**Final report**

**Business Analytics   
Course: Project Big data (X\_400645).**

**Students:**

Yannick Hogebrug – 2626424 - y.r.hogebrug@student.vu.nl  
Jesse Schouten – 2621562 - [j7.schouten@student.vu.nl](mailto:j7.schouten@student.vu.nl)

Submission date: 28-06-2018



**Introduction**

This study was focused on the data analysis of a bedtime procrastination study. Bedtime procrastination is related to sleep problems, to which severe outcomes including memory and health problems have been related by several studies (Kroese et al., 2014). By analyzing personal sleeping data, as well as poststudy questionnaire data filled out by the participant, the following question was attempted to be answered:

* *Can bedtime procrastination be significantly influenced by experiment?*
* *How well can bedtime procrastination be predicted?*

These questions will be answered during this report. Furthermore, a few visualizations will provide extra insight into the data.

**Data description and exploration**

After merging the two datafiles, the final data has multiple variables for each participant in the study. 5 participants didn’t fill in the questionnaire and as a result they were excluded from the data in the merged dataframe. These variables are shown below including a short explanation if necessary:

- Gender: Male (=1) or female (=2)

- Age:

- Chronotype: 7 point scale if you are more a morning person (1) or an evening person (7)

- Bp\_scale: Bed procrastination scale; the higher, the more procrastination

- Motivation: Going to bed on time each night (1 = not motivated, 7 = very motivated)

- Daytime\_sleepiness: 4-point scale from 0-3; 8 questions, values summed

- Self\_reported\_effectiveness: do you feel more rested since the intervention (range 0-7)

- Group: Control- (0) or experimental group (1)

- Delay\_nights: Number of nights a participant delayed their bedtime (range 0-12)

- Delay\_time: Mean time in seconds a participant delayed their bedtime

- Sleep\_time: The mean bedtime in seconds

Because procrastination is defined as: “voluntary delay of an intended course of action despite expecting to be worse off for the delay”, delay\_time is considered the most valuable variable to answer the earlier stated research question (Kroese et al., 2014).

To get an impression of all variables, we calculated some descriptive statistics, which are shown in Table 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **standard deviation** | **median** | **min** | **max** |
| **gender** | 42 | 1,5714 | 0,4949 | 2 | 1 | 2 |
| **age** | 42 | 31,7381 | 12,1500 | 27 | 18 | 61 |
| **chronotype** | 42 | 4,9762 | 1,8450 | 5 | 1 | 7 |
| **bp\_scale** | 42 | 5,0690 | 0,9051 | 5.165 | 2,67 | 6,67 |
| **motivation** | 42 | 4,4524 | 1,1170 | 5 | 1 | 6 |
| **daytime\_sleepiness** | 42 | 16,0476 | 3,8480 | 16 | 8 | 26 |
| **self\_reported\_effectiveness** | 42 | 2,6190 | 1,3619 | 2 | 1 | 6 |
| **group** | 42 | 0,4524 | 0,4977 | 0 | 0 | 1 |
| **delay\_nights** | 42 | 7,2143 | 3,2699 | 8 | 0 | 12 |
| **delay\_time** | 38 | 2354,6842 | 1438,2945 | 1974 | 0 | 5482 |
| **sleep\_time** | 38 | 28822,3158 | 2848,8786 | 29190 | 21929 | 34644 |

Table 1: Measurements per participant of variables in dataset

Table 1 shows all appropriate measures to introduce the dataset as a whole. The count stands out as some variables have empty cells, according to the facts not all counts of the variables are the same. For the analysis of the variable with an empty cell, it was decided to remove the corresponding row. So, for some tests more data might have been used then others, as it depends on the selected variables having empty cells. We see at least one strange value, which is the maximum of the daytime sleepiness. This variable is a 4 point scale of 0-3, containing 7 questions. This means it could be at most 24. Further investigation turned out there is a total of two observations with a daytime sleepiness larger then 24. Despite this, we decided to include those variables in the analyses. Also, the maximum of age and the maximum of motivation seem to be relatively high compared to the mean and/or median, indicating there might be some outliers. For the variables we end up emphasizing in this report we will show visualizations later on in the report to give more insight in the distribution and outliers.

Because the delay time is of importance is this research, we further investigated this variable with the help of a plot. In this research we also look for differences between the two groups of the study, namely the control and the experimental group. The only difference between the groups is the fact that for the experimental group, the lights automatically dim at the intended bedtime. Because of this, we expect to see less delay time for the experimental than for the control group. Figure 1 shows the distribution of the delay time for both the control as the experimental group.



Figure 1: Delay time per participant group

Some difference is spotted, but it is not clear whether this is significant due to lack of observations.

|  |  |
| --- | --- |
|  | P-value |
| Delay nights | 0.7425 |
| Sleep time | 0.8838 |
| Delay time | 0.05367 |

**Data analysis**

Table 2: Results of Wilcoxon signed rank test for difference in groups

In order to check whether the experiment had significant difference on sleeping behavior, an analysis of sleeptime and delaytime variables was made by comparing both the experimental group with the control group.

