Group 7 Analysis of German Apartments

Zach Gordon, Jason Hu, Tyler Dennis

12/1/2021

```
## Loading required package: data.table

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

Introduction

This report covers our findings on German rent data throughout the span of February to October 2020. It goes into detail about the small effect COVID-19 had on prices, as well as the creation of models to predict rent rates based on apartment attributes. It also echos a classification method that aims to determine whether the apartment is located in East or West Germany.

The dataset we chose consisted of 268,850 observations of apartment attributes throughout Germany which were scraped at four separate times: February, May, September, and October of 2020. The attributes consisted of the total monthly rent of the apartment, as well as accommodations such as having a kitchen, balcony, etc. It also echod apartment qualities like number of rooms, which floor it is located on, and amount of living space.

The first research question we had came from the date at which the data was scraped, which was pre-COVID/during the beginning of the COVID pandemic. We wanted to know if apartment rent in Germany was affected in any way by the outbreak of the virus, similar to how rent rates in the United States were affected. Our second question involved the prediction of rent prices based on attribute data. We knew that there was a strong relation between apartment attributes and rent price, but wanted to find the most significant variables, as well as know if that relationship was linear, or logistic, or another form. Our third question came from our curiosity about the lasting effects of World War II and the Soviet Division of Germany. Was there a significant difference between the development of living spaces in old Soviet Germany and not Soviet Germany? To answer this we wanted to see if German apartments were able to be correctly classified into East or West locations based on their attributes.

Data Overview

The data set consists of rental apartment listings throughout Germany. The creators of the data set scraped it from https://www.immobilienscout24.de/, which is similar to Zillow. We retrieved the data from Kaggle.

The original data set had 268,850 observations and 49 variables, which describe the apartments. Notable variables echo tota rent, living space, location, and when the apartment was constructed.

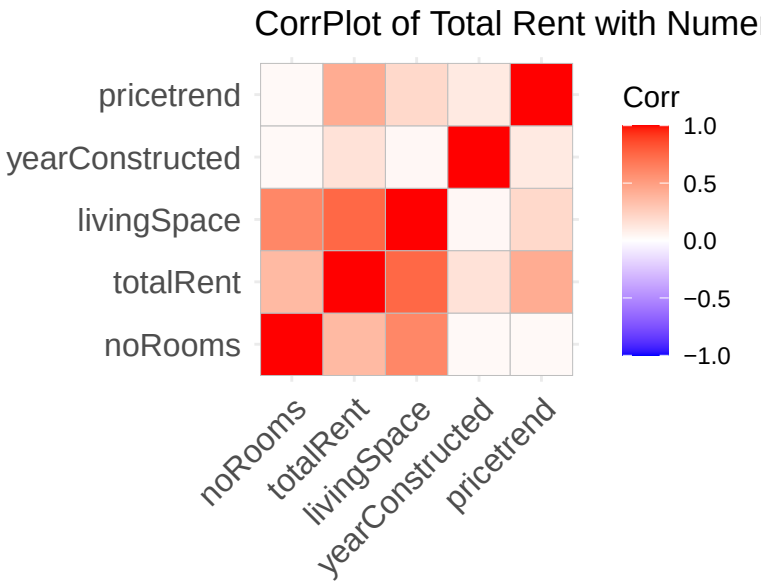
We found this data set interesting for a few reasons. The first was that the data was scraped at the beginning of the COVID-19 pandemic. This means the rents for the apartments could have been influenced by the pandemic. The second reason we found it interesting was the possibility of using the variables to build a rent prediction model. Variables like living space and number of rooms correlates with rent which means it could be possible to get a well fitting regression model to predict rent. The last reason we were interested in the data set was that it was from Germany. Until the early 1990's Germany was divided into two separate countries, East Germany and West Germany. East Germany was under the influence of the Soviet Union and West Germany was under the influence of NATO. The two countries had very different housing policies. As a result their housing stock looked quite different. This means we could identify if an apartment was located in East Germany or West Germany based on the variables in the data set.

The data contained a very high number of NA values. We used this to inform our variable selection. We looked at the percentage of values that were not NA for each variable and eliminated all variables with over 40 percent missing. We then used the na.omit() function to remove the remaining observations with missing values. This left us with 79,867 observations, which is approximately 35 percent of the original data.

To prepare the data for use in our models, we also eliminated variables that were uninterpretable. These variables echod "description", "facilities", and "streetName" which were long strings of German text. We checked variables that seemed irrelevant, like "houseNumber" for correlation with price, then eliminated them if they were not high enough. "firingType" proved to be too unwieldy a format to use, so we eliminated it. We decided to pivot all of our unordered categorical variables so that each value of a categorical variable now had a new indicator column. The column corresponding to the value that sample was would have a "1" as its value while other indicator columns would have a "0". This would allow us to more accurately represent unordered categorical variables in our models, as well as allow us to use them in PCA. | The data was then filtered to exclude enormous outlier variables, for example, an observation having a total rent above 50,000 while the second highest was 5 times less than i. The filters we applied were: Year Constructed after 1830, Total rent less than 5000 and greater than 100, Living Space under 300 square meters and greater than 5 square meters, and Number Of Rooms less than 40. | The last step of pre-processing was to create a new variable called "east_vs_west". This new variable was built by categorizing all Western states as "west", and all eastern states as "east". Berlin was dealt with by dividing the city based on zip codes. The Wikipedia page for German zip codes was consulted for this. We then subsetted the data and took 20,000 random observations from these 79,000. We made sure half of them were from East Germany and half were from West Germany.

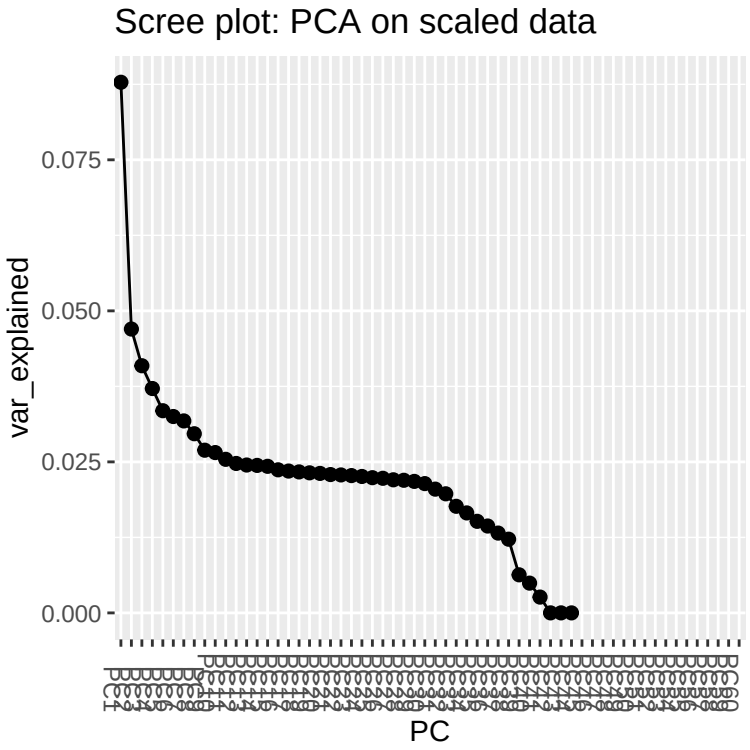
Exploratory Data Analysis

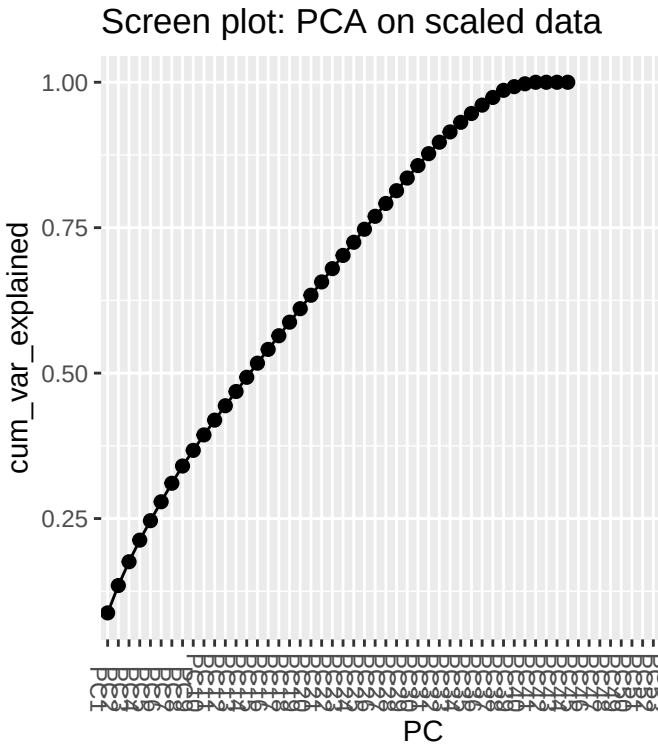
For the purpose of the rent prediction model, we explored which numerical predictors correlated with rent. We found that the number of rooms, the price trend, and the living space in square meters were all strongly correlated with the price of rent.



In terms of categorical variables, we wanted to see if the presence of certain features had statistically significant effects on rent prices. We found that the differences in the price of rent were not statistically significant for the presence of cellars, elevators, and kitchens. However, newly-constructed properties did have a statistically significant increase in the price of rent.

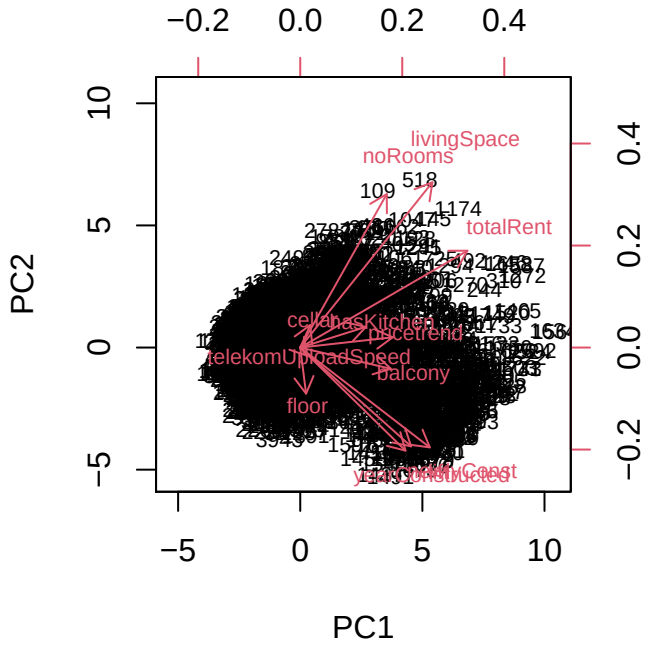
We did a scaled PCA analysis of our numerical data and plotted the variance explained and cumulative variance of each PC. Scaled analysis was better due to the difference in magnitudes of many variables.





The variance explained by the PCs do not look out of the ordinary (PC1 does not explain an unexpectedly high amount of variance), although the variance explained by the first few PCs are a little low, which could mean there is a smaller chance of finding significant findings.

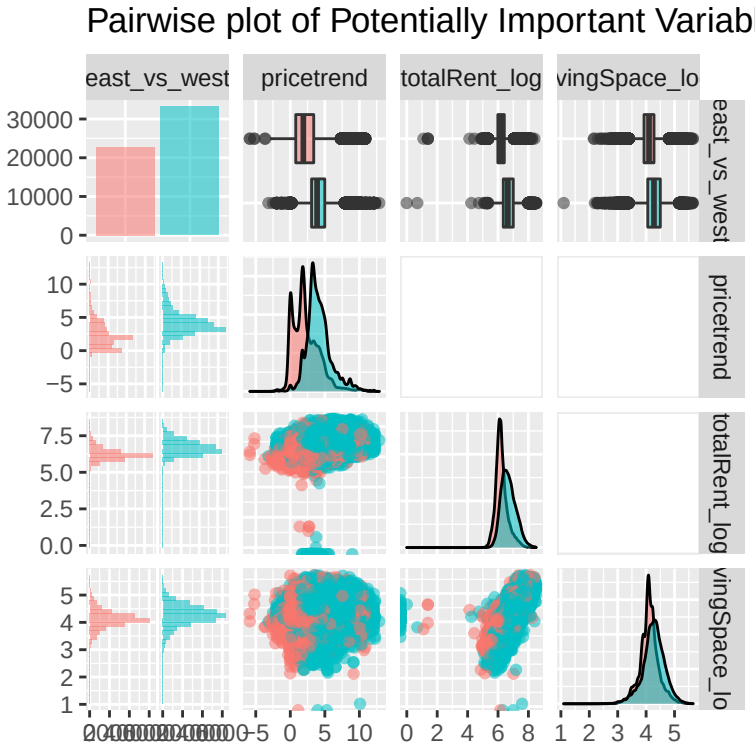
We then plotted a biplot of PC1 and PC2 along with factor loadings of variables:

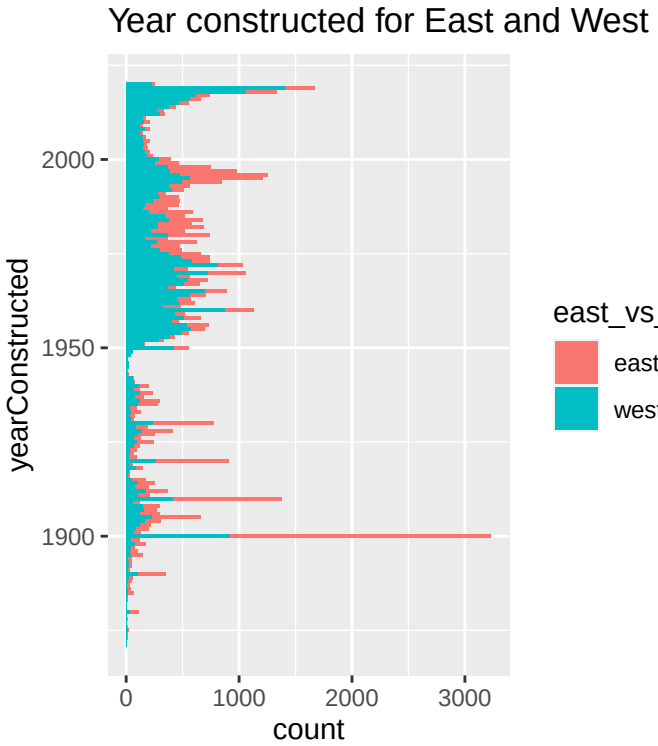


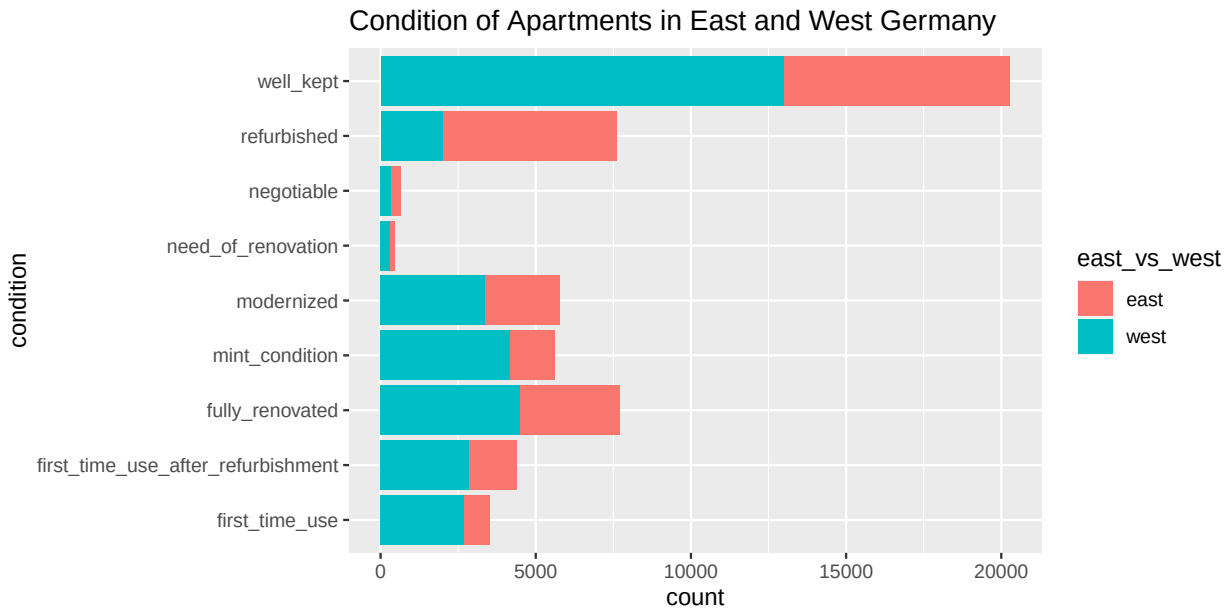
The black numbers represent samples and are plotted on the graph based on their PC1 and PC2 scores. The red arrows represent the first 13 variables, which are plotted based on their factor loadings (correlation) with PC1 and PC2. The arrows point in the direction that a point will be moved if that observation increases its value for that variable. The red arrows can also be interpreted as showing how much correlation variables have with each other. If two variables have red arrows pointing the same direction that means they are correlated the same with PC1 and PC2, which means that they are also likely correlated with each other. From this the biplot suggests that livingSpace, noRooms, pricetrend, and hasKitchen all seem to have a rather strong correlation with totalRent, the variable we will be predicting with our models.

