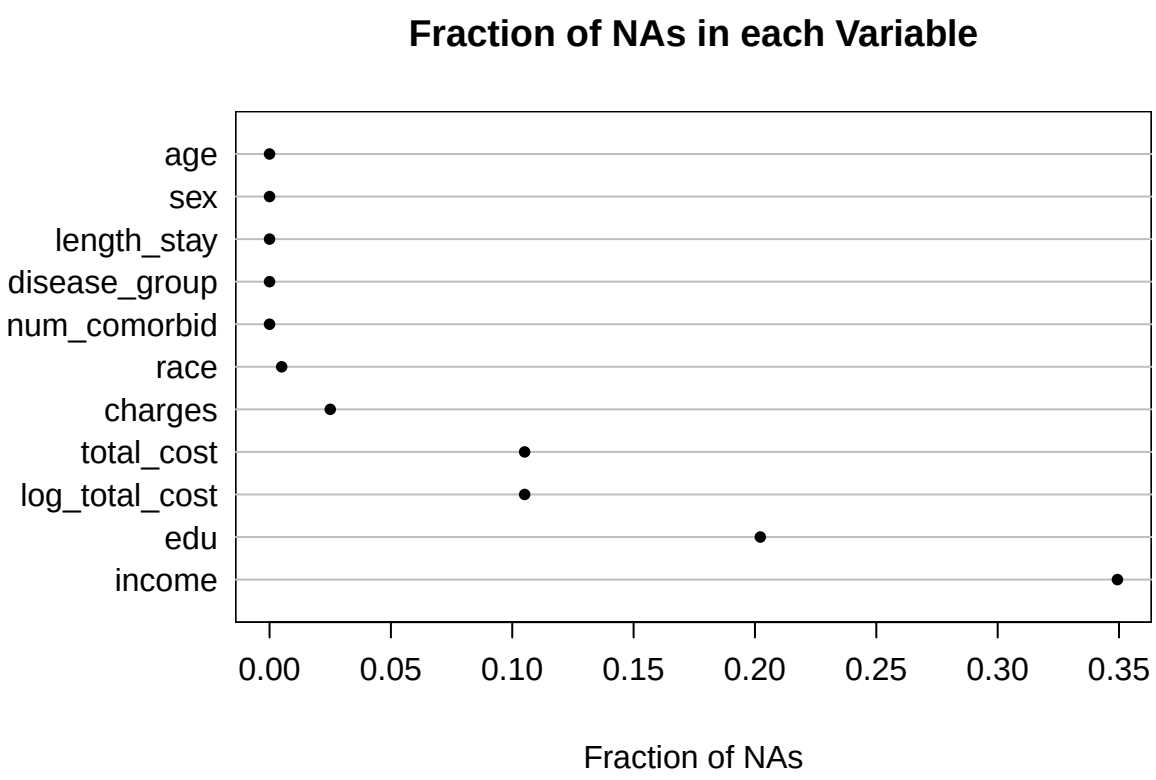
Age vs Hospital Charges: EDA with Hypothesis Testing

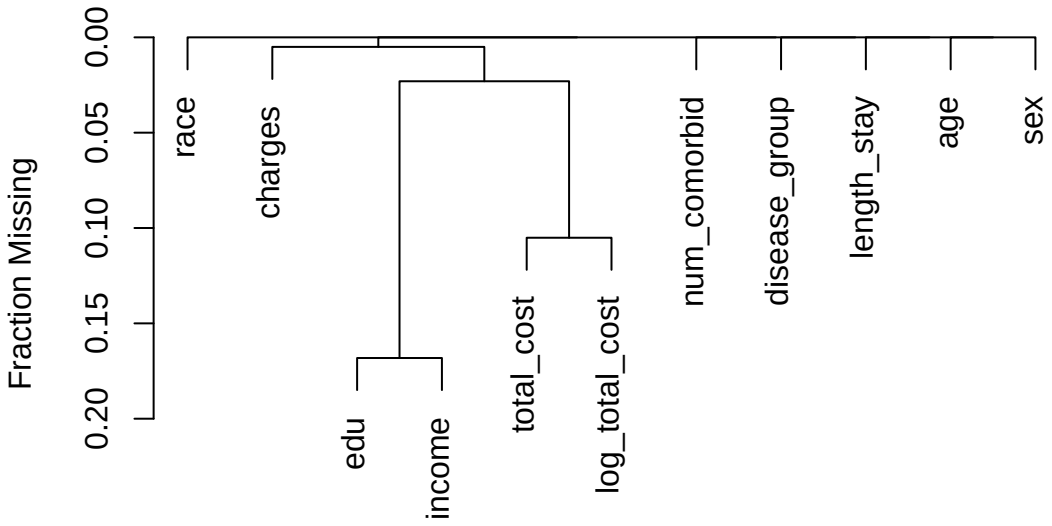
Kreesh Rajani

PART - 1 Loading and cleaning the dataset

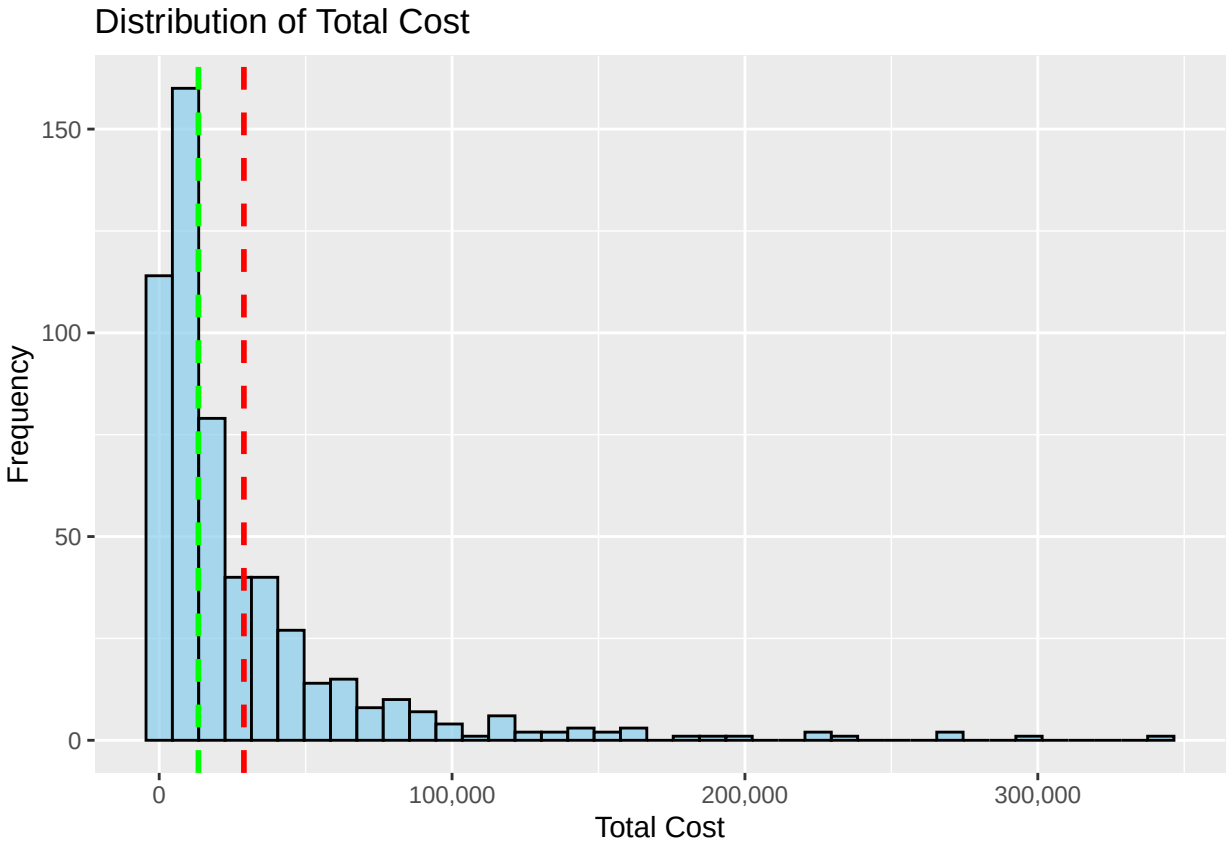
```
library(dplyr)
library(ggplot2)
library(knitr)
library(tidyr)
library(naniar)
library(Hmisc)

# loading the dataset
df <- read.table("support.txt",header = TRUE)

# renaming the columns
df <- df %>%
  rename(length_stay = slos, disease_group = dzgroup, num_comorbid = num.co, total_cost = totcst)

#finding which column type is character and use it further
cate_columns <- sapply(df, is.character)

# Convert categorical variables to factors
df[cate_columns] <- lapply(df[cate_columns], as.factor)

# capitalized the first letter of the categorical variable
df[cate_columns] <- lapply(df[cate_columns], function(x) {
  levels(x) <- capitalize(levels(x))
  x
  })

# new variable which is log(base e) of total_cost
df$log_total_cost <- log(df$total_cost)
```

Part 2: Exploratory Data Analysis

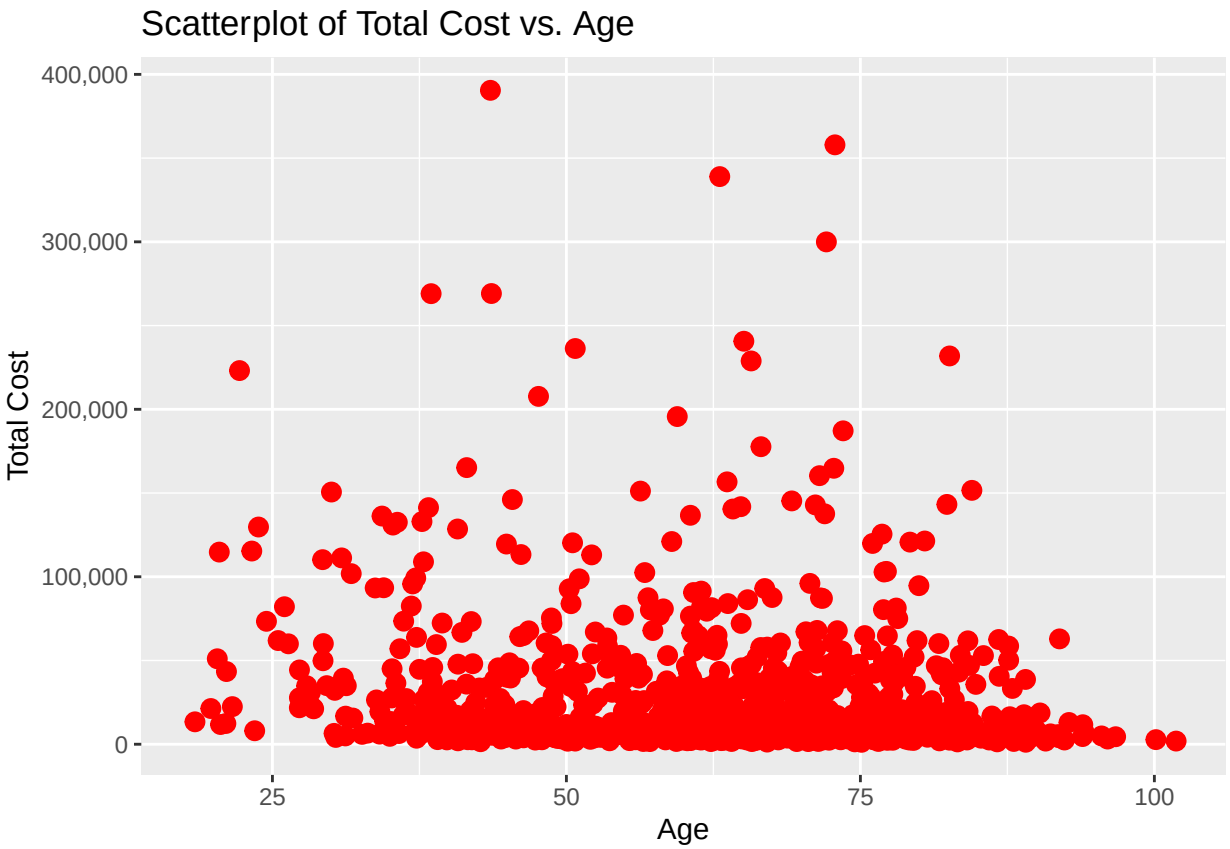

Fraction of NAs in each Variable
age
sex
length_stay
disease_group
num_comorbid
race
charges
total_cost
log_total_cost
edu
income
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Fraction of NAs



- By observing both the plots, I notice that the income variable has the highest percentage of missing data at 35%. This could be attributed to factors such as age, where individuals may be retired and lack a source of income, or may not be well-educated, leading to unemployment. Additionally, the charges variable has 2.5% null values, possibly indicating that some individuals have insurance coverage through their employers, allowing for adjustments by insurance companies, so they might end up with zero payment. The sensitivity of race as personal information may also explain why some individuals choose not to disclose this information.



- By visualizing the graph it can be seen that the total cost appears right-skewed (positively skewed) shape, which shows that there is a long tail on the right side of the distribution, suggesting that there are relatively few instances with very high total costs. The red line indicate the mean (center of plot) which is around 26000 dollar as a average total cost paid by individual and green line indicate the median which is 13000 dollar. Further more the cost range is between 1162.42871, 3.904605×10^5 and the standard deviation is 4.3478705×10^4.



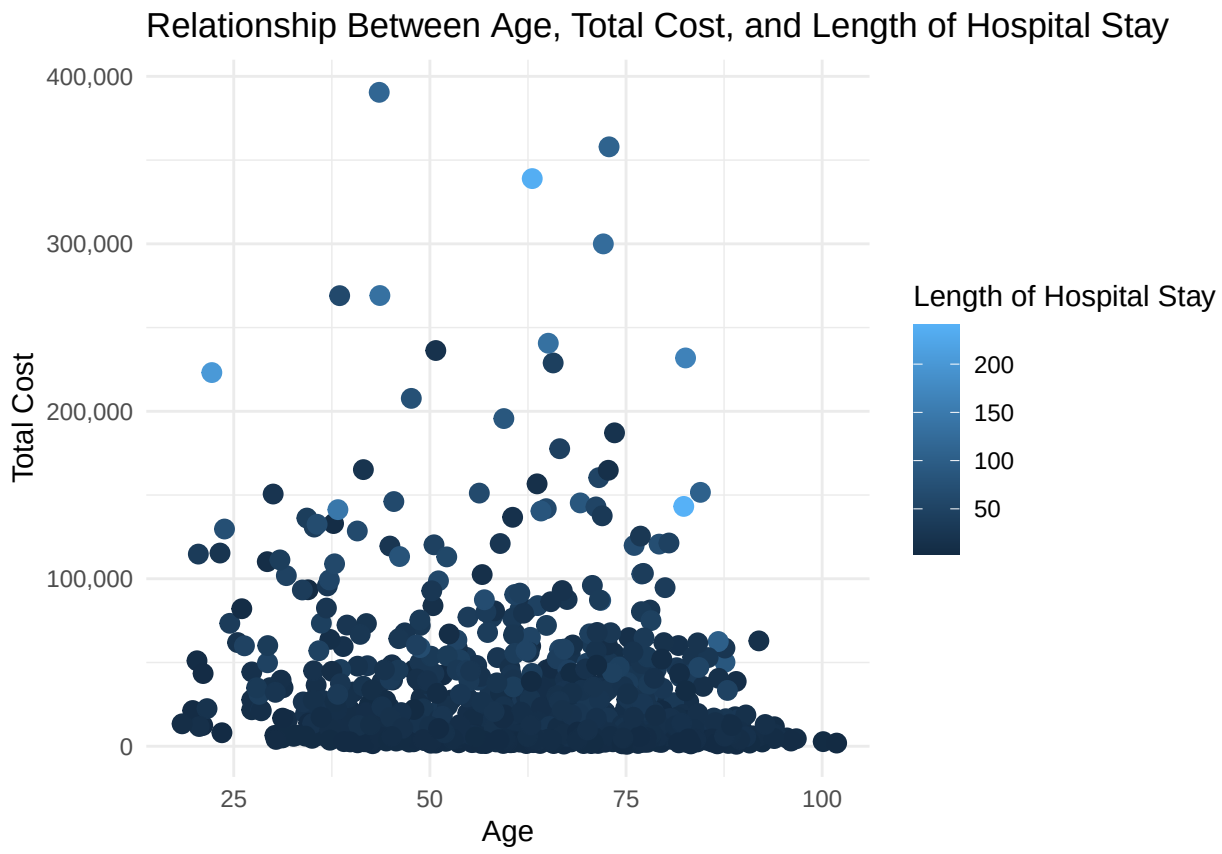
- The potential confounding variable could be "length of hospital stay" (slos) as it related to both response variable (age) and the explanatory variable (total cost)

