Analysis and modelling of rent data

Nathan van Rensburg
38287420

Table of Contents

1. Data cleanliness

The dataset appeared to be clean and well-prepared, with no major inconsistencies or missing values. This indicates that the data was likely collected and processed with care, which could lead to reliable results and accurate conclusions. The cleanliness of the dataset suggests that it is suitable for future statistical analysis with confidence. Overall, the dataset's cleanliness does not seem to be a significant obstacle and is a positive indication for potential future analysis.

2. Exploratory analyses using graphical techniques

Boxplots showing how city and price are related were drawn. These boxplots provide a nice comparison showing how prices in the three cities compare to each other (see Figure 1).

Figure 1: Boxplots showing price per city

According to the boxplots, there is no significant difference in the median price in the different cities.

For the average buyer, there isn't going to be a significant price difference between renting a place in Stellenbosch, Cape Town, or Johannesburg. The plots show that the median rental price is around R6,000 per month.

It is worth noting, however, that the graphs shown here don't consider how big each apartment is or how modern the kitchen is. They merely show the price distribution for each city and how they compare to the other two cities.

A scatterplot was used to visualize the relationship between apartment size and price (see Figure 2).

According to the plot, a relationship between apartment size and price seems to exist. Generally, the bigger the apartment, the higher the price.

There seem to be three rough apartment sizes, namely 60 square meters, 80 square meters, and 100 square meters. They rent for around R5,000, R9,000, and R13,000, respectively.
Interestingly, Cape Town seemingly has quite a few lower-priced large apartments.

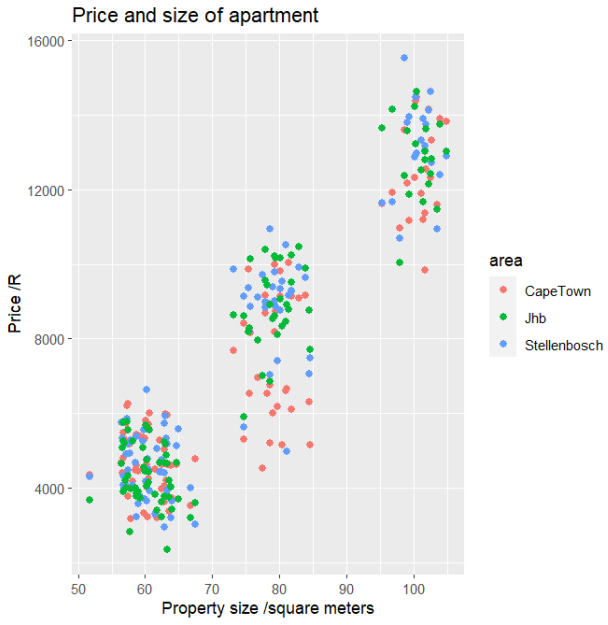

Figure 2: Scatterplot showing how price and size of apartment are related.

Another factor that could determine the price is the state of the kitchen, i.e. is the kitchen old or modern?

Two boxplots were drawn to show the price distributions for old and modern kitchens (see Figure 3).

Unsurprisingly, apartments with a modern kitchen have a higher median price than apartments with an old kitchen. The median "modern kitchen" apartment price is around R9,000, whereas the median "old kitchen" apartment has a R5,000 price tag.

Figure 3: Boxplots showing price by kitchen

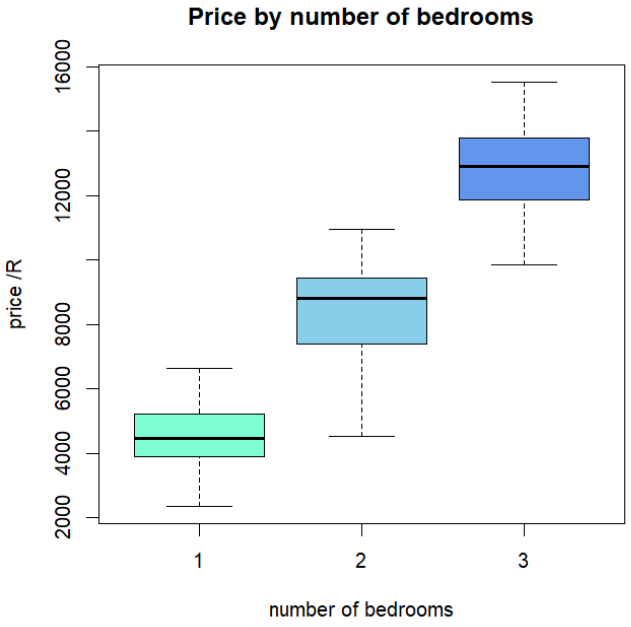

Figure 4: Boxplots showing price by number of bedrooms

The price of apartments also showed considerable variation when taking into account the number of bedrooms in each apartment.

