Dataset Name: Chronic Kidney Disease

Group Name: Thu-12 On Campus/cloud: On Campus

STUDENT ID	STUDENT FULL NAME	Individual contribution *
217452473	RONIT BHADRESHKUMAR SHAH	5
217508259	GAUTAM BEDI	5
217443566	KEVIN MANOJKUMAR PATEL	5
217428992	XIAOPENG CHEN	4

* 5 – Contributed significantly, attended all meetings
4 – Partial contribution, attended all meetings
3 – Partial contribution, attended few meetings
1 – No contribution, attended few meetings
0 – No contribution, did not attended any meetings

NOTE: IF ANY OF THE CELLS IN INDIVIDUAL CONTRIBUTION MARK IS EMPTY ALL STUDENTS WOULD GET 3 MARK BY DEFAULT

Section 1: Introduction and getting to know your data (max 2 pages)

1. Data set Inspection

1.1 Data collection

The aim of collecting this data is to classify whether the patient is suffering from chronic kidney disease or not. The chronic kidney disease data is a multivariate dataset, consisting of 25 different attributes (including classification/target) such as blood pressure, diabetes mellitus, pedal edema, appetite etc. in the time interval of 2 months period for around 400 patients. This data was collected from hospitals in Tamil Nadu, India. Dr. P. Soundarapandian open-sourced this dataset in 2015 for the public.

id	age	blood pressure	specific gravity	albumin	sugar	red blood cells	pus cell	pus cell clumps	bacteria	blood glucose random	...
0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...
1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...
2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	...
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	...
4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	...

5 rows × 25 columns

The data consists of both numerical and nominal data attributes. There are in total 11 numerical data attributes and 14 nominal data attributes. Out of the 400 patients we have complete record of 158 patients while other 242 patients have some missing values in their records. When we first looked at our data, we were not familiar with some of the attributes like sg, bu, dm, cad, etc. so we researched about them and found out their meanings. The target of our data is to predict that whether a person is suffering from chronic kidney diseases or not based on input features.

1.2 Data Initial observations

In our initial observations, we figured out the number of missing values in the dataset for every column. For instance, bp - 12 missing values, sg - 47 missing values, al - 46 missing values, etc. We also observed that we have three different data types such as decimal (including integers), boolean, and string. Out of 25 data attributes, 11 are decimal, 9 are string, and 5 are boolean.

id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	48	80	1.02	1	0		normal	notpresen	notpresen	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	7	50	1.02	4	0		normal	notpresen	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
2	62	80	1.01	2	3	normal	normal	notpresen	notpresen	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
3	48	70	1.005	4	0	normal	abnormal	present	notpresen	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd

After going through our data, we observed that:

- ⇨ Most of the values for ‘al’ and ‘su’ are 0
- ⇨ Most of the values for ‘rbc’ and ‘pc’ are “normal”
- ⇨ Most of the values for ‘pcc’ and ‘ba’ are “notpresent”

- ⇨ Most of the values for ‘htn’, ‘dm’, ‘cad’, ‘pe’, and ‘ane’ are “no”
- ⇨ Most of the values for ‘appet’ is “good”
- ⇨ The ‘rbc’ column has maximum number of missing values which is 152
- ⇨ We have “?” instead of values in some columns

1.3 Data Exploratory Analysis Plan

Task	Description
Data Cleaning	In this part, we will focus on handling missing or physically impossible values for each column, removing extra tabs or whitespaces, correcting data types, sanity checks, outlier checks and check for typos
Pearson Correlation	Examining the Pearson correlation between all the features to check for linear relationships
Exploratory and Confirmatory Data Analysis	In this part, we will analyse the relationship of variables/features with respect to other variables/features: ⇨ Packed cell volume and hemoglobin ⇨ blood urea and hemoglobin ⇨ blood pressure and class ⇨ Hypertension, coronary artery disease, diabities with respect to class ⇨ Analysing body diseases grouped by age groups ⇨ Analysing of 'blood pressure', 'sodium', 'potassium', 'hemoglobin' with each other ⇨ Average 'rbc count', 'wbc count', 'sodium', 'potassium', 'blood glucose' levels based on age group ⇨ Analysing peoples' appetite (mood) with respect to CKD ⇨ Anemia and rbc count ⇨ Hypertension and Serum creatinine
Conclusion	Deriving conclusions on basis of exploratory data analysis

