HW 7

Brandon Keck

2024-11-24

2.

```
colnames(support) # Lists all column names in the data
```

```
##  [1] "age"     "sex"     "slos"    "dzgroup" "num.co"  "edu"     "income" 
##  [8] "charges" "totcst"  "race"
```

```
colnames(support)[c(3, 4, 5, 9)] <- c("length_stay",
                                      "disease_group",
                                      "num_comorbid",
                                      "total_cost") # Select the specific column names
colnames(support)
```

```
##  [1] "age"           "sex"           "length_stay"   "disease_group"
##  [5] "num_comorbid"  "edu"           "income"        "charges"      
##  [9] "total_cost"    "race"
```

```
# Check to make sure we get the right columns
```

```
support$sex <- factor(support$sex, levels = c("male", "female"),
                         labels = c("Male", "Female"))
# Capitalizing both variables
support$race <- factor(support$race, levels = c("white", "hispanic", "black"),
                       labels = c("White", "Hispanic", "Black"))
# Converts race column to factor
support$disease_group <- as.factor(support$disease_group)
# Converts disease group to factor
```

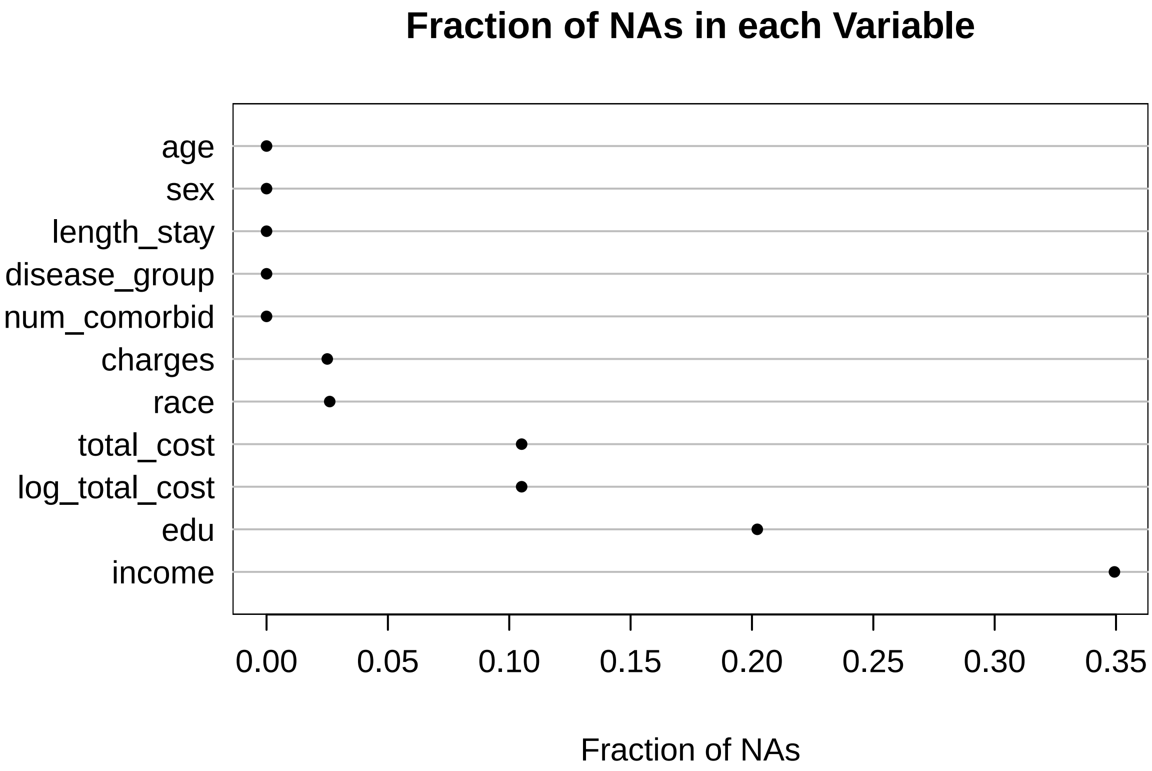
```
support <- support %>%
  mutate(log_total_cost = log(total_cost)) # create the new variable log total cost

# head(support) # check to make sure that the new variable is created
```

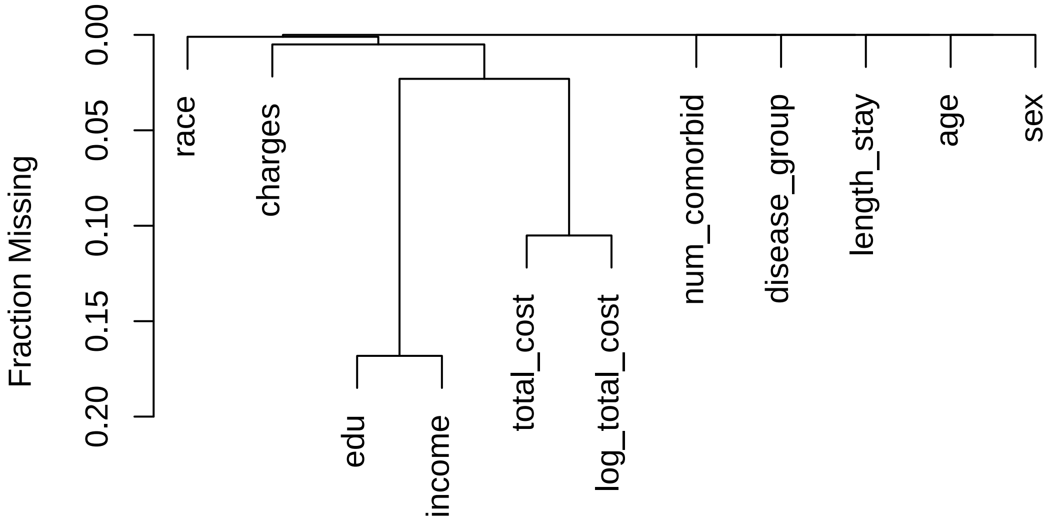
Part 2: Exploratory Data Analysis

3.

```
# From Dr. Moore
na_patterns <- naclus(support)
naplot(na_patterns, "na per var")
```



```
plot(na_patterns)
```

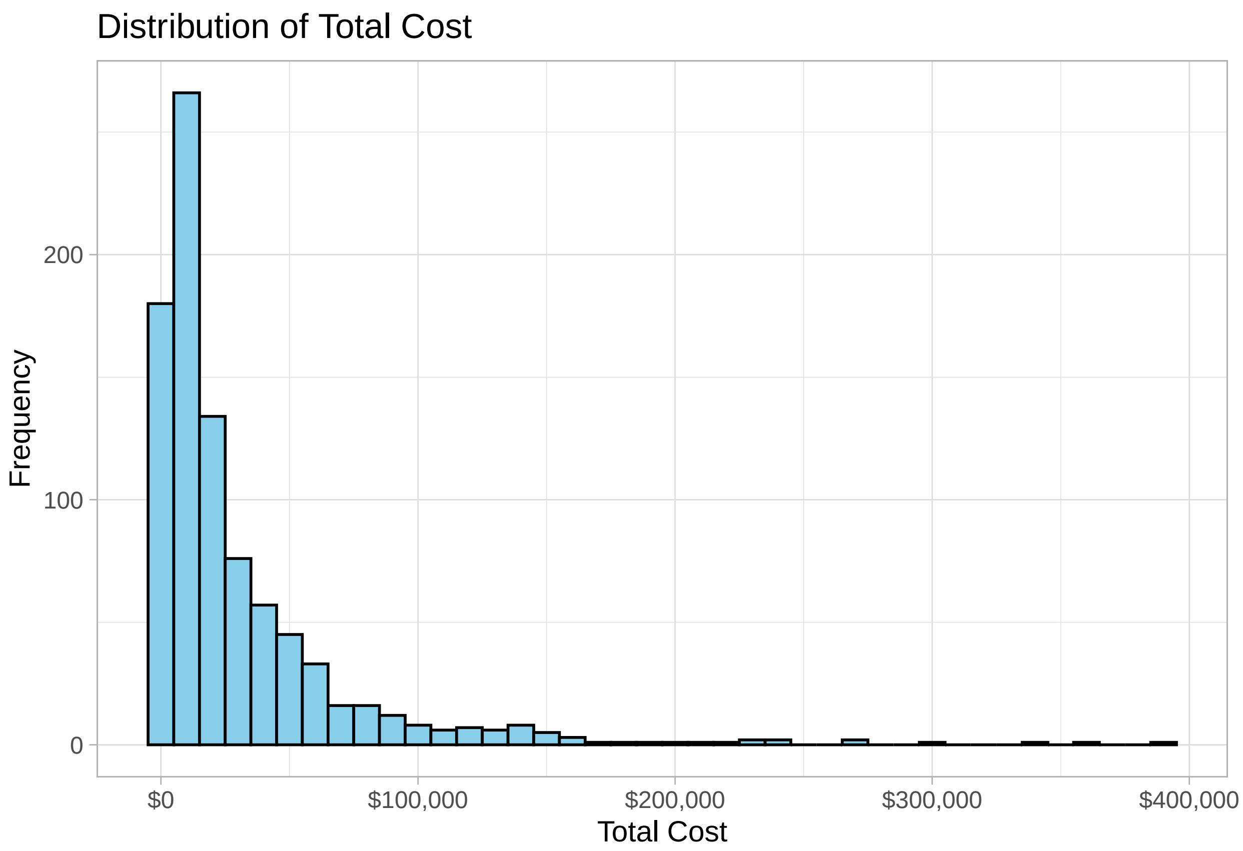


4.

In the branching diagram, the variables race and charges are closely related. This could be because some patients choose not to disclose their race or may have been unclear about the questions being asked. Similarly, education and income are also closely related, which might suggest that patients who leave the education field blank on the form are likely to leave the income field blank as well. The relationship between total_cost and log_total_cost is expected, as log_total_cost is derived directly from total_cost. Lastly, variables such as the number of comorbidities, the disease causing hospitalization, length of stay, age, and sex have no missing values. This is likely because these are critical pieces of information that the hospital is required to maintain in its records, ensuring completeness and accuracy.

In the scatter plot of Fraction of NA's we observe that education and income have the highest proportion of values. These variables could have either been left blank by the hospital for privacy reasons or by the patient themselves feeling that that particular information is irrelevant or unnecessary. Again we notice that race and charges exhibit missing values.

```
ggplot(support, aes(x = total_cost)) +
  geom_histogram(binwidth = 10000, fill = "skyblue", color = "black") +
  scale_x_continuous(labels = dollar_format()) + # Ensures proper placement
  labs(
    title = "Distribution of Total Cost",
    x = "Total Cost",
    y = "Frequency"
  ) +
  theme_light()
```



