Question 1

a) Counterfactual is an unobserved scenario that helps understand what would have happened (potential outcome) if an individual/group was or wasn't treated. For example, for a treated individual (i) we know the outcome (y) given they were treated (x=1) but can't observe their outcome if they were not treated (x=0). This unobserved outcome is a counterfactual.

b) Selection refers to systematic differences between the treated and control group even before the treatment has occurred. It is a component of the difference in outcomes between the treated and untreated that is not related to the treatment itself. For treatment to have a causal effect, selection should be eliminated, that is, the assignment of the treatment should be random.

c) The causal effect of variable a (explanatory) on variable b (outcome) is the change in variable b associated with a change in variable a, given all other variables are controlled for (no selection).

d) The "back door" is a term used to describe unobserved variables that may be associated with the treatment and outcome, biasing the true relationship between them. Identifying and controlling for unobserved variables helps compute the true causal effect.

e) Randomized Controlled Trial (RCT) is a research method where the assignment into treatment and control groups is random. Random assignment eliminates (or minimizes) selection bias, which allows a more accurate assessment of the treatment's causal effect on the outcome.

f) Pre-experimental balance refers to the process of making sure the treatment and control groups are as similar as possible in their characteristics before the experiment. It ensures that the effect of other confounding variables is eliminated. One example, would be the proportion of sick people in a clinical trial should be similar amongst the treatment and control groups

Question 2

a) There are 320 students included in this dataset: 108 students with "AML" treatment, 106 students with "ANL" treatment, and 106 students with "B" treatment.

```
> nrow(HW.1.a)
[1] 320
> sum(HW.1.a$treatment == 'AML')
[1] 108
> sum(HW.1.a$treatment == 'ANL')
[1] 106
> sum(HW.1.a$treatment == 'B')
[1] 106
```

b) Descriptive statistics for the change in scores for all students and student groups based on treatment:

```
> HW.1.a$scorechange <- HW.1.a$postscore - HW.1.a$prescore
> summary(HW.1.a$scorechange)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -16.60   33.20   49.80   45.39   61.70   91.30
> aggregate(scorechange ~ treatment, data = HW.1.a, FUN = summary)
  treatment scorechange.Min. scorechange.1st Qu. scorechange.Median scorechange.Mean
1       AML        -4.700005           28.500000          45.099998        42.948148
2       ANL       -16.600000           33.199997          49.799998        45.258491
3         B       -16.600006           33.200005          49.799999        47.999056
  scorechange.3rd Qu. scorechange.Max.
1           53.400003        86.599998
2           66.399999        83.000003
3           64.324998        91.300003
```

c) Yes, the average change in scores is smaller for students who were subject to ANL condition (improved by 45.3 points) compared to students with B condition (improved by 48 points). However, the relationship is not statistically significant, as shown below.

```
> HW.1.a$female <- as.numeric(HW.1.a$female)
> anl_b <- HW.1.a[HW.1.a$treatment != 'AML', ]
> View(anl_b)
> anl_b$treatment <- ifelse(anl_b$treatment == 'ANL', 1, 0)
> ols <- lm(scorechange ~ treatment, data = anl_b)
> ols

Call:
lm(formula = scorechange ~ treatment, data = anl_b)

Coefficients:
(Intercept)    treatment
     47.999       -2.741

> summary(ols)

Call:
lm(formula = scorechange ~ treatment, data = anl_b)

Residuals:
    Min      1Q  Median      3Q     Max
-64.599 -12.058   1.801  18.401  43.301

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   47.999      2.171  22.114   <2e-16 ***
treatment     -2.741      3.070  -0.893    0.373
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.35 on 210 degrees of freedom
Multiple R-squared:  0.003781,  Adjusted R-squared:  -0.0009624
F-statistic: 0.7971 on 1 and 210 DF,  p-value: 0.373
```

