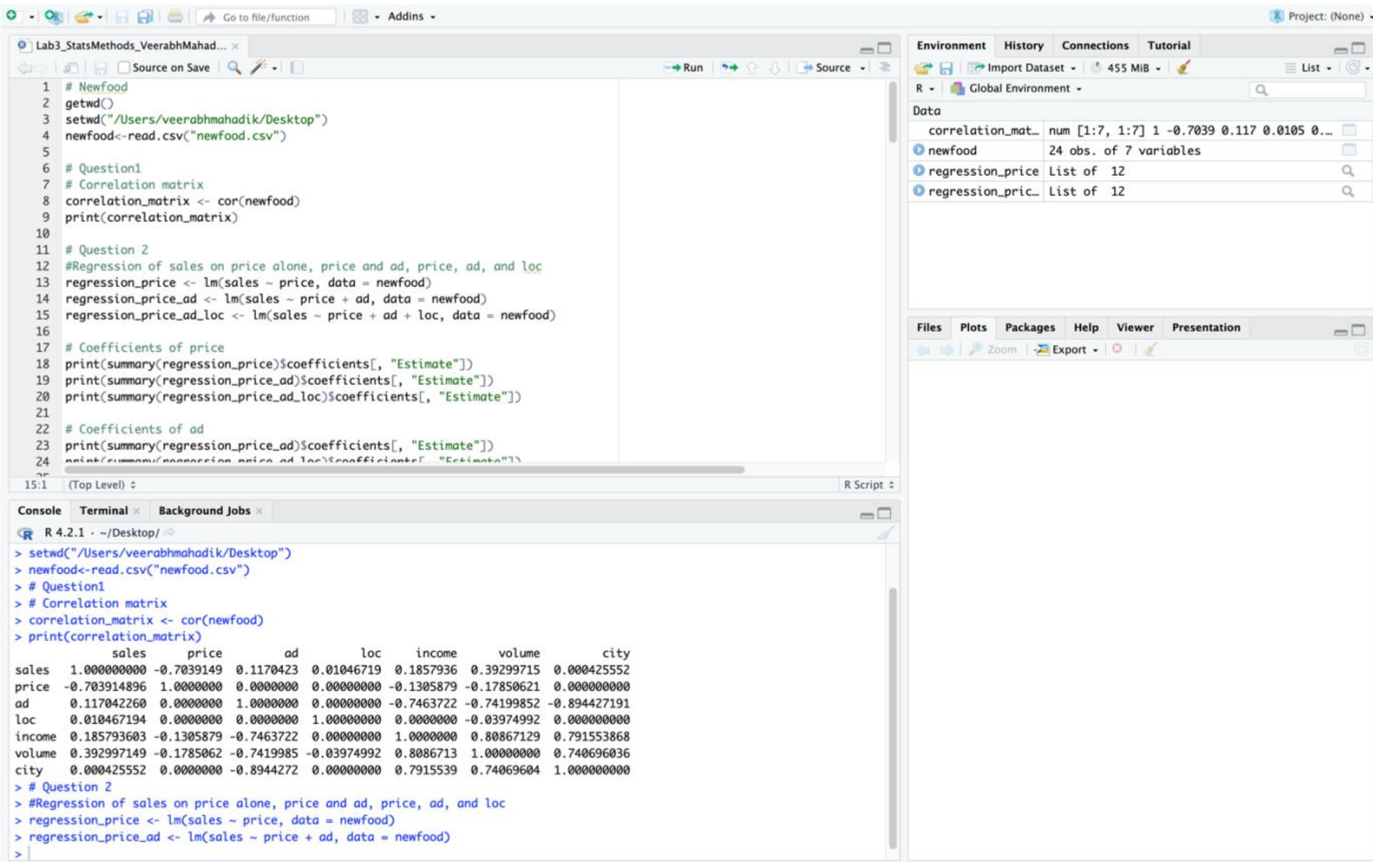
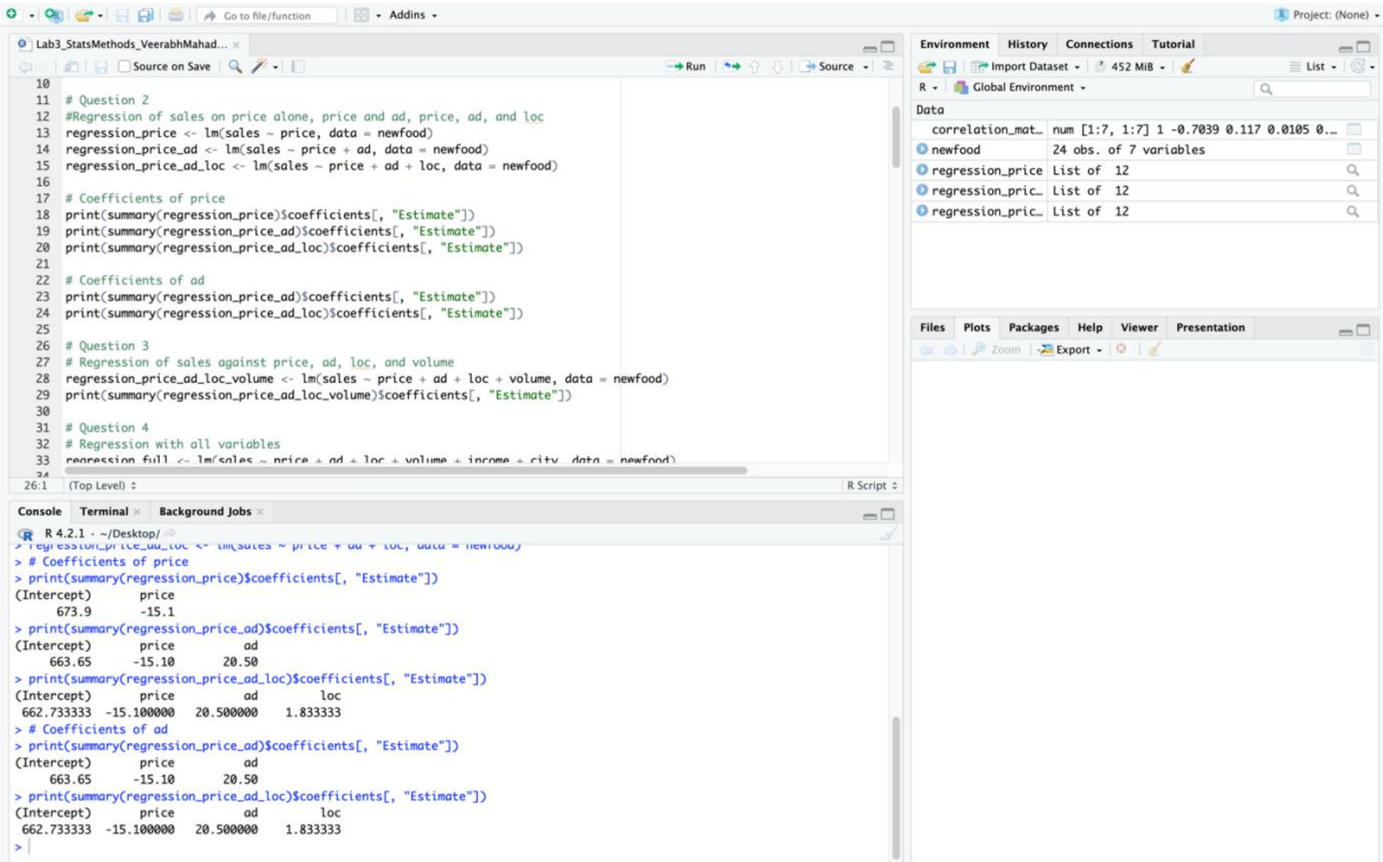
Newfood

Question 1



The correlation matrix will show correlations between variables. Zero correlations between location and advertising indicate no linear relationship between those variables.

Question 2



1. Regression of sales on price alone:

- The coefficient of price (-15.1) suggests that for every unit increase in price, sales decrease by 15.1 units, holding other variables constant.

2. Regression of sales on price and ad:

- The coefficient of price remains consistent (-15.1), indicating that its effect on sales remains the same when advertising expenditure is included in the model.

- The coefficient of ad (20.5) indicates that for every unit increase in advertising expenditure, sales increase by 20.5 units, holding other variables constant.

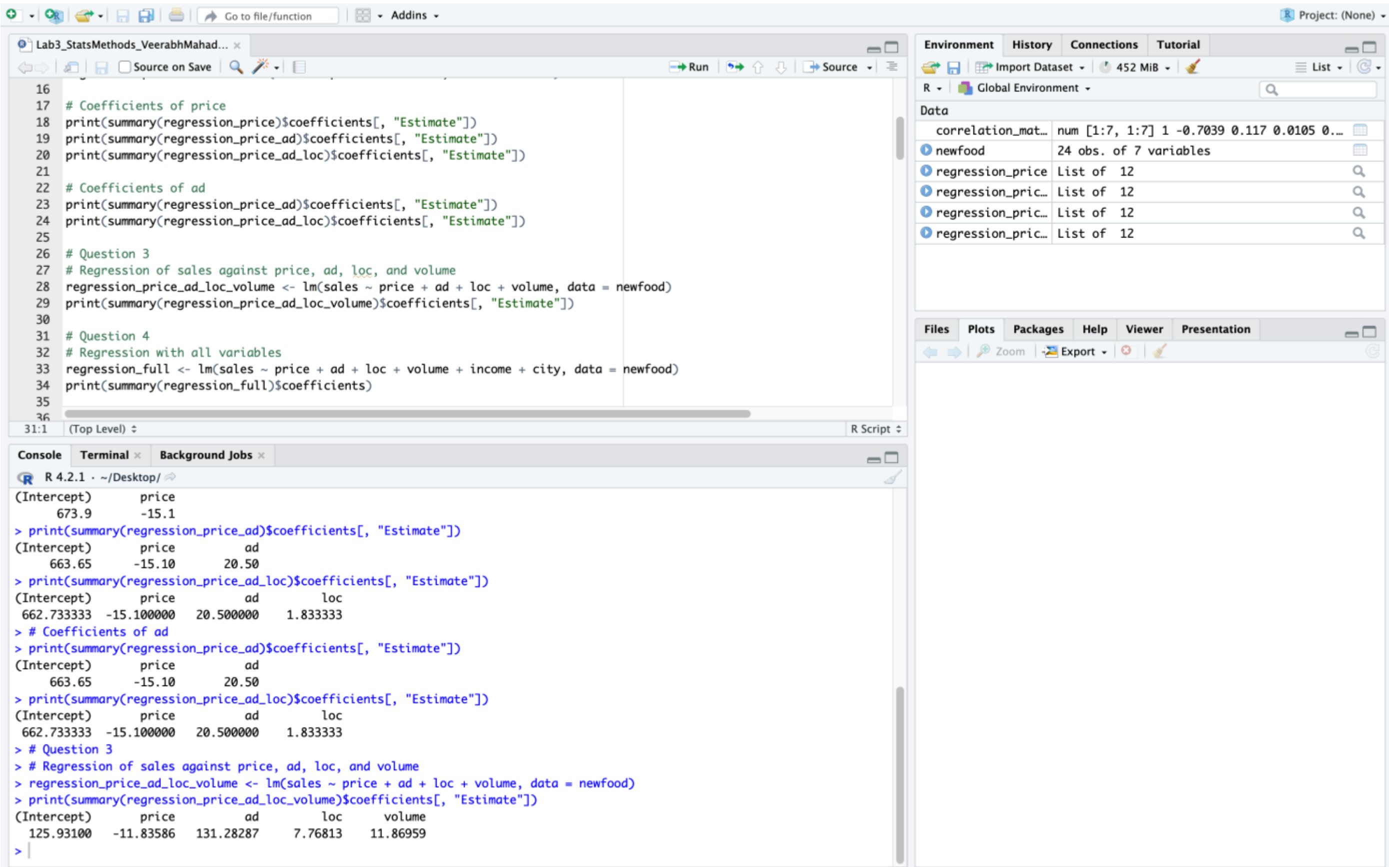
3. Regression of sales on price, ad, and loc:

- The coefficient of price remains consistent (-15.1), suggesting its effect on sales remains unchanged even with the inclusion of location and advertising.

- The coefficient of ad remains consistent (20.5), indicating its consistent effect on sales even when accounting for location.

- The coefficient of loc (1.833) indicates that placing the product in the instant breakfast section increases sales by 1.833 units compared to placing it in the bread section, holding other variables constant.

Question 3:



Regression Coefficients:

- For regression with price, ad, loc, and volume (regression_price_ad_loc_volume):

  - Intercept: 125.931

  - Price: -11.836

  - Ad: 131.283

  - Loc: 7.768

  - Volume: 11.870

- For regression with price, ad, and loc (regression_price_ad_loc):

  - Intercept: 662.733

- Price: -15.100

- Ad: 20.500

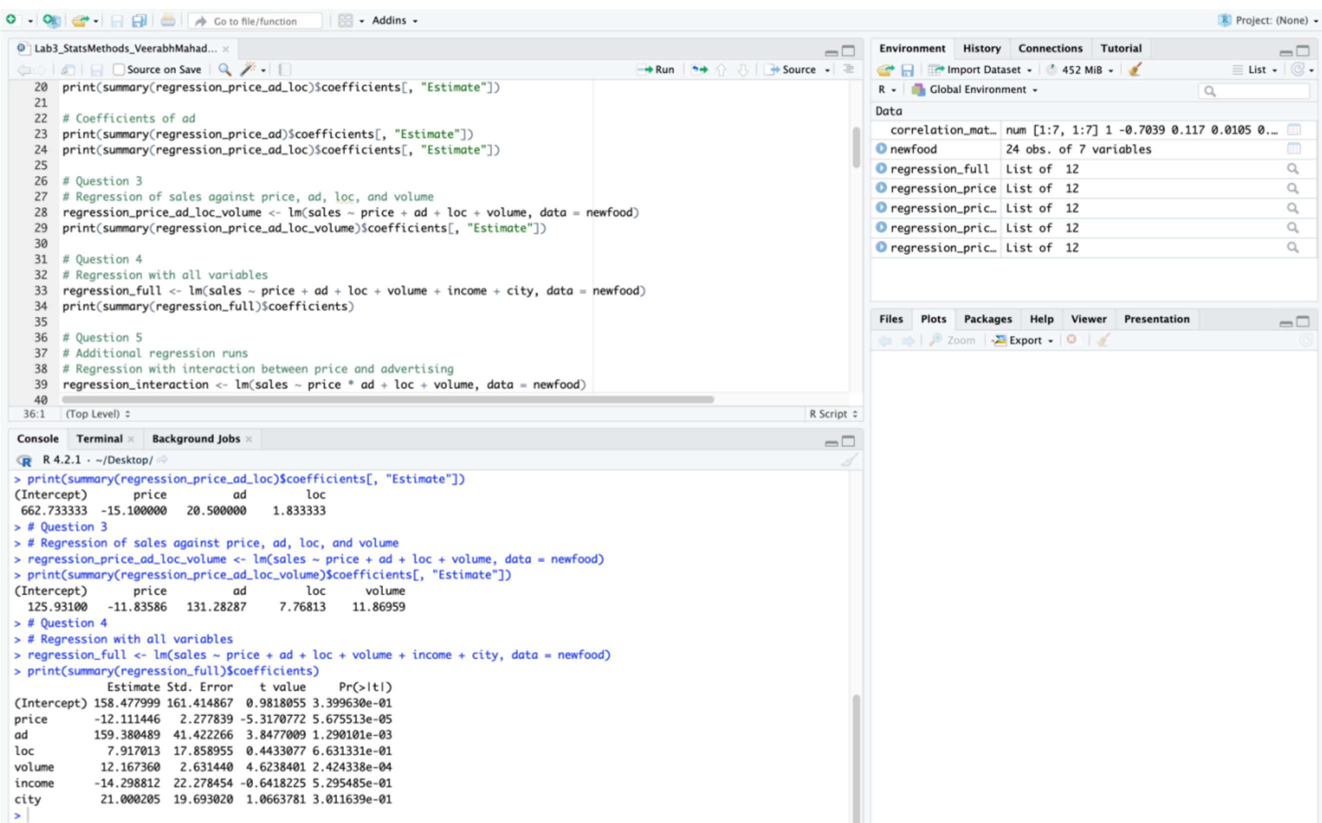
- Loc: 1.833

- The coefficients of price and ad change slightly while intercept changes from 662.733 to 125.931

- The coefficient of loc remains approximately the same in both regressions.

- In regression_price_ad_loc, there is no volume variable, so its introduction in regression_price_ad_loc_volume leads to a significant change in its coefficient. This happens because the volume variable captures additional variation in sales that was not explained by the previous variables (price, ad, and loc).

Question 4



- The coefficients for price, ad, loc, and volume are relatively consistent between the two models.

There are slight variations in the estimates and standard errors, but the overall trends remain similar.

- The regression_full model includes additional variables for income and city, which were not present in the regression_price_ad_loc_volume model.

- The coefficient for income is estimated to be -14.30, indicating a negative relationship with sales, although it is not statistically significant (p-value > 0.05).

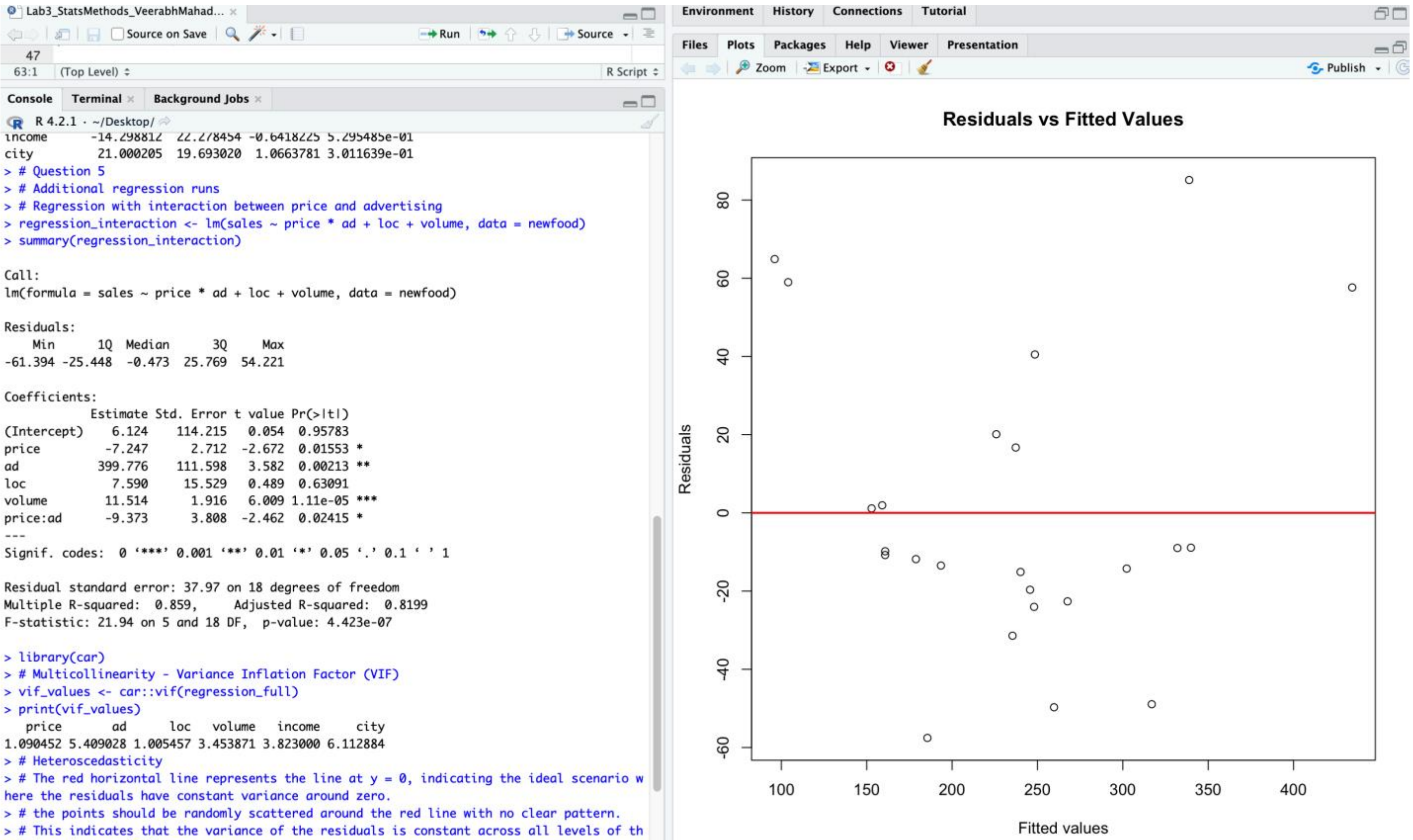
- The coefficient for city is estimated to be 21.00, indicating a positive relationship with sales, but again, it is not statistically significant (p-value > 0.05).

- The intercept in the regression_full model is estimated to be 158.48, which is different from the intercept in the regression_price_ad_loc_volume model (125.93). This difference is a result of the inclusion of additional variables.

Question 5

Lab3_StatsMethods_VeerabhMahad...

```r

# Question 4
# Regression with all variables
regression_full <- lm(sales ~ price + ad + loc + volume + income + city, data = newfood)
print(summary(regression_full)$coefficients)

# Question 5
# Additional regression runs
# Regression with interaction between price and advertising
regression_interaction <- lm(sales ~ price * ad + loc + volume, data = newfood)
summary(regression_interaction)

library(car)

# Multicollinearity - Variance Inflation Factor (VIF)
vif_values <- car::vif(regression_full)
print(vif_values)

# Heteroscedasticity
# The red horizontal line represents the line at y = 0, indicating the ideal scenario where the residuals have constant
# the points should be randomly scattered around the red line with no clear pattern.
# This indicates that the variance of the residuals is constant across all levels of the independent variable.
# Obtain residuals from the regression model
residuals <- residuals(regression_full)

# Plot residuals against fitted values
plot(fitted(regression_full), residuals,
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residuals vs Fitted Values")

# Add a horizontal line at y = 0 for reference
abline(h = 0, col = "red", lwd = 2)

```



- The VIF values reveal the extent of multicollinearity for each predictor variable in the regression model..

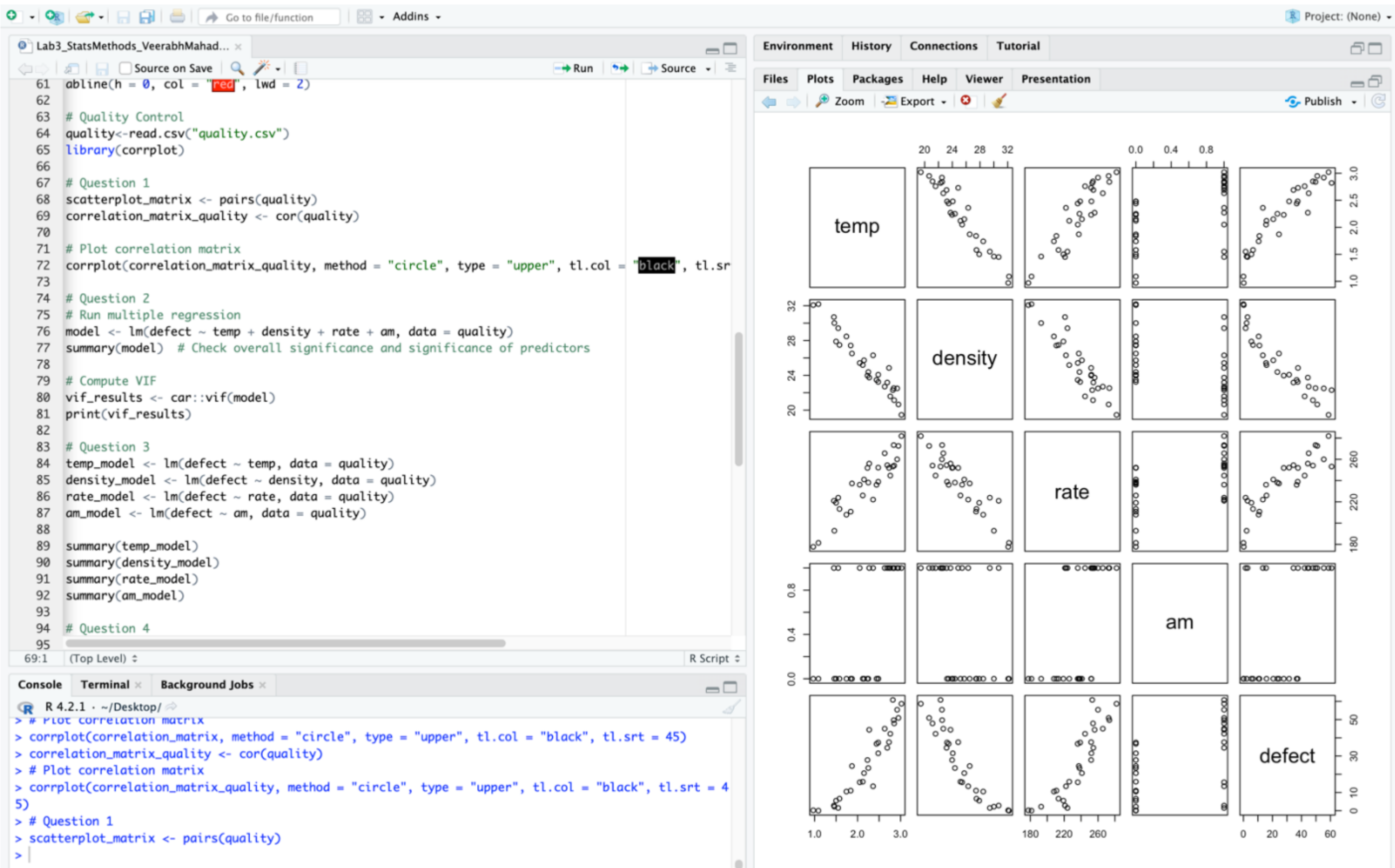
- A VIF value close to 1 implies minimal multicollinearity between the predictor variable and other variables in the model. Variables with VIF values near 1, such as "price," "loc," and "volume," indicate minimal multicollinearity with other variables.

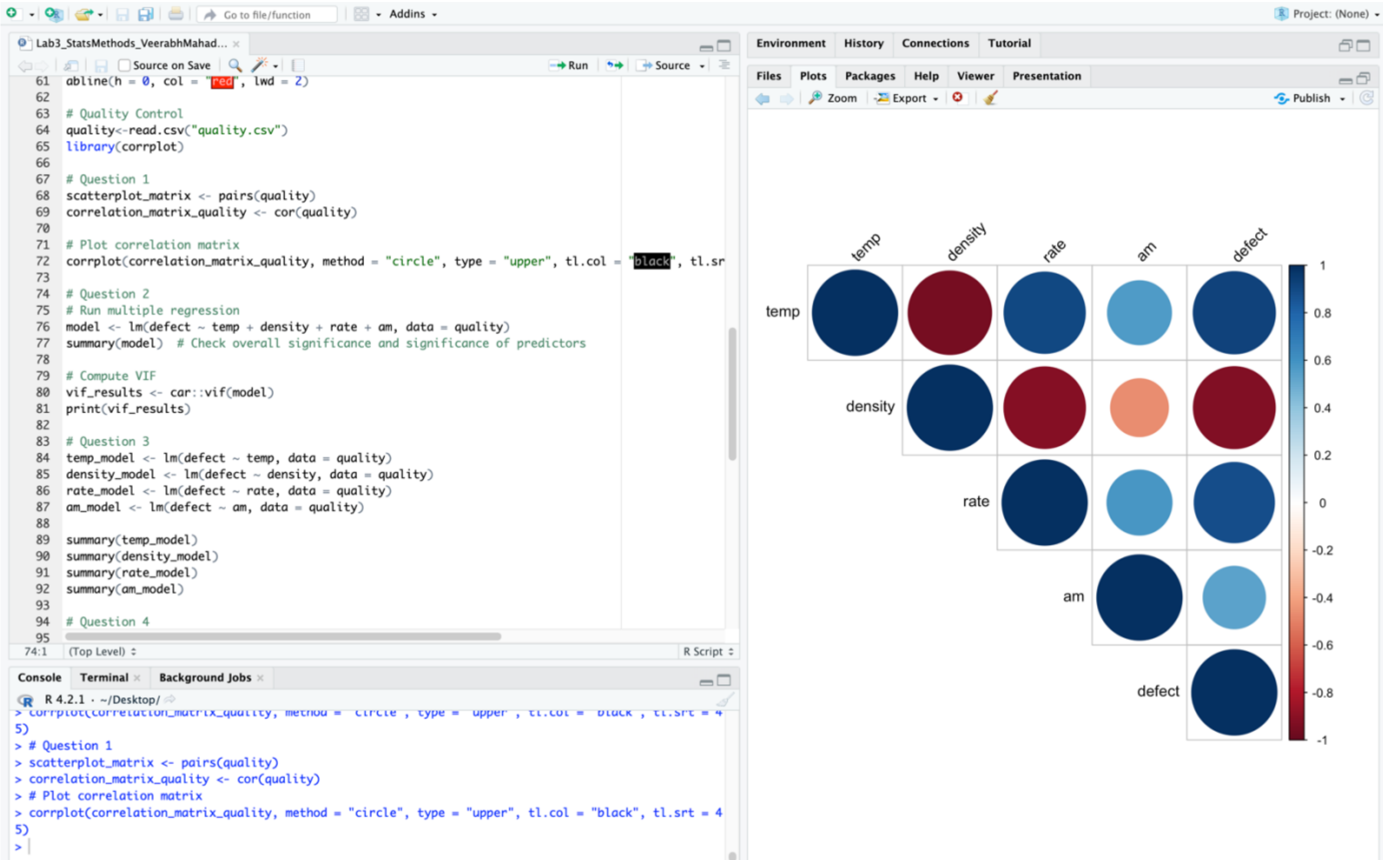
- Typically, a VIF exceeding 10 indicates substantial multicollinearity, signifying a significant increase in the variance of the coefficients. Predictor variables like "ad," "income," and "city" have VIF values ranging from 3.823 to 6.113, suggesting moderate multicollinearity but not to a concerning extent.

