Group 1 Project Final Paper

Real-Estate agents and managers

December 5, 2023

BAN 4550 – Analytics Programming

Professor Yoon

Members:

Sitanshu Bakshi

Parimala Yerraguntla

Samuel Owusu

Introduction

Team member names: Samuel Owusu, Parimala Yerraguntla, Sitanshu Bakshi
Summary of questions: The questions revolve around the influence of economic indicators on the financial performance of businesses in the real estate agents and manager's sector.

Background

Overview of the industry

SIC code and Industry

6531 – Real Estate Agents & Managers

Industry Description

This industrial classification code is for businesses predominantly involved in helping individuals or other organizations rent, buy, sell, manage, and value real estate. Real estate agents, brokers, rentals, and escrow agents are in this type of business.

https://www.referenceforbusiness.com/industries/Finance-Insurance-Real-Estate/Real-Estate-Agents-Managers.html

Companies in the interested industry and its SEC EDGAR page

1. **RE/MAX Holdings, Inc.**
   SEC EDGAR: https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=0001581091&owner=include&count=40&hidefilings=0
   Description: RE/MAX is a Denver, CO based real estate firm that operates on a franchise business model. It operates globally and has a presence in over 110 countries. At present, RE/MAX has more than 140,000 agents in over 9,000 office locations around the globe.

   https://www.remaxholdings.com/

2. **JONES LANG LASALLE INC**
   SEC EDGAR: https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=0001037976&owner=include&count=40&hidefilings=0
   Description: A well-known international provider of real estate services, JLL (Jones Lang LaSalle Incorporated) was founded in the UK and currently has operations in 80 different nations. In addition to its core real estate services, the company provides global investment management services to a wide range of clients, including high-net-worth individuals, institutional investors, and retail investors. Through JLL Technologies, JLL expands its range of services to include technology solutions, and JLL Spark, its PropTech fund, invests in startup capital. JLL is listed as the 190th largest company on the Fortune 500 list.

https://en.wikipedia.org/wiki/JLL_(company)

Research question with supporting evidence

Description of the interested economic indicator.

We are a group of three and we have the following three economic indicators in our research:

1. **Interest Rate** (https://fred.stlouisfed.org/series/DFF)
   Description: The interest rate is the agreed-upon rate of interest that the borrower must pay back to the lender on top of the principal amount (amount owed). In the United States financial sector, the federal funds rate is the fundamental interest rate.
   https://fred.stlouisfed.org/series/DFF

2. **Population Level** (https://fred.stlouisfed.org/series/LNU00000060)
   Description: The term "population" refers to all citizens who are either permanently residing in a country or who are just passing through. This indicator reveals how many people typically reside in a certain area. Growth rates are the population changes that occur each year because of births, deaths, and net migration.

   https://data.oecd.org/pop/population.htm#:~:text=Population%20is%20defined%20as%20all,net%20migration%20during%20the%20year

3. **Labor Force Flows Employed to Unemployed**.
   https://fred.stlouisfed.org/series/LNS17400000
   Description: Labor force flows depict the underlying motions of net over-the-month changes in employment, unemployment, or absence from the labor force.
   https://fred.stlouisfed.org/series/LNS17400000

Research questions and corresponding evidence with the summary

We are trying to find answers to the following three research questions:

1. The performance of firms in the real estate industry is inversely proportional to the Federal fund effective rate (interest rate), so if the effective rate goes down, it positively affects the real estate managers and agents.
2. The performance of firms in the Real-estate agents and managers industry is positively affected by the population of the US economy.
3. Does the labor force flow from employed to unemployed have a significant impact on the nation of America and the real estate industry?

We have taken revenue as the dependent variable and our expectations on revenue are the following:

1. **The interest rate (independent variable) has a negative effect on the revenue (dependent variable) of the companies, so an increase in interest rate decreases the**

revenue of the Real-Estate businesses.

2. **The population (independent variable) has a positive impact on the revenue (dependent variable) of the companies, i.e. the population increase has an impact on the revenue of organizations.**

3. **The household income (independent variable) has a positive impact on the revenue (dependent variable) of the companies; as the household income increases, the revenue of organizations in Real-Estate also increases.**

Data

What is FRED?**:** FRED is an abbreviation for Federal Reserve Economic Data, a database managed and maintained by the Federal Reserve Bank. It is an online resource containing economic information gathered from a wide range of government, public, and private sources. https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/

What is Compustat? Compustat, published by Standard and Poor's (S&P), is a financial database and information provider serving global organizations. It offers information on business operations, pricing, earnings, and financial data, with historical records dating back to the 1950s. https://www.investopedia.com/terms/c/compustat.asp

Data Preparation

We import economic data using the Fredapi package and Computstat data from the WRDS package. First, we import the WRDS data from 1980 as the data from 2010 was not sufficient for our research, WRDS data is common for our research, so it was fetched only once. Here is the

snapshot of WRDS data import.

Bringing Data from 1980

For further analysis, we are brining data from 1980 to 2022 so we can perform further regression analysis.

Import WRDS data

In [68]:
```
# import WRDS package
import wrds
```

In [69]:
```
# Setup connection with WRDS server.
conn = wrds.Connection()
```

```
Enter your WRDS username [sitanshubakshi]: sbakshi
Enter your password: ·····
WRDS recommends setting up a .pgpass file.
Create .pgpass file now [y/n]?:  n
You can create this file yourself at any time with the create_pgpass_file() function.
Loading library list...
Done
```

In [70]:
```
# Read data from 1980 to 2022.
wrds_1980_2022 = conn.raw_sql('''select cik, gvkey, datadate, conm, revt, ni from comp.funda
                        where sich=6531 and datadate>='01/01/1980' and datadate<='12/31/2022'
                        and  datafmt = 'STD' and consol = 'C'and indfmt = 'INDL'
                        ''', date_cols=['datadate'])

# Summary of data
# info method prints information about DataFrame (wrds_1980_2022) - datatypes, range, data columns.
wrds_1980_2022.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 572 entries, 0 to 571
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   cik       528 non-null    object
 1   gvkey     572 non-null    object
 2   datadate  572 non-null    datetime64[ns]
 3   conm      572 non-null    object
 4   revt      565 non-null    float64
 5   ni        565 non-null    float64
dtypes: datetime64[ns](1), float64(2), object(3)
memory usage: 26.9+ KB
```

Figure 1: Import WRDS data

Second, we import FRED data, FRED contains data for our economic indicators – interest rate, population levels, and flow of labor from employment to unemployment. We import FRED data separately for each economic indicator. Here is the snapshot of FRED data import for interest

rate.

Import FRED data

In [74]:
```
from fredapi import Fred # Import Fred package from fredapi
mykey='ea53013a18f3522101c4235350d71709' # Fred API key
fred = Fred(api_key=mykey) # Set API key
econ_var = 'DFF' # name of abbriviation of interest rate (FRB rates) economic data
```

Import data

Import `Federal Funds Effective Rate` as `dff_1980_2022` from '1980-01-01'

In [75]:
```
# import data from 1980-01-01
dff_1980_2022=fred.get_series(econ_var, observation_start='1980-01-01')

# Summary of data for DFF economic variable.
dff_1980_2022.info() # info method provides the summary of data.
```

```
<class 'pandas.core.series.Series'>
DatetimeIndex: 16040 entries, 1980-01-01 to 2023-11-30
Series name: None
Non-Null Count  Dtype
--------------  -----
16040 non-null  float64
dtypes: float64(1)
memory usage: 250.6 KB
```

Review data from FRED

In [76]:
```
dff_1980_2022=dff_1980_2022.to_frame().reset_index() # convert to dataframe.
dff_1980_2022=dff_1980_2022.rename(columns={0:'effective_rate', 'index':'date'}) # name columns as column names are not defined.
dff_1980_2022.head() # head method displays the first five rows of data.
```

Out[76]:

	date	effective_rate
0	1980-01-01	14.77
1	1980-01-02	14.00
2	1980-01-03	13.89
3	1980-01-04	14.00
4	1980-01-05	14.00

Figure 2: Import FRED data for interest rate

Labor flows to unemployment

In [118…
```
# Labor flows economic var.
labor_econ_var = 'LNS17400000'
```

In [119…
```
# Import FRED
from fredapi import Fred
```

In [120…
```
# FRED API Key
my_key = 'ea53013a18f3522101c4235350d71709'
fred = Fred(api_key=my_key)
```

In [121…
```
# Get labor flows data from FRED since 1980.
econ_labor=fred.get_series(labor_econ_var,observation_start='1980-01-01')
```

In [122…
```
# Convert to dataframe.
econ_labor=econ_labor.to_frame().reset_index()
```

In [123…
```
# Analyze data
econ_labor.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 405 entries, 0 to 404
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   index   405 non-null    datetime64[ns]
 1   0       405 non-null    float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 6.5 KB
```

Figure 3: Import FRED data for household income (labor flows from employment to unemployment)

Population

In [146…
```
# Population economic var.
pop_econ_var = 'LNU00000060'
```

In [147…
```
# Get data from 1980 to 2022 from FRED.
econ_pop = fred.get_series(pop_econ_var, observation_start='1980-01-01', observation_end='2022-12-31')

# Convert to dataframe
econ_pop=econ_pop.to_frame().reset_index()
```

In [148…
```
# Analyze info
econ_pop.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 516 entries, 0 to 515
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   index   516 non-null    datetime64[ns]
 1   0       516 non-null    float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 8.2 KB
```

Figure 4: Import FRED data for population

The primary key of data from WRDS and FRED individually

Primary key**:** The primary is evaluated on the merged dataset, and we check the uniqueness of data before predicting the primary key; Since, the dataset is large and there are panel data on companies, so we have to take the composite primary key on 'GvKey' and 'Datadate' columns

Primary key.

Evaluate the uniqueness of the composite primary key.

```
[180]: # True: gvkey and datadate can be composite primary key.
       # False: they are not.
       dff_wrds_1980_2022.set_index(['gvkey', 'datadate']).index.is_unique
```

```
[180]: True
```

```
[182]: # True: year and gvkey can be composite primary key.
       # False: they are not.
       dff_wrds_1980_2022.set_index(['year', 'gvkey']).index.is_unique
```

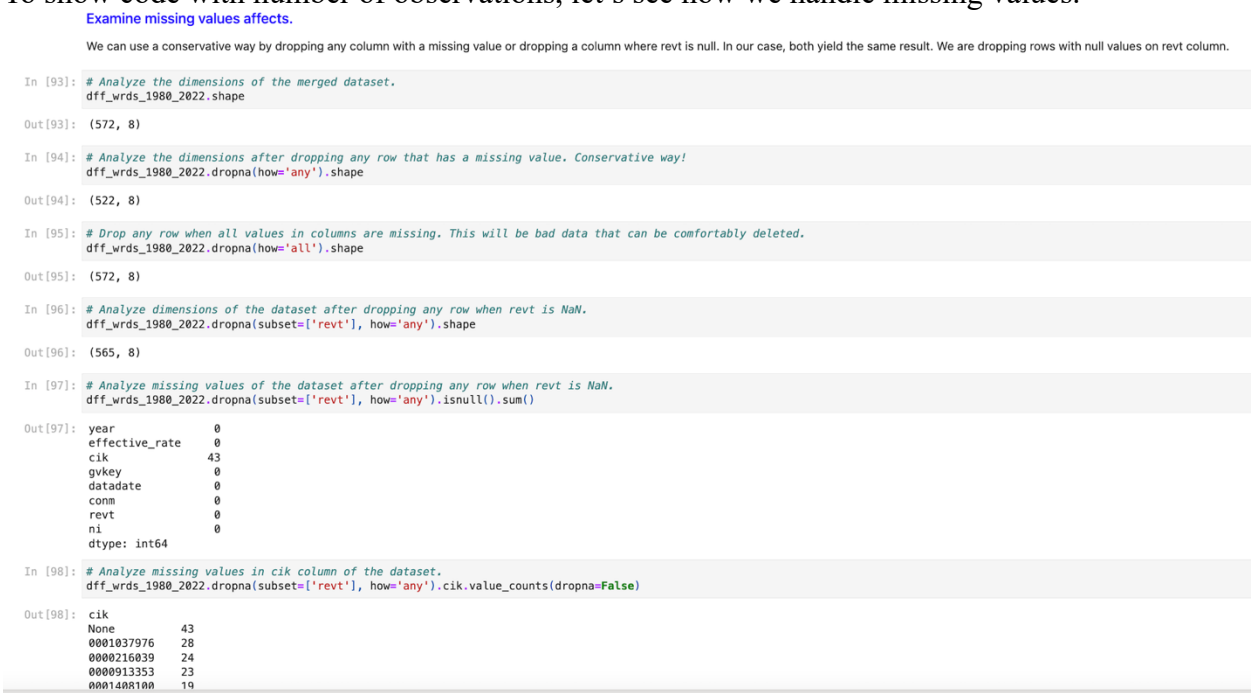
```
[182]: False
```

Primary Key

We have a composite primary key on: **GvKey, datadate**.