Section 2: Exploratory Data Analysis and Results (max 7 pages)

1. Data Cleaning

1.1 Handling missing values (Imputation)

The most important aspect of data preparation is how we treat/handle the null data values of the observations. As significant chunk of dataset is nominal data, we could not use the standard statistical methods like mean, median to replace them.
To get a better understanding of missing/nullified data, we firstly visualised the missing values for each data attribute using heat-map and stacked bar chart. From those observations, we found out that 'red blood cell', 'sodium', 'potassium', 'red blood cell count' and 'white blood cell count' had more than 20% missing data. Apart from those attributes, we decided to replace missing values in floating point attributes using mean(), provided that their data distribution was almost normal (i.e. not skewed) (which was verified by plotting histograms). In addition, statistical method - median() was used for floating point attributes with skewed distribution for imputation purposes. The attributes with more than 20% missing data were kept as-is to avoid compromising the integrity of the data.

1.2 Removing extra-whitespaces

After completing data set inspection, we could see that there are few columns with tabs/leading and trailing spaces. So, to make sure that our data has no extra whitespace we used strip() function to each column with string data type.

```
df = df.applymap(lambda column:column.strip() if type(column) == str else column)
```

1.3 Checking Data Types

The next thing which we did was checked for the data attributes data types. We have corrected data types for almost all the columns as recommended by the data attribute information mentioned in the original dataset source website. We divided our dataset into three data types which are int64, float64, and category.

⇨ Changed 'age' to int64

```
df['age'] = df['age'].astype('int64')
```

⇨ Changed 'specify gravity' to category

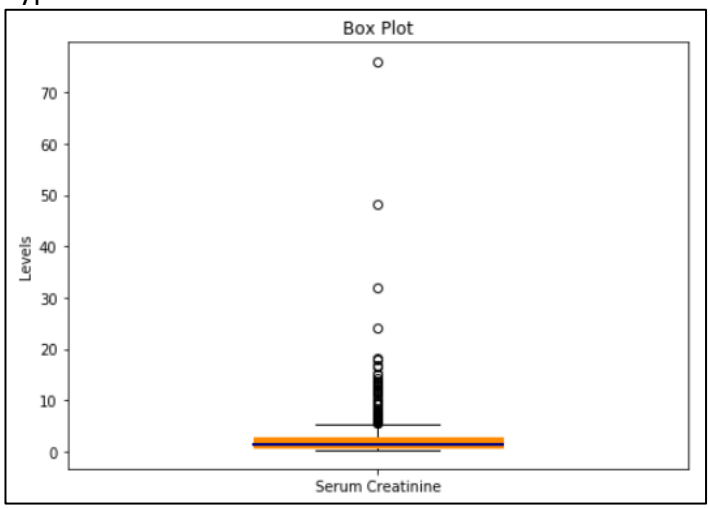
```
df['specific gravity'] = df['specific gravity'].astype(pd.api.types.CategoricalDtype(categories=specific_gravity_category))
```

1.4 Checking typos

We have also checked the complete data for any typos, physically impossible values and incorrect capitalisation if there are any. For instance, typo like '\t?' value was found in 'white blood cell count'.

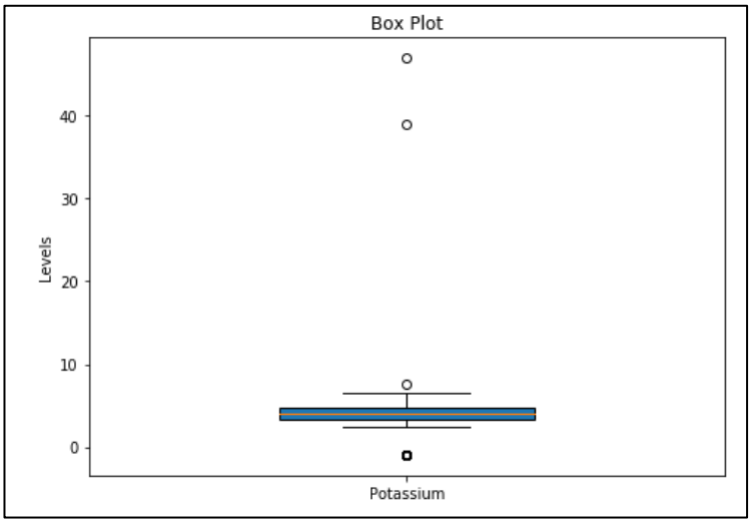
```
white blood cell count :  ['7800' '6000' '7500' '6700' '7300' nan '6900' '9600' '12100' '4500'
 '12200' '11000' '3800' '11400' '5300' '9200' '6200' '8300' '8400' '10300'
 '9800' '9100' '7900' '6400' '8600' '18900' '21600' '4300' '8500' '11300'
 '7200' '7700' '14600' '6300' '\t6200' '7100' '11800' '9400' '5500' '5800'
 '13200' '12500' '5600' '7000' '11900' '10400' '10700' '12700' '6800'
 '6500' '13600' '10200' '9000' '14900' '8200' '15200' '5000' '16300'
 '12400' '\t8400' '10500' '4200' '4700' '10900' '8100' '9500' '2200'
 '12800' '11200' '19100' '\t?' '12300' '16700' '2600' '26400' '8800'
 '7400' '4900' '8000' '12000' '15700' '4100' '5700' '11500' '5400' '10800'
 '9900' '5200' '5900' '9300' '9700' '5100' '6600']
```

We found out that there are some physically impossible values like very high value for bp, high potassium level, high serum creatinine level etc. which can be strong indications of a typo error.



1.5 Checking for outliers

We have checked for outlies in all the columns and we found that most of the columns have outliers present in them, so we visualised them using boxplots. One or two columns has outliers which may be typing error (impossible values), for example, bp = 180 mm/Hg which is a medical emergency and for serum creatinine there are few values which is more than 10 mgs/dl and one of the value is 76 mgs/dl which is extremely dangerous and can lead to kidney damage. Another outlier is present in potassium which is 47 mEq/L and it is very dangerous because it is far outside the range of normal body potassium levels.



2. Data Exploration Analysis

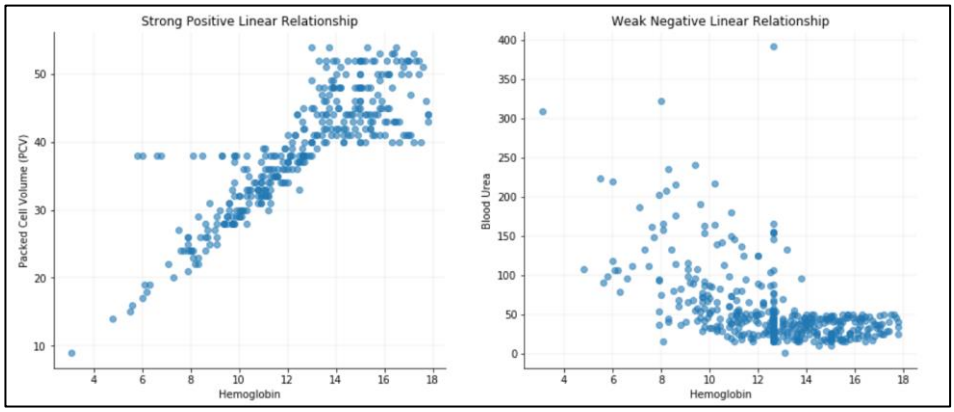
2.1 Examine Pearson correlation between features

We are calculating Pearson correlation coefficients between all the features to find out the linear relationship between each other features.

	age	blood pressure	blood glucose random	blood urea	serum creatinine	sodium	potassium	hemoglobin	packed cell volume	white blood cell count	red blood cell count
age	1.00	0.15	0.23	0.19	0.13	-0.02	0.03	-0.17	-0.21	-0.01	-0.03
blood pressure	0.15	1.00	0.15	0.18	0.14	-0.07	0.02	-0.28	-0.29	-0.04	-0.08
blood glucose random	0.23	0.15	1.00	0.12	0.07	-0.13	-0.03	-0.25	-0.26	0.01	-0.11
blood urea	0.19	0.18	0.12	1.00	0.58	0.07	0.32	-0.54	-0.53	-0.02	-0.16
serum creatinine	0.13	0.14	0.07	0.58	1.00	-0.04	0.15	-0.34	-0.35	-0.08	-0.15
sodium	-0.02	-0.07	-0.13	0.07	-0.04	1.00	0.64	0.14	0.13	0.20	0.37
potassium	0.03	0.02	-0.03	0.32	0.15	0.64	1.00	-0.02	-0.05	0.07	0.17
hemoglobin	-0.17	-0.28	-0.25	-0.54	-0.34	0.14	-0.02	1.00	0.86	0.09	0.46
packed cell volume	-0.21	-0.29	-0.26	-0.53	-0.35	0.13	-0.05	0.86	1.00	0.03	0.41
white blood cell count	-0.01	-0.04	0.01	-0.02	-0.08	0.20	0.07	0.09	0.03	1.00	0.65
red blood cell count	-0.03	-0.08	-0.11	-0.16	-0.15	0.37	0.17	0.46	0.41	0.65	1.00

With the help of this data, we figured out that either two features are positively, negatively or weakly correlated with each other. Pearson correlation is always between -1 to 1 and if the correlation is between -1 to 0 that means it is negatively correlated, if the correlation is 0 that means it is not correlated, if the correlation is between 0 to 1 that means it is positively correlated.

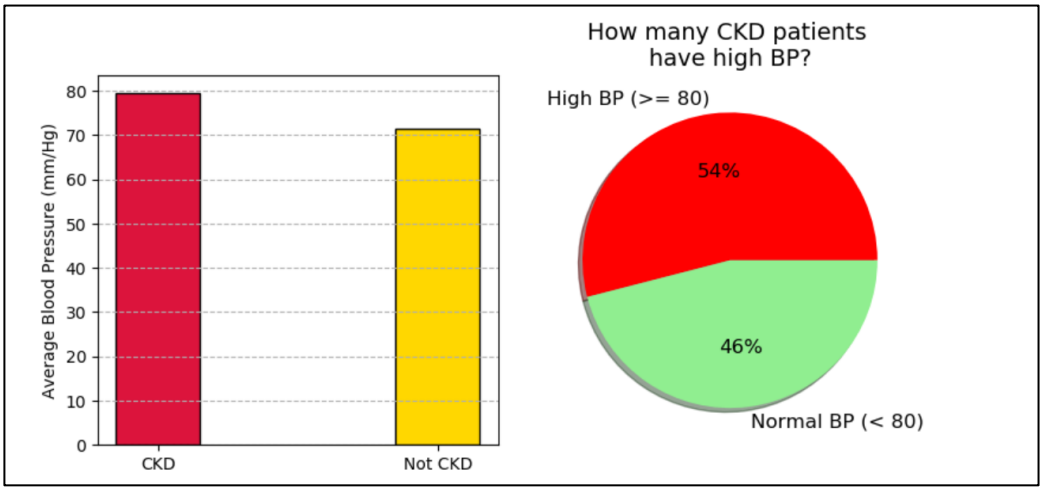
For example, 'haemoglobin' (measures the amount of protein in RBC cells) and 'packed cell volume' (measures the proportion of blood made up of cells) is strong positively correlated (as packed cell volume increases, RBC count increases, which in turn increase the haemoglobin levels) while 'haemoglobin' and 'blood urea' is weak negatively correlated.



2.2 Analysing relationship between 'blood pressure' and 'class'

"*The second leading cause of CKD is high blood pressure. Blood vessels in your kidney can be damaged by high blood pressure. Out of 5 adults almost 1 adult has CKD due to high blood pressure*" (Information et al., 2020)

In order to analyse this relationship, we plotted a bar graph depicting people count suffering from 'CKD' or 'NOTCKD' with their average blood pressure. We also plotted a pie chart to show how many people with CKD have high blood pressure greater than 80 mm/Hg.



As a result, we can see that average blood pressure of a patient suffering from CKD has higher BP as compared to someone who does not. 54% people of those who have CKD also have high blood pressure while 46% people of those who have CKD have normal blood pressure. More than half of the patient with CKD have BP values greater than 80, which is in line with our hypothesis.

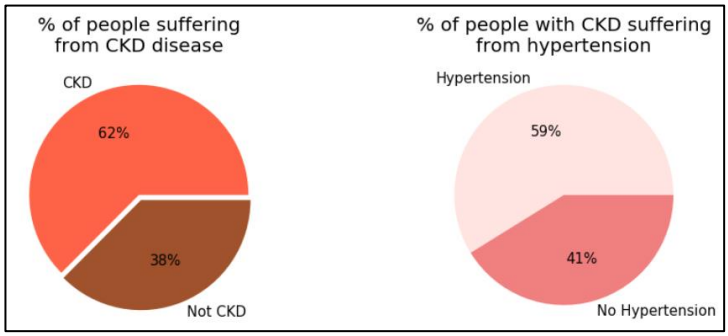
2.3 Analysing relationship between 'hypertension', 'coronary artery disease', 'diabities' with respect to 'class' columns

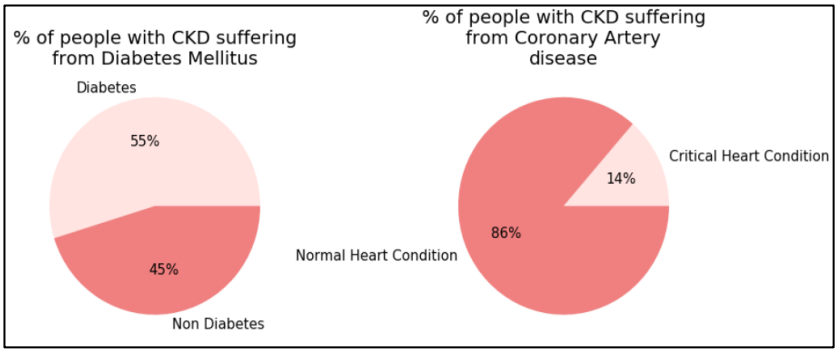
"*Another leading cause of CKD is diabities. The blood vessels present in your kidney can be damaged from high blood sugar. Out of 3 adults almost 1 adult has CKD due to Diabities*" (Information et al., 2020)

"*Hypertension also known as high blood pressure is also the leading cause of CKD*" (Information et al., 2020)

"*Coronary artery disease is also the cause for CKD*" (Aline Milane, 2020)

To analyse, we have plotted a pie chart to show how many percentages of people with CKD also have diabities, hypertension, and coronary artery disease.

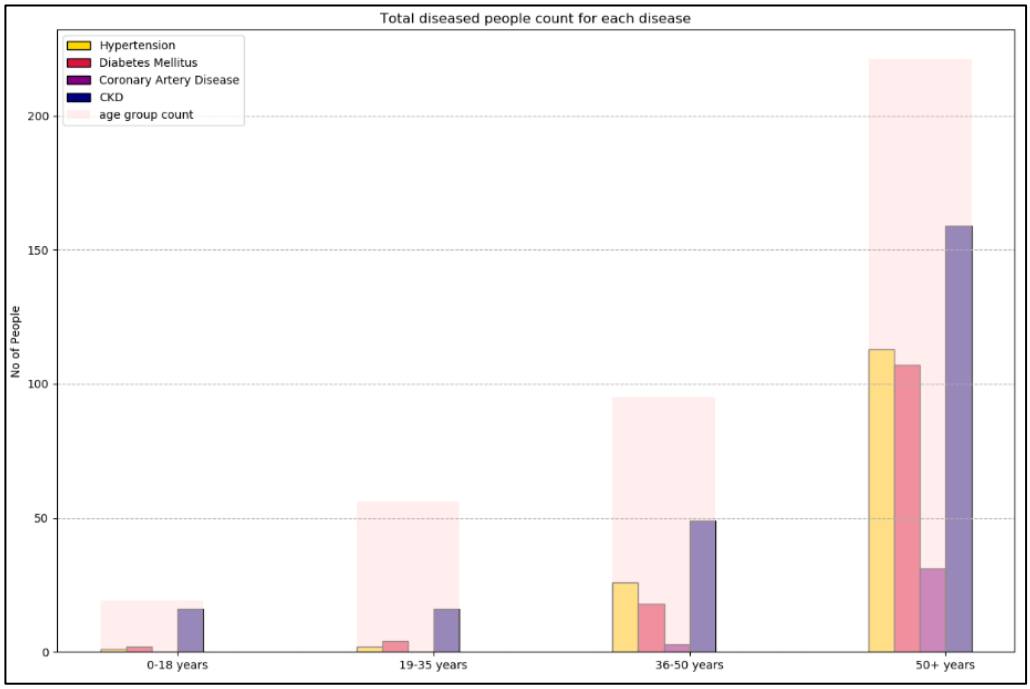




We can clearly see that 59% people with hypertension are suffering from CKD whereas 41% people with no hypertension are suffering from CKD, 55% people with diabities are suffering from CKD whereas 45% people with non diabities are suffering from CKD and only 14% people with critical heart condition are suffering from CKD whereas 86% people with normal heart condition are suffering from CKD. Diabetes, Hypertension is in line with our hypothesis whereas coronary artery diseases is not in line with our hypothesis according to our data set.

2.4 Analysing body diseases grouped by age groups (hypertension, coronary artery disease, diabities with respect to class columns)

In order to analyse we have plotted a grouped bar graph for total count of diseased people grouped by age group as 0-18 years, 19-35 years, 36-50 years and 50+ years.



As we can clearly see from this graph that between 0-18 years there are less than 20 people having CKD with no people suffering from coronary artery disease and very few suffer from hypertension and diabities. Between 19-35 years there are also less than 20 people having CKD with no people suffering from coronary artery diseases and very few suffer from hypertension and diabities. Between 36-50 years there are less than 60 people having CKD with most of them suffering from hypertension and few of them suffer from diabities and coronary artery diseases. Between 50+ years almost more than 160 people are having CKD with most of them are suffering from hypertension and diabetes and very few of them are suffering from coronary artery disease.

Insights: From 0-18 years we could conclude that CKD is very prevalent, among 35+ age group more than 50% of people are suffering from CKD. Among youngsters (19-35 years) the disease rate is the lowest as body immunity peaks at this age group.

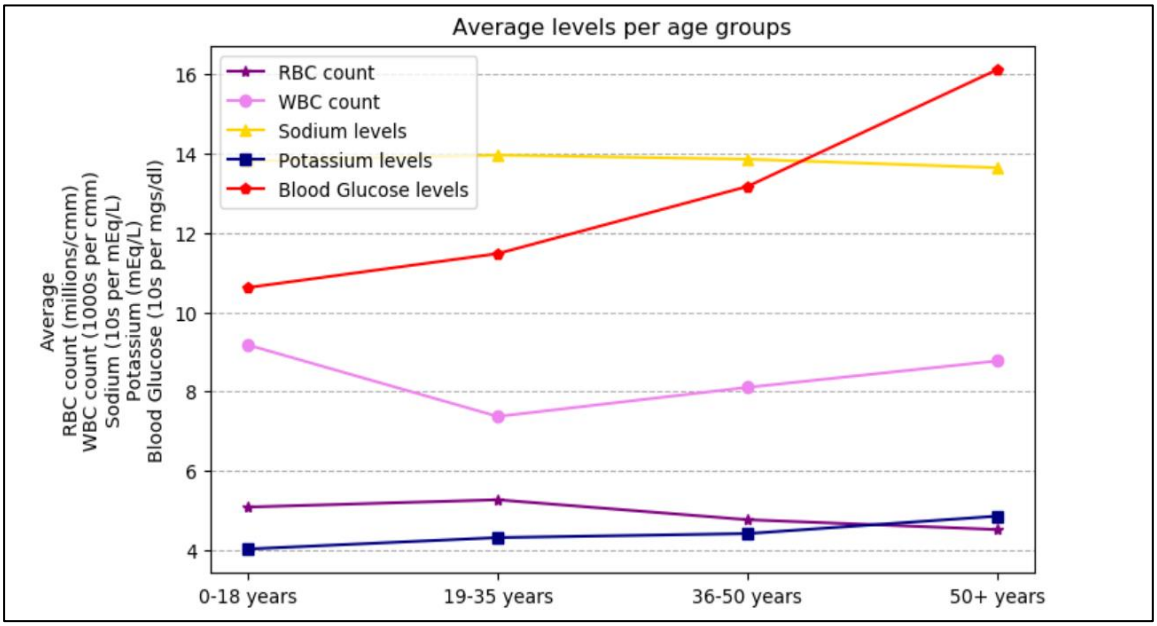
2.5 Analysing linear relationships of 'blood pressure', 'sodium', 'potassium', 'hemoglobin' columns with each other

In order to analyse the linear relationship of these features with each other we are using pairplot which is part of seaborn which is python library based on matplotlib. We can clearly see in the pairplot that some features are positively correlated with each other and some features are negatively correlated with each other.

```
columns = ['blood pressure', 'sodium', 'potassium', 'hemoglobin', 'age group']
sns.pairplot(df_binned[columns])
```

2.6 Average level of 'rbc count', 'wbc count', 'sodium', 'potassium', 'blood glucose' based on age group

In order to show avg count of all these features based on age group we used line graph to visualise changes in levels categorised by age group in 4 different categories which is 0-18 years, 19-35 years, 36-50 years, 50+ years.



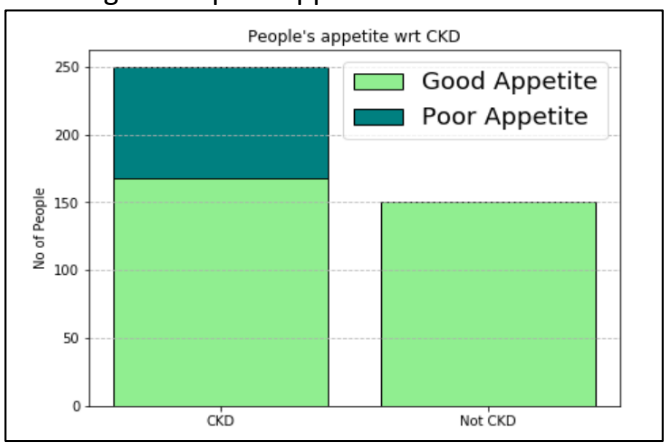
As we can clearly see from the figure that the avg level for potassium increases from 4-5 mEq/L as age group increase. Avg level for RBC count is slightly increasing from 5- 5.5 millions/cmm for age group of 0-18 years to 19-35 years and then decreasing from 5.5 to 4.2 millions/cmm for age group of 36-50 years to 50+ years. Avg level for WBC count is drastically decreasing from almost 9500 to 7800 (in 1000s per cmm) for age group of 0-18 years to 19-35 years and then increasing from 7800 to 8800 (in 1000s per cmm) for age group of 36-50 years to 50+ years. Avg level for blood glucose is increasing from 105-160 (in 10s per mgs/dl) as age group increase. Avg level for Sodium is slightly increasing from 137- 140 (in 10s per mEq/L) for age group of 0-18 years to 19-35 years and then decreasing from 140 to 135 (in 10s per mEq/L) for age group of 36-50 years to 50+ years.

Insights: WBC count is the least for age group 19-35 while RBC count is the highest as people are the fittest and healthiest in these times. As 50+ age group has alarmingly high levels of blood glucose, they are more prone to type 2 diabetes.

2.7 Analysing people's appetite (mood) with respect to CKD

"*One of the symptoms of CKD is poor appetite which becomes the cause for CKD*" (Information et al., 2020)

In order to analyse the appetite, we have used bar graph to show people count suffering from CKD have good or poor appetite.

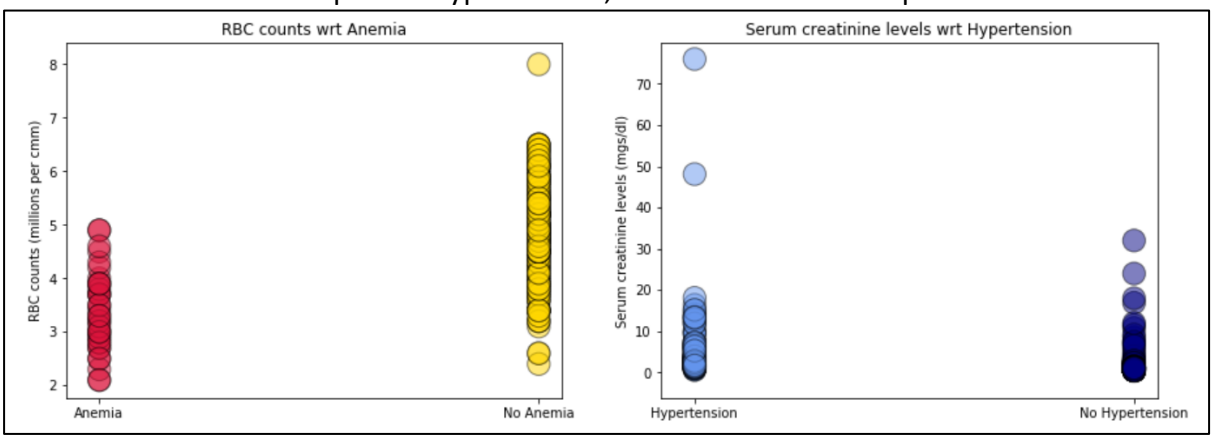


As a result, we can clearly see that people with NOTCKD has a good appetite and more than 150 people with CKD has a good appetite and almost 90-95 people with CKD has a poor appetite which is in line with our hypothesis. However, the interesting pattern in this analysis is that if a person is having a poor appetite then there is a 100% chance that he/she will have CKD.

