**你先看 等等我会发第二版 //////下面是答案**

**Question 11 pts**

Below are five descriptions of variables. Which of the following are quantitative continuous variables? Select all that apply.

Group of answer choices

The recorded weight gain of rats which are eating either grains, seeds or fruit.

The temperature in Adelaide at midday everyday, recorded for 3 months.

The brands of chocolate available at the local supermarket.

The number of people attending a seminar on space travel.

The annual income of an employee recorded as: 1 = less than $30,000, 2 = greater than $30,000 but less than $50,000, 3 = greater than $50,000 but less than $100,000, 4 = greater than $100,000

///////////////////////

To identify quantitative continuous variables, we need to look for variables that can take on an infinite number of values within a given range. These values are typically measurements or counts.

Let's evaluate each description:

The recorded weight gain of rats which are eating either grains, seeds, or fruit.

This is a measurement and can take on a variety of values within a range. It is quantitative and continuous.

The temperature in Adelaide at midday every day, recorded for 3 months.

Temperature can take on a variety of values within a given range. It is quantitative and continuous.

The brands of chocolate available at the local supermarket.

This is a categorical variable as it represents different categories or types of chocolate brands. It is not continuous.

The number of people attending a seminar on space travel.

While this is a quantitative variable, it's discrete because the number of people can only be whole numbers (e.g., 1, 2, 3, etc.), not a continuous range of values.

The annual income of an employee recorded as: 1 = less than $30,000, 2 = greater than $30,000 but less than $50,000, 3 = greater than $50,000 but less than $100,000, 4 = greater than $100,000

This variable is ordinal and represents different income categories. It is quantitative but not continuous as it groups incomes into specific brackets.

From the descriptions provided, the following are quantitative continuous variables:

The recorded weight gain of rats which are eating either grains, seeds, or fruit.

The temperature in Adelaide at midday every day, recorded for 3 months.

[Flag question: Question 2](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 21 pts**

Below are five descriptions of variables. Which of these variables are categorical variables?  Select all that apply.

Group of answer choices

Species of bird in a tropical rainforest.

Average waiting time at a set of traffic lights.

Seriousness of injury of arrivals at an emergency department, where 0 = no injury, 1 = minor, 2 = major, 3 = fatal.

The number of arrivals at the emergency department of a hospital.

Country of birth for players in the English national cricket team.

 ///////////////////////////////

Categorical variables take on a limited, fixed number of possible values and represent categories or labels.

Let's evaluate each description:

Species of bird in a tropical rainforest.

This variable represents different categories of bird species. It is categorical.

Average waiting time at a set of traffic lights.

This is a quantitative variable, representing a measurement of time. It is not categorical.

Seriousness of injury of arrivals at an emergency department, where 0 = no injury, 1 = minor, 2 = major, 3 = fatal.

This represents categories of injury severity. It is categorical, albeit ordinal (it has an inherent order).

The number of arrivals at the emergency department of a hospital.

This is a quantitative discrete variable, as it represents a count of arrivals. It is not categorical.

Country of birth for players in the English national cricket team.

This represents different categories of countries. It is categorical.

From the descriptions provided, the following are categorical variables:

Species of bird in a tropical rainforest.

Seriousness of injury of arrivals at an emergency department, where 0 = no injury, 1 = minor, 2 = major, 3 = fatal.

Country of birth for players in the English national cricket team.

[Flag question: Question 3](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 31 pts**

Match the following R function to its description. There is only one possible option for each function.

Group of answer choices

str\_detect()

           [ Choose ]             split each type of variable into a separate plot             identifies the possible entries within a column, with duplicate entries removed             examine the numerical variables in a data frame             examine the missing data in a data frame             find the presence or absence of a pattern in a sentence             add a new variable to a data frame whilst preserving existing variables

unique()

           [ Choose ]             split each type of variable into a separate plot             identifies the possible entries within a column, with duplicate entries removed             examine the numerical variables in a data frame             examine the missing data in a data frame             find the presence or absence of a pattern in a sentence             add a new variable to a data frame whilst preserving existing variables

facet\_wrap()

           [ Choose ]             split each type of variable into a separate plot             identifies the possible entries within a column, with duplicate entries removed             examine the numerical variables in a data frame             examine the missing data in a data frame             find the presence or absence of a pattern in a sentence             add a new variable to a data frame whilst preserving existing variables

mutate()

           [ Choose ]             split each type of variable into a separate plot             identifies the possible entries within a column, with duplicate entries removed             examine the numerical variables in a data frame             examine the missing data in a data frame             find the presence or absence of a pattern in a sentence             add a new variable to a data frame whilst preserving existing variables

inspect\_num()

           [ Choose ]             split each type of variable into a separate plot             identifies the possible entries within a column, with duplicate entries removed             examine the numerical variables in a data frame             examine the missing data in a data frame             find the presence or absence of a pattern in a sentence             add a new variable to a data frame whilst preserving existing variables

inspect\_na()

           [ Choose ]             split each type of variable into a separate plot             identifies the possible entries within a column, with duplicate entries removed             examine the numerical variables in a data frame             examine the missing data in a data frame             find the presence or absence of a pattern in a sentence             add a new variable to a data frame whilst preserving existing variables

/////////////////////////////////

Here's the matching of each R function to its description:

str\_detect()

find the presence or absence of a pattern in a sentence

unique()

identifies the possible entries within a column, with duplicate entries removed

facet\_wrap()

split each type of variable into a separate plot

mutate()

add a new variable to a data frame whilst preserving existing variables

inspect\_num()

examine the numerical variables in a data frame

inspect\_na()

examine the missing data in a data frame

So, the answers are:

str\_detect() → find the presence or absence of a pattern in a sentence

unique() → identifies the possible entries within a column, with duplicate entries removed

facet\_wrap() → split each type of variable into a separate plot

mutate() → add a new variable to a data frame whilst preserving existing variables

inspect\_num() → examine the numerical variables in a data frame

inspect\_na() → examine the missing data in a data frame

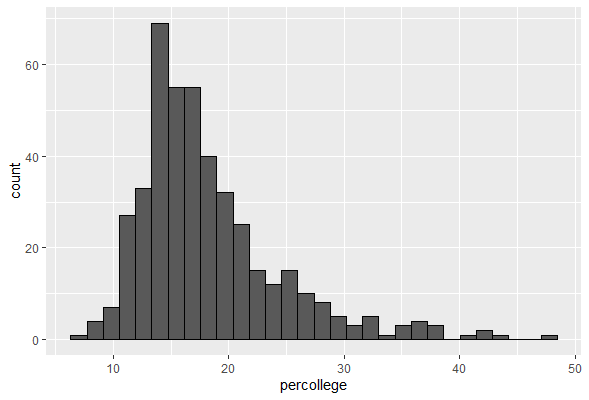
[Flag question: Question 4](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 41 pts**

Below is a histogram of percentages of people with college degrees in towns in the Midwest United States. It was obtained by loading the midwest data set and using

ggplot(midwest, aes(x=percollege))+geom\_histogram(colour="black")

What is true of this histogram?



Group of answer choices

It is right-skewed and unimodal

It has no outliers

It is symmetric

It is left-skewed and unimodal

 ///////////////////////////////////////

Based on the provided histogram:

The majority of the data seems to be concentrated on the left side of the graph, with fewer data points (towns) as the percentage of people with college degrees increases. This indicates that the histogram is right-skewed.

The histogram appears to have one primary peak, which means it is unimodal.

Therefore, the correct answer choice is:

It is right-skewed and unimodal.

[Flag question: Question 5](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 51 pts**

Load the msleep data, and produce a scatterplot with sleep\_total on the x-axis and sleep\_rem on the y-axis. Which of the below options would best describe the relationship?

Group of answer choices

Moderate, positive, linear

Strong, negative, linear

Weak, negative, non-linear

There is no relationship

/////////////////////////////

Based on the provided scatterplot:

The best description for the relationship between sleep\_total on the x-axis and sleep\_rem on the y-axis is:

Moderate, positive, linear.

[Flag question: Question 6](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 61 pts**

Load the diamonds data set and produce side-by-side box plots of price for each colour. Which colour is the most expensive, on average?

Group of answer choices

J

H

I

G

 ////////////////////////////

Based on the provided boxplot, the median (the horizontal line within the box) of the diamond prices is highest for color "J".

[Flag question: Question 7](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 71 pts**

Using the same boxplots as the previous question, which colour has the smallest interquartile range of prices?

Group of answer choices

G

E

D

F

//////////////////////////////////////

The interquartile range (IQR) is the range between the first quartile (bottom of the box) and the third quartile (top of the box).

Based on the provided boxplot, color "E" has the smallest box, indicating the smallest interquartile range of prices.

So, the color with the smallest interquartile range is:

E.

[Flag question: Question 8](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 81 pts**

Continuing with the diamonds data set, what kind of plot would be appropriate for comparing the variables depth and price?

Group of answer choices

Box plots

Scatter plot

Faceted histograms

Stacked bar charts

/////////////////////////////////

To compare two quantitative variables, such as depth and price, a scatter plot is typically the most appropriate choice.

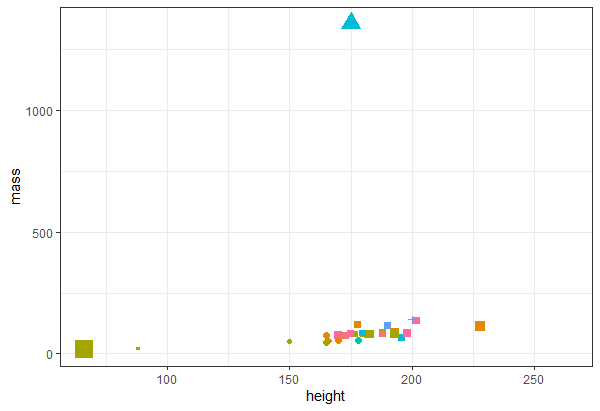
So, the correct choice is:

Scatter plot.

[Flag question: Question 9](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 91 pts**

The following plot has been produced in a similar manner to those you have produced in the previous questions. What aesthetics have been changed from the default? Select all that apply.



Group of answer choices

Shape

Colour

Alpha

Size

Facet\_wrap

////////////////////////////

Therefore, the changed aesthetics from the default are:

Shape

Colour

[Flag question: Question 10](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 101 pts**

Which of the following is NOT true about a 95% confidence interval?

Group of answer choices

On average, 1 out of every 20 95% confidence intervals will not contain the population mean.

If a large number of samples was taken, and a confidence interval calculated for each, 95% of confidence intervals would contain the population mean.

A 95% confidence interval is narrower than a 99% confidence interval.

If a sample is taken, and a confidence interval is produced, 95% of population means will fall within the confidence interval.

//////////////////////////////////

The correct statement that is NOT true about a 95% confidence interval is:

If a sample is taken, and a confidence interval is produced, 95% of population means will fall within the confidence interval.

This statement is incorrect because it reverses the logic of a confidence interval. It's not that 95% of all possible population means will fall within a particular confidence interval, but rather that if we were to repeatedly draw samples and compute confidence intervals, about 95% of those intervals would contain the true population mean.

[Flag question: Spacer](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

For the remaining questions you will be using the midwest data.

The response variable you will use is percollege.

The predictor variables you may use are log(poptotal) and state; where poptotal is the number of people in the county and state is a state in the US.  We take a logarithm of the number of people in the county as poptotal is substantially right-skewed.

If you need to know anything about the data you can use ?midwest in R.

[Flag question: Question 11](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 111 pts**

For the midwest data, fit a simple linear regression model (also known as an identical lines model)  to predict the percentage of people in a Midwestern US county who have a college degree (percollege)using the predictor log(poptotal).

According to this model, the intercept is  and the slope is  ?

Provide your answer to 3 decimal places.

 ////////////////////////////////

To answer this question, we need to fit a simple linear regression model in R using the lm() function. Although I can't run the code directly in this environment, I can guide you on how to do it.

To fit a simple linear regression model in R with percollege as the response variable and log(poptotal) as the predictor, you can use the following code:

R

Copy code

# Load the necessary library

library(ggplot2)

# Fit the linear regression model

model <- lm(percollege ~ log(poptotal), data=midwest)

# Display the summary of the model to get coefficients

summary(model)

After running the above code, you will get a summary of the regression model. Look for the "Coefficients" section in the summary output. The intercept and slope (coefficient for log(poptotal)) will be displayed there.

The value next to (Intercept) will be the intercept, and the value next to log(poptotal) will be the slope. Make sure to round your answers to three decimal places as instructed.

[Flag question: Question 12](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 121 pts**

According to this model, what do we expect the percentage of college graduates to be in a county with population 10,000?

Provide your answer to 3 decimal places.

To find the expected percentage of college graduates in a county with a population of 10,000 based on the model, you'll need to plug in the log-transformed value of 10,000 into the linear regression equation.

The linear regression equation is of the form:

percollege=intercept+slope×log(poptotal)

Given a population (poptotal) of 10,000: log⁡(10,000)=4log(10,000)=4 (Because the logarithm base 10 of 10,000 is 4)

Plug this value into the regression equation: Expected percollege=intercept+slope×4Expected percollege=intercept+slope×4

Now, you would substitute the intercept and slope values you obtained from the regression summary into this equation to get the expected **percollege** for a county with a population of 10,000.

[Flag question: Question 13](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 131 pts**

Now you need to build a multiple linear regression model with both the predictors log(poptotal) and state. Do not include any interaction terms in your model.

Fit the model in R. What is the slope of all the lines?

Provide your answer to 3 decimal places.

 //////////////////////////////////////////

In a multiple linear regression model, when we include a categorical predictor like state, R will create dummy variables for each level of that predictor (excluding one reference level). As a result, there will be multiple slope coefficients corresponding to each level of the state, but only one slope for the continuous predictor log(poptotal).

To determine the slope for log(poptotal) in the context of a multiple linear regression with both log(poptotal) and state, you would run:

# Fit the multiple linear regression model

model\_mult <- lm(percollege ~ log(poptotal) + state, data=midwest)

# Display the summary of the model to get coefficients

summary(model\_mult)

From the summary output, locate the coefficient for log(poptotal). That coefficient will be the slope for log(poptotal) in the context of the multiple regression model. The other coefficients corresponding to each level of state represent the differences in the intercepts from the reference state.

Look for the value next to log(poptotal) to get the slope. Remember to round your answer to three decimal places.

[Flag question: Question 14](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 141 pts**

For a county in Michigan, if we increase log(poptotal) by 3, how much do we expect percollege to change by?          [ Select ]      ["10.8984", "9.9834", "8.8968"]

Would this be different if the county was in Illinois?          [ Select ]      ["No", "Yes"]

[Flag question: Question 15](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 151 pts**

What do we expect the percentage of college graduates to be in a county with population 10,000, in Ohio?

Provide your answer to 3 decimal places.

[Flag question: Question 16](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 161 pts**

Fit a 95% confidence interval for the percentage of college graduates in a town with 10,000 residents in Ohio. How should this interval be expressed?

We are 95% confident that the percentage of college graduates in         [ Select ]      ["the average", "an individual"]  county in Ohio with 10,000 residents is between         [ Select ]      ["3.6819", "6.9784", "8.8364", "12.8676"]  and         [ Select ]      ["11.3284", "23.5730", "11.2064", "14.3873"]  .

[Flag question: Question 17](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 171 pts**

Now fit a multiple regression modal and include the interaction term (i.e. separate lines model). Which state has the largest slope?

Group of answer choices

Ohio

Wisconsin

Illinois

Indiana

Michigan

Texas

[Flag question: Question 18](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 181 pts**

Using anova(), determine whether separate lines are necessary.

Group of answer choices

Separate lines are necessary as the p-value is greater than 0.05

Separate lines are not necessary as the p-value is less than 0.05

Separate lines are not necessary as the p-value is greater than 0.05

Separate lines are necessary as the p-value is less than 0.05

[Flag question: Question 19](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 191 pts**

Using the separate lines model, and keeping the interaction term (even if you decided it was not necessary), answer the following:

For a county in Michigan, if we increase log(poptotal) by 3, how much do we expect percollege to change by?

Provide your answer to 3 decimal places.

[Flag question: Question 20](https://myuni.adelaide.edu.au/courses/84813/quizzes/167546/take)

**Question 201 pts**

Using the separate lines model, and keeping the interaction term (even if you decided it was not necessary) fit a 95% prediction interval for the percentage of college graduates in a town with 10,000 residents in Ohio. How should this interval be expressed?

We are 95% confident that the percentage of college graduates in         [ Select ]      ["an individual", "the average"]  county in Ohio with 10,000 residents is between         [ Select ]      ["-0.3735", "3.6819", "8.8364", "0.7584"]  and         [ Select ]      ["19.4064", "23.5730", "18.5583", "11.3284"]