Business Statistics End of Term Assessment IB94X0 2024-2025 #1

5624055

Contents

Academic Integrity Statement **2**

Use of AI Statement **2**

Question 1 **2**

- **Data Dictionary** . . . 2
  - **Overview** . . . 2
  - **Variable Definitions** . . . 2
- **Data Cleaning and Overview** . . . 3
- **Exploratory Correlation Analysis prior to Multiple Linear Regression** . . . 8
- **Multiple Linear Regression** . . . 11
- **Inference** . . . 13
- **Future Work Prospect** . . . 14

Question 2 **16**

- **Data Dictionary** . . . 16
  - **Overview** . . . 16
  - **Variable Definitions** . . . 17
- **Exploratory Correlation Analysis prior to Multiple Linear Regression** . . . 21
- **Multiple Linear Regression** . . . 23
- **Effect of Delivery time upon Customer Satisfaction across Socio-Economic-Status** . . . 25
  - **Inference** . . . 30

```r
library(tidyverse)
library(gridExtra)
library(grid)
library(dplyr)
library(knitr)
library(corrplot)
library(Hmisc)
library(gridGraphics)
library(car)
library(lmtest)
library(nortest)
library(emmeans)
library(mgcv)
```

Academic Integrity Statement

I hereby declare that the work presented in this report is my own original work. I am fully aware of the University of Warwick's regulations concerning plagiarism and collusion, and I confirm that no part of this work has been submitted for assessment in any other course or institution.

Use of AI Statement

This assignment utilized AI tools (ChatGPT, Gemini) for the following purposes:

- **Code Optimization:** AI assisted in refining code for better readability, efficiency, and conciseness through code minimization and nesting.
- **Concept Exploration:** AI was used to explore and understand complex concepts, aiding in the interpretation of research findings and cross-referencing information from various sources.

Question 1

Data Dictionary

Overview

- **Dataset Name:** Cardio Vascular Disease in England
- **Source:** myWBS

Variable Definitions

Variable Name	Description	Data Type	Valid Values
area_name	Name of the Area	Character	
area_code	Area Code	Character	
Population	Total Population living in each area	Number/Numeric	1960-1056970
Poverty	Proportion of people who meet the definition of living in poverty	Number/Numeric	12.9-30.7

Variable Name	Description	Data Type	Valid Values
CVD	Percentage of people living in the area who have recently experienced Cardiovascular Disease	Number/Numeric	7.9- 17.8
overweight	Proportion of people who are overweight	Number/Numeric	10.2432-40.21633
smokers	Proportion of people who smoke	Number/Numeric	3.2-27.8
wellbeing	Average wellbeing score of people living in the area	Number/Numeric	6.61-8.17

Data Cleaning and Overview

```r
#Importing the data
data1 <- read_csv("~/Downloads/BS Assignment/Cardio_Vascular_Disease.csv")
```

```
## Rows: 385 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): area_name, area_code
## dbl (6): Population, Poverty, CVD, overweight, smokers, wellbeing
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
#Checking data for duplicates
duplicates <- data1[duplicated(data1), ]
duplicates
```

```
## # A tibble: 0 x 8
## # i 8 variables: area_name <chr>, area_code <chr>, Population <dbl>,
## #   Poverty <dbl>, CVD <dbl>, overweight <dbl>, smokers <dbl>, wellbeing <dbl>
```

No duplicate values were found in the data.

```r
#Checking the data for missing values
print(summary(is.na(data1)))
```

```
##  area_name       area_code       Population       Poverty
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:385       FALSE:385       FALSE:309       FALSE:309
##                                   TRUE :76        TRUE :76
##     CVD          overweight       smokers         wellbeing
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:309       FALSE:313       FALSE:378       FALSE:370
##  TRUE :76        TRUE :72        TRUE :7         TRUE :15
```

Missing values were identified in the following columns: Population, Poverty, CVD, Overweight, Smokers, and Wellbeing.

```r
#Further examining the data before discarding missing values
#Checking the number of unique area names and area codes and comparing with the length of the column
length(unique(data1$area_name)) == length(data1$area_name)
```

```
## [1] TRUE
```

```r
length(unique(data1$area_code)) == length(data1$area_code)
```

```
## [1] TRUE
```

A total of 385 unique areas were identified in the dataset. Initially, we considered using median imputation to address missing values. However, testing revealed that median imputation significantly skewed the data distribution compared to the original, suggesting it could introduce substantial bias. Therefore, we opted to remove missing values in this analysis.

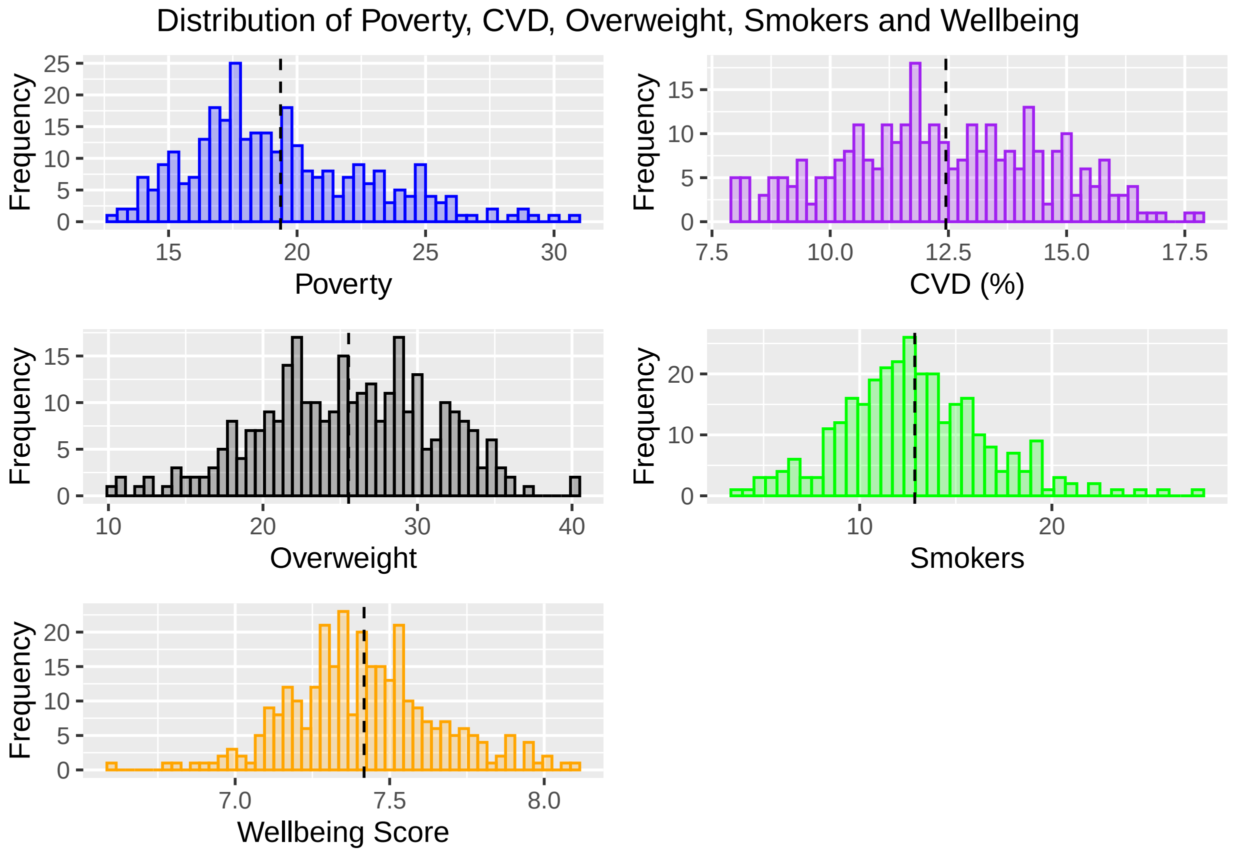
```r
#Dropping missing values in the columns that are important for our analyses
data1_clean <- data1 %>%
  drop_na(Poverty, CVD, overweight, smokers, wellbeing)
```

```r
#Z-Score Method to Identify Outliers
#Declaring the variables under consideration
variables <- c("Poverty", "CVD", "overweight", "smokers", "wellbeing")
#Defining the dataframe
outliers_df <- data.frame(Variable = character(), Outlier_Values = character())

for (var in variables) {
  # Checking if the variable exists and is numeric
  if (var %in% colnames(data1_clean) && is.numeric(data1_clean[[var]])) {
    z_scores <- scale(data1_clean[[var]])
    sigma <- 3
    outliers <- data1_clean[[var]][abs(z_scores) > sigma]

    if (length(outliers) > 0) {
      outliers_df <- rbind(outliers_df, data.frame(Variable = var, Outlier_Values = paste(outliers, col
    } else {
      outliers_df <- rbind(outliers_df, data.frame(Variable = var, Outlier_Values = "No outliers found")
    }
  } else {
    outliers_df <- rbind(outliers_df, data.frame(Variable = var, Outlier_Values = "Variable not found o
  }
}
#Displaying the dataframe containing the outliers
kable(outliers_df)
```

Variable	Outlier_Values
Poverty	30, 30.7
CVD	No outliers found
overweight	No outliers found
smokers	27.8, 25.8, 24.8
wellbeing	6.61

While the data contains a few outliers, discarding them would introduce bias due to the unique nature of each area. Therefore, the analysis will proceed with the full dataset, including outliers.

```r
#Visualising the Data
grid.arrange(
  # Poverty
  ggplot(data1_clean, aes(Poverty)) +
    geom_histogram(binwidth = 0.4, color = "blue", fill = "blue", alpha = 0.25)  +
    geom_vline(data = data1_clean, aes(xintercept = mean(Poverty)), linetype = "dashed") +
    labs(x = "Poverty", y = "Frequency"),
  # CVD
  ggplot(data1_clean, aes(CVD)) +
    geom_histogram(binwidth = 0.2, color = "purple", fill = "purple", alpha = 0.25)  +
    geom_vline(data = data1_clean, aes(xintercept = mean(CVD)), linetype = "dashed") +
    labs(x = "CVD (%)", y = "Frequency"),
  # Overweight
  ggplot(data1_clean, aes(overweight)) +
    geom_histogram(binwidth = 0.6, color = "black", fill = "black", alpha = 0.25) +
    geom_vline(data = data1_clean, aes(xintercept = mean(overweight)), linetype = "dashed") +
    labs(x = "Overweight", y = "Frequency"),
  # Smokers
  ggplot(data1_clean, aes(smokers)) +
    geom_histogram(binwidth = 0.6, color = "green", fill = "green", alpha = 0.25)  +
    geom_vline(data = data1_clean, aes(xintercept = mean(smokers)), linetype = "dashed") +
    labs(x = "Smokers", y = "Frequency"),
  # Well being
  ggplot(data1_clean, aes(wellbeing)) +
    geom_histogram(binwidth = 0.03, color = "orange", fill = "orange", alpha = 0.25)  +
    geom_vline(data = data1_clean, aes(xintercept = mean(wellbeing)), linetype = "dashed") +
    labs(x = "Wellbeing Score", y = "Frequency"),
  ncol = 2, nrow = 3,
  top = textGrob("Distribution of Poverty, CVD, Overweight, Smokers and Wellbeing")
)
```



A preliminary visual inspection of the histograms revealed the following:

- **Poverty:** The histogram exhibits moderate symmetry with a slight skew (not normal).
- **CVD:** The distribution appears approximately bell-shaped, suggesting potential normality, although slight skewness or heavier tails cannot be ruled out.
- **Overweight:** The histogram is relatively symmetrical and bell-shaped, indicating a possible normal distribution.
- **Smokers:** The distribution appears skewed towards lower values, suggesting non-normality.
- **Wellbeing:** The histogram is moderately symmetrical with a central peak, potentially indicating a normal distribution.

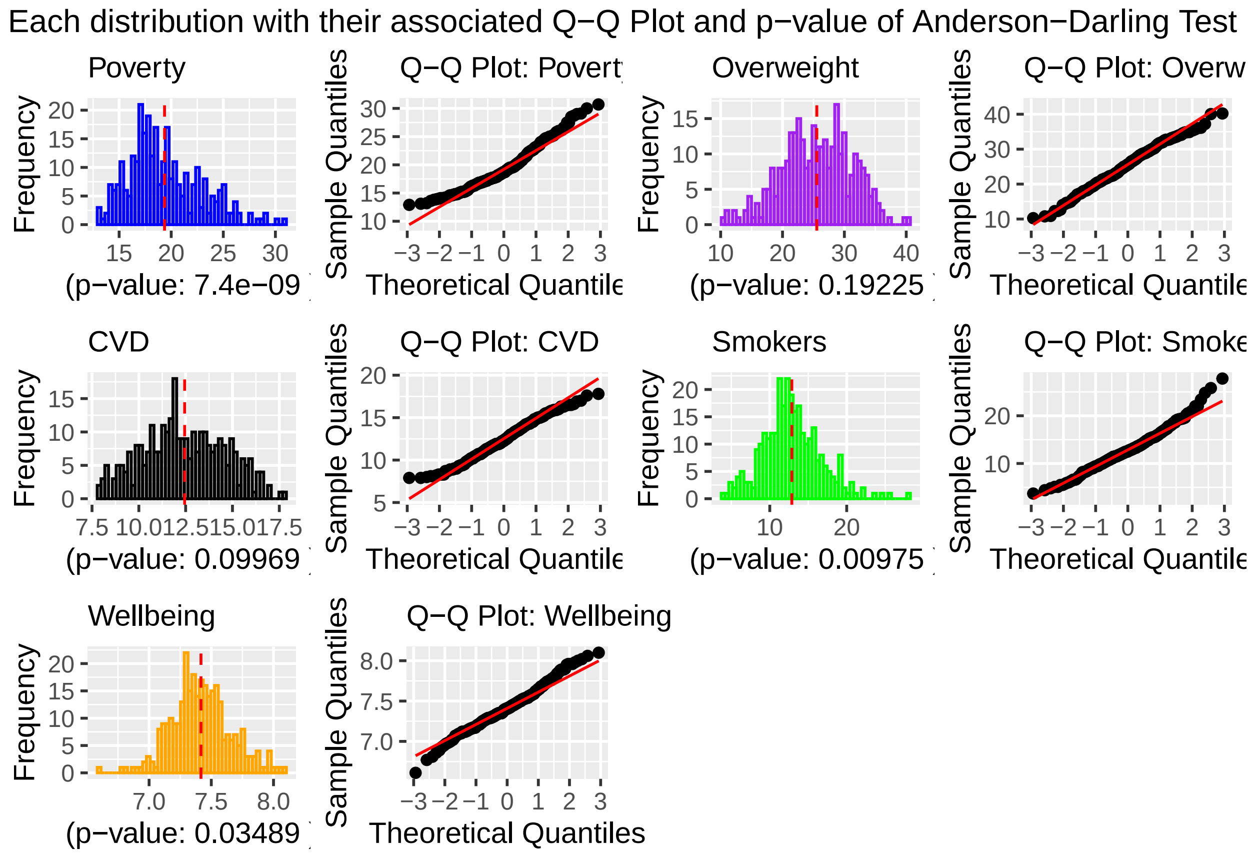
While the preliminary visual inspection provides initial insights, a more rigorous statistical analysis is necessary to confirm the distributional assumptions. To this end, we will employ the Anderson-Darling test in conjunction with Q-Q plots to assess the normality of each dataset.

```r
# Poverty
ad_test_result_1 <- ad.test(data1_clean$Poverty)
ad_p_value_1 <- round(ad_test_result_1$p.value, 10)

# Overweight
ad_test_result_2 <- ad.test(data1_clean$overweight)
ad_p_value_2 <- round(ad_test_result_2$p.value, 5)

# CVD
ad_test_result_3 <- ad.test(data1_clean$CVD)
ad_p_value_3 <- round(ad_test_result_3$p.value, 5)
```

```r
# Smokers
ad_test_result_4 <- ad.test(data1_clean$smokers)
ad_p_value_4 <- round(ad_test_result_4$p.value, 5)

# Wellbeing
ad_test_result_5 <- ad.test(data1_clean$wellbeing)
ad_p_value_5 <- round(ad_test_result_5$p.value, 5)

# Function to create both histogram and Q-Q plot for a variable
plot_variable <- function(data, variable_name, ad_p_value, color, subtitle) {
  # Histogram
  hist_plot <- ggplot(data, aes_string(variable_name)) +
    geom_histogram(bins = 50, color = color, fill = color, alpha = 0.25) +
    geom_vline(aes_string(xintercept = paste0("mean(", variable_name, ", na.rm = TRUE)")), linetype = "
    labs(
      x = paste("(p-value:", ad_p_value, ")"),
      y = "Frequency",
      subtitle = subtitle
    )

  # Q-Q Plot
  qq_plot <- ggplot(data, aes_string(sample = variable_name)) +
    stat_qq() +
    stat_qq_line(color = "red") +
    labs(x = "Theoretical Quantiles", y = "Sample Quantiles", subtitle = paste("Q-Q Plot:", subtitle))

  return(list(hist_plot, qq_plot))
}

# Generating plots for each variable
plots_poverty <- plot_variable(data1_clean, "Poverty", ad_p_value_1, "blue", "Poverty")
```

```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with ‘aes()‘.
## i See also ‘vignette("ggplot2-in-packages")‘ for more information.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```r
plots_overweight <- plot_variable(data1_clean, "overweight", ad_p_value_2, "purple", "Overweight")
plots_cvd <- plot_variable(data1_clean, "CVD", ad_p_value_3, "black", "CVD")
plots_smokers <- plot_variable(data1_clean, "smokers", ad_p_value_4, "green", "Smokers")
plots_wellbeing <- plot_variable(data1_clean, "wellbeing", ad_p_value_5, "orange", "Wellbeing")

# Arranging all plots in a grid
grid.arrange(
  plots_poverty[[1]], plots_poverty[[2]],
  plots_overweight[[1]], plots_overweight[[2]],
  plots_cvd[[1]], plots_cvd[[2]],
  plots_smokers[[1]], plots_smokers[[2]],
  plots_wellbeing[[1]], plots_wellbeing[[2]],
  ncol = 4,
  top = textGrob("Each distribution with their associated Q-Q Plot and p-value of Anderson-Darling Test
)
```



The results indicate the following:

- **Poverty:** The Q-Q plot shows significant deviations from the diagonal line, particularly in the tails. The p-value of 7.4×10^{-9} from the Anderson-Darling test strongly supports the conclusion that the distribution is non-normal.
- **Cardiovascular Disease (CVD):** The Q-Q plot closely follows the diagonal line, indicating normality. The p-value of 0.09969 from the Anderson-Darling test suggests that the distribution is approximately normal.
- **Overweight:** The Q-Q plot shows that the points generally follow the diagonal line, indicating approximate normality. The p-value of 0.19225 from the Anderson-Darling test further supports this conclusion.
- **Smokers:** The Q-Q plot exhibits deviations from the diagonal line, particularly in the tails. The p-value of 0.00975 from the Anderson-Darling test suggests that the distribution is not normally distributed.
- **Wellbeing:** The Q-Q plot shows deviations from the diagonal line, indicating non-normality. The p-value of 0.03489 from the Anderson-Darling test supports this conclusion.

Exploratory Correlation Analysis prior to Multiple Linear Regression

Next, we examine the correlations between the variables: CVD, overweight, smokers, wellbeing, and poverty.

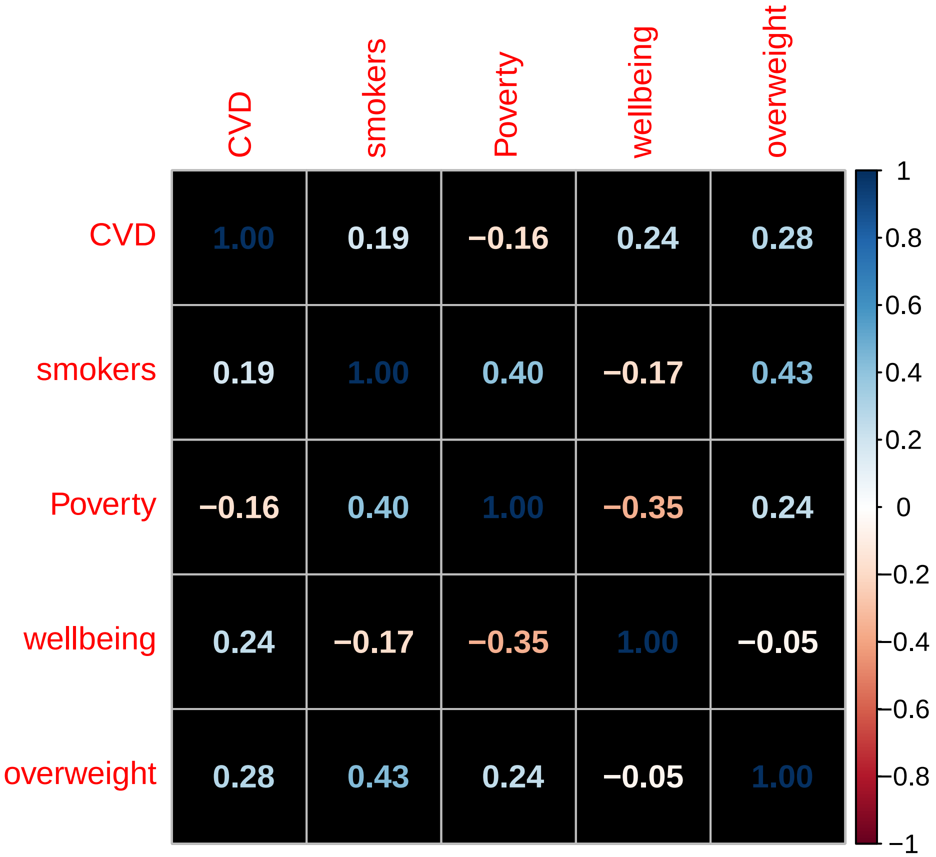
Since only the overweight and CVD variables were previously determined to be normally distributed, we use Spearman's rank correlation coefficient to assess the relationships between the other variables and CVD. Spearman's correlation is appropriate here because it does not assume normality.

```
# Calculating correlations and checking the statistical significance
corr <- rcorr(as.matrix(data1_clean %>% select(CVD, smokers, Poverty, wellbeing, overweight)), type = "
#Extracting and comparing the p-values (TRUE indicates statistically significant and FALSE indicates st
kable(corr$P<0.05)
```

	CVD	smokers	Poverty	wellbeing	overweight
CVD	NA	TRUE	TRUE	TRUE	TRUE
smokers	TRUE	NA	TRUE	TRUE	TRUE
Poverty	TRUE	TRUE	NA	TRUE	TRUE
wellbeing	TRUE	TRUE	TRUE	NA	FALSE
overweight	TRUE	TRUE	TRUE	FALSE	NA

We observe that most of the associated p-values are less than 0.05, indicating that most of the correlations are statistically significant. However, the correlation between overweight and wellbeing was not statistically significant (p>0.05), indicating that this relationship may be due to chance.

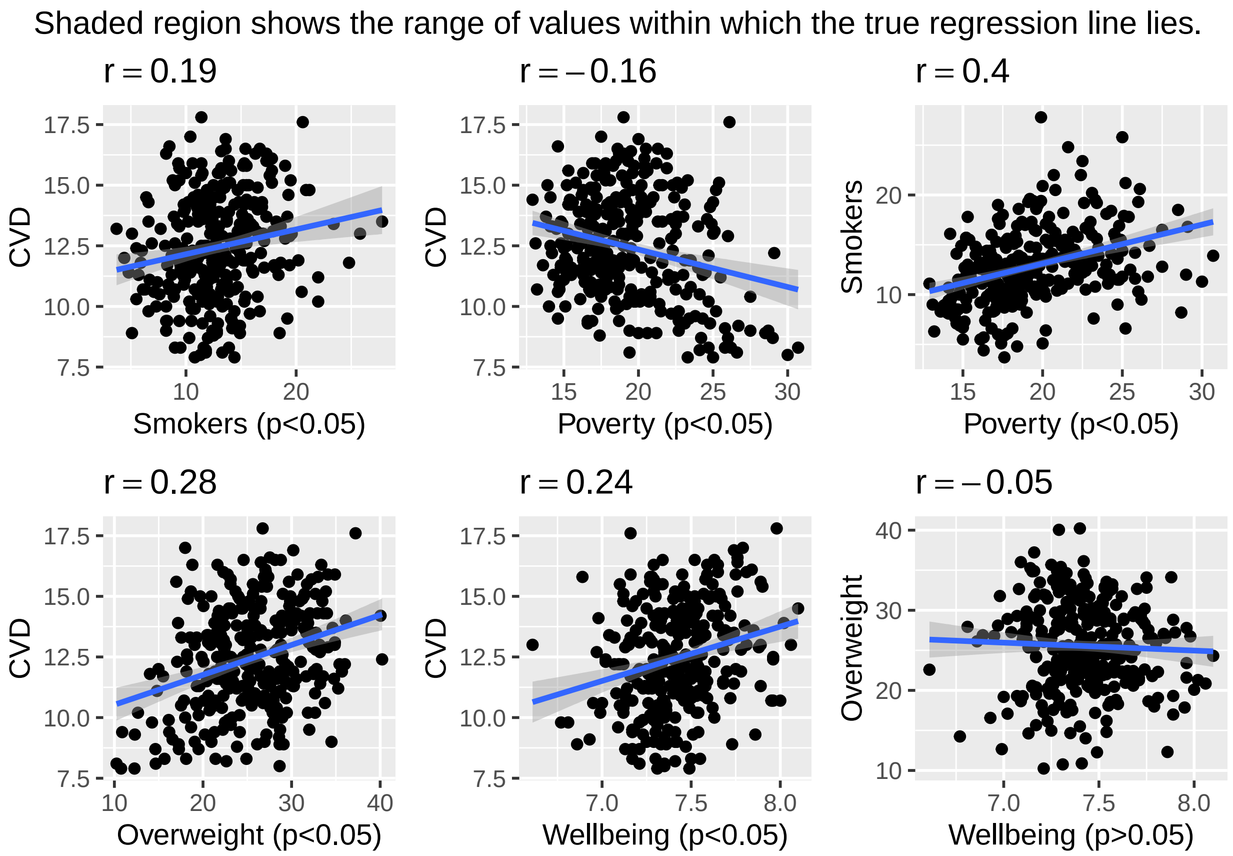
```
#Extracting the coefficients (r) in a matrix
corr_matrix <- corr$r
#Correlation Heatmap
corrplot(corr_matrix, method = "number", bg = "black")
```



	CVD	smokers	Poverty	wellbeing	overweight
CVD	1.00	0.19	−0.16	0.24	0.28
smokers	0.19	1.00	0.40	−0.17	0.43
Poverty	−0.16	0.40	1.00	−0.35	0.24
wellbeing	0.24	−0.17	−0.35	1.00	−0.05
overweight	0.28	0.43	0.24	−0.05	1.00

```r
#Visualising the correlation between each variable
grid.arrange(
  #CVD vs Smokers
    ggplot(data1_clean, aes(y=CVD, x=smokers)) + geom_point() + labs(x="Smokers (p<0.05)", y="CVD", titl
    #CVD vs Poverty
    ggplot(data1_clean, aes(y=CVD, x=Poverty)) + geom_point() + labs(x="Poverty (p<0.05)", y="CVD", titl
    #Smokers vs Poverty
    ggplot(data1_clean, aes(y=smokers, x=Poverty)) + geom_point() + labs(x="Poverty (p<0.05)", y="Smoke
    #CVD vs Overweight
    ggplot(data1_clean, aes(y=CVD, x=overweight)) + geom_point() + labs(x="Overweight (p<0.05)", y="CVD
    #CVD vs Wellbeing
    ggplot(data1_clean, aes(y=CVD, x=wellbeing)) + geom_point() + labs(x="Wellbeing (p<0.05)", y="CVD",
    #Overweight vs Wellbeing
     ggplot(data1_clean, aes(y=overweight, x=wellbeing)) + geom_point() + labs(x="Wellbeing (p>0.05)", y
    ncol=3, top = textGrob("Shaded region shows the range of values within which the true regression li
)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



- **CVD and Smokers:** A weaker positive correlation ($r = 0.19$, $p < 0.05$) was observed between CVD and smokers, suggesting a potential link between smoking and CVD risk.

- **CVD and Poverty:** A weak negative correlation (r = -0.16, $p < 0.05$) was detected between CVD and poverty, implying a slightly lower prevalence of CVD among individuals in poverty.
- **Smokers and Poverty:** A strong positive correlation (r = 0.40, $p < 0.05$) was found between smoking and poverty, suggesting that individuals in poverty are more likely to be smokers.
- **CVD and Overweight:** A moderate positive correlation (r = 0.28, $p < 0.05$) was found between CVD and overweight, indicating that individuals with higher body weight are more likely to experience CVD.
- **CVD and Wellbeing:** A moderate positive correlation (r = 0.24, $p < 0.05$) was identified between CVD and wellbeing, indicating a slight positive association between the two. As one variable increases, the other tends to increase slightly.
- **Overweight and Wellbeing:** A negligible negative correlation (r = -0.05, $p > 0.05$) was observed between overweight and wellbeing, suggesting no statistically significant relationship between the two.

Now, we cannot use correlation alone to predict, as correlation does not imply causation or say anything much about the shape of the relationship between the variables. In order to do that, we will have to perform regression. Since, we are interested in identifying which of these factors (overweight, smokers, wellbeing, and poverty) affect the prevalence of CVD in an area, we will be using Multiple Linear Regression.

Multiple Linear Regression

```r
#Fitting the model for Multiple Linear Regression to check the effects of Overweight, Smokers, Wellbein
data1_clean_reg <- lm(CVD ~ overweight + smokers + wellbeing + Poverty, data = data1_clean)
#Checking for Multicollinearity
kable(vif(data1_clean_reg))
```

	x
overweight	1.196546
smokers	1.359206
wellbeing	1.145506
Poverty	1.263094

Since all the values are close to 1, it indicates little or no multicollinearity.

