Assignment 1: Comprehensive Analysis and Visualization of Public Datasets

DSC 415: Data Analysis and Visualization

Nupur Upadhyay
IMS22191
Indian Institute of Science Education and Research, Thiruvananthapuram
Kerala, India
nupur22@iisertvm.ac.in

***Abstract*—This report consists of a comprehensive Analysis and Visualization of three datasets: the Keras datasets- California Housing price regression dataset, Reuters Newswire classification dataset and Kaggle- Breast Cancer Wisconsin (Diagnostic) Dataset. The objective was to perform an exploratory data analysis for these datasets and to visualize the information about the features and their relationships in these datasets.**

***Index Terms*—datasets, EDA, vizualization, analysis, Python**

I. Introduction

Exploratory Data Analysis (EDA) helps uncover patterns, trends, and anomalies in the data before applying advanced models. In this assignment, EDA was performed on three datasets: **the Reuters Newswire classification dataset** for text classification, **the California Housing price regression dataset** for socio-economic housing insights, and the **Breast Cancer Wisconsin (Diagnostic) Dataset** for medical diagnostics. The goal was to visualize distributions, detect outliers, and gain insight about various features of these datasets to guide further analysis.

II. The Reuters Newswire classification dataset

A. About the dataset

The Reuters Newswire dataset is a popular benchmark for text classification and natural language processing tasks. It contains thousands of news articles labeled with multiple topics such as economics, politics, and business. The dataset provides textual features and categorical labels, making it ideal for exploring word distributions, topic frequencies, and text length variations. Through EDA, insights such as the most frequent topics, word usage patterns, and class imbalances can be found that may influence text classification models.

B. Sample article

To see what a sample article from the dataset looks like, a random article (with index value 15) was decoded to words from the integer encoding (Box 1). The label for this particular article is 8, meaning "grain", which covers news related to grain markets and commodities.

Box 1: Sample article (Index 15)

Sample article: ? commercial and industrial loans on the books of the 10 major new york banks excluding acceptances fell 572 mln dlrs to 64 297 billion in the week ended march 11 the federal reserve bank of new york said including acceptances loans fell 475 mln dlrs to 65 16 billion commercial paper outstanding nationally increased 2 98 billion dlrs to 339 00 billion national business loan data are scheduled to be released on friday reuter 3
Label: 8

C. No. of articles per topic

To find how many articles exist for each label topic, a bar graph was plotted between the two variables using the 'Matplotlib' library (Figure 1). Topic number three, corresponding to the topic "earn", which is the category for earnings-related news. Most news articles are classified under this label because earnings announcements are very common in financial news. The second highest is label four, which stands for "acq" or acquisition of companies.

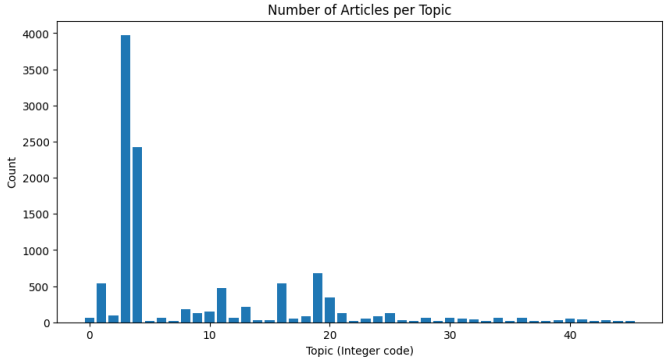

Fig. 1. Bar graph of No. of articles per topic

D. Distribution of Article lengths

Next, to analyze how long each article is, a histogram was plotted between the number of words (or article length) and the number of articles with those many words (Figure 2). The number of words was extracted as the ’length’ of the text (or the x parameter) of the dataset. The majority of the articles are between zero and five hundred words, with the highest fraction of articles falling in the zero to two-hundred-word range.

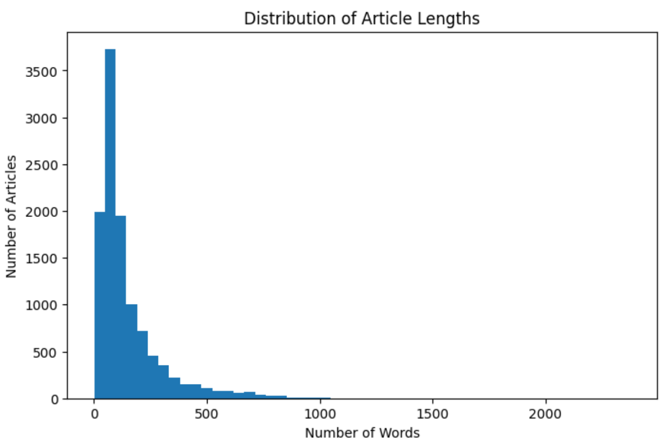

Fig. 2. Article length distribution histogram

E. Distribution of text length and topic

Further, to find out how the article length (or the number of words) correlates to the topic, a scatter plot visualization was made for the article length and topic variables (Figure 3). It can be seen that some topics, like 20, have longer articles compared to the other topics.

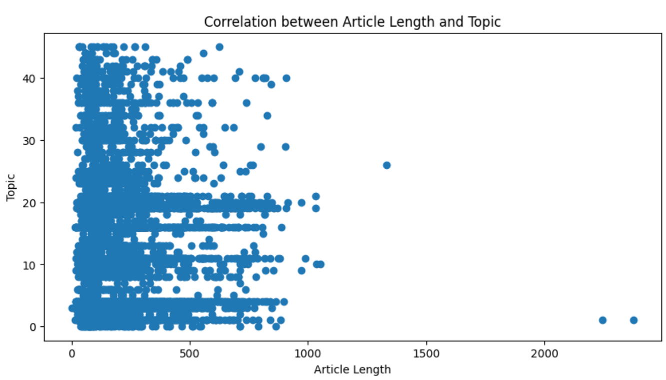

Fig. 3. The distribution of article lengths per topic

F. Word Cloud Representation

A word cloud is a visual representation of text data where the size of each word indicates its frequency or importance in the dataset. Words that appear more frequently in the text are displayed in a larger font so as to quickly identify the prominent themes or keywords in a dataset. A word cloud was plotted for this dataset using the ’WordCloud’ library (Figure

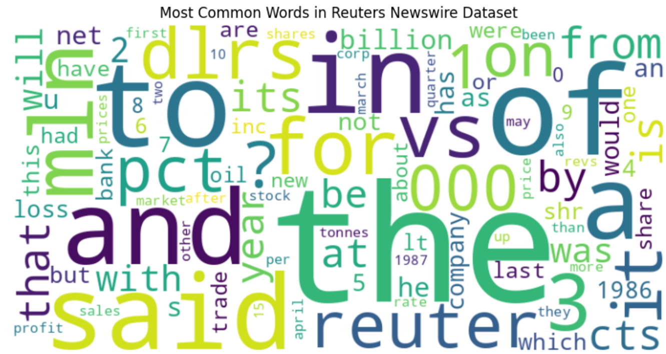

Fig. 4. Word cloud of the most frequent words in the dataset

4). For this dataset, words like the, vs, of, and, to, etc. are the most frequent words in these news articles.

G. Average article length

To find out how long the average article in the dataset is, the number of words in each article was counted, and then these word counts were averaged (Box 2).

Box 2: Average article length

Average article length: 145.96419665122906 words

III. The California Housing Price Regression Dataset

A. About the dataset

The California Housing dataset offers real-world data on housing prices and related demographic factors in California districts. It includes features such as median income, house age, population, and average number of rooms, along with the median house value as the target variable. EDA on this dataset helps reveal the relationships between housing prices and socioeconomic indicators, highlights geographical trends, and detects any anomalies or outliers in the housing data. Figure 5 shows the variables and the stastistics about these variables.

