DS-100 Final Exam, Version A

Fall 2018

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Instructions:

- This final exam must be completed in the **170 minute** time period ending at **2:30 PM**, unless you have accommodations supported by a DSP letter.
- Note that some questions have bubbles to select a choice. This means that you should only **select one choice**. Other questions have boxes. This means you should **select all that apply**.
- When selecting your choices, you must **fully shade** in the box/circle. Check marks will likely be mis-graded. We reserve the right to deny regrade requests if an answer choice is not completely filled in.
- Write **clearly and legibly** when filling in free response questions.
- You may use a two-sheet (each two-sided) study guide.

Honor Code:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam and I completed this exam in accordance with the honor code.

Signature: _____

Syntax Reference

Regular Expressions

"|" matches expression on either side of symbol. Has lowest priority.

"\" match the following character literally.

"?" match preceding literal or sub-expression 0 or 1 times.

"+" match preceding literal or sub-expression *one* or more times.

"*" match preceding literal or sub-expression *zero* or more times.

"." match any character except new line.

"[ ]" match any one of the characters inside, accepts a range, e.g., "[a-c]". All characters inside treated literally.

"( )" used to create a sub-expression.

"{n}" preceding expression repeated n times.

Some useful Python functions and syntax

re.findall(pattern, st) returns the list of all non-overlapping sub-strings in st that match pattern.

np.random.choice(a, replace, size) Generates a random sample from a consisting of size values (with replacement if replace=True). a can be 1-D array-like or int.

Useful Pandas Syntax

```
df.loc[row_selection, col_list]  # row selection can be boolean
df.iloc[row_selection, col_list] # row selection can be boolean
pd.get_dummies(data) # Convert categorical variable into indicator values
df.groupby(group_columns)[['colA', 'colB']].agg(agg_func)
df.groupby(group_columns)[['colA', 'colB']].filter(filter_func)
```

Variance and Expected Value

The expected value of X is $\mathbb{E}[X] = \sum_{j=1}^{m} x_j p_j$. The variance of X is $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. The standard deviation of X is $SD[X] = \sqrt{Var[X]}$.

Misc

For calculations involving percentiles of collections of numbers, we will use the following convention from Data 8: Let p be a number between 0 and 100. The p^{th} percentile is the smallest number in the collection that is at least as large as $p\%$ of all the values.

The logistic equation is $\sigma(x) = \frac{1}{1+\exp(-x)}$ and the KL divergence for two distributions is $D(P||Q) = \sum_{k=1}^{K} P(k) \log(P(k)/Q(k))$

Score Breakdown

Page	Points
4	8
5	3
6	6
7	12
8	8
9	12
11	12
12	7
13	9
14	4
15	4
16	9
17	3.5
18	4
19	11.5
20	5
21	11
22	7
23	2
24	8
25	6
26	6
27	11
28	15
30	7
31	9
Total:	200

Tabular Data

1. For this section, we will be working with the UC Berkeley Undergraduate Career Survey dataset. Each year, the UC Berkeley career center surveys graduating seniors for their plans after graduating. Below is a sample of the full dataset. The full dataset contains many thousands of rows.

j_name	c_name	c_location	m_name
Llama Technician	Google	MOUNTAIN VIEW	EECS
Software Engineer	Salesforce	SF	EECS
Open Source Maintainer	Github	SF	Computer Science
Big Data Engineer	Microsoft	REDMOND	Data Science
Data Analyst	Startup	BERKELEY	Data Science
Analyst Intern	Google	SF	Philosophy

Table 1: `survey` Table

Each record of the `survey` table is an entry corresponding to a student. We have the student's major information (`m_name`), company information (`c_name, c_location`), and the job title (`j_name`).

(a) [3 Pts] Write a SQL query that selects all data science major graduates that got jobs in Berkeley. The result generated by your query should include all 4 columns.

```
_____ FROM survey
_____
_____
```

Solution:

```
SELECT * FROM survey
WHERE m_name = 'Data Science'
AND c_location = 'Berkeley'
```

(b) [5 Pts] Write a SQL query to find the top 5 popular companies that data science graduates will work at, from most popular to 5th most popular.

```
SELECT c_name, _____ as count
FROM survey
WHERE _____ = 'Data Science'
GROUP BY _____
ORDER BY _____
LIMIT 5
```

Solution:

```
SELECT c_name, COUNT(*) AS count
FROM survey
WHERE m_name = 'Data Science'
```

```
GROUP BY c_name
ORDER BY count DESC
LIMIT 5;
```

(c) [3 Pts] Suppose our table has 9,000 rows, with 3,000 unique job names, 1,700 unique company names, 817 unique locations, and 105 unique major names. The table above has many redundancies. Suppose we wanted to instead use the star schema idea from lecture, where we have one fact table and many dimension tables. How many dimension tables would we end up with? How many rows would there be in our fact table? How many columns would there be in our fact table? There may be more than one correct answer.

i. Number of dimension tables: **4**

ii. Number of rows in fact table: **9000**

iii. Number of columns in fact table: **4**

(d) [3 Pts] Consider the pandas expression below, where `nunique` returns the number of unique elements in a group.

`survey.groupby('c_name')['m_name'].nunique().max().`

What does it return?

◯ **A. One value: The number of unique majors for the company with the most unique majors.**

◯ B. One value: The number of unique companies for the major with the most hires.

◯ C. Many values: For each company, the count of the number of hires for the most popular major.

◯ D. Many values: For each major, the count of the number of hires by the most popular company.

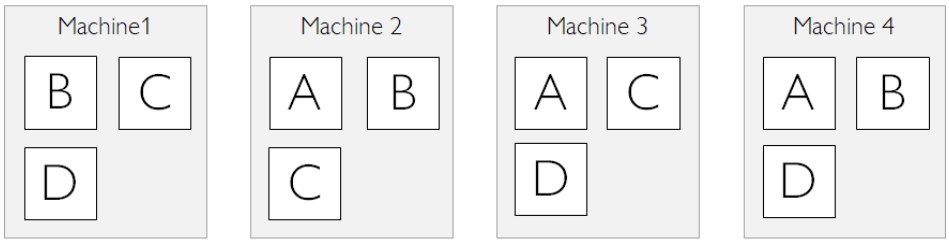
(e) [3 Pts] Which of the SQL expressions below is equivalent to the pandas code from above?

