Introduction to the Tidyverse and Data Wrangling

Nancy Carmona

2024-04-25

Session 4. Introduction to the Tidyverse and Data Wrangling - Learning Objectives

You will learn:

- What the "tidyverse" is
- What data "wrangling" is
- Packages and functions used to tidy and wrangle data in R
- How to tidy and wrangle data with R

A) Introduction to the Tidyverse Package

Before we get into data wrangling, let's look at the Tidyverse

```
library(tidyverse)

tidyverse_logo()
```

```
## * __  _    __   .    o           *  .
##  / /_(_)__/ /_ ___  _____ _______ ___
## / __/ / _  / // / |/ / -_) __(_-</ -_)
## \__/_/\_,_/\_, /|___/\__/_/ /___/\__/
##      *  . /___/      o      .       *
```

- "The tidyverse is an opinionated collection of R packages designed for data science."
- "All packages share an underlying design philosophy, grammar, and data structures."
- There are tidyverse packages for data wrangling, modeling, and visualization.

Why would you want to use the tidyverse?

One of the biggest reasons to learn the tidyverse is consistency. Throughout these packages, consistency comes in three primary forms:

1. The first formal argument of tidyverse functions is always a data frame that provides the function's input.
2. The idea of tidy data: a data frame where each row is an observation and each column contains the value of a single variable.
3. The pipe operator, %>%, guides the flow of operations on data. (more on this soon..)

What are “Pipes”?

Pipes let you compose a sequence of function calls in a more readable way. The following two examples of code do the same thing.

```
# First, let's look at the standard functional form in R using nested functions:

print(head(iris))
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
# Using pipes makes this more readable as a sequence of operations:

iris %>% head() %>% print()
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

The pipe supplies the result of the previous function as the first argument to the next function.

Activity 1: Installing tidyverse

```
# Install from CRAN

# install.packages("tidyverse")

# Load the package into your current session

library(tidyverse)
library(readxl)
```

What lives in the tidyverse package?

The tidyverse package actually contains other packages (dplyr,tidyr, ggplot2, etc.) and you’ll see that when you load the tidyverse package using library().

Remember the package must be installed to your device before it can be loaded into your libraries!

Core Packages in Tidyverse

- dplyr provides a grammar of data manipulation
- set of "verbs" that solve most common data manipulation challenges
- tidyr provides a set of functions that help you get to tidy data
- tidy data = every variable goes in a column, and every column is a variable.
- ggplot2 is a system for declaratively creating graphics
- based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.
- Next week!

B) Data Wrangling

What does "data wrangling" mean?

"The process of of getting your data into a useful structured format." This can include *tidying* and transforming. Prepare data for further analysis including: * Summarizing data * Modeling data * Visualizing data

Why would you need to "wrangle" data?

Many datasets, especially if you were involved in the data collection, will have exactly the variables you need in exactly the right format and data type.

But often we import data from the electronic medical record, a database, or the Centers for Disease Control, and the data may not be in quite the format we want.

What is tidy data?

While messy data can be messy in myriad ways, all tidy data follows the same structure, allowing us to easily manipulate and transform our data however we want.

For a dataset to be considered tidy, it needs to follow 3 key rules: 1) Every different variable in our dataset gets a column to itself. 2) Every different observation or object measured in our dataset gets a row to itself. 3) Every different value in our dataset gets its own cell.

C) Tidyverse Functions

Data wrangling with the package *dplyr*

"*dplyr* is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges"

Common functions include:

- Choosing variables based on their name with `select()`
- Choosing rows based on a condition with `filter()`
- Creating new variables or changing old variables with `mutate()`
- Rneaming existing variables with `rename()`

D) Wrangling Demo & Activity

Let's say I am working on a project on physical activity among college students at SFSU. I can use data from the American College Health Association – National College Health Assessment (ACHA-NCHA) which I have saved as an Excel file.

While file types like *.csv and* .tsv are common, it is also common to use Microsoft Excel or an equivalent for data entry. There are a lot of reasons that this is not a great idea, but Excel is so ubiquitous that it is often used for data entry.

Step 1) Install & Load the *readxl* package

```
# install package for reading Excel files

install.packages("readxl")
```

```
## Warning: package 'readxl' is in use and will not be installed
```

```
# load the package "readxl" which is used to import excel documents

library(readxl)
```

Step 2) Once you have your new data file in your "Data" folder you are ready to read the data. Read in the data subset that is stored as an Excel file. This will load the file into your environment.

```
# read the excel file and assign the name "NCHA2021_RWS_Subset" to the new dataset in your environment

NCHA2021_RWS_Subset <- read_excel("~/PINC 2025/Bilaoen-PSP-2025/Onramp/Data/NCHA2021_RWS_Subset.xlsx")
```

```
# OR name it a shorter name when you load it so that it is easy to use for coding!
# ncha_subset <- read_excel("Data/NCHA2021_RWS_Subset.xlsx")
```

What do you see in your Environment pane when you open the dataset?

Step 3) Explore the NCHA dataset! What does your data structure look like?

```
# View the structure of the dataset after it has been imported and saved to your working environment

str(NCHA2021_RWS_Subset)
```

```
## tibble [2,358 x 41] (S3: tbl_df/tbl/data.frame)
##  $ Overall Health
##  $ Leisure time activities - Participating in physical exercise, team sports, recreational sports, o
##  $ Leisure time activities - Socializing with friends
##  $ Leisure time activities - Watching TV, streaming movies/TV, or other media for entertainment
##  $ Leisure time activities - Using social media
##  $ Leisure time activities - Commuting to school and/or to work
##  $ Leisure time activities - Working for pay
##  $ Leisure time activities - Performing unpaid household responsibilities
##  $ Leisure time activities - Taking care of children or other family members (unpaid)
##  $ Self described weight
##  $ Minutes Moderate PA - minutes
##  $ Minutes Vigorous PA - minutes
##  $ Days Strengthening Exercises
##  $ Last 7 days, usual sugar-sweetened beverages per day - servings
##  $ Last 7 days, usual fruit servings per day
##  $ In the last week, usual vegetable servings per day
##  $ Substances ever used - Tobacco or nicotine delivery products (cigarettes, e-cigarettes, Juul or o
##  $ Substances ever used - Alcoholic beverages (beer, wine, liquor, etc.)
##  $ Substances ever used - Cannabis (marijuana, weed, hash, edibles, vaped cannabis, etc.) [nonmedica
##  $ Last 3 months: frequency of substances used - Tobacco or nicotine delivery products (cigarettes,
##  $ Last 3 months: frequency of substances used - Alcoholic beverages (beer, wine, liquor, etc.)
##  $ Last 3 months: frequency of substances used - Cannabis (marijuana, weed, hash, edibles, vaped cann
##  $ Gender identity - Selected Choice
##  $ Age in years - Years
##  $ How do you usually describe yourself? - Selected Choice American Indian or Native Alaskan
##  $ How do you usually describe yourself? - Selected Choice Asian or Asian American
##  $ How do you usually describe yourself? - Selected Choice Black or African American
##  $ How do you usually describe yourself? - Selected Choice Hispanic or Latino/a
##  $ How do you usually describe yourself? - Selected Choice Arab/Middle Eastern/North African Origin
##  $ How do you usually describe yourself? - Selected Choice Native Hawaiian or Other Pacific Islander
##  $ How do you usually describe yourself? - Selected Choice White
##  $ How do you usually describe yourself? - Selected Choice Biracial or Multiracial
##  $ Are you: - Selected Choice Mexican, Mexican American, Chicano
##  $ Are you: - Selected Choice Puerto Rican
##  $ Are you: - Selected Choice Cuban
##  $ Are you: - Selected Choice Another Hispanic, Latino/a/x, or Spanish origin (please specify)
##  $ Are you: - Selected Choice East Asian (for example: Chinese, Japanese, or Korean)
##  $ Are you: - Selected Choice Southeast Asian (for example: Cambodian, Vietnamese, Hmong, or Filipin
##  $ Are you: - Selected Choice South Asian (for example: Indian, Pakistani, Nepalese, or Sri Lankan)
##  $ Are you: - Selected Choice Other Asian (please specify)
##  $ Approximate GPA
```

Step 4) Does the dataset need to be changed to a different type? Yes, a "data.frame" will be easier to wrangle than a "tibble"

To continue our wrangling, we should transform our "tibble" into a "data.frame"

```r
# use the base function "data.frame" to transform a "tibble" into a "data.frame" class

ncha_df <- data.frame(NCHA2021_RWS_Subset)
```

Step 5) Check that the dataset is now a "data.frame" using the function `class()`

```r
# use the base function "class" to make sure your transformation worked

class(ncha_df)
```

```
## [1] "data.frame"
```

Step 6) Continue exploring the new data frame object. In addition to `str()` and `class()` we can explore the data with `dim()` and `head()`:

```r
# View the dimensions of the new data frame

dim(ncha_df)
```

```
## [1] 2358   41
```

```r
# View the first 6 rows of the new data frame

head(ncha_df)
```

```
##   Overall.Health
## 1      Excellent
## 2           Good
## 3           Good
## 4           Good
## 5           Good
## 6      Very Good
##   Leisure.time.activities...Participating.in.physical.exercise..team.sports..recreational.sports..or
## 1
## 2
## 3
## 4
## 5
## 6
##   Leisure.time.activities...Socializing.with.friends
## 1                                          1-5 hours
## 2                                          1-5 hours
## 3                                        11-15 hours
## 4                                          1-5 hours
## 5                                            0 hours
## 6                                          1-5 hours
##   Leisure.time.activities...Watching.TV..streaming.movies.TV..or.other.media.for.entertainment
## 1                                                                                   6-10 hours
## 2                                                                                    1-5 hours
## 3                                                                                  26-30 hours
## 4                                                                                    1-5 hours
## 5                                                                                  21-25 hours
## 6                                                                                    1-5 hours
##   Leisure.time.activities...Using.social.media
## 1                                   6-10 hours
## 2                                   6-10 hours
## 3                                  11-15 hours
## 4                                    1-5 hours
## 5                                  21-25 hours
## 6                                   6-10 hours
##   Leisure.time.activities...Commuting.to.school.and.or.to.work
## 1                                                      0 hours
## 2                                                      0 hours
## 3                                                      0 hours
## 4                                                   6-10 hours
## 5                                                      0 hours
## 6                                                    1-5 hours
##   Leisure.time.activities...Working.for.pay
## 1                               21-25 hours
## 2                                6-10 hours
## 3                               16-20 hours
## 4                        More than 30 hours
## 5                               11-15 hours
## 6                        More than 30 hours
##   Leisure.time.activities...Performing.unpaid.household.responsibilities
## 1                                                             6-10 hours
## 2                                                             6-10 hours
## 3                                                            11-15 hours
## 4                                                             6-10 hours
## 5                                                            16-20 hours
## 6                                                                0 hours
##   Leisure.time.activities...Taking.care.of.children.or.other.family.members..unpaid.
## 1                                                                            0 hours
## 2                                                                          1-5 hours
## 3                                                                            0 hours
## 4                                                                 More than 30 hours
## 5                                                                            0 hours
## 6                                                                            0 hours
##    Self.described.weight Minutes.Moderate.PA...minutes
## 1 About the right weight                             8
## 2       Very underweight                            90
## 3 About the right weight                           400
## 4    Slightly overweight                           360
## 5 About the right weight                            20
## 6 About the right weight                            25
##   Minutes.Vigorous.PA...minutes Days.Strengthening.Exercises
## 1                             2                       2 days
## 2                            90                       0 days
## 3                            60                       0 days
## 4                           160                       3 days
## 5                             0                       0 days
## 6                            25                       0 days
##   Last.7.days..usual.sugar.sweetened.beverages.per.day...servings
## 1                                                               2
## 2                                                               6
## 3                                                              10
## 4                                                               0
## 5                                                               1
## 6                                                              12
##   Last.7.days..usual.fruit.servings.per.day
## 1                      3-4 servings per day
## 2                        0 servings per day
## 3                      1-2 servings per day
## 4                      5-6 servings per day
## 5                        0 servings per day
## 6                      1-2 servings per day
##   In.the.last.week..usual.vegetable.servings.per.day
## 1                               1-2 servings per day
## 2                               1-2 servings per day
## 3                               1-2 servings per day
## 4                               1-2 servings per day
## 5                               1-2 servings per day
## 6                               1-2 servings per day
##   Substances.ever.used...Tobacco.or.nicotine.delivery.products..cigarettes..e.cigarettes..Juul.or.oth
## 1
## 2
## 3
## 4
## 5
## 6
##   Substances.ever.used...Alcoholic.beverages..beer..wine..liquor..etc..
## 1                                                                   Yes
## 2                                                                    No
## 3                                                                   Yes
## 4                                                                    No
## 5                                                                   Yes
## 6                                                                   Yes
##   Substances.ever.used...Cannabis..marijuana..weed..hash..edibles..vaped.cannabis..etc....nonmedical
## 1
## 2
## 3
## 4
## 5
## 6
##   Last.3.months..frequency.of.substances.used...Tobacco.or.nicotine.delivery.products..cigarettes..e
## 1
## 2
## 3
## 4
## 5
## 6
##   Last.3.months..frequency.of.substances.used...Alcoholic.beverages..beer..wine..liquor..etc..
## 1                                                                                Once or twice
## 2                                                                                         <NA>
## 3                                                                                Once or twice
## 4                                                                                         <NA>
## 5                                                                                      Monthly
## 6                                                                                       Weekly
##   Last.3.months..frequency.of.substances.used...Cannabis..marijuana..weed..hash..edibles..vaped.canna
## 1
## 2
## 3
## 4
## 5
## 6
##   Gender.identity...Selected.Choice Age.in.years...Years
## 1                               Man                   39
## 2                             Woman                   32
## 3                             Woman                   27
## 4                             Woman                   21
## 5                             Woman                   21
## 6                               Man                   26
##   How.do.you.usually.describe.yourself....Selected.Choice.American.Indian.or.Native.Alaskan
## 1                                                                              Not selected
## 2                                                                              Not selected
## 3                                                                              Not selected
## 4                                                                              Not selected
## 5                                                                              Not selected
## 6                                                                              Not selected
##   How.do.you.usually.describe.yourself....Selected.Choice.Asian.or.Asian.American
## 1                                                                        Selected
## 2                                                                        Selected
## 3                                                                    Not selected
## 4                                                                    Not selected
## 5                                                                    Not selected
## 6                                                                    Not selected
##   How.do.you.usually.describe.yourself....Selected.Choice.Black.or.African.American
## 1                                                                      Not selected
## 2                                                                      Not selected
## 3                                                                      Not selected
## 4                                                                      Not selected
## 5                                                                      Not selected
## 6                                                                      Not selected
##   How.do.you.usually.describe.yourself....Selected.Choice.Hispanic.or.Latino.a
## 1                                                                 Not selected
## 2                                                                 Not selected
## 3                                                                 Not selected
## 4                                                                 Not selected
## 5                                                                 Not selected
## 6                                                                 Not selected
##   How.do.you.usually.describe.yourself....Selected.Choice.Arab.Middle.Eastern.North.African.Origin
## 1                                                                                     Not selected
## 2                                                                                     Not selected
## 3                                                                                     Not selected
## 4                                                                                         Selected
## 5                                                                                     Not selected
## 6                                                                                     Not selected
##   How.do.you.usually.describe.yourself....Selected.Choice.Native.Hawaiian.or.Other.Pacific.Islander.
## 1                                                                                             Not sel
## 2                                                                                             Not sel
## 3                                                                                             Not sel
## 4                                                                                             Not sel
## 5                                                                                             Not sel
## 6                                                                                             Not sel
##   How.do.you.usually.describe.yourself....Selected.Choice.White
## 1                                                  Not selected
## 2                                                  Not selected
## 3                                                      Selected
## 4                                                  Not selected
## 5                                                      Selected
## 6                                                      Selected
##   How.do.you.usually.describe.yourself....Selected.Choice.Biracial.or.Multiracial
## 1                                                                    Not selected
## 2                                                                    Not selected
## 3                                                                    Not selected
## 4                                                                    Not selected
## 5                                                                    Not selected
## 6                                                                    Not selected
##   Are.you....Selected.Choice.Mexican..Mexican.American..Chicano
## 1                                                          <NA>
## 2                                                          <NA>
## 3                                                          <NA>
## 4                                                          <NA>
## 5                                                          <NA>
## 6                                                          <NA>
##   Are.you....Selected.Choice.Puerto.Rican Are.you....Selected.Choice.Cuban
## 1                                    <NA>                             <NA>
## 2                                    <NA>                             <NA>
## 3                                    <NA>                             <NA>
## 4                                    <NA>                             <NA>
## 5                                    <NA>                             <NA>
## 6                                    <NA>                             <NA>
##   Are.you....Selected.Choice.Another.Hispanic..Latino.a.x..or.Spanish.origin..please.specify.
## 1                                                                                        <NA>
## 2                                                                                        <NA>
## 3                                                                                        <NA>
## 4                                                                                        <NA>
## 5                                                                                        <NA>
## 6                                                                                        <NA>
##   Are.you....Selected.Choice.East.Asian..for.example..Chinese..Japanese..or.Korean.
## 1                                                                          Selected
## 2                                                                      Not selected
## 3                                                                              <NA>
## 4                                                                              <NA>
## 5                                                                              <NA>
## 6                                                                              <NA>
##   Are.you....Selected.Choice.Southeast.Asian..for.example..Cambodian..Vietnamese..Hmong..or.Filipino
## 1                                                                                        Not selecte
## 2                                                                                            Selecte
## 3                                                                                                <NA
## 4                                                                                                <NA
## 5                                                                                                <NA
## 6                                                                                                <NA
##   Are.you....Selected.Choice.South.Asian..for.example..Indian..Pakistani..Nepalese..or.Sri.Lankan.
## 1                                                                                     Not selected
## 2                                                                                     Not selected
## 3                                                                                             <NA>
## 4                                                                                             <NA>
## 5                                                                                             <NA>
## 6                                                                                             <NA>
##   Are.you....Selected.Choice.Other.Asian..please.specify. Approximate.GPA
## 1                                            Not selected              A-
## 2                                            Not selected               D
## 3                                                    <NA>              B+
## 4                                                    <NA>               B
## 5                                                    <NA>              A-
## 6                                                    <NA>               C
```

Step 7) Do the variable names make sense? They seem a bit cumbersome…

