Supply and Demand Factors Affecting Housing Prices in US

Introduction:

I have used quarterly data from 1st January 2000 till date.

We are considering the following national factors for finding out how much the affect the housing prices in US.

- NASDAQ 100 Index
- Producer Price Index by Commodity: Special Indexes: Construction Materials
- Interest Rates and Price Indexes; Commercial Real Estate Price Index, Level
- Producer Price Index by Commodity: Special Indexes: Construction Materials
- Population
- Unemployment Rate
- Per Capita Income
- Mortgage
- Housing units

Analysis Description:

All the data was imported in python. Correlation of the features with housing index was calculated. A random forest regression model was trained with housing index as the independent variable and the above features as dependent variables.
The model was then used to find out the feature importance.

Analysis:

In [94]:
```
import numpy as np
import pandas as pd
from functools import reduce
```

In [96]:
```
df_index=pd.read_excel('data/CSUSHPINSA wb.xlsx')
df_market=pd.read_excel('data/NASDAQ100 wb.xlsx')
df_pop=pd.read_excel('data/POPTHM wb.xlsx')
df_construction=pd.read_excel('data/WPUSI012011 wb.xlsx')
df_unemp=pd.read_excel('data/UNRATE wb.xlsx')
df_interest=pd.read_excel('data/BOGZ1FL075035503Q wb.xlsx')
df_mortgage=pd.read_excel('data/MORTGAGE30US wb.xlsx')
df_percapinc=pd.read_excel('data/percapinc.xls')
df_hunits=pd.read_excel('data/houseunits.xls')
```

In [107]:
```
data=[df_market,df_pop,df_construction,df_unemp,df_interest,df_mortgage,df_percapinc,df
```

In [108]:
```
for dataset in data:
    print(dataset.shape)
```

```
(90, 2)
(90, 2)
```

```
(90, 2)
(90, 2)
(90, 2)
(90, 2)
(90, 2)
(90, 2)
```

In [109]:
```
for dataset in data:
    print(dataset.columns)
```

```
Index(['Quarter', 'NASDAQ100'], dtype='object')
Index(['Quarter', 'POPTHM'], dtype='object')
Index(['Quarter', 'WPUSI012011'], dtype='object')
Index(['Quarter', 'UNRATE'], dtype='object')
Index(['Quarter', 'Interest_rate'], dtype='object')
Index(['Quarter', 'MORTGAGE30US'], dtype='object')
Index(['Quarter', 'percapinc'], dtype='object')
Index(['Quarter', 'housunits'], dtype='object')
```

In [ ]:
```
```

In [115]:
```
dfs = [df_index,df_market, df_interest, df_construction,df_mortgage,df_pop,df_unemp,df_

#merge all DataFrames into one
df = reduce(lambda  left,right: pd.merge(left,right,on=['Quarter'],
                                            how='outer'), dfs)
```

In [116]:
```
df.head()
```

Out[116]:

	Quarter	CSUSHPINSA	NASDAQ100	Interest_rate	WPUSI012011	MORTGAGE30US	POPTHM	UNR
0	2000 Q1	100.679000	4046.825397	127696	144.733333	8.256923	281304.333333	4.033
1	2000 Q2	103.698667	3629.497460	126924	145.166667	8.316154	282002.000000	3.933
2	2000 Q3	106.459000	3779.890000	139947	143.833333	8.020000	282768.666667	4.000
3	2000 Q4	108.270000	2941.222857	144433	142.833333	7.620769	283518.666667	3.900
4	2001 Q1	109.749333	2168.620952	143811	142.266667	7.006923	284168.666667	4.233

In [131]:
```
df.rename(columns={'CSUSHPINSA':'Index','WPUSI012011':'cons_mat'})
```

Out[131]:

	Quarter	Index	NASDAQ100	Interest_rate	cons_mat	MORTGAGE30US	POPTHM	UNRATE
0	2000 Q1	100.679000	4046.825397	127696	144.733333	8.256923	281304.333333	4.033333
1	2000 Q2	103.698667	3629.497460	126924	145.166667	8.316154	282002.000000	3.933333
2	2000 Q3	106.459000	3779.890000	139947	143.833333	8.020000	282768.666667	4.000000
3	2000 Q4	108.270000	2941.222857	144433	142.833333	7.620769	283518.666667	3.900000
4	2001 Q1	109.749333	2168.620952	143811	142.266667	7.006923	284168.666667	4.233333
...	...	...	...	...	...	...	...	...
85	2021 Q2	255.562000	13797.364762	320999	304.466667	3.003846	332021.000000	5.900000
86	2021 Q3	268.652333	15113.542812	337230	314.587000	2.872143	332296.666667	5.100000
87	2021 Q4	276.157000	15843.419219	351927	328.697333	3.079231	332583.666667	4.233333
88	2022 Q1	288.006000	14576.935968	355629	345.059000	3.822308	332748.666667	3.800000
89	2022 Q2	305.348000	12733.244839	360001	348.737333	5.266154	332939.666667	3.600000

90 rows × 10 columns

In [117]:
```
df.shape
```

Out[117]:
```
(90, 10)
```

In [118]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90 entries, 0 to 89
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Quarter        90 non-null     object 
 1   CSUSHPINSA     90 non-null     float64
 2   NASDAQ100      90 non-null     float64
 3   Interest_rate  90 non-null     int64  
 4   WPUSI012011    90 non-null     float64
 5   MORTGAGE30US   90 non-null     float64
 6   POPTHM         90 non-null     float64
 7   UNRATE         90 non-null     float64
 8   percapinc      90 non-null     int64  
 9   housunits      90 non-null     int64  
dtypes: float64(6), int64(3), object(1)
memory usage: 7.7+ KB
```

In [119]:
```
df.describe(include="all")
```

Out[119]:

	Quarter	CSUSHPINSA	NASDAQ100	Interest_rate	WPUSI012011	MORTGAGE30US	POPTH
count	90	90.000000	90.000000	90.000000	90.000000	90.000000	90.00000
unique	90	NaN	NaN	NaN	NaN	NaN	Na
top	2000 Q1	NaN	NaN	NaN	NaN	NaN	Na
freq	1	NaN	NaN	NaN	NaN	NaN	Na
mean	NaN	169.041648	4100.429369	216455.144444	200.884230	4.985707	310426.67407
std	NaN	41.061039	3610.055121	58960.683183	43.961087	1.359509	16100.87705
min	NaN	100.679000	945.332813	126924.000000	142.033333	2.760714	281304.33333
25%	NaN	141.962583	1633.297222	169301.500000	170.400000	3.887349	296526.66666
50%	NaN	165.653000	2588.831357	212705.500000	201.950000	4.717692	311658.33333
75%	NaN	184.360417	4814.456295	254024.000000	215.683333	6.084038	325399.33333
max	NaN	305.348000	15843.419219	360001.000000	348.737333	8.316154	332939.66666

In [120]:
```
#Finding correlation
df.corr()['CSUSHPINSA'].sort_values(ascending=False)
```

Out[120]:
```
CSUSHPINSA       1.000000
Interest_rate    0.956532
WPUSI012011      0.896744
percapinc        0.877490
NASDAQ100        0.849974
housunits        0.814834
POPTHM           0.771812
UNRATE          -0.262429
MORTGAGE30US    -0.583233
Name: CSUSHPINSA, dtype: float64
```

In [121]:
```
#No missing values
df.isnull().sum()
```

Out[121]:
```
Quarter          0
CSUSHPINSA       0
NASDAQ100        0
Interest_rate    0
WPUSI012011      0
MORTGAGE30US     0
POPTHM           0
UNRATE           0
percapinc        0
housunits        0
dtype: int64
```

In [122]:
```
df.to_excel('data/data.xlsx',index=False)
```

In [123]:
```
x=df.drop(["CSUSHPINSA","Quarter"],axis=True)
y=df["CSUSHPINSA"]
```

In [141]:
```
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
import matplotlib.pyplot as plt
```

In [128]:
```
#training random forest model
reg = RandomForestRegressor(n_estimators=100)
reg.fit(x,y)
```

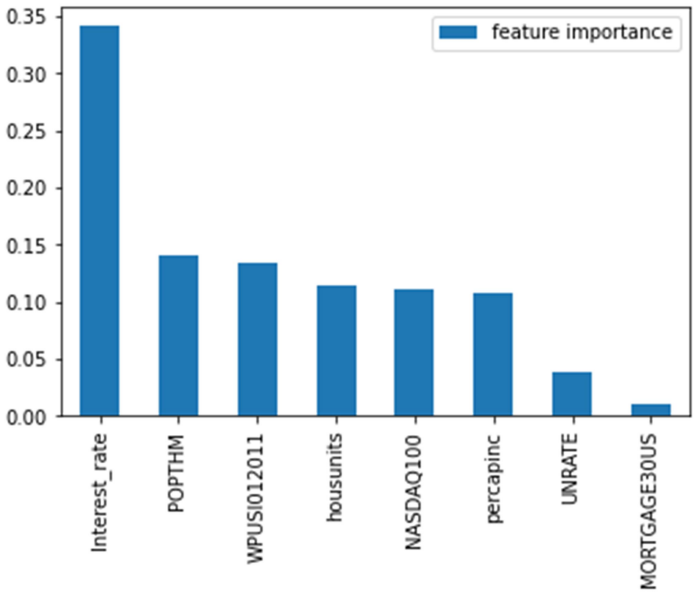
Out[128]:
```
RandomForestRegressor()
```

In [129]:
```
#feature importance
df_feature_importance = pd.DataFrame(reg.feature_importances_, index=x.columns, columns
df_feature_importance
```

Out[129]:

	feature importance
Interest_rate	0.341093
POPTHM	0.140881
WPUSI012011	0.134077
housunits	0.114712
NASDAQ100	0.111757
percapinc	0.107653
UNRATE	0.038619
MORTGAGE30US	0.011209

```
In [130]: df_feature_importance.plot(kind='bar');
```



Results:

To find the most important supply and demand factors that affect the housing prices in the US, I created a decision tree regression model in Python. That model is used to find out the importance of all the features in the model.

A few Important Demand Factors are: Population, economic growth, per capita income

A few Important Supply Factors are: Interest Rates, Construction Materials, Housing units

Further Scope of Improvement:

For improvement we can consider more features like housing affordability and inflation etc. We can also try using more models like linear regression, XGBregressor and then find feature importance for those models.

References:

Data: https://fred.stlouisfed.org/