◯ **A.**
```
SELECT MAX(count)
FROM (
  SELECT c_name, COUNT(DISTINCT m_name) AS count
  FROM survey
  GROUP BY c_name
  );
```

◯ B.
```
SELECT c_name, MAX(COUNT(DISTINCT m_name)) AS count
FROM survey
GROUP BY c_name;
```

◯ C.
```
SELECT c_name, COUNT(DISTINCT m_name) AS count
FROM survey
GROUP BY c_name
HAVING MAX(count);
```

◯ D.
```
SELECT MAX(count)
FROM (
  SELECT c_name, COUNT(DISTINCT m_name) AS count
  FROM survey
  GROUP BY c_name
  )
WHERE count >= MAX(count);
```



Big Data

2. [6 Pts] The figure above from class shows four distinct file blocks labeled A, B, C, and D spread across four machines, where each machine holds exactly 3 blocks.
   (a) For the figure above, at most, how many of our machines can fail without any data loss? **2**
   (b) Suppose that instead of 4 machines, we have only 3 machines that can store 3 blocks each. Suppose we want to be able to recover our data even if two machines fail. What is the maximum total number of distinct blocks we can store? **3**
   (c) Same as part b, but now suppose we only need to be able to recover our data if one machine fails. What is the maximum total number of distinct blocks we can store? **4**

3. [3 Pts] Suppose we use the map reduce paradigm to compute the total lab scores for each student in DS100. Suppose there are 800 students and 12 labs, and exactly 1 submission per student. Each execution of the map operation is an execution of the autograder, i.e. it will compute the score for a single lab for a single student. The reduce operation computes the total score for each student by adding up all of the lab scores.
   (a) How many key value pairs will be generated in total after all map operations have completed execution? **9600**
   (b) How many distinct keys will there be? **800**
   (c) How many final key value pairs will remain after all reduce operations have completed? **800**

4. [3 Pts] As described in class, the traditional data warehouse is a large tabular database that is periodically updated through the ETL process, which combines data from several smaller data sources into a common tabular format. The alternative is a data lake, where data is stored in its original natural form. Which of the following are good reasons to use a data lake approach?
   ☐ A. The data is sensitive, e.g. medical data or government secrets.
   ☐ B. To maximize compatibility with commercial data analysis and visualization tools.
   ☐ **C. When there is no natural way to store the data in tabular format.**
   ☐ D. To ensure that the data is clean.

Bootstrap and the Sampling Distribution

5. Note: The following problem is stylized to fit on paper, meaning that the sample size and number of replicates are much smaller than they should be in practice.

   In order to infer the population mean, variance, and median of a discrete random variable X, a single simple random sample of size 5 is drawn from the population.

x	11	11	92	34	53

The mean, variance, and median of the values above are 40.2, 918.16, and 34, respectively. You decide to perform a bootstrap analysis of your statistics. The replicates and (rounded) summary statistics are given to you below. Bootstrap replicates are displayed either as rows or columns–you must infer which from the structure of this problem. `rowmeans`, `rowvars`, and `rowmedians` are the means, variances, and medians of the rows respectively.

replicates	0	1	2	3	4	rowmeans	rowvars	rowmedians
0	11	11	34	11	92	32	985	11
1	53	92	92	92	92	84	243	92
2	11	11	11	53	34	24	290	11
3	53	92	11	34	11	40	918	34
colmeans	32	52	37	48	57			
colvars	441	1640	1096	881	1274			
colmedians	11	11	11	34	34			

The summary statistics are loaded into a Python session as numpy arrays.

(a) [1 Pt] What is the sample mean? 40.2

(b) [3 Pts] Which of the following is the bootstrap estimate for the variance of the sample mean?

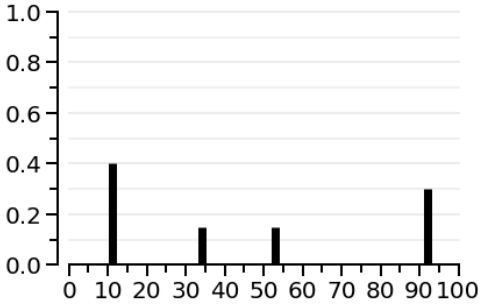
- ◯ **A. `np.var(rowmeans)`**
- ◯ B. `np.mean(rowvars)`
- ◯ C. `np.var(colmeans)`
- ◯ D. `np.mean(colvars)`
- ◯ E. None of the above

(c) [4 Pts] Which of the following changes would decrease the variance of the sample mean? Select all that apply.

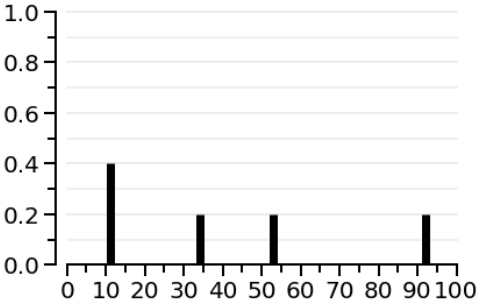
- ☐ **A. Increasing the sample size**
- ☐ B. Increasing the size of each bootstrap replicate
- ☐ C. Increasing the number of bootstrap replicates
- ☐ D. Combining all the bootstrap replicates into one array before estimating the variance

(d) [3 Pts] Which of the following plots displays the approximate sampling distribution of the sample median according to the results of the bootstrap?

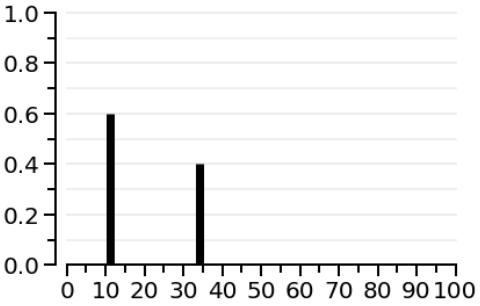
◯ A.



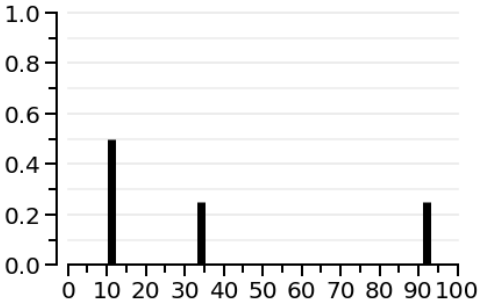
◯ B.



