# 코멘토 직무부트캠프_3주차

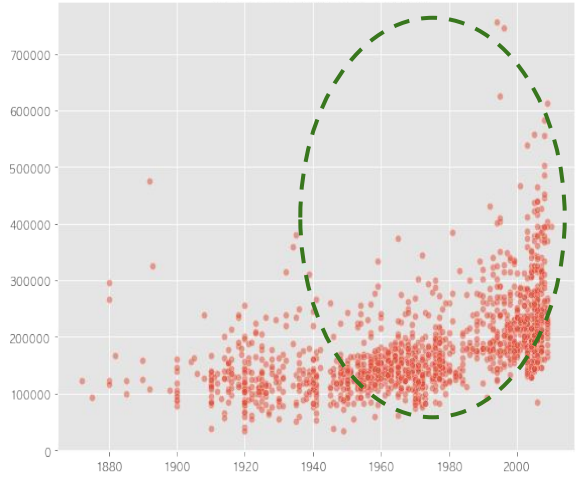
양재영

2021.04.25

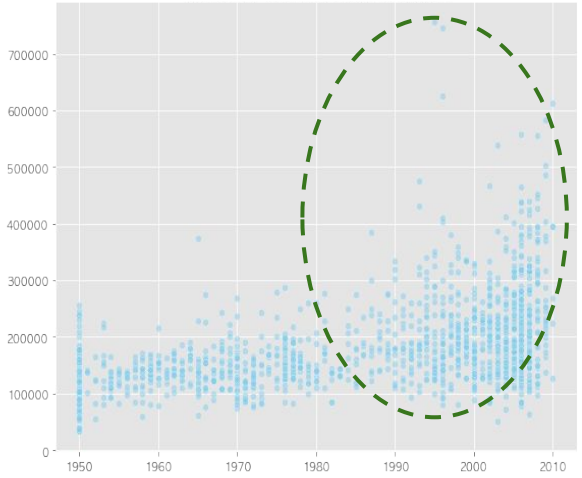
# No. 1 가설 검정

**리모델링 여부에 따라 건물의 가격에 차이가 있을까?**

[Scatterplot of YearBuilt & SalePrice]



[Scatterplot of YearRemodAdd & SalePrice]

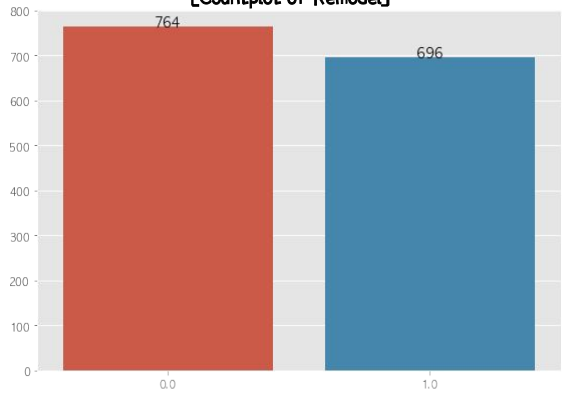


✔ 나중에 건축된 건물일수록, 최근에 리모델링을 한 건물일수록 가격이 높게 분포되어있음을 확인할 수 있다.

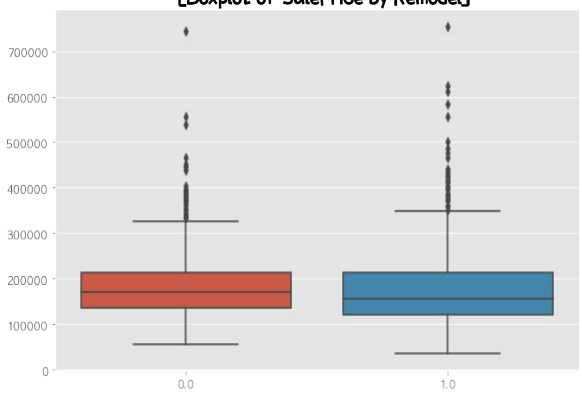
✔ 하지만 YearRemodAdd의 값이 YearBuilt 값과 같으면 리모델링을 하지 않은 건물이다.

✔ 따라서, 리모델링 여부를 나타내는 변수를 만들어 확인해보도록 하자.

