Quiz 1 - EDA

1 1 / 1 point

Data science is often an iterative process that can generate new questions

- ✓ ● True
- ○ False

2 1 / 1 point

Most data scientists spend the majority of their time developing new models

- ○ True
- ✓ ● False

3 1 / 1 point

The structure of data includes its formatting (e.g., JSON, CSV, XML, raw text) as well as the fields and organization of records.

- ✓ ● True
- ○ False

4 0 / 1 point

Consider the following CSV file on crime rates provided by the US Department of Justice. Assume you have been told that the row delimiter is "newline" and the column delimiter is ",". Which of the following statements are true? Select all that apply.

data.csv

```
Reported crime in Alabama,
,
2004,4029.3
2005,3900
2006,3937
2007,3974.9
2008,4081.9
,
Reported crime in Alaska,
,
2004,3370.9
2005,3615
2006,3582
2007,3373.9
2008,2928.3
2009,3639.8
<End of File>
```

- [x] In its current form, this data cannot be loaded into a dataframe

Selected Answer - Incorrect

- [ ] If we load this data into a rectangular table without transformation, we will lose important information

Missed option - incorrect

- [ ] We could transform the structure of this data into a matrix of real numbers, with a row for each state and column for each year

Missed option - incorrect

- [ ] 2009 is an outlier
- [ ] This data appears to have limited scope

Missed option - incorrect

- [ ] None of the above

Feedback

General Feedback

- data can be entered into a dataframe, however if done without caution information will be lost, e.g. order

- 2009 is not an outlier (not enough data here)

- limited scope is true (we only have 2 states and few years)

5 0 / 1 point

Suppose we want to make a scatter plot for the houses sold in the Maastricht area in 2020. The *x*-axis is the size of the house in square-meters and the *y*-axis is the cost of the house per square-meter. Over 20,000 homes were sold last year. What techniques would you employ to avoid problems related to over-plotting. Select all that apply.

- [ ] jitter the values for cost
- [ ] plot a smoothed curve of average cost for the size of the house

Missed option - incorrect

- [x] make the plotting symbols semi-transparent
- [ ] use color to distinguish the area where the house was sold
- [ ] none of the above

Feedback

General Feedback

tricky question since we do not have access to the plot, however it helps you think in advance. note that we do have ~200K records meaning that the plot is seriously going to be over-plotted.

- jittering could be a solution, however in this case due to the high # of points it won't help much.

- color won't help either (when does it?)

- semi-transparency will help avoiding the over-concentration of points

- plotting a smoothed curve of average cost vs. size, will also help (note that here we are changing the plot)

6 0 / 1 point

Which visualization would be appropriate for examining the relationship between the birth weight of a baby and the number of siblings of the baby? (Assume there are a few hundred observations) Select all that apply.

- [x] side-by-side box plots of weight by number of siblings

Missed option - incorrect

- [x] scatter plot of weight by number of siblings

Selected Answer - Incorrect

- [ ] bar plot of weight by number of siblings
- [x] overlaid density curves of weight, one for each number of siblings

Missed option - incorrect

- [ ] none of the above

Feedback

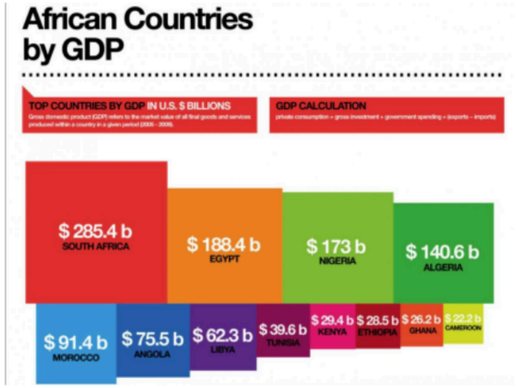
General Feedback

tricky question since we do not have the plot, however it helps you *think* in advance how you should plot. note that we hav hundreds of obervations (but not thousands).

- We have two quantitative variables, however one of them is discrete (# of siblings) meaning that a scatterplot is not the best choice. Same goes for the barplot?

- Side-by-side boxplots (one boxplot for each number of siblings) would allow for comparisons. Same for the density curves (check slides/notebook for how these would look like)

Consider the figure below. Which of the following suggestions would better facilitate comparisons of the GDP for African countries. Select all that apply.



- [ ] arrange the countries in alphabetical order to make it easier to find a country's GDP
- [x] choose a sequential color palette to match size of the GDP

Selected Answer - Incorrect

- [ ] make a box plot of GDP to show the skew and spread in GDP
- [ ] make a dot chart (e.g like this) of GDP

Missed option - incorrect

Feedback

General Feedback

- alphabetical order does not help our goal here

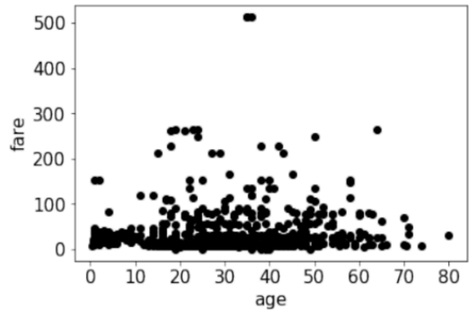
- picking a color palette (like the one in the figure) won't help either - there are too many countries (thus colors) and with color we do not have a size perception

- with the boxplot we can see the distribution of GDP, however we cannot focus on all countries (if we had way more countries, things would be different)

- a dot plot (as seen in the linked example) seems the most plausible option here

8 0 / 1 point

Consider the plot below. What are some ways to improve the plot? Choose all that apply. Assume each is done individually.