```
# Using the function "names()" we can view a list of all of the variable names within the entire data f
names(ncha_df)
```

```
##  [1] "Overall.Health"
##  [2] "Leisure.time.activities...Participating.in.physical.exercise..team.sports..recreational.sports
##  [3] "Leisure.time.activities...Socializing.with.friends"
##  [4] "Leisure.time.activities...Watching.TV..streaming.movies.TV..or.other.media.for.entertainment"
##  [5] "Leisure.time.activities...Using.social.media"
##  [6] "Leisure.time.activities...Commuting.to.school.and.or.to.work"
##  [7] "Leisure.time.activities...Working.for.pay"
##  [8] "Leisure.time.activities...Performing.unpaid.household.responsibilities"
##  [9] "Leisure.time.activities...Taking.care.of.children.or.other.family.members..unpaid."
## [10] "Self.described.weight"
## [11] "Minutes.Moderate.PA...minutes"
## [12] "Minutes.Vigorous.PA...minutes"
## [13] "Days.Strengthening.Exercises"
## [14] "Last.7.days..usual.sugar.sweetened.beverages.per.day...servings"
## [15] "Last.7.days..usual.fruit.servings.per.day"
## [16] "In.the.last.week..usual.vegetable.servings.per.day"
## [17] "Substances.ever.used...Tobacco.or.nicotine.delivery.products..cigarettes..e.cigarettes..Juul.or
## [18] "Substances.ever.used...Alcoholic.beverages..beer..wine..liquor..etc.."
## [19] "Substances.ever.used...Cannabis..marijuana..weed..hash..edibles..vaped.cannabis..etc....nonmedi
## [20] "Last.3.months..frequency.of.substances.used...Tobacco.or.nicotine.delivery.products..cigarettes
## [21] "Last.3.months..frequency.of.substances.used...Alcoholic.beverages..beer..wine..liquor..etc.."
## [22] "Last.3.months..frequency.of.substances.used...Cannabis..marijuana..weed..hash..edibles..vaped.c
## [23] "Gender.identity...Selected.Choice"
## [24] "Age.in.years...Years"
## [25] "How.do.you.usually.describe.yourself....Selected.Choice.American.Indian.or.Native.Alaskan"
## [26] "How.do.you.usually.describe.yourself....Selected.Choice.Asian.or.Asian.American"
## [27] "How.do.you.usually.describe.yourself....Selected.Choice.Black.or.African.American"
## [28] "How.do.you.usually.describe.yourself....Selected.Choice.Hispanic.or.Latino.a"
## [29] "How.do.you.usually.describe.yourself....Selected.Choice.Arab.Middle.Eastern.North.African.Origi
## [30] "How.do.you.usually.describe.yourself....Selected.Choice.Native.Hawaiian.or.Other.Pacific.Island
## [31] "How.do.you.usually.describe.yourself....Selected.Choice.White"
## [32] "How.do.you.usually.describe.yourself....Selected.Choice.Biracial.or.Multiracial"
## [33] "Are.you....Selected.Choice.Mexican..Mexican.American..Chicano"
## [34] "Are.you....Selected.Choice.Puerto.Rican"
## [35] "Are.you....Selected.Choice.Cuban"
## [36] "Are.you....Selected.Choice.Another.Hispanic..Latino.a.x..or.Spanish.origin..please.specify."
## [37] "Are.you....Selected.Choice.East.Asian..for.example..Chinese..Japanese..or.Korean."
## [38] "Are.you....Selected.Choice.Southeast.Asian..for.example..Cambodian..Vietnamese..Hmong..or.Filip
## [39] "Are.you....Selected.Choice.South.Asian..for.example..Indian..Pakistani..Nepalese..or.Sri.Lankan
## [40] "Are.you....Selected.Choice.Other.Asian..please.specify."
## [41] "Approximate.GPA"
```

Step 8) Let's rename the variables to make it easier to code. The simplified names will improve our workflow if we are not typing out complex names that require us to use single quotes when variables have a space.

```r
# rename variables with a simpler name using "rename()"

ncha_df <- rename(ncha_df,
    "leisure_sports" = "Leisure.time.activities...Participating.in.physical.exercise..team.sports..recre


# We can also rename multiple variables at the same time

ncha_df <- rename(ncha_df,
                              "overall_health" = "Overall.Health",
                              "exercise_mod_min" = "Minutes.Moderate.PA...minutes",
                              "exercise_vig_min" = "Minutes.Vigorous.PA...minutes")
```

Step 9) Let's check the variable names and see if our transformation worked.

```r
# view names of variables in the data frame

names(ncha_df)
```

```
##  [1] "overall_health"
##  [2] "leisure_sports"
##  [3] "Leisure.time.activities...Socializing.with.friends"
##  [4] "Leisure.time.activities...Watching.TV..streaming.movies.TV..or.other.media.for.entertainment"
##  [5] "Leisure.time.activities...Using.social.media"
##  [6] "Leisure.time.activities...Commuting.to.school.and.or.to.work"
##  [7] "Leisure.time.activities...Working.for.pay"
##  [8] "Leisure.time.activities...Performing.unpaid.household.responsibilities"
##  [9] "Leisure.time.activities...Taking.care.of.children.or.other.family.members..unpaid."
## [10] "Self.described.weight"
## [11] "exercise_mod_min"
## [12] "exercise_vig_min"
## [13] "Days.Strengthening.Exercises"
## [14] "Last.7.days..usual.sugar.sweetened.beverages.per.day...servings"
## [15] "Last.7.days..usual.fruit.servings.per.day"
## [16] "In.the.last.week..usual.vegetable.servings.per.day"
## [17] "Substances.ever.used...Tobacco.or.nicotine.delivery.products..cigarettes..e.cigarettes..Juul.o
## [18] "Substances.ever.used...Alcoholic.beverages..beer..wine..liquor..etc.."
## [19] "Substances.ever.used...Cannabis..marijuana..weed..hash..edibles..vaped.cannabis..etc....nonmedi
## [20] "Last.3.months..frequency.of.substances.used...Tobacco.or.nicotine.delivery.products..cigarette
## [21] "Last.3.months..frequency.of.substances.used...Alcoholic.beverages..beer..wine..liquor..etc.."
## [22] "Last.3.months..frequency.of.substances.used...Cannabis..marijuana..weed..hash..edibles..vaped.
## [23] "Gender.identity...Selected.Choice"
## [24] "Age.in.years...Years"
## [25] "How.do.you.usually.describe.yourself....Selected.Choice.American.Indian.or.Native.Alaskan"
## [26] "How.do.you.usually.describe.yourself....Selected.Choice.Asian.or.Asian.American"
## [27] "How.do.you.usually.describe.yourself....Selected.Choice.Black.or.African.American"
## [28] "How.do.you.usually.describe.yourself....Selected.Choice.Hispanic.or.Latino.a"
## [29] "How.do.you.usually.describe.yourself....Selected.Choice.Arab.Middle.Eastern.North.African.Origi
## [30] "How.do.you.usually.describe.yourself....Selected.Choice.Native.Hawaiian.or.Other.Pacific.Islan
## [31] "How.do.you.usually.describe.yourself....Selected.Choice.White"
## [32] "How.do.you.usually.describe.yourself....Selected.Choice.Biracial.or.Multiracial"
## [33] "Are.you....Selected.Choice.Mexican..Mexican.American..Chicano"
## [34] "Are.you....Selected.Choice.Puerto.Rican"
## [35] "Are.you....Selected.Choice.Cuban"
## [36] "Are.you....Selected.Choice.Another.Hispanic..Latino.a.x..or.Spanish.origin..please.specify."
## [37] "Are.you....Selected.Choice.East.Asian..for.example..Chinese..Japanese..or.Korean."
## [38] "Are.you....Selected.Choice.Southeast.Asian..for.example..Cambodian..Vietnamese..Hmong..or.Fili
## [39] "Are.you....Selected.Choice.South.Asian..for.example..Indian..Pakistani..Nepalese..or.Sri.Lanka
## [40] "Are.you....Selected.Choice.Other.Asian..please.specify."
## [41] "Approximate.GPA"
```

```r
# Or View the entire data frame

View(ncha_df)
```

Yes, it is a much more tidy name!

Step 10) Let’s use the function `select()` to only keep the columns (variables) we are interested in analyzing.

```
# create a data subset by selecting variables with cleaned up names
# we can use pipes [%>%] (shortcut for pipes = command + shift + M)

ncha_df_sub <- ncha_df %>% select(overall_health, leisure_sports, exercise_mod_min, exercise_vig_min)
```

Step 11) Let's use the function `filter()` to only keep rows (observations) for students who self-reported 0 hours of leisure sports.

```
# select observations for students who self-reported 0 hours of leisure sports
ncha_df_sub <- filter(ncha_df_sub, leisure_sports == "0 hours")

# we can look at a table of two variables, to see how they overlap
# check what the overlap is between overall health categories and minutes of leisure sports
table(ncha_df_sub$leisure_sports, ncha_df_sub$overall_health)
```

```
##
##             Excellent Fair Good Poor Very Good
##   0 hours          48  147  249   26       190
```

Step 11) Let's create a new variable using `mutate()`

```
# view a summary of the observations for "exercise_mod_min"
summary(ncha_df_sub$exercise_mod_min)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    45.0   115.0   225.2   210.0  5040.0       8
```

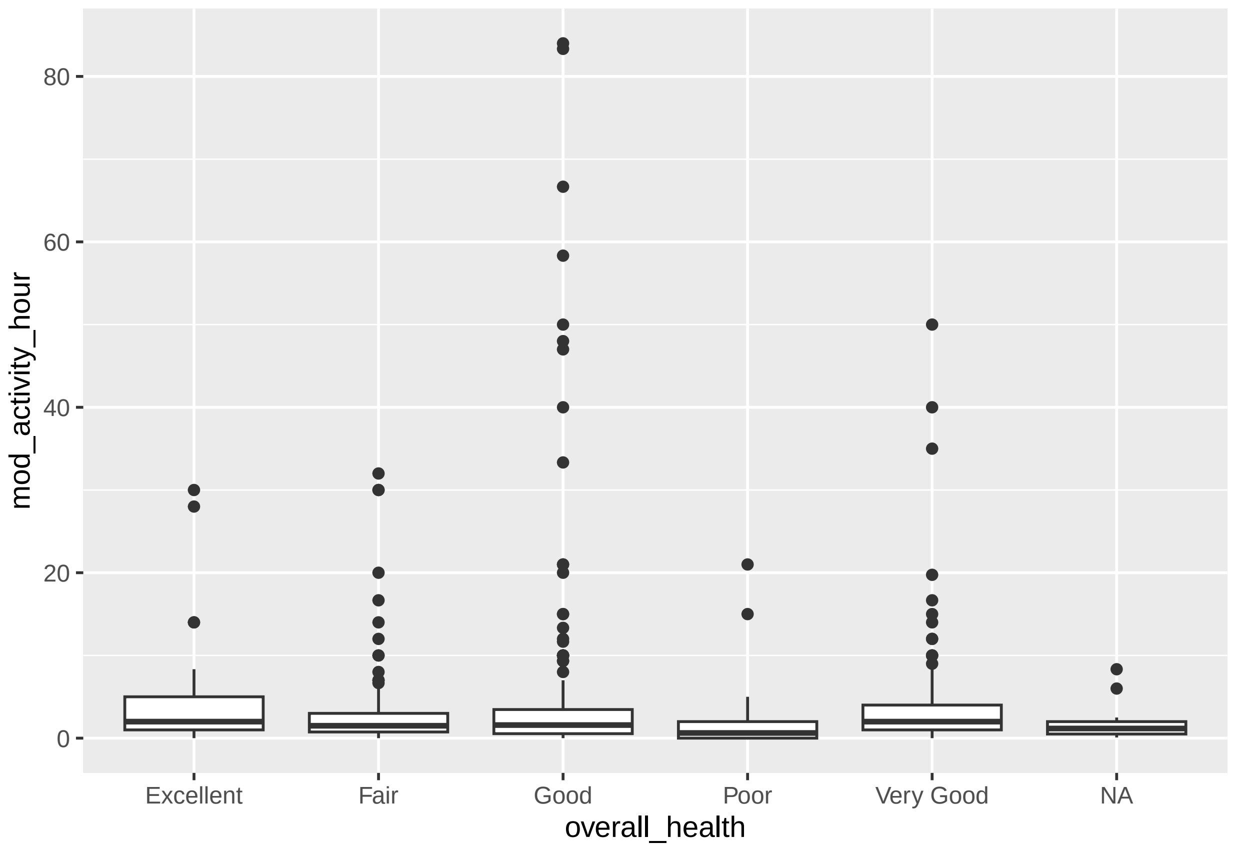
```
# make a new variable that will be part of the existing data frame
# we can make a variable for the number of hours of activity

ncha_df_sub <- ncha_df_sub %>% mutate(mod_activity_hour = exercise_mod_min / 60 )

# check what the updated data frame looks like

View(ncha_df_sub)
```

Step 12) Let's create a simple plot to visualize our exercise data.

We can make a bivariate plot to look at the relationship between overall health and hours of moderate physical activity.

```
# make a simple plot can be made using ggplot2 package

ggplot(data = ncha_df_sub, aes (x = overall_health, y = mod_activity_hour, colour) ) +
    geom_boxplot()
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```

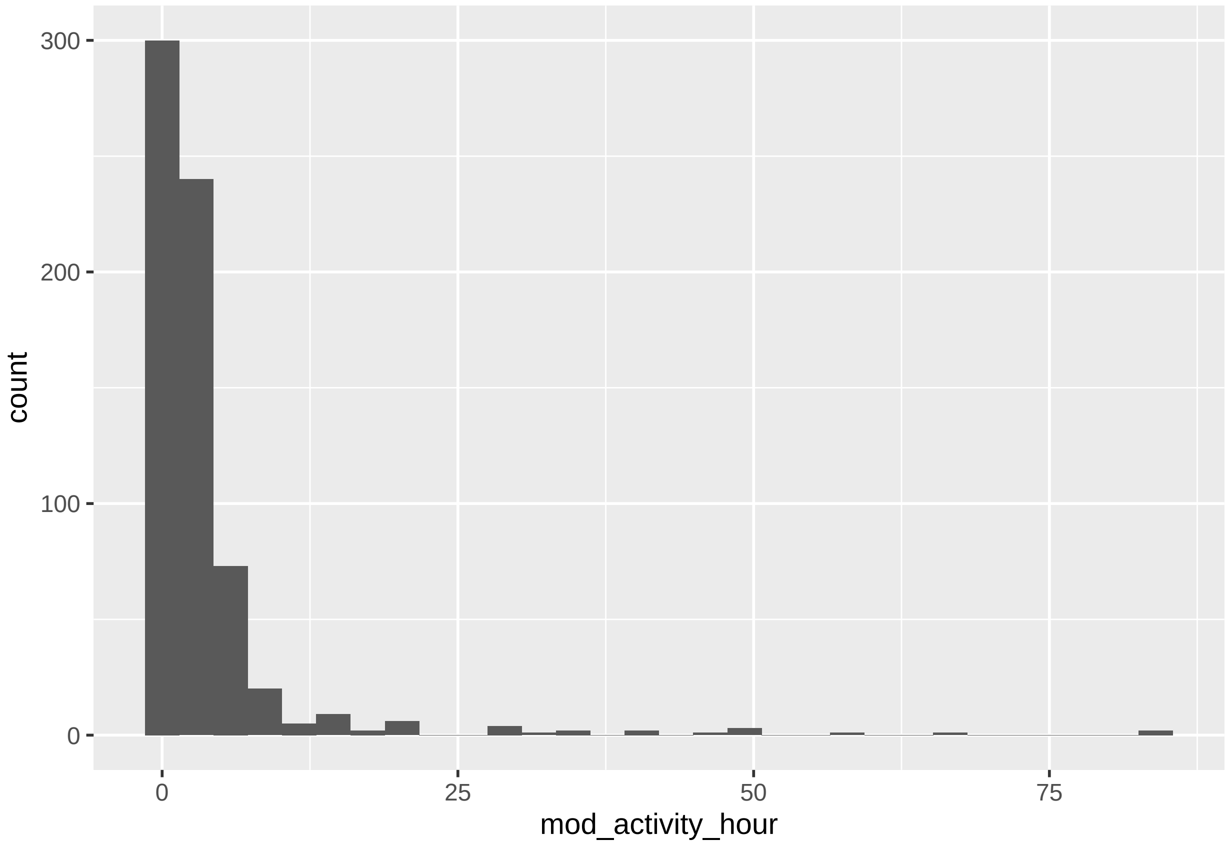


```
# OR you can use pipes instead of [+] symbol

# ggplot(ncha_df_sub, aes(overall_health, mod_activity_hour) ) %>% geom_boxplot()
```

We can also make univariate plots to look the distribution of our data.

```
# make a simple histogram plot using ggplot2 package

ggplot(ncha_df_sub, aes(mod_activity_hour) ) + geom_histogram()
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



E) Activity: Data Wrangle using Tidyverse!

- Start with a clean version of the data!

```
# load the original data subset

NCHA2021_RWS_Subset <- read_excel("Data/NCHA2021_RWS_Subset.xlsx")

# make the tibble into a data frame
library(tidyverse)
ncha_df <- data.frame(NCHA2021_RWS_Subset)
```

- Check that your data import went well – typically done in an "exploratory analysis" – should always do this!
- What class is the data? Character
- How many rows? 2358
- How many columns? 41
- How many observations? 2358 (?)

```
# Check that your data import worked as intended
glimpse(ncha_df)
```

```
## Rows: 2,358
## Columns: 41
## $ Overall.Health
## $ Leisure.time.activities...Participating.in.physical.exercise..team.sports..recreational.sports..or
## $ Leisure.time.activities...Socializing.with.friends
## $ Leisure.time.activities...Watching.TV..streaming.movies.TV..or.other.media.for.entertainment
## $ Leisure.time.activities...Using.social.media
## $ Leisure.time.activities...Commuting.to.school.and.or.to.work
## $ Leisure.time.activities...Working.for.pay
## $ Leisure.time.activities...Performing.unpaid.household.responsibilities
## $ Leisure.time.activities...Taking.care.of.children.or.other.family.members..unpaid.
## $ Self.described.weight
## $ Minutes.Moderate.PA...minutes
## $ Minutes.Vigorous.PA...minutes
## $ Days.Strengthening.Exercises
## $ Last.7.days..usual.sugar.sweetened.beverages.per.day...servings
## $ Last.7.days..usual.fruit.servings.per.day
## $ In.the.last.week..usual.vegetable.servings.per.day
## $ Substances.ever.used...Tobacco.or.nicotine.delivery.products..cigarettes..e.cigarettes..Juul.or.oth
## $ Substances.ever.used...Alcoholic.beverages..beer..wine..liquor..etc..
## $ Substances.ever.used...Cannabis..marijuana..weed..hash..edibles..vaped.cannabis..etc....nonmedical
## $ Last.3.months..frequency.of.substances.used...Tobacco.or.nicotine.delivery.products..cigarettes..e
## $ Last.3.months..frequency.of.substances.used...Alcoholic.beverages..beer..wine..liquor..etc..
## $ Last.3.months..frequency.of.substances.used...Cannabis..marijuana..weed..hash..edibles..vaped.canna
## $ Gender.identity...Selected.Choice
## $ Age.in.years...Years
## $ How.do.you.usually.describe.yourself....Selected.Choice.American.Indian.or.Native.Alaskan
## $ How.do.you.usually.describe.yourself....Selected.Choice.Asian.or.Asian.American
## $ How.do.you.usually.describe.yourself....Selected.Choice.Black.or.African.American
## $ How.do.you.usually.describe.yourself....Selected.Choice.Hispanic.or.Latino.a
## $ How.do.you.usually.describe.yourself....Selected.Choice.Arab.Middle.Eastern.North.African.Origin
## $ How.do.you.usually.describe.yourself....Selected.Choice.Native.Hawaiian.or.Other.Pacific.Islander.
## $ How.do.you.usually.describe.yourself....Selected.Choice.White
## $ How.do.you.usually.describe.yourself....Selected.Choice.Biracial.or.Multiracial
## $ Are.you....Selected.Choice.Mexican..Mexican.American..Chicano
## $ Are.you....Selected.Choice.Puerto.Rican
## $ Are.you....Selected.Choice.Cuban
## $ Are.you....Selected.Choice.Another.Hispanic..Latino.a.x..or.Spanish.origin..please.specify.
## $ Are.you....Selected.Choice.East.Asian..for.example..Chinese..Japanese..or.Korean.
## $ Are.you....Selected.Choice.Southeast.Asian..for.example..Cambodian..Vietnamese..Hmong..or.Filipino
## $ Are.you....Selected.Choice.South.Asian..for.example..Indian..Pakistani..Nepalese..or.Sri.Lankan.
## $ Are.you....Selected.Choice.Other.Asian..please.specify.
## $ Approximate.GPA
```

- Rename the variables: "Overall.Health", "Minutes.Moderate.PA..minutes", "Minutes.Vigorous.PA..minutes"

```
# We can rename one variable at a time, OR we can also rename multiple variables at the same time
ncha_df <- rename(ncha_df,
                  "overall_health" = "Overall.Health",
                  "exercise_mod_min" = "Minutes.Moderate.PA...minutes",
                  "exercise_vig_min" = "Minutes.Vigorous.PA...minutes")
```

- Select the newly renamed variables: "overall_health", "exercise_mod_min", "exercise_vig_min"

```
# Select the variables using pipes [ %>% ]
ncha_df_sub <- ncha_df %>% select(overall_health, exercise_mod_min, exercise_vig_min)
```

- Check to see what our new smaller data frame looks like!

```
# View the data frame in a pop up window
view(ncha_df_sub)
```

- Filter the dataset to only keep rows for students with self-reported "Excellent" overall health.

```
# Only keep observations for students who self-report "Excellent" health
ncha_df_sub <- filter(ncha_df_sub, overall_health == "Excellent")
```

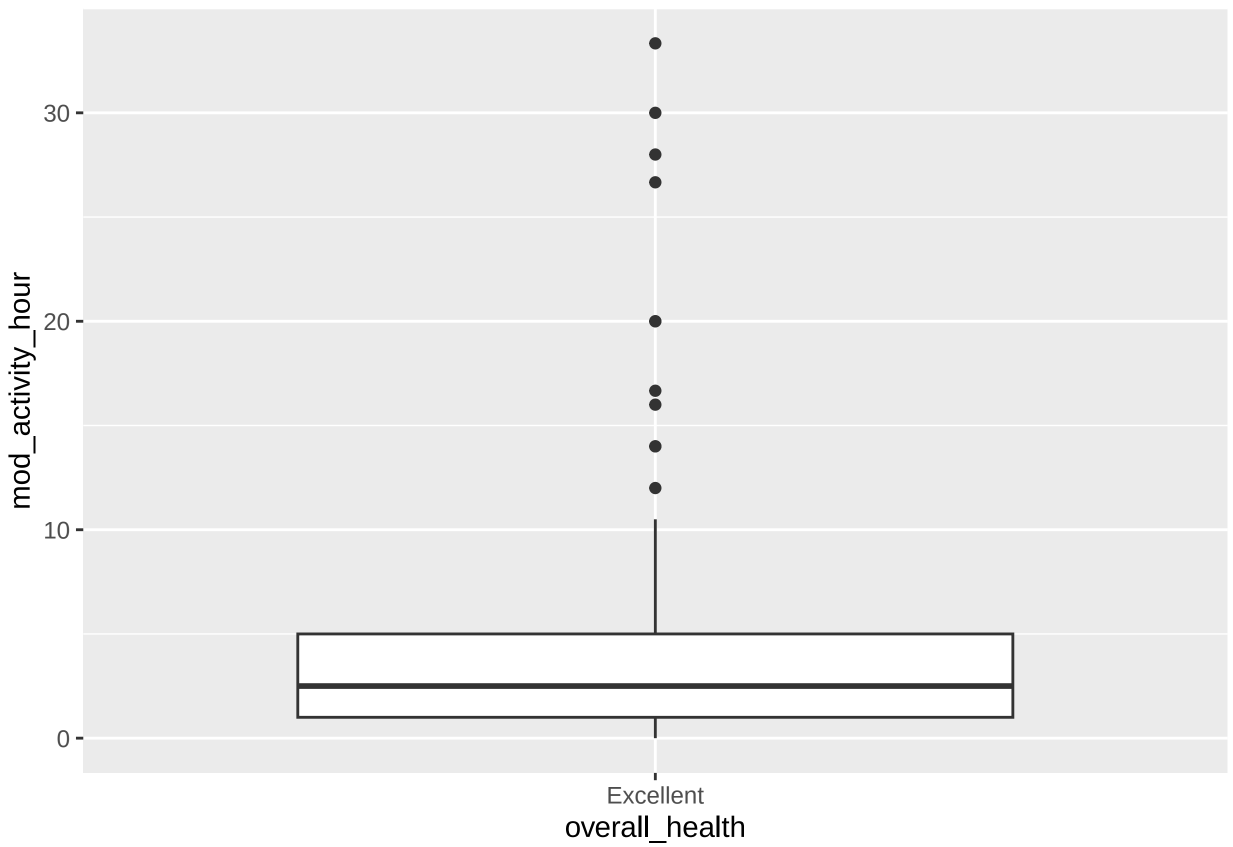
- Make a new variable: make a new variable from "exercise_vig_min" to "vig_activity_hour".

```
# make a new variable using mutate()
ncha_df_sub <- ncha_df_sub %>% mutate(vig_activity_hour = exercise_vig_min / 60)
ncha_df_sub <- ncha_df_sub %>% mutate(mod_activity_hour = exercise_mod_min/60)
```

- Make a simple univariate plot of "mod_activity_hour"

```
# make a simple plot using ggplot()

ggplot(ncha_df_sub, aes(x = overall_health, y = mod_activity_hour)) + geom_boxplot()
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```


mod_activity_hour
30
20
10
0
Excellent
overall_health