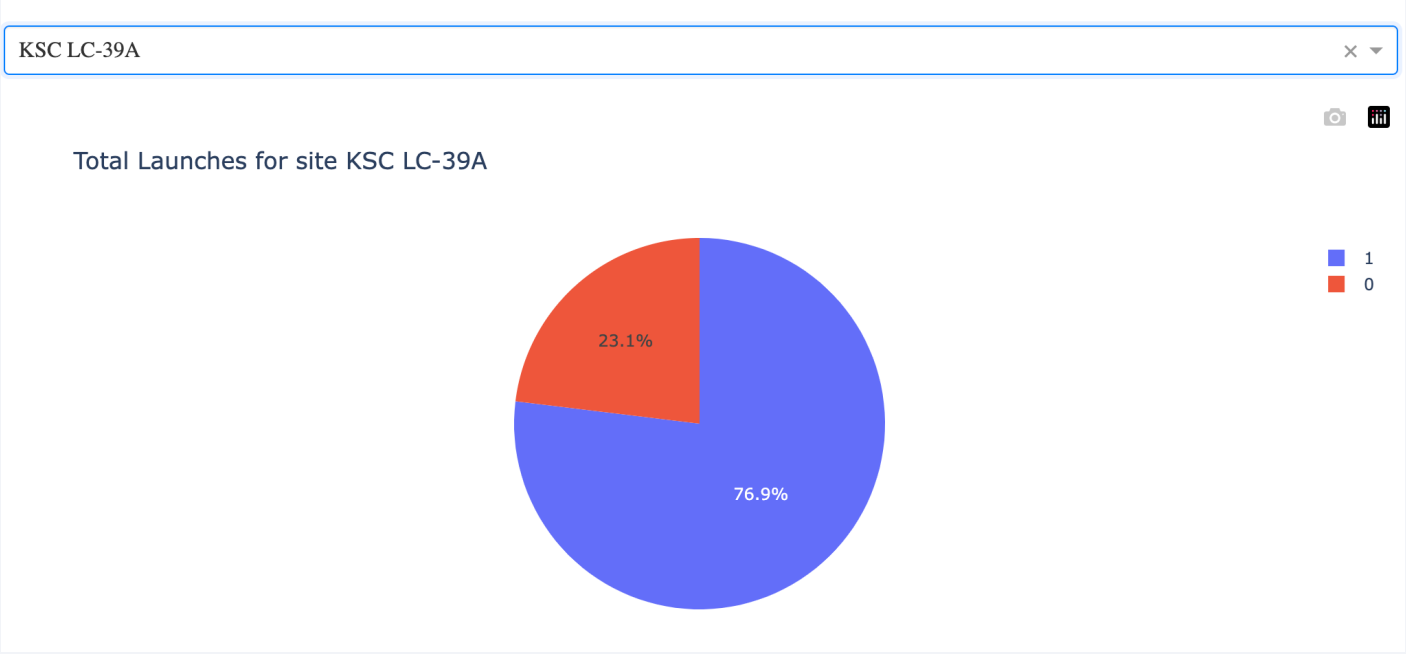

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



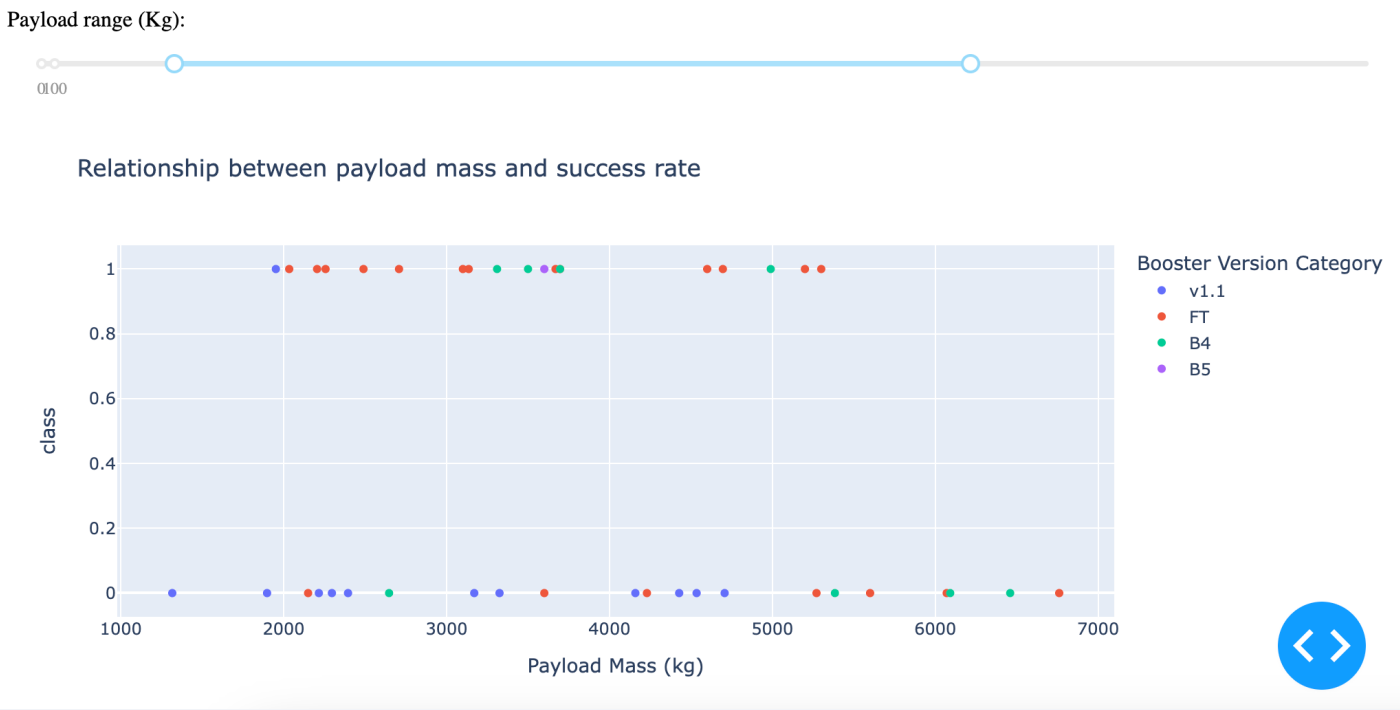
Most successful launches occurred at launch site KSC LC-39A.

<Dashboard Screenshot 2>



Launch site KSC LC-39A has the highest percentage of successful launches (76.9%).

Payload vs. Launch Outcome



Rockets with payload mass between 2000 kg and 6000 kg in combination with booster category FT have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Test data

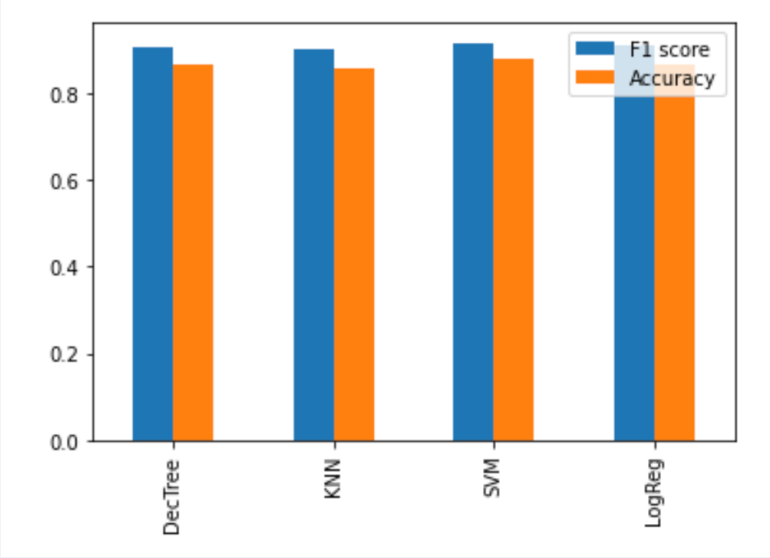
Based on the test data, it is not possible to select the best model as all have the same F1 and accuracy scores. This could be due to the small test dataset (18 observations).

