

Parallel Computation Framework for Discrete Copula Modeling

Dhyey Mavani



Amherst College

Submitted to the Department of Mathematics and Statistics
of Amherst College in partial fulfillment of the requirements
for the degree of Bachelor of Arts with honors.

Advisor(s):
Professor Shu-Min Liao

November 1, 2024

Table of contents

Abstract	iv
Acknowledgements	v
1 Introduction	1
1.1 What's in this project directory?	1
1.2 Quarto workflow	1
1.3 Recommended reading before you begin editing	2
2 Unraveling Notion of Dependence through Copulas	4
2.1 A Motivating Example: Comparison of Dependence	6
2.2 Copulas	7
2.3 Figures	9
2.4 Tables	9
2.5 Chapter 2 Code	9
References	16
Appendices	17
A Code availability	17
B Corrections	20

Abstract

Understanding regression dependencies among discrete variables in categorical data—especially with ordinal responses—is a significant challenge in fields like finance, where the natural order of variables can unlock deeper insights into underlying distributions. While numerous model-based methods have been developed to examine these structures, there is a notable lack of flexible, model-free approaches. To address this gap, SCCRAM, a novel model-free measure based on the checkerboard copula, was introduced to identify and quantify regression dependence in multivariate categorical data involving both ordinal and nominal variables. Building upon this foundation, my thesis focuses on developing scalable implementations of SCCRAM in R and Python, utilizing parallel computing to enhance efficiency and accessibility for large-scale data analysis. Initial simulations confirm the effectiveness of these tools, providing researchers with a powerful resource for exploratory modeling and a deeper investigation into regression dependence structures within complex categorical datasets.

Acknowledgements

I would like to thank everyone who made my experience nothing short of amazing during my thesis journey. Firstly, I'd like to convey my gratitude to Professor Shu-Min Liao for advising me through this journey, and for believing in me to take on statistical software component of her most recent research work along with her collaborators. From my first research experience on campus building R-Blocks, she played a pivotal role in my journey.

I am also incredibly grateful to my college advisor, and statistics major advisor, Professor Nicholas Horton for always advocating for me and for supporting me throughout the Amherst College experience. I would also like to thank Professor Jun Ishii for teaching me Advanced Econometrics, which helped me gain a really clear understanding on the foundational tools on which I was able to build on in my work.

Finally, I would like to thank my friends and family for their constant belief in my abilities. Special thanks to my mom, dad, sister, grandfather, and grandmother for lending me and making me capable for this opportunity to study abroad. Last, but not the least, I would like to thank my friends, peers, and colleagues on campus, who took courses, worked, and played sports with me. This journey of academic, personal, and professional growth wouldn't be possible without their support.

Chapter 1

Introduction

XXX

1.1 What's in this project directory?

XXX

- fig XXX

i Note

XXX


1.2 Quarto workflow

1. XXX

- XXX

1.3 Recommended reading before you begin editing

XXX

 Caution

XXX

XXX

Chapter 2

Unraveling Notion of Dependence through Copulas

In this chapter, we introduce the concept of copulas and the necessary preliminary concepts, including probability transformations, quantile functions, and dependence measures. Let's start by defining some mathematical notation & objects that we will use throughout this chapter.

Define \mathbb{R} as the real line $(-\infty, \infty)$ and \mathbb{R}^2 as the real plane $\mathbb{R} \times \mathbb{R}$. A rectangle in the real plane is defined by the Cartesian product of two closed intervals: $[x_1, x_2] \times [y_1, y_2]$, where the vertices of the rectangle are $(x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2)$. We denote the unit interval as $\mathbb{I} = [0, 1]$, with $\mathbb{I}^2 = [0, 1] \times [0, 1]$ representing the unit square.

2.0.1 Probability Integral Transformation

Lemma 1 (Probability Integral Transformation):

Let F be a continuous distribution function, and let $X \sim F$. Then the transformed variable

$F(X)$ follows a standard uniform distribution, i.e., $F(X) \sim U(0, 1)$.