2.8 Analysing relationship b/w 'anemia' vs 'rbc count' columns and 'hypertension' vs 'serum creatinine' columns

"*Anemia is condition where the number of red blood cells drops below average levels*" ((COVID-19) et al., 2020)

In order to analyse relationship between rbc count with respect to anemia and serum creatinine with respect to hypertension, we have used scatter plots.



As you can clearly see in the figure that rbc count ranges from 2-5.2 (millions/cmm) (red colour) when patient is suffering from anemia and it ranges from 2-8 (millions/cmm) (yellow colour) when patient is not suffering from anemia.
Serum creatinine ranges from 0.5-76 mgs/dl (light blue colour) when patient is suffering from hypertension and it ranges from 0.5-40 mgs/dl (dark blue colour) when patient is not suffering from hypertension.

Insights: Average rbc count is lower in individuals suffering from anemia, which is in line with the hypothesis.

Section 3: Conclusions (max 1 page)

1. Main Observations

The main observation after preforming exploratory data analyses are as follows:

- The interesting pattern between coronary artery diseases, diabetes and hypertension is that if any one of them is 'yes' in the patient, then that patient is suffering from CKD
- If an individual is not suffering from CKD, then that person is having a good appetite (100% chance as per the data) but if a person is suffering from CKD then there is 30-35% chance that the person is having a poor appetite
- Out of all the people in the age group of 0-18 years, more than 90% of the people are suffering from CKD based on our dataset
- Blood glucose level is almost twice for age group of 50+ people as compared to 0-18 years, which is the reason why senior citizens are more prone to type 2 diabetes

2. Problems for Machine Learning Project

As the dataset was originally donated to the public with the aim to promote analysing the root causes for chronic kidney disease, which is a prevalent disease in India, and uncover insights and information on early-stage indications of people suffering from it, one of the major problem that could be considered would be the prediction of chronic kidney disease in humans. It is a binary classification algorithm in Machine Learning. Algorithms like Decision Tree, Random Forest, Logistic Regression and Support Vector Machines could be potential ML algorithms that could be used for predicting CKD.

Section 4: References (max 1 page)

1. Aline Milane, A., 2020. 'Association Of Coronary Artery Disease And Chronic Kidney Disease In Lebanese Population.', retrieved 25 April 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4658979/

2. (COVID-19), C., Health, E., Disease, H., Disease, L., Management, P., Conditions, S., Problems, S., Disorders, S., Checker, S., Blogs, W., Boards, M., Answers, Q., Guide, I., Doctor, F., A-Z, C., A-Z, S., Medications, M., Identifier, P., Interactions, C., Drugs, C., Pregnant, T., Management, D., Obesity, W., Recipes, F., Exercise, F., Beauty, H., Balance, H., Relationships, S., Care, O., Health, W., Health, M., Well, A., Sleep, H., Teens, H., Pregnant, G., Trimester, F., Trimester, S., Trimester, T., Baby, N., Health, C., Vaccines, C., Kids, R., Cats, H., Dogs, H., Here, G., Now, C., Home, H., Surfaces?, H., Spread, S., Boards, M., Blogs, W., Center, N. and Guides, A., 2020. 'Anemia.', retrieved 25 April 2020, https://www.webmd.com/a-to-z-guides/understanding-anemia-basics

3. Information, H., Disease, K., (CKD), C., Disease?, W., Disease?, W., Center, T. and Health, N., 2020. 'What Is Chronic Kidney Disease? | NIDDK.', retrieved 25 April 2020, https://www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd/what-is-chronic-kidney-disease

4. Medium. 2020. 'Customizing Plots With Python Matplotlib.', retrieved 24 April 2020, https://towardsdatascience.com/customizing-plots-with-python-matplotlib-bcf02691931f