Final project documentation

Nadiia Kuzmenko (2202854)

Berlin International University Of Applied Sciences

Analytics lab 1

February 11, 2024

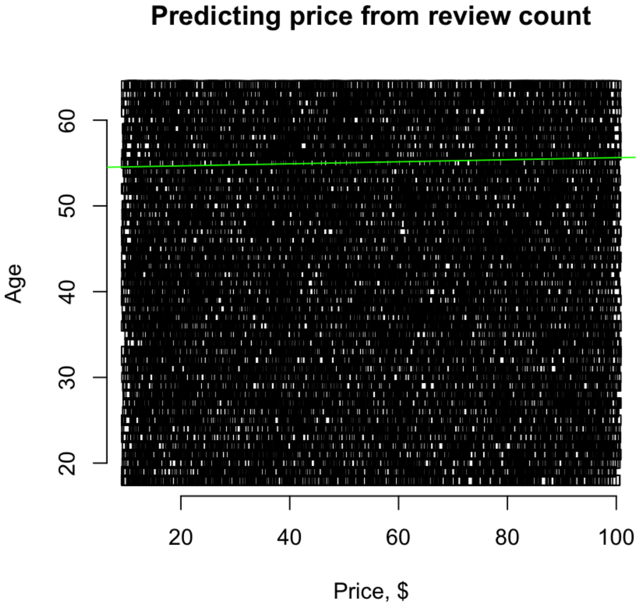
During my final analytics lab 1 assignment I worked with a data set containing information about the retail industry in the UK and the US. The data set is designed to mimic the real sales process in the fashion industry. It offers a realistic representation of sales transactions, customer preferences, and product characteristics.

Part 0: Preparation

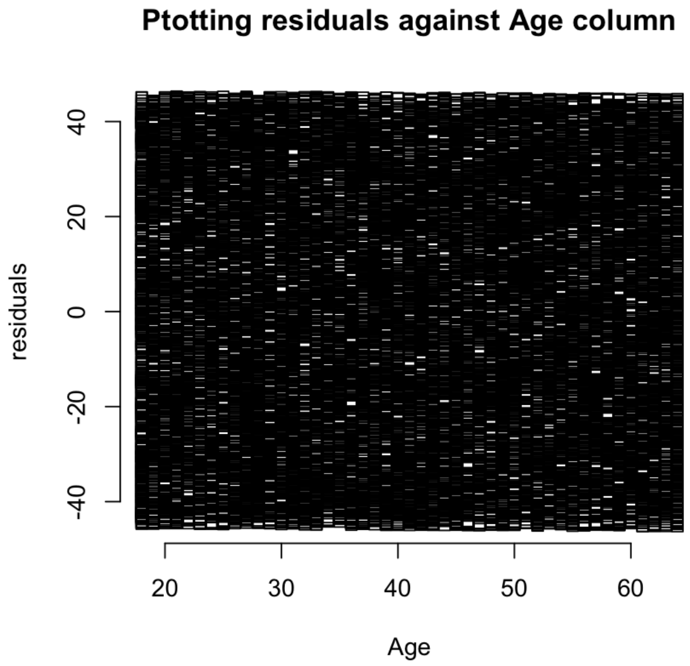
1. I started by uploading the data set into R Studios, omitting the NA values and sampling 20000 values, as the data set was too heavy for the type of research I aimed for.
2. I changed the column “Review Count” to be numeric and renamed it for easier use.
3. I explored the basic information about the data set, discovering the data set has 20 columns. After checking the class of each column I discovered that 4 of them are numeric: “Price”, “Rating”, “Age” and “Review_Count”. I am going to focus on those columns for my model.
4. I proceeded to calculate the means and standard deviations of each numerical column.

Part 1: Linear regression

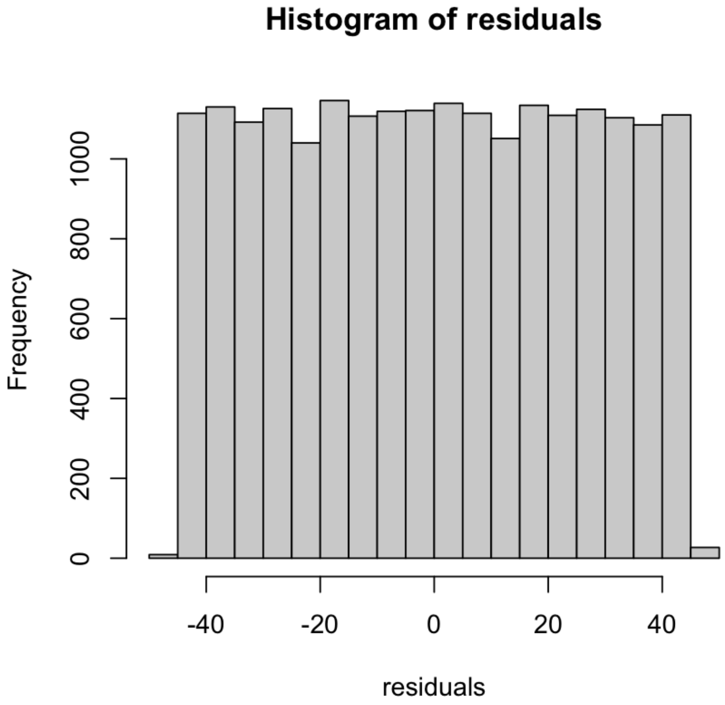
1. I picked “Price” as my target column for a linear regression model because it was interesting for me to learn how other features influence how expensive the item is.
2. Computed the correlation between my target column and all other numerical columns, and found out that column “Age” has the strongest correlation (0,006). It suggests a very weak positive linear relationship between the two variables. Nevertheless, I am going to try to build a linear model with those variables.
3. As I predicted, the linear model showed a very weak linear relationship between the Price of the item and its Age.
4. Made a scarred plot, plotting Price column against Age column. There is no pattern and, therefore no outliers.
5. Added a line predicted by the linear model. It intercepts the y-axis at 54.42, which is very close to the mean of price (54.92).



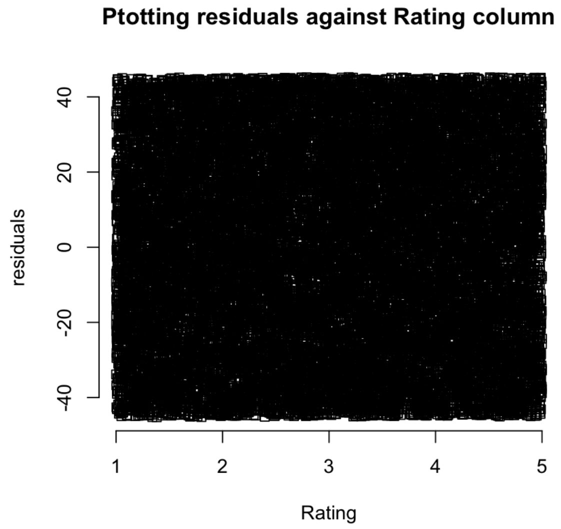
6. The reported R-squared value of 3.948e-05 indicates that the independent variables in the model explain only a very small fraction of the variance in the dependent variable.
7. Plot the residuals against the Age. No pattern is visible.

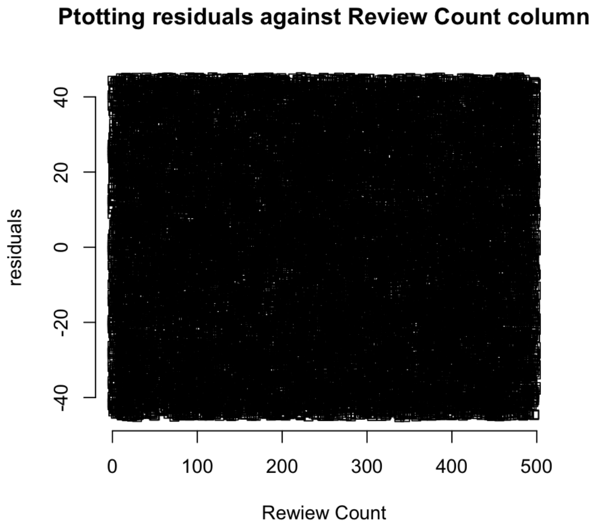


8. Plot the histogram of residuals. The distribution is uniform, there is no systematic pattern in the residuals as the Age changes.



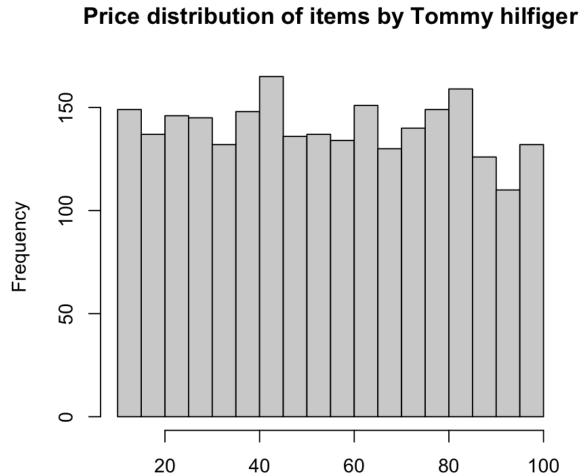
9. Included the other two columns in a multivariate linear regression. Made a new model, where Price is explained by Age, Review Count and Rating
10. R-squared = 4.818e-05, which is bigger than before, but still very small, it did not improve much.
11. Plotted review count and rating against residuals. No pattern was found.

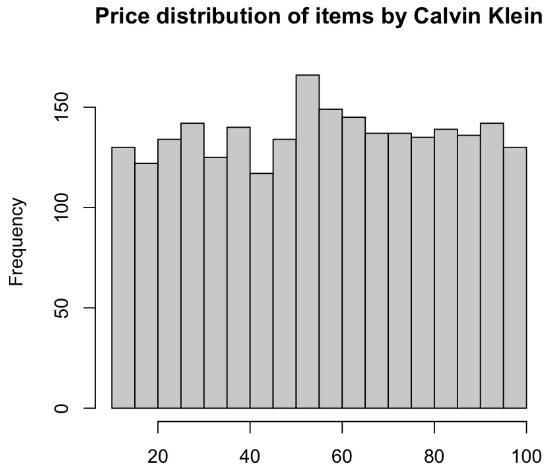

Ptotting residuals against Rating column


Ptotting residuals against Review Count column

Part 2: t-test

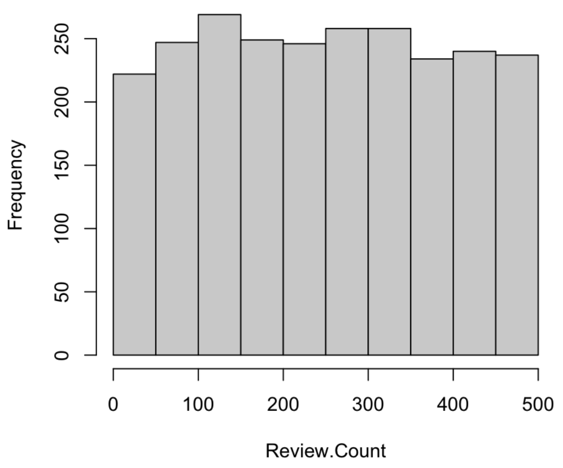
1. To decide on 2 groups of at least 100 data points, I chose 2 brands and checked if there were enough items produced by each of them.
2. Subselected the data frame into those two distinct groups: “hilfiger” - containing info about clothing items produced by Tommy Hilfiger; “calvin_klein” - containing clothing info about items produced by Calvin Klein.
3. Checked the standard deviations, medians, and means of all the numerical columns for those two groups.
4. Formulated a null hypothesis regarding the price distribution of Calvin Klein and Tommy Hilfiger items.
“H0: There is no difference in the price distribution between Brand A and Brand B”
5. To check if the data is normally distributed, I plotted histograms for the Price distribution in 2 selected brands.
6. As we can see, data is not normally distributed but rather has uniform distribution. Still, I am going to attempt to do a two-sided t-test to test my hypothesis.


Price distribution of items by Tommy hilfiger

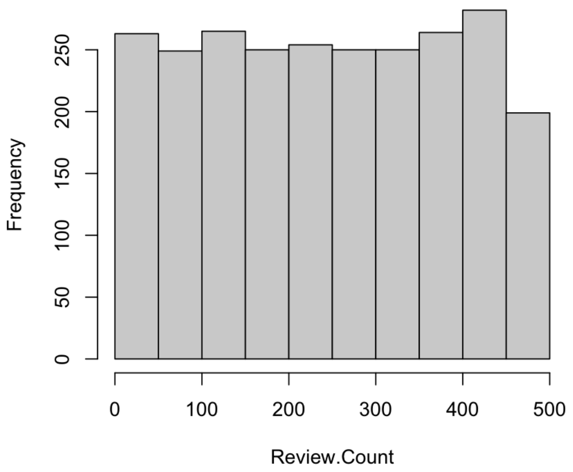

Price distribution of items by Calvin Klein

7. Before doing a t-test also check the variances of both groups. They are approximately equal.
8. To test the H0 I am using a t-test. It is useful to determine if there is a significant difference in means of prices, therefore, if the average price is significantly different.
9. The estimated means of the two groups are close to each other (mean of x = 54.13922; mean of y = 55.47728).
10. p-value = 0.06638, which is bigger than the significance level (0.05). We can conclude, that there is a significant difference in the average price for those brands' items. Therefore, I reject the null hypothesis.
11. Formulated a null hypothesis regarding the review count of Calvin Klein and Tommy Hilfiger items.
    "H0: There is no difference in the review count between Brand A and Brand B"
12. To check if the data is normally distributed, I plotted histograms for the review count distribution in 2 selected brands.
13. As we can see, data is not normally distributed but rather has a uniform distribution. Still, I am going to attempt to do a two-sided t-test to test my hypothesis.

Review.Count distribution of items by Calvin Klein



Review.Count distribution of items by Tommy hilfige



14. Before doing a t-test also check the variances of both groups. They are approximately equal.
15. To test the H0 I am using a t-test. It is useful to determine if there is a significant difference in means of prices, therefore, if the average price is significantly different.
16. The estimated means of the two groups are close to each other (mean of x = 246.8226; mean of y = 250.0252).
17. p-value =0.4278, which is bigger than the significance level (0.05). We can conclude, that there is a significant difference in the average review for those brands' items. Therefore, I reject the null hypothesis.
18. The results of the t-tests performed could not be influential, as the groups of data I examined do not have a normal distribution.