HOUSE PRICE PREDICTION AND ANALYSIS CASE STUDY

1. Problem Statement

MUNA is a Real Estate Company that successfully launched a new branch in the state of Iowa in 2020. The company, which has been in the Real Estate Business for the past 30 years, noticed an unprecedented population growth in Ames city located in the State of Iowa.
In order to figure out if the city is worth investing in, this case study will analyze the sales price of houses in the past years and predict the possible current and the future price, based on factors such as Location, house design and Sale conditions.

2. Questions From This Case Study

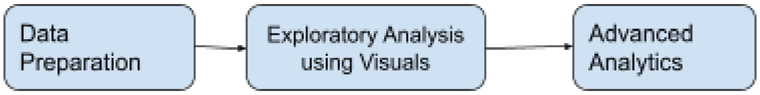
1. How does Location affect Sales Price?
2. Does Month and Season affect Sales Price?
3. What is the best Sales condition to use

3. Data Collection

This data was compiled by Dean De Cock for use in Data Science education

4. Data Processing and Analysis

To successfully process my data, i will carry out the following method:



Data Preparation:
The dataset which contains 1406 rows was split into train data with 81 columns and the test data with 80 columns. The datasets were imported as a table into the HomePrediction Database I created on Microsoft SQl Management studio. This was to

enable me to create views, a Virtual table that will enable me to visualize the Sales price with respect to Location, Construction types and House Features.

Exploratory Analysis With Visuals

This is the Descriptive Analysis stage, that answers the **WHAT** question.
The tool used for this analysis was Power BI. Before i started, I did some Data cleaning process which include:

1. Change of data types to ensure that the my analysis are accurate
2. Check for blank rows and remove them
3. Change the month Which were in numbers to month name using the DATE FORMAT function

Insights

1. The data has 1460 entries of houses sold between January 2006 to September 2010 with an average sale price of $180,000 within that time frame.
2. Over 79% of property were made in normal sales condition which means houses were completed and properly assessed before they were put up for sale.
3. House prices were over the roof in the year 2007, but a great fall in price was seen in 2010.
4. The price of a house located in the city of Ames in Iowa is twice the House Price in the city of Gilbert, a town in the state of Arizona.
5. The bigger the Area Square meter of a building, the higher the Sales Price Value
6. From my analysis, the sale value of one story building is generally higher than the 2 story building. This could possibly be because One story Buildings have a bigger Area Square meter and hence takes in more foundation and roofing materials for its construction, compared to 2 story Buildings.

Advanced Analytics

The goal of this stage of analysis is to get more insights using statistical tools.
I used the R programming Language for this analysis. Before i started, I imported these library packages;

1.corrplot to check for relationships
2. Odbc for connecting my database to my R studio
3. Dplyr and Tidyverse for data Manipulation and wrangling etc.

Then, I imported my Train and test dataset after connecting with my HomePrediction database from SQL.

Next, I selected only the columns that contained the numerical values and stored them as df_new.

```
df_new = select(TrainData, c(2, 4, 5,
18,19,20,21,27,35,37,38,39,44,45,46,47,48,49,50,51,52,53,55,57,60,62,63,67,
68,69,70,71,72,76,77,78,81))
View(df)
```

The names of the column selected include;
The house types, linear feet of street connected to property, Area Square meter, Overall Quality of the building, year the house was built etc.

Data cleaning

- Changed the data type from Character to numeric using the `as.numeric`

Function.

- Rename the IstFlrSf and 2ndFlrsf column to remove code error from my analysis

```
TrainData %>%
  rename(firstfloorsquarefeet = 44,
         secondfloorsquarefeet = 45,
         )
```

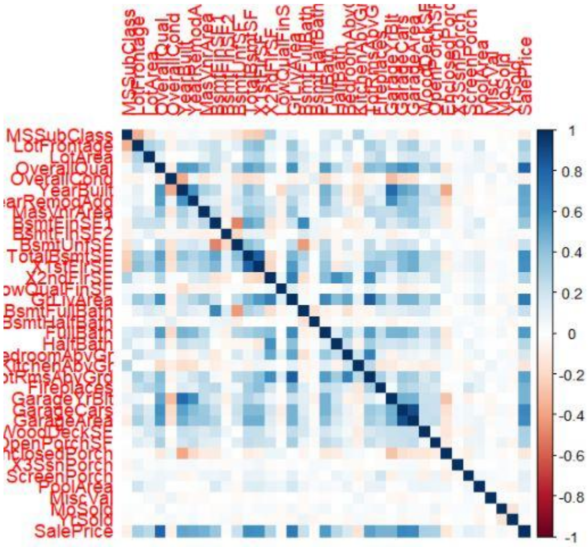
- Check for missing values and replace it with the median values using the code below

```
df_new <- df_new %>%
  mutate_if(is.numeric, function(x) ifelse(is.na(x),
median(x, na.rm = T), x))
```

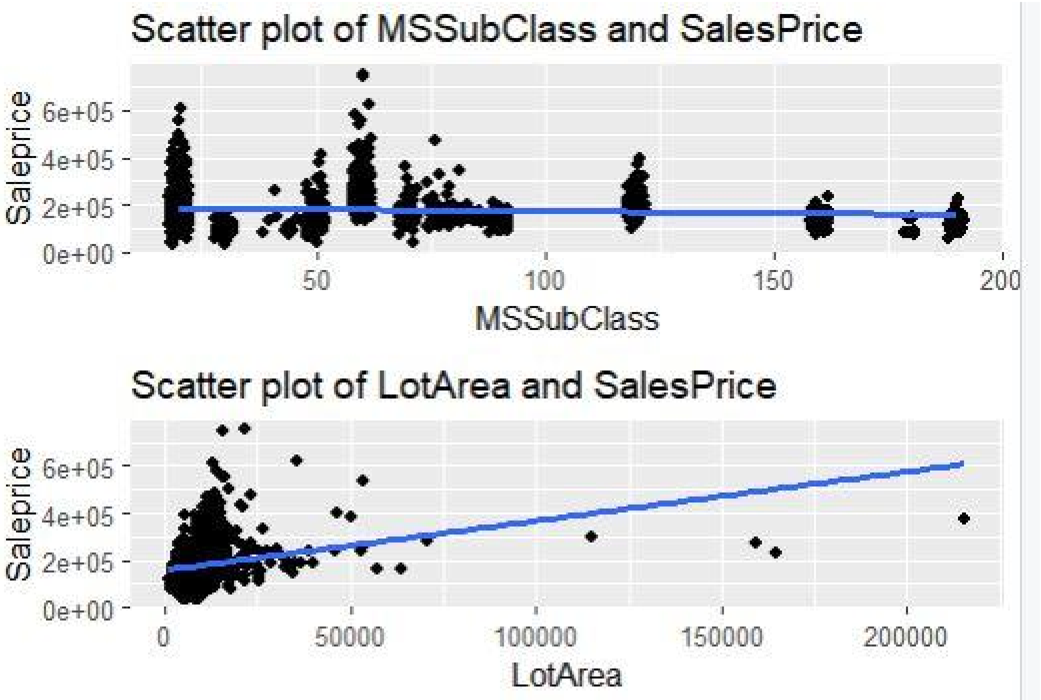
This will help for better model prediction results as I am using the numeric values which are not a normal distribution.

Exploratory Data Analysis (EDA)

- **Correlation Analysis**



From the corrplot above, the salesPrice has a strong positive correlation with House types, the Square Area, Overall condition of the building while having a strong negative correlation with month and year of sales, screen porch area and Value of miscellaneous features.



From the scatter plot fig above, shows the relationship between MSsub Class which represents House Types and the sales price. 1- 2 story buildings built in the 1940s but renovated were in more demand. Are square meter increases with the House Sales price

5. Model Building

I used the Multiple Linear Regression Algorithm to predict the sale price with more than one variable.

However, to get an accurate prediction result, I used the stepwise regression function to get the variables with most significance for my model.

Hypothesis:

H_0 = There is no significant difference between sales price and the multiple numerical variables

H_1 = There is significant difference betweetween sales price and multiple numerical variables