index	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.0	20640.0	20640.0	20640.0	20640.0	20640.0	20640.0	20640.0	20640.0
mean	-119.56970977783203	35.63186264038086	28.63948631286621	2635.76318359375	537.8980102539062	1425.4766845703125	499.5396728515625	3.870671033859253	206855.8125
std	2.003523588180542	2.135953903198242	12.585662841796875	2181.60986328125	421.2468566894531	1132.464599609375	382.3299560546875	1.8998239040374756	115395.375
min	-124.3499984741211	32.540000915527344	1.0	2.0	1.0	3.0	1.0	0.4999000132083893	14999.0
25%	-121.80000305175781	33.93000030517578	18.0	1447.75	295.0	787.0	280.0	2.5633999705314636	119600.0
50%	-118.48999786376953	34.2599983215332	29.0	2127.0	435.0	1166.0	409.0	3.5347999334335327	179700.0
75%	-118.01000213623047	37.709999084472656	37.0	3148.0	647.0	1725.0	605.0	4.743250131607056	264725.0
max	-114.30999755859375	41.95000076293945	52.0	39320.0	6445.0	35682.0	6082.0	15.000100135803223	500001.0

Fig. 5. Information about the dataset.

B. Scatter plots- house value vs other variables

To check how the median house value varies with respect to the other variables like house age, latitude, longitude, etc., in the dataset, scatter plots were plotted using the ’Matplotlib’ library and a collage was created using the ’PIL’ library (Figure 6). It can be seen from these plots that house age has no correlation with the house value, i.e. old and new houses both can have higher or lower values. Another interesting trend about the household income and house value can also be seen.

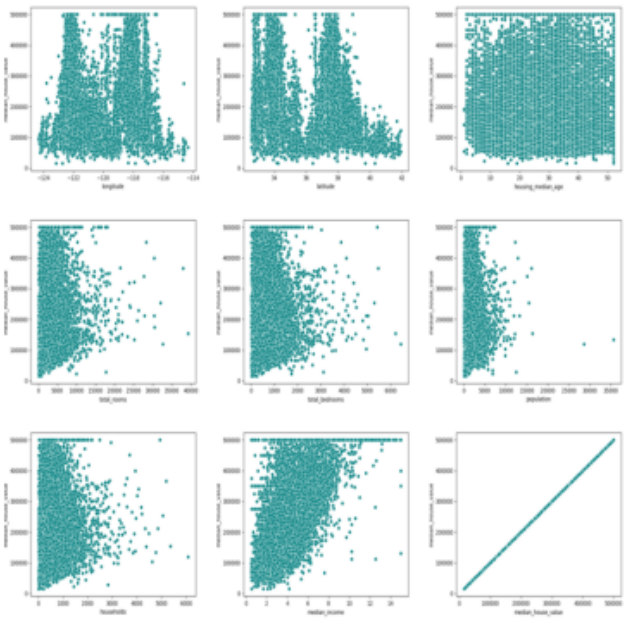

Fig. 6. Scatter plots of house value vs other variables.

C. Correlation matrix of the variables

Then, to know the correlation between different variables of the dataset better, a correlation matrix was plotted and visualized using the "coolwarm" palette (Figure 7). Here, in the image, the correlation coefficients for each pair of variables are shown. Correlation coefficients with high positive correlation are in red and the ones with more negative in darker blue. It can be seen that the median income of the household has a positive correlation with the house value, meaning that for houses with higher values, the household income is higher. Also, latitude and longitude do not have a strong correlation with the house value, meaning that the location of the house tends to have a little effect on the house value.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
longitude	1.000000	-0.924664	-0.108197	0.044568	0.068378	0.099773	0.055310	-0.015176	-0.045967
latitude	-0.924664	1.000000	0.011173	-0.036100	-0.066318	-0.108785	-0.071035	-0.079809	-0.144160
housing_median_age	-0.108197	0.011173	1.000000	-0.361262	-0.320485	-0.296244	-0.302916	-0.119034	0.105623
total_rooms	0.044568	-0.036100	-0.361262	1.000000	0.929893	0.857126	0.918484	0.198050	0.134153
total_bedrooms	0.068378	-0.066318	-0.320485	0.929893	1.000000	0.878026	0.979829	-0.008093	0.050594
population	0.099773	-0.108785	-0.296244	0.857126	0.878026	1.000000	0.907222	0.004834	-0.024650
households	0.055310	-0.071035	-0.302916	0.918484	0.979829	0.907222	1.000000	0.013033	0.065843
median_income	-0.015176	-0.079809	-0.119034	0.198050	-0.008093	0.004834	0.013033	1.000000	0.688075
median_house_value	-0.045967	-0.144160	0.105623	0.134153	0.050594	-0.024650	0.065843	0.688075	1.000000

Fig. 7. Correlation matrix for all pairs of variables.

