DS-100 Practice Final Questions

Fall 2017

Name: ______________________________

Email address: ______________________________

Student id: ______________________________

Instructions:

- These are a random selection of previous final exam questions.
- This is not representative of the length of the final (its too long!).
- You may use a single page (two-sided) cheat sheet.

1 Loss Minimization

1. In a petri dish, yeast populations grow exponentially over time. In order to estimate the growth rate of a certain yeast, you place yeast cells in each of n petri dishes and observe the population y_i at time x_i and collect a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Because yeast populations are known to grow exponentially, you propose the following model:

$$\log(y_i) = \gamma x_i \tag{1}$$

where γ is the growth rate parameter (which you are trying to estimate). We would like to derive the L_2 regularized estimator least squares estimator.

(1) [4 Pts.] Write the *regularized least squares loss function* for γ under this model. Use λ as the regularization parameter.

(2) [8 Pts.] Solve for the optimal γ as a function of the data and λ

2. Suppose we observe a dataset $\{x_1, \ldots, x_n\}$ of independent and identically distributed samples from the exponential distribution. Suppose we give you a "probability model" parameterized by λ:

$$f_\lambda(x) = \lambda e^{-\lambda x}$$

that estimates the probability of a particular data point. In addition we give you the "log-likelihood" loss function as the following:

$$L(\lambda) = -n \log(\lambda) + \lambda \sum_{i=1}^{n} x_i \tag{10}$$

Derive the parameter value λ that minimizes this loss function. **Circle your answer.**

3. Suppose we collect a dataset of n observations $\{x_1, \ldots, x_n\}$ which we believe are drawn from a distribution with the following PDF:

$$f_\mu(x) = C \exp\left(-\frac{(x-\mu)^6}{6}\right) \tag{18}$$

where C is a constant that does not depend on μ. As before we are given the loss function:

$$L(\mu) = -n \log C + \frac{1}{6}\sum_{i=1}^{n}(x_i - \mu)^6 \tag{19}$$

(1) [4 Pts.] Compute the derivative of the derivative of the loss with respect to μ.

(2) [3 Pts.] Because there is no closed form solution for μ in $\frac{\partial}{\partial \mu} L(\mu) = 0$, we would likely use gradient *descent* to approximately compute $\hat{\mu}$. Given the gradient function:

$$g(\mu) = \frac{\partial}{\partial \mu} \log L(\mu), \tag{23}$$

and a step size $\rho(t)$, what is the gradient descent update rule to go from $\mu^{(t)}$ to $\mu^{(t+1)}$? *(Hint: your answer should contain only the variables $g(\mu^{(t)})$, $\mu^{(t)}$, $\mu^{(t+1)}$, and $\rho(t)$.)*

2 Wrangling and Querying Data

2.1 SQL

For the questions in this subsection, assume we have a massive database in the cloud with the following schema:

```
-- A simple digital media store database
CREATE TABLE media
    (mid integer PRIMARY KEY,
     name text, type char, year_released integer, length integer,
     buy_cost float, rent_cost float, avg_rating float);

CREATE TABLE customers
    (cid integer PRIMARY KEY,
     name text, joined date, nation_id integer,
     activity_level integer);

CREATE TABLE transactions
    (tid integer PRIMARY KEY,
     tdate date, item integer, customer integer,
     rent_or_buy integer, price_paid float, percent_viewed float,
     FOREIGN KEY (item) REFERENCES media,
     FOREIGN KEY (customer) REFERENCES customers);

CREATE VIEW stats AS
SELECT min(length) AS len_min, max(length) AS len_max,
       avg(length) AS len_mu, stddev(length) AS len_sigma,
       min(avg_rating) AS ar_min, max(avg_rating) AS ar_max,
       avg(avg_rating) AS ar_mu, stddev(avg_rating) AS ar_sigma
 FROM media;
```

4. [4 Pts.] In the `media` table above, the `type` column encodes the type of media as a unique character code (e.g., `'S'` for song, `'M'` for movie, `'E'` for episode, etc.). Suppose we wanted to modify the `stats` view to display the stats for each `type` of media. Which of the following are true? **(Select *all* that apply.)**

   A. We need to change the granularity of the view to be finer than it is above.

   B. We need to add a **`GROUP BY`** `type` clause to the view.

   C. It would be helpful to add `media.type` to the list of columns in the **`SELECT`** clause of the view.

   D. The modified view should have more rows than the original view above.

   E. None of the above.

5. [3 Pts.] Which of the following queries finds the ids of media that are 2 standard deviations longer than the mean length? **(Select *only one*.)**

   A.

```
SELECT media.mid
  FROM media, stats
 WHERE media.mid = stats.mid
   AND media.length >= stats.len_mu
                       + 2*(stats.len_sigma);
```

   B.

```
SELECT media.mid
  FROM media, stats
 WHERE media.length >= stats.len_mu
                       + 2*(stats.len_sigma);
```

   C.

```
SELECT media.mid
  FROM media
 WHERE media.length >= avg(media.length)
                       + 2*stddev(media.length);
```

   D. None of the above.

2.2 SQL Sampling

The `transactions` table has 30 million (30×10^6) rows. It is too large to load into the memory of our laptop. We will extract a sample from the database server to process on our laptop in Python.

```
SELECT *
  FROM transactions TABLESAMPLE Bernoulli(.0001);
```

6. [2 Pts.] In expectation, how many rows will there be in the answer to this query?

7. [4 Pts.] Your friend Emily Engineer tells you to avoid Bernoulli sampling, and use the following query instead:

```
SELECT *
  FROM transactions
 LIMIT XX;
```

(where `XX` is replaced by the correct answer to the previous question). **Select all the true statements:**

A. Emily's `LIMIT` query will probably run faster than the `TABLESAMPLE` query. For Emily's query, the database engine can simply access the first `XX` rows it finds in the table, and skip the rest.

B. Emily's query result may be biased to favor certain rows.

C. The output of the `TABLESAMPLE` query provides a hint about how many rows there are in the `transactions` table while Emily's `LIMIT` query does not.

D. Emily's `LIMIT` query may run fast, but it will swamp the memory on your laptop, since it doesn't sample the database.

E. None of the above.

8. [2 Pts.] You will recall from Homework 5 that it is possible to do bootstrap sampling in SQL by constructing a `design` table with two columns. Each of the columns used in that scheme is described by a single choice below. **Identify the *two* correct choices**:

A. A foreign key to the table being sampled.

B. A `count` column to capture the number of tuples in each bootstrap sample.

C. An identifier to group rows together into bootstrap samples.

D. A regularization column to prevent overfitting.

2.3 Pandas

For the questions in this subsection, assume that we have pandas dataframes with the same schemas as described in the previous section on SQL. That is, we have a `media` dataframe with columns `mid`, `name`, `type`, `year`, et cetera. Assume that the index column of each dataframe is meaningless—the primary key is represented as a regular column.

