Winning Space Race with Data Science

<Shaily Singh>
<19/04/2024>

Outline

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

Executive Summary

The methodologies used for this project are

- API
- Webscraping
- Perform data wrangling
- Sampling Data for a particular falcon (Falcon 9)
- Dealing with Nulls –calculate mean and replace null values
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models like KNN,SVM,Logistic Regression and DecisionTree

- Summary of all results
  - Data will be extracted using API and Webscraping and then prepared by removing nulls and data wrangling .Using exporatory data analysis and SQL we will understand the data and prepare for machine learning model using train and test set and predict the accuracy of the models.

Introduction

- Project background and context
  - Data science project to collect data from various sources, improve the data quality and exploring the processed data.
  - Insights will be provided by some basic statistical analysis and data visualization for a private space launch company called SpaceX.
  - 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.
  - We will predict if the Falcon 9 first stage will land successfully.
  - If we can determine if the first stage will land, we can determine the cost of a launch
- Problems you want to find answers
  - This information can be used if an alternate company wants to bid against SpaceX for a rocket launch
  - Rocket company SpaceY which wants to compete with Space X.
  - Use data methodologies and machine learning models to check if SpaceX will reuse its first stage.

Section 1

Methodology

Methodology

- Data collection methodology:
  - API and Webscraping
- Data wrangling
  - Wrangling data using API
  - Sampling Data –exclude data for falcon 1 – we want only falcon 9
  - Dealing with Nulls –find and replace null values with mean values
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  The training data is divided into validation data, a second set used for training data; then the models are trained and hyperparameters are selected using the function GridSearchCV.

  find the model fit using training data

  Predict the data using LogisticRegression,KNN,decision tree and find the accuracy score

  We have finally - total 83 rows – training 72 and test 18

Data Collection using API

Past Launch data API -https://api.spacexdata.com/v4/launches/past

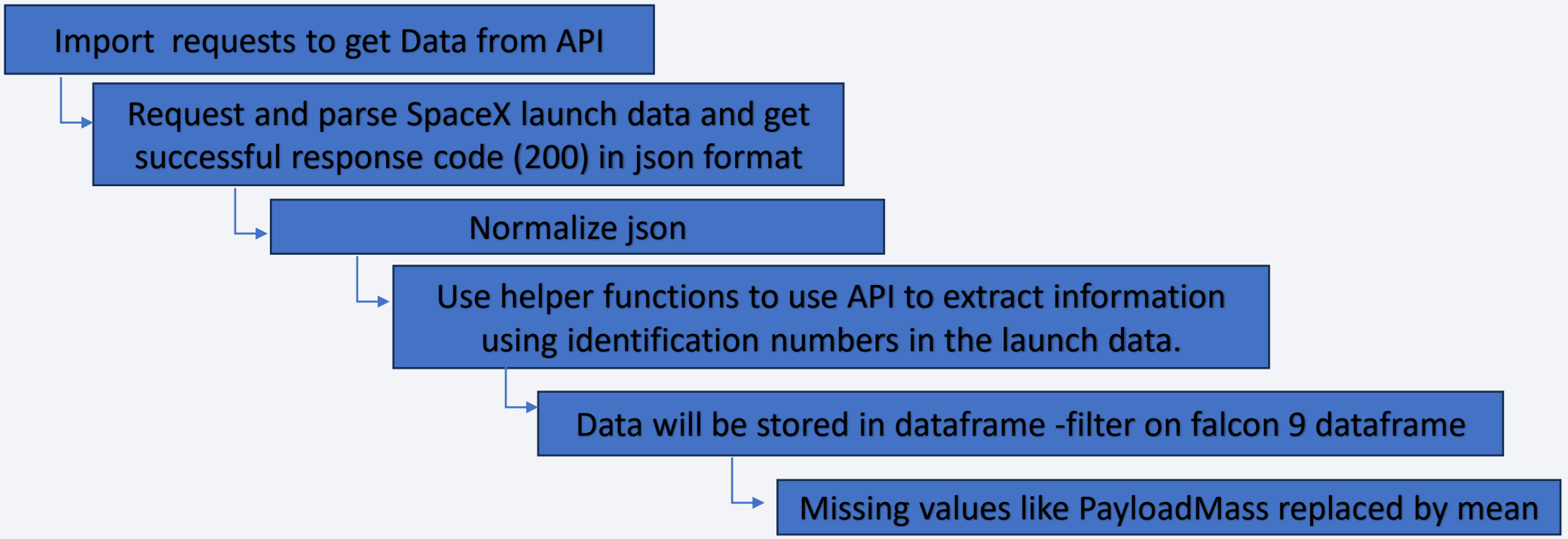
boosterName from rocket column(API - https://api.spacexdata.com/v4/rockets/)

Launch site, latitude andlongitude being used from launchpad API(https://api.spacexdata.com/v4/launchpads/)

Payloadmass and Orbit from payload API (https://api.spacexdata.com/v4/payloads/")

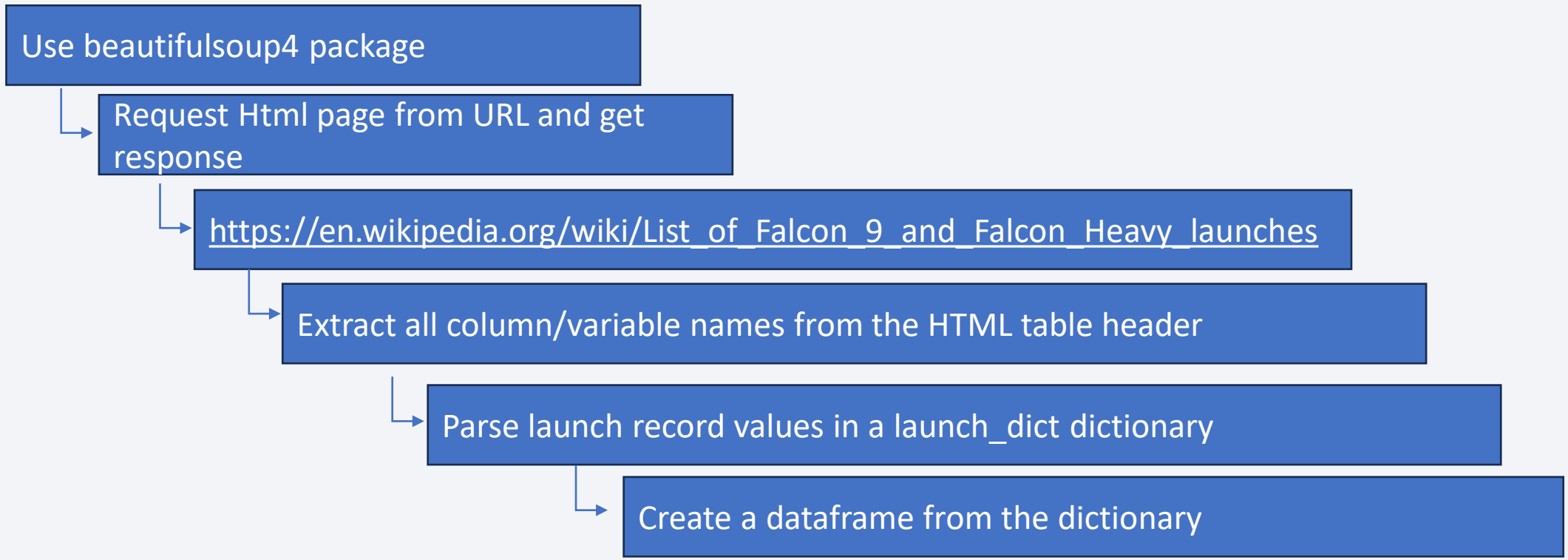
outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to seperate version of cores, the number of times this specific core has been reused, and the serial of the core.API https://api.spacexdata.com/v4/cores/

Data Collection – SpaceX API



https://github.com/ssingh2024/Coursera-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/ssingh2024/Coursera-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