To illustrate this concept, let's generate a random sample from a normal distribution, apply the transformation, and plot the results.

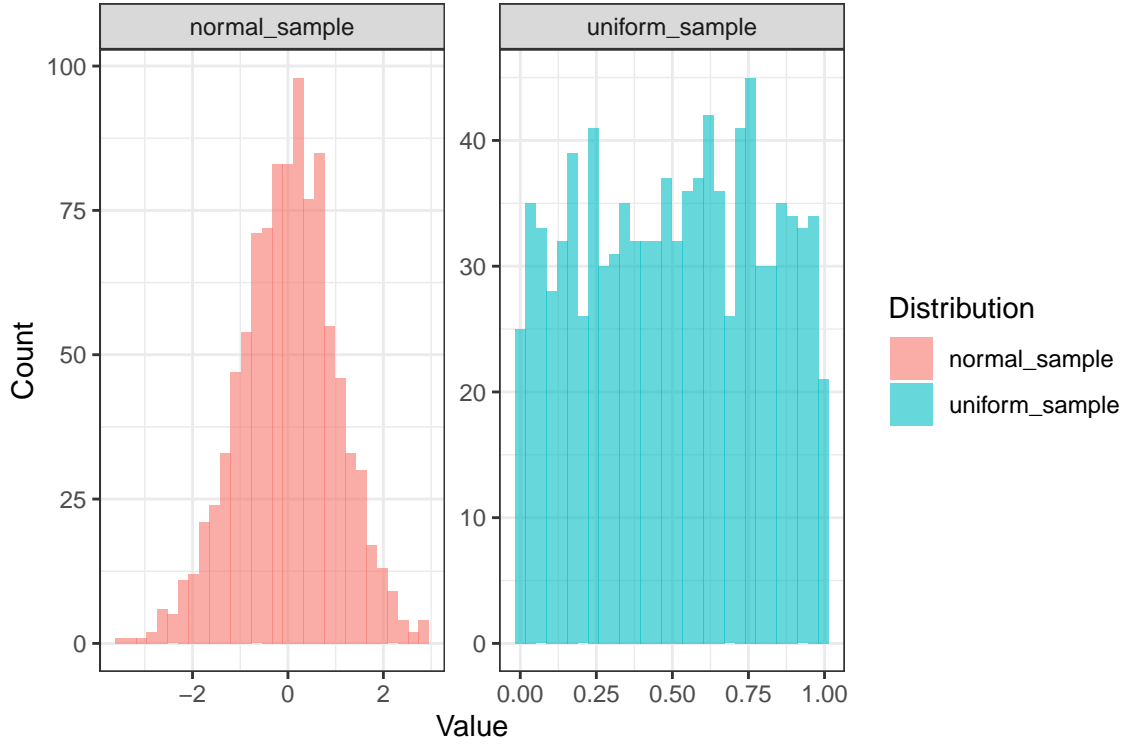


Figure 2.1: Probability Integral Transformation of a Normally Distributed Sample.

2.0.2 Quantile Function

The quantile function F^{-1} is defined as:

$$F^{-1}(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}, \quad y \in [0, 1].$$

For continuous and strictly increasing distribution functions F , $F^{-1} = F^{-1}$. However, if F is not strictly increasing, it may not have an inverse in the usual sense.

Lemma 2 (Quantile Transform):

Let $U \sim U(0, 1)$ and let F be any distribution function. Then $F^{-1}(U) \sim F$.

To demonstrate this, we can generate a uniform sample, apply the inverse transformation of a normal distribution, and compare the results.

