#!/usr/bin/env python

# coding: utf-8

# # Learning Curves

# In[4]:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import learning\_curve

# In[5]:

coursework=pd.read\_csv(r"C:\Users\tiwar\OneDrive\Desktop\CourseWork\_cleaned\_data.csv")

# In[6]:

coursework.head()

# In[7]:

X=["MSSubClass\_N","MSZoning\_N","LotFrontage\_log","LotArea\_log","LotShape\_N","HouseStyle\_N","OverallQual","OverallCond","Foundation\_N","BsmtQual\_N","BsmtCond\_N","Heating\_N","HeatingQC\_N","SaleCondition\_N","GarageCond\_N","GarageQual\_N","FireplaceQu\_N","Electrical\_N","KitchenQual\_N","CentralAir\_N"]

# In[8]:

y='sales\_log'

# In[9]:

train\_sizes = range(1,1200,50)

# In[11]:

train\_sizes, train\_scores, test\_scores = learning\_curve(estimator= LinearRegression(),

X=coursework[X], y=coursework[y], cv=10, scoring='neg\_root\_mean\_squared\_error',

# 50 different sizes of the training set

train\_sizes=train\_sizes)

# In[12]:

train\_mean = np.mean(-train\_scores, axis=1)

train\_std = np.std(-train\_scores, axis=1)

# In[13]:

test\_mean = np.mean(-test\_scores, axis=1)

test\_std = np.std(-test\_scores, axis=1)

# In[14]:

# Drawing lines

plt.subplots(1, figsize=(10,10) )

plt.plot(train\_sizes, train\_mean, '--', color="#110011", label="Training score")

plt.plot(train\_sizes, test\_mean, color="#834651", label="Cross-validation score")

# Drawing bands

plt.fill\_between(train\_sizes, train\_mean - train\_std, train\_mean + train\_std, color="#DDDDDD")

plt.fill\_between(train\_sizes, test\_mean - test\_std, test\_mean + test\_std, color="#DDDDDD")

# Creating plot

plt.title("Learning Curve")

plt.xlabel("Training Set Size"), plt.ylabel("RMSE Score"), plt.legend(loc="best")

plt.tight\_layout();

plt.show()

# In[ ]:

# In[ ]: