Determining Risk Factors for Diabetes

A STATISTICAL REPORT

By

Isioma Clement

Contents

Question 1: Introduction

The diabetes dataset comprises data on a variety of individual health-related characteristics, including age, pregnancy, blood pressure, skin thickness, insulin levels, and diabetes pedigree function, which may or may not influence whether an individual is diabetic. However, the dataset is incomplete, with missing entries represented by 0s. In the following steps, we will examine the problematic nature of missing entries represented by 0s in this dataset.

Method

To determine the extent of missing entries in the dataset, the number of missing entries for each column in the dataset excluding the "Outcome" variable was examined. The 0s in the outcome column were excluded due to the nature of the variable. By counting the number of 0s in each column, the number of missing entries was calculated.

Result

The analysis showed that there are a total of 763 missing entries in the dataset. This amount represents approximately 11% percent of the entire data frame and this missingness can be observed in varying degrees in 6 out 9 of the columns of the dataset. The significant number of missing entries poses a huge challenge and can impact the accuracy of statistical analyses, as a lot of statistical models perform better on complete observations (Kang, 2013). Incorrect conclusions or an overestimation of the frequency of the values in the dataset may also result from this. A major problem is in the reduction of the statistical power of tests (Enders, 2022), that is, a reduction in the likelihood of finding differences in the data when they exist. Missing data can also introduce bias (Graham, 2009), although this largely depends on the type of missingness present in the data. Also, given the number of observations present in the dataset, the relative amount of missing data present may result in one carrying out analyses on a sample that is not the best representation of the population of interest (Gelman & Hill, 2006).

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Total
111	5	35	227	374	11	0	0	763

Table 1. Missing values by category

Discussion

The presence of missing entries in a dataset can lead to several problems. One of the primary issues is that the presence of missing entries can skew statistical analyses and affect the accuracy of the results (Schafer, 2002).

In the case of bias, using Insulin as an example, it is possible that during the collection of this data, the device may have failed to detect the insulin levels of people with very low insulin, resulting in an underestimation of insulin levels in the dataset. As a result, assumptions concerning the association between insulin levels and other health indicators in the dataset may be incorrect.

There are several ways to deal with missing data based on the type of missingness present in the data. Data can be missing either completely at random(MCAR), at random(MAR), or not at random(MNAR) (Graham, 2009). As mentioned earlier, since the performance of most statistical models is influenced by the completeness of data, it would be useful to examine the nature of missingness in the data so as to decide on the best way to handle the missing data. This could help mitigate the problems mentioned.

Question 2: Introduction

The instruction requires us to discard the "Pregnancies" variable and to assume that there are no missing data points in the "Outcome" variable. Additionally, we are told to clean the data so that missing entries in other variables become NaN values, so that we can do statistical tests without any problems. We will outline the procedures used to clean the data in the next steps and provide the solution code.

Method

The “Pregnancies” column was removed using the select() function from the dplyr package which allows the subsetting of columns using their names. The replace_with_na_at() function from the naniar package was used to clean the data by passing in the necessary arguments so that the 0s were replaced as required.

Results

i) The dimension of the original diabetes dataset consisted of 768 rows and 9 columns. By removing this variable, the resulting data frame now consists of 768 rows and 8 columns.

Code to discard the pregnancy variable:

```
cut_diabetes = select(diabetes, -Pregnancies)
```

ii) To ensure that the 0s present in the remaining variables of the dataset excluding the “Outcome” variable get treated as missing, they are converted to NA (not available). This is a data cleaning step necessary to perform statistical tests without issues as suggested in the instructions.

Code to clean data so that missing entries are changed to NA values:

```
clean_diabetes = replace_with_na_at(cut_diabetes, .vars = c("Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "DiabetesPedigreeFunction", "Age"), ~.x == 0)
```

Discussion

The cleaning of a dataset is a crucial preliminary step in analysing data. The use of NA values instead of 0s ensure that missing entries are recognized as such and do not influence the results of the analyses.

It should be noted that even though the instruction explicitly states that the missing entries be recoded to NaN, they have been transformed to NA instead.

Question 3: Introduction

This section aims to visualize the distributions of each variable in the dataset and provide their summary statistic. It is important to visualize the distributions of variables in a dataset and to show their summary statistics because it provides a comprehensive understanding of the data. These methods provide a quick and easy way to get an overview of the data, including measures of central tendency and variability. It can also be used to help spot outliers and skewness.

Method

Histograms were created to visualize the distributions of the variables in the dataset. A bar plot was also created for the binary variable Outcome. Using both the summary() and describe() functions in R, the summary statistics were generated. The summary statistics gotten for each variable include measures of central tendency, spread, and measures of skewness.

Result

i) The figure below presents the visualization of each variable in the dataset.

