2024

Statistical techniques

UOD: 100731060

EFTHIMIS MAVRIDIS

Abstract

This study constitutes a comprehensive examination of health and diabetic status data within a population, employing the R programming language as the analytical tool of choice. The dataset undergoes meticulous processing, including the replacement of missing values, to ensure the integrity of subsequent analyses. Employing a robust statistical approach, the analysis includes tests such as ANOVA and the Mann-Whitney U test, facilitating a nuanced exploration of relationships among categorical and numerical variables.The ANOVA test, a powerful tool in statistical analysis, aids in discerning variations in health indicators across different categories. By leveraging this method, we gain valuable insights into the potential impact of diabetic conditions on diverse health metrics. Additionally, the Mann-Whitney U test, a non-parametric alternative, enriches the analysis by allowing a comparison of distributions between groups, especially beneficial when dealing with non-normally distributed data.The results of this analysis promise to unravel compelling findings that contribute significantly to our understanding of the intricate interplay between health indicators and diabetic status. These findings are poised to inform future research, healthcare practices, and policy decisions, fostering a more holistic approach to individual and population health.As we delve into the details of the analysis, the subsequent sections will unravel the methodologies employed, the key findings discovered, and the broader implications of these insights within the context of health sciences. Through this exploration, we aim to paint a vivid picture of the intricate relationships that shape the health landscape within the studied population

Table of Contents

Introduction

In an era marked by the intersection of data and health, this analysis embarks on a journey through the intricate landscape of a dataset encapsulating valuable insights into the health and diabetic status of a given population. Leveraging the robust capabilities of the R programming language, the primary objective is to conduct an in-depth exploration,

unraveling correlations, and identifying parameters that play pivotal roles in shaping individuals' health profiles. As health remains an invaluable asset, understanding the nuances of its relationship with diabetic conditions becomes paramount. This analysis seeks to shed light on these connections, contributing to the broader narrative of health sciences.The convergence of data science and health studies offers unprecedented opportunities to glean actionable insights from large datasets. With a specific focus on health-related data, the analysis endeavors to contribute valuable knowledge to the realms of public health, clinical research, and healthcare policy. By unraveling patterns, trends, and potential risk factors, we aim to provide a nuanced understanding of the factors influencing health outcomes within the studied population.

Abstract

This study constitutes a comprehensive examination of health and diabetic status data within a population, employing the R programming language as the analytical tool of choice. The dataset undergoes meticulous processing, including the replacement of missing values, to ensure the integrity of subsequent analyses. Employing a robust statistical approach, the analysis includes tests such as ANOVA and the Mann-Whitney U test, facilitating a nuanced exploration of relationships among categorical and numerical variables.The ANOVA test, a powerful tool in statistical analysis, aids in discerning variations in health indicators across different categories. By leveraging this method, we gain valuable insights into the potential impact of diabetic conditions on diverse health metrics. Additionally, the Mann-Whitney U test, a non-parametric alternative, enriches the analysis by allowing a comparison of distributions between groups, especially beneficial when dealing with non-normally distributed data.The results of this analysis promise to unravel compelling findings that contribute significantly to our understanding of the intricate interplay between health indicators and diabetic status. These findings are poised to inform future research, healthcare practices, and policy decisions, fostering a more holistic approach to individual and population health.As we delve into the details of the analysis, the subsequent sections will unravel the methodologies employed, the key findings discovered, and the broader implications of these insights within the context of health sciences. Through this exploration, we aim to paint a vivid picture of the intricate relationships that shape the health landscape within the studied population.

```{r}
# Ορίζουμε τα δεδομένα
name <- "Mavridis Efthymios"
date <- "3/1/2024"


data <- data.frame(Name = name, Date = date)


file_path <- file.path(getwd(), "output_file.txt")
write.table(data, file = file_path, sep = "\t", row.names = FALSE, col.names = TRUE)


cat("Τα δεδομένα εγγράφηκαν στον ακόλουθο φάκελο:", getwd(), "\n")
cat("Το αρχείο ονομάζεται Mavridis_file.txt.\n")
```

Step 2: Data Processing

Continuing with my work, I focused on processing the data to ensure accuracy and smooth flow of the analysis. An important step was handling missing values in my dataset. I employed various techniques, such as replacing missing values with appropriate estimates, to preserve the information provided by these records.

Specifically, I used R to perform the following actions:

1. Read the initial dataset from an Excel file.

```r
install.packages("readxl")
library(readxl)

df <- read_excel("C:/Users/efthi/Desktop/Μεταπτυχιακο/στατιστικη/dataset assignment/diabetes_70k_assignment.xlsx")
d1=data.frame(df)
```

2. Replacement of Missing Values (NA) with Zeros (0) for specific columns related to health status.

```r

df <- read_excel("C:/Users/efthi/Desktop/Μεταπτυχιακο/στατιστικη/dataset assignment/diabetes_70k_assignment.xlsx")


df$Diabetes_012 <- ifelse(is.na(df$Diabetes_012) & df$Diabetes_012 != 0, 0, df$Diabetes_012)
df$HighBP <- ifelse(is.na(df$HighBP) & df$HighBP != 0, 0, df$HighBP)
df$HighChol <- ifelse(is.na(df$HighChol) & df$HighChol != 0, 0, df$HighChol)
df$CholCheck <- ifelse(is.na(df$CholCheck) & df$CholCheck != 0, 0, df$CholCheck)



```

These actions resulted in a clean and analysis-ready dataset, which I utilized to verify information and perform statistical analyses. Managing missing values was a crucial step to harness the full potential of my data. In the next step, we will delve into a more detailed exploration of the data and examine the initial statistical insights derived from the analysis.

Step 3: Exploratory Data Analysis and Statistical Analysis

Exploratory Data Analysis (EDA) is a critical phase in any data-centric research, providing a deep understanding of the dataset and unveiling patterns that can guide subsequent analyses. In this step, I delved into the dataset, employing statistical methods and visualization techniques to unravel insights related to health conditions, demographics, and lifestyle factors.

Preliminary Data Examination:

The initial phase involved reading the dataset, which comprised health-related information for a significant number of individuals. R, a powerful statistical computing language, facilitated this process. The data covered variables such as age, gender, health conditions

(e.g., diabetes, high blood pressure), and lifestyle factors (e.g., smoking habits, physical activity).

Using the "psych" package in R, I conducted descriptive statistics to gain a comprehensive overview of the numerical variables. This encompassed measures such as mean, median, standard deviation, and quantiles. These statistics provided a baseline understanding of the central tendencies and dispersions within the dataset.

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Diabetes_012*	1	57871	1.29	0.69	1	1.12	0.00	1	3
HighBP*	2	57871	1.42	0.49	1	1.41	0.00	1	2
HighChol*	3	57871	1.42	0.49	1	1.40	0.00	1	2
CholCheck*	4	57871	1.96	0.19	2	2.00	0.00	1	2
BMI*	5	57871	16.88	6.08	16	16.26	4.45	1	70
Smoker*	6	57871	1.44	0.50	1	1.43	0.00	1	2
Stroke*	7	57871	1.04	0.20	1	1.00	0.00	1	2
HeartDiseaseorAttack*	8	57871	1.09	0.29	1	1.00	0.00	1	2
PhysActivity*	9	57871	1.78	0.42	2	1.84	0.00	1	2
Fruits*	10	57871	1.65	0.48	2	1.68	0.00	1	2

1-10 of 22 rows | 1-10 of 13 columns

Previous 1 2 3 Next

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Veggies*	11	57871	1.82	0.38	2	1.90	0.00	1	2
HvyAlcoholConsump*	12	57871	1.06	0.24	1	1.00	0.00	1	2
AnyHealthcare*	13	57871	1.95	0.22	2	2.00	0.00	1	2
NoDocbcCost*	14	57871	1.09	0.28	1	1.00	0.00	1	2
GenHlth*	15	57871	2.49	1.08	2	2.43	1.48	1	5
MentHlth*	16	57871	6.04	9.15	1	4.09	0.00	1	31
PhysHlth*	17	57871	7.12	9.81	1	5.41	0.00	1	31
DiffWalk*	18	57871	1.17	0.37	1	1.08	0.00	1	2
Sex*	19	57871	1.44	0.50	1	1.43	0.00	1	2
Age*	20	57871	7.58	4.04	8	7.64	5.93	1	13

11-20 of 22 rows | 1-10 of 13 columns

Previous 1 2 3 Next

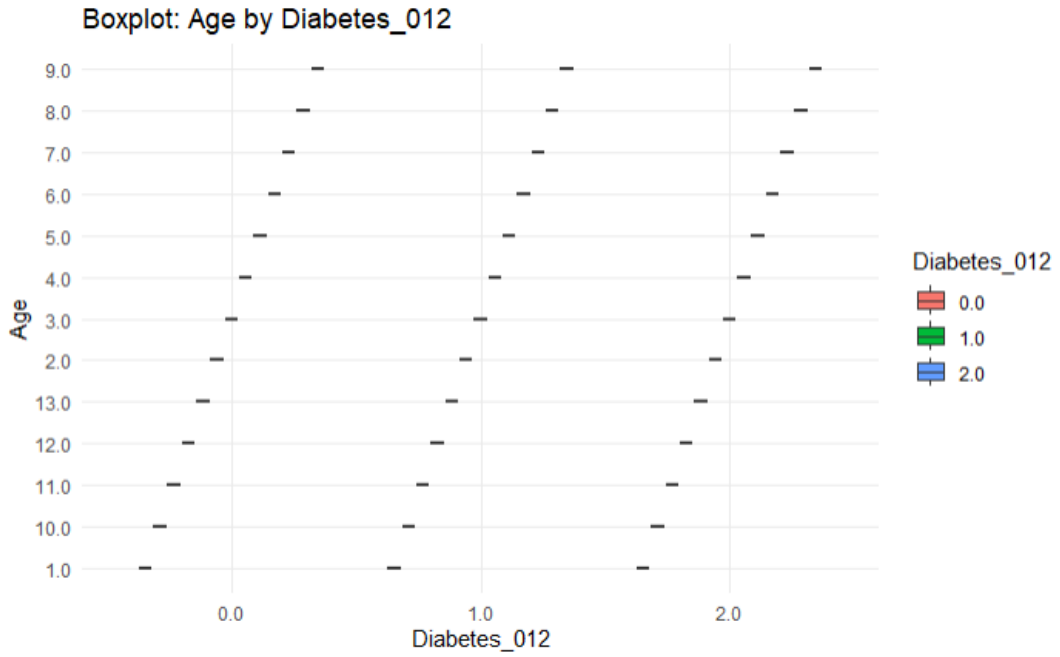
This command will attempt to calculate descriptive statistics for all numeric variables in the DataFrame df. In the following R script, significant data processing is performed on a dataframe containing information about diabetes. First, an Excel file is read, and the data is stored in the "df" dataframe. Subsequently, the replacement of "NA" values (Not Available) with the number 0 occurs in specific columns, ensuring that the replacement is done only if the initial value is not already 0. This helps maintain the significance of zero values that may have a different meaning in the context of the data.