- [x] Remove outliers and then plot on a different scale

Missed option - incorrect

- [ ] Plot as a line plot instead of a scatterplot.
- [ ] Jitter the data with noise sampled from a uniform distribution of (-1, 1)
- [x] Utilize transparency
- [ ] none of the above

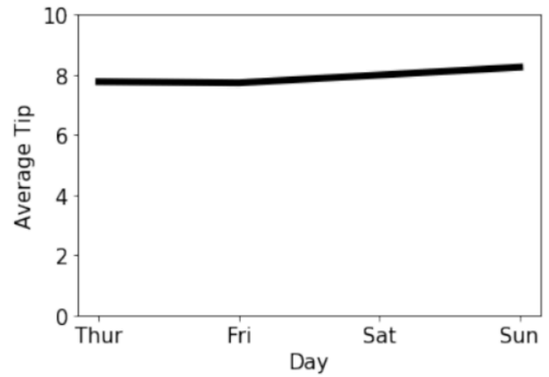
Feedback

General Feedback

removing the outliers (on top) and applying e.g. a log scale might be a good idea. Same with transparency. Jittering here won't help (because of the high concentration of points). A line plot could be clearly wrong.

9 1 / 1 point

Consider the plot below which visualizes day of the week versus the average tip given in euros. What are serious visualization errors made with this plot? Choose all that apply.



- [ ] Area perception
- [ ] Jittering
- [ ] Overplotting
- [ ] Stacking
- [x] None of the above

Feedback

General Feedback

- The problem here is none of the listed, rather than the empty plot region

10 1 / 1 point

Let's take a look at the California Air Quality Index (AQI) for 2017. The following cells and outputs are for your reference.
Select all that apply. You can assume that the records shown are indicative of the complete dataset.

```
> 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())
51
```

- [ ] Grouping by COUNTY is equivalent to grouping by LAT, LON
- [ ] DATE can be used to uniquely identify every record in the dataset
- [x] Supposing that there is a one to one mapping from COUNTY CODE to COUNTY, we can extrapolate the value of COUNTY for index 4 (NB: index here refers to the dataframe index)
- [ ] LAT/LON values seem to be correct and have no errors

11 0 / 1 point

(this question continues from the previous one about AQI)
Select the best plot to visualize the AQI for Los Angeles, San Diego, San Francisco, Humboldt, and Inyo counties over the first 7 days of January 2017.

- ( ) stacked bar plot
- ( ) side-by-side line plot
- (x) KDE (kernel density estimation) plot (similar to histograms)

Correct Answer: **side-by-side line plot**

- ( ) side-by-side violin plot

Feedback

General Feedback

since we have a temporal dimension (clearly time), the only answer is lineplot.

12 0 / 1 point

(this question continues from the previous one about AQI)
Select the best plot to visualize the distribution of site locations by latitude and longitude.

- ( ) histogram
- ( ) scatterplot
- (x) barplot

Correct Answer: **scatterplot**

- ( ) 1-dimensional KDE plot

Quiz 2 - Regression, Classification, Regularization

2 1 / 1 point

A professors runs a regression to see how students' exam scores (Y) are related to their homework grades (X). The R^2 of the regression is 21%. What does R^2 tell us?

- 21% of students have their grades accurately predicted by the regression equation.
- 21% of each student's exam grade will be determined by their homework grade.
- ✓ 21% of the variation in the exam scores is explained by the regression analysis.
- Exam scores are not related to homework grades since 21% is greater than 5%.
- None of the listed options are correct

1 1 / 1 point

Which of the following statements is NOT an assumption of inference for a regression model?

- The dependent variable is linearly related to the explanatory variables.
- The errors around the idealized regression line follow a Normal model.
- The errors around the idealized regression line have equal variability.
- The errors around the idealized regression line are independent of each other.
- ✓ The errors around the idealized regression line are linearly related.

3 1 / 1 point

The problem of collinearity occurs when

- more than one predictor variable is linearly related to the response variable.
- there is an influential observation in the data set.
- at least one predictor var. has a nonlinear relationship with the response variable
- ✓ two or more predictor variables are linearly related to each other.
- none of the listed options are correct

4 0 / 1 point

Which of the following are characteristics of a good regression model?

- relatively few predictor variables
- ✗ a relatively high R^2

Correct Answer: **all of the above**

- relatively small p-values for the F- and t-statistics
- a relatively low value of s (the standard deviation of the residuals)
- all of the above

Feedback

General Feedback

you might disagree with the "few predictor" variables, but overall, try to think about the generalizeability of the model. fewer predictors means lower variance.

5 3 / 3 points

Nearly all the algorithms we have learned about in this course have a tuning parameter for regularization that adjusts the bias/variance tradeoff, and can be used to protect against overfitting. More regularization tends to cause less overfitting.

For each of the following algorithms, we point out one such tuning parameter. If you increase the parameter, does it lead to more or less regularization? Write down "more" or "less" (without the quotes") in the gaps.

Logistic regression with regularization parameter \lambda --> Increasing \lambda means ✓ more regularization. Decision tree with an upper limit on the number of nodes in the tree (let's call it n). Higher n means ✓ less regularization. Boosting with n number of iterations. Higher n means ✓ less regularization.

6 1 / 1 point

In terms of feature selection, L2 regularization is preferred since it comes up with sparse solutions.

- True
- ✓ False

Feedback

General Feedback

L1 regularization (LASSO) comes up with sparse solutions.