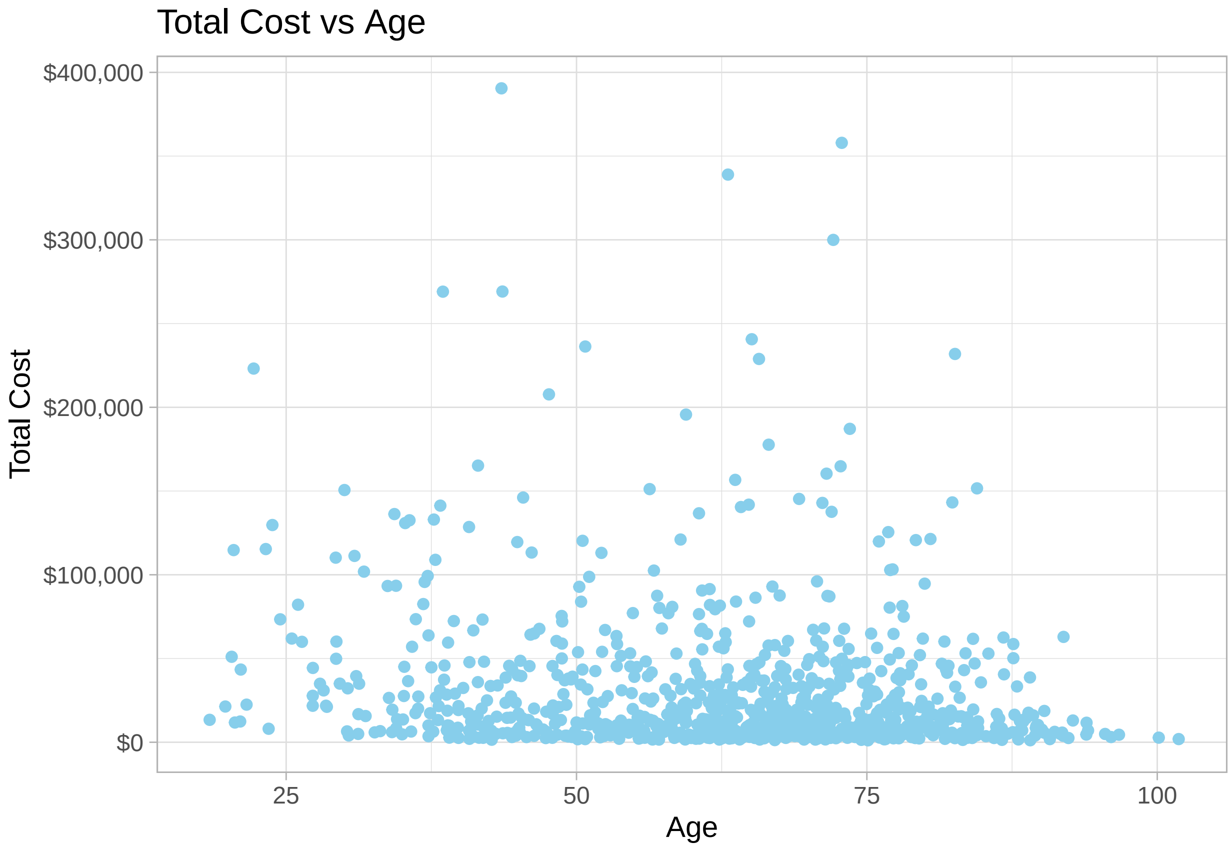
```
# binwidth = 10000 seems to work well here
# scale_x_continuous adjust x-axis scale
# dollar format function from scales
```

6.

The histogram is right-skewed, with most values concentrated in the $0–$50,000 range. A small number of outliers extend beyond $200,000, contributing to the long tail of the distribution. The center of the distribution appears to be around $25,000, indicating that many patients incur costs near this amount. The spread is quite wide, ranging from $0 to over $400,000, reflecting significant variability in total costs.

7.

```
ggplot(support, aes(x = age, y = total_cost)) +
  geom_point(color = "skyblue") +
  labs(
    title = "Total Cost vs Age",
    x = "Age",
    y = "Total Cost"
  ) + scale_y_continuous(labels = dollar_format()) +
  theme_light()
```



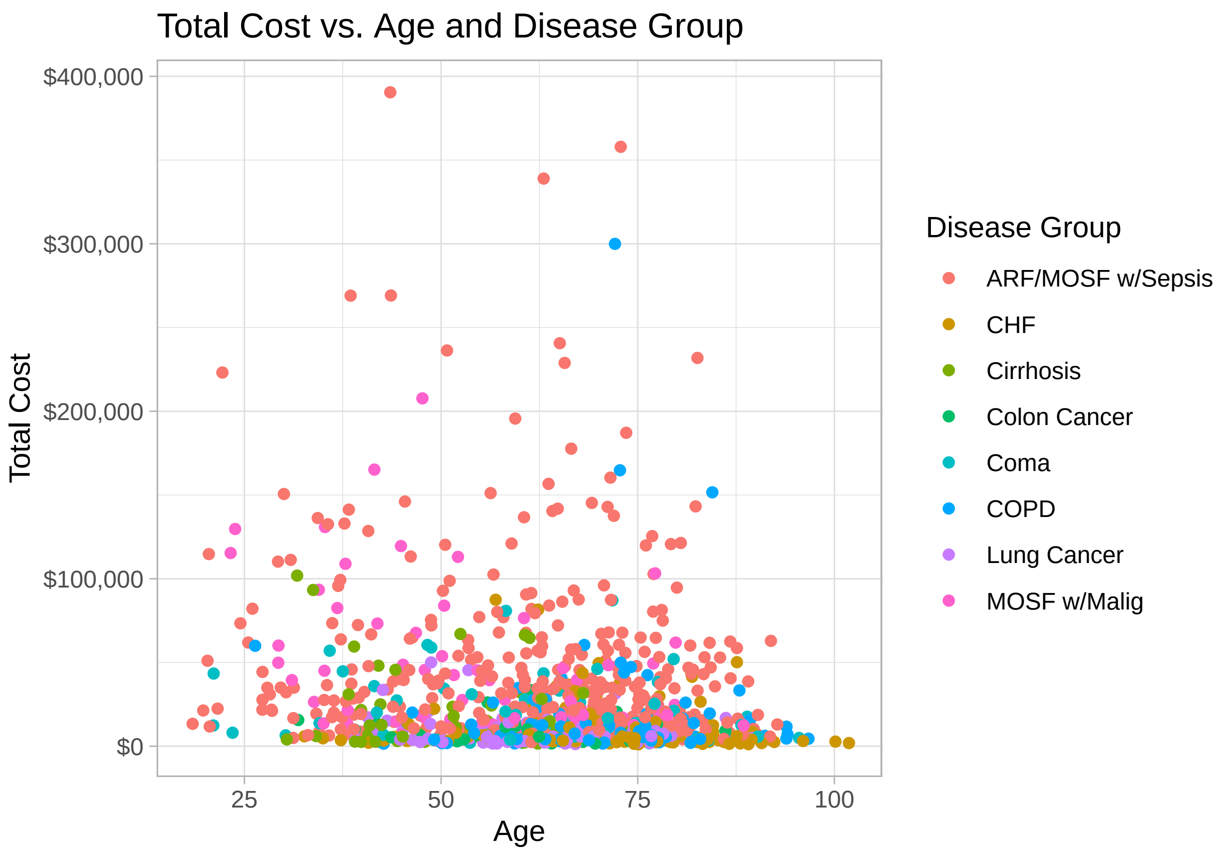
```
# scale_y_continuous adjusting the y-axis
# dollar format function from scales
```

8.