9. [3 Pts.] Consider the following code snippet:

```
def get_average_price_paid(join_method):
    return (customers
        .merge(transactions, how=join_method,
               left_on='cid', right_on='customer')
        .loc[:,'price_paid']
        .fillna(0)            # <- Important
        .mean()
    )

inner = get_average_price_paid('inner')
outer = get_average_price_paid('outer')
left = get_average_price_paid('left')
right = get_average_price_paid('right')
```

Assume that all item *prices are positive*, all `transactions` refer to valid customers in the `customers` table, but some customers may have no transactions.

(1) How are `inner` and `outer` related? **Pick *one* best answer.**

A. `inner` $<$ `outer`
B. `inner` $\leq$ `outer`
C. `inner` $=$ `outer`
D. `inner` $\geq$ `outer`
E. `inner` $>$ `outer`

(2) How are `left` and `right` related? **Pick *one* best answer.**

A. `left` $<$ `right`
B. `left` $\leq$ `right`
C. `left` $=$ `right`
D. `left` $\geq$ `right`
E. `left` $>$ `right`

(3) How are `left` and `outer` related? **Pick *one* best answer.**

A. `left` $<$ `outer`
B. `left` $\leq$ `outer`
C. `left` $=$ `outer`
D. `left` $\geq$ `outer`

10. [3 Pts.] We wish to write a python expression to find the largest amount of money spent by one person on any single date. We will use the following code:

```
biggie = transactions.groupby(______)['price_paid'].sum().max()
```

What should we be pass in as our `groupby` predicate? **Select *only one* answer.**

A. `'tdate'`

B. `'customer'`

C. `['item', 'tdate']`

D. `['customer', 'tdate']`

E. `['customer', 'item']`

11. [6 Pts.] Fill in the following python code that finds the names of every customer who has spent over $100.

```
merged = customers.merge(__A__, left_on=__B__, right_on=__C__)
grouped = merged.groupby(__D__).__E_()
names = grouped[__F__].index
```

12. [4 Pts.] We wish to find years where the average `price_paid` (over all time) for products released in that year is greater than the average `price_paid` across all transactions; from those years we want to return the *earliest* (smallest). We have the following code:

```
merged = transactions.merge(media, left_on="item", \
                            right_on="mid")
mean_price = merged.groupby("year_released")\
                   .mean().price_paid.mean() # Line A
by_year = merged.groupby("year_released").count() # Line B
is_greater = by_year[by_year.price_paid > mean_price] # Line C
result = is_greater.sort_index(ascending=False).index[0] # Line D
```

Some of these lines need to be modified in order for the code to work properly. We have suggested replacements for each line below. Which lines need to be *replaced*? **Select *all* that apply.**

A. `mean_price = merged.price_paid.mean()`

B. `by_year = merged.groupby("year_released").mean()`

C. `is_greater = by_year.where(by_year.price_paid > mean_price)`

D. `result = is_greater.sort_index(ascending=True).index[0]`

E. All the lines are correct.

3 Feature Engineering

For this question you were given the following sales data and asked to build a model to predict units sold based on the *the product attributes* to guide the design of future products.

ProdID	Name	Desc	Price	Category	Units Sold
13	Errorplane	*"A truly uncaught exception ... "*	404.00	Toy	9
42	Rock Kit	*"Launch into minerology with ... "*	123.45	Toys	1
54	Punative Jokes	*"Jokes that will get you fined ... "*	1.00	Books	30
		...			

13. Write down a reasonable schema for this data.

14. Suppose we are interested in building a linear predictive model. For each of the columns indicate which (one or more) of the feature transformations could be appropriate.

(1) The `ProdID` column:

A. Drop the column
B. One-Hot Encoding
C. Leave as is

(2) The `Name` column:

A. The length of the text in characters
B. One-Hot Encoding
C. Bag-of-words Encoding
D. Leave as is

(3) The `Desc` column:

A. The length of the text in characters
B. One-Hot Encoding
C. Bag-of-words Encoding
D. bi-gram Encoding
E. Leave as is

(4) The `Price` column:

A. The length of the text in characters
B. One-Hot Encoding
C. Bag-of-words Encoding
D. Convert the price to an indicator indicating if it is less than 19.99.
E. Leave as is

(5) The `Category` column:

A. The length of the text in characters.

B. One-Hot Encoding
C. Bag-of-words Encoding
D. N-Gram Encoding
E. Leave as is

15. It might be reasonable to assume that the relationship between units sold and price differs for each category (e.g., an expensive toy might be less likely to sell than expensive jewelry). Which of the following feature functions might capture this intuition?

A. $\phi(\texttt{category, price}) = \texttt{category} + \texttt{price}$
B. $\phi(\texttt{category, price}) = \texttt{price} \times \texttt{category}$
C. $\phi(\texttt{category, price}) = \textbf{OneHot}\,(\texttt{category}) + \texttt{price}$
D. $\phi(\texttt{category, price}) = \texttt{price} \times \textbf{OneHot}\,(\texttt{category})$
E. $\phi(\texttt{category, price}) = \texttt{category} \times \textbf{OneHot}\,(\texttt{price})$
F. $\phi(\texttt{category, price}) = \textbf{Concatenate}(\textbf{OneHot}(\texttt{category}), \texttt{price})$

4 Feature Engineering 2

For this problem we collected the following data on the new social networking app *UFace*.

PostID	UTC Time	Text	Num. Responses	State
3	08:10 PM	*"Checkout my breakfast . . . "*	2	VA
13	11:00 AM	*"Studied all night for . . . "*	5	CA
14	12:04 PM	*"Hello world!"*	0	NY
17	11:35 PM	*"That exam was lit . . . "*	42	CA
		. . .		

