Real Estate Prices Comparative Analysis
Ames, Iowa Neighborhoods: 2007-2010

Aiden Dickson, Khoa Pham

2022-07-06

I. INTRODUCTION:

This is a data set of 2,930 Real Estate properties sold between years 2006 and 2010 in Ames, Iowa. Each observation contains descriptive and quantitative property features including "Sale Price", "Number of Bedrooms", "Number of Full Bathrooms", "Neighborhood" and more, including 82 in total. This data contains information from the Ames Assessor's Office and was compiled by Dean De Cock for use in real estate prediction. More information about the data set description can be found here: https://www.openintro.org/data/index.php?data=ames

Research Question:

How has the cost of housing changed over time in Ames, Iowa? How have the differences fluctuated between particular neighborhoods? In pursuing the answer to these questions, we have started by comparing the mean real estate sale prices for specific neighborhoods of Ames, Iowa, for a timeframe chosen between the years 2007 - 2010 before 'zoning in' on a few neighborhoods of interest.

We do this by exploring price mean differences between these neighborhoods and within their own trajectory over time. Initially we put limitations on some factors, such as comparing only those with <= 2 Bedrooms and <= 1 Bath. We eventually found more practicality in removing these constraints, in order to have a demonstration that would meet ANOVA assumptions. Our research questions are important to ask not only if one happens to have a stake in Ames, Iowa, but because there is an opportunity in finding the answer, to observe real estate price behavior that are universal to all markets.

3 Rows of Dataset Preview:

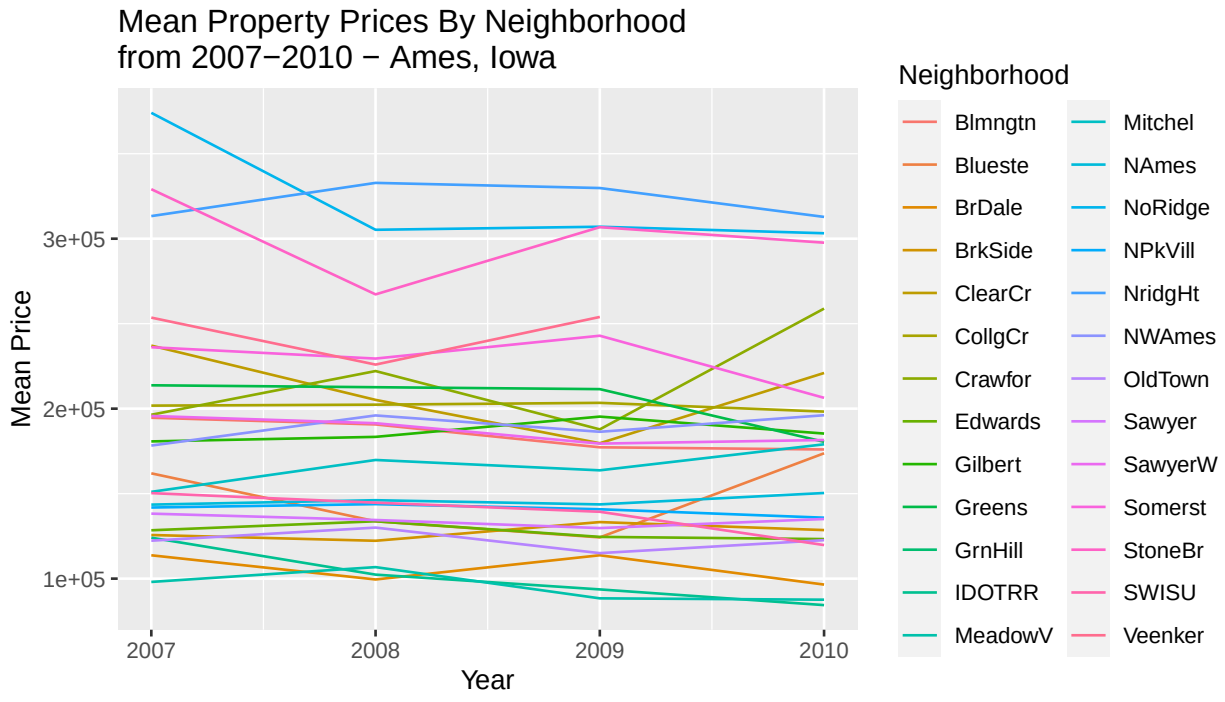
```
## # A tibble: 3 x 82
##   Order      PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street
##   <dbl>    <dbl> <dbl>  <dbl>       <dbl> <chr>            <dbl>    <dbl> <chr>
## 1     1   5.26e8  1656 215000          20 RL                 141    31770 Pave
## 2     2   5.26e8   896 105000          20 RH                  80    11622 Pave
## 3     3   5.26e8  1329 172000          20 RL                  81    14267 Pave
## # ... with 73 more variables: Alley <chr>, Lot.Shape <chr>, Land.Contour <chr>,
## #   Utilities <chr>, Lot.Config <chr>, Land.Slope <chr>, Neighborhood <chr>,
## #   Condition.1 <chr>, Condition.2 <chr>, Bldg.Type <chr>, House.Style <chr>,
## #   Overall.Qual <dbl>, Overall.Cond <dbl>, Year.Built <dbl>,
## #   Year.Remod.Add <dbl>, Roof.Style <chr>, Roof.Matl <chr>,
## #   Exterior.1st <chr>, Exterior.2nd <chr>, Mas.Vnr.Type <chr>,
## #   Mas.Vnr.Area <dbl>, Exter.Qual <chr>, Exter.Cond <chr>, ...
```