◯ C.



◯ **D.**



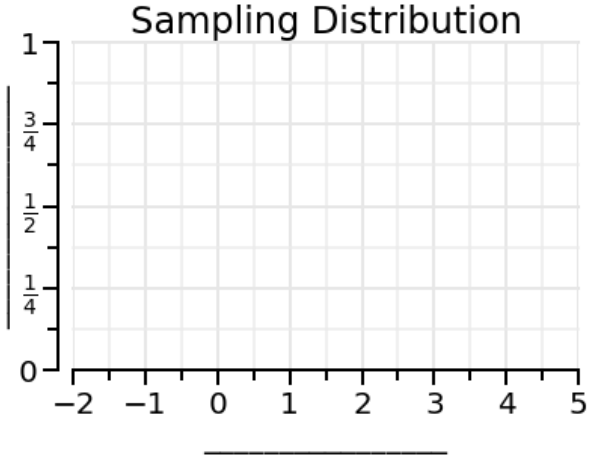
(e) [3 Pts] Using the results of the bootstrap, construct a 50% confidence interval for the population variance.

[**243** , **918**]

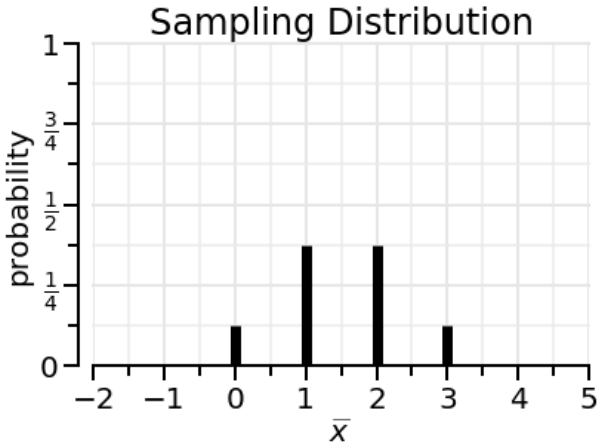
(f) [2 Pts] Above, we drew R bootstrap replicates. If we draw many more sets of bootstrap replicates, say $M = 10000$ sets of R replicates each, and calculate 10000 50% confidence intervals for the population variance the same way we did in the part above, then roughly 50% of those intervals will contain the true population variance.

◯ True √ **False**

6. [4 Pts] This problem is unrelated to the previous problem. Suppose you have a box containing prize tickets. Half of the tickets are worth \$0, and other half are worth \$3. Suppose you draw a simple random sample of size 3 with replacement from the box. Plot the sampling distribution of the sample mean and provide titles for the axes.



Solution:


Sampling Distribution
probability
1
3/4
1/2
1/4
0
−2
−1
0
1
2
3
4
5
$\bar{x}$

Hypothesis Testing

7. A mysterious stranger on Sproul Plaza stops you on your way to class and claims that she has learned to flip any coin such that it lands on heads more often than the 50% you'd expect from random chance. To demonstrate, she takes a penny from her wallet, flips it 10 times, and gets heads nine times and only gets tails once.

   (a) [4 Pts] The null hypothesis is that this was pure random chance, and that the probability of getting heads was 50% for each flip. What is the p-value under the null hypothesis of getting 1 or fewer tails out of 10 flips? You may leave your answer as a fraction if necessary.

   Solution: $\frac{11}{2^{10}}$
   There is 1 way to get all heads, and n ways to get one tails. This is out of a total of 2^n permutations possible. Therefore the chances of getting 1 or fewer heads is $\frac{11}{2^{10}}$.

   (b) [4 Pts] Suppose the stranger flips the coin 28 more times, and they all end up heads. The resulting p value including all 38 flips under the null hypothesis is approximately $p_b = 10^{-10}$. Which of the following are true? Select all that apply.

   - ☐ **A. It is extremely unlikely that the stranger just happened to get 37 heads by randomly getting heads on 50/50 coin flips.**
   - ☐ B. p_b is the probability that the null hypothesis is true.
   - ☐ C. $1 - p_b$ is the probability that the stranger has the skill to flip any arbitrary coin and get heads.
   - ☐ **D. If you flipped a fair coin 38 times, p_b is the chance that you'd get at least 37 heads by random chance.**
   - ☐ E. The stranger has proven beyond any reasonable doubt that she has the skill to flip any coin to land on heads with high probability.
   - ☐ F. None of the above.

8. The DS100 staff is trying to test whether going to discussion improves a students grade in the class. In order to test this, they consider an observational study to measure possible effects of going to discussion on a students grade in the class.

   (a) [4 Pts] For Fall 2019, suppose A is a Series containing final exam grades for students who attended section 7 or more times, and B is a Series containing final exam grades for students who attended section 3 or fewer times. Assume that there are no students who attended section 4, 5, or 6 times. The staff wishes to evaluate the null hypothesis that attending discussion has no relationship with a student's score on the final exam. Which of the following are reasonable test statistics to evaluate the null hypothesis? Select all that apply.

☐ **A. `A.mean() - B.mean()`**

☐ B. `A.sum() - B.sum()`

☐ **C. `A.median() - B.median()`**

☐ D. `A.max() - B.max()`

(b) [2 Pts] Suppose the Fall 2019 staff selects the first statistic above: `A.mean() - B.mean()`. Suppose this difference is 8 points out of 100. From this information alone, what can the staff conclude? Select all that apply.

☐ A. It is very unlikely that this large difference in performance was merely due to chance.

☐ B. Attending section helps improve a student's performance.

☐ **C. Neither of these.**

(c) [2 Pts] A staff member suggests using the bootstrap to create a confidence interval for the test statistic from part b. Another staff member disagrees and says that the bootstrap would be useless for this purpose because the data is already a census, not a sample. Who is right?

◯ A. The bootstrap would yield a useful confidence interval.

◯ B. The bootstrap confidence interval would be useless since the data is already a census.

◯ C. Neither is correct. The bootstrap confidence interval would be useless, but for a different reason.

Solution: A was awarded 2 points, B was awarded 1 point.

(d) [3 Pts] Another staff member suggests using a permutation test. Which of the following could a permutation test help with? Select all that apply.

☐ **A. Can be used to provide a p value for the null hypothesis.**

☐ **B. Can provide strong evidence that the difference in performance was not merely due to chance.**

☐ C. Can establish a causal relationship that attending section helps improve a student's performance.

