Amlan Jyoti Patnaik

June 1, 2021

Original Article

Leveraging Data Analytics Methods For Statistical (ANOVA) Analysis of Factors Influencing US Housing Prices

Contents

Introduction

REITs (Real Estate Investment Trusts) own approximately $3.5 trillion in gross real estate assets, with more than $2 trillion of that total from public listed and non-listed REITs and the remainder from privately held REITs. The economic and investment impact of those assets is felt by millions of Americans all across the country. REITs focus on income-producing and highly appreciating commercial and residential real estate in order to provide high returns to their investors.

REITs assess several factors for determining the true value and the future growth potential of the investment properties. Investment decisions are made based on the outcomes of the research and analysis of the factors influencing the returns on the investment.

In the journey to model housing prices, two approaches have been widely used. The first approach is the monocentric model which assumes that the housing price is a function of its proximity to a single employment center or workplace. The relative housing prices then reflect the relative savings in commuting costs associated with different locations. However, unlike other consumer goods, the housing market is unique because it manifests the characteristics of durability, heterogeneity, and spatial fixity. Thus, to model this differentiation effectively, the second approach of the hedonic price model has been introduced. The hedonic price model posits that goods are typically sold as a package of inherent attributes (Rosen 1974). Therefore, the price of one house relative to another will differ with the additional unit of the different attributes inherent in one house relative to another house. The relative price of a house is then the summation of all its marginal or implicit prices estimated through the regression analysis. The hedonic price model, derived

from Lancaster's (1966) consumer theory and Rosen's theoretical (1974) model, has been used extensively in the scientific investigation of various aspects of housing markets.

Most of the previous researches used regression analysis and were mainly focused on the structural, locational and neighborhood factors and their impact on housing prices. Studies have revealed that the number of rooms and bedrooms (Fletcher, et al. 2000; Li & Brown 1980), the number of bathrooms (Garrod & Willis 1992; Linneman 1980), and the floor area (Carroll, Clauretie, & Jensen 1996; Rodriguez & Sirmans 1994) are positively related to the sale price of houses. This is because buyers are willing to pay more for more functional space. Researchers also proved that building age is negatively related to property prices (Clark & Herrin 2000; Kain & Quigley 1970). Ketkar (1992) observed that whites in New Jersey tended to be sensitive about the proportion of non-whites in their neighborhood. Proximity to shopping centers and the size of shopping centers have both been found to exert an influence on the value of the surrounding residential properties (Des Rosiers, et al. 1996).

While all the factors determined in the previous researches have a significant influence on the housing prices and the expected future return on the investments, the research of the state graduation attainment level factor and its impact on housing prices has not been studied earlier.

This research will utilize ANOVA to analyze the impact of the state graduation attainment levels on the housing prices. The contribution of this research to the field of Data Analytics is to explore whether the graduation attainment levels of the states lead to a significant statistical difference in housing prices which will help the Real Estate Investment Trusts to make informed property investment decisions by considering the state graduation

attainment level as one of the factors for identifying and investing in properties in the states that offer a possibility of higher real estate price appreciation.

The state graduation attainment level is selected for analysis as it may be an indicator of the availability of high paying job opportunities and lower unemployment rates in the state leading to higher housing prices due to higher affordability.

Research Question

This analysis aims to answer this research question – "Is there a statistical difference in mean housing prices based on the graduation attainment levels of the states?"

The following are the Hypothesis for this analysis:

(Null) H0: There is no statistical difference in mean housing prices based on the graduation attainment levels of the states

(Alternate) H1: There is a significant statistical difference in mean housing prices based on the graduation attainment levels of the states

Data Collection

The graduation attainment levels of all the US states and the mean housing prices across all major cities in US for the year 2019 are required for this analysis. Data for this analysis comes from two sources. The data for graduation attainment level is available in data.ers.usda.gov and the mean housing prices across major cities are available in the online real estate marketplace company www.zillow.com. The data is available for public use.

The data from https://data.ers.usda.gov/reports.aspx?ID=17829 includes the following variables for state name and the graduation rate for all US states:

Field	Type	Dependency Type
State	Categorical	Independent
Graduation Rate	Continuous	Independent

The data from www.zillow.com contains 27000 records with the following variables of State, County, City and the target variable of mean housing price for 27000 cities across all US states for multiple years and months including September 2019.

Field	Type	Dependency Type
State Name	Categorical	Independent
County	Categorical	Independent
City	Categorical	Independent
Mean House Price	Continuous	Dependent

Across the country, home sales are recorded by county governments and reported to Internet Data Exchange (IDX) through third-party data providers. Real estate agents also update the home sale price data into the Multiple listing service (MLS) database which in turn updates the data in IDX (mlslistings.com, 2021). This is publicly available data and Zillow receives the data feed from Internet Data Exchange (IDX) (Zillow, 2021). Zillow provides an easier interface to obtain the data in a single platform. Zillow receives price history information from a data feed provided by the MLS or the county government through third-party data providers. While Zillow can edit and/or remove erroneous sales data, it is not able to manually add sales data that has not been reported publicly. The accuracy of the data depends on the data reported by the county and the third-party data providers.

Data Extraction and Preparation

The data from the two data sources will be merged into one final dataset for analysis. Python programming language is used for the analysis as it contains an array of packages to perform statistical analysis and machine learning with very few lines of code with a relatively simpler and readable syntax.

The data for the mean housing prices across 27000 major cities will be programmatically downloaded using the API made available by the online real estate marketplace company www.zillow.com. Any entries that have missing inputs for the variable of mean housing prices will be removed as the number of cases with missing data are negligible and will not impact the analysis (Kang, 2013).

The state graduation rate data is manually extracted from the https://data.ers.usda.gov/reports.aspx?ID=17829 website and saved to a local directory in the computer. The conversion of the continuous variable to categorical variable is required to perform the ANOVA analysis (DeCostera, et al. 2011). The graduation rate for each of the states will be converted to a categorical variable called graduation attainment levels based on the below ranges:

State Graduation Rate	Graduation Attainment Level
0% to 20%	LOW
21% to 40%	MEDIUM
41% to 100%	HIGH

Before converting the graduation rate to the above-mentioned categories, some cleanup would be required to remove unwanted characters and the data type of the column needs to be converted from string to float.

As the first step, all the libraries required for the data collection, preparation and analysis will be imported into the jupyter notebook environment.

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
from scipy import stats
import researchpy as rp

from statsmodels.formula.api import ols
```

Once the required libraries are imported, then pandas read_csv method will be used to read the state graduation rate data and the basic information about the dataset and all the fields will be printed for a quick review of the available fields and to ensure that there are no missing values in the dataset.

As shown in the below screenshot, the data for 52 states are present in the dataset and there are no missing values in the field 'State_Graduation_Rate':

```
grad_attainment = pd.read_csv(r"C:\Users\Amlan\Desktop\WGU Capstone\grad_rates.csv")
```

```
grad_attainment.shape
```

```
(52, 3)
```

```
grad_attainment.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52 entries, 0 to 51
Data columns (total 3 columns):
State_Code              52 non-null object
State                   52 non-null object
State_Graduation_Rate   52 non-null object
dtypes: object(3)
memory usage: 1.3+ KB
```

As we can see from the above information, there are no missing values for the 'State_Graduation_Rate' field

```
grad_attainment['State_Graduation_Rate'].isnull().sum()
```

```
0
```

As can be seen in the above screenshot, the 'State_Graduation_Rate' is a string type object. This field will be converted to float to use the pandas numeric functions to perform the transformation.

The data presented in this dataset is printed for quick review.

```
grad_attainment
```

	State_Code	State	State_Graduation_Rate
0	AL	Alabama	20.50%
1	AK	Alaska	28.20%
2	AZ	Arizona	22.95%
3	AR	Arkansas	20.75%
4	CA	California	28.80%
5	CO	Colorado	40.65%
6	CT	Connecticut	40.50%
7	DE	Delaware	32.00%
8	DC	District of Columbia	58.50%
9	FL	Florida	20.50%
10	GA	Georgia	25.55%

11	HI	Hawaii	31.60%
12	ID	Idaho	26.35%
13	IL	Illinois	27.70%
14	IN	Indiana	23.00%
15	IA	Iowa	27.40%
16	KS	Kansas	30.50%
17	KY	Kentuck	23.05%
18	LA	Louisiana	20.35%
19	ME	Maine	30.90%
20	MD	Maryland	33.90%
21	MA	Massachusetts	42.00%
22	MI	Michigan	26.00%
23	MN	Minnesota	31.25%
24	MS	Mississippi	20.80%
25	MO	Missouri	25.30%
26	MT	Montana	32.50%
27	NE	Nebraska	29.85%
28	NV	Nevada	20.45%
29	NH	New Hampshire	36.40%
30	NJ	New Jersey	41.70%
31	NM	New Mexico	25.30%
32	NY	New York	30.20%
33	NC	North Carolina	27.25%
34	ND	North Dakota	30.20%
35	OH	Ohio	24.20%
36	OK	Oklahoma	24.15%
37	OR	Oregon	28.25%
38	PA	Pennsylvania	25.75%
39	RI	Rhode Island	34.20%
40	SC	South Carolina	23.90%
41	SD	South Dakota	28.90%
42	TN	Tennessee .	20.85%

43	TX	Texas	24.20%
44	UT	Utah	31.25%
45	VT	Vermont	44.50%
46	VA	Virginia	41.90%
47	WA	Washington	31.45%
48	WV	West Virginia	19.85%
49	WI	Wisconsin	27.45%
50	WY	Wyoming	26.95%
51	PR	Puerto Rico	20.80%

The graduation data will now be cleaned and converted to the type float:

```
# remove the special character from the graduation attainment rate
grad_attainment['State_Graduation_Rate_Cleaned']=grad_attainment['State_Graduation_Rate'].str.replace(r'%', '')
```

```
# convert graduation rate to float
grad_attainment['State_Graduation_Rate_Cleaned'] = grad_attainment['State_Graduation_Rate_Cleaned'].astype(float)
```

The basic statistics for the dataset are printed to review the range, high and low values and the mean of the graduation rates.

```
grad_attainment.describe()
```

	State_Graduation_Rate_Cleaned
count	52.000000
mean	29.065385
std	7.623657
min	19.850000
25%	24.087500
50%	27.575000
75%	31.487500
max	58.500000

Now the continuous variable 'State_Graduation_Rate' will be converted into a categorical variable and stored in a new field 'Grad_Attainment_Level' based on the previously discussed ranges using the pandas cut method:

```
# Convert the graduate attainment levels from continuous variable to categorical
# Three categories of 'LOW', 'MEDIUM' and ' HIGH' will be create based on the ranges

grad_attainment['Grad_Attn_Level']=pd.cut(grad_attainment['State_Graduation_Rate_Cleaned'],\
                                          bins=[0,21,40,100],labels=['LOW','MEDIUM','HIGH'])
```

A sample of the converted data is printed below for a quick review:

```
grad_attainment
```

	State_Code	State	State_Graduation_Rate	State_Graduation_Rate_Cleaned	Grad_Attn_Level
0	AL	Alabama	20.50%	20.50	LOW
1	AK	Alaska	28.20%	28.20	MEDIUM
2	AZ	Arizona	22.95%	22.95	MEDIUM
3	AR	Arkansas	20.75%	20.75	LOW
4	CA	California	28.80%	28.80	MEDIUM
5	CO	Colorado	40.65%	40.65	HIGH
6	CT	Connecticut	40.50%	40.50	HIGH
7	DE	Delaware	32.00%	32.00	MEDIUM
8	DC	District of Columbia	58.50%	58.50	HIGH
9	FL	Florida	20.50%	20.50	LOW
10	GA	Georgia	25.55%	25.55	MEDIUM
11	HI	Hawaii	31.60%	31.60	MEDIUM
12	ID	Idaho	26.35%	26.35	MEDIUM
13	IL	Illinois	27.70%	27.70	MEDIUM
14	IN	Indiana	23.00%	23.00	MEDIUM
15	IA	Iowa	27.40%	27.40	MEDIUM
16	KS	Kansas	30.50%	30.50	MEDIUM
17	KY	Kentuck	23.05%	23.05	MEDIUM

The graduation data is now ready to join with the housing price data from Zillow.

The second dataset required for the analysis is the housing price data across all major cities across the United States. The data is manually downloaded from the www.zillow.com API and stored in a local drive for data transformation and analysis.

Pandas read_csv method will be used to read the Zillow home price data. The basic information about the dataset and the fields are printed for a quick review of the available fields.

```
zillow_housing_data = pd.read_csv(r"C:\Users\Amlan\Desktop\WGU Capstone\USA_Cities_Home_Prices\USA_Cities_Home_Prices.csv")
```

```
zillow_housing_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27330 entries, 0 to 27329
Columns: 300 entries, Unnamed: 0 to 2020-03-31
dtypes: float64(291), int64(3), object(6)
memory usage: 62.6+ MB
```

```
zillow_housing_data.columns
```

```
Index(['Unnamed: 0', 'RegionID', 'SizeRank', 'RegionName', 'RegionType',
       'StateName', 'State', 'Metro', 'CountyName', '1996-01-31',
       ...
       '2019-06-30', '2019-07-31', '2019-08-31', '2019-09-30', '2019-10-31',
       '2019-11-30', '2019-12-31', '2020-01-31', '2020-02-29', '2020-03-31'],
      dtype='object', length=300)
```

```
zillow_housing_data.head()
```

StateName	State	Metro	CountyName	1996-01-31	...	2019-06-30	2019-07-31	2019-08-31	2019-09-30	2019-10-31	2019-11-30	2019-12-31
NY	NY	New York-Newark-Jersey City	Queens County	196258.0	...	659421.0	659007.0	658239.0	656925.0	655613.0	654394.0	653930.0
CA	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	185649.0	...	712660.0	713807.0	715688.0	718245.0	721896.0	725180.0	730358.0
TX	TX	Houston-The Woodlands-Sugar Land	Harris County	93518.0	...	186844.0	187464.0	188070.0	188496.0	189125.0	189612.0	190179.0
IL	IL	Chicago-Naperville-Elgin	Cook County	130920.0	...	248372.0	248646.0	248725.0	248483.0	248278.0	248090.0	248029.0
TX	TX	San Antonio-New Braunfels	Bexar County	94041.0	...	182732.0	183350.0	183930.0	184846.0	185490.0	186244.0	186420.0

The dataset is pretty extensive and contains several fields for multiple years and quarters of home prices. As mentioned previously, the data for September 2019 is required for this analysis. A new dataset is created with the fields required for the analysis and the basic information for the new dataset is printed for review:

```
# Select the fields required for this analysis

df_home_prices = zillow_housing_data[['State','CountyName','2019-09-30']]
```

```
df_home_prices.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27330 entries, 0 to 27329
Data columns (total 3 columns):
State         27330 non-null object
CountyName    27330 non-null object
2019-09-30    27330 non-null float64
dtypes: float64(1), object(2)
memory usage: 640.7+ KB
```

```
# Rename the column "2019-09-30" to a more intuitive name called "Home_Price"

df_home_prices = df_home_prices.rename(columns={"2019-09-30": "Home_Price"})
df_home_prices['Home_Price'] = df_home_prices['Home_Price'].astype(int)
```

```
df_home_prices.describe()
```

	Home_Price
count	2.733000e+04
mean	2.046374e+05
std	2.321621e+05
min	1.134900e+04
25%	9.886250e+04
50%	1.515350e+05
75%	2.373352e+05
max	1.332819e+07

The 'Home_Price' column is checked for the presence of any missing or null values. As can be seen in the below screenshot, all the values are present:

```
# No missing values in the Home_Price column

df_home_prices['Home_Price'].isnull().sum()
```

```
0
```

```
df_home_prices
```

	State	CountyName	Home_Price
0	NY	Queens County	656925
1	CA	Los Angeles County	718245
2	TX	Harris County	188496
3	IL	Cook County	248483
4	TX	Bexar County	184846
...	...	...	...
27325	MN	Saint Louis County	68537
27326	MS	Jones County	79930
27327	TX	Clay County	181188
27328	GA	Candler County	98225
27329	PA	Bedford County	89628

27330 rows × 3 columns

MERGE the two datasets:

The two datasets will be joined using state code as the common key to create a new dataframe with all the fields required for the analysis:

```
df = pd.merge(grad_attainment_level,df_home_prices, left_on='State_Code', right_on='State')
```

```
df.head()
```

	State_Code	Grad_Attn_Level	State	CountyName	Home_Price
0	AL	LOW	AL	Mobile County	124175
1	AL	LOW	AL	Jefferson County	61482
2	AL	LOW	AL	Montgomery County	90788
3	AL	LOW	AL	Madison County	171348
4	AL	LOW	AL	Tuscaloosa County	164150

```
df.head()
```

	State_Code	Grad_Attn_Level	State	CountyName	Home_Price
0	AL	LOW	AL	Mobile County	124175
1	AL	LOW	AL	Jefferson County	61482
2	AL	LOW	AL	Montgomery County	90788
3	AL	LOW	AL	Madison County	171348
4	AL	LOW	AL	Tuscaloosa County	164150

The new dataframe 'df' contains the data from both datasets. The data in 'df' belongs to one of the 3 categories (LOW, MEDIUM, HIGH) -

```
df['Grad_Attn_Level'].unique()
```

```
[LOW, MEDIUM, HIGH]
Categories (3, object): [LOW < MEDIUM < HIGH]
```

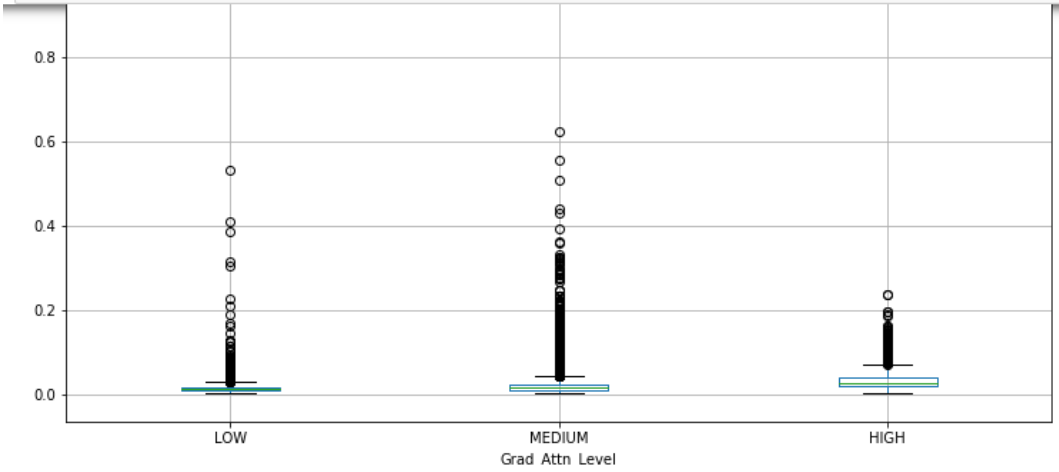
```
df.groupby('Grad_Attn_Level')['Home_Price'].describe().T
```

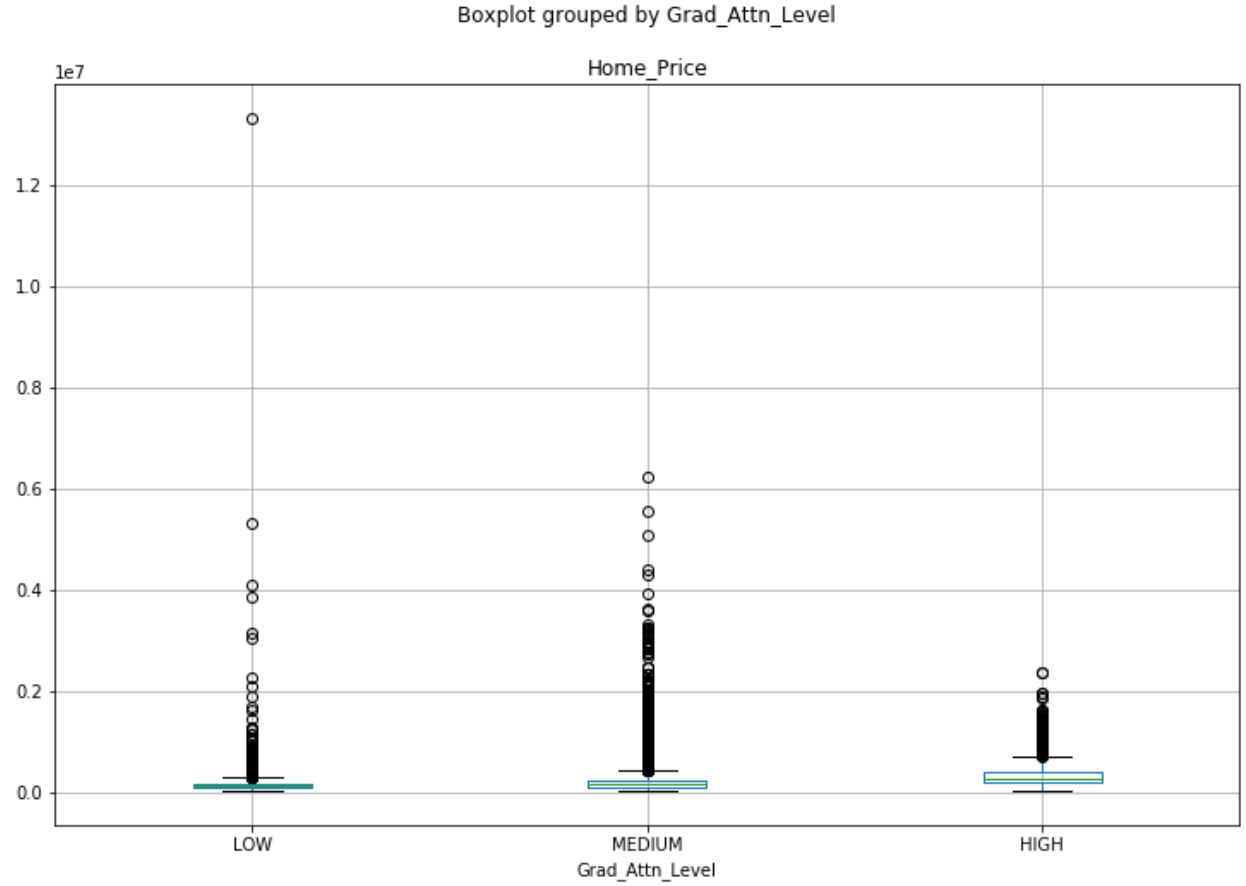
Grad_Attn_Level	LOW	MEDIUM	HIGH
count	3.701000e+03	2.118100e+04	2.448000e+03
mean	1.546723e+05	2.000820e+05	3.195921e+05
std	2.898021e+05	2.169814e+05	2.242582e+05
min	1.134900e+04	1.348800e+04	2.890900e+04
25%	7.993000e+04	9.964500e+04	1.831882e+05
50%	1.133430e+05	1.500710e+05	2.644690e+05
75%	1.697380e+05	2.302520e+05	3.955600e+05
max	1.332819e+07	6.228628e+06	2.360928e+06

A boxplot is generated to review the data distribution of data by Graduation Attainment Level. Boxplot helps to easily detect the differences in means between different Graduation Attainment Levels:

```
# Generate a boxplot to see the data distribution by Graduation Attainment Level
# Boxplot helps to easily detect the differences between different Graduation Attainment Level

df.boxplot(column=['Home_Price'], by='Grad_Attn_Level', figsize=(12, 8))
```





As we can see in the above boxplots, the mean home prices are slightly higher for the states with HIGH graduation attainment level.

Analysis

ANOVA analysis will be performed to answer the below mentioned research question. ANOVA or Analysis of Variance is a statistical test that assumes that the mean across 2 or more groups is equal. If the results don't support the assumption, then the null hypothesis is rejected and the alternate hypothesis is accepted. The test will only determine whether the mean home prices are the same or not across the different groups of 'Graduation Attainment Level'; it cannot determine which groups are different or by how much.

The purpose of a one-way analysis of variance (one-way ANOVA) is to compare the means of two or more groups (the independent variable) on one dependent variable to see if the group means are significantly different from each other (Urdan, 2010).

Research Question: "Is there a statistical difference in mean housing prices based on the graduation attainment levels of the states?"

Hypothesis Testing

(Null) H0: There is no statistical difference in mean housing prices based on the graduation attainment levels of the states

(Alternate) H1: There is a significant statistical difference in mean housing prices based on the graduation attainment levels of the states

The stats method from the SciPy library is used to perform the ANOVA analysis. Since there is only one categorical predictor variable and one continuous outcome variable, one-way ANOVA will be performed.

ANOVA with `scipy.stats`

```
# stats f_oneway functions takes the groups as input and returns ANOVA F and p value

fvalue, pvalue = stats.f_oneway(df['Home_Price'][df['Grad_Attn_Level'] == "LOW"],
                                df['Home_Price'][df['Grad_Attn_Level'] == "MEDIUM"],
                                df['Home_Price'][df['Grad_Attn_Level'] == "HIGH"],)

print(f'F-value : {fvalue}')
print(f'p-value : {pvalue}')
```

```
F-value : 401.3013327628494
p-value : 1.6880447224325337e-172
```

Interpretation of the results

```
# interpretation of the results
alpha = 0.05
if pvalue > alpha:
    print('Same distributions (fail to reject H0)')
else:
    print('Different distributions (reject H0)')
```

```
Different distributions (reject H0)
```

Results from the above analysis clearly show that the p-value obtained from ANOVA is significant ($p < 0.05$), and therefore, we conclude that there are significant differences in home prices based on the graduation attainment levels.

The same results can be achieved using the statsmodel library (shown below).

ANOVA with `statsmodels`

```
model = ols('Home_Price ~ C(Grad_Attn_Level)', data = df).fit()

anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```

	sum_sq	df	F	PR(>F)
C(Grad_Attn_Level)	4.202845e+13	2.0	401.301333	1.688045e-172
Residual	1.430984e+15	27327.0	NaN	NaN

```
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Home_Price   R-squared:                       0.029
Model:                            OLS   Adj. R-squared:                  0.028
Method:                 Least Squares   F-statistic:                     401.3
Date:                Sat, 05 Jun 2021   Prob (F-statistic):          1.69e-172
Time:                        22:42:12   Log-Likelihood:            -3.7605e+05
No. Observations:               27330   AIC:                         7.521e+05
Df Residuals:                   27327   BIC:                         7.521e+05
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                     1.547e+05   3761.507     41.120      0.000    1.47e+05    1.62e+05
C(Grad_Attn_Level)[T.MEDIUM]  4.541e+04   4076.911     11.138      0.000    3.74e+04    5.34e+04
C(Grad_Attn_Level)[T.HIGH]    1.649e+05   5961.539     27.664      0.000    1.53e+05    1.77e+05
==============================================================================
Omnibus:                    50991.347   Durbin-Watson:                   1.491
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        280855771.398
Skew:                          13.636   Prob(JB):                         0.00
Kurtosis:                     498.874   Cond. No.                         6.79
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Data Summary and Implications

One-way ANOVA analysis was performed to assess whether there is a statistical difference in mean housing prices based on the graduation attainment level of the states. The outcome of the analysis proved that the graduation attainment levels of the states significantly influence the home prices. Real Estate Investment Trusts can make investment decisions based on the findings of this study.

The limitation of this research is that it proves that the home prices vary based on the states' graduation attainment levels, but it doesn't determine which groups are different and by how much.

One recommendation for a future course of action is to expand this study to perform a posthoc analysis to determine which groups are different and also join additional housing market variables to perform a regression analysis to quantify the variations in home prices across the states. The study can be further

expanded to a more granular level by performing the analysis at the county or city level to get more actionable insights for making investment decisions.

References

Garrod, G. & Willis, K. (1992). Valuing the goods characteristics – an application of the hedonic price method to environmental attributes, Journal of Environmental Management, vol. 34, no. 1, pp. 59-76.

Linneman, P. (1980). Some empirical results on the nature of the hedonic price function for the urban housing market, Journal of Urban Economics, vol. 8, no. 1, pp. 47 – 68.

Carroll, T. M., Clauretie, T. M. & Jensen, J. (1996). Living next to godliness: Residential property values and churches, Journal of Real Estate Finance and Economics, vol. 12, pp. 319-330.

Rodriguez, M. & Sirmans, C. F. (1994). Quantifying the value of a view in single-family housing markets, Appraisal Journal, vol. 62, pp. 600-603.

Ketkar, K. (1992). Hazardous waste sites and property values in the state of New Jersey, Applied Economics, vol. 24, pp. 647-659.

Clark, D. E. & Herrin, W. E. (2000). The Impact of public-school attributes on home sale price in California, Growth and Change, vol. 31, pp. 385-407.

Des Rosiers, F., Lagana, A., Theriault, M. & Beaudoin, M. (1996). Shopping centers and house values: An empirical investigation, Journal of Property Valuation & Investment, vol. 14, no. 4, pp. 41-62.

Sirpal, R. (1994). Empirical modeling of the relative impacts of various sizes of shopping centers on the value of surrounding residential properties, Journal of Real Estate Research, vol. 9, no. 4, pp. 487-505.

DeCostera, J., Galluccib, M. & Iselinc A. R. (2011). Best Practices for Using Median Splits, Artificial Categorization, and their Continuous Alternatives, vol. 2

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). Applied linear statistical models (5th). New York, NY: McGraw-Hill Irwin.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition, Journal of Political Economy, vol. 82, no. 1, pp. 35-55.

Lancaster, K. J. (1966). A new approach to consumer theory, Journal of Political Economy, vol. 74, pp. 132-157.

Comparison of Python and R: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

Shah, A. (2016). R, Python or SAS. Retrieved May 16, 2021, from https://www.datasciencecentral.com/profiles/blogs/r-python-or-sas-which-one-should-you-learn-first

Multiple Listing Services – MLS (2021) - https://support.mlslistings.com/s/article/MLS-Rules-and-Regulations-FAQ

Zillow.com (2021) - https://zillow.zendesk.com/hc/en-us/articles/214908558-How-does-Zillow-receive-price-history-information.

Kumar, S. (2021). How to know which Statistical Test to use for Hypothesis Testing? https://towardsdatascience.com/how-to-know-which-statistical-test-to-use-for-hypothesis-testing-744c91685a5d

Urdan, T. (2010). Statistics in Plain English, Third Edition, pp. 105.