Analysis of Hospitalized Patient Data

Athul Prakash

30/11/2021

Introduction

We are presented with demographic data and clinical test data of 1177 patients with heart-failure, who were admitted to the ICU of a teaching hospital in Boston, Massachusetts. The data was sourced from the MIMIC-III database.
Our aim is to explore the patient data we have and summarize it visually.

Data Description

The following data is a subset extracted from the MIMIC-III database, of ICU-admitted Heart-Failure patients. There are 11 variables of 1177 patients. The data collected includes demographic data, such as age and gender, as well as clinically measured data, such as Blood Pressure, Heart-Rate and Urea-Nitrogen concentration. The outcome of the treatment is whether the patient succumbed or survived while in ICU care.

The variables are summarized below. All of them are numeric type.

```
knitr::kable(summ)
```

VariableName	Type	MeasuredQuantity
age	Discrete - Count	Age of the Patient
gendera	Categorical - Nominal (Dichotomous)	Gender (M/F)
BMI	Continuous	Body Mass Index
heart.rate	Continuous	Heart beats per minute
Diastolic.blood.pressure	Continuous	Blood Pressure
diabetes	Categorical - Nominal (Dichotomous)	Having Diabetes (YES/NO)
Renal.failure	Categorical - Nominal (Dichotomous)	Having renal failure (YES/NO)
Lymphocyte	Continuous	Blood lymphocyte count
Neutrophils	Continuous	Blood neutrophil count
Urea.nitrogen	Continuous	Blood Urea Nitrogen level
outcome	Categorical - Nominal (Dichotomous)	Patient Mortality while admitted (YES/NO)

Note:-
* Gender: Male=1, Female=2
* Outcome: Survived=0, Died=1
* BMI = height/(weight^2) in m/kg^2
* Renal Failure: YES=1, NO=1
* Diabetes: YES=1, NO=0

Data Source Description

The MIMIC-III database (version 1.4, 2016) is a publicly available critical care database containing de-identified data on 46,520 patients and 58,976 admissions to the ICU of the Beth Israel Deaconess Medical Center, Boston, USA, between 1 June, 2001 and 31 October, 2012.
These data include comprehensive information, such as demographics, admitting notes, International Classification of Diseases-9th revision (ICD-9) diagnoses, laboratory tests, medications, procedures, fluid balance, discharge summaries, vital sign measurements undertaken at the bedside, caregivers notes, radiology reports, and survival data12.
The MIMIC data is available to those who complete a web-based training course of the same institute.

Data Challenges

- The data was collected manually by combining records from different departments and procedures. For this reason, not all variables have been measured for all individuals. Thus, there are many NAs for each column, as is common for medical databases,
- The MIMIC-III database is available only after obtaining a certificate of training in the use of medical data. Since our data is a subset of the MIMIC database, we could not source the descriptions of the quantities being measured. As such, variables such as Neutrophil count are presented without knowing the real-world units.

Univariate Analysis

Discrete Variable - Age

A quick look at summary statistics reveals :-

```
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   65.00   77.00   74.05   85.00   99.00
```

- The mean age is 74 years with a median of 77 years.
- This indicates a relatively old population, which could be expected for ICU admissions.

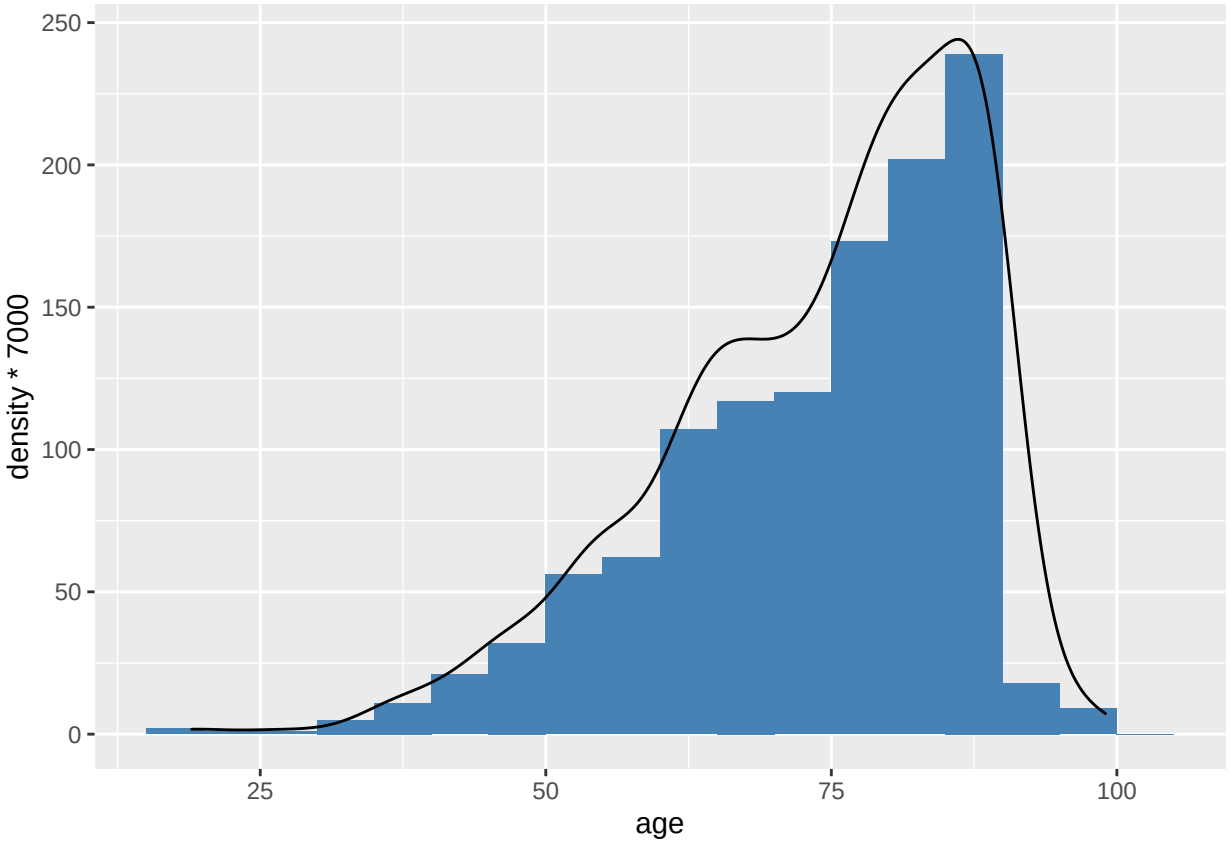
```
sd(data$age)
```

```
## [1] 13.43724
```

- The standard deviation is 13.43 years

For a deeper look, we can consider age brackets with 5 year gap. The following histogram shows the number of people in each age bracket, with density overlaid.

```
ggplot(data=data) + geom_histogram(mapping=aes(x=age), breaks = seq(15,105,by=5), fill="steelblue") +
  geom_density(mapping=aes(x=age, y=..density..*7000))
```



- We can summarize that the frequency increases with increasing age till 90 years and then sharply declines.

Categorical Variables

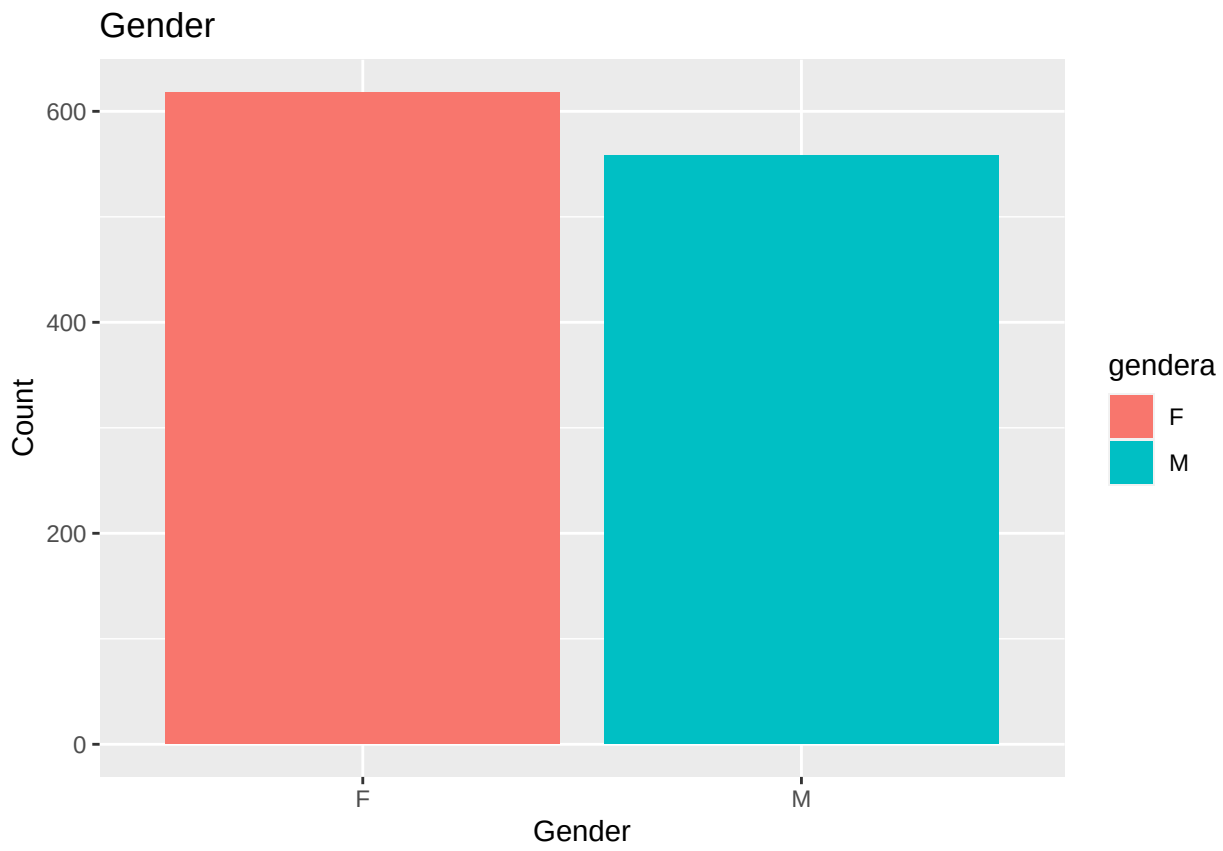
For dichotomous variables, we can show the comparative frequencies of each using a histogram.

We have 4 categorical variables

- Gender
- Diabetic
- Renal Failure
- Mortality

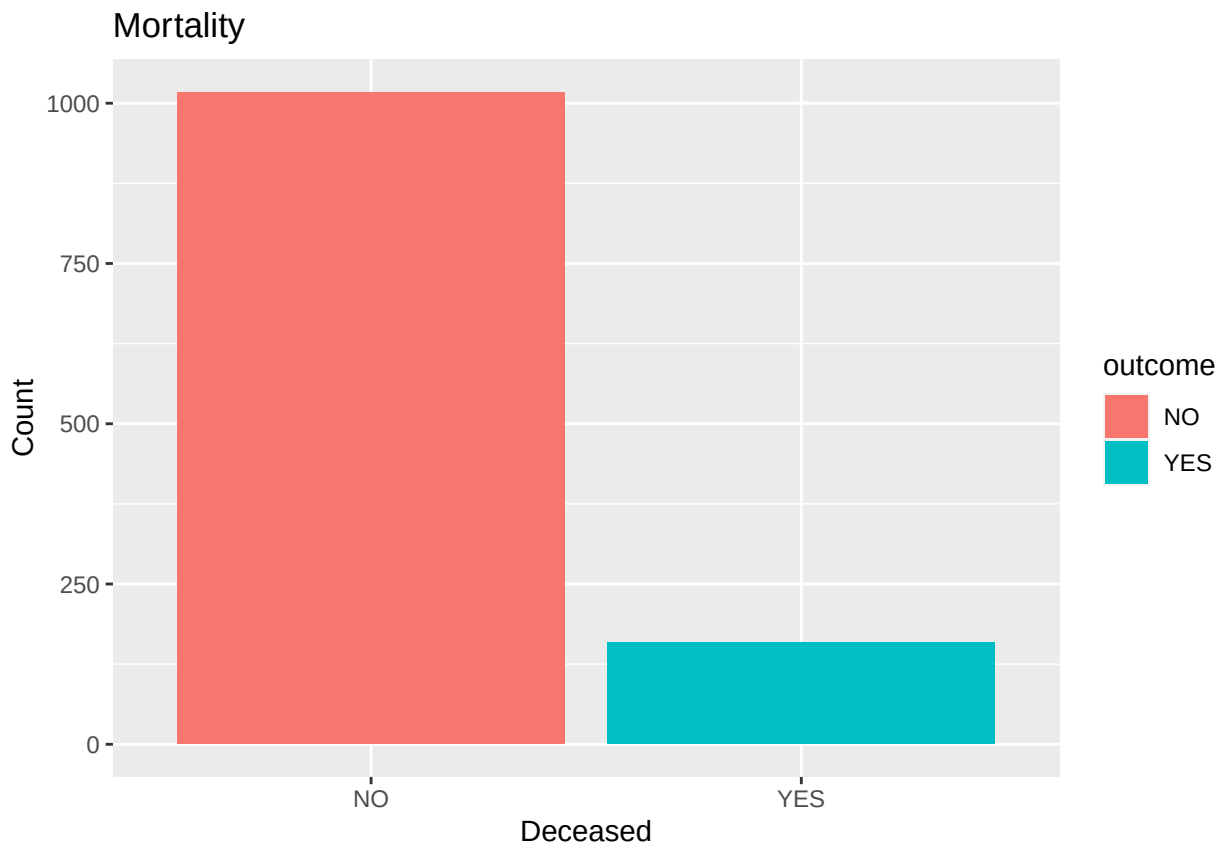
Gender We can visualize the same in a histogram.

```
ggplot(data=data) + layer(mapping = aes(x=gendera,fill=gendera), stat="count", geom="bar", position="id
  labs(x="Gender", y="Count", title="Gender")
```



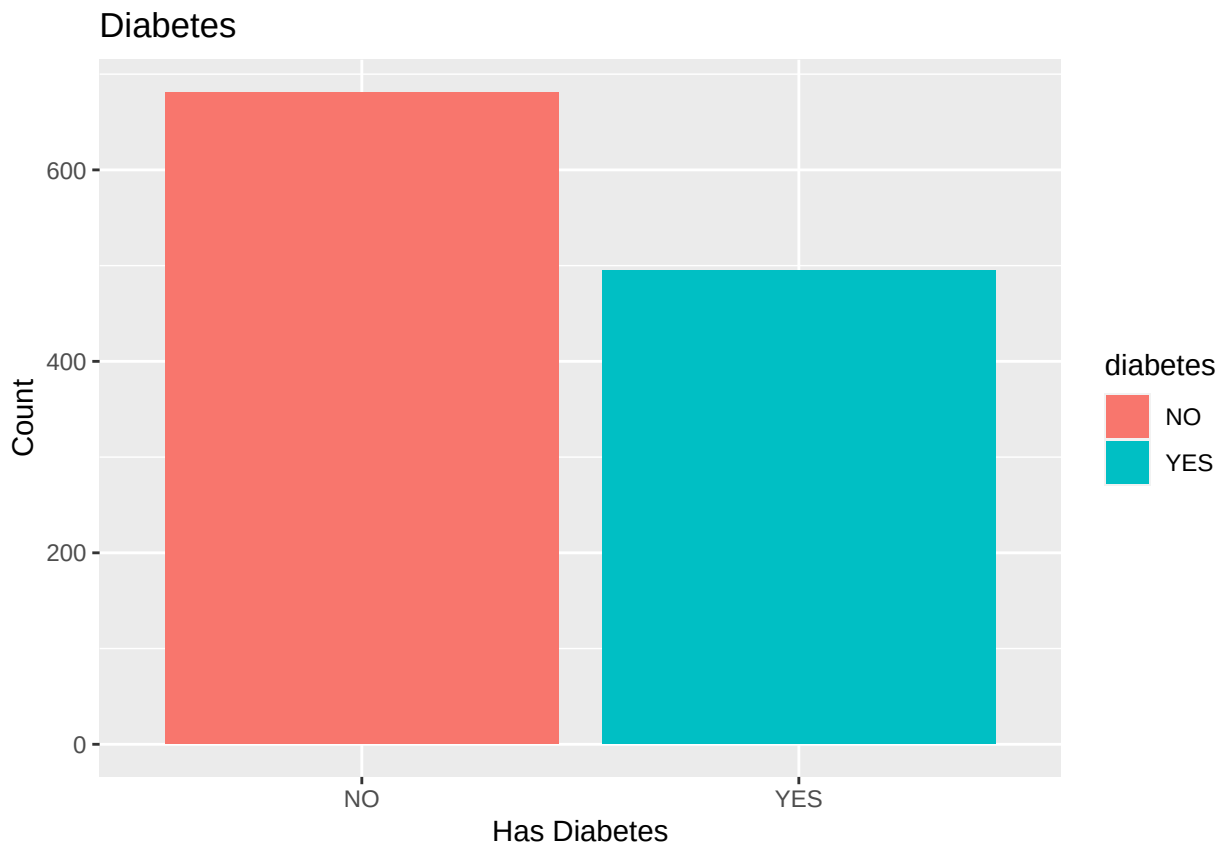
Mortality We can visualize the same in a histogram.

```
ggplot(data=data) + layer(mapping = aes(x=outcome,fill=outcome), stat="count", geom="bar", position="id
  labs(x="Deceased", y="Count", title="Mortality")
```



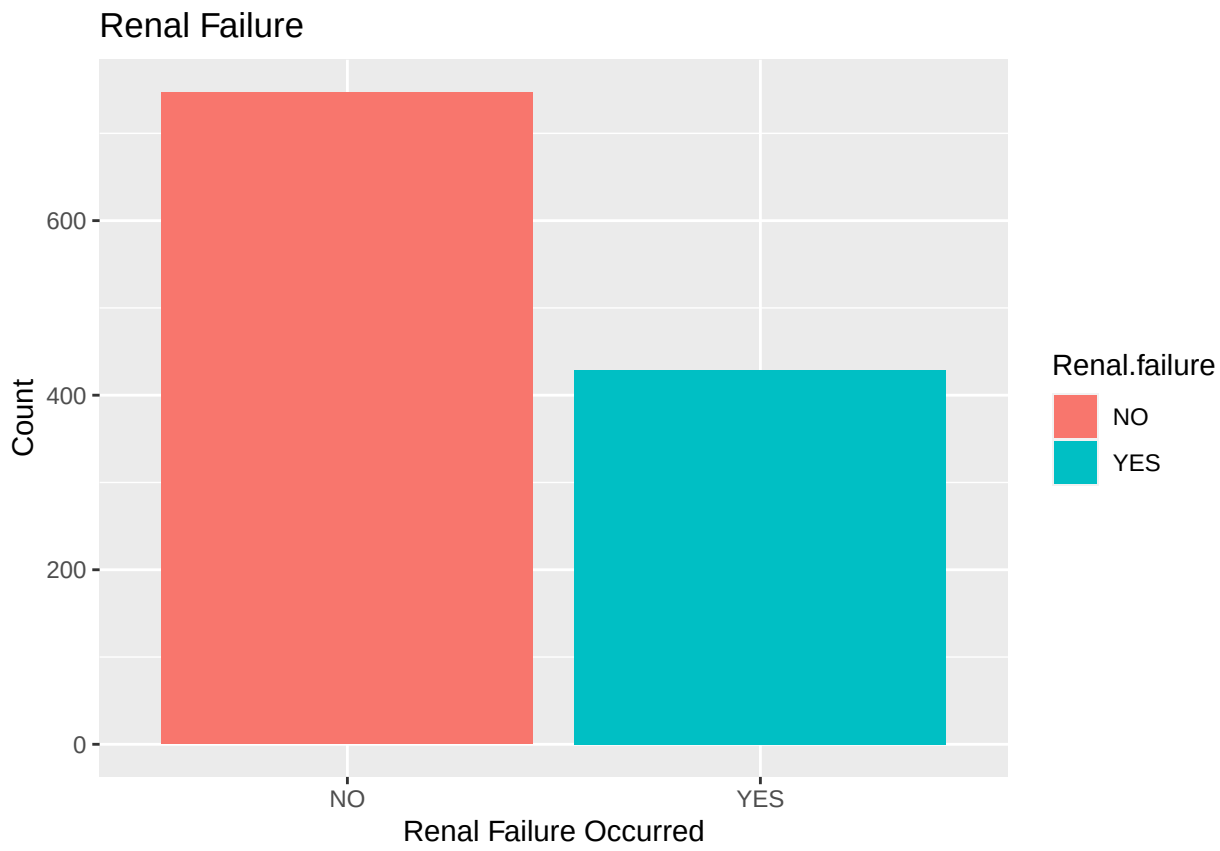
Diabetes We can visualize the same in a histogram.

```
ggplot(data=data) + layer(mapping = aes(x=diabetes,fill=diabetes), stat="count", geom="bar", position="
  labs(x="Has Diabetes", y="Count", title="Diabetes")
```



Renal Failure We can visualize the same in a histogram.

```
ggplot(data=data) + layer(mapping = aes(x=Renal.failure,fill=Renal.failure), stat="count", geom="bar", p
  labs(x="Renal Failure Occurred", y="Count", title="Renal Failure")
```

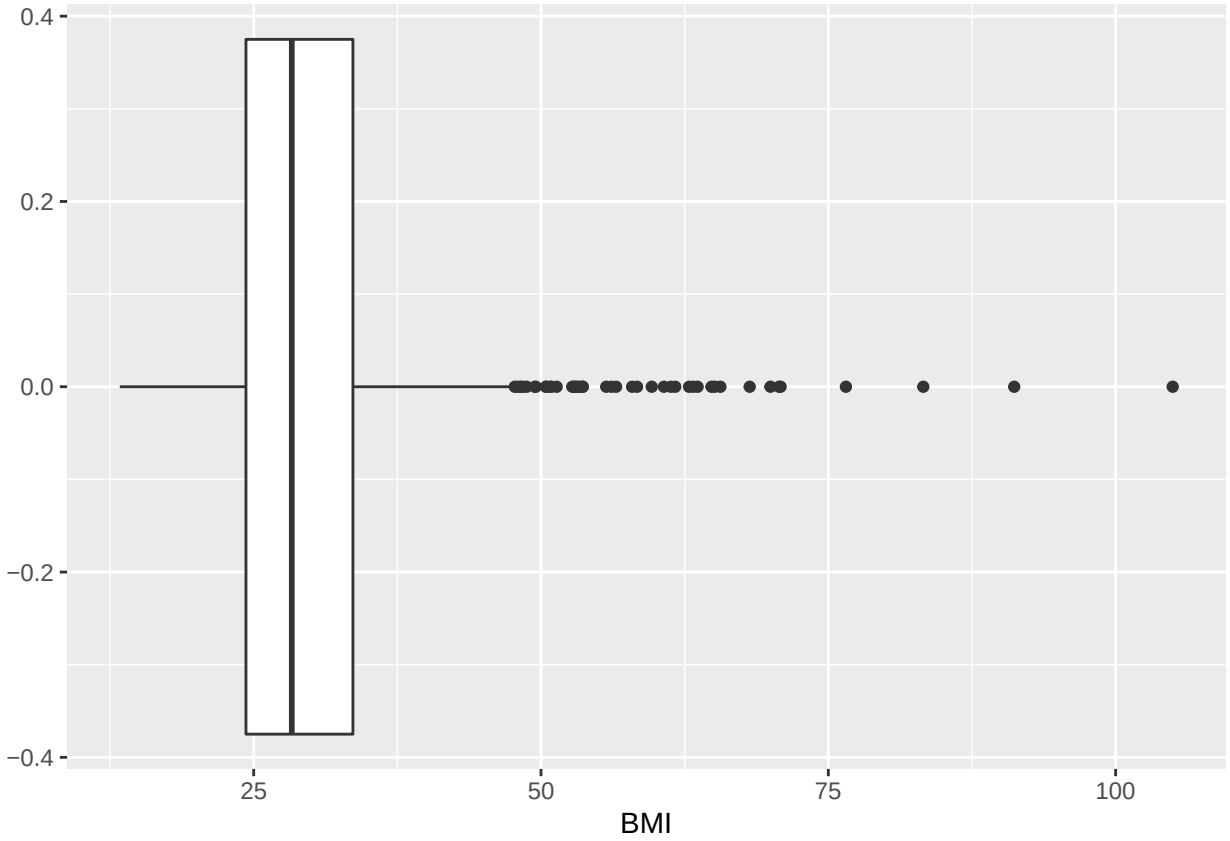


Blood Pressure

We can plot the boxplot summarizing BMI

```
ggplot(data=data, mapping=aes(x=BMI)) + geom_boxplot()
```

```
## Warning: Removed 214 rows containing non-finite values (stat_boxplot).
```



The mean and median are as follows:

```
summary(data$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   13.35   24.33   28.31   30.19   33.63  104.97     214
```

Heart Rate

The mean and median are as follows:

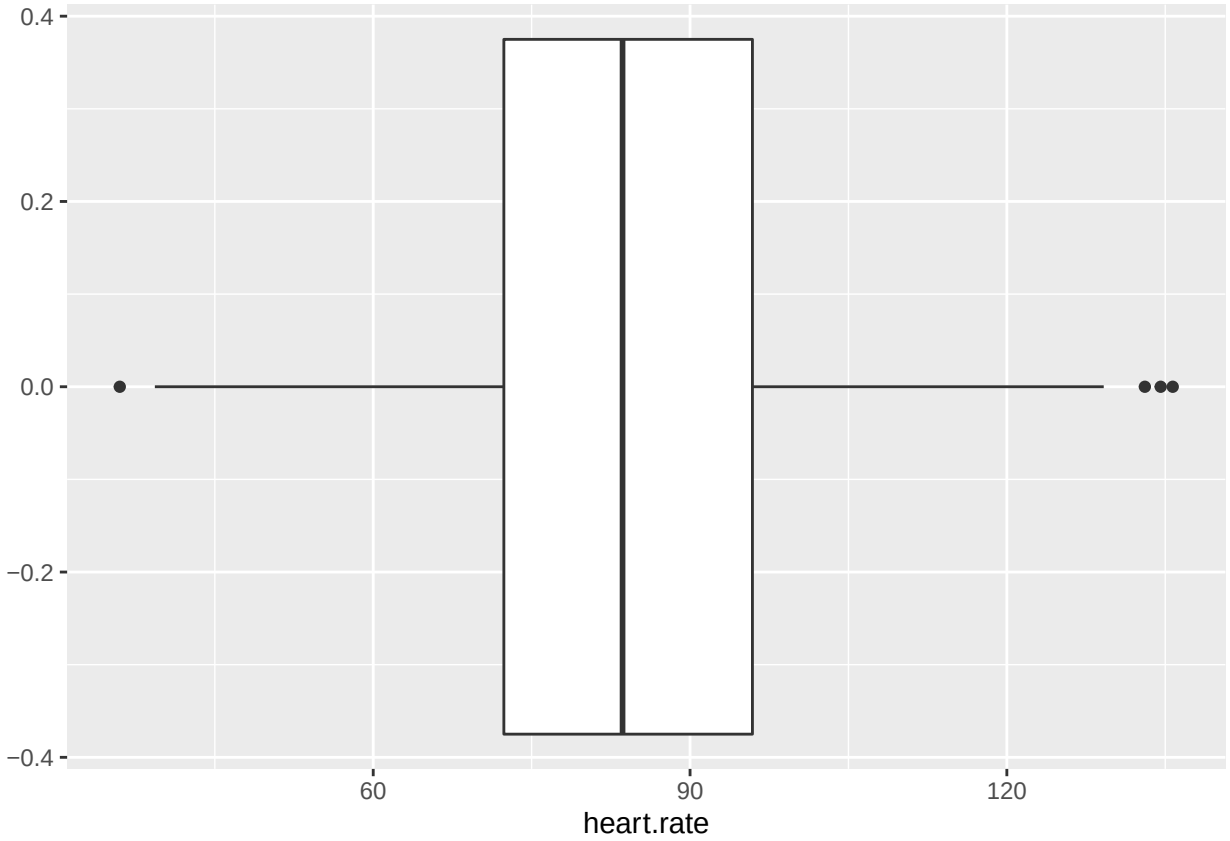
```
summary(data$heart.rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   36.00   72.37   83.61   84.58   95.91  135.71      12
```

We can plot the boxplot summarizing Heart Rate

```
ggplot(data=data, mapping=aes(x=heart.rate)) + geom_boxplot()
```

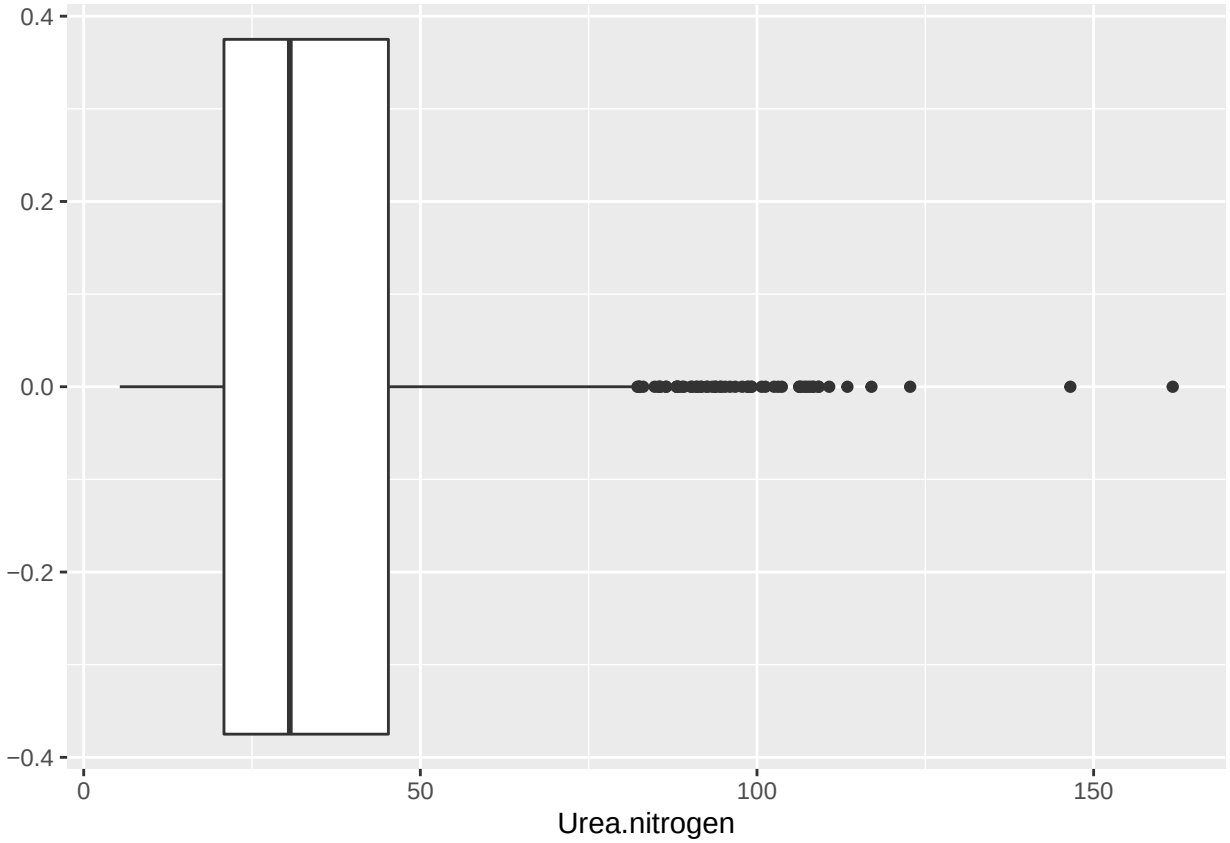
```
## Warning: Removed 12 rows containing non-finite values (stat_boxplot).
```



Urea Nitrogen

We can plot the boxplot summarizing Urea Nitrogen Level

```
ggplot(data=data, mapping=aes(x=Urea.nitrogen)) + geom_boxplot()
```

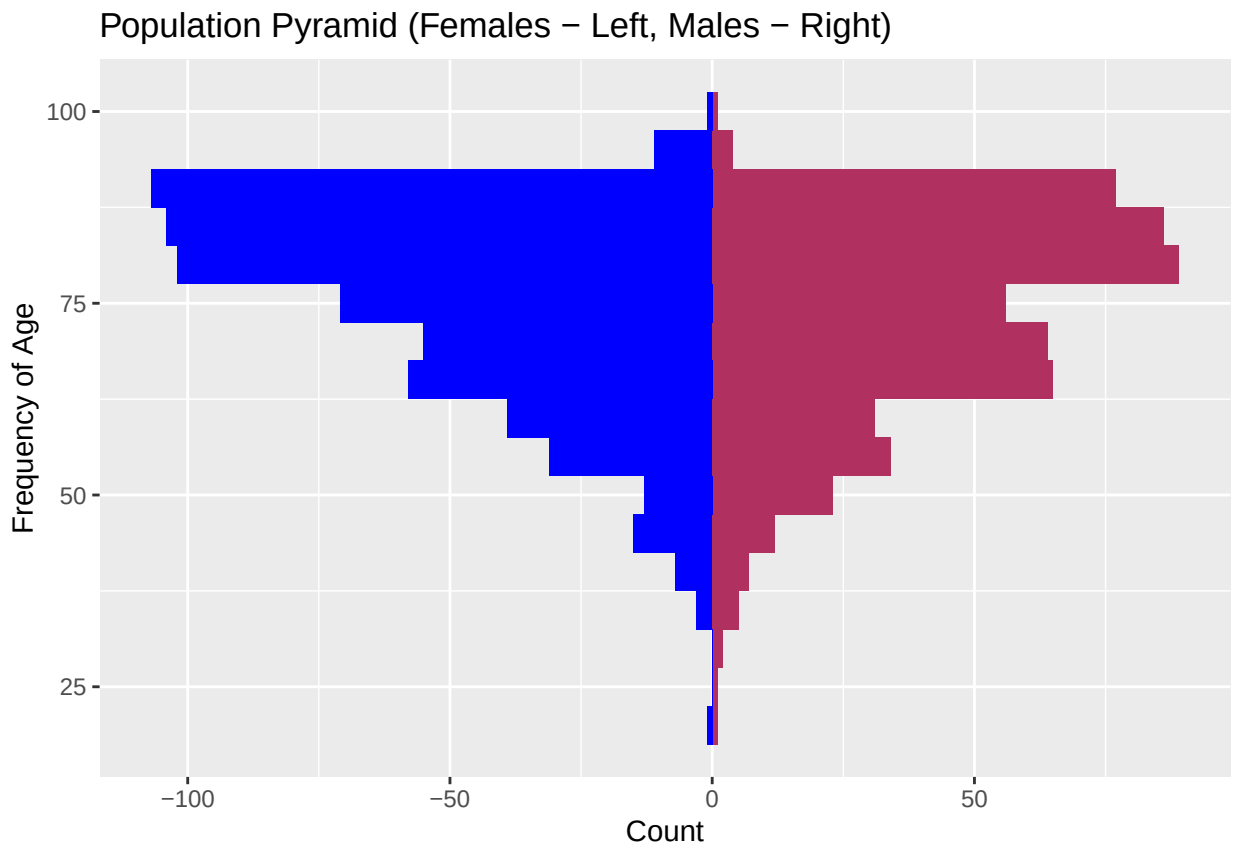


Mutlivariate Analysis

Demographic Data - Age + Gender

The following population pyramid summarizes the frequency of each gender among age brackets of 5 years' span.

```
ggplot() +
  layer(data = data[data$gendera=="M",], geom="bar", stat="bin", position = "identity", mapping=aes(y=a
  layer(data = data[data$gendera=="F",], geom="bar", stat="bin", position = "identity", mapping=aes(y=a
  ggtitle("Population Pyramid (Females - Left, Males - Right)") +
  ylab("Frequency of Age") +
 xlab("Count")
```



Renal Failure vs Mortality

Looking at the raw distribution
* Renal Failure along rows, Mortality along columns

```
table(data$Renal.failure, data$outcome)
```

```
##
##        NO YES
##   NO  625 122
##   YES 392  37
```

Performing a Chi-Squared analysis to establish correlation:

```
chisq.test(data$Renal.failure, data$outcome, correct=FALSE)
```

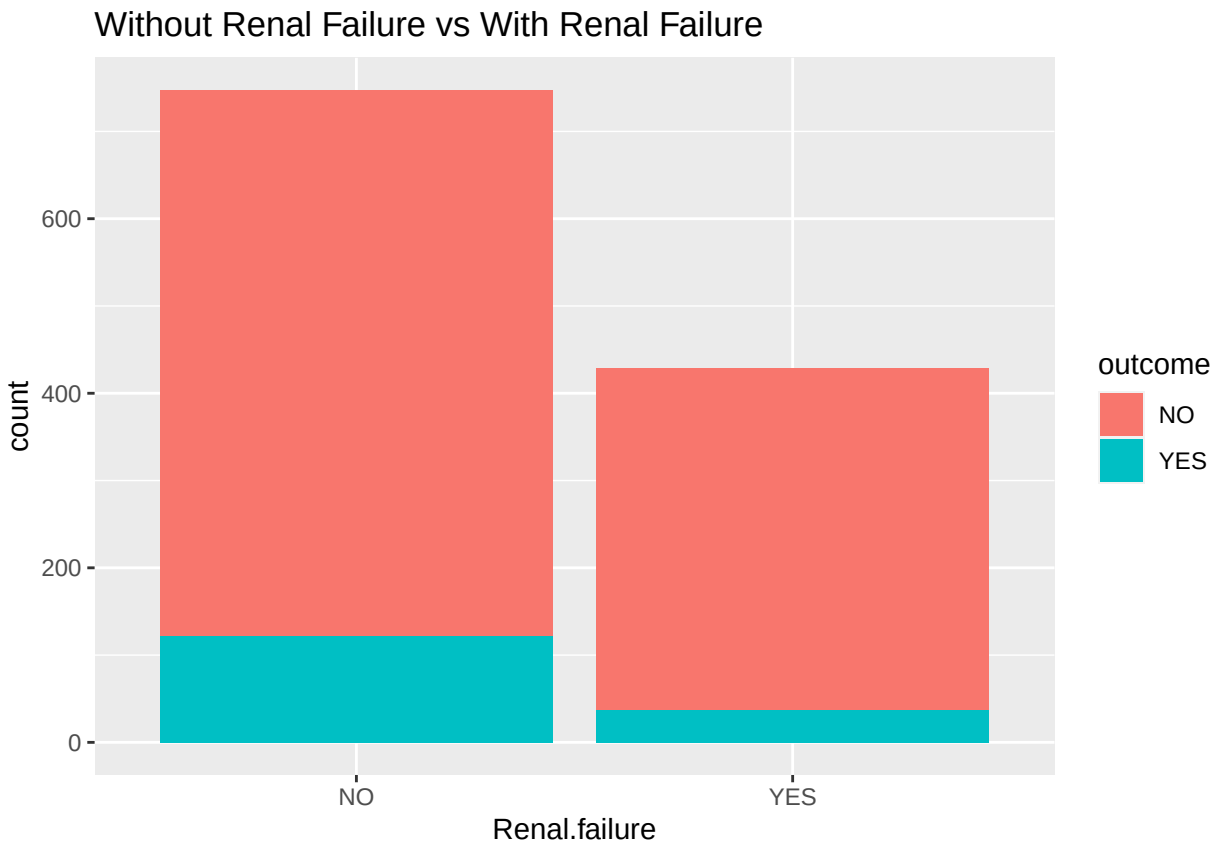
```
##
##  Pearson's Chi-squared test
##
## data:  data$Renal.failure and data$outcome
## X-squared = 13.844, df = 1, p-value = 0.0001986
```

A p-value less than 0.05 with a high X-squared value indicates a strong correlation b/w Renal failure and mortality.

We can plot the histograms of outcome for the groups with and without Renal Failure

```
ggplot(data=data) + geom_histogram(mapping=aes(x=Renal.failure, fill=outcome), stat="count") +
  ggtitle("Without Renal Failure vs With Renal Failure")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Gender vs Mortality

Looking at the raw distribution
* Gender along rows, Mortality along columns

```
table(data$gendera, data$outcome)
```

```
##
##      NO YES
##   F 539  79
##   M 478  80
```

Performing a Chi-Squared analysis to establish correlation:

```
chisq.test(data$gendera, data$outcome, correct=FALSE)
```

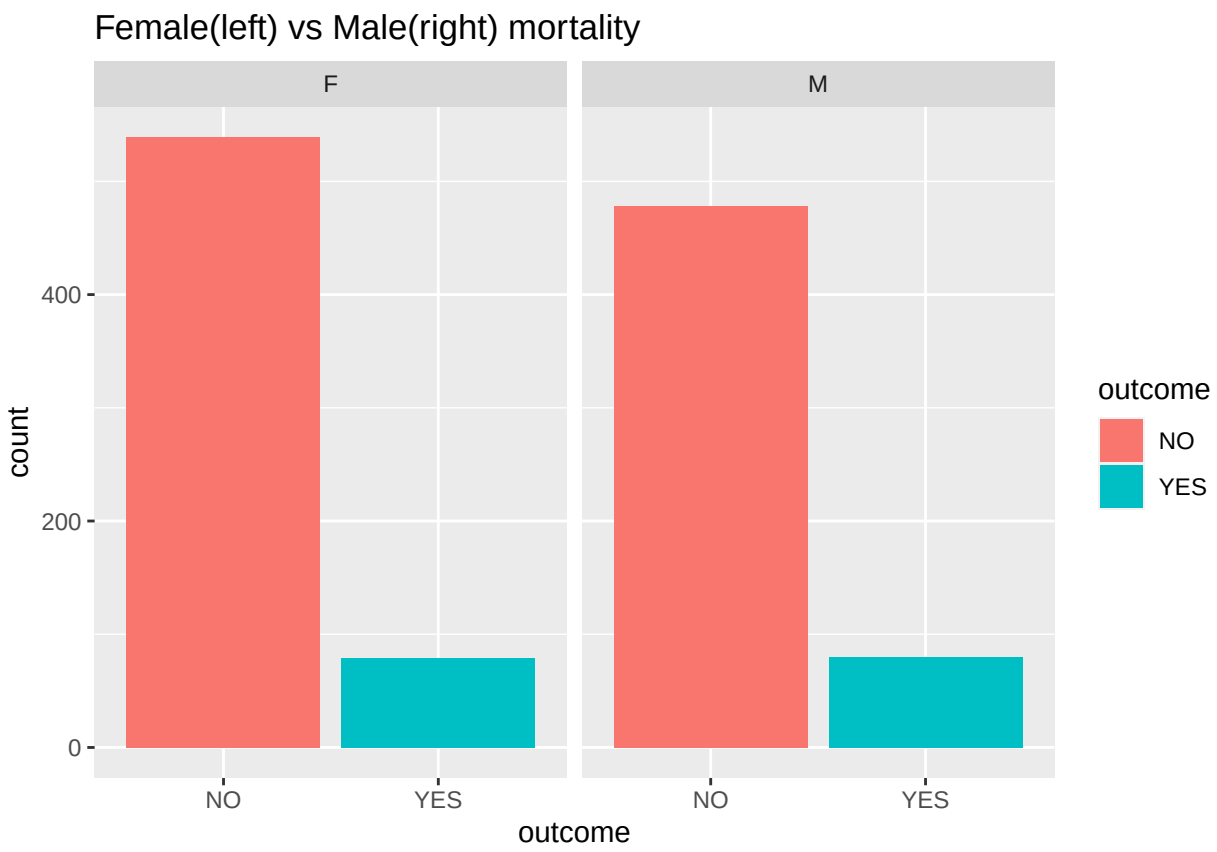
```
##
##  Pearson's Chi-squared test
##
## data:  data$gendera and data$outcome
## X-squared = 0.60544, df = 1, p-value = 0.4365
```

Thus, mortality is largely independent of gender.

We can plot the hi

```
ggplot(data=data) + geom_histogram(mapping=aes(x=outcome, fill=outcome), stat="count") + facet_wrap(data
  ggtitle("Female(left) vs Male(right) mortality")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

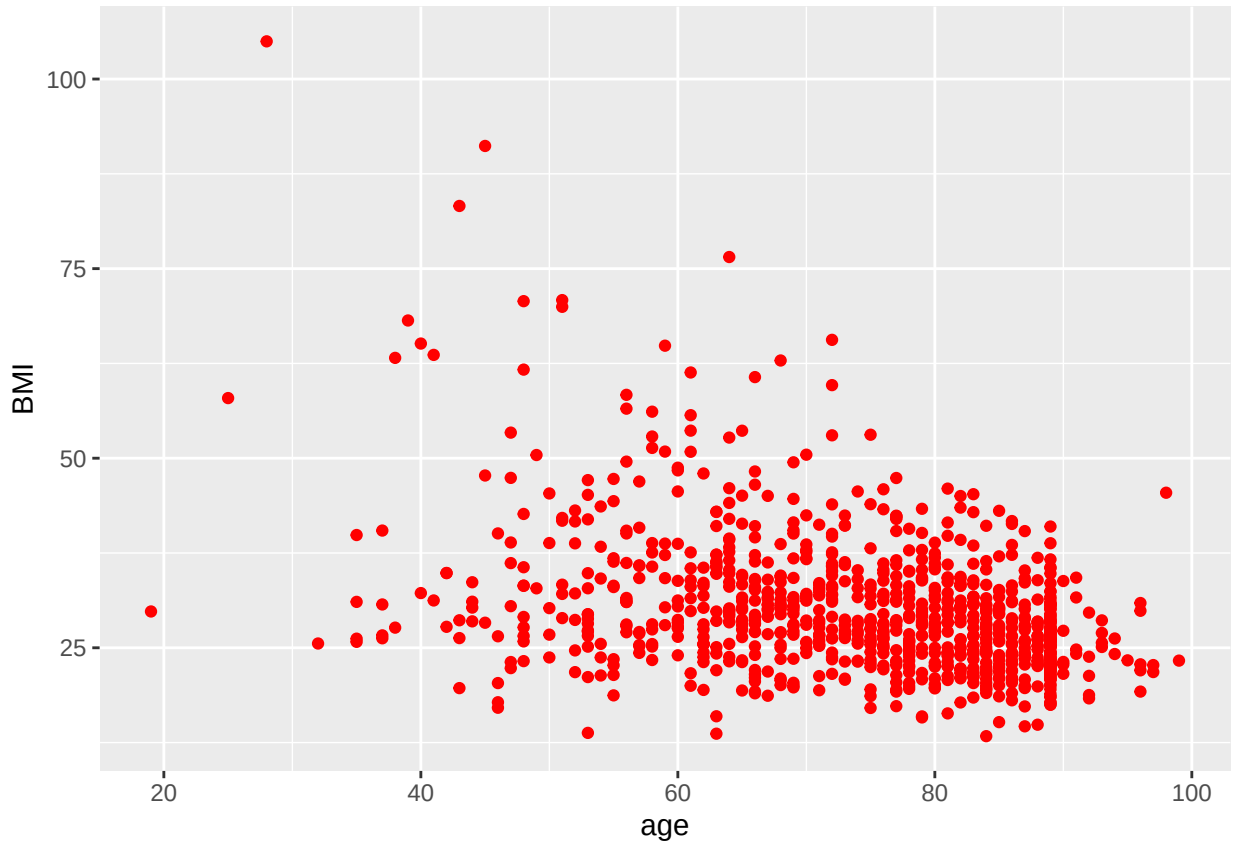


Age vs BMI

We can draw a scatterplot to investigate any relationship b/w these variables

```
ggplot(data=data) + geom_point(mapping=aes(x=age,y=BMI), color="red")
```

```
## Warning: Removed 214 rows containing missing values (geom_point).
```

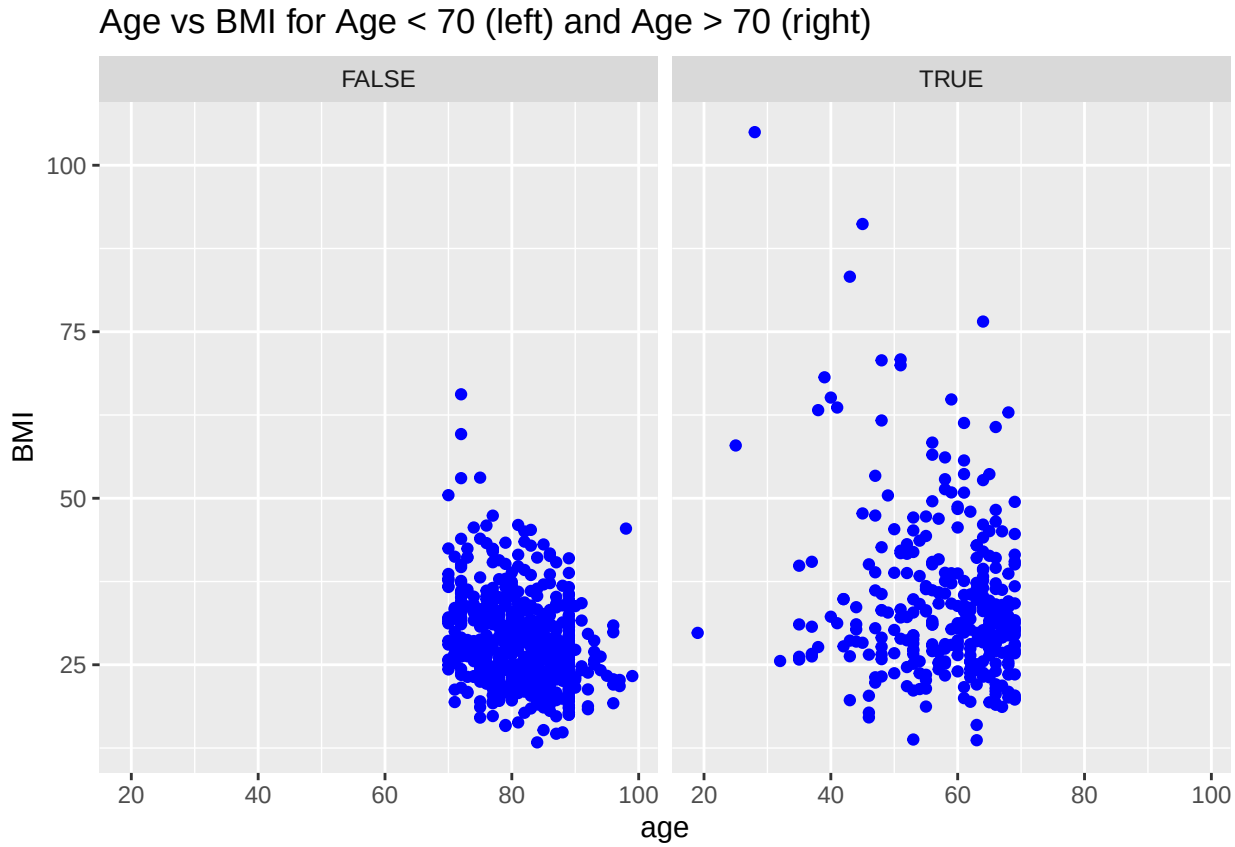


There is a moderate correlation. Particularly, we can see that young patients have a large variance in BMI whereas older patients have a uniform and lower average BMI.

We can split our dataset to observe the data separately for younger and older patients

```
ggplot(data=data) + geom_point(mapping=aes(x=age,y=BMI), color="blue") + facet_wrap(data$age<70) +
  ggtitle("Age vs BMI for Age < 70 (left) and Age > 70 (right)")
```

```
## Warning: Removed 214 rows containing missing values (geom_point).
```



Neutrophils vs Lymphocytes (and outcome)

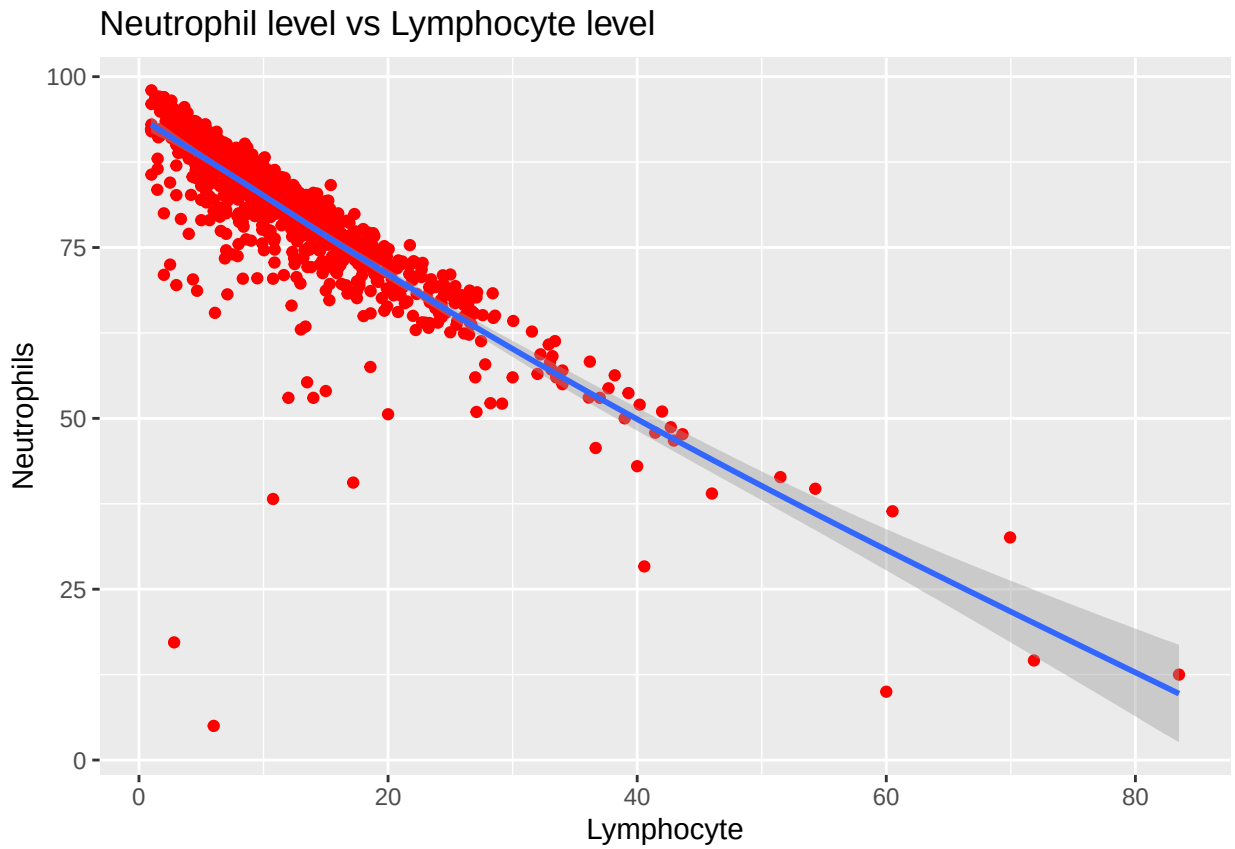
We can make a scatterplot overlaid with a smoothing line

```
ggplot(data=data, mapping=aes(x=Lymphocyte, y=Neutrophils)) + geom_point(color="red") + geom_smooth() +
  ggtitle("Neutrophil level vs Lymphocyte level")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 145 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 145 rows containing missing values (geom_point).
```



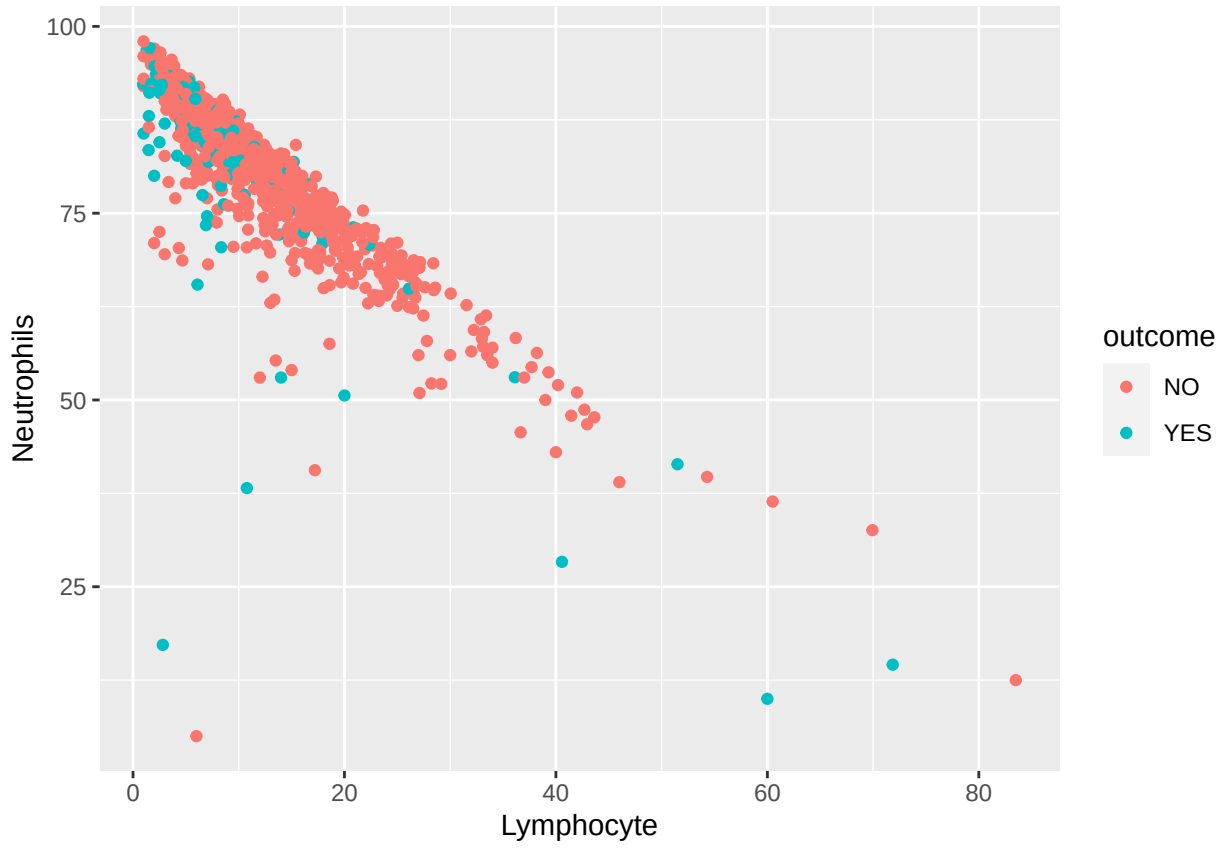
Undoubtedly, Blood Neutrophil levels drop steadily and linearly with Blood Lymphocyte levels.

Furthermore...
In medical investigations, the ratio of Neutrophils to Lymphocytes is often used as a predictor of a patient's health condition. With this motivation, we can investigate the impact of outcome on the correlation.

Let us color the scatterplot differently for surviving and deceased patients.

```
ggplot(data=data) + geom_point(mapping=aes(x=Lymphocyte,y=Neutrophils, color=outcome))
```

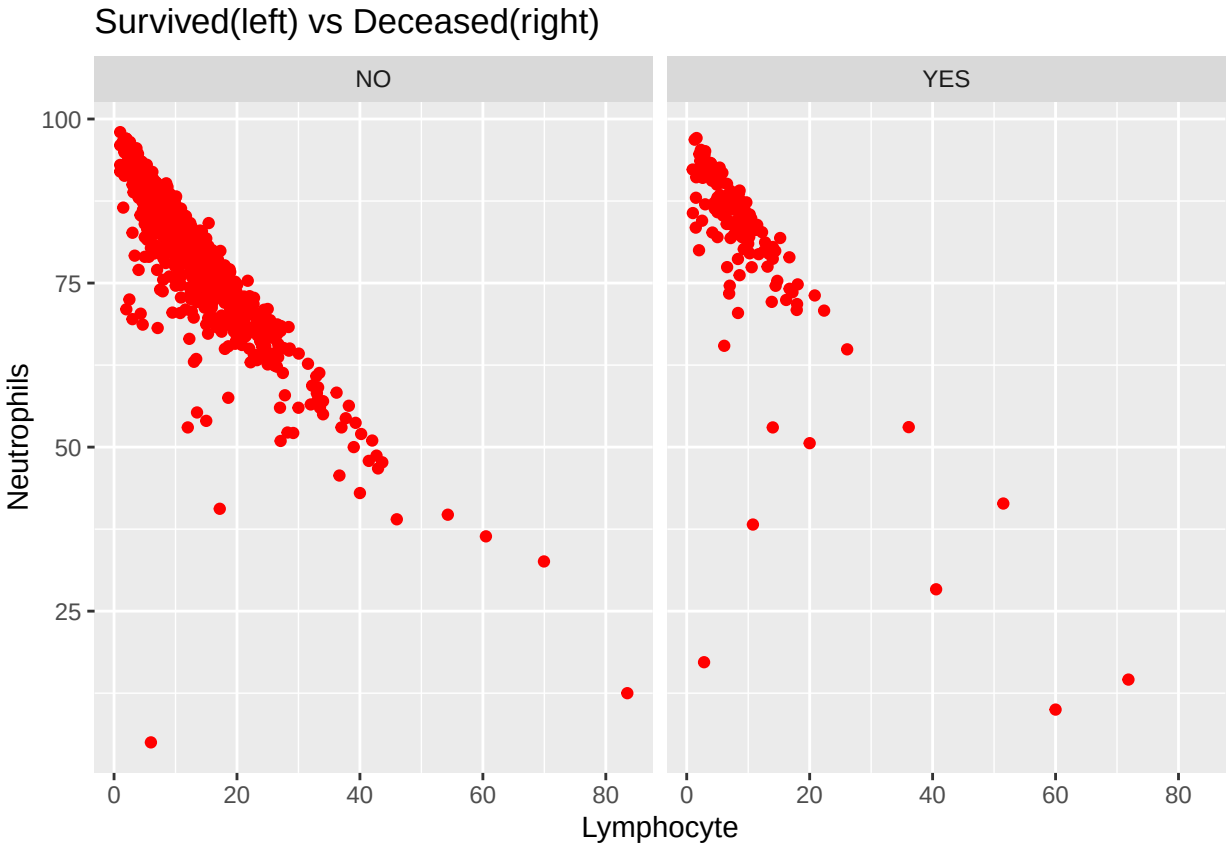
```
## Warning: Removed 145 rows containing missing values (geom_point).
```



Nexy, We separate the plots and facet as follows

```
ggplot(data=data) + geom_point(mapping=aes(x=Lymphocyte,y=Neutrophils),color="red") + facet_wrap(data$ou
  ggtitle("Survived(left) vs Deceased(right)")
```

```
## Warning: Removed 145 rows containing missing values (geom_point).
```



Conclusions

We investigated the patient variables and their dependencies from the MIMIC-III dataset.
* Age was found to have a marked effect on a person's BMI. * Neutrophils and Lymphocytes have a strong dependence * Most of the variables are independent and show no relationship with each other.

Domain Learning

- Neutrophils and Lymphocytes are examples of White Blood Cells
- Neutrophil-Lymphocyte ratio is used as a measure of
- Medical science is replete with opportunities to make an impact with data analysis.
- The MIMIC-III dataset is an example of a large medical database of clinical treatments.
- More data can be obtained from physionet.org

Future Ideas

- Use a linear/logistic regression model to predict mortality from the other variables.
- Add a variable of Neutrophil-Lymphocyte ratio and demonstrate a correlation with patient mortality.
- Use the MIMIC-III full database and compare to the results from this subset.