TEAM 3 PHASE 4 PROJECT CRISP - DM REPORT

PROJECT:

A TIME SERIES ANALYSIS OF REAL ESTATE PRICES IN TOP 5 ZIP CODES

TEAM MEMBERS:

- **Anthony Ngatia**
- **Jessyca Aperi**
- **Joy Chepchumba**
- **Naomi Rotich**

DATE: 20 / 06/ 2023

BUSINESS UNDERSTANDING

BUSINESS OVERVIEW

Real estate investment is a lucrative and dynamic industry that requires careful analysis and decision-making. The fictional real estate investment firm is seeking guidance on identifying the top 5 zip codes for investment opportunities. To address this question, historical data from Zillow Research is utilized. The dataset contains information on various attributes, including RegionID, RegionName, City, State, Metro, SizeRank, CountyName, and value (real estate prices).

BUSINESS OBJECTIVES

To identify the top 5 zip codes that offer the best investment potential in terms of real estate prices.

Specific Objectives

- .To analyze Historical Data: The project involves analyzing the historical data of real estate prices across different zip codes. This includes understanding the trends, patterns, and fluctuations in property values over time.
- To identify promising Zip Codes: Using the analysis of historical data, the project aims to identify the zip codes that have shown consistent growth, stability, or potential for future appreciation. These zip codes are considered the most favorable for investment.
- To consider location factors: In addition to the historical performance, the project also takes into account location-specific factors such as city, state, and metro. This information helps assess the overall desirability and attractiveness of the investment opportunities.
- To evaluate market SizeRank: The SizeRank attribute provides insights into the relative size and competitiveness of the real estate market in each zip code. This factor helps gauge the potential opportunities and risks associated with investing in a particular area.

Business Success Criteria

To determine the top five best zip codes to invest in for a real estate investment firm.

ASSESSING THE SITUATION

Resource Inventory

Datasets

- Zillow Housing Dataset - `time-series/zillow_data.csv`

Software Used

- Google Colab
- Pandas
- Numpy
- GitHub

ASSUMPTIONS

The data provided is correct and up to date.

CONSTRAINTS

Forecasts are solely based on historic monthly returns, and past performance does not necessarily predict future results.

PROJECT PLAN

The project was conducted by a team of four members. The project was divided into major sections i.e. Data importation and analysis, Data Understanding and cleaning, Data Modelling and forecasting, Non-technical presentation and reporting. The teammates were allocated tasks and expected to deliver within the timelines specified.

DATA MINING GOALS

Our data mining goals for this project are as follows;

- To import and download the dataset from Zillow Housing: - The datasets were already provided and hence we did not perform data scrapping.

DATA MINING SUCCESS CRITERIA

Our success will be measured by the following criteria.

- Downloading the datasets from the data sources i.e.Zillow Housing Website
- Loading the datasets successfully onto the Google Colab workspace

DATA UNDERSTANDING

Overview

For this project, we are using the available dataset in Zillow Housing Website.

The dataset is the Zillow Housing Data. The link is attached below;

https://www.zillow.com/research/data

DATA DESCRIPTION

The column definitions are presented below:

- RegionID -This is unique Id for the Regions
- SizeRank -This is the ranking done based on the size of the region
- RegionName - This field contains the zip code of the region.
- RegionType- Type of region is Zip.
- StateName - State
- City - This column provide the specific City Name of Housing Data
- Metro - This provide the name of the metro city around that region
- County Name - This is the county name for that region
- Months Column - These columns contains the prices of region for every month

VERIFYING DATA QUALITY

The data did not meet our required quality because after checking through we found that the dataset had missing values but we performed linear interpolation to handle the missing values. Moreover, we dropped some columns that were not relevant to the case study.

DATA PREPARATION

In this phase, we conducted data munging which involved preparing the final dataset for modeling.

Data Selection - The dataset was provided on the Zillow Research Data website.

