**Chronic Kidney Disease**

**Exploratory Analysis**

Dylan Tulett, dtulett@bellarmine.edu

1. **INTRODUCTION**

The Chronic Kidney Disease dataset is a comprehensive dataset that includes data on both patients with and without chronic kidney disease (CKD). It includes many variables that are biomarkers for CKD (Hb count, WBC count, RBC count, etc.), as well as general information on patients (age, blood pressure, etc.). 400 patients were tested and recorded, 250 of which had CKD and 150 of which did not. The dataset can be found on the UCI machine learning data repository: <https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease>

I chose this dataset because of my interest in the natural sciences. The variable types and amounts that this dataset provided were also useful for analysis, which made it very attractive for use.

1. **DATA SET DESCRIPTION**

This data set contains 400 samples with 25 columns with various data types(ratio,nominal, and ordinal). A complete listing is shown in Table 1.

**Table 1: Data Types and Missing Data**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variable Name** | **Data Type**  **(data type, pandas dtype)** | **Missing Data**  **(%)** |
| **0** | age | Ratio/float64 | 2.25 |
| **1** | bp | Ratio/float64 | 3.0 |
| **2** | sg | Ordinal/object | 11.75 |
| **3** | al | Nominal/object | 11.5 |
| **4** | su | Nominal/object | 12.25 |
| **5** | rbc | Nominal/object | 38.0 |
| **6** | pc | Nominal/object | 16.25 |
| **7** | pcc | Nominal/object | 1.0 |
| **8** | ba | Nominal/object | 1.0 |
| **9** | bgr | Ration/float64 | 11.0 |
| **10** | bu | Ratio/float64 | 4.75 |
| **11** | sc | Ratio/float64 | 4.25 |
| **12** | sod | Ratio/float64 | 21.75 |
| **13** | pot | Ratio/float64 | 22.0 |
| **14** | hemo | Ratio/float64 | 13.0 |
| **15** | pcv | Ratio/float64 | 17.75 |
| **16** | wbcc | Ratio/float64 | 26.5 |
| **17** | rbcc | Ratio/float64 | 32.75 |
| **18** | htn | Nominal/object | 0.5 |
| **19** | dm | Nominal/object | 0.5 |
| **20** | cad | Nominal/object | 0.5 |
| **21** | appet | Nominal/object | 0.25 |
| **22** | pe | Nominal/object | 0.25 |
| **23** | ane | Nominal/object | 0.25 |
| **24** | class | Nominal/object | 0.0 |

1. **Data Set Summary Statistics**

Statistically significant data on each column of numerical data type (ratio), proportion data on categorical columns, and correlation data relating each numerical variable.

**Table 2: Summary Statistics for Chronic Kidney Disease**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **age** | 391 | 51.48337595907930 | 17.16971408926220 | 2.0 | 42.0 | 55.0 | 64.5 | 90.0 |
| **bp** | 388 | 76.46907216494850 | 13.683637493525300 | 50.0 | 70.0 | 80.0 | 80.0 | 180.0 |
| **bgr** | 356 | 148.0365168539330 | 79.28171423511780 | 22.0 | 99.0 | 121.0 | 163.0 | 490.0 |
| **bu** | 381 | 57.425721784776900 | 50.5030058492225 | 1.5 | 27.0 | 42.0 | 66.0 | 391.0 |
| **sc** | 383 | 3.0724543080939900 | 5.741126066859780 | 0.4 | 0.9 | 1.3 | 2.8 | 76.0 |
| **sod** | 313 | 137.52875399361000 | 10.408752051798800 | 4.5 | 135.0 | 138.0 | 142.0 | 163.0 |
| **pot** | 312 | 4.627243589743590 | 3.1939041765567000 | 2.5 | 3.8 | 4.4 | 4.9 | 47.0 |
| **hemo** | 348 | 12.526436781609200 | 2.9125866088267600 | 3.1 | 10.3 | 12.650 | 15.0 | 17.8 |
| **pcv** | 329 | 38.88449848024320 | 8.990104814740940 | 9.0 | 32.0 | 40.0 | 45.0 | 54.0 |
| **wbcc** | 294 | 8406.122448979590 | 2944.474190410340 | 2200.0 | 6500.0 | 8000.0 | 9800.0 | 26400.0 |
| **rbcc** | 269 | 4.707434944237920 | 1.0253232655721800 | 2.1 | 3.9 | 4.8 | 5.4 | 8.0 |

Table 3a: Proportions for ‘Albumin (‘al’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| 0 | 199 | 49.75 |
| 1 | 44 | 11.0 |
| 2 | 43 | 10.75 |
| 3 | 43 | 10.75 |
| 4 | 24 | 6.0 |
| 5 | 1 | 0.25 |
| ? | 46 | 11.5 |

Table 3b: Proportions for Anemia(‘ane’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 1 | 0.25 |
| no | 339 | 84.75 |
| yes | 60 | 15.0 |

Table 3c: Proportions for Appetite (‘appet’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 1 | 0.25 |
| good | 317 | 79.25 |
| poor | 82 | 20.5 |

Table 3d: Proportions for Bacteria (‘ba’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 4 | 1.0 |
| notpresent | 374 | 93.5 |
| present | 22 | 5.5 |

Table 3e: Proportions for Coronary Artery Disease (‘cad’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 2 | 0.5 |
| no | 364 | 91.0 |
| yes | 34 | 8.5 |

Table 3f: Proportions for Class (‘class’) (n=400) – if the patient has ckd or not

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ckd | 250 | 62.5 |
| notckd | 150 | 37.5 |

Table 3g: Proportions for Diabetes Mellitus (‘dm’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 2 | 0.5 |
| no | 261 | 65.25 |
| yes | 137 | 34.25 |

Table 3h: Proportions for Hypertension (‘htn’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 2 | 0.5 |
| no | 251 | 62.75 |
| yes | 147 | 36.75 |

Table 3i: Proportions for Pus Cell (‘ps) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 65 | 16.25 |
| abnormal | 76 | 19.0 |
| normal | 259 | 64.75 |

Table 3j: Proportions for Pus Cell Clumps (‘pcc’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 4 | 1.0 |
| notpresent | 354 | 88.5 |
| present | 42 | 10.5 |

Table 3k: Proportions for Pedal Edema (‘pe’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 1 | 0.25 |
| no | 323 | 80.75 |
| yes | 76 | 19.0 |

Table 3l: Proportions for Red Blood Cells (‘rbc) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| ? | 152 | 38.0 |
| abnormal | 47 | 11.75 |
| normal | 201 | 50.25 |

Table 3m: Proportions for Specific Gravity (‘sg) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| 1.005 | 7 | 1.75 |
| 1.010 | 84 | 21.0 |
| 1.015 | 75 | 18.75 |
| 1.020 | 106 | 26.5 |
| 1.025 | 81 | 20.25 |
| ? | 47 | 11.75 |

Table 3n: Proportions for Sugar (‘su’) (n=400)

|  |  |  |
| --- | --- | --- |
| **Category** | **Frequency** | **Proportion (%)** |
| 0 | 290 | 72.5 |
| 1 | 13 | 3.25 |
| 2 | 18 | 4.5 |
| 3 | 14 | 3.50 |
| 4 | 13 | 3.25 |
| 5 | 3 | 0.75 |
| ? | 49 | 12.25 |

Table 4a: Correlation Table/Tables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **age** | **bp** | **bgr** | **bu** |
| **age** | 1.0 | 0.15947969344545300 | 0.2449921996169060 | 0.1969848707760300 |
| **bp** | 0.1594796934454530 | 1.0 | 0.1601934618058240 | 0.1885172453103120 |
| **bgr** | 0.2449921996169060 | 0.16019346180582400 | 1.0 | 0.1433220200458160 |
| **bu** | 0.1969848707760300 | 0.18851724531031200 | 0.1433220200458160 | 1.0 |
| **sc** | 0.1325308652423880 | 0.1462220201137380 | 0.1148749998409420 | 0.5863678207097760 |
| **sod** | -0.100045983075957 | -0.1164220357241530 | -0.267847586157687 | -0.323054235289353 |
| **pot** | 0.0583771200100528 | 0.07515106557193240 | 0.0669657945529565 | 0.3570490813524030 |
| **hemo** | -0.192928338968053 | -0.3065398884398660 | -0.306189281504202 | -0.610360278485781 |
| **pcv** | -0.242119409268333 | -0.3263193050079200 | -0.3013847434100510 | -0.6076213553320020 |
| **wbcc** | 0.1183385204297380 | 0.029753296531023400 | 0.15001468216754800 | 0.05046201708217970 |
| **rbcc** | -0.268896285187890 | -0.2619358101479630 | -0.2815407123595220 | -0.5790865252805200 |

Table 4b: Correlation Table/Tables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **sc** | **sod** | **pot** | **hemo** |
| **age** | 0.1325308652423880 | -0.1000459830759570 | 0.0583771200100528 | -0.192928338968053 |
| **bp** | 0.1462220201137380 | -0.11642203572415300 | 0.0751510655719324 | -0.306539888439866 |
| **bgr** | 0.1148749998409420 | -0.2678475861576870 | 0.0669657945529565 | -0.306189281504202 |
| **bu** | 0.5863678207097760 | -0.32305423528935300 | 0.3570490813524030 | -0.610360278485781 |
| **sc** | 1.0 | -0.6901578920579920 | 0.326107131869241 | -0.401669624435810 |
| **sod** | -0.6901578920579920 | 1.0 | 0.0978867154783519 | 0.365182652257529 |
| **pot** | 0.326107131869241 | 0.09788671547835190 | 1.0 | -0.133746041764821 |
| **hemo** | -0.4016696244358100 | 0.36518265225752900 | -0.1337460417648210 | 1.0 |
| **pcv** | -0.4041930644204250 | 0.37691355206594400 | -0.1631822837746950 | 0.8953817669928050 |
| **wbcc** | -0.0063899103348037 | 0.007277275891274250 | -0.1055762183345210 | -0.169413065804814 |
| **rbcc** | -0.4008519768697740 | 0.34487348202530800 | -0.1583093171278070 | 0.7988802467445650 |

Table 4c: Correlation Table/Tables

|  |  |  |  |
| --- | --- | --- | --- |
|  | **pcv** | **wbcc** | **rbcc** |
| **age** | -0.2421194092683330 | 0.1183385204297380 | -0.2688962851878900 |
| **bp** | -0.3263193050079200 | 0.029753296531023400 | -0.2619358101479630 |
| **bgr** | -0.3013847434100510 | 0.15001468216754800 | -0.2815407123595220 |
| **bu** | -0.6076213553320020 | 0.05046201708217970 | -0.5790865252805200 |
| **sc** | -0.4041930644204250 | -0.00638991033480371 | -0.4008519768697740 |
| **sod** | 0.37691355206594400 | 0.007277275891274250 | 0.34487348202530800 |
| **pot** | -0.1631822837746950 | -0.10557621833452100 | -0.1583093171278070 |
| **hemo** | 0.8953817669928050 | -0.16941306580481400 | 0.7988802467445650 |
| **pcv** | 1.0 | -0.19702236172906500 | 0.7916252713729910 |
| **wbcc** | -0.1970223617290650 | 1.0 | -0.1581627625264120 |
| **rbcc** | 0.7916252713729910 | -0.15816276252641200 | 1.0 |

Chart

Description automatically generated

**Figure 1: Correlation Heatmap**

1. **DATA SET GRAPHICAL EXPLORATION**

For most of the data, when focusing on the differences between ckd and non-ckd patients, the ckd group had a much larger range and interquartile range across almost all continuous variables. Whether the mean was elevated or lowered compared to the non-ckd group, the ckd group was more dispersed.

*Chart, box and whisker chart

Description automatically generated*

**Figure 2: Comparison of Blood Glucose Random in CKD group, and control group**

**Chart, box and whisker chart

Description automatically generated**

**Figure 3: Comparison of Blood Urea in CKD group, and control group**

*Chart, box and whisker chart

Description automatically generated*

**Figure 4: Comparison of Serum Creatinine in CKD group, and control group**

Notice the dispersiveness in the CKD groups of figures 2, 3, and 4. This signifies that these biomarkers become very unbalanced in patients with CKD. It is likely that other patients with CKD will have a hard time keeping these variables within a normal range, which means that these can be used as screening markers for CKD detection.

**Calendar

Description automatically generated**

**Figure 5: Pairplot of CKD data**

The purpose of producing Figure 5 was to get a quick look at trends and what variables look to be related. Some plots that look nice to expand upon are hemo vs. pcv and hemo vs. rbcc. These would be good to observe together because they look to have similar trends, and this would be good to go deeper into.

*Chart, scatter chart

Description automatically generated*

**Figure 6: Hemoglobin relationship with Packed Cell Volume (a) and Red Blood Cell Count (b).**

These two variables had very similar patterns with hemoglobin. Packed Cell Volume does not have a unit because it is the ratio of volume of blood before centrifugation to volume of packed blood cells in the pellet after centrifugation. This is very similar to red blood cell count, but different because it involves all cells and solid matter in the blood serum.

Chart, bar chart, histogram

Description automatically generated

**Figure 7: Number of patients in each serum albumin group**

1. **SUMMARY OF FINDINGS**

The dataset has shown that many factors and biomarkers are significantly changed in individuals with Chronic Kidney Disease. A lot of blood related disorders come about, from hypertension to blood urea levels and serum creatinine. Though many of the mean values differ from the healthy mean, the biggest change is simply the dispersion of the values. It seems that CKD causes a lot of unreliability and it would be hard to come up with specific measurable values that signify its onset. Instead, it would be more useful to understand that any of these biomarkers outside the normal range could be a sign of CKD, and though it isn’t an automatic diagnosis, it can be put on the radar for future testing.