LUT School of Business and Management
Master's degree in International Business and Entrepreneurship

Data analysis project
Part 1: Predicting glucose levels
Part 2: Clustering e-commerce users' profiles

Author: Matteo Bulleri

10/11/2021

Contents

Part 1

Dataset description (1)

To begin this analysis, we shall describe the dataset's characteristics and structure. The **kidney** dataset is composed by 156 observations of patients and 16 numeric variables, which represent medical data recorded at the moment of the medical visit such as blood pressure, presence of diabetes, sodium level and other vital parameters.
There are no missing values or incomplete observations, and this makes the dataset easier to treat and manipulate for analysis. In the dataset there are 5 dummy (or binary) variables, whose value can only be either 1 or 0. They are: *red_blood, bacteria, diabetes, pedal_edema, anemia*. When their value is 1, it means that the condition assessed is present, 0 otherwise. All the other explanatory variables, namely *age, gravity, sugar, sodium, potassium, hemoglobin, cell_volume, white_bloodcount, red_bloodcount* are either discrete or continuous ones.

Table 1: Data summary

Name	Piped data
Number of rows	156
Number of columns	16
Column type frequency:	
numeric	16
Group variables	None

Variable type: numeric

Variable name	mean	sd	p0	p25	p50	p75	p100
age	49.44	15.52	6.0	39.00	50.50	60.00	83.00
blood_pressure	73.85	11.04	50.0	60.00	75.00	80.00	110.00
gravity	1.02	0.01	1.0	1.02	1.02	1.02	1.02
sugar	0.26	0.82	0.0	0.00	0.00	0.00	5.00
red_blood	0.11	0.31	0.0	0.00	0.00	0.00	1.00
bacteria	0.07	0.26	0.0	0.00	0.00	0.00	1.00
glucose	131.42	65.35	70.0	96.75	113.50	131.25	490.00
sodium	138.91	7.51	111.0	135.00	139.50	144.00	150.00
potassium	4.63	3.50	2.5	3.70	4.50	4.90	47.00
hemoglobin	13.72	2.88	3.1	12.90	14.30	15.80	17.80
cell_volume	41.99	9.14	9.0	39.75	44.00	48.25	54.00
white_bloodcount	8491.03	3141.63	3800.0	6575.00	7800.00	9800.00	26400.00
red_bloodcount	4.88	0.99	2.1	4.50	4.95	5.60	6.50
diabetes	0.17	0.38	0.0	0.00	0.00	0.00	1.00
pedal_edema	0.12	0.33	0.0	0.00	0.00	0.00	1.00
anemia	0.10	0.30	0.0	0.00	0.00	0.00	1.00

The purpose of this analysis is to determine which factors have a statistically significant potential to affect the glucose level in blood, so that the interested users (e.g. Doctors) reading this report will be able to adjust their recommendations accordingly. Therefore, a regression analysis will be conducted, having *glucose* as the dependent variable.

Explanatory data analysis for selected variables (2)

4 explanatory variables and the dependent one will be analysed in more detail. Table 3 highlights the key descriptive statistics for the selected variables under examination.

Table 3: Descriptive summary statistics of Selected variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	156	49.442	15.524	6	39	60	83
blood_pressure	156	73.846	11.040	50	60	80	110
sugar	156	0.256	0.818	0	0	0	5
glucose	156	131.417	65.348	70	96.8	131.2	490
diabetes	156	0.173	0.380	0	0	0	1

Figure 1 shows a visualisation of the distribution of the selected variables through histograms for an easier and quicker understanding of the data at hand with respect to the previous table.
It can be noticed that all the selected variables are discrete ones in that the only permitted values are positive integers. This has some slight implications on the calculus but it is nothing of causing concern.

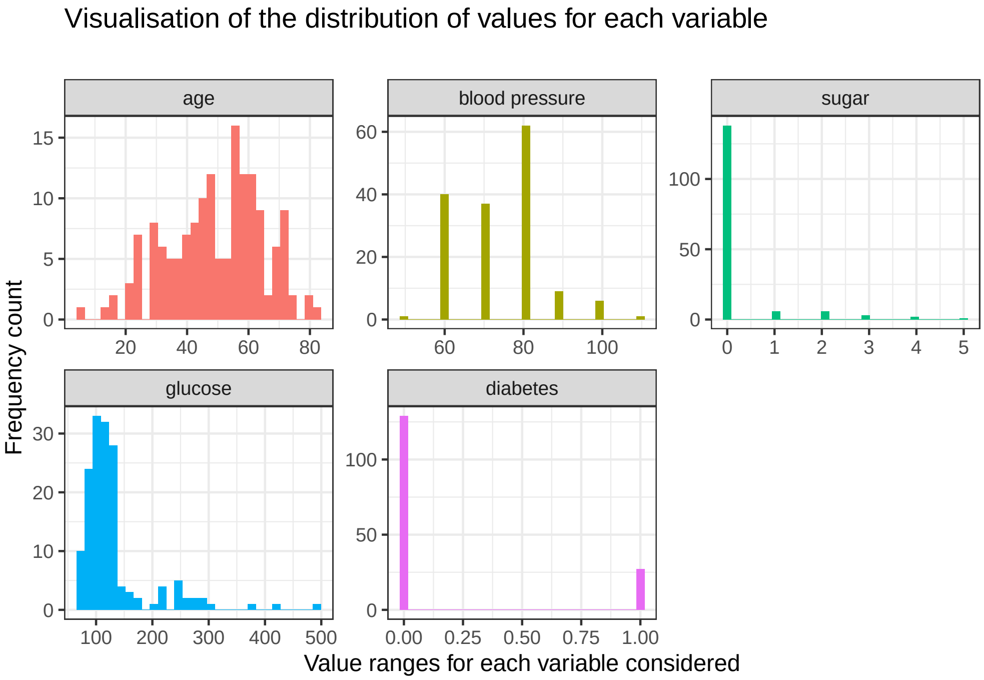

Figure 1: Data distributions

Figure 2 shows the boxplots for the selected discrete variables. Boxplots show in a visual manner 5 key descriptive statistics and highlight the presence of potential outliers. These are the median, the 25th and 75th percentile, the Interquartile Range, and the minimum and maximum values expected. The observations falling outside the whiskers are considered to be outliers.

Binary and categorical variables (i.e. *diabetes* and *sugar*) have not been considered in this instance because their distribution cannot be meaningfully represented by the boxplots.

Based on this, it seem that *age* is normally distributed and with 1 outlier. Blood pressure instead seems to be skewed to the right, having no outliers. Finally, the dependent variable *glucose* is heavily concentrated to the right and many outliers seem to exist for higher values. However, these should not be considered as such since these levels of glucose in the blood can be found in patients suffering from diabetes, which are a relatively small portion of the population.

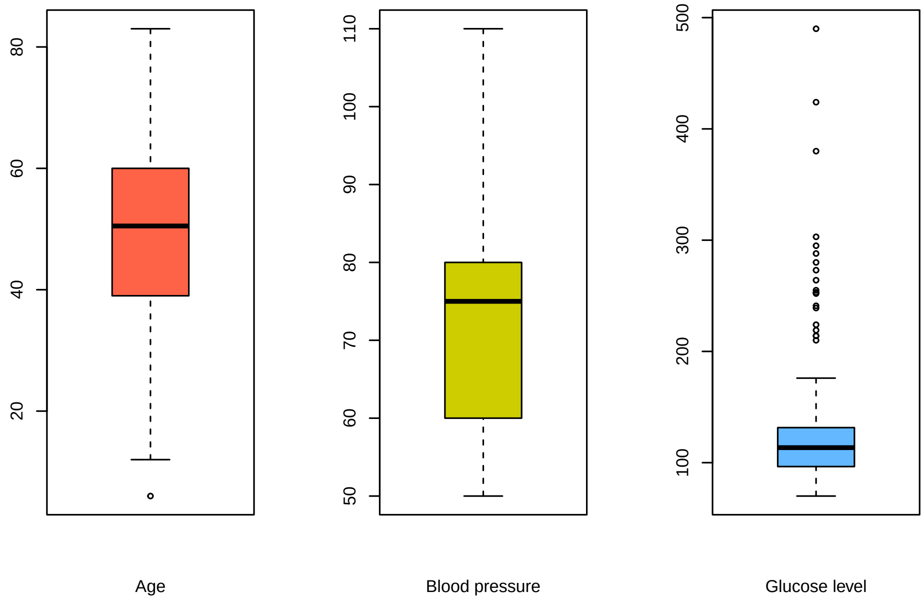

Figure 2: Boxplot distribution

Going deeper in the EDA, more specific questions about the data may be asked. For instance, are there any differences in summary statistics between different age groups?

Figure 2 and table 4 provides some statistics attempting to explore this aspect by dividing sample data into three (3) age groups: group 1 for people whose age is less or equal to 35, group 2 for people whose age is included between 35 and 60, and group 3 for people whose age is greater or equal than 60.

The main finding is that people from group 1 have not been found with diabetes and sugar in their blood. Also, they tend to have lower levels of glucose on average, thereby recording the lowest average glucose level.

Group 2 instead, the most numerous, was found with somewhat greater levels of glucose, blood pressure and sugar recorded, although the average value for blood pressure is approximately the same among all three groups. Moreover, there have been some cases of diabetes too.

Finally, group 3 has been found to have the greatest (average) levels of sugar, glucose and percentage of patients with diabetes. These results are somewhat expected and confirm the fact that the older a person is, the more likely health-related problems will be present.

Table 4: Summary statistics for selected variables by Age group

	age_group	n	avg_bp	avg_sugar	avg_gluc	avg_diab
1	<= 35	33	72.73	0	100.27	0
2	>35 & <60	80	74.63	0.23	131.91	0.16
3	>= 60	43	73.26	0.51	154.40	0.33

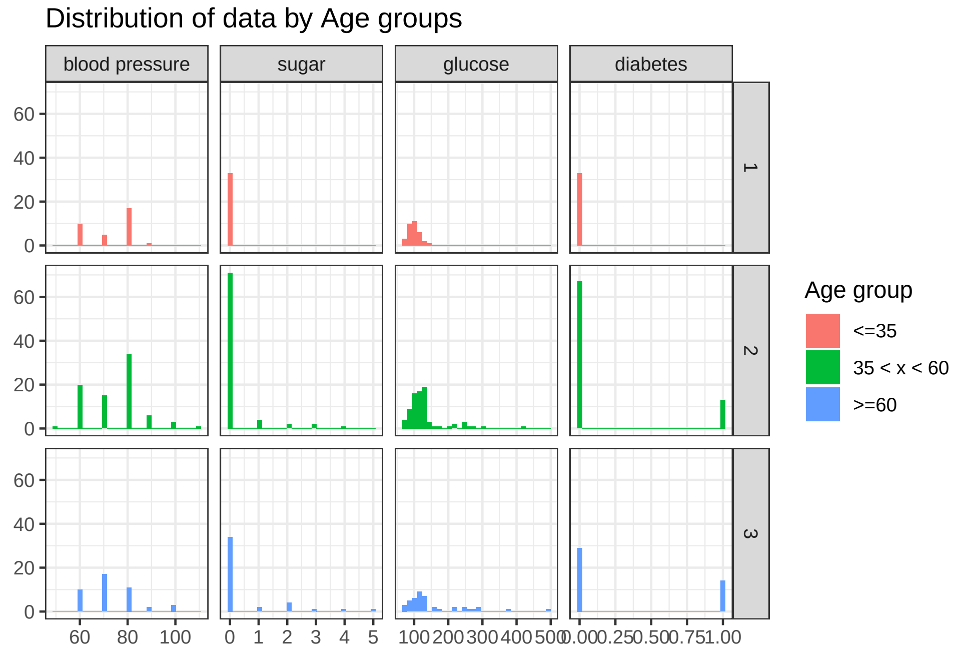

Figure 3: Focused data distribution

Correlation analysis (3-4-5)

In this section, we shall perform a correlation analysis in order to observe and understand the relationship between and among dependent and independent variables.
Figure 4 is a visualisation of the correlation matrix between the variables in the dataset. In the lower part we can see the correlation between the specific column and row in the form of a number, while in the upper part the direction of the relationship is highlighted by the circle's colour and its intensity by the circle's size. A useful remark: a strong correlation is present when the absolute value of the coefficient is >= 0.80.

The 5 highest absolute linear associations with blood glucose level are:

- *sugar*: 0.73
- *diabetes*: 0.67
- *gravity*: -0.56

- *red_blood*: 0.51
- *cell_volume*: -0.45

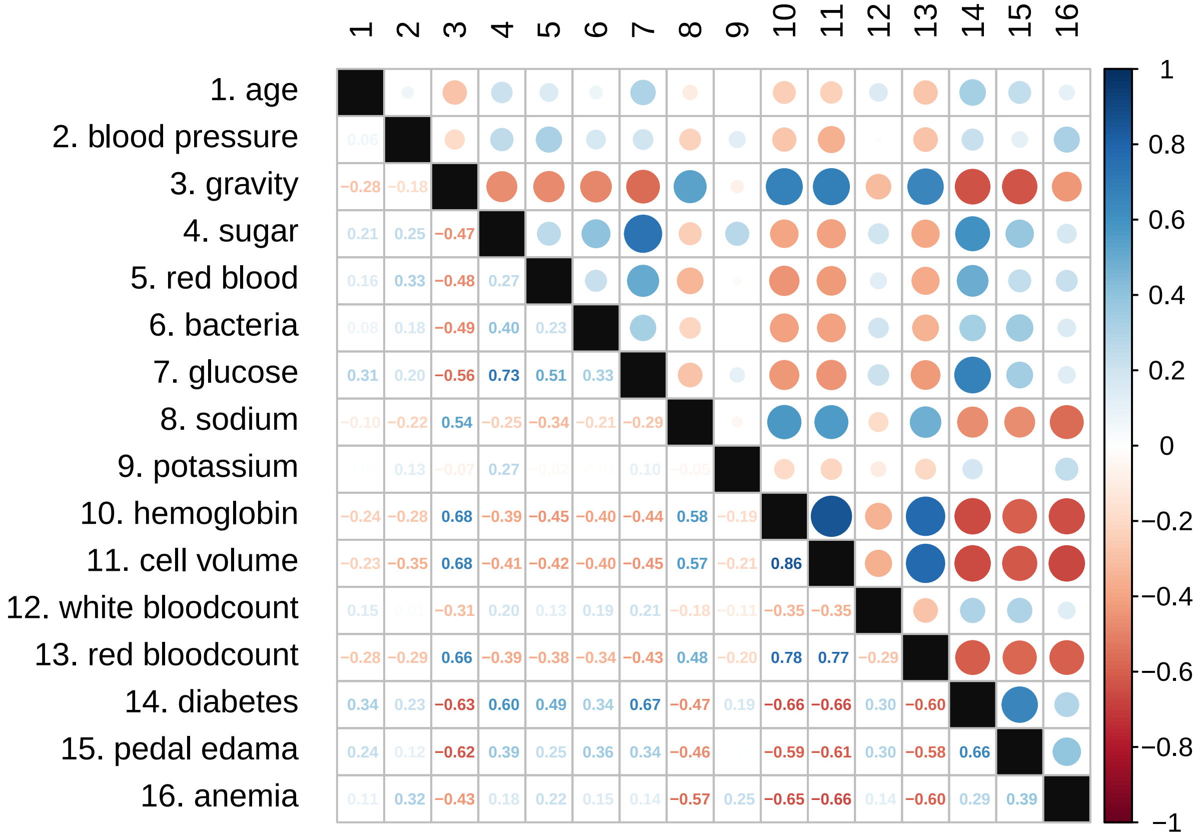

Figure 4: Correlation matrix

These outcomes make sense for the most part in that blood glucose is somehow related to them from a scientific point of view. For instance, diabetes (the pathology) can be the result of very high levels of glucose for an extensive period of time, so the correlation makes logical sense. In general however, none of the explanatory variables seem to be strongly correlated with *glucose*.

Proceeding with the correlation analysis, it is worth investigating which are the pairs of variables that are strongly correlated with each other.
The only pair with a correlation greater than 0.80 is *hemoglobin* and *cell_volume*, which have a correlation coefficient of 0.86. Other pairs with high absolute correlations but smaller than 0.80 nonetheless are: *red_bloodcount* with both *hemoglobin* and *cell_volume* (0.78 and 0.77, respectively), *pedal_edema* and *diabetes* (0.66), *gravity* with *red_bloodcount, diabetes* and *pedal_edema* (0.66, -0.63, -0.62 respectively), and in general *hemoglobin* and *cell_volume* with most of the other variables. The last finding can be the source of some issues when it comes to regression analysis and so, before proceeding, pre-processing has to be conducted.

Pairs of explanatory variables that are strongly correlated among each need to be taken care of. In particular, is is necessary to remove one of the two from the aforementioned pair as this may lead to what is know for "Multicollinearity". The basic problem is that a fractional change in a variable composing the pair would be associated with a very similar change in the other, and if they are both included in the regression model, then the effects on the dependent variable would be substantially amplified without getting any benefit in return. Therefore, this needs to be avoided and when such pairs are found, then one needs to decide which variable of the pair has to be removed.

As mentioned earlier, only the pair *hemoglobin; cell_volume* were strongly correlated (i.e. >=0.80), and so I proceeded with those. In this case, the criterion I used consisted in identifying and removing the explanatory variable that had the highest average correlation with all the other ones. The final result was that *cell_volume* was removed for the purposes of the regression analysis.

Regression analysis (6-7-8)

As initially highlighted, the purpose of this report is to give actionable recommendations regarding patients' treatment of high levels of glucose. In this regard, knowing how to predict it based on some other data is useful for Doctors wanting to, for instance, prevent the degeneration of certain conditions. Regression analysis is a tool which enables to predict the value of a feature (i.e. dependent variable) given the value(s) of others which are somewhat related to and are believed to have an effect on it (i.e. explanatory variables). So, we start by using all the variables in the dataset to create the first regression model.

Table 5 presents on the left the names of the explanatory variables, while on the right side the associated regression coefficients, standard errors and p-values. Regression coefficients represent the estimated, isolated effect of the specific feature on the dependent variable keeping all the other attributes unchanged (i.e.Ceteris paribus). The standard error (in parenthesis) indicates the extent to which the estimated coefficient could reasonably vary, on average, and finally the p-value is the probability of obtaining test results at least as extreme as the results actually observed, assuming the null hypothesis is correct. Below this section, there are some metadata describing the regression and its quality: R^2 represents the extent to which the variance of the dependent variable is explained by the independent variables employed, while the Adjusted R^2 does the same by also taking into account the number of features used. Another statistics we will look at is the *F-statistic*, which measures the degree to which the variables of the model are jointly statistically significant. Finally, as it can be read by the note at the bottom, the asterisk(s) represent the degree of statistical significance of the estimated coefficient, with more asterisks indicating a lower p-value and thus a greater statistical significance.

This model suggests that only some features are statistically significant, thus useful in determining the glucose level in blood. These are *gravity, sugar, red_blood, potassium, diabetes*, and *pedal_edema*.

The *Curse of dimensionality* has been firstly introduced by Bellman in 1961 and it refers to the fact that more features in a decision model can make the identification of (relevant) dependencies more difficult, leading to the phenomena of data sparsity. Thus, dimensionality reduction, and in particular feature extraction, has to be conducted with respect to the regression model at hand because it can help us to eliminate the non-statistically significant variables, thereby reducing noise.

To do so, we shall proceed by removing from the model one variable at once. The extraction criterion will be to remove the one with the highest p-value (or, similarly, with the lowest t-statistic in absolute terms) at each iteration, eventually reaching a regression model where all the explanatory variables are statistically significant. In parallel, particular attention will be paid to the R^2 and F-statistic items, making sure they do not deprecate as variables get removed.

Table 5: Comparison of initial and second regression models

	Dependent variable:	
	glucose	
	(1)	(2)
age	0.270 (0.207)	0.270 (0.206)
	$p = 0.194$	$p = 0.192$
blood_pressure	-0.318 (0.305)	-0.316 (0.304)
	$p = 0.299$	$p = 0.300$
gravity	$-2,332.764^{**}$ (943.457)	$-2,317.571^{**}$ (937.183)
	$p = 0.015$	$p = 0.015$
sugar	42.606^{***} (4.978)	42.707^{***} (4.936)
	$p = 0.000$	$p = 0.000$
red_blood	41.273^{***} (12.308)	40.959^{***} (12.165)
	$p = 0.002$	$p = 0.001$
bacteria	-10.676 (14.170)	-11.222 (13.856)
	$p = 0.453$	$p = 0.420$
sodium	0.213 (0.548)	0.223 (0.544)
	$p = 0.698$	$p = 0.682$
potassium	-1.810^{*} (0.979)	-1.824^{*} (0.974)
	$p = 0.067$	$p = 0.064$
hemoglobin	0.419 (2.105)	
	$p = 0.843$	
white_bloodcount	-0.0003 (0.001)	-0.0004 (0.001)
	$p = 0.760$	$p = 0.729$
red_bloodcount	-3.101 (5.220)	-2.725 (4.850)
	$p = 0.554$	$p = 0.576$
diabetes	52.714^{***} (14.952)	51.965^{***} (14.424)
	$p = 0.001$	$p = 0.0005$
pedal_edema	-41.431^{***} (13.944)	-41.503^{***} (13.892)
	$p = 0.004$	$p = 0.004$
anemia	-10.675 (15.486)	-11.796 (14.379)
	$p = 0.492$	$p = 0.414$
Constant	$2,494.192^{***}$ (942.843)	$2,481.667^{***}$ (937.557)
	$p = 0.010$	$p = 0.010$
Observations	156	156
R^2	0.712	0.712
Adjusted R^2	0.683	0.685
Residual Std. Error	36.779 (df = 141)	36.655 (df = 142)
F Statistic	24.880^{***} (df = 14; 141)	26.973^{***} (df = 13; 142)
Note:		*p<0.1; **p<0.05; ***p<0.01

So, the first dimension to be removed is *hemoglobin* since $p = 0.843$ and table 5 summarises the results. As it can be noted, R^2 remained unchanged and the F-statistic has slightly improved

(+2.09). This means that the model has improved its overall fitness without whilst reducing its complexity. This means we can continue the iterations as long as improvements occur. In the appendix of part 1 all the regression models tried out and their respective results can be found. In general, the process followed is the same at each step, and reporting each and every iteration would be wastefully redundant. The dimensionality reduction followed was (in order): *white_bloodcount, sodium, red_bloodcount, bacteria, anemia, blood_pressure, age.*

Table 6 compares the initial model with the final one. As it can be observed, all the variables that were originally statistically significant remained as such, improving in certain instances. In particular, although *sugar, red_blood, diabetes* and *pedal_edama* were already significant at a level of statistical significance of 99%, as well as *gravity* at 95% significance level, the *potassium* variable has improved its significance from 90% to 95%.