[Countplot of Remodel]




[Boxplot of SalePrice by Remodel]



[T-test를 수행하는 코드]

```python
# T_test를 수행하는 함수 생성
def t_test(x = None, y = None, data = None, alpha = None):

    # X의 고유한 값 추출
    first = data[x].unique()[0]
    second = data[x].unique()[1]

    # X의 고유한 값에 해당되는 데이터 추출
    first_data = data[data[x] == first]
    second_data = data[data[x] == second]

    # 표본의 크기가 충분하기 때문에 정규성은 충족한다는 가정하에 'Levene' 등분산 검정을 수행해주도록 한다.
    value, pval = stats.levene(first_data[y], second_data[y])

    # 등분산을 만족할 경우
    if pval > alpha:
        print('등분산 검정 결과 유의확률이 {}으로 유의수준 {}보다 크므로 대립가설을 기각할 수 있다.'.format(round(pval, 5), a

        # T-test 수행
        t_value, t_pval = stats.ttest_ind(first_data[y], second_data[y], equal_var = True)

        print('statistic =', round(t_value, 3), '\n')
        print('P-value =', round(t_pval, 3))


    # 등분산을 만족하지 않을 경우
    else:
        print('등분산 검정 결과 유의확률이 {}으로 유의수준 {}보다 작으므로 대립가설을 기각할 근거가 없다.'.format(pval, alpha),

        # T-test 수행
        t_value, t_pval = stats.ttest_ind(first_data[y], second_data[y], equal_var = False)

        print('statistic =', round(t_value, 3), '\n')
        print('P-value =', round(t_pval, 3))

# 함수 생성 후 적용
t_test(x = 'Remodel', y = 'SalePrice', data = final_data, alpha = 0.05)
```

```
등분산 검정 결과 유의확률이 0.0002890049470259897으로 유의수준 0.05보다 작으므로 대립가설을 기각할 근거가 없다.

statistic = 0.829

P-value = 0.407
```

✔ 리모델링을 한 건물은 696채 이고, 리모델링을 하지 않은 건물은 764채 임을 확인할 수 있다.

✔ 시각화를 통해서 리모델링을 하지 않았을 때 평균이 더 높음을 확인할 수 있으며, 보다 정확한 검증을 위해 T-test를 수행해주도록 하자.

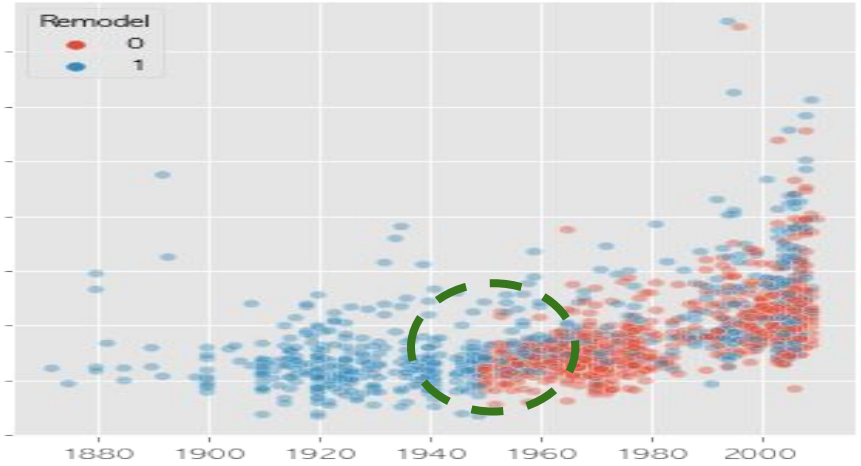
✔ T-test를 수행할 수 있는 함수를 생성한 뒤, T검정을 수행하였고 검정 결과, t통계량 값은 약 0.829이고 유의확률은 0.407 이다.

✔ 즉, 유의확률 0.407이 유의수준 0.05보다 크므로 대립가설을 기각할 수 있는 근거가 있다. 따라서, 리모델링 여부에 따른 가격의 평균 차이는 없다고 할 수 있다.

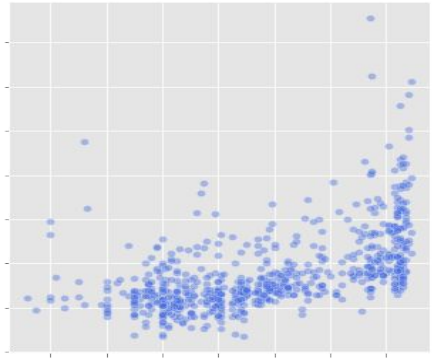
# No. 2 가설 검정

**그렇다면, 언제 지어진 건물들이 리모델링을 했을까?**

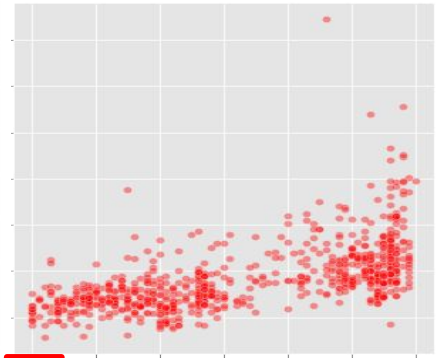
[Scatterplot of YearBuilt & SalePrice by Remodel]




[리모델링을 한 건물]




[리모델링을 하지 않은 건물]

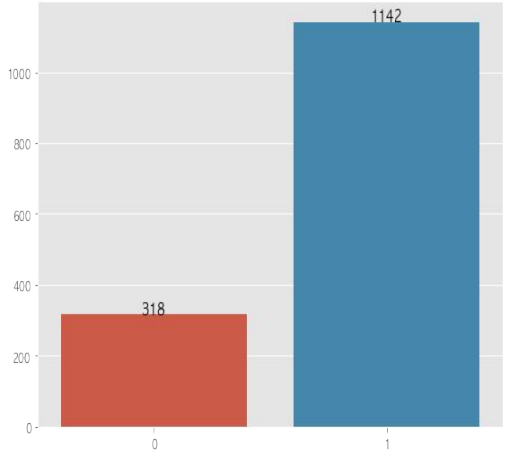



✔ 위의 시각화를 통하여, 1950년도 이전에 지어진 건물은 모두 리모델링이 되었음을 알 수 있다.

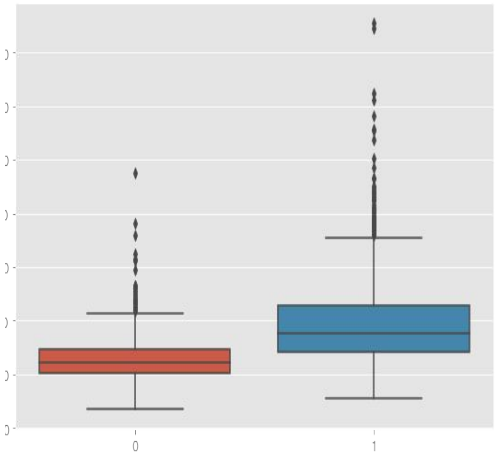
✔ 그렇다면, 1950년 이전에 지어진 건물과 이후에 지어진 건물의 가격에 차이가 있을까?

✔ 만약, 차이가 있다면 1950년 이후에 지어진 건물에서는 리모델링 여부에 따라 가격 차이가 있을까?

[Countplot of Year_1950]




[Boxplot of SalePrice by Year_1950]




[1950년 이전에 지어진 건물과 이후에 지어진 건물에 따른 가격 T-test 수행 코드]

```
1 t_test(x = 'Year_1950', y = 'SalePrice', data = final_data, alpha = 0.05)
```

등분산 검정 결과 유의확률이 9.480014882005363e-08으로 유의수준 0.05보다 작으므로 대립가설을 기각할 근거가 없다.

statistic = 16.092

P-value = 0.0

[1950년 이후에 지어진 건물에서의 리모델링에 따른 가격 T-test 수행 코드]

```
1 t_test(x = 'Remodel', y = 'SalePrice', data = final_data[final_data['Year_1950'] == 1], alpha = 0.05)
```

등분산 검정 결과 유의확률이 3.4735319594033 29e-05으로 유의수준 0.05보다 작으므로 대립가설을 기각할 근거가 없다.

statistic = -6.672

P-value = 0.0

✔ 1950년도 이전과 이후에 지어진 건축물을 나타내는 'Year_1950' 변수를 만들어 주었고, 1950년 이전에 지어진 건물은 318채, 이후에 지어진 건물은 1142채 임을 확인할 수 있다.

✔ 아래의 Boxplot을 통해 1950년 이후에 지어진 건물의 가격 평균이 더 높음을 확인할 수 있으며. T검정 결과 유의확률이 매우 작은 값을 나타내고 있고 유의수준 0.05보다 작으므로 대립가설을 기각할 수 있는 근거가 없다.

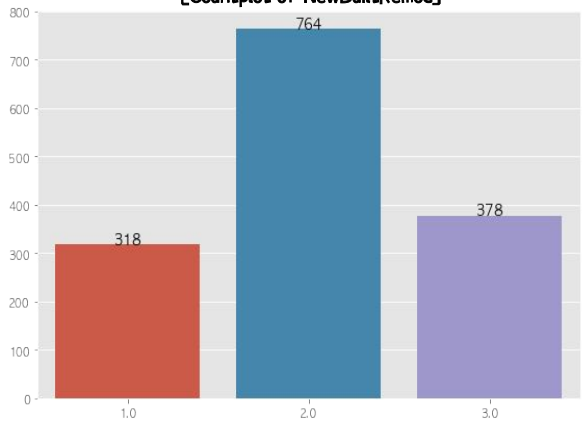
✔ 따라서, 1950년 이전과 이후에 지어진 건물에 따라 가격의 평균에 차이가 있다고 할 수 있다.

✔ 차이가 있음을 확인했으므로, 1950년 이후에 지어진 건물에서의 리모델링에 따른 T검정을 수행해주었다.

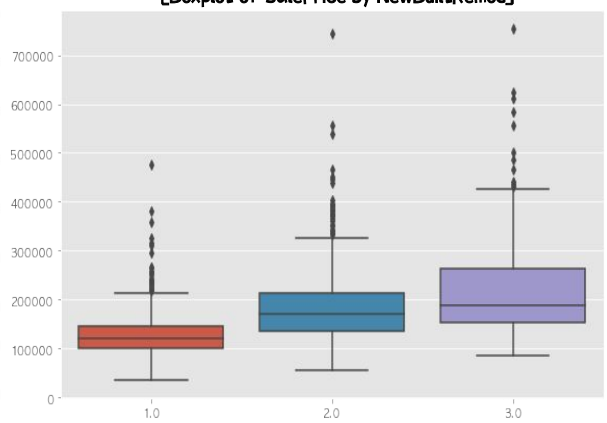
✔ T검정 결과, 유의확률이 매우 작은 값을 나타내고 있고 유의수준 0.05보다 작으므로 대립가설을 기각할 수 있는 근거가 없다. 따라서, 1950년대 이후에 지어진 건물에서의 리모델링에 따른 가격의 평균에 차이가 있다고 할 수 있다.

# No. 3 가설 검정

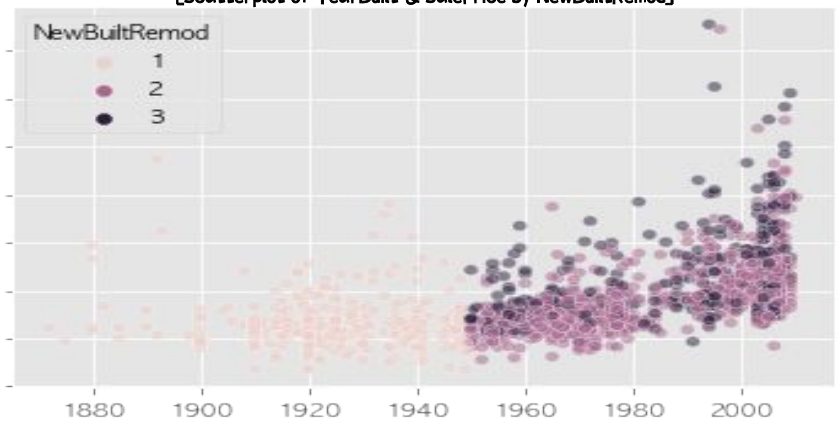
[Countplot of NewBuiltRemod]




[Boxplot of SalePrice by NewBuiltRemod]




[Scatterplot of YearBuilt & SalePrice by NewBuiltRemod]




[분산분석을 수행하는 코드]

```python
# 분산분석을 수행하는 함수 만들기
def anova_test(*args, x = None, y = None, alpha = None, data = None):

    # 우선 정규성 검정을 먼저 해주도록 한다.
    normal_list = []

    for data in args:

        # 정규성 검정을 수행
        n_stats, n_pval = stats.shapiro(data)

        # 정규분포를 따르지 않는 경우
        if n_pval < alpha:
            normal_list.append(True)

        # 정규분포를 따르는 경우
        else:
            normal_list.append(False)

    # 'normal_list'의 합을 구해 정규성을 만족하는지 판단
    sum_normal = np.sum(normal_list)

    # 'normal_list' 값들의 합이 0이면 정규성을 만족한다.
    if sum_normal == 0:

        # 정규성을 만족하는 경우, 'Levene' 등분산 검정
        l_stats, l_pval = stats.levene(*args)

        # 등분산을 만족하는 경우, 'stats' 라이브러리 'f_oneway' 수행
        if l_pval > alpha:
            f_stats, f_pval = stats.f_oneway(*args)
            print('F Statistic :', round(f_stats, 4), '\n')
            print('F P-value :', round(f_pval, 4))

        # 등분산을 만족하지 않는 경우, 'pingouin' 라이브러리의 'welch_anova' 수행
        else:
            pg.welch_anova(dv = y, between = x, data = data)

    # 정규성을 만족하지 않는 경우, 'Kruskal-Wallis H-test' 수행
    else:
        k_stats, k_pval = stats.kruskal(*args)
        print('Kruskal Statistic :', round(k_stats, 4), '\n')
        print('Kruskal P-value :', round(k_pval, 4))

# 분산분석 수행하기 위한 데이터 생성
data1, data2, data3 = [modify_data[modify_data['NewBuiltRemod'] == num]['SalePrice'] for num in [1, 2, 3]]

# 함수 적용
anova_test(data1, data2, data3, alpha = 0.05)
```

```
Kruskal Statistic : 285.1593

Kruskal P-value : 0.0
```

✔ 앞의 결과들을 통하여, 1950년 이전에 지어진 건물은 모두 리모델링을 했음을 알 수 있었다. 또한 1950년 이전에 지어진 건물과 이후에 지어진 건물에 따라 가격에 차이가 있음을 확인했으며, 1950년 이후에 지어진 건물에서 리모델링 여부에 따라 가격에 차이가 있음을 확인하였다.

✔ 따라서, 1950년 이전에 지어진 건물은 1, 1950년 이후에 지어지고 리모델링을 하지 않았다면 2, 1950년 이후에 지어지고 리모델링을 했다면 3을 의미하는 'NewBuiltRemod' 변수를 생성해 주었다.

✔ Boxplot과 Scatterplot을 통해 NewBuiltRemod 범주에 따라 가격에 차이가 있음을 확인할 수 있었고, 보다 정확한 검증을 위해 분산분석을 수행해주었다.

✔ 정규성을 만족하지 않아 'Kruskal-Wallis H-test'를 수행하였으며, 검정 결과 유의확률이 매우 작은 값을 나타내고 있고 유의수준 0.05보다 작으므로 대립가설을 기각할 근거가 없다. 따라서 세 집단에 따라 가격에 차이가 있다고 할 수 있다.

**결과를 종합하면, 리모델링 여부에 따라 가격에 차이가 없는 것으로 밝혀졌지만 그 이유는**
**1950년 이전에 지어진 건물은 모두 리모델링을 하였고, 1950년 이전에 지어진 건물과 이후에 지어진 건물에 따라**
**가격 차이가 있었기 때문이다.**
**또한, 1950년 이후에 지어진 건물에서는 리모델링 여부에 따라 가격 차이가 있음을 확인하였다.**
**즉, 1950년 이전에 지어진 건물은 리모델링을 해야함을 알 수 있고, 1950년 이후에 지어진 건물은 리모델링을 하면 가격이 올라감을**
**알 수 있다.**

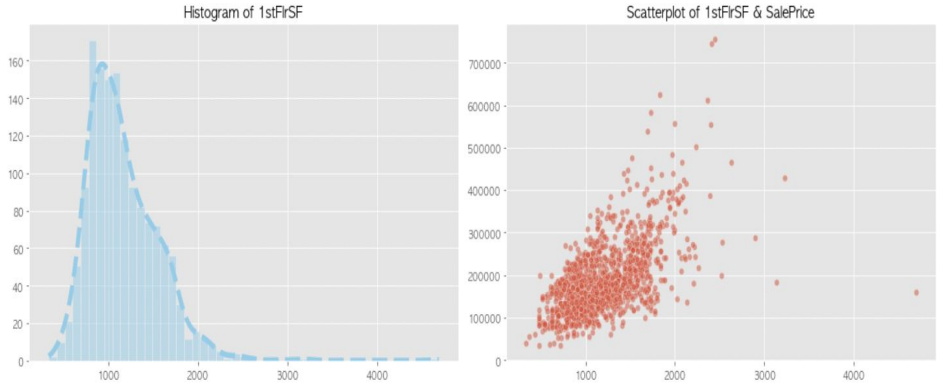
# No. 4 가설 검정

## 2층 여부에 따라 건물의 가격에 차이가 있을까?

[1stFlrSF & SalePrcie]

Correlation cofficient between two variables : 0.606

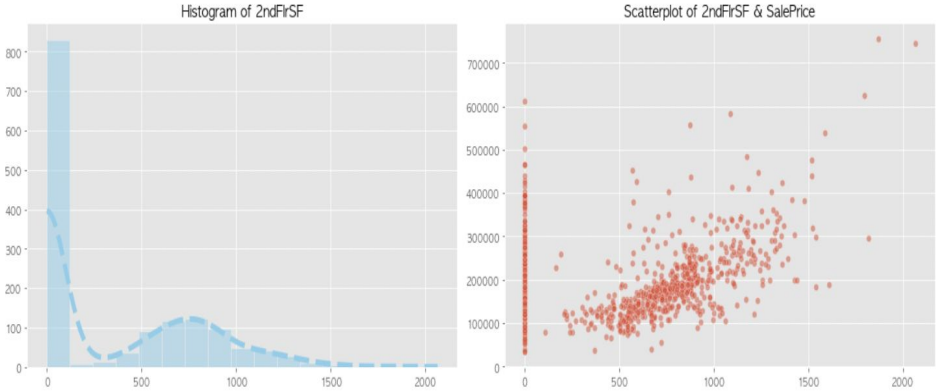
P-value of correlation coefficient between two variables : 0.0




[2ndFlrSF & SalePrcie]

Correlation cofficient between two variables : 0.319

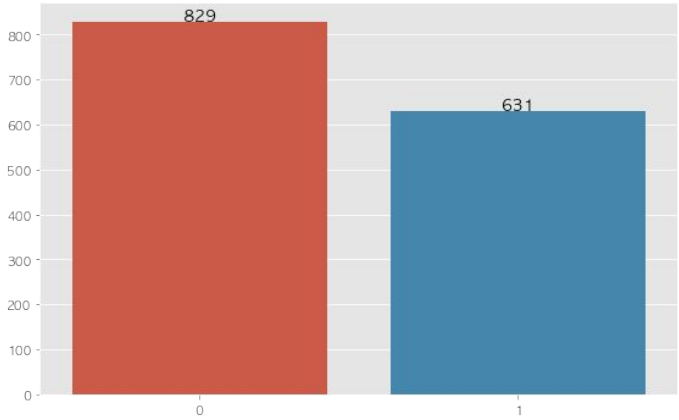
P-value of correlation coefficient between two variables : 0.0



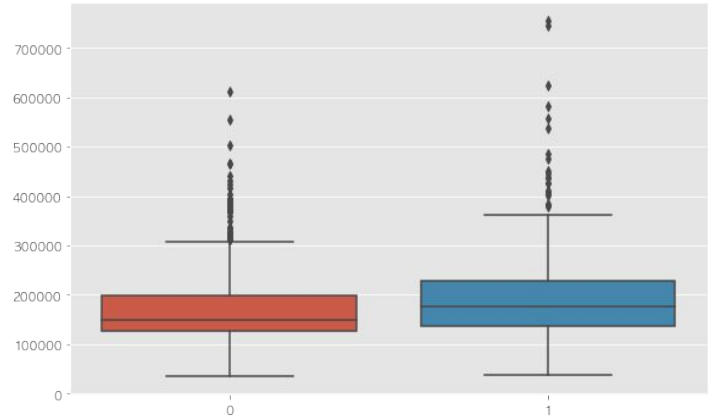

✔ 1stFlrSF, 2ndFlrSF 두 변수 각각 SalePrice와의 상관계수는 0.606, 0.319 이며, 상관분석 결과 두 변수 모두 유의확률이 매우 작으므로 유의수준 0.05 내에서 양의 상관관계가 있다고 할 수 있다.

✔ 2층이 존재하면 당연히 가격이 높을 것이라 생각하였고, 2층 여부를 나타내는 변수를 만들어 시각화와 T검정을 수행해 주었다.

[Countplot of Exist2nd]




[Boxplot of SalePrice by Exist2nd]



```
# 'Exist2nd'에 따른 'SalePrice' 가설 검정 수행
t_test(x = 'Exist2nd', y = 'SalePrice', data = modify_data, alpha = 0.05)
```

등분산 검정 결과 유의확률이 **0.03437581397709266**으로 유의수준 **0.05**보다 작으므로 대립가설을 기각할 근거가 없다.

**statistic = 5.211**

**P-value = 0.0**

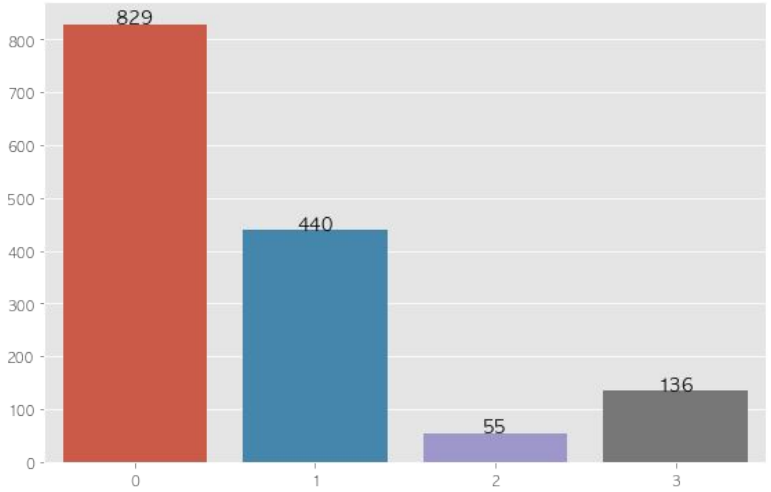
✔ 2층이 있는 건물은 631채이고, 2층이 없는 건물은 829채이다. 또한, 시각화를 통하여 2층이 있는 건물 가격의 평균이 높음을 확인할 수 있다.

✔ T검정 결과, 유의확률이 매우 작은 값을 나타내고 있고 유의수준 0.05보다 작으므로 대립가설을 기각할 근거가 없다. 따라서 2층 여부에 따른 건물 가격의 평균에 차이가 있다고 할 수 있다.

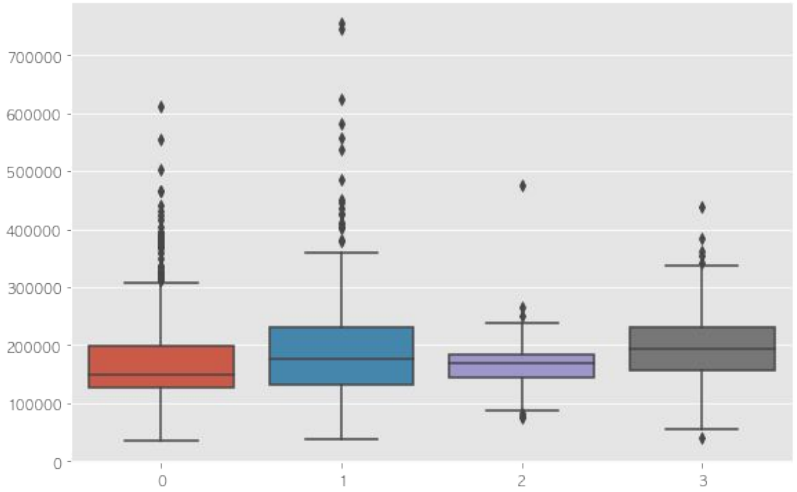
# No. 5 가설 검정

**그렇다면, 1층과 2층의 크기가 변함에 따라 가격에 차이가 있을까?**

[Countplot of ChangeFlr]




[Boxplot of SalePrice by ChangeFlr]




✔ 2층의 여부에 따라 가격에 차이가 있음을 확인하였고, 각 층의 크기에 따라 차이가 있을 것으로 생각하여 1층과 2층의 크기를 비교한 'ChangeFlr' 변수를 생성해 주었다.

✔ 2층을 가지고 있지 않는 경우 '0', 1층의 크기가 2층보다 큰 경우는 '1', 1층과 2층의 크기가 같은 경우는 '2', 2층의 크기가 1층보다 큰 경우는 '3'을 의미한다.

✔ 시각화를 통하여 ChangeFlr이 3일 때, 가격의 평균이 가장 높은 것으로 보이지만 뚜렷한 차이를 확인할 수 없으므로 분산분석을 수행해주도록 하자.

[ChangeFlr에 따른 분산분석 수행 코드]

```
# 분산분석 수행
data1, data2, data3, data4 = [modify_data[modify_data['ChangeFlr'] == num]['SalePrice'] for num in range(4

# 함수 적용
anova_test(data1, data2, data3, data4, alpha = 0.05)

# 사후분석 수행
pval_data = sp.posthoc_conover(a = modify_data, val_col = 'SalePrice',
                               group_col = 'ChangeFlr', p_adjust = 'holm')
display(pval_data)
```

[분산분석 결과]

```
Kruskal Statistic : 46.4186

Kruskal P-value : 0.0
```

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **0** | 1.000000e+00 | 0.000005 | 0.568673 | 3.858987e-08 |
| **1** | 5.174730e-06 | 1.000000 | 0.568673 | 4.256482e-02 |
| **2** | 5.686730e-01 | 0.568673 | 1.000000 | 4.256482e-02 |
| **3** | 3.858987e-08 | 0.042565 | 0.042565 | 1.000000e+00 |

✔ ChangeFlr 집단 중에서 정규성을 만족하지 않는 집단이 존재해 'Kruskal-Wallis H-test' 를 수행하였으며, 검정 결과 유의확률이 유의수준 0.05보다 작으므로 대립가설을 기각할 근거가 없다. 따라서 네 개의 집단에 따라 가격에 차이가 있다고 할 수 있다.

✔ 차이가 존재하여 사후분석을 수행하였고, '0 & 2', '1 & 2'는 통계적으로 유의하지 않음을 알 수 있고, 나머지는 통계적으로 유의미함을 확인할 수 있다.

**결과를 종합하면, 2층 여부에 따라 가격에 차이가 있음을 확인할 수 있었으며,**
**1층과 2층의 크기가 변함에 따라서도 가격에 차이가 있음을 발견하였다.**
**즉, 1층과 2층의 크기에 따른 가격의 분포를 보았을 때 뚜렷한 차이를 확인할 수 없었지만**
**2층이 존재하고 크기가 클 때, 가격이 올라감을 알 수 있다.**

# No. 6 모델링

## 주택 판매가격 예측 모델 구축

[명목형 변수의 변수 선택 과정 코드]

```
# 모델 생성 후 학습시키기
nor_model = sm.OLS(y, X).fit()

# 'P-value'를 저장한 DataFrame 생성
nor_model_data = pd.DataFrame(nor_model.pvalues, columns = ['P-value'])

# 유의수준이 0.05보다 낮은 변수만 저장
nor_model_data = nor_model_data[nor_model_data['P-value'] < 0.05]

# 데이터 확인
nor_model_data.sort_index()
```

[순서형 변수의 변수 선택 과정 코드]

```
# 모델 생성 후 학습시키기
or_model = sm.OLS(y, X).fit()

# 'P-value'를 저장한 DataFrame 생성
or_model_data = pd.DataFrame(or_model.pvalues, columns = ['P-value'])

# 유의수준이 0.05보다 낮은 변수만 저장
or_model_data = or_model_data[or_model_data['P-value'] < 0.05]

# 데이터 확인
or_model_data.sort_index()
```

[수치형 변수의 변수 선택 과정 코드]

```
# 모델 생성 후 학습시키기
nu_model = sm.OLS(y, X).fit()

# 'P-value'를 저장한 DataFrame 생성
nu_model_data = pd.DataFrame(nu_model.pvalues, columns = ['P-value'])

# 유의수준이 0.05보다 낮은 변수만 저장
nu_model_data = nu_model_data[nu_model_data['P-value'] < 0.05]

# 데이터 확인
nu_model_data.sort_index()
```

[명목형 변수 중 유의한 변수]

| | P-value |
|---|---|
| BldgType_Duplex | 1.396778e-02 |
| BldgType_Twnhs | 7.594745e-04 |
| BldgType_TwnhsE | 5.281941e-04 |
| BsmtFinType1_GLQ | 4.539355e-03 |
| BsmtFinType1_No have | 1.485590e-03 |
| BsmtFinType1_Unf | 1.526713e-03 |
| CentralAir_Y | 1.784886e-13 |
| Electrical_SBrkr | 2.957056e-02 |
| Foundation_PConc | 6.052737e-03 |
| Foundation_Stone | 2.750469e-02 |
| GarageFinish_No have | 6.411762e-04 |
| GarageFinish_RFn | 4.549519e-04 |
| GarageFinish_Unf | 9.724526e-11 |
| GarageType_No have | 6.411762e-04 |
| HouseStyle_1.5Unf | 2.506287e-02 |
| HouseStyle_1Story | 1.509706e-07 |
| HouseStyle_2.5Fin | 3.676376e-06 |
| HouseStyle_2.5Unf | 9.068857e-06 |
| HouseStyle_SFoyer | 3.641468e-07 |
| HouseStyle_SLvl | 7.591730e-06 |

[순서형 변수 중 유의한 변수]

| | P-value |
|---|---|
| BedroomAbvGr | 1.072339e-06 |
| BsmtFinSF1 | 9.483817e-03 |
| BsmtTotalBath | 5.355493e-05 |
| ExistPorch | 1.832283e-03 |
| ExistWoodDeck | 6.874801e-07 |
| Fireplaces | 2.813013e-05 |
| GarageArea | 4.480260e-33 |
| GrLivArea | 5.957841e-22 |
| GradeTotalBath | 1.956017e-42 |
| KitchenAbvGr | 2.897138e-24 |
| LogLotArea | 2.825699e-02 |
| MasVnrArea | 4.325122e-02 |
| TotalBsmtSF | 7.446427e-27 |
| const | 0.000000e+00 |

[수치형 변수 중 유의한 변수]

| | P-value |
|---|---|
| FireplaceQu_2 | 1.300058e-03 |
| FireplaceQu_3 | 4.488010e-28 |
| FireplaceQu_4 | 8.012096e-20 |
| FireplaceQu_5 | 5.082575e-05 |
| GarageBuiltRemod_1 | 1.208207e-08 |
| GarageBuiltRemod_2 | 2.233854e-07 |
| GarageBuiltRemod_3 | 1.700893e-07 |
| KitchenQual_4 | 1.440901e-02 |
| KitchenQual_5 | 4.429247e-04 |
| NewBuiltRemod_2 | 3.189815e-08 |
| NewBuiltRemod_3 | 1.323002e-12 |
| NewHeatingQC_1 | 1.211088e-03 |
| OverallQual_10 | 9.681799e-09 |
| OverallQual_5 | 1.785537e-02 |
| OverallQual_6 | 4.851404e-03 |
| OverallQual_7 | 3.153415e-04 |
| OverallQual_8 | 4.944261e-06 |
| OverallQual_9 | 2.143234e-07 |
| const | 0.000000e+00 |

✔ 'Id' 변수를 제외한 변수들을 명목형, 순서형, 이산형, 연속형에 따라 나눠서 분석을 수행하였고, 각 타입에서 시각화와 검정 등을 통하여 유의하다고 판단되는 변수를 선택하였다.

✔ 'statsmodels' 패키지를 사용하여 각 타입에서 유의하다고 판단한 변수와 로그를 취한 종속변수 'SalePrice' 와의 다중회귀분석을 수행하였다.

✔ 회귀분석 수행 후, 유의수준 0.05보다 작은 유의확률을 가지고 있는 변수들만 선택하였으며, 명목형 변수는 227개 중 187개, 순서형 변수는 97개 중 67개, 이산형과 연속형 변수는 167개 중 137개의 변수를 선택하여 총 377개의 독립변수를 선택하였다.

✔ 마지막으로 범주형 변수는 One-Hot Encoding을 수행하였고, 예측 모델을 구축하는 데 있어서 사용할 변수는 총 1247개이다.