Project I

Question 1:

1. **What is your method to solve this question?**

One-way ANOVA

One-way ANOVA:  is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups (although **you** tend to only see it used when there are a minimum of three, rather than two groups).  
**Reason**: Because we compute statistic test between each factory -> blood lead levels among 5 factories are the same or not. Error of variances follow N(0, sigma^2).

Choose **1-way** instead of **2-way**: because the factor level is the only factor in this table.

Choose **1-way** instead of **Chi-square**: this table does not represent frequency and we do not test the independence between columns and rows.

1. **header = T**: first row in csv is the header.  
   **colClasses**: classify columns into types: factor and numeric.

**attach**: get the data frame, extract value, avoid $ sign.

1. **Conclusion from boxplot**: cannot conclude anything from data, just can identify the outliers to take care (can remove outlier or can keep it), in this case we keep the outliers because it is few data.

**Conclusion from boxplot in general**: 5 factors: min, max, median (second quartile), first quartile, third quartile. Determine max extreme by from range (Q3, Q3 + 1.5IQR), the max extreme is the max value in that range. Min extreme range (Q1 – 1.5IQR, Q1).

**Talking about boxplot not consider about the MEAN**

1. **Explain about ANOVA1 table pg6**

* Factory is treatment, Residuals is error
* Df is degree of freedom: 24 = 28 – 4 (error degree of freedom)
* 0.00729: SStr
* 0.02673: SSE
* Mean Sq follows formula of MSTr & MSE
* F value = MSTr / MSE

1. **Multiple Comparison**  
   **Reason to use**: in fact, fail to reject H\_0 but want to check if blood lead levels among 5 factories are really the same -> LSD (tính theo từng cặp)  
   **y=ANOVA1**: use model ANOVA1  
   **trt=’Factory’**: treatment is Factory  
   **p.adj=’none’**: chỉ ra cách nào so sánh giữa 2 factory, none means default -> default means using t test  
   **console=TRUE**: to print result to the console
2. **Explain the LSD table in pg7**  
   **FactorLevel**: mean (TBC của từng column)  
   **std** (sample standard deviation): tính theo formula sample standard deviation = sqrt(sample variance) (sample variance đã học)  
   **r**: number of observations  
   **LCL** (Lower Control Limit): tính bằng formula mean – t(alpha/2, N-a)\*sqrt(MSE/r) (N-a là df error)  
   **UCL** (Upper Control Limit): tính bằng formula mean + t(alpha/2, N-a)\*sqrt(MSE/r)   
   **Critical value of t**: t(alpha/2, n-1)

Question 2:

1. **What method?**  
   Chi-square test

**Reason**: table is frequency

1. **Why not use two sample t-test over chi-square?**

* Each columns will have same proportions
* **t-test**: (Zinc Concentration)

+ cùng 1 column giống nhau các dk, chỉ khác ở 1 dk  
+ cùng 1 row thì cùng 1 dk

* **Chi-square**:  
  + câu 2 đáng ra chỉ ảnh hưởng bởi age group nhg ở đây mỗi cell trg cùng 1 hàng chịu ảnh hưởng bởi 2 yếu tố.

1. **Using chi-square**:

* Assume ban đầu là row & column independent
* keyword: H\_0 là row & column are independent

1. **Objective of Chi-square**: expected và observed có match với nhau tốt không. Nếu p-value > alpha thì observed gần expected -> fail to reject H\_0.

Question 3:

1. **Schools** **is confounding variable** because đề kêu tính days thui.
2. **days = factor(rep(c(), each = 4))**: tạo ra vector với trình tự Mon,Mon,Mon,Mon,Tue,Tue,…  
   **schools = factor(rep(c(), 4))**: tạo ra vector A,B,C,D,A,B,C,D,…

Project II

1. **What method?**  
   Linear Regression Model
2. **How do you know linear?**  
   Do not know.Assume each variable such as G1, G2, studytime, … have linear relationship with G3. Then build a linear model by using lm() -> If Adjusted R-squared is close to 1, assumption linear is correct.
3. **What is your solution to NA status ?**  
   Eliminate row having NA status using na.omit()
4. **Why use frequency table for sex, studytime, failures, higher?**

Because these variables are categorical variables.

1. **lm() ?**lm(): linear model. G3 ~ . means G3 depends all other variables except G3.
2. **Null hypothesis for each p-value in table pg33?**H\_0: hệ số B = 0, which means that if reject H\_0 then G3 depends on that variable (y = Bx)
3. **Adjusted R-squared ?**If that number is near to 1, assumption on all regressors inside linear model is linearly related to predictor is correct.
4. **Anova 3 models ?**  
   H\_0: 2 models are the same  
   3 p-values > 0.05 -> fail to reject H\_0 -> 3 models are the same -> prefer the simpler one, M3
5. **Đường chéo ở figure 18 ?**  
   Theo histogram của G3, đi qua tọa độ (0,0) và chiều đi xuống, mà ta có cthuc error = G3 – G3^ => error = -G3^ or residual = -fitted. Đoán dc vậy do đặc trưng dataset có nhiều hs G3 = 0.  
   **Reason**: theo research, the data set it outlying because dropout!

Tất cả giá trị đều mang fitted < 10 -> model đoán dựa trên thông tin history thì hs fail, dựa theo thông tin như vậy thì hs bik mình rớt nên không muốn học nữa.

1. **Explain table in pg43**Have difference because G3 >= 10 and 10 is the threshold to pass, lots of students approximately 10 and our predicted model can have some biases, errors -> lead to a failed student 9.9 can predicted to pass!