LAB WEEK 1

DESCRIBING DATA

UVEAL MELANOMA AND DEPRESSION

Uveal melanoma, a rare eye cancer, presents potential vision loss and life threat. This prospective, longitudinal study interrogated the predictive utility of visual impairment, as moderated by optimism/pessimism, on depressive symptoms in 299 adults undergoing diagnostic evaluation.

Annette Stanton

James MacDonald

MacDonald, J.J., Jorge-Miller, A., Enders, C.K., McCannel, T., Beran, T., & Stanton, A.L. (2021). Perceived and objective visual impairment predicting depressive symptoms across one year in uveal melanoma diagnostic biopsy: Optimism and pessimism as moderators. *Health Psychology*, *40*, 408-417.

PRELIMINARIES: LOADING PACKAGES

- A package is a collection of functions bundled together to extend R's capabilities beyond its basic capabilities

```
# LOAD R PACKAGES ----

# load R packages
library(ggplot2)
library(psych)
library(summarytools)
```

IMPORTING DATA

= data frame name
= variable name
= raw data file name

- Import the raw data (in a .csv file on a website) and into an R data frame called Cancer

```
# READ DATA ----

# url for raw data
filepath <-
  'https://raw.githubusercontent.com/craigenders/psych250a/main/data/CancerData.csv'

# import CancerData.csv from the url filepath into an R data frame called Cancer
# stringsAsFactors converts alphanumeric variables to "factors" (categorical
variables)
Cancer <- read.csv(filepath, stringsAsFactors = T)
```

SUMMARIZING DATA

= data frame name

= variable name

- The dfSummary function gives numeric and visual summaries of a data frame's variables, and the describe function gives descriptive statistics

```
# INSPECT DATA ----

# summarize entire data frame (summarytools package)
dfSummary(Cancer)

# DESCRIPTIVE STATISTICS ----

# descriptive statistics for entire data frame (psych package)
describe(Cancer)
```