1) Relationship with Age (Response Variable):

- People who are older usually face more complicated health issues. This often leads to them needing to stay in the hospital for a longer time. Hence, a longer hospital stay is associated with older age.

2) Relationship with Total Cost (Explanatory Variable):

- When someone stays in the hospital for a longer period, it involves more medical services and treatments. This results in higher charges, causing the overall cost to go up.



Part 3: Data Analysis

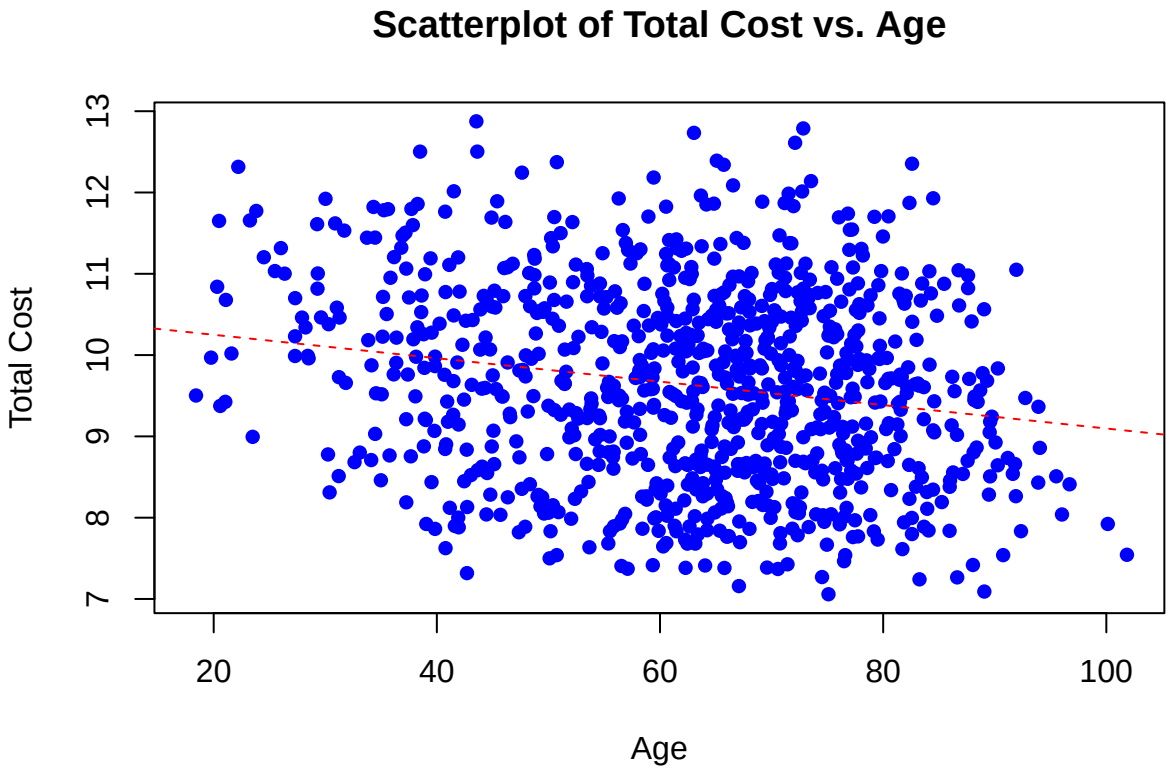
Linear Hypothesis Testing

- Null Hypothesis (H0): There is no linear relationship between total cost and age.
- Alternative Hypothesis (H1): There is a linear relationship between total cost and age.

H_0: $\beta_0 = 0$
H_1: $\beta_1 \neq 0$

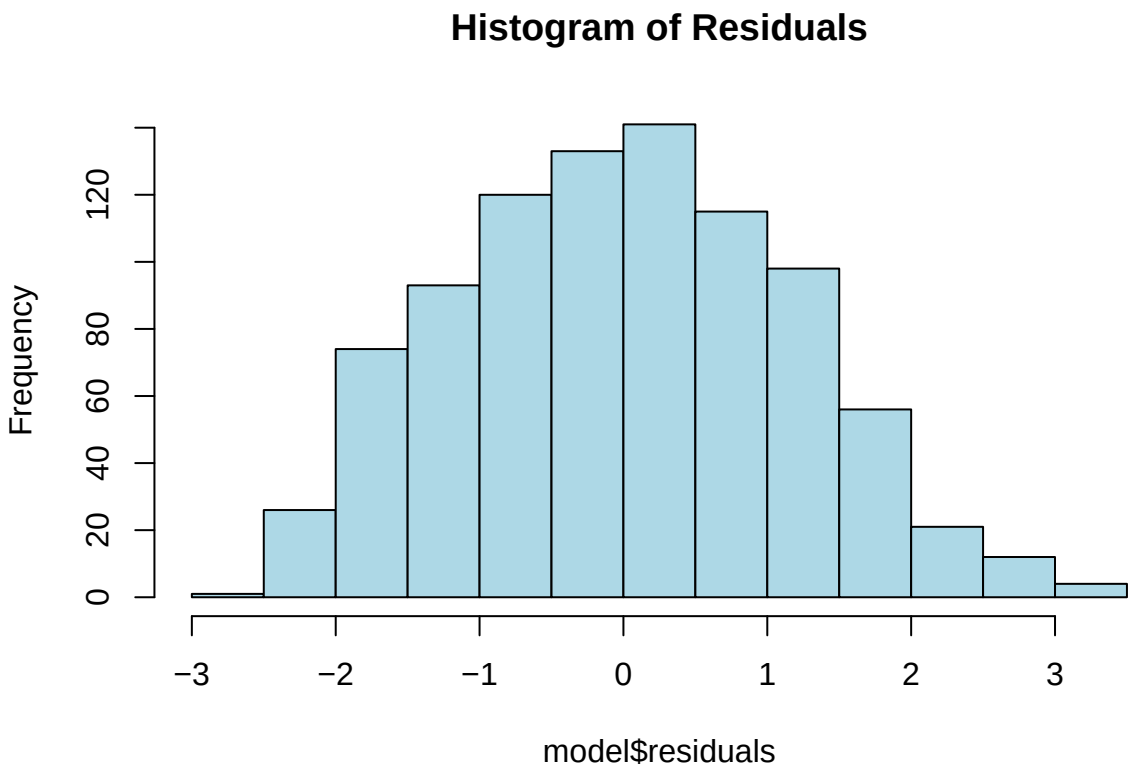
Model Assumptions:

1) Linearity: The connection between age (independent variable) and total cost (dependent variable) should follow a straight-line pattern.