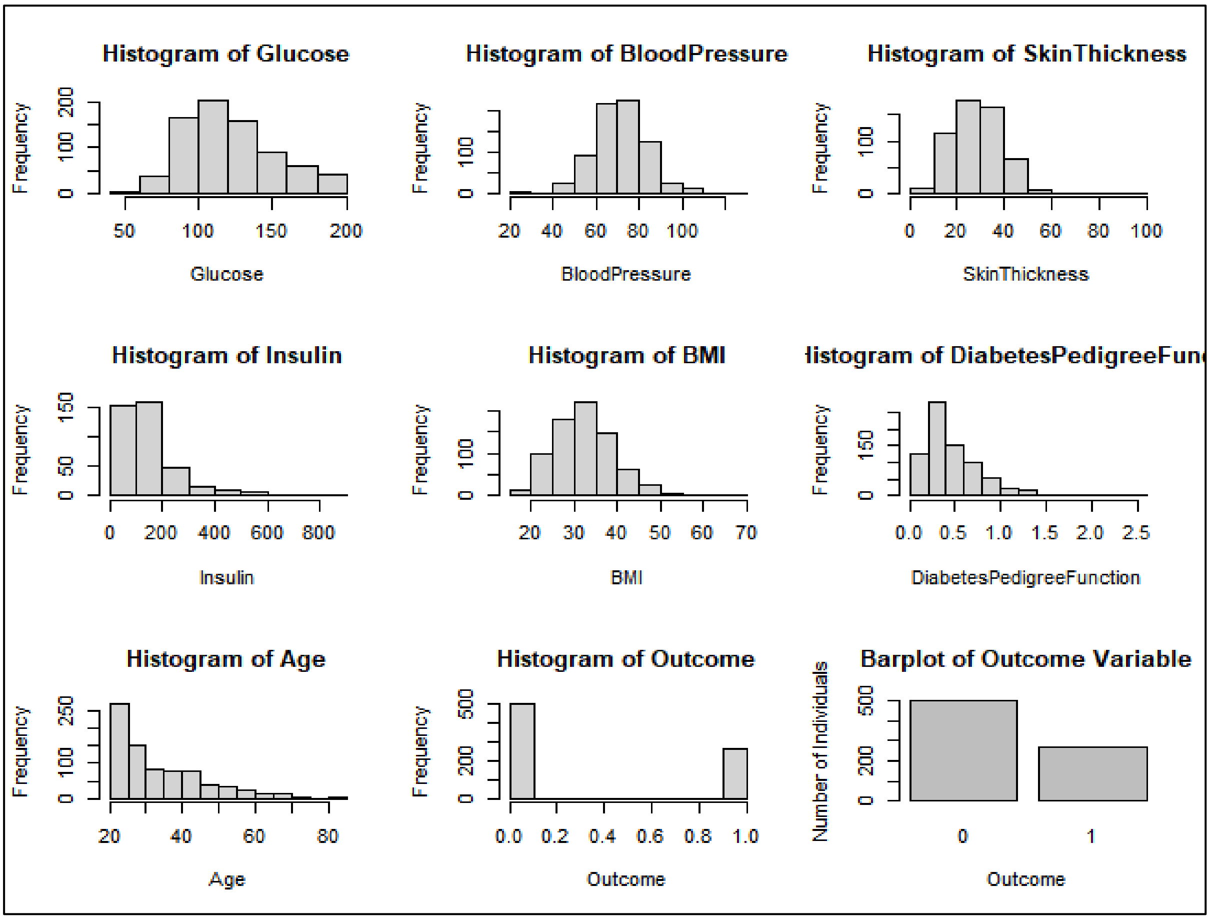

Figure 1. Visual Representation of the Distribution of all Variables in the Diabetes Dataset

ii) Summary Statistics:

The summary statistics for each variable are presented in the following table:

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
Min.	44	24	7	14	18.2	0.078	21	0
1st Qu.	99	64	22	76.25	27.5	0.2437	24	0
Median	117	72	29	125	32.3	0.3725	29	0
Mean	121.7	72.41	29.15	155.55	32.46	0.4719	33.24	0.349
3rd Qu.	141	80	36	190	36.6	0.6262	41	1
Max.	199	122	99	846	67.1	2.42	81	1
Range	155	98	92	832	48.9	2.342	60	1
Std.	30.54	12.38	10.48	118.78	6.92	0.33	11.76	0.48

Variance	932.69	153.26	109.83	14108.69	47.89	0.11	138.30	0.23
Skew	0.53	0.13	0.69	2.15	0.59	1.91	1.13	0.63
NA's	5	35	227	374	11	0	0	0

Table 2. Summary statistics for all the variables in the data.

The distribution of the binary outcome variable appears to be imbalanced, with 65.1% of observations having an outcome of 0 (no diabetes) and 34.9% having an outcome of 1.

Glucose has a mean of 121.7 and a median of 117, with a range of 155.