Classification

9. Suppose we train a binary classifier on some dataset. Suppose y is the set of true labels, and $\hat{y}$ is the set of predicted labels.

y	0	0	0	0	0	1	1	1	1	1
$\hat{y}$	0	1	1	1	1	1	1	0	0	0

Determine each of the following quantities.

(a) [1 Pt] The number of true positives

Solution: 2

(b) [1 Pt] The number of false negatives

Solution: 3

(c) [1 Pt] The precision of our classifier. Write your answer as a simplified fraction.

Solution: $\frac{2}{2+4} = \frac{1}{3}$

10. You have a classification data set, where x is some value and y is the label for that value:

x	y
2	1
3	0
0	1
1	0

Suppose that we're using a logistic regression model to predict the probability that $Y = 1$ given x:

$$\mathbb{P}(Y = 1|x) = \sigma(\phi^T(x)\theta)$$

(a) [6 Pts] Suppose that $\phi(x) = [\phi_1 \quad \phi_2 \quad \phi_3]^T = \begin{bmatrix}1 & x & x^2\end{bmatrix}^T$ and our model parameters are $\theta^* = \begin{bmatrix}1 & 0 & -2\end{bmatrix}^T$. For the following parts, leave your answer as an expression (do not numerically evaluate $\log$, e, π, etc).

i. Compute $\hat{\mathbb{P}}(y = 1|x = 0)$.

Solution: $\frac{1}{1+\exp(-1)}$

ii. What is the loss for this single prediction $\hat{\mathbb{P}}(y = 1|x = 0)$, assuming we are using KL divergence as our loss function (or equivalently that we are using the cross entropy as our loss function)?

Solution: $\log(1 + \exp(-1))$

(b) [4 Pts] Suppose $\phi(x) = \begin{bmatrix} 1 & x & x\%2 \end{bmatrix}^T$, where $\%$ is the modulus operator. Are the data from part a linearly separable with these features? If so, give the equation for a separating plane, e.g. $\phi_2 = 3\phi_3 + 1$. Use 1-indexing, e.g. we have ϕ_1, ϕ_2, and ϕ_3. If not, just write "no".

Solution: Yes, they can be separated by the hyperplane $\phi_3 = 0.5$.

11. [4 Pts] Suppose we have the dataset below.

x	y
1	1
-1	0

Suppose we have the feature set $\phi(x) = [\phi_1 \quad \phi_2]^T = [1 \quad x]^T$. Suppose we use gradient descent to compute the θ which minimizes the KL divergence under a logistic model without regularization, i.e.

$$\arg\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n}(y_i\phi(x_i)^T + log(\sigma(-\phi(x_i)^T\theta)))$$

Select all that are true regarding the data points and the optimal theta value θ.

☐ **A. The data is linearly separable.**

☐ **B. The optimal θ yields an average cross entropy loss of zero.**

☐ C. The optimal θ diverges to $-\infty$

☐ **D. The optimal θ diverges to $+\infty$**

☐ **E. The equation of the line that separates the 2 classes is $\phi_2 = 0$.**

☐ F. None of the above.

Solution:

☐ **A. True. When drawn in the 2-D feature space, the points are linearly separable.**

☐ **B. True. If the data is linearly separable, we can achieve an average cross entropy loss of zero and our parameter value θ will diverge.**

☐ C. False. The optimal theta value θ diverges to $+\infty$

☐ **D. True. The optimal theta value θ diverges to $+\infty$**

☐ **E. True. If we draw the line $\phi_2 = 0$ in the 2-D feature space, this separates the points.**

☐ F. False. 4 choices were true above.

12. Suppose we have the dataset below.

x	y
-3	1
-1	0
1	0
3	1

Suppose we have the feature set $\phi(x) = \begin{bmatrix}1 & x^2\end{bmatrix}^T$. Suppose we use gradient descent to compute the θ which minimizes the KL divergence under a logistic model without regularization, i.e.

$$\arg\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n}(y_i\phi(x_i)^T + log(\sigma(-\phi(x_i)^T\theta)))$$

(a) [3 Pts] Explain in 10 words or fewer why the magnitudes of θ_1 and θ_2 will be very large.

Solution: Because the data is linearly separable.

(b) [3 Pts] Will the sign of θ_2 be negative or positive?

- ◯ A. Could be either, it depends on where our gradient descent starts
- ◯ **B. Positive**
- ◯ C. Negative
- ◯ D. Neither, θ_2 will be zero

(c) [3 Pts] If we use L_1 regularization, which of our θ values would you expect to be zero?

- ◯ **A. Neither of them**
- ◯ B. θ_1
- ◯ C. θ_2
- ◯ D. Both θ_1 and θ_2

Bias Variance Tradeoff

13. In class, we showed that the expected squared error can be decomposed into several important terms:

$$\mathbb{E}[(Y - f_{\hat{\theta}}(x))^2] = \sigma^2 + (h(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2 + \mathbb{E}[(\mathbb{E}[f_{\hat{\theta}}(x)] - f_{\hat{\theta}}(x))^2].$$

(a) [1 Pt] For which of the following reasons are we taking an expectation? In other words, what are the sources of randomness that we are considering in the derivation of the bias-variance tradeoff?

- ☐ A. We chose arbitrary features when doing feature engineering.
- ☐ **B. We drew random samples from some larger population when we built our training set.**
- ☐ **C. There is some noise in the underlying process that generates our observations Y from our features.**
- ☐ D. Our x values could have had missing or erroneous data, e.g. participants misreading a question on a survey.
- ☐ E. None of the Above.

(b) [1.5 Pts] Which of the following do we treat as fixed? Select all that apply.

