

Statistical Assignment

Dimitri Saridakis | 20083467| 19.04.2020

Table of Contents

[Part A 2](#_Toc39524983)

[Q1: Do houses and apartments tend to differ in rent amount paid? 2](#_Toc39524984)

[Q2: Is population size a factor in rent level? 5](#_Toc39524985)

[Q3: Is the ratio of house occupants to apartment occupants the same for all occupant types? (E.g. if for families, 60% are in houses and 40% in apartments, is the 60:40 ratio the same for other occupant types?) 7](#_Toc39524986)

[Q4: Is the percentage of houses managed by property management companies about the same as for 9](#_Toc39524987)

[apartments? 9](#_Toc39524988)

[Q5: Excluding “others”, is there evidence of differences in average rents between family, house share and single occupant? 11](#_Toc39524989)

[Q6: Have rents increased from where they were one year ago? 14](#_Toc39524990)

[Part B: 16](#_Toc39524991)

[Q7: Simple Regression: 16](#_Toc39524992)

[Q8: Multiple Regression 20](#_Toc39524993)

[R Code: 25](#_Toc39524994)

[#question1 25](#_Toc39524995)

[#question2 25](#_Toc39524996)

[#question3 27](#_Toc39524997)

[#question5 29](#_Toc39524998)

[#question6 31](#_Toc39524999)

[# Q7 Simple Regression 33](#_Toc39525000)

[# Q8 Mulitple Regression 34](#_Toc39525001)

# Part A

## Q1: Do houses and apartments tend to differ in rent amount paid?

Below are some summary statistics for the rent:

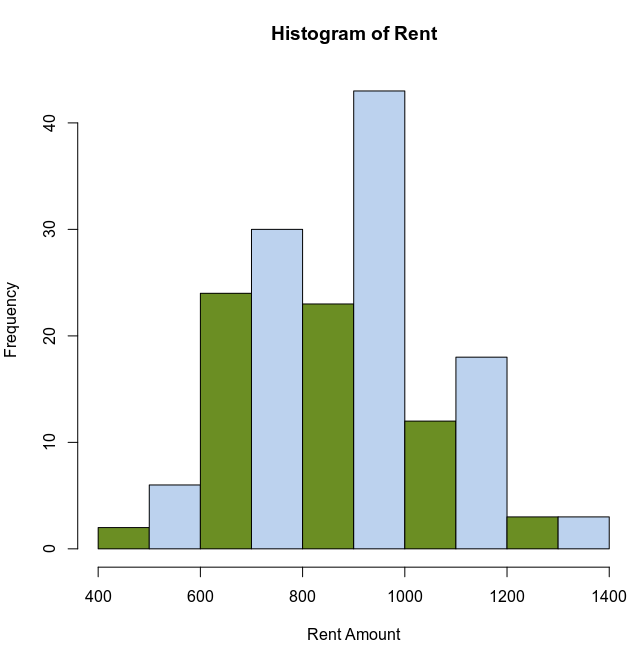
|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Standard Deviation | Sample Size |
| House | 853.8961 | 189.9387 | 80 |
| Apartment | 911.0345 | 184.1197 | 89 |
| Combined | 884.2073 | 188.4811 | 169 |

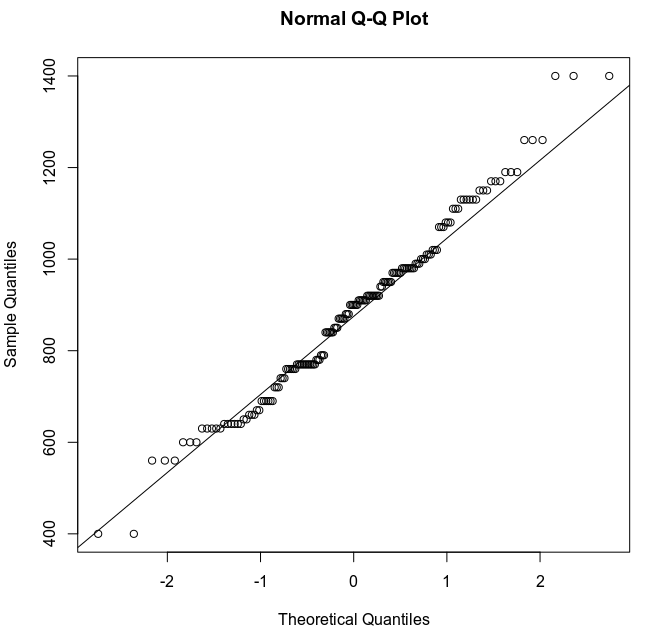
The Standard Deviation, as shown in the table, is a number that shows how spread out each particular sample value is in relation to the mean of that sample. A low value for SD indicates that most numbers are close to the mean whilst a high value for standard deviation indicates that the variables are quite spread out.

If the values, when plotted on a histogram form the shape of a bell curve (roughly), with the highest bars in the middle of the graph then the data sample is said to have a normal distribution. Or if the data Q-Q plot is in the shape of a straight(relatively) diagonal line this is also another visual que for normality of a distribution. The mean, mode and median are equal (or close) and there is a sense of symmetry of the rest of the values corresponding to the aforementioned bell-shaped histogram.

Using the independent samples t test in R, the output for this test yields a p value of 0.05289. As p > than 0.05 we can conclude there isn’t a statistically significant difference between the amount of rent paid for houses and apartments using our chosen confidence interval of 95%. In essence, this means that 95% of experiments like we just did will include the true mean, but 5% will not. So, there is a 5% chance that our Confidence Interval doesn’t include the true mean. Using a different R function, namely boxplot.stats(rentTypeQ1), we can garner that there are 5 outliers that may affect the data. These are two rent values of 400 and 3 rent values of 1400.

We want to use a t-test in R, which assumes a normal distribution, as it is the accurate test to use with this sample data. We can find this out in several ways, visually we can get an indication in a histogram by the shape of the graph (bell shape), whilst in a Q-Q plot the dots should be a diagonal line.





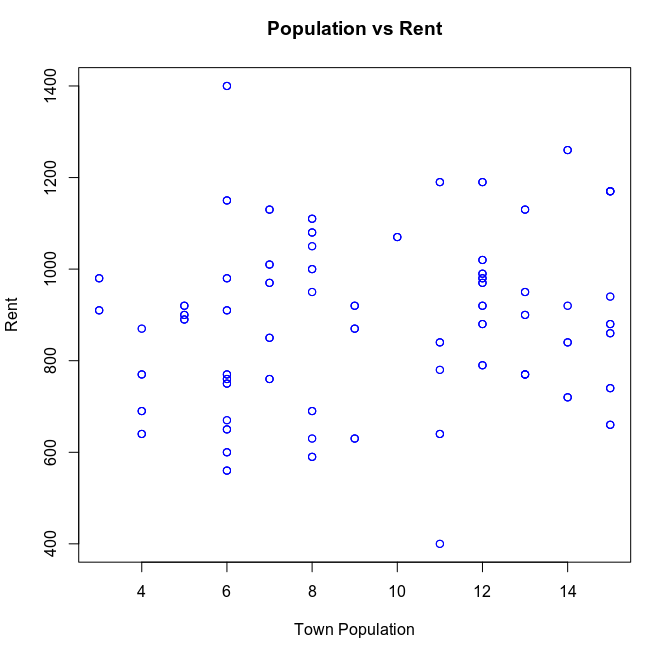
As we can see the distribution seems like it may be normal, but we confirm with either a Shapiro-Wilks test or a Kolmogorov-Smirnov test for normality.

Seeing as the sample size is larger than 50 will use the KS test. With a p value derived from the KS test of p < 2.2e-16, which coincidentally is the lowest number that can be stored on a floating-point computer system, as the p-value is under 0.05 this means that we reject the null hypothesis, we don’t believe that our variable follows a normal distribution.

In conclusion we cannot definitively say that houses and apartments differ in rent being paid.

## Q2: Is population size a factor in rent level?

This analysis is between two variables that are both numerical/scaler, so we begin by getting a visual representation of the data. For this we use a scatter plot.



We can arguably see a very vague trend toward a linear increase in rent amount as population increases. However, as the results are visually inconclusive, some more analysis is needed to see if there is any correlation between the two. As both our variables are scalers, we will use a correlation test to determine the magnitude of the correlation.

The correlation r is 0.1520635. Whilst the r2 is computed as 0.02312331. We shall use the r2 figure. This figure of 0.02 indicates that a mere 2% of the variation in rent price depends on the population. We will conduct a further test to determine if any correlation exists. Here we have two different scenarios.

H0: There is zero correlation between our variables. (null hypothesis)

HA: There is a correlation between the population and the price (alternate hypothesis)

We will use a Pearson's product-moment correlation to determine whether we accept or reject the null hypothesis. As our p value is 0.03672, which is less than 0.05, we reject the null hypothesis that there is no correlation. In conclusion we can say that there is a correlation even if the evidence of a correlation is very weak.

These tests assume that these two variables follow a joint normal distribution, but we must confirm this using the Henze-Zirkler test. Using this test, we obtain a p value of 7.001226e-08. As the p value < 0.05 we reject the null hypothesis of the assumption of a normal distribution. We also get a HZ value of 3.35226 from this value the p-value is calculated for us.

## Q3: Is the ratio of house occupants to apartment occupants the same for all occupant types? (E.g. if for families, 60% are in houses and 40% in apartments, is the 60:40 ratio the same for other occupant types?)

Some summary statistics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Family | House Share | Single Occupant | Other | Total |
| House | 39 | 9 | 25 | 5 | 78 |
| Apartment | 39 | 18 | 9 | 22 | 88 |
| Other | 20 | 0 | 2 | 3 | 25 |
| Total | 98 | 27 | 36 | 30 |  |

We lay out our null hypothesis that there is no difference between the ratios for the different occupant groups and the type of dwelling where you’d find them inhabiting.

For this question our samples to use are categorial therefore the appropriate test to use in this instance is the χ2 test of independence. Deploying this test in r returns a p-value = 1.115e-05**.** As p-value < than 0.05 we can conclude with a 95% confidence interval that we can reject the null hypothesis of independence that the ratio is not the same for all occupant types in dwelling types. However, the χ2 test does also display a warning, namely;

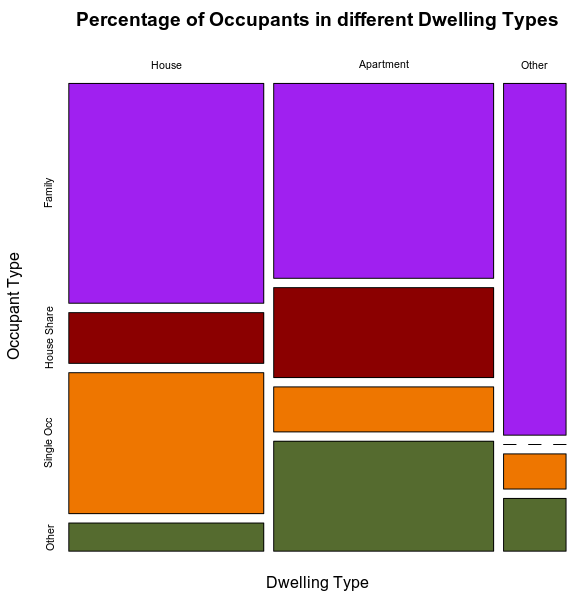
‘Chi-squared approximation may be incorrect’

And as such to get a more accurate answer we would need to alter the data. This goes out of scope of the question though so we will produce a bar plot to visualise the data where one can also judge the differences.

We make a note here that the χ2 test of independencies a non-parametric test. It applies to two nominal / binary (categorial) variables. Such variables are never normally distributed so checks on normality of distribution do not apply for this test as it not assumed by the test.

In conclusion the ratio is not the same the same for all occupant types but we cannot prove this as our data won’t let us.

We can visualise the data like this:



## Q4: Is the percentage of houses managed by property management companies about the same as for

## apartments?

First, we shall start off with some summary statistics from our data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Managed | Not Managed | Total | % Managed |
| Houses | 27 | 52 | 79 | 34.18 |
| Apartments | 23 | 64 | 87 | 26.43 |
| Difference |  |  |  | 7.75 |

We lay out our hypothesis that is, we test that the percentage of houses under management is independent of percentage of apartments under management.

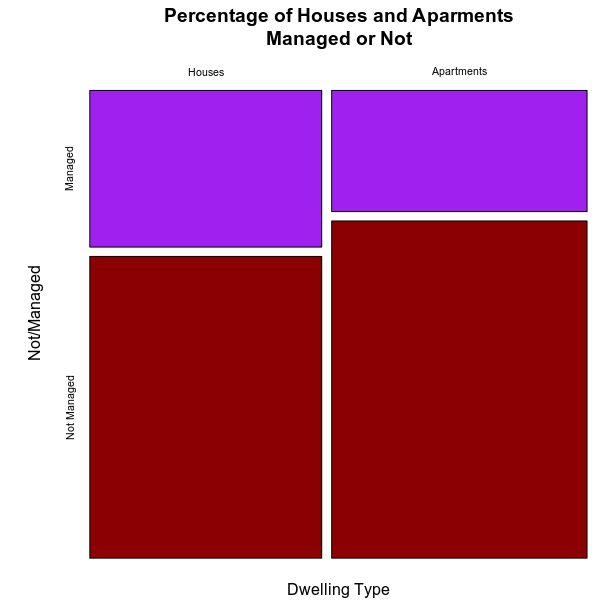
H0: Assumes there is no association between the two variables. They are independent.

HA: Assumes there is an association between the variables. They are **not** independent.

As we have two categorial variables, the appropriate test to use is the χ2 test of independence. Running this test in r returns a p-value of 0.3595. As p-value > 0.05 we can say with a 95% confidence interval that we do not reject the null hypothesis that the variables are independent/ there is no association between the variables. As such we can say that a statistically significant difference may exist.

In conclusion we can say that the percentages of houses managed by property management companies differs for apartments.

We will use a bar plot to visualise the data:



## Q5: Excluding “others”, is there evidence of differences in average rents between family, house share and single occupant?

Some statistics from sample:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Standard Deviation | Sample Size |
| Family | 859.7917 | 175.9185 | 99 |
| House Share | 823.3333 | 154.7703 | 27 |
| Single Occupant | 953.7143 | 241.1259 | 37 |

First, we lay out our hypothesis that we are testing.

H0: The average spend on rent is the same (no variation) between the three groups.

HA: The average spend on rent is **not** the same in all three groups.

As we have three different groups to compare in our sample the correct test to use is an ANOVA test, assuming the criteria of which the ANOVA test assumes is satisfied, namely normality of distribution and homoscedasticity of the data (I will elaborate on this later). Running the ANOVA test in R gives back a p-value = 0.0365. As the p-value is below 0.05 we can reject the null hypothesis and accept the alternate (subject to normality testing) that the average rent is not the same in all three groups, and that there is a statistically significant difference. I’d further like to carry out a Post Hoc test to try and determine the significant difference between each group. The test I will use is the Tukey’s honest significant difference. Running it in R gives the following output:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Difference | Lower | Upper | P-Value |
| A: House Share vs Family | -36.45833 | -133.913605 | 60.99694 | 0.6503840 |
| B: Single Occ vs Family | 93.92262 | 5.586912 | 182.25833 | 0.0342468 |
| C: Single Occ vs House Share | 130.38095 | 15.790029 | 244.97188 | 0.0213592 |

In the above data table, the ‘Difference’ is the difference between means of each grouping.

Taking case A’s p-value of 0.6503, as this p-value > 0.05 we can say there is not a statistically significant difference. The [-133.91, 61] is a 95% confidence interval for the mean price of rent paid between the two class of occupants. We accept the null hypothesis.

When we look at both cases B and C, their p-values are both below 0.05 and as such we can classify the variance in rent paid as statistically different. In these two cases we reject the null hypothesis.

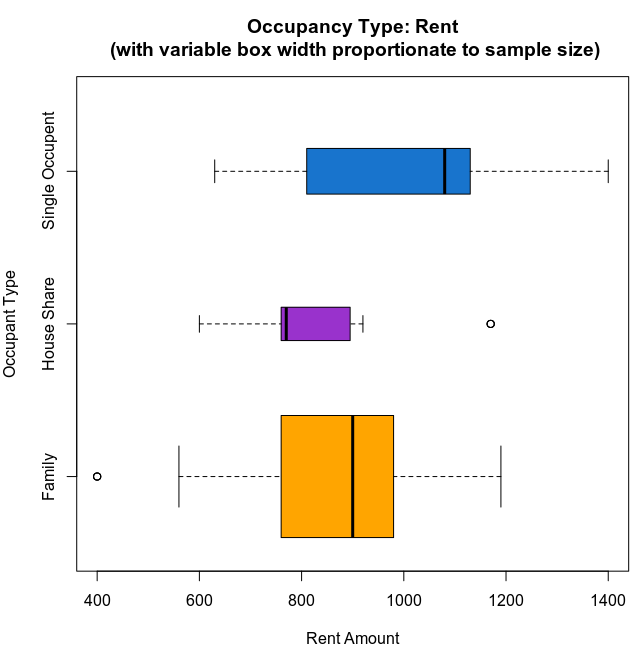
In case B the given p-value of 0.034 < 0.05. [5.59, 182.26] is a 95% confidence interval for the levels of rent paid between single occupants and Families.

Whilst in case C a p-value of 0.021 < 0.05. [15.80, 244.97] is a 95% confidence interval for variance in rent levels paid between single occupants and those in a house share situation.

These results do come with a caveat, in that for the use of an ANOVA test some assumptions must be made about the dataset, as previously stated.

1. Normality of distribution, as seen previously.
2. The variance in each sample should be homoscedastic. That is, the error term (i.e. the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same finite variance across all values of the independent variables.

This is one of the reasons the standard deviations are calculated and displayed in the first table. It is also why a box plot is chosen to represent the data visually.



However, we can use a formal hypothesis test to more accurately gauge the homoscedasticity of the data. For this we shall use the Levene’s Test and we set out our hypothesis as follows:

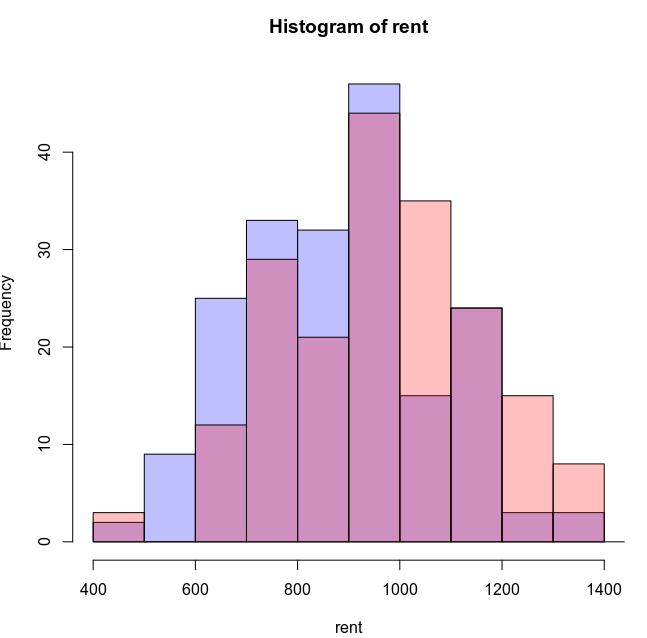
H0: The variance is same for all three groups.

HA: The variance is **not** the same, i.e. it differs between groups.

Running Levene’s test on the data returns a p-value of 0.0054949. As the p-value < 0.05 we must reject the null hypothesis and even without completing a test to confirm the normality of the distribution we cannot report a statistically significant difference. Since the assumptions of the ANOVA are not met, further nonparametric tests can be carried out such as the Kruskal-Wallis test, but these tests are outside of the scope of this report.

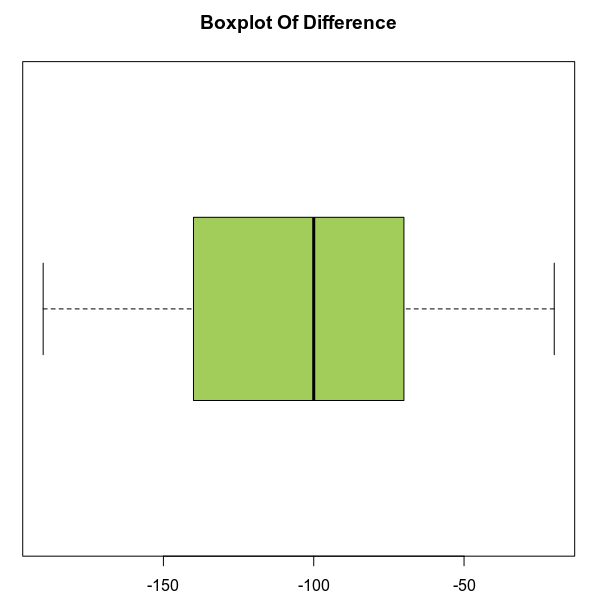
## Q6: Have rents increased from where they were one year ago?

To answer this question, we must compare rent levels as they currently stand with rent levels from one year previous. We will use a clustered alpha-blending histogram to visualise the data. In this Histogram the blue is the value for ‘rent’ (current), the red is the value for ‘previousRent’ whilst the purple is the overlap of both. Just by looking at the Histogram we can judge that the rent levels have not only not increased but that they appear to have indeed fallen compared to 12 months prior. Further analysis will reveal more.



Below are some summary statistics for our variables:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Standard Deviation | Sample Size |
| Rent(current) | 888.25 | 187.62 | 189 |
| Previous Rent | 986.88 | 199.19 | 189 |
| Difference | -98.62 | -11.56 | 0 |



The correct test to use in this scenario is the paired t test. The hypotheses we test are.

H0: The true mean difference between the two samples is zero (no variation).

HA: The true mean difference is **not** zero. There is a difference between the two samples.

Generated by r, the p-value < 2.2e-16. As the p value is less than 0.05 we can conclude that there is a statistically significant difference between the two variables. With a 95% confidence interval for rent variation given by [-104.52, -92.73]. This range does not include zero, so we can conclude (with a presumption of normality in our distribution) that there is a statistically significant difference between current rent levels and previous rent levels and a result we can reject the null hypothesis.

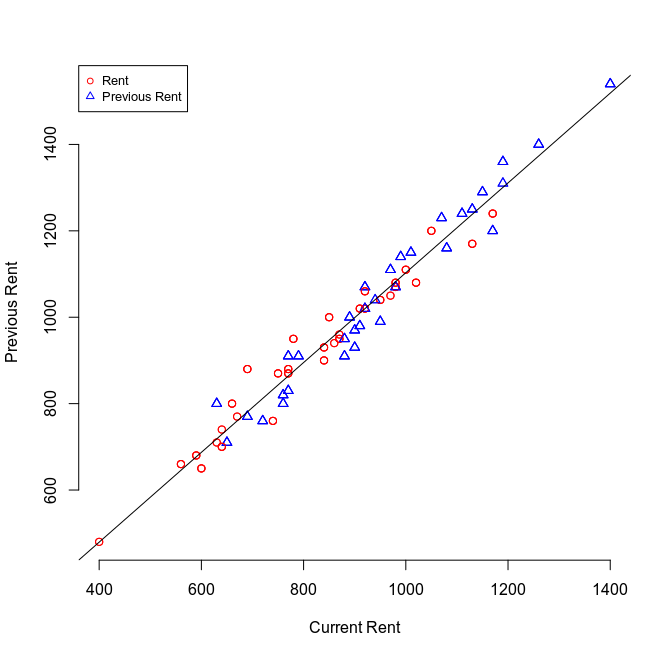
As we made a presumption of normality in using the paired t test, we will now test the variables to confirm that presumption. We have a sample size of over fifty variables in our sample data, we will use the KS test to check for a normal distribution.

Upon running the KS test in r, we get a p value of 2.746e-05. As this value is below 0.05 we reject the hypothesis of normality in the distribution. I further ran the Shapiro-Wilks test on the rent difference which returned a p-value = 0.0008901. This corroborates the finding in the KS test.

In conclusion no rents have not increased from 12 months previous.

# Part B:

## Q7: Simple Regression:



The two values I have chosen to use for my regression model are Rent and Previous Rent as displayed on the above scatter plot. Here the scatter plot suggests a strong correlational relationship. Using the correlation test in r we get a r value of 0.98. A value of 1 would indicate that the data is fully correlated whilst a value of 0 would indicate no correlation (as seen in the correlation between rent and population). An r2 value of 0.96 corroborates the strong correlation and indicates that 96% of the change in rent can be attributed to the previous rent levels.

The equation for regression models comes from the slope of the regression line fitted to the data. The equation of a line is given as y = mx + c. In regression models we will use this formula to complete our model. The intercept (or ‘a’) will go in place of the ‘c’ in the equation of a line formula, whilst the coefficient is calculated by taking the range of values and dividing it by the step. Using the lm function in r gives us back these values for us. The regression model is:

rent = 0.92previousRent – 21.98

The coefficient of rent (0.92) indicates the average effect a unit change of previous rent will have on current rent levels per a single unit change in rent.

We will deploy ‘summary (lm(rent ~ previousRent)’ to our dataset. The ‘previousRent’ in this instance is known as the predictor/ explanatory variable whilst the ‘rent’ variable is known as the target/response variable. This deployment returns a p-value, of a test if the coefficient were set to zero, is <2.2e-16. Since the p-value is < 0.05 we can interpret this to conclude that the variables are **not** independent, and the results are statistically significant. The function also gives us some other data as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max | Est. Std. | Std. Error | t-value |
| -99.67 | -27.35 | 1.98 | 25.76 | 85.18 | -21.98 | 14.07 | -1.56 |
| 0.922 | 0.014 | 65.967 |

The values in gold are known as the Residuals, these are essentially the difference between actual observed data and the values that our model has predicted. A symmetrical distribution across the median gives us some insight that our data fits our model well. 50% of the residual values are within +- 28 of the median. Further tests of normality could be carried out using the residuals to check for normality of the distribution for confirmation.

The values in red are coefficient values. Top row are the values for the intercept, the expected values for when considering the average rent of all data in the dataset. Whilst the bottom are the slope values or predictions based on unit changes.

The ‘Est. Std’ is the level of value of our response variable when our predictor carriable is set to zero in our data set this means that the average rent in our dataset is -21.98 below average previousRent level when the previousRent level is zero. Whilst the 0.922 is the coefficient of rent as previously explained.

‘Std. Error’ shows the average amount that the given coefficient number varies from the actual average of the response variable. The lower this number is in relation to the coefficient number the better our model is. In this our case we can expect the difference/ error in coefficient to be out by a mere 0.014. This suggests our model fits our data very well.

The ‘t-value’ is a measure of how many standard deviations our coefficient is away from zero, it is also used in the computation of the p-value.

We also have a Residual standard error of 38.18. This is a measure of how much deviation there is from the true regression line. If we make a call to scale(previousRentQ7) we can scale the previousRent so that the mean of our new variable is 0. Using this technique we can use the new coefficient of 888.254 and divide it by the residual standard error(38.18) to obtain a new value of 23.26 which is the possible percentage error of a prediction based on our model.

The multiple r2 value (0.9588) shows that 96% of the variation of rent can be attributed to the previous rent. We have also derived this value previously using the cor2 test. These stats make logical sense.

We can take an example based on this model, if your lease is nearly up and you’re looking to calculate the amount of your new rent we can say with a 95% confidence interval that if your rent is 600 then this model reads that 530.02 will be your new rent going forward.

A further example of 900 would leave our new rent level to be 806.02. This example of 900 is within our actual data set, as this is so we are interpolating to make our prediction.

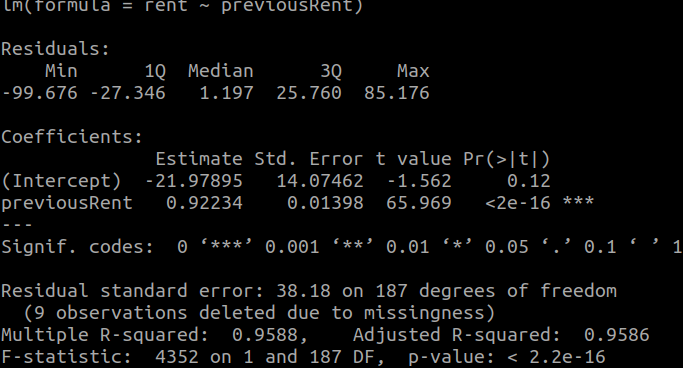
A value of 100 into our model would yield a value of 70.02. as the initial value of 100 is not within the range of our data said we are said to be extrapolating to make our prediction. This is not as credible as interpolation.

Using the mean of the previous rent (981.557) with our model (0.92\*(981.86) - 21.98) returns the value 881.33. The value of 888.81 is the true mean of rent. This gives us a difference of 7.48. Our model appears to be quite accurate.

Yes, I believe the predictions match up with the quite well with the observed values. The factor I have used in this model is definitely a useful predictor of rent. Short of external economic factors, drastic changes to geography or the dwelling state the previous rent is one of the major if not the major prediction variable.

I would have more confidence in the values that are within the range of the dataset(interpolated) as once rent levels go low they aren’t really going to get cheaper although we must use caution with this model as it indicates rent prices declining from previous levels, whilst this may happen in a market correction after a bull run into a bear market this isn’t generally the way true values go due to inflation. I would be recalculating the model frequently to take the other external factors(not in the data set) into consideration or consider using a multiple regression model.

Summary lm output from r for regression model.

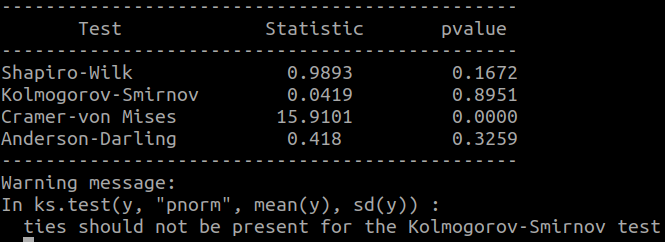


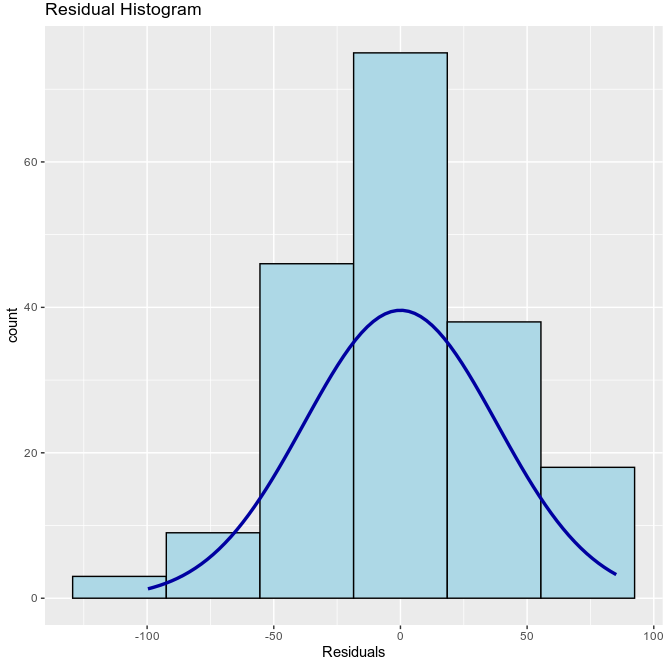
There are a number of assumptions that regression models are based on. For the purpose of this report I’ll just be using these the normality of distribution.

Some other assumptions include:

1. The errors have mean zero.
2. The errors have same but unknown variance (homoscedasticity assumption).
3. The error are independent of each other (independent errors assumption).

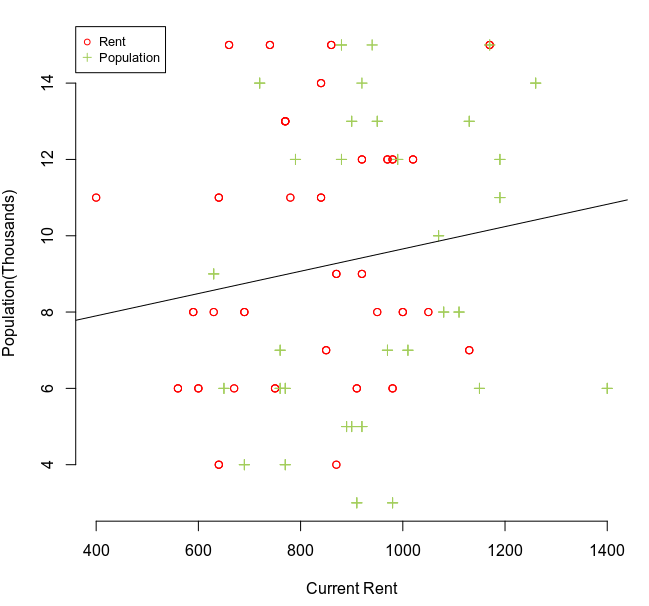
I’ll start with normality and will be availing of the olsrr package for r. This is the results for normality assumptions. P-values > 0.05 therefore we can conclude with a 95% confidence variable that the data is normally distributed. Calling ols\_test\_normality(model) returns:



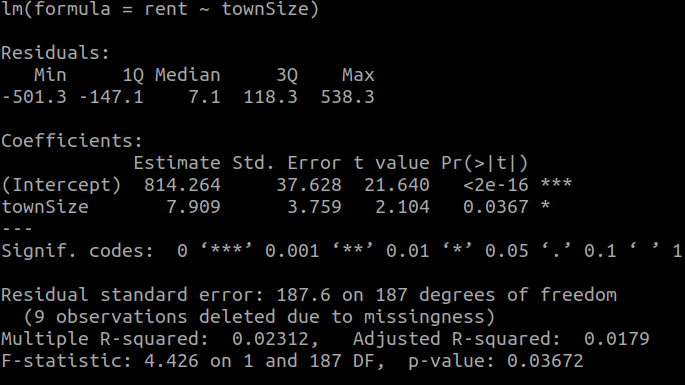


## Q8: Multiple Regression

To build the next model for rent predictions I will use a scaler - townSize (population) along with the categorial variable management. I will create a model based on these variables then briefly look at the other categorial variables to explore if any could have potentially been more suitable fits for the data. We can visualise our data on the following scatter plot.



Calling summary(lm(rent~ townSize)) returns a summary as follows:

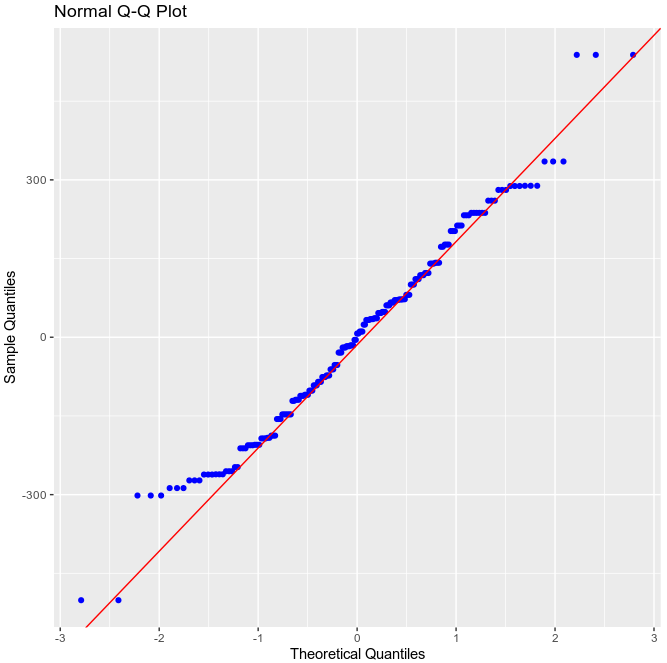


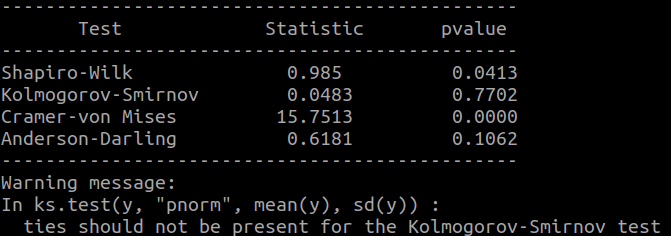
Noting both visually from the plot and from the r output, that with a r2 value of 0.0179 we can conclude that just 1.8% points of variance in rent levels can be attributed to these two variables collectively and with a cor test value of 0.1520635 there isn’t much of a correlation either.

Our model is;

Rent = 7.909(population) + 814.26

Next I will check the normality of the distribution:





The main test I will be paying attention to here is the Kolomogorov-Smirnov test given the sample size of the data set. With p-value > 0.05 therefore we can conclude with a 95% confidence variable that the data is normally distributed.

Using this model with 8 for population (interpolated) our given rent value is: 870.9872

Using the mean population value (9.314433), also interpolated.

Rent = 7.0909\*(9.314433) + 814.26 = 880.3077

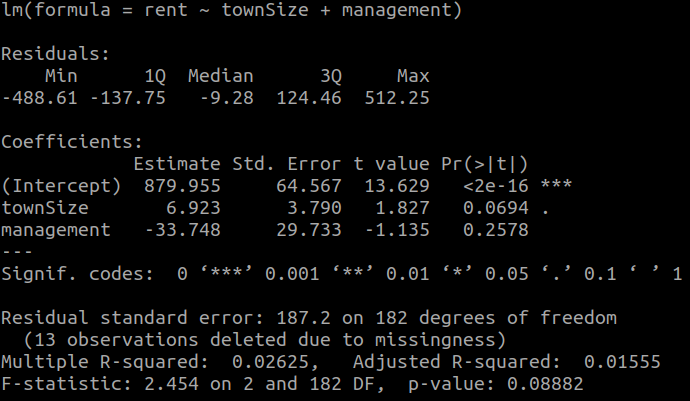
Given that the mean rent is 888.81, that isn’t too far off, in fact it’s approx. 1 off the same calculation carried out with the *rent-previousRent* model.

Using an extrapolated value;

Rent = 7.0909\*(30) + 814.26 = 1026.987

I wouldn’t have too much confidence in the model for predictions with such low r2 values although the predictions do seem to come close to expected values.

Next we will add another variable to the model, Management. A summary(lm(rent~ townSize+ Management)) call returns the following output:



In this case the combination of these two variables result in p-values > 0.05 that are not statistically significant. The r2 value comes in at 0.016 or 1.6% of the variance of rent levels is determined by our chosen variables. Our new regression model is

Rent = 6.92(townSize) -33.75(management) + 880

Where townSize is the population in thousands and management is either a 0 or a 1 as it’s a binary variable. We will be using some similar figures to test our model.

Interpolated:

7.909\*(9.314433)-33.75\*(1)+880 = 919.92

7.909\*(9.314433)-33.75\*(0)+880 = 953.67

Extrapolation:

7.909\*(30)-33.75\*(1)+880 = 1083.52

7.909\*(30)-33.75\*(0)+880 = 1117.27

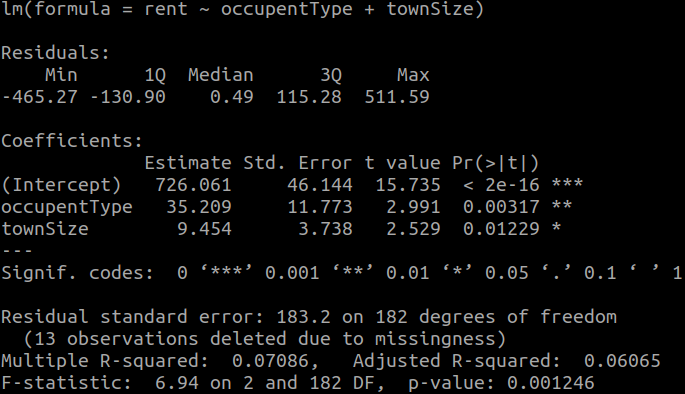
As we can see there is a difference in rent between when the property is managed when compared to not being managed. However, this model appears to be drifting further away from actual rent prices.

I shall take a look at some of the other variables available to us in our data set namely, dwelling type and occupant type to see if we can find any evidence of a better suiting variables by deploying summary(lm(rent ~ <variables>)) of all independent variables I have from the data set. As shown below this gives me back very lowr2 and adjusted r2 values for all variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Town Size | Management | Dwelling type | Occ Type |
| R2 | 0.02312 | 0.007759 | 0.02423 | 0.04155 |
| Adj. R2 | 0.0179 | 0.002453 | 0.01904 | 0.03643 |

Briefly looking at these figures gives us some insight that the dataset is skewed. I will quickly see if other variables could have given us a better fit to the data and if we can cover more percentage of variance in the rent with these variables.

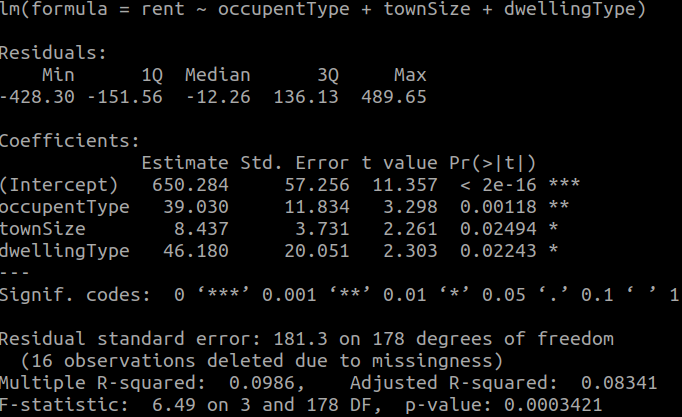
summary(lm(rent~ occupentType + townSize)) returns:



The Pr(>|t|) gives the p-value per each variable and whether we can accept or reject the null hypothesis for that variable to determine if the variable is statistically significant or not. The \*s beside the value define significance values. The lower the p-value the more significant the finding, this is further shown by having more stars, subsequently allowing us to reject the null hypothesis

We notice that collectively townSize and occupant type only have an adjusted R2 value of 0.061, that is 6% variation in rent can is due to these two variables whilst both are statistically significant with p-values less than 0.05.

Including dwellingType as a variable in our model returns this output:



In all 3 cases the p-value < 0.05 meaning that the variables are statistically significant to the price of rent. With a r2 value of just 0.083, collectively the variables only contribute to approximately 8%-point variance of rent levels.

It seems management and townSize were a poor choice of predictor variables for the model, although the remaining variables showed a little more potential, they collectively represented a small percentage variance in rent.

The best model would have previousRent and potentially a combination of occupant type and dwelling in the model.

# R Code:

========================================================== #Part A

==========================================================

## #question1

-----------------------------------------------------------------------------------------------------------------------------

# to plot scatter graph

plot(townSize, rent, xlab = "Town Population", ylab = "Rent")

# subset of data without na values

rentTypeQ2 <- subset(rent, !(is.na(townSize) | (is.na(rent))))

popTypeQ2 <- subset(townSize, !(is.na(townSize) | (is.na(rent))))

# correlation tests

cor(popTypeQ2,rentTypeQ2)

cor(popTypeQ2,rentTypeQ2)^2

# pearson correlation test

cor.test(popTypeQ2,rentTypeQ2)

# HZ test for Multivariate Normality

HZ.test(data.frame(popTypeQ2,rentTypeQ2))

-----------------------------------------------------------------------------------------------------------------------------

## #question2

-----------------------------------------------------------------------------------------------------------------------------

# for q1 to subset dwellingType column to allow for just vars 1(house) or 2(apartment)

dwellTypeQ1 <- subset(dwellingType,dwellingType < 3, na.rm=TRUE )

# for q1 to subset rent column to exclude the same var rows that were excluded in dwellTypeQ1

rentTypeQ1 <- subset(rent,dwellingType < 3, na.rm=TRUE )

# using the two data columns in a t.test and saving the output to q1Ttest

t.test(rentTypeQ1~dwellTypeQ1) -> q1Ttest

# to create a boxplot comparing the rents

boxplot(rentTypeQ1~dwellTypeQ1,names=c("House", "Apartment"),xlab = "Rent", ylab = "Dwelling", horizontal = T)

# to generate some stats that detail the outliers from the boxplots($out)

bxstats <- boxplot.stats(rentTypeQ1)

#saves the outliers in the var q1outliers

q1outliers <- bxStats[4]

# pastes outliers over graph

#mtext(paste("Outliers: ", paste(outliers, collapse=", ")))

# get the standard deviation of houses and apartments first must subset each

q1house <- subset(rentTypeQ1,dwellTypeQ1 == 1, na.rm=T)

q1apartment <- subset(rentTypeQ1,dwellTypeQ1 == 2, na.rm=T)

#then get SD

q1sdHouse <- sd(q1house,na.rm=T)

q1sdApartment <- sd(q1apartment,na.rm=T)

# sd of total rent

q1sdRent <- sd(rentTypeQ1, na.rm=T)

#getting sample size of data set

q1ApartmentLength <- length(q1apartment)

q1HouseLength <- length(q1house)

#getting means

q1HouseMean <- mean(q1house, na.rm=T)

q1ApartmentMean <- mean(q1apartment, na.rm=T)

# ks test of normality of the distribution

ks.test(rentTypeQ1,dwellTypeQ1)

# checking visually for normality

qqnorm(rentTypeQ1); qqline(rentTypeQ1)

qqnorm(rentTypeQ1)

hist(rentTypeQ1, xlab="Rent Amount", ylab="Frequency", main="Histogram of Rent")

-----------------------------------------------------------------------------------------------------------------------------

## #question3

-----------------------------------------------------------------------------------------------------------------------------

# to subset of occupantTypes in relation to dwellingType

houseQ3 <- subset(occupentType, dwellingType == 1)

apartmentQ3 <- subset(occupentType, dwellingType == 2)

otherQ3 <- subset(occupentType, dwellingType == 3)

# to obtain statistics

table1Q3 <- table(dwellingType, occupentType)

# reorganise table labels

colnames(table1Q3) = c("Family", "House Share", "Single Occ", "Other")

rownames(table1Q3) = c("House", "Apartment", "Other")

# to construct a par plot to visualise the data

plot(table1Q3, col=c("purple", "darkred", "darkorange2", "darkolivegreen"), xlab="Dwelling Type", ylab="Occupant Type", main="Percentage of Occupants in different Dwelling Types")

# to conduct the chi sq test

chisq.test(dwellingType, occupentType)

-----------------------------------------------------------------------------------------------------------------------------

#question4

----------------------------------------------------------------------------

# to subset apartments and houses

dwellQ4 <- subset(dwellingType, dwellingType < 3)

mgmtQ4 <- subset(management, dwellingType < 3)

houseQ4 <- subset(dwellQ4, dwellQ4 == 1)

apartmentQ4 <- subset(dwellQ4, dwellQ4 == 2)

# to get summary stats

table1Q4 <- table(dwellQ4, mgmtQ4)

# to alter new tables row and column names

colnames(table1Q4) = c("Managed", "Not Managed")

rownames(table1Q4) = c("Houses", "Apartments")

# to plot graph

plot(table1Q4, col=c("purple", "darkred"), xlab="Dwelling Type", ylab="Not/Managed", main="Percentage of Houses and Aparments\nManaged or Not")

# to carry out the chi sq test

chisq.test(dwellQ4, mgmtQ4)

-----------------------------------------------------------------------------------------------------------------------------

## #question5

-----------------------------------------------------------------------------------------------------------------------------

# to seperate 'others' from rent column

rentTypeQ5 <- subset(rent, occupentType != 4)

#seperate 'others' fromt occupent col

occTypeQ5 <- subset(occupentType, occupentType < 4)

# subset rent paid based on occupancy type

familyRentQ5 <- subset(rent,occupentType ==1)

houseShareRentQ5 <- subset(rent,occupentType ==2)

singleOccRentQ5 <- subset(rent,occupentType ==3)

# create var for use in box plot

names <- c(rep("Family",length(familyRentQ5)), rep("House Share", length(houseShareRentQ5)), rep("Single Occupent", length(singleOccRentQ5 )))

#create var for values for boxplot

values <- c(sample(familyRentQ5, replace=T), sample(houseShareRentQ5, replace=T), sample(singleOccRentQ5, replace=T))

# data frame for boxplot

dfQ5 <- data.frame(names,values)

# calculate proportion for width of boxplot in relation to sample size

propQ5 <- table(dfQ5$names)/nrow(dfQ5)

# draw boxplot

boxplot(dfQ5$value ~ dfQ5$names , width=propQ5, col=c("orange", "darkorchid", "dodgerblue3"), ylab="Occupant Type", xlab="Rent Amount", horizontal=T, main="Occupancy Type: Rent\n (with variable box width proportionate to sample size)")

# calculate the means

familyMeanQ5 <- mean(subset(familyRentQ5, !(is.na(familyRentQ5))))

# alternate mean

tapply(rentTypeQ5, occTypeQ5, mean, na.rm=T)

# calculate standard deviation

tapply(rentTypeQ5, occTypeQ5, na.rm=T)

# calculate sample sizes

tapply(rentTypeQ5, occTypeQ5, length)

# appying ANOVA test

summary(aov(rentTypeQ5 ~ occTypeQ5)

# organising data for Tukey test

q5Rent <- aov(rentTypeQ5 ~ as.factor(occTypeQ5))

# Tukey test

TukeyHSD(q5Rent)

# vars for Levene's test

rentSubQ5 <- subset(rentTypeQ5, !(is.na(rentTypeQ5) | (is.na(occTypeQ5))))

occSubQ5 <- subset(occTypeQ5, !(is.na(rentTypeQ5) | (is.na(occTypeQ5))))

# using Levene's test

levene.test(rentSubQ5,occSubQ5)

-----------------------------------------------------------------------------------------------------------------------------

## #question6

-----------------------------------------------------------------------------------------------------------------------------

# to plot histogram

plot(h1, col="blue")

plot(h2, col="red", add=T)

plot(h1, col=rgb(0,0,1,1/4))

plot(h2, col=rgb(1,0,0,1/4) add=T)

#to get rent difference

rentDifference <- (rent - previousRent)

# paired t test

t.test(rent,previousRent,paired=T)

# to subset the two vars with na values exluded

rentTypeQ6 <- subset(rent, !(is.na(previousRent) | (is.na(rent))))

previousRentTypeQ6 <- subset(previousRent, !(is.na(previousRent) | (is.na(rent))))

# optain the means

rentMeanQ6 <- mean(rentTypeQ6)

previousRentmean <- mean(previousRentTypeQ6)

# obtain sd

rentSdQ6 <- sd(rentTypeQ6)

previousRentSdQ6 <- sd(previousRentTypeQ6)

# sample size calculations

length(rentTypeQ6)

length(previousRentTypeQ6)

# for the boxplot of change

boxplot(rentDifference, col="darkolivegreen3", horizontal=TRUE, main="Boxplot Of Difference")

# test for normality of distribution

ks.test(rentTypeQ6,previousRentQ6)

==========================================================

#Part B

==========================================================

-----------------------------------------------------------------------------------------------------------------------------

## # Q7 Simple Regression

-----------------------------------------------------------------------------------------------------------------------------

# scatter plot for simple Regression

plot(rent,previousRent,col=c("red","blue"), xlab="Previous Rent", ylab="Current Rent", frame.plot=F,pch=c(1,2)) ; abline(lm(previousRent~rent)) ; legend("topleft",c("Rent","Previous Rent"),cex=.8,col=c("red","blue"), pch=c(1,2))

# removing na values from data

rentTypeQ7 <- subset(rent, !(is.na(previousRent) | (is.na(rent))))

previousRentTypeQ7 <- subset(previousRent, !(is.na(previousRent) | (is.na(rent))))

# correlation tests

cor(rentTypeQ7, previousRentTypeQ7)

cor(rentTypeQ7, previousRentTypeQ7)^2

# to find values for model

model = lm(rentTypeQ7~previousRentTypeQ7)

# to get stats

summary(model)

# call to scale previousRent

previousRentScaleQ7 <- scale(previousRent, center=T, scale=F)

# call to get stats based on new scaled variable

summary(lm(rentTypeQ7~previousRentScaleQ7))

# dividing new coefficient by residual standard error for prediction % error

888.254/38.18

#mean

mean(previousRent,na.rm=T)

# normality check from olsrr package

ols\_plot\_resid\_hist(model)

ols\_test\_normality(model)

-----------------------------------------------------------------------------------------------------------------------------

## # Q8 Mulitple Regression

-----------------------------------------------------------------------------------------------------------------------------

# to visualise rent and population

plot(rent,townSize,col=c("red","darkolivegreen3"), xlab="Current Rent", ylab="Population(Thousands)", frame.plot=F,pch=c(1,3)) ; abline(lm(townSize~rent)) ; legend("topleft",c("Rent","Population"),cex=.8,col=c("red","darkolivegreen3"), pch=c(1,3))

# stats on for our model

Model1Q8 <- summary(lm(rent~ townSize))

summary(lm(rent~ management))

# normality check from olsrr package

ols\_plot\_resid\_qq(model1Q8)

ols\_test\_normality(model1Q8)

# to remove na

popQ8 <- subset(townSize, !(is.na(townSize) | (is.na(rent))))

rentTypeQ8 <- subset(rent, !(is.na(townSize) | (is.na(rent))))

# cor test

cor(rentTypeQ8, popQ8)

# get means

mean(townSize, na.rm=T)

mean(rent, na.rm=T)

# getting highest value in pop for extrpolated value model test

max(townSize, na.rm=T)

# using multiple summaries of lm vars to quickly gage the best variables to start buildinf the model with

summary(lm(rent~ dwellingType))

summary(lm(rent~ occupantType))

summary(lm(rent~ management))

# adding variables together for our model

summary(lm(rent~ occupentType + townSize))

summary(lm(rent~ occupentType + townSize + dwellingType))

-----------------------------------------------------------------------------------------------------------------------------