|  |  |  |
| --- | --- | --- |
|  | **Use Case** | **Tasks** |
| 1 | Use NYC 311 dataset  Download from  <https://www.kaggle.com/datasets/sapnilbhatnagar/nyc-311-service-request-analysis/data>  (200 MB)  Code also provided (github) | * Perform EDA as in the code (enhance wherever you see feasible) * Deduce if the complaint type has influence on the time to resolve the complaints (types) |
| 2 | Objective: To demonstrate essential statistical techniques and their importance in data science using the CA Housing dataset. (in the Github folder) CA\_housing.csv  The analysis will cover central measures, dispersion, sampling theory, symmetry of data (including skewness and kurtosis), outliers, and multicollinearity using the Variance Inflation Factor (VIF) method.  Dataset Overview:  The CA Housing dataset contains information about housing prices and related features in California. It includes attributes such as median house value, median income, housing median age, total rooms, total bedrooms, population, households, and proximity to the ocean.  1. Central Measures  Objective: To determine the central tendency of housing prices and other numerical features.  Mean: Calculate the mean of median house values to understand the average housing price.  Median: Calculate the median to understand the middle point of housing prices.  Mode: Determine the mode of categorical features (if any) to understand the most frequent category.  Outcome: The central measures provide insights into the typical value of housing prices, helping to understand the average and most common values in the dataset.  2. Dispersion  Objective: To measure the variability or spread of housing prices and other numerical features.  - Variance: Calculate the variance of median house values.  - Standard Deviation: Determine the standard deviation to understand the average deviation from the mean.  - Range: Find the range to understand the difference between the highest and lowest values.  - Interquartile Range (IQR): Calculate the IQR to understand the spread of the middle 50% of the data.    Outcome: Measures of dispersion provide information on how much variability exists in the data, which is crucial for understanding the spread and consistency of housing prices.  3. Sampling Theory  Objective: To understand the properties of different samples from the dataset.  - Random Sampling: Draw random samples to estimate population parameters.  - Stratified Sampling: Use stratified sampling to ensure representation from different strata (e.g., proximity to the ocean).    Outcome: Sampling helps in making inferences about the population from the dataset, ensuring that different segments are adequately represented.  4. Symmetry of Data, Skewness, Kurtosis  Objective: To understand the distribution shape of housing prices and other numerical features.  - Skewness: Calculate skewness to determine the asymmetry of the distribution.  - Kurtosis: Calculate kurtosis to understand the peakedness of the distribution.    Outcome: Analyzing skewness and kurtosis helps in understanding the distribution's shape, identifying deviations from normality, and potential data transformations needed.  5. Outliers  Objective: To identify and analyze outliers in the housing prices and other numerical features.  - Boxplot: Use boxplots to visually identify outliers.  - Z-score: Calculate Z-scores to quantitatively identify outliers.    Outcome: Detecting outliers is essential for ensuring data quality and robustness, as outliers can significantly impact statistical analyses and model performance. | * Perform house price using KNN regression and Linear regression * perform CV (K-Fold, repeated K-Fold * compute MAE, MSE, MAPE, MPE, R2 * state observations on the results * save/loading the models * prediction |
|  |  |  |
|  |  |  |