**Interpretation of Data**

**Assignment-2**

**Submitted By: \_\_\_\_\_\_\_\_\_\_\_\_\_**

**Student No : \_\_\_\_\_\_\_\_\_\_\_\_\_**

**Part-2 : Using PropertyAssignment2.csv dataset**

Based on the provided use case for this assessment, A detailed study has been carried out with necessary Exploratory Data Analysis, Feature Engineering of Data has been performed after analysing several variables containing in the dataset of property prediction of houses sold in Ireland.

**Data Understanding:**

**Brief Description about Data :**

This Dataset has totally 16 Variables, with total records of more than 43000 data,

**Geohash.** This is a public domain, geocode system which encodes a

geographic location into a short string of letters and digits. Can be less accurate

than Latitude and Longitude but can be, more efficient to work with. May be more robust for duplicate detection.

**lat :**  Latitude Co-ordinates

**lon:**  Longitude Co-ordinates

**sale\_date**: Date of the sold property

**price** : Price of the house

**formatted\_address:**  Formatted Address, which has been modified based on the **Raw Address**

**raw\_address**: Raw Address- original address without any modification

**post\_code** Post Code, only for Dublin Limits

**county**: Total list of county in Ireland.

**num\_bedrooms:**  Number of bedrooms in house(1,2,3,4,5,6,7,8,14)

**num\_bathrooms** : Number of bathrooms in house (1,2,3)

**property\_description:**  Description of property(New Dwelling House or Apartment, Second-Hand Dwelling house /Apartment, Teach) 3 types , most are common between the **Formatted Description**

**property\_size\_description:**  Size of the property in Meter Square ( greater than 125 sq. metres, greater than or equal to 38 sq. metres and less than 125 sq. metres, less than 38 sq. metres)

**formatted\_description** : The Detailed Description of type of house around 10 types of house such as ( House, Apartment, Duplex, Townhouse, Second-Hand Dwelling house /Apartment, Terraced House, End of Terrace House, Semi-Detached House, Terraced House, Semi-Detached House)

**not\_full\_market\_price** Whether it is full market price or not(True /False)

**vat exclusive** Whether vat is included or excluded ((True /False))

**Business Understandings:**

The Ultimate aim of this Data is to predict the House price prediction with the available data resources., To perform an effective prediction of the Property price and to fetch better results of the model, data must fully refine and Feature Engineered.

**Exploratory Data Analysis:**

As mentioned in the Business Understanding**,** few techniques for EDA (Exploratory Data Analysis) are Such as Removing Missing Values, Duplicates, Deleting Variable which are repetitive( with same meaning), Restructuring the entire data, Data Augmentation and lot of techniques are to be followed before handing over to the Data Scientists or feeding to the model.

**Study Performed:**

Individual Study of Variables has been done, and found several variables which are repetitive, and they are not much informative

**Removing Duplicate Variables :**

Duplicate variable of the entire dataset has been analysed using the R programming, and in this dataset around 543 Duplicates records has been eliminated at the first stretch., Found other duplicates values based on the rows but they are essential records, and such are kept not disturbed, below section of variables removed (Geohash) explains the conflict.

**Variables Removed :**

1. **Geohash :** Which has the information of the Both latitudes and Longitude, Majority of missing values and the values has a huge duplicate variables. For example, as it contains Latitude and Longitude ( it calculates for the locality and the Geohash has is similar with near locality each other which brings in the conflict with other variables, as they are very different Address, Price , Date, House details, Bedroom/ Bath. And it has same Geohash Values.
2. **Latitude , Longitude:** These are the points like the **Geohash,** and it has also been removed
3. **Formatted \_Address:** This is the address which has been modified based on the raw \_address and it has missing values, incomplete and empty values, on comparing with the Raw-Address, which is much more precise and perfect.
4. **Post-Code :** This Feature is only available for the Dublin County and its nearby localities like (Dublin, 3,5,6,8, 16, 15, etc) And this locality has been added with the respective address to make the process effective. As it has roughly very less data points comparatively with 40000 records of other county and this have the columns like the properties type(second-hand dwellings) on the Dublin County alone.

**Example Code:**

drop1 <- c("formatted\_description”, property\_description" )

df3 <- df2[,!(names(df2) %in% drop1)]

which drops the specific column

**Detailed Analysis :**

The main Aim of the data is to predict the prices of the housing ,and the data set will be in the form of Numerical and Categorical values are to be changed, to make the ease of the Data Scientist so that they have their flow of modelling process smoothly and they can avoid less time in analysing further and doing Encoding Techniques to convert (Categorical values into Numerical) such as One Hot Encoding, Label Encoding, etc.

1. Bedroom, Bathroom categorical variable has been changed to binary format (0 and 1) , there are (1,2,3,4,5,6,7,8,14 Bedrooms) and (1,2,3) Bathrooms available and they are made as an individual Column with (1 as Yes and 0 as No) based on each category has been followed.
2. Property Size description has also created as new column (greater than 125 sq. metres, greater than or equal to 38 sq. metres and less than 125 sq. metres and less than 38 sq. metres) with (1 as Yes and 0 as No) based on each category has been followed.
3. Property Description (variables has been modified as the individual columns) House, Apartment, Duplex, Townhouse, Second-Hand Dwelling house /Apartment, Terraced House, End of Terrace House, Semi-Detached House, Terraced House, Semi-Detached House with (1 as Yes and 0 as No) based on each category has been followed.
4. Not Full Value and Vat exclusive are Boolean values (TRUE/FALSE) which is converted into (1 For True and O for False)

**Example of Code:**

df2$Is\_1\_Bedroom<- ifelse(grepl("1 Bedroom", df$num\_bedrooms), "1", "0")

df2$Is\_2\_Bedroom<- ifelse(grepl("2 Bedrooms", df$num\_bedrooms), "1", "0")

this function looks for the specific variable in the column if present it makes a new column with respective title and in (Binary) 0 and 1

**Excess Columns Removal :**

After splitting the individual categorical variable into multiple column variable, next we must remove the categorical column and I have removed few columns like Bedroom, Bathroom, Property Size description, Property Description and Not Full Value and Vat exclusive.

**Blank Values :**

**Few majority of the blank values of the property belongs to the** second-hand dwellings and they are available in property section, and they are clearly mentioned in the r program comments.

**Shuffling the data and fill N/A for Blanks**

Shuffling the data has been done and the Blanks which are not appropriate is marked NA and then it has been dropped.

**Part 3- Load in the csv file from Moodle called “CarsAssignment2.csv” into R Studio. Clean this dataset**. **Only full rows must be kept (No missing values). Explain every step of the cleaning process.**

1. Dataset is loaded and the basic study of the data has been carried out,
2. As the Data set lacks in important Fields which are missing , at this point we cannot do any data augmentation for filling in the blanks , which may create the data in a bad shape and model performance can reduce.
3. Process is to find the Blanks and replace them with N/A and then drop them using .is dropna () function.
4. In the Date and Time, it has in different format and changing of the time stamps correctly
5. Shuffling the data

**Part-1 : [Using PropertyAssignment2.csv dataset]**

1. **Filter all the houses from the dataset that were sold in 2017 that have 3 bedrooms and 2 bathrooms and save it as a data frame.**

Select the (date , bedrooms, and Bathrooms ) save in a data frame

And filter based on the condition of ( 3 bedrooms and 2 bathrooms, date on 2017)

1. **b) Filter all of the data from the dataset for premises that were sold in Cork or Galway and are greater than 125 square metres. Order the data so that the price is listed from highest to lowest and save it as a data frame**

select the county, property size , price variable ) filter based on the properties on specific condition and save result In data frame similarly for the other condition and combine the both data frame using rbind function and we need to do the descending based on the price

**code:**

data3 <-select(data1, county, property\_size\_description, price ) # %>% filter(property\_size\_description > "greater than 125 sq metres")

select\_home <- filter(data3, county == "Cork" & property\_size\_description == "greater than 125 sq metres")

select\_home1 <- filter(data3, county == "Galway" & property\_size\_description == "greater than 125 sq metres")

final <-rbind(select\_home , select\_home1 )

desc<- arrange(final, desc(price))

desc

1. **Create a data frame that contains the average price, maximum price and minimum price of all of the new dwelling houses or apartments sold in Dublin 18, Dublin 16, Dublin 15, Dublin 8 or Dublin 4.**

Select the county , and filter based on the conditions and compute the max, min, average(mean(), min(), max()

1. **Using the “ggplot2” package, create a bar chart with all the property sold in the different counties in Ireland. Colour each of the bars based on the type of properties sold in each county.**

Select the county, formatted\_proprties, and use the ggplot lib for bar plot with specific x, y values of county and formatted\_properties.

1. **Remove all of the columns from “num\_bedrooms” to “not\_full\_market\_price” in the dataset and keep all of the rows that are from either Dublin, Cork or Galway and are a Semi-detached house and arrange the prices from lowest to highest.**

Remove columns from 10 to 15 from data , store in separate data frame, filter the county, formatted\_properties store in data frame and combine and make a lowest to highest price calculation(ascending).

1. **Filter all of the houses greater than €200,000 from Galway, Roscommon, Mayo or Sligo. Use ggplot to create a line graph. The line graph should consist of 4 lines representing the price of each of the 4 counties over the years in the data set. The lines must all be different colours on the plot.**

Select the price, county and filter based on the condition of price and the specific county and using ggplot for graphs to plot the line graph.