Predicting House Prices

With OP Bank

By:

Anna Lebedeff

Mohamed Amine Mkhaibou

Khoa Nguyen

Sami Valkamaa

Table of Contents:

[**Abstract**](#_5fqrdx20qvnd) **2**

[**1. Data**](#_t2jmlz2x4d8j) **3**

[**2. Exploratory Data Analysis**](#_8n0mon1qczpl) **4**

[**3. Methods**](#_20sxoom57w57) **8**

[3.1 Linear Regression](#_pwb7z4ts5dj4) 9

[3.2 Facebook Prophet](#_kq4sdibnz4mc) 9

[3.3 SARIMA Model](#_4tm5ljq9yhn5) 9

[**4. Comparison of the Algorithms: Accuracy and parameter tuning**](#_9pj3ziaegrzs) **10**

[4.1 Linear Regression and Prophet](#_bidbahjrjzha) 11

[4.2 SARIMA Model](#_la7i2yah0fha) 14

[4.3 Discussion](#_k9ywz8w0ups3) 16

[**5. Result and Visualization**](#_eqv01kbg273h) **17**

[5.1 Linear Regression and Prophet](#_45a13llxeher) 17

[5.2 SARIMA](#_v3tzp0l46w02) 18

[**6. Ethical Issues**](#_5cqw6iugs7di) **20**

[**7. Conclusion and Future Prospects**](#_36508ki0pse2) **21**

[**8. Roles of the team members**](#_6z30r58dbl1) **21**

# Abstract

This project is part of the course CS-C3250 - Data Science Project 2020 and conducted in collaboration with OP Bank, supervised mainly by Juha Vesanto.

OP Financial Group is the largest housing lender in Finland. Therefore, future house prices reflect on OP's ability to gain revenue, analyze risks, and evaluate its solvency position. By predicting future house prices, OP is also able to make better budget and other business decisions related to home loans.

Therefore, the goal of the project was to predict housing prices until the end of 2021. The project's requirements included that OP can follow house prices on a quarter-to-quarter basis when provided a region and a housing type.

In order to yield the necessary predictions, the following steps were taken. First, the data was gathered from Statistics Finland's PxWeb databases. The gathered data was then explored with traditional exploratory data analysis (EDA) methods, after which the data was preprocessed and cleaned, and missing values were imputed to gain a full data set. Lastly, three models were applied to predict the future house prices: Linear Regression, Facebook Prophet, and SARIMA model. The results of each prediction model were visualized and gathered into a table. Additionally, to improve the readability of house price trends in each region, an index was provided to indicate how the house prices have changed in the region compared to the pivot year 2018.

From the three models, linear regression provides reasonably accurate predictions for short time predictions. The Facebook prophet is highly sensitive for specific parameters but for most regions the model provided reasonably accurate results even in the long-run. Additionally, the SARIMA model seemed to have good predictions but the problem with this is that it is computationally very expensive.

From our side, the project's motivation was to gain experience in utilizing different Python libraries in Data Science as well as applying our current knowledge about the field to solve some real-life problems. Additionally, for us the goal was to learn how versatile Data Science can be as a field, what different processes go into a Data Science project, and what it is like to work for a client. OP’s objective on the ther hand was to gain predictions about the house prices as they have an effect on the organization’s profits, risks and solvency.

# 1. Data

Our client suggested we study the datasets from Statistics Finland’s PxWeb database. However, before settling into a database, we examined different websites to establish which database suits our needs the best and is the most credible. Some of the reviewed databases included Statistics Finland, [Asuntojen.hintatiedot.fi](https://asuntojen.hintatiedot.fi/), and asuntojenhinnat.fi . We concluded to gather our data from Statistics Finland since the database is the only Finnish public authority established for statistics, produces the vast majority of Finnish official statistics, and is a significant international operator in statistics. Additionally, the PxWeb database from Statistics Finland’s PxWeb databases had the most data (starting from as early as 2006) with more parameter variety than the other examined databases. Therefore, the data in this project was gathered from Statistics Finland's PxWeb databases.

In the PxWeb database, there exist eight different datasets in which the data is provided quarterly. We perceived that the gathered data must be in this format as our client requested us to predict Finland's house prices quarterly. From the eight options that provided data quarterly, we chose to utilize a dataset called "112r -- Average prices of old dwellings in housing companies and number of transactions, quarterly, 2006-, 2006Q1-2020Q3\*". We concluded with this dataset as it has the earliest starting year, providing us the most data to work with, and the dataset presents the average prices per mˆ2 for each region and house-type.

From the Statistics Finland's PxWeb databases, the data was downloaded into three separate CSV files (one-room flats, two-room flats, and three-room flats). After obtaining the data, the data was examined. We observed that one-room flats had 22.59% of missing data, and therefore, imputations would be required. Despite that two-room and three-room flats had less than 4% of missing data, we also decided to impute these datasets to have a full set of data to work with. However, before imputating the data, it was decided with the client that a threshold of 70% would be applied to the datasets. This way, we were able to filter out the regions that had less than 30% of data which could result in inaccurate imputations. For one-room flats, 13 regions were filtered out. For two-room flats only 1 region was required to be removed and 2 regions were removed in three-room flats as they missed over 70% of the data.

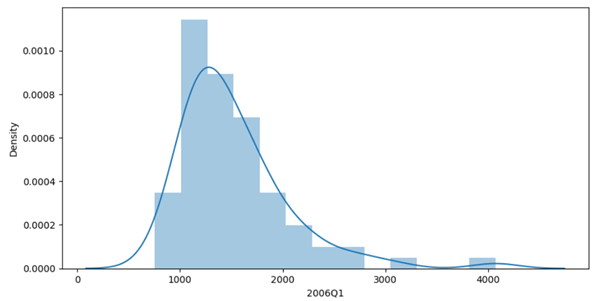
The deleted regions are listed below for each house type:

* 'Vantaa 1', 'Järvenpää', 'Kerava', 'Rauma', 'Hämeenlinna 2', ‘Riihiäki’, 'Joensuu 2', 'Southern Ostrobothnia', 'Seinäjoki', 'Vaasa 2', 'Central Ostrobothnia', 'Kokkola', and 'Kajaani'
* Two-room flats: ‘Joensuu 2’
* Three-room flats: 'Hämeenlinna 2', 'Joensuu 2'

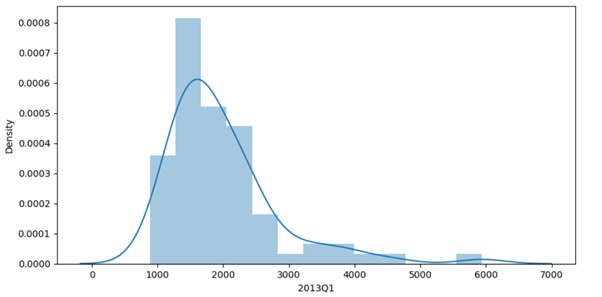
The imputations for each CSV file were conducted with a Sklearn's iterative imputer. The imputer estimates the missing values by creating a function based on the other features. More specifically, the iterative imputer treats each region as output, and each quarter is as the feature based on which regression is constructed to estimate the missing values. The regression is constructed iteratively for each feature. After establishing a full data set for each CSV file, we went through the imputed values to ensure that there will not be any outliers or unaccurate imputations. Imputations that differed from their neighbouring values by 450 or more were replaced with the mean of the neighbouring values.

# 2. Exploratory Data Analysis

We are going to show you one out of three EDAs because they are quite similar. We have therefore decided to present the two-room flat EDA:

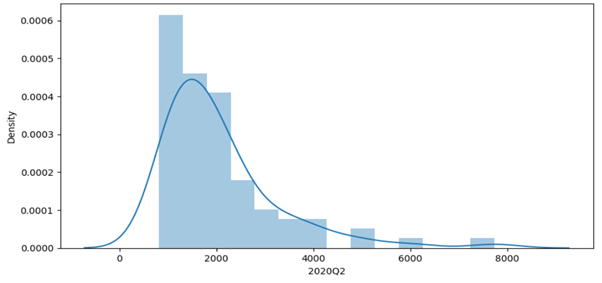
Firstly, the axes plot of two-room flat in 2006Q1:

* We can see that most of the values are between 1000 and 1500.

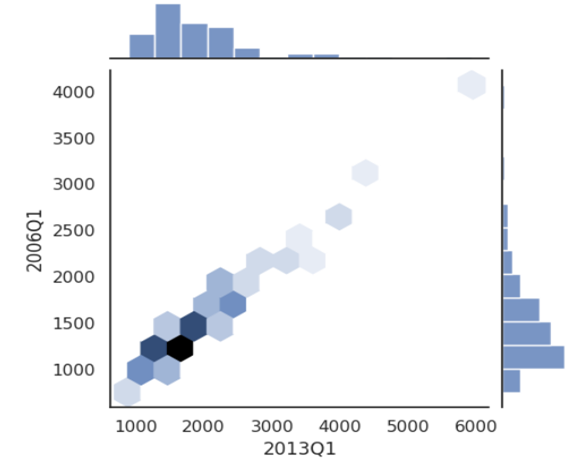
Secondly, the axes plot of two-room flat in 2013Q1:

* We see that the prices are increased because values close to 5000 and 6000 have appeared, and most of the values are between 1500 and 2000.

