Preview Test: Mid-semester Test 2021

Test Information

Description	This is an open book test. This test paper contains **20 multiple choice questions**, each is worth 1 point. **The total is 20 points**. You have **1 hour and 30 minutes** to complete the test. You can click on the greenish blue colour text **"Question Completion Status:"** to turn On/Off to see which questions you have or have not attempted during the entire test period. Before submission, ensure that you have attempted all the questions. --- **If you encounter any issue/problem during the test, please** • **send a private chat message on Microsoft Teams to the invigilator, and, if needed,** • **wave/raise your hand in front of your video camera to draw attention of the invigilator.** **The invigilator will try to communicate with you about your problem in private chat messages. If needed, the invigilator can also talk to you in another breakout room or over the phone.**
Instructions	The password for opening the test paper is: test-2021 You can attempt the test 2 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 2 attempts. This is attempt number 2.
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, Big Data, and/or Data Science is the most accurate?

- a. AI, Big Data, and Data Science are often used interchangeably to refer to mining data to extract knowledge and understanding of the data.
- b. Data Science is a multi-disciplinary area encompassing AI, Big Data, and Data Mining.
- ✓ c. Both Big Data and AI have significant overlap with Data Science.
- d. AI and Big Data focus more on algorithm development while Data Science focuses more on data visualization.

QUESTION 2 — **1 points** — Save Answer

Given the R code below:

```
a <- matrix(seq(1,12), 3, 4)
b <- c(1,2,3)
```

which of the following statements is correct?

- a. `a * b` will cause an error as the dimensions mismatch.
- ✓ b. `a * b` and `b * a` both give 3 x 4 matrices.
- c. `b * a` will cause an error but `b %*% a` will not.
- d. `a * b` will cause an error but `a %*% b` will not.

QUESTION 3 — **1 points** — Save Answer

Given the following function which takes in two arguments: `x`, a vector of values, and `t`, a real number, which of the following statements about the function is TRUE?

```
secret <- function(x, t) {
   L <- length(x)
   val <- Inf
   for (i in seq(1,L))
      if (x[i] >= t & x[i] < val) {
     ind <- i
     val <- x[i]
      }
   c(val,ind)
}
```

- a. It returns the smallest value in `x` that is larger than `t` and its index location.
- b. It returns the largest value in `x` that is smaller than `t` and its index location.
- ✓ c. The function will crash for certain values of `t`.
- d. The function will crash as variable `Inf` is not defined.

QUESTION 4 1 points Save Answer

Given the function `func` below which takes in two arguments: a vector `x` and a floating-point number `p`:

```
func <- function(x, p=50) {
   x <- sort(x)
   L <- length(x)
   ind <- ceiling(p * L / 100)
   c(x[ind], ind)
}
```

which of the following statements is (are) TRUE?

1. For the default value of `p`, the function returns the median and the index of the median in the vector.
2. The function works correctly but will crash and give an error message for certain values of `p`.
3. The function returns the `p` percentile of `x` but the index returned is wrong.

- a. 1 and 2
- b. 1 only.
- c. 2 only.
- ✓ d. 3 only.

QUESTION 5 1 points Save Answer

The data frame (`df`) given below is about the jobs undertaken by some graduates who completed their degrees in the last few years.

	person	job	degree	grad.year
1	P1	Engineer	Engineering	2020
2	P2	Data Analyst	Engineering	2017
3	P3	Programmer	Computing	2017
4	P4	Data Analyst	Computing	2019
5	P5	Data Analyst	Physics	2014
6	P6	Contractor	Computing	2018
7	P7	Physicist	Physics	2020
8	P8	Engineer	Computing	2016

Which one of the following R statements is equivalent to the statement below?

```
a <- df[df$degree == "Computing" & df$grad.year < 2018,]
```

- a. `a <- df[df$degree == "Computing" & df$grad.year < 2018]`
- b. `a <- df[degree == "Computing" && grad.year < 2018,]`
- ✓ c. `a <- subset(df, grad.year < 2018 & degree == "Computing")`
- d. `a <- subset(df, degree == "Computing" && grad.year < 2018, select=names(df))`

QUESTION 6 1 points Save Answer

The same data frame `df` from the previous question is shown below again:

	person	job	degree	grad.year
1	P1	Engineer	Engineering	2020
2	P2	Data Analyst	Engineering	2017
3	P3	Programmer	Computing	2017
4	P4	Data Analyst	Computing	2019
5	P5	Data Analyst	Physics	2014
6	P6	Contractor	Computing	2018
7	P7	Physicist	Physics	2020
8	P8	Engineer	Computing	2016

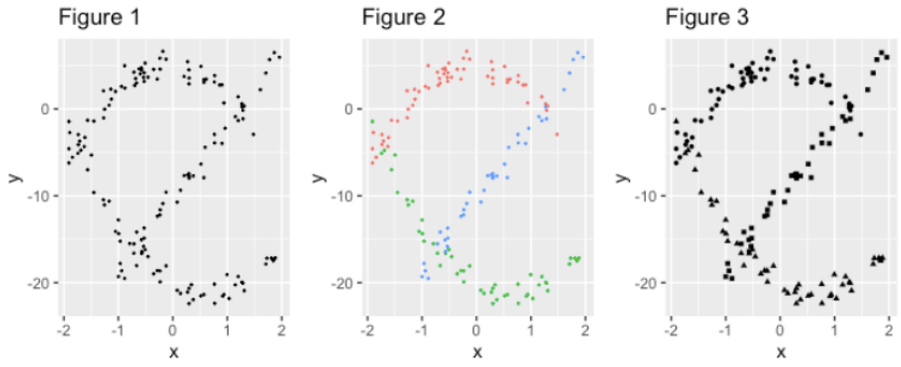
which of the following statements is the most sensible for visualizing the relationship between the two variables `degree` and `job`?

- a. We can plot two boxplots side-by-side, one for each variable.
- b. We can use `geom_tile` and `geom_histogram` to visualize their relationship.
- ✓ c. We can use `geom_count` to visualize their relationship.
- d. It is most suitable to use `hexbin` to visualize their relationship.

QUESTION 7

1 points Save Answer

Consider the three plots shown below for a data frame `df` having two continuous variables `x` and `y` and a categorical variable `z`:



and the code template:

```
ggplot(df) + geom_???(aes(x, y, ?????=z))
```

Which of the following statements is TRUE?

- ✓ a. All of them are `geom_point` functions with `?????` being the `group`, `colour`, and `shape` aesthetic mappings, respectively, for Figures 1, 2, and 3.
- b. All of them are `geom_jitter` with `?????` being the `group`, `colour`, and `type` aesthetic mappings, respectively, for Figures 1, 2, and 3.
- c. All of them are `geom_point` functions. Figures 1 and 3 use the `group` and `shape` aesthetics for the `?????` part. However, the template won't work for Figure 2 as it needs an additional aesthetic mapping to get different colours displayed.
- d. Figures 1 and 3 use `geom_jitter` but Figure 2 uses `geom_point`; The `?????` aesthetics are `group`, `color`, and `type`, respectively, for Figures 1, 2, and 3.

QUESTION 8

1 points Save Answer

A local Council is interested in finding the distributions of primary school students of different age groups within the Council. Suppose that the age range is divided into 3 groups:

- Group 1: `6 <= age < 8`
- Group 2: `8 <= age < 10`
- Group 3: `10 <= age <= 12`

and there are 4 primary schools, labelled as `"A"`, `"B"`, `"C"`, and `"D"`, in the Council. What type of charts would be suitable to visualize the total number of students for each age group in each school?

- ✓ a. A side-by-side bar chart with *school* as the primary variable (the x-axis) and *age group* as the aesthetic mapping.
- b. A stacked bar chart with *school* as the primary variable (the x-axis) and *age group* as the aesthetic mapping.
- c. A filled bar chart with *age group* as the primary variable (the x-axis) and *school* as the aesthetic mapping.
- d. A filled bar chart with *school* as the primary variable (the x-axis) and *age group* as the aesthetic mapping.

QUESTION 9

1 points Save Answer

Suppose that we have a data frame `df` for the primary school students' age data mentioned in the previous question. This data frame has three columns: `stud.name`, `age`, and `school.name`. Each row of the data frame stores a student's name, her/his age, and the name of the school she/he is in. Using the grouping of ages as described in the previous questions:

- Group 1: `6 <= age < 8`
- Group 2: `8 <= age < 10`
- Group 3: `10 <= age <= 12`

which of the following R statements will correctly insert a column `age.group` having values 1, 2, and 3 defined above?

1.
```
df$age.group <- ifelse(df$age >= 6 & df$age < 8, 1,
                       ifelse(df$age >= 8 & df$age < 10, 2,
                              ifelse(df$age >= 10 & df$age <= 12, 3, NA)))
```

2.
```
breaks <- c(6, 8, 10, 12)
df$age.group <- as.numeric(cut(df$age, breaks=breaks, labels=c(1,2,3)))
```

3.
```
df <- within(df, {
    age.group <- NA
    age.group[age >= 6 & age < 8] <- 1
    age.group[age >= 8 & age < 10] <- 2
    age.group[age >= 10 & age <= 12] <- 3
})
```

- a. 1 and 2.
- ✓ b. 1 and 3.
- c. 2 and 3.
- d. All of them.

QUESTION 10 1 points Save Answer

Assuming that you now have an `age.group` column correctly created for the data frame `df` for the previous question. So `head(df)` returns something like:

```
      stud.name age school.name age.group
1     John Clark   8           A         2
2 Rosemary Smith   7           C         1
3      Mark Ford   9           D         2
4      Bing Wong  10           C         3
5    Stewart Lee  12           C         3
6    Betty Walsh  10           B         3
```

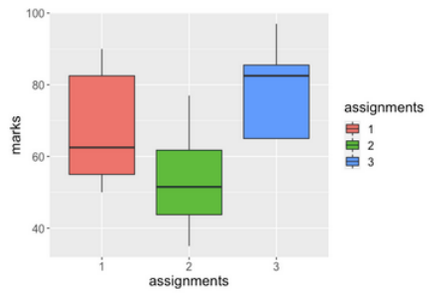
Which of the following R statements will give you the number of students falling into age group 1 for school "A"?

1. `sum( df$age.group == 1 & df$school.name == "A" )`
2. `length( which(df$age.group == 1 & df$school.name == "A") )`
3. `sum( which(df$age.group == 1 & df$school.name == "A") )`
4. `length( df$age.group == 1 & df$school.name == "A" )`

- a. 2 and 3
- b. 3 and 4
- c. 1 and 2 ✓
- d. 1 and 3

QUESTION 11 1 points Save Answer

Given the marks of three assignments of a class of 100 students shown in the boxplots below:



which of the following statements are correct?

1. The marks for assignments 1 and 3 are more skewed than the marks for assignment 2.
2. The blue box for assignment 3 not having a bottom whisker indicates that its lowest mark is above the bottom edge of the box.
3. Almost 50% of students in the class failed assignment 2.

- a. 1 and 2 are correct.
- b. 1 and 3 are correct. ✓
- c. 2 and 3 are correct.
- d. They are all correct.

QUESTION 12 1 points Save Answer

Assuming that the data frame for the boxplots in the previous question is `df`, which has a categorical column `assignments` (containing values `1`, `2`, or `3`) and a numerical column `marks`, which R statements below will generate the boxplots above?

1.
```
ggplot(data=df, mapping=aes(x=assignments, y=marks, fill=assignments)) +
    geom_boxplot()
```
2.
```
ggplot(df) +
    geom_boxplot(mapping=aes(x=assignments, y=marks), fill=assignments)
```
3.
```
ggplot(df) +
    geom_boxplot(aes(x=assignments, y=marks, fill=assignments))
```

- a. 2 and 3.
- b. 1 only.
- c. 1 and 3. ✓
- d. 1 and 2.

QUESTION 13 1 points Save Answer

When a numerical variable in a data frame has missing values. which of the following data cleaning practices is the **least** sensible?

- a. Impute the missing values by the median value and create an indicator column to mark the locations of the missing values.
- b. Use other variables to train a model to predict the missing values and impute them using the predicted values.
- c. Apply a missing value treatment plan such as those provided by the `vtreat` library.
- ✔ d. The best and cleanest practice is to drop the observations that have missing values.

QUESTION 14 1 points Save Answer

Given two data frames `df1`, which contains the prices of four products, and `df2`, which contains four vitamins found in various products:

	product	price
1	P1	10
2	P2	15
3	P3	12
4	P4	20

	vitamin	found.in
1	A	P1
2	A	P2
3	B	P1
4	B	P2
5	B	P4
6	C	P5
7	D	NA

how many records will we get from the *inner join* operation of these data frames on the `product` and `found.in` columns?

- a. 3.
- b. 4.
- ✔ c. 5.
- d. 6.

QUESTION 15 1 points Save Answer

Using the same two data frames from the previous question, how many records will we get after applying the *semi-join* operation on `df1` and `df2` using `product` and `found.in` as the two matching columns?

- ✔ a. 3.
- b. 4.
- c. 5.
- d. 6.

QUESTION 16 1 points Save Answer

Again using the data in `df1` and `df2` above, how many columns do the output records have after applying the *anti-join* and *full-join* operations on the data frames using `product` and `found.in` as the two matching columns?

- a. 1 and 8.
- b. 2 and 2.
- ✔ c. 2 and 3.
- d. 3 and 3.

QUESTION 17 1 points Save Answer

Which of the following statements about data cleaning and data transformation is FALSE?

- a. We can convert a continuous variable to a discrete variable if the range of values is of more interest than the absolute values of the variable.
- ✔ b. Applying z-normalization to a variable allows us to identify outliers in our data.
- c. Even though we have a categorical variable of character type having two distinct values, we may not be able to use `as.logical()` to convert it into logical type.
- d. The `isBad` variables created from the missing data treatment plan should not be mixed up with the original variables of the data in model fitting.

QUESTION 18 1 points Save Answer

Suppose that we sample 4 elements one-by-one from a vector of 6 elements, how many different combinations of outcomes are there if we

- sample with replacement?
- sample without replacement?

- ✔ a. Samplig with replacement: 1296; sampling without replacement: 360.
- b. Samplig with replacement: 360; sampling without replacement: 1296.
- c. Samplig with replacement: 720; sampling without replacement: 4096.
- d. Samplig with replacement: 4096; sampling without replacement: 720.

QUESTION 19

1 points Save Answer

A data frame `df` kept by a home loan company has three main columns as shown below:

```
> head(df)
  custid loan.start.date loan.end.date
1  19536      2012-11-05          <NA>
2  40459      2004-06-27    2012-12-02
3  47031      2010-02-07          <NA>
4  55447      2008-10-20          <NA>
5  56806      2003-10-20    2015-09-18
6  65386      2009-05-10    2018-01-13
```

(for columns of character type, the `head()` function displays missing values as `<NA>` rather than `NA`)
where `custid` is the customer ID, `loan.start.date` is the starting date that the customer took the loan, and `loan.end.date` is the date that the customer paid off the loan. If the loan is still on-going, then it has the `NA` value for `loan.end.date`. The types of the three columns of the data frame are shown below:

```
> str(df)
'data.frame': 1000 obs. of  3 variables:
 $ custid        : num  19536 40459 47031 55447 56807 ...
 $ loan.start.date: chr  "2012-11-05" "2004-06-27" "2010-02-07" "2008-10-20" ...
 $ loan.end.date  : chr  NA "2012-12-02" NA NA ...
```

Which R statement below will add a new column called `loan.duration` containing the number of weeks (which can be floating point numbers) taken by each customer to pay off the loan (if a `loan.end.date` value is `NA`, then the corresponding `loan.duration` value should be `NA` also)?

1. `df$loan.duration <- (as.Date(df$loan.end.date) - as.Date(df$loan.start.date))`
2. `df$loan.duration <- (as.Week(df$loan.end.date) - as.Week(df$loan.start.date))`
3. `df$loan.duration <- difftime(df$loan.end.date, df$loan.start.date, units="weeks")`

- ○ a. 1 only.
- ○ b. 2 only.
- ✓ c. 3 only.
- ○ d. None of them as we have NA in the `loan.end.date` column which should be dealt with separately.

QUESTION 20

1 points Save Answer

The census data collected every 5 years by the *Australian Bureau of Statistics* contains a lot of information about each household in the country. Among this information is the number of persons living in each household. Combining with information about the land area of each suburb, the population density per suburb can be easily estimated. Suppose that we have two data frames:

- `df1`, which has three columns: `house.address`, `suburb`, and `number.persons`
- `df2`, which has two columns: `suburb` and `land.area`. The rows in this data frame are in alphabetical order of the suburb names.

Which of the following R statements will add to `df2` a new column `population.density` to store the population density of each suburb?

1.
```
out <- aggregate(df1[,"number.persons"], FUN=sum, by=df1$suburb)
df2$population.density <- out$x / land.area
```
2.
```
out <- df1 %>% subset(select="number.persons") %>% aggregate(sum, by=list(df1$suburb))
df2 <- within(df2, {population.density <- out$number.persons / land.area})
```

(**Hint:** If you want to experiment with the R statements above in RStudio, you can make up some data for the two data frames `df1` and `df2`. Alternatively, you can modify the customer dataset (`custdata.tsv`) or the `mpg` dataset by renaming some variables. The two data frames do not need to be large.)

- ○ a. 1 only.
- ✓ b. 2 only.
- ○ c. Both of them.
- ○ d. None of them.