- Heteroscedasticity: The red horizontal line represents the line at y = 0, indicating the ideal scenario where the residuals have constant variance around zero. The points are randomly scattered around the red line with no clear pattern. This indicates that the variance of the residuals is constant across all levels of the independent variable.

Quality Control

Question 1





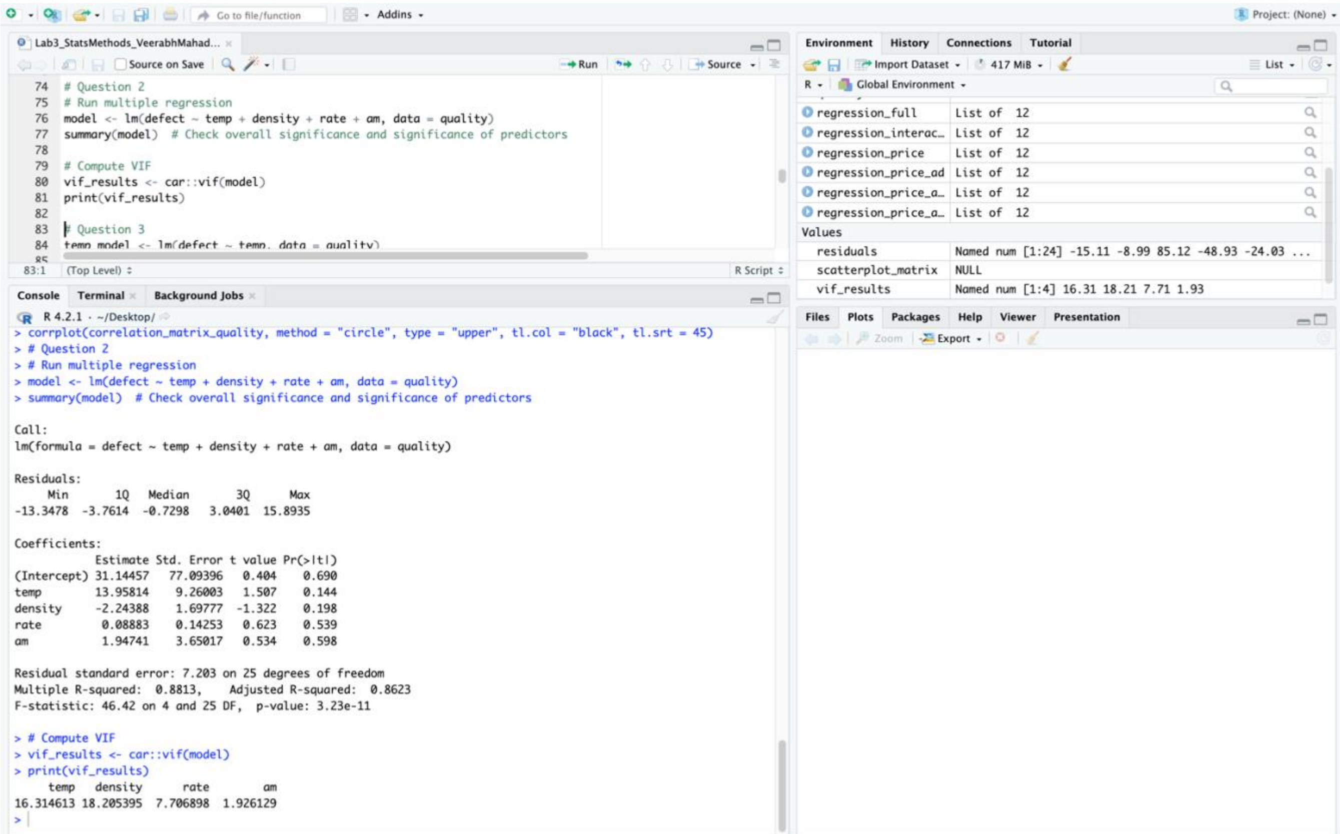
  - Strong correlations exist between temp and density, temp and rate, and density and rate. This indicates a tight interrelation among these variables.

  - temp, rate, and am show moderate positive correlations with defect, suggesting that increases in these variables are associated with higher defect rates.

  - In contrast, density exhibits a strong negative correlation with defect, implying that higher density correlates with lower defect rates.

  - The strong correlations between temp, density, and rate hint at potential multicollinearity issues in regression analysis.

Question 2



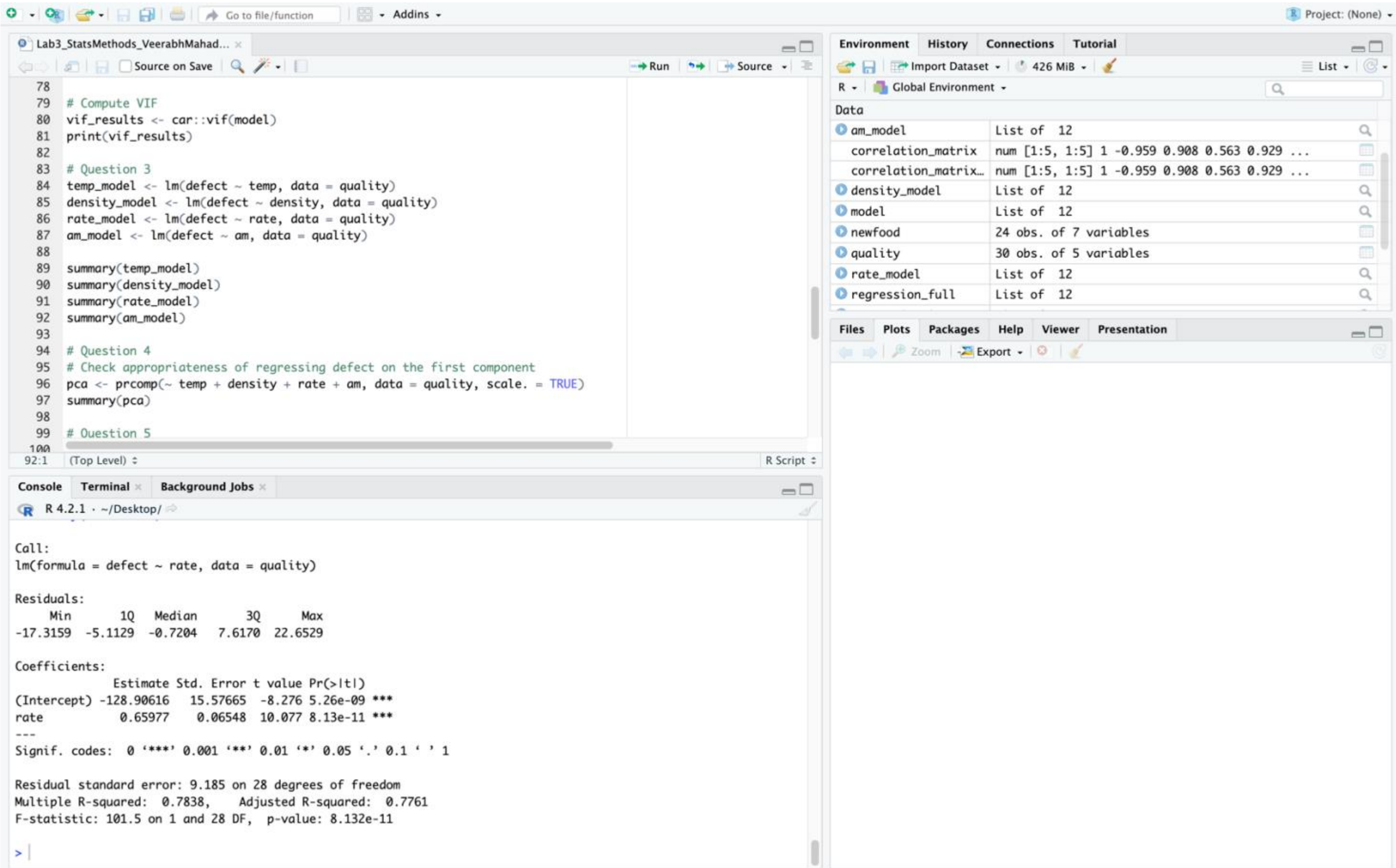
- The overall model is statistically significant (F-statistic = 46.42, p-value = 3.23e-11), indicating that at least one predictor has a significant relationship with the defect variable.

- None of the individual predictors (temp, density, rate, am) are statistically significant at the 0.05 level, as indicated by their p-values.

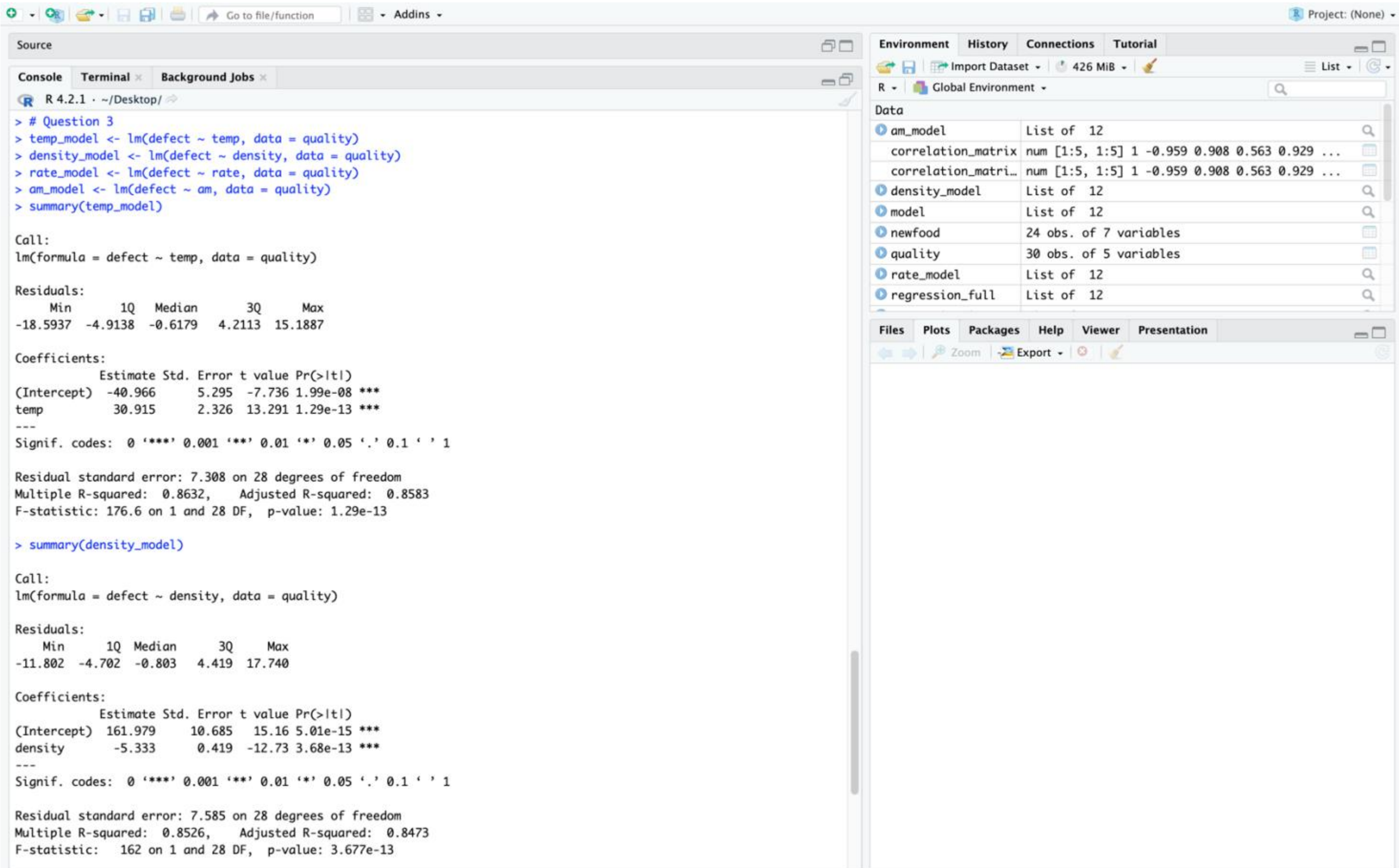
- VIF values for temp, density, and rate are considerably high, suggesting multicollinearity issues among these predictors.

- While the overall model is significant, the lack of significance among individual predictors and high VIF values indicate potential issues with model specification.

Question 3

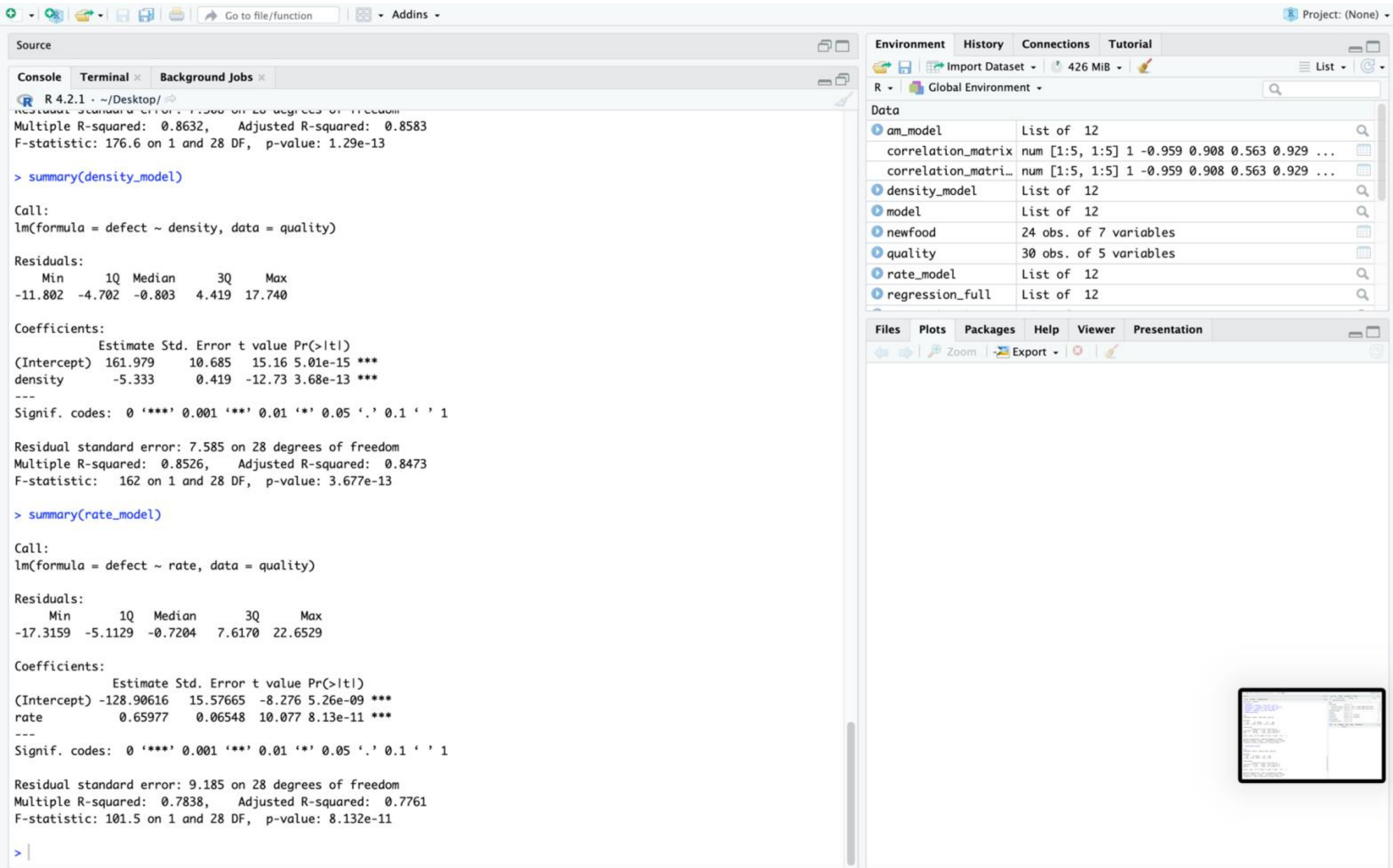

Go to file/function
Addins
Project: (None)
Lab3_StatsMethods_VeerabhMahad...
Source on Save
Run
Source
# Compute VIF
vif_results <- car::vif(model)
print(vif_results)
# Question 3
temp_model <- lm(defect ~ temp, data = quality)
density_model <- lm(defect ~ density, data = quality)
rate_model <- lm(defect ~ rate, data = quality)
am_model <- lm(defect ~ am, data = quality)
summary(temp_model)
summary(density_model)
summary(rate_model)
summary(am_model)
# Question 4
# Check appropriateness of regressing defect on the first component
pca <- prcomp(~ temp + density + rate + am, data = quality, scale. = TRUE)
summary(pca)
# Question 5
92:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.2.1 · ~/Desktop/
Call:
lm(formula = defect ~ rate, data = quality)
Residuals:
Min 1Q Median 3Q Max
-17.3159 -5.1129 -0.7204 7.6170 22.6529
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -128.90616 15.57665 -8.276 5.26e-09 ***
rate 0.65977 0.06548 10.077 8.13e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 9.185 on 28 degrees of freedom
Multiple R-squared: 0.7838, Adjusted R-squared: 0.7761
F-statistic: 101.5 on 1 and 28 DF, p-value: 8.132e-11
Environment
History
Connections
Tutorial
Import Dataset
426 MiB
List
Global Environment
Data
am_model List of 12
correlation_matrix num [1:5, 1:5] 1 -0.959 0.908 0.563 0.929 ...
correlation_matrix... num [1:5, 1:5] 1 -0.959 0.908 0.563 0.929 ...
density_model List of 12
model List of 12
newfood 24 obs. of 7 variables
quality 30 obs. of 5 variables
rate_model List of 12
regression_full List of 12
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export


Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Background Jobs
R 4.2.1 · ~/Desktop/
> # Question 3
> temp_model <- lm(defect ~ temp, data = quality)
> density_model <- lm(defect ~ density, data = quality)
> rate_model <- lm(defect ~ rate, data = quality)
> am_model <- lm(defect ~ am, data = quality)
> summary(temp_model)
Call:
lm(formula = defect ~ temp, data = quality)
Residuals:
Min 1Q Median 3Q Max
-18.5937 -4.9138 -0.6179 4.2113 15.1887
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -40.966 5.295 -7.736 1.99e-08 ***
temp 30.915 2.326 13.291 1.29e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.308 on 28 degrees of freedom
Multiple R-squared: 0.8632, Adjusted R-squared: 0.8583
F-statistic: 176.6 on 1 and 28 DF, p-value: 1.29e-13
> summary(density_model)
Call:
lm(formula = defect ~ density, data = quality)
Residuals:
Min 1Q Median 3Q Max
-11.802 -4.702 -0.803 4.419 17.740
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 161.979 10.685 15.16 5.01e-15 ***
density -5.333 0.419 -12.73 3.68e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.585 on 28 degrees of freedom
Multiple R-squared: 0.8526, Adjusted R-squared: 0.8473
F-statistic: 162 on 1 and 28 DF, p-value: 3.677e-13
Environment
History
Connections
Tutorial
Import Dataset
426 MiB
List
R
Global Environment
Data
am_model List of 12
correlation_matrix num [1:5, 1:5] 1 -0.959 0.908 0.563 0.929 ...
correlation_matri… num [1:5, 1:5] 1 -0.959 0.908 0.563 0.929 ...
density_model List of 12
model List of 12
newfood 24 obs. of 7 variables
quality 30 obs. of 5 variables
rate_model List of 12
regression_full List of 12
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export



- The coefficient estimate for temp is highly significant (p-value < 0.001), indicating a strong relationship between temperature and defect rate.

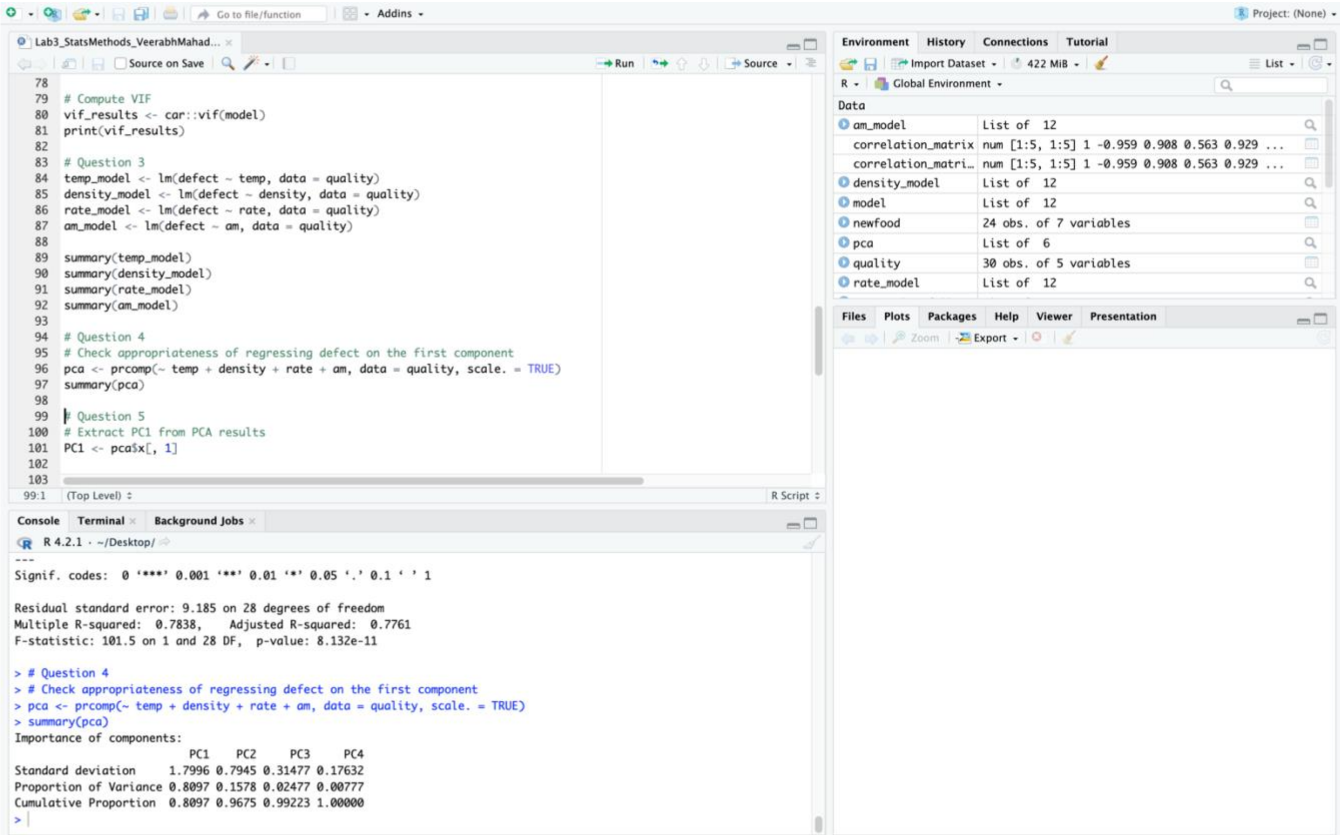
- Similarly, the coefficient estimate for density is highly significant (p-value < 0.001), suggesting a significant association between density and defect rate.

- The coefficient estimate for rate is also highly significant (p-value < 0.001), indicating a significant relationship between rate and defect rate.

- The coefficient estimate for am is not significant (p-value > 0.05), suggesting that there is insufficient evidence to conclude that am has a significant impact on defect rate in this model.

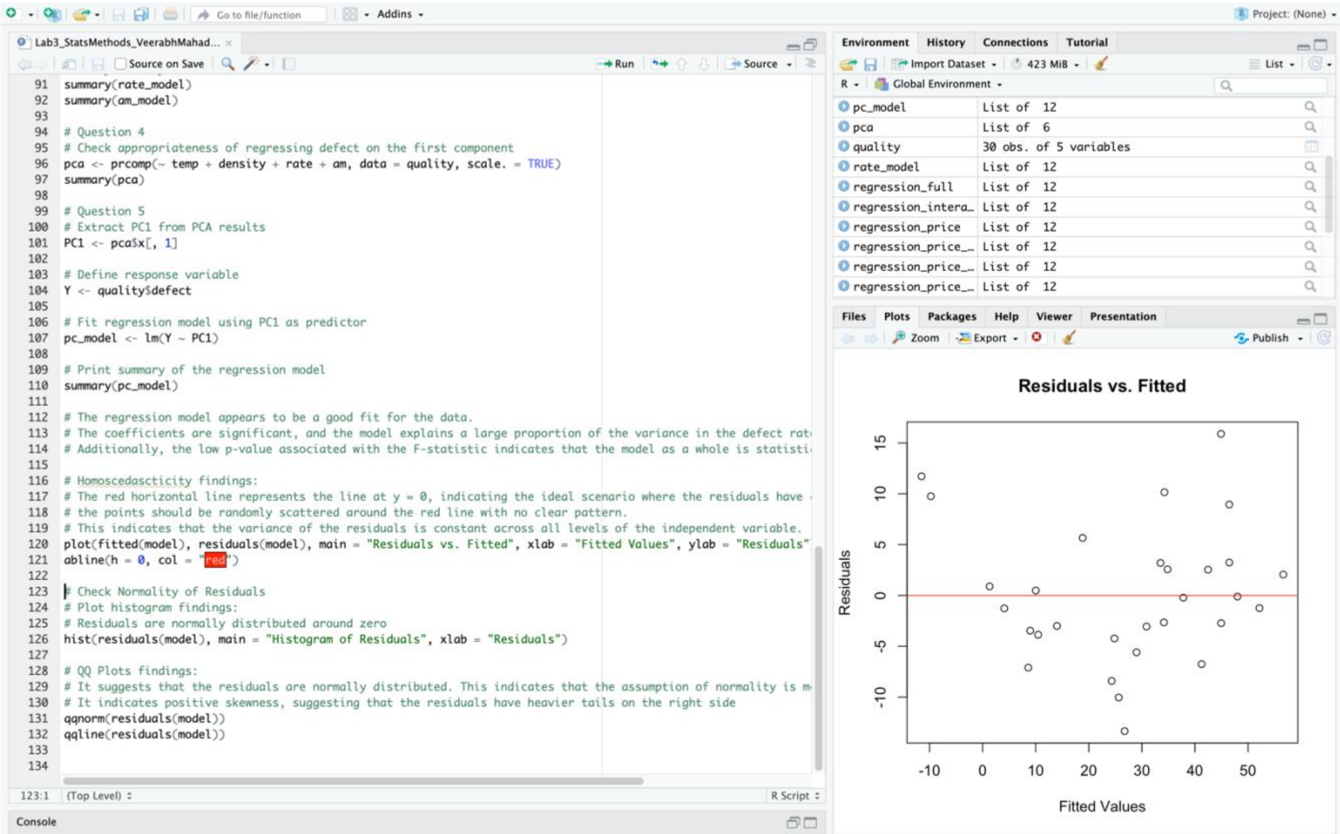
In summary, temp, density, and rate are significant predictors of defect rate, while `am` is not significant in this analysis.

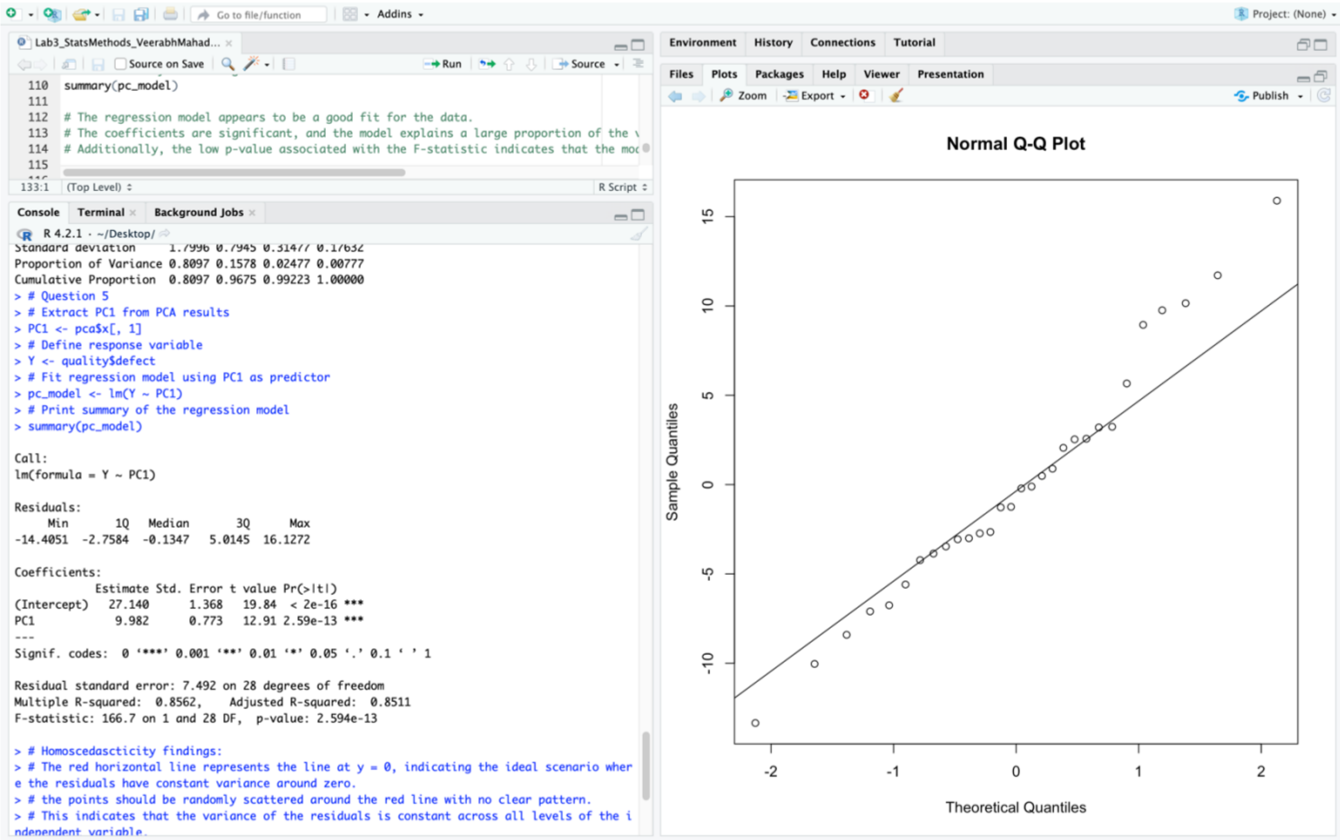
Question 4



PC1 accounts for approximately 80.97% of the total variance in the data. Regressing defect on the first principal component (PC1) may be appropriate given its high proportion of variance.

Question 5

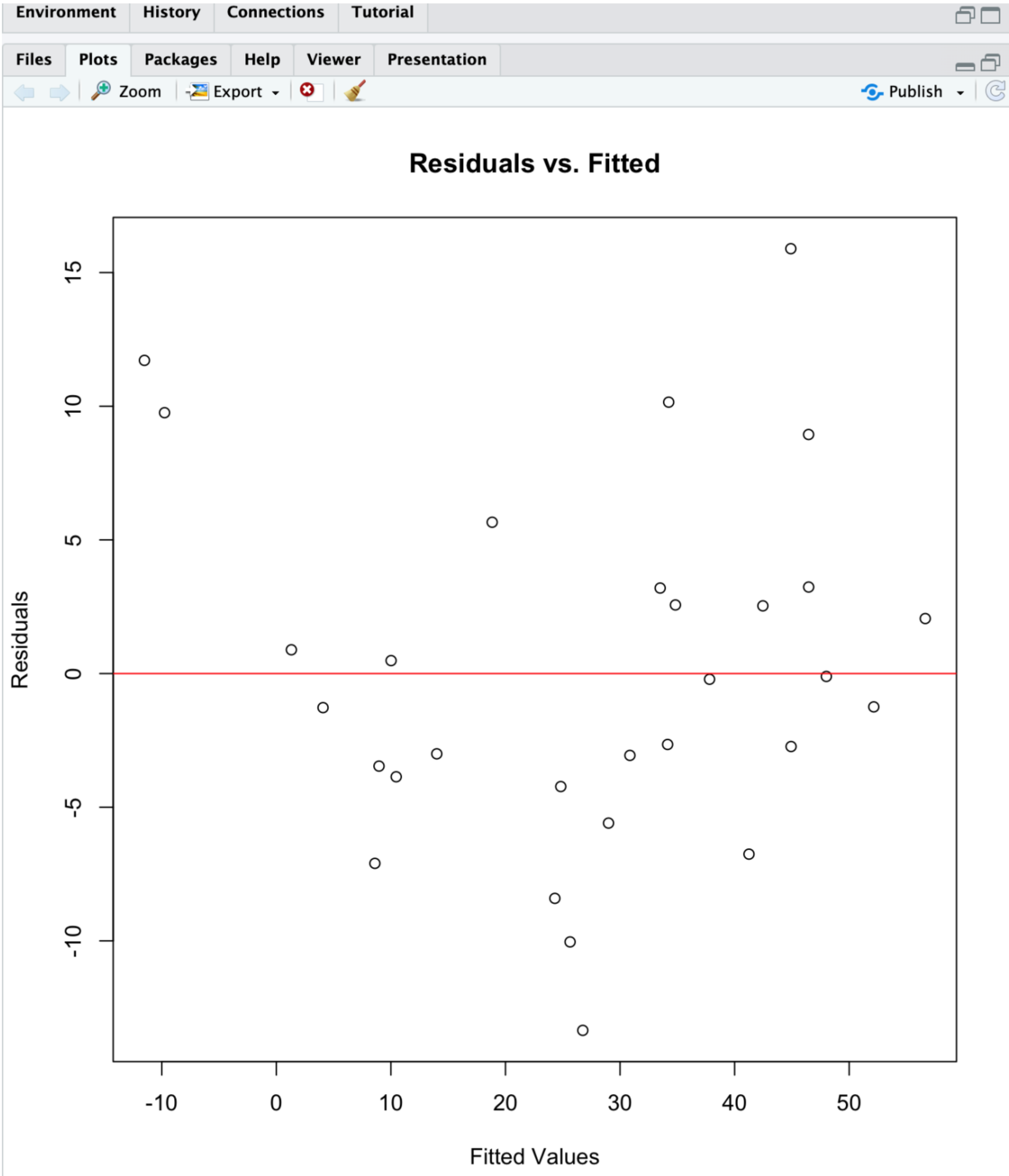

Go to file/function
Addins
Project: (None)
Lab3_StatsMethods_VeerabhMahad...
Source on Save
Run
Source
summary(rate_model)
summary(am_model)
# Question 4
# Check appropriateness of regressing defect on the first component
pca <- prcomp(~ temp + density + rate + am, data = quality, scale. = TRUE)
summary(pca)
# Question 5
# Extract PC1 from PCA results
PC1 <- pca$x[, 1]
# Define response variable
Y <- quality$defect
# Fit regression model using PC1 as predictor
pc_model <- lm(Y ~ PC1)
# Print summary of the regression model
summary(pc_model)
# The regression model appears to be a good fit for the data.
# The coefficients are significant, and the model explains a large proportion of the variance in the defect rat
# Additionally, the low p-value associated with the F-statistic indicates that the model as a whole is statisti
# Homoscedascticity findings:
# The red horizontal line represents the line at y = 0, indicating the ideal scenario where the residuals have
# the points should be randomly scattered around the red line with no clear pattern.
# This indicates that the variance of the residuals is constant across all levels of the independent variable.
plot(fitted(model), residuals(model), main = "Residuals vs. Fitted", xlab = "Fitted Values", ylab = "Residuals"
abline(h = 0, col = "red")
# Check Normality of Residuals
# Plot histogram findings:
# Residuals are normally distributed around zero
hist(residuals(model), main = "Histogram of Residuals", xlab = "Residuals")
# QQ Plots findings:
# It suggests that the residuals are normally distributed. This indicates that the assumption of normality is m
# It indicates positive skewness, suggesting that the residuals have heavier tails on the right side
qqnorm(residuals(model))
qqline(residuals(model))
123:1
(Top Level)
R Script
Console
Environment
History
Connections
Tutorial
Import Dataset
423 MiB
List
R
Global Environment
pc_model List of 12
pca List of 6
quality 30 obs. of 5 variables
rate_model List of 12
regression_full List of 12
regression_intera... List of 12
regression_price List of 12
regression_price_... List of 12
regression_price_... List of 12
regression_price_... List of 12
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Residuals vs. Fitted
Residuals
15
10
5
0
-5
-10
-10
0
10
20
30
40
50
Fitted Values



The regression model appears to be a good fit for the data.

