USBusinessReport

Simone Trindade Steel

02/08/2020

US Business Report: an analysis of economic sentiment and new housing construction

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”.

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 (a) the time lag between economic indicators and housing offers and (b) 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").

All dates indicate the begining 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.

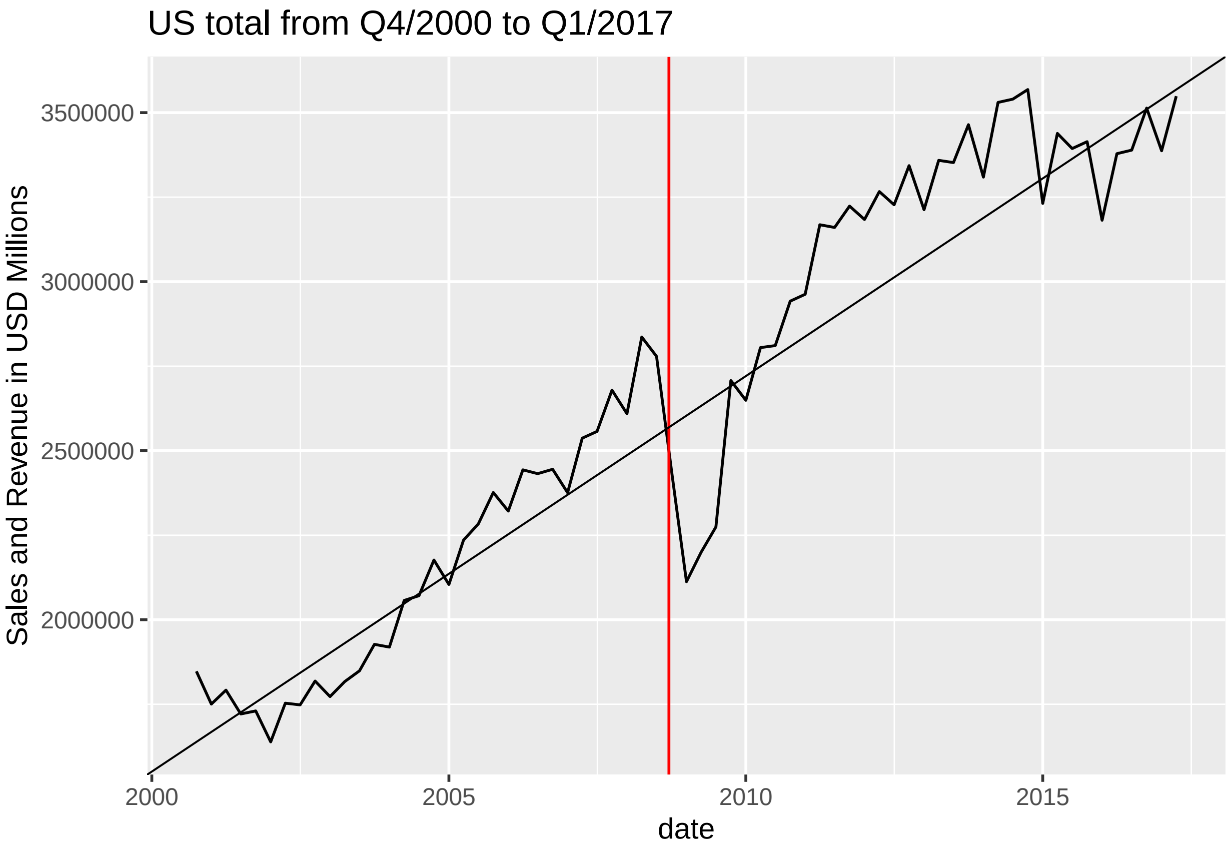
The analysis will also consider variations of revenue generation per industry, as these also exhibit geographic characteristics.

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.

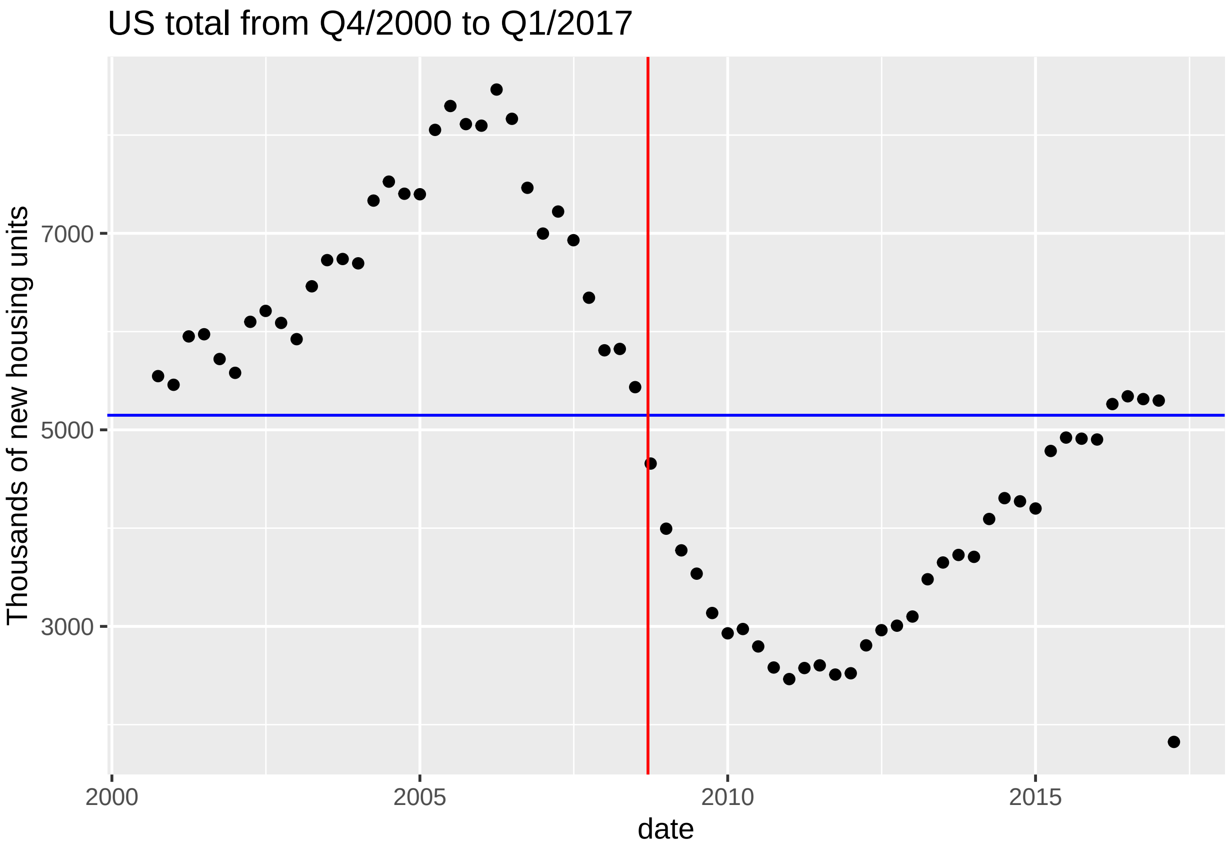
In order to understand what drives an increase and decrease of new housing, this analysis will categorise above average new housing offers as 1, and below average as 0, relative to each regional average.

The structure and sample data below shows the relevant metadata used in this study. For ease of reference, the script is commented as "Understanding and preparing data" in the Report.R file.

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

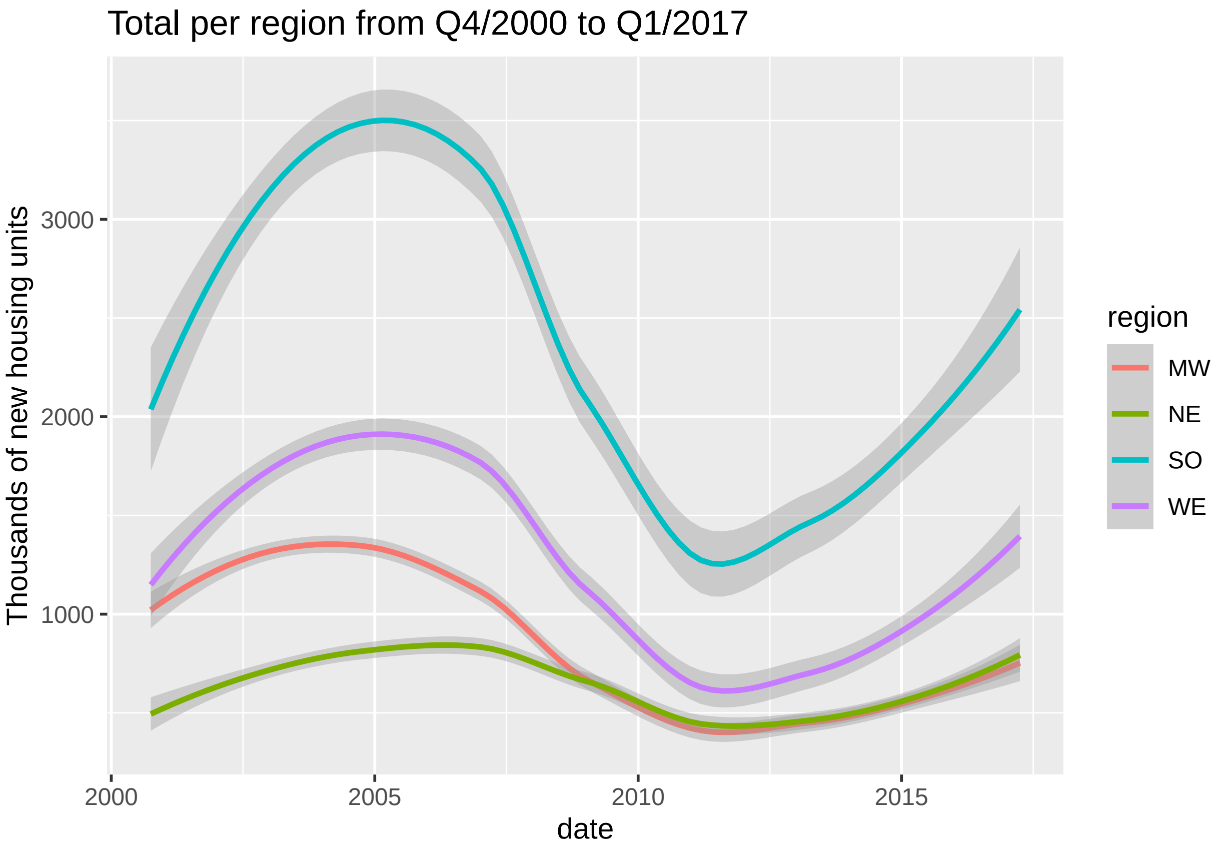


```
## ‘summarise()‘ regrouping output by ’region’ (override with ‘.groups‘ argument)
```



```
## ‘summarise()‘ regrouping output by ’region’ (override with ‘.groups‘ argument)
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



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.

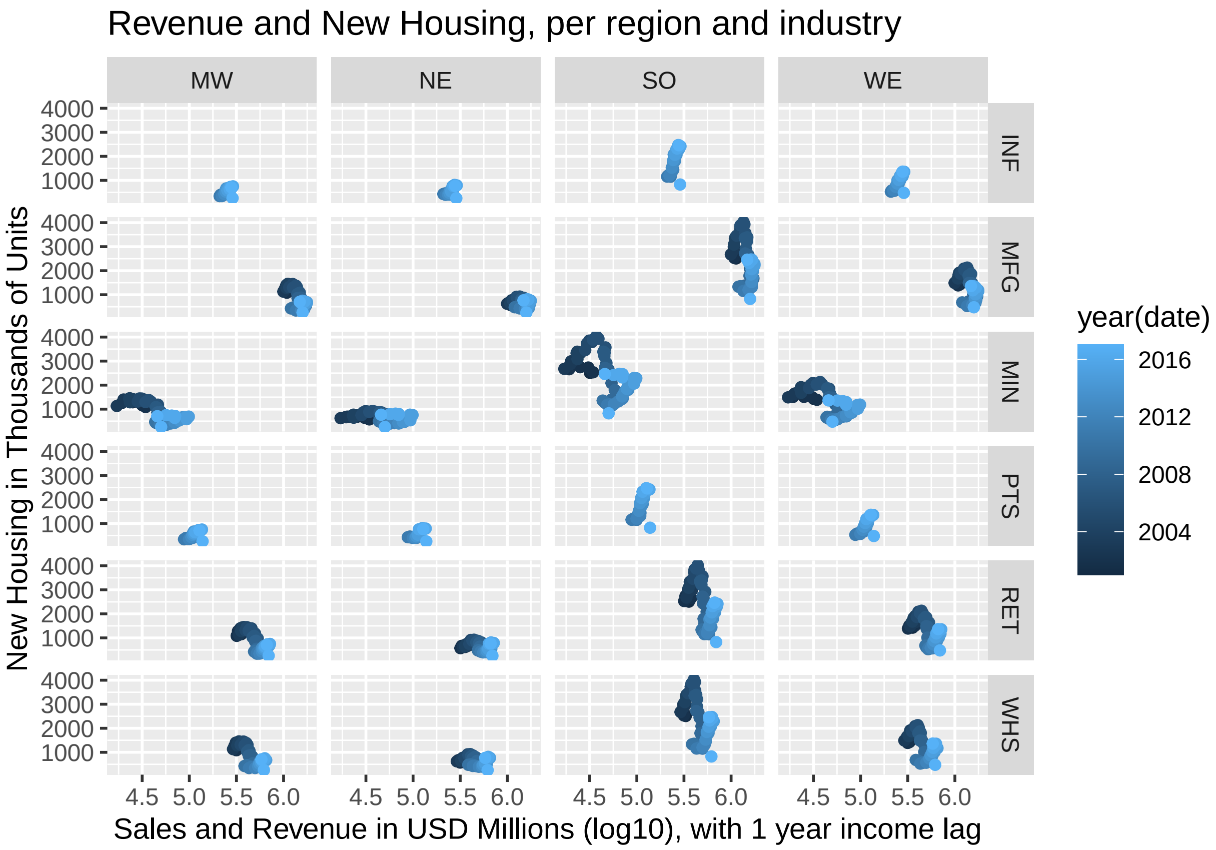
Industry codes are:

Code	Description
MFG	All Manufacturing
MIN	All Mining
RET	All Retail Trade
