***Shape

Description automatically generated with medium confidence***Name: Mohammad Siddiq

Reg. no.: 12018057

Roll no.: RK20CHA13

Course: EDA Project

Course Code: INT353

CA1

Introduction to EDA

**Exploratory Data Analysis (EDA)** is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. It is the very first step in data analytics stream and was developed by ***“John Tukey”*** , in 1970s. EDA gives a robust understanding of the data, issues associated with either the info or process.

TYPES OF EXPLORATORY DATA ANALYSIS:

1. Univariate Non-graphical
2. Multivariate Non-graphical
3. Univariate graphical
4. Multivariate graphical

Python and EDA are often used together to spot missing values in the data set, which is vital so you’ll decide the way to handle missing values for machine learning. There are often many open-source tools in existence which automate the steps of data modelling, data cleaning and data visualization such as, Excel, Weka, Tableau, etc.

EDA in python is done usually by :-

* Pandas
* NumPy
* Matplotlib
* Jupyter Notebook
* Seaborn

Context

***Melbourne Housing Market***

This dataset contains 34,858 rows of observations and 21 columns of variables (with 8 variables as characters or strings) for the house market pricing in Melbourne. Every row of the dataset denotes information of the property, price it was sold for, seller information no. of bedroom, bathrooms and cars, location(latitude and longitude, formal address), land size and others.

Acknowledgements

This data was scraped from publicly available results posted every week from Domain.com.au, I've cleaned it as best I can, now it's up to you to make data analysis magic. The dataset includes Address, Type of Real estate, Suburb, Method of Selling, Rooms, Price, Real Estate Agent, Date of Sale and distance from C.B.D.

Why did I choose this Dataset ?

There was no particular reason for choosing this dataset initially but later I found out some interesting things about this dataset.

Melbourne is currently experiencing a housing bubble (some experts say it may burst soon). Maybe someone can find a trend or give a prediction? Which suburbs are the best to buy in? Which ones are value for money? Where is the expensive side of town? And more importantly where should I buy a 2-bedroom unit?

From this dataset we can find out the address of the house, no. of rooms in’t , expensive or cheap as compared to nearby location and area of the house, geographical location, etc. in order to make a right decision before buying a house.

Abbreviations Used for in the Columns

Suburb: Suburb

Address: Address

Rooms: Number of rooms

Price: Price in Australian dollars

Method:

S - property sold

SP - property sold prior

PI - property passed in

PN - sold prior not disclosed

SN - sold not disclosed

NB - no bid

VB - vendor bid

W - withdrawn prior to auction

SA - sold after auction

SS - sold after auction price not disclosed.

N/A - price or highest bid not available.

Type:

br - bedroom(s)

h - house,cottage,villa, semi,terrace

u - unit, duplex

t - townhouse

dev site - development site

o res - other residential

SellerG: Real Estate Agent

Date: Date sold

Distance: Distance from CBD in Kilometres

Regionname: General Region (West, North-West, North, North-east …etc.)

Propertycount: Number of properties that exist in the suburb.

Bedroom2 : Scraped # of Bedrooms (from different source)

Bathroom: Number of Bathrooms

Car: Number of carspots

Landsize: Land Size in Meters

BuildingArea: Building Size in Metres

YearBuilt: Year the house was built

CouncilArea: Governing council for the area

Lattitude: Self explanitory

Longitude: Self explanitory

What are the insights I want to retrieve

from the dataset?

* Room availability and average price are two most essential factors regarding to the booking. It firstly determines the income for the hosts. Secondly, it is important for the benefits for itself, because benefits of a person is dependent on the income and amount.
* Cleaning Data: For clustering, your data must be indeed integers. Moreover, since k-means is using Euclidean distance, having categorical column is not a good idea. I normalized my dataset which would have a certain set of variables chosen. All the nonvalues were either imputed or removed in the previous step. The variable ‘Date’ in its standard format will not be fed well in the model so I made a new column which encodes the month of the date present and dropped the former.
* Data processing and descriptive analysis. The goal of this analysis is to predict the Price variable, summarize each variable in the data set and provide simple descriptive statistics using tables/plots. You do not have to use the original scale of the variables. Any transformations, such as log, square root etc. are allowed. The variables come in different forms (text, date, categorical, continuous, etc.), you should provide a detailed discussion on how to transform them into a format that can be analyzed by regression models.
* Perform regression models to predict the Price variable.
* What property types are the most expensive?
* Most affordable price range in the city? Which house is most expensive and least expensive and what is the average price range for the house in a particular city