Data Analysis Section

Qiheng(Michael) Yan

2023-11-21

```
library(tidyverse)
library(stats)
library(ggplot2)
library(expss)
set.seed(1028)
```

Epi 3 Data Analysis

This section ignores the inclusion/exclusion criteria and uses all 7122 subjects (using complete_data)

Import data

```
# Import the dataset
complete_data<-readRDS("combined_data.rds")
subset_1168_data<-readRDS("subset_1168.rds")
#write.csv(complete_data, "complete_data.csv", row.names=FALSE)
#write_labelled_csv(complete_data, "complete_data_labeled.csv", row.names=FALSE)
#write.csv(subset_1168_data, "subset_1168_data.csv", row.names=FALSE)
#write_labelled_csv(subset_1168_data, "subset_1168_data_labeled.csv", row.names=FALSE)
```

Bivariate Analysis

In bivariate analysis, we'll use Chi-squared tests for categorical variables like health insurance status, race/ethnicity, gender, and a T-test for continuous variables like age.

```
# Chi-squared test for medication adherence and family income
chisq_income <- chisq.test(table(complete_data$adherence, complete_data$income_cat))

# Chi-squared test for medication adherence and total insurance status
chisq_insurance <- chisq.test(table(complete_data$adherence, complete_data$ins_classif))
```

```
## Warning in chisq.test(table(complete_data$adherence,
## complete_data$ins_classif)): Chi-squared approximation may be incorrect
```

```r
# Chi-squared test for medication adherence and race
chisq_race <- chisq.test(table(complete_data$adherence, complete_data$race_6cat))

# Chi-squared test for medication adherence and sex
chisq_gender <- chisq.test(table(complete_data$adherence, complete_data$sex))

# T-test for medication adherence and age
t_test_age <- t.test(complete_data$age ~ complete_data$adherence)

# Print the results
print(chisq_income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(complete_data$adherence, complete_data$income_cat)
## X-squared = 8.3414, df = 2, p-value = 0.01544
```

```r
print(chisq_insurance)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(complete_data$adherence, complete_data$ins_classif)
## X-squared = 54.283, df = 8, p-value = 6.085e-09
```

```r
print(chisq_race)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(complete_data$adherence, complete_data$race_6cat)
## X-squared = 15.478, df = 5, p-value = 0.008503
```

```r
print(chisq_gender)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(complete_data$adherence, complete_data$sex)
## X-squared = 2.4827, df = 1, p-value = 0.1151
```

```r
print(t_test_age)
```

```
##
##  Welch Two Sample t-test
##
## data:  complete_data$age by complete_data$adherence
## t = -11.074, df = 901.08, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
```

```
## 95 percent confidence interval:
##  -9.332160 -6.522391
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            56.91137            64.83864
```

The results from the bivariate analyses provide valuable insights into the relationships between adherence to prescribed cholesterol medication and various factors such as family income, health insurance status, race/ethnicity, gender, and age.

1. Family Income vs. Adherence: Chi-squared test result:

   $$X^2 = 8.3414, df = 2, p - value = 0.01544$$

   Interpretation: There is a statistically significant association between family income category and adherence to cholesterol medication. Since the p-value is less than 0.05, we can conclude that the differences in adherence rates across different income categories are not due to random chance.

2. Health Insurance Status vs. Adherence: Chi-squared test result:

   $$X^2 = 54.283, df = 8, p - value = 6.085 \times 10^{-9}$$

   Interpretation: There is a highly statistically significant association between health insurance classification and medication adherence. The extremely low p-value indicates strong evidence against the null hypothesis of no association.

3. Race/Ethnicity vs. Adherence: Chi-squared test result:

   $$X^2 = 15.478, df = 5, p - value = 0.008503$$

   Interpretation: Race/ethnicity shows a statistically significant association with medication adherence. The p-value below 0.05 suggests that different racial/ethnic groups have different adherence rates to cholesterol medication.

4. Gender vs. Adherence: Chi-squared test result:

   $$X^2 = 2.4827, df = 1, p - value = 0.1151$$

   Interpretation: There is no statistically significant association between gender and medication adherence. The p-value is greater than 0.05, indicating that any observed differences in adherence between genders could be due to chance.

