Peter Mamaev

MGMT-4600 Data Analytics Assignment 3

> # Filtering out null Age values for most of my datasets because a 0 age is impossible.

> nyt3\_age <- filter(nyt3, Age!=0)

> head(nyt3\_age)

Age Gender Impressions Clicks Signed\_In

1 46 1 3 0 1

2 75 0 9 0 1

3 39 0 2 0 1

4 54 0 4 0 1

5 15 1 3 0 1

6 50 0 8 0 1

> nyt14\_age <- filter(nyt14, Age!=0)

> nyt15\_age <- filter(nyt15, Age!=0)

> nyt16\_age <- filter(nyt16, Age!=0)

> nyt17\_age <- filter(nyt17, Age!=0)

> nyt28\_age <- filter(nyt28, Age!=0)

> nyt29\_age <- filter(nyt29, Age!=0)

# Create boxplots for all 7 datasets for each of two key variables (you choose the

# variables), i.e. two figures (one for each variable) with 7 boxplots (for the

# 7 different datasets) in each. Describe and run summary statistics on the two

# chosen variable and explain them in your words. min. 3-4 sentences

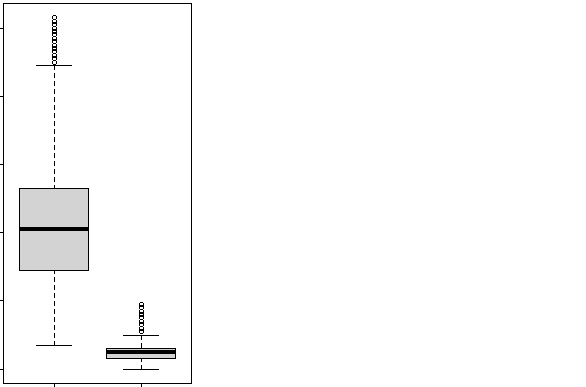
# I'm using the \_age dataset, where I removed the null age values, ONLY for the boxplots, because it is the only

# time I'm directly comparing the Age and Impressions against each other.

# I use the original, unfiltered dataset for Impressions after this point since it's more complete on its own.

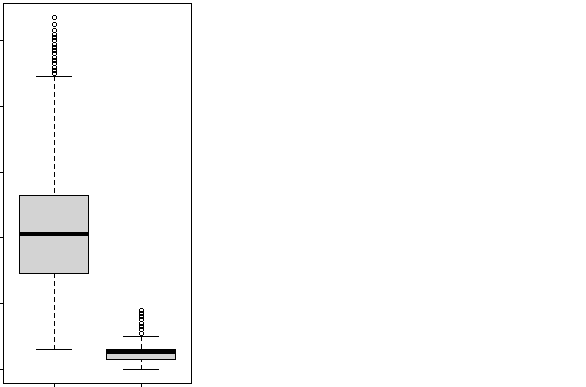
> # Average at roughly 36 for Age, much lower for impressions at about 5?

> nyt3boxplot = boxplot(nyt3\_age$Age, nyt3\_age$Impressions)



> # Highest average and upper quartile of the boxplots.

> nyt14boxplot = boxplot(nyt14\_age$Age, nyt14\_age$Impressions)

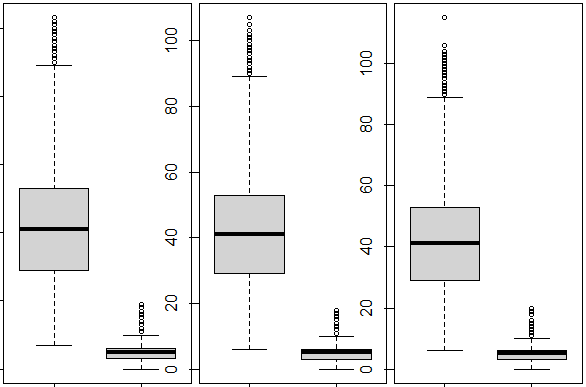


> # Boxplot distributions for nyt15-17 roughly identical.

> nyt15boxplot = boxplot(nyt15\_age$Age, nyt15\_age$Impressions)

> nyt16boxplot = boxplot(nyt16\_age$Age, nyt16\_age$Impressions)

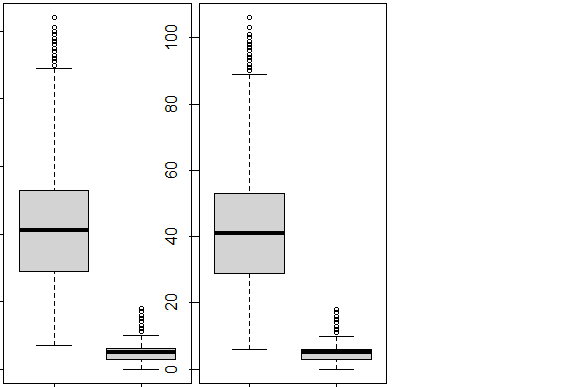
> nyt17boxplot = boxplot(nyt17\_age$Age, nyt17\_age$Impressions)



> # Median roughly 20 for Age,

> nyt28boxplot = boxplot(nyt28\_age$Age, nyt28\_age$Impressions)

> nyt29boxplot = boxplot(nyt29\_age$Age, nyt29\_age$Impressions)



# Initially I wanted to compare Age data from the filtered data and Impressions from the unfiltered data.

# For the time being I'm using filtered data for both.

# Impressions distribution is significantly lower when using data where null age values are filtered out.

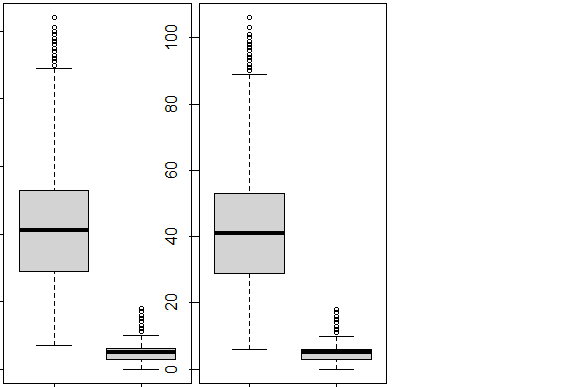
A diagram of a graph

Description automatically generated

> # Median roughly 20 for Age,

> nyt28boxplot = boxplot(nyt28\_age$Age, nyt28\_age$Impressions)

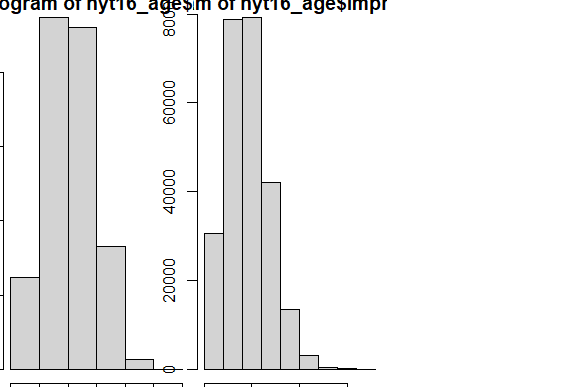
> nyt29boxplot = boxplot(nyt29\_age$Age, nyt29\_age$Impressions)



|  |
| --- |
| > # For the time being I'm using filtered data for both.  > # Impressions distribution is significantly lower when using data where null age values are filtered out.  > summary(nyt3\_age$Age)  Min. 1st Qu. Median Mean 3rd Qu. Max.  5.00 29.00 41.00 42.09 53.00 109.00  > summary(nyt3\_age$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.997 6.000 19.000  > summary(nyt14\_age$Age)  Min. 1st Qu. Median Mean 3rd Qu. Max.  3.00 29.00 41.00 42.11 53.00 105.00  > summary(nyt14\_age$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0 3 5 5 6 20  > summary(nyt15\_age$Age)  Min. 1st Qu. Median Mean 3rd Qu. Max.  7.00 29.00 41.00 42.13 53.00 103.00  > summary(nyt15\_age$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 5.008 6.000 19.000  > summary(nyt16\_age$Age)  Min. 1st Qu. Median Mean 3rd Qu. Max.  6.00 29.00 41.00 42.11 53.00 107.00  > summary(nyt16\_age$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0 3 5 5 6 18  > summary(nyt17\_age$Age)  Min. 1st Qu. Median Mean 3rd Qu. Max.  6.00 29.00 41.00 42.02 53.00 115.00  > summary(nyt17\_age$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.998 6.000 20.000  > summary(nyt28\_age$Age)  Min. 1st Qu. Median Mean 3rd Qu. Max.  7.00 29.00 41.00 42.08 53.00 104.00  > summary(nyt28\_age$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 5.003 6.000 18.000  > summary(nyt29\_age$Age)  Min. 1st Qu. Median Mean 3rd Qu. Max.  6.00 29.00 41.00 42.08 53.00 106.00  > summary(nyt29\_age$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.997 6.000 18.000 |
|  |
| |  | | --- | |  | |
| > # Using data from the unfiltered dataset now, notably the distribution of Impressions is roughly the same.  > # We don't know what a zero distribution means or whether there is a maximum value, so we can't technically  > # compare it to the Age or draw conclusions on whether the Impressions are "good" or "bad".  > # Since we understand what a rough possible average age for a human is.  > summary(nyt3$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.996 6.000 19.000  > summary(nyt14$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.998 6.000 20.000  > summary(nyt15$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 5.001 6.000 19.000  > summary(nyt16$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.999 6.000 19.000  > summary(nyt17$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.999 6.000 20.000  > summary(nyt28$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.999 6.000 19.000  > summary(nyt29$Impressions)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 5.000 4.998 6.000 18.000 |
|  |
| |  | | --- | | # Conduct the applicable normality test (i.e Shapiro Wilk, Anderson Darling,  # Kolmogorov-Smirnov) for the chosen two variables  # for all 7 datasets for two key variables – can be the same variables  # in 1a or different. Create histograms for those two variables in the 7 datasets  # (you choose the histogram bin width). Describe the distributions in terms of  # known parametric distributions and similarities/ differences among them.  >install.packages('nortest')  >library(nortest)  > # Extremely small p-values for all these values, meaning they are not normally distributed.  > ad.test(nyt3\_age$Age)  Anderson-Darling normality test  data: nyt3\_age$Age  A = 1242.2, p-value < 2.2e-16  > ad.test(nyt3\_age$Impressions)  Anderson-Darling normality test  data: nyt3\_age$Impressions  A = 3250, p-value < 2.2e-16  > ad.test(nyt14\_age$Age)  Anderson-Darling normality test  data: nyt14\_age$Age  A = 1242.9, p-value < 2.2e-16  > ad.test(nyt14\_age$Impressions)  Anderson-Darling normality test  data: nyt14\_age$Impressions  A = 3204.8, p-value < 2.2e-16  > ad.test(nyt15\_age$Age)  Anderson-Darling normality test  data: nyt15\_age$Age  A = 967.91, p-value < 2.2e-16  > ad.test(nyt15\_age$Impressions)  Anderson-Darling normality test  data: nyt15\_age$Impressions  A = 2497.4, p-value < 2.2e-16  > ad.test(nyt16\_age$Age)  Anderson-Darling normality test  data: nyt16\_age$Age  A = 991.15, p-value < 2.2e-16  > ad.test(nyt16\_age$Impressions)  Anderson-Darling normality test  data: nyt16\_age$Impressions  A = 2605.6, p-value < 2.2e-16  > ad.test(nyt17\_age$Age)  Anderson-Darling normality test  data: nyt17\_age$Age  A = 1033.8, p-value < 2.2e-16  > ad.test(nyt17\_age$Impressions)  Anderson-Darling normality test  data: nyt17\_age$Impressions  A = 2552.9, p-value < 2.2e-16  > ad.test(nyt28\_age$Age)  Anderson-Darling normality test  data: nyt28\_age$Age  A = 968.71, p-value < 2.2e-16  > ad.test(nyt28\_age$Impressions)  Anderson-Darling normality test  data: nyt28\_age$Impressions  A = 2507.5, p-value < 2.2e-16  > ad.test(nyt29\_age$Age)  Anderson-Darling normality test  data: nyt29\_age$Age  A = 1029, p-value < 2.2e-16  > ad.test(nyt29\_age$Impressions)  Anderson-Darling normality test  data: nyt29\_age$Impressions  A = 2679.6, p-value < 2.2e-16  > # Randomly sampling 5000 values from the 44K-ish rows for the Shapiro-Wilks test applicability.  > # Thanks to the sample() function the mean stays roughly similar.  > # I'm performing the Shapiro-Wilks test on a sample of a sample technically speaking.  > # It's only here to reaffirm the tiny p-value.  > # Acceptable values for the Shapiro-Wilks test are between 3 and 5000 exclusive, so I'm using the maximum acceptable sample size for it (4999).  > nyt3age = sample(nyt3\_age$Age, 4999, replace = FALSE, prob = NULL)  > nyt3impressions = sample(nyt3\_age$Impressions, 4999, replace = FALSE, prob = NULL)  > nyt14age = sample(nyt14\_age$Age, 4999, replace = FALSE, prob = NULL)  > nyt14impressions = sample(nyt14\_age$Impressions, 4999, replace = FALSE, prob = NULL)  > nyt15age = sample(nyt15\_age$Age, 4999, replace = FALSE, prob = NULL)  > nyt15impressions = sample(nyt15\_age$Impressions, 4999, replace = FALSE, prob = NULL)  > nyt16age = sample(nyt16\_age$Age, 4999, replace = FALSE, prob = NULL)  > nyt16impressions = sample(nyt16\_age$Impressions, 4999, replace = FALSE, prob = NULL)  > nyt17age = sample(nyt17\_age$Age, 4999, replace = FALSE, prob = NULL)  > nyt17impressions = sample(nyt17\_age$Impressions, 4999, replace = FALSE, prob = NULL)  > nyt28age = sample(nyt28\_age$Age, 4999, replace = FALSE, prob = NULL)  > nyt28impressions = sample(nyt28\_age$Impressions, 4999, replace = FALSE, prob = NULL)  > nyt29age = sample(nyt29\_age$Age, 4999, replace = FALSE, prob = NULL)  > nyt29impressions = sample(nyt29\_age$Impressions, 4999, replace = FALSE, prob = NULL)  # Shapiro-Wilks Test as verification  # Test statistic W extremely close to 1.  # Hypothesis of a normal distribution is rejected if the critical value P for the test statistic W is less than 0.05.  # All p-values are extremely small regardless of whether the data used is sampled or not.  > shapiro.test(nyt3age)  Shapiro-Wilk normality test  data: nyt3age  W = 0.98116, p-value < 2.2e-16  > shapiro.test(nyt3impressions)  Shapiro-Wilk normality test  data: nyt3impressions  W = 0.97049, p-value < 2.2e-16  > shapiro.test(nyt14age)  Shapiro-Wilk normality test  data: nyt14age  W = 0.9802, p-value < 2.2e-16  > shapiro.test(nyt14impressions)  Shapiro-Wilk normality test  data: nyt14impressions  W = 0.97179, p-value < 2.2e-16  > shapiro.test(nyt15age)  Shapiro-Wilk normality test  data: nyt15age  W = 0.97999, p-value < 2.2e-16  > shapiro.test(nyt15impressions)  Shapiro-Wilk normality test  data: nyt15impressions  W = 0.97118, p-value < 2.2e-16  > shapiro.test(nyt16age)  Shapiro-Wilk normality test  data: nyt16age  W = 0.98074, p-value < 2.2e-16  > shapiro.test(nyt16impressions)  Shapiro-Wilk normality test  data: nyt16impressions  W = 0.96945, p-value < 2.2e-16  > shapiro.test(nyt17age)  Shapiro-Wilk normality test  data: nyt17age  W = 0.98114, p-value < 2.2e-16  > shapiro.test(nyt17impressions)  Shapiro-Wilk normality test  data: nyt17impressions  W = 0.97313, p-value < 2.2e-16  > shapiro.test(nyt28age)  Shapiro-Wilk normality test  data: nyt28age  W = 0.98348, p-value < 2.2e-16  > shapiro.test(nyt28impressions)  Shapiro-Wilk normality test  data: nyt28impressions  W = 0.97129, p-value < 2.2e-16  > shapiro.test(nyt29age)  Shapiro-Wilk normality test  data: nyt29age  W = 0.9818, p-value < 2.2e-16  > shapiro.test(nyt29impressions)  Shapiro-Wilk normality test  data: nyt29impressions  W = 0.96925, p-value < 2.2e-16  > # Checking that the distribution remains roughly the same even when cut down to 5000 samples for the Shapiro-Wilk test.  > mean(nyt3age)  [1] 41.91598  > mean(nyt3\_age$Age)  [1] 42.09335  > ks.test(nyt3\_age$Age, nyt3\_age$Impressions)  Asymptotic two-sample Kolmogorov-Smirnov test  data: nyt3\_age$Age and nyt3\_age$Impressions  D = 0.99454, p-value < 2.2e-16  alternative hypothesis: two-sided  Warning message:  In ks.test.default(nyt3\_age$Age, nyt3\_age$Impressions) :  p-value will be approximate in the presence of ties  > ks.test(nyt14\_age$Age, nyt14\_age$Impressions)  Asymptotic two-sample Kolmogorov-Smirnov test  data: nyt14\_age$Age and nyt14\_age$Impressions  D = 0.99474, p-value < 2.2e-16  alternative hypothesis: two-sided  Warning message:  In ks.test.default(nyt14\_age$Age, nyt14\_age$Impressions) :  p-value will be approximate in the presence of ties  > ks.test(nyt15\_age$Age, nyt15\_age$Impressions)  Asymptotic two-sample Kolmogorov-Smirnov test  data: nyt15\_age$Age and nyt15\_age$Impressions  D = 0.9947, p-value < 2.2e-16  alternative hypothesis: two-sided  Warning message:  In ks.test.default(nyt15\_age$Age, nyt15\_age$Impressions) :  p-value will be approximate in the presence of ties  > ks.test(nyt16\_age$Age, nyt16\_age$Impressions)  Asymptotic two-sample Kolmogorov-Smirnov test  data: nyt16\_age$Age and nyt16\_age$Impressions  D = 0.99476, p-value < 2.2e-16  alternative hypothesis: two-sided  Warning message:  In ks.test.default(nyt16\_age$Age, nyt16\_age$Impressions) :  p-value will be approximate in the presence of ties  > ks.test(nyt17\_age$Age, nyt17\_age$Impressions)  Asymptotic two-sample Kolmogorov-Smirnov test  data: nyt17\_age$Age and nyt17\_age$Impressions  D = 0.99455, p-value < 2.2e-16  alternative hypothesis: two-sided  Warning message:  In ks.test.default(nyt17\_age$Age, nyt17\_age$Impressions) :  p-value will be approximate in the presence of ties  > ks.test(nyt28\_age$Age, nyt28\_age$Impressions)  Asymptotic two-sample Kolmogorov-Smirnov test  data: nyt28\_age$Age and nyt28\_age$Impressions  D = 0.99474, p-value < 2.2e-16  alternative hypothesis: two-sided  Warning message:  In ks.test.default(nyt28\_age$Age, nyt28\_age$Impressions) :  p-value will be approximate in the presence of ties  > ks.test(nyt29\_age$Age, nyt29\_age$Impressions)  Asymptotic two-sample Kolmogorov-Smirnov test  data: nyt29\_age$Age and nyt29\_age$Impressions  D = 0.99439, p-value < 2.2e-16  alternative hypothesis: two-sided  Warning message:  In ks.test.default(nyt29\_age$Age, nyt29\_age$Impressions) :  p-value will be approximate in the presence of ties  # The break values in the Impressions histogram are larger than in the Age histogram.  # This is to keep the graphs distinct and more detailed.  hist(nyt3\_age$Age, breaks=5, xlab = "3 Age Distribution")  hist(nyt3\_age$Impressions, breaks=7, xlab = "3 Impressions Distribution")    > hist(nyt14\_age$Age, breaks=5, xlab = "14 Age Distribution")  > hist(nyt14\_age$Impressions, breaks=7, xlab = "14 Impressions Distribution")    > hist(nyt15\_age$Age, breaks=5, xlab = "15 Age Distribution")  > hist(nyt15\_age$Impressions, breaks=7, xlab = "15 Impressions Distribution") | |

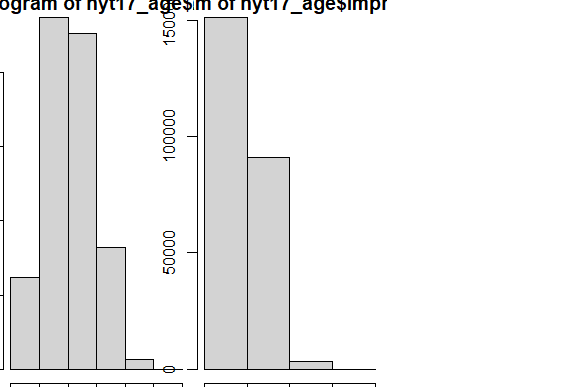
> hist(nyt16\_age$Age, breaks=5, xlab = "16 Age Distribution")

> hist(nyt16\_age$Impressions, breaks=7, xlab = "16 Impressions Distribution")



> hist(nyt17\_age$Age, breaks=5, xlab = "17 Age Distribution")

> hist(nyt17\_age$Impressions, breaks=7, xlab = "17 Impressions Distribution")



> hist(nyt28\_age$Age, breaks=5, xlab = "28 Age Distribution")

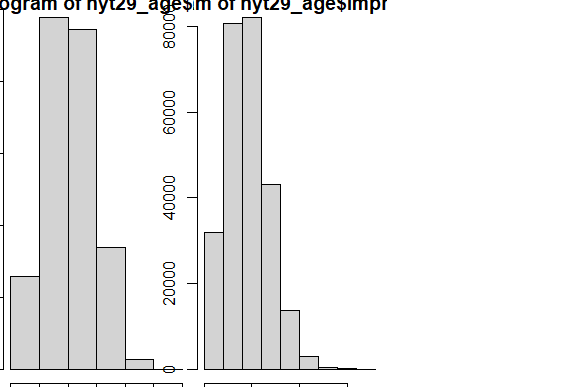
> hist(nyt28\_age$Impressions, breaks=7, xlab = "28 Impressions Distribution")

A graph of a number of people

Description automatically generated with medium confidence

> hist(nyt29\_age$Age, breaks=5, xlab = "Age Distribution")

> hist(nyt29\_age$Impressions, breaks=7, xlab = "Impressions Distribution")



# Plot the ECDFs (Empirical Cumulative Distribution Function for your two key

# variables. Plot the quantile-quantile distribution using a suitable parametric

# distribution you chose in 1b. Describe features of these plots.

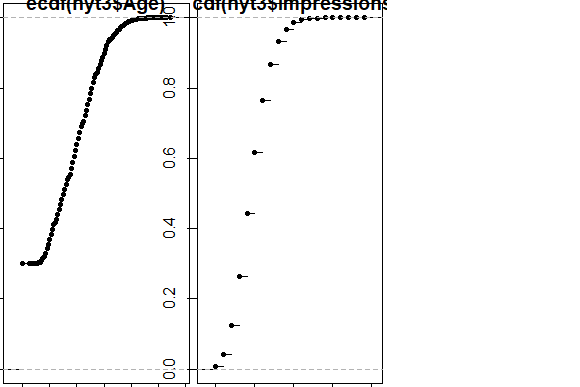
# ECDF plots selected data feature in order from least to greatest and see the whole feature as if is distributed across the data set.

# Naturally Impressions have a much greater spread than Data, and seem more concentrated towards the highest value.

# While difficult to judge, the ECDF distributions between the same categories in different datasets are extremely similar.

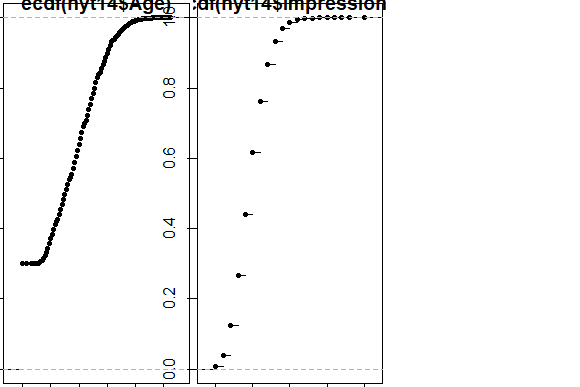
> plot(ecdf(nyt3$Age))

> plot(ecdf(nyt3$Impressions))



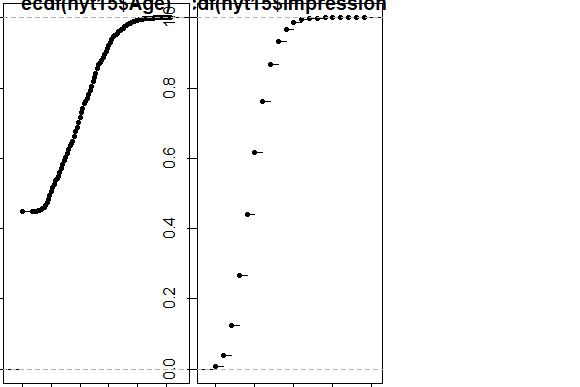
> plot(ecdf(nyt14$Age))

> plot(ecdf(nyt14$Impressions))



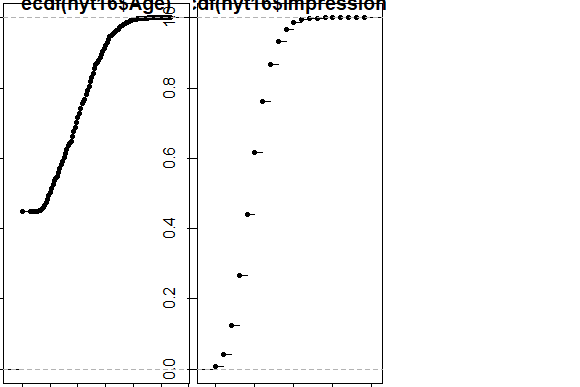
> plot(ecdf(nyt15$Age))

> plot(ecdf(nyt15$Impressions))



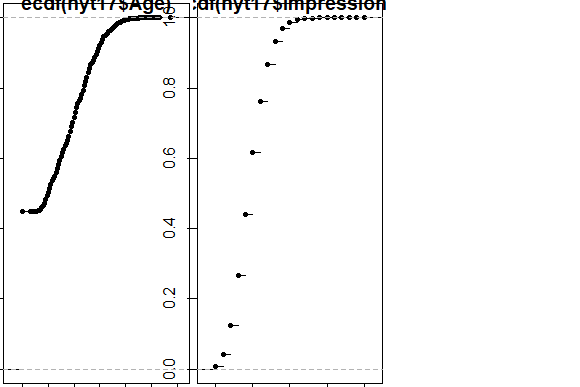
> plot(ecdf(nyt16$Age))

> plot(ecdf(nyt16$Impressions))



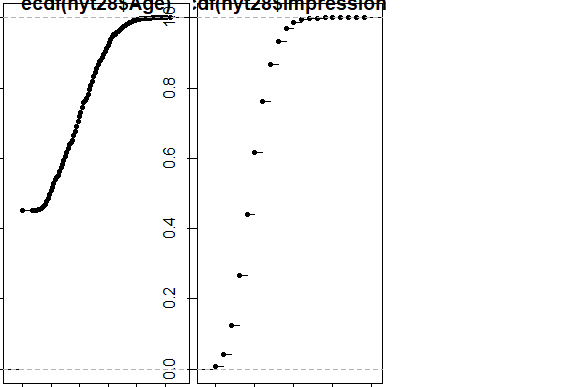
> plot(ecdf(nyt17$Age))

> plot(ecdf(nyt17$Impressions))



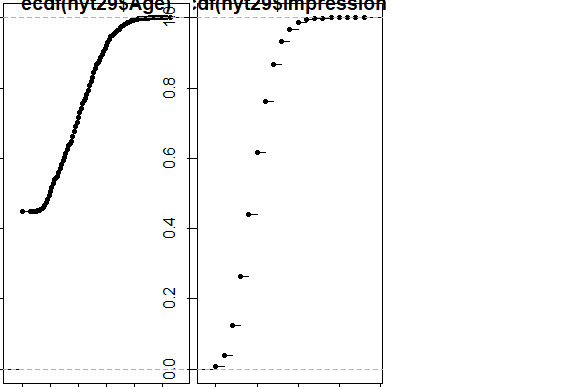
> plot(ecdf(nyt28$Age))

> plot(ecdf(nyt28$Impressions))



> plot(ecdf(nyt29$Age))

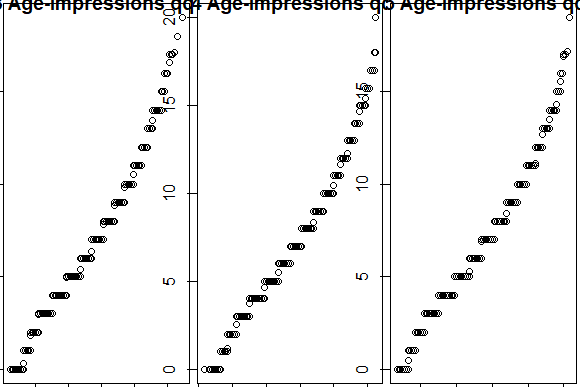
> plot(ecdf(nyt29$Impressions))



> qqplot(nyt3\_age$Age, nyt3$Impressions) + title("nyt3 Age-Impressions qqplot")

> qqplot(nyt14\_age$Age, nyt14$Impressions) + title("nyt14 Age-Impressions qqplot")

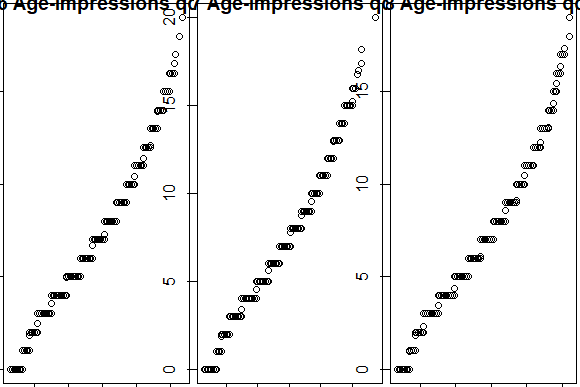
> qqplot(nyt15\_age$Age, nyt15$Impressions) + title("nyt15 Age-Impressions qqplot")



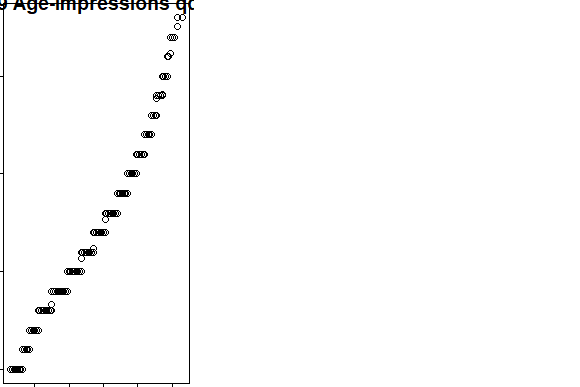
> qqplot(nyt16\_age$Age, nyt16$Impressions) + title("nyt16 Age-Impressions qqplot")

> qqplot(nyt17\_age$Age, nyt17$Impressions) + title("nyt17 Age-Impressions qqplot")

> qqplot(nyt28\_age$Age, nyt28$Impressions) + title("nyt28 Age-Impressions qqplot")



> qqplot(nyt29\_age$Age, nyt29$Impressions) + title("nyt29 Age-Impressions qqplot")



# Significance test should give out a p-value

# A t-test is used to compare means of two different variables

# Comparing sample and population test is z-test

# Comparing means of different values of NYTimes

# p-values of the Kolmgorodov-Smirnov, Shapiro-Wilks, and Anderson-Darling Test all

# consistently demonstrate that the data is not normally distributed due to the small p-values.

|  |
| --- |
| > t.test(nyt3\_age$Age)  One Sample t-test  data: nyt3\_age$Age  t = 1430.9, df = 308286, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  42.03570 42.15101  sample estimates:  mean of x  42.09335  > t.test(nyt14\_age$Age)  One Sample t-test  data: nyt14\_age$Age  t = 1428.4, df = 308038, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  42.04799 42.16354  sample estimates:  mean of x  42.10577  > t.test(nyt15\_age$Age)  One Sample t-test  data: nyt15\_age$Age  t = 1264.3, df = 240785, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  42.06839 42.19902  sample estimates:  mean of x  42.1337  > t.test(nyt16\_age$Age)  One Sample t-test  data: nyt16\_age$Age  t = 1283.5, df = 247311, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  42.04919 42.17781  sample estimates:  mean of x  42.1135  > t.test(nyt17\_age$Age)  One Sample t-test  data: nyt17\_age$Age  t = 1273.5, df = 245573, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  41.95350 42.08284  sample estimates:  mean of x  42.01817  > t.test(nyt28\_age$Age)  One Sample t-test  data: nyt28\_age$Age  t = 1273, df = 243498, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  42.01731 42.14689  sample estimates:  mean of x  42.0821  > t.test(nyt29\_age$Age)  One Sample t-test  data: nyt29\_age$Age  t = 1299.1, df = 254671, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  42.01335 42.14032  sample estimates:  mean of x  42.07684  # Heavy variation in degrees of freedom (df).  # Mean values extremely close, at around 42.  # Miniscule p-value indicating abnormal distribution.  # Low confidence intervals for all.  # Data varies little between distributions - makes sense since most of these spreadsheets come from "synthetic" extrapolated data.  > t.test(nyt3\_age$Impressions)  One Sample t-test  data: nyt3\_age$Impressions  t = 1242.8, df = 308286, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  4.98897 5.00473  sample estimates:  mean of x  4.99685  > t.test(nyt14\_age$Impressions)  One Sample t-test  data: nyt14\_age$Impressions  t = 1242.7, df = 308038, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  4.992104 5.007876  sample estimates:  mean of x  4.99999  > t.test(nyt15\_age$Impressions)  One Sample t-test  data: nyt15\_age$Impressions  t = 1100.8, df = 240785, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  4.998692 5.016525  sample estimates:  mean of x  5.007608  > t.test(nyt16\_age$Impressions)  One Sample t-test  data: nyt16\_age$Impressions  t = 1111.4, df = 247311, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  4.991025 5.008659  sample estimates:  mean of x  4.999842  > t.test(nyt17\_age$Impressions)  One Sample t-test  data: nyt17\_age$Impressions  t = 1109.4, df = 245573, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  4.988915 5.006573  sample estimates:  mean of x  4.997744  > t.test(nyt28\_age$Impressions)  One Sample t-test  data: nyt28\_age$Impressions  t = 1105.6, df = 243498, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  4.994498 5.012238  sample estimates:  mean of x  5.003368  > t.test(nyt29\_age$Impressions)  One Sample t-test  data: nyt29\_age$Impressions  t = 1128.3, df = 254671, p-value < 2.2e-16  alternative hypothesis: true mean is not equal to 0  95 percent confidence interval:  4.987845 5.005205  sample estimates:  mean of x  4.996525  # My initial hypothesis was that data varies has great variation across different datasets - Age and Impression distributions are both extremely close.  # I also believed that data was normally distributed, due to its "artificial" extrapolated nature. The low p-values prove me wrong. |
|  |
| |  | | --- | |  | |