D. The location of the houses

The locations of the houses (top 1000) in this dataset were plotted on a map using the 'folium' library using the variables-longitude and latitude (Figure 8). The map shows that the houses are spread across the state with two major clusters.

E. The cheapest houses

To find the locations with the cheapest houses, the dataset was segmented into the lowest 25 per cent of house values

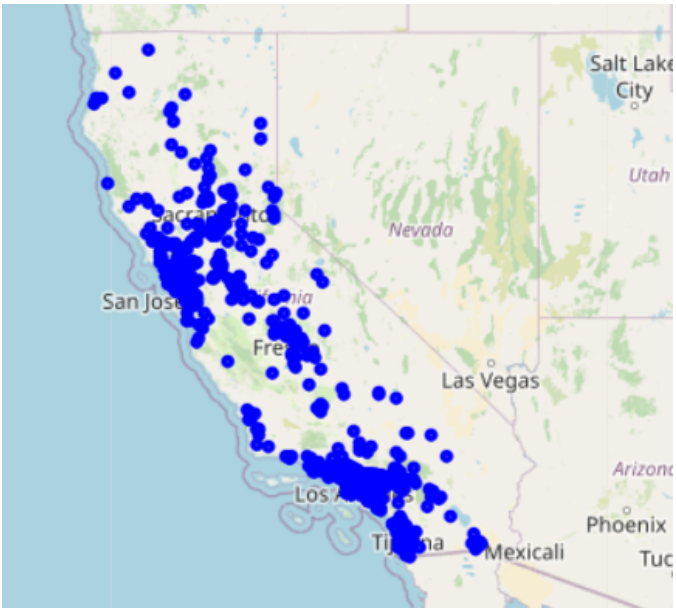

Fig. 8. The map of the house locations in California.

or the first quartile. Then a plot of the latitude and longitude was plotted with the house value as the hue (Figure 9). It can be seen that among these cheapest houses, the houses with relatively higher and lower values exist in the same place, implying that the location might not be affecting the value of these houses, matching with the observation made for the houses in the whole dataset.

Fig. 9. Location plot of the cheapest (lowest 25 per cent house value) houses.

Then, to see the locations of these houses on the map of California, a map was created using the 'folium' library (Figure 10). It can be seen that the cheapest houses are spread across the state and not grouped in any particular location.

Fig. 10. Location plot of the cheapest (lowest 25 per cent house value) houses.

Then, to check if there was any correlation between the house value and house age for the cheapest houses, a box plot between the two variables was plotted using the 'Seaborn' library (Figure 11). It can be seen that the cheap houses range from very new to very old and thus, the age of the houses seems to have no significant effect on the house being cheaper or not.

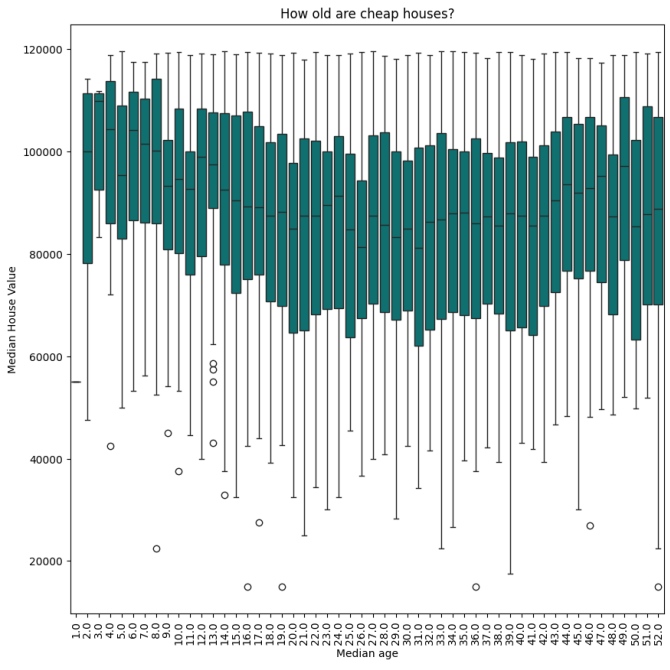

Fig. 11. Location plot of the cheapest (lowest 25 per cent house value) houses.

