IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

NGOV KHEANG
2023 07 13

Outline

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

Executive Summary

❑ Summary of methodologies

- Launch data was acquired using two methods: connecting to the SpaceX API and webscraping the SpaceX wiki page.
- In the EDA phase, visualizations were built using the matplotlib and seaborn packages.
- Data was mapped using Folium and a dashboard was made using Plotly Dash.
- Created four distinct categorization models and gained accuracy ratings for each.

❑ Summary of all results

- The launch site, cargo mass, and orbit type are critical variables in determining whether a landing is successful.
- All classification models had a high accuracy score; but, as shown in the confusion matrices, they tend to anticipate false positive outcomes.
- LOGREG model shows the best accurate model

Introduction

- **Project background**

  - Challenge in high cost of space traveling
  - Compete with other companies for affordable space traveling
  - First stage of launching cost is very large, data science are needed to reuse the data from SpaceX about the successful of first stage

- **Problems**

  - Determine the price of each launch by gathering information from multiple sources from each companies such as SpaceX, Rocket Lab, and Virgin Galactic
  - Predict whether SpaceX will reuse the first stage
  - Use trained classification model to make informed decisions about the success of launch project

Section 1

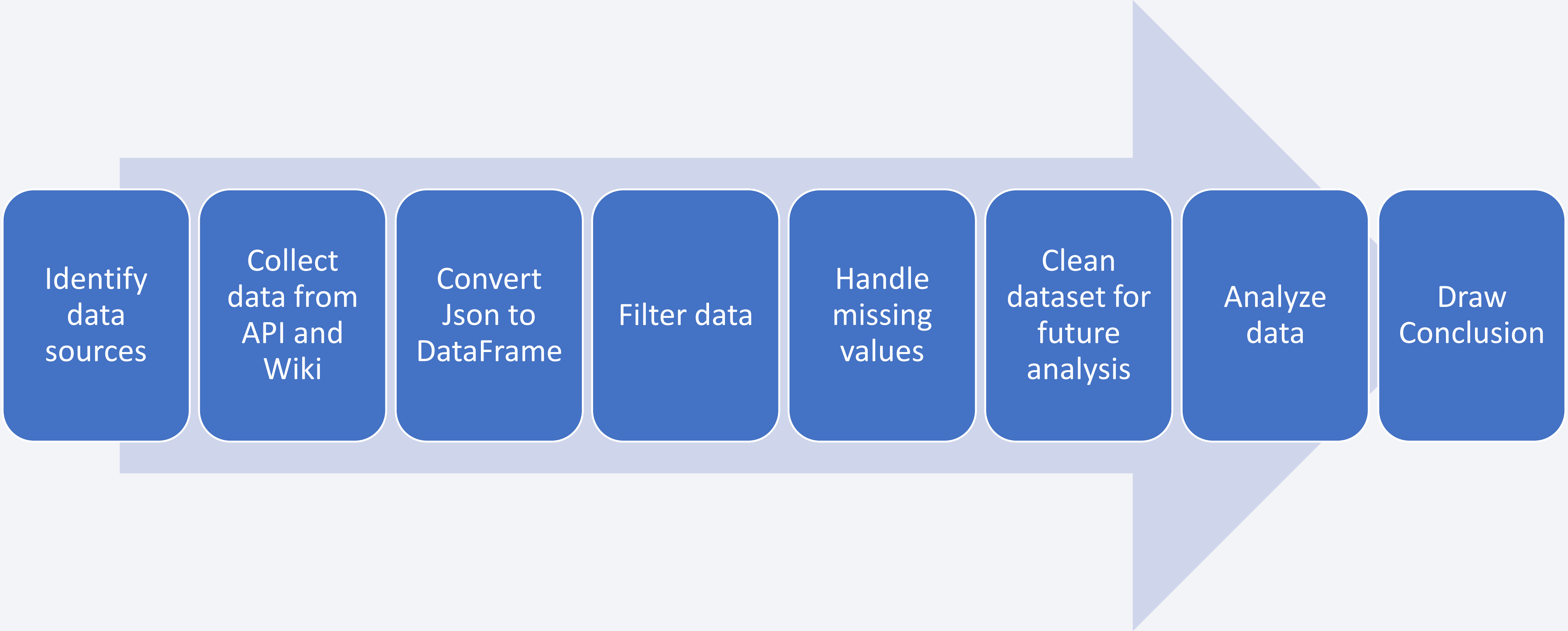
Methodology

Methodology

- Data collection methodology:
  - Data were collected from multiple sources
  - Web scrapping method were applied to collect some information from SpaceX
  - Sampling was also done to make the practice more convenient
- Perform data wrangling
  - Merging some output and input from different source
  - Preparing and clearing missing values

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistics regression, SVM, Decision Tree, and KNN were selected to experiment
  - Cross-Validation method using 10-Folds
  - Each methods, hyperparameters were tuned through GridSearch
  - Evaluation using confusion Matrix and accuracy score.

Data Collection


Identify data sources

Collect data from API and Wiki

Convert Json to DataFrame

Filter data

Handle missing values

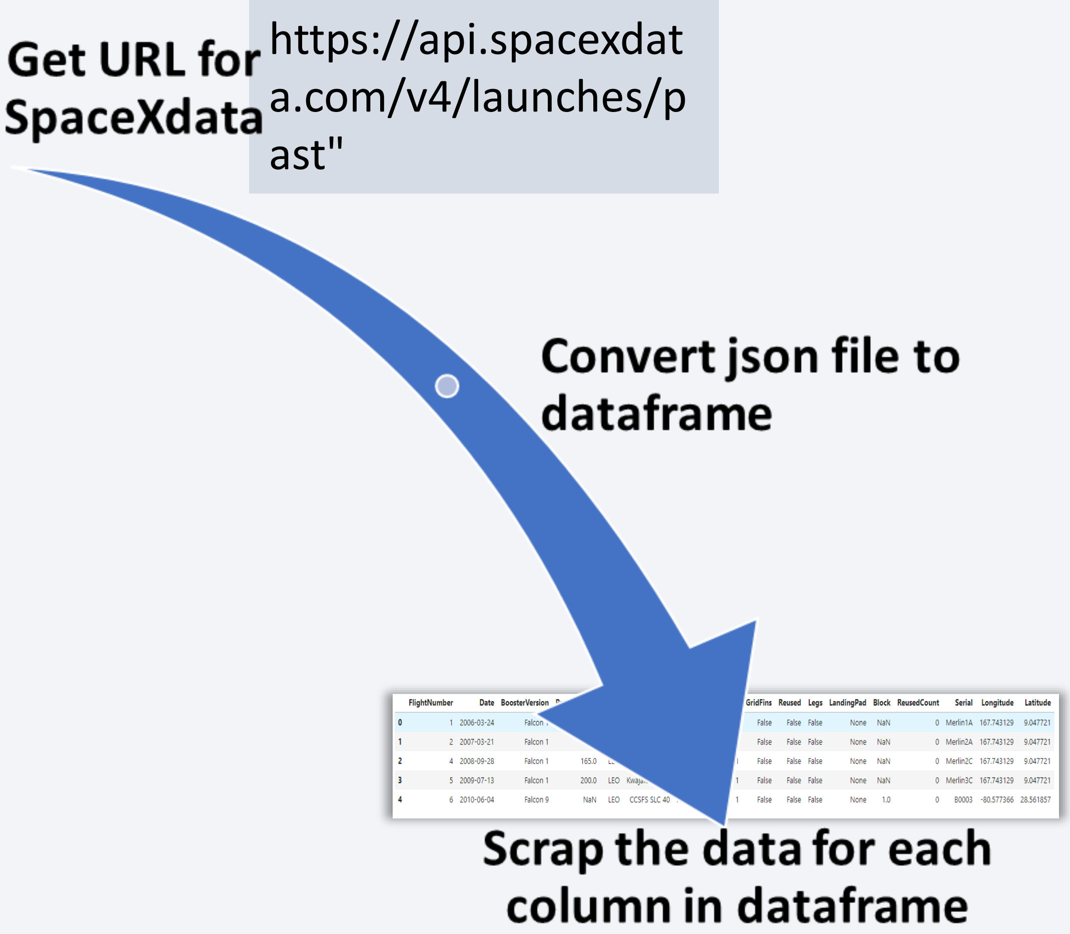
Clean dataset for future analysis

Analyze data

Draw Conclusion

Data Collection – SpaceX API

- **Send a GET request to the SpaceX API endpoint to retrieve information about previous launches.**
- **Using the pandas package, convert the JSON response into a data frame.**
- **Create routines to retrieve specific launch data (for example, launch location and payload mass).**



Get URL for SpaceXdata

https://api.spacexdata.com/v4/launches/past"

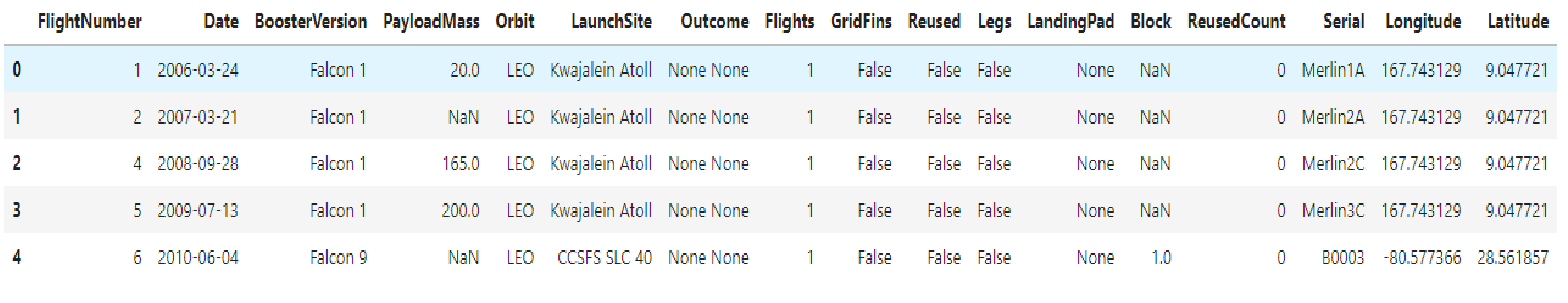
Convert json file to dataframe

	FlightNumber	Date	BoosterVersion	P[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	[illegible]	[illegible]	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwaja[illegible]	[illegible]	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	[illegible]	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Scrap the data for each column in dataframe

Data Collection - Scraping

- **Save the extracted data in a dictionary and then convert it back to a data frame.**
- **Filter the data frame such that only Falcon 9 launches are included.**
- **The mean value should be used to replace any missing values in the data frame.**



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data Wrangling

-Clean Missing Values in Dataframe

-One hot encoding for categorical variables is applied

-Identify target class for label for classification model

- 90 rows and 83 columns are derived.

