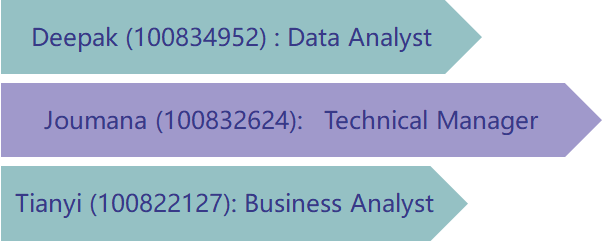
## 





03.04.2022



### Executive Summary

Real-estate developers may not have key insights into future housing prices. The objective of our capstone is to help facilitate those insights, and example is predicting which methods could better predict housing prices (based on key variables) and the variables associated with price, which will benefit not only the real-estate developers but also the community. Using these methods, we can tell if a potential spot is closer to the main road or if it is a preferred area.

The data is from the Kaggle Platform, which provides information about housing prices based on various factors. 13 columns (12 variables contributing to cost) and 546 rows of data (different housing prices entries).

We have used Excel and Python for Data Cleaning and manipulation. We used PowerBI to analyze the current state while using python for predictions. We aim to leverage Python Code for prediction in a mobile app to benefit from widespread ease of use on mobile phones.

By the end of the project, we will be recommending the best algorithms used to predict the future housing prices, whether it is a prefered area or not and if it is the main road. Some of the algorithms have higher accuracy than others. We used two approaches, one with selecting all futures and choosing essential features. The main four important features as we are going to see later on are: “area”,” bedroom”,” bathroom”, and” main road”. For pricing, if you have a limited budget, you should buy a smaller house that is not close to the main road, has more minor bedrooms, bathrooms, stories, parking, is semi-furnished or not furnished, etc.

Graphical user interface, application, table, Excel

Description automatically generated

### Problem Description

Our business goals are:

1.    Ability to predict future housing prices & best type of units to construct.

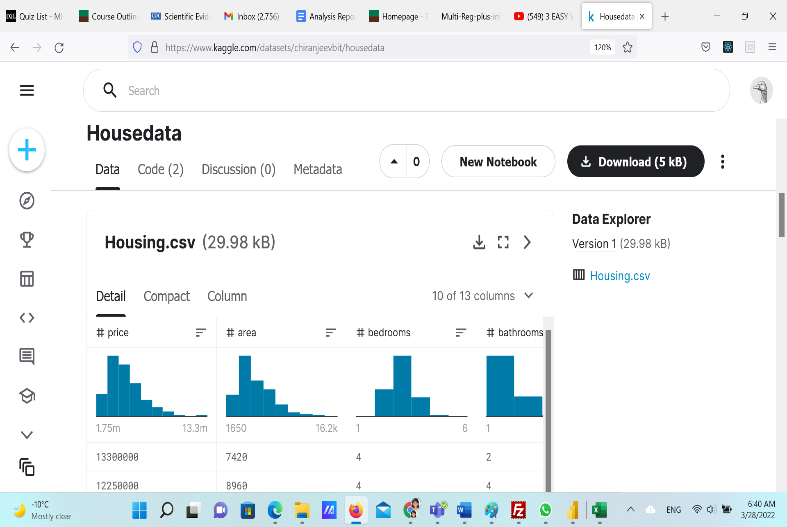
2.    Utilizing the use of space & locations, especially in crowded areas.

3.    No waste of time and money in zero or low demand.

### Our data analysis goal is to find a better method to analyze the housing price influence variables and predict those factors that have more profound influences.

