**Part I**

Exercise 1:

a. Downloading the data from the course site and read it into R under the name "flint".

**Command:**

**A close up of a text

Description automatically generated**

**Output:**

**A screenshot of a computer

Description automatically generated**

b. The EPA states a water source is especially dangerous if the lead level is 15 PPB or greater.

**Command:**

A black text with a dollar sign

Description automatically generated

**Output:**

**A close-up of numbers

Description automatically generated**

**🡪** The proportion of locations tested that have dangerous lead levels is **0.04436229.**

c. Reporting the mean copper level for only test sites in the North region.

**Command:**

A black and white text

Description automatically generated

**Output:**

A close-up of a sign

Description automatically generated

d. Reporting the mean copper level for only test sites with dangerous lead levels (at least 15 PPB).

**Command:**

**A close-up of a sign

Description automatically generated**

**Output:**

**A close-up of a sign

Description automatically generated**

e. Reporting the mean lead and copper levels.

**Command:**

**A black text on a white background

Description automatically generated**

**Output:**

**A screenshot of a computer

Description automatically generated**

f. Creating a box plot with a good title for the lead levels.

**Command:**

****

**Output:**

**A diagram of a box plot

Description automatically generated**

g. Based on what you see in part (f), does the mean seem to be a good measure of center for the data? Report a more useful statistic for this data.

**Command:**

median(flint$Pb)

**Output:**

[1] 0

🡪 By looking at the box plot, the data for lead levels is skewed to the left, which means the mean is not a good measure of center for the data. A more useful statistic would be the median, which has a value of 0.

Exercise 2:

life = read.table("https://ucla.box.com/shared/static/rqk4lc030pabv30wknx2ft9jy848ub9n.txt", header = TRUE)

a. Constructing a scatterplot of Life against Income.

**Command:**

plot(y = life$Life, x = life$Income, xlab = "Income", ylab = "Life Expectancy", main = "Life Expectancies vs. per Capita Income")

**Output:**

A graph of a chart

Description automatically generated with medium confidence

🡪 As income increases, so does life expectancy, showing a positive relationship.

b. Constructing the boxplot and histogram of Income.

**Command:**

hist(life$Income, xlab = "Income", main = "Life Expectancies vs. per Capita Income")

**Output:**

**A graph of life expectancy

Description automatically generated**

🡪 Yes, there are many outliers past an approximate income of 2500.

c. Splitting the data set into two parts: One for which the Income is strictly below $1000, and one for which the Income is at least $1000.

**Command:**

**A close up of words

Description automatically generated**

d. Using the data for which the Income is below $1000. Plotting Life against Income and computing the correlation coefficient.

**Command:**

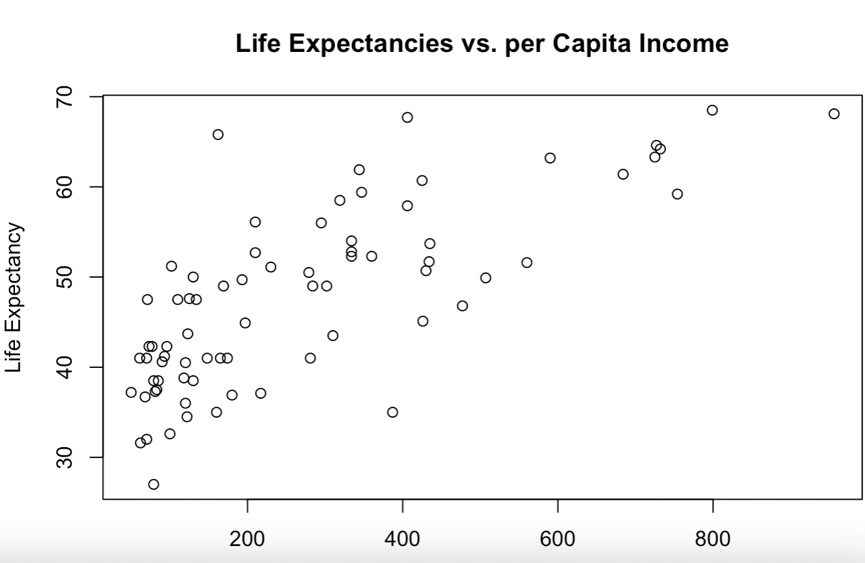
plot(lowerincome$Life~lowerincome$Income, xlab = "Income", ylab = "Life Expectancy", main

= "Life Expectancies vs. per Capita Income")

cor(x = lowerincome$Life, y = lowerincome$Income)

**Output:**

[1] 0.752886



Exercise 3:

maas <- read.table("https://ucla.box.com/shared/static/tv3cxooyp6y8fh6gb0qj2cxihj8klg1h.txt", header = TRUE)

a. Computing the summary statistics for lead and zinc using the summary() function.

**Command:**

**A close-up of a logo

Description automatically generated**

**Output:**

**A close-up of numbers

Description automatically generated**

b. Plotting two histograms: one of lead and one of log(lead).

**Command:**

**A close-up of a computer code

Description automatically generated**

**Output:**

1. **A graph of a person with a bar graph

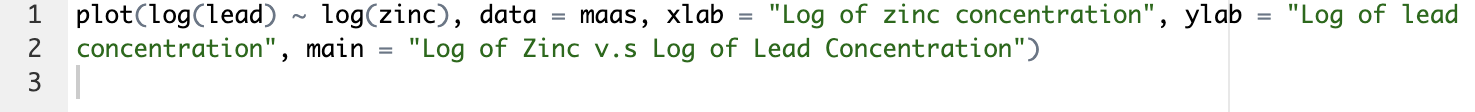
Description automatically generated with medium confidence**

2. **A graph of a graph

Description automatically generated**

c. Plotting log(lead) against log(zinc).

**Command:**

****

**Output:**

**A graph of a graph of zinc

Description automatically generated with medium confidence**

d. Giving different colors and sizes to lead concentrations in 155 locations based on level of risk for surface oil.

**Command:**

**A computer code with black text

Description automatically generated**

**Output:**

**A chart showing the growth of soil

Description automatically generated**

Exercise 4:

**Command:**

LA <- read.table("https://ucla.box.com/shared/static/d189x2gn5xfmcic0dmnhj2cw94jwvqpa.txt", header=TRUE)

find.package("maps")

**Output:**

#[1] "/Library/Frameworks/R.framework/Versions/4.3-x86\_64/Resources/library/maps"

a. Plotting the data point locations with good formatting for the axes and title. Then adding the outline of LA County.

**Command:**

A close-up of a white background

Description automatically generated

**Output:**

A map of the neighborhood

Description automatically generated

b. Plot the variable Schools against the variable Income and describe what you see. Ignore the data points on the plot for which Schools = 0. Use what you learned about subsetting with logical statements to first create the objects you need for the scatter plot. Then, create the scatter plot.

**Command:**

**A close-up of a white background

Description automatically generated**

**Output:**

**A graph of a number of people

Description automatically generated with medium confidence**

Do you see any relationship between income and school performance?

🡪 There's a moderate positive linear relationship between income and LA school performance. That means that neighbourhoods with higher incomes tend to have better performing schools. However, this does not mean that one is the effect of the other.

Exercise 5:

customer\_data <- read.csv("https://ucla.box.com/shared/static/y2y8rcie7mjw2h5t92x9dfcp133tc90h.csv")

a. Are there any missing values in the dataset? If so, how many are there and which variables have missing values?

🡪 There are 22 missing values from the data set. 10 of them were part of the age variable, 5 were part of the income variable, and the remaining 7 were part of the purchase amount variable.

b. What is the data type of each variable? Are there any variables that should be converted to a different data type?

🡪 Customer ID is an continuous numerical variable. Age is a continuous numerical variable. Gender is a binary categorical variable. Income is a continuous numerical variable. Education is an ordinal categorical variable. Marital status is a nominal categorical variable. Purchase amount is a continuous numerical variable. There are no variables that should be converted into another data type except for gender, which should be nominal rather than binary.

**Part II**

Exercise 1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Two-way table** | | **Body Image** | | | |
| **About**  **right** | **Overweight** | **Underweight** | **Total** |
| **Gender** | **Female** | 310 | 130 | 30 | **470** |
| **Male** | 290 | 68 | 72 | **430** |
| **Total** | **600** | **198** | **102** | 900 |

a. Total students that are happy with their body weight (“About right”) = 600

Percentage of students happy with their body weight = (600/900) x 100% = **66.67%**

Since 66.67% of the students in the data set have “About right” body image, the students in general **are happy** with their body weight.

b. The best choice of graph to visualize this data of differences in body image between females and males would be a **grouped bar chart**, since it is a comparison of categorical 🡪 categorical data. This graph can represent multiple categories (“About right”, “Overweight”, and “Underweight”) for each gender side by side, making it easier to compare the distribution between genders.

c. Percentage of female students who feel “About right” = (310/470) x 100% = **65.96%**

Percentage of male students who feel “About right” = (290/430) x 100% = **67.44%**

It can be observed that 67.44% of male students feel their weight is about right, compared to 65.96% of female students. Therefore, male students are slightly more likely to feel they are about right than female students.

d. Number of female students who do not feel “About right” = 470 - 310 = 160

Number of male students who do not feel “About right” = 430 - 290 = 140

Percentage of female students who feel they are overweight = (130/160) × 100% = **81.25%**

Percentage of male students who feel they are overweight = (68/140) × 100% = **48.57%**

Percentage of female students who feel they are underweight = (30/160) × 100% = **18.75%**

Percentage of male students who feel they are underweight = (72/140) × 100% = **51.43%**

It can be observed that female students who do not feel ‘about right’ with their body image are more likely to feel they are overweight (81.25% vs. 48.57%), whereas male students who do not feel ‘about right’ with their body image are more likely to feel they are underweight (51.43% vs. 18.75%). Therefore, there **are differences** between the two gender groups regarding how they perceive their body image when they do not feel ‘about right'. **Female students are more likely to perceive themselves as overweight, while male students are more likely to perceive themselves as underweight.**

Exercise 2:

a. *U.S. Census Data on Median Family Income vs. Percentage with a BA*

**Form:** The relationship appears to be **linear**. As the percentage of people with a BA increases, the median family income also increases in a relatively straight-line manner.

**Direction:** The direction is **positive/increasing** because as one variable (percentage with a BA) increases, the other variable (median family income) also increases.

**Strength:** The association is **moderately strong**. While there's some scatter and vertical variation around the line, the points mostly follow a clear upward trend.

**Outliers:** There **is a** potential outlier towards the extreme right of the scatterplot where the percentage with a BA is much higher than the general trend.

**Contextual Meaning:** In the context of the data, states with a higher percentage of individuals holding a BA degree generally have a higher median family income. This suggests an association between higher educational attainment and higher family incomes in those states. However, it's important to note that this association does not necessarily imply causation. There could be other confounding variables or factors at play that contribute to this observed relationship.

b*. Relationship between Fuel Consumption and Speed*

**Form:** The relationship appears to be **non-linear,** specifically U-shaped. Initially, as speed increases, fuel consumption decreases, but after a certain point as speed continues to increase, fuel consumption starts to increase again.

**Direction:** The direction is **neither** positive nor negative since it's decreasing initially and then increasing after a certain speed.

**Strength:** The association is **strong**, especially given the clear U-shaped pattern with very little scatter around the curve.

**Outliers:** There **is a** potential outlier at the far left in the provided scatterplot, representing very low speed but high fuel consumption, which stands apart from the general trend of the data.

**Contextual Meaning:** At lower speeds, cars might be consuming more fuel due to reasons like stop-and-go driving or inefficiencies. As speed increases, the car becomes more efficient up to a certain point. However, at very high speeds, the fuel consumption starts to increase again, possibly due to reasons like increased air resistance and engine strain. However, it's important to note that this association does not necessarily imply causation.

Exercise 3:

a. The **explanatory variable** (or independent variable) is the **"Start Median Salary"** since it is on the **x-axis** and we are using it to help explain or predict the response. The **response variable** (or dependent variable) is the **"Mid-Career Median Salary"** since it is on the **y-axis** and we are trying to predict or explain its values based on the start median salary.

b. The median salary is used instead of the mean since it is **less sensitive to outliers or extreme values**. If there are a few very high or very low salaries, the mean would be skewed by these values. The median, being the middle value when all are sorted in order, is not affected by these extremes. Thus, using the median can provide a more representative measure of the central tendency of the data, especially when the data might have skewed distributions.

c. Yes, the median mid-career salary can be estimated using the given regression equation:

Mid-Career = −7.699 + 1.989 × Start Med

Plugging in a start median salary of 60, we get:

Mid-Career = −7.699 + 1.989 × 60 = −7.699 + 119.34 = 111.641

So, given a median starting salary of 60 (in thousands of dollars), the estimated median mid-career salary would be approximately **$111,641 per year.**

d. Yes, using the regression equation, we can estimate the median mid-career salary for a starting salary of 100:

