




Exploratory Data Analysis

SGA07_DATASCI

28th January 2020



Module Overview

- Statistics 101
- Hypotheses Testing
- Understand Data through Graphs



Book Keeping

- Direct any technical questions to TA
- Spend some time to build programming skills
- Expect 2 mini group-based projects
- Catch up on Tasks/Practice Labs so far



Outcome

After this Module, you will;

- Get a refresher on basic concepts of statistics
- Understand the differences between population and sample
- The importance of counting
- Brief overview of estimates (mean, median, mode & quantiles)
- Understand relationship in data through graphical methods



Population

- All new born babies in Nigeria
- All loan application in Union Bank
- Energy consumption of each appliance within a house

“

Population is the entire pool from which statistical inference maybe drawn. It refers to an entire group of events, entities, measurements or objects that is being observed for particular features/attributes.

”



Sample

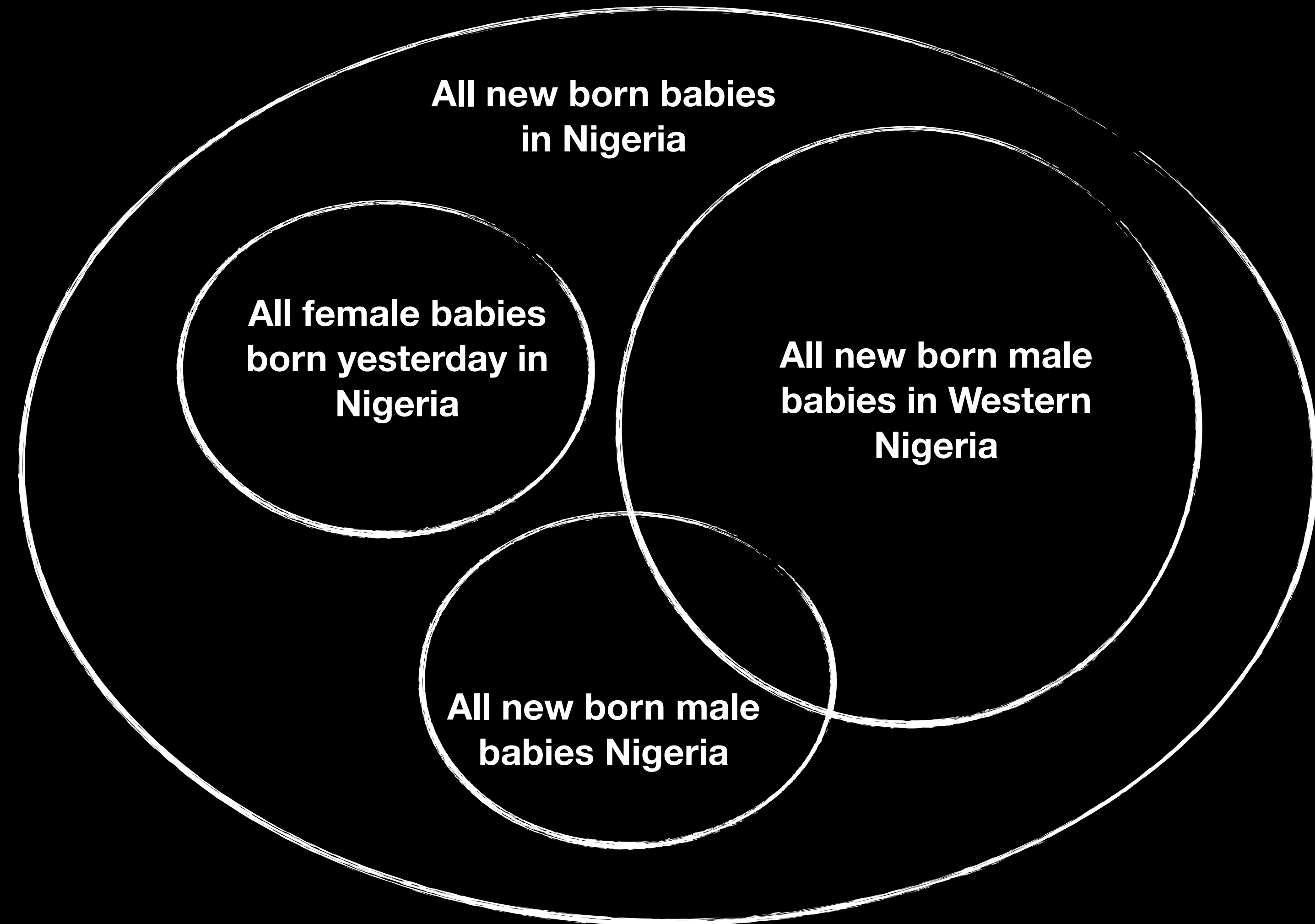
- All new male born babies in Western Nigeria
- Small ticket and payday loan application no later than 2018 in Union Bank

“

A sample is a random or non-random selection of members of a population. It refers to a smaller group drawn from the population that provides a representation of the characteristics of the population.

”

Sample



Tables

“

The rectangular data or table contains features and records in essentially a two-dimensional matrix. The feature (attribute or variable) refers to column of the matrix while the record (observation or instance) refers to rows of the matrix.

”

	mpg	cyl	displacement	horsepower	drat	weight	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4

Counting

“

The number of instances of a single attributes or the instances that meet the conditions of a number of attributes.

”

	Var1	Freq
1	0	18
2	1	14

	Var1	Freq
1	1	7
2	2	10
3	3	3
4	4	10
5	6	1
6	8	1

	Var1	Freq
1	4	11
2	6	7
3	8	14



Five Number Summary

- Mean
- Median
- Mode
- Percentile
- Variance
- Standard Deviation
- Correlation Matrix
- Minimum / Maximum

Minimum

“

This is the lowest/smallest/least
element in a sample

”

$$\text{Min}(2, 8, 10, 6, 2, 5, -1, 7, 34, 2, 0, -3) = -3$$



Maximum

“

This is the highest/biggest/
greatest element in a sample

”

$$\text{Max}(2, 8, 10, 6, 2, 5, -1, 7, 34, 2, 0, -3) = 34$$

Mean

“

This provides the most basic estimate of location. It is the sum of all the values divided by the number of observations

”

$$\text{Mean}(2, 8, 10, 6, 2, 5, -1, 7, 34, 2, 0, -3) = \frac{\sum_1^n x_i}{n} = \frac{72}{12} = 6$$

Percentile

“

This helps to indicate the value below which a given percentage of observations in a group of observation falls. First is to rank your data in ascending order and then apply the percentile formula

”

$$n = \left[\frac{P}{100} * N \right]$$

Percentile	Rank Order	Order Rank n	Value
5th	(-3, -1, 0, 2, 2, 2, 5, 6, 7, 8, 10, 34)	$n = \left[\frac{5}{100} * 12 \right] = 0.6 = 1$	-3

Median

“

The median provides the estimate centre of location using the 50th percentage of observations in a group of aberrations. Similar to the percentile, first you rank your data in ascending order and then apply the percentile formula

”

$$\text{Median}(2, 8, 10, 6, 2, 5, -1, 7, 34, 2, 0, -3) = \left[\frac{50}{100} * 12 \right] = 2$$

Mode

“

The mode of a set of data values is the value that appears most often (i.e the element with the highest frequency)

”

$$\text{Mode}(2, 8, 10, 6, 2, 5, -1, 7, 34, 2, 0, -3) = 2$$



Practice Lab

Carryout basic statistics on mtcars data In R

Use the following Instructions:

- Create a new script edm,r in your root directory
- Get the mtcars data into a data frame
- Get the dimension and names of attributes in the dataframe
- Get sample of the dataframe
- Get count of categorical attributes and five number summary of quantitative attributes of the data frame

Hypotheses Testing

- State the initial research hypothesis of which the truth is unknown.
- State the relevant null and alternative hypotheses. This is important, as misstating the hypotheses will muddy the rest of the process.

“

Statistical inference allows us to make propositions about a population based on estimates of samples. This is done through hypothesis testing which compares the statistical relationship between sample dataset towards a generalised idea/fact about the populations as observed.

”



Hypotheses Testing (Contd.)

- Consider the statistical assumptions being made about the sample in doing the test; for example, assumptions about the statistical independence or about the form of the distributions of the observations. This is equally important as invalid assumptions will mean that the results of the test are invalid.
- Decide which test is appropriate, and state the relevant test statistic T .



Hypotheses Testing (Contd.)

- Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. For example, the test statistic might follow a Student's t distribution or a normal distribution.
- Select a significance level (α), a probability threshold below which the null hypothesis will be rejected. Common values are 5% and 1%.



