**Final Report: Data Wrangling and Modeling in R; A Simple Linear Regression Lesson for Researchers.**

**Introduction:**

This project is dedicated to demonstrating the principles and application of simple linear regression analysis using the R programming language, specifically tailored for researchers across various fields. The objective is to provide a comprehensive guide that empowers researchers to leverage simple linear regression in their analytical work, enabling them to uncover relationships between two continuous variables in their datasets. The material in this project will cover essential techniques for data handling and preparation in R, to ensure participants can manage and prepare their datasets for regression analysis effectively. Using R, a powerful tool for statistical computing and graphics, this hands on project demonstrate how to perform simple linear regression analysis step-by-step including data preparation, model fitting, and diagnostics; this project assumes the user has a prior solid theoretical foundation, outlining the key concepts of simple linear regression such as understanding the assumptions behind the model, the interpretation of the coefficients, and the significance of the model fit.

Finally, two techniques, ‘Base R’ and ‘ggplot2’ for visualizing regression results will be presented, using R’s powerful graphical capabilities to create plots that effectively communicate the findings of the analysis. From the output in R, the researcher will be able to interpret the results obtained from the regression analysis. This includes understanding the output provided by R, such as regression coefficients, R-squared values, p-values, and diagnostic plots.

**Methods:**

The methodological step by step approach combines data cleaning, visualization, and statistical modeling. It provides a robust framework for analyzing and interpreting the relationship between two key variables in a clinical dataset. This approach can be adapted for various research scenarios, particularly in clinical and health-related research.

* **Step 0: Load Libraries**
  + Essential R libraries are loaded, including ‘**readr’** for reading CSV files, ‘**tidyverse’** for data manipulation and visualization, **‘mice’** for missing data imputation, **‘dplyr’** for data transformation, and **‘ggplot2’** for advanced graphing.
* **Step 1: Upload Data and Save into Data Frame**
* Data is uploaded from a CSV file named "Patient\_List\_aaliyah-washington-md.csv" into the R environment as a DataFrame. The read\_csv function from the readr package is used, with the argument ‘show\_col\_types = FALSE’ to simplify initial data viewing.
* **Step 2: Data Examination and Cleaning**
  + The data is examined and cleaned using functions from the dplyr package. Specifically, the argument ‘mutate\_all()’ is used to replace 'None' values with NA (missing values).
  + A preliminary data summary is obtained using the summary function.
  + Additional functions are employed for more detailed examination of specific columns (e.g., 'wt' and 'hgba1c') and counting NA values.
  + **Mean/Median Imputation: A simple example of Imputation of missing values.** 
    - Mean imputation is one method used to handle missing data, appropriate for when missing data is not significant and is missing completely at random (MCAR). This includes converting relevant columns to numeric and then applying the imputation.
    - A custom function, ‘impute\_data’, is defined and applied to the dataset for this purpose. ( Refer to .R file in the repository).
  + The DataFrame is further cleaned for plotting by removing rows with NA values and zeros in key columns ('wt' and 'hgba1c').
* **Step 3A: Plot Data Using Base R**
  + A scatter plot is created using the base R plot function. The plot illustrates the relationship between 'wt' (weight) and 'hgba1c', with labels and a main title.
* **Step 3B: Plot Data Using ggplot2**
  + An alternative scatter plot is created using ggplot2. This approach includes a trend line (using geom\_smooth) and customized aesthetics like centered title and minimal theme.
* **Step 4A and 5A: Simple Linear Regression and Output Summarization**
* A simple linear regression model is fitted using the ‘lm’ function in R, with 'hgba1c' as the response variable and 'wt' as the predictor.
* The ‘summary’ function is used to summarize the output of the linear regression model, providing insights into the model coefficients, statistical significance, and model fit.
* **Step 4B and 5B: Simple Linear Regression and Output Summarization with Cleaned Data**
  + A separate linear regression analysis is performed on the cleaned dataset, following similar steps as above. This allows for comparison between models using the original and cleaned data.