d) Yes, the average change in scores is smaller for students who were subject to AML condition (improved by 42.9 points) compared to students with B condition (improved by 48 points). The average improvement with condition AML is even smaller than for students with condition ANL. However, the difference is not statistically significant, as shown below.

```
> aml_b <- HW.1.a[HW.1.a$treatment != 'ANL', ]
> View(aml_b)
> aml_b$treatment <- ifelse(aml_b$treatment == 'AML', 1, 0)
> ols <- lm(scorechange ~ treatment, data = aml_b)
> summary(ols)

Call:
lm(formula = scorechange ~ treatment, data = aml_b)

Residuals:
    Min      1Q  Median      3Q     Max
-64.599 -14.448   1.801  10.452  43.652

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   47.999      1.953  24.572   <2e-16 ***
treatment     -5.051      2.750  -1.837   0.0676 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.11 on 212 degrees of freedom
Multiple R-squared:  0.01567,   Adjusted R-squared:  0.01102
F-statistic: 3.374 on 1 and 212 DF,  p-value: 0.06763
```

e) No, we cannot consider the relationships found in c) and d) as causal. There is no information on the randomness in the assignment of conditions, so selection bias might have affected the results. Chi-Squared Test for Independence illustrates that there is insufficient evidence to conclude a significant association between treatment and gender in the observed data. However, there were statistically significant differences in prescores between the treatment groups.

```
> ols <- lm(treatment ~ female, data = aml_b)
> summary(ols)

Call:
lm(formula = treatment ~ female, data = aml_b)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5407 -0.5407  0.4593  0.4593  0.5570

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44304    0.05626   7.874 1.75e-13 ***
female       0.09770    0.07084   1.379    0.169
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5001 on 212 degrees of freedom
Multiple R-squared:  0.008893,  Adjusted R-squared:  0.004218
F-statistic: 1.902 on 1 and 212 DF,  p-value: 0.1693
```

```
> ols <- lm(treatment ~ female, data = anl_b)
> summary(ols)

Call:
lm(formula = treatment ~ female, data = anl_b)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51111 -0.49180 -0.00146  0.50820  0.50820

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.51111    0.05295   9.654   <2e-16 ***
female      -0.01931    0.06979  -0.277    0.782
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5023 on 210 degrees of freedom
Multiple R-squared:  0.0003643, Adjusted R-squared:  -0.004396
F-statistic: 0.07653 on 1 and 210 DF,  p-value: 0.7823
```

```
> ols <- lm(treatment ~ female, data = aml_anl)
> summary(ols)

Call:
lm(formula = treatment ~ female, data = aml_anl)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5489 -0.5489  0.4511  0.4511  0.5679

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.43210    0.05546   7.792 2.91e-13 ***
female       0.11677    0.07034   1.660   0.0984 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4991 on 212 degrees of freedom
Multiple R-squared:  0.01283,   Adjusted R-squared:  0.008176
F-statistic: 2.756 on 1 and 212 DF,  p-value: 0.09838
```

```
> contingency_table <- table(HW.1.a$treatment, HW.1.a$female)
> chi_squared_test <- chisq.test(contingency_table)
> print(chi_squared_test)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 3.1125, df = 2, p-value = 0.2109
```

```
> aggregate(prescore ~ treatment, data = HW.1.a, FUN = summary)
  treatment prescore.Min. prescore.1st Qu. prescore.Median prescore.Mean prescore.3rd Qu.
1       AML      13.00000         21.30000        37.90000       41.28148          62.80000
2       ANL       0.00000          8.30000        24.90000       26.07453          41.50000
3         B       0.00000          8.30000        24.90000       27.48396          39.42500
  prescore.Max.
1     104.30000
2      91.30000
3      91.30000
> anova_result <- aov(prescore ~ treatment, data = HW.1.a)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
treatment     2  15153    7577    15.1 5.47e-07 ***
Residuals   317 159114     502
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question 3

a) How many students participated in the experiment in each condition? How many of them are female?

