Data Visualization Assignment

Your Name

Getting Started

This assignment will test your ability to use the tidyverse to perform a complete data analysis workflow: from data wrangling to publication-quality visualization. You will be working with datasets from the psych package, which is commonly used in psychological research.

Your goal for each problem is to write a single, clean dplyr pipe (%>%) that creates the requested data, and then pipe that data directly into a ggplot() call to create a publication-quality plot.

This assignment uses datasets from the `psych` package. For each problem: - Briefly **describe your wrangling** (1–2 sentences). - Complete the r code in the `# TODO` sections. - Use tidy verbs (`filter`, `mutate`, `select`, `summarise`, `pivot_*`). - Choose **appropriate plots** (when asked) that match variable types and questions. - Be sure to polish chart elements! For example, label numeric levels when necessary (e.g., 1 = Male, 2 = Female) - Be sure to comment your code!

Install/Load Required Packages

```r
if (!require(psych)) install.packages("psych")
if (!require(tidyverse)) install.packages("tidyverse")
if (!require(corrplot)) install.packages("corrplot")
if (!require(ggplot2)) install.packages("ggplot2")

library(psych)
library(tidyverse)
library(corrplot)
library(ggplot2)
```

Part 1: Exploring Distributions (bfi Dataset)

The bfi dataset contains 25 personality items from the Big Five Inventory (BFI), along with gender, education, and age.

---

1) Distribution of Agreeableness

Data: ‘psych::bfi

Goal Create a density plot showing the distribution of the mean “Agreeableness” score.

Wrangling: Start with the bfi data. Use mutate() to create a new agreeableness column by calculating the row mean of the five agreeableness items (A1, A2, A3, A4, A5). Make sure to handle missing values (na.rm = TRUE).

Plotting: Pipe the wrangled data into ggplot(). * Use geom_density() to plot the agreeableness distribution.

- Add a fill to the geom (e.g., “skyblue”) and set alpha = 0.7.

- Add appropriate labels using labs() (title, subtitle, x-axis).
- Use theme_minimal().

```
#YOUR CODE HERE
```

---

2) Neuroticism by Gender

Data: ‘psych::bfi

Goal Create a boxplot comparing the distribution of mean “Neuroticism” scores for males and females.

Wrangling: Start with bfi. Create a neuroticism column (mean of N1…N5). Then, filter() out any NA values for gender. Finally, mutate() the gender column into a factor with descriptive labels (e.g., 1 = “Male”, 2 = “Female”).

Plotting: Use geom_boxplot() to map gender to the x-axis and neuroticism to the y-axis.

- Add fill = gender to the aes() mapping.
- Add labs() (e.g., title, subtitle, x=“Gender”, y=“Mean Neuroticism Score”).
- Use theme_minimal() and remove the legend (theme(legend.position = “none”)).

```
#YOUR CODE HERE
```

---

3) Conscientiousness by Education Level

Data: `psych::bfi`

Goal: Create a column chart showing the average “Conscientiousness” score by education level and showing the 95% confidence interval.

Wrangling: Create a conscientiousness column (mean of C1…C5). filter() out NA education levels. group_by() education. summarise() to find the mean_c_score.

Plotting: Use geom_col() to map education to x and mean_c_score to y. (Note: education should be a factor to prevent ggplot from treating it as a continuous number. Use factor(education) in your aes() mapping).

- Add fill = factor(education) to aes().
- Add labs() (title, subtitle, x=“Education Level”, y=“Mean Conscientiousness”).
- Use geom_errorbar() to display the 95% CI for each mean.
- Use scale_x_discrete() to provide meaningful labels for the education levels (1=HS, 2=Finished HS, 3=Some College, 4=College Grad, 5=Grad School).

```
# Example use of scale_x_discrete()
#scale_x_discrete(
#    limits = c("3", "2", "1"), # limits controls order of factors
#    labels = c("Subcompact", "Compact", "Midsize") # create labels
#  )
#YOUR CODE HERE
```

Part 2: Visualizing Relationships (sat.act Dataset)

---

4) SAT Verbal vs. SAT Quantitative

Data: `psych::sat.act`

Goal: Create a scatterplot to visualize the relationship between SAT Verbal (SATV) and SAT Quantitative (SATQ) scores.

Wrangling: None needed, just pipe sat.act directly into ggplot().

Plotting: Use geom_point() to map SATV to x and SATQ to y. Set alpha = 0.5 to handle overplotting.

- Add a linear regression line using geom_smooth.
- Add labs() (e.g., title, subtitle, x="SAT Verbal", y="SAT Quantitative").
- Use theme_bw().

```
#YOUR CODE HERE
```

---

5) Faceting the SAT Relationship by Gender

Data: `psych::sat.act`

Goal: Expand on Problem 4. Does the relationship between SATV and SATQ differ by gender?

Wrangling: mutate() the gender column to a factor with labels (1="Male", 2="Female").

Plotting: Create the same plot as Problem 4 (geom_point + geom_smooth).

- Add color = gender to the aes() in the main ggplot() call.
- Add facet_wrap(~ gender) to create two separate plots.
- Use labs() and theme_bw().

```
#YOUR CODE HERE
```

---

6) ACT Scores by Education

Data: `psych::sat.act`

Goal: Create a bar chart showing the average ACT score by education level, ordered from highest to lowest.

Wrangling: group_by(education), summarise(mean_act = mean(ACT, na.rm = TRUE)). Use mutate() to reorder() the education factor by mean_act (e.g., education = reorder(factor(education), mean_act)).

Plotting: Use geom_col() to map education to x and mean_act to y.

- Add fill = education to the aes() and remove the legend.
- Add labs() and theme_minimal().

```
#YOUR CODE HERE
```

---

7) BFI: Trait Tradeoffs (E vs N)

Data: `psych::bfi`

Goal: Compute mean **Extraversion** and **Openness** and visualize their **bivariate relationship**.
Pick one: **2D density (filled)**, or **scatter + ggMarginal**. Briefly interpret the pattern.

Wrangling: summarize() (e.g., mean of E1…E5).

Plotting: Use geom_density_2d_filled() or ggMarginal().

```
#YOUR CODE HERE
```

---

8) Visualizing Intercorrelations Among Cognitive Ability Tests

Data: ability.cov

Goal: Convert the ability.cov covariance matrix in `$cov` to a correlation matrix using cov2cor() and create a correlogram using `corrplot`.

- The top half of the matrix should represent the correlations using shapes or colors and the bottom half of the matrix should contain the values.
- Change the default colors from the default pallette

```
#YOUR CODE HERE
```

---

9) SAT & ACT: Relationships

Data: `psych::sat.act` (columns typically `SATV`, `SATQ`, `ACT`)

Goal: Tidy column names, then visualize relationships:

Plotting: Set up the aes() to plot `SATV` vs `SATQ` with `ACT` as size.

- Add a smooth trend. Comment on any nonlinearity.
- Add labels
- Add a theme

```
#YOUR CODE HERE
```

---

10) Harman74: Correlogram with Numbers

Data: `psych::Harman74.cor` (correlation matrix + N)

Goal: Plot a **mixed correlogram**: symbols/colors on one triangle and numbers on the other.

```
data(Harman74.cor)
#YOUR CODE HERE
```

Part 3: Reshaping and Comparing (msq and bfi Datasets)

The msq (Motivation State Questionnaire) dataset is in a "wide" format, with different mood scales as columns.

---

11) Tidying the msq Data

Data: `psych::msq`

Goal The msq dataset is wide. Use pivot_longer() to make it tidy.

Wrangling: Use pivot_longer() to gather all the mood scale columns (from Active to Sleepy) into two new columns: mood_scale and score.

Plotting: No plot needed. Just show the head() of your new tidy msq_tidy data frame.

```
#YOUR CODE HERE
```

---

12) Faceted Density of Moods

Data: Your msq_tidy data from #9

Goal: Using your msq_tidy data, create faceted density plots for the "Calm", "Energetic", and "Tense" mood scales.

Wrangling: Start with msq_tidy. filter() so that mood_scale is one of "Calm", "Energetic", or "Tense".

Plotting: Use geom_density(aes(x = score, . . .).

- Add facet_wrap(~ mood_scale).
- Use labs() and theme_minimal(). Remove the legend.

```
#YOUR SCORE HERE
```

---

13) All BFI Traits by Gender

Data: `psych::bfi`

Goal: Create a grouped bar chart comparing the mean scores for all five BFI traits (A, C, E, N, O) grouped by gender.

Wrangling: This will be a multi-step process.

- Start with bfi.
- mutate() to create 5 new mean-score columns (agree, consc, extra, neuro, open).
- select() only the gender column and your 5 new score columns.
- pivot_longer() to gather the 5 score columns into trait and score.
- group_by(gender, trait) and summarise(mean_score = mean(score, na.rm = TRUE)).
- filter() out NA genders and mutate(gender = factor(gender, labels = c("Male", "Female"))).

Plotting:

- ggplot(aes(x = trait, y = mean_score, fill = gender)).
- Use geom_col(position = "dodge").
- Use geom_errorbar() to display your choice of measure of uncertainty.
- Use labs() (title, subtitle, x="Personality Trait", y="Mean Score").
- Use scale_fill_brewer(palette = "Set1") for a nice color palette.
- Use theme_minimal().

```r
#YOUR CODE HERE
```

Part 4: Open Ended

---

14) BFI: Age Trends in Traits

Data: `psych::bfi`

Goal: For any **two traits** of your choice, examine how scores change with **age**.

```r
#YOUR CODE HERE
```

---

15) SAT & ACT: Choose the View

Data: `psych::sat.act`

Goal: Decide on an **appropriate plot** to communicate how **SATQ** relates to **ACT**, conditioned on **SATV** (e.g., faceting SATV bins or using color gradients). Justify your choice.

```r
#YOUR CODE HERE
```

---

16) BFI: Trait Interactions by Gender

Data: `psych::bfi`

Goal: Investigate whether the relationship between two traits (you choose) **differs by gender**.

```r
#YOUR CODE HERE
```

17) "Tell a Data Story"

Use your own data.Ask a meaningful question, **tidy** the data, and produce **1–2 plots** that answer it. Write a short caption (3–5 sentences) explaining your findings and why the charts fit the question.

If you do not have any of your own data, you can pick a dataset from `psych` (`bfi`, `msq`, `msqR`, `sat.act`, `ability.cov`, `Harman74.cor`), but your analysis must be sufficiently different from those covered in this assignment!.

Your "story" should include:

- At least two continuous DVs and two categorical IVs with 2 or more levels
- Publication quality plots of the data distributions (use aes() and/or faceting)
- Publication quality plots of the data summaries with estimates of uncertainty (e.g., geom_col)
- All appropriate labels

```r
# TODO: choose data, tidy, visualize, and caption
```

About the datasets in the psych package

Here's a summary of some prominent datasets within the psych package:

bfi (Big Five Inventory):

- Content: Contains item responses from 2800 participants on 25 personality self-report items designed to measure the Big Five personality traits (Agreeableness, Conscientiousness, Extraversion, Neuroticism, Openness). Also includes demographic variables like gender, education, and age.

epi.bfi:

- Content: Data from 231 participants who completed both the Eysenck Personality Inventory (EPI - measuring Extraversion and Neuroticism) and the Big Five Inventory (bfi items).

sat.act:

- Content: Self-reported scores on the SAT Verbal (SATV), SAT Quantitative (SATQ), and ACT for 700 students. Also includes gender and education level.

How to Explore All Datasets in psych: You can see the full list of datasets available in the currently loaded version of the psych package and access their documentation directly within R:

```r
# Make sure the psych package is loaded
library(psych)

# List all datasets in the psych package
data(package = "psych")

# Get detailed help documentation for a specific dataset (e.g., bfi)
?bfi
help(bfi)
```