Customer Behaviour Analysis

Thi Thu Hien LE

a) Test whether shopping frequency is independent of store location. Make sure you:

i. Visualize the data
ii. Show all the key steps of hypothesis testing
iii. Interpret the results, both statistically and in the context of the case. In no more than 3 sentences explain what the results reveal about how shopping frequency is associated with store location, and how might these insights inform marketing strategies or store management decisions?

H0: shopping frequency is independent of store location HA: shopping frequency is dependent of store location

```
# Load necessary libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

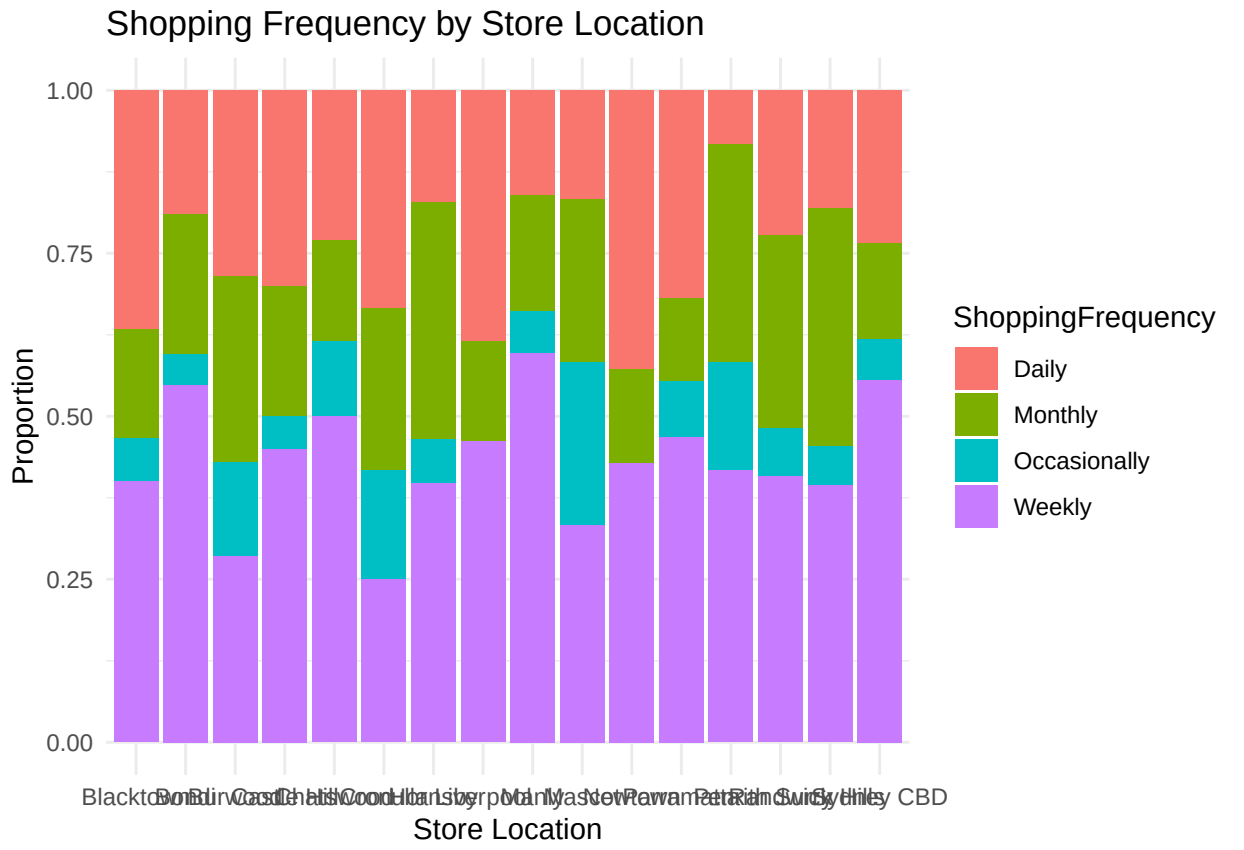
```
# Load the dataset
data <- read.csv("Retail.csv")

# Create a contingency table of Store Location and Shopping Frequency
contingency_table <- table(data$StoreLocation, data$ShoppingFrequency)

# Perform the Chi-Square Test of Independence
chi_test <- chisq.test(contingency_table)
```

```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

```r
# Print the result of the Chi-Square test
print(chi_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 48.262, df = 45, p-value = 0.3424
```

```r
# Visualize the data
ggplot(data, aes(x=StoreLocation, fill=ShoppingFrequency)) +
  geom_bar(position="fill") +
  labs(title="Shopping Frequency by Store Location", x="Store Location", y="Proportion") +
  theme_minimal()
```



