Winning the Race to Space with Applied Data Science

Varun Bansal

21-Aug-2022

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Overall Findings
- Conclusion
- Thank You

EXECUTIVE SUMMARY

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

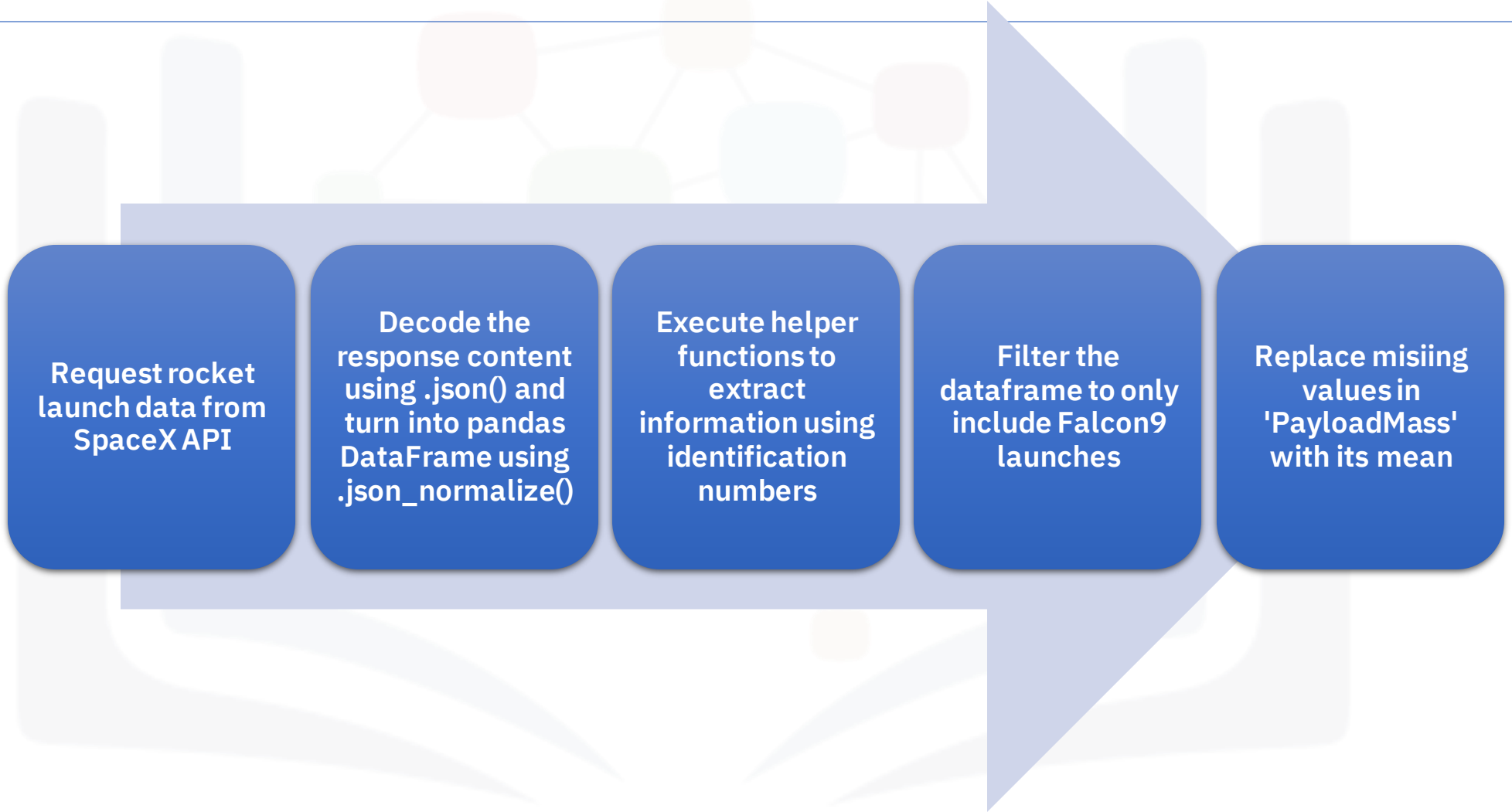
INTRODUCTION

- Project Background and context
  - SpaceX advertises Falcon9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each.
  - First stage is much larger than the second stage and does most of the work; it is expensive too.
  - SpaceX is able to recover and reuse the first stage.
- Problem Statement
  - Determine the price of each launch by gathering information about SpaceX
  - Train a machine learning model to determine if SpaceX will reuse the first stage

