Applied Competitive Lab in Data Science

Exercise 1

Deadline: 3.11.2022 at 23:55

0 Submission Guidelines

Your grade would be made from the top 3 questions you've answered.

Submit a single Jupyter notebook file containing your answers(including your plots, code, and textual answers). Your notebook should be run from beginning to end without errors, and conform to the style shown in class.

1 Question 1

(a) Identify the different features in the exercise dataset(ex1.csv) - which features are categorical? Are there any ordinal features? Which are continuous?

(b) Select two categorical features, and plot their value distributions. How many different unique values exist? Are the features uniformly distributed?

2 Question 2

(a) Do any null values exist? Which columns have null values? Fill those values with the mode (most common column value) before continuing, we'll discuss different ways to perform this imputation later.

(b) Find at least 3 general types of data inaccuracies, errors or illogical values (to be coherent, we do not mean 3 rows, but 3 different issues in different columns). Explain what they are, and write functions to fix them.

3 Question 3

(a) Visualize the amount of crimes and their severity throughout time- did the amount of gun violence increase over time? Did crimes become more severe? Several different features can be used for the purpose of reporting 'severity'- choose one and explain your choice.

(b) Visualize the effect of location on the amount of reported crimes. Are there more 'dangerous' areas where crimes are more prevalent? Does the trend change when observing the amount of injured victims? Note that there are several indicators of location in the dataset. Select one and explain your choice- what are the positives and negatives of your choice?

4 Question 4

(a) Next week, we're going to talk about feature engineering. You should already be familiar with One Hot encoding- try and use this encoding technique for two of the categorical features you selected in question 1. Explore the correlation between different features. Which features are closely correlated? Use a heatmap or any other visualization technique you see fit- just make sure it's comprehensible/

(b) Say we wanted a simple heuristic to determine how many people were injured in a violent crime, given all existing features (without the amount of injured and dead, of course). Use the exploration you've performed, as well as any other explorative ideas you might have, to select such a heuristic or logic. Explain your choice, write a function performing it and visualize its results.