Review for: Dennis Scheper's Research log
Reviewed by: Wytze Bekker (380875)

Introduction
It is immediately clear which research we will be looking at and what the conclusion of that research was. Some wording feels a little strange, however, like "further enlargement" could just be "enlarge" or "improve" and "diabetes tactics" could be "diabetes treatments", to improve readability.

Dataset and attributes information
Not loading in the codebook for space saving and readability reasons feels a bit odd as later on, you do show all a 50 column table. Showing the codebook might however indeed be interruptive in reading the paper, so your choice of looking at a few essential attributes is a good alternative.
You explain everything extensively which is great, but does cause a wall of text. Dividing this into separate paragraphs will most likely improve ease of reading, while not sacrificing any part of your very thorough explanation.

Research question
The research question appears to be slightly wrongly worded, I think you mean to find out if you can predict a patients time in hospital from their HbA1c test, not only if it is linked.

EDA
You start with clearly describing what will happen in this part.
You mention you load in a codebook, why not show it? You focused on the main attributes earlier which gave information, but you specifically mention loading the codebook but do not show it.
Next, you show an example piece of the dataset, which is useful. The summaries of some of the rows are however not very useful, for example summaries of "encounter_id" and "patient_number". I assume these patient numbers are just randomly assigned and do not have any numerical worth; showing the average patient number does not really give me any information.

Missing values
A nice clean table showing exactly what you are talking about and again a clear explanation of what conclusions and next steps you derive from this. (You refer to it as "Table 1" but nowhere on/near the table does it say "Table 1".) You checked to see if you could just remove all NA's, but were paying attention as this removed too much data so you wisely decided not to go that way. You use good logic to see if you can remove NA data from "gender".
Yet again, you describe what will happen in the next session, that really improves readability.
Under the table about records and duplicates, you appear to have used some code to get the correct numbers from the table in the text. In my pdf version however, they just say "NA". You could just type the numbers from the table, as they will most likely not change.

Categorial attributes
Description of what will happen, talked about it, love it.

It starts of with a strange sentence: “Since observe and use this attribute, it is essential to remove the abundant value.” I have no idea what this is supposed to be, but most likely just a typo.
A clear explanation of figure 1, but it is on the next page, which makes it difficult to look at the figure while reading the explanation. The explanation flows into an explanation of figure 2, but maybe it would be convenient to put figures 1 and 2 on the same page, maybe by making both narrower, which should not be a problem if you rotate the X-axis labels by using “+ theme(axis.text.x = element_text(angle = 90))”.

The same for figures 3 and 4, even though this might create problems with the (bigger) legends of the figures.

Figure 5 looks clear but might benefit from giving it some more vertical space, to improve readability of the bars. Also, some of the labels overlap so rotating them might prove helpful. In the corresponding text, you mention what interesting things we can see in them and how the authors of the original dataset did not use this data, but also how you might still use it, if you deem this necessary, which shows critical thinking.

Figure 6 looks a bit overwhelming, is it possible to merge the smaller groups which are now slightly messy? This also applies to figure 7 (which is not actually named).

Distribution - numeric attributes
Is it really necessary to show the figures twice(big and small)?

Correlation - numeric attributes
For complicated figures such as heatmaps, it would again be convenient to have the interpretation and explanation on the same page as the figure. The explanation itself is clear. For figure 14 and 15 the figured could do with nicer axis names, but because the names of the parameters are already quite clear, this is not strictly necessary for interpretation of the figures.

References
No comments really, just a proper reference list.

Overall
It looks great, you explain everything beautifully thoroughly, it could do with some small fixes in word usage and placement, but overall it looks good.