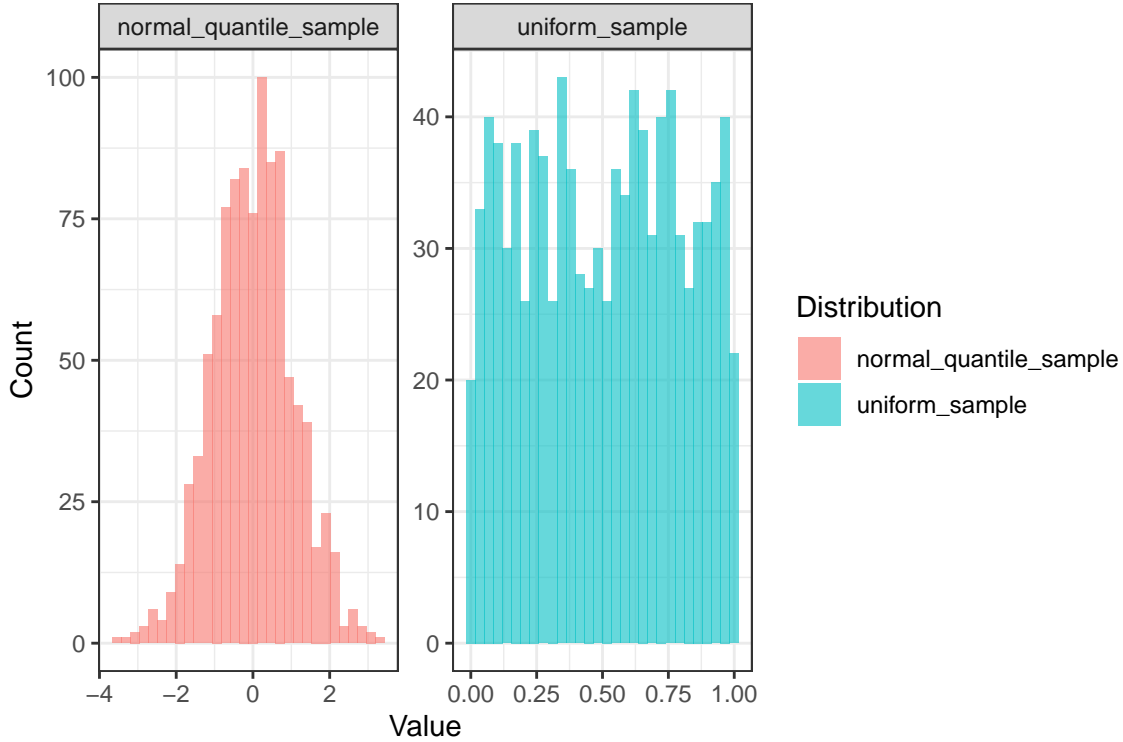


Figure 2.2: Quantile Transformation from Uniform to Normal Distribution.

2.1 A Motivating Example: Comparison of Dependence

Suppose we have two bivariate datasets, each consisting of 1000 independent observations from a bivariate random vector (X_1, X_2) and (Y_1, Y_2) , respectively. We aim to analyze the dependence between the components of each dataset.

To explore this, let's calculate and compare the Pearson correlation coefficient for each dataset.

Table 2.1: Pearson Correlation Coefficients for Bivariate Datasets (X_1, X_2) and (Y_1, Y_2) .

Variable Pair	Pearson Correlation
(X_1, X_2)	0.91956301
(Y_1, Y_2)	-0.05467525

2.1.1 Visualizing Dependence Structures

Let's plot scatterplots of (X_1, X_2) and (Y_1, Y_2) to visually examine the dependence structures.

