Data Analytics and Dashboarding

Faisal Jina

22 February 2021

Contents

1 Dashboard

The dashboard to accompany this report can be found at: https://faisaljina.shinyapps.io/da taanalyticsdashboard/

2 Background

This project will examine the trends of the housing market in Britain. The dataset used will be the most recent iteration of the UK House Price Index data. This dataset is released monthly as a CSV file, available from the gov.uk website. This is read in to R and the structure examined.

3 Exploratory Analysis

```
## Rows: 131,222
## Columns: 54
## $ Date                   <chr> "01/01/2004", "01/02/2004", "01/03/2004", "0...
## $ RegionName             <chr> "Aberdeenshire", "Aberdeenshire", "Aberdeens...
## $ AreaCode               <chr> "S12000034", "S12000034", "S12000034", "S120...
## $ AveragePrice           <dbl> 81693.67, 81678.76, 83525.10, 84333.68, 8637...
## $ Index                  <dbl> 40.86421, 40.85676, 41.78032, 42.18478, 43.2...
## $ IndexSA                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ X1m.Change             <dbl> NA, -0.01824784, 2.26048321, 0.96807069, 2.4...
## $ X12m.Change            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ AveragePriceSA         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ SalesVolume            <int> 388, 326, 453, 571, 502, 525, 652, 512, 497,...
## $ DetachedPrice          <dbl> 122490.1, 121280.9, 123395.4, 122334.0, 1244...
## $ DetachedIndex          <dbl> 43.61098, 43.18047, 43.93332, 43.55543, 44.3...
## $ Detached1m.Change      <dbl> NA, -0.9871659, 1.7435088, -0.8601624, 1.769...
## $ Detached12m.Change     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ SemiDetachedPrice      <dbl> 70563.16, 70804.42, 72689.07, 74484.23, 7663...
## $ SemiDetachedIndex      <dbl> 40.82189, 40.96146, 42.05176, 43.09029, 44.3...
## $ SemiDetached1m.Change  <dbl> NA, 0.3419153, 2.6617665, 2.4696403, 2.89121...
## $ SemiDetached12m.Change <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ TerracedPrice          <dbl> 55319.63, 55720.08, 57362.86, 59193.40, 6120...
## $ TerracedIndex          <dbl> 38.30567, 38.58295, 39.72049, 40.98803, 42.3...
## $ Terraced1m.Change      <dbl> NA, 0.7238695, 2.9482802, 3.1911554, 3.39463...
## $ Terraced12m.Change     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FlatPrice              <dbl> 48016.07, 49030.18, 50349.45, 51736.22, 5323...
## $ FlatIndex              <dbl> 42.43355, 43.32975, 44.49564, 45.72118, 47.0...
## $ Flat1m.Change          <dbl> NA, 2.1120161, 2.6907230, 2.7543015, 2.88773...
## $ Flat12m.Change         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ CashPrice              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ CashIndex              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
```

```
## $ Cash1m.Change          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ Cash12m.Change         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ CashSalesVolume        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ MortgagePrice          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ MortgageIndex          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ Mortgage1m.Change      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ Mortgage12m.Change     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ MortgageSalesVolume    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FTBPrice               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FTBIndex               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FTB1m.Change           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FTB12m.Change          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FOOPrice               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FOOIndex               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FOO1m.Change           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ FOO12m.Change          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ NewPrice               <dbl> 88436.14, 88606.45, 90296.91, 90319.88, 9198...
## $ NewIndex               <dbl> 40.26725, 40.34479, 41.11451, 41.12496, 41.8...
## $ New1m.Change           <dbl> NA, 0.19257621, 1.90783778, 0.02543242, 1.84...
## $ New12m.Change          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ NewSalesVolume         <int> 103, 107, 140, 180, 167, 164, 163, 130, 142,...
## $ OldPrice               <dbl> 81043.95, 80965.30, 82903.24, 84003.99, 8622...
## $ OldIndex               <dbl> 40.88337, 40.84369, 41.82130, 42.37659, 43.4...
## $ Old1m.Change           <dbl> NA, -0.0970528, 2.3935490, 1.3277553, 2.6412...
## $ Old12m.Change          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ OldSalesVolume         <int> 285, 219, 313, 391, 335, 361, 489, 382, 355,...
```

Fig. 1 - Full Dataset structure

Figure 1 shows the data that is recorded on house prices across the UK. This is evidently a large dataset that will need reducing for the purpose of a pragmatic analysis. Looking at the head and tail of this data, it can be seen that a new data row is available each month for each region, with data running from 1968 to 2020.

