EX3

roi hezkiyahu

10 3 2022

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Q.1

שאלה 1

הוכיחו כי תיקון בונפרוני שולט באופן חזק על ה FWER (family wise error rate) (מתוך מצגת 3 שקף 5).

lets take a look at the following expression:

$$Pr(\cup_{m=1}^{M} A_m) \leq \sum_{m=1}^{M} Pr(A_m)$$

$$the\, RHS = LHS\ iff\, A_i, A_j\, are\, independet\ \forall i, j$$

thus we only need to prove for the case where A_i, A_j are independet

in this case:

$$FWER = Pr_{H_0}(R > 0) = 1 - \Pi_{i=1}^{M} Pr_{H_0}(R_i = 0) = 1 - (1 - \alpha^*)^m \Rightarrow 1 - FWER = (1 - \alpha^*)^m \Rightarrow$$

$$(1 - FWER)^{\frac{1}{m}} = 1 - \alpha^* \Rightarrow \alpha^* = 1 - (1 - FWER)^{\frac{1}{m}} > 1 - (1 - FWER) = FWER$$

Q.2

שאלה 2

בשאלה זו נשתמש בנתונים בשם Ricci מחבילת R בשם Stat2Data (ראו בתרגיל בית 1 איך לטעון אותם).

א. נרצה לדעת האם תוחלת הציון המשוקלל (Combined) שונה בין קבוצות התפקיד (Position) והגזע (Race) השונות. בדקו זאת בעזרת ניתוח שונות דו כיווני, עם אינטראקציה (רק אם יש עדות בנתונים שיש בה צורך).

I. פרטו אילו הנחות הנחתם. האם נראה כי הן מתקיימות?

II. האם הנתונים מאוזנים (balanced)? מה זה אומר על סכומי הריבועים השונים?

III. דווחו את הממוצעים של הקבוצות השונות לפי כל אחד מהגורמים וגם לפי השילוב שלהם.

IV. דווחו את התוצאות שקיבלתם. מה המסקנות עבור $\alpha = 0.05$?

ב. עבור כל אפקט מובהק שמצאתם בניתוח השונות, השתמשו בניתוח פוסט הוק על מנת לבדוק מאיפה הוא נובע. בחנו את כל ההשואות הזוגיות, בשתי השיטות הבאות. עבור כל שיטה, דווחו מה הadjusted pvalue המתקבל לכל השוואה ומה המסקנות המתקבלות עבור FWER=0.05

I. על ידי תיקון בונפרוני.

II. על ידי שיטת טוקי (או שיטת טוקי-קרמר, מה שמתאים כאן).

ג. אם מצאתם אפקט מובהק של גזע (Race) על הציון המשוקלל (Combined),

I. בידקו האם זה נובע מההבדל בין הקבוצה עבורה Race = H לבין הקבוצה עבורה Race = B.

II. השוו בין התוצאה שקיבלתם בסעיף ג, לבין התוצאות שקיבלתם בסעיף ב בנוגע להבדל בין הקבוצה עבורה Race = H לבין הקבוצה עבורה Race = B. האם יש הבדל? מדוע?

a

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------
----------------------- tidyverse 1.3.0 --
```

```
## <U+221A> ggplot2 3.3.2     <U+221A> purrr   0.3.4
## <U+221A> tibble  3.0.3     <U+221A> dplyr   1.0.2
## <U+221A> tidyr   1.1.2     <U+221A> stringr 1.4.0
## <U+221A> readr   1.3.1     <U+221A> forcats 0.5.0
```

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

```
library(Stat2Data)
library(glue)
```

```
##
## Attaching package: 'glue'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(lawstat)
#see if intercation is necceserty
data(Ricci)
full_model <- lm(Combine ~ Position*Race,data = Ricci)
no_inter_model <- lm(Combine ~ Position + Race,data = Ricci)
anov_test <- anova(no_inter_model,full_model)
pv <- anov_test[["Pr(>F)"]][2]
glue("interaction relevence p value is: {pv} therefore we will not reject the null, meaning the
 interaction is not needed")
```

```
## interaction relevence p value is: 0.644670365912923 therefore we will not reject the null, me
aning the interaction is not needed
```

```
#assumptions
glue("we assume equal variance")
```

```
## we assume equal variance
```

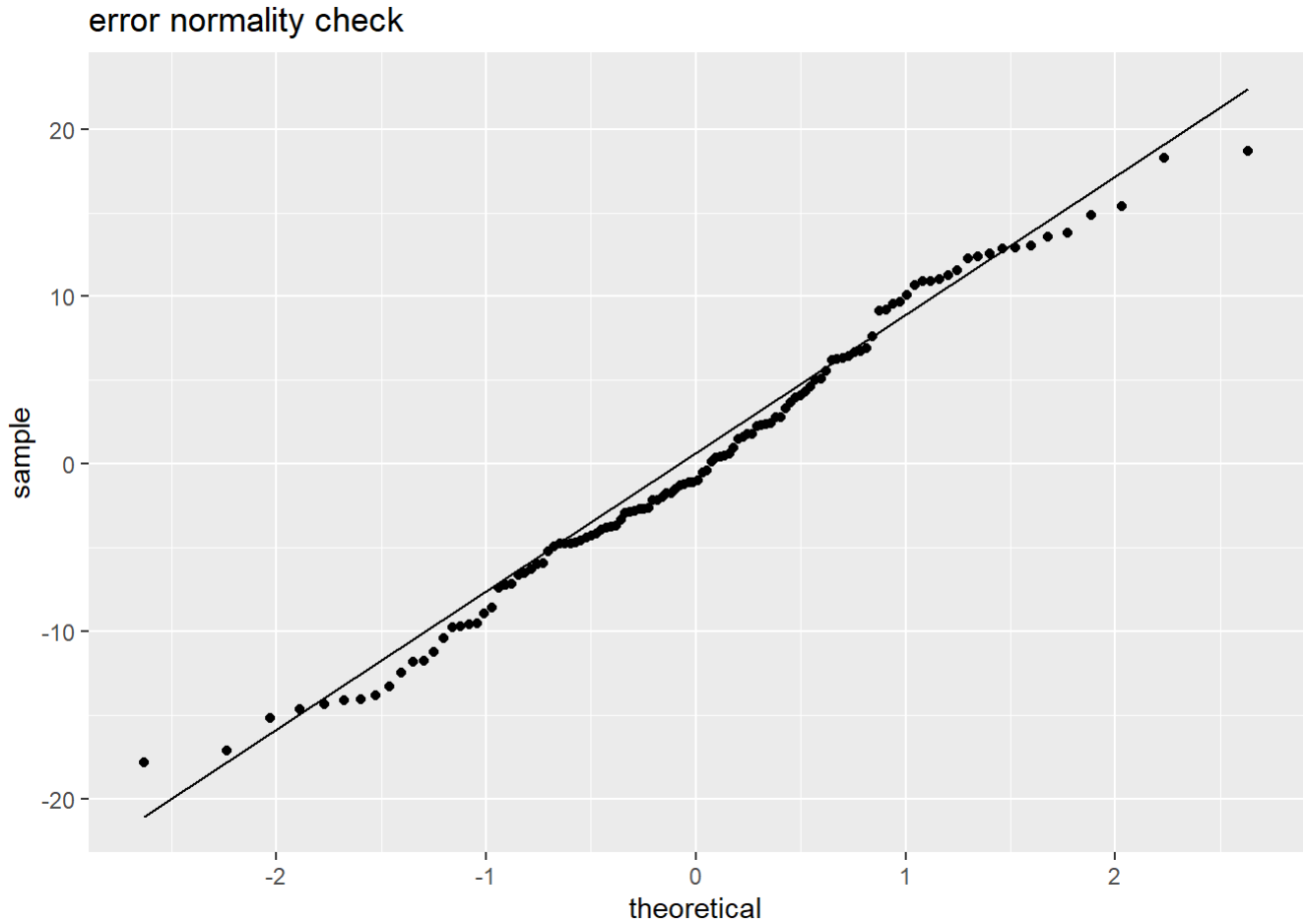
```
levene_test <-levene.test(Ricci$Combine,Ricci$Race:Ricci$Position,location = "mean")
glue("pvalue for levene test is {round(levene_test$p.value,4)} therefore we will not reject the
 null and conclude equal varicane")
```

```
## pvalue for levene test is 0.5363 therefore we will not reject the null and conclude equal var
icane
```

```
glue("we assume normality of the residuals")
```

```
## we assume normality of the residuals
```

```
ggplot(Ricci,aes(sample = Combine - predict(no_inter_model,Ricci[c("Position","Race")]))) +
  stat_qq()+
  stat_qq_line()+
  ggtitle("error normality check")
```




```
glue("looks ok")
```

```
## looks ok
```

```
glue("we also assume independece but we cant check it")
```

```
## we also assume independece but we cant check it
```

```
#balance check
Ricci %>%
  group_by(Position,Race) %>%
  summarize(n())
```

```
## `summarise()` regrouping output by 'Position' (override with `.groups` argument)
```

```
## # A tibble: 6 x 3
## # Groups:   Position [2]
##   Position   Race  `n()`
##   <fct>      <fct> <int>
## 1 Captain    B         8
## 2 Captain    H         8
## 3 Captain    W        25
## 4 Lieutenant B        19
## 5 Lieutenant H        15
## 6 Lieutenant W        43
```

```
glue("we can see that the groups are unbalanced")
```

```
## we can see that the groups are unbalanced
```

```
glue("it mneas that SST =/= SSA +SSB + SSAB + SSE")
```

```
## it mneas that SST =/= SSA +SSB + SSAB + SSE
```

```
#show means
Ricci %>%
  group_by(Position,Race) %>%
  summarize(mean(Combine))
```

```
## `summarise()` regrouping output by 'Position' (override with `.groups` argument)
```

```
## # A tibble: 6 x 3
## # Groups:   Position [2]
##   Position   Race  `mean(Combine)`
##   <fct>      <fct>           <dbl>
## 1 Captain    B                63.8
## 2 Captain    H                68.5
## 3 Captain    W                74.1
## 4 Lieutenant B                63.7
## 5 Lieutenant H                63.6
## 6 Lieutenant W                71.8
```

```
Ricci %>%
  group_by(Position) %>%
  summarize(mean(Combine))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   Position   `mean(Combine)`
##   <fct>                <dbl>
## 1 Captain               71.0
## 2 Lieutenant            68.2
```

```
Ricci %>%
  group_by(Race) %>%
  summarize(mean(Combine))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   Race  `mean(Combine)`
##   <fct>           <dbl>
## 1 B                63.7
## 2 H                65.3
## 3 W                72.7
```

```
#anova for Position
Pos_model <- lm(Combine ~ Race ,data = Ricci)
anov_test_pos <- anova(Pos_model,no_inter_model)
pv_pos <- anov_test_pos[["Pr(>F)"]][2]
glue("interaction relevence p value is: {pv_pos} therefore we will not reject the null, meaning
 that position has no effect on test results")
```

```
## interaction relevence p value is: 0.159918072118673 therefore we will not reject the null, me
aning that position has no effect on test results
```

```
#anova for Race
Race_model <- lm(Combine ~ Position ,data = Ricci)
anov_test_R <- anova(Race_model,no_inter_model)
pv_R <- anov_test_R[["Pr(>F)"]][2]
glue("interaction relevence p value is: {pv_R} therefore we will reject the null, meaning that r
ace has an effect on test results")
```

```
## interaction relevence p value is: 6.52696400342208e-06 therefore we will reject the null, mea
ning that race has an effect on test results
```

b

```
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
contr <- rbind(
  "Race H - Race B" = c(-1,1,0),
  "Race W - Race B" = c(-1,0,1),
  "Race W - Race H" = c(0,-1,1)
)
aov_model <-aov(Combine ~ Race,data = Ricci)
pair_wise <- glht(aov_model, linfct = mcp(Race = contr))
glue("bonferroni method results:")
```

```
## bonferroni method results:
```

```
summary(pair_wise, test = adjusted(type = "bonf"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = Combine ~ Race, data = Ricci)
##
## Linear Hypotheses:
##                      Estimate Std. Error t value Pr(>|t|)
## Race H - Race B == 0    1.600      2.416   0.662  1.00000
## Race W - Race B == 0    8.941      1.937   4.616 3.08e-05 ***
## Race W - Race H == 0    7.341      2.054   3.574  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- bonferroni method)
```

```
glue("Tukey method results:")
```

```
## Tukey method results:
```

```
summary(glht(aov_model, mcp(Race = "Tukey")), test = univariate())
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = Combine ~ Race, data = Ricci)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## H - B == 0    1.600      2.416   0.662 0.509167
## W - B == 0    8.941      1.937   4.616 1.03e-05 ***
## W - H == 0    7.341      2.054   3.574 0.000515 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Univariate p values reported)
```

```
glue("we can conculde from both methods that there is a difference between: whites and blacks, w
hites and hispanics for FWER = 0.05 ")
```

```
## we can conculde from both methods that there is a difference between: whites and blacks, whit
es and hispanics for FWER = 0.05
```

c

```
W_C <- Ricci %>%
  filter(Race == "W") %>%
  pull(Combine)
B_C <- Ricci %>%
  filter(Race == "B") %>%
  pull(Combine)
glue(" the difference is {mean(W_C) - mean(B_C)} which is the same as in the test, because it is
calculated the same way")
```

```
##  the difference is 8.94120261437908 which is the same as in the test, because it is calculate
d the same way
```

Q.3

שאלה 3

בשאלה זו נשתמש בנתונים בשם Ricci מחבילת R בשם Stat2Data (ראו בתרגיל בית 1 איך לטעון אותם).

בתרגיל בית 1 התמקדנו בתת המדגם של התצפיות עבורן מתקיים Position = Lieutenant ובדקנו האם יש עדות לכך שהנחת שוויון השונויות אינה מתקיימת עבור הציון המשוקלל (Combined) בין קבוצות הגזע (Race) השונות, על ידי מבחן לוין.

בשאלה זו נשתמש באותם הנתונים ונבצע ניתוח שונות חד כיווני ולאחר מכן גם ניתוחי פוסט הוק.

א. בצעו ניתוח שונות חד כיווני הבוחן האם בתת מדגם זה, הציון המשוקלל (Combined) שונה בין קבוצות הגזע (Race) השונות.

   I. פרטו אילו הנחות הנחתם והאם נראה שהן מתקיימות.

   II. דווחו את התוצאות שקיבלתם ומה המסקנות עבור $\alpha = 0.05$.

ב. אם דחיתם את השערת האפס בסעיף א', המשמעות היא שמצאתם עדות לכך שהציון המשוקלל (Combined) שונה בין קבוצות הגזע (Race) השונות. נרצה לבצע ניתוחי המשך על מנת לבחון מאיפה ההבדל הזה נובע.

   I. בחנו את שלושת ההשוואות הזוגיות בשתי השיטות הבאות. עבור כל שיטה, דווחו מה ה- adjusted pvalue המתקבל לכל השוואה ומה המסקנות המתקבלות עבור FWER=0.05:

      (1) על ידי תיקון בונפרוני.

      (2) על ידי שיטת טוקי (או שיטת טוקי-קרמר, מה שמתאים כאן).

ג. סכמו את המסקנות שקיבלתם. מה למדתם על הנתונים? האם ניתן לומר על סמך הנתונים הללו כי המבחן מפלה קבוצה מסויימת?

...זה זהה לשאלה 2

Q.4

שאלה 4

בשאלה זו נשתמש בנתונים בשם bike הזמינים כקובץ csv במודל.

אלו נתונים יומיים על השכרת אופניים בעיר סיאול שבדרום קוריאה, אשר נאספו במהלך פברואר ומרץ של 2018. כל שורה מייצגת יום אחר, והעמודות הנדרשות לנו בשאלה זו הן -

- Rented_Bikes - כמה אופניים הושכרו באותו היום (נומרי)
- Temperature - מה הייתה הטמפרטורה באותו היום (חם, נוח או קר)
- Humidity_percent - מה היה אחוז הלחות הממוצע באותו היום (גבוה או נמוך)

בסעיפים א-ב נעבוד עם תת המדגם עבורו Humidity_percent = high.

תוכלו ליצור את תת המדגם הזה למשל כך,

```
df_high <- df[df$Humidity_percent == "high",]
```

א. נרצה לדעת האם תוחלת מספר האופניים שהושכרו (Rented_Bikes) שונה בין ימים בהם יש טמפרטורה (Temperature) שונה.
בדקו זאת בעזרת ניתוח שונות חד כיווני.

I. פרטו אילו הנחות הנחתם. האם נראה כי הן מתקיימות?

II. דווחו את התוצאות שקיבלתם. מה המסקנות עבור $\alpha = 0.05$?

ב. בנו רווחי סמך סימולטניים עם רמת סמך משפחתית של 95% עבור כל ההשוואות הזוגיות, בשתי השיטות הבאות:

I. על ידי תיקון בונפרוני.

II. על ידי שיטת טוקי (או שיטת טוקי-קרמר, מה שמתאים כאן).

III. אם הייתם צריכים לדווח על רווחי סמך (למשל ללקוח, או למאמר שאתם כותבים), רווחי סמך של איזו שיטה הייתם בוחרים? הסבירו.

ג. נחזור למדגם המלא ונרצה לדעת האם תוחלת מספר האופניים שהושכרו (Rented_Bikes) שונה בין ימים בהם יש טמפרטורה (Temperature) ולחות (Humidity_percent) שונות.
בדקו זאת בעזרת ניתוח שונות דו כיווני, עם אינטאקרציה (רק אם יש עדות בנתונים שיש בה צורך).

I. פרטו אילו הנחות הנחתם. האם נראה כי הן מתקיימות?

II. הציגו גרף אינטארקציה (interaction plot). הסבירו מה רואים בו.

III. דווחו את התוצאות שקיבלתם. מה המסקנות עבור $\alpha = 0.05$?

a

```
bike <- read_csv("bike.csv")
```

```
## Parsed with column specification:
## cols(
##   Date = col_date(format = ""),
##   Rented_Bikes = col_double(),
##   Humidity_percent = col_character(),
##   Temperature = col_character(),
##   Wind_speed = col_double(),
##   Rainfall = col_character(),
##   Snowfall = col_character(),
##   Holiday = col_character(),
##   Functioning.Day = col_character(),
##   Visibility = col_double(),
##   Solar_Radiation = col_double()
## )
```

```
bike <- bike %>%
  dplyr::select(Rented_Bikes,Temperature,Humidity_percent) %>%
  filter(Humidity_percent == "high") %>%
  mutate(across(c(Temperature,Humidity_percent),as.factor))
bike_aov <- aov(Rented_Bikes~Temperature,data = bike)
#assumptions
glue("we assume equal variance")
```

```
## we assume equal variance
```

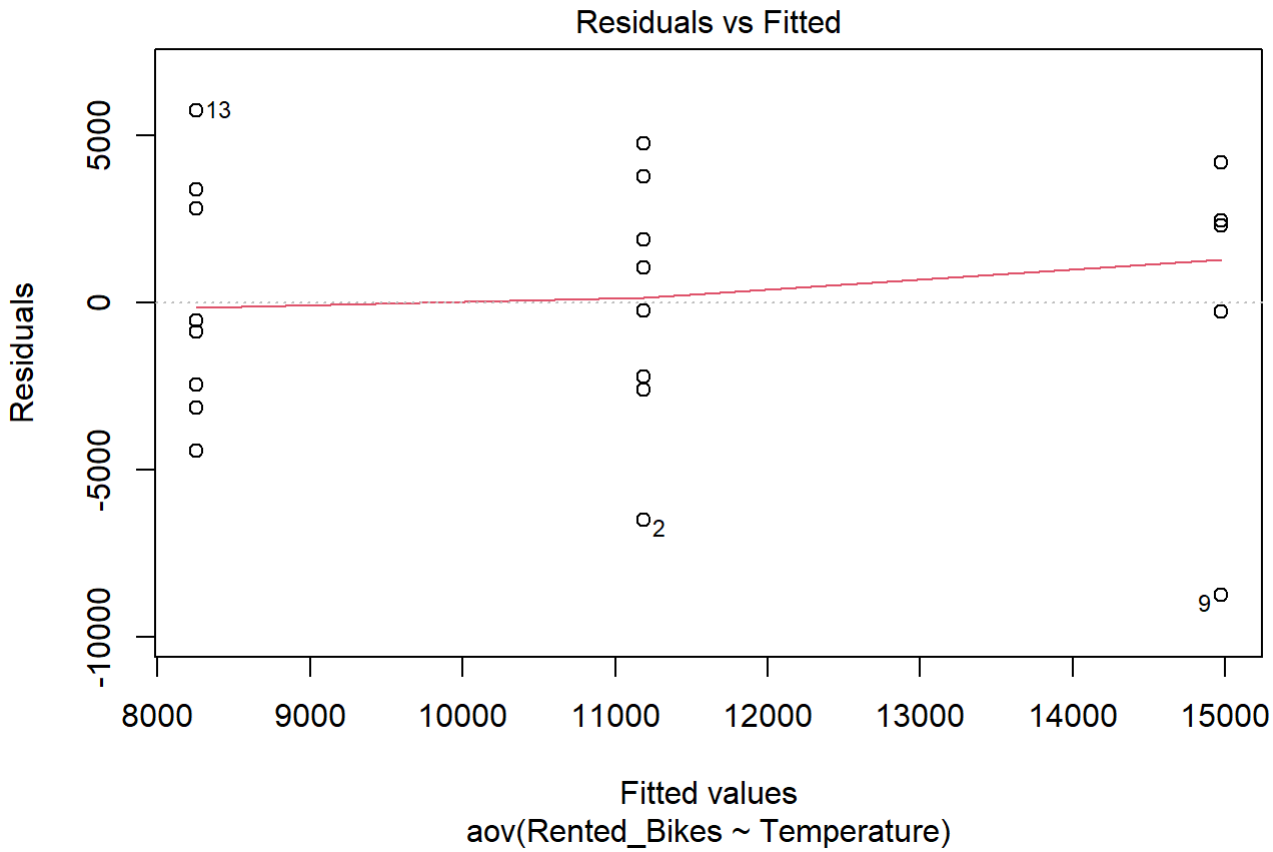
```
levene_test <-levene.test(bike$Rented_Bikes,bike$Temperature,location = "mean")
glue("pvalue for levene test is {round(levene_test$p.value,4)} therefore we will not reject the
 null and conclude equal varicane")
```

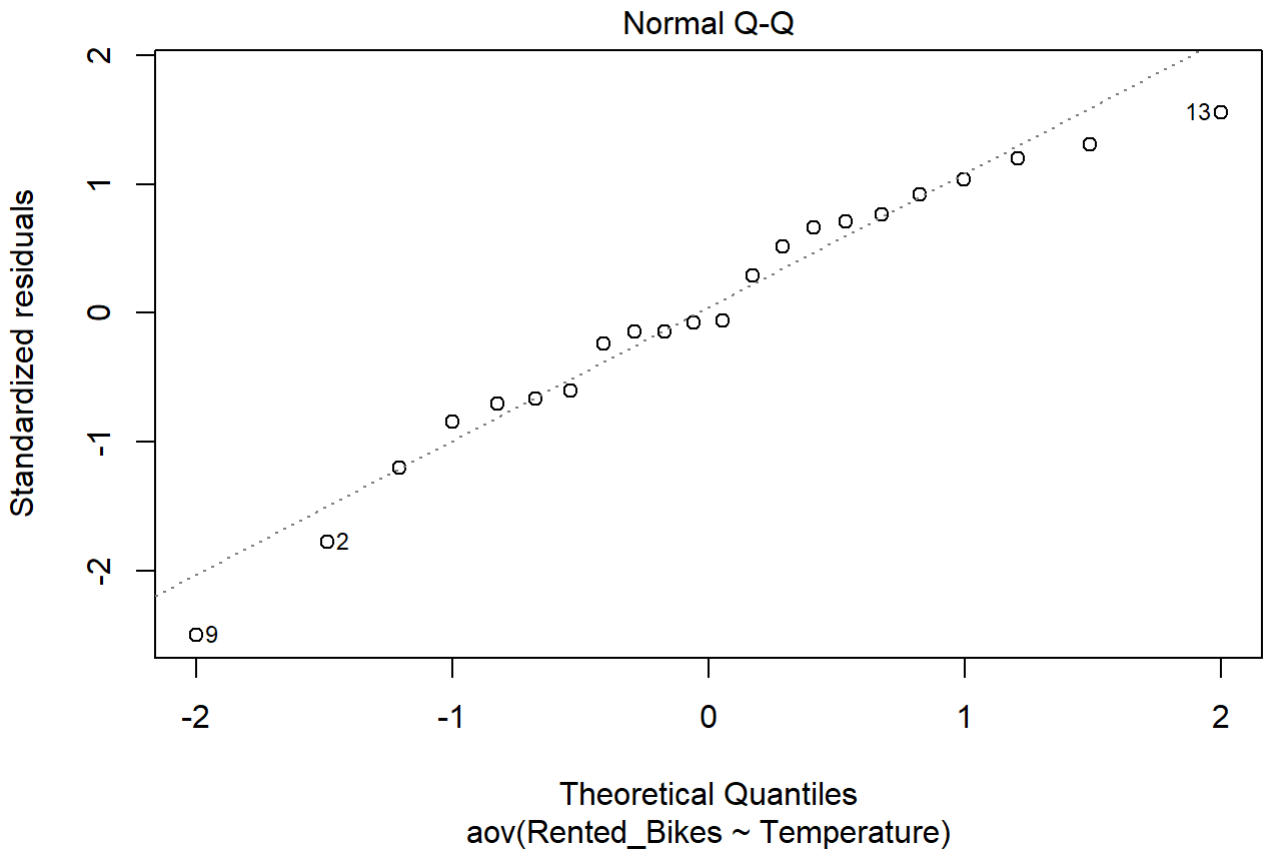
```
## pvalue for levene test is 0.7505 therefore we will not reject the null and conclude equal var
icane
```

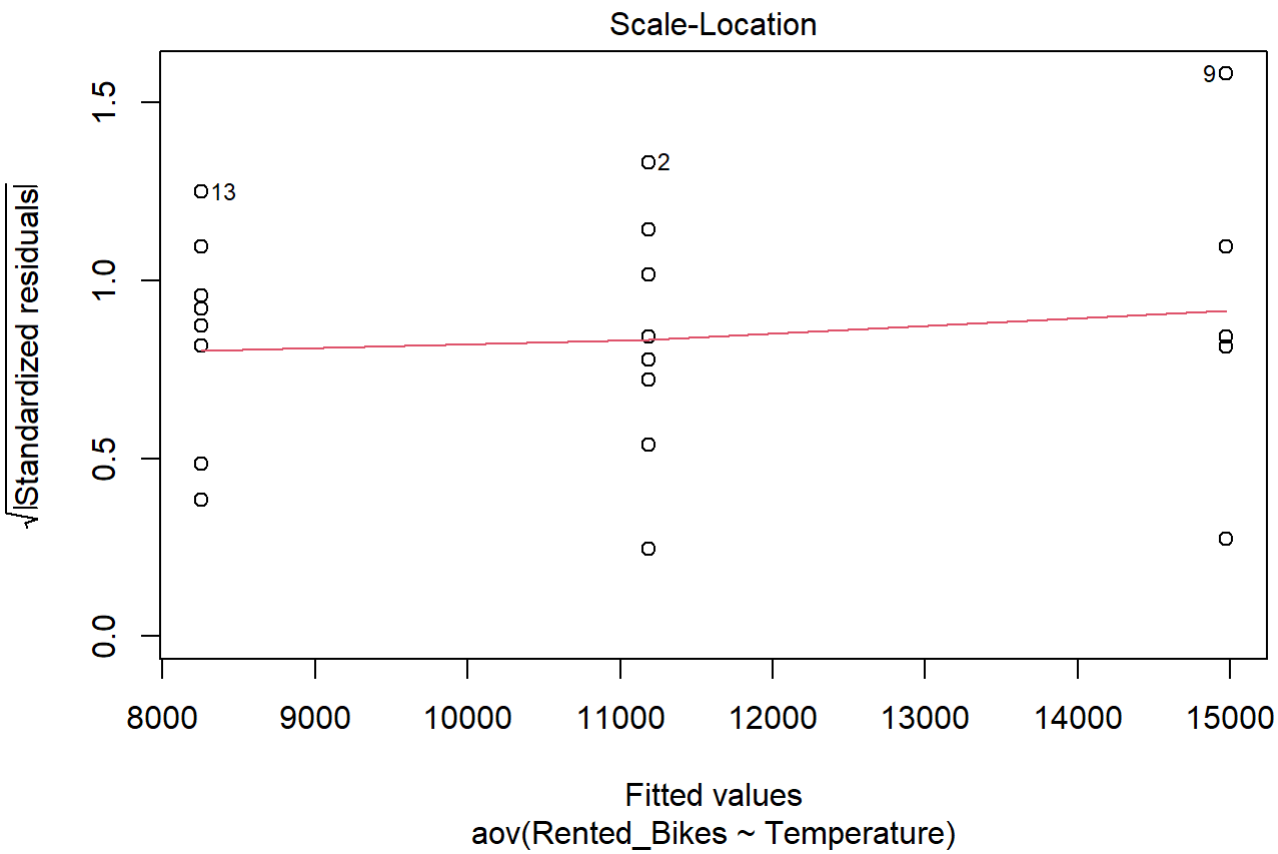
```
glue("we assume normality of the residuals")
```

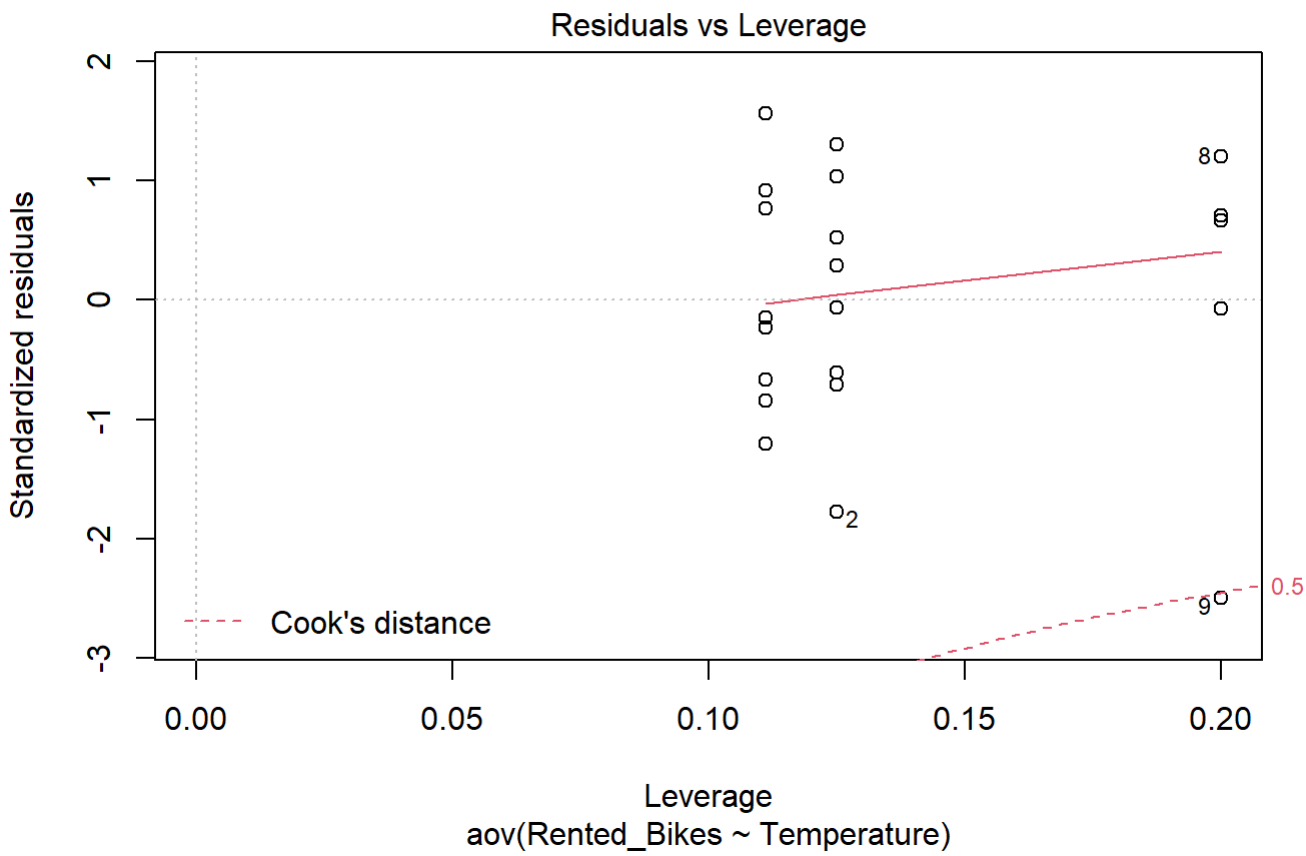
```
## we assume normality of the residuals
```

```
plot(bike_aov)
```


Residuals vs Fitted
Residuals
Fitted values
aov(Rented_Bikes ~ Temperature)
5000
0
-5000
-10000
8000
9000
10000
11000
12000
13000
14000
15000
13
2
9


Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(Rented_Bikes ~ Temperature)
2
1
0
-1
-2
-2
-1
0
1
2
13
2
9


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
8000
9000
10000
11000
12000
13000
14000
15000
Fitted values
aov(Rented_Bikes ~ Temperature)
13
2
9


Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
-3
0.00
0.05
0.10
0.15
0.20
Leverage
aov(Rented_Bikes ~ Temperature)
Cook's distance
0.5
8
2
9

```
glue("qqplot looks ok")
```

```
## qqplot looks ok
```

```
glue("we also assume independece but we cant check it")
```

```
## we also assume independece but we cant check it
```

```
#pvalue
pv_bike <- summary(bike_aov)[[1]][["Pr(>F)"]][1]
glue("pvalue is: {pv_bike} therefore we will reject the null and conclude that there is a differ
ence in bike rentals for different Temperatures")
```

```
## pvalue is: 0.0206880232217566 therefore we will reject the null and conclude that there is a
difference in bike rentals for different Temperatures
```

b

```
contr <- rbind(
  "Temperature hot - Temperature cold" = c(-1,1,0),
  "Temperature nice - Temperature cold" = c(-1,0,1),
  "Temperature nice - Temperature hot" = c(0,-1,1)
)
glue("bonferonni ci's method")
```

```
## bonferonni ci's method
```

```
pairwise_res <- glht(bike_aov,linfct = mcp(Temperature = contr))
pairwise_ci <- confint(pairwise_res,level = 1- (0.05/3) ,calpha = univariate_calpha())
pairwise_ci
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = Rented_Bikes ~ Temperature, data = bike)
##
## Quantile = 2.6251
## 98.3333333333333% confidence level
##
##
## Linear Hypotheses:
##                                          Estimate   lwr        upr
## Temperature hot - Temperature cold == 0   6720.2889   992.0423 12448.5355
## Temperature nice - Temperature cold == 0  2932.2639 -2057.9844  7922.5122
## Temperature nice - Temperature hot == 0  -3788.0250 -9642.7382  2066.6882
```

```
glue("Tukey - Kramer method")
```

```
## Tukey - Kramer method
```

```
TukeyHSD(bike_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Rented_Bikes ~ Temperature, data = bike)
##
## $Temperature
##                diff       lwr       upr     p adj
## hot-cold   6720.289  1176.770 12263.808 0.0162093
## nice-cold  2932.264 -1897.056  7761.584 0.2942483
## nice-hot  -3788.025 -9453.932  1877.882 0.2315525
```

```
glue("i would report Tukey - Kramer CI because they are smaller")
```

```
## i would report Tukey - Kramer CI because they are smaller
```

c

```
bike <- read_csv("bike.csv")
```

```
## Parsed with column specification:
## cols(
##   Date = col_date(format = ""),
##   Rented_Bikes = col_double(),
##   Humidity_percent = col_character(),
##   Temperature = col_character(),
##   Wind_speed = col_double(),
##   Rainfall = col_character(),
##   Snowfall = col_character(),
##   Holiday = col_character(),
##   Functioning.Day = col_character(),
##   Visibility = col_double(),
##   Solar_Radiation = col_double()
## )
```

```
bike <- bike %>%
  dplyr::select(Rented_Bikes,Temperature,Humidity_percent) %>%
  mutate(across(c(Temperature,Humidity_percent),as.factor))

bike %>%
  group_by(Temperature,Humidity_percent) %>%
  summarize(n())
```

```
## `summarise()` regrouping output by 'Temperature' (override with `.groups` argument)
```

