Scientific Programming with Python - Exercise 3

Attached to this exercise is a CSV dataset, “mobile_prices_1.csv” pertaining to mobile phone prices. The data set contains specifications of mobile phones, each with an id, a set of features, and a price. You will be tasked with presenting a number of visualizations required for this data set, as well as some questions regarding the behavior of the data.

Your submission may be in pptx/pdf/docx format (your choice), and each answer must contain a visualization of the data that supports it (for example, if you claim that price grows exponentially with device speed, show a 2-d plot that for price as a function of speed). Note that a part of the grade will be given for the overall look (esthetics) of the submission – choose plot colors and sizes that allow for the reader to easily understand the graphs).

Additionally, you must submit your python code. Document for each question the section of code used to answer that question.

Following are the feature descriptions for this data set:

battery_power: mAh capacity of the battery (numerical)
m_dep: the thickness of the phone (numerical)
mobile_wt: the weight of the phone (numerical)
px_height: screen height in pixels (numerical)
px_width: screen width in pixels (numerical)
ram: amount of available RAM in MB (numerical)
sc_h: screen height in centimeters (numerical)
sc_w: screen width in centimeters (numerical)
talk_time: length of maximum call time in hours for 100% charged battery (numerical)
bluetooth: does the device have Bluetooth (categorical – Yes/No)
gen: what maximum generation network is supported (categorical – 2/3/4)
cores: core architecture (categorical – single/dual/triple/quad/penta/hexa/hepta/octa)
speed: what is the processor speed level (categorical - low/medium/high)
sim: is the sim support single or dual (categorical - Single/Dual)
f_camera: megapixel quality of the front camera (numerical, empty if no camera)
camera: megapixel quality of the camera (numerical, empty if no camera)
memory: the internal memory in MB (numerical)
screen: the screen type (categorical – Touch/LCD)
wifi: the top wifi standard supported (categorical – none/b/a/g/n)

Task 1

1. Load the data into a Pandas Dataframe.
2. Which of the categorical features are nominal and which are ordinal?
3. Add a column that holds the total screen resolution for each device. Name it resolution.
4. Add a column that holds the DPI (dots per inch) of the screen width and name it DPI_w.
5. Add a column that holds the ratio battery_power/talk_time and name it call_ratio.
6. Change the memory column to hold the memory in GB instead of MB.
7. Include the output of the `describe()` function of the dataframe.
8. Include a histogram of the prices.

Task 2

1. Plot a correlation heatmap of the data set and include it.
2. Which features would you say are correlated with the device price?
3. Are there features not shown in the correlation matrix that are correlated with the price? If so, what are they?
4. For each feature correlated with the price, plot its relationship with price.
5. Select 3 features that are correlated with price and create a pivot table showing average price with relation to cross sections of those 3 features (remember to divide numerical features into cuts, for example quartile cuts).

Task 3

1. For each ordinal feature <O>, add a column to the dataframe which holds the ordered values representing each original value of F. This new column will be named <O>_ord. (without the triangle brackets)
2. For each nominal feature <N>, add a binary column **OR** one-hot encoding (whichever is relevant for that feature) to the dataframe representing the original values. Name binary columns <N>_bin, and prefix one-hot encodings with <N>. (without the triangle brackets)
3. Plot a correlation heatmap of the modified data set and include it.
4. Save the entire dataframe to a csv file named “mobile_prices_converted.csv” and include it in the submission. Make sure you don’t add a redundant index column.

Task 4

1. Choose 4 features and use a 2-d plot to show the relationships between each pair. This should be done in the form of a 4x4 plot matrix as shown in class.
2. We have shown in class how to plot 4 dimensions of data in a 2-dimensional plot. Use this method to plot the relationship between px_width, px_height, price and core. Px_width and px_height should be the X and Y coordinates respectively.
3. There is an additional file named “mobile_prices_2.csv” distributed alongside the exercise. This file contains a mapping of id to price. This price is a transformation of the price in the original data set. The transformation has been made based on a single feature from the data set. Which feature was used and how do you know? Include any relevant plots and tables.

Do not forget to write clean, readable code, and document your functions!

Do not forget to title your graphs and add legends and color bars!

Grading:

60 points for correctness.

20 points for esthetics (visualization)

20 points for code cleanliness and readability

Task 5 – Bonus Task!

This task is worth 25 bonus points and can bring your exercise grade to above 100.

HOWEVER, these bonus points will be awarded to up to 3 student pairs. When grading submissions, if we find that more than 3 pairs have successfully answered this bonus question, points will go to the earliest submissions (based on the last submission time of each pair). Do not share your answer with other students!

1. In Task 4 question 3 you were asked to find the feature on which the transformation of price into price_2 was made. You must now find that exact manner in which this transformation was done. Analyze the data and present the algorithm used when transforming price into price_2.