```r
#Regression Results
summary(data1_clean_reg)
```

```
##
## Call:
## lm(formula = CVD ~ overweight + smokers + wellbeing + Poverty,
##     data = data1_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3966 -1.4313 -0.1097  1.3905  4.7779
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.69780    3.93361  -0.432 0.666335
## overweight   0.10985    0.02123   5.174 4.22e-07 ***
## smokers      0.12030    0.03366   3.574 0.000410 ***
## wellbeing    1.80025    0.49096   3.667 0.000291 ***
## Poverty     -0.18400    0.03515  -5.234 3.13e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.894 on 298 degrees of freedom
## Multiple R-squared:  0.2494, Adjusted R-squared:  0.2393
## F-statistic: 24.75 on 4 and 298 DF,  p-value: < 2.2e-16
```

```
#Confidence Intervals for Null Hypothesis Significance Testing
kable(cbind(coefficient=coef(data1_clean_reg), confint(data1_clean_reg)))
```

	coefficient	2.5 %	97.5 %
(Intercept)	-1.6977972	-9.4389651	6.0433707
overweight	0.1098530	0.0680676	0.1516385
smokers	0.1203042	0.0540571	0.1865512
wellbeing	1.8002479	0.8340507	2.7664450
Poverty	-0.1839990	-0.2531784	-0.1148195

Model

$$\text{CVD} = \beta_{\text{Intercept}} + \beta_{\text{Overweight}} \times \text{Overweight} + \beta_{\text{Smokers}} \times \text{Smokers} + \beta_{\text{Wellbeing}} \times \text{Wellbeing} + \beta_{\text{Poverty}} \times \text{Poverty} + \epsilon$$

Intercept: The model predicts a baseline CVD value of -1.6978 when all predictor variables are zero. However, this intercept is not statistically significant (p > 0.001), suggesting it may not have practical significance.

Predictor Variables

- **Overweight:** There is a significant positive association between being overweight and CVD. A one-unit increase in the overweight variable is associated with a 0.11 unit increase in CVD, holding other factors constant $(\beta = 0.11, t(298) = 5.17, p < 0.001)$ (95% CI: [0.07, 0.15]).
- **Smokers:** There is a strong positive association between smoking and CVD. A one-unit increase in the smokers variable is associated with a 0.12 unit increase in CVD, holding other factors constant $(\beta = 0.12, t(298) = 3.57, p < 0.001)$(95% CI: [0.05, 0.19]).
- **Wellbeing:** There is a strong positive association between wellbeing and CVD. A one-unit increase in the wellbeing variable is associated with a 1.80 unit increase in CVD, holding other factors constant $(\beta = 1.8, t(298) = 3.67, p < 0.001)$(95% CI: [0.83, 2.77]).
- **Poverty:** There is a significant negative association between poverty and CVD. A one-unit increase in the poverty variable is associated with a 0.18 unit decrease in CVD, holding other factors constant $(\beta = 0.18, t(298) = -5.23, p < 0.001)$(95% CI: [-0.25, -0.11]).

Model Fit

- **Residual Standard Error:** The model's predictions deviate from the actual values by an average of 1.894 units.

- **Multiple R-squared:** Approximately 25% of the variability in CVD can be explained by the combined effects of the predictor variables.
- **Adjusted R-squared:** After adjusting for the number of predictors, the model explains approximately 24% of the variability in CVD.
- **F-statistic:** The overall model is statistically significant, indicating that at least one of the predictor variables is significantly associated with CVD. $(F(4, 298) = 24.75, p < 0.001)$

Inference

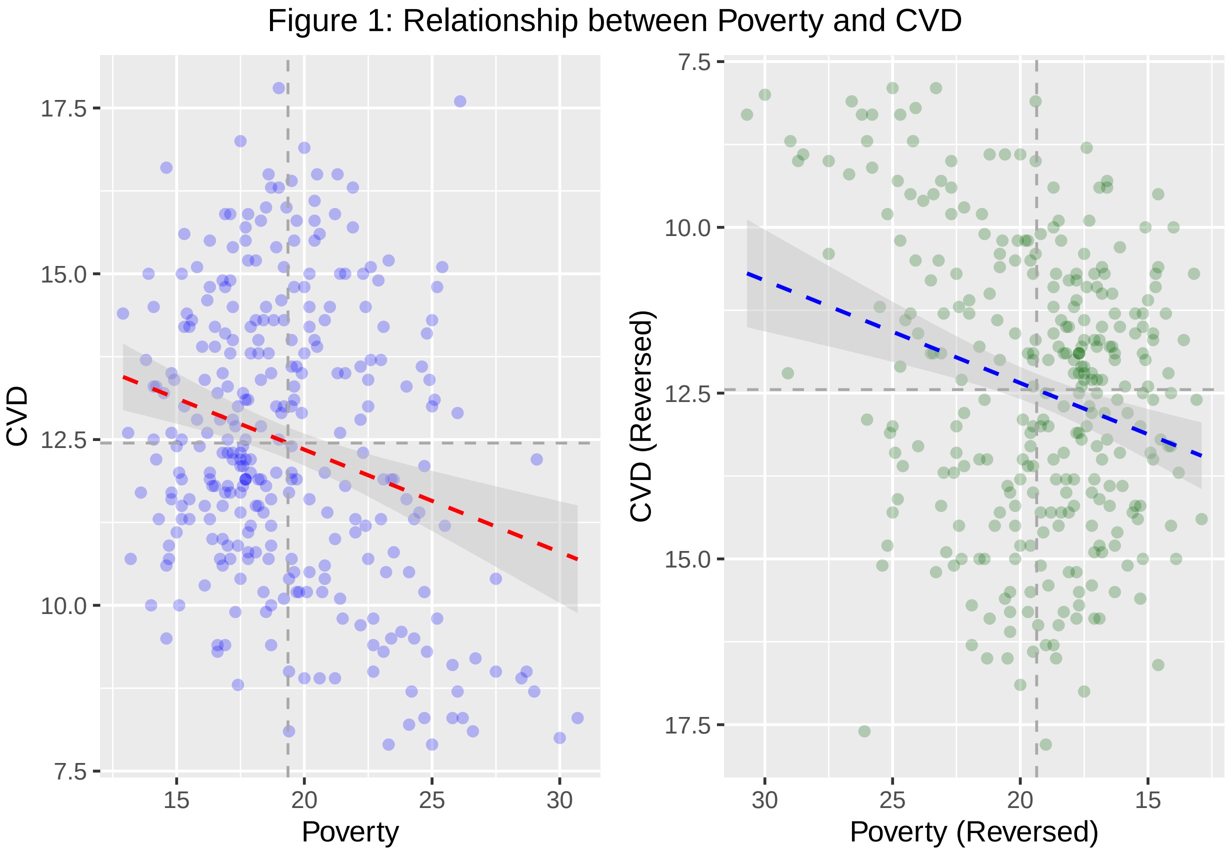
Our analysis reveals that being overweight and a smoker significantly increases the risk of cardiovascular disease (CVD), as expected. We also encounter a problem that only 25% of the variability in CVD can be explained by the combined effects of the predictor variables. However, two unexpected findings warrant further investigation:

- **Poverty and CVD:** We observed a significant positive correlation between poverty and smoking, as well as a moderate positive correlation between CVD and smoking. However, contrary to expectations, our model indicates that individuals in poverty have a lower risk of CVD. This discrepancy suggests a complex relationship between poverty, smoking, and CVD, highlighting the need for further research to understand the underlying factors.

```
#Visualization of the relationship of CVD and Poverty using Scatter Plots
grid.arrange(
  #Both scales increasing
  ggplot(data1_clean, aes(x= Poverty, CVD)) +
    geom_point(alpha = 0.25, color = "blue") +
    geom_smooth(method=lm, color = "red", fill = "grey", linetype = 2, size = 0.7) +
    geom_vline(data = data1_clean, aes(xintercept = mean(Poverty)), color = "dark grey", linetype = "da
    geom_hline(data = data1_clean, aes(yintercept = mean(CVD)), color = "dark grey", linetype = "dashed
    labs(x = "Poverty", y = "CVD"),
  #Both scales decreasing (Reverse X & Y-Axis)
   ggplot(data1_clean, aes(x= Poverty, CVD)) +
    geom_point(alpha = 0.25, color = "darkgreen") +
    geom_smooth(method=lm, color = "blue", fill = "grey", linetype = 2, size = 0.7) +
    geom_vline(data = data1_clean, aes(xintercept = mean(Poverty)), color = "dark grey", linetype = "da
    geom_hline(data = data1_clean, aes(yintercept = mean(CVD)), color = "dark grey", linetype = "dashed
    scale_y_reverse() + scale_x_reverse() +
    labs(x = "Poverty (Reversed)", y = "CVD (Reversed)"),
  ncol = 2,
  top = textGrob("Figure 1: Relationship between Poverty and CVD"
))
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

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



- **Wellbeing and CVD:** Surprisingly, higher wellbeing scores were associated with an increased risk of CVD. This could be attributed to various factors, such as increased access to unhealthy foods, sedentary lifestyles, and stress associated with modern living. Additional research is needed to explore this intriguing relationship and identify potential mediating factors.

Future Work Prospect

Given the conflicting results from the regression model, a more in-depth analysis is warranted. A potential cause for these discrepancies could be heteroscedasticity.

```
#Breusch-Pagan Test to check for heteroscedasticity
bptest(data1_clean_reg)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  data1_clean_reg
## BP = 9.6764, df = 4, p-value = 0.04625
```

The observed p-value of 0.046 is less than the significance level of 0.05, leading to the rejection of the null hypothesis of homoscedasticity,indicating our model has heteroscedasticity.

```
#Checking the relationship of smokers and poverty
data1_clean_reg_2 <- lm(smokers ~ Poverty, data = data1_clean)
summary(data1_clean_reg_2)
```

```
##
## Call:
## lm(formula = smokers ~ Poverty, data = data1_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5519 -2.3794 -0.1947  2.0462 14.7225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.28882    1.14512   4.619 5.73e-06 ***
## Poverty      0.39139    0.05822   6.723 8.96e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.525 on 301 degrees of freedom
## Multiple R-squared:  0.1305, Adjusted R-squared:  0.1277
## F-statistic: 45.19 on 1 and 301 DF,  p-value: 8.958e-11
```

```
#Checking the relationship of CVD with Poverty and smokers, assuming an interaction between them
data1_clean_reg_3 <- lm(CVD ~ Poverty * smokers, data = data1_clean)
summary(data1_clean_reg_3)
```

```
##
## Call:
## lm(formula = CVD ~ Poverty * smokers, data = data1_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3837 -1.4724 -0.2688  1.4757  5.5552
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.053827   2.386174   6.728 8.77e-11 ***
## Poverty         -0.303727   0.123166  -2.466   0.0142 *
## smokers          0.052630   0.186479   0.282   0.7780
## Poverty:smokers  0.006287   0.009280   0.677   0.4986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.017 on 299 degrees of freedom
## Multiple R-squared:  0.1451, Adjusted R-squared:  0.1366
## F-statistic: 16.92 on 3 and 299 DF,  p-value: 3.54e-10
```

- **Poverty and Smoking:** The first model confirms a significant positive relationship between poverty and smoking, aligning with the observed correlation.
- **CVD and Poverty:** The second model shows a significant negative relationship between poverty and CVD, which is unexpected and suggests a complex relationship that may involve other mediating or confounding factors.

