# **Project Report**

## GitHub URL

[PranavManiyar/UCDPA\_PranavManiyar (github.com)](https://github.com/PranavManiyar/UCDPA_PranavManiyar)

## Abstract

This project is to showcase the knowledge and skills that are gained throughout the last few months. This project aims to showcase different skills like data importing, cleaning and manipulating, analysing and providing various insights on chosen real-world scenarios. This project is broken down into two parts. The first part of the project is analysing UK Housing Market. Further down at a later stage it will focus on the "Slough" area as this is my residential area

For Machine Learning, this project is looking at the set of data for Students Grading and trying to predict one of the 3rd grading using linear regression.

## Introduction

The UK housing market has always been an interest of mine. I am interested in investigating how the Sales Volume and Average Sale Price have changed over the years in this area for different types of properties. I was also interested in understanding the difference in the sale price of Cash Vs Mortgage. For this part of the exercise, various data sources have been used to source different sets of information by reading CSV and Web scrapping methods. This little analysis work has given me some very interesting insights about this market. For the machine learning part, I have chosen to predict student grades which can be quite useful if worked well.

## Dataset

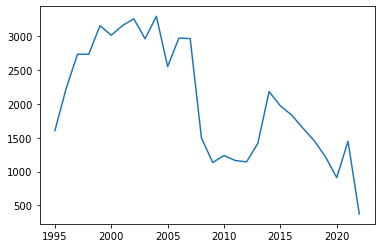
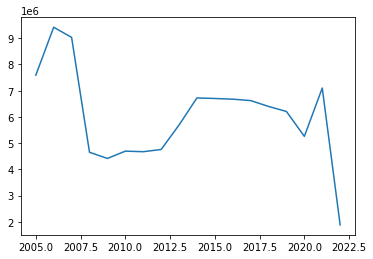
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **Description** | **File Name** | **Source Link** | |
| **1** | UK Housing Price History from 1968 by Regions and by house type | UK HPI Full File | | [UK House Price Index: data downloads July 2022 - GOV.UK (www.gov.uk)](https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-july-2022) |
| **2** | MonthFile to create mmmyyy field within the dataframe | MonthFile | | Self Created |
| **3** | Slough Crime Rate | EPL\_standings\_2000-2021.csv | | [Slough Westminster Constituency (ukcrimestats.com)](https://www.ukcrimestats.com/Constituency/65680) |
| **4** | UK Average Salary | UK Average Salary | | [EARN01: Average weekly earnings - Office for National Statistics (ons.gov.uk)](https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/averageweeklyearningsearn01) |
| **5** | **UK Inflation Rate** | CPI Index | | [CPI-consistent inflation rate estimates for UK household groups - Office for National Statistics](https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/cpiconsistentinflationrateestimatesforukhouseholdgroups/2005to2021) |
| **6** | **Student Grades Files** | Student-mat | | [UCI Machine Learning Repository: Student Performance Data Set](https://archive.ics.uci.edu/ml/datasets/Student+Performance) |

## Implementation Process

**UK Housing Market**

* Imported the required modules for this particular exercise.
* Imported UK HPI Full File.csv and read that file in and created a Pandas DataFrame called df\_house.
* Investigated the dataframe by using different functions like info, describe, head and tail
* Removed some unnecessary variables from the dataframe by choosing only selected columns by their names
* Change the type for the Date field from object to datetime64ns by using the to\_datetime function.
* Converted date to mmmyy format to merge later with other data sources. To this, I have created 2 new variables "Year" and "Month" by extracting relevant information from the date
* Did further investigations of the variables
* Read in MonthFile.csv file and created another Pandas Dateframe called Monthfile. This is then merged with the house dataframe on the month variable. The new dataframe is called df\_house\_clean. After creating this dataframe further manipulation was done to create the date in mmmyy format.
* UK Sales Volume and Average Sale Prices by the different property types have been investigated to understand the historic trends. Sales volume by year was plotted on the chart using matplotlib to understand how the market has been performing over the years. Another multiline chart was created to plot AveragePrice, DetachedPrice, SemiDetachedPrice, TerracedPrice and FlatPrice by year. This has given some valuable insight in terms of how the market reacted to the 2008-2009 financial crisis, then how it revived and performed and how it is tracking now.
* After looking at the overall UK Market, the focus was turned specific to the “Slough” area.
* A new dataframe called df\_house\_clean was created by using Regular Expression to extract the “Slough” area from the Region variable within the dataframe. I then cross-checked this by creating a similar dataframe using a different method to make sure the Regex worked as intended. The new dataframe had 331 observations. I then further investigated the data and then removed some unnecessary variables by using the Pandas drop function. Checked the shape of the new dataframe and did the further investigation of the dataframe. I then tried to understand the null values within the datafarme and then filled them using different methods like with integer 0, median value, bfill
* I then created a NumPy array using AveragePrice, CashPrice, and MortgagePrice to investigate the difference in prices between the CashPrice and MortgatePrice. Different NumPy functions were used to answer this. This has also given me some valuable insights.
* I then performed a similar exercise for the “Slough” market which I did for the “UK” market to understand if “Slough” has followed the same trend as the UK or different. Looking at the trend it shows that the “Slough” market has performed very similar to UK Market by and large however it has seen some steep fall from around 2013 to 2019 compared to the overall UK Market.

Figure 1UK Market Figure 2Slough Market

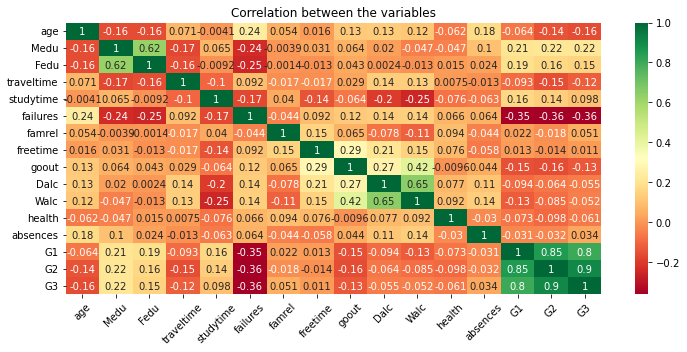
* 
* I then ploted the New Sales and Old Sales on the bar graph to show the variance in the different property types sales which confirms that majority of UK property sales is old property.
* I then investiaged the Prices for the Slough by different types of property but this time plotting these on separate graphs and that has also given some interesting facts about how the different types of properties have performed over the years in terms of their price.
* I then tried to fill the NA values by using the different methods as before.
* I then created Crime data by web scrapping UKCrimestats.com website and storing that in pandas DataFrame called Crime. Used pd.read\_html function and used index 0 to get the correct table from the webpage.
* I then renamed the the variables using .rename function to get the desired names of the column. Then dropped unnecessary columns.
* I then read in Imported UK Average Salary.csv and CPI Index.csv files,read them in and created a Pandas DataFrames called Salary and CPI respectively
* I then merged those DataFrames with the House dataframe by using the .merge function.

**Machine Learning on Students' Grades**

* Imported the extra modules required for this exercise.
* Imported student-mat.csv' and read that file in and created a Pandas DataFrame called df\_student. This file was separated by “;” so specified that while reading to get the desired format.
* The file has 33 different variables below is the more details on different variables and their attributes. The data is sourced from [UCI Machine Learning Repository: Student Performance Data Set](https://archive.ics.uci.edu/ml/datasets/Student+Performance)

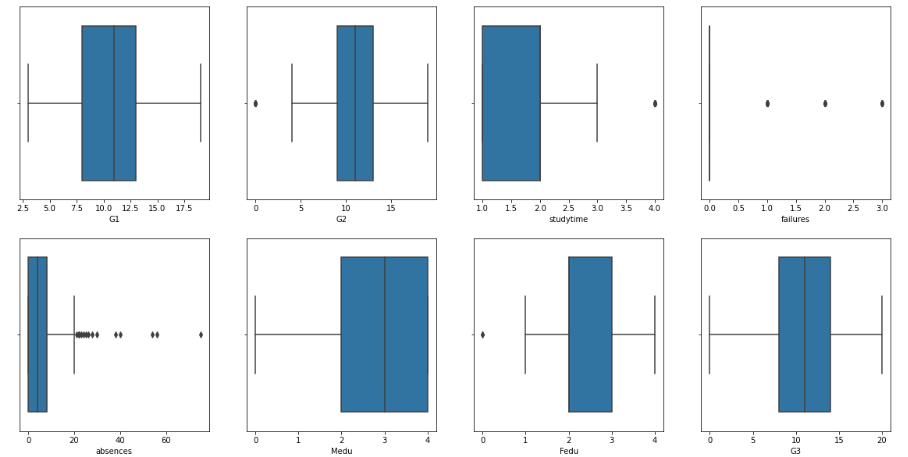


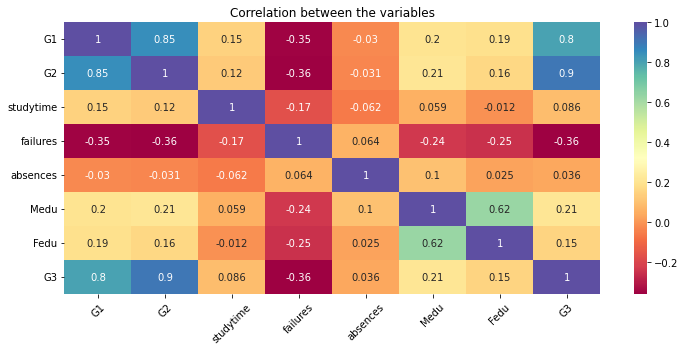
* Investigated the dataframe by using different functions like info, describe, head and tail
* Plot these variables on Seaborn heatmap to see the correlation between the different variables.