Data Cleaning - The dataset was corrected by performing linear interpolation for the missing values. There was trend and seasonality observed on the data and hence detrending and deseasonalizing was performed.

DATA MODELLING

The tasks involved in this phase were as follows:

- Selecting modeling techniques
- Performing Train Test splits
- Building the model
- Model Assessment
- Forecasting

We were to use two techniques namely ARIMA, SARIMA and Prophet Modelling, and choose the model with the best metrics of performance.

We performed ACF and PACF on the top 10 detrended and deseasonalized region names. On performing the ARIMA and SARIMA modelling we obtained the following results;

```
SARIMAX Results
====================================================================
========== Dep. Variable: seasonal No. Observations: 36 Model:
ARIMA(1, 1, 1) Log Likelihood -111.620 Date: Tue, 20 Jun 2023 AIC
229.239 Time: 17:11:03 BIC 233.905 Sample: 05-01-2015 HQIC 230.850 -
04-01-2018 Covariance Type: opg
====================================================================
========== coef std err z P>|z| [0.025 0.975]
--------------------------------------------------------------------
---------- ar.L1 -0.1313 0.297 -0.441 0.659 -0.714 0.452 ma.L1
-0.6924 0.213 -3.245 0.001 -1.111 -0.274 sigma2 33.6579 8.259 4.075
0.000 17.471 49.845
====================================================================
=============== Ljung-Box (L1) (Q): 0.05 Jarque-Bera (JB): 0.41
Prob(Q): 0.82 Prob(JB): 0.81 Heteroskedasticity (H): 1.58 Skew:
-0.23 Prob(H) (two-sided): 0.44 Kurtosis: 3.27
====================================================================
===============
```

We modeled using SARIMAX module and obtained our results as follows:

```
SARIMAX Results
Dep. Variable:        seasonal     No. Observations:   36
Model:          SARIMAX      Log Likelihood        -116.825
Date:   Tue, 20 Jun 2023    AIC     235.651
Time:   17:19:53        BIC     237.234
Sample:         05-01-2015    HQIC   236.203
- 04-01-2018
Covariance Type:    opg
coef    std err z         P>|z|   [0.025 0.975]
sigma2          38.5677         6.964  5.538  0.000  24.919         52.216
Ljung-Box (L1) (Q):  1.40    Jarque-Bera (JB):    3.58
Prob(Q):        0.24    Prob(JB):       0.17
Heteroskedasticity (H):      3.23    Skew: -0.07
Prob(H) (two-sided):         0.05    Kurtosis:       4.54
```

The heteroskedasticity statistic (H) in this case, has a statistic of 3.23, and the associated p-value (0.05) suggests some evidence of heteroskedasticity in the residuals.

We proceeded to calculate the RMSE of the train data and obtained a value of 5.4439.

The train data was fit to the model and iterated over the test data for each of the region names and the RMSE values obtained. The model was then used for forecasting using the Prophet Model and a period of three years was selected for forecasting.

This aided in visually determining the Region Names which show an upward future trend in the housing market.

DATA EVALUATION

We conclusively identified the top 5 best Zipcodes to invest in as:

- 34102
- 81611
- 81615
- 89413
- 96141

CONCLUSION

After performing time series analysis on the 10 zip codes and forecasting total returns for up to three years, we recommend the company to invest in the following 3 zipcodes:

- 81611 - Location: Aspen, CO (R.O.I - 132.378817)
- 10021 - Location: New York, NY (R.O.I - 111.795552)
- 34102 - Location: Naples, FL (R.O.I - 7.605307)

As for the other 6 zip codes, they are not fit for investment given the negative returns.

REFERENCES:

1. https://www.zillow.com/research/data/ accessed on 13th June 2023 at 1700hrs
2. https://www.dcc.fc.up.pt/~ltorgo/Papers/DFRBR/DFRBR-4.html accessed on 17th June 2023at 1000 hrs.
3. Canvas Content.