Iowa State Real Estate

A Data Analysis Project

Term Project - Group 3

SCS 3250 - Section 035

Fall 2018

Report submitted by

Group 3

Name of the members (in alphabetical order):

- Andrew Pang (andrew.wl.pang@gmail.com)
- Anuj Gupta (anuj0510@gmail.com)
- Jessica Lee (dayea.lee@gmail.com)
- Kerim Terzioglu (kterzioglu@yahoo.com)
- Ziauddin Ahmed (zia.ahmed28@gmail.com)

Preface

The following analysis work forms the term project, submitted as a combined *group* effort by the aforementioned members of University of Toronto - School of Continuing Studies. This *group* has undertaken the project as a mandatory requirement for the course - Fundamentals of Data Science, Section 035 in the Fall 2018 semester.

In order to avoid any possible copyright infringements, the analysis has been done on a data set publicly available on Kaggle, named - House Prices: Advanced Regression Techniques. The data set was compiled by Dean De Cock for use in data science education, and approved by the instructor Larry Simon to be used for the given purpose in the current context.

Any questions or concerns regarding the project can be directed to any of the group members listed on the previous page.

Objective

As a home buyer, prospective home buyer, investor, home owner who wants to sell his/her house, or an agent working in the area of buying and selling houses, all of us are, have been, or likely to be a part of the real estate market. While finding the dream house remains the primary concern of a home buyer who has just set foot into the market, the next big challenge being faced by everyone involved in the process is - what is the right price for that dream house that you just found?

While the price remains a major concern for nearly everyone involved in the house buying & selling process, there is another set of questions that worries you if you were an investor or broker. And these questions are - what is the right market to put your efforts and money into? Which neighbourhood is hot and expected to give you maximum value for your money? Is it worth putting your money into an old house or a new property? Does the age of construction and renovations matter, and if so, to what extent?

The objective of this study is to look at the available data set and to attempt to answer some of these questions for the benefit of the readers. And hopefully deliver some insights for those of us who 'are' in process of buying or selling a house! So let's proceed further and look at what this report has to offer.

Data Preparation

The data set used for the analysis is the comprehensive sale and purchase data for Iowa State, gathered between the year 2006 and 2010. It may be important to mention here that 2010 is a partial year with last quarter missing. The data features a total of 81 parameters recorded for 1,460 transactions that took place in the said period in different neighbourhoods of the city. *Few* of the key features of the data, that have been used extensively in the analysis, are presented below:

1. Types of properties:

C	Commercial
FV	Floating Village Residential
RH	Residential High Density
RL	Residential Low Density
RM	Residential Medium Density

2. Neighbourhood - There are 25 neighbourhoods within Ames city limits. A few examples (but not limited to) are:

Blmngtn	Bloomington Heights
Blueste	Bluestem
BrDale	Briardale
BrkSide	Brookside
ClearCr	Clear Creek
CollgCr	College Creek
Crawfor	Crawford
Edwards	Edwards
Gilbert	Gilbert

IDOTRR	Iowa DOT and Rail Road
MeadowV	Meadow Village
Mitchel	Mitchell
Names	North Ames

3. OverallCond: Rates the overall condition of the house

10	Very Excellent
9	Excellent
8	Very Good
7	Good

And so on…

4. YearBuilt: Original construction date

Missing values

The data gathered was quite comprehensive with minimal missing values. There were a few parameters that were missing for a large portion of the data, namely - PoolQC (Pool Quality), MiscFeature (Miscellaneous features), Alley (Type of Alley access to property) and Fence (Fency Quality). There were few more parameters that were missing for less than 50% of the data. All the missing values were defaulted to a '0' value.

As an exception to the above rule:

- Lot Frontage - The missing values were filled in with the 'median' values for the specific neighbourhood
- Electrical - Missing values were filled-in with the 'mode' value

Glossary of terms used

Correlation: Pearson's correlation coefficient gives a measure of the relationship between two variables on a scale of -1 to 1. +1 represents a perfect positive correlation whereas -1 would represent a perfect negative correlation. In the analysis, focus has been on positive correlations observed in the heat map

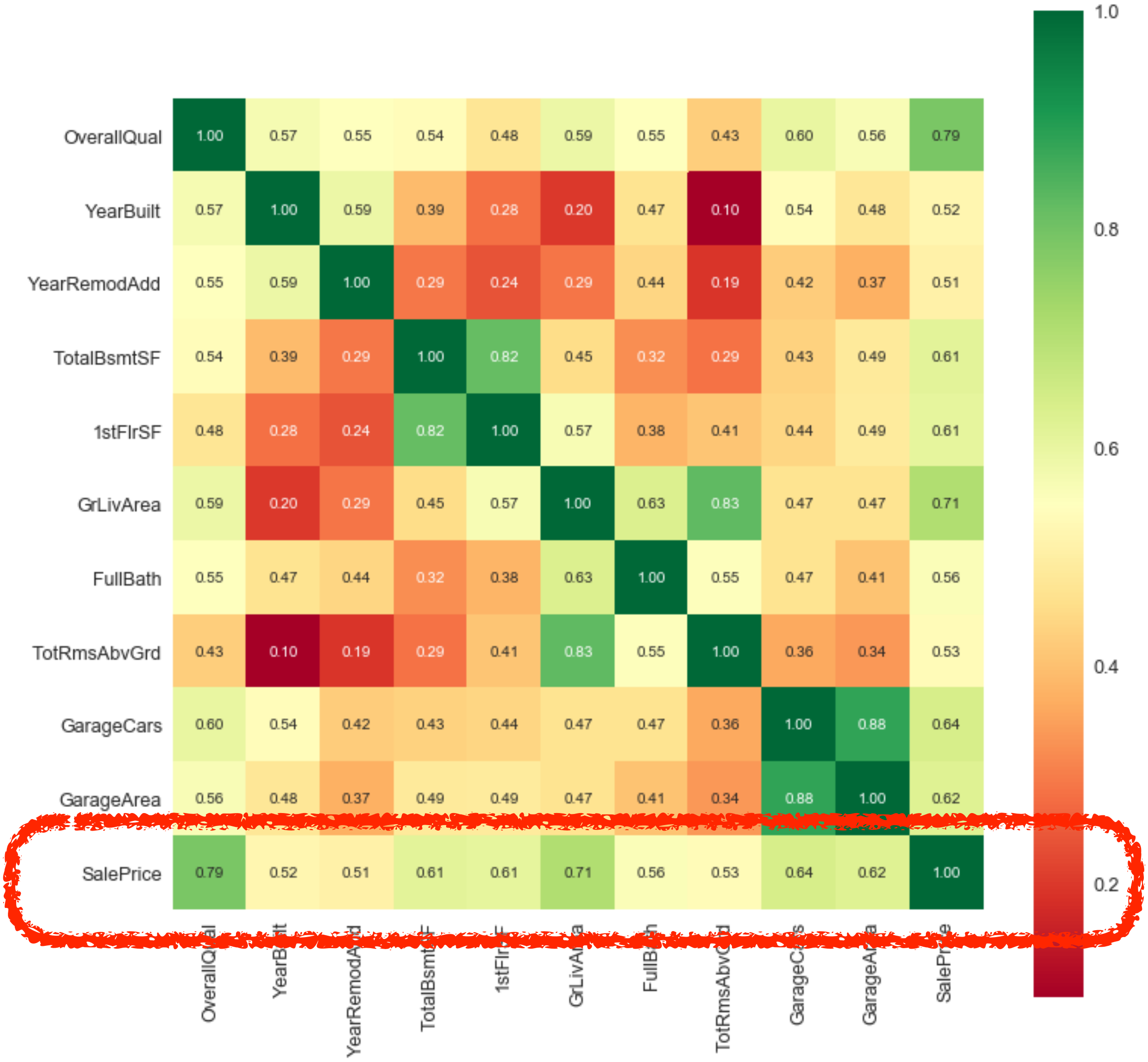
Time Series: Analyzing a set of observations taken at different points in time to extract meaningful statistics. This is achieved using Python's autocorrelation_plot tool to find the effect of lag vs SalesPrice. This helped identify these patterns:

- **Trends**: Long-term increase or decrease (Page 10)
- **Seasonality**: Where there is an influence that varies with the time of year or other calendar period (Page 11)
- **Cycles**: Patterns of repeated increase and decrease of varying period (Page 16)

Heat Map: A graphical representation of data where the individual values contained in a matrix are represented as colours

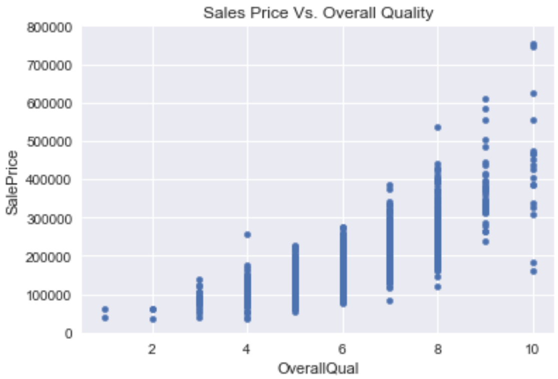
Analysis

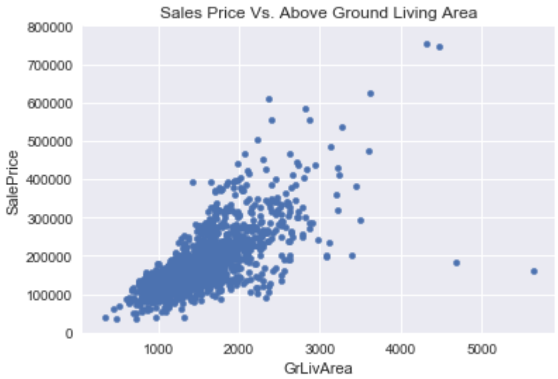
A preliminary correlation “heat-map” was built to determine the factors with maximum influence on the sale price of a property. This gave us a clear indication as to which parameters were to be kept at the core of the analysis, and which parameters could be pushed to the back seat.

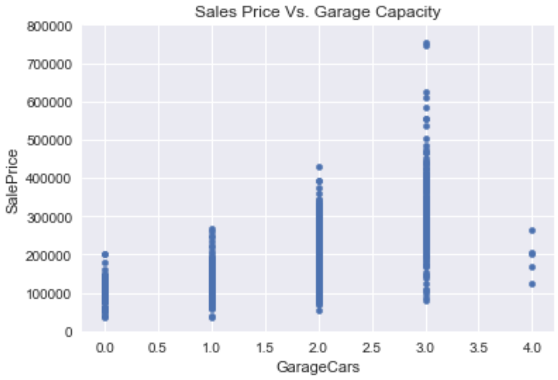


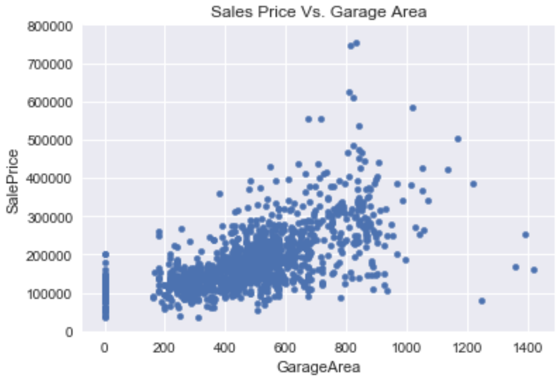
Based on the derivations from the above figure, one-step deeper relationships were explored between the Sale Price and variables that have highest correlation with the Sale Price - Overall Quality, Ground Floor Living Area, Garage for

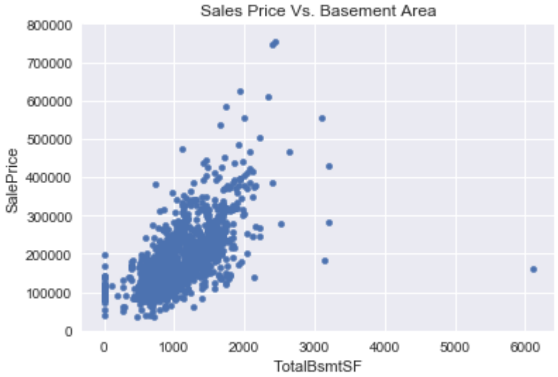
Cars, Garage Area, Total Basement Square Footage, First Floor Square Footage, Total number of Rooms above Ground Floor and Year Built.

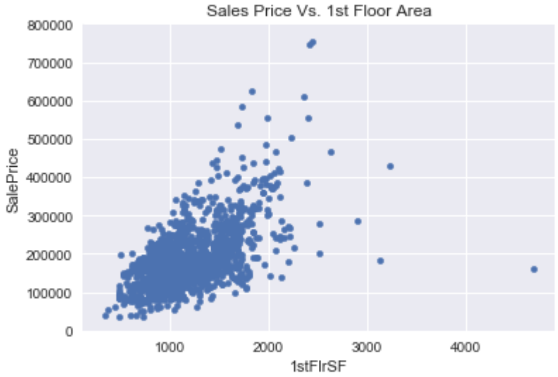


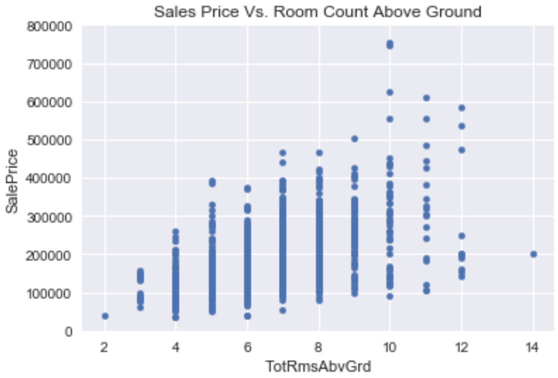


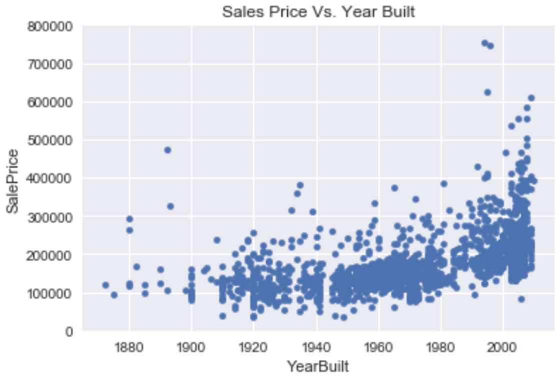










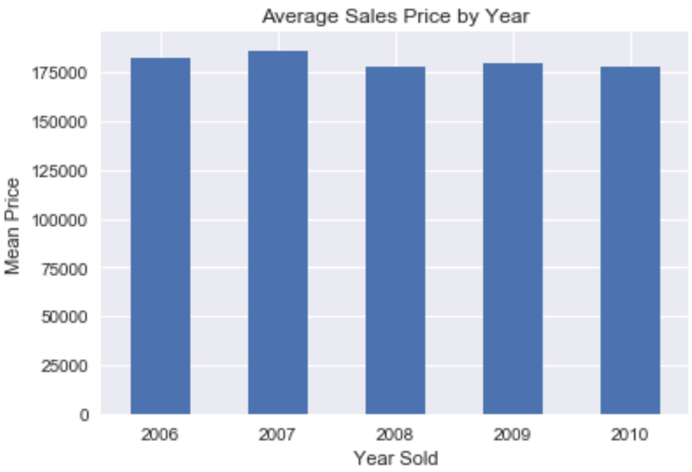


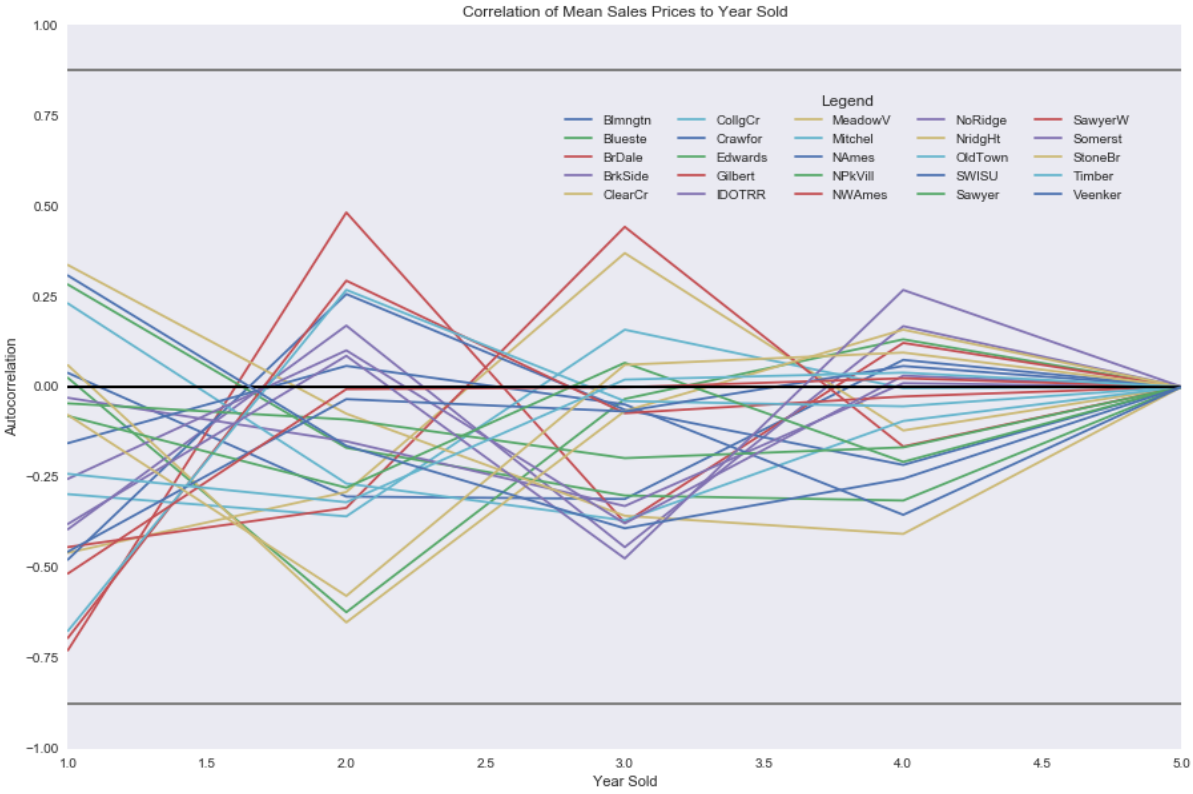
As can be generally seen, all the above listed variables showed a positive correlation with the Sale Price of the house - forming the first plank of the analysis. This leads to the next phase of the analysis - average sale price!

Average Sale Price

Let's take a first look at the Year-on-Year average sale price across our data set.

The graph to the right suggests a uniform sale price across neighbourhoods throughout our study period. This called for an insight into individual neighbourhoods - whether they show similar trends or is there a different theme there?

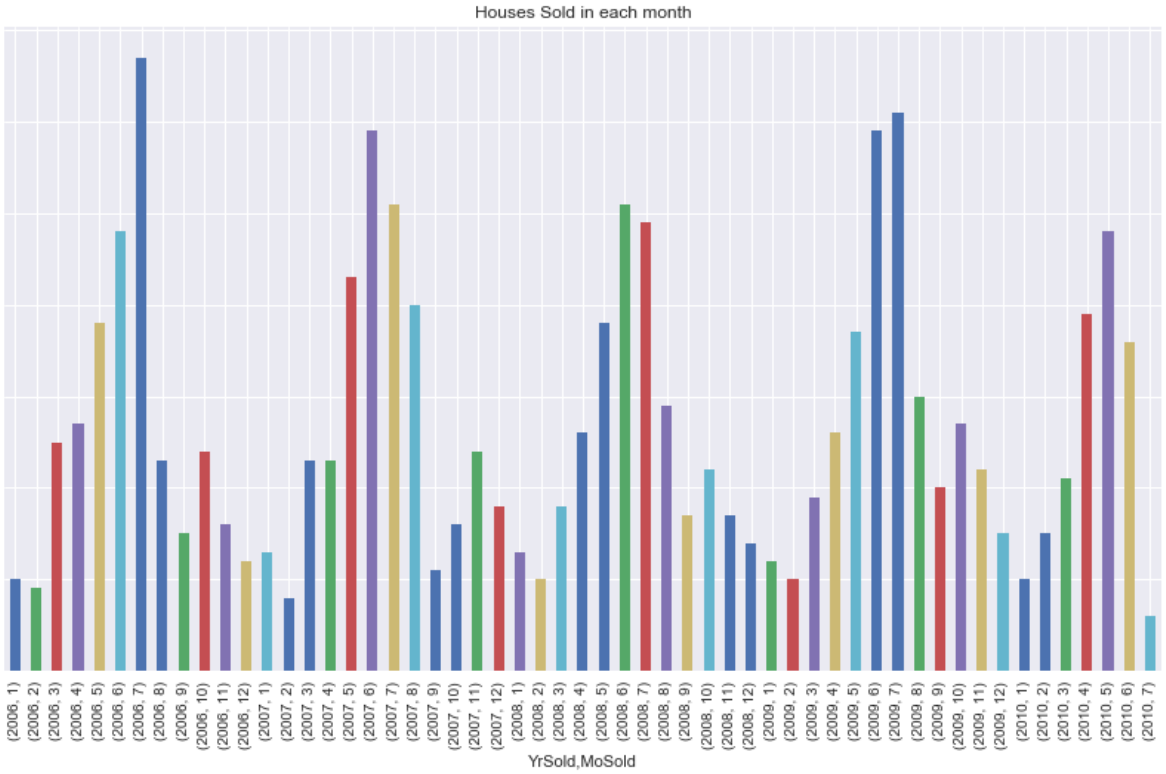




And thus, a quick look at the neighbourhood-wise sales prices tell a completely different story - while the overall sales prices averaged out similarly, most neighbourhoods went through some massive turbulence in the average sale prices year after year.

Seasonality and Real Estate Market

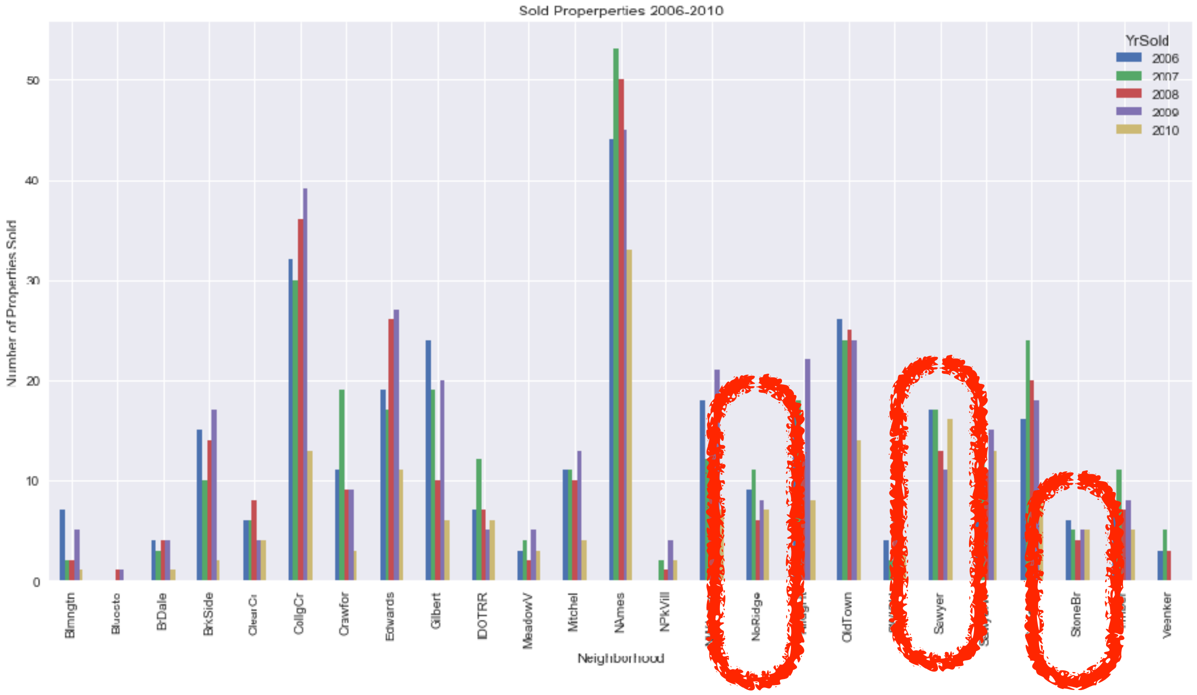
An important aspect that impacts house sales is *seasonality*. A general expectation, as in other domains like shopping and traveling, would be - high level of activity in the summer months and a generally cooled down market in the cold winter months. Let's take a look at what the data set has to offer:



The above graph that reflects the number of houses sold in a monthly fashion, also reflects and reaffirms the notion that we set forth in the previous paragraph. The house sales invariable go up during the summer months, and the shows an alternative trend of high-&-low sales between winters and summers.

Number of Houses Sold - By Neighbourhood

Taking a lead from the above, we look further into the number of sales by neighbourhoods and how the numbers fluctuate YoY. Here we go:

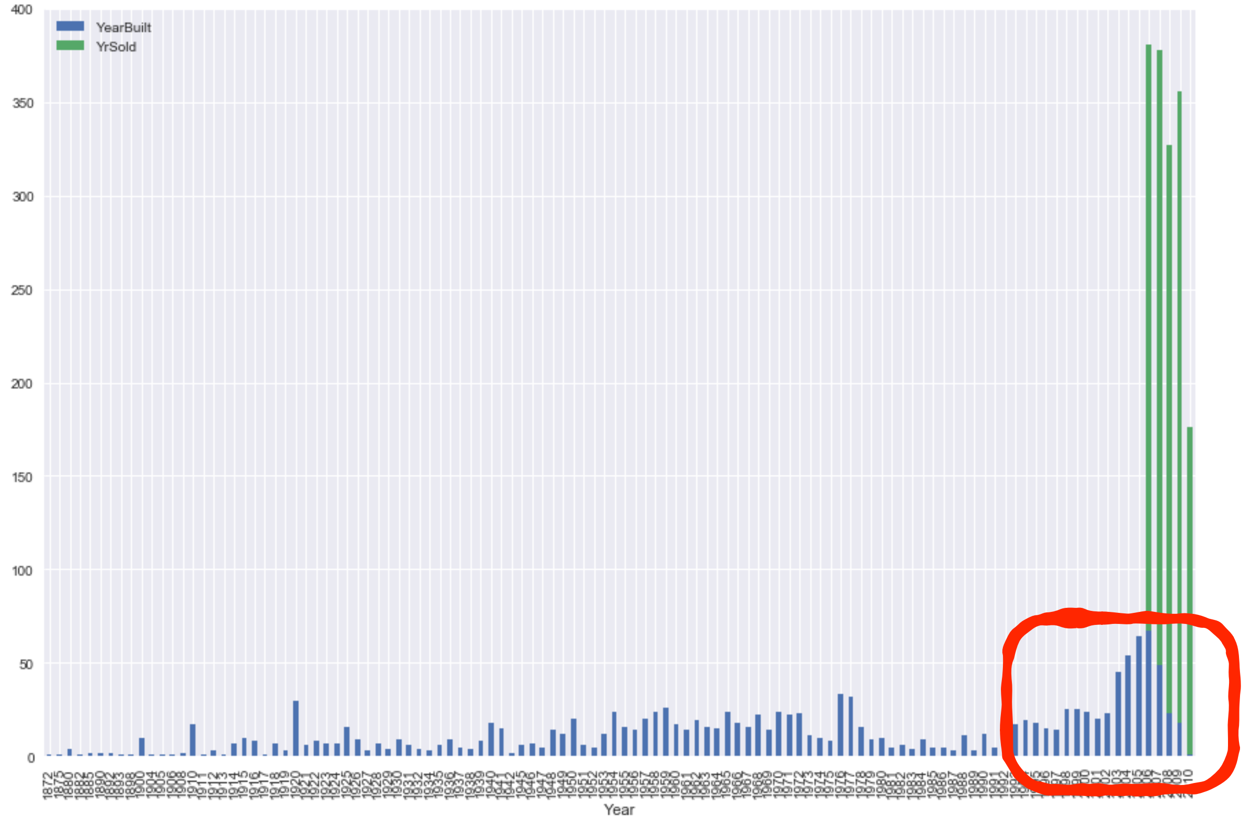


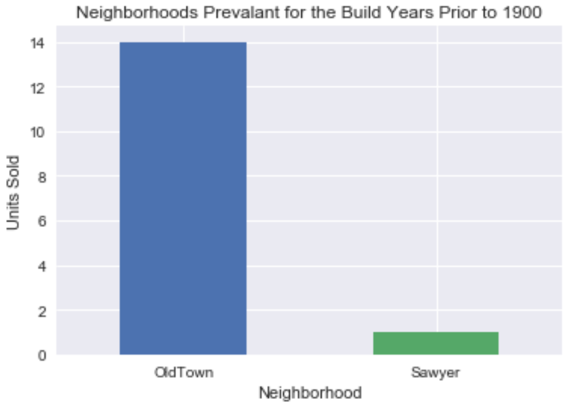
Interestingly, while majority of the neighbourhoods showed a decline in sales in the year 2010, three of those stood apart with higher than average sales from the past four years - Sawyer, Northridge and Stone Brook.

Number of Houses Sold - By Age

We further look at the the number of houses sold as a function of age of the houses, i.e., number of sales vis-a-vis the year of built.

The above graph clearly shows an inclination towards new houses. The number of sales for houses built in recent years is visibly higher than those built farther in the past.

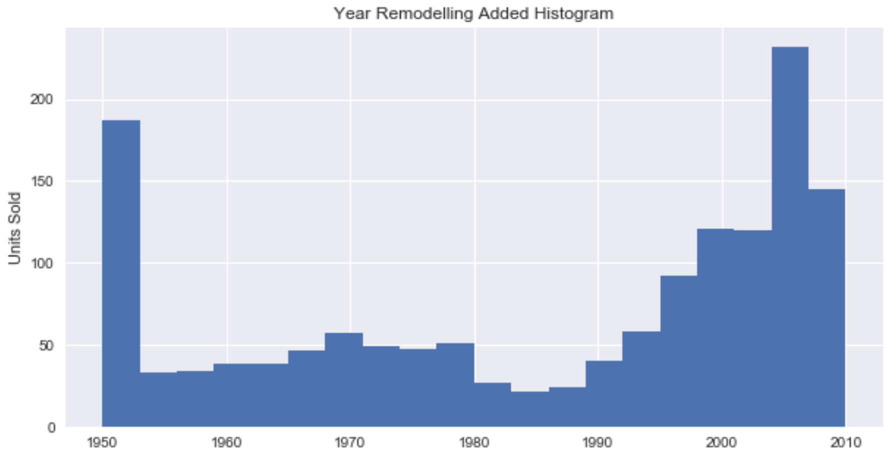


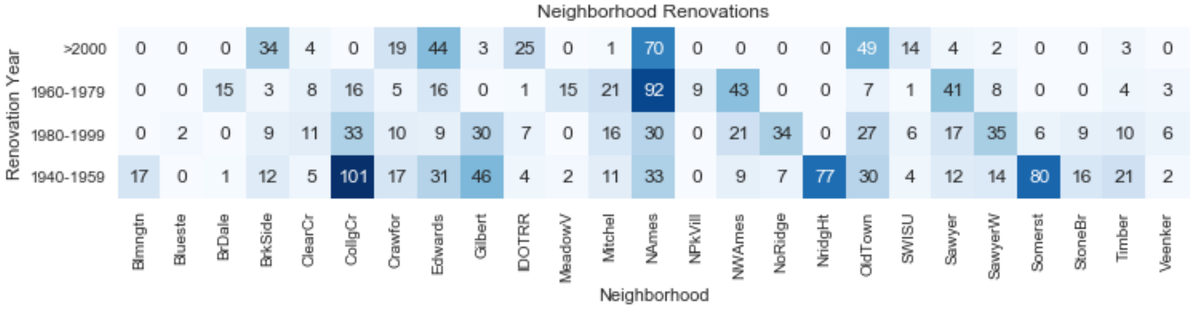


However, there seems to be an exception to this rule:

Among the old neighbourhoods of the region, Old Town seems to be of special interest to buyers with the number of sales being much higher than its counterpart Sawyer, for houses built in the 1800's.

