Introduction To Data Science

Assignment Three

Delivery Instructions:

1. Cheaters will be graded by ***–ve points, Don't copy any code from anywhere.***
2. Due Date: **June 4, 2022**
3. No late submission will be accepted.
4. upload your python code as well as screenshots of the outcomes in zip file **YourGroup-Student1ID-Assignment2.zip**

You will be given historical sales information for various stores in various geographies. Each store features a variety of categories, each with its own set of weekly sales.

Note

- Date attributes represents the week
- The term "holiday" relates to whether the week is a special holiday week.

Question One [Data Cleaning]

Load the sales data from the supplied file"sales .csv", which contains historical sales data from different categories. Load the weather data from the supplied file"weather .csv", shows the average temperature in each retail region over time. Load the fuel pricing data from the supplied file"fuel .csv", which contains historical fuel prices for the region.

Then Perform the following functions:

1. Examine your datasets with Pandas, which displays all columns and their data types, the top ten for each dataset, and basic statistics for numeric columns (Count, mean, std, min, max). **Add your comments about the data**
2. Show the missing data and incorrect values for each column, such as zeros or negative sales.

3. Decide how you want to handle missing and incorrect values and implement it.
4. Merge all datasets into data frame based on the date and store

Introduction To Data Science

Assignment Three

Question Two [Visualization]

1. make a chart to illustrate if weekly sales are increasing or decreasing over time.
2. Make a chart to show how much each brand sells.
3. Determine the top ten selling stores.
4. Make a histogram to show the top 10 stores sales.
5. Create a chart that compares average weekly sales for the top ten selling stores during holidays and non-holidays.
6. Create a chart that displays the average weekly sales for each brand department for the top 10 selling stores.
7. Make a line chart to show the relationship between weekly sales and weather Temperature.
8. Make a line chart to show the relationship between the cost of fuel and weather weekly sales.
9. For each possible pair, create a pair plot to show different correlations.
   Hint: ”using sns.pairplot”
10. **Bonus**: plot and save a correlation matrix between all of the numerical attributes and discuss various correlations that you've discovered.
    Hint: “https://datagy.io/python-correlation-matrix/”

Question Three [Modeling]

To forecast weekly sales, we need to create a machine learning model:

1. Choose the optimal attributes to be input into this model based on Data Visualization and correlations in Question 2 and delete irrelevant features. Justify Your answer.
2. Divides the data into training and testing categories (80 percent training data and 20 percent testing data).
3. Create two separate supervised learning models to forecast weekly sales based on specific characteristics.
4. Compare the accuracy of the two models (in percentages).
5. Create a clustering model to group together store categories with similar sales. Which number of Clusters is the best? Why?