**Logo

Description automatically generated**

**DSCI 5180**

**Introduction To the Business Decision Process**

**Final Project Fall 2021**

**Submitted to: Prof. Arunachalam Narayanan**

**Topic: Amex House Prices**

Submitted by: Preeti Chaudhry  
Student Id - 11559290

1. Data Source:

This dataset is available on Kaggle.com

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

1. Data Description:

Considering the data to be normally distributed.

The dataset has the details on different houses in the Amex town in Iowa State. The dataset is split into test and train dataset, however, for the purpose of this project we are only considering the training data. In the Training dataset there is a record of 1,460 different houses in 25 distinct neighborhoods in the Amex town. There are 81 variables which allow us to test various relationships with the dataset.

Key Insights from the data:

Mean of Sale Price (Column “CC”) – 1,80,921

Standard Deviation of Sale Price – 79,442

Median of Sale Price – 1,63,000

Below are the key variables which we will be using for our analysis amongst others:

| Variable | Description |
| --- | --- |
| Neighborhood | Physical locations within Ames city limits |
| SalePrice | The sale price of the house |
| GarageType | Type of Garage – Attached or Detached |

1. Project Details:

* Module 1: Normal Distribution (Percentile, distribution of means, and chance of occurrence if we assume normal distribution)

Determine 75th percentile of house prices with and without garage.

Solution:

Percentile indicates the value below which a given percentage of observations in a group of observations fall. 75th percentile of the sale price of the house is the value below which 75% of house prices with and without garage will be found. By applying the percentile formula in excel for the sale prices we get value as 213,000.

Hence, 75th percentile of house prices with and without garage is 213,000

* Module 2: Confidence Interval Estimation (Including Sample Size determination)

Estimating the population mean of sale prices with a 95% confidence interval.

Solution:

Since we want to calculate 95% confidence interval of sale prices of houses.

Mean = 180,921

Standard deviation s = 79,442.5

n = 1460

Degrees of freedom associated with the problem df = 1460 - 1 = 1459

α = 0.05

α /2 = 0.025

Calculate T-value using Ti-84 calculator (Invt function) we get T = 1.962

As we know that the general form of the confidence interval for the sample mean when s is unknown



Finding value of confidence interval from Ti-84 calculator we get 176,846 and 184,996

Thus, we are 95% confident that the mean sale price of houses lies between 176,846 and 184,996

* Module 3: Inferences from data (Hypothesis testing, i.e., confirming or checking a claim made about the data.)

To test hypothesis whether the sale price of the houses with an attached garage is more than the houses with detached garage.

By running the test on data with mean of attach garage = 202,892.66

Standard Deviation = 77,146.622

n (for attached garage) = 870

Conduct the test at 0.05 Significance level.

Solution:

* Define Null and Alternative Hypothesis i.e.

Null Hypothesis: Houses with attached garage does not have higher prices.

Alternative Hypothesis: Houses with attached garage have higher prices.

μa > μo

* Here we are interested in determining that prices of attach garage are higher, therefore it’s a one-sided hypothesis.
* Since the population standard deviation is unknown, we will be using T statistics with α 0.05

Using Invt function, t-critical value = 1.646

Find T-statistic using formula or Excel, we get this value as 16.516

* Make the decision

Since, the T-value falls in the rejection region, it indicates that the sample mean is too far from the hypothesized value to believe it is due to ordinary sampling.

The null hypothesis must be rejected in favor of alternative hypothesis.

Hence, Sale price of house with attached garage is higher than with detach garage.

* Module 4: More Inferences from data (Multiple samples)

Comparing the mean sale prices of 2 neighborhoods Northridge and Timber in the city.

Mean sale price of Northridge = 335,295.3, Standard deviation = 121,412.7

Mean sale price of Timber = 242,247.4, Standard deviation = 64,845.65

a. Calculate at 95% confidence interval for the difference in average sale price between the two neighborhoods.

b. Test to see if Northridge neighborhood has a higher mean sales price than Timber neighborhood. Use α = 0.01.

Please refer the Excel: Raw\_Module\_dataSheet\_FinalProjectPreeti Chaudhry\_1159290

Tab: NorthridgeAndTimber\_Data

Column: ‘M’ (Neighborhood)

Solution:

Given: N1=41, N2=38

S1 = 121,412.7, S2 = 64,845.65

Mean X1 = 335,295.3 and X2 = 242,247.4

Assuming that the data follow a normal distribution and that the population variances are equal, and estimate the population standard deviation with the sample standard deviations. Since we want a 95% confidence interval, we need to find the value of



In the data there are 41 records of neighborhood Northridge and 38 of Timer.

Therefore, Df = N1 + N2 - 2

= 41 + 38 - 2

= 77

Thus, t-critical value = 1.991

95% confidence interval for the difference in average sale prices between two neighborhoods is given by



By calculating the mean sale price, we get lower range as 49,887.218 and upper as 136,208.582

Thus, we are 95% confident that the true mean difference in sale price of houses is between 49,887 and 136,208

* Define Hypothesis

Null Hypothesis: Northridge neighborhood doesn’t have higher mean sales price than Timber neighborhood.

Alternative Hypothesis: Northridge neighborhood has a higher mean sales price than Timber neighborhood.

μa > μo

* We are interested to know whether the average sales price of Northridge is higher than the average sales of Timber. Thus, the alternative hypothesis is one-sided and the test is one-tailed.
* Finding test statistic



Calculating t-critical from calculator we get 2.376

By calculating the value of T-statistic, we get it as 4.2

The value of the test statistic does fall in the rejection region. There is sufficient evidence at the 0.01 level to conclude that the average sales price for Northridge are significantly higher than the average sales of Timber neighborhood.

* Module 5: Regression analysis (Both simple and multiple, apart from basic ANOVA)

To find how sale price of the house varies with GarageArea, LotArea and YearBuilt using regression.

A multiple regression model is a linear regression model using two or more explanatory variables to predict a response variable.

Independent Variables: x1 = GarageArea, x2 = LotArea, x3 = YearBuilt

The variable we want to predict is called the dependent variable, y = Sale price of the House



Running multiple regression in excel by using independent variables (x) and dependent variable (y) at 95% confidence interval, we get the below output

| SUMMARY OUTPUT |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| Regression Statistics | |  |  |  |  |  |
| Multiple R | 0.696338409 |  |  |  |  |  |
| R Square | 0.484887179 |  |  |  |  |  |
| Adjusted R Square | 0.48382582 |  |  |  |  |  |
| Standard Error | 57075.67502 |  |  |  |  |  |
| Observations | 1460 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | df | SS | MS | F | Significance F |  |
| Regression | 3 | 4.4648E+12 | 1.48827E+12 | 456.8550841 | 3.9066E-209 |  |
| Residual | 1456 | 4.74311E+12 | 3257632679 |  |  |  |
| Total | 1459 | 9.20791E+12 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| Intercept | -1508421.996 | 109825.8312 | -13.73467407 | 1.86162E-40 | -1723855.756 | -1292988.236 |
| LotArea | 1.428162138 | 0.15273732 | 9.350446511 | 3.17458E-20 | 1.128553433 | 1.727770843 |
| YearBuilt | 809.8157853 | 56.55642934 | 14.31872193 | 1.31812E-43 | 698.8749976 | 920.7565731 |
| GarageArea | 164.8265148 | 8.121828925 | 20.29426085 | 7.77777E-81 | 148.8947788 | 180.7582507 |

To consider model to be statistically significant p < 0.05

Here p-value ≈ 3.9066 x 10^-209 we reject the null hypothesis and conclude that there is significance evidence at the 0.05 level of significance to support the claim that this multiple regression model fits the data well and can be used for predictions.

Inferences for the Regression Model:

* All the independent variable used in the model are significant.
* We are 95% confident that 1 sq. ft. increase in the GarageArea will result in the increase of the sale price of the house between $148.89 and $180.75
* The value of adjusted R square is far from 1 and hence other variable can be used to improve the model fitment.