---
title: "NSW-500"
author: "Paris Heard"
date: "`r Sys.Date()`"
output: github_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Learning Objectives

-   Understand the relationship between reported effective collaboration and remote work
-   Explore the relationship between leadership, team culture, and the effectiveness of collaboration
-   Gain insight into the desires of the average employee, by industry

Package Installations

```{r}
set.seed(2)

# EDA
library(tidyverse)
library(ggplot2)
library(dplyr)
library(reshape2)

# Regression
library(infer)
library(MASS)
library(foreign)
library(moderndive)
library(Hmisc)
library(reshape2)
library(mosaicCore)
library(car)

# Plotting
library(HH)
library(likert)
library(colorspace)
library(RColorBrewer)
library(grid)

# Output/Resources
library(knitr)
library(readr)
library("data.table")
```

Creating Data-frame

```{r cars}
sdata <- read.csv("data/data.csv", header = TRUE, sep = ",")
```

Cleaning Data

```{r}
head(sdata)
```

Crafting a new data frame for only those employees who have worked remotely in the past six months, and removing all ID numbers.

```{r}
recent_RW <- sdata %>%
  filter(worked_RW_past_six_months == "Yes") %>%
  dplyr::select(birth_year, gender, yr_experience, household_size, metro_regional,
frequency_RW, perc_required_onsite, six_months_collaborate_easily_RW,
hybrid_biggest_barriers_one, hybrid_smallest_barriers_one, feel_better_when_RW,
feel_better_seeing_colleagues_onsite, more_active_when_RW, team_works_well_RW,
productivity_RW, easy_to_contact_employees_RW, pared_collab, pared_wellness) %>%
  mutate(across(where(~all(str_detect(.,"%"))), parse_number))

head(recent_RW)
```

Plotting for EDA (Exploratory Data Analysis)

Remote Work Frequency & Remote Team Collaboration

In order to visualize the collaboration responses in likert format, we will need to do a bit of data wrangling. Let's select our target variables and pivot the likert responses to a wider format. The code below selects our target variables, groups them, and then pivots our likert responses wider- one column for each response, and the cells populated with counts, grouped by remote work frequency. Finally, we will sort this data-frame in descending order.

```{r}
collab_freq <- recent_RW %>%
  dplyr::select(six_months_collaborate_easily_RW, frequency_RW) %>%
  group_by(six_months_collaborate_easily_RW, frequency_RW) %>%
  pivot_wider(names_from = "six_months_collaborate_easily_RW",
              values_from = "six_months_collaborate_easily_RW",
              values_fn = function(x) sum(!is.na(x)), values_fill = 0) %>%
  arrange(desc(frequency_RW))

collab_freq_use <- collab_freq %>%
  dplyr::relocate(6, 2, 4, 3, 5)
```

Next, let's visualize this data! Using the **HH** package, we'll create a likert plot.

As observed, the reported ease of remote collaboration is split, however there is a larger percentage of responses reflecting that collaboration was ***not*** easier when working remotely. Further, there appears to be a larger number of responses as the frequency of remote work decreases, indicating a shift back to in-person work. Additionally, it seems that those who work remotely less tend to report greater frustrations with remote collaboration.

```{r}
coloring <- likertColor(nc=5, ReferenceZero=NULL,
                          colorFunction="diverge_hcl",
                          colorFunctionArgs= list(h=c(327, 164), c=100, l=c(75,95,100),
power=1.5))

likert1 <- HH::likert(frequency_RW~., collab_freq_use, ReferenceZero=3, ylab = "Remote
Work Frequency", main =
list("Ease of Remote Collaboration", x=unit(.62, "npc")), auto.key = list(columns = 2,
reverse.rows = T), colorFunction = "diverge_hcl", col = coloring, data.order=TRUE)

likert1

png("../NSW-500/output/likert_1.png", height=720, width=1080)
```

````
likert1
dev.off()
```

```{r}
use <- collab_freq_use %>%
  dplyr::relocate(frequency_RW)

# chart design elements
neutral_color <- "gray90"
my_breaks <- seq(-100, 100, 10)
my_vline  <- geom_vline(xintercept = my_breaks, color = "white", size = 0.25)
my_hline  <- geom_hline(yintercept = my_breaks, color = "white", size = 0.25)

# ggplot() theme settings
my_theme_elements <- theme(panel.background = element_blank(),
                           legend.key.size = unit(4, "mm"),
                           legend.title = element_blank(),
                           axis.ticks = element_blank(),
                           legend.justification = 0.5,
                           legend.position = "top")

# labeling vectors
opinion_labels <- c("Strongly Disagree",
                    "Disagree",
                    "Neutral",
                    "Agree",
                    "Strongly Agree")

# question_labels <- c("Beyond the content",
#                     "Analyze errors",
#                     "Provide facts",
#                     "Develop writing",
#                     "Independent learning")

# rename opinion columns
setnames_opinion_labels <- function(x) {
  setnames(use,
           old = c("Strongly disagree", "Somewhat disagree", "Neither agree nor disagree",
"Somewhat agree", "Strongly agree"),
           new = opinion_labels,
           skip_absent = TRUE)
}

setnames(use, "frequency_RW", "Item", skip_absent = TRUE)

# create the likert list
likert_list <- likert(summary = use)

# set scale limits
my_limits <- c(0, 100)

# recode the opinion options
setnames_opinion_labels(likert_list$results)

# create the chart
plot(likert_list,
     centered = FALSE,              # 100% stacked bars
     include.center  = TRUE,        # include neutral
     plot.percent.low     = FALSE,
     plot.percent.neutral = FALSE,
     plot.percent.high    = FALSE) +
  scale_y_continuous(limits = my_limits,
````

```
                     breaks = my_breaks,
                     sec.axis = sec_axis( # second scale
                       transform = function(z) z - 100,
                       breaks = my_breaks,
                       labels = as.character(abs(my_breaks)))) +
  my_theme_elements +
  my_hline
```

