WELCOME TO THE NDACAN SUMMER TRAINING SERIES!

National Data Archive on Child Abuse and Neglect

Duke University, Cornell University, University of California San Francisco, & Mathematica

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

SUMMER TRAINING SERIES SCHEDULE

- **July 2nd, 2025**
  - Developing a research question & exploring the data
- **July 9th, 2025**
  - Data management
- **July 16th, 2025**
  - Linking data
- **July 23rd, 2025**
  - Exploratory Analysis
- **July 30th, 2025**
  - Visualization and finalizing the analysis

LIFECYCLE OF AN NDACAN RESEARCH PROJECT

This session is being recorded.

Please submit questions to the Q&A box.

See ZOOM Help Center for connection issues: https://support.zoom.us/hc/en-us

SESSION AGENDA

- STS Review
  - Regression Review

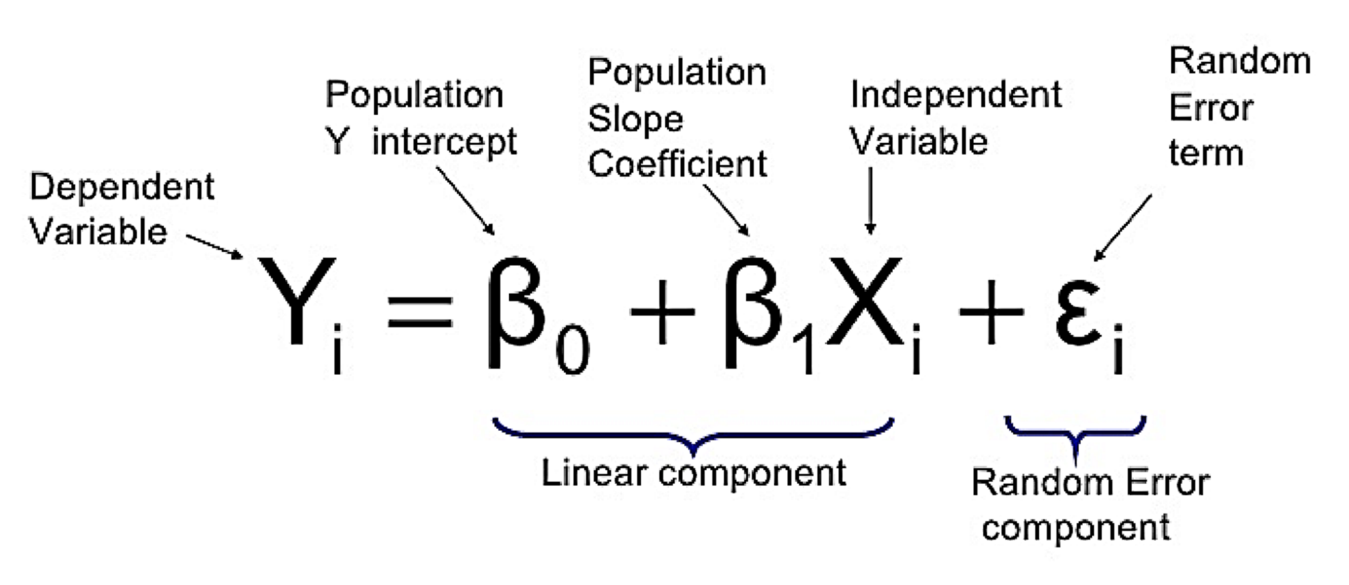
- Data Visualization
  - Univariate Plots
  - Bivariate Plots

- Regression
  - Variable Types
  - Assumption Assessment

STS REVIEW

REGRESSION

- Regression analysis is a statistical method for estimating the relationship between two (or more) random variables
- Equation:



$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Confounders are variables that affect both the independent and dependent variable

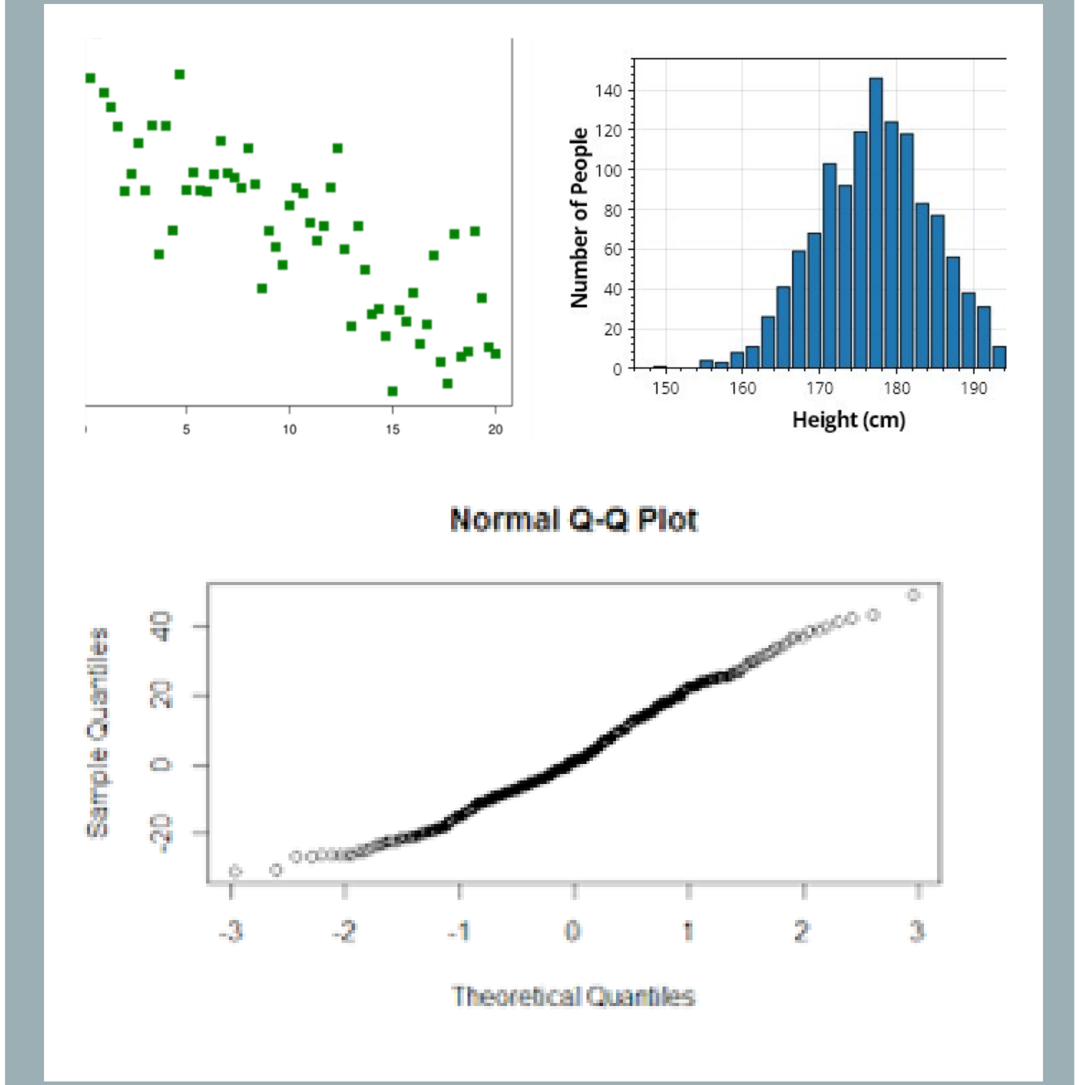
Expansions

- Stratification
- Fixed-State Effect Models

DATA VISUALIZATION

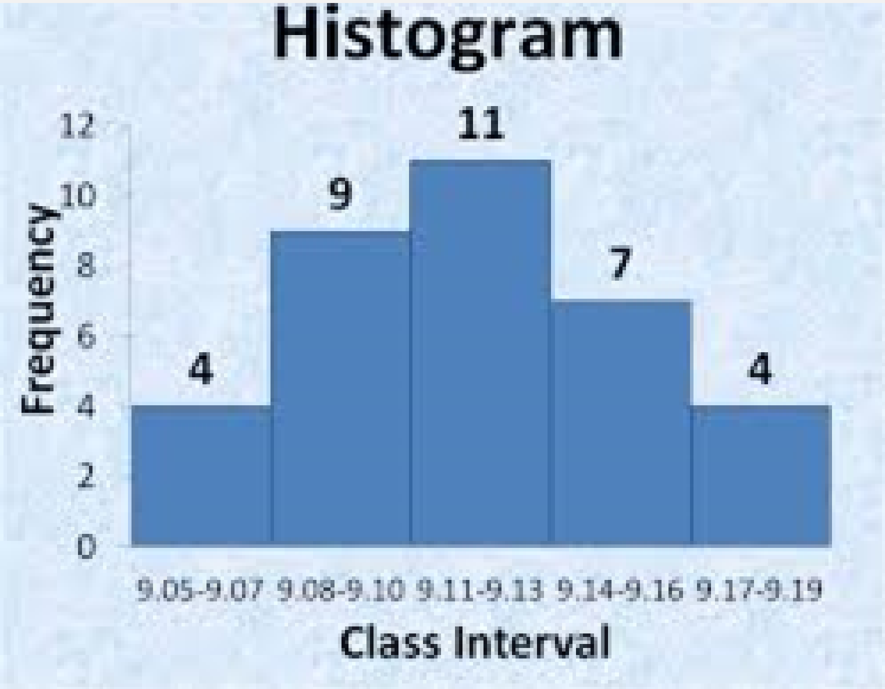
USES FOR DATA VISUALIZATION

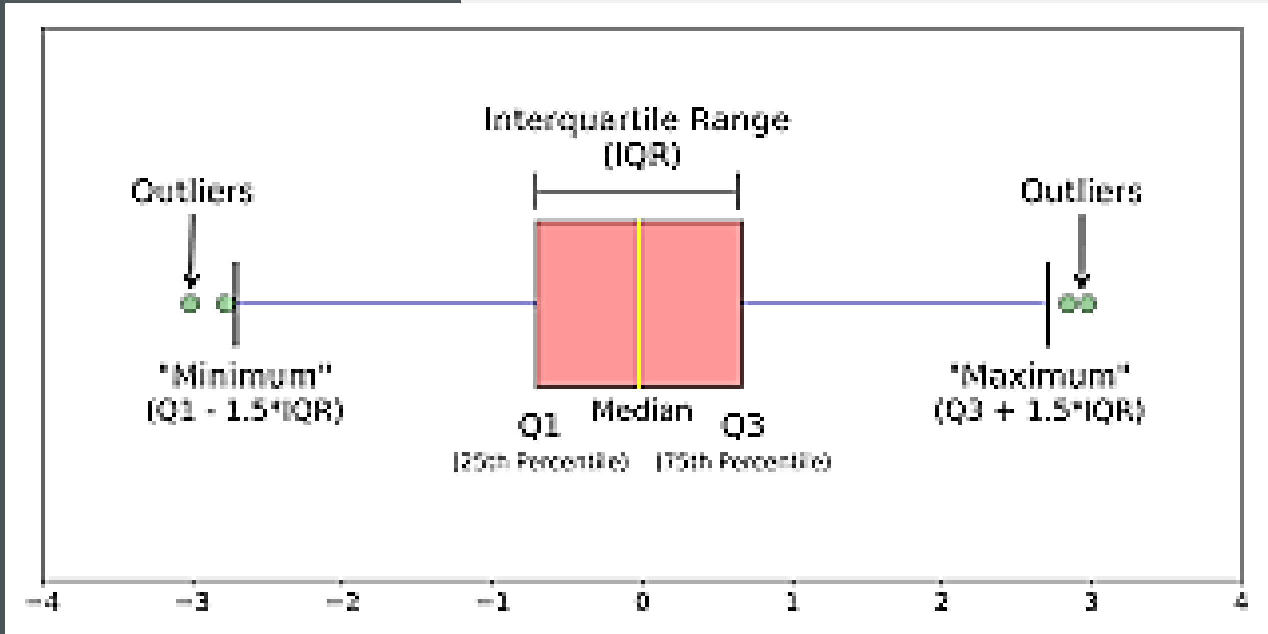
- Holistic Overview
  - Provides a quick, concise, visual summary of data
- Association at a Glance
  - Reveal trends or pattern in data
- Identify General Nature of Relationship
  - e.g. Linear, Quadratic, Cubic Splines
- Assumption Evaluation
  - Aids in the validation of assumption testing