Blood pressure has a mean of 72.41 and a median of 72 and a standard deviation of 12.38. Its range is 98. The skewness is 0.53.

SkinThickness has close mean and median values of 29.15 and 29 respectively and a range of 92.

Insulin is quite peculiar, with a large difference in mean and median (155.55 and 125 respectively). Its standard deviation is 118.78 and it has a huge range of 832.

The results of all the other variables are equally displayed in the table above (Table 2).

Discussion:

The distributions of Glucose, BloodPressure, SkinThickness, and BMI appear to be roughly normally distributed. Evidence of this can also be seen in their skew values which are relatively close to zero. The Insulin, DiabetesPedigreeFunction, and Age variables are highly skewed to the right, suggesting a greater deviation from a normal distribution.

In summary, it can be observed from the results that the distributions of the variables in the dataset vary widely in shape and range. The range of values is particularly wide for insulin levels. The dataset contains missing values for some variables.

Question 4: Introduction

This section investigates whether there are significant differences in central tendencies of predictor variables with respect to the diabetes outcome. Following the instructions, we assume that all the independent variables are independent of one another. Checking for these significant differences in central tendencies is important to help identify which variables are most strongly associated with the outcome.

Method

The Kolmogorov-Smirnov test was used to check the normality assumption of the data before conducting the difference test. The test result determined whether a parametric or non-parametric test was suitable for analysis. The independent two–sample t-test was employed for variables that satisfied the normality assumption, while the Mann-Whitney U test was used for variables that failed the normality test. The significance level chosen for the tests is 0.005. Box plots were created to visualize differences in the central tendencies of the predictor variables between diabetic and non-diabetic groups.

The null and alternative hypotheses for the statistical test for differences are thus;

H_0 = there is no significant difference between the groups being compared

H_1 = there is a significant difference between the groups being compared.

Results

The results of the statistical tests are summarized in the table below:

Variable	p-values
Glucose	<2.20E-16
Blood Pressure	3.97E-06
Skin Thickness	1.83E-09
Insulin	7.48E-14
BMI	<2.20E-16
DiabetesPedigreeFunction	1.20E-06
Age	<2.20E-16

Table 3. p-values of the statistical test for differences for differences in central tendencies.

The boxplots below visualize the differences in central tendencies between the two diabetes outcome groups for each predictor variable:

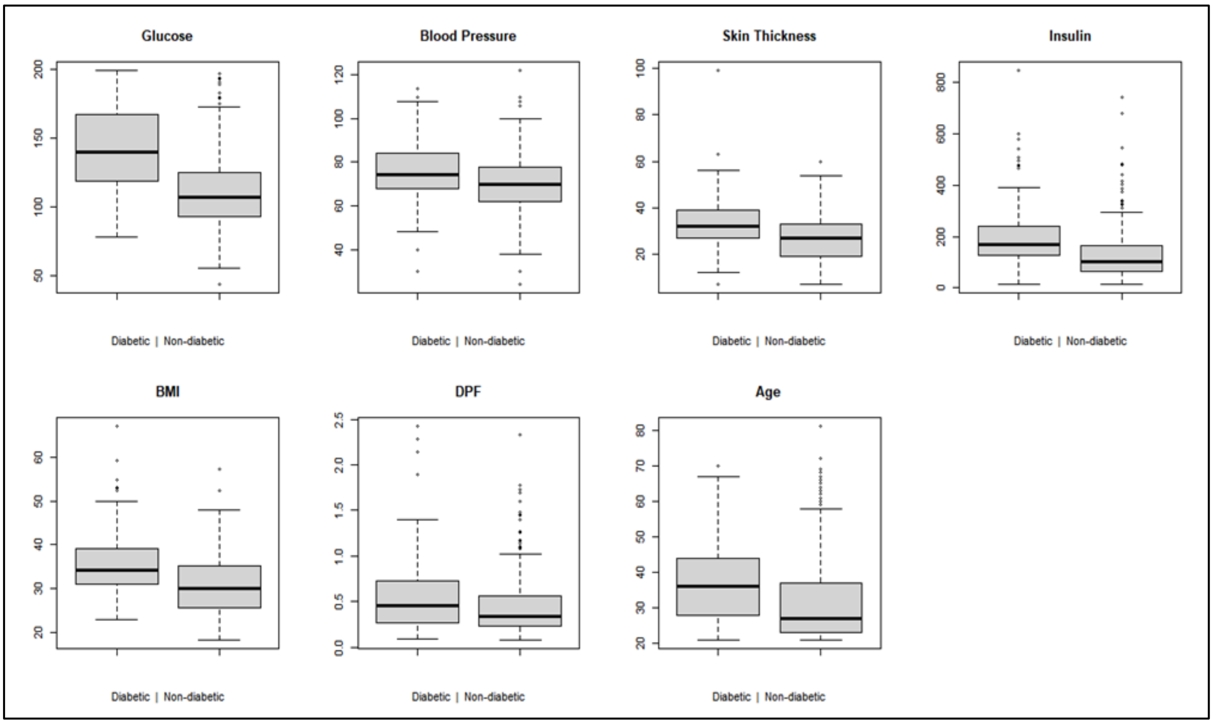

Figure 2. Paired boxplots illustrating differences in central tendencies of the variable with respect to the outcome.

The null hypothesis for our statistical test for differences in central tendencies is that there is no difference in central tendencies of the variables with respect to the diabetes outcome. After conducting statistical tests to compare the central tendencies of the predictor variables with respect to the diabetes outcome, we obtained p-values below the threshold ($p < 0.005$), leading to the rejection of the null hypothesis that there is no difference in central tendencies. This suggests that there is evidence to support the conclusion that the means of the diabetic and non-diabetic groups are significantly different for all predictor variables.

Discussion

The results from the analysis suggest that the predictor variables are useful in distinguishing between individuals with and without diabetes. The boxplots provide a clear visualization of the differences in central tendencies between the two diabetes outcome groups for each predictor variable. The predictors with the largest differences in mean between the two groups are Glucose, Insulin, BMI, and Age, suggesting that these predictors may be particularly important for predicting the outcome.

If there are significant differences in the central tendencies of a predictor variable for diabetic and non-diabetic individuals, this may suggest that the predictor variable is an important factor in determining diabetes status. This information can be used to build more accurate predictive models for diabetes. Further investigations may consider exploring the relationship between predictor variables and diabetes outcome, taking into account potential confounding factors.

Question 5: Introduction

This section aims to test the assumption that all predictor variables are independent using correlation coefficients. Testing for correlation is important in this dataset because it helps to understand the relationship between the predictor variables and the outcome variable. It is particularly useful in determining whether any predictors have a high correlation with one another (Tabachnick, et al., 2013), which may result in problems like multicollinearity in certain regression models (Field, et al., 2012).

Method

Spearman's Rank correlation coefficient method was used to calculate the correlation coefficients between all possible pairs of predictor variables. The magnitude of the correlation coefficient (rho) was used to determine the strength of the association, which ranged from -1 to 1. The p-value was used to assess the statistical significance of the correlations. p-values of less than 0.05 were considered statistically significant. Earlier visualizations of the data showed that not all variables in the dataset were normally distributed and that some variables contained outliers. Consequently, Spearman's correlation method was selected as the preferred correlation test because it is a non-parametric test and has greater resistance to the influence of outliers.

Result

The figure below shows the correlation between variables in the dataset through a scatterplot and displays the values of their correlation coefficient (Spearman).

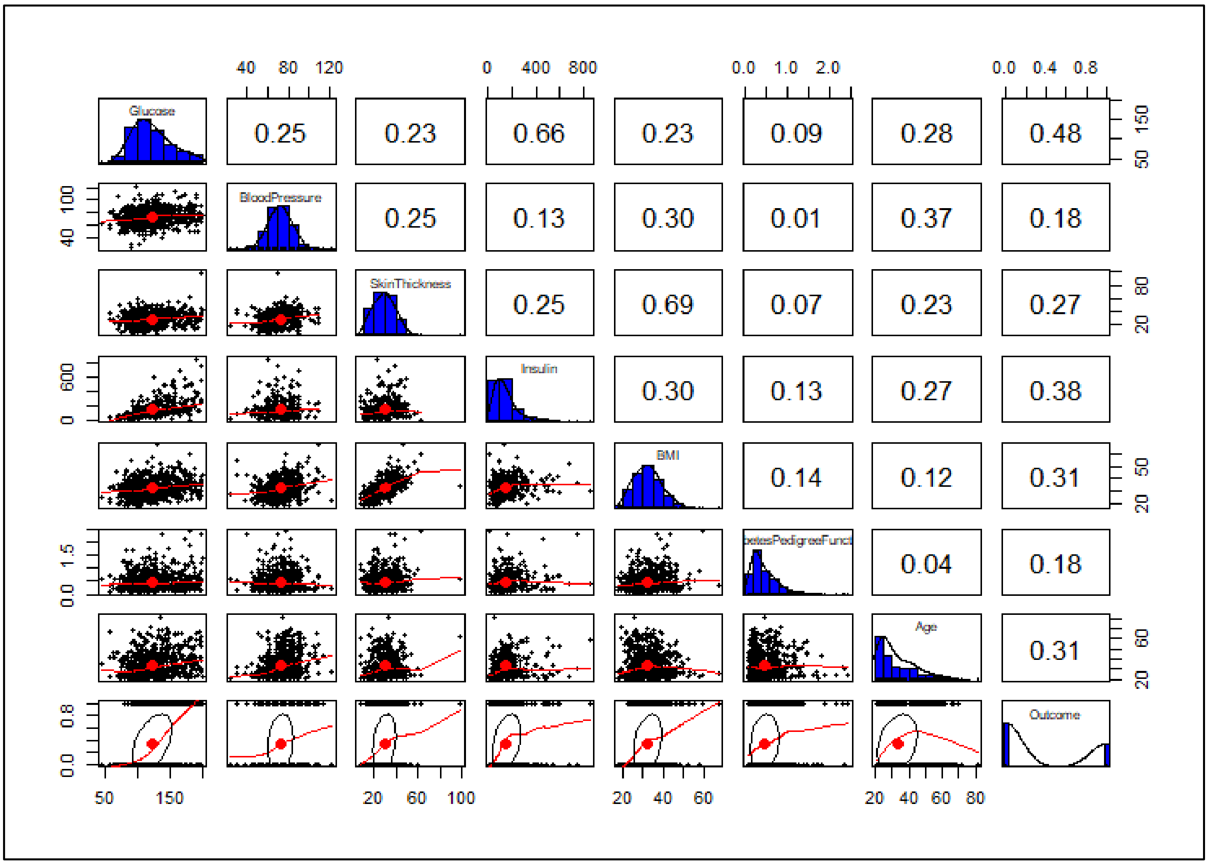

Figure 3. Visualization to show correlation among all the variables in the data.

The table below gives the p-values for the correlation coefficients.

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age
Glucose		8.72E-12	7.43E-08	0.00E+00	2.64E-10	0.012296	1.55E-15
BloodPressure	8.72E-12		5.34E-09	9.53E-03	0.00E+00	0.818058	0.00E+00
SkinThickness	7.43E-08	5.34E-09		8.36E-07	0.00E+00	0.081443	5.41E-08
Insulin	0.00E+00	9.53E-03	8.36E-07		8.08E-10	0.009641	7.04E-08
BMI	2.64E-10	0.00E+00	0.00E+00	8.08E-10		0.000172	8.40E-04
DPF	1.23E-02	8.18E-01	8.14E-02	9.64E-03	1.72E-04		2.35E-01
Age	1.55E-15	0.00E+00	5.41E-08	7.04E-08	8.40E-04	0.234941	

Table 4. The p-values gotten from the test for correlation.

The results of the correlation analysis showed that there were statistically significant correlations between many of the predictor variables. Specifically, Glucose was positively correlated with Blood Pressure (rho = 0.25, p < .00), Skin thickness (rho = 0.23, p < .005), Insulin (rho = 0.66, p < .005), BMI (rho = 0.23, p < .001), Age (rho = 0.28, p < .001), and Outcome (rho = 0.48, p < .001). BloodPressure was positively correlated with SkinThickness (rho = 0.25, p < .001), BMI (rho = 0.30, p < .001), and Age (rho = 0.37, p < .001), but not significantly correlated with DiabetesPedigreeFunction (rho = 0.01, p = 0.818). SkinThickness was positively correlated with Insulin (rho = 0.25, p < .001), BMI (rho = 0.69, p < .001), Age (rho = 0.23, p < .001), and Outcome (rho = 0.27, p < .001). Insulin was positively correlated with BMI (rho = 0.30, p < .001), Age (rho = 0.27, p < .001), and Outcome (rho = 0.38, p < .001). BMI was positively correlated with DiabetesPedigreeFunction (rho = 0.14, p = 0.001) and Age (rho = 0.12, p = 0.006), and Outcome (rho = 0.31, p < .001). Age was positively correlated with

Outcome (rho = 0.31, p < .001). It is also worth noting that DiabetesPedigreeFunction was not significantly correlated with other variables except BMI and Outcome.

Discussion

The results of the correlation analysis suggest that many of the predictor variables in the dataset are correlated with each other. Specifically, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Age, and Outcome were all found to be correlated with at least one other variable in the dataset. This means that these variables may not be independent predictors of the outcome variable. One way to look at this could also be this – following the question, if we are to proceed with an assumption that all the predictor variables are independent, the null hypothesis in such case would be that the variable are not independent of each other. Hence, finding correlation between at least one variable and another could be enough grounds to fail to reject the null hypothesis as evidence suggest that the predictor variables may be dependent. Therefore, it is important to consider the potential effects of multicollinearity when building models that use these predictor variables. To assess whether these correlations are strong enough to indicate dependence between the predictor variables, it would be important to consider the context of the data and the specific question being asked (Jackson, et al., 2009). In some cases, even weak correlations may be meaningful and important, while in other cases, strong correlations may not be relevant (Agresti & Finlay, 2018). Additionally, it is important to note that correlation does not necessarily imply causation, and there may be other factors at play that are influencing the relationships between the predictor variables (Peng & Matsui, 2016). Visualizing the results using scatter plots may help to identify nonlinear relationships and provide more insight into the nature of the correlations between the variables.

Question 6: Introduction

The objective of this section is to identify variables that have a significant influence on diabetes using the appropriate regression model. Identifying risk factors for diabetes can help in the development of effective strategies for the prevention and management of the disease.

Method

Based on the binary categorical nature of the Outcome variable, the Binary Logistic regression model was chosen to model the relationship between the predictor and outcome variables (Hosmer Jr, et al., 2013). The model was created using the glm() function in the R and then specifying the dependent and independent variables. The alpha value was chosen to be 0.05.

After modelling, we then proceeded to identify the influential valuables and ascertain the goodness of fit of the model using the Hosmer-Lemeshow test. The chi-squared test was also used to test for the differences in deviance. The assumptions of the logistic regression were also checked.

Result

The table below displays the results gotten from the binary logistic model.

Coefficients:				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-1.02E+01	1.21E+00	-8.409	< 2e-16
Glucose	3.82E-02	5.78E-03	6.605	3.97E-11
BloodPressure	-1.09E-03	1.17E-02	-0.092	0.926379
SkinThickness	1.17E-02	1.72E-02	0.681	0.495593
Insulin	-9.42E-04	1.33E-03	-0.71	0.477683
BMI	6.66E-02	2.71E-02	2.456	0.014046

DiabetesPedigreeFunction	1.08E+00	4.23E-01	2.551	0.010729
Age	5.20E-02	1.43E-02	3.652	0.000261

Table 5. Binary logistics regression model coefficients.

The logistic regression model found that glucose, BMI, diabetes pedigree function, and age were significant predictors of diabetes with p-values < 0.05. Specifically, the odds of having diabetes increased by a factor of exp(0.0382) = 1.039 for every one-unit increase in glucose, holding all other variables constant. Similarly, for every one-unit increase in BMI, the odds of having diabetes increased by a factor of exp(0.0666) = 1.068. For every one-unit increase in diabetes pedigree function, the odds of having diabetes increased by a factor of exp(1.08) = 2.94. Finally, for every one-unit increase in age, the odds of having diabetes increased by a factor of exp(0.052) = 1.054. The other predictor variables, including blood pressure, skin thickness, and insulin levels, were not found to be significant predictors of diabetes.

After testing the difference in null and residual deviance using the chi-square test, a very small p-value (<0.05) was obtained. Therefore, the null hypothesis that the model did not provide a better fit than an intercept-only model was rejected in favour of the alternative hypothesis that the model does provide a better fit than an intercept-only model.

Furthermore, the Hosmer-Lemeshow goodness of fit test produced a test statistic of X-squared = 4.8231 with 8 degrees of freedom and a p-value of 0.7763. Since the p-value was greater than the alpha level of 0.05, we failed to reject the null hypothesis that there is no statistically significant difference between the observed and expected values of the outcome variable. Thus, we accept that the model fits the data well, as there is no evidence to suggest otherwise.

Model Assumptions

We assumed that the observations in the data were independent, based on the nature of the data. To check for multicollinearity, we examined the Variance Inflation Factor (VIF) of the variables. All VIF values were less than 2, indicating that there is little to no evidence of a high correlation between the independent variables, based on the general rule of thumb of VIF > 5 suggesting a high level of linearity (James, et al., 2013). Thus, we conclude that the assumption of multicollinearity in the logistic regression model is satisfied.

However, the logistic regression model assumes that there are no influential outliers in the data (Fox, 2015). Unfortunately, using Cook’s distance plot, it was observed that there was a significant number of influential outliers in the data, indicating the violation of this assumption. Additionally, the linearity assumption was not fully satisfied as we observed instances of non-linearity between some of the predictor variables and the logit of the outcome variable.

Discussion

The logistic regression model developed suggests that Glucose, BMI, Diabetes Pedigree Function, and Age are significant variables that have an influence on diabetes. Glucose has the highest coefficient estimate, indicating that it has the most significant impact on diabetes. BMI and DiabetesPedigreeFunction also have a significant impact, indicating that body weight and genetic factors play a role in diabetes. Age has a moderate impact, indicating that the risk of diabetes increases with age. However, blood pressure and Insulin were not significant in the model, indicating that they have little to no impact on diabetes. These findings suggest that lifestyle interventions that focus on maintaining a healthy weight and regulating blood sugar levels may help prevent diabetes. The logistic regression model demonstrated good predictive ability, as indicated by the relatively small residual deviance and AIC values. However, the assumptions of logistic regression were also checked, and some violations were found, such as the non-linearity of some predictors, and the presence of influential outliers which may affect the validity of the model. Additionally, the model assumes independence of observations, which was satisfied in this analysis. Overall, this analysis provides valuable insights into the variables that have an influence on diabetes.

Question 7: Introduction

In this section, a simple linear regression model is used to predict the glucose levels for the missing entries. The regression model assumes that Glucose levels are dependent only on age. In this section, we will discuss the methods employed, the results obtained, and the assumptions of the linear regression model.

Method

A simple linear regression model was chosen to perform the prediction on the missing entries in Glucose. The lm() function in R was used to fit a linear regression model to the data and the predict() function was used to get the required values. The model assumptions, including linearity, normality, and homoscedasticity were checked to ensure that the model was appropriate for the data. the chosen alpha value is 0.05.

Results

The table below displays the p-value and estimates of the coefficient of the linear regression model.

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	98.63245	3.19767	30.845	< 2e-16
Age	0.69292	0.09061	7.647	6.21E-14
Multiple R-squared:	0.07136			
p-value:	6.21E-14			

Table 6. The coefficients of the linear regression model.

Age	22	21	22	37	41
Predicted Glucose	113.877	113.184	113.877	124.270	127.042

Table 7. The predicted glucose levels for the missing values based on age.

The linear regression model provided a significant result, with an intercept of 98.63, indicating that the predicted Glucose level when Age is zero is 98.63. The coefficient for Age is 0.69, indicating that for every unit increase in Age, the predicted Glucose level increases by 0.69. The R-squared value was 0.07136, indicating that only 7.14% of the variability in Glucose levels could be explained by Age. The p-value of < 0.05 suggests that the Age variable is significant in predicting Glucose levels.

Model Assumptions

Linearity: A scatter plot of Glucose versus Age did not show any obvious curvature and the points appeared to cluster around a line that indicated a positive association between the two variables. Therefore, it was concluded that the linearity assumption was met.

Normality: The normal probability plot of the residuals was approximately normal, with no obvious deviations from normality. Therefore, we concluded that the normality assumption was met.

Homoscedasticity: This assumption was checked by examining the scatterplot of the residual versus the fitted values. There was no obvious pattern in the residuals, indicating that the errors were independent. Therefore, we concluded that the assumption was met.

Independent observations: It was decided that the assumption of independent observations is met by considering the context of what each entry in the data represented.

Overall, we conclude that the assumptions of the linear regression model were satisfied in this analysis.

The figure below contains several plots used to check whether the necessary assumptions were met.

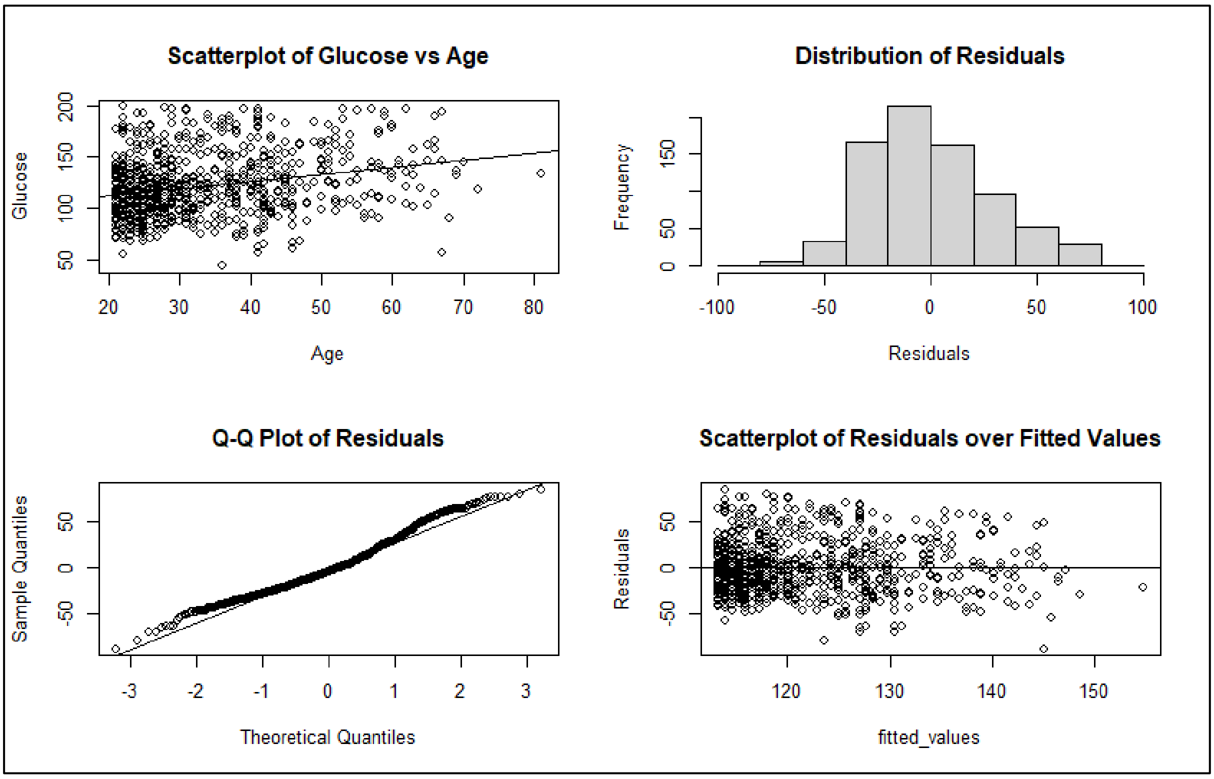

Figure 4. Visualization to show the results from checking assumptions of the linear regression model.

Discussion

Overall, the results of this study suggest that Age is a statistically significant predictor of Glucose levels. However, the low R-squared value of 0.07 suggests that only a small proportion of the variation in Glucose levels can be explained by Age, indicating that other factors may also play a role in predicting Glucose levels. Therefore, future studies may benefit from including additional variables in the model, such as body mass index or the diabetes Pedigree function, to better understand the factors that contribute to Glucose levels. The linear regression model developed in this analysis satisfactorily met all the assumptions for the regression model.

Conclusion

This paper aimed to perform a series of statistical tests on a dataset containing variables related to diabetes. The analyses carried out included checking the distribution and summary statistics, testing for differences in central tendencies between the groups in the data based on the diabetes outcome, and performing correlation tests on the variables. Appropriate regression models were also built to detect variables that had an influence on diabetes and to predict glucose levels based on age. The report considered the presence of missing entries in the dataset and explored its implications for the analyses. An alpha value of 0.05 was chosen as the significance level for all the tests in this study.

The distributions of the variables in the dataset varied widely in shape and range. Secondly, we found that there was a significant difference in central tendencies between the groups in the data based on the diabetes outcome suggesting that the predictor variables were useful in distinguishing between individuals with and without diabetes. Further analysis on correlation of the data suggested that the variables may not be independent predictors of the Outcome variable. The logistic regression model developed suggested that Glucose, BMI, Diabetes Pedigree Function, and Age were significant variables influencing the outcome variable with Glucose being the most significant. It should however be noted that due to the violation of certain assumptions of the model – notably, no influential outliers and linearity – the validity and reliability of the model may be questionable. Lastly, the linear regression model for determining glucose levels based on age, predicted an increase in glucose levels by 0.69 for every unit rise in age. However, the low R-squared value of 0.07 hints at the existence of other confounding factors that may be important in the prediction of glucose levels.

In summary, this analysis has identified important variables that influence diabetes and provided insights into the relationship between age and glucose levels. These findings could be useful in the prevention and management of diabetes. The key takeaway points are the importance of cleaning data, checking data distribution, carrying out correlation tests, and building appropriate regression models to gain insights into the relationships that exist between variables in a dataset.

Bibliography

Agresti, A. & Finlay, B., 2018. *Statistical methods for the social sciences.* 5 ed. s.l.:Pearson Education Limited .

Angrist, J. D. & Pischke, J.-S., 2015. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton, New Jersey: Princeton University Press .

Devore, J., 2012. *Probability and Statistics for Engineering and the Sciences.* 8 ed. Boston,: Brooks/Cole.

Enders, C. K., 2022. *Applied missing data analysis.* 2 ed. s.l.:Guilford Publications.

Field, A., Miles, J. & Field, Z., 2012. *Discovering statistics using R.* s.l.:Sage publications.

Fox, J., 2015. *Applied regression analysis and generalized linear models.* 3 ed. s.l.:Sage Publications.

Gelman, A. & Hill, J., 2006. *Data analysis using regression and multilevel/hierarchical models.* s.l.:Cambridge university press.

Graham, J. W., 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology,* Volume 60, pp. 549-576.

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X., 2013. *Applied logistic regression.* 3 ed. s.l.:John Wiley & Sons.

Jackson, D. L., Gillaspy Jr, J. A. & Purc-Stephenson, R., 2009. Reporting practices in confirmatory factor analysis: an overview and some recommendations.. *Psychological methods,* 14(1), p. 6.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An introduction to statistical learning.* New York: Springer.

Kang, H., 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology,* 64(5)(May 24, 2013), pp. 402-406..

Kim, J., 2019. *How to Choose the Level of Significance: A Pedagogical Note.* [Online]
Available at: https://mpra.ub.uni-muenchen.de/id/eprint/66373
[Accessed 12 March 2023].

Miller, M., 2020. *Towards Data Science.* [Online]
Available at: https://towardsdatascience.com/statistical-thinking-understanding-correlation-5f7c63934699
[Accessed 9 March 2023].

Peng, R. D. & Matsui, E., 2016. *The Art of Data Science: A guide for anyone who works with Data.* s.l.:Skybrude consulting LLC.

Schafer, J. L. a. G. J. W., 2002. Missing data: our view of the state of the art.. *Psychological methods,* Volume 7, p. 147.

Tabachnick, B. G., Fidell, L. S. & Ullman, J. B., 2013. *Using multivariate statistics.* 6 ed. s.l.:pearson Boston, MA.