PH290 | R for Public Health

Midterm Exam

Name:

INSTRUCTIONS

We suggest that you first download the PDF form to your computer and then open it with Adobe Reader or Adobe Acrobat Pro and fill it. **Do not complete the form on your internet browser.** For the short answer questions, you can either type information directly into each field, or copy and paste text. You can save your responses and re-open the file later to modify or enter additional information. Submit your completed exam to Gradescope.

SECTION 1 Multiple Choice [1 pt each, 10 total]

1. Which of the following will return a value of FALSE?
   x <- 5
   y <- 25

   ○ x != y
   ○ x*x == y
   ○ x^2>y
   ○ x<y

2. For the following character string:
   my_date <- “July 31, 2020”

   Which code would yield a result five months prior to this date?

   ○ mdy(my_date) - months(5)
   ○ mdy(my_date) %m-% months(5)
   ○ as_date(my_date) %m-% months(5)
   ○ as_date(my_date) - months(5)

3. Which of these statements about vectors and lists is <u>false</u>?

   ○ All vectors are lists, but not all lists are vectors
   ○ A single vector can contain both numeric and character values
   ○ A single list can contain both numeric and character values
   ○ Vectors and lists can be indexed using [ ]

4. Which code will create a vector with the following contents?
   2, 4, 6, 8, 10

   ○ seq(2,10,by=2)
   ○ 2:10
   ○ seq(2,10,length.out=4)
   ○ even(2,10)

5. For the following vector:

   ```
   v <- c(NA, 5, 10, NA, NaN)
   ```

   What is the output of is.na(v)?

   ○ TRUE
   ○ TRUE FALSE FALSE TRUE TRUE
   ○ 2
   ○ TRUE FALSE FALSE TRUE FALSE

6. For the list that is generated by this code:

   ```
   multi_list <- list(
   “Numbers” = seq(3,21,by=3),
    “Matrix” = matrix(c(-3,9,6,12,3,21),
     nrow = 2),
     “Words”=list(“one”,“two”,“three”))
   ```

   Which of the following will not return a single value of 3?

   ○ multi_list[["Numbers"]][1]
   ○ multi_list[[2]][1,3]
   ○ length(multi_list[["Words"]])
   ○ multi_list["Numbers"][1]

7. Data frames and tibbles are two options for storing tabular data in R. Which of the following is false?

   ○ Tibbles do not default to converting character values to factors, whereas data frames do
   ○ Tibbles have more flexibility than data frames for naming columns (i.e. allowing spaces and symbols)
   ○ Tibbles cannot be indexed or subset in the same way as data frames (i.e. using [ ], [[]], or $)
   ○ Tibbles do not require row names, whereas data frames do

8. Using the df below, which of the following will return a vector of the values of cases?

state	year	cases
CA	2019	34
CA	2020	23
AZ	2019	89
AZ	2020	27

○ df$cases
○ df[3]
○ df[[cases]]
○ df["cases"]

9. Using the df below, which of the following will not return this subset data frame:

Original df:

state	year	cases
CA	2019	34
CA	2020	23
AZ	2019	89
AZ	2020	27

Subset dataframe:

state	year	cases
CA	2020	23
AZ	2020	27

○ df[which(df$year==2020),]
○ df[c(2,4)]
○ df[df$year==2020,]
○ subset(df,year==2020)

10. There is a need for developing a function that calculates the volume of a rectangular shipping container (length*width*height). If the volume is less than or equal to 1000 cubic feet, then the function should return “too small”, if it is greater than or equal to 2000 cubic feet it should return “too big”, and if it is between 1000-2000 cubic feet it should return “just right”.

Which function will return the correct value when

Length (l)= 12
Width (w) = 5
Height (h)= 6

○

```
check_volume <- function(l, w, h) {
 if(volume >= 2000) {
  return("too big")
 } else if (volume <= 1000){
   return("too small")
 } else {
   return("just right")
 }
}
```

○

```
check_volume <- function(l, w, h) {
  volume <- l*w*h

  if(volume >= 2000) {
   return("too big")
  } else if (volume < 2000){
    return("just right")
  } else {
    return("too small")
  }
}
```