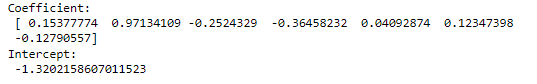
* Removed some unnecessary variables from the DataFrame by choosing only selected columns by their names and created a new DataFrame called Student.
* I then did a Univariate Analysis and then removed any outliers
* I then did the investigative analysis to confirm the quality of the data columns that were fed into the model and make sure there were no null values.
* I then plotted the newly selected columns on a heatmap to understand the correlation between them.
* Two NumPy arrays were created to create the features and labels. These were then split into 90% – 10% ratios into training and test sets.
* linear regression algorithm was chosen to model this data by using linear\_model.LinearRegression()
* I then ran the model which had 86% accuracy with a very strong 98% coefficient with the G2 variable.
* I then checked my predictions against the actual results to see how the model performed and it almost predicted the correct score for the chosen sample.

## Results

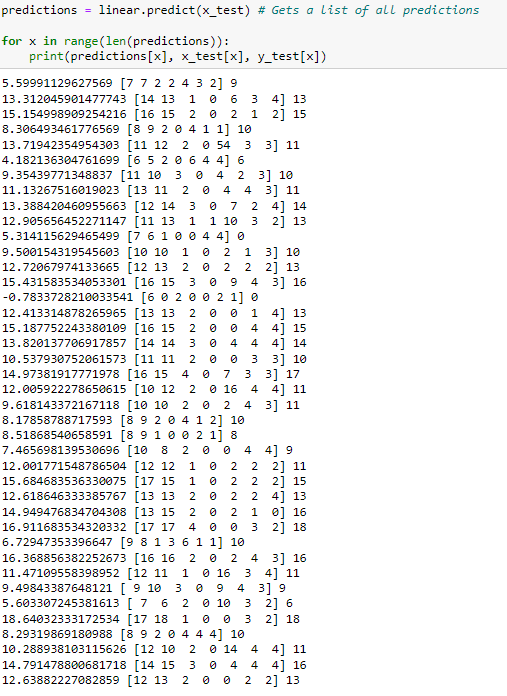
* Univariate Analysis Results
* Correlation between the key variables shows some strong relationships between first period grade and second period grade with the Final grade



* Coefficients are confirming that G3 has very stong relationship with G2 with 97% Coefficient



* The mode has 85% accuracy and Final result is showing that the model has performed quite well in predicting the scores.



## Insights

* The UK housing market was booming around mid-2005 to 2006 in terms of sales volumes and increase in average sales prices. This clearly took a massive hit at the time of the 2008 financial crisis where you can see a deep dive in the graph. Since 2010 both Sales volume and prices have started going up. The Sales volume is still to hit before the 2008 financial crisis mark but the average sale price has gone up significantly. Interestingly all different types of houses have followed a similar pattern in terms of increase in their prices but Flats have seen some massive growth in price around 2016 to 2018 which has then slowed down and is now in line with other types of properties.
* Looking at the trend it shows that the “Slough” market has performed very similar to UK Market by and large however it has seen some steep fall from around 2013 to 2019 compared to the overall UK Market.
* Within the Slough property market, Cash Price for the property was down by an average of 6.75% compared to the average sale price. Similarly, Mortgage Prices were 1.35% higher than the average sale price
* Within Slough area Detached houses, Semi Detached Houses and Terraced houses all have performed similarly but Flats have had a slightly different trajectory in terms of price over the last few years.
* Average Sale Price has consistently increased over the last years with average UK salary.
* From the machine learning project, the linear regression algorithm has provided with 85% accuracy and has performed very well in predicting the final grades when cross-checked with the data. It is very evident that the second-period grade has the most influence on the final grades.

## Reference

* [UK House Price Index: data downloads July 2022 - GOV.UK (www.gov.uk)](https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-july-2022)
* [Slough Westminster Constituency (ukcrimestats.com)](https://www.ukcrimestats.com/Constituency/65680)
* [EARN01: Average weekly earnings - Office for National Statistics (ons.gov.uk)](https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/averageweeklyearningsearn01)
* [CPI-consistent inflation rate estimates for UK household groups - Office for National Statistics](https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/cpiconsistentinflationrateestimatesforukhouseholdgroups/2005to2021)
* [UCI Machine Learning Repository: Student Performance Data Set](https://archive.ics.uci.edu/ml/datasets/Student+Performance)