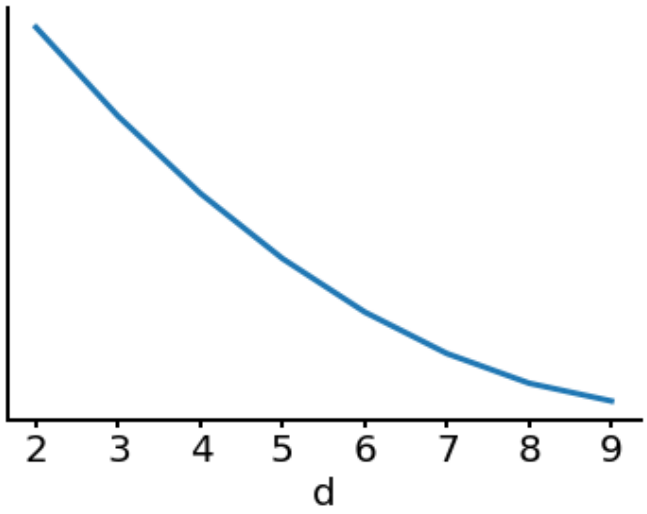
- ☐ A. $\hat{\theta}$
- ☐ **B.** σ^2
- ☐ **C.** $h(x)$

(c) [1 Pt] By decreasing model complexity, we are able to decrease σ^2.

- ◯ A. True
- ◯ **B. False**

14. Your team would like to train a machine learning model in order to predict the next YouTube video that a user will click on based on m features for each of the previous d videos watched by that user. In other words, the total number of features is $m \times d$. You're not sure how many videos to consider.

(a) [2 Pts] Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



Which of the following could the y axis represent? Select all that apply.

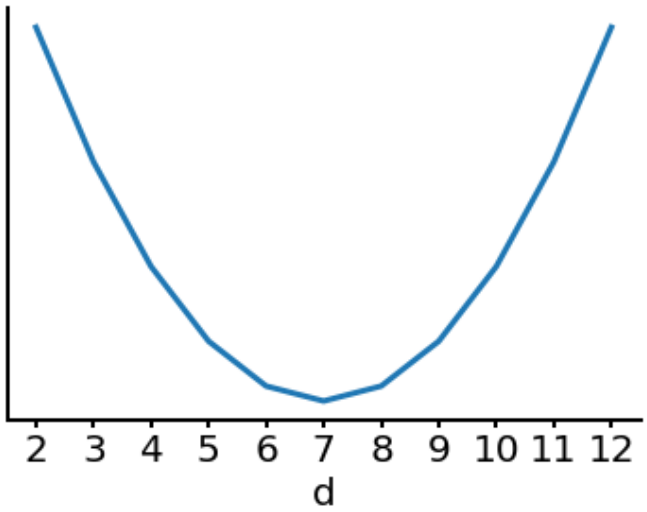
☐ **A. Training Error**

☐ **B. Validation Error**

☐ **C. Bias**

☐ D. Variance

(b) [2 Pts] Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



Which of the following could the y axis represent? Select all that apply.

☐ A. Training Error

☐ **B. Validation Error**

☐ C. Bias

☐ D. Variance

Cross Validation

15. [2.5 Pts] Aman and Ed built a model on their data with two regularization hyperparameters λ and γ. They have 4 good candidate values for λ and 3 possible values for γ, and they are wondering which λ, γ pair will be the best choice. If they were to perform five-fold cross-validation, how many validation errors would they need to calculate?

Solution: 60

16. [2 Pts] In the typical setup of k-fold cross validation, we use a different parameter value on each fold, compute the mean squared error of each fold and choose the parameter whose fold has the lowest loss.
◯ True √ **False**

17. [2 Pts] Suppose we have m data points in our training set and n data points in our test set. In *leave-one-out* cross validation, we only use one data point for validation while the rest are used for training. Which of the following is *leave-one-out* cross validation equivalent to?

◯ **A. m-fold cross validation**
◯ B. n-fold cross validation
◯ C. (m + n)-fold cross validation
◯ D. 1-fold cross validation

18. [5 Pts] Suppose we have a linear regression model with L2 regularization that we'd like to train. Recall that ridge regression has a single hyperparameter λ. Suppose we are trying to pick a λ value from $[0, 0.1, 0.2, 0.3, 0.4]$. In class, we discussed cross validation, but there are other ways we could attempt to compute λ. Let λ_{CV} be the optimal λ that would be calculated using 5-fold cross validation. Let λ_X be the optimal λ that would be computed using procedure X below.

Procedure X: Don't create any sort of validation set. Instead, for every candidate λ value, compute the theta that minimizes the average loss over the **entire training set** including the regularization term $\lambda \sum_{i=1}^{d} \theta_i^2$. Return the λ that yields the lowest training loss.

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

☐ A. λ_X will require vastly more computation time to compute than λ_{CV}.
☐ **B. $\lambda_X \leq \lambda_{CV}$.**
☐ C. Procedure X reduces the risk of overfitting even more than cross validation.
☐ D. If we computed both, we should use the smaller of λ_X and λ_{CV} to train our final model.
☐ E. None of the above are true.

19. Suppose you are working with a partner to train a model with one hyperparameter λ. Together, you and your partner run 5-fold cross validation and compute mean squared errors for each fold and value of λ from a set of 4 candidate values for λ. However, your partner forgets to send you the results for the last two folds! The table below contains the mean squared errors for the first three of five total folds.

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	64.2	60.1	77.7	79.2	70.3
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0

(a) [3 Pts] Your partner uses the full table containing data for all five folds to create a final model to use on test data. Given the information above, what can you conclude about the final model? Select all that apply.

☐ A. Our final model should use $\lambda = 0.4$.

☐ B. Our final model should be trained on fold 1, since it achieves the lowest row average.

☐ C. Our final model should be trained on fold 2, since it achieves the highest row average.

☐ **D. None of the above.**

(b) [2 Pts] Let's say we know the row averages for all 5 folds. Which of the following are valid conclusions we can draw? Select all that apply.

☐ A. We can determine which fold number to use for our model.

☐ B. We can determine which λ value to use in our model.

☐ **C. None of the above.**

Regularization

20. [3.5 Pts] Which of the following are indications that you should regularize? Select all that apply.

- ☐ **A. Our training loss is 0.**
- ☐ B. Our model bias is too high.
- ☐ **C. Our model variance is too high.**
- ☐ **D. Our weights are too large.**
- ☐ E. Our model does better on unseen data than training data.
- ☐ **F. We have linearly dependent features.**
- ☐ **G. We are training a classification model and the data is linearly separable.**

21. [7.5 Pts] Suppose we have a data set which we divide into 3 equally sized parts, A, B, and C. We fit 3 linear regression models with L2 regularization (i.e. ridge regression), X, Y, and Z, all on A. Each model uses the same features and training set, the only difference is the λ used by each model. Select all below that are **always true**.