- **Interaction:** The lack of significant interaction between poverty and smoking in affecting CVD suggests that the relationship between these variables and CVD might not be straightforward and could involve other unmeasured variables.

```
#Checking for possible non-linear relationship
gam_model <- gam(CVD ~ s(Poverty, smokers), data = data1_clean)
summary(gam_model)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## CVD ~ s(Poverty, smokers)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4469     0.1119   111.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df     F p-value
## s(Poverty,smokers) 9.603  13.37 5.507  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.195   Deviance explained =   22%
## GCV = 3.9339  Scale est. = 3.7963    n = 303
```

Conclusion

- **Significant Interaction:** The combined effect of poverty and smokers on CVD is significant and non-linear.
- **Model Fit:** The model provides a moderate explanation of the variance in CVD, emphasizing the importance of considering complex interactions.

Question 2

Data Dictionary

Overview

- **Dataset Name:** Customer Satisfaction Data from a furniture retail company.
- **Source:** myWBS

Variable Definitions

Variable Name	Description	Data Type	Valid Values
SES_category	The Company’s categorization of store type by local socio-economic-status (low, medium and high).	Character	Low, Medium and High
customer.satisfaction	Average customer satisfaction score	Number/Numeric	3.76-9.67
staff.satisfaction	Average staff job satisfaction score	Number/Numeric	4.85-8.86
delivery.time	Average delivery time of large and custom items	Number/Numeric	32.96-92.48
CVD	Percentage of people living in the area who have recently experienced Cardiovascular Disease	Number/Numeric	7.9- 17.8
new_range	Whether the store carries a new range of products	Logical	TRUE and FALSE

```r
#Importing the data
data2 <- read_csv("~/Downloads/BS Assignment/cust_satisfaction.csv")
```

```
## Rows: 300 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): SES_category
## dbl (3): customer.satisfaction, staff.satisfaction, delivery.time
## lgl (1): new_range
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
#Checking data for duplicates
duplicates_2 <- data2[duplicated(data2), ]
duplicates_2
```

```
## # A tibble: 0 x 5
## # i 5 variables: SES_category <chr>, customer.satisfaction <dbl>,
## #   staff.satisfaction <dbl>, delivery.time <dbl>, new_range <lgl>
```

No duplicate values were found in the data.

```r
#Checking the data for missing values
summary(is.na(data2))
```

```
##  SES_category    customer.satisfaction staff.satisfaction delivery.time
##  Mode :logical   Mode :logical         Mode :logical      Mode :logical
##  FALSE:300       FALSE:300             FALSE:300          FALSE:300
##  new_range
##  Mode :logical
##  FALSE:300
```

No missing values were identified.

```r
#Checking the structure of the data
str(data2)
```

```
## spc_tbl_ [300 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ SES_category         : chr [1:300] "Medium" "Medium" "Medium" "High" ...
##  $ customer.satisfaction: num [1:300] 7.27 7.93 7.12 6.35 6.78 ...
##  $ staff.satisfaction   : num [1:300] 6.88 7.44 7.15 6.47 7.06 ...
##  $ delivery.time        : num [1:300] 66.7 68.2 70.7 61.4 57.7 ...
##  $ new_range            : logi [1:300] FALSE FALSE FALSE TRUE TRUE TRUE ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   SES_category = col_character(),
##   ..   customer.satisfaction = col_double(),
##   ..   staff.satisfaction = col_double(),
##   ..   delivery.time = col_double(),
##   ..   new_range = col_logical()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

We observed that 'SES_category' is a character variable and 'new_range' is a logical variable. Typically, we would convert 'SES_category' to an ordinal variable and 'new_range' to a binary variable for ease of analysis. However, due to our small data size, we will use only the numeric data for the correlation check. For the rest of the analysis, we will avoid using the converted numerical data to maintain greater accuracy.

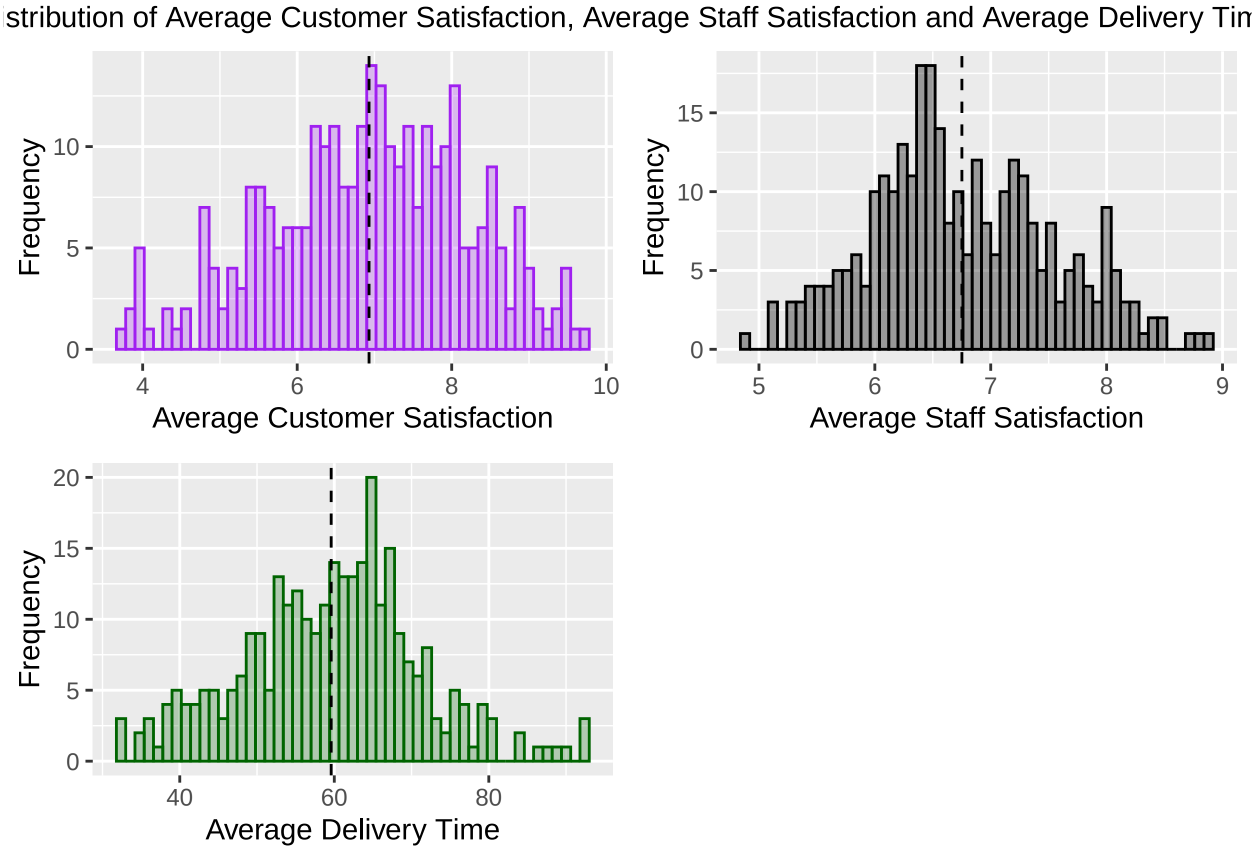
```r
#Converting SES_category to ordinal variable with 0, 1 & 2 signifying Low, Medium and High respectively
data2_new <- data2 %>%
  mutate(SES_category = case_when(
  SES_category == "Low" ~ 0,
  SES_category == "Medium" ~ 1,
  SES_category == "High" ~ 2
  ))
#Converting new_range to binary variable with 0 & 1 signifying FALSE & TRUE respectively.
data2_new <- data2_new %>%
  mutate(new_range = case_when(
  new_range == "TRUE" ~ 1,
  new_range == "FALSE" ~ 0
  ))


#Z-Score Method to check for Outliers
```

```r
#Declaring the variables under consideration (Skipping discrete variables)
variables_2 <- c("customer.satisfaction", "staff.satisfaction", "delivery.time")
#Defining the dataframe
outliers_df_2 <- data.frame(Variable = character(), Outlier_Values = character())

for (var in variables_2) {
  # Checking if the variable exists and is numeric
  if (var %in% colnames(data2) && is.numeric(data2[[var]])) {
    z_scores <- scale(data2[[var]])
    sigma <- 3
    outliers <- data2[[var]][abs(z_scores) > sigma]

    if (length(outliers) > 0) {
      outliers_df_2 <- rbind(outliers_df_2, data.frame(Variable = var, Outlier_Values = paste(outliers,
    } else {
      outliers_df_2 <- rbind(outliers_df_2, data.frame(Variable = var, Outlier_Values = "No outliers fou
    }
  } else {
    outliers_df_2 <- rbind(outliers_df_2, data.frame(Variable = var, Outlier_Values = "Variable not fou
  }
}
#Displaying the dataframe containing the outliers
kable(outliers_df_2)
```

Variable	Outlier_Values
customer.satisfaction	No outliers found
staff.satisfaction	No outliers found
delivery.time	No outliers found

No Outliers were identified.

```r
#Visualising the data (Skipping SES_category and new_range as they are discrete variables)
grid.arrange(
  # Customer Satisfaction
  ggplot(data2, aes(customer.satisfaction)) +
    geom_histogram(binwidth = 0.12, color = "purple", fill = "purple", alpha = 0.25) +
    geom_vline(aes(xintercept = mean(customer.satisfaction, na.rm = TRUE)), linetype = "dashed") +
    labs(x = "Average Customer Satisfaction", y = "Frequency"),

  # Staff Satisfaction
  ggplot(data2, aes(staff.satisfaction)) +
    geom_histogram(binwidth = 0.08, color = "black", fill = "black", alpha = 0.35) +
    geom_vline(aes(xintercept = mean(staff.satisfaction, na.rm = TRUE)), linetype = "dashed") +
    labs(x = "Average Staff Satisfaction", y = "Frequency"),

  # Delivery Time
  ggplot(data2, aes(delivery.time)) +
    geom_histogram(binwidth = 1.2, color = "darkgreen", fill = "darkgreen", alpha = 0.25) +
    geom_vline(aes(xintercept = mean(delivery.time, na.rm = TRUE)), linetype = "dashed") +
    labs(x = "Average Delivery Time", y = "Frequency"),
  ncol = 2,
  top = textGrob("Distribution of Average Customer Satisfaction, Average Staff Satisfaction and Average
)
```



A preliminary visual inspection of the histograms revealed the following:

- **Average Customer Satisfaction:** The distribution is roughly symmetric with a central peak, resembling a bell shape. It appears fairly normal, though there are some irregularities at the tails.
- **Average Staff Satisfaction:** The distribution is somewhat symmetric but exhibits more variability and irregularities compared to customer satisfaction. It is less normal, with noticeable fluctuations.
- **Average Delivery Time:** The distribution is symmetric and bell-shaped, centered around the mean. This distribution appears the most normal, closely following a typical bell curve.

To confirm the normality of the distributions, we performed the Kolmogorov-Smirnov test.

