US Economic Analysis

Simone Trindade Steel

02/08/2020

An analysis of national economic sentiment effect on new housing offers

1. Introduction and executive summary

This reports is based on a study of business and industry trends in the United States using the US Census Bureau data available on https://www.kaggle.com/census/business-and-industry-reports.

The inspiration for this study is the hypothesis that US National level of value generation, as an indication of economic optimism, drives the anticipation of new property demand.

The analysis looks for regional variations in the US and attempts to predict the regional growth of new home construction and sales.

The main objective is to unveil whether or not there is an identifiable time lag between the perception of a buoyant economic environment and the offer of newly built houses on the market. Given that the construction sector does not have the same agility as other parts of the economy, such as retail and services, it could be important for investors, government agencies and policy makers to prevent over or under investment in the construction sector.

By identifying a time lag between stimulus and result, investment in construction could be self-regulated to avoid large oscillation in supply, which may contribute to distortions in the market, such as the so-called "housing bubbles".

This report will cover the exploration and preparation of data, the construction of alternative predictive models, and the results analysis leading to the report conclusion.

2. Method and analysis

2.1. Initial considerations

House construction is a relatively slow-moving activity. Securing funds, land and necessary permissions are traditionally extended processes. An important assumption in this analysis is that there is an interval between decision to build and houses becoming available, i.e. time lag.

This report includes the analysis of several values for time lag, aiming to detect the interval that produces the best predictive capability for investment expansion and contraction trajectories. The time lag values chosen for this analysis will be based on a short Fibonacci sequence (1, 2, 3, 5, 8, 13 expressed and calendar years).

Another dimension of analysis is region within the United States. This is to verify the hypothesis that regional differences in dominant industries and variations in economic cycles lead to different behaviour towards real estate investment.

In summary:

a. This report uses the standard mechanism for randomly splitting the training and test sets using probability of 0.9 and 0.1, respectively.

b. The outcome of the prediction will be categorical: 1 representing above average housing offers for the region and 0 representing below average.

c. Overall accuracy was chosen as the best method for measuring success of the predictive model, i.e. maximising the proportion of correct predictions on the test set.

d. The objectives are to identify (1) the time lag between economic indicators and housing offers and (2) the best performing predictive model or ensemble of models that maximises overall accuracy.

2.2. Understanding the data

The original dataset has many economic factors that will not be used in this study.

The time series that are relevant to the subject of this report are:

a. The macroeconomic indicators (named "Financial Reports") and

b. The new housing indicators (named "New Home Sales", "New Residential Construction").

The metadata file for these two subjects provide the time series code to access the available data:

report	number_of_timeseries
New Home Sales	10
New Residential Construction	25

report	number_of_timeseries
Quarterly Financial Report	52

```
## [1] "First 10 rows of relevant Housing Data"
```

time_series_code	date	value
PERMITS_TOTAL_US	1959-01-01	75.7
PERMITS_TOTAL_NE	1959-01-01	11.4
PERMITS_TOTAL_MW	1959-01-01	11.8
PERMITS_TOTAL_SO	1959-01-01	26.8
PERMITS_TOTAL_WE	1959-01-01	25.6
STARTS_TOTAL_US	1959-01-01	96.2
STARTS_TOTAL_NE	1959-01-01	15.4
STARTS_TOTAL_MW	1959-01-01	19.8
STARTS_TOTAL_SO	1959-01-01	34.5
STARTS_TOTAL_WE	1959-01-01	26.6

```
## [1] "Total rows for Housing Data: " "21465"
```

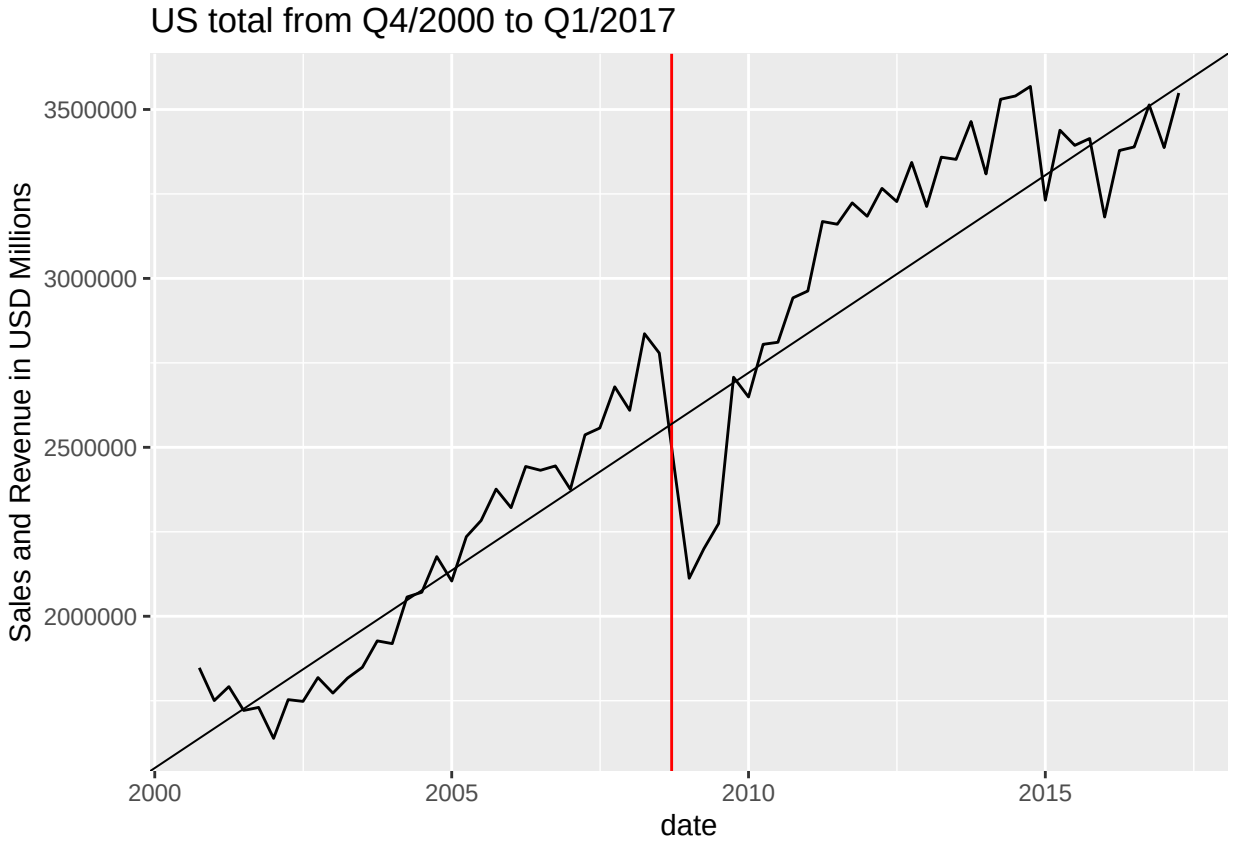
```
## [1] "First 10 rows of relevant Financial Data"
```

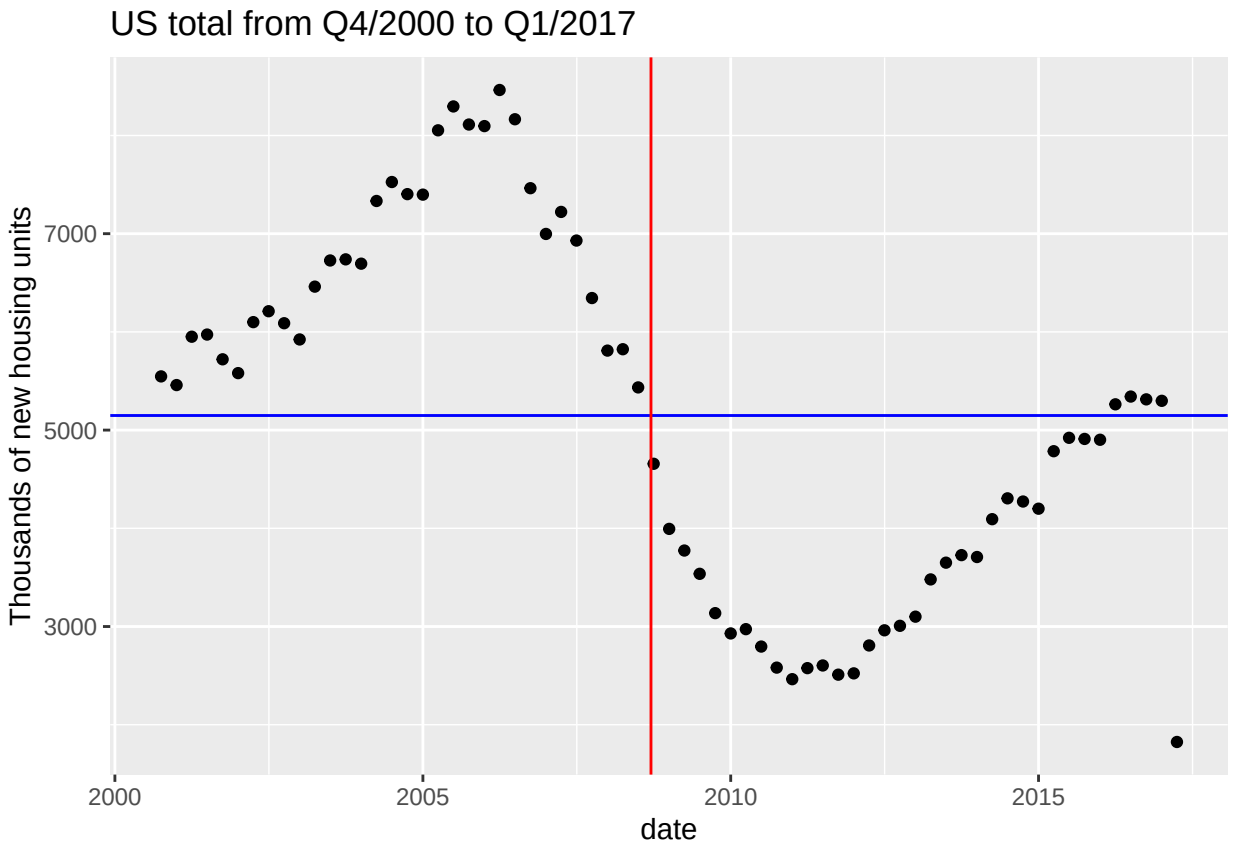
time_series_code	date	value	industry
MFG_101_US	2000-10-01	1128790	MFG
MIN_101_US	2000-10-01	31852	MIN
RET_101_US	2000-10-01	356432	RET
WHS_101_US	2000-10-01	330326	WHS
MFG_101_US	2001-01-01	1082233	MFG
MIN_101_US	2001-01-01	34380	MIN
RET_101_US	2001-01-01	318014	RET
WHS_101_US	2001-01-01	315767	WHS
MFG_101_US	2001-04-01	1116597	MFG
MIN_101_US	2001-04-01	30551	MIN

```
## [1] "Total rows for Financial Data: " "330"
```

All dates indicate the beginning of the analysis period, and they have been used to align the time series into quarterly periods - January, April, July and October.

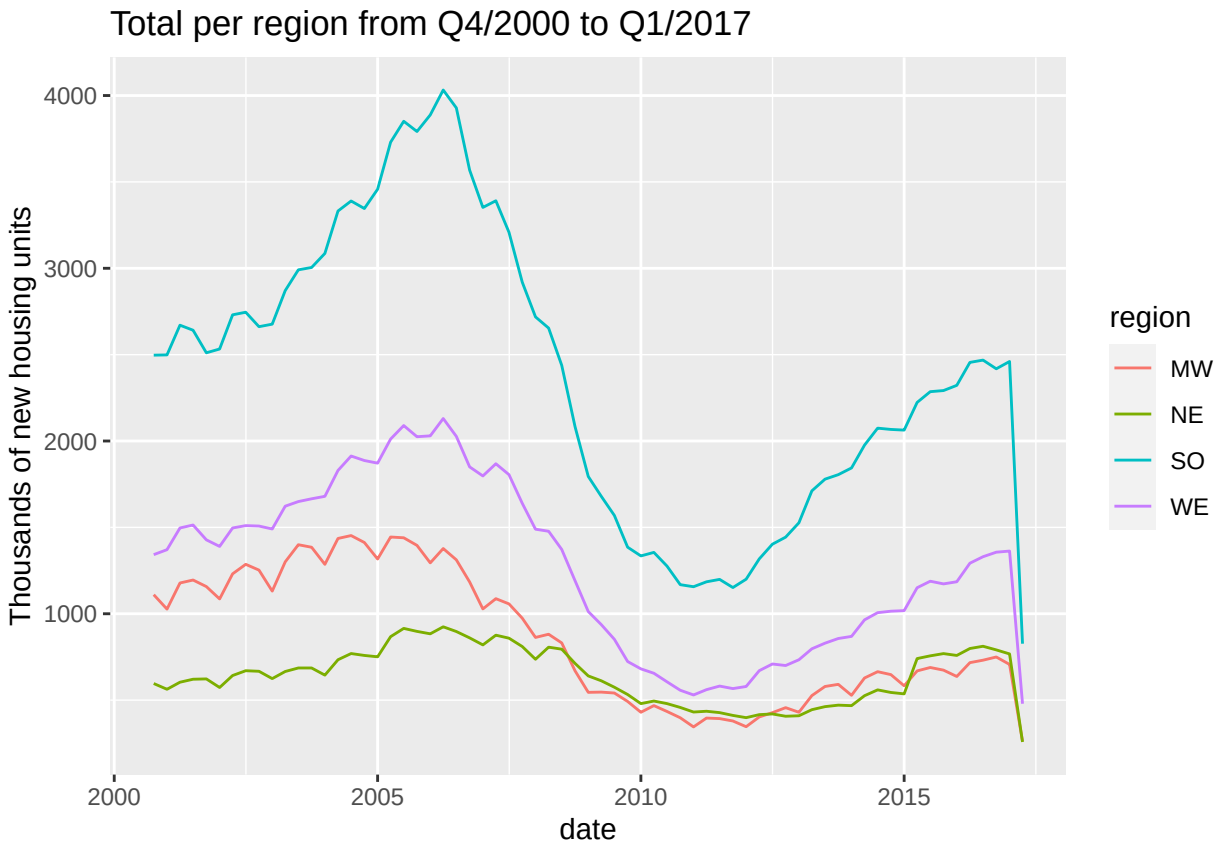
The graphs below summarise the US national totals, and show the disruption created by the 2008 crisis. Following the crisis, revenue resumed its upwards trajectory, as seen with broader economic recovery.





The regional graphs, however, show that new housing construction and offering vary along each individual regional average that appears not related to the broader revenue generation at national level.

Looking closer at each region, the house offer variability in the South is far greater than the other regions. The Northeast is stable relatively to the national income fluctuations.



In order to understand what drives an increase and decrease of new housing, this analysis will categorise new housing offers above average as 1, and below average as 0, relative to each regional average.

2.3. Preparing data for analysis

Several steps are performed here:

(1) Creating different time lag data and aligning them with income generation and house construction. As explained earlier, a short Fibonacci sequence is going to be used to explore different time lags: 1, 2, 3, 5, 8 and 13 years.

(2) Removing aggregated US National data points in order not to double count properties.

(3) Preparing regional averages, given that the variations are significant between regions.

(4) Introducing industry classification, given their regional variation.

Industry codes are:

```
## [1] "Industry codes and description"
```

Code	Description
MFG	All Manufacturing
MIN	All Mining
RET	All Retail Trade

Code	Description
WHS	All Wholesale Trade
INF	All Information
PTS	All Professional and Technical Services, Except Legal Services

```
## [1] "First 10 rows of relevant Financial Data, with extra columns representing time lag"
```

```
##    time_series_code       date   value industry     lag_1y     lag_2y
## 1        MFG_101_US 2000-10-01 1128790      MFG 2001-10-01 2002-10-01
## 2        MIN_101_US 2000-10-01   31852      MIN 2001-10-01 2002-10-01
## 3        RET_101_US 2000-10-01  356432      RET 2001-10-01 2002-10-01
## 4        WHS_101_US 2000-10-01  330326      WHS 2001-10-01 2002-10-01
## 5        MFG_101_US 2001-01-01 1082233      MFG 2002-01-01 2003-01-01
## 6        MIN_101_US 2001-01-01   34380      MIN 2002-01-01 2003-01-01
## 7        RET_101_US 2001-01-01  318014      RET 2002-01-01 2003-01-01
## 8        WHS_101_US 2001-01-01  315767      WHS 2002-01-01 2003-01-01
## 9        MFG_101_US 2001-04-01 1116597      MFG 2002-04-01 2003-04-01
## 10       MIN_101_US 2001-04-01   30551      MIN 2002-04-01 2003-04-01
##        lag_3y     lag_5y     lag_8y    lag_13y
## 1  2003-10-01 2005-10-01 2008-10-01 2013-10-01
## 2  2003-10-01 2005-10-01 2008-10-01 2013-10-01
## 3  2003-10-01 2005-10-01 2008-10-01 2013-10-01
## 4  2003-10-01 2005-10-01 2008-10-01 2013-10-01
## 5  2004-01-01 2006-01-01 2009-01-01 2014-01-01
## 6  2004-01-01 2006-01-01 2009-01-01 2014-01-01
## 7  2004-01-01 2006-01-01 2009-01-01 2014-01-01
## 8  2004-01-01 2006-01-01 2009-01-01 2014-01-01
## 9  2004-04-01 2006-04-01 2009-04-01 2014-04-01
## 10 2004-04-01 2006-04-01 2009-04-01 2014-04-01
```

```
## [1] "Number of data points available by industry when the different time lags were introduced"
```

industry	sum1y	sum2y	sum3y	sum5y	sum8y	sum13y
INF	135	115	95	55	0	0
MFG	315	295	275	235	175	75
MIN	315	295	275	235	175	75
PTS	135	115	95	55	0	0
RET	315	295	275	235	175	75
WHS	315	295	275	235	175	75

Now that the datasets are aligned in quarterly periods and by each of the hypothesis of time lag, the following graphs show some interesting relationships that will inform choices for model-based predictions.

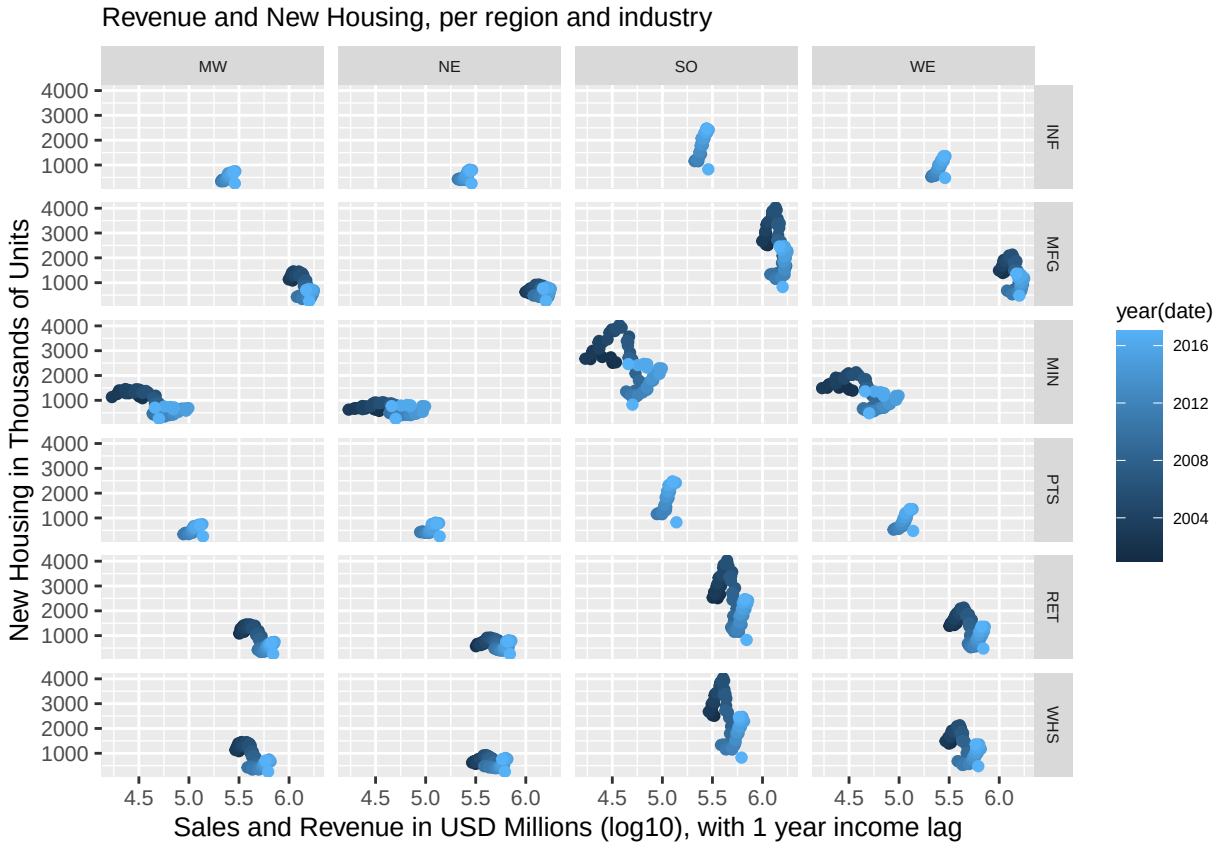
Each graph is organised to show the relationship between income (x axis) and new building activity (y axis), in vertical facets for each region and horizontal facets by industry. The shades in the graph represent time in calendar years.

When looking at the same point in time, pre-2008 (darker shades) data shows new housing activity was higher in the South and West regions. But over time, it seems negatively influenced by income.

Does this support the notion of a time lag between income from sales and revenue rising and new housing

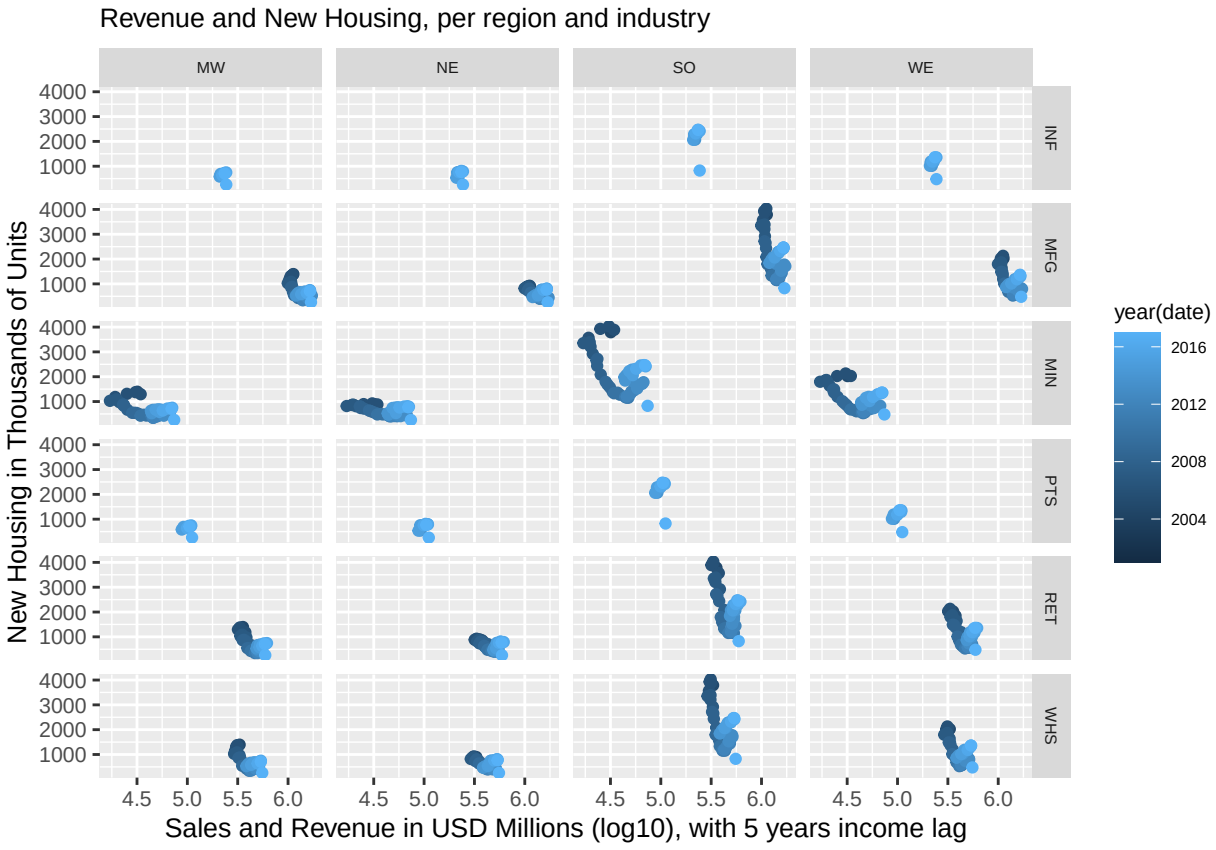
activity occurring?

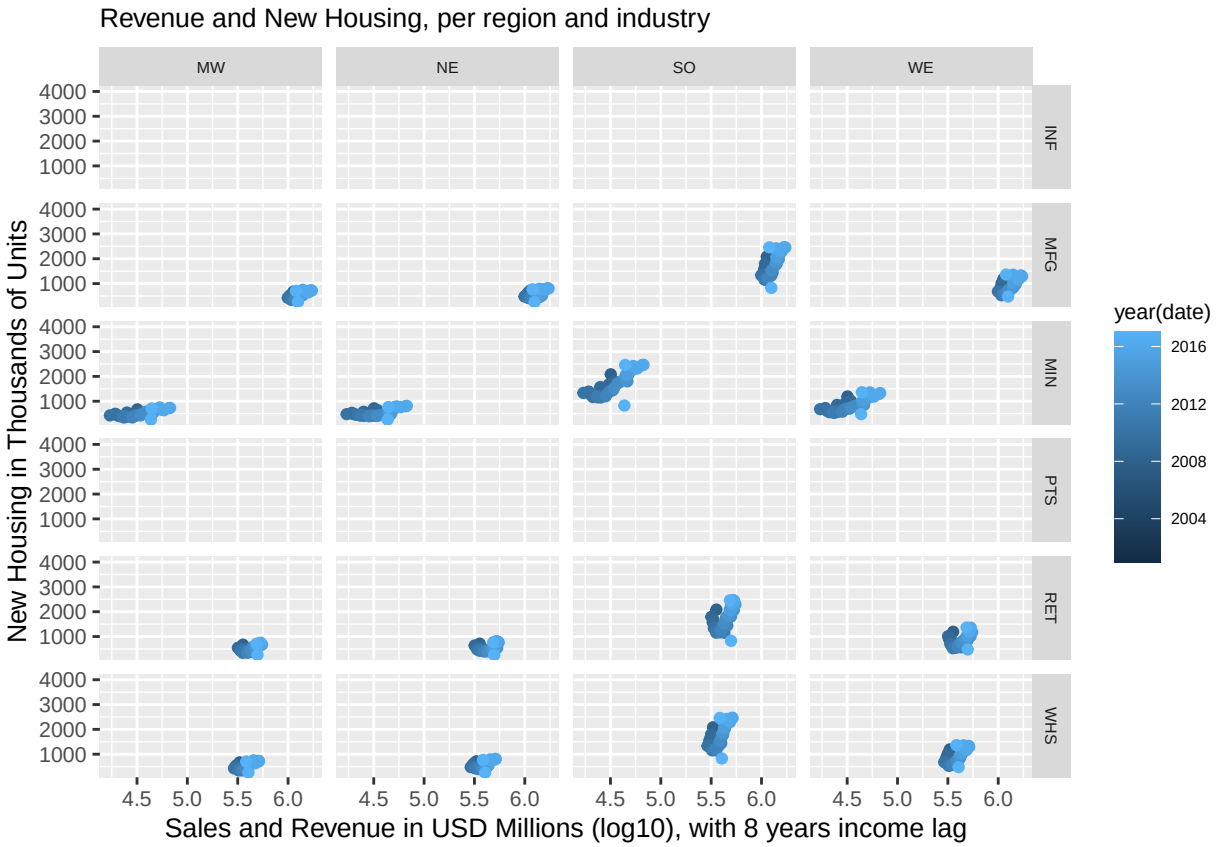
The graphs also show (just as the table above) that there are no data points for 8- and 13-year lag for Information and Professional and Technical services, as these are relatively new industries from the point of view of census records. Due to the lack of significant amount of data, 8- and 13-year lag will not be considered in the building of predictive models.




Revenue and New Housing, per region and industry
MW
NE
SO
WE
INF
MFG
MIN
PTS
RET
WHS
4000
3000
2000
1000
4.5
5.0
5.5
6.0
New Housing in Thousands of Units
Sales and Revenue in USD Millions (log10), with 2 years income lag
year(date)
2016
2012
2008
2004


Revenue and New Housing, per region and industry
MW
NE
SO
WE
INF
MFG
MIN
PTS
RET
WHS
4000
3000
2000
1000
4.5
5.0
5.5
6.0
New Housing in Thousands of Units
Sales and Revenue in USD Millions (log10), with 3 years income lag
year(date)
2016
2012
2008
2004


Revenue and New Housing, per region and industry
MW
NE
SO
WE
INF
MFG
MIN
PTS
RET
WHS
4000
3000
2000
1000
4.5
5.0
5.5
6.0
New Housing in Thousands of Units
Sales and Revenue in USD Millions (log10), with 5 years income lag
year(date)
2016
2012
2008
2004


Revenue and New Housing, per region and industry
MW
NE
SO
WE
INF
MFG
MIN
PTS
RET
WHS
4000
3000
2000
1000
4.5
5.0
5.5
6.0
New Housing in Thousands of Units
Sales and Revenue in USD Millions (log10), with 8 years income lag
year(date)
2016
2012
2008
2004

Given that there are considerable differences in new housing per region, the trend will be added to the dataset relative to regional means.

```
## [1] "First 10 rows showing the addition of Trend (0 or 1) to the dataset for each region"
```

```
##         date region industry new_houses_K nat_income_M_1y nat_income_M_2y
## 1 2001-10-01     MW      MFG         1157         1128790              NA
## 2 2001-10-01     MW      MIN         1157           31852              NA
## 3 2001-10-01     MW      RET         1157          356432              NA
## 4 2001-10-01     MW      WHS         1157          330326              NA
## 5 2001-10-01     NE      MFG          623         1128790              NA
## 6 2001-10-01     NE      MIN          623           31852              NA
##   nat_income_M_3y nat_income_M_5y avg trend
## 1              NA              NA 788     1
## 2              NA              NA 788     1
## 3              NA              NA 788     1
## 4              NA              NA 788     1
## 5              NA              NA 627     0
## 6              NA              NA 627     0
```

2.4. Making predictions and finding optimal time lag parameter

This report aims to demonstrate that the regional trends of new house construction is lagging the sentiment of a positive economic position, as observed by financial reports on sales, invoicing and revenue at the national level.

In order to do that, the concept of positive or negative trend was introduced. Positive trend means that regional construction is above historical average, and negative trend means below average.

The models that appeared most appropriate for this exercise were GLM, KNN and Random Forest. An ensemble of models was also introduced to verify if it could improve on the predictions of individual models.

As mentioned at the beginning of this report, the dataset for training the models and testing them was randomly split from the original with probability of 0.9 and 0.1, respectively. Each of the models was fit and tested for time lags of 1, 2, 3 and 5 years.

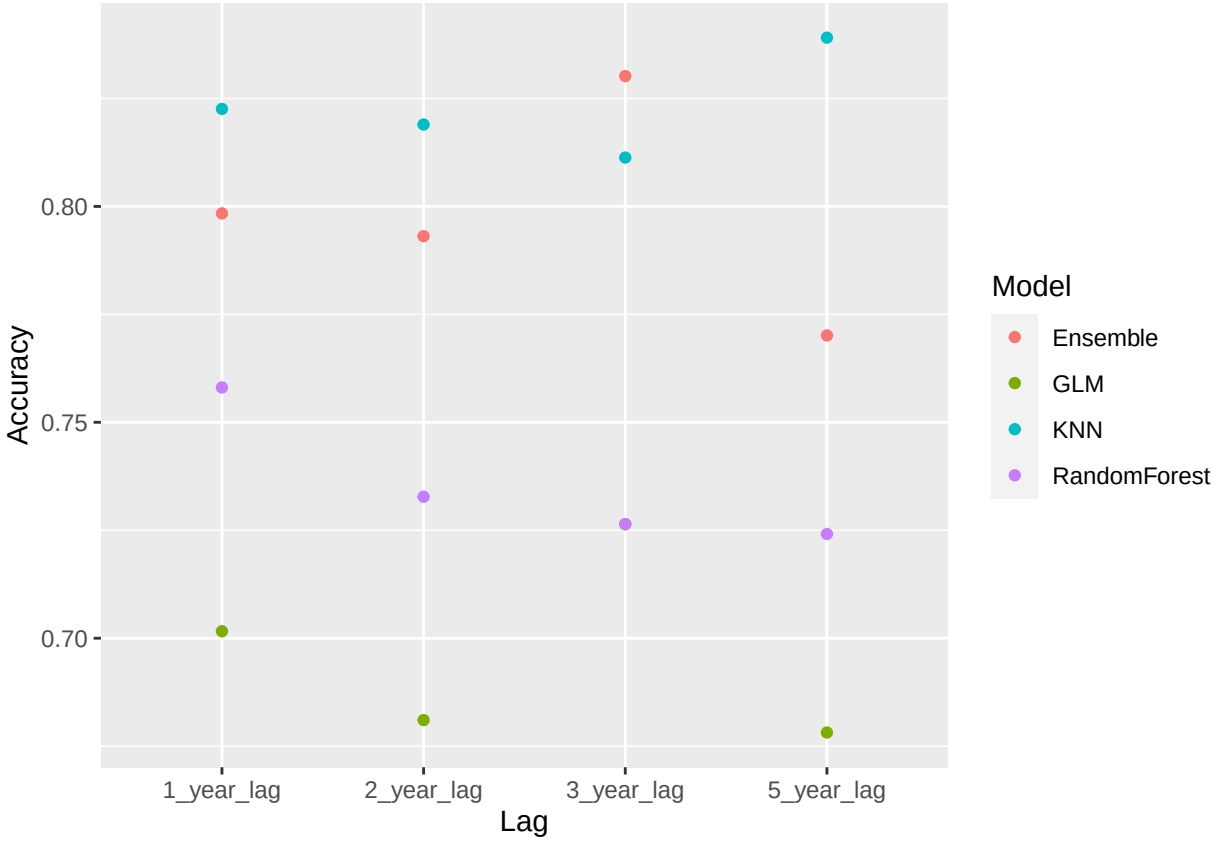
3. Results

The maximum accuracy of 0.839 was obtained by the KNN model with k=4 and a time lag of 5 years between income and new housing activity.

The accuracy of each model and time lag combination were gathered during the execution of the R script and are displayed below as a table and a plot graph.

```
## [1] "Accuracy results of predictive models built and tested in this report"
```

	1_year_lag	2_year_lag	3_year_lag	5_year_lag
GLM	0.702	0.681	0.726	0.678
RandomForest	0.758	0.733	0.726	0.724
KNN	0.823	0.819	0.811	0.839
Ensemble	0.798	0.793	0.830	0.770



4. Conclusion

This report shows that a predictive model can be used to determine the volume, relative to the regional average, of new houses that will be offered 5 years later based on the economic outlook in the United States. The models were built to take into account regional and industry differences and the best performing one was able to produce a maximum accuracy 0.839.

Although accuracy is not particularly high, this may still have useful applications for investors and policy makers to prevent over and under investment, which can have damaging effects in the economy because of the illiquid nature of real estate assets.

This report has several limitations, notably:

(a) As with any macroeconomic analysis, the inter-dependencies and correlations between indicators are complex and were not fully explored in this report.

(b) There is a relatively small number of data points available, especially for recently developed industries such as Information, Professional and Technical Services sectors.

(c) Other predictive models and techniques not explored in this report may yield better accuracy results.

Future studies based on this approach could enrich the understanding between perception of prosperity and attitude towards real estate investment, such as tax structure changes, specific cyclical nature of each industry and internal migration of workforce.