Part I

A. Question

A1. Using a t-test can we determine if there is a significant difference between the average age of patients who were readmitted and those who were not?

A2. It is important to identify patients who are at risk for readmission. There are a growing number of insurance companies that will not reimburse hospitals for the care they provide for a patient who is readmitted within a certain time after discharge. In addition, readmission can stain the trust a patient has in a healthcare provider. Patients and their families may choose to receive care elsewhere if they feel that they were not treated fully during their initial admission. Any hospitalization puts a patient at risk for hospital acquired infections, medication errors, and death. So, we want to prevent hospitalization if at all possible, especially re-hospitalization. Stakeholders will benefit from this analysis as we may uncover an important correlation with readmission. If a correlation is uncovered then we will be able to further analyze this variable and potentially find ways to prevent readmission.

A3. The relevant data for the question presented in A1 includes the ReAdmis and the Age field. The independent variable is readmission status, which is contained in the ReAdmis field. ReAdmis is nominal, qualitative, categorical data with yes or no responses. A yes in this field indicates that the patient was readmitted within three months of discharge. The dependent variable is age. The Age field contains quantitative, continuous, numeric data representing the patient's age. This data is ratio data as there is a "true" zero, or the day that the patient was born.

B. Description of Data Analysis

B1. A paired t-test was run. The code is shown below:

```
#Load packages
library('tidyverse')
library(visdat)
library(ggplot2)
library(hrbrthemes)
library(dplyr)

#confirm working directory location
getwd()

#load file
df <- read_csv('medical_clean.csv')

#view data frame
spec(df)
head(df,5)

#verify no duplication
str(df)
duplicated(df)
sum(duplicated(df))

#verify no NA
colSums(is.na(df))
```

```
#Filter for readmission status
Readmit <- df %>% filter(ReAdmis == "Yes")
Noreadmit <- df %>% filter(ReAdmis == "No")

#Save ages of readmitted patients
x <- Readmit$Age

#Save Ages of patients not readmitted
y <- Noreadmit$Age

#count how many rows are in each group
nrow(Readmit)
nrow(Noreadmit)

#run paired ttest
t.test(x, y, var.equal=FALSE)
```

B2. The results of running t.test(x, y, var.equal=TRUE) as included below.

Results of Two Sample t-test

```
data:  x and y
t = 1.5837, df = 7700, p-value = 0.1133
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1609549  1.5149612
sample estimates:
mean of x mean of y
 53.94031  53.26331
```

I also calculated the value of T utilizing the following code:

```
x1 <- mean(Readmit$Age)
x2 <- mean(Noreadmit$Age)
n1 <- nrow(Readmit)
n2 <- nrow(Noreadmit)
s1 <- sd(Readmit$Age)
s2 <- sd(Noreadmit$Age)

xv<- (x1-x2)
a <- s1^2/n1
b <- s2^2/n2

t<- xv/sqrt(a+b)
print(t)
[1] 1.583743
```

This t value of 1.583743 is equal to what the ttest function calculated so this is acceptable.

A z-test was performed to validate the t-test findings, as demonstrated here:

```
z.test(Readmit$Age, Noreadmit$Age,
        alternative = "two.sided",
        mu = 0,
        sigma.x = sd(Readmit$Age),
        sigma.y = sd(Noreadmit$Age),
        conf.level = 0.95
)
```

Results:

Two-sample z-Test

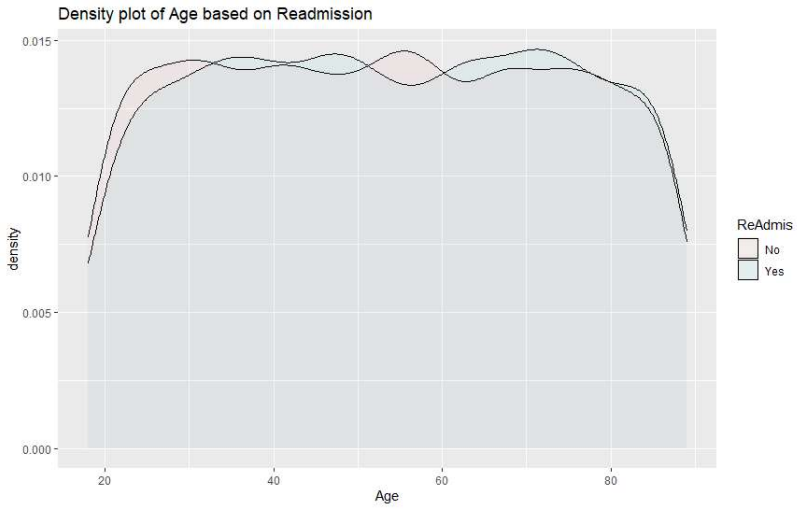
```
data:  Readmit$Age and Noreadmit$Age
z = 1.5837, p-value = 0.1133
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1608231  1.5148295
sample estimates:
mean of x mean of y
 53.94031  53.26331
```