While looking at the overall age of the houses, another aspect that calls for attention is the year houses were renovated or remodelled, and whether that has any bearing on the sales. Let's take a look:





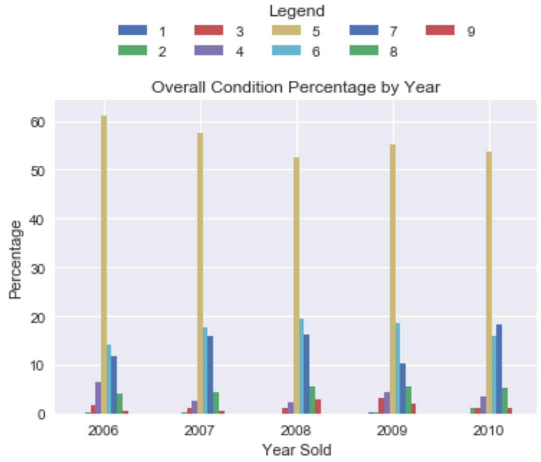
Renovation Year	Blmngtn	Blueste	BrDale	BrkSide	ClearCr	CollgCr	Crawfor	Edwards	Gilbert	IDOTRR	MeadowV	Mitchel	NAmes	NPkVill	NWAmes	NoRidge	NridgHt	OldTown	SWISU	Sawyer	SawyerW	Somerst	StoneBr	Timber	Veenker
>2000	0	0	0	34	4	0	19	44	3	25	0	1	70	0	0	0	0	49	14	4	2	0	0	3	0
1960-1979	0	0	15	3	8	16	5	16	0	1	15	21	92	9	43	0	0	7	1	41	8	0	0	4	3
1980-1999	0	2	0	9	11	33	10	9	30	7	0	16	30	0	21	34	0	27	6	17	35	6	9	10	6
1940-1959	17	0	1	12	5	101	17	31	46	4	2	11	33	0	9	7	77	30	4	12	14	80	16	21	2

The data here reflects a few noteworthy insights:

- Houses remodelled prior to 1950's were in demand nearly as much as those done in recent years
- College Creek, Northridge Heights and Somerset are three most active old neighbourhoods with high sales and remodelling done over 50 years ago

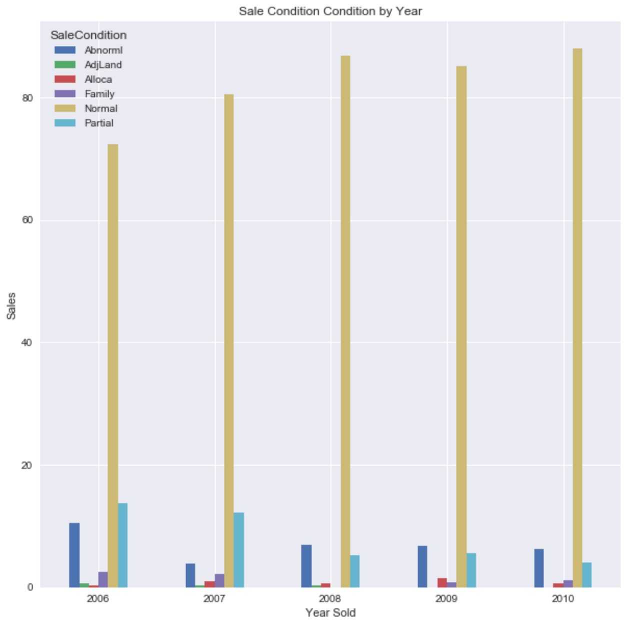
Number of Houses sold - by Overall Condition

Here we look at the overall condition of the house that people preferred buying, and based on a percentage analysis of the units sold across categories, where 10 means excellent condition and 0 means worst, people showed a tendency to stick to the 5-7 range of overall condition across the years.



With some fluctuations within the bands, 5-7 band witnessed more than 80% sales in any given year, while the sub-5 and above-7 bands constituted the rest.

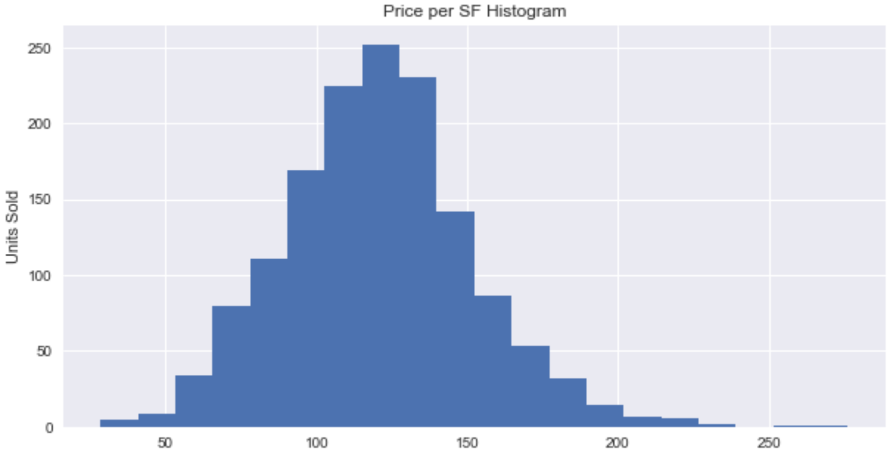
Number of Houses sold - by Sale Condition



There is a clear upward trend in the "normal sales" and a significant rise in this category in the 5-year period. As against that, abnormal and partial sales grew less in popularity falling drastically - proportionately speaking. Another noteworthy insight here is absence of "family" sales in 2008 - the year that the worst financial crisis in recent years hit the global markets.

Price Per Square Foot (PPSF)

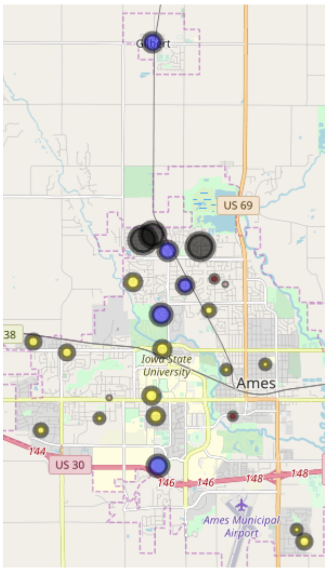
Another important factor in the house buying process, that helps make an apple-to-apple comparison between two house prices is, of course, Price Per Square



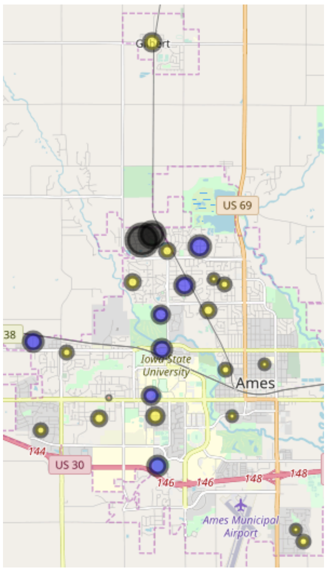
Foot. PPSF is an engineered attribute derived from the sale price and ground floor living area, both of which are *not* normally distributed w.r.t. units sold. However, the engineered attribute PPSF is shown to be normally distributed.