Another point to consider relates to the direction of the effects (i.e. sign of the coefficients). As a matter of fact, none of them has changed, meaning that we can be particularly confident with respect to their impact on glucose level. In other words, variables with positive coefficients indicate a positive relationship with glucose levels, and vice versa. Notice that I have not used intentionally the concept of "causation" because while proving correlation is straightforward, things are much more complex when it comes to demonstrating causal links. For a regression model, or even a single variable, to isolate a causal relationship between an independent and dependent variable certain properties and assumptions must be fulfilled. The actual presence of this properties will be assessed in the next section.

Other considerations concerning the findings from an econometric point of view relate to the fact that, in the final model, the actual value of coefficients has changed over the dimensionality reduction process. For instance, the influence of *gravity* in predicting glucose level has downplayed by 240.41, from $-2,332.76$ to $-2,092.35$. Vice versa, other variables such as the *diabetes* or *potassium* display greater effects than at the beginning as a result of the decreased noise previously due to the presence of many statistically insignificant variables.

So, the final model is represented with the following expression:

$$\begin{aligned} glu\hat{c}ose_i =& \hat{\beta}_0 + \hat{\beta}_1 gravity_i + \hat{\beta}_2 sugar_i + \hat{\beta}_3 red\_blood_i \\ &+ \hat{\beta}_4 potassium_i + \hat{\beta}_5 diabetes_i + \hat{\beta}_6 pedal\_edema_i \end{aligned}$$

Similarly, it can be represented by inputting the actual coefficient values (mathematical model):

$$\begin{aligned} glu\hat{c}ose_i =& 2,256.109 - 2,092.353 * gravity_i + 41.385 * sugar_i + 35.914 * red\_blood_i \\ &- 2.057 * potassium_i + 58.193 * diabetes_i - 45.692 * pedal\_edema_i \end{aligned}$$

The first thing that can be noticed is that both the *intercept* and *gravity* coefficients have particularly high values. This is due to the fact that the values of the *gravity* variable are discrete and range in $[1.005; 1.025]$. This tends to over-represent the actual impact on glucose level. Moreover, the consequence of this input data is that the intercept coefficient has to accommodate for it, thereby assuming an even larger value. Nevertheless, linear regression does not loose anything in terms of validity because of this, because the coefficients are the explanatory variables are robust to intra-variable measures.

diabetes, a binary variable, seems to be associated to higher levels of glucose, and this makes logical sense since diabetes is a pathology originating from exceptionally high levels of glucose in the blood.

Table 6: Comparison of initial and final regression models

	Dependent variable:	
	glucose	
	(1)	(2)
age	0.270 (0.207)	
	$p = 0.194$	
blood_pressure	-0.318 (0.305)	
	$p = 0.299$	
gravity	$-2,332.764^{**}$ (943.457)	$-2,092.353^{**}$ (802.550)
	$p = 0.015$	$p = 0.011$
sugar	42.606^{***} (4.978)	41.385^{***} (4.671)
	$p = 0.000$	$p = 0.000$
red_blood	41.273^{***} (12.308)	35.914^{***} (11.625)
	$p = 0.002$	$p = 0.003$
bacteria	-10.676 (14.170)	
	$p = 0.453$	
sodium	0.213 (0.548)	
	$p = 0.698$	
potassium	-1.810^{*} (0.979)	-2.057^{**} (0.895)
	$p = 0.067$	$p = 0.023$
hemoglobin	0.419 (2.105)	
	$p = 0.843$	
white_bloodcount	-0.0003 (0.001)	
	$p = 0.760$	
red_bloodcount	-3.101 (5.220)	
	$p = 0.554$	
diabetes	52.714^{***} (14.952)	58.193^{***} (13.389)
	$p = 0.001$	$p = 0.00003$
pedal_edema	-41.431^{***} (13.944)	-45.692^{***} (13.287)
	$p = 0.004$	$p = 0.001$
anemia	-10.675 (15.486)	
	$p = 0.492$	
Constant	$2,494.192^{***}$ (942.843)	$2,256.109^{***}$ (820.204)
	$p = 0.010$	$p = 0.007$
Observations	156	156
R^2	0.712	0.699
Adjusted R^2	0.683	0.687
Residual Std. Error	36.779 (df = 141)	36.548 (df = 149)
F Statistic	24.880^{***} (df = 14; 141)	57.754^{***} (df = 6; 149)
Note:		*p<0.1; **p<0.05; ***p<0.01

The same holds true for *sugar*: since glucose originates from sugar, presence or higher levels of sugar are expected to be found with higher levels of glucose.

With respect to the presence of red blood in cells, it appears that if these are present, then the glucose level is greater than the cases in which these are not present. It is worth pointing out that, in general, it is extremely hard and often nonsense to impute a causal link between a dummy variable and the dependent one, and this seems to apply here.

On the opposite side, *gravity, potassium* and *pedal_edema* appear to be associated with lower levels of glucose in blood.

In general terms, this final regression model performs better than both the initial model including all the variables and a "intercept-only" model. In the first case, we can conclude so by comparing the initial and final F-statistics: while the "all-inclusive" model had a F-statistic of 24.88, the final one scored 57.75, an improvement of 32.87. Also, the R^2 measure did not deprecate significantly, passing from 71.18% to 69.93% (-1.25 percentage points). With respect to the "intercept-only" case, we are confident that our model is dramatically better at modelling the isolated effects of the explanatory variables on the dependent one due to the p-value being practically zero.

Figure 5 displays the fitted *glucose* values of our model with respect to the original ones, from which we can observe that there is a very high level of fitness.

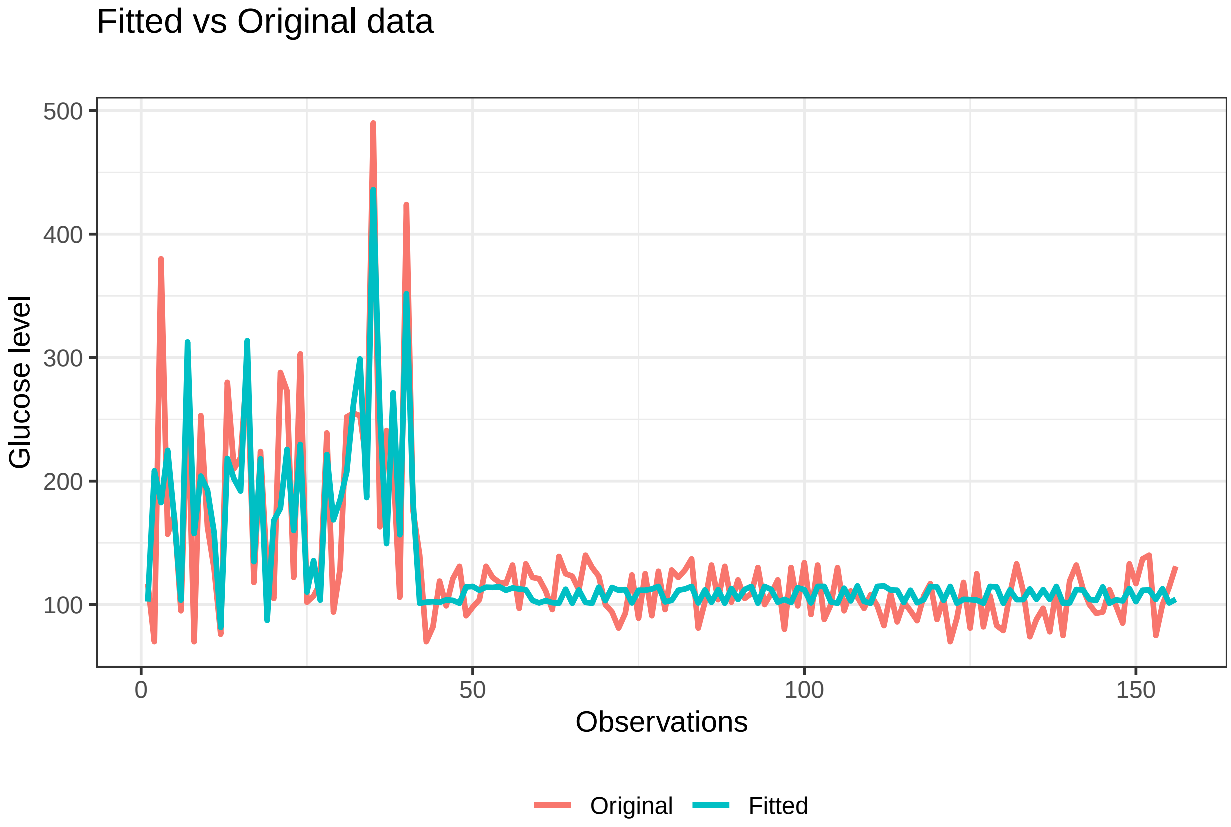

Figure 5: Fitted distribution of glucose

Residual analysis (9)

As it was previously mentioned, correlation is not causation. In order for the OLS model we developed to isolate some causal links between explanatory and explained variables some properties of OLS must hold true. If one or more of these is not fulfilled, then the reliability and usefulness of the regression model is deficient.

The five (5) properties of Ordinary Least Squares (OLS) method that are related to variance residuals are:

1) Residuals have zero-mean, namely their expected value is zero –> $E(e) = 0$

2) The variance of the residuals is constant and finite –> $var(e) = \sigma^2 < \infty$

3) The residuals are linearly independent of one another –> $cov(e_i, e_j) = 0$

4) There is no relationship between the residuals and each of the explanatory variables –> $cov(e, x_i) = 0$

5) The residuals are normally distributed –> $e \sim N(0, \sigma^2)$

These 5 properties can be assessed by using different methodologies and tests. Table 8 shows the tests and the statistics aimed at testing properties 1, 2, 3 and 5. Table 7 provides evidence for property 4.

To begin with, to check the first property we only need to calculate the expected value (i.e. mean) of the residuals. As it can be noted, the expression $E(e) = 0$ holds true, so the first property is confirmed.