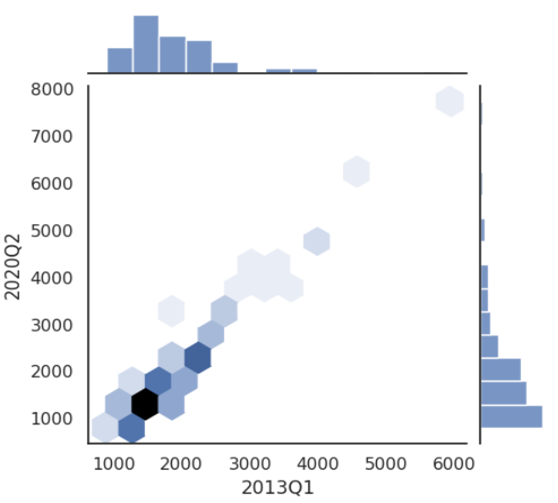
Finally, the axes plot of two-room flat in 2020Q2:



* It is seen that the prices are increased because some values reach 6000 and almost 8000 have appeared, but most of the values are between 1000 and 1500.

We can see that more clearly with the ‘join plot’ between the first quarter of 2006 and 2013:

* We can see that in 2006Q1 the top value was 4000, and most values are between 1000 and 1500. For 2013Q1 the top value is 6000, and most values are between 1500 and 2000.

Next, the ‘join plot’ between the first quarter of 2013 and the second quarter of 2020:

* We can see that in 2013Q1 the top value was 6000, and most values are between 1500 and 2000. For 2020Q2 the top value is 8000, and most values are between 1000 and 1500.

To see the evolution of the price per square meter and the difference between the regions in the prices, we can expose some statistics:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **2006Q1** | **2006Q2** | **2006Q3** | **2006Q4** | **2013Q1** | **2013Q2** | **2013Q3** | **2013Q4** | **2020Q1** | **2020Q2** |
| **mean** | 1524.99 | 1548.27 | 1535.67 | 1579.40 | 1979.62 | 2014.37 | 2012.13 | 2010.55 | 2075.26 | 2069.32 |
| **std** | 536.95 | 548.05 | 580.63 | 580.87 | 847.42 | 858.59 | 878.77 | 880.52 | 1157.91 | 1225.41 |
| **min** | 755.00 | 885.00 | 713.00 | 869.00 | 885.00 | 904.00 | 988.00 | 984.00 | 679.26 | 809.93 |
| **max** | 4070.00 | 4099.00 | 4120.00 | 4295.00 | 5936.00 | 6040.00 | 6214.00 | 6178.00 | 7548.00 | 7735.00 |

* You can see that the average difference between 2006 Q1 & 2020Q2 = 544.33

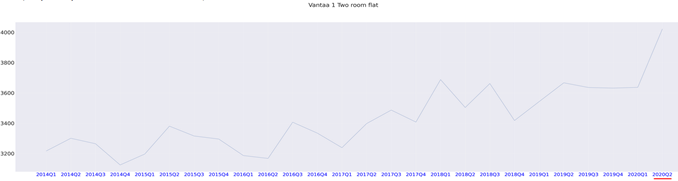
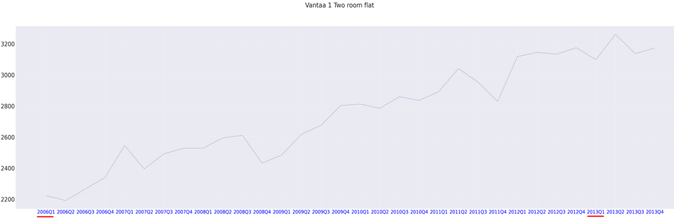
Which means that over this entire period the price has increased by around 544 euros.

* And the standard deviation difference between 2006 Q1 & 2020Q2 = 688.46

Which means that over this whole period, the price difference between the regions has increased by around 688 euros.

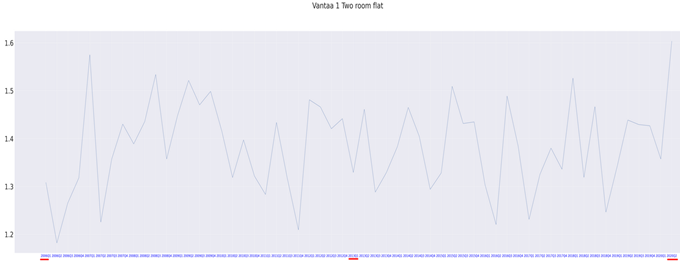
We choose Vantaa 1 as an example, to show you the whole price evolution in this region:

The chart which contains all quarters is not clear for this reason we have split the chart in half. The red lines are under 2006Q1, 2013Q1 and 2020Q2:



* It seems that there is a continuous increase in the price throughout this period with some drops from time to time. And generally this is the general case for all regions and all types of flats.

As an option, after that we have normalized the quarters with the Z score method, we can see the rises and drops clearly:



# 3. Methods

In this project, three different time series models were applied to predict house prices in Finland: linear regression, Facebook Prophet, and the SARIMA model. With the data consisting of the house prices of different regions in Finland, each region was treated as an individual time series, and therefore, the predictions were carried out separately for each time series. However, when starting the project, it was realized that each region is somewhat a non-stationary time series with different trends, seasonality, and substantial fluctuations. Therefore, for each model, a unique benchmark/ grid search algorithm was implemented to provide the best result possible for each time series. The results of each model are to be displayed and discussed in this session.

### 3.1 Linear Regression

Linear Regression models the linear relationship between a dependent variable and one or more explanatory variables. Since the data in the project is working with a time series, the dependent variable in this case is the housing price and time is the explanatory variable. In other words, the house price is the value Y that is defined by X which takes in value 2006Q1, 2006Q2 and so on multiplied with the regression coefficient. .The sklearn Linear Regression model calculates the best parameters/ trend from the data and applies the parameter to estimate future values from the future years.

In order to find the most optimal start year, the data is splitted into a train (Before 2018) and a test set (2018 onwards) with the train set starts at different years. Linear Regression is then applied to predict the interval of 2018 to 2020Q2 and the result is compared to the test set. The start year with the lowest Mean Absolute Error is then chosen to be the start year to predict the future housing price of that region.

### 3.2 Facebook Prophet

Prophet is a procedure for forecasting time series data based on an additive

model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. The Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.” (Source: [facebook.github](https://facebook.github.io/prophet/))

Since the Prophet model not only returns the prediction value but also an error interval, finding the optimal start year for the Prophet is different from Linear Regression. The train set in this case is the whole time series itself with different start years. The “start year” with the smallest error interval is chosen to be the result of the model itself.

### 3.3 SARIMA Model

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

The autoregressive (AR) part refers to the utilization of lagged values of our target variable to make our prediction. The order of the autoregression, denoted with p, tells us how far in the past we are looking to make future predictions. The equation for calculating the current value of a series is given by Y = B0 + B1\*Y\_lag1 + B2\*Y\_lag2 + … Bp\*Y\_lagp, where p is the order of autoregression and B0…Bp are the regression betas that we fit to our model.

Differencing or “integrated” (I) part of the model changes the autoregressive function to Y\_next – Y = B0 + B1\*(Y - Y\_lag1) + B2\*(Y - Y\_lag2) + … Bp\*(Y - Y\_lagp). Differencing brings a time series to a more stationary form with constant variance and mean zero. Models built on stationary data tend to be more robust.

A moving average (MA) model is represented by the following equation: Y = B0 + B1\*E\_lag1 + B2\*E\_lag2 + … + Bq\*E\_lagq. Here, B0…Bq are the betas we are trying to fit, q is the order of the moving average, and E is the error which represents the random residual changes between our model and the target variable.

Finally, SARIMA adds three new hyperparameters to the ARIMA model to specify the autoregression, differencing and moving average for the seasonal component of the series, as well as an additional parameter for the period of the seasonality. In this project we have a seasonality of four as we have four data points, one per quarter of a year.

# 4. Comparison of the Algorithms: Accuracy and parameter tuning

Before introducing the predictions generated by the three models, we studied the performance of each model on our datasets. To do so, the data was split into train set (2006Q1-2017Q4) and test set (2018Q1-2020Q2) and each model attempted to predict the house prices from the original dataset for quarters 2018Q1 to 2020Q2 starting at different years. The predicted values were then compared to the original data to see which model predicts the values the most accurately. Additionally, the experiment tested the effect of parameter tuning on the models. The data contains several “ugly” regions with spikes and strong fluctuation which proves to be challenging for the models. Especially since there are too few data points in the time series because the data is divided into quarters, not months or days that can presumably enrich the data, parameter tuning for the models themselves is important.

In this project, parameter tuning is done for both the Prophet and SARIMA. For the Prophet, the parameter that is needed for tuning is changepoint\_prior\_scale. Since the data’s unit is quarter, other factors such as holidays can be ignored, alongside with n\_changepoint (number of changepoint) because the length of the actual time series is too short. Therefore, only changepoint\_prior\_scale is tuned. changepoint\_prior\_scale indicates how flexible the changepoints are allowed to be, therefore the higher the more flexible but also more chance of overfitting. Three values were tested, 0.05 (default of Prophet), 0.1 and 0.2 (0.2 is the maximum value tested since the time series are too short, increasing the value can lead to overfitting) to see which gives the lowest Mean Absolute Error between the predicted values and the real values.

### 4.1 Linear Regression and Prophet

For Linear Regression and Prophet, with the parameter tuning, histogram was chosen to be the plot for interpreting the Mean Absolute Error from Prophet with different parameters as well as the difference between Linear Regression and Prophet. Each region is predicted with different start years and the prediction is compared with the true values. A table is created to save the optimal start year of each region, which is the year that gives the best prediction/ smallest Mean Absolute Error. The result table is then plotted as histogram to visualize the distribution of the MAE values and optimal start year.

One-room flats:

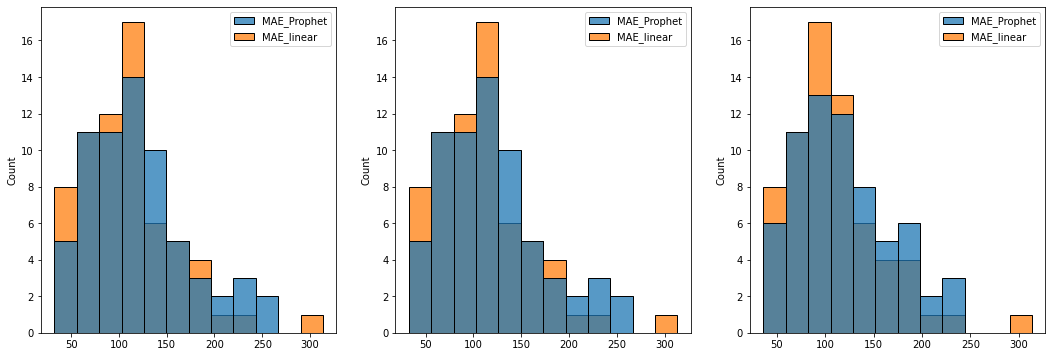
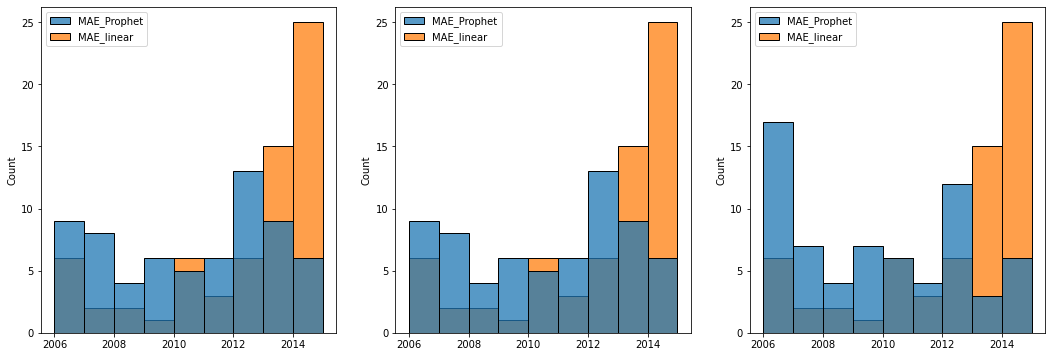


Fig.Distribution of MAE values between changepoint\_prior\_scale 0.05, 0.1 and 0.2 for One Room Flat

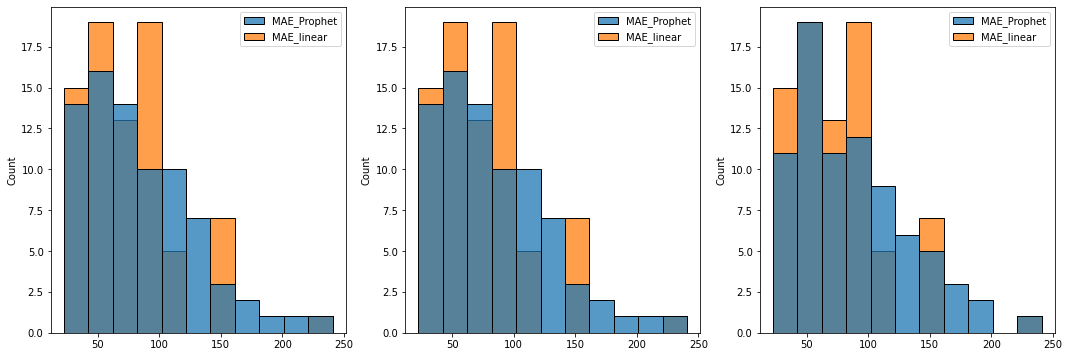
For One Room Flat, the Mean Absolute Error(MAE), MAE decreases as the changepoint\_prior\_scale increases from 0.05 to 0.2. The majority of regions gives predictions with error of around 100 which is acceptable with house prices being more than 1000 in average.

The 0.2 changepoint\_prior\_scale is the best plot with the values are less than 250 in error except for one outlier.

Fig.Distribution of optimal start year for One Room Flat

For One Room Flat, the optimal years of Linear Regression heavily skewed towards 2013, 2014 and 2015 as expected due to spikes in the data as well as the drop in house prices in 2008. The optimal years of the Prophet are more spread out. The most notable difference between the three plots is the changepoint\_prior\_scale of 0.2 tends to use the whole time series, with 2006 as the start year. This can be due to the fact that there are more changepoints that give better prediction, especially the changepoints are more flexible and therefore adapt to fluctuations really well to give more accurate prediction.

Two Room Flats:

Fig. Distribution of MAE values between changepoint\_prior\_scale 0.05, 0.1 and 0.2 for Two Room Flat

For Two Room Flat, the Mean Absolute Error are heavily left skewed. The majority of MAE values are less than 100. The difference between the three plots is the fact that for changepoint\_prior\_scale 0.2, there are some less regions which prediction gives MAE value more than 200 and there is even one region with MAE value close to 250.

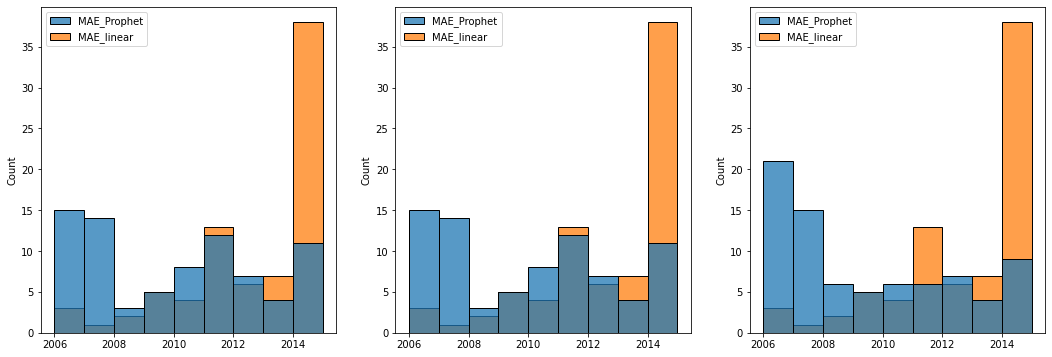
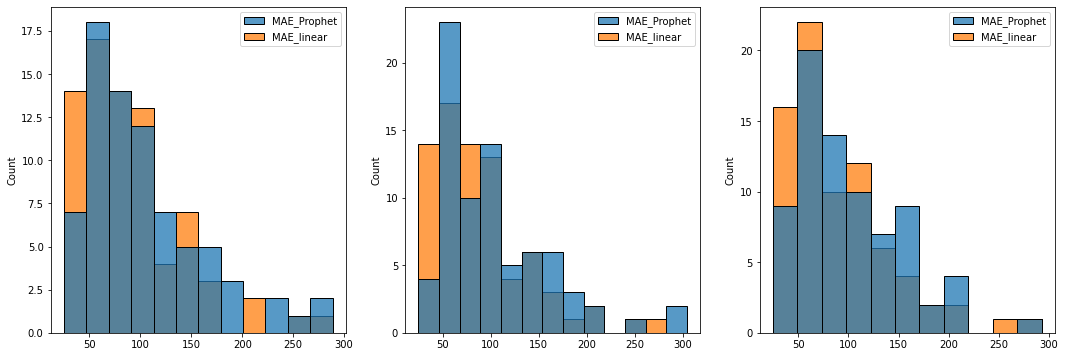


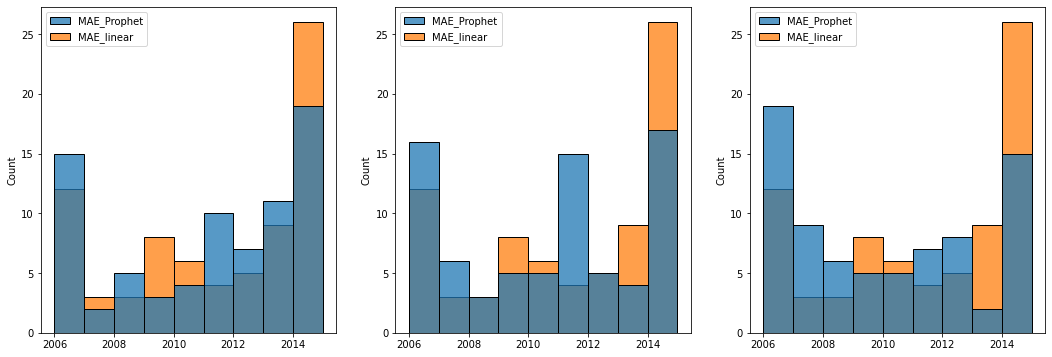
Fig.Distribution of optimal start year for Two Room Flat

For Two Room Flat, the optimal start years for Linear Regression and Prophet behaves the same as for One Room Flat.

Three Room Flat:

Fig. Distribution of MAE values between changepoint\_prior\_scale 0.05, 0.1 and 0.2 for Three Room Flat

For Three Room Flat, the distribution of Mean Absolute Error is the same as in Two Room Flat. The most notable difference in between the third histogram is that the number of region that has MAE value more than 225 is less than the other two.

Fig. Distribution of optimal start year for Three Room Flat

For Three Room Flat, the behavior of Linear Regression stays the same. For the Prophet, although there is an increase in the number of regions using 2015 as the optimal year, the increase changepoint\_prior\_scale reduces this effect similarly to the other two apartment types.

### 4.2 SARIMA Model

The order of both non-seasonal and seasonal parts of the model are usually acquired from autocorrelation and partial autocorrelation functions. Autocorrelation function gives the total direct and indirect correlation of a past value and the current value whereas partial autocorrelation function tells the direct correlation of these values. Looking at the non-zero lags of the autocorrelation function, we can normally determine the order of the MA part of the model. The autocorrelation function is presented below (not all lags are presented in the picture).

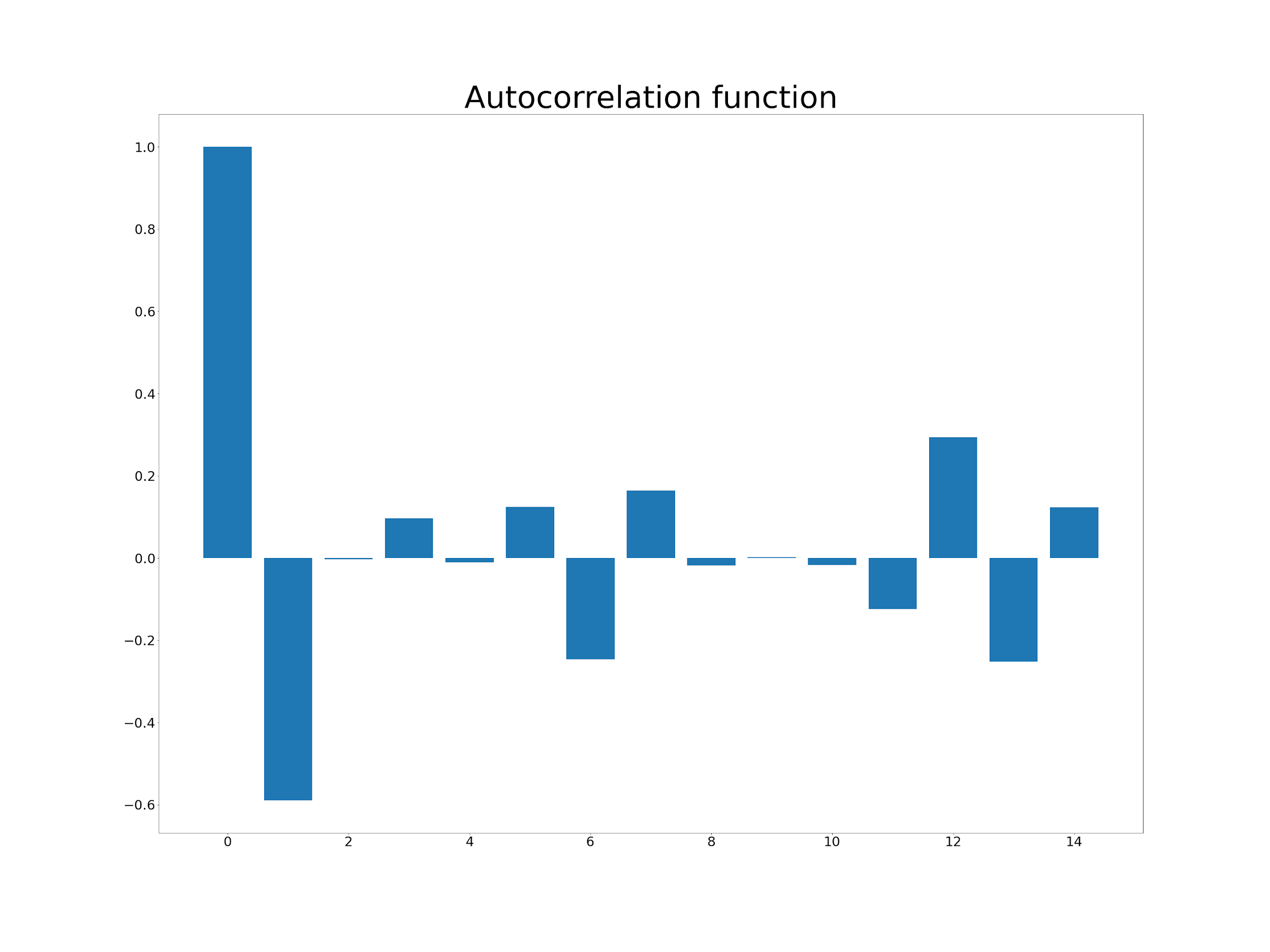


Fig. The autocorrelation function of the complete time series

On the other hand, the partial autocorrelation function gives us the order of the AR part as relevant lags on the plot. As with the autocorrelation, the partial autocorrelation function also presented in the figure below. However, in this project this approach is not applied as the time series is complicated and thus accurate orders are not obtained.

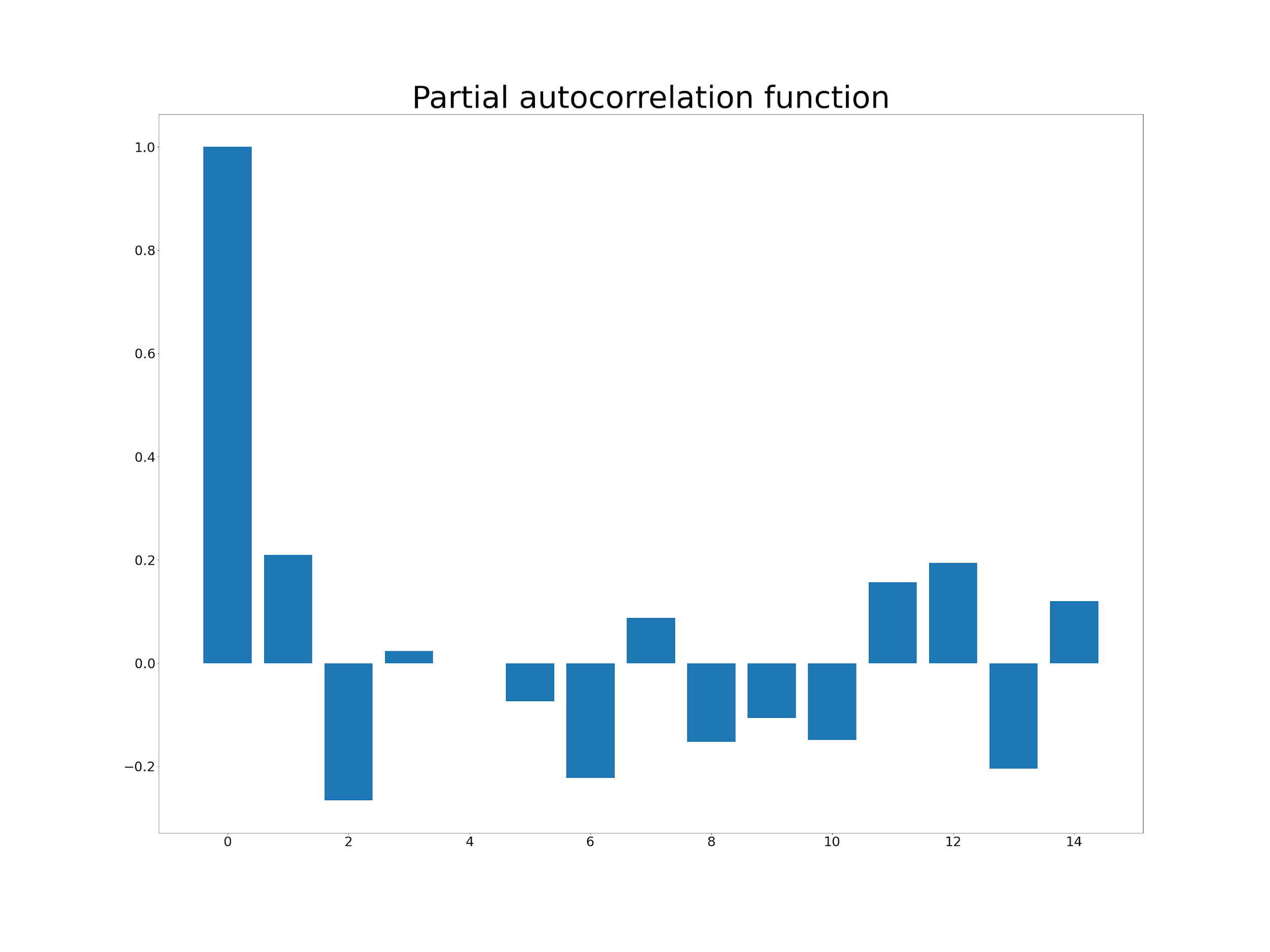


Fig. The partial autocorrelation function of the complete time series

A grid search over the parameters - orders, seasonal orders, and start year for the use of data - was constructed instead to find the best orders for the model to minimize the mean absolute error and the akaike information criterion in training the model on known data. An example of a training prediction can be seen in the figure below. It is of note that due to the computation time of the grid search performed (~8 hours), it was not feasible to provide complete tables or plots to contrast the performance of the model between different regions.