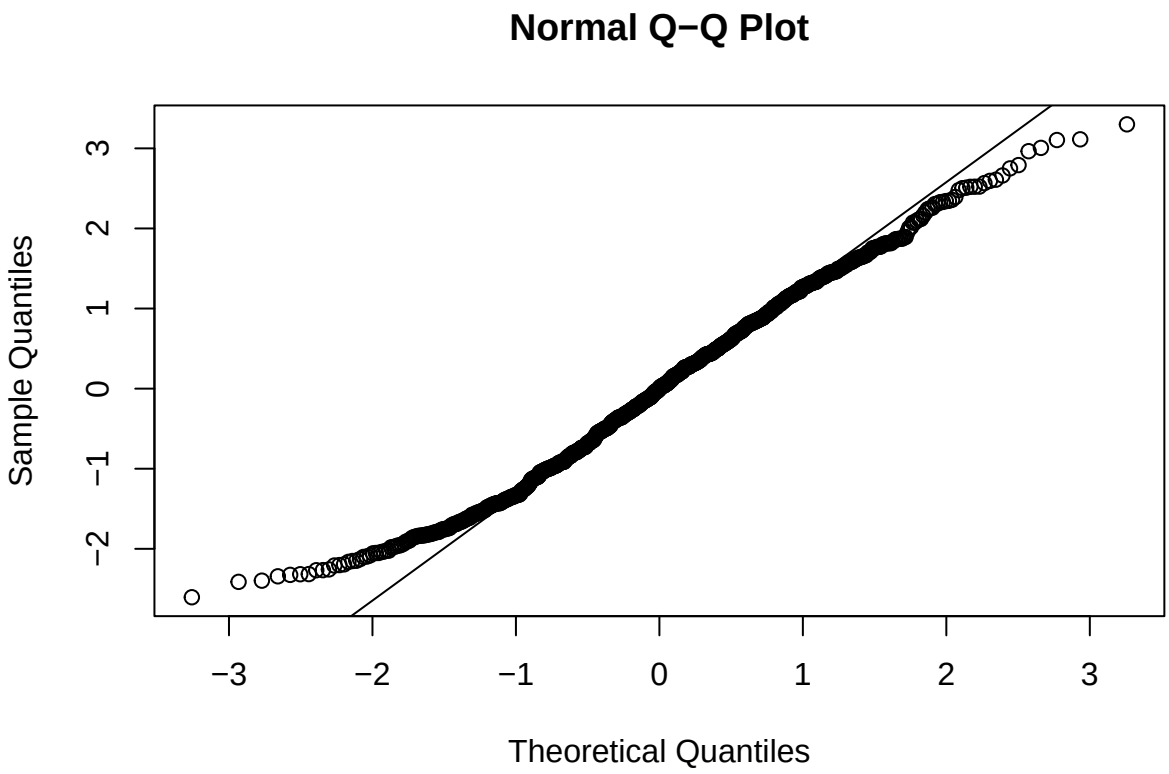


By above scatter plot it shows a slight downward trend, which indicate the negative linear association and the correlation coefficient is -0.1935673, the negative sign confirms the inverse relationship between age and total cost, which means as age increase the total cost decreases. - But still we can proceed with a linear regression analysis.

2) Nearly normal residuals: To check this condition, we can look at a histogram



- The histogram of the residuals appears to be roughly symmetric and bell-shaped, it suggests that the residuals may follow a normal distribution. This is a positive indication for the normality assumption in linear regression.



- The majority of points in the Quantile-Quantile (QQ) plot follow a straight line, it suggests that the residuals are approximately normally distributed. So, this satisfy the Nearly normal residuals condition.

3) Constant variability (homoscedasticity): By going through scatter plot we can indicate homoscedasticity in the residuals of linear regression model, it means that the variability of the residuals is approximately constant across the range of predicted values. So, this satisfy the constant variability or homoscedasticity condition.

Three step linear hypothesis test.

```
##
## Call:
## lm(formula = log_total_cost ~ age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6047 -0.9151 -0.0011  0.8468  3.2997
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.537899   0.158379  66.536  < 2e-16 ***
## age         -0.014410   0.002445  -5.893 5.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.167 on 892 degrees of freedom
##   (105 observations deleted due to missingness)
## Multiple R-squared:  0.03747,    Adjusted R-squared:  0.03639
## F-statistic: 34.72 on 1 and 892 DF,  p-value: 5.385e-09
```

P-Value of the liner regression model is 5.3854804×10^{-9} and the r square value is 0.0374683.

- log_total_cost = 10.54 - 0.0144 $\times$ age
- The slope coefficient for 'age' is -0.0144.
- For each one-unit increase in 'age', the log-transformed total cost is expected to decrease by approximately 0.0144 units.
- On the original scale, this corresponds to an expected decrease of about 1.43% in total cost for each one-unit increase in 'age'.

Part 4: Result

- The analysis indicates a linear association between age and log_total_cost. The statistical significance of the coefficients (both intercept and age) with p-values < 0.05 implies that, we have sufficient evidence to reject the null hypothesis which means that there is a linear relationship between log total cost and age . The negative slope coefficient (-0.0144) suggests that, on average, as age increases, the log-total cost decreases by about 1.43%. This implies a linear relationship where older age is associated with lower total costs.