The second property requires us to check for residuals' variance and the fact it is finite and constant, namely $var(e) = \sigma^2 < \infty$. In other words, we want to check that the assumption of Homoscedasticity holds true. We can conduct this test both visually (figure 6) and mathematically (using the Breusch-Pagan test).

It seems that a pattern in residual variance exists. When sorting residuals according to the glucose level, we can notice a positive relationship between glucose and residual variance. This may represent some evidence against the homoskedasticity assumption of OLS. Nonetheless, we can compute the Breusch-Pagan test to assess whether heteroskedasticity is present or not. Table 21 shows a test statistic of approximately 42 and a p-value close to 0. Given this p-value, we reject the null hypothesis that the error term's variance is homoskedastic, and therefore heteroscedasticity is present. This means that the standard errors of our regression model are understated, and thus the statistical significance of regressors' coefficients is overestimated.
One way to remedy would be to use heteroskedasticity-robust standard errors (i.e. wider).

The Durbin-Watson test aims to assess whether residuals are autocorrelated, meaning that the value of one residual affects (to some extent) the value of another residual. In other words, the covariance between the residuals of two observations shall be 0 , or similarly: $var(e) = \sigma^2 < \infty$. For this test, the null hypothesis is that the residuals are not autocorrelated and the possible values that can be assumed range between 0 and 4, with values less than 2 when residuals are positively autocorrelated and greater than 2 when negatively autocorrelated. Table 21 shows a test statistic of 2.7 and a p-value close to zero.
Given these results, then we reject the null hypothesis and conclude that they are somewhat negatively autocorrelated.

The fourth OLS property states that residuals and explanatory variables must not be correlated among each other, or put differently: $cov(e, x_i) = 0$. This can be uncovered by means of correlation

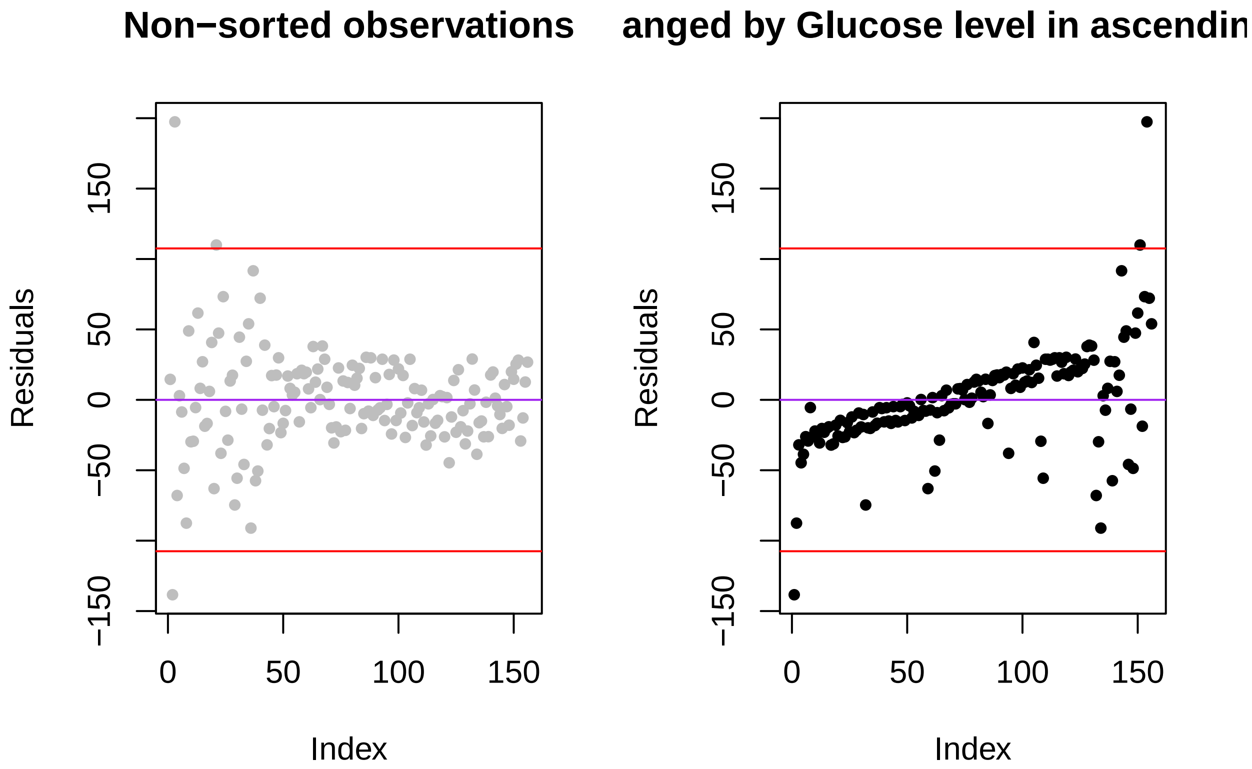

Figure 6: Residuals' distribution

analysis between explanatory variables of the regression model and the residuals.
As it can be noted from table 7, it can be concluded that this assumption holds true in that the correlation is basically zero for all the independent variables.

Table 7: Correlation Matrix - Residuals vs Explanatory variables

gravity	sugar	red_blood	potassium	diabetes	pedal_edema
0	0	0	−0	−0	0

The fifth OLS property requires us to investigate that the distribution of the residuals fulfills the conditions for normality. To do so, residuals' density function can be graphically visualised in the figure 7.

It seems that the distribution is slightly skewed to the right where there is a greater concentration of values, while also having a longer right tail.
It also appears to have a positive excess kurtosis (i.e. more peaked than a normal distribution).
Nonetheless, mathematical testing can be performed as well. The Jarque-Bera test aims to assess whether sample data, which in our case is the residual variance of the final regression model, have a normal distribution based on their skewness and kurtosis. Its null hypothesis is that sample data are normally distributed The test statistic is 309.95 (p-value ~ 0). Additional statistics related

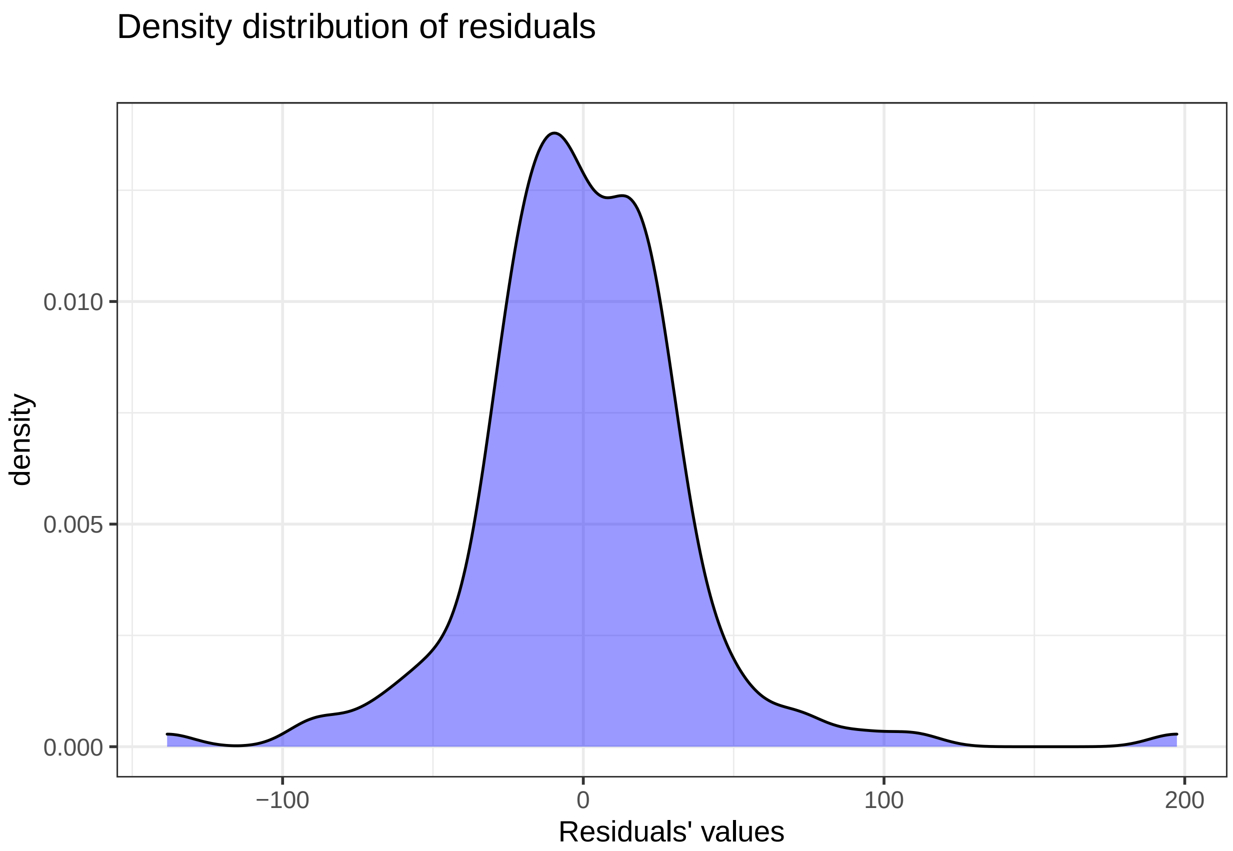

Figure 7: Density data distribution

to Skewness: 0.77 and Kurtosis: 9.73 (= excess kurtosis of 6.73). In general, the farther the test statistic is from 0, the less data are normally distributed. Therefore, we can conclude that the residuals are not normally distributed.

Table 8: Assessment of OLS properties

Test	Output statistic	Output p-value	Rejection rule
Expected value of residuals	0		<>0
Breusch-Pagan test	41.774	0	<0.05
Durbin-Watson test	2.727	0	<0.05
Jarque-Bera test	309.953	0	<0.05

Conclusions (10)

Throughout this analysis, the **kidney** dataset has been explored and some insights from the regression models have been drawn in order to pursue the initially stated goal: formulating useful recommendations to improve glucose levels' treatment, both addressed for doctors and patients. In general, the reliability of this model is high, looking at the F-statistic and R-squared measures, but it has to be pointed out that not all the properties of OLS were fulfilled, and so the reliability in terms of establishing a causal link between the independent and dependent variables lacks.

We found that sugar levels, as well as the presence of blood in patients' analyses are statistically significant for predicting higher levels of glucose. In particular, when blood is found, patients tend to have 35.9 mgs/dl more glucose than patients without blood in their analyses. Doctors may use this information to suggest their patients to do more frequent check-ups aimed at uncovering its presence, and take action preemptively. With respect to sugar levels, each additional unit in patients' blood is associated with 41.4 extra mgs/dl of glucose, thereby strongly increasing the risk of developing serious pathologies such as diabetes. On the one hand, doctors can use this information to proactively educate patients to reduce their added sugar intake, while on the other hand patients themselves could factor in this information when making food consumption choices, thus bearing in mind the effects on one's health and general lifestyle.

It was also found that a patient suffering of diabetes has, on average, glucose levels 58.2 mgs/dl higher than the same person not suffering of this condition. Being aware of this difference might be useful to doctors when it comes to identifying and thus proactively treat the pathology itself.

The other statistically significant variables have been found to decrease glucose levels, implying a negative relationship between the two. In particular, it is worth pointing out that each mEq/L of potassium is associated with a decrease of 2.057 mgs/dl of glucose in blood.

Overall, doctors and patients alike can use these insights to conduct healthier lifestyles and to keep high levels of alertness towards certain conditions such as pedal edema or diabetes.

Part 2

Dataset description (1)

The dataset used for this analysis is **dataShopping** and it contains information on customers and how they used the website of our case e-commerce platform. It is composed by 12,330 observations, where each row represents a session on the website. Moreover, there were 8 variables, such as: *administrative_duration, informational_duration, productrelated_duration, bouncerates, specialday, newvisitor, pagevalues* and *revenue.* Since in the instructions the dimension *pagevalues* was not indicated, I decided to exclude it from the analysis, thereby leaving 7 variables. All the variables are numeric and indicate a specific action when browsing the website.

The goal of this analysis is to identify the optimal number of visitors' groups so that these clusters can be understood and described. Furthermore, insights on clusters and their members' behaviour shall be drawn so that actionable recommendations on how to improve website's performance (i.e. conversion of visitors into buying customers) can be communicated to its owners.

Exploratory Data Analysis (2)

Table 9 gives us a general understanding of the assessed dimensions and some generic information on their typology and characterisics.
revenue and *newvisitor* are binary variables, with the former indicating whether the browsing section has led to a transaction. The 3 *_duration* dimensions measure the time (in seconds) spent on the associated web pages. *specialday* indicates the extent to which the navigation session was conducted in proximity to an event or holiday.

Table 9 can provide us with some broad insights on the sample. For instance, the percentage of transactions concluded within the sample is 15.5%, and an average of 13.7% of visitors are first-timers. Moreover, we can notice that both the mean and the median (p50) of *productrelated_duration* is sensibly higher than the other two *_duration*, suggesting that product-related pages are more visited (on average), and perhaps more decisive in converting a visitor into customer, although this requires a more thorough investigation by means of (e.g.) regression analysis.

Table 9: Data summary

Name	Piped data
Number of rows	12330
Number of columns	7
Column type frequency:	
numeric	7
Group variables	None

Variable name	mean	sd	p0	p25	p50	p75	p100
administrative_duration	80.82	176.78	0	0.00	7.50	93.26	3398.75
informational_duration	34.47	140.75	0	0.00	0.00	0.00	2549.38
productrelated_duration	1194.75	1913.67	0	184.14	598.94	1464.16	63973.52
bouncerates	0.02	0.05	0	0.00	0.00	0.02	0.20
specialday	0.06	0.20	0	0.00	0.00	0.00	1.00
newvisitor	0.14	0.34	0	0.00	0.00	0.00	1.00
revenue	0.15	0.36	0	0.00	0.00	0.00	1.00

We can conduct EDA also from a more visual perspective, thereby benefiting from quicker data elaboration to get a general understanding of the data.

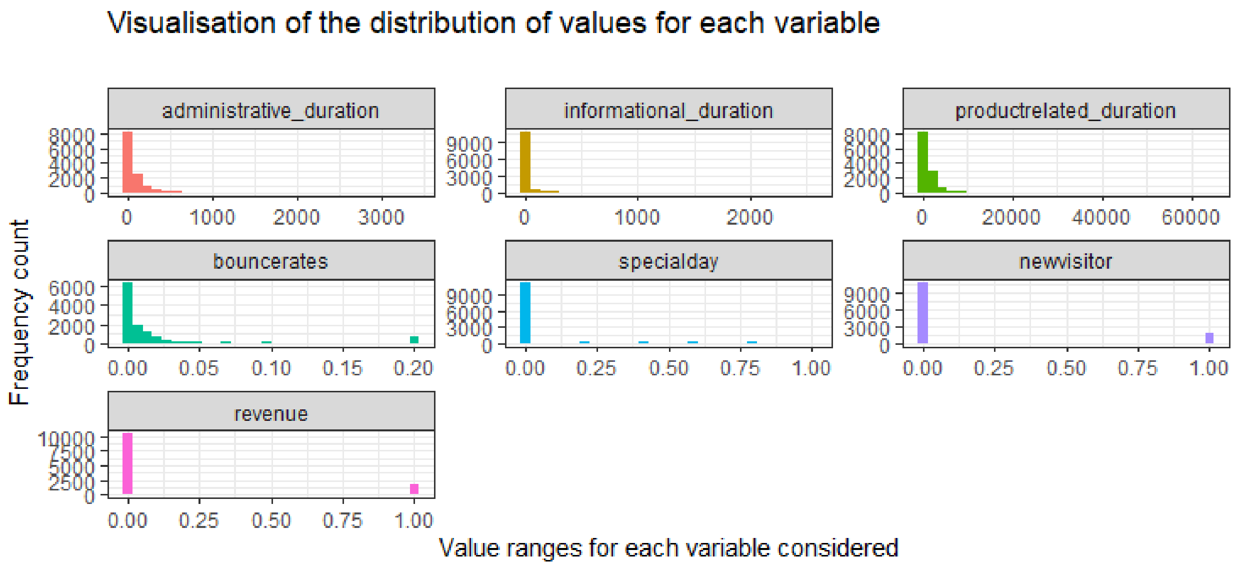

Figure 8: Data distribution

However, figure 8 appears to be a little too confusing because of the excessive amount of zeros. By focusing on values greater than zero, we may get more insightful visualisations. The next graphics (figure 9) will present the same variables whose observations' value is greater than 0, while excluding *revenue* and *newvisitor* in that they are binaries and visualising only "1" values would not make sense.

It is interesting to see that *specialday* presents a sort of normal distribution whenever there a special event or holiday. In particular, the website tends to receive its peak visits at mid-period (with respect to the start of the special day attribution). This may suggest a well-known consumer behaviour where people begin to think about gifts and presents for holidays or festivities sometime before them, and this results in peak visits in (e.g.) late November for Christmas holiday.

The boxplots in figure 10 visualise the data from figure 9, but from another perspective. Boxplots may be helpful in identifying potential outliers (in red) but in this case I feel that they are not useful for this purpose because of the typology of the data and the fact that they are not normally distributed.

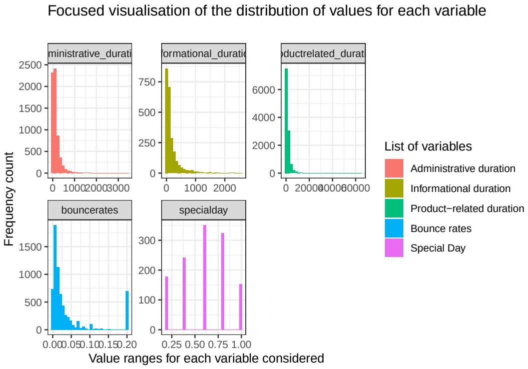

Figure 9: Focused data distribution

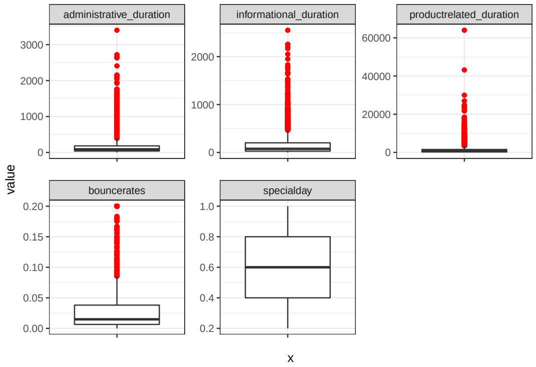

Figure 10: Boxplot visualisation

Correlation analysis (3)

In this section, we will conduct a correlation analysis in order to explore the relationships between explanatory and dependent variables, and among explanatory variables themselves.

Figure 11 shows the correlation matrix, both in a numerical and visual form. It emerges that none of the variables are strongly correlated, where as a matter of fact the highest associations are low to moderate (0.36 and 0.35) between *productrelated_duration* and both *administrative_duration* and *informational_duration*. This means that all these features and dimensions are rather independent among each other and this makes the analysis and the formulation of actionable recommendations particularly difficult. Nonetheless, additional insights may be found when controlling for some macro differences between visitors (i.e. clustering), which might uncover some particular relationships covered by their overlay.

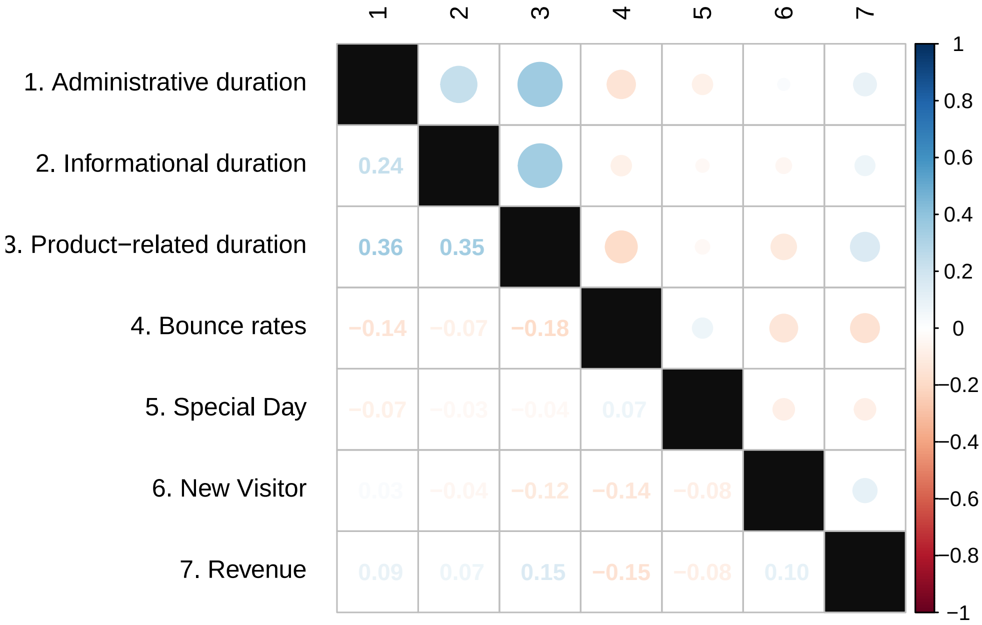

Figure 11: Correlation matrix

Data Normalisation (4)

In this section, all data points are going to be normalised by column using the min-max method. This entails to rescale the values for each feature within a pre-specified range (e.g. from 0 to 1) so that all the characteristics of the observed variables are preserved such as the distribution of data. However, the key reason to scale data is to avoid implicitly over-emphasising a variable rather than another. This may occur when different variables have extremely different value ranges, and so high absolute differences in values, although being relatively small, might be wrongly assumed to be relevant.