In figure 2 we see almost no rough straight lines, maybe except the sleep time in the control group. This indicates we can’t assume normality for the statistical tests, delay time and delay nights, as the variables are not from the same location scale family as the standard normal distribution (University of Iowa, 2016). For this reason, the Wilcoxon rank sum test was used as this a distribution free test, which resulted in the p-values as described in table 2. The Wilcoxon rank sum test has hypothesis:  
  
H0: F=G  
H1: F≠G

Non of the variables turned out to be significant at a 95% confidence interval, although delay time was at the verge of rejecting.

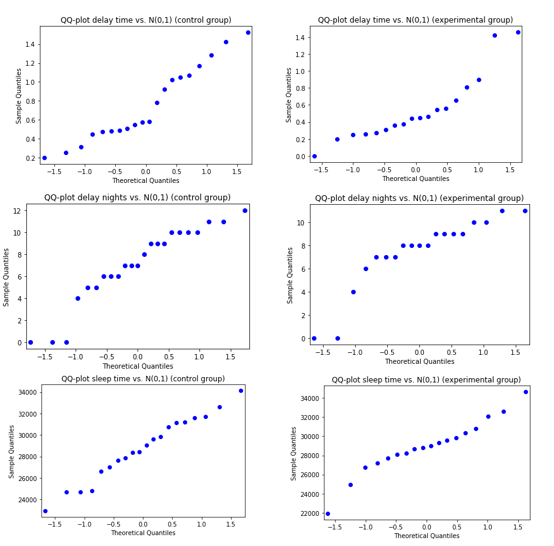


Figure 2: QQ-plots of delay time, delay nights & sleep time per group against N(0,1)

|  |  |  |  |
| --- | --- | --- | --- |
| **‘Mean delay time’ vs.:** | **‘bp scale’ (pearson)** | **‘Age’ (kendall)** | **‘Daytime sleepiness‘ (pearson)** |
| **Correlation coefficient** | 0.6118 | -0.02746 | 0.08328 |

Table 3: Correlation tests for ‘mean delay time’ vs. variables

In order to get some insight in correlation between delay time and other variables, plots are provided in figure 3. Delay time and bp\_scale seem to be related the closest from this figure, as a straight line could be drawn relatively easiest through the cloud of points. The scatterplots provide an image of correlation between delay time and age, bp\_scale and daytime sleepiness (by looking at the top row figures for example). As specified in table 3, kendall’s correlation coefficient and pearson’s correlation coefficient were calculated between mean delay time and three other variables. These measures are at least -1 and at most 1, indicating a respectively perfect negative or positive correlation in these cases. A value close to 0 indicates no correlation. It is noted Pearson just checks linear correlation, and Kendall’s checks more kinds of relationships (Statistics solutions, 2018). From table 2 it can be noted the correlation coefficient of delay time and bp scale are relatively close to 1, as the other variables turn out to be close to 0, indicating no (linear) correlation.

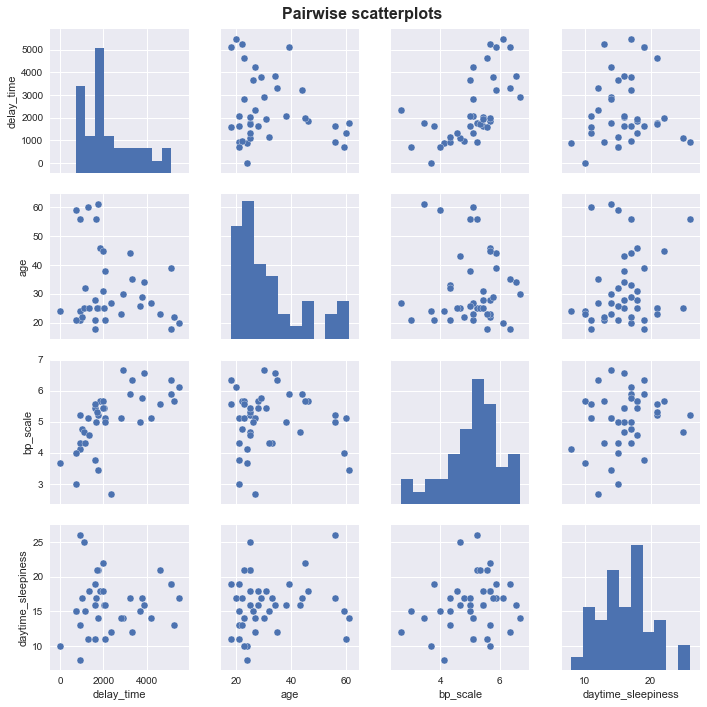


Figure 3: Pairwise plots of delay\_time vs. three other variables

For the multiple regression model, information of prior analysis was used as much as possible to predict delay time. This resulted in a model with the following explanatory variables:

* ***bp\_scale:*** a relative high correlation with delay time was found earlier.
* ***group:*** a possible statistical significance was found as a p-value almost lower then 0.05 can be seen in table 2.
* ***chronotype:*** by context it would make sense evening people are more likely to delay bedtime, as the evening is the time they probably feel food.

