Merve Karacaoğlu 32282

This pdf explains all the steps of the project. I made changes to the STEP 2 of the project.

CS210 Project

First Data set:

The first data set is “2019-2024 US Stock Market Data” from Kaggle. This data set includes dates, stock prices, and trading volumes of various companies and goods. I have focused specifically on Amazon. To prepare the data for analysis, I performed several preprocessing steps. I found the mean, variance, etc., converted dates to DateTime format, and set the “Date” column as the index. Additionally, I made 2 plots to show the change in Amazon Stock Prices and Amazon Trading Volume over the years by day.

Second Dataset:

I have utilized a second dataset from Our World in Data, which provides information on stay-at-home requirement levels across various dates and locations between the years 2020-2023. Similar to the first data set, I performed several preprocessing steps. I found the mean, variance, etc., converted dates to DateTime format, and set the “Day” column as the index. I made a bar plot to show from which countries the data set have the most data.

NaN-Values:

There aren’t any missing values in both of my data sets.

Monthly Data:

I resampled both data sets to take the monthly averages and made 2 plots to show the change in Amazon Stock Prices and Amazon Trading Volume over the years by month. Additionally, I made another plot to show the change in stay-at-home requirement levels over the years by month.

Scatter Plots:

I scaled both data sets before plotting the scatter plots, to be able to visually assess the distribution and identify any potential outliers.

I generated 2 scatter plots to show the relations between Amazon Stock Prices and Amazon Trading Volume compared to stay-at-home requirement levels.

Hypothesis Testing

Hypothesis 1: Impact of Stay-Home-Requirement Levels on Amazon Stock Prices

Null Hypothesis (H0): There is a significant correlation between monthly Amazon Stock Prices and the stay-home-requirement levels.

Alternative Hypothesis (H1): There is no correlation between monthly Amazon Stock Prices and the stay-home-requirement levels.

Hypothesis 2: Impact of Stay-Home-Requirement Levels on Amazon Trading Volume.

Null Hypothesis (H0): There is a significant correlation between monthly Amazon stock traded and the stay-home-requirement levels.

Alternative Hypothesis (H1): There is no correlation between monthly Amazon stock traded and the stay-home-requirement levels.

I merged the two data sets to find the correlation coefficients, and p-values and perform a t-test.

Amazon Stock Price vs. Stay-At-Home Requirement Levels

The correlation coefficient is 0.578. There is a moderate positive linear relationship between Amazon Stock Prices and Stay-At-Home Requirement Levels. As stay-home requirements increase, Amazon prices tend to increase as well.

The p-value is 0.0002. The p-value is very low (much less than 0.05), indicating that the correlation is statistically significant. The likelihood that this correlation is due to random chance is very small.

Amazon Trading Volume vs. Stay-At-Home Requirements Levels

The correlation coefficient is 0.085. There is a very weak positive linear relationship between Amazon Trading Volume and Stay-At-Home Requirement Levels. The correlation is so weak that it suggests almost no linear relationship.

The p-value is 0.623. The p-value is high (greater than 0.05), indicating that the correlation is not statistically significant. This means that any observed correlation is likely due to random chance.

Defining Hypothesis:

Null Hypothesis (H0): There is no difference in the mean quantity of stay-at-home requirement levels between months with high Amazon Stock prices and months with low Amazon Stock Prices.

Alternative Hypothesis (H1): There is a significant difference in the mean quantity of stay-at-home requirement levels between months with high Amazon Stock Prices and months with low Amazon Stock Prices.

The T-Statistic of 3.389 indicates that the estimated coefficient is 3.39 standard errors away from zero. This is quite far, indicating that there is a large difference in the means of the two groups(high and low prices).

The P-value of 0.0025 is much lower than the common significance level of 0.05. This suggests that the observed difference in mean quantities of stay-at-home requirements between high and low-price months is statistically significant at the %5 level.

Based on the two-sample t-test, we have sufficient evidence to reject the null hypothesis. Therefore, based on this test, we conclude that there is a significant effect of stay-home-requirement levels on the Amazon Stock Prices

Linear Regression

I plotted a single linear regression model for the prediction of stay-at-home requirement levels.

Machine Learning

I applied 3 different machine learning algorithms: Random Forest, kNN, and Decision Tree. I tuned and tested their parameters for the best values and fewer errors. Since all the data are normalized, and all the values are between 0-1, I calculated the RMSE values. kNN mean squared value is the lowest value compared to others. So kNN is the best ML model for this data set.

kNN MAE: 0.122

An MAE of 0.122 indicates that the model's predictions are quite accurate on average. The error magnitude is relatively low, which is a positive indication of the model's performance.

kNN R^2: 0.664

An R^2 of 0.664 suggests that the model explains a substantial portion of the variance in the target variable. This indicates a reasonably strong relationship between Stay-at-Home Requirements and Amazon Prices.