LUT School of Business and Management

A220A0010 Free Analytics Environment R

Christoph Lohrmann

Second Assignment

Autumn 2021

TABLE OF CONTENTS

1. Part 1 (Regression analysis)

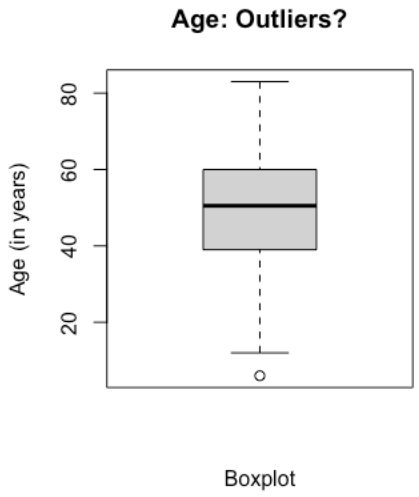
Working as a data scientist for a medical research laboratory, author has received a dataset consisting of historical patient information. Focusing on the glucose level as the dependent variable, a machine learning approach is employed to determine which factors are influencing blood glucose levels in order to help doctors and patients adjust their recommendations.

Concretely, the dataset is evaluated from different perspectives and a regression model is built to predict the blood glucose level of patients with the help of given variables.

1.1. Evaluating the dataset and conducting exploratory data analysis

This section contains steps 1 and 2 from the assignment part 1 (Regression analysis).

When imported, loaded and inspected, the dataset is found to contain 156 observations where each observation is described by 16 variables or features. Data can be considered of high quality as there are no missing values contained in any of the observations (being NA or null). This means that no treatment of missing values by imputation or other means is necessary. Variables **Age**, **Blood_Pressure**, **Sugar**, **Diabates**, as well as the dependent variable **Glucose** are selected for closer inspection.



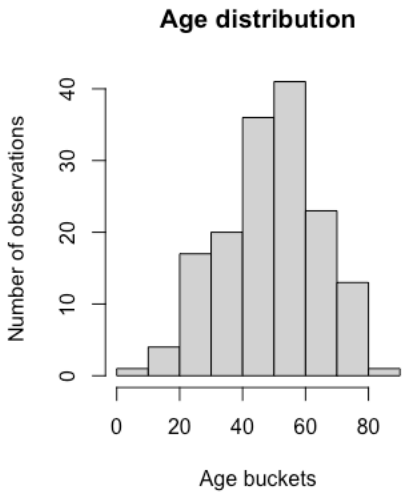


Figure 1. Exploratory data analysis, variable: Age.

Variable **Age** which is considered in the context of this report as continuous variable measured in years, contains a minimum age of 6 which is a relative outlier to the rest of observations as shown in Figure 1. The maximum observation is identified to be 83. The average age of all observations in the dataset is 49,44 years with the median being 50,5. Because of the insignificant impact of the outlier on the mean, author considers that deleting the observation from the dataset is currently not needed. Based on the histogram displayed in Figure 1, the distribution of ages takes somewhat of an evenly distributed "bell" form, showing that the great majority of observations are between the ages of 20 and 80.

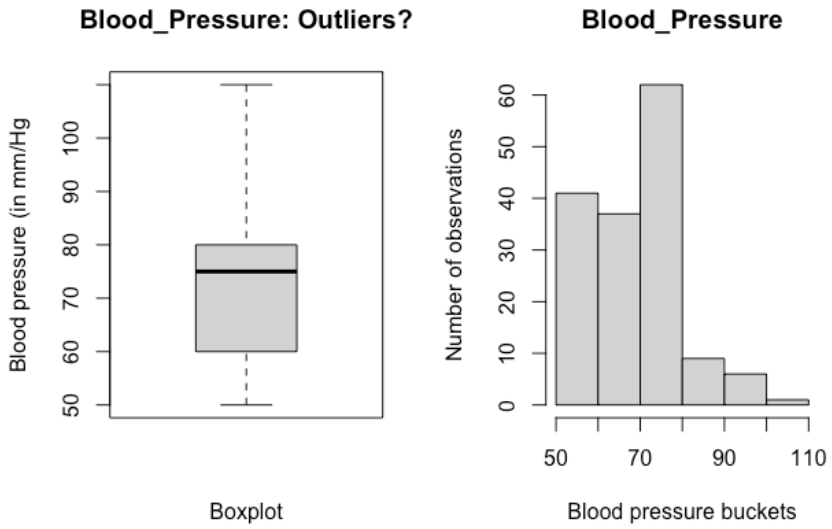


Figure 2. Exploratory data analysis, variable: Blood_Pressure.

Variable **Blood_Pressure** measured in mm/Hg is found to have a minimum observation of 50 and a maximum of 110, with the mean blood pressure for all observations lying at approximately 73,85. The distribution is skewed to the right, but does not contain any significant outliers.

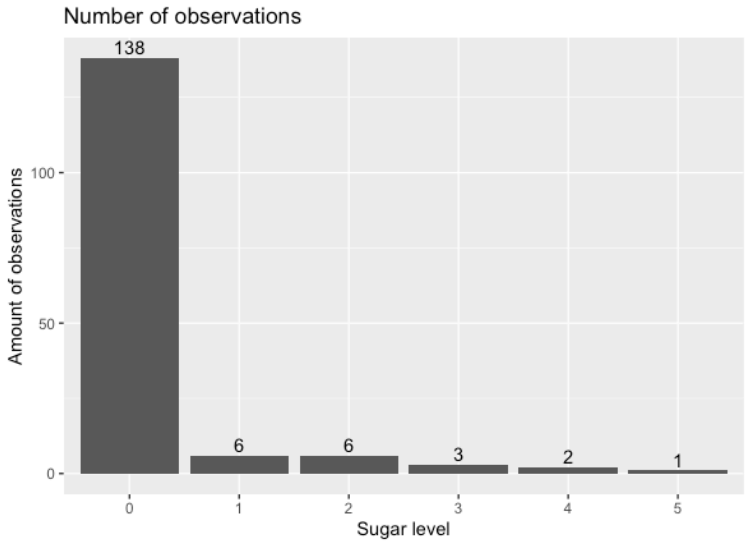


Figure 3. Exploratory data analysis, variable: Sugar.

Categorical (Ordinal) variable **Sugar** is measured on a scale 0-5. Vast majority of observations are categorized to be zero, while just a single observation is categorized to be 5. Only 17 observations are found to be somewhere between the maximum and minimum values.

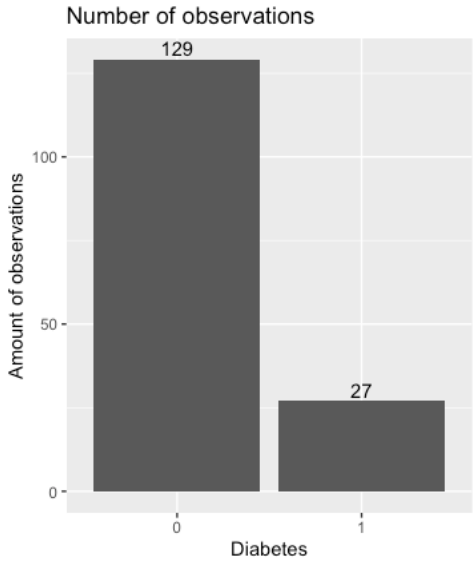


Figure 4. Exploratory data analysis, variable: Diabetes.

Diabetes consists of two categories 0 and 1, which the author assumes to be an equivalent to a Boolean meaning yes or no (nominal). Majority of observations are categorized as “No” and only 27 observations as “Yes” equating to a proportion of 17,3% out of the total 156 observations in the dataset.

Figure 5. Exploratory data analysis, variable: Glucose.

Variable **Glucose**, which in this context is treated as dependent and continuous, has a minimum observation of 70 and a maximum observation of 490. It is measured in mgs/dl. The mean of all observations stands at approximately 131,42 while the median is at 113,5. The fairly significant difference between mean and the median can be attributed to a considerable amount of outliers that skew the distribution to the right as can be seen in Figure 5.

1.2. Correlation analysis and pre-model preparations

This section contains steps 3, 4 and 5 from the assignment part 1 (Regression analysis).

Variables **Diabetes** (0,67) and **Sugar** (0,73) carry the highest positive correlations with **Glucose**. Variable **Gravity** (-0,56), on the other hand, carries the highest negative correlation with **Glucose** out of all variables in the dataset as demonstrated below in Figure 6.

It is widely accepted that an absolute correlation of approximately 0,7 to 0,8 and higher is considered a relatively strong correlation. **Hemoglobin** is strongly correlated with **Cell_Volume** (0,86) and **Red_BloodCount** (0,78), while being less strongly inversely correlated with variables **Diabetes** (-0,66) and **Anemia** (-0,65).

The pair **Hemoglobin** and **Cell_Volume** can thus be considered to be strongly correlated. Additionally, the pair **Hemoglobin** and **Red_BloodCount** are strongly correlated with each other, albeit less so at 0,78. It is worth noting that **Gravity** seems to come close to strongly correlating with both **Hemoglobin** and **Cell_Volume**, each having correlation values of 0,68 respectively and **Red_BloodCount** having a value of 0,66, while at the same time somewhat negatively correlating with **Diabetes** (-0.63) and **Pedal_Edema** (-0.62).

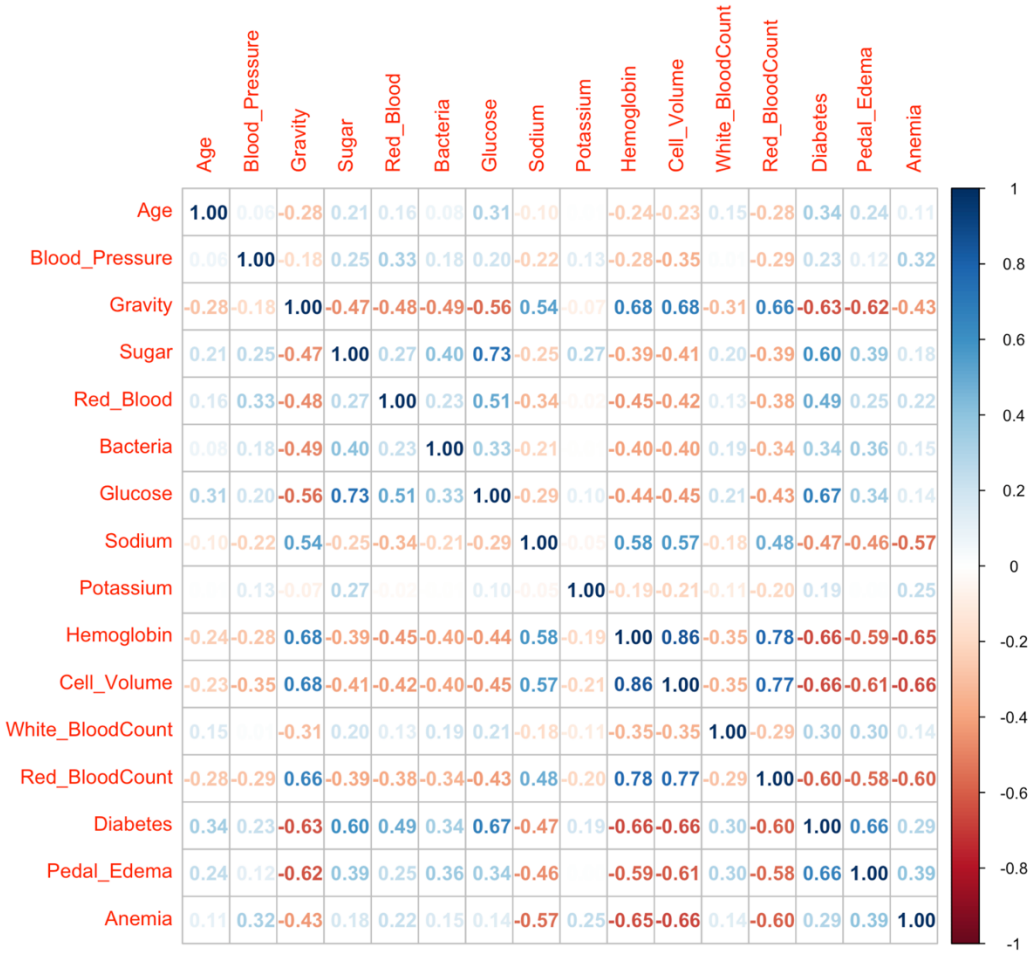


Figure 6. Correlations of all variables visualized in a single plot.

Author suspects that variables which are highly positively correlated with **Glucose**, can be attributed to “unhealthy” type of indicators in any person’s health (**Diabetes**, **Sugar**). Meanwhile correlating attributes such as **Hemoglobin**, **Cell_Volume** and **Red_BloodCount** can be

considered as attributes that are together signaling of good overall health of the person in question.

Only one variable with a high absolute correlation of 0,8 or more was removed at this step of the process, and it was **Cell_Volume** standing at 0,86 correlation with **Hemoglobin**. It was chosen to be deleted because it had the highest average correlation to all other variables among the set of independent variables, albeit only slightly so compared to **Hemoglobin** which was kept in the dataset (0,4797614 compared to 0,4729466).

If not removed, highly correlated explanatory variables can result in a less stable model with erratic behavior, where the coefficient estimates can change unpredictably just with small changes in the model or data. The phenomenon that author is trying to avoid by excluding **Cell_Volume** is called “Multicollinearity” or “Collinearity”, meaning that one variable can be well predicted by another one with high accuracy.

As the purpose of regression is to learn the weights from the training data and use these weights for prediction purposes, if there are mutually correlated explanatory variables, the model may not perform well on actual test data.

After removing **Cell_Volume**, author proceeded to implement the linear regression model.

1.3. Implementing and improving linear regression

This section contains steps 6, 7 and 8 from the assignment part 1 (Regression analysis).

After initially implementing linear regression to estimate the coefficients of the model, residuals were found to be distributed summetrically as the median was found to be close to zero while both of the quartiles were close to each other in magnitude. Likewise, both Min and Max had mutually similar magnitudes in terms of residuals implying an overall symmetrical distribution. From the get-go, some particularly small p-values were observed for variables

Sugar and **Diabetes**. Additionally at least **Red_Blood** and **Pedal_Edema** seemed to be impacting **Glucose** as the dependent variables. Additionally, it is worth noting separately that the coefficient of **Gravity** had a very high negative value, meaning that a unit change in this feature has a large expected inverse change in the response variable which the model is attempting to predict.

The R-squared value measuring the fit of the model stood originally at 0,7118 (implying a better fit than only taking the mean of the response variable alone), while adjusted R-squared was 0,6832 describing some possible noise in the data. Author expected the adjusted R-squared value to slightly improve were the model to be simplified while at the same time trying to decrease the R-squared value by only as little as possible. The F-statisic was originally 24,88. The p-value of the F-statistic already from the first time after model creation indicated that the model's relationship to the response variable was highly significantly better than just that of an intercept only model.

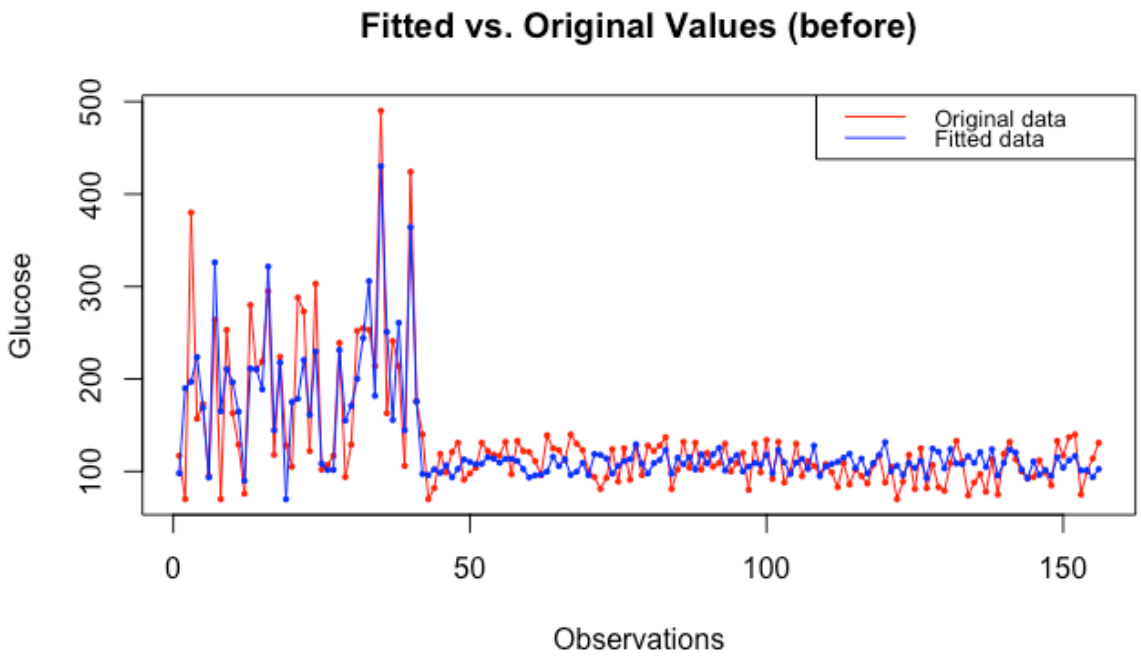


Figure 7. Fitted vs. Original Values before simplifying the model.

Simplification of the model started by the careful removal of variables one after another while observing the changes in abovementioned indicators. A drop of a variable was justified by a high respective p-value. Thus, based on a p-value of above 0,1 the variables that eventually got removed were **Sodium**, **White_BloodCount**, **Hemoglobin**, **Red_BloodCount**, **Bacteria**, **Anemia** and **Age**.

After improvements to the model, author was now able to ensure with a confidence of 95% that variables selected for the model actually did have an impact on the response variable, thus decreasing the noise and resulting in a more optimal model as can be seen in Figure 8.

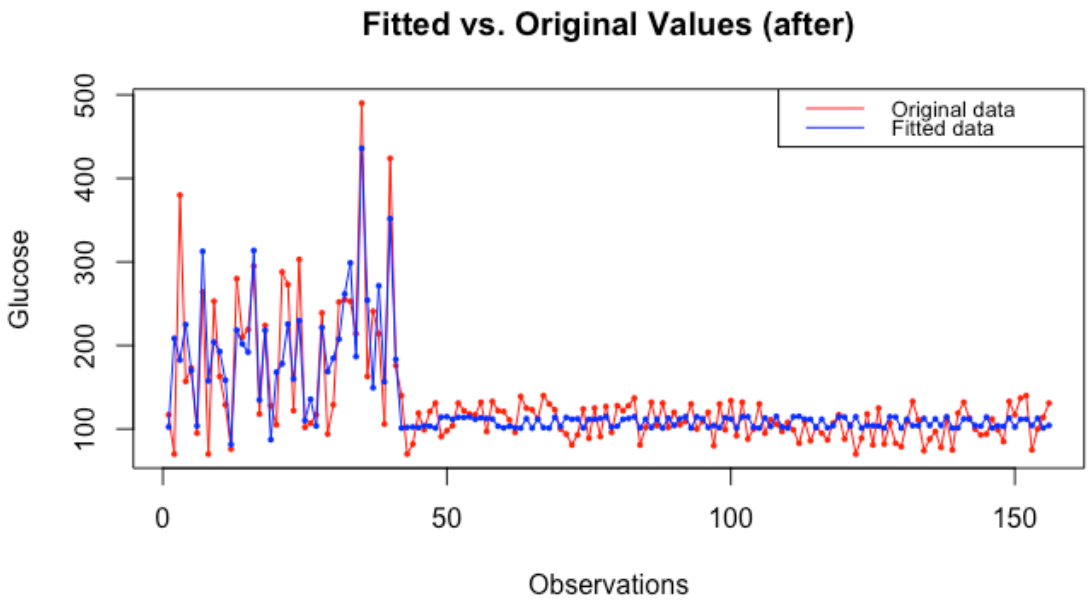


Figure 8. Fitted vs. Original Values after simplifying the model to six variables plus intercept.

The newly improved F-statistic was observed to be now 57,75 (up from 24,88), with an adjusted R-squared value of 0,6872 (up from 0,6832), while the R-squared itself was decrased by just 0,0125 down to 0,6993 from original 0,7118, continuing to indicate a good overall fit for the model. The final regression equation can be seen below.

$$\text{Glucose} = \theta_0 + \theta_1 * \text{Gravity} + \theta_2 * \text{Sugar} + \theta_3 * \text{Red\_Blood} + \theta_4 * \text{Potassium} + \theta_5 * \text{Diabetes} + \theta_6 * \text{Pedal\_Edema} + \theta_7$$

Where **Glucose** is the dependent variable to be predicted, θ_0 is the intercept at 2256,1087 and θ_1 is the "slope" or "impact" of the independent variable **Gravity** at -2092,3532 in the final model. The slopes of other variables in the final model are 41.3851 (for θ_2, **Sugar**), 35.9141 (for θ_3, **Red_Blood**), -2.0572 (for θ_4, **Potassium**), 58.1930 (for θ_6, **Diabetes**), -45.6918 (for θ_7, **Pedal_Edema**).

1.4. Evaluating the resulting model

This section contains step 9 from the assignment part 1 (Regression analysis).

Author proceeded to evaluate the model according to five properties of linear regression using OLS in regards to the residuals that are important for assessing whether the results are reliable.

First of all, the mean of the residuals in the model is 4,355935e-16, which is very close to zero. This means there is no systematic error in the predictions in this regard. The residuals or "error terms" are assumed to be random, and author's finding above supports the assumption for the validity of the model.

The variance of the residuals should be constant (and finite), or as often talked about in the field of statistics, "Homoskedastic". A formal approach would be the Breusch-Pagan test, but in this case a visual representation can be observed in the Figure 9 below. The residuals should be linearly independent of one another and this could be formally tested with the Durbin-Watson test for autocorrelation. Visually there does seem to be a pattern where the variance is wider in the beginning and less so over time which would suggest of a possibility for an existing systematic error in the model.

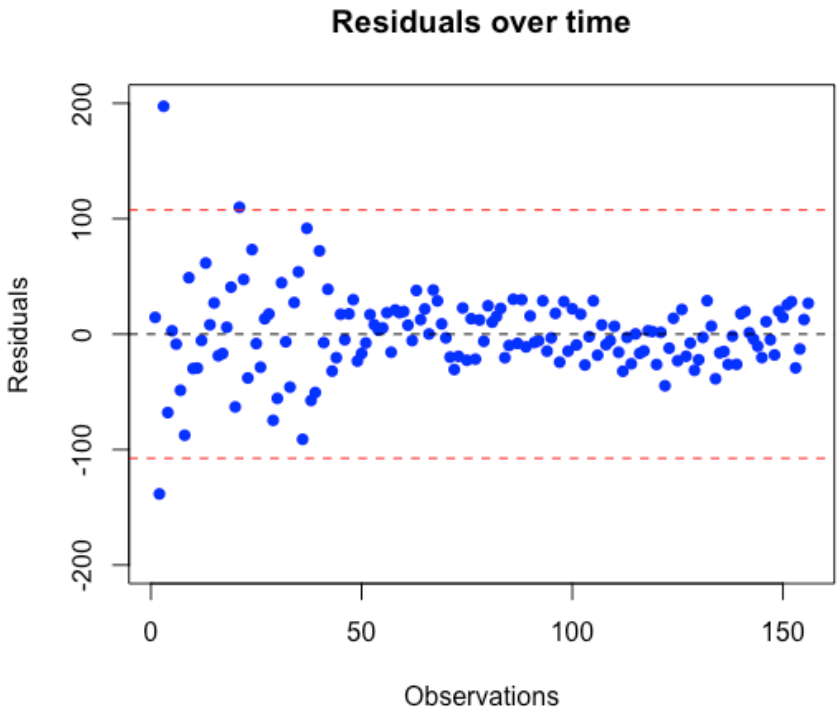


Figure 9. Variance of the residuals in the final model

In a linear regression models using OLS, the residuals should not be correlated in any way or be dependent with explanatory variables. According to the observation by the author, the correlations are very close to zero implying a non-existent correlation, meaning that this indeed is the case with the final model.

Finally, the residuals should be normally distributed in an OLS regression model. This was checked formally by the author using the Jarque-Bera test for normality of the distribution. Jarque-Bera test uses two “moments” of a distribution called Skewness and Kurtosis. In the case of a normal distribution, both Skewness and excess Kurtosis should be close to zero, so the test overall should be also close to zero. When evaluated against the model, the aggregate test result with X-squared of 309,95 and a very small p-value (close to zero) did however indicate that the residuals are not normally distributed in the model implying some systematic error wich results in a less reliable model by the author. The statistic for Skewness was 0,77486 and Kurtosis 9,7293.

1.5. Findings and conclusion

This section contains the 10th and last step from the assignment part 1 (Regression analysis).

It seems that for the purpose of predicting the glucose level of a patient, residuals in the author's model are somewhat "heteroskedatstic", meaning that the model does somewhat systematically deviate from a normal distribution. Overall this means that the final model is not very suitable to fit the data as-is, because at least one of the five properties for the residuals in linear regression using OLS are not being met.

This could mean that author's final model still contains a wider range of values than is needed, making the model more prone to heteroskedasticity. This implies that differences between the smallest and largest values in the dataset chosen for the model can be very significant, resulting in a decreased reliability of the model. Perhaps the variables chosen are too "cross-sectional" which often results in Heteroskedasticity when conducting regression modelling.

Possible remedies to address Heteroskedasticity more generally could be:

- Transforming the dependent variables
- Redefining the dependent variables
- Using weighted regression

In order to improve the model to be more usable by doctors and patients in real-world scenarios for their recommendations, author would like to receive more observations for further studying of the issue of Heteroskedasticity in the currently generated or future models.

Nonetheless, author could already start giving predictions of the Glucose level of patients by the means of employing the generated model and the "predict" function available in base R.

An example observation fed into the final model by the author in RStudio is as follows: Gravity (1.010), Sugar (1), Red_Blood (1), Potassium (3,4), Diabetes (1) and Pedal_Edema (1)

The above observation results in an exact prediction of the following Glucose level for the patient: 225,6378. The prediction given by the model has a 95% confidence of being within the lower 203,8617 and upper 247,4139 bounds of the probable true glucose level for this particular patient.

Above conclusions could already help doctors and patients with predicting glucose levels in a fairly reliable manner in situations where there are no other methods available. Again, improvements in the underlying model would be necessary for more accurate predictions going forward. At this stage author concludes the recommendations for this task.

2. Part 2 (Clustering)

As a data scientist employed by an e-commerce platform business, author has received a dataset containing information on customers and how they have used the website. By exploring the dataset, the goal is to find similar types of users based on their characteristics, e.g. "user behavior" and to understand what defines these groups while feeding recommendations back to the business. Each observation or row in the dataset represents a "session" of a unique customer on the website.

2.1. Evaluating the dataset and conducting exploratory data analysis

This section contains the steps 1, 2 and 3 from the assignment part 2 (clustering).

When loaded, the dataset is found to contain 12 330 observations consisting of 8 variables. The data does not seem to contain any missing or null values. This means that no treatment of missing variables by imputation or other means is needed in the beginning stages of the process.

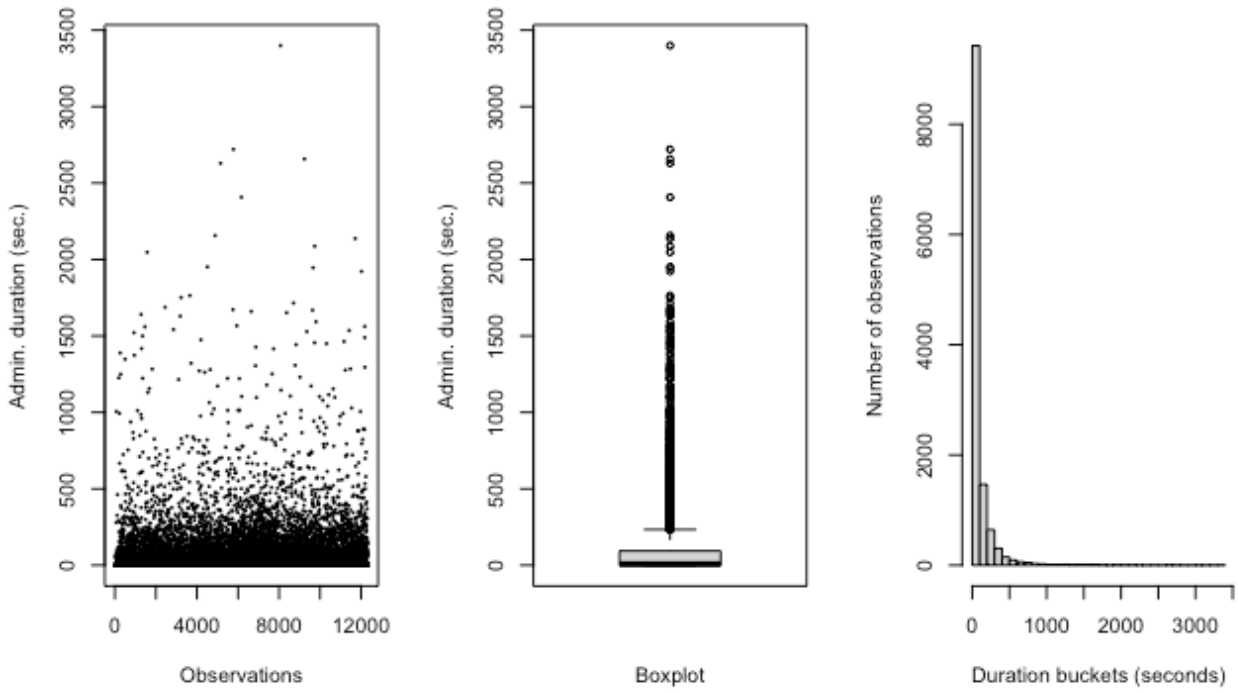


Figure 10. Exploratory data analysis, distributions for variable: Administrative_Duration.

After visualizing variables **Administrative_Duration**, **Informational_Duration** and **ProductRelated_Duration**, author found them to be distributed in mutually similar manners leading to a decision to only include an example distribution visualization of **Administrative_Duration** in this report as can be seen in Figure 10. Meanwhile, some of the basic descriptive statistics for these variables can be observerd below in Table 1.

Table 1. Descriptive statistics of the three first variables in the dataset.

	Administrative_Duration	Informational_Duration	ProductRelated_Duration
Min.	0,00	0,00	0,00
Max.	3 398,75	2 549,38	63 973,50
Mean	80,82	34,47	1 194,80
Med.	7,50	0,00	598,90

Regarding the variable **BounceRates**, author found them to be distributed between 0,00 and 0,20 meaning that the maximum value for percentage of visitors who entered the website

from the same webpage as the current user and then left is capped to 20%, leading to a somewhat problematic distribution that can be observed below in Figure 10.

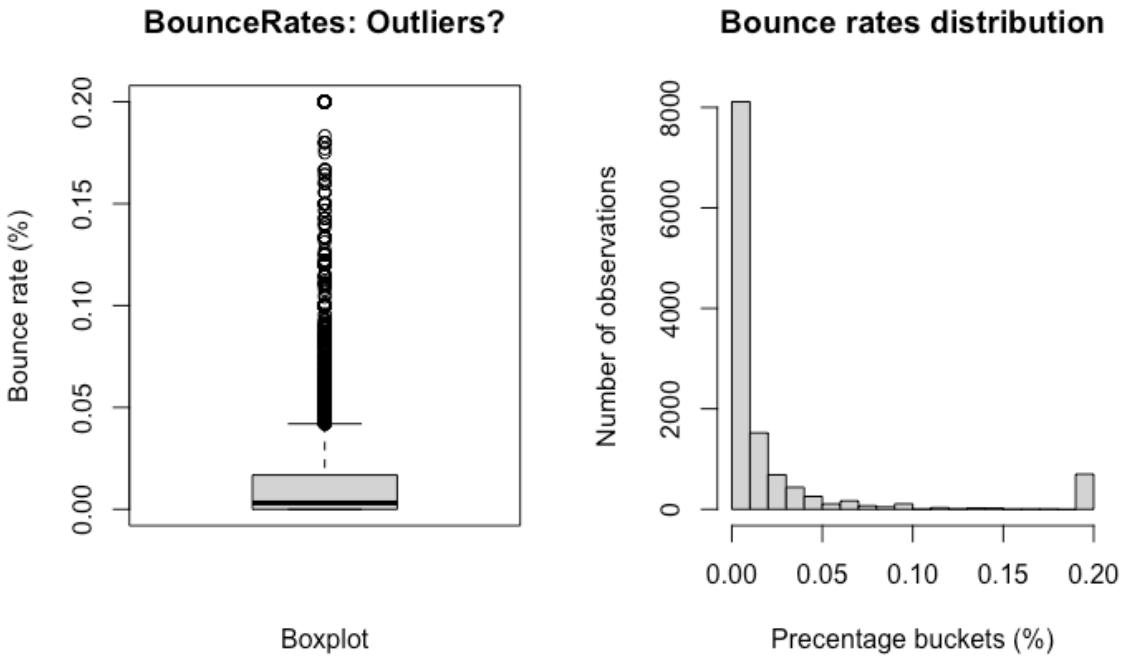


Figure 10. Exploratory data analysis, distributions for variable: BounceRates.

According to author's knowledge, it is a known weakness that k-means model in general is sensitive to outliers, often resulting in a less reliable model. Nonetheless, a decision was made to proceed without any further treatment of this variable at this stage.

Variable **SpecialDay** descirbes the closeness of the time at which user visits the website to a special holiday, like Christmas for example. There being a total of 12 330 observations in the dataset, a great majority with 11 079 observations, or 89,85% of sessions were not even close to a special day. Only approximately 8,9% observations from the entire dataset were close, e.g. somewhere between 1-4 days from a special day, and just 1,25% of all observations were happening on a special day itself.

When looking into **NewVisitor**, author knew that all sessions were classified whether the visitor was coming to the website for the first time or not. In the dataset there is a total of 1 649 observations (15,93%) classified as visiting for the first time. The proportion can be seen

visualized below in Figure 11. Based on this variable, author concluded that many more sessions come from recurring platform visitors rather than new ones.

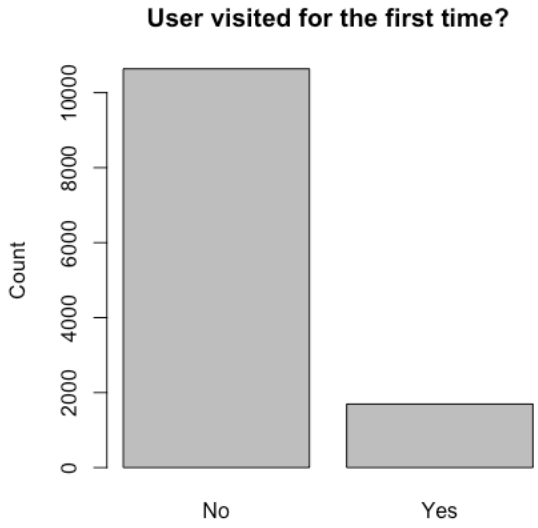


Figure 11. Exploratory data analysis, propotions of variable: NewVisitor.

Additionally, it is worth mentioning that there is a mystery variable **PageValues** included in the dataset that author was not made aware of. A distribution of this variable can be seen below in Figure 12. Nonetheless, author has decided to remove this mystery variable out of the dataset to not interfere with further analysis.

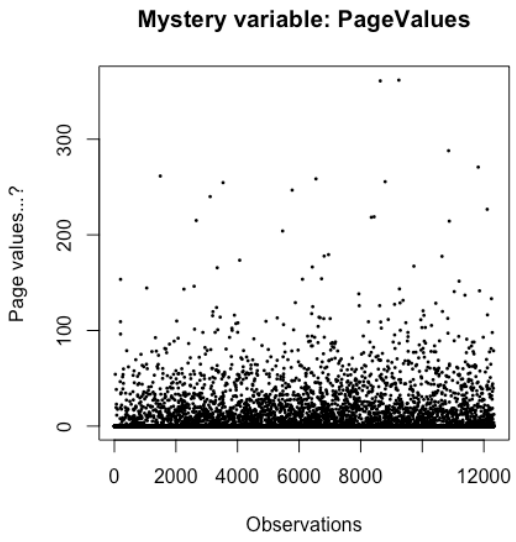


Figure 12. A distribution visualization of the mystery variable before deletion: PageValues.

Variable **Revenue** describes whether a session in question ended with a commercial transaction, e.g. a purchase on the platform. A good amount of 1 908 sessions ended in a purchase, which is a proportion of 18,3% of all observations in the dataset. A visualization of this proportion can be found below in Figure 13.

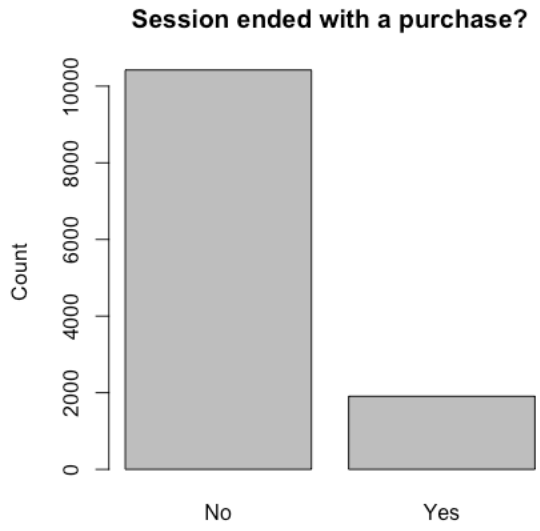


Figure 13. Exploratory data analysis, propotions of variable: Revenue.

According to author's correlation analysis, no strong correlations were found within the dataset at this stage. Only between the values of time spent on various parts of the platform were associated in a minimal manner with each other, but this was not considered to be in any way significant. The correlations can be seen in a correlation plot displayed below in Figure 14.

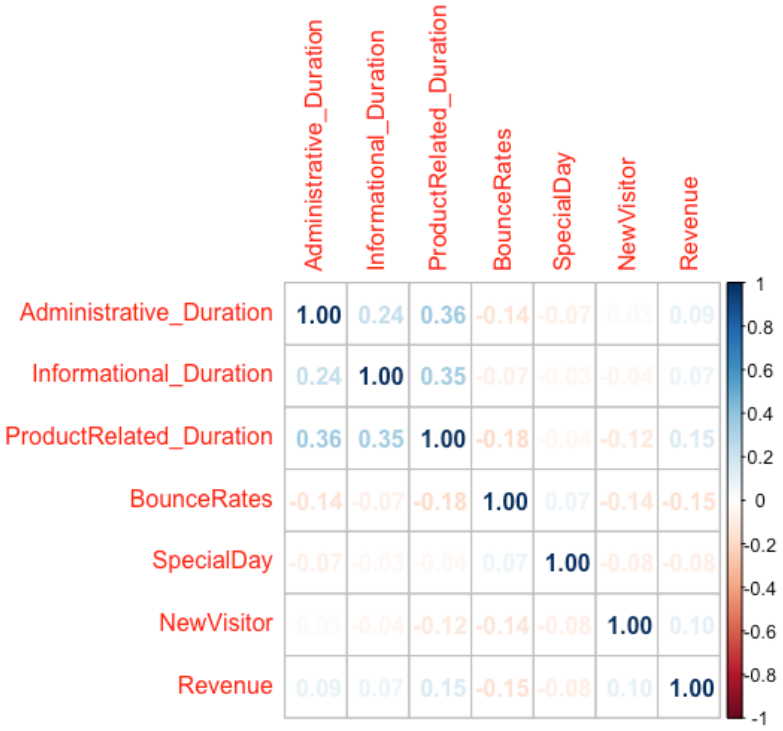


Figure 14. Correlation analysis of all variables after PageValues was deleted

2.2. Data preprocessing

This section contains the step 4 from the assignment part 2 (clustering).

In some algorithms like clustering, variables that are measured on differing scales will end up contributing unequally to the model, rendering the results unreliable. This is because many ML algorithms try to find trends in the data by comparing the datapoints of features. Features that have larger scales, like the duration variables (**Administrative_Duration**, etc.) in author's case, end up dominating the other ones which is an unwanted behavior. Normalization or so called "feature scaling" should be conducted each time when an algorithm is used that is sensitive to differing scales in the underlying data and when the data has vastly different values in different features of the dataset.

After removing variable **Revenue** from the dataset, the three remaining duration variables, as well as **BounceRates**, **SpecialDay** and **NewVisitor** were all normalized by author in accordance with the min-max method. This means that all values were scaled to be in the range between 0 and 1 to avoid disproportionate impact of high versus low values during the cluster analysis later on in the process. Author was now ready to proceed with the analysis.

2.3. Performing cluster analysis

This section contains steps 5 and 6 from the assignment part 2 (clustering).

Author directly ran the first cluster analysis with the k-means algorithm by arbitrarily deciding to have a go with a random amount of 3 clusters (which was later found out to be a good guess), and an "nstart" of 5. The initial model (**first_model**) in this case ended up with clusters of 1 693, 1 642 and 8 985 observations each, as well as respective cluster sum of squares per cluster (dispersions) of 56,98, 540,28 and 153, 07. The total within cluster sum of squares for the model was thus 750,34. Having these initial results, author started looking for an actual optimal amount of clusters with the help of multiple methods.

According to the "Elbow method" alone, an optimal amount of clusters in the model according to author's opinion would have been 4 as can be observed below in Figure 15.

The Elbow method measures an optimal amount of clusters by checking for the difference in dispersions towards the mean of each cluster at varying cluster amounts. According to this method, when the "improvement" in the dispersion (within sum of squares) of each cluster starts to decrease with each additional cluster added, the optimal amount of clusters can be inferred from when the decreases in dispersion start to become small versus when they are large with each new cluster added, e.g. when the "gain" of adding each new cluster decreases significantly.

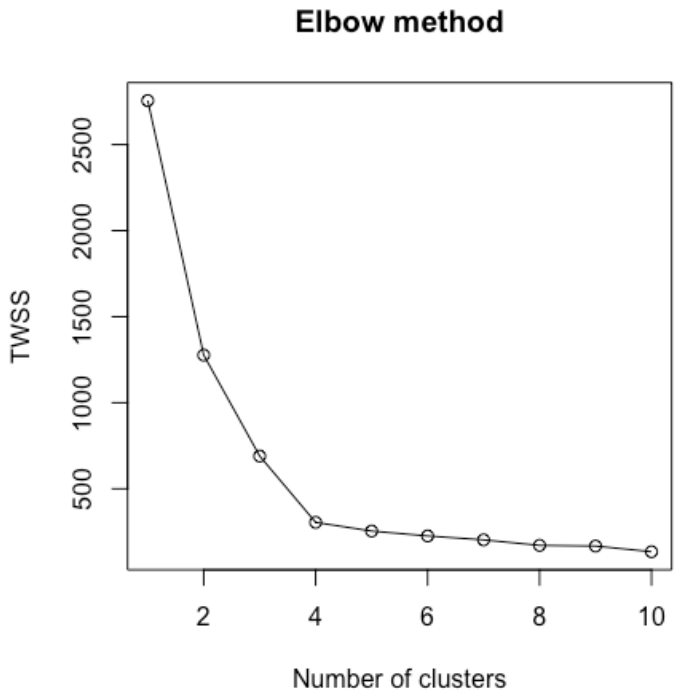


Figure 15. Determining an optimal amount of clusters with the Elbow method.

However, the Elbow method is not the only approach that can be employed in finding the optimal amount of clusters for a model, and thus the use of additional sophisticated methods is warranted. Such methods employed by the author in the context of this analysis are additionally the Silhouette method, the GAP statistic and the Calinski-Harabasz Index as can be seen visualized below in Figure 16.

The Calinski-Harabasz index which is also known as a "Variance Ratio Criterion" measures the ratio between the Between-Group-Sum-of-Squares (BGSS) and Within-Group-Sum-of-Squares (WGSS). It aims to look for dense and well separated clusters and is calculated by dividing the variance of sum of squares and the distances of objects to their respective cluster centers, as well as the distances between the cluster centers among each other. The higher this index, the better is the clustering model.

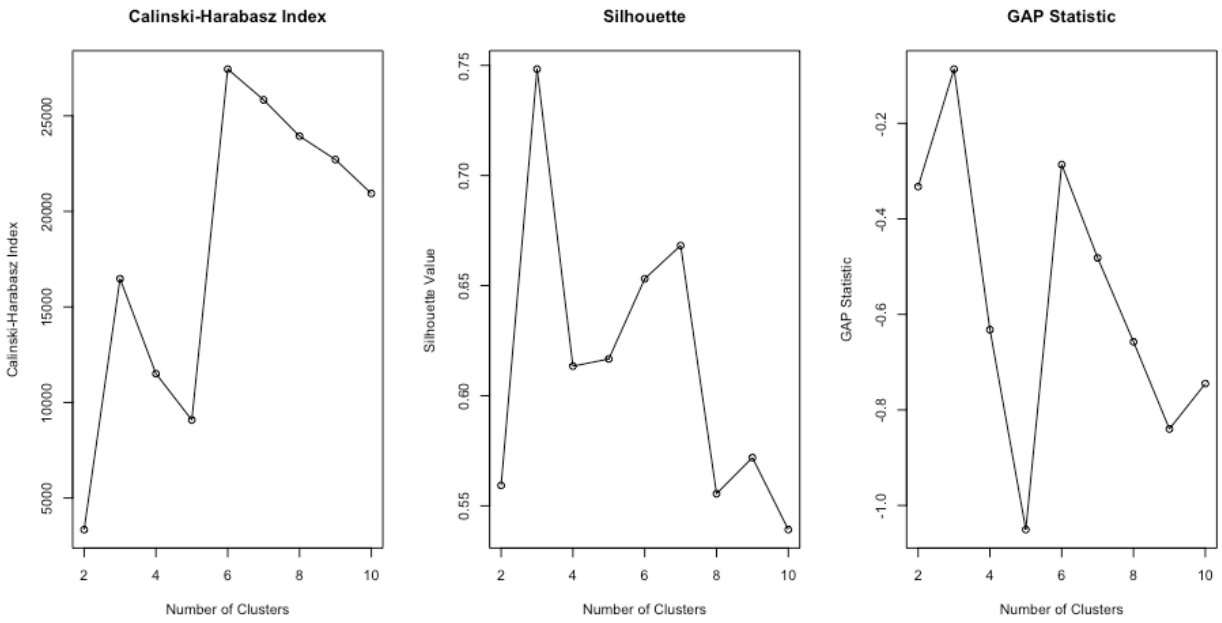


Figure 16. Determining the optimal cluster count, Silhouette, GAP Statistic and the CH Index.

When looked at more broadly and taking into account all four methods, a less clear picture for the optimal amount of clusters emerges. According to author's judgement, two contenders for the optimal amount of clusters must be evaluated, those being the peaking amounts of 3 and 6 clusters for the final k-means model displayed in all the visualizations above.

Three clusters are well supported by the GAP Statistic and Silhouette methods indicating this amount of clusters as their highest values and dropping down significantly afterwards, only to peak again, albeit less so at around six clusters, just like the Calinski-Harabasz Index at its respective globally highest point among all cluster amounts. Although 6 clusters is best supported by the Calinski-Harabasz Index and somewhat supported by the GAP statistic and the Silhouette method, the Elbow method disagrees with this amount by a quite significant margin.

Finally, also being a weighting factor in the decision by the author is the added complexity of analyzing six groups of observations and classifying them later to provide tangible and valuable business insights for the e-commerce platform business, a decision is made to stay with an optimal amount of 3 clusters as supported somewhat in a balanced manner by all four

methods. The simpler and clearer analysis for the business owners in the findings and conclusions is worth the decision to keep things simple instead of introducing additional complexity.

After running the k-means model with the newly acquired (or properly re-confirmed) optimal amount of three clusters on the dataset at “ntstart” of 100, the new results for the model are as follows:

- Cluster sizes (observations as visualized in Figure 17.): 9 750, 1 693, 887.
- Within cluster sum of squares (dispersion) by cluster: 553,07, 56,98, 79,90.
- Total cluster sum of squres for the model: 689,94.

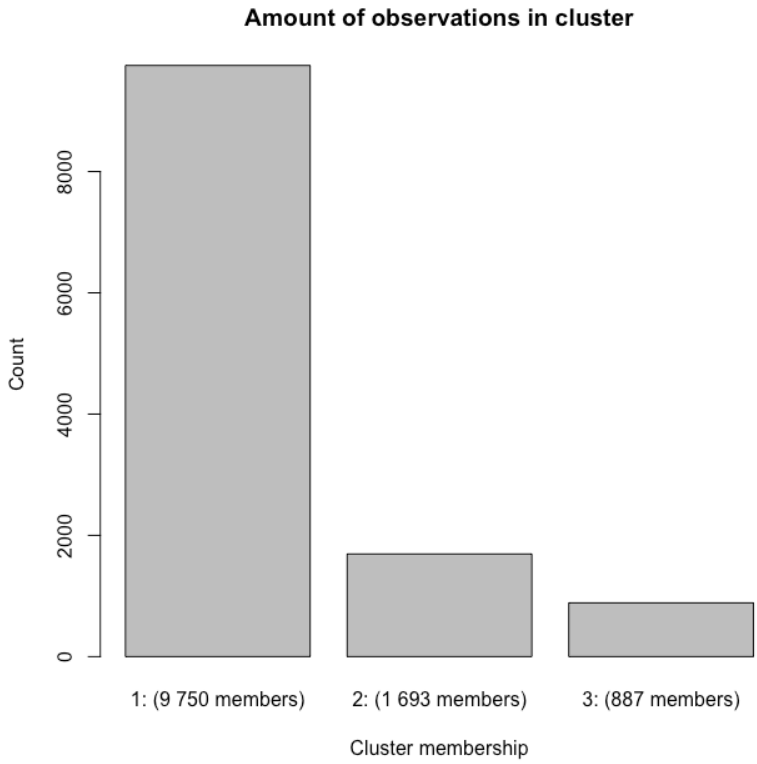


Figure 17. Amount of cluster memberships by cluster

In k-means clustering, the fact of where clustering is started may impact outcome. This means that different runs can give different results. When author sets an “ntstart” of 100 initializations, the best result is based on the smallest amount of sum of squares result in all of

those runs and it is thus less likely that the model has sub-optimal results. The negative side effect of setting a high “ntstart” may be that it takes longer for the clustering to run on weaker hardware, but in author’s case this was not a problem.

The final model does not significantly differ by its characteristics from the initial model where a guess was made to “start and see” with three clusters, albeit the member amounts are assigned differently among clusters. Author does consider the final model to be more reliable as it was run and selected after a significantly higher amount of random initializations (ntstart 100 versus 5).

By now author was formally assured that three was the right amount of clusters and the model was more reliable after it was selected from a larger amount of random initializations. It was time to combine the data and the respective memberships of each observation and give brief findings and conclusions back to the management of the e-commerce platform.

2.4. Findings and conclusions

According to author’s findings, the middle group number 2 (size of 1 693 observations), where there are predominantly new visitors, has by far the highest chance of resulting in a commercial transaction (group membership 2, revenue mean 0,25), versus the largest group where most sessions are not from new users but current ones (group membership 1, revenue mean 0,15). The e-commerce platform should definitely invest in acquiring more new visitors as showing growth in first time users will disproportionately highly reflect on the “top line” (revenue) of this particular e-commerce business.

Nonetheless, the absolute vast majority of sessions are from the non first time visitors (group number 1, size of 9 750 observations), and albeit their revenue potential can be considered just over half that of new visitors (revenue mean of 0,15 versus 0,25), perhaps they should not be ignored either as they probably serve as the committed “backbone” clientele of the entire business. Author would recommend to look into this more deeply as there

could be some “untapped potential” and a chance for improvement by developing some marketing and sales strategies to upsell the current user base.

It is worth noting that if any given session starts from a page, or part of a page with a high bounce rate, as is often the case in group 3 with 887 observations, the chances of those sessions resulting in revenue is miniscule (revenue mean 0,01). It seems as if some of the pages almost scare away visitors and the revenue potential of those sessions is lost. Sessions in this group that tend to start on a page where the mean bounce rates are high, barely if at all, spend any time on any of the parts of the platform (Administrative, informational or product-related pages). These sessions simply appear and disappear. The closeness of a special day of this group (highest mean of variable **SpecialDay**) among all of the three groups and their miniscule chance of ending up in a transaction, as well as their high average mean of starting from a page with a high bounce rate points towards an existing problem of current marketing campaigns and that they are currently tied to special days, e.g. Christmas.

Author infers that there may be problems with current landing pages tied to campaigns that are based on special days of the year. Marketing team should either improve their campaigns so that they are more convertive, or drop campaigns that are related to special days altogether as they don’t result in commercial transactions (purchases). Perhaps a lot of resources are being lost here that could be directed towards acquiring new users in other, more convertive ways, or even towards campaigns of upselling current users which both have a much more significant chance of ending in commercial transactions.

In summary for the business:

- First time sessions have the highest chance of ending in a commercial transaction.
- The current large user base may have an untapped revenue potential (upsell).
- Current marketing campaigns tied to special days like holidays are not effective:
    - Remedy 1: Drop such campaigns.
    - Remedy 2: Improve the landing pages or targeting of these campaigns.
    - Remedy 3: Focus the limited marketing resources somewhere else with a better return on investment.