Chronic Kidney Disease

By Elias Fedai

Bellevue University

Variables

- **Specific Gravity**- is the density of a solution substance in a ratio to a known standard density.
- **Albumin**- Protein responsible for many functions. One being, ensuring fluid doesn’t leak out of blood vessels.
- **Hemoglobin**- Protein responsible for moving/transporting oxygen in blood.
- **Red blood Cell**- blood cells that carries oxygen throughout the body.
- **Blood Pressure**- Is the force of how strong blood is pushing against the walls of blood vessels.
- **Diagnosis**- this variable indicates if the patient has or doesn’t have chronic kidney disease.

Central Tendency

	Specific Gravity	Albumin	Hemoglobin	Blood Pressure	Red Blood Cell count	Diagnosis
count	233.000000	233.000000	233.000000	233.000000	233.000000	233.000000
mean	1.018584	0.854077	13.185408	75.493562	4.749785	0.446352
std	0.005736	1.375648	2.877102	11.738575	1.004049	0.498184
min	1.005000	0.000000	3.100000	50.000000	2.100000	0.000000
25%	1.015000	0.000000	11.100000	70.000000	4.000000	0.000000
50%	1.020000	0.000000	13.700000	80.000000	4.800000	0.000000
75%	1.025000	2.000000	15.400000	80.000000	5.500000	1.000000
max	1.025000	5.000000	17.800000	110.000000	8.000000	1.000000

Note: Good overall spread, blood pressure has a rather high standard deviation.

Correlation

	Specific Gravity	Albumin	Hemoglobin	Blood Pressure	Red Blood Cell count	Diagnosis
Specific Gravity	1.000000	-0.561656	0.652512	-0.296861	0.579624	-0.773388
Albumin	-0.561656	1.000000	-0.693287	0.282081	-0.579846	0.692950
Hemoglobin	0.652512	-0.693287	1.000000	-0.325872	0.787477	-0.796864
Blood Pressure	-0.296861	0.282081	-0.325872	1.000000	-0.261385	0.411777
Red Blood Cell count	0.579624	-0.579846	0.787477	-0.261385	1.000000	-0.696941
Diagnosis	-0.773388	0.692950	-0.796864	0.411777	-0.696941	1.000000

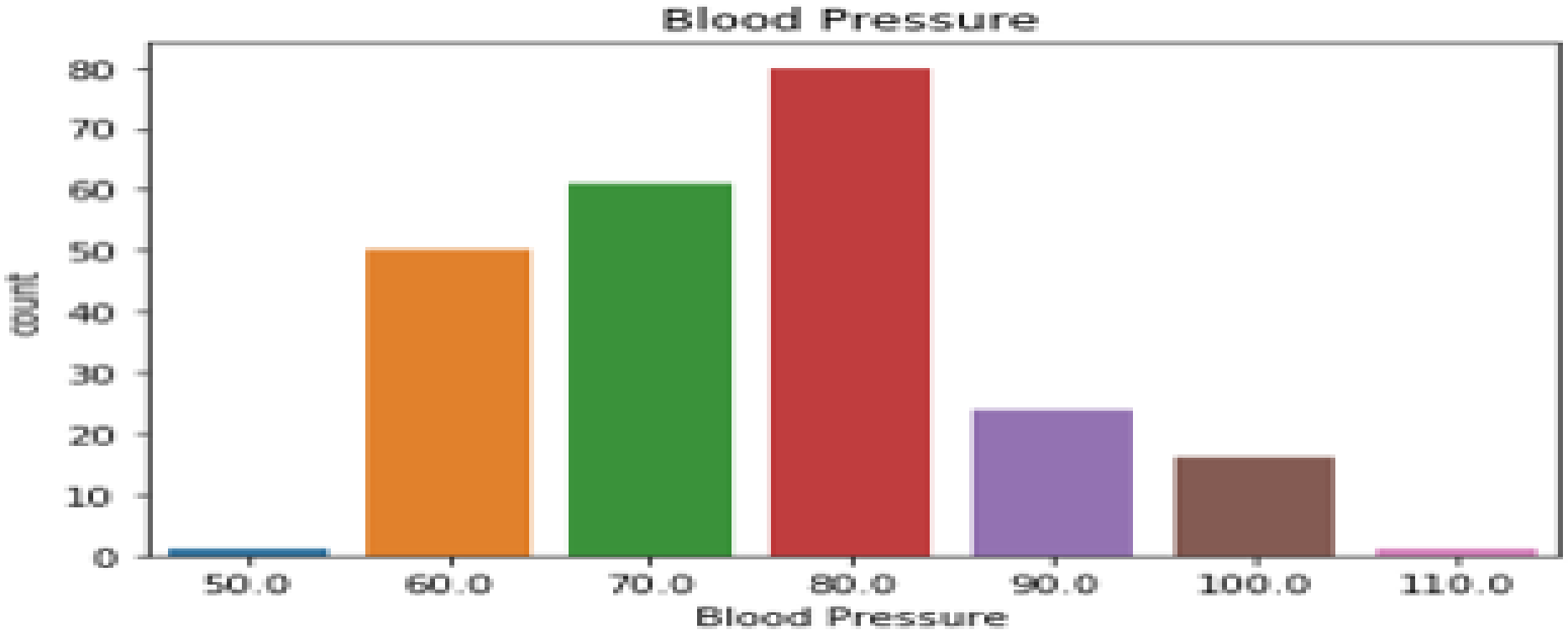
Note: Our diagnosis variable appears to have pretty good correlations with all variables except blood pressure. Hemoglobin and red blood cell count also correlate well.

Covariance

	Specific Gravity	Albumin	Hemoglobin	Blood Pressure	Red Blood Cell count	Diagnosis
Specific Gravity	0.000033	-0.004432	0.010768	-0.019988	0.003338	-0.002210
Albumin	-0.004432	1.892408	-2.743949	4.555091	-0.800894	0.474898
Hemoglobin	0.010768	-2.743949	8.277717	-11.005698	2.274824	-1.142166
Blood Pressure	-0.019988	4.555091	-11.005698	137.794139	-3.080713	2.408058
Red Blood Cell count	0.003338	-0.800894	2.274824	-3.080713	1.008114	-0.348611
Diagnosis	-0.002210	0.474898	-1.142166	2.408058	-0.348611	0.248187

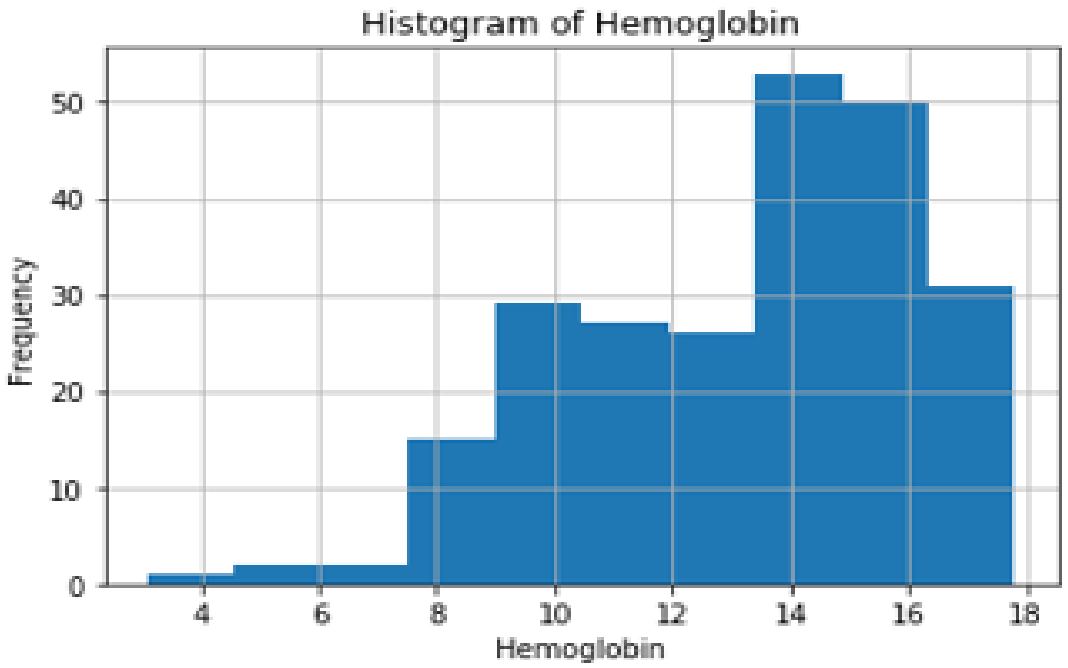
Note: Specific gravity, Hemoglobin, RBC count display a negative relationship with our diagnosis. While all others display a positive relationship.

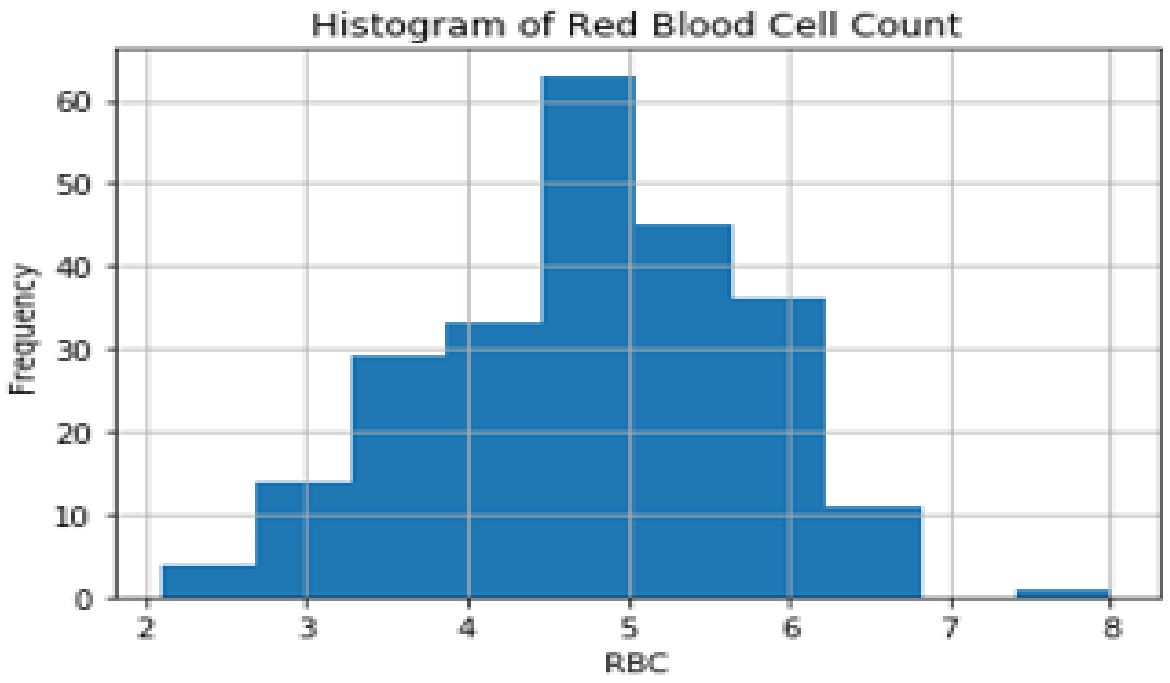
Blood Pressure Frequency



Note: Appears majority of participants appear to have 80mm/Hg blood pressure. Not much correlation or relationship between blood pressure and the other variables. No data given to indicate diastolic or systolic.

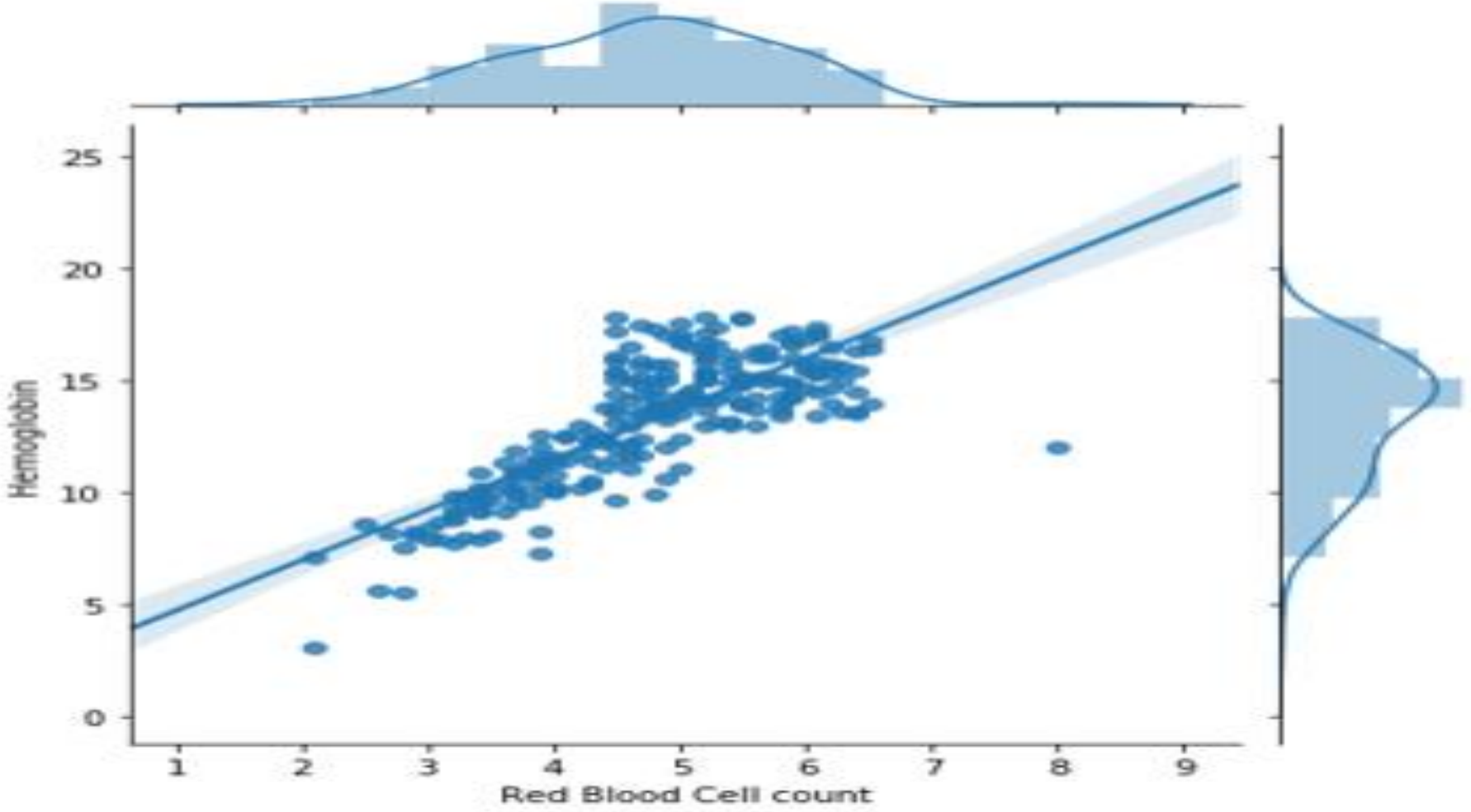
Histograms





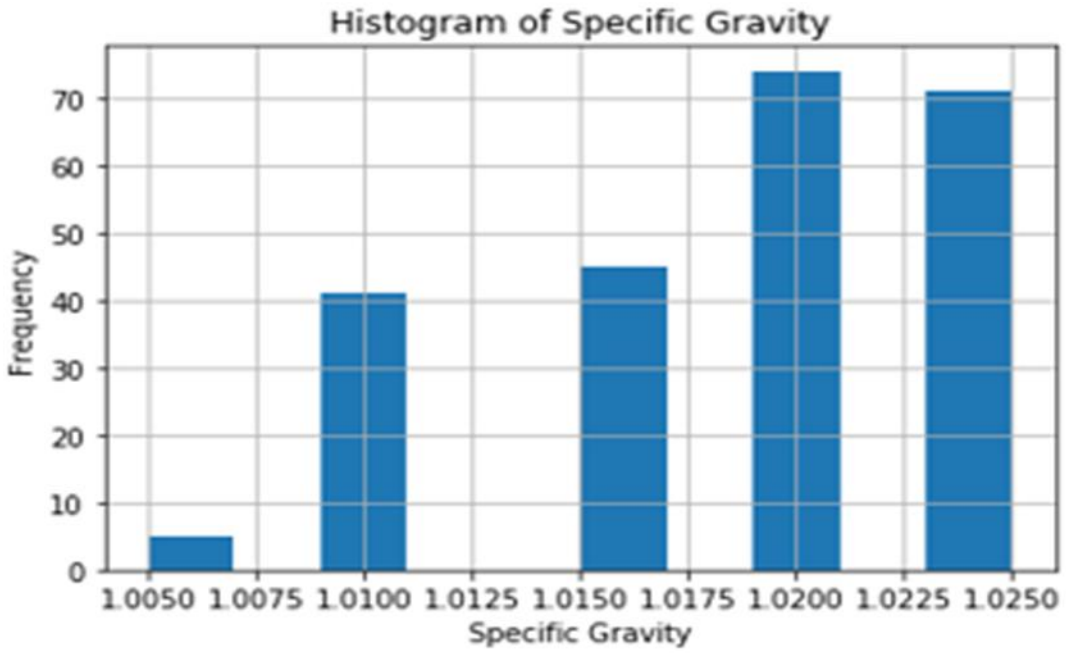
Note: both display skewness of some degree to the left. Both parameters correlate very well and display a positive relationship. Appears there may be few outliers but this may be due to smaller sample size or true rare occurrence, also the sample source was never noted. In this situation I would leave these outliers in place and not remove.

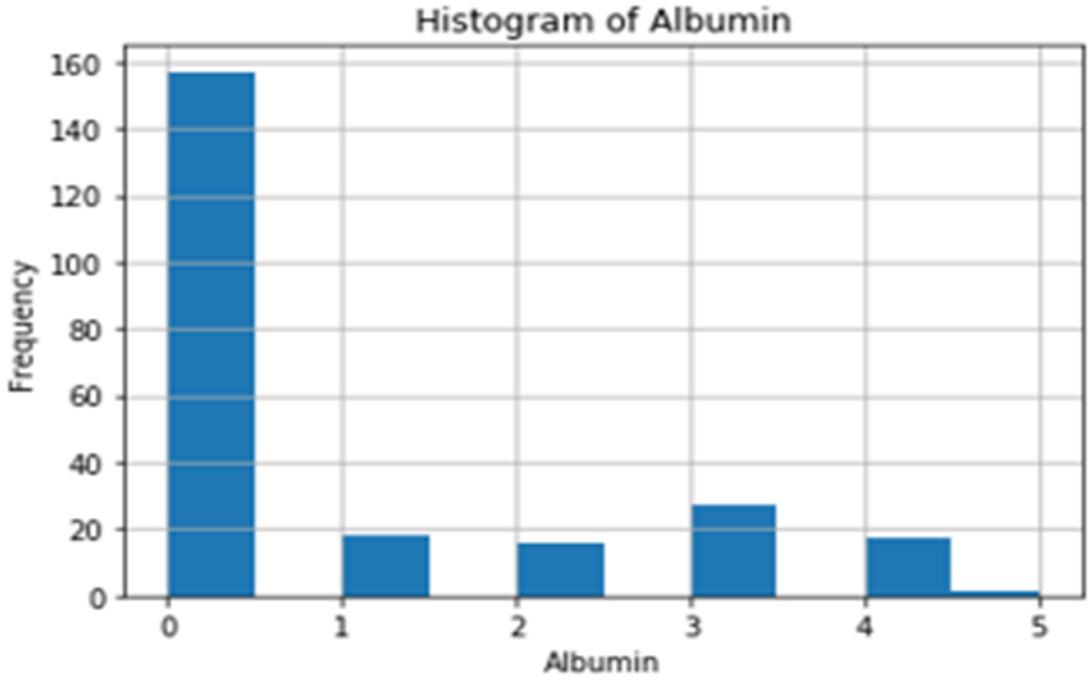
Hemoglobin and RBC



Note: Strong positive linear relationships appears to be present between these variables. This could cause problems in our regression analysis.

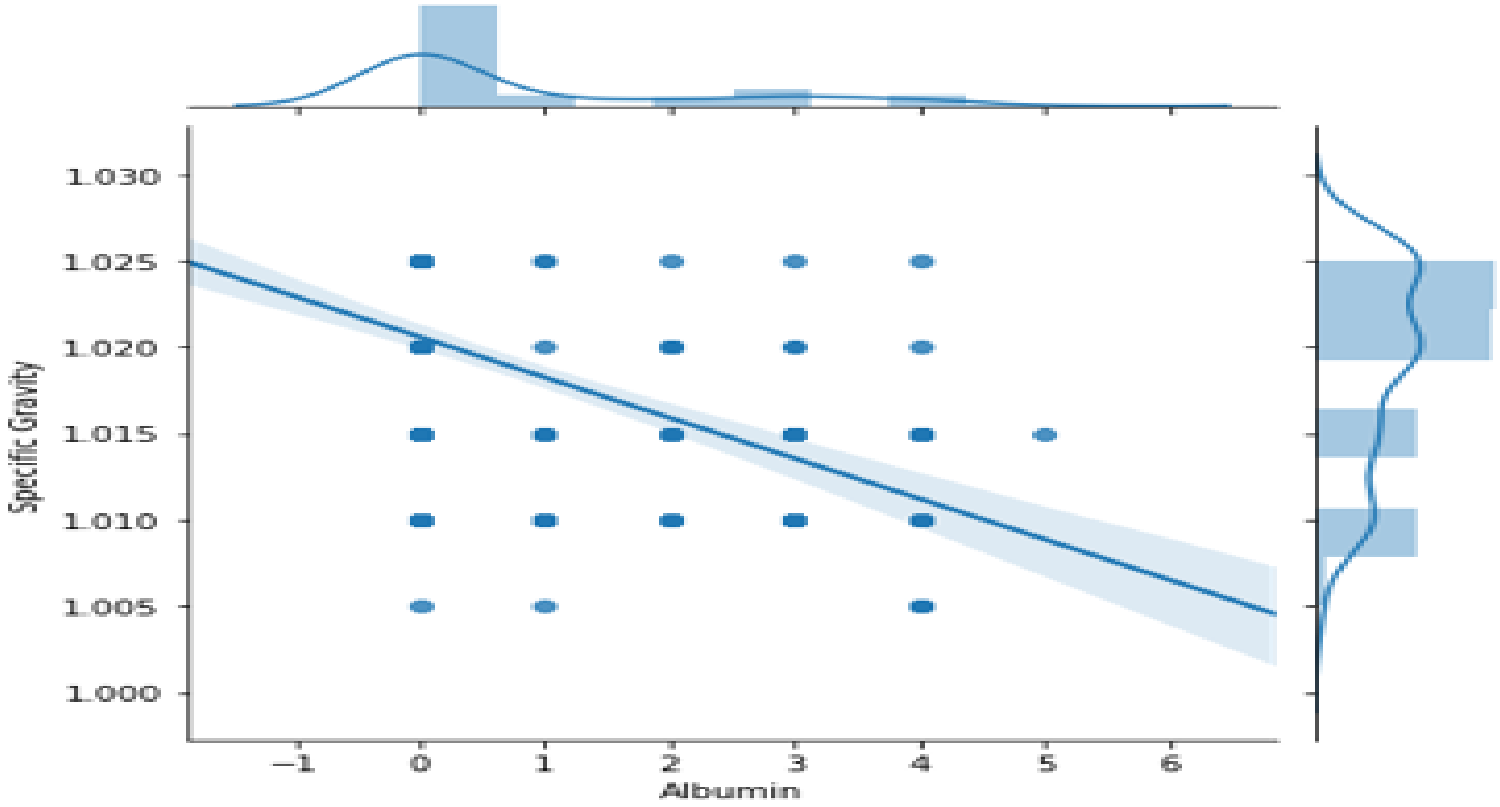
Histograms





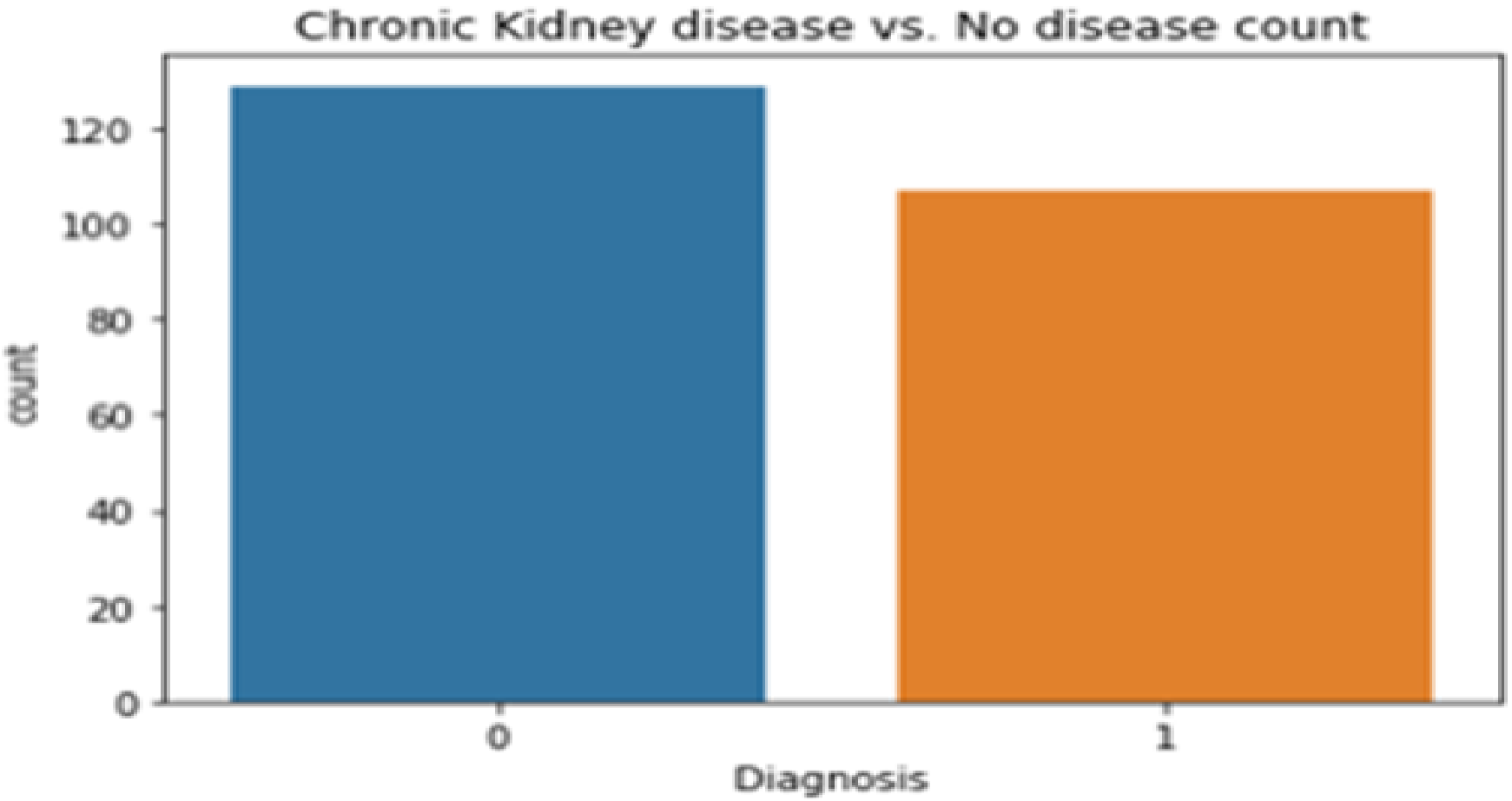
Note: both histograms display some degree of skewness along with some degree of correlation and a negative relationship.

Specific gravity vs Albumin



Note: there appears to be a negative relationship, but very weak if any linear relationship.

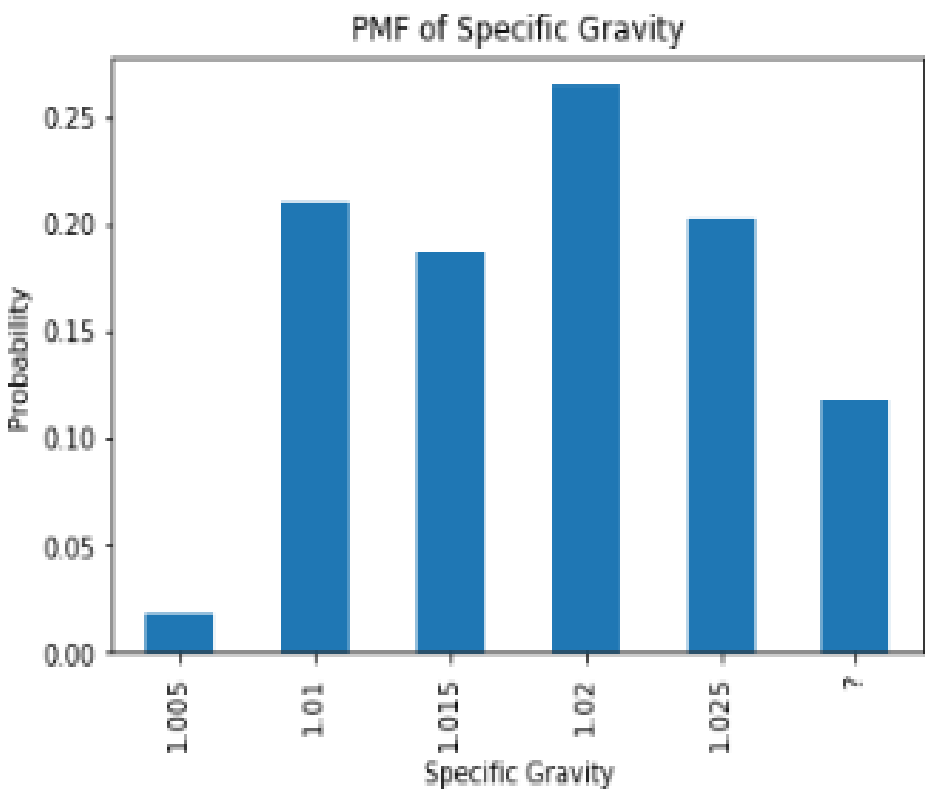
Binary variable graph



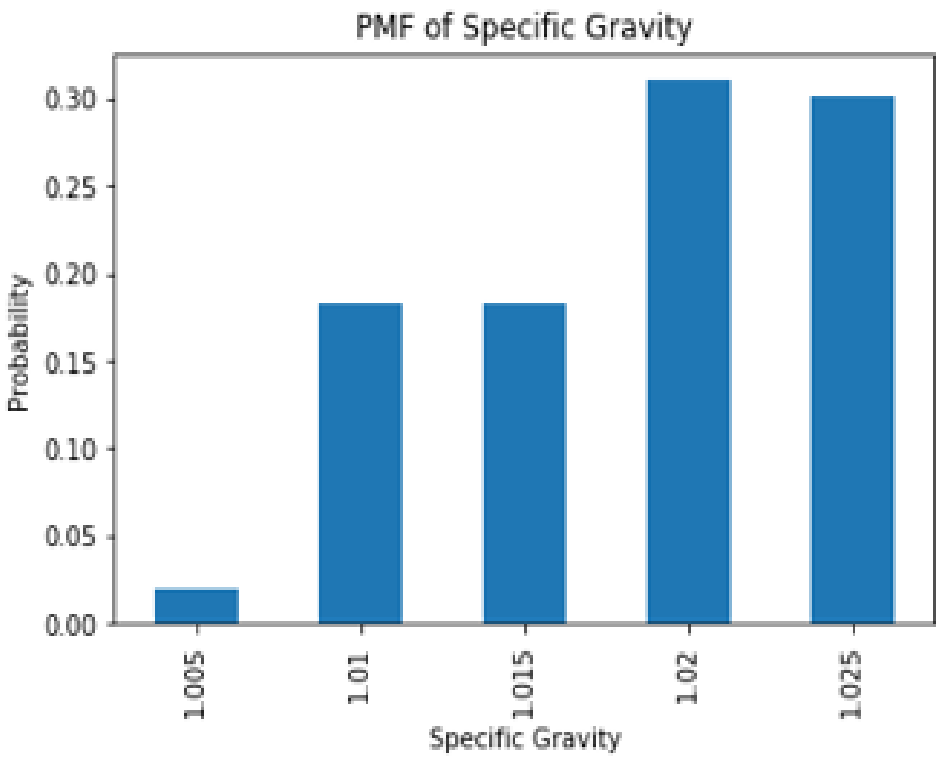
Note: 0 refers to no disease , and 1 refers to disease (more non disease in data set.)

PMF- Pre and Post data clean

Pre

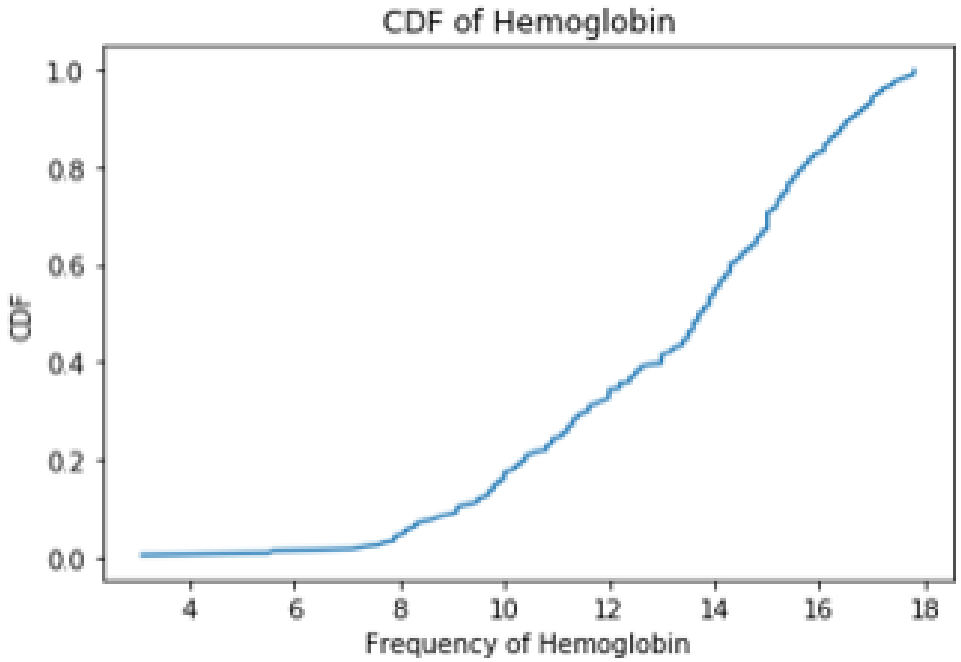


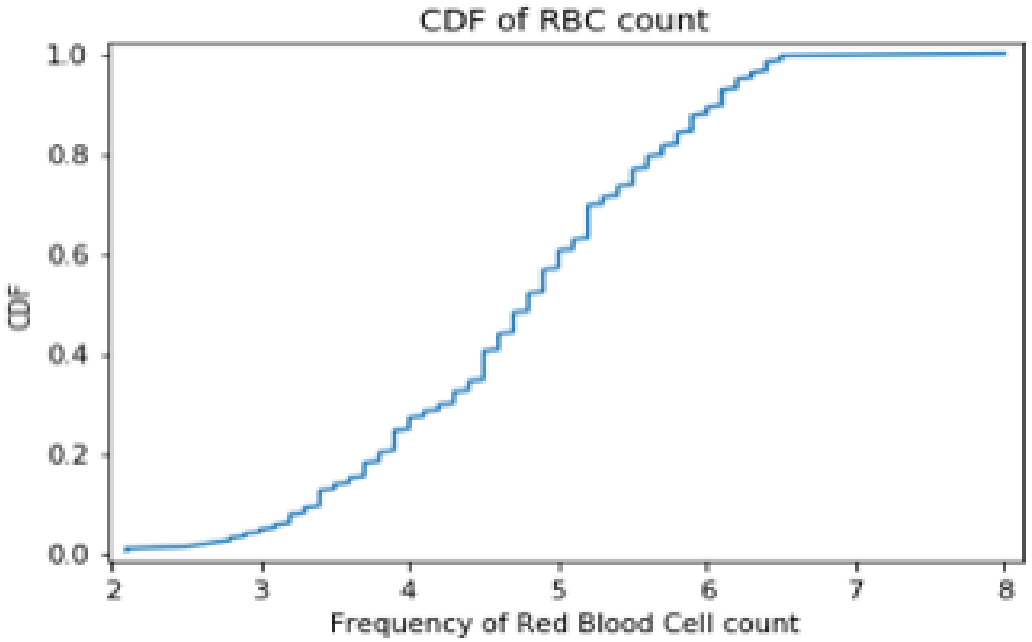
Post



Note: graph (left) displays false obscured data due to duplicates, etc., and graph on the right is post cleaning, very evident differences.

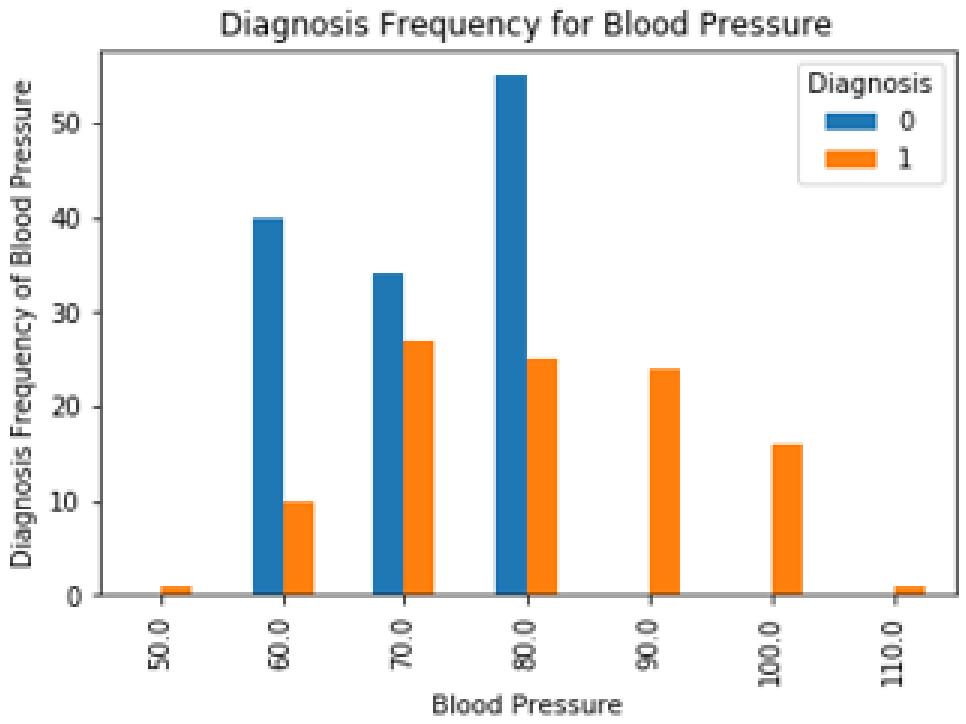
CDF for Hemoglobin and RBC

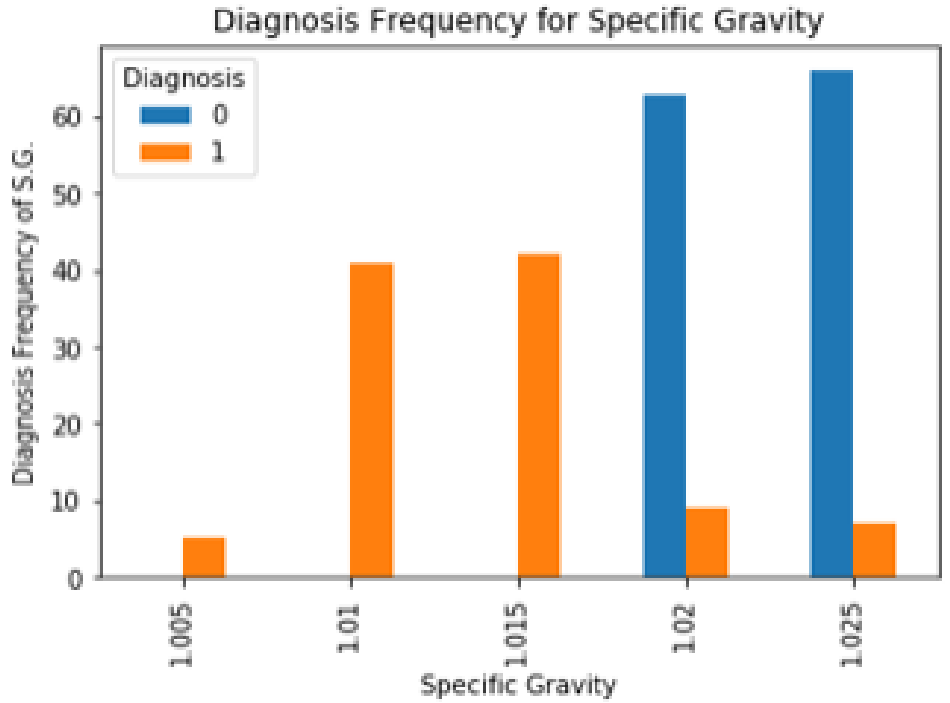




Note: Both CDF's appears to display that 80% of patients appear to have hemoglobin of approximately 16 or less and an RBC count of approximately 5.8 or less.

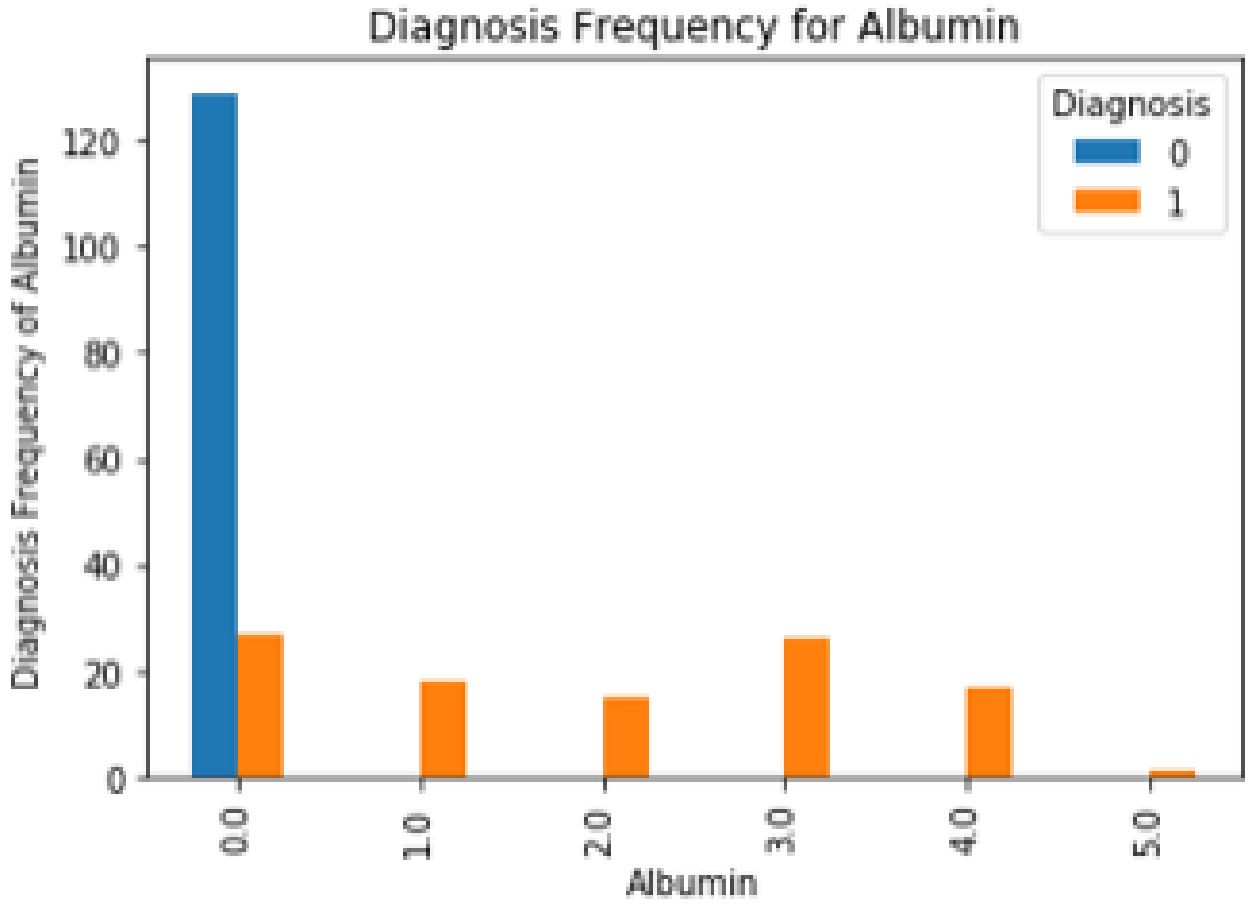
Bi-variable analysis





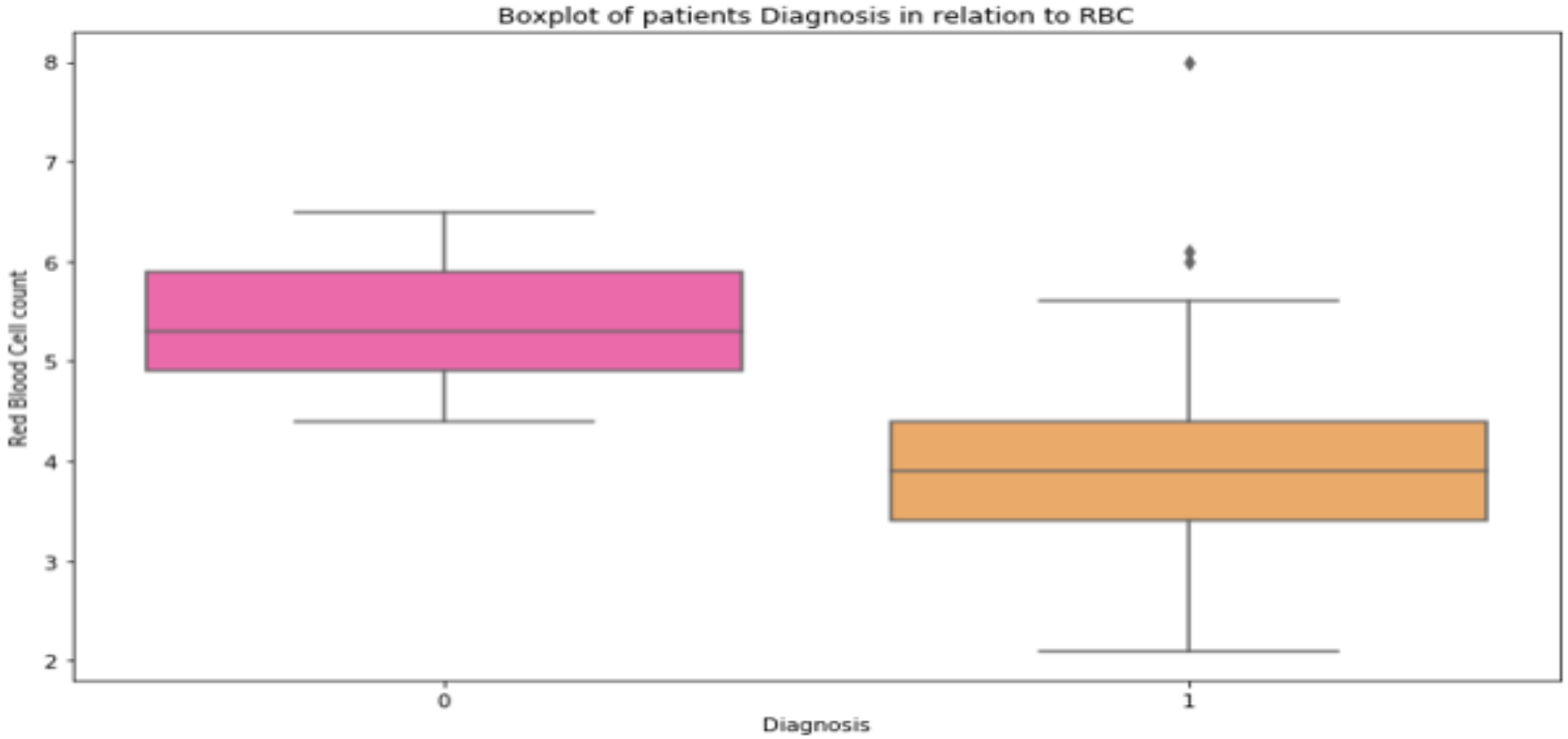
Note: Frequency graph for blood pressure displays a higher number of kidney disease for higher blood pressure values. Frequency graph for specific gravity displays higher number of kidney disease for lower values.

Bi-Variable analysis diagnosis/Albumin



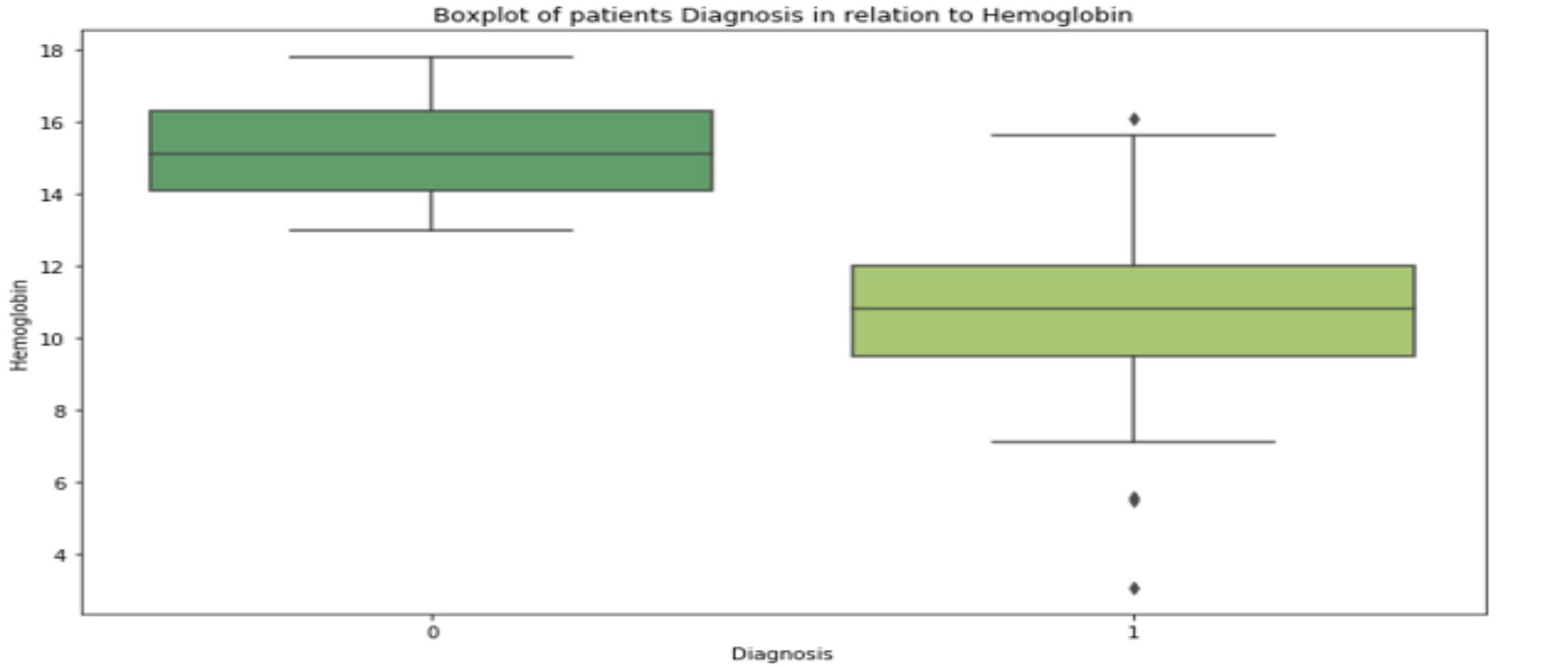
Note: appears majority of patients with chronic kidney disease report having higher albumin values.

Box plot Diagnosis in relation to RBC



Note: Appears patients with no kidney disease display higher RBC counts. Also, few patients with disease display higher values than the majority (outliers).

Box plot Diagnosis in relation to hemoglobin



Note: Appears patients with no disease also display higher hemoglobin which correlates well with the RBC count. Few outliers remain on the patients with the disease hemoglobin on both the low and high end.

Logistic Regression Analysis

Logit Regression Results

Dep. Variable:	Diagnosis	**No. Observations:**	233
Model:	Logit	**Df Residuals:**	227
Method:	MLE	**Df Model:**	5
Date:	Sun, 09 Feb 2020	**Pseudo R-squ.:**	0.9644
Time:	18:14:06	**Log-Likelihood:**	-5.6975
converged:	False	**LL-Null:**	-160.16
Covariance Type:	nonrobust	**LLR p-value:**	1.207e-64

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2548.2848	1532.923	1.662	0.096	-456.189	5552.758
SG	-2452.5303	1478.920	-1.658	0.097	-5351.160	446.100
Albumin	24.0014	3205.779	0.007	0.994	-6259.210	6307.212
Hemoglobin	-4.0534	2.263	-1.791	0.073	-8.489	0.382
RBC	-3.1356	2.537	-1.236	0.216	-8.108	1.837
BP	0.3021	0.184	1.639	0.101	-0.059	0.663

Note: This analysis appears to be statistically significant, thus leading us not able to reject the null hypothesis.(nonrobust displays possible outliers(remember the RBC count/hemoglobin boxplot), p-value less than 0.05 and R^2 of 96.4%).

Final Thoughts

- Based on Analysis, Hemoglobin(Hgb) and Red blood cell count(Rbc) displayed a strong positive relationship. Which is expected as Hgb is a component of Rbc.
- The Strong correlation between Hgb and Rbc is of concern to our model. Recommend reanalysis by removing one of these variables out of the dataset.
- All variables appeared to display correlation to our target variable(diagnosis) except for blood pressure.
- Our logistic analysis displayed that our model indicted that our variables are significant in predicting chronic kidney disease.
- Our data source never indicated where the source of these results came from(i.e. blood stream, body fluid, urine, etc.) or methodology.
- Based on our results, we cannot reject our null hypothesis.