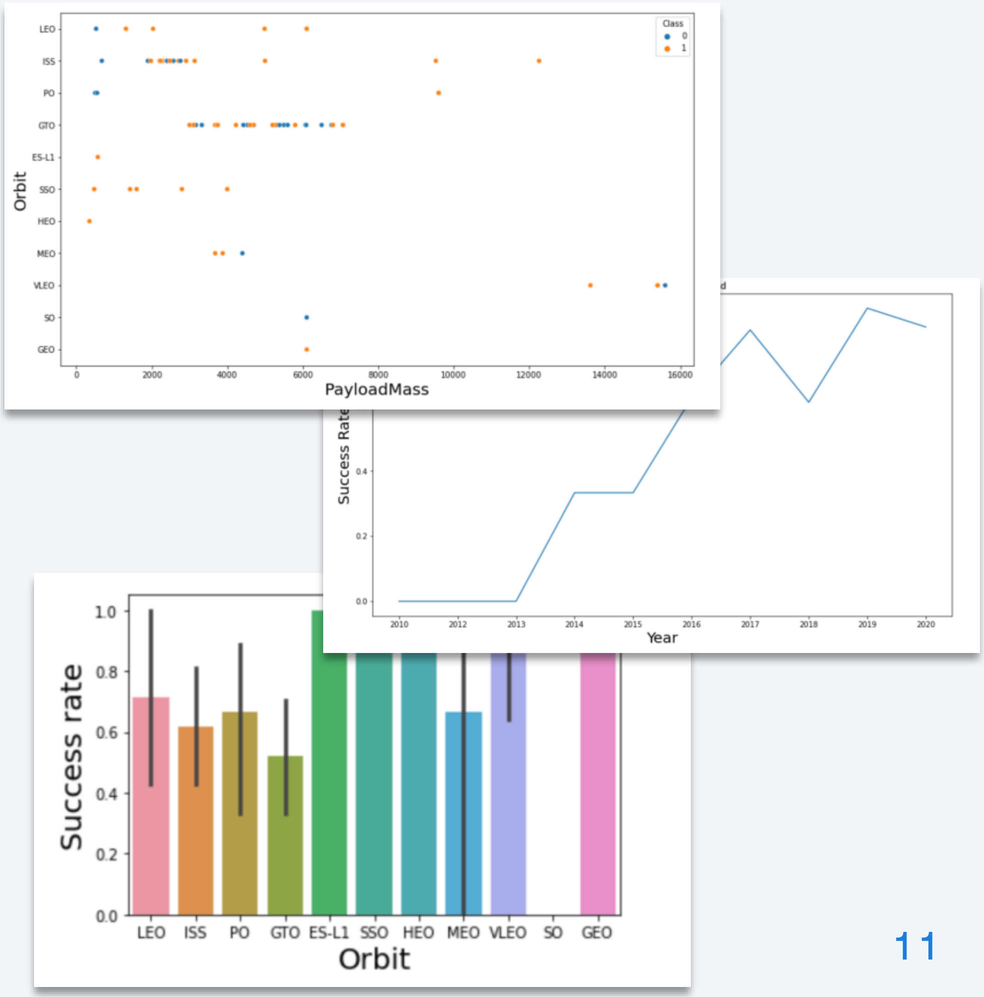
EDA with Data Visualization

Below plots are used in datavisulization:

- Scatter plot : useful for visualizing patterns, correlations, or clusters in the data.

- Line graph: show trends or changes in data

- Bar chatt:comparing data across different categories and identifying relative differences or trend



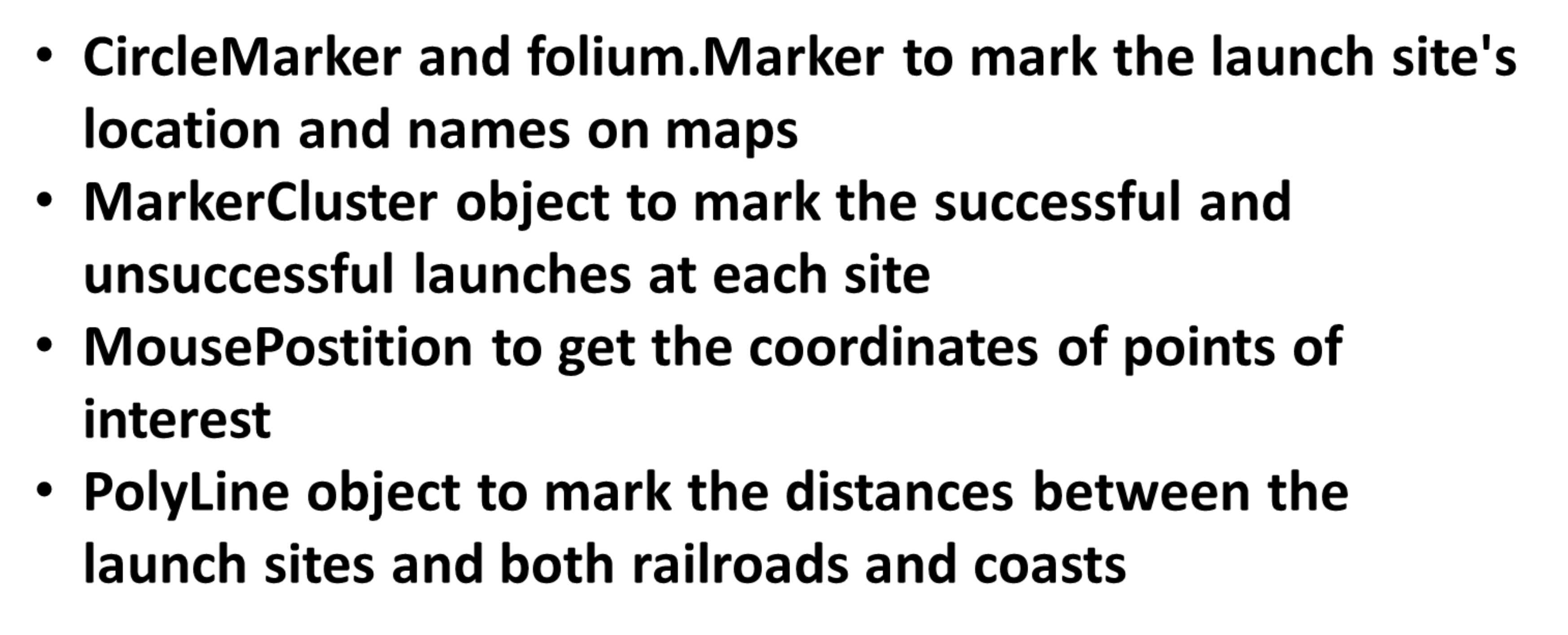
EDA with SQL

- cur.execute('PRAGMA table_info(SPACEXTBL)').fetchall()
- cur.execute('SELECT DISTINCT "Launch_Site" FROM SPACEXTBL').fetchall()
- cur.execute('SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE "CCA%" LIMIT 5').fetchall()
- cur.execute('SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Customer" LIKE "NASA (CRS)"').fetchall()
- cur.execute('SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" LIKE "F9 v1.1"').fetchall()
- cur.execute('SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)"').fetchall()
- cur.execute('SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)" AND "Payload_Mass__kg_" > 4000 AND "Payload_Mass__kg_" < 6000').fetchall()
- cur.execute('SELECT "Mission_Outcome", COUNT(*) as "Total" FROM SPACEXTBL GROUP BY "Mission_Outcome"').fetchall()
- cur.execute('SELECT "Booster_Version" FROM SPACEXTBL WHERE "Payload_Mass__kg_" = (SELECT MAX ("Payload_Mass__kg_") FROM SPACEXTBL)').fetchall()
- "')

EDA with SQL

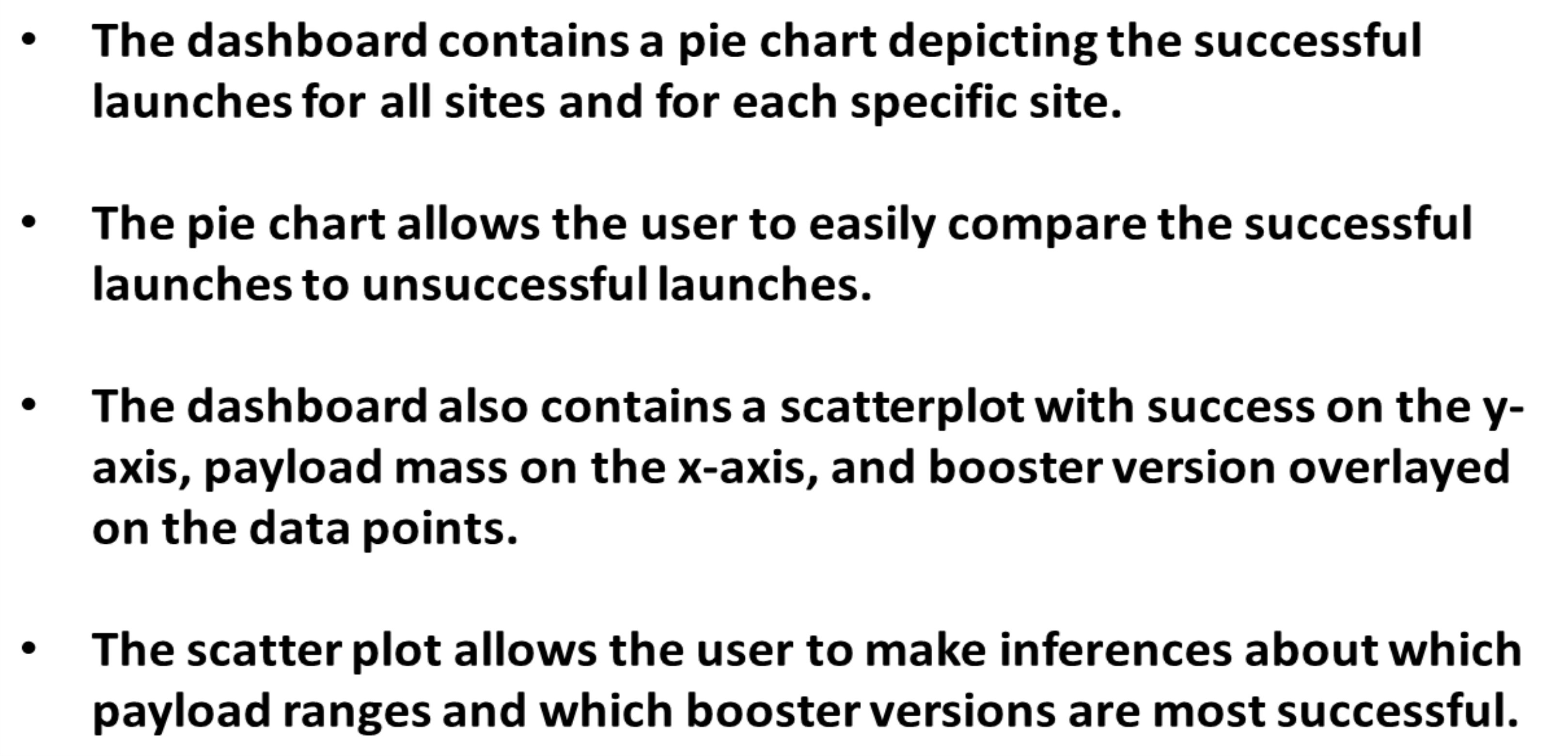
```
cur.execute('''
    SELECT
        CASE SUBSTR("Date", 4, 2)
            WHEN '01' THEN 'January'
            WHEN '02' THEN 'February'
            WHEN '03' THEN 'March'
            WHEN '04' THEN 'April'
            WHEN '05' THEN 'May'
            WHEN '06' THEN 'June'
            WHEN '07' THEN 'July'
            WHEN '08' THEN 'August'
            WHEN '09' THEN 'September'
            WHEN '10' THEN 'October'
            WHEN '11' THEN 'November'
            WHEN '12' THEN 'December'
        END AS MonthName,
        "Landing_Outcome",
        "Booster_Version",
        "Launch_Site"
    FROM SPACEXTBL
    WHERE SUBSTR("Date", 7, 4) = '2015'
        AND "Landing_Outcome" LIKE 'Failure (drone ship)'
''')
```