A potential confounding variable in the relationship between age and total cost is disease group. Different diseases or medical conditions in the dataset are often associated with specific age groups. For instance, certain types of cancer are more prevalent in older patients, whereas younger patients may experience conditions that are less complex or costly to treat. These medical conditions can directly impact total medical costs. For example, cancer treatments often involve advanced medical care, such as specialized procedures, chemotherapy, or prolonged hospital stays, which can significantly increase total costs.

Bonus graph

```
ggplot(support, aes(x = age, y = total_cost, color = disease_group)) +
  geom_point() +
  labs(
    title = "Total Cost vs. Age and Disease Group",
    x = "Age",
    y = "Total Cost",
    color = "Disease Group"
  ) +
  scale_y_continuous(labels = scales::dollar_format()) +
  theme_light()
```



```
# dollar format function from scales
# scale_y_continuous to adjust y-axis
# color = Disease group to fill by disease type
```

Part 3: Data Analysis

9.

$y = \beta_0 + \beta_1 x + \epsilon$

H_0 : There is no linear relationship between total cost and age. $H_0 : \beta_1 = 0$

H_A : There is a linear relationship between total cost and age. $H_A : \beta_1 \neq 0$

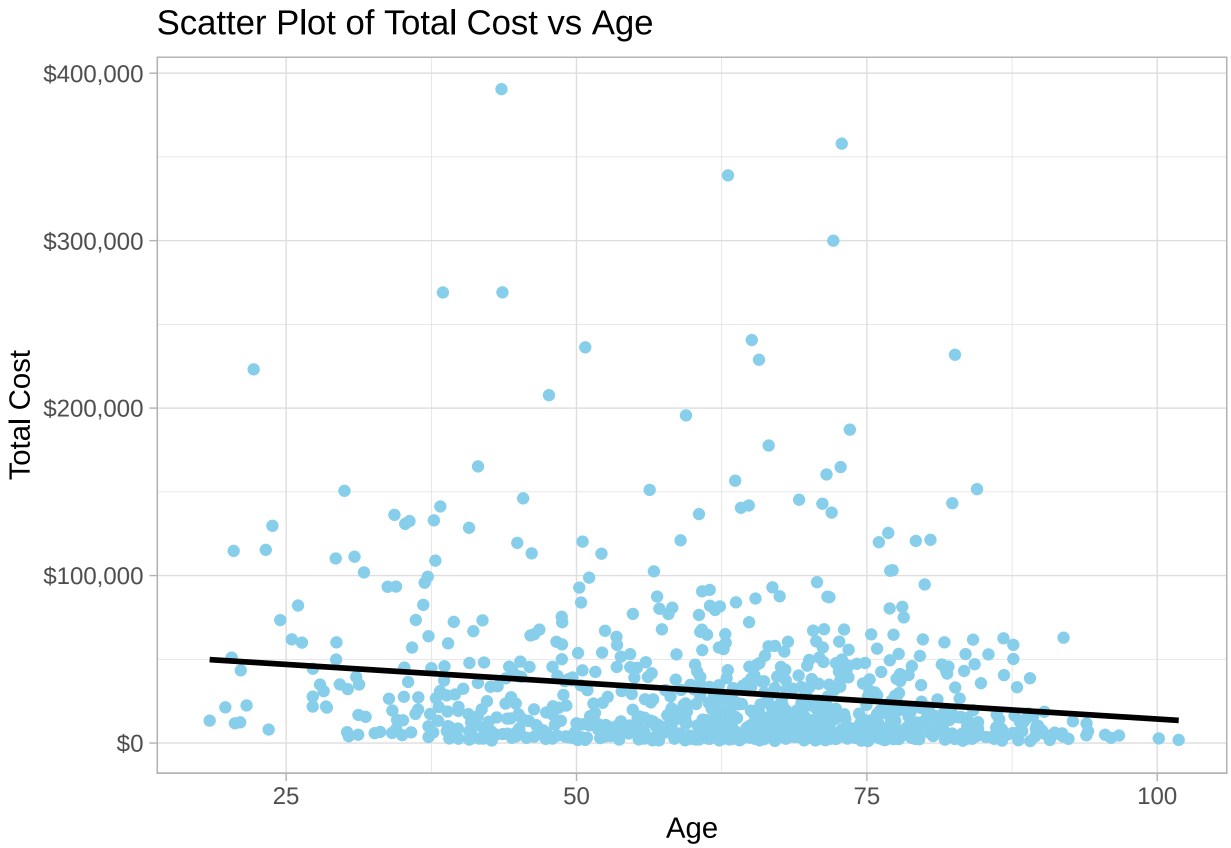
10.

```
model1 <- lm(total_cost ~ age, data = support) # Create the linear model
```

Check for Linearity:

```
ggplot(support, aes(x = age, y = total_cost)) +
  geom_point(color = "skyblue") +
  geom_smooth(method = "lm", color = "black", se = FALSE) +
  labs(
    title = "Scatter Plot of Total Cost vs Age",
    x = "Age",
    y = "Total Cost"
  ) +
  scale_y_continuous(labels = dollar_format()) +
  theme_light()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

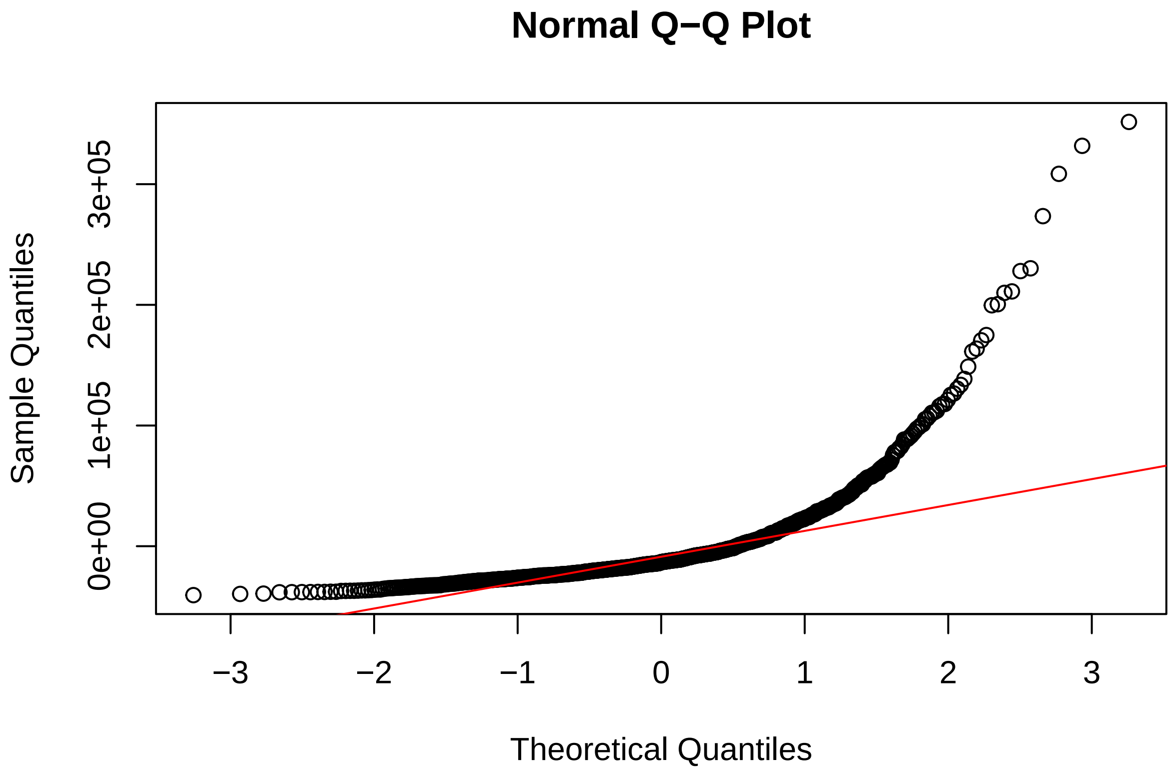


```
# dollar format function from scales
# scale_y_continuous to adjust y-axis
```

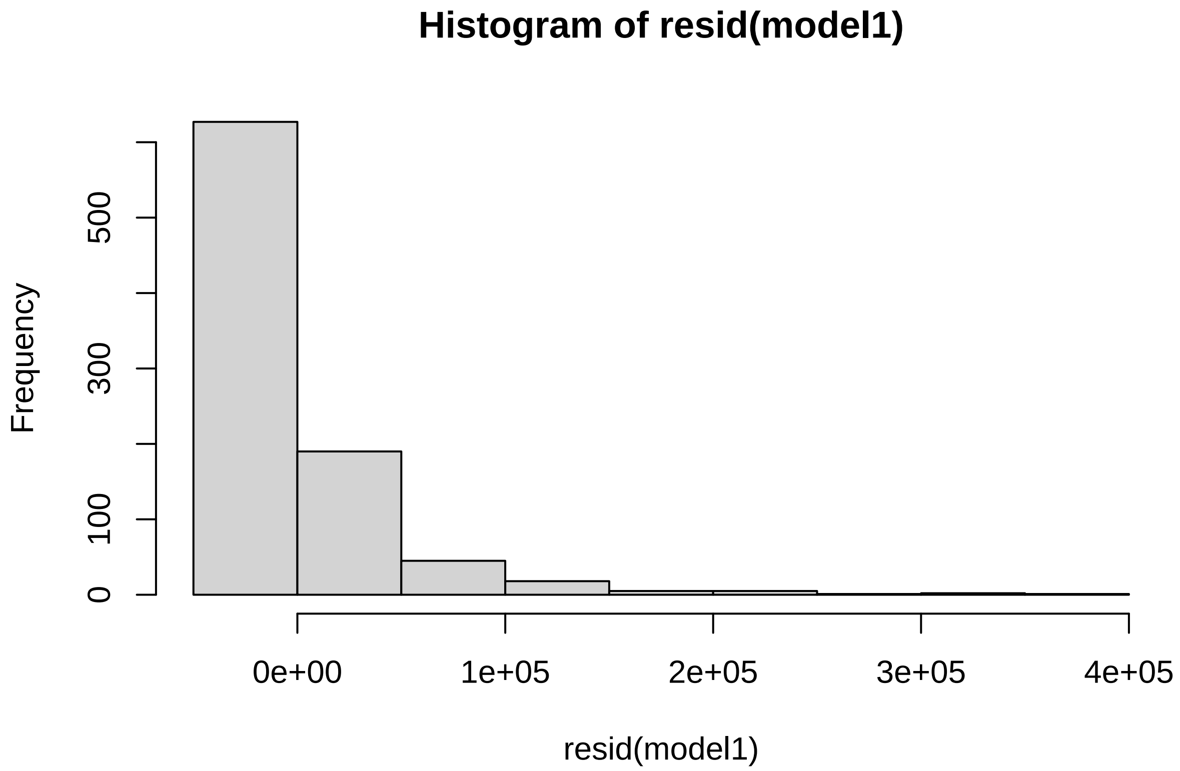
Independence is met because this is a random sample.

Check for Normality

```
# From Dr. Moore
qqnorm(resid(model1))
qqline(resid(model1), col = "red")
```



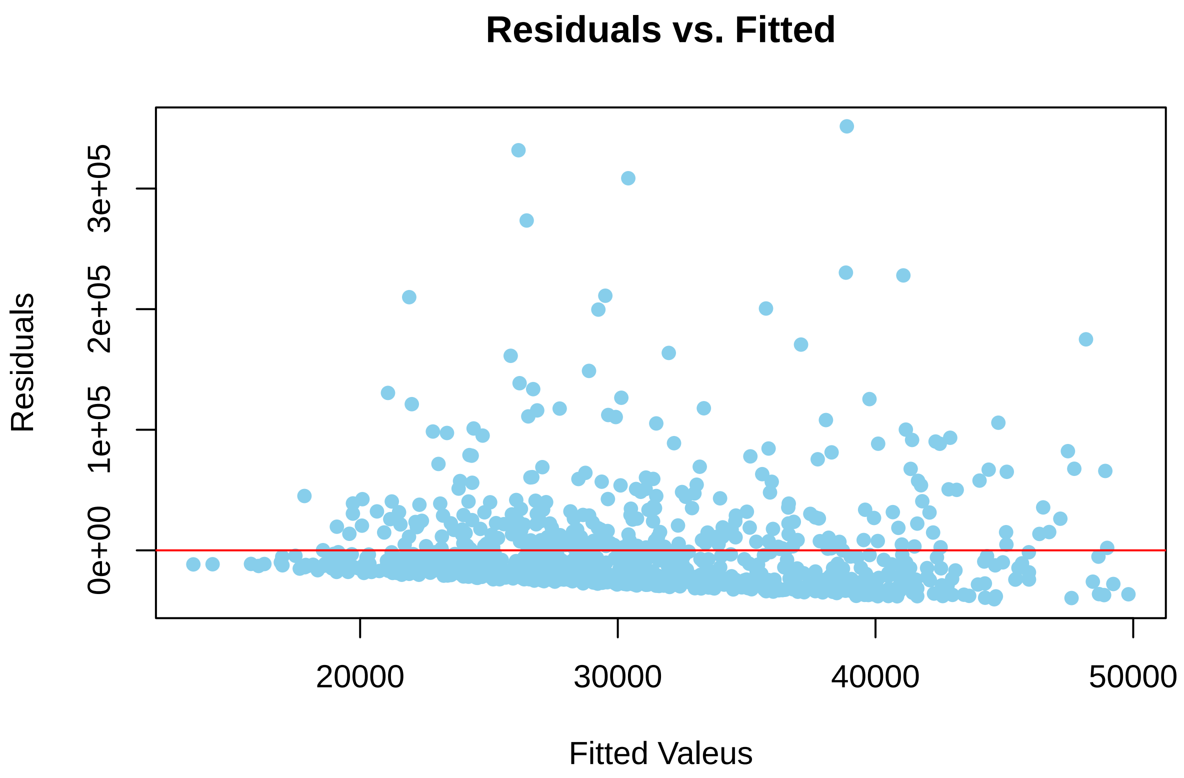
```
# From Dr. Moore
hist(resid(model1))
```



(Equal) Constant

```
#From Dr. Moore
plot(resid(model1) ~ fitted(model1),
     main = "Residuals vs. Fitted",
     xlab = "Fitted Valeus",
     ylab = "Residuals",
     pch = 16,
     col = "skyblue")

abline(h = 0, col = "red")
```



```
# adding horizontal line
```

The Q-Q plot shows significant deviations from the red diagonal line, particularly at both ends of the tails. This indicates that the residuals are not normally distributed. The histogram further supports this conclusion, as it reveals a right-skewed distribution of residuals, with most values clustered near zero and a long tail extending toward higher values.

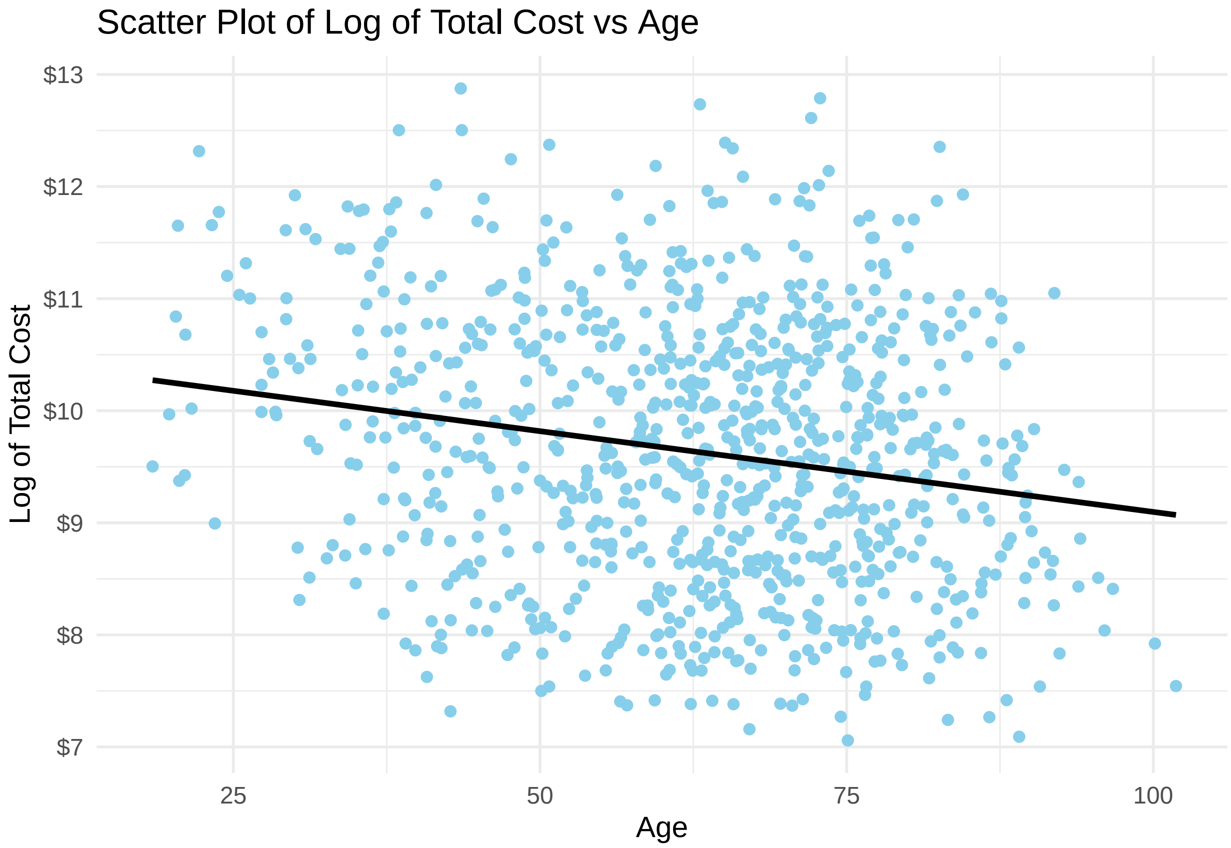
11.

```
# model the relationship between the log of total cost and age
model2 <- lm(log_total_cost ~ age, data = support)
```

Check for Linearity

```
ggplot(support, aes(x = age, y = log_total_cost)) +
  geom_point(color = "skyblue") +
  geom_smooth(method = "lm", color = "black", se = FALSE) +
  labs(
    title = "Scatter Plot of Log of Total Cost vs Age",
    x = "Age",
    y = "Log of Total Cost"
  ) +
  scale_y_continuous(labels = dollar_format()) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

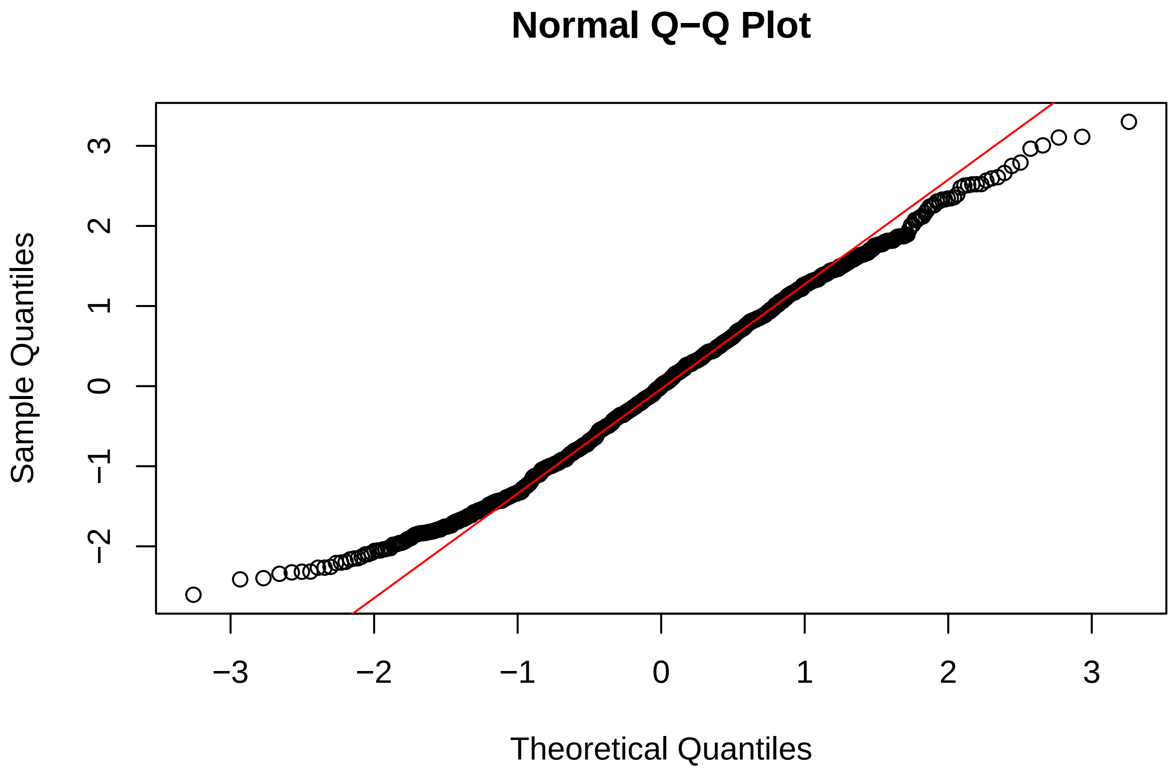


```
# scale_y_continuous adjust y-axis
# dollar format function from scales
```

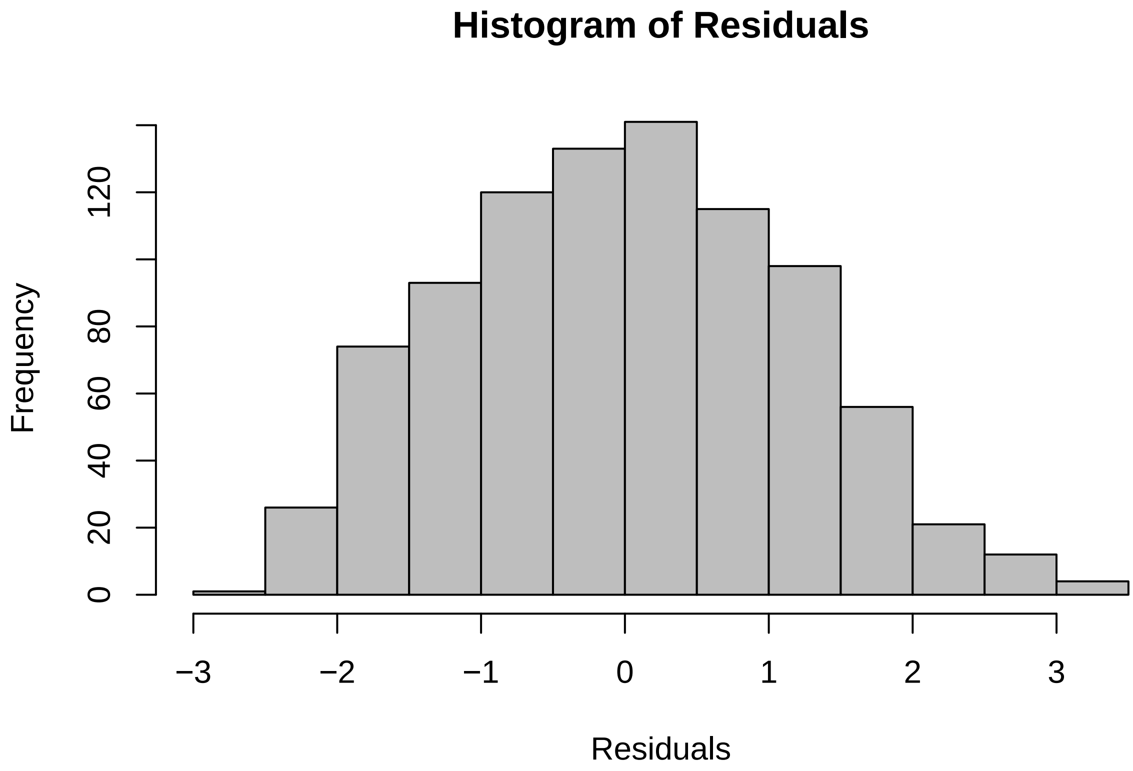
Independence is met because this is a random sample

Check for Normality

```
#From Dr. Moore notes
qqnorm(resid(model2))
qqline(resid(model2), col = "red")
```

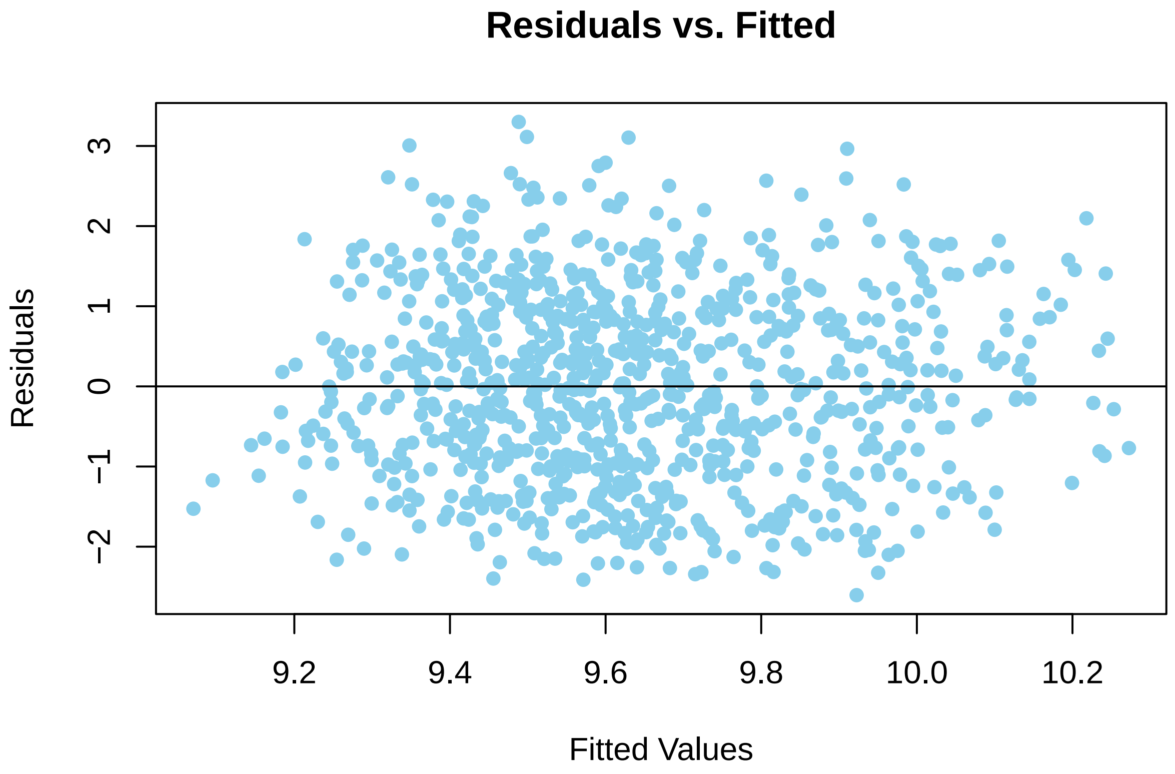


```
hist(resid(model2),
     main = "Histogram of Residuals",
     xlab = "Residuals",
     col = "grey",
     border = "black")
```



(Equal) Constant

```
# From Dr. Moore notes
plot(resid(model2) ~ fitted(model2),
     main = "Residuals vs. Fitted",
     xlab = "Fitted Values",
     ylab = "Residuals",
     pch = 16,
     col = "skyblue")
abline(h = 0, col = "black")
```



12.

```
model2 <- lm(log_total_cost ~ age, data = support)
summary(model2)
```

```
##
## Call:
## lm(formula = log_total_cost ~ age, data = support)
##
## 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
```

```
# creating new linear model with log
# summary to get summary statistics
```

Decision: With a p-value of 5.39e-09 which is smaller than our significance level of $alpha = 0.05$ we reject the null hypothesis.

Conclusion:We have enough evidence to conclude that there is relationship between age and the log of total cost.

13.

The slope of the coefficient for age is -0.014410. For every 1 year increase in age, the total cost decreases by approximately 1.44% on average.

14.

Yes there is evidence that age is linearly associated with the (log) of total cost. The relationship between age and (log) total cost is a negative relationship. What this means is that for every year that age increases the log of total cost will decrease by 0.014410 on average. What this suggests is that older individuals have lower medical costs.

Rating E