Final Exam - DSC 10, Spring 2022

Instructions:

- This exam consists of 18 questions. A total of ~~102~~ 105 points are available, but the exam will be graded out of 100, so there are ~~2~~ 5 points of extra credit possible.
- Write your PID or name in the top right of each page in the space provided.
- Please write neatly in the provided answer boxes. We will not grade work that appears elsewhere.
- Completely fill in bubbles and square boxes.

  ○ A bubble means that you should only **select one choice**.

  □ A square box means you should **select all that apply**.
- You may refer to the DSC 10 reference sheet only. No other resources or technology (including calculators) are permitted.

Full Name:

PID:

Lecture: ○ A00, 10-10:50am. Final Exam in Center 216.
○ B00, 11-11:50am. Final Exam in Center 214.

By signing below, you are agreeing that you will behave honestly and fairly during and after this exam. You should not discuss any part of this exam with anyone enrolled in the course who has not yet taken the exam (this includes posting questions about the exam on Campuswire!)

Signature:

Please do not open your exam until instructed to do so.

Version B

Question 1 (4 points)

Complete the expression below so that it evaluates to a DataFrame indexed by "category" with one column called "price" containing the median cost of the products in each category.

```
ikea.___(a)___.get(___(b)___)
```

a) What goes in blank (a)?

b) What goes in blank (b)?

Question 2 (4 points)

Complete the expression below so that it evaluates to the name of the product for which the average assembly cost per package is lowest.

```
(ikea.assign(assembly_per_package = ___(a)___)
     .sort_values(by="assembly_per_package").___(b)___)
```

a) What goes in blank (a)?

b) What goes in blank (b)?

Question 3 (5 points)

In the `ikea` DataFrame, the first word of each string in the `"product"` column represents the product line. For example the HEMNES line of products includes several different products, such as beds, dressers, and bedside tables.

The code below assigns a new column to the `ikea` DataFrame containing the product line associated with each product.

```
(ikea.assign(product_line = ikea.get("product")
                                .apply(extract_product_line)))
```

a) What are the input and output types of the `extract_product_line` function?

○ takes a Series as input, returns a string
○ takes a Series as input, returns a Series
○ takes a string as input, returns a string
○ takes a string as input, returns a Series

b) Complete the return statement in the `extract_product_line` function below.

```
def extract_product_line(x):
    return _________
```

What goes in the blank?

Question 4 (8 points)

Recall that we have the complete set of currently available discounts in the DataFrame `offers`.

The DataFrame `with_offers` is created as follows.

```
(with_offers = ikea.take(np.arange(6))
                        .merge(offers, left_on="category",
                                        right_on="eligible_category"))
```

a) How many rows does `with_offers` have?

b) How many rows of `with_offers` have a value of 20 in the `"percent_off"` column?

c) If you can use just one offer per product, you'd want to use the one that saves you the most money, which we'll call the best offer.

True or False: The expression below evaluates to a Series indexed by `"product"` with the name of the best offer for each product that appears in the `with_offers` DataFrame.

```
with_offers.groupby("product").max().get("offer")
```

○ True

○ False

d) You want to add a column to `with_offers` containing the price after the offer discount is applied.

```
with_offers = with_offers.assign(after_price = _________)
with_offers
```

Which of the following could go in the blank? Select all that apply.

□ `(with_offers.get("price") - with_offers.get("price")*with_offers.get("percent_off")/100)`

□ `with_offers.get("price")*(100 - with_offers.get("percent_off")/100)`

□ `with_offers.get("price") - with_offers.get("percent_off")/100`

□ `with_offers.get("price")*(100 - with_offers.get("percent_off"))/100`

Question 5 (9 points)

Recall that an IKEA fan created an app for people to log the amount of time it takes them to assemble an IKEA product. We have this data in `app_data`.

a) Suppose that when someone downloads the app, the app requires them to choose a username, which must be different from all other registered usernames.

True or False: If `app_data` had included a column with the username of the person who reported each product build, it would make sense to index `app_data` by username.

○ True

○ False

b) What does the code below evaluate to?

```
(app_data.take(np.arange(4))
          .sort_values(by="assembly_time")
          .get("assembly_time")
          .iloc[0])
```

Hint: The `"assembly_time"` column contains strings.

c) Complete the implementation of the `to_minutes` function below. This function takes as input a string formatted as "x hr, y min" where x and y represent integers, and returns the corresponding number of minutes, **as an integer** (type `int` in Python).

For example, `to_minutes("3 hr, 5 min")` should return 185.

```
def to_minutes(time):
    first_split = time.split(" hr, ")
    second_split = first_split[1].split(" min")
    return _________
```

What goes in the blank?

d) You want to add to `app_data` a column called `"minutes"` with integer values representing the number of minutes it took to assemble each product.

```
app_data = app_data.assign(minutes = _________)
app_data
```

Which of the following should go in the blank?

○ `to_minutes("assembly_time")`

○ `to_minutes(app_data.get("assembly_time"))`

○ `app_data.get("assembly_time").apply(to_minutes)`

○ `app_data.get("assembly_time").apply(to_minutes(time))`

Question 6 (4 points)

We want to use `app_data` to estimate the average amount of time it takes to build an IKEA bed (any product in the "bed" category). Which of the following strategies would be an appropriate way to estimate this quantity? Select all that apply.

□ Resample with replacement many times. For each resample, first query to keep only the beds and then take the mean of the `"minutes"` column. Compute a 95% confidence interval based on those means.

□ Resample with replacement many times. For each resample, first query to keep only the beds. Then group by `"product"` using the mean aggregation function, and finally take the mean of the `"minutes"` column. Compute a 95% confidence interval based on those means.

□ Query to keep only the beds. Then resample with replacement many times. For each resample, take the mean of the `"minutes"` column. Compute a 95% confidence interval based on those means.

□ Query to keep only the beds. Group by `"product"` using the mean aggregation function. Then resample with replacement many times. For each resample, take the mean of the `"minutes"` column. Compute a 95% confidence interval based on those means.

Question 7 (3 points)

Laura built the LAPPLAND TV storage unit in 2 hours and 30 minutes, and she thinks she worked at an average speed. If you want to see whether the average time to build the TV storage unit is indeed 2 hours and 30 minutes using the sample of assembly times in `app_data`, which of the following tools **could** you use to help you? Select all that apply.

□ bootstrapping

□ confidence interval

□ hypothesis testing

□ permutation testing

□ Central Limit Theorem

□ regression

Question 8 (13 points)

For this question, let's think of the data in `app_data` as a random sample of all IKEA purchases and use it to test the following hypotheses.

Null Hypothesis: IKEA sells an equal amount of beds (category "bed") and outdoor furniture (category "outdoor").

Alternative Hypothesis: IKEA sells more beds than outdoor furniture.

The DataFrame `app_data` contains 5000 rows, which form our sample. Of these 5000 products,

- 1000 are beds,
- 1500 are outdoor furniture, and
- 2500 are in another category.

a) Which of the following **could** be used as the test statistic for this hypothesis test? Select all that apply.

□ Among 2500 beds and outdoor furniture items, the number of beds.

□ Among 2500 beds and outdoor furniture items, the number of beds plus the number of outdoor furniture items.

□ Among 2500 beds and outdoor furniture items, the absolute difference between the proportion of beds and the proportion of outdoor furniture.

□ Among 2500 beds and outdoor furniture items, the proportion of beds.

b) Let's do a hypothesis test with the following test statistic: among 2500 beds and outdoor furniture items, the proportion of outdoor furniture minus the proportion of beds.

Complete the code below to calculate the observed value of the test statistic and save the result as `obs_diff`.

```
outdoor = (app_data.get("category")=="outdoor")
bed = (app_data.get("category")=="bed")
obs_diff = ( ___(a)___ - ___(b)___ ) / ___(c)___
```

The table below contains several Python expressions. Choose the correct expression to fill in each of the three blanks. Three expressions will be used, and two will be unused.

Python expression	Which blank does this go in?
`app_data.shape[0]`	○ (a) ○ (b) ○ (c) ○ None
`app_data[outdoor].shape[0]`	○ (a) ○ (b) ○ (c) ○ None
`app_data[bed].shape[0]`	○ (a) ○ (b) ○ (c) ○ None
`app_data[outdoor & bed].shape[0]`	○ (a) ○ (b) ○ (c) ○ None
`app_data[outdoor | bed].shape[0]`	○ (a) ○ (b) ○ (c) ○ None

c) Which of the following is a valid way to generate one value of the test statistic according to the null model? Select all that apply.