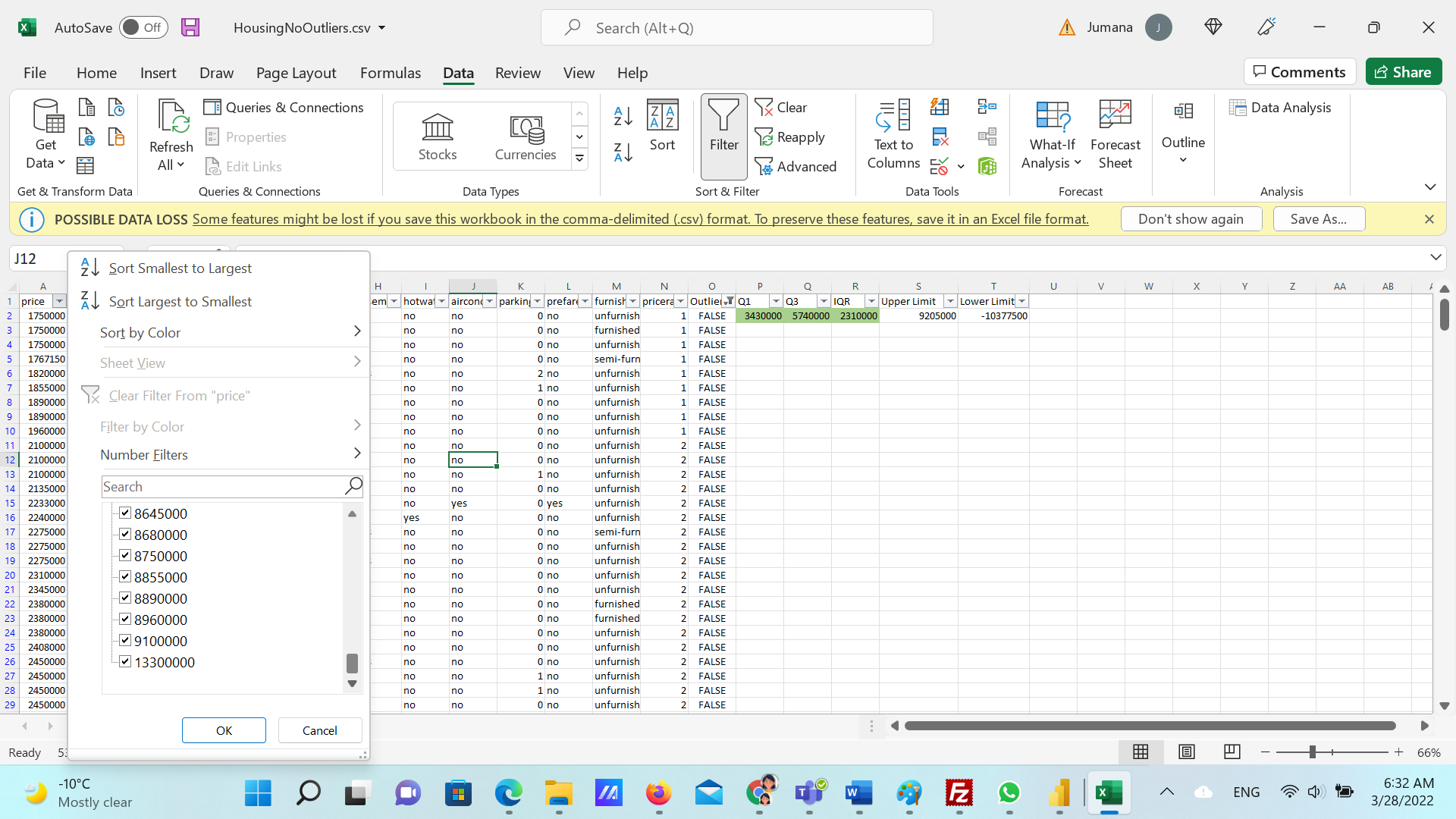
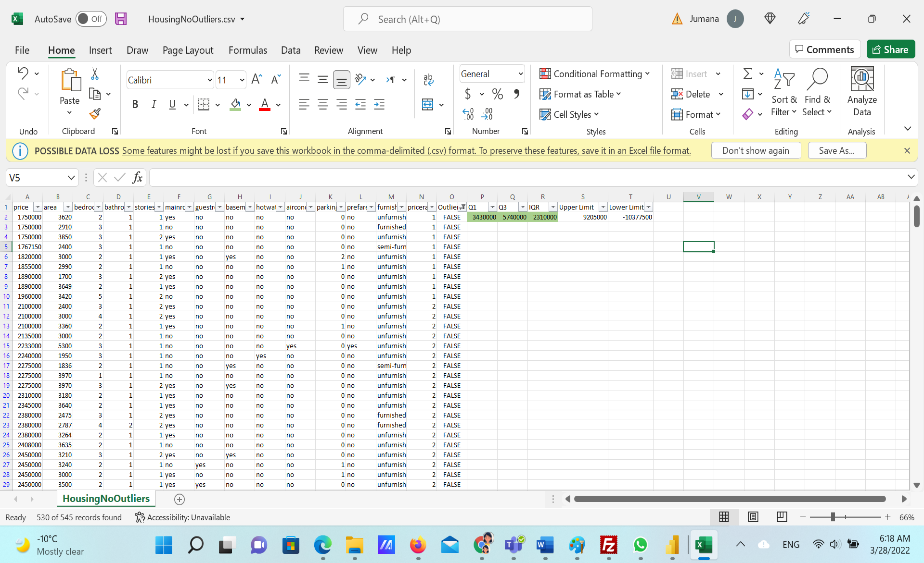
### Data Description

Here are the following variables to determine the **price**:

1. **area**: Unit Area.
2. **bedrooms**: Number of bedrooms.
3. **bathrooms**: Number of Bathrooms.
4. **stories**: Number of Stories.
5. **mainroad**: Does it lie on the main road or not?
6. **guestroom**: Does it have a guest room or not.
7. **basement**: Does it have a basement or not.
8. **hot water heating**: Does it include hot water heating or not?
9. **air conditioning**: Does it have air conditioning or not?
10. **parking**: Does it have parking or not?
11. **prefarea**: Does it lie in a preferred area or not?
12. **furnishing status**: Is it Furnished/Semi-furnished/Unfurnished.
13. **pricerange**: Calculated field for predicting price ranges.

### Data preparation details

* Remove Duplicates.
* Remove Nulls.
* Remove Outliers.



### Data analysis solution

### POWERBI to analyze current state

1. By analyzing the **current state**, we conclude that the maximum price ranges from 10 M to 13M, while the minimum price range is 2M.
2. The **highest** demand is for three bedrooms for all furnished, semi-furnished and unfurnished.
3. There is an **average price** **increase** from unfurnished to semi-furnished by 0.9 M unfurnished to furnished by 1.49 M.

Graphical user interface, application, table, Excel

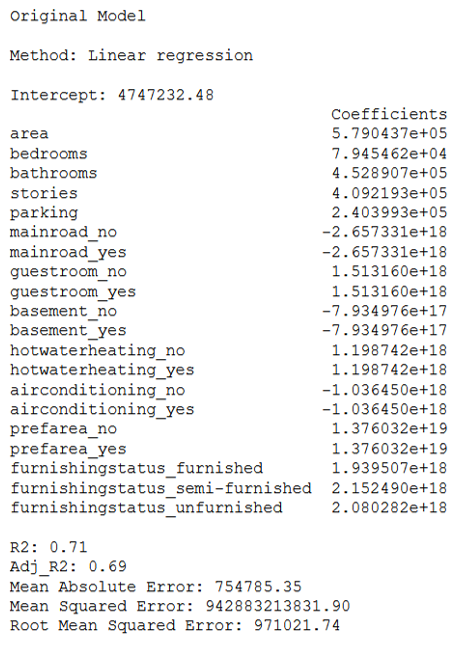
Description automatically generatedGraphical user interface, application, table, Excel

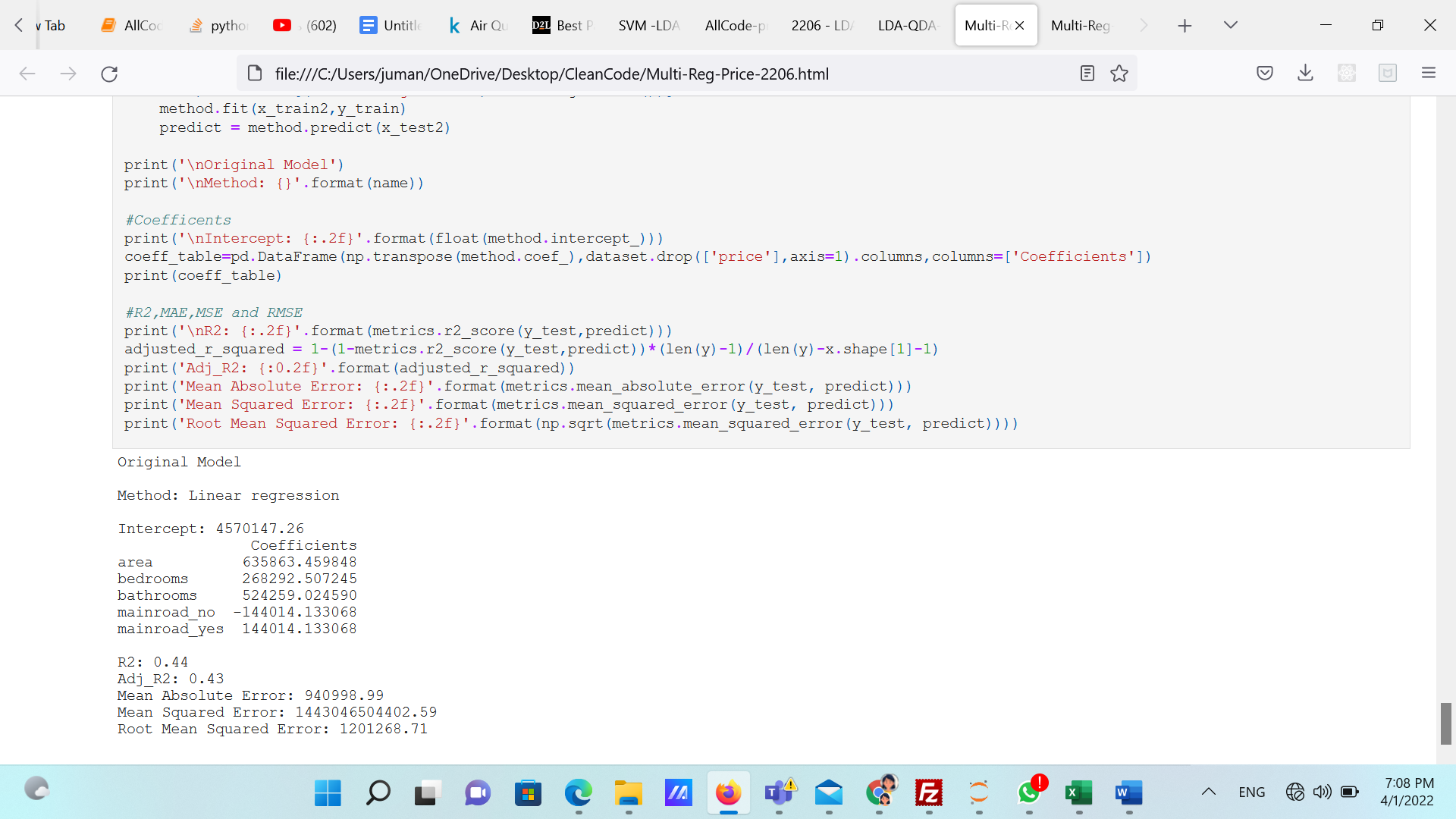
Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

### Linear Regression to predict housing prices

* Multi-linear Regression to predict the relationship between price and independent variables.
* We have attempted Linear Regression first to help create a prediction model for Housing Prices.
* WE CONCLUDE from R2 and adjusted R2 that the regression model explains 70% of the data.
* Root Mean Squared Error. as we previously mentioned: is **the standard deviation of the residuals** (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how to spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. And it is too high.



* We went one step further to see how the model would perform if we selected essential features.
* By applying Backward Selection, we found that the important features are area”, “bedroom”,” bathroom”,” mainroad”. However, the algorithm with all features compound performed better.
* The main for features scored less R2, adjusted R2 and a higher RMSE.

### LDA, QDA, Logistic Regression, SVM and Naïve Bayse to predict if the house is close to Main Road or not.

* Limitations of Logistic Regression
  + 1. Two-Class Problems. Logistic regression is intended for two-class or binary classification problems. It can be extended for multi-class classification but is rarely used for this purpose.
    2. Unstable with Well Separated Classes. Logistic regression can become unstable when the classes are well separated.
    3. Unstable with Few Examples. Logistic regression can become unstable when there are few examples from which to estimate the parameters.
* Best Prediction Method: LDA
* From the scores, we see that LDA did better predict the assumption.

. The performance evaluation suggests that LDA outperformed Logistic Regression in terms of accuracy (F1 Score 83% vs 73%). F1 score is a harmonic mean of both Precision & Recall and measures model performance. And basically, better than any other algorithm.

Graphical user interface, application

Description automatically generated

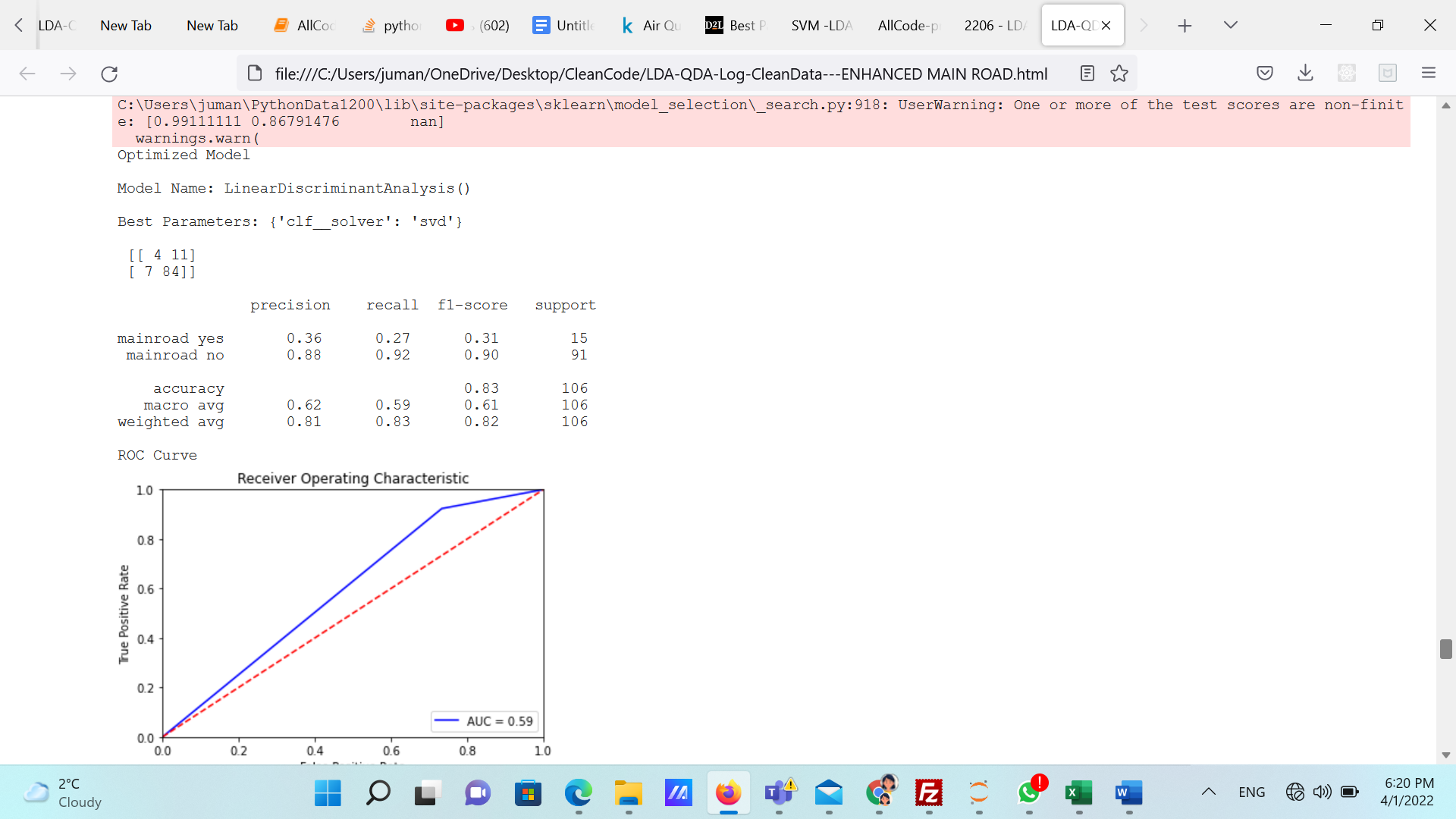
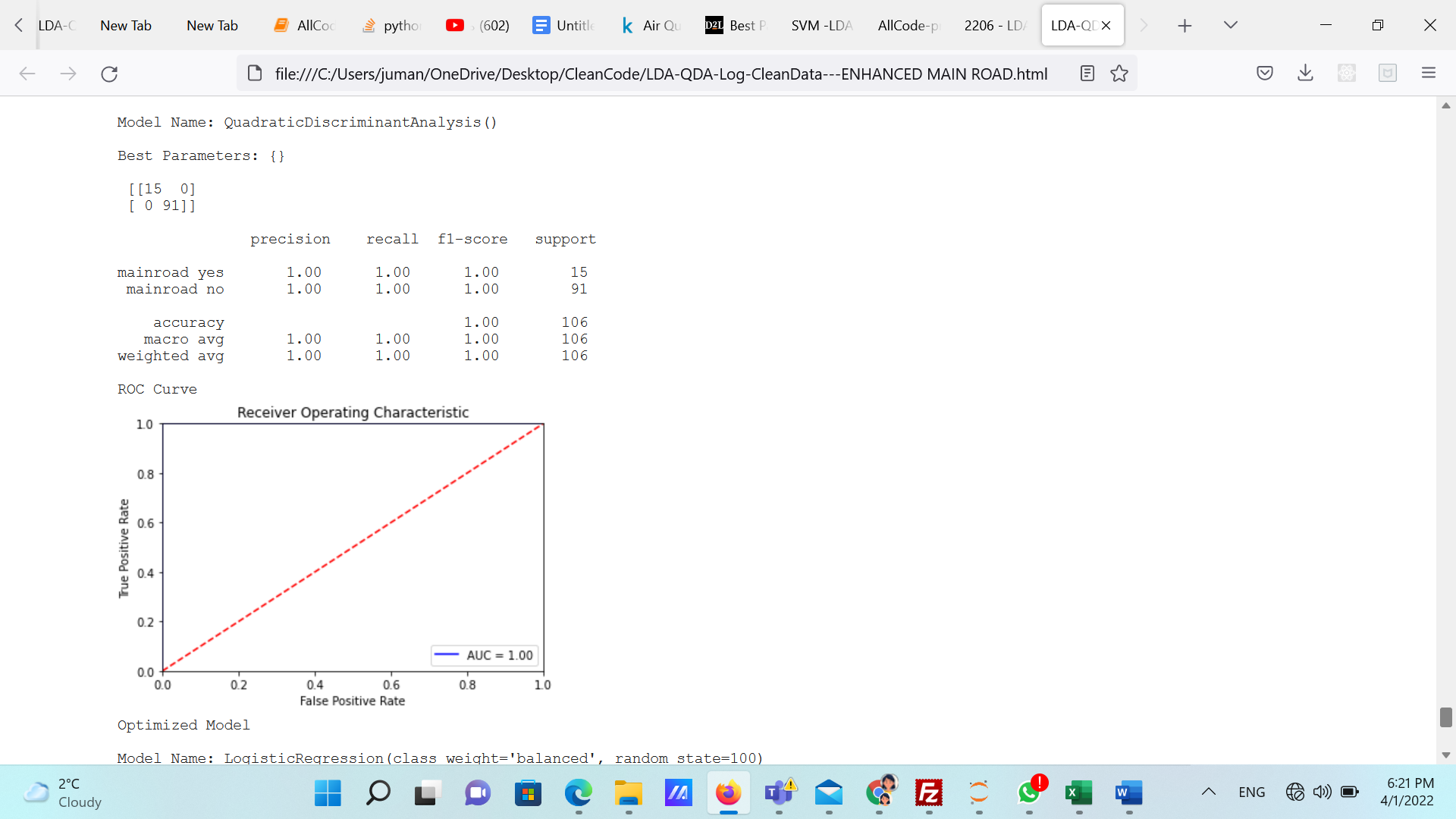
* SVM:

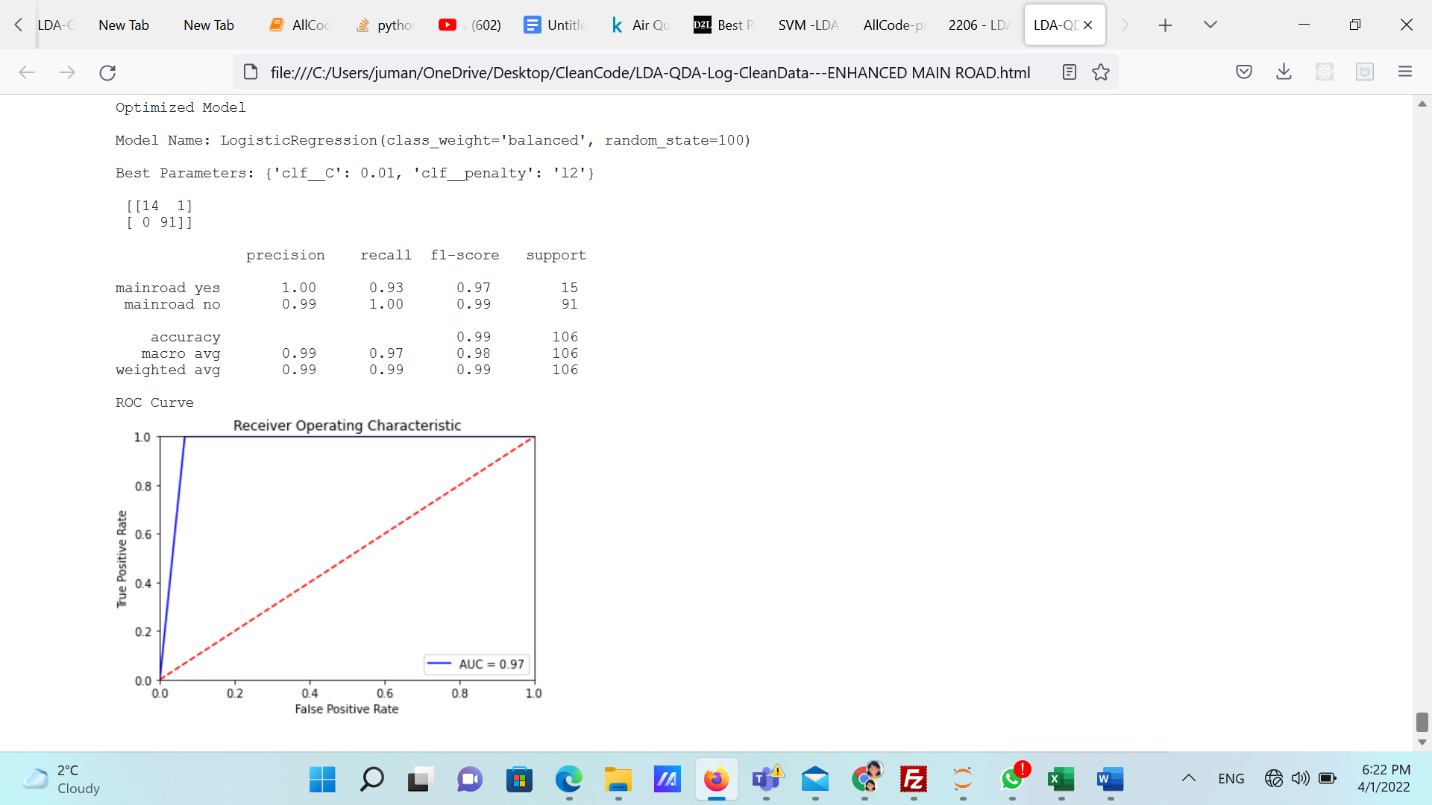
SVM works by mapping data to a high-dimensional feature space to categorize data points, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane.

