1. Assignment 4

Note: This assignment was done using RStudio with R language.
Please find here for the code https://github.com/abin733/Assignment-4

2. Section A

Task 1

Table 1 Cohort Characteristics

	Total	*Female*	*Male*
n	470	111 (23.62%)	359 (76.38%)
Age, years	Mean: 35.75 (SD: 7.79)	Mean: 36.07 (SD: 7.71)	Mean: 35.65 (SD: 7.84)
	Median: 35 (IQR: 30-41)	Median: 35 (IQR: 31-40.5)	Median: 35 (IQR: 30-41)
Race group	Black: 218 (46.38%)	Black: 56 (50.45%)	Black: 162 (45.13%)
	Hispanic: 51 (10.85%)	Hispanic: 10 (9.01%)	Hispanic: 41 (11.42%)
	Other: 28 (5.96%)	Other: 8 (7.21%)	Other: 20 (5.57%)
	White: 173 (36.81%)	White: 37 (33.33%)	White: 136 (37.88%)
Homeless	Homeless: 219 (46.60%)	Homeless: 42 (37.84%)	Homeless: 177 (49.30%)
Number of hospitalizations for medical problems in their lifetime	Mean: 3.10 (SD: 2.47)	Mean: 3.50 (SD: 2.04)	Mean: 2.97 (SD: 2.58)
	Median: 2 (IQR: 1-3)	Median: 2 (IQR: 1-4)	Median: 2 (IQR: 1-3)
Number of times entered a detox programme in the past 6 months	Mean: 2.47 (SD: 2.49)	Mean: 2.04 (SD: 1.59)	Mean: 2.66 (SD: 2.66)
	Median: 2 (IQR: 1-3)	Median: 2 (IQR: 1-2)	Median: 2 (IQR: 1-3)
Average number of drinks consumed per day in the past 30 days	Mean: 18.26 (SD: 20.13)	Mean: 15.00 (SD: 18.47)	Mean: 19.27 (SD: 20.54)
	Median: 13 (IQR: 5-26)	Median: 8 (IQR: 1-24)	Median: 13 (IQR: 4-26)
Primary substance of abuse	Alcohol: 185 (39.36%)	Alcohol: 37 (33.33%)	Alcohol: 148 (41.23%)
	Cocaine: 156 (33.19%)	Cocaine: 41 (36.94%)	Cocaine: 115 (32.03%)
	Heroin: 128 (27.23%)	Heroin: 32 (28.83%)	Heroin: 96 (26.74%)
Any substance abuse post-detox	196 (41.70%)	40 (36.04%)	156 (43.45%)
Time to first substance abuse post-detox, days	Mean: 75.13 (SD: 79.52)	Mean: 83.77 (SD: 85.57)	Mean: 72.61 (SD: 77.71)
	Median: 33 (IQR: 5-164.25)	Median: 33 (IQR: 7-179)	Median: 33 (IQR: 4-154)

Note: The cell highlighted in yellow (Female and Alcohol) represents the percentage of females in the cohort who have alcohol as their primary substance of abuse, which is 33.33%.

Task 2

a) Homelessness by Sex:

- In the total section, the data shows that approximately 46.59% of the total sample reported being homeless at some point.
- When broken down by sex, the data reveals that among females, about 37.84% reported being homeless, while among males, the percentage is higher at approximately 49.30%.
- This suggests that a higher proportion of males in the sample experienced homelessness compared to females.

b) Primary Substance of Abuse by Sex:

- When examining the primary substance of abuse, we can see that the total percentage of individuals who reported alcohol as their primary substance of abuse is approximately 39.36%.
- In the female section, about 33.33% of females reported alcohol as their primary substance of abuse.
- Among males, the percentage reporting alcohol as their primary substance of abuse is higher, at around 41.23%.

In summary, regarding homelessness, a higher percentage of males in the sample reported experiencing homelessness compared to females. In terms of primary substance of abuse, a slightly higher percentage of males identified alcohol as their primary substance of abuse compared to females. This indicates that a slightly higher proportion of males in the sample identified alcohol as their primary substance of abuse compared to females.