○

```
check_volume <- function(l, w, h) {
  volume <- l*w*h

  if(volume >= 2000) {
   return("too big")
  } else if (volume <= 1000){
    return("too small")
  } else {
    return("just right")
  }
}
```

```
○check_volume <- function(l, w, h) {
  volume <- l*w*h

  if(volume >= 2000) {
    output <- "too big"
  } else if (volume <= 1000){
    output <- "too small"
  } else {
    output <- "just right"
  }

  return(volume)
}
```

SECTION 2: Short Answer

For all questions below, use the “inpatient_payer_ca.csv” file that is saved here: ~/PHW290_Fall2020.git/midterm. This is a real dataset from the California Health and Human Services Open Data Portal, but has been altered slightly for purpose of this exam. The dataset contains counts of inpatient stays by expected payer source.

All questions below should be answered using R. Please paste your code at the end of the exam. Your code will only be used to evaluate for partial credit on problems that are missed.

The file includes the following columns:

- Year
- OSHPD Facility Number
- Facility Name
- County Name
- Expected Payer
- Cout - count of inpatient stays per expected payer
- Total Inpatient Stays - total count of inpatient stays per facility per year

11. Import the csv data file. **[2 pts]**
    a. How many rows and columns are there?
        i. Rows =

        ii. Columns =

b. The “OSHPD Facility Number” column is reading in as character which is causing some ID’s to have leading zeros, creating some inconsistencies in the data. Force it to read in as numeric to drop the leading 0.

   Paste the argument used at import here:

c. Notice the column names are not reading in a very user-friendly way. Rename all columns to align with best practices for naming.

   Paste new column names here:

12. The column that originally read in as “Expected Payer” contains categories that are inconsistently named, but obviously mean the same thing. Clean up the values to include 9 categories. **[1 pt.]**

    Paste the 9 unique categories here:

13. The values in the County column contain a mixture of upper case, lower case, and title case. Change the values to all be in a consistent case. **[1 pt.]**

    Enter the number of unique county values <u>before</u> changes here:

    Enter the number of unique county values <u>after</u> changes here:

    **Hint*: Number of unique values can be obtained by applying the length() function to a vector of values.

14. There are some facilities with duplicate records for a year and payer type; in some cases the counts on these records differ. Retain only one record per year, facility id, and payer type; for instances where there are records with different counts, retain the record with the highest number. **[1 pt.]**

    a. **How many rows remain?**

    b. **What is the count of Private Coverage inpatient stays at Kern Medical Center for 2013?**

15. There are some facilities that have used negative numeric values to indicate unknown for the "Counts" column; these values are invalid. Replace any invalid values with missing. **[1 pt.]**

    a. **After replacing these values, how many rows are missing a value for "Counts"?**

16. Limit dataset to only contain rows for Medi-Cal coverage. Create a new column called **pct_medi_cal** that calculates the percent of inpatient stays that were covered by Medi-Cal. Display the percentage as multiplied by 100 and rounded to 1 decimal (for example, 35.1% as 35.1); replace any missing or invalid values with 0. **[1 pt.]**

    a. **How many records are in the subsetted dataset?**

    b. **What is the value of pct_medi_cal for Eden Medical Center in 2015?**

17. Create a new column called **quartiles** that categorizes percent based on quartile of percent paid by Medi-Cal. **[1 pt.]**

    a. **What is the lower bound for the top quartile?**

    b. **What quartile is Woodland Memorial Hospital in for 2010?**

18. Order the table to display rows by most recent year and percent of inpatient visits that were covered by Medi-Cal. **[1 pt]**

    a. **Enter the Facility Name(s) with the highest proportion of inpatient visits covered by Medi-Cal in 2015.**

19. There is interest in knowing specifically about payers for Kaiser facilities. Create a new variable to flag observations for Kaiser facilities. Restrict the dataset to only include records for Kaiser, Medi-Cal, Counts >0, and year is 2015. **[1 pt]**

    a. **How many records remain?**

    b. **What Kaiser facility has the highest percent of Medi-Cal covered inpatient stays?**

EXTRA CREDIT [2 pts]

20. Using the dataset created in question #18, use dplyr functions to answer the following questions:

    a. **What county has the highest total number of Medi-Cal covered inpatient stays in 2015?**

    b. **In 2014, which county had the highest mean facility-level percent of inpatient stays covered by Medi-Cal?**

Paste your code here: