# PS3

### September 26, 2025

- 0.1 Problem set 3
- 0.2 Name: [Yawen Tan]
- 0.3 Link to your PS3 github repo: [https://github.com/IsabellaTan/Brown-DATA1030-HW3.git]
- 0.3.1 Problem 0

-2 points for every missing green OK sign.

Make sure you are in the DATA1030 environment.

```
[1]: from __future__ import print_function
     from packaging.version import parse as Version
     from platform import python_version
     OK = ' \times 1b[42m[OK] \times 1b[Om']
     FAIL = "\x1b[41m[FAIL]\x1b[0m"]
     try:
         import importlib
     except ImportError:
         print(FAIL, "Python version 3.12.10 is required,"
                     " but %s is installed." % sys.version)
     def import_version(pkg, min_ver, fail_msg=""):
         mod = None
         try:
             mod = importlib.import_module(pkg)
             if pkg in {'PIL'}:
                 ver = mod.VERSION
             else:
                 ver = mod.__version__
             if Version(ver) == Version(min_ver):
                 print(OK, "%s version %s is installed."
                        % (lib, min_ver))
             else:
                 print(FAIL, "%s version %s is required, but %s installed."
                       % (lib, min_ver, ver))
```

```
except ImportError:
        print(FAIL, '%s not installed. %s' % (pkg, fail_msg))
    return mod
# first check the python version
pyversion = Version(python_version())
if pyversion >= Version("3.12.10"):
    print(OK, "Python version is %s" % pyversion)
elif pyversion < Version("3.12.10"):</pre>
    print(FAIL, "Python version 3.12.10 is required,"
                " but %s is installed." % pyversion)
else:
    print(FAIL, "Unknown Python version: %s" % pyversion)
print()
requirements = {'numpy': "2.2.5", 'matplotlib': "3.10.1", 'sklearn': "1.6.1",
                'pandas': "2.2.3", 'xgboost': "3.0.0", 'shap': "0.47.2",
                'polars': "1.27.1", 'seaborn': "0.13.2"}
# now the dependencies
for lib, required version in list(requirements.items()):
    import_version(lib, required_version)
```

#### OK Python version is 3.12.10

```
[ OK ] numpy version 2.2.5 is installed.
[ OK ] matplotlib version 3.10.1 is installed.
[ OK ] sklearn version 1.6.1 is installed.
[ OK ] pandas version 2.2.3 is installed.
[ OK ] xgboost version 3.0.0 is installed.
[ OK ] shap version 0.47.2 is installed.
[ OK ] polars version 1.27.1 is installed.
[ OK ] seaborn version 0.13.2 is installed.
```

#### 0.4 Problem 1: EDA and visualizations

### 0.4.1 Problem 1a: EDA (5 points)

One of the datasets we will be working with this semester is the kaggle house price dataset. The goal of PS3 is to use this dataset to practice dataframe manipulations and perform EDA. The dataset, and its description, are located in the data folder.

Carefully read the dataset description. Whenever you work with a dataset, it is highly recommended that you prepare a similar description if it is not readily available. Specific things to note:

• each feature is described in full detail,

- the meaning of continuous features is explained and their unit is provided (e.g., lot size is measured in square feet),
- each category in a categorical or ordinal feature is spelled out and explained.

Answer the following EDA-related questions.

The sequence of questions here are typical things to ask when you perform EDA on a new dataset. First, you always want to know how many data points and features you have, and whether they are continuous, ordinal, or categorical. You should then take a closer look at the target variable. We will study the properties of the features and the relationships between the features and the target variable in 1b.

Q0 First, read the data into a data frame and display the columns of the data frame below. You might encounter error messages and other issues along the way. Please diagnose and resolve them.

```
[]: # your code here
     import pandas as pd
     df = pd.read_excel('C:/Users/DELL/Desktop/zy/OneDrive/Brown/DATA1030/
      assignment3/Brown-DATA1030-HW3/data/train.xlsx', sheet_name='data')
     print(df.columns)
    Index(['MSSubClass', 'MSZoning', 'LotFrontage', 'LotArea', 'Street', 'Alley',
           'LotShape', 'LandContour', 'Utilities', 'LotConfig', 'LandSlope',
           'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle',
           'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'RoofStyle',
           'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'MasVnrArea',
           'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual', 'BsmtCond',
           'BsmtExposure', 'BsmtFinType1', 'BsmtFinSF1', 'BsmtFinType2',
           'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', 'Heating', 'HeatingQC',
           'CentralAir', 'Electrical', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF',
           'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath',
           'BedroomAbvGr', 'KitchenAbvGr', 'KitchenQual', 'TotRmsAbvGrd',
           'Functional', 'Fireplaces', 'FireplaceQu', 'GarageType', 'GarageYrBlt',
           'GarageFinish', 'GarageCars', 'GarageArea', 'GarageQual', 'GarageCond',
           'PavedDrive', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorch', '3SsnPorch',
           'ScreenPorch', 'PoolArea', 'PoolQC', 'Fence', 'MiscFeature', 'MiscVal',
           'MoSold', 'YrSold', 'SaleType', 'SaleCondition', 'SalePrice'],
          dtype='object')
```

Q1 How many rows and columns do we have in the dataframe?

```
[]: # your code here

print("The number of row is " + str(df.shape[0]))
print("The number of columb is " + str(df.shape[1]))
```

```
The number of row is 1460
The number of columb is 80
```

**Q2** What are the data types of the columns? Make sure that the output is not truncated and you

see the type of each column.

```
# your code here

# Create a for loop to iterate through each column and print its name and data
type
for col, dtype in df.dtypes.items():
    print(f"{col}: {dtype}")
```

MSSubClass: int64 MSZoning: object LotFrontage: float64

LotArea: int64 Street: object Alley: object LotShape: object LandContour: object Utilities: object LotConfig: object LandSlope: object Neighborhood: object Condition1: object Condition2: object BldgType: object HouseStyle: object OverallQual: int64 OverallCond: int64 YearBuilt: int64 YearRemodAdd: int64 RoofStyle: object RoofMatl: object Exterior1st: object Exterior2nd: object

BsmtQual: object
BsmtCond: object
BsmtExposure: object
BsmtFinType1: object
BsmtFinSF1: int64
BsmtFinType2: object
BsmtFinSF2: int64
BsmtUnfSF: int64
TotalBsmtSF: int64
Heating: object

MasVnrType: object MasVnrArea: float64 ExterQual: object ExterCond: object Foundation: object HeatingQC: object CentralAir: object Electrical: object 1stFlrSF: int64 2ndFlrSF: int64 LowQualFinSF: int64 GrLivArea: int64 BsmtFullBath: int64 BsmtHalfBath: int64 FullBath: int64 HalfBath: int64 BedroomAbvGr: int64 KitchenAbvGr: int64 KitchenQual: object TotRmsAbvGrd: int64 Functional: object Fireplaces: int64 FireplaceQu: object GarageType: object GarageYrBlt: float64 GarageFinish: object GarageCars: int64 GarageArea: int64 GarageQual: object GarageCond: object PavedDrive: object WoodDeckSF: int64 OpenPorchSF: int64 EnclosedPorch: int64 3SsnPorch: int64 ScreenPorch: int64 PoolArea: int64 PoolQC: object Fence: object

MiscFeature: object

MiscVal: int64 MoSold: int64 YrSold: int64 SaleType: object SaleCondition: object

SalePrice: int64

Q3 The ML target variable in this dataset is the sale price. We will develop ML pipelines to predict this variable based on the other features.

Is this column continuous or categorical? Please use .describe or .value\_counts to take a quick look at this feature.

```
[]: # your code here

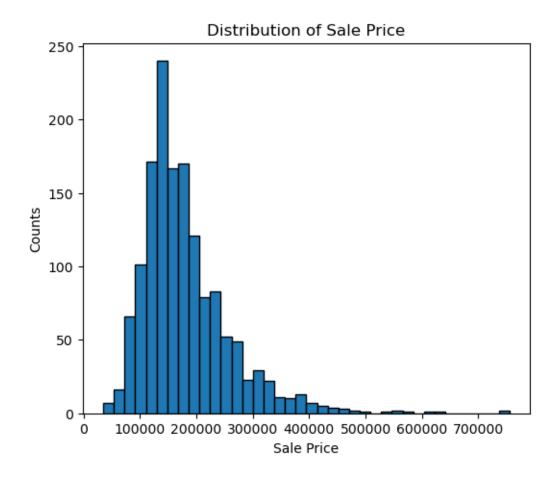
print(df["SalePrice"].describe())

# 'SalePrice' is countinuous variable
```

```
1460.000000
count
mean
         180921.195890
std
          79442.502883
min
          34900.000000
25%
         129975.000000
50%
         163000.000000
75%
         214000.000000
         755000.000000
max
Name: SalePrice, dtype: float64
```

Q4 Visualize the target variable. Don't forget the axis labels and graph title. Make sure to use appropriate arguments to best display the data.

```
[]: # your code here
     import matplotlib
     from matplotlib import pylab as plt
     import numpy as np
     # Set the figure size
     plt.figure(figsize=(6,5))
     # Create the histogram
     df['SalePrice'].plot.hist(
         bins = int(np.sqrt(df.shape[0])), # Let bins be the square root of the
      ⇔number of rows
         edgecolor='black', # Color of the edge of the bars
         linewidth=1 # Width of the edge of the bars
         )
     # Add labels and title
     plt.xlabel('Sale Price')
     plt.ylabel('Counts')
     plt.title('Distribution of Sale Price')
    plt.show()
```



## 0.4.2 Problem 1b: visualization (10 points)

Find one continuous, one ordinal, and one categorical feature that strongly correlates with the sale price. Create figures that illustrate your selected features and the sale price.

Don't forget to add axis labels and titles, and find appropriate arguments. Write figure captions to explain what the figure shows.

If you know how to quantitatively assess correlation strengths between variables, feel free to use those techniques. Qualitative/visual assessment also works for now.

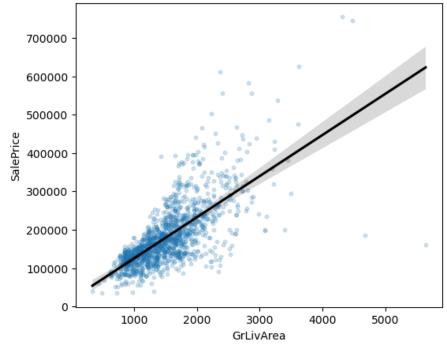
```
[]: # your code here
import matplotlib.pyplot as plt
import seaborn as sns

# Features that we choose:
# continuous feature: GrLivArea which is Above grade (ground) living area
square feet
# ordinal feature: OverallQual which is Overall material and finish of the house
# categorical feature: Alley which is Type of alley access to property
```

```
# Create a scatter plot for GrLivArea vs SalePrice
df.plot.scatter('GrLivArea', 'SalePrice', alpha=0.1, s=10, figsize=(6,5))
# Add labels and title
plt.xlabel('Above grade (ground) living area square feet')
plt.ylabel('Sale Price')
plt.title('Scatter plot of Sale Price vs Above grade (ground) living area⊔
⇔square feet', weight='bold')
# Add a regression line
sns.regplot(x='GrLivArea', y='SalePrice', data=df, scatter_kws={'alpha':0.1,__
 plt.show()
# Calculate the Pearson correlation coefficient
corr1 = round(df['GrLivArea'].corr(df['SalePrice']),2)
print('The Pearson correlation coefficient between GrLivArea and SalePrice is ⊔
print('Based on the scatter plot and the correlation coefficient, there is a_{\sqcup}
 \hookrightarrowstrong positive linear relationship between GrLivArea and SalePrice. As the \sqcup
 →above grade living area increases, the sale price tends to increase as well.
 ' )
# Create a dictionary to map the ordinal values to their corresponding labels
cond_labels = {
   1: 'Very Poor',
   2: 'Poor',
   3: 'Fair',
   4: 'Below Average',
   5: 'Average',
   6: 'Above Average',
   7: 'Good',
   8: 'Very Good',
   9: 'Excellent',
   10: 'Very Excellent'}
# Create a box plot for OverallQual vs SalePrice
df[['OverallQual','SalePrice']].boxplot(by='OverallQual',figsize=(5,5))
# Add labels and title
plt.ylabel('Sale Price')
plt.xlabel('Overall material and finish of the house')
```