Remote Work Frequency & Employee Wellness

```{r}
wellness_freq <- recent_RW %>%
  dplyr::select(feel_better_when_RW, frequency_RW) %>%
  group_by(feel_better_when_RW, frequency_RW) %>%
  pivot_wider(names_from = "feel_better_when_RW",
              values_from = "feel_better_when_RW",
              values_fn = function(x) sum(!is.na(x)), values_fill = 0) %>%
  arrange(desc(frequency_RW))

wellness_freq_use <- wellness_freq %>%
  relocate(6, 5, 2, 4, 3)
```

```{r}
likert2 <- HH::likert(frequency_RW~., wellness_freq_use, ReferenceZero=3, ylab = "Remote
Work Frequency", main =
list("Improved Wellness while Working Remotely", x=unit(.62, "npc")), auto.key =
list(columns = 2,
reverse.rows = T), colorFunction = "diverge_hcl", col = coloring, data.order=TRUE)

likert2

png("../NSW-500/output/likert_2.png", height=720, width=1080)

likert2
dev.off()
```

Remote Work Frequency & Remote Team Collaboration/Biggest Barriers to Remote Work

Next, we'll analyze ease of remote collaboration in relation to remote work frequency and their reported largest barrier to working remotely.

```{r}
likert_gg_1 <- ggplot(recent_RW, aes(x = six_months_collaborate_easily_RW, y =
frequency_RW, color = hybrid_biggest_barriers_one)) +
  geom_boxplot(size = 0.75) +
  geom_jitter(alpha = 0.5) +
  facet_wrap(~hybrid_biggest_barriers_one) +
  theme(axis.text.x = element_text(angle=55, hjust=1, vjust=1), panel.spacing.x = unit(4,
"lines")) +
  labs(x = "Was collaboration easy while working remotely?", y = "Remote work frequencey
(%)", color = "Biggest barrier while working remotely")

likert_gg_1

png("../NSW-500/output/likert_gg_pplot_1.png", height=720, width=1080)

likert_gg_1
dev.off()

```

Remote Team Collaboration & Biggest Barriers to Remote Work

Next, we'll analyze ease of remote collaboration in relation to their reported largest barrier to working remotely.

```{r}
likert_gg_2 <- ggplot(recent_RW, aes(x = six_months_collaborate_easily_RW, fill =
hybrid_biggest_barriers_one)) +
  geom_bar() +
#  facet_wrap(~frequency_RW) +
  theme(axis.text.x = element_text(angle=55, hjust=1, vjust=1), panel.spacing.x = unit(4,
"lines")) +
  labs(title = "Responses to Ease of Remote Collaboration", subtitle = "Ranked by Largest
Reported Barrier", x = "Was collaboration easy while working remotely?", y = "Number of
Responses", fill = "Biggest barrier while working remotely")

likert_gg_2

png("../NSW-500/output/likert_gg_pplot_2.png", height=720, width=1080)

likert_gg_2
dev.off()
```