4 Aims of Analysis

This report will look at the current trends in house prices of the 3 countries of Great Britain. A 10-year period is selected to allow for establishing an idea of longer-term trends, which can be further filtered as required for short-term trends. Columns of interest selected include Average Price and Region data, as well as the 12-month Percentage Change column. This will help with like-for-like comparison between regions.

```
##         Date    RegionName  AreaCode AveragePrice X12m.Change SalesVolume
## 1 2010-12-01 Aberdeenshire S12000034     179526.1   2.8978449         335
## 2 2011-01-01 Aberdeenshire S12000034     174067.8   0.6663602         214
## 3 2011-02-01 Aberdeenshire S12000034     174751.1   0.8346159         189
## 4 2011-03-01 Aberdeenshire S12000034     174139.6   2.1862236         329
## 5 2011-04-01 Aberdeenshire S12000034     178523.2   3.2983790         401
## 6 2011-05-01 Aberdeenshire S12000034     181756.7   3.3686994         363
```

Fig. 2 - First rows of data

```
##             Date               RegionName  AreaCode AveragePrice X12m.Change
## 51235 2020-06-01 Yorkshire and The Humber E12000003     168183.9    2.765173
## 51236 2020-07-01 Yorkshire and The Humber E12000003     169321.7    2.625537
## 51237 2020-08-01 Yorkshire and The Humber E12000003     172339.2    4.102662
## 51238 2020-09-01 Yorkshire and The Humber E12000003     173057.1    4.600707
## 51239 2020-10-01 Yorkshire and The Humber E12000003     177412.5    6.729787
## 51240 2020-11-01 Yorkshire and The Humber E12000003     180855.5    9.691853
##       SalesVolume
## 51235        4654
## 51236        5011
## 51237        5146
## 51238        5594
## 51239          NA
## 51240          NA
```

Fig. 3 - Last rows of data

Checking the head and tail of the filtered data, it now appears to be uniform with a manageable and relevant structure from which to continue the analysis. As the focus is on the 3 countries of Great Britain, these 3 regions as well as Great Britain are extracted from the the data. This leaves us with a dataset of 480 rows x 4 columns.

```
## Rows: 480
## Columns: 4
## $ Date          <date> 2010-12-01, 2011-01-01, 2011-02-01, 2011-03-01, 2011...
## $ RegionName    <fct> England, England, England, England, England, England,...
## $ AveragePrice  <dbl> 176035.9, 174442.3, 173810.6, 173045.6, 175490.5, 174...
## $ `%12m.Change` <dbl> 1.090987796, -0.008770055, -0.819974837, -0.983825702...
```

Fig. 4 - Glimpse of the filtered data

To observe the changes in house prices over time, it is useful to view this data diagrammatically.

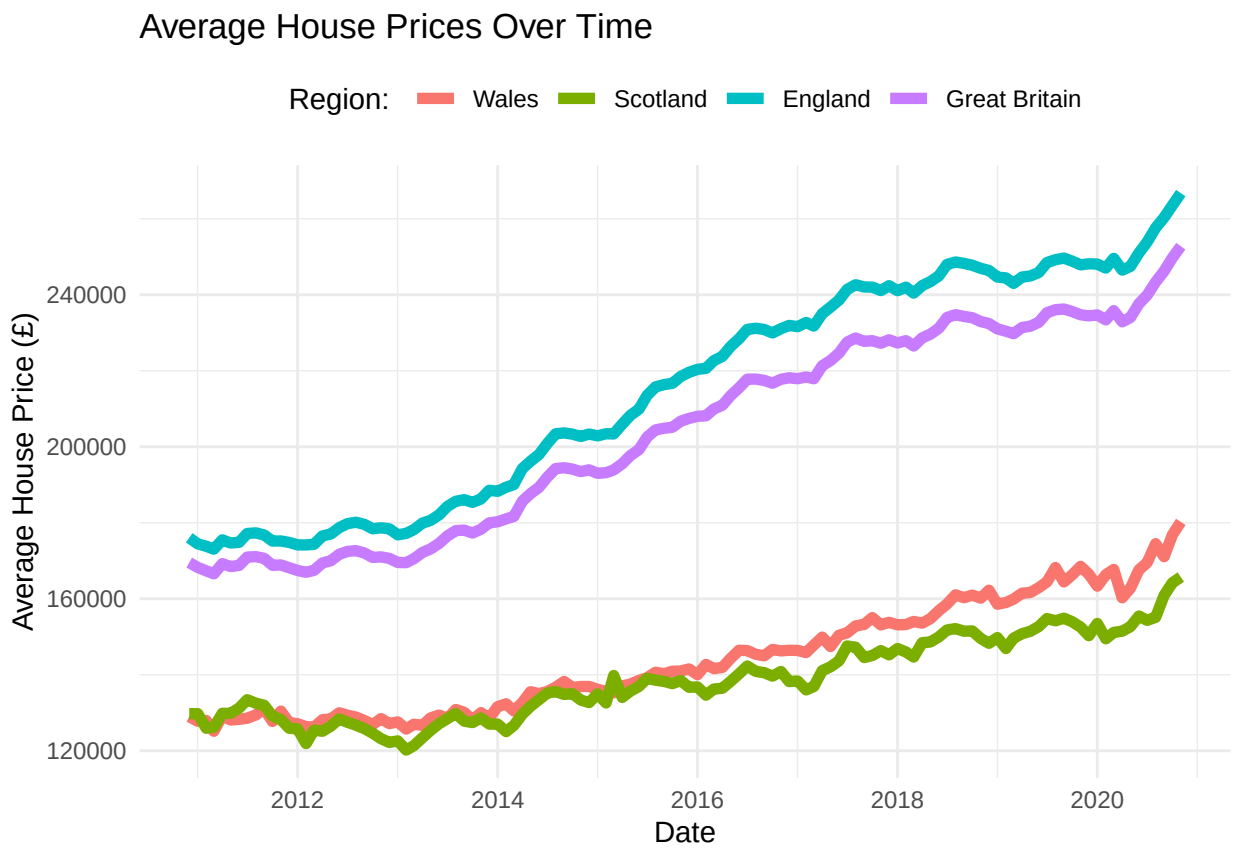

Fig. 5 - Regional House Price Graph

Figure 5 gives an idea of what the data looks like in each region of interest. In particular, it shows that the 3 countries largely follow the general trend shown by Great Britain. This is examined in more detail in section 5.

5 Long-term Trend

For long-term modelling, it is useful for businesses to know the general trend of the national housing market to inform pricing forecasts where real estate is involved. Whilst the average national house price appears to be largely linear with respect to time, a simple linear regression between these variables can help to determine if a linear model does indeed represent house prices well over the long-term.

5.1 Linear Regression Hypotheses

Null hypothesis: There is not a linear relationship between the average house price and time
Alternative hypothesis: There is a linear relationship between the average house price and time

```
##
## Call:
## lm(formula = AveragePrice ~ Date, data = dfGB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10542.4  -3716.3      2.8   4336.7  10125.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.090e+05  7.538e+03  -27.73   <2e-16 ***
## Date         2.465e+01  4.490e-01   54.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5186 on 118 degrees of freedom
## Multiple R-squared:  0.9623, Adjusted R-squared:  0.962
## F-statistic:  3014 on 1 and 118 DF,  p-value: < 2.2e-16
```

Fig. 6 - Simple Linear Model

5.2 Results

The linear regression model and coefficients were all significant at the 95% level ($p < 0.05$). The null hypothesis is therefore rejected and the alternative hypothesis accepted, and this confirms with a high degree of confidence that a linear relationship exists between the average house price and time. The adjusted R-squared value shows that 96% of the variance in the average house price is captured by this simple linear model based on this data, which is a very high result.

5.3 Outcome

Based on this linear trend, one could reasonably assume persistence of this trend for forecasting purposes. Whilst there is volatility around this trendline, this is a long-term trend with a high R-squared, so there is high degree of confidence in extrapolating this going forward. A company using forecasting models on house prices would typically make HPI calculations at least every quarter, so it is suggested that this trendline should be recalculated at the same time to look for changes over time, and update forecasting models as appropriate. The linear model is displayed in the graph below (also available in the dashboard).

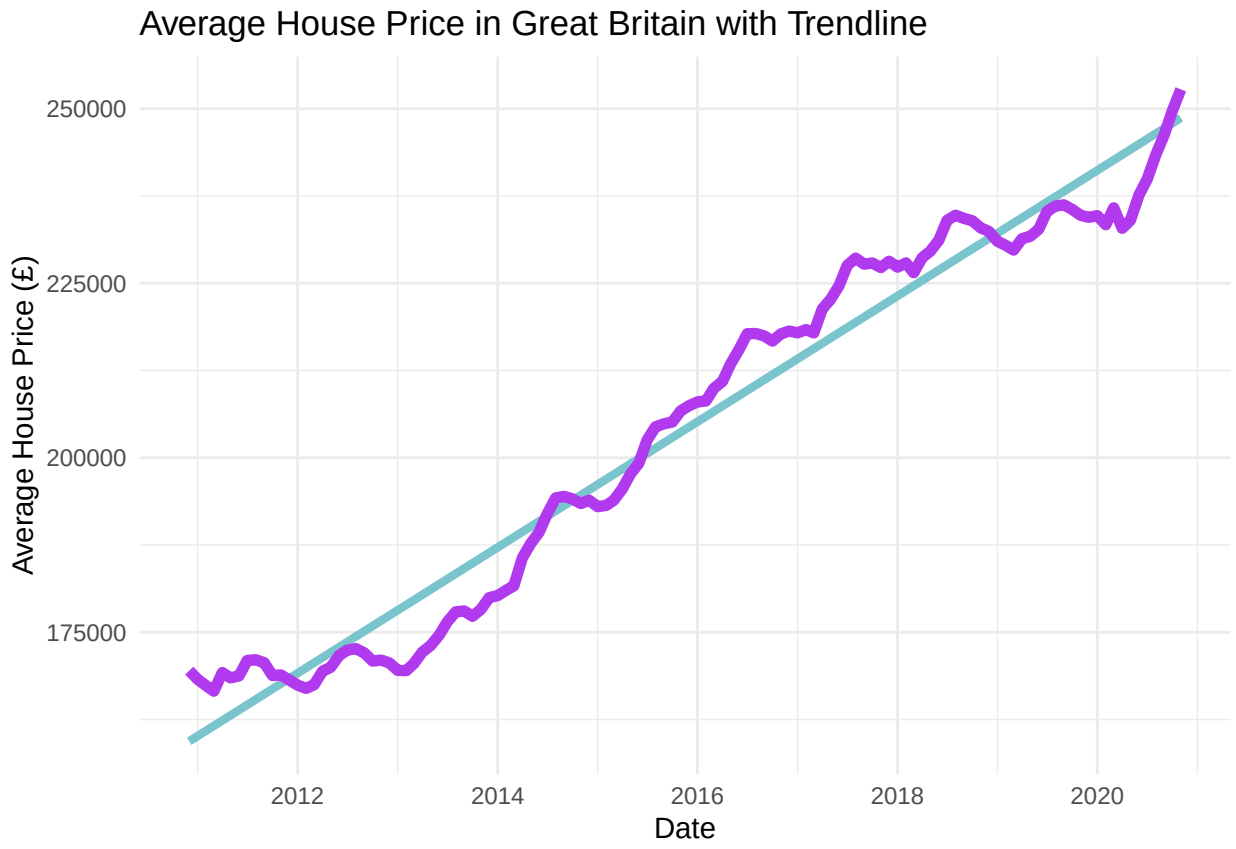

Fig. 7 - Graph of British House Price Trend

6 Regional Variation

Whilst house prices in Great Britain were linear over the longer term, the regional graph shows that the 3 countries may exhibit slightly different trends in the shorter term, particularly more recently. Boxplots are plotted to help identify any trend deviations in the 12-monthly price changes.