R OUTPUT

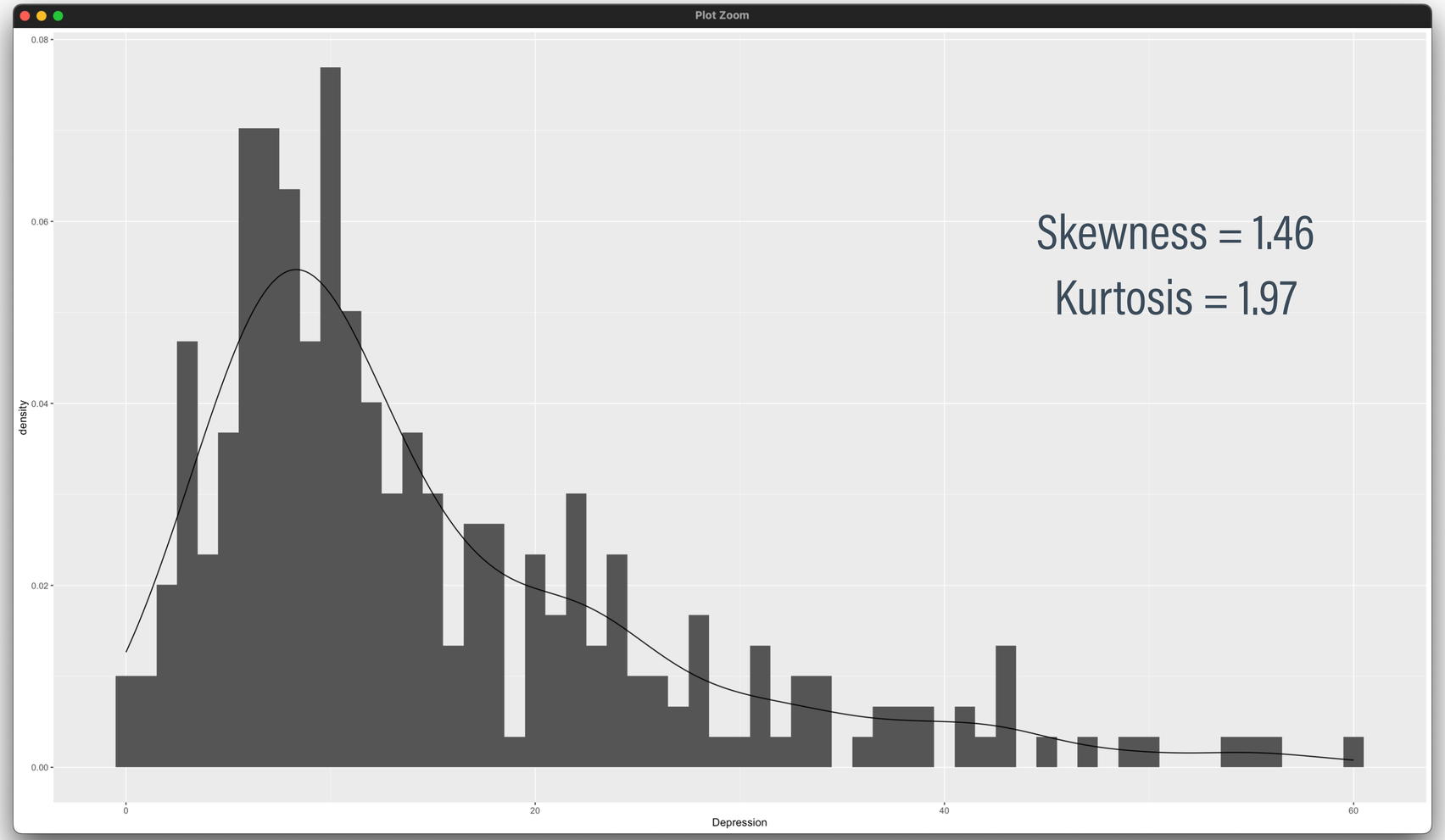
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Participant	1	299	150.00	86.46	150.00	150.00	111.19	1.00	299.00	298.00	0.00	-1.21	5.00
Diagnosis*	2	299	1.64	0.48	2.00	1.68	0.00	1.00	2.00	1.00	-0.59	-1.66	0.03
Age	3	299	58.99	13.94	60.00	59.47	11.86	19.00	91.00	72.00	-0.33	-0.15	0.81
Gender*	4	299	1.54	0.50	2.00	1.55	0.00	1.00	2.00	1.00	-0.15	-1.98	0.03
Comorbids	5	299	0.94	0.74	1.00	0.93	1.48	0.00	2.00	2.00	0.10	-1.17	0.04
Optimism	6	299	8.65	3.06	9.00	8.87	2.97	0.00	14.00	14.00	-0.62	-0.23	0.18
Depression	7	299	14.85	11.44	11.00	13.12	7.41	0.00	60.00	60.00	1.46	1.97	0.66
VisImpair	8	299	5.09	2.23	4.85	4.98	2.21	0.36	12.18	11.82	0.43	-0.03	0.13

