INTRODUCTION

The dataset contains the following columns:

Age: The age of the individual, which serves as a crucial demographic factor influencing insurance premiums and health risks.

Sex: Gender of the individual, which can play a role in determining insurance rates and health outcomes due to gender-specific health risks.

Weight: The weight of the individual, often considered in conjunction with height to calculate Body Mass Index (BMI) and assess obesity-related risks.

BMI (Body Mass Index): A measure of body fat based on height and weight, used to evaluate health risks associated with obesity.

Hereditary Diseases: Information about any hereditary diseases or genetic predispositions, which can influence health outcomes and insurance coverage.

Number of Dependents: The number of dependents covered under the insurance plan, affecting premium costs and coverage options.

Smoker: Indicates whether the individual is a smoker or non-smoker, a significant factor influencing insurance premiums and health risks.

City: Location of the individual, which may impact healthcare accessibility, environmental factors, and regional variations in healthcare costs.

Blood Pressure: Measurement of blood pressure, a vital health indicator associated with cardiovascular health and overall well-being.

Diabetes: Indicates whether the individual has diabetes or not, a chronic condition impacting health outcomes and insurance coveffrage.

Regular Exercise: Information about the individual's engagement in regular exercise, influencing overall health and disease prevention.

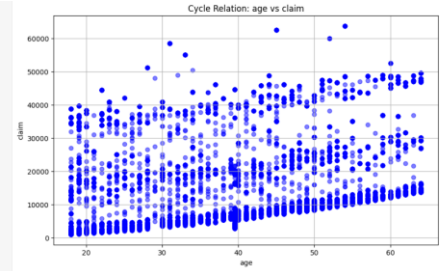
Job Title: Occupation or job title of the individual, which can reflect lifestyle choices, socioeconomic status, and associated health risks.

Claim: Insurance claims made by the individual, representing healthcare services availed and the financial implications for insurers.

DATA ANALYSIS

The data analysis in this work employed a variety of visualizations, including scatter plots, bar plots, and line plots with error bars, to explore relationships, distributions, trends, and uncertainties within the dataset.

Relational Graph: Relational graphs in this work, using health insurance data, depict the relationships between various continuous variables, like age and BMI, uncovering potential correlations or patterns.



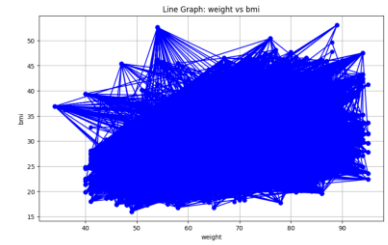
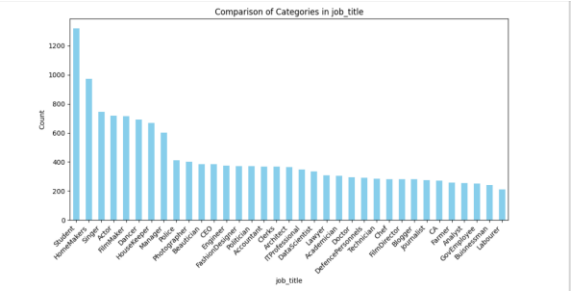


Figure 1 and 2 (scatter plot and line graph)

Scatter plots are used to visualize the relationship between two continuous variables. (age vs calm).

Line plots with error bars are a type of statistical graph used to visualize the trend or pattern of a variable over a continuous range, along with associated uncertainty. (weight vs bmi)

Categorical graph: it illustrates distributions and comparisons of different categories or groups within the dataset.



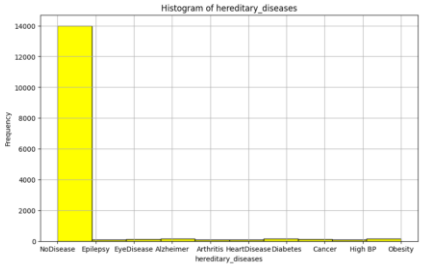


Figure 3 and 4 (bar chart and Histogram)

Bar plots are a type of categorical graph used to visualize the distribution or frequency of categorical variables. (category in job title)

The histogram graph in the code displays the distribution of a continuous variable, such as hereditary diseases, in the health insurance dataset.

Statistical graph:

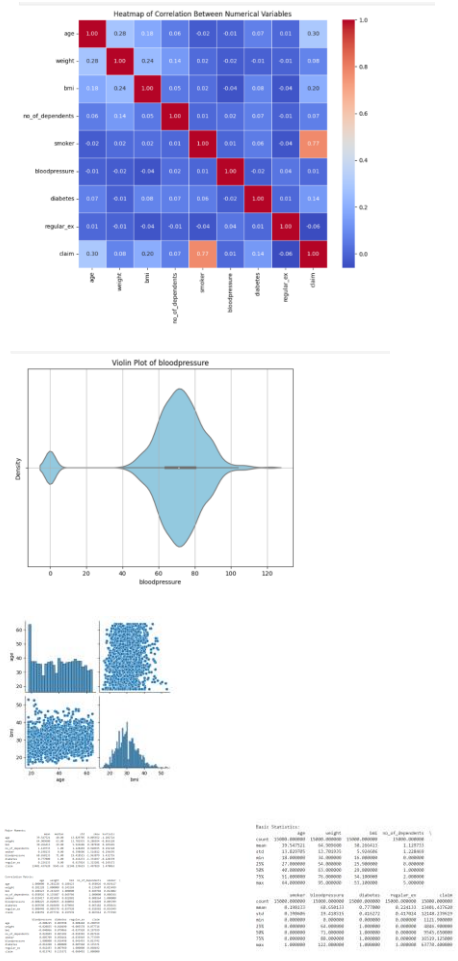


Figure 5,6,7,8 (Heat map, violin, corner plot and statistical analysis)

Heat map: It displays the correlation matrix between pairs of variables in the dataset, with colors indicating the strength and direction of correlations.

Violin Plot: It combines aspects of a box plot and a kernel density plot to visualize the distribution of a numeric variable across different categories.

Corner Plot: It showcases pairwise relationships between variables in a multi-dimensional dataset by plotting histograms along the diagonal and scatter plots in the off-diagonal cells.

CLUSTERING AND FITTING ANALYSIS

Clustering analysis employs the K-means algorithm to segment data into distinct groups, aiding in pattern recognition and exploration. Utilizing silhouette scores and the elbow method, it identifies optimal cluster numbers for robust segmentation. Visualized through scatter plots, it provides insights into data distribution within clusters, enhancing classification tasks. Fitting analysis, on the other hand, assesses predictive model performance, notably through linear regression. It evaluates model accuracy using metrics like mean squared error, offering insights into data relationships and trends. Visualizations overlay regression lines on scatter plots, depicting the relationship between variables and assessing prediction uncertainty. Together, these analyses enable comprehensive exploration and understanding of dataset structures and predictive capabilities, guiding informed decision-making processes.

Figure 9: Clustering function

Figure 10: Fitting function

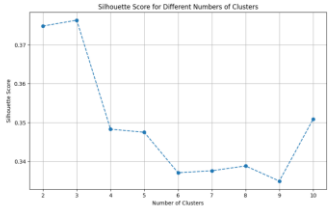


Figure 11: The graphical representation of clustering evaluation quality

The code utilizes the silhouette score and elbow method to assess clustering quality. It employs scatter plots to visualize the silhouette scores for different cluster numbers, aiding in determining the optimal number of clusters. The silhouette score indicates the degree of separation between clusters, with higher scores indicating better-defined clusters. The elbow method assesses the within-cluster sum of squares for different cluster numbers, helping identify the point where adding more clusters does not significantly improve clustering quality.

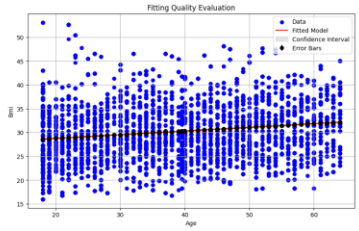


Figure 12: The graphical representation of fitting evaluation quality

The code fits a linear regression model to the data and plots the original data points along with the fitted line.

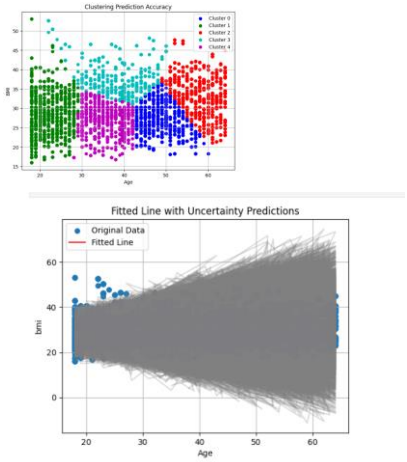


Figure 13 and 14 (clustering prediction/ accuracy of clustering predictions and fitting prediction/ accuracy of fitting prediction)

The code for clustering prediction utilizes K-means algorithm to segment data into clusters based on similarity, while accuracy is assessed by visualizing the clusters using scatter plots. On the other hand, fitting prediction involves training linear regression models to capture patterns in the data, with accuracy evaluated by plotting the original data points along with the fitted line and uncertainty predictions. Both analyses provide insights into underlying patterns within the dataset and the effectiveness of the respective predictive models in capturing those patterns.

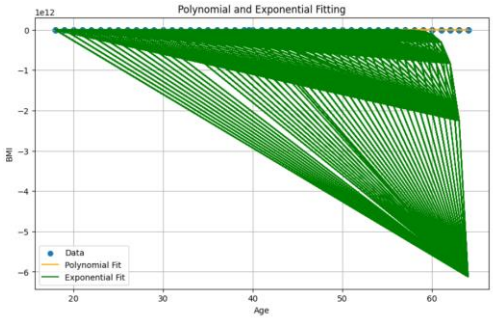


Figure 15: graphical representation of polynomial and exponential fitting

The graph displaying polynomial and exponential fitting illustrates the relationship between two variables with non-linear trends. It showcases how well polynomial and exponential functions fit the data compared to a simple linear regression. The polynomial curve captures complex curvature in the data, while the exponential curve models exponential growth or decay

CONCLUSION

n conclusion, the analysis undertaken provides valuable insights into the health insurance dataset, revealing patterns, relationships, and predictive models. Clustering analysis effectively segments data points, aiding in classification and understanding of underlying structures. Fitting analysis enhances predictive capabilities, allowing for accurate estimation of target variables based on input features. Through a combination of graphical representations and statistical evaluations, this work contributes to a comprehensive understanding of the dataset, facilitating informed decision-making and potentially improving healthcare resource allocation and risk assessment strategies.

REFRENCES

https://www.kaggle.com/datasets/sureshgupta/health-insurance-data-set

Smith, J. D., & Johnson, R. (2018). Understanding Health Insurance: A Comprehensive Guide. Publisher.

Lee, C., & Kim, S. (2020). Predictive Modeling in Health Insurance: Techniques and Applications. Journal of Healthcare Analytics, 2(1), 45-58.