7 2 / 2 points

When doing least-squares regression with regularisation (assuming that the optimisation can be done exactly), increasing the value of the regularisation parameter λ

- [x] will never decrease the training error.
- [ ] will never increase the training error.
- [ ] will never decrease the testing error.
- [ ] will never increase the testing error.
- [ ] may either increase or decrease the training error.
- [x] may either increase or decrease the testing error

8 1 / 1 point

Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting? You may select more than one.

- [ ] Increase the amount of training data.
- [x] Improve the optimisation algorithm being used for error minimisation.
- [ ] Decrease the model complexity.
- [ ] Reduce the noise in the training data.

9 1 / 1 point

You are reviewing papers for the World's Fanciest Machine Learning Conference, and you see submis- sions with the following claims. Which ones would you consider accepting?

- My method achieves a training error lower than all previous methods!
- My method achieves a test error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise test error.)
- ✓ My method achieves a test error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise cross-validaton error.)
- My method achieves a cross-validation error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise cross-validaton error.)

10 0 / 1 point

The training error of 1-NN (1-nearest-neighbor) classifier is 0

- True
- ✗ False

Correct Answer: **True**

Feedback

General Feedback

Each point is its own neighbor, so 1-NN classifier achieves perfect classification on training data.

11 1 / 1 point

Which of the following would justify applying regularization to logistic regression? Select all that apply.

- [ ] the training error is too high
- [ ] the testing error is too low
- [x] the data are high-dimensional
- [ ] there is a large class imbalance
- [ ] none of the above justify regularization for logistic regression

12 1 / 1 point

If we get 100% training accuracy with a logistic regression model, then the data is linearly separable

- (x) Always true
- ( ) Sometimes true
- ( ) Never true

15 0 / 1 point

Suppose you have picked the parameter θ for a model using 10-fold cross validation. The best way to pick a final model to use and estimate its error is to:

- ( ) pick any of the 10 models you built for your model; use its error estimate on the held-out data
- ( ) pick any of the 10 models you built for your model; use the aver- age CV error for the 10 models as its error estimate
- (x) average all of the 10 models you got; use the average CV error as its error estimate

Correct Answer: **train a new model on the full data set, using the θ you found; use the average CV error as its error estimate**

- ( ) average all of the 10 models you got; use the error the combined model gives on the full training set
- ( ) train a new model on the full data set, using the θ you found; use the average CV error as its error estimate

13 0.333 / 1 point

Suppose we train a binary classifier on some dataset. Suppose y is the set of true labels, and yˆ 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.

The number of true positives is ✓ 2. The number of false negatives is ✗ 4. The precision of our classifier is: ✗ 3/10

Correct Answer: **3**

Correct Answer: **1/3**

14 1 / 1 point

Suppose we want to compute 10-Fold Cross-Validation error on 100 training examples. We need to compute error N1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size N2, and test the model on the data of size N3. What are the appropriate numbers for N1, N2, N3?

- ✓ ◉ N1 = 10, N2 = 90, N3 = 10
- ○ N1 =1, N2 =90, N3 =10
- ○ N1 = 10, N2 = 100, N3 = 10
- ○ N1 = 10, N2 = 100, N3 = 100

Understanding "Positive" and "Negative" in Confusion Matrix Terms

In **True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN)**:

- **The words "Positive" and "Negative" refer to the model's prediction.**
- **The words "True" and "False" indicate whether the model was correct.**

Breaking it Down:

Term	What it Means
True Positive (TP)	Model **predicted 1** (positive) and was **correct** (true label was 1)
False Positive (FP)	Model **predicted 1** (positive) but was **wrong** (true label was actually 0)
True Negative (TN)	Model **predicted 0** (negative) and was **correct** (true label was 0)
False Negative (FN)	Model **predicted 0** (negative) but was **wrong** (true label was actually 1)

Quiz 3 - Tree Based Models

2 1 / 1 point

Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.

✓ ● True

○ False

Feedback

General Feedback

The number of iterations in boosting controls the complexity of the model, therefore, a model selection procedure like cross validation can be used to select the appropriate model complexity and reduce the possibility of overfitting.

3 0 / 1 point

We learn a classifier f by boosting weak learners h. The functional form of f's decision boundary is the same as h's, but with different parameters. (e.g., if h was a linear classifier, then f is also a linear classifier).

✗ ● True

Correct Answer: **False**

○ False

Feedback

General Feedback

For example, the functional form of a decision stump is a single axis-aligned split of the input space, but the functional form of the boosted classifier is linear combinations of decision stumps which can form a more complex (piecewise linear) decision boundary.

4 1 / 1 point

The depth of a learned decision tree can be larger than the number of training examples used to create the tree.

- True
- ✓ False

Feedback

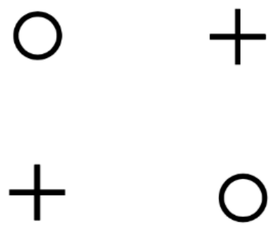
General Feedback

Each split of the tree must correspond to at least one training example, therefore, if there are n training examples, a path in the tree can have length at most n.

Note: There is a pathological situation in which the depth of a learned decision tree can be larger than number of training examples n - if the number of features is larger than n and there exist training examples which have same feature values but different labels. So, you may have been correct if you selected true and had in mind this explanation.

5 0 / 1 point

Consider the following dataset:



Select all classifiers that will achieve zero training error on this dataset. You may select more than one

- Logistic Regression
- ✓ Decision Tree of depth 2
- Decision Tree of depth 1
- ✗ k-NN classifier with k=3

Selected Answer - Incorrect

4. k-NN Classifier with k=3 ❌ (Incorrectly Selected)

- **k-Nearest Neighbors (k-NN) with k=3** classifies a point based on the **majority vote** of its 3 nearest neighbors.
- Since there are **only four points**, choosing **k=3** means every point considers **3 out of 4** total points to determine its class.
- Because of the checkerboard pattern:
  - Any point will have **two neighbors of the opposite class** and only **one neighbor of its own class**.
  - The majority vote (2 vs. 1) will assign the **wrong** label.
- This results in **misclassification**, meaning k-NN **fails to achieve zero training error**.
- **Conclusion: Incorrectly selected.** The correct answer would have been **not selecting k-NN**.

21

Training error here is the error you'll have when you input your training set to your KNN as test set. When K = 1, you'll choose the closest training sample to your test sample. Since your test sample is in the training dataset, it'll choose itself as the closest and never make mistake. For this reason, the training error will be zero when K = 1, irrespective of the dataset. There is one logical assumption here by the way, and that is your training set will not include same training samples belonging to different classes, i.e. conflicting information. Some real world datasets might have this property though.

6 1 / 1 point

In each round of Boosting (let's take Adaboost), the misclassification penalty for a particular training observation is increased going from round t to round t + 1 if the observation was:

A) classified incorrectly by the weak learner trained in round t
B) classified incorrectly by the full ensemble trained up to round t
C) classified incorrectly by a majority of the weak learners trained up to round t
D) B and C are correct
E) A, B, and C are correct

- ✓ ● A
- ○ B
- ○ C
- ○ D
- ○ E

7 3 / 3 points

For the following scenarios argue whether bagging and/or boosting would be your preferred choice and justify your choice shortly.

Noisy data : ✓ Bagging

A very big dataset: ✓ Bagging

Inbalanced data: ✓ Boosting

Feedback

General Feedback

- For noisy data: Bagging would be preferred since with the multiple splits of the data we can avoid too much noise. Boosting could amplify this noise.

- For a very big dataset: Bagging due to the ability to run things in parallel. With boosting we have to use the whole dataset

For inbalanced data: Boosting will do a good job - it can focus on the minority samples compared to bagging

8 1 / 2 points

Which of the following are advantages to using AdaBoost with *short* trees (e.g. max depth 4) over random forests with an equal number of tall trees (i.e. refined until the leaves are pure)? Select all that apply.

- ✓ [x] AdaBoost is better at reducing bias than a random forest
- [ ] AdaBoost is better at reducing variance than a random forest.
- ✗ [ ] AdaBoost is faster to train.

Missed option - incorrect

- [ ] AdaBoost is more robust against overfitting outliers in the training data.

Feedback

General Feedback

- Adaboost is design to reduce bias (improve accuracy)
- Bagging and randomizing the split choices are designed to reduce variance (in RF), however Adaboost doesn't actively reduce variance.
- We have the same number of trees for both examples and since Adaboost uses shorter trees, it is expected to be faster
- AdaBoost is expected to be worse at handling outliers because it will assign to them a high misclassification error, causing subsequent weak learners to focus on them.

9 1 / 2 points

Which of the following statements are true about bagging (note: not in a random forest; just bagging alone). Select all that apply.

- ✓ [x] Bagging without replacement is more likely to overfit than bagging with replacement
- ✓ [x] Even with bagging, sometimes decision trees still end up looking very similar
- ✗ [x] Bagging is often used with decision trees because it helps increase their training accuracy

Selected Answer - Incorrect

- [ ] Bagging involves using different learning algorithms on different subsamples of the training set

10 0 / 2 points

In functional gradient descent, each weak learner is trained to approximate the ✗ loss function

Correct Answer: **negative gradient of the loss function**

In gradient boosting with exponential loss , the negative gradient at each step is proportional to ✗ the residual y-f

Correct Answer: **the weighted probability of misclassification**

Feedback

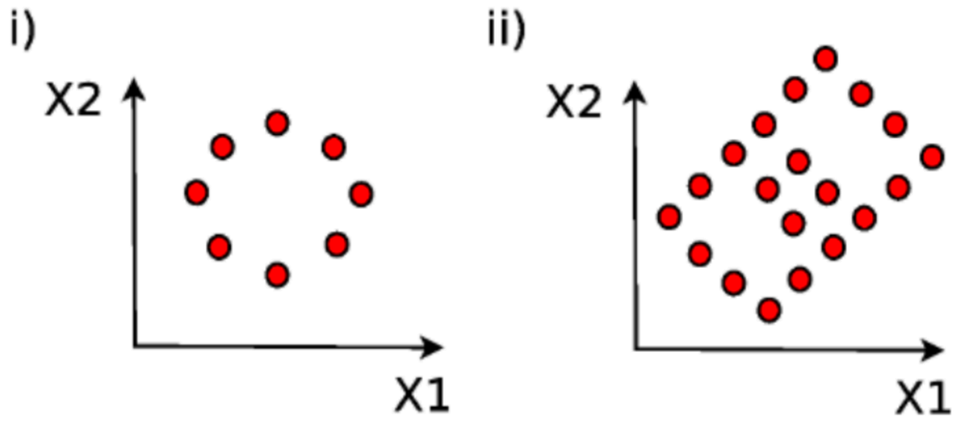
General Feedback

- In functional gradient descent, each weak learner is trained to **correct the errors of the previous model** by following the **steepest descent direction** of the loss function. This direction is given by the **negative gradient**, which tells the model how to adjust its predictions to minimize loss most effectively.
- The negative gradient of exponential loss is **largest for misclassified points** and **smallest for well-classified points**. This means the model assigns more weight to misclassified samples in the next iteration, just like AdaBoost. Since misclassification probability determines how much a point should be upweighted, the gradient is proportional to it

Quiz 4 - Timeseries and DR

We have two datasets (see below) that we apply PCA.