```
#Kolmogorov-Smirnov Test
#Customer Satisfaction
k1 <- ks.test(data2_new$customer.satisfaction, "pnorm", mean=mean(data2_new$customer.satisfaction, sd=s
#Checking if p-value is less than 0.05 (TRUE indicates normality)
k1$p.value < 0.05
```

```
## [1] TRUE
```

```r
#Staff Satisfaction
k2 <- ks.test(data2_new$staff.satisfaction, "pnorm", mean=mean(data2_new$staff.satisfaction, sd=sd(data2
#Checking if p-value is less than 0.05 (TRUE indicates normality)
k2$p.value < 0.05
```

```
## [1] TRUE
```

```r
#Delivery Time
k3 <- ks.test(data2_new$delivery.time, "pnorm", mean=mean(data2_new$delivery.time, sd=sd(data2$delivery
```

```
## Warning in ks.test.default(data2_new$delivery.time, "pnorm", mean =
## mean(data2_new$delivery.time, : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
```

```r
#Checking if p-value is less than 0.05 (TRUE indicates normality)
k3$p.value < 0.05
```

```
## [1] TRUE
```

We confirmed that all the aforementioned columns are normally distributed.

Exploratory Correlation Analysis prior to Multiple Linear Regression

Next, we examine the correlations between the following variables and customer satisfaction: SES_category, staff satisfaction, delivery time, and new_range.

Since SES_category and new_range are discrete variables, we will use Spearman's rank correlation coefficient to assess their relationships. For the other variables that were previously determined to be normally distributed, we will use Pearson's correlation coefficient.

```r
# Calculating correlations for discrete variables with Customer Satisfaction and checking the statistic
corr_non_normal <- rcorr(as.matrix(data2_new %>%
  select(SES_category, customer.satisfaction, new_range)), type = "spearman")
#Extracting and comparing the p-values (TRUE indicates statistically significant and FALSE indicates st
kable(corr_non_normal$P<0.05)
```

	SES_category	customer.satisfaction	new_range
SES_category	NA	FALSE	FALSE
customer.satisfaction	FALSE	NA	FALSE
new_range	FALSE	FALSE	NA

We observe that all of the associated p-values are greater than 0.05, indicating that the correlations are not statistically significant.

```r
# Calculating correlations for normally distributed data and checking the statistical significance
corr_normal <- rcorr(as.matrix(data2_new %>%
  select(customer.satisfaction, staff.satisfaction, delivery.time)), type = "pearson")
#Extracting and comparing the p-values (TRUE indicates statistically significant and FALSE indicates st
kable(corr_normal$P<0.05)
```

	customer.satisfaction	staff.satisfaction	delivery.time
customer.satisfaction	NA	TRUE	TRUE
staff.satisfaction	TRUE	NA	FALSE
delivery.time	TRUE	FALSE	NA

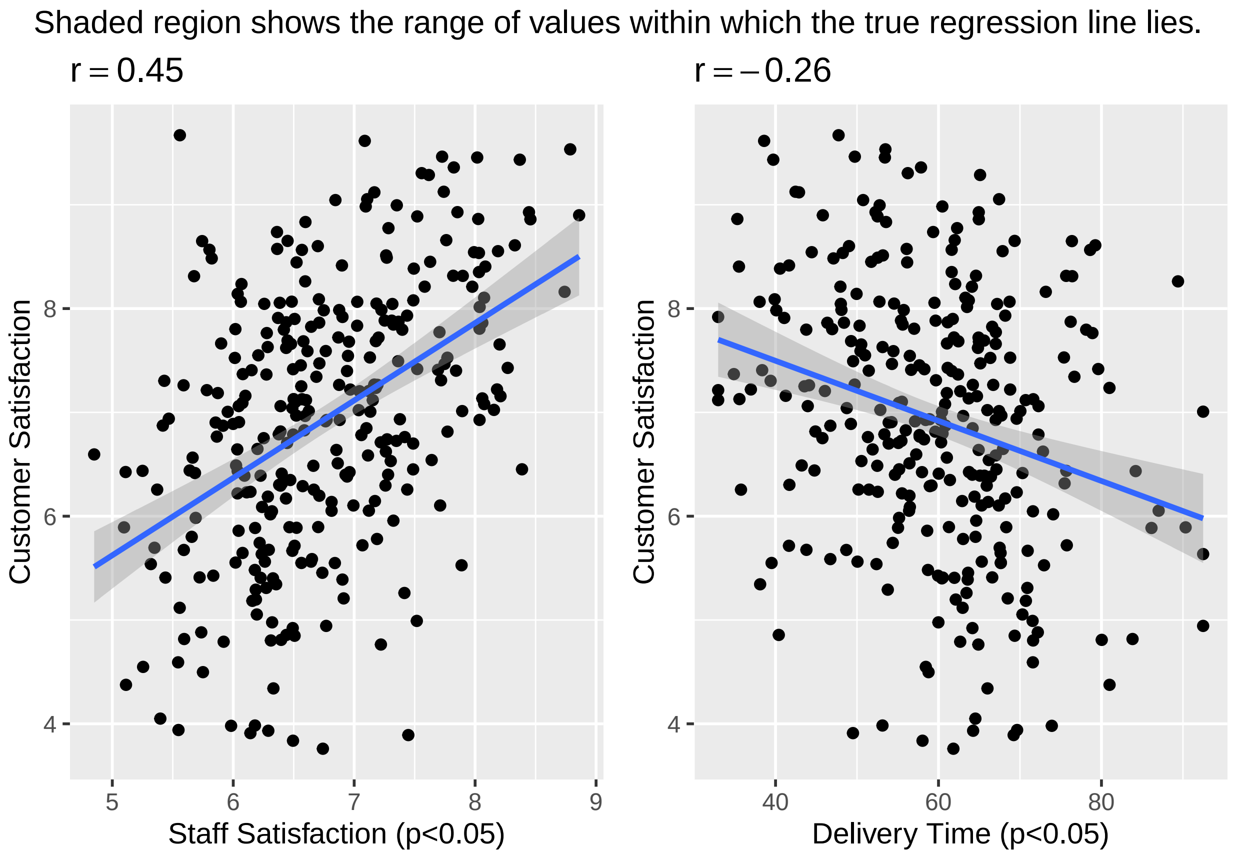
We observed that most of the associated p-values are less than 0.05, indicating that the majority of the correlations are statistically significant. However, the correlation between delivery time and staff satisfaction was not statistically significant (p > 0.05), suggesting that this relationship may be due to chance.

```
#Displaying the coefficient (r) values after rounding off to 2 decimal places
kable(round(corr_normal$r,2))
```

	customer.satisfaction	staff.satisfaction	delivery.time
customer.satisfaction	1.00	0.45	-0.26
staff.satisfaction	0.45	1.00	-0.07
delivery.time	-0.26	-0.07	1.00

```
#Visualising the correlation of statistically significant variables
grid.arrange(
  #Staff Satisfaction vs Customer Satisfaction
    ggplot(data2_new, aes(y=customer.satisfaction, x=staff.satisfaction)) + geom_point() + labs(x="Staf
    #Customer Satisfaction vs Delivery Time
    ggplot(data2_new, aes(y=customer.satisfaction, x=delivery.time)) + geom_point() + labs(x="Delivery
    ncol=2, top = textGrob("Shaded region shows the range of values within which the true regression li
)
```

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



- **Customer Satisfaction and Staff Satisfaction:** A strong positive correlation (r = 0.45, p < 0.05) was found between customer satisfaction and staff satisfaction, suggesting that customer satisfaction tends to increase when staff are satisfied.
- **Customer Satisfaction and Delivery Time:** A moderate negative correlation (r = -0.26, p < 0.05) was identified between customer satisfaction and delivery time, indicating a slight negative association between the two. As one variable increases, the other tends to decrease slightly.

Correlation alone cannot be used to make predictions, as it does not imply causation or provide detailed information about the nature of the relationship between variables. To explore these relationships further, we will perform regression analysis. Specifically, we will use Multiple Linear Regression to identify which variables—SES_category, staff satisfaction, new_range, and delivery time—affect customer satisfaction.

Multiple Linear Regression

```
#Fitting the model for Multiple Linear Regression to check the effects of SES_category, staff satisfact
data2_reg <- lm(customer.satisfaction ~ SES_category + staff.satisfaction + delivery.time + new_range, d
#Checking for Multicollinearity
kable(vif(data2_reg))
```

	GVIF	Df	GVIF^(1/(2*Df))
SES_category	1.319039	2	1.071678

	GVIF	Df	GVIF^(1/(2*Df))
staff.satisfaction	1.274924	1	1.129125
delivery.time	1.041958	1	1.020763
new_range	1.004263	1	1.002129

Since all the values are close to 1, it indicates little or no multicollinearity.

