In [28]:
```
import math
import json
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
%matplotlib inline
```

In [20]:
```
home_prices = pd.read_csv("data/home_price_change_between_2010_2020.csv", usecols=["RegionName", "home_price_ch
home_prices = home_prices.rename(columns={"RegionName":"zipcode"})
```

In [10]:
```
age_sex_race_2015 = pd.read_csv("data/age_sex_race_ACS_2015_cleaned.csv")
age_sex_race_2015

educ_2015 = pd.read_csv("data/educational_attainment_ACS_2015_cleaned.csv", usecols=["zipcode", "total_pop_2015
educ_2015

data_2015 = pd.merge(age_sex_race_2015, educ_2015)
data_2015
```

Out[10]:

	zipcode	male_2015	white_2015	median_age_2015	total_pop_2015	college_pop_2015	college_percent_2015
0	60002	0.52	0.93	40.60	35087.0	9861.0	0.28
1	60004	0.48	0.85	41.80	75286.0	40022.0	0.53
2	60005	0.50	0.84	42.50	45452.0	22056.0	0.49
3	60007	0.49	0.82	42.80	51689.0	18480.0	0.36
4	60008	0.48	0.71	37.20	32497.0	9484.0	0.29
...	...	...	...	...	...	...	...
1378	62996	0.50	0.79	53.20	684.0	66.0	0.10
1379	62997	0.42	0.99	38.90	638.0	10.0	0.02
1380	62998	0.53	1.00	59.10	473.0	12.0	0.03
1381	62999	0.51	0.99	37.70	2501.0	261.0	0.10
1382	63673	0.54	0.97	44.00	3251.0	236.0	0.07

1383 rows × 7 columns

In [11]:
```
age_sex_race_2019 = pd.read_csv("data/age_sex_race_ACS_2019_cleaned.csv")
age_sex_race_2019

educ_2019 = pd.read_csv("data/educational_attainment_ACS_2019_cleaned.csv", usecols=["zipcode", "total_pop_2019
educ_2019

data_2019 = pd.merge(age_sex_race_2019, educ_2019)
data_2019
```

Out[11]:

	zipcode	male_2019	white_2019	median_age_2019	total_pop_2019	college_pop_2019	college_percent_2019
0	60002	0.50	0.91	43.30	35142.0	11452.0	0.33
1	60004	0.48	0.84	43.00	75641.0	43973.0	0.58
2	60005	0.51	0.83	44.40	46307.0	23208.0	0.50
3	60007	0.48	0.83	43.80	51855.0	18224.0	0.35
4	60008	0.50	0.74	37.80	32249.0	11018.0	0.34
...	...	...	...	...	...	...	...
1377	62996	0.48	0.74	42.80	775.0	156.0	0.20
1378	62997	0.51	1.00	32.90	772.0	52.0	0.07
1379	62998	0.50	1.00	35.90	478.0	82.0	0.17
1380	62999	0.49	0.97	37.30	2488.0	208.0	0.08
1381	63673	0.45	0.99	52.20	2684.0	282.0	0.11

1382 rows × 7 columns

In [19]:
```
change_data = pd.merge(pd.merge(data_2015, data_2019), home_prices)
change_data = change_data.astype('float64')
change_data.dtypes

new_cols = ["male","white","median_age","college_percent"]

#change_data["male_change"] = change_data.loc[:, "median_age_2019"] - change_data.loc[:, "median_age_2015"]
for i in new_cols:
    change_data[i+"_change"] = (change_data.loc[:, i+"_2019"] - change_data.loc[:, i+"_2015"])/change_data.loc[

change_data = change_data.dropna()

change_data.head()
```

Out[19]:

	zipcode	male_2015	white_2015	median_age_2015	total_pop_2015	college_pop_2015	college_percent_2015	male_2019	white_
0	60601.0	0.51	0.71	36.2	20688.0	15785.0	0.76	0.47	
1	60602.0	0.50	0.85	25.9	2221.0	1504.0	0.68	0.56	
4	60605.0	0.48	0.60	32.7	44023.0	30963.0	0.70	0.45	
5	60606.0	0.51	0.77	31.5	5561.0	5107.0	0.92	0.51	
6	60607.0	0.50	0.60	30.3	43061.0	29610.0	0.69	0.50	

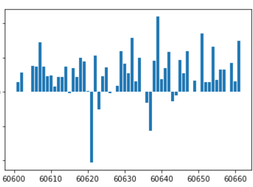
In [30]:
```
change_data = change_data.sort_values("home_price_change_percent") #sort change in home prices in ascending orc

fig, ax = plt.subplots()

y_ax = np.array(change_data['home_price_change_percent'])
x_ax = np.array(change_data["zipcode"])
#x_ax = range(len(change_data['home_price_change_percent']))

ax.bar(x_ax, y_ax)
```

Out[30]:
```
<BarContainer object of 54 artists>
```



In [25]:
```
change_data['home_price_change_percent'].describe()
```

Out[25]:
```
count    54.000000
mean     11.038262
std      13.585594
min     -41.298622
25%       5.893871
50%      11.188944
75%      17.899045
max      43.954938
Name: home_price_change_percent, dtype: float64
```

The highest change in home prices is 43.95%, and this is in zipcode 60639. The lowest change (most negative) in home prices is -41.29%, and this is in zipcode 60621.

In [43]:
```
change_data[["male_change", "white_change", "median_age_change", "college_percent_change"]].describe()
```

Out[43]:

	male_change	white_change	median_age_change	college_percent_change
count	54.000000	54.000000	54.000000	54.000000
mean	0.003445	0.120296	0.033108	0.133410
std	0.035975	0.353610	0.060753	0.145550
min	-0.078431	-0.247059	-0.071823	-0.176471
25%	-0.020306	-0.025183	0.000663	0.039044
50%	0.000000	0.005618	0.019554	0.103914
75%	0.020408	0.116667	0.038244	0.190616
max	0.120000	2.000000	0.316602	0.666667

In [35]:
```
train = change_data.iloc[:40]
test = change_data.iloc[40:]
```

In [44]:
```
X = np.array(train[["male_change", "white_change", "median_age_change", "college_percent_change"]])
y = np.array(train["home_price_change_percent"])
model = LinearRegression().fit(X, y)


r_sq = model.score(X, y)
print('coefficient of determination:', r_sq)

# Print the Intercept:
print('intercept:', model.intercept_)

# Print the Slope:
print('slope:', model.coef_)
```

```
coefficient of determination: 0.17158207068220865
intercept: 10.035233582489628
slope: [ 13.81504412  -3.93741661   1.56428335 -29.70748188]
```

In [48]:
```
y_hat = 10.035233582489628 + 13.81504412*train['male_change'] + -3.93741661*train["white_change"] + 1.56428335*
```

In [51]:
```
from sklearn.metrics import mean_squared_error
```

In [52]:
```
mean_squared_error(train["home_price_change_percent"], y_hat)
```

Out[52]:
```
102.03364106986339
```

In [54]:
```
y_hat_test = 10.035233582489628 + 13.81504412*test['male_change'] + -3.93741661*test["white_change"] + 1.564283
```

In [56]:
```
mean_squared_error(test["home_price_change_percent"], y_hat_test)
```

Out[56]:
```
510.4463699614198
```