IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

George Mathew
15th February 2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection through API, SpaceX launch data is gathered from SpaceX REST API and used to predict whether SpaceX will attempt to land a rocket or not.
  - Data Collection of Falcon9 launch data by Web Scraping HTML tables in related Wiki pages using BeautifulSoup package
  - Data Wrangling after converting into Pandas data frame and using API, sampling data and dealing with Nulls. And convert landing outcomes to classes y
  - Exploratory Data Analysis with SQL combining attributes correlated with successful landing and preparing data for ML model to predict outcome of 1st stage
  - Exploratory Data Analysis with Data Visualization to explore and manipulate data in an interactive and real time way using Folium and Plotly dash
  - Machine Learning Pipeline to predict successful landing of Falcon9 1st stage
- **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 upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. The first stage is quite large and expensive. Unlike other rocket providers, SpaceX's Falcon9 can recover the first stage. Sometimes the first stage does not land, sometimes it will crash, other times, SpaceX will sacrifice the first stage due to the mission parameters like payload, orbit and customer.
- **Problems you want to find answers**
- **Problem no. 1**
- How can the success rate of landing help determine a realistic price of each launch ?
- **Problem no. 2**
- Which all variables affect the success rate of landing and how can the learnings be used to improvise upon the launches for increasing the successful landings ?
- **Problem no. 3**
- Which is the best machine learning model that can be used based on the accuracy so as to ensure the best successful landing rate ?

Section 1

Methodology

Methodology
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API (api.spaceXdata.com/v4/) and webscraping related wiki pages
- Perform data wrangling
  - Wrangling done using an API, Sampling data and dealing with Nulls, (by getting booster version, launch site, Payload Data, Core Data) and converting landing outcomes to classes y
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - By preprocessing, allowing us to standardize our data and train_test_split, allowing us to train split our data into training and testing data. Find hyperparameters and determine the model with best accuracy.

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts
- Datasets were collected by
  - Working with SpaceX launch data that is gathered using the SpaceX REST API.
  - API gives us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.
  - Data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup.

Data Collection – SpaceX API

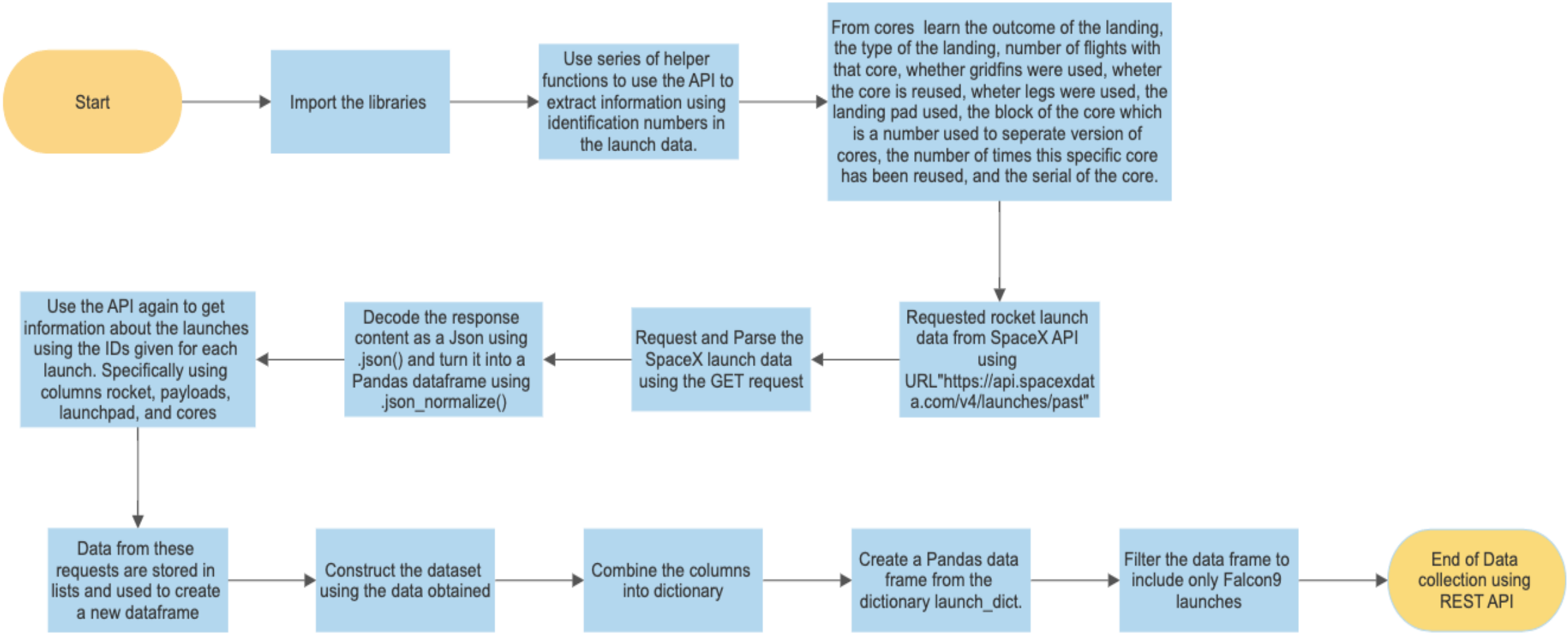
- **Present your data collection with SpaceX REST calls using key phrases and flowcharts**
- Individual slides added for key phrases and flowcharts
- **Add the GitHub URL of the completed SpaceX API calls notebook**
- https://github.com/georgebhilai/Capstone-Project/blob/b7780d00d4d379a771258afbb3cfc83f527e594b/jupyter-labs-spacex-data-collection-api%20(2).ipynb

Place your flowchart of SpaceX API calls here

Presentation of Data Collection using Key phrases

- Import the Libraries
- Use series of helper functions to use the API to extract information using identification numbers in the launch data
- From cores learn the 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.
- Use the API again to get information about the launches using the IDs given for each launch. Specifically using columns rocket, payloads, launchpad, and cores
- Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()
- Request and Parse the SpaceX launch data using the GET request
- Requested rocket launch data from SpaceX API using URL"https://api.spacexdata.com/v4/launches/past"
- Data from these requests are stored in lists and used to create a new dataframe
- Construct the dataset using the data obtained
- Combine the columns into dictionary
- Create a Pandas data frame from the dictionary launch_dict.
- Filter the data frame to include only Falcon9 launches
- End of Data collection using REST API

Presentation of Data Collection using Flowchart



Data Collection - Scraping

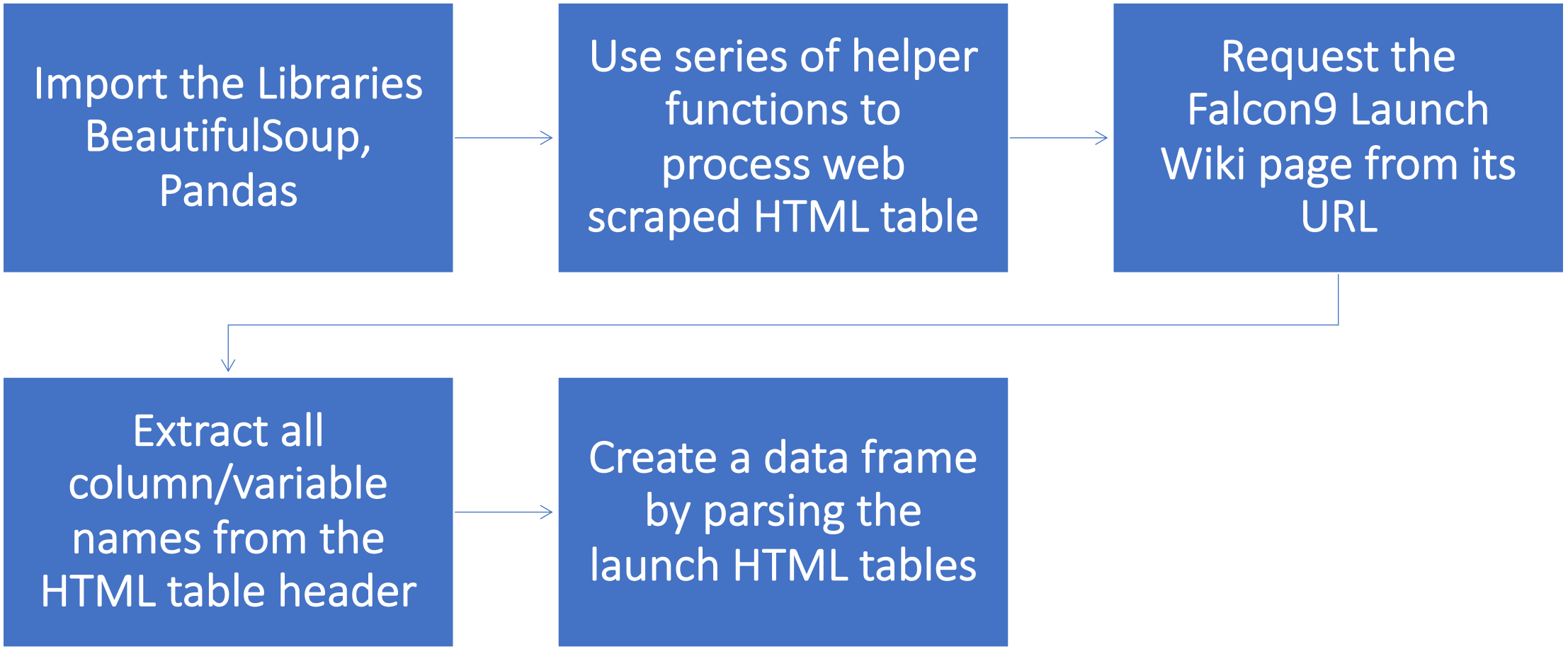
- **Present your web scraping process using key phrases and flowcharts**
- Individual slides added for key phrases and flowcharts
- **Add the GitHub URL of the completed web scraping notebook.**
- https://github.com/georgebhilai/Capstone-Project/blob/cfca724fa2e592e03380c56a340a42c2fa3acc2b/jupyter-labs-webscraping.ipynb

Place your flowchart of web scraping here

Presentation of Web Scraping process using Key phrases

- Import the Libraries BeautifulSoup, Pandas
- Use series of helper functions to process web scraped HTML table
- Request the Falcon9 Launch Wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

Presentation of Web Scraping process using Flowchart



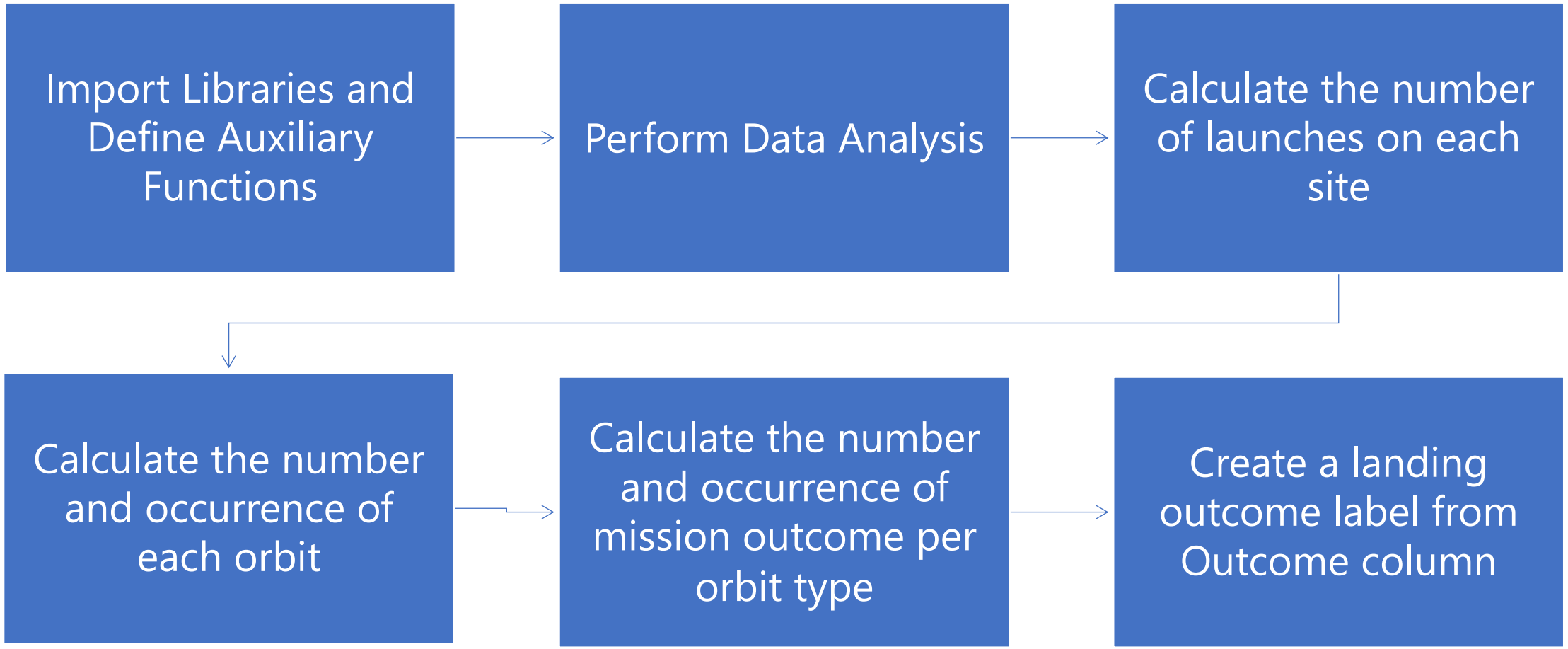
Data Wrangling

- **Describe how data were processed**
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- **You need to present your data wrangling process using key phrases and flowcharts**
- Individual slides added for key phrases and flowcharts
- **Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose**
- https://github.com/georgebhilai/Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Presentation of Data Wrangling process using Key phrases

- Import Libraries and Define Auxiliary Functions
- Perform Data Analysis
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column

Presentation of Data Wrangling process using Flowchart



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

- **Scatter Graphs**:
- **Flight Number Vs Payload Mass**
- >> After plotting the graph and overlaying the outcome of the launch we observed that as the flight number increases, the first stage is more likely to land successfully. The graph was used to find out the effect of change in flight number on the successful landing of the first stage.
- >> The graph was also used to find out the effect of Payload Mass on the successful landing of the first stage and it appeared that the more massive the Payload, the less likely the first stage will return
- **Flight Number Vs Launch Site**
- >> To find out the success rates of different launch sites. We find that different launch sites have different success rates. CCAFS LC-40 has a success rate of 60% while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- **Payload Vs Launch Site**
- >> On observing the scatter plot chart we find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000). In respect of the other two launch sites, while in the case of CCAFS LC-40 launch site there is concentration of launches with comparatively lower payload in case of KSC LC-39A there is almost uniform distribution of the payload-wise launches.

EDA with Data Visualization

- **Scatter Graphs**:
- **Orbit Vs Flight Number**
- >> On observing the scatter plot chart we see that in the LEO orbit the Success appears related to the number of flights, on the other hand, there seems to be no relationship between flight number whenin GTO orbit.
- **Payload Vs Orbit Type**
- >> On observing the scatter plot chart we see that with heavy payloads the successful landing or positive landing rate are more for Polar LET and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.
- Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation . Scatter plots usually consist of a large body of data.
- **Bar Graph:**
- **Success rate Vs Orbit type**
- >> On observing the Bar graph we see that ES-L1, GEO, HEO and SSO orbit types have almost similar and highest success rate while GTO orbit type has the lowest success rate.
- A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.

EDA with Data Visualization

- **Line Graph:**
- **Success Rate Vs Year**
- We observe that the success rate kept increasing from 2013 till 2020
- Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded
- **Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose**
- https://gist.github.com/georgebhilai/9b4ac86a64eac1d38de5ba1e8e2c585b

EDA with SQL

Using bullet point format, summarize the SQL queries you performed

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the first successful landing outcome in drone ship was acheived.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- Listing the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

EDA with SQL

Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

https://github.com/georgebhilai/Data-Science-and-Machine-Learning-Capstone-Project/blob/bdf6cfc78d969a4da3f66d136e3064b252fa8a8b/jupyter-labs_drone%20ship%20landing%20variation-eda-sql-edx.ipynb

Build an Interactive Map with Folium

- **Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map**

- >> Launch success may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and some of the factors can be determined by analysing the existing launch site locations (some geographical patterns about launch sites can be found).
- **Markers created**
- >> Markers were created for all launch sites location on the map using site's latitude and longitude co-ordinates
- >> Markers were created for each success / failed launches for each site on the map and were coloured as red and green
- **Circles created**
- >> Using folium.Circle function, highlighted circle areas with text label was created in specific coordinates
- >> Circles were created for each launch site on the site map

Build an Interactive Map with Folium (contd/-)

- **Lines created**
- >> for calculating the distances between a launch site to its proximities viz. railway, highway, coastline and city
- **Other pointers created**
- >> MousePosition on the map to get coordinate for a mouse over a point on the map
- **Explain why you added those objects**
- >> To explore, analyse and understand the proximity of the launch sites to the coast, cities, railway, highway
- >> To understand which sites have high success rates
- >> From the colour labeled markets in market clusters, it is easy to identify which launch sites have relatively high success rates
- >> MousePoint added to help easy navigation and conveniently determining the coordinates of the different locations viz. launch sites, railway, coastline, highway etc.

Build an Interactive Map with Folium (contd/-)

Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

- https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/13ac1bc3-dd38-414d-9ec9-b842a1f7cb27/view?access_token=1c82015285ccfbd0921445044628e79c1016038eb5ec904aa1839a88d0c08a64
- https://gist.github.com/georgebhilai/561c0789843ea0526eb3aa41a39778bd#file-interactive-visual-analytics-with-folium-lab-ipynb
- https://nbviewer.org/github/georgebhilai/Data-Science-and-Machine-Learning-Capstone-Project/blob/d559c00a0490b2c5f66f57dff85087def0d76619/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite_trusted.ipynb

Build a Dashboard with Plotly Dash

- **Summarize what plots/graphs and interactions you have added to a dashboard**

>> Pie chart graph to show the total success launches for all sites

>> If a specific launch site is selected, a pie chart graph to show the success launches and failed launches for the selected site.

>> Scatter plot with the x axis to be the payload and the y axis to be the launch outcome to visually observe how payload may be correlated with mission outcomes for selected site(s).

>> Color-labeling the Booster version on each scatter point so that mission outcomes with different boosters can be observed

- **Explain why you added those plots and interactions**

>> Pie chart graph to show the success launches and failed launches for the selected site.

>> Scatter plot to visually observe how payload may be correlated with mission outcomes for selected site(s)

>> Color-labeling the Booster version on each scatter point so that mission outcomes with different boosters can be observed

- **Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose**
- **https://github.com/georgebhilai/Data-Science-and-Machine-Learning-Capstone-Project/blob/865eb3a1e858d168a3406b8f4c0f2f158ed81b03/Building_Dasboard_Application.py**

Predictive Analysis (Classification)

- **Summarize how you built, evaluated, improved, and found the best performing classification model**

- **BUILDING MODEL**
- Load our dataset into NumPy and Pandas
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.
- **EVALUATING MODEL**
- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix
- **IMPROVING MODEL**
- Feature Engineering
- Algorithm Tuning
- **FINDING THE BEST PERFORMING CLASSIFICATION MODEL**
- The model with the best accuracy score wins the best performing model
- In the notebook there is a dictionary of algorithms and scores are displayed at the end of the notebook

Predictive Analysis (Classification)

- You need present your model development process using key phrases and flowchart
- Individual slides added for key phrases and flowcharts

- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- https://gist.github.com/georgebhilai/02c03c57b50acc43d53ccdba3c9d5dc1

Presentation of Model Development Process using Key phrases

>> Import the libraries viz. Pandas, NumPy, Matplotlib, Seaborn, sklearn
>> Load the dataframe
>> Create a NumPy array
>> Standardise the data through Preprocessing using sklearn
>> Split the data into training and testing data using the function train_test_split
>> Models are trained and hyperparameters are selected using the function GridSearchCV

- Create a logistic regression object

>> Then create a GridSearchCV object logreg_cv with cv = 10
>> Fit the object to find the best parameters from the dictionary parameters
>> Output the GridSearchCV object for logistic regression
>> Display the best parameters using the data attribute best_params_ and the accuracy on the validation data using the data attribute best_score
>> Calculate the accuracy on the test data using the method score
>> Plot the confusion matrix

Presentation of Model Development Process using Key phrases contd/-

- Create a support vector machine object

>> Then create a GridSearchCV object svm_cv with cv = 10

>> Fit the object to find the best parameters from the dictionary parameters

>> Output the GridSearchCV object for support vector machine

>> Display the best parameters using the data attribute best_params_ and the accuracy on the validation data using the data attribute best_score

>> Calculate the accuracy on the test data using the method score

>> Plot the confusion matrix

Presentation of Model Development Process using Key phrases contd/-

❖Create a decision tree classifier object

>> Then create a GridSearchCV object decision tree _cv with cv = 10

>> Fit the object to find the best parameters from the dictionary parameters

>> Output the GridSearchCV object for decision tree classifier

>> Display the best parameters using the data attribute best_params_ and the accuracy on the validation data using the data attribute best_score

>> Calculate the accuracy on the test data using the method score

>> Plot the confusion matrix

Presentation of Model Development Process using Key phrases contd/-

❖Create a k nearest neighbors object

>> Then create a GridSearchCV object knn_cv with cv = 10

>> Fit the object to find the best parameters from the dictionary parameters

>> Output the GridSearchCV object for decision tree classifier

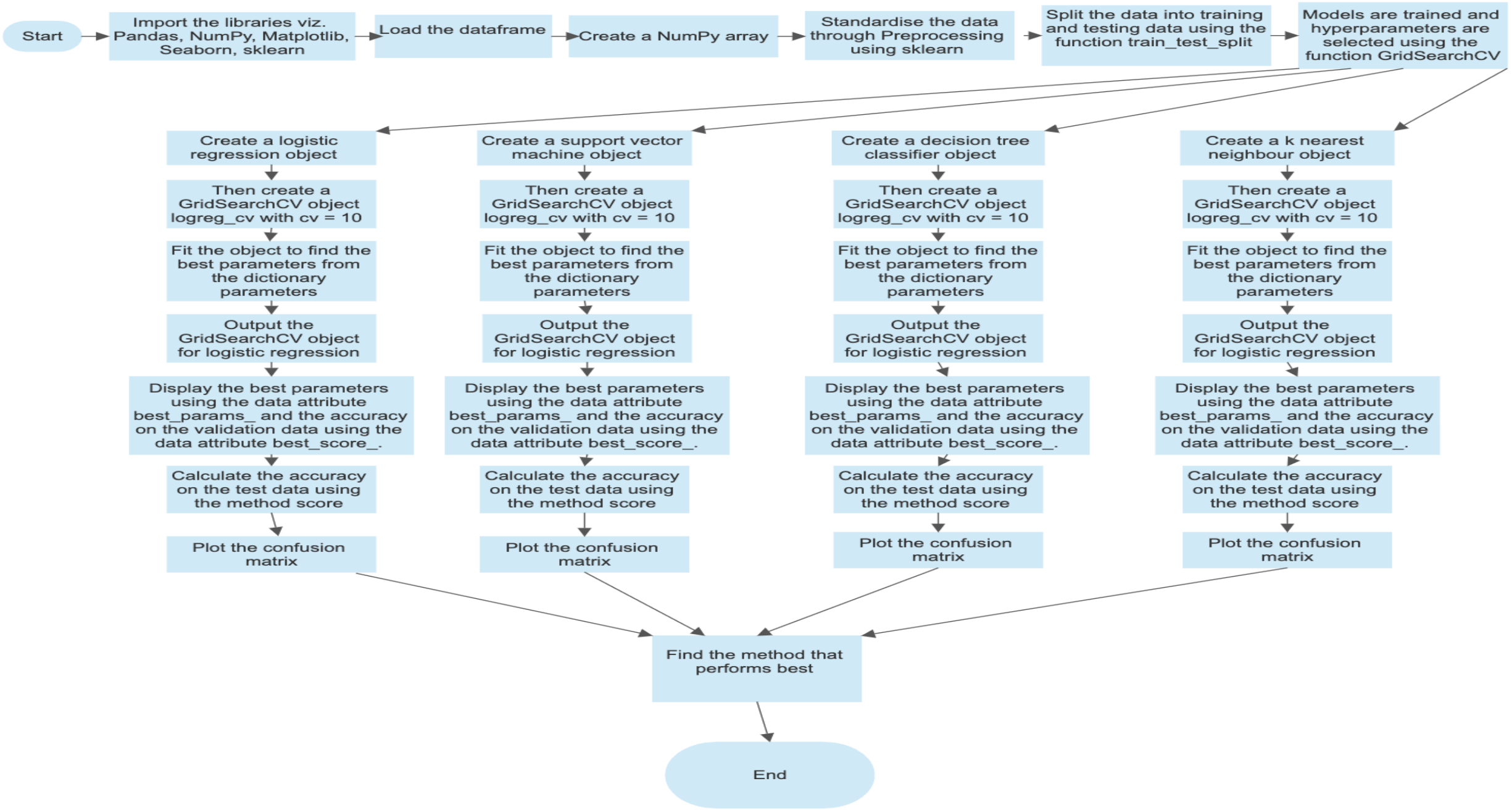
>> Display the best parameters using the data attribute best_params_ and the accuracy on the validation data using the data attribute best_score

>> Calculate the accuracy on the test data using the method score

>> Plot the confusion matrix

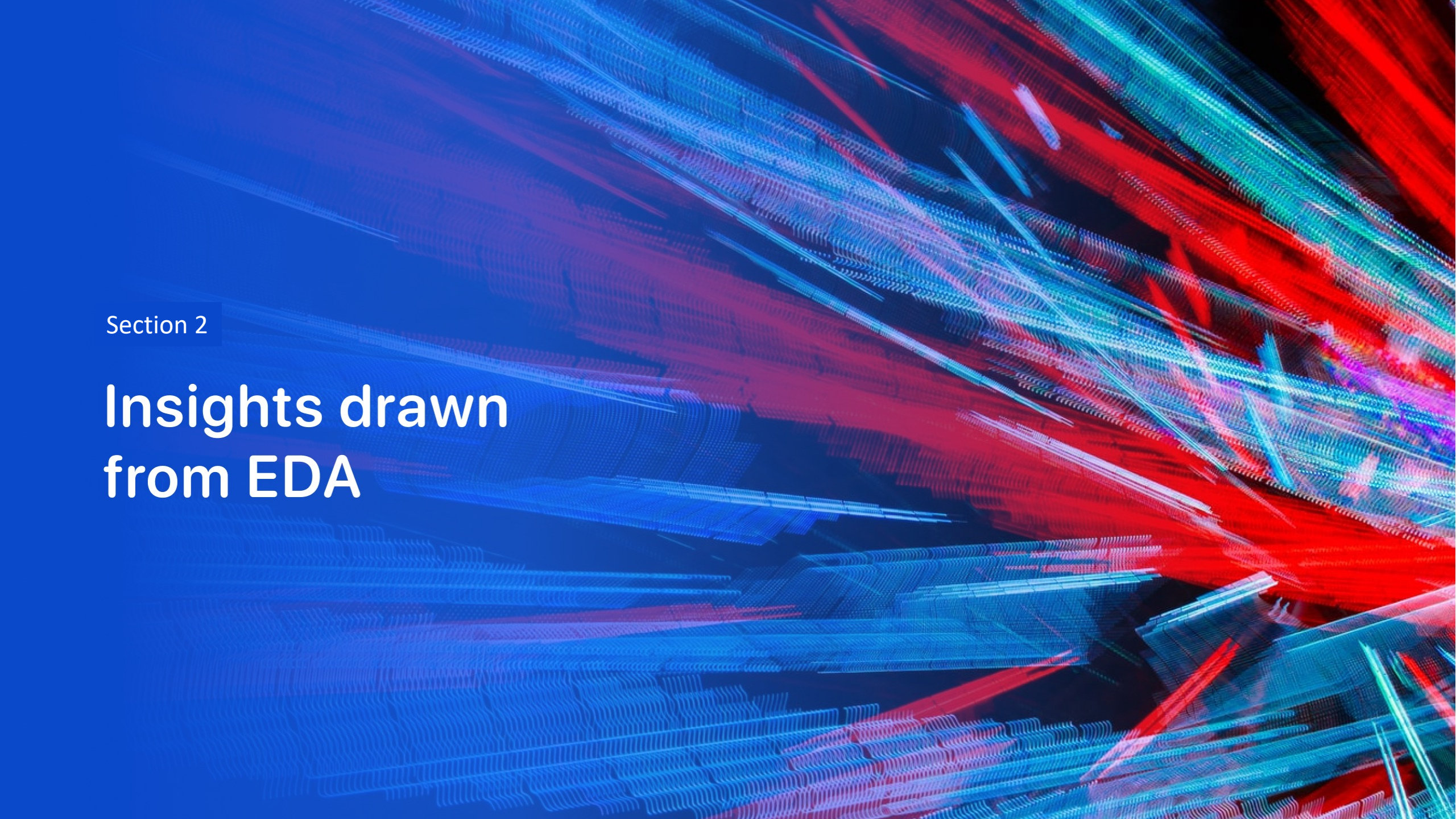
❖Find the method that performs best

Presentation of Model Development Process using Flowchart



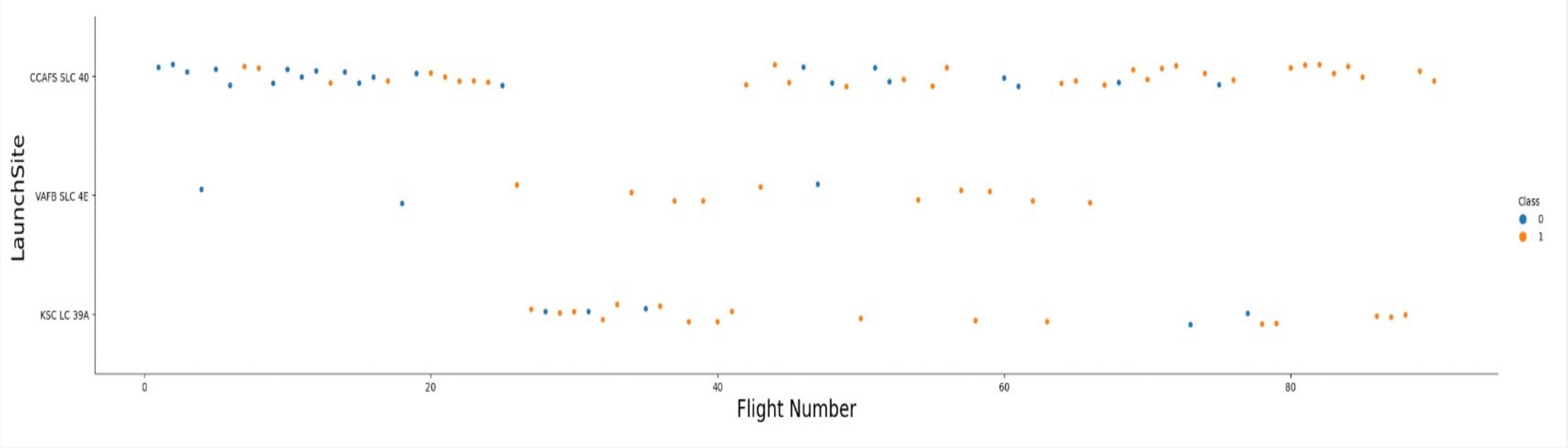
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

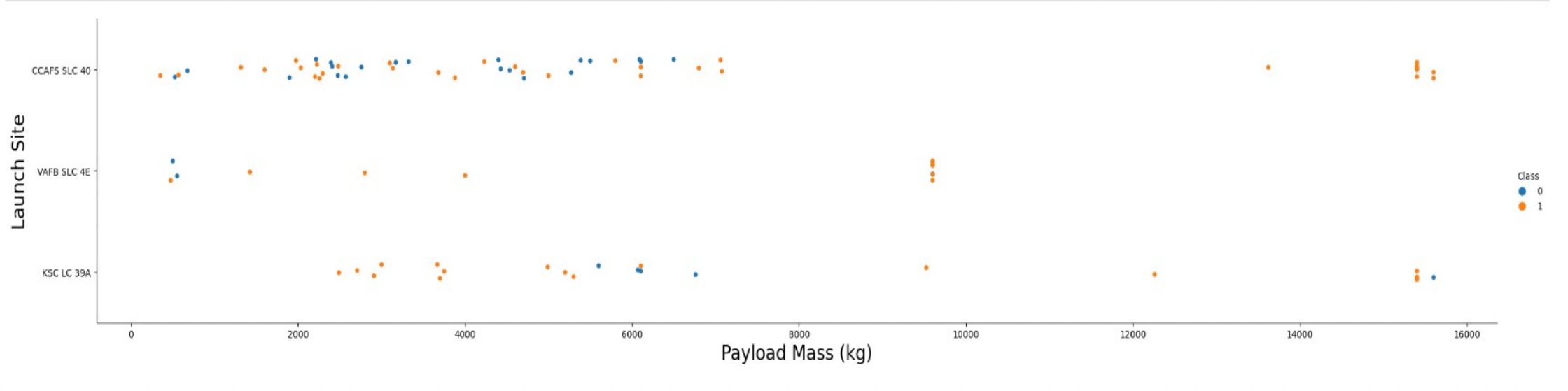
- **Show a scatter plot of Flight Number vs. Launch Site**



- **Show the screenshot of the scatter plot with explanations**
- **EXPLANATION**
- >> The scatter plot gives a visual presentation of the success rates of different launch sites. We find that different launch sites have different success rates. CCAFS LC-40 has a success rate of 60% while KSC LC-39A and VAFB SLC 4E has a success rate of 77%. As the number of flight increases the absolute number of successful launches also increases.

Payload vs. Launch Site

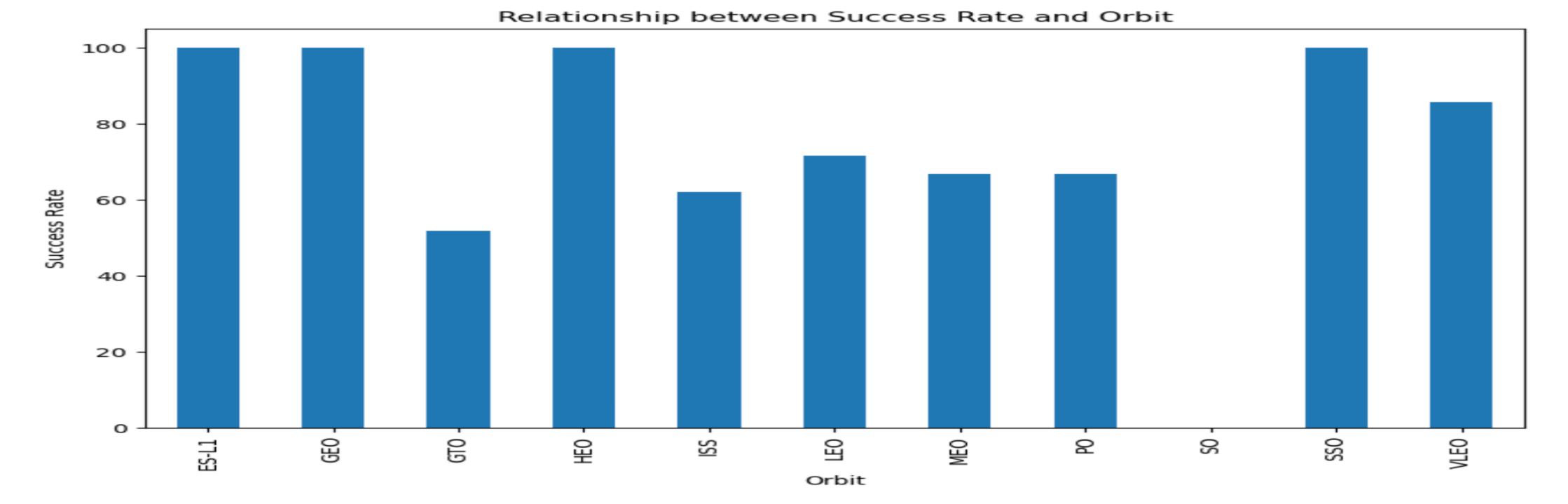
- **Show a scatter plot of Payload vs. Launch Site**



- **Show the screenshot of the scatter plot with explanations**
- **EXPLANATION**
- >> For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000). In respect of the other two launch sites, while in the case of CCAFS LC-40 launch site there is concentration of launches with comparatively lower payload in case of KSC LC-39A there is almost uniform distribution of the payload-wise launches. Selection of launch sites has a relation to the Payload mass.

Success Rate vs. Orbit Type

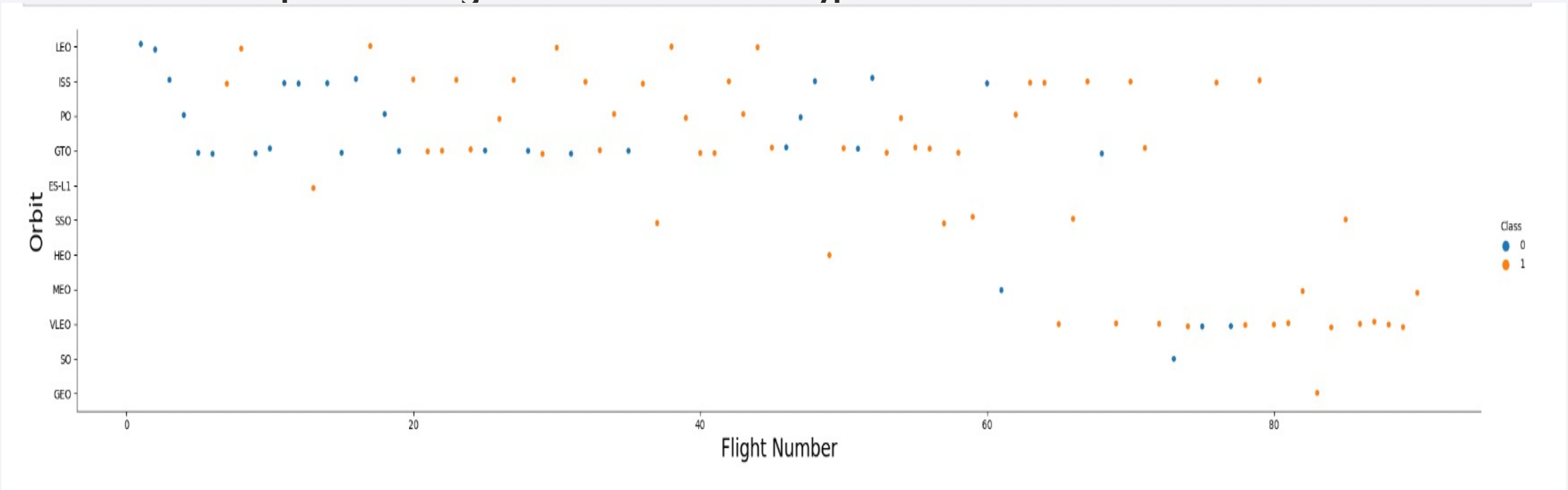
- Show a bar chart for the success rate of each orbit type



- Show the screenshot of the scatter plot with explanations
- **EXPLANATION**
- The Bar graph shows that ES-L1, GEO, HEO and SSO orbit types have almost similar and highest success rate while GTO orbit type has the lowest success rate.

Flight Number vs. Orbit Type

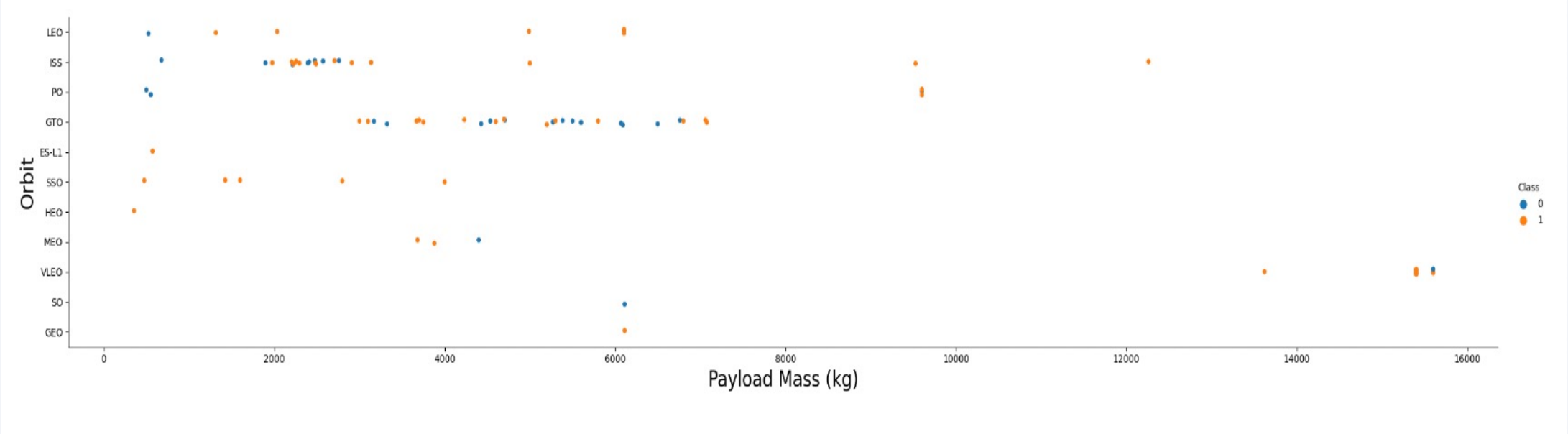
- Show a scatter point of Flight number vs. Orbit type



- Show the screenshot of the scatter plot with explanations
- **EXPLANATION**
- The scatter plot chart shows that in the LEO orbit the Success is related to the number of flights, on the other hand, there seems to be no relationship between flight number in GTO orbit.