```
# Assuming you have your data loaded in a DataFrame called 'data'

# Count the number of students in each condition
condition_counts = data['treatment'].value_counts()

# Count the number of female students in each condition
female_counts = data[data['female'] == 1]['treatment'].value_counts()

# Print the results
print("Number of students in each condition:")
print(condition_counts)

print("\nNumber of female students in each condition:")
print(female_counts)
```

```
Number of students in each condition:
AML    139
B      130
ANL    128
Name: treatment, dtype: int64

Number of female students in each condition:
AML    88
B      79
ANL    71
Name: treatment, dtype: int64
```

b) Show suggestive evidence that this data came from a well executed RCT

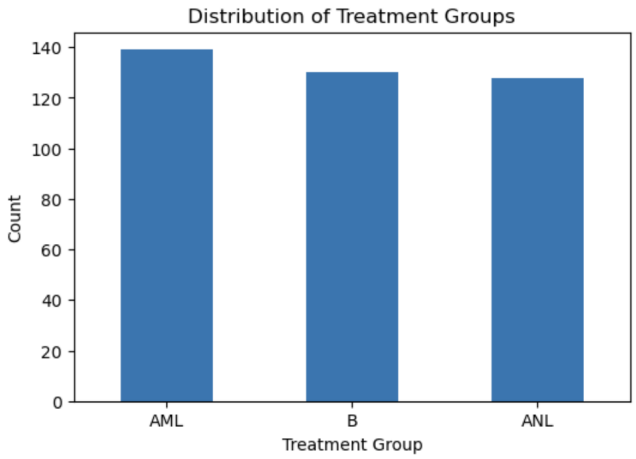
In a well-executed RCT, you are expected to see a relatively even distribution of subjects across treatment groups, similar summary statistics for baseline measures, and no significant differences in baseline covariates between groups. Any deviations from these expectations may raise questions about the randomization process or the quality of the trial.

In [5]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
# Load the data into a DataFrame
data = pd.read_csv('HW-1-b.csv')  # Replace 'your_data_file.csv' with the actual file path
print(data.head())
```

```
    stu_id   prescore  postscore treatment  female
0  2108062  24.900000  83.000000         B       1
1  2108238  33.200001  74.699997         B       0
2  2108181  33.200001  83.000000         B       1
3  2108160  41.500000  91.300003         B       1
4  2108213  24.900000  91.300003         B       1
```

In [6]:
```
#Question 2

#Step 1: Check the distribution of treatment to check if it is randomized
plt.figure(figsize=(6, 4))
data['treatment'].value_counts().plot(kind='bar', rot=0)
plt.xlabel('Treatment Group')
plt.ylabel('Count')
plt.title('Distribution of Treatment Groups')
plt.show()
```

