Task 1: Blocking:

(a) How does blocking affect your results? Specifically, describe your choice of blocking method and choice of blocking keys.

Discuss which attributes and/or attribute combination(s) in the given data sets were useful as blocking keys and which were not, and why.

(b) If there is a trade-off between performance (reduction ratio, pairs completeness, and pairs quality) and the quality of the final record linkage results, where do you think the optimal balance is, and why?

(c) Do you think this trade-off would change on different data sets with different levels and characteristics of data quality?

If so, how and why?

Write a maximum of 20 lines of text (around 500 words) in total in the corresponding answer field on Wattle. Clearly indicate

your answers to (a) to (c).

===============================================================

Task 2: Comparison and Classification:  
(a) How do different comparison techniques affect linkage results? Discuss and justify how you selected appropriate comparison functions for different attributes, and why these selected functions are suitable while others were not.  
(b) How do different classification techniques using different parameter settings affect linkage quality? Discuss and justify how you selected an appropriate classification function and corresponding parameter settings to obtain high linkage quality.  
(c) As discussed in the lectures in week 4, for suitable linkage quality measures, describe how the final record linkage quality changes with the choice of different parameters and techniques?  
(d) Are there any evaluation measures that are not useful? Describe why these measures are not useful evaluating the performance of a record linkage project.  
(e) Is the record linkage quality particularly sensitive to certain parameters, or choice of comparison or classification techniques? If so, why is this the case?  
(f) Provide the numerical linkage evaluation results for other (not optimal, see below) parameter settings that you have used (you only have to provide the output file for your best obtained linkage results { see next task).  
Write a maximum of 20 lines of text (around 500 words) in total in the corresponding answer field on Wattle. Clearly indicate your answers to (a) to (f) [about 70 -100 words each sub question.]

Task 3: Optimal Settings:  
(a) What is the best linkage quality result you are able to achieve, both in the blocking and the classification steps? Why do you think this combination of parameters and techniques worked well for your data set pair?  
(b) Are the results good for all evaluation measures discussed in the lectures in week 4, or only for some? If the results are good only for some measures, why do you think the results are not good for other measures?  
Write a maximum of 250 words (around 10 lines of text) in the corresponding answer field on Wattle. Clearly indicate your answers to (a) and (b).  
In addition to answering this task in Wattle, you must also submit the output file which contains the linked and classified matching record pairs (as a CSV file) for the best linkage result you were able to obtain.  
You must use the Python program saveLinkResult.py which we use in lab 5 to write linkage output into a file. Your submitted output file must exactly follow this CSV file format! We will use a program to check linkage quality using this file to validate what you write in your answers in Wattle. If our program does not work with your submitted file because it does not follow the required file structure then you will loose marks.

Task 4: Data Quality:  
(a) How dirty are these new data sets you generated compared to all the data sets you have worked with in labs 3 to 5?   
Describe your impression after having conducted the linkage on the different data sets used in the labs.  
(b) How can you determine this? Describe the methodology you used to assess the data quality of the data sets we provided for this assignment (such as any calculations you used, or how you determined the data quality using data exploration and profiling).  
Write a maximum of 250 words (around 10 lines of text) in the corresponding answer field on Wattle. Clearly indicate your answers to (a) and (b).