In which of the above cases will the first principal component capture a larger proportion of the variance in the data?



- I
- ✓ II
- both the same
- can't say without more information

Feedback

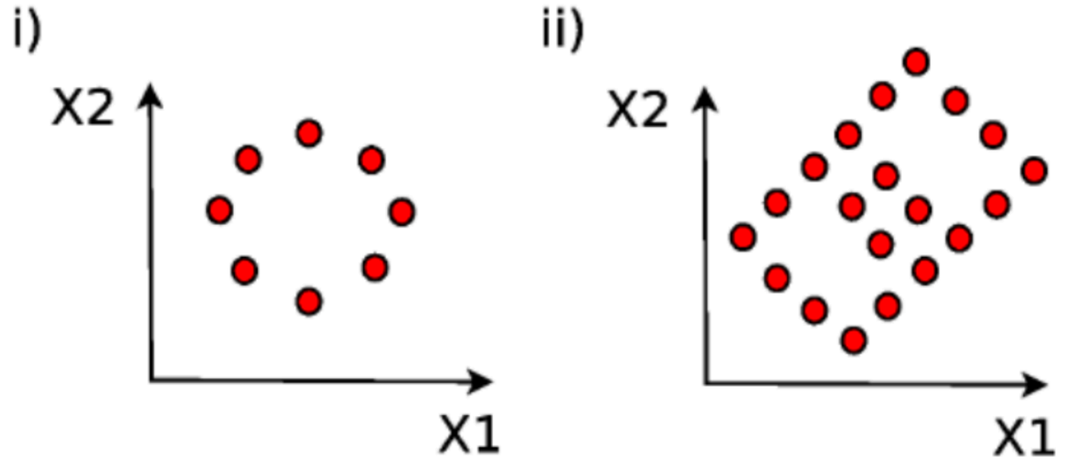
General Feedback

in case II variance on the 1st PC will be larger than in case I. The data is 2D, meaning that the direction of the largest variance will show the direction of the 1st PC

2 0 / 1 point

We have two datasets (see below) that we apply PCA.

In which of the above cases will the first two principal components (combined) capture a larger proportion of the variance?



- I
- ✕ II

Correct Answer: **both the same**

- both the same
- can't say without more information

Feedback

General Feedback

Data is 2-dimensional, meaning that with 2 PCs we can explain 100% of the variance

3 2 / 2 points

Why is PCA sometimes used as a preprocessing step before regression?

- ✓ ☑ To reduce overfitting by removing poorly predictive features
- ☐ For inference and scientific discovery, we prefer features that are not axis-aligned.
- ☐ To expose information missing from the input data.
- ✓ ☑ To make computation faster by reducing the dimensionality of the data.

4 1 / 2 points

A low-rank approximation of a matrix can be useful for

- ✓ ☑ matrix compression
- ✗ ☐ filling in unknown values

Missed option - incorrect

- ✓ ☑ removing noise
- ✗ ☐ discovering latent categories in the data

Missed option - incorrect

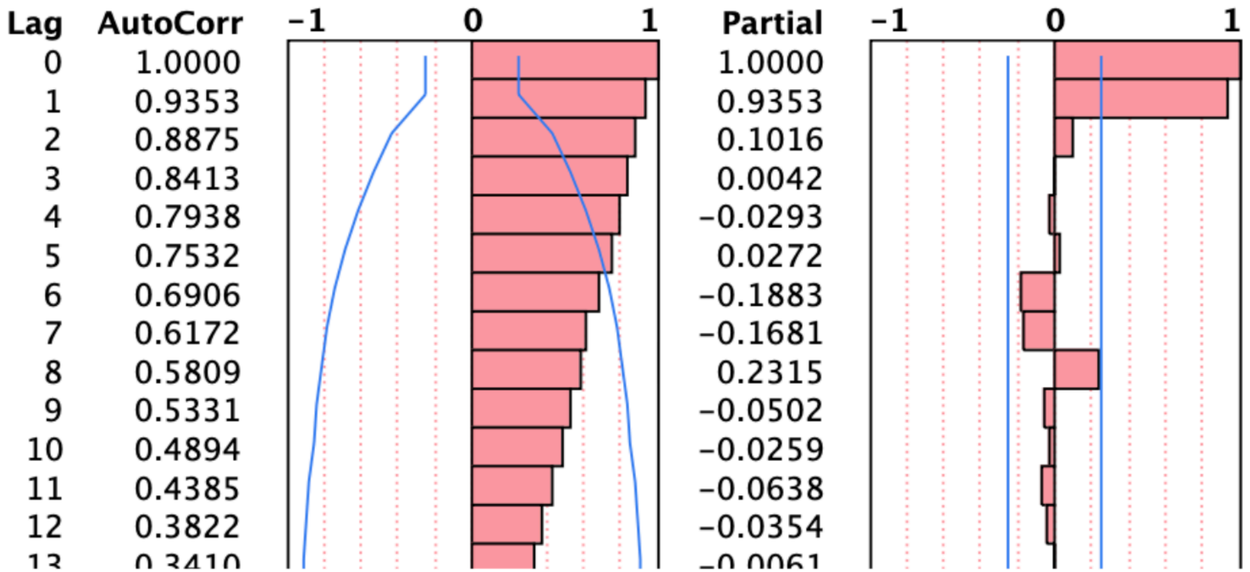
5 1 / 1 point

The largest eigenvector of the covariance matrix is the direction of minimum variance in the data.

- ○ True
- ✓ ◉ False

6 0 / 1 point

The figure below shows the estimated autocorrelation and partial autocorrelations of a time series of n=60 observations. Based on these plots, we should:



Lag	AutoCorr	Partial
0	1.0000	1.0000
1	0.9353	0.9353
2	0.8875	0.1016
3	0.8413	0.0042
4	0.7938	−0.0293
5	0.7532	0.0272
6	0.6906	−0.1883
7	0.6172	−0.1681
8	0.5809	0.2315
9	0.5331	−0.0502
10	0.4894	−0.0259
11	0.4385	−0.0638
12	0.3822	−0.0354
13	0.3410	−0.0061

- [ ] Transform data by taking logs
- [x] Difference the series to obtain stationary data
- [x] Fir an AR(1) model to the timeseries

Selected Answer - Incorrect

- [ ] Fir an MA(1) model to the timeseries
- [ ] Fit a linear trend to the timeseries

Missed option - incorrect

Feedback

General Feedback

Multiple answers might be correct here (well, some maybe more than others). The "gradual dying out" of the ACF suggests that there is a trend, so we have to remove it: In that sense, taking differences of fitting a model for the trend are logical choices.

Given the large lag at t=1, fitting an AR(1) model might be a reasonalbe choice - although would not be our first choice here.

7 1 / 1 point

Following the steps to run a PCA's algorithm, why is so important standardize your data?

- Standardize data allows other people understand better your work
- ✓ We control for variance due to e.g. different units
- Make the training time more fast
- It's a good pre-processing technique

8 0.667 / 1 point

Which of the following are true about principal components analysis (PCA)? Assume that no two eigenvectors of the sample covariance matrix have the same eigenvalue.

- ✓ [x] Appending a 1 to the end of every sample point doesn't change the results of performing PCA (except that the useful principal component vectors have an extra 0 at the end, and there's one extra useless component with eigenvalue zero).
- ✗ [ ] If you use PCA to project d-dimensional points down to j principal coordinates, and then you run PCA again to project those j-dimensional coordinates down to k principal coordinates, with d > j > k, you always get the same result as if you had just used PCA to project the d-dimensional points directly down to k principle coordinates.

Missed option - incorrect

- [ ] If you perform an arbitrary rigid rotation of the sample points as a group in feature space before performing PCA, the principal component directions do not change.
- ✓ [x] If you perform an arbitrary rigid rotation of the sample points as a group in feature space before performing PCA, the largest eigenvalue of the sample covariance matrix does not change.

Feedback

General Feedback

Appending an extra dimension with the same values introduces no variance in the extra dimension, so PCA will ignore that dimension. PCA discards the eigenvector directions associated with the largest eigenvalues; as the eigenvectors are mutually orthogonal, this does not affect the variance in the surviving dimensions, so your results depend solely on how many directions you discard. Rotating the sample points rotates the principal components, but it doesn't change the variance along each of those (rotated) component directions.

9 2 / 2 points

Which of the following statements about the Singular Value Decomposition (SVD) are true?

- [x] Every real matrix has an SVD.
- [ ] SVD and PCA always produce the same result
- [x] A matrix with rank r will have exactly r singular values that are greater than 0.

Does Seasonality Mean It's Within the Same Period?

Yes! **Seasonality means that a pattern repeats at fixed, predictable intervals**. These intervals are typically tied to natural or economic cycles, such as:

- **Daily (e.g., website traffic peaking during the day and dropping at night)**
- **Weekly (e.g., increased sales on weekends)**
- **Monthly (e.g., electricity usage rising in winter months)**
- **Yearly (e.g., holiday shopping spikes in December, flu season in winter)**

So, in this case, the **students' holiday pattern follows a yearly seasonality**.

Is Seasonality the Same as a Cycle?

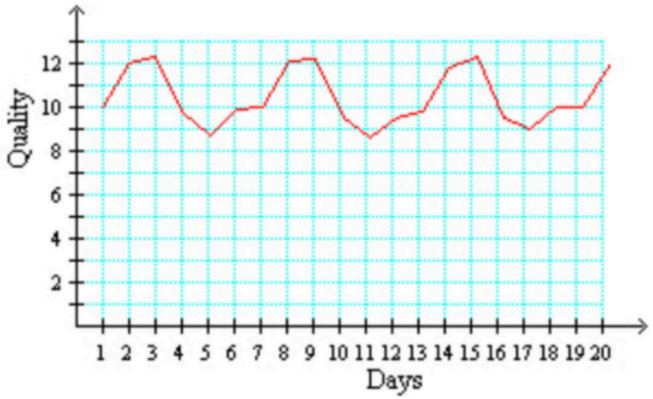
No, **seasonality and cycles are different**:

Feature	Seasonality	Cycle
Repetitive?	Yes ✅	Yes ✅
Fixed Intervals?	Yes (e.g., every year, every month) ✅	No (varies) ❌
Caused by?	Natural or human-made calendar effects (seasons, holidays)	Economic or external forces (business cycles, market trends)
Example	Summer travel peaks every July	Stock market fluctuations over several years

Key Difference:

- **Seasonality has fixed, regular intervals (e.g., every year, every month).**
- **Cycles occur at irregular intervals and are often influenced by external factors (e.g., economic recessions, business cycles).**

This timeseries plot is an example of



- random variation
- a positive trend
- a cycle trend (that we cannot identify)
- ✓ seasonal variation

Key Takeaways

- **Autocorrelation measures how past values in a time series influence future values.**
- It is **used in time series forecasting** to identify repeating patterns or seasonality.
- It is different from **cross-correlation**, which measures relationships between two different series.

Let me know if you need more clarification! 😊

12 1 / 1 point

What does autocorrelation (or autocovariance) measure?

- Linear dependence between multiple points on the different series observed at different times
- Linear dependence between two points on different series observed at same time
- ✓ Linear dependence between two points on the same series observed at different times
- none of the above

Understanding Autocorrelation (or Autocovariance)

Autocorrelation measures how related a time series is to itself at different time points.
In simpler terms, it checks whether past values in the series influence future values.

Why is the correct answer:

✅ **"Linear dependence between two points on the same series observed at different times"?**

- **Autocorrelation (ACF) analyzes a single time series** and examines how much the current value depends on its past values.
- It does this by **computing the correlation between a value at time t and a value at a previous time $t - k$ (with a lag of k).**
- If there is a high correlation at some lags, this suggests a **pattern or trend in the time series**.

Example: Temperature Forecasting

If you measure the **temperature** every day and see that today's temperature is highly correlated with yesterday's temperature, this suggests **a strong autocorrelation** in the series.

13 1 / 1 point

Imagine, you are working on a time series dataset. Your manager has asked you to build a highly accurate model. You started to build two types of models which are given below.

Model 1: Decision Tree model (on the lagged data)

Model 2: Time series autoregression model

At the end of evaluation of these two models, you found that model 2 is better than model 1. What could be the possible reason for your inference?

- Model 1 will always be better than Model 2
- ✓ Model 1 couldn't map the linear relationship as good as Model 2
- You can't compare decision tree with time series regression
- None of the listed options is true

Why is this the Correct Answer?

1. **Decision Trees are Poor at Capturing Linear Trends**
   - Decision Trees are **piecewise models**, meaning they **split** data at certain points but do not naturally model continuous relationships.
   - They work best with **categorical or highly non-linear relationships**, but they struggle to approximate smooth **linear dependencies**.
2. **Autoregressive (AR) Models are Designed for Linear Time Series**
   - AR models are specifically designed to capture **linear dependencies** over time.
   - They use past values (**lags**) of the same variable to predict the next value, making them idea for **time-dependent patterns**.
3. **If the Time Series is Mostly Linear, Decision Trees Struggle**
   - If the data follows a **strong linear pattern**, a decision tree will **fail to fit the trend well**, while an AR model will **capture it naturally**.

Why Are the Other Options Incorrect?

❌ **"Model 1 will always be better than Model 2"**

- **False.** Decision Trees are **not always superior** to autoregressive models, especially in linear time series forecasting.

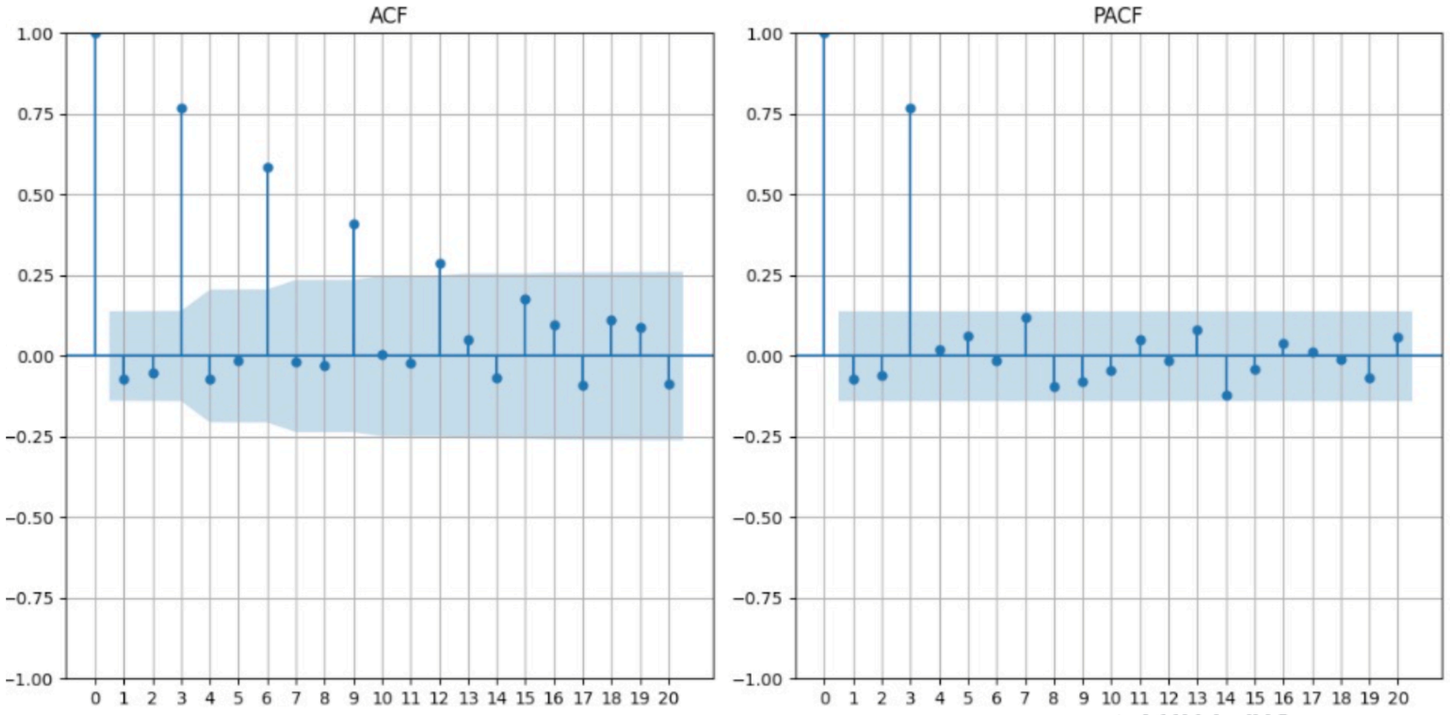
❌ **"You can't compare decision trees with time series regression"**

- **False.** You **can** compare them because both can be used for forecasting.
- Decision trees can be applied to **lagged features**, while AR models are directly designed for **time-dependent data**.

❌ **"None of the listed options is true"**

- **False.** The correct reason is that **Decision Trees struggle with linear relationships**, making Model 2 (AR) better in this case.

Which ARIMA model can generate such ACF/PACF plots?



- ARIMA(3,0,0)
- ARIMA(0,0,3)
- ✕ ARIMA(3,0,3)

Correct Answer: **ARIMA(3,0,0)**

- can't say

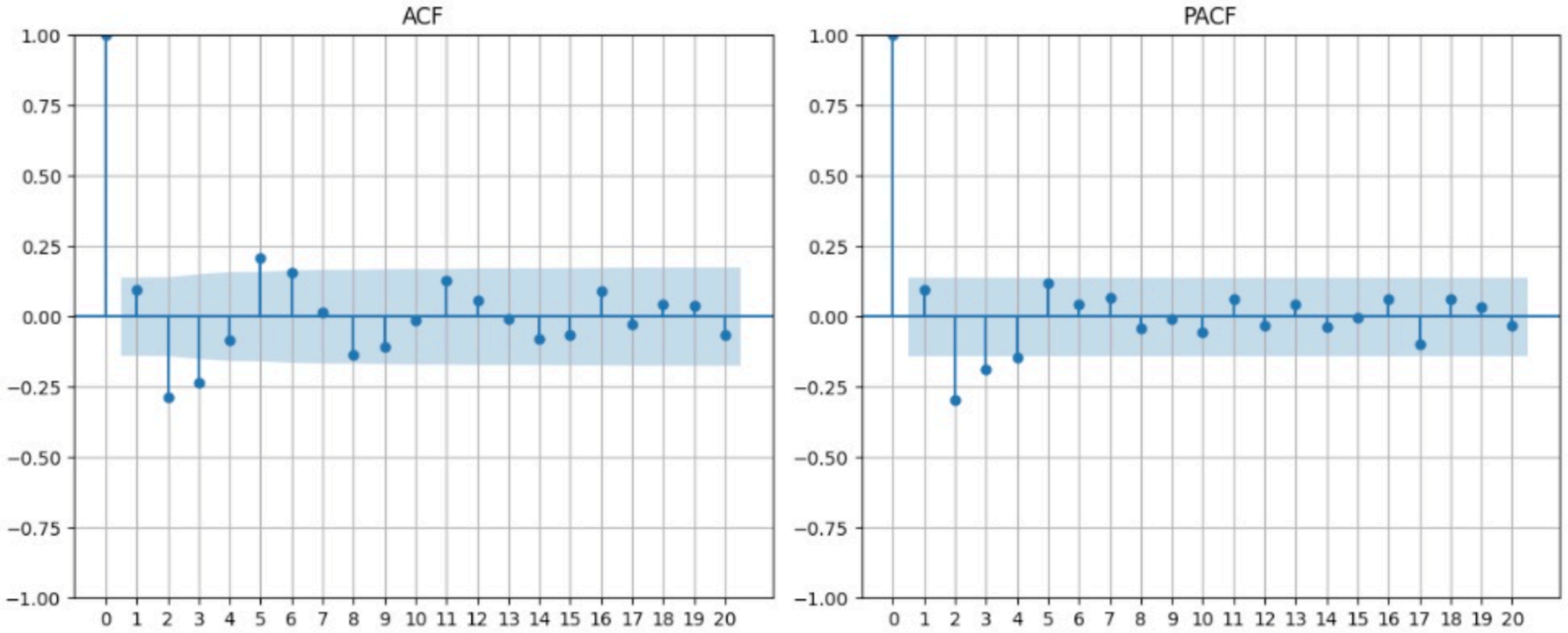
Feedback

General Feedback

- **PACF:** Clear cut-off at lag 3 --> suggesting AR terms
- **ACF:** Smooth decay, lags at 3xn because of the AR terms

15 0 / 1 point

Which ARIMA model can generate such ACF/PACF plots?



× ARIMA(2,0,0)

Correct Answer: **ARIMA(2,0,2)**

- ARIMA(0,0,2)
- ARIMA(2,0,2)
- ARIMA(1,0,1)
- ARIMA(1,0,0)
- ARIMA(0,0,1)

Feedback

General Feedback

Tricky selection (and more options might be correct, although this was generated by an ARIMA (2,0,2).

Since the ACF, PACF do not exhibit any clear behavior, we cannot pick either a sole AR or MA model. A mix of the two is suggested.

Because we see the peaks at lag 2 for both, that might suggest an ARIMA(2,0,2) and not an ARIMA(1,0,1).