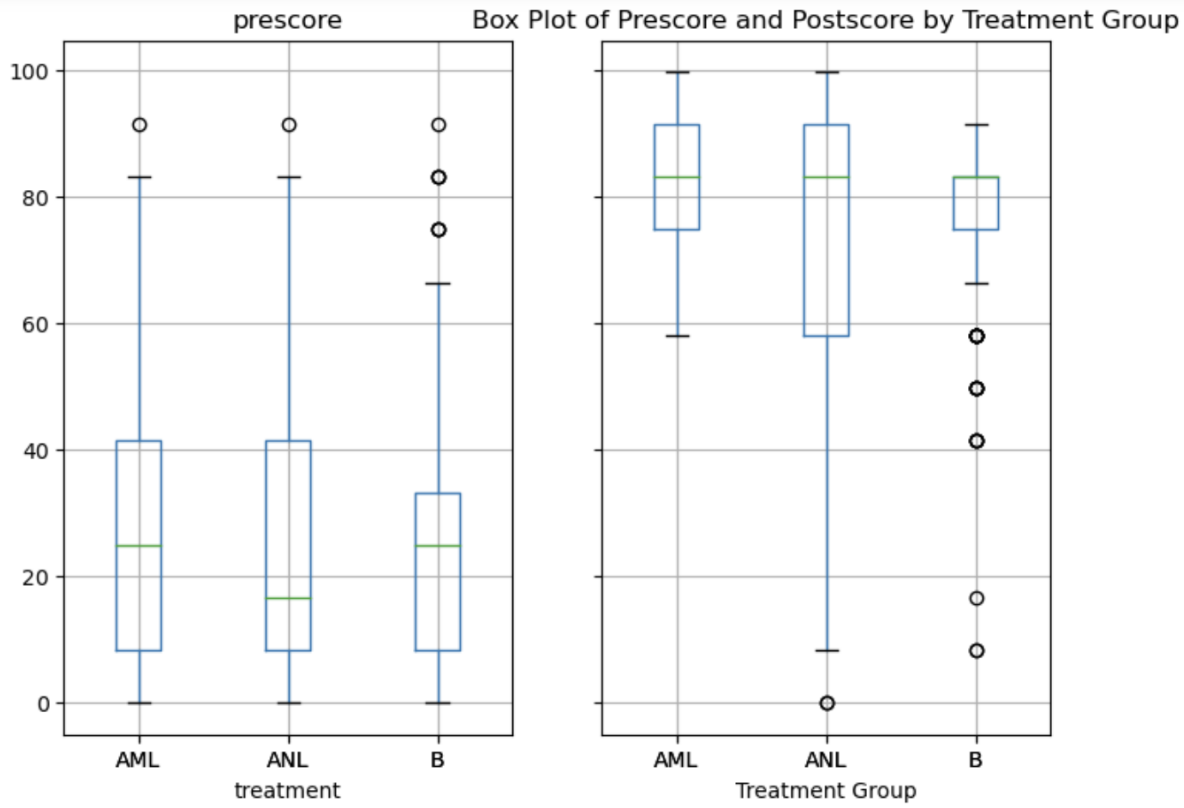


Based on the graph, the distribution of subjects amongst the various treatment groups is roughly even.

In [7]:
```
# Calculate summary statistics for 'prescore' and 'postscore' by treatment group
summary_stats = data.groupby('treatment')[['prescore', 'postscore']].describe()

# Print the summary statistics
print(summary_stats)

# Box plot of 'prescore' and 'postscore' by treatment group
data.boxplot(column=['prescore', 'postscore'], by='treatment', figsize=(8, 6))
plt.title('Box Plot of Prescore and Postscore by Treatment Group')
plt.suptitle('')  # Remove the default title
plt.xlabel('Treatment Group')
plt.ylabel('Score')
plt.show()
```

```
          prescore                                                    \
             count       mean        std  min  25%   50%        75%
treatment
AML          139.0  27.467626  22.583944  0.0  8.3  24.9  41.500000
ANL          128.0  25.418750  22.064468  0.0  8.3  16.6  41.500000
B            130.0  26.113077  21.557507  0.0  8.3  24.9  33.200001

                   postscore                                          \
                 max     count       mean        std        min        25%
treatment
AML        91.300003     139.0  84.373381  11.153407  58.099998  74.699997
ANL        91.300003     128.0  70.355469  23.887440   0.000000  58.099998
B          91.300003     130.0  75.402308  15.571996   8.300000  74.699997
```

```
             50%        75%        max
treatment
AML         83.0  91.300003  99.599998
ANL         83.0  91.300003  99.599998
B           83.0  83.000000  91.300003
```



In [8]:
```
# Calculate the proportion of females in each treatment group
proportion_female_by_treatment = data.groupby('treatment')['female'].mean()

# Print the proportion of females - the balance of baseline covariates (e.g., 'female') between treatment groups:
print(proportion_female_by_treatment)
```

```
treatment
AML    0.633094
ANL    0.554688
B      0.607692
Name: female, dtype: float64
```

