## Data Wrangling Exercise 1: Basic Data Manipulation

"Data Wrangling Excercise 1: Refine" library(dplyr) library(tidyr) #0: Load the data in RStudio purchase2 <- read.csv("refine\_original.csv")

# 1: Clean up brand names

# create a vector of company names

company\_names <- purchase2$company

# identify the vector positions for the companies

phillips\_names <- grep("p|f", company\_names, ignore.case = TRUE) akzo\_names <- grep("^a", company\_names, ignore.case = TRUE) van\_houten\_names <- grep("^v", company\_names, ignore.case = TRUE) unilever\_names <- grep("^u", company\_names, ignore.case = TRUE)

# replace with correct names

company\_names[phillips\_names] <- "Phillips" company\_names[akzo\_names] <- "Akzo" company\_names[van\_houten\_names] <- "Van Houten" company\_names[unilever\_names] <- "Unilever"

# insert back into data frame

purchase2$company <- company\_names purchase2

# 2: Separate product code and number

purchase\_df <- separate(purchase2, Product.code...number, c("product\_code", "product\_number"), sep = "-", remove = TRUE, convert = FALSE, extra = "warn", fill = "warn")

# 3: Add product categories

product\_category <- factor(purchase\_df$product\_code, c("p", "v", "x","q"),labels = c("Smartphone", "TV", "Laptop", "Tablet"))

purchase\_df <- cbind(purchase\_df, product\_category)

# 4: Add full address for geocoding

full\_address <- as.character(paste(purchase\_dfcity, purchase\_df$country, sep = ",")) purchase\_df <- cbind(purchase\_df, full\_address)

# 5: Create dummy variables for company and product category

company\_phillips <- factor(purchase\_df$company, "Phillips",labels = "1,1") purchase\_df <- cbind(purchase\_df, company\_phillips) company\_azko <- factor(purchase\_df$company, "Azko",labels = "1,0") purchase\_df <- cbind(purchase\_df, company\_azko) company\_van\_houten <- factor(purchase\_df$company, "Van Houten",labels = "0,1") purchase\_df <- cbind(purchase\_df, company\_van\_houten) company\_unilever <- factor(purchase\_df$company, "Unilever",labels = "0,0") purchase\_df <- cbind(purchase\_df, company\_unilever) product\_smartphone <- factor(purchase\_df$product\_category, "Smartphone",labels = "1,1") purchase\_df <- cbind(purchase\_df, product\_smartphone) product\_tv <- factor(purchase\_df$product\_category, "TV",labels = "1,0") purchase\_df <- cbind(purchase\_df, product\_tv) product\_laptop <- factor(purchase\_df$product\_category, "Laptop",labels = "0,1") purchase\_df <- cbind(purchase\_df, product\_laptop) product\_tablet <- factor(purchase\_df$product\_category, "Tablet",labels = "0,0") purchase\_df <- cbind(purchase\_df, product\_tablet)

write.table(purchase\_df, file = "refine\_clean.csv", sep = ",") write.csv(purchase\_df, file = "refine\_clean.csv")