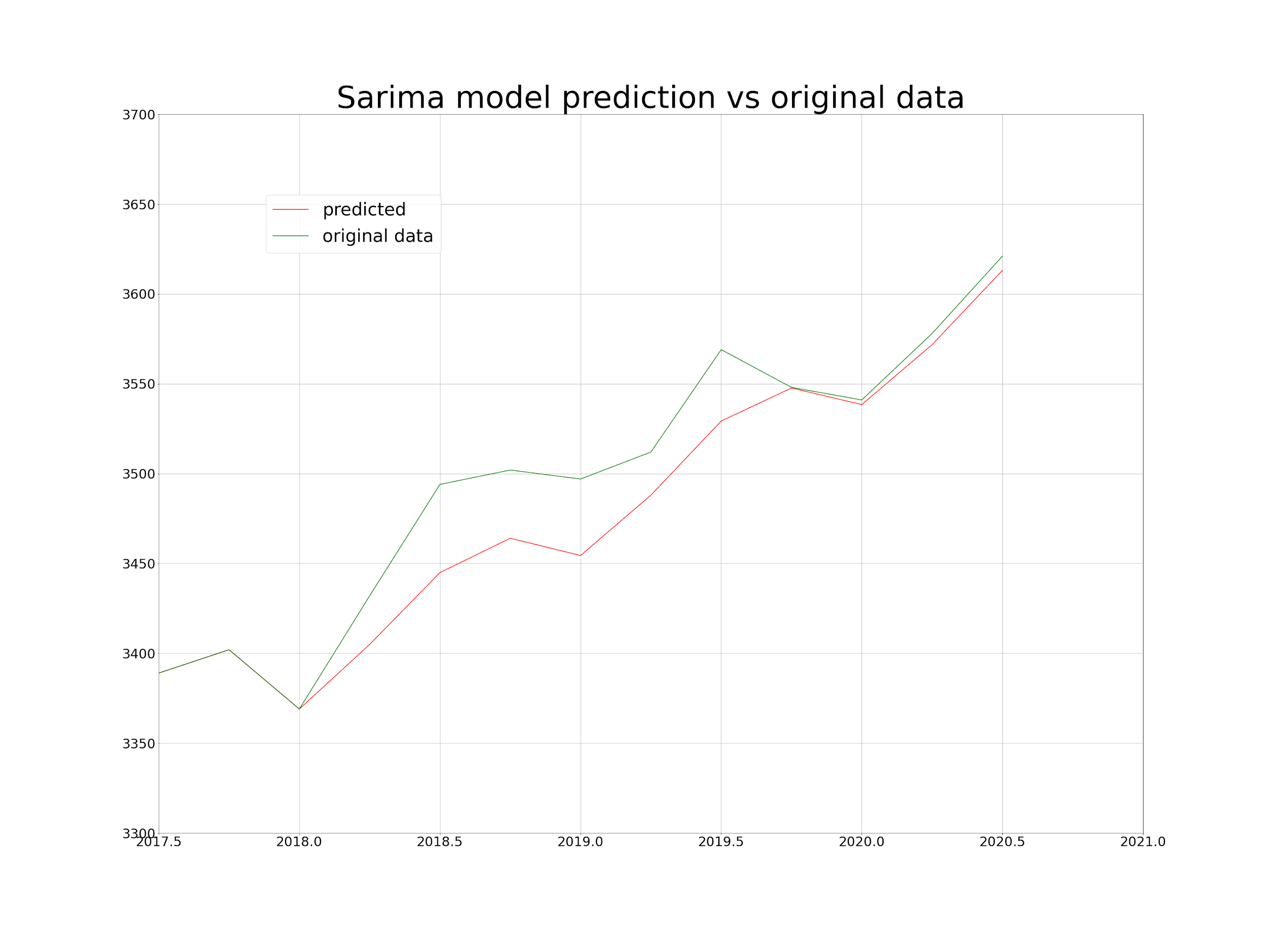


Fig. Sarima model training prediction example on total flats in Uusimaa region

With the prediction interval in training (and also later in the predictions) being so short, the best starting year for the prediction also seemed to be short with 2016 achieving the lowest mean absolute error of approximately 6.8 €/m^2. However, lengthening the prediction interval, more years are needed to make an accurate prediction, and the prediction mean absolute error increases.

### 4.3 Discussion

Both models, SARIMA and Prophet require more data in order to strengthen their predictions while Linear Regression does not behave the same. Also, with the correct combination of parameters, Prophet and SARIMA proved to be viable models that can predict the house prices accurately. Mean absolute error-wise, SARIMA proves to be better than Prophet. However, this conclusion is not absolute since the run-time that SARIMA requires to predict just one region is considerably larger than Prophet which makes it harder to retrieve how the model predicts with the whole dataset. By testing the model performance on the whole dataset, histograms can be plotted in order to help better compare SARIMA with Prophet which provides a much better insight to which model performs better and with which combination of parameters.

# 5. Result and Visualization

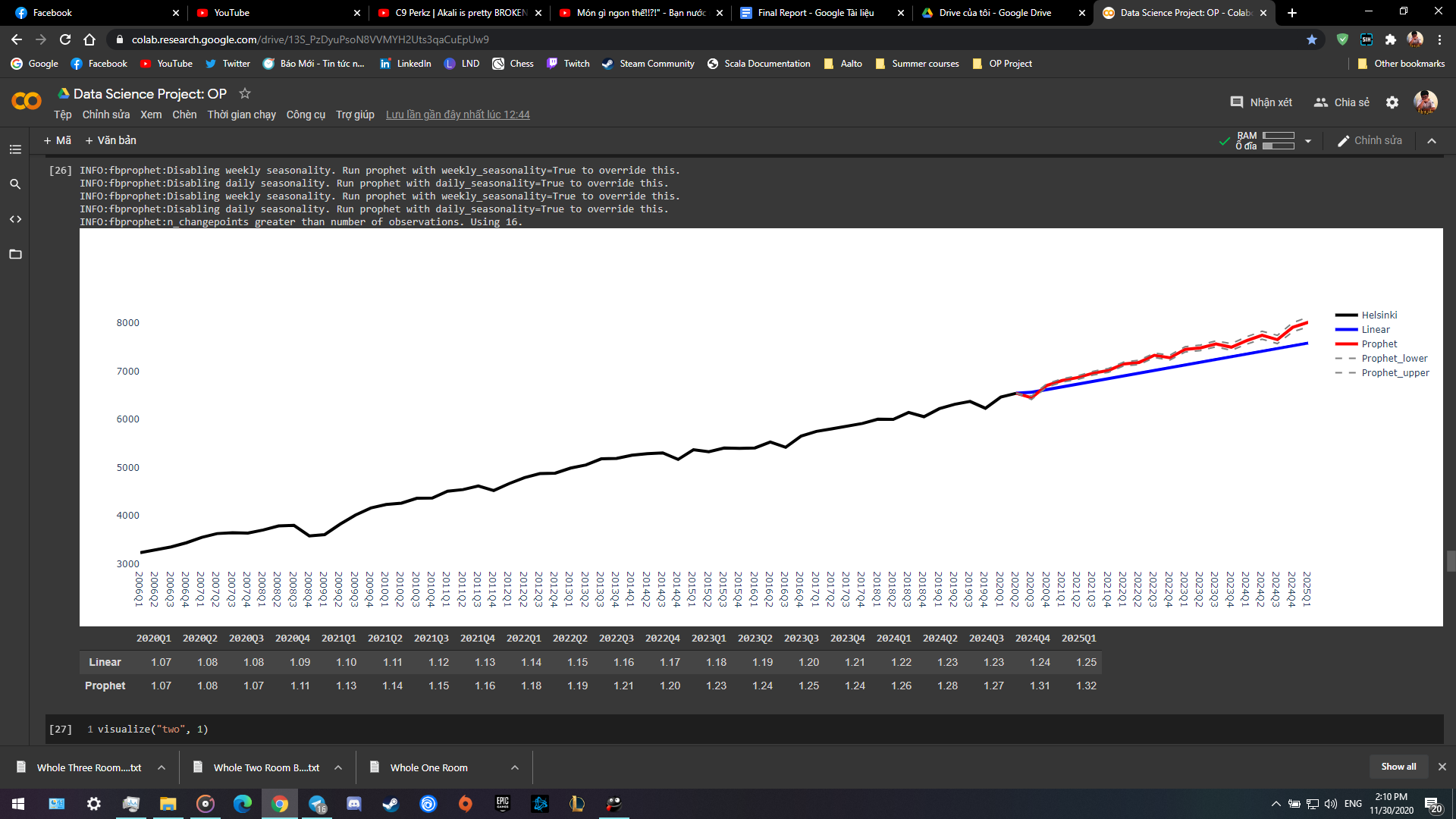
### 5.1 Linear Regression and Prophet

For visualization, a function is implemented where users can type in the type of apartment and the region they want to predict, the output is an interactive plot as well as a table that shows the predicted value as index with 2018 the pivot. The majority of regions return a smooth and very closed and reasonable prediction. However, there are still “outliers” that give mediocre predictions. This effect is caused by mostly the fluctuation of house prices in those regions are rough, mostly due to Iterative Imputer. In order to solve this problem, not only data pre-processing needs to be improved such that smoother time series can be produced and more data should be gathered so that the model have more training data to work with but also more complicated model or Prophet and SARIMA must be tuned accordingly to each region in order to predict the most accurately.

Predictions for all regions can be viewed from the following link: <https://drive.google.com/drive/folders/1K9Ue3oSryTES8MSYCcl2ceZhtrnKwddM?usp=sharing>

From the provided Google Drive, one finds three types of files: price predictions as text files, price predictions as Excel files and price indexes as Excel files. The price indexes utilize 2018 as their pivot year. The price for the pivot year has been calculated as the average of the four quarters in 2018 which has then been compared to each quarter separately to see the trend of the region as percentage. For each region, a unique pivot year has been computed.

Some result of the visualization function:



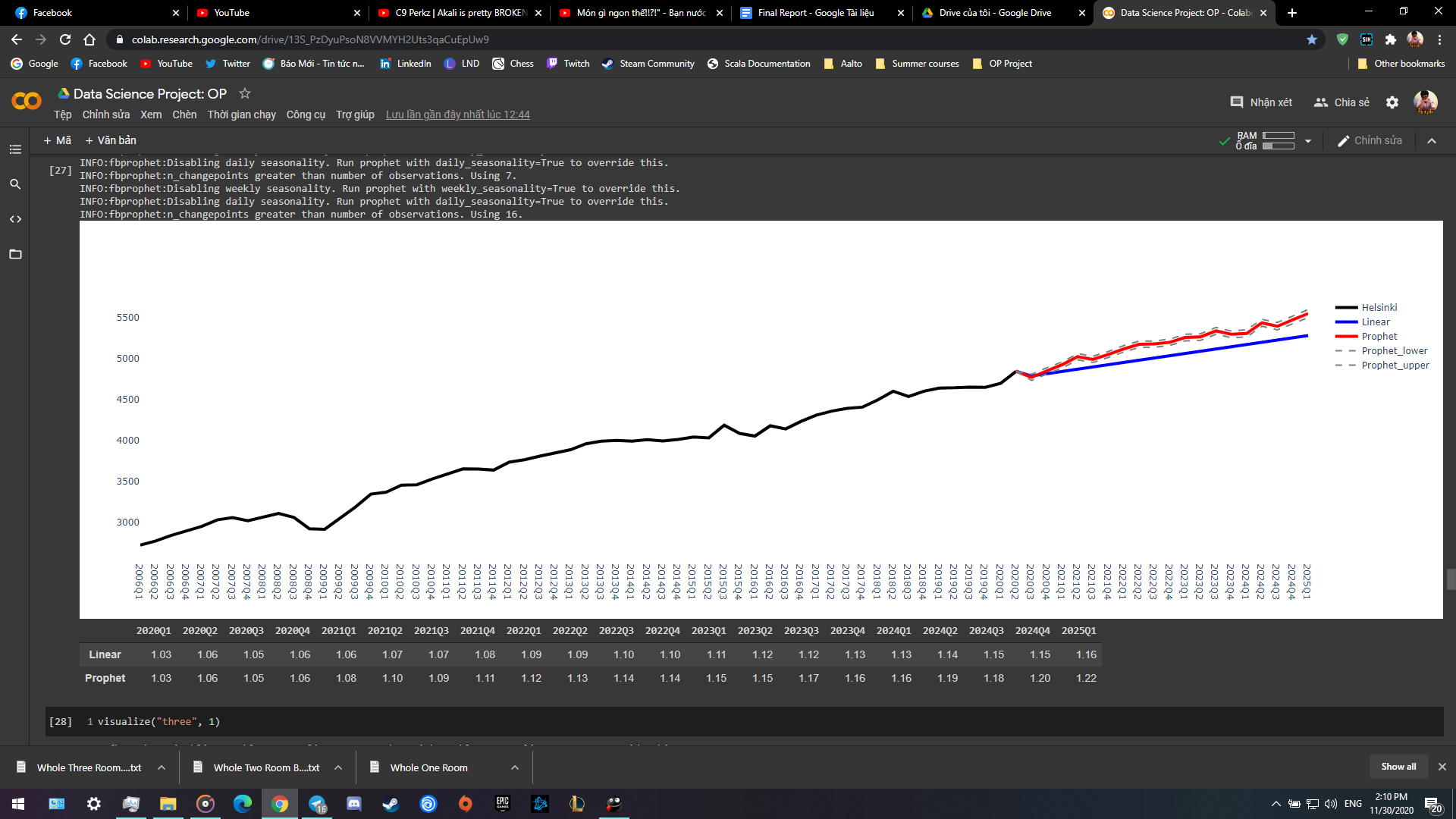
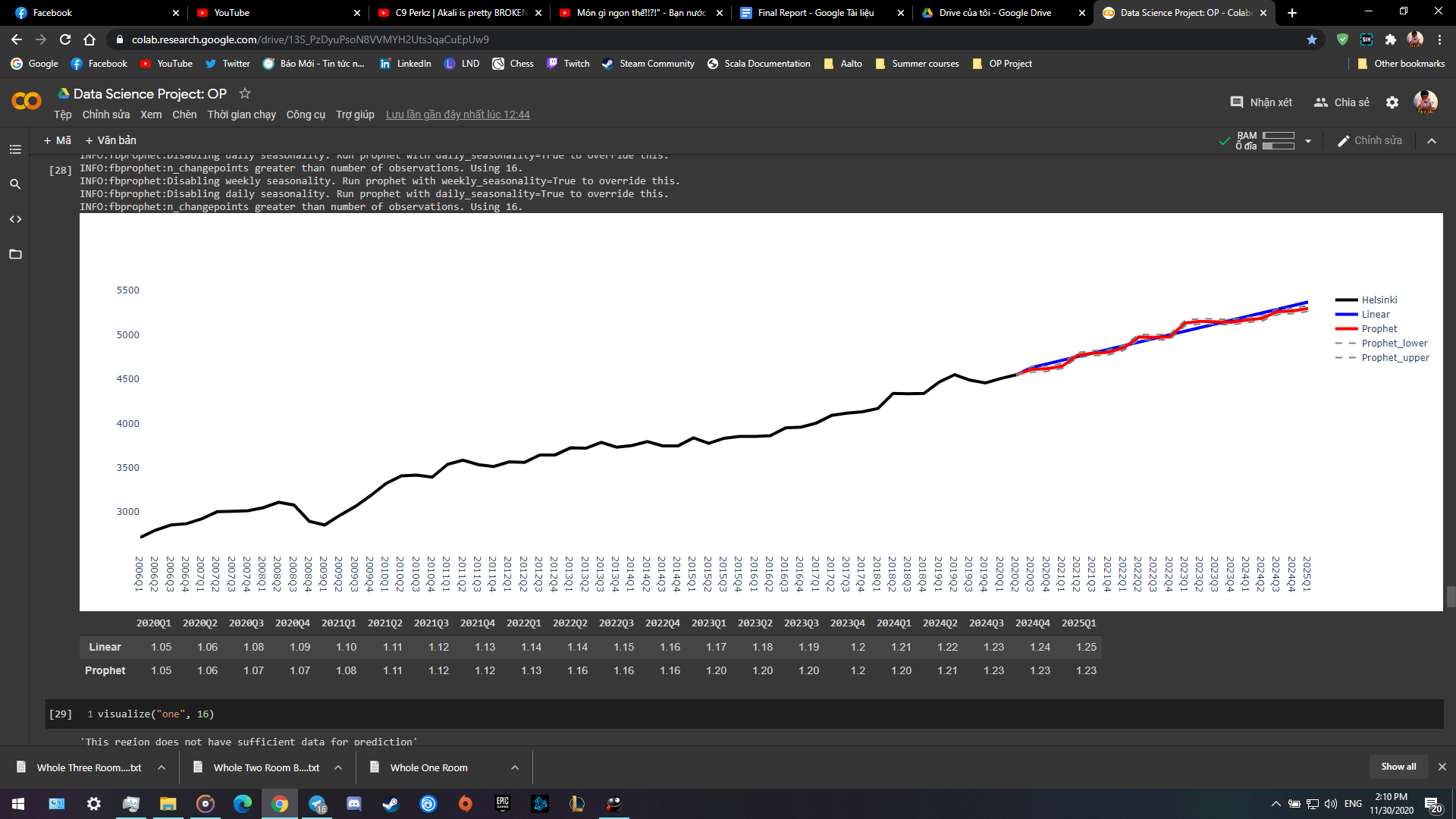
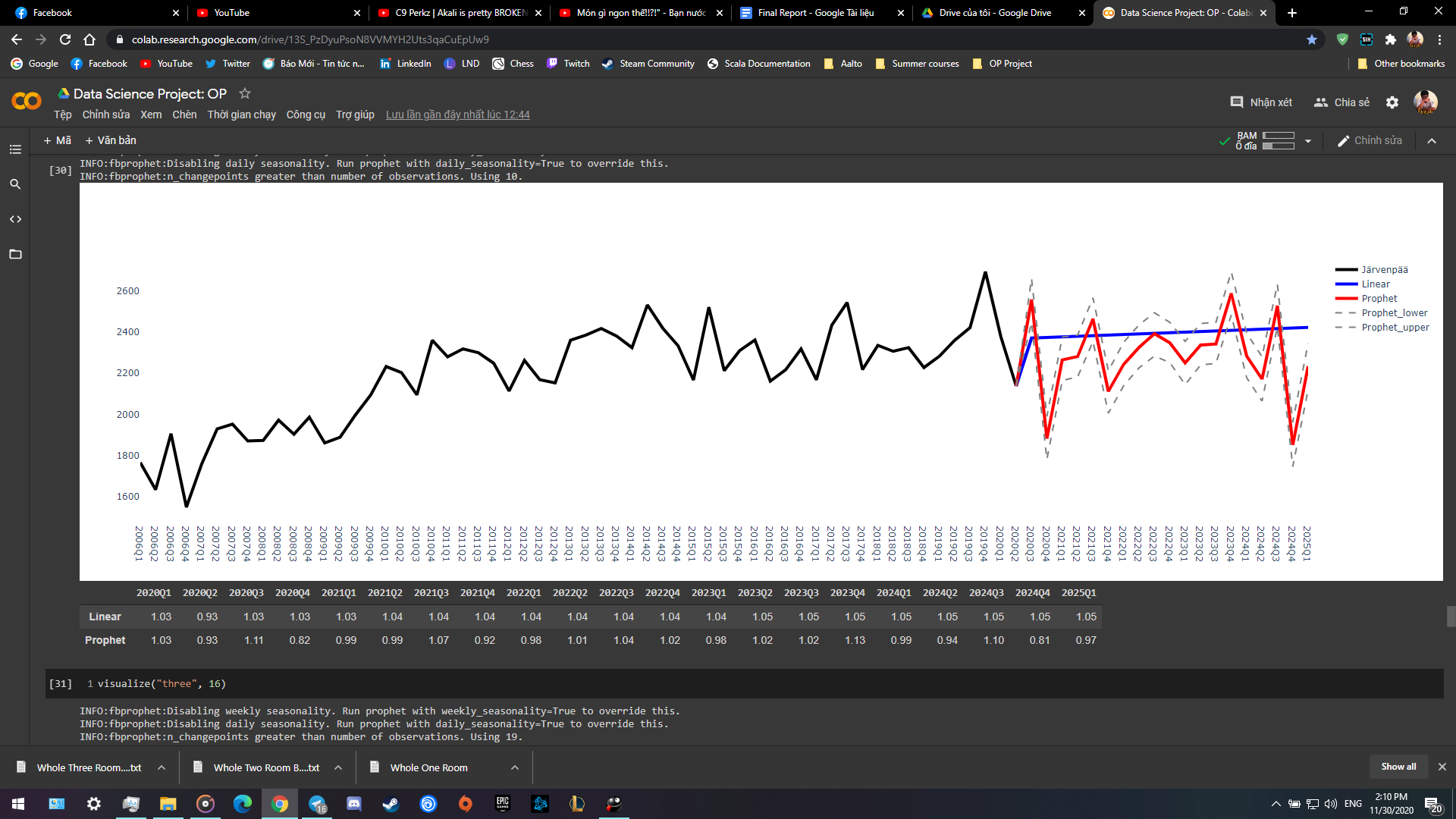
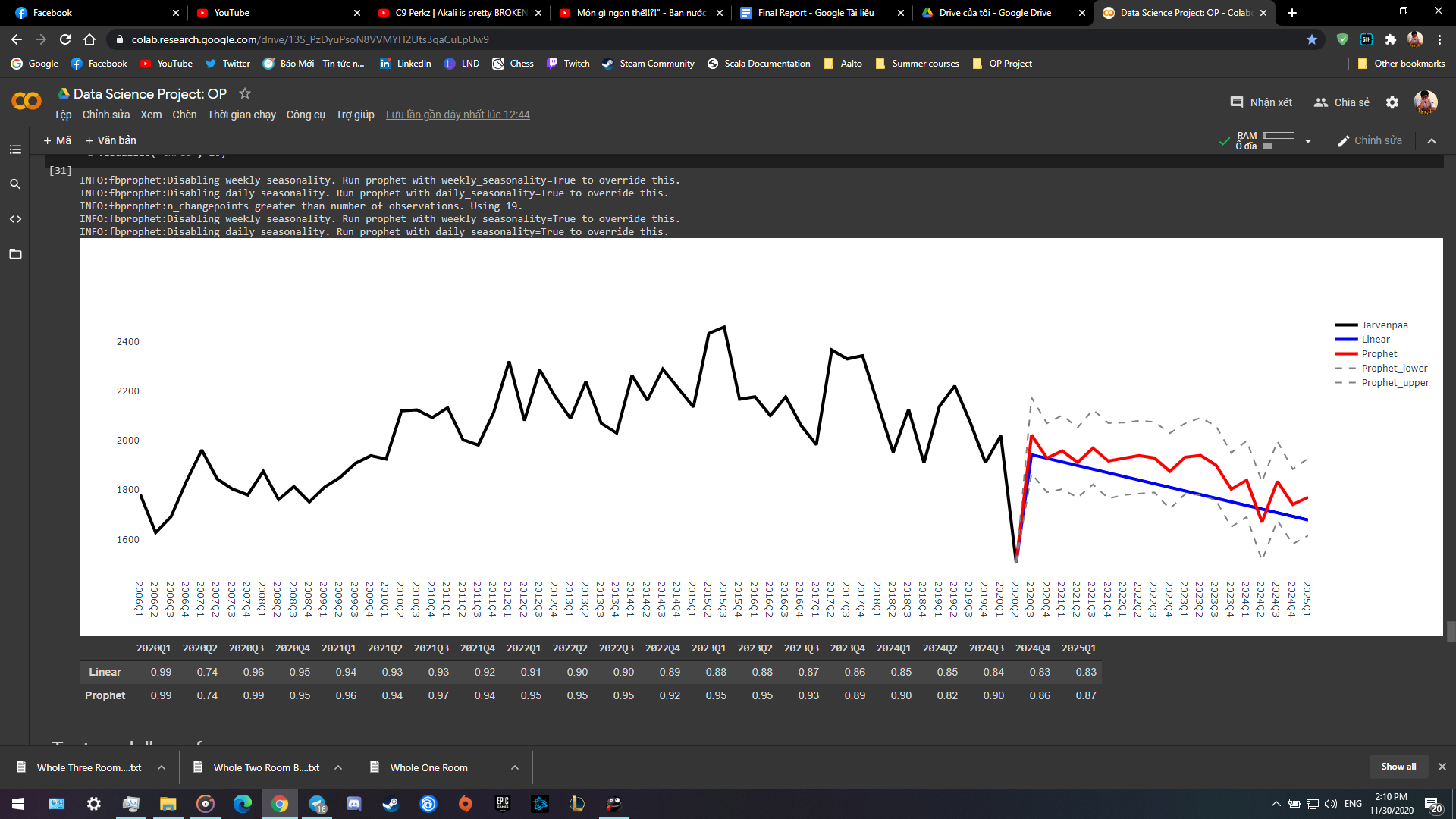