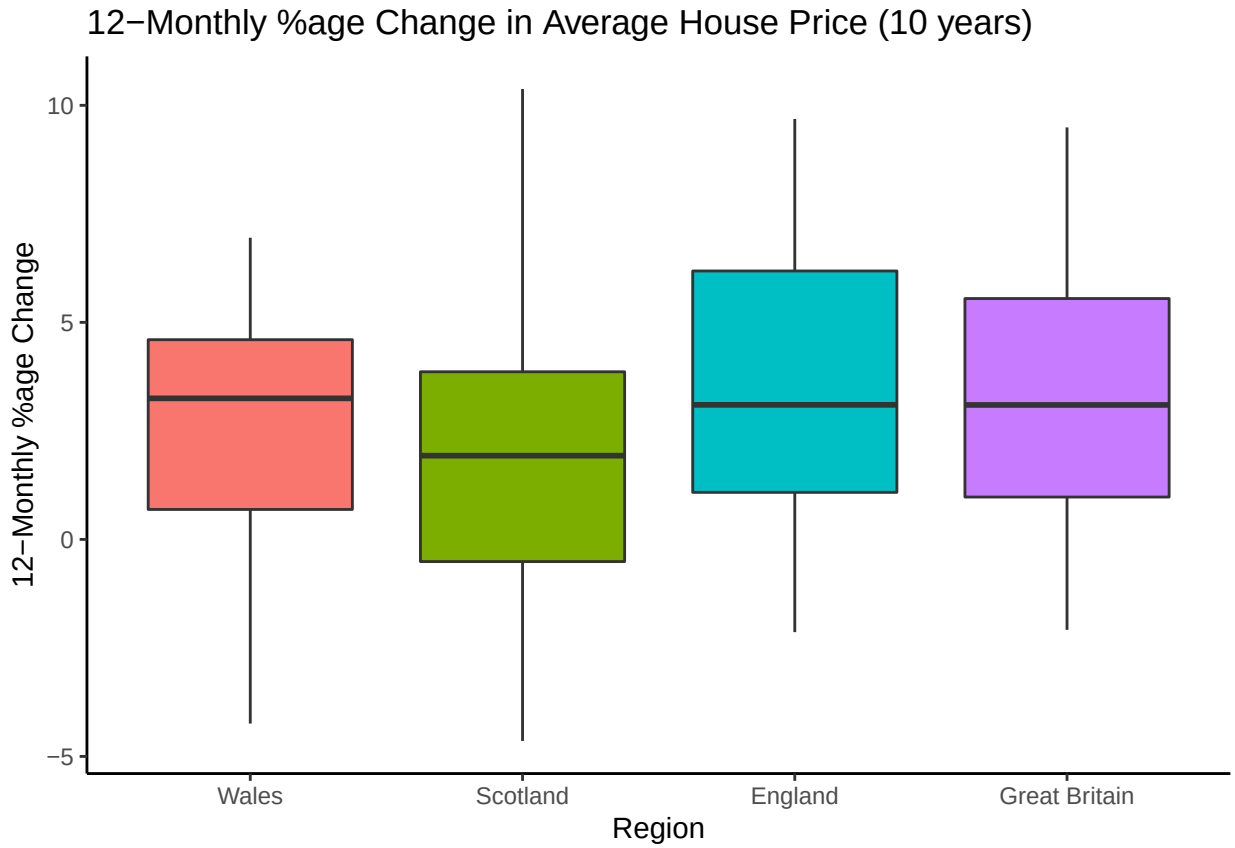

Fig. 8 - Regional Boxplots of Average House Price - 10 Year

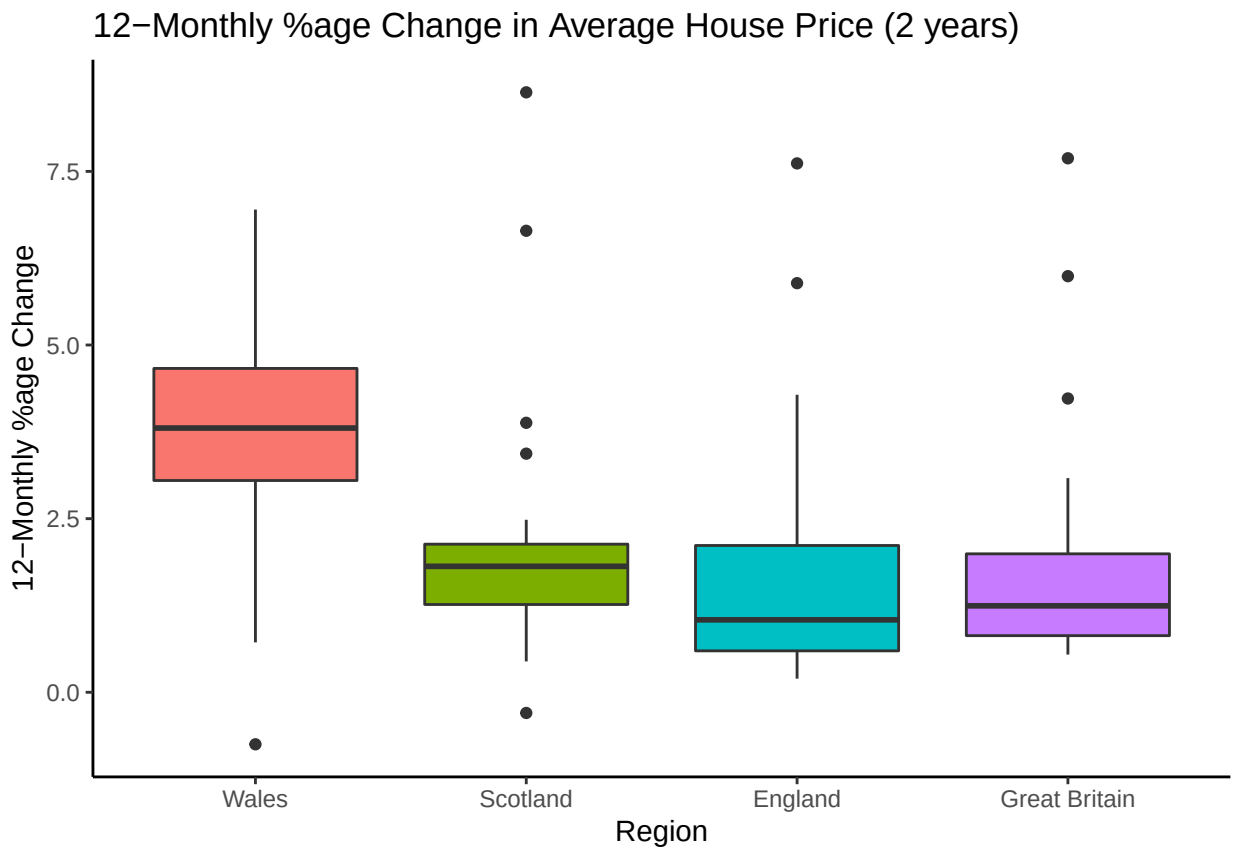

Fig. 9 - Regional Boxplots of Average House Price - 2 Year

As evidenced by the long-term regional graph (Fig.5) and 10-year boxplot (Fig. 8), all 3 countries appear to show a similar pattern of average house price change over the long term, with England and Great Britain appearing to be most similar. However, from the boxplots it is evident that as the time period is narrowed to a more recent subset e.g. the past 2 years (Fig. 9), it appears that Wales increasingly stands out as potentially having greater average house price changes than the other regions. (Boxplots also available on dashboard - select a region of the graph to see the corresponding boxplots).
It may be useful to examine Wales and England to determine if the difference observed is significant - if so, these markets may need to be treated differently in financial modelling.