Next compute for the regression value between the Sales price and the numerical values and get the summary

```
reg =lm(SalePrice~ ., data= df_new)
summary(reg)


*******************RESULT***********************************

Call:
lm(formula = SalePrice ~ ., data = df2)

Residuals:
    Min      1Q  Median      3Q     Max 
-442865  -16873   -2581   14998  318042 

Coefficients: (2 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -3.232e+05  1.701e+06  -0.190 0.849317    
MSSubClass    -2.005e+02  3.449e+01  -5.814 8.03e-09 ***
LotFrontage   -1.161e+02  6.124e+01  -1.896 0.058203 .  
LotArea        5.454e-01  1.573e-01   3.466 0.000548 ***
OverallQual    1.870e+04  1.478e+03  12.646  < 2e-16 ***
OverallCond    5.227e+03  1.367e+03   3.824 0.000139 ***
YearBuilt      3.170e+02  8.762e+01   3.617 0.000311 ***
```

```
YearRemodAdd   1.206e+02  8.661e+01   1.392 0.164174    
MasVnrArea     3.160e+01  7.006e+00   4.511 7.15e-06 ***
BsmtFinSF1     1.739e+01  5.835e+00   2.980 0.002947 ** 
BsmtFinSF2     8.362e+00  8.763e+00   0.954 0.340205    
BsmtUnfSF      5.006e+00  5.275e+00   0.949 0.342890    
TotalBsmtSF           NA         NA      NA       NA    
`1stFlrSF`     4.591e+01  7.356e+00   6.241 6.21e-10 ***
`2ndFlrSF`     4.668e+01  6.099e+00   7.654 4.28e-14 ***
LowQualFinSF   3.415e+01  2.788e+01   1.225 0.220788    
GrLivArea             NA         NA      NA       NA    
BsmtFullBath   8.980e+03  3.194e+03   2.812 0.005018 ** 
BsmtHalfBath   2.490e+03  5.071e+03   0.491 0.623487    
FullBath       5.390e+03  3.529e+03   1.527 0.126941    
HalfBath      -1.119e+03  3.320e+03  -0.337 0.736244    
BedroomAbvGr  -1.023e+04  2.154e+03  -4.750 2.30e-06 ***
KitchenAbvGr  -2.193e+04  6.704e+03  -3.271 0.001105 ** 
TotRmsAbvGrd   5.440e+03  1.486e+03   3.661 0.000263 ***
Fireplaces     4.375e+03  2.188e+03   2.000 0.045793 *  
GarageYrBlt   -4.914e+01  9.093e+01  -0.540 0.589011    
GarageCars     1.679e+04  3.487e+03   4.815 1.68e-06 ***
GarageArea     6.488e+00  1.211e+01   0.536 0.592338    
WoodDeckSF     2.155e+01  1.002e+01   2.151 0.031713 *  
OpenPorchSF   -2.315e+00  1.948e+01  -0.119 0.905404    
EnclosedPorch  7.233e+00  2.061e+01   0.351 0.725733    
`3SsnPorch`    3.458e+01  3.749e+01   0.922 0.356593    
ScreenPorch    5.797e+01  2.040e+01   2.842 0.004572 ** 
PoolArea      -6.126e+01  2.984e+01  -2.053 0.040326 *  
MiscVal       -3.850e+00  6.955e+00  -0.554 0.579980    
MoSold        -2.240e+02  4.227e+02  -0.530 0.596213    
YrSold        -2.536e+02  8.454e+02  -0.300 0.764216    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36790 on 1086 degrees of freedom
Multiple R-squared:  0.8095,	Adjusted R-squared:  0.8036 
F-statistic: 135.7 on 34 and 1086 DF,  p-value: < 2.2e-16
```

Hence to select the variables with the most significance, I used the Step_reg function. This selects variables that have a P- value < 0.05, which means the variable with P-value, greater than 0.05 is taken out.

Here is the code for the step_reg analysis.

```
step_reg = step(reg)
summary(step_reg)
```

columns selected from the result are was used to build my prediction model using the regression model function (lm) on myTrain data.

```
model=lm(data=TrainData,SalePrice~MSSubClass+LotArea + OverallQual +
YearBuilt + BsmtFinSF1 + `1stFlrSF` + `2ndFlrSF` + BsmtFullBath +
FullBath + BedroomAbvGr +  KitchenAbvGr + Fireplaces + ScreenPorch)
summary(model)
```

Prediction:

```
prediction =predict(newdata=TestData,model)
```

Using my Prediction model above, I was able to get the following mathematical representation of how the predicted sales Price (the dependent variable) will be calculated using my independent variables.

Predicted Sales Price = -7.796 e5 - 1.94 e2 MSSubclass + 4.49e-1LotArea - 1.917 OverallQual + 0.011yearBult - 0.0132BsmtFinsf1 + 0.037IstFlrf + 6.807BsmtFullbath -6.5Fullbath +65.5Bedroomabovegrade - 3.814KitchenAboveGrade -3.033Fireplaces + 0.061ScreenPorch

RECOMMENDATIONS

- MUNA real estate should invest more at Northern part of Ames city in the state of Iowa
- There is high demand for House and property during the month of June and July
- Contract based Sales conditions should be discouraged because it returns less value
- Neighborhoods situated in the urban area should be prioritized more even if it costs more in investment but the returns are worth it .