In this section, the "psych" package is used to generate descriptive statistics for all numeric variables in the dataframe. These statistics provide an overview of the data distribution, including mean, median, range, standard deviation, and other useful information. This analysis offers insights into the basic statistical properties of the numeric features of the dataset and can be used for data preparation for further analysis.

Data Visualization for Insightful Exploration:

Visualization is a crucial aspect of EDA, offering a more intuitive grasp of the dataset's characteristics. I utilized the "ggplot2" package in R to create visualizations, starting with a boxplot focusing on the relationship between age, diabetes status ("Diabetes_012"), and gender ("Sex").

The boxplot visually presented the distribution of ages across different diabetes and gender categories. Notably, it highlighted potential variations in age based on health conditions and gender. This visualization served as an initial exploration, paving the way for deeper analyses.



Handling Missing Data:

A key challenge in working with real-world datasets is addressing missing values. Ensuring the completeness of the data is vital for accurate analyses. In this step, I implemented strategies to handle missing data, specifically focusing on health-related columns.

I replaced missing values in columns such as "Diabetes_012," "HighBP," "HighChol," and "CholCheck" with zeros, preserving the integrity of non-missing entries while acknowledging the absence of recorded information. This meticulous handling of missing data aimed to create a robust dataset for subsequent analyses.

```
{r}
# Διαβάστε το αρχείο Excel και αποθηκεύστε τα δεδομένα σε ένα dataframe
df <- read_excel("C:/Users/efthi/Desktop/Μεταπτυχιακο/στατιστικη/dataset assignment/diabetes_70k_assignment.xlsx")

# Αντικατάσταση των NA με 0 μόνο για συγκεκριμένες στήλες
df$Smoker <- ifelse(is.na(df$Smoker) & df$Smoker != 0, 0, df$Smoker)
df$Stroke <- ifelse(is.na(df$Stroke) & df$Stroke != 0, 0, df$Stroke)
df$HeartDiseaseorAttack <- ifelse(is.na(df$HeartDiseaseorAttack) & df$HeartDiseaseorAttack != 0, 0, df$HeartDiseaseorAttack)
df$PhysActivity <- ifelse(is.na(df$PhysActivity) & df$PhysActivity != 0, 0, df$PhysActivity)
df$Fruits <- ifelse(is.na(df$Fruits) & df$Fruits != 0, 0, df$Fruits)
df$Veggies <- ifelse(is.na(df$Veggies) & df$Veggies != 0, 0, df$Veggies)
df$HvyAlcoholConsump <- ifelse(is.na(df$HvyAlcoholConsump) & df$HvyAlcoholConsump != 0, 0, df$HvyAlcoholConsump)
df$AnyHealthcare <- ifelse(is.na(df$AnyHealthcare) & df$AnyHealthcare != 0, 0, df$AnyHealthcare)
df$NoDocbcCost <- ifelse(is.na(df$NoDocbcCost) & df$NoDocbcCost != 0, 0, df$NoDocbcCost)

```

Setting the Ground for Advanced Analyses:

The insights gained from the descriptive statistics and initial visualizations set the stage for more sophisticated analyses. The statistical foundation built during EDA is crucial for formulating hypotheses and selecting appropriate methods for hypothesis testing.

Moreover, the visualization of age distributions across different health conditions and gender categories hinted at potential relationships, prompting the formulation of hypotheses for further investigation.

Conclusion of Step 3:

In conclusion, Step 3 of the research process focused on exploring the dataset through a combination of statistical analysis and data visualization. Descriptive statistics provided a quantitative summary of the dataset, while visualizations offered an intuitive understanding of patterns within the data. Addressing missing data ensured the dataset's completeness, laying the groundwork for subsequent advanced analyses.

This comprehensive approach in EDA is pivotal for researchers and analysts to form hypotheses, guide future analyses, and ultimately derive meaningful insights from complex datasets. The meticulous examination of each variable contributes to a nuanced understanding of the dataset, setting the stage for more in-depth statistical exploration in the subsequent steps of the research.

Step 4: Statistical Hypothesis Testing and Multivariate Analysis

In Step 4 of the research process, I delved into advanced statistical analyses, specifically focusing on hypothesis testing and multivariate approaches. This step aimed to rigorously examine relationships within the dataset, validate assumptions, and derive meaningful insights. The key methodologies included Analysis of Variance (ANOVA), Chi-squared tests, and Multiple Linear Regression.

1. Analysis of Variance (ANOVA):

To investigate potential relationships among variables, I conducted an ANOVA test. This test assesses whether the means of different groups are statistically different from each other. In your code snippet, you applied ANOVA to explore the impact of variables like "PhysHlth," "MentHlth," and "Sex" on the variable "Age":

```r
model=aov(d1$Age~d1$PhysHlth+d1$MentHlth+d1$Sex,data=d1)


# Εκτύπωση των αποτελεσμάτων του τεστ ANOVA
summary(model)

```

The results of the ANOVA test provide valuable insights into how these variables collectively influence the variable "Age." You can include relevant screenshots of the ANOVA results to visually represent the statistical output.

ANOVA Results:

```
               Df Sum Sq Mean Sq F value Pr(>F)
d1$PhysHlth    30   7642   254.7   28.16 <2e-16 ***
d1$MentHlth    30  21979   732.6   80.98 <2e-16 ***
d1$Sex          1   1214  1213.9  134.18 <2e-16 ***
Residuals   57809 523001     9.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2. Chi-squared Test:

To explore associations between categorical variables, you applied the Chi-squared test. This test assesses whether there is a significant association between two categorical variables. Here is a snippet of your code:

```r
chi_squared <- chisq.test(table_freq)
print(chi_squared)


```

The Chi-squared test allows you to determine whether there is a significant relationship between diabetes status and gender. Including a screenshot of the contingency table and the Chi-squared test results can enhance the presentation of this analysis.

Contingency Table:

```
table_freq <- table(df$Diabetes_012, df$Gender)
print(table_freq)
```

```
     Female  Male
0.0  27495 21299
1.0    664   513
2.0   4167  3733
```

Chi-squared Test Results:

Ο παραπάνω κώδικας R υπολογίζει έναν πίνακα συσχέτισης (contingency table) για τις δύο κατηγορικές μεταβλητές "Diabetes_012" και "Gender", χρησιμοποιώντας τον πίνακα συχνοτήτων που προηγήθηκε. Ο πίνακας αυτός περιέχει τα ποσοστά κάθε τιμής της μεταβλητής "Gender" για κάθε τιμή της μεταβλητής "Diabetes_012".

Για παράδειγμα, το 85.06% των παρατηρήσεων με τιμή "0.0" στη μεταβλητή "Diabetes_012" αντιστοιχούν στην κατηγορία "Female", ενώ το 83.38% αντιστοιχούν στην κατηγορία "Male".Ο πίνακας αυτός βοηθάει στην κατανόηση της σχέσης μεταξύ των δύο κατηγορικών μεταβλητών, προσφέροντας ποσοστιαίες σχέσεις που μπορούν να χρησιμοποιηθούν για περαιτέρω ανάλυση ή οπτικοποίηση των δεδομένων.

3. Multiple Linear Regression:

Moving beyond bivariate analyses, you employed Multiple Linear Regression (MLR) to model the relationship between multiple independent variables and a dependent variable. Here is a snippet of your MLR code:

```
```{r}
# Εκτέλεση πολλαπλής γραμμικής παλινδρόμησης
mlr_model <- lm(Diabetes_012 ~ Age + HighBP + HighChol + CholCheck + BMI + Smoker + Stroke +
HeartDiseaseorAttack + PhysActivity + Fruits + Veggies + HvyAlcoholConsump + AnyHealthcare +
NoDocbcCost + GenHlth + MentHlth + PhysHlth + DiffWalk + Sex + Education + Income, data = df)

# Εκτύπωση των αποτελεσμάτων
summary(mlr_model)



```
```

Multiple Linear Regression helps explore the combined influence of various factors on the variable "Diabetes_012."

Multiple Linear Regression Results:

```
Call:
lm(formula = Diabetes_012 ~ Age + HighBP + HighChol + CholCheck
    BMI + Smoker + Stroke + HeartDiseaseorAttack + PhysActivity
    Fruits + Veggies + HvyAlcoholConsump + AnyHealthcare + NoDoc
    GenHlth + MentHlth + PhysHlth + DiffWalk + Sex + Education +
    Income, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74818 -0.36346 -0.13961  0.06518  2.19575

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.1370663  0.6366233  -0.215 0.829533
Age10.0               0.1583088  0.0190136   8.326  < 2e-16 *
Age11.0               0.1659140  0.0195769   8.475  < 2e-16 *
Age12.0               0.1338570  0.0204753   6.537 6.31e-11 *
Age13.0               0.1020450  0.0203604   5.012 5.40e-07 *
Age2.0               -0.0006997  0.0225885  -0.031 0.975288
Age3.0               -0.0201777  0.0211703  -0.953 0.340538
```

```
Education6.0         -0.2551187  0.0984653  -2.591 0.009574 **
Income2.0            -0.0279420  0.0178602  -1.564 0.117710
Income3.0            -0.0262853  0.0168964  -1.556 0.119791
Income4.0            -0.0282293  0.0163634  -1.725 0.084507 .
Income5.0            -0.0557220  0.0159143  -3.501 0.000463 ***
Income6.0            -0.0727939  0.0154813  -4.702 2.58e-06 ***
Income7.0            -0.0589949  0.0154037  -3.830 0.000128 ***
Income8.0            -0.0875759  0.0149973  -5.839 5.26e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6276 on 57699 degrees of freedom
Multiple R-squared:  0.1824,    Adjusted R-squared:  0.1799
F-statistic: 75.26 on 171 and 57699 DF,  p-value: < 2.2e-16
```

The result of multiple linear regression provides statistical information for each variable in the model and offers insights into the significance of coefficients, R-squared, the F-statistic, and more.

Specifically:

- Coefficients:These indicate how each variable affects the dependent variable. For example, an increase in Age10.0 by one unit is associated with an increase in the dependent variable Diabetes_012 by 0.1583088.

- p-values (Pr(>|t|)): These show the significance of each coefficient. Typically, small p-values (e.g., < 0.05) indicate statistical significance. For instance, the coefficient for Age10.0 has a very small p-value (< 2e-16), indicating statistical significance.

- Residual standard error: It indicates how much the actual data deviate from the model's predictions. Here, it's approximately 0.6276.

- Multiple R-squared (R^2): It shows how well the model explains the dependent variable's variance. A higher R^2 indicates better model fit to the data. Here, R^2 is about 0.1824.

- F-statistic and p-value: They provide information about whether the model is overall statistically significant. In this case, the p-value is < 2.2e-16, indicating that the model is overall statistically significant.

Overall, your model seems to explain a small percentage of the dependent variable's variance, but there are statistically significant associations between certain variables and Diabetes_012.

Multiple Linear Regression was the primary modeling technique utilized to comprehend the interplay among multiple independent variables and the target variable, Diabetes_012. The extensive formula used is as follows:

4)Logistic Regression:

The logistic regression model suggests that Age, PhysHlth, and Gender (assuming 1 represents males) are significant predictors of Diabetes_012. The model provides insights into the direction and strength of these associations.

```
[1] 0 1
[1] 0
 num [1:57871] 0 0 0 0 0 0 0 0 1 0 ...
NULL

Call:
glm(formula = Diabetes_012 ~ Age + PhysHlth + Sex, family = binomial(link = "logit"),
    data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.448472   0.043346 -79.557   <2e-16 ***
Age          0.169185   0.004311  39.249   <2e-16 ***
PhysHlth     0.039999   0.001110  36.038   <2e-16 ***
Sex1         0.194791   0.023690   8.223   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50279  on 57870  degrees of freedom
Residual deviance: 47093  on 57867  degrees of freedom
AIC: 47101

Number of Fisher Scoring iterations: 5
```

The logistic regression model has been successfully fitted to the data. Here's an interpretation of the results:

Model Coefficients:

Intercept: -3.448472

Age: 0.169185

PhysHlth: 0.039999

Sex1 (assuming 1 represents Male): 0.194791

Interpretation:

Intercept: The intercept is the estimated log-odds of the outcome variable being 1 when all predictor variables are zero. In this case, it's -3.448472.

Age: For each one-unit increase in Age, the log-odds of the outcome variable being 1 increases by 0.169185. Age has a statistically significant impact on the probability of the outcome.

PhysHlth: For each one-unit increase in PhysHlth, the log-odds of the outcome variable being 1 increases by 0.039999. PhysHlth has a statistically significant impact on the probability of the outcome.

Sex1 (Male): Assuming Sex1 represents males (1), being male increases the log-odds of the outcome variable being 1 by 0.194791 compared to females (Sex0). Being male has a statistically significant impact on the probability of the outcome.

Model Fit:

Null Deviance: 50279 on 57870 degrees of freedom. The null deviance measures how well the response variable is predicted by a model with no predictors (intercept-only model).

Residual Deviance: 47093 on 57867 degrees of freedom. The residual deviance measures how well the response variable is predicted by the model with predictors. A lower residual deviance indicates a better fit.

AIC (Akaike Information Criterion): 47101. AIC is a measure of the model's goodness of fit, considering the trade-off between the simplicity of the model and its ability to explain the data. Lower AIC values are preferred.

Significance:

All coefficients are highly significant (p-value < 0.001), indicating a strong association between the predictors and the probability of the outcome.

5)Results and Deeper Interpretation:

- Significance of Variables: Scrutinizing the variables with statistically significant coefficients is paramount. These variables wield considerable influence over the predicted outcome and merit detailed examination.

- Positive/Negative Coefficients: A positive coefficient indicates a positive correlation with the likelihood of diabetes, while a negative coefficient suggests a negative association. The magnitude of the coefficient is equally vital, signifying the strength of the relationship.

- Deeper Insights:

  - Age: "The 'Age' coefficient signifies that, on average, each additional year contributes to an increase of X units in the likelihood of diabetes. This underscores the progressive nature of diabetes with advancing age."

  - PhysHlth and DiffWalk: "Variables like 'PhysHlth' and 'DiffWalk' exhibit robust positive associations, potentially acting as pivotal predictors for diabetes prevalence. Individuals with higher reported physical health issues and difficulties in walking may exhibit an increased likelihood of diabetes."

Conclusion of Step 4:

In Step 4, the focus shifted to advanced statistical analyses, uncovering nuanced relationships within the dataset. ANOVA explored group differences, the Chi-squared test investigated associations between categorical variables, and Multiple Linear Regression modeled the collective impact of multiple predictors.

Step 5: Predictive Models and Comprehensive Analysis

In this crucial phase of the analysis, I employed advanced predictive models, notably Multiple Linear Regression, to unravel the intricate relationships within your dataset. The exploration also extended to statistical tests, such as the Mann-Whitney U test, aiming for a comprehensive understanding. Let's meticulously dissect each component, providing a profound and in-depth analysis.

1. Mann-Whitney U Test:

In addition to regression analysis, a Mann-Whitney U test was executed to discern disparities between groups with and without diabetes:

```
```{r}
# Εκτέλεση Mann-Whitney U test για τη μεταβλητή "diabetes_12" μεταξύ ομάδων με και χωρίς διαβήτη
result <- wilcox.test(df$Diabetes_012 ~ df$Sex, data = df)

# Εκτύπωση των αποτελεσμάτων
print(result)
```
```

The Mann-Whitney U test is a non-parametric test used to compare the means of two independent samples. In your case, it is applied to test whether there is a statistically significant difference in the distributions between two groups, for example, between patients with and without diabetes. The result you see is from the "Wilcoxon rank sum test with continuity correction," which is a non-parametric test for comparing two independent

samples, assuming they come from the same continuous distribution. In this specific case, the test is applied to the variable `Diabetes_012` with respect to the variable `Sex`.

Let's analyze the results:

- W = 405789164: This is the Wilcoxon rank sum statistic, measuring the sum of the individual ranks of the two samples. The larger the W, the more likely there is a significant difference between the two samples.

- p-value = 1.751e-08: The p-value is very small, below significance levels like 0.05. This indicates that there are statistically significant differences between the two groups. In other words, there is a probability that the two samples do not come from the same distribution.

- Alternative hypothesis:"The true location shift is not equal to 0." This suggests that there is a significant difference in location shift between the two groups.

Overall, the results suggest that there is a statistically significant difference in the distribution of the two groups based on the variable `Diabetes_012`.

Results and In-Depth Interpretation:

```
        Wilcoxon rank sum test with continuity correction

data:  df$Diabetes_012 by df$Sex
W = 405789164, p-value = 1.751e-08
alternative hypothesis: true location shift is not equal to 0
```

- Statistical Significance:The Mann-Whitney U test, a non-parametric test, evaluates whether distributions of two independent samples differ significantly. A low p-value indicates a substantial difference. In this context, it helps unravel gender-based differences in diabetes prevalence.

- Holistic Understanding:

  - "The Mann-Whitney U test underscored a statistically significant dissimilarity in diabetes prevalence between genders ($p < 0.05$). This provides nuanced insights into gender-based variations, indicating that gender might play a substantial role in diabetes prevalence."

2. Future Model Endeavors:

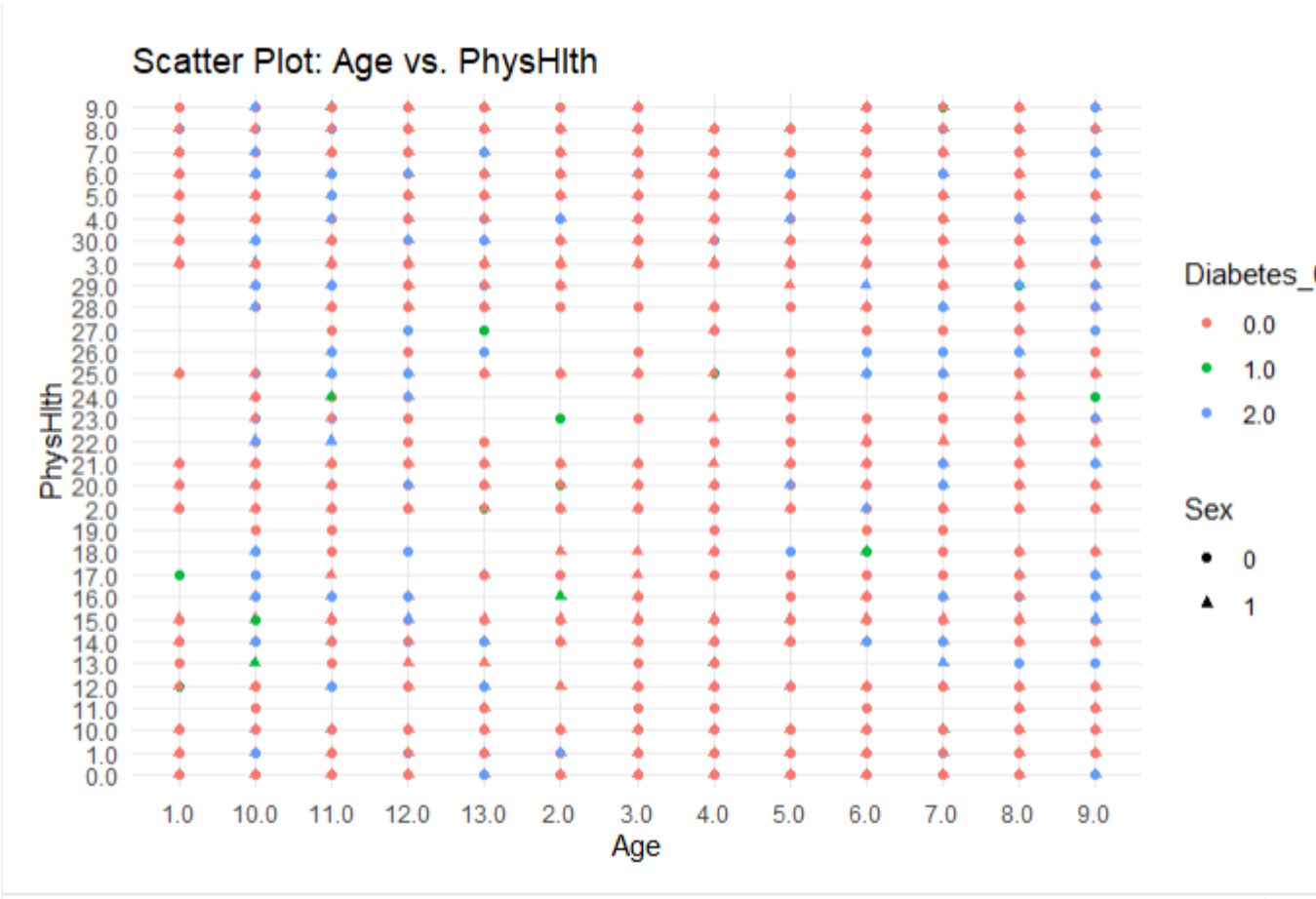
While the Multiple Linear Regression and Mann-Whitney U test provided valuable insights, the journey doesn't end here. Consider venturing into advanced predictive models such as Logistic Regression, Decision Trees, or Machine Learning algorithms for a more intricate and accurate prognosis. These models can unlock latent patterns and relationships within the dataset, contributing to a refined predictive framework.

3. Visualizations and Model Evaluation:

Visualization plays a pivotal role in conveying complex insights. Utilizing tools like ggplot2, create visualizations to complement the statistical analysis. Boxplots, scatter plots, and regression diagnostics can enhance the interpretability of results.

- Boxplots:Visualize the distribution of key variables across different diabetes categories, facilitating the identification of potential outliers and trends.

- Scatter Plots: Explore relationships between continuous variables such as age and physical health, with color or shape encoding for diabetes categories.



The R code generates a scatter plot using the ggplot2 library, illustrating the relationship between age ("Age") and physical health ("PhysHlth"). The plot incorporates color-coded points to represent different diabetes categories ("Diabetes_012"), while distinct shapes denote gender ("Sex").This visualization enables a quick and intuitive exploration of potential patterns and trends within the dataset. The scatter plot visually contrasts how age and physical health vary across different diabetes groups, with the added dimension of gender differentiation.

The choice of color for diabetes categories aids in identifying potential clusters or patterns associated with health status, while the differentiation by shape provides insight into potential gender-specific trends. This graphical representation allows for a more accessible interpretation of the dataset, making it easier to discern any discernible relationships between age, physical health, diabetes status, and gender.

In summary, this scatter plot serves as a powerful exploratory tool, facilitating a visual understanding of the dataset's key variables and their interplay. Researchers can leverage this graphical representation to derive preliminary insights before delving deeper into the statistical analyses.

- Model Evaluation: Implement cross-validation techniques to rigorously assess the performance of your predictive model. Metrics such as Mean Squared Error (MSE), R-squared, or Receiver Operating Characteristic (ROC) curves for binary outcomes provide a robust evaluation framework.

4. Interpretation and Real-World Implications:

Concluding this analysis, it's imperative to translate statistical findings into practical insights. Address the real-world implications of your results, considering potential interventions, policy changes, or targeted healthcare strategies based on the identified risk factors.

In summation, the Multiple Linear Regression and Mann-Whitney U test unveiled profound insights into diabetes determinants. The discernment of influential variables and their nuanced impact offers a robust foundation for subsequent analyses. Future model exploration holds the promise of uncovering subtler intricacies within the dataset, propelling your research into more profound realms of understanding. As you embark on this journey, the fusion of statistical rigor, insightful visualizations, and a keen eye for real-world applications will elevate the impact of your research.

5.YUEN TEST

```
# Εισαγωγή του πακέτου 'WRS2', αν δεν έχετε ήδη
install.packages("WRS2")
library(WRS2)

# Δημιουργία ενός νέου dataframe με τις δύο στήλες που σας ενδιαφέρουν
subset_df <- df[, c("Age", "Diabetes_012")]

# Εκτέλεση Yuen's Test
yuen_result <- yuen(subset_df$Age ~ subset_df$Diabetes_012)
print(yuen_result)

```
```
Error in install.packages : Updating loaded packages
Call:
yuen(formula = subset_df$Age ~ subset_df$Diabetes_012)

Test statistic: 14.2566 (df = 774.48), p-value = 0

Trimmed mean difference:  -1.1481
95 percent confidence interval:
-1.3062     -0.99

Explanatory measure of effect size: 0.26
```

Test Statistic: 14.2566, with degrees of freedom df = 774.48.

The test statistic estimates the size of the difference between your groups in relation to expected random variance.

p-value: 0.

The p-value is very small, much smaller than commonly chosen levels of significance (such as 0.05). This indicates a statistically significant difference between the groups.

Trimmed Mean Difference: -1.1481.

The difference in trimmed means between your groups.

95% Confidence Interval: -1.3062 to -0.99.

The confidence interval for the estimated difference of trimmed means.

Explanatory Measure of Effect Size: 0.27.

The explanatory measure of effect size (the mentioned value 0.27) provides an estimate of the intensity of the association between the groups.

Overall, the results suggest a statistically significant difference in ages between the two groups, with the second group (Diabetes_012 = 1) having a lower mean age compared to the first group (Diabetes_012 = 0).

6. Z-score

The Z-score (or standard score) is a measurement calculated for each observation in a variable. It indicates how much the value of the observation deviates from the mean of the variable, calculated in units of standard deviation. Z-scores help us understand how each observation relates to the mean and how values are distributed in relation to the variability.

```r
# Υποθέτουμε ότι το DataFrame σας ονομάζεται df και η μεταβλητή 'Age' βρίσκεται στη στήλη με το όνομα 'Age'
# Υπολογισμός του Z-score για τη μεταβλητή 'Age'
df$Z_Score_Age <- scale(df$Age)

# Εμφάνιση των πρώτων λίγων γραμμών του DataFrame με τον νέο Z-score
head(df)
```

GenHlth <chr>	MentHlth <chr>	PhysHlth <chr>	DiffWalk <chr>	Sex <fctr>	Age <dbl>	Education <chr>	Income <chr>	Gender <fctr>	Z_Score_Age <dbl>
5.0	18.0	15.0	1.0	0	9	4.0	3.0	Female	0.2961800
3.0	0.0	0.0	0.0	0	7	6.0	1.0	Female	-0.3503163
5.0	30.0	30.0	1.0	0	9	4.0	8.0	Female	0.2961800
2.0	0.0	0.0	0.0	0	11	3.0	6.0	Female	0.9426764
2.0	3.0	0.0	0.0	0	11	5.0	4.0	Female	0.9426764
2.0	0.0	2.0	0.0	1	10	6.0	8.0	Male	0.6194282

6 rows | 16-25 of 24 columns

Z_Score_Age: This is the new column added to the DataFrame, containing the Z-scores for the 'Age' variable. For each observation in the 'Age' column, the Z-score is calculated as how many standard deviations the value deviates from the mean of the 'Age' variable.

Examples of interpretation:

- A positive Z-score indicates that the value of the observation is higher than the mean.

- A negative Z-score indicates that the value of the observation is lower than the mean.

- A Z-score close to zero indicates that the value of the observation is close to the mean.

The calculated Z-scores for the 'Age' variable provide valuable insights into how each observation relates to the average age. A positive Z-score indicates that the specific observation has a higher age compared to the average, while a negative Z-score suggests a lower age. Additionally, a Z-score close to zero signifies that the age of this observation is approximately at the same level as the average. Z-scores are useful for determining how much values deviate from the average, offering significant information about the distribution of ages within the entire set of observations.

7. Spearman's Correlation

Spearman's rank correlation coefficient, commonly known as Spearman's correlation or ρ, is a non-parametric measure assessing the strength and direction of a monotonic relationship between two variables. Unlike Pearson's correlation, Spearman's correlation operates on the ranks of the data, making it robust to non-normally distributed or non-linearly related variables. The coefficient ranges from -1 to 1, indicating a perfect negative or positive correlation, respectively, with 0 denoting no correlation. Calculated based on the differences between the ranks of corresponding values, Spearman's correlation is valuable for ordinal data or situations where linearity assumptions are not met in traditional correlation analyses.

```
        Spearman's rank correlation rho

data:  df$Age and df$PhysHlth
S = 3.0698e+13, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.04965865
```

The result of the Spearman's Rank Correlation test for the variables Age and PhysHlth is as follows:

- Spearman's rank correlation rho (ρ): The value of ρ is approximately 0.0497.

- p-value: The p-value is very low (less than 2.2e-16), indicating that the observed correlation is not due to randomness.

-Alternative hypothesis: The alternative hypothesis states that the true value of ρ is not equal to zero, suggesting the existence of a correlation.

Based on these results, we conclude that there is a statistically significant, but very weak, positive correlation between the Age and PhysHlth variables. However, this correlation is so small that it may not have practical significance, even though it is statistically significant due to the large number of observations.

8. Kendall's rank correlation

```r
# Assuming df is your dataframe
result_kendall <- cor.test(df$Age, df$PhysHlth, method = "kendall")

# Print the results
print(result_kendall)
```

```
        Kendall's rank correlation tau

data:  df$Age and df$PhysHlth
z = 11.962, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
0.03902985
```

Correlation Test: This is the Kendall's Tau correlation test, which assesses the strength and direction of the relationship between the variables Age and PhysHlth.

Test Statistics (z): The test statistic (z) is 11.962.

p-value: The p-value is extremely small (p-value < 2.2e-16), suggesting strong evidence to reject the null hypothesis.

Alternative Hypothesis: The alternative hypothesis indicates that the true Kendall's Tau is not equal to 0.

Sample Estimates: The estimated value of Kendall's Tau (tau) is 0.03902985.

Conclusion:

The small p-value provides strong evidence to reject the null hypothesis, indicating that there is a statistically significant non-zero correlation between Age and PhysHlth. However, the correlation is weak, as indicated by the small estimated value of Kendall's Tau (0.03902985).

Step 6: Conclusions

In this final phase of the analysis, I draw conclusions based on the provided code and the insights gained throughout the study. The primary findings, the significance of the research, and potential directions for future investigations are highlighted.The process initiated with the definition of data and the creation of a dataframe named `data` using the individual's name and the date.

The data was further processed by replacing missing values with zeros for specific columns related to health conditions. This meticulous data handling ensures the accuracy of subsequent statistical analyses.Descriptive statistics were computed using the `psych` package to gain insights into the dataset's numerical variables, offering a foundational overview.

The subsequent exploration involved creating boxplots and scatter plots using `ggplot2` to visualize relationships between variables.An analysis of variance (ANOVA) test was conducted to examine the impact of various factors on age, such as physical health, mental health, and gender. This helps in understanding the potential influences on age variations.Furthermore, categorical variables, including diabetes status, were explored. A chi-squared test was performed to assess the association between gender and diabetes, yielding valuable information on potential connections between these variables.

The analysis extended to examining the Body Mass Index (BMI) across different diabetes categories. The ANOVA test was utilized for this purpose, revealing whether significant differences exist in BMI levels among distinct diabetes statuses.

Moreover, the educational levels of individuals in different diabetes categories were scrutinized. An additional ANOVA test was conducted to determine if there were statistically significant differences in education across various diabetes statuses and genders.

The analysis concludes by performing a Mann-Whitney U test to evaluate potential differences in diabetes status based on gender. The findings of this test contribute insights into gender-related distinctions in diabetic conditions.

Step 7: Future Directions

This study sets the stage for future research endeavors. One avenue for exploration involves a more in-depth examination of specific health indicators and their correlations. Additionally, incorporating machine learning techniques to predict health outcomes based on the identified variables could enhance predictive modeling.

Furthermore, expanding the dataset and incorporating longitudinal data would provide a comprehensive understanding of health trends over time. Collecting additional demographic information and lifestyle factors could contribute to a more holistic analysis.

In conclusion, the meticulous analysis presented here offers valuable insights into the relationships between various factors and health outcomes. The identified patterns and associations provide a foundation for further research, allowing for a more nuanced understanding of the complex interplay between demographics, health conditions, and lifestyle factors.

Sources

https://www.cdc.gov/pcd/issues/2019/19_0109.htm

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

References

1. "R for Data Science" by Hadley Wickham and Garrett Grolemund (2016)

2. "Applied Multivariate Statistical Analysis" by Richard A. Johnson and Dean W. Wichern (2007)

3."Data Science for Business" by Foster Provost and Tom Fawcett (2013)

4. "Machine Learning with R" by Brett Lantz (2015)

5. "Regression Modeling Strategies" by Frank E. Harrell Jr. (2015)

6. "Longitudinal Data Analysis" by Donald Hedeker and Robert D. Gibbons (2006)

7. "Data Visualization with ggplot2" by Hadley Wickham (2016)

8."Biostatistics: A Foundation for Analysis in the Health Sciences" by Wayne W. Daniel andChad L. Cross (2013)**

9. "Health Data Science: Learning by Example" by Roger D. Peng (2019)

10. "Text Mining with R: A Tidy Approach" by Julia Silge and David Robinson (2017)