Final assignment of Data Analysis

!!! General notice !!!
Please make sure you write down your elaboration on the assignments extensively; meaning what you have done and why and substantiate the decisions you made.
In total you can achieve 100 points. You will pass this course if you achieve 55 points or more.

For these assignments the Saxion Report Scan applies, see appendix 1 of this document

Authors: M. Wesselink
L. Hofste

Table of Contents

1. Business Understanding & Data Understanding

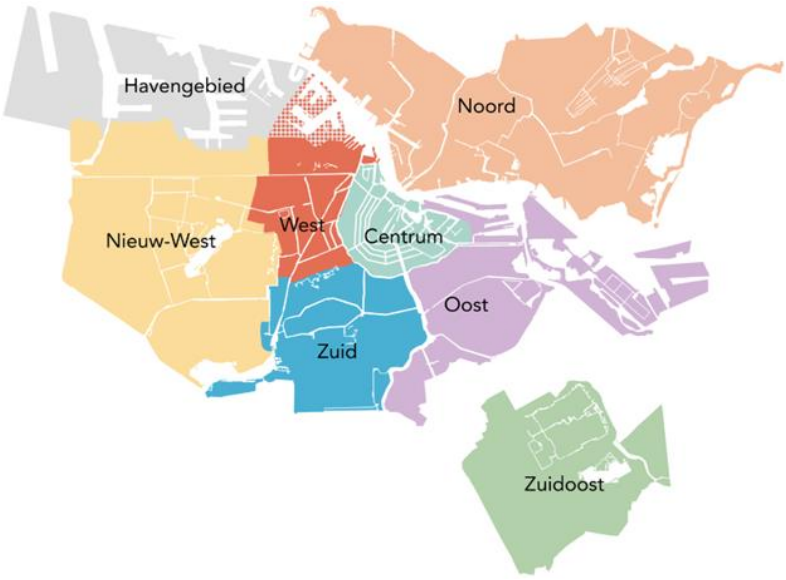
During the exercises we want to do research on the housing market in Amsterdam. We have a file with data from the internet site of 'Funda'. We want to make a **prediction model** for the selling prices of houses in Amsterdam.
Therefore we are using the CRISP-DM method, with the steps:

- **Step 1: Business Understanding**
- **Step 2: Data Understanding**
- Step 3: Data Preparation
- Step 4: Model Building
- Step 5: Testing and Evaluation
- Step 6: Deployment

Exercise 1.1 (max. 30 pt.): Business Understanding and Data Understanding

With this assignment we are going to do research on the housing market in Amsterdam. This is **step 1**, **Business Understanding**, from the CRISP-method. Unfortunately, you cannot make a prediction model for the whole city because there are too many differences in determining the house pricing in Amsterdam. Therefore, you will have to decide what part of Amsterdam you want to make your prediction model. We are only interested in houses and apartments in Amsterdam.

a. Analyse the layout of Amsterdam and the house prices in the different areas.
b. Also do a research on how the WOZ and the selling price of real estate are determined of Amsterdam.
c. Download the 'Amsterdam.csv'-file from Blackboard.
d. Define **the goal for your prediction model**. Decide on which part of Amsterdam you want make the prediction model. You can choose one of the areas (Dutch: 'stadsdeel') from the map below. Also decide whether you want to predict the prices of houses or apartments. Explain your choices.
From now on we **only use the term houses**, but this can **also refer to apartments**!



e. Define price categories that are suitable for this model, based on the house prices. Explain the choice of the categories.

f. **Create a first intuitive model** for determining the house price (category) in a certain part of Amsterdam. Indicate which variables you think influence the final sale price of a house.

With the second part of this exercise we want to understand the data from the file 'Amsterdam'. This is **step 2, Data Understanding**, from the CRISP-method.

g. Make a thorough analysis of the supplied dataset 'Amsterdam' and then provide a detailed description and meaning of the data set and the variables.
h. To better understand the data, make some **statistical calculations** and figures. For this, you can use Excel.
   Use this file 'Amsterdam' to calculate the following statistical results for houses in the chosen part of Amsterdam:
   - The mean value of the variables price, area, volume, photos, rooms and year_build.
   - The standard deviation of the variables price, area, volume, photos, rooms and year_build.
   - The range, minimum and maximum values of all the variables.
   - The boxplot of the variables price, area, volume, photos, rooms and year_build.
   - What are the outliers of each of the variables price, area, volume, photos, rooms and year_build?
   - Draw the frequency polygon of the variables price, area, volume, photos, rooms and year_build.
   - We use the normal distribution and the z-score. Find the boundaries for which 95% of the values of the variables price, area, volume, photos, rooms and year_build are between these boundaries.
   - From the results above, draw conclusions regarding the usefulness of **each** variable for the prediction model.

2. Data Preparation and Modelling

During the exercises we want to do research on the housing market in Amsterdam. We have a file with data from the internet site of 'Funda'. We want to make a **prediction model** for the selling prices of houses in Amsterdam.
Therefore, we are using the CRISP-DM method, with the steps:

- Step 1: Business Understanding
- Step 2: Data Understanding
- **Step 3: Data Preparation**
- **Step 4: Model Building**
- Step 5: Testing and Evaluation
- Step 6: Deployment

In exercise-1 we have already done step-1 and step-2 of the CRISP-DM method. For step-3 we use a special Data Pre-processing method (look below). The result of step-3 is a correct datafile which we can use for Model Building(step-4)

Exercise 2.1 (max. 20 pt.): Data Preparation and Modelling

With **Data Preparation** we use the Data Pre-processing method as described in the book 'Business Intelligence, Analytics, and Data science'. This method consists of the following four steps:

- Step 1: Data Consolidation
- Step 2: Data Cleaning
- Step 3: Data Transformation
- Step 4: Data Reduction

Step 1: Data Consolidation
It is important to select variables that are useful in predicting house prices.
You can use a scatterplot to visualize a possible relationship between the variable price and the variables area, volume, photo's, rooms and year_build. Furthermore, you can calculate Pearson's correlation coefficient (r) to measure the strength of a (linear) relationship between the variables.
Note that you use the **original price** variable, **not** the **price category** variable you created.

a. Use Orange to draw scatterplots for the variable price (y-axis) and each of the variables area, volume, photo's, rooms and year_build (x-axis).
b. Let Orange also calculate Pearson's r.
c. Which variables are you going to use for your prediction model? Give a good explanation for your choices using the scatterplots and the calculations of Pearson's r.

Step 2: Data Cleaning
In a dataset there are missing values and data which is incorrect.
Sometimes it's better for the result (prediction model) to repair the missing or incorrect data instead of removing them from the dataset.

d. Fill in the missing values, if this is possible.
e. Clean your dataset and give an explanation how you cleaned the dataset and why you did it. Identify the incorrect values in your data. Give an explanation, if you eliminate incorrect data.
f. Identify the outliers with simple statistical techniques. Watch out, the statistical calculations for categories are sometimes different from statistical calculations with the original data. Make a choice which outliers you are still using for the model and which not. Give an explanation.

Step 3: Data Transformation
This step does not apply to the assignment.

Step 4: Data Reduction
This step does not apply to the assignment (we don't have too much data).

Now your dataset is ready for the next phase, CRISP-DM - Model building

Exercise 2.2 (max. 20 pt.): Classification Model Building

We use the **DM-method Classification**.

g. Create a classification model in Orange using the dataset you prepared. To obtain a suitable model, perform the following:
   - Describe what types of classifiers are available in Orange.
   - Choose to most appropriate classifier. Motivate why.
   - Use this classifier and assess the **predictive performance** using measures that are relevant. Motivate why.

h. Create a classification model with a **decision tree** in Orange and perform the following:
   - Create a decision tree using the default settings ('min. instances per leave ' is 2) and visualize the tree model.
   - Describe which variables are useful for predicting using the tree model.
   - Describe and explain the results of the confusion matrix for the tree model.
   - Give the predictive performance of the tree model using the accuracy measure. Check the result using the formula for accuracy.
   - Minimize the number of leaves of the decision tree (change 'min. instances per leave ' to 15) and visualize the new tree model.
   - Compare the new tree visualization with the previous tree visualization. Describe and explain the differences between the two visualizations.
   - Compare the confusion matrix for the new tree model with the confusion matrix for the previous tree model. Describe and explain the differences between the two matrices.
   - Give the predictive performance of the new tree model using the accuracy measure.

3. Modelling, Clustering and Testing

Exercise 3.1 (max. 20 pt.): Data Mining with Orange; Clustering

The **DM-method Clustering** is used to find groups of similar objects or persons, f.i. customer groups with similar consumers' behaviors or similar demographic characteristics (like age, sex, marital status, family size, etc). Clustering can be used as a separate method. But in this exercise, you will use clustering to investigate whether there may exist a classification model that makes better predictions.

The approach is that you cluster the dataset on the **independent** variables of your classification model. Houses within a cluster will have similar variable values and are therefore expected to have comparable prices. By aligning price categories more closely with the ranges of house prices in the clusters, a classification model using these price categories may have better predictive performance.
Of course, the **new price categories** must be **adjacent** and **not overlapping.**

We are using the CRISP-DM method, with the steps:
- Step 4: Model Building
- Step 5: Testing and Evaluation

Step 4: Model Building (clustering)

a. Cluster the dataset you prepared in Orange using the widgets 'Distances' and **'Hierarchical clustering**'.
b. Analyse the results of this clustering. Use also the scatterplot visualization to analyse the data in combination with the hierarchical clusters.

Step 5: Testing and Evaluation

c. Decide if there is a reason to change the price categories in the dataset.
d. Important is that you give a good explanation: why you **do** or **don't change** the price categories.

Step 4: Model Building (classification, if applicable)

e. If you have changed the price categories, create a classification model in Orange using the dataset you changed.
f. Look if the predictive performance has improved.

We repeat these steps using **K-means clustering** in stead of **Hierarchical clustering**

Step 4: Model Building (clustering)

a. Cluster the dataset you prepared in Orange using the widget **'K-means**' and different number of clusters.
b. Analyse the results of this clustering. Use different boxplots to analyse the data in combination with the clusters.

Step 5: Testing and Evaluation

c. Decide if there is a reason to change the price categories in the dataset.
d. Important is that you give a good explanation: why you **do** or **don't change** the price categories.

Step 4: Model Building (classification, if applicable)

e. If you have changed the price categories, create a classification model in Orange using the dataset you changed.
f. Look if the predictive performance has improved.

Exercise 3.2 (max. 10 pt.): Data Mining with Excel & Orange; Time series

As a final assignment we would like to know what the average housing price in Amsterdam will most likely be in quarter 3 and 4 of 2021. Therefore, you will have to make a time series forecasting based on the 'timeline housing price Amsterdam.csv' file.

We are using the CRISP-DM method, with the steps:
- Step 4: Model Building
- Step 5: Testing and Evaluation

a. Remove unnecessary attributes from the dataset. Make sure that the data is nice and clean and ready to use for time series analysis.
b. Make three time series forecasting models in Excel using **Moving Averages**, **Exponential Smoothing** and **Holt's Method**. Try different parameter settings (α and β) to improve **the forecast accuracy** of the model for each method. Make sure you tweak the necessary attributes in order to get a realistic forecast for quarter 3 and 4 of 2021.
c. Create a times series forecasting model in Orange using **ARIMA**. Try different parameter settings (p, d and q) to improve the forecast accuracy of the model.
d. Compare the forecast accuracy of the four models. Describe and explain the differences between the accuracy.
e. Document all the previous step and motivate what you did and why.

Appendix 1: Report scan

Conditions on reports

We use the following six rules as conditional on reports:

- The file name and format are in accordance with the assignment.
- The names of the students are mentioned on the cover page.
- The scope of the report is in accordance with the assignment and contains all the requested parts.
- Citation is consistently cited according to a common standard, such as APA, IEEE or Harvard.
- The report is formatted and structured (font, chapter layout, etc.) in such a way that it looks neat and pleasant to read.
- There are on average no more than x spelling and / or grammar errors per 450 words (= approximately 1 A4). In addition, x is ten errors in year 1, seven errors in year 2 and five errors in years 3 and 4.

In addition to the report, all files created to obtain the results in the report must be submitted:

- For each result in the report, you state which file and part of it (for example, a particular worksheet in an Excel file) produced the result.
- It is mandatory that you describe all results and answers in the report. You may not refer to any of the files for a result or answer (e.g. “in Excel file xxx on the first worksheet you will find the answer”).

The additional files to be submitted include:

- All Orange workflows.
- All files (f.i. CSV files) to run the Orange workflows.
- All Excel files.
- The settings used in widgets of Orange workflows must be identical to the settings used to obtain the results presented in the report. Therefore, running a workflow should return the same results as presented (unless there is random behavior as in for example cross validation).
- Similar, opening an Excel file should display the same results as presented in the report.

Note: Reproducing someone else's work without citing the source is not permitted (see regulations regarding plagiarism in the EER).