**ITWS - 6600 Data Analytics**Assignment 3

# Assignment 3 Question 1

setwd("C:/Users/Shrey Jain/Documents/Study/Data Analytics/DataAnalyticsFall2022\_SHREY\_JAIN/Lab/DataAnalytics\_A3\_SHREY\_JAIN/nytimes/")

nyt3 <- read.csv("nyt3.csv")

nyt4 <- read.csv("nyt4.csv")

nyt5 <- read.csv("nyt5.csv")

nyt6 <- read.csv("nyt6.csv")

nyt7 <- read.csv("nyt7.csv")

nyt8 <- read.csv("nyt8.csv")

nyt9 <- read.csv("nyt9.csv")

# Question 1 (a)

boxplot(nyt3$Age, nyt3$Clicks)

boxplot(nyt4$Age, nyt4$Clicks)

boxplot(nyt5$Age, nyt5$Clicks)

boxplot(nyt6$Age, nyt6$Clicks)

boxplot(nyt7$Age, nyt7$Clicks)

boxplot(nyt8$Age, nyt8$Clicks)

boxplot(nyt9$Age, nyt9$Clicks)

# After plotting the boxplots for age and clicks, it seems that the median value of age is about 30.

# This goes for all 7 datasets from nyt3 to nyt9.

# Secondly, the max-age value is roughly about 50 for all 7 datasets.

# For clicks, it seems the average value or the most encountered value is 0. Meaning that most users have not clicked.

# Secondly, the max value for clicks is around 5. That's the max click count basis for the given datasets.

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

# Question 1 (b)

hist(nyt3$Age, col='blue')

hist(nyt3$Impressions, col='green', add=TRUE)

hist(nyt4$Age, col='blue')

hist(nyt4$Impressions, col='green', add=TRUE)

hist(nyt5$Age, col='blue')

hist(nyt5$Impressions, col='green', add=TRUE)

hist(nyt6$Age, col='blue')

hist(nyt6$Impressions, col='green', add=TRUE)

hist(nyt7$Age, col='blue')

hist(nyt7$Impressions, col='green', add=TRUE)

hist(nyt8$Age, col='blue')

hist(nyt8$Impressions, col='green', add=TRUE)

hist(nyt9$Age, col='blue')

hist(nyt9$Impressions, col='green', add=TRUE)

# It seems age and impression give a little better insight as compared to age and clicks.

# That is primarily because most of the users didn't click making the majority of values in clicks column 0.

# This goes for all 7 datasets from nyt3 to nyt9.

# Looking at the histograms, both columns seem to follow a normal distribution. Especially, impressions.

# For Age, it seems the data is a bit left skewed until age 10, however, from 10 to 90, the distribution is normal.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

# Question 1 (c)

plot(ecdf(nyt3$Age), col='blue')

plot(ecdf(nyt3$Impressions), col='green', add=TRUE)

qqplot(nyt3$Age, nyt3$Impressions)

plot(ecdf(nyt4$Age), col='blue')

plot(ecdf(nyt4$Impressions), col='green', add=TRUE)

qqplot(nyt4$Age, nyt4$Impressions)

plot(ecdf(nyt5$Age), col='blue')

plot(ecdf(nyt5$Impressions), col='green', add=TRUE)

qqplot(nyt5$Age, nyt5$Impressions)

plot(ecdf(nyt6$Age), col='blue')

plot(ecdf(nyt6$Impressions), col='green', add=TRUE)

qqplot(nyt6$Age, nyt6$Impressions)

plot(ecdf(nyt7$Age), col='blue')

plot(ecdf(nyt7$Impressions), col='green', add=TRUE)

qqplot(nyt7$Age, nyt7$Impressions)

plot(ecdf(nyt8$Age), col='blue')

plot(ecdf(nyt8$Impressions), col='green', add=TRUE)

qqplot(nyt8$Age, nyt8$Impressions)

plot(ecdf(nyt9$Age), col='blue')

plot(ecdf(nyt9$Impressions), col='green', add=TRUE)

qqplot(nyt9$Age, nyt9$Impressions)

# Looking at the qqplots across all 7 datasets between age and impression, it seems, both the values do come from a population with a common distribution.

# Similarly, the ECDF plots also convey that the although the curve is different, many values tend to reach a tangent to y=1

# It again seems age and impression give a little better insight as compared to age and clicks. This goes for all 7 datasets from nyt3 to nyt9.

Chart

Description automatically generated Chart

Description automatically generated Chart

Description automatically generated Chart

Description automatically generated   
  
 Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart

Description automatically generated

# Question 1 (d)

# Shapiro test for checking normal distribution

shapiro.test(nyt4$Age[0:5000])

shapiro.test(nyt4$Impressions[0:5000])

# Both columns are not normally distributed as the p-value is very low

# Wilcox test for checking co-relation

wilcox.test(nyt4$Age, nyt4$Impressions, data=nyt4)

# Both are independent as the p-value is very low

# Question 1 (e)

# All the 7 datasets seems to be evenly split meaning that after plotting all the various plots accross all the 7 dataframes,

# I see that almost all the plots follow the same patten. Some of the columns also seem to have not much insights.

# Assignment 3 Question 2

nyt3\_new <- nyt3[nyt3$Gender == "1", ]

nyt4\_new <- nyt4[nyt4$Gender == "1", ]

nyt5\_new <- nyt5[nyt5$Gender == "1", ]

nyt6\_new <- nyt6[nyt6$Gender == "1", ]

# Histograms

hist(nyt3\_new$Age, col='blue')

hist(nyt3\_new$Impressions, col='green', add=TRUE)

hist(nyt4\_new$Age, col='blue')

hist(nyt4\_new$Impressions, col='green', add=TRUE)

hist(nyt5\_new$Age, col='blue')

hist(nyt5\_new$Impressions, col='green', add=TRUE)

hist(nyt6\_new$Age, col='blue')

hist(nyt6\_new$Impressions, col='green', add=TRUE)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

# ECDF and QQPlots

plot(ecdf(nyt3\_new$Age), col='blue')

plot(ecdf(nyt3\_new$Impressions), col='green', add=TRUE)

qqplot(nyt3\_new$Age, nyt3\_new$Impressions)

plot(ecdf(nyt4\_new$Age), col='blue')

plot(ecdf(nyt4\_new$Impressions), col='green', add=TRUE)

qqplot(nyt4\_new$Age, nyt4\_new$Impressions)

plot(ecdf(nyt5\_new$Age), col='blue')

plot(ecdf(nyt5\_new$Impressions), col='green', add=TRUE)

qqplot(nyt5\_new$Age, nyt5\_new$Impressions)

plot(ecdf(nyt6\_new$Age), col='blue')

plot(ecdf(nyt6\_new$Impressions), col='green', add=TRUE)

qqplot(nyt6\_new$Age, nyt6\_new$Impressions)

Chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart

Description automatically generated Chart, histogram

Description automatically generated

# Shapiro test and Wilcox test

shapiro.test(nyt4\_new$Age[0:5000])

shapiro.test(nyt4\_new$Impressions[0:5000])

wilcox.test(nyt4\_new$Age, nyt4\_new$Impressions, data=nyt4\_new)

# Histograms seem to follow normal distribution without any skew when we filter data with Gender==1.

# ECDFs and QQplots still follow the same trend, as the characteristics of the data have not changed through filtering. However, some outliers were removed.

# Shapiro test still results in roughly the same results meaning both age and impressions are independent as the p-value is very low

# Similarly, the Wilcox test also results in the same. And this is in fact obvious, as we have just filtered the data, this does not change the characteristics of the data