DS 5110 - Fall 2019 Exam 2

Name:_____ **Year/Progam:**_____

Instructions

Answer each problem to the best of your ability. Fully read instructions for all sections. Justify or explain your answers when appropriate. Partial credit will be given for answers that are partially correct. Points will be deducted for incorrect statements even if all other parts of your answer are correct.

All data tables can be found at the end of the exam. You may remove that page from the exam for your convenience when referencing them.

No notes are permitted for this exam. Use of computers, smartphones, or other unauthorized aides while taking the exam is strictly prohibited.

You may ask the instructor or a TA to display the help/documentation page of any function from base R or `tidyverse` packages.

Honor statement

I promise I will not cheat on this exam. I will neither give nor receive any unauthorized assistance. I will not share information about the exam with anyone who may be taking it at a different time. I have not been told anything about the exam by someone who has taken it earlier.

Signature:_____ **Date:**_____

Part	Points Possible	Points Received
A	30	
B	30	
C	40	
Total	100	

Part A

This section uses multiple choice. For each problem, circle the **best** answer for each question.

All datasets referenced can be found at the end of the exam.

1. (3 pts) What is the primary key of the `customers` table in Appendix A?

a. `name`
b. `email`
c. `customer_id`
d. `item_id`
e. No primary key

2. (3 pts) What is the primary key of the `inventory` table in Appendix A?

a. `name`
b. `order_id`
c. `customer_id`
d. `item_id`
e. No primary key

3. (3 pts) What is the primary key of the `orders` table in Appendix A?

a. `name`
b. `order_id`
c. `customer_id`
d. `item_id`
e. No primary key

4. (3 pts) Which of the following variables is a foreign key?

a. `customers$customer_id`
b. `customers$email`
c. `inventory$item_id`
d. `inventory$description`
e. None of the above

5. (3 pts) Which of the following variables is a foreign key?

a. `customers$customer_id`
b. `inventory$item_id`
c. `orders$order_id`
d. `orders$item_id`
e. None of the above

6. (3 pts) Which of the following variables is a foreign key?

a. `customers$email`
b. `inventory$price`
c. `orders$customer_id`
d. `orders$data`
e. None of the above

7. (3 pts) Why do we perform cross-validation?

a. To make sure the model assumptions are valid.
b. To report a realistic predictive accuracy
c. To compare different models and parameters
d. Both (a) and (b)
e. Both (b) and (c)

8. (3 pts) What is the purpose of a validation set?

a. To make sure the model assumptions are valid.
b. To report a realistic predictive accuracy
c. To compare different models and parameters
d. To train the regression model
e. None of the above

9. (4 pts) What assumptions do we make when we fit a linear regression model?

a. The relationship between the response variable and the explanatory variables is linear
b. The residuals are randomly distributed following an approximate normal distribution
c. The explanatory variables are independent (i.e., not correlated with each other)
d. All of the above
e. None of the above

