HW 20 Solution

The table below gives the energy use of five gas ranges for seven menu days. (The units are equivalent kilowatt-hours; $0.238\text{kWh}=1\text{ft}^3$ of natural gas.) The energy use calculate is an average over several cooks. Your goal is to describe the variation in the table below in terms of the effects of different ranges and different menu days.

```
mydata <- read.csv("https://raw.githubusercontent.com/ProfSuzy/Stat61/main/gasranges.csv",
                   header=T)
mydata
```

```
##   Day Range1 Range2 Range3 Range4 Range5
## 1   1   8.25   8.26   6.55   8.21   6.69
## 2   2   5.12   4.81   3.87   4.81   3.99
## 3   3   5.32   4.37   3.76   4.67   4.37
## 4   4   8.00   6.50   5.38   6.51   5.60
## 5   5   6.97   6.26   5.03   6.40   5.60
## 6   6   7.65   5.84   5.23   6.24   5.73
## 7   7   7.86   7.31   5.87   6.64   6.03
```

```
##Put data table into proper form (each unit on one row)
kWh = as.vector(unlist(mydata[,2:6]))
menu_day=rep(1:7, 5)
range = c(rep("Range1", 7), rep("Range2", 7), rep("Range3", 7),
          rep("Range4", 7), rep("Range5", 7))
## Double check that the lengths of these vectors match
length(kWh); length(menu_day); length(range)
```

```
## [1] 35
```

```
## [1] 35
```

```
## [1] 35
```

```
## Create the data frame to analyze
unit_data_energy <- data.frame(kWh, menu_day, range)
unit_data_energy
```

```
##     kWh menu_day  range
## 1  8.25        1 Range1
## 2  5.12        2 Range1
## 3  5.32        3 Range1
## 4  8.00        4 Range1
## 5  6.97        5 Range1
## 6  7.65        6 Range1
## 7  7.86        7 Range1
## 8  8.26        1 Range2
## 9  4.81        2 Range2
## 10 4.37        3 Range2
## 11 6.50        4 Range2
## 12 6.26        5 Range2
## 13 5.84        6 Range2
```

```
## 14 7.31        7 Range2
## 15 6.55        1 Range3
## 16 3.87        2 Range3
## 17 3.76        3 Range3
## 18 5.38        4 Range3
## 19 5.03        5 Range3
## 20 5.23        6 Range3
## 21 5.87        7 Range3
## 22 8.21        1 Range4
## 23 4.81        2 Range4
## 24 4.67        3 Range4
## 25 6.51        4 Range4
## 26 6.40        5 Range4
## 27 6.24        6 Range4
## 28 6.64        7 Range4
## 29 6.69        1 Range5
## 30 3.99        2 Range5
## 31 4.37        3 Range5
## 32 5.60        4 Range5
## 33 5.60        5 Range5
## 34 5.73        6 Range5
## 35 6.03        7 Range5
```

Problem 1

(a) Find and interpret the grand mean.

```
grand_mean = mean(unit_data_energy$kWh)
```

(b) Find and interpret the differential effect of a menu day, for each of the seven days.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.2      v dplyr   1.0.7
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   2.0.1      v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
## Warning: package 'readr' was built under R version 4.1.1
```

```
## Warning: package 'forcats' was built under R version 4.1.1
```

```
## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Effects of menu day
for(i in 1:7){
  res = (unit_data_energy[(unit_data_energy$menu_day==i),1] %>% mean - grand_mean)
  print(paste0("Day ", i, ": ", round(res,3)))
}
```

```
## [1] "Day 1: 1.601"
## [1] "Day 2: -1.471"
## [1] "Day 3: -1.493"
```

```
## [1] "Day 4: 0.407"
## [1] "Day 5: 0.061"
## [1] "Day 6: 0.147"
## [1] "Day 7: 0.751"
```

(c) Find and interpret the differential effect of the range, for each of the five ranges.

```
## Effects of range
for(i in 1:5){
  res = (unit_data_energy[(unit_data_energy$range==paste0("Range",i)),1] %>% mean - grand_mean)
  print(paste0("Range ", i, ": ", round(res,3)))
}
```

```
## [1] "Range 1: 1.033"
## [1] "Range 2: 0.201"
## [1] "Range 3: -0.893"
## [1] "Range 4: 0.22"
## [1] "Range 5: -0.561"
```

Problem 2

Report the estimated additive model for energy use based on this data. Complete the following table of the differences of the observed energy use and the predicted energy use from your model above. Do these residuals indicate a possible interaction effect between the ranges and the menu days?

Menu day	Range 1	Range 2	Range 3	Range 4	Range 5
1					
2					
3					
4					
5					
6					
7					

```r
resids = matrix(rep(NA, 7*5), ncol=5)
for(j in 1:5){
   beta_j = unit_data_energy[(unit_data_energy$range==paste0("Range",j)),1] %>% mean - grand_mean
  for(i in 1:7){
    obs_val = mydata[i,j]
    alpha_i =  unit_data_energy[(unit_data_energy$menu_day==i),1] %>% mean - grand_mean
    fit_val = grand_mean + alpha_i + beta_j
    resids[i,j] = obs_val - fit_val
  }
}
round(resids, 3)
```

```
##         [,1]  [,2]  [,3]   [,4]  [,5]
## [1,] -7.625 0.457 1.561 -1.262 1.179
## [2,] -3.553 0.399 1.183 -0.870 0.851
## [3,] -2.531 0.621 0.765 -0.958 0.733
## [4,] -3.431 1.401 0.995 -1.238 0.673
## [5,] -2.085 0.717 1.101 -1.242 0.909
## [6,] -1.171 1.311 0.595 -1.128 0.663
## [7,] -0.775 0.917 1.461 -1.092 0.459
```

Interpretation:

If we consider that the units are KWh, then we might reasonably think a difference ± 1 kWh is not practically meaningful. From this perspective, there does appear to be numerical evidence of an interaction effect especially for Range 1 but possibly for other ranges over the different menu days as well.

Problem 3

Complete the following 2-way ANOVA table for this data "by hand" (just show your work). Then calculate the F-test statistic and corresponding p-value for a test of the significance of an interaction effect between the ranges and the menu days.

Source	df	SS	MS	F
Menu day				
Range				
Interaction				
Error				
Total				

**Note: It is not possible to answer the question as stated since we don't have the entire data set but rather, we have means over several "cooks" for each range and menu day. Instead, we can fit a 2-way ANOVA table without an interaction effect for $K = 1$ with the following table:

```
mod = lm(kWh ~ range + menu_day, data = unit_data_energy)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: kWh
##           Df Sum Sq Mean Sq F value  Pr(>F)
## range      4 15.877  3.9693  2.8633 0.04092 *
## menu_day   1  0.896  0.8960  0.6463 0.42797
## Residuals 29 40.202  1.3863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

More generally, the ANOVA table for a 2-way model without interaction effects and with one observation per combination of factor level is:

Source	df	SS	MS	F
FactorA	$(I-1)$	$SSA = J\sum_j(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$MSA = \frac{SSA}{I-1}$	$F = \frac{MSA}{MSE}$
FactorB	$(J-1)$	$SSB = I\sum_i(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$	$MSB = \frac{SSB}{J-1}$	$F = \frac{MSB}{MSE}$
Error	$(I-1)(J-1)$	$SSE = \sum_i\sum_j(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$	$MSE = \frac{SSE}{(I-1)(J-1)}$	
Total	$(IJ-1)$	$SS_{Tot} = \sum_i\sum_j(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$		

Here, since $K = 1$ the two-way ANOVA model becomes

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}.$$