Build an Interactive Map with Folium



- CircleMarker and folium.Marker to mark the launch site's location and names on maps
- MarkerCluster object to mark the successful and unsuccessful launches at each site
- MousePostition to get the coordinates of points of interest
- PolyLine object to mark the distances between the launch sites and both railroads and coasts

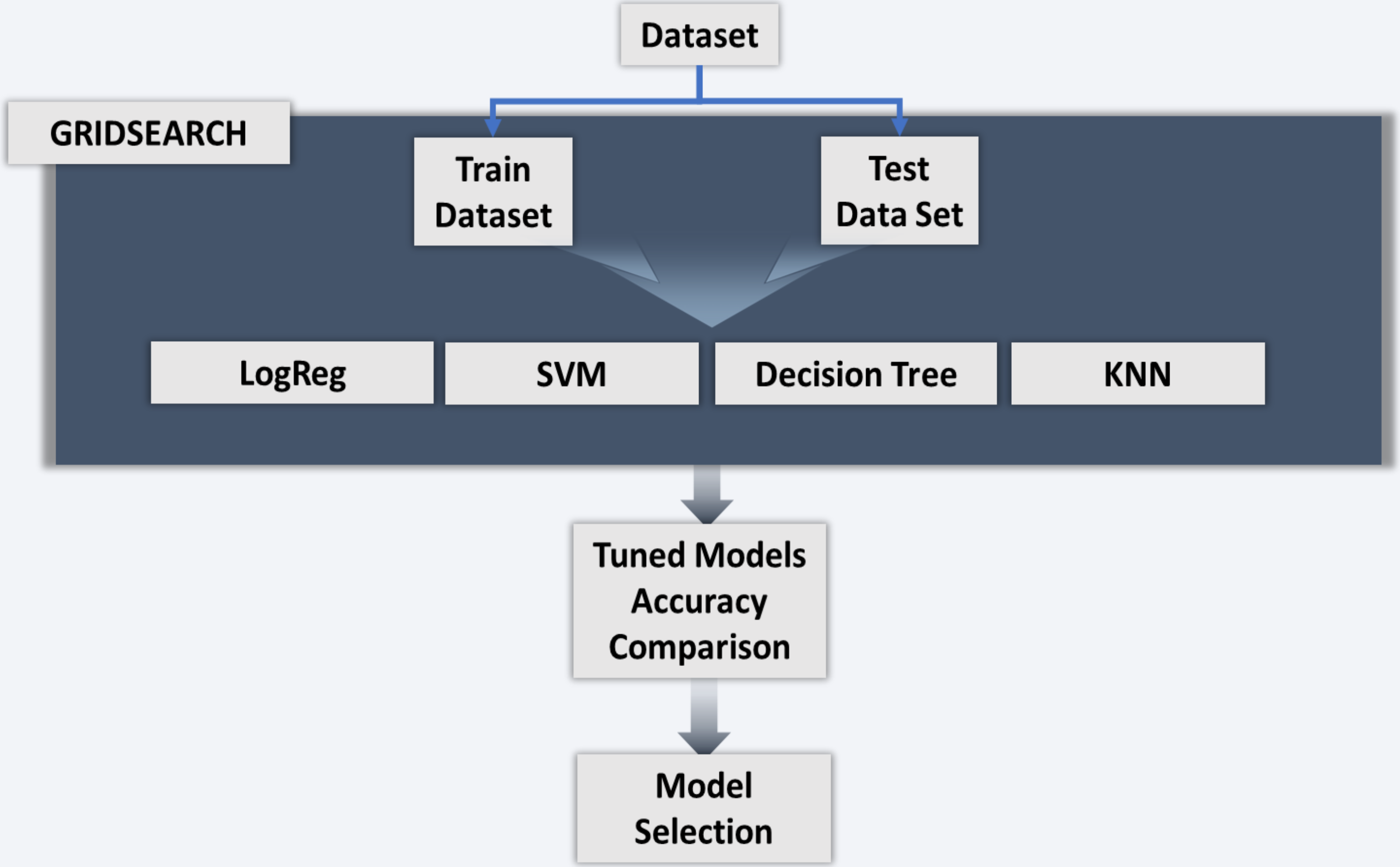
Build a Dashboard with Plotly Dash



- The dashboard contains a pie chart depicting the successful launches for all sites and for each specific site.
- The pie chart allows the user to easily compare the successful launches to unsuccessful launches.
- The dashboard also contains a scatterplot with success on the y-axis, payload mass on the x-axis, and booster version overlayed on the data points.
- The scatter plot allows the user to make inferences about which payload ranges and which booster versions are most successful.

Predictive Analysis (Classification)

- Split data into training and testing data
- Train the 4 models using LogReg, SVM, Decision Tree, KNN
- Hyperparameter tuning using Gridsearch method
- Compare the accuracy of each model
- Select best classification model

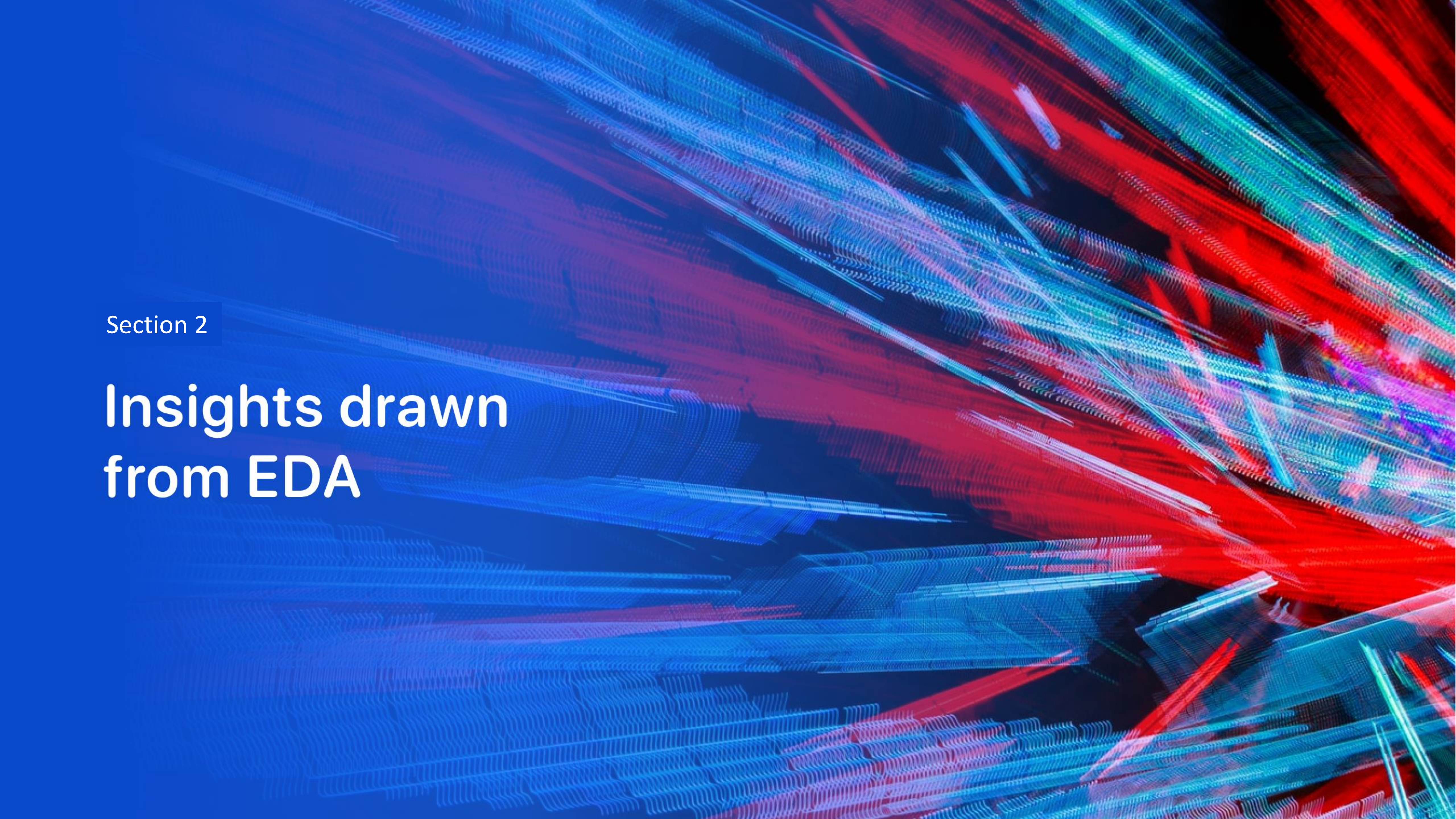


Results

- The initial stage is more likely to land successfully as the flight number increases. The cargo mass is also significant; it appears that the larger the payload, the less probable the first stage would return.

- No rockets have been launched for heavy payload mass (more than 10,000) from the VAFB-SLC launch site, and the success rate for this big payload mass is normally high. Polar, LEO, and ISS have a higher successful or positive landing rate with heavier payloads. However, for GTO, we can't tell the difference because both positive and negative landing rates (missed missions) are present.