Fig: Prediction for One Room, Two Room and Three Room of Helsinki region

Fig: Prediction for Two Room and Three Room of Järvenpää region (“Outlier”)

### 5.2 SARIMA

The sarima model does not produce a single model that can be used to make predictions on all of the regions or housing types, and a new model with new parameters must be trained each time a different dataset is used. This prevents us from making a prediction for each region or apartment type. For the predictions, another grid search was implemented to find which parameters would minimize the value of the akaike information criterion (aic) provided that the orders are so that all the included lags are significant (P-value less than 0.05). A trade-off between minimizing the aic and using significant lags to choose the best model was one of the problems that took the most effort trying to solve. Often used to estimate the quality of a statistical model, aic measures the information lost in the prediction and thus minimizing its value decreases the amount of information lost.

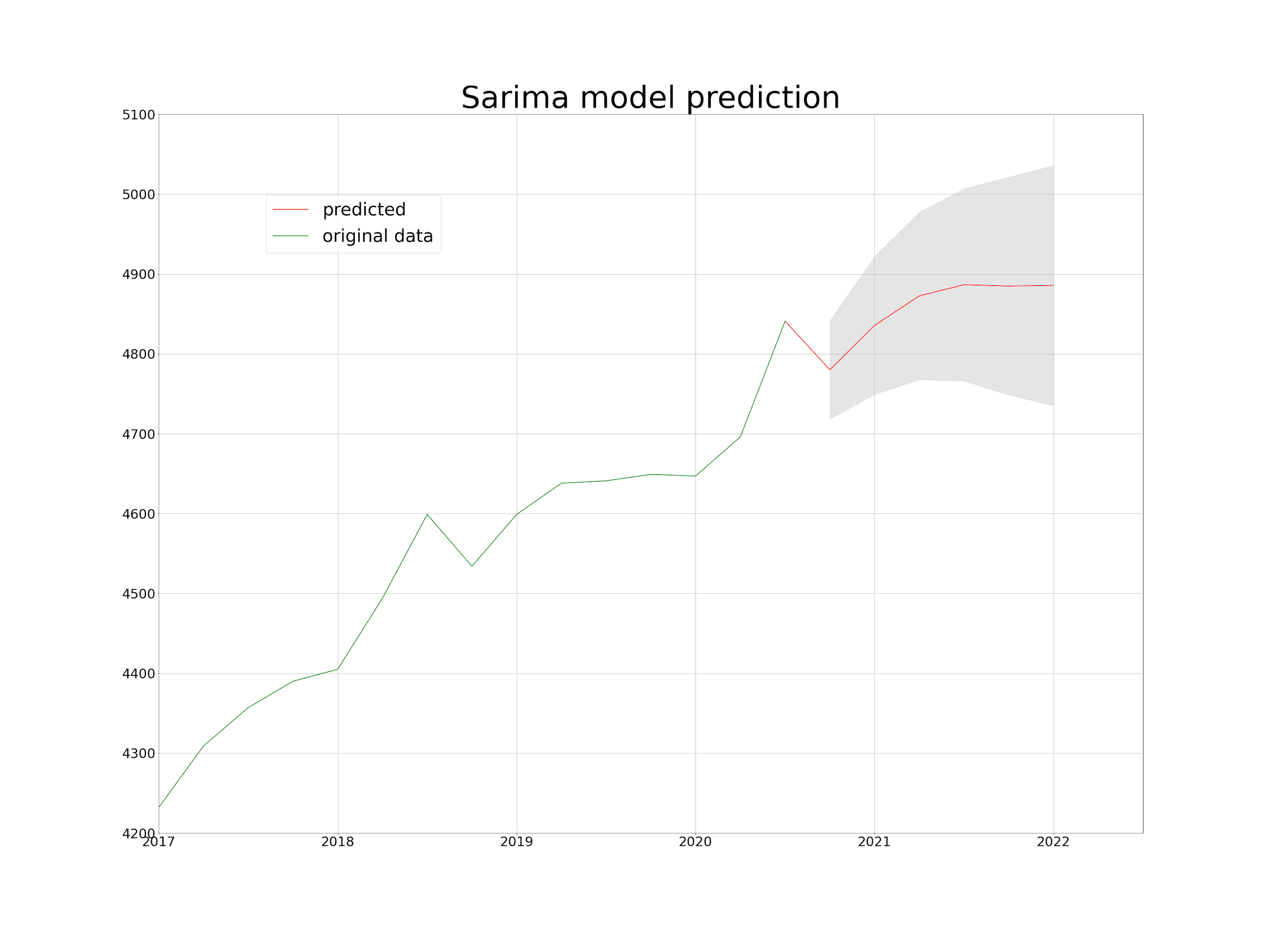


Fig. Sarima model prediction in Helsinki region with two-room flats

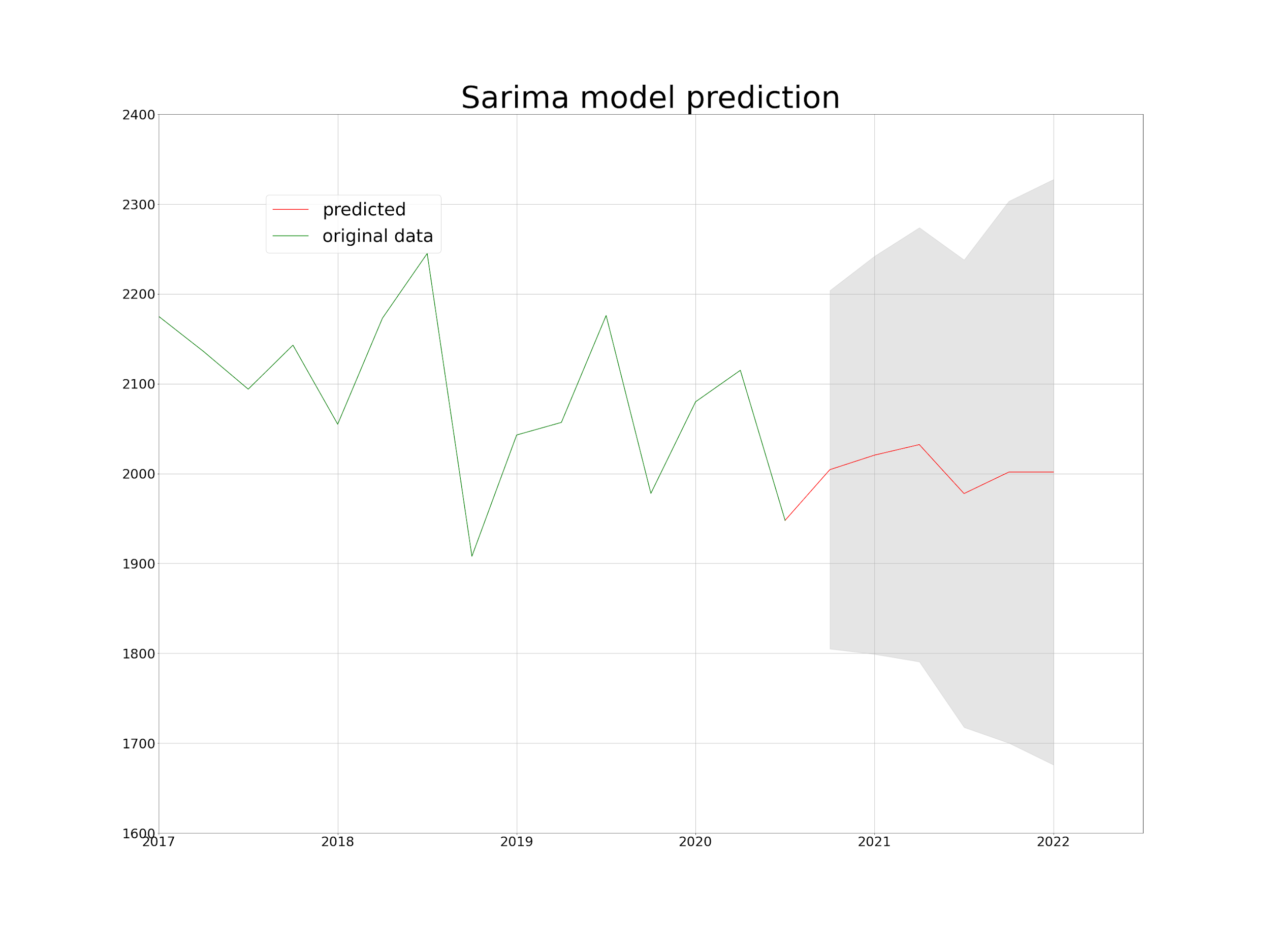


Fig. Sarima model prediction in Vantaa 1 region with three-room flats+

Above we have two prediction examples: first one is a prediction made in Helsinki region with two-room flats and the second a prediction in Vantaa 1 region with three-room and bigger flats. The first prediction is made with data from 2016 onwards whereas the second is generated with all of the data starting in 2006. This is where we see the difference in predicting using independent datasets, the approach can vary greatly depending on the data.

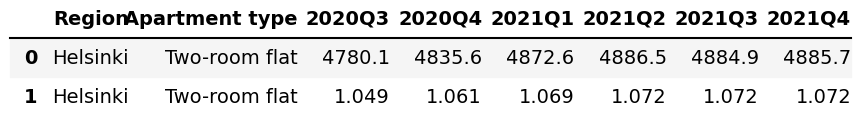


Table. Helsinki two-room flats predictions as a table with price indexes

In the table above, we have also included the physical values of the prediction and the corresponding price indexes.

# 6. Ethical Issues

When conducting any research, the primary ethical issues include: a) Informed consent, b) Beneficence, c) Respect for anonymity and confidentiality, d) Respect for privacy, and e) Integrity. This section examines the possible ethical issues relating to our project that predicted house prices in Finland.

As the data was gathered from the official website of Statistics Finland that produces the vast majority of Finnish official statistics, it is assumed that the data was collected, examined, processed, and presented following all ethical guidelines. Some examples of the ethical guidelines that we assume have been followed include the house owners' consent to gather data, ensuring the anonymity of the individuals providing data, and not falsifying data points to fit the original hypothesis better, to name a few. However, Statistic Finland does not specify how they have gathered the data, so we need to rely on their reputation as a reliable organization.

A point that arose when analyzing the original data and the sparsity of some regions was data masking. Data masking is the act of hiding the original data with modified value to preserve the privacy of data considered personally identifiable, sensitive, or personal. Hereby, we hypothesized that some regions that contained NaN values in their columns might have been masked if too few homes were sold in that region to preserve those deals' privacy. For this reason, the data that was used to formulate the predictions might not be 100% accurate, leading to some inaccuracies in the predictions. Additionally, the predictions might have some small inaccuracies as results of data imputations.

# 7. Conclusion and Future Prospects

This project was conducted as part of the course CS-C3250 - Data Science Project 2020, and its objective was to predict housing prices in Finland until the end of 2021. The predictions were provided on a quarter-to-quarter basis given a region and a house type. In order to predict the house prices, three models were implemented: linear regression, Facebook prophet, and the SARIMA model. For short term predictions that study house prices within the next year, it was concluded that the linear regression provides accurate enough estimations. However, for longer forecasts, the Facebook Prophet and SARIMA were seen as the optimal choices. These two models provided similar results. However, SARIMA was computationally more expensive, and therefore, if predicting house prices for multiple regions simultaneously, the Prophet is seen as the preferable choice. The specific results for each house-type gained by utilizing the linear model and the Facebook prophet can be found from the following Google Drive: <https://drive.google.com/drive/folders/1K9Ue3oSryTES8MSYCcl2ceZhtrnKwddM?usp=sharing>

As for future prospects, the project could be extended to include an interactive visualization or an application. This way it could be easier for a user to study the results. Additionally, after obtaining real values for the predicted data points, the estimations and real values could be compared to see how accurately our models predicted the house prices in Finland.

# 8. Roles of the team members

Anna: Data cleaning and imputations, research of accuracy measurements, index tables

Sami: Data gathering, pre-processing, and implementation of the sarima model

Mohamed: Initial gathering of data & pre-processing, EDA of each DB.

Khoa: Data pre-processing, implement and study the performance of Linear Regression and Prophet