**Results: Analysis and Visualization of Simple Linear Regression: (Shown for Cleaned Data):**

The simple linear regression analysis was conducted to explore the relationship between patients' weight (wt) and HgbA1c levels(%), using the cleaned dataset. The results of the regression model are as follows:

**Graph 1: Using Base R to Model the Relationship Between Weight and HgbA1c.**

**A graph of weight loss

Description automatically generated**

**Graph 2: Using ggplot2 to Model the Relationship Between Weight and HgbA1c.**

**A graph with black dots and red line

Description automatically generated**

**Model Summary: ( Cleaned Data Only) :**

summary(patregclean)

Call:

lm(formula = hgba1c ~ wt, data = cleaned\_data)

Residuals:

Min 1Q Median 3Q Max

-2.0329 -0.6297 -0.2983 0.1424 6.0257

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.955553 0.214665 23.085 < 2e-16 \*\*\*

wt 0.012452 0.002531 4.919 1.16e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.228 on 528 degrees of freedom

Multiple R-squared: 0.04383, Adjusted R-squared: 0.04202

F-statistic: 24.2 on 1 and 528 DF, p-value: 1.161e-06

**Figure 2: Formula and Residuals**

|  |
| --- |
| * **Formula: hgba1c ~ wt**. |
| * **Residuals :** ranged from a minimum of -2.0329 to a maximum of 6.0257, with a median of -0.2983, indicating the variance in the differences between observed and predicted HgbA1c levels. |

**Figure 3: Coefficients Analysis:**

* + **Using the formula Y= AX+B, where A=slope and B= Intercept**

|  |  |
| --- | --- |
| **Intercept** | The intercept, representing the expected HgbA1c level when weight is zero, was estimated at 4.955553 with a standard error of 0.214665. This intercept was significantly different from zero (t = 23.085, p < 2e16). |
| **Slope** | The slope coefficient for weight (wt) was 0.012452 with a standard error of 0.002531. This indicates that for each one unit increase in weight, the HgbA1c level is expected to increase by an average of 0.012452 units. This relationship was statistically significant (t = 4.919, p = 1.16e |

**Figure 4: Model Fit and Significance:**

|  |
| --- |
| **Residual Standard Error value :** 1.228 on 528 degrees of freedom. |
| **Multiple R-squared: value: 0.04383, and the adjusted R-squared was 0.04202**, suggesting that approximately 4.4% of the variability in HgbA1c is explained by the model. |
| **F-statistic :** for the overall model fit was 24.2 on 1 and 528 degrees of freedom, with a highly significant p-value of 1.161e-06, indicating that the model is statistically significant. |

**Interpretation:**

The results suggest a statistically significant but relatively small relationship between weight and HgbA1c levels. The positive coefficient for weight indicates that higher weight is associated with higher HgbA1c levels. However, given the small proportion of variance in HgbA1c explained by weight, other factors not included in this model also play a significant role in determining HgbA1c levels.

**Limitations:**

This model explains a small percentage of the variance in HgbA1c levels, indicating that other variables may also be important in predicting HgbA1c. The nature of the data and analysis is correlational, and as such, causality cannot be inferred from these findings. Overall, this regression model provides valuable insights into the relationship between weight and HgbA1c levels, although it does not conspicuously show the complexity of factors influencing HgbA1c; this is beyond the scope of this project. The mean imputation method is just one example on how to handle missing values in a data set. Other methods are available for use, i.e. k-nearest neighbors but they are not shown in this project. Selection of method depends on the nature of the data, the extent of the missingness, and the type of analysis. It is also important to understand the reason behind the missing data (missing at random, missing completely at random, or missing not at random) as it can influence the imputation strategy. Another limitation is that this project only demonstrates univariate analysis. For simplicity, multivariable and logistic regression are not included in this project; depending on the dataset and nature of the data, these types of regressions may be more appropriate models.

**Conclusion:**

In this project, using R for data wrangling and linear regression modeling revealed a modest correlation between weight and HgbA1c levels. The analysis, while informative, highlighted the model's limitations in explaining variance, suggesting the influence of other factors. Critically, the project emphasized the nuanced approach required in handling missing values, underscoring that techniques vary with the dataset and analysis type (e.g., multivariate, or logistic regression). This study underscores the importance of tailored methodologies in data analysis and paves the way for more comprehensive future research for deeper insights.