Mean Prices Sold - by Years

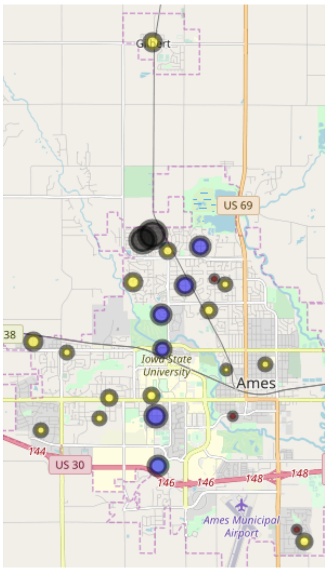
The above analysis calls for a geographic study of the average sale prices across years, and below is a snapshot of what we see.



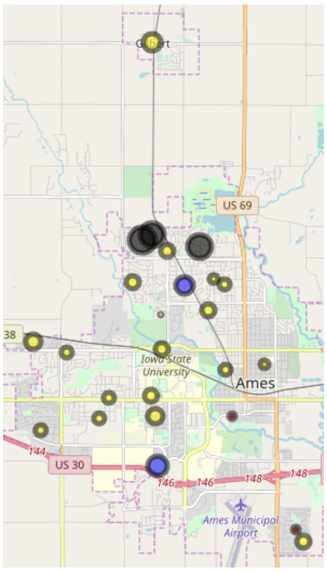
2006

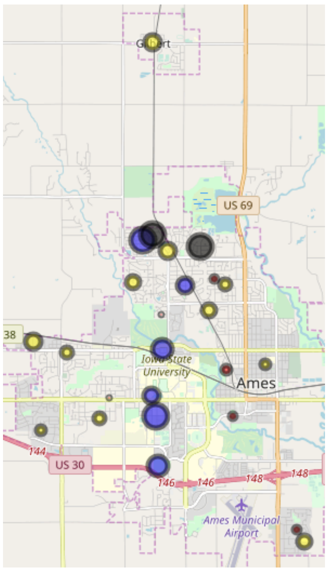


2007



2008





🔴 RED: Mean average < 100k

🟡 YELLOW: Mean average > 100k and < 200k

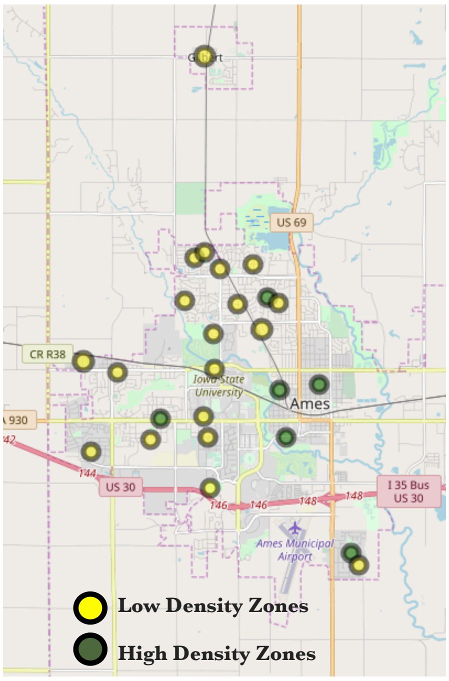
🔵 BLUE: Mean average > 200k and < 300k

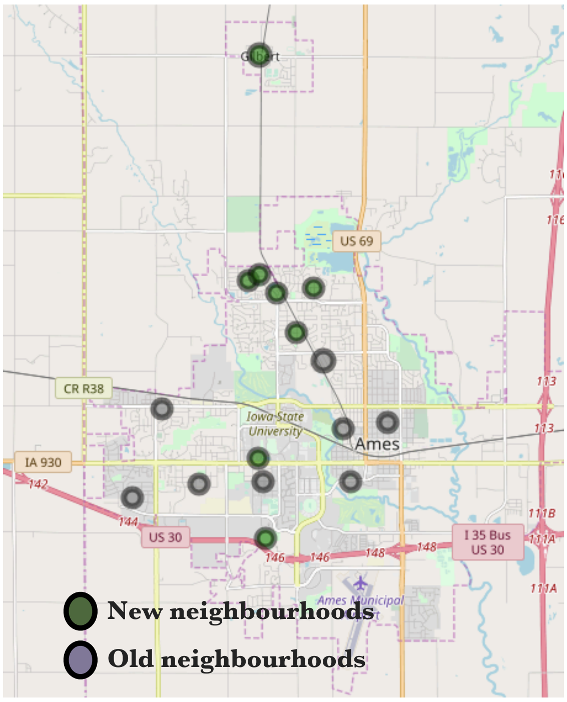
⚫ BLACK: Mean average > 300K

A Study into Neighbourhoods

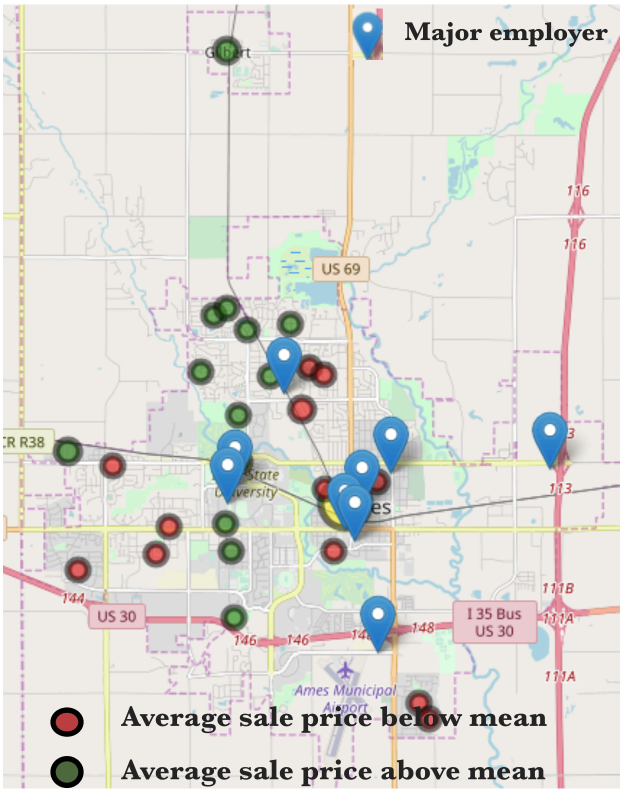
Since the analysis presented in this reports dwells significantly upon neighbourhoods, it is imperative to have a better understandings of this parameter. This will not only help us understand how it may have impacted the data, but may also help us understanding the housing market as such.

Figure to the right is an attempt to categorize the neighbourhoods in two zones - Low Density and High Density - depending upon the majority of sales that took place in those areas. As is visible, majority of the central, western, northern and southern parts of the city are low density zones - hinting towards smaller families. Only a few neighbourhoods on the eastern and south eastern parts of the city of medium to high density - possibly those occupied by larger families.





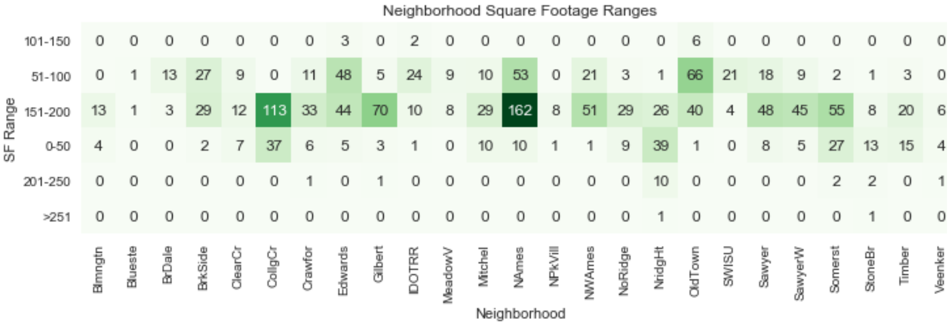
Taking a cue from the densities, another aspects that demands attention is the age of the neighbourhoods. A quick look on the left indicates a North-to-South trajectory of new neighbourhoods, while the East-West of the city core constitutes the older ones. Quite expected, we'd say! However, of special interest here is a new neighbourhood in the core of the city, to the South of Iowa Sate University, that also happens to be the biggest employer of the city. Perhaps because of more people moving in to uninhabited parts of the city core?



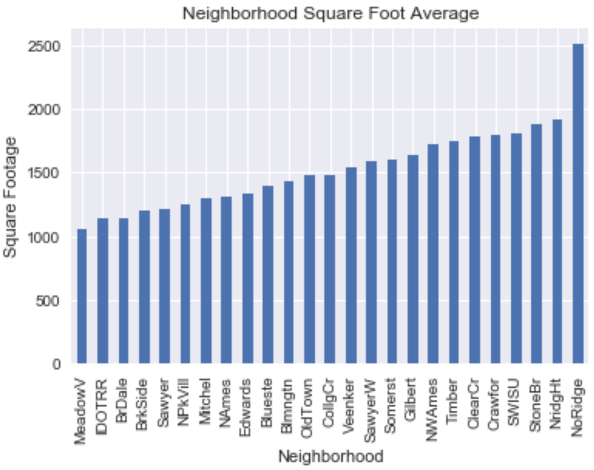
And since we all understand that the housing market moves primarily with the commercial activity, it is of utmost importance to see where are the major employers of the city located and how that impacts the housing market.

The picture to the right maps the major employers, and it is worth noting that Iowa State University is the biggest employers giving jobs to nearly 1/4th of the city's population! A closer look at these three maps hints at the following:

- Old neighbourhoods of the city to the east of the city are also the ones with below average sale prices - seemingly less activity there
- New neighbourhoods on the North-South trajectory are also the ones with above average sale prices - clearly high in demand by the workforce
- This North-South trajectory also happens to be the low density housing zone, further testifying the last point



SF Range	Blmngtn	Blueste	BrDale	BrkSide	ClearCr	CollgCr	Crawfor	Edwards	Gilbert	IDOTRR	MeadowV	Mitchel	NAmes	NPkVill	NWAmes	NoRidge	NridgHt	OldTown	SWISU	Sawyer	SawyerW	Somerst	StoneBr	Timber	Veenker
101-150	0	0	0	0	0	0	0	3	0	2	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0
51-100	0	1	13	27	9	0	11	48	5	24	9	10	53	0	21	3	1	66	21	18	9	2	1	3	0
151-200	13	1	3	29	12	113	33	44	70	10	8	29	162	8	51	29	26	40	4	48	45	55	8	20	6
0-50	4	0	0	2	7	37	6	5	3	1	0	10	10	1	1	9	39	1	0	8	5	27	13	15	4
201-250	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	10	0	0	0	0	2	2	0	1
>251	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0



Another interesting aspect worth highlighting, while talking about the neighbourhoods, is about those with premium lot sizes. The 'heat map' on the left suggests there are two neighbourhoods, Northridge Heights and Stonebrook, with premium square foot plots of over 251 sq ft, while it is only Northridge Heights that appears in the list of most expensive neighbourhoods with a sell price of over $300,000's. Also worth noting is the fact that the two adjacent neighbourhoods - Northridge and Northridge Heights - have an average selling price of over $300K, despite the former having much smaller houses than the latter one. This clearly indicates that *Northridge is a more established and premium neighbourhood with a high Per Square Foot price compared to its geographical neighbour*.

Conclusion

The data set studied in this analysis, spanning a period of five years and the length & breadth of the Ames city, presented us with quite a few surprising insights. From an initial indication of uniform average sale prices to understanding the employment centres and how it moved the housing market. In order to summarize the key takeaways, we could say the following:

1. Over time, satellite markets rotate in popularity and sales prices with a resulting trend to move out of the core and back in
2. Houses in long established neighbourhoods are more likely to increase in value. These houses have more space and bigger backyards, desired by growing young families as well as older families that don't want to downsize
3. Inner city homes are closer to downtown core surrounded by employers, transportation hubs and artillery roads, and are considered starter homes
4. Year 2008 saw a jump in sales for some neighbourhoods with old houses, like that of Old Town, with a subsequent increase in average sale prices
5. Northridge and Northridge Heights - neighbourhoods next to each other - one has high square footage and lower PPSF while the other has premium houses with a high PPSF; both tend to have some of the highest sale prices

Appendix

Appendix A

Yearly Average Sale Price by Neighbourhood

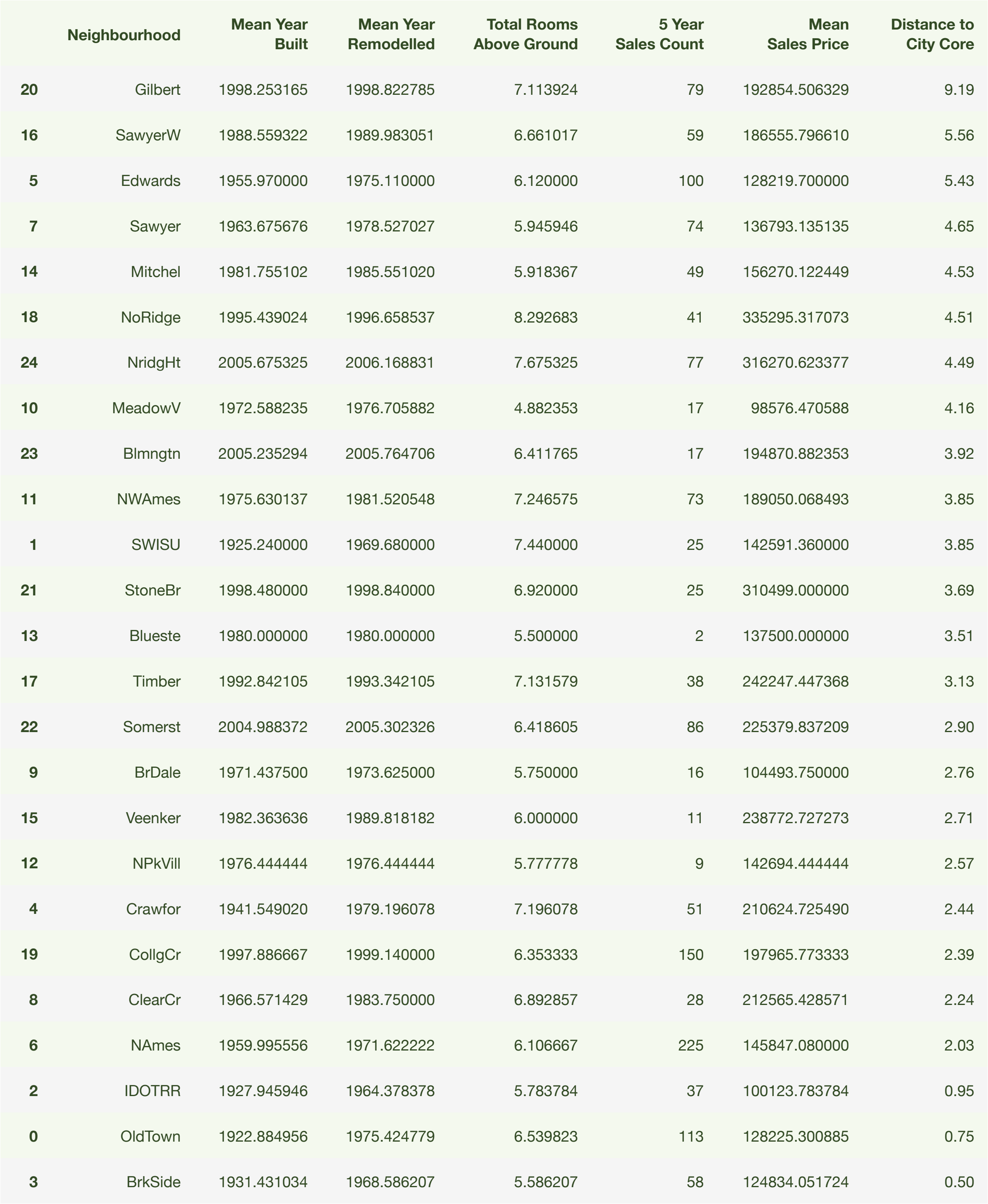
Neighbourhood	Mean Sales Price
Blmngtn	194870.882353
Blueste	137500.000000
BrDale	104493.750000
BrkSide	124834.051724
ClearCr	212565.428571
CollgCr	197965.773333
Crawfor	210624.725490
Edwards	128219.700000
Gilbert	192854.506329
IDOTRR	100123.783784
MeadowV	98576.470588
Mitchel	156270.122449
NAmes	145847.080000
NPkVill	142694.444444
NWAmes	189050.068493
NoRidge	335295.317073
NridgHt	316270.623377
OldTown	128225.300885
SWISU	142591.360000
Sawyer	136793.135135
SawyerW	186555.796610
Somerst	225379.837209
StoneBr	310499.000000
Timber	242247.447368
Veenker	238772.727273

