Danmarks Tekniske Universitet

Assignment 4

Author

Emil Johannesen Haugstvedt - s211818

December 15, 2021

Contents

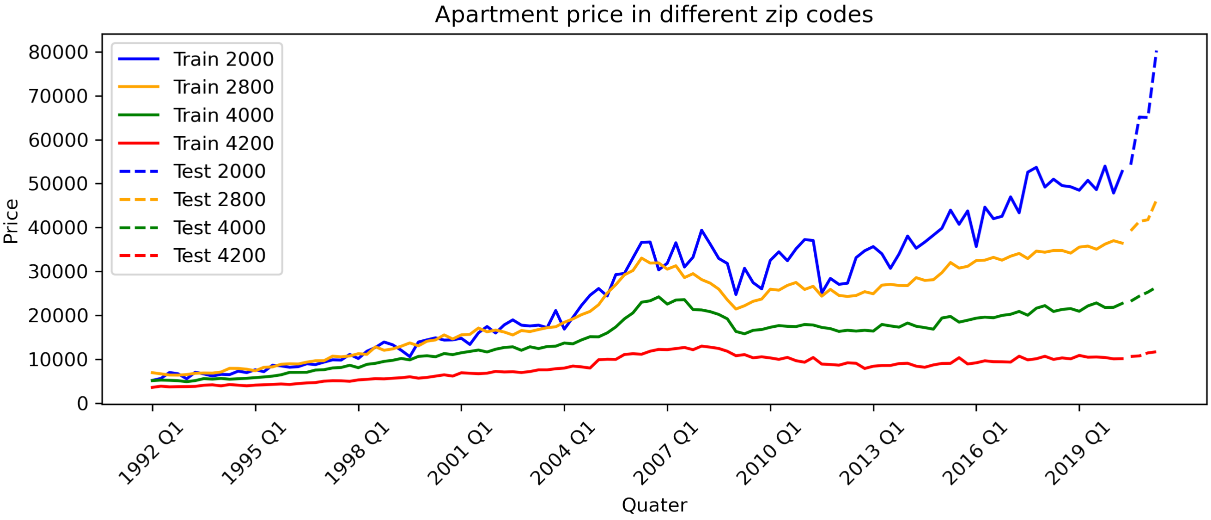

Figure 1: Prices for the different zip codes. Both training data and test data are shown in this plot.

Question 4.1

In Figure 1 the prices for the four different zip codes, 2000, 2800, 4000 and 4200 are plotted, both for the training data and test data. For the rest of the assignment I will only use the training data if not specified.

As you can see from the plot most of the prices seem to follow the same pattern, when one prices goes up, so does the others. There are also no signs of any seasonality in the different prices.

Due to the fact that all prices increases through the time series all of them looks to be non-stationary. For the zip code 2000 the variance also is clearly increasing through time. For the three other zip codes the change in variance is not that clear, but by looking at the price in the very start and very and the variance of all those time series also seem to increase.

It is clear that the mean of all different prices increases through time. And by looking at the different prices in the very start and towards the end it is clear that the variance is also increasing for all prices. Thus the prices are non-stationary.

Since both the mean and variance changes through time both differencing and log transformation are needed to make the time series stationary. As far as I can see the trend in training part of the prices are linear, thus a differencing of one combined with a normal log transformation would make the prices stationary.

The two different interest rates can be seen in Figure 2, both training and test. Here you can see that, as for the prices, the interest rates also seem to follow the same overall pattern, with one laying a bit higher than the other and with less fluctuation. Also for this time series the mean is clearly shifting through time. When it comes to the variance of the rates, it seems to be a bit higher in the start than in the end. This implies that the rates are non-stationary.

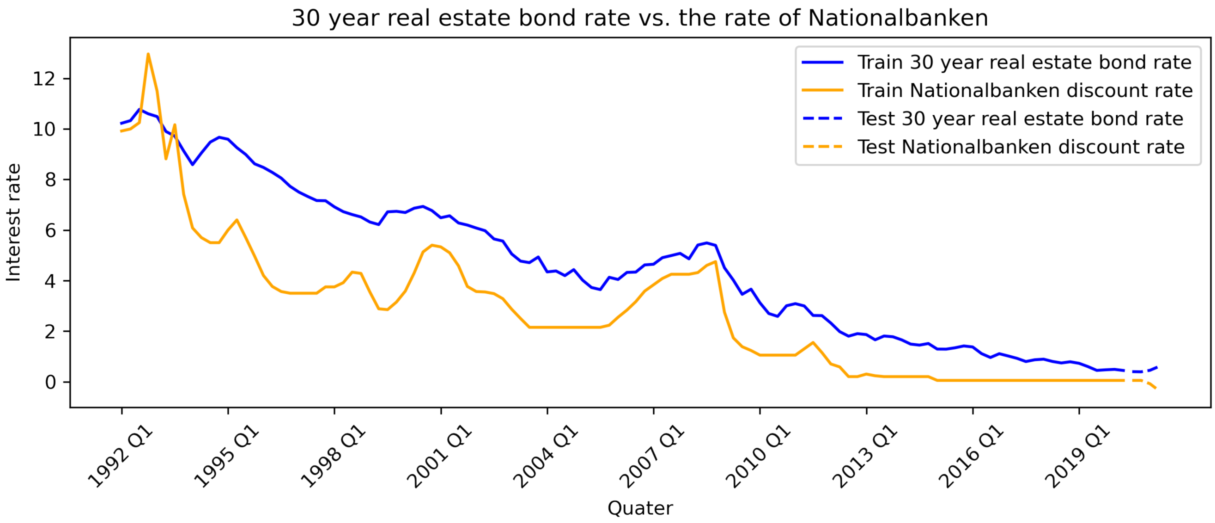

Figure 2: The 30 year real estate bond rate and Nationalbanken discount rate plotted from 1992 to 2020, both test data and training data. This plot is showing how the mean and variance of both interest rates are changing through time.

These time series can be transformed in the same way as the prices to achieve a stationary time series. This is by using differencing to make the mean stationary and use log transformation to make the variance more constant.

Lastly, Figure 3 shows the consumer price index (CPI). Both for training and test. The mean of the CPI changes massively through time making the time series non-stationary. But when it comes to the variance it seem to have been constant for the whole time period. Thus the time series may only need to be differenced in order to be made stationary.

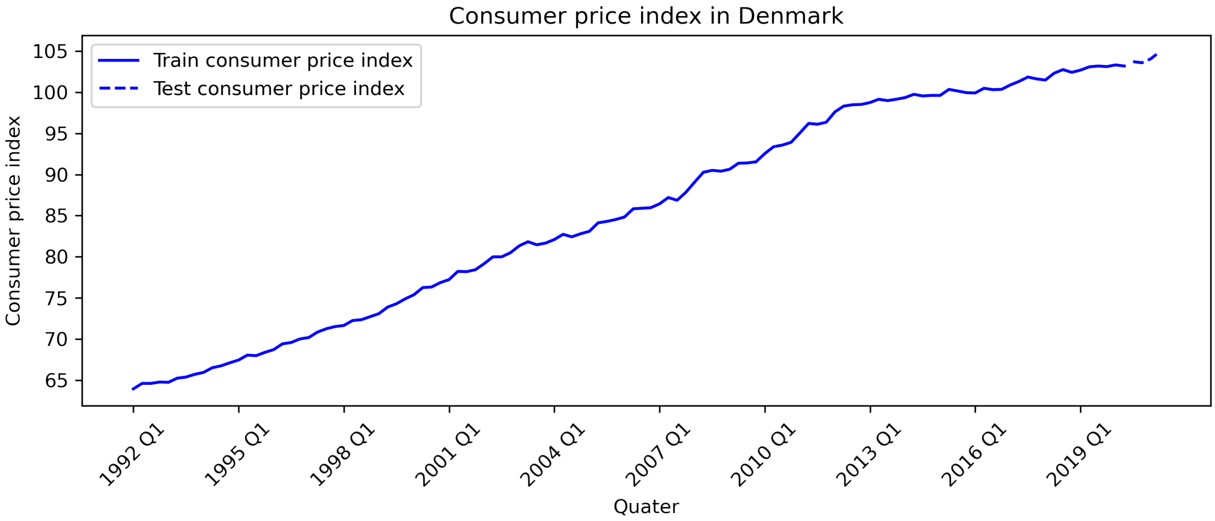

Figure 3: Consumer price index from 1992 to 2020, both test and training. The mean is heavily shifted upwards for the whole period, making the time series non-stationary. But when it comes to the variance it seems to have remained the same for the whole period.

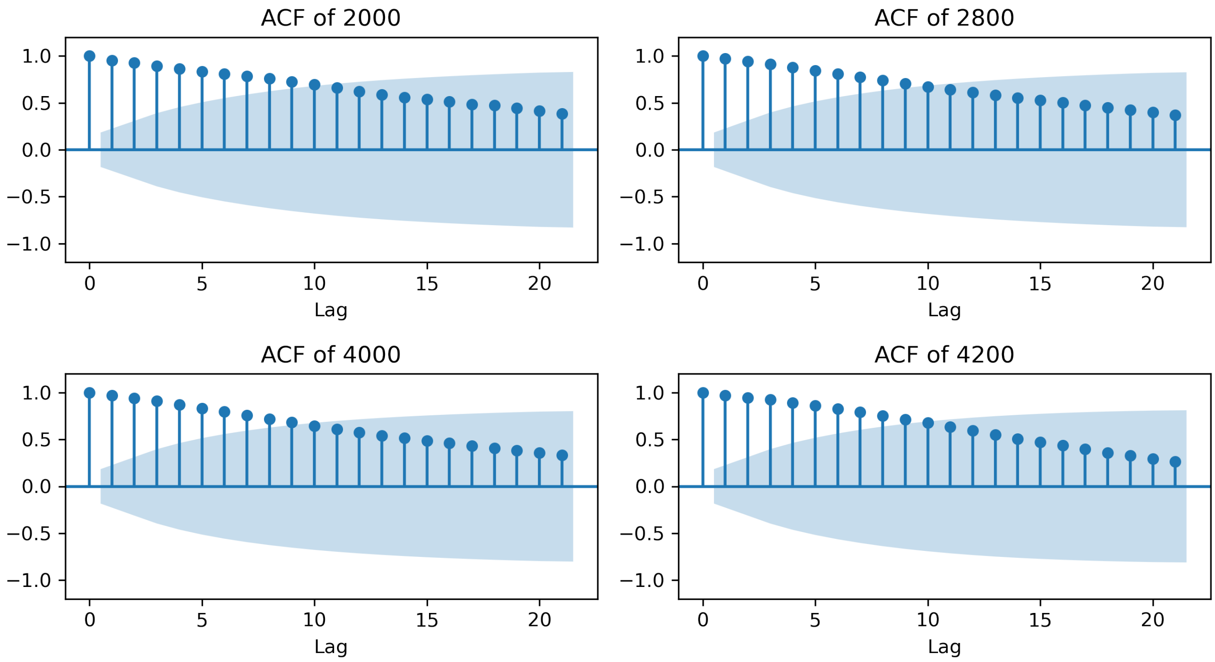

Figure 4: Plot of the autocorrelation function of the four different zip codes. All shows large positive correlations for many lags, indicating non-stationary time series.

Question 4.2

Plots of the autocorrelation functions (ACF) of all the different ship codes are shown in Figure 4. As suggested in the last question the ACF of the zip codes proves the non-stationarity of the time series. The reason for the non-stationarity is the consecutive large positive correlations in all the different ACF plots. The partial autocorrelation function of all the different zip codes are shown in Figure 5.

To cope with the non-stationarity of the time series differencing and log transformations are performed. In Figure 6 the differenced and transformed prices are shown. Now the mean of all the zip code prices are around zero and the variance seem to be the same throughout the whole time series. This results in the time series now being stationary and it is suitable for predictions, shown in Figure 7. Here you can see how the positive correlations decays to zero after just a couple of lags.

The cross-correlation function (CCF) between the prices in the four zip codes are shown in Figure 9. The CCF is a measure of the similarity between two time series. And as you can see there are high correlations between the different prices. This indicates that it is possible to use the other prices in the other zip codes to predict the future price of a zip.

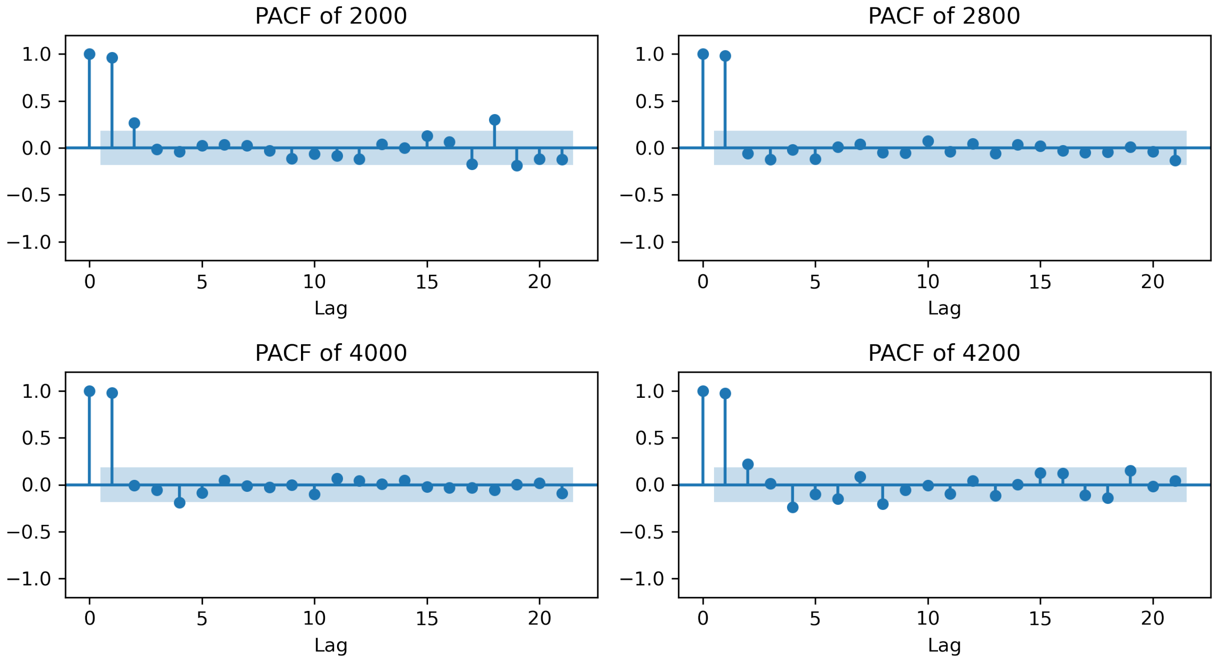

Figure 5: Plot of the partial autocorrelation function of the four different zip codes, 2000, 2800, 4000 and 4200.

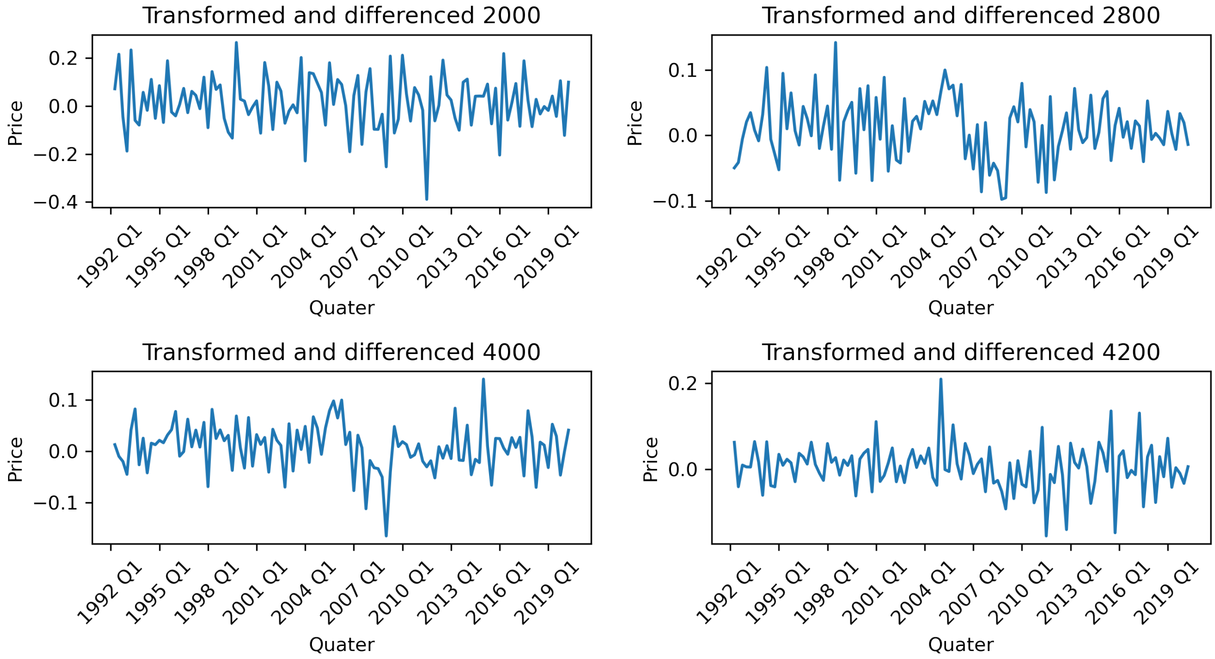

Figure 6: The differenced and transformed prices for the different zip codes. Now the mean of all the zip code prices looks to be 0 and the variance seem to be constant throughout the time series.

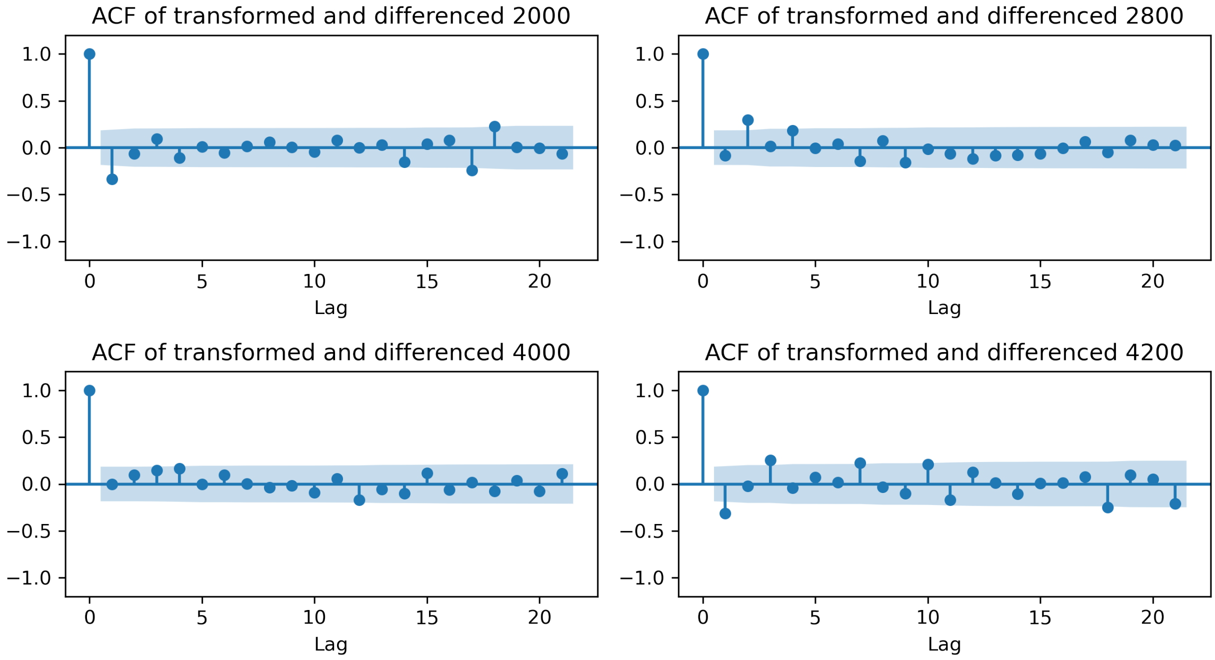

Figure 7: Auto correlation function of the prices in the different zip codes after both differencing and transformation. All ACF plots now decays quickly to zero a sign of stationarity.

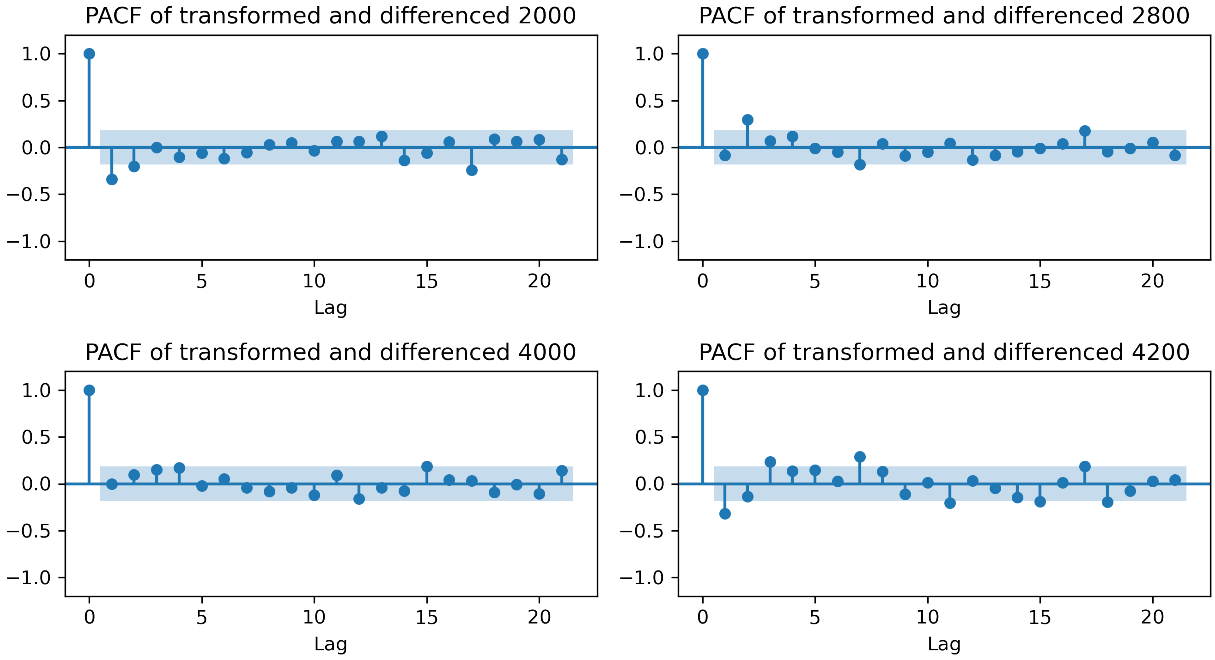

Figure 8: The PACF of the differenced and transformed prices in the differenced zip codes. The also goes more quickly to 0 than the PACH of the original time series.

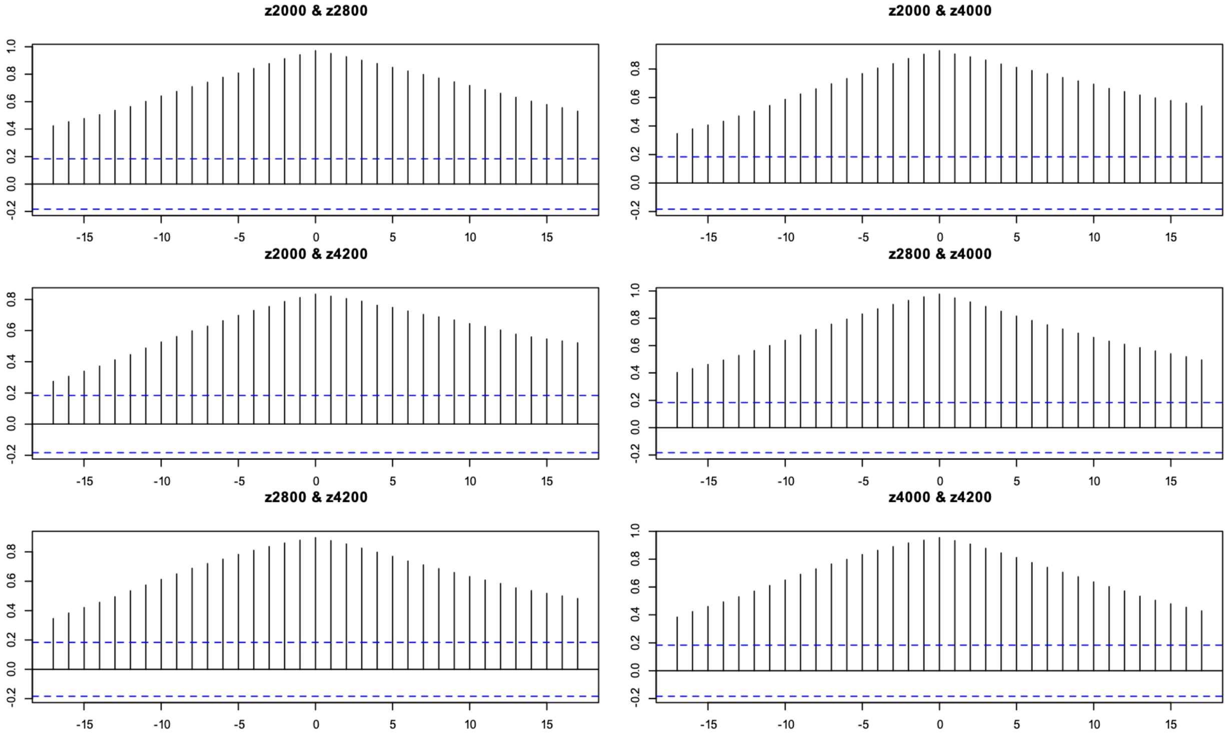

Figure 9: The cross correlation function between the different zip codes.

Question 4.3

Now that the zip code prices time series are made stationary it is possible to start predicting. The predictions are made using ARIMA models. The order of the models for the different zip codes are determined by testing different model orders, using the ACF and PACF for prices as a starting point. First a model is chosen based on the AIC of the different ones tested, then the order is adjusted after residual analysis. All models are differenced once the be made stationary, and the different orders tested for the AR and MA parts are $p, q \in [0, 3]$.

Using the approach stated above the following orders were found for the different ARIMA models for the prices:

- **Zip 2000:** (0, 1, 1)
- **Zip 2800:** (2, 1, 1), changed to (0, 1, 1) after residual analysis.
- **Zip 4000:** (1, 1, 1), changed to (0, 1, 1) after residual analysis.
- **Zip 4200:** (1, 1, 2), changed to (0, 1, 2) after residual analysis.

After analysing the residuals of the models suggested above some adjustments had to be made. For the zip 2800, zip 4000 and zip 4200 models the residuals had high autocorrelations for the first lags. Since the AIC is known for suggesting a too high model order I reduced the AR term by two for zip 2800, one for zip 4000 and one for zip 4200 and ended up with the orders (0, 1, 1), (0, 1, 1) and (0, 1, 2), respectively.

The predictions with the final models are seen in Figure 10. Here you can see how all predictions are way below the actual values. When looking at Figure 1 this is not surprising as all the prices experiences a big growth right after the training data ends.

The residuals on the training data of the four final models are plotted in Figure 11. The residuals for the zip 2000 looks i.i.d, but for zip 2800, zip 4000 and zip 4200 there is a big residual in one of the first lags. It is hard to say why this big residual is there, but it may be some abnormalities in the training data in the first samples resulting in wrong predictions. Since the rest of the residuals looks i.i.d and slightly changing the model order did not help I keep the orders.

To further check for autocorrelations in the residual space for the models I performed the Ljung-Box test on the residuals. The p-values for the first 20 lags of the residuals models are shown in Figure 12. These plots show that the p-value for zip 2000 all lies above 0.05, something that indicates no autocorrelations in the residuals. But for zip 2800, zip 4000 and zip 4200 there are correlations in the first lags of the residuals. This is not surprising, as the residuals has this big dip in Figure 11. As chaining the model order did not improve the residuals I will keep the model and hope that the MARIMA performs better!

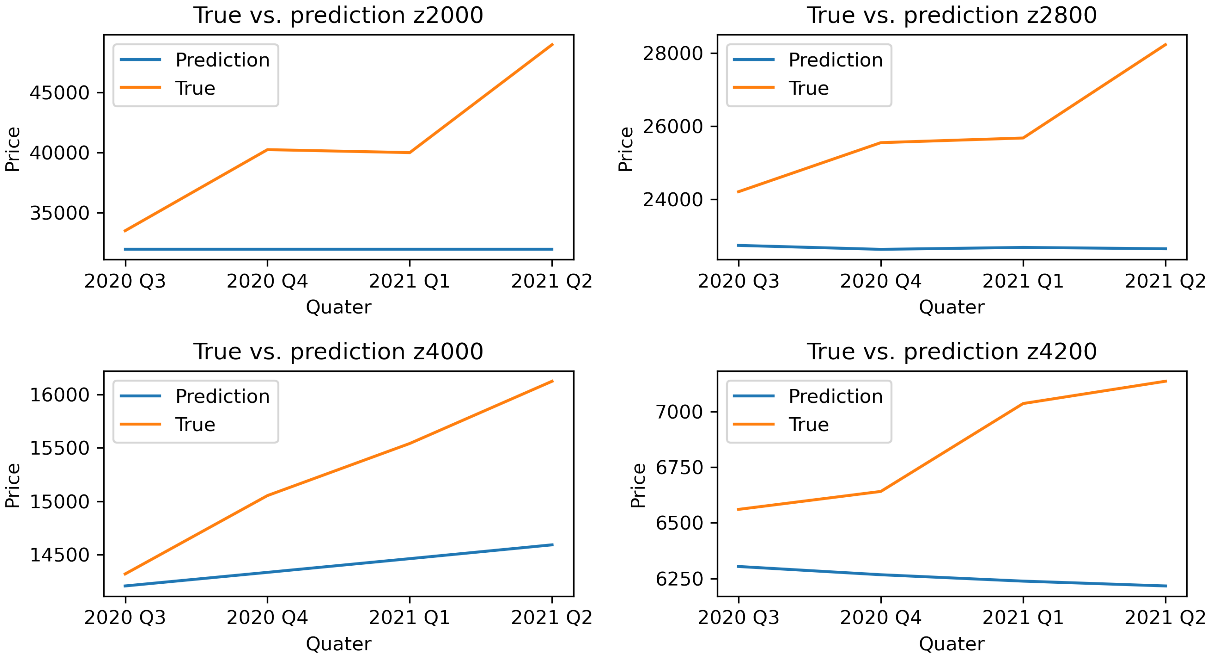

Figure 10: The predictions and the true prices in the test set. There are quite a big difference between the prediction and the true data. When looking at Figure 1 this is not a surprise as the data has a big growth right after the training data ends.