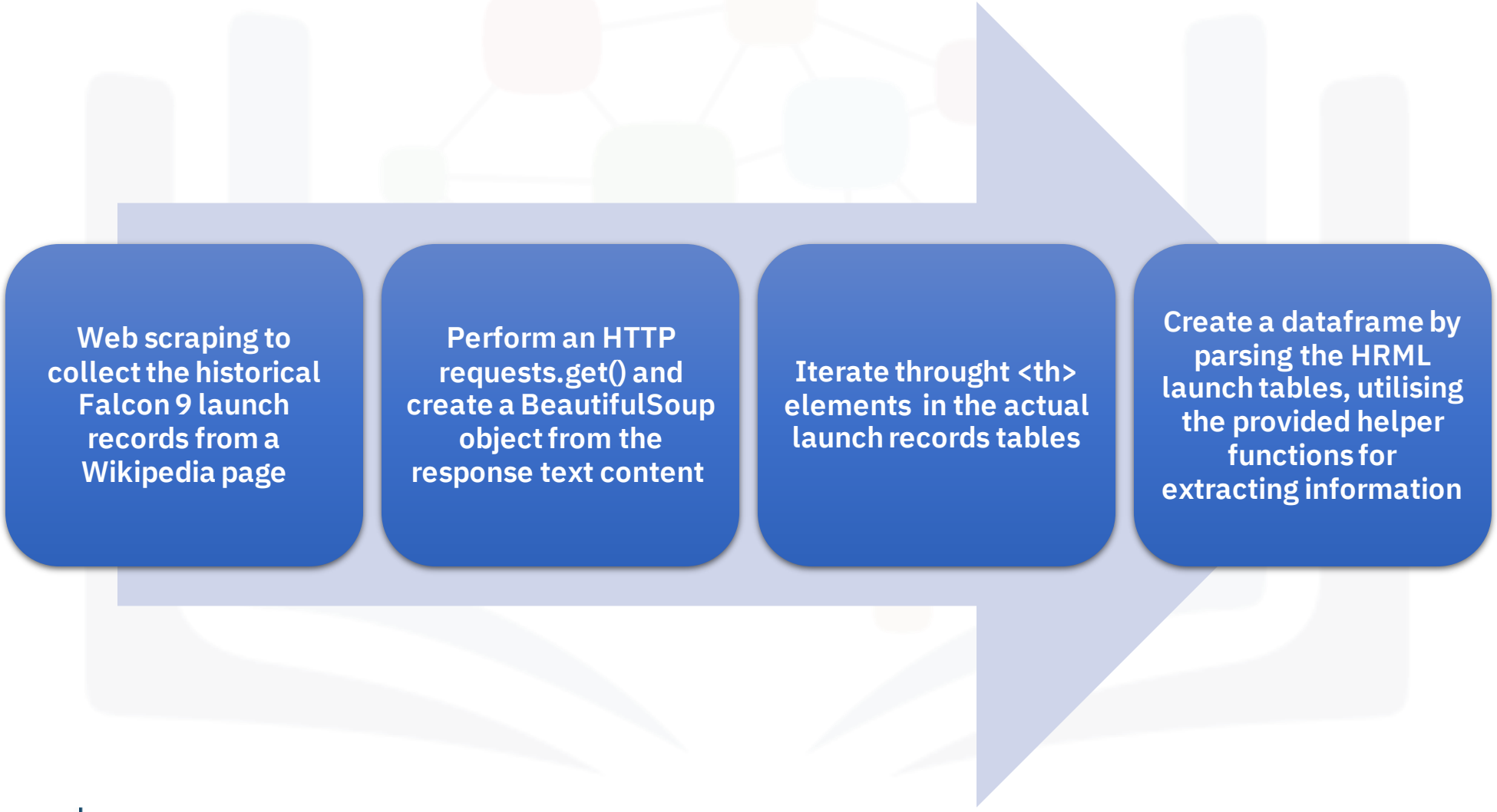
METHODOLOGY

- Data collection
  - REST APIs
  - Web scraping
- Data Wrangling
- Exploratory Data Analysis
  - Using SQL
  - Using pandas and matplotlib
- Interactive visual analytics with Folium
- Interactive dashboard with plotly dash
- Machine learning predictive analysis