Number of Observations Per Neighborhood:

```
##
## Blmngtn Blueste  BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert  Greens
##      28      10      30     108      44     267     103     194     165       8
## GrnHill  IDOTRR Landmrk MeadowV Mitchel   NAmes NoRidge NPkVill NridgHt  NWAmes
##       2      93       1      37     114     443      71      23     166     131
## OldTown  Sawyer SawyerW Somerst StoneBr   SWISU  Timber Veenker
##     239     151     125     182      51      48      72      24
```

II. Time-Series Visual Analyses

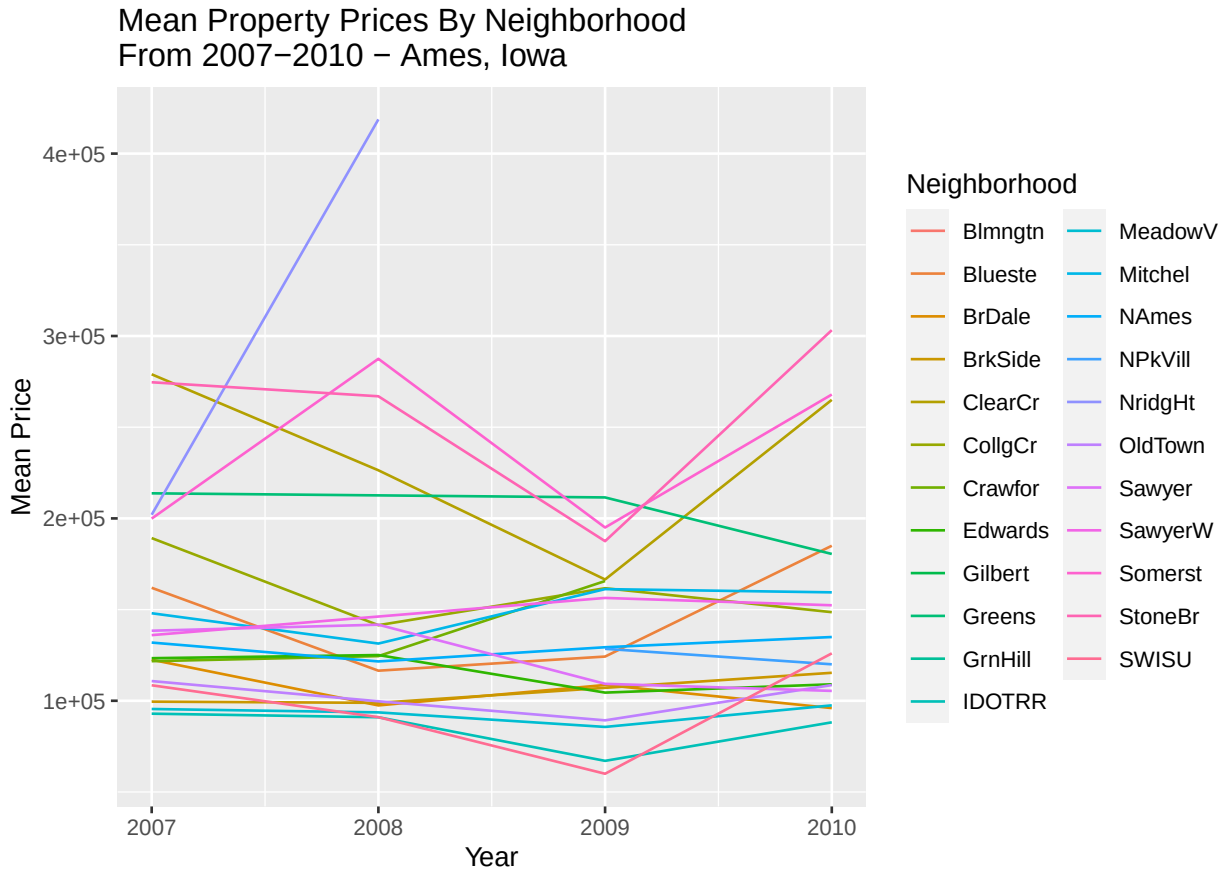
In the following 3 Time Series infographs, the fluctuation of the neighborhood means from year 2007-2010 are illustrated 'side-by-side' for all neighborhoods included in the Ames, Iowa dataset. This 1st time series graph is before filtering towards the neighborhoods of interest. This first one (below), is a rough overview of all, to visually target a few interesting neighborhoods, in contrast to the others.



In the 2nd time series graph we narrow our inquiry, by filtering the comparison to only houses with less than or equal to 2 bedrooms (above ground) and less than or equal to 1 full bath. Since the aim is to also see a trajectory from 2007-2010, for an optimally analogous comparison, we remove the truncated timeline of neighborhood "Veenker" (shorter red line in previous graph).

Code Lines Edited:

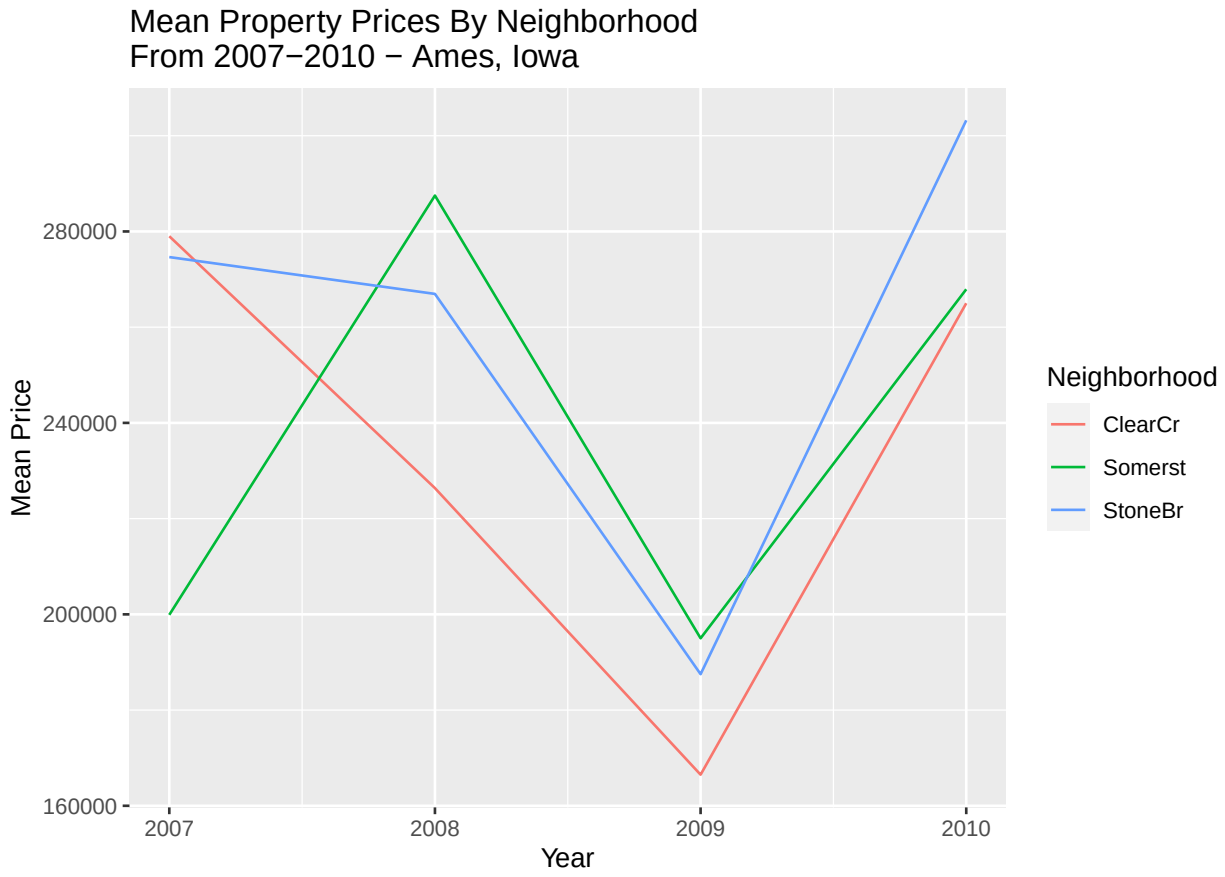
1. #Neighborhood == "Veenker") %>% Commented out.
2. filter(Bedroom.AbvGr <= "2") %>% Added in filters.
3. filter(Full.Bath <= "1") %>% Added in filters.



The 3rd time-series (below), has retained only 3/4 of the top priced neighborhoods, within the constraints of houses with <= 2 bedrooms (above ground) and <= 1 full bath in Ames, Iowa. I will remove the steepest one 'NridgHt' as it moves out of range in 2008. This allows us to take a closer look at potentially significant differences in mean as they dramatically fluctuate over time, controlled by a couple of factors.