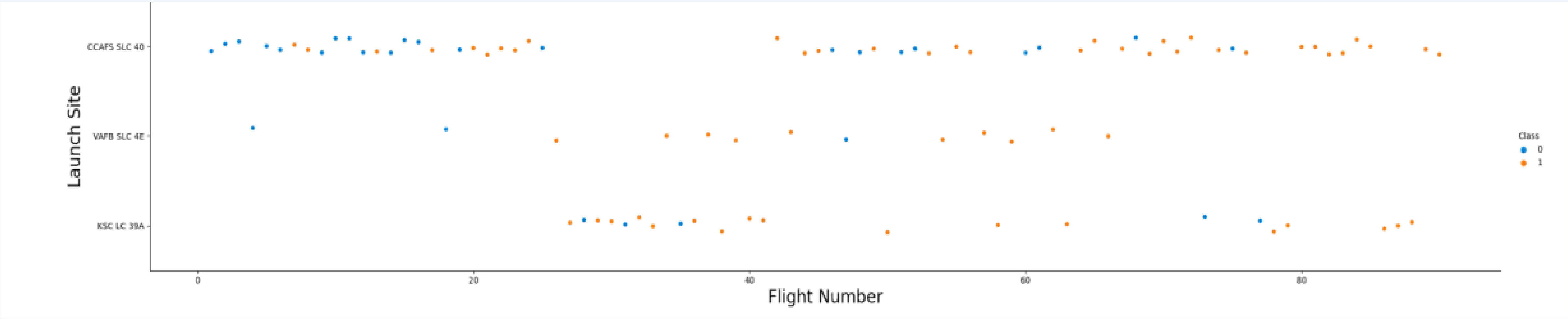
- The success rate has been steadily growing since 2013, and it is expected to continue through 2020.



Section 2

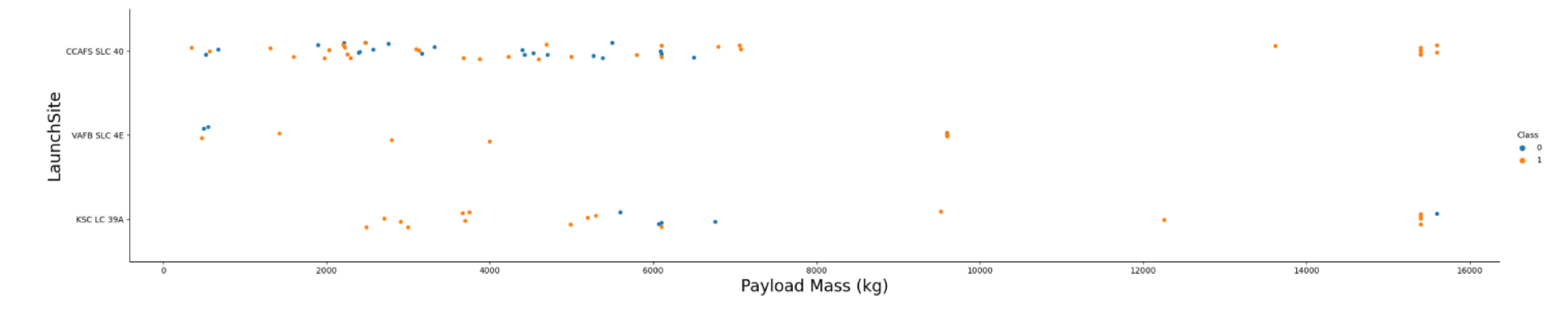
Insights drawn from EDA

Flight Number vs. Launch Site



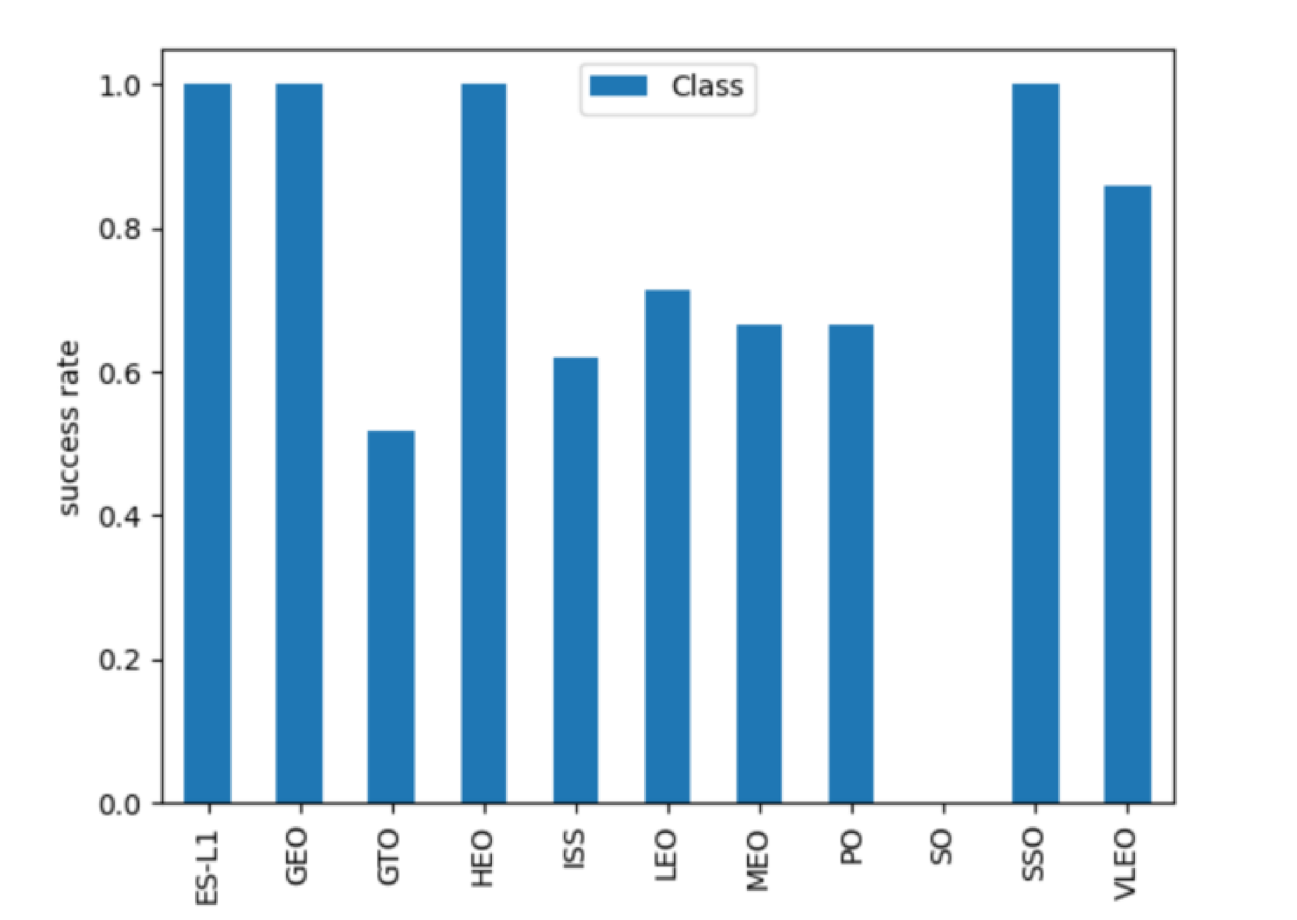
- Class 0 denotes fail, Class 1 denote success
- CCAFS SLC 40 half are fail and half are success
- VAFB SLC 4E has the fewest data points but has the highest rate of success
- KSC LC 39 A shows similar result of that of VAFB SLC 4E

Payload vs. Launch Site



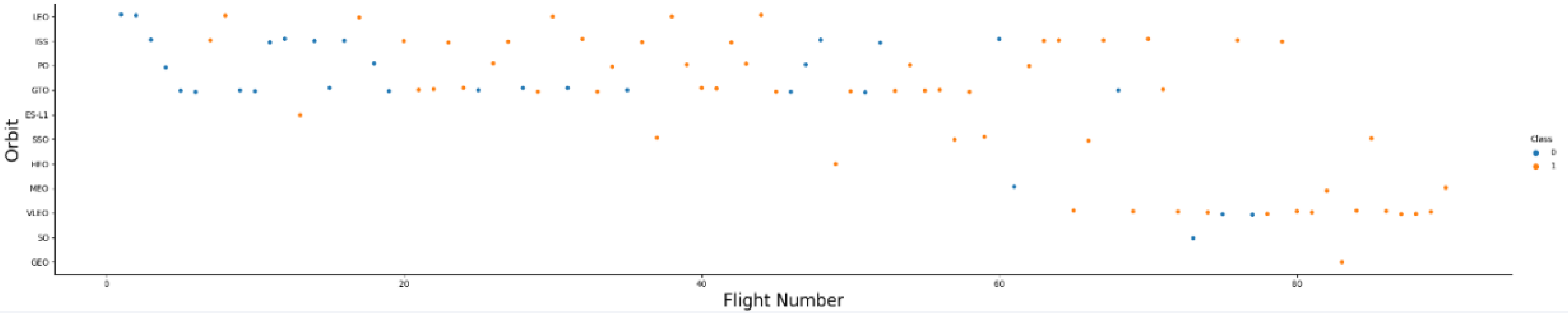
- Payload mass varies by launch location; for example, VAFB SLC 4E does not have a payload larger than 1000 KG.
- The data points are grouped below the 8000 KG line, implying that payloads larger than 8000 KG are not as common.

Success Rate vs. Orbit Type



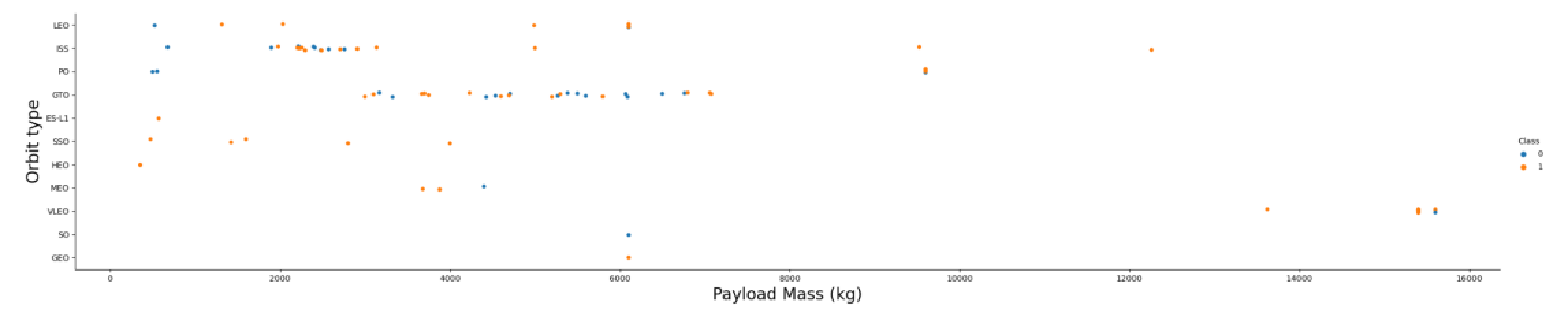
- The orbits ES-L1, GFO, HEO. SSO and VLEO had the best success rate
- While other are below 0.8

Flight Number vs. Orbit Type



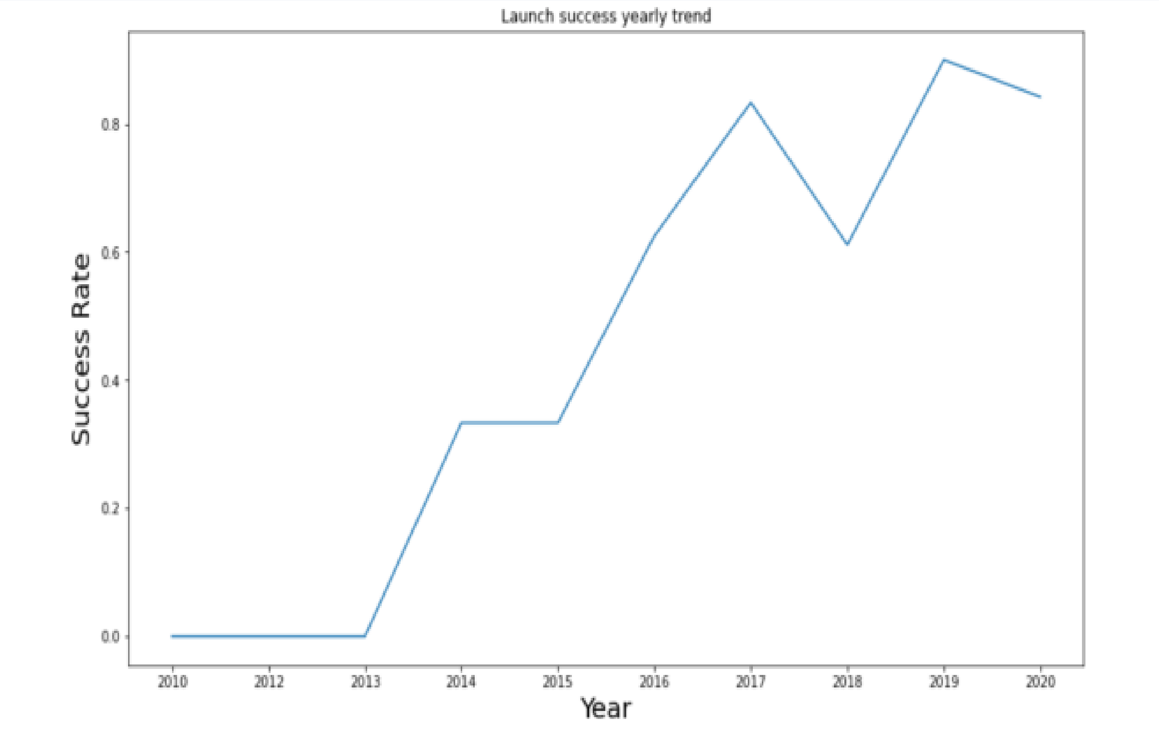
Flight number has no relationship with the success rate of each orbit

Payload vs. Orbit Type



- **The payload mass is larger than 8000KG, PO, LEO, and VLEO have better landing success percentages than other orbit types. This indicates higher the pay, higher the success rate.**
- **ES-L1, SSO, and HEO orbit types tend to have a higher success rate with lighter payload weights.**

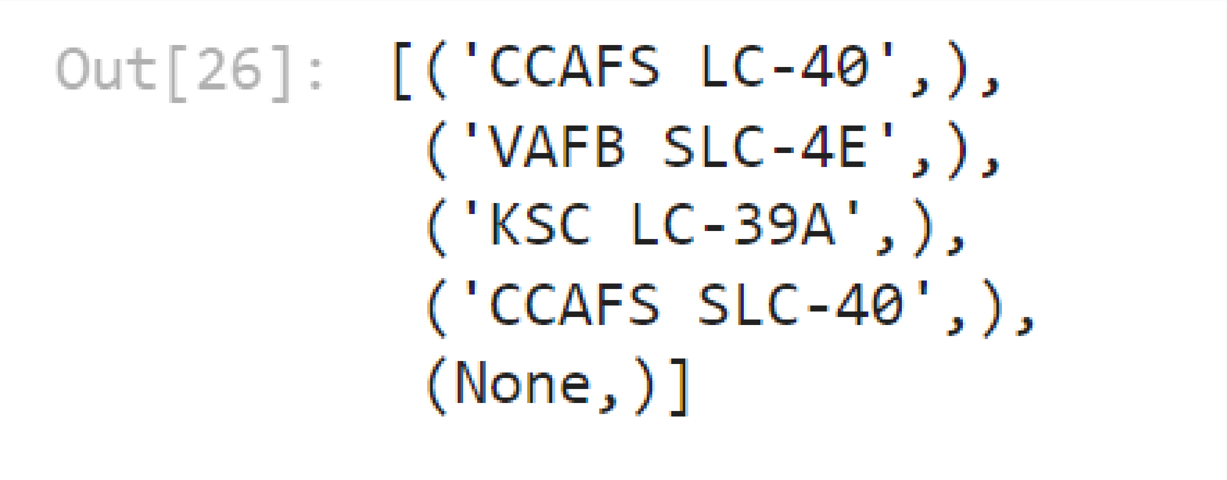
Launch Success Yearly Trend



- A drastic increase of success rate is shown when time is going forward.
- It can be predicted that the future can hold better success rate of launching.

All Launch Site Names

```
cur.execute('SELECT DISTINCT "Launch_Site" FROM
SPACEXTBL').fetchall()
```

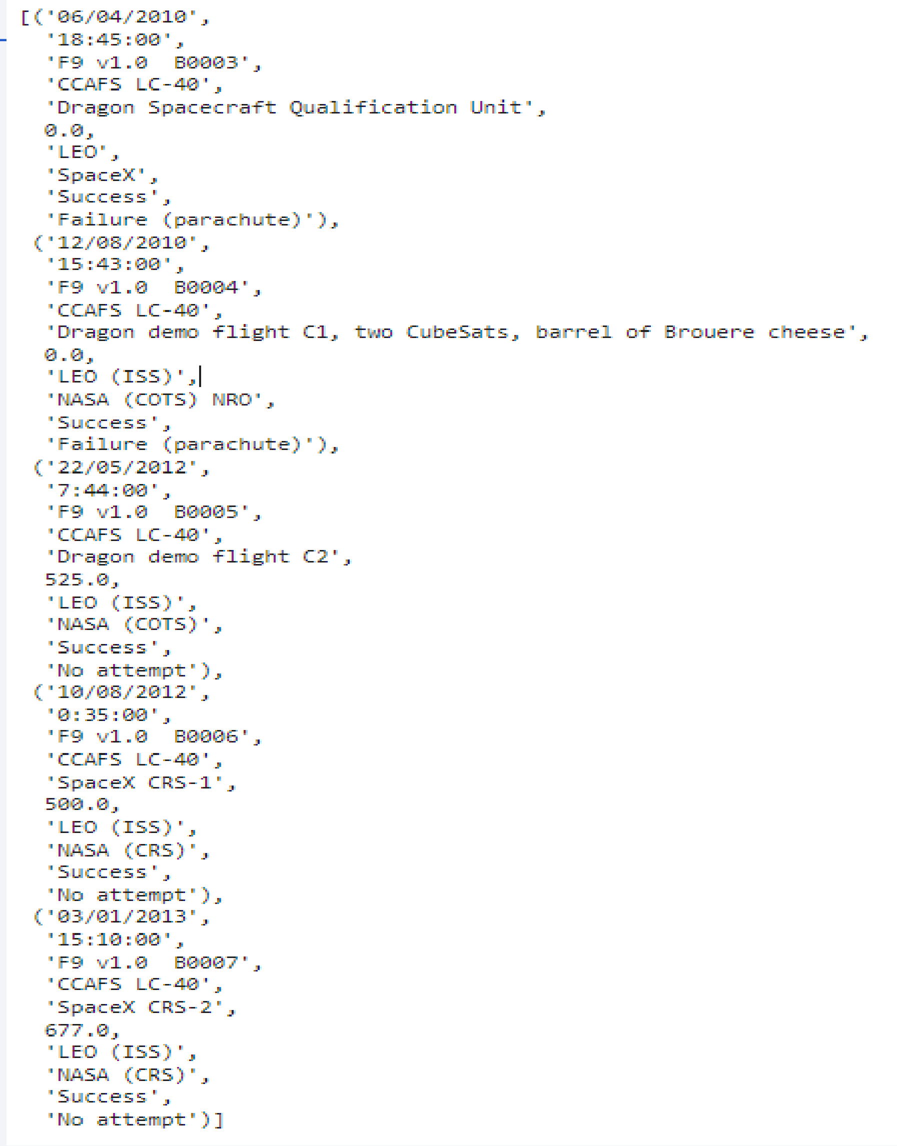


```
Out[26]: [('CCAFS LC-40',),
          ('VAFB SLC-4E',),
          ('KSC LC-39A',),
          ('CCAFS SLC-40',),
          (None,)]
```

- **Use Select distinct so it can avoid showing all the same items**
- **Target Launch site because we want the name of each one**

Launch Site Names Begin with 'CCA'

cur.execute('SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE "CCA%" LIMIT 5').fetchall()



```
[('06/04/2010',
  '18:45:00',
  'F9 v1.0  B0003',
  'CCAFS LC-40',
  'Dragon Spacecraft Qualification Unit',
  0.0,
  'LEO',
  'SpaceX',
  'Success',
  'Failure (parachute)'),
 ('12/08/2010',
  '15:43:00',
  'F9 v1.0  B0004',
  'CCAFS LC-40',
  'Dragon demo flight C1, two CubeSats, barrel of Brouere cheese',
  0.0,
  'LEO (ISS)',
  'NASA (COTS) NRO',
  'Success',
  'Failure (parachute)'),
 ('22/05/2012',
  '7:44:00',
  'F9 v1.0  B0005',
  'CCAFS LC-40',
  'Dragon demo flight C2',
  525.0,
  'LEO (ISS)',
  'NASA (COTS)',
  'Success',
  'No attempt'),
 ('10/08/2012',
  '0:35:00',
  'F9 v1.0  B0006',
  'CCAFS LC-40',
  'SpaceX CRS-1',
  500.0,
  'LEO (ISS)',
  'NASA (CRS)',
  'Success',
  'No attempt'),
 ('03/01/2013',
  '15:10:00',
  'F9 v1.0  B0007',
  'CCAFS LC-40',
  'SpaceX CRS-2',
  677.0,
  'LEO (ISS)',
  'NASA (CRS)',
  'Success',
  'No attempt')]
```

Total Payload Mass

```
cur.execute('SELECT SUM("PAYLOAD_MASS__KG_") FROM
SPACEXTBL WHERE "Customer" LIKE "NASA (CRS)"').fetchall()
```

```
[(45596.0,)]
```

- **Calculate the total by SUM function on Payload mass**
- **Filter using WHERE "Customer" LIKE "NASA (CRS)"**

Average Payload Mass by F9 v1.1

```
cur.execute('SELECT AVG("PAYLOAD_MASS__KG_") FROM
SPACEXTBL WHERE "Booster_Version" LIKE "F9
v1.1"').fetchall()
```

```
[(2928.4,)]
```

Use function AVG for calculating the average of payload mass from the table

Filter using WHERE "Booster_Version" LIKE "F9 v1.1

First Successful Ground Landing Date

```
cur.execute('SELECT MIN("Date") FROM SPACEXTBL WHERE
"Landing_Outcome" = "Success (ground pad)"').fetchall()
```

```
[('01/08/2018',)]
```

-Use MIN function of that to get the first successful date
- WHERE "Landing_Outcome" = "Success (ground pad)“ to get the first success on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
('SELECT "Booster_Version" FROM SPACEXTBL WHERE
"Landing_Outcome" = "Success (drone ship)" AND
"Payload_Mass__kg_" > 4000 AND
"Payload_Mass__kg_“>6000
```

```
[('F9 FT B1022',), ('F9 FT B1026',), ('F9 FT  B1021.2',), ('F9 FT  B1031.2',)]
```

- Booster version is for finding the name of drone ship landing
- Given WHERE "Landing_Outcome" = "Success (drone ship)“ for the success in drone ship
- Plus, Payload_Mass__kg_" > 4000 AND "Payload_Mass__kg_“>6000 for limiting the payload mass between 4000 and 6000
- Result show 4 which is listed as above

Total Number of Successful and Failure Mission Outcomes

```
cur.execute('SELECT "Mission_Outcome", COUNT(*) as "Total" FROM SPACEXTBL
GROUP BY "Mission_Outcome"').fetchall()
```

```
[(None, 898),
 ('Failure (in flight)', 1),
 ('Success', 98),
 ('Success ', 1),
 ('Success (payload status unclear)', 1)]
```

- Find the outcome number of each mission outcome
- The result shows the number of success distinctly from outcome column element

Boosters Carried Maximum Payload

```
cur.execute('SELECT "Booster_Version" FROM SPACEXTBL
WHERE "Payload_Mass__kg_" = (SELECT MAX
("Payload_Mass__kg_") FROM SPACEXTBL)').fetchall()
```

```
[('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 ',)]
```

- **Select booster version as targe when the payload has the max value**
- **The result shows those booster version that has highest values in Payload mass kg column**

2015 Launch Records

```
cur.execute('''
    SELECT
        CASE SUBSTR("Date", 4, 2)
            WHEN '01' THEN 'January'
            WHEN '02' THEN 'February'
            WHEN '03' THEN 'March'
            WHEN '04' THEN 'April'
            WHEN '05' THEN 'May'
            WHEN '06' THEN 'June'
            WHEN '07' THEN 'July'
            WHEN '08' THEN 'August'
            WHEN '09' THEN 'September'
            WHEN '10' THEN 'October'
            WHEN '11' THEN 'November'
            WHEN '12' THEN 'December'
        END AS MonthName,
        "Landing_Outcome",
        "Booster_Version",
        "Launch_Site"
    FROM SPACEXTBL
    WHERE SUBSTR("Date", 7, 4) = '2015'
        AND "Landing_Outcome" LIKE 'Failure (drone ship)'
''')

# Fetch all the results
results = cur.fetchall()
results
```

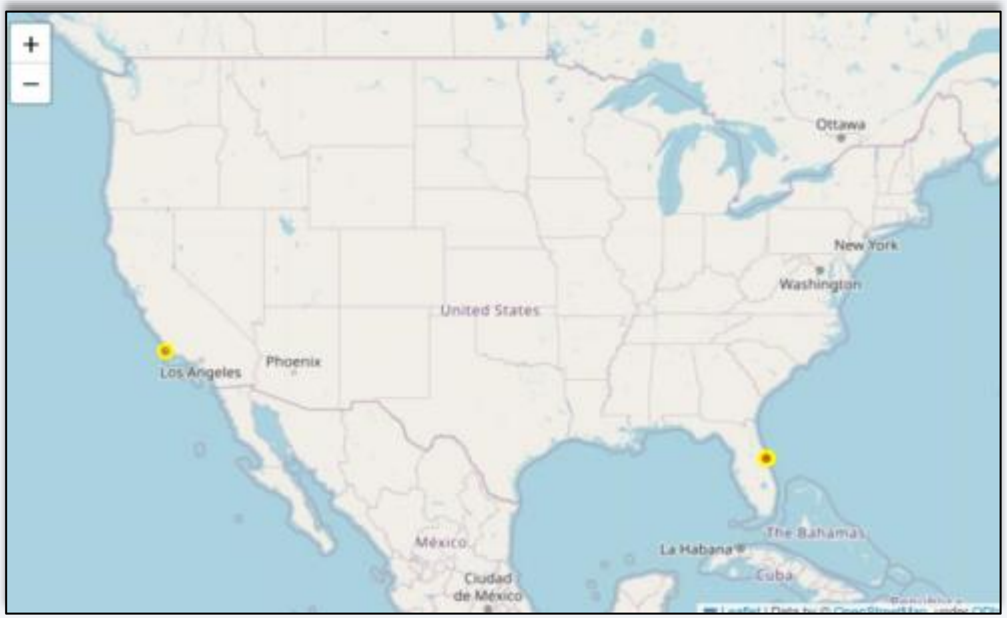
```
[('October', 'Failure (drone ship)', 'F9 v1.1 B1012', 'CCAFS LC-40'),
 ('April', 'Failure (drone ship)', 'F9 v1.1 B1015', 'CCAFS LC-40')]
```

- **Use When to replace number from data format to string of month**
- SUBSTR("Date", 7, 4) = '2015’ represent as year =2015
- AND "Landing_Outcome" LIKE 'Failure for a failure record in that year

Section 3

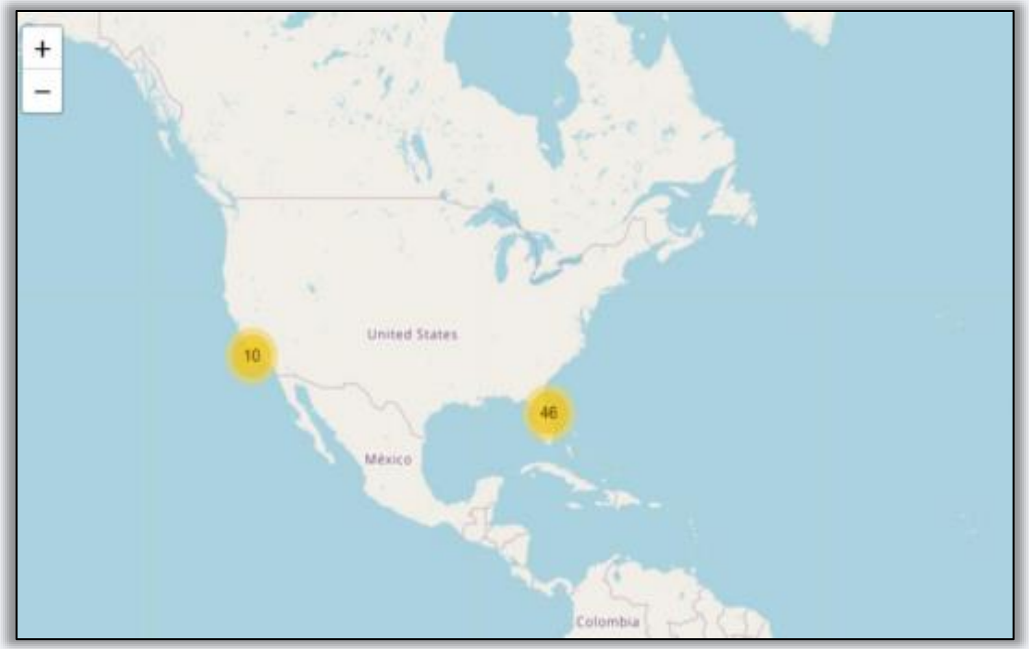
Launch Sites Proximities Analysis

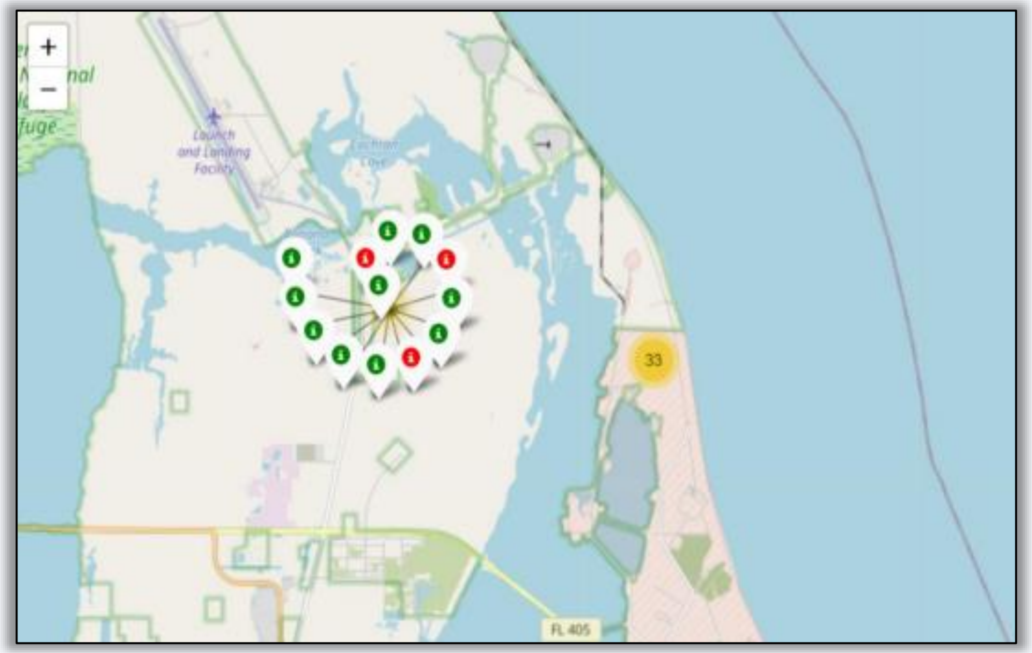
Location of SpaceX Launch Sites



- Notably, all sites are placed along the coast.
- the locations of the four SpaceX launch sites. When the user clicks on the site marker, a pop-up label displaying the site name appears, as illustrated in the rightmost image.
- Only one launch location is in California, while the other three are in Florida.

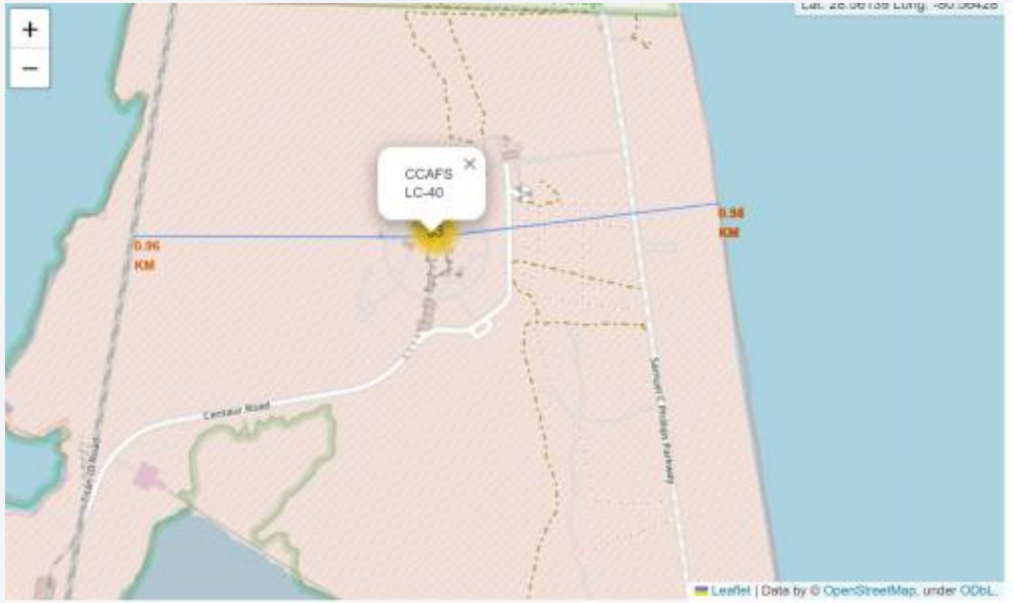
Successful and Fail Launches by launch Sites





- Launch location CCAFS LC-40 has the lowest success rate around 27%, while launch site KSC LC-39A has the highest success rate around 77%.

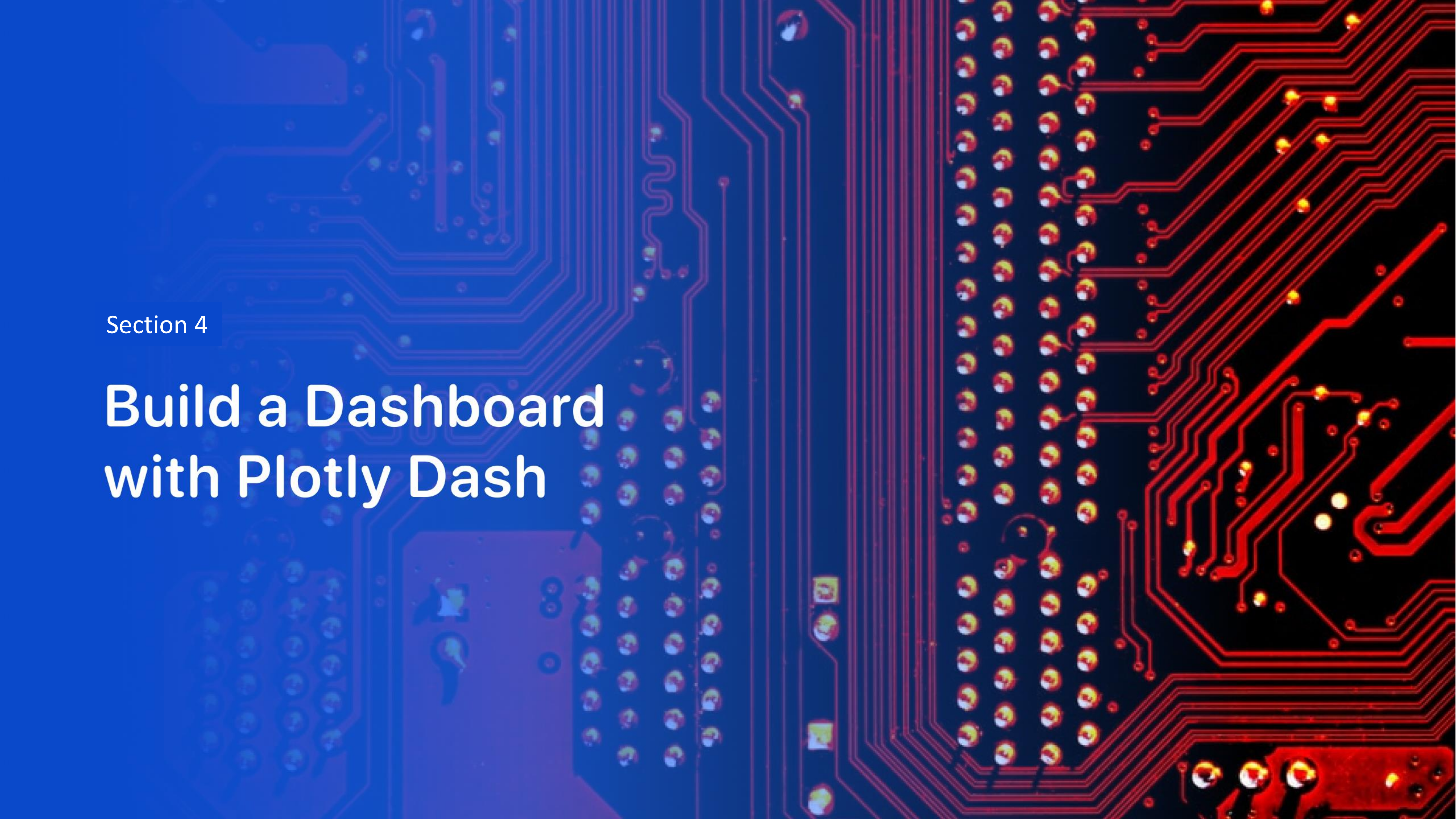
<Folium Map Screenshot 3>


+
−
CCAFS
LC-40
0.96
KM
0.98
KM
Centaur Road
Samuel C Phillips Parkway
Leaflet | Data by © OpenStreetMap, under ODbL.

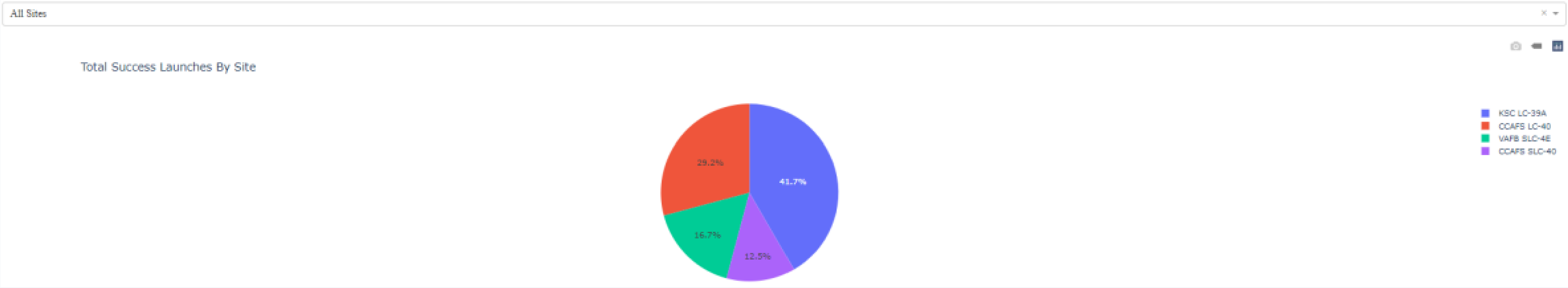
- The CCAFS LC-40 launch location is 0.98 kilometers from the nearest beach and 0.96 kmfrom the nearest railway.
- The closeness of the location to trains and coasts shows that these are major variables considered by SpaceX when selecting launch sites.

Section 4

Build a Dashboard with Plotly Dash

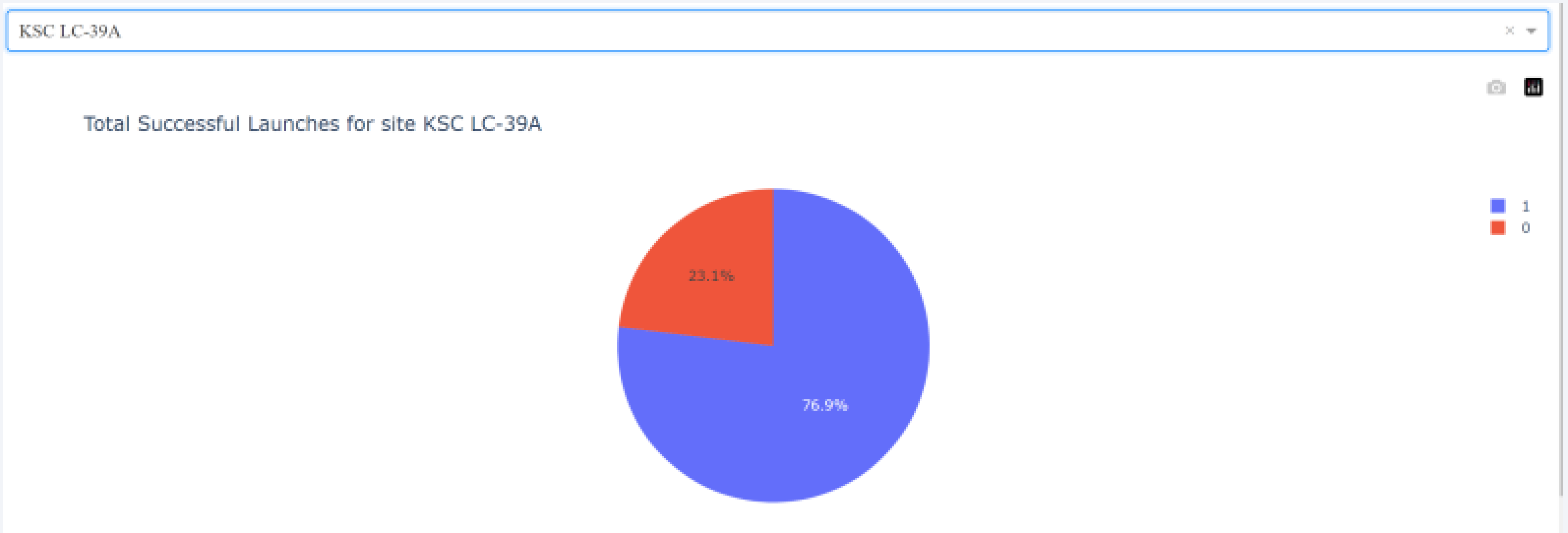


Success rate for all sites



- Pie Chart show percentages of success rates of each site
- The launch location CCAFS SLC-40 had the fewest successful launches.
- In comparison to the other launch sites, KSC LC-39A had the most successful launches.

KSC LC-39A

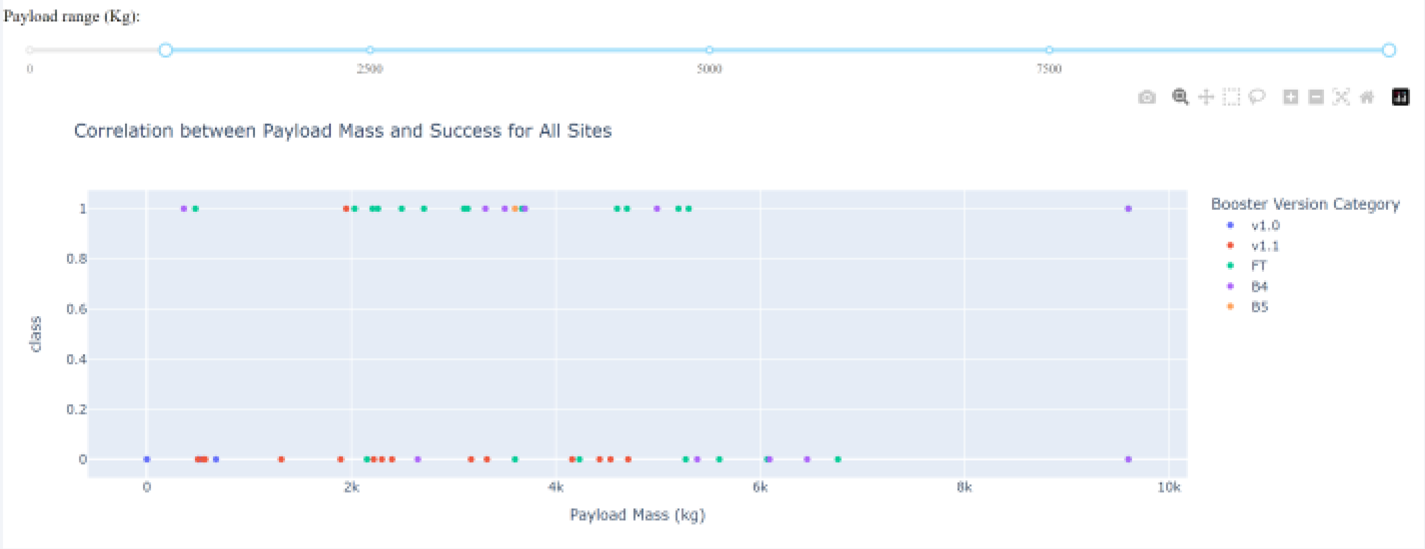


KSC LC-39A

Total Successful Launches for site KSC LC-39A

- Success vs Fail ratio of KSC LC-39 after filtering using dropdown

Relationship between Payload and Launch Outcome



According to this graph, booster version FT has the best success rate when compared to other booster versions.
Version 1.1 of the booster appears to have the lowest success rate. The majority of successful launches occur between 2000KG and 4000KG.

Section 5

Predictive Analysis (Classification)

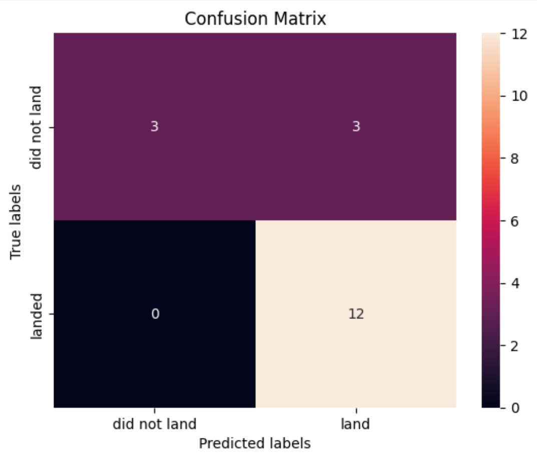
Classification Accuracy

MODEL	Train Accuracy	Test Accuracy
Log Reg	0.833	0.916
SVM	0.833	0.847
DECISION TREE	0.833	0.861
KNN	0.722	1.0

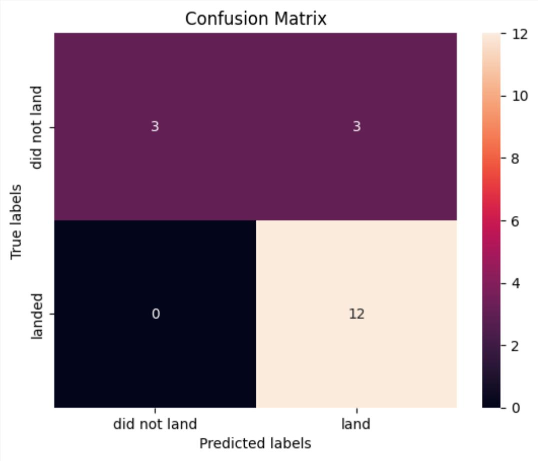
- Best model is log reg because it has the highest test and train accuracy.
- KNN shows test accuracy is 1 while train accuracy is 0.7222, indicate overfitting behaviour.

Confusion Matrix

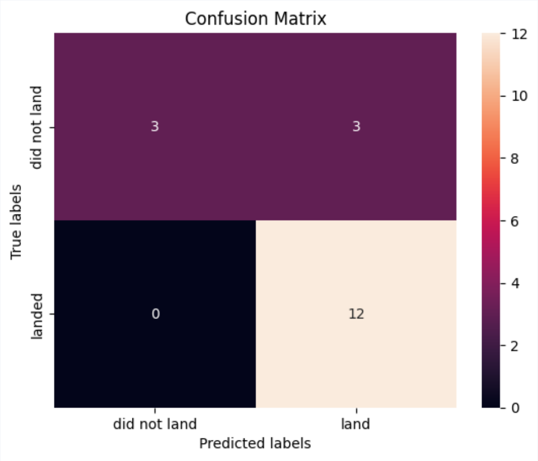
LOG REG



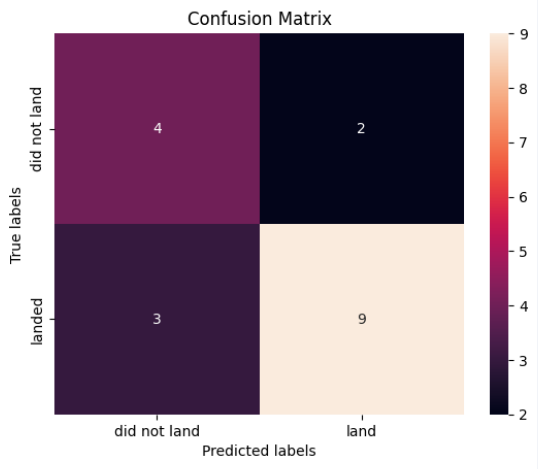
SVM



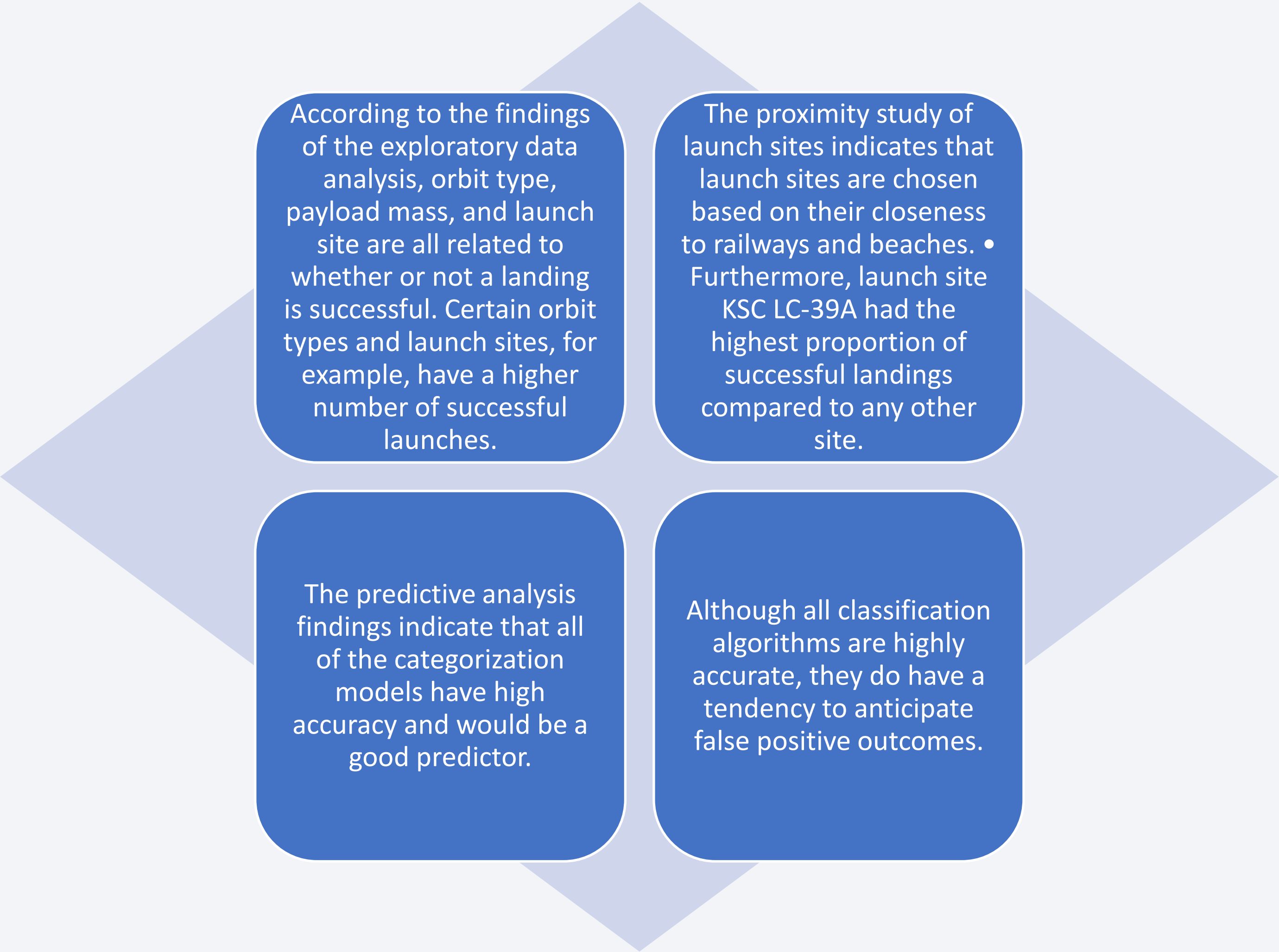
TREE



KNN



Conclusions



According to the findings of the exploratory data analysis, orbit type, payload mass, and launch site are all related to whether or not a landing is successful. Certain orbit types and launch sites, for example, have a higher number of successful launches.

The proximity study of launch sites indicates that launch sites are chosen based on their closeness to railways and beaches. • Furthermore, launch site KSC LC-39A had the highest proportion of successful landings compared to any other site.

The predictive analysis findings indicate that all of the categorization models have high accuracy and would be a good predictor.

Although all classification algorithms are highly accurate, they do have a tendency to anticipate false positive outcomes.

Appendix

GITHUB LINK:
https://github.com/keeeen5678/IBM-DATA-SCIENCE-CAPSTONE.git

Thank you!