Cousera Applied Data Science Capstone Project

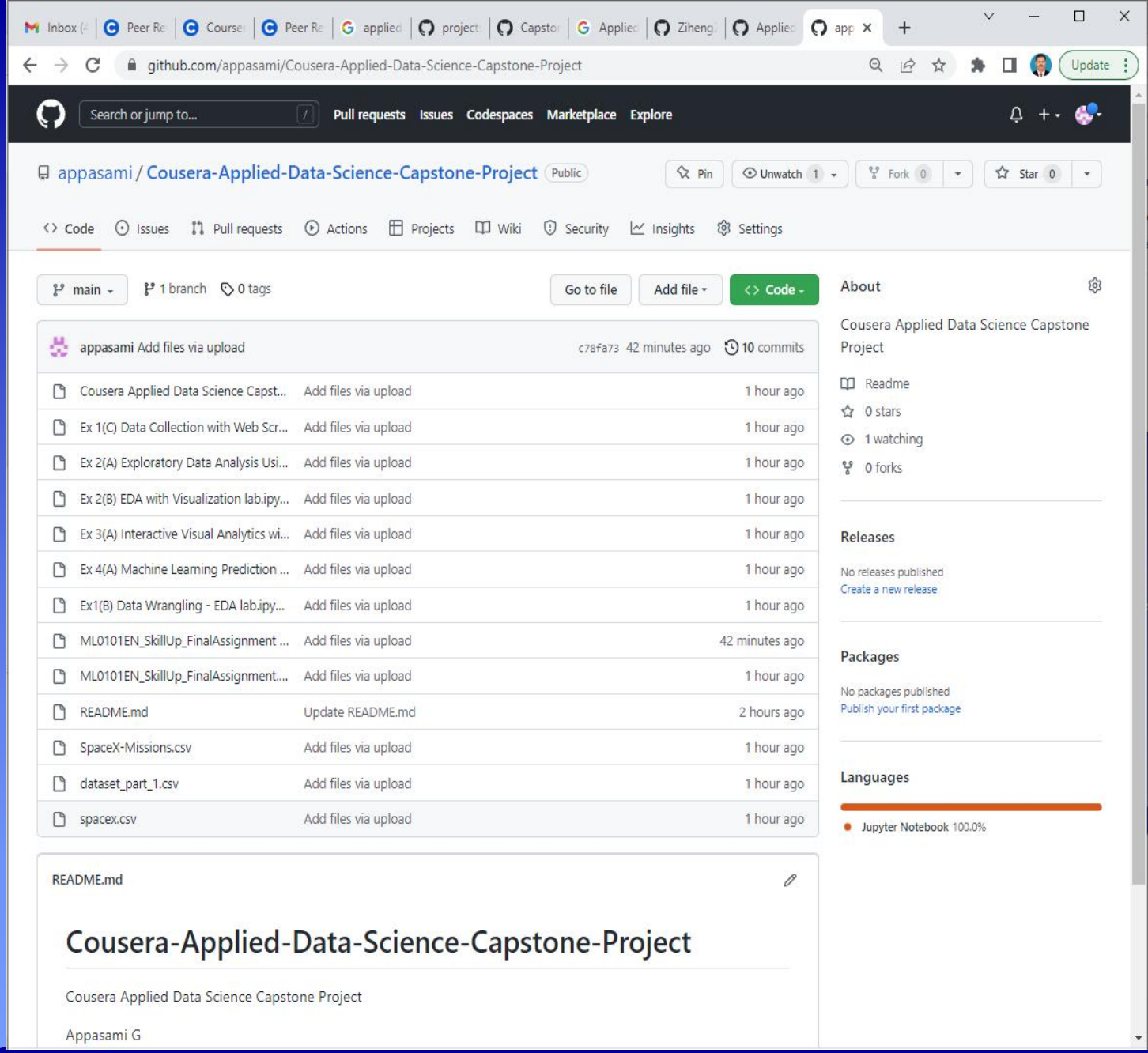
Appasami G
21/02/2023

Outline

1. Uploaded the URL of your GitHub repository including all the completed notebooks and Python files (1 pt)
2. Uploaded your completed presentation in PDF format (1 pt)
3. Completed the required Executive Summary slide (1 pt)
4. Completed the required Introduction slide (1 pt)
5. Completed the required data collection and data wrangling methodology related slides (1 pt)
6. Completed the required EDA and interactive visual analytics methodology related slides (3 pts)
7. Completed the required predictive analysis methodology related slides (1 pt)
8. Completed the required EDA with visualization results slides (6 pts)
9. Completed the required EDA with SQL results slides (10 pts)
10. Completed the required interactive map with Folium results slides (3 pts)
11. Completed the required Plotly Dash dashboard results slides (3 pts)
12. Completed the required predictive analysis (classification) results slides (6 pts)
13. Completed the required Conclusion slide (1 pts)
14. Applied your creativity to improve the presentation beyond the template (1 pts)
15. Displayed any innovative insights (1 pts)

Screen shots

1. GitHub repository URL



https://github.com/appasami/Cousera-Applied-Data-Science-Capstone-Project

Screen shots

2. Presentation in PDF format

Screen shots

3. Executive Summary

- **Summary of methodologies**
- **Summary of all results**
- **Important Discusions**

Screen shots

4. Introduction

- **Project background and context**
- **Problems you want to find answers**

Screen shots

5. Data collection and data wrangling methodology

Data Analysis

Load Space X dataset, from last section.

In [2]:
```
df=pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv")
df.head(10)
```

Out[2]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Lat
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.56
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.56
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.56
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.63
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.56
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.56
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1006	-80.577366	28.56
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1007	-80.577366	28.56
8	9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1008	-80.577366	28.56
9	10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1011	-80.577366	28.56

Identify and calculate the percentage of the missing values in each attribute

In [3]:
```
df.isnull().sum()/df.count()*100
```

Out[3]:
```
FlightNumber      0.000
Date              0.000
BoosterVersion    0.000
PayloadMass       0.000
Orbit             0.000
LaunchSite        0.000
Outcome           0.000
Flights           0.000
GridFins          0.000
Reused            0.000
Legs              0.000
```

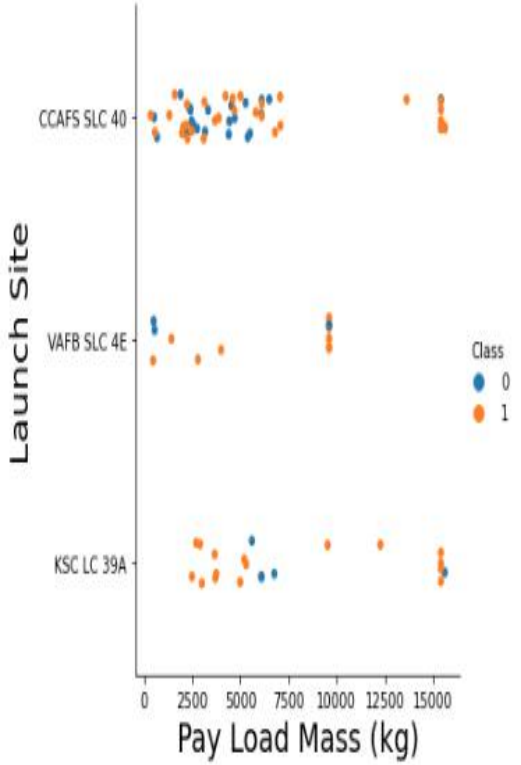
Screen shots

6. EDA and interactive visual analytics methodology

TASK: Visualize the relationship between Payload and Launch Site

We also want to observe if there is any relationship between launch sites and their payload mass.

In [5]:
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



Now try to explain any patterns you found in the Payload Vs. Launch Site scatter point chart.

We can infer that Launch Site-CCAFS SLC 40 is suitable for launching rockets with payload mass varying from low to very high. Launch Site VAFB SLC 4E is preferred for medium payload mass.

Screen shots

7 Predictive analysis methodology

Q9) Create and train a Decision Tree model called Tree using the training data (`x_train` , `y_train`).

In [ ]:
```
#Enter Your Code, Execute and take the Screenshot
```

In [20]:
```
Tree = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
Tree.fit(x_train, y_train)
```

Out[20]:
```
DecisionTreeClassifier(criterion='entropy', max_depth=4)
```

Q10) Now use the `predict` method on the testing data (`x_test`) and save it to the array `predictions` .

In [ ]:
```
#Enter Your Code, Execute and take the Screenshot
```

In [21]:
```
predictions = Tree.predict(x_test)
```

Q11) Using the `predictions` and the `y_test` dataframe calculate the value for each metric using the appropriate function.

In [ ]:
```
#Enter Your Code, Execute and take the Screenshot
```

In [22]:
```
Tree_Accuracy_Score = metrics.accuracy_score(y_test, predictions)
Tree_JaccardIndex = metrics.jaccard_score(y_test, predictions)
Tree_F1_Score = metrics.f1_score(y_test, predictions)
Tree_Log_Loss = metrics.log_loss(y_test, predictions)
print("Tree accur_acy score: ", Tree_Accuracy_Score)
print("Tree JaccardIndex : ", Tree_JaccardIndex)
print("Tree_F1_Score : ", Tree_F1_Score)
print("Tree Log Loss : ", Tree_Log_Loss)
```

```
Tree accur_acy score:  0.8183206106870229
Tree JaccardIndex :  0.48034934497816595
Tree_F1_Score :  0.6489675516224188
Tree Log Loss :  6.275038737219435
```

Logistic Regression

Q12) Use the `train_test_split` function to split the `features` and `Y` dataframes with a `test_size` of `0.2` and the `random_state` set to `1` .

In [ ]:
```
#Enter Your Code, Execute and take the Screenshot
```

In [24]:
```
x_train, x_test, y_train, y_test = train_test_split(features, Y, test_size = 0.2, random_state =1)
print ('Train set:', x_train.shape,  y_train.shape)
print ('Test set:', x_test.shape,  y_test.shape)
```

Screen shots

8. EDA with visualization

TASK: Visualize the relationship between success rate of each orbit type

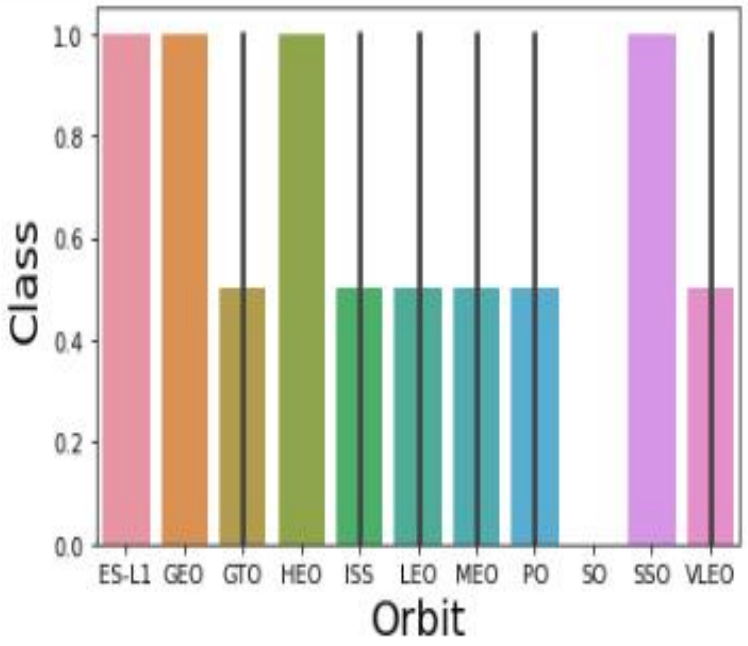
Next, we want to visually check if there are any relationship between success rate and orbit type.

Let's create a `bar chart` for the sucess rate of each orbit

In [15]:

```
# HINT use groupby method on Orbit column and get the mean of Class column
t = df.groupby(['Orbit', 'Class'])['Class'].agg(['mean']).reset_index()
sns.barplot(y="Class", x="Orbit", data=t)

plt.xlabel("Orbit",fontsize=20)
plt.ylabel("Class",fontsize=20)
plt.show()
```



Analyze the ploted bar chart try to find which orbits have high sucess rate.

Screen shots

9. EDA with SQL results

Task 1

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

In [39]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

```
 * ibm_db_sa://sdk38546:***@dashdb-txn-sbox-yp-lon02-07.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[39]:

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

Task 2

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

In [40]:
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://sdk38546:***@dashdb-txn-sbox-yp-lon02-07.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[40]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Task 3

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

In [41]:
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

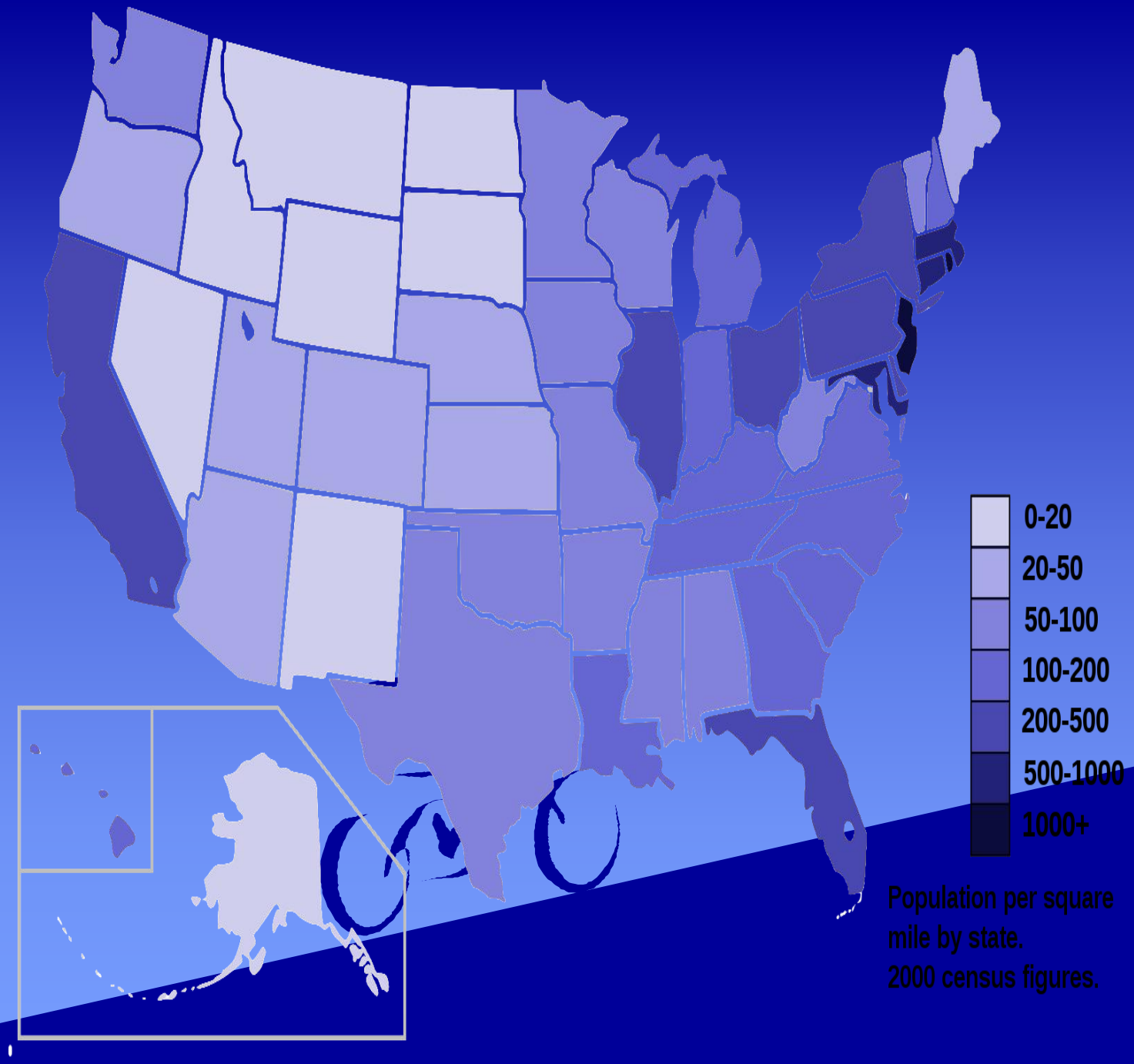
```
 * ibm_db_sa://sdk38546:***@dashdb-txn-sbox-yp-lon02-07.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[41]:

1

Screen shots

10. interactive map with Folium results



Screen shots

11. Plotly Dash dashboard

Dash uses Plotly.js for charting. About 50 chart types are supported, including maps.

Dash isn't just for dashboards. You have full control over the look and feel of your applications. Here's a Dash App that's styled to look like a PDF report.

Screen shots

12. predictive analysis (classification) results

TASK

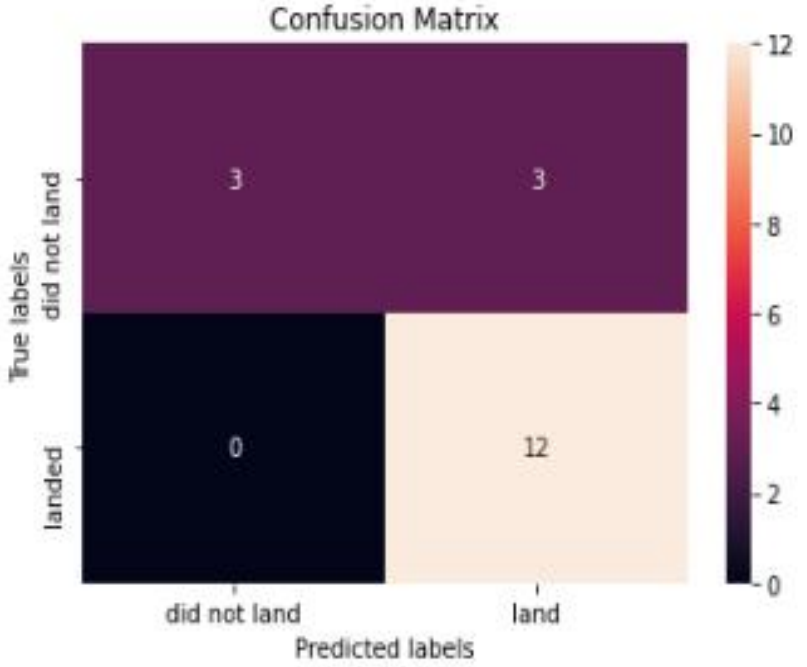
Calculate the accuracy of tree_cv on the test data using the method `score` :