Code Lines Edited: 1. Multi-Line - Neighborhood == <All but 3 depicted> Removed.

```
house_features_ames  %>%
  filter(
          Neighborhood == "Somerst" |
          Neighborhood == "StoneBr" |
          Neighborhood == "ClearCr" ) %>%
  filter(Bedroom.AbvGr <= "2") %>%
  filter(Full.Bath <= "1") %>%
  filter(Yr.Sold >= "2007") %>%
  group_by(Neighborhood, Yr.Sold) %>%
  summarise(mean = mean(price)) %>%
  ggplot(aes(x = Yr.Sold, y = mean)) +
  labs (title = str_wrap("Mean Property Prices By Neighborhood From
                         2007-2010 - Ames, Iowa", 40),
        x = "Year",
        y = "Mean Price") +
  geom_line(aes(color = Neighborhood))
```



III. ANOVA Assumptions Check, Adjustments, and Analyses of 3 Target Neighborhoods:

In this section we take a look at an "ANOVA" analysis, comparison of the means of the 3 target neighborhoods, to get a sense of the difference in means between theses in the overall time span of 2007- 2010, after checking Assumptions and making adjustments if necessary.

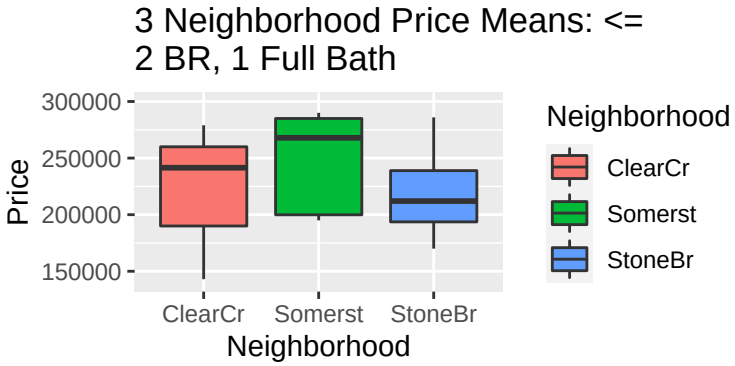
Table 1: Summary Statistics for 'Somerset', 'Stone Brook' and 'Clear Creek' Neighborhoods

Neighborhood	Mean Price	Standard Deviation of Price	Observations
ClearCr	227500.0	47560.49	9
Somerst	247563.2	46506.19	5
StoneBr	268970.1	99508.13	8

Prelim Summary:

```
house_features_ames  %>%
  filter(
          Neighborhood == "Somerst" |
          Neighborhood == "StoneBr" |
          Neighborhood == "ClearCr" ) %>%
  filter(Bedroom.AbvGr <= "2") %>%
  filter(Full.Bath <= "1") %>%
  filter(Yr.Sold >= "2007") %>%
  group_by(Neighborhood) %>%
  summarise(mean = mean(price),
            sd = sd(price),
            count = count(Neighborhood)) %>%
  kbl(caption = "Summary Statistics for 'Somerset', 'Stone Brook' and 'Clear
      Creek' Neighborhoods",
      col.names = c("Neighborhood",
                    "Mean Price",
                    "Standard Deviation of Price",
                    "Observations"),
      digits = 2) %>%
  kable_material(c("grey",
                   "hover"))
```

Box Plot Distribution:



ANOVA Assumptions Check:

Before conducting an ANOVA test to compare the mean Price for 3 neighborhoods in Ames between 2007 and 2010, we need to ensure the assumptions are met:

1. Random samples - we can assume that we are having random samples.
2. Normal population - we can see that the first two box plots and density plots pretty skewed to the left, and the last one slightly to the right. While ANOVA is quite robust against violations of normality, we will play it conservative and narrow the timerange to == 2010, and found it necessary as well to substitute these neighborhoods to "Somerset, Greens and Crawford" to ensure visual normalcy as well as below variance assumptions met for the sake of ANOVA. Additionally needing to remove the room number constraints for enough data.
   -NOT Met.-
3. Equal variance - there looks to be variance in IQR as well as some outliers in the box plot. max(SD)/min(SD) = 99508.13^2 / 46506.19^2 = 4.58 > 2, so we found it necessary to substitute these three neighborhoods with "Somerset, Greens and Crawford", to ensure this assumption met as well. Additionally needing to remove the room number constraints for enough data.
   -NOT Met.-
4. Independence within and between groups - We will check the independence assumption once we have fitted a model to our data.

Prelim Summary - 2nd Trial - ANOVA Assumptions Now Met:

We have changed 2/3 of the neighborhoods (this code correlates to Summary Table 2), removed the room number constraints, and limited the year to 2010 to meet the assumptions for variance required to do ANOVA. Our Variance Assumption check is now $\max(SD)^{2/\min(SD)}2$ = 44647.31^2 / 36062.45^2 = 1.53 <2. So the Variance Assumption is met for ANOVA.

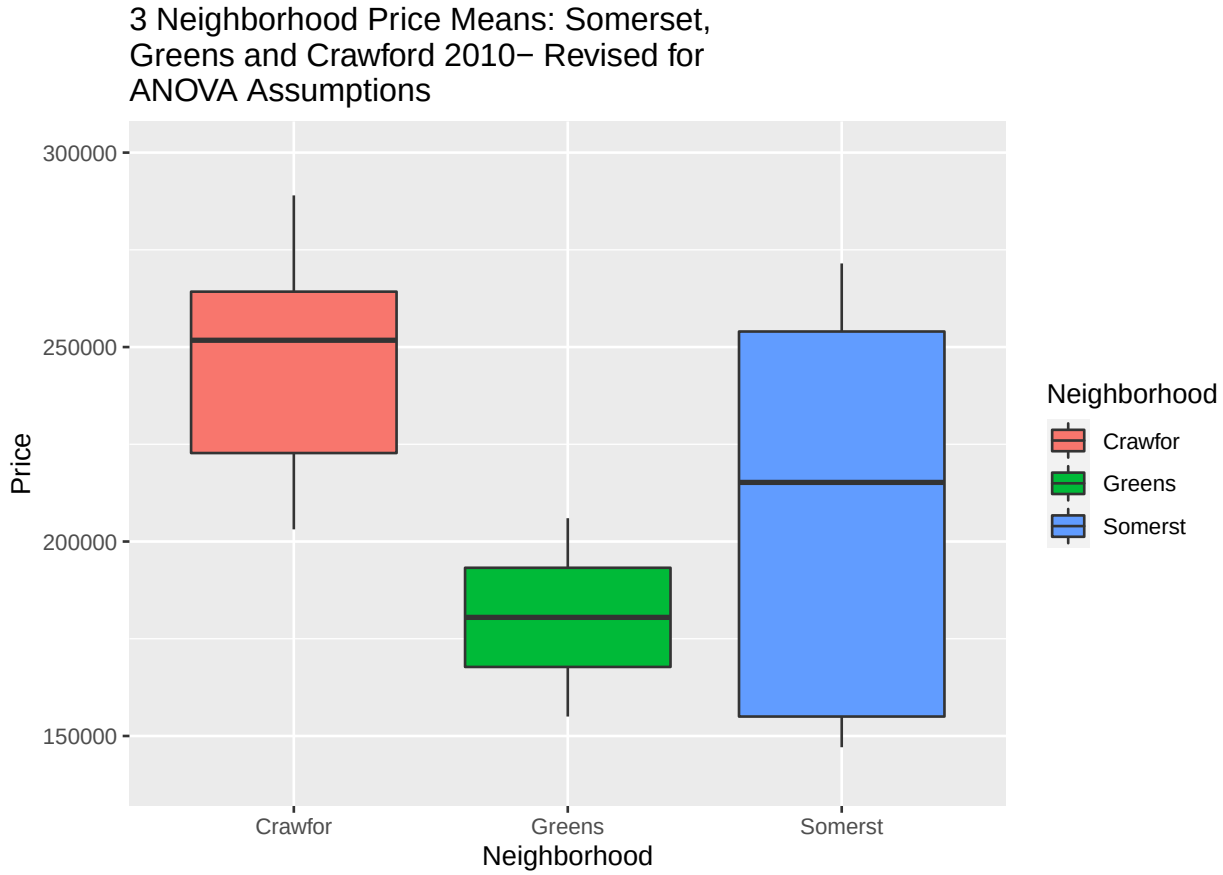
```
house_features_ames  %>%
  filter(
          Neighborhood == "Somerst" |
          Neighborhood == "Greens" |
          Neighborhood == "Crawfor" ) %>%
  filter(Yr.Sold == "2010") %>%
  group_by(Neighborhood) %>%
  summarise(mean = mean(price),
            sd = sd(price),
            count = count(Neighborhood)) %>%
  kbl(caption =  str_wrap("3 Neighborhood Statistical Summaries: Somerset,
                          Greens and Crawford 2010 - Revised for ANOVA", 20),
      col.names = c("Neighborhood",
                    "Mean Price",
                    "Standard Deviation of Price",
                    "Observations"),
      digits = 2) %>%
  kable_material(c("grey",
                   "hover"))
```

Table 2: 3 Neighborhood Statistical Summaries: Somerset, Greens and Crawford 2010 - Revised for ANOVA

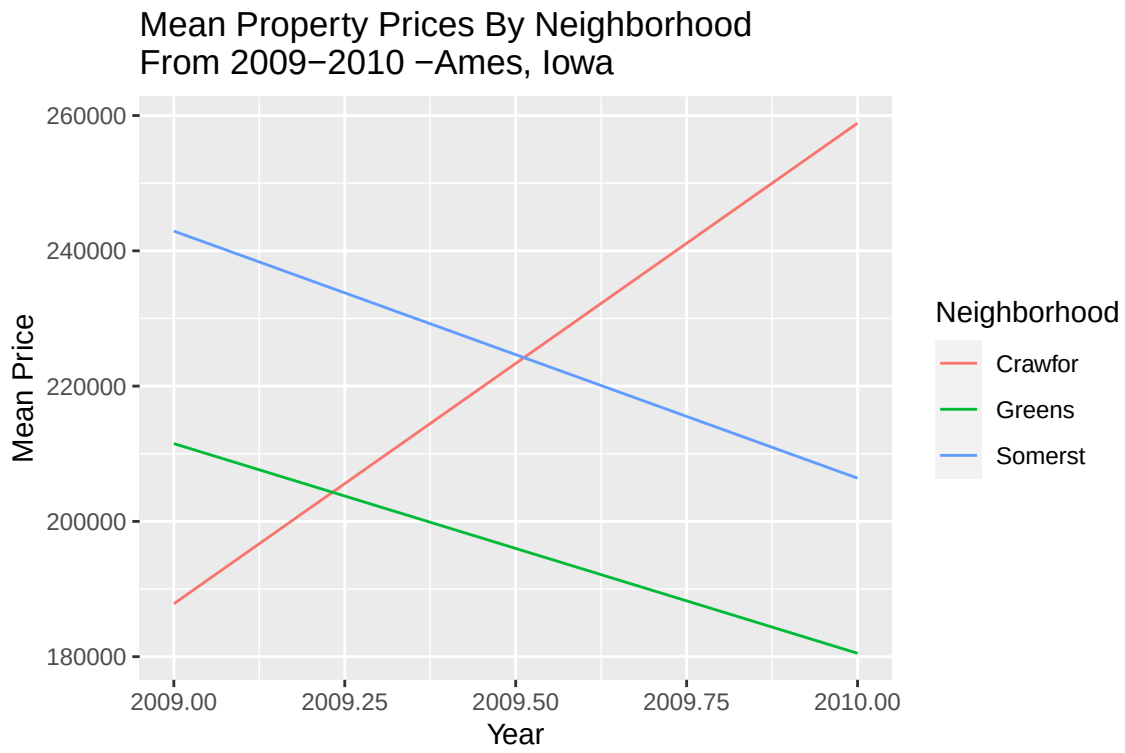
Neighborhood	Mean Price	Standard Deviation of Price	Observations
Crawfor	258876.4	44647.31	7
Greens	180500.0	36062.45	2
Somerst	206387.0	44663.09	21

Box Plot Distribution - 2nd Trial - ANOVA Assumptions Now Met:

Below we have changed 2/3 of the neighborhoods, removed the room number constraints, and limited the year to 2010 to meet the assumptions for normalcy required to do ANOVA. Skew has been minimized.



Quick Time Series Revisit for 3 Neighborhoods Now That Assumptions Are Met- With Trajectories Leading Up to this Point (2009-2010):



ANOVA Analysis:

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Neighborhood  2 1.731e+10 8.653e+09   4.395 0.0223 *
## Residuals    27 5.316e+10 1.969e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Null Hypothesis (H0) here being tested via ANOVA is that the mean price value of each neighborhood "Somerset", "Greens" and "Crawford", are all equal and the Alternative Hypothesis (Ha) is that there is at least one price comparison that is not equal. F-Val is (SST/df1)/(SSE/df2)

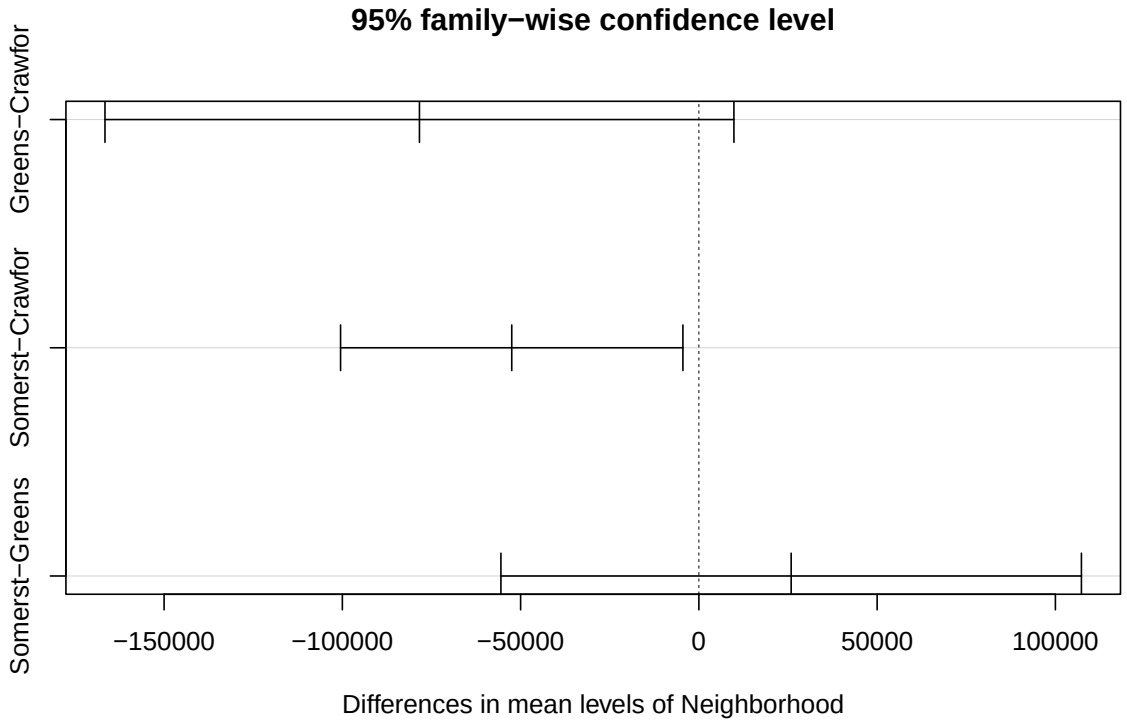
Here we can verify $F_{2,27}$ = (1.731 * 10^10 /2) / (5.316 * 10^10 /27) = 4.395
We then calculate (or observe from results) P Value: $P(F_{2,27} > 4.395) = 0.0223$, which is small enough to reject the Null Hypothesis within .05 probability of error, alpha = .05. To recall: This indicates that at least one of these price mean comparisons between neighborhoods is NOT equal for 2010.

IV. Closer Comparison: Tukey and Confidence Intervals

Using a "Tukey" Multiple Mean Comparison, we can see the price mean relationships on a case-by-case basis that are the reason for the Null Hypothesis rejection in the ANOVA. The second relationship, Somerset and Crawford, is the one 95% Confidence Interval that does not have a zero within its range. This appears to be the only one that justifies rejection of the Null Hypothesis. The other two have zeros within their intervals.

Tukey Multiple Mean Comparison

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $Neighborhood
##                       diff        lwr        upr     p adj
## Greens-Crawfor  -78376.43 -166583.57   9830.709 0.0887532
## Somerst-Crawfor -52489.48 -100503.36  -4475.592 0.0300297
## Somerst-Greens   25886.95  -55524.55 107298.454 0.7132266
```



Somerset Mean Confidence Interval - T Distribution for Small Sample

For the Confidence Intervals regarding individual neighborhood means (below), we are 95% confident that the housing price mean of the neighborhood of Somerset for year 2010 is between $186,057 and $226717

```
206387-c(-1,1)*qt(0.025, 20)*(44663.09/sqrt(21))
```

```
## [1] 186056.6 226717.4
```

Crawford Mean Confidence Interval - T Distribution for Small Sample

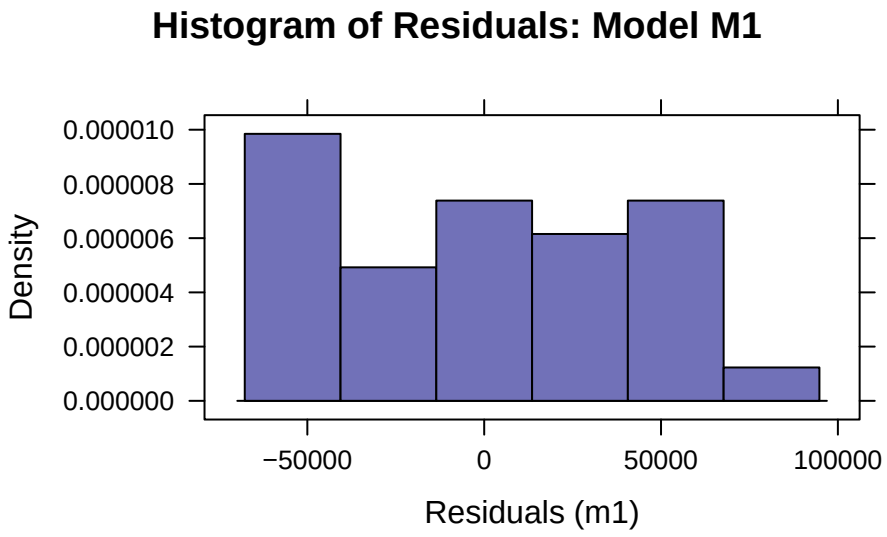
As for Crawford we are 95% confident that the housing price mean in this neighborhood for year 2010 is between $217585 and $300168

```
258876.4-c(-1,1)*qt(0.025, 6)*(44647.31/sqrt(7))
```

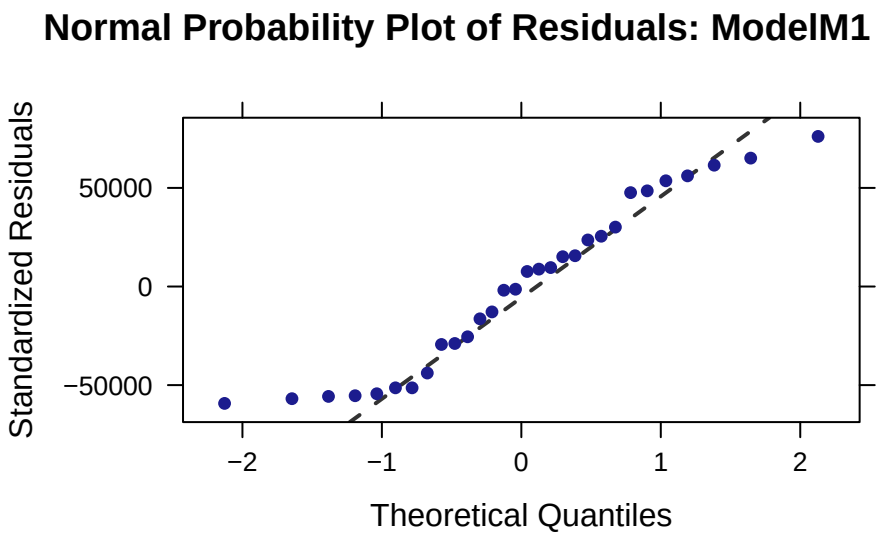
```
## [1] 217584.5 300168.3
```

Revisit - ANOVA Assumptions Normality/Variance Analysis via Infograph

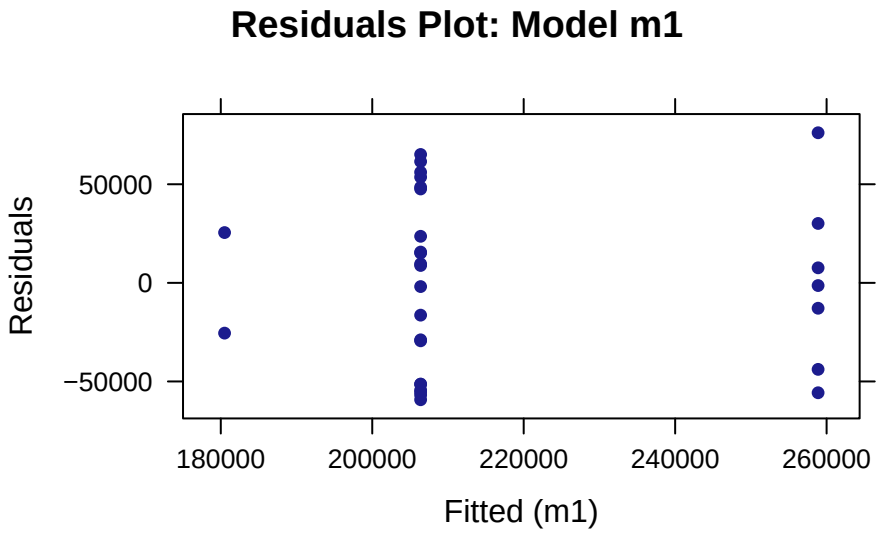
The 'Normality' requirement for our study is best depicted in the second diagnostic infograph, while, not so much a 'bell shape' for the bar graph.



Interesting to note that we have reassuringly close adherence of points to the linear model in the second, with minor deviation at the ends. Indicating high Normality, what we need for our ANOVA assumption to be met.



The 'Variance test' component of our study looks to have a good balance. We might do better to rely on the calculation that demonstrates its < 2, because from the 'sparse' availability of points, it is difficult to determine if any 'coning' or other concerning pattern is actually taking place.



V. CONCLUSION

It should be noted that the Greens Neighborhood has only 2 observations, while we realize, technically the min acceptable for ANOVA is '3'. This shortcoming was noticed late, and while out of the time scope for replacement and re-factor, our study did yield a valuable finding:

Both indicated and corroborated via ANOVA and Tukey, we found a statistically significant difference in means between the neighborhoods Somerset and Crawford for the year 2010. Furthermore, utilizing the same methods could shed light on the interesting trend of pronounced decline for Somerset and growth for Crawford in the year leading to this difference (2009). During this time it appears by visual assessment of the graph that almost 'traded places' with regards to housing value. This could be shown by a likewise ANOVA and Tukey 95% Confidence Interval comparison for their difference in means in 2009, which could substantiate what is indicated visually, that their positions in one year time (2009-2010), are almost inverted (but with higher min and max.)