Mid-Career = −7.699 + 1.989 × 100 = −7.699 + 198.9 = 191.201

So, given a median starting salary of 100 (in thousands of dollars), the estimated median mid-career salary would be approximately **$191,201 per year.**

Exercise 4:

Given:

Mean of calories (ȳ) = 141.67

Mean of % alcohol (x̄) = 11.03

Standard Deviation of calories = 46.34

Standard Deviation of % alcohol = 2.32

Correlative Coefficient (r) = 0.95

a. Using the formula for slope:

b = r x (SD of y / SD of x) = 0.95 x (46.34/2.32)

**b ≈ 19**

Using the formula for intercept (a):

a = Mean of y – b x Mean of x = 141.67 – (19 x 11.03)

**a ≈ -68**

b. Equation of regression line: **y = a + bx**

*Interpretation:*

This equation represents the estimated linear relationship between the percentage of alcohol content in a wine and the corresponding calories in a five-ounce serving. It suggests that when the % alcohol is zero, the expected calorie count in a five-ounce serving of wine is approximately -68, which doesn't make practical sense. This is why we don't usually use the y-intercept value in isolation for interpretation. However, the slope of 19, indicates that for each 1% increase in alcohol content, the calories in a five-ounce serving of wine increase by approximately 19, when all other factors are held constant. This suggests a positive relationship or association between alcohol content and calorific content of wine.

c. The coefficient of determination (r²):

r² = (0.95)2 = **0.9025**

*Interpretation:*

The coefficient of determination measures how much the variation in response variable 𝑦 is explained by the predictor 𝑥. In this case, the coefficient of determination (r²) is 0.9025 or 90.25%. This means that approximately 90.25% of the variability in the calories in a five-ounce serving of wine can be explained by the % alcohol content. The remaining variability might be due to other factors not included in the model.

d. Adding a wine with 20% alcohol and 0 calories would be considered an outlier, given the existing data set and the observed relationship between % alcohol and calories. This new point will have the following effects:

1. **Value of r (correlation coefficient):** The addition of an outlier can significantly affect the correlation between the two variables. In this case, since the new point deviates strongly from the established general trend, the value of r will decrease, meaning the strength of the linear relationship between the two variables will weaken.

2. **Slope of the regression line:** The outlier will influence the angle and positioning of the least squares regression line, potentially causing it to tilt and adjust towards this new point. As a result, the slope of the regression line may decrease compared to the original line and likely be less steep, given that this new point suggests that higher alcohol content doesn't necessarily mean more calories. This point contradicts the trend observed in the rest of the data.

Exercise 5:

a. Yes, if the doctor places the most severe patients in the antidepressants group, it will affect his ability to compare the effectiveness of the antidepressants and the "talk therapy." This is because the initial severity of the patients' conditions can be a confounding variable. If the most severe patients show more or less improvement than the other group, it might not be because of the effectiveness of the antidepressants alone but could be influenced by the initial severity of their conditions. This design introduces bias into the study and makes the results less generalizable or potentially skewed.

b. It is not recommended for the doctor to know which treatment allocation each patient receives since it could introduce observer bias. The doctor might, even subconsciously, evaluate a patient’s progress differently based on the knowledge of their treatment. For example, he might interpret a small improvement in a patient on antidepressants more favourably than the same improvement in a patient on talk therapy because of his inherent belief, thereby possibly skewing the results and reducing the validity of the study's findings.

c. *Improvements to the plan:*

1) **Randomization:** Patients should be randomly assigned to each treatment group. This ensures that both observed and unobserved confounding variables are equally distributed between the two groups and reduces bias.

2) **Blinding:** Ideally, the study should be double-blind, meaning neither the patients nor the doctor knows which treatment each patient is receiving. This helps in eliminating biases in treatment administration and outcome assessment.

3) **Matching:** If there's a concern about the severity of depression affecting the outcomes, the patients can be matched based on the severity of their depression. For every severely depressed patient assigned to the antidepressants group, another severely depressed patient should be assigned to the talk therapy group, and so forth.

4) **Control for Confounders:** Ensure that other potential confounding variables, like age, gender, or duration of depression, are accounted for in the analysis.

5) **Multiple Evaluations:** Instead of evaluating patients just after six months, consider multiple evaluations at different time points to understand the progression and consistency of improvement.