Set2

Function Understanding Tasks

distinct

```
beaver1_unq = distinct(beaver1, day, activ)
```

Function: distinct
Text: Remove duplicate rows on day and activ

1. Regarding the operation performed by this function, which of the following statements are correct:
    a. The function does not affect the number of rows in the input table
    b. The function does not affect the number of columns in the input table
    c. There are no cells with the same value in the column *day* of the table *beaver1_unq*
    d. There may be cells with the same value in the column *activ* of the table *beaver1_unq*
    e. None of the above is correct

Answer: d

filter

```
fish_encounters_filter = filter(fish_encounters, station=="BCW", fish>4850)
```

Function: filter
Text: Keep rows where station is ”BCW” and fish > 4850

2. Regarding the operation performed by this function, which of the following statements are correct:
    a. The function does not affect the number of rows in the input table
    b. The function does not affect the number of columns in the input table
    c. The values of the column *fish* in the table *fish_encounters_filter* are all greater than 4850
    d. The function means to filter out rows that satisfy any one of the two conditions
    e. None of the above is correct

Answer: bc

select

```
USArrests_select = select(USArrests, -2, -4)
```

Function: select

Text: Delete Assault, Rape

3. Regarding the operation performed by this function, which of the following statements are correct:
   a. Keep columns from 2 to 4
   b. Delete columns from 2 to 4
   c. Delete the second and fourth rows
   d. Delete the second and fourth columns
   e. None of the above is correct

Answer: d

merge

```
table_merge = merge(table1, table2, by.x = "country", by.y = "country")
```

Function: merge

Text: Merge table1 and table2 on country==country

4. Regarding the operation performed by this function, which of the following statements are correct:
   a. The number of rows in *table_merge* is equal to the sum of the number of rows in *table1* and the number of rows in *table2*
   b. The number of columns in *table_merge* is equal to the sum of the number of columns in *table1* and the number of columns in *table2*
   c. Any value of the column *country* in *table_merge* can be found in the column *country* in *table1* and *table2*
   d. None of the above is correct

Answer: d

gather

```
sleep_gather = gather(sleep, key=name, value=num, extra, group)
```

Function: gather

Text: Convert extra and group into rows

5. Regarding the operation performed by this function, which of the following statements are correct:
   a. The function does not affect the number of rows in the input table
   b. The table *sleep_gather* contains 2 more columns than the table *sleep*
   c. The value of the column *name* in the table *sleep_gather* is either 'extra' or 'group'
   d. The values of the columns *extra* and *group* in the table *sleep* are used as the value of the column *num* in the table *sleep_gather*
   e. None of the above is correct

Answer: cd

Script Understanding Tasks

repo: baltimore-sun-data/baltimore-police-overtime
script: cleaning.R, table: fy2018.csv

```
1   library(dplyr)
2
3   fy2018 <- read.csv('fy2018.csv')
4   fy_overtime = arrange(fy2018, desc(date))
5   fy_overtime = distinct(fy_overtime, emplid, name)
6   fy_overtime = group_by(fy_overtime, emplid)
7   fy_overtime = mutate(fy_overtime, n = row_number())
8   overtime.names.2018 = filter(fy_overtime, n == 1)
9   overtime.names.2018 = select(overtime.names.2018, emplid, name.standardized
10  fy2018 = merge(fy2018, overtime.names.2018, by = 'emplid', all = T)
```

Functions:

1. read.csv
2. arrange
3. distinct
4. group_by
5. mutate
6. row_number
7. filter
8. select
9. merge

Text:

```
fy2018(L3,115R*24C): Create table from "fy2018.csv"
fy_overtime(L4,115R*24C): Sort rows by -date in fy2018(L3)
fy_overtime(L5,11R*2C): Remove duplicate rows on emplid and name in
  fy_overtime(L4)
fy_overtime(L6,11R*2C): Convert fy_overtime(L5) into a grouped table
  by emplid
fy_overtime(L7,11R*3C): Create n from row_number() in fy_overtime(L6)
overtime.names.2018(L8,8R*3C): Keep rows where n is 1 in
  fy_overtime(L7)
overtime.names.2018(L9_1,8R*2C): Keep emplid and name in
  overtime.names.2018(L8)
overtime.names.2018(L9_2,8R*2C): Rename name to "name.standardized" in
  overtime.names.2018(L9_1)
fy2018(L10,115R*25C): Merge fy2018(L3) and overtime.names.2018(L9_2)
  on emplid==emplid
```

Questions:

1. Is overtime.names.2018(L9_1) created by fy_overtime(L5) in one or more data transformations?

a. Yes

b. No

2. How many data transformations are performed from table fy_overtime(L6) to overtime.names.2018(L9_2)?

   a. 2

   b. 3

   c. 4

   d. 5

3. From fy_overtime(L4) to overtime.names.2018(L8), which columns are created?

   a. date

   b. emplid

   c. name

   d. n

   e. name.standardized

4. From the beginning of the script execution, which data tables contribute to the creation of overtime.names.2018(L8)?

   a. fy_overtime(L4)

   b. fy_overtime(L5)

   c. fy_overtime(L6)

   d. fy_overtime(L7)

   e. overtime.names.2018(L9_2)

5. Which data tables in the script are used as input tables for data transformations more than once (at least twice)?

   a. fy2018(L3)

   b. fy_overtime(L4)

   c. fy_overtime(L6)

   d. overtime.names.2018(L8)

   e. overtime.names.2018(L9_1)

Answers:

1. a
2. c
3. d
4. abcd
5. a

1. How helpful were those textual/visual descriptions for completing the tasks?

○ 1 (Not Helpful) ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 (Extremely Helpful)

2. How interpretable were those textual/visual descriptions?

○ 1 (Not Interpretable) ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 (Extremely Interpretable)