The p-value of 0.1133 is the same as our t-test.

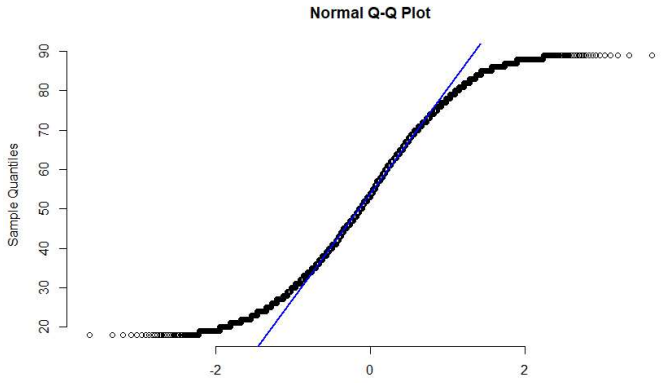
B3. A two-sample t-test was chosen to analyze this data. The question is; is the mean age of readmitted patients statistically different from the mean age of patients who were not readmitted? The analysis needs to compare the mean of two independent groups so a 2 sample t-test is appropriate. Anova would not work because Anova requires three or more groups. A one-sample t-test would also not be appropriate as we have two groups. Chi-squared requires ordinal data for the independent variable, since the independent variable chosen here (age) is ratio-level data, a t-test is more appropriate. By creating a QQ plot and density plot we can see that the data for Age is not perfectly normal but it is relatively normally distributed. Again this is appropriate for a t-test.

The following code was used to create a density plot to assess the normality:

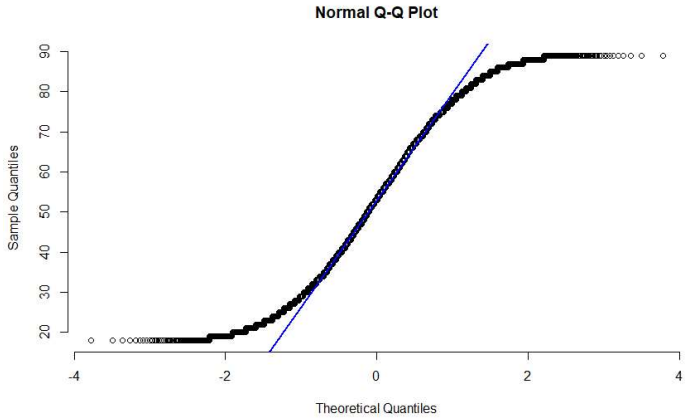
```
ggplot(df, aes(x=Age, fill=ReAdmis))+
  geom_density(alpha=0.05)+
  labs(title = "Density plot of Age based on
Readmission")+
  xlab("Age")
```



QQ plot of Ages of Readmitted Patients
code:
qqnorm(Readmit$Age, pch=1, frame=FALSE)
qqline(Readmit$Age, col="blue", lwd=2)



QQ plot of Ages of Patients who were not Readmitted
qqnorm(Noreadmit$Age, pch=1, frame=FALSE)
qqline(Noreadmit$Age, col="blue", lwd=2)



C. Univariate Statistics

Univariate Statistics of Age:
describe(df$Age)

```
    vars     n  mean    sd median trimmed  mad min max range skew kurtosis   se
X1     1 10000 53.51 20.64     53   53.49 26.69  18  89    71 0.01    -1.19 0.21
```

Univariate Statistics of Initial_days:
describe(df$Initial_days)

```
    vars     n  mean    sd median trimmed  mad min   max range skew kurtosis   se
X1     1 10000 34.46 26.31  35.84   34.04 39.83   1 71.98 70.98 0.07    -1.75 0.26
```

Univariate Statistics of Item1:
print(table(df$Item1))

```
   1    2    3    4    5    6    7    8
 213 1315 3404 3455 1377  225   10    1
```

cat("The mode is:", names(sort(-table(df$Item1)))[1])

```
The mode is: 4
```

Univariate Statistics of Item2:
print(table(df$Item2))

```
   1    2    3    4    5    6    7
 213 1360 3439 3351 1421  204   12
```

cat("The mode is:", names(sort(-table(df$Item2)))[1])

```
The mode is: 3
```

C1. **Continuous Variables**: Age, Initial_days.

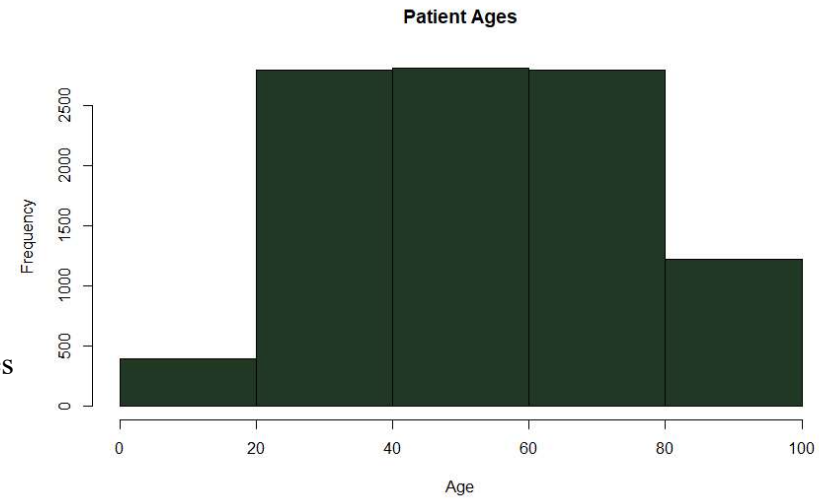
Categorical Variables: Item1(Patient ranking of the importance of timely admission), Item2 (Patient ranking of the importance of timely treatment). Despite the numeric output the ranking of importance of timely admission and timely treatment, this is ordinal categorical data.

Graph of Continuous Variable 1: Age

Graph Code:

```
hist(df$Age,
    main="Patient Ages",
    xlab="Age",
    border="black",
    col="#203824",
    breaks= 5)
```

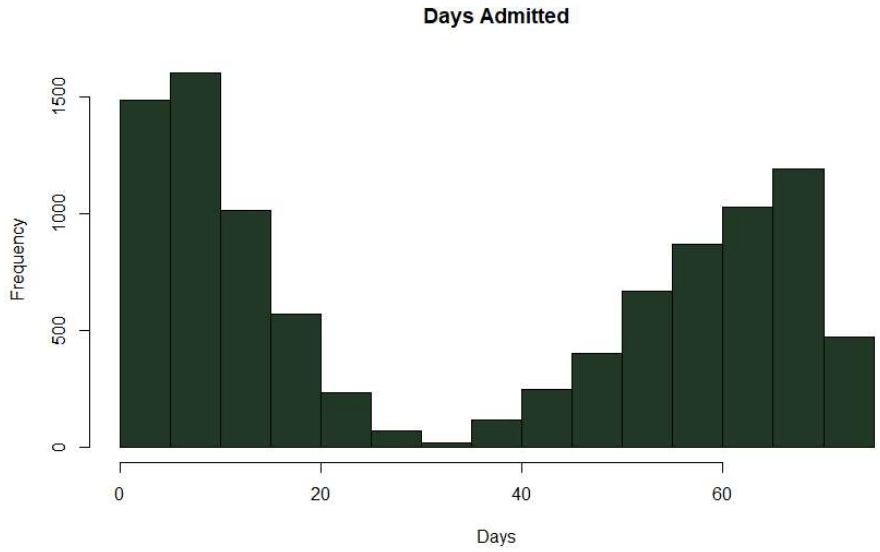
Distribution: The graph of patient ages shows a normal distribution.



Graph of Continuous Variable 2: Initial_Days

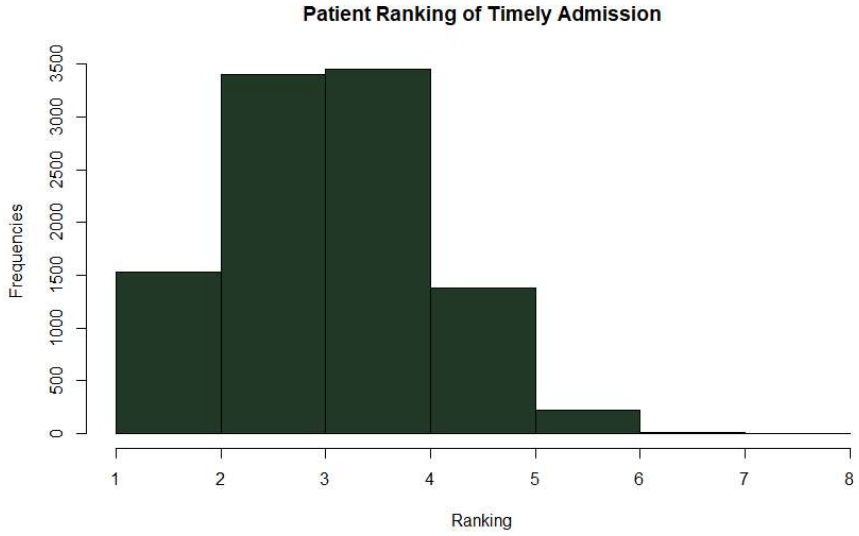
```
#Histogram of  Intial_days
hist(df$Initial_days,
    main="Days Admitted",
    xlab="Days",
    ylab = "Frequency",
    border="black",
    col="#203824",
    breaks=20)
```

Distribution: The graph of Initial_days shows bimodal distribution.



Graph of Categorical Variable 1: Item1

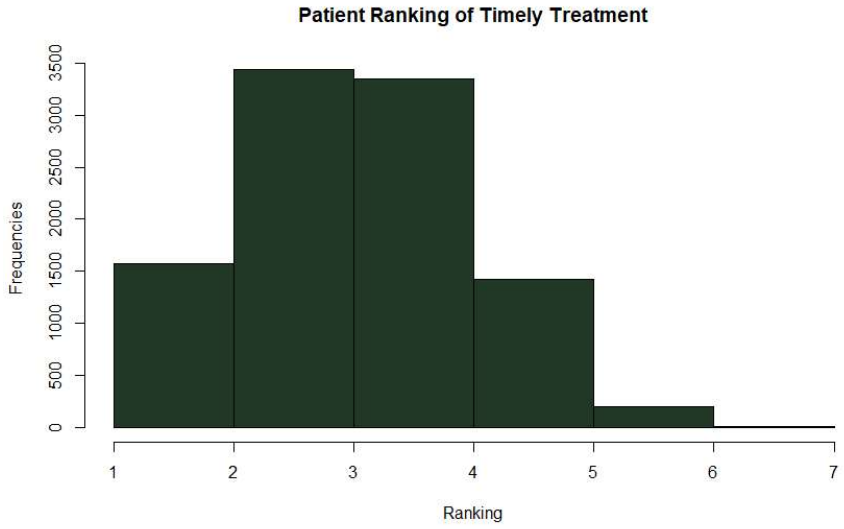
```
#Histogram of Item1
hist(df$Item1,
    main="Patient Ranking of
Timely Admission",
    xlab="Ranking",
    border="black",
    col="#203824",
    ylab = "Frequencies",
    breaks=9)
```



Distribution: The graph of Item1 shows a left-skewed distribution.

Graph of Categorical Variable 2: Item2

```
#Histogram of Item2
hist(df$Item2,
    main="Patient Ranking of
Timely Treatment",
    xlab="Ranking",
    border="black",
    col="#203824",
    ylab = "Frequencies",
    breaks=8)
```



Distribution: The graph of Item2 also demonstrates a left-skewed distribution.

D. Bivariate Statistics

Bivariate Statistics of continuous variables Age and Initial_days:

```
df %>%
  regress(Age, Initial_days)
```

Variable	B	StdErr	beta	t	p
* <chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 (Intercept)	53.1	0.340	NA	156.	0
2 Initial_days	0.0128	0.00784	0.0163	1.63	0.104

 # F(1, 9998) = 2.645442, p = 0.103879, R-square = 0.000265

Bivariate statistics of categorical variables Item1 and Item2:

```
df %>%
  regress(Item1, Item2)
```

Variable	B	StdErr	beta	t	p
* <chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 (Intercept)	1.23	0.0275	NA	44.5	0
2 Item2	0.654	0.00753	0.656	86.8	0

 # F(1, 9998) = 7535.660208, p = 0.000000, R-square = 0.429782

D1. **Continuous Variables**: Age vs. Initial_days.

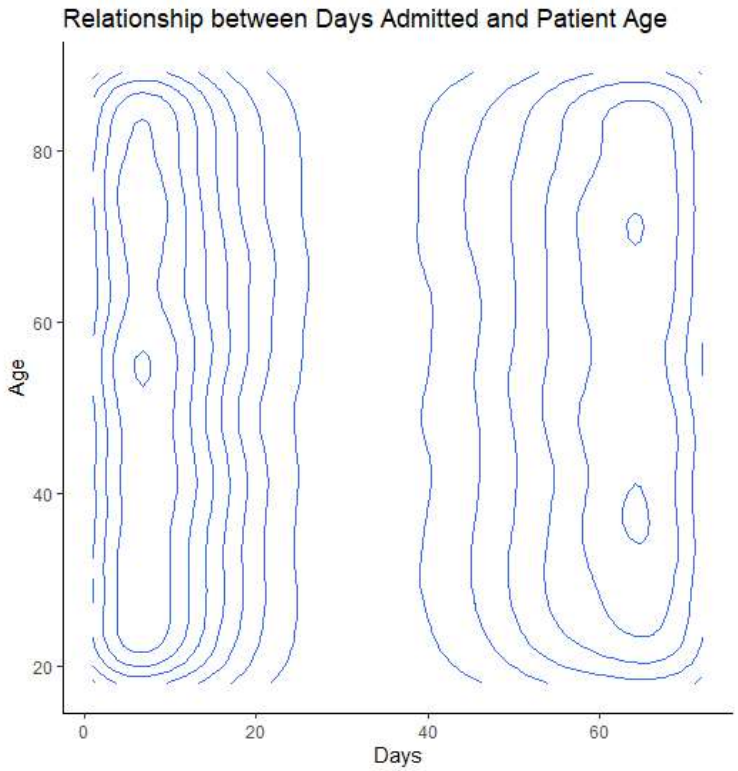
Categorical Variables: Item1(Patient ranking of the importance of timely admission) vs Item2 (Patient ranking of the importance of timely treatment). Item1 and Item2 are ordinal categorical variables.

Density plot of Continuous Variables: Age and Initial_days.

Plot code:

```
my.theme <- theme_classic()  +
theme(aspect.ratio = 1)

ggplot(df, mapping =
aes(x=Initial_days, y=Age))+
  geom_density2d() +
  labs(x = "Days", y = "Age", title =
"Relationship between Days Admitted
and Patient Age") +
  my.theme
```



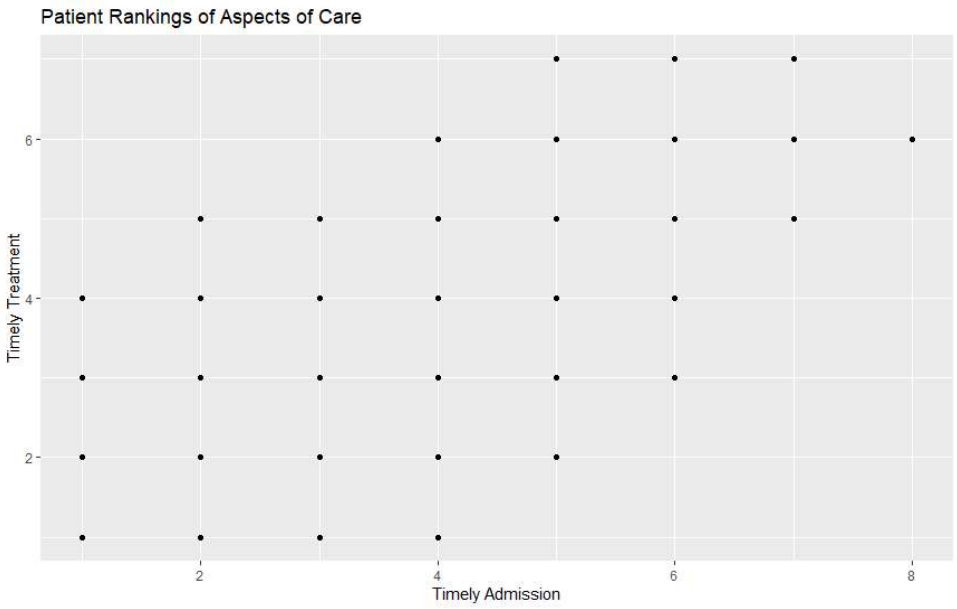
Distribution: Demonstrates no correlation. This is supported by the low r-squared value of 0.000265 from the table above.

Scatter plot of Categorical Variable: Item1 (Patient ranking of timely admission) and Item2 (Patient ranking of timely treatment).

Plot code:

```
Item.labels <- labs(x =
"Timely Admission", y =
"Timely Treatment",
                    title = "Patient
Rankings of Aspects of
Care")

ggplot(df)  +
  geom_point(aes(x = Item1,
y = Item2)) +
  Item.labels
```



Distribution: This demonstrates a mild positive correlation. This is supported by the r-squared value of 0.429782 from the table above.

E. Summary

E1. The hypothesis that was tested is: there is a statistical difference in mean ages of patients who were readmitted and those who were not readmitted. The null hypothesis is that there is no statistical difference between the mean ages of these two patient populations. T-testing was performed to test these hypothesis. The t-testing suggests that we should reject the hypothesis and accept the null hypothesis. For a two-sample t-test a t value greater than 1.960 would be required at alpha 0.05. Our T value is 1.5811 which is less than the 1.960 cutoff. In addition, the p-value from our t-test 0.1139 is greater than alpha (0.05) so we can reject our hypothesis. There is no statistically significant difference between the mean age of patients who were readmitted and those who were not readmitted.

E2. There are several limitations to the t-test performed. To start with the t-test assumes the data is normally distributed. As shown in the density plots of age compared to readmission status, the data is not perfectly normally distributed so this may impact the accuracy of this test. The t-test is also typically used for sample sizes less than 100. In this case, the sample size is much larger so a z-test is necessary to validate the t-test findings. The z-test produced an identical p-value so, the t-test is valid. Other limitations to our testing include a lack of known population means. It would benefit our testing if we knew the mean age of hospitalized patients at this hospital. This way we could validate that the mean age of this sample is representative of the whole population. The whole sample may not be representative of the whole population, for instance, our sample may be younger or older than the mean of the population and this may be the reason there was no difference in the mean age of the two groups.

E3. The recommended course of action based on these results would be to continue to look for factors that may correlate with readmission. In this case, we could find no correlation between age and readmission but there may be other factors that do correlate with readmission and this should be

investigated. A chi-squared test would be a reasonable next step to evaluate multiple factors against patient readmission.

F. Panopto video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6f970944-35ad-4978-a285-b1680146141e

G. Code Sources

"Bivariate Analysis of Continuous And/or Categorical Variables." *The Comprehensive R Archive Network*, 22 Feb. 2024, cran.r-project.org/web/packages/tidycomm/vignettes/v02_bivariate.html.

Kabacoff, Robert. "Quick-R: Descriptive Statistics." w*ww.statmethods.net*, 2017, www.statmethods.net/stats/descriptives.html.

Paul, Magwene. "Creating Bivariate Plots with Ggplot2." *Bio304-Class.github.io*, bio304-class.github.io/bio304-fall2017/ggplot-bivariate.html. Accessed 6 May 2024.

Sakshi. "Two Sample Z-Test in R with Examples." *Statistics Tutorial*, 24 Dec. 2021, statstutorial.com/two-sample-z-test-in-r-with-examples/.

Schoonjans, Frank. "T-Distribution Table (Two-Tailed)." *MedCalc*, www.medcalc.org/manual/t-distribution-table.php.

H. Sources

Sewell, Dr. William. "D207 Exploratory Data Analysis Webinar Episodes 1-6." *WGU.edu*, wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=9c39f90e-4457-4613-b4cf-b109004601ed. Accessed 4 May 2024.

I. Professionalism

This paper is intended to be profession in manner and appearance.