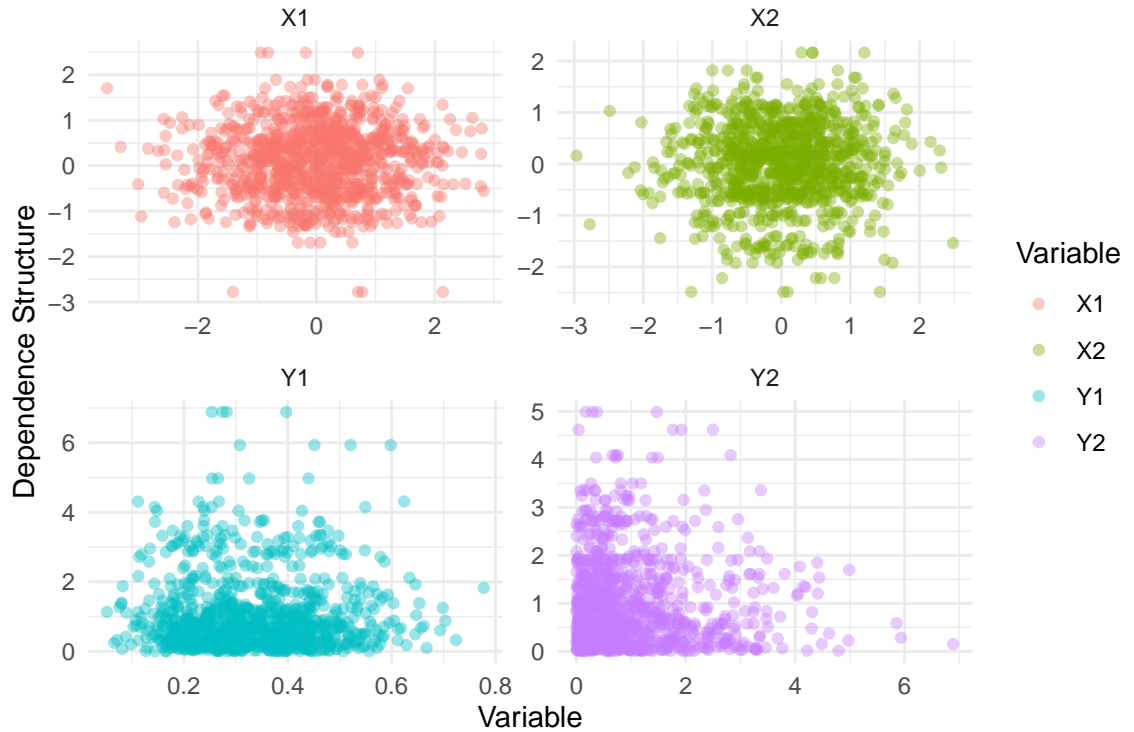


Figure 2.3: Scatterplots of (X_1, X_2) and (Y_1, Y_2) showing different dependence structures.

2.2 Copulas

Copulas enable the study of dependence independent of marginal distributions. A copula is a distribution function with standard uniform marginals. By transforming marginals to a standard uniform scale, copulas can be used to compare different dependency structures on an equal footing.

2.2.1 Subcopula Definition

A 2-dimensional subcopula C^S is defined on $D_1 \times D_2 \rightarrow [0, 1]$, where: 1. **Grounded**: $C^S(u, 0) = 0 = C^S(0, v)$ for all $u, v \in D_1 \times D_2$. 2. **2-increasing**: $C^S(u_2, v_2) - C^S(u_1, v_2) - C^S(u_2, v_1) + C^S(u_1, v_1) \geq 0$, for $u_1 \leq u_2$ and $v_1 \leq v_2$.

2.2.2 Visualizing Uniform Marginals using Copula Transform

To illustrate this concept, we can apply the copula transformation to both datasets and observe the transformed data. The transformed data for each dataset should now follow a standard uniform distribution on both axes.