5. Age vs. Adherence: T-test result:

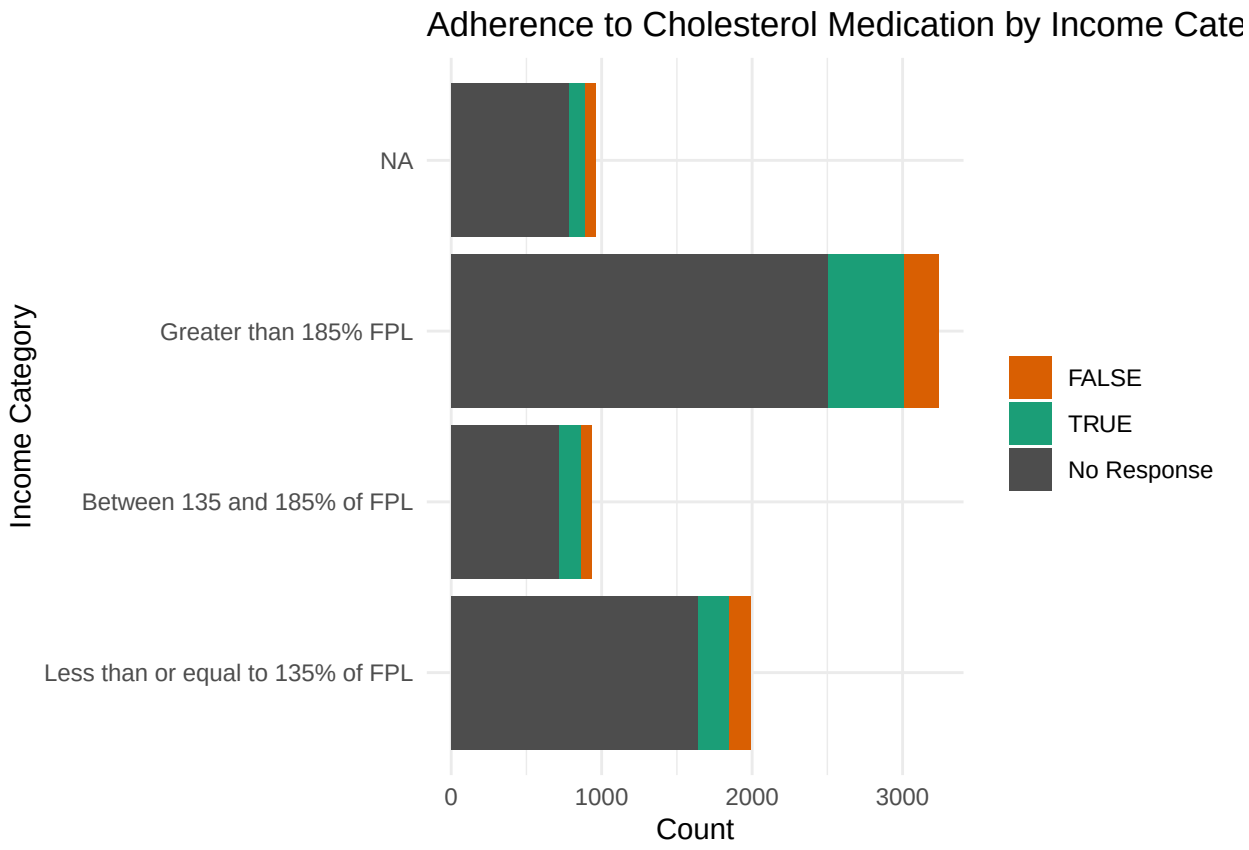
   $$t = -11.074, df = 901.08, p - value = 2.2 \times 10^{-16}$$

   Interpretation: There is a highly statistically significant difference in the average age between those who adhere to their medication and those who do not. The negative t-value indicates that the mean age of the group adhering to the medication (mean = 64.84 years) is higher than that of the non-adhering group (mean = 56.91 years). The extremely low p-value provides strong evidence against the null hypothesis of no difference in means.

These results suggest that socioeconomic factors (like income and insurance status), as well as demographic characteristics (like race/ethnicity and age), are associated with adherence to cholesterol medication. Gender, however, does not seem to show a significant association in this context. These findings can inform further multivariate analysis to understand the independent effect of each factor while controlling for others.

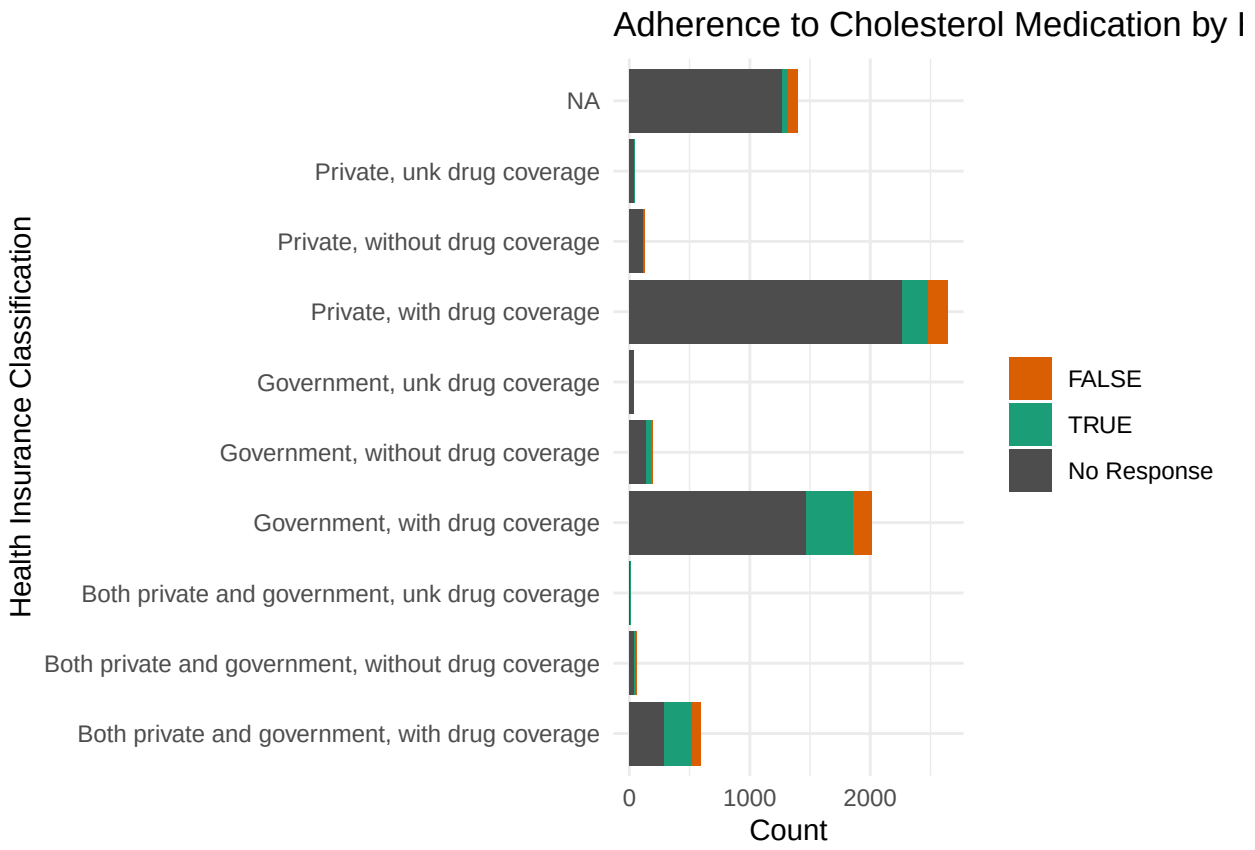
Bivariate Analysis Visualization

```r
# Ensure NA is a factor level and label it as "No Response"
complete_data$adherence_factor <- factor(complete_data$adherence, levels = c(FALSE, TRUE))
complete_data$adherence_factor <- addNA(complete_data$adherence_factor)
levels(complete_data$adherence_factor)[is.na(levels(complete_data$adherence_factor))] <- "No Response"

# Define new colors for the bars, including NA
new_colors <- c("TRUE" = "#1b9e77", "FALSE" = "#d95f02", "No Response" = "#4D4D4D")

# Create the plots, making sure to use scale_fill_manual to include NA values
# and set the axis titles correctly after coord_flip()

# Income Category vs Adherence
ggplot(complete_data, aes(x = income_cat, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Income Category",
       y = "Count",
       x = "Income Category") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```

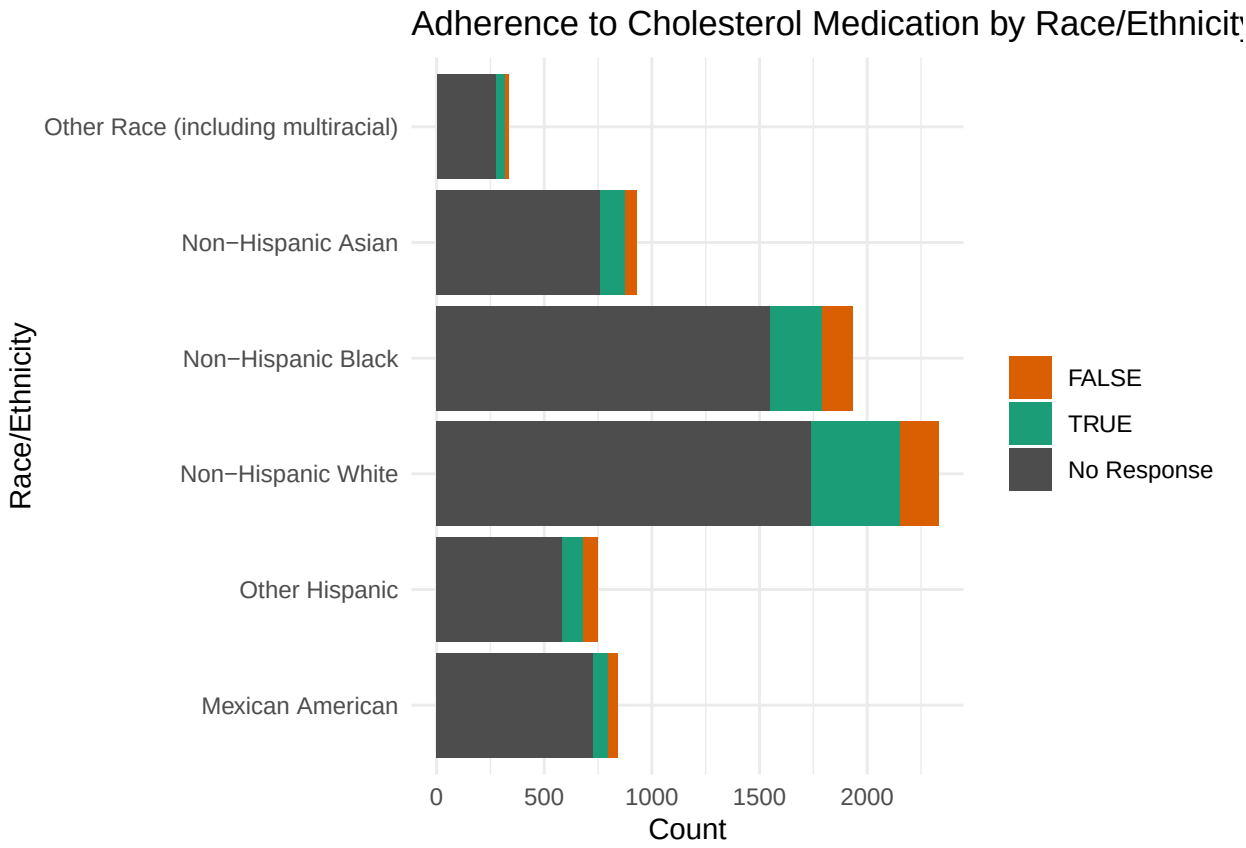



```r
# Health Insurance Classification vs Adherence
ggplot(complete_data, aes(x = ins_classif, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
```

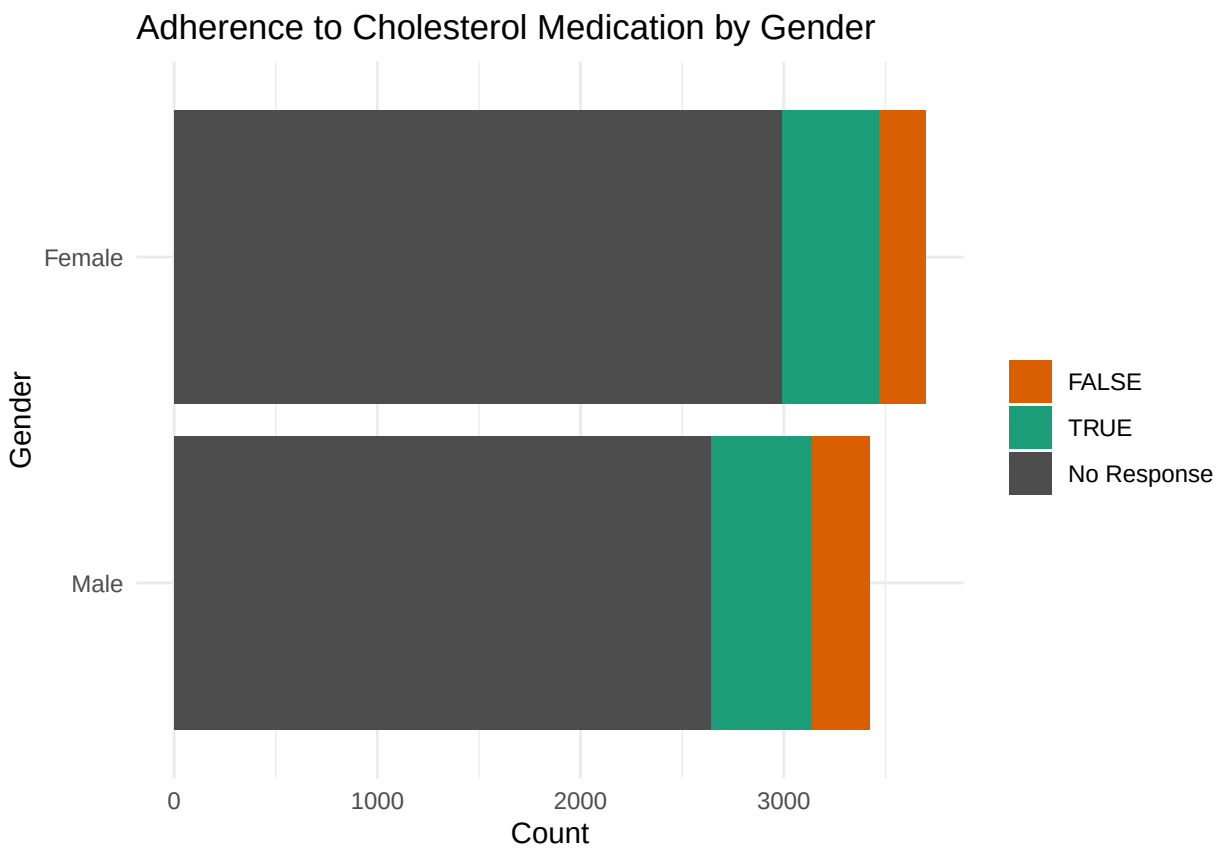
```
  labs(title = "Adherence to Cholesterol Medication by Health Insurance Status",
       y = "Count",
       x = "Health Insurance Classification") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```