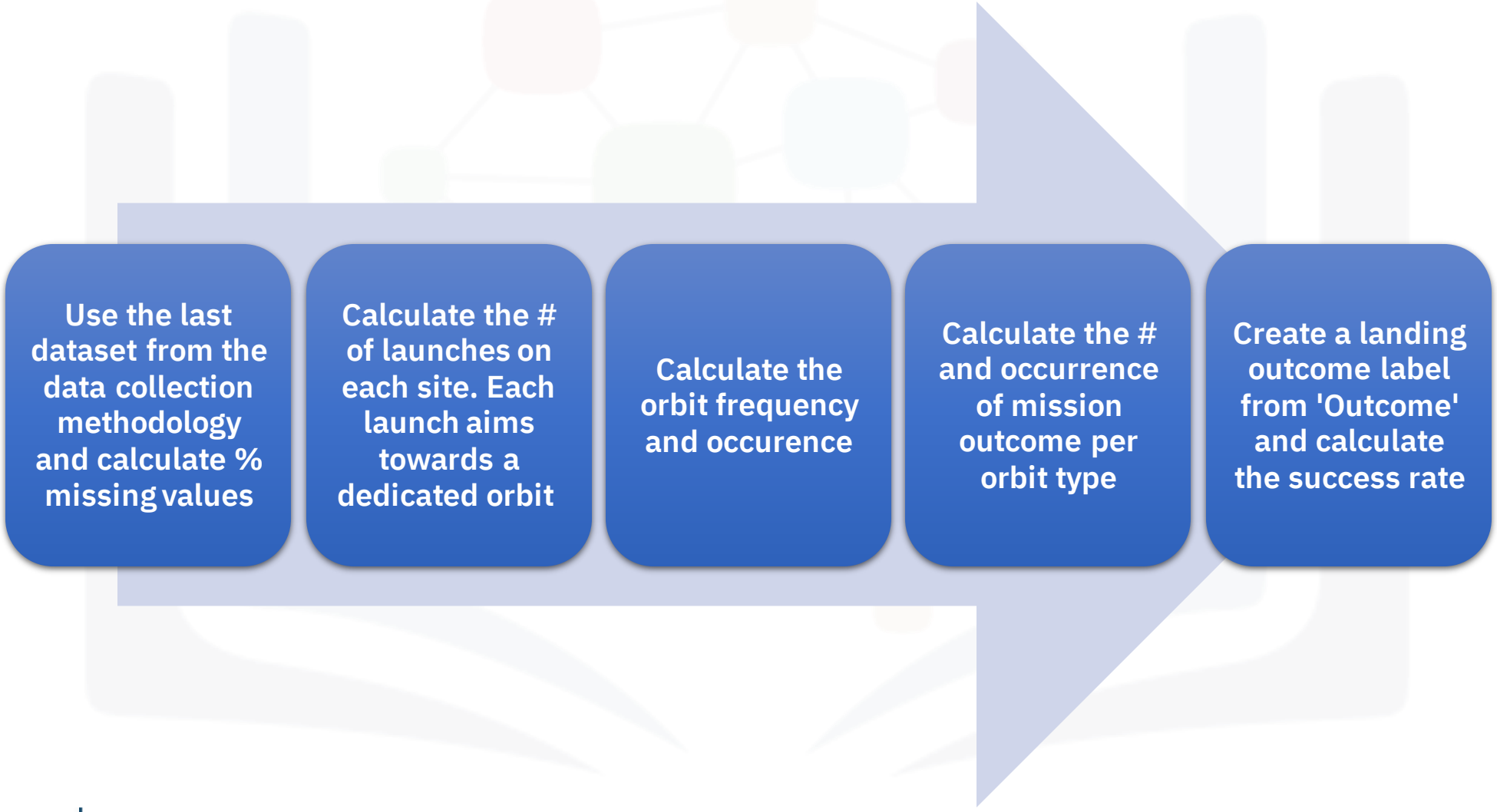
DATA COLLECTION WITH REST APIs



DATA COLLECTION WITH WEB SCRAPING



DATA WRANGLING



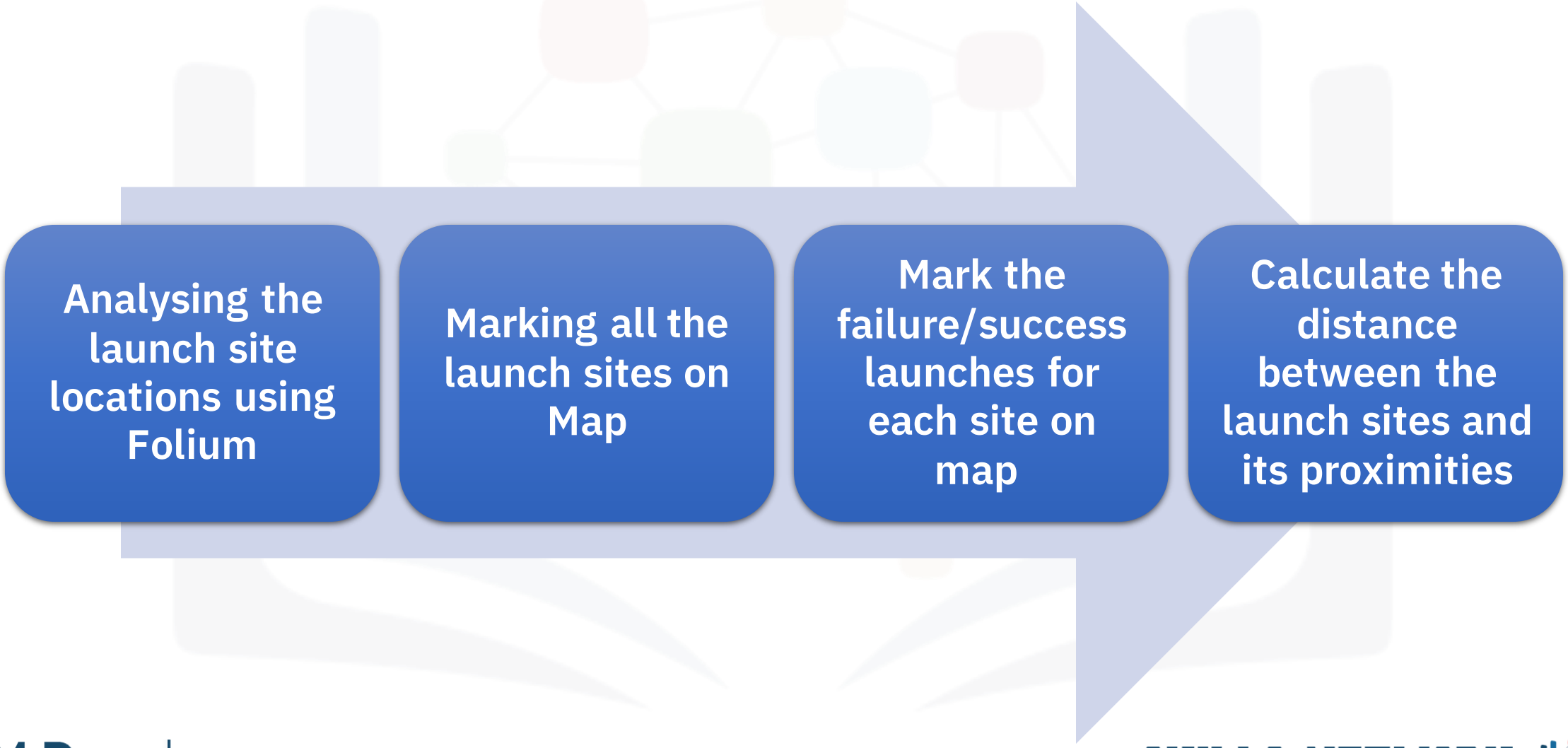
EXPLORATORY DATA ANALYSIS USING SQL

1. The SpaceX dataset was loaded into PL SQL database using the Juppyter notebook.
2. The following 10 SQL queries were executed for EDA:

- Unique launch sites in the space mission
- 5 records where launch sites begin with 'CCA'
- Booster names with payload mass between 4000-6000kg and have success in drone ship
- Names of booster versions which have carried the maximum payload mass
- Total payload mass carried by boosters launched by NASA(CRS)
- Average payload mass carried by booster version F9 v1.1
- Total successful and failure mission outcomes
- Records displaying month names, failure landing outcomes in drone ship, booster version and launch sites in the year 2015
- Date: when the first successful landing outcome in ground pad was achieved
- Rank the count of successful landing outcomes between 4th June 2010 and 20th March 2017 in descending order

EXPLORATORY DATA ANALYSIS USING PANDAS AND MATPLOTLIB

INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

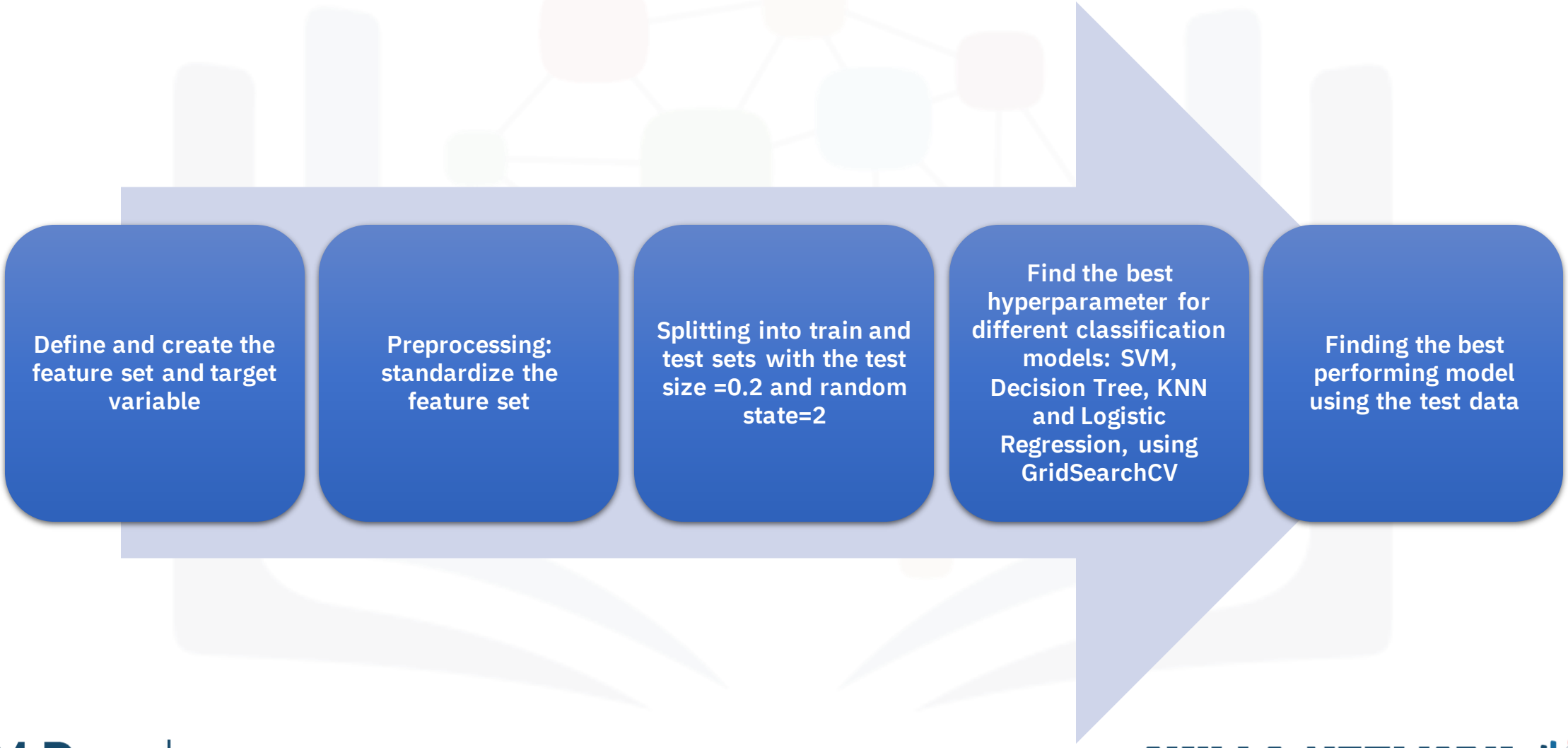


INTERACTIVE DASHBOARD WITH PLOTLY DASH

The following methods were adopted to build an interactive dashboard with Plotly dash:

- Pie charts showing the launch distribution for the different launch sites
- Scatter charts showing the relationships for the Outcome and Payload mass(kg) for different Booster versions

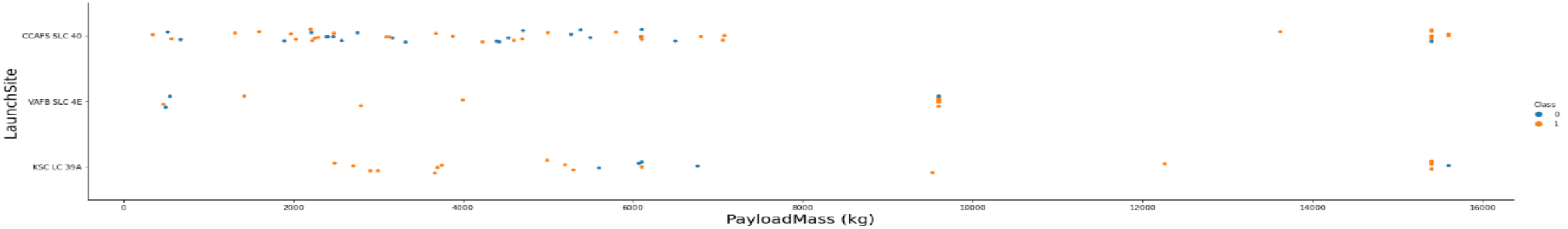
MACHINE LEARNING PREDICTIVE ANALYSIS



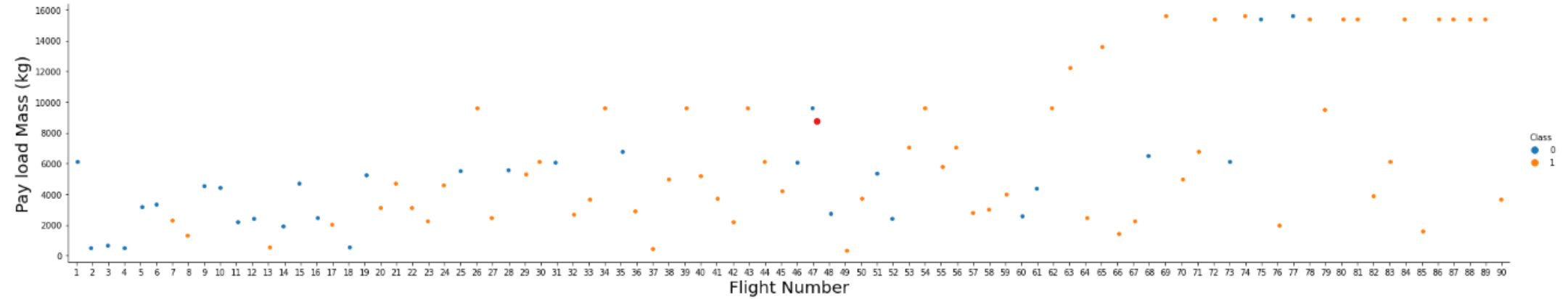
RESULTS

- Exploratory Data Analysis
  - Using SQL
  - Using pandas and matplotlib
- Launch site location analysis
- Dashboard with plotly dash
- ML models predictive analysis

EDA using pandas and matplotlib(1/3)

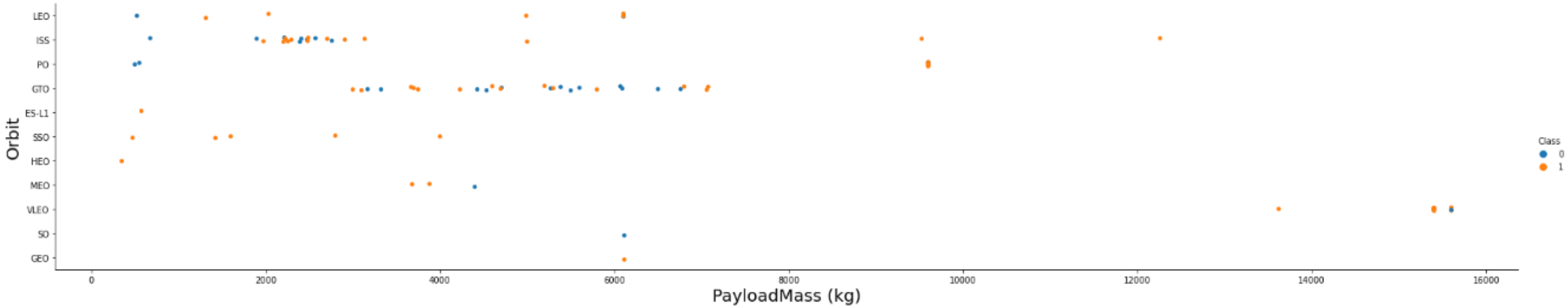


For the launch site VAFB-SLC no rockets launched for heavy payload mass(>10,000 kg)



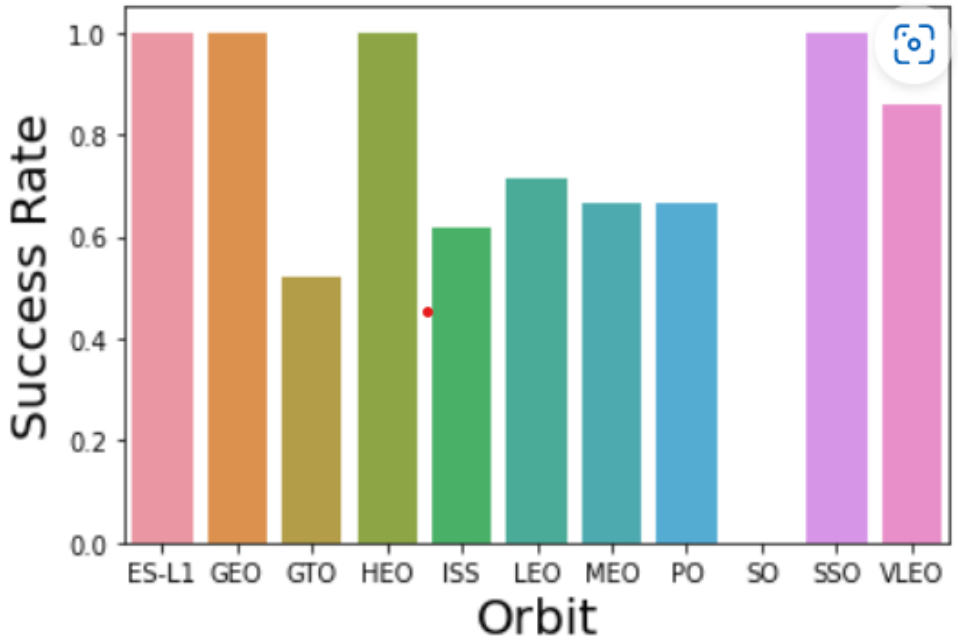
As the flight number increases , the first stage is more likely to land successfully

EDA using pandas and matplotlib(2/3)

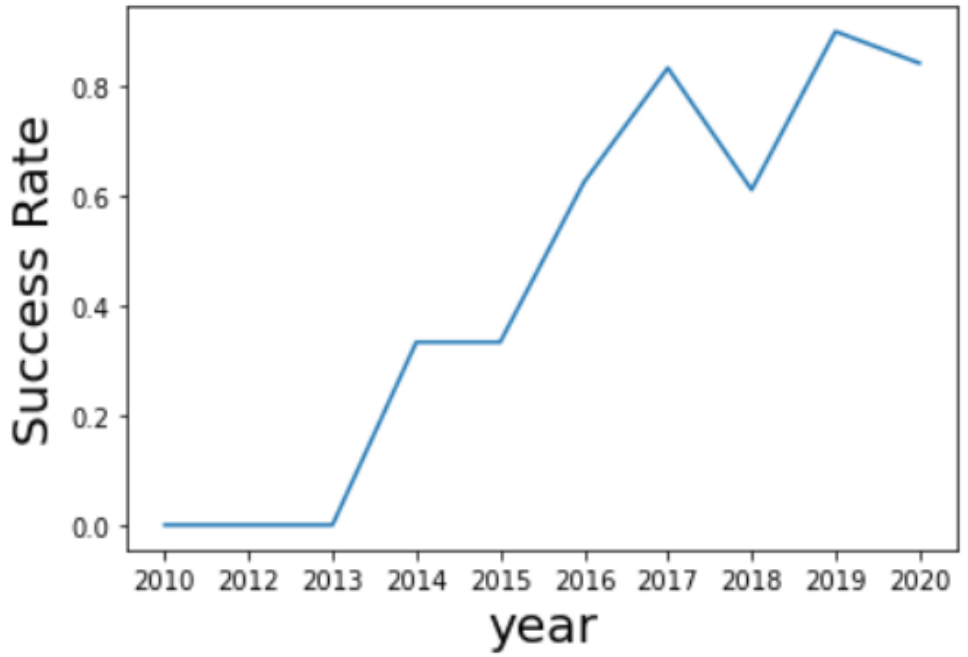


The successful landing rates with heavy payloads are more positive for Polar, LEO and ISS orbit types. For GTO, we have both positive and negative landing rates

EDA using pandas and matplotlib(3/3)



SSO, GEO, HEO orbit types have 100% success rate



- Yearly success rate increased from 2013 till 2020
- Success rate for the first 3 years was 0 (2010-2013)

EDA using SQL (1/5)

Task 1

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

```
In [10]: sql Select distinct(Launch_Site) from SPACEXTBL
```

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

Out[10]:

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

Task 2

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

```
In [13]: sql Select * from SPACEXTBL where Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677

EDA using SQL (2/5)

Task 3

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

In [14]:
```
sql Select SUM(CAST([PAYLOAD_MASS__KG_] AS INT)) AS NASA_CRS_TOTAL FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

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

Out[14]:

NASA_CRS_TOTAL
45596

Task 4

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

In [16]:
```
sql Select AVG(CAST([PAYLOAD_MASS__KG_] AS INT)) AS F9_V1_AVG from SPACEXTBL WHERE Booster_Version='F9 v1.1'
```

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

Out[16]:

F9_V1_AVG
2928.4

EDA using SQL (3/5)

Task 5

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

Hint:Use min function

In [17]:
```
sql Select min([Date]) from SPACEXTBL WHERE [Landing _Outcome]='Success (ground pad)'
```

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

Out[17]:

min([Date])
01-05-2017

Task 6

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

In [18]:
```
sql Select distinct(Booster_Version) from SPACEXTBL where [Landing _Outcome]='Success (drone ship)' AND (CAST([PAYLOAD_MASS__KG_
```

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

Out[18]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

EDA using SQL (4/5)