F. The most expensive houses

Similarly, to know about the locations of the most expensive houses (top 25 per cent of the house values), the part of the dataset with the last quartile of the house value was used. A scatter plot between the latitude and longitude, with the house value as hue, was plotted (Figure 12). To see the locations on the map, a map was created using the 'folium' library (Figure 13). While there was no particular trend for the cheapest houses, the most expensive houses are clustered in two regions, as seen from the scatter plot and the map.

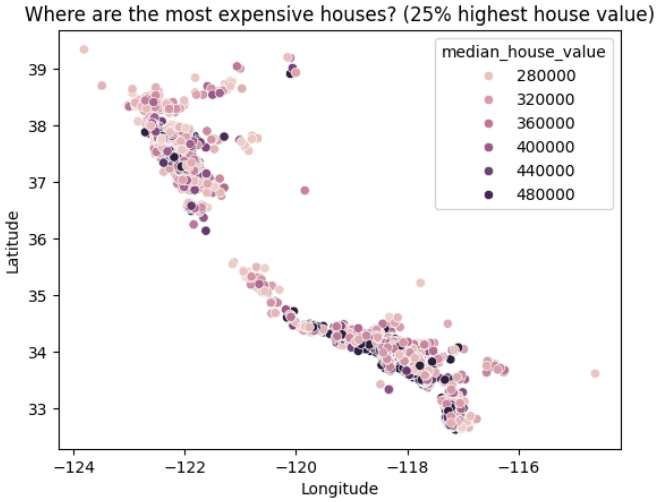

Fig. 12. Location plot of the most expensive (top 25 per cent house value) houses.

Fig. 13. Location of the most expensive (top 25 per cent house value) houses

G. Income vs House value

Lastly, to check whether there was any difference between the household income and the house value in that region, a

scatter plot between these variables was plotted for the most expensive and the cheapest houses (Figure 14). It can be seen that the household income for the cheapest houses is low and clustered, whereas for the most expensive houses, the household incomes are diverse.

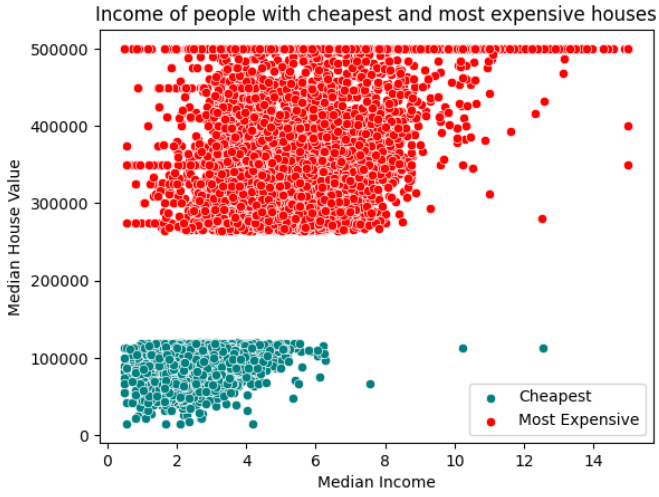

Fig. 14. Scatter plot of the household incomes and the house value for the cheapest and the most expensive houses.

IV. Breast Cancer Wisconsin (Diagnostic) Dataset

A. About the dataset

The Wisconsin Breast Cancer (Diagnostic) dataset is widely used in medical diagnostics and machine learning classification tasks. It contains measurements from digitized images of breast mass samples, including features such as radius, texture, perimeter, and smoothness, each derived from cell nuclei characteristics. The dataset is labeled as either benign or malignant, enabling EDA to explore differences between the two classes, examine feature distributions, and identify the most informative attributes for distinguishing cancerous from non-cancerous samples.