The proportion of females is roughly even

```
> aggregate(prescore ~ treatment, data = HW.1.b, FUN = summary)
  treatment prescore.Min. prescore.1st Qu. prescore.Median prescore.Mean prescore.3rd Qu.
1       AML       0.00000          8.30000        24.90000       27.46763         41.50000
2       ANL       0.00000          8.30000        16.60000       25.41875         41.50000
3         B       0.00000          8.30000        24.90000       26.11308         33.20000
  prescore.Max.
1      91.30000
2      91.30000
3      91.30000
> anova_result <- aov(prescore ~ treatment, data = HW.1.b)
> summary(anova_result)
             Df Sum Sq Mean Sq F value Pr(>F)
treatment     2    292   145.9   0.299  0.742
Residuals   394 192163   487.7
> aggregate(female ~ treatment, data = HW.1.b, FUN = summary)
  treatment female.Min. female.1st Qu. female.Median female.Mean female.3rd Qu. female.Max.
1       AML    0.0000000       0.0000000     1.0000000   0.6330935       1.0000000   1.0000000
2       ANL    0.0000000       0.0000000     1.0000000   0.5546875       1.0000000   1.0000000
3         B    0.0000000       0.0000000     1.0000000   0.6076923       1.0000000   1.0000000
> anova_result <- aov(female ~ treatment, data = HW.1.b)
> summary(anova_result)
             Df Sum Sq Mean Sq F value Pr(>F)
treatment     2   0.42  0.2113   0.877  0.417
Residuals   394  94.90  0.2409
```

c) Show descriptive statistics for the improvement in test scores from the pre-test to the post-test

In [9]:
```
#Question 3
# Calculate the improvement in test scores
data['improvement'] = data['postscore'] - data['prescore']

# Calculate descriptive statistics for the improvement
improvement_stats = data['improvement'].describe()

# Print the descriptive statistics
print(improvement_stats)
```

```
count    397.000000
mean      50.552645
std       21.946801
min      -24.899998
25%       41.499996
50%       49.800003
75%       66.400002
max       99.599998
Name: improvement, dtype: float64
```

```
> HW.1.b$scorechange <- HW.1.b$postscore - HW.1.b$prescore
> aggregate(scorechange ~ treatment, data = HW.1.b, FUN = summary)
  treatment scorechange.Min. scorechange.1st Qu. scorechange.Median scorechange.Mean
1       AML          8.299995           41.500003          58.100000        56.905755
2       ANL        -24.899998           31.124998          49.799998        44.936719
3         B        -16.600006           41.499996          49.800002        49.289231
  scorechange.3rd Qu. scorechange.Max.
1           66.400003        99.599998
2           66.400000        91.300003
3           66.400000        91.300003
```

d) Show whether the improvement in the scores (as defined above) reduced when smartphones were allowed into the classroom and not used to assist instruction (compared to when they were banned from the classroom)?

```
#Question 4
# Filter the data for the "B" (banned) and "ANL" (not used to assist instruction) conditions
banned_condition = data[data['treatment'] == "B"]
anl_condition = data[data['treatment'] == "ANL"]

# Calculate the improvement in scores for each condition
banned_improvement = banned_condition['improvement']
anl_improvement = anl_condition['improvement']

# Perform a t-test for the difference in means
t_statistic, p_value = stats.ttest_ind(anl_improvement, banned_improvement, equal_var=False)

# Print the results
print(f"T-statistic: {t_statistic}")
print(f"P-value: {p_value}")

# Check if the p-value is less than your chosen significance level (e.g., 0.05) to determine significance
if p_value < 0.05:
    print("The difference in improvement between ANL and B conditions is statistically significant.")
else:
    print("There is no statistically significant difference in improvement between ANL and B conditions.")
```

```
T-statistic: -1.5113577772364484
P-value: 0.13194808378600495
There is no statistically significant difference in improvement between ANL and B conditions.
```

The difference in improvement scores between ANL and B conditions is not statistically significant and hence, we cannot confirm that the improvement in scores reduced when smartphones were allowed into the classroom and not used to assist instruction.

