ADD Coursera IBM Advanced Data Science

Why have I chosen a specific method for data quality assessment?
I have used the standard methods, in order to secure that all assumptions are correct. I went specially through the input data that I needed and rather choose the leave out incomplete data since I know that some might dilute the data: some missing dimensions suggest for example a round measurement (e.g. radius).

Why have I chosen a specific method for feature engineering?
I choose to use dummies for the "categories" a.k.a. product areas as these are a multi class. This would mean that I can integrate it as a feature rather than model per category. Besides this I choose to standard scale rather than normalize. The reason for this is that I did not expect any outliers after removing nan values in sizes.

Why have I chosen a specific algorithm?
I tried my standard set of algorithms:

- Linear regression as a baseline
- Gradient booster as versatile model which generally performs best in my cases
- Simple neural network, in this case with MSE metric and linear activation

The final decision is that GBR has the best performance. Probably with some extra time the model will perform better when the overfitting is reduced. Also GBR is faster than a neural net on this size of data.

Why have I chosen a specific framework?
I used:

- Python as code language as it is most integrated.
- Seaborn and Pandas to present data in a way that is suitable for data analysis using PD dataframes.
- Keras for the neural nets as it is the most straightforward but allows for specifying the nodes.
- Sklearn for both linear regression and gradient booster regressor. It contains all needed models and metrics.

Why have I chosen a specific model performance indicator?
I used MSE as a performance indicator as it is easy to use for all different models and easiest to interpret.