Homework 1

Data Analysis and Classification 2019-2020
EDA and Linear Regression

RealEstate.csv dataset (iCorsi)

This homework has to be developed on a Jupyter Notebook. Each question needs to have at least a Code Cell (implementation) and a Markdown Cell (explanation and/or answer). The notebook developed, named as **<surname_homework_1>.ipynb** has to be sent via email at michela.papandrea@supsi.ch by sunday 20.10.2019.

1. Load the dataset RealEstate.csv
2. How many features are therein the dataset? Which is the target value?
3. Explore the data and identify the datatype of each column
4. Per each column, show the statistics of its values (value distribution, min, max, mean, mode, percentiles, ...)
5. Are there any missing values? If yes, handle them
6. Is there any outlier? If yes, what can we say about them
7. Are the values ranges comparable? Can we do anything about it?
8. Are the features correlated among them? (Visualize both the scatter plots and the correlation matrix).
9. Are the features correlated to the target value.
10. Does it make sense to reduce the dimensionality of the feature space? If yes, which are the most relevant features, considering "Price/SQ.Ft" as target value?

11. **MODEL 1**. Let's call the target values **y**. Take into account only the first most relevant feature and call it **X1**. Train a Simple Linear Regression model able to predict the value target y given X1. Use 80% of the data for training, and 20% for testing the model.
12. Plot on the same graph the scatter plot of the data X1-y, and the Least Square trained fitting line. For this line show the parameters m='slope' and q='intercept'.
13. Which is the SST (Total Sum of Squares)? Which is the SSR (or SSE, Sum of Squared Residuals)? Which is the value of R-square, and what does it mean?

14. **MODEL 2**. Let's call the target values **y**. Take into account the two most relevant features and call them <**X1, X2**>. Train a Linear Regression model able to predict the value target y given <X1, X2>. Use 80% of the data for training, and 20% for testing the model.
15. Print the model coefficients **W** and the value of 'intercept'.

16. Which is the SST (Total Sum of Squares)? Which is the SSR (or SSE, Sum of Squared Residuals)? Which is the value of R-square, and what does it mean?

17. **MODEL 3**. Let's call the target values **y**. Take into account all the features vector <**Xi>**. Train a Linear Regression model able to predict the value target y given the vector <Xi>. Use 80% of the data for training, and 20% for testing the model.
18. Print the model coefficients **W** and the value of 'intercept'.
19. Which is the SST (Total Sum of Squares)? Which is the SSR (or SSE, Sum of Squared Residuals)? Which is the value of R-square, and what does it mean?