Task 7

List the total number of successful and failure mission outcomes

In [19]:
```
sql SELECT Mission_Outcome,COUNT(BOOSTER_VERSION) AS OUTCOME_COUNT FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

Out[19]:

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

Task 8

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

In [24]:
```
sql Select distinct Booster_Version from SPACEXTBL where CAST([PAYLOAD_MASS__KG_] AS IN
```

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

Out[24]:

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

EDA using SQL (5/5)

Task 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.

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 [27]:
```
sql Select substr([Date],4,2) as [Month], Booster_Version, Launch_Site from SPACEXTBL WHERE substr([Date],7,4)='2015' AND [Landir
```

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

Out[27]:

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

In [28]:
```
sql Select [Landing _Outcome],count(booster_version) as success_count from SPACEXTBL where [Landing _Outcome] like '%success%' gr
```

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

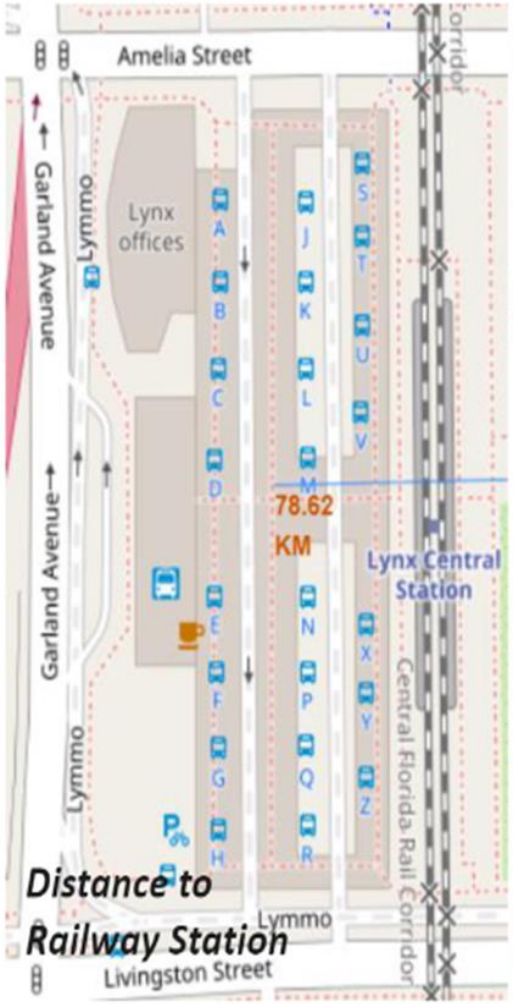
Out[28]:

Landing _Outcome	success_count
Success (ground pad)	9
Success (drone ship)	14
Success	38

Launch site location analysis(1/3)

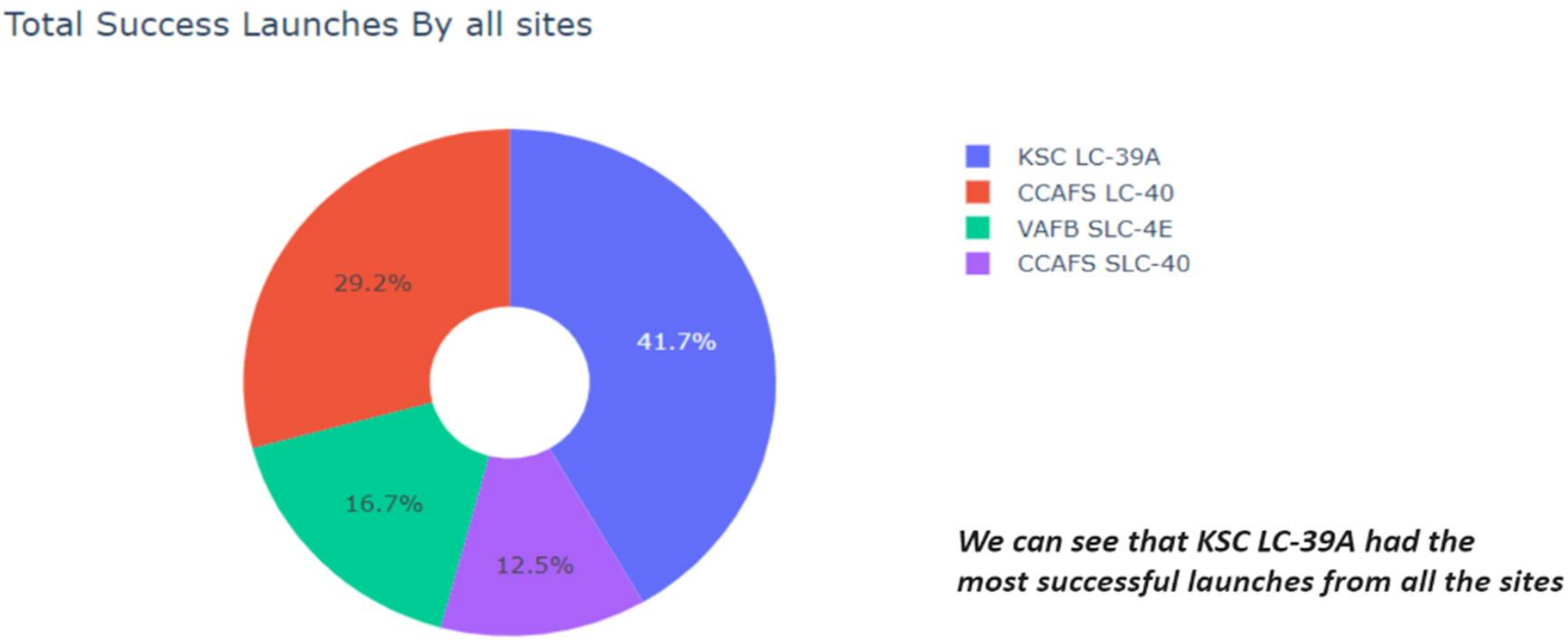
Launch site location analysis(2/3)

Launch site location analysis(3/3)



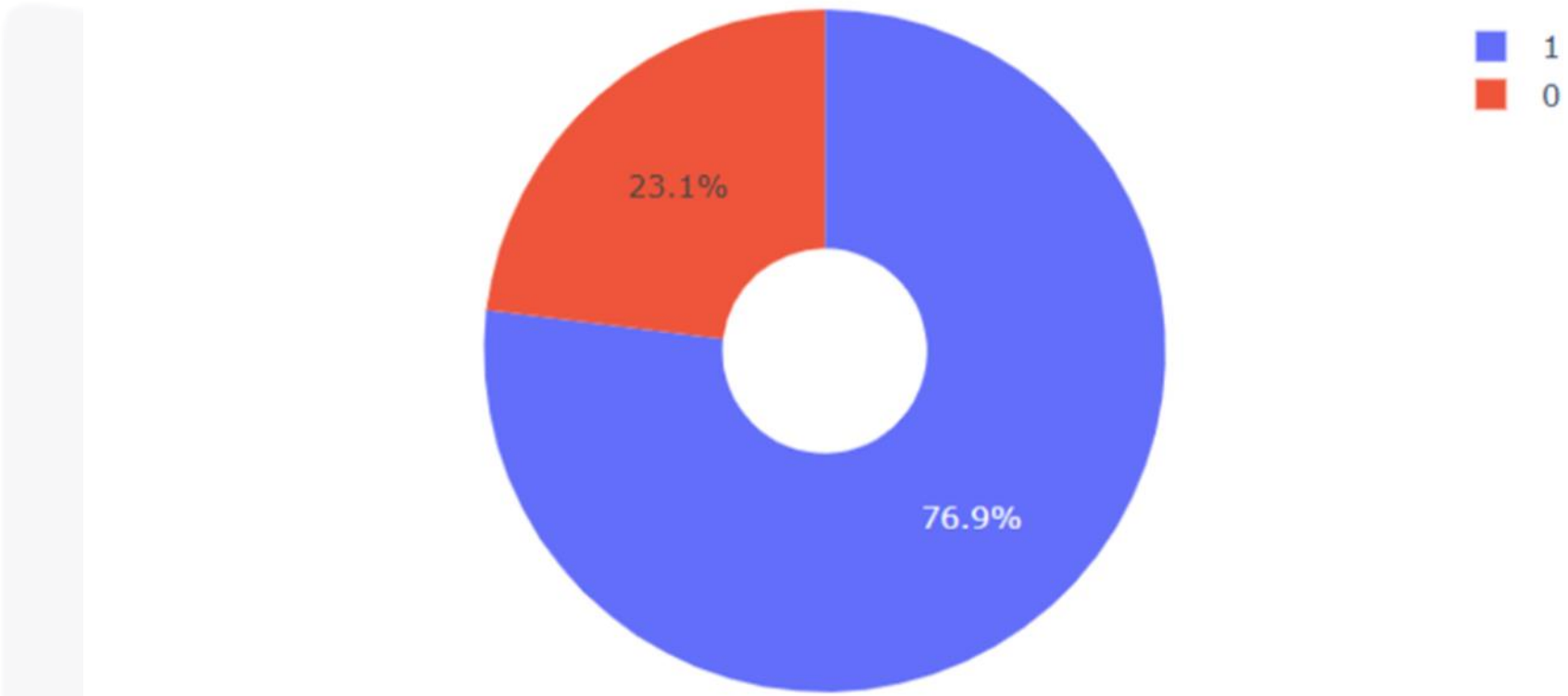
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Dashboard with Plotly dash(1/3)



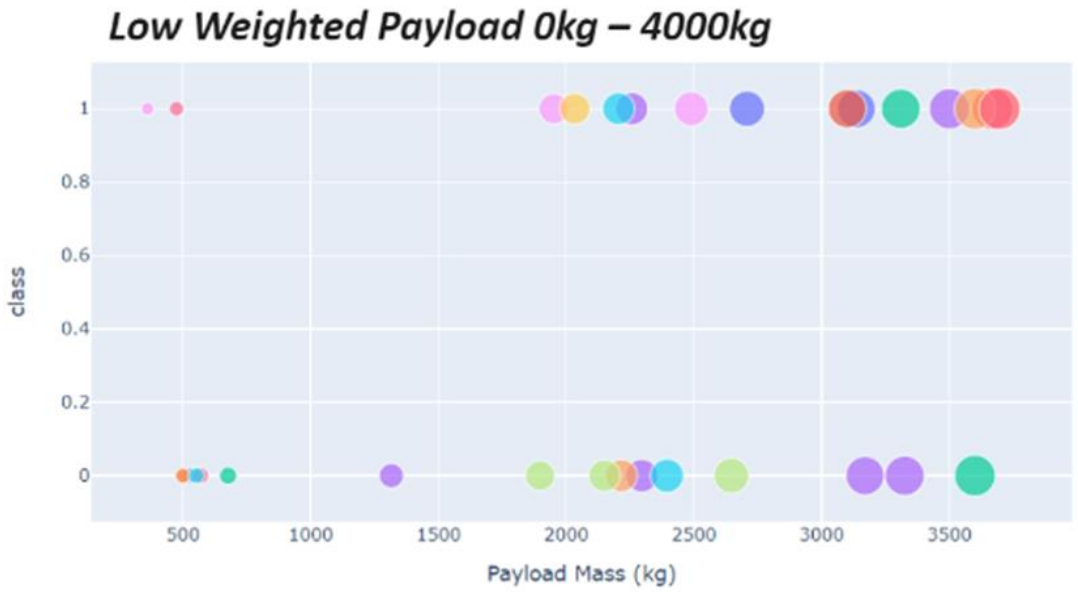
We can see that KSC LC-39A had the most successful launches from all the sites

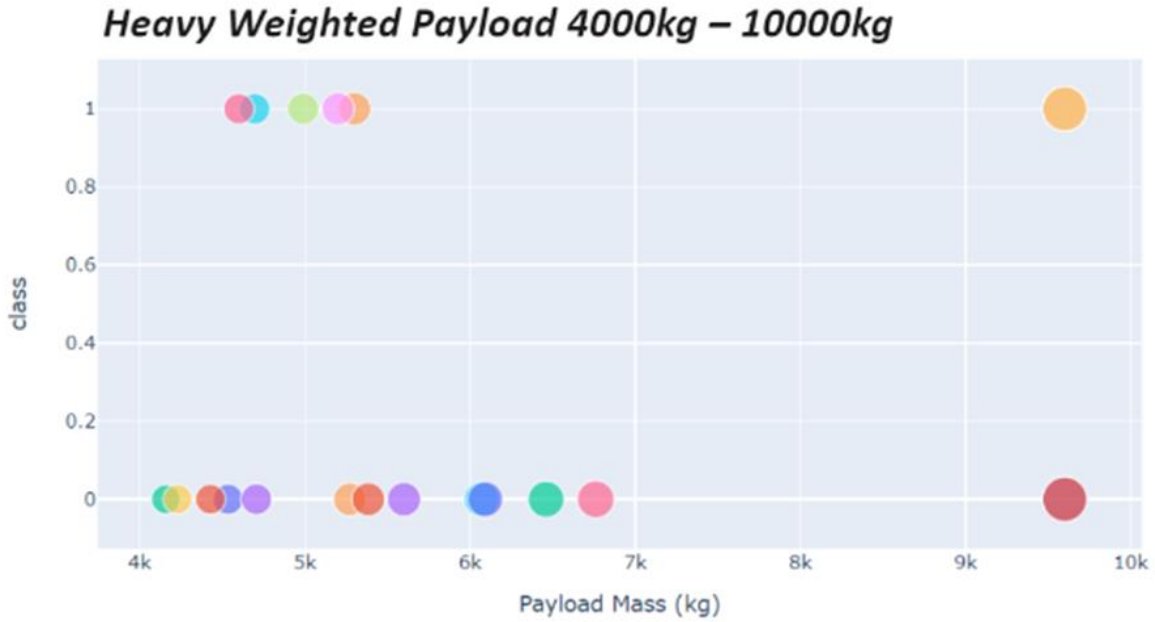
Dashboard with Plotly dash(1/3)



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Dashboard with Plotly dash(1/3)

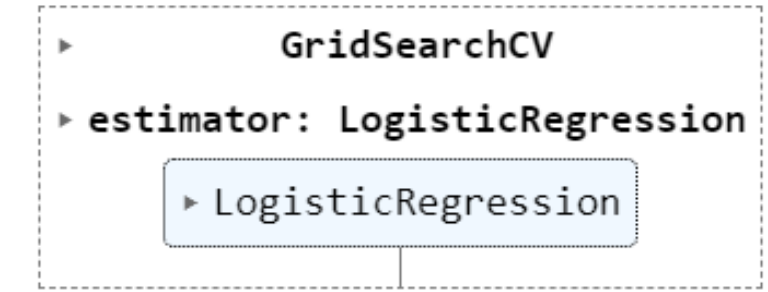




We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

ML models predictive analysis(1/4)

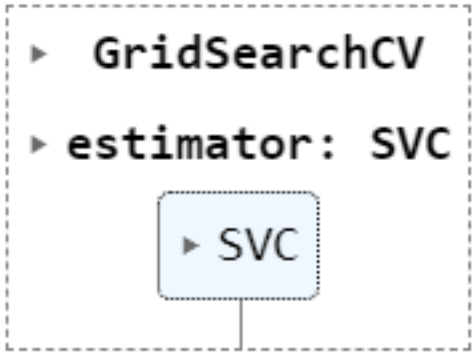
Out[15]:



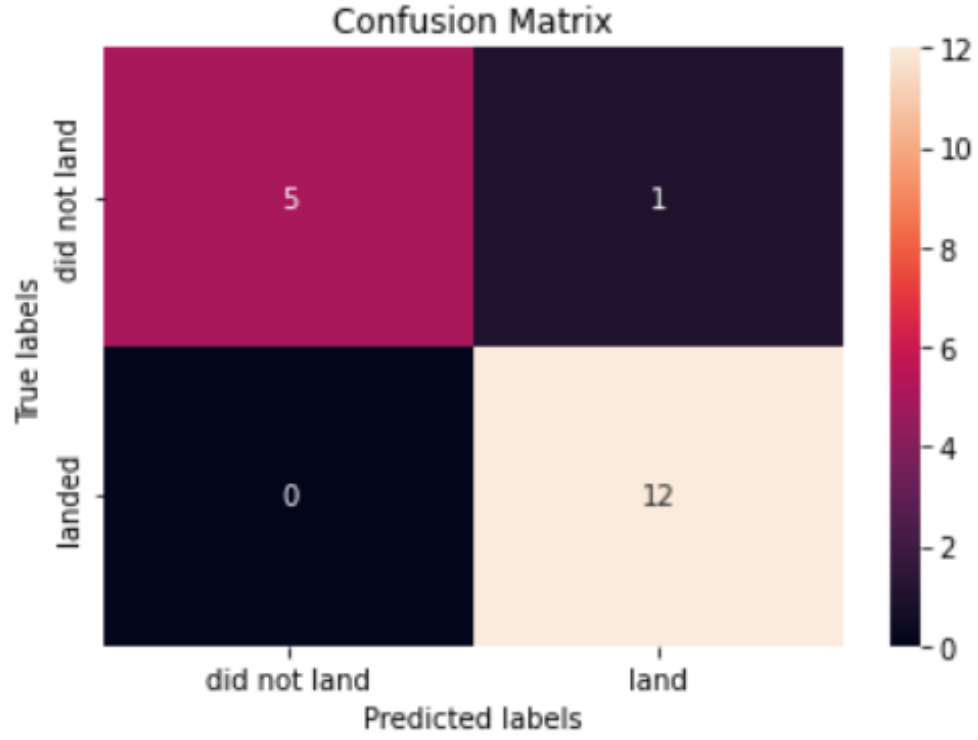

```
[[ 6  0]
 [ 0 12]]
```

ML models predictive analysis(2/4)

Out[23]:

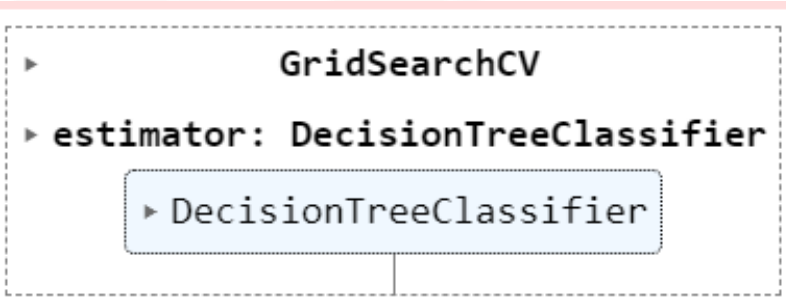



```
In [26]: yhat=svm_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



ML models predictive analysis(3/4)

Out[28]:



In [29]:
```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 4,
'min_samples_split': 5, 'splitter': 'best'}
accuracy : 0.875
```

In [39]:
```
from sklearn.metrics import accuracy_score
tree = DecisionTreeClassifier(criterion='gini',max_depth=4, max_features='auto',min_samples_leaf=4, min_samples_split=5, splitter
tree.fit(X_train,Y_train)
yhat=tree.predict(X_test)
print(accuracy_score(Y_test,yhat))
```

ML models predictive analysis(4/4)

Out[33]:

GridSearchCV

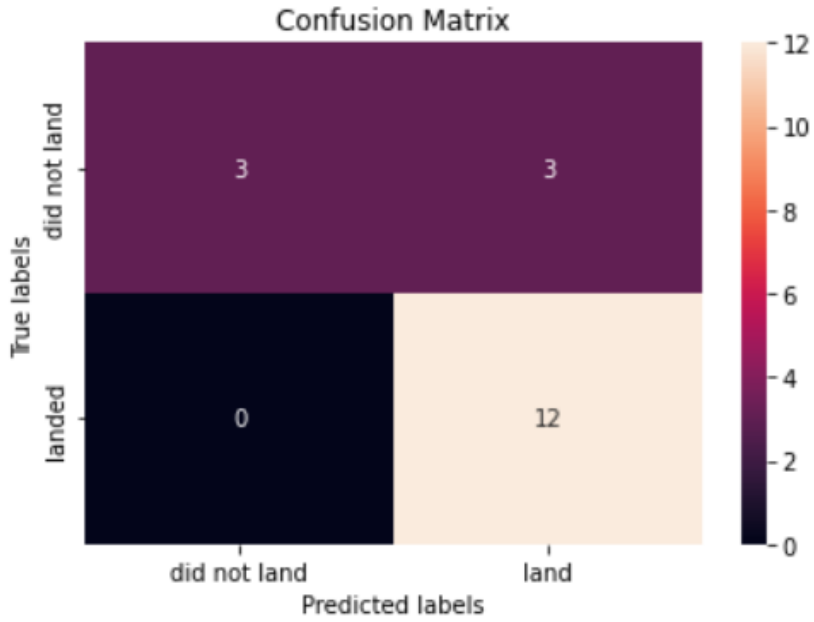
estimator: KNeighborsClassifier

KNeighborsClassifier

```
accuracy : 0.9
```

We can plot the confusion matrix

In [35]:
```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



OVERALL FINDINGS

- Findings
  - KSC LC-39A and VAFB SLC 4E has a success rate of 77%
  - The major problem with logistic regression classifier is false positives
  - Sigmoid kernel had the best results for SVM on the validation dataset
  - Decision Tree classifier has the highest accuracy > 83% for the test set, after selecting the best hyperparameters

CONCLUSION

We are able to establish the following conclusions after completing the applied data science capstone:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate ~100%.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

THANK YOU!!