As we got p-value = 0.3424 > 0.05, we do not reject the null hypothesis. This suggests that there is no statistically significant association between shopping frequency and store location. This implies that customers have similar shopping habits across the various store locations, and factors other than location (e.g., demographics, product offerings) might play a more critical role in influencing shopping frequency.

b) Compute a 95% confidence interval for the difference in the mean satisfaction score between customers who buy baby items and those who buy personal care items? Make sure you:

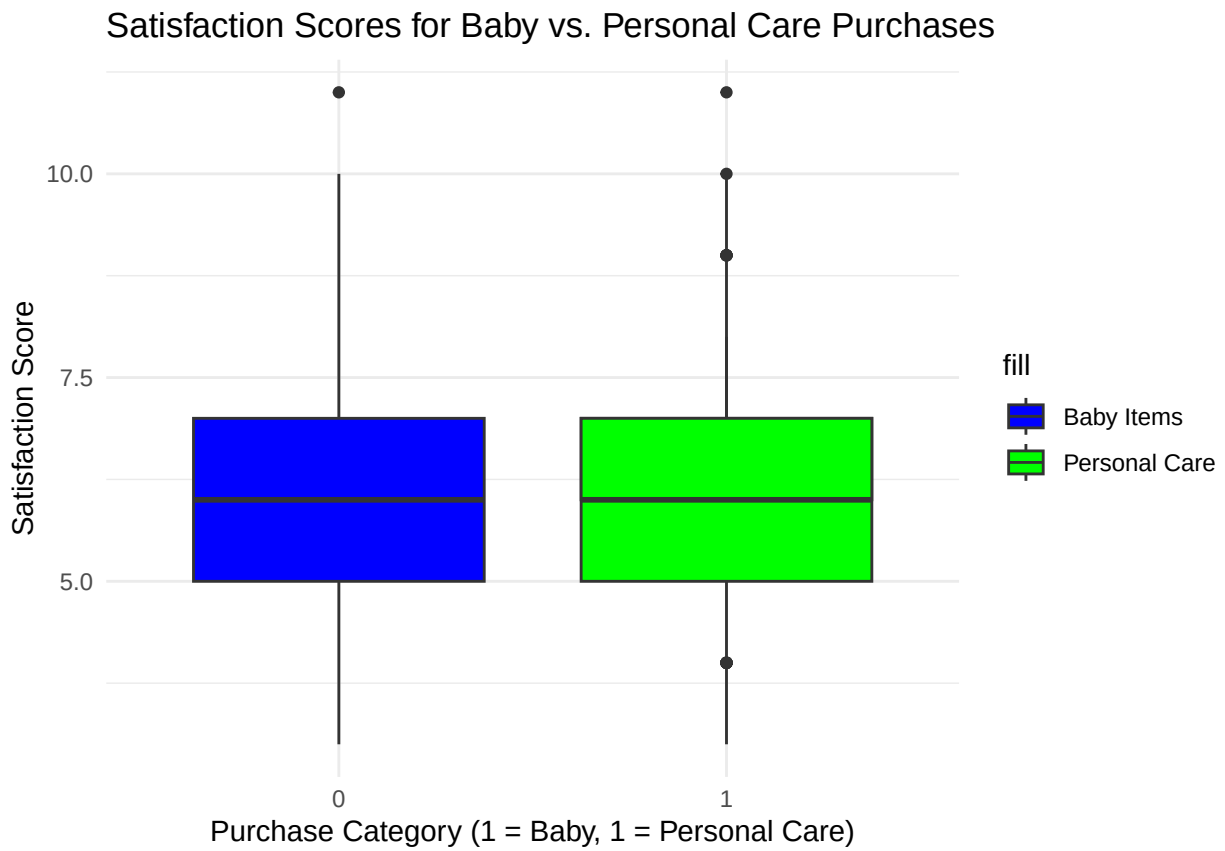
i. Visualize the data

ii. Interpret the results, both statistically and in the context of the case. In no more than three sentences, explain what the results reveal about the difference in mean satisfaction score between customers who buy baby items and those who buy personal care items, and how these findings could be relevant for marketing or product placement strategies.

```r
# Separate the satisfaction scores based on Baby Purchase and Personal Care Purchase
baby_items <- data[data$BabyPurchase == 1, "SatisfactionScore"]
personal_care_items <- data[data$PersonalCarePurchase == 1, "SatisfactionScore"]

# Compute the means and standard deviations for both groups
mean_baby <- mean(baby_items)
mean_personal_care <- mean(personal_care_items)
sd_baby <- sd(baby_items)
sd_personal_care <- sd(personal_care_items)

# Perform a t-test to compute the confidence interval for the difference in means
t_test <- t.test(baby_items, personal_care_items, conf.level = 0.95)

# Print the confidence interval
print(t_test$conf.int)
```

```
## [1] -0.2337688  0.1803445
## attr(,"conf.level")
## [1] 0.95
```

```r
# Visualize the data using box plots
ggplot(data, aes(x=factor(BabyPurchase), y=SatisfactionScore)) +
  geom_boxplot(aes(fill="Baby Items")) +
  geom_boxplot(data = data[data$PersonalCarePurchase == 1, ], aes(x=factor(PersonalCarePurchase), fill=
  labs(title="Satisfaction Scores for Baby vs. Personal Care Purchases",
       x="Purchase Category (1 = Baby, 1 = Personal Care)", y="Satisfaction Score") +
  scale_fill_manual(values=c("blue", "green")) +
  theme_minimal()
```



Since the confidence interval includes 0, we cannot conclude that there is a statistically significant difference in the mean satisfaction scores between the two groups. The difference is likely to be small and could be due to random variation rather than a true effect. The data suggests that customers who buy baby items and those who buy personal care items have similar satisfaction scores. From a marketing or product placement perspective, this means that satisfaction may not be the key factor differentiating these customer groups.

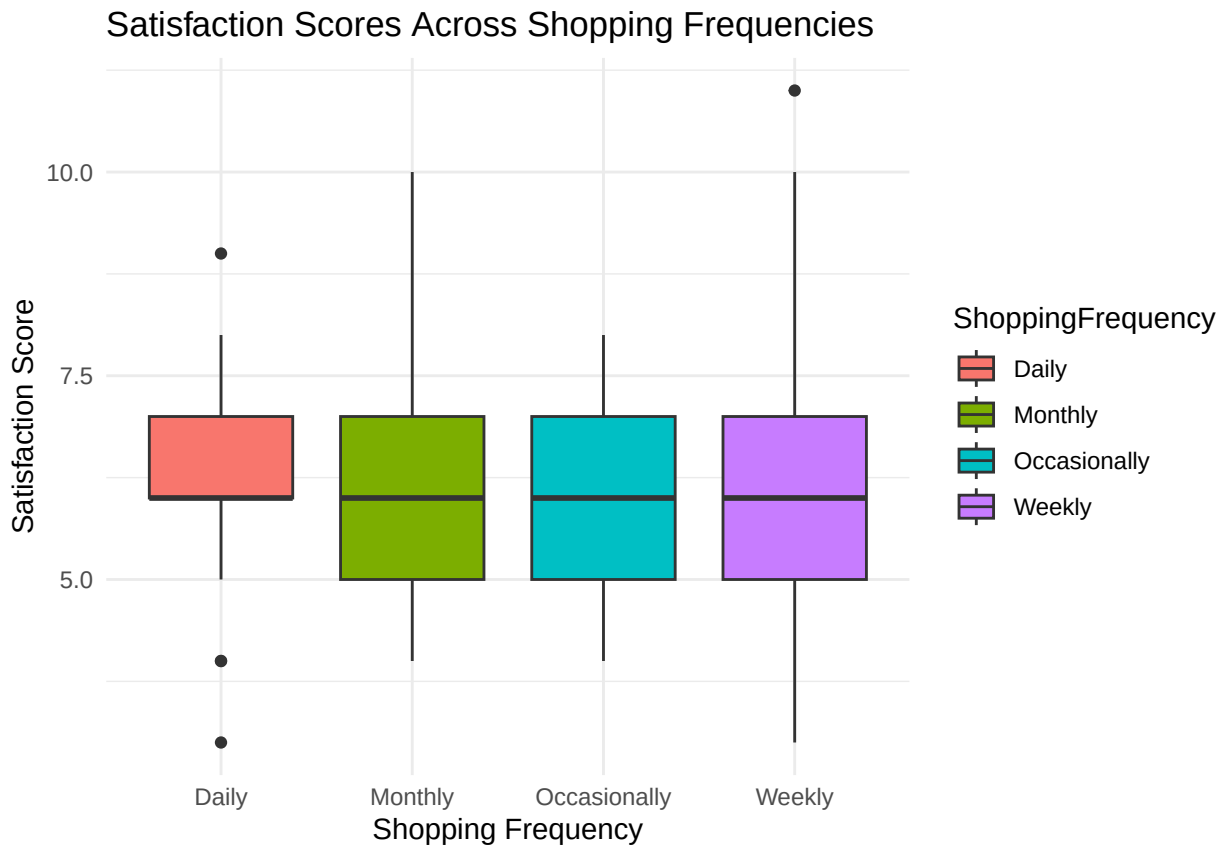
c) Test whether the satisfaction score varies significantly across different shopping frequency. Make sure you:

i. Visualize the data
ii. Show all the key steps of hypothesis testing,
iii. Interpret the results, both statistically and in the context of the case.
iv. In no more than three sentences, explain what the results reveal about the difference mean satisfaction score across different shopping frequency.
v. Additionally, identify which specific shopping frequency show significant differences and explain how these insights could be used to make informed business decisions (only print the significant results).

H0: satisfaction score is the same across different shopping frequency HA: satisfaction score varies significantly across different shopping frequency

```
# Visualize the data using boxplots
ggplot(data, aes(x=ShoppingFrequency, y=SatisfactionScore, fill=ShoppingFrequency)) +
  geom_boxplot() +
  labs(title="Satisfaction Scores Across Shopping Frequencies",
```

```
      x="Shopping Frequency", y="Satisfaction Score") +
  theme_minimal()
```




```
# Perform ANOVA to test for significant differences in satisfaction score across shopping frequency gro
anova_result <- aov(SatisfactionScore ~ ShoppingFrequency, data = data)

# Summary of ANOVA results
summary(anova_result)
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## ShoppingFrequency   3    0.9  0.2893    0.23  0.875
## Residuals         492  618.5  1.2571
```

```
# If the ANOVA shows significant results, perform a Tukey post-hoc test to find which groups differ
tukey_result <- TukeyHSD(anova_result)

# Print only significant results from Tukey post-hoc test
significant_results <- tukey_result$ShoppingFrequency[tukey_result$ShoppingFrequency[, "p adj"] < 0.05,
print(significant_results)
```

```
##      diff lwr upr p adj
```

As the p-value = 0.875 which is much greater than 0.05, we do not reject the null hypothesis. There is no statistically significant difference in the mean satisfaction scores across different

shopping frequencies. Since the ANOVA result is not significant (p-value = 0.875), there is no need for post-hoc testing, and no specific pairs of shopping frequency groups show significant differences. This implies that customer satisfaction does not depend heavily on how frequently they shop and businesses can focus on other factors like product quality, customer service, or pricing to improve satisfaction levels rather than targeting specific frequency groups.

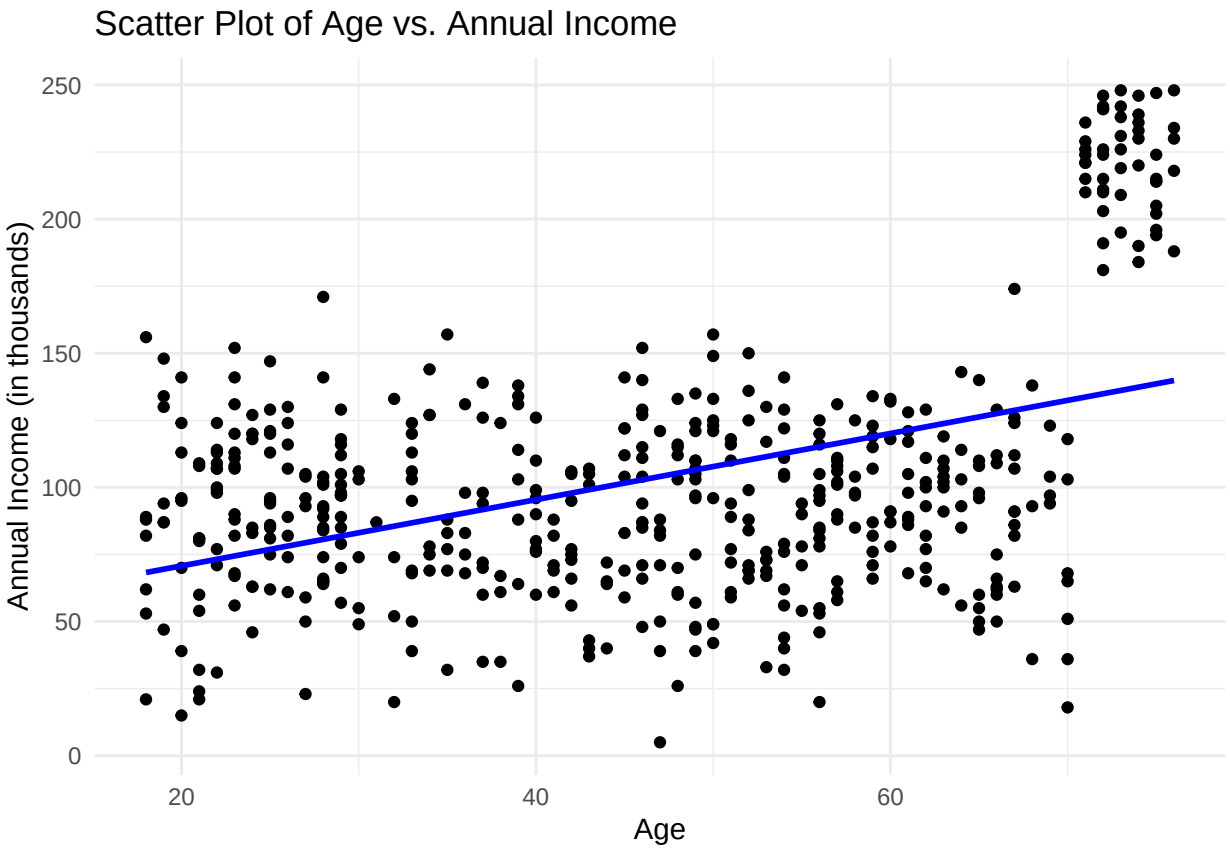
d) Test if age can be used to predict annual income. Assess the strength of the predictive power of age for annual income. Make sure you:

i. Visualize the data
ii. Show all the key steps of hypothesis testing
iii. Interpret the results, both statistically and in the context of the case.
iv. In no more than three sentences, explain what the results reveal about the relationship between age and annual income. Additionally, use your model to predict the annual income for a customer who is 34 years old. Comment on the appropriateness of this prediction and justify your answer. Examine the residuals of your regression model: What do they indicate about the model's performance and any potential issues with the assumptions of your model?

H0: age can not be used to predict annual income HA: age can be used to predict annual income

```
# Visualize the relationship between Age and AnnualIncome using a scatter plot
ggplot(data, aes(x=Age, y=AnnualIncome)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE, color="blue") +
  labs(title="Scatter Plot of Age vs. Annual Income",
       x="Age", y="Annual Income (in thousands)") +
  theme_minimal()
```

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



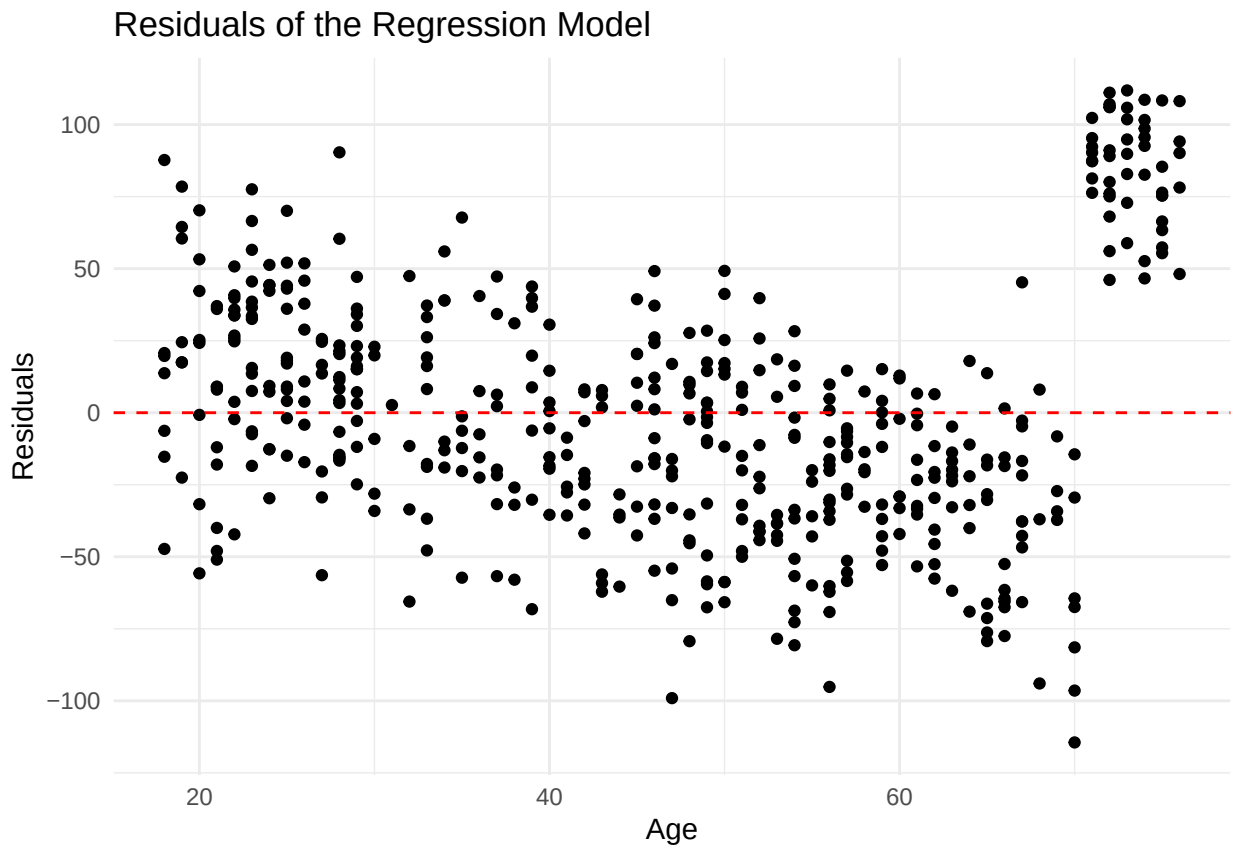

```r
# Fit a linear regression model: AnnualIncome as a function of Age
model <- lm(AnnualIncome ~ Age, data = data)

# Summary of the linear regression model
summary(model)
```

```
##
## Call:
## lm(formula = AnnualIncome ~ Age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -114.445  -31.843   -4.542   25.236  111.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.0761     5.7732   7.981 1.02e-14 ***
## Age           1.2338     0.1166  10.582  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.17 on 494 degrees of freedom
## Multiple R-squared:  0.1848, Adjusted R-squared:  0.1831
## F-statistic:   112 on 1 and 494 DF,  p-value: < 2.2e-16
```

```r
# Predict annual income for a customer who is 34 years old
age_34_income <- predict(model, newdata = data.frame(Age = 34))
print(age_34_income)
```

```
##        1
## 88.02661
```

```r
# Plot residuals to assess the performance of the model
ggplot(data, aes(x=Age, y=residuals(model))) +
  geom_point() +
  geom_hline(yintercept=0, linetype="dashed", color="red") +
  labs(title="Residuals of the Regression Model",
       x="Age", y="Residuals") +
  theme_minimal()
```



As the p-value for Age (< 2e-16), we reject the null hypothesis. This indicates that the relationship between age and annual income is highly significant. With intercept of 46.08, we can create a fomular: AnnualIncome=46.08+1.23×Age. The predicted annual income for a customer aged 34 years is $88,026, however, R^2 = 0.1848 means that only 18.48% of the variation in annual income can be explained by age; This suggests that age is a weak to moderate predictor of annual income, and other factors likely play a more significant role in determining income => The prediction may not be highly reliable, as other factors influencing income are not considered. The range of residuals is quite wide, from -114.445 to 111.854, indicating some large deviations from the predicted values, which could signal issues like potential outliers or other missing explanatory variables.

e) Describe the potential sampling process that Southern Cross Goods might have used to collect this data. Consider how they might have ensured the sample is representative of their entire customer base and discuss any sampling methods that could have been employed.

Since the data was already collected using stratified sampling, Southern Cross Goods likely ensured representation by dividing their customer base into distinct subgroups (strata) based on key characteristics like age, gender, income, shopping frequency, and store location. Each subgroup was sampled in a manner proportional to its size in the overall population or with equal representation if certain groups were smaller.

This method would have helped them address potential biases, ensuring that insights drawn from the data reflect the broader customer base accurately. The stratified sampling approach is especially beneficial when customer behaviors differ across various subgroups, enabling more granular and targeted analysis of trends and behaviors across the company's diverse clientele.

By ensuring a representative sample, Southern Cross Goods can derive insights that are more actionable for improving marketing strategies, customer service, and operational decisions, such as tailoring promotions to specific demographic groups or optimizing inventory at different store locations.