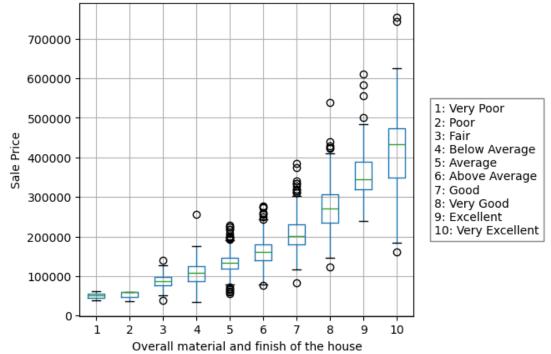
```
plt.title('Box plot of Sale Price vs Overall material and finish of the house', u
 ⇔weight='bold')
# Delete the automatic 'Boxplot grouped by group_by_column' title
plt.suptitle('')
# Add the mapping of ordinal values to their corresponding labels on the right_{\sqcup}
⇔side of the plot
textstr = '\n'.join([f"{k}: {v}" for k,v in sorted(cond_labels.items())])
plt.gcf().text(0.95, 0.5, textstr, fontsize=10, va='center',
 ⇒bbox=dict(facecolor='white', alpha=0.5))
plt.show()
# Calculate the Spearman correlation
corr2 = round(df['OverallQual'].corr(df['SalePrice'], method='spearman'),2)
print('The Spearman correlation coefficient between OverallQual and SalePrice⊔
 ⇔is ' + str(corr2))
print('Based on the box plot and the correlation coefficient, there is a strong⊔
 \hookrightarrowpositive monotonic relationship between OverallQual and SalePrice. As the \sqcup
 \hookrightarrowoverall quality of the house increases, the sale price tends to increase as \sqcup
 ⇔well.')
# Create a dictionary to map the categorical values to their corresponding □
 \hookrightarrow labels
alley_labels = {
    'Grvl': 'Gravel',
    'Pave': 'Paved'}
# Create a violin plot for Alley vs SalePrice
plt.figure(figsize=(5,5))
sns.violinplot(x='Alley', y='SalePrice', data=df, color='lightblue')
# Add labels and title
plt.ylabel('Sale Price')
plt.xlabel('Type of alley access to property')
plt.title('Violin plot of Sale Price vs Type of alley access to property', u
 ⇔weight='bold')
# Add the mapping of categorical values to their corresponding labels on the
⇔right side of the plot
textstr2 = '\n'.join([f"{k}: {v}" for k,v in sorted(alley_labels.items())])
plt.gcf().text(0.95, 0.5, textstr2, fontsize=10, va='center',
 ⇒bbox=dict(facecolor='white', alpha=0.5))
plt.show()
print('Based on the violin plot, houses with paved alley access tend to have⊔
 \hookrightarrowhigher sale prices compared to those with gravel alley access. The type of \sqcup
 →alley access appears to have an impact on the sale price of the house.')
```

Scatter plot of Sale Price vs Above grade (ground) living area square feet



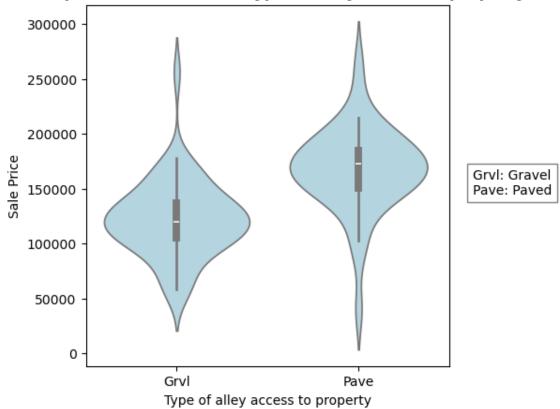
The Pearson correlation coefficient between GrLivArea and SalePrice is 0.71 Based on the scatter plot and the correlation coefficient, there is a strong positive linear relationship between GrLivArea and SalePrice. As the above grade living area increases, the sale price tends to increase as well.





The Spearman correlation coefficient between OverallQual and SalePrice is 0.81 Based on the box plot and the correlation coefficient, there is a strong positive monotonic relationship between OverallQual and SalePrice. As the overall quality of the house increases, the sale price tends to increase as well.





Based on the violin plot, houses with paved alley access tend to have higher sale prices compared to those with gravel alley access. The type of alley access appears to have an impact on the sale price of the house.

### 0.5 Problem 2: basic splitting strategy (15 points)

Write a general function that performs basic splitting on a dataset, while also conducting integrity tests on both its inputs and outputs. The function is called basic\_split, it is outlined in the cell below. It takes the following arguments as inputs: feature matrix (X), a target variable (y), train\_size, val\_size, test\_size, and random\_state. The output of the function should be: X\_train, y\_train, X\_val, y\_val, X\_test, y\_test, the three sets split according to the train, val, test sizes.

This function is general purpose, you'll be able to reuse it for any project if you want to perform basic split on your data.

```
[91]: def basic_split(feature_matrix,target_variable,train_size = 0.6,val_size=0.

$\times 2$,test_size=0.2,random_state=42):

Split dataframes (feature matrix X and target variable y) into random_

$\times train$, validation and test sets
```

```
Parameters:
   _____
  feature_matrix: a dataframe that contains your feature matrix
   target_variable: a series that contains your target variable
   train_size: a float between 0.0 and 1.0, it represents the proportion
       of the dataset to include in the training set
  val_size: a float between 0.0 and 1.0, it represents the proportion
       of the dataset to include in the validation set
   test_size: a float between 0.0 and 1.0, it represents the proportion
       of the dataset to include in the test set
  random_state: an int, it controls the shuffling applied to the data
       before applying the split
  NOTE: train_size+val_size+test_size must be equal to 1.
  Returns:
       a tuple containing the train, validation, and test sets
  Example:
   -----
  >>> import numpy as np
  >>> from sklearn.model_selection import train_test_split
  \Rightarrow > X, y = np.arange(10).reshape((5, 2)), range(5)
  >>>
  >>> X_train, y_train, X_val, y_val, X_test, y_test = basic_split(X,y)
   111
  # *********
  # TODO: test the inputs first
  # *********
  # write an if statement to perform each of these checks
  # Important: raise a ValueError with a descriptive error message if_{\sqcup}
⇔something is off
   \# call basic_split with incorrect arguments to make sure all of the tests_{\sqcup}
⇔work as intended!
  import pandas as pd
  import polars as pl
   # test 1: if feature_matrix is not a dataframe (pandas or polars), raise_
\hookrightarrow ValueError
  if not isinstance(feature_matrix, (pd.DataFrame, pl.DataFrame)):
      raise ValueError("Feature_matrix is not a dataframe (pandas or polars)")
```

```
# test 2: if the target-variable is not a series (pandas or polars), raise
\hookrightarrow ValueError
  if not isinstance(target_variable, (pd.Series, pl.Series)):
      raise ValueError("Target variable is not a series (pandas or polars)")
  # test 3: if the number of rows in feature_matrix is not equal to the
→length of target variable, raise a ValueError
  if feature_matrix.shape[0] != len(target_variable):
      raise ValueError("The number of rows in feature matrix is not equal to ...
→the length of target_variable")
  # test 4: if train_size is less than 0.0 or larger than 1.0, raise_
\hookrightarrow ValueError
  if train_size < 0.0 or train_size > 1.0:
      raise ValueError("train_size is less than 0.0 or larger than 1.0")
  # test 5: if val_size is less than 0.0 or larger than 1.0, raise ValueError
  if val_size < 0.0 or val_size > 1.0:
      raise ValueError("val_size is less than 0.0 or larger than 1.0")
  # test 6: if test_size is less than 0.0 or larger than 1.0, raise ValueError
  if test_size < 0.0 or test_size > 1.0:
      raise ValueError("test_size is less than 0.0 or larger than 1.0")
  # test 7: if train size+val size+test size is not equal to 1.0, raise
\hookrightarrow ValueError
  if train_size + val_size + test_size != 1.0:
      raise ValueError("train size+val size+test size is not equal to 1.0")
  # test 8: if random state is not an integer, raise ValueError
  if not isinstance(random_state, int):
      raise ValueError("random_state is not an integer")
  # ************
  # TODO: implement the splitting strategy
  # ***********
  # as we discussed in class, use sklearn's train_test_split twice
  from sklearn.model selection import train test split
  X_train, X_other, y_train, y_other = train_test_split(feature_matrix,_
utarget_variable, train_size=train_size, random_state=random_state)
  X_val, X_test, y_val, y_test = train_test_split(X_other, y_other,_
-test_size=test_size/(test_size + val_size), random_state=random_state)
  # *******
  # TODO: test the outputs
  # ********
  # same as above, write an if statement to perform each of these checks
  # raise a ValueError with a descriptive error message if something is off.
```

```
# test 1: the number of rows in X train divided by the number of rows in X_{\sqcup}
 ⇔should be close to train_size
    # think why we sometimes cannot achieve exact equality
    # and how you should express this as a condition in the if statement
    if not abs((X_train.shape[0] / feature_matrix.shape[0]) - train_size) < 0.</pre>
 ⇔05:
        raise ValueError("The number of rows in X_train divided by the number ⊔

→of rows in X is not close to train_size")
    # test 2: the number of rows in X_{-}val divided by the number of rows in X_{-}
 ⇔should be close to val size
    if not abs((X val.shape[0] / feature matrix.shape[0]) - val_size) < 0.05:</pre>
        raise ValueError("The number of rows in X val divided by the number of L
 ⇔rows in X is not close to val_size")
    # test 3: the number of rows in X_test divided by the number of rows in X_{\sqcup}
 ⇒should be close to test_size
    if not abs((X_test.shape[0] / feature_matrix.shape[0]) - test_size) < 0.05:</pre>
        raise ValueError("The number of rows in X test divided by the number of \sqcup
 →rows in X is not close to test_size")
    # test 4: make sure that the length of y_train, y_val, y_test is equal to
    # the number of rows in X_train, X_val, X_test, respectively
    if not (len(y_train) == X_train.shape[0] and len(y_val) == X_val.shape[0]_u
 →and len(y_test) == X_test.shape[0]):
        raise ValueError("The length of y_train, y_val, y_test is not equal to⊔
 →the number of rows in X_train, X_val, X_test, respectively")
    return X_train, y_train, X_val, y_val, X_test, y_test
# Call the function and preform tests here
# test 1: Apply the function to the house price dataset from problem 1 with
\Rightarrow train_size = 0.6, val_size = 0.2, and test_size = 0.2.
X = df.drop(columns=['SalePrice'])
y = df['SalePrice']
X_train, y_train, X_val, y_val, X_test, y_test = basic_split(X,y)
# test 2: Print out the head of X train, X val, and X test.
print("Train/Val/Test sizes:", len(X_train), len(X_val), len(X_test))
print("X_train head:\n", X_train.head())
print("X_val head:\n", X_val.head())
print("X_test head:\n", X_test.head())
# test 3: Make sure that you get the same points in each set every time you_
 \rightarrowrerun the cell (a.k.a., test for reproducability).
X_train1, y_train1, X_val1, y_val1, X_test1, y_test1 = basic_split(X, y)
X_train2, y_train2, X_val2, y_val2, X_test2, y_test2 = basic_split(X, y)
print('Checking reproducibility:')
print('If X_train is the same:', X_train1.equals(X_train2))
print('If y_train is the same:', y_train1.equals(y_train2))
```

```
print('If X_val is the same:', X_val1.equals(X_val2))
print('If y_val is the same:', y_val1.equals(y_val2))
print('If X_test is the same:', X_test1.equals(X_test2))
print('If y_test is the same:', y_test1.equals(y_test2))
# test 4: Try a couple of other train, val, test sizes here. make sure to test \Box
 ⇔the possible extreme values!
X_train_extrem, y_train_extrem, X_val_extrem, y_val_extrem, X_test_extrem,_

y_test_extrem = basic_split(X, y, train_size = 0.8,val_size=0.1,test_size=0.
 →1)
X train_extrem2, y_train_extrem2, X val_extrem2, y_val_extrem2, X_test_extrem2, \( \text{\text} \)
 \hookrightarrow05,test size=0.05)
# notice how most of the lines in basic split are about testing the inputs and
 ⇔outputs and
# testing the inputs ensures that the user correctly calls the function. if \Box
 they do not, a descriptive error message is returned.
# testing the outputs ensures that your code correctly performs the intended,
 operation anticipating edge cases and potential issues.
# this is pretty typical in software engineering and this is the key to writing
 ⇔reusable code.
Train/Val/Test sizes: 876 292 292
     MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape \
```