Appendix B
5-Yearly Neighbourhood Mean Sale Prices

	2006	2007	2008	2009	2010
Neighbourhood					
Blmngtn	217087.000000	183350.500000	175447.500000	176720.000000	192000.000000
Blueste	0.000000	0.000000	151000.000000	124000.000000	0.000000
BrDale	96750.000000	113833.333333	95225.000000	118625.000000	88000.000000
BrkSide	112746.666667	135737.500000	121707.142857	134994.117647	96500.000000
ClearCr	199166.666667	236333.333333	208991.500000	169875.000000	246850.000000
CollgCr	199016.406250	213999.933333	187718.055556	192317.769231	203700.000000
Crawfor	196635.181818	198777.578947	254411.111111	180211.111111	296833.333333
Edwards	134403.684211	132588.235294	132473.076923	123855.555556	111445.454545
Gilbert	200250.625000	181967.947368	186000.000000	199955.000000	185500.000000
IDOTRR	95758.714286	118933.333333	91642.857143	89580.000000	86278.166667
MeadowV	123466.666667	105850.000000	98000.000000	88400.000000	81333.333333
Mitchel	150036.363636	136731.818182	165280.000000	167860.461538	166950.000000
NAmes	138985.454545	142962.264151	151553.160000	143880.000000	153665.909091
NPkVill	0.000000	141500.000000	140000.000000	146937.500000	136750.000000
NWAmes	199463.888889	175267.083333	193820.000000	185133.333333	187428.571429
NoRidge	322333.333333	399730.909091	304750.000000	323875.000000	289938.285714
NridgHt	305491.882353	310833.111111	332422.833333	323143.500000	308281.125000
OldTown	135963.807692	114794.625000	147670.000000	116378.291667	122464.285714
SWISU	130125.000000	187500.000000	139612.500000	141048.000000	141333.333333
Sawyer	149735.294118	133935.294118	128900.692308	136925.727273	132400.000000
SawyerW	164787.500000	209300.000000	184080.000000	183934.133333	184076.923077
Somerst	210268.875000	233248.916667	225631.000000	236315.000000	206762.500000
StoneBr	365046.666667	279585.200000	245000.000000	319967.400000	318886.400000
Timber	264485.714286	229470.545455	234361.000000	245437.500000	245160.000000
Veenker	273333.333333	214900.000000	244000.000000	0.000000	0.000000

Appendix C

Summary Statistics



	Neighbourhood	Mean Year Built	Mean Year Remodelled	Total Rooms Above Ground	5 Year Sales Count	Mean Sales Price	Distance to City Core
20	Gilbert	1998.253165	1998.822785	7.113924	79	192854.506329	9.19
16	SawyerW	1988.559322	1989.983051	6.661017	59	186555.796610	5.56
5	Edwards	1955.970000	1975.110000	6.120000	100	128219.700000	5.43
7	Sawyer	1963.675676	1978.527027	5.945946	74	136793.135135	4.65
14	Mitchel	1981.755102	1985.551020	5.918367	49	156270.122449	4.53
18	NoRidge	1995.439024	1996.658537	8.292683	41	335295.317073	4.51
24	NridgHt	2005.675325	2006.168831	7.675325	77	316270.623377	4.49
10	MeadowV	1972.588235	1976.705882	4.882353	17	98576.470588	4.16
23	Blmngtn	2005.235294	2005.764706	6.411765	17	194870.882353	3.92
11	NWAmes	1975.630137	1981.520548	7.246575	73	189050.068493	3.85
1	SWISU	1925.240000	1969.680000	7.440000	25	142591.360000	3.85
21	StoneBr	1998.480000	1998.840000	6.920000	25	310499.000000	3.69
13	Blueste	1980.000000	1980.000000	5.500000	2	137500.000000	3.51
17	Timber	1992.842105	1993.342105	7.131579	38	242247.447368	3.13
22	Somerst	2004.988372	2005.302326	6.418605	86	225379.837209	2.90
9	BrDale	1971.437500	1973.625000	5.750000	16	104493.750000	2.76
15	Veenker	1982.363636	1989.818182	6.000000	11	238772.727273	2.71
12	NPkVill	1976.444444	1976.444444	5.777778	9	142694.444444	2.57
4	Crawfor	1941.549020	1979.196078	7.196078	51	210624.725490	2.44
19	CollgCr	1997.886667	1999.140000	6.353333	150	197965.773333	2.39
8	ClearCr	1966.571429	1983.750000	6.892857	28	212565.428571	2.24
6	NAmes	1959.995556	1971.622222	6.106667	225	145847.080000	2.03
2	IDOTRR	1927.945946	1964.378378	5.783784	37	100123.783784	0.95
0	OldTown	1922.884956	1975.424779	6.539823	113	128225.300885	0.75
3	BrkSide	1931.431034	1968.586207	5.586207	58	124834.051724	0.50

Appendix D
Demographics

Source: https://en.wikipedia.org/wiki/Ames,_Iowa

In 2017, Ames had a population of 66,498.[6] Iowa State University is home to 36,321 students (Fall 2017), [7] which make up approximately one half of the city's population.

2010 census[edit]

As of the census[3] of 2010, there were 58,965 people, 22,759 households, and 9,959 families residing in the city. The population density was 2,435.6 inhabitants per square mile (940.4/km^2). There were 23,876 housing units at an average density of 986.2 per square mile (380.8/km^2). The racial makeup of the city was 84.5% White, 3.4% African American, 0.2% Native American, 8.8% Asian, 1.1% from other races, and 2.0% from two or more races. Hispanic or Latino of any race were 3.4% of the population.

There were 22,759 households of which 19.1% had children under the age of 18 living with them, 35.6% were married couples living together, 5.4% had a female householder with no husband present, 2.7% had a male householder with no wife present, and 56.2% were non-families. 30.5% of all households were made up of individuals and 6.2% had someone living alone who was 65 years of age or older. The average household size was 2.25 and the average family size was 2.82.

The median age in the city was 23.8 years. 13.4% of residents were under the age of 18; 40.5% were between the ages of 18 and 24; 22.9% were from 25 to 44; 15% were from 45 to 64; and 8.1% were 65 years of age or older. The gender makeup of the city was 53.0% male and 47.0% female.

Top employers[edit]

According to Ames's 2015 Comprehensive Annual Financial Report, the top employers in the city are:

#	Employer	# of Employees
1	Iowa State University	15,695
2	Mary Greeley Medical Center	1,287
3	City of Ames	1,226
4	Iowa Department of Transportation	920
5	McFarland Clinic	910
6	Hy-Vee	790
7	Ames Community School District	679
8	Danfoss	650
9	Wal-Mart	435
10	Ames Laboratory	432