AIML | MODULE PROJECT

Python,EDA and Statistics

TOTAL SCORE 60

General Instructions:

1. Submission of all the parts is expected in 1 notebook only
2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only
3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.
4. If output for any code cell is missing, 50% marks will be deducted.

Submission Format:

1. '.ipynb' (Jupyter Notebook) and
2. '.html' (Jupyter Notebook > File > Download as > HTML)
5 Marks will be deducted if submission in any of the

- **DOMAIN :** Agriculture
- **CONTEXT :** In 2006, global concern was raised over the rapid decline in the honeybee population, an integral component to American honey agriculture. Large numbers of hives were lost to Colony Collapse Disorder, a phenomenon of disappearing worker bees causing the remaining hive colony to collapse. Speculation to the cause of this disorder points to hive diseases and pesticides harming the pollinators, though no overall consensus has been reached. Twelve years later, some industries are observing recovery but the American honey industry is still largely struggling. The U.S. used to locally produce over half the honey it consumes per year. Now, honey mostly comes from overseas, with 350 of the 400 million pounds of honey consumed every year originating from imports. This dataset provides insight into honey production supply and demand in America by state from 1998 to 2012

- **DATA DESCRIPTION :**

  Useful metadata on certain variables of the honeyproduction dataset is provided below :

  - **numcol**:
    Number of honey producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies which did not survive the entire year
  - **yieldpercol**: Honey yield per colony. Unit is pounds
  - **totalprod**: Total production (numcol x yieldpercol). Unit is pounds
  - **stocks**: Refers to stocks held by producers. Unit is pounds
  - **priceperlb**: Refers to average price per pound based on expanded sales. Unit is dollars.
  - **prodvalue**: Value of production (totalprod x priceperlb). Unit is dollars.
  - **Other useful information**:
    Certain states are excluded every year (ex. CT) to avoid disclosing data for individual operations. Due to rounding, total colonies multiplied by total yield may not equal production. Also, summation of states will not equal U.S. level value of production
    This is real commercial data, it has been anonymized, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

- **PROJECT OBJECTIVE :** Perform various EDA & statistical analysis to understand about the data and represent the same

- **STEPS AND TASKS :**

Q1. Data Collection and Data Preprocessing [20 Marks]

1) Import required libraries and import csv file into a dataframe. **[2Marks]**
2) Show the column names of all individual datasets. **[2Marks]**
3) Change the datatype of all columns expect price per lb. **[2Marks]**
4) Share 5-point summary and the details about the dataset. **[2Marks]**
5) Find the average production per state and shape of dataset. **[2Marks]**
6) Check the distribution of records for every year. **[2Marks]**
7) Find top 10 years with highest totalprod. **[2Marks]**
8) Find years with highest and lowest totalprod. **[2Marks]**
9) Find top 10 states with highest totalprod. **[2Marks]**

Q2. Data Visualization [20 Marks]

1) Visualize Q 1.9 and Q1.10 **[4 Marks]**
2) Find states with minimum and max price per lb. **[4 Marks]**
3) Visualize Q 2.2 **[4 Marks]**
4) Visualize the totalprod with respect to year. **[4 Marks]**
5) Visualize pairplot and share your insights. **[4 Marks]**

Q3 – Statistical Analysis [20 Marks]

1) Visualize distribution of numcol, yieldpercol, priceperlb, stocks in one single frame using subplots and share your insights. **[4 Marks]**
2) Visualize, boxplot and confirm if there are any outliers. **[4 Marks]**
3) Find Skewness and Kurtosis of complete dataframe. **[4 Marks]**
4) Make a new copy of dataset and use for further analysis. **[4 Marks]**
5) Try to make the skewed data as normal as possible & visualize the same. **[4 Marks]**