The scatter plot in figure 3 already provided information on the linear relationship between delay time and bp\_scale. Figure 4 provides some more insight in the relationship between delay time and chronotype and group.

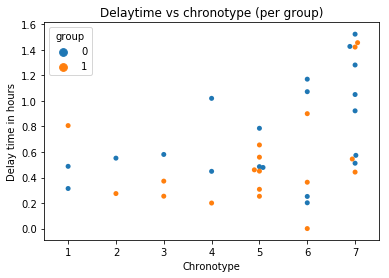


Figure 4: Plot of remaining variables in multiple linear regression model

The fact that the blue points seem to lay relatively higher then the orange points in the plot stands out. Also, the delay time seems to be higher on average as we move to the right in the figure, so when a relatively higher chronotype is observed, perhaps a higher delay time can be expected. It is noted some imagination has to be used to draw this conclusion.

The final model resulted in a R squared 0.475.

**Discussion**

During this report, we performed multiple tests on the data. The focus was on two research questions, which were mentioned in the introduction section.

The first question was about the differences between the two investigated groups to measure the influence of the experimental setup. We tested this with the help of the Wilcoxon rank-sum test. For all the tested variables, the null hypothesis was rejected. However, the p-value of the variable ‘delay-time’ was on the verge of rejecting by a significance level of 5%. Regarding the small number of observations, it is difficult to draw strong conclusions.

Besides, we looked for different variables if they were correlated with the delay time. The one that striked out was the variable ‘bp\_scale’ with an correlation value of 0.61. Using the guide that Evans (1996) suggested for the absolute value of the correlation coefficient r, this correlation was considered to be strong. As shown in the data analysis part, the other variables were not highly correlated.

In the continuation of our study, we build a linear regression model to predict the delay time, so we could answer our second research question with the help of other variables. Here, we used the earlier obtained knowledge. We included the variables bp\_scale, group and chronotype in the model. This resulted in a R2-value of 0.475, which means that 47,5% of the variance in the dependent variables (delay time) can be explained by the variance of the independent variables (bp\_scale, group and chronotype). In contrast to the value of r, there is no specific ‘good’ or ‘bad’ value for R2, this really depends on the context (Nau, n.d.).

Although the R2-value did increase after adding the variables ‘group’ and ‘chronotype’, these variables didn’t seem to have significant effect on the model (p-values of respectively 0.108 and 0.064). For a linear model it is always preferred to include as less variables as possible, which will prevent problems like collinearity (Enders, n.d.). It seemed that only the variable bp\_scale would be sufficient in the model.

Altogether, it is difficult to draw strong conclusions, based on this analysis. This is especially associated with the lack of observations. In another research, we advise to measure more participants or combine the outcome of that research with the one discussed in this report. Furthermore, we picked variables partly based on intuition. For another research, one could use for example the step-up strategy for determining the ‘best’ linear model.

Finally, more data of the participants could be collected. The sleep behavior of someone could also be influenced by the extent of effort that day or the smartphone use before sleeping (Scutti, 2017).

**Conclusion**

We can draw the following conclusions from our research:

* There are no clear reasons to assume that bedtime procrastination could be significantly influenced by experiment. Conducted tests did not show any significance between the groups, although one of them was on the verge of rejecting (p-value of 0.053).
* With the help of a linear regression analysis, we concluded that bp\_scale seemed to be sufficient for predicting the delay time. However, these model doesn’t seem really strong, regarding the small R2-value. It is difficult to draw a strong conclusion due to the lack of observations and because there is no real ‘good’ or ‘bad’ value for R2, though.

**Literature**:

Enders, F.B. (n.d.). *Collinearity.* Retrieved 27-06-2018, from <https://www.britannica.com/topic/collinearity-statistics>

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole Publishing

Kroese, F. M., De Ridder, D. T., Evers, C., & Adriaanse, M. A. (2014). *Bedtime procrastination: introducing a new area of procrastination.*Frontiers in psychology, 5, 611.

Nau, R. (n.d.). *What’s a good value for R-squared?* Retrieved 26-06-2018, from <https://people.duke.edu/~rnau/rsquared.htm>

Scutti, S. (2017). *Your smartphone may be hurting your sleep.* Retrieved 27-06-2017, from <https://edition.cnn.com/2016/11/09/health/smartphones-harm-sleep/index.html>

Statistics solutions. (2018). *Correlation (Pearson, Kendall and Spearman),* from <http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/>

University of Iowa. (2016). *QQ-plots and PP-plots, from* [*https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/qqpp.html*](https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/qqpp.html)