UNIVARIATE PLOTS

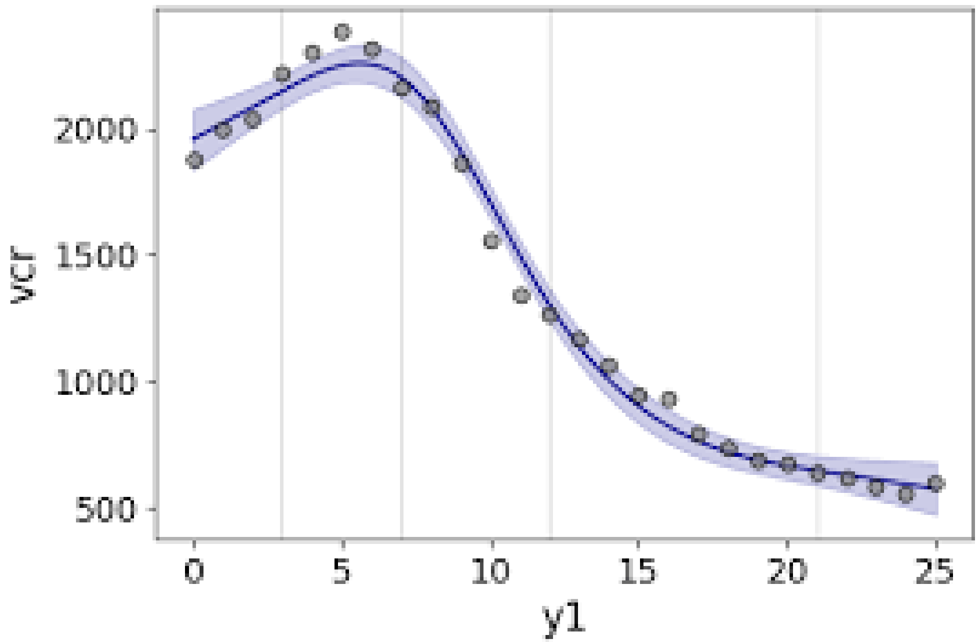
- Helpful for showing distribution of a single variable
- Types
  - Histograms – Discrete
    - Geom_histogram (in R)
  - Dot Plots - Discrete
    - Geom_dotplot (in R)
  - Box and Whisker – Continuous
    - Geom_boxplot (in R)





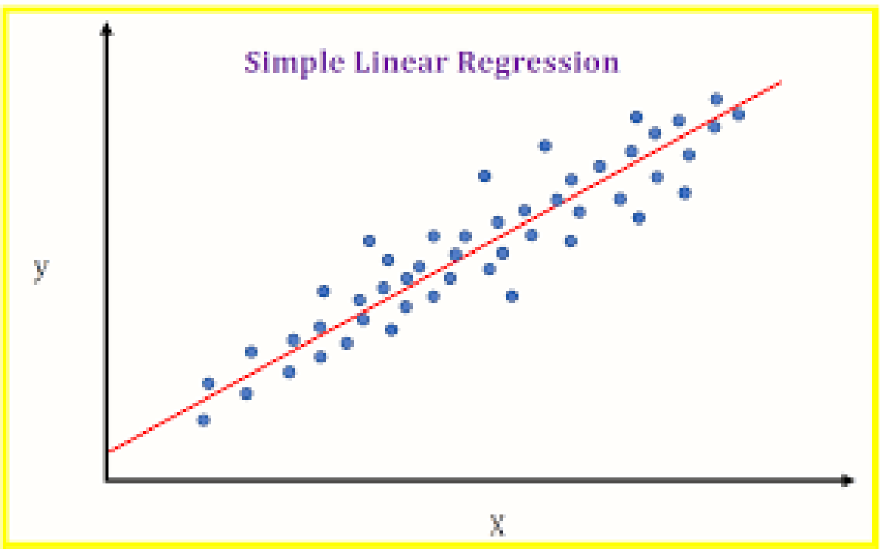
BIVARIATE PLOTS

- Helpful for showing relationship between two variables
  - Linear, Quadratic?
- Types
  - Scatterplot – Two Continuous Variables
    - Geom_point
  - Cubic Splines – Two Continuous Variables
    - Geom_smooth

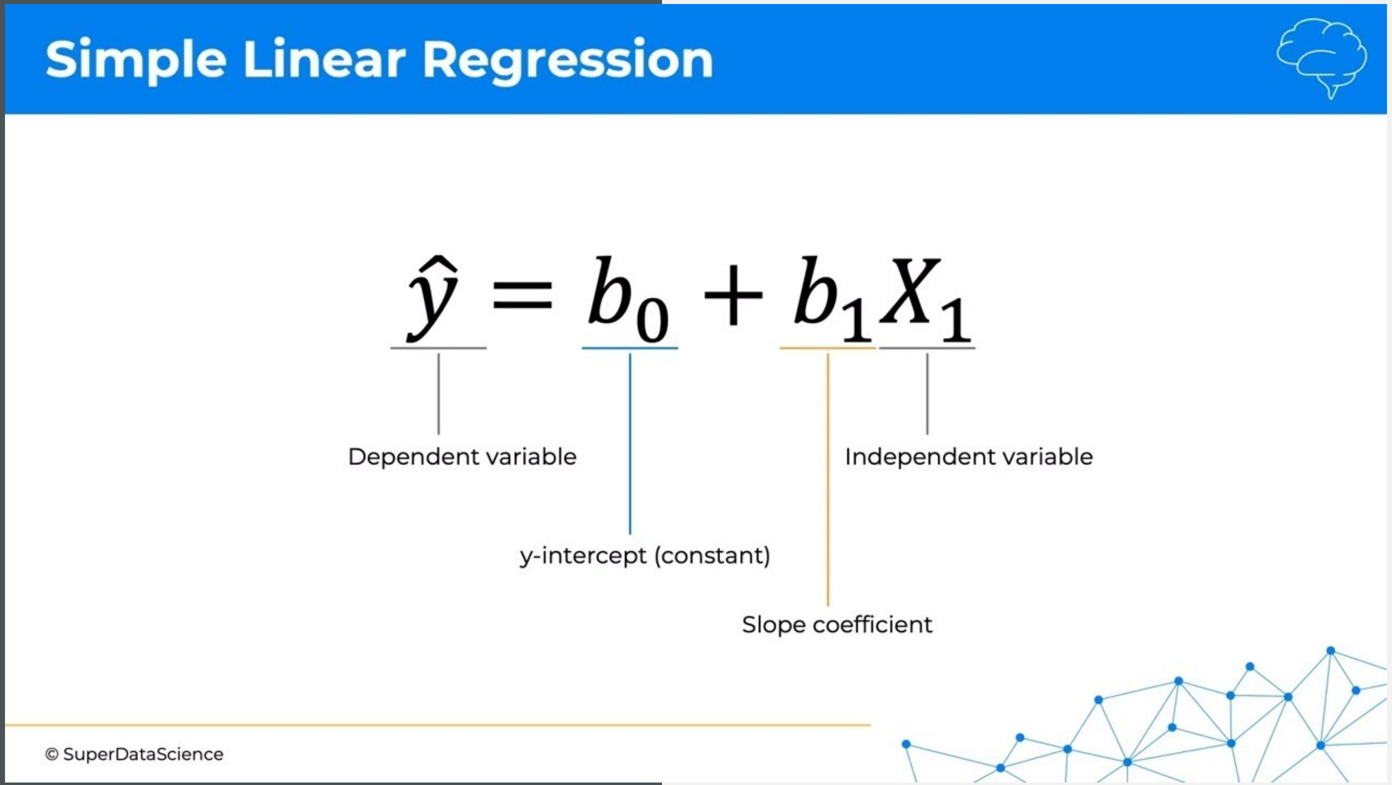


REGRESSION REVIEW

WHAT IS REGRESSION



- **Simple Linear Regression models LINEAR relationship between a dependent and independent variable**
- **Composition**
  - **Right Side**
    - **Intercept: b0**
    - **Slope Intercept: b1**
    - **Independent Variable: X1**
- **Multiple Linear Regression**