Now, we'll analyse the relationship between collaboration, wellness, and colleague interaction.

```{r}
likert_gg_3 <- ggplot(recent_RW, aes(x = six_months_collaborate_easily_RW, fill =
feel_better_when_RW)) +
  geom_bar() +
  facet_wrap(~feel_better_seeing_colleagues_onsite) +
  theme(axis.text.x = element_text(angle=55, hjust=1, vjust=1), panel.spacing.x = unit(4,
"lines")) +
  labs(title = "Responses to Ease of Remote Collaboration", subtitle = "Ranked by
Agreeance: Do you feel better seeing your colleagues onsite?", x = "Was collaboration easy
while working remotely?", y = "Number of Responses", fill = "Do you feel better when
working remotely?")

likert_gg_3

png("../NSW-500/output/likert_gg_pplot_3.png", height=720, width=1080)

likert_gg_3
dev.off()
```

Ordinal Logistic Regression - Remote Work Frequency and Collaboration

Slice New Data Frame

First, we need to grab a section of our original data frame for this usage. We'll select our target variables, remote work frequency and ease of collaboration, and rename them freq_RW and collaborate, respectively. Display the head of the data to ensure it populated correctly.

```{r}
regdata <- recent_RW %>%
  dplyr::select(frequency_RW,
                six_months_collaborate_easily_RW,
                feel_better_when_RW) %>%
  rename(freq_RW = frequency_RW,
         collaborate = six_months_collaborate_easily_RW,
         wellness = feel_better_when_RW)
```

```
head(regdata)
```

Mutate & Factor

Next, we will mutate our collaborate variable to ensure it is in factor form, and in the correct ordering. We'll convert it to a factor and give it the correct levels, beginning with "Strongly disagree" and ending with "Strongly agree". We will do the same to the freq_RW variable, with the levels beginning at "10" (10% remote work frequency) and ending with "100". We'll display the levels of these two variables to ensure that they were successfully manipulated.

```{r}
regdata <- mutate_at(regdata, vars(collaborate), as.factor)
regdata$collaborate <- factor(regdata$collaborate,
                              levels = c("Strongly disagree",
                                         "Somewhat disagree",
                                         "Neither agree nor disagree",
                                         "Somewhat agree",
                                         "Strongly agree"))

regdata <- mutate_at(regdata, vars(freq_RW), as.factor)
regdata$freq_RW <- factor(regdata$freq_RW,
                          levels = c("10", "20", "30", "40", "50",
                                     "60", "70", "80", "90", "100"))

regdata <- mutate_at(regdata, vars(wellness), as.factor)
regdata$wellness <- factor(regdata$wellness,
                              levels = c("Strongly disagree",
                                         "Somewhat disagree",
                                         "Neither agree nor disagree",
                                         "Somewhat agree",
                                         "Strongly agree"))
levels(regdata$collaborate)
levels(regdata$wellness)
levels(regdata$freq_RW)
```

Analyze

Quickly, we will table this data, to ensure that our values are still valid after factor conversion.

```{r}

lapply(regdata[, c("collaborate", "freq_RW")], table)
```

Regression

First, we will fit our model. Collaborate (ease of remote collaboration) will be our outcome variable, while freq_RW (remote work frequency) will be our exploratory variable. We will pull this data from the data frame we created earlier, and mark Hess=TRUE to make summary calls more efficient. P-values are not included in this summary.

```{r}
m <- MASS::polr(collaborate ~ freq_RW, data = regdata, Hess=TRUE)
summary(m)
```

Intercepts

The outcome variable, collaborate, has five levels and four intercepts. These are stored in a variable called zeta within the regression model. They are log-odds of cumulative probabilities, which we will compare to the raw cumulative probabilities at the freq_RW predictor level of 100.

```{r}
ilogit(m$zeta)
```

The first intercept above represents the log odds of *P* (Strongly disagree ≤ Somewhat disagree / *P* (Strongly disagree \> Somewhat disagree). This is similar for the remainder of the intercepts.

```{r}
cumsum(
  prop.table(
    table(regdata$collaborate[regdata$freq_RW == 100])
  )
)
```

As a result, we may conclude that the estimated proportion of individuals who

-   "Strongly disagree" is 0.03

-   "Somewhat disagree" or "Neither agree nor disagree" is 0.15

-   "Neither agree nor disagree" or "Somewhat agree" is 0.31

-   "Somewhat agree" or "Strongly agree" is 0.73

The predictor variable at reference "100" indicates the change in this probability for each one-unit increase in that predictor.

Predictor Coefficients

Now, we'll compute a 95% confidence interval for the OR. OR \> 1 signifies the risk of greater probability of increased levels of our outcome variable, or collaborate.

```{r}
CI <- confint(m)

data.frame(
  OR = exp(m$coefficients),
  lower = exp(CI[,1]),
  upper = exp(CI[,2])
)
```

As evidenced above, we can observe slightly significant OR levels greater than 1 for the following remote work frequencies: 30%, 60%, 70%, 80%, 90%, and 100%. The 100% level is extremely significant at 2.18, implying that those who work entirely from home are more likely to answer at a higher level of agreeableness regarding remote work collaboration.

This could be due to their increased use of remote technologies. Interestingly, those who work hybrid, or 50%, were not more likely to agree on ease of collaboration. Those who work from home 100% have 118% greater odds of agreeableness.

Calculate P-Value

```{r}
adjm <- MASS::polr(collaborate ~ freq_RW,
                  data = regdata,
                  Hess = T)
```

```{r}
CIp <- confint(adjm)
TSTAT <- summary(adjm)$coef[1:nrow(CI), "t value"]

data.frame(
  AOR = exp(adjm$coefficients),
  lower = exp(CI[,1]),
  upper = exp(CI[,2]),
  p = 2*pnorm(abs(TSTAT), lower.tail = F)
)
```

Using a significance level of 0.05, we have **two** statistically significant results:

- 50% RW: p-value = 0.04

- 100% RW: p-value = 0.01

If we wanted to use more than one predictor variable, we could add it to our adjusted regression model above and compute an ANOVA chart (below).

```{r}
car::Anova(adjm, type = 3)
```

Ordinal Logistic Regression - Remote Work Frequency and Collaboration

Slice New Data Frame

First, we need to grab a section of our original data frame for this usage. We'll select our target variables, remote work frequency and increased wellness working from home, and rename them freq_RW and wellness, respectively. Display the head of the data to ensure it populated correctly.

```{r}
regdata1 <- recent_RW %>%
  dplyr::select(frequency_RW,
                feel_better_when_RW) %>%
  rename(freq_RW = frequency_RW,
         wellness = feel_better_when_RW)

head(regdata1)
```

Mutate & Factor

Next, we will mutate our wellness variable to ensure it is in factor form, and in the correct ordering. We'll convert it to a factor and give it the correct levels, beginning with "Strongly disagree" and ending with "Strongly agree". We will do the same to the freq_RW variable, with the levels beginning at "10" (10% remote work frequency) and ending with "100". We'll display the levels of these two variables to ensure that they were successfully manipulated.

```{r}
regdata1 <- mutate_at(regdata1, vars(freq_RW), as.factor)
regdata1$freq_RW <- factor(regdata1$freq_RW,
                           levels = c("10", "20", "30", "40", "50",
                                      "60", "70", "80", "90", "100"))

regdata1 <- mutate_at(regdata1, vars(wellness), as.factor)
regdata1$wellness <- factor(regdata1$wellness,
                             levels = c("Strongly disagree",
                                        "Somewhat disagree",
                                        "Neither agree nor disagree",
                                        "Somewhat agree",
```

```
                                         "Strongly agree"))
levels(regdata1$wellness)
levels(regdata1$freq_RW)
```

Analyze

Quickly, we will table this data, to ensure that our values are still valid after factor conversion.

```{r}
lapply(regdata1[, c("wellness", "freq_RW")], table)
```

Regression

First, we will fit our model. Wellness (increased wellness working from home) will be our outcome variable, while freq_RW (remote work frequency) will be our exploratory variable. We will pull this data from the data frame we created earlier, and mark Hess=TRUE to make summary calls more efficient. P-values are not included in this summary.

```{r}
m1 <- MASS::polr(wellness ~ freq_RW, data = regdata1, Hess=TRUE)
summary(m1)
```

Intercepts

The outcome variable, wellness, has five levels and four intercepts. These are stored in a variable called zeta within the regression model. They are log-odds of cumulative probabilities, which we will compare to the raw cumulative probabilities at the freq_RW predictor level of 100.

```{r}
ilogit(m1$zeta)
```

The first intercept above represents the log odds of *P* (Strongly disagree ≤ Somewhat disagree / *P* (Strongly disagree \> Somewhat disagree). This is similar for the remainder of the intercepts.

```{r}
cumsum(
  prop.table(
    table(regdata1$wellness[regdata1$freq_RW == 100])
  )
)
```

As a result, we may conclude that the estimated proportion of individuals who

-   "Strongly disagree" is 0.02

-   "Somewhat disagree" or "Neither agree nor disagree" is 0.14

-   "Neither agree nor disagree" or "Somewhat agree" is 0.49

-   "Somewhat agree" or "Strongly agree" is 0.80

The predictor variable at reference "100" indicates the change in this probability for each one-unit increase in that predictor.

Predictor Coefficients

Now, we'll compute a 95% confidence interval for the OR. OR \> 1 signifies the risk of greater probability of increased levels of our outcome variable, or wellness.

```{r}
CI_1 <- confint(m1)

data.frame(
  OR = exp(m1$coefficients),
  lower = exp(CI_1[,1]),
  upper = exp(CI_1[,2])
)
```

As evidenced above, we can observe slightly significant OR levels greater than 1 for *all* work frequencies. This indicates that across all remote work frequencies, **employees perceive their wellness to increase when working from home.** The 70% and 100% levels are extremely significant at 2.18 and 2.82, implying that those who work mostly or entirely from home are much more likely to answer at a higher level of agreeableness regarding remote wellness.

This could be due to a perceived sense of work-life balance and freedom.

Calculate P-Value

```{r}
adjm1 <- MASS::polr(wellness ~ freq_RW,
                  data = regdata1,
                  Hess = T)

CIp1 <- confint(adjm1)
TSTAT1 <- summary(adjm1)$coef[1:nrow(CI_1), "t value"]

data.frame(
  AOR = exp(adjm$coefficients),
  lower = exp(CI_1[,1]),
  upper = exp(CI_1[,2]),
  p = 2*pnorm(abs(TSTAT1), lower.tail = F)
)
```

Using a significance level of 0.05, we have **three** statistically significant results:

-   50% RW: p-value = 0.005

-   70% RW: p-value = 0.03

-   100% RW: p-value = 0.0007 (**WOW**!)

If we wanted to use more than one predictor variable, we could add it to our adjusted regression model above and compute an ANOVA chart (below).

```{r}
car::Anova(adjm1, type = 3)
```

Spearman's Rank Coefficient

```{r}
rank <- recent_RW %>%
  dplyr::select(frequency_RW, six_months_collaborate_easily_RW, pared_collab,
feel_better_when_RW, pared_wellness, productivity_RW) %>%
  rename(freq_RW = frequency_RW,
         collaborate = six_months_collaborate_easily_RW,
         prod = productivity_RW,
```

         wellness = feel_better_when_RW)
head(rank)
```

***Null hypothesis:*** There is no significant relationship between the level of ease of remote collaboration and the level of perceived increased wellness working remotely.

***Alternative hypothesis:*** There is a significant relationship between the level of ease of remote collaboration and the level of perceived increased wellness working remotely.

***Result:*** Reject the null hypothesis. There is a significant positive weak relationship.

```{r}
cor.test(rank$pared_collab, rank$pared_wellness, method = c("spearman"))
```

```{r}
ggplot(data = rank, aes(x = pared_collab, y = pared_wellness)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y~x)
```

Null hypothesis: There is no significant relationship between the level of perceived increased wellness working remotely and the frequency of remote work.

Alternative hypothesis: There is a significant relationship between the level of perceived increased wellness working remotely and the frequency of remote work.

Result: Reject the null hypothesis. There is a significant positive very weak relationship.

```{r}
cor.test(rank$pared_collab, rank$freq_RW, method = c("spearman"))
```

```{r}
ggplot(data = rank, aes(x = pared_collab, y = freq_RW)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y~x)
```

Null hypothesis: There is no significant relationship between the level of ease of remote collaboration and the frequency of remote work.

Alternative hypothesis: There is a significant relationship between the level of ease of remote collaboration and the frequency of remote work.

Result: Reject the null hypothesis. There is a significant positive very weak relationship.

```{r}
cor.test(rank$pared_wellness, rank$freq_RW, method = c("spearman"))
```

```{r}
ggplot(data = rank, aes(x = pared_wellness, y = freq_RW)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y~x)
```

Output Data for Datawrapper
```

````
```{r}
write.csv(collab_freq_use,"~/Downloads/collab_freq_use.csv", row.names = TRUE)
```
````