*Description of the dataset*

The datafile **Online Purchases.xslx** contains data about customers who have made purchases at a web shop. Data about the products was gathered and averaged out. Furthermore, some information about the behavior of the customer on the website was recorded and the customers were also asked to fill in a short survey at the end of their purchase. You need to do some statistical analyses to infer some conclusions about the behavior of the customers.

**Assignment 1 (10%)**

As an experiment, the web developers created three versions of the website: the versions are identical except that the background color is white, black, or yellow (the variable is called Front\_Page\_Type). The versions of the websites were randomly assigned to visitors of the websites. The web developers now want to know whether the color of the background influences the time customers spend on the website (variable Time\_Spent\_on\_Website). Carry out the following tasks:

1. Create a bar chart where you show the differences between the three colors when it comes to time spent on the website. Make sure to include all the required elements.
2. Perform an ANOVA to test whether background color matters. Also carry out a Tukey test to see how the three colors differ from each other. Clearly explain what the results of the analyses mean.

**Assignment 2 (10%)**

You want to know the time customers spend on the website correlates with other variables, specifically with the number of products looked at (variable: Number\_of\_products\_browsed) and the average number of pictures (variable: Pictures) for the products purchased. Carry out the following tasks:

1. Create two scatterplots: one scatterplot with Time spent on website versus number of products looked at, and one scatterplot with time spent on website versus average number of pictures.
2. Create and present a correlation matrix for three variables. Also calculate the p-value for the three correlations using the Pearson correlation and show these in the matrix.

**Assignment 3 (40%)**

As a final task, you will need to create a regression model that can predict how much a customer spends on your website. Include the following variables in your regression model:

1. **Purchase\_Amount:** (dependent variable) How much money a customer spent in total
2. **Number\_of\_products\_browsed**: Number of items the customer looked at before final purchase
3. **Device**: Whether the customer used a mobile device or a PC
4. **Pictures:** Average number of pictures for the products purchased
5. **Shipping\_Time**: Average days of shipping time for the products purchased
6. **Review\_Rating**: Average rating for the products purchased (1 = lowest, 5 = highest)
7. **Find\_website:** Survey question 1: How did you find the website?
   1. Through a social media advertisement
   2. Through a search engine
   3. Through family or friends
8. **Ease\_of\_purchase:** Survey question 2: How would you rate the ease of your purchase? (1 = worst, 5 = best)
9. **Age:** Survey question 3: How old are you? (measured in years)

Before creating the model Take the following steps:

1. There are missing values on the three survey questions. Delete them for now.
2. Make dummy variables out of the categorical variables

**Assignment 3a:** Take the following steps:

1. Present the model described above in an APA style table.
2. For each variable in the model, explain clearly what the result means.

**Assignment 3b:** Answer the following questions about this regression model:

1. Why is it unwise to include both the variables Time\_Spent\_on\_Website and Number\_of\_products\_browsed in the same model? Use relevant statistics from your data to support your argument.
2. Standardize the relevant variables. Which standardized variable has the largest influence on the amount customers purchase?

**Assignment 3c:**

1. Can you use this model to accurate predictions about how much a customer will spend? Explain clearly why or why not using relevant statistics from your data.
2. Use your model to make a prediction of the amount purchased by a customer with the following characteristics:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Products Viewed | Device | Pictures | Shipping Time | Review Rating | Find website | Ease of Purchase | Age |
| 17 | PC | 3.5 | 2 | 4.3 | Search engine | 4 | 27 |

Keep in mind the principle of parsimony when making the prediction

**Assignment 3d:**

There are missing values on the survey questions. Deal with these appropriately either by creating a dummy variable for the missing values or by using MICE. Present the results of the model with the imputed data next to the original model. Are there any substantial changes?