HISTOGRAM AND KERNEL DENSITY PLOTS

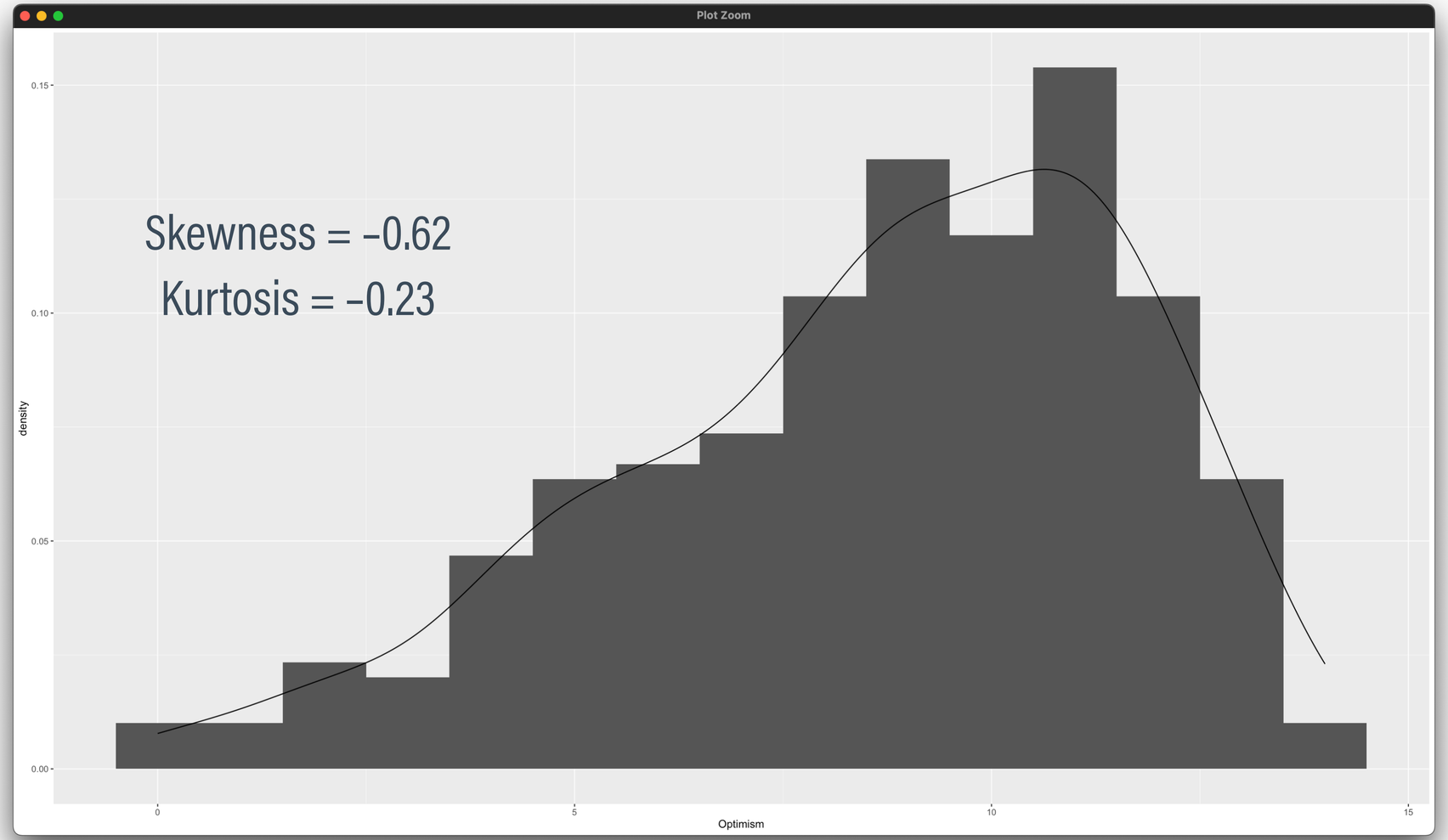
= data frame name
= variable name

```
# HISTOGRAMS AND KERNEL DENSITY PLOTS FOR NUMERIC VARIABLES ----

# histogram and kernel density plot for numeric variable (ggplot2 package)
ggplot(Cancer, aes(x = Depression)) +
  geom_histogram(aes(y = after_stat(density)), bins = 61) +
  geom_density()

# histogram and kernel density plot for numeric variable (ggplot2 package)
ggplot(Cancer, aes(x = Optimism)) +
  geom_histogram(aes(y = after_stat(density)), bins = 15) +
  geom_density()

# histogram and kernel density plot for numeric variable (ggplot2 package)
ggplot(Cancer, aes(x = VisImpair)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50) +
  geom_density()
```

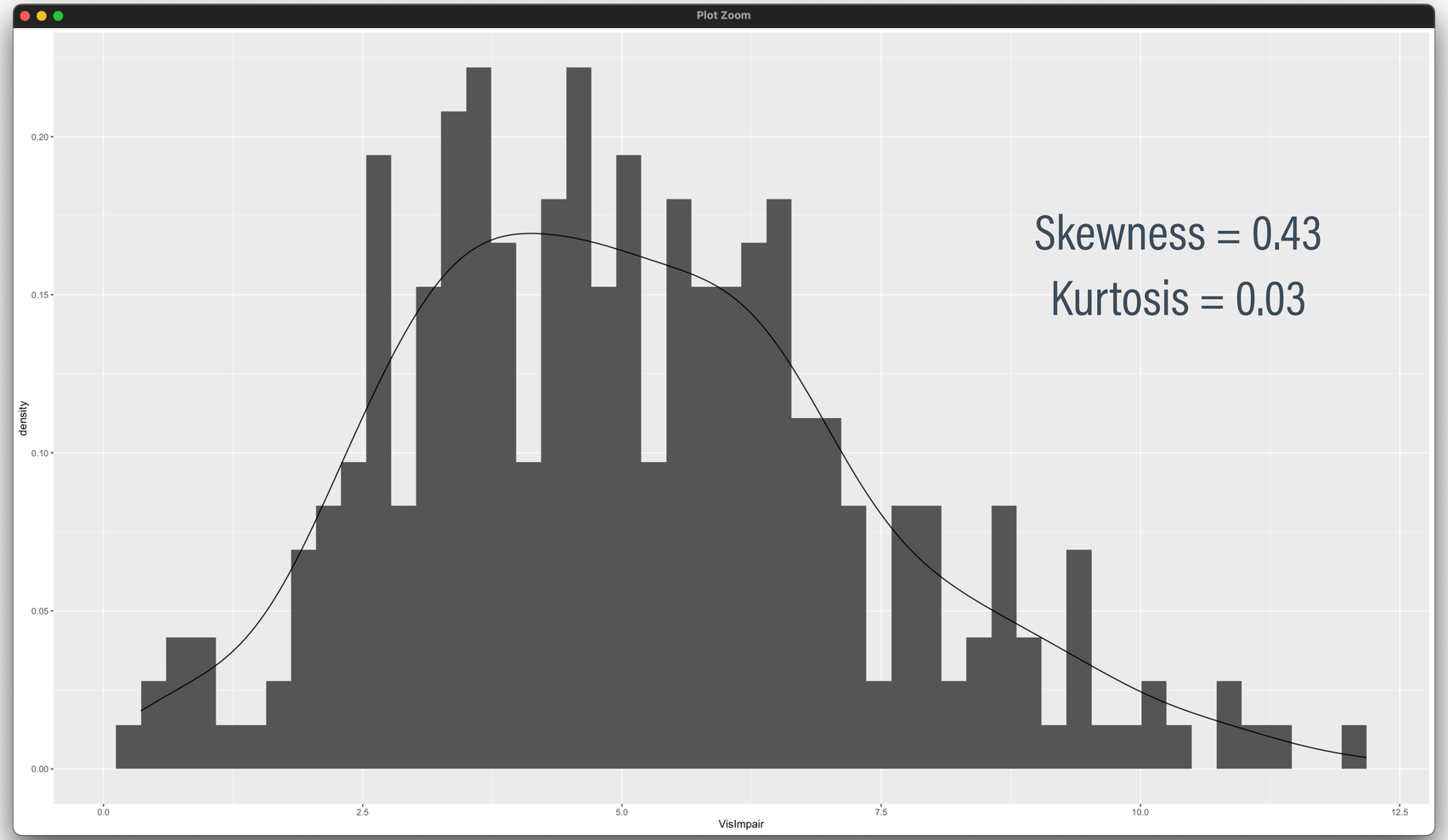
DEPRESSION DISTRIBUTION



OPTIMISM DISTRIBUTION



VISUAL IMPAIRMENT DISTRIBUTION

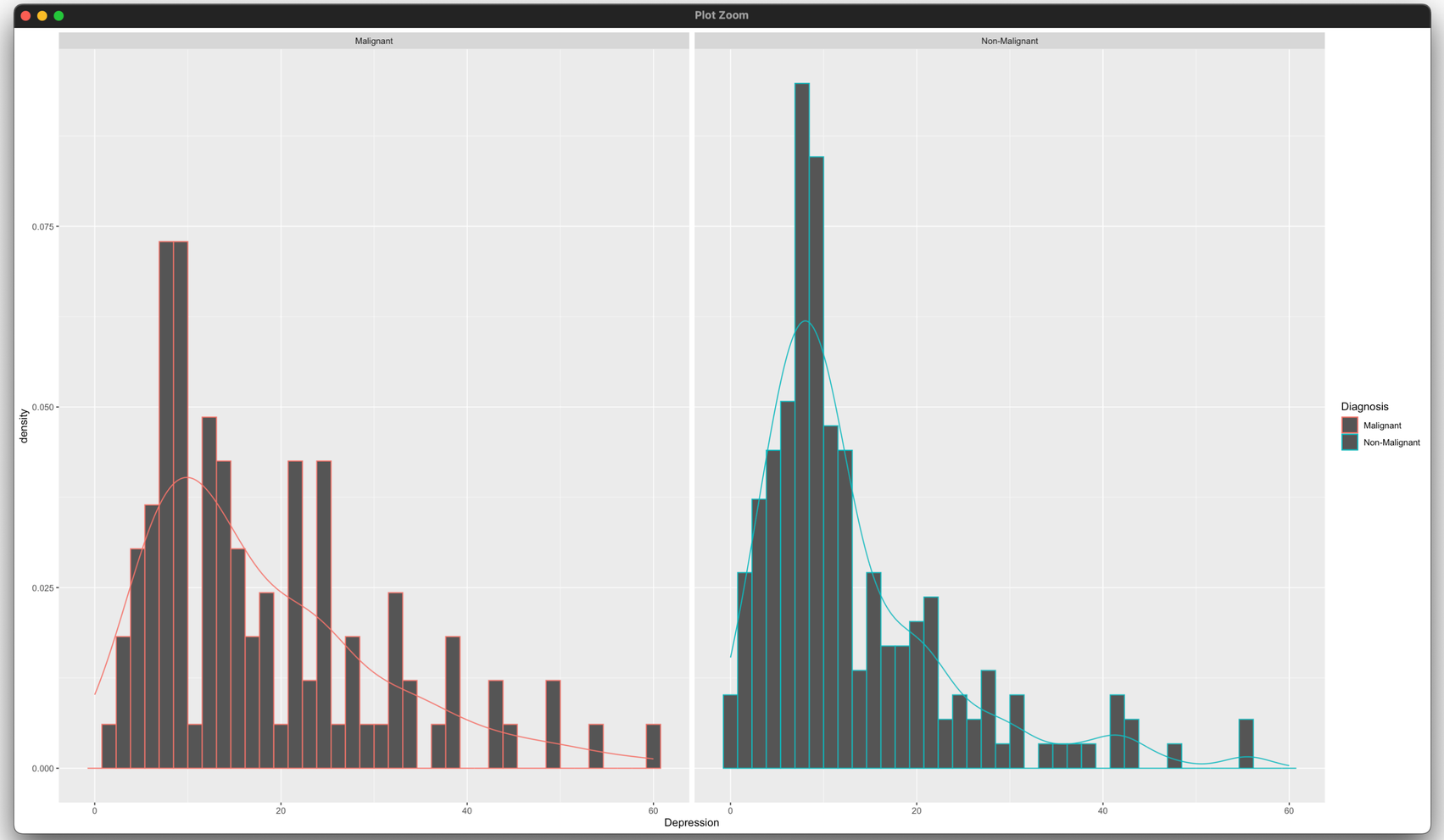


DISTRIBUTION PLOTS BY GROUP

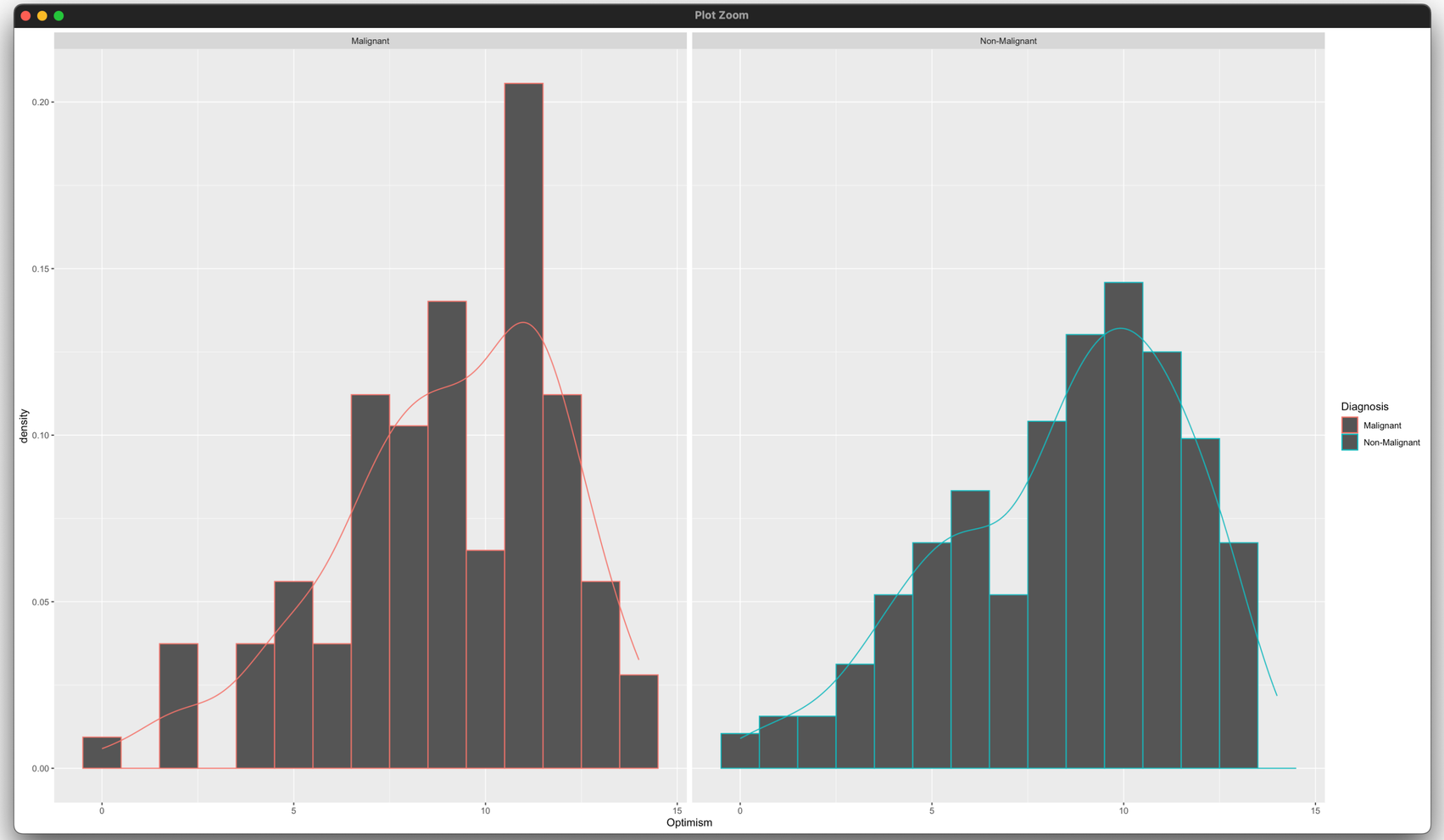
= data frame name
= variable name
= grouping variable

```
# HISTOGRAMS AND KERNEL DENSITY PLOTS BY GROUP ----

# histogram and kernal density plot separately by group (ggplot2 package)
ggplot(Cancer, aes(x = Depression, colour = Diagnosis)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40) +
  geom_density() +
  facet_wrap(~ Diagnosis)

# histogram and kernal density plot separately by group (ggplot2 package)
ggplot(Cancer, aes(x = Optimism, colour = Diagnosis)) +
  geom_histogram(aes(y = after_stat(density)), bins = 15) +
  geom_density() +
  facet_wrap(~ Diagnosis)
```

DEPRESSION DISTRIBUTIONS



OPTIMISM DISTRIBUTIONS



DEFINE FACTOR (NOMINAL) VARIABLES

= data frame name

= variable name

- Frequency distributions are best suited for variables with relatively few score values

```
# FREQUENCY DISTRIBUTIONS FOR DISCRETE VARIABLES  ----

# frequency distributions for discrete or categorical variables (summarytools
package)
freq(Cancer$Diagnosis)
freq(Cancer$Optimism)
freq(Cancer$Comorbids)
```

R OUTPUT

Frequencies
Cancer$Diagnosis
Type: Factor

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Malignant	107	35.79	35.79	35.79	35.79
Non-Malignant	192	64.21	100.00	64.21	100.00
<NA>	0			0.00	100.00
Total	299	100.00	100.00	100.00	100.00

R OUTPUT

Frequencies
Cancer$Optimism
Type: Integer

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
0	3	1.00	1.00	1.00	1.00
1	3	1.00	2.01	1.00	2.01
2	7	2.34	4.35	2.34	4.35
3	6	2.01	6.35	2.01	6.35
4	14	4.68	11.04	4.68	11.04
5	19	6.35	17.39	6.35	17.39
6	20	6.69	24.08	6.69	24.08
7	22	7.36	31.44	7.36	31.44
8	31	10.37	41.81	10.37	41.81
9	40	13.38	55.18	13.38	55.18
10	35	11.71	66.89	11.71	66.89
11	46	15.38	82.27	15.38	82.27
12	31	10.37	92.64	10.37	92.64
13	19	6.35	99.00	6.35	99.00
14	3	1.00	100.00	1.00	100.00
<NA>	0			0.00	100.00
Total	299	100.00	100.00	100.00	100.00

R OUTPUT

Frequencies
Cancer$Comorbids
Type: Integer

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
0	91	30.43	30.43	30.43	30.43
1	135	45.15	75.59	45.15	75.59
2	73	24.41	100.00	24.41	100.00
<NA>	0			0.00	100.00
Total	299	100.00	100.00	100.00	100.00

BAR PLOTS

= data frame name
= variable name

- Bar plots are histogram-like graphs for discrete or categorical variables (spaces between bars convey that the variable is not continuous)

```
# BAR PLOTS FOR DISCRETE VARIABLES  ----

# bar plots for discrete or categorical variable (ggplot2 package)
ggplot(Cancer, aes(x = factor(Diagnosis))) + geom_bar()
ggplot(Cancer, aes(x = factor(Gender))) + geom_bar()
```

DIAGNOSIS DISTRIBUTION

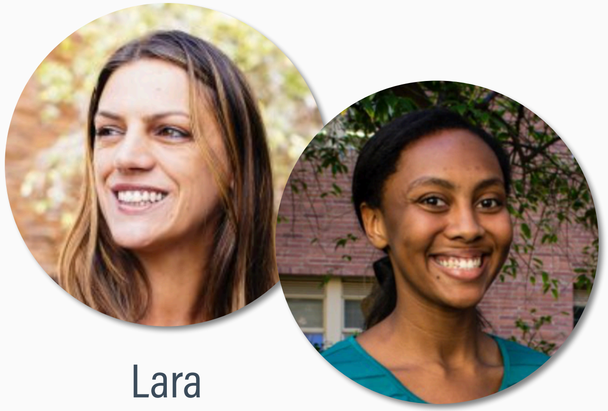
GENDER DISTRIBUTION

SMALL GROUP EXERCISE

Download two files from Bruin Learn: “Week 1 Lab. Describing Data.R” and “Week 1 Small Group Exercise.R”. The Lab script contains the R code we just discussed. The Exercise script contains only the URL for a different data set, ClinicalTrialData.csv. In groups of two or three, you will complete a series of R tasks that provide practice for the next assignment. There is no need to write code from scratch; instead, you can copy and paste code chunks from the Lab file into your Exercise script, modifying the data and variable names as needed. The ClinicalTrialData.csv file for this exercise contains data from a clinical trial investigating the impact of two medication regimens on smoking and drinking behavior.

SMOKING AND DRINKING CESSATION TRIAL

Pharmacological treatments that can concomitantly address cigarette smoking and heavy drinking stand to improve health care delivery for these highly prevalent co-occurring conditions. This superiority trial compared the combination of varenicline and naltrexone against varenicline alone for smoking cessation and drinking reduction among heavy-drinking smokers.



Lara Ray

ReJoyce Green

Ray, L.A., Green, R., Enders, C., et al. (2021). Efficacy of combining varenicline and naltrexone for smoking cessation and drinking reduction: A randomized clinical trial. *American Journal of Psychiatry*, *178*, 818–828.

SMALL GROUP EXERCISE TASK 1

- Use the provided URL to import the ClinicalTrialData.csv file into an R data frame (import method #3 from the Week 0 lab script).
- Use the dfSummary function to get numeric and visual summaries of the data frame's variables.

SMALL GROUP EXERCISE TASK 2

- Use the describe function to get descriptive statistics for all variables in the data frame.

SMALL GROUP EXERCISE TASK 3

- One of the main variables, breath carbon monoxide, is a biomarker used to objectively measure recent smoking behavior, as CO levels rise with tobacco inhalation and decrease after abstinence. Use the ggplot function to get a histogram and kernel density plot of breath CO at week 8 (COWeek8).

- How would you describe the shape of the COWeek8 distribution?

SMALL GROUP EXERCISE TASK 4

- Consider the skewness and kurtosis statistics for the COWeek8 variable. Interpret each statistic in practical terms (e.g., what it says about tail heaviness or asymmetry). How do these statistics align with your visual impression from the plot?

SMALL GROUP EXERCISE TASK 5

- Use the ggplot2 function to get histograms and kernel density plots separately for males and females.

- Suppose it is of interest to determine whether males and females differ in their smoking. Identify specific features of the distributions (e.g., location, spread, shape) that suggest differences between groups, as well as features that suggest similarities. Conclude with a brief statement about whether the visual evidence alone suggests there are differences.