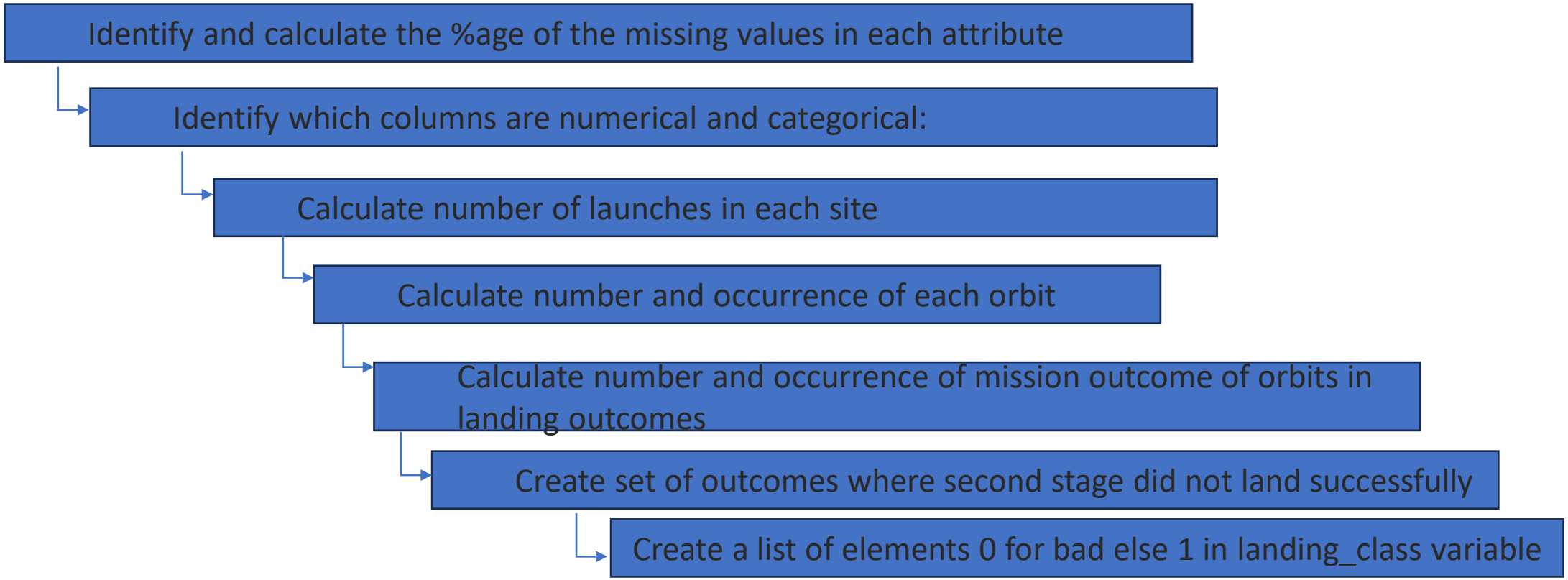
Data Wrangling

- Output or Rsult for the data can be follows (using Outcome)
  - Landing outcome Class (0 or 1) ,0 is bad outcome 1 is good outcome

Type of Landing	Description	Outcome
True Ocean	successfully landed to a specific region of the ocean	1
False Ocean	unsuccessfully landed to a specific region of the ocean	0
True RTLS	successfully landed to a ground pad	1
False RTLS	unsuccessfully landed to a ground pad	0
TRUE ASDS	successfully landed on a drone ship	1
FALSE ASDS	Unsuccessfully landed on a drone ship	0
None ASDS and None	Failure to land	0

https://github.com/ssingh2024/Coursera-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Data Wrangling



EDA with Data Visualization

- **Charts were plotted**
  - Show relation between Flight number and Launch site
  - Show relation between Payload and Launch site
  - Show success rate of Orbit Types
  - Show relation of Flight number and Orbit type
  - Show relation of Payload and Orbit type
  - Show line chart of yearly average success rate
- **scatter plots and bar charts .**
  - Scatter plot -It provides a visual and statistical means to test the strength of a relationship between two variables. Scatter plots can be effective in measuring the strength of relationships.
  - Bar chart-Bar charts enable us to compare numerical values like integers and percentages. They use the length of each bar to represent the value of each variable. used to see success rate of orbit type.
  - Line chart to see the trend

- https://github.com/ssingh2024/Coursera-Project/blob/main/edadataviz%20(1).ipynb

EDA with SQL

- SQL queries performed were
  - Find unique launch site
  - 5 records where launch site name like %CCA
  - total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - date when the first successful landing outcome in ground pad was achieved
  - names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - total number of successful and failure mission outcomes
  - names of the booster_versions which have carried the maximum payload mass
  - records which display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - 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
- https://github.com/ssingh2024/Coursera-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map objects used
    - FoliumMap,Folium.Circle,Foilum.Map.marker,MarkerCluster,MosuePosition,Polyline

- These map objects were used to identify
    - Map with centre location as NASA Johnspace Space Center
    - Create a blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name
    - Create a blue circle at NASA Johnson Space Center's coordinate with a icon showing its name
    - circle for each launch site in data frame `launch_sites`
    - For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name as a popup label
    - create markers for all launch records.If a launch was successful `(class=1)`, then we use a green marker and if a launch was failed, we use a red marker `(class=0)`
    - Mouse Position to get the coordinate (Lat, Long) for a mouse over on the map
    - Display the distance between coastline point and launch site using the icon property
- https://github.com/ssingh2024/Coursera-Project/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- plots/graphs and interactions added to a dashboard
  - Pie Chart
  - Scatter plot
  - Ranger bar
- These graphs show the details
  - Pie Chart will show success failure outome for a launch site
  - Pie chart will show success of all launch sites
  - Scatter plot will show launch outcomes(class) wrt Payload mass
- https://github.com/ssingh2024/Coursera-Project/blob/main/Dash%20Assignment.docx

Predictive Analysis (Classification)

- Perform code tasks in Jupyter notebook using pandas and numpy
- Separate data in training and testing set
- Build different machine learning models and tune hyperparameters using GridSearchCV
- Accuracy as metrics for model used
- Find the best performing classification model
- https://github.com/ssingh2024/Coursera-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(2).ipynb

Predictive Analysis (Classification)

Logistic regression

GridSearch CV=10

dictionary parameters ={'C':[0.01,0.1,1], 'penalty':['l2'], 'solver':['lbfgs']}

tuned hyperparameters :(best parameters) {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}

Best score accuracy : 0.8464285714285713

Score 0.8333333333333334

Support Vector Machine

GridSearch CV=10

parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'), 'C': np.logspace(-3, 3, 5), 'gamma':np.logspace(-3, 3, 5)}

tuned hyperparameters :(best parameters) {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}

Best Score accuracy 0.8482142857142856

Score 0.8333333333333334

Decision Tree

GridSearch CV=10

parameters = {'criterion': ['gini', 'entropy'], 'splitter': ['best', 'random'], 'max_depth': [2*n for n in range(1,10)], 'max_features': ['auto', 'sqrt'], 'min_samples_leaf': [1, 2, 4], 'min_samples_split': [2, 5, 10]}

tuned hpyerparameters :(best parameters) {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}

Best score accuracy : 0.8892857142857142

Score 0.6666666666666666

K Nearest Neighbours

GridSearch CV=10

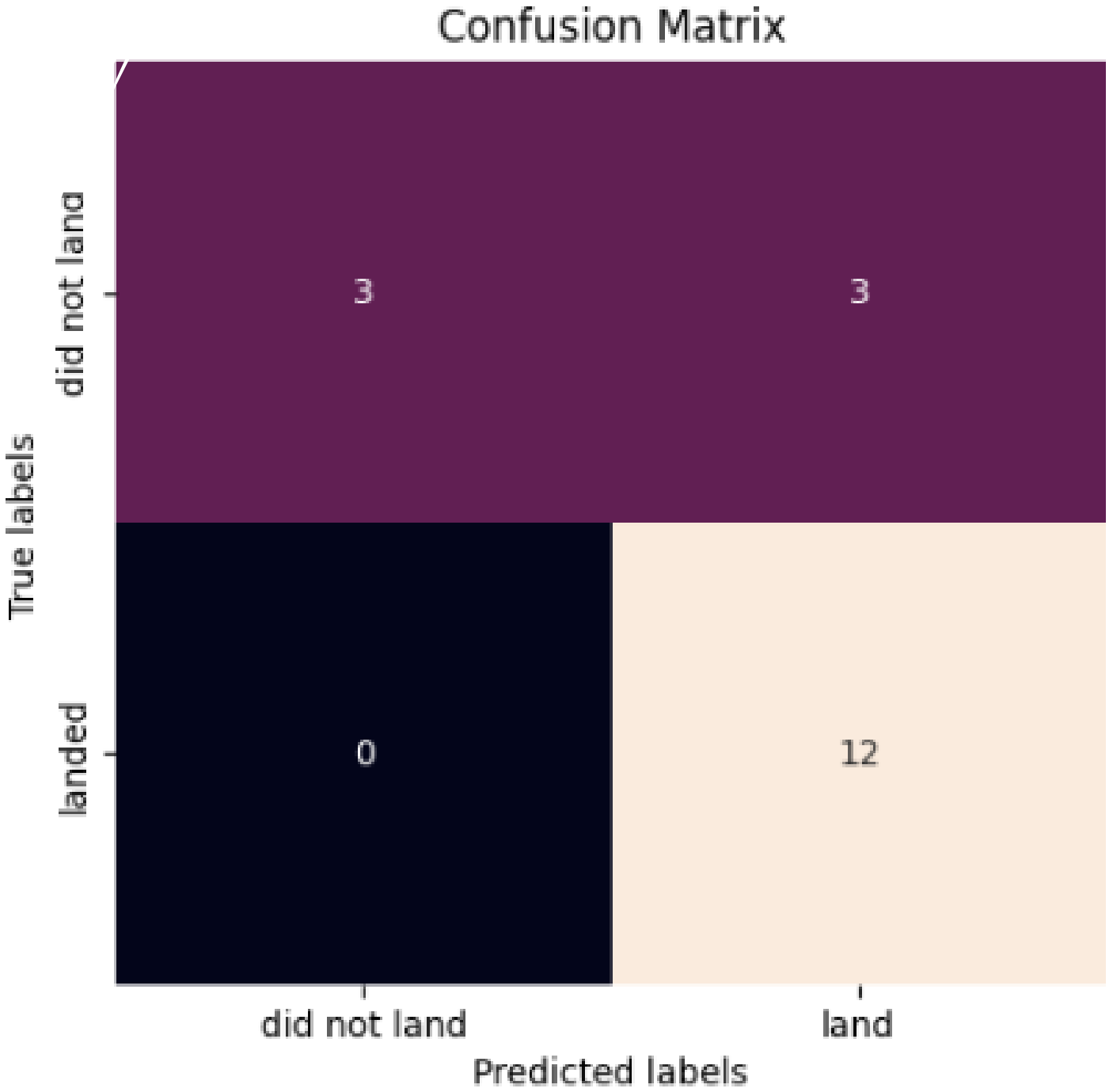
d parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'], 'p': [1,2]}

tuned hpyerparameters :(best parameters) {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}

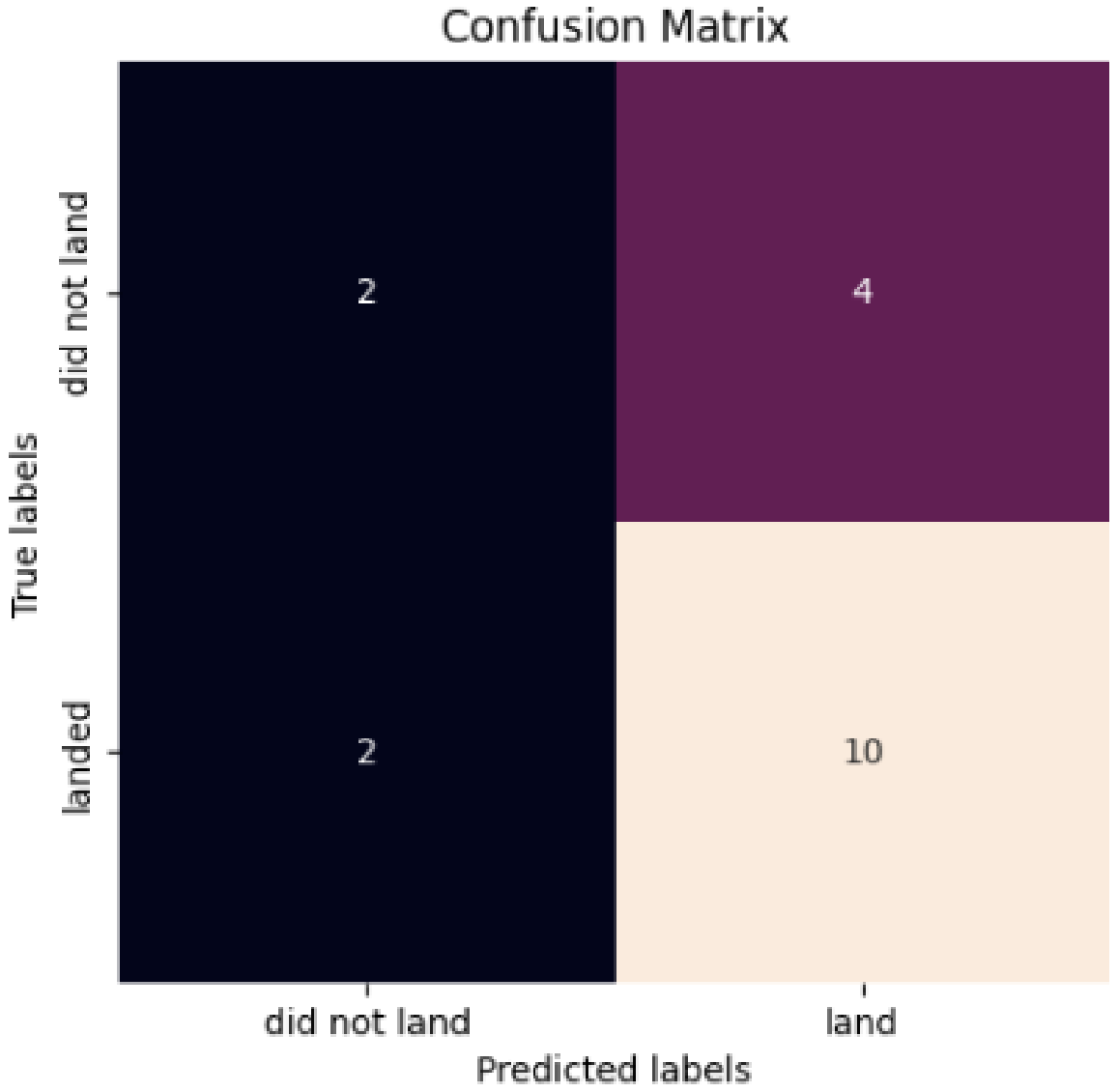
Best Score accuracy : 0.8482142857142858

Score 0.8333333333333334

Predictive Analysis (Classification) for Logistic Regression, SVM and KNN



Predictive Analysis (Classification) for Decision Tree



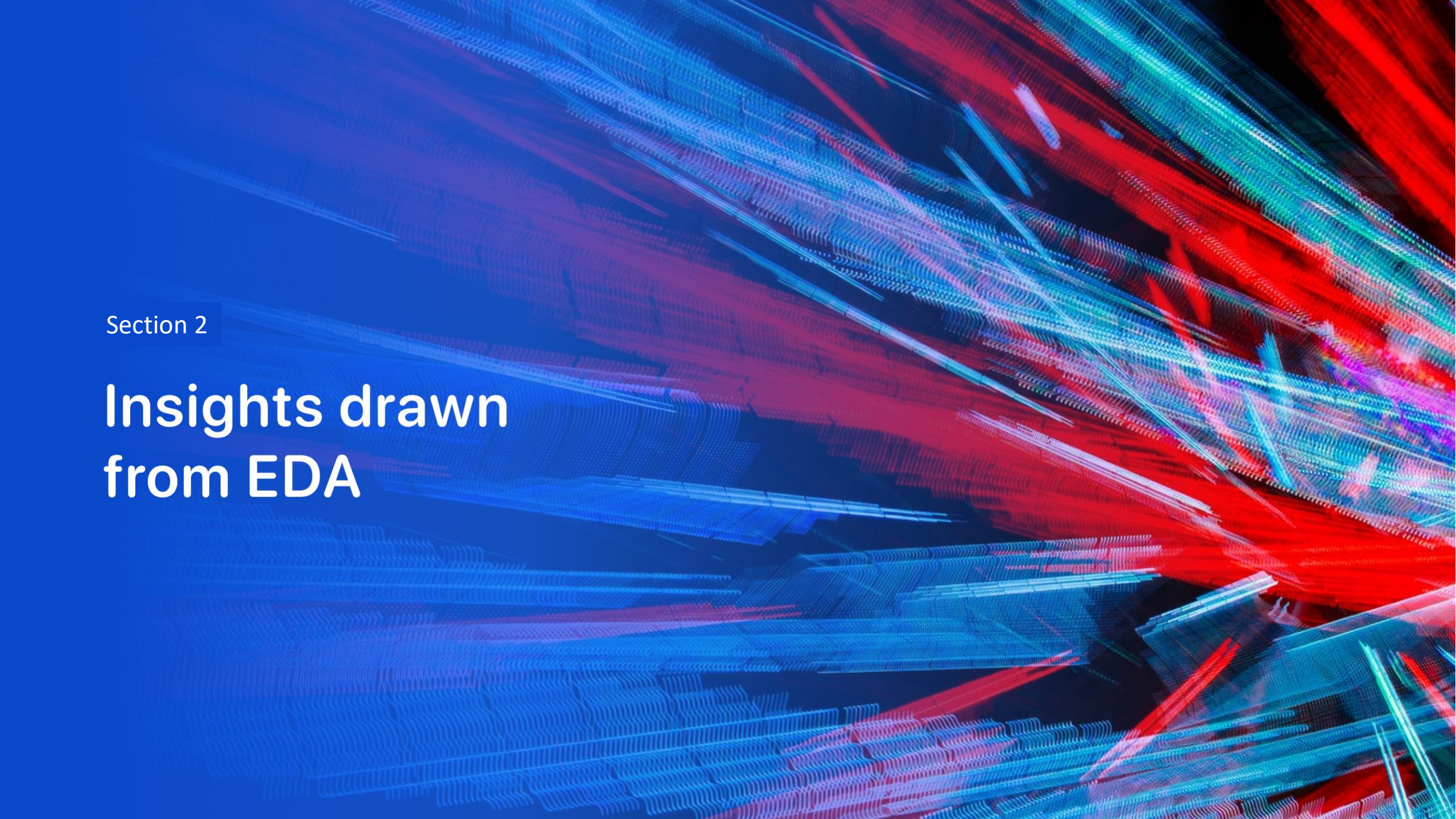
Results

Exploratory data analysis results

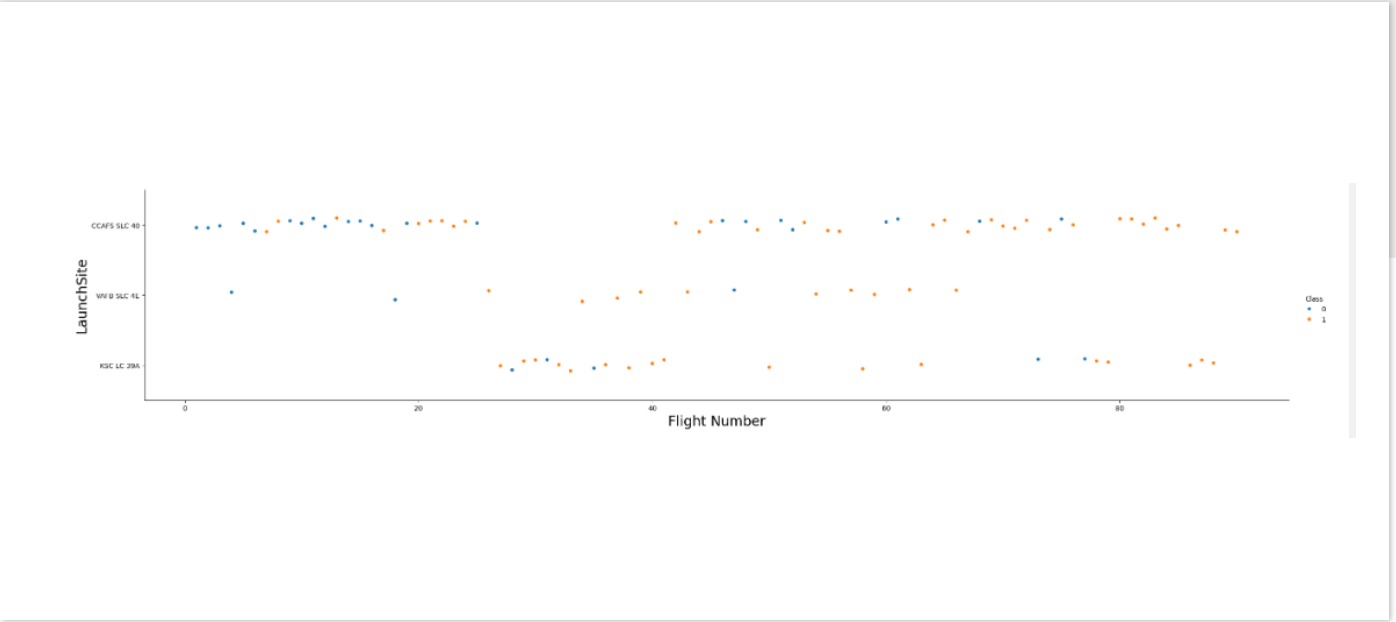
-

Interactive analytics demo in screenshots

Predictive analysis results


Section 2
Insights drawn from EDA

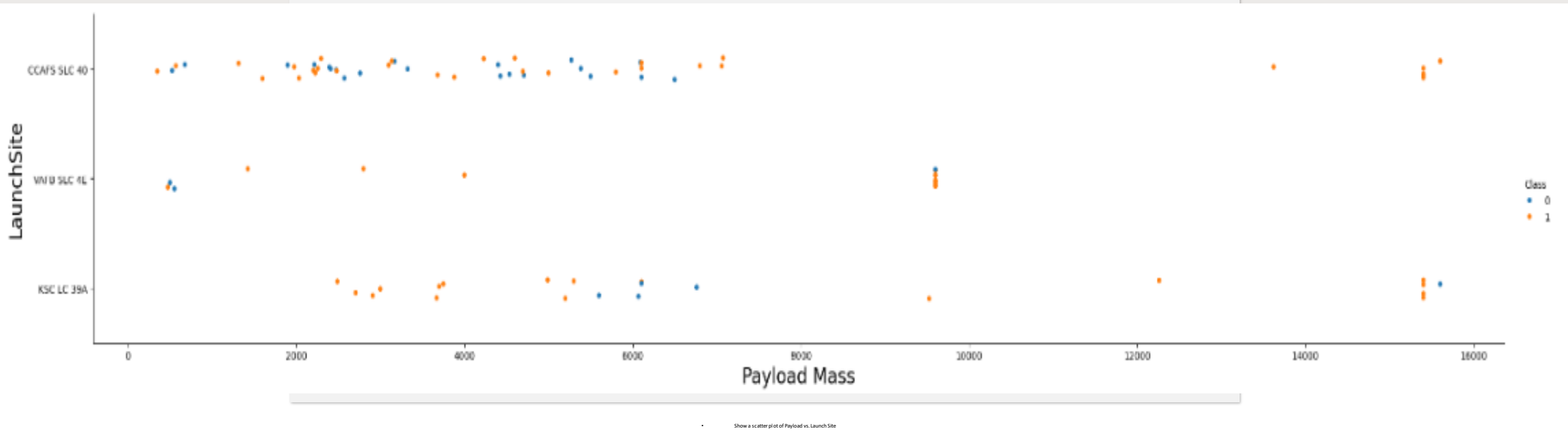
Flight Number vs. Launch Site



- Show a scatter plot of Flight Number vs. Launch Site

- Show the screenshot of the scatter plot with explanations
- We see that as the flight number increases, the first stage is more likely to land successfully

Payload vs. Launch Site



Show a scatter plot of Payload vs. Launch Site

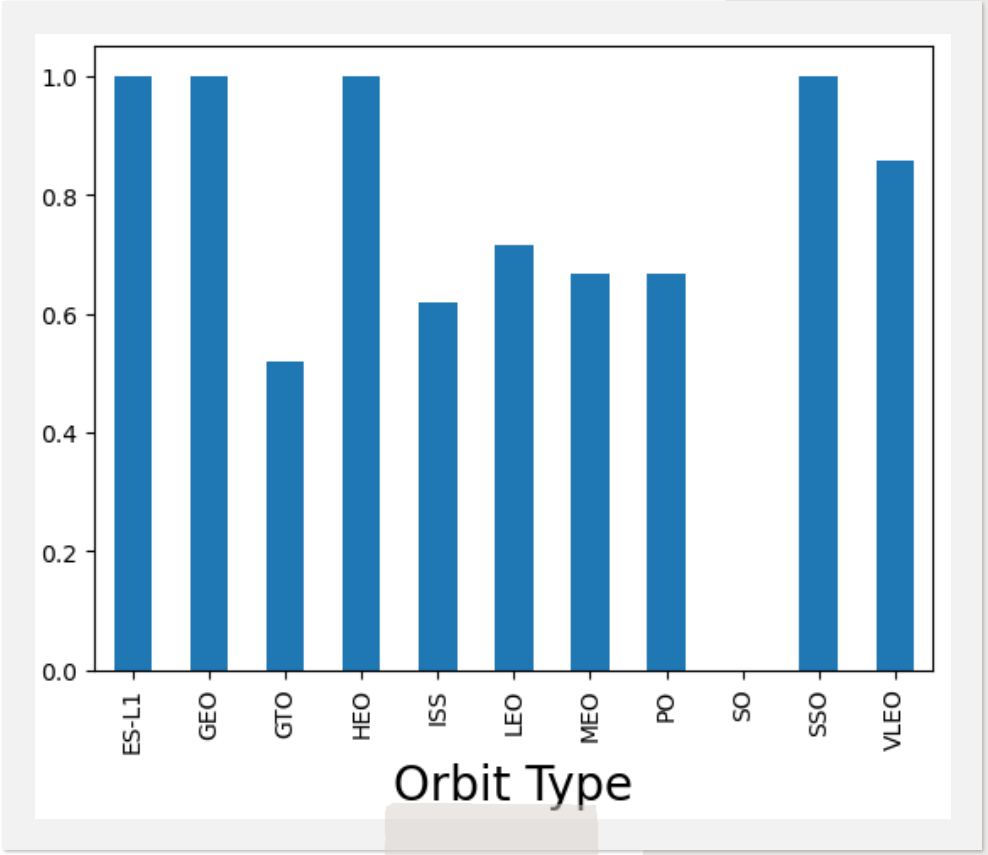
- Show the screenshot of the scatter plot with explanations

- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- For CCAFS outcome is 1 for payload greater than 12000

