```
In [ ]:
          Type Markdown and LaTeX: \alpha^2
 In [ ]: ▶
In [12]: ▶ # importing libraries
              import pandas as pd # data science essentials
              import matplotlib.pyplot as plt # data visualization
              import seaborn as sns # enhanced data visualization
              import statsmodels.formula.api as smf # regression modeling
              from sklearn.model_selection import train_test_split # train/test split
              import sklearn.linear_model # linear modeling in scikit-learn
              # setting pandas print options
              pd.set_option('display.max_rows', 500)
             pd.set_option('display.max_columns', 500)
              pd.set_option('display.width', 1000)
              # specifying the path and file name
              file = 'Ames Housing Dataset.xlsx'
              # reading the file into Python
              df = pd.read_excel(file)
              # checking the file
              df.head(n = 5)
   Out[12]:
                 Order Lot_Area Street Lot_Config Neighborhood Overall_Qual Overall_Cond Mas_Vnr_Area Total_Bsmt_SF First_Flr_SF Second_Flr_SF Gr_Liv_Area
              0
                          31770
                                 Pave
                                                       NAmes
                                                                       6
                                                                                   5
                                                                                             112.0
                                                                                                          1080.0
                                                                                                                       1656
                                                                                                                                       0
                                                                                                                                                16
                                          Corner
                    2
                          11622
                                 Pave
                                           Inside
                                                       NAmes
                                                                       5
                                                                                   6
                                                                                               0.0
                                                                                                           882.0
                                                                                                                       896
                                                                                                                                       0
                                                                                                                                                 89
                    3
                          14267
                                 Pave
                                          Corner
                                                       NAmes
                                                                                   6
                                                                                             108.0
                                                                                                          1329.0
                                                                                                                       1329
                                                                                                                                       0
                                                                                                                                                132
                     4
                          11160
                                 Pave
                                          Corner
                                                       NAmes
                                                                                   5
                                                                                               0.0
                                                                                                          2110.0
                                                                                                                       2110
                                                                                                                                       0
                                                                                                                                                21
                     5
                          13830
                                 Pave
                                           Inside
                                                       Gilbert
                                                                       5
                                                                                   5
                                                                                               0.0
                                                                                                           928.0
                                                                                                                       928
                                                                                                                                     701
                                                                                                                                                162
In [18]:  df.isna().sum()
   Out[18]: Order
              Lot Area
                                 0
              Street
                                 0
              Lot_Config
                                 0
              Neighborhood
                                 0
              Overall_Qual
              Overall_Cond
                                0
              Mas_Vnr_Area
                                23
              Total_Bsmt_SF
              First_Flr_SF
              Second Flr SF
              Gr_Liv_Area
              Full_Bath
                                 0
              Half_Bath
              Kitchen_AbvGr
              TotRms_AbvGr
              Fireplaces
                                 a
              Garage_Cars
                                 1
              Garage_Area
              Porch Area
                                 0
              Pool_Area
                                 0
              Sale_Price
              dtype: int64
```

```
In [16]: ► df.info()
               <class 'pandas.core.frame.DataFrame'>
               RangeIndex: 2930 entries, 0 to 2929
               Data columns (total 22 columns):
                                     Non-Null Count Dtype
                    Column
                0
                    Order
                                      2930 non-null
                                                        int64
                    Lot_Area
                                      2930 non-null
                                                        int64
                    Street
                                      2930 non-null
                                                        object
                    Lot_Config
                                      2930 non-null
                3
                                                        object
                    Neighborhood
                                      2930 non-null
                                                        object
                    Overall_Qual
                                      2930 non-null
                                                        int64
                    Overall_Cond
                                      2930 non-null
                                                        int64
                    Mas_Vnr_Area
                                      2907 non-null
                                                        float64
                8
                    Total Bsmt SF
                                      2929 non-null
                                                        float64
                    First_Flr_SF
                                      2930 non-null
                                                        int64
                10
                    Second_Flr_SF
                                      2930 non-null
                                                        int64
                11
                    Gr_Liv_Area
                                      2930 non-null
                                                        int64
                12
                    Full Bath
                                      2930 non-null
                                                        int64
                    Half_Bath
                                      2930 non-null
                13
                                                        int64
                14
                    Kitchen_AbvGr
                                     2930 non-null
                                                        int64
                15
                     TotRms_AbvGr
                                      2930 non-null
                                                        int64
                16
                    Fireplaces
                                      2930 non-null
                                                        int64
                                      2929 non-null
                17
                    Garage Cars
                                                        float64
                18
                    Garage_Area
                                      2929 non-null
                                                        float64
                19
                    Porch_Area
                                      2930 non-null
                                                        int64
                20
                    Pool_Area
                                      2930 non-null
                    Sale Price
                21
                                      2930 non-null
                                                        int64
               dtypes: float64(4), int64(15), object(3)
               memory usage: 503.7+ KB
In [13]: ► df.describe()
    Out[13]:
                                                                                                                                                  Full_Bath
                           Order
                                       Lot_Area Overall_Qual Overall_Cond Mas_Vnr_Area Total_Bsmt_SF First_Flr_SF Second_Flr_SF Gr_Liv_Area
                                                                                                                                                             н
                count 2930.00000
                                    2930.000000
                                                2930.000000
                                                              2930.000000
                                                                            2907.000000
                                                                                           2929.000000
                                                                                                       2930.000000
                                                                                                                      2930.000000
                                                                                                                                  2930.000000
                                                                                                                                               2930.000000
                                                                                                                                                           2930
                mean 1465.50000
                                   10147.921843
                                                   6.094881
                                                                 5.563140
                                                                             101.896801
                                                                                           1051.614544 1159.557679
                                                                                                                       335.455973 1499.690444
                                                                                                                                                  1.566553
                  std
                       845.96247
                                    7880.017759
                                                    1.411026
                                                                 1.111537
                                                                              179.112611
                                                                                            440.615067
                                                                                                        391.890885
                                                                                                                       428.395715
                                                                                                                                    505.508887
                                                                                                                                                  0.552941
                                    1300.000000
                                                   1.000000
                                                                 1.000000
                                                                               0.000000
                                                                                              0.000000
                                                                                                                         0.000000
                                                                                                                                    334.000000
                 min
                         1.00000
                                                                                                        334.000000
                                                                                                                                                  0.000000
                 25%
                       733 25000
                                    7440 250000
                                                   5.000000
                                                                 5.000000
                                                                               0.000000
                                                                                            793.000000
                                                                                                        876 250000
                                                                                                                         0.000000
                                                                                                                                   1126 000000
                                                                                                                                                  1.000000
                 50%
                     1465.50000
                                    9436.500000
                                                   6.000000
                                                                 5.000000
                                                                               0.000000
                                                                                            990.000000
                                                                                                       1084.000000
                                                                                                                         0.000000
                                                                                                                                   1442.000000
                                                                                                                                                  2.000000
                 75% 2197.75000
                                   11555.250000
                                                   7.000000
                                                                 6.000000
                                                                             164.000000
                                                                                           1302.000000 1384.000000
                                                                                                                       703.750000 1742.750000
                                                                                                                                                  2.000000
                 max 2930.00000 215245.000000
                                                   10.000000
                                                                 9.000000
                                                                             1600.000000
                                                                                           6110.000000 5095.000000
                                                                                                                      2065.000000 5642.000000
                                                                                                                                                  4.000000
In [19]: ► df.columns
    Out[19]: Index(['Order', 'Lot_Area', 'Street', 'Lot_Config', 'Neighborhood', 'Overall_Qual', 'Overall_Cond', 'Mas_Vnr_Area', 'Total_B smt_SF', 'First_Flr_SF', 'Second_Flr_SF', 'Gr_Liv_Area', 'Full_Bath', 'Half_Bath', 'Kitchen_AbvGr', 'TotRms_AbvGr', 'Firepla
               ces', 'Garage_Cars', 'Garage_Area', 'Porch_Area', 'Pool_Area', 'Sale_Price'], dtype='object')
```

```
In [17]:
                   #nLot heatman
                   plt.figure(figsize=(10,8))
                   sns.heatmap(df.corr(),annot=True)
                   plt.show()
                                                                                                                                 - 1.0
                             Order - 1 0.0340.049.01-0.0340.029.040300094.00900450.040.01000026.0140.036.03050068.0530.03
                                   0.031 1 0.0970.0350.13 0.25 0.330.0330.29 0.130.0350.02 0.22 0.26 0.18 0.21 0.2 0.0940.27
                      Overall Qual 0.049.097 1 0.0950.43 0.55 0.48 0.24 0.57 0.52 0.27 0.16 0.38 0.39 0.6 0.56 0.29 0.03 0.8
                                                                                                                                  0.8
                      Overall_Cond -0.01-0.03-0.09: 1 -0.140.17-0.16.00620.120.210.088.0860.090.0320.180.150.0380.017-0.1
                     Mas_Vnr_Area -0.0310.13 0.43-0.14 1 0.4 0.4 0.12 0.4 0.26 0.190.0510.28 0.27 0.36 0.37 0.170.004 0.5
                     Total_Bsmt_SF -0.0290.25 0.55-0.17 0.4 1 0.8 0.21 0.44 0.320.059.0390.28 0.33 0.44 0.49 0.280.072
                                                                                                                                  0.6
                       First_Flr_SF -0.0130.33 0.48-0.16 0.4 0.8 1 -0.25 0.56 0.37 -0.10.0760.39 0.41 0.44 0.49 0.3 0.12
                     Second_Flr_SF -000420330.240.00620.12-0.21-0.25 1 0.66 0.4 0.610.0690.59 0.17 0.18 0.13 0.170.0450.23
                       Gr_Liv_Area = .0098.29 0.57-0.12 0.4 0.44 0.56 0.66 1 0.63 0.43 0.12 0.81 0.45 0.49 0.48 0.37 0.14
                                                                                                                                  0.4
                         Full_Bath -0.0450.13 0.52-0.21 0.26 0.32 0.37 0.4 0.63 1 0.16 0.17 0.53 0.23 0.48 0.41 0.2 0.028
                         Half_Bath -0.040.0350.270.0880.190.0550.1 0.61 0.43 0.16 1 0.0420.35 0.2 0.23 0.18 0.14 0.0150.29
                    0.2
                     TotRms_AbvGr -.002 0.22 0.38-0.09 0.28 0.28 0.39 0.59 0.81 0.53 0.35 0.29 1 0.3 0.36 0.33 0.240.072 0.5
                         Fireplaces -0.0190.26 0.390.0320.27 0.33 0.41 0.17 0.45 0.23 0.2 -0.11 0.3 1 0.32 0.29 0.310.0980.43
                      Garage Cars -0.0360.18 0.6 -0.18 0.36 0.44 0.44 0.18 0.49 0.48 0.23 0.037 0.36 0.32 1 0.89 0.24 0.03 0
                                                                                                                                  0.0
                      Garage_Area -0.0350.21 0.56-0.15 0.37 0.49 0.49 0.13 0.48 0.41 0.180.0580.33 0.29 0.89 1 0.270.053
                       Porch_Area = .00680.2 0.290.0380.17 0.28 0.3 0.17 0.37 0.2 0.14-0.11 0.240.31 0.240.27 1 0.15 0.38
                         Pool Area -0.05 ib 0940.030.0170046.0720.120.0450.140.026.0015.01 ib 0770.0980.030.0530.15 1 0.06
                                                                       Gr_Liv_Area
                                                                           Full_Bath
                                                  Overall_Cond
                                                      las_Vnr_Area
                                                          btal_Bsmt_SF
                                                               First_FIr_SF
                                                                   econd Flr_SF
 In [ ]: ▶
```

#### Step 1: Hypothesize on Features to Engineer

Write a 2-3 sentence hypothesis on the effect you believe each new feature will have on the response variables (Sale\_Price AND log\_Sale\_Price) in a markdown cell.

## **New Feature 1**

Neighbourhood cannot be handled as it is, because it is an object. However, there are so many unique values that it is not a good idea to do one hot encodign. Therefore, here we create a new feature, the average house price for each Neighbourhood. As location is a major determinant of house prices, it will be correlated with the response variable.

```
In [ ]: ▶
In [ ]: ▶
```

### Step 2: Code the New Features

Use the code cell below to develop the features you have hypothesized.

```
In [45]: ► #New Feature 1
```

```
In [40]:
          mapping = df.groupby('Neighborhood')['Sale_Price'].mean().to_dict()
             mapping
   Out[40]: {'Blmngtn': 196661.67857142858,
               Blueste': 143590.0,
               'BrDale': 105608.333333333333,
               'BrkSide': 124756.25,
               'ClearCr': 208662.0909090909,
               'CollgCr': 201803.43445692884,
               'Crawfor': 207550.83495145632,
               'Edwards': 130843.38144329897,
               'Gilbert': 190646.57575757575,
               'Greens': 193531.25,
               'GrnHill': 280000.0,
               'IDOTRR': 103752.90322580645,
               'Landmrk': 137000.0,
               'MeadowV': 95756.48648648648,
               'Mitchel': 162226.63157894736,
               'NAmes': 145097.34988713317,
               'NPkVill': 140710.86956521738.
               'NWAmes': 188406.90839694656,
               'NoRidge': 330319.1267605634,
               'NridgHt': 322018.265060241,
               'OldTown': 123991.89121338913,
               'SWISU': 135071.9375,
               'Sawyer': 136751.1523178808,
               'SawyerW': 184070.184,
               'Somerst': 229707.32417582418,
               'StoneBr': 324229.1960784314,
               'Timber': 246599.54166666666,
               'Veenker': 248314.58333333334}
In [42]: M df['Neighborhood_New'] = df['Neighborhood'].map(mapping)
```

#### Step 3: Check the Results

Develop a correlation matrix or a heatmap to show the linear relationships between your five new features and the response variables (Sale\_Price AND log\_Sale\_Price).

```
In [48]: ▶
                 #plot heatmap
                 plt.figure(figsize=(10,8))
                 sns.heatmap(df.corr(),annot=True)
                 plt.show()
                                                                                                                           -10
                              Order - 1 0.03-0.049.014.034.029.043000442093.049.040.0180026.019.036.0350063805-0.034.06
                            Lot_Area -0.031 1 0.0940.0350.130.25 0.330.0330.29 0.130.0350.020.220.260.18 0.21 0.20.0940.270.16
                        Overall_Qual -0.049.097 1 0.095.430.550.480.240.570.520.27-0.160.380.39 0.6 0.560.290.03 0.8 0.
                                                                                                                           - 0.8
                        Overall_Cond 0.010.036.09: 1 0.140.170.16.006@.120.210.086.0860.090.03@.180.150.038.0170.1-0.24
                       Total_Bsmt_SF_0.029.250.55-0.17 0.4 1 0.8 0.210.44 0.320.058.0390.280.330.44 0.490.280.0720.630.49
                         First_Fir_SF -0.0130.330.48-0.16 0.4 0.8 1 -0.250.56 0.37 -0.10.0760.390.410.44 0.49 0.3 0.12 0.620.45
                                                                                                                           0.6
                       Second_Fir_SF -00002/330.24.00620.12-0.21-0.25 1 0.66 0.4 0.610.0650.59 0.170.18 0.13 0.170.0450.270.15 

Gr_Liv_Area - .0098.29 0.57-0.12 0.4 0.44 0.560.66 1 0.63 0.43 0.12 0.81 0.45 0.49 0.48 0.370.14 0.71 0.47
                           Full_Bath -0.049.130.52-0.210.260.320.37 0.4 0.63 1 0.160.17 0.530.230.48 0.41 0.20.028
                           Half_Bath -0.040.0350.270.0880.190.0550.1 0.610.43 0.16 1 0.0420.35 0.2 0.23 0.180.14.0015 290.23
                       TotRms_AbvGr -0026.220.38-0.090.280.280.390.590.810.530.350.29 1 0.3 0.360.330.240.0720.5 0.3
                                                                                                                            0.2
                           Fireplaces -0.019.260.390.0320.270.33 0.410.170.45 0.23 0.2 -0.11 0.3 1 0.32 0.290.310.0980.470.32
 In [ ]: ▶
```

#### Step 4: Explain the Results

Did your engineered features have the effect that was expected? Explain in 1-2 sentences for each engineered feature.

# **Neighborhood New and Sale Price**

As expected, these two new features are correlated(r=0.76). When actually building machine learning models, you need to worry about over fitting, but that is out of scope, so this is fine here.

In [ ]: M