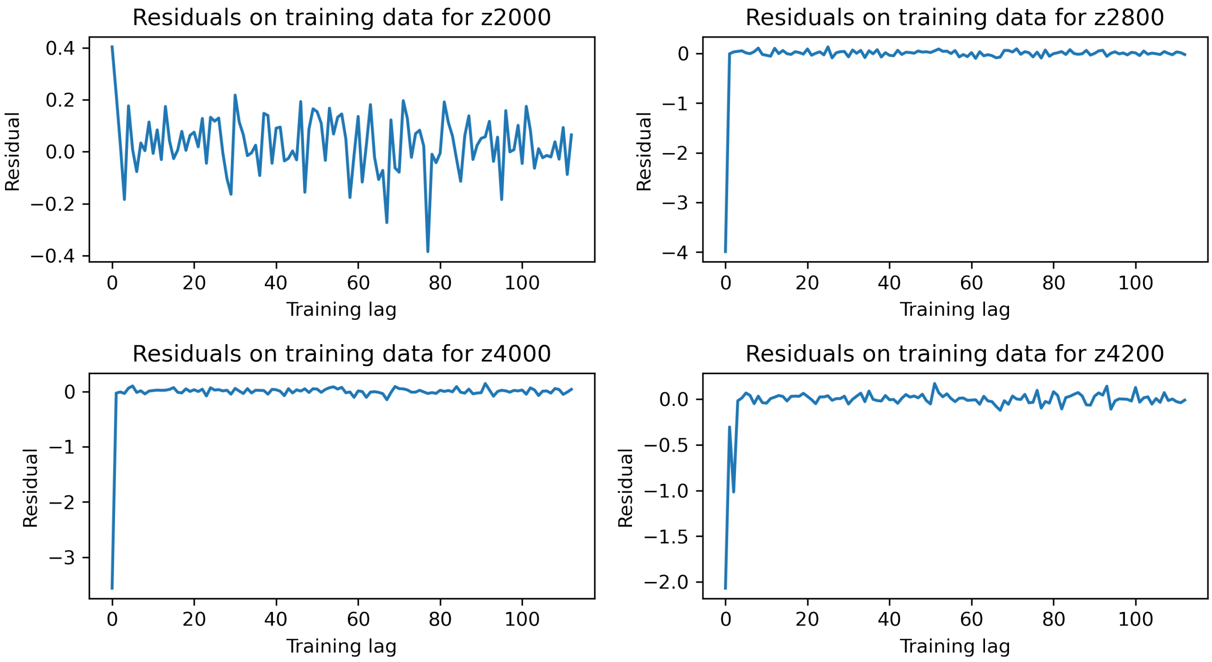

Figure 11: The residuals on the training data for the different models. All residuals seem to be i.i.d with a constant mean and variance. The first lag i ignored for all the residuals as it had a much higher residuals than the other lags. This high values is due to no prior terms to make the prediction on and is thus expected.

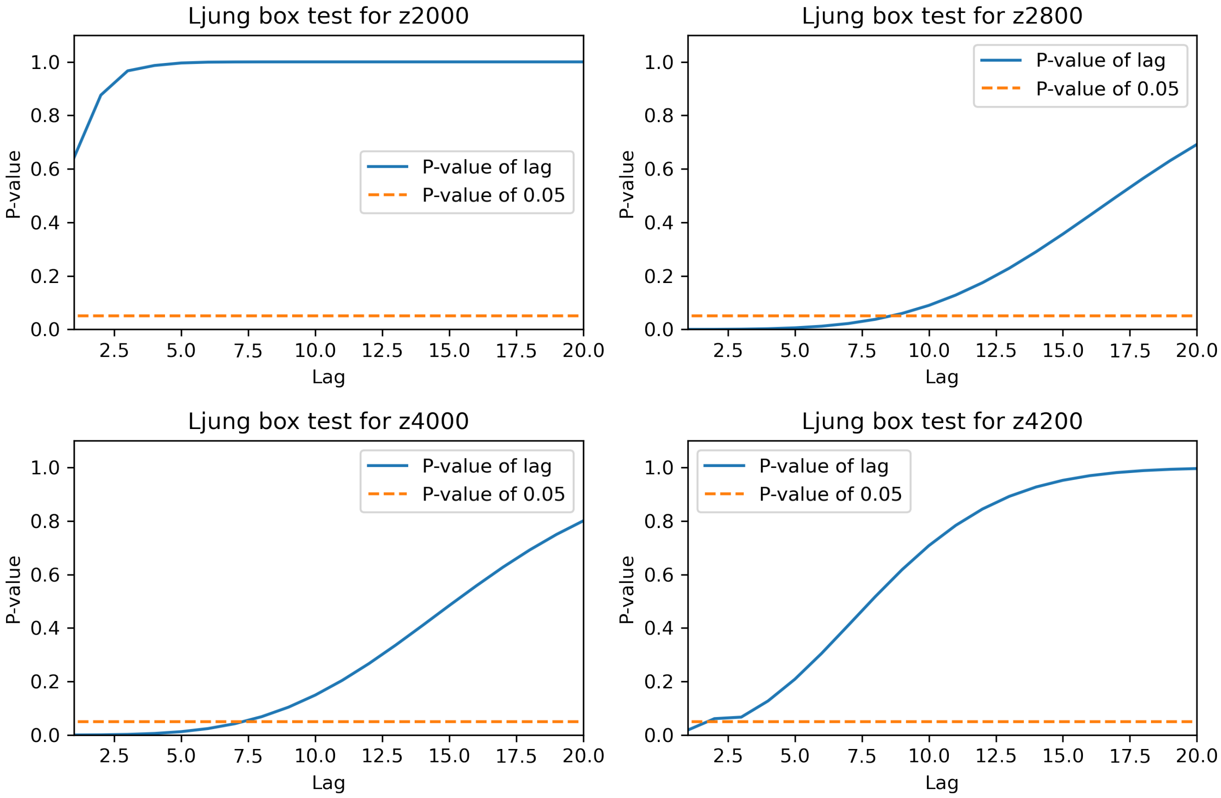

Figure 12: Ljung-Box test of the residuals for of the models for the different prices. All models shows no significant autocorrelations between any of the residuals for the first 20 samples. This shows that the models manages to catch. most of the information available.

AR parameters

	z2000:	z2800:	z4000:	z4200:	CPI:	30 year rate:	Nat rate:
Estimated z2000	0	-0.551	0	0	0	0	0
Estimated z2800	0	0.214	0	0	0.023	-0.012	0.012
Estimated z4000	-0.111	0	0	-0.158	0	0	0
Estimated z4200	0	0	-0.347	0	0	0	0

Table 1: The estimated AR parameters for the MARIMA model.

MA parameters

	z2000:	z2800:	z4000:	z4200:	CPI:	30 year rate:	Nat rate:
Estimated z2000	-0.631	0	0	0	0	0	0
Estimated z2800	0	0	0	0	0	0	0
Estimated z4000	0	0.265	-0.279	0	0	0	0
Estimated z4200	0	0	0	-0.478	0	0	0

Table 2: The estimated AR parameters for the MARIMA model.

Question 4.4

The estimated AR and MA parameters for the MARIMA model are shown in Table 1 and Table 2, respectively. The tables shows how the different prices are predicted using both the prices from other zip codes and input variables, namely the CPI and the two interest rates.

For example will the MARIMA model predict the price for zip 2800 using a AR model consisting of its prior prices, the CPI, the 3 year bond rate and the Nationalbanken rate, all with parameters from Table 1. And to predict price in zip 4200 the MARIMA will use a ARMA model with a AR part consisting of its prior prices combined with a MA part consisting of the prior prices in zip 4200, with parameter values given in Table 1 and Table 2, respectively.

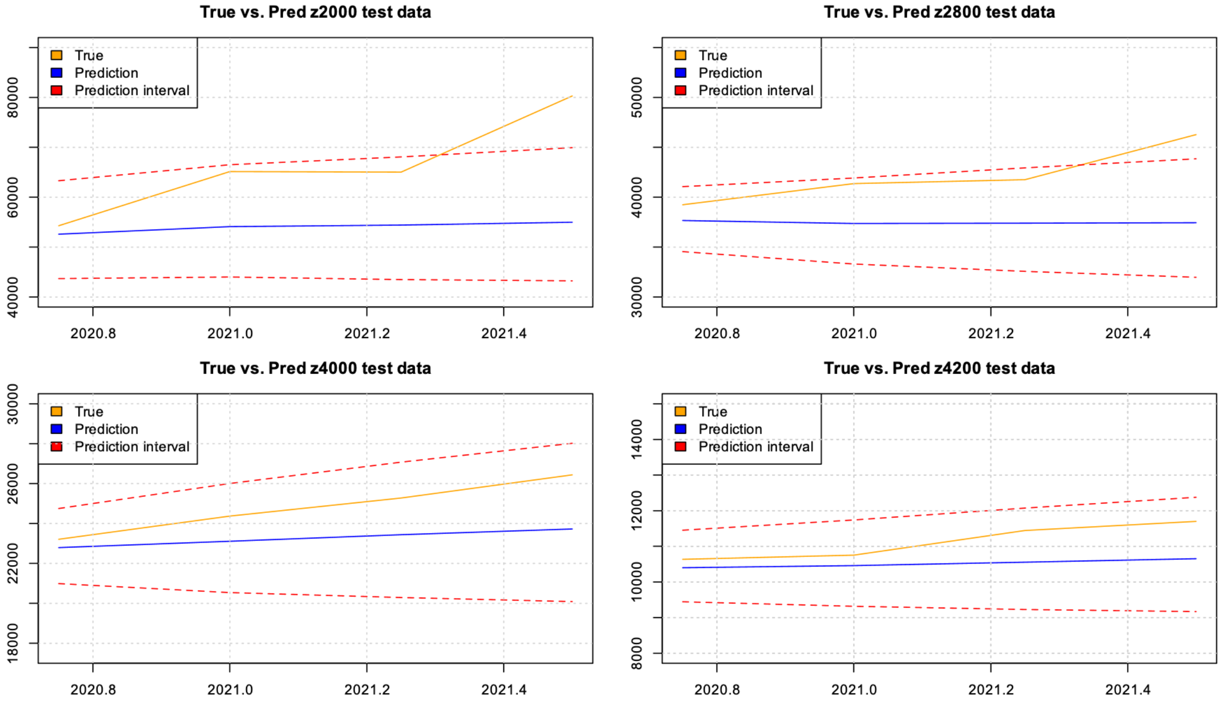

Figure 13: Predictions on test data with prediction intervals.

Zip 2000

Quarter:	True	Predicted:	Upper:	Lower:
2020 Q3	54290	52597.61	63310.74	43697.29
2020 Q4	65148	54111.76	66517.07	44020.02
2021 Q1	65039	54421.01	68078.78	43503.22
2021 Q2	80295	54995.67	69932.63	43249.11

Table 3: Table showing the true and predicted prices for zip 2000. It also includes an upper and lower 95% prediction interval.

Question 4.5

The predictions with prediction intervals for zip 2000, zip 2800, zip 4000 and zip 4200 are shown in Table 3, Table 4, Table 5 and Table 6 respectively. Plots of the true values, predictions and predictions interval are also shown in Figure 13.

As you can see from Figure 13 the model predicts lower prices in all different zip codes. The MARIMA predictions are made from the prices for the different zip codes as well as the CPI and two rates. And by looking at the test data, consisting of the last four quarters, the prices experiences a big increase. Such a big increase are not seen in any of the other inputs, thus the model has no way of knowing that such a big increase in price is going to happen. That is way we see such a big difference between the prediction and the true price.

Zip 2800

Quarter:	True	Predicted:	Upper:	Lower:
2020 Q3	39249	37668.53	41064.74	34553.21
2020 Q4	41368	37371.48	41928.97	33309.37
2021 Q1	41761	37401.7	42943.72	32574.9
2021 Q2	46275	37449.75	43853.29	31981.27

Table 4: Table showing the true and predicted prices for zip 2800. It also includes an upper and lower 95% prediction interval.

Zip 4000

Quarter:	True	Predicted:	Upper:	Lower:
2020 Q3	23211	22792.9	24750.87	20989.82
2020 Q4	24372	23111.85	26005.13	20540.46
2021 Q1	25277	23439.53	27072.5	20294.09
2021 Q2	26437	23725.44	28017.92	20090.59

Table 5: Table showing the true and predicted prices for zip 4000. It also includes an upper and lower 95% prediction interval.

Zip 4200

Quarter:	True	Predicted:	Upper:	Lower:
2020 Q3	10636	10400.79	11452.6	9445.58
2020 Q4	10753	10459.35	11739.58	9318.73
2021 Q1	11444	10555.44	12072.76	9228.82
2021 Q2	11700	10653.09	12376.31	9169.81

Table 6: Table showing the true and predicted prices for zip 4200. It also includes an upper and lower 95% prediction interval.

AR parameters

	z2000:	z2800:	z4000:	z4200:	CPI:	30 year rate:	Nat rate:
Estimated z2000:	0	-0.566	0	0	0	0	0
Estimated z2800:	0	0.21	0	0	0.025	-0.011	0.011
Estimated z4000:	-0.112	0	0	0.165	0	0	0
Estimated z4200:	0	0	-0.359	0	0	0	0

Table 7: AR parameters for the new MARIMA model trained on the whole data set and made for predicting the future.

MA parameters

	z2000:	z2800:	z4000:	z4200:	CPI:	30 year rate:	Nat rate:
Estimated z2000:	-0.606	0	0	0	0	0	0
Estimated z2800:	0	0	0	0	0	0	0
Estimated z4000:	0	0.261	-0.278	0	0	0	0
Estimated z4200:	0	0	0	-0.478	0	0	0

Table 8: MA parameters for the new MARIMA model trained on the whole data set and used for prediction the future.

Question 4.6

In this part we are to remake the model using the same approach as before, but using the whole data set for estimating. The new AR and MA parameters are shown in Table 7 and Table 8, respectively.

Compared to the prior model all estimated prices are built up by combinations of the same input with almost the same parameters. The biggest difference between the parameter value for this model compared to the prior one is that the AR parameters related to the prices increases in magnitude, while the MA parameters mostly decreases in magnitude.

The AR part of the model describes how much the future prices are a function of the past prices, while the MA term is modelling the error term as a combination of the prior error term. Thus an increase in the magnitude of the parameters of the AR part and a decrease in the magnitude of the parameters of the MA part implies that the new model relies more on the prior prices, and less on the error term.

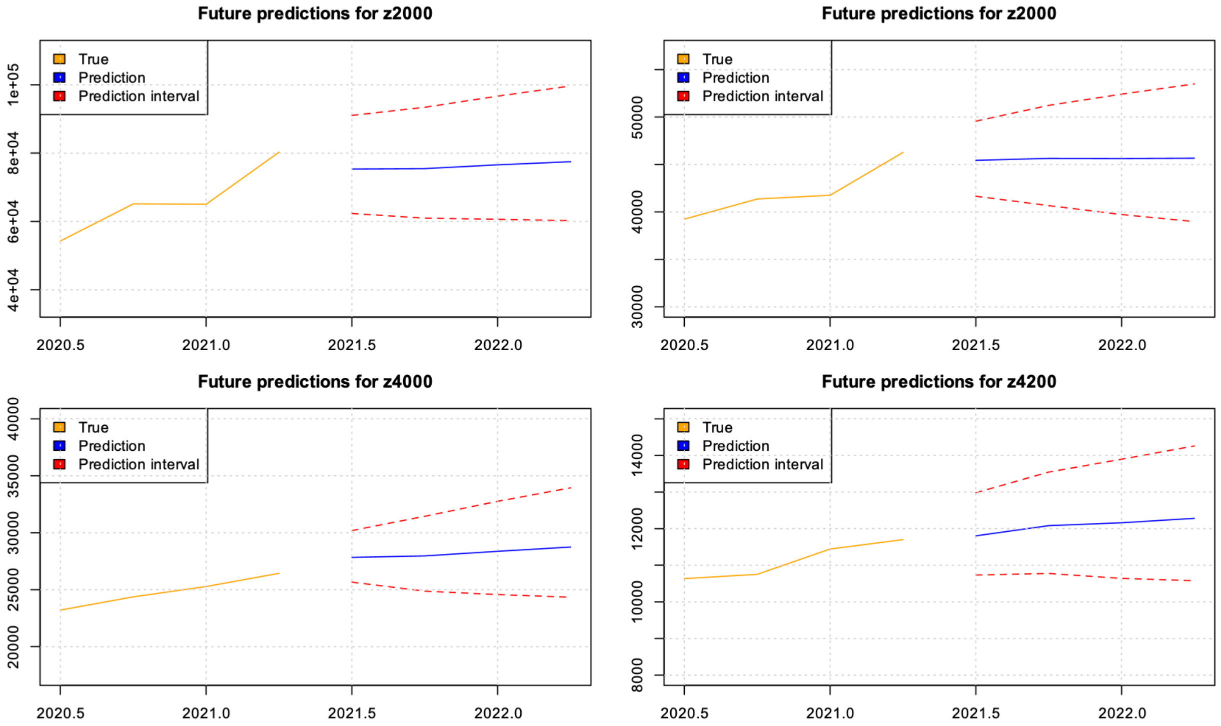

Figure 14: Future price predictions for the different zip codes. All are made for one year into the future. The plot also includes a 95% prediction interval.

Question 4.7

The new price predictions are shown with a 95% prediction interval in Figure 14. They are also presented in Table 9, Table 10, Table 11 and Table 12.

By looking at Figure 14 and comparing the tables showing the predicted values, Table 9, Table 10, Table 11 and Table 12, and the tables showing the true values, Table 3, Table 4, Table 5 and Table 6 it is clear that I should invest my savings in zip 4000 or zip 4200. This is because these are the only zip codes where my model predicts an increase in the housing prices. For zip 2000 and zip 2800 my model predicts that the housing prices will decrease the next year.

Future prediction for zip 2000

Quarter:	Prediction:	Upper:	Lower:
Q3 2021	64436.96	91174.96	62398.2
Q4 2021	75426.48	93592.17	60997.24
Q1 2022	75557.03	96875.27	60667.61
Q2 2022	76662.84	99866.77	60260.09

Table 9: The future prediction for the prices in zip 2000 with a 95% prediction interval.

Future prediction for zip 2800

Quarter:	Prediction:	Upper:	Lower:
Q3 2021	42647.02	49568.2	41660.32
Q1 2022	45442.57	51233.88	40670.48
Q2 2022	45647.63	52415.18	39732.89
Q3 2022	45635.58	53494.3	38988.93

Table 10: The future prediction for the prices in zip 2800 with a 95% prediction interval.

Future prediction for zip 4000

Quarter:	Prediction:	Upper:	Lower:
Q3 2021	25754.26	30188.83	25663.78
Q1 2022	27834.5	31449.15	24873.77
Q2 2022	27968.89	32760.41	24574.07
Q3 2022	28373.52	33944.31	24343.38

Table 11: The future prediction for the prices in zip 4000 with a 95% prediction interval.

Future prediction for zip 4200

Quarter:	Prediction:	Upper:	Lower:
Q3 2021	11408.42	12981.55	10733.01
Q1 2022	11803.86	13544.93	10776.65
Q2 2022	12081.76	13897.63	10641.39
Q3 2022	12161.01	14260.73	10579.07

Table 12: The future prediction for the prices in zip 4200 with a 95% prediction interval.

List of Figures

References

[1] Madsen H.(2007) *Time Series Analysis*, Chapman & Hall/CRC.

[2] Spliid, H. (2016) *Estimation of Multivariate Time Series with Regression Variables*, DTU Statistical Consulting Center, DSCC