Simple Linear Regression

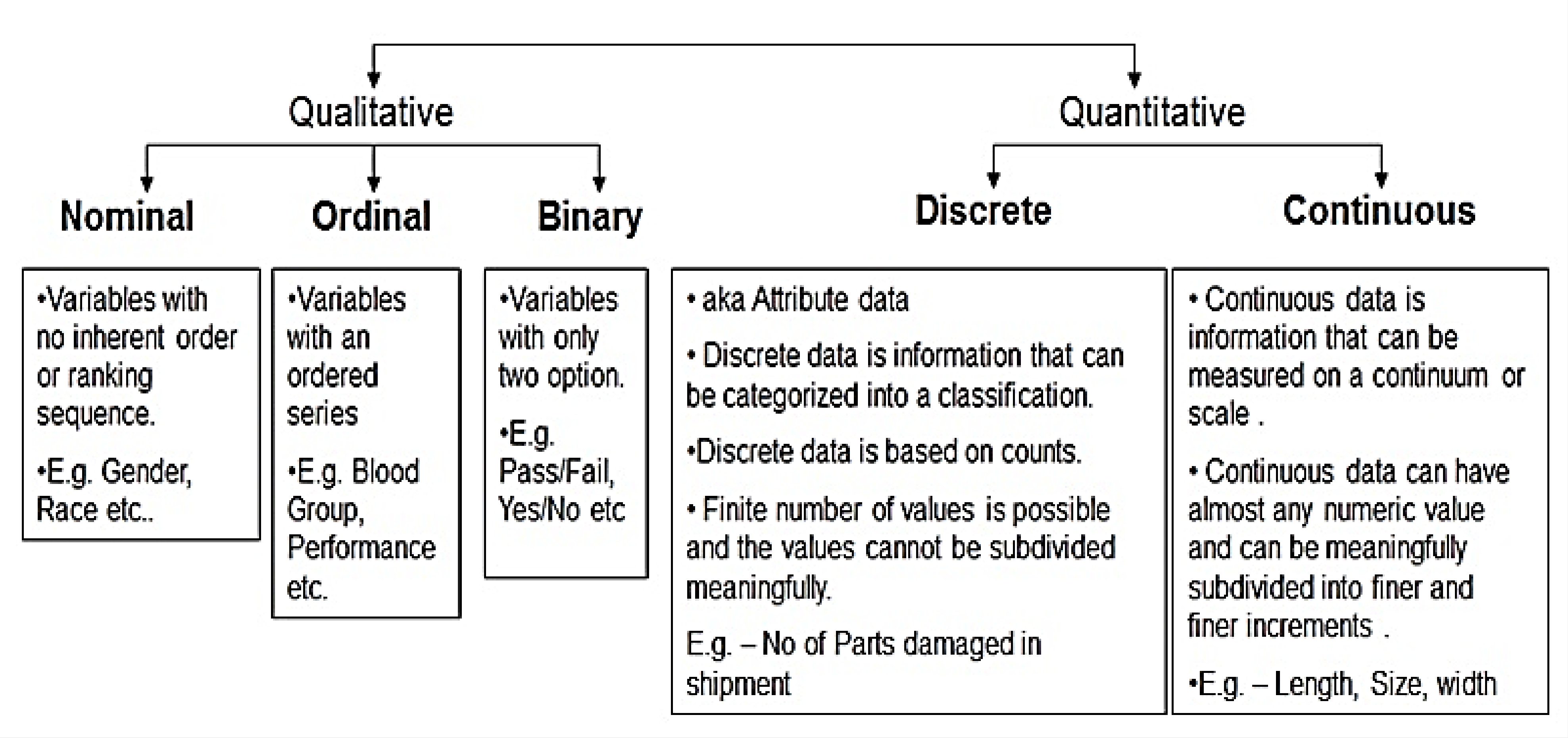
$$\underbrace{\hat{y}}_{\text{Dependent variable}} = \underbrace{b_0}_{\text{y-intercept (constant)}} + \underbrace{b_1}_{\text{Slope coefficient}}\underbrace{X_1}_{\text{Independent variable}}$$

© SuperDataScience

VARIABLE TYPES

- **Quantitative**
  - Discrete- Poisson Regression
  - **Continuous** - Simple or Multiple Linear Regression

- **Qualitative**
  - Binary - Logistic Regression
  - Nominal – Multinomial Logistic Regression
  - Ordinal – Ordinal Logistic Regression



Qualitative

Nominal

•Variables with no inherent order or ranking sequence.

•E.g. Gender, Race etc..

Ordinal

•Variables with an ordered series

•E.g. Blood Group, Performance etc.

Binary

•Variables with only two option.

•E.g. Pass/Fail, Yes/No etc

Quantitative

Discrete

• aka Attribute data

• Discrete data is information that can be categorized into a classification.

•Discrete data is based on counts.

• Finite number of values is possible and the values cannot be subdivided meaningfully.

E.g. – No of Parts damaged in shipment

Continuous

• Continuous data is information that can be measured on a continuum or scale .

• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .

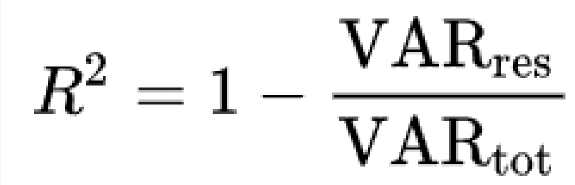
•E.g. – Length, Size, width

ASSUMPTION ASSESSMENT

- Core Assumptions
  - Homoscedasticity – Variance of residuals is constant for all levels of all independent variables
    - Use plots of residuals
  - Independence - Each observation is independent of others
    - Verify with study design
  - Linearity
    - Inspect scatter plots of independent and dependent variables
  - No Multicollinearity
    - Independent variables are not correlated with each other
    - Check correlation tables

