Farhoud Khoshnoud

[Company Name]

R

Table of Contents

[Complete Guide to R: Wrangling, Visualizing, and Modeling Data 2](#_Toc199170778)

[# Intro 2](#_Toc199170779)

[#1 What is R 2](#_Toc199170780)

[#2 Getting Started 3](#_Toc199170781)

[#3 Importing Data 3](#_Toc199170782)

[#4 Visualizing Data With ggplot 4](#_Toc199170783)

[#5 Wrangling Data 4](#_Toc199170784)

[#6 Recoding Data 9](#_Toc199170785)

[#7 An R for Data Science Case Study 11](#_Toc199170786)

[#8 Exploring Data 12](#_Toc199170787)

[#9 Analysing Data 15](#_Toc199170788)

[#10 Predicting Outcomes 18](#_Toc199170789)

[#11 Clustering and Classifying Cases 21](#_Toc199170790)

[R for Excel Users 24](#_Toc199170791)

[#2 Descriptive Statistics in Excel 24](#_Toc199170792)

[Question 2 of 3 29](#_Toc199170793)

[R for Data Science: Analysis and Visualization 31](#_Toc199170794)

[# 31](#_Toc199170795)

[# 31](#_Toc199170796)

[# 31](#_Toc199170797)

[# 31](#_Toc199170798)

# Complete Guide to R: Wrangling, Visualizing, and Modeling Data

=============================

## # Intro

=============================

## #1 What is R

**Question 1 of 4**

You are a data scientist working on a project that involves creating complex statistical models. Which tool should you consider moving beyond spreadsheets for this task?

* a basic text editor with no data functionality
* a graphic design software for enhanced visual representation
* an audio editing tool for data representation
* a statistical application like Jamovi or SPSS

Correct

**Question 2 of 4**

While exploring the relationship between happiness and marital status, what variable transformation was applied to simplify the analysis?

* Marital status was ignored in favor of analyzing age groups.

Incorrect

* Marital status was collapsed into a dichotomous married vs. not married variable.
* Marital status categories were reordered based on happiness levels.

Incorrect

* Marital status was expanded to include more specific categories.

Incorrect



Replay

Review this video

Data science with R: A case study

11m 46s

**Question 3 of 4**

Which of the following is a reason why spreadsheets are considered effective tools for certain data tasks?

* They automatically cleanse data for errors and inconsistencies.
* Spreadsheets are the only tool capable of handling large datasets.
* They are the most effective way to perform machine learning tasks.
* They allow for data to be organized, sorted, and visually represented through graphs.

Correct

**Question 4 of 4**

As a researcher examining data from the General Social Survey, you're interested in understanding the impact of financial status on happiness. After visualizing the data, what conclusion might you draw?

* Financial status has no discernible impact on happiness.
* There appears to be a linear relationship where as financial status increases, so does the proportion of individuals reporting they are "very happy."

Correct

* Individuals with below-average financial status report the highest levels of happiness.
* The relationship between financial status and happiness is inversely proportional.

=============================

## #2 Getting Started

Control + l to clear

rm(list = ls())

# default pat: /Users/farhoudkhoshnoud

setwd()

setwd('/Users/farhoudkhoshnoud/Library/CloudStorage/OneDrive-Personal/Self Study/LinkedIn Learn Exercise/R/Complete Guide to R Wrangling\_ Visualizing\_ and Modeling Data/Exercise Files')

pipe

%>%

=============================

## #3 Importing Data

=============================

## #4 Visualizing Data With ggplot

=============================

## #5 Wrangling Data

tibbles

data.table

Wide Tall – package MASS

**Question 1 of 21**

Maria is preparing a dataset for analysis that she obtained from a public API. The data is formatted in JSON, which makes it difficult for her to apply traditional data analysis methods directly. What step should Maria take first to make this data more manageable for analysis in R?

* Convert the JSON data into a tidy format with rows and columns

Correct

* Immediately start analyzing the JSON data using R's JSON parsing capabilities
* Use a text editor to manually extract the values from JSON into a CSV file
* Transfer the JSON data into a spreadsheet manually and then import it into R

**Question 2 of 21**

Which command is used to convert a data frame into a tibble in R?

* as\_tibble()

Correct

* convert\_to\_tibble()
* to\_tibble()
* df\_to\_tibble()

**Question 3 of 21**

You are working on cleaning a dataset that contains employee names in a single column, with first and last names separated by a space. What R function would you primarily use to separate these names into two distinct columns for "First Name" and "Last Name"?"

* gather()
* unite()
* spread()
* separate()

Correct

**Question 4 of 21**

While analyzing the trend of sunspot occurrences over decades, you decide to calculate the mean number of sunspots for each decade to simplify your analysis. What process should you follow to achieve this in R?

* Use a linear regression model to predict sunspot numbers for each decade
* Group the data by decades and calculate the mean for each group

Correct

* Concatenate the yearly data into decade-long strings and then calculate an overall mean
* Apply a filtering function to select rows corresponding to each decade start

**Question 5 of 21**

In data analysis, why is the ability to reshape data considered critical?

* It simplifies the data to only numerical values for easier calculation.
* It allows for the data to be adjusted to meet the needs of specific analyses or graphics.

Correct

* It enhances the security of the data by encoding variables.
* It increases the overall amount of data available for analysis.

**Question 6 of 21**

You are analyzing demographic data for a research project and realize the format in which satisfaction levels are recorded does not suit your analysis. Your dataset lists satisfaction levels alongside each demographic group, but you need to compare the satisfaction levels across all groups simultaneously. Which tidyverse function would you most likely use to adjust the dataset to fit your needs?

* spread
* gather

Incorrect

* filter

Incorrect

* mutate

Incorrect



Replay

Review this video

Converting data from wide to tall and from tall to wide

4m 13s

**Question 7 of 21**

You are plotting the growth of various orange trees over time and want to ensure the tree ID numbers are in sequence to make analysis easier. Which R function would you most likely use based on the script?

* sort()
* sequence()
* arrange()
* reorder()

Correct

**Question 8 of 21**

Which R package is highlighted for its efficiency in handling large datasets?

* tidyr
* data.table

Correct

* dplyr
* ggplot2

**Question 9 of 21**

You're analyzing a dataset related to the United States, specifically focusing on the state level data including demographics, political leanings, and Google search trends. Which R package facilitated the import of StateData.xlsx for further analysis?

* rio

Correct

* tidyr
* dplyr
* ggplot2

**Question 10 of 21**

In conducting a study on U.S. states based on their Google search preferences, Anaya is interested in selecting states that demonstrate a significantly higher interest in entrepreneurship compared to the national average. Which statistical marker indicates a state's searches for "entrepreneur" are above the national average?"

* Median search values

Incorrect

* Standard deviation from search mean

Incorrect

* Average search percentage

Incorrect

* Z scores of one or above



Replay

Review this video

Filtering cases and subgroups

7m 32s

**Question 11 of 21**

What is the primary reason for converting a character variable into a factor within the R programming environment?

* It allows for the application of functions specific to categorical data.

Correct

* It increases the processing speed of data analysis.
* It reduces the memory usage of the dataset.
* It converts all textual data into numerical data, facilitating mathematical operations.

**Question 12 of 21**

You are analyzing a dataset about a school's student population and decide to group the students by age into 'Younger' and 'Older'. If you want 'Younger' to appear first when the data is analyzed or visualized, which step is essential after converting the age groups into factors?

* Convert age from a numerical to a character variable before creating age groups.
* Arrange the dataset in ascending order based on age before factor conversion.
* Use the ifelse statement to assign 'Younger' and 'Older' categories.

Incorrect

* Manually specify the order of the factor levels with 'Younger' first.

Correct

**Question 13 of 21**

In managing time series data for analysis, which method involves transforming the data into a format where each row represents a unique time period (e.g., a month or a year) and each column represents a single variable?

* Utilizing merged cells in spreadsheets for time-based data
* Converting data into a tidy time series format

Correct

* Applying formulas across spreadsheets to aggregate time data
* Merging multiple data sets based on time intervals

**Question 14 of 21**

Zara is analyzing the daily closing prices of major European stock indices between 1991 and 1998. She needs to transform the dataset's date format for better readability and to chart the data effectively. Which approach should Zara take?

* Convert the dataset into a JSON file and parse the dates with JavaScript.
* Leave the dates in their original format since it is already suitable for analysis.
* Save the dataset as a tsibble and separate the date into year, month, and day.

Correct

* Store the dates as strings and use regular expressions to extract the date parts.

**Question 15 of 21**

What is the primary reason for converting lists to a more structured data format, such as a tibble, in R?

* to reduce the amount of memory needed for data storage
* to simplify the data for analysis and make insights easier to obtain

Correct

* to increase the processing speed for data manipulation tasks
* to enable the creation of complex data types not supported by lists

**Question 16 of 21**

You, a data scientist, are working on the UCB Admissions dataset to identify trends in graduate admissions. What function would you use to simplify the process of converting the frequency of admissions into individual observations?

* uncount

Correct

* repeat
* lapply
* tibble

**Question 17 of 21**

Which R package is specifically highlighted for working with dates and times to facilitate data analysis?

* lubridate

Correct

* tidyverse

Incorrect

* tsibble

Incorrect

* ggplot

**Question 18 of 21**

You are analyzing financial information from a dataset similar to the Missouri data portal. Your goal is to compare sales tax rates between counties. Which initial step should you prioritize to handle the hierarchically structured data effectively?

* Immediately convert the entire dataset into a numeric variable type.
* Import the XML data and saving it as a list.

Correct

* Create a histogram to observe the distribution of sales tax rates.
* Squish white space in the county column to ensure data cleanliness.

**Question 19 of 21**

After obtaining fluency information in R for a data set representing languages known by job applicants, what method is used to calculate the number of languages an applicant is fluent in?

* Sum the logical TRUE values representing fluency.

Correct

* Use a frequency table to manually count the TRUE values.
* Apply a filter function to count rows where fluency is marked as TRUE.
* Run a complex algorithm to parse through the data frame and count languages.

**Question 20 of 21**

Why is the unnest\_wider function utilized in the process of handling XML data in R?

* to increase the performance speed when processing large XML files
* to convert XML data directly into a data frame without needing lists
* to transform nested lists into separate variables for easier analysis

Correct

* to filter out unnecessary data before starting the analysis

**Question 21 of 21**

Why is it necessary to convert data from table format to a row-by-row format for analysis in R?

* It allows for broader analytical approaches and graphics creation.

Correct

=============================

## #6 Recoding Data

Question 1 of 9

In an effort to analyze time to graduation data, which method would be most appropriate to handle outliers that excessively extend beyond the typical range of years students take to graduate?

* Ignore outliers and compute the average time to graduation.

Incorrect

* Exclude all data points that represent graduation times beyond four years.

Incorrect

* Transform all time to graduation data points by squaring them.

Incorrect

* Winsorize the data to cap the graduation times at a maximum value.



Replay

Review this video

Transforming outliers

8m 49s

Question 2 of 9

You work in a data analytics team and are tasked with investigating the influence of cultural interests across states in the U.S.A. How could the approach demonstrated with the "likeArts" variable assist your research?"

It could serve as a tool to predict future trends in arts interest on a state-by-state basis.

It can precisely calculate the economic impact of arts on each state.

It would allow for the direct comparison of cultural interests between U.S. states and other countries.

It could help identify states with high interest in arts by creating a yes/no variable based on z-scores of search terms.

Correct

Question 3 of 9

In the process of averaging scores from various measurements, what is the primary benefit sought by researchers?

to identify the highest scoring variables for focused analysis

to reduce the idiosyncratic variation and clarify the intended signal

Correct

to make the data easier to analyze through simple statistical methods

to increase the overall scores by combining different scales

Question 4 of 9

What is the significance of z-scores being positive or negative in the context of analyzing state-by-state popularity of search terms?

Positive scores are used exclusively for arts-related searches, while negative scores apply to science-related queries.

Positive scores denote a high level of accuracy in the data, while negative scores highlight errors.

Positive scores indicate the term is trending downwards, while negative scores suggest an upward trend.

Positive scores indicate more searches for a term compared to other states, while negative scores indicate fewer searches.

Correct

Question 5 of 9

In the analysis of customer engagement, what method did the instructor highlight as an effective way to gain insights?

Compare engagement rates to competitors.

Count how many times they have been engaged.

Correct

Perform complex predictive modeling on engagement data.

Segment customers into groups based on geographic location.

Incorrect

Question 6 of 9

In data preparation, if a researcher needs to adjust scores on a 1-7 scale where a low score indicates more of something but wants the direction aligned so that a high score indicates more, what calculation should they perform?

Subtract the score from 8.

Correct

Subtract the score from 1.

Divide the score by 7.

Add 7 to the score.

Question 7 of 9

You are conducting a study with Lin, who is analyzing personality tests like the MMPI. Lin wants to adjust the scores to a scale where the mean is 50 and the standard deviation is 10. After standardizing the data to Z-scores, what should Lin's next step be?

Add 50 to the Z-scores and then divide by 10.

Multiply the Z-scores by 10 and then add 50.

Correct

Divide the Z-scores by 10 and then subtract 50.

Multiply the Z-scores by 50 and then add 10.

Question 8 of 9

In the process of data cleaning and manipulation, why would you want to re-level categorical variables?

to eliminate outliers from the dataset completely

to rearrange the order in which categories appear for analysis or visualization

Correct

to increase the computational efficiency of data analysis

to convert numerical data into categorical data

Incorrect

Question 9 of 9

You are analyzing a dataset on the popularity of mobile operating systems in 2023. You notice that there are several systems with a very small market share that are not of interest to your current analysis. Which function would you use to combine these into a single 'Other' category to simplify your analysis?

* as\_factor to convert numeric data pertaining to operating systems into categorical data.
* fct\_lump with n=3 to keep the top three operating systems distinct and lump the rest into 'Other'.

Correct

=============================

## #7 An R for Data Science Case Study

Question 1 of 2

You are examining the relationship between extroversion and openness in a dataset that includes both traits among others. To assess if there is a significant difference in extroversion between individuals high and low in openness, which statistical test would be most appropriate?

* ANOVA
* Chi-square test
* Factor analysis
* T-test

Correct

Question 2 of 2

Mia is working on a large dataset pertaining to individual personality traits and decides to use a machine learning approach to predict levels of openness based on other factors. She splits her data into training and testing sets and wants to use a method that creates multiple decision trees to improve prediction accuracy. Which method should Mia use?

* K nearest neighbors (KNN)

Incorrect

* Random Forest

Correct

=============================

## #8 Exploring Data

Frequency

Descriptive statistic

Principle Component Analysis or Factor Analysis

Item Analysis

Confirmatory Factor Analysis (CFA)

Question 1 of 13

When conducting an exploratory data analysis on a dataset measuring large personality factors, what statistical measure is used to determine if items within a factor can be reliably averaged together?

* Mean Square Error
* Cronbach's Alpha

Correct

* Standard Deviation
* Pearson Correlation Coefficient

Question 2 of 13

What is indicated by a correlation coefficient value close to 0?

* There is a perfect negative linear association between the variables.
* The variables are inversely proportional to each other.
* There is a strong positive linear association between the variables.
* There is no straight line association between the variables.

Correct

Question 3 of 13

You are analyzing data from a new marketing campaign. After creating a correlation matrix, you find a coefficient of -0.9 between the number of clicks and the number of unsubscribes. What does this indicate about the relationship between these two variables?

* The number of clicks and unsubscribes are perfectly correlated.
* Higher number of clicks is associated with a lower number of unsubscribes.

Correct

* Higher number of clicks is associated with a higher number of unsubscribes.
* The number of clicks has no association with the number of unsubscribes.

Question 4 of 13

When analyzing a data set in R, what function would you use to select a single variable for summary statistics?

* mutate
* select

Correct

* summary
* pull

Question 5 of 13

What does the box plot statistic "third quartile" represent in data analysis?"

* the 75th percentile score
* the median value

Incorrect

* the lowest non-outlier score

Incorrect

* the highest outlier score

Incorrect



Replay

Review this video

Computing descriptive statistics

9m 42s

Question 6 of 13

In the context of analyzing personality data, which method was applied to determine the appropriate number of components to use while accounting for each variable contributing to only one component?

* Principal Component Analysis (PCA)
* Hierarchical Clustering
* Very Simple Structure (VSS) analysis

Correct

* Gradient Projection Algorithm Rotation (GPA rotation)

Question 7 of 13

Chioma has collected data on 50 variables relating to employee satisfaction within her organization. She intends to reduce the dimensionality of her dataset for a clearer analysis of the contributing factors. Which of the following methods should she consider using to summarize the variability in her data effectively?

* T-test
* Analysis of Variance (ANOVA)
* Principal Component Analysis (PCA)

Correct

* Regression Analysis

Question 8 of 13

What function in R is used to provide row percentages in a contingency table analysis?

* prop\_table with the margin argument set to 1

Correct

* freq() with a row specification

Incorrect

* rowSums on the contingency table
* factor\_recode to adjust row data

Question 9 of 13

You are conducting a study on regional preferences for coffee flavors in the United States and decide to use a contingency table for your analysis. After compiling your data, which R function would be ideal for comparing the percentage distribution across different regions?

* chisq.test for initial data compilation
* apply() with a sum function
* table() to first compile the raw frequencies
* prop\_table with an appropriate margin argument

Correct

Question 10 of 13

Which function in R provides a list of frequencies for character variables?

* table

Correct

* select

Incorrect

* mutate
* summary

Incorrect

Question 11 of 13

You are analyzing a dataset that contains information on different US states, including categorical variables like region and psychRegion. To ensure R treats these variables appropriately for your analysis, what should you transform these variables into?

* factors

Correct

* integers
* strings
* booleans

Question 12 of 13

What is the main purpose of using the Lavaan package in R for data analysis?

* to simplify data organization and cleaning processes
* to increase the speed of large database processing
* to enhance the graphical representation of data analysis results
* to perform confirmatory factor analysis and structural equation modeling

Correct

Question 13 of 13

You are reviewing the results of a confirmatory factor analysis that shows comparative fit index (CFI) and Tucker-Lewis index (TLI) values of 0.735 and 0.748, respectively. What can you infer about the model fit based on these values?

* The model fit is not ideal and could indicate discrepancies between the theoretical model and the observed data.

Correct

* The values indicate a perfect fit between the theoretical model and the observed data.
* Such values suggest that the model is overfitting to the dataset.
* These indices do not provide useful information regarding model fit.

=============================

## #9 Analysing Data

Question 1 of 12

When creating a dataset in R that represents hypothetical sales of products before and after a marketing campaign, what function is used to generate the sales data with a normal distribution?

* tibble
* rnorm

Correct

* ggparcoord
* ggplot

Question 2 of 12

You are examining the effects of a new study method on student performance over a semester. To measure the difference in scores from the beginning to the end of the semester for the same group of students, which statistical test would be most suitable?

* Paired samples t-test

Correct

* Chi-square test

Incorrect

* ANOVA
* Independent samples t-test

Question 3 of 12

Which of the following is a key benefit of plotting a histogram and box plot before conducting a t-test on earthquake magnitude data?

* It automatically corrects any errors in the dataset.
* It reduces the need for inferential statistics.
* It helps in understanding the distribution and identifying outliers.

Correct

* It directly calculates the mean magnitude for the dataset.

Question 4 of 12

In analyzing lung cancer study data, what is the primary purpose of recoding the variables "status" and "sex" into more readable forms?"

* to reduce the size of the data set for faster processing
* to increase the computational efficiency of the analysis
* to facilitate easier interpretation and analysis of the data

Correct

* to make the data suitable for non-statistical analyses

Question 5 of 12

In the context of performing an independent samples T-test, what does specifying an alternative hypothesis as "less" indicate about the researcher's expectation?"

* The researcher anticipates the sample size of one group to be less than the other.
* The researcher expects that there will be no difference between the group means.
* The researcher believes the means of both groups will be equal.
* The researcher expects the mean of one group to be less than the mean of the other group.

Correct

Question 6 of 12

You are given a dataset similar to the NCCTG Lung Cancer Study and decide to assess the relationship between survival status and gender using R. Which test would you initially utilize to determine if there is a statistically significant difference between genders in terms of survival?

* a linear regression analysis between gender and survival time
* chi-squared test on a two-by-two frequency table

Correct

* a multivariate analysis of variance (MANOVA) on all available variables
* an independent samples t-test comparing age differences by gender

Question 7 of 12

Why would one use an Analysis of Variance (ANOVA) instead of a t-test when comparing groups?

* because ANOVA is simpler to perform than a t-test
* to prevent overfitting in a machine learning model
* to compare the means of more than two groups simultaneously

Correct

* aNOVA requires less data to compare groups effectively

Question 8 of 12

What is the primary purpose of converting the drug variable in the Student's Sleep Data dataset to a factor in R?

* to allow the categorization of drug types for analysis

Correct

* to reduce the dataset size by simplifying variables
* to increase the computational speed of analysis
* to convert the dataset to a tibble format

Question 9 of 12

You are conducting a study on the impact of two different study methods on students' test scores. Following the example of the independent samples T-test used in the sleep study data, what would be your initial approach to analyze the difference in mean test scores between the two study methods?

* Use a chi-square test to see the distribution of scores across the study methods.
* Apply a Pearson correlation to determine the relationship between study methods and test scores.
* Generate density plots for each individual's test scores without comparing groups.

Incorrect

* Use an independent samples T-test to compare the mean test scores of the two groups.

Correct

Question 10 of 12

When analyzing earthquake magnitudes in Fiji using R, what is the first step after extracting the magnitude vector from the dataset?

* Immediately start performing the t-test.
* Use glimpse or print to view a section of the data.

Correct

* Calculate the mean and median of the entire dataset.
* Generate a scatter plot to identify trends.

Question 11 of 12

What is the primary purpose of using a factorial analysis of variance in data analysis?

* to reduce the computational complexity of the analysis
* to evaluate the effects of more than one factor on an outcome variable

Correct

* to ensure a higher significance level in the results
* to solely predict the outcome variable based on one factor

Question 12 of 12

Akira is analyzing data to understand how different types of soil (sandy, loamy, and clay) and amounts of water (low, medium, high) affect plant growth. She intends to identify whether the interaction between soil type and water amount significantly impacts growth. Which statistical method should Akira employ for her analysis?

* linear regression analysis
* single-variable analysis of variance
* pearson correlation coefficient
* factorial analysis of variance

Correct

=============================

## #10 Predicting Outcomes

Question 1 of 8

During an analysis, Priya notices that the adjusted R squared value in her multiple regression model is 0.6181. What does this indicate about her model's ability to predict the outcome?

* The model fails to account for more than 60% of the variance in the outcome.
* It can account for approximately 62% of the variance in the outcome.

Correct

* It predicts the outcome with 38% accuracy.
* It can predict 61.81% of the outcomes exactly.

Incorrect

Question 2 of 8

Nia is preparing to standardize her dataset, which has variables on vastly different scales, for a regression analysis. The mean of her variables ranges from 0.52 to 8.32. Which function should Nia use to ensure all variables have a mean of zero and a standard deviation of one?

* scale function

Correct

* lars function
* rename\_all function
* make.names function

Question 3 of 8

In an international study examining personality traits, Sarah, a researcher, aims to understand how well age, gender, and speaking English as a native language predict a person's openness to experience. After integrating her demographic data, she plans to incrementally add other personality traits to the analysis. To determine the unique contribution of each added block of variables, what statistical approach should Sarah employ?

* blocked regression

Correct

* multivariate analysis of variance
* canonical correlation analysis
* principal component analysis

Question 4 of 8

Which statement correctly explains the purpose of the GLM function in logistic regression?

* The GLM function is used to calculate the correlation coefficient between the predictor and outcome variables.
* The GLM function is primarily designed to transform categorical variables into quantitative ones for analysis.
* The GLM function specifies the titles and labels for visualizations in logistic regression analysis.
* The GLM function is used for fitting the logistic regression model to predict a binary outcome based on predictors.

Correct

Question 5 of 8

You are working on a digital marketing campaign targeting users with high openness scores on the Big Five personality test. To initially filter potential targets, you decide to use basic demographic information. Which method should you use to assess the effectiveness of adding more detailed personality data later on?

* stepwise regression
* lasso regression

Incorrect

* blocked regression

Correct

* ridge regression

Question 6 of 8

What is the primary purpose of using quantile regression in data analysis?

* to reduce the influence of outliers on the regression analysis

Correct

* to allow for a higher number of variables in the model
* to increase the speed of the regression analysis
* to simplify the steps required to interpret the regression model

Question 7 of 8

In Utah, the Google search popularity for scrapbook and modern dance showed a unique relationship. If you were analyzing this data, how would quantile regression benefit your analysis?

* by increasing the influence of the outlier on the regression line
* by providing a more accurate representation of the relationship by reducing the effect of the bivariate outlier

Correct

* by making the outlier the focal point of the analysis
* by ensuring that the outlier is removed from the dataset entirely

Question 8 of 8

You're analyzing a dataset with multiple variables in R, aiming to predict a single outcome. Which type of regression model allows the inclusion of several predictor variables for this purpose?

* logistic regression

Incorrect

* hierarchical regression
* multiple regression

Correct

=============================

## #11 Clustering and Classifying Cases

Need refresher on (z-score, stdv)

Question 1 of 10

You are planning a targeted marketing campaign and decide to use hierarchical clustering to identify distinct customer groups based on their interest patterns. How should you proceed after identifying the optimal number of clusters for your campaign?

* Combine all clusters into a single group for a uniform marketing strategy.
* Increase the number of clusters to cover a wider range of customer interests.
* Focus solely on the largest cluster and ignore the smaller ones.
* Develop different marketing strategies for each identified cluster.

Correct

Question 2 of 10

Why is it important to scale variables before performing k-means clustering on a dataset?

* to uniformly distribute all data points across the clusters
* to increase the computational speed of the k-means algorithm
* to prevent variables with larger scales from dominating the algorithm

Correct

* to convert all categorical variables to numerical

Question 3 of 10

In conducting hierarchical clustering analysis, why is it important to ensure that data are on similar scales before computing the clusters?

* Adding variables with vastly different scales can dominate the clustering algorithm.

Correct

* Clustering algorithms always require data to be normalized to a mean of zero.

Incorrect

* Data with similar scales are required to perform the scale function.
* Similar scales on data prevent the formation of too many clusters.

Incorrect

Question 4 of 10

In trying to classify gender based on personality variables, Ibrahim finds his model misclassifies 59% of males as female. What might be a reason for this high misclassification rate?

* Only one decision tree was used in the modeling process.
* The model was not trained with a random forest algorithm.
* The predictor variables have a weak association with the outcome.

Correct

* Too many predictor variables were used, causing overfitting.

Question 5 of 10

You are Jerusha, a data scientist, and you're tasked with choosing an ideal method for classifying emails as spam or not spam based on their content's similarity to known examples. Given your knowledge, which method is most suitable for this task?

* Support vector machines
* Linear regression
* K-nearest neighbors

Correct

* Decision trees

Question 6 of 10

Why might a researcher choose to partition their dataset into a training set and a test set?

* to focus exclusively on qualitative data for model building
* to increase the speed of the data processing by reducing the dataset size
* to avoid using graphs or visual representations in their analysis
* to build the model using the training set and validate its accuracy with the test set

Correct

Question 7 of 10

When choosing the number of clusters for k-means clustering, which method involves looking for a bend in the graph of within cluster sum of squares?

* the elbow method

Correct

* the dendrogram method
* the hierarchical clustering method
* the silhouette method

Question 8 of 10

Which feature of the big five personality factors was specifically targeted for classification into high or low scores in the presented methodology?

* Conscientiousness
* Openness to experience

Correct

* Extroversion
* Neuroticism

Question 9 of 10

You are analyzing personality data and want to predict an individual's gender based on their scores on the big five personality factors. Which of the following methods is most appropriate for creating a model that is easily interpretable and visual?

* decision trees

Correct

* neural networks
* support vector machines
* linear regression

Question 10 of 10

Which of the following best describes the purpose of using a random forest approach in data analysis?

* to improve prediction accuracy by averaging the results of multiple decision trees

Correct

=============================

# R for Excel Users

=============================

#1 Install R

=============================

## #2 Descriptive Statistics in Excel

Use of excel “Data Analysis” tool pack,

=============================

#3 Working Between R and Excel

library(DescTools)

Question 1 of 3

Which of these does R recognize as the name of a package?

* Desctools
* desctools
* DescTools

Correct

* Desc Tools

Question 3 of 3

Which function saves R data in the csv format?

* the write() function

Correct

=============================

#4 DescTools Univariate Analyses

Desc(d.pizza$temperature)

Desc(d.pizza$temperature, plotit = FALSE)

Fmt(abs=structure(list(big.mark=","),class="fmt"))

Question 1 of 1

Which character (slash, tilde, dollar sign or ampersand) is used to separate the name of a data frame from one of its variables?

* d.pizza $ driver

Correct

=============================

#5 Desc Output

Question 1 of 2

A variable in a data frame that can take on both numeric and text values is often termed a \_\_\_\_\_.

* Atomic
* Variant
* Factor

Correct

* Field

Incorrect

Question 2 of 2

Which type of chart shows the relationship between the IQR and any outliers that might exist?

* Treemap
* Sunburst
* Waterfall
* Box & Whisker

Correct

=============================

#6 Bivariate Analysis in DescTools

A blue and white box with text

AI-generated content may be incorrect.

Numeric ~ Numeric -> Correlation

A graph of a pizza delivery

AI-generated content may be incorrect.

Numeric ~ Factor/Tex -> Mean Breakdown

A diagram of a pizza delivery

AI-generated content may be incorrect.

Factor/Text ~ Factor/Text -> Contingency Table

inferential statistics

A diagram of a driver's area

AI-generated content may be incorrect.

A chart of a pizza driver

AI-generated content may be incorrect.

{  
Contingency tables from Desc

Selecting transcript lines in this section will navigate to timestamp in the video

- [Instructor] One more bivariate analysis to go, and so far, we've looked at a numeric-by-numeric analysis for which we get correlations, and we've looked at the analysis of a numeric variable by a factor from which we get the breakdowns of means, medians, and other statistics by the levels of that factor. The third and remaining type of bivariate analysis covered in this course examines one factor by another factor and returns a contingency table of the two factors. You get that using the Descfunction in the DescTools package. This lesson shows an example by breaking down the relationship between area and driver. Again, the data comes from the data frame named d.pizza and it's calling for the plotted argument to be true which means that we want it to show the charts. That is, by the way, the default for plotted. Notice that the main difference in the function syntax is that it calls for two nominal variables, area and driver, not two numeric variables and not a numeric variable and a nominal variable or factor. This is in text, which treats one factor as a contingent on another factor by means of the tilde gets you the contingency table. If we come up to the beginning of the output, then we get our inferential statistics from the analysis, a little bit of information regarding the number of cases and how many rows and columns are on the table. There's seven columns and three rows. Seven columns because there are seven drivers and three rows because there are three areas. The first set of inferential statistics includes the Pearson chi-square test, the likelihood ratio, and the Mantel-Haenszel chi-square. These are inferential statistics that tell us the likelihood of getting the degree of an association that we observe between the two variables assuming that as a hypothesis, there is no association in the population. If the likelihood is small enough to suit us, we reject the hypothesis of no association in the population. So under that hypothesis, we have an extremely small p value, at least as compared to the chi-square value based on 12 degrees of freedom. Such a small p value is highly significant in a statistical sense and it's telling us that there's a strong association between the two variables that define the contingency table. Finally, we get to the contingency table itself, the frequencies for each cell in the table and the marginals alongside, the marginals on the right and along the bottom, the percent of the rule that belongs to the cell and the percent of the column that belongs to the cell. Now, if you accept the default true value for the plotted argument, you can also get some further information that you may find productive. These charts will give you a sense of the association between the area delivered to and the driver who's making a delivery. Or in the terms of the right-hand chart, which area is being serviced most by which drivers. So in this case, we've got the Brent area being serviced mostly by Carter and secondarily by Hunter and the Westminster area being served primarily by Carpenter and secondarily by Miller. The chart on the right shows which of the three areas gets the most attention paid by each driver. So that Butcher is spending most of his time in the Brent area and Carpenter is spending most of his time in the Westminster area.  
}

**Question 2 of 3**

You have a data frame that contains data on voters, including their age and political affiliation. Which R function would you use to get the mean age of each party in your database?

* Breakdown()
* Desc()

Correct

**Question 3 of 3**

A contingency table is often used to quantify the strength of the relationship between two or more \_\_\_\_\_.

* ordinal variables
* nominal variables

Correct

* integer variables
* ratio variables

=============================

#7 Bivariate Analysisin Excel

Use of pivot table

#

# R for Data Science: Analysis and Visualization

=============================

#Intro

=============================

## #1 What is R

=============================

## #2 Getting Started

RStudio

Google Colab, colab.to/r

Jamovi.org

https://www.r-project.org/

https://cloud.r-project.org/

https://cloud.r-project.org/web/views/

=============================

## #3 Data Visualization

=============================

## #4 Data Wrangling

=============================

## #5 Data Analysis

=============================

#Conclusion