Figure 5: Primary key on merged wrds and FRED dataset.

Number of observations**:** 565: Since we take data from 1980 for merged datasets after handling missing values, we get 565 observations in the merged dataset after handling missing values. Our primary columns for research were – revt, net income, effective_rate, population, house_income. So, we removed any row that has any of the missing data for these columns.
To show code with number of observations, let's see how we handle missing values.



Examine missing values affects.

We can use a conservative way by dropping any column with a missing value or dropping a column where revt is null. In our case, both yield the same result. We are dropping rows with null values on revt column.

```
In [93]: # Analyze the dimensions of the merged dataset.
         dff_wrds_1980_2022.shape
```

```
Out[93]: (572, 8)
```

```
In [94]: # Analyze the dimensions after dropping any row that has a missing value. Conservative way!
         dff_wrds_1980_2022.dropna(how='any').shape
```

```
Out[94]: (522, 8)
```

```
In [95]: # Drop any row when all values in columns are missing. This will be bad data that can be comfortably deleted.
         dff_wrds_1980_2022.dropna(how='all').shape
```

```
Out[95]: (572, 8)
```

```
In [96]: # Analyze dimensions of the dataset after dropping any row when revt is NaN.
         dff_wrds_1980_2022.dropna(subset=['revt'], how='any').shape
```

```
Out[96]: (565, 8)
```

```
In [97]: # Analyze missing values of the dataset after dropping any row when revt is NaN.
         dff_wrds_1980_2022.dropna(subset=['revt'], how='any').isnull().sum()
```

```
Out[97]: year              0
         effective_rate    0
         cik              43
         gvkey             0
         datadate          0
         conm              0
         revt              0
         ni                0
         dtype: int64
```

```
In [98]: # Analyze missing values in cik column of the dataset.
         dff_wrds_1980_2022.dropna(subset=['revt'], how='any').cik.value_counts(dropna=False)
```

```
Out[98]: cik
         None          43
         0001037976    28
         0000216039    24
         0000913353    23
         0001408100    19
```

Figure 6: Examine missing values

In [98]:
```
# Analyze missing values in cik column of the dataset.
dff_wrds_1980_2022.dropna(subset=['revt'], how='any').cik.value_counts(dropna=False)
```

Out[98]:
```
cik
None          43
0001037976    28
0000216039    24
0000913353    23
0001408100    19
              ..
0000820906     1
0000017221     1
0000946733     1
0000890151     1
0001138118     1
Name: count, Length: 90, dtype: int64
```

In [99]:
```
# Verify the data that has no cik
dff_wrds_1980_2022.dropna(subset=['revt'], how='any')[dff_wrds_1980_2022.dropna(subset=['revt'], how='any').cik == None]
```

Out[99]:

	year	effective_rate	cik	gvkey	datadate	conm	revt	ni

Analysis on missing values affects.

We used multiple ways to analyze how we can effectively handle null values. We used the following approaches:

1. Dropping any column with a missing value: This is a conservative approach and may lead to the loss of valuable information if a large number of records are dropped. In our case, we ended up with 522 records.
2. Dropping records where **revt** is missing: This approach focuses on a specific key column (revt), which may be critical for our analysis. Since we ended up with 565 records. Now, we have 43 records where only cik has a null value.
3. Dropping records where all values are missing: This approach targets only entirely empty rows, which do not contribute any information to our analysis. In our case, no records were removed, indicating that there were no completely empty rows in the dataset.

Based on these observations, we decided to drop rows with null values in the revt column. It targets the main source of missing values without causing excessive data loss. Additionally, as the revt column seems to be a key column in our analysis, removing rows with missing revt values would ensure a more reliable and consistent dataset for further analysis. cik can be ignored if we have value in revt column.

In [100…
```
# Drop if the value in the revt column is missing as revt is the most essential column.
dff_wrds_1980_2022 = dff_wrds_1980_2022.dropna(subset=['revt'], how='any')

# Shape of DataFrame after dropping null value rows.
dff_wrds_1980_2022.shape
```

Out[100… (565, 8)

Figure 7: Analysis of missing values and handle missing values

Finally, we get to number of observations in the dataset after handling missing values.

Number of observations

[184]:
```
# Get the number of records and columns in the merged dataset.
dff_wrds_1980_2022.shape
```

[184]: (565, 10)

Number of observations = 565

Figure 8: Number of observations

Change in data frame**:** The data in WRDS is yearly panel data for companies based on fiscal period and FRED has irregular time series data (monthly, daily). The data are not compatible to establish a relation. So, we had to bring them on common ground by adding a year column to both datasets, group the irregular time series data in the FRED dataset by year and calculating the “mean.

Combined analysis of WRDS and FRED datasets based on initial review.

The data in WRDS is yearly panel data for companies based on fiscal period and FRED has irregular time series data (monthly, daily). The data are not compatible to establish a relation. So, we have to bring them on common ground by doing the following:

1. Add a **year** column to both datasets:
   - For the WRDS dataset, calculate the year from the **datadate** column.
   - For the FRED dataset, calculate the year from the **date** column.
2. Aggregate the irregular time series data in the FRED dataset:
   - Group the data by year and calculate the **mean** value of the effective_rate column.

Adding year column to WRDS dataset.

```
In [79]: # Add year column by using the DatetimeIndex method of Pandas, to get the year component of datadate column of the data frame.
         wrds_1980_2022['year'] = pd.DatetimeIndex(wrds_1980_2022['datadate']).year
         wrds_1980_2022.head() # Verify the year column is added.
```

Out[79]:

	cik	gvkey	datadate	conm	revt	ni	year
0	0000017221	002733	1997-07-31	CAPITAL INVESTMENT OF HAWAII	1.818	-0.847	1997
1	0000216039	005357	1987-12-31	GRUBB & ELLIS CO	336.989	0.250	1987
2	0000216039	005357	1988-12-31	GRUBB & ELLIS CO	370.838	-1.946	1988
3	0000216039	005357	1989-12-31	GRUBB & ELLIS CO	357.566	0.521	1989
4	0000216039	005357	1990-12-31	GRUBB & ELLIS CO	319.022	-29.751	1990

Adding year column to FRED dataset.

```
In [80]: dff_1980_2022['year'] = pd.DatetimeIndex(dff_1980_2022['date']).year # Add year column by getting year from date column.
         dff_1980_2022[dff_1980_2022['year']==1998].head(12) # Verify that data for the year 1998 appears in dataset.
```

Out[80]:

	date	effective_rate	year
6575	1998-01-01	5.84	1998
6576	1998-01-02	6.06	1998
6577	1998-01-03	6.06	1998

Figure 9: Data analysis before merging and adding year column.

Adding year column to FRED dataset.

```
In [80]: dff_1980_2022['year'] = pd.DatetimeIndex(dff_1980_2022['date']).year # Add year column by getting year from date column.
         dff_1980_2022[dff_1980_2022['year']==1998].head(12) # Verify that data for the year 1998 appears in dataset.
```

Out[80]:

	date	effective_rate	year
6575	1998-01-01	5.84	1998
6576	1998-01-02	6.06	1998
6577	1998-01-03	6.06	1998
6578	1998-01-04	6.06	1998
6579	1998-01-05	5.51	1998
6580	1998-01-06	5.35	1998
6581	1998-01-07	5.29	1998
6582	1998-01-08	5.42	1998
6583	1998-01-09	5.48	1998
6584	1998-01-10	5.48	1998
6585	1998-01-11	5.48	1998
6586	1998-01-12	5.48	1998

Aggregate FRED data yearly using mean method.

Create a new variable having an annual interest rate - `dff_1980_2022_ann`

```
In [81]: # Find the mean or average of the yearly interest rate.
         dff_1980_2022_ann = dff_1980_2022.groupby('year')['effective_rate'].mean()

         # Verify that data is accurate.
         dff_1980_2022_ann
```

Out[81]:

```
year
1980    13.349727
1981    16.386356
1982    12.237671
1983     9.090274
1984    10.225082
1985     8.099671
1986     6.799479
```

Figure 10: Handle data change in FRED

Merging Data

Relation between datasets**:** We have two datasets: one from WRDS with annual panel data for various companies, and the other with irregular (monthly, daily) time-series interest rate data from FRED. First, add a 'year' column to both datasets to serve as a common key. To align the time periods, we convert the irregular time-series dataset to a yearly one by calculating the mean value for each year. This results in a one-to-many relationship between the FRED and WRDS datasets.

Reason: The FRED data is consolidated into annual data, resulting in one record for each year's interest rate. In contrast, the annual data in the WRDS database contains multiple entries for each year due to the presence of various companies' data.

```
In [82]: # Convert series data to a data frame.
         dff_1980_2022_ann2 = dff_1980_2022_ann.to_frame().reset_index()

         # Verify the summary of data in the data frame.
         dff_1980_2022_ann2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44 entries, 0 to 43
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   year            44 non-null     int32
 1   effective_rate  44 non-null     float64
dtypes: float64(1), int32(1)
memory usage: 660.0 bytes
```

Figure 11: FRED data after aggregation - 44 records

The following snapshot shows the merged dataset, where we use inner join to merge FRED and WRDS datasets, the combined dataset results in 572 records. Since FRED dataset has only 44 records and inner joining with WRDS yields 572 records, it shows that we have one-to-many mapping between FRED and WRDS datasets.

Analysis before merging

It seems that merging records using the inner join yields less number of null values, so we are going ahead with inner join.

```
In [89]: # Merge data frames using inner join on the year column.
         dff_wrds_1980_2022 = pd.merge(
             dff_1980_2022_ann2, wrds_1980_2022, how="inner", on=["year"]
         )

         # Verify the summary of data after merging datasets.
         dff_wrds_1980_2022.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 572 entries, 0 to 571
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   year            572 non-null    int32
 1   effective_rate  572 non-null    float64
 2   cik             528 non-null    object
 3   gvkey           572 non-null    object
 4   datadate        572 non-null    datetime64[ns]
 5   conm            572 non-null    object
 6   revt            565 non-null    float64
 7   ni              565 non-null    float64
dtypes: datetime64[ns](1), float64(3), int32(1), object(3)
memory usage: 33.6+ KB
```

```
In [90]: # Tuple representing the dimension of the combined dataset.
         dff_wrds_1980_2022.shape
```

```
Out[90]: (572, 8)
```

Figure 12: Merged dataset showing 572 records - One-to-Many mapping.

How do you merge datasets? We first explored available join options to understand the effect on the merged dataset. We examined the following join options: inner join, left join, and outer join

using the common column between the datasets. We merged the datasets using the inner join on the "**year**" column.

Merge datasets

Merge wrds and FRED datasets from 1980 to 2022 data. Before merging, explore different join types by reviewing how they affect the final dataset. We are joining using the 'year' column.

In [83]:
```
# Review the shape of data after left join.
pd.merge(
    dff_1980_2022_ann2, wrds_1980_2022, how="left", on=["year"]
).shape
```

Out[83]: (579, 8)

In [84]:
```
# Verify null values
pd.merge(
    dff_1980_2022_ann2, wrds_1980_2022, how="left", on=["year"]
).isnull().sum()
```

Out[84]:
```
year               0
effective_rate     0
cik               51
gvkey              7
datadate           7
conm               7
revt              14
ni                14
dtype: int64
```

In [85]:
```
# Review the shape of data after outer join.
pd.merge(
    dff_1980_2022_ann2, wrds_1980_2022, how="outer", on=["year"]
).shape
```

Out[85]: (579, 8)

In [86]:
```
# Verify null values after outer join
pd.merge(
    dff_1980_2022_ann2, wrds_1980_2022, how="outer", on=["year"]
).isnull().sum()
```

Out[86]:
```
year               0
effective_rate     0
cik               51
gvkey              7
datadate           7
conm               7
revt              14
```

Figure 13: Validate joins before merging

In [87]:
```
# Review the shape of data after inner join.
pd.merge(
    dff_1980_2022_ann2, wrds_1980_2022, how="inner", on=["year"]
).shape
```

Out[87]:
```
(572, 8)
```

In [88]:
```
# Verify null values after inner join
pd.merge(
    dff_1980_2022_ann2, wrds_1980_2022, how="inner", on=["year"]
).isnull().sum()
```

Out[88]:
```
year               0
effective_rate     0
cik               44
gvkey              0
datadate           0
conm               0
revt               7
ni                 7
dtype: int64
```

Analysis before merging

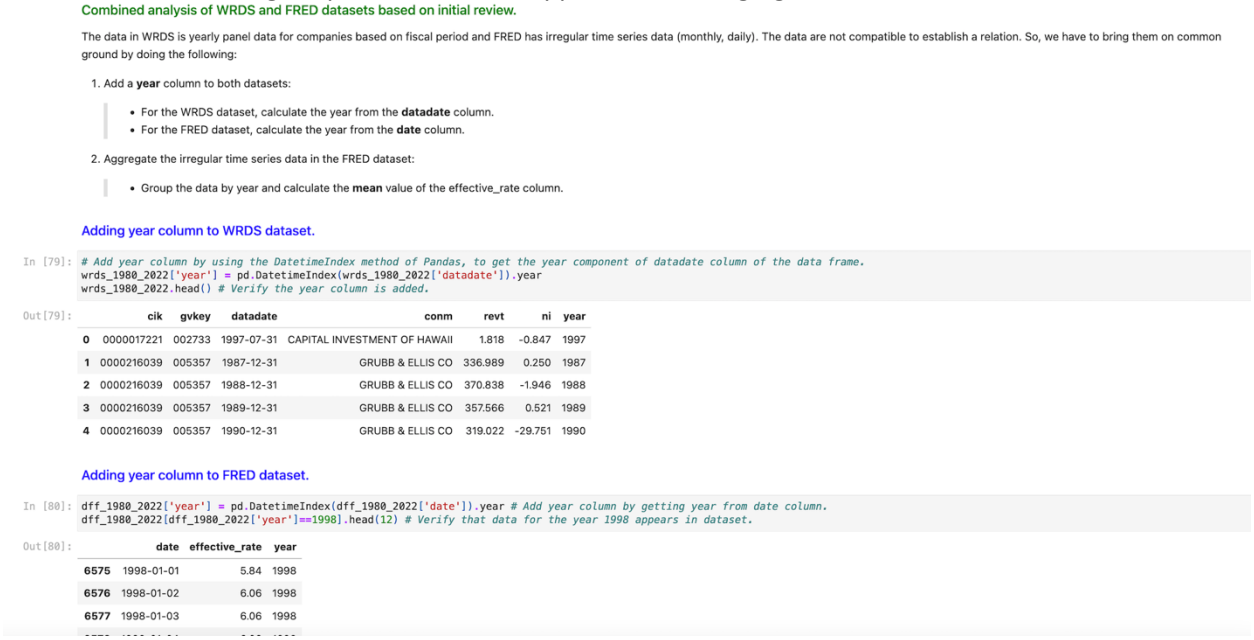
It seems that merging records using the inner join yields less number of null values, so we are going ahead with inner join.

In [89]:
```
# Merge data frames using inner join on the year column.
dff_wrds_1980_2022 = pd.merge(
    dff_1980_2022_ann2, wrds_1980_2022, how="inner", on=["year"]
)

# Verify the summary of data after merging datasets.
dff_wrds_1980_2022.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 572 entries, 0 to 571
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   year            572 non-null    int32
 1   effective_rate  572 non-null    float64
 2   cik             528 non-null    object
 3   gvkey           572 non-null    object
 4   datadate        572 non-null    datetime64[ns]
 5   conm            572 non-null    object
 6   revt            565 non-null    float64
```

Figure 14: Analysis before merging and merged datasets.

Reason for selection: The "**year**" column provides the common basis or matching column for both datasets. The following snapshot shows our approach for merging.



Combined analysis of WRDS and FRED datasets based on initial review.

The data in WRDS is yearly panel data for companies based on fiscal period and FRED has irregular time series data (monthly, daily). The data are not compatible to establish a relation. So, we have to bring them on common ground by doing the following:

1. Add a **year** column to both datasets:
   - For the WRDS dataset, calculate the year from the **datadate** column.
   - For the FRED dataset, calculate the year from the **date** column.
2. Aggregate the irregular time series data in the FRED dataset:
   - Group the data by year and calculate the **mean** value of the effective_rate column.

Adding year column to WRDS dataset.

```
In [79]: # Add year column by using the DatetimeIndex method of Pandas, to get the year component of datadate column of the data frame.
         wrds_1980_2022['year'] = pd.DatetimeIndex(wrds_1980_2022['datadate']).year
         wrds_1980_2022.head() # Verify the year column is added.
```

Out[79]:

	cik	gvkey	datadate	conm	revt	ni	year
0	0000017221	002733	1997-07-31	CAPITAL INVESTMENT OF HAWAII	1.818	-0.847	1997
1	0000216039	005357	1987-12-31	GRUBB & ELLIS CO	336.989	0.250	1987
2	0000216039	005357	1988-12-31	GRUBB & ELLIS CO	370.838	-1.946	1988
3	0000216039	005357	1989-12-31	GRUBB & ELLIS CO	357.566	0.521	1989
4	0000216039	005357	1990-12-31	GRUBB & ELLIS CO	319.022	-29.751	1990

Adding year column to FRED dataset.

```
In [80]: dff_1980_2022['year'] = pd.DatetimeIndex(dff_1980_2022['date']).year # Add year column by getting year from date column.
         dff_1980_2022[dff_1980_2022['year']==1998].head(12) # Verify that data for the year 1998 appears in dataset.
```

Out[80]:

	date	effective_rate	year
6575	1998-01-01	5.84	1998
6576	1998-01-02	6.06	1998
6577	1998-01-03	6.06	1998

Figure 15: Reason for year column selection for merging

Description of merged data

We evaluate the uniqueness of the merged dataset on the following parameters before concluding the primary key. First, verify if gvkey and datadate columns from the WRDS dataset help in uniquely identifying each row as it was a composite primary key for the WRDS dataset, we did not. Finally, we tested uniqueness on year and gvkey, which does not provide uniqueness.

Primary key.

Evaluate the uniqueness of the composite primary key.

```
[180]: # True: gvkey and datadate can be composite primary key.
       # False: they are not.
       dff_wrds_1980_2022.set_index(['gvkey', 'datadate']).index.is_unique
```

```
[180]: True
```

```
[182]: # True: year and gvkey can be composite primary key.
       # False: they are not.
       dff_wrds_1980_2022.set_index(['year', 'gvkey']).index.is_unique
```

```
[182]: False
```

Primary Key

We have a composite primary key on: **GvKey, datadate**.

Figure 16: Check uniqueness

Count of unique companies in merged dataset

```
[201]: dff_wrds_1980_2022.loc[:,'conm'].value_counts()
```

```
[201]: conm
JONES LANG LASALLE INC          28
GRUBB & ELLIS CO                24
COLLIERS INTL GROUP INC         23
KENNEDY-WILSON HOLDINGS INC     19
ANYWHERE REAL ESTATE INC        18
                                ..
WUNONG ASIA PACIFIC CO           1
BONAMOUR INC                     1
ASTRALIS LTD                     1
MEDIA TECHNOLOGIES INC           1
CBRE GROUP INC                   1
Name: count, Length: 99, dtype: int64
```

Figure 17: Number of unique companies in dataset: 99

Information about fields

```
[202]: dff_wrds_1980_2022.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 565 entries, 0 to 571
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   year            565 non-null    int32
 1   effective_rate  565 non-null    float64
 2   cik             522 non-null    object
 3   gvkey           565 non-null    object
 4   datadate        565 non-null    datetime64[ns]
 5   conm            565 non-null    object
 6   revt            565 non-null    float64
 7   ni              565 non-null    float64
 8   norm            565 non-null    float64
 9   log_revt        564 non-null    float64
dtypes: datetime64[ns](1), float64(5), int32(1), object(3)
memory usage: 62.5+ KB
```

Figure 18: Fields information

Data Analysis

Descriptive Statistics

The variables of interest in the combined dataset include revenues, net income, interest rate, population, household income, and company. These variables are represented by the following columns in our dataset: revt (revenues), ni (net income), effective_rate (interest rate), pop (population), house_income (household income), and comn (company).

Here is the snapshot of mean values of our dependent variable: revenue (we use logarithmic value of revenue as the distribution of data was better than revenue) and other independent variables: population, interest rate, household income, and net income.

File Edit View Run Kernel Settings Help

Code JupyterLab

Average values of dataset

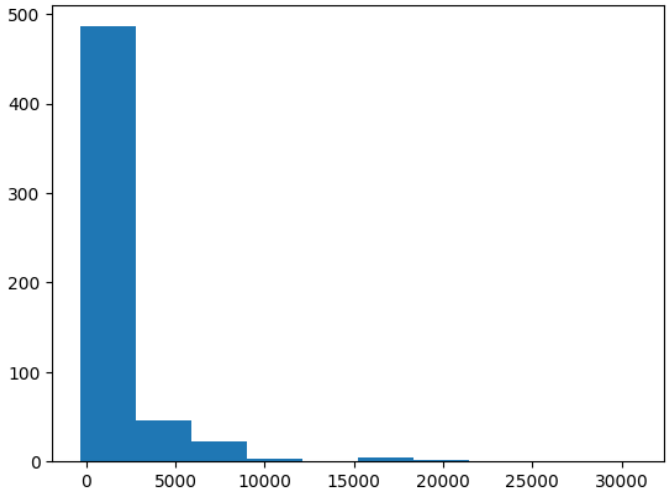
[187]:
```
# mean
mean_df = dff_labor_pop_wrds_1980_2022.loc[:,['log_revt', 'ni', 'effective_rate', 'pop','house_income']].mean()

print("Mean:\n", mean_df)
```

```
Mean:
 log_revt              4.570239
ni                    15.926841
effective_rate         2.561556
pop               121762.355535
house_income        1917.065510
dtype: float64
```

Here is snapshot of histogram using the dependent variable: revenue (revt)

Bakshi_Sitanshu_Project_Combined_Analysis-3.html

In [103…
```
# Understand the distribution of data among revt (dependent variable) using a Histogram
plt.hist(dff_wrds_1980_2022['revt'])
plt.show()
```



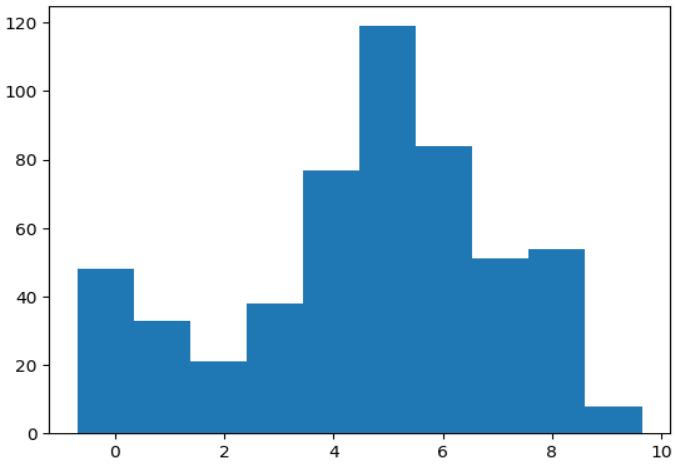
Result interpretation

Revenue is significantly positively skewed, so it is difficult to do regression.

We performed the frequency distribution of the data by using histograms throughout our analysis on log-transformed revenue (dependent variable).

In [155…

```
# Understand the distribution of data among log_revt (dependent variable) using a Histogram
plt.hist(dff_labor_pop_wrds_1980_2022['log_revt'])
plt.show()
```



Result interpretation

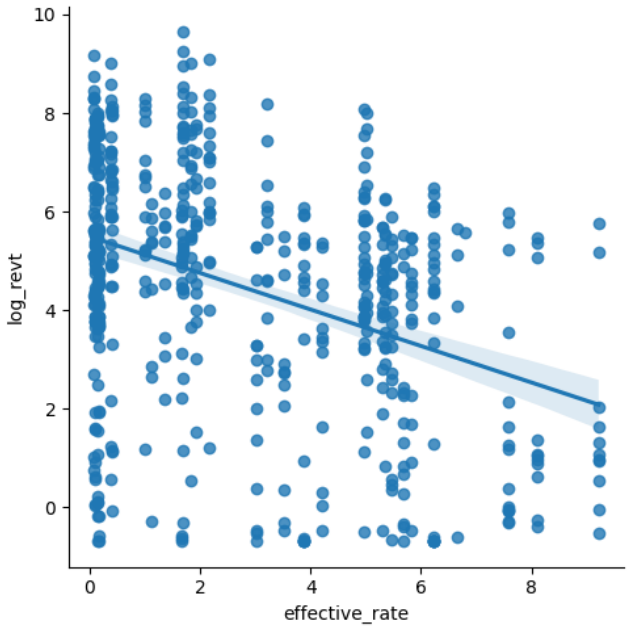
The log-transformed revenues are quite evenly distributed; however, it has some variations. Can be considered for regression analysis.

Trends of interested variables

We use scatter plots to represent the trend of variables.

Scatter Plot of revenue with effective_rate

```
In [136…  # Regression line in the scatter plot using revenue and effective_rate
          sns.lmplot(x='effective_rate',y='log_revt',data=dff_wrds_1980_2022,fit_reg=True)

Out[136…  <seaborn.axisgrid.FacetGrid at 0x129167f50>
```

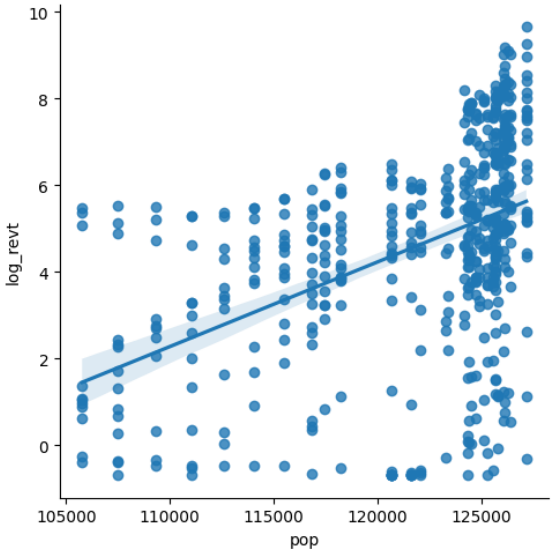


Plot interpretation

Increase in effective rate cause decline in company's revenue.

Scatter plot and regression line across revenue and population.

```
In [159…  # Regression line in the scatter plot using log-transformed revenue and population
          sns.lmplot(x='pop',y='log_revt',data=dff_labor_pop_wrds_1980_2022,fit_reg=True)

Out[159…  <seaborn.axisgrid.FacetGrid at 0x1293ba210>
```

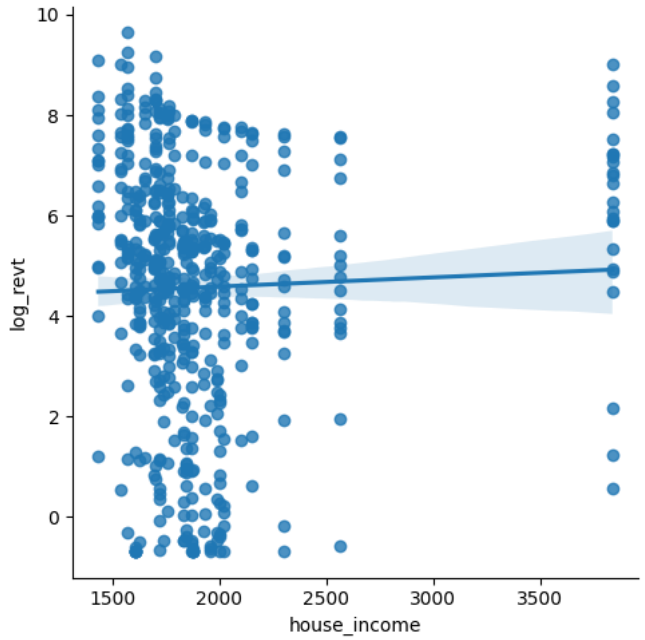


Graph interpretation

The increase in population increases the revenue of organizations in Real-estate industry.

Scatter plot on house income against the revenue.

```
In [157…  # Regression line in the scatter plot using revenue and house_income
          sns.lmplot(x='house_income',y='log_revt',data=dff_labor_wrds_1980_2022,fit_reg=True)

Out[157…  <seaborn.axisgrid.FacetGrid at 0x1293086d0>
```



Correlation

The following correlation analysis suggests that effective_rate is negatively correlated with the revenue. However, all the correlations are weak.

Bakshi_Sitanshu_Project_Combined_Analysis-3.html

Correlation

In [109…

```
# Correlation metrics.
dff_wrds_1980_2022_2 = dff_wrds_1980_2022[['log_revt', 'effective_rate','ni','year']]
dff_wrds_1980_2022_cor = dff_wrds_1980_2022_2.corr()
dff_wrds_1980_2022_cor
```

Out[109…

	log_revt	effective_rate	ni	year
log_revt	1.000000	-0.378722	0.106421	0.504652
effective_rate	-0.378722	1.000000	-0.035560	-0.817935
ni	0.106421	-0.035560	1.000000	0.041027
year	0.504652	-0.817935	0.041027	1.000000

Result interpretation on correlation

The above correlation metrics show a strong correlation between log_revt and year. However, log_revt and ni show a weak relationship.

In [110…

```
# Correlation metrics.
dff_wrds_1980_2022_3 = dff_wrds_1980_2022[['revt','effective_rate', 'ni','year']]
dff_wrds_1980_2022_3_cor = dff_wrds_1980_2022_3.corr()
dff_wrds_1980_2022_3_cor
```

Out[110…

	revt	effective_rate	ni	year
revt	1.000000	-0.250996	0.248194	0.399956
effective_rate	-0.250996	1.000000	-0.035560	-0.817935
ni	0.248194	-0.035560	1.000000	0.041027
year	0.399956	-0.817935	0.041027	1.000000

Result interpretation on correlation

The above correlation metrics show a weak correlation between all variables.

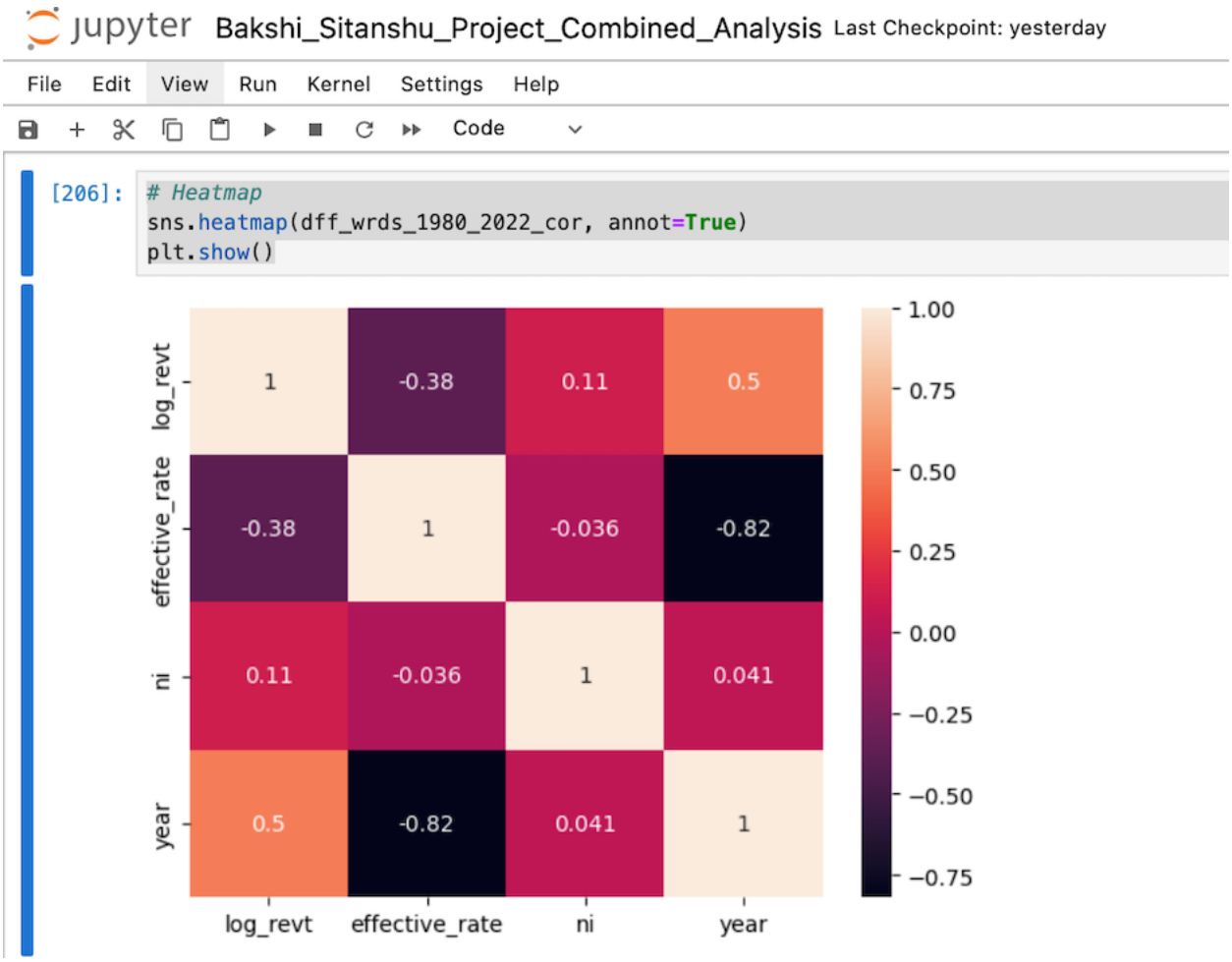


Figure 19: Heat map of effective rate

The correlation between the revenue (log-transformed) and house income is not strong.

Correlation

In [133…

```
# Correlation metrics.
dff_labor_wrds_1980_2022_n = dff_labor_wrds_1980_2022[['log_revt','house_income','effective_rate','ni','year']]
dff_labor_wrds_1980_2022_cor = dff_labor_wrds_1980_2022_n.corr()
dff_labor_wrds_1980_2022_cor
```

Out[133…

	log_revt	house_income	effective_rate	ni	year
log_revt	1.000000	0.035113	-0.321925	0.106501	0.477305
house_income	0.035113	1.000000	-0.325540	-0.066220	0.160605
effective_rate	-0.321925	-0.325540	1.000000	-0.022331	-0.780524
ni	0.106501	-0.066220	-0.022331	1.000000	0.038902
year	0.477305	0.160605	-0.780524	0.038902	1.000000

Result interpretation on correlation

The above correlation metrics suggest a positive correlation between log_revt and house_income. This is not a strong correlation.

The negative correlation between log_revt and effective_rate. Again, this is not a strong correlation. However, log_revt and ni show a weak relationship.

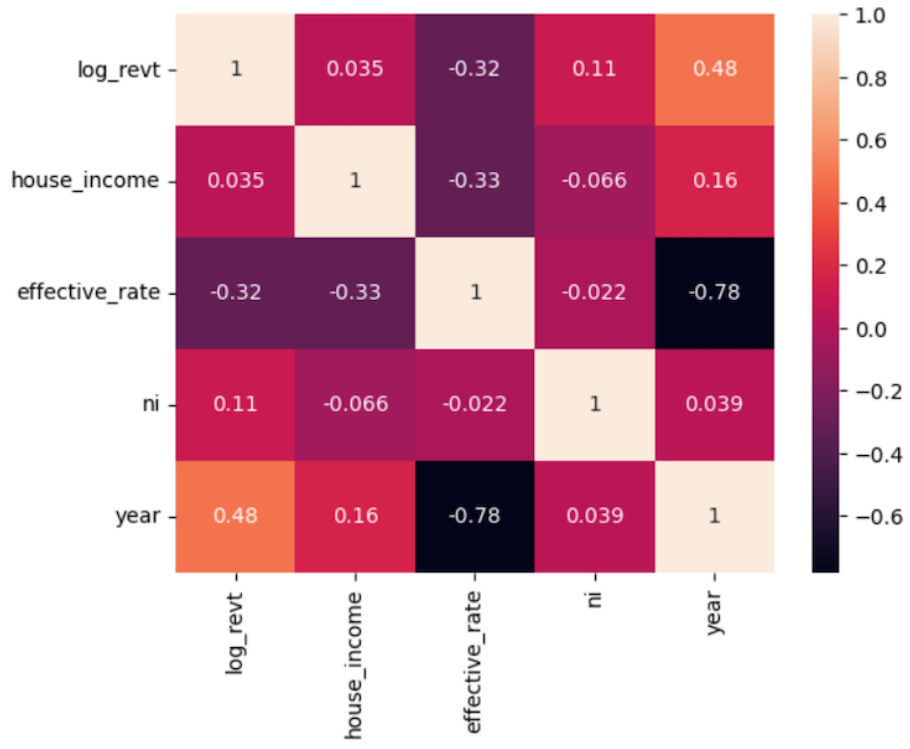
jupyter Bakshi_Sitanshu_Project_Combined_Analysis Last Checkpoint: yesterday

File Edit View Run Kernel Settings Help

Code

[208]:

```
# Heatmap
sns.heatmap(dff_labor_wrds_1980_2022_cor, annot=True)
plt.show()
```



The correlation between revenue (log-transformed) and population is also not strong.

In [160…

```
# Correlation metrics.
dff_labor_pop_wrds_1980_2022_n = dff_labor_pop_wrds_1980_2022[['log_revt','pop','ni','year']]
dff_labor_pop_wrds_1980_2022_cor = dff_labor_pop_wrds_1980_2022_n.corr()
dff_labor_pop_wrds_1980_2022_cor
```

Out[160…

	log_revt	pop	ni	year
log_revt	1.000000	0.447405	0.106501	0.477305
pop	0.447405	1.000000	0.025262	0.885001
ni	0.106501	0.025262	1.000000	0.038902
year	0.477305	0.885001	0.038902	1.000000

Result interpretation on correlation

The above correlation metrics suggest a positive correlation between log_revt and population. This is not a strong correlation.

However, log_revt and ni show a weak relationship.

Figure 20: Correlation metrics on population

jupyter Bakshi_Sitanshu_Project_Combined_Analysis Last Checkpoint: yesterday

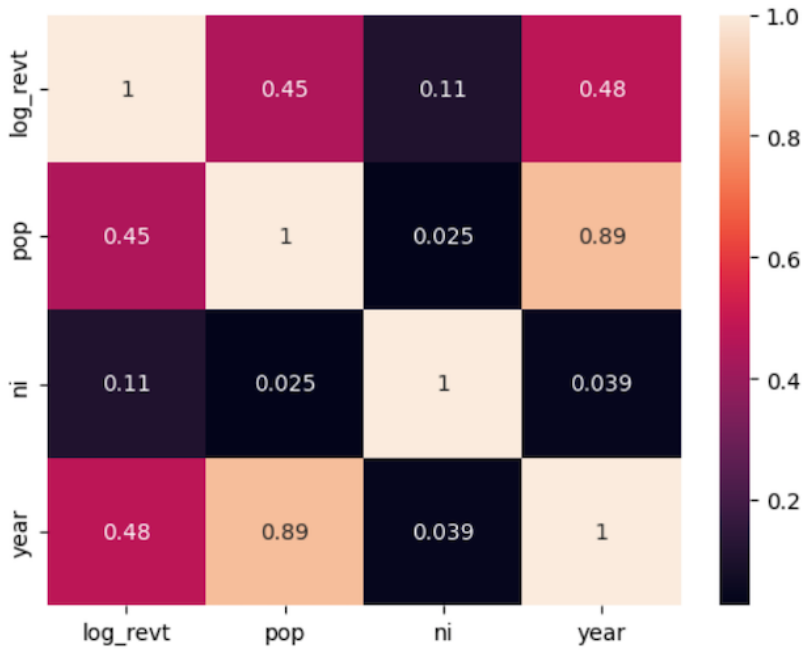
File Edit View Run Kernel Settings Help

Code

However, log_revt and ni show a weak relationship.

[209]:

```
# Heatmap
sns.heatmap(dff_labor_pop_wrds_1980_2022_cor, annot=True)
plt.show()
```



Regression Analysis

We conducted more than 35 regression analyses, incorporating a broad range of combinations including individual economic indicators (interest rate, population, labor), combinations of these indicators, subsets of data on specific companies, and distinct time periods such as the 1990s, 2000s, 2010s, pre-Covid-19, post-Covid-19, and the effects of US elections in the previous decade. Below, we present a selection of notable regression results and provide an interpretation of their implications.

The following important terms helps in reading and understanding our following regression analysis:

- R-squared: This is the statistical metric which qualifies the proportion of variation in the dependent variable that can be accounted for by the independent variable.
- Prob(F-Statistic): This conveys the overall significance of regression. The Prob(F-Statistic) evaluates the significance of all variables collectively. The Prob(F-Statistic) represents the likelihood of "all coefficients close to zero". The probability is close to zero, indicating that overall regression is meaningful.
- Coefficient: It is the value that multiplies the independent variable in a linear regression equation. The equation combines the product of the coefficient and the independent variable with a constant term to estimate the value of the dependent variable.
- P>|t| (p-value): It is a statistical measure used in hypothesis testing to determine the probability of observing results as extreme as the ones obtained, assuming that the null hypothesis is true. In the context of linear regression, the p-value helps assess the significance of the relationship between the independent and dependent variables.

Reference: Professor Yoon's in-class10.shared.ipynb file

Following are the regressions highlighting the impact of interest rate (effective_rate) on revenue of organizations in real estate sector.

Regression 6: Effect on log-transformed revenue on effective rate and net income

In [112…

```
# Regression
model = smf.ols(formula='log_revt ~ effective_rate+ni', data=dff_wrds_1980_2022)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_revt   R-squared:                       0.153
Model:                            OLS   Adj. R-squared:                  0.150
Method:                 Least Squares   F-statistic:                     50.52
Date:                Sat, 02 Dec 2023   Prob (F-statistic):           6.71e-21
Time:                        10:55:25   Log-Likelihood:                -1275.7
No. Observations:                 564   AIC:                             2557.
Df Residuals:                     561   BIC:                             2570.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.4699      0.147     37.255      0.000       5.182       5.758
effective_rate -0.3679      0.038     -9.672      0.000      -0.443      -0.293
ni             0.0011      0.000      2.466      0.014       0.000       0.002
==============================================================================
Omnibus:                       28.755   Durbin-Watson:                   1.655
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               31.750
Skew:                          -0.568   Prob(JB):                     1.28e-07
Kurtosis:                       2.750   Cond. No.                         326.
==============================================================================

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

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 15.3% - i.e. 15.3% of revenue is explained by effective rate and net income.

The coefficient on effective_rate is -0.3679, which means effective_Rate is negatively associated with retailers' revenues. In other words, the increase in effective_rate causes a decrease in revenue. **This coefficient is statistically significant at p<0.01 level**.

For net income (ni), the coefficient is 0.0011 means that ni is also positively associated with retailers' revenues. The p-value of 0.014 indicates that the **relationship between 'ni' and log_revt(dependent variable) is also statistically significant**.

Figure 21: Regression on effect of effective_rate and net income on revenue.

In the given regression analysis, the impact on log-transformed revenue (dependent variable) is assessed in relation to the independent variable: interest rate (effective_rate). Since the F-stat test score is below 0.1, the regression analysis is considered valid. The r-squared stands at 0.153, signifying that 15.3% of the revenue can be explained by the effective_rate and net income (ni). The effective_rate has a coefficient of -0.3679, indicating a negative association with the company's revenue. Furthermore, this coefficient is statistically significant at a p<0.01 level.

Regression 7: Revenue in 1990's based on effective_rate

In [115…

```
model2 = smf.ols(formula='log_revt ~ effective_rate+y_1990+y_1991+y_1992+y_1993+y_1994+y_1995+y_1996+y_1997+y_1998+y_1999', data=dff_wrds_1980_2022_cv)
results2 = model2.fit()
print(results2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_revt   R-squared:                       0.191
Model:                            OLS   Adj. R-squared:                  0.175
Method:                 Least Squares   F-statistic:                     11.88
Date:                Sat, 02 Dec 2023   Prob (F-statistic):           4.12e-20
Time:                        10:57:17   Log-Likelihood:                -1262.5
No. Observations:                 564   AIC:                             2549.
Df Residuals:                     552   BIC:                             2601.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.5462      0.147     37.833      0.000       5.258       5.834
y_1990[T.True]    -0.9305      0.753     -1.236      0.217      -2.409       0.548
y_1991[T.True]    -1.8053      0.645     -2.801      0.005      -3.071      -0.539
y_1992[T.True]    -1.8416      0.704     -2.617      0.009      -3.224      -0.459
y_1993[T.True]    -2.2835      0.647     -3.527      0.000      -3.555      -1.012
y_1994[T.True]    -1.2695      0.679     -1.871      0.062      -2.602       0.063
y_1995[T.True]    -0.0158      0.646     -0.024      0.981      -1.285       1.254
y_1996[T.True]     0.0778      0.603      0.129      0.897      -1.106       1.261
y_1997[T.True]    -0.4948      0.573     -0.863      0.389      -1.621       0.631
y_1998[T.True]     0.5665      0.603      0.939      0.348      -0.618       1.751
y_1999[T.True]     0.4314      0.554      0.779      0.436      -0.657       1.520
effective_rate    -0.3320      0.046     -7.279      0.000      -0.422      -0.242
==============================================================================
Omnibus:                       31.625   Durbin-Watson:                   1.746
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.036
Skew:                          -0.616   Prob(JB):                     1.50e-08
Kurtosis:                       2.877   Cond. No.                         35.1
==============================================================================

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

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.191 - i.e. 19.1% of revenue is explained by independent variables.

For effective rate, the coefficient for the effective rate is negative i.e. -0.3320 means that an increase in effective_rate causes a decrease in retailers' revenues. The p-value of 0.000 indicates that the relationship between '**effective_rate' and log_revt(dependent variable) is statistically significant**. The analysis of years indicates that years 1991, 1992, and 1993 have a significant negative relationship with revenue, while other years do not show a statistically significant relationship.

In the regression under consideration, we incorporate control variables for the years 1980 to 2022. The analysis evaluates the impact of the interest rate (effective_rate) as an independent variable on log-transformed revenue (dependent variable) during the 1990s, considering control variables from 1990 to 1999. With an F-stat test score below 0.1, the regression analysis is deemed reliable. The r-squared is 0.191, suggesting that the effective_rate accounts for 19.1% of the revenue in the 1990s. The coefficient for the effective_rate is -0.3320, signifying a negative relationship with the company's revenue. Additionally, the coefficient is statistically significant at a $p<0.01$ level. The analysis of years indicates that the years 1991, 1992, and 1993 have a significant negative relationship with revenue, while other years do not show a statistically significant relationship.

Regression 8: Revenue of companies in 2000's based on effective_rate.

In [116…

```
model3 = smf.ols(formula='log_revt ~ effective_rate+y_2000+y_2001+y_2002+y_2003+y_2004+y_2005+y_2006+y_2007+y_2008+y_2009', data=dff_wrds_1980_2022_cv)
results3 = model3.fit()
print(results3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_revt   R-squared:                       0.182
Model:                          OLS   Adj. R-squared:                  0.166
Method:               Least Squares   F-statistic:                     11.20
Date:              Sat, 02 Dec 2023   Prob (F-statistic):           7.09e-19
Time:                      10:59:24   Log-Likelihood:                -1265.6
No. Observations:                564   AIC:                             2555.
Df Residuals:                    552   BIC:                             2607.
Df Model:                         11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.6048      0.159     35.337      0.000       5.293       5.916
y_2000[T.True]     0.2247      0.505      0.445      0.657      -0.767       1.217
y_2001[T.True]    -1.0659      0.557     -1.913      0.056      -2.160       0.029
y_2002[T.True]    -1.2948      0.628     -2.061      0.040      -2.529      -0.061
y_2003[T.True]    -0.9164      0.780     -1.175      0.240      -2.448       0.615
y_2004[T.True]    -0.3373      0.779     -0.433      0.665      -1.868       1.193
y_2005[T.True]     1.0839      0.676      1.605      0.109      -0.243       2.411
y_2006[T.True]     1.7816      0.681      2.614      0.009       0.443       3.120
y_2007[T.True]     1.4566      0.656      2.219      0.027       0.168       2.746
y_2008[T.True]     0.4629      0.607      0.763      0.446      -0.729       1.655
y_2009[T.True]    -0.8119      0.615     -1.320      0.187      -2.020       0.396
effective_rate    -0.4121      0.041    -10.024      0.000      -0.493      -0.331
==============================================================================
Omnibus:                       31.793   Durbin-Watson:                   1.728
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               33.499
Skew:                          -0.565   Prob(JB):                     5.32e-08
Kurtosis:                       2.615   Cond. No.                         32.3
==============================================================================

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

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.182 - i.e. 18.2% of revenue is explained by independent variables.

For effective rate, the coefficient for the effective rate is negative i.e. -0.4121 means that an increase in effective_rate causes a decrease in retailers' revenues. The p-value of 0.000 indicates that the relationship between **'effective_rate' and log_revt(dependent variable) is statistically significant**.

The analysis of years indicates that years 2001 and 2002 have a significant negative relationship with revenue, and 2006 and 2007 have a significant positive relation with revenue, while other years do not show a statistically significant relationship.

In Regression 8, the analysis examines the influence of the interest rate (effective_rate) as an independent variable on the log-transformed revenue (dependent variable) throughout the 2000s, considering control variables from 2000 to 2009. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.182, which implies that the effective_rate explains 18.2% of the revenue during the 2000s. The effective_rate has a coefficient of -0.4121, indicating a negative association with the company's revenue. Moreover, this coefficient is statistically significant at a $p<0.01$ level. The year-wise analysis reveals that 2001 and 2002 have a significant negative relationship with revenue, while 2006 and 2007 exhibit a significant positive connection. Other years, however, do not demonstrate a statistically significant relationship.

Regression 9: Revenue of company's in 2010's based on effective rate

In [117…

