IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vijay Kumar
July 16, 2024

Outline

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

Executive Summary

- Data related to SpaceX rocket launches was collected using the SpaceX API and by scraping table content from Wikipedia pages.
- Multiple Learning algorithms such as Logistic Regression, SVM, Decision Trees, & KNN are evaluated to predict the whethere launch will be successful or not.
- Overall results show that Decision Tree models performed the best in being able to classify whether a Falcon 9 launch would be successful or not.

Introduction

- The Projects aim is to predict Future Space X Falcon 9 first stage landing. Using data of past launches which contains features related to specification of each rocket launched, a classifier is built.
- Knowing that there are multiple versions and design of the Falcon 9 Booster, we would like to understand which version has the most success rate, which version can carry the most payload, which type is best suitable for which application (Low loads, Mid-loads, High-Loads), and which features contribute towards building a successful classifier.

Section 1

Methodology

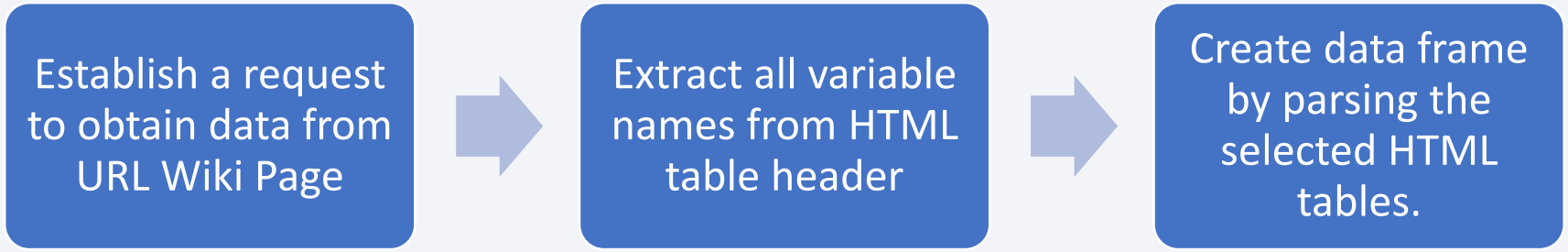
Methodology

Executive Summary

- Data collection methodology:
  - Data is scraped from Wikipedia pages by extracting content from tables
- Perform data wrangling
  - Based on the outcome of the launch, binary categorical labels were created.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearchCV and multiple learning algorithms, each model was tested with varying parameters, and the best performing model was selected.

Data Collection

- Using BeautifulSoup, tables from Wikipedia pages were extracted
- Data Collection process can be summarized as such:



Data Collection – SpaceX API

1. Send Request to connect to the SpaceX API and obtain launch data.
2. Decode Response as Json format and convert to pandas dataframe for easy manipulation
3. Filter data based on information required and convert to appropriate format.
4. Define functions to extract and filter required information
5. Iterate and organize data into pandas dataframe with labeled column names.
6. Filter dataset to contain Falcon 9 Booster Version
7. Check & Handle missing values by replacing with the mean value of that column.

- https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/1-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

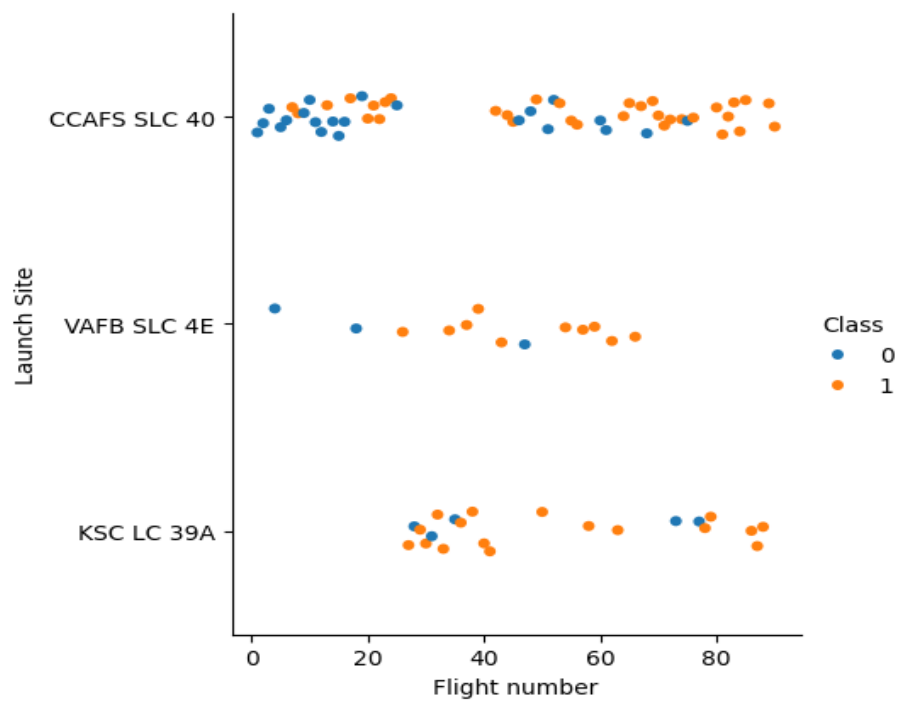
1. Request Falcon9 Launch Data from Wiki page
2. Extract all tables in HTML document.
3. Extract all column names from selected HTML table header
4. Parse HTML tables and create a data frame

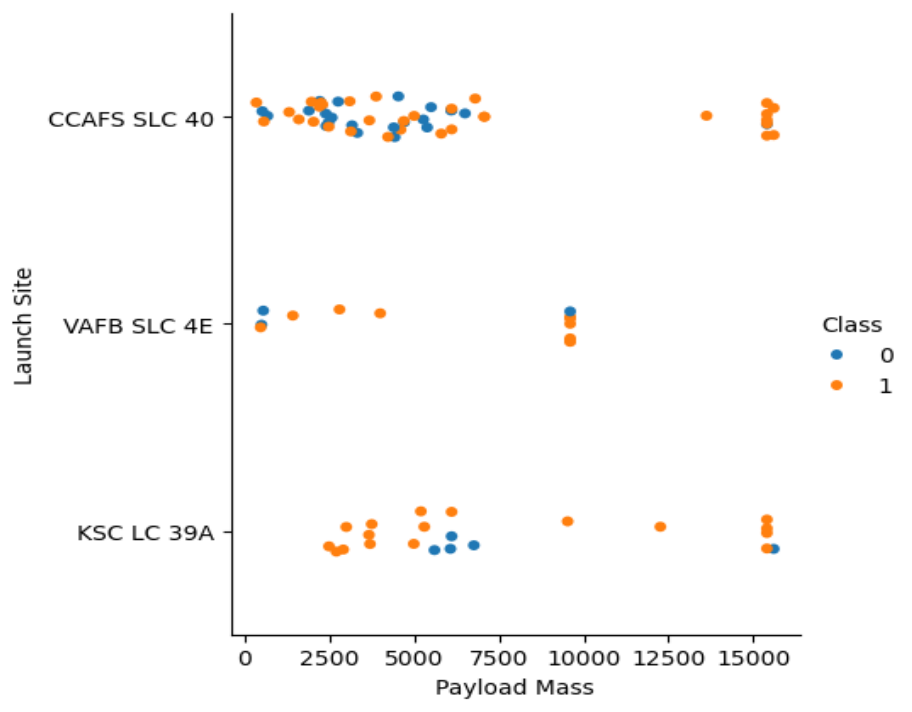
- https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/1-jupyter-labs-webscraping.ipynb

Data Wrangling

1. Identifying categorical and continuous variables.
2. Verifying each feature has correct datatype assigned.
3. Checking for missing values
4. Compute number of launches per site
5. Compute number of each orbit
6. Compute number of successful missions of the orbits.
7. Organize missions into 2 categories by assigning labels as such: success (1) or fail (0).
8. Save data to CSV file.

EDA with Data Visualization

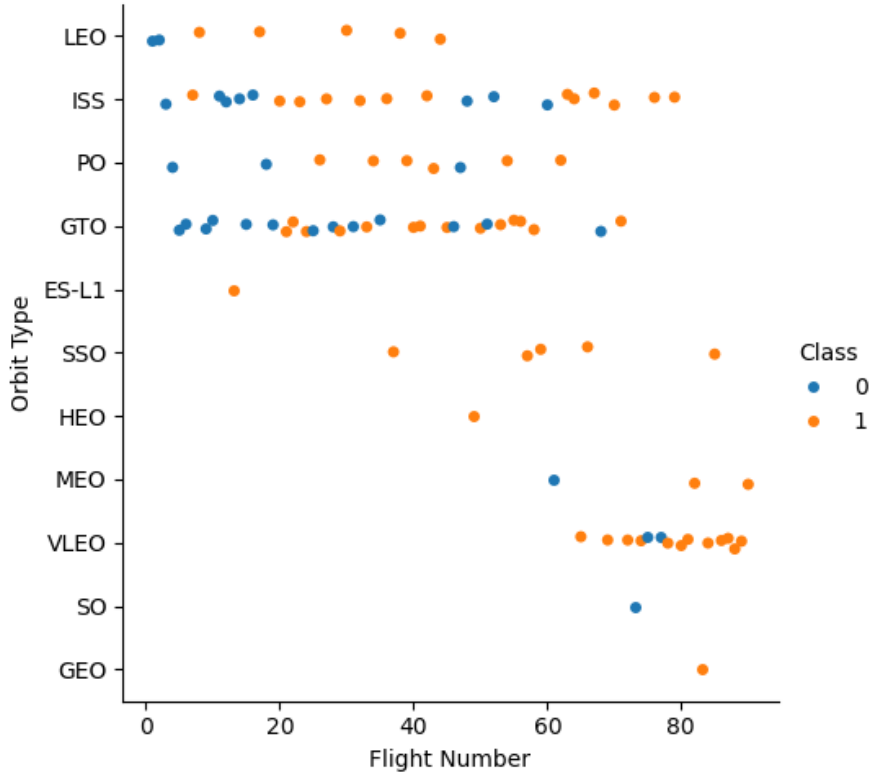


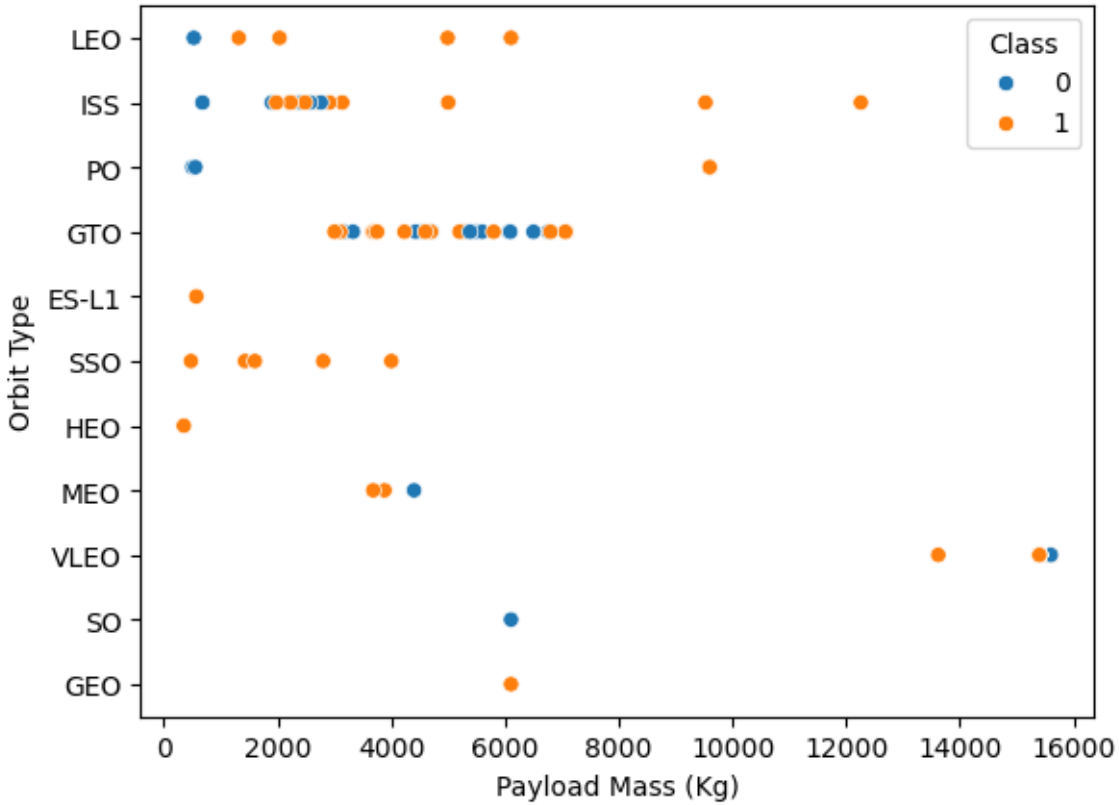


- Flight Number VS Launch Site: As number of flights increase there is an increase in successful launches and decrease in failed launches.
- Payload Mass (kg) VS Launch Site: Site CCAFS SLC 40 is optimal for Heavy Payloads which is shown by the high success rate, whereas VAFB SLC 4E & KSC LC 39A both are optimal for low to mid payloads.

- https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/4-edadataviz.ipynb

EDA with Data Visualization

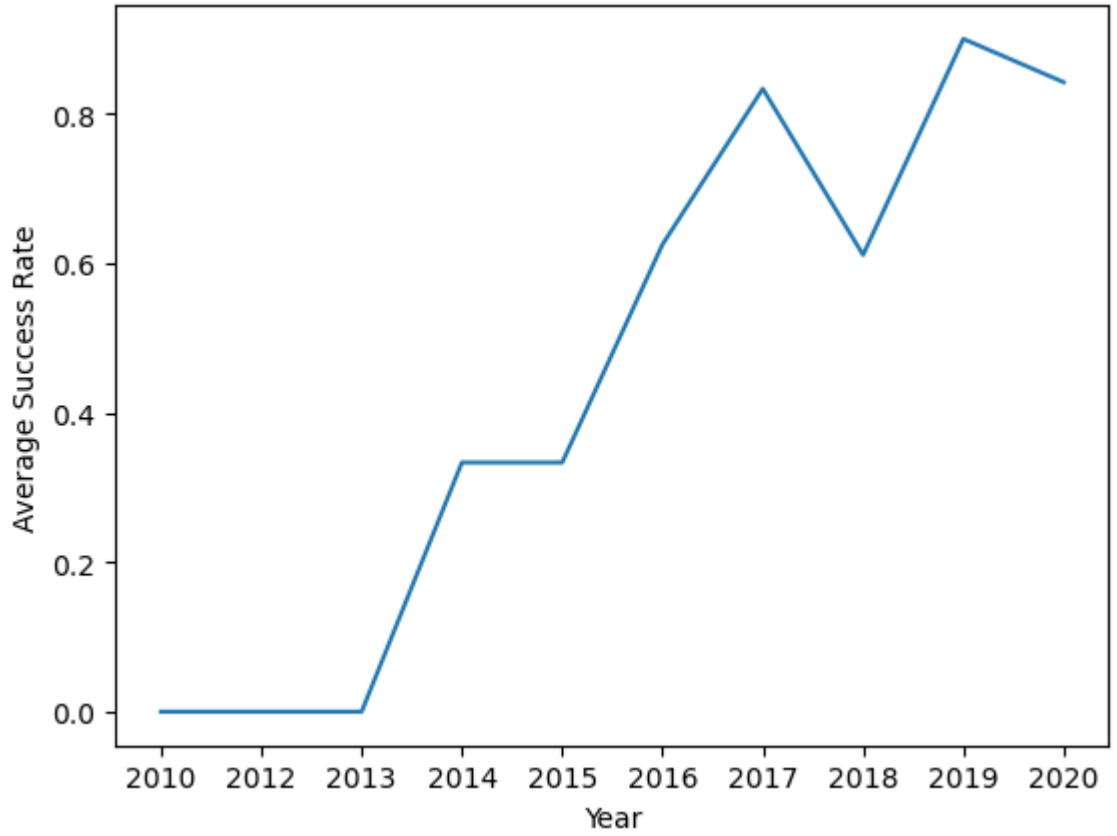




- Flight Number VS Orbit Type: ES-L1, SSO, HEO, MEO, VLEO, & GEO all have a high success rate whereas other orbit types have an average success rate.
- Payload Mass (Kg) VS Orbit Type: observation can be made that orbit types have success rate based on payload range. LEO, ISS, PO & VLEO have a high success rate with heavy payloads, GTO is unpredictable, whereas the others have a high success rate with low payloads.

- https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/4-edadataviz.ipynb

EDA with Data Visualization



- Over time the success rate of each launch has improved over the years indicating more successful landings of the first stage.

- https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/4-edadataviz.ipynb

EDA with SQL

- To extract unique names from launch site with no repetition, the DISTINCT keyword is used in queries.
- To extract data with a given a conditional requirement, the WHERE keyword is used at the end of the query.
- To apply a function on a range of data, the function is applied to the column and value is computed accordingly. AVG, MIN, SUM, COUNT, and SUBSTR are functions used in the queries.
- To extract data of unknown value given the condition, subqueries are used in combination with conditional statements that use the keyword WHERE in queries.

```
%sql select DISTINCT(Launch_Site) from SPACEXTABLE
```

```
%sql select * FROM SPACEXTABLE WHERE Launch_Site Like "CCA%" LIMIT 5
```

```
%sql SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) AS "Average payload mass" FROM SPACEXTABLE
WHERE Booster_Version = "F9 v1.1"
```

```
%sql select DISTINCT(Booster_Version), PAYLOAD_MASS__KG_ as "Max payload" from SPACEXTABLE WHERE
PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```


https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/3-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To highlight key points and areas of interest on the map, certain objects were used such as:

1. Folium Circle: used to highlight a zone of radius “r” with a circular pattern.
2. Folium Marker: used to create a marker object to be used as a visual indicator to highlight a certain point with text.
3. Folium MarkerCluster: creates an object that shares the same coordinate of multiple markers.
4. Folium PolyLine: Creates an object that draws a line between a starting and ending point.

https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/5-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

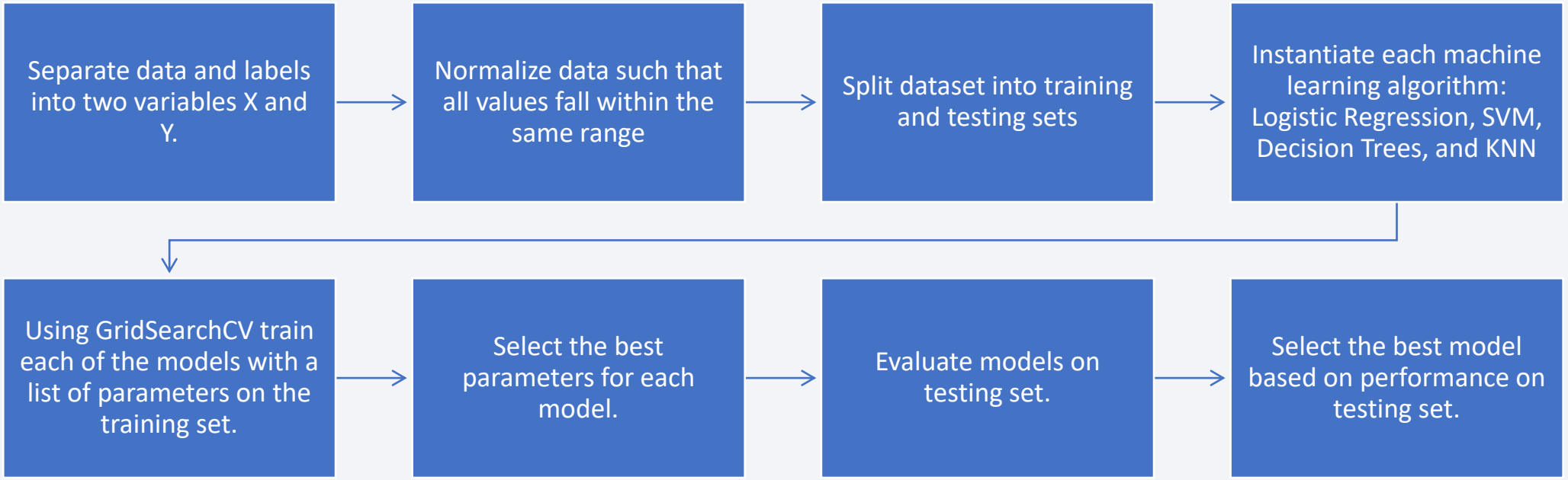
- To derive useful insights from data, it should be presented in such a way that it communicates directly to the end user and delivers the information with ease. The following visualizations are used:

1. Pie Chart: A pie chart displays categorical data, highlighting the proportions of each class in the given set of data. Pie charts give an overview of the data, displaying the dominant and least dominant classes based on area size.

2. Scatter Chart: is a visualization tool used to compare two variables with respect to each other and understand their distribution, each data point can be visually represented by the class to which they belong to highlighting the outcome of each point.

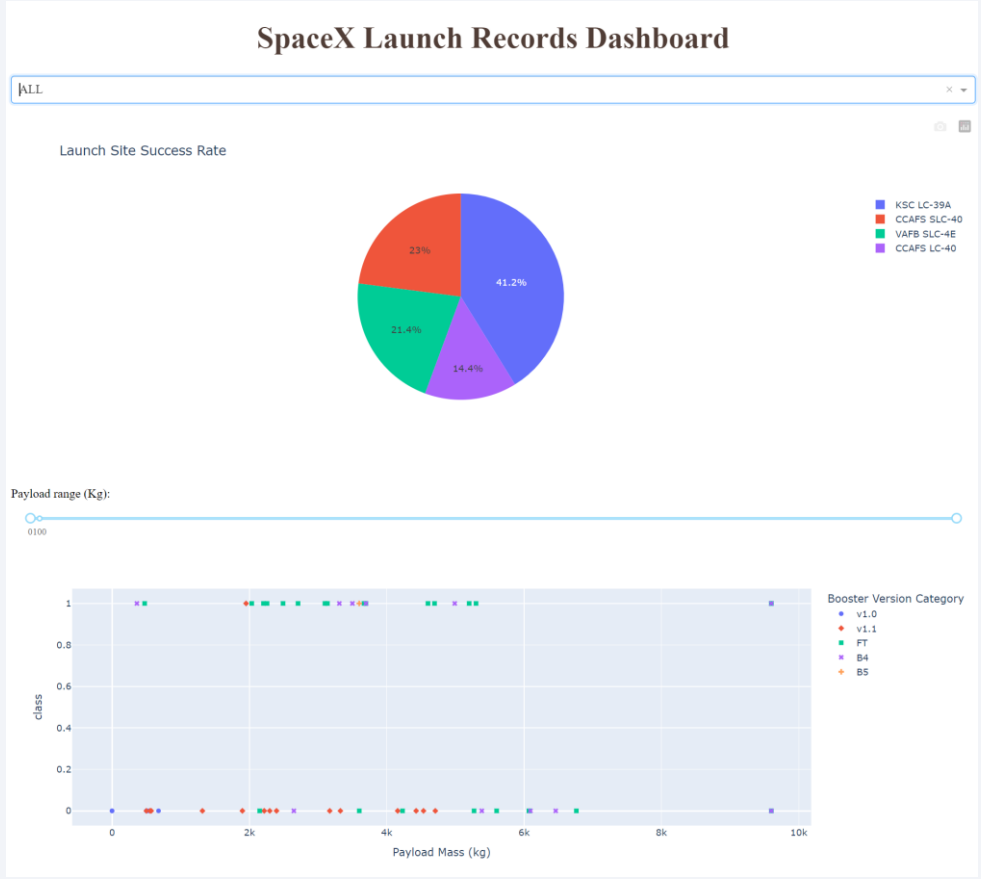
https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/6-spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/vijaykumar1799/SpaceX-First-Stage-Launch-Classifier/blob/main/6-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results



	Logistic Regression	SVM	Tree	KNN
Training	0.834286	0.834286	0.861905	0.860952
Testing	0.833333	0.833333	0.944444	0.833333

```
log_report = classification_report(Y_test, logreg_cv.predict(X_test))
print(log_report)
```

```
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         6
           1       0.80      1.00      0.89        12

    accuracy                           0.83        18
   macro avg       0.90      0.75      0.78        18
weighted avg       0.87      0.83      0.81        18
```

```
svm_report = classification_report(Y_test, svm_cv.predict(X_test))
print(svm_report)
```

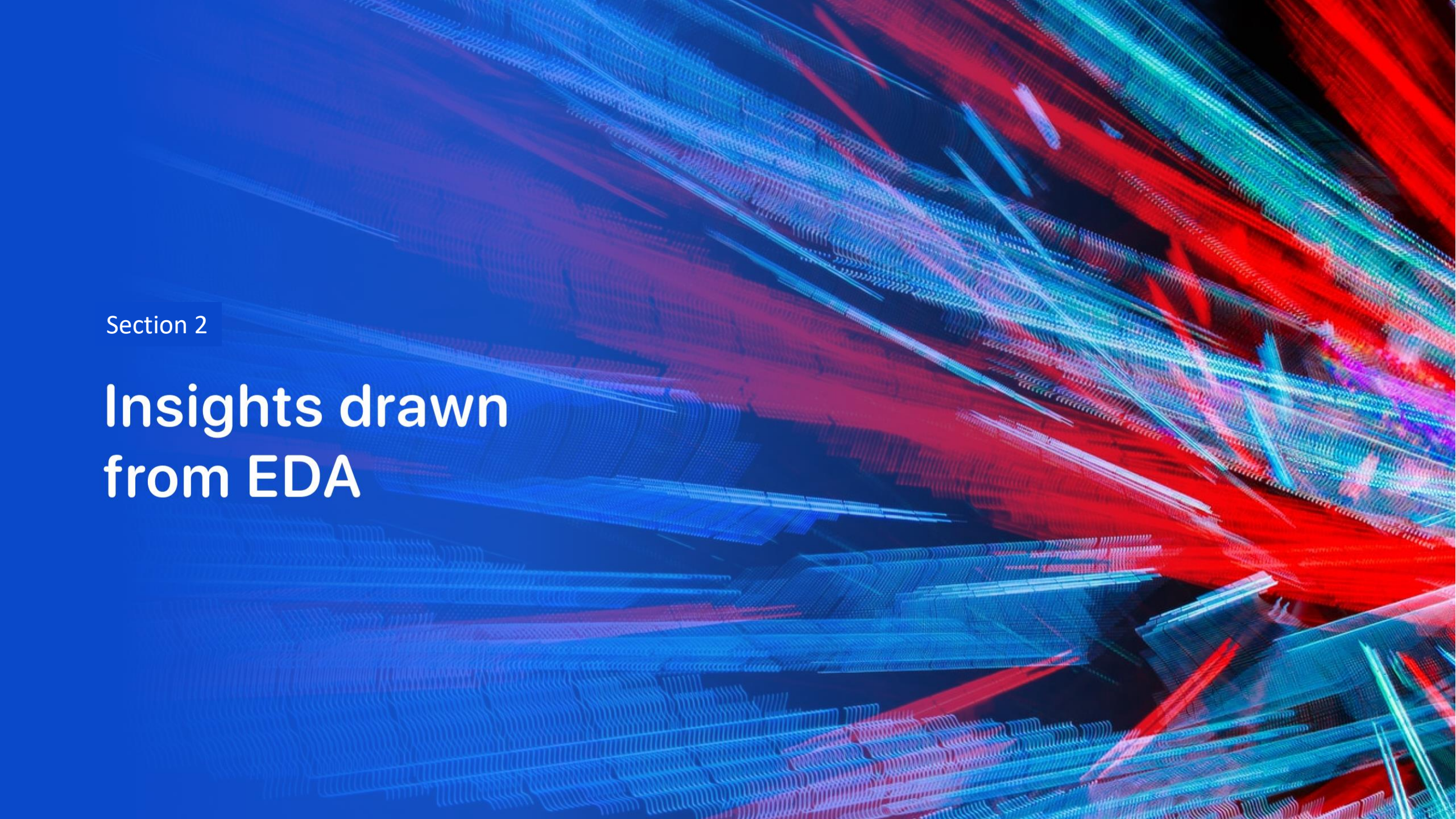
```
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         6
           1       0.80      1.00      0.89        12

    accuracy                           0.83        18
   macro avg       0.90      0.75      0.78        18
weighted avg       0.87      0.83      0.81        18
```

```
tree_report = classification_report(Y_test, tree_cv.predict(X_test))
print(tree_report)
```

```
              precision    recall  f1-score   support

           0       1.00      0.83      0.91         6
           1       0.92      1.00      0.96        12

    accuracy                           0.94        18
   macro avg       0.96      0.92      0.93        18
weighted avg       0.95      0.94      0.94        18
```

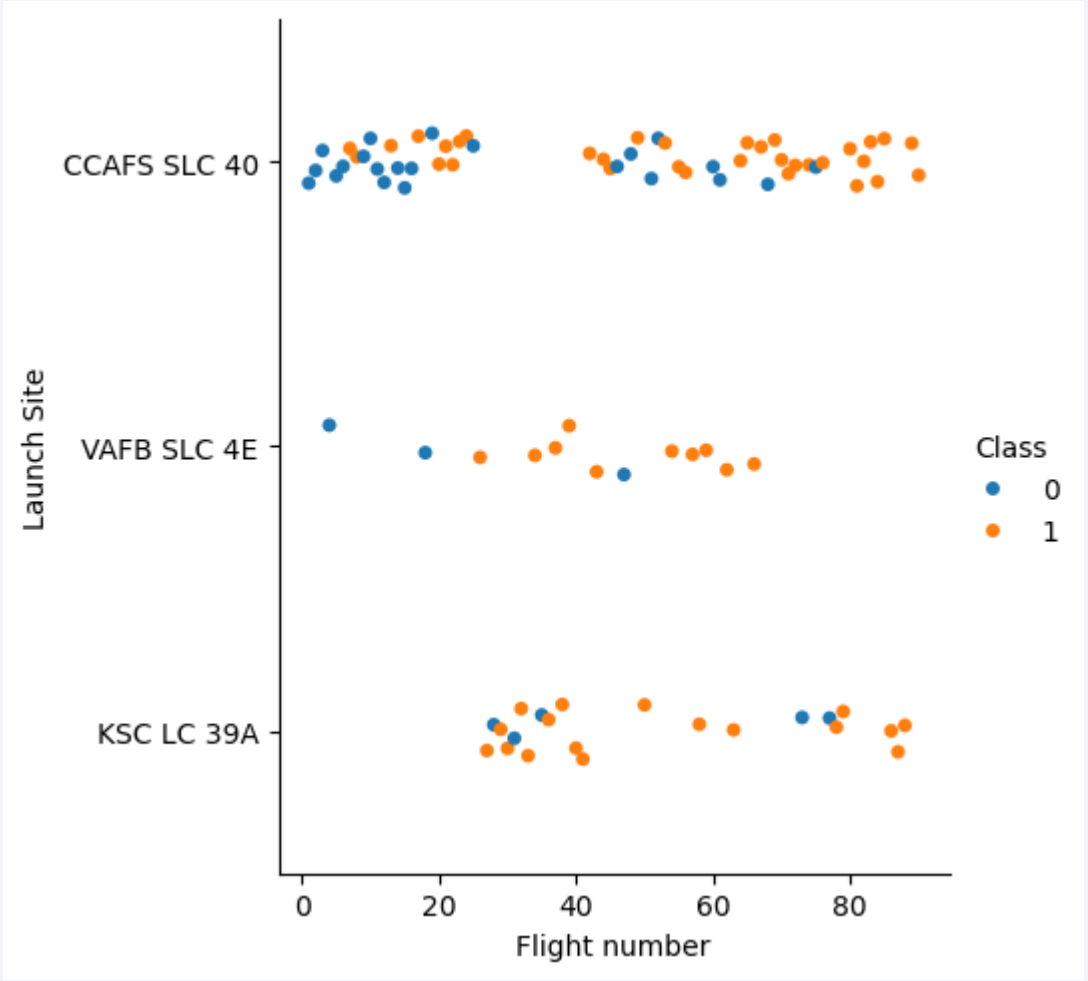
```
knn_report = classification_report(Y_test, knn_cv.predict(X_test))
print(knn_report)
```

```
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         6
           1       0.80      1.00      0.89        12

    accuracy                           0.83        18
   macro avg       0.90      0.75      0.78        18
weighted avg       0.87      0.83      0.81        18
```


Section 2
Insights drawn from EDA

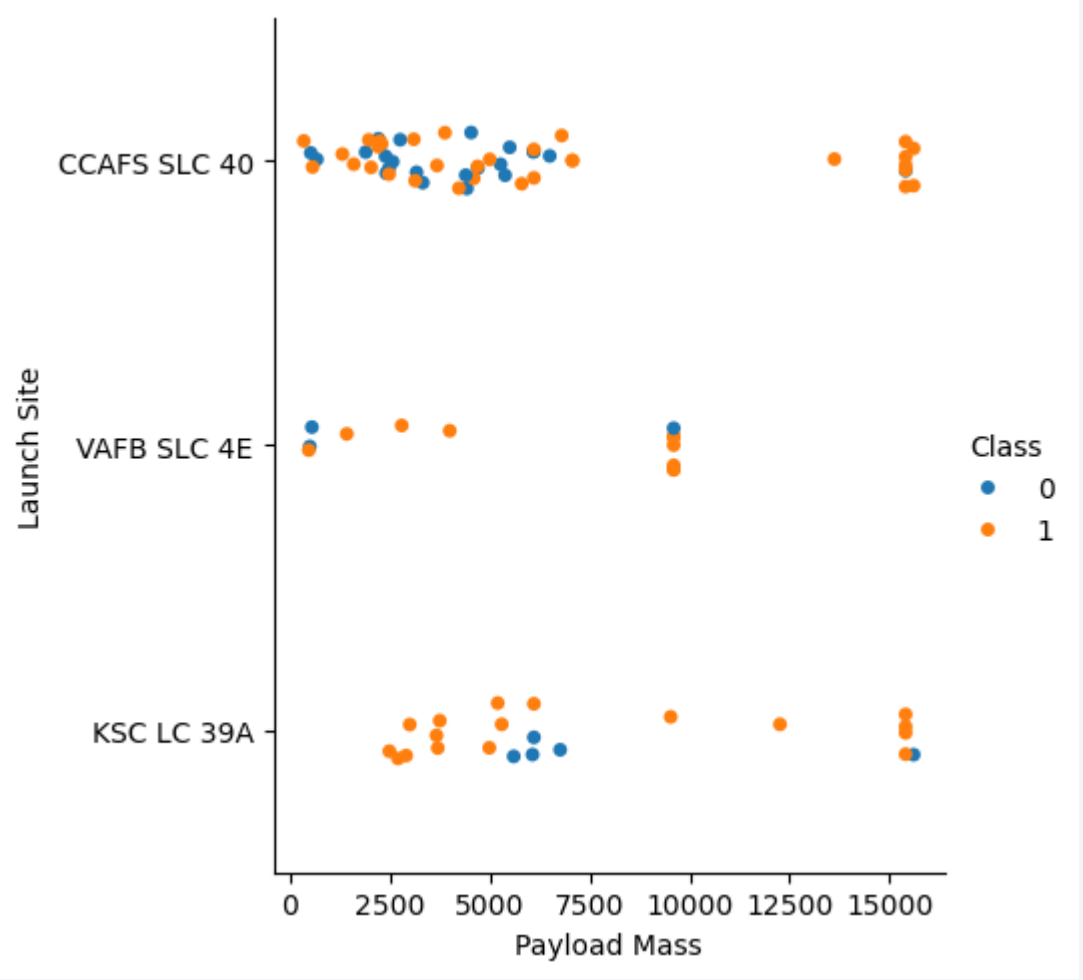
Flight Number vs. Launch Site

- CCAFS SLC 40 launch site has an unpredictable success performance as flight numbers increase.
- VAFB SLC 4E & KSC LC 39A both have an increase in success rate as flight number increases.
- Overall as flight number increases, there is an increase in successful launches for each of the launch sites.



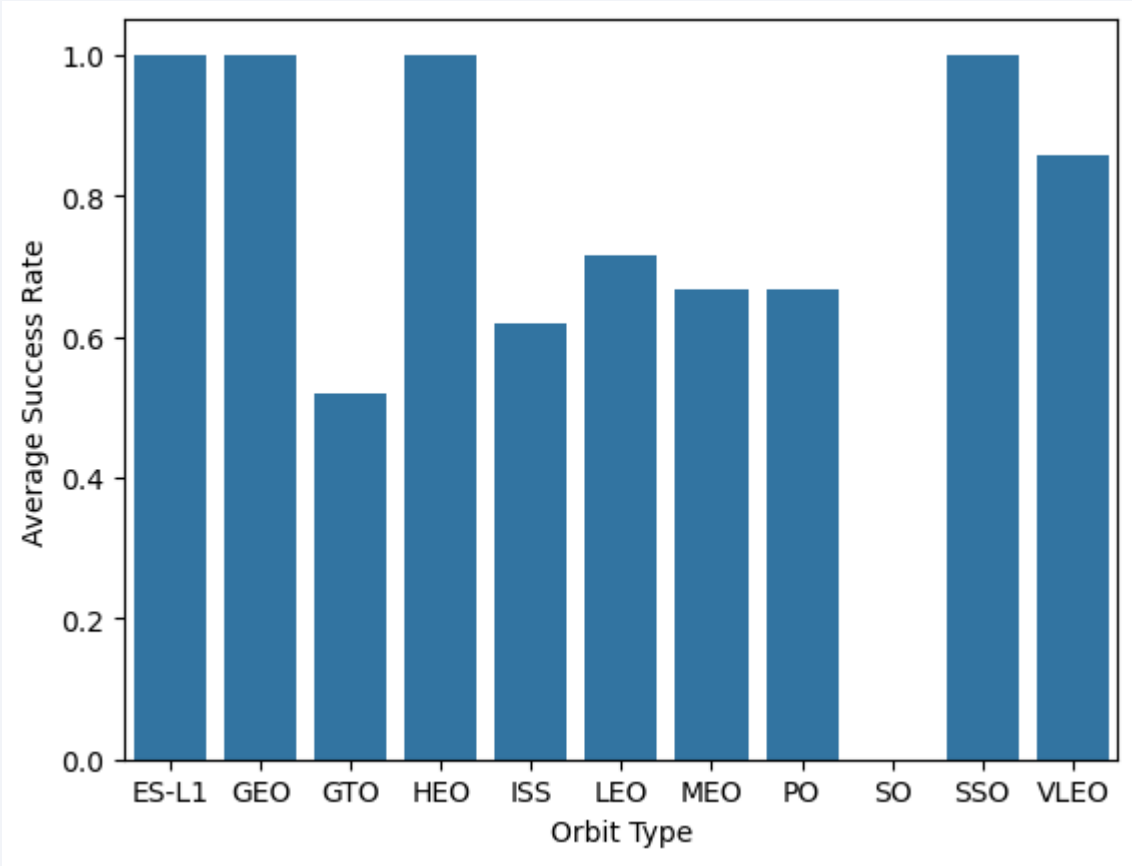
Payload vs. Launch Site

- CCAFS SLC 40 has a high success rate for heavy payloads
- VAFB SLC 4E & KSC LC 39A both have are compatible with a wide range of payload mass but ideally have a higher success rate with low to mid range payloads.



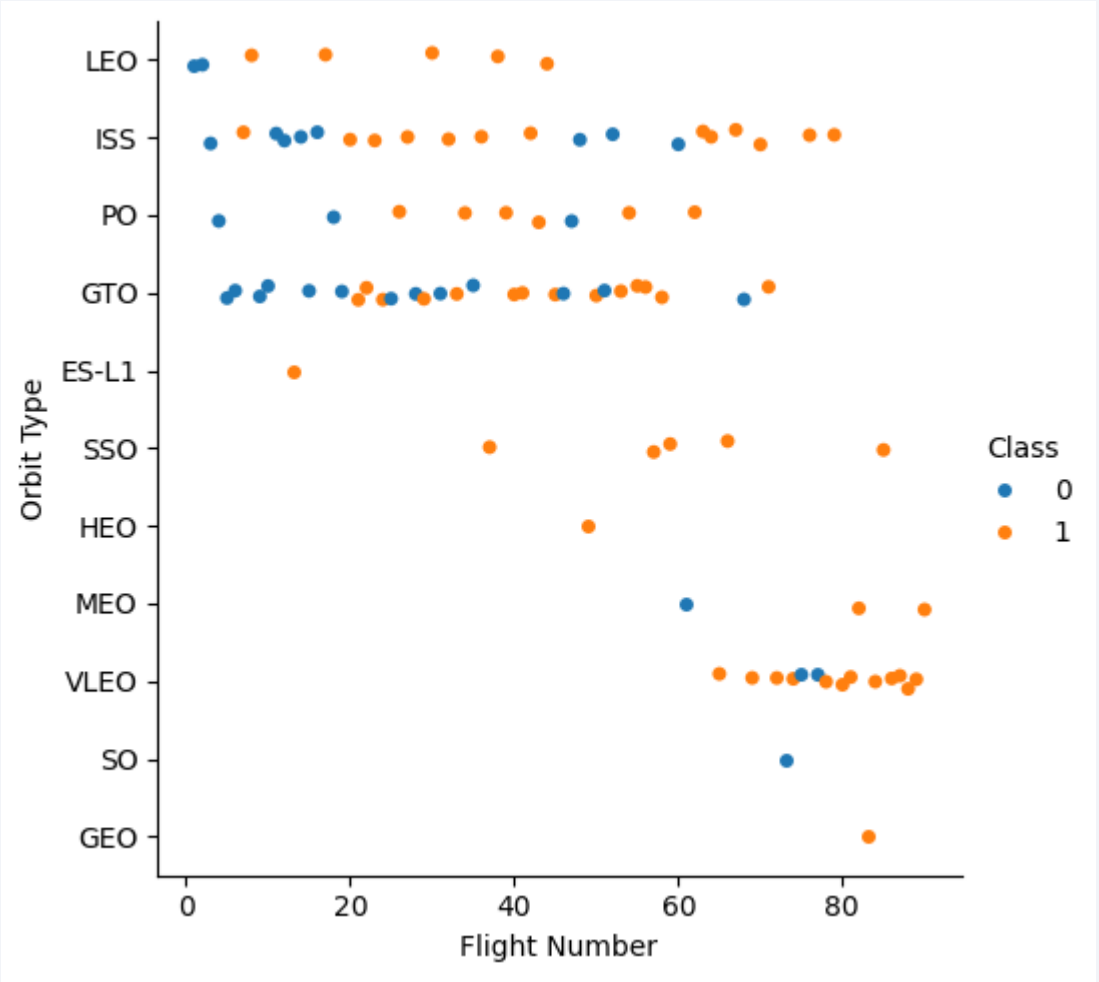
Success Rate vs. Orbit Type

- The top performing orbit types are ES-L1, GEO, HEO, SSO & VLEO with a success rate of 80% and higher.
- Orbit types GTO, ISS, LEO, MEO, & PO have an average success rate of 50%-70%.
- Least performing orbit is the SO orbit type with 0%.
- Overall the top performing orbit types which have a success rate of 80% and above will be the ideal selection.



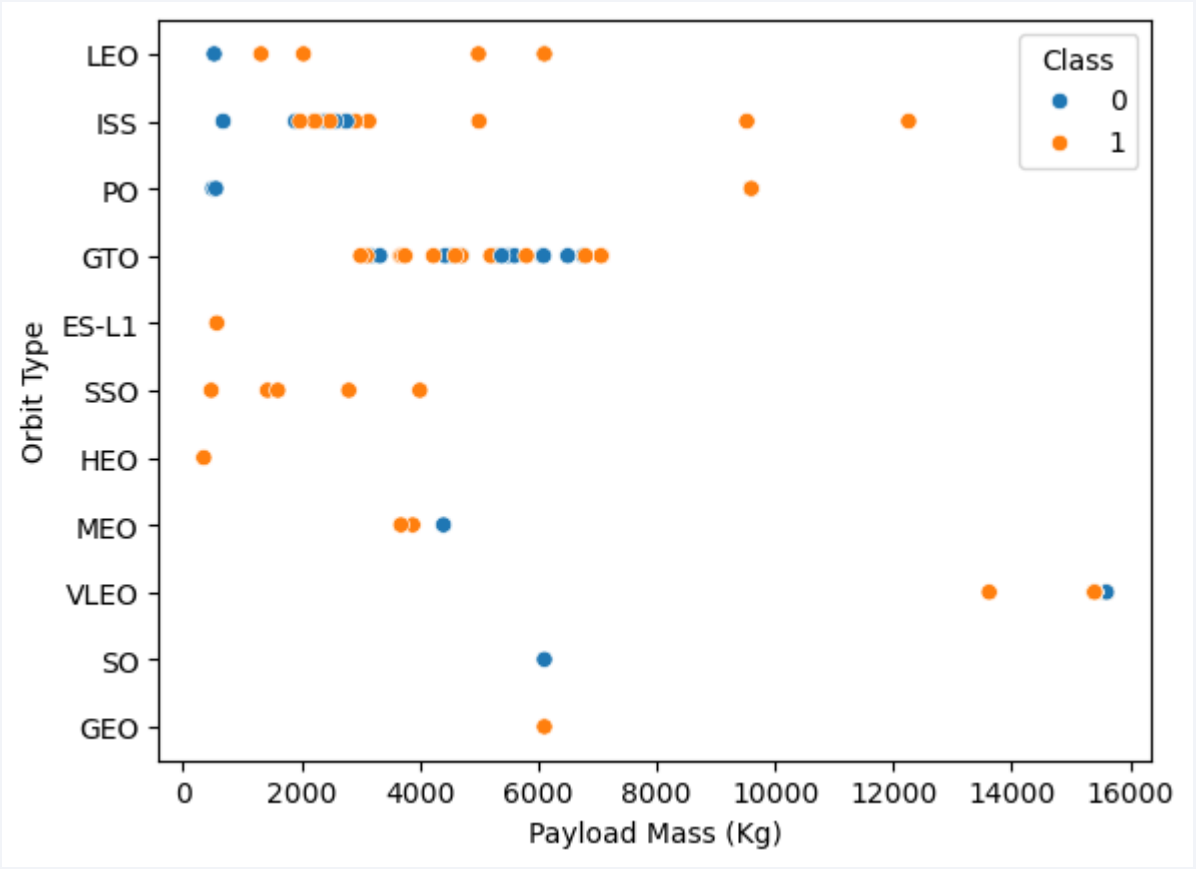
Flight Number vs. Orbit Type

- We can observe that for the majority of orbit types as flight number increases there is an increase in success rate as well.
- In the case of GTO and SO orbit types, these have a low success rate.



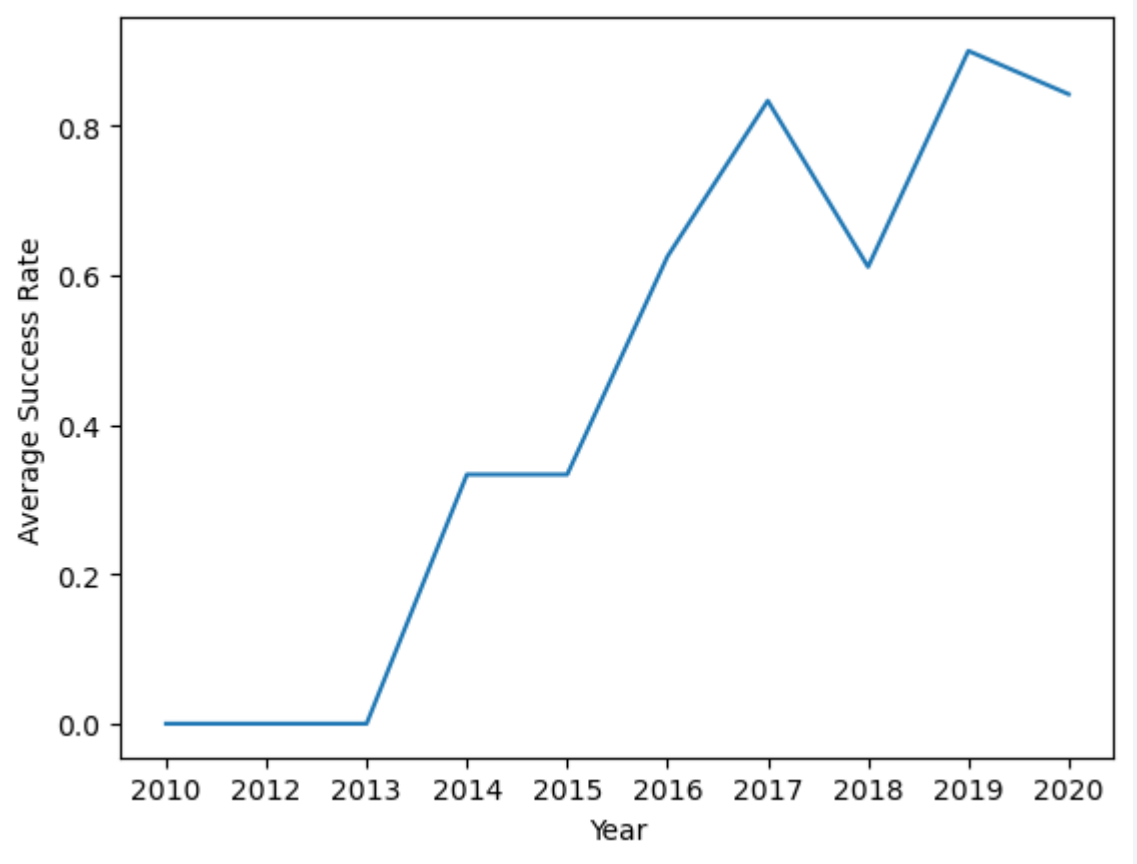
Payload vs. Orbit Type

- LEO, ISS, & PO have a higher success rate with heavy payloads.
- ES-L1, SSO, & HEO have a higher success rate with low payload mass.
- Whereas GTO and others cannot be distinguished as the success rate varies as mass increases or decreases.
- Overall for heavy payloads LEO, ISS & PO are favorable. For low mass payloads ES-L1, SSO, & HEO are favorable.



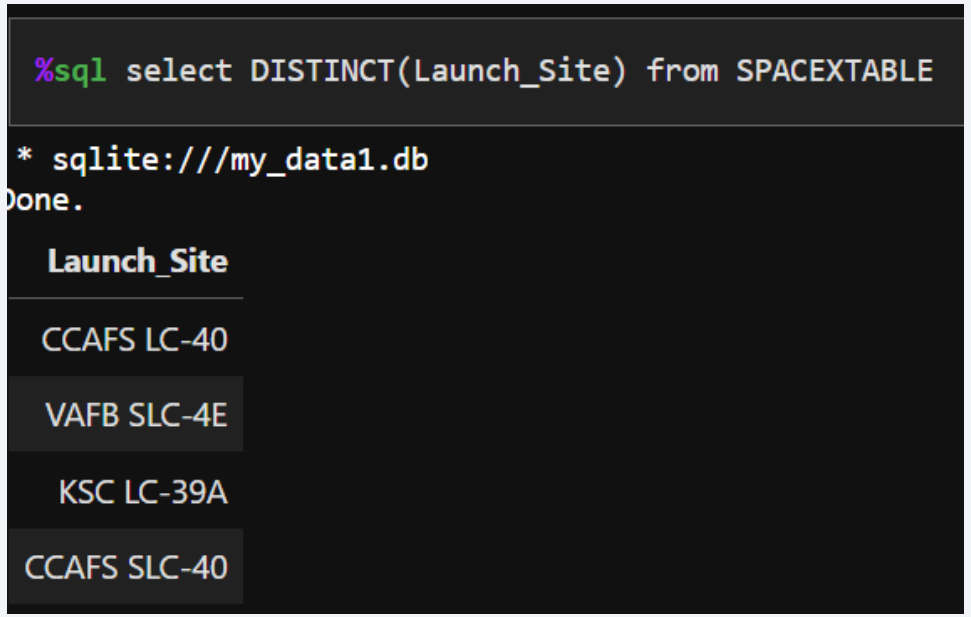
Launch Success Yearly Trend

As the years pass by, the success rate has been increasing, indicating an improvement in launches of the first stage of the falcon 9 rocket.



All Launch Site Names

- The names of the launch site can observed in the figure.
- In order to extract unique non-repetitive launch site names, the DISTINCT keyword is used such that there are no redundant launch sites.



Launch Site Names Begin with 'CCA'

- The sql query uses the WHERE and LIKE keywords to apply a conditional statement when extracting the required data, since the goal is to iterate over launch sites and select the ones that begin with “CCA” the query defined does as such.

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
one.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

- To compute the total payload mass launched by a specific customer, a conditional statement is induced. To compute the total payload the values should be added which in this case the function SUM is used to compute the total for the values extracted.

Task 3

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

```
%sql select SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass" FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
```

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

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

- To compute the Average payload by booster version F9 v1.1 a conditional statement has been induced along with the use of the function AVG to compute the average.

Task 4

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

```
1 SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) AS "Average payload mass" FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1"
```

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

Booster_Version	Average payload mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- To extract the first successful landing achieved on the ground pad, a filter is applied to search for the case, and the MIN function which computes the minimum is applied on the launch date to obtain the first successful landing.

Task 5

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

Hint:Use min function

```
%sql select MIN(DATE) as "First Successful landing", * from SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)"
```

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

First Successful landing	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_(
2015-12-22	2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Succes

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters landed are the F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

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

```
from spacextable WHERE (Landing_Outcome = "Success (drone ship)" AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000)
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The Total number of successful missions are 100 missions whereas the total number of failed missions is 1.

Task 7

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, COUNT(Mission_Outcome) AS "Total Number" from spacextable GROUP BY  Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- The maximum payload is 15600, and the boosters are shown below in the figure.
- Given that the maximum payload is to found and is not defined a sub query is used to iterate over the data to find the max and use that value as the value to compare with. The query can be defined as shown below.

```
%sql select DISTINCT(Booster_Version), PAYLOAD_MASS__KG_ as "Max payload" from SPACEXTABLE WHERE
PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

Month	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

```
%sql select SUBSTR(DATE, 6, 2) AS "Month", Landing_Outcome, Booster_Version, Launch_Site from spacextable WHERE DATE LIKE "2015%" AND Landing_Outcome = "Failure (drone ship)"
```

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

- The query to rank the count of landing outcomes between the given dates in descending order is defined as shown below, and the extracted data shows the highest landing outcome count which is 10 all the way to 1which is the lowest.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	Landing Outcome Count
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt	10
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)	5
2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)	5
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)	3
2014-04-18	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)	NASA (CRS)	Success	Controlled (ocean)	3
2013-09-29	16:00:00	F9 v1.1 B1003	VAFB SLC-4E	CASSIOPE	500	Polar LEO	MDA	Success	Uncontrolled (ocean)	2
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)	2
2015-06-28	14:21:00	F9 v1.1 B1018	CCAFS LC-40	SpaceX CRS-7	1952	LEO (ISS)	NASA (CRS)	Failure (in flight)	Precluded (drone ship)	1

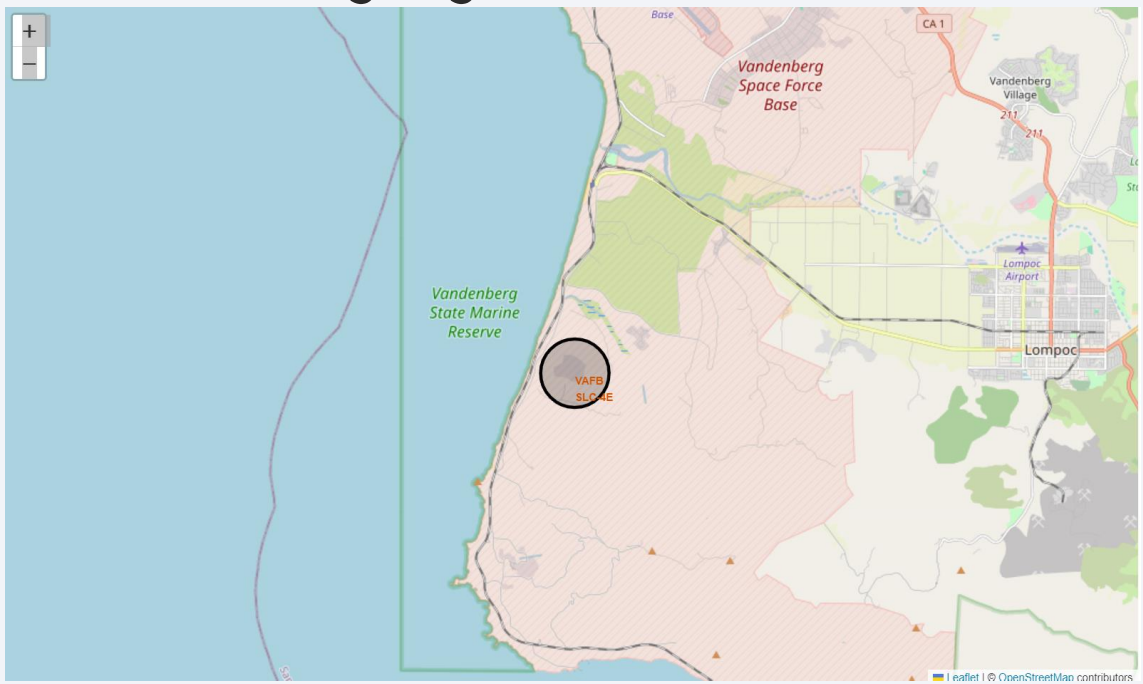
```
%sql SELECT *, COUNT(Landing_Outcome) AS "Landing Outcome Count" FROM SPACEXTABLE WHERE (DATE >= "2010-06-04" AND DATE <= "2017-03-20") GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC
```

Section 3

Launch Sites Proximities Analysis

Launch Sites marked on map

The figures shown below marks the 4 launch sites, on the left figure we have 1 launch site in the Los Angeles area, and on the right figure we have 3 launch sites in the Florida region. There are two launch sites that overlap hence the two intersecting regions.

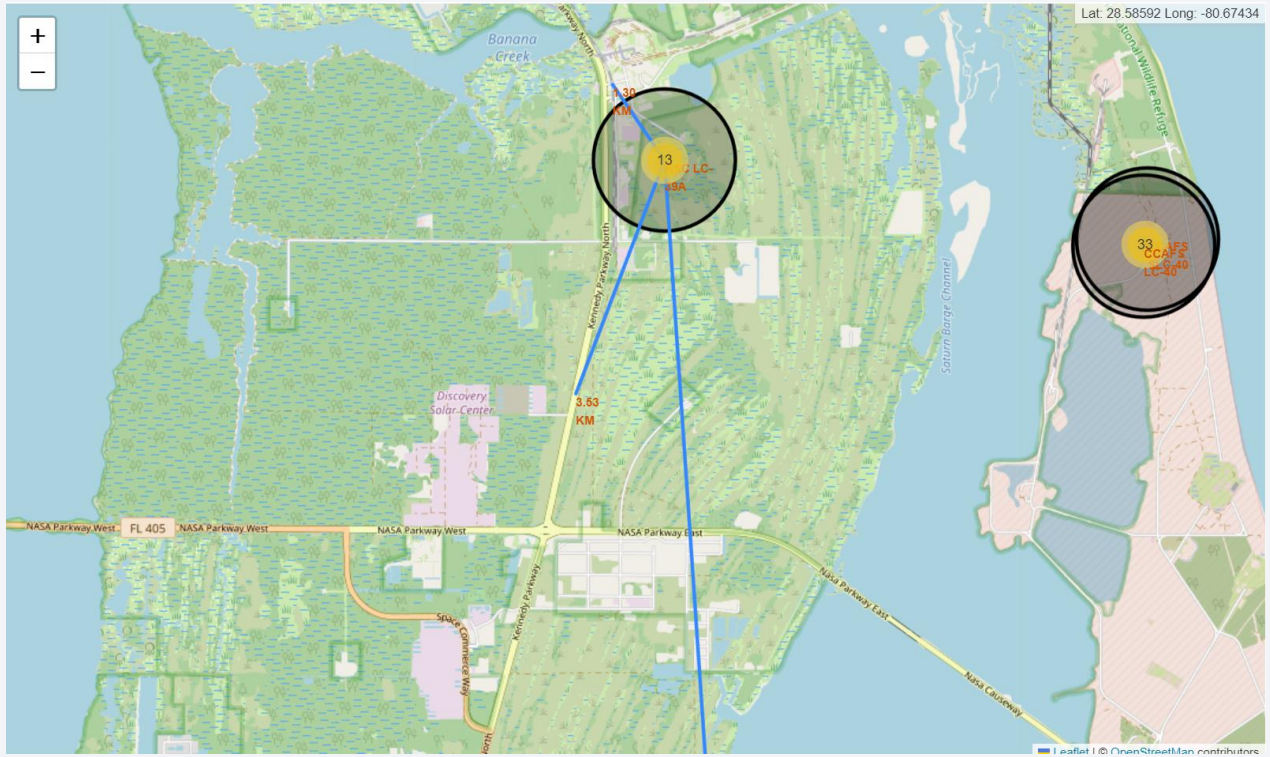


Successful/Failed Launches Mapped

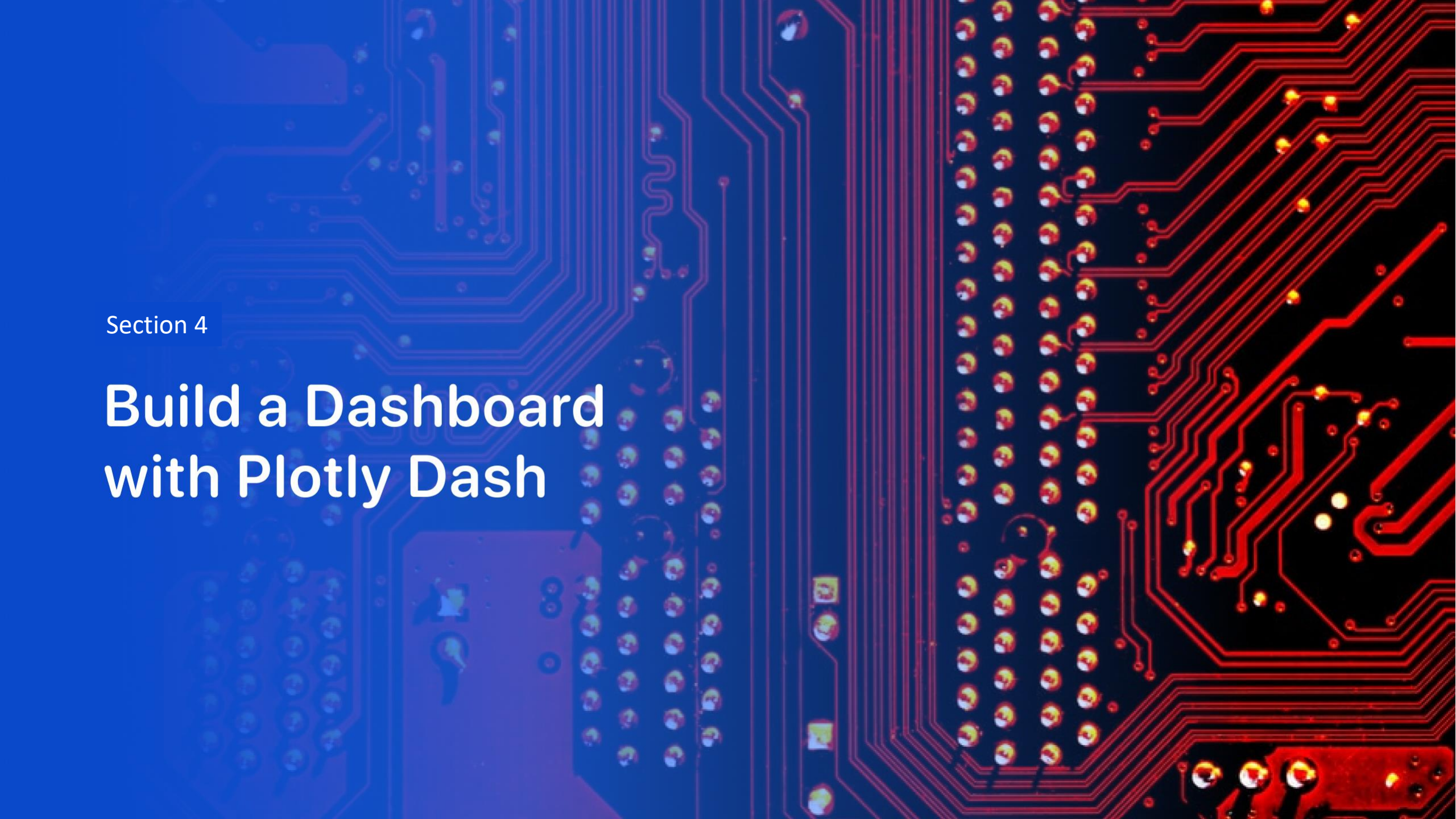
The figures shown below represent 3 out of the 4 launch sites which have been marked with markers highlighting successful and failed launches.

Launch Site Areas proximity distance measurements

- The figure shown on the right is a representation highlighting the distances computed between one of the launch sites with respect to a railway, highway, and city.
- The closest is the railway, then highway, and the furthest is the city.

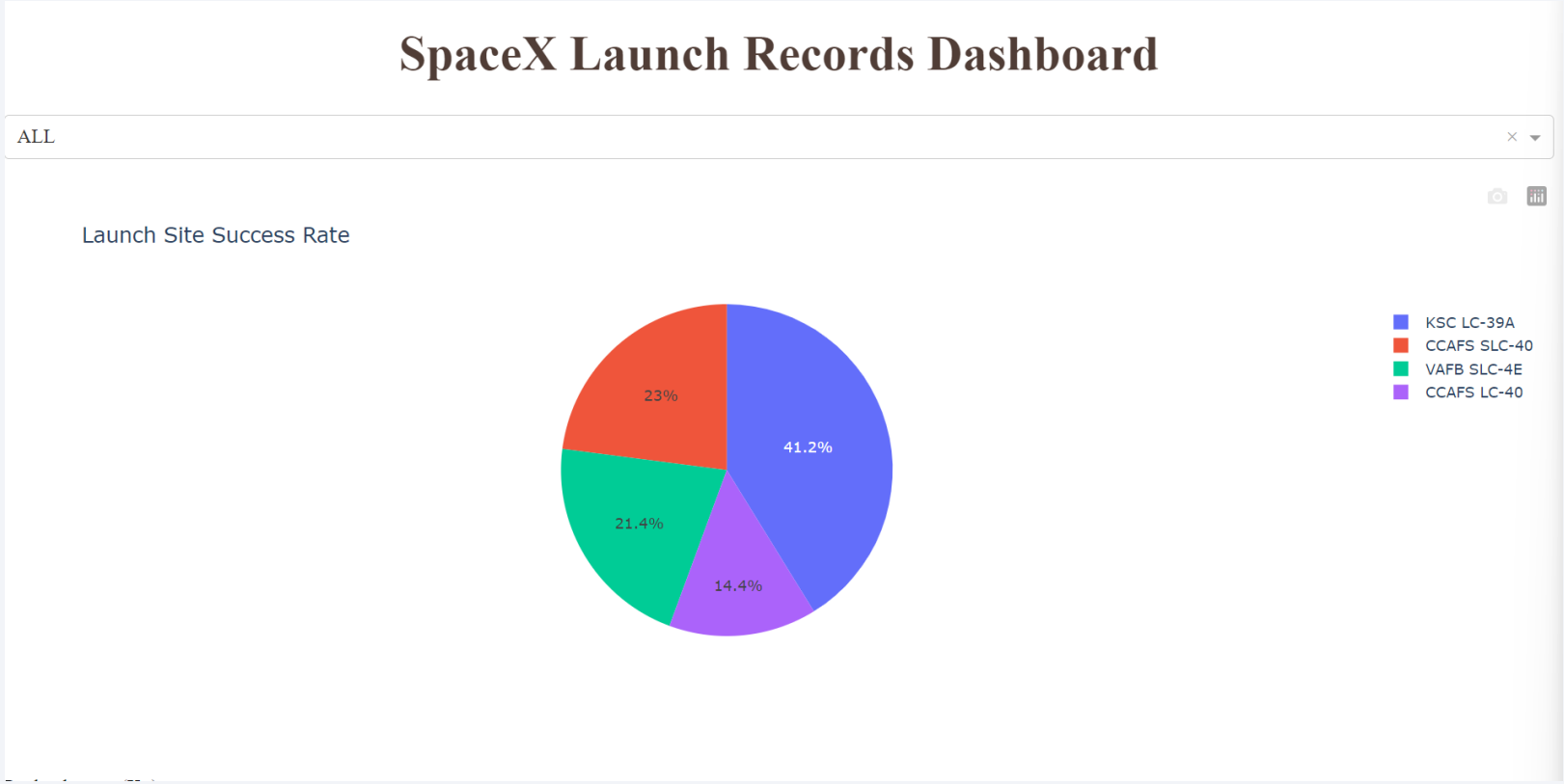


Location	Distance from Launch site KSC LC-39A
Railway	1.3 KM
Highway	3.53 KM
City	55.66 KM


Section 4
Build a Dashboard with Plotly Dash

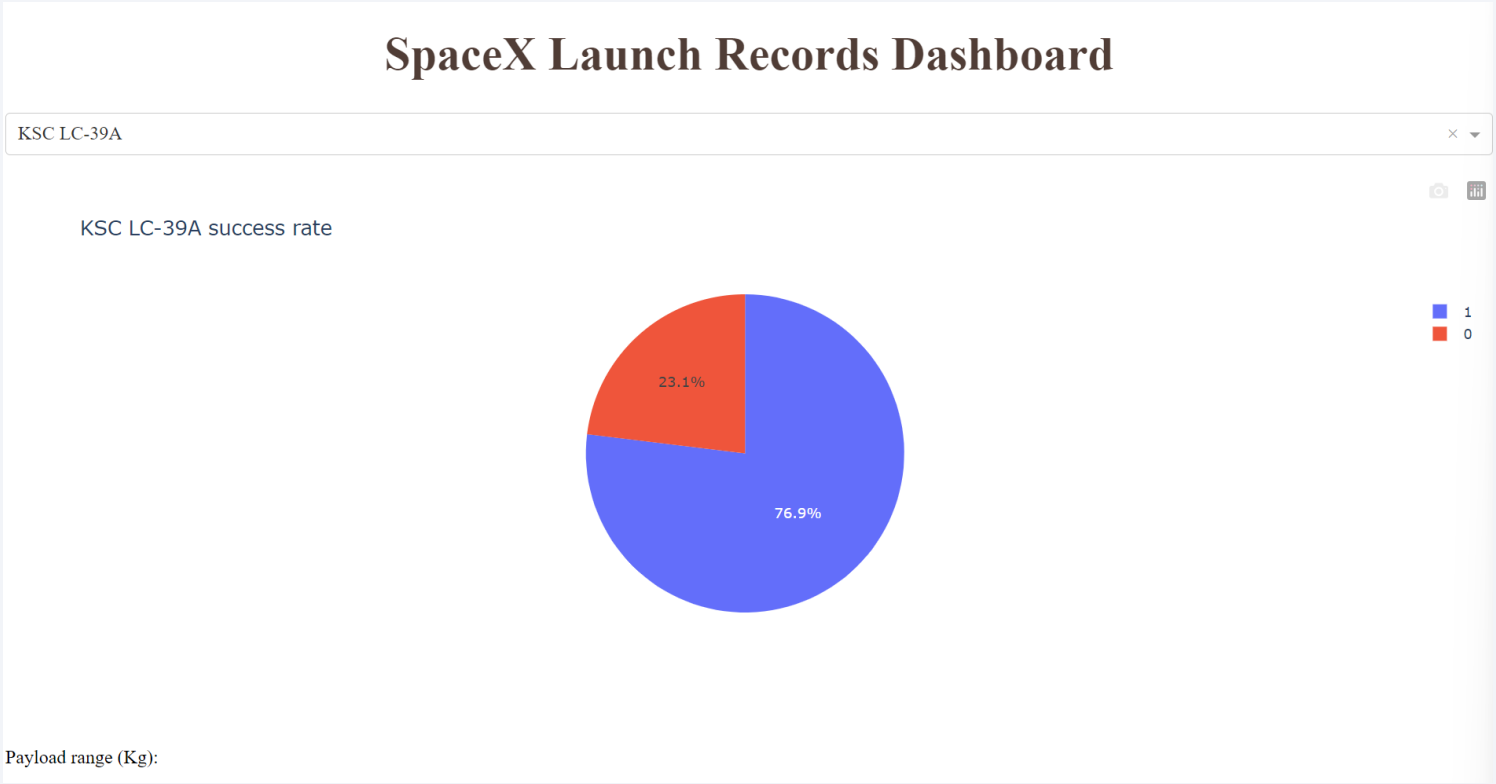
Launch Site Success Rate

- KSC LC-39A is the launch site with the most successful launches
- CCAFS LC-40 is the launch site with the least successful launches.



KSC LC-39A Success Rate Pie Chart

- KSC LC-39A is the launch site with the highest success rate, and as shown below it has a 76.9% success rate with a 23.1% of failed launches.

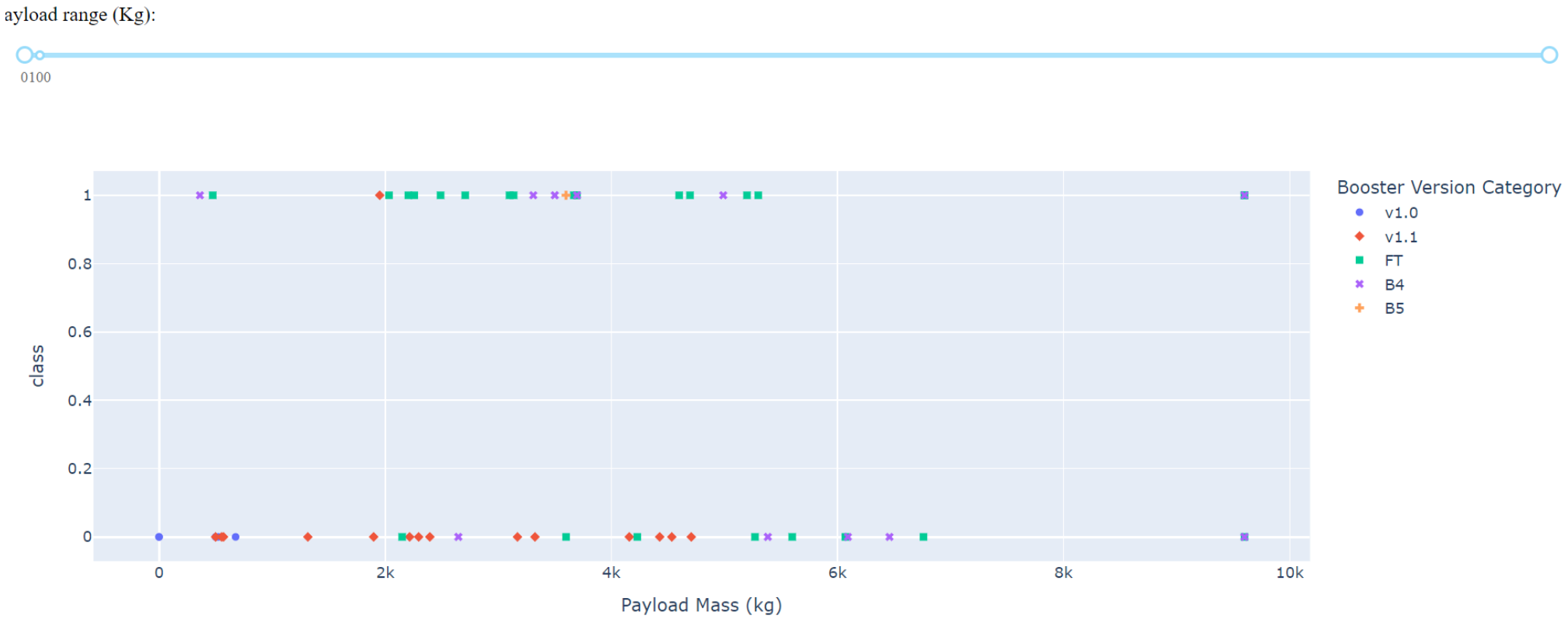


Payload Vs Launch Outcome Scatter plot – Light Payload

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

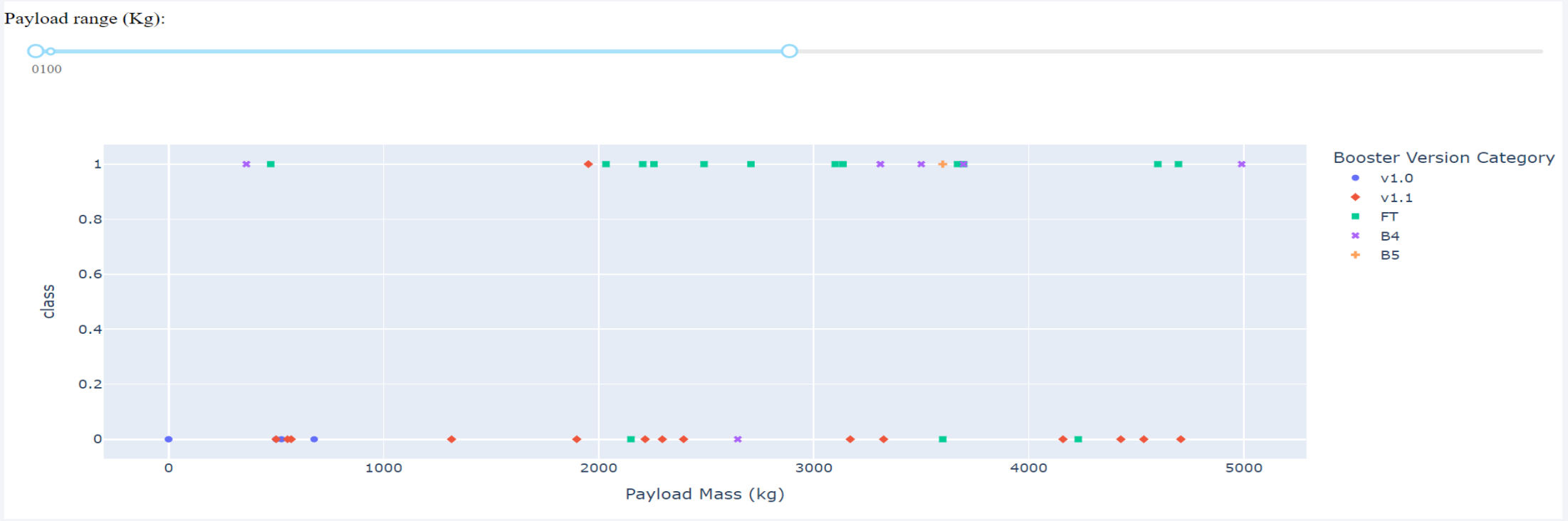
Payload Vs Launch Outcome Scatter plot

- Overall, Booster version FT and B4 have a high success rate when compared to others.



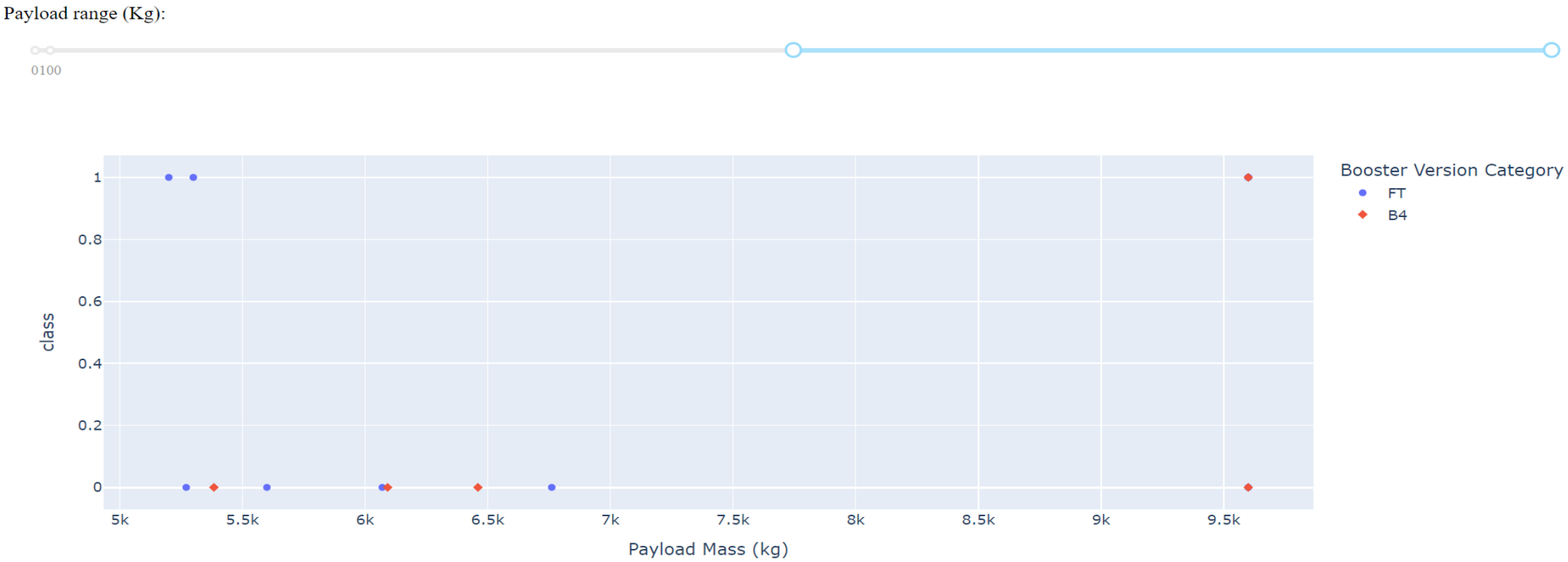
Payload Vs Launch Outcome Scatter plot – Light Payload

- Considering the light payloads ranging from 0 to 5000 Kg, booster version FT and B4 are the top performing boosters.



Payload Vs Launch Outcome Scatter plot – Heavy Payload

- Considering the Heavy payloads, ranging from 5000 to 10000 Kgs, both booster versions perform badly.



Section 5

Predictive Analysis (Classification)

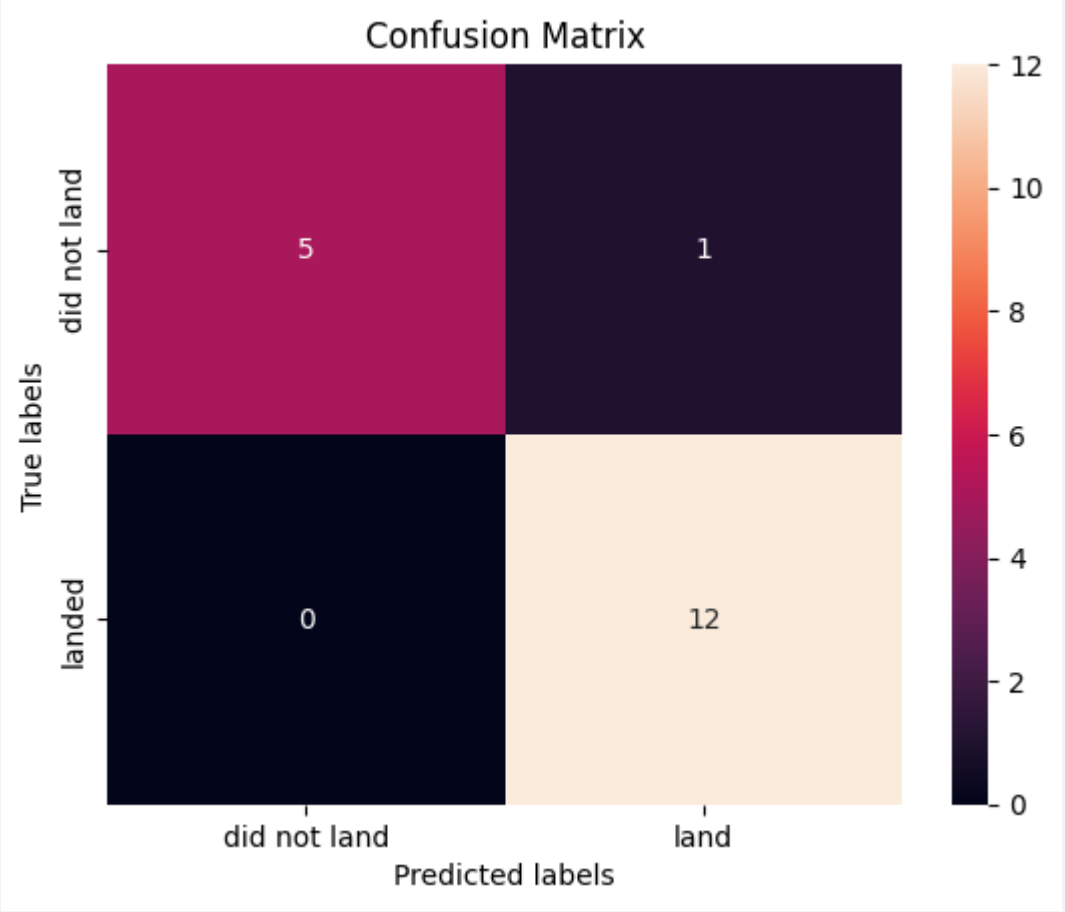
Classification Accuracy

- The figure shown represents the training and testing accuracy of each of the defined models, the best performing model is the Decision Tree with a traingin accuracy of 86% and testing accuracy of 94 %.

	Logistic Regression	SVM	Tree	KNN
Training	0.834286	0.834286	0.861905	0.860952
Testing	0.833333	0.833333	0.944444	0.833333

Confusion Matrix

- The matrix shown represents the performance of the Decision Tree model.
- We can observe that there is a 100% accuracy in classifying landing. Whereas a 83% accuracy of did not land. So there seems to be some confusion in the datapoints where a sample has been classified as landed which is incorrect.



Conclusions

- The top performing orbit types are ES-L1, GEO, HEO, SSO & VLEO with a success rate of 80% and higher. And as flight numbers increase the success rate increases as well.
- KSC LC-39A Launch Site is the best with a high success rate.
- Launch sites should be situated close to railways, coastlines, and highways facilitating transportation of vehicle, and being the furthest to cities to ensure safety of the citizens.
- Booster Version FT and B4 best performing.
- Decision Trees are best suitable for predicting the outcome of each landing based on the given parameters.

Thank you!