* Naïve Bayes:

A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points. Popular uses of naive Bayes classifiers include spam filters, text analysis, and medical diagnosis.

* We went one step further and enhanced the algorithm for the main road for LDA, QDA and Logistic regression, and the output was as below:



The methods that we used to enhance our model are the **smote** function that balances the classes. However, this enhancement caused the model to have high variance, a problem that may cause overfitting. We also did **a grid search, which finds the best hyperparameters for the algorithm.**

The f1 scores are very close to one. As we can see, The ROC and AUC are almost perfect.

### LDA, QDA, Logistic Regression, SVM and Naïve Bayse to predict if the house is a Preferred Area or not.

* LDA again performed better than other classifiers with 83% F1 score.

Graphical user interface, application

Description automatically generated

Table

Description automatically generated with low confidence

### Conclusion

This decision is beneficial for Realtor's stakeholders. We decided to use Linear regression with all features possible to find and include in our data set because the model performed way better than the model, which has only some features.

As on our models for predicting the prefered area or not and main road or not, (LDA) performed better than any other classifier.

Unfortunately, the model is such that it's not automated, and updated data has to be plugged in to get the latest trends/potential housing prices. The operational recommendation would be to integrate software (Power Automate) to sync the latest data and update the model accordingly with the latest data on real estate.

We can also create a mobile and web app for recommendations.