```
#Regression Results
summary(data2_reg)
```

```
##
## Call:
## lm(formula = customer.satisfaction ~ SES_category + staff.satisfaction +
##     delivery.time + new_range, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59866 -0.67952 -0.01176  0.65469  2.89231
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.218333   0.610361   8.550 6.87e-16 ***
## SES_categoryLow    -0.255765   0.138541  -1.846 0.065878 .
## SES_categoryMedium  1.209293   0.147773   8.183 8.43e-15 ***
## staff.satisfaction  0.351113   0.080457   4.364 1.77e-05 ***
## delivery.time      -0.017220   0.004948  -3.480 0.000577 ***
## new_rangeTRUE       0.093878   0.112249   0.836 0.403643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9687 on 294 degrees of freedom
## Multiple R-squared:  0.447,  Adjusted R-squared:  0.4376
## F-statistic: 47.53 on 5 and 294 DF,  p-value: < 2.2e-16
```

```
#Confidence Intervals for Null Hypothesis Significance Testing
kable(cbind(coefficient=coef(data2_reg), confint(data2_reg)))
```

	coefficient	2.5 %	97.5 %
(Intercept)	5.2183329	4.0171019	6.4195639
SES_categoryLow	-0.2557651	-0.5284229	0.0168927
SES_categoryMedium	1.2092932	0.9184667	1.5001196
staff.satisfaction	0.3511132	0.1927684	0.5094580
delivery.time	-0.0172198	-0.0269580	-0.0074816
new_rangeTRUE	0.0938782	-0.1270351	0.3147916

Model

$$\text{Customer Satisfaction} = \beta_{\text{Intercept}} + \beta_{\text{SES Category}} \times \text{SES Category} + \beta_{\text{Staff Satisfaction}} \times \text{Staff Satisfaction} + \beta_{\text{Delivery Time}} \times \text{Delivery}$$

Intercept: The model predicts a baseline customer satisfaction value of 5.2183 when all predictor variables are at their reference levels. The 95% confidence interval for this estimate is [4.017, 6.420].

Predictor Variables

- **SES_categoryLow:** (95% CI: [-0.528423, 0.016893]) Customers in the Low SES category have an average satisfaction score that is approximately 0.256 points lower than those in the High SES category, holding other variables constant. However, since the confidence interval includes zero, this effect is not statistically significant at the 95% confidence level. This means we do not have strong evidence that being in the Low SES category significantly decreases satisfaction compared to the High SES category.
- **SES_categoryMedium:** Customers in the Medium SES category have an average satisfaction score that is about 1.21 points higher than those in the High SES category, holding other variables constant ($\beta = 1.21, t(294) = 8.18, p < 0.001$). (95% CI: [0.92, 1.5]).
- **Staff Satisfaction:** Each one-unit increase in staff.satisfaction is associated with an increase of about 0.351 points in customer.satisfaction, holding other factors constant ($\beta = 0.351, t(294) = 4.36, p < 0.001$). (95% CI: [0.19, 0.51]).
- **Delivery Time:** Each additional unit of delivery.time is associated with a decrease of about 0.017 points in customer.satisfaction, holding other factors constant ($\beta = 0.017, t(294) = -3.48, p < 0.001$). (95% CI: [-0.03, -0.01]).
- **new_rangeTRUE:** (95% CI: [-0.13, 0.31]) When new_range is available, customer satisfaction increases by approximately 0.094 points compared to when new_range is not available, holding other factors constant. However, since the confidence interval includes zero, this effect is not statistically significant.

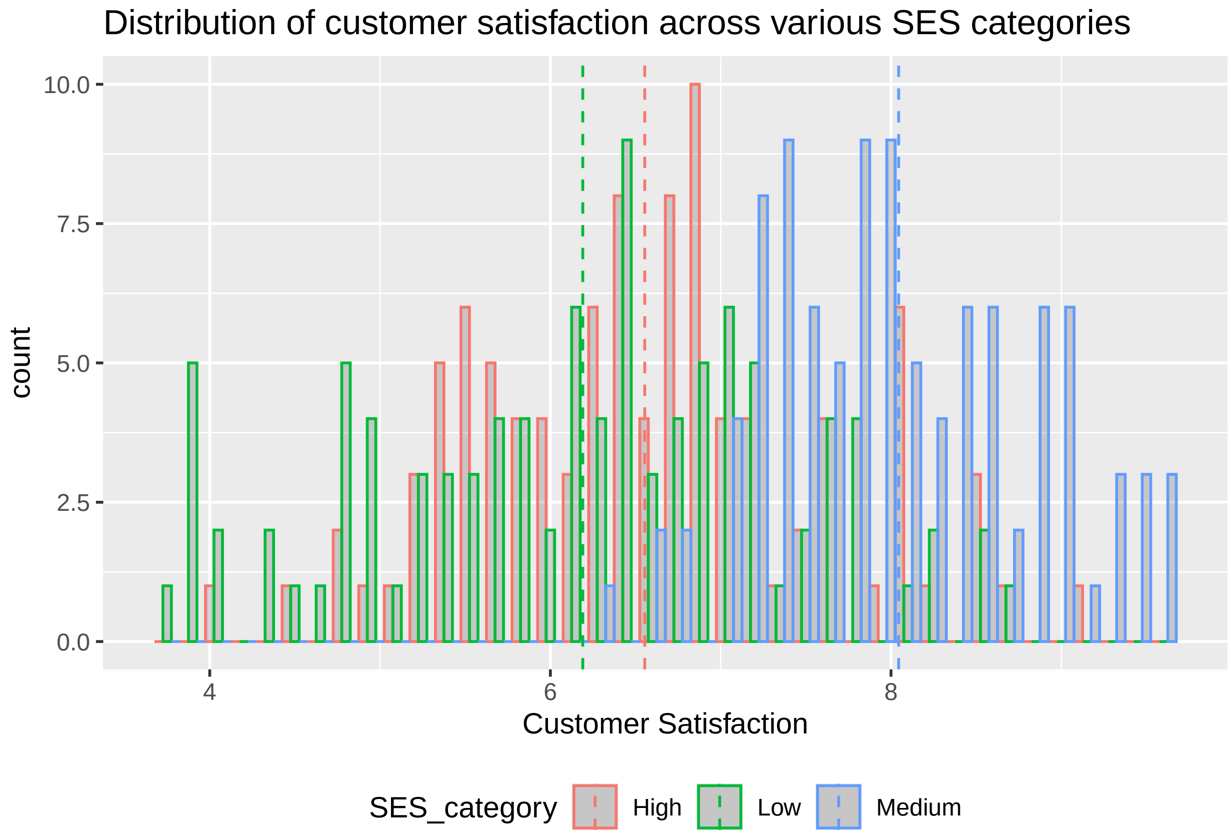
Model Fit

- **Residual Standard Error:** The model's predictions deviate from the actual values by an average of 0.9687 units.
- **Multiple R-squared:** Approximately 44.7% of the variability in Customer Satisfaction can be explained by the combined effects of the predictor variables.
- **Adjusted R-squared:** After adjusting for the number of predictors, the model explains approximately 43.76% of the variability in Customer Satisfaction.
- **F-statistic:** The overall model is statistically significant, indicating that at least one of the predictor variables is significantly associated with Customer Satisfaction ($F(5, 294) = 47.53, p < 0.001$).

Effect of Delivery time upon Customer Satisfaction across Socio-Economic-Status

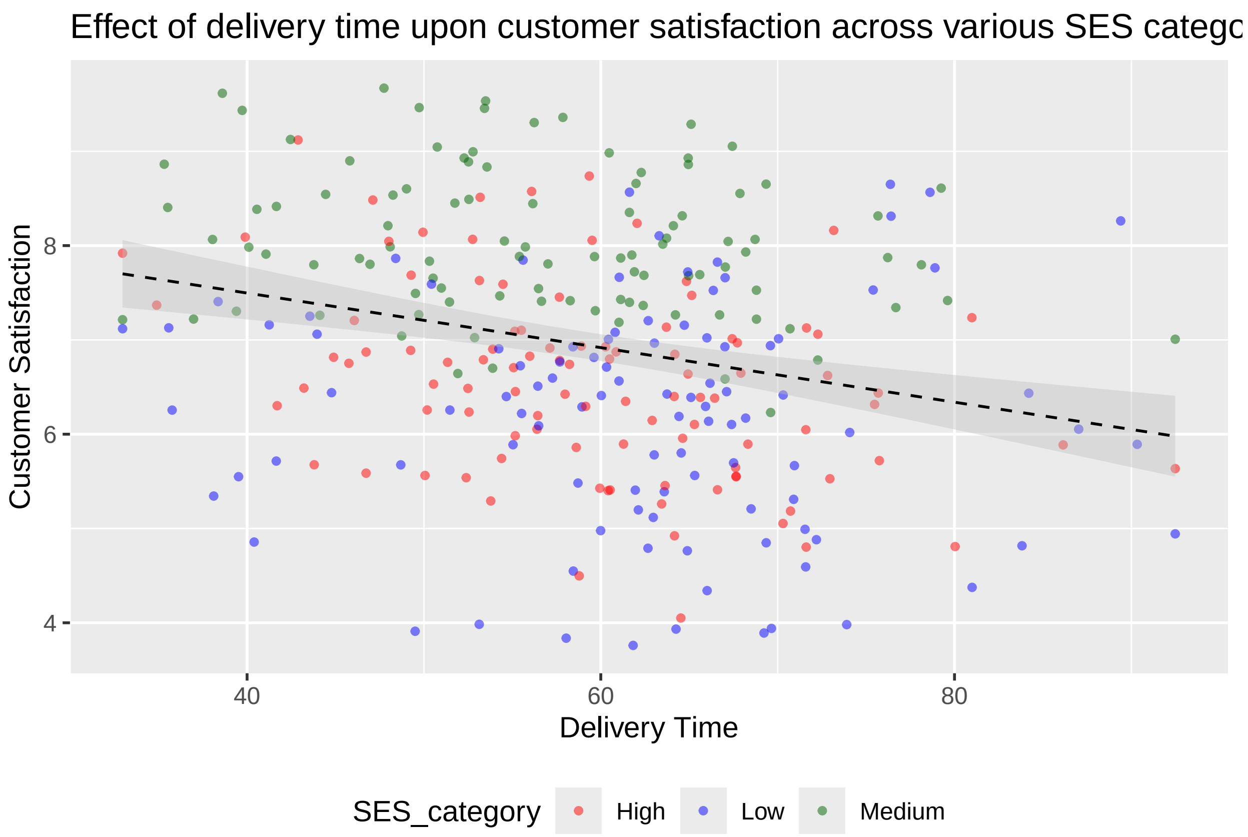
```r
#Storing the average customer satisfaction across various SES categories
avg.customer.satisfaction <- data2 %>%
  group_by(SES_category) %>%
  summarise(
    avg_satisfaction = mean(customer.satisfaction, na.rm = TRUE)
  )
```

```r
#Visualising the distribution of customer satisfaction across various SES categories
ggplot(data2, aes(customer.satisfaction, color = SES_category)) +
  geom_histogram(binwidth = 0.15, alpha = 0.25, position = "dodge") +
  geom_vline(data = avg.customer.satisfaction, aes(xintercept = avg_satisfaction, color=SES_category), 
  labs(x = "Customer Satisfaction", title = "Distribution of customer satisfaction across various SES ca
  theme(legend.position = "bottom")
```



```r
#Visualising the effect of delivery time upon customer satisfaction across various SES categories
ggplot(data2, aes(x= delivery.time, y = customer.satisfaction, color = SES_category)) +
  geom_point(alpha = 0.5, size = 1) +
  geom_smooth(method=lm, color = "black", fill = "grey", linetype = 2, size = 0.5) +
  scale_color_manual(values= c("red", "blue", "darkgreen")) +
  labs(x = "Delivery Time", y = "Customer Satisfaction", title = "Effect of delivery time upon customer
  theme(legend.position = "bottom")
```

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



```
#Individual terms
data2_reg_2 <- lm(customer.satisfaction ~ delivery.time + SES_category, data = data2)
#Assuming Interaction
data2_reg_2_int <- lm(customer.satisfaction ~ delivery.time * SES_category, data = data2)
#Checking which model is a better fit
anova(data2_reg_2, data2_reg_2_int)
```

```
## Analysis of Variance Table
##
## Model 1: customer.satisfaction ~ delivery.time + SES_category
## Model 2: customer.satisfaction ~ delivery.time * SES_category
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    296 294.65
## 2    294 289.06  2    5.5932 2.8444 0.05977 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test indicates that the model with interaction terms provides a better fit compared to the model without them.

```
#Checking regression results
summary(data2_reg_2_int)
```

```
##
## Call:
```

```
## lm(formula = customer.satisfaction ~ delivery.time * SES_category,
##     data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.43290 -0.63000  0.00057  0.72673  2.52903
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       8.62092    0.57199  15.072  < 2e-16 ***
## delivery.time                    -0.03471    0.00946  -3.669 0.000289 ***
## SES_categoryLow                  -2.12232    0.77457  -2.740 0.006519 **
## SES_categoryMedium                0.29635    0.76383   0.388 0.698310
## delivery.time:SES_categoryLow     0.02976    0.01253   2.374 0.018221 *
## delivery.time:SES_categoryMedium  0.01937    0.01287   1.505 0.133336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9916 on 294 degrees of freedom
## Multiple R-squared:  0.4205, Adjusted R-squared:  0.4107
## F-statistic: 42.67 on 5 and 294 DF,  p-value: < 2.2e-16
```

```
#CI
kable(cbind(coefficient=coef(data2_reg_2_int), confint(data2_reg_2_int)))
```

	coefficient	2.5 %	97.5 %
(Intercept)	8.6209247	7.4952166	9.7466327
delivery.time	-0.0347063	-0.0533237	-0.0160888
SES_categoryLow	-2.1223232	-3.6467252	-0.5979212
SES_categoryMedium	0.2963538	-1.2069192	1.7996269
delivery.time:SES_categoryLow	0.0297618	0.0050929	0.0544308
delivery.time:SES_categoryMedium	0.0193725	-0.0059566	0.0447016

Model

$$\text{Customer Satisfaction} = \beta_{\text{Intercept}} + \beta_{\text{Delivery Time}} \times \text{Delivery Time} + \beta_{\text{SES Category Low}} \times \text{SES Category Low} + \beta_{\text{SES Category Mediun}}$$

Intercept: The model predicts a baseline customer satisfaction value of 8.6209 when delivery time is zero and SES is not low or medium. ($\beta = 8.6209, t(294) = 15.072, p < 0.001$) (95% CI = [7.495,9.747]).

Predictor Variables

- **Delivery Time:** Each additional unit of delivery time is associated with a decrease of about 0.035 points in customer satisfaction, holding other factors constant ($\beta = -0.03471, t(294) = -3.669, p < 0.001$) (95% CI = [-0.053,-0.016]).
- **SES_categoryLow:** Customers in the Low SES category have an average satisfaction score that is approximately 2.122 points lower than those in the High SES category, holding other variables constant ($\beta = -2.12232, t(294) = -2.740, p < 0.01$) (95% CI = [-3.647,-0.598]).

- **SES_categoryMedium:** Customers in the Medium SES category have an average satisfaction score that is about 0.296 points higher than those in the High SES category, holding other variables constant. However, this effect is not statistically significant ($p = 0.698$).

- **Delivery Time: SES_categoryLow:** The interaction term indicates that for low SES, the negative effect of delivery time on satisfaction is mitigated by 0.030 points ($\beta = 0.02976, t(294) = 2.374, p < 0.05$) (95% CI = [0.005,0.054]).

- **Delivery Time: SES_categoryMedium:** The interaction term suggests a slight mitigation of the delivery time effect for medium SES, but it is not statistically significant ($p = 0.133$).

Model Fit

- **Residual Standard Error:** The model's predictions deviate from the actual values by an average of 0.9916 units.

- **Multiple R-squared:** Approximately 42.05% of the variability in Customer Satisfaction can be explained by the combined effects of the predictor variables.

- **Adjusted R-squared:** After adjusting for the number of predictors, the model explains approximately 41.07% of the variability in Customer Satisfaction.

- **F-statistic:** The overall model is statistically significant, indicating that at least one of the predictor variables is significantly associated with Customer Satisfaction ($F(5, 294) = 42.67, p < 0.001$).

```r
# Calculating estimated marginal means for the interaction
emm_interaction <- emmeans(data2_reg_2_int, ~ delivery.time | SES_category)
summary(emm_interaction)
```

```
## SES_category = High:
##  delivery.time emmean     SE  df lower.CL upper.CL
##           59.6   6.55 0.0992 294     6.36     6.75
## 
## SES_category = Low:
##  delivery.time emmean     SE  df lower.CL upper.CL
##           59.6   6.20 0.1020 294     6.00     6.40
## 
## SES_category = Medium:
##  delivery.time emmean     SE  df lower.CL upper.CL
##           59.6   8.00 0.1020 294     7.80     8.20
## 
## Confidence level used: 0.95
```

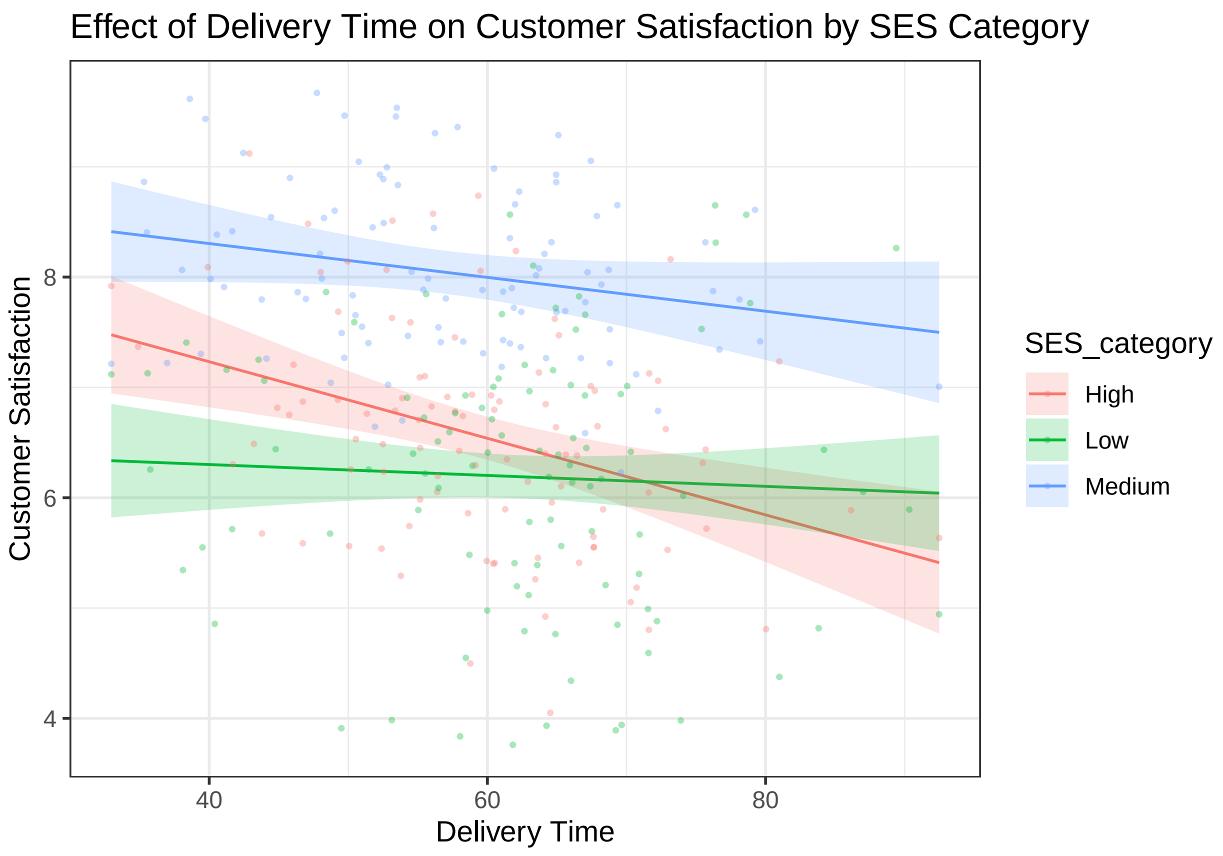
- **High SES Category:** The average customer satisfaction for the high SES group is 6.55. (SE = 0.0992, 95% CI = [6.36, 6.75])

- **Low SES Category:** The average customer satisfaction for the low SES group is 6.20. (SE = 0.102, 95% CI = [6, 6.4])

- **Medium SES Category:** The average customer satisfaction for the medium SES group is 8.00. (SE = 0.102, 95% CI = [7.8, 8.2])

Conclusion: Medium SES category shows the highest customer satisfaction, followed by high and then low SES categories.

```
# Generating predictions and including confidence intervals
predicted <- predict(data2_reg_2_int, newdata = data2, interval = "confidence")

# Adding predicted values and intervals to a data frame
predicted.customer.satisfaction <- data2 %>%
  mutate(
    predict.customer.satisfaction = predicted[, "fit"],
    lwr = predicted[, "lwr"],
    upr = predicted[, "upr"]
  )

# Creating the plot
ggplot() +
  geom_line(data = predicted.customer.satisfaction, aes(x = delivery.time, y = predict.customer.satisfac
  geom_ribbon(data = predicted.customer.satisfaction, aes(x = delivery.time, ymin = lwr, ymax = upr, fil
  geom_point(data = data2, aes(x = delivery.time, y = customer.satisfaction, color = SES_category), alph
  labs(y = "Customer Satisfaction", x = "Delivery Time", title = "Effect of Delivery Time on Customer Sa
  theme_bw()
```



Inference

In High SES category stores, customer satisfaction decreases with longer delivery times. In Low SES category stores, customer satisfaction remains relatively stable regardless of delivery time. In Medium SES category stores, customer satisfaction decreases slightly as delivery times increase. The interaction between delivery time and SES category is marginally significant, suggesting some combined effect on satisfaction. Delivery strategies might need to be adjusted based on SES to maintain customer satisfaction.