	DecTree	KNN	SVM	LogReg
F1 score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Entire dataset

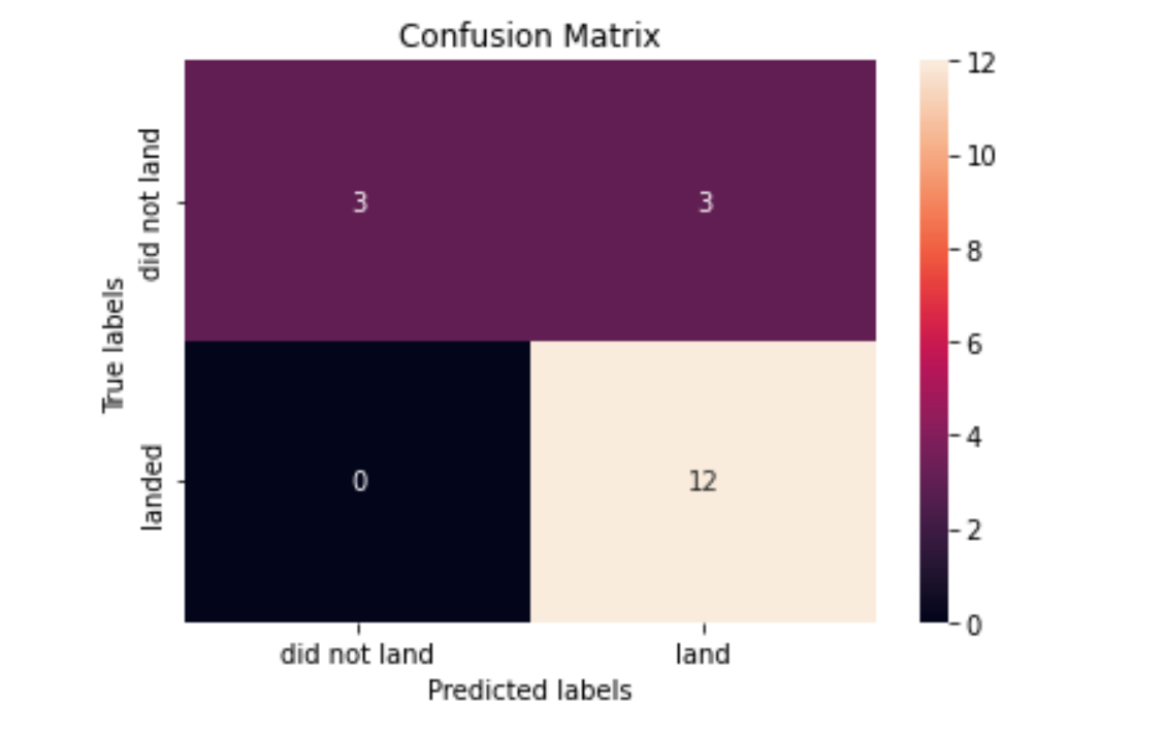
If we calculate F1 and accuracy scores on the entire dataset, the SVM model has the highest scores, but they do not differ significantly from those of other models.

	DecTree	KNN	SVM	LogReg
F1 score	0.906250	0.900763	0.916031	0.909091
Accuracy	0.866667	0.855556	0.877778	0.866667



Confusion Matrix

SVM does not produce any false negatives, but its major problem are false positives (i.e. it is predicted that the first stage will land but it does not).



Conclusions

- Launch site KSC LC-39A has the highest percentage of successful launches.
- All launch sites are close to the equator line and coastline. They are relatively close to highways and railways.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- As the payload mass increases, launches tend to be more successful. Most launches with payload > 9000kg were successful; however, all launches at KSC LC-39A with mass < 6000kg succeeded.
- SVM seems to be the most accurate model to predict the likelihood of successful launches (accuracy: 88,%, F1-score: 92%) but it does not perform significantly better than other models (SVM, KNN, Decision Tree).

Thank you!