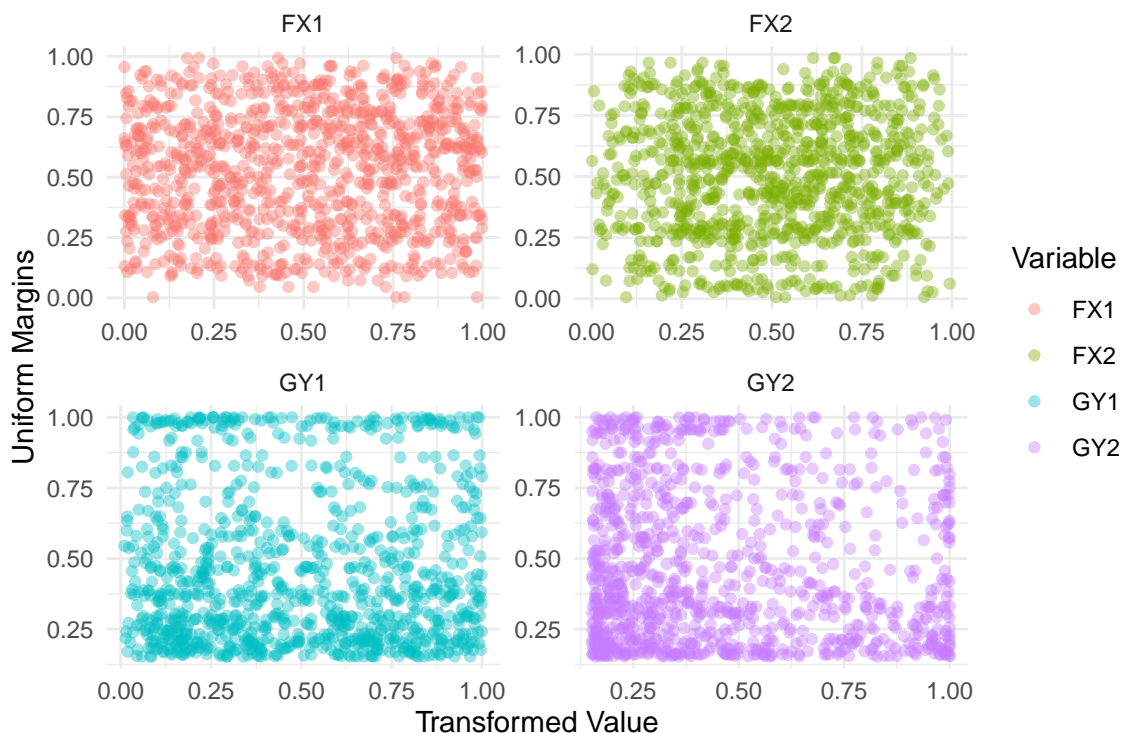


Figure 2.4: Copula Transformation on (X1, X2) and (Y1, Y2) to Uniform Margins.

2.3 Figures

Consider customizing your plot themes per-plot—as we do below to create `?@fig-wnba-ht`—or changing the default `ggplot()` theme in your document within your setup code chunks using `ggplot2::theme_set()`.

If **ggplot2** is loaded, the following code sets the default `ggplot()` theme to `theme_classic()`.

```
1 theme_set(theme_classic())
```

2.4 Tables

Your tables should be publication quality. Consider using [gt](#) (Iannone et al. 2024) or [kableExtra](#) (Zhu 2024) to customize your tables. The [gtsummary](#) package (Sjoberg et al. 2021) may also come in handy.

`?@tbl-ht-by-pos` shows the average heights of WNBA players by position.

2.5 Chapter 2 Code

The following code was used to create Chapter 2.

2.5.1 Code within chapter

```
1 # Load packages
2 library(tidyverse)
3 library(gt)
4
5 # Set default ggplot theme for document
```

```

6  theme_set(theme_classic())

7  # If using kableExtra tables, print blank cells instead of `NA`

8  options(knitr.kable.NA = "")

9  # Generate random normal sample

10 set.seed(8990)

11 normal_sample <- rnorm(1000)

12

13 # Apply transformation

14 uniform_sample <- pnorm(normal_sample)

15

16 # Plot both original and transformed samples

17 data.frame(normal_sample, uniform_sample) %>%

18   pivot_longer(cols = everything(), names_to = "Distribution", values_to = "Value") %>%

19   ggplot(aes(x = Value, fill = Distribution)) +

20   geom_histogram(bins = 30, alpha = 0.6, position = "identity") +

21   facet_wrap(~ Distribution, scales = "free") +

22   labs(x = "Value", y = "Count") +

23   theme_bw()

24 # Generate uniform sample

25 set.seed(8990)

26 uniform_sample <- runif(1000)

27

28 # Apply inverse transform (quantile function of normal)