Х	train	head:

	Masubclass	s MPTOIITING	LOCLI	ontage	Lotarea	prieer	иттеу	LOCOIIS	ipe /	
314	70	RM		60.0	9600	Pave	Grvl	Re	eg .	
442	50	RM		52.0	6240	Pave	NaN	Re	eg	
319	80	RL		NaN	14115	Pave	NaN	Re	eg	
767	50	RL		75.0	12508	Pave	${\tt NaN}$	IF	R1	
756	60	RL		68.0	10769	Pave	${\tt NaN}$	IF	R1	
	${\tt LandContour}$	Utilities	LotCon	fig …	ScreenPo	cch Pool	lArea :	PoolQC	Fence	\
314	Lvl	AllPub	Ins	ide		0	0	NaN	${\tt NaN}$	
442	Lvl	AllPub	Ins	ide		0	0	NaN	${\tt NaN}$	
319	Lvl	AllPub	Ins	ide		0	0	NaN	${\tt NaN}$	
767	Lvl	AllPub	Ins	ide		0	0	NaN	${\tt NaN}$	
756	Lvl	AllPub	Ins	ide		0	0	NaN	${\tt NaN}$	
	${\tt MiscFeature}$	MiscVal N	MoSold	YrSold	SaleTyp	pe Sal	eCondi	tion		
314	NaN	0	8	2006		√D	No	rmal		
442	NaN	0	6	2008	3 7	<b>V</b> D	No	rmal		

2009

2008

2009

6

7

[5 rows x 79 columns] X\_val head:

NaN

Shed

NaN

0

1300

319

767

756

WD

WD

WD

Normal

Normal

Normal

	MSSubClass	s MSZoning	LotFr	ontage	LotArea	Street	Alley	LotSha	.pe \	
1336	90	RL		87.0	9246	Pave	NaN	IR	.1	
178	20	RL		63.0	17423	Pave	NaN	IR	.1	
619	60	RL		85.0	12244	Pave	NaN	Re	g	
548	20	RM		49.0	8235	Pave	NaN	IR	.1	
1046	60	RL		85.0	16056	Pave	NaN	IR	.1	
I	LandContour	Utilities	LotCon	fig S	ScreenPo	rch Poo	lArea i	PoolQC	Fence	\
1336	Lvl	AllPub	Ins	ide …		0	0	NaN	NaN	
178	Lvl	AllPub	CulD	Sac		0	0	NaN	NaN	
619	Lvl	AllPub	Ins	ide …		0	0	NaN	NaN	
548	HLS	AllPub	Ins	ide …		0	0	NaN	${\tt MnPrv}$	
1046	Lvl	AllPub	Ins	ide …		0	0	NaN	NaN	
N	MiscFeature	MiscVal N	MoSold	YrSold	SaleTy	pe Sal	eCondi	tion		
1336	NaN	0	11	2008	7	٧D	No	rmal		
178	NaN	0	7	2009	Ne	w	Par	tial		
619	NaN	0	8	2008	7	٧D	No	rmal		
548	NaN	0	6	2008	7	٧D	No	rmal		
1046	NaN	0	7	2006	Ne	ew	Par	tial		
[5 rot	ws x 79 colu									
LO TO	WB X 19 COIL	mmsl								
	t head:	mmsj								
			LotFr	ontage	LotArea	Street	Alley	LotSha	.pe \	
	t head:		LotFr	ontage NaN	LotArea 14112	Street Pave	Alley NaN	LotSha IR	-	
X_test	t head: MSSubClass	s MSZoning	LotFr	_	14112		•		.1	
X_test 1290	t head: MSSubClass 80	s MSZoning RL	LotFr	NaN	14112	Pave	NaN NaN	IR	.1 eg	
X_test 1290 282	t head: MSSubClass 80 120	s MSZoning RL RL	LotFr	NaN 34.0 90.0 88.0	14112 5063	Pave Pave	NaN NaN	IR Re	.1 g g	
X_test 1290 282 836	t head: MSSubClass 80 120 30	s MSZoning RL RL RM	LotFr	NaN 34.0 90.0	14112 5063 8100	Pave Pave Pave	NaN NaN Pave	IR Re Re	.1 eg eg .1	
X_test 1290 282 836 594 544	t head: MSSubClass 80 120 30 20 60	s MSZoning RL RL RM RL RL		NaN 34.0 90.0 88.0 58.0	14112 5063 8100 7990 17104	Pave Pave Pave Pave	NaN NaN Pave NaN NaN	IR Re Re IR	25 25 26 27 21	
X_test 1290 282 836 594 544	t head: MSSubClass 80 120 30 20	s MSZoning RL RL RM RL RL Utilities		NaN 34.0 90.0 88.0 58.0	14112 5063 8100 7990 17104	Pave Pave Pave Pave	NaN NaN Pave NaN NaN	IR Re Re IR	25 25 26 27 21	\
X_test 1290 282 836 594 544	t head:  MSSubClass  80  120  30  20  60  LandContour  Lv1	s MSZoning RL RL RM RL RL Utilities AllPub	LotCon Cor:	NaN 34.0 90.0 88.0 58.0 fig S	14112 5063 8100 7990 17104	Pave Pave Pave Pave	NaN NaN Pave NaN NaN	IR Re Re IR	25 25 26 27 21	\
X_test 1290 282 836 594 544	t head:  MSSubClass  80  120  30  20  60  LandContour	s MSZoning RL RL RM RL RL Utilities	LotCon Cor Ins	NaN 34.0 90.0 88.0 58.0 fig S	14112 5063 8100 7990 17104	Pave Pave Pave Pave Pave	NaN NaN Pave NaN NaN	IR Re Re IR IR PoolQC NaN NaN	eg eg .1 .1 .1 Fence NaN	\
1290 282 836 594 544 1290 282 836	t head:  MSSubClass  80  120  30  20  60  LandContour  Lv1	s MSZoning RL RL RM RL RL Utilities AllPub	LotCon Cor Ins Ins	NaN 34.0 90.0 88.0 58.0 fig ? ner ide	14112 5063 8100 7990 17104	Pave Pave Pave Pave Pave Cch Pool	NaN NaN Pave NaN NaN	IR Re Re IR IR PoolQC NaN	g g g .1 .1 Fence	\
1290 282 836 594 544 1290 282 836 594	t head:  MSSubClass  80  120  30  20  60  LandContour  Lvl  Lvl  Lvl  Lvl	MSZoning RL RL RM RL RL Utilities AllPub AllPub AllPub	LotCon Cor Ins Ins Ins	NaN 34.0 90.0 88.0 58.0 fig s ner ide ide	14112 5063 8100 7990 17104	Pave Pave Pave Pave Pave Ch Pool	NaN NaN Pave NaN NaN  LArea  0 0	IR Re Re IR IR PoolQC NaN NaN NaN	g g g d d f f ence NaN NaN GdWo MnPrv	\
1290 282 836 594 544 1290 282 836	t head:  MSSubClass  80  120  30  20  60  LandContour  Lvl  Lvl  Lvl	s MSZoning RL RL RM RL RL Utilities AllPub AllPub	LotCon Cor Ins Ins	NaN 34.0 90.0 88.0 58.0 fig s ner ide ide	14112 5063 8100 7990 17104	Pave Pave Pave Pave Pave Cch Pool 0 0	NaN NaN Pave NaN NaN  LArea  0 0 0	IR Re Re IR IR PoolQC NaN NaN	gg gg .1 .1 Fence NaN NaN GdWo	\
1290 282 836 594 544 1290 282 836 594 544	t head:  MSSubClass  80  120  30  20  60  LandContour  Lvl  Lvl  Lvl  Lvl  Lvl	MSZoning RL RM RL RL Utilities AllPub AllPub AllPub AllPub	LotCon Cor Ins Ins Ins Ins	NaN 34.0 90.0 88.0 58.0 fig 3 ide ide ide ide ide ide	14112 5063 8100 7990 17104 ScreenPor	Pave Pave Pave Pave Cch Pool 0 0 0 0	NaN NaN Pave NaN NaN  LArea  0 0 0 0	IR Re Re IR IR PoolQC NaN NaN NaN NaN	g g g d d f f ence NaN NaN GdWo MnPrv	\
1290 282 836 594 544 1290 282 836 594 544	t head:  MSSubClass  80  120  30  20  60  LandContour  Lvl  Lvl  Lvl  Lvl  Lvl  Lvl  MiscFeature	S MSZoning RL RL RM RL RL Utilities AllPub AllPub AllPub AllPub AllPub	LotCon Cor Ins Ins Ins Ins	NaN 34.0 90.0 88.0 58.0 fig 3 ner ide ide ide YrSold	14112 5063 8100 7990 17104 ScreenPor	Pave Pave Pave Pave Cch Pool 0 0 0 0 0 coe Sale	NaN NaN Pave NaN NaN  Area  0 0 0 0 0 0 0	IR Re Re IR IR PoolQC NaN NaN NaN NaN	g g g d d f f ence NaN NaN GdWo MnPrv	\
1290 282 836 594 544 1290 282 836 594 544	t head:  MSSubClass 80 120 30 20 60  LandContour Lvl Lvl Lvl Lvl Lvl NiscFeature NaN	S MSZoning RL RM RL RL Utilities AllPub AllPub AllPub AllPub AllPub AllPub AllPub	LotCon Cor Ins Ins Ins Ins 4oSold	NaN 34.0 90.0 88.0 58.0 fig s ner ide ide ide ide YrSold 2010	14112 5063 8100 7990 17104 ScreenPor	Pave Pave Pave Pave O O O O O O O O O O O O O O O O O O O	NaN NaN Pave NaN NaN  Area  0 0 0 0 0 0 0 0 0	IR Re Re IR IR PoolQC NaN NaN NaN NaN	g g g d d f f ence NaN NaN GdWo MnPrv	\
1290 282 836 594 544 1290 282 836 594 544	t head:  MSSubClass 80 120 30 20 60  LandContour Lvl Lvl Lvl Lvl Lvl NaN NaN	S MSZoning RL RM RL RL Vtilities AllPub AllPub AllPub AllPub AllPub O O	LotCon Cor Ins Ins Ins Ins 4 40Sold 4	NaN 34.0 90.0 88.0 58.0 fig 3 ner ide ide ide YrSold 2010 2009	14112 5063 8100 7990 17104 ScreenPon	Pave Pave Pave Pave Pave O O O O O O O D D D D S S S S S S S S S	NaN NaN Pave NaN NaN  LArea  0 0 0 0 0 0 No:	IR Re Re IR PoolQC NaN NaN NaN tion	g g g d d f f ence NaN NaN GdWo MnPrv	\
1290 282 836 594 544 1290 282 836 594 544	t head:  MSSubClass 80 120 30 20 60  LandContour Lvl Lvl Lvl Lvl Lvl NaN NaN	MSZoning RL RL RM RL RL Utilities AllPub AllPub AllPub AllPub AllPub O O O	LotCon Cor Ins Ins Ins Ins 4 4 4 6	NaN 34.0 90.0 88.0 58.0 fig s ner ide ide ide YrSold 2010 2009 2007	14112 5063 8100 7990 17104 ScreenPor	Pave Pave Pave Pave O O O O O O O D D D D D D D D D D D D	NaN NaN Pave NaN NaN  Area  0 0 0 0 0 0 0 No:	IR Re Re IR IR PoolQC NaN NaN NaN NaN tion rmal rmal rmal	g g g d d f f ence NaN NaN GdWo MnPrv	\
1290 282 836 594 544 1290 282 836 594 544	t head:  MSSubClass 80 120 30 20 60  LandContour Lvl Lvl Lvl Lvl Lvl NaN NaN	S MSZoning RL RM RL RL Vtilities AllPub AllPub AllPub AllPub AllPub O O	LotCon Cor Ins Ins Ins Ins 4 40Sold 4	NaN 34.0 90.0 88.0 58.0 fig 3 ner ide ide ide YrSold 2010 2009	14112 5063 8100 7990 17104 ScreenPor SaleTyj Conl	Pave Pave Pave Pave Pave O O O O O O O D D D D S S S S S S S S S	NaN NaN Pave NaN NaN  Area  0 0 0 0 0 0 0 0 No:	IR Re Re IR PoolQC NaN NaN NaN tion	g g g d d f f ence NaN NaN GdWo MnPrv	\

[5 rows x 79 columns]

Checking reproducibility:

If X\_train is the same: True

If y\_train is the same: True

```
If X_val is the same: True
If y_val is the same: True
If X_test is the same: True
If y_test is the same: True
```

# 1 Problem 3: kfold splitting (20 points)

Write a function that performs kfold splitting. We provided the input arguments and the header of the function. You need test the inputs, implement the algorithm, and test the outputs.

```
[]: def basic_kfold(feature_matrix, target_variable, k = 5, shuffle = True, u
      ⇒random state = 42):
        111
        Split dataframes (feature matrix X and target variable y) into \hat{k} number \hat{k}
      ⇔of equal size folds.
        One fold is used as the test set.
        Iterate over the remaining k-1 folds. Fold i is used as the validation set,
        the remaining folds are used as the training set.
        Parameters:
        feature_matrix: a dataframe that contains your feature matrix
        target_variable: a series that contains your target variable
        k: an int, the number of folds
        shuffle: boolean variable. If True, the feature matrix and the target \sqcup
      \neg variable are shuffled
            before the folds are created to randomize the sets
        random_state: an int, it controls the shuffling applied to the data
        Returns:
            →train and validation sets
        >>> import numpy as np
        >>> from sklearn.model_selection import train_test_split
        >>>
        \Rightarrow > X, y = np.arange(10).reshape((5, 2)), range(5)
        >>>
        >>> X_test, y_test, train_val_sets = basic_kfold(X,y)
        import pandas as pd
        import polars as pl
        import numpy as np
```

```
# one of the outputs, it will contain the train and validation sets
   # k-1 train and val -> tuple1(train, val), tuple2(train, val), .....
   \# train\_val\_sets = [(train1, val1) = ((x\_train, y\_train), (x\_val, y\_val))
                         (traink-1, valk-1)
  train_val_sets = []
  # ********
  # TODO: test the inputs!
  # ********
  \# test each of the input arguments. consider their types and what values \sqcup
→are possible as you come up with the tests
  # come up with at least 8 tests
  # among other things, consider what the smallest and largest k we can have
  # test 1: if feature matrix is not a dataframe (pandas or polars), raise
\hookrightarrow ValueError
  if not isinstance(feature_matrix, (pd.DataFrame, pl.DataFrame)):
      raise ValueError("Feature_matrix is not a dataframe (pandas or polars)")
   # test 2: if the target-variable is not a series (pandas or polars), raise_{\sqcup}
\hookrightarrow ValueError
  if not isinstance(target variable, (pd.Series, pl.Series)):
       raise ValueError("Target_variable is not a series (pandas or polars)")
   # test 3: if the number of rows in feature_matrix is not equal to the
→length of target_variable, raise a ValueError
  if feature_matrix.shape[0] != len(target_variable):
      raise ValueError("The number of rows in feature matrix is not equal to⊔
→the length of target_variable")
  # test 4: if k is not an integer, raise ValueError
  if not isinstance(k, int):
      raise ValueError("k is not an integer")
  # test 5: if shuffle is not a boolean, raise ValueError
  if not isinstance(shuffle, bool):
      raise ValueError("shuffle is not a boolean")
   # test 6: if random_state is not an integer, raise ValueError
  if not isinstance(random_state, int):
      raise ValueError("random_state is not an integer")
  # test 7: if k is less than 3 or larger than the number of rows in
⇔feature_matrix, raise ValueError
  if k < 3 or k > feature matrix.shape[0]:
      raise ValueError("k is less than 2 or larger than the number of rows in \Box
# test 8: if feature_matrix is empty, raise ValueError
  if feature_matrix.shape[0] == 0 or feature_matrix.shape[1] == 0:
      raise ValueError("feature_matrix is empty")
```

```
# test 9: if target_variable is empty, raise ValueError
  if len(target_variable) == 0:
      raise ValueError("target_variable is empty")
  # ************
  # TODO: implement the splitting strategy
  # **********
  # you can use numpy, pandas or polars. do not use sklearn here!
  n_samples = len(feature_matrix)
  # Shuffle indices if needed
  indices = np.arange(n_samples)
  if shuffle:
      rng = np.random.default_rng(seed=random_state)
      rng.shuffle(indices)
  # Split indices into k folds
  fold_sizes = [n_samples // k] * k
  for i in range(n_samples % k):
      fold_sizes[i] += 1 # distribute remainder
  folds = []
  start = 0
  for size in fold_sizes:
      folds.append(indices[start:start + size])
      start += size
  # Pick first fold as test set
  test_idx = folds[0]
  X_test = feature_matrix.iloc[test_idx]
  y_test = target_variable.iloc[test_idx]
  # Remaining folds for train/validation
  train_val_folds = folds[1:]
  # Prepare train val sets
  train_val_sets = []
  for i, val_idx in enumerate(train_val_folds):
      # validation set
      X_val = feature_matrix.iloc[val_idx]
      y_val = target_variable.iloc[val_idx]
      # training set = all other folds
      other_idx = np.hstack([train_val_folds[j] for j in_
→range(len(train_val_folds)) if j != i])
```

```
X_train = feature_matrix.iloc[other_idx]
      y_train = target_variable.iloc[other_idx]
       # append tuple ((X_train, y_train), (X_val, y_val))
      train_val_sets.append(((X_train, y_train), (X_val, y_val)))
   # ********
  # TODO: test the outputs!
   # ******
   # test 1: check whether each point is in exactly one set (no point is \Box
→duplicated and no point is left out)
  # All original sample indices
  all_indices = set(range(len(feature_matrix)))
  # Indices in the test set
  test_indices = set(X_test.index)
  # Indices in all train and validation sets
  train_val_indices = set()
  for train_set, val_set in train_val_sets:
      X train, y train = train set # get train features and labels
      X_val, y_val = val_set # get validation features and labels
      train_val_indices.update(X_train.index) # add train indices
      train_val_indices.update(X_val.index) # add validation indices
  # Check the intersection between test and train/val
  if test_indices & train_val_indices:
      raise ValueError("Some points are duplicated between test and train/
⇔validation sets!")
  # Check that all original samples are included
  if all_indices != test_indices.union(train_val_indices):
      raise ValueError("Some points are missing from the output sets!")
   # test 2: check whether you preserve the row-wise alignment between the
→feature matrix and the target variable in each set
   # hint: for row-wise alignment, it may be useful to note that row indices_
\hookrightarrowdo not change when subsetting dfs.
  # i.e. if a row is index id 10 in dataframe 1, it is still given the index_{\sqcup}
\hookrightarrow id of 10 in a subsetted df.
  # Loop over each train/validation fold
  for i, (train_set, val_set) in enumerate(train_val_sets):
      X_train, y_train = train_set # get features and labels for train set
      X_val, y_val = val_set # # get features and labels for validation set
       # Check that row indices match between features and labels in train set
      if not X_train.index.equals(y_train.index):
           raise ValueError(f"Train set row alignment broken in fold {i+1}")
```

```
# Check that row indices match between features and labels in
\hookrightarrow validation set
      if not X_val.index.equals(y_val.index):
          raise ValueError(f"Validation set row alignment broken in fold | 1
# Check that row indices match between features and labels in test set
  if not X_test.index.equals(y_test.index):
      raise ValueError("Test set row alignment broken!")
  # test 3: check whether the order of the columns is the same in each set.
  # We don't check y because y only has 1 column
  # Check train and validation
  # Loop over each train/validation fold
  for i, (train_set, val_set) in enumerate(train_val_sets):
      X_train, _ = train_set # get feature matrix from train set
      X_val, _ = val_set # get feature matrix from validation set
      # Check if train columns match original feature matrix
      if not X_train.columns.equals(feature_matrix.columns):
          raise ValueError(f"Train set column order broken in fold {i+1}")
      # Check if validation columns match original feature matrix
      if not X_val.columns.equals(feature_matrix.columns):
          raise ValueError(f"Validation set column order broken in fold⊔
# Check Test set
  if not X_test.columns.equals(feature_matrix.columns):
      raise ValueError("Test set column order broken!")
  # test 4: perform output tests similar to those in basic_split
  # Check if X_test is Dataframe
  if not isinstance (X_test, (pd.DataFrame,pl.DataFrame)):
      raise ValueError('The output X_test is not a Dataframe')
  # Check if y-test is Series
  if not isinstance(y_test, (pd.Series, pl.Series)):
      raise ValueError('The output y_test is not a Series')
  # Check if train val sets is List
  if not isinstance(train_val_sets, list):
      raise ValueError('The output train_val_sets is not a List')
  return X_test, y_test, train_val_sets
```

```
# TODO: call the function and preform tests here
# test 1: Apply the function to the house price dataset from problem 1 with 4
⇔folds.
X = df.drop(columns=['SalePrice'])
y = df['SalePrice']
X_test, y_test, train_val_sets = basic_kfold(X,y,k = 4)
# test 2: Print out the head of the sets.
# print the head of test sets
print('Head of X_test:')
print(X_test.head())
print('Head of y_test')
print(y_test.head())
# make for loop to print the head of training and validation sets for every fold
for fold_idx, (train_set, val_set) in enumerate(train_val_sets):
   X_train, y_train = train_set
   X_val, y_val = val_set
   print(f"=== Fold {fold_idx+1} ===")
   print("X_train head:")
   print(X_train.head())
   print("y_train head:")
   print(y_train.head())
   print("X_val head:")
   print(X_val.head())
   print("y_val head:")
   print(y_val.head())
   print("\n")
# test 3: Make sure that you get the same points in each set every time you_
\rightarrowrerun the cell (a.k.a., test for reproducability).
X_test1, y_test1, train_val_sets1 = basic_kfold(X, y)
X_test2, y_test2, train_val_sets2 = basic_kfold(X, y)
print('Checking reproducibility:')
print('If X_test is the same:', X_test1.equals(X_test2))
print('If y_test is the same:', y_test1.equals(y_test2))
```

```
# Loop over each fold in two separate runs of k-fold to check reproducibility
for i, ((train1, val1), (train2, val2)) in enumerate(zip(train_val_sets1,__

→train_val_sets2)):
    X train1, y train1 = train1
    X_val1, y_val1 = val1
    X train2, y train2 = train2
    X val2, y val2 = val2
    if not X_train1.index.equals(X_train2.index) or not y_train1.index.
  ⇔equals(y_train2.index):
         raise AssertionError(f"Fold {i+1} train indices differ!")
    if not X_val1.index.equals(X_val2.index) or not y_val1.index.equals(y_val2.
  ⇒index):
         raise AssertionError(f"Fold {i+1} validation indices differ!")
print("All train/validation fold indices are same.")
# test 4: Try a couple of other `k` values. test the extreme values!
X test_extreme, y_test_extreme, train_val_sets_extreme = basic_kfold(X,y,k = 3)
X_test_extreme2, y_test_extreme2, train_val_sets_extreme2 = basic_kfold(X,y,k =_
 \rightarrowlen(y))
Head of X_test:
      MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape \
1257
              30
                                             4060
                       RL
                                   56.0
                                                    Pave
                                                           NaN
                                                                     Reg
612
              60
                       R.I.
                                    NaN
                                           11885
                                                    Pave
                                                           NaN
                                                                    Reg
1270
              40
                       RL
                                    NaN
                                           23595
                                                    Pave
                                                           NaN
                                                                    Reg
1061
                  C (all)
                                  120.0
              30
                                           18000
                                                    Grvl
                                                           NaN
                                                                    Reg
184
              50
                       RL
                                   92.0
                                            7438
                                                    Pave
                                                                     IR1
                                                           NaN
     LandContour Utilities LotConfig LandSlope Neighborhood Condition1
1257
             Lvl
                    AllPub
                               Corner
                                            Gtl
                                                      Edwards
                                                                   Feedr
                    AllPub
                               Inside
612
             T.v.T
                                            Gt.1
                                                      CollgCr
                                                                    Norm
1270
             Low
                    AllPub
                               Inside
                                            Sev
                                                      ClearCr
                                                                     Norm
             Low
                    AllPub
                               Inside
                                                       IDOTRR
                                                                    Norm
1061
                                            Gt.1
184
             Lvl
                    AllPub
                               Inside
                                            Gtl
                                                      BrkSide
                                                                    RRAn
     Condition2 BldgType HouseStyle OverallQual OverallCond
                                                                 YearBuilt \
                              1Story
1257
           Norm
                    1Fam
                                                 5
                                                              8
                                                                       1922
                    1Fam
                                                 8
                                                              5
                                                                       2001
612
           Norm
                              2Story
                                                 7
                                                              6
1270
           Norm
                    1Fam
                              1Story
                                                                       1979
1061
           Norm
                    1Fam
                              1Story
                                                 3
                                                              4
                                                                       1935
                              1.5Fin
                                                 5
                                                              8
184
          Feedr
                    1Fam
                                                                       1908
      YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType \
1257
              1950
                       Gable CompShg
                                           Wd Sdng
                                                        Wd Sdng
                                                                        NaN
612
              2001
                                           VinylSd
                                                        VinylSd
                       Gable CompShg
                                                                   BrkFace
```

1270	1979		WdShake	Plywood	Plywoo		NaN	
1061	1950	) Gable	CompShg	MetalSd	MetalS	d	NaN	
184	1991	Gable	CompShg	AsbShng	Plywoo	od	NaN	
	Maayaan Aasaa T		C	.d.ti Dem	+01 D	.+0		
1057	MasVnrArea E				-			
1257 612	0.0 108.0	TA	TA	PConc	Fa	TA TA		
1270	0.0	Gd Gd	TA TA	PConc PConc	Gd Gd	TA		
1061	0.0	Fa	TA	CBlock	TA	TA		
184	0.0	ra TA	TA	PConc	Fa	TA		
104	0.0	IA	1 A	PCOIIC	Га	1 A		
	BsmtExposure	BsmtFinType1	l BsmtFinSH	F1 BsmtFinT	Type2 Bsm	tFinSF2	\	
1257	No	Unf	Ē	0	Unf	0		
612	Av	GLO	99	90	Unf	0		
1270	Gd	GLO	125	58	Unf	0		
1061	No	Unf	Ē	0	Unf	0		
184	No	Unf	Ē	0	Unf	0		
		CotalBsmtSF F	_	_				
1257	864	864	GasA	Ex	Y	SBrkr		
612	309	1299	${\tt GasA}$	Ex	Y	SBrkr		
1270	74	1332	${\tt GasA}$	TA	Y	SBrkr		
1061	894	894	${\tt GasA}$	TA	Y	SBrkr		
184	504	504	GasA	Gd	Y	SBrkr		
	1stFlrSF 2n	ndFlrSF Low(	DualFinSF (	GrLivArea	Rsmt.FulllE	Rath Rsm	tHalfRath	\
1257					BsmtFullE			
1257 612	864	0	0	864	BsmtFullE	0	0	
612	864 1299	0 573	0 0	864 1872	BsmtFullE	0 1	0	
612 1270	864 1299 1332	0 573 192	0 0 0	864 1872 1524	BsmtFullE	0 1 2	0 0 0	
612 1270 1061	864 1299 1332 894	0 573 192 0	0 0 0	864 1872 1524 894	BsmtFullE	0 1 2 0	0 0 0	
612 1270	864 1299 1332	0 573 192	0 0 0	864 1872 1524	BsmtFullE	0 1 2	0 0 0	
612 1270 1061	864 1299 1332 894 936	0 573 192 0 316	0 0 0 0	864 1872 1524 894		0 1 2 0 0	0 0 0	
612 1270 1061	864 1299 1332 894 936	0 573 192 0 316	0 0 0 0	864 1872 1524 894 1252		0 1 2 0 0	0 0 0	
612 1270 1061 184	864 1299 1332 894 936 FullBath Ha	0 573 192 0 316	0 0 0 0 0	864 1872 1524 894 1252	ir Kitcher	0 1 2 0 0	0 0 0	
612 1270 1061 184	864 1299 1332 894 936 FullBath Ha	0 573 192 0 316 alfBath Bedr	0 0 0 0 0 roomAbvGr F	864 1872 1524 894 1252	r Kitcher 1	0 1 2 0 0 0	0 0 0	
612 1270 1061 184 1257 612	864 1299 1332 894 936 FullBath Ha	0 573 192 0 316 alfBath Bedr 0	0 0 0 0 0 roomAbvGr P	864 1872 1524 894 1252	r Kitcher 1 1	0 1 2 0 0 .Qual \ TA Ex	0 0 0	
612 1270 1061 184 1257 612 1270	864 1299 1332 894 936 FullBath Ha 1 2 0	0 573 192 0 316 alfBath Bedr 0 1	0 0 0 0 0 coomAbvGr F	864 1872 1524 894 1252	Fr Kitcher 1 1	0 1 2 0 0 TA Ex Gd	0 0 0	
612 1270 1061 184 1257 612 1270 1061	864 1299 1332 894 936 FullBath Ha 1 2 0	0 573 192 0 316 alfBath Bedr 0 1 1 0	0 0 0 0 0 coomAbvGr P 2 3 0 2 3	864 1872 1524 894 1252 KitchenAbvG	Gr Kitcher 1 1 1 1	0 1 2 0 0 TA Ex Gd TA	0 0 0 0	
612 1270 1061 184 1257 612 1270 1061 184	864 1299 1332 894 936  FullBath Ha 1 2 0 1 1 TotRmsAbvGrd	0 573 192 0 316 alfBath Bedr 0 1 1 0 0	0 0 0 0 0 coomAbvGr F 2 3 0 2 3	864 1872 1524 894 1252 KitchenAbvG	Gr Kitcher 1 1 1 1 1 1 eQu Garage	0 1 2 0 0 TA Ex Gd TA TA TA TA	0 0 0 0 rageYrBlt	\
612 1270 1061 184 1257 612 1270 1061 184	864 1299 1332 894 936  FullBath Ha 1 2 0 1 1 TotRmsAbvGrd	0 573 192 0 316 alfBath Bedr 0 1 1 0 0	0 0 0 0 0 roomAbvGr F 2 3 0 2 3	864 1872 1524 894 1252 KitchenAbvG	Gr Kitcher 1 1 1 1 1 2Qu Garage	0 1 2 0 0 TA Ex Gd TA TA TA TA NaN	0 0 0 0 rageYrBlt NaN	
612 1270 1061 184 1257 612 1270 1061 184 1257 612	864 1299 1332 894 936  FullBath Ha 1 2 0 1 1 TotRmsAbvGrd	0 573 192 0 316 alfBath Bedr 0 1 1 0 0	0 0 0 0 0 coomAbvGr F 2 3 0 2 3 Fireplaces	864 1872 1524 894 1252 KitchenAbvG	Gr Kitcher 1 1 1 1 1 2Qu Garage JaN Bui	0 1 2 0 0 CQual \ TA Ex Gd TA Type Ga NaN .ltIn	0 0 0 0 0 0 rageYrBlt NaN 2001.0	\
612 1270 1061 184 1257 612 1270 1061 184 1257 612 1270	864 1299 1332 894 936  FullBath Ha 1 2 0 1 1 TotRmsAbvGrd	0 573 192 0 316 alfBath Bedr 0 1 1 0 0 0	0 0 0 0 0 roomAbvGr P 2 3 0 2 3 Fireplaces	864 1872 1524 894 1252 KitchenAbvG	Gr Kitcher 1 1 1 1 1 EQu Garage JaN TA Bui	0 1 2 0 0 TA Ex Gd TA TA TA A Type NaN ltIn tchd	0 0 0 0 0 0 1 2001.0 1979.0	\
612 1270 1061 184 1257 612 1270 1061 1257 612 1270 1061	864 1299 1332 894 936  FullBath Ha 1 2 0 1 1 TotRmsAbvGrd	0 573 192 0 316  alfBath Bedr 0 1 1 0 0 0  Functional Typ Typ Typ Typ Typ	0 0 0 0 0 roomAbvGr A 2 3 0 2 3 Fireplaces	864 1872 1524 894 1252 KitchenAbvG	Fr Kitcher 1 1 1 1 1 EQu Garage JaN TA Bui TA At	0 1 2 0 0 IQual \ TA Ex Gd TA TA TA SType Ga NaN ltIn ctchd	rageYrBlt NaN 2001.0 1979.0	\
612 1270 1061 184 1257 612 1270 1061 184 1257 612 1270	864 1299 1332 894 936  FullBath Ha 1 2 0 1 1 TotRmsAbvGrd	0 573 192 0 316  alfBath Bedr 0 1 1 0 0 0  Functional Typ Typ Typ Typ	0 0 0 0 0 roomAbvGr P 2 3 0 2 3 Fireplaces	864 1872 1524 894 1252 KitchenAbvG	Gr Kitcher 1 1 1 1 1 2Qu Garage JaN TA Bui TA At	0 1 2 0 0 TA Ex Gd TA TA TA A Type NaN ltIn tchd	0 0 0 0 0 0 1 2001.0 1979.0	\
612 1270 1061 184 1257 612 1270 1061 184 1257 612 1270 1061 184	864 1299 1332 894 936  FullBath Ha 1 2 0 1 1 TotRmsAbvGrd	0 573 192 0 316  alfBath Bedr 0 1 1 0 0 Vertical Typ	0 0 0 0 0 coomAbvGr P 2 3 0 2 3 Fireplaces	864 1872 1524 894 1252 KitchenAbvG	Gr Kitcher 1 1 1 1 1 1 EQu Garage JaN Bui TA Bui TA At JaN De	0 1 2 0 0 CQual \ TA Ex Gd TA TA TA CType Ga NaN Ottln ctchd	0 0 0 0 0 0 0 1979.0 1994.0 1986.0	\
612 1270 1061 184 1257 612 1270 1061 184 1257 612 1270 1061 184	864 1299 1332 894 936  FullBath Ha 1 2 0 1 1 TotRmsAbvGrd	0 573 192 0 316  alfBath Bedr 0 1 1 0 0 0  Functional Typ Typ Typ Typ Typ	0 0 0 0 0 0 coomAbvGr H 2 3 0 2 3 Fireplaces	864 1872 1524 894 1252 KitchenAbvG	Gr Kitcher  1  1  1  1  1  1  TA Bui  TA At  JaN De	0 1 2 0 0 CQual \ TA Ex Gd TA TA TA CType Ga NaN Ottln ctchd	0 0 0 0 0 0 0 1979.0 1994.0 1986.0	\

612 RF	n 2	53	1	TA		TA	Y
1270 Fi				TA		TA	Y
1061 RF				TA		TA	Y
184 Un				TA		TA	Y
WoodDeckSF	OpenPorchSF	EnclosedP	orch 38	SsnPorch	Scre	enPorch	\
1257 0	96	;	0	C	)	0	
612 160	122	!	0	C	)	0	
1270 268	0	1	0	C	)	0	
1061 0	20	1	0	C	)	0	
184 104	0	1	0	C	)	0	
	oolQC Fence					Sold Sa	
1257 0	NaN NaN	NaN		0		2009	WD
612 0	NaN NaN	NaN		0		2009	WD
1270 0	NaN NaN	NaN		0	4	2010	WD
1061 0	NaN NaN	Shed	. 56	30	8	2008	ConLD
184 0	NaN MnPrv	NaN		0	6	2006	WD
Q 1 Q 1:							
SaleConditi							
1257 Norm							
612 Norm							
1270 Norm							
1061 Norm							
184 Norm	al						
Head of y_test							
1257 99900							
612 261500							
1270 260000							
1061 81000							
184 127000							
Name: SalePrice,	dtype: int64	:					
=== Fold 1 ===							
X_train head:							
MSSubClass	MSZoning Lo	tFrontage	LotArea	Street	Alley	LotShap	e \
541 60	RL	NaN	11000	Pave	NaN	Re	g
341 20		60.0	8400	Pave	NaN	Re	_
1074 20		74.0	8556	Pave	NaN	Re	_
521 20		90.0	11988		NaN	IR	_
850 120		36.0	4435	Pave	NaN	Re	
120	1411	23.3	1100	1475	11411	100	5
LandContour	Utilities Lo	tConfig Lan	dSlope N	Veighbor	hood C	Condition	n1 \
541 Lvl	AllPub	FR2	Gtl	NoF	lidge	No	rm
341 Lvl	AllPub	Inside	Gtl	Sav	yerW	Fee	dr
1074 Lvl	AllPub	Inside	Gtl	Col	.lgCr	No	rm
521 Lvl	AllPub	Corner	Gtl		Ames	Fee	dr
850 Lvl		Inside	Gtl		lgCr	No	
					•		

	Condition2	BldgTvpe	HouseStyle	OverallQual	OverallCond	YearBuilt	\
541	Norm		2Story	8	5	2000	
341	Norm	1Fam	1Story	4	4	1950	
1074	Norm	1Fam	1Story	7	5	2006	
521	Norm	1Fam	1Story	6	6	1957	•
850	Norm	TwnhsE	1Story	6	5	2003	
			J				
	YearRemod	Add RoofSt	yle RoofMatl	L Exterior1st	Exterior2nd	MasVnrType	
541	2	000 Ga	ble CompShg	g VinylSd	VinylSd	BrkFace	!
341	1	950 Ga	ble CompShg	g Wd Sdng	AsbShng	NaN	
1074	2	006 Ga	ble CompShg	g VinylSd	VinylSd	NaN	
521	1	957 Ga	ble CompShg	g VinylSd	VinylSd	NaN	
850	2	003 Ga	ble CompShg	g VinylSd	VinylSd	BrkFace	
				Foundation B			
541	72.			PConc	Gd	TA	
341	0.	0 F		CBlock	TA	Fa	
1074				PConc	Gd	TA	
521	0.			CBlock	TA	TA	
850	170.	O G	d TA	PConc	Gd	TA	
	BsmtExposu			FinSF1 BsmtFi		FinSF2 \	
541		No	Unf	0	Unf	0	
341		No	Unf	0	Unf	0	
1074		Av	Unf	0	Unf	0	
521		No	Rec	777	Unf	0	
850		Av	GLQ	659	Unf	0	
	D . II COD	m			. 34: 53		
E 4 4	BsmtUnfSF		•	HeatingQC Cer			
541	969		969 GasA	Ex	Y	SBrkr	
341	721		721 GasA	Gd	Y	SBrkr	
1074	1240		240 GasA	Ex	Y	SBrkr	
521	467		244 GasA	Ex	Y	FuseA	
850	189	•	848 GasA	Ex	Y	SBrkr	
	1stFlrSF	2ndFlrSF	LowQualFin9	SF GrLivArea	BsmtFullBat	th BsmtHal	fBath \
541	997	1288	Londaari III	0 2285		0	0
341	841	0		0 841		0	0
1074		0		0 1240		0	0
521	1244	0		0 1244		0	0
850	848	0		0 848		1	0
	0.10	•		0 20		_	·
	FullBath	HalfBath	BedroomAbv(	Gr KitchenAb	vGr KitchenQı	ual \	
541	2	1		4	1	Gd	
341	1	0		2	1	TA	
1074		0		2	1	Gd	
521	1	1		3	1	TA	
850	1	0		1	1	Gd	

```
TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt \
                                                                               2000.0
541
                  8
                                                        TA
                                                               BuiltIn
                                            1
                             Тур
341
                  4
                             Тур
                                            0
                                                       NaN
                                                               CarPort
                                                                               1950.0
                  5
1074
                                            0
                                                       NaN
                                                                Attchd
                                                                               2006.0
                             Тур
                                            2
521
                  6
                                                        Gd
                                                                Attchd
                                                                               1957.0
                             Тур
                  3
850
                             Тур
                                            0
                                                       NaN
                                                                Attchd
                                                                               2003.0
     GarageFinish GarageCars
                                  GarageArea GarageQual GarageCond PavedDrive
541
               Fin
                                          648
                                                       TA
                                                                   TA
                               3
                                          294
341
               Unf
                               1
                                                       TΑ
                                                                   TA
                                                                                 N
1074
               RFn
                               3
                                          826
                                                       TΑ
                                                                   TA
                                                                                 Y
521
               Unf
                               1
                                          336
                                                       TΑ
                                                                   TA
                                                                                 Y
850
               Fin
                               2
                                                       TΑ
                                                                                 Υ
                                          420
                                                                   TA
      WoodDeckSF
                    OpenPorchSF
                                  {\tt EnclosedPorch}
                                                   3SsnPorch
                                                               ScreenPorch
541
                0
                              56
                                               0
                                                            0
341
              250
                              0
                                              24
                                                            0
                                                                          0
1074
              140
                              93
                                               0
                                                            0
                                                                          0
                                               0
                                                            0
                                                                          0
521
                0
                              40
850
                               0
                                               0
                                                            0
              140
                                                                          0
      PoolArea PoolQC Fence MiscFeature MiscVal
                                                                YrSold SaleType
                                                       MoSold
541
                    NaN
                          NaN
                                        NaN
                                                    0
                                                             6
                                                                   2007
341
              0
                   NaN
                          NaN
                                        NaN
                                                    0
                                                             9
                                                                  2009
                                                                               WD
                                                             5
1074
              0
                   NaN
                          NaN
                                        NaN
                                                    0
                                                                  2007
                                                                               WD
521
              0
                                                    0
                                                             5
                                                                  2007
                                                                               WD
                   {\tt NaN}
                          NaN
                                        NaN
850
              0
                    NaN
                          NaN
                                        NaN
                                                    0
                                                            11
                                                                  2007
                                                                               WD
     SaleCondition
541
             Normal
             Normal
341
1074
             Normal
521
             Normal
850
             Normal
y_train head:
541
        248000
341
         82000
1074
        194000
521
         150000
850
         131500
Name: SalePrice, dtype: int64
X_val head:
      MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape \
895
               60
                         RL
                                     71.0
                                               7056
                                                       Pave
                                                               NaN
                                                                         Reg
1414
               50
                         RL
                                     64.0
                                              13053
                                                       Pave
                                                              Pave
                                                                         Reg
25
               20
                         RL
                                    110.0
                                              14230
                                                       Pave
                                                               NaN
                                                                         Reg
0
               60
                         RL
                                     65.0
                                               8450
                                                       Pave
                                                               NaN
                                                                         Reg
```

959	160	FV	24.0	2572	Pave NaN	Reg	
	LandContour 1	Utilities I	otConfig L	andSlope Nei	ighborhood C	ondition1 \	
895	Lvl	AllPub	Inside	Gtl	NAmes	Norm	•
1414	Bnk	AllPub	Inside	Gtl	BrkSide	Norm	
25	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	
0	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	
959	Lvl	AllPub	FR2	Gtl	Somerst	Norm	
	Condition2 B		seStyle O	verallQual	OverallCond	YearBuilt	\
895	Norm	1Fam	2Story	6	5	1963	
1414	Norm	1Fam	1.5Fin	6	7	1923	
25	Norm	1Fam	1Story	8	5	2007	
0	Norm	1Fam	2Story	7	5	2003	
959	Norm	Twnhs	2Story	7	5	1999	
	YearRemodAdd	d RoofStvle	RoofMatl	Exterior1st	Exterior2nd	MasVnrTvpe	\
895	1963	•		HdBoard	HdBoard	BrkFace	•
1414	2000	-		Wd Sdng	Wd Sdng	NaN	
25	200		1 0	VinylSd	VinylSd		
0	2003		1 0	VinylSd	VinylSd		
959	1999		1 0	MetalSd	MetalSd		
		-	1 0				
	MasVnrArea l	ExterQual E	ExterCond F	oundation Ba	smtQual Bsmt	Cond \	
895	415.0	TA	TA	CBlock	TA	TA	
1414	0.0	TA	TA	BrkTil	TA	TA	
25	640.0	Gd	TA	PConc	Gd	TA	
0	196.0	Gd	TA	PConc	Gd	TA	
959	0.0	Gd	TA	PConc	Gd	TA	
	D . I	D	4 5 .5.	GE4 5 . E.		n:	
	BsmtExposure			nSF1 BsmtFir		FinSF2 \	
895	No		BLQ 	400	Unf	0	
1414	No		Inf	0	Unf	0	
25	No		Inf	0	Unf	0	
0	No		LQ	706	Unf	0	
959	No	P	LQ	604	Unf	0	
	BsmtUnfSF 7	TotalBsmtSF	Heating H	eatingQC Cer	ntralAir Ele	ctrical \	
895	380	780	GasA	TA	Y	SBrkr	
1414	833	833	GasA	Gd	Y	SBrkr	
25	1566	1566	GasA	Ex	Y	SBrkr	
0	150	856	GasA	Ex	Y	SBrkr	
959	92	696	GasA	Ex	Y	SBrkr	
	4 . 53 . 65 -	183 GE -	0 35. ~=	a	D :====	. 1	
005			wQualFinSF		BsmtFullBa <sup>-</sup>		
895	983	813	0	1796		1	0
1414	1053	795	0	1848		0	0
25	1600	0	0	1600		0	0

0 959	856 696	854 720		0		1710 1416		1 1			0	
		HalfBath	Dodr	oomAbvGr			·Cr Vit	ch on Ou o l	\			
895				oomadvGr 4	VICCII	епару		chenQual TA	\			
1414	1 1	1 1		4			1 1	Gd				
25	2	0		3			1	Gd				
0	2	1		3			1	Gd				
959	2	1		3			1	Gd				
303	2	1		3			1	ďď				
005	TotRmsAbvG			_		eplac			Gara	ageYrI		\
895		8	Тур		1		TA	Attchd		1963		
1414		8	Тур		1		Gd	Detchd		1922		
25		7	Тур		1		Gd	Attchd		2007		
0		8	Тур		0		NaN N-N	Attchd		2003		
959		6	Тур		0		NaN	Detchd		1999	9.0	
	GarageFinish		eCars			ageQu	al Gara	ageCond 1	Paved	Drive	\	
895	RFı		2	48			TA	TA		Y		
1414	Uni		2	37			TA	TA		N		
25	RFı		3	89			TA	TA		Y		
0	RFı		2	54			TA	TA		Y		
959	Uni	f	2	48	4		TA	TA		Y		
	WoodDeckSF	OpenPo	rchSF	EnclosedP	orch	3Ssn	Porch	ScreenPo	orch	\		
895	0		50		0		0		0			
1414	0		0		0		0		220			
25	0		56		0		0		0			
0	0		61		0		0		0			
959	0		44		0		0		0			
	PoolArea Po	oolQC Fe	nce Mi	scFeature	Misc	Val	MoSold	YrSold	Sale'	Гуре	\	
895	0		NaN	NaN		0	10	2008		WD		
1414	0	NaN	NaN	NaN		0	6	2008		WD		
25	0		NaN	NaN		0	7	2009		WD		
0	0		NaN	NaN		0	2	2008		WD		
959	0		NaN	NaN		0	5	2010		WD		
	SaleCondition	on.										
895	Norma											
1414	Norma											
25	Norma											
0	Norma											
959	Norma											
	head:	<b></b>										
9_vai 895	140000											
1414	207000											
25	256300											

0 208500 959 155000

Name: SalePrice, dtype: int64

===	Fold	1 2	===
X_tı	rain	hea	ad:

X_tra	ain head:						
	MSSubClass	MSZoning Lo	otFrontage	LotArea	Street Alley	LotShape \	
895	60	RL	71.0	7056	Pave NaN	I Reg	
1414	50	RL	64.0	13053	Pave Pave	e Reg	
25	20	RL	110.0	14230	Pave NaN	I Reg	
0	60	RL	65.0	8450	Pave NaN	I Reg	
959	160	FV	24.0	2572	Pave NaN	I Reg	
	${\tt LandContour}$	Utilities Lo	otConfig L	andSlope N	eighborhood	Condition1 \	
895	Lvl	AllPub	Inside	Gtl	NAmes	Norm	
1414	Bnk	AllPub	Inside	Gtl	BrkSide	Norm	
25	Lvl	AllPub	Corner	Gtl	${ t NridgHt}$	Norm	
0	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	
959	Lvl	AllPub	FR2	Gtl	Somerst	Norm	
	Condition2 B	ldgType Hous	seStyle O	verallQual	OverallCon	nd YearBuilt	\
895	Norm	1Fam	2Story	6		5 1963	
1414	Norm	1Fam	1.5Fin	6		7 1923	
25	Norm	1Fam	1Story	8		5 2007	
0	Norm	1Fam	2Story	7		5 2003	
959	Norm	Twnhs	2Story	7		5 1999	
	YearRemodAd	d RoofStyle	RoofMatl	Exterior1s	t Exterior2r	nd MasVnrType	\
895	196	3 Hip	CompShg	HdBoar		rd BrkFace	
1414	200	0 Gambrel	CompShg	Wd Sdn	g Wd Sdr	ng NaN	
25	200	7 Gable	CompShg	VinylS	d VinylS	Sd Stone	
0	200	Gable	CompShg	VinylS	d VinylS	d BrkFace	
959	199	9 Hip	CompShg	MetalS	d MetalS	Sd NaN	
	${ t MasVnrArea}$	ExterQual Ex	xterCond F	oundation	BsmtQual Bsm	ntCond \	
895	415.0	TA	TA	CBlock	TA	TA	
1414	0.0	TA	TA	${\tt BrkTil}$	TA	TA	
25	640.0	Gd	TA	PConc	Gd	TA	
0	196.0	Gd	TA	PConc	Gd	TA	
959	0.0	Gd	TA	PConc	Gd	TA	
	BsmtExposure			nSF1 BsmtF	V -	ntFinSF2 \	
895	No		LQ	400	Unf	0	
1414	No		nf	0	Unf	0	
25	No		nf	0	Unf	0	
0	No		LQ	706	Unf	0	
959	No	) Al	LQ	604	Unf	0	

	${\tt BsmtUnfSF}$	TotalBsm	ntSF I	Heating I	Heati	ngQC Cer	ntralAi	r Electr	ical	\		
895	380		780	${\tt GasA}$		TA	7	Y S	Brkr			
1414	833		833	${\tt GasA}$		Gd	7	Y S	Brkr			
25	1566	1	566	${\tt GasA}$		Ex	7	Y S	Brkr			
0	150		856	${\tt GasA}$		Ex	7	Y S	Brkr			
959	92		696	${\tt GasA}$		Ex	7	Y S	Brkr			
	4 . 53 . 65	0 183 88						775 . 1				,
005	1stFlrSF	2ndFlrSF	Lowl	QualFinSl		LivArea	BsmtFi		Bsmt	HalfBa		\
895	983	813			0	1796		1			0	
1414	1053	795			0	1848		0			0	
25	1600	0			0	1600		0			0	
0 959	856 696	854 720			) )	1710 1416		1 1			0	
959	090	120		(	J	1410		1			U	
	FullBath	HalfBath	Bedi	roomAbvG	r Ki	tchenAb	vGr Kit	chenQual	\			
895	1	1		4	4		1	TA				
1414	1	1		4	4		1	Gd				
25	2	0		;	3		1	Gd				
0	2	1		;	3		1	Gd				
959	2	1		;	3		1	Gd				
	m . D . A1 . /	a 1 E	,			n. 1	0 0	m	a	V D	٠.	,
005	TotRmsAbv(			Firepia		Firepla				ageYrB		\
895		8	Тур		1		TA	Attchd		1963		
1414 25		8	Тур		1		Gd Cd	Detchd		1922 2007		
25 0		7 8	Тур		1		Gd NaN	Attchd Attchd		2007		
959		6	Тур		0		NaN	Detchd		1999		
303		O	Тур		U		Ivaiv	Detcha		1999	.0	
	GarageFinis	sh Garage	Cars	Garage	Area	GarageQı	ual Gara	ageCond 1	Paved	Drive	\	
895	RI	Fn	2		483		TA	TA		Y		
1414	Uı	nf	2		370		TA	TA		N		
25	RI	Fn	3		890		TA	TA		Y		
0	Rl	Fn	2		548		TA	TA		Y		
959	Uı	nf	2		484		TA	TA		Y		
	UdDl-Cl	C 0D	- L CF	Engles	- 4D	-h 20	-Db	C D	h	`		
895	WoodDeckSl	F OpenPor O	cnsr 50	Enclose	earor	ch 3Ssi 0	nPorcn 0	ScreenP	orcn O	\		
1414		0	0			0	0		220			
25		0	56			0	0		0			
0		0	61			0	0		0			
959		0	44			0	0		0			
909	,	J	44			U	U		U			
	PoolArea l	PoolQC Fer	ce M:	iscFeatu	re M	liscVal	MoSold	YrSold	Sale	Туре	\	
895	0	NaN N	IaN	Na	aN	0	10	2008		WD		
1414	0	NaN N	JaN	Na	aN	0	6	2008		WD		
25	0	NaN N	IaN	Na	aN	0	7	2009		WD		
0	0	NaN N	IaN	Na	aN	0	2	2008		WD		

959	0	NaN NaN	NaN	0	5 2	2010 WD
	SaleCondition	on				
895	Norma					
1414	Norma					
25	Norma					
0	Norm					
959	Norma					
	in head:	Q				
895	140000					
1414	207000					
25	256300					
0	208500					
959	155000					
	SalePrice,	dtwne: int	64			
	head:	dtype. Int	04			
Λ_VαΙ		MS7oning	LotFrontage	IntArea S	treet Alley	LotShape \
541	60	RL	NaN	11000	Pave NaN	Reg
341	20	RH	60.0	8400	Pave NaN	_
1074	20	RL	74.0	8556	Pave NaN	Reg
						Reg
521	20	RL	90.0	11988	Pave NaN	IR1
850	120	RM	36.0	4435	Pave NaN	Reg
	LandContour	Utilities	LotConfig La	ndSlope Ne	ighborhood (	Condition1 \
541	Lvl		FR2	Gtl	NoRidge	Norm
341	Lvl	AllPub	Inside	Gtl	SawyerW	Feedr
1074	Lvl		Inside	Gtl	CollgCr	Norm
521	Lvl		Corner	Gtl	NAmes	Feedr
850	Lvl		Inside	Gtl	CollgCr	Norm
000	271	niii ub	IIIDIGO	401	0011801	WOIM
	Condition2	BldgType Ho	useStyle Ov	erallQual	OverallCond	d YearBuilt \
541	Norm	1Fam	2Story	8		5 2000
341	Norm	1Fam	1Story	4	4	1950
1074	Norm	1Fam	1Story	7	Ę	5 2006
521	Norm	1Fam	1Story	6	6	5 1957
850	Norm	TwnhsE	1Story	6	Ę	5 2003
	YearRemodA	dd RoofStyl	e RoofMatl E	Exterior1st	Exterior2nd	d MasVnrType \
541	200	00 Gabl	e CompShg	VinylSd	VinylSo	d BrkFace
341	19	50 Gabl	e CompShg	Wd Sdng	; AsbShng	g NaN
1074	200	06 Gabl	e CompShg	VinylSd	VinylSo	d NaN
521	19	57 Gabl	e CompShg	VinylSd	VinylSo	d NaN
850	200	03 Gabl	e CompShg	VinylSd	VinylSo	d BrkFace
			ExterCond Fo			
541	72.0	Gd	TA	PConc	Gd	TA
341	0.0		Fa	CBlock	TA	Fa
1074	0.0	Gd	TA	PConc	Gd	TA

521	0.0	TA	TA	CBlock	TA	TA			
850	170.0	Gd	TA	PConc	Gd	TA			
	Remt Evnosure	e BsmtFinType:	l RemtFinGl	F1 RemtF	inTwne?	Ram+FinS	F2 \		
541	No			0	Unf	Domet. THO	0	`	
341	No			0	Unf		0		
1074	Av			0	Unf		0		
521	No			77	Unf		0		
850	Av			 59	Unf		0		
		<u> </u>	•		V				
		TotalBsmtSF	_	tingQC C	entralAi	r Electri	cal	\	
541	969	969	GasA	Ex		Y SB	rkr		
341	721	721	GasA	Gd		Y SB	rkr		
1074	1240	1240	GasA	Ex		Y SB	rkr		
521	467	1244	GasA	Ex			seA		
850	189	848	GasA	Ex		Y SB	rkr		
	1stFlrSF 2	2ndFlrSF Low	QualFinSF (	GrLivAre	a BsmtF	ullBath	BsmtH	MalfBath	\
541	997	1288	0	228		0		0	
341	841	0	0	84		0		0	
1074	1240	0	0	124		0		0	
521	1244	0	0	124		0		0	
850	848	0	0	84		1		0	
	FullBath H	HalfBath Bedi	roomAbvGr I	KitchenA	bvGr Kit	chenQual	\		
541	2	1	4		1	Gd			
341	1	0	2		1	TA			
1074	2	0	2		1	Gd			
521	1	1	3		1	TA			
850	1	0	1		1	Gd			
	TotRmsAbvGr	d Functional	Fireplaces	s Firepl	aceQu Ga	rageTvpe	Gara	ngeYrBlt	\
541				_		BuiltIn		2000.0	•
341		4 Typ		0	NaN	CarPort		1950.0	
1074		5 Typ		0	NaN	Attchd		2006.0	
521		6 Typ		2	Gd	Attchd		1957.0	
850		3 Typ		0	NaN	Attchd		2003.0	
	GarageFinish	•	GarageArea	_		-	avedD		
541	Fin		648		TA	TA		Y	
341	Unf		294		TA	TA		N	
1074	RFn		826		TA	TA		Y	
521	Unf		336		TA	TA		Y	
850	Fin	n 2	420	0	TA	TA		Y	
	WoodDeckSF	OpenPorchSF	EnclosedPo	orch 3S	snPorch	ScreenPo	rch	\	
541	0	56		0	0	22230111 0	0	•	
341	250	0		24	0		0		

1074 521	140 0	93 40	0 0	0		0 0	
	140	0	0	0		0	
541 0 341 0 1074 0 521 0	a PoolQC Fenc O NaN Na O NaN Na O NaN Na O NaN Na	N Na N Na N Na N Na	N O N O N O N O	MoSold 6 9 5 5	YrSold 2007 2009 2007 2007 2007	SaleType WD WD WD WD WD	
SaleCond:	ition						
541 No 341 No 1074 No 521 No	ormal ormal ormal ormal ormal ormal						
Name: SalePri	ce, dtype: in	.T04					
Name: SalePrid		.to4					
=== Fold 3 === X_train head:	=						
=== Fold 3 === X_train head: MSSubCla	= ass MSZoning	LotFrontage			•	-	
=== Fold 3 === X_train head: MSSubCla	= ass MSZoning 60 RL	LotFrontage 71.0	7056	Pave 1	NaN	Reg	
=== Fold 3 === X_train head: MSSubCla 895 1414	ass MSZoning 60 RL 50 RL	LotFrontage 71.0 64.0	7056 13053	Pave Pave Pa	NaN ave	Reg Reg	
=== Fold 3 === X_train head: MSSubCl: 895 1414 25	ass MSZoning 60 RL 50 RL 20 RL	LotFrontage 71.0 64.0 110.0	7056 13053 14230	Pave Pave Pave Pave	VaN ave VaN	Reg Reg Reg	
=== Fold 3 === X_train head: MSSubCla 895 1414 25	ass MSZoning 60 RL 50 RL 20 RL	LotFrontage 71.0 64.0 110.0 65.0	7056 13053 14230 8450	Pave Pave Pave Pave Pave	NaN ave NaN NaN	Reg Reg Reg Reg	
=== Fold 3 === X_train head: MSSubCla 895 1414 25 0 959	ass MSZoning 60 RL 50 RL 20 RL 60 RL	LotFrontage 71.0 64.0 110.0 65.0 24.0	7056 13053 14230 8450 2572	Pave Pave Pave Pave Pave Pave Pave Pave	VaN ave VaN VaN	Reg Reg Reg Reg Reg	
=== Fold 3 === X_train head:	ass MSZoning 60 RL 50 RL 20 RL 60 RL 160 FV	LotFrontage 71.0 64.0 110.0 65.0 24.0	7056 13053 14230 8450 2572 andSlope Ne	Pave Pave Pave Pave Pave Pave Pave Pave	JaN ave JaN JaN JaN	Reg Reg Reg Reg Reg	
=== Fold 3 === X_train head:	ass MSZoning 60 RL 50 RL 20 RL 60 RL 160 FV	LotFrontage 71.0 64.0 110.0 65.0 24.0 LotConfig L Inside	7056 13053 14230 8450 2572 andSlope Ne	Pave Pave Pave Pave Pave Pave Pave Nave Nave	JaN nve JaN JaN JaN od Condi	Reg Reg Reg Reg Reg Ltion1 \	
=== Fold 3 === X_train head:	ass MSZoning 60 RL 50 RL 20 RL 60 RL 160 FV our Utilities Lvl AllPub	LotFrontage 71.0 64.0 110.0 65.0 24.0 LotConfig L Inside Inside	7056 13053 14230 8450 2572 andSlope Ne	Pave Pave Pave Pave Pave Pave Name	JaN ave JaN JaN JaN od Condi	Reg Reg Reg Reg Reg .tion1 \ Norm	
=== Fold 3 === X_train head:	ass MSZoning 60 RL 50 RL 20 RL 60 RL 160 FV cour Utilities Lv1 AllPub Bnk AllPub	LotFrontage 71.0 64.0 110.0 65.0 24.0 LotConfig L Inside Inside Corner	7056 13053 14230 8450 2572 andSlope Ne Gtl Gtl Gtl	Pave Pave Pave Pave Pave Pave Name BrkSic	JaN ave JaN JaN JaN od Condi es de	Reg Reg Reg Reg Reg Norm Norm	
=== Fold 3 === X_train head:	ass MSZoning 60 RL 50 RL 20 RL 60 RL 160 FV our Utilities Lvl AllPub Lvl AllPub Lvl AllPub	LotFrontage 71.0 64.0 110.0 65.0 24.0 LotConfig L Inside Inside Corner Inside	7056 13053 14230 8450 2572 andSlope Ne Gtl Gtl Gtl Gtl	Pave Pave Pave Pave Pave Pave Pave Pave	JaN ave JaN JaN JaN JaN od Condi es de It	Reg Reg Reg Reg Rom Norm Norm Norm	
=== Fold 3 === X_train head:	ass MSZoning 60 RL 50 RL 20 RL 60 RL 160 FV cour Utilities Lv1 AllPub Bnk AllPub	LotFrontage 71.0 64.0 110.0 65.0 24.0 LotConfig L Inside Inside Corner Inside	7056 13053 14230 8450 2572 andSlope Ne Gtl Gtl Gtl	Pave Pave Pave Pave Pave Pave Name BrkSic	JaN ave JaN JaN JaN JaN od Condi es de It	Reg Reg Reg Reg Reg Norm Norm	
=== Fold 3 === X_train head:	ass MSZoning 60 RL 50 RL 20 RL 60 RL 160 FV  Dur Utilities Lvl AllPub Enk AllPub Lvl AllPub Lvl AllPub Lvl AllPub rm 1Fam rm 1Fam	LotFrontage 71.0 64.0 110.0 65.0 24.0 LotConfig L Inside Inside Corner Inside FR2	7056 13053 14230 8450 2572 andSlope Ne Gtl Gtl Gtl Gtl	Pave Pave Pave Pave Pave Pave Pave Pave	JaN ave JaN JaN JaN JaN od Condi es de Ht st	Reg Reg Reg Reg Rom Norm Norm Norm	\

	YearRemodAd	ld RoofSty	le RoofMatl	Exterior1st	Exterior2nd	d MasVnrTy	pe \
895	196	3 H	ip CompShg	HdBoard	HdBoard	d BrkFa	.ce
1414	200	00 Gambr	el CompShg	Wd Sdng	Wd Sdng	g N	aN
25	200	7 Gab	le CompShg	VinylSd	VinylSo	i Sto	ne
0	200	3 Gab	le CompShg	VinylSd	VinylSo	d BrkFa	.ce
959	199	99 H	ip CompShg	MetalSd	MetalSo	i N	aN
		ExterQual	ExterCond	Foundation Ba	smtQual Bsm	tCond \	
895	415.0	TA	TA	CBlock	TA	TA	
1414	0.0	TA	TA	BrkTil	TA	TA	
25	640.0	Gd		PConc	Gd	TA	
0	196.0	Gd	TA	PConc	Gd	TA	
959	0.0	Gd	TA	PConc	Gd	TA	
	ъ . п	D . E. E	4 5 . 7		<b></b>	n: ano \	
005	_		-	inSF1 BsmtFir			
895	No		BLQ	400	Unf	0	
1414	No		Unf	0	Unf	0	
25	No		Unf	0	Unf	0	
0	No		GLQ	706	Unf	0	
959	No	)	ALQ	604	Unf	0	
	BsmtUnfSF	TotalRemt	SF Heating	HeatingQC Cer	ntralAir Fla	actrical	\
895	380		80 GasA	TA	Y	SBrkr	•
1414	833		33 GasA	Gd	Y	SBrkr	
25	1566	15		Ex	Y	SBrkr	
0	150		56 GasA	Ex	Y	SBrkr	
959	92		96 GasA	Ex	Y	SBrkr	
	1stFlrSF 2	ndFlrSF	LowQualFinS	F GrLivArea	BsmtFullBa	ath BsmtH	alfBath \
895	983	813		0 1796		1	0
1414	1053	795		0 1848		0	0
25	1600	0		0 1600		0	0
0	856	854		0 1710		1	0
959	696	720		0 1416		1	0
	FullBath H	MalfBath :	${ t BedroomAbvG}$	r KitchenAby	vGr Kitchen(	Qual \	
895	1	1		4	1	TA	
1414	1	1		4	1	Gd	
25	2	0		3	1	Gd	
0				3	1	Gd	
•	2	1		O	_		
959	2 2	1 1		3	1	Gd	
	2	1		3	1		
959		1 rd Function	nal Firepl	3 aces Fireplac	1 ceQu Garage	Гуре Gara	geYrBlt \
959 895	2	1 rd Function 8	nal Firepl Typ	3 aces Fireplac 1	1 ceQu Garage TA Att	Гуре Gara cchd	1963.0
959 895 1414	2	1 rd Function 8	nal Firepl Typ Typ	3 aces Fireplac 1 1	1 ceQu Garage TA At Gd De	Type Gara cchd cchd	1963.0 1922.0
959 895	2	1 rd Function 8 8 7	nal Firepl Typ	3 aces Fireplac 1	1 ceQu Garage TA Att Gd Det Gd Att	Гуре Gara cchd	1963.0

959	6	Гур	0	NaN	Detchd		1999	).(
GarageFinis	h GarageCa	ars GarageAre	ea Garage	Qual Gara	ageCond	PavedDı	cive	`
895 RF	n	2 48	33	TA	TA		Y	
1414 Un	f	2 37	0	TA	TA		N	
25 RF	'n	3 89	90	TA	TA		Y	
0 RF		2 54		TA	TA		Y	
959 Un		2 48		TA	TA		Y	
	_		_				_	
WoodDeckSF	OpenPorch	nSF EnclosedF	orch 3S	snPorch	ScreenP	orch \	\	
895 0		50	0	0		0		
1414 0		0	0	0		220		
25 0		56	0	0		0		
0 0		61	0	0		0		
959 0		44	0	0		0		
PoolArea P	oolQC Fence	e MiscFeature	MiscVal	MoSold	YrSold	SaleTy	уре	\
895 0	NaN Nal	NaN	0	10	2008		WD	
1414 0	NaN Nal	NaN	0	6	2008		WD	
25 0	NaN Nal		0		2009		WD	
0 0	NaN Nal		0		2008		WD	
959 0	NaN Nal		0		2010		WD	
SaleConditi	on							
895 Norm								
1414 Norm								
25 Norm								
0 Norm								
959 Norm								
y_train head:	.ui							
895 140000								
1414 207000								
25 256300								
0 208500								
959 155000								
Name: SalePrice,	dtypo: int	-64						
X_val head:	dtype. In	004						
_	MCZoning	I atEmantama	T a+ Amas	Street A	110 10+	Chana	`	
	_	LotFrontage NaN		Pave	v	-	\	
			4426		NaN NaN	Reg		
1217 20		72.0	8640	Pave	NaN	Reg		
869 60		80.0	9938	Pave	NaN NaN	Reg		
855 20		NaN	6897	Pave	NaN NaN	IR1		
178 20	RL	63.0	17423	Pave	NaN	IR1		
IandCantaum	II+ili+ioo	LotConfig Lar	ndSlope ™	aighharh	ood Cond	ition1	\	
1441 Lvl		Inside	Gtl	_			\	
				Colla	-	Norm		
1217 Lvl		Inside	Gtl C+l	Some		Norm		
869 Lvl	AllPub	Inside	Gtl	Sawye	≒T.M	Norm		

855 178	Lv Lv			Gtl Gtl	Sawyer StoneBr	Norm Norm	
1441 1217	Condition2 Norm	TwnhsE	HouseStyle 1Story 1Story	OverallQual 6 8	OverallCond 5 5	YearBuilt 2004 2009	\
869	Norm		2Story	7	5	1993	
855	Norm		1Story	5	8	1962	
178	Norm		1Story	9	5	2008	
170	NOTIII	11 alli	15001 y	9	3	2008	
	YearRemod	Add RoofSt	yle RoofMati	l Exterior1st	Exterior2nd	MasVnrType	\
1441	2	004 Ga	ble CompSh	g VinylSd	l VinylSd	${\tt BrkFace}$	
1217	2	009 Ga	ble CompSh	g CemntBo	CmentBd	Stone	
869	1	994 Ga	ble CompSh	g MetalSd	l MetalSd	BrkFace	
855	2	010 Ga	ble CompSh		l HdBoard	NaN	
178	2	009	Hip CompSh		l VinylSd	Stone	
				y v	J		
					SsmtQual Bsmt		
1441	147.		d TA	PConc	Gd	TA	
1217	72.		d TA	PConc	Gd	TA	
869	246.		d TA	PConc	Gd	TA	
855	0.		TA Gd	CBlock	TA	TA	
178	748.	0 E	Ex TA	PConc	Ex	TA	
						<b>\</b>	
	BsmtExposu		V -	FinSF1 BsmtFi	V -	FinSF2 \	
1441		Αv	GLQ	697	Unf	0	
1217		Mn	GLQ	936	Unf	0	
869		No	GLQ	750	Unf	0	
855		No	ALQ	659	Unf	0	
178		No	GLQ	1904	Unf	0	
	BsmtUnfSF	TotalBsm	ntSF Heating	HeatingQC Ce	entralAir Ele	ctrical \	
1441	151		848 GasA	•		SBrkr	
1217	364		.300 GasA	Ex	Y	SBrkr	
869	300		.050 GasA	Ex	Y	SBrkr	
855	381		.040 GasA	Ex	Y	SBrkr	
178	312		216 GasA	Ex	Y	SBrkr	
					_		
	1stFlrSF	2ndFlrSF	LowQualFins	SF GrLivArea	BsmtFullBa	th BsmtHali	fBath \
1441	848	0		0 848	}	1	0
1217	1314	0		0 1314		1	0
869	1062	887		0 1949		1	0
855	1040	0		0 1040		1	0
178	2234	0		0 2234		1	0
•	1	· ·					-
	FullBath	HalfBath	BedroomAbv	Gr KitchenAb	vGr KitchenQ	ual \	
1441	1	0		1	1	Gd	
1217	2	0		3	1	Gd	

869	2		1	3		1	Gd		
855	1		1	3		1	TA		
178	2		0	1		1	Ex		
	TotRmsAbv	Grd Fur	nctional	Fireplaces	Firepla	ceQu Ga	arageTvpe	GarageYrB	lt \
1441		3	Тур	1	r	TA	Attchd	2004	
1217		6	Тур	0		NaN	Attchd	2009	
869		8	Тур	1		TA	Attchd	1993	
855		6		0		NaN	Detchd	1962	
178		9	Тур			Gd	Attchd	2009	
170		9	Тур	1		Gu	Accella	2009	.0
	GarageFini	sh Gar	ageCars	GarageArea	GarageQ	ual Gai	rageCond P	avedDrive	\
1441	R.	Fn	2	420		TA	TA	Y	
1217	R.	Fn	2	552		TA	TA	Y	
869	F	in	2	574		TA	TA	Y	
855	U:	nf	1	260		TA	TA	Y	
178		in	3	1166		TA	TA	Y	
	WoodDeckS	-	PorchSF	EnclosedPo			ScreenPo		
1441	14		0		0	0		0	
1217	13	5	112		0	0		0	
869	15	6	90		0	0		0	
855		0	104		0	0		0	
178		0	60		0	0		0	
	<b>5 7 6</b>							a	,
	PoolArea			MiscFeature	MiscVal	MoSo		SaleType	\
1441	0	NaN	NaN	NaN	0		5 2008	WD	
1217	0	NaN	NaN	NaN	0		9 2009	New	
869	0	NaN	GdPrv	NaN	0		6 2010	WD	
855	0	NaN	NaN	NaN	0		4 2010	WD	
178	0	NaN	NaN	NaN	0		7 2009	New	
	SaleCondit	ion							
1441	Nor								
1217	Part								
869	Nor								
855	Nor								
178	Part								
		ıaı							
y_vai 1441	head: 149300								
1217	229456								
869	236000								
855	127000								
178	501837	5.							
Name:	SalePrice	, dtype	e: int64						

 ${\tt Checking\ reproducibility:}$ 

If X\_test is the same: True If y\_test is the same: True

All train/validation fold indices are same.