```
## # A tibble: 6 x 3
## # Groups:   Temperature [3]
##   Temperature Humidity_percent `n()`
##   <fct>       <fct>            <int>
## 1 cold        high                 9
## 2 cold        low                 28
## 3 hot         high                 5
## 4 hot         low                  2
## 5 nice        high                 8
## 6 nice        low                  7
```

```
glue("we have unblaced groups")
```

```
## we have unblaced groups
```

```
#see if intercation is necceserty
full_model <- lm(Rented_Bikes ~ Temperature*Humidity_percent,data = bike)
no_inter_model <- lm(Rented_Bikes ~ Temperature + Humidity_percent,data = bike)
anov_test <- anova(no_inter_model,full_model)
pv <- anov_test[["Pr(>F)"]][2]
glue("interaction relevence p value is: {pv} therefore we will reject the null, meaning the inte
raction is needed")
```

```
## interaction relevence p value is: 0.0295053593256452 therefore we will reject the null, meani
ng the interaction is needed
```

```
#assumptions
glue("we assume equal variance")
```

```
## we assume equal variance
```

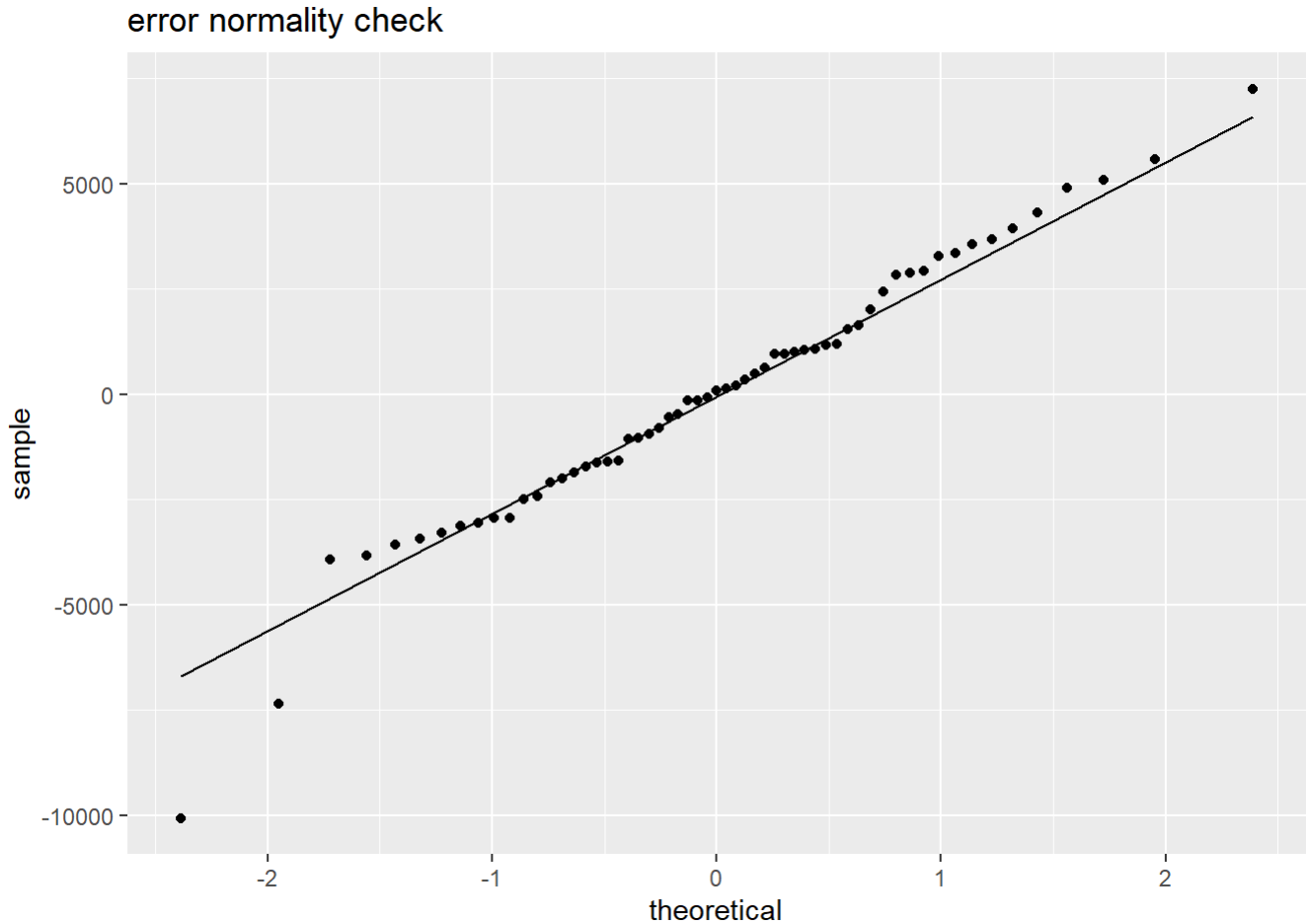
```
levene_test <-levene.test(bike$Rented_Bikes,bike$Temperature:bike$Humidity_percent,location = "m
ean")
glue("pvalue for levene test is {round(levene_test$p.value,4)} therefore we will not reject the
 null and conclude equal varicane")
```

```
## pvalue for levene test is 0.1324 therefore we will not reject the null and conclude equal var
icane
```

```
glue("we assume normality of the residuals")
```

```
## we assume normality of the residuals
```

```
ggplot(bike,aes(sample = Rented_Bikes - predict(no_inