# Preview Test: Mid-semester Test 2020

## Test Information

Description **To get yourself familiar with the test, please try the online test paper for 2020 here. The instruction given below are for the 2020 students.**

**NOTE:** You can click on the greenish blue colour text **"Question Completion Status:"** to toggle On/Off to see which questions you have or have not attempted.

Instructions Things to note:

- The mid-semester test, same as this mock test, is 90 minutes, which include time to work on questions as well as any trouble shooting time during the test.
- The test is **open-book.** You can bring textbooks, lecture notes, and make use of R or RStudio. However, **you must complete the test on your own**. We will conduct interviews with randomly sampled students after the test, to make sure that students answer the questions by themselves.
- Please ensure that you have your student card with you. Use your first name + surname + student number to join the Zoom session, to help our teaching team save some time on ID checking.
- On the real midterm test date, we will allocate 30mins to do identity checking. You will be allocated in a waiting room, and to be admitted by the host to one of the two meeting sessions. So be on time - the ID checking starts at 1pm, the test finishes at 3pm. Late attendees may be refused to enter the Zoom meeting and attempts at the test will be considered as invalid.
- Please don't use background picture and you are required to keep your video on throughout the test. This will help us monitor whether you are doing the test on your own.
- You can attempt the tests three (3) times but only the last attempt will be graded. Questions during the test should be communicated to the host via private chat.

The password for opening the test paper is: **test-2020**

You can attempt this test 3 times.

Timed Test This test has a time limit of 1 hour and 30 minutes.This test will save and be submitted automatically when the time expires.
Warnings appear when **half the time**, **5 minutes**, **1 minute**, and **30 seconds** remain.*[The timer does not appear when previewing this test]*

Multiple Attempts This test allows 3 attempts. This is attempt number 1.

Force Completion Once started, this test must be completed in one sitting. Do not leave the test before clicking **Save and Submit**.

Your answers are saved automatically.

Question Completion Status:

### QUESTION 1

**1 points** Save Answer

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

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

[✓ b. Data Science makes uses of machine learning techniques to turn data into useful information.]

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

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

### QUESTION 2

**1 points** Save Answer

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. `as.factor(cars)`

○ b. `count(as.factor(cars))`

[✓ c. `table(cars)`]

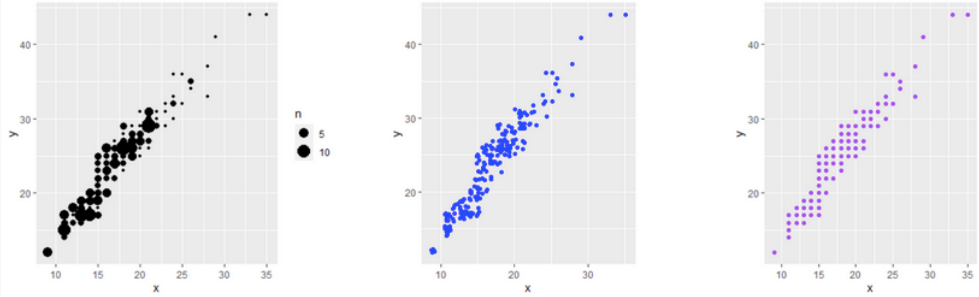
○ d. `levels(as.factor(cars))`

### QUESTION 3

**1 points** Save Answer

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 scatterplots (`geom_point`) produced with different aes mappings.

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

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

[✓ d. Each of the three plots uses a different geom, namely, `geom_count`, `geom_jitter`, and `geom_point`.]

**QUESTION 4**

**1 points** Save Answer

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. The right most one, as each circle clearly represents where each data point is.]

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

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

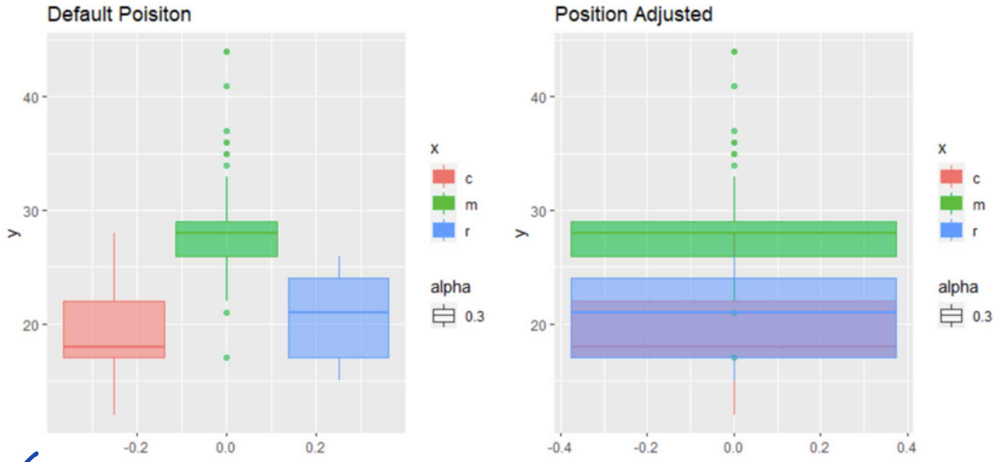
○ d. All of them are good because they all represent the data distribution.

**QUESTION 5**

**1 points** Save Answer

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 and the options such as `fill, alpha`, etc.

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



○ [✓ a. `geom_boxplot(aes(...), position="identity")`]

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

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

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

**QUESTION 6**

**1 points** Save Answer

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

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

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

○ [✓ d. The Metro area residents are mostly in agreement with their opinion as compared with those in the other two areas.]

**QUESTION 7**

**1 points** Save Answer

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 filled bar chart with *gender* as the primary variable (the x-axis) and *profession* as the aesthetic mapping.

○ [✓ b. A filled bar chart with *profession* as the primary variable (the x-axis) and *gender* as the aesthetic mapping.]

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

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

**QUESTION 8**

**1 points** Save Answer

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

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

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

○ [✓ d. 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.]

### QUESTION 9

1 points Save Answer

If we are to draw 3 samples one-by-one from a vector of 6 elements, how many different combinations are there for the sampled outputs

- if we sample with replacement and
- if we sample without replacement?

✓ a. Sampling with replacement: 216; Sampling without replacement 120.

○ b. Sampling with replacement and without replacement are roughly the same, the numbers of combinations for both are 216.

○ c. Sampling with replacement and without replacement are roughly the same, the numbers of combinations for both are 120.

○ d. Sampling with replacement: 120; Sampling without replacement: 216.

### QUESTION 10

1 points Save Answer

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

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

Which of the following statements is FALSE?

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

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

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

✓ d. The function returns the median and its index of the input vector.

### QUESTION 11

1 points Save Answer

Assuming that we have a new function `MyNewMedian` which correctly returns the median and the corresponding index (location) of the median of an odd-sized vector. What is calculated for the `result` variable in the code below?

```
     index <- myNewMedian(x)[2]
     lq <- myNewMedian(sort(x)[1:(index-1)])[1]
     uq <- myNewMedian(sort(x)[(index+1):length(x)])[1]
     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 Save Answer

Which of the following statements ***cannot*** select a subset of the data frame `mydata`, which contains the `sex` and `age` columns? 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 Save Answer

Suppose that we have a data frame `df` and one of its column has the name `col1`. Consider the line below:

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

Which of the following is correct about this line?

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

○ b. It removes `col1` from `df`.

○ c. It should be `df[!col1]`.

○ d. It should be `df[, -"col1"]`.

### QUESTION 14

1 points Save Answer

When a data frame contains 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 Department 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 `cbind()` as it will automatically detect and merge the same variables and add an extra column to the observations in `df1`.

○ [✓ b. Use `rbind()` but we need to firstly add an extra column to `df1`, and populate the column with NAs.]

○ c. 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.

○ d. None of the above.

**QUESTION 16**

1 points Save Answer

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

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

○ b. 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.

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

○ [✓ d. `runif()` follows a normal distribution, which is more powerful in selecting highly probable values.]

**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. 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.

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

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

○ d. 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.

**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 left outer join: `merge(customer, house, all.x=TRUE)`]

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

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

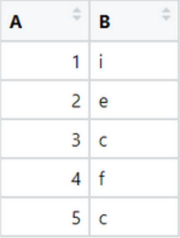
○ d. Use right outer join: `merge(customer, house, all.y=TRUE)`

**QUESTION 19**

1 points Save Answer

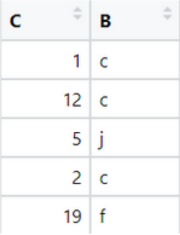
Given two data frames in the picture below, what's the number of records for left outer join on the commonly named column B?

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. 7

○ b. 8

○ [✓ c. 9]

○ d. 10

**QUESTION 20**

1 points Save Answer

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

○ a. 2

○ [✓ b. 3]

○ c. 7

○ d. 4