In [38]:
```
tree_cv.score(X_test, Y_test)
```

Out[38]: 0.6666666666666666

We can plot the confusion matrix

In [39]:
```
yhat = svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Screen shots

13. Conclusion

Study

Result

Summary

Whole

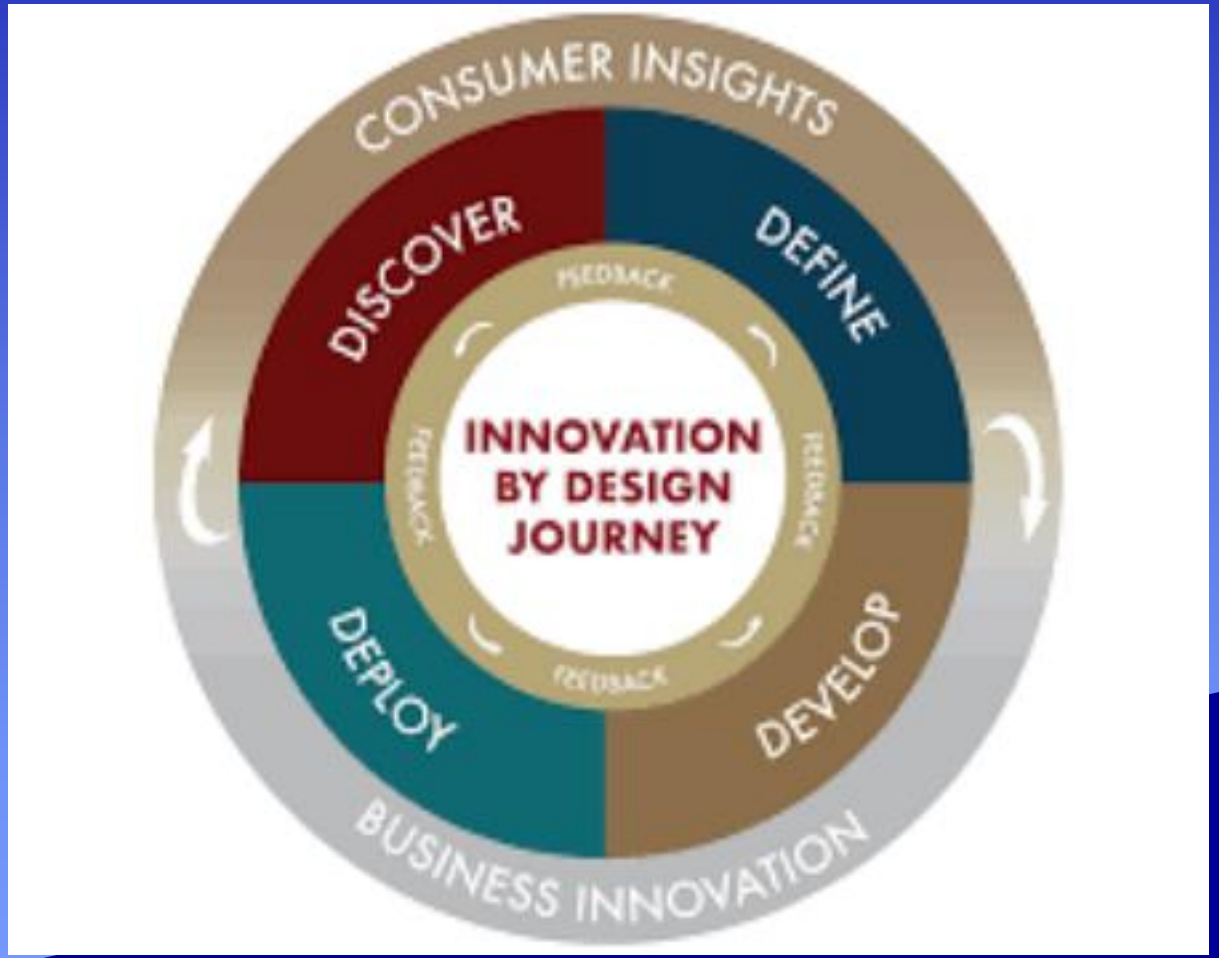
Nutshel

Screen shots

14. creativity to improve the presentation

Screen shots

14. innovative insights



Introduction

- **Project background and context**
- **Problems you want to find answers**

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - **Describe how data was collected**
- **Perform data wrangling**
  - **Describe how data was processed**
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - **How to build, tune, evaluate classification models**

Data Collection

- **Describe how data sets were collected.**
- **You need to present your data collection process use key phrases and flowcharts**

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

- **Present your web scraping process using key phrases and flowcharts**

- **Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose**

Place your flowchart of web scraping here

Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results

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

Flight Number vs. Launch Site

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

- Show the screenshot of the scatter plot with explanations

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

- Show the screenshot of the scatter plot with explanations

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

Flight Number vs. Orbit Type

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

- Show the screenshot of the scatter plot with explanations

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

- Show the screenshot of the scatter plot with explanations

Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations

All Launch Site Names

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Total Payload Mass

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

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

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

- Present your query result with a short explanation here

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

2015 Launch Records

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

- Present your query result with a short explanation here

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

- Present your query result with a short explanation here

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

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

- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

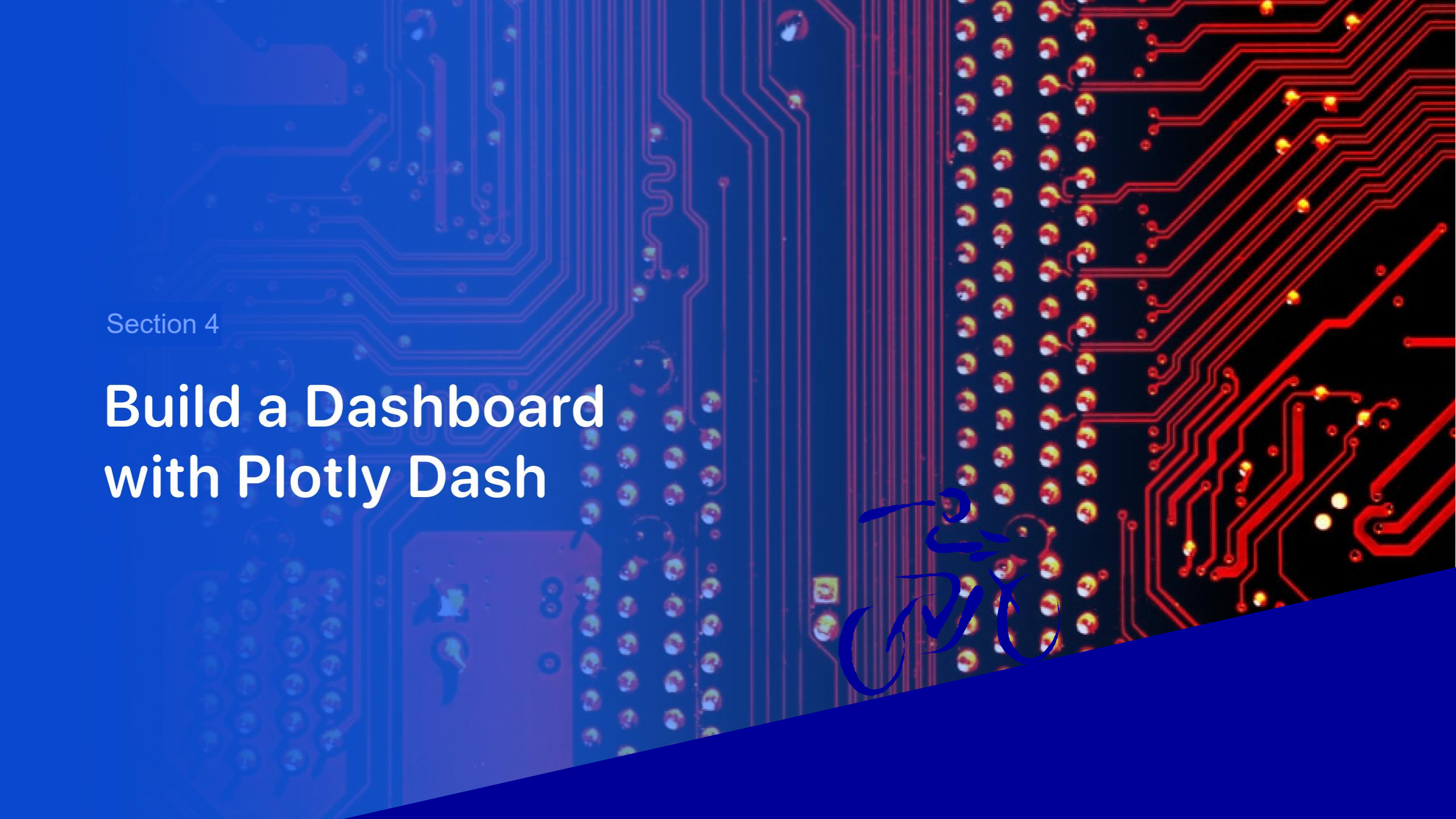
- Replace <Folium map screenshot 3> title with an appropriate title

- 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

- Explain the important elements and findings on the screenshot

Section 4

Build a Dashboard with Plotly Dash



<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

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

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

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation

Conclusions

- Point 1
- Point 2
- Point 3
- Point 4
- …

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

Thank you!