```
model4 = smf.ols(formula='log_revt ~ effective_rate+y_2010+y_2011+y_2012+y_2013+y_2014+y_2015+y_2016+y_2017+y_2018+y_2019', data=dff_wrds_1980_2022_cv)
results4 = model4.fit()
print(results4.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                log_revt   R-squared:                       0.194
Model:                            OLS   Adj. R-squared:                  0.178
Method:                 Least Squares   F-statistic:                     12.08
Date:                Sat, 02 Dec 2023   Prob (F-statistic):           1.77e-20
Time:                        11:00:38   Log-Likelihood:                -1261.6
No. Observations:                 564   AIC:                             2547.
Df Residuals:                     552   BIC:                             2599.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.7495      0.212     27.182      0.000       5.334       6.165
y_2010[T.True]    -1.2963      0.592     -2.190      0.029      -2.459      -0.134
y_2011[T.True]    -0.9886      0.578     -1.709      0.088      -2.125       0.147
y_2012[T.True]    -1.5863      0.552     -2.874      0.004      -2.671      -0.502
y_2013[T.True]    -0.8949      0.593     -1.509      0.132      -2.060       0.270
y_2014[T.True]    -0.6941      0.646     -1.074      0.283      -1.964       0.576
y_2015[T.True]    -0.3539      0.609     -0.581      0.561      -1.549       0.842
y_2016[T.True]    -0.2831      0.589     -0.481      0.631      -1.440       0.874
y_2017[T.True]     0.7415      0.617      1.203      0.230      -0.469       1.952
y_2018[T.True]     1.1948      0.575      2.079      0.038       0.066       2.324
y_2019[T.True]     1.5696      0.543      2.888      0.004       0.502       2.637
effective_rate    -0.4314      0.046     -9.306      0.000      -0.522      -0.340
==============================================================================
Omnibus:                       29.607   Durbin-Watson:                   1.751
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               32.819
Skew:                          -0.577   Prob(JB):                     7.47e-08
Kurtosis:                       2.749   Cond. No.                         34.1
==============================================================================

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

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 19.4% - i.e. 19.4% of revenue is explained by independent variables.

For effective rate, the coefficient for the effective rate is negative i.e. -0.4314 means that an increase in effective_rate causes a decrease in retailers' revenues. The p-value of 0.000 indicates that the relationship between '**effective_rate' and log_revt(dependent variable) is statistically significant**.

The analysis of years indicates that years 2010 and 2012 have a significant negative relationship with revenue, and 2018 and 2019 have a significant positive relation with revenue, while other years do not show a statistically significant relationship.

In Regression 9, the analysis examines the influence of the interest rate (effective_rate) as an independent variable on the log-transformed revenue (dependent variable) throughout the 2010s, considering control variables from 2010 to 2019. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.194, which implies that the effective_rate explains 19.4% of the revenue during the 2010s. The effective_rate has a coefficient of -0.4314, indicating a negative association with the company's revenue. Moreover, this coefficient (effective_rate) is statistically significant at a $p<0.01$ level. The year-wise analysis reveals that 2010 and 2012 have a significant negative relationship with revenue, while 2018 and 2019 exhibit a significant positive connection. Other years, however, do not demonstrate a statistically significant relationship.

Interpretation of the Overall Results for the Impact of Effective Rate on Revenue Across Decades (1990s, 2000s, 2010s)

The regression analyses reveal that the effective rate plays a crucial role in influencing the revenue of companies within the Real Estate sector. The relationship between the effective rate and revenue is negative, indicating that an increase in the effective rate leads to a decrease in the organizations' revenue across the examined decades.

Overall Results Interpretation of effective rate on revenue for decade results (1990's, 2000's, 2010's)

From the above regressions, it is evident that effective rate has a significant impact on the revenue of the organizations in the Real Estate, where an increase in effective_rate decreases the revenue of organizations.

The following regressions showcase the impact on revenue (dependent variable) on household income.

Regression 10: Revenue in 1990's based on house income

```
In [139…
model5 = smf.ols(formula='log_revt ~ house_income+y_1990+y_1991+y_1992+y_1993+y_1994+y_1995+y_1996+y_1997+y_1998+y_1999', data=dff_labor_wrds_1980_2022)
results5 = model5.fit()
print(results5.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_revt   R-squared:                       0.149
Model:                            OLS   Adj. R-squared:                  0.131
Method:                 Least Squares   F-statistic:                     8.298
Date:                Sat, 02 Dec 2023   Prob (F-statistic):           1.71e-13
Time:                        11:26:45   Log-Likelihood:                -1195.2
No. Observations:                 533   AIC:                             2414.
Df Residuals:                     521   BIC:                             2466.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.0284      0.438     11.489      0.000       4.169       5.888
y_1990[T.True]    -3.1468      0.705     -4.464      0.000      -4.532      -1.762
y_1991[T.True]    -3.2247      0.627     -5.144      0.000      -4.456      -1.993
y_1992[T.True]    -2.5422      0.705     -3.608      0.000      -3.926      -1.158
y_1993[T.True]    -2.8153      0.650     -4.331      0.000      -4.092      -1.538
y_1994[T.True]    -2.1938      0.676     -3.246      0.001      -3.522      -0.866
y_1995[T.True]    -1.4806      0.627     -2.360      0.019      -2.713      -0.248
y_1996[T.True]    -1.2075      0.590     -2.048      0.041      -2.366      -0.049
y_1997[T.True]    -1.8330      0.558     -3.285      0.001      -2.929      -0.737
y_1998[T.True]    -0.7345      0.591     -1.244      0.214      -1.895       0.426
y_1999[T.True]    -0.7421      0.546     -1.359      0.175      -1.815       0.331
house_income    2.481e-05      0.000      0.115      0.908      -0.000       0.000
==============================================================================
Omnibus:                       40.465   Durbin-Watson:                   1.735
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               47.975
Skew:                          -0.730   Prob(JB):                     3.82e-11
Kurtosis:                       3.165   Cond. No.                     1.49e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.49e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.149 - i.e. 14.9% of revenue is explained by independent variables.

For house income, the coefficient for the house income is so small. The p-value of 0.908 indicates that the relationship between '**house_income' and log_revt(dependent variable) is not statistically significant**.

In Regression 10, the analysis explores the impact of household income as an independent variable on the log-transformed revenue (dependent variable) during the 1990s, taking into account control variables from 1990 to 1999. The F-stat test score falls below 0.1, signifying that the regression analysis is deemed reliable. The r-squared value is 0.149, suggesting that household income accounts for 14.9% of the revenue in the 1990s. The coefficient for household income is 0.00002481, which is quite small. Additionally, this coefficient (household income) is not statistically significant with a value of 0.908. The analysis of individual years indicates that the majority of years exhibit a significant negative correlation with revenue.

Regression 13: Revenue in 2000's based on house income

In [143…

```
model7 = smf.ols(formula='log_revt ~ house_income+y_2000+y_2001+y_2002+y_2003+y_2004+y_2005+y_2006+y_2007+y_2008+y_2009', data=dff_labor_wrds_1980_2022)
results7 = model7.fit()
print(results7.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                log_revt   R-squared:                       0.041
Model:                            OLS   Adj. R-squared:                  0.021
Method:                 Least Squares   F-statistic:                     2.016
Date:                Sat, 02 Dec 2023   Prob (F-statistic):             0.0251
Time:                        11:28:48   Log-Likelihood:                -1227.1
No. Observations:                 533   AIC:                             2478.
Df Residuals:                     521   BIC:                             2530.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept          4.4653      0.465      9.595      0.000       3.551       5.380
y_2000[T.True]    -1.3774      0.519     -2.651      0.008      -2.398      -0.357
y_2001[T.True]    -1.7238      0.590     -2.922      0.004      -2.882      -0.565
y_2002[T.True]    -1.0498      0.665     -1.577      0.115      -2.357       0.258
y_2003[T.True]    -0.4456      0.825     -0.540      0.589      -2.066       1.175
y_2004[T.True]     0.0518      0.825      0.063      0.950      -1.569       1.673
y_2005[T.True]     0.7101      0.718      0.989      0.323      -0.700       2.120
y_2006[T.True]     0.6875      0.718      0.958      0.339      -0.723       2.098
y_2007[T.True]     0.3386      0.690      0.490      0.624      -1.018       1.695
y_2008[T.True]     0.5862      0.645      0.909      0.364      -0.681       1.853
y_2009[T.True]    -0.0097      0.661     -0.015      0.988      -1.308       1.289
house_income       0.0001      0.000      0.454      0.650      -0.000       0.001
==============================================================================
Omnibus:                       32.381   Durbin-Watson:                   1.545
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.270
Skew:                          -0.458   Prob(JB):                     1.97e-06
Kurtosis:                       2.412   Cond. No.                     1.58e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.58e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.041 - i.e. 4.1% of revenue is explained by independent variables.

For house income, the coefficient for the house income is 0.0001. The p-value of 0.650 indicates that the relationship between '**log_house_income' and log_revt(dependent variable) is not statistically significant**. However, on years, **2000 and 2001 are statistically significant** to revenue.

In Regression 13, the analysis explores the impact of household income as an independent variable on the log-transformed revenue (dependent variable) during the 2000s, taking into account control variables from 2000 to 2009. The F-stat test score falls below 0.1, signifying that the regression analysis is deemed reliable. The r-squared value is 0.041, suggesting that household income accounts for 4.1% of the revenue in the 2000s. The coefficient for household income is 0.0001, which is small. Additionally, this coefficient (household income) is not statistically significant with a value of 0.650. The year-wise analysis reveals that 2000 and 2001 have a significant negative relationship with revenue, while other years do not demonstrate a statistically significant relationship.

Interpretation of the Overall Results for the Impact of Household Income on Revenue Across Decades (1990s, 2000s, 2010s)

Based on the regression analyses, it is clear that household income does not substantially influence the revenue generated by companies in the Real Estate sector throughout the examined decades.

Overall Results Interpretation of house income on revenue for decade results (1990's, 2000's, 2010's)

From the above regressions, it is evident that house income does not have a significant impact on the revenue of the organizations in the Real Estate.

The following regression analysis shows the impact of population (independent variable) on the revenue of businesses in the real estate industry.

Regression 16: Revenue during 1990's based on population.

In [161…

```
model9 = smf.ols(formula='log_revt ~ pop+y_1990+y_1991+y_1992+y_1993+y_1994+y_1995+y_1996+y_1997+y_1998+y_1999', data=dff_labor_pop_wrds_1980_2022)
results9 = model9.fit()
print(results9.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                log_revt   R-squared:                       0.250
Model:                            OLS   Adj. R-squared:                  0.234
Method:                 Least Squares   F-statistic:                     15.80
Date:                Sat, 02 Dec 2023   Prob (F-statistic):           8.43e-27
Time:                        11:50:45   Log-Likelihood:                -1161.5
No. Observations:                 533   AIC:                             2347.
Df Residuals:                     521   BIC:                             2398.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -62.1441      8.024     -7.745      0.000     -77.908     -46.381
y_1990[T.True]     7.1124      1.392      5.109      0.000       4.378       9.847
y_1991[T.True]     6.1034      1.259      4.847      0.000       3.630       8.577
y_1992[T.True]     5.8040      1.196      4.854      0.000       3.455       8.153
y_1993[T.True]     4.6032      1.075      4.280      0.000       2.491       6.716
y_1994[T.True]     4.3813      1.009      4.341      0.000       2.399       6.364
y_1995[T.True]     4.3362      0.910      4.763      0.000       2.548       6.125
y_1996[T.True]     3.8100      0.815      4.676      0.000       2.209       5.411
y_1997[T.True]     2.4697      0.732      3.372      0.001       1.031       3.909
y_1998[T.True]     3.2354      0.728      4.445      0.000       1.806       4.665
y_1999[T.True]     2.8229      0.663      4.255      0.000       1.519       4.126
pop            0.0005   6.43e-05      8.378      0.000       0.000       0.001
==============================================================================
Omnibus:                       35.819   Durbin-Watson:                   1.965
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.998
Skew:                          -0.687   Prob(JB):                     7.59e-10
Kurtosis:                       2.924   Cond. No.                     1.10e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.1e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.250 - i.e. 25% of revenue is explained by independent variables.

For population, the coefficient for the population is 0.0005. The p-value of 0.000 indicates that the relationship between **'pop' and log_revt(dependent variable) is statistically significant** and so is the relationship of revenue with years based on population.

In Regression 16, the analysis examines the influence of the population as an independent variable on the log-transformed revenue (dependent variable) throughout the 1990s, considering control variables from 1990 to 1999. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.250, which implies that the population explains 25% of the revenue during the 1990s. The population has a coefficient of 0.0005, indicating a positive association with the company's revenue. Moreover, this coefficient (pop) is statistically significant at a $p<0.01$ level. The year-wise analysis reveals that all years have a significant positive relationship with revenue.

Regression 17: Revenue during 2000's based on population.

In [162…

```
model10 = smf.ols(formula='log_revt ~ pop+y_2000+y_2001+y_2002+y_2003+y_2004+y_2005+y_2006+y_2007+y_2008+y_2009', data=dff_labor_pop_wrds_1980_2022)
results10 = model10.fit()
print(results10.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_revt   R-squared:                       0.237
Model:                            OLS   Adj. R-squared:                  0.221
Method:                 Least Squares   F-statistic:                     14.72
Date:                Sat, 02 Dec 2023   Prob (F-statistic):           5.86e-25
Time:                        11:51:39   Log-Likelihood:                -1166.1
No. Observations:                 533   AIC:                             2356.
Df Residuals:                     521   BIC:                             2407.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept        -19.5202      2.090     -9.339      0.000     -23.627     -15.414
y_2000[T.True]    -1.3098      0.459     -2.854      0.004      -2.211      -0.408
y_2001[T.True]    -1.8173      0.526     -3.455      0.001      -2.851      -0.784
y_2002[T.True]    -1.2292      0.594     -2.070      0.039      -2.395      -0.063
y_2003[T.True]    -0.8701      0.737     -1.181      0.238      -2.317       0.577
y_2004[T.True]    -0.4074      0.737     -0.553      0.580      -1.854       1.040
y_2005[T.True]     0.0917      0.642      0.143      0.886      -1.169       1.352
y_2006[T.True]    -0.0727      0.643     -0.113      0.910      -1.335       1.190
y_2007[T.True]    -0.5813      0.620     -0.937      0.349      -1.799       0.637
y_2008[T.True]    -0.2913      0.579     -0.503      0.615      -1.429       0.847
y_2009[T.True]    -0.8211      0.579     -1.418      0.157      -1.959       0.316
pop                0.0002   1.72e-05     11.588      0.000       0.000       0.000
==============================================================================
Omnibus:                       37.564   Durbin-Watson:                   1.939
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               44.431
Skew:                          -0.706   Prob(JB):                     2.25e-10
Kurtosis:                       2.923   Cond. No.                     2.70e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.7e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.237 - i.e. 23.7% of revenue is explained by independent variables.

For population, the coefficient for the population is 0.0002. The p-value of 0.000 indicates that the relationship between '**pop**' **and log_revt(dependent variable) is statistically significant**.

In Regression 17, the analysis examines the influence of the population as an independent variable on the log-transformed revenue (dependent variable) throughout the 2000s, considering control variables from 2000 to 2009. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.237, which implies that the population explains 23.7% of the revenue during the 2000s. The population has a coefficient of 0.0002, indicating a positive association with the company's revenue. Moreover, this coefficient (pop) is statistically significant at a $p<0.01$ level. The year-wise analysis reveals that only 2000 and 2001 are statistically significant with negative association and all other years are not statistically significant.

Regression 18: Revenue during 2010's based on population.

```
In [163…  model11 = smf.ols(formula='log_revt ~ pop+y_2010+y_2011+y_2012+y_2013+y_2014+y_2015+y_2016+y_2017+y_2018+y_2019', data=dff_labor_pop_wrds_1980_2022)
          results11 = model11.fit()
          print(results11.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_revt   R-squared:                       0.221
Model:                            OLS   Adj. R-squared:                  0.205
Method:                 Least Squares   F-statistic:                     13.48
Date:                Sat, 02 Dec 2023   Prob (F-statistic):           8.48e-23
Time:                        11:51:57   Log-Likelihood:                -1171.5
No. Observations:                 533   AIC:                             2367.
Df Residuals:                     521   BIC:                             2418.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -18.9394      2.243     -8.442      0.000     -23.347     -14.532
y_2010[T.True]    -0.8904      0.555     -1.603      0.109      -1.981       0.201
y_2011[T.True]    -0.4373      0.539     -0.811      0.418      -1.496       0.622
y_2012[T.True]    -0.9765      0.512     -1.906      0.057      -1.983       0.030
y_2013[T.True]    -0.2886      0.553     -0.522      0.602      -1.374       0.797
y_2014[T.True]    -0.0998      0.606     -0.165      0.869      -1.290       1.090
y_2015[T.True]     0.1054      0.571      0.185      0.854      -1.016       1.226
y_2016[T.True]    -0.0617      0.557     -0.111      0.912      -1.156       1.032
y_2017[T.True]     0.7122      0.590      1.207      0.228      -0.447       1.871
y_2018[T.True]     0.6727      0.559      1.203      0.230      -0.426       1.771
y_2019[T.True]     0.9294      0.531      1.750      0.081      -0.114       1.973
pop            0.0002   1.87e-05     10.359      0.000       0.000       0.000
==============================================================================
Omnibus:                       41.436   Durbin-Watson:                   1.895
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.603
Skew:                          -0.746   Prob(JB):                     1.69e-11
Kurtosis:                       3.096   Cond. No.                     2.89e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.89e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.221 - i.e. 22.1% of revenue is explained by independent variables.

For population, the coefficient for the population is 0.0002. The p-value of 0.000 indicates that the relationship between '**pop**' **and log_revt(dependent variable) is statistically significant**.

In Regression 18, the analysis examines the influence of the population as an independent variable on the log-transformed revenue (dependent variable) throughout the 2010s, considering control variables from 2010 to 2019. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.221, which implies that the population explains 22.1% of the revenue during the 2010s. The population has a coefficient of 0.0002, indicating a positive association with the company's revenue. Moreover, this coefficient (pop) is statistically significant at a $p<0.01$ level. The year-wise analysis reveals that years are not statistically significant.

Interpretation of the Overall Results for the Impact of Population on Revenue Across Decades (1990s, 2000s, 2010s)

Based on the regression analyses, population plays a significant role in influencing the revenue generated by companies in the Real Estate sector. An increase in population leads to an increase in the revenue of these organizations across the examined decades.

Overall Results Interpretation of population on revenue for decade results (1990's, 2000's, 2010's)

From the above regressions, it is evident that population has a significant impact on the revenue of organizations in Real Estate, where an increase in population increases the revenue of organizations.

The following regression analysis on ReMax Holdings is based on economic indicators. We only have Compustat data from 2011, so regression is based on small dataset.

In [167…
```
# Check years of data available for ReMax
rev_remax_1980_2022.loc[:,'year']
```

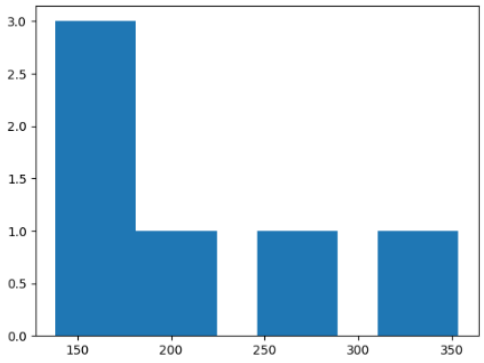
Out[167…
```
302    2011
321    2012
341    2013
358    2014
372    2015
388    2016
404    2017
419    2018
436    2019
455    2020
479    2021
506    2022
Name: year, dtype: int32
```

Analysis on ReMax Holding data.

Data for ReMax Holding only exists from 2011 onwards.

Histogram on ReMax Holding data.

In [168…
```
# Understand the distribution of data among revt (dependent variable) using a Histogram
plt.hist(rev_remax_1980_2022['revt'])
plt.show()
```



Interpretation on ReMax Holding histogram based on revenue

The data is positively skewed

Regression 20: Revenue of ReMax Holdings based on population.

```
In [173…  model12 = smf.ols(formula='log_revt ~ pop', data=rev_remax_1980_2022)
          results12 = model12.fit()
          print(results12.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_revt   R-squared:                       0.765
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     32.57
Date:                Sat, 02 Dec 2023   Prob (F-statistic):           0.000196
Time:                        11:57:57   Log-Likelihood:                 5.9620
No. Observations:                  12   AIC:                            -7.924
Df Residuals:                      10   BIC:                            -6.954
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -33.2797      6.645     -5.008      0.001     -48.086     -18.473
pop            0.0003   5.29e-05      5.707      0.000       0.000       0.000
==============================================================================
Omnibus:                        0.004   Durbin-Watson:                   1.267
Prob(Omnibus):                  0.998   Jarque-Bera (JB):                0.167
Skew:                           0.023   Prob(JB):                        0.920
Kurtosis:                       2.423   Cond. No.                     1.79e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.79e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/scipy/stats/_stats_py.py:1806: UserWarning: kurtosistest only valid for n>=20 ... continuing any
way, n=12
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.765 - i.e. 76.5% of revenue is explained by independent variables.

For population, the coefficient for the population is 0.0003. The p-value of 0.000 indicates that the relationship between **'pop' and log_revt(dependent variable) is statistically significant**.

In Regression 20, the analysis examines the influence of the population as an independent variable on the log-transformed revenue (dependent variable) of ReMax Holdings. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.765, which implies that the population explains 76.5% of the revenue of ReMax Holdings. The population has a coefficient of 0.0003, indicating a positive association with the company's revenue. Moreover, this coefficient (pop) is statistically significant at a $p<0.01$ level.

Regression 21: Revenue of ReMax Holdings based on population and effective rate.

```
In [174…
model12 = smf.ols(formula='log_revt ~ pop+effective_rate', data=rev_remax_1980_2022)
results12 = model12.fit()
print(results12.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_revt   R-squared:                       0.789
Model:                           OLS   Adj. R-squared:                  0.742
Method:                Least Squares   F-statistic:                     16.84
Date:               Sat, 02 Dec 2023   Prob (F-statistic):           0.000907
Time:                       11:58:28   Log-Likelihood:                 6.6102
No. Observations:                  12   AIC:                            -7.220
Df Residuals:                       9   BIC:                            -5.766
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -40.5418      9.768     -4.151      0.002     -62.638     -18.446
pop            0.0004   7.82e-05      4.611      0.001       0.000       0.001
effective_rate    -0.0926      0.091     -1.013      0.337      -0.299       0.114
==============================================================================
Omnibus:                        0.631   Durbin-Watson:                   1.541
Prob(Omnibus):                  0.730   Jarque-Bera (JB):                0.628
Skew:                          -0.329   Prob(JB):                        0.730
Kurtosis:                       2.092   Cond. No.                     2.64e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.64e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/scipy/stats/_stats_py.py:1806: UserWarning: kurtosistest only valid for n>=20 ... continuing any
way, n=12
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

R-squared is 0.789 - i.e. 78.9% of revenue is explained by independent variables.

For population, the coefficient for the population is 0.0004. The p-value of 0.001 indicates that the relationship between '**pop**' **and log_revt(dependent variable) is statistically significant**.

For effective_rate, the coefficient of -0.0926 suggests a negative relationship with revenue. However, due to the high p-value 0.337, the relation between '**effective_rate**' **and log_revt(dependent variable) is not statistically significant**.

In Regression 21, the analysis examines the influence of the population and interest rates (effective_rate) as independent variables on the log-transformed revenue (dependent variable) of ReMax Holdings. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.789, which implies that the population explains 78.9% of the revenue of ReMax Holdings. The population has a coefficient of 0.0003, indicating a positive association with the company's revenue. Moreover, this coefficient (pop) is statistically significant at a $p<0.01$ level. However, effective_rate with the coefficient of –0.0926 suggests a negative relationship, moreover, due to high p-value 0.337, effective_rate is not statistically significant.

Regression of effect on revenue of ReMax Holding based on effective rate and house hold income.

Regression 23: Revenue for ReMax Holdings based on effective rate and house income.

In [176…

```
model12 = smf.ols(formula='log_revt ~ effective_rate+house_income', data=rev_remax_1980_2022)
results12 = model12.fit()
print(results12.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_revt   R-squared:                       0.335
Model:                            OLS   Adj. R-squared:                  0.187
Method:                 Least Squares   F-statistic:                     2.266
Date:                Sat, 02 Dec 2023   Prob (F-statistic):              0.160
Time:                        12:00:35   Log-Likelihood:                -0.28271
No. Observations:                  12   AIC:                             6.565
Df Residuals:                       9   BIC:                             8.020
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.2552      0.334     12.757      0.000       3.501       5.010
effective_rate     0.2526      0.119      2.129      0.062      -0.016       0.521
house_income       0.0001      0.000      0.771      0.461      -0.000       0.000
==============================================================================
Omnibus:                        8.727   Durbin-Watson:                   0.665
Prob(Omnibus):                  0.013   Jarque-Bera (JB):                4.363
Skew:                           1.347   Prob(JB):                        0.113
Kurtosis:                       4.211   Cond. No.                     8.36e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.36e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/scipy/stats/_stats_py.py:1806: UserWarning: kurtosistest only valid for n>=20 ... continuing any
way, n=12
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

Result interpretation on regression

Prob (F-statistic) is greater than 0.1, this regression analysis cannot be considered as it is not statistically significant.

R-squared is 0.335 - i.e. 33.5% of revenue is explained by independent variables.

The coefficients effective_rate or house_income do not have any significant relationship with log_revt.

In Regression 23, the analysis examines the influence of the interest rates (effective_rate) and household income(house_income) as independent variables on the log-transformed revenue (dependent variable) of ReMax Holdings. The F-stat test score is above 0.1, indicating that the regression analysis cannot be considered. Even though, the r-squared value is 0.335. The coefficient are not statistically significant.

Regression of effect on revenue of ReMax Holding based on effective rate, poplulation and house hold income.

Regression 24: Revenue during for ReMax Holdings based on effective rate, house income and population.

In [177…

```
model12 = smf.ols(formula='log_revt ~ effective_rate+house_income+pop', data=rev_remax_1980_2022)
results12 = model12.fit()
print(results12.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               log_revt   R-squared:                       0.790
Model:                            OLS   Adj. R-squared:                  0.711
Method:                 Least Squares   F-statistic:                     10.00
Date:                Sat, 02 Dec 2023   Prob (F-statistic):            0.00440
Time:                        12:01:31   Log-Likelihood:                 6.6208
No. Observations:                  12   AIC:                            -5.242
Df Residuals:                       8   BIC:                            -3.302
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept       -40.2189     10.701     -3.759      0.006     -64.895     -15.543
effective_rate   -0.0869      0.108     -0.804      0.444      -0.336       0.162
house_income  1.084e-05   9.11e-05      0.119      0.908      -0.000       0.000
pop               0.0004    8.6e-05      4.157      0.003       0.000       0.001
==============================================================================
Omnibus:                        0.528   Durbin-Watson:                   1.599
Prob(Omnibus):                  0.768   Jarque-Bera (JB):                0.569
Skew:                          -0.275   Prob(JB):                        0.752
Kurtosis:                       2.086   Cond. No.                     2.73e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.73e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/scipy/stats/_stats_py.py:1806: UserWarning: kurtosistest only valid for n>=20 ... continuing any
way, n=12
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

Result interpretation on regression

Prob (F-statistic) is less than 0.1, this regression analysis can be considered.

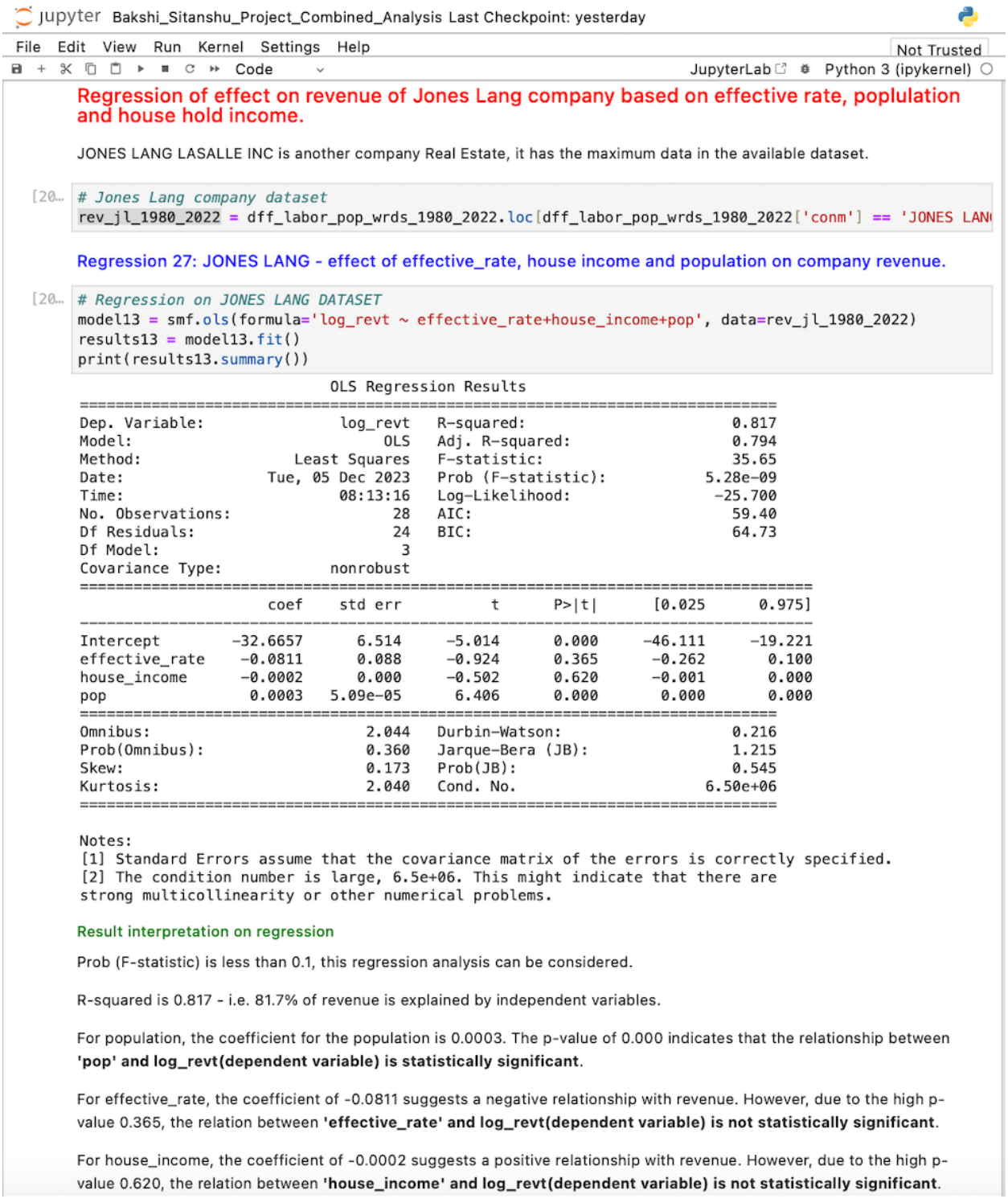
R-squared is 0.790 - i.e. 79% of revenue is explained by independent variables.

For population, the coefficient for the population is 0.0004. The p-value of 0.003 indicates that the relationship between '**pop**' **and log_revt(dependent variable) is statistically significant**.

For effective_rate, the coefficient of -0.0869 suggests a negative relationship with revenue. However, due to the high p-value 0.908, the relation between '**effective_rate**' **and log_revt(dependent variable) is not statistically significant**.

For house_income, the coefficient of 1.084e-05 suggests a positive relationship with revenue. However, due to the high p-value 0.444, the relation between '**house_income**' **and log_revt(dependent variable) is not statistically significant**.

In Regression 24, the analysis examines the influence of all the economic variables - the population, interest rates (effective_rate) and house income as independent variables on the log-transformed revenue (dependent variable) of ReMax Holdings. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.79, which implies that the population explains 79% of the revenue of ReMax Holdings. The population has a coefficient of 0.0004, indicating a positive association with the company's revenue. Moreover, this coefficient (pop) is statistically significant at a p<0.01 level. However, effective_rate with the coefficient of –0.0926 suggests a negative relationship, moreover, due to high p-value 0.337, effective_rate is not statistically significant. Additionally, house_income with high p-value is not statistically significant for the model.

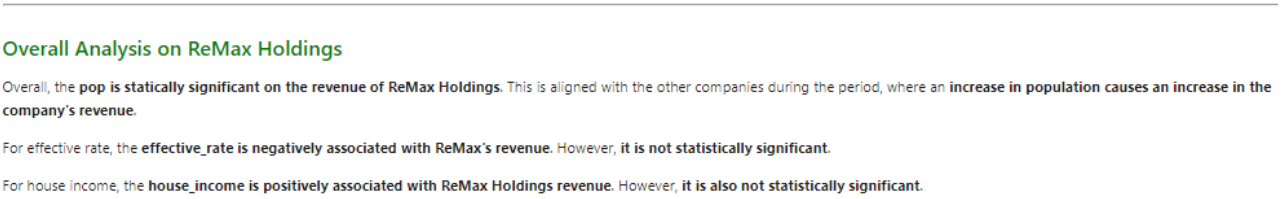


In Regression 27, the analysis examines the influence of all the economic variables - the population, interest rates (effective_rate) and house income as independent variables on the log-transformed revenue (dependent variable) of JONES LANG (another company in industry) to confirm our findings. The F-stat test score is below 0.1, indicating that the regression analysis is considered valid. The r-squared value is 0.817, which implies that the population explains 81.7% of the revenue of JONES LANG. The

population has a coefficient of 0.0003, indicating a positive association with the company's revenue. Moreover, this coefficient (pop) is statistically significant at a p<0.01 level. However, effective_rate with the coefficient of –0.0811 suggests a negative relationship, moreover, due to high p-value 0.365, effective_rate is not statistically significant. Additionally, house_income with high p-value is not statistically significant for the model.

The overall analysis on ReMax Holdings suggests that population is statistically significant on the revenue of ReMax Holdings. Increase in population causes an increase in revenue.

The effective rate is negatively associated with ReMax Holding's revenue but is not statistically significant. Likewise, house_income with positive association is also not statistically significant.



Conclusion

Business implication

What do we learn from this analysis?

In the analysis for Clark & Co consulting company, factors such as interest rates, population, and household income were considered for the real estate agents and managers industry.

The trends on interest rates revealed that a rise in effective rates led to a decrease in the organization's revenue. This implies that when effective rates increase, the revenue pattern of the real estate industry declines, and the opposite is also true. During periods of high population growth along with increased employment opportunities for the labor force, there is a surge in demand for estate homes, resulting in a corresponding increase in revenue for the industry. However, household income alone does not significantly impact the revenues of businesses in the real estate industry.

Limitations of this research

- One of limitation is the insufficient data for RE/MAX Holdings Inc, as the data available in Compustat (WRDS) for the company only spans from 2011 to 2022. Due to lack of sufficient data, the analysis for ReMax holdings could not be performed for a longer time frame. However, the past decade has seen significant fluctuations in both the world and the US economy, making the analysis still relevant in the current context.
- Analysis done on industry as a whole and specifically on RE/MAX holdings, so the findings or the results from RE/MAX holdings cannot be generalized for other real-estate agents and manager industry Companies.

- This research included only three economic indicators- interest rate, population and household income. These factors alone do not determine the overall performance of RE/MAX Holdings.

Potential project

Prescriptive analytics

Definition: Prescriptive analytics refers to the method of utilizing data to identify the best course of action. This kind of research produces suggestions for further actions by taking into account all relevant elements. For this reason, prescriptive analytics is an important tool in data-driven decision-making.

https://online.hbs.edu/blog/post/prescriptive-analytics

Prescriptive analysis method involves using data to determine the best course of action. In the ReMax Holdings case, the analysis showed a statistically significant relationship between population growth and revenue.

Research question: Our future research question for prescriptive analytics would be, "What strategies ReMax holdings can adopt to move their business to high population growth areas to adopt maximum revenue?"

Predictive analytics

Definition: By using previous data to predict future patterns and occurrences, predictive analytics is a potent tool that helps firms make well-informed strategic decisions. Predictive analytics helps businesses anticipate possible outcomes, proactively solve problems, and take advantage of opportunities by examining historical trends and behaviors.

https://online.hbs.edu/blog/post/predictive-analytics

Research question: This approach focuses on predicting future outcomes based on historical data. In the case of ReMax Holdings, our research question for predictive analytics will be, "How will changes in interest rates, population growth, and household income affect the future revenue of ReMax Holdings?"