- ☐ A. Suppose Z has the lowest average loss on B. Model Z will have the lowest average loss when evaluated on C.
- ☐ B. If A and B have the same exact mean and variance, the average loss of model Y on B will be exactly equal to the average loss of Y on A.
- ☐ **C. If $\lambda = 0$ for model X, $Loss(X, A) \leq Loss(Y, A)$ and $Loss(X, A) \leq Loss(Z, A)$.**
- ☐ **D. If $\lambda_Y < \lambda_Z$, then $Loss(Y, A) \leq Loss(Z, A)$.**
- ☐ E. If $\lambda_Y > \lambda_Z$, then $Loss(Y, B) \geq Loss(Z, B)$.
- ☐ F. None of the above.

Solution:
A: Not guaranteed since we don't know the distributions of B, C.
B: Having the same mean and variance does not imply that the data are the same.
C: Since increasing λ increases bias, the loss of X must be less than or equal to the loss of Y, Z on A.
D: Since Y and Z were trained on A, and Y is less restricted than Z, the loss of Y on A must be less than the loss of Z on A. E: Even though Z is a more restricted (i.e. simpler) model, it is possible that the dataset B is slightly better for Z. In other words, minimizing training error with a regularized model does not guarantee minimized error on unseen datasets.

Probability and Potpourri

Some of the problems in this section were explicitly (or near explicitly) covered in a lecture or discussion section. As a result, they are worth fewer points than you might expect given the amount of work needed to derive them from scratch. If you find yourself spending a ton of time on one of these, come back to it later.

22. Recall from lecture 21 that Jensen's Inequality states that for a random variable X and a **convex function** f, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.

(a) [2 Pts] In class, we showed that the Kullback-Leibler divergence $D_{KL}(P||Q) \geq 0$. To prove this, we applied Jensen's inequality with which of the following four functions?

◯ **A.** $f(X) = -\log X$
◯ B. $f(X) = \log X$
◯ C. $f(X) = -X^2$
◯ D. $f(X) = X^2$

Solution: $-\log X$. We did this proof in class, and $-\log X$ is the appropriate convex function.

(b) [3 Pts] As we know, the variance of a random variable is always greater than or equal to 0, i.e. $\text{Var}(X) \geq 0$. Give a function f that lets us trivially prove this statement using Jensen's inequality. For example, if we can prove that the variance is always non-negative by plugging in $f(x) = e^x$, write e^x.

Solution: X^2. Since X^2 is convex, Jensen's inequality tells us that $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$. Since $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, we know that $\text{Var}(X) \geq 0$.

(c) [2 Pts] For which of the following functions f will equality (instead of inequality) hold for Jensen's inequality regardless of the random variable X? You may assume a, b, and c are constants. Select all that apply.

☐ **A.** $f(X) = a$
☐ **B.** $f(X) = aX$
☐ **C.** $f(X) = aX + b$
☐ D. $f(X) = aX^2 + bX + c$
☐ E. None of the Above

Solution: This follows directly from the linearity of expectation. Note that in general, $\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$.

23. A/B Testing

(a) [2 Pts] You have a coin which lands heads on average 50% of the time, and tails 50% of the time. You know that the coin is fair, having flipped it millions of times in the past. Suppose your most recent four flips have all been heads. How does the probability that a fifth toss will again be heads compared to the probability that a fifth toss will be tails?

◯ A. The odds of heads are greater than 50% for the fifth flip.

◯ **B. The odds of heads are 50% for the fifth flip.**

◯ C. The odds of heads are less than 50% for the fifth flip.

(b) [4 Pts] You want to know whether or not a coin is fair. As you flip the coin, you keep track of your test statistic D, which is the total number of heads observed minus the total number of tails observed. After each flip, you compute the p value under the null hypothesis (that the coin is fair) for your observed value of D. If the p value ever falls below 0.05, you stop and announce the coin is unfair. If you reach T flips, you announce the coin is fair. Suppose you have a fair coin, what is the probability p_{FD} of a false discovery with this procedure using that fair coin?

◯ A. At most 5%, independent of the value of T.

◯ B. As T grows, p_{FD} asymptotically decreases to 0%.

◯ **C. As T grows, p_{FD} asymptotically increases to 100%.**

24. [2 Pts] Which of the following statements are true? Select all that apply.

☐ **A. A matrix with a condition number of 10^{20} will magnify numerical issues more than a matrix with a condition number of 10^{10}.**

☐ **B. Suppose a 3 dimensional sphere is inscribed in a 3 dimensional cube. The volume inside the sphere is greater than the volume outside the sphere but inside the cube.**

☐ C. Suppose a 20 dimensional sphere is inscribed in a 20 dimensional hypercube. The volume inside the sphere is greater than the volume outside the sphere but inside the hypercube.

☐ **D. Suppose $f_1(x) = rx(1-x)$, and $f_2(x) = rx - rx^2$. For $r = 1.9$ and $x = 0.8$, f_1 and f_2 will return two different numbers in Python.**

Solution: These were all discussed in the numerical issues lecture.

25. [2 Pts] For this problem, recall that stochastic gradient descent is very similar to normal gradient descent, except that the gradient of the loss function is computed on a random sample of the data instead of the entire dataset. Which of the following are true? Select all that apply.

☐ A. At a particular iteration, stochastic gradient descent will often update θ more accurately compared to an update with regular gradient descent.

☐ B. For a convex loss function, a single step of gradient descent always decreases the loss.

☐ C. For a convex loss function, a single step of stochastic gradient descent always decreases the loss.

☐ **D. Suppose it takes t seconds for one update of regular gradient descent, Stochastic gradient descent can usually perform more than one update in t seconds.**

☐ E. None of the Above

Linear Regression (Hard Problem)

Throughout this section we refer to "least squares regression", which is the process of minimizing the average L2 loss using a linear regression model. Ordinary least squares is the version of least squares regression where **we do not use regularization**. Assume throughout that **our model includes a bias term**. Warning: Parts of this problem are pretty hard!

26. [3 Pts] What is always true about the residuals in least squares regression? Select all that apply.

- ☐ **A. They are orthogonal to the column space of the features.**
- ☐ **B. They represent the errors of the predictions.**
- ☐ C. Their sum is equal to the mean squared error.
- ☐ **D. Their sum is equal to zero.**
- ☐ E. None of the above.

Solution: (a), (b)

(c) is supposed to be a trick since the mean squared error is the *mean* of the sum of the *squares* of the residuals. So I guess this tests whether they understand what the acronym MSE represents or if they just regurgitate it mindlessly.

(e) is wrong since (c) is wrong obviously

27. [3 Pts] What are possible disadvantages of ordinary least squares (OLS) regression compared to ridge or LASSO? Select all that apply.

- ☐ A. The OLS estimate selects too few features as being important.
- ☐ B. OLS has unrealistically small variance compared to LASSO or ridge.
- ☐ C. OLS is computationally much more expensive.
- ☐ **D. OLS is more prone to overfitting.**
- ☐ E. None of the above.

Solution: (e)

(a) LASSO tends to identify fewer features as important

(b) LASSO and ridge are biased, but often have smaller MSE than OLS (that's why people use them after all), which means that OLS (often) has larger variance compared to LASSO and ridge,

not smaller

(c) OLS has the simplest loss function and the simplest closed form solution

28. [3 Pts] What differentiates LASSO compared to OLS? Select all that apply.

☐ A. LASSO uses the mean absolute error (MAE) loss, while OLS uses mean squared error (MSE).

☐ B. LASSO tends to identify more features as relevant.

☐ C. LASSO typically has lower average error on the training data.

☐ D. All weights in a LASSO model must be less than 1.

☐ **E. None of the above.**

Solution: (a) is false, since it's the penalization of the coefficients, not the residuals, which uses $L1$ in LASSO.

(b) LASSO tends to identify fewer features as relevant (the second s in LASSO is selection after all).

(c) The error for a LASSO model is always greater than or equal to an OLS model.

(d) the LASSO estimate is biased whereas OLS is unbiased.

(e) is wrong since (a), (b), (c), and (d) are.

29. [3 Pts] Which are true about the predictions made by OLS? Select all that apply.

☐ **A. They are projections of the observations onto the column space of the features.**

☐ **B. They are linear in the chosen features.**

☐ **C. They are orthogonal to the residuals.**

☐ D. They are orthogonal to the column space of the features.

☐ E. None of the above.

Solution: (a), (b), (c)

(a) is correct because they are linear projections onto the column space. This fact also makes (c) correct and (e) incorrect and is what makes (d) incorrect.

(b) is also correct because even in e.g. polynomial regression the resulting predictions are linear in the new/transformed features. But admittedly this is somewhat awkwardly worded.

30. [3 Pts] Which of the following would be true if you chose mean absolute error (L1) instead of mean squared error (L2) as your loss function? Select all that apply.

☐ A. The results of the regression would be more sensitive to outliers.

☐ B. You would not be able to use gradient descent to find the regression line.

☐ **C. You would not be able to use the normal equation to calculate your parameters.**

☐ D. The sum of the residuals would now be zero.

☐ E. None of the above.

Solution: (e)

(a) is false because using $L1$ loss increases robustness to outliers.

(b) is false because you can still use (sub)gradient descent given the convexity of $L1$ loss.

(c) is true, the normal equation only works if we're minimizing the L2 loss.

(d) is false because the sum of the residuals was zero in OLS, so if this happened it wouldn't be a change from OLS.

31. Let $\hat{\boldsymbol{y}} \in \mathbb{R}^n$ be the vector of fitted values in the ordinary least squares regression of $\boldsymbol{y} \in \mathbb{R}^n$ on the full column-rank feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n \times d}$ with n much larger than d. Denote the fitted coefficients as $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ and the vector of residuals as $\boldsymbol{e} \in \mathbb{R}^n$.

(a) [4 Pts] What is $\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$?

◯ A. $\mathbf{0}$ ◯ **B. $\hat{\boldsymbol{y}}$** ◯ C. $\boldsymbol{e}$ ◯ D. $\hat{\boldsymbol{\beta}}$ ◯ E. 1 ◯ F. None of the above

Solution: We discussed this in discussion 6, where we called $\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$ the hat matrix. It projects y into the feature space.

(b) [4 Pts] What is $\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\hat{\boldsymbol{y}}$? Notice: This problem has a hat in $\hat{\boldsymbol{y}}$.

◯ A. $\mathbf{0}$ ◯ **B. $\hat{\boldsymbol{y}}$** ◯ C. $\boldsymbol{e}$ ◯ D. $\hat{\boldsymbol{\beta}}$ ◯ E. 1 ◯ F. None of the above

Solution: Since $\hat{y}$ is already in the feature space, projecting it into the feature space has no effect.

Suppose $\boldsymbol{e} \neq \mathbf{0}$. Define a new feature matrix $\boldsymbol{\Psi}$ by appending the residual vector $\boldsymbol{e}$ to the feature matrix $\boldsymbol{\Phi}$. In other words,

$$\boldsymbol{\Psi} = \begin{bmatrix} | & | & \vdots & | & | \\ \boldsymbol{\Phi}_{:,1} & \boldsymbol{\Phi}_{:,2} & \cdots & \boldsymbol{\Phi}_{:,d} & \boldsymbol{e} \\ | & | & \vdots & | & | \end{bmatrix}$$

(c) [4 Pts] We now want to fit the model $\boldsymbol{y} = \boldsymbol{\Psi\gamma} = \gamma_1\boldsymbol{\Phi_{:,1}} + \gamma_2\boldsymbol{\Phi_{:,2}} + \cdots + \gamma_d\boldsymbol{\Phi_{:,d}} + \gamma_{d+1}\boldsymbol{e}$ by choosing $\hat{\boldsymbol{\gamma}} = [\hat{\gamma}_1 \ldots \hat{\gamma}_{d+1}]^T$ to minimize the L_2 loss. What is $\hat{\gamma}_{d+1}$?

◯ A. 0 ◯ **B.** 1 ◯ C. $\boldsymbol{e}^T\boldsymbol{y}$ ◯ D. $1 - \hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}$
◯ E. $(\Phi^T\Phi)^{-1}\Phi^T$ ◯ F. None of the above

Solution: We're effectively memorizing all of our regression values here. This is the equivalent (in a roundabout way) of using someone's weight, age, and height to predict their height. It'll work perfectly, but the model is useless.

32. We collect some data $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ and decide to model the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$ as

$$\boldsymbol{y} = \beta_1\boldsymbol{\Phi_{:,1}} + \beta_2\boldsymbol{\Phi_{:,2}}$$

where $\boldsymbol{\Phi_{i,:}} = \begin{bmatrix}1 & x_i\end{bmatrix}$ We found the estimates $\hat{\beta}_1 = 2$ and $\hat{\beta}_2 = 5$ for the coefficients by minimizing the L_2 loss. Given that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \begin{bmatrix}4 & 2\\ 2 & 5\end{bmatrix}$, answer the following problems. If not enough information is given, write "Cannot be determined."

(a) [4 Pts] What was the sample size n? Hint: Consider the form of the feature matrix.

Solution:

$$[\boldsymbol{\Phi}^T\boldsymbol{\Phi}]_{1,1} = \sum_{i=1}^{n} 1 \times 1 = n = 4$$

(b) [7 Pts] What must $\boldsymbol{\Phi}^T\boldsymbol{y}$ be for this data set?

Solution: $\hat{\boldsymbol{\beta}}$ comes from the normal equations

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi}\hat{\boldsymbol{\beta}} = \boldsymbol{\Phi}^T\boldsymbol{y}$$

Therefore, we have

$$\boldsymbol{\Phi}^T\boldsymbol{y} = \begin{bmatrix}4 & 2\\ 2 & 5\end{bmatrix}\begin{bmatrix}2\\ 5\end{bmatrix} = \begin{bmatrix}18\\ 29\end{bmatrix}$$

Cleaning, EDA, Visualization

Let's take a look at the California Air Quality Index (AQI) for 2017. The following cells and outputs are for your reference.

```
aq = pd.read_csv("./air_quality_final.csv", index_col=0)
aq.head()
```

	Date	AQI	COUNTY_CODE	COUNTY	LAT	LON
0	01/01/2017	24.0	1	Alameda	37.687526	-121.784217
1	01/02/2017	19.0	1	Alameda	37.687526	-121.784217
2	01/03/2017	NaN	1	Alameda	37.687526	-121.784217
3	01/04/2017	15.0	1	Alameda	0.000000	0.000000
4	01/05/2017	20.0	1	NaN	37.687526	-121.784217

```
aq.iloc[49437:49442]
```

	Date	AQI	COUNTY_CODE	COUNTY	LAT	LON
49437	01/01/2017	NaN	113	Yolo	38.534450	-121.773400
49438	01/02/2017	15.0	113	Yolo	38.534450	-121.773400
49439	01/03/2017	36.0	113	Yolo	38.534450	-121.773400
49440	01/04/2017	18.0	113	Yolo	37.995239	-121.756812
49441	01/05/2017	16.0	113	NaN	38.534450	-121.773400

```
aq.describe()
```

	AQI	COUNTY_CODE	LAT	LON
count	49810.000000	49812.000000	49812.000000	49812.000000
mean	38.270167	56.169678	36.294056	-119.859815
std	24.750558	30.486150	2.235560	2.099002
min	0.000000	1.000000	0.000000	-124.203470
25%	21.000000	29.000000	34.144350	-121.618549
50%	35.000000	65.000000	36.487823	-119.828400
75%	52.000000	77.000000	37.960400	-118.147294
max	537.000000	113.000000	41.756130	0.000000

```
print(aq['COUNTY'].nunique())
Output: 51
```

33. [3 Pts] Select all that apply.

☐ **A. Supposing that there is a one to one mapping from COUNTY_CODE to COUNTY, we can extrapolate the value of COUNTY for index 4.**

☐ B. Grouping by COUNTY is equivalent to grouping by LAT, LON.

☐ C. The primary key in this dataset is the DATE.

☐ D. None of the above

Solution:
A: True
B: No, there are different latitude and longitudes for a county
C: Dates are not unique.

For all following questions, assume we have finished cleaning the dataset (filled in or removed missing, NaN, etc.).

34. [2 Pts] Which of the following correctly verifies that the mapping from COUNTY_CODE to COUNTY is 1 to 1? Select only one.

◯ A.
```
len(aq['COUNTY'].value_counts()) ==
len(aq['COUNTY_CODE'].value_counts())
```

◯ B.
```
len(set(aq['COUNTY'])) ==
len(set(aq['COUNTY_CODE']))
```

◯ C.
```
len(aq['COUNTY'].unique()) ==
len(aq['COUNTY_CODE'].unique())
```

◯ **D.**
```
len(aq.groupby(['COUNTY', 'COUNTY_CODE'])) ==
len(set(aq['COUNTY'])) and
len(set(aq['COUNTY'])) == len(set(aq['COUNTY_CODE']))
```

◯ E. None of the above

Solution:
A-C: Having 51 unique COUNTY values and 51 COUNTY_CODE values does not imply a 1 to 1 mapping
D: Correct

35. [2 Pts] In the questions below, select the best plot to visualize a certain aspect of our data.

(a) visualize the AQI for Los Angeles, San Diego, San Francisco, Humboldt, and Inyo counties over the first 7 days of January 2017.

◯ A. Stacked bar plot

◯ **B. Side by side line plot**

◯ C. KDE plot

◯ D. Side by side violin plot

(b) visualize the distribution of site locations by latitude and longitude.

◯ A. Histogram

◯ **B. Scatter plot**

◯ C. Bar plot

◯ D. 1D KDE plot

(c) visualize the average AQI over all counties for each day of January.

◯ A. Overlaid line plot

◯ **B. Line plot**

◯ C. Side by side histogram

◯ D. Side by side box plot

36. [9 Pts] We wish to visualize the mean AQI measurements taken for Alameda, San Francisco and Yolo county over the entire period. Fill in the code below to accomplish this. Use choices from the following table.

aq	'Date'	:	'AQI'	'COUNTY_CODE'
'COUNTY'	'LAT'	'LON'	Alameda	San Francisco
Yolo	str	apply	match	groupby
agg	findall	count	sum	mean
|	==	or	and	filter

```
reg = r'___________ ___________ ___________ ___________ ___________'
temp = (___________
        .loc[
                ___________[___________].str.___________(___________),
                ___________
        ]
        .___________(___________)
        .___________()
        .reset_index())
sns.barplot(x = ___________, y=___________, data=data);
```

Solution:

```
reg = r'Alameda|San Francisco|Yolo'
data = aq.loc[aq['COUNTY'].str.match(reg), :].groupby('COUNTY').mean().res
sns.barplot(x ='COUNTY', y='AQI', data=temp);
```