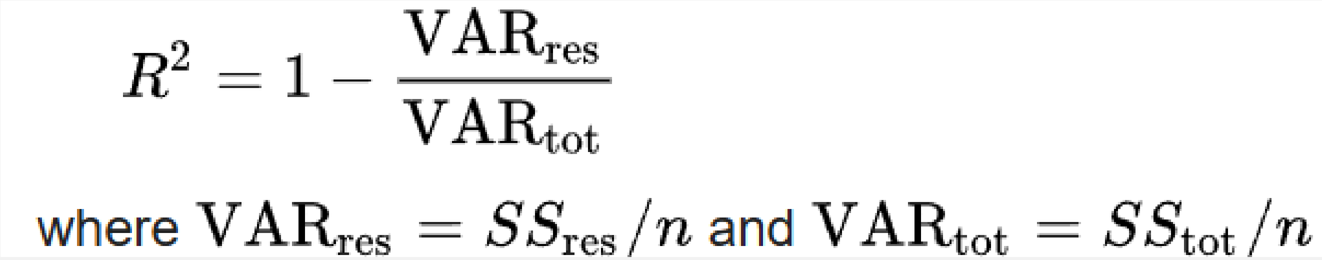
COEFFICIENTS OF DETERMINATION

- How do we determine how well our model performs?
- Coefficients of Determination
  - R-square



$$R^2 = 1 - \frac{\mathrm{VAR}_{\mathrm{res}}}{\mathrm{VAR}_{\mathrm{tot}}}$$

  - Adj R-square



$$R^2 = 1 - \frac{\mathrm{VAR}_{\mathrm{res}}}{\mathrm{VAR}_{\mathrm{tot}}}$$

where $\mathrm{VAR}_{\mathrm{res}} = SS_{\mathrm{res}}/n$ and $\mathrm{VAR}_{\mathrm{tot}} = SS_{\mathrm{tot}}/n$

HELPFUL RESOURCE

- GGplot2 Cheat Sheet
  - https://posit.co/wp-content/uploads/2022/10/data-visualization-1.pdf
- GGplot2 Documentation
  - https://ggplot2.tidyverse.org/reference/index.html

QUESTIONS?

ALEX ROEHRKASSE
AROEHRKASSE@BUTLER.EDU

NOAH WON
NOAH.WON@DUKE.EDU

PAIGE LOGAN PRATER
PAIGE.LOGANPRATER@UCSF.EDU

R CODE PAGE 1 OF 6

```
#########
# NOTES #
#########

# This program file demonstrates strategies discussed in
# session 5 of the 2025 NDACAN Summer Training Series
# "Data Visualization."

# For questions, contact the presenter
# Noah Won (noah.won@duke.edu).

# Note that because of the process used to anonymize data,
# all unique observations include partially fabricated data
# that prevent the identification of respondents.
# As a result, all descriptive and model-based results are fabricated.
# Results from this and all NDACAN presentations are for training purposes only
# and should never be understood or cited as analysis of NDACAN data.


#####################
# TABLE OF CONTENTS #
#####################

# 0. SETUP
# 1. Univariate Plots
# 2. Bivariate Plots
# 3. Logistic Regression
```

R CODE PAGE 2 OF 6

```
############
# 0. SETUP #
############

# Clear environment
rm(list=ls())

# Installs packages if necessary, loads packages
if (!requireNamespace("pacman", quietly = TRUE)){
  install.packages("pacman")
}
pacman::p_load(data.table, tidyverse, mice)

# Defines filepaths working directory
project <- "C:/Users/nhwn1/Downloads/STS5/data"
data <- "C:/Users/nhwn1/Downloads/STS5/data"

# Set working directory
setwd(project)

# Set seed
set.seed(1013)
```

R CODE PAGE 3 OF 6

```
####################
# 2. Univariate Plots #
####################

# Let's read in our cleaned, anonymized
# versions of the 2020 AFCARS files
afcars <- fread(paste0(data,'/afcars_clean_anonymized_linear.csv'))
head(afcars, 20)

# Running frequency tables of predictors of interest
table(afcars$SEX)
table(afcars$RaceEthn)
table(afcars2$FCMntPay)

# Creating Dummy Variables and Age Variables for Predictors
# Also filtering out those older than 30
afcars2 <- afcars %>%
        mutate(SEX_d = case_when(
          SEX == "Male" ~ 1,
          SEX == "Female" ~ 0),
        CLINDIS_d = case_when(
            CLINDIS == "Yes" ~ 1,
            CLINDIS == "No" ~ 0),
        Hispanic = case_when(
          RaceEthn == "Hispanic" ~ 1,
          TRUE ~ 0),
        age = as.numeric(difftime(Sys.Date(), DOB, units = "days")) / 365.25
        ) %>%
        filter(age <= 30)
```

R CODE PAGE 4 OF 6

```
# Checking new derived variables
table(afcars2$SEX_d)
table(afcars2$CLINDIS_d)
table(afcars2$Hispanic)
table(afcars2$age)

#Let's plot the distribution of age using a VERY basic histograms
hist1 <- ggplot(afcars2, aes(x = age)) + geom_histogram()
hist1

#Let's add some titles using the labs() function
hist2 <- ggplot(afcars2, aes(x = age)) +
  geom_histogram() +
  labs(title = "Histogram of Age",
       x = "Age",
       y = "Frequency")
hist2

#Let's add some color, a theme, and increase total number of bins
hist3 <- ggplot(afcars2, aes(x = age)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white") +
  labs(title = "Histogram of Age",
       x = "Age",
       y = "Frequency")
hist3

#Let's do this for a Box and Whisker plot by SEX
box1 <- ggplot(afcars2, aes(x = SEX, y = age)) +
  geom_boxplot(fill = "skyblue", color = "darkblue") +
  labs(title = "Box and Whisker Plot of Age by Sex",
       y = "Age", x = "Sex") +
  theme_minimal()
box1
```

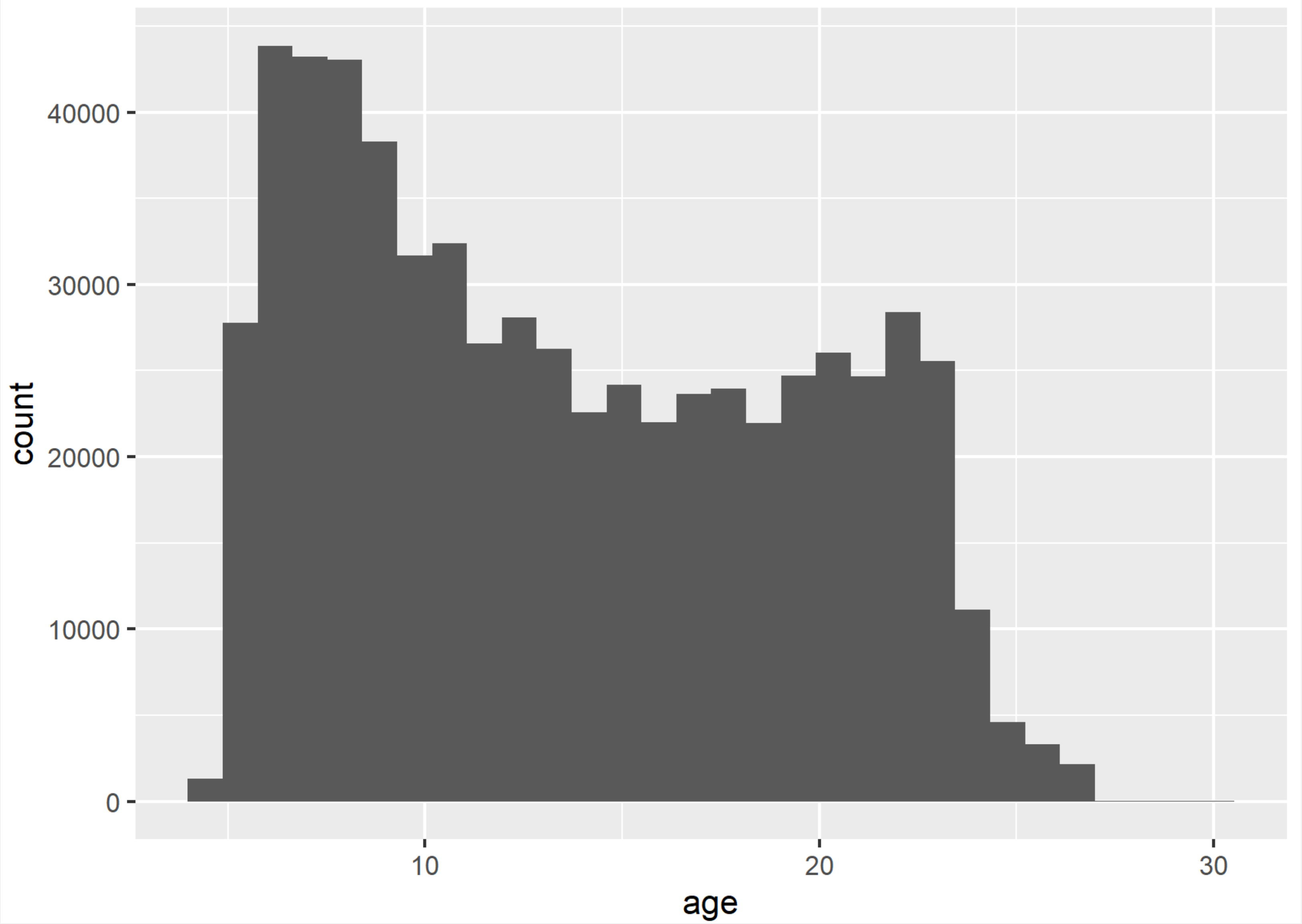
R CODE PAGE 5 OF 6

```
##################################################
# 3. Bivariate Plots #
# Let's plot the relationship between two continous variables using a scatterplot
scatter1 <- ggplot(afcars2, aes(x = age, y = FCMntPay)) +
  geom_point(color = "steelblue", alpha = 0.6, size = 2) +
  labs(title = "Scatterplot of Foster Care Monthly Payment vs Age",
       x = "Age",
       y = "Foster Care Monthly Payment (FCMntPay)") +
  theme_minimal()
scatter1

#There seems to be a positive relationship between age and Monthly Foster Care Payment but
it is hard to tell
#Let's fit a cubic spline in the data to see
scatter2 <- ggplot(afcars2, aes(x = age, y = FCMntPay)) +
  geom_point(color = "steelblue", alpha = 0.6, size = 2) +
  geom_smooth(method = "gam", formula = y ~ s(x, bs = "cs"), color = "darkred", se = FALSE) +
  labs(title = "Scatterplot of FCMntPay vs Age with Cubic Spline Fit",
       x = "Age",
       y = "Foster Care Monthly Payment (FCMntPay)") +
  theme_minimal()
scatter2

#The gam method stand for Generalized Additive Model, which is a regression model that fits
cubic splines onto data
#The formula section fits a smooth cubic spline of x onto y
#This scatterplot can be used to hollistically evaluate linearity between two continous variables
#There exists hypothesis tests for testing linearity but this falls outside the scope of this lecture
```

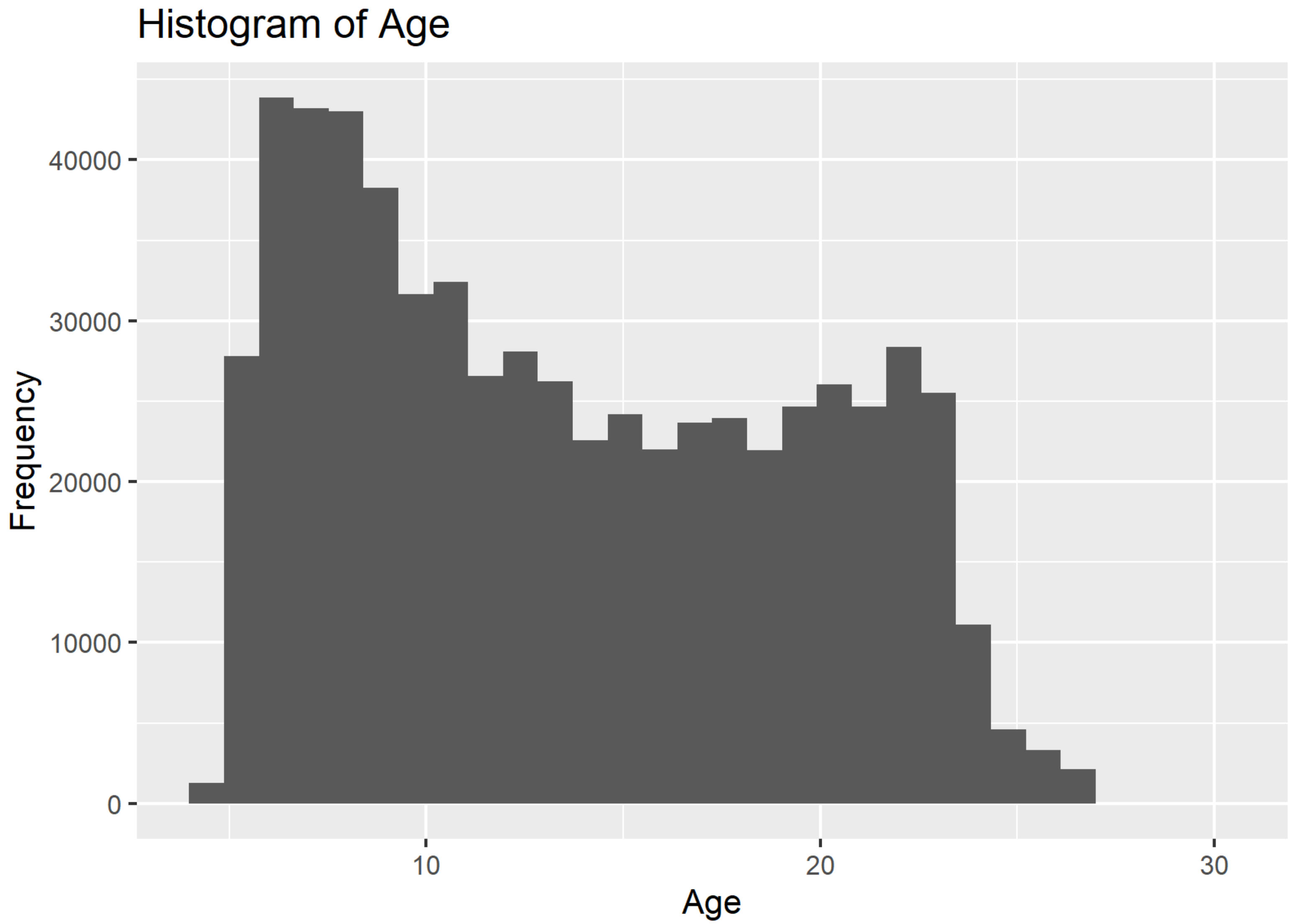
R CODE PAGE 6 OF 6

```
#################################################
# 4. Logistic Regression #
#################################################
#What if we want to model and outcome that is NOT continuous but binary (i.e. has two different values, yes/no, male/female, etc.)
#We can use a logistic regression model which converts the outcome variable to log odds
#Odds is a ratio of outcomes that we can use to model chance
#What is the probability we don't roll a 6 on a fair die? 5/6. What are the odds? 5 to 1.
logmodel <- glm(CLINDIS_d ~ age, data = afcars2, family = binomial)
summary(logmodel)

#The estimates we received are log odds. To determine odds, we need to exponentiate the estimate. The odds estimate
#for age is e^.0944796 = 1.099
#Since our p-value is significant, we can say that the odds of being clinically diagnosed with a disability
#increase 1.099 times per year of age
```

