In[1]:

**import**pandas**as**pd

**import**numpy**as**np

**import**matplotlib.pyplot**as**plt

**import**seaborn**as**sns

**import**plotly.express**as**px

In[2]:

**%**matplotlibinline

**import**warnings

warnings.filterwarnings('ignore')

In[3]:

train **=** pd.read\_csv("../input/house-prices-advanced-regression-techniques/train.csv")test**=**pd.read\_csv("../input/house-prices-advanced-regression-techniques/test.csv")

In[4]:

print(train.head())

IdMSSubClass MSZoningLotFrontageLotArea Street Alley LotShape\

1. 1 60 RL 65.0 8450 Pave NaN Reg
2. 2 20 RL 80.0 9600 Pave NaN Reg2 3 60 RL 68.0 11250 Pave NaN IR13 4 70 RL 60.0 9550 Pave NaN IR14 5 60RL84.014260PaveNaNIR1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LandContour | Utilities ... PoolArea | PoolQC | Fence | MiscFeature | MiscVal | MoSold | \ |
| 0 Lvl | AllPub ... 0 | NaN | NaN | NaN | 0 | 2 |  |
| 1 Lvl | AllPub ... 0 | NaN | NaN | NaN | 0 | 5 |  |
| 2 Lvl | AllPub ... 0 | NaN | NaN | NaN | 0 | 9 |  |
| 3 Lvl | AllPub ... 0 | NaN | NaN | NaN | 0 | 2 |  |
| 4 Lvl | AllPub ... 0 | NaN | NaN | NaN | 0 | 12 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| YrSold | SaleType | SaleCondition | SalePrice |
| 0 2008 | WD | Normal | 208500 |
| 1 2007 | WD | Normal | 181500 |
| 2 2008 | WD | Normal | 223500 |
| 3 2006 | WD | Abnorml | 140000 |
| 4 2008 | WD | Normal | 250000 |

[5 rows x 81 columns]

In[5]:

print(test.head())

Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape \0 1461 20 RH 80.0 11622 Pave NaN Reg

1 1462 20 RL 81.0 14267 Pave NaN IR1

2 1463 60 RL 74.0 13830 Pave NaN IR1

3 1464 60 RL 78.0 9978 Pave NaN IR1

4 1465 120 RL 43.0 5005 Pave NaN IR1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| LandContour | Utilities ... ScreenPorch | PoolArea | PoolQC | Fence | MiscFeature | \ |
| 0 Lvl | AllPub ... 120 | 0 | NaN | MnPrv | NaN |  |
| 1 Lvl | AllPub ... 0 | 0 | NaN | NaN | Gar2 |  |
| 2 Lvl | AllPub ... 0 | 0 | NaN | MnPrv | NaN |  |
| 3 Lvl | AllPub ... 0 | 0 | NaN | NaN | NaN |  |
| 4 HLS | AllPub ... 144 | 0 | NaN | NaN | NaN |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MiscVal | MoSold | YrSold | SaleType | SaleCondition |
| 0 0 | 6 | 2010 | WD | Normal |
| 1 12500 | 6 | 2010 | WD | Normal |
| 2 0 | 3 | 2010 | WD | Normal |
| 3 0 | 6 | 2010 | WD | Normal |
| 4 0 | 1 | 2010 | WD | Normal |

[5 rows x 80 columns]

