QUESTION 1

1 points ✓ Save Answer

Which of the following statements about AI, Machine Learning, and/or Data Science is the most accurate?

○ a. Data Science makes uses of machine learning techniques to turn data into useful information.

○ b. Data Science concerns more about visualisation than Machine Learning and AI.

○ c. Data Science is a sub-branch of Machine Learning.

● d. AI, machine learning and Data Science are often used interchangeably to refer to building intelligent programs that learn from data.

QUESTION 2

1 points ✓ Saved

Which of the following code returns a count for each unique category in a categorical variable `cars` which has the following summary?

```
> summary(cars)
   Length     Class      Mode 
      234 character character 
```

○ a. `table(cars)`

● b. `as.factor(cars)`

○ c. `count(as.factor(cars))`

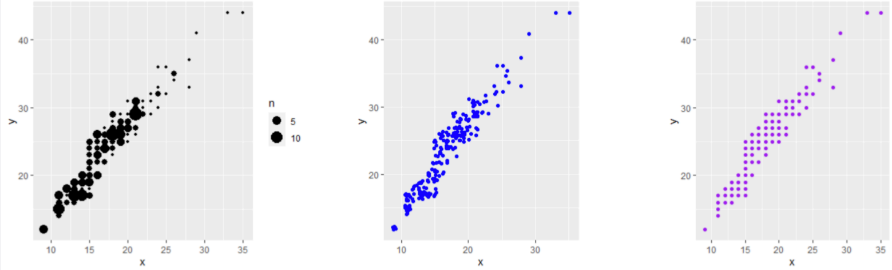
○ d. `levels(as.factor(cars))`

QUESTION 3

1 points ✓ Saved

Given the three plots below for the same data frame, and the same pair of variables, which of the following statements is TRUE? We know the right-most figure is a scatter plot. All three plots are produced in the same code template as below:

`ggplot(df, aes(x, y)) +  geom_???(color="???", ...)`



○ a. All of them are produced by `geom_jitter` with different amount of jittering.

○ b. All of them are `geom_count` plots with different aes mappings, only the left most one has legend turned on.

○ c. Each of the three plots is using a different geom, namely, `geom_count`, `geom_jitter` and `geom_point`.

● d. All of them are scatterplots (`geom_point`) produced with different aes mappings.

QUESTION 4

1 points ✓ Saved

Referring to the three plots in Question 3, if both `x` and y are numerical variables, which of the plots is best for visualising the co-variation between `x` and `y`? Why?

○ a. All of them are good because they all represent the data distribution.

○ b. The right most one as each circle clearly represents where each data point is.

● c. The middle one as we can see better clustering of the data.

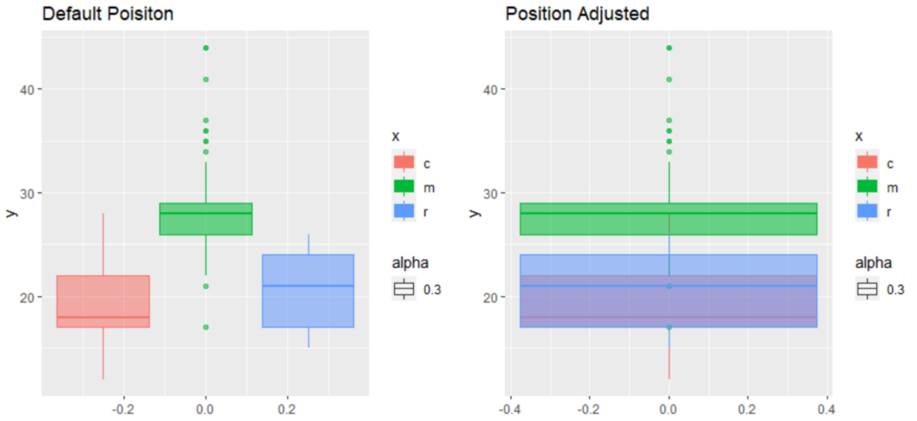
○ d. The left most one as we can see how many data points overlap.

QUESTION 5

1 points ✓ Saved

Same as bar charts, in `ggplot`, one can apply position adjustment to boxplots. Below are two boxplots of the same numerical variable (`y`) , split according to a categorical variable (`x`) . Which of the following code (to replace the ???) will produce the Position Adjusted figure below? Note the `...` in the answers are to be replaced by the correct aesthetic mapping such as `fill, alpha` andetceta.

`ggplot(df, aes(y=y)) +  ???`


Default Poisiton

Position Adjusted

○ a. `geom_boxplot(aes(...), position="fill")`

○ b. `geom_boxplot(aes(...), position="dodge")`

● c. `geom_boxplot(aes(...), position="identity")`

○ d. `geom_boxplot(aes(...), position="stack")`

QUESTION 6

1 points ✓ Saved

Referring to the boxplots in the question above. If the categorical variable `x` records the location of residence, and the numerical variable `y` records the level of satisfaction of Internet Speed in their areas. Which of the following statement is **most** sensible? (Note: `c` is short for CBD, `m` for Metro area, `r` for Rural area).

○ a. The outliers above the upper whisker are the main reasons why CBD residents have higher satisfactions than those in the other two regions.

● b. The Metro area residents are mostly in agreement with their opinion as compared with those in the other two areas.

○ c. The average satisfaction score of CBD and Rural residents are roughly the same.

○ d. There are more residents living in the Rural area than in CBD.

QUESTION 7

1 points ✓ Saved

A government agency wants to analyse the gender distribution of each profession, regardless of the distribution of various professions. Which type of charts would you recommend?

○ a. A stacked bar chart with profession as the primary variable (the x-axis), gender as the aesthetic mapping.

○ b. A filled bar chart with gender as the primary variable (the x-axis), profession as the aesthetic mapping.

● c. A filled bar chart with profession as the primary variable (the x-axis), gender as the aesthetic mapping.

○ d. A stacked bar chart with gender as the primary variable (the x-axis), profession as the aesthetic mapping.

QUESTION 8

1 points ✓ Saved

Given the code below, which of the following statements is TRUE?

```
ggplot(data = a, mapping = aes(x = x,  y=..density..)) +
  geom_histogram(binwidth=1) +
  geom_density(colour = "blue", alpha=0.5 )
```

○ a. The code will work if `..density..` is replaced by `..count..`

● b. The code plots a density plot onto a histogram with `y` values representing the raw count of each bin normalised by the total number of observations.

○ c. The code won't work because you cannot have a variable name (`..density..`) starting with a dot.

○ d. The code won't work because you cannot plot a density plot with histogram because the `y` axis is not of the same scale.

QUESTION 9

1 points ✓ Saved

If we are to draw 3 samples one-by-one from a vector of 6 elements, what are the sampling probabilities if we sample with replacement and if we sample without replacement?

○ a. Sampling with replacement: 1/120; Sampling without replacement 1/216.

○ b. Sampling with replacement and without replacement are roughly the same, both with probability 1/216.

○ c. Sampling with replacement and without replacement are roughly the same, both with probability 1/120.

● d. Sampling with replacement: 1/216; Sampling without replacment 1/120.

QUESTION 10

1 points ✓ Saved

Below is a function that attempts to find the median of an odd-number sized vector (x) of numerical values:

```
myMedian <- function(x) {
    i <- floor(length(x)/2) + 1
    return(c(i, x[i]))
 }
```

Which of the following statements is FALSE?

○ a. The function returns the index and the number in the middle position of an input vector.

○ b. The function works but does not produce the correct median.

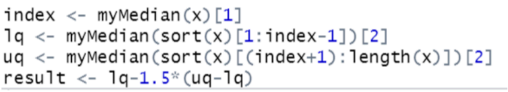
● c. The function returns the median and its index of an input vector.

○ d. The function can be fixed by inserting a sorting function.

QUESTION 11

1 points ✓ Saved

Assuming the `myMedian` function works correctly. What is calculated for the `result` variable?



```
index <- myMedian(x)[1]
lq <- myMedian(sort(x)[1:index-1])[2]
uq <- myMedian(sort(x)[(index+1):length(x)])[2]
result <- lq-1.5*(uq-lq)
```

○ a. It calculates the IQR (Inter-Quartile Range).

○ b. It calculates the lower quartile of `x`.

● c. It calculates the lower whisker of `x`.

○ d. None of the above.

QUESTION 12

1 points ✓ Saved

Which of the following statements ***cannot*** select a subset of the data frame `mydata`? Assuming the use of `attach()` and `detach()`

```
attach(mydata)
```

1. FemaleOver60 <- mydata[which(sex=='F' & age > 60),]
2. FemaleOver60 <- mydata[, sex=='F' & age > 60]
3. if (sex=='F' & age > 60) { FemaleOver60 <- mydata }

```
detach(mydata)
```

○ a. 1) Only

○ b. 1) and 2)

● c. 2) and 3)

○ d. 1) and 3)

QUESTION 13

1 points ✓ Saved

Which of the following is correct about the following code, where `df` is a data frame, `col1` is the name of a column in `df`.

```
df[-"col1"]
```

○ a. It should be `df[, -which(colnames(df)=="col1")]`.

○ b. It should be `df[!col1]`.

● c. It removes `col1` from `df`.

○ d. It should be `df[, -"col1"]`.

QUESTION 14

1 points ✓ Save Answer

When a data frame contain variables that have outliers and numerical values used as flags or codes, which of the following is the best practice?

○ a. Find outliers and replace with NA.

● b. Determine if outliers are nonsensical or sentinel values, replace with NA, but create binary variables for each sentinel values.

○ c. Consult a data dictionary for sentinel values, impute using a meaningful estimate (e.g. mean or median of the corresponding variable).

○ d. Perform list-wise deletion of the observations containing outliers and sentinel values.

QUESTION 15

1 points ✓ Save Answer

Two departments of the same company merged into one after restructuring. The HR team needs to combine two data tables (`df1` for Department 1 and `df2` for Dearpartment 2) together. Apart from one extra column in `df2`, the rest of the variables in the two data frames are the same. What is the most sensible suggestion here?

○ a. Use `rbind()` but we need to first remove the extra column from `df2`, and rearrange the columns for both data frames into a matching order.

○ b. Use `cbind()` as it will automatically detect and merge the same variables and add an extra column to the observations in `df1`.

● c. Use `rbind()` but we need to first add an extra column to `df1`, and populate the column with NAs.

QUESTION 16

1 points ✓ Save Answer

Which of the following about re-producible sampling is `FALSE`?

● a. `runif()` follows a normal distribution, which is more powerful in selecting highly probable values.

○ b. It is essential as we often need to split datasets into training and testing for training and evaluating machine learning models, respectively.

○ c. We can use the `set.seed()` function to ensure the random sampling functions (e.g. `sample()` or `runif()`) to produce the same values each time.

○ d. We can add an extra column to the data frame to store the grouping information, typically obtained through `runif()`.

QUESTION 17

1 points ✓ Save Answer

State Road Authority monitors traffic speed at all major road sections. They use GPS or Wheel Speed Sensors to measure speed as numerical readings. When the data for such variables are missing, which one of the following is the ***least*** reasonable strategy?

○ a. Discretise the numerical values into categories, and then add a separate category for missing values.

○ b. Use clustering or regression models to make use of other variables for imputation.

○ c. You can create a treatment plan using the `vtreat` package, which adds extra columns to flag the missingness and differentiate imputed values from measured ones.

● d. These missing data tend to occur randomly due to sensor failure, we can replace the NAs with the average or median of each numerical variable.

QUESTION 18

1 points ✓ Save Answer

When we assess the possibility of a customer buying a health insurance, where they live could also be a good indicator. So the plan is to combine two data frames, one for customer suburb information (`customer`), one for real estate records of the median house price (`house`) for each suburb. Note some customers may choose not to disclose their residential suburbs. Assuming the only commonly named column of the two data frames is `suburb`. How do we incorporate the median house price while keeping all records in the customer table?

○ a. Use full outer join: `merge(customer, house, all=TRUE)`

○ b. Use right outer join: `merge(customer, house, all.y=TRUE)`

● c. Use left outer join: `merge(customer, house, all.x=TRUE)`

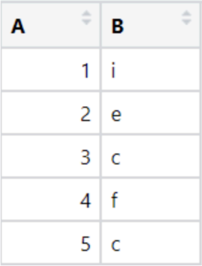
○ d. Use natural join: `merge(customer, house)`

QUESTION 19

1 points ✓ Saved

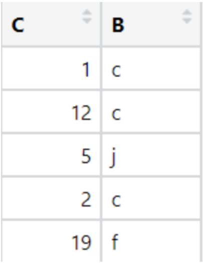
Given two data frames in the picture below, what's the number of records for left outer join on the commonly named column B?

df1



A	B
1	i
2	e
3	c
4	f
5	c

df2



C	B
1	c
12	c
5	j
2	c
19	f

○ a. 8

○ b. 7

● c. 9

○ d. 10

QUESTION 20

1 points ✓ Saved

Using the same two data frames as the previous question, how many records we will get after `semi_join(df1, df2)`?

○ a. 2

○ b. 3

● c. 4

○ d. 7