Project 1 - Exploring prisoner data

1. Data Loading:

a. Load the dataset “prisoners.csv” using pandas and display the first and last five rows in the dataset.
[Hint: Refer to read_csv, head and tail methods in pandas]
b. Use describe method in pandas and find out the number of columns. Can you say something about those rows who have zero inmates?
[Hint: Use the loc attribute of dataframe combined with conditional checks]

2. Data Manipulation:

a. Create a new column -’total_benefitted’ that is a sum of inmates benefitted through all modes.
[Hint: Use sum method with appropriate axis]
b. Create a new row - “totals” that is the sum of all inmates benefitted through each mode across all states.

3. Plotting:

a. Make a bar plot with each state name on the x -axis and their total benefitted inmates as their bar heights. Which state has the maximum number of beneficiaries?
[Hint: Use bar method of pyplot]
b. Make a pie chart that depicts the ratio among different modes of benefits.
[Hint: Use pie method of pyplot]

Project 2 - Exploring cereal dataset

1. Data Loading:

Load the data from “cereal.csv” and plot histograms of sugar and vitamin content across different cereals.
[Hint: Extract values of a specific column using their labels and use hist method of pyplot]

2. Data Manipulation:

- The names of the manufactures are coded using alphabets, create a new column with their full name using the below mapping.
  'N': 'Nabisco',
  'Q': 'Quaker Oats',
  'K': 'Kelloggs',
  'R': 'Raslston Purina',
  'G': 'General Mills' ,
  'P' :'Post' ,
  'A':'American Home Foods Products'
- Create a bar plot where each manufacturer is on the y axis and the height of the bars depict the number of cereals manufactured by them.

[Hint: Try using countplot this time or bar method of pyplot]

- Extract the rating as your target variable ‘y’ and all numerical parameters as your predictors ‘x’. Separate 25% of your data as test set.

2. Training and testing:

Fit a linear regression module and measure the mean squared error on test dataset.
[Hint: Explore linear models and metrics section of sklearn documentation]

Project 3:

Challenge/Requirement:

Fyntra is the largest online clothing company in USA. It sells clothing online, but they also have in-store style and clothing advice sessions. Customers come into the store, have meetings with a personal stylist, then can go home and order either on a mobile app or website for the clothes they want. Company wants to decide whether to focus the effort on mobile app experience or its website. As a drastic measure it is also evaluating to shut down the website.

As a MLexpert in the team, you will help the company make the right decision

Key issues

Clearly establish a correlation among the parameters supplied in data

Business benefits

Increase in profits as the focus on the optimal sales channel will result into the higher top line and the higher bottom line

Approach to solve

1. Compute and use seaborn to create a jointplot to compare the Time on Website and Yearly Amount Spent columns. Is there a correlation?
2. Do the same as above but now with Time on App and Yearly Amount Spent. Is this correlation stronger than 1st One?
3. Compute and explore types of relationships across the entire data set using pairplot . Based off this plot what looks to be the most correlated feature with Yearly Amount Spent?
4. Compute and create linear model plot of Length of Membership and Yearly Amount Spent. Does the data fits well in linear plot?
5. Using sklearn. train_test_split to split the data -- What is the use of random_state=85?
5. Train and Test the data and answer multiple questions
6. Compute the predict the data and do a scatter plot. Check if actual and predicted data match?
7. What is the value of Root Mean Squared Error?