16. Suppose we are interested in predicting the number of responses *for future posts*. For each of the columns, indicate which **(one or more)** of the given feature transformations could be informative. **Select *all* that apply.**

(1) [2 Pts.] The `PostID` column:

A. Drop the column
B. One-Hot encoding
C. Leave as is

(2) [2 Pts.] The `Time` column:

A. Take the hour as a float
B. One-Hot encoding
C. Bag-of-words encoding
D. Time since midnight in seconds

(3) [2 Pts.] The `Text` column:

A. The length of the text
B. One-Hot encoding
C. Bag-of-words encoding
D. Leave as is

(4) [2 Pts.] The `State` column:

A. The length of the text
B. One-Hot encoding
C. Bag-of-words encoding
D. Leave as is

17. [4 Pts.] Suppose we believe that people are more likely to respond to tweets in the *afternoon* (roughly from hours 13 to 17). Which of the following feature functions would help capture this intuition? Assume that the function **localHour** takes a time and a state as its arguments and returns the hour of the day (in 24-hour time) in the state's time zone. Also assume that any boolean-valued feature is encoded as 0 (false) or 1 (true). **Select *all* that apply.**

    A. $\phi(\texttt{time, state}) = \textbf{localHour}(\texttt{time, state})$

    B. $\phi(\texttt{time, state}) = 13 < \textbf{localHour}(\texttt{time, state}) < 17$

    C. $\phi(\texttt{time, state}) = \exp\left(-\left(\textbf{localHour}(\texttt{time, state}) - 15\right)^2\right)$

    D. $\phi(\texttt{time, state}) = \exp\left(\textbf{localHour}(\texttt{time, state}) - 15\right)$

    E. None of the above.

18. [2 Pts.] Given the following text from a BigData Borat post:

    "Data Science is statistics on a Mac."

    Which of the following is the *bi-gram* encoding *including stop-words*? **(Select *only one.*)**

    A. {('data', 1), ('science', 1), ('statistics', 1), ('mac', 1)}

    B. {('data science', 1), ('science statistics', 1), ('statistics mac', 1)}

    C. { ('data science', 1), ('science is', 1), ('is statistics', 1), ('statistics on', 1), ('on a', 1), ('a mac', 1)}

    D. {('data science', 1), ('is statistics', 1), ('on a', 1), ('mac', 1)}

5 Least Squares Regression and Regularization

19. **Binary Features** You are part of a team that is analyzing data from a clinical trial. Let X be a full-column-rank $n \times 2$ design matrix with the following columns:

    1. X_0 is a column of 1s. This generates an intercept term.
    2. X_1 is a binary treatment indicator vector taking on values 0 or 1. $X_{i1} = 1$ means that y_i represents the response of a treated individual. $X_{i1} = 0$ means that y_i represents the response of an untreated individual

    You propose the following linear model:

    $$y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

    You solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ (estimates of β_0 and β_1, respectively) by minimizing the residual sum of squares. Show that

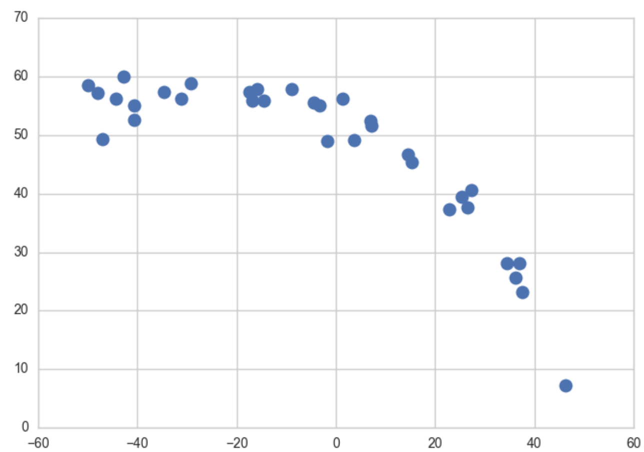
    $$\hat{\beta}_1 = \bar{y}_T - \bar{y}_C,$$

    where:

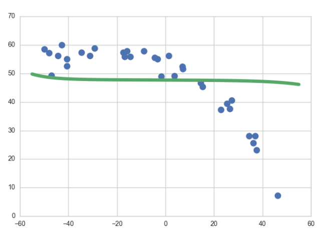
    - $\bar{y}_T$ is the average response of all the treated individuals, and
    - $\bar{y}_C$ is the average response over all untreated or “control” individuals.

    Hint: Think about the meaning of $\sum_{i=1}^{n} X_{i1}$ and how $\bar{y}$ is related to $\bar{y}_T$ and $\bar{y}_C$

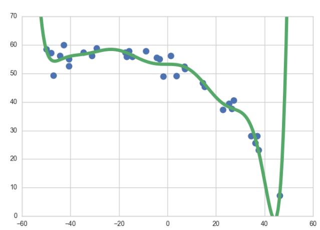
20. For this question we use the following toy dataset:



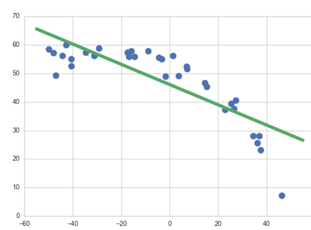
(1) [3 Pts.] We have fit several models depicted as curves in the following plots:



(a)



(b)

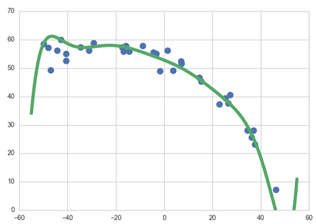


(c)

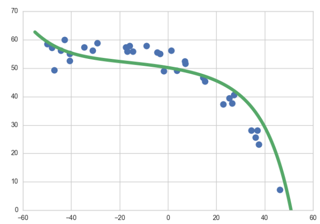
Select the plot that best matches each of the models below. **Each plot is used exactly once.**

1. Linear regression model
   ◯ (A) ◯ (B) ◯ (C)
2. Linear regression with degree 10 polynomial features
   ◯ (A) ◯ (B) ◯ (C)
3. Ridge regression with degree 10 polynomial features and substantial regularization.
   ◯ (A) ◯ (B) ◯ (C)

(2) [2 Pts.] We fit two more models to these data. Again, the solid curves display the predictions made by each model.



(a)



(b)

Select the plot that best matches each of the models below. **Each plot is used exactly once.**

1. Ridge regression with degree 10 polynomial features, $\lambda = 0.1$.
   ◯ (A) ◯ (B)
2. Ridge regression with degree 10 polynomial features, $\lambda = 1.0$.
   ◯ (A) ◯ (B)

21. Suppose you are given a dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}$ is a one dimensional feature and $y_i \in \mathbb{R}$ is a real-valued response. To model this data you choose a model characterized by the following objective function:

$$J(\theta) = \sum_{i=1}^{n} \left(y_i - \theta_0 - x_i\theta_1 - x_i^2\theta_2\right)^2 + \lambda \sum_{i=1}^{2} |\theta_i| \tag{32}$$

(1) [7 Pts.] **Select *all* the true statements** for the above objective function (Equation 32).

A. This loss function likely corresponds to a classification problem.
B. θ is the regularization parameter.
C. This is an example of L_1 regularization.
D. This is not a linear model in θ.
E. This model includes a bias/intercept term.
F. This model incorporates a non-linear feature transformation.
G. Large values of λ would reduce the model to a constant θ_0.
H. None of the above are true.

(2) [2 Pts.] Suppose in our implementation we accidentally forget to square the first term:

$$J(\theta) = \sum_{i=1}^{n} \left(y_i - \theta_0 - x_i\theta_1 - x_i^2\theta_2 \right) + \lambda \sum_{i=1}^{2} |\theta_i| \tag{33}$$

What would change if we tried to train a model using gradient descent on this objective function rather than the original objective function? **(Select only one)**

A. The training code would raise an error due to a matrix/vector dimension problem.

B. The training process would diverge with $\theta_0 \to -\infty$

C. The training process would diverge with $\theta_0 \to \infty$

D. The training process would converge to a different regression line.

E. Nothing; the training process would eventually converge to the same regression line.

22. [5 Pts.] Let X be a $n \times p$ design matrix with full column rank and y be a $n \times 1$ response vector. Let $\hat{\beta}$ be the optimal solution to the least squares problem and r be its associated error. In other words,

$$y = X\hat{\beta} + r \tag{34}$$

Consider X_2 the second column of X.

(1) [1 Pt.] **True or False.** Without any additional assumptions,

$$r \cdot X_2 = 0$$

where $\cdot$ denotes the usual dot product?

(2) [4 Pts.] Provide a short proof or counter example.

You may use the following scratch space but we will only grade what you put on the answer sheet.

6 Classification

23. For each of the following circle **T** for true or **F** for false.

    (1) [1 Pt.] Binary or multi-class **classification** techniques are most appropriate when making predictions about **continuous responses**.

    (2) [1 Pt.] In a setting with extreme class imbalance in which 95% of the training data have the same label it is always possible to get at least 95% **training accuracy**.

    (3) [1 Pt.] In a setting with extreme class imbalance in which 95% of the training data have the same label it is always possible to get at least 95% **test accuracy**.

    (4) [1 Pt.] In logistic regression, predictor variables (X) are continuous, with values from 0 to 1.

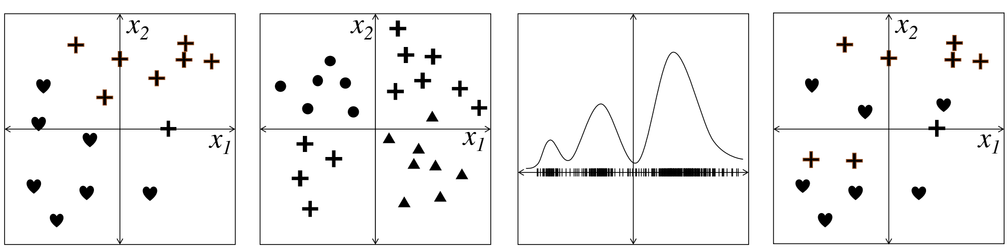
    (5) [1 Pt.] In two-class logistic regression, the response variable (y) is continuous, with values from 0 to 1.

    (6) [1 Pt.] In logistic regression, the outputs of the sigmoid function are continuous, with values from 0 to 1.

    (7) [1 Pt.] In two-class logistic regression, the output of the sigmoid function for each data point represents the category of that data point.

    (8) [1 Pt.] In logistic regression, we calculate the weights $\hat{\theta}$ as $\hat{\theta} = \left(X^T X\right)^{-1} X^T y$, and then fit responses as $\hat{y}_i = \sigma\left(x_i^T \hat{\theta}\right)$.

24. Using the following figure to answer each of the following questions:



(A) (B) (C) (D)

    (1) Which of the above plots represents a **linearly separable binary classification** task?

    ◯ (A) ◯ (B) ◯ (C) ◯ (D)

    (2) Which of the above plots represents a **binary classification** task that is **not linearly separable**?

    ◯ (A) ◯ (B) ◯ (C) ◯ (D)

    (3) Which of the above plots represents a **multi-class classification task**?

    ◯ (A) ◯ (B) ◯ (C) ◯ (D)

    (4) Which of the above plots depicts a **1-dimensional Gaussian mixture model**?

    ◯ (A) ◯ (B) ◯ (C) ◯ (D)

25. Consider the following buggy Python implementation of gradient descent.

```
1  def grad_descent(X, Y, theta0,
2                   grad_function, max_iter = 1000000):
3      """X: A 2D array, the feature matrix.
4      Y: A 1D array, the response vector.
5      theta0: A 1D array, the initial parameter vector.
6      grad_function: Maps a parameter vector, a feature matrix, and
7        a response vector to the gradient of some loss function at
8        the given parameter value.  The return value is a 1D array."""
9      theta = theta0
10     for t in range(1, max_iter+1):
11         grad = grad_function(theta, X, Y)
12         theta = theta0 + 1/t * grad
13     return theta
```

Select all the issues with this Python implementation

A. **Line 11** `theta` should be replaced by `theta0`.

B. **Line 12** `theta0` should be replaced by `theta`.

C. **Line 12** `1/t` should be replaced by `t`.

D. **Line 12** + should be replaced by −.

26. Suppose we collect a binary classification dataset consisting of $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$. Recall that the the probability mass function for a Bernoulli random variable $y \in \{0, 1\}$ is:

$$\mathbf{P}(y \,|\, \theta) = \theta^y (1-\theta)^{(y-1)} \tag{35}$$

and the sigmoid function is given by:

$$\sigma(t) = \frac{1}{1+e^{-t}} \tag{36}$$

Which of the following is the **loss function** for the logistic regression model with L_2 regularization?

A. $J(\theta) = \sum_{i=1}^n \theta_i^y (1-\theta)^{(y_i-1)} + \lambda|\theta|$

B. $J(\theta) = -\sum_{i=1}^n (\theta x_i)^y (1-\theta x_i)^{(y_i-1)} + \lambda\theta^2$

C. $J(\theta) = \sum_{i=1}^n \sigma(\theta x_i)^y \left(1-\sigma(\theta x_i)\right)^{(y_i-1)}$

D. $J(\theta) = \lambda\theta^2 - \sum_{i=1}^n \left[y_i \log \sigma(\theta x_i) + (y_i - 1)\log\left(1-\sigma(\theta x_i)\right)\right]$

E. $J(\theta) = \lambda\theta^2 + \sum_{i=1}^n \left[y_i \log \sigma(\theta x_i) + (y_i - 1)\log\left(1-\sigma(\theta x_i)\right)\right]$

27. [4 Pts.] Which of the following can help deal with overfitting in a logistic regression model?

A. Adding additional features.

B. Obtaining additional training data.

C. Performing regularization.

D. Removing data until your classes are linearly separable.

7 Classification 2

28. For each of the following select **T** for true or **F** for false on the answer sheet.

    (1) [1 Pt.] A binary or multi-class **classification** technique should be used whenever there are **categorical features**.

    (2) [1 Pt.] Logistic regression is actually used for classification.

    (3) [1 Pt.] The logistic regression loss function was derived by modeling the observations as noisy observations with a Gaussian noise model.

    (4) [1 Pt.] Class imbalance can be a serious problem in which the number of training data points from one class is much larger than another.

    (5) [1 Pt.] A broken *binary* classifier that *always* predicts 0 is likely to get a test accuracy around 50% on all prediction tasks.

    (6) [1 Pt.] The root mean squared error is the correct metric for evaluating the prediction accuracy of a binary classifier.

29. Consider the following binary classification dataset

    (1) [3 Pts.] Draw a reasonable approximation of the logistic regression probability estimates for $\mathbf{P}(Y = 1 \mid x)$ on top of the figure on the answersheet.

    (2) [1 Pt.] Are these data linearly separable?

        A. Yes

        B. No

30. [3 Pts.] Suppose you are given θ for the logistic regression model to predict whether a tumor is malignant ($y = 1$) or benign ($y = 0$) based on features of the tumor x. If you get a new patient x_* and find that $x_*^T \theta > 0$, what can you say about the tumor? **Select *only one.***

    A. The tumor is benign

    B. The tumor is more likely benign

    C. The tumor is more likely to be malignant

    D. The tumor is malignant

31. [4 Pts.] Which of the following explanations that applying regularization to a logistic regression model? **Select *all* that apply.**

    A. The training error is too high.

    B. The test error is too low.

    C. The data are high-dimensional.

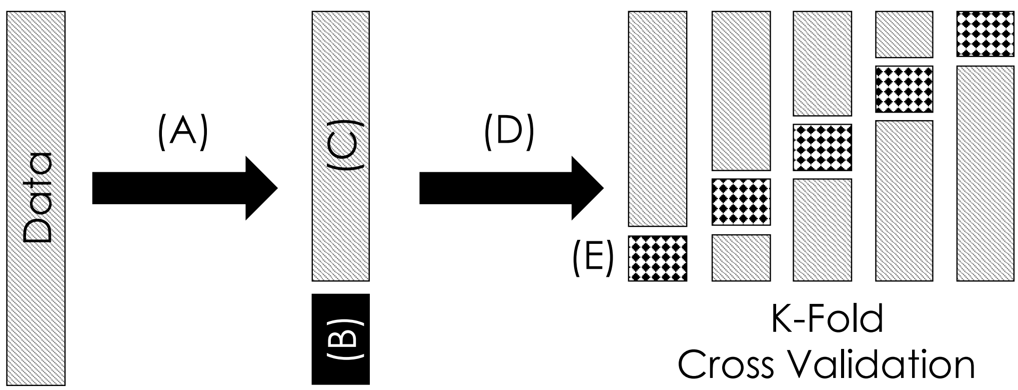
    D. There is a large class imbalance.

    E. None of the above justify regularization for logistic regression.

8 Bias-Variance Tradeoff

32. For each of the following circle **T** for true or **F** for false.
    (1) Increasing the regularization penalty decreases bias.
    (2) Without taking precautions, reducing bias often leads to increased variance.
    (3) In the bias variance trade-off the variance refers to the variability in predictions across different training datasets.
    (4) As we improve the model to reduce bias we often run the risk of **under-fitting**.

33. In this question we complete the following figure describing the train-test split and k-fold cross validation. Note the data table with many records and a few columns is depicted on the left as a tall rectangle.



    (1) This part of the figure refers to the **validation** data.
    ◯ (A) ◯ (B) ◯ (C) ◯ (D) ◯ (E)
    (2) This part of the figure refers to the **testing** data
    ◯ (A) ◯ (B) ◯ (C) ◯ (D) ◯ (E)
    (3) This part of the figure refers to the process of constructing the *train-test* split.
    ◯ (A) ◯ (B) ◯ (C) ◯ (D) ◯ (E)
    (4) Select all the following statements that apply to the above figure.
        A. This figure illustrates 5-fold cross validation.
        B. This figure illustrates 6-fold cross-validation.
        C. Assuming all the data points are distinct each of the validation data sets are also distinct.
        D. The test data should be used during cross-validation to fully evaluate the model.

34. For each of the following select **T** for true or **F** for false on the answer sheet.

(1) [1 Pt.] Regularization can be used to manage the bias-variance trade-off.

(2) [1 Pt.] When conducting linear regression, adding polynomial features to your data often decreases the variance of your fitted model.

(3) [1 Pt.] When conducting linear regression, adding polynomial features to your data often decreases the bias of your fitted model.

(4) [1 Pt.] Suppose your data are an i.i.d. sample from a population. Then collecting a larger sample for use as a *training set* can help reduce *bias*.

(5) [1 Pt.] Suppose your data are an i.i.d. sample from a population. Then collecting a larger sample for use as a *training set* can help reduce *variance*.

(6) [1 Pt.] Training error is typically larger than test error.

(7) [1 Pt.] If you include the test set in your training data, your accuracy as measured on the test set will probably increase.

(8) [1 Pt.] It is important to frequently evaluate models on the test data throughout the process of model development.

35. [2 Pts.] A colleague has been developing models all quarter and noticed recently that her *test* error has started to gradually increase while her training error *has been decreasing*. Which of the following is the most likely explanation for what is happening? **Select *only one*.**

A. She is starting to over-fit to her training data.

B. She is starting to under-fit to her training data.

C. The model is overly biased.

D. None of the above.

36. [5 Pts.] Given the following general loss formulation:

$$\arg\min_{\theta} \left[\sum_{i=1}^{n} \left(y_i - x_i^T \theta \right)^2 + \lambda \sum_{p=1}^{d} \theta_p^2 \right] \quad (37)$$

Which of the following statements are true? **Select *all* that apply.**

A. There are d data points.

B. There are n data points.

C. The data is d dimensional.

D. This is a classification problem.

E. This is a linear model.

F. This problem has LASSO regularization.

G. Larger values of λ imply increased regularization.

H. Larger values of λ will increase variance.

I. Larger values of λ will likely increase bias.

J. None of the above are true.

37. [3 Pts.] In class we broke the least-squares error into three separate terms:

$$\mathbf{E}\left[(y - f_\theta(x))^2\right] = \mathbf{E}\left[(y - h(x))^2\right] + \mathbf{E}\left[(h(x) - f_\theta(x))^2\right] + \mathbf{E}\left[(f_\theta(x) - \mathbf{E}\left[f_\theta(x)\right])^2\right] \quad (38)$$

where $y = h(x) + \epsilon$, $h(x)$ is the true model and ϵ is zero-mean noise. For each of the following terms, indicate its usual interpretation in the bias variance trade-off:

1. $\mathbf{E}\left[(y - h(x))^2\right]$: A. Bias B. Variance C. Noise
2. $\mathbf{E}\left[(h(x) - f_\theta(x))^2\right]$: A. Bias B. Variance C. Noise
3. $\mathbf{E}\left[(f_\theta(x) - \mathbf{E}\left[f_\theta(x)\right])^2\right]$: A. Bias B. Variance C. Noise

9 Big Data

38. Which of the following are true:
    - A. Star schemas are designed to decrease redundancy.
    - B. A Data Warehouse is typically updated every time a change occurs in a related Operational Data Store.
    - C. A typical Data Warehouse favors cleanliness over completeness: it rejects data that does not conform to the warehouse schema.
    - D. A typical Data Lake favors completeness over cleanliness: it allows you to store any data you like, without even requiring a schema.
    - E. The "T" in ETL involves many of the same tasks as Data Wrangling.
39. Consider a data warehouse of automobile sensor readings, which records information on sensors, readings, and vehicles where the sensors are placed. Which of the following are true:
    - A. Because a traditional ETL process only loads data into the warehouse periodically, it will lose sensor information recorded in the operational data store.
    - B. Each sensor reading should be timestamped in the data warehouse.
    - C. There is no reason for the data warehouse to record timestamps for information on the vehicles.
40. Which of the following features are typical of a distributed file system:
    - A. It can store large volumes of data.
    - B. It is optimized to store data as compactly as possible.
    - C. It can keep serving files even after a certain number of machine failures.
    - D. After a crash, if any data can be recovered at all, then all the data can be recovered.
41. In class, we asserted that MapReduce is being used less and less in practice. It is being replaced by what other programming interfaces? Why?
42. Which of the following are true?
    - A. Because people can store any file in a Data Lake, it is harder to assess data quality in a Lake than in a Warehouse.
    - B. The raw data in a Data Lake will likely require more wrangling than the data in a well-governed Data Warehouse.
    - C. The lack of a unifying schema in a Data Lake makes it difficult to get a global view of information being captured.
    - D. Relative to traditional Data Warehouses, Data Lakes make it easier to secure data in a well-governed way.

43. Consider the following simple Data Warehouse schema from a Cellular Service Provider, which records activity on a cell phone network:

```
CREATE TABLE devices (
  did integer, customer_id integer,
  phone_number varchar(13),
  firstname text, lastname text,
  zip varchar(12), registered_on varchar(2),
  PRIMARY KEY(did),
  UNIQUE (customer_id) -- a ''candidate'' key
  );

CREATE TABLE billing (
  rate_code char PRIMARY KEY,
  description text, base_fee float, per_minute float,
  max_minutes integer,  overage_fee float,
  PRIMARY KEY (rate_code));

CREATE TABLE calls (
  caller_handset_id integer, callee_handset_id integer,
  cell_tower_id integer, call_start datetime, call_end datetime,
  billing_code char,
  PRIMARY KEY (caller_handset_id, call_start),
  FOREIGN KEY (caller_handset_id) REFERENCES devices,
  FOREIGN KEY (billing_code) REFERENCES billing;
```

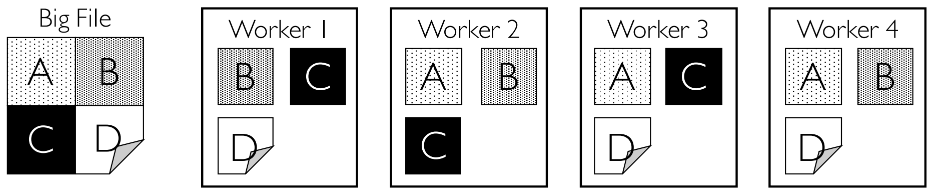
(1) [3 Pts.] Which of these tables is a dimension table? **Select *all* that apply.**

A. `devices`
B. `calls`
C. `billing`
D. None of the above.

(2) [3 Pts.] Which of the following statements are true? **Select *all* that apply.**

A. The `calls.billing_code` column violates star schema design because any update to a single billing fee requires updates to many call records.
B. If we want to look for correlations between a device's average call length and the time since it was registered, we have to perform a join.
C. If the cell service provider implemented a Data Lake, it would make it easier for them to load audio recordings of calls for subsequent analysis.
D. None of the above statements are true.

44. [3 Pts.] The figure below depicts a distributed file system with one logical "big file" partitioned into 4 "shards" (A, B, C, D) and replicated across multiple worker machines (1, 2, 3, 4).



Suppose workers 1 AND 2 both fail. Which of the following statements are true? **Select *all* that apply.**

A. The full file will remain available since worker 3 and worker 4 are both still running.

B. The system can tolerate one more worker failure without losing data.

C. If every request requires all 4 shards of the file, then worker 3 and worker 4 can share the work evenly.

D. None of the above statements are true.

45. Consider only the mechanism of *partitioning* files into shards, and storing different shards on different machines. Which of the following statements are true? **Select *all* that apply.**

A. Partitioning enhances the ability of the system to store large files.

B. Partitioning allows the system to tolerate machine failures without losing data.

C. Partitioning allows the system to read files in parallel.

D. None of the above statements are true.

46. [2 Pts.] Recall the statistical query pattern discussed in class for computing on very large data sets. Which of the following statements are true? **Select *all* that apply.**

A. It eliminates the need for the end-user device (e.g. a laptop) to acquire all the data.

B. It pushes the computational task closer to the large-scale data storage.

C. It is well suited to both MapReduce and SQL interfaces.

D. An alternative to the statistical query pattern for big data is to acquire a sample of the full dataset on the end-user device.

E. None of the above statements are true.

10 EDA and Visualization

47. [2 Pts.] Consider the following statistics for infant mortality rate. According to these statistics, which transformation would best symmetrize the distribution? **(Select only one.)**

Transformation	lower quartile	median	upper quartile
x	13	30	68
$\sqrt{x}$	3.5	5	8
$\log(x)$	1.15	1.5	1.8

A. no transformation

B. square root

C. log

D. not possible to tell with this information

48. [5 Pts.] For each of the following scenarios, determine which plot type is *most* appropriate to reveal the distribution of and/or the relationships between the following variable(s). **For each scenario, select only one plot type. Some plot types may be used multiple times.**

A. histogram

B. pie chart

C. bar plot

D. line plot

E. side-by-side boxplots

F. scatter plot

G. stacked bar plot

H. overlaid line plots

I. mosaic plot

(1) [1 Pt.] sale price and number of bedrooms (assume integer) for houses sold in Berkeley in 2010.