Task 3

Table 2 Cohort Characteristics: Substance Abuse Post-Detox

Cohort Characteristics: Substance Abuse Post-Detox	*Total*	*Female*	*Male*
n	196	40 (36.04%)	156 (43.45%)
Age, years	Mean: 35.75 (SD: 7.79)	Mean: 36.07 (SD: 7.71)	Mean: 35.65 (SD: 7.84)
	Median: 35 (IQR: 30-41)	Median: 35 (IQR: 31-40.5)	Median: 35 (IQR: 30-41)
Primary Substance of Abuse	Alcohol: 185 (39.36%)	Alcohol: 37 (33.33%)	Alcohol: 148 (41.23%)
	Cocaine: 156 (33.19%)	Cocaine: 41 (36.94%)	Cocaine: 115 (32.03%)
	Heroin: 128 (27.23%)	Heroin: 32 (28.83%)	Heroin: 96 (26.74%)
Time to First Substance Abuse Post-Detox, Days	Mean: 75.13 (SD: 79.52)	Mean: 83.77 (SD: 85.57)	Mean: 72.61 (SD: 77.71)
	Median: 33 (IQR: 5-164.25)	Median: 33 (IQR: 7-179)	Median: 33 (IQR: 4-154)

The cell highlighted in yellow (Male and Cocaine) represents the percentage of males in the cohort who have cocaine as their primary substance of abuse, which is 32.03%.

Task 4

A. Plot the distributions of age by sex. Is the difference in mean age statistically significant? Explain how you came to your answer.

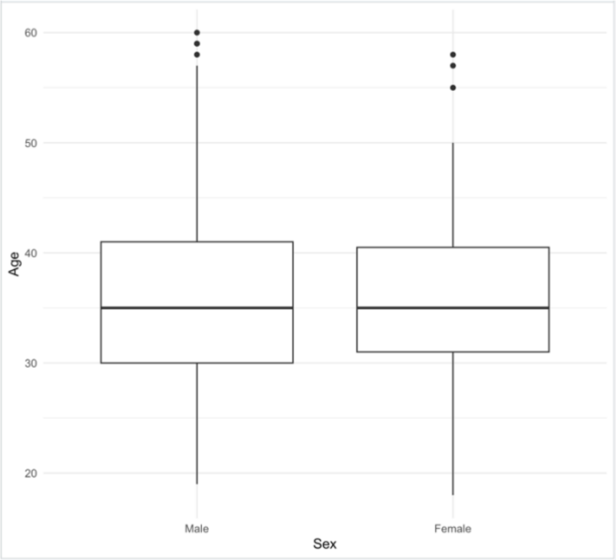


Figure 1 Boxplot for the distributions of age by sex.

The analysis aimed to determine whether there is a statistically significant difference in the mean age between females and males among a total sample of 470 individuals. The results of a two-sample t-test with a significance level (alpha) of 0.05 indicated that the p-value (p > alpha) was greater than the chosen alpha level. Therefore, the null hypothesis (H0), which posited that the mean age of females is equal to the mean age of males, was not rejected. In other words, there was insufficient statistical evidence to conclude that the mean age of females differs from that of males in this sample. The sample data showed that the mean age of females was 36.07 years (SD: 7.71), and for males, it was 35.65 years (SD: 7.84), suggesting that the age difference between the two groups was not statistically significant.

B. Plot a histogram of 'lifetime number of hospitalisations for medical problems' and describe the distribution.

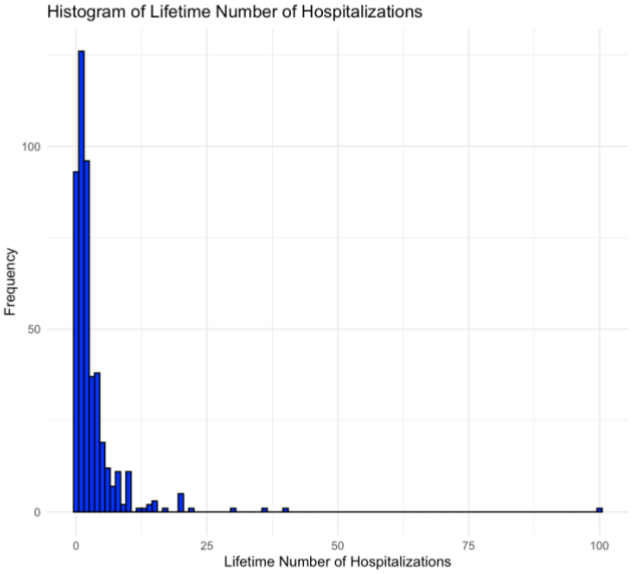


Figure 2 Histogram

In this dataset of 470 observations regarding the "Lifetime Number of Hospitalizations for Medical Problems," several key statistics provide valuable insights. The mean number of hospitalizations is 3.1, suggesting that, on average, individuals in the dataset have experienced around 3 hospitalizations in their lifetime. The relatively high standard deviation of 6.21 indicates considerable variability, signifying that hospitalization counts vary widely among individuals. The median value of 2, which is less than the mean, implies a right-skewed distribution. The range spans from 0 to 100, demonstrating the broad spread of values. Positive skewness (skew = 9.41) indicates a right-skewed distribution with potential outliers on the higher end. Additionally, the high positive kurtosis (kurtosis = 128.98) suggests the presence of heavy tails or extreme values. Overall, the data reflects a positively skewed distribution with substantial variability and potential outliers on the higher end of hospitalization counts.

C. Plot a figure containing boxplots that show the maximum number of drinks consumed per day stratified by primary substance of abuse.

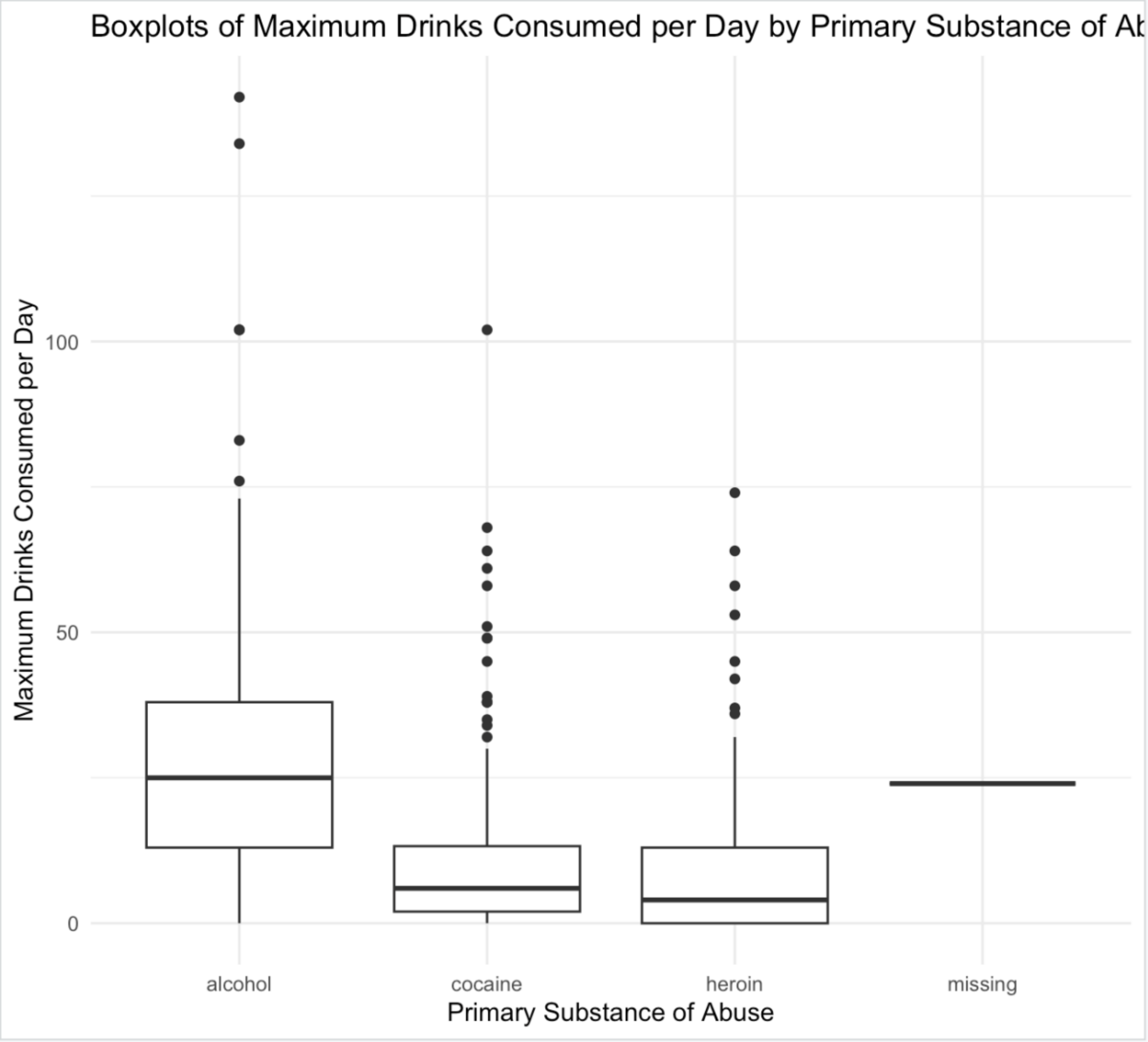


Figure 3 Boxplots of Maximum Drinks Consumed per Day by Primary Substance of Abuse

1.

