FINAL PROJECT

Project Description

You are hired by the American Cancer Society (https://www.cancer.org/) to write a white paper exploring the factors related to cancer incidences/deaths in the US. They will use your analysis to identify regions (and associated partners) for cancer interventions across the US. The dataset provided includes information at the county level in the US.

Data

You will use the **Cancer.xlsx** dataset. Below is the data description:

Variable	**Description**
countyCode	Unique county code
State	State Abbreviation
PovertyEst	Total number of people below the poverty line in the county
medIncome	Median household income in the county
Name	County name
popEst2015	Estimated population in the county in 2015
incidenceRate	Cancer (all cancers) age-adjusted incidence per 100K in the county
avgAnnCount	2009-2013 mean incidences in the county
fiveYearTrend	incidence five year trend
recentTrend	incidence recent trend
deathRate	deaths from cancer per 100K in the county
avgDeathsPerYear	Average number of deaths in the county per year (2009-2013)
recTrend	Cancer mortality recent trend

Objectives

1. **Data Visualization and Summary Measures**
   a. Create a series of visualizations that effectively communicate the geographic distribution of cancer incidence and death rates across counties and states within the dataset.
   b. Develop an interactive dashboard using Tableau to allow users to explore variations in cancer rates by median income and poverty levels.

   c. Utilize advanced visualization techniques to represent multi-variable relationships, such as the association between population estimates and cancer rates, while considering the state as a categorical variable.
   d. Design visualizations that highlight any significant trends or outliers in the dataset, such as counties with exceptionally high or low cancer rates.
   e. Compute descriptive statistics including mean, median, mode, range, variance, and standard deviation for key variables such as median income, poverty estimates, and cancer rates.
   f. Analyze the central tendency and dispersion measures by state to see if regional patterns exist.
   g. Generate summary tables that categorize counties by quantiles based on their cancer rates and median income, to identify socio-economic patterns.

2. **Regression Analysis**
   a. Construct a linear regression model to examine the impact of median income and poverty estimates on cancer incidence rates.
   b. Interpret the coefficients of the regression model to understand the magnitude and direction of the relationship between socio-economic factors and cancer rates.
   c. Assess the model's goodness-of-fit using R-squared, adjusted R-squared, and analyze the residuals to check for homoscedasticity and normality.
   d. Conduct hypothesis testing on regression coefficients to determine the statistical significance of the predictors, using p-values and confidence intervals.
   e. Investigate the assumptions of linear regression, including linearity, independence, and absence of multicollinearity, and discuss how any violations may impact the model's validity.

Deliverables

1. **Report**
   a. Must include:
      i. *Introduction:* Brief on objectives, data, and significance of analysis
      ii. *Methodology:* Analytical models/concepts and tools applied.
      iii. *Results:* Key findings from visualizations, summary measures, and regression.
      iv. *Discussion:* Insights and implications from the analysis.
      v. *Conclusion:* Recap of findings and their broader impact.
   b. Format: Max 10 pages with charts and tables.
2. **Source Files**
   a. Submit R scripts, Excel files, and Tableau workbooks used in the analysis.
   b. Label and organize files for clarity.

Evaluation Criteria

The project will be evaluated based on the following criteria:

- **Comprehensiveness:** Coverage of all required analysis and concepts.
- **Accuracy:** Correct application of concepts.
- **Insights and Interpretation:** Quality of insights drawn and their interpretation in the business context.
- **Report Presentation:** Clarity, structure, and presentation of the report.

- **Application of Course Concepts:** Effective application of the topics covered in the course.

Notes To Students

- Ensure to cite any external sources of data or information used.
- Focus on applying the concepts learned in the course to derive meaningful business insights.
- Pay attention to the quality of your visualizations and ensure they are self-explanatory.
- Ensure the report is well-structured and free of grammatical errors.

Ensure that all submitted files are well-organized and clearly labeled, facilitating easy navigation and understanding of your work.