```



```

29 normal_quantile_sample <- qnorm(uniform_sample)
30
31 # Plot both original and transformed samples
32 data.frame(uniform_sample, normal_quantile_sample) %>%
33   pivot_longer(cols = everything(), names_to = "Distribution", values_to = "Value") %>%
34   ggplot(aes(x = Value, fill = Distribution)) +
35   geom_histogram(bins = 30, alpha = 0.6, position = "identity") +
36   facet_wrap(~ Distribution, scales = "free") +
37   labs(x = "Value", y = "Count") +
38   theme_bw()
39 # Generate synthetic data
40 set.seed(8990)
41 x1 <- rnorm(1000)
42 x2 <- 0.7 * x1 + rnorm(1000, sd = 0.3)
43 y1 <- rbeta(1000, 5, 10)
44 y2 <- rexp(1000, rate = 1)
45
46 # Calculate correlation
47 correlations <- data.frame(
48   Pair = c("(X1, X2)", "(Y1, Y2)"),
49   Correlation = c(cor(x1, x2), cor(y1, y2))
50 )
51

```

```

52 correlations %>%
53   gt() %>%
54   cols_label(
55     Pair = "Variable Pair",
56     Correlation = "Pearson Correlation"
57   )
58 data.frame(X1 = x1, X2 = x2, Y1 = y1, Y2 = y2) %>%
59   pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
60   ggplot(aes(x = Value, y = ifelse(Variable %in% c("X1", "X2"), x2, y2), color = Variable)) +
61   geom_point(alpha = 0.4) +
62   facet_wrap(~ Variable, scales = "free") +
63   labs(x = "Variable", y = "Dependence Structure") +
64   theme_minimal()
65 # Apply copula transform
66 transformed_x1 <- pnorm(x1)
67 transformed_x2 <- pnorm(x2)
68 transformed_y1 <- pnorm(y1, mean = mean(y1), sd = sd(y1))
69 transformed_y2 <- pnorm(y2, mean = mean(y2), sd = sd(y2))
70
71 # Plot transformed datasets
72 data.frame(FX1 = transformed_x1, FX2 = transformed_x2, GY1 = transformed_y1, GY2 = transformed_y2) %>%
73   pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
74   ggplot(aes(x = Value, y = ifelse(Variable %in% c("FX1", "FX2"), transformed_x2, transformed_y2), color = Variable)) +

```

```

75   geom_point(alpha = 0.4) +
76   facet_wrap(~ Variable, scales = "free") +
77   labs(x = "Transformed Value", y =
78
79   "Uniform Margins") +
80   theme_minimal()
81   # =====
82   # Sample R script for thesis template
83   #
84   # Cleans temp_raw_wnba.csv dataset, which contains data pulled from
85   # https://www.espn.com/wnba/stats/player on 2024/06/19
86   #
87   # Last updated: 2024/06/19
88   # =====
89   library(tidyverse)
90
91   wnba <- read_csv("data/temp_raw_wnba.csv") |>
92     janitor::clean_names() |>
93     # Pull jersey numbers off of names and
94     # turn height text into msmt (6'4" = 6.3333)
95     mutate(jersey = str_extract(name, "[0-9]+$"),
96            name = str_remove(name, "[0-9]+$"),
97            ht_ft = parse_number(str_extract(ht, "^[0-9]")),

```

```

98     ht_in = parse_number(str_extract(ht, '[0-9]+\\\\"$')),
99     height = ht_ft * 12 + ht_in,
100    weight = parse_number(wt),
101    position = factor(pos,
102                      levels = c("G", "F", "C"),
103                      labels = c("Guard", "Forward", "Center")))) |>
104    select(-c(ht, wt, ht_ft, ht_in, pos))
105
106    save(wnba, file = "data/temp_wnba.RData")

```

2.5.2 Code sourced from external scripts

```

1  # =====
2  # Sample R script for thesis template
3  #
4  # Cleans temp_raw_wnba.csv dataset, which contains data pulled from
5  # https://www.espn.com/wnba/stats/player on 2024/06/19
6  #
7  # Last updated: 2024/06/19
8  # =====
9  library(tidyverse)
10
11 wnba <- read_csv("data/temp_raw_wnba.csv") |>
12   janitor::clean_names() |>
13   # Pull jersey numbers off of names and

```

```

14   # turn height text into msmt (6'4" = 6.3333)
15   mutate(jersey = str_extract(name, "[0-9]+$"),
16          name = str_remove(name, "[0-9]+$"),
17          ht_ft = parse_number(str_extract(ht, "^[0-9]")),
18          ht_in = parse_number(str_extract(ht, "[0-9]+\\\"$")),
19          height = ht_ft * 12 + ht_in,
20          weight = parse_number(wt),
21          position = factor(pos,
22                             levels = c("G", "F", "C"),
23                             labels = c("Guard", "Forward", "Center")))) |>
24   select(-c(ht, wt, ht_ft, ht_in, pos))
25
26   save(wnba, file = "data/temp_wnba.RData")

```

References

- Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., Seo, J., Brevoort, K., and Roy, O. (2024), *Gt: Easily create presentation-ready display tables*.
- Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A., and Larmarange, J. (2021), “Reproducible summary tables with the gtsummary package,” *The R Journal*, 13, 570–580. <https://doi.org/10.32614/RJ-2021-053>.
- Ushey, K., and Wickham, H. (2024), *Renv: Project environments*.
- Zhu, H. (2024), *kableExtra: Construct complex table with 'kable' and pipe syntax*.

Appendix A

Code availability

This thesis is written using Quarto with **renv** (Ushey and Wickham 2024) to create a reproducible environment. All materials (including the data sets and source files) required to reproduce this document can be found at the Github repository github.com/GITHUB-USERNAME/THESIS-REPO-NAME.

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

```
1 # =====
2 # Sample R script for thesis template
3 #
4 # Cleans temp_raw_wnba.csv dataset, which contains data pulled from
5 # https://www.espn.com/wnba/stats/player on 2024/06/19
6 #
7 # Last updated: 2024/06/19
```

```

8 # =====
9 library(tidyverse)
10
11 wnba <- read_csv("data/temp_raw_wnba.csv") |>
12   janitor::clean_names() |>
13   # Pull jersey numbers off of names and
14   # turn height text into msmt (6'4" = 6.3333)
15   mutate(jersey = str_extract(name, "[0-9]+$"),
16          name = str_remove(name, "[0-9]+$"),
17          ht_ft = parse_number(str_extract(ht, "^[0-9]")),
18          ht_in = parse_number(str_extract(ht, '[0-9]+\\\\"$')),
19          height = ht_ft * 12 + ht_in,
20          weight = parse_number(wt),
21          position = factor(pos,
22                             levels = c("G", "F", "C"),
23                             labels = c("Guard", "Forward", "Center")))) |>
24   select(-c(ht, wt, ht_ft, ht_in, pos))
25
26 save(wnba, file = "data/temp_wnba.RData")

```

```

1 # =====
2 # Sample R script for thesis template
3 #
4 # Doesn't do anything useful

```



```
5  #  
6  # Last updated: 2024/08/24  
7  # =====  
8  
9  print("Hello, Amherst!")
```

Appendix B

Corrections

This section may be excluded if no corrections are made to your thesis after initial submission to the department and before final submission to the college.

Per the [Statistics Honors Thesis Regulations](#):

Corrections to theses may be made after the date on which they are due in the Department's hands. Corrections may be made to the body of the thesis, but every such correction will be acknowledged in a list under the heading "Corrections," along with the statement "When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected." This list will be given on a sheet or sheets to be appended to the thesis. Corrections to spelling, grammar, or typography may be acknowledged by a general statement such as "30 spellings were corrected in various places in the thesis, and the notation for definite integral was changed in approximately 10 places." However, any correction that affects the meaning of a sentence or paragraph should be described in careful detail, and substantial

additions to the thesis will not be allowed. Questions about what should appear in the “Corrections” should be directed to the Chair. Electronic versions of the thesis, technical appendix, and necessary data and supplemental files must all be updated at the time of correction as well.

When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected.

1. ...
2. ...