WHS	All Wholesale Trade
INF	All Information
PTS	All Professional and Technical Services, Except Legal Services

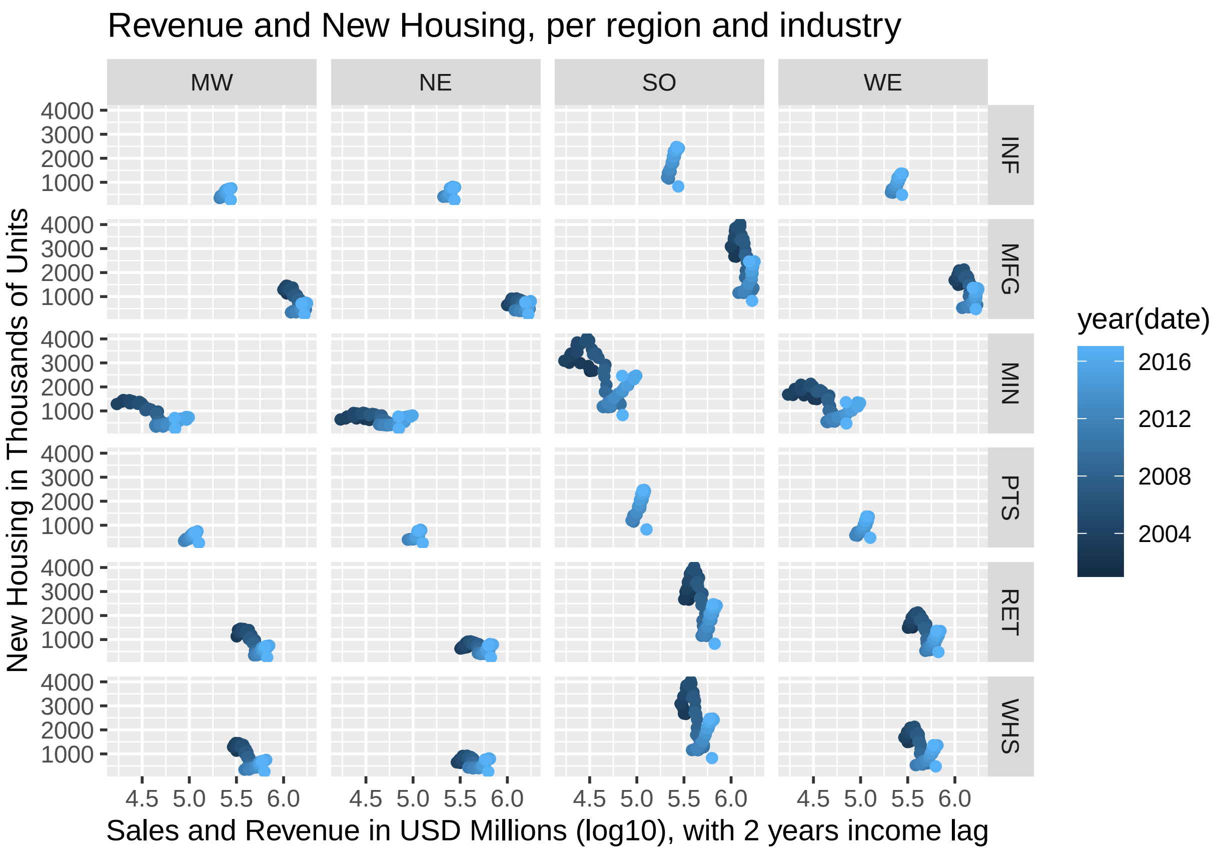
```
##    time_series_code       date    value industry     lag_1y     lag_2y     lag_3y
## 1       MFG_101_US 2000-10-01 1128790      MFG 2001-10-01 2002-10-01 2003-10-01
## 2       MIN_101_US 2000-10-01   31852      MIN 2001-10-01 2002-10-01 2003-10-01
## 3       RET_101_US 2000-10-01  356432     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

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

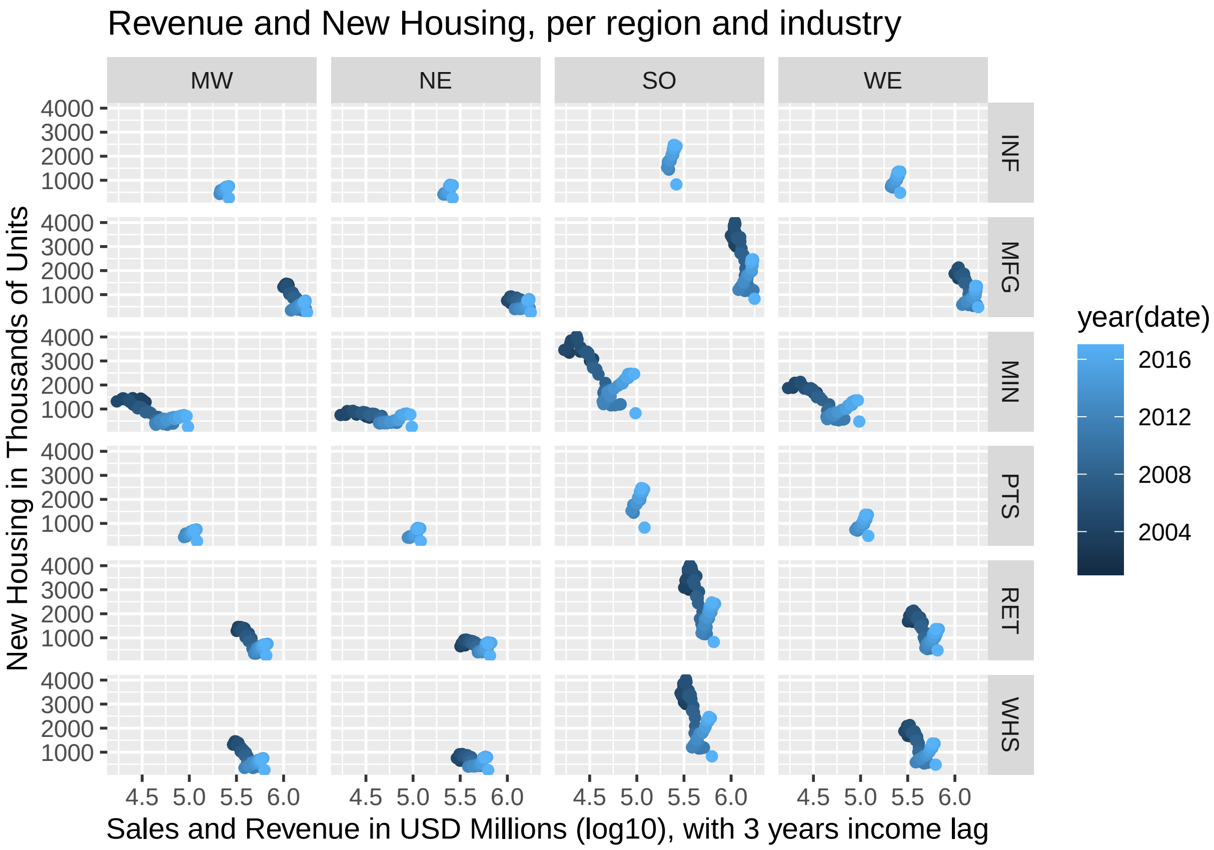
industry	sum1y	sum2y	sum3y	sum5y	sum8y	sum13y
INF	135	115	95	55	0	0
MFG	315	295	275	235	175	75
MIN	315	295	275	235	175	75
PTS	135	115	95	55	0	0
RET	315	295	275	235	175	75
WHS	315	295	275	235	175	75



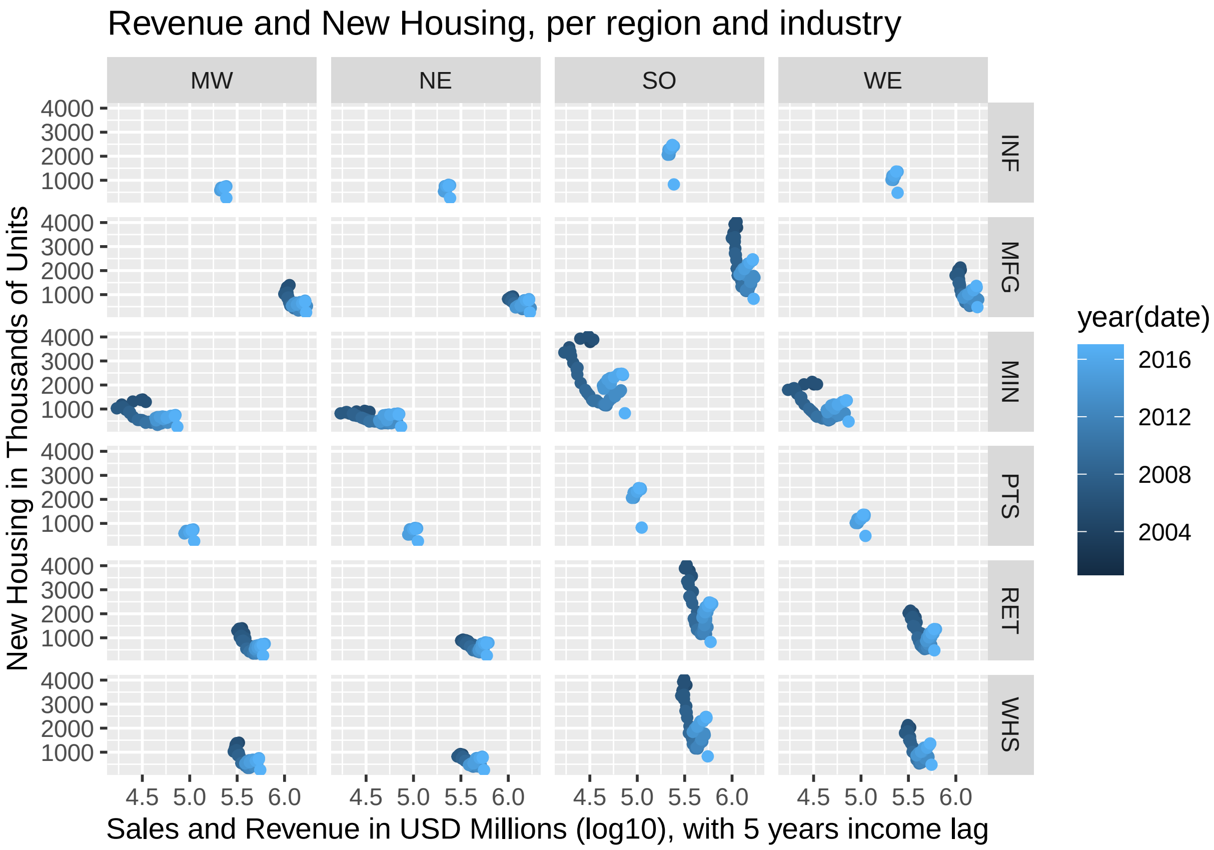
```
## Warning: Removed 96 rows containing missing values (geom_point).
```



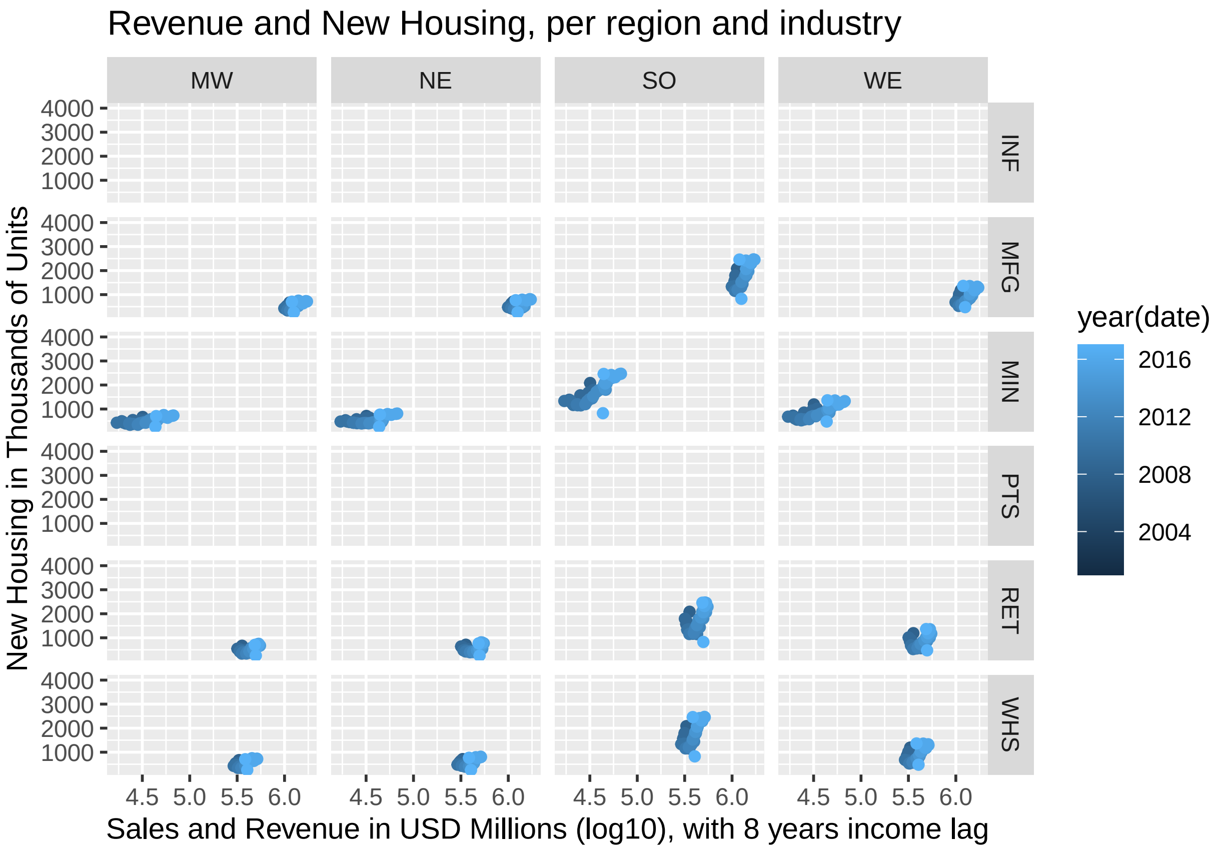
```
## Warning: Removed 192 rows containing missing values (geom_point).
```



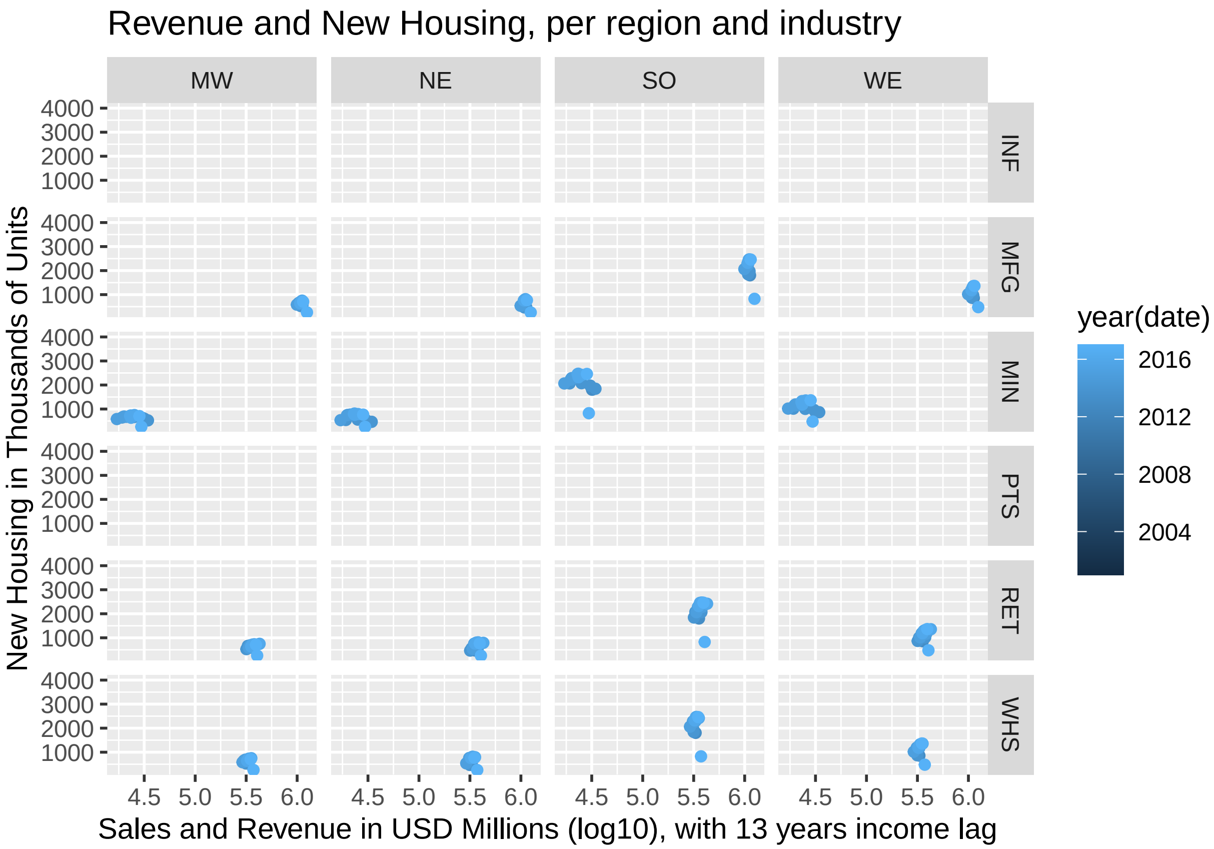
```
## Warning: Removed 384 rows containing missing values (geom_point).
```



```
## Warning: Removed 664 rows containing missing values (geom_point).
```



```
## Warning: Removed 984 rows containing missing values (geom_point).
```



```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
##         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.

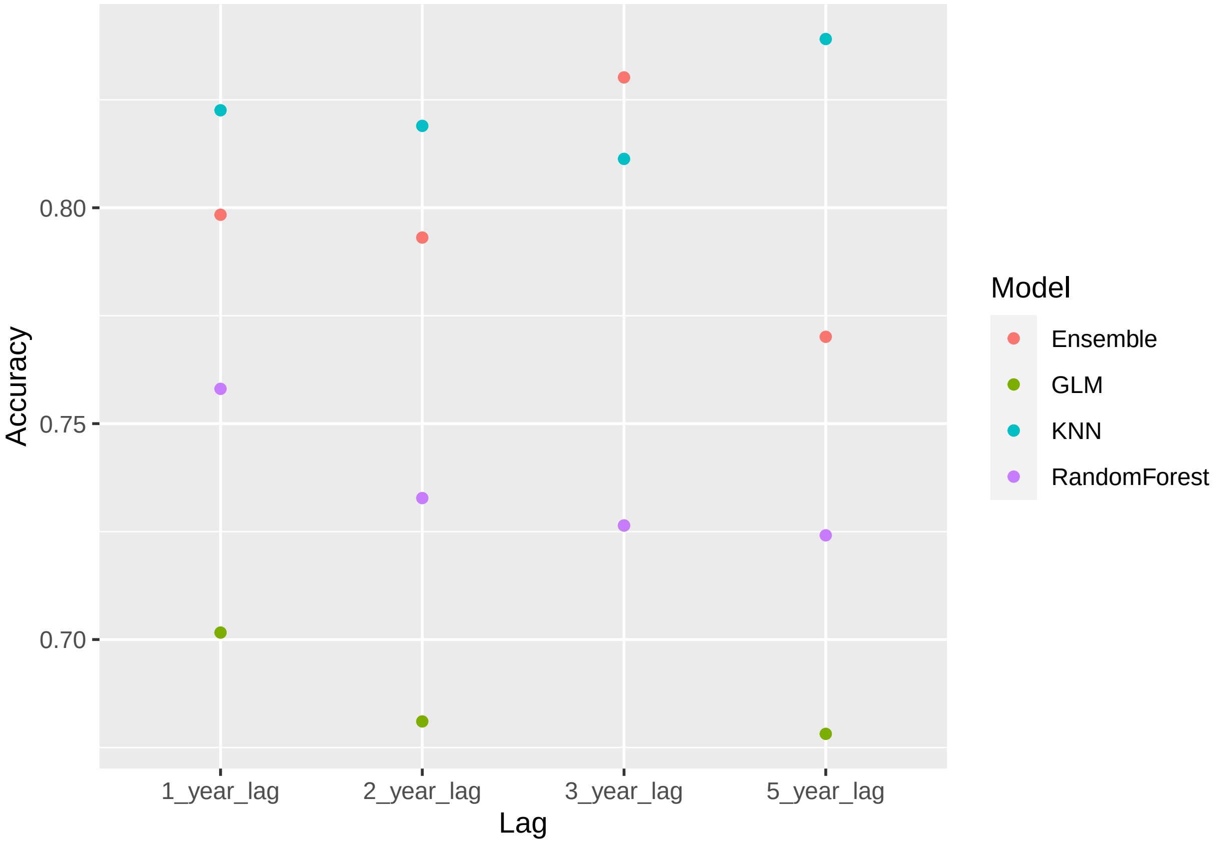
The models that appeared most appropriate for this exercise were GLM, KNN and Random Forest. The ensemble of those models did not improve on the prediction made via Random Forest.

3. Results

Optimal Time lag Comparison of predictive models Ensemble result

```
##       V1    V2    V3    V4
## 1: 0.702 0.681 0.726 0.678
## 2: 0.758 0.733 0.726 0.724
## 3: 0.823 0.819 0.811 0.839
## 4: 0.798 0.793 0.830 0.770
```

	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 x years later based on the economic outlook of a particular region in the United States. This may 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.

There is a reasonably high predictive power of house construction by US Region based on the past year financial performance of each industry. The maximum accuracy obtained with the Random Forest model was y.

As with any macroeconomic analysis, the interdependencies and correlations between indicators are complex and were not fully explored in this report. However, 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 and specific cyclical nature of each industry.