□ Way 1:

```
choice = np.random.choice([0, 1], 2500, replace=True)
choice_sum = choice.sum()
(choice_sum - (2500 - choice_sum))/2500
```

□ Way 2:

```
choice = np.random.choice(["bed", "outdoor"], 2500, replace=True)
bed = np.count_nonzero(choice=="bed")
outdoor = np.count_nonzero(choice=="outdoor")
outdoor/2500 - bed/2500
```

□ Way 3:

```
multi = np.random.multinomial(2500, [0.5,0.5])
(multi[0] - multi[1])/2500
```

□ Way 4:

```
outdoor = np.random.multinomial(2500, [0.5,0.5])[0]/2500
bed = np.random.multinomial(2500, [0.5,0.5])[1]/2500
outdoor - bed
```

□ Way 5:

```
outdoor = (app_data.get("category")=="outdoor")
bed = (app_data.get("category")=="bed")
samp = app_data[outdoor|bed].sample(2500, replace=True)
(samp[samp.get("category")=="outdoor"].shape[0]/2500 -
 samp[samp.get("category")=="bed"].shape[0]/2500)
```

□ Way 6:

```
outdoor = (app_data.get("category")=="outdoor")
bed = (app_data.get("category")=="bed")
samp = (app_data[outdoor|bed].groupby("category").count()
        .reset_index().sample(2500, replace=True))
(samp[samp.get("category")=="outdoor"].shape[0]/2500 -
 samp[samp.get("category")=="bed"].shape[0]/2500)
```

d) Suppose we generate 10,000 simulated values of the test statistic according to the null model and store them in an array called `simulated_diffs`. Complete the code below to calculate the p-value for the hypothesis test.

```
np.count_nonzero(simulated_diffs _________ obs_diff)/10000
```

What goes in the blank?

○ >

○ >=

○ <

○ <=

Question 9 (4 points)

You are browsing the IKEA showroom, deciding whether to purchase the BILLY bookcase or the LOMMARP bookcase. You are concerned about the amount of time it will take to assemble your new bookcase, so you look up the assembly times reported in `app_data`. Thinking of the data in `app_data` as a random sample of all IKEA purchases, you want to perform a permutation test to test the following hypotheses.

Null Hypothesis: The assembly time for the BILLY bookcase and the assembly time for the LOMMARP bookcase come from the same distribution.

Alternative Hypothesis: The assembly time for the BILLY bookcase and the assembly time for the LOMMARP bookcase come from different distributions.

a) Suppose we query `app_data` to keep only the BILLY bookcases, then average the `"minutes"` column. In addition, we separately query `app_data` to keep only the LOMMARP bookcases, then average the `"minutes"` column. If the null hypothesis is true, which of the following statements about these two averages is correct?

○ These two averages are the same.

○ The difference between these averages is statistically significant.

○ Any difference between these two averages cannot be ascribed to random chance alone.

○ Any difference between these two averages is due to random chance.

b) For the permutation test, we'll use as our test statistic the average assembly time for BILLY bookcases minus the average assembly time for LOMMARP bookcases, in minutes.

Complete the code below to generate one simulated value of the test statistic in a new way, without using `np.random.permutation`.

```
billy = (app_data.get("product") ==
         "BILLY Bookcase, white, 31 1/2x11x79 1/2")
lommarp = (app_data.get("product") ==
           "LOMMARP Bookcase, dark blue-green, 25 5/8x78 3/8")
billy_lommarp = app_data[billy|lommarp]
billy_mean = np.random.choice(billy_lommarp.get("minutes"),
                              billy.sum()).mean()
lommarp_mean = _________
billy_mean - lommarp_mean
```

What goes in the blank?

○ `np.random.choice(billy_lommarp.get("minutes"), lommarp.sum()).mean()`

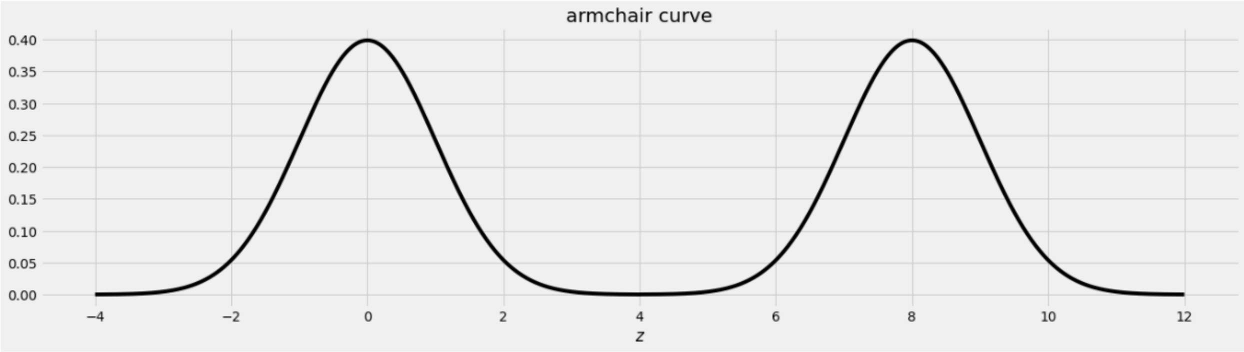
○ `billy_lommarp.get("minutes").mean() - billy_mean`

