Farhoud Khoshnoud

[Company Name]

R

Table of Contents

[Complete Guide to R: Wrangling, Visualizing, and Modeling Data 2](#_Toc200365445)

[# Intro 2](#_Toc200365446)

[#1 What is R 2](#_Toc200365447)

[#2 Getting Started 3](#_Toc200365448)

[#3 Importing Data 3](#_Toc200365449)

[#4 Visualizing Data With ggplot 3](#_Toc200365450)

[#5 Wrangling Data 4](#_Toc200365451)

[#6 Recoding Data 8](#_Toc200365452)

[#7 An R for Data Science Case Study 11](#_Toc200365453)

[#8 Exploring Data 11](#_Toc200365454)

[#9 Analysing Data 15](#_Toc200365455)

[#10 Predicting Outcomes 18](#_Toc200365456)

[#11 Clustering and Classifying Cases 20](#_Toc200365457)

[R for Excel Users 24](#_Toc200365458)

[#2 Descriptive Statistics in Excel 24](#_Toc200365459)

[Question 2 of 3 29](#_Toc200365460)

[R for Data Science: Analysis and Visualization 31](#_Toc200365461)

[#1 What is R 31](#_Toc200365462)

[#2 Getting Started 31](#_Toc200365463)

[#3 Data Visualization 34](#_Toc200365464)

[Creating scatterplots 34](#_Toc200365465)

[Creating line charts 35](#_Toc200365466)

[Creating cluster charts 36](#_Toc200365467)

[Quiz 38](#_Toc200365468)

[#4 Data Wrangling 40](#_Toc200365469)

[Selecting cases and subgroups 40](#_Toc200365470)

[Recoding variables 41](#_Toc200365471)

[Computing new variables 42](#_Toc200365472)

[#5 Data Analysis 44](#_Toc200365473)

[Computing frequencies 44](#_Toc200365474)

[Computing descriptives 45](#_Toc200365475)

[Computing correlations 46](#_Toc200365476)

[Computing a linear regression 47](#_Toc200365477)

[Computing contingency tables 50](#_Toc200365478)

# 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

Descriptive Statistics, R and Excel, DescTools Univariate Analyses, Desc Output, Bivariate Analysis in DescTools, Bivariate Analysis in Excel

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

#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/

Question 1 of 14

In order to output the sequence of numbers below, what will you need to input?

30 27 24 21 18 15 12 9 6 3 0

* seq(30, 3, by = -2)
* seq(30:0)
* seq(30-0)
* seq(30, 0, by = -3)

**Correct**

Question 2 of 14

Where are the objects created in RStudio kept and displayed?

* in the Viewer
* in the Environment pane

**Correct**

* in the console
* in the Terminal

Question 3 of 14

The default structure in R is a(n) \_\_\_\_\_.

* matrix
* vector

**Correct**

* data frame
* array

**Incorrect**

Question 4 of 14

Which data types are double precision by default?

* logical variables
* character variables
* complex numbers
* numerical variables

**Correct**

Question 5 of 14

When installing RStudio for this course, which version should you download?

* RStudio Server
* RStudio Team
* RStudio Desktop

**Correct**

* RStudio Cloud

Question 6 of 14

If you want to interact with R and other languages in an entirely online format, which environment works best?

* the R application

**Incorrect**

* RStudio

**Incorrect**

* Microsoft Azure Notebooks

**Correct**

* the terminal

Question 7 of 14

To install R, which URL should you use?

* r-project.net
* r-project.com
* r-project.org

**Correct**

* r-project.edu

Question 8 of 14

When viewing datasets that come in the datasets package, you are given their name and a few words to describe them. If you want more information, what can you select?

* Index

**Correct**

* Files
* Help

**Incorrect**

* Packages

**Incorrect**

Question 9 of 14

How can the following code be rewritten using the piping function?

round(prop.table(margin.table(UCBAdmissions, 3)), 2) \* 100

* UCBAdmissions (3) %>% margin.table(2) %>% prop.table %>% round %>% multiply\_by(2)
* UCBAdmissions %>% margin.table(3) %>% prop.table %>% round(2) %>% multiply\_by(100)

**Correct**

* round %>% margin.table(2) %>% prop.table %>% UCBAdmissions(3) %>% multiply\_by(100)
* prop.table %>% margin.table(3) %>% UCBAdmissions %>% round(2) %>% multiply\_by(100)

Question 10 of 14

If you are using a .csv file, how can you easily import your data into R?

* the read\_csv() command

**Correct**

* the read\_csvfile() command
* the read\_excel() command
* the read.table() command

Question 11 of 14

When can you use a matrix?

* when you have variables of different types
* when you have data with different dimensions
* when you are working with unstructured text
* when you have data that is all of the same type

**Correct**

Question 12 of 14

How can you generate an array so it creates the following?

, , 1 [,1] [,2] [,3] [,4] [,5] [,6] [1,] 1 4 7 10 13 16 [2,] 2 5 8 11 14 17 [3,] 3 6 9 12 15 18

* a1 <- array(c( 1:18), c(3, 6, 1)) a1

**Correct**

* a1 <- array(c( 1-18), c(6, 3, 1)) a1
* a1 <- array(c(18), c(3, 6, 1) a1
* a1 <- array(c( 19-1), c(6, 3, 1)) a1

Question 13 of 14

If you want to sort packages into topics on the R project website, what should you select?

* CRAN Task Views

**Correct**

* CRAN Mirrors
* CRAN Search
* CRAN Repository Policy

Question 14 of 14

How can you indicate that something is a comment, and not a command that should be run?

* with the hashtag (#) symbol

**Correct**

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

## #3 Data Visualization

Colours https://datalab.cc/rcolors

### Creating scatterplots

- [Instructor] Scatter plots are one of the best ways of looking at associations between two quantitative or measured or scaled variables. To demonstrate this in R, let's load our libraries. And then from our data set, which is state trends, let's bring in just the five terms about searches for sports. I do that by reading in the CSV and then selecting basketball through hockey. That brings in five sports. And then we'll use a glimpse. It's a slightly different way of looking at the data from what we've used before. And when we do that, you can see here that we have read in. It tells us which variables are character, which ones are double, but then when we specify just the five variables from basketball through hockey, they're all double precision. And here are the initial values and that's great. That's what we're looking for. Now, to make a scatter plot, we have an option of simply plotting all of the associations. We only have five variables in our dataset right now, so that's not too many. If we had 20, it would be a nightmare. It'd take a long time to plot it and it would be very hard to read it. But five is manageable so I just do DF and I feed it into plot. And we can zoom in on this scatter plot matrix. So basketball is on the left and on the top, then football then baseball, then soccer, and then hockey is across the bottom and on the right. And it's symmetrical. And so this chart right here shows the association between football and basketball. And this one also shows the association, except in this one basketball is across the bottom and football is up the side. And in this one, football's across the bottom and basketball is up the side. So they're sort of mirror images of each other. From this, we can see a few things. Number one of basketball, football, and baseball, there are generally uphill trends. So it looks like states that search more for one of these sports also tend to search more for the other ones. Soccer, on the other hand, you can see it's got a little different trend. It's either no trend or possibly even downhill. And hockey has some major outliers which throw things off but it looks like maybe it's a downhill association except between soccer and hockey. And this gives you an idea of where you can look for things. Now, if you're doing marketing, that tells you that there are important similarities possibly in your audiences for basketball, football, and baseball. But that soccer and hockey operate a little bit differently. Let's take a closer look by doing a bivariate, which means two variables scatter plot, using just the defaults. So we'll take our data frame, we'll select just soccer and hockey since they seem to be operating a little differently than the others. And we'll make a plot for those two. And now what you can see is we have a couple of outliers which we happen to know from our previous one on box boss that's Minnesota and that's North Dakota. And there are the rest where it's basically uphill where there's a slight trend that states that search more for soccer, relatively speaking are also a little more likely to search for hockey but it's not really pronounced. On the other hand, we can come back here and we can do our bivariate scatter plot again with a few options. Mostly here we are adding a title, we're adding labels, we're going to color the points. So instead of being these empty circles, we'll make them red and we're going to change the plotting character set it to a small circle. Now this involves using PCH which is for plotting character and we're using character number 20, the small circle. But now, that looks good. If you want to know what the other characters are, by the way we can do help on PCH. And you do have a number of options. It'll show 'em when we get down here. Here is the list. And here's the small point number 20 that I'm using. I try to keep it pretty simple. I would really personally only use the medium or the large or the small circle, but that one's up to you. So let's go back to our plot. And we can also do one very common thing which is adding a regression line. So that is a fit line, a predictor line. And the formula for this is LM which stands for linear model, that's regression. And you tell it what's your outcome variable what's the one going up on the Y axis? And then use a tilde to say is a function of or as predicted by the other variable. So here I say use hockey from our data frame as a function or as predicted by soccer in our data frame 'cause soccer is on the X and hockey's on the Y. And then AB line which means regression line with your A and B coefficients. And there you go. It is a slight uphill and you can see it a little more clearly here. It doesn't get really big. But we have these states that just have very little interest in both of these sports up to ones that show a little more interest. And again, with our superstar outliers of Minnesota and North Dakota that we identified earlier. And so if you are looking at where you can invest your advertising dollar for a hockey or a soccer or whether there's any crossover between those markets this kind of analysis is going to give you a great start on planning your strategy and finding the people who are going to be most interested in what you have to offer them.

### Creating line charts

- [Instructor] I'm a firm believer in trying to get more done with less work, less effort, and also less code. And I tell people when they are starting to work with data that you're going to get 90% of your value from bar charts and from line charts because bar charts show you how common something is. And line charts show you change over time. And if you're in a business setting, that's especially interesting. Let's take a look at line charts in R by first loading the data sets. I'm going to be using some of the built-in data sets for this example. And we're going to take a look at a few of them. The first one is US population, so it's uspop. Let's get the help on that one. And as population in millions from 1790 to 1970, it's a time series of 19 values. And we can actually see the entire dataset by just calling uspop. And the frequency 0.1 lets us know that these are decades apart from each other. And so that's a population of growth. And fortunately it's really easy to plot a single time series element here. And it's important to point out that it is a time series. It tells R that this data set is time series, the year that it starts, the year that it ends and the frequency in between. It's a special kind of data set. But let's graph this one. When we use plot with a time series data set, it knows to make this line. And you can tell from this that the US population has been going up over time. Let's zoom in just for a second. In fact, it's really smooth exponential growth with a little glitch here and here. It slowed down and then it resumed what it was doing. And in fact, we can make the same plot with a few options. Let's give it a title and some labels and then I'm going to add some separate elements. So you have to first make the chart, then you can add these things. One is I'm going to add a vertical line at 1930 in light gray, and then I'm going to put some text at the bottom that says that's 1930. Then I'll do another vertical line at 1940 and add some text there. When we zoom in on that, you can tell this is a different population that corresponded to the depression and then picked up back after World War II. And those are reasonable times to mark as a way of an explanation for this little glitch in the exponential growth of the population. Now, when you're working with time series data, you do have other options in R. We did just the generic plot command and it worked fine. There is one that is specific to time series. It's called ts.plot, and it can be used for a single time series, but it's going to look exactly the same. Here we go. And the general plot is a little easier and I would use it instead. There is also a more powerful alternative than instead of being ts.plot, it is plot.ts. And that gives you some more options though for our simple data set, once again, the chart's going to look exactly the same. Things vary, however, when you're showing more than one time series at a time. So you want to look at the changes of several different things. In this case, I'm going to use a data set that is about the European Union stock markets and it gives us four different stock markets the DAX in Germany, the SMI in Switzerland the CAC in France, and the FTSE in the UK. So let's get a little bit of information about that data set. Again, it's a time series with 1,860 observations on four variables. And it is a object of class mts which is for multiple time series. And we can take a look at that data set by simply calling its name. And it's only showing us the first bits and pieces of it but let's see if we can get up to the top here. So we have the year and then we have these fractional years. That's how it's indicating the daily values in this particular data set. And then we have the four indexes with their values below 'em. But this is going to be a whole lot easier to make sense of if we make a graph. And so I'm going to use the ts.plot. I'm going to tell it that we're charting the EU stock markets and I'm going to tell it to make the lines in four different colors by using the first four colors of the rainbow palette that we looked at a while ago. And when we do that, you can see the exponential growth over time and that they all tend to show roughly the same percentage growth over time. I'm also going to add a legend and I do that with a separate legend function. I'm going to tell it to put it in the top left 'cause we have a big empty space there. And I'm going to tell it to use the column names from the EU stock markets. And then I'm going to give it the same four rainbow colors and tell it to use solid lines. And when I do that, that's quick and easy. Let's zoom in on it just for a minute. And so this might be a sufficient graph for your internal presentation. It shows the growth over time. You can see for instance that the purple FTSE from the UK drops down and the bright green SMI from Switzerland picks up around 1997. And that there might be something that is of value to you and the people that you work within that. But this is one great way of looking at data that goes over time. The time series line charts in R.

### Creating cluster charts

- [Instructor] A common task in working with data is finding more or less naturally occurring groups in your data of observations or cases or people who can be grouped together. Say, for instance, you're going to do marketing campaigns and you have enough money for three different campaigns, then you want to find three different groups of people in your audience who can each receive one of the campaigns. And the nice way to do this is with a visual diagram, a cluster chart, which is very easy to do in R. And we actually have a few different options. Now I'm going to start by opening this file, 03\_07\_clustercharts.R. And when I do that, we get this yellow warning across the top that says there are two packages that are required but not installed. You can either click Install right here and it'll take care of it. But I'm also going to just run through it with the code itself. Now we're going to be using a package that we haven't used before called ggdendro, which stands for grammar of graphics dendrograms. That's the kind of chart that you make for clustering. I'm going to do cmd or ctrl + return to install that package. And here I see it's downloaded. It's a small package, it's there now. And then to make that available, I have to load it using the library command. And I do the same thing for the tidyverse. So I'm going to do cmd or ctrl + return to load the tidyverse. And then the same thing for ggdendro. And now let's look at some data. We're going to use the state dataset, so we're going to use read\_csv to read it and we're going to put in the state code and a bunch of the search terms. We'll drop missing cases, even though there aren't any missing cases, but we'll save that to df. Great, and so that's now in there. Now one of the things you want to do when you're working with data is you want to make sure that things are on the same scale. Now we're already pretty good because these are Google Trends terms, and theoretically they go from zero up to 100. But sometimes you might have things that indicate population or some other preference, and you'll have variables that are on different scales that can really wreak havoc with a cluster analysis. And so one of the first things that we're going to do is we're going to scale the data set. We're going to standardize it and so that it takes each of the variables and rescales them so that they have a mean of zero and a standard deviation of one. Put them on comparable scales, which is important for getting a more accurate cluster analysis. We're going to select everything except state code will remove that and we'll just use the scale function and save it in a new object called df\_scaled. And so you see that's over here now. And you can see these numbers that are now Z scores. Again, they are a mean of zero and extended deviation of one for each of the variables. And then we're going to put the row names back on. So we have the state codes. Now let's come down here and let's do the cluster analysis itself. We're going to do this in a couple of steps. The first one is we're going to use h cluster for hierarchical clustering, which is an agglomerative method. So it starts with everything separate. All the cases are separate, and then it starts putting them together until the very last step. All the cases are in one group. This is as opposed to a method that splits them up. And so there are a couple of different methods, but we're going to use this one. And we use couple of functions. We start with our scaled data. That's the one where we make sure that things are on the same range of the data. And we're going to calculate distances that is a distance or what's called a dissimilarity matrix. How far is each point from the other points based on the data that we gave it. And then we use that to compute hierarchical clusters. And we're going to save that into an object called hc for hierarchical clusters. We do that and it's a list over here, but we can then plot it. So we take hc and we feed it into the command plot. And we're going to give it some labels and a title. And let's run that one. And now we get a graph. And let's zoom in on it for just a moment. This is what a hierarchical cluster looks like. We have the individual 48 states listed across the bottom. And then these are lines that join two of them at a time. Remember, 'cause we're agglomerating. We're simply adding two at a time. And as it goes through the data we see, for instance, it looks like Arizona and Florida get joined very first 'cause they're the very first step. On the other hand, we have other places like Indiana, Kentucky get joined a little bit later. And then we start joining each of those different groups. And so that is one way to look at your data and maybe you already see some sort of pattern in there that's going to make sense to you. It's going to be a little easier to interpret if we draw some boxes around it. So I'm going to use this function called rect.hclust. And I'm going to do it first by saying k = 2, that means I want two boxes and I'm going to draw them with a gray border. And when I do that cmd or ctrl + return, now we can zoom in on it and you can see it says, well, we've got this group of states right here. And then over here we've got all the other states. And it does it by splitting them where they are at these last two points, two might be helpful, but we can also do three and we can do it with different colors. I'm doing border as two through four for the colors. And when we do that, it just lays it over the top of it. And now you can see three different groups. So if you're doing a marketing campaign with three groups and you want to do states, this might be one way of divvying up your states. Now there is alternative visualization to this. If you're looking for something that's more presentation polished and that's using ggdendro. Again, 'cause this kind of chart here is called a dendrogram, which is like icicles. And we're going to use this one, the function ggdendrogram. And we're going to tell it to rotate it to horizontal and turning off the stock theme. But let's run that one and see what it looks like. And you see here, it looks a lot more like ggplot2 graphics as we go through. And it lists the states here, it's a little squished. You'd probably want to make it taller, but it is the same organization of the data as you go through. And what both of these have in common is a way of finding naturally occurring similarities based on the data that you gave to the algorithm. And that's important to say. This is simply based on a dozen or two Google search terms. It does not include things like education and income and whatever else you might want to put in there. But based on the limited data that I provided it, it's grouping states in this particular way. And again, that might be appropriate depending on your purposes. And so a cluster diagram, a cluster chart can be one really good way of getting a quick initial impression of the groupings in your data that can help you meet your own particular analytical goals.

Standardize Quantetative Variables 🡪 mean of zero and standard-deviation of 1

### Quiz

**Question 1 of 7**

How can the following code be changed to reflect a histogram, with a title of "History of the Price of Diamonds", labeling the x-axis "Price of Diamonds" and the y-axis "Frequency"?

hist(diamonds$price)

* hist(diamonds$price, main = "History of the Price of Diamonds", ylab = "Frequency", xlab = "Price of Diamonds")

**Correct**

* hist(diamonds$price, title = "History of the Price of Diamonds", x-axis = "Frequency", y-axis = "Price of Diamonds")
* hist(diamonds$price, name = "History of the Price of Diamonds", y-axis = "Frequency", x-axis = "Price of Diamonds")
* hist(diamonds$price, main = "History of the Price of Diamonds", xlabel = "Frequency" ylabel = "Price of Diamonds", )

**Question 2 of 7**

You want to arrange bars in a chart from greatest to least in value. Which command would replace the BLANK in the code below to achieve this?

diamonds %>% select (clarity) %>% table() %>% BLANK Barplot()

* sort(increasing = T) %>%
* sort(greatest - least) %>%
* sort(greatest = T) %>%
* sort(decreasing = T) %>%

**Correct**

**Question 3 of 7**

Which two procedures are needed to calculate the distance and hierarchy of the clusters in your cluster chart created with the code below?

hc %>% plot(labels = df$state\_code)

* grp and dist
* dist and hclust

**Correct**

* library and clustplot
* kmeans and plot

**Question 4 of 7**

You make a barplot with the following code. Using color names, how can you color the barplot red?

`x = c(24, 13, 7, 5, 3, 2) barplot(x)

* barplot(x, col = rgb(.80, 0, 0, max = 255))
* barplot(x, col = "red3")

**Correct**

* barplot(col, xl = "red3")
* barplot(x, col = rgb(.80, 0, 0))

**Incorrect**

**Question 5 of 7**

Using the pipes process from tidyverse, which code correctly shows a boxplot that compares price and color?

* diamonds %>% select(color, price) %>% plot()

**Correct**

* diamonds %>% select(price, color) %>% plot() %>%
* diamonds %>% select(color) %>% select(price) %>> plot()
* diamonds %>% select(price) %>% select(color) %>% boxplot() %>%

**Question 6 of 7**

How can the following code be changed to reflect a line chart, with a title of "US Population 1790 - 1970", labeling the x-axis "Price of Diamonds" and the y-axis "Frequency"?

plot(uspop)

* uspop- plot( main = "US Population 1790-1970", xlab = "Year", ylab = "Population (in millions)" )
* lc(uspop)%>% plot() main = "US Population 1790-1970", xlab = "Year", ylab = "Population (in millions)"
* uspop%>% plot( main = "US Population 1790-1970", xlab = "Year", ylab = "Population (in millions)" )

**Correct**

* lc(uspop, plot( main = "US Population 1790-1970", xlab = "Year", ylab = "Population (in millions)" )

**Question 7 of 7**

If you want to include quotes in your labels when creating a scatterplot, which character must you include?

* the ampersand or at (@) character
* the star or asterisk (\*) character
* the caret or control (^) character
* the escape or backslash (\) character

**Correct**

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

## #4 Data Wrangling

### Selecting cases and subgroups

- [Instructor] Sometimes you don't need to see everything all at once and having a little bit of focus can give you better, more useful insights in your data. One way to do this in R is through selecting cases. Actually, we're going to be using something called filtering to do the selection, but the idea is you can select individual cases or subgroups and give them your special attention. Let's start by loading our two packages and then we're going to load our data set and I'm going to convert a few of the variables to factors in that process. And so we've got our data set right here. I've kept state, the region, the geographical region, the psych\_region, and their scores on one Google Trend search, and that is data\_analysis. So this is the relative popularity of data\_analysis as a search term for that state compared to other states. So let's start by filtering out some of these cases. We'll first look for the state where data\_analysis has a score on Google Trends of over 50. So I do filter and then I say data\_analysis greater than 50. And then I'm actually going to sort the cases, sort the states, so that the ones with the highest scores are at the top and lower scores are at the bottom. And when I do that, we have only four states, in this particular variable, with scores greater than 50. By the way, the way Google Trends work is that the highest area will have a score of a hundred. The reason we don't have a hundred here is because Washington DC was the highest at a hundred and then Maryland was substantially lower than that. But because we don't have the personality data for Washington DC, or Alaska, or Hawaii, they're not included in this particular data set. Next, we can look, for instance, at just the cases where psych\_region is Relaxed and Creative. So the way I do that is, I say filter and then psych\_region, and when you're searching for a variable that has text as its data, you do need to do this one thing, use two equal signs. That is a common thing in computer programming languages. And it means equivalent to, as opposed to when you're assigning a value to something. Also, it needs to be in quotes. And I'm going to do that and then I'm going to sort it again in descending values by data\_analysis. And when I do that, we have 10 states this time, that are Relaxed and Creative, according to this personality analysis. And the one with the highest scoring data\_analysis is Virginia and it goes down to Nevada at 27. Now, you can filter by more than one variable at a time. So for instance the vertical pipe character, which is over the return key, is used as an OR. And so, this time I say take the data frame and then search for either region is equal to south, and then the pipe here means OR, psych\_region is equal to, or is equivalent to, Relaxed and Creative. And let's take a look at those and I'm going to sort them by both region and then psych\_region. When I do that, you can see our states that are in the south, here at the top, and then, in addition to the two that are Relaxed and Creative, we have additional states that are Relaxed and Creative and we also have their values for data\_analysis. If you want to do a conjunction and combine two search terms, you can say, show me the states that are both in the south, so region is equivalent to South, and, all you need's the ampersand, psych\_region is equal to Relaxed and Creative. And when I do that, we have just two states that meet both of those search requirements. It's North Carolina and Virginia. And then, you can also do an exclusive criterion, where you use the exclamation point, which is equal to NOT. So what this one says is, show me the states that are in the South region, is equal to or equivalent to South, and NOT psych\_region is equal to Relaxed and Creative. So show me the other states in the South. And then, we're going to arrange those by psych\_region. And then, we'll do descending values on data\_analysis. Now, when you look at these, all of them are in the South, 'cause that was one of the criteria I used. And then, they are either Friendly and Conventional, or Temperamental and Uninhibited. Not Relaxed and Creative, 'cause I specifically excluded that. And then we sorted it by psych\_region alphabetically. And then within that, data\_analysis in descending values. And so this is a good way to start focusing in on the particular cases and groups that'll be most interesting and informative to the projects that you're doing. It's a very simple process to select cases using the command filter and then combining things with the OR pipe, the AND ampersand, or the NOT exclamation point.

### Recoding variables

- [Instructor] The data rule of thumb is that 80% of any project time is spent preparing the data, and one of the major tests that can be included in that is recoding the data, so taking it as existing categories or scores and creating new categories or scores based on that, and I want to show you a few ways of doing this very briefly in R. We're going to do this by opening up 04\_02\_Recoding.R, and we're going to be using two packages. Number one is the general purpose tidyverse, 'cause that makes a lot of other functions possible, so we're going to load that one. We've installed it before. The other one is readxl. Now, interestingly, this is part of the tidyverse, but it's not loaded by default. You have to load it explicitly, and so I'm going to load that one as well, and it's an improved way of reading CSV and Excel files. So the first thing we're going to do is we're going to read our dataset in. Now, we're going to do a couple of things here. We're going to use the read CSV because our data's in a CSV format. We're specifying the name of it, and then one of the things we're going to do right now is we're going to convert all of the character variables to as factor. We're going to switch them to factors, because that changes the way some of the other functions, including graphs, work, but let's come right here and run that command, and now we've saved it over here into DF for data frame, and if you want to see it, you can click on it right here and see where it looks like a spreadsheet, but we also got a short output of it down here in the console when we ran the command. Now, the first thing I want to show you is how to combine some of the categories, so let's take this variable right here, psych region, where things are relaxed and creative, friendly and conventional, or temperamental and uninhibited. What I'm going to do is I'm going to take the data frame, DF, and then I'm going to use the command mutate, which is what we use for changing the data, and I'm going to say, we're going to create a new variable called relaxed, and we'll use the recode command as a way of doing this, and the way we're going to recode is we're going to start with the variable psych region, and then if psych region is relaxed and creative, and you have to put the spelling and the spacing and the capitalization exactly as it appears in the dataset and put it in quotation marks, then I tell it to make that equal to yes on our new variable, relaxed. And then I'm doing something I don't really need to do. I'm just showing you that you can specify explicitly what the no would be. In this case, I'm saying friendly and conventional is equal to no, because it's not the relaxed and creative, or you can just do this, and you can say .default, so everything that you didn't automatically already specify becomes no, so it would work the same. We could remove this line, and friendly and conventional would still go to no because of this default, but I just want you to see that there are different ways of doing it. And then we're going to select just the state code, the psych region, and this new variable we created called relaxed, and so I'm going to run that command by doing command or control return, and then let's zoom up and look at this one a little better. Here, you have the state codes, and here, you have the psych region and whether they are relaxed or not, and you can see that it's taking the different states, and if they're relaxed and creative, it says yes. If they're something else, it says no, and so that's a very basic kind of recoding. Now, if you want to get a little fancier, if you've got a lot of variables and you're creating these nuanced conditional things, you can use something called case when, and I'm going to get help on that one, so we'll do question mark case when, and you can see it is the general vectorized if else, and it's got a lot of information you can look up here. But what we're going to do is we're going to come over here to DF, start with our data frame again, and we're going to use mutate because we're changing things, and in this case, I'm going to create a new variable called like arts, and instead of recode, I'm going to use case\_when, and I'm going to say that a state likes art if their score on art in Google Trends is greater than 75, or if their score on dance is over 75, or, remember, the vertical pipe means or, if their score on museum is over 75, then I use the tilde to say, that means they get yes. And then it says true is no for all other values, so this is a very different kind of syntax than we had for specifying the values and defaults, but this is how it works. And then I'm going to select those variables, arrange them by like art, and print the entire thing, so let's run that particular command, and then I'm going to zoom in on this again, and you can see we have all 48 states, and the states that like arts, because this is over 75 or this or this, is going to get them at the top. It's not scoring them or averaging them, it's just saying, really, it's doing it alphabetically, but these are the yeses, and these are the ones who did not go above 75 on any of those variables. And so this is a way of looking at possible combinations of variables as a way of creating sort of an index or a customized grouping that can best meet your needs, but these two examples will give you an idea of what is possible when recoding your data in R.

### Computing new variables

- [Instructor] So, I should probably be a little more explicit about some of the things that you've already seen me do, and I may have explained a little bit, and that's about creating new variables in R, and this is actually a really simple process when you use the Tidyverse approach. Let me show you how this works and do some common tasks in working with data of the kind that I'm accustomed to. So, I'm going to just load the Tidyverse, and what I'm going to do then is I'm going to create a little toy data set with three variables. I'm going to save it as a tibble, that's a kind of data frame that is used in the Tidyverse, it makes certain things easier, and I'm going to have X be the values 1 through 5, I'm going to have Y be the value is 7 going down to 3. And then for Z, third one, I'm using the C for concatenate combined, and then I just typed in a few different values, a 2, a 4, a 3, a 7, and then an NA, which is a missing value and not available, and then I'm going to save that to DF, and I'm also going to print it to the console. So, here is my data set down here, you see X, Y, and Z as we go across, and you see the NA in the bottom right, and then we're going to average across a few variables. So, one thing that's really common in my work, which is in the social sciences, is that you'll give somebody a questionnaire, and then you need to summarize across those different variables. Well, the easiest way to do that in R is with rowMeans, and so I'm going to come here to my data frame DF, and then I'm going to use mutate, which means I'm changing something in the data. and then I'm going to create three new variables. I'm calling them mean\_xy, xyz, and xz, and I'm using the function rowMeans, and then I also have to add this thing that says across, 'cause it means I'm going across the columns to make this happen, but it's pretty easy, aside from that. RowMeans across, and this says X to Y. This one says X to Z, which means X, Y, and Z, and this one I'm specifying just X and Z, and when we do that, you can see how it works down here. So, it's averaging them. One trick, however, is that if you have an NA, a not available, because it's missing data value in there, it just kind of shuts down the process, and R won't compute anything that has an NA in it. Now, that may be appropriate in certain situations. On the other hand, I've had lots of situations where you give somebody a questionnaire with 20 questions and they don't answer one or two of them, but you can still get close enough to what you need by averaging the rest of what's available, and so there's a very simple way of telling R to do that. And as you add the argument, "na.rm = T", and what that stands for is not available or missing data, rm means remove, and then the T is for true. So, it says, please remove any missing values but otherwise do things the same, and when we do that, let me make it so you can see both of these together. You can see we still have the NA and the original data, but it has gone ahead and created these mean values, just ignoring that one particular value, and again, depending on your circumstances this may be a very useful approach when you have nearly complete data, you have enough to make valid conclusions, and this allows you to get through that process. Another thing that comes up sometimes is having to reverse code a variable. So, for instance, maybe sometimes you have it so 7 means more of something on a 1 to 7 scale, but sometimes 1 means more of something. So, to do that, we're going to again use mutate, and I'm going to create a new variable called y\_r, that means Y reversed. By the way, it's easy to simply overwrite Y, but I know from experience that it's easy to forget whether you have done that or not, and so I always create a new variable, and the \_r indicates that I have reverse coded, and because I set up my data as though it were from a 1 to 7 rating scale, like strongly disagree to strongly agree, well, when you have a 1 to 7 what you do is you take your value and you subtract it from eight, one more, and that will create a reverse variable. So, this will create my new reversed score on Y, and then I'm going to use select to tell it that I want to see the values for X and then y\_r, and then Z. So I'm actually replacing Y with y\_r. Y is still in the dataset, I'm just not going to see it at this particular moment, and then I'm going to tell it to go ahead and calculate the means, except this time using y\_r as we go through, and here you see we have the reversed values, and now, in fact, they're identical to X, and it's able to go ahead and calculate the means. This can be an extremely useful approach, a very common one in this social science research that I do, and the people I've collaborated with. Just as a few notes, if you're using a 1 to 5, or a 1 to 7, or 1 to 10 scale, you start with one more than your high number and you subtract your value. So, if it's a 1 to 7 use 8 - X. If it's a 1 to 10, use 11 - X. For a 0 to 5 scale, or a 0 to 10, or whatever, simply take the high value and subtract it. So, for 0 to 5 to 5 - X, for 0 to 10, do 10 - X, and if you're using a scale that goes from negative to positive, all you need to do is multiply it times negative one to reverse it. Now, if you're going to do a lot of this, and if you're doing psychometrics work, then there is a package called Psych, which can make this this whole process much easier, much faster and give you a lot of additional insight into what's going on, say for instance, with your scale scores, and if you open this up, it'll get you that link in your browser, and then you can see what you can do extra with Psych as a way of getting the aggregate value out of your data, and so any one of these approaches should be a great start on you finding the meaning as you go across your data to help you in your project.

**Question 1 of 3**

Which function adds new variables and preserves existing ones?

* switch()
* transmute()
* mutate()

**Correct**

* rescale()

**Question 2 of 3**

How can you filter by multiple variables using piping to return data where "regions" are equal to "South" or "psychRegions" are equal to "Relaxed and Creative"?

* df %>% filter(region = "South" | psychRegions = "Relaxed and Creative") %>% print ()
* df %>% filter(region == "South" & psychRegions == "Relaxed and Creative") %>% print ()
* df %>% filter(region == "South" & !psychRegions == "Relaxed and Creative") %>% print ()
* df %>% filter(region == "South" | psychRegions == "Relaxed and Creative") %>% print ()

**Correct**

**Question 3 of 3**

Which code takes the variable "psychRegions" and recodes it so "Relaxed" displays a value of yes if "psychRegions" says "Relaxed and Creative"?

* df %>% mutate(psychRegions = recode(Relaxed, "Relaxed and Creative" = "yes", "Friendly and Conventional" = "no", .default = "no")) %>% select(state\_code, psychRegions, relaxed)
* df %>% mutate(Relaxed = recode(psychRegions, "Relaxed and Creative" = "Relaxed" "Friendly and Conventional" = "no", .default = "no")) %>% select(state\_code, psychRegions, relaxed)

**Incorrect**

* df %>% mutate(Relaxed = recode(psychRegions, "Relaxed and Creative" = "yes", "Friendly and Conventional" = "no", .default = "no")) %>% select(state\_code, psychRegions, relaxed)

**Correct**

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

## #5 Data Analysis

### Computing frequencies

- [Instructor] The simplest analysis you can do is simply counting how often things happen in your dataset, and strangely, that can be really informative, and is the basis for a lot of very sophisticated approaches, but let's start with just getting the simple frequencies. How common is each score or category in your dataset? We'll do this by opening the script 05\_01\_frequencies.r. And then the first step is to load the tidy first, 'cause we're going to rely on some of its functions. And then we're going to load our dataset. We're going to read the CSV of state trends. We're going to select a few variables. We're going to mutate or change several of them to be factors. That's going to be important in a minute, using the as.factor command. And then we will print it and also save it to the environment. So we've saved it to the environment. And we also see the three variables that we've got right here. And so we can start looking at the frequencies. How often do each one of these things occur? Now if we come right here and we hit summary(df), so that's a built-in command summary of the data frame. Well, it actually works really nicely if you have factors. It tells us now region was not a factor, and it just simply tells us there were 48 scores. But these other ones, which were defined as factors, it works wonderfully. But let's come down here and let's try a few other options for a categorical variable. Let's do different just region using select and summary. And again, all it does is it tells us is that we're 48 observations in there. That's not very useful. Table, so you select the region. Instead of using summary, you use table. We can do that one, and look, it sets it out in a really clean, easy way, 12 missed Western states, nine in the northeast, 16 in the south, 11 in the west. Now we can also go to psych region, which is a factor, and we can use the summary here. And when we do that, you see it lays it out differently, it organizes it vertically, and then it's got the colon and then the values here. You can use either one of these, whatever works better. And then we can also try it with table instead of summary. Now I think this one is a little harder to read. I wouldn't want to do it. We can convert region to a factor and then we can see what that looks like. Now looking at the data itself, it doesn't look very different. You got the region over here, but now we can summarize all of the factors in the data frame all at once, summary, and then df. And look, it's a beautiful, concise summary of what's in there, and it's a great way to get started in terms of understanding your data and guiding you in further analysis for insight.

### Computing descriptives

- [Instructor] For your quantitative variables, you're going to want to start with basic descriptive statistics. That includes things like the mean, the standard deviation, and maybe the quartiles. These, of course, are easy to do in R, which is a language designed for working with data. I'm going to load a few packages, and then I'm going to load a data set, and I am taking several of the variables, region, psych\_region, psy\_reg, and then several of the sports variables at the end, and I'm going to convert them all to factors. I'll save that as df and then print it here in the bottom. So, here you can see that we have our data set. It's giving us the first 10 lines. All right, and so that looks good. Now what we're going to do is we're just going to do summary. Now, I did this for frequencies also, and the nice thing is it's actually a general purpose when you include quantitative or numerical variables as well. Here I'm using a pipe to send DF into summary. I could also do summary and then in parentheses, df. It accomplishes the same thing. But let's come up and see what we have here. Now we've got a lot of variables. Our first one, state, is a character variable. So it just tells us that there are 48 observations. That's what the length means. And the same thing for state code. But for population, which is a numerical variable, it gives us actual numbers: the minimum, the first quartile, and the median, the mean, that's the average, and the third quartile and the maximum. It gives us that for every one of the numerical variables. And for the factors like region and psy\_region and psy\_reg, it gives us how many cases there are for each of those. We scroll through the Google trends data, and we get those same summaries, the quartiles, and minimum/maximum, the mean. And then we have a series of yes no questions at the end, that because I coded them as factors, it simply tells us how many yeses and nos there are. So this actually is a great way to get a first look at the dataset. If you want to look at just one variable, then all you need to do here is select that one variable and ask for summary. And here we have the same thing, minimum, first quartile, median, mean, third quartile, and maximum. And that's a good starting point for your descriptives. Now, if you're familiar with box plots, you know that you can also get this five-number summary. It was developed by John Tukey, and we have a built-in function called fivenum, which is what he called the minimum, the lower hinge, median, upper hinge, and maximum, where the lower hinge and the upper hinge are usually identical to the first and third quartiles. The trick is it doesn't label them. So let's pick our one variable here. And we see we just have these numbers. That's the minimum, that's the maximum, and these are the quartiles in between. It makes a little more sense if you make a box plot first. So let's do that. And then we can ask for boxplot.stats for the same thing. And that's going to give us the summary numbers here that are used in creating the box plot. The important thing here is that these numbers right here correspond to, that's the 41, that's the minimum. This is the 53, it's the first quartile score. That's the 55, the median. This is the third quartile score, 62. And then this is the maximum, which is 73. We also have the number of observations. There's 48 altogether, and then we have the confidence intervals that correspond to the notches for the median. So it goes down to about 53 and up to about 57. And it turns out that there are no outliers in the data. So it doesn't have anything that it needs to report right here. But that's a quick and easy way of describing your numerical variables. If you want additional options, or if you want more control over what's going on, you can try using the psych package. Again, it's a free package that you can download and install which gives a huge additional range of descriptive statistics, and many other ways that you can look at your data.

### Computing correlations

- [Instructor] Some of the most exciting work in looking at data comes in finding associations, these patterns, differences between groups or ways that one variable can be used to predict another. That starts to give you some really actionable insights. And probably one of the easiest ways to do this is with the correlation coefficient. Now I'm going to show you some basic correlations using the quantitative variables in our dataset. So I'm taking some of the Google Trends data, and that is again, the state by state relative popularity of some search terms. And I'm using several that all have to do with data. Specifically, I'm using data science, artificial intelligence, machine learning, data analysis, business intelligence, spreadsheet, and statistics. And what I'm going to do right here is I'm using the select command which normally I use to just choose the variables. But select is really handy because it does a couple of other things. One is it lets you reorder the variables. Now I'm keeping them in the order that they came in, but if you put the variables in in a different order and select, it will simply create 'em in that order. But even more, it allows you to rename them. And I'm doing that because we're going to be making some matrices and really long names like artificial intelligence get hard to read. And so by going to the abbreviated versions, a lot of which are very familiar, DS for data science, AI for artificial intelligence, and so on. I'm using SS for spreadsheet and stats for statistics. But when we do that, you'll see that it makes for a very compact, concise data set and even more important, when we do some graphics or plots. So let's start off by making a scatter plot matrix. Let's take the data frame I just created and we'll just give it to the generic plot command. And because I have all quantitative variables, it knows to make a scatter plot matrix. So let's run that one and zoom in on it. And so what we have here is a scatter plot matrix and we have the name of each variable going down the diagonal from top left to bottom right. And this first column is data science as a search term going across the bottom, and the first row is data science as the search term going up the vertical Y axis. And these plots show us the combination of scores. So for instance, there's a couple of things that we can see. Number one is data science, artificial intelligence and machine learning, and in fact, even data analysis have very strong linear trends. It's almost a perfect diagonal as you go up. And what that means is states that search a lot for one of these things, relatively speaking, search a lot for all of them. And they seem to occupy the same behavioral or cognitive elements. So data science, AI, ML, machine learning and data analysis, very similar to each other. Business intelligence works a little bit differently. You see things spreading out here and the patterns are not so clear and spreadsheet operates almost completely independently. On the other hand, there appears to be some relationship between business intelligence and spreadsheet. And then stats or statistics can mean a lot of different things 'cause remember, as a general search term, it doesn't mean that they're wanting to know about statistical analysis. They might be looking for statistics about their favorite sports team. And so it can be a very different kind of search. And the scatter plots here give us an indication of that. But let's go back to our commands here and let's get a correlation matrix. Let's get numbers that index the strength of the linear relationship in each of those squares. So I'm going to do that by taking our data frame and feeding it into the cor for correlation command. And when I do that, because I gave things a short name, it all fits down here very nicely. These are correlations, but it's going to be even easier to read if we round it off to two decimal places. So I do the same command or I take our data frame, I feed it into cor, and then I use round two to two decimal places. And when we do that, we can see the actual numbers that go with each of these. So down the diagonal, we have the ones thats each variable correlated with itself. But you can see here, for instance, AI and DS, so thats artificial intelligence and data science have a correlation of 0.91 with each other. Data science and machine learning have a correlation of 0.7. That's nearly a perfect correlation. A positive one would be perfect. These are very, very strong. On the other hand, you can see that the association between statistics and business intelligence, well that's just a 0.31. By the way, this is symmetrical so this is the same 0.31 right here. That is a much weaker relationship. It might still be enough to do something useful with but it's almost nothing compared to these identity relationships between data science, artificial intelligence and machine learning. But let's take a look at something else we can do and that is if you want to look closer at a single correlation, you can use the cor.test function. You have to specify just the two variables. And here, I'm going to say take the variable DS from our data frame and take the variable DA, data analysis from our data frame and let's look at just those two. It's going to give us the correlation coefficient hypothesis test result with a P value and a confidence interval. And I'll open this up a little bit. So it's doing the standard Pearson's product-moment correlation coefficient, and we have a T-test here that gives us a P value that's very, very small, which says that the true correlation is not equal to zero. So we're rejecting the null hypothesis, and in fact, we get a confidence interval that says the 95% confidence interval, it ranges from 0.80 to 0.93, so it's very far from zero. It's a strong positive correlation and it's all the more interesting considering it's based on only 48 data points. And then finally, we have the actual correlation coefficient itself which is 0.88. Again, a strong positive association. Now there's obviously a lot more you can do with correlations. One thing that you probably would want to do is get a correlation matrix with the probability values for every correlation in that matrix. The standard core function in R doesn't do that but there are packages you can download such as Hmisc, or Rstatics that are able to do those along with additional functions for analyzing correlations. If you're interested, take a look at each one of them. You'll find a wide range of additional functionality and power to get insight into the relationships in your data.

### Computing a linear regression

- [Instructor] When working with data, one of the most interesting things you can do is start building models, that is statistical models, where you're trying to use a group of variables to predict a particular outcome. And the simplest and most flexible and often the most powerful way of doing this is with linear regression, and this is a very simple thing to do in R. Let's start by loading a couple of packages, and then let's read our dataset. Now what I'm going to do here is I'm going to pick the personality variables on a state-by-state basis, and all of the Google Trends variables. And when I do that, you can see that they've all loaded here, we've got a pretty big dataset, it's got 20 variables at the moment. Let's make a scatter plot of a subset of this. Let's take just the search term data science, so how much that term pops up relatively speaking from state to state, along with the five personality variables on a state-by-state basis. So data science and extroversion through openness, and let's make a scatter plot of those. When we zoom in on that, here's our term "data science," which in this first column is going to cross the bottom on the X-axis, and on the top row is going up the side on the Y-axis. I'm interested in whether any of these personality variables can predict the relative interest in data science and I see a lot of not really strong patterns, but there is an interesting one here at the end. This is with openness on the X-axis across the bottom and data science up the side on the Y-axis. And so, there might be something that we can look at more closely here. So what I'm going to do is I'm going to make another scatter plot, but this time with just openness and data science. And it looks like there's a pattern here that it starts in the bottom left and it goes to the upper right. Now, there is an issue here of what's called heteroscedasticity, where things fan out as it goes out, but in any case, we can tell that the really high values for data science are associated with states that have relatively higher levels of openness. One of the nice things in a scatter plot is to add a regression line, or a straight line that can be used to predict the association between them. What you do in R is you first create the scatter plot and then you add the line. We use the LM, which is for linear model function, and I'm going to tell it that data science as predicted by openness. Now please note, this is the opposite of the order of when we made the plot. The plot does the X first and the Y second, but when you're making a regression, you do Y as predicted by X, so you have to flip them around and the tilde means as predicted by, and then we use AB line, which has to do with the coefficient for the regression line, so we simply run that command, and you can see there is a positive or uphill relationship between these two variables, openness and data science. Now if you want to actually get a numerical description of that association, then you can do the linear regression, and I take that same command, LM and then data science and then tilde, meaning as a function of openness. And I'm going to save it to an object called Fit1. By the way, if you've worked much with R, you'll notice it's very common to take models like regression models and save them as fit, because you are fitting a model to the data. So let's make one that we're going to call Fit1 and it's called Fit1 because I have more than one that I'm going to do. And then let's show the model. Now, when you show the model, it's a very basic thing. It shows you the command and it gives you the intercept and the slope of any variables you included. I only have two variables, the outcome and the predictor, so the intercept is -8, that's where this line would cross, and the slope is 0.8. So for each step it goes up in openness, it goes up 8/10 of a step on data science. So that's good. If you want more information, we can simply ask for a summary. So summary of Fit1, and I'll make this a little bigger here, it gives us the residuals, which are an important way of diagnosing what's going on with the model, and you'll be able to tell that things will be come much more spread out as the values go up. We also get the coefficients, there's the intercept and there's the slope, along with the standard error, a T-test for each of them, and a probability. And so the slope is functionally equivalent to, excuse me, and so the intercept is functionally equivalent to zero, but the slope is highly significant, very different from zero. And then we also have some information about the standard error, the multiple R squared, which is about .3 in this situation, we adjusted R squared, we have a small dataset, only 48 observations, so not surprising that there's this big difference, as well as additional information about the significance of the overall model. That would be more important if we had many variables in the predictor. Now we can also get confidence intervals, so we use our model that we created, Fit1, but we run the conf int command on that, and when we do that, we get the 95% confidence intervals, the lower bound and the upper bound, for both the intercept and for the slope. You can also predict values for individual cases, and if you have more than 48, this is going to be an awful lot, so you may want to do that one graphically instead. And you can get prediction intervals for the values. And again, we only have 48 observations, so it's possible to scroll through the whole thing, but you have the predicted value along with the lower and upper confidence bounds for that particular prediction. And then in terms of regression diagnostics, you can run LM.influence, where you get a lot of information about the residuals and other information having to do, you can see it's a lot of output, the predicted values, and then influence.measures. And I'll just open this one up so we can see that one a little more. Again, a lot of different measures you can look at, and depending on the kind of model you're building, you may want to invest some time looking into what these individual measures mean, and how they influence the interpretation of your model. But let's move onto multiple regression, which is the more common thing, where you have more than one predictor. Here I'm going to create a new dataset, I'm going to redo DF, our data frame, where I do data science and then I do the five predictor variables, that is, the five personality variables. So, here's the first bit of our dataset. It's important to point out a lot of this is easier if you have your outcome variable, the thing you're trying to predict at the very front, that's your Y variable, and then everything that comes after that are the predictor variables. So that can be the capital X, it's a matrix of variables. And that makes it a lot easier to specify things in R. So here what I'm doing is I am going to not just select the variables, but I'm going to put them in order, so data science first and then the personality variables. By the way, another way to do this is you can also do select and then you can have your outcome variable, and then if you want to keep everything else in order, you just use everything open and closed parentheses. Now let's come down to specifying a model. There are three different ways that we can do this. The easiest by far is to just run the command LM, and as long as you have your predictor variable listed first, all of your outcomes after that, and nothing else in the dataset, just runs really easily. And here we have a multiple regression and we have the slopes associated with each one of these predictor variables. Another way to do this is to specifically label what the outcome is. Here I'm going to say data science, and then use the tilde as predicted by, dot means everything else, and then I tell it that the dataset is DF, that's the data frame that I've saved in memory. We do that, we'll get the exact same output, and then if you want to write a lot, you can put out every single thing, data science, tilde as predicted by extroversion plus agreeableness, plus conscientiousness, plus neuroticism, plus openness, and then it all comes from the data frame, right? Then it all comes from the dataset DF. We run that, you'll see we get again the exact same output, and you can save the model, here I'm saving it to Fit2, then we can look at the model by simply calling the name, and then here we have the coefficients again, but the summarized regression model's going to be the thing that's probably most interesting to people, and let's zoom in on that one. And here we have the residuals and we also have the coefficients, their standard errors, their T-tests, and the probability values along with the asterisks. And so we see from this one that conscientiousness is statistically significantly correlated within the context of this multiple regression, with the relative popularity of data science as a search term, but it's negative. The higher a state is in conscientiousness, the less interested in data science. Again, this is a multiple correlation, so if we would look at them individually, we might get different things, but within the context of these five predictors, it's got a negative coefficient, and then openness has a highly significant positive association, the more open a state is, based on the one research project, the more likely they are to search for data science as a term. And we have a multiple R squared of .46, so about 46% of the variance in data science is explained by the model, but we have a small sample size, so it's important to let the adjusted R squared, so it's about 40% total. Anyhow, we could give several entire courses on conducting linear regression in general and even just in R, but this is a great way to get started in building models to try to better understand and even predict what is happening to the things of interest in your dataset.

### Computing contingency tables

- [Instructor] The final procedure for analyzing data that I want to show to you is contingency tables, where you're dealing with two nominal or categorical variables. And in the marketing world, this is extremely common, and fortunately, it's easy to do in R. So, let's come and load our two libraries, and let's come down to our data set. Now, I'm going to convert all the variables that I leave as factors, on the other hand, I'm only saving two, I'm saving region, that's the geographical variable, and then also psy\_reg, which is the psychological region. I'm saving just the two of those. And then I'm using this command to change everything, all both of my variables, to factors. And we'll take a quick look at it. So, let's come right here. And we have region and psy\_reg, we have south, friendly, west, creative, and so on. Great. Now let's make a contingency table. All you need to do is use the Table command and tell it the two variables that you want. I tell it's in the data frame and I want the variable called region and in our data frame, df and psy\_reg. And we're going to save it to ct for contingency table. And as you can see here, we have a table that has shown up. And let's take a quick look at it by just calling the name of it. And there it is arranged in rows and columns. So, we can see, for instance, that we have 10 states that are in the South that are friendly, we have one in the Midwest that is uninhibited, and so on. Now, you can also get the cell row and column percentages and we can do rounding to get two decimal places, multiply them times a hundred to make it like a percentage. So, I'm going to take the contingency table and I'm going to ask for a prop.table. That's how we get the proportions of the cases in a particular part of the table. And one means I want row percentages. Then I'm going to round those to two decimal places and multiply it times a hundred. So here we see that the row percentages, so this is going across, so 62% of the southern states are friendly, 12% are creative, and 25% are uninhibited. In the Midwest, 92% are friendly, 0% go into the creative category, and 8% go into the uninhibited category. If you want column percentages that sum up to a hundred going vertically, you just need to change this argument to 2. So that gives us column percentages. And here you can see that the creative, 20% are in the South, 80% in the West, and none in the Northeast or the Midwest. And then total percentages, you just leave out the argument, you don't say anything there. And you can see that of the 48 states in our data set, 23% of them are Midwest and friendly. You can also do an inferential test. In this case, we're going to do a chi-square test. And I have to start by saying, doing a 4 by 3 chi-square test with only 48 observations and with a lot of empty cells is statistically irresponsible. You would need much better coverage to make this work properly, but the command will still function and you can adapt it to situations where your data better match the assumptions of the chi-square test. So, with that caveat in mind, all we need to do is do chi-square or chisq.test, and we apply that to the table. And I'm going to feed that into a new object called tchi, T-C-H-I. And when we call that one, it tells us that our value of chi-squared is 50.002 with 6 degrees of freedom. And the p-value that is very, very small. And if you want to do this in one step, you can simply say table of this variable and that variable and feed that into chi-square test. And see, we get the same information as well as our thing that the approximation may be incorrect, because we have not enough data going into this. You can also get additional tables that are used for diagnosing the table, help you with the interpretation. So for instance, you have the observed values, that's our original data. You have the expected values, because remember, chi-squared looks at the difference between the expected and the observed values. You have the residuals, which is the difference between the observed and the expected, and then you also have standardized residuals. Any one of those could give you a little additional insight into how your data are functioning. And so, that's going to be our final method for numerically analyzing the data in this short course on R. I hope that it's been enough to really whet your appetite to get you excited and interested in learning more about how you can use the free, open-source programming language, R, to get more insight and actionable insights out of your data.

**Question 1 of 5**

Using the table command, how can you create a contingency table specifying two variables in R?

* cor.test(df$variable, df$variable) ct
* fit1 <- lm(df$variable, df$variable) ct
* tchi <- chisq.test(df$variable, df$variable) ct
* ct <- table(df$variable, df$variable) ct

**Correct**

**Question 2 of 5**

Which function enables you to get the values that are used for locating each of the elements in your graph?

* the fivenum() function

**Incorrect**

* the summary() function

**Incorrect**

* the describe() function

**Incorrect**

* the boxplots.stats() function



**Replay**

Review this video

Computing descriptives

4m 9s

**Question 3 of 5**

If you have categorial variables or factors, which functions help you calculate the frequency of your variables

* the summary and cor() functions

**Incorrect**

* the fivenum() and describe() functions

**Incorrect**

* the rrcor() and boxplots() functions
* the summary() and table() functions

**Correct**

**Question 4 of 5**

How can you add a regression line that goes through your data and superimposes it on top of the scatter plot?

* lm(df$volunteering ~ df$museum) %>% abline()

**Correct**

* fit1 <- lm(df$volunteering ~df$museum) %>%
* lm.influence(fit1, interval = "volunteering, "museum") %>%
* summary("volunteering", "museum") %>% abline()

**Question 5 of 5**

Choosing two variables from a data frame, how can you look at the correlation of those variables?

* rrcor(dfvariable, dfvariable)
* cor(df(variable, variable))
* cor.test(df$variable, df$variable)

**Correct**

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

#Conclusion