The east-west classification question required extra exploratory analysis. We compared key variables for both the east and west labels. The first plot we made looked at was a pair plot which compared "pricetrend", "totalRent", and "livingSpace". "totalRent" and "livingSpace" were logged to make it easier to see the trends. "Pricetrend" showed the largest separation between east and west. totalRent and "livingSpace" had some separations but less than "pricetrend". One explanation for this is that more people want to live in the west than in the east which leads to increasing prices. As shown in other analyses there is a correlation between "livingSpace" and "totalRent". Both "totalRent" and "livingSpace" follow a normal distribution in the east and the west. The next two variables we checked were the year constructed and the condition of the apartment. Based on the graph of "yearConstructed" there appears to be clusters of years with a higher proportion of east labels than others. However, it is not the case that as the year increases the ratio of east to west increases or decreases. This means there does not appear to be a linear relationship between location and yearConstructed. The analysis of the condition of the apartments revealed that "well_kept" and "refurbished" had the highest ratio of east to west labels, with "refurbished" being the only label with a majority in the east. This is an indicator of separation and shows promise for the classification model.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```







Learning Methods

We chose a variety of methods to answer the research questions. Principal Component Analysis and ordinary least squares regression were used to examine the effect COVID had on rental prices. The rental price predictor was built using Ridge and LASSO regression. A support was the method used to classify apartments as being in either the east or the west. Both the rent predictor and the classifier took advantage of cross validation to protect against overfitting.

PCA

Principal Component Analysis is mainly used as a dimension-reductionality tool during exploratory data analysis. After reducing the dimensionality of the data, we can use the principal components of the data to explore relationships between variables as well as the clustering of observations. PCA reduces dimensionality by finding a line in the data that results in the greatest variation when all observations are projected onto that line. This line, called the first principal component, becomes a new "axis" that allows us to see the largest chunk of variability in data possible by looking at only one dimension. There is then a new "axis" created for every variable that is orthogonal to every other line, each sequenced line explaining less and less variability in the data. After the principal components are made, we can plot samples based on PC values and look at those graphs to find clusters in data. We can also look at factor loadings for each variable to each principal component, which shows the correlation between that variable and that principal component. If two variables are both positively correlated to the same principal component, then it is likely that those two variables are also correlated. For relating samples to variable factor loadings, if an observation increases in the value of a variable, and that variable has a positive factor loading on a principal component, then that observation has an increased PC value for the respective PC. For our research we mainly used PCA as a clustering analysis tool, specifically for our research on East and West divide, as well as on COVID affecting rent price.

OLS

Ordinary Least Squares regression is the simplest of linear regression models, which aims to find the one-dimensional line in data that will result in the minimum of the sum of all squared distances of observation

points from that line. It is used to model linear regression and predict a response variable by predictor variables, assuming they have a linear relationship. This method is used in our research on COVID affecting rent price, specifically, to see if there is a linear relationship between date and total rent, and if there is, to show the degree to which COVID affected rent.

Cross-Validation

K-Folds Cross-Validation is a method of cross-validation which shuffles the data and splits it into k folds. This is performed by reserving one fold as the testing set and using the other k-1 groups as training sets for fitting a model. This is then repeated k times, with each fold serving as the test data once and all the Cross-Validated Errors are averaged. As the value of k increases, bias decreases and variance increases. This method prevents overfitting and sample bias.

Ridge Regression

Ridge Regression is a method of linear regression that adds a penalty term that is equal to the square of the coefficient of each predictor. There is also a coefficient added to the penalty term that penalizes large predictor coefficients. If the penalty term is zero, then the method as OLS. As we increase the value of the penalty term, it causes the value of the coefficient to trend towards zero. This leads to lower variance and low training bias.

LASSO Regression

LASSO Regression, short for Least Absolute Shrinkage and Selection Operator Regression, is a linear regression model that, in a similar fashion to Ridge Regression, adds a penalty term and a regularization . It adds a penalty term to the cost function. This term is the sum of the absolute value of the coefficients. As the value of coefficients increases from 0 this term increases, causing the model to decrease the value of coefficients in order to reduce loss. As opposed to Ridge Regression, which lowers the value of coefficients but won't reduce dimensionality, LASSO Regression tends to set coefficients equal to zero.

SVM

The new supervised learning method we chose was a support vector machine or SVM. SVM is a non-stochastic learning method that uses the shape of the data to create a binary classification. It uses a hyperplane to divide the data and create a decision boundary. It chooses the hyperplane by maximizing the minimum distance of the observations to the hyperplane. The observations with the minimum distance to the hyperplane are referred to as support vectors. If the data is linearly separable then the optimization problem is solvable as is and a decision boundary is found that perfectly divides the data. If the data is not linearly separable then slack variables are added to some of the variables in order to make the problem solvable. This is called soft margin and it allows for misclassifications in the data. Because the data is almost never linearly separable, including our data, soft margin is frequently implemented. There is also a method called the "kernal trick" which adds dimensionality to make the data linearly separable, or at least more linearly separable. The "kernel trick" brings the risk of overfitting if too complex. Under soft margin SVM uses a hyper parameter called degree of tolerance or C. The point of this parameter is to penalize the model for misclassifications. The higher the parameter the higher the penalty for misclassification. Too high of a parameter can lead to overfitting. We use SVM to classify apartments as being either in East Germany or West Germany. Our model uses a soft max approach and tests multiple C values to find the optimal parameter. We also implement a 10-fold cross validation to check for overfitting.

Results and Discussion

Effects of the Covid-19 Pandemic on Rent Pices

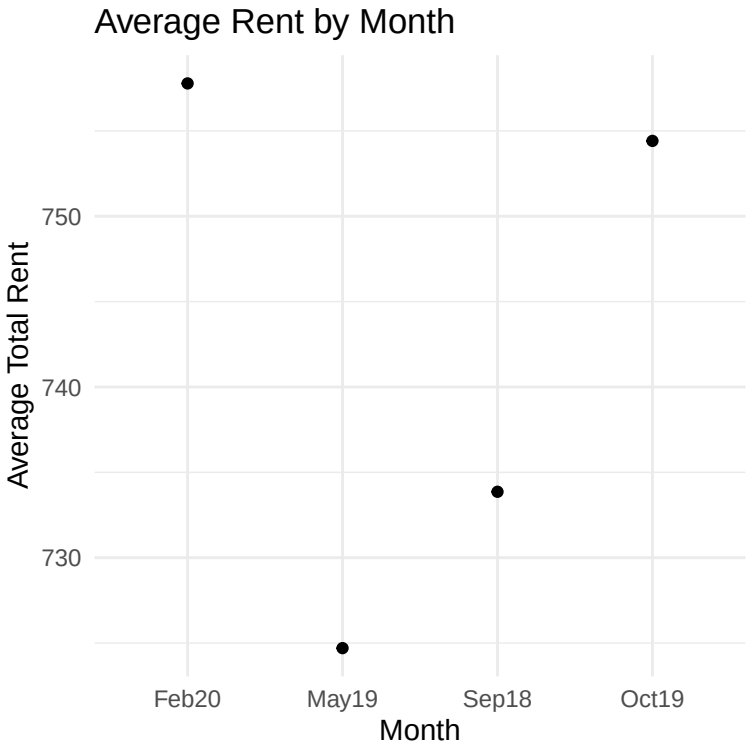
With our PCA described in the Exploratory Data Analysis section, we plotted pairwise plots of observations based on their PC1, PC2, and PC3 scores. We then color-coded observations by the month they were scraped in to see if any apparent clusters appeared.

From this, there seems to be no apparent clusters formed when coloring by the four different dates. We then decided to color-code by months before the COVID pandemic and months after to see if there was any clustering.

It seems from this too that there are not any apparent clusters.

We then moved on to OLS regression to see if we could find a linear trend in total rent price based on date.

The output of the OLS regression shows that rent was not very significantly affected by changes in date, although some dates did change more than others. This also implied that a linear model was probably not the ideal method for looking at price changes. Although this regression did show us that there was a trend in the data, with rent going down in May, but then increasing again to be even higher than pre-COVID in October. From this we decided to look at a visual representation of average prices during specific months.



From this we clearly see a trend in the short time period between February and October, with it seemingly looking quadratic. The average rent price dropped 1.62% in May, then increased by 1.09% in September, then increased by 2.82% in October.

So based on the data, rent price dropped at the beginning of COVID, then increased and spiked up in October as COVID went on. There are many ways to interpret this. One way is to infer that COVID did play a hand in these price changes. At the beginning of the pandemic people could have been less likely to be out looking for apartments, so rent price would drop, but lower interest rates and low prices could have caused a surplus of buying after the initial drop in price. This would be parallel to what happened in the United States, where lower interest rates on living spaces caused a large surplus of houses sold, causing price

increases during COVID. However, it is also possible that these price changes are not caused by COVID. For example, landlords could want to rent out their apartments during the summer, a time when many people are looking for a new place, so they would lower their prices. Then, as more people rented apartments and/or the sale season ended, prices went back up.

Rent Prediction Models

To make rent prediction models, we used both Ridge and LASSO Regression on the training sets. To explore the effects of the relative sizes of training sets and testing sets, we divided the n = 20,000 observations into five different sets of training and testing data divided as follows: 50-50, 60-40, 70-30, 80-20, and 90-10. The cv.glmnet function was used to find the value of lambda that minimized the Mean Cross-Validated Error for each training set and then the glmnet function was used to fit the data.The optimal values of were found to be 0.4021 and 0.3664 for Ridge and LASSO Regression, respectively. The optimization for both is visualized below.

To measure the predictive power of the ten models, we used the Normalized Root of the Mean Square Error (NRMSE), which was calculated by taking the root of the Norrmalized SSE, in which each observations squared error is divided by the value of the observation. Tables showing the error of each ratio are provided below for Ridge and LASSO Regression, respectively.

Data Split <chr>	Training NRMSE <chr>	Testing NRMSE <chr>	Training RMSE <chr>	Testing RMSE <chr>
50-50	0.171079084678195	0.273705741248402	131.898814627273	208.85152277188
60-40	0.173816926851353	0.271536965003428	133.825499130838	207.089796227934
70-30	0.177583845061154	0.268109231124805	135.902321364251	206.727095025308
80-20	0.182325998309759	0.253420718862364	139.597080252828	195.767526861938
90-10	0.183729979729441	0.256825934062786	140.569314785531	201.450739027034

Figure 1: Ridge

Data Split <chr>	Training NRMSE <chr>	Testing NRMSE <chr>	Training RMSE <chr>	Testing RMSE <chr>
50-50	0.180390116868234	0.269313065485674	139.07744848034	205.499685803081
60-40	0.182209905767757	0.268192108951327	140.287439363177	204.538815523584
70-30	0.181686605428695	0.265310290354108	139.042103914618	204.568956373213
80-20	0.185420303039726	0.250499441242169	141.966220746891	193.510839652028
90-10	0.183729979729441	0.256825934062786	140.569314785531	201.450739027034

Figure 2: LASSO

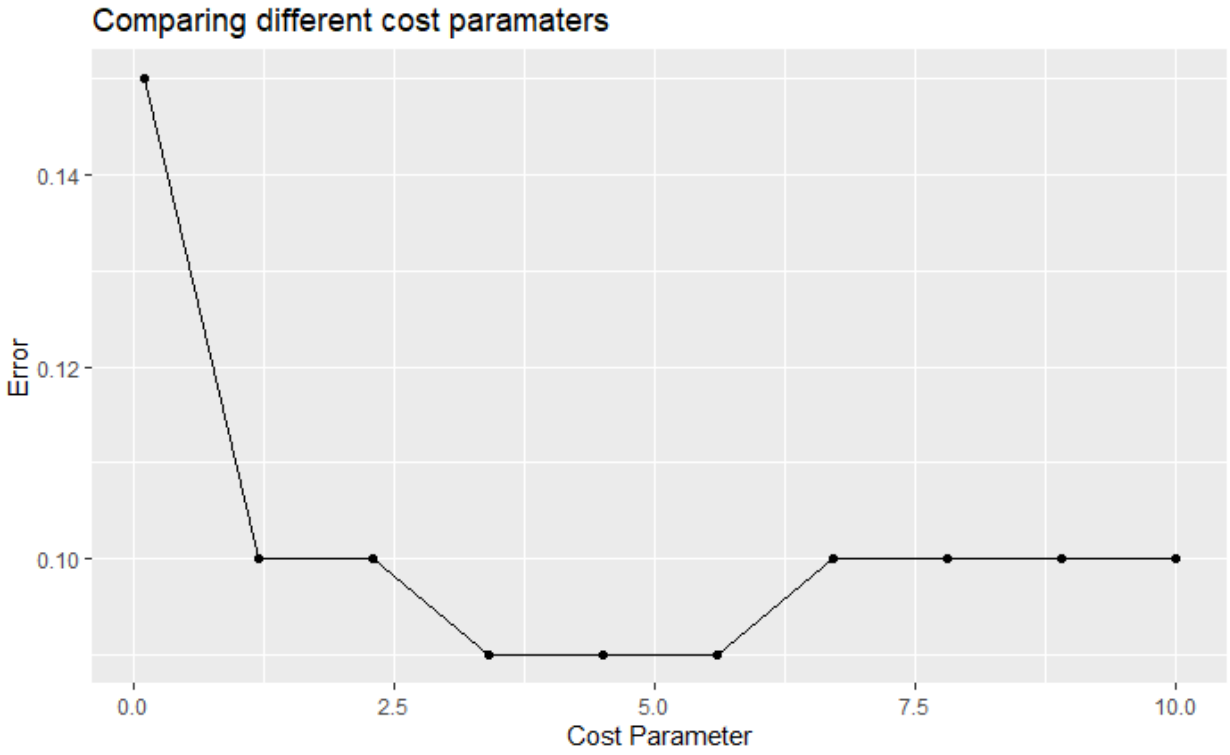
As can be seen, the 80-20 training and testing set ratio provided the lowest NRMSE on the testing sets for both regression methods. Unfortunately, the best model for Ridge Regression had an NRMSE of 25.34% and the best model for LASSO Regression had one of 25.05%. A rent prediction model with 25% error indicate a weak predictive power.

East or West German Classification

The results of the east-west classifier were relatively strong given the nature of the problem and the quality of the data. The first step in the process was to find an optimal C parameter; this was done by training models using different parameters and choosing the parameter that produced the lowest error. While doing this we used a training sample size of 18,000 and a testing sample size of 2,000. We tested 10 C values in a range from .0001 to 10. The chart shows the change in error for each parameter. The highest C was at point .0001 with an error of 15 percent. The lowest was between 3.4 and 5.6 with an error of 9 percent. Based on this graph we chose a C parameter of 4.75, which is the middle point of the lowest Cs.

The next step in the process was to run the 10-fold cross validation. The point of this was to use as much of the data for training as possible without overfitting. The training folds contained 18,000 data points and the

testing folds contained 2,000 data points. The maximum error produced by the cross-validation was 16.25 percent, the minimum error was 13.4 percent, and the mean error was 14.42 percent. This is noticeably higher than when finding the optimal C value which is intriguing and may be due to being a different split.



Conclusion

In Regards to COVID affecting rent prices, we came to the conclusion that the effect of COVID on rent price in Germany in 2020 was negligible. The largest change in price was only 2.82% in between September and October. This is also assuming that COVID was the cause of these price changes, it is possible that other factors are in play that changed the price, such as summertime being a hotspot for apartment sales. For seeing if COVID affected rent price, it would be beneficial if we had another set of data from years before COVID that we could compare results from this data to, to see if the trends are the same or different, implying whether or not COVID really was a factor in this trend.

Using Ridge and LASSO Regression methods, we were unable to make a strong rent prediction model. The ratio between training and testing data that minimized the NRMSE on the testing set was the 80-20 split for both models. A future attempt could be made with other regression methods, both linear and nonlinear. Additional features from another dataset regarding the localities in which the rental properties are located could supplement this data to make a stronger model.

Given that Germany has been united for three decades, the housing stock in the two regions has likely become more similar. This makes the classification harder and the data set we have is not as well suited to the problem. Given this fact the error rates our model found are actually pretty good. Future research should build on this by building a classification model on a more complete data set with a wider range of variables. Researchers could also try different models on larger data sets if they have more computing capacity.