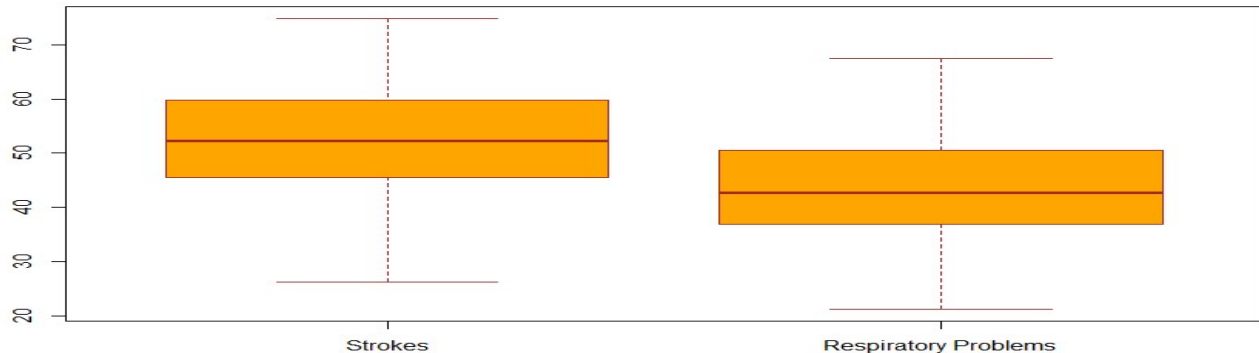


DATA PROJECT 5:

The data set is based on the death rates from various causes in all the states in America. From these measurements, I am focusing on the death rates from strokes and respiratory problems in America. This data set is important to analyze as it will give us a pretty good understanding of the mortality rate of the population in all the states of America.

The length of the data sets of the death rates from strokes and respiratory problems are above 40 since there are more than 40 states in America, hence, it is safe to assume **that the distributions are standard normal distributions.**



We have the boxplots of both the data sets being discussed on one plot. The five-number summary tells us the following:

	OUTLIERS/EXTREME OUTLIERS?	FIRST QUARTILE	MEDIAN	THIRD QUARTILE	INTERQUARTILE RANGE
STROKES	NONE	45.55	52.30	59.85	14.3
RESPIRATORY PROBLEMS	NONE	36.85	42.70	50.50	13.65

The distribution of data looks similar as the boxplots overlap, which means that the death rates from strokes and respiratory problems might be similar in every state in America.

The **null hypothesis** chosen is that the difference between the mean death rates from strokes and the mean death rates from respiratory problems is equal to 2, that is, $\mu_1 - \mu_2 = 2$.

The **alternate hypothesis** is that the difference between the mean death rates from strokes and the mean death rates from respiratory problems is greater than 2, that is, $\mu_1 - \mu_2 > 2$.

Through RStudio, we find that the **test statistic value, z**, is **3.35**.

The p value is the probability that the null hypothesis is true. The **p value**, through RStudio, is **0.00041**. Since the p value is below 0.01, we have **very strong evidence to support the claim** that the difference between the mean death rates from strokes and the mean death rates from respiratory problems is **greater than 2**.