B. Correlation between tissue Features and Malignancy

To find the correlation between tissue features (radius, concavity, etc.) and Malignancy, a correlation matrix was visualized for the first 10 variables (due to limit of visualization for more variables) as shown in the Figure 15. There is a significant negative correlation between sample radius, perimeter, concave points and Malignancy.

Fig. 15. Correlation between features and Malignancy (Target) for 10 variables

C. Ratio of Malignant vs Benign tissues

To check the ratio of Malignant vs Benign tissue samples, the counts of each index of target were extracted (0 for Malignant and 1 for Benign) and plotted using the 'Matplotlib' library (Figure 16). In this dataset, there are more samples classified as Malignant.

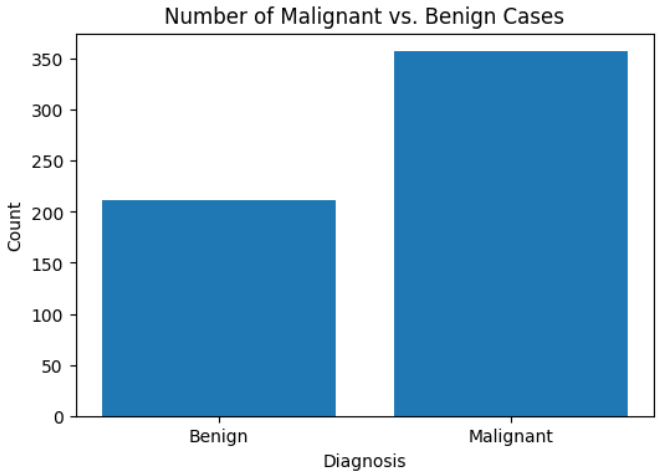

Fig. 16. Counts of Malignant and Benign samples

D. Relationship between Mean concave points and Malignancy

To better understand the relationship between concave points and Malignancy, a box plot between the two variables was plotted using the 'Seaborn' library (Figure 17). There is a difference between the range and inter-quartile range in the number of mean concave points and malignancy. Thus, this variable can be a good marker for the diagnosis of Breast cancer.

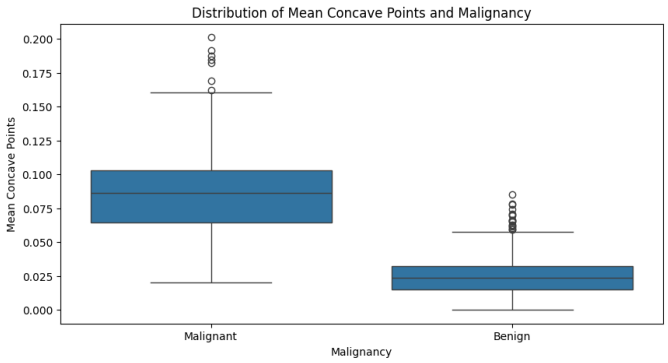

Fig. 17. Mean concave points vs Malignancy box plot

E. Relationship between Mean Radius and Malignancy

Similarly, to find the relationship between radius and Malignancy, a box plot between the two variables was plotted using the 'Seaborn' library (Figure 18). Again, there is a clear difference between the range and inter-quartile range for malignant and benign samples. Thus, like mean concave

points, this variable can also be a good marker for the diagnosis of Breast cancer.

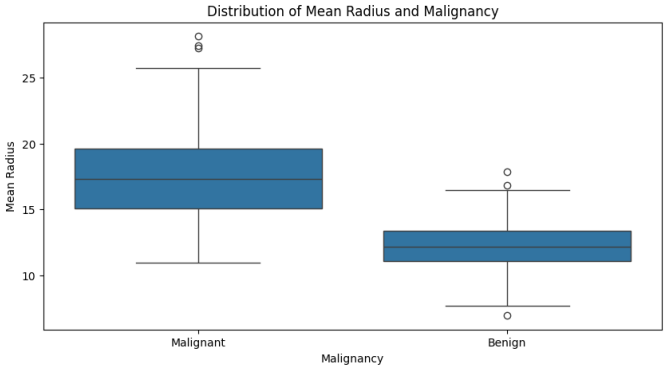

Fig. 18. Mean Radius vs Malignancy box plot