Boxplots of the prices of the various numbers of bedrooms show that the higher the number of bedrooms, the higher the monthly rental price is, generally (see Figure 4)
The median price for a one-bedroom apartment seems to be around R4,500, for a two-bedroom it's around R9,000, and for a three-bedroom, it seems to be around R13,000.

The number of bathrooms also significantly affected the price of an apartment. More bathrooms generally indicated a higher price (see the boxplot in Figure 5).

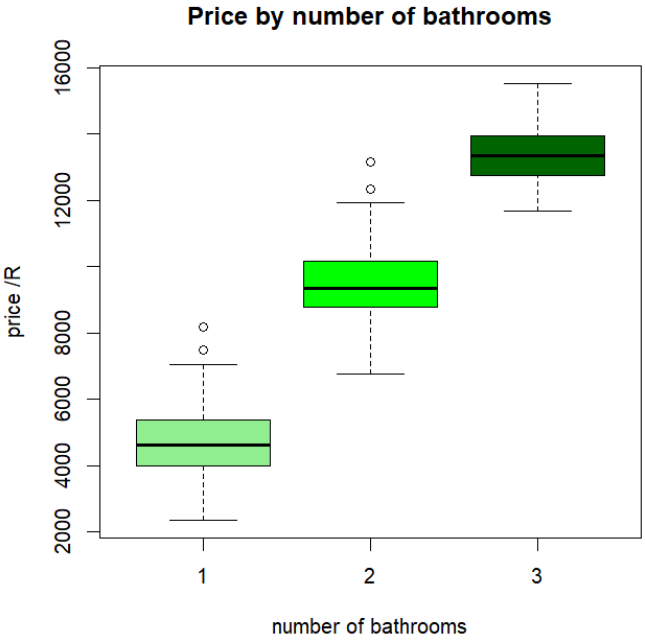

Figure 5: Boxplots showing price by number of bathrooms

The dataset included data regarding the distance to the nearest academic bookstore and the distance to campus, and scatterplots of these two variables showed no meaningful relationship between these variables and the price. Please see Figures 6 & 7.

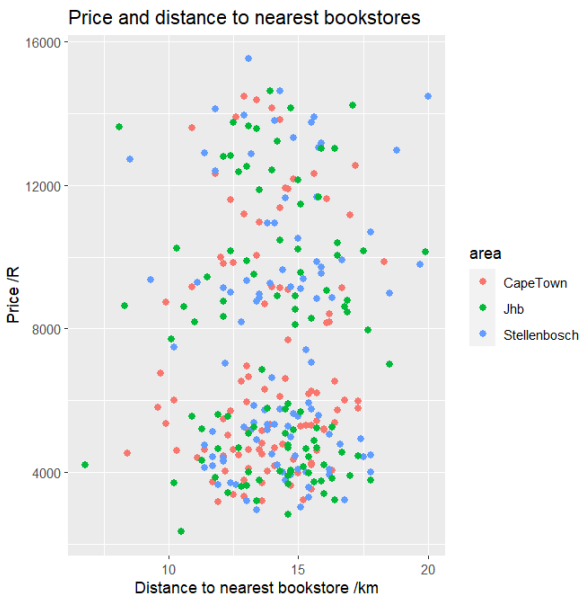

Figure 7: Scatterplot showing price by distance to nearest academic bookstore

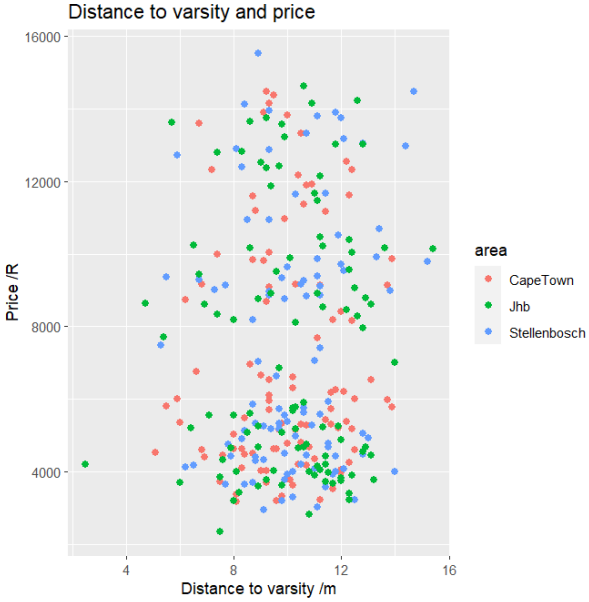

Figure 6: Scatterplot showing price by distance to varsity

3. Correlations

A correlation matrix of the quantitative variables (number of bedrooms, number of bathrooms, property square meterage, distance to varsity, distance to nearest academic bookshop, and price) shows which variables are highly correlated and which are not. This will give some insight into which variables are related to which other variables.

The value to indicate correlation lies between -1 and 1. A value close to -1 shows that the two values are negatively correlated, which means that as one increases, the other one decreases by a similar amount and vice versa. A value close to 1 shows that the two variables are positively correlated, which means that as the one increases, so does the other one (by a similar amount). The main diagonal of the matrix below will have all ones because each value is perfectly correlated with itself.

The results are as follows:

	No of bedrooms	No of bathrooms	Square meters	Distance to varsity	Distance to bookstore	Price
No of bedrooms	1	0.903	0.982	0.048	0.025	0.940
No of bathrooms	0.903	1	0.892	0.036	0.009	0.948
Square meters	0.982	0.892	1	0.031	0.009	0.922
Distance to varsity	0.048	0.036	0.031	1	0.959	0.055
Distance to bookstore	0.025	0.009	0.009	0.959	1	0.031
Price	0.940	0.948	0.922	0.055	0.031	1

From the above matrix, we can see that the following variables have a strong correlation:

- Number of bedrooms and number of bathrooms
- Number of bedrooms and square meterage of the apartment
- Number of bedrooms and price
- Number of bathrooms and square meterage of the apartment
- Number of bathrooms and price
- Square meters of the apartment and price
- Distance to varsity and distance to the nearest academic bookstore

4. Preliminary regression

The variables used to build Model 1 to predict the monthly rental price are: the number of bedrooms, number of bathrooms, square meterage, distance to the university, and the distance to the nearest academic bookstore. The city that the apartments are in and the state of the kitchen variables are not included in the model.

In the table below are each variable's coefficients and the p-value for each coefficient. The p-value is used to test the hypothesis that the coefficient is significant. If the p-value is less than 0.05, then we reject the hypothesis that the coefficient is not significant and conclude that it is, in fact, significant to the model.

Variable	**Coefficient**	**P-value**
Intercept term	668.061	0.417
Bedrooms (X_1)	2455.221	6.11x10^11
Bathrooms (X_2)	2518.138	2x10^-16
Square meterage (X_3)	-23.935	0.157
Distance to varsity (X_4)	-7.601	0.929
Distance to a bookshop (X_5)	27.726	0.738

This leads to the following linear model:

$$Y = 668.061 + 2455.221X_1 + 2518.138X_2 - 23.935X_3 - 7.601X_4 + 27.726X_5$$

These coefficients can be interpreted as follows:

The baseline apartment price is R668 per month. For each bedroom, the price increases by around R2 400. For each bathroom, the price increases by about R2 500.

For each unit increase in square meterage (i.e. for each square meter that the floor space increases by), the price decreases by R24. This was a surprising result because the price seemed highly correlated with the square meterage. The price decrease is not nearly as big as the increase that comes with the addition of a bedroom or a bathroom, so overall the effect of a bigger apartment on price is an increase.

For each additional kilometer between the apartment and the university, the price decreased by R7. This is not a very big decrease.
For each additional kilometer between the apartment and the closest bookshop, the price increased by R28. This is not a very significant increase.

The ANOVA (analysis of variance) table is as follows:

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
Number of bedrooms	1	3078364830	3078364830
Number of bathrooms	1	183920754	183920754
Square meterage	1	1650061	1650061
Distance to varsity	1	494121	494121
Distance to bookshop	1	83161	83161
Residuals	294	218510668	743234

The adjusted R^2 value tells us how much of the variation in the model is described by the variables in the model. The adjusted R^2 value is 0.9362. This means that 93% of the variation in the data can be described by the model.

5. Multicollinearity

Multicollinearity occurs when the variables are correlated with one another. This is a problem because it undermines our ability to note the contribution made by each unique variable. As a result, we want to identify the variables that cause multicollinearity and eliminate them from the model.

This can be achieved by first identifying which variables have the highest VIF (Variance Inflation Factor). This tells us how much the variances of the estimated regression coefficients were increased due to multicollinearity. Ideally, we want this to be as low as possible. The VIF for the variables in Model 1 is as follows:

Variable	**VIF**
Number of bedrooms	32.168946
Number of bathrooms	5.454416
Square meterage	28.983010
Distance to varsity	12.502307
Distance to the nearest bookstore	12.472153

As we can see from the output, the VIF for the number of bedrooms variable is the highest, followed by the square meterage.

Another way to detect multicollinearity is to use the condition indices and something called Spectral Decomposition to figure out how linearly dependent the predictor variables are, i.e. how "redundant" the information in the data is. Ideally, we want to eliminate the variables that create high degrees of linear dependence among the predictors.

The output for that is as follows:

Condition index	**Bedroom**	**Bathroom**	**Square meterage**	**Distance to varsity**	**Distance to bookshop**
1	3.7009e-03	0.02039	4.068e-03	9.3746	0.00005
1.2091	1.9111	0.0002	3.803e-05	2.0314e-02	0.02045
1.6903	0	0	0	0	0
4.6754	3.049e-02	0.9495	5.115e-02	3.912e-04	0.0002937
8.3586	5.187e-04	0.005395	9.127e-05	9.761e-01	0.9783
12.9408	9.653e-01	0.0245	9.447e-01	3.105e-03	0.0008677

From this output, we see that the highest condition index is 12.9408. We also identify the number of bedrooms as the variable with the highest dependence among the other variables, so we remove it.

After doing this and checking the results, the variable Distance to the nearest bookstore was also found to be a source of much multicollinearity. Therefore, it was also removed. Finally, we get to the following output:

Variable	**VIF**
Number of bathrooms	4.8913
Square meterage	4.8895
Distance to varsity	1.0013

Condition index	Bathroom	Square meterage	Distance to varsity
1.000	0.0538	0.05382	0.0015
1.3764	0	0	0
1.3781	0.00024	0.00033	0.9984
4.1864	0.9459	0.9458	0.0002

It makes sense that an apartment with more bedrooms will have a higher square meterage, so it seems redundant to include both the size of the apartment and how many bedrooms each apartment has when trying to model the price of each apartment. The one implies the other.

It also makes sense that the distance from the apartment to the university and the distance to the nearest academic bookstore are correlated because most academic bookstores are built near universities.

Both these variables have been removed from the model, and the model we're left with contains the variables Number of Bathrooms, Square Meterage, and Distance to Varsity.

6. All possible subsets regression

"All possible subsets" regression is a process in which we run tests on all the possible combinations of predictor variables (e.g. number of bedrooms, number of bathrooms, etc) and find which of the combinations of variables satisfy certain requirements.

In this case, we ran "all possible subsets" regression to find which of them had the maximum adjusted R^2 value. This value gives an indication of how much of the variation in the model is explained by the variables in the model, and it penalizes adding 'junk' variables into the model. Thus, it is generally considered a good indication of how "good" a model is.

Using "all possible subsets" regression on the model with distance to varsity, number of bedrooms, area, and kitchen as predictors, we found that the model with the highest possible adjusted R^2 value is the model using only bedrooms, area, and kitchen as predictors.

7. Influential values

In this part of the analysis, the influential and severe values were identified and deleted from the data frame.

8. Final model

We created the final model with the following predictors: distance to varsity, number of bedrooms, area, and kitchen. This model does not include the values identified as influential and severe.

9. Testing assumptions

Before we can use this model, we have to test it for the following assumptions:

- The first assumption is that the errors of the model all have the same variance.
  After statistical testing, it was found that the errors do, in fact, have equal variance. We can continue to the next test.
- We assume that the model is normally distributed. This means that it follows what's called a 'normal' distribution. This is useful because the normal distribution has certain properties that we can make predictions with.
  After formal statistical testing, it was found that the mode
- We thus conclude that we can continue with using the model because it follows a normal distribution (which is important for in linear modelling; we can't assume that the model is accurate)

10. Predictions

Using, the final model, we can predict the price per month for the following apartment:

Square meters	300m²
Distance to varsity	5km
Distance to bookshop	4.8km
Number of bedrooms	3
Number of bathrooms	2
Kitchen	modern

The model is: $Y = 882.83 + 8.33X_1 + 3842.56X_2$

We only substituted the following values in: 5km into X_1 (distance to varsity), and 3 into X_2 (number of bedrooms). The model is built with the assumption that the apartment is in Cape Town and the kitchen is modern.

The predicted price is **R12 452.15 per month**

11. Conclusions

To summarize, the regression analysis using the independent variables of "dist_to_varsity", "bedr", "area", and "kitchen" produced an accurate prediction of the monthly rental price for the apartment. According to the model, the anticipated rental price per month for the apartment is R12 452.15, which can serve as valuable information for individuals involved in the student apartment rental market (e.g. potential students, investors, and even businesses who operate in the areas in which student apartments are found) . Property owners and real estate agents can benefit from the insights provided by an regression model to inform their pricing and marketing strategies,

while potential renters can gain valuable insights into the rental market and make more informed decisions about renting an apartment that aligns with their preferences and budget.

12. Appendix

Code for assessing data cleanliness

```
df <- data.frame(read.csv("rent_data.csv"))
```

The fact that the data was successfully loaded into R-Studio properly is a good indication that no string was entered as numbers and vice versa, etc.

Code for exploratory analysis using graphical techniques

```
# price and city

boxplot(price~area, main="Price by city", xlab="City", ylab="Price",
col=c("khaki","gold","yellow"))

# price and size of apartment

ggplot(df, aes(x=prop_sqr_m, y=price, color=area)) +
geom_point(size=2, ) + ggtitle("Price and size of apartment") +

  labs(x="Property size /square meters", y="Price /R")

# distance to nearest academic bookshop and price

ggplot(df, aes(x=dist_to_bookshop, y=price, color=area)) +
geom_point(size=2) + ggtitle("Price and distance to nearest
bookstores") +

  labs(x="Distance to nearest bookstore /km", y="Price /R")

# distance to varsity and price

ggplot(df, aes(x=dist_to_varsity, y=price, color=area)) +
geom_point(size=2) + labs(x="Distance to varsity /m", y="Price /R")
+ ggtitle("Distance to varsity and price")

# kitchen and price

boxplot(price~kitchen, main="Price by kitchen", xlab="Kitchen",
ylab="Price /R", col=c("aquamarine","darkgrey"))

# bedroom, bathroom, and price

boxplot(price~bedr, main="Price by number of bedrooms", xlab="number
of bedrooms", ylab="price /R",
col=c("aquamarine","skyblue","cornflowerblue"))
```

```
boxplot(price~bathr, main="Price by number of bathrooms",
xlab="number of bathrooms", ylab="price /R",
col=c("lightgreen","green","darkgreen"))
```

Code for correlations analysis

```
df2 <- subset(df, select = c(-kitchen,-area))

round(cor(df2),3)
```

Code for preliminary regression

```
Model1 <- lm(price ~ bedr + bathr + prop_sqr_m + dist_to_varsity +
dist_to_bookshop, data=df)

summary(Model1)

anova(Model1)
```

Code for multicollinearity analysis

```
df_scaled <- as.data.frame(scale(subset(df, select=c(-area, -
kitchen))))

mod <- lm(price ~ bedr + bathr + prop_sqr_m + dist_to_varsity +
dist_to_bookshop, data=df_scaled)

ols_coll_diag(mod)

Model2 <- lm(price ~ bathr +  prop_sqr_m + dist_to_varsity +
dist_to_bookshop, data=df_scaled)

ols_coll_diag(Model2)
```

Code for all possible subsets regression

```
Model3 <- lm(price ~ dist_to_varsity + bedr + area + kitchen,
data=df)

summary(df)

result <- ols_step_all_possible(Model3)

maxrsq <- max(result$adjr)

i <- which(result$adjr == maxrsq)

result$predictors[i]
```

Code for influential values analysis

```
influence.measures(Model3)

influence <- influence.measures(Model3)

influence

b <- influence$infmat[,1:6]

dfbetas_infl <- b>2/sqrt(300)

sum_infl <- c()

for (i in 1:300){

  cnt=0

  for(j in 1:6){

    if(b[i,j]>2/sqrt(300)){

      cnt = cnt + 1

    }

  sum_infl[i] <- cnt

  }

}

sum_infl

del_indx <- c(which(sum_infl>0))

del_vals <- c(Model3$residuals[del_indx])


rstud <- c((del_vals-mean(del_vals))/sd(del_vals))

new_rent_df <- df[-del_indx,]
```

Code for final model

```
Model4 <- lm(price ~ dist_to_varsity + bedr + area + kitchen,
data=new_rent_df)
```

Code for testing assumptions

```
res4 <- Model4$residuals

# formal hypothesis test

# H0: errors are normal vs HA: errors not normal
```

```
shapiro.test(res4)

# at significance level alpha = 0.05 we accept H0 and conclude that
the data normally distributed

qqnorm(res4)

qqline(res4)

hist(res4, freq=FALSE)

d <- density(res4)

lines(d, lty=2)

r41 <- res4[1:(268/2)]

r42 <- res4[(length(res4)/2):268]

var.test(r41,r42)

# at significance level alpha = 0.05 we accept H0 and conclude that
the data are homoskedastic
```

Code for predictions

```
new_data <- data.frame(dist_to_varsity = 5, bedr = 3, area =
"CapeTown", kitchen = "modern")

predicted_price <- predict(Model4, newdata = new_data)

predicted_price

summary(Model4)
```