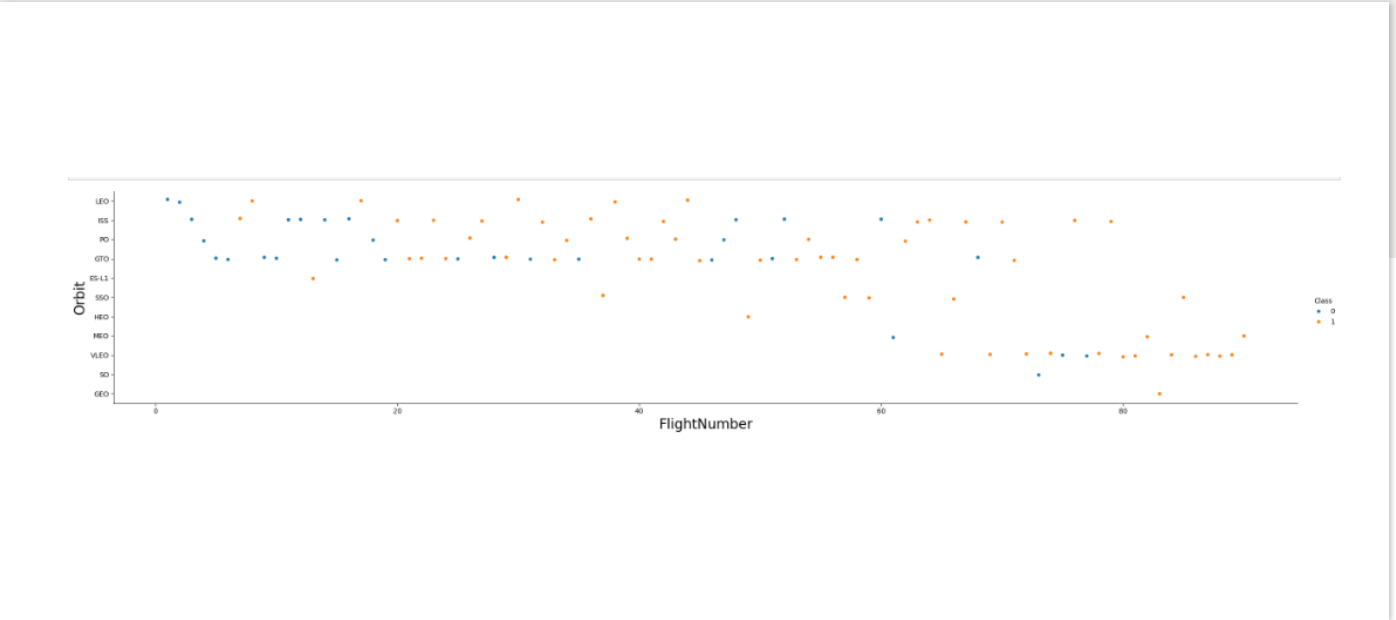
Success Rate vs. Orbit Type

Bar chart for the success rate of each orbit type

- ES-L1,GEO ,HEO,SSO have high success rates
- GTO has low success rate

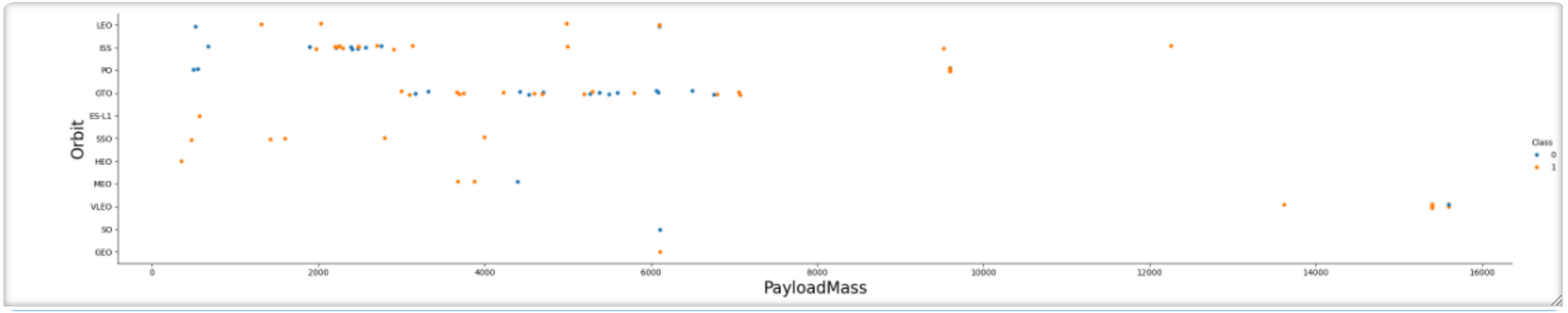


Flight Number vs. Orbit Type



scatter point of Flight number vs. Orbit type

- 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.

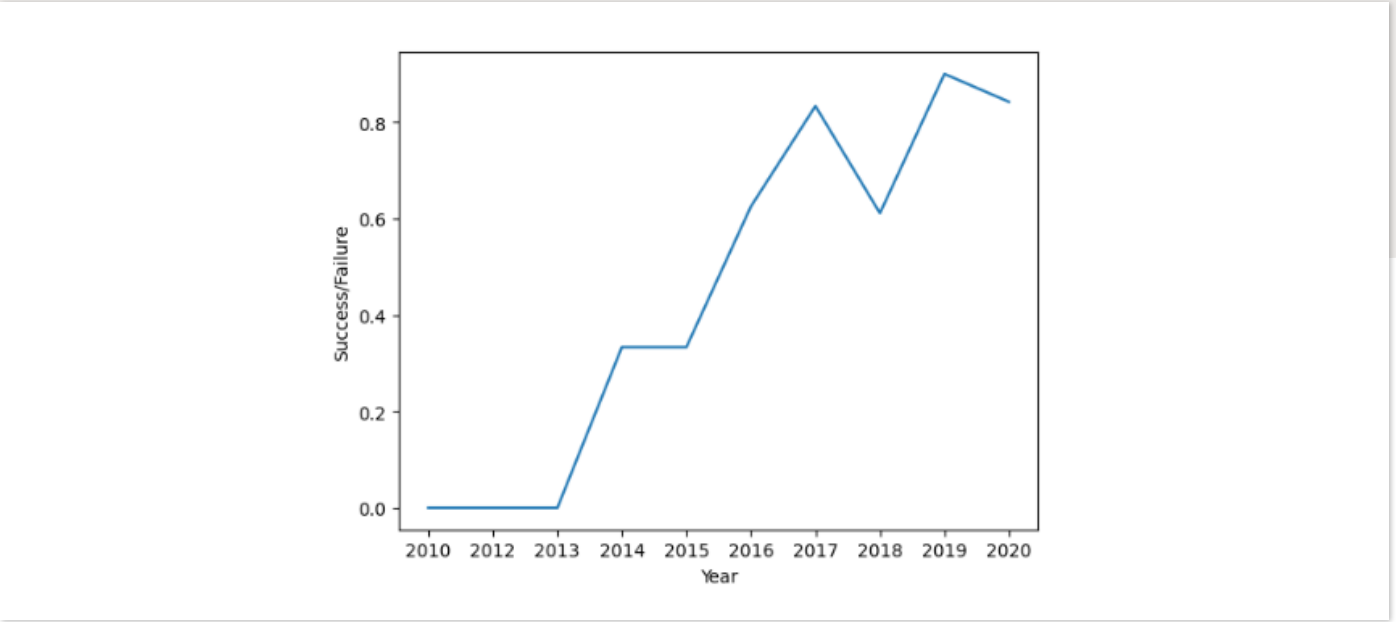


Payload vs. Orbit Type

scatter point of 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



Line chart of yearly average success rate

- sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites

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

- Unique launch sites show total number of launch sites in total

Task 1

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

```
[11]: %sql select distinct(Launch_Site) from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

[11]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

We have Launch sites with CCA, the table gives only 5 records.

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

[12]:
```
%sql select * from SPACEXTABLE where Launch_site like 'CCA%' limit 5
```

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

[12]:

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

- Calculate the total payload carried by boosters from NASA
- The query gives total mass payload of NASA (CRS).

Task 3

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

```
[13]: %sql select  sum(PAYLOAD_MASS__KG_),Customer from SPACEXTABLE where Customer='NASA (CRS)' group by Customer
```

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

[13]:

sum(PAYLOAD_MASS__KG_)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

This query shows average payload mass carried by booster version F9 v1.1

Task 4

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

```
[28]: %%sql select  sum(PAYLOAD_MASS__KG_)/count(Booster_Version) average_payload_mass,Booster_Version from SPACEXTABLE
sql where Booster_Version='F9 v1.1';
```

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

[28]:

average_payload_mass	Booster_Version
2928	F9 v1.1

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

This shows first successful landing outcome on ground pad.

Task 5

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

Hint:Use min function

```
[15]: %sql select  min(date) from SPACEXTABLE where landing_outcome like 'Success%'
```

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

[15]:

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

[17]:
```
%sql select * from SPACEXTABLE where PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000 and landing_outcome='Success (drone ship)'
```

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

[17]:

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

- Calculate the total number of successful and failure mission outcomes

Task 7 ¶

List the total number of successful and failure mission outcomes

[18]:
```
%sql select  count(Mission_Outcome),Mission_Outcome from SPACEXTABLE group by Mission_Outcome
```

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

[18]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Task 8

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

```
[31]: %%sql select  distinct(Booster_Version) from SPACEXTABLE where PAYLOAD_MASS__KG_
      in (select max(PAYLOAD_MASS__KG_)from SPACEXTABLE)
```

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

[31]:

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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[32]:
```
%%sql select  substr(Date, 6,2) Month,Landing_Outcome ,Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015'
and Landing_Outcome='Failure (drone ship)'
```

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

[32]:

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

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.
- We can see Success in drone ship is more.

Task 10

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.

[33]:
```
%%sql select  count(Landing_Outcome) my_count,Landing_Outcome from SPACEXTABLE
where date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by my_count desc
```

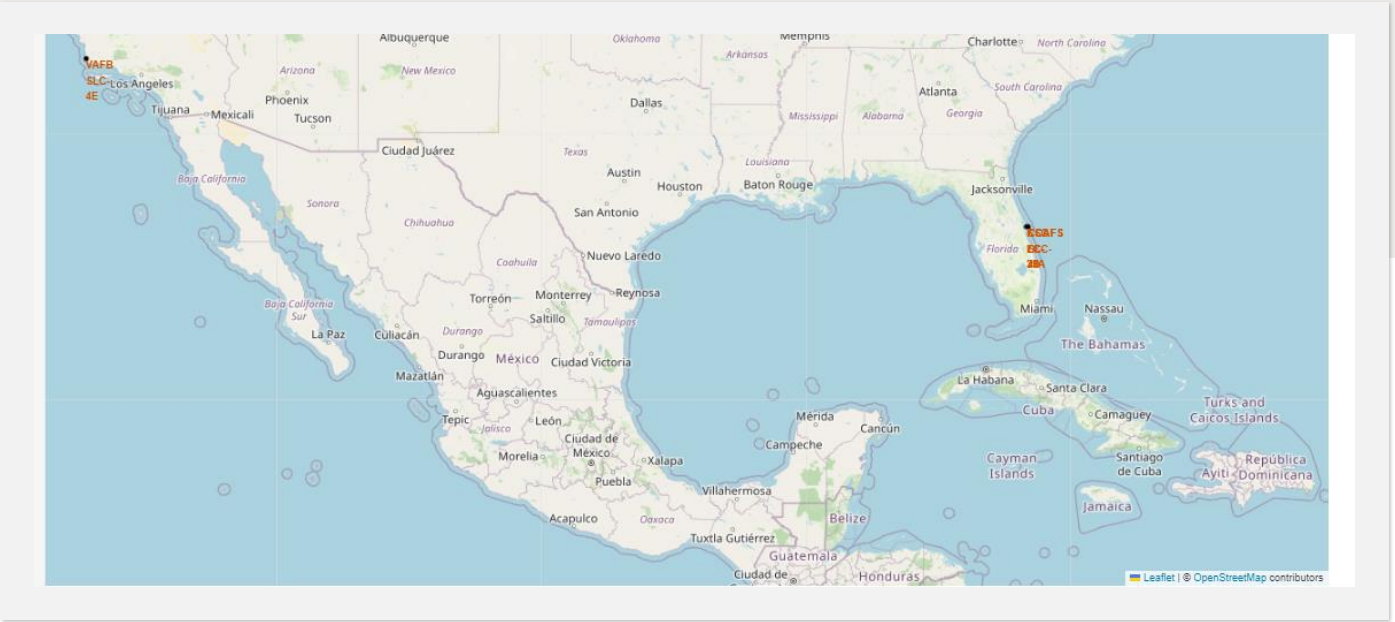
```
 * sqlite:///my_data1.db
Done.
```

[33]:

my_count	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)


Section 3
Launch Sites
Proximities Analysis

Falcon 9 Launch site Locations

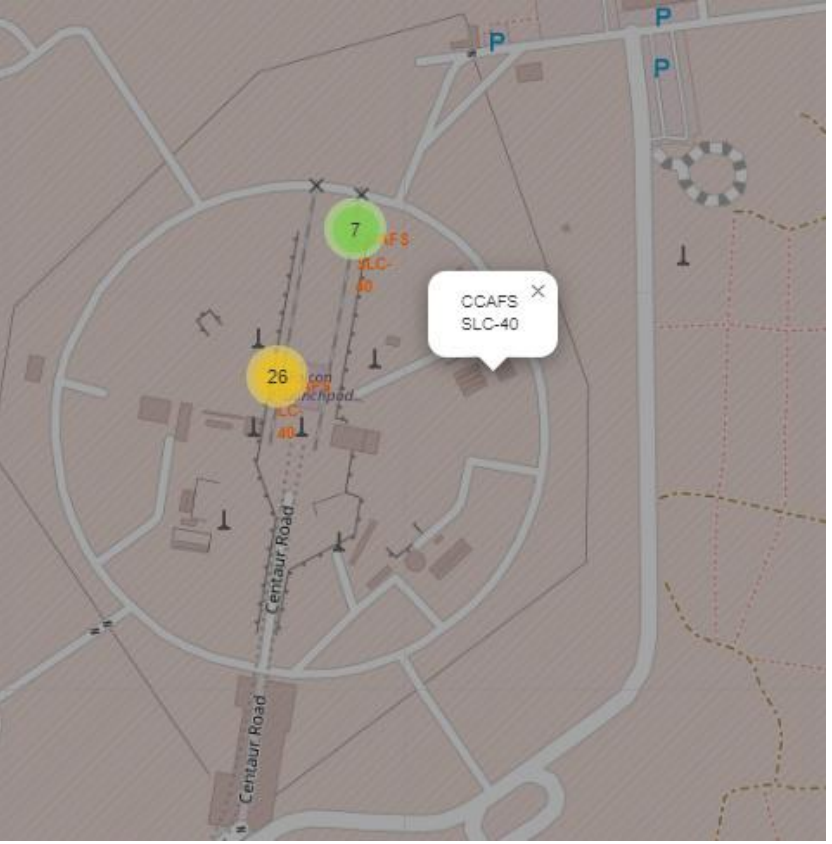


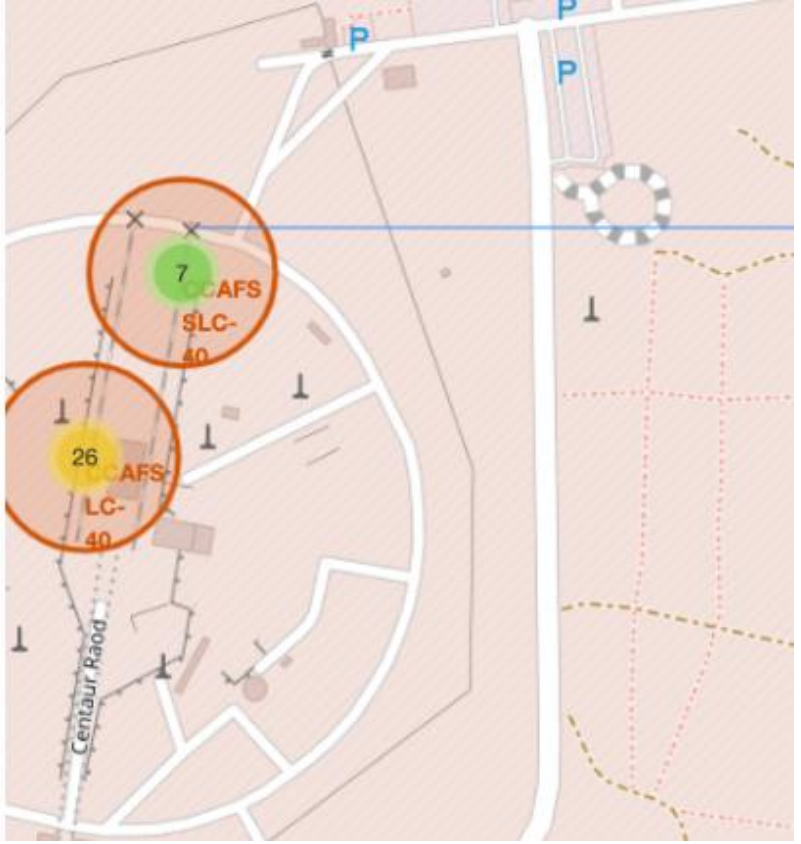
Replace <Folium map screenshot 1> title with an appropriate title

- Launch sites in proximity to the equator
- Launch sites close proximity to the coast

Map Markers for Success/Failure Landings

The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.

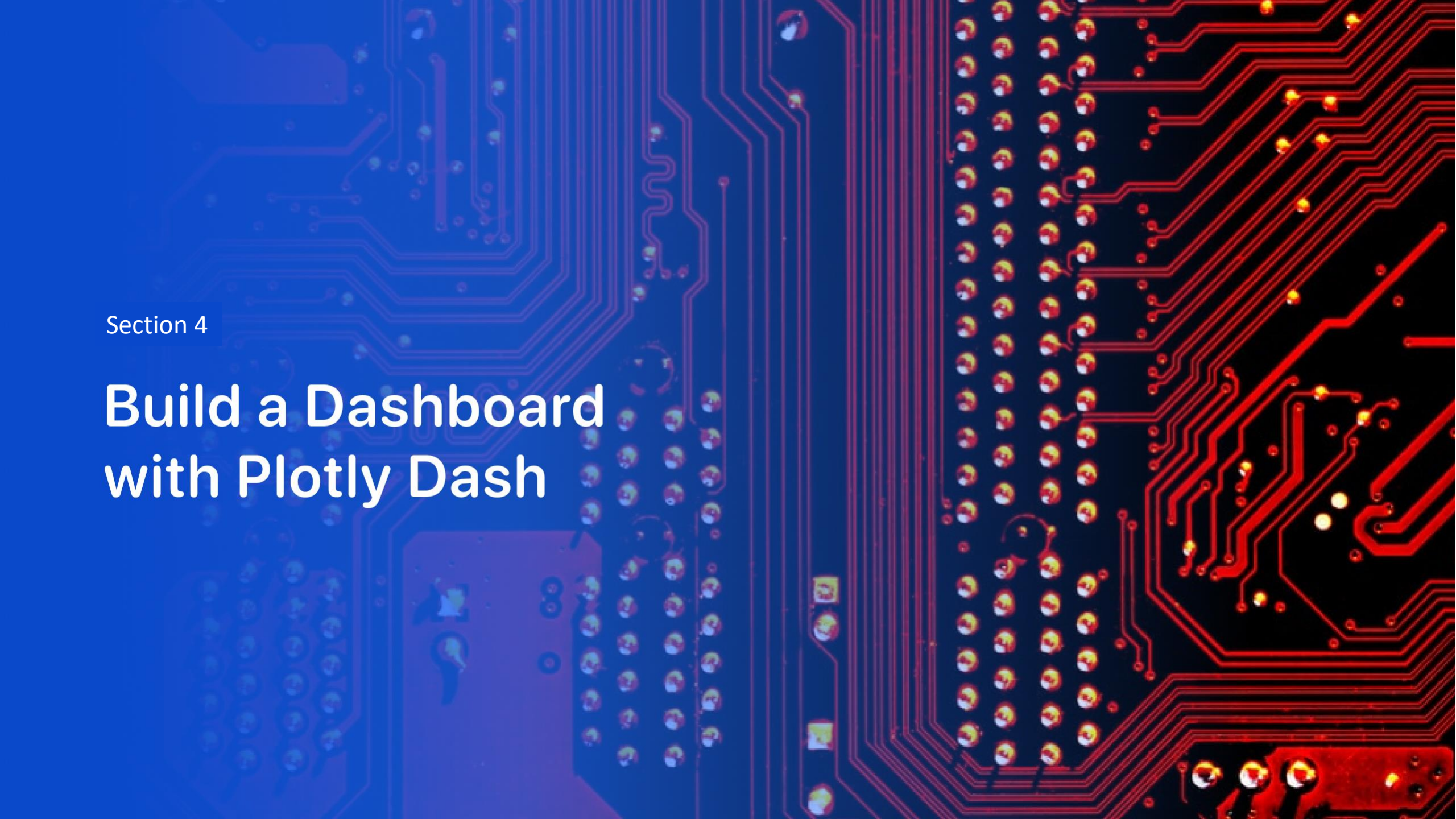




Distance from Launch site to Proximities

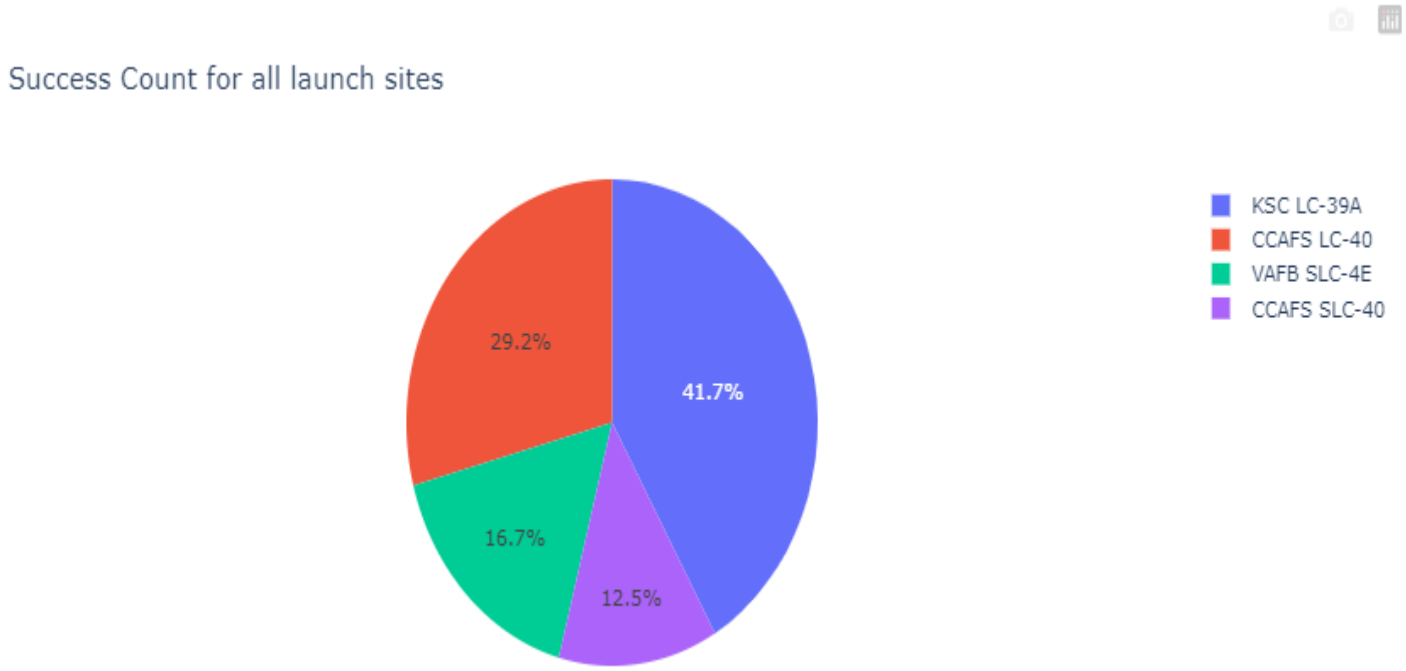
```
distance_highway = calculate_distance(launch_site_lat,
print('distance_highway =',distance_highway, ' km')
distance_railroad = calculate_distance(launch_site_lat
print('distance_railroad =',distance_railroad, ' km')
distance_city = calculate_distance(launch_site_lat, la
print('distance_city =',distance_city, ' km')
```

```
distance_highway = 0.5834695366934144  km
distance_railroad = 1.2845344718142522  km
distance_city = 51.434169995172326  km
```


Section 4
Build a Dashboard with Plotly Dash

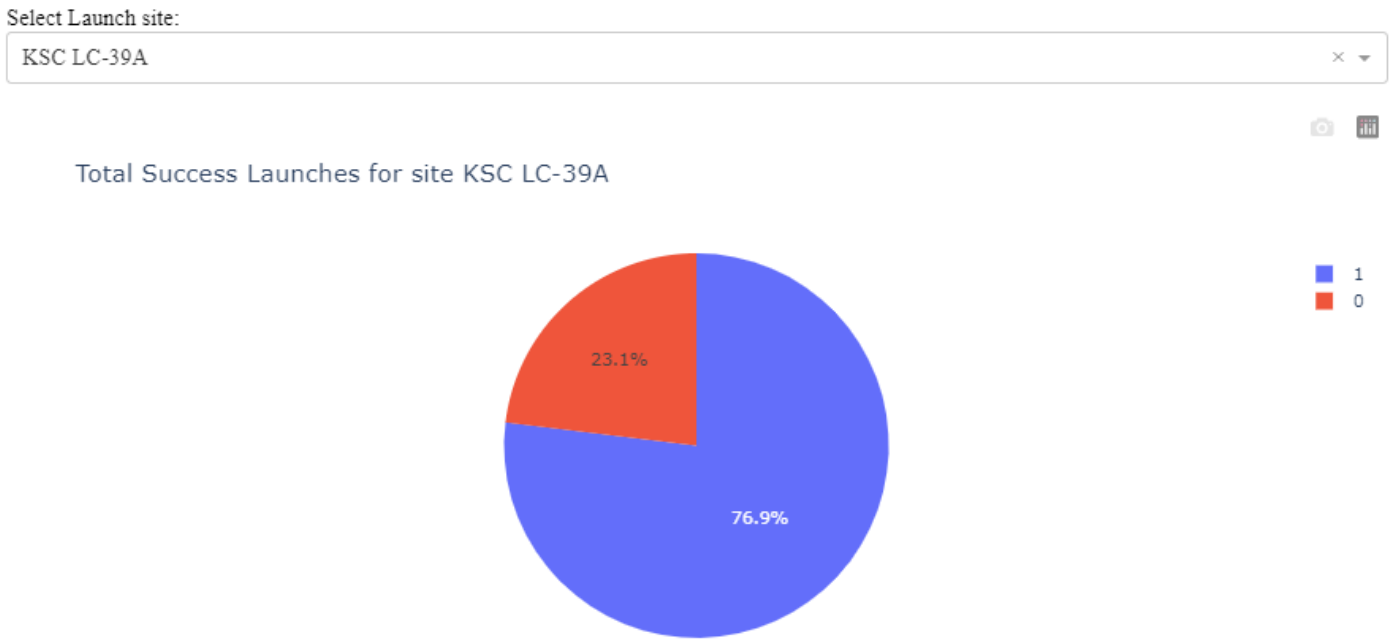
Launch success count for all sites

- With all launch sites selected, the pie chart displays the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites. •
- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A.

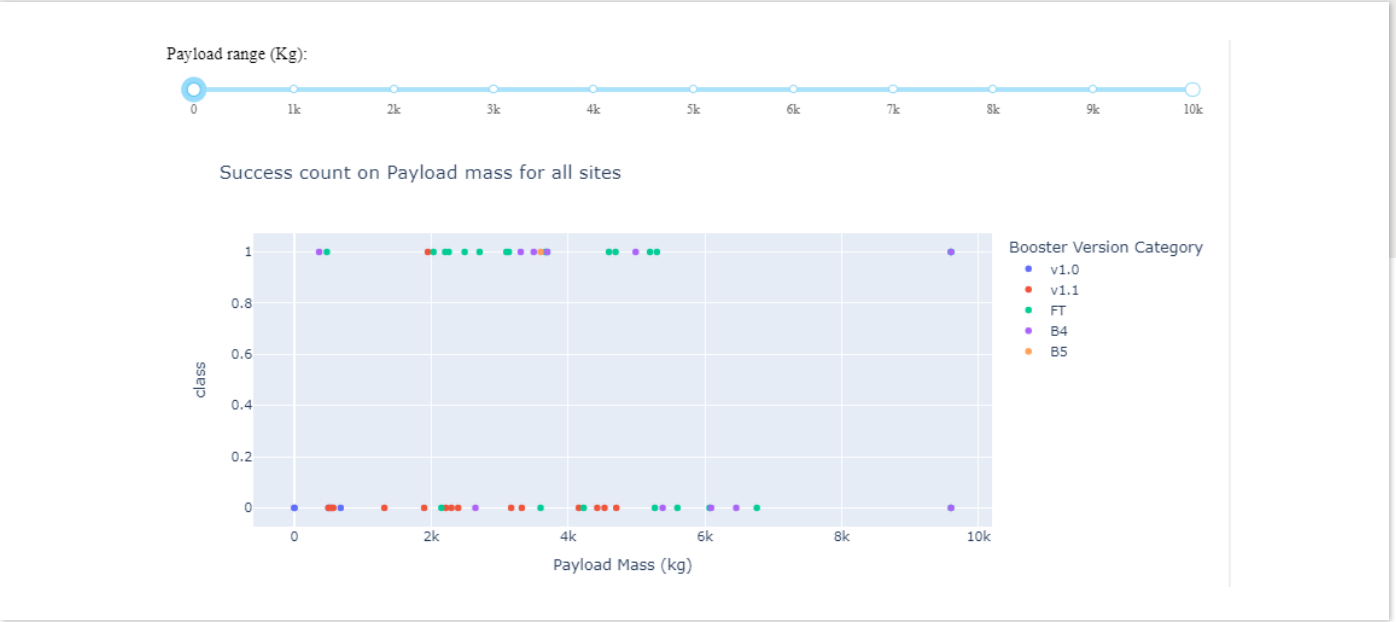


Launch site with highest success ratio

- KSC LC-39A was the launch site that had the highest Falcon9 first stage landing success rate which is 76.9%



Payload versus Launch Outcome

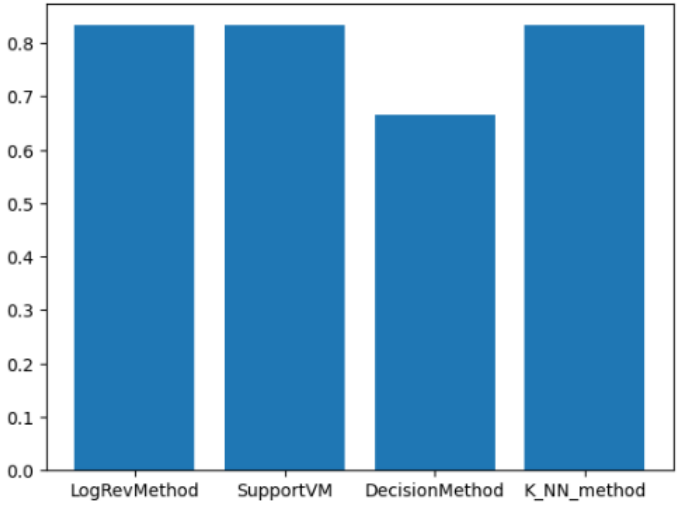


- This screenshots is of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider.
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.
- • The ‘FT’ booster version category has the largest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



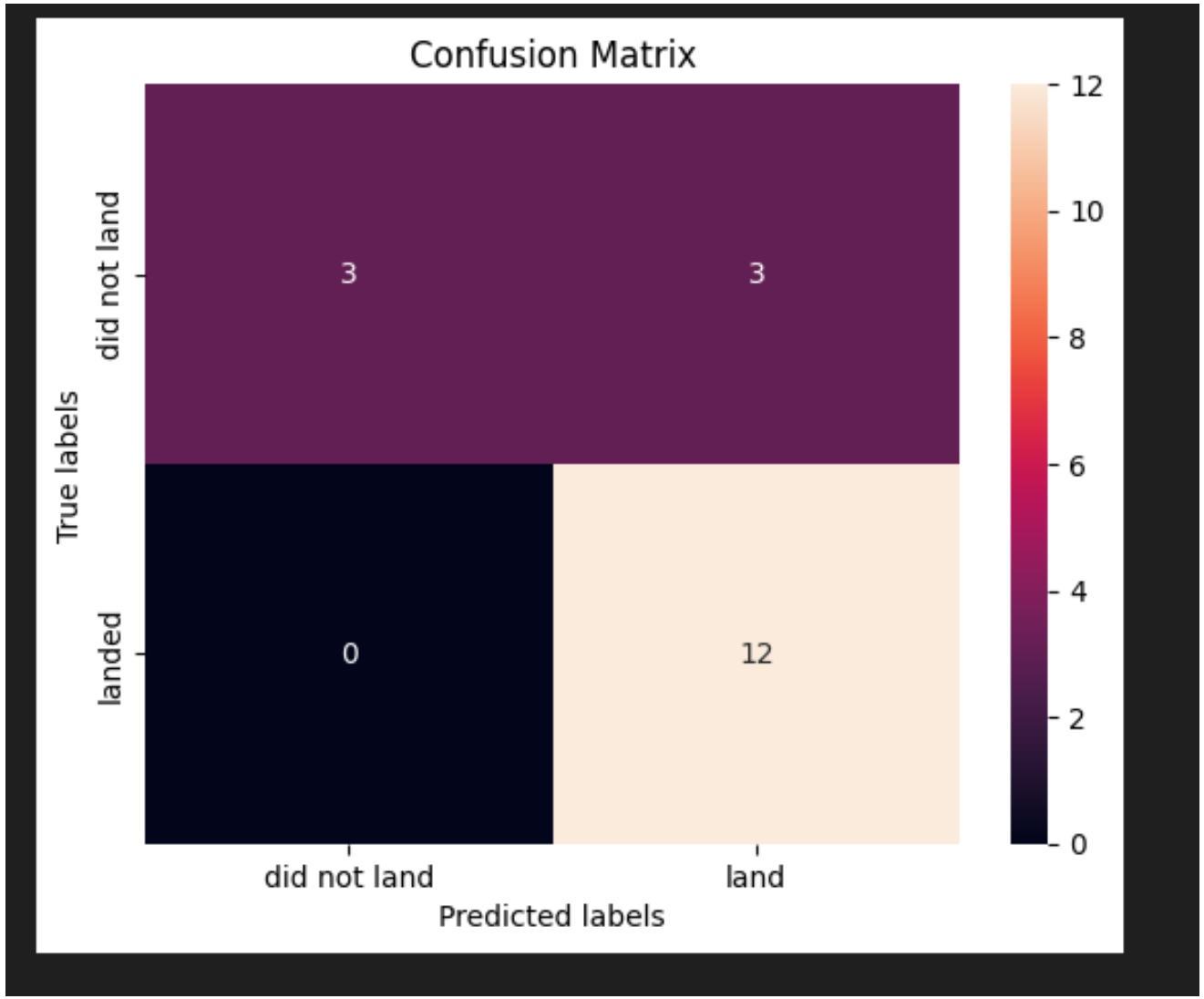
Accuracy

Logistics Regression method: 0.8333333333333334

Support Vector Machine method: 0.8333333333333334

Decision tree method: 0.6666666666666666

K nearest neighbors method: 0.8333333333333334



Confusion Matrix

- Shown here is the confusion matrix for the Logistic Regression model.

- Prediction Breakdown:

  12 True Positives and 3 True Negatives • 3 False Positives and 0 False Negatives

Conclusions

- SpaceX does not have a perfect track record of Falcon 9 first stage landing outcomes.
- SpaceX's Falcon 9 first stage landing outcomes have been trending towards greater success as more launches are made.
- The machine learning models can be used to predict future SpaceX Falcon 9 first stage landing outcomes.

Appendix

- SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- Wikipedia (Webpage): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- SpaceX (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321ENSkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetworkChannel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
- Launch Geo (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv • Launch Dash (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

Thank you!