In the context of clustering, data normalisation is useful because it allows to disentangle (dis)similarities caused by different scales in data, and since clustering is all about finding groups with high intra-class similarity and low inter-class similarity.

Choice of the optimal number of clusters K (5)

As mentioned earlier, clustering data in similar groups may reduce the enable us to conduct a more precise and clear analysis of the factors leading to transactions on our e-commerce, especially when looking at different visitors' profiles.
Although several clustering methods exist, the one with which this analysis will be conducted is the k-means method, which consists of dividing the data points and assigning them to a specific cluster based on their distance to the cluster's mean, which is calculated iteratively.

To do this, though, this method requires us to first identify the optimal number of clusters, where "optimality" is found with the cluster's centers that maximises both intra-class similarity and inter-class dissimilarity.
Different methods and calculations exist but they often differ between each other. The elbow method tackles this problem by visualising the total within sum of squares (TWSS) associated with each attempt having different number of clusters. According to this method, the optimal number is the one after which the marginal improvements in TWSS (i.e. marginal decrease) become small. We look at the marginal improvements rather than the absolute value of TWSS because when adding more clusters a trade-off emerges: intra-class similarity increases (good) and, after a certain point (i.e. the optimum), inter-class dissimilarity decreases.
The Calinski-Harabasz index instead employs a variance-ratio criterion where the optimum is found by maximising the ratio between BGSS (Between-Group-Sum-of-Squares) and WGSS (Within-Group-Sum-of-Squares). Put simply it is about maximising the ratio between inter-class dissimilarity and intra-class similarity. This can be done by maximising the numerator, that is to increase the distance between clusters' means, and minimising the denominator, that is to decrease the distance between the specific cluster's members and the cluster's mean. Additional ways of determining the optimal number of clusters which are being used are the Silhouette method and the Gap statistic.

After having performed the required calculations, the results are displayed in picture 12 . As it can be observed, different methods provide somewhat different results. The elbow method suggests that 4 clusters' centers would be ideal, while both the Silhouette and Gap statistic methods suggest 3 as the optimal. On the other side, the Calinski-Harabasz method indicates that 6 would be the best.

When it comes to this point where results indicate divergent answers, a reasonable judgement is called. The simplest way of dealing with this issue would be to choose 3 centers because both the Silhouette and Gap statistic method suggest that, while taking into consideration that 3 is

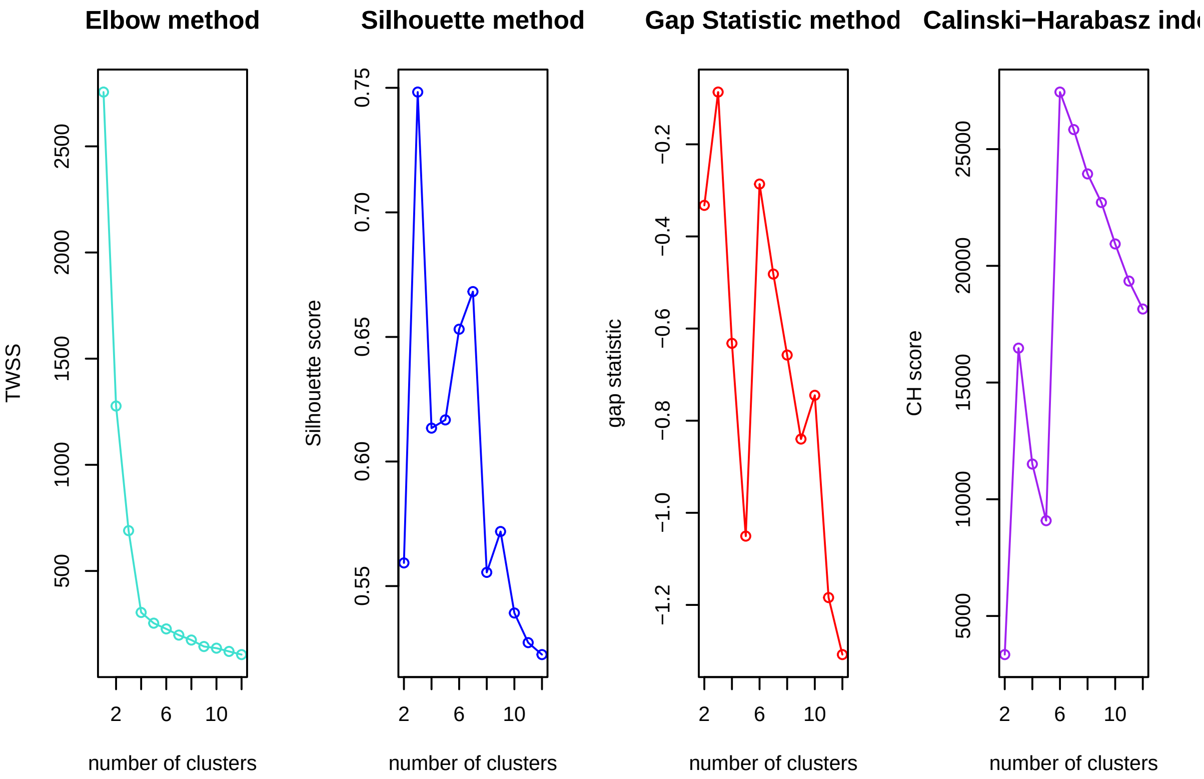

Figure 12: Methods to identify optimal number of clusters

the second-best outcome according to the elbow method. Another option could be to create an algorithm to assign a certain value (or rank) to each point in the graph and then choose the one with the highest combined value. However, this would be overly complicated and still it would be an heuristic way of making a choice, therefore I chose to follow a majority rule and picked 3 as clusters' centers.

Clusters' analysis (6)

In this section, we will conduct the actual cluster analysis aimed at understanding and getting insights about each particular cluster and its members. To do so, we will first use the chosen number of clusters in the `kmeans` function and then we will extract the cluster's membership for each observation to finally join it to the original dataset. One important parameter to set is `nstart`. As the k-means algorithm is sensitive to the initial positions of the cluster centers, setting a sufficiently high number for `nstart` enables to decrease the probability of getting very unlikely or distorded results due to the the aforementioned initial randomness. In this way, the function will run (e.g. `nstart = 100`) 100 times and only the best result will be provided, discarding the other 99 attempts. However, one cannot set an exorbitant value (e.g. `nstart = 9,000,000`) because it would computationally expensive for a result which would still remain uncertain. From table 11 the key statistics of the data are presented for each cluster. Some basic considerations involve the fact that across the groups, cluster 1 (C1) is characterised by the greatest share of transactions (`mean = 0.25`), while cluster 2 (C2) has a smaller share of transactions (`mean = 0.15`) and cluster 3 (C3) the smallest (`mean = 0.01`).

In general, it seems that C1 is characterised by new visitors only that, on average, spend quite some time in product-related webpages, as well as in the other two types. Additionally, C1 is characterised by low bounce rates and by a low propensity to visit the e-commerce during or close to special days. I would describe this group as “Resolute” because its members know what they want and do not spend much time worrying about getting the best possible deal.

With respect to C2, the most numerous group, this group is composed by old visitors only who tend to spend the most average time on product-related and informational webpages. Also, they tend to prefer to a somewhat greater degree visiting the website during or close to special days, which are likely to offer some sort of discount or special offer. In one word, I would describe this group as “Diligent” in that, conversely to the “Resolute” group, they spend quite some investigating the product (roughly twice as much) and searching it also in periods close to holidays.

Finally, with C3 being the smallest of the groups, it is characterised by the highest average bounce rates and visits during a special day, as well as by the lowest various durations and, again, transactions concluded. Also, it can be said that they are basically old visitors (mean *newvisitor* of 0.001). I would describe the members of this group as “Tap-and-Leave” since they tend to make very short sessions not buying anything, but just for the sake of surfing on the website.

Table 11

Variable name	member	mean	sd	p0	p25	p50	p75	p100
administrative_duration	1	91.97	161.58	0.00	0.00	48.35	113.67	1946.00
administrative_duration	2	86.13	185.25	0.00	0.00	11.00	98.92	3398.75
administrative_duration	3	1.19	22.58	0.00	0.00	0.00	0.00	613.67
informational_duration	1	19.25	86.54	0.00	0.00	0.00	0.00	1779.17
informational_duration	2	40.25	153.53	0.00	0.00	0.00	0.00	2549.38
informational_duration	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
productrelated_duration	1	636.77	766.41	0.00	166.71	415.50	846.25	12983.79
productrelated_duration	2	1396.69	2076.28	0.00	270.57	756.19	1720.95	63973.52
productrelated_duration	3	39.93	155.35	0.00	0.00	0.00	0.00	1787.50
bouncerates	1	0.01	0.03	0.00	0.00	0.00	0.00	0.20
bouncerates	2	0.01	0.02	0.00	0.00	0.00	0.01	0.10
bouncerates	3	0.18	0.04	0.09	0.20	0.20	0.20	0.20
specialday	1	0.02	0.12	0.00	0.00	0.00	0.00	1.00
specialday	2	0.07	0.21	0.00	0.00	0.00	0.00	1.00
specialday	3	0.10	0.24	0.00	0.00	0.00	0.00	1.00
newvisitor	1	1.00	0.00	1.00	1.00	1.00	1.00	1.00
newvisitor	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
newvisitor	3	0.00	0.03	0.00	0.00	0.00	0.00	1.00
revenue	1	0.25	0.43	0.00	0.00	0.00	0.00	1.00
revenue	2	0.15	0.36	0.00	0.00	0.00	0.00	1.00
revenue	3	0.01	0.07	0.00	0.00	0.00	0.00	1.00

Since we have successfully clustered the dataset, it is now possible to draw some deeper insights on the statistically significant predictors of transactions in the e-commerce for each identified segment. This is why regression analysis has been conducted.

After having checked that there is no multicollinearity among explanatory variables in each cluster, an iterative process to reduce the amount of dimensions taken into account in the final regression model for each cluster has been followed. Table 12 presents the final models for each cluster, hereby indicated by C1, C2 and C3. It is possible to notice that although the independent variables are highly statistically significant and the F-statistic has a significance level of 99%, the R^2 measure is basically zero, suggesting that the dimensions considered explain only approximately 4% of the variance in the results. This poses immense doubts on the reliability and practical application of this results.

That being clarified, what is interesting to consider is the opposite effect associated with administrative duration on C1 and C2, where in the first instance it counterintuitively reduces the probability of customers to make a purchase. Conversely *productrelated_duration* has a positive effect on both C1 and C2, although the measures are different. Another similarity is the negative effect of people entering and leaving the web page immediately (i.e. bounce rate), which, whenever high, may entail a low level of engagement with the page's contents, and thus the lower probability of making a purchase. A feature that is missing in all the models is *informational_duration*, which I found counterintuitive. An example of these web pages are informative blogs and this result may indicate that are not directly useful in driving up e-commerce sales. However, before drawing a definitive conclusion on this, I would recommend to gather more data and undertake a more detailed analysis investigating particularly this issue.

Table 12: Final regression models for each cluster

	Dependent variable:		
	revenue		
	(**C1**)	(**C2**)	(**C3**)
Administrative duration	-0.0002^{***}	0.0001^{***}	0.001^{***}
	(0.0001)	(0.0000)	(0.0001)
Product-related duration	0.0001^{***}	0.0000^{***}	
	(0.0000)	(0.0000)	
Bounce rates	-1.19^{***}	-2.23^{***}	
	(0.40)	(0.22)	
Special day		-0.11^{***}	
		(0.02)	
Constant	0.21^{***}	0.14^{***}	0.005^{**}
	(0.01)	(0.01)	(0.002)
Observations	1,693	9,750	887
R^2	0.04	0.04	0.03
Adjusted R^2	0.04	0.04	0.03
Residual Std. Error	0.42 (df = 1689)	0.35 (df = 9745)	0.07 (df = 885)
F Statistic	23.77^{***} (df = 3; 1689)	108.91^{***} (df = 4; 9745)	25.44^{***} (df = 1; 885)
Note:			*p<0.1; **p<0.05; ***p<0.01

Now, let us look at the key characteristics distinguishing buyers and non buyers across C1 and C2, thereby excluding C3 as it is both the smallest group and the one that does not commit to purchase through the e-commerce.

Table 13 distinguishes between C1 buyers' and non-buyers' key characteristics. No substantial differences have been found, and this somewhat confirms the idea of clustering and high intra-class similarity. The only dimension worth investigating from a visual point of view seems to be *productrelated_duration*, in that the related mean of buyers is 56.7% higher than non-buyers'.

Table 13

Variable name	revenue	mean	sd	p0	p25	p50	p75	p100
administrative_duration	0	95.69	164.66	0	0.00	54.00	122.00	1946.00
administrative_duration	1	80.75	151.55	0	0.00	28.60	92.38	1592.92
informational_duration	0	19.08	87.89	0	0.00	0.00	0.00	1779.17
informational_duration	1	19.76	82.42	0	0.00	0.00	0.00	763.00
productrelated_duration	0	557.93	741.52	0	137.88	335.67	702.92	12983.79
productrelated_duration	1	874.22	791.65	0	355.65	689.38	1167.50	5969.57
bouncerates	0	0.01	0.03	0	0.00	0.00	0.00	0.20
bouncerates	1	0.00	0.00	0	0.00	0.00	0.00	0.04
specialday	0	0.02	0.12	0	0.00	0.00	0.00	1.00
specialday	1	0.02	0.11	0	0.00	0.00	0.00	1.00

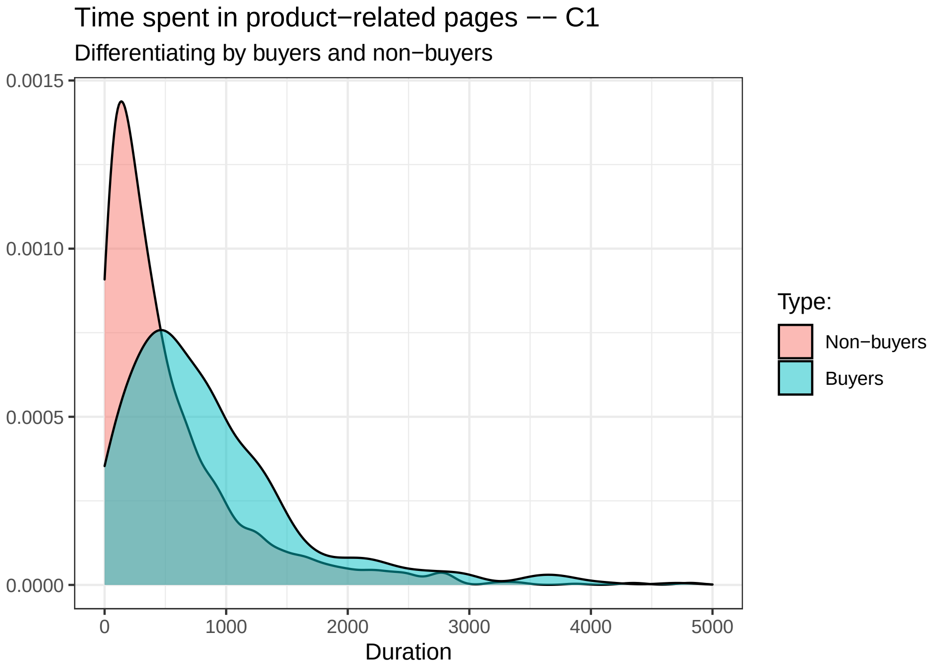

Figure 13: Density distribution of C1 buyers and non-buyers for *productrelated_duration*.

The visualisation confirms that the mean is higher because buyers' distribution has a lower kurtosis (i.e. peak) than non-buyers and therefore the mass concentration is skewed to the right, thereby also having a longer right tail. So, this is indicative that Resolute paying visitors tend to stay more on product-related pages than non-paying ones. In other words, this may indicate a the existence of a conversion factor which can be leveraged by e-commerce owners to increase sales.

Moreover, it should be recalled that this cluster is characterised by new visitors only, thereby reinforcing the idea that quality and well-indexed copywriting is capable of attracting new visitors.

With respect to C2 (i.e. Diligent group), table 14 presents the same key statistics for this group.

Table 14

Variable name	revenue	mean	sd	p0	p25	p50	p75	p100
administrative_duration	0	78.14	178.86	0	0.00	2.00	87.50	3398.75
administrative_duration	1	130.75	212.08	0	0.00	59.83	168.20	2086.75
informational_duration	0	35.18	145.85	0	0.00	0.00	0.00	2549.38
informational_duration	1	68.59	188.36	0	0.00	0.00	36.00	1767.67
productrelated_duration	0	1258.60	1955.71	0	232.57	667.17	1575.22	63973.52
productrelated_duration	1	2167.75	2515.63	0	651.00	1300.04	2671.07	27009.86
bouncerates	0	0.01	0.02	0	0.00	0.00	0.02	0.10
bouncerates	1	0.01	0.01	0	0.00	0.00	0.01	0.08
specialday	0	0.07	0.22	0	0.00	0.00	0.00	1.00
specialday	1	0.02	0.13	0	0.00	0.00	0.00	1.00

Again, although intra-class similarity is high, there are nonetheless some interesting differences, especially with respect to *productrelated_duration* and *administrative_duration*. Respectively, the difference in means is 72.24% and 67.33% more for the buyers' subgroup.

From a visual perspective, figures 14 and 15 enable to assess the data distribution too. The smaller and rightwards-shifted density peak is indicative of a greater propensity of buyers to spend some more time on the specific pages. Therefore, it may be recommended to find ways of increasing the time that C2 visitors spend on the platform, for instance by means of gamification.

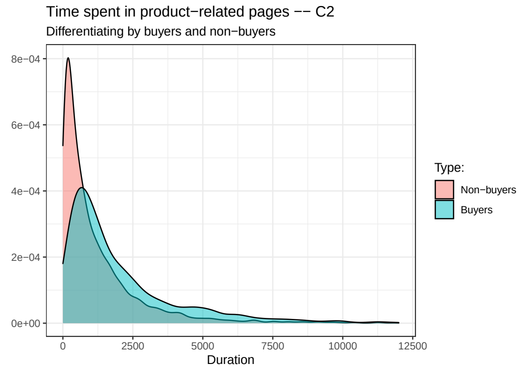

Figure 14: Density distribution of C2 buyers and non-buyers for *productrelated_duration*.

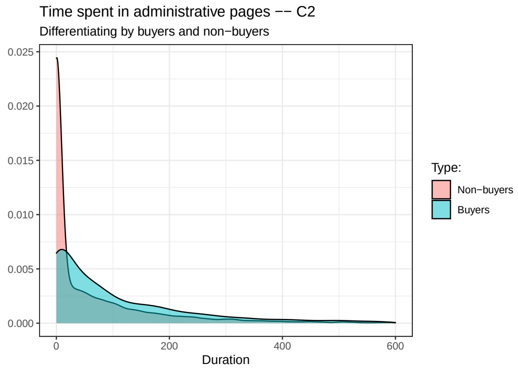

Figure 15: Density distribution of C2 buyers and non-buyers for *administrative_duration*.

However, it is also worth investigating whether there are some differences among buyers of both C1 and C2 because we may want to understand to what extent the two clusters differ on specific dimensions such as the time spent on different pages, or the general browsing behaviour. These insights might become the basis for developing some recommendation, bearing in mind that C1 is mainly characterised by new visitors and C2 by old ones, eventually aimed at enhancing the conversion power of the website, which can substantially increase the revenues for the e-commerce.

Table 15

Variable name	member	mean	sd	p0	p25	p50	p75	p100
administrative_duration	1	80.75	151.55	0	0.00	28.60	92.38	1592.92
administrative_duration	2	130.75	212.08	0	0.00	59.83	168.20	2086.75
informational_duration	1	19.76	82.42	0	0.00	0.00	0.00	763.00
informational_duration	2	68.59	188.36	0	0.00	0.00	36.00	1767.67
productrelated_duration	1	874.22	791.65	0	355.65	689.38	1167.50	5969.57
productrelated_duration	2	2167.75	2515.63	0	651.00	1300.04	2671.07	27009.86
revenue	1	1.00	0.00	1	1.00	1.00	1.00	1.00
revenue	2	1.00	0.00	1	1.00	1.00	1.00	1.00
bouncerates	1	0.00	0.00	0	0.00	0.00	0.00	0.04
bouncerates	2	0.01	0.01	0	0.00	0.00	0.01	0.08
specialday	1	0.02	0.11	0	0.00	0.00	0.00	1.00
specialday	2	0.02	0.13	0	0.00	0.00	0.00	1.00

I decided to visually compare only the dimensions that were statistically significant in both regression models (i.e. *productrelated_duration, administrative_duration, bouncerates*) as these are what might lead to some fruitful insights and data-driven recommendations. Figures 16 and 17 show that on average buying, new visitors in C1 spend less time on product-related and administrative pages than buying, old visitors in C2. The reasons associated with this tendency may be several but I would argue that the main one, thereby characterising also the main behavioural difference between the two clusters, is that C1 members are more impulsive and, as previously mentioned, resolute when navigating and deciding to make a purchase (reflected by the greater rate of transactions compared to C2). Vice versa, C2 members are more hesitant and enjoy taking their time to commit to the purchase. These insights can be useful when it comes to A/B testing or proposing personalised user interfaces to visitors depending on their profile.

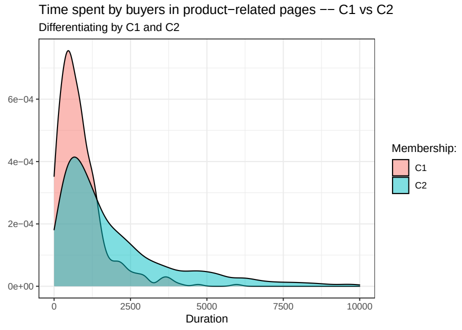

Figure 16: Comparison of *productrelated_duration*.

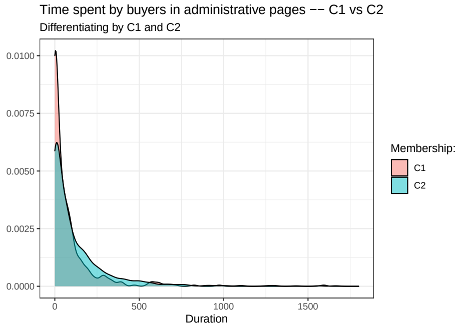

Figure 17: Comparison of *administrative_duration*.

Finally, it was was shown by both the regression analysis (table 12) and table 15 that differences in both the effects and the actual distributions of bounce rates across C1 and C2, are minimal and the visualisation in figure 18 does not provide additional insights other than the confirmation of the lower mean bounce rate of C1.

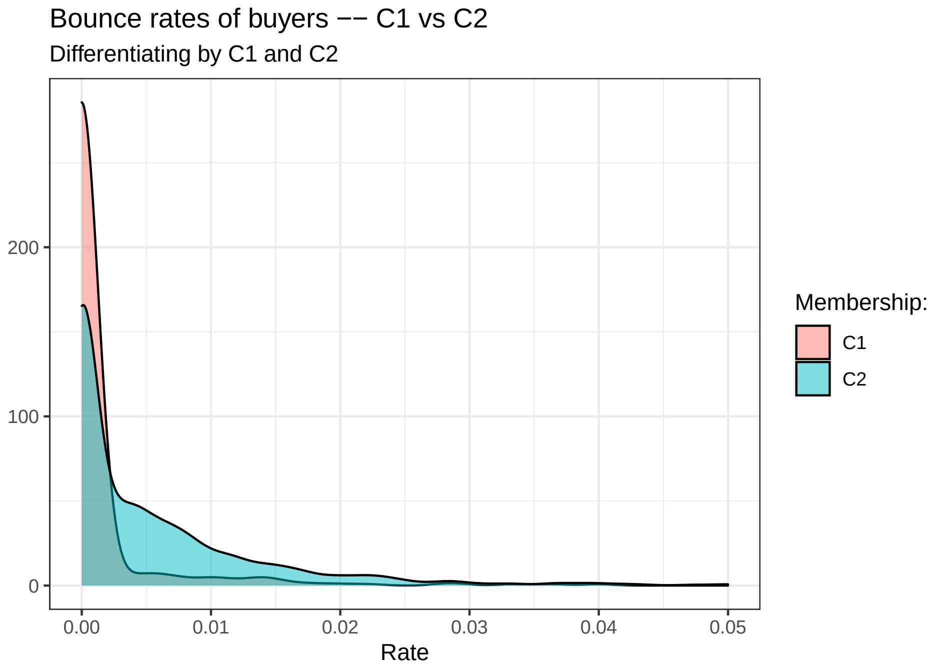

Figure 18: Comparison of *bouncerates*.

Conclusions (7)

The purpose of this analysis was to identify and cluster the visitors of the e-commerce depending on the data and variables available, and subsequently study their navigation behaviour so as to define their key characteristics and develop actionable recommendations for better targeting, advertising and branding, among the others.

3 groups were found: C1, the "Resolutes", which are first-time visitors who tend to make the most purchases (compared in relative terms to the other groups) while spending less time on browising the website. The epithet "Resolutes" was defined because of these two characteristics: their decision-making process (concerning whether to buy or not) seems to be quite fast, especially taking into consideration that being new users, they do not have previous experience with the website's design and structure. From the regression analysis, it was found that their attitude to make a purchase is significantly positively affected by the time spent in browising product-related pages. This outlined profile may be used by the e-commerce owners in a number of ways to increment transactions rate. A more pressing personalised user interface may be designed so as to transmit a sense of urgency to the Resolute visitor. Also, since they enjoy spending relatively little time on the website, all the relevant information should be clearly and quickly provided on top of the product-related page so as to increase C1 visitors' satisfaction.

The second group identified C2, the "Diligents" is characterised by old visitors who spend more time than C1 members in surfing the website and make purchases less frequently than C1 as well. One reason why the Diligents spend more time than Resolutes navigating, eventually purchasing less than the latter group, can be found in the fact that they find the web pages confusing and not responsive to their queries. Based on this, a possible recommendation might be to improve the design or user interface and contents provided so as to match their need to get exhaustive information about the product. Another reason could relate to the fact that Diligents are old visitors, and so tend to make more substantial purchases, although less frequently than new visitors. However, a major limitation of the dataset is that we do not know which share of the transactions recorded are repeat-purchases and which are first-purchases. This information may be important to better cluster visitors. Finally, a peculiar aspect is that this group is not incetivised by making purchasing in periods close to special days, as it was found a statistically significant negative relationship with the probability of concluding a transaction.

The third and last group identified C3 is characterised by what I named "Tap-and-leave" visitors, who are either new or old users who do not make purchases and have very high bounce rates. For the purposes of this analysis, I decided not to investigate them as they do not bring any value or revenue to the e-commerce.

Appendix part 1

Table 16: Third regression model

	Dependent variable:
	glucose
age	0.269 (0.205)
	$p = 0.193$
blood_pressure	-0.310 (0.302)
	$p = 0.308$
gravity	$-2,294.326^{**}$ (931.919)
	$p = 0.016$
sugar	42.636^{***} (4.916)
	$p = 0.000$
red_blood	41.232^{***} (12.103)
	$p = 0.001$
bacteria	-11.351 (13.808)
	$p = 0.413$
sodium	0.219 (0.542)
	$p = 0.687$
potassium	-1.764^{*} (0.955)
	$p = 0.067$
red_bloodcount	-2.577 (4.816)
	$p = 0.594$
diabetes	51.460^{***} (14.306)
	$p = 0.0005$
pedal_edema	-41.628^{***} (13.845)
	$p = 0.004$
anemia	-11.913 (14.331)
	$p = 0.408$
Constant	$2,454.264^{***}$ (931.363)
	$p = 0.010$
Observations	156
R^2	0.712
Adjusted R^2	0.687
Residual Std. Error	36.542 (df = 143)
F Statistic	29.391^{***} (df = 12; 143)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 17: Fourth regression model

	Dependent variable:
	glucose
age	0.279 (0.203)
	$p = 0.174$
blood_pressure	-0.312 (0.301)
	$p = 0.303$
gravity	$-2,216.000^{**}$ (908.896)
	$p = 0.016$
sugar	42.739^{***} (4.895)
	$p = 0.000$
red_blood	41.054^{***} (12.059)
	$p = 0.001$
bacteria	-11.040 (13.747)
	$p = 0.424$
potassium	-1.716^{*} (0.945)
	$p = 0.072$
red_bloodcount	-2.697 (4.793)
	$p = 0.575$
diabetes	50.290^{***} (13.969)
	$p = 0.0005$
pedal_edema	-41.854^{***} (13.793)
	$p = 0.003$
anemia	-14.319 (12.998)
	$p = 0.273$
Constant	$2,405.275^{***}$ (920.757)
	$p = 0.010$
Observations	156
R^2	0.711
Adjusted R^2	0.689
Residual Std. Error	36.436 (df = 144)
F Statistic	32.236^{***} (df = 11; 144)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 18: Fifth regression model

	Dependent variable:
	glucose
age	0.290 (0.202)
	$p = 0.154$
blood_pressure	-0.299 (0.300)
	$p = 0.320$
gravity	$-2,346.758^{***}$ (876.620)
	$p = 0.009$
sugar	42.606^{***} (4.878)
	$p = 0.000$
red_blood	41.203^{***} (12.028)
	$p = 0.001$
bacteria	-10.539 (13.685)
	$p = 0.443$
potassium	-1.669^{*} (0.939)
	$p = 0.078$
diabetes	51.867^{***} (13.653)
	$p = 0.0003$
pedal_edema	-40.979^{***} (13.673)
	$p = 0.004$
anemia	-11.391 (11.883)
	$p = 0.340$
Constant	$2,523.115^{***}$ (894.515)
	$p = 0.006$
Observations	156
R^2	0.711
Adjusted R^2	0.691
Residual Std. Error	36.350 (df = 145)
F Statistic	35.596^{***} (df = 10; 145)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 19: Sixth regression model

	Dependent variable:
	glucose
age	0.304 (0.201)
	$p = 0.132$
blood_pressure	-0.322 (0.298)
	$p = 0.282$
gravity	$-2,133.777^{**}$ (830.686)
	$p = 0.012$
sugar	41.735^{***} (4.739)
	$p = 0.000$
red_blood	41.434^{***} (12.008)
	$p = 0.001$
potassium	-1.606^{*} (0.934)
	$p = 0.088$
diabetes	52.572^{***} (13.603)
	$p = 0.0002$
pedal_edema	-41.849^{***} (13.607)
	$p = 0.003$
anemia	-10.652 (11.827)
	$p = 0.370$
Constant	$2,305.906^{***}$ (847.695)
	$p = 0.008$
Observations	156
R^2	0.709
Adjusted R^2	0.691
Residual Std. Error	36.299 (df = 146)
F Statistic	39.595^{***} (df = 9; 146)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 20: Seventh regression model

	Dependent variable:
	glucose
age	0.304 (0.201)
	$p = 0.132$
blood_pressure	-0.397 (0.286)
	$p = 0.167$
gravity	$-1,936.282^{**}$ (800.705)
	$p = 0.017$
sugar	42.323^{***} (4.690)
	$p = 0.000$
red_blood	41.220^{***} (11.998)
	$p = 0.001$
potassium	-1.846^{**} (0.895)
	$p = 0.041$
diabetes	53.769^{***} (13.529)
	$p = 0.0002$
pedal_edema	-44.813^{***} (13.195)
	$p = 0.001$
Constant	$2,110.083^{**}$ (818.810)
	$p = 0.011$
Observations	156
R^2	0.708
Adjusted R^2	0.692
Residual Std. Error	36.276 (df = 147)
F Statistic	44.500^{***} (df = 8; 147)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 21: Eight regression model

	Dependent variable:
	glucose
age	0.305 (0.201)
	$p = 0.132$
gravity	$-1,976.703^{**}$ (802.694)
	$p = 0.015$
sugar	41.337^{***} (4.651)
	$p = 0.000$
red_blood	36.707^{***} (11.586)
	$p = 0.002$
potassium	-1.971^{**} (0.893)
	$p = 0.029$
diabetes	54.385^{***} (13.565)
	$p = 0.0001$
pedal_edema	-45.283^{***} (13.232)
	$p = 0.001$
Constant	$2,123.176^{**}$ (821.331)
	$p = 0.011$
Observations	156
R^2	0.704
Adjusted R^2	0.690
Residual Std. Error	36.390 (df = 148)
F Statistic	50.265^{***} (df = 7; 148)
Note:	*p<0.1; **p<0.05; ***p<0.01