e) Show whether the improvement in the scores (as defined above) increased when smartphones were allowed into the classroom and used to assist instruction (compared to when they were banned from the classroom)?

```
import pandas as pd
from scipy import stats

# Assuming you have your data loaded in a DataFrame called 'data'

# Filter the data for the "B" (banned) and "AML" (used to assist instruction) conditions
banned_condition = data[data['treatment'] == "B"]
aml_condition = data[data['treatment'] == "AML"]

# Calculate the improvement in scores for each condition
banned_improvement = banned_condition['improvement']
aml_improvement = aml_condition['improvement']

# Perform a t-test for the difference in means
t_statistic, p_value = stats.ttest_ind(aml_improvement, banned_improvement, equal_var=False)

# Print the results
print(f"T-statistic: {t_statistic}")
print(f"P-value: {p_value}")

# Check if the p-value is less than your chosen significance level (e.g., 0.05) to determine significance
if p_value < 0.05:
    print("The difference in improvement between AML and B conditions is statistically significant.")
else:
    print("There is no statistically significant difference in improvement between AML and B conditions.")
```

```
T-statistic: 3.112178718353024
P-value: 0.002073386758845547
The difference in improvement between AML and B conditions is statistically significant.
```

The difference in improvement scores between AML and B conditions is statistically significant and hence, the improvement in scores increased when smartphones were allowed into the classroom and used to assist instruction.

f) Show whether the results obtained in d) and e) above are different for female and male students

```
import pandas as pd
from scipy import stats

# Assuming you have your data loaded in a DataFrame called 'data'

# Filter the data for the "B" (banned) and "AML" (used to assist instruction) conditions
banned_condition = data[data['treatment'] == "B"]
aml_condition = data[data['treatment'] == "AML"]

# Calculate the improvement in scores for each condition and each gender
banned_female_improvement = banned_condition[banned_condition['female'] == 1]['improvement']
banned_male_improvement = banned_condition[banned_condition['female'] == 0]['improvement']

aml_female_improvement = aml_condition[aml_condition['female'] == 1]['improvement']
aml_male_improvement = aml_condition[aml_condition['female'] == 0]['improvement']

# Perform separate t-tests for each gender and condition
female_t_statistic, female_p_value = stats.ttest_ind(aml_female_improvement, banned_female_improvement, equal_var=False)
male_t_statistic, male_p_value = stats.ttest_ind(aml_male_improvement, banned_male_improvement, equal_var=False)

# Print the results for females
print("Results for Female Students:")
print(f"T-statistic: {female_t_statistic}")
print(f"P-value: {female_p_value}")
if female_p_value < 0.05:
    print("The difference in improvement for females between AML and B conditions is statistically significant.")
else:
    print("There is no statistically significant difference in improvement for females between AML and B conditions.")
```

```
# Print the results for females
print("Results for Female Students:")
print(f"T-statistic: {female_t_statistic}")
print(f"P-value: {female_p_value}")
if female_p_value < 0.05:
    print("The difference in improvement for females between AML and B conditions is statistically significant.")
else:
    print("There is no statistically significant difference in improvement for females between AML and B conditions.")

# Print the results for males
print("\nResults for Male Students:")
print(f"T-statistic: {male_t_statistic}")
print(f"P-value: {male_p_value}")
if male_p_value < 0.05:
    print("The difference in improvement for males between AML and B conditions is statistically significant.")
else:
    print("There is no statistically significant difference in improvement for males between AML and B conditions.")
```

```
Results for Female Students:
T-statistic: 0.9046076532393857
P-value: 0.36707031175069693
There is no statistically significant difference in improvement for females between AML and B conditions.

Results for Male Students:
T-statistic: 4.146568730927063
P-value: 7.890490223601153e-05
The difference in improvement for males between AML and B conditions is statistically significant.
```

The improvement in scores between AML and B conditions is statistically significant for males, but it is not statistically significant for females. The improvement in scores between AML and B conditions increased for males but we cannot confirm that it increased for females.

References:

- Lecture notes
- ChatGPT-4 for coding in R: Linear regression, Chi-Squared Test for Independence

- ChatGPT-4 for helping with code on t-tests in python