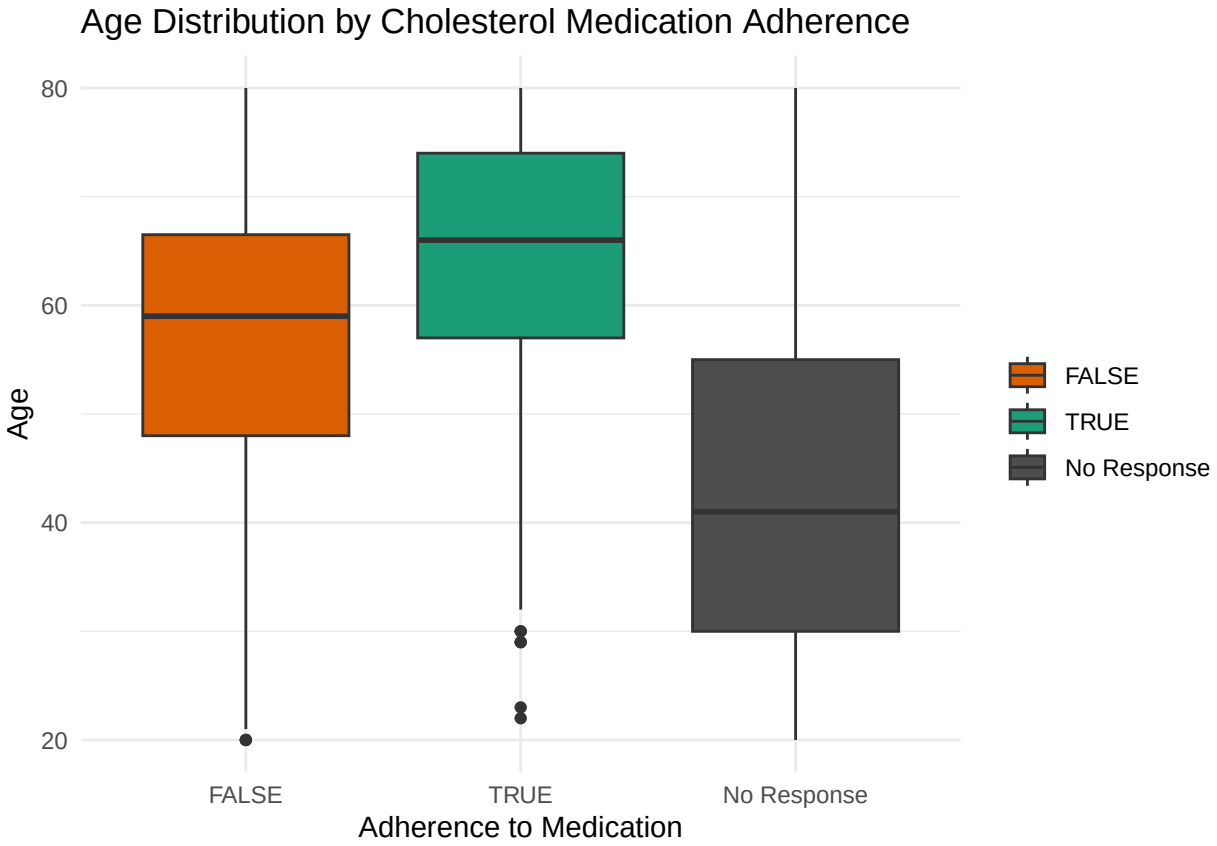

```
# Race/Ethnicity vs Adherence
ggplot(complete_data, aes(x = race_6cat, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Race/Ethnicity",
       y = "Count",
       x = "Race/Ethnicity") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```




```r
# Gender vs Adherence
ggplot(complete_data, aes(x = sex, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Gender",
       y = "Count",
       x = "Gender") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```




```
# Age vs Adherence Box Plot (does not use coord_flip())
ggplot(complete_data, aes(x = adherence_factor, y = age, fill = adherence_factor)) +
  geom_boxplot() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Age Distribution by Cholesterol Medication Adherence",
       x = "Adherence to Medication",
       y = "Age") +
  theme_minimal() +
  theme(legend.title = element_blank())
```



The resulting plots reveal several patterns related to adherence to cholesterol medication among adults:

1. Adherence by Income Category: The first bar chart suggests that individuals with higher income (greater than 185% FPL - Federal Poverty Level) show the highest adherence to cholesterol medication. This may indicate that financial stability plays a crucial role in the ability to maintain prescribed medication regimens. The "No Response" category could reflect missing data or respondents who did not answer the adherence question, highlighting the need to address potential barriers in data collection or survey response.

2. Adherence by Health Insurance Status: The second bar chart shows that individuals with both private and government insurance, especially with drug coverage, have higher adherence levels. This reinforces the importance of comprehensive health insurance in supporting medication adherence. The presence of "No Response" in this category similarly underscores the potential for missing data or non-responses that could influence the study's findings.

3. Adherence by Race/Ethnicity: The third bar chart indicates variability in adherence among different racial and ethnic groups. Notably, the "Non-Hispanic White" group exhibits a higher adherence compared to other groups. Such disparities may point to underlying social, economic, or cultural factors that affect health behaviors.

4. Adherence by Gender: The fourth bar chart illustrates that female respondents exhibit slightly higher adherence to cholesterol medication than male respondents, which could suggest gender-specific factors influencing health behavior, though the difference is not huge.

5. Age Distribution by Adherence: The box plot shows the age distribution for each adherence group. Individuals who are adherent to their cholesterol medication appear to be older on average than those who are non-adherent. This could be due to older adults having more established routines, a greater prevalence of chronic conditions necessitating adherence, or a higher likelihood of experiencing the

consequences of non-adherence. The "No Response" group does not provide a clear age distribution due to the nature of missing data.

These visualizations highlight critical associations between socioeconomic factors, insurance coverage, demographic characteristics, and medication adherence. They serve as an essential complement to the statistical analyses, providing a clear and interpretable depiction of the data that can guide targeted interventions and policy-making to improve adherence rates and reduce health disparities.

Multivariate Analysis

Next, we'll conduct a logistic regression analysis to assess the relationship between family income and medication adherence, controlling for confounders.

```
# Logistic regression model
lr_model <- glm(adherence ~ income_cat + ins_classif + race_6cat + sex + age + educ_level,
             family = binomial(link = "logit"),
             data = complete_data)

# Summary of the model
summary(lr_model)
```

```
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     sex + age + educ_level, family = binomial(link = "logit"),
##     data = complete_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.753143  84.179697  -0.021   0.9834
## income_cat.L    0.276106   0.124303   2.221   0.0263 *
## income_cat.Q   -0.129567   0.155199  -0.835   0.4038
## ins_classif.L  -6.388729 169.500182  -0.038   0.9699
## ins_classif.Q   2.233497 169.501504   0.013   0.9895
## ins_classif.C   9.521359 284.658013   0.033   0.9733
## ins_classif^4  -6.866729 175.728782  -0.039   0.9688
## ins_classif^5  -9.356423 214.728917  -0.044   0.9652
## ins_classif^6   5.847728 306.244257   0.019   0.9848
## ins_classif^7   1.139671 362.100094   0.003   0.9975
## ins_classif^8  10.973410 264.223598   0.042   0.9669
## race_6cat.L     0.303860   0.279379   1.088   0.2768
## race_6cat.Q     0.443528   0.248841   1.782   0.0747 .
## race_6cat.C     0.209811   0.222654   0.942   0.3460
## race_6cat^4     0.011743   0.192939   0.061   0.9515
## race_6cat^5    -0.059934   0.140616  -0.426   0.6699
## sex.L           0.006823   0.094580   0.072   0.9425
## age             0.035953   0.006249   5.754 8.73e-09 ***
## educ_level.L   -0.087949   0.220589  -0.399   0.6901
## educ_level.Q    0.076210   0.192952   0.395   0.6929
## educ_level.C   -0.052161   0.170468  -0.306   0.7596
## educ_level^4    0.185227   0.148058   1.251   0.2109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.8  on 1203  degrees of freedom
## Residual deviance: 1399.1  on 1182  degrees of freedom
##   (5918 observations deleted due to missingness)
## AIC: 1443.1
##
## Number of Fisher Scoring iterations: 13
```

The output from the logistic regression model provides several pieces of information that can be interpreted to understand the factors associated with adherence to cholesterol medication:

1. Income Category (income_cat.L): The linear term for income category is significant ($p = 0.0263$), suggesting that as income increases, the log-odds of being adherent to cholesterol medication also increase. The positive coefficient (log-odds) (Estimate = 0.276106) indicates a positive relationship between higher income levels and adherence.

2. Health Insurance Classification (ins_classif): None of the terms for health insurance classification are statistically significant, as all p-values are well above the conventional alpha level of 0.05. This suggests that within this model, health insurance classification is not a significant predictor of medication adherence when controlling for other factors.

3. Race/Ethnicity (race_6cat): The terms for race/ethnicity are not statistically significant, with p-values greater than 0.05. However, the quadratic term (race_6cat.Q) approaches significance ($p = 0.0747$), indicating there might be a complex relationship between race/ethnicity and medication adherence that warrants further investigation.

4. Gender (sex.L): Gender is not a significant predictor of medication adherence in this model ($p = 0.9425$), indicating that the difference between males and females is not statistically significant when other factors are controlled for.

5. Age: Age is a highly significant predictor ($p < 0.001$), with a positive coefficient (Estimate = 0.035953). This indicates that for each additional year of age, the log-odds of being adherent to cholesterol medication increase, suggesting that older individuals are more likely to adhere to their medication.

6. Education Level (educ_level): The education level terms are not significant predictors of medication adherence, with all p-values above 0.05, indicating no clear association between education level and adherence within this model.

7. Model Fit: The difference between the null deviance and the residual deviance indicates that the model with predictors fits the data better than a model without any predictors. However, given the relatively small decrease, there may still be room for model improvement.

8. AIC: The Akaike Information Criterion (AIC) for the model is 1443.1. This metric helps compare different models, with lower values indicating a better fit to the data.

9. Missingness: A large number of observations were deleted due to missingness (5918 observations), which could significantly affect the results. It's important to investigate the missing data pattern to ensure that it is not biasing the results.

Overall, income and age seem to be significant factors associated with adherence to cholesterol medication in this multivariate context. The significance of income suggests a possible socioeconomic gradient in medication adherence. The relationship with age could reflect better health habits or more regular healthcare usage among older individuals. It is crucial to consider the context of these results and the potential impact of missing data on the study's findings. Further analysis might involve exploring interactions between variables,

considering non-linear relationships, and addressing the issue of missing data, possibly through imputation methods or sensitivity analyses.

```r
# Calculate Odds Ratios and 95% Confidence Intervals
or <- exp(coef(lr_model))
# Wald Confidence Intervals and p-values
se <- sqrt(diag(vcov(lr_model)))
wald_ci_lower <- exp(coef(lr_model) - 1.96 * se)
wald_ci_upper <- exp(coef(lr_model) + 1.96 * se)
p_values <- summary(lr_model)$coefficients[, "Pr(>|z|)"]

# Create a data frame to nicely format the results
results <- data.frame(
  OR = exp(coef(lr_model)),
  LowerCI = wald_ci_lower,
  UpperCI = wald_ci_upper,
  PValue = p_values
)

# View the results
print(results)
```

```
##                          OR       LowerCI       UpperCI       PValue
## (Intercept)   1.732286e-01  3.831909e-73  7.831119e+70 9.833843e-01
## income_cat.L  1.317988e+00  1.033005e+00  1.681589e+00 2.633492e-02
## income_cat.Q  8.784761e-01  6.480697e-01  1.190798e+00 4.038075e-01
## ins_classif.L 1.680391e-03 8.789074e-148 3.212755e+141 9.699336e-01
## ins_classif.Q 9.332446e+00 4.868585e-144 1.788909e+145 9.894867e-01
## ins_classif.C 1.364815e+04 6.749655e-239 2.759725e+246 9.733170e-01
## ins_classif^4 1.041879e-03 2.719308e-153 3.991868e+146 9.688300e-01
## ins_classif^5 8.640864e-05 1.430912e-187 5.217970e+178 9.652446e-01
## ins_classif^6 3.464463e+02 7.232049e-259 1.659627e+263 9.847653e-01
## ins_classif^7 3.125740e+00 1.858362e-308           Inf 9.974887e-01
## ins_classif^8 5.830305e+04 7.145364e-221 4.757274e+229 9.668728e-01
## race_6cat.L   1.355080e+00  7.837054e-01  2.343024e+00 2.767592e-01
## race_6cat.Q   1.558195e+00  9.567619e-01  2.537696e+00 7.468830e-02
## race_6cat.C   1.233445e+00  7.972471e-01  1.908301e+00 3.460290e-01
## race_6cat^4   1.011812e+00  6.932131e-01  1.476838e+00 9.514694e-01
## race_6cat^5   9.418263e-01  7.149510e-01  1.240696e+00 6.699413e-01
## sex.L         1.006846e+00  8.364797e-01  1.211911e+00 9.424912e-01
## age           1.036607e+00  1.023989e+00  1.049381e+00 8.731580e-09
## educ_level.L  9.158074e-01  5.943407e-01  1.411149e+00 6.901126e-01
## educ_level.Q  1.079189e+00  7.393567e-01  1.575220e+00 6.928645e-01
## educ_level.C  9.491758e-01  6.795813e-01  1.325720e+00 7.596134e-01
## educ_level^4  1.203491e+00  9.003538e-01  1.608692e+00 2.109201e-01
```

```r
####### Uncomment this chunk to generate the multivariate results table #######
# # Load the required packages
# library(knitr)
# library(kableExtra)
#
# # Create a nice looking table to present the above ORs, CIs, and p-values
# nice_table <- kable(results,
```

```
#                     format = "latex",  # Use "latex" for PDF output or "pipe" for Markdown or "html" f
#                     digits = 3,       # Number of decimal places
#                     align = 'c',      # Center align the columns
#                     caption = "Multivariate Analysis of Factors Associated with Adherence to Cholestero
#   kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
#                 full_width = F,
#                 position = "center") %>%
#   column_spec(1, bold = T) %>%         # Make the OR column bold
#   column_spec(2:4, color = "blue") %>% # Color the CI columns blue
#   scroll_box(width = "100%", height = "2000px") # Add a scroll box if the table is too large
#
# # Print the table
# nice_table
#
# # To display this table outside of an R Markdown document, save it to an HTML file and open it in a w
# save_kable(nice_table, file = "NiceTable.html")
```