Hypotheses Testing (Contd.)

- The distribution of the test statistic under the null hypothesis partitions the possible values of T into those for which the null hypothesis is rejected—the so-called critical region—and those for which it is not. The probability of the critical region is α .
- Compute from the observations the observed value of the test statistic T .



Hypotheses Testing (Contd.)

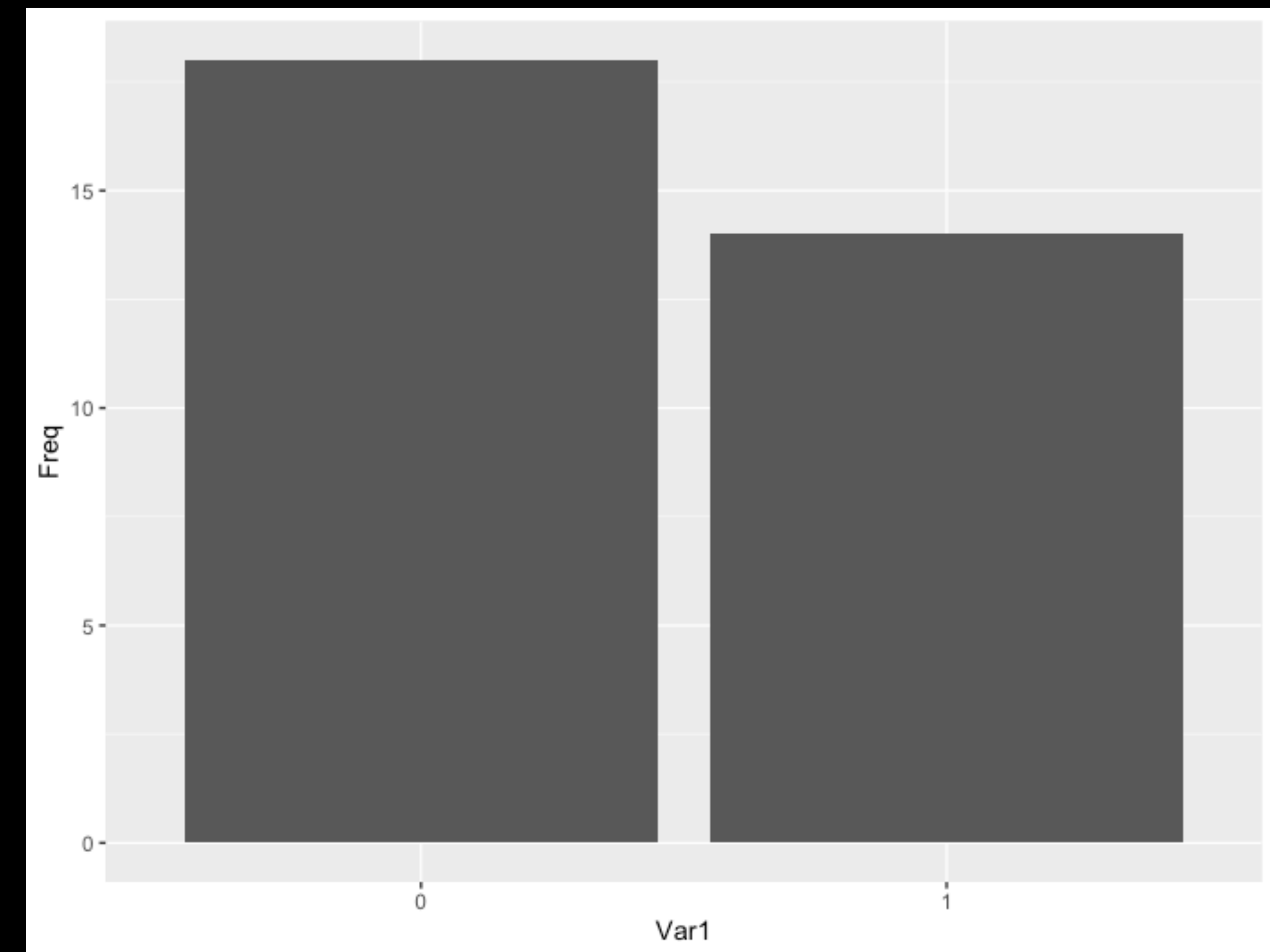
- Decide to either reject the null hypothesis in favour of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 if the observed value is in the critical region, and to accept or "fail to reject" the hypothesis otherwise.

Barchart

“

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

”

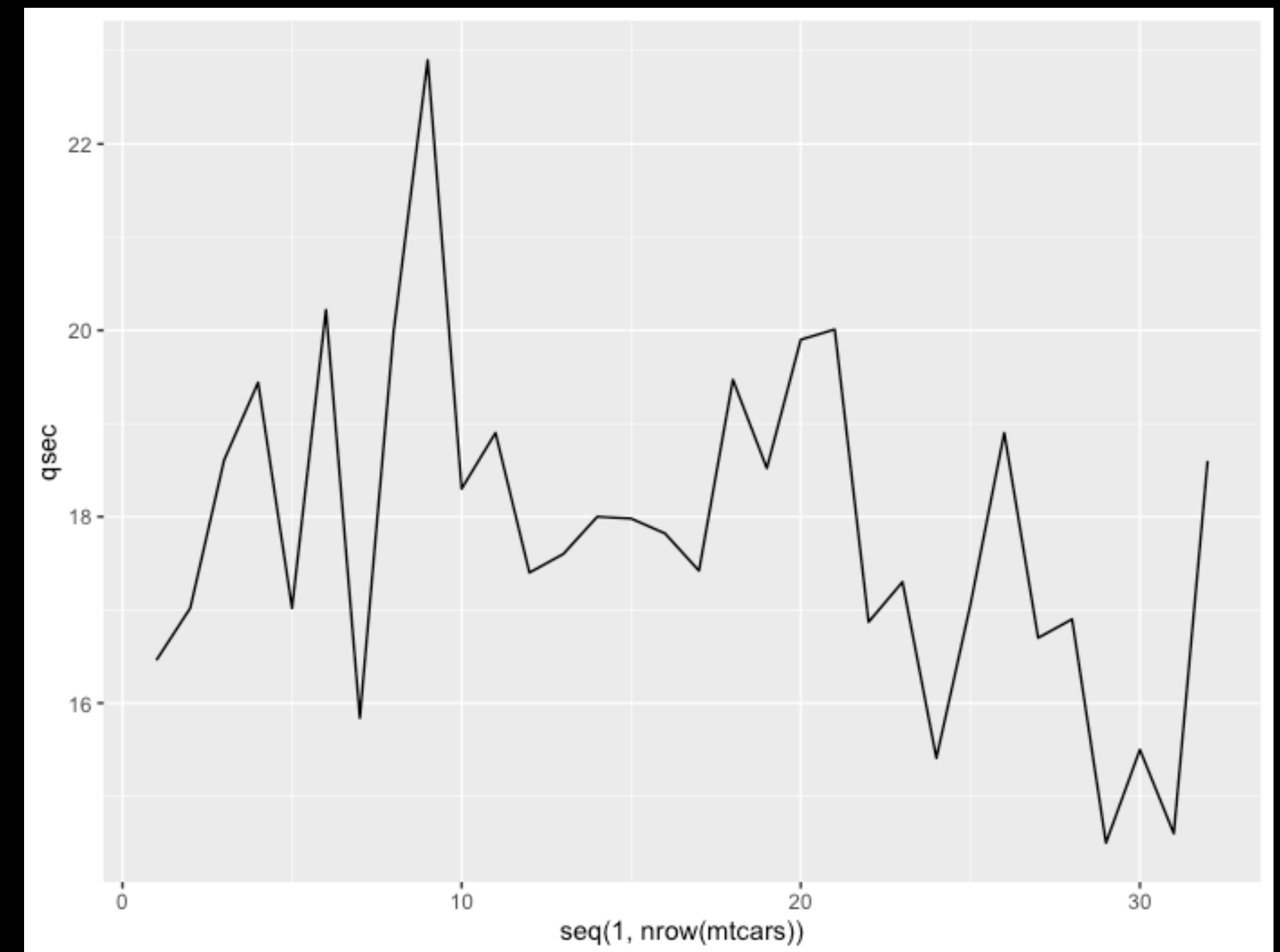


Line chart

“

A line chart or line plot or line graph or curve chart[1] is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.

”

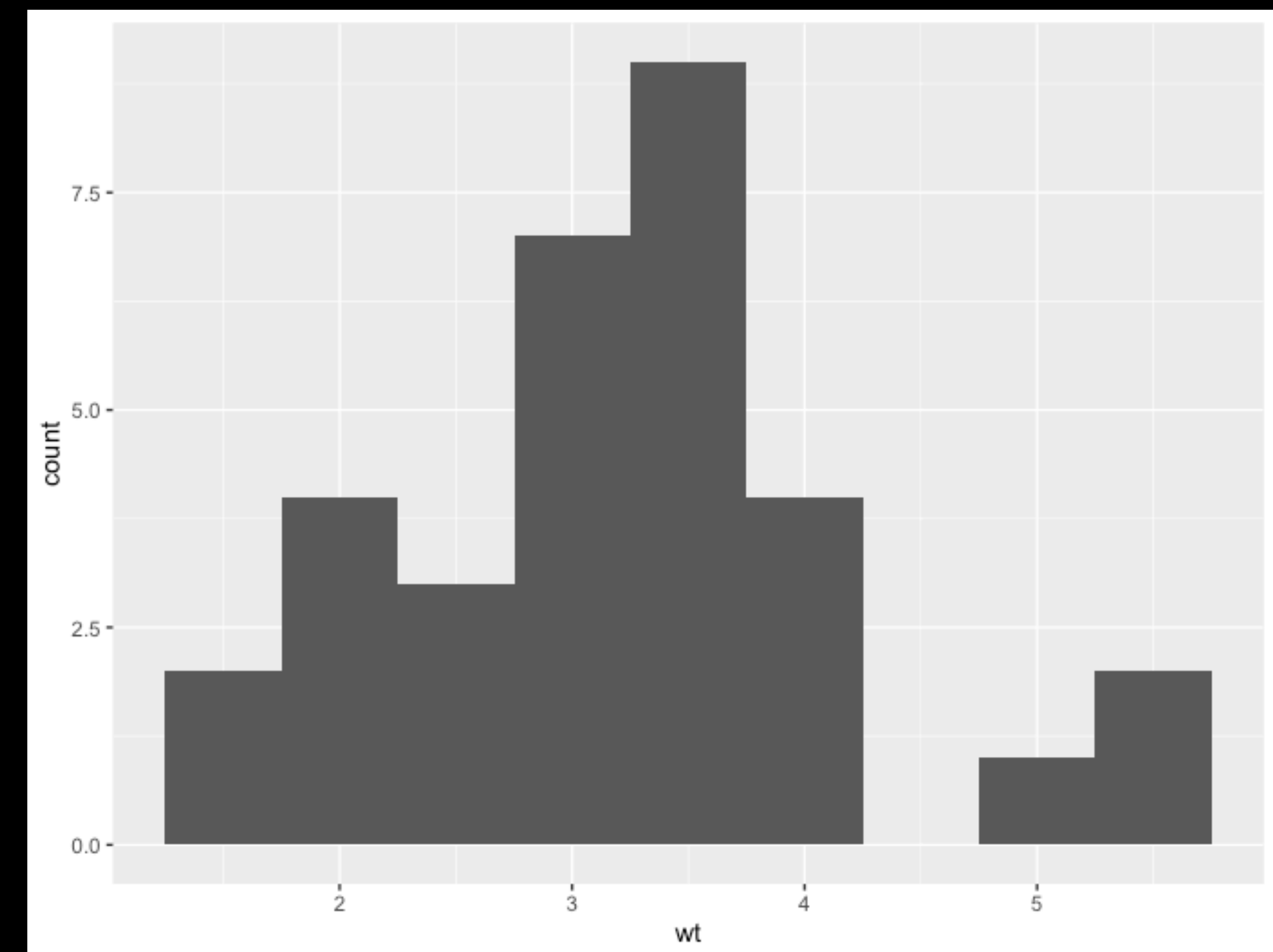


Histogram

“

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable.

”

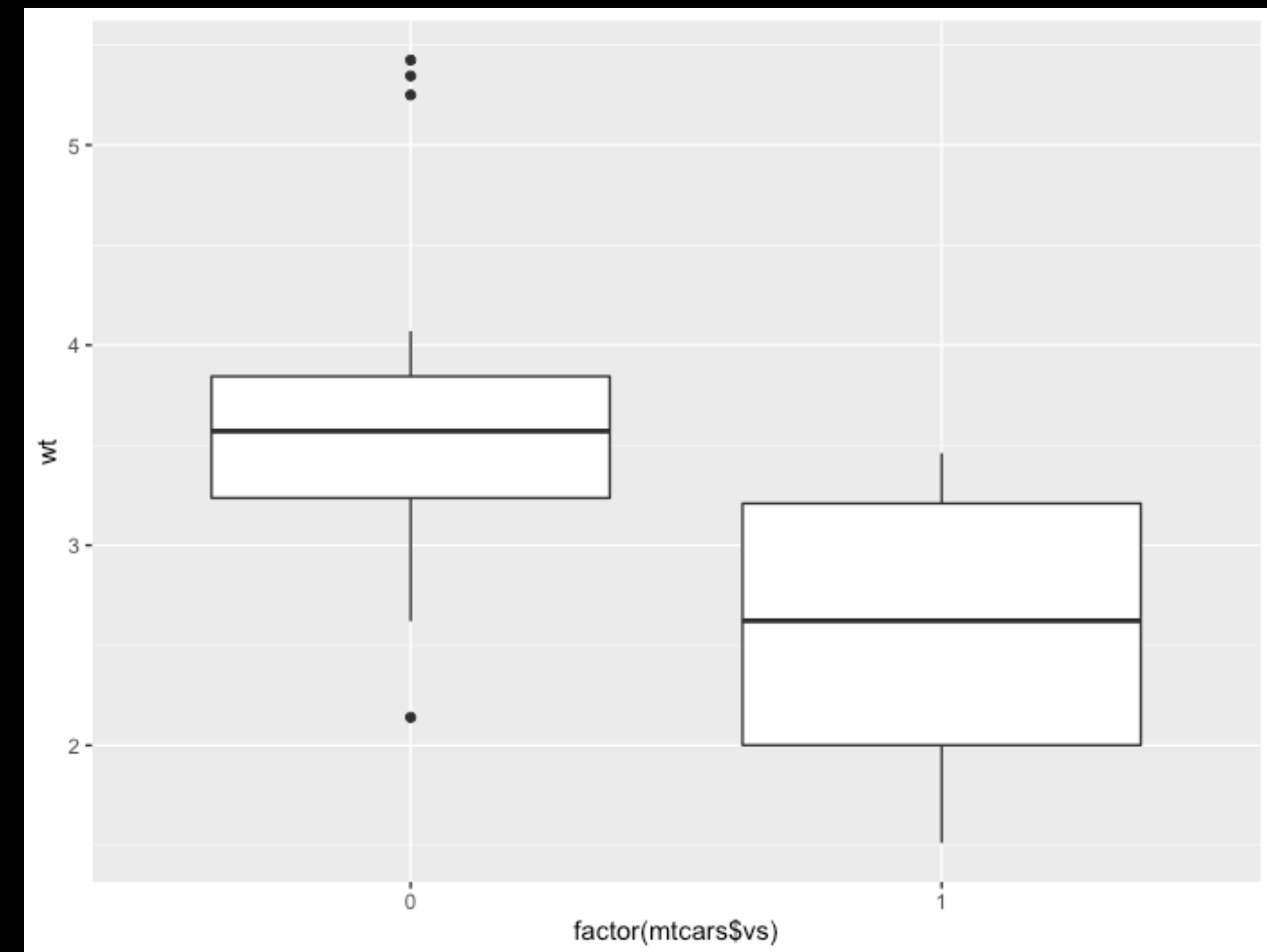


Boxplot

“

A box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles.

”

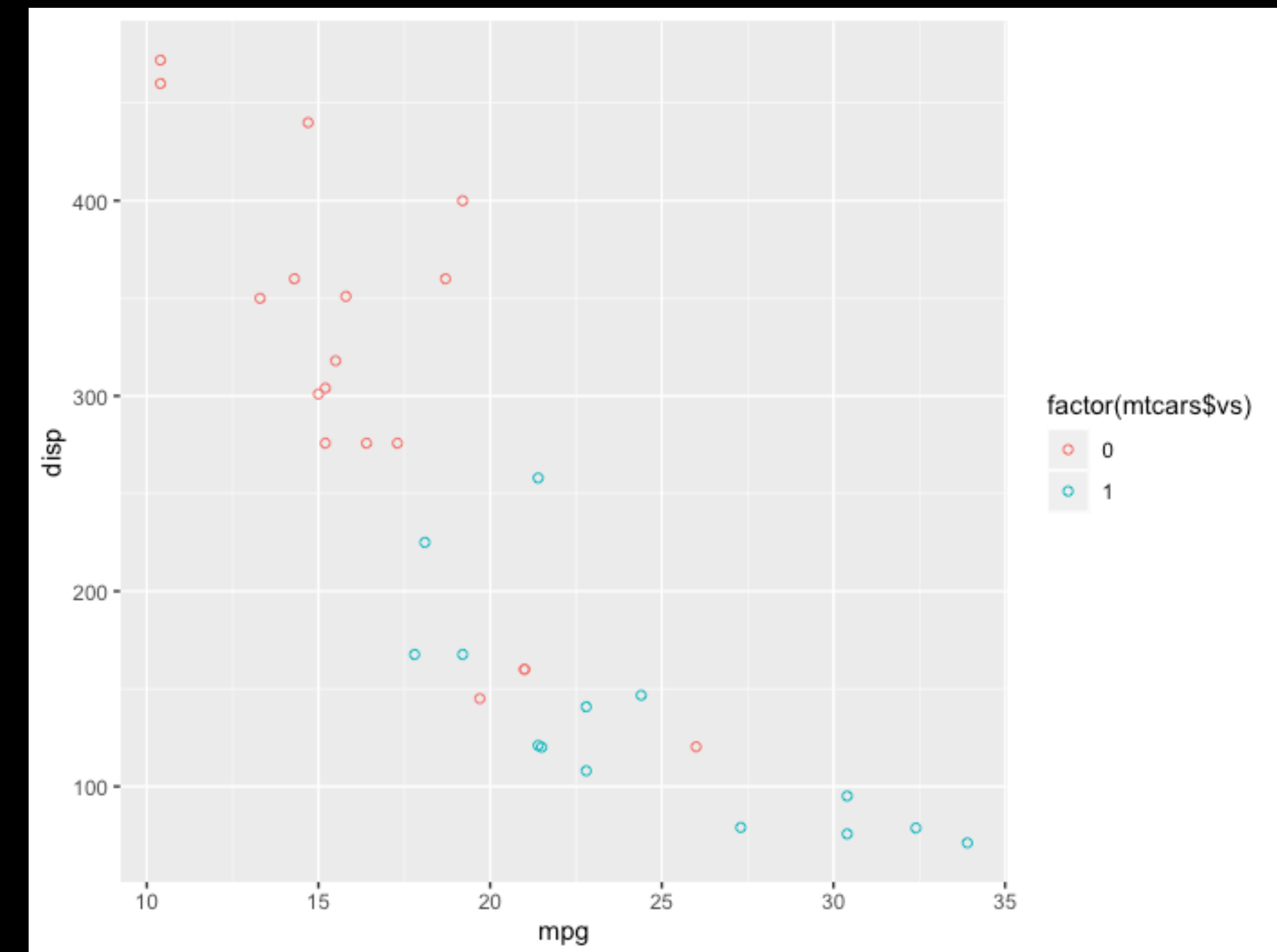


Scatterplot

“

A scatter plot (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data

”





Practice Lab

Use graphical methods to better understand your data

Use the following Instructions:

- Use graphical methods to gain better intuition of the mtcars dataset
- Explore univariate analysis for categorical and quantitative attributes
- Explore bivariate analysis for categorical and quantitative attributes
- Explore multivariate analysis for categorical and quantitative attributes
- Draft an Exploratory Data Analysis report of insights derived from mtcars dataset



Recap/Summary

At the end of this Module, you should understand;

- Get a refresher on basic concepts of statistics
- Understand the differences between population and sample
- The importance of counting
- Brief overview of estimates (mean, median, mode & quantiles)
- Understand relationship in data through graphical methods



Suggested Material (Programming)

- <https://www.edx.org/course/cs50s-introduction-to-computer-science>
- <https://www.hackerearth.com/practice/python/functional-programming/functional-programming-1/tutorial/>
- <https://www.coursera.org/learn/python>



Suggested Material

- Exploratory Data Analysis by John Tukey (Pearson)
- The Future of data Analysis by John Tukey, Annals of Mathematical Statistics, Volume 33, Number 1 (1962), 1-67
- Freedman, David, Robert Pisani, & Roger Pervis (2007). Statistics. New York: W.W. Norton.
- James, Gareth, Daniela Witten, Trevor Hastie, & Robert Tibshirani (2013). An Introduction to Statistical Learning: With Applications in R. New York: Springer.
- The Elements of Graphing Data by William S. Cleveland (Hobart Press)