I. Summarise the distribution of people whose primary substance of abuse is alcohol.

The "boxplot_summary" for individuals with alcohol as their primary substance of abuse reveals key insights: the median consumption of alcohol per day is approximately 13 drinks, and the interquartile range spans from 3 to 26 drinks. This suggests considerable variability in alcohol consumption within this group. Furthermore, the presence of several outliers indicates that some individuals in this category consume significantly more alcohol daily than the majority, implying a positively skewed distribution with extreme consumption levels among a subset of individuals.

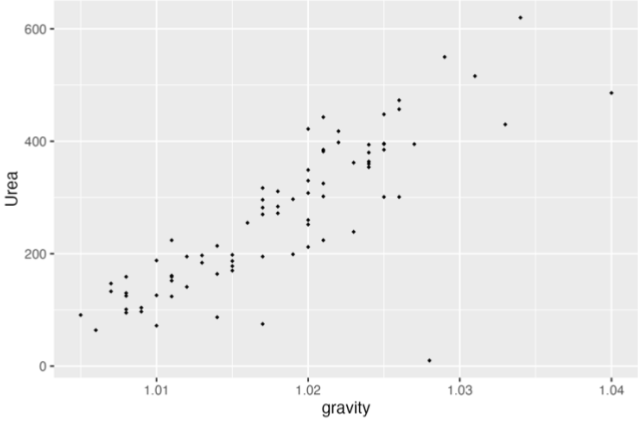
II. What was the primary substance of abuse for the person who drank the highest number of drinks in the last six months, and how many drinks was it?

Alcohol was the primary substance. It is 142, as this outlier represents an exceptionally high number of drinks consumed.

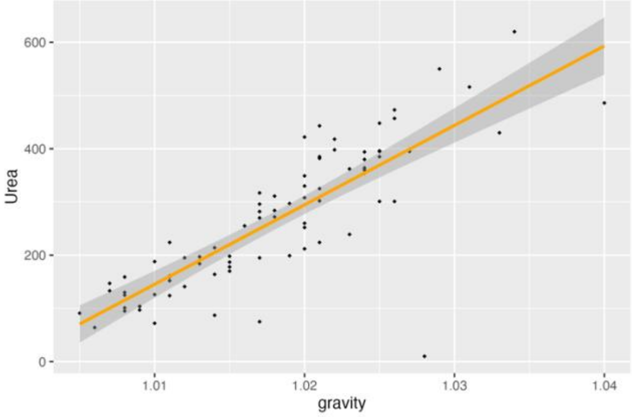
3. Section B

Task 5

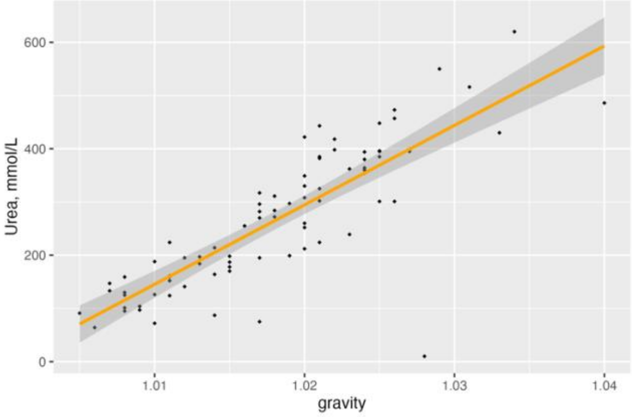
- Change the symbol being plotted to diamonds and enlarge it to a point size of 1.



- Add a linear trend line coloured orange.



- Change the y-axis title to 'Urea, mmol/L'.



Task 6

With a correlation coefficient of 0.823477, there is a strong positive linear relationship between these two variables. This means that as the "urea" concentration in urine increases, the "specific gravity" of urine tends to increase as well. The closer the correlation coefficient is to 1, the stronger the positive linear relationship between the variables. In this case, a value of 0.823477 indicates a robust positive association between "urea" and "specific gravity."

Task 7 (4 marks): Filter the dataset to only include people with a urea of <100 mmol/L.

A. How many people are left in the data?

Number of people with urea < 100 mmol/L: 8

B. What is the median pH in this subset of people?

Median pH in this subset: 6.745

4. Section C

Task 8

Median for Group A (<= medianAllBili): 0.8

Median for Group B (> medianAllBili): 3.5

Task 9

Table 3 Table for Treatment Category

D-penicillamine	*Placebo*
158	154

Table 4 Table for Outcome Category

Censored	*Transplant*	*Dead*
232	25	161

A. How many people died and what percentage of those who died were in the placebo group?

Total who died is 161 percentage of those who died in the placebo group: 37.26708 %

B. What percentage of the treatment group received a transplant?

Percentage of the treatment group (D-penicillamine) that received a transplant: 6.329114 %

Task 10

A. What is the risk ratio of having a transplant or dying in the treatment compared to the placebo group?

Risk Ratio of having a transplant or dying in the treatment group compared to the placebo group: 0.92

B. How would you report this in words?

The risk of having a transplant or dying in the treatment group (D-penicillamine) was 0.92 times lower compared to the placebo group. In other words, individuals in the treatment group had a 0.92-fold decreased risk of experiencing a transplant or death compared to those in the placebo group.

C. What is the odds ratio of having a transplant or dying in the treatment compared to the placebo group?

Odds Ratio of having a transplant or dying in the treatment group compared to the placebo group: 0.8983529

D. How would you report this in words?

After merging the outcome statuses of transplantation and death into a single group using the 'outcome.binary' variable, where 0 represents individuals who were censored (i.e., neither had a transplant nor died), and 1 represents those who experienced either transplantation or death, we calculated the odds ratio. The odds ratio for having a transplant or dying in the treatment group (D-penicillamine) compared to the placebo group was found to be approximately 0.898. This implies that individuals in the treatment group had approximately 0.898 times the odds of experiencing a transplant or death compared to those in the placebo group. In other words, the treatment group had a slightly lower odds of this outcome when compared to the placebo group.

Task 11

A. Which stage has the most people in it?

The stage with the most people is: 3

B. What is the mean (SD) of age for each of the 4 stages?

Table 5 Stages Summary

stage_summary	*Mean_Age*	*SD_Age*
1	46.841107	9.545687
2	49.465941	9.624315
3	48.962540	10.118584
4	53.765572	10.823320

C. To further assess mean age and stage, select Inference and Hypothesis testing using ANOVA. Interpret the result of the ANOVA F-test that is shown in the output.

Table 6 ANOVA F-test Results on RStudio

Source	*Df*	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
stage_category	3	2273	757.6	7.227	9.8e-05
Residuals	408	42771	104.8		

The ANOVA F-test results show a statistically significant difference in mean age across different stages of the disease. The F-statistic is 7.227 with a p-value of 9.8e-05, indicating that there is strong evidence to reject the null hypothesis that the mean ages of individuals in all disease stages are equal. In other words, at least one stage differs significantly in terms of mean age from the others. This suggests that disease stage plays a significant role in influencing the mean age of individuals in the study. Further post-hoc tests should be conducted to identify which specific stage(s) differ from the others in terms of mean age. Additionally, it's worth noting that 6 observations were deleted due to missing data, which may have impacted the analysis.

D. ~~One of the results in the output in iNZight shows the difference in mean age of people in stage 3 with those in stage 4~~. Report the findings of that analysis.
If you are **not using iNZight software**, assess the paired differences in means, paste the output you get, and report the findings of the difference in mean age of people in stage 3 with those in stage 4.

Tukey multiple comparisons is a statistical method used to compare the difference in mean age between two groups, specifically in this case, people in stage 3 and stage 4. This method is chosen because it allows for the comparison of multiple groups while controlling for the family-wise error rate (FWER) (Bretz et al., 2016). The FWER is the probability of making at least one Type I error (rejecting a true null hypothesis) when conducting multiple hypothesis tests.

Table 7 Table for Tukey multiple comparisons of means.

Comparison	*Difference*	*Lower Bound*	*Upper Bound*	*p-value*
2-1	2.6248337	-3.7627804	9.012448	0.7140310
3-1	2.1214331	-4.0201996	8.263066	0.8095151
4-1	6.9244647	0.7549062	13.094023	0.0207410
3-2	-0.5034005	-3.9794962	2.972695	0.9821972
4-2	4.2996310	0.7744303	7.824832	0.0095698
4-3	4.8030315	1.7460592	7.860004	0.0003516

Note: Tukey multiple comparisons of means 95% family-wise confidence level Fit: aov(formula = age ~ stage_category, data = data)

The comparison between Stage 3 and Stage 4 in terms of mean age reveals a statistically significant difference. People in Stage 4 have a significantly higher mean age than those in Stage 3, with a mean age difference of approximately 4.803 years. The 95% confidence interval for this difference ranges from approximately 1.746 years to 7.860 years. This difference is supported by a very low p-value of 0.0003516, which is well below the typical significance level of 0.05. Therefore, there is strong evidence to suggest that the mean age of individuals in Stage 4 is significantly higher than that of individuals in Stage 3.

5. References

1) Bretz, F., Hothorn, T., & Westfall, P. (2016). Multiple comparisons using r.. https://doi.org/10.1201/9781420010909