# Midterm Exam

Mike Kozlowski

2023-04-04

Load the data file “IBM_HR-Employee-Attrition.csv”

Generate the summary and determine if there are NAs in the data set.

Create a bar plot that shows the continuous variables DailyRate, MonthlyIncome, MonthlyRate on a single plot, the y-axis should be log transformed.

Hint: There is an example of this is the powerpoints, early in the semester, I think, or a google search on “r plot multiple boxplot in one graph. You will also need to do some digging in ggplot to figure out how to log transform the y axis

```
ibm_infile="F:\\Chrome Downloads\\IBM_HR-Employee-Attrition.csv"
ibmdata=read.csv(ibm_infile,stringsAsFactors=TRUE)
summary(ibmdata)
```

```
##       Age        Attrition            BusinessTravel   DailyRate     
##  Min.   :18.00   No :1233   Non-Travel       : 150   Min.   : 102.0  
##  1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.: 465.0  
##  Median :36.00              Travel_Rarely    :1043   Median : 802.0  
##  Mean   :36.92                                       Mean   : 802.5  
##  3rd Qu.:43.00                                       3rd Qu.:1157.0  
##  Max.   :60.00                                       Max.   :1499.0  
##                                                                      
##                    Department  DistanceFromHome   Education    
##  Human Resources       : 63   Min.   : 1.000   Min.   :1.000  
##  Research & Development:961   1st Qu.: 2.000   1st Qu.:2.000  
##  Sales                 :446   Median : 7.000   Median :3.000  
##                               Mean   : 9.193   Mean   :2.913  
##                               3rd Qu.:14.000   3rd Qu.:4.000  
##                               Max.   :29.000   Max.   :5.000  
##                                                                
##            EducationField EmployeeCount EmployeeNumber   EnvironmentSatisfaction
##  Human Resources : 27    Min.   :1     Min.   :   1.0   Min.   :1.000          
##  Life Sciences   :606    1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000          
##  Marketing       :159    Median :1     Median :1020.5   Median :3.000          
##  Medical         :464    Mean   :1     Mean   :1024.9   Mean   :2.722          
##  Other           : 82    3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000          
##  Technical Degree:132    Max.   :1     Max.   :2068.0   Max.   :4.000          
##                                                                                
##     Gender      HourlyRate     JobInvolvement    JobLevel    
##  Female:588   Min.   : 30.00   Min.   :1.00   Min.   :1.000  
##  Male  :882   1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000  
##               Median : 66.00   Median :3.00   Median :2.000  
##               Mean   : 65.89   Mean   :2.73   Mean   :2.064  
##               3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000  
##               Max.   :100.00   Max.   :4.00   Max.   :5.000  
##                                                              
##                       JobRole    JobSatisfaction  MaritalStatus MonthlyIncome  
##  Sales Executive          :326   Min.   :1.000   Divorced:327   Min.   : 1009  
##  Research Scientist       :292   1st Qu.:2.000   Married :673   1st Qu.: 2911  
##  Laboratory Technician    :259   Median :3.000   Single  :470   Median : 4919  
##  Manufacturing Director   :145   Mean   :2.729                  Mean   : 6503  
##  Healthcare Representative:131   3rd Qu.:4.000                  3rd Qu.: 8379  
##  Manager                  :102   Max.   :4.000                  Max.   :19999  
##  (Other)                  :215                                                 
##   MonthlyRate    NumCompaniesWorked Over18   OverTime   PercentSalaryHike
##  Min.   : 2094   Min.   :0.000      Y:1470   No :1054   Min.   :11.00    
##  1st Qu.: 8047   1st Qu.:1.000               Yes: 416   1st Qu.:12.00    
##  Median :14236   Median :2.000                          Median :14.00    
##  Mean   :14313   Mean   :2.693                          Mean   :15.21    
##  3rd Qu.:20462   3rd Qu.:4.000                          3rd Qu.:18.00    
##  Max.   :26999   Max.   :9.000                          Max.   :25.00    
##                                                                          
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000  
##  1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000  
##  Median :3.000     Median :3.000            Median :80    Median :1.0000  
##  Mean   :3.154     Mean   :2.712            Mean   :80    Mean   :0.7939  
##  3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000  
##  Max.   :4.000     Max.   :4.000            Max.   :80    Max.   :3.0000  
##                                                                           
```

```
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
##  1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
##  Median :10.00     Median :3.000         Median :3.000   Median : 5.000
##  Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008
##  3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000
##  Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000
##                                                                         
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000     Min.   : 0.000          Min.   : 0.000      
##  1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000      
##  Median : 3.000     Median : 1.000          Median : 3.000      
##  Mean   : 4.229     Mean   : 2.188          Mean   : 4.123      
##  3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000      
##  Max.   :18.000     Max.   :15.000          Max.   :17.000      
##                                                                 
```

```
sum(is.na(ibmdata))
```

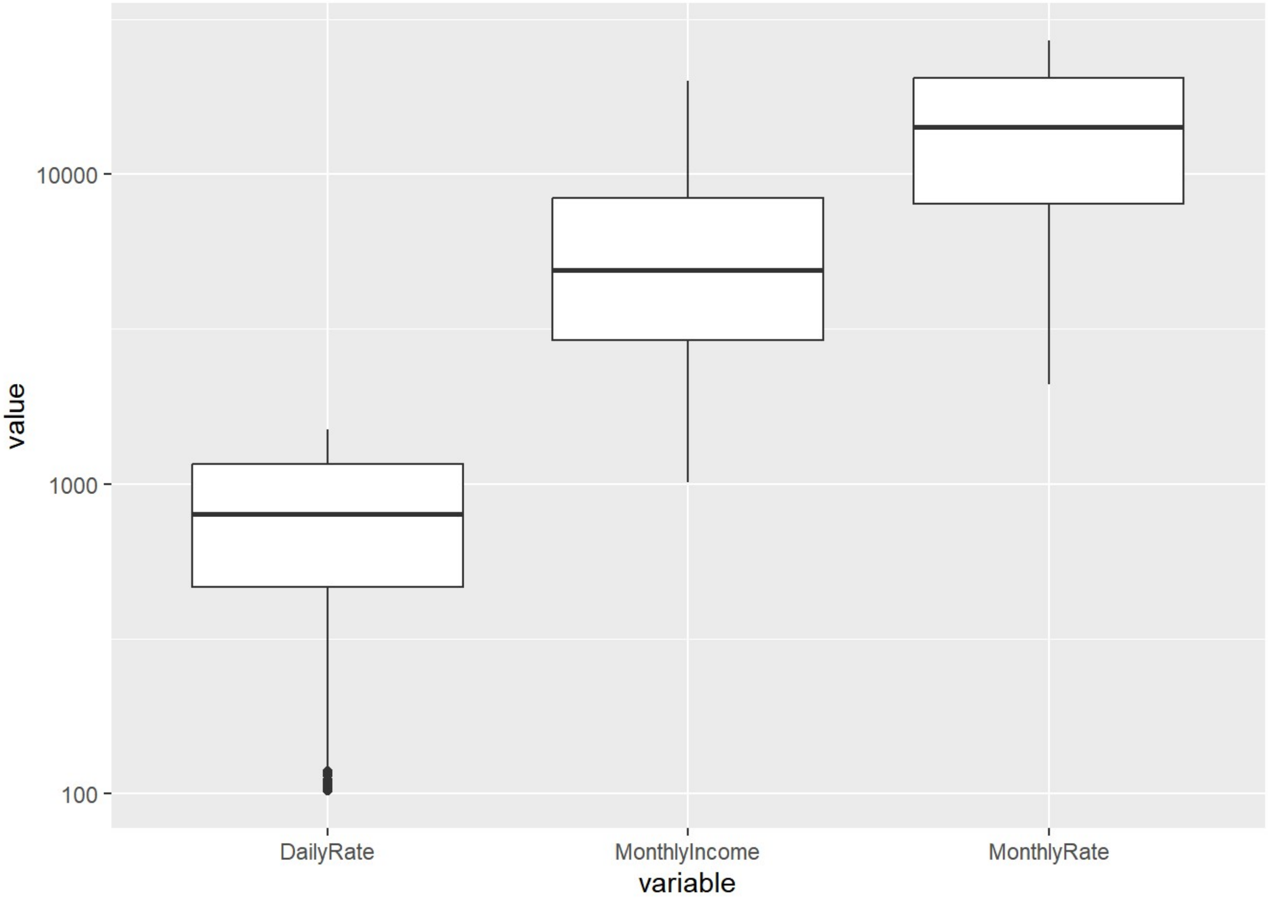
```
## [1] 0
```

```
#There are no values set as NA in the data set based on the results of the summary and the sum of is.na be
ing 0.
```

```
ibmdata_long <- melt(ibmdata, measure.vars = c("DailyRate", "MonthlyIncome", "MonthlyRate"))
```

```
## Warning in melt(ibmdata, measure.vars = c("DailyRate", "MonthlyIncome", : The
## melt generic in data.table has been passed a data.frame and will attempt to
## redirect to the relevant reshape2 method; please note that reshape2 is
## deprecated, and this redirection is now deprecated as well. To continue using
## melt methods from reshape2 while both libraries are attached, e.g. melt.list,
## you can prepend the namespace like reshape2::melt(ibmdata). In the next
## version, this warning will become an error.
```

```
ggplot(ibmdata_long, aes(x = variable, y = value)) + geom_boxplot() + scale_y_log10()
```



Load the mtcars dataset (one of the built-in data sets).

a. Create a table that shows the mean hp of cars within each category of cylinders
b. Create a second table that shows for each combination of category of cylinders and type of transmission, the mean quarter-mile time.

Hint: there is a function called fct_cross in tidyverse that may help here with part b- do look it up!

```
data(mtcars)

sapply(split(mtcars$hp, mtcars$cyl), mean)
```

```
##         4         6         8 
##  82.63636 122.28571 209.21429
```

```
aggregate(qsec~am*cyl, data=mtcars, mean)
```

```
##   am cyl     qsec
## 1  0   4 20.97000
## 2  1   4 18.45000
## 3  0   6 19.21500
## 4  1   6 16.32667
## 5  0   8 17.14250
## 6  1   8 14.55000
```

```
#I was unable to figure out how to get fct_cross working, so I had to take a different approach to buildin
g the second table.
```

Create an S3 structure that will hold the following information about a fish and chips shop: -the name -the owner -A list of the number of fish sold per month for the last 12 months -The pounds of potatoes used per month for the last 12 months -The income per month

Create a member function called plot(x) that will plot a graph of fish or potatoes over time (the last 12 months) depending on whether x is “fish” or “potatoes”

Pick values at random for the fish and potatoes entry or use rnorm() to fill them in

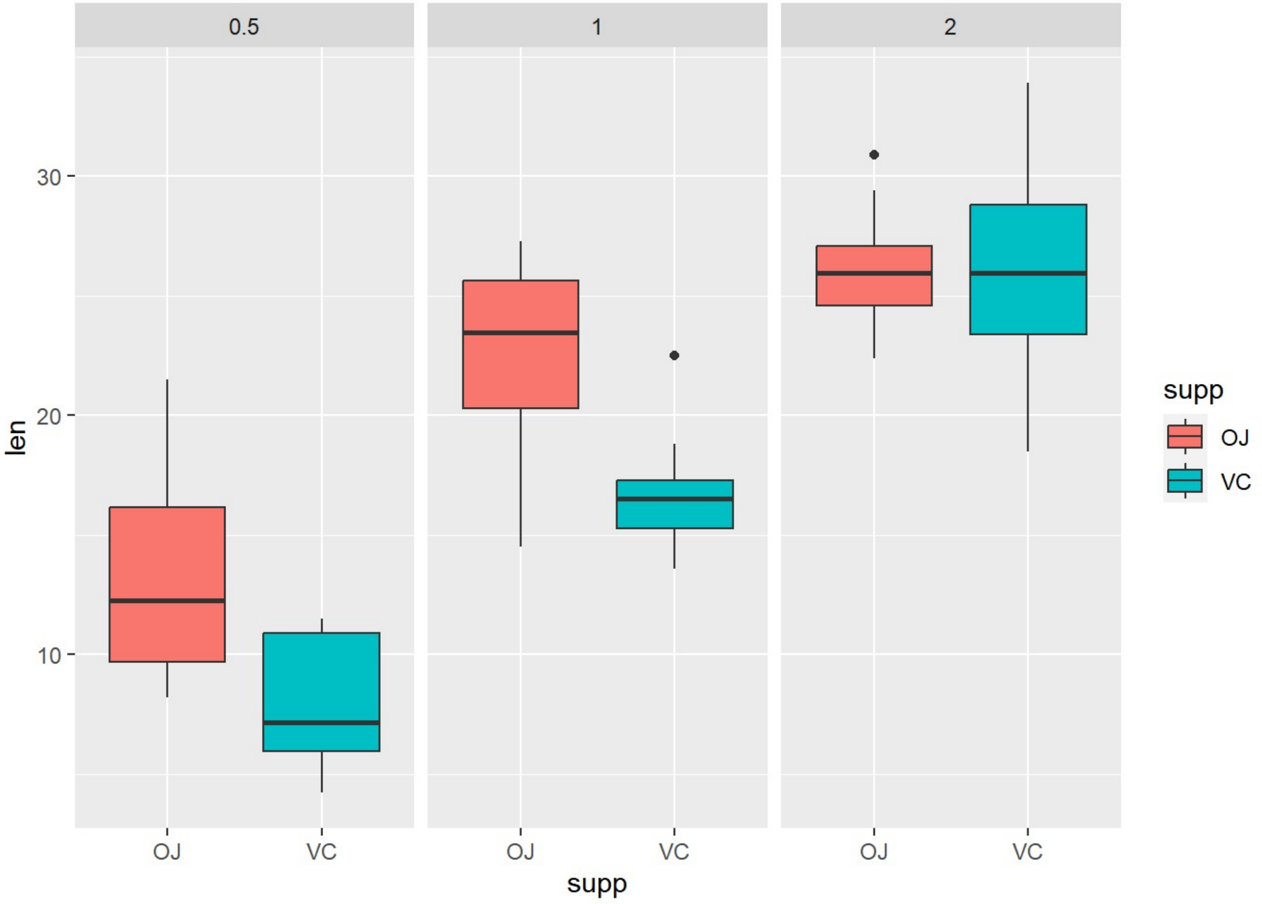
```
mys3 <- list(name = "Long John Silver's", owner="Bob", fishMonthly=300, potatoesMonthly=200, incomeMonthly
=10000)
class(mys3)<-"fishNChips"

plot.fishNChips <- function(x)
{
  if(x == "fish")
  {
    plot(x=1:nrow(mys3), y=mys3$fishMonthly)
  } else if (x == "potatoes") {
    plot(x=1:nrow(mys3), y=mys3$potatoeshMonthly)
  }
}
#plot.fishNChips("fish")

#I'm not really sure how to proceed here in getting this to function
```

Using the built-in data set “Tooth Growth”, produce a graph or table that shows how the growth of pig’s teeth (len) is influenced by the dose of a drug (dose) and the way the drug was delivered (sup). Produce a single plot or table that clearly shows the impact of these two factors

```
data("ToothGrowth")
ggplot(ToothGrowth, aes(x = supp, y = len)) + geom_boxplot(aes(fill = supp))+facet_grid(. ~ dose)
```


0.5
1
2
len
30
20
10
OJ
VC
OJ
VC
OJ
VC
supp
supp
OJ
VC