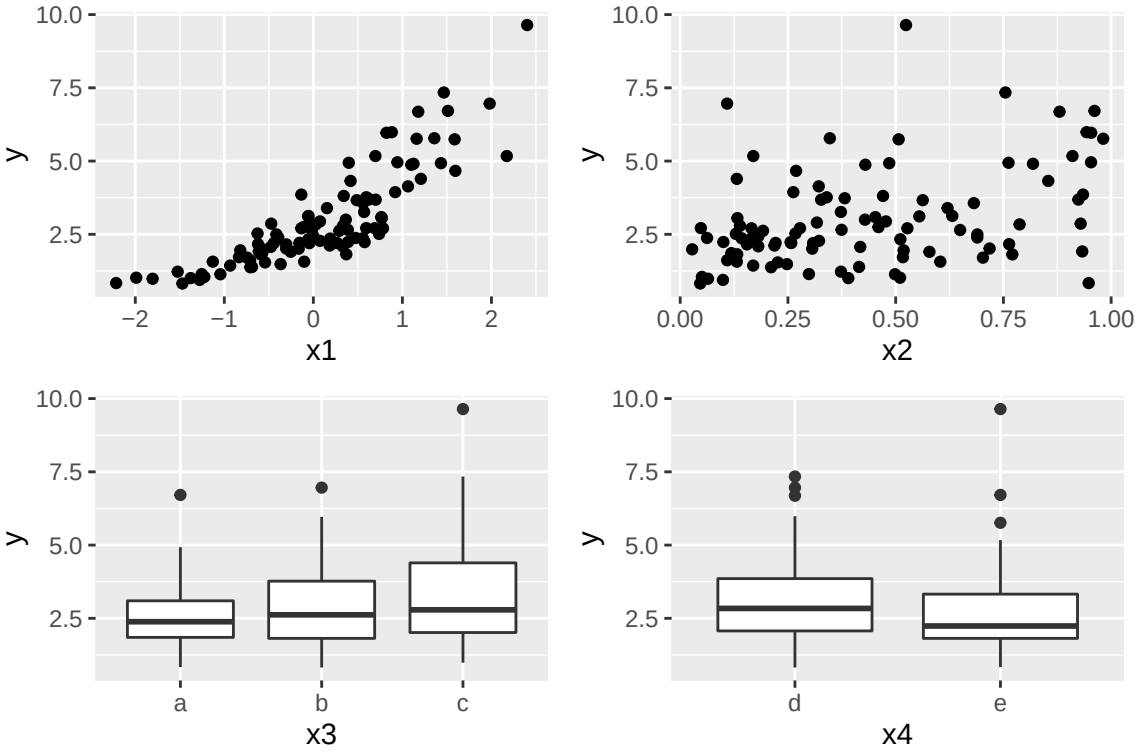
10. (3 pts) Suppose we wish to fit a linear model to predict the temperature using the month of the year and the hour of the day. How many total parameters must be estimated? (There are 12 months in a year, and 24 hours in a day.)

a. 2
b. 3
c. 34
d. 35
e. 36

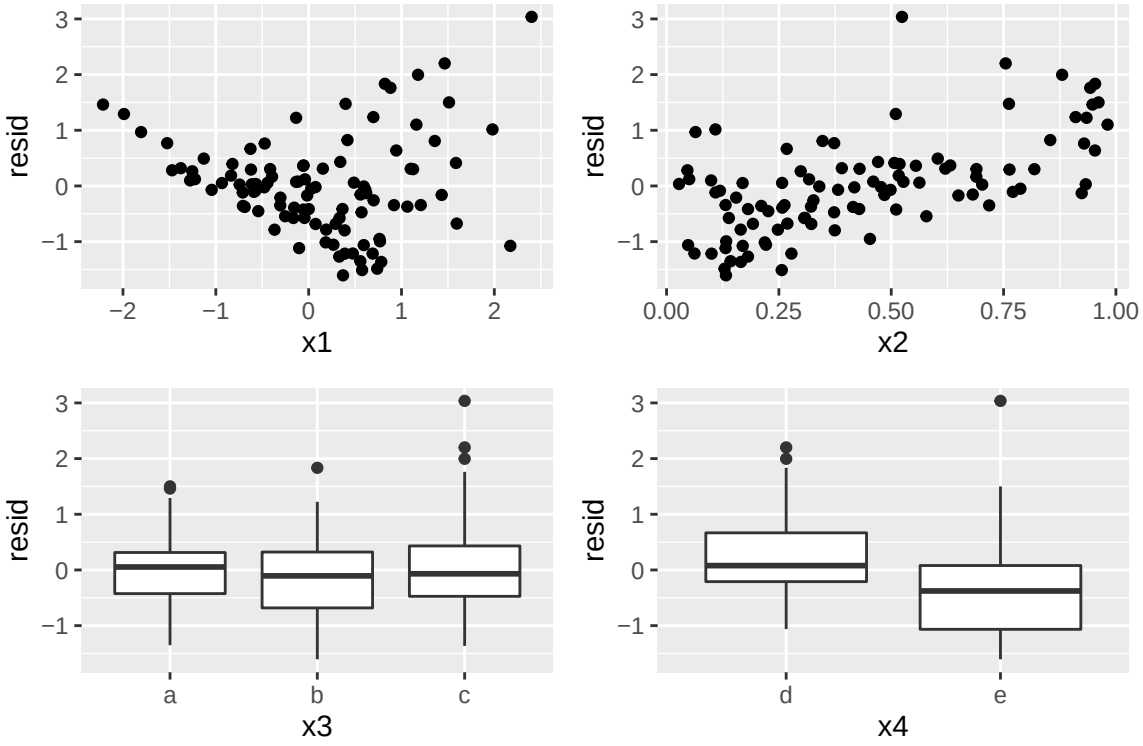
Part B

The problems in this section are free response. For each problem, you are given a set of plots. Answer the questions based on the plots, using the plots the justify your answer.

11. (15 pts) We would like to build a model for predicting `y`. It is plotted below against potential predictors `x1`, `x2`, `x3`, and `x4`. State your conclusions about each potential predictor and its inclusion in the model. Then propose an initial model you would fit for `y`.



12. (15 pts) The model `lm(y ~ x1)` was fit. The residuals are plotted below against `x1`, `x2`, `x3`, and `x4`. State your conclusions about each variable and its inclusion in the model. Then propose a new model you would fit for `y`.



Part C

In this section, provide a pseudocode strategy (using relational data concepts such as `group_by()`, `summarise()`, joins such as `left_join()` and `right_join()`, etc.) for solving each problem.

All datasets referenced can be found at the end of the exam. You do not need to account for missing data or other special cases. You do not need to calculate anything.

Answers may vary

13. (10 pts) Customer and transaction information for a certain online vendor are given in Appendix B. Calculate the total price of each order in `orders`.

14. (10 pts) Calculate the average price of items from each department.

15. (10 pts) Calculate the total amount of revenue for each department across all orders.

16. (10 pts) Create a new table with all items that have not yet been ordered by anyone.

Appendix A

The following three data tables describe the customer information, inventory items, and online orders for a certain vendor:

```
customers
```

```
## # A tibble: 7 x 3
##   customer_id name           email
##         <dbl> <chr>          <chr>
## 1           1 John Smith     john@thedude.com
## 2           2 Kelly Shay     kt598@h0tmail.com
## 3           3 Simone Arnold  coolchick99@geemail.com
## 4           4 Denise Sanchez dsanchez@outlooook.org
## 5           5 Shirley Grace  noreply@somedomain.edu
## 6           6 John Smith     jsmith@harvrad.org
## 7           7 Aiden Shu      noreply@somedomain.edu
```

```
inventory
```

```
## # A tibble: 7 x 4
##   item_id description        department       price
##     <dbl> <chr>              <chr>            <dbl>
## 1       7 Black wood chair   Furniture        60.9
## 2       8 XE Laptop computer Electronics    2200.
## 3      11 Sandalwood desk    Furniture       111.
## 4      13 Shiny thing        Toys and Games 1000.
## 5     113 Mini screwdriver   Tools             5.76
## 6     213 Black wood chair   Furniture       161.
## 7     226 Deck playing cards Toys and Games    5.76
```

```
orders
```

```
## # A tibble: 10 x 4
##    order_id customer_id item_id date
##       <dbl>       <dbl>   <dbl> <chr>
##  1     1001           2       7 10/3/18
##  2     1001           2       7 10/3/18
##  3     1001           2      11 10/3/18
##  4     1004           4       8 10/3/18
##  5     1022           5     113 10/6/18
##  6     1022           5       8 10/6/18
##  7     1103           1     226 10/6/18
##  8     1103           1     213 10/6/18
##  9     1268           5     226 10/8/19
## 10     1299           4       7 10/8/18
```