Ejercicios Preprocesamiento

Linear regresión con dataset Housing Prices

1. Carga el dataset “train_housing_prices.csv”. **Abajo estan explicadas las variables.**

2. **Missings**. Primero vamos a realizar el tratamiento de missings:
    a. Observa el % de missings y el tipo de cada variable del dataframe.
    b. Elimina aquellas variables con más de un 80% de missings
    c. Imputa los missings de las **variables categóricas** a la categoría “Missing” en cada variable.
    d. Imputa los missings de las **variables continuas** a la media de cada variable.

3. **Análisis variable respuesta**: La variable respuesta es *SalePrice,* precio de venta.
    a. Describe la variable, verifica que no tenga valores negativos y analiza la distribución empírica (histograma).
    b. ¿Cumple las características de una distribución normal?
        i. ¿Está centrada?
        ii. ¿Dónde se acumulan la mayor parte de los valores?
        iii. ¿Las colas son iguales?
    c. Piensa y busca qué transformación simple se puede aplicar para conseguir una variable más parecida a una normal (raíz cuadrada, potencia al cuadrado, exponencial, logaritmo…?). Aplica la transformación y verifica que la nueva distribución se parece más a una normal. **Continua todo el análisis con la variable *SalePrice* transformada.**

4. **Análisis gráfico y transformación de variables**: Ahora vamos a realizar el análisis gráfico de las 9 variables que están más correlacionadas con *SalePrice.* ***Al final de los enunciados hay un codigo que podéis utilizar.*** **Para cada variable:**
    a. Gráfica la distribución/histograma y verifica el tipo de variable. Algunas variables categóricas vienen como tipo continuo.
    b. Si son variables categóricas, verifica si es posible unificar categorías. Para ello realiza el análisis bivariante con los gráficos boxplot de cada categoría(con el target) . [Por ejemplo, puedes agrupar categorías de la variable *OverallQual*]. Une categorías que tengan una distribución del target parecida.
    c. Si son variables continuas, verifica que no tengan outliers (a simple vista). Para ello realiza el análisis bivariante con el scatter(con el target). [Por ejemplo, puedes eliminar 2 outliers de la variable *GrLivArea*]
    d. Si son variables continuas, verifica que no tengan demasiada asimetría y en caso de que tengan demasiada, realiza la transformación log. Para ello observa la distribución y calcula la skewness(número que indica el nivel de asimetría). [Por ejemplo, *GrLivArea* presenta mucha asimetría]
    Df.column.skew

https://en.wikipedia.org/wiki/Skewness (skewness=Asímetria)

5. **Estimación modelo**: Ahora vamos a estimar una regresión lineal con las 9 variables anteriores.
    a. Crea las dummy asociadas a las variables categóricas. Recuerda transformarlas a categórica antes [Por ejemplo, *OverallQual*]
    b. Escala todas las variables continuas con la transformación StandardScaler.
    c. Estima la regresión lineal y calcula el RMSE y el R2.
    d. Calcula los errores.
        i. Grafíca el histograma, ¿se parece a una normal? ¿Hay colas pesadas?
        ii. Grafíca el scatter de la predicción vs el error. ¿En qué zona observas mayor dispersión?
        iii. Grafíca el scatter de la predicción vs el valor real. ¿Los residuos(errores) tienen varianza constante?

6. **Revisiones**: A continuación listamos **una serie de opciones** para intentar mejorar nuestra regresión lineal:
    a. Probar otras transformaciones para escalar las variables continuas.
    b. Realizar una imputación de missings más realista y detallada.
    Muchas de las variables que presentan missing se puede asumir que el missing es igual a 0, en algunas categóricas se podrían asignar a alguna categoría (por ejemplo si no tiene sótano, pues missings de variables relacionadas con sótano poner en categoría No).
    c. Crear variables adicionales a través de combinaciones de otras variables.
    Por ejemplo, sumar Baños y 0.5*MediosBaños para obtener BañosTotales.
    d. Crear variables adicionales (X^2 y X^3) de las principales variables.
    e. Trabajar con todas las variables.

Código:

```
# correlacion y mapa de calor
corrmat = X.corr() # X son los datos numericos
sns.heatmap(corrmat, vmax=.8, square=True)

#Ampliación matriz, para variable que se selecciona muestra las k variables más correlacionadas, Substituye 'NOMBREVARIABLE' por el nombre de la variable a estudiar
k = 10 #numero de variables
cols = corrmat.nlargest(k, 'NOMBREVARIABLE')['NOMBREVARIABLE'].index
cm = np.corrcoef(X[cols].values.T)
sns.set(font_scale=1.25)
```

```
hm = sns.heatmap(cm, cbar=True, annot=True, square=True, fmt='.2f', annot_kws={'size': 10}, yticklabels=cols.values, xticklabels=cols.values)
plt.show()
```

Data fields

Here's a brief version of what you'll find in the data description file.

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)

- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale