Student (s) Number as per student card:

10595913	Pavan Seerapu

Course Title: MSc in Data Analytics

Lecturer Name: Dr. Shahram Azizi Sazi

Module/Subject Title: Statistics for Data Analytics

Assignment Title: Design and implementation of Statistical Learning methods

No of Words: 3350

Contents

Individual Contribution

As a team leader, I assist my teammates in understanding the problem statements given to us and guide them throughout the assignment using the statistical analysis concepts taught by our lecturer Dr Shahram Azizi Sazi. Each team member is assigned to a specific task in generating a report, and I am assigned with descriptive analytical techniques and dataset implementation. Our first meeting was on November 2, 2021, and we worked out the criteria for the issue statement. Since then, my expertise has grown by self-learning, reading articles, and books.

Later, I learned that we may model a variable using sample data using probability distributions such as the Binomial, Normal, and Poisson distributions, which will be covered at our second meeting on November 9th, 2021.In terms of hypothesis testing, I discovered that we could determine whether or not there is a link between the independent variables. Furthermore, when researching which testing strategies to utilize, we came across the Wald test, Shapiro-Walk test, and Anova test on November 16th, 2021. Despite the fact that we battled while learning these distributions and hypothesis testing, we eventually caught up and finished this report.

I obtained summaries of the datasets that we worked on using descriptive statistics, and based on the summaries, we can go to the following phases, which are probability distributions and hypothesis testing. I even attempted to display a graphical depiction of the data in order to have a better understanding of the data. I use the binomial distribution, which was explained in class.

Conclusion

Our first meeting took place on November 2, 2021, and we developed criteria for the declaration of issue. I later learned that we can model a variable using sample data using probability distributions, such as the Binomial, Normal, and Poisson distributions, which we will address at our second meeting on November 9, 2021. Despite the fact that we are learning these distributions and testing hypotheses. we finally caught up and finished this report. I obtained the summaries of the datasets we worked on using descriptive statistics, and based on the summaries, we can move on to the following phases, which are probability distributions and hypothesis testing.

Code

https://colab.research.google.com/drive/1dxKKX8nLljg-6kK_fNCSmkwJPgRrgpvt?usp=sharing

Note

Below are our group meetings

2nd November 2021 – 1st Meeting

9th November 2021 – 2nd Meeting

16th November 2021 – 3rd Meeting

ICU Data Analysis using Descriptive statistics

Abstract

Analysing the mortality rate of intensive care unit patients has been a challenging task. Due to the unique characteristics of the records, the prediction of this mortality rate is difficult. Although data mining is becoming more prevalent in the healthcare industry, a comprehensive literature review on the subject has not yet been completed. This study aims to analyse the early hospital mortality of patients in an intensive care unit. It is proposed that a hybrid framework is used to improve prediction performance. During the first 48 hours of an intensive care unit admission, we analysed the performance of various data mining approaches.

Introduction

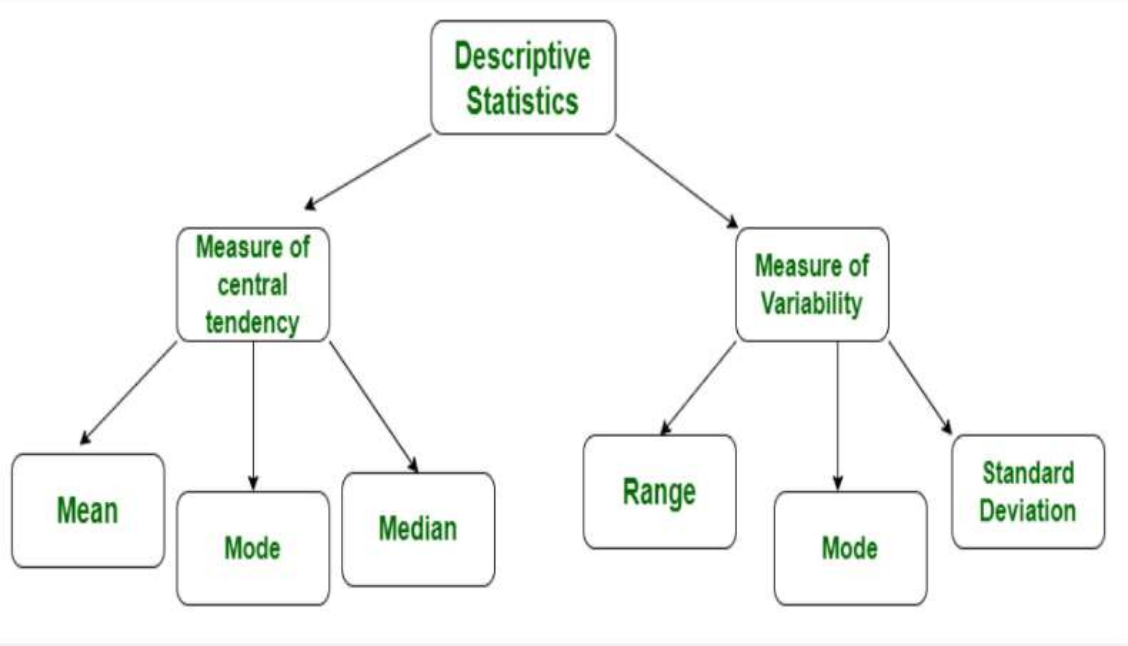
ICUs are an essential component of today's health-care system. Improving ICU performance needs a paradigm change away from a focus on individual performance and toward a focus on analyzing and improving ICU systems and procedures. This is the first installment in a two-part series on the subject. It discusses contemporary concerns in intensive care as well as ways for defining and measuring ICU performance. Descriptive statistics give concise descriptions of the sample and results. These summaries might be numerical, such as summary statistics, or visual, such as simple graphics. These summaries might serve as the foundation for an initial description of the data as part of a larger statistical research, or they could suffice for a single study.

Mortality after ICU discharge accounted for around 20–30% of all fatalities. We sought to examine if the presence and degree of organ dysfunction/failure shortly before to ICU discharge is associated with post-ICU mortality. Patients and medical procedures: The EURICUS-II database was used in the study, which comprised 4621 patients, 2958 of whom were released alive to general wards (post-ICU mortality 8.6 percent). Over a four-month period, we collected clinical and demographic data, including the Simplified Acute Physiology Score (SAPS II), Nine Equivalents of Nursing Manpower Use Score, and Sequential Organ Failure Assessment (SOFA) score.

Descriptive statistics provides concise descriptions of the sample and results. These summaries might be numerical, such as summary statistics, or visual, such as simple graphics. These summaries might serve as the foundation for an initial description of the data as part of a larger statistical research, or they could suffice for a single study. Statistics is a scientific process that collects, organizes, analyzes, and interprets data in order to characterize and make conclusions. There are two types of statistical techniques.

- **Descriptive Statistics** - This branch of statistics deals with the display of numerical facts or data in tables or graphs, as well as the methods for analysing the data.
- **Inferential Statistics** - This category includes techniques for inferring information about the entire population from observations acquired from samples.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton worked together to construct a model comparable to Yann LeCun's.



Some Standard Terminologies

- **Population** - A population is a group of people from whom information will be gathered.
- **Sample** - A sample is a small portion of a larger group.
- **Variable** - A variable is a quality or quantity trait that differs from person to person in a population.
- **Quantitative variable** - A quantitative variable, such as a person's weight or the number of passengers in a car, varies in magnitude.
- **Qualitative variable** - A qualitative variable, often known as an attribute, is a variable that varies in quality, such as the color of a car after an accident.
- **Discrete variable** - A discrete variable is one that cannot be anticipated to have a value between two defined values, such as the number of children in a household.
- **Continuous variable** - A continuous variable can have any value between two given values, for as the time for a 100-meter run.

Approach

Analysis of Univariate Data

We'll utilize Univariate analysis to define the distribution of a single variable, including its central tendency (mean, median, and mode) and dispersion (including the data set and spread measurements like variance and standard deviation). Indices such as skewness and kurtosis can also be used to define the shape of the distribution. Histograms and stem-and-leaf displays are graphical or tabular representations of properties of a variable's distribution.

```
[ ]  #Categorical Variables Univariate Analysis
     %%R
     table(ICUData_df$surgery)
```

```
cardiothoracic gastrointestinal            neuro            other
           223               79               46              121
       trauma
           31
```

```
[ ]  %%R
     table(ICUData_df$surgery)/nrow(ICUData_df)
```

```
cardiothoracic gastrointestinal            neuro            other
         0.446             0.158            0.092            0.242
       trauma
```

Bivariate Analysis

When a sample comprises more than one variable, descriptive statistics may be performed to characterize the relationship between the variables. In this example, descriptive data include:

Descriptive Statistics for Categorical Data

Descriptive statistics are the initial data points used to comprehend and portray a dataset. In essence, their purpose is to use concise summaries to highlight the important aspects of numerical and categorical data. These summaries can take many different forms, such as a single quantitative measure, summary tables, or a graphical depiction. The most frequent kind of descriptive statistics for categorical data is represented here, however there are various ways to describe and portray essential data properties.

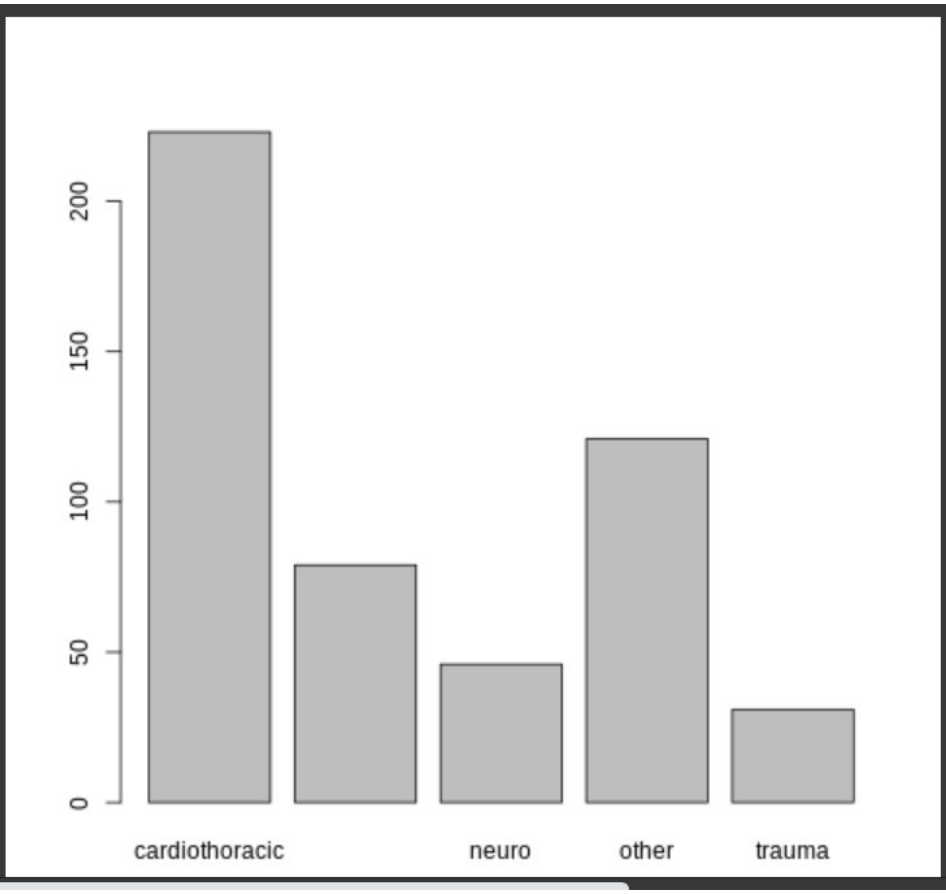
Frequencies

Use R's table () function to produce contingency tables that generate counts for each combination of category variables. We'd want to know, for example, how many female and male clients there are:

```
[ ] %%R
    Freq(ICUData_df$surgery)
```

	level	freq	perc	cumfreq	cumperc
1	cardiothoracic	223	44.6%	223	44.6%
2	gastrointestinal	79	15.8%	302	60.4%
3	neuro	46	9.2%	348	69.6%
4	other	121	24.2%	469	93.8%
5	trauma	31	6.2%	500	100.0%

```
[ ] %%R
    barplot(table(ICUData_df$surgery))
```



Proportions

We may also generate contingency tables with the percentages (proportions) of each category or group of categories. Table () creates frequency tables, which are then supplied to the prop. table () is a function. The tables that follow are identical to the ones that came before them, except that they compute proportions rather than counts.

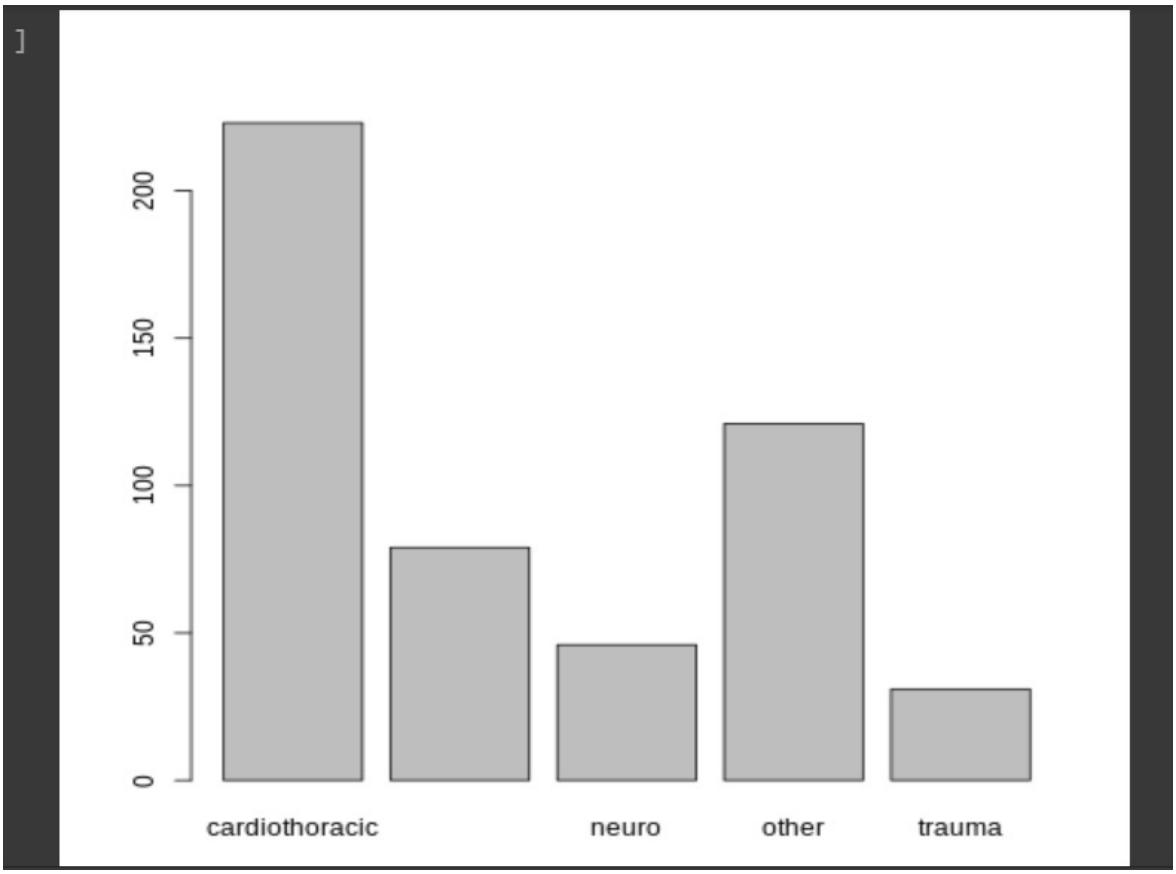
Marginals

In a contingency table, margins show the total counts or percentages across columns or rows. Consider table3, which shows the gender and marital status cross-classification data. Using the margin parameter, we may compute marginal frequencies using margin. table () and prop were used to calculate the percentages for these marginal frequencies. table ().

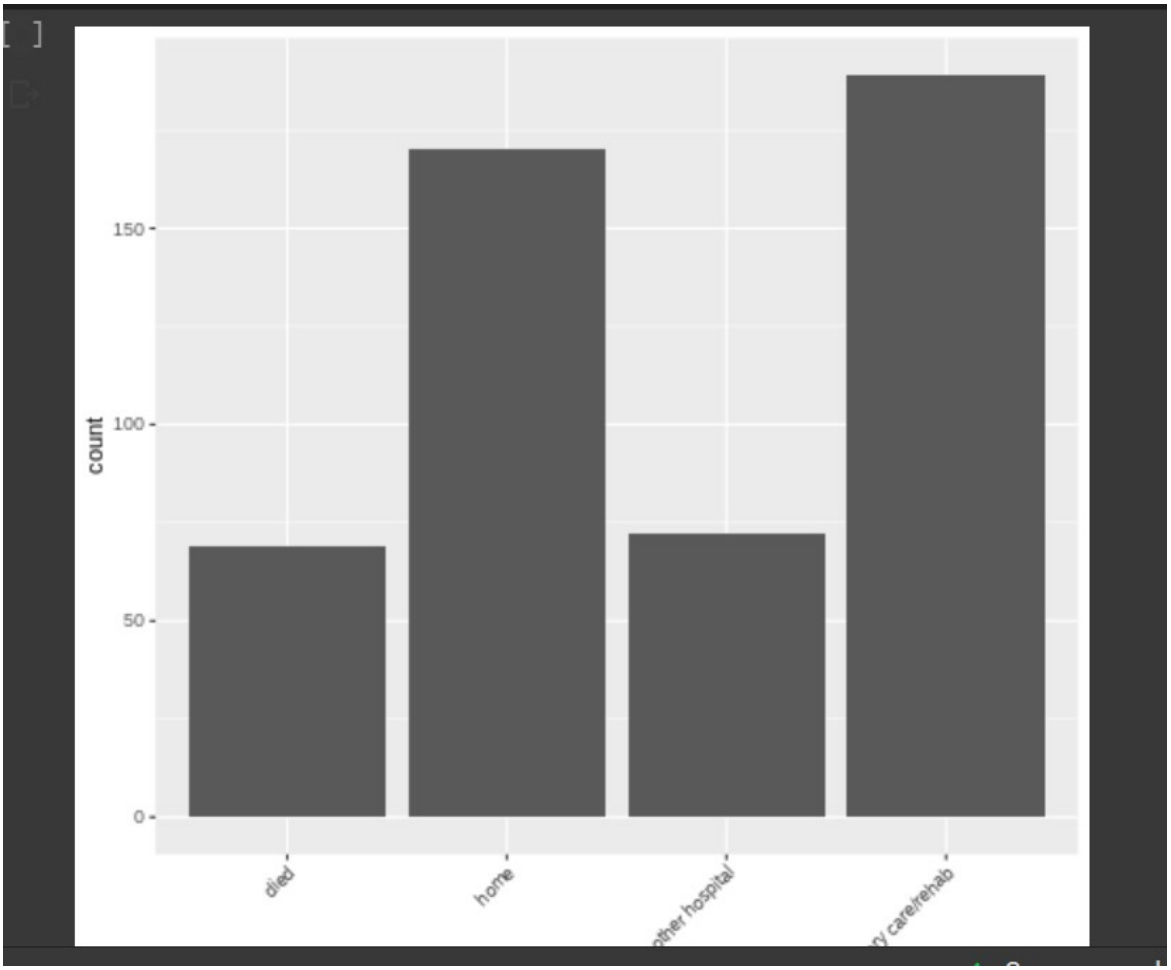
Visualization of categorical data

A bar chart is the most frequent technique to depict category data. Here's how we can figure out how many customers are at each location.

```
[ ]  %%R
     barplot(table(ICUData_df$surgery))
```

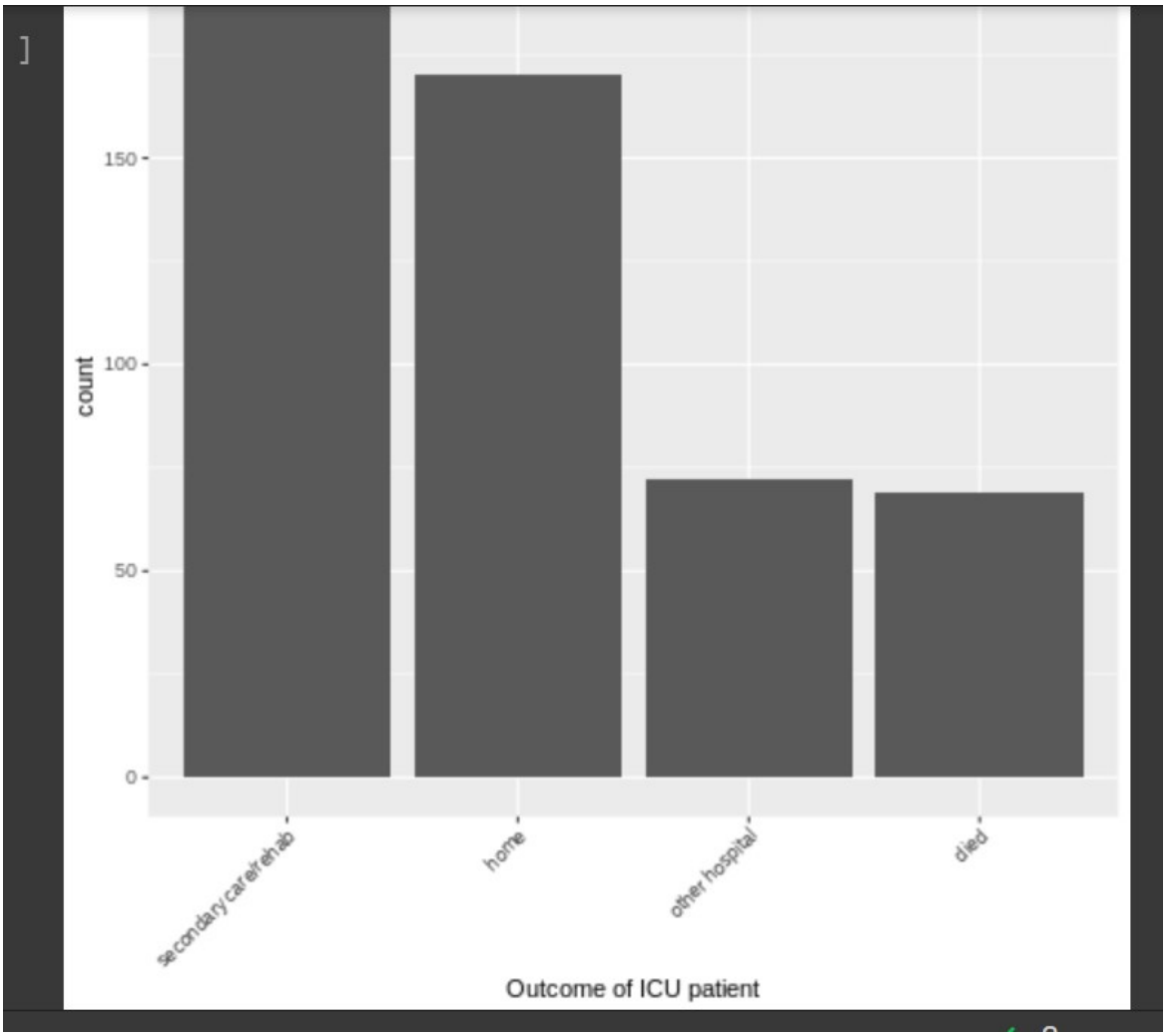


```
[ ]  %%R
     ggplot(ICUData_df, aes(x = `outcome`)) +
             geom_bar() +
             theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

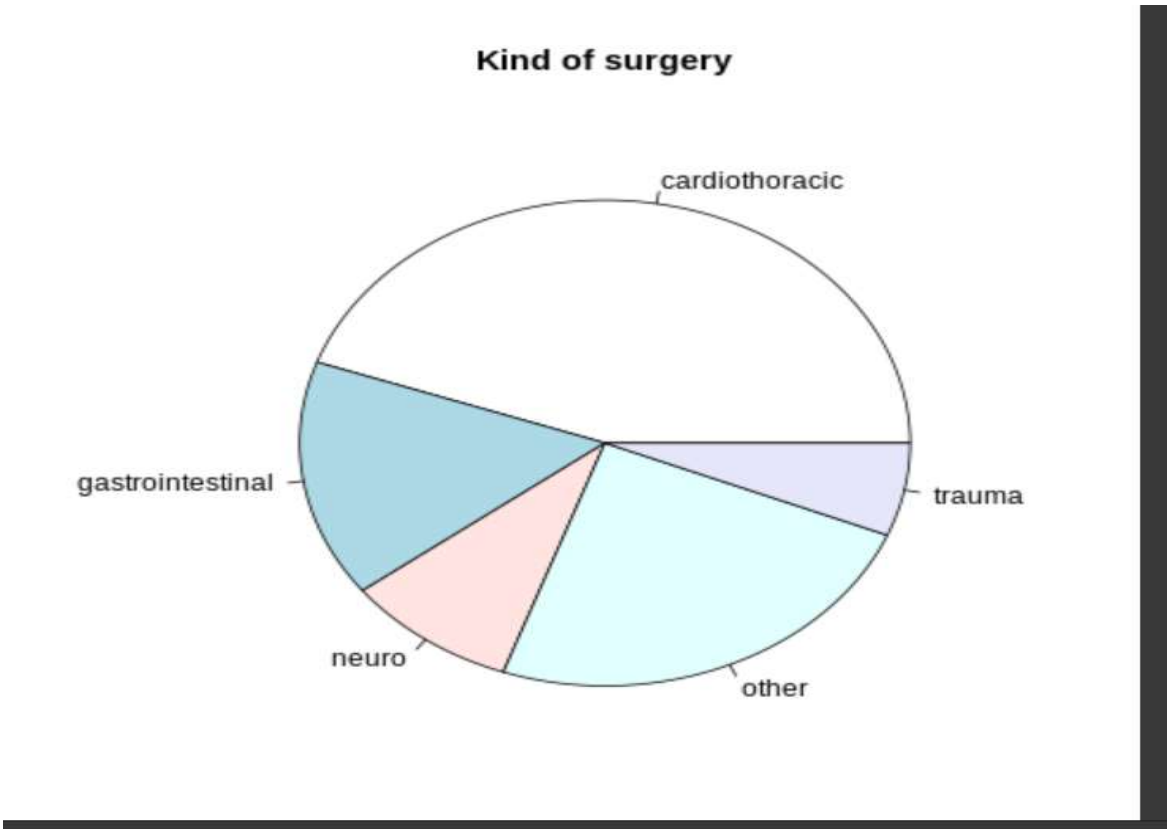



```
[ ]  # re-order levels
     %%R
     reorder_size <- function(x) {
             factor(x, levels = names(sort(table(x), decreasing = TRUE)))
     }

     ggplot(ICUData_df, aes(x = reorder_size(`outcome`))) +
             geom_bar() +
             xlab("Outcome of ICU patient") +
             theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
[ ]  %%R
     pie(table(ICUData_df$surgery), main = "Kind of surgery")
```



Descriptive Statistics for Numeric Data

Descriptive statistics are the initial data points used to comprehend and portray a dataset. In essence, their purpose is to use concise summaries to highlight the important aspects of numerical and categorical data. These summaries can take many different forms, such as a single quantitative measure, summary tables, or a graphical depiction. The most frequent sort of descriptive statistics for numerical data is depicted here, however there are other more techniques of explaining and depicting important data characteristics.

Central Tendency Metrics

There are three basic central tendency metrics, each of which attempts to answer the issue of which number is the most "typical." The mean (average of all observations), median (middle observation), and mode are the three metrics (appears most often). These metrics can be computed for a single variable or for the entire dataframe.

Metric Variable's Univariate Analysis

```
[ ]  %%R
     mean(ICUData$temperature)
```

```
     [1] 4.880826
```

```
[ ]  %%R
     round(mean(ICUData_df$temperature), 1)
```

```
     [1] 37.7
```

```
[ ]  %%R
     median(ICUData_df$temperature)
```

```
     [1] 37.7
```

Variability

The central trends indicate the most common values (in this example, salaries), but they do not indicate how variable those values are. Variability may be expressed in a few ways, each of which offers a distinct perspective on the distribution of values.

```
[ ]  %%R
     log(Gsd(ICUData_df$bilirubin))
     sd(log(ICUData_df$bilirubin))
```

```
     [1] 0.7238379
```

```
[ ]  %%R
     Skew(ICUData_df$temperature[-398])
```

```
     [1] 0.3142909
```

Range

A simple measure of variation that gives the maximum and minimum values, as well as the difference between them. To compute range summaries, use the following formula:

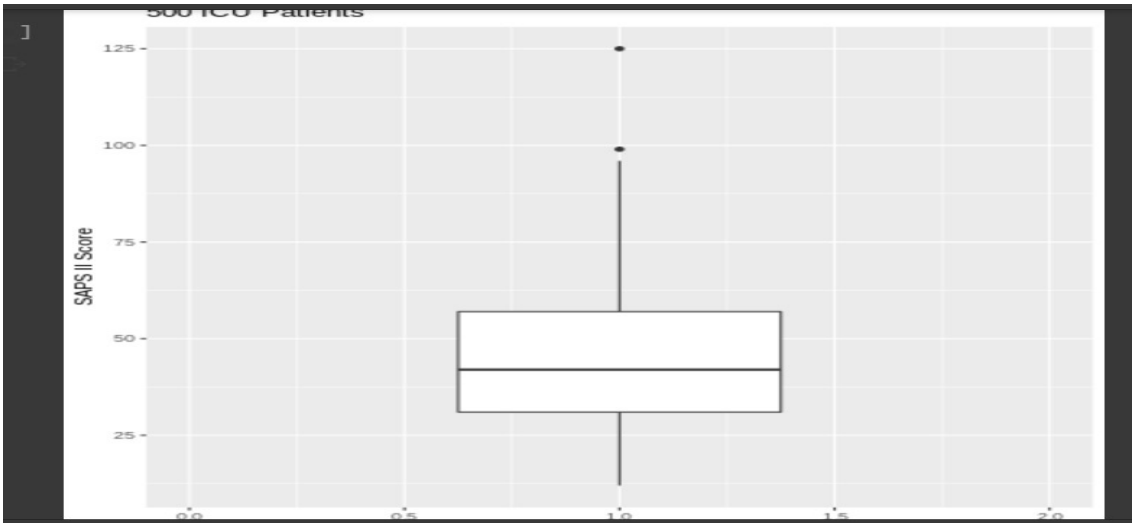
Percentiles

What is the monetary worth of compensation for a given percentage, such as 25%, such that this percentage of wage is less than it? Percentiles and quartiles are employed in this type of inquiry. For every percentage p, the pth percentile is the value at which a percentage p of all values is less than it. The first, second, and third quartiles correspond to the percentiles corresponding to p=25%, 50%, and 75%, respectively. These three integers split the data into four groups, each containing (about) a quarter of all observations. By definition, the median equals the second quartile. These metrics are easy to compute in R:

Outliers

Forecasts can be influenced by data outliers, which reduces their accuracy. As a result, understanding whether outliers exist and, if so, whether observations are classified as outliers is critical. The outliers package includes a variety of helpful methods for removing outliers in a systematic manner. Outlier () and scores are the most commonly utilized functions (). The outlier () method locates the most severe outlier in a collection of data. The scores () function generates a normalized (z, t, chisq, etc.) score that may be used to identify observations that are out of range.

```
[ ]  %%R
     ggplot(ICUData_df, aes(x = 1, y = SAPS.II)) +
       geom_boxplot() + xlim(0, 2) + ylab("SAPS II Score") +
       ggtitle("500 ICU Patients")
```



Normalization

It's a rescaling and shifting method that moves and rescales data points till they're in the 0 to 1 range. This strategy is described by the phrase min-max scaling.

The formula for computing the normalized score is as follows:

X new = (X — X min)/ (X max — X min)

Xmax and Xmin denote the maximum and lowest values of the feature, respectively.

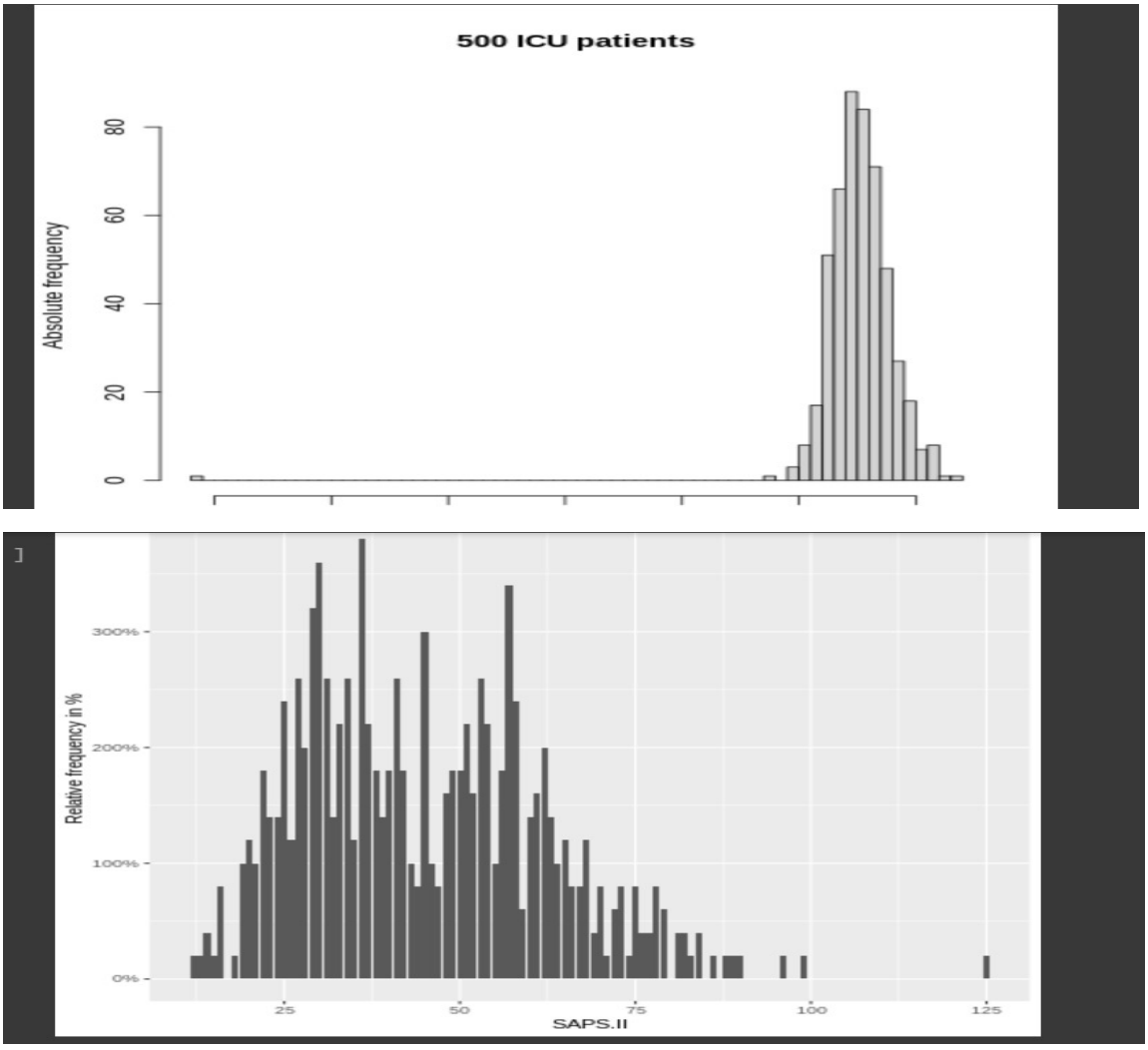
If Xmin=Xnew, Xnew equals 0.

The denominator will be 0 since the numerator will be Xmin –Xmin.

If Xmax=Xnew, Xnew equals 1.

In this scenario, both the numerator and denominator will be equal, canceling each other out to give us Xnew =1.

Isn't it a little too complicated? Let's have a look at an example to understand how this works in practice.



Standardization

Another scaling approach is standardization, in which data are centered around the mean with a single standard deviation. The mean and standard deviation of standard scores will be 0 and 1, respectively, if we compute them.

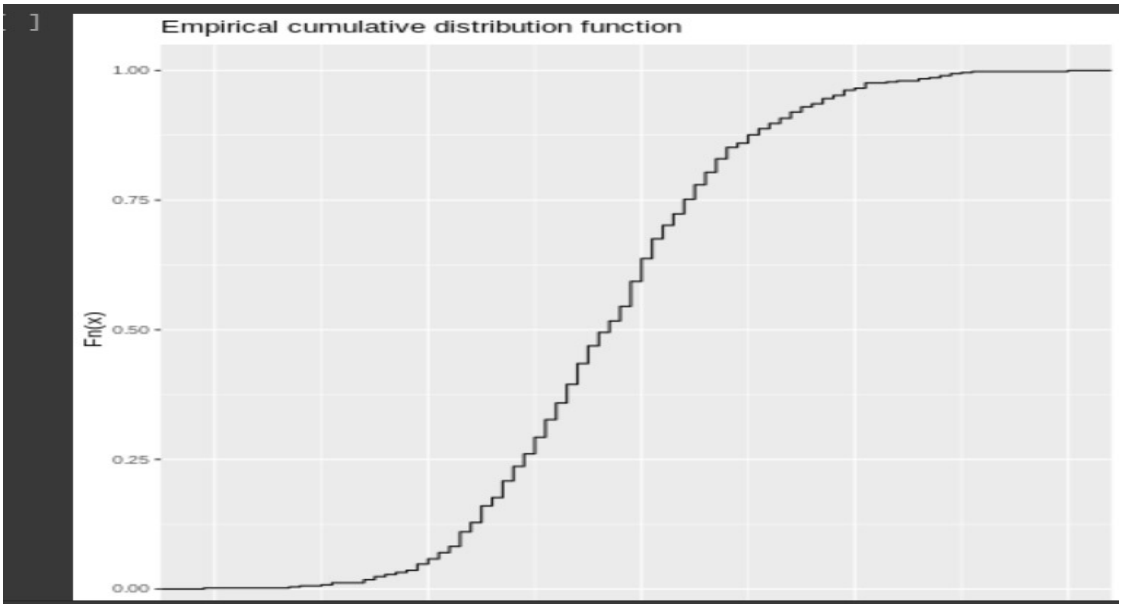
The following formula is used to calculate standardized values:

$$z = \frac{x - \mu}{\sigma}$$

Probability Distributions

Probability distributions are statistical functions that represent the chance of a random variable taking any of its potential values. To put it another way, the variable's values are determined by the underlying probability distribution. Let's say you're measuring the heights of a random group of people. You may make a height distribution by measuring people's heights. When you need to know which outcomes are most likely, the range of potential values, and the probability of distinct possibilities, this form of distribution is important.

```
[ ]  %%R
     ggplot(ICUData_df[-398,], aes(x = temperature)) + stat_ecdf() +
       xlab("Maximum body temperature") + ylab("Fn(x)") +
       ggtitle("Empirical cumulative distribution function")
```



Binomial Distribution

The probability distribution of X, the number of successes in N independent experiments, each of which yields success with probability p, is called the binomial distribution. The binomial distribution is the discrete probability distribution of the number of successes in a sequence of N independent experiments, each with probability of success p. The number of successes in a sequence is called a Bernoulli trial. The binomial distribution is a special case of a binomial distribution. It is a discrete random variable denoting the number of successes in a sequence of n Bernoulli trials, each with probability p of success.

The binomial distribution formula is calculated as:

$$P_{(x:n,p)} = {}_nC_x \, x \, p^x (1-p)^{n-x}$$

```
[ ]  #Binomial distribution
     %%R
     X=ICUData_df$heart.rate
     T=table(X)
     T
     pmf=dbinom(T,10,0.3) # pmf
     cdf=pbinom(T,10,0.3)  # cdf
     df=data.frame(T,pmf,cdf)
     head(df)
```

```
      X Freq  X.1     Freq.1  X.2     Freq.2
1 59.2    1 59.2 0.1210608 59.2 0.1493083
2   62    1   62 0.1210608   62 0.1493083
3   63    2   63 0.2334744   63 0.3827828
4   64    1   64 0.1210608   64 0.1493083
5 67.4    1 67.4 0.1210608 67.4 0.1493083
6 67.5    1 67.5 0.1210608 67.5 0.1493083
```

Normal Distribution

Data in a normal distribution is symmetrically distributed with no skew. When displayed on a graph, the data has a bell shape, with most values clustering in a core region and tapering off as they go away from the center.

The Empirical rule

The empirical rule, often known as the 68-95-99.7 rule, indicates where the majority of your values fall in a normal distribution:

- Approximately 68 percent of the results are within one standard deviation of the mean.
- 95 percent of the data are within two standard deviations of the mean.
- Approximately 99.7 percent of data are within three standard deviations of the mean.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $f(x)$ = probability
- x = value of the variable
- μ = mean
- σ = standard deviation
- σ^2 = variance

```
%%R
#Normal Probability distribution
pnorm(104, heart_avg, heart_ssd)
```

```
[1] 0.4349846
```

Poisson Distribution

The Poisson Process is the model we use to describe randomly occurring events, even though it is not particularly helpful on its own. We need the Poisson Distribution to accomplish interesting things like calculate the chance of a lot of events occurring in a certain time period or calculate the likelihood of waiting a certain amount of time until the next event occurs.

Given the duration of the period and the average number of events per time, the Poisson Distribution probability mass function calculates the likelihood of seeing k events in a time period:

$$P(k \text{ events in time period}) = e^{-\frac{\text{events}}{\text{time}} * \text{time period}} * \frac{(\frac{\text{events}}{\text{time}} * \text{time period})^k}{k!}$$

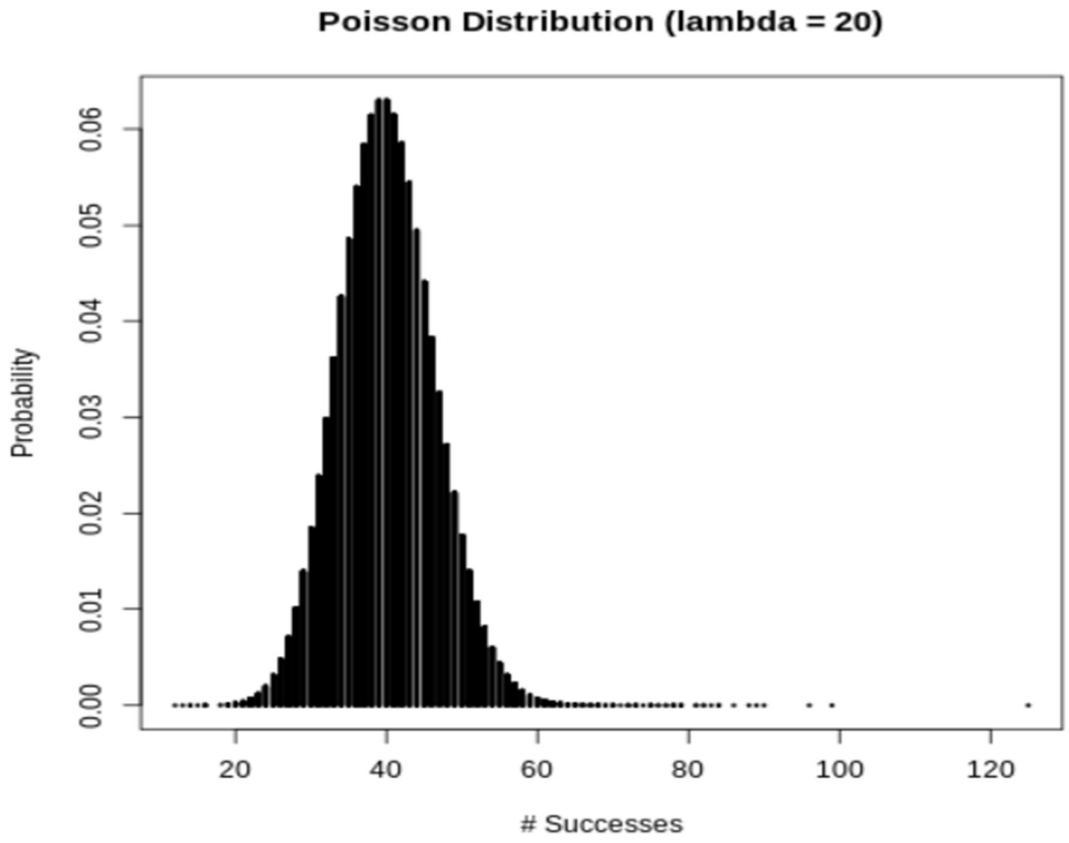
Poisson distribution for probability of k events in time period.

This is a little complicated, and the rate parameter, lambda, is frequently condensed into a single parameter, events/time * time period. The Poisson Distribution probability function now contains one parameter as a result of this change:

$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Poisson distribution probability of k events in an interval.

```
%%R
plot(ICUData_df$SAPS.II
, dpois(ICUData_df$SAPS.II
, lambda=40),
     type='h',
     main='Poisson Distribution (lambda = 20)',
     ylab='Probability',
     xlab ='# Successes',
     lwd=3)
```



Hypothesis testing

Hypothesis testing is a statistical approach in which an analyst confirms a hypothesis about a population parameter. The analyst's technique is influenced by the type of data and the study's purpose. The practice of evaluating the plausibility of a hypothesis using sample data is known as hypothesis testing. This sort of information might originate from a bigger population or a data collection technique. In the following examples, the term "population" will be used to describe both of these conditions.

Wald Test

The Wald test is a statistical hypothesis test used to compare two treatment groups. It is a nonparametric test, which means that it does not assume a normal distribution of data. It can be used when the groups being compared have different variances.

The Wald Test statistic formula is:

$$W_T = \frac{\left[\hat{\theta} - \theta_0\right]^2}{1/I_n(\hat{\theta})} = I_n(\hat{\theta})\left[\hat{\theta} - \theta_0\right]^2$$

Where:

- $\hat{\Theta}$ = Maximum Likelihood Estimator (MLE),
- $I_n(\hat{\theta})$ = expected Fisher information (evaluated at the MLE).

Wald test statistic was used to do hypothesis testing

H0: There is no difference in mean heart rate of died and home patient in ICU

H1: Mean heart rate of the 2 patients of ICU's is different

The W statistic is 0.05574695

The p value is 0.9555434

At α=0.05, this implies that we have enough evidence not to reject the null that there is no dif ference in mean heart rate of two type of ICU Patient

```
%%R
died=ICUData_df$heart.rate[ICUData_df$outcome=="died"]
home=ICUData_df$heart.ratee[ICUData_df$outcome=="home"]
n.died=length(died)
n.died
```

```
[1] 69
```

```
%%R
mu.died=mean(died)
sigma.died<-sd(died)
n.home=length(home)
n.home
```

```
[1] 0
```

```
%%R
mu.home=mean(home)
sigma.home<-sd(home)
mu_hat=mu.died-mu.home
mu_hat
```

```
[1] NA
```

Shapiro-Wilk

The Shapiro-Wilk test first calculates the degree of similarity between the observed and normal distributions as a single number by superimposing a normal curve over the observed distribution, as illustrated below. It then computes the proportion of our sample that overlaps with it, resulting in a similarity percentage.

Finally, the Shapiro-Wilk test computes the likelihood of obtaining this observed -or a lower - similarity percentage. It does so under the null hypothesis, which assumes that the population distribution is perfectly normal.

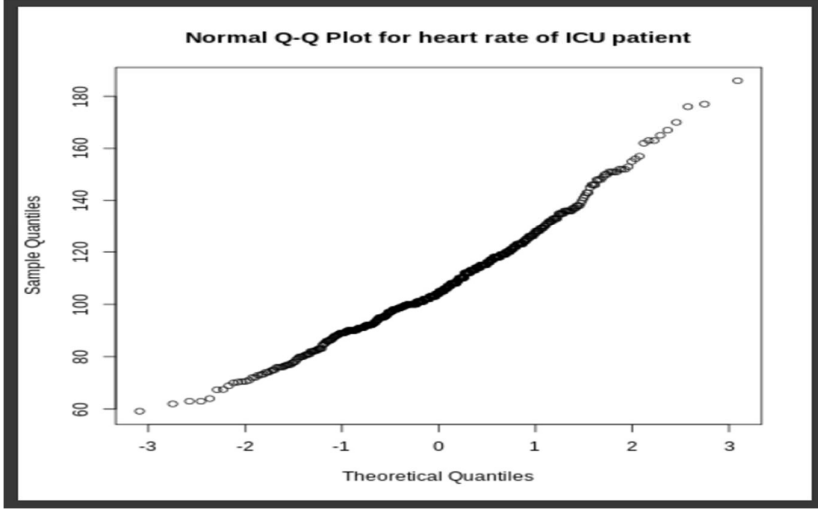
$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

```
%%R
shapiro.test(ICUData_df$heart.rate)

        Shapiro-Wilk normality test

data:  ICUData_df$heart.rate
W = 0.97873, p-value = 1.123e-06
```

```
%%R
qqnorm(ICUData_df$`heart.rate`, main = "Normal Q-Q Plot for heart rate of ICU patient")
```



Anova Testing

ANOVA is a test of variances. ANOVA determines whether differences between groups of data are statistically significant. It analyzes the levels of variation within the groups using samples drawn from each of them.

The approach for testing for a difference in more than two independent means is an extension of the previously stated two independent samples process, which applies when there are precisely two separate comparison groups. When there are two or more independent groups, the ANOVA approach is used. However, because there are more than two groups, the computation of the test statistic is more complicated. The sample sizes, means, and standard deviations must all be included in the test statistic.

The test statistic for testing H_0: $\mu_1 = \mu_2 = ... = \mu_k$ is:

$$F = \frac{\Sigma n_j(\bar{X}_j - \bar{X})^2/(k-1)}{\Sigma\Sigma(X - \bar{X}_j)^2/(N-k)}$$

```
%%R
#One way Anova test
one.way <- aov(age ~ heart.rate, data = ICUData_df)

summary(one.way)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
heart.rate    1     44   44.07   0.202  0.653
Residuals   498 108764  218.40
```

```
%%R
#Two way Anova test
two.way <- aov(age ~ heart.rate + temperature, data = ICUData_df)

summary(two.way)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
heart.rate    1     44    44.1   0.202 0.6529
temperature   1    601   600.5   2.759 0.0973 .
Residuals   497 108164   217.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Simulations:

The Dataset's Introduction:

Kaggle was used to acquire the input data for the modeling. Each data collection contains around 500 unique patients. The parameters that may be used to predict a patient's outcome were examined.

```
[4]  %%R
     ICUData_df = read.csv('ICUData.csv')
     head(ICUData_df)
```

```
       ID    sex age          surgery heart.rate temperature bilirubin SAPS.II
     1  1 female  76            other       98.0        36.5  6.512142       57
     2  2 female  60 gastrointestinal       80.0        38.1 14.523197       52
     3  3   male  66   cardiothoracic       99.6        37.4 22.972480       57
     4  4   male  74            other      110.0        39.1 19.299346       45
     5  5 female  68            other       94.1        38.5 39.076485       49
     6  6   male  68   cardiothoracic       88.8        35.1 14.805941       53
       liver.failure LOS               outcome
     1             0   1                  died
     2             0   2                  home
     3             0   1 secondary care/rehab
     4             0   2                  home
     5             0   1                  home
     6             0   1 secondary care/rehab
```

Inspecting and evaluating the data We will plot numerous graphs and evaluate the data graphically.

Dim of data: The dim function returns the total number of columns and rows in the data.

```
[6]  %%R
     #Checking number of rows and column in data
     dim(ICUData_df)
```

```
     [1] 500  11
```

Data str: We have str () function to get the data information.

```
[5]  %%R
     str(ICUData_df)
```

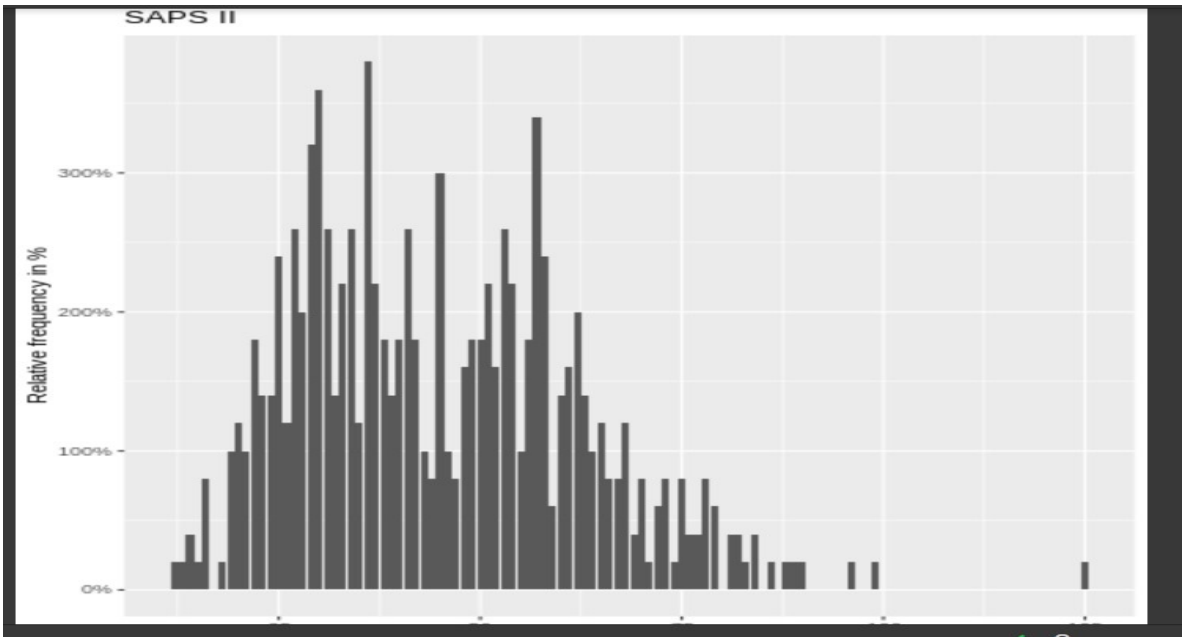
```
     'data.frame':   500 obs. of  11 variables:
      $ ID           : int  1 2 3 4 5 6 7 8 9 10 ...
      $ sex          : chr  "female" "female" "male" "male" ...
      $ age          : int  76 60 66 74 68 68 70 55 75 71 ...
      $ surgery      : chr  "other" "gastrointestinal" "cardiothoracic" "other" ...
      $ heart.rate   : num  98 80 99.6 110 94.1 88.8 102 106 109 102 ...
      $ temperature  : num  36.5 38.1 37.4 39.1 38.5 35.1 36.7 39.8 39.9 38.4 ...
      $ bilirubin    : num  6.51 14.52 22.97 19.3 39.08 ...
      $ SAPS.II      : int  57 52 57 45 49 53 25 19 58 56 ...
      $ liver.failure: int  0 0 0 0 0 0 0 0 0 0 ...
      $ LOS          : int  1 2 1 2 1 1 1 1 1 3 ...
      $ outcome      : chr  "died" "home" "secondary care/rehab" "home" ...
```

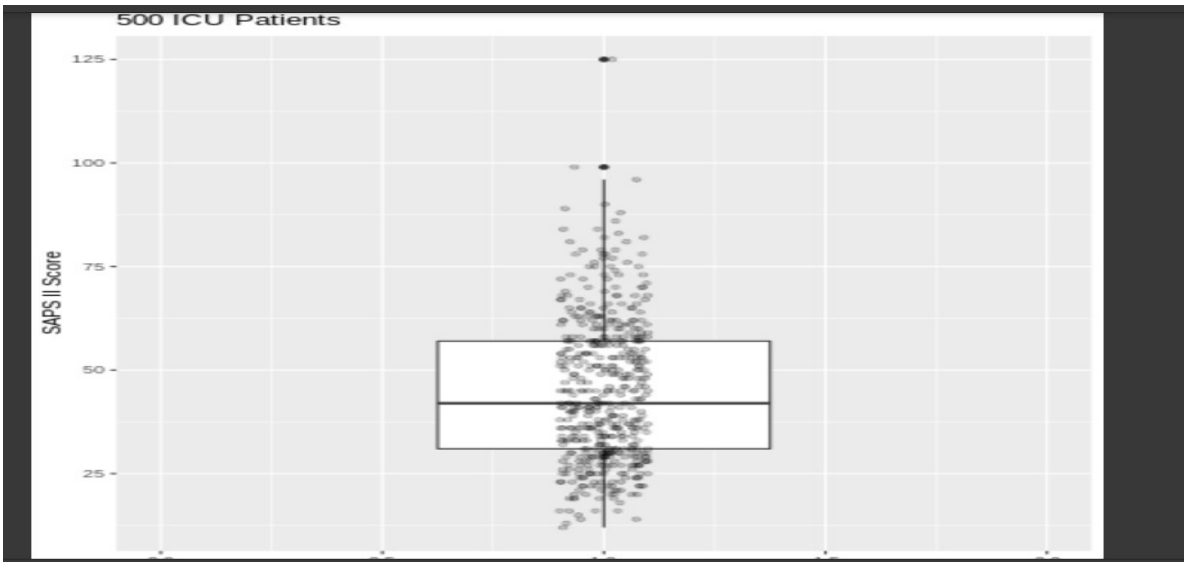
Data Summary - The summary () method provides us with statistical information regarding data.

```
%%R
#statical information of data
summary(ICUData_df)
```

```
       ID              sex                 age          surgery
 Min.   :  1.0   Length:500         Min.   :18.00   Length:500
 1st Qu.:125.8   Class :character   1st Qu.:55.00   Class :character
 Median :250.5   Mode  :character   Median :66.00   Mode  :character
 Mean   :250.5                      Mean   :63.06
 3rd Qu.:375.2                      3rd Qu.:73.00
 Max.   :500.0                      Max.   :98.00
   heart.rate      temperature       bilirubin          SAPS.II
 Min.   : 59.20   Min.   : 9.10   Min.   :  3.679   Min.   : 12.00
 1st Qu.: 92.72   1st Qu.:36.90   1st Qu.: 10.705   1st Qu.: 31.00
 Median :104.00   Median :37.70   Median : 15.312   Median : 42.00
 Mean   :107.41   Mean   :37.66   Mean   : 25.744   Mean   : 44.88
 3rd Qu.:119.00   3rd Qu.:38.40   3rd Qu.: 23.427   3rd Qu.: 57.00
 Max.   :186.00   Max.   :42.00   Max.   :647.305   Max.   :125.00
 liver.failure        LOS            outcome
 Min.   :0.00   Min.   :  1.00   Length:500
 1st Qu.:0.00   1st Qu.:  1.00   Class :character
 Median :0.00   Median :  1.00   Mode  :character
 Mean   :0.04   Mean   :  5.29
 3rd Qu.:0.00   3rd Qu.:  5.00
 Max.   :1.00   Max.   :105.00
```

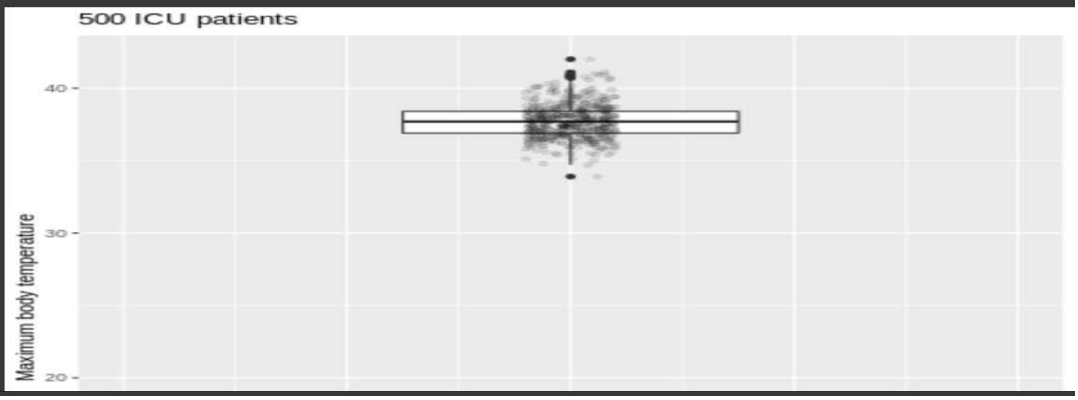
Data Explore and Visualization






```
[ ]  df.groupby(["MultipleLines", "Churn"]).size().unstack().plot(kind='bar', stacked=True, figsize=(5,5))
```

```
## Box-and-whisker plot at position x = 1
%%R
ggplot(ICUData_df, aes(x = 1, y = temperature)) +
  geom_boxplot() + xlim(0, 2) + ylab("Maximum body temperature
  geom_jitter(height = 0, width = 0.1, alpha = 0.1) +
  ggtitle("500 ICU patients")
```



Critical Analysis of results

Images were captured in a range of lighting settings and with various lenses. To address these challenges, data augmentation may be employed during model training. The introduction of Convolution layers, as observed, boosted the accuracy to 84.7 percent. By employing a bigger training dataset and fine-tuning the hyper parameters, the model's accuracy may be increased even more. According to the categorization study, cheetah and jaguar recall are lower than in other classes.

Conclusion

Improving ICU performance demands a paradigm change away from individual performance and toward ICU system and process development. Cumulative data must be gathered in order to create meaningful summary measures of performance in order to define and quantify performance in the ICU. Detecting unfavorable occurrences (errors) on an individual basis is inadequate. The intensive care unit (ICU) is a data-rich environment that lends itself nicely to analysis. Several scoring systems and DM methods have been developed to predict clinical worsening and death in the ICU. The bulk of these techniques, on the other hand, are designed to predict outcomes one or more days following admission. There have been no definitive studies to our knowledge that compare mortality prediction per hour over the first 48 hours of a patient's admission in order to designate to clinicians when is the optimal period for ICU data analysis.

References

A. https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

B. https://www.kaggle.com/grfiv4/plot-a-confusion-matrix

C. https://www.superdatascience.com/blogs/artificial-neural-networks-the-neuron

D. *Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in* Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.*

E. https://www.kaggle.com/iluvchicken/cheetah-jaguar-and-tiger

F. G. Chen, T. X. Han, Z. He, R. Kays and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 858-862, doi: 10.1109/ICIP.2014.7025172.

G. W. Xue, T. Jiang and J. Shi, "Animal intrusion detection based on convolutional neural network," 2017 17th International Symposium on Communications and Information Technologies (ISCIT), 2017, pp. 1-5, doi: 10.1109/ISCIT.2017.8261234.