EDUC 641 Lab: Applied Statistics in Education and Human Services I

Lab 3: 10/11 and 10/12

## Installing packages, exploring tidyverse/ggplot2, and creating visualizations using geom\_bar()

**Goals:**

1. Install and load packages
2. Use ggplot() to describe one categorical variable
3. Use ggplot() to describe two categorical variable

***Packages* background**

**Tidyverse** is a collection of R packages designed to make data manipulation and analysis more efficient, consistent, and user-friendly. It's not a single package but a suite of packages that work together cohesively.

1. **ggplot2: A powerful package for creating data visualizations and plots.**
2. **dplyr:** Used for data manipulation, including filtering, summarizing, and arranging data.
3. **tidyr:** Designed for tidying and reshaping data, such as converting data from wide to long format.
4. **readr:** A package for efficient reading and writing of data from various file formats.
5. **purrr:** Provides tools for working with functions and vectors.
6. **tibble:** A modern data frame package with improved features over the default R data frames.
7. **stringr:** Offers functions for working with strings and text data.
8. **forcats:** Useful for working with categorical data and factors.

**Worksheet**

**Installing & Loading Packages**

1. *Install* package using [quotes around package name] - do this only once

**install.packages(“package\_name”)** \* # or delete this line of code after it has been run

1. *Load* your package using library() [no quotes] - you have to do this every time, preferably in the beginning of your script

**library(tidyverse)**

**library(here)** [if you use this to import data]

A screenshot of a computer

Description automatically generated

**Try yourself:** Install the package “modelsummary.” We will use this later in the term.

**ggplot2()**

* ggplot2 is one of the packages of the "tidyverse" ecosystem in R, which provides a consistent and efficient way to work with data and create visualizations.
* You do not need to *load* this package if you are using tidyverse, it is already a part of it. 

1. **Describe One Categorical Variable from the cat.csv dataset**
2. Suppose you have a dataset with one categorical variable, let's say "Category," and you want to create a bar chart to visualize the counts of each category.

**ggplot(data = name\_of\_dataset, aes(x = Category)) +**

**geom\_bar() +**

**labs(**

**title = "Bar Chart Example",**

**x = "Categories",**

**y = "Counts") +**

**theme\_minimal()** [this is a function that improves the visualization of your graph]

Note the + and not the %>%

1. Suppose you want to visualize proportions instead of count for a categorical variable.

**ggplot(data = name\_of\_dataset, aes(x = name\_of\_category\_column, y = after\_stat(prop)) +**

**geom\_bar() +**

**labs(**

**title = "Bar Chart Example",**

**x = "Categories",**

**y = "Proportions") +**

**theme\_minimal()**

**Try yourself:** Make a bar chart to represent the counts and proportion of cgender variable from the cat.csv

1. **Describe Two Categorical Variables from the cat.csv dataset**

Let's say you have a dataset with two categorical variables, "Category" and "Group," and you want to create a grouped bar chart to visualize the counts of each category within each group. You can do this as follows:

**ggplot(data = name\_of\_dataset, aes(x = Category, fill = Group)) +**

**geom\_bar(position = "dodge") +**

**labs(title = "Grouped Bar Chart for Two Categorical Variables",**

**x = "Categories",**

**y = "Counts") +**

**theme\_minimal()**

When you set **position = "dodge"**, it means that the geoms will be positioned side by side, allowing you to create grouped or clustered bar charts where each category is divided by another categorical variable.

NO (position = “dodge”) example

A graph with a red and blue rectangle

Description automatically generated

YES (position= “dodge”)

A graph with different colored squares

Description automatically generated with medium confidence

**Try yourself:** Make a grouped bar chart to represent the counts of cgender and absenteeism variables from the cat.csv data in two ways:

1. x = cgender, fill = absenteeism
2. x = absenteeism, fill = cgender

How are they different?