First of all we gathered the data. We used a dataset containing data about 400 patients with/ without chronic kidney disease (ckd).

The dataset contained the following features:
- Age (numerical) - age in years
- Blood Pressure (numerical) - bp in mm/Hg
- Specific Gravity (nominal) - sg - (1.005,1.010,1.015,1.020,1.025)
- Albumin (nominal) - al - (0,1,2,3,4,5)
- Sugar (nominal) - su - (0,1,2,3,4,5)
- Red Blood Cells (nominal) - rbc - (normal,abnormal)
- Pus Cell (nominal) - pc - (normal,abnormal)
- Pus Cell clumps (nominal) - pcc - (present,notpresent)
- Bacteria (nominal) - ba - (present,notpresent)
- Blood Glucose Random (numerical) - bgr in mgs/dl
- Blood Urea (numerical) -bu in mgs/dl
- Serum Creatinine (numerical) - sc in mgs/dl
- Sodium (numerical) - sod in mEq/L
- Potassium (numerical) - pot in mEq/L
- Hemoglobin (numerical) - hemo in gms
- Packed Cell Volume (numerical)
- White Blood Cell Count (numerical) - wc in cells/cumm
- Red Blood Cell Count (numerical) - rc in millions/cmm
- Hypertension (nominal) - htn - (yes,no)
- Diabetes Mellitus (nominal) - dm - (yes,no)
- Coronary Artery Disease (nominal) - cad - (yes,no)
- Appetite (nominal) - appet - (good,poor)
- Pedal Edema (nominal) - pe - (yes,no)
- Anemia (nominal) - ane - (yes,no)
- Classification (nominal)- classification - (ckd,notckd)

Then we accessed the data visually, and we discovered the following quality issues:
- 'sg', 'al', and 'su' attributes should be ‘object’ not ‘float64’.
- 'rbc' attribute has about 38% missing values (null values).
- 'classification' attribute has 'ckd\t' value which has '\t' as noise.
- 'pcv' attribute has \t43 value which has '\t' as noise.
- 'pcv' attribute has \t? value which is noise.
- 'dm' attribute has '\tno' value which has '\t' as noise.
- 'dm' attribute has '\tyes' value which has '\t' as noise.
- 'dm' attribute has ' yes' value which has ' ' as noise.
- 'cad' attribute has '\tno' value which has '\t' as noise.
- 'pcv' attribute should be ‘float64’ not ‘object’.
- There exist missing values in 'age', 'bp', 'bgr', 'bu', 'sc', 'hemo', and 'pcv' attributes.

- There exist missing values in 'sg', 'al', 'su', 'pc', 'pcc', 'ba', 'htn', 'dm', 'cad', 'appet', 'pe', and 'ane' attributes.
- 'rc' attribute has 'May-00', 'Apr-00', '\t?', 'Jun-00', and 'Mar-00' values which are noise.
- 'wc' attribute has \t6200 value which has '\t' as noise.
- 'wc' attribute has \t8400 value which has '\t' as noise.
- 'wc' attribute has \t? value which is noise.
- The type of 'rc' and 'wc' should be ‘float64’ instead of ‘object’.
- There exist missing values in 'sod', 'pot', 'rc', and 'wc'.

Then we cleaned the data by doing the following:

1) Converting the data types of 'sg', 'al', and 'su' to ‘object’.
2) Dropping 'rbc' column as the missing values occupy about 38%.
3) Replacing 'ckd\t' with 'ckd' in the 'classification' column.
4) Replacing \t43 with 43 in the 'pcv' column.
5) Replacing \t? with the median value in the 'pcv' column.
6) Replacing '\tno' with 'no' in the 'dm' column.
7) Replacing '\tyes' with 'yes' in the 'dm' column.
8) Replacing ' yes' with 'yes' in the 'dm' column.
9) Replacing '\tno' with 'no' in the 'cad' column.
10) Converting 'pcv' to float64.
11) Replacing missing values in 'age', 'bp', 'bgr', 'bu', 'sc', 'hemo', and 'pcv' by median value.
12) Replacing missing values in 'sg', 'al', 'su', 'pc', 'pcc', 'ba', 'htn', 'dm', 'cad', 'appet', 'pe', and 'ane' by most frequent value.
13) Separating 'sod-pot' column into two columns: 'sod' and 'pot'.
14) Separating 'rc / wc' column into 'rc' (red blood cell count) and 'wc' (white blood cell count) columns.
15) Replacing 'May-00', 'Apr-00', '\t?', 'Jun-00', and 'Mar-00' values with median value in 'rc' column.
16) Replacing \t6200 with 6200 in the 'wc' column.
17) Replacing \t8400 with 8400 in the ‘'wc' column.
18) Replacing \t? with the median value in the 'wc' column.
19) Converting 'rc' and 'wc' to float64.
20) Replacing missing values in 'sod', 'pot', 'rc', and 'wc' by median value.

Then we stored the cleaned data frame in a CSV file ('kidney_disease_cleaned').