| index | name | | geocode\_name | | street | | neighborhood | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1812 Theatre | | 1812 Theatre | | Rose Street | | Upper Ferntree Gully | | |
| 3 | ACMI | | ACMI | | Federation Square | | Melbourne | | |
| 5 | Adelphi Theatre | | Mason Theatre | | 14 | | Sir Christopher Ondaatje Avenue | | |
| 7 | Alexander Theatre | | Alexander Theatre (7) (CL48Exh) | | Ancora Imparo Way | | Clayton | | |
| 8 | Altona Drive-In | | Altona Drive | | Mandurah | | Australia | | |
| 9 | Apollo Theatre | | Steve Brittain Performing Arts and Physical Education | | 16 Sharpes Road | | Sharpes Road | | |
| 11 | Argosy Theatre | | Meadow Street | | Guildford | | Midland | | |
| 12 | Armadale Theatre | | Cromwell Street Theatre | | Wilson Street | | South Yarra | | |
| 15 | Arts Theatre | | Lingo Lin Performing Arts Theatre | | Sturgeon Street | | Ormiston | | |
| 16 | Ascot Vale Cinema | | Sun Theatre | | 8 | | Ballarat Street | | |
| 18 | Astor Theatre | | Astor Theatre | | Walcott Street | | Highgate | | |
| 20 | Athenaeum Hall | | Athenaeum Place | | Willsmere | | Kew | | |
| 21 | Athenaeum Theatre | | Athenaeum Theatre | | Collins Street | | Melbourne | | |
| 22 | Austral Theatre | | Amphitheatre | | Shale Hills Loop | | West Hoxton | | |
| 28 | Balwyn Cinema Melbourne | | Stop 46: Balwyn Cinema | | Whitehorse Road | | Balwyn | | |
| 43 | Cameo Cinemas Melbourne | | Cameo Belgrave | | Burwood Highway | | Belgrave | | |
| 50 | Cinema Como Melbourne | | Village Cinemas | | 500 | | Chapel Street | | |
| 52 | Cinema Nova Melbourne | | Cinema Nova | | Cochrane Place | | Carlton | | |
| 53 | Cinema Orient Melbourne | | Village Cinemas | | Overland Drive | | Fountain Gate | | |
| 54 | Cinema le Boulevarde Melbourne | | Village Cinemas | | Williamsons Road | | Doncaster Hill | | |
| 58 | Circle Theatre Melbourne | | Theatre | | 28 | | Isaac Road | | |
| 64 | Coburg Town Hall Melbourne | | Coburg Town Hall | | Urquhart Street | | Coburg | | |
| 65 | Comedy Theatre Melbourne | | Comedy Theatre | | Lonsdale Street | | Melbourne | | |
| 100 | Grand Theatre Melbourne | | Theatre | | 28 | | Isaac Road | | |
| 106 | Hoyts Broadmeadows Melbourne | | Hoyts Cinema | | Pearcedale Parade | | Broadmeadows | | |
| 112 | Hoyts Cinema Centre Melbourne | | 140 Bourke | | 140 | | Bourke Street | | |
| 118 | Hoyts Forest Hill 6 Melbourne | | Hoyts | | 270 | | Canterbury Road | | |
| 119 | Hoyts Frankston Melbourne | | Hoyts Cinema | | Wells Street | | Frankston | | |
| 122 | Hoyts Greensborough Melbourne | | Hoyts Cinema | | Church Street | | Apollo Parkways | | |
| 127 | Hoyts Melbourne Central Melbourne | | Hoyts Melbourne Central | | Queen Elizabeth Walk | | Melbourne | | |
| 140 | Hoyts Richmond Cinema Melbourne | | Hoyts Victoria Gardens | | David Street | | Richmond | | |
| 146 | Hoyts Victoria Gardens Melbourne | | Hoyts Victoria Gardens | | David Street | | Richmond | | |
| index | | name | | geocode\_name | | street | | neighborhood | city | | county | lng |
| 2 | | Churchill National Park | | Churchill National Park | | Stonemasons Track | | Lysterfield South |  | |  | 145.2615914 |
| 3 | | You Yangs Regional Park | | You Yangs Regional Park | | West Track | | Little River |  | |  | 144.4160884 |
| 4 | | Albert Park | | Albert Park | | Canterbury Place | | Middle Park |  | |  | 144.9620409 |
| 5 | | Dandenong Ranges National Park | | Dandenong Ranges National Park | | Old Coach Road | | Mount Dandenong |  | |  | 145.35846 |
| 7 | | Warrandyte State Park | | Warrandyte State Park | | Parkside Court | | Warrandyte Bridge |  | |  | 145.2275017 |
| 9 | | Batman Park | | Batman Park | | Melbourne | | City of Melbourne |  | |  | 144.9564064 |
| 11 | | Point Nepean National Park | | Point Nepean National Park | | Butlers Track | | Portsea |  | |  | 144.6856017 |
| 12 | | Jells Park | | Jells Park | | Wheelers Hill | | City of Monash |  | |  | 145.2012927 |
| 15 | | Victoria Park | | Victoria Park | | Adelaide | | City of Adelaide |  | |  | 138.6207513 |
| 17 | | Flagstaff Gardens | | Flagstaff Gardens | | West Melbourne | | Melbourne |  | |  | 144.9546075 |
| 18 | | Royal Park | | Royal Park | | Adelaide | | South Australia |  | |  | 138.506709 |
| 19 | | Queens Park | | Queens Park | | Perth | | Western Australia |  | |  | 115.9521192 |
| 22 | | Churchill National Park | | Churchill National Park | | Stonemasons Track | | Lysterfield South |  | |  | 145.2615914 |
| 23 | | You Yangs Regional Park | | You Yangs Regional Park | | West Track | | Little River |  | |  | 144.4160884 |
| 24 | | Albert Park | | Albert Park | | Canterbury Place | | Middle Park |  | |  | 144.9620409 |
| 25 | | Dandenong Ranges National Park | | Dandenong Ranges National Park | | Old Coach Road | | Mount Dandenong |  | |  | 145.35846 |
| 27 | | Warrandyte State Park | | Warrandyte State Park | | Parkside Court | | Warrandyte Bridge |  | |  | 145.2275017 |
| 29 | | Batman Park | | Batman Park | | Melbourne | | City of Melbourne |  | |  | 144.9564064 |
| 31 | | Point Nepean National Park | | Point Nepean National Park | | Butlers Track | | Portsea |  | |  | 144.6856017 |
| 32 | | Jells Park | | Jells Park | | Wheelers Hill | | City of Monash |  | |  | 145.2012927 |
| 35 | | Victoria Park | | Victoria Park | | Adelaide | | City of Adelaide |  | |  | 138.6207513 |
| 37 | | Flagstaff Gardens | | Flagstaff Gardens | | West Melbourne | | Melbourne |  | |  | 144.9546075 |
| 38 | | Royal Park | | Royal Park | | Adelaide | | South Australia |  | |  | 138.506709 |
| 39 | | Queens Park | | Queens Park | | Perth | | Western Australia |  | |  | 115.9521192 |
| 42 | | Churchill National Park | | Churchill National Park | | Stonemasons Track | | Lysterfield South |  | |  | 145.2615914 |
| 43 | | You Yangs Regional Park | | You Yangs Regional Park | | West Track | | Little River |  | |  | 144.4160884 |
| 44 | | Albert Park | | Albert Park | | Canterbury Place | | Middle Park |  | |  | 144.9620409 |
| 45 | | Dandenong Ranges National Park | | Dandenong Ranges National Park | | Old Coach Road | | Mount Dandenong |  | |  | 145.35846 |
| 47 | | Warrandyte State Park | | Warrandyte State Park | | Parkside Court | | Warrandyte Bridge |  | |  | 145.2275017 |
| 49 | | Batman Park | | Batman Park | | Melbourne | | City of Melbourne |  | |  | 144.9564064 |
| 51 | | Point Nepean National Park | | Point Nepean National Park | | Butlers Track | | Portsea |  | |  | 144.6856017 |
| 52 | | Jells Park | | Jells Park | | Wheelers Hill | | City of Monash |  | |  | 145.2012927 |
| 55 | | Victoria Park | | Victoria Park | | Adelaide | | City of Adelaide |  | |  | 138.6207513 |
| 57 | | Flagstaff Gardens | | Flagstaff Gardens | | West Melbourne | | Melbourne |  | |  | 144.9546075 |
| 58 | | Royal Park | | Royal Park | | Adelaide | | South Australia |  | |  | 138.506709 |
| 59 | | Queens Park | | Queens Park | | Perth | | Western Australia |  | |  | 115.9521192 |
| 62 | | Churchill National Park | | Churchill National Park | | Stonemasons Track | | Lysterfield South |  | |  | 145.2615914 |
| 63 | | You Yangs Regional Park | | You Yangs Regional Park | | West Track | | Little River |  | |  | 144.4160884 |
| 64 | | Albert Park | | Albert Park | | Canterbury Place | | Middle Park |  | |  | 144.9620409 |
| 65 | | Dandenong Ranges National Park | | Dandenong Ranges National Park | | Old Coach Road | | Mount Dandenong |  | |  | 145.35846 |
| 67 | | Warrandyte State Park | | Warrandyte State Park | | Parkside Court | | Warrandyte Bridge |  | |  | 145.2275017 |
| 69 | | Batman Park | | Batman Park | | Melbourne | | City of Melbourne |  | |  | 144.9564064 |
| 71 | | Point Nepean National Park | | Point Nepean National Park | | Butlers Track | | Portsea |  | |  | 144.6856017 |
| 72 | | Jells Park | | Jells Park | | Wheelers Hill | | City of Monash |  | |  | 145.2012927 |
| 75 | | Victoria Park | | Victoria Park | | Adelaide | | City of Adelaide |  | |  | 138.6207513 |
| 77 | | Flagstaff Gardens | | Flagstaff Gardens | | West Melbourne | | Melbourne |  | |  | 144.9546075 |
| 78 | | Royal Park | | Royal Park | | Adelaide | | South Australia |  | |  | 138.506709 |
| 79 | | Queens Park | | Queens Park | | Perth | | Western Australia |  | |  | 115.9521192 |
| 82 | | Churchill National Park | | Churchill National Park | | Stonemasons Track | | Lysterfield South |  | |  | 145.2615914 |
| 83 | | You Yangs Regional Park | | You Yangs Regional Park | | West Track | | Little River |  | |  | 144.4160884 |
| 84 | | Albert Park | | Albert Park | | Canterbury Place | | Middle Park |  | |  | 144.9620409 |
| 85 | | Dandenong Ranges National Park | | Dandenong Ranges National Park | | Old Coach Road | | Mount Dandenong |  | |  | 145.35846 |

