POL501 - Problem Set 1

Solutions

2024-10-06

Contents

Answers Quiz 1

Answer to Question 1

A study exploring the effects of political corruption on tax compliance found that in countries with less political corruption, private firms were less likely to engage in tax fraud.

What type of relationship is described in the study?

- (a) Positive relationship
- (b) Negative relationship
- (c) Non-linear relationship
- (d) No relationship

Answer: (a) Positive relationship

Justification: The study states that in countries with less political corruption, private firms are less likely to engage in tax fraud. This indicates that as corruption decreases, tax fraud decreases, showing a positive relationship because both variables change in the same direction.

However, note that the relationship described in the study depends on how we frame the variables. If we measure tax fraud (or non-compliance) against political corruption, a *positive relationship* means that as corruption decreases, tax fraud also decreases. However, if we instead focus on *tax compliance* as the variable, the relationship becomes *negative*: as corruption decreases, tax compliance increases. This reversal occurs because tax compliance and tax fraud are opposites. It's crucial to clearly define which variables we are measuring and their directions to avoid confusion when interpreting relationships between them. Clear identification ensures accurate understanding of the study's findings.

Answer to Question 2

A study found that as the number of community events increases from a few to a moderate number, community cohesion improves. However, when the number of events increases from moderate to high, community cohesion decreases.

What kind of association is this?

- (a) Positive association
- (b) Negative association
- (c) Non-linear association
- (d) The variables are independent of each other

Answer: (c) Non-linear association

Justification: The study found that community cohesion improves with an increase in community events up to a moderate number, but then decreases as the number of events becomes high. This indicates a non-linear association, where the relationship between community events and cohesion changes direction at different levels.

Answer to Question 3

A study of one recent primary for the Republican party revealed the following data. Researchers were surprised to observe the results.

Candidate Names	Total Votes Won	Campaign Spending ($)
Candidate A	10,000	500,000
Candidate B	15,000	300,000
Candidate C	8,000	700,000
Candidate D	12,000	400,000

According to this data table, what kind of association is found between "spending" and "votes"?

- (a) Positive association
- (b) Negative association
- (c) Non-linear association
- (d) The variables are independent of each other

Answer: (b) Negative association

Justification: By examining the data table and sorting the candidates by spending shows that higher spending generally corresponds to fewer votes: - Candidate B: $300,000, 15,000 votes - Candidate D: $400,000, 12,000 votes - Candidate A: $500,000, 10,000 votes - Candidate C: $700,000, 8,000 votes

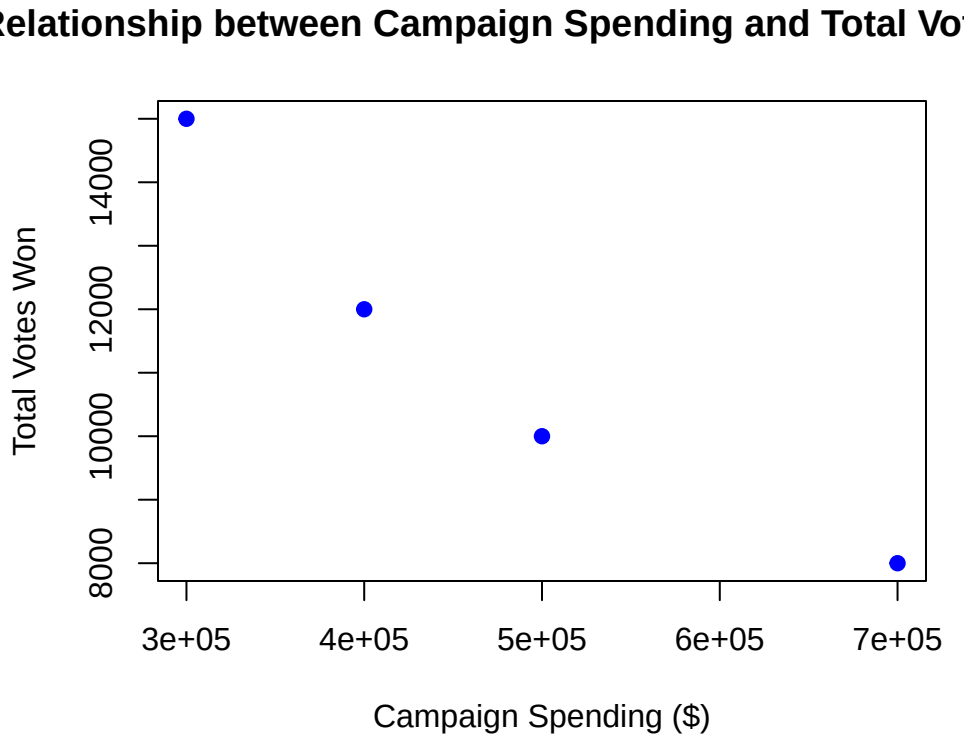
This pattern suggests a negative association where higher spending is associated with fewer votes. Therefore, there is a negative relationship between the two variables.

We can confirm this by also plotting the two variables in a scatter plot.

```r
# Create the dataframe
candidate_data <- data.frame(
  Candidate_Names = c("Candidate A", "Candidate B", "Candidate C", "Candidate D"),
  Total_Votes_Won = c(10000, 15000, 8000, 12000),
  Campaign_Spending = c(500000, 300000, 700000, 400000)
)

# View the dataframe
print(candidate_data)
```

```
##   Candidate_Names Total_Votes_Won Campaign_Spending
## 1     Candidate A           10000             5e+05
## 2     Candidate B           15000             3e+05
## 3     Candidate C            8000             7e+05
## 4     Candidate D           12000             4e+05
```

```r
# Set plot dimensions (width, height) in inches
par(pin = c(4, 2.5))  # Example: 4 inches wide, 2.5 inches tall

# Plot the relationship between Total Votes and Campaign Spending
plot(
  candidate_data$Campaign_Spending, # variable on the x-axis
  candidate_data$Total_Votes_Won, # y-axis variable
  xlab = "Campaign Spending ($)",
  ylab = "Total Votes Won",
  main = "Relationship between Campaign Spending and Total Votes Won",
  pch = 19, col = "blue"
)
```



Answer to Question 4

Consider the level of satisfaction with local government services as expressed through a survey using ratings: very unsatisfied, unsatisfied, neutral, satisfied, very satisfied. This question assesses the perceived effectiveness of services such as public transportation, parks, and emergency responses.

Which kind of variable type is this?

- (a) Regular Categorical (Nominal)
- (b) Ordinal Categorical (Ordinal)
- (c) Numerical (Discrete)
- (d) Numerical (Continuous)

Answer: (b) Ordinal Categorical (Ordinal)

Justification: The levels of satisfaction (very unsatisfied, unsatisfied, neutral, satisfied, very satisfied) are ordered categories. This makes the variable ordinal since it has a meaningful order but the intervals between levels are not necessarily equal.

Answer to Question 5

In a psychological study examining stress triggers, participants were categorized by their primary work environment settings, such as 'open-plan offices', 'private offices', and 'remote work from home'. Researchers sought to determine if these settings influenced reported stress levels during work hours.

Which is the explanatory variable and which is the response?

- (a) Work environment setting is the response, and stress level is the explanatory variable.
- (b) Stress level is the response, and work environment setting is the explanatory variable.

Answer: (b) Stress level is the response, and work environment setting is the explanatory variable.

Justification: The study examines how different work environment settings (open-plan offices, private offices, remote work from home) influence the reported stress levels of participants. Therefore, the work environment setting is the explanatory variable (independent variable), and the stress level is the response variable (dependent variable).

Answer to Question 6

Part (a)

Mean, Median, and Standard Deviation

1. We will calculate the mean, median, and standard deviation for both the number of social activities per month and the social trust levels across the neighborhoods.
2. The algebraic formulas are:

$$\text{Mean}(\mu) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Median: Arrange the data in ascending order and find the middle value.

$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

We can create a dataframe and use dyplyr or base R to do the calculations.

```
# Create the data vectors
neighborhood_id <- c("N1", "N2", "N3", "N4", "N5", "N6", "N7", "N8", "N9", "N10", "N11")
number_of_social_activities <- c(1.0, 2.9, 4.8, 6.7, 8.6, 10.5, 12.4, 14.3, 16.2, 18.1, 20.0)
social_trust_index <- c(21.48, 48.90, 76.20, 96.73, 96.87, 98.58, 102.14, 85.35, 59.21, 37.10,

# Create the dataframe
df <- data.frame(
  neighborhood_id,
  number_of_social_activities,
  social_trust_index
)

# View the dataframe
print(df)
```

```
##    neighborhood_id number_of_social_activities social_trust_index
## 1               N1                         1.0              21.48
## 2               N2                         2.9              48.90
## 3               N3                         4.8              76.20
## 4               N4                         6.7              96.73
## 5               N5                         8.6              96.87
## 6               N6                        10.5              98.58
```

```
## 7               N7                      12.4            102.14
## 8               N8                      14.3             85.35
## 9               N9                      16.2             59.21
## 10             N10                      18.1             37.10
## 11             N11                      20.0              1.20
```

```r
# Using Base R:

# Calculate mean, median, and standard deviation for number_of_social_activities
mean_activities <- mean(df$number_of_social_activities)
median_activities <- median(df$number_of_social_activities)
sd_activities <- sd(df$number_of_social_activities)

# Calculate mean, median, and standard deviation for social_trust_index
mean_trust <- mean(df$social_trust_index)
median_trust <- median(df$social_trust_index)
sd_trust <- sd(df$social_trust_index)

# Print the results
cat("Base R Summary Statistics:\n")
```

```
## Base R Summary Statistics:
```

```r
cat("Number of Social Activities - Mean:", mean_activities,
    ", Median:", median_activities, ", SD:", sd_activities, "\n")
```

```
## Number of Social Activities - Mean: 10.5 , Median: 10.5 , SD: 6.301587
```

```r
cat("Social Trust Index - Mean:", mean_trust,
    ", Median:", median_trust, ", SD:", sd_trust, "\n")
```

```
## Social Trust Index - Mean: 65.79636 , Median: 76.2 , SD: 34.78167
```

Alternatively we can use the package `dyplyr` and get the output in a dataframe. Importantly, the code here allows to get the summary statistics for each variable organized by rows. This is format is easy to read and good for reports.

```r
# Install dplyr if it's not already installed
# install.packages("dplyr")

# Load dplyr
library(dplyr)


# Compute summary statistics for 'number_of_social_activities'
stats_activities <- df %>%
  summarise(
    Variable = "Number of Social Activities",
    Mean = mean(number_of_social_activities),
    Median = median(number_of_social_activities),
    SD = sd(number_of_social_activities)
```

```
  )

# Compute summary statistics for 'social_trust_index'
stats_trust <- df %>%
  summarise(
    Variable = "Social Trust Index",
    Mean = mean(social_trust_index),
    Median = median(social_trust_index),
    SD = sd(social_trust_index)
  )

# Combine the two data frames
summary_stats_dplyr <- bind_rows(stats_activities, stats_trust)

# View the summary statistics
print(summary_stats_dplyr)
```

```
##                       Variable     Mean Median        SD
## 1 Number of Social Activities 10.50000   10.5  6.301587
## 2          Social Trust Index 65.79636   76.2 34.781671
```

5. Final Results:
    - Number of Social Activities (per month):
        - Mean: 10.5
        - Median: 10.5
        - Standard Deviation: 6.01
    - Social Trust Level (out of 100):
        - Mean: 65.48
        - Median: 76.2
        - Standard Deviation: 33.80

Interpretation: - Number of Social Activities: - The mean and median being equal (10.5) indicates a symmetric distribution. - The standard deviation of 6.01 suggests moderate variability around the mean. - Social Trust Level: - The mean (65.48) and median (76.2) indicate a skewed distribution, likely left-skewed because the mean is less than the median. - The high standard deviation (33.80) indicates high variability in social trust levels among the neighborhoods.

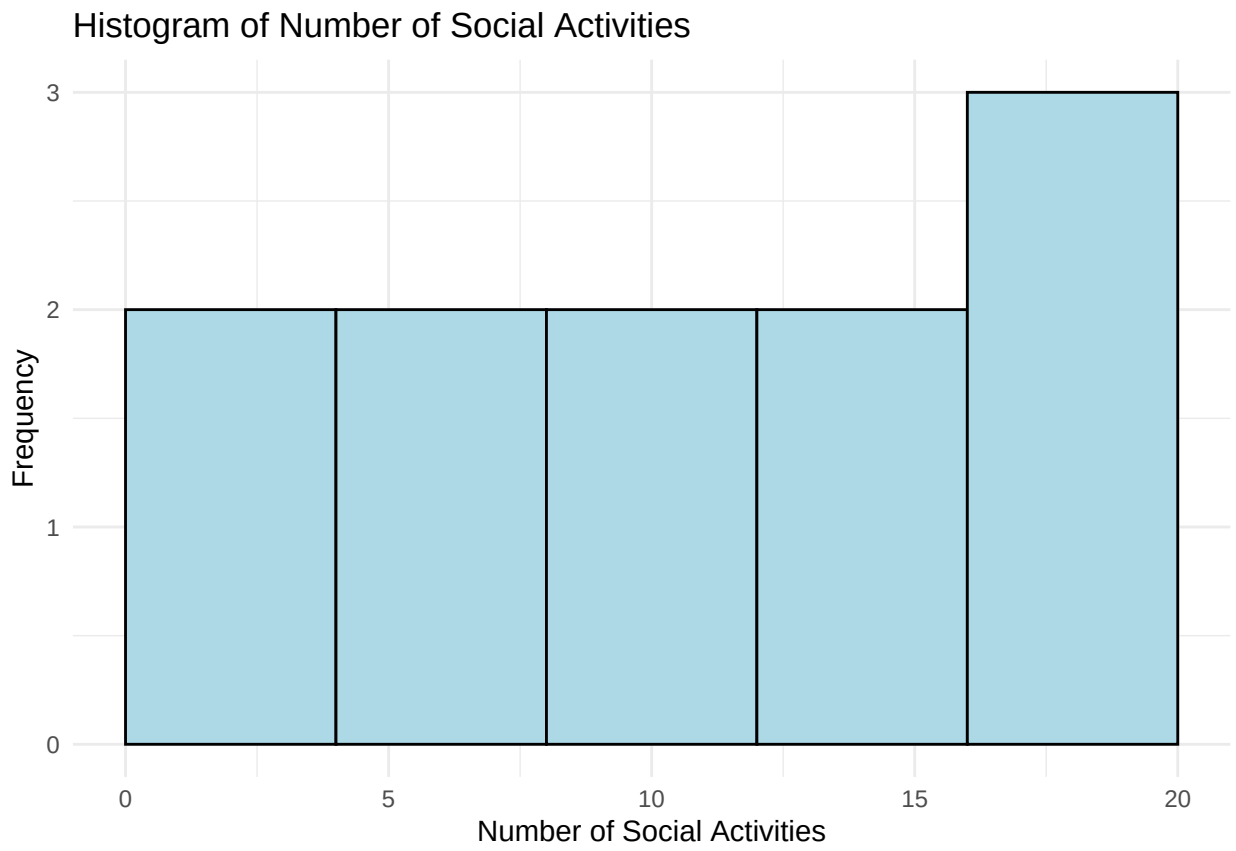
Note: `bind_rows()` in R combines multiple data frames or data-like objects (such as tibbles) by stacking them vertically, i.e., adding rows from one below the other. It automatically matches columns by name, filling in missing values with `NA` if a column is not present in all data frames.

Part (b)

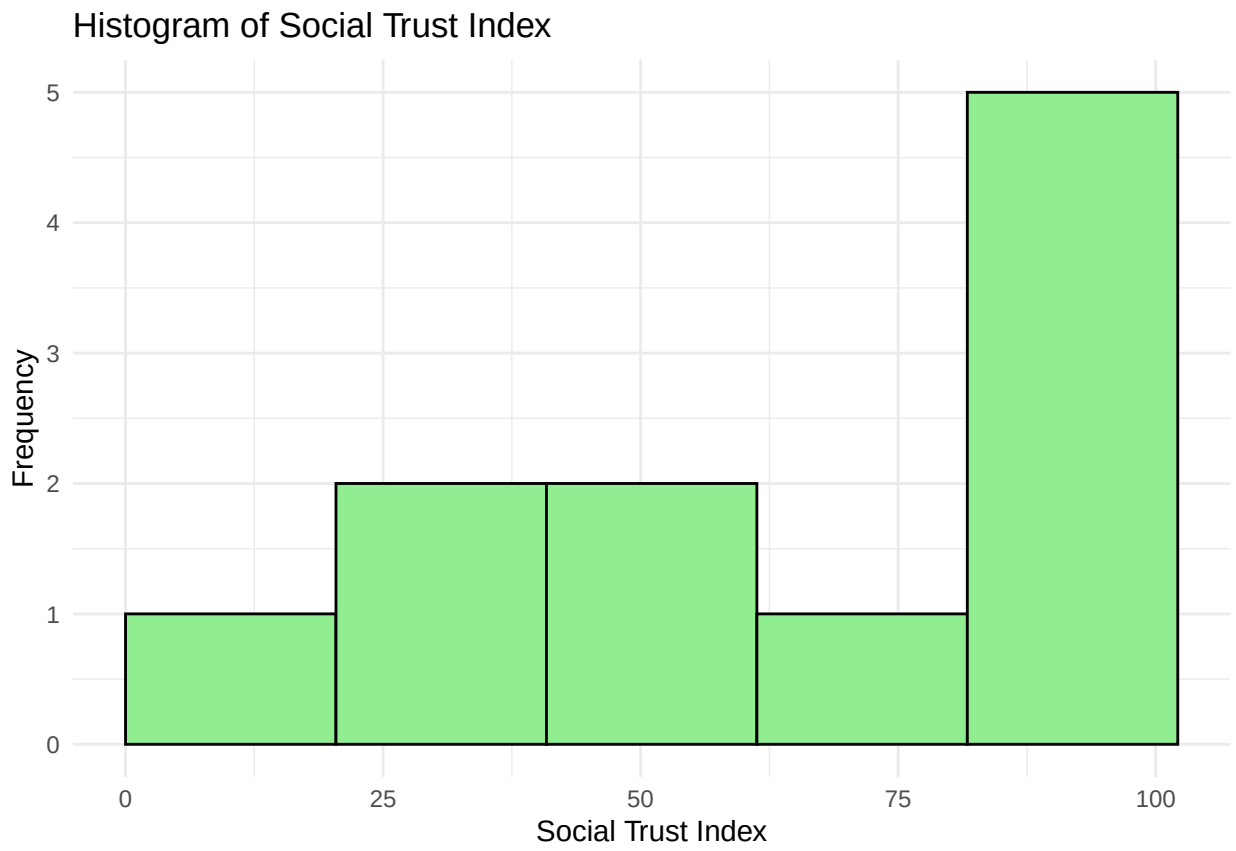
Histogram

We will create a histogram for both the number of social activities per month and the social trust levels with five bins of equal size. We will divide the data range into five equal-sized bins for both variables. We count how many observations fall unto each bin. We use these frequency counts as the values for the y-axis.

```r
# Install ggplot2 if not already installed
# install.packages("ggplot2")

# Load ggplot2
library(ggplot2)

# Manually specify bin boundaries for Number of Social Activities
min_activities <- 0  # Minimum theoretical value
max_activities <- max(df$number_of_social_activities)  # Maximum value from the data
bin_breaks_activities <- seq(min_activities, max_activities, length.out = 6)  # 5 bins require

# Manually specify bin boundaries for Social Trust Index
min_trust <- 0  # Minimum theoretical value
max_trust <- max(df$social_trust_index)  # Maximum value from the data
bin_breaks_trust <- seq(min_trust, max_trust, length.out = 6)  # 5 bins require 6 boundaries


# Histogram for Number of Social Activities with manually specified bins
ggplot(df, aes(x = number_of_social_activities)) +
  geom_histogram(breaks = bin_breaks_activities, fill = "lightblue", color = "black") +
  labs(title = "Histogram of Number of Social Activities",
       x = "Number of Social Activities",
       y = "Frequency") +
  theme_minimal()
```




```
# Histogram for Social Trust Index with manually specified bins
ggplot(df, aes(x = social_trust_index)) +
  geom_histogram(breaks = bin_breaks_trust, fill = "lightgreen", color = "black") +
  labs(title = "Histogram of Social Trust Index",
       x = "Social Trust Index",
       y = "Frequency") +
  theme_minimal()
```



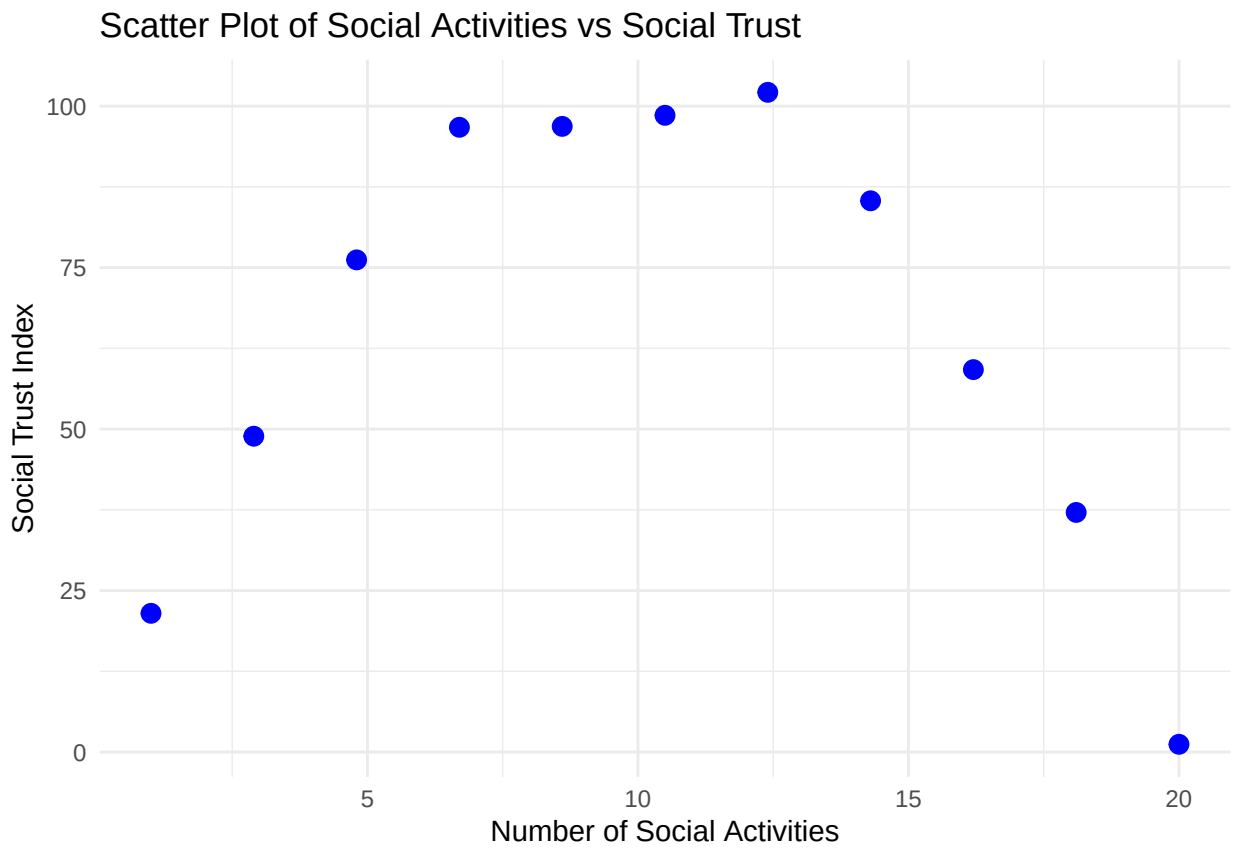
Final Results: - Number of Social Activities: - The histogram shows a roughly symmetric distribution. - The data is concentrated around the mean. - Social Trust Levels: - The histogram shows a skewed distribution, likely left-skewed. - There is a high concentration of data points at higher trust levels, with a tail extending towards lower trust levels.

Part (c)

Scatter Plot

- We will create a scatter plot between the number of social activities and social trust levels.
- We will plot each pair (x_i, y_i) where x_i is the number of social activities and y_i is the social trust level.

```
# Note: Because we already loaded ggplot in a previous code chunk, we dont need to do it again

# Scatter plot using ggplot2
ggplot(df, aes(x = number_of_social_activities, y = social_trust_index)) +
  geom_point(color = "blue", size = 3) +
  labs(title = "Scatter Plot of Social Activities vs Social Trust",
       x = "Number of Social Activities",
       y = "Social Trust Index") +
  theme_minimal()
```



Interpretation: - The scatter plot shows a clear pattern where the social trust level increases with the number of social activities up to a point, and then it appears to decrease. - This suggests a non-linear relationship between the number of social activities and social trust levels.

Part (d)

Hypothesis: There is a non-linear relationship between the number of social activities and social trust levels. Initially, more social activities correlate with higher social trust, but after reaching a peak, further increases in social activities are associated with a decrease in social trust. - Initially, increased social activities may help build community bonds and trust as more interactions occur. - However, beyond a certain threshold, too many social activities could become overwhelming, leading to stress or a feeling of obligatory participation, which might reduce overall trust levels. - This indicates the importance of balancing the number of social activities to maintain optimal levels of social trust within communities.