GROUP 13:

**Time series - Zillow**

1. Business Understanding

1.1 Introduction

* + We introduced the real estate industry in the USA, its valuation etc
  + We then proceeded to talk about the ROI of residential and commercial returns.
  + We showed why it’s important in making investments decisions.

1.2 Problem statement

* + We can up with the problem statement where a stakeholder aims to
  + construct residential homes in the USA.
  + To achieve this, we were tasked to use time series analysis using zillow
  + dataset to find the top 10 locations where they can have the highest
  + ROI and have the lowest price volatility.

"The stakeholder for this project is a real estate investment firm aiming

to construct residential homes in the United States' most lucrative locations,

ensuring a high return on their investment. To achieve this goal, the project

focuses on conducting a time series analysis utilizing a Zillow dataset that

encompasses various locations across the country.

Specifically, the project involves analyzing house sale prices spanning

from 1996 to 2018. By leveraging this data, the aim is to identify the top

five locations that offer the highest Return on Investment (ROI) potential.

Additionally, considering the stakeholder's risk-averse nature, the project

aims to recommend locations characterized by low price volatility.

This quality will enable the model to provide reliable predictions,

enhancing the stakeholder's ability to make informed investment decisions.

Overall, the project centers on utilizing time series analysis techniques and

the Zillow dataset to pinpoint the top ten locations in the US that not only

offer a high ROI but also exhibit low price volatility, aligning with the

risk preferences of the real estate investment firm."

1.3 Metrics of Success:

* + We will use MAPE(Mean Absolute Percentage Error) here.
  + MAPE is chosen due to its ability to provide weighted error values
  + It is good at handling outliers.
  + By utilizing MAPE, we aim to obtain a comprehensive evaluation

of the model's performance that accounts for both the magnitude

of errors and the relative proportion, they represent.

* + This metric offers a more robust assessment and ensures that

outliers do not unduly influence the perception of the model's

effectiveness in generating accurate predictions.

1.4 Alternative:

* + Use RMSE (Root Mean Squared Error) instead
  1. Problem Questions?
* What are the 10 best locations to invest/construct in the USA?

(We use the zipcode as the location point)

* What makes these locations valuable?

(Is it the city, metro, the state, the county, ROI, value of properties)

* What affect does urbanization have on value of houses
* The locations with the highest/lowest volatility?

(this is to ensure that the value of property remain stable)

(Use Zipcode and the duration of time passed)

Can we predict the value of property?

(The point of this project)

1.6 Objectives?

* Provide effective real estate investment.

recommendations to the stakeholder.

(Through predicting what the value of property may be)

* Increase the real estate investor’s customer base.

(Through constructing in areas with stable and growing value

we will attract survey investors/ home buyers who treat their homes as assets.)

2.0 Data Understanding

* + We explore the various columns.
  + The dataset utilized in this project consists of historical median house prices sampled from various states in the United States. The data spans a period of 22 years, specifically, from April 1996 to April 2018. The information was obtained from the official website of Zillow, a prominent real estate and rental marketplace.
  + The dataset comprises a total of 14,723 rows and 272 columns.
  + Among these columns, four are categorical variables, while the remaining columns are numerical in nature.

Here is a brief overview of the column names and their descriptions:

\* `RegionID`: A unique identifier assigned to each region.

\* `RegionName`: The names of the regions, typically representing zip codes.

\* `City`: The corresponding city names associated with the regions.

\* `State`: The names of the states to which the regions belong.

\* `Metro`: The names of the metropolitan areas to which the regions are affiliated.

\* `County Name`: The names of the counties where the regions are located.

\* `Size Rank`: The ranking of zip codes based on their level of urbanization.

\* `Date Columns (265 Columns) `: These columns represent median house prices recorded across different years.

By utilizing this dataset, the project aims to extract insights and patterns from the historical median house prices in order to predict the top investment locations with the highest Return on Investment (ROI).

3.0 Data Preparation

* + Imported the libraries
  + Loaded the data

3.1 Data cleaning

1. Completeness:

* + Check for missing datasets.
  + Display the missing data.
  + Metro had 1043, filled with 'missing' as can be an important indicator.
  + Date had 1039, interpolated them.

2. Consistency:

* + Check for duplicates:
  + There were no duplicates

3. Validity:

* + Make sure that columns is accurate and appropriate
  + Changed 'RegionName' to 'Zipcode'

4. Uniformity:

* + Make sure that the datatype are appropriate for the analysis
  + Converted Zipcode(integer) to a string that is categorical
  + Zipcode is attached to a physical location.
  + This means that, it should be treated as a location(categorical)
  + Also made them all have 5 digits.

Alternative, treat it as a numerical?

* + Multicollinearity?
  + Outliers?

4.0 Data Preprocessing

* Added the columns ROI and CV
* Reshaped the dataset and using the melt data function
* Made a new column (Date) using the melt dataset and made it the index
* Renamed value column from melted dataset to median house price for easier understanding.

1. Exploratory Data Analysis
2. As per our objectives, we will use EDA to determine:

The top 10 zipcodes/locations with the highest ROI:

* We will analyze the ROI values for different zipcodes/locations and identify the top 10 areas with the highest returns on investment.

2. The zipcodes/locations with high price volatility:

* We will examine the price volatility of properties in different

zipcodes/locations to identify areas with high fluctuations in

property prices. This information will help us find locations with stable

property values.

3. The cities/states with the highest valued property:

* We will explore the dataset to determine which cities or states

have the highest property values. This analysis will provide insights

into areas where property prices are generally high.

4. Whether urbanization affects property values:

* We will investigate the relationship between urbanization

and property values. By analyzing relevant variables such as

population density, city size, or proximity to urban centers,

we can determine whether urbanization has an impact on property values.

5. The trend of property prices over the years:

* We will analyze the historical trend of property prices over time.
* This analysis will help us identify any patterns, trends, or fluctuations

in property prices that may be useful for predicting future values.