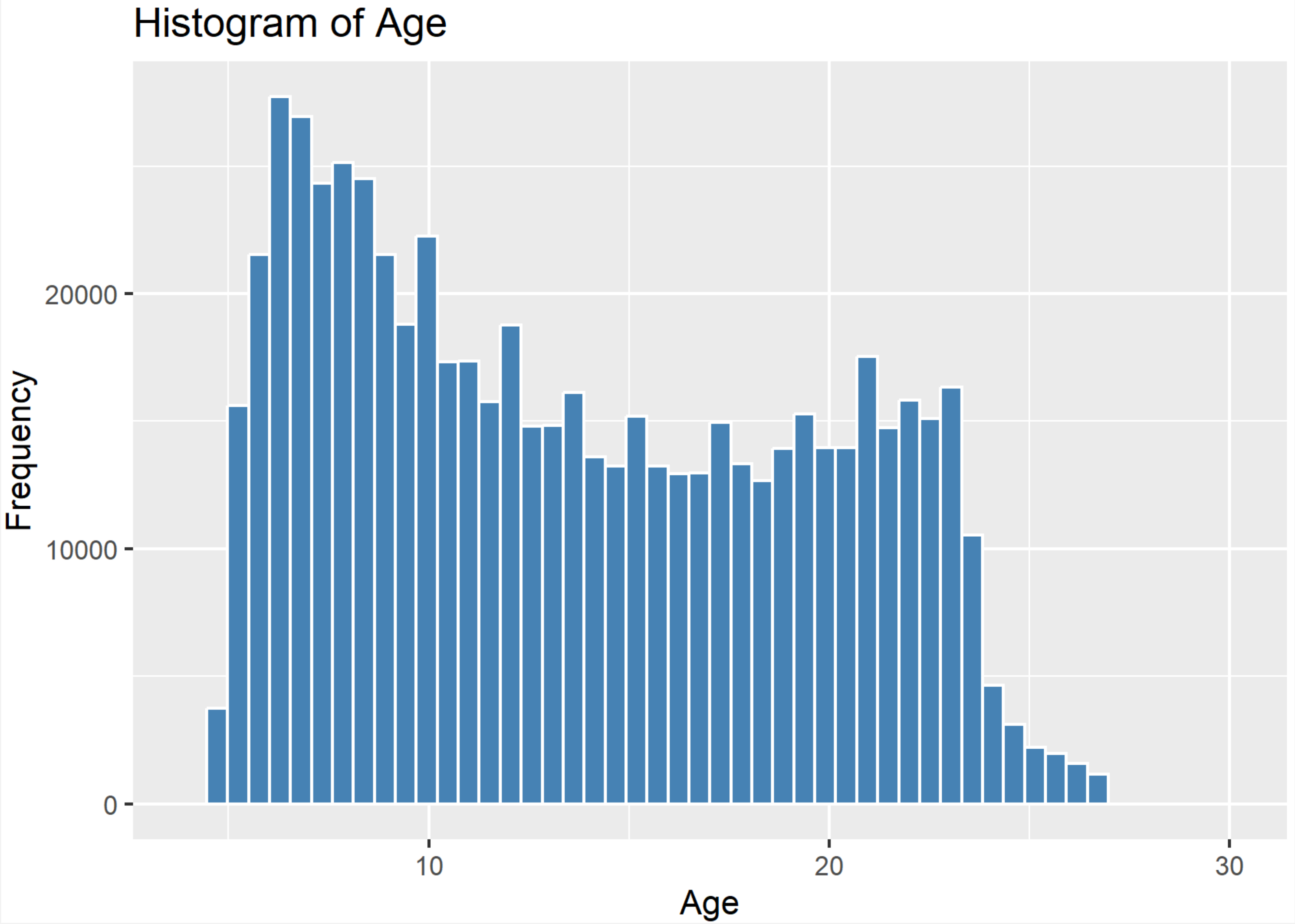
R CODE IMAGE 01: BASIC HISTOGRAM



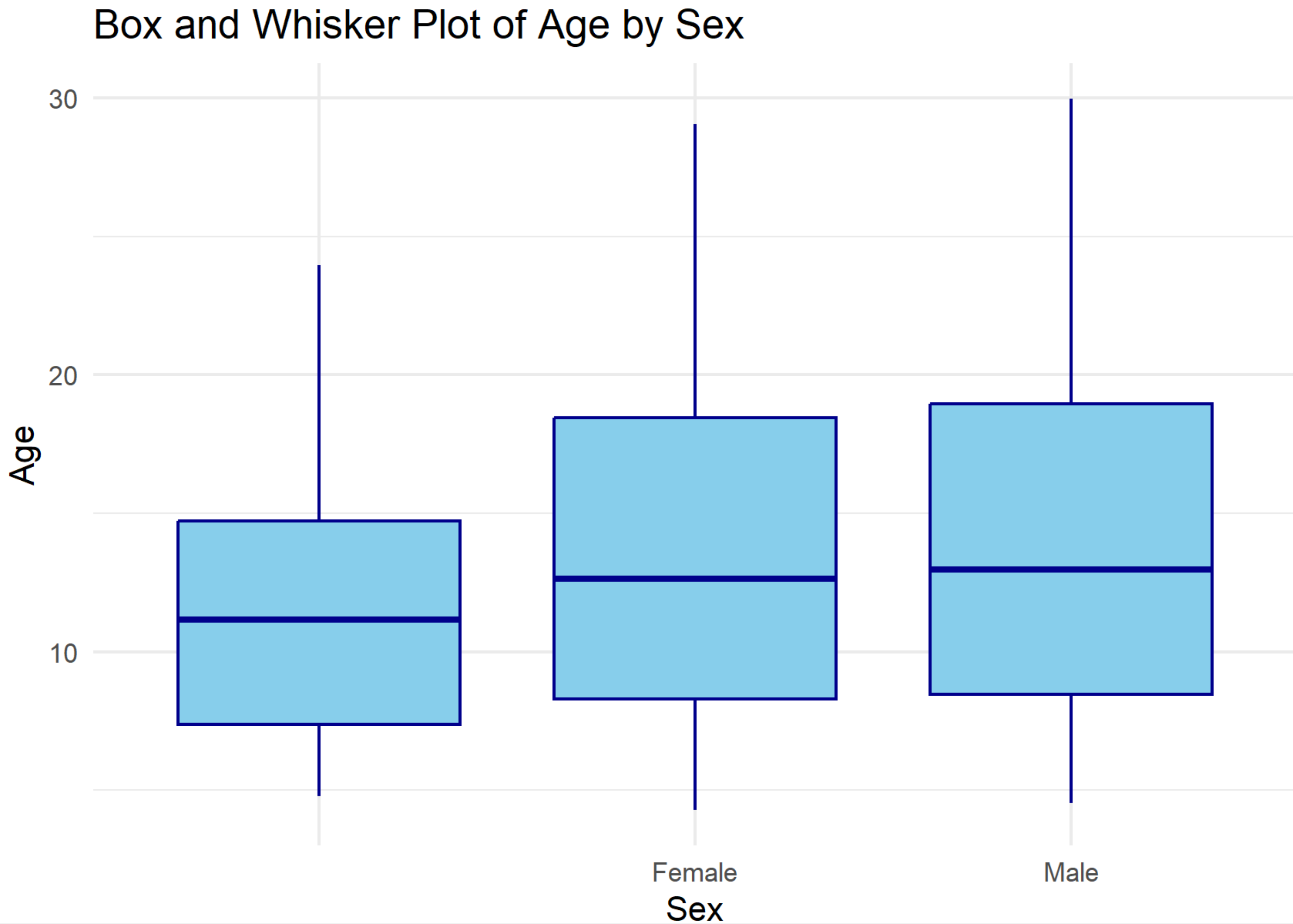
R CODE IMAGE 02: HISTOGRAM WITH TITLE AND LABELS



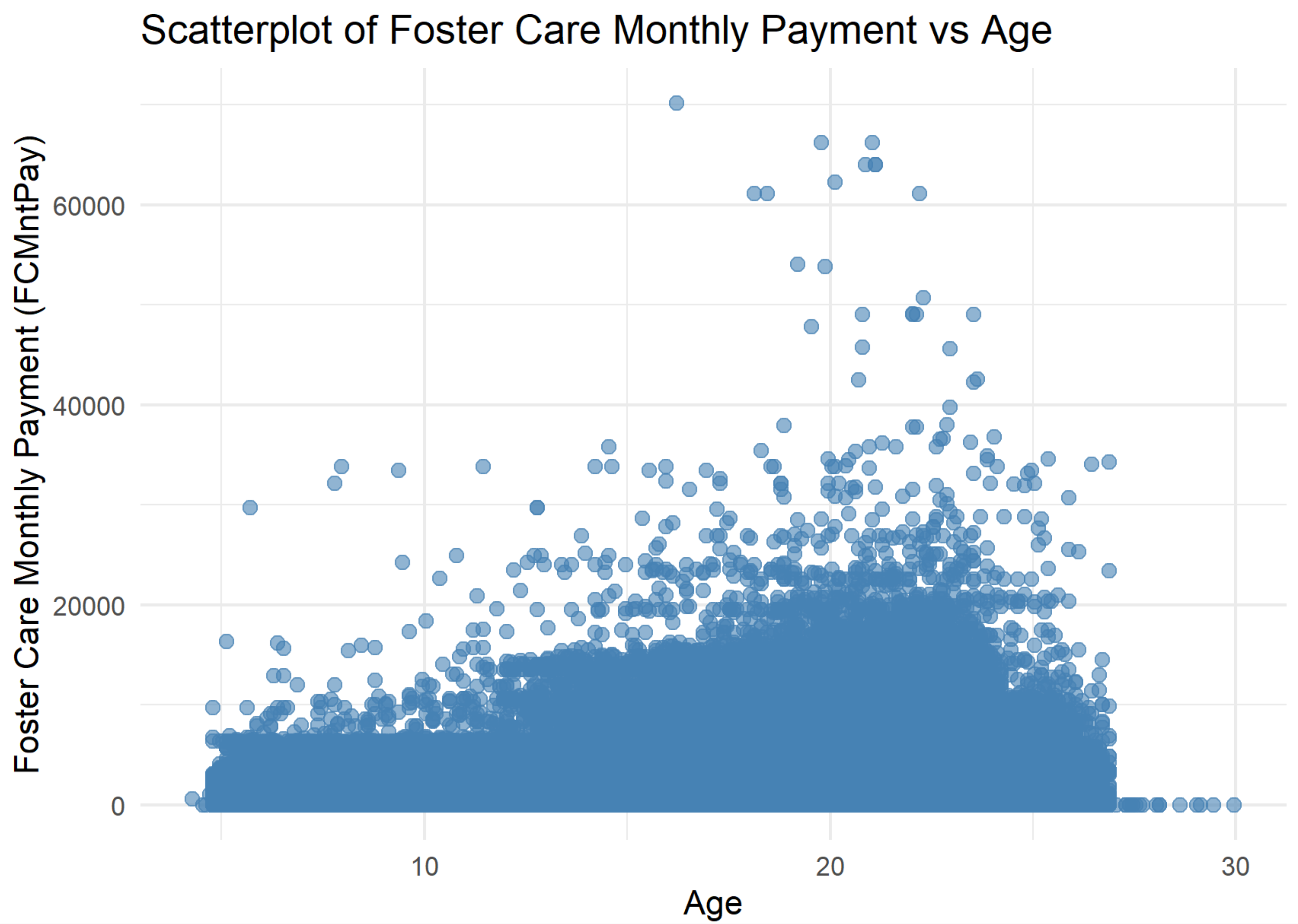
R CODE IMAGE 03: HISTOGRAM IN COLOR WITH BIN ADJUSTMENT



R CODE IMAGE 04: BOX AND WHISKER PLOT OF AGE BY SEX



R CODE IMAGE 05: SCATTERPLOT OF FOSTER CARE MONTHLY PAYMENT VS. AGE


Scatterplot of Foster Care Monthly Payment vs Age
Foster Care Monthly Payment (FCMntPay)
Age