(2) [1 Pt.] sale price and date of sale for houses sold in Berkeley between 1995 and 2015.

(3) [1 Pt.] infant birth weight (grams) for babies born at Alta Bates hospital in 2016.

(4) [1 Pt.] mother's education-level (highest degree held) for students admitted to UC Berkeley in 2016

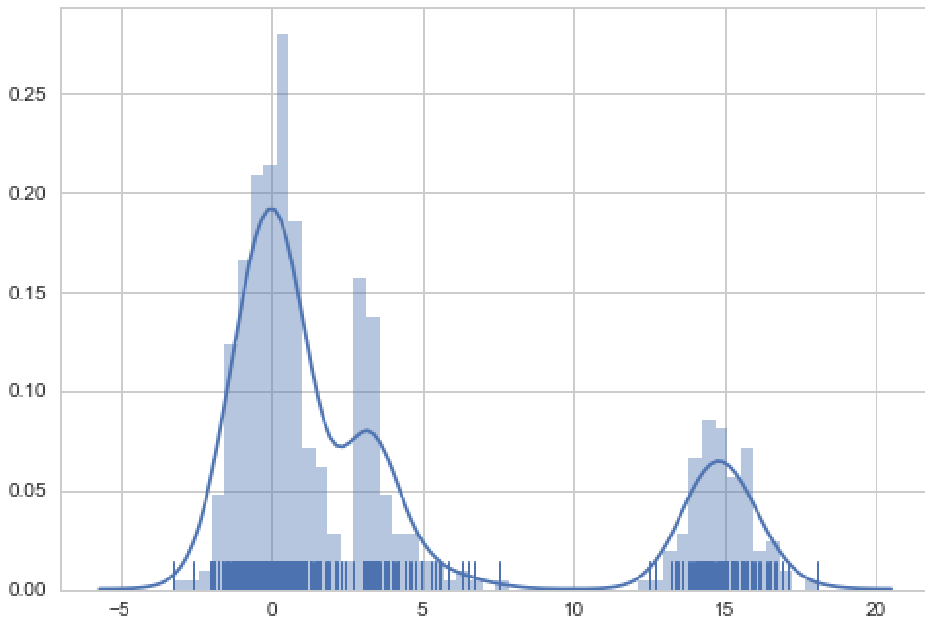
(5) [1 Pt.] SAT score and HS GPA of students admitted to UC Berkeley in 2016

(6) [1 Pt.] race and gender of students admitted to UC Berkeley in 2016

(7) [1 Pt.] The percentage of female student admitted to UC Berkeley each year from 1950 to 2000.

(8) [1 Pt.] SAT score for males and females of students admitted to UCB from 1950 to 2000

49. [4 Pts.] Consider the following empirical distribution:



(1) [1 Pt.] The distribution has ____________ mode(s).

A. 1 B. 2 C. 3 D. 4

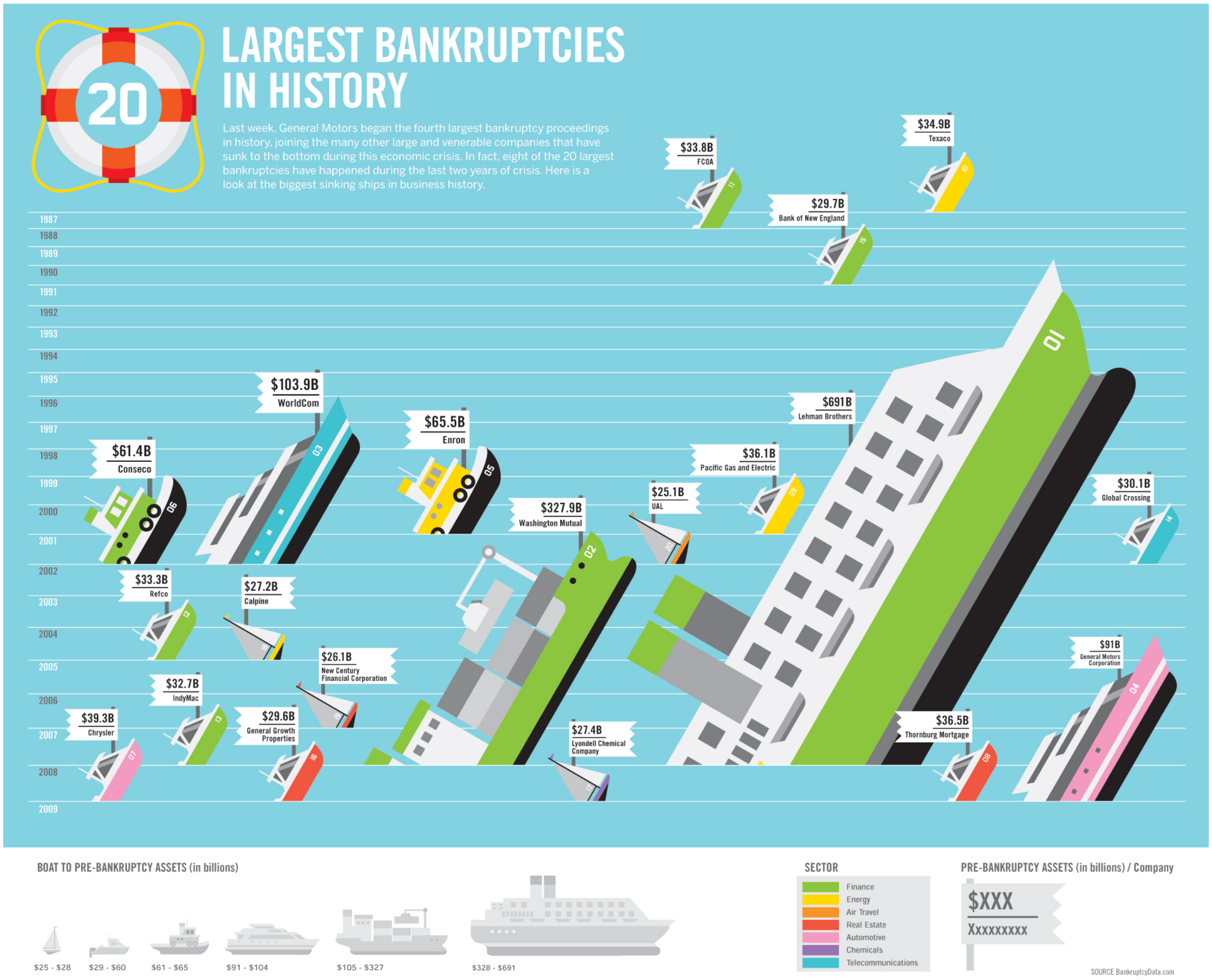
(2) [1 Pt.] The distribution is:

A. Skewed left
B. Symmetric
C. Skewed right

(3) [2 Pts.] Select **all** of the following properties displayed by the distribution:

A. gaps
B. outliers
C. normal left tail
D. None of the above

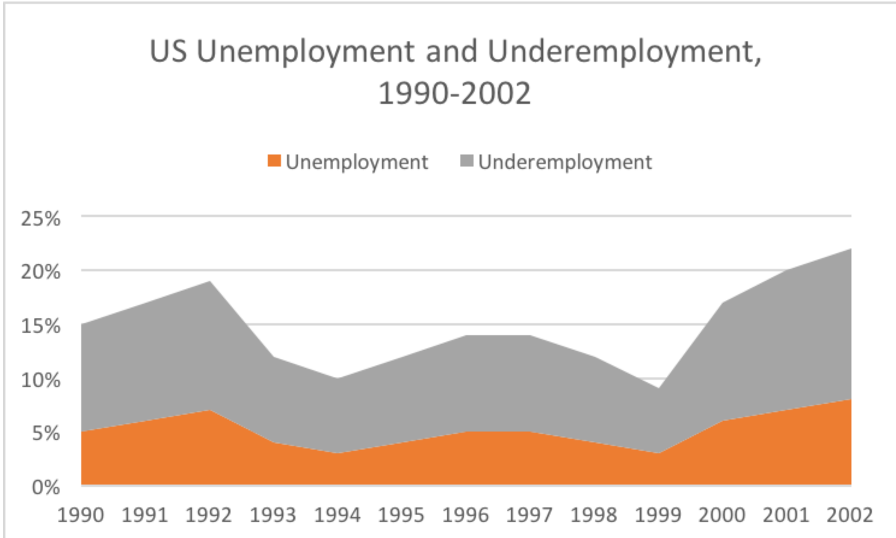
50. [4 Pts.] Select all of the problems associated with the following plot (there may be more than one problem):



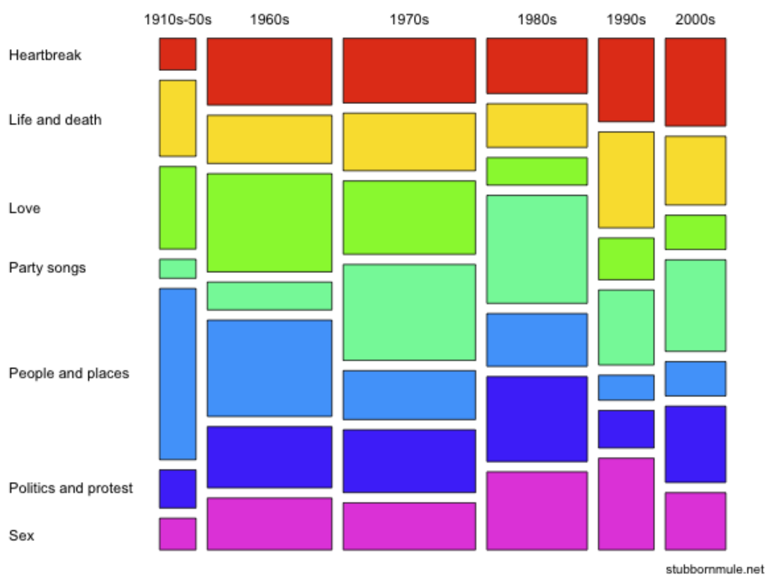
A. Over-plotting
B. Use of chart junk
C. Vertical axis should be in log scale
D. Missing vertical axis label
E. Poor use of the horizontal dimension
F. Graph elements interfere with data
G. Stacking
H. Use of angles to convey information
I. None of the above are problems with this awesome plot.

51. In the odd questions, name the plot's type (for example, "scatter" or "box"). In the even questions, answer whether the plot is useful for answering the given query.

(1) [1 Pt.]



(2) [1 Pt.] *True* or *false*: This plot is useful for answering the question: "Did underemployment generally increase when unemployment increased?"

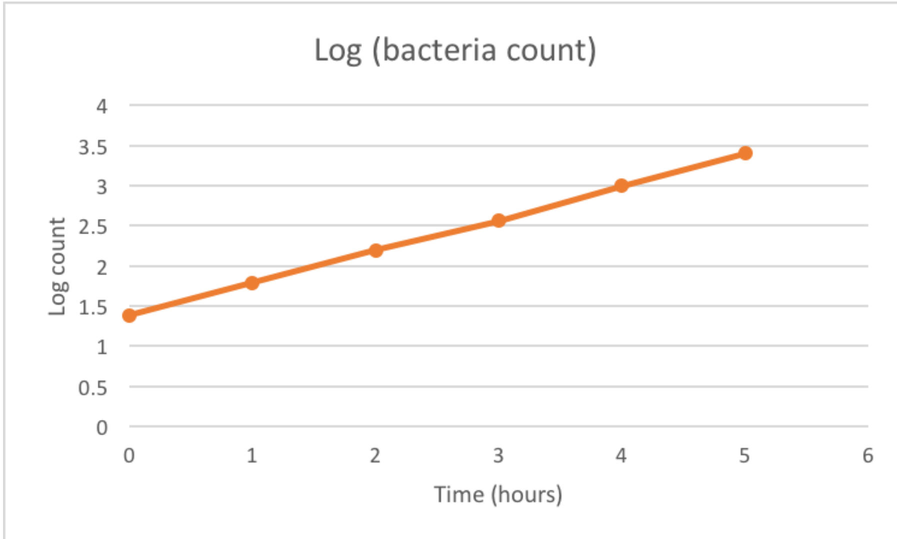


(3) [1 Pt.]

(The plot is from `http://www.stubbornmule.net/2009/07/love-old-fashioned/` and displays topics of selected popular music over time. Not all pop songs are represented in the dataset.)

(4) [1 Pt.] *True* or *false*: This plot is useful for answering the question: "Among the songs in this dataset, how many were released in each of the five decades from 1960 to 2010?"

(5) [1 Pt.]



(6) [1 Pt.] *True* or *false*: This plot is useful for answering the question: “Assuming the bacteria population grew linearly over time, what was the rate of increase?”

End of Exam