○ `billy_lommarp[lommarp].get("minutes").mean()`

○ `(billy_lommarp.get("minutes").sum() - billy_mean*billy.sum())/lommarp.sum()`

Question 10 (8 points)

An IKEA chair designer is experimenting with some new ideas for armchair designs. She has the idea of making the arm rests shaped like bell curves, or normal distributions. A cross-section of the armchair design is shown below.



This was created by taking the portion of the standard normal distribution from $z = -4$ to $z = 4$ and adjoining two copies of it, one centered at $z = 0$ and the other centered at $z = 8$. Let's call this shape the armchair curve.

Since the area under the standard normal curve from $z = -4$ to $z = 4$ is approximately 1, the total area under the armchair curve is approximately 2.

Complete the implementation of the two functions below:

1. `area_left_of(z)` should return the area under the armchair curve to the left of `z`, assuming `-4 <= z <= 12`, and
2. `area_between(x, y)` should return the area under the armchair curve between `x` and `y`, assuming `-4 <= x <= y <= 12`.

```
import scipy

def area_left_of(z):
    '''Returns the area under the armchair curve to the left of z.
       Assume -4 <= z <= 12'''
    if ___(a)___:
        return ___(b)___
    return scipy.stats.norm.cdf(z)

def area_between(x, y):
    '''Returns the area under the armchair curve between x and y.
    Assume -4 <= x <= y <= 12.'''
    return ___(c)___
```

a) What goes in blank (a)?

b) What goes in blank (b)?

c) What goes in blank (c)?

Question 11 (4 points)

Suppose you have correctly implemented the function `area_between(x, y)` so that it returns the area under the armchair curve between `x` and `y`, assuming the inputs satisfy `-4 <= x <= y <= 12`.

Note: You can still do this question, even if you didn't know how to do the previous one.

a) What is the approximate value of `area_between(1.79, 9.79)`?

○ 0.68
○ 0.95
○ 1
○ 1.5

b) What is the approximate value of `area_between(-2, 10)`?

○ 1.9
○ 1.95
○ 1.975
○ 2

Question 12 (~~7~~ 14 points)

IKEA is piloting a new rewards program where customers can earn free Swedish meatball plates from the in-store cafe when they purchase expensive items. Meatball plates are awarded as follows. Assume the item price is always an integer.

item price	number of meatball plates
less than 99 dollars	0
100 to 199 dollars	1
200 dollars or more	2

We want to implement a function called `calculate_meatball_plates` that takes as input an array of several item prices, corresponding to the individual items purchased in one transaction, and returns the total number of meatball plates earned in that transaction.

Select all correct ways of implementing this function from the options below.

□ Way 1:

```
def calculate_meatball_plates(prices):
    meatball_plates = 0
    for price in prices:
        if price >= 100 and price <= 199:
            meatball_plates = 1
        elif price >= 200:
            meatball_plates = 2
    return meatball_plates
```

□ Way 2:

```
def calculate_meatball_plates(prices):
    meatball_plates = 0
    for price in prices:
        if price >= 200:
            meatball_plates = meatball_plates + 1
        if price >= 100:
            meatball_plates = meatball_plates + 1
    return meatball_plates
```

□ Way 3:

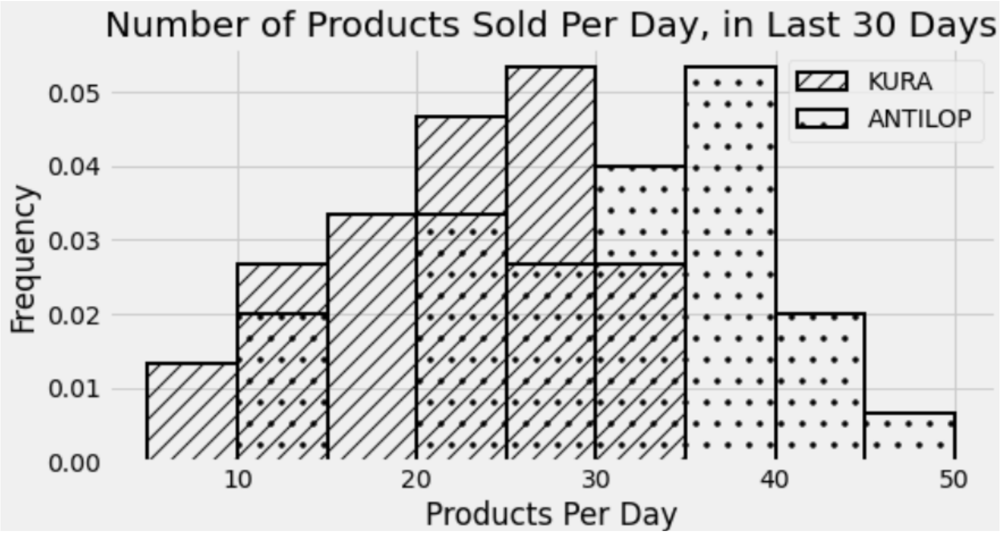
```
def calculate_meatball_plates(prices):
    return ((prices >= 200).sum()*2 +
            ((100 <= prices) & (prices <= 199)).sum()*1)
```

□ Way 4:

```
def calculate_meatball_plates(prices):
    return (np.count_nonzero(prices >= 100) +
            np.count_nonzero(prices >= 200))
```

Question 13 (6 points)

The histogram below shows the distribution of the number of products sold per day throughout the last 30 days, for two different IKEA products: the KURA reversible bed, and the ANTILOP highchair.



a) For how many days did IKEA sell more KURA reversible beds than ANTILOP highchairs? Enter an **integer** in the box or select "not enough information", but **not both**.

□ ○ not enough information

b) For how many days did IKEA sell between 20 (inclusive) and 30 (exclusive) KURA reversible beds per? Enter an **integer** in the box or select "not enough information", but **not both**.

□ ○ not enough information

c) Determine the relative order of the three quantities below.

1. The number of days for which IKEA sold at least 35 ANTILOP highchairs.
2. The number of days for which IKEA sold less than 25 ANTILOP highchairs.
3. The number of days for which IKEA sold between 10 (inclusive) and 20 (exclusive) the KURA reversible beds.

○ (1) < (2) < (3)
○ (1) < (3) < (2)
○ (2) < (1) < (3)
○ (2) < (3) < (1)
○ (3) < (1) < (2)
○ (3) < (2) < (1)

Question 14 (6 points)

The HAUGA bedroom furniture set includes two items, a bed frame and a bedside table. Suppose the amount of time it takes someone to assemble the bed frame is a random quantity drawn from the probability distribution below.

time to assemble bed frame	probability
10 minutes	0.1
20 minutes	0.4
30 minutes	0.5

Similarly, the time it takes someone to assemble the bedside table is a random quantity, independent of the time it takes them to assemble the bed frame, drawn from the probability distribution below.

time to assemble bedside table	probability
30 minutes	0.3
40 minutes	0.4
50 minutes	0.3

a) What is the probability that Stella assembles the bed frame in 10 minutes if we know it took her less than 30 minutes to assemble? Give your answer as a decimal between 0 and 1.

b) What is the probability that Ryland assembles the bedside table in 40 minutes if we know that it took him 30 minutes to assemble the bed frame? Give your answer as a decimal between 0 and 1.

c) What is the probability that Jin assembles the complete HAUGA set in at most 60 minutes? Give your answer as a decimal between 0 and 1.

Question 15 (6 points)

An IKEA employee has access to a data set of the purchase amounts for 40,000 customer transactions. This data set is roughly normally distributed with mean 150 dollars and standard deviation 25 dollars.

a) Why is the distribution of purchase amounts roughly normal?

○ because of the Central Limit Theorem

○ for some other reason

b) Shiv spends 300 dollars at IKEA. How would we describe Shiv's purchase in standard units?

○ 0 standard units

○ 2 standard units

○ 4 standard units

○ 6 standard units

c) Give the endpoints of the CLT-based 95% confidence interval for the mean IKEA purchase amount, based on this data.

Left endpoint = □ Right endpoint = □

Question 16 (2 points)

There are 52 IKEA location in the United States, and there are 50 states.

Which of the following describes how to calculate the total variation distance between the distribution of IKEA locations by state and the uniform distribution?

○ For each state, take the absolute difference between 1/50 and the number of IKEAs in that state divided by the total number of IKEA locations. Sum these values across all states and divide by two.