The coefficients are significant, and the model explains a large proportion of the variance in the defect rate.

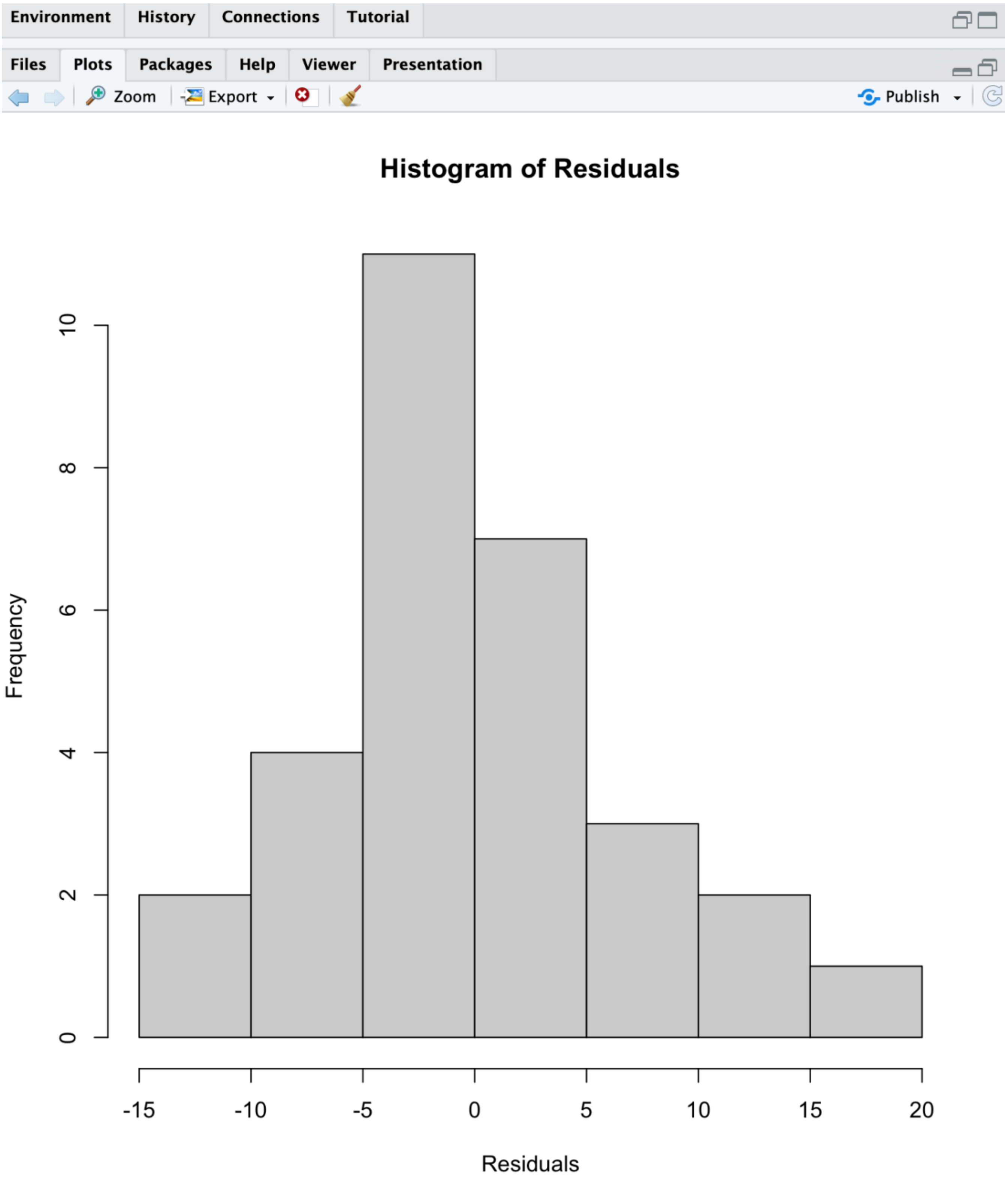
Additionally, the low p-value associated with the F-statistic indicates that the model as a whole is statistically significant.



Homoscedasticity:

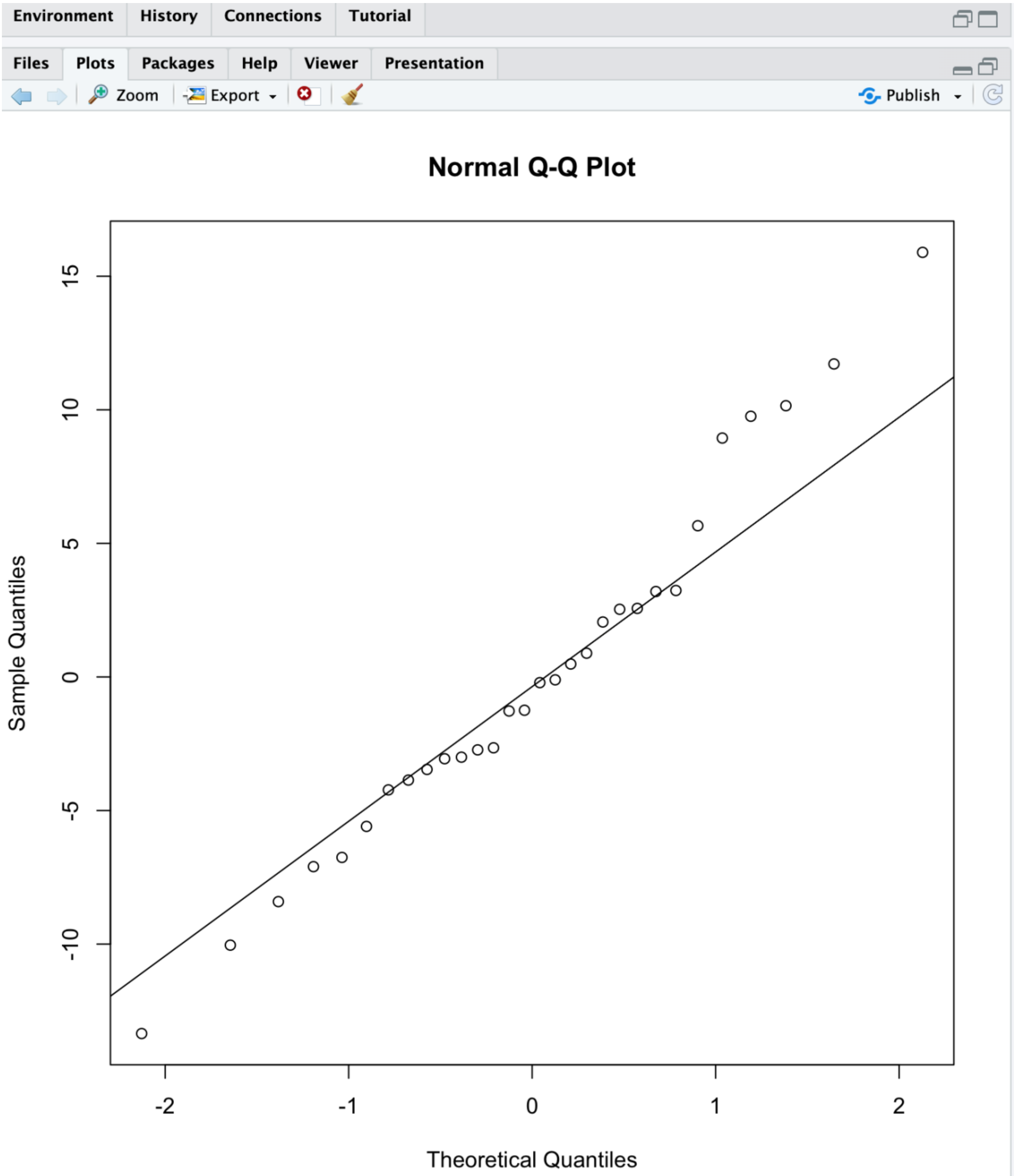
- The red horizontal line represents the line at y = 0, indicating the ideal scenario where the residuals have constant variance around zero.

- The points are randomly scattered around the red line with no clear pattern. This indicates that the variance of the residuals is constant across all levels of the independent variable.



Check Normality of Residuals

-Residuals are normally distributed around zero



QQ Plots findings:

-It suggests that the residuals are normally distributed. This indicates that the assumption of normality is met.

- It indicates positive skewness, suggesting that the residuals have heavier tails on the right side

Question 6

- The simple linear regression models also reveal that both temperature (temp) and density individually have a significant impact on the defect, as indicated by their low p-values and high t-values.

- The multiple regression model has an Adjusted R-squared value of 0.8623, indicating that approximately 86.23% of the variance in the defect can be explained by the predictors included in the model.

- The F-statistic's low p-value (3.23e-11) suggests that the model is statistically significant, meaning that at least one of the predictors has a non-zero effect on the defect.

- The VIF values for each predictor indicate that multicollinearity might be an issue, especially for temperature (temp) and density, which have high VIF values above 10.

- The reliability of the model can be ensured by QQ plot, normality of residuals, PCA, heteroscedasticity.