Payload vs. Orbit Type

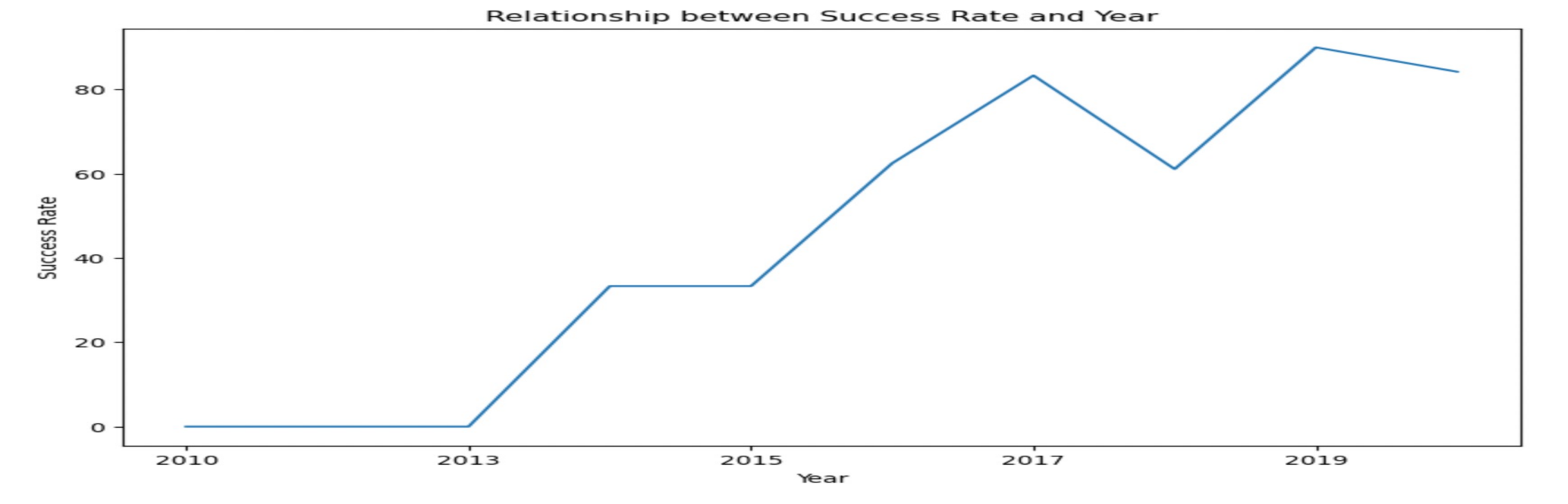
- **Show a scatter point of payload vs. orbit type**



- **Show the screenshot of the scatter plot with explanations**
- **EXPLANATION**
- >> 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

- Show a line chart of yearly average success rate



- Show the screenshot of the scatter plot with explanations
- **EXPLANATION**
- The success rate kept increasing from 2013 till 2020 based on the learnings and improvements.

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

Task 1

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

In [7]:
```
%sql select Unique(LAUNCH_SITE) from SPACEXTBL;
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[7]:

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

- Using the Unique keyword in the query with select statement ensures selection of only unique records from the LAUNCH_SITE column from the table SPACEXTBL.

Launch Site Names Begin with 'KSC'

- **Find 5 records where launch sites' names start with `KSC`**
- Present your query result with a short explanation here

Task 2

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

In [13]:
```
%sql SELECT * from SPACEXTBL where (LAUNCH_SITE) LIKE 'KSC%' LIMIT 5;
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[13]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- Using the LIMIT clause in the SQL restricts how many rows are returned from a query. Here the LIMIT 5 represents that we want to retrieve five records from SPACEXTBL and LIKE operator has a wild card with the string KSC, the percent in the end suggests that the LAUNCH_SITE name must start with KSC.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

Task 3

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

In [14]:
```
%sql SELECT sum(payload_mass__kg_) as sum_payload from SPACEXTBL where (customer) = 'NASA (CRS)'
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[14]:

sum_payload
45596

- **QUERY EXPLANATION**
- Using the function sum summates the total in the column payload_mass_kg_
- The where clause filters the dataset to only perform calculations on customer, NASA (CRS)

Average Payload Mass by F9 v1.1

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

Task 4

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

In [15]:
```
%sql SELECT avg(payload_mass__kg_) as average_payload from SPACEXTBL where (booster_version) = 'F9 v1.1'
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[15]:

average_payload
2928

- Present your query result with a short explanation here
- **QUERY EXPLANATION**
- Using the function avg works out the average in the column payload_mass_kg_
- The where clause filters the dataset to only perform calculations on booster_version, F9 v1.1

First Successful Ground Landing Date

- **Find the dates of the first successful landing outcome on drone ship.**

Task 5

List the date where the first successful landing outcome in drone ship was acheived.

Hint:Use min function

In [16]:
```
%sql SELECT min(date) from SPACEXTBL where landing__outcome = 'Success (drone ship)'
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[16]:

1
2016-04-08

- **Present your query result with a short explanation here**
- **QUERY EXPLANATION**
- Using the function min works out the minimum date in the column date
- The where clause filters the dataset to only perform calculations on landing_outcome, Success (drone ship)

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

```
[13]: %sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

[13]:

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

- Present your query result with a short explanation here
- **QUERY EXPLANATION**
- For getting the data of BOOSTER_VERSION we used select query with where clause and operator AND for ensuring conditions LANDING_OUTCOME = Success (ground pad) and PAYLOAD_MASS_KG_ BETWEEN 4001 and 5999 being fulfilled.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Task 7

List the total number of successful and failure mission outcomes

In [18]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[18]:

mission_outcome	outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- **QUERY EXPLANATION**
- Using the function COUNT for column MISSION_OUTCOME and GROUP BY clause in the statement gives the desired result.

Boosters Carried Maximum Payload

- List the names of the boosters 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

In [21]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[21]:

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

- **QUERY EXPLANATION**
- Selecting the BOOSTER_VERSION from the dataset using the where clause for the function max in sub query to select maximum from the PAYLOAD_MASS_KG_ column of the dataset gives the desired results.

2015 Launch Records

- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

Task 9

List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

In [41]:
```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)' AND YEAR(I
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[41]:

DATE	booster_version	launch_site	landing__outcome
2017-02-19	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
2017-05-01	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
2017-06-03	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
2017-08-14	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
2017-09-07	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
2017-12-15	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

- For getting the data of DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME we used select query with where clause and operator AND for ensuring conditions LANDING_OUTCOME = Success (ground pad) and YEAR(DATE) = 2017 being fulfilled

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

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

Task 10

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

In [42]:
```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_
```

```
 * ibm_db_sa://wjn90813:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90108kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[42]:

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

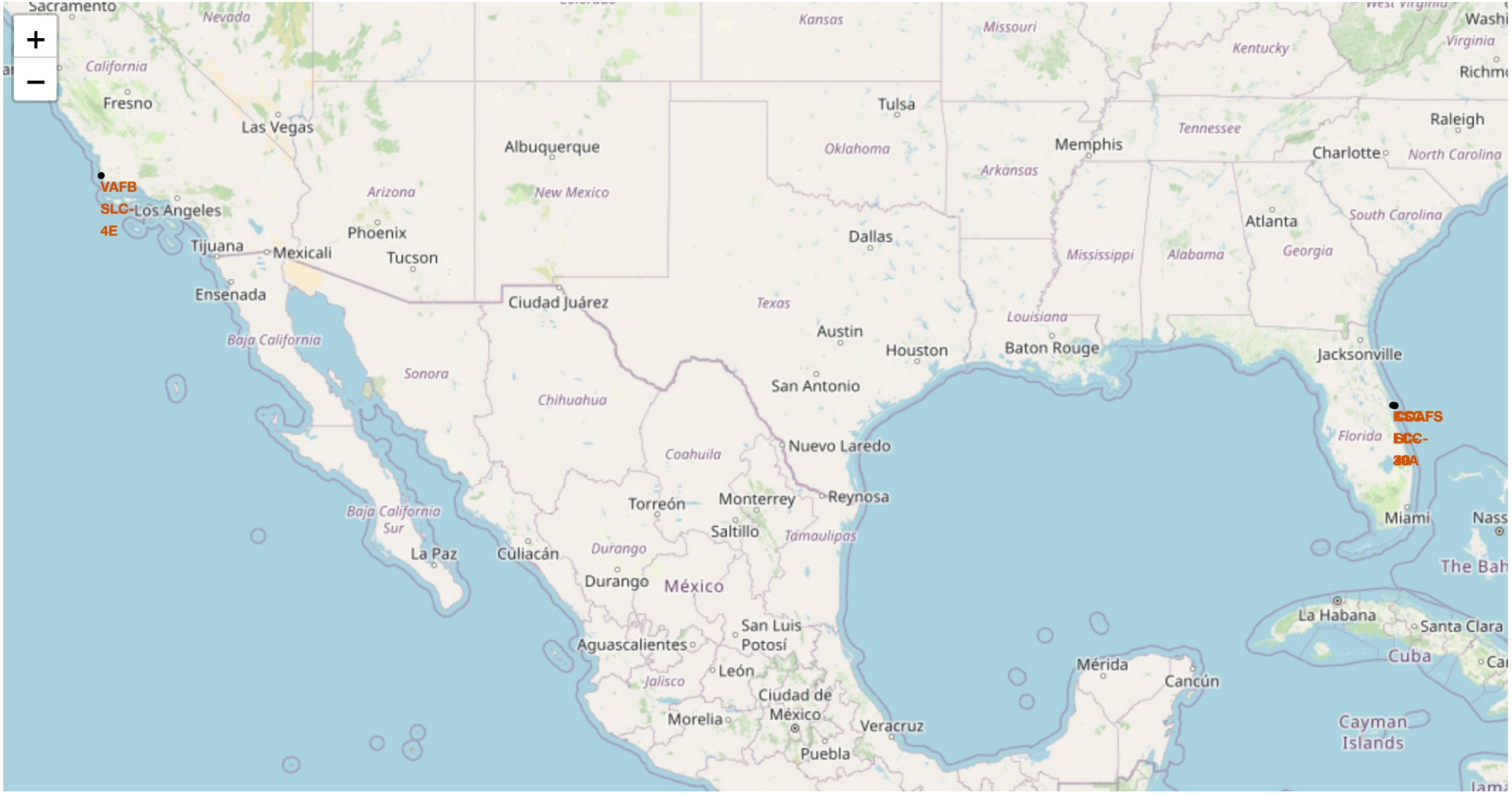
- **QUERY EXPLANATION**

- For getting the data of landing_outcome, count_launches we used select query with count function and where, group by and order by clause. We used operator AND for ensuring conditions DATE BETWEEN '2010-06-04' **AND** '2017-03-20' being fulfilled. To sort the records in descending order (order by clause), we used the DESC keyword