○ For each state, take the absolute difference between the number of IKEAs in that state and the average number of IKEAs in each state. Sum these values across all states and divide by two.

○ For each IKEA location, take the absolute difference between 1/50 and the number of IKEAs in the same state divided by the total number of IKEA locations. Sum these values across all locations and divide by two.

○ For each IKEA location, take the absolute difference between the number of IKEAs in the same state and the average number of IKEAs in each state. Sum these values across all locations and divide by two.

○ None of the above.

Question 17 (6 points)

Suppose the price of an IKEA product and the cost to have it assembled are linearly associated with a correlation of 0.8. Product prices have a mean of \$140 and a standard deviation of \$40. Assembly costs have a mean of \$80 and a standard deviation of \$10. We want to predict the assembly cost of a product based on its price using linear regression.

a) The NORDMELA 4-drawer dresser sells for \$200. How much do we predict its assembly cost to be?

□ dollars

b) The IDANÄS wardrobe sells for \$80 more than the KLIPPAN loveseat, so we expect the IDANÄS wardrobe will have a greater assembly cost than the KLIPPAN loveseat. How much do we predict the difference in assembly costs to be?

□ dollars

c) If we create a 95% prediction interval for the assembly cost of a \$100 product and another 95% prediction interval for the assembly cost of a \$120 product, which prediction interval will be wider?

○ The one for the \$100 product.

○ The one for the \$120 product.

Question 18 (9 points)

For each IKEA desk, we know the cost of producing the desk, in dollars, and the current sale price of the desk, in dollars. We want to predict sale price based on production cost using linear regression.

a) For this scenario, which of the following most likely describes the slope of the regression line when both variables are measured in dollars?

○ less than 0

○ between 0 and 1, exclusive

○ more than 1

○ none of the above (exactly equal to 0 or 1)

b) For this scenario, which of the following most likely describes the slope of the regression line when both variables are measured in standard units?

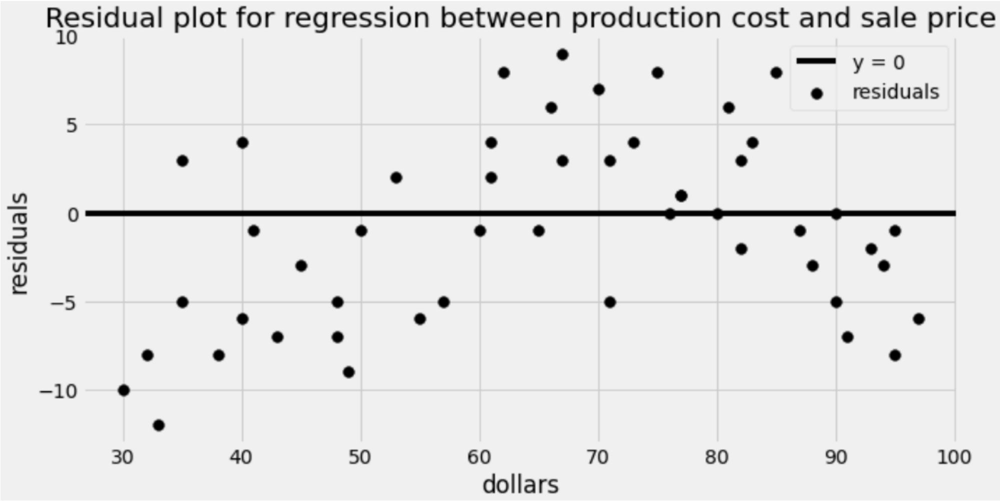
○ less than 0

○ between 0 and 1, exclusive

○ more than 1

○ none of the above (exactly equal to 0 or 1)

c) The residual plot for this regression is shown below.



What is being represented on the horizontal axis of the residual plot?

○ actual sale price
○ actual production cost
○ predicted sale price
○ predicted production cost

d) Which of the following is a correct conclusion based on this residual plot? Select all that apply.

□ We don't have enough data to do regression.
□ The regression line is not the best-fitting line for this data set.
□ The data set is not representative of the population.
□ The correlation between production cost and sale price is weak.
□ It would be better to fit a nonlinear curve.
□ Our predictions will be more accurate for some inputs than others.

Before turning in your exam, please make sure that your PID or name is on every page.

In the space below, feel free to draw some IKEA-inspired wordless instructions for how to do something! Our favorite instructions will earn a small amount of extra credit, just for fun.