erial)MasVnrType:Masonry veneer type

SaleCondition: Condition of sale

In[6]:

print("train shape:",train.shape)print("testshape:",test.shape)

train shape: (1460, 81)

test shape: (1459, 80)

In[7]:

df**=**pd.concat([train,test],ignore\_index**=True**)print("Combineddatashape",df.shape)

print("\n")

print("Combineddatainfo",df.info())

Combined data shape (2919, 81)

<class 'pandas.core.frame.DataFrame'>RangeIndex: 2919 entries, 0 to 2918

Data columns (total 81 columns):

# Column Non-Null Count Dtype

* 1. Id 2919 non-null int64
  2. MSSubClass 2919 non-null int64
  3. MSZoning 2915 non-null object
  4. LotFrontage 2433 non-null float64
  5. LotArea 2919 non-null int64
  6. Street 2919 non-null object
  7. Alley 198 non-null object
  8. LotShape 2919 non-null object
  9. LandContour 2919 non-null object
  10. Utilities 2917 non-null object
  11. LotConfig 2919 non-null object
  12. LandSlope 2919 non-null object
  13. Neighborhood 2919 non-null object
  14. Condition1 2919 non-null object
  15. Condition2 2919 non-null object
  16. BldgType 2919 non-null object
  17. HouseStyle 2919 non-null object
  18. OverallQual 2919 non-null int64
  19. OverallCond 2919 non-null int64
  20. YearBuilt 2919 non-null int64
  21. YearRemodAdd 2919 non-null int64
  22. RoofStyle 2919 non-null object
  23. RoofMatl 2919 non-null object
  24. Exterior1st 2918 non-null object
  25. Exterior2nd 2918 non-null object
  26. MasVnrType 2895 non-null object
  27. MasVnrArea 2896 non-null float64
  28. ExterQual 2919 non-null object
  29. ExterCond 2919 non-null object

4

In[8]:

train.describe()

Out[8]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Id** | **MSSubClass** | **LotFrontage** | **LotArea** | **OverallQual** | **OverallCond** | **YearBuilt** | **YearRemodAdd** | **MasVnrArea** | **Bsm** |
| **count** | 1460.000000 | 1460.000000 | 1201.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1452.000000 | 1460 |
| **mean** | 730.500000 | 56.897260 | 70.049958 | 10516.828082 | 6.099315 | 5.575342 | 1971.267808 | 1984.865753 | 103.685262 | 443 |
| **std** | 421.610009 | 42.300571 | 24.284752 | 9981.264932 | 1.382997 | 1.112799 | 30.202904 | 20.645407 | 181.066207 | 456 |
| **min** | 1.000000 | 20.000000 | 21.000000 | 1300.000000 | 1.000000 | 1.000000 | 1872.000000 | 1950.000000 | 0.000000 | 0 |
| **25%** | 365.750000 | 20.000000 | 59.000000 | 7553.500000 | 5.000000 | 5.000000 | 1954.000000 | 1967.000000 | 0.000000 | 0 |
| **50%** | 730.500000 | 50.000000 | 69.000000 | 9478.500000 | 6.000000 | 5.000000 | 1973.000000 | 1994.000000 | 0.000000 | 383 |
| **75%** | 1095.250000 | 70.000000 | 80.000000 | 11601.500000 | 7.000000 | 6.000000 | 2000.000000 | 2004.000000 | 166.000000 | 712 |
| **max** | 1460.000000 | 190.000000 | 313.000000 | 215245.000000 | 10.000000 | 9.000000 | 2010.000000 | 2010.000000 | 1600.000000 | 5644 |

8 rows × 38 columns

In[9]:

test.describe()

Out[9]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Id** | **MSSubClass** | **LotFrontage** | **LotArea** | **OverallQual** | **OverallCond** | **YearBuilt** | **YearRemodAdd** | **MasVnrArea** | **Bsmt** |
| **count** | 1459.000000 | 1459.000000 | 1232.000000 | 1459.000000 | 1459.000000 | 1459.000000 | 1459.000000 | 1459.000000 | 1444.000000 | 1458. |
| **mean** | 2190.000000 | 57.378341 | 68.580357 | 9819.161069 | 6.078821 | 5.553804 | 1971.357779 | 1983.662783 | 100.709141 | 439. |
| **std** | 421.321334 | 42.746880 | 22.376841 | 4955.517327 | 1.436812 | 1.113740 | 30.390071 | 21.130467 | 177.625900 | 455. |
| **min** | 1461.000000 | 20.000000 | 21.000000 | 1470.000000 | 1.000000 | 1.000000 | 1879.000000 | 1950.000000 | 0.000000 | 0. |
| **25%** | 1825.500000 | 20.000000 | 58.000000 | 7391.000000 | 5.000000 | 5.000000 | 1953.000000 | 1963.000000 | 0.000000 | 0. |
| **50%** | 2190.000000 | 50.000000 | 67.000000 | 9399.000000 | 6.000000 | 5.000000 | 1973.000000 | 1992.000000 | 0.000000 | 350. |
| **75%** | 2554.500000 | 70.000000 | 80.000000 | 11517.500000 | 7.000000 | 6.000000 | 2001.000000 | 2004.000000 | 164.000000 | 753. |
| **max** | 2919.000000 | 190.000000 | 200.000000 | 56600.000000 | 10.000000 | 9.000000 | 2010.000000 | 2010.000000 | 1290.000000 | 4010. |

8 rows × 37 columns

|  |  |  |
| --- | --- | --- |
| In[10]: | df.isnull().sum() |  |
| Out[10]: | Id | 0 |
|  | MSSubClass | 0 |
|  | MSZoning | 4 |
|  | LotFrontage | 486 |
|  | LotArea | 0 |
|  |  | ... |
| MoSold | | 0 |
| YrSold | | 0 |
| SaleType | | 1 |
| SaleCondition | | 0 |
| SalePrice | | 1459 |

Length: 81, dtype: int64

In[11]:

plt.figure(figsize**=**(12,7))sns.heatmap(df.isnull())

Out[11]:<AxesSubplot:>

In[16]:

df.head()

Out[16]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **MSSubClass** | **MSZoning** | **LotFrontage** | **LotArea** | **Street** | **LotShape** | **LandContour** | **Utilities** | **LotConfig** | **LandSlope... EnclosedPorch** | **3S** |
| **0** 60 | RL | 65.0 | 8450 | Parking | Reg | Lvl | AllPub | Inside | Gtl ... 0 |  |
| **1** 20 | RL | 80.0 | 9600 | Shopping | Reg | Lvl | AllPub | FR2 | Gtl ... 0 |  |
| **2** 60 | RL | 68.0 | 11250 | Shorping centre | IR1 | Lvl | AllPub | Inside | Gtl ... 0 |  |
| **3** 70 | RL | 60.0 | 9550 | Road | IR1 | Lvl | AllPub | Corner | Gtl ... 272 |  |
| **4** 60 | RL | 84.0 | 14260 | Parkings | IR1 | Lvl | AllPub | FR2 | Gtl ... 0 |  |

5 rows × 75 columns

In[17]:

df['LotFrontage']**=**df['LotFrontage'].fillna(df['LotFrontage'].mean())

In[18]:

df['BsmtQual']**=**df['BsmtQual'].fillna(df['BsmtQual'].mode()[0])

df['BsmtCond']**=**df['BsmtCond'].fillna(df['BsmtCond'].mode()[0])

df['GarageType']**=**df['GarageType'].fillna(df['GarageType'].mode()[0])

In[19]:

plt.figure(figsize**=**(12,7))sns.heatmap(df.isnull())

Out[19]:<AxesSubplot:>

In[20]:

df['BsmtExposure'] **=** df['BsmtExposure'].fillna(df['BsmtExposure'].mode()[0])df['BsmtFinType1'] **=** df['BsmtFinType1'].fillna(df['BsmtFinType1'].mode()[0])df['BsmtFinType2'] **=** df['BsmtFinType2'].fillna(df['BsmtFinType2'].mode()[0])df['GarageYrBlt']**=**df['GarageYrBlt'].fillna(df['GarageYrBlt'].mean())

df['GarageFinish'] **=** df['GarageFinish'].fillna(df['GarageFinish'].mode()[0])df['GarageQual']**=**df['GarageQual'].fillna(df['GarageQual'].mode()[0])

df['GarageCond']**=**df['GarageCond'].fillna(df['GarageCond'].mode()[0])

In[21]:

plt.figure(figsize**=**(12,7))sns.heatmap(df.isnull())

Out[21]:<AxesSubplot:>

In[22]:

y**=**df['SalePrice']

df**=**df.drop('SalePrice',axis**=**1)

In[25]:

plt.figure(figsize**=**(12,7))sns.heatmap(df.isnull())

Out[25]:<AxesSubplot:>

In[26]:

df**=**df.join(y,how**=**'left')df

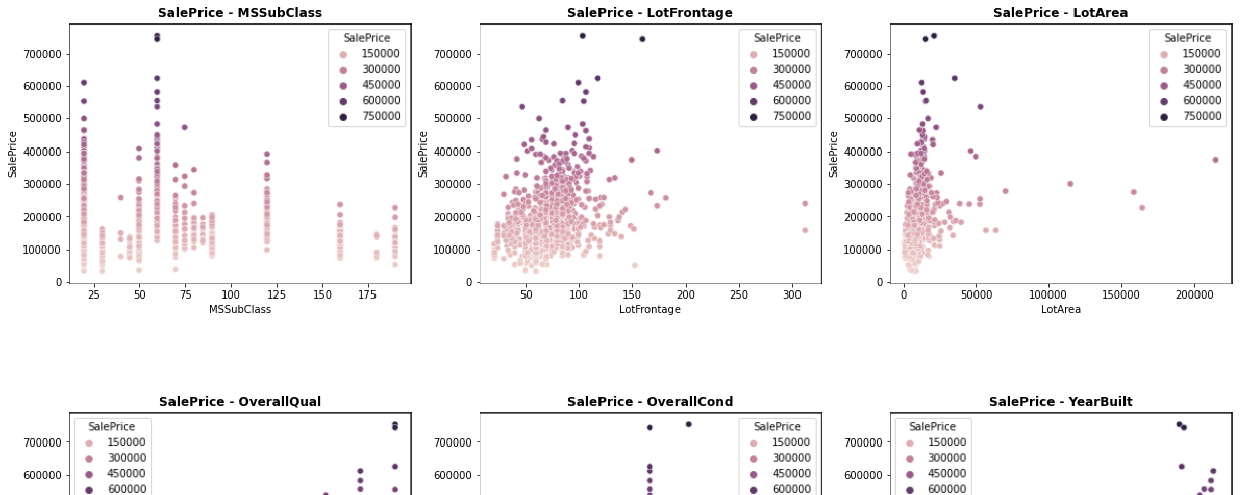
Out[26]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MSSubClass** | **MSZoning** | **LotFrontage** | **LotArea** | **Street** | **LotShape** | **LandContour** | **Utilities** | **LotConfig** | **LandSlope... EnclosedPorch** |
| **0** | 60 | RL | 65.0 | 8450 | Pave | Reg | Lvl | AllPub | Inside | Gtl ... 0 |
| **1** | 20 | RL | 80.0 | 9600 | Pave | Reg | Lvl | AllPub | FR2 | Gtl ... 0 |
| **2** | 60 | RL | 68.0 | 11250 | Pave | IR1 | Lvl | AllPub | Inside | Gtl ... 0 |
| **3** | 70 | RL | 60.0 | 9550 | Pave | IR1 | Lvl | AllPub | Corner | Gtl ... 272 |
| **4** | 60 | RL | 84.0 | 14260 | Pave | IR1 | Lvl | AllPub | FR2 | Gtl ... 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... ... ... |
| **2914** | 160 | RM | 21.0 | 1936 | Pave | Reg | Lvl | AllPub | Inside | Gtl ... 0 |
| **2915** | 160 | RM | 21.0 | 1894 | Pave | Reg | Lvl | AllPub | Inside | Gtl ... 0 |
| **2916** | 20 | RL | 160.0 | 20000 | Pave | Reg | Lvl | AllPub | Inside | Gtl ... 0 |
| **2917** | 85 | RL | 62.0 | 10441 | Pave | Reg | Lvl | AllPub | Inside | Gtl ... 0 |
| **2918** | 60 | RL | 74.0 | 9627 | Pave | Reg | Lvl | AllPub | Inside | Mod ... 0 |

2919 rows × 75 columns

# RecognisingOutliers

In[27]:



df\_number**=**df.select\_dtypes(include**=**['int64','float64'])fig,axs**=**plt.subplots(12,3,figsize**=**(20,80))

fig.subplots\_adjust(hspace**=**0.5)

**for**i,ax**in**zip(df\_number.columns,axs.flatten()):

sns.scatterplot(x**=**i,y**=**'SalePrice',hue**=**'SalePrice',data**=**df\_number,ax**=**ax)plt.xlabel(i,fontsize**=**12)

plt.ylabel('SalePrice',fontsize**=**12)

ax.set\_title('SalePrice'**+**'-'**+**str(i),fontweight**=**'bold',fontsize**=**12)

From the observations above we understand

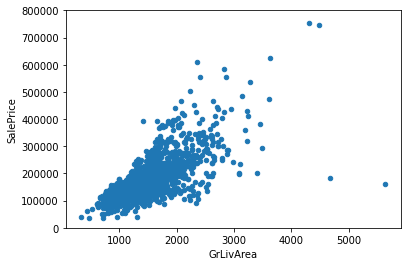
LotAreaforafewisvastinsizebutitscostingsismuchless,thereasonofthiscouldbeunfavourableland,weather,crimesetc.Mostsalesdepends onthearea ofits firstfloor,and secondfloor butmosthouses areobserved tonota secondfloor.

Wesee 2 outliers in Ground living areawhere the costing it too low forits ground area, the same reason oflot area can be used here

In[28]:

df.plot.scatter(x**=**'GrLivArea',y**=**'SalePrice',ylim**=**(0,800000),cmap**=**'viridis')

Out[28]:<AxesSubplot:xlabel='GrLivArea', ylabel='SalePrice'>



The two values with bigger 'GrLivArea' seem strange and they are not following the crowd. We can speculate why this is happening.Maybe they refer to agricultural area and that could explain the low price. I'm not sure about this but I'm quite confident that these twopoints are not representative of the typical case. Therefore, we'll define them as outliers and delete them.

In[29]:

drop\_indexes**=**df[df['SalePrice']**>**700000]df**=**df.drop(drop\_indexes.index,axis**=**0)

In[30]:

sum\_missing\_values**=**df.select\_dtypes(include**=**[np.number]).isnull().sum()sum\_missing\_values[sum\_missing\_values**>**0].sort\_values(ascending**=False**)

Out[30]:SalePrice 1459

MasVnrArea 23

BsmtFullBath 2

BsmtHalfBath 2

BsmtFinSF1 1

BsmtFinSF2 1

BsmtUnfSF 1

TotalBsmtSF 1

GarageCars 1

GarageArea 1

dtype: int64

In[31]:

df.shape

Out[31]:(2917, 75)

**SplittingTestandTrain**

In[32]:

X**=**df.iloc[:1458,:]

test\_1**=**df.iloc[1458:,:]

In[33]:

X.head()

Out[33]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **MSSubClass** | **MSZoning** | **LotFrontage** | **LotArea** | **Street** | **LotShape** | **LandContour** | **Utilities** | **LotConfig** | **LandSlope** | **... EnclosedPorch 3S** |
| **0** 60 | RL | 65.0 | 8450 | Pave | Reg | Lvl | AllPub | Inside | Gtl | ... 0 |
| **1** 20 | RL | 80.0 | 9600 | Pave | Reg | Lvl | AllPub | FR2 | Gtl | ... 0 |
| **2** 60 | RL | 68.0 | 11250 | Pave | IR1 | Lvl | AllPub | Inside | Gtl | ... 0 |
| **3** 70 | RL | 60.0 | 9550 | Pave | IR1 | Lvl | AllPub | Corner | Gtl | ... 272 |
| **4** 60 | RL | 84.0 | 14260 | Pave | IR1 | Lvl | AllPub | FR2 | Gtl | ... 0 |

5 rows × 75 columns

In[34]:

test\_1.head()

Out[34]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MSSubClass** | **MSZoning** | **LotFrontage** | **LotArea** | **Street** | **LotShape** | **LandContour** | **Utilities** | **LotConfig** | **LandSlope... EnclosedPorch** |
| **1460** | 20 | RH | 80.0 | 11622 | Pave | Reg | Lvl | AllPub | Inside | Gtl ... 0 |
| **1461** | 20 | RL | 81.0 | 14267 | Pave | IR1 | Lvl | AllPub | Corner | Gtl ... 0 |
| **1462** | 60 | RL | 74.0 | 13830 | Pave | IR1 | Lvl | AllPub | Inside | Gtl ... 0 |
| **1463** | 60 | RL | 78.0 | 9978 | Pave | IR1 | Lvl | AllPub | Inside | Gtl ... 0 |
| **1464** | 120 | RL | 43.0 | 5005 | Pave | IR1 | HLS | AllPub | Inside | Gtl ... 0 |

5 rows × 75 columns

# DataVisualisation

In[35]:

fig,ax**=**plt.subplots(figsize**=**(14,14))

sns.heatmap(X.corr(),linewidths**=**.5,cmap**=**"YlGnBu")

Out[35]:<AxesSubplot:>

In[36]:

correlation**=**X.select\_dtypes(include**=**[np.number]).corr()

print(correlation['SalePrice'].sort\_values(ascending**=False**))

| index | name | geocode\_name | street | neighborhood | city | county | lng |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Victoria Gardens Shopping Centre | Victoria Gardens Shopping Centre | IKEA Pickup | Richmond |  |  | 145.0107759 |
| 1 | Melbourne Central | Melbourne Central | La Trobe Street | Melbourne |  |  | 144.9625944 |
| 2 | Highpoint Shopping Centre | Highpoint Shopping Centre | 120-200 | Rosamond Road |  |  | 144.888802 |
| 4 | Mountain Gate Shopping Centre | Mountain Gate Shopping Centre | Mountain Gate Drive | Ferntree Gully |  |  | 145.2757407 |
| 5 | Bourke Street Mall | Bourke Street Mall | Centrepoint | Melbourne |  |  | 144.9652982 |
| 7 | Northland Shopping Centre | Northland Shopping Centre | Hannah Street | Preston |  |  | 145.0299296 |
| 9 | Box Hill Central | Box Hill Central (South Precinct) | Carrington Road | Box Hill |  |  | 145.1211564 |
| 10 | Gladstone Park Shopping Centre | Gladstone Park Shopping Centre | North Circular Road | Gladstone Park |  |  | 144.8850525 |
| 11 | Brimbank Shopping Centre | Brimbank Central Shopping Centre | Neale Road | Deer Park |  |  | 144.7750742 |
| 13 | Sanctuary Lakes Shopping Centre | Sanctuary Lakes Shopping Centre | Jamieson Way | Point Cook |  |  | 144.7522644 |
| 14 | Chirnside Park Shopping Centre | Chirnside Park Shopping Centre | Maroondah Highway | Chirnside Park |  |  | 145.312461 |
| 15 | Westfield Fountain Gate | Westfield Fountain Gate Shopping Centre | Overland Drive | Fountain Gate |  |  | 145.3041576 |
| 16 | Wyndham Village Shopping Centre | Wyndham Village Shopping Centre | Sayers Road | Duck Haven |  |  | 144.7048327 |
| 17 | Vermont South Shopping Centre | Vermont South Shopping Centre | Hanvover Road | Vermont South |  |  | 145.1828554 |
| 18 | Broadmeadows Central | Broadmeadows | Terang Street | Dallas |  |  | 144.9401593 |
| 20 | Victoria Gardens Shopping Centre | Victoria Gardens Shopping Centre | IKEA Pickup | Richmond |  |  | 145.0107759 |
| 21 | Melbourne Central | Melbourne Central | La Trobe Street | Melbourne |  |  | 144.9625944 |
| 22 | Highpoint Shopping Centre | Highpoint Shopping Centre | 120-200 | Rosamond Road |  |  | 144.888802 |
| 24 | Mountain Gate Shopping Centre | Mountain Gate Shopping Centre | Mountain Gate Drive | Ferntree Gully |  |  | 145.2757407 |
| 25 | Bourke Street Mall | Bourke Street Mall | Centrepoint | Melbourne |  |  | 144.9652982 |
| 27 | Northland Shopping Centre | Northland Shopping Centre | Hannah Street | Preston |  |  | 145.0299296 |
| 29 | Box Hill Central | Box Hill Central (South Precinct) | Carrington Road | Box Hill |  |  | 145.1211564 |
| 30 | Gladstone Park Shopping Centre | Gladstone Park Shopping Centre | North Circular Road | Gladstone Park |  |  | 144.8850525 |
| 31 | Brimbank Shopping Centre | Brimbank Central Shopping Centre | Neale Road | Deer Park |  |  | 144.7750742 |
| 33 | Sanctuary Lakes Shopping Centre | Sanctuary Lakes Shopping Centre | Jamieson Way | Point Cook |  |  | 144.7522644 |
| 34 | Chirnside Park Shopping Centre | Chirnside Park Shopping Centre | Maroondah Highway | Chirnside Park |  |  | 145.312461 |
| 35 | Westfield Fountain Gate | Westfield Fountain Gate Shopping Centre | Overland Drive | Fountain Gate |  |  | 145.3041576 |
| 36 | Wyndham Village Shopping Centre | Wyndham Village Shopping Centre | Sayers Road | Duck Haven |  |  | 144.7048327 |
| 37 | Vermont South Shopping Centre | Vermont South Shopping Centre | Hanvover Road | Vermont South |  |  | 145.1828554 |
| 38 | Broadmeadows Central | Broadmeadows | Terang Street | Dallas |  |  | 144.9401593 |
| 40 | Victoria Gardens Shopping Centre | Victoria Gardens Shopping Centre | IKEA Pickup | Richmond |  |  | 145.0107759 |
| 41 | Melbourne Central | Melbourne Central | La Trobe Street | Melbourne |  |  | 144.9625944 |
| 42 | Highpoint Shopping Centre | Highpoint Shopping Centre | 120-200 | Rosamond Road |  |  | 144.888802 |
| 44 | Mountain Gate Shopping Centre | Mountain Gate Shopping Centre | Mountain Gate Drive | Ferntree Gully |  |  | 145.2757407 |
| 45 | Bourke Street Mall | Bourke Street Mall | Centrepoint | Melbourne |  |  | 144.9652982 |
| 47 | Northland Shopping Centre | Northland Shopping Centre | Hannah Street | Preston |  |  | 145.0299296 |
| 49 | Box Hill Central | Box Hill Central (South Precinct) | Carrington Road | Box Hill |  |  | 145.1211564 |
| 50 | Gladstone Park Shopping Centre | Gladstone Park Shopping Centre | North Circular Road | Gladstone Park |  |  | 144.8850525 |
| 51 | Brimbank Shopping Centre | Brimbank Central Shopping Centre | Neale Road | Deer Park |  |  | 144.7750742 |
| 53 | Sanctuary Lakes Shopping Centre | Sanctuary Lakes Shopping Centre | Jamieson Way | Point Cook |  |  | 144.7522644 |
| 54 | Chirnside Park Shopping Centre | Chirnside Park Shopping Centre | Maroondah Highway | Chirnside Park |  |  | 145.312461 |
| 55 | Westfield Fountain Gate | Westfield Fountain Gate Shopping Centre | Overland Drive | Fountain Gate |  |  | 145.3041576 |
| 56 | Wyndham Village Shopping Centre | Wyndham Village Shopping Centre | Sayers Road | Duck Haven |  |  | 144.7048327 |
| 57 | Vermont South Shopping Centre | Vermont South Shopping Centre | Hanvover Road | Vermont South |  |  | 145.1828554 |
| 58 | Broadmeadows Central | Broadmeadows | Terang Street | Dallas |  |  | 144.9401593 |
| 60 | Victoria Gardens Shopping Centre | Victoria Gardens Shopping Centre | IKEA Pickup | Richmond |  |  | 145.0107759 |
| 61 | Melbourne Central | Melbourne Central | La Trobe Street | Melbourne |  |  | 144.9625944 |
| 62 | Highpoint Shopping Centre | Highpoint Shopping Centre | 120-200 | Rosamond Road |  |  | 144.888802 |

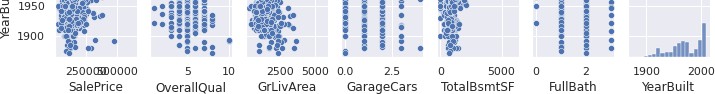
In[37]:

sns.set()

cols**=**['SalePrice','OverallQual','GrLivArea','GarageCars','TotalBsmtSF','FullBath','YearBuilt']sns.pairplot(X[cols],size**=**1.5)

plt.show();



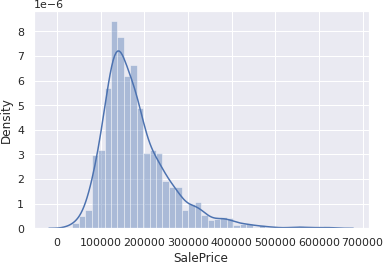


In[38]:

sns.distplot(X['SalePrice'])

print('Skewnessis:%f'**%**X['SalePrice'].skew())print('Kurtosisis:%f'**%**X['SalePrice'].kurt())

|  |  |  |
| --- | --- | --- |
| Skewness | is : | 1.567321 |
| Kurtosis | is : | 3.894915 |



From the pyplot we can clearly understand the following relations

Sales Price and Ground Living Area are directly proportionate that means, the larger the area of the living room more the costing ofthe house.

ThesteepslopeagainstsalespriceandTotalBasementinSquareFeet,thatacertainBasementareaispreferredthusitscostingisthe most

GroundLivingareashowsadirectproportionwithTotalBasementinsquarefeet,thusindicatingthatusuallybiggerareahomeshavebigger basements

In[39]:

fig**=**px.scatter(X,x**=**'TotRmsAbvGrd',y**=**'SalePrice')fig.show()

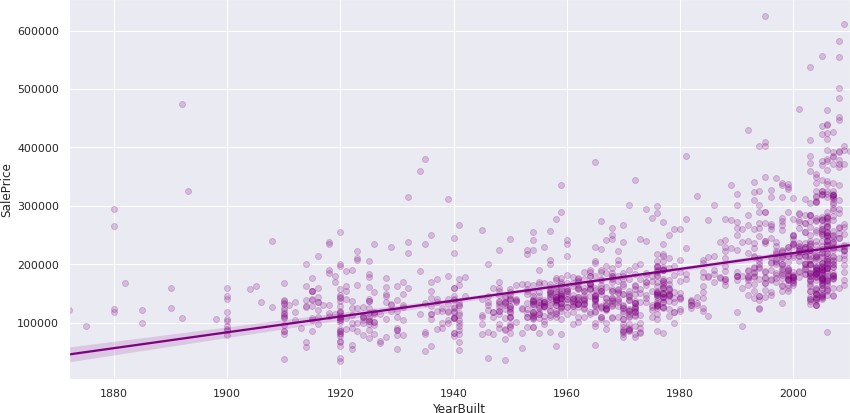
Usually 5-9 have similar Price Ranges. This thus inidicates that a lot of people do not consider the total rooms when looking at house saleprice

In[40]:

fig,ax**=**plt.subplots(figsize**=**(14,7))

sns.regplot(data**=**X,x**=**'YearBuilt',y**=**'SalePrice',scatter\_kws**=**{'alpha':0.2},color**=**'purple')

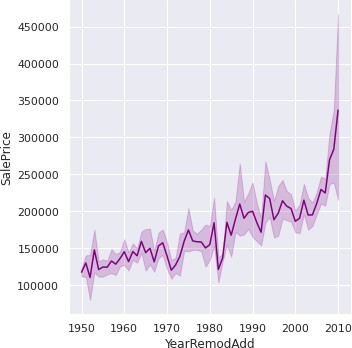
Out[40]:<AxesSubplot:xlabel='YearBuilt', ylabel='SalePrice'>



In[41]:

sns.relplot(data**=**X,x**=**'YearRemodAdd',y**=**'SalePrice',color**=**'purple',kind**=**"line")

Out[41]:<seaborn.axisgrid.FacetGrid at 0x7f8370f58b50>



In[42]:

fig**=**px.pie(X,names**=**'Neighborhood',title**=**"Neighborhoodarea")fig.update\_layout(margin**=**dict(t**=**45,b**=**0,l**=**0,r**=**0))

fig.show()

In[43]:

In[44]:

fig**=**px.pie(X,names**=**'SaleType',title**=**"TypeofSale")fig.update\_layout(margin**=**dict(t**=**45,b**=**0,l**=**0,r**=**0))

fig.show()

In[46]:

**from**sklearn.preprocessing**import**StandardScaler

saleprice\_scaled**=**StandardScaler().fit\_transform(X['SalePrice'][:,np.newaxis]);low\_range**=**saleprice\_scaled[saleprice\_scaled[:,0].argsort()][:10]

high\_range**=** saleprice\_scaled[saleprice\_scaled[:,0].argsort()][**-**10:]print('outerrange (low) of the distribution:')

print(low\_range)

print('\nouter range (high) of the distribution:')print(high\_range)

outer range (low) of the distribution:[[-1.89560696]

[-1.8902428 ]

[-1.85645247]

[-1.83818037]

[-1.82904432]

[-1.67242635]

[-1.6659006 ]

[-1.63327185]

[-1.63327185]

[-1.62031172]]

outer range (high) of the distribution:[[3.73741957]

[3.84835731]

[3.97887229]

[4.19862036]

[4.67060168]

[4.89247715]

[4.91311156]

[5.25704464]

[5.63193587]

[5.80608201]]

1. the low or the negative values are close to zero so that means they might be within the range of distribution
2. The necessary outliers are already removed so leaving these values untouched

# Modeling

In[47]:

x**=**X.drop(['SalePrice'],axis**=**1)y**=**X['SalePrice']

In[48]:

**from**sklearn.preprocessing**import**LabelEncoderle**=**LabelEncoder()

x**=**x.apply(le.fit\_transform)

In[49]:

x.head()

Out[49]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **MSSubClass** | **MSZoning** | **LotFrontage** | **LotArea** | **Street** | **LotShape** | **LandContour** | **Utilities** | **LotConfig** | **LandSlope... OpenPorchSF** | **Enc** |
| **0** 5 | 3 | 36 | 327 | 1 | 3 | 3 | 0 | 4 | 0 ... 49 |  |
| **1** 0 | 3 | 52 | 498 | 1 | 3 | 3 | 0 | 2 | 0 ... 0 |  |
| **2** 5 | 3 | 39 | 702 | 1 | 0 | 3 | 0 | 4 | 0 ... 30 |  |
| **3** 6 | 3 | 31 | 489 | 1 | 0 | 3 | 0 | 0 | 0 ... 24 |  |
| **4** 5 | 3 | 56 | 925 | 1 | 0 | 3 | 0 | 2 | 0 ... 70 |  |

5 rows × 74 columns

In[50]:

y.head()

|  |  |  |
| --- | --- | --- |
| Out[50]: | 0 | 208500.0 |
|  | 1 | 181500.0 |
|  | 2 | 223500.0 |
|  | 3 | 140000.0 |
|  | 4 | 250000.0 |

Name: SalePrice, dtype: float64

In[51]:

**from**sklearn.model\_selection**import**train\_test\_split

X\_train,X\_test,y\_train,y\_test**=**train\_test\_split(x,y,test\_size**=**0.33,random\_state**=**42)

In[52]:

**from**sklearn.preprocessing**import**StandardScalerss**=**StandardScaler()

X\_train**=**ss.fit\_transform(X\_train)X\_test**=**ss.fit\_transform(X\_test)

In[53]:

**from**sklearn.metrics**import**r2\_score,mean\_absolute\_error

**from**sklearn.model\_selection**import**cross\_val\_score

**def** model\_eval(y\_true,y\_pred):r2**=**r2\_score(y\_true,y\_pred)

mae**=**mean\_absolute\_error(y\_true,y\_pred)

**return**r2,mae

In[54]:

# LinearRegression

|  |  |  |
| --- | --- | --- |
| **from**sklearn**import**linear\_model  lr**=**linear\_model.LinearRegression().fit(X\_train,y\_train)cv**=**cross\_val\_score(lr,X\_train,y\_train,cv**=**10)  y\_pred**=**lr.predict(X\_test)cv**=**np.mean(cv)  print("Crossvalscore="**+**str(cv))  r2,mae**=**model\_eval(y\_test,y\_pred) *# model\_eval is the function in the above code block in order to calculat*  print("R^2 score = "**+**str(r2))print("MAEscore="**+**str(mae))  test\_accuracy\_score**=**lr.score(X\_test,y\_test)  print("testaccuracyscore="**+**str(test\_accuracy\_score)) | | |
|  |  |  |

Cross val score = 0.8349802357507636R^2 score = 0.8126923695971509

MAE score = 20721.900796082104

test accuracy score = 0.8126923695971509

# GradientBooster

|  |  |  |
| --- | --- | --- |
| **from** sklearn.ensemble **import** GradientBoostingRegressorgbr **=** GradientBoostingRegressor().fit(X\_train,y\_train)cv**=**cross\_val\_score(gbr,X\_train,y\_train,cv**=**10)  y\_pred**=**gbr.predict(X\_test)cv**=**np.mean(cv)  print("Crossvalscore="**+**str(cv))  r2,mae**=**model\_eval(y\_test,y\_pred) *# model\_eval is the function in the above code block in order to calculat*  print("R^2 score = "**+**str(r2))print("MAEscore="**+**str(mae))  test\_accuracy\_score**=**gbr.score(X\_test,y\_test)  print("testaccuracyscore="**+**str(test\_accuracy\_score)) | | |
|  |  |  |

Cross val score = 0.8590918409453145R^2 score = 0.8614501945688788

MAE score = 16613.224819008876

test accuracy score = 0.8614501945688788

# Predictingtestdata

In[60]:

test\_1.head()

Out[60]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MSSubClass** | **MSZoning** | **LotFrontage** | **LotArea** | **Street** | **LotShape** | **LandContour** | **Utilities** | **LotConfig** | **LandSlope... EnclosedPorch** |
| **1460** | 20 | RH | 80.0 | 11622 | Pave | Reg | Lvl | AllPub | Inside | Gtl ... 0 |
| **1461** | 20 | RL | 81.0 | 14267 | Pave | IR1 | Lvl | AllPub | Corner | Gtl ... 0 |
| **1462** | 60 | RL | 74.0 | 13830 | Pave | IR1 | Lvl | AllPub | Inside | Gtl ... 0 |
| **1463** | 60 | RL | 78.0 | 9978 | Pave | IR1 | Lvl | AllPub | Inside | Gtl ... 0 |
| **1464** | 120 | RL | 43.0 | 5005 | Pave | IR1 | HLS | AllPub | Inside | Gtl ... 0 |

5 rows × 75 columns

In[61]:

test\_2**=**test\_1.drop('SalePrice',axis**=**1)

In[62]:

test\_2**=**test\_2.apply(**lambda**col:le.fit\_transform(col.astype(str)),axis**=**0,result\_type**=**'expand')

In[63]:

test\_2

Out[63]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MSSubClass** | **MSZoning** | **LotFrontage** | **LotArea** | **Street** | **LotShape** | **LandContour** | **Utilities** | **LotConfig** | **LandSlope... OpenPorchSF** |
| **1460** | 5 | 2 | 96 | 187 | 1 | 3 | 3 | 0 | 4 | 0 ... 0 |
| **1461** | 5 | 3 | 97 | 374 | 1 | 0 | 3 | 0 | 0 | 0 ... 134 |
| **1462** | 10 | 3 | 90 | 357 | 1 | 0 | 3 | 0 | 4 | 0 ... 131 |
| **1463** | 10 | 3 | 94 | 1103 | 1 | 0 | 3 | 0 | 4 | 0 ... 134 |
| **1464** | 0 | 3 | 58 | 631 | 1 | 0 | 1 | 0 | 4 | 0 ... 187 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... ... ... |
| **2914** | 2 | 4 | 39 | 491 | 1 | 3 | 3 | 0 | 4 | 0 ... 0 |
| **2915** | 2 | 4 | 39 | 488 | 1 | 3 | 3 | 0 | 4 | 0 ... 103 |
| **2916** | 5 | 3 | 36 | 500 | 1 | 3 | 3 | 0 | 4 | 0 ... 0 |
| **2917** | 14 | 3 | 77 | 57 | 1 | 3 | 3 | 0 | 4 | 0 ... 128 |
| **2918** | 10 | 3 | 90 | 1052 | 1 | 3 | 3 | 0 | 4 | 1 ... 151 |

1459 rows × 74 columns