Section 3
Launch Sites
Proximities Analysis

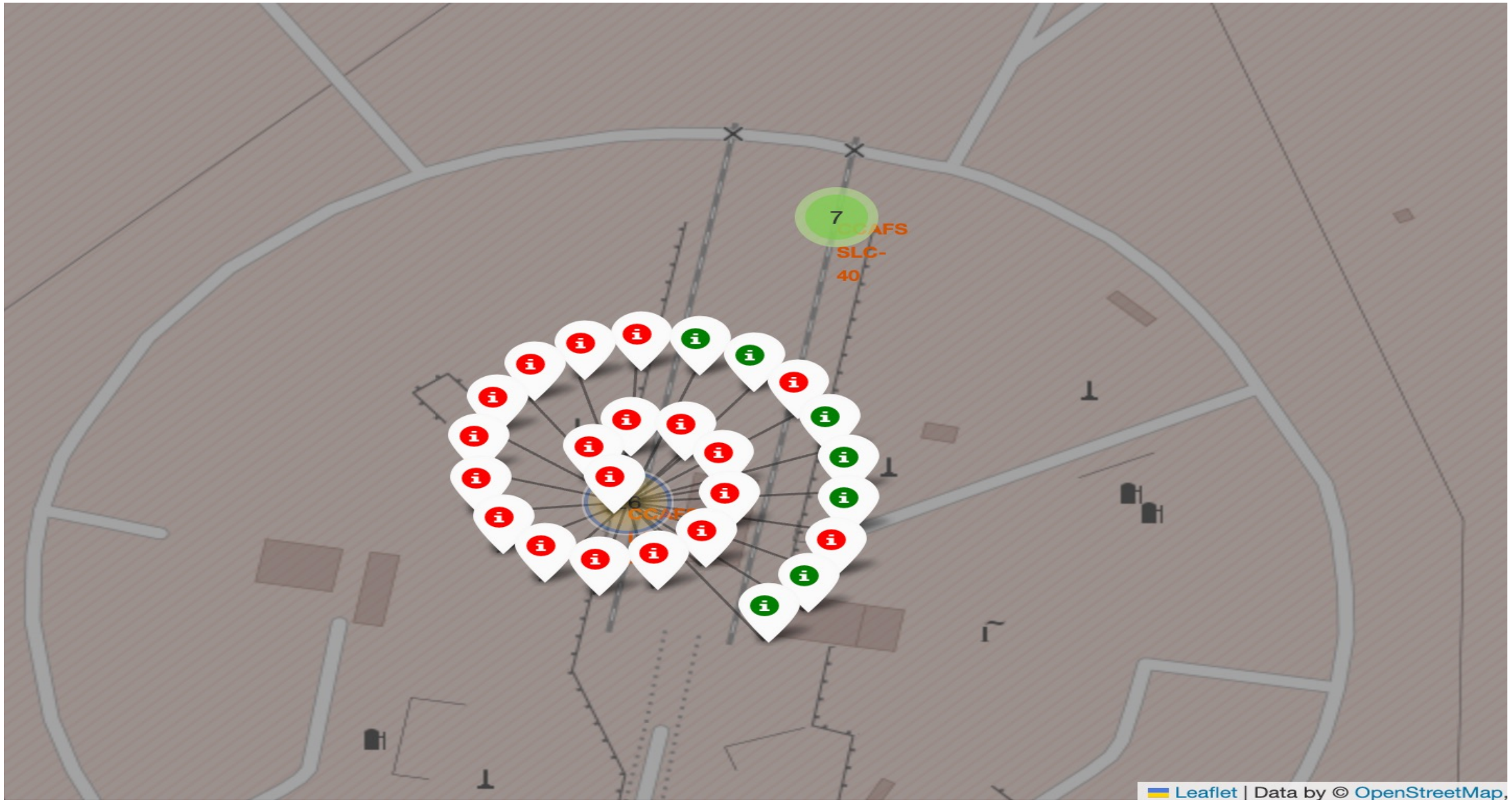
Screenshot of Folium Map with Marked Launch sites

- **Replace <Folium map screenshot 1> title with an appropriate title**
- Appropriate Title given ‘Screenshot of Folium Map with Marked Launch sites’
- **Explore the generated folium map and make a proper screenshot to include all launch sites’ location markers on a global map**
- Screenshot is pasted in the following slide
- **Explain the important elements and findings on the screenshot**
- All the four launch sites viz., CC AFS LC-40, CC AFS SLC-40, KSC LC-39A and VAFB SLC-4E are marked and labelled in the map.


VAFB
SLC-4E
KSC
CCAFS
SLC-40
Sacramento
Nevada
California
Fresno
Las Vegas
Albuquerque
Kansas
Missouri
Kentucky
Virginia
Tulsa
Oklahoma
Arkansas
Memphis
Tennessee
Raleigh
Charlotte
North Carolina
Arizona
New Mexico
Los Angeles
Phoenix
Tucson
Tijuana
Mexicali
Ensenada
Dallas
Mississippi
Alabama
Georgia
Atlanta
South Carolina
Ciudad Juárez
Texas
Austin
Houston
Louisiana
Baton Rouge
Jacksonville
Baja California
Sonora
Chihuahua
San Antonio
Florida
Coahuila
Nuevo Laredo
Torreón
Monterrey
Reynosa
Saltillo
Tamaulipas
Miami
Baja California Sur
La Paz
Culiacán
Durango
México
The Bah
La Habana
Santa Clara
Cuba
San Luis Potosí
Aguascalientes
León
Jalisco
Mérida
Cancún
Ciudad de México
Morelia
Veracruz
Puebla
Cayman Islands

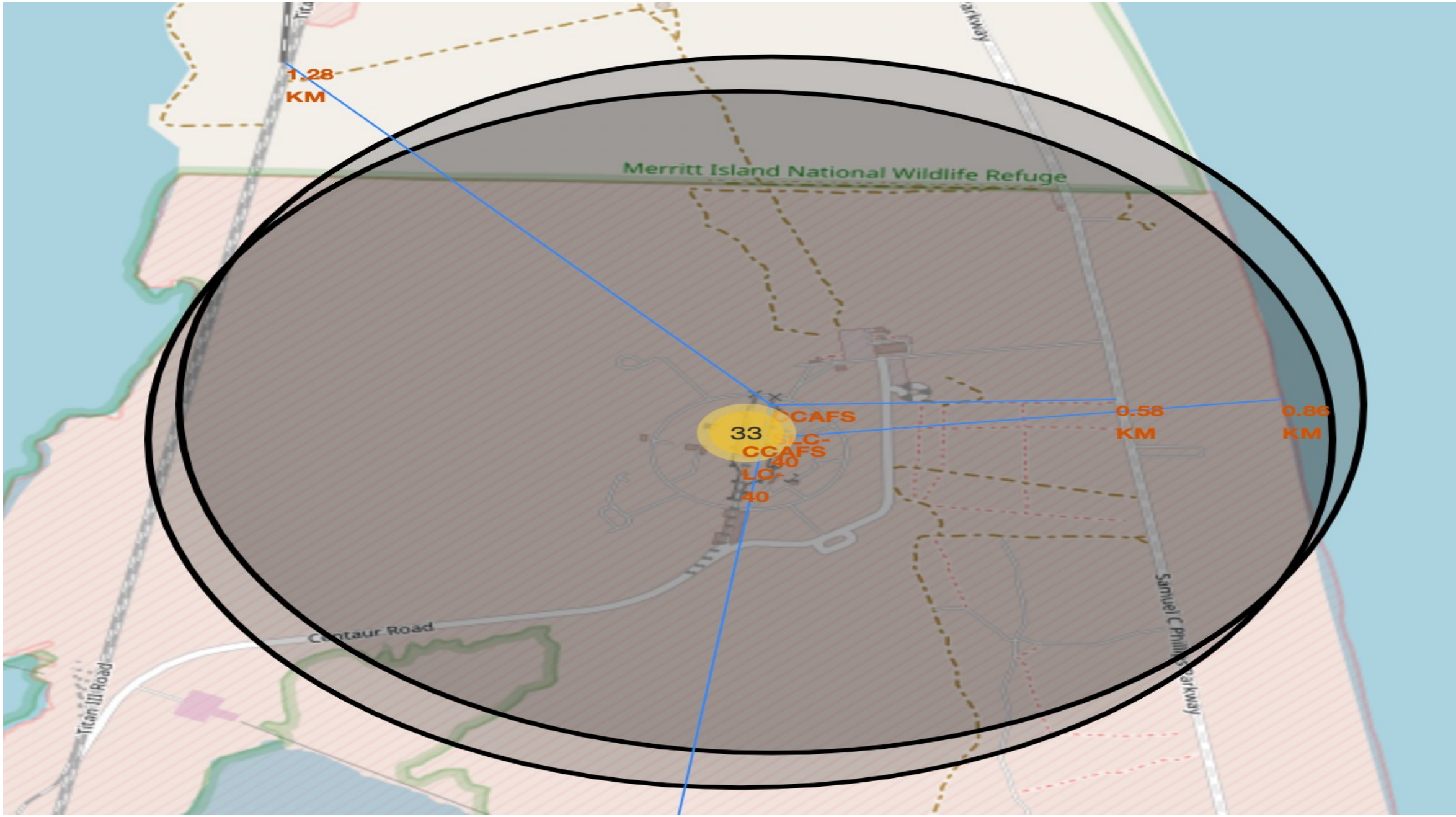
Screenshot of Folium Map with Color Labeled Launch Outcomes

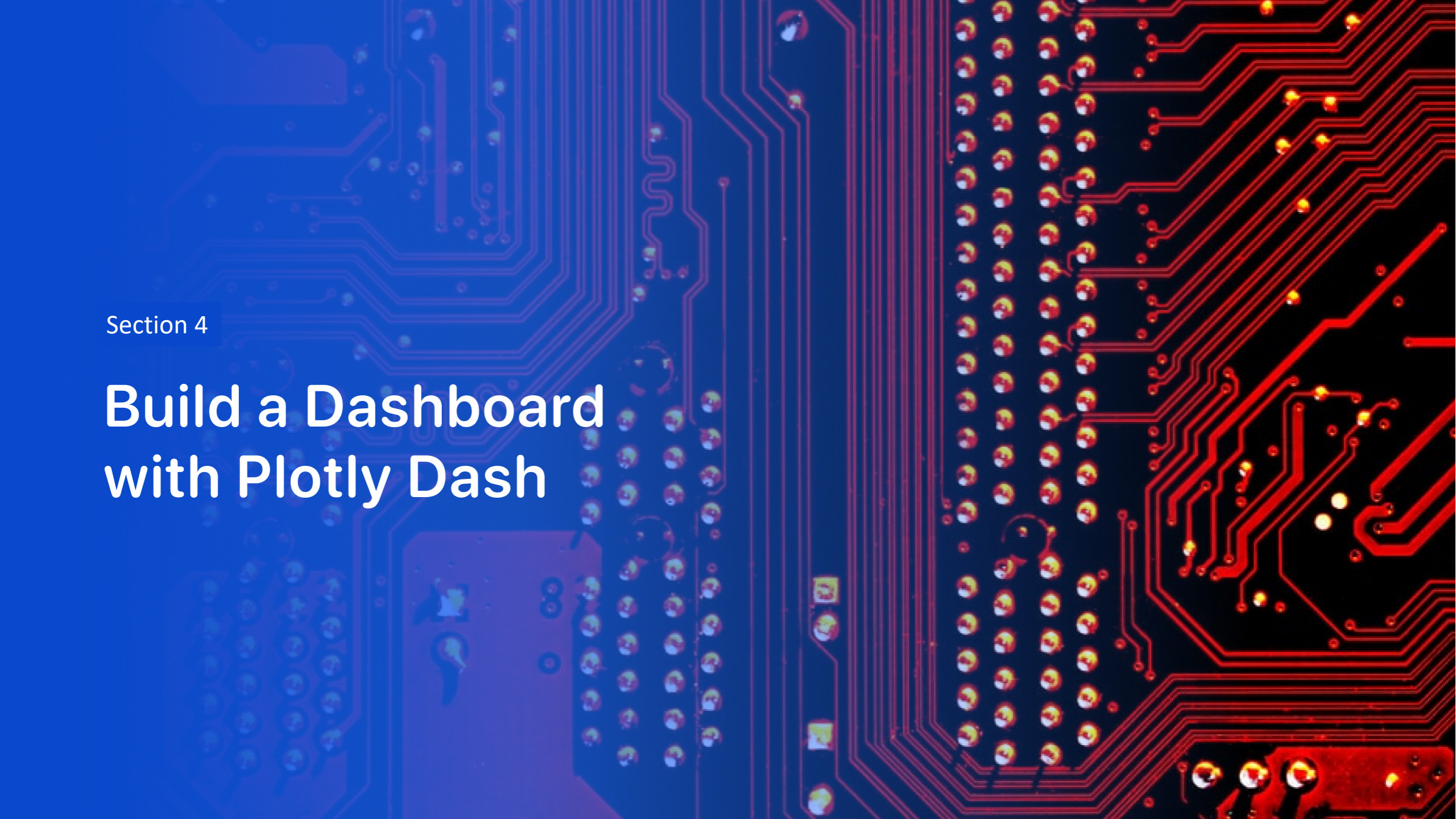
- **Replace <Folium map screenshot 2> title with an appropriate title**
- Appropriate Title given ‘Screenshot of Folium Map with Color Labeled Launch Outcomes’
- **Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map**
- Screenshot is pasted in the following slide
- **Explain the important elements and findings on the screenshot**
- The launch outcomes of launches from the launch site ‘CC AFS SLC-40‘ is displayed in the screen shot and it shows 7 successful launches and 19 unsuccessful launches.


7
CCAFS
SLC-
40
Leaflet | Data by © OpenStreetMap,

Screenshot of Folium Map showing proximities such as railway, highway and coastline with distance displayed

- **Replace <Folium map screenshot 3> title with an appropriate title**
- Appropriate Title given ‘Screenshot of Folium Map showing proximities such as railway, highway and coastline with distance displayed’
- **Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed**
- Screenshot is pasted in the following slide
- **Explain the important elements and findings on the screenshot**
- We find that the launch site has proximity to the railway, highway and coastline with distance of 1.28 kms 0.58 kms and 0.86 kms respectively.


1.28
KM
Merritt Island National Wildlife Refuge
33
CCAFS
LC-
CCAFS
40
LC-
40
0.58
KM
0.86
KM
Centaur Road
Titan III Road
Samuel C Phillips Parkway


Section 4
Build a Dashboard with Plotly Dash

Screenshot of Interactive Dashboard of launch success for all sites

- **Replace <Dashboard screenshot 1> title with an appropriate title**
- Appropriate Title given ‘Screenshot of Interactive Dashboard of launch success for all sites’
- **Show the screenshot of launch success count for all sites, in a piechart**
- Screenshot is pasted in the following slide
- **Explain the important elements and findings on the screenshot**
- The different percentages of launch success for different sites viz. KSC LC-39A, CC AFS LC-40, VAFB SLC-4E and CC AFS SLC-40 are 41.7%, 29.2%, 16.7% and 12.5% respectively. The findings from the screen shot are that the highest launch success happened for KSC LC-39A site.

SpaceX Launch Records Dashboard

ALL SITES

Total Launches for All Sites

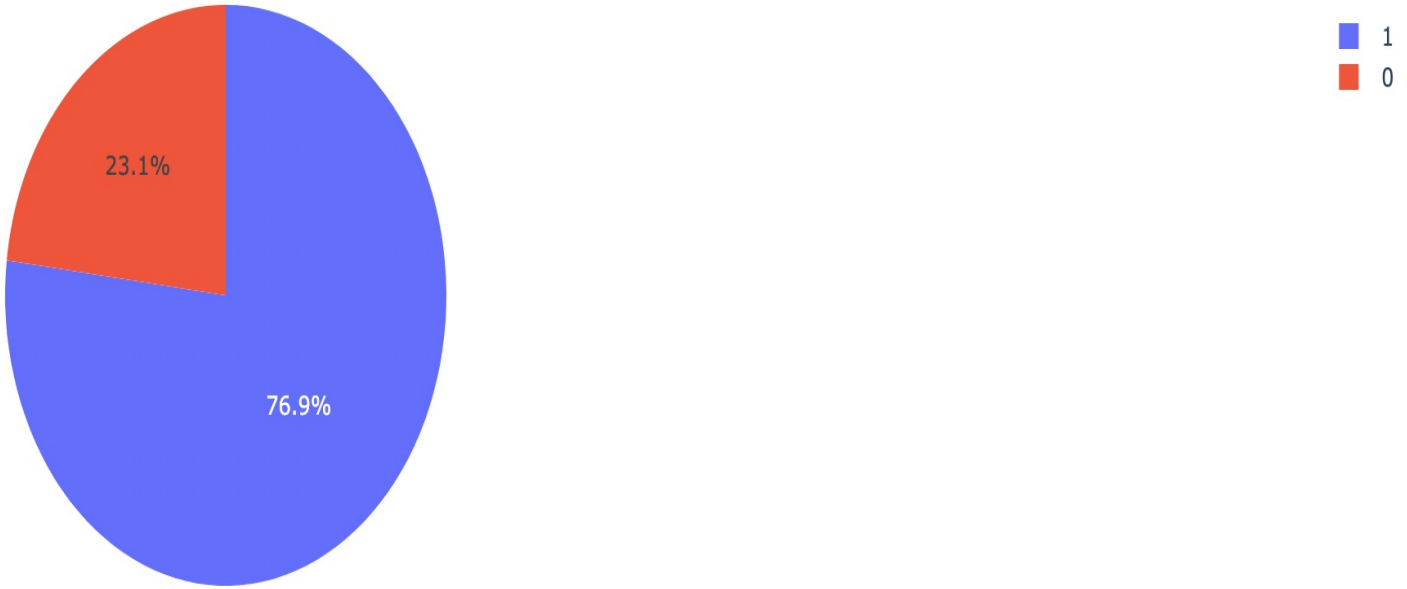
Screenshot of piechart for launch site with highest launch success ratio

- **Replace <Dashboard screenshot 2> title with an appropriate title**
- Appropriate Title given ‘Screenshot of piechart for launch site with highest launch success ratio’
- **Show the screenshot of the piechart for the launch site with highest launch success ratio**
- Screenshot is pasted in the following slide
- **Explain the important elements and findings on the screenshot**
- The launch site with highest launch success ratio is KSC LC-39A with the successful launches being 76.9% and the unsuccessful launches being 23.1%.

SpaceX Launch Records Dashboard

KSC LC-39A

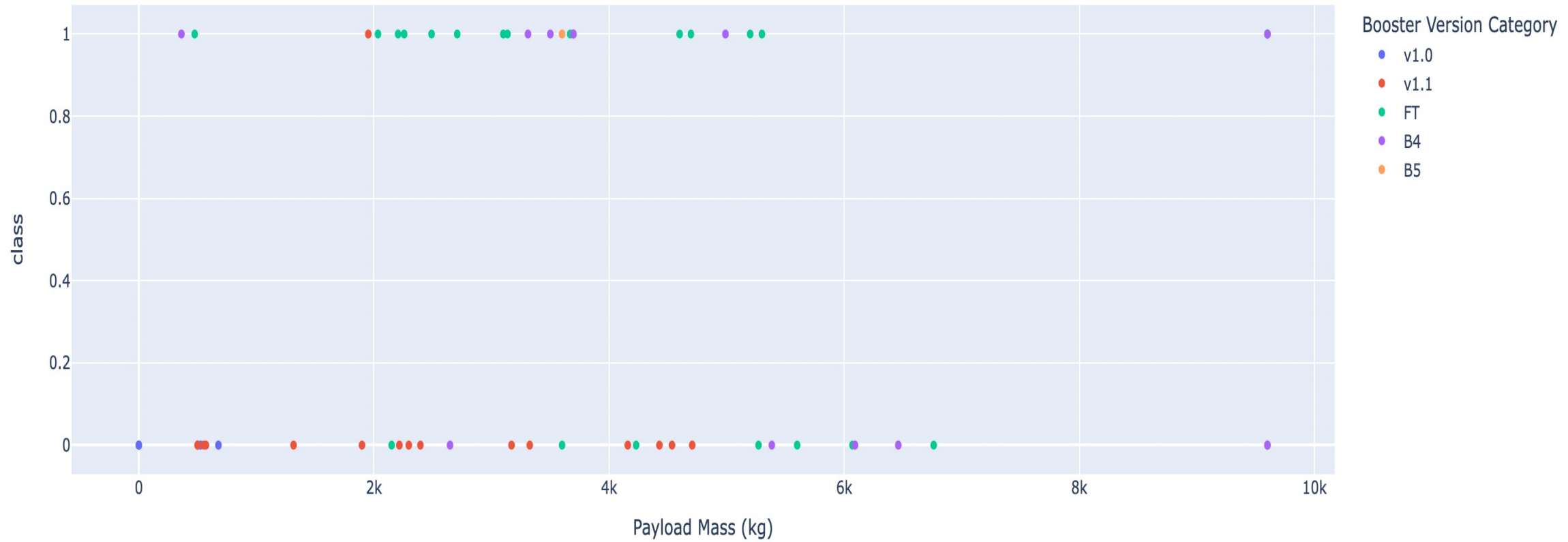
Total Launch for a Specific Site



Screenshot of Payload vs. Launch Outcome scatter plot for all sites

- **Replace <Dashboard screenshot 3> title with an appropriate title**
- Appropriate Title given ‘Screenshot of Payload vs. Launch Outcome scatter plot for all sites’
- **Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider**
- Screenshot is pasted in the following slide
- **Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.**
- Our findings are that the payload range ‘2500 to 5000’ and booster version ‘FT’ has the largest success rate.

Payload range (Kg):

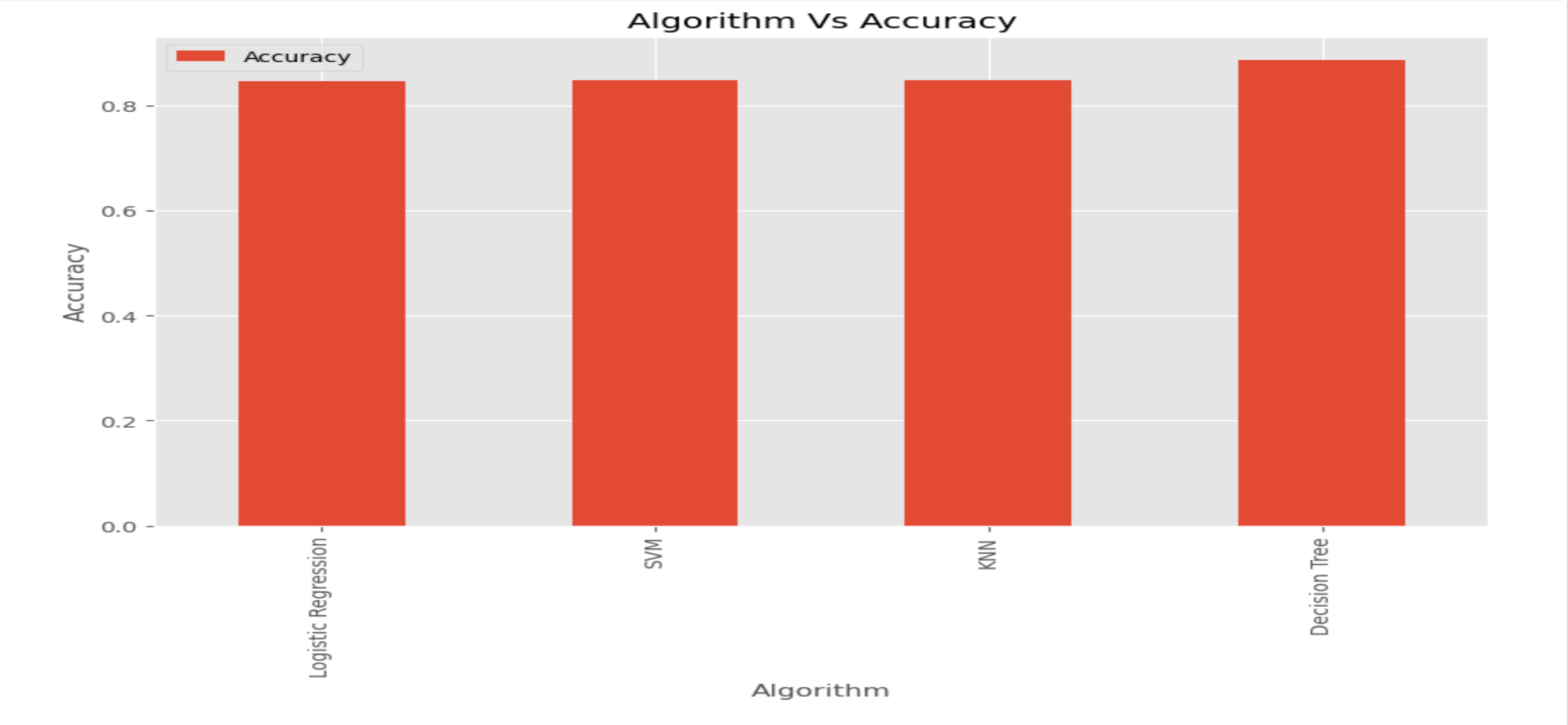


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- **Visualize the built model accuracy for all built classification models, in a bar chart**

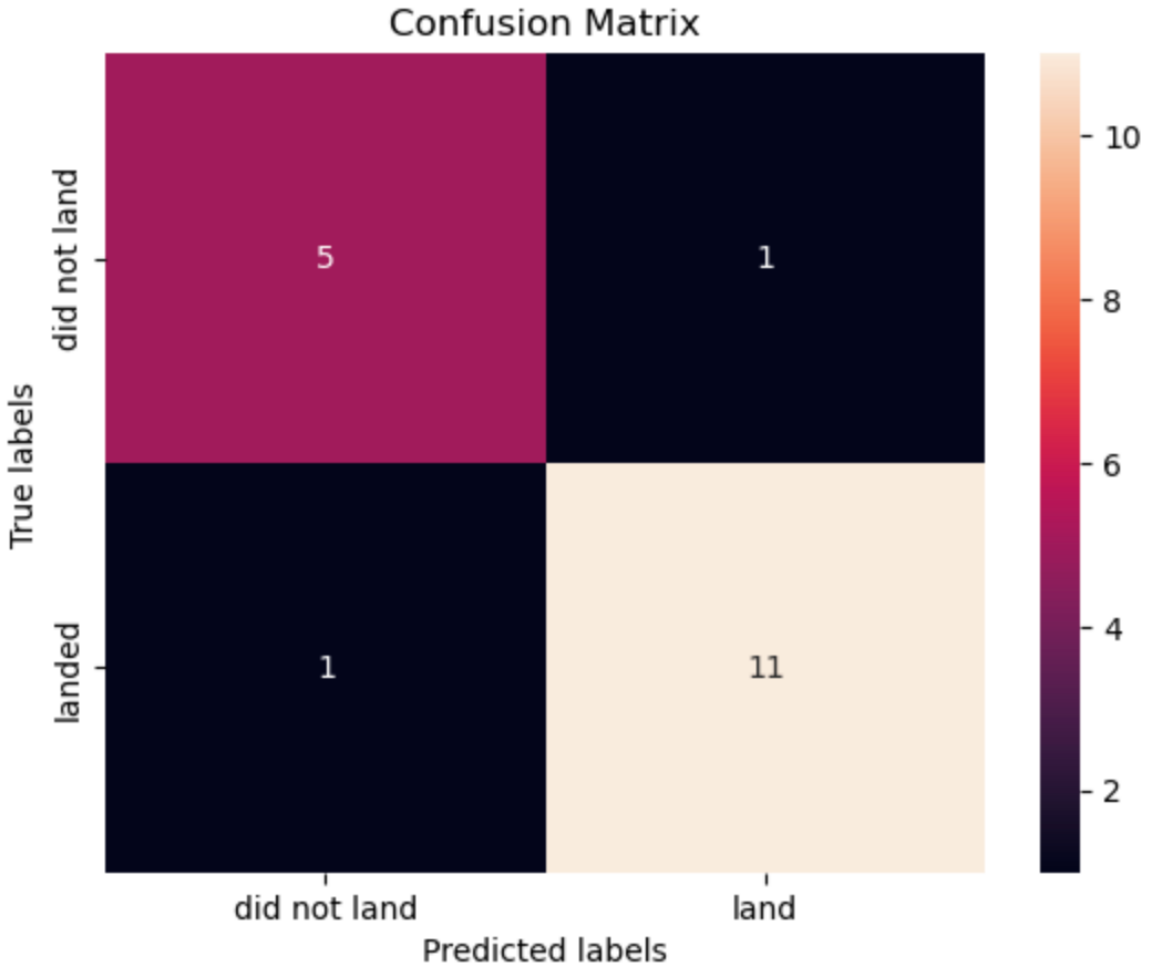


- **Find which model has the highest classification accuracy** : Decision Tree is the model that has the highest classification accuracy

Confusion Matrix

- **Show the confusion matrix of the best performing model with an explanation**
- Confusion matrix of the best performing model i.e. Decision Tree placed alongside.
- From the table below we can observe that the it is the best on account of highest accuracy of 0.8875.
- The false positives are low indicating a good performance of the ML model.

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



	Algorithm	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	KNN	0.848214
3	Decision Tree	0.887500

Conclusions

At the beginning of the project we defined the Problem for which we want to find answers. The problems are as under :

Problem no. 1

How the success rate of landing can help determine a realistic price of each launch?

Problem no. 2

Which all variables affect the success rate of landing and how can the learnings be used to improvise upon the launches for increasing the successful landings ?

Problem no. 3

Which is the best machine learning model that can be used based on the accuracy so as to ensure the best successful landing rate ?

Conclusions contd/-

Problem no. 1

Cost incurred on each launch can be determined by using the costing method. However, when it comes to pricing, the cost of first stage in cases of successful landing can be reduced from the overall cost. But the apportionment ratio will depend on the percentage of successful first stage landing.

- After using the public data and creating the dashboard, on an analysis of the data using the dashboard we can conclude that there are many instances where the first stage will land successfully. The successful landing itself means that SpaceX will reuse the first stage.
- Most of the launch sites are near the equator for an additional natural boost due to the rotational speed of earth which helps save the cost of putting in extra fuel and boosters which will also help in determining the cost based on the launch site.
- All the launch sites are close to the coast. All launch sites have closer proximities to coastlines and farther away from populous cities. Reduction of risk factors thus help avoid costs to mitigate risks and cost of insurance which help in better forecasting / prediction of the price leading to fair price determination.
- The launch sites are strategically located near coastline, highways and railways for transportation of personnel and cargo. These factors contribute to the optimisation of cost and help determine the price of launch.

Conclusions contd/-

Problem no. 2

- There is a positive correlation between the flight number and success rate as the flight number increases, the first stage is more likely to land successfully.
- The Success Rate Vs Year variable shows that the success rate kept increasing from 2013 till 2020. A trend line can help to make predictions about the results of data not yet recorded.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate. ES-L1, GEO, HEO and SSO have a 100% success rate.
- Based on the relationship between Orbit types and success rate we find that GTO orbit type has the lowest success rate and the launch to this orbit needs to be avoided at all costs.
- Based on the relationship between payload-wise launch to different Orbit types and success rate we find that GTO orbit type has the lowest success rate.
- Similarly launch site and success rate also has a relationship. CCAFS LC-40 launch site has a success rate of 60% while KSC LC-39A and VAFB SLC 4E launch sites has a success rate of 77%. KSC LC-39A had the most successful launches of any sites. Indeed having a 100% success rate for launches with payload mass less than 5,500 kg.

Conclusions contd/-

Problem no. 2 contd/-

- Based on the relationship between payload mass and success rate we find that low weighted payloads perform better than the heavier payloads. Success rate can be linked to payload mass as the lighter payloads generally proved to be more successful than the heavier payloads, Or more massive the Payload, the less likely the first stage will return.
- The success rates for SpaceX launches is directly proportional time in years which means that the learnings from previous failures / success needs to be put into application to perfect the launches.
- Success is defined as a successful recovery process of Falcon9 booster. As more success occurred to rockets having booster version of FT, the F9 FT booster version needs to be used.
- As more success occurred to rockets having payload mass between 1950-5200 kg., while deciding about the payload mass it needs to be seen that it falls in the range of 2000 to 5000.
- In spite of variation in the payload mass in respect of the launches from KSC LC-39A launch site the success rate has been high at 77%, therefore, maximum launches needs to be made from this launch site.

Conclusions contd/-

- ***Problem no. 3***

- The best predictive model (machine learning model) used for this dataset is the Decision Tree Classifier as it had the highest accuracy with 89%. Using this model best prediction of the outcome of the landing can be made which will help in determining the parameters for each launch attempts. Predictability and certainty will help in proper appropriation of the cost of first stage to cost price of each launch and a competitive pricing can be arrived at.

- Finally, we can conclude that the problems that we had encountered at the beginning of the project have been answered satisfactorily.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

- GitHub URL of the CSV file containing the data used for the Capstone Project.
- https://github.com/georgebhilai/Data-Science-and-Machine-Learning-Capstone-Project/blob/20d65054e71644e29c49be0728ef1d411
- GitHub URL of the uploaded files and PPT presentation and PDF of the Capstone Project
- https://github.com/georgebhilai/Data-Science-and-Machine-Learning-Capstone-Project/upload/main

Thank you!