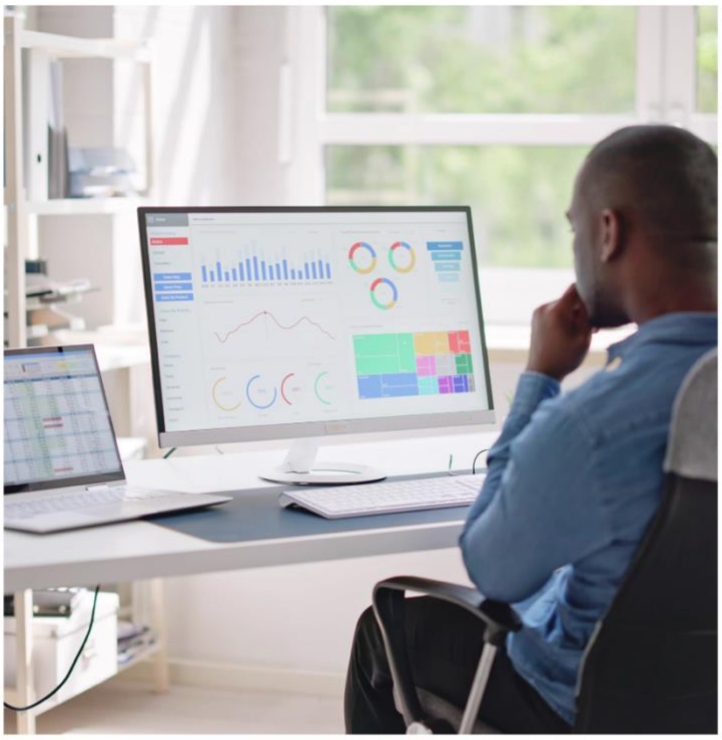
HANDBOOK



HANDBOOK • ONLINE • R

R FOR BEGINNERS

Basic KPI You're Probably Using Wrong: for R Users

THE STATISTICAL CONSULTING FIRM OF SOUTH CAROLINA

Basic KPI You're Probably Using Wrong: *for R Users*

Section 1: What's the Most Representative Value of My Dataset?

When to Use: Measures of center describe the typical value in your dataset. This is the most representative value.

What to keep in mind: When choosing one, consider the shape of the distribution and robustness of the measure itself.

What are my options: Commonly used KPI for center are shown in the table below:

KPI Name	What is it?	When to Use it?	R Code
Mean	The arithmetic average of the data	Use this when you have a symmetric distribution	mean(insert_dataset_name)
Median	The middle-most number of the data	Use this when your distribution is highly skewed (asymmetric)	median(insert_dataset_name)
Mode	The most frequently occurring number or class	Use this with categorical data or when you just want to know the most common value in the data.	No base function exists.
Trimmed Mean	The arithmetic average after removing n% of the highest and lowest values in a dataset	Use this when you have highly skewed or heavy-tailed distributions and you don't want to use the median	mean(insert_dataset_name, trim=_insert_proportion_to_trim)

Section 2: How Much Variability is in My Dataset?

When to Use: Measures of dispersion describe the spread of your dataset.

What to keep in mind: When choosing one, consider the shape of the distribution and robustness of the measure itself

What are my options: Commonly used KPI for dispersion are shown in the table below

KPI Name	What is it?	When to Use it?	R Code
Variance	A measure of how far observations (as a group) vary from the mean	Use this when you have a symmetric distribution. This measure is influenced by outliers, so avoid or use with caution if you have outliers.	var(insert_dataset_name)
Standard Deviation	A measure of how far observations vary from the mean. This is the square root of the variance.	Use this when you have a symmetric distribution. This measure is influenced by outliers, so avoid or use with caution if you have outliers.	sd(insert_dataset_name)
Interquartile Range	The spread of the middle 50% of the data	Use this when your data is highly skewed (asymmetric). This measure is not highly influenced by outliers.	IQR(insert_dataset_name)
Range	The spread of the entire dataset	Use this with a symmetric dataset. This measure is influenced by outliers, so avoid or use with caution if you have outliers.	max(insert_dataset_name) – min(insert_dataset_name)
Midrange	The arithmetic average of the maximum and minimum of a dataset	Use this when you want to find the center of your dataset. This measure is influenced by outliers, so avoid or use with caution if you have outliers.	(max(insert_dataset_name)+min(insert_dataset_name))/2

Section 3: How do the Individual Values Compare to Each Other?

When to Use: Measures of relativity all you to see how data values compare to others within your dataset

What to keep in mind: When choosing a measure, be mindful that these measures only tell a part of your data story. Often you will need to use more than one.

What are my options: Commonly used KPI for relative positioning are shown in the table below:

KPI Name	What is it?	When to Use it?	R Code
Q1	A measure that divides the lower 25% of the data from the upper 75%	This can be used for any numeric distribution	quantile(insert_dataset_name,probs=0.25)
Q2	A measure that divides the data in half	This can be used for any numeric distribution	quantile(insert_dataset_name,probs=0.50)
Q3	A measure that divides the lower 75% of the data from the upper 25%	This can be used for any numeric distribution	quantile(insert_dataset_name,probs=0.75)
Min	The smallest value in the dataset	This can be used for any numeric distribution. This could be an outlier in your dataset.	min(insert_dataset_name)
Max	The largest value in the dataset	This can be used for any numeric distribution. This could be an outlier in your dataset.	max(insert_dataset_name)

Note: you can get all relative positioning measures with fivenum(insert_dataset_name) as well.