Influencing factors of the rents in Stuttgart

Applied Data Science Capstone Project on Coursera

Dennis Netzer

February 12, 2019

Contents

1 Project Report

1.1 Introduction

This is the capstone project of the Applied Data Science Course from Coursera. In this project we had to think about an own problem, that could be solved using data from free internet sources as well as from the foursquare database. Since I live in Stuttgart and the rents are pretty high in some areas in Stuttgart, I thought about looking into some characteristics that influence the rent.
For the sake of training, what this project is supposed to be, I will also build a regression model, that can predict the average rent in an area, given the specified features.
Stakeholders: This could also be interesting for city planners. The politicians don't want the rents to soar in some areas, so nearly anybody can afford them. With this project, the city planners can see what makes the rents rise and can maybe take action if necessary.

1.2 Objective

The goal of this project is to figure out which features have a big impact on the rent and which are just incidental. Therefore, we will look for online sources with data regarding the population and other relevant factors in the different areas. The foursquare database will be used to get information about local venues. Statistical methods will then be used to find dependencies between the extracted features and the rent.
The final objective is a set of features, that mainly influence the rent.
Recommendation: - explore other cities to see if the correlations are the same, take more features into consideration, explore older data and see the trends

1.3 Data

To find out what characteristics of an borough make rents rise, we have to get as much data about these boroughs. This can include data about the residents, the flats, the infrastructure, opportunities for different activities, crime rates and other features related to the quality of living. Of course we also need the current rents for each borough.

The current rents can be downloaded from following website: https://www.wohnungsboerse.net/mietspiegel-Stuttgart/972
To use the foursquare database, we also need the longitudes and latitudes for the boroughs. We can extract them from following website: https://www.suche-postleitzahl.org/stuttgart-plz-70173-70629.608e
I searched a while online to get data about the boroughs. Following website offers free statistics about the different boroughs in Stuttgart: https://statistik.stuttgart.de/statistiken/statistikatlas/atlas/atlas.html?indikator=i0&select=00
Further, we can use the foursquare database to get data about the venues in the boroughs. Having great opportunities for leisure activities for example, can also influence the rent. The venues will be clustered into some categories, that can be used for analysis.

1.4 Methodology

The methodology section will cover following points:

1. Loading, extracting and preparing the data
2. Exploratory Data Analysis
3. Modeling

1.4.1 Loading, extracting and preparing the data

Rent data

The website mentioned in the Data section to get the current rents, provides the rents in a pdf file. Since there are only 23 boroughs in Stuttgart, it was easier to prepare an excel file to load the data into a data frame than scrap the pdf document. So all the boroughs and their corresponding average rents were manually written to an excel file.

Coordinates

The website mentioned in the Data section has the postcodes together with the longitude and latitude of each borough in form of a table on the website. So with the beautiful soup library, I could extract the necessary data from the website and load it into a data frame.

Statistics about the Boroughs

The website mentioned in the Data section provides an excel file with a bunch of statistics about the borough. The statistics that could have the biggest impact on the rent were prepared in another excel file that could be read into a data frame.

Foursquare data

Having the respective coordinates of each borough, I could make some calls to the foursquare API using the developer account. I decided to get the 150 most popular venues in a radius of three kilometers around the given coordinate for each borough. I extracted the short name of the category for each venue out of the json file and saved it into a list.
Now I could explore the different short names that existed in the foursquare database and cluster them a bit. For the best features that could be extracted were: Number of restaurants, number of leisure activities, number of transport possibilities, number of sport activities and number of shopping possibilities.
Therefore, I created lists with the most common short names for each of these categories. Now I could count the number of venues, that fell into one of the categories. This data, together with all previous gathered data was then added to one data frame.

1.4.2 Exploratory Data Analysis

First of all, I had a look at the different features that I had gathered. I created histograms for each feature with the matplotlib library too see how they were distributed just to get a better understanding of them.
After that I created linear regression plot for each feature using the seaborn library. With these regression plot I could see how each feature was related to the dependent variable rent. Having analyzed them visually, I also calculated the Pearson Correlation and the p-value for each feature.
Now I could point out the features that had the biggest influence on the rent. A Pearson Correlation coefficient close to one indicates a strong positive correlation and a coefficient close to -1 indicates a strong negative correlation. The p-value indicates whether this correlation is statistically significant or not. A value lower than 0.005 indicates significance.

Following correlations were found:

feature	**Pearson Correlation Coefficient**	**p-value**
amount of married residents (%)	-0.723	$9.701 * 10^{-5}$
residents per household	-0.706	$1.67 * 10^{-4}$
number of restaurants	0.711	$1.697 * 10^{-5}$
number of leisure	0.738	$5.747 * 10^{-5}$

1.4.3 Modeling

With the defined relevant features I could now build a model, that can predict the average rents for other boroughs given these features.

The model that is most adequate for this situation is Multiple Linear Regression. We have four feature that all have a linear correlation to the depended variable. To build the model I used the sklearn library. For building and later evaluating the model, I also used the train-test-split function of the sklearn library. This function automatically splits the data into a training and a testing set. The training set was then used for fitting the model. With the remaining testing values for the features, we could predict the rents.

These predicted rents were then measured against the actual values. For the evaluation, I used R-squared and the mean-squared-error. The calculation of R-squared results in a value between 0 and 1. Where 0 indicates that the used variables are very bad in predicting the dependent variable and 1 means they explain the dependent variable perfectly.

The mean-squared-error measures the distance between the actual and the predicted points, takes the square root of the distances and then calculates the mean of these values. So it's a measure of how far our predictions are away from reality.

Since there is so little data, the values for R-squared and the mean-squared-error highly depend on what data is used for testing and what is used for evaluating. Because of this, I also used cross validation. In cross validation, the data is split into so called folds. A fold is just a part of the data. If we split it into three folds, we will have three data sets, that have the same size. Now two of the folds will be used for training and one for testing in turn, until every fold was used for testing once. That way we can calculate R-squared mean. For this model the R-squared mean was 0.24.

1.5 Results and Discussion

As the table in section 1.4.2 shows, the analysis provided four features, that have a correlation to the rent. Let's interpret these insights with words.

Amount of married residents(%): The less married residents live in an area, the higher the rents will be.

Residents per household: The less residents live together in one household in an area, the higher the rents will be.

Number of restaurants: The more restaurants are in an area, the higher the rents will be.

Number of leisure: The more possibilities for leisure activities there are in an area, the higher the rents will be.

From this we can induce that, in areas where people live in smaller flats with less people in a household, the rents are higher. In these areas there are more restaurants and more possibilities for leisure activities like bars or museums. We can also say, that in areas with bigger households and a lot of married residents the rents per square meter are lower. This means that families are more likely to move away from the well-frequented places.

The cross-validation showed a result of R-squared mean of 0.24 that indicates a bad model. If we really want a reliable model, we need more data. We could use all the data that we have to train the model and then apply it to other cities to see if it results in good accuracy. If we want to develop the model further, we could also take previous years into account or include data from other cities for training purposes as well.

1.6 Conclusion

This project showed how data about the boroughs of a city can be used to make statements about the average rent in this borough. It showed what characteristics of a borough have influence on the rent.

For me personally it was a good practice. I learned a lot in how we can use data to approach a problem that I have defined myself. I learned how to use stackoverflow and other online sources to solve problems that made you stuck. Even if I didn't include much data and the insights weren't that great, the used methodology could also be used for data sets much bigger, so it fulfilled its purpose and I am ready for more challenges.