ADTA 5130 SPRING 2022   
DATA ANALYTICS 1 FINAL PROJECT REPORT

**Ames Housing Data**

Professor: Denise R. Philpot, PhD

Individual Project: Manisha Adulla (11525405)

University of North Texas   
Spring 2022   
Date Submitted: 05/12/2022

# Introduction:

Statistical Analysis on a dataset is one of the most conventional approaches to understand the mathematical implications. ANOVA is one such statistical techniques in which the numerical variables are analyzed against a categorical group. The Linear Regression model using statistics are designed to predict the dependent variable which is continuous. The insights generated performing these statistical approaches supports in building better predictive models.

# About the Dataset:

The given dataset is about the Ames housing to analyze the buying and selling activities in the market. The dataset is about 25 fields and 2930 records.

Importing Libraries and reading the given excel dataset into python:

Graphical user interface, application

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with medium confidence

Text

Description automatically generated

It has Categorical variables - Building Type, Neighborhood, and House Style, Binary Variables and Numerical Variables.

Below are the details of datatypes for each variable:

Table

Description automatically generated

Description of each Column in the dataset:

* Order – Unique id number.
* Lot Area - Size of lot.
* Neighborhood - Name of neighborhood.
* Bldg Type -Type of the Building. Single family, duplex, condo, townhouse, townhouse end unit.
* House Style- Style of the House. 1 story, 1.5 story, 2 stories
* Year Built – Year when house was built.
* BsmtFin SF 1- Finished square footage in basement area 1.
* BsmtFin SF 2- Finished square footage in basement area 2.
* Bsmt Unf SF- Unfinished square footage in basement.
* Total Bsmt SF - Total square footage in basement.
* Central Air- Does the house have central air conditioning. Values are in the form of Yes or No.
* Gr Liv Area - Gross living area of house
* Bsmt Full Bath - Number of full baths in basement
* Bsmt Half Bath - Number of half baths in basement
* Full Bath - Number of full baths above ground
* Half Bath - number of half baths above ground
* Bedroom AbvGr - Number of bedrooms above ground
* Kitchen AbvGr - Number of kitchens above ground
* TotRms AbvGrd - Total rooms above ground
* Fireplaces - Number of fireplaces
* Garage Cars - Number of garage spaces for cars.
* Garage Area - Square footage of garage
* Mo Sold – Month when house was sold
* Yr Sold – Year when house was sold
* SalePrice - Sales price of the house.

# Exploratory Data Analysis:

Data Cleaning: Used Python to clean the data.

1. Checking for missing values:

Text, table

Description automatically generated

Columns which have null values are:

Graphical user interface, application, Word

Description automatically generated

* Replacing the missing values for the above columns with **Median** value.

Graphical user interface

Description automatically generated

All the places where there is no value in the above Columns, they are replaced by Median of those columns. Now there are no nulls or missing values.

Missing values are replaced by Median value of those columns instead of Mean to avoid dependency of data on outliers. Mode is used for categorical data.

1. Export the cleaned data from Python to an Excel file



# **Visualization Graphs:**

Below are some charts to visualize the given dataset in a user-friendly manner. Here Tableau is used to create various graphs. Imported the given dataset to Tableau to analyze the data graphically.

Graph 1: Below Stacked Bar graph shows the Total Sales Price for each Building Type based on the year sold.

Chart, bar chart

Description automatically generated

Graph 2: Below Line graph describes the Average Gross Living Area for each Building Type.

Graphical user interface, chart, line chart

Description automatically generated

Graph 3: Below Bubble graph illustrates the total number of houses for each House style with and without Central Air.

Chart, bubble chart

Description automatically generated

# **Research Question 1:**

### ANOVA Analysis 1:

Are mean Sales Prices for houses with and without Central Air same?

Null Hypothesis - mean Sales Prices for houses with and without Central Air are same.

Alternate Hypothesis - mean Sales Prices for houses with and without Central Air are not same.

* Filter the prices with and without Central Air into two columns.

Table

Description automatically generated

* Perform ANOVA Analysis on the above two columns.



**Analysis**: There is significance difference in the average sale price with respect to the categorical column ‘Central Air’. It is observed that the average sale price when the ‘Central Air’ being Yes is almost double compared to the ‘Central Air’ being No. Considering the performed ANOVA analysis, the p-static value is lesser than 0.05. So, we reject the Null hypothesis.

### ANOVA Analysis 2:

Are mean Sales Prices for different house styles same?

Null Hypothesis - mean Sales Prices for different house styles are same.

Alternate Hypothesis - mean Sales Prices for different house styles are not same.

* Filter the prices for different House Styles into columns.

Table

Description automatically generated

* Perform ANOVA Analysis on the above 8 columns.



**Analysis**: Considering the performed ANOVA analysis, the p-static value is lesser than 0.05 and F value is greater than F critical value. So, we reject the Null hypothesis.

# **Research Question 2:**

Correlation Analysis:

A picture containing text

Description automatically generated

A picture containing scatter chart

Description automatically generated

The correlation analysis is performed on the columns of the data set and above is the heatmap generated with their correlation values. The color pattern has the importance of understanding the correlation values. The box with the darkest blue values is highly correlated. The lesser the intensity of blue color, lesser the correlation values that the two columns possess.

Text

Description automatically generated

In the given dataset, the columns {Garage Area, Garage Cars} are highly correlated with a correlation coefficient of about 0.889 followed by the columns {Gr Liv Area, TotRms AbvGrd}, {Gr Liv Area, SalePrice} with correlation coefficient values 0.807772 and 0.706780 respectively.

### Regression Analysis:

How to predict the Sales price?

A linear regression model was performed using all the numerical variables of the dataset to predict the Sales price of the houses.

Graphical user interface, application, table, Excel

Description automatically generated



**Analysis**: The R Square for the performed Regression model is 0.7547, which can be interpreted as the model explains 75.47% variability of the sales price when predicted with 16 independent variables. The R Square value ranges from 0 to 1 and closer the value to 1 better the model.

# **Conclusion:**

The imported Housing Data file is preprocessed, and various Exploratory Data Analysis (EDA) techniques are performed. The implemented ANOVA Analysis 1 on the Sale price against the categorical variable Central Air depicted the significant change in variance and the p-value obtained is less than 0.05. Hence the Null hypothesis is rejected. The performed ANOVA Analysis 2 on the Sale price against the categorical variable House Style with the p-value less than 0.05 rejecting the Null hypothesis. The major limitation of ANOVA is that multivariate grouping analysis is time consuming and losing the variable level interpretation. The designed Linear Regression model obtained the R Square value of 75.47% which is a good fit. One limitation with the R Square evaluation is that it is not the only metric to interpret the performance of the model.