6.1 Short-term Trend

As the housing stock in Wales and England may be different, these regions are treated as independent groups. A 3-year subset of the 12-month changes is taken for these regions - this ensures the period is small enough to be relevant, but large enough to have enough data to draw reasonably reliable conclusions.

6.2 Normality Test

Firstly, a Shapiro-Wilk test is made to see if the differences between these groups is normal, to determine the testing going forward.

Null hypothesis: The distribution is normally distributed.
Alternative hypothesis: The distribution is not normally distributed.

```
##
##  Shapiro-Wilk normality test
##
## data:  diff$`%12m.Change`
## W = 0.98124, p-value = 0.7865
```

Fig. 10 - Test of Normality on 'Wales minus England' data

The Shapiro-Wilk p-value » 0.05, so the null hypothesis is accepted, which indicates that the distribution of the difference between these groups is not significantly different from the normal. We therefore assume normality, and a two sample t-test can be run.

6.3 T-test

The two-sample t-test examines the difference in means of the 12-monthly percentage change of average house prices between Wales and England. This is run as unpaired, as these groups are assumed to be independent, and a Welch test is used as the variance in these groups may be different.

Null hypothesis: There is no difference between the means of the two groups.
Alternative hypothesis: There is a difference between the means of the two groups.

```
##
##  Welch Two Sample t-test
##
## data:  Wal3 and Eng3
## t = 4.9388, df = 67.993, p-value = 5.386e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.089308 2.566344
## sample estimates:
## mean of x mean of y
##  4.037337  2.209511
```

Fig. 11 - Two-sample Unpaired t-test

The t-test p-value « 0.05, indicating that we should reject the null hypothesis and accept the alternative hypothesis - that the difference in means between these groups is significant at the 95% level. The mean of the Wales group was also calculated as being higher than the England group.

6.4 Outcome

The result of the t-test informs us that over the past 3 years, the average 12-month percentage change in Wales has been greater than that in England. Whilst this was not apparent in the Average House Price graph, this may be due to the absolute difference in house price between the regions. This disparity can be resolved by indexing these regions both to an arbitrary value of 100 at a point 4 years ago (4 years chosen as previous tests were on a 3 year sample using a 12-month price change).

6.4.1 Indexed House Prices - Wales vs England

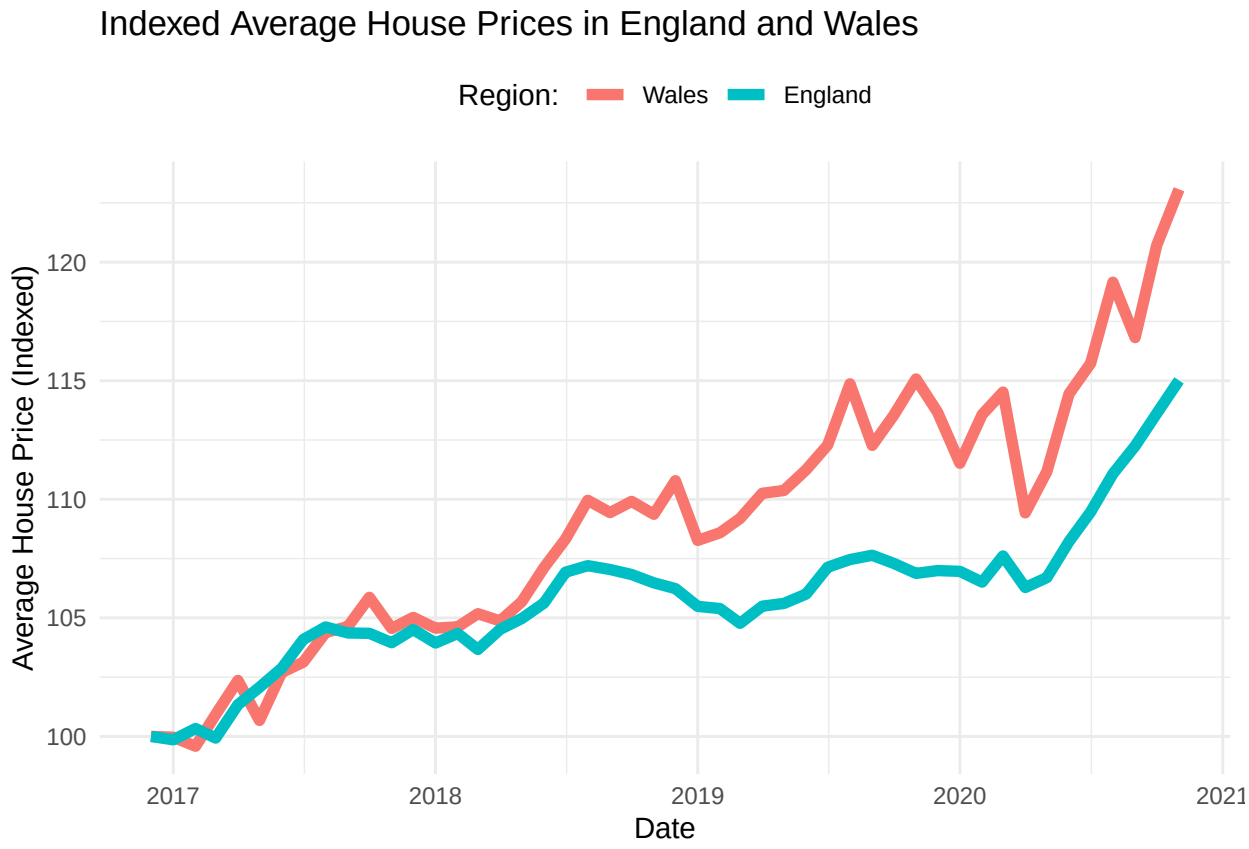

Fig. 12 - Graph of Indexed House Prices - Wales and England

Figure 12 (also available on dashboard) shows that Wales' average house prices have indeed risen faster (in percentage terms) than those in England over this time period, and the tests show that this difference was significant.

The effect of this finding for businesses is that shorter-term growth in these 2 markets can not each be assumed to be uniform across Britain. The differences between the markets of Wales and England may need to be modelled separately, which in turn will impact forecasts and risk profiles of any assets/liabilities linked to real estate in these regions. Again, this finding should be monitored over time to see if the discrepancy between these regions persists into the future. The skewing of the housing market in this way may present an opportunity for businesses to take advantage of increased demand in Wales with the view of a greater

increase in property values over time relative to England.

7 References

1. 'flexdashboard: Easy interactive dashboards for R'. Flexdashboard. Available at: https://rmarkdown.rstudio.com/flexdashboard/. Accessed: 20 Feb 2021.
2. 'ggplot2 Brushing'. JJ Allaire. Available at: https://jjallaire.shinyapps.io/shiny-ggplot2-brushing/. Accessed: 21 Feb 2021.
3. 'ggplot2 Quick Reference: colour (and fill)'. Sape Research Group. Available at: http://sape.inf.usi.ch/quick-reference/ggplot2/colour. Accessed: 18 Feb 2021.
4. 'Markdown Basics'. RStudio. Available at: https://rmarkdown.rstudio.com/authoring_basics.html. Accessed: 22 Feb 2021
5. 'Paired vs Unpaired T-Test: Differences, Assumptions and Hypotheses'. Nicole Gleichmann. 14 Feb 2020. Available at: https://www.technologynetworks.com/informatics/articles/paired-vs-unpaired-t-test-differences-assumptions-and-hypotheses-330826. Accessed: 17 Feb 2021.
6. 'Smoothed conditional means'. ggplot2. Available at: https://ggplot2.tidyverse.org/reference/geom_smooth.html. Accessed: 16 Feb 2021.
7. 'Statistical tools for high-throughput data analysis'. STHDA. Available at: http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r. Accessed: 16 Feb 2021.
8. 'Themes'. ggplot2. Available at: https://ggplot2-book.org/polishing.html. Accessed: 16 Feb 2021.
9. 'UK House Price Index: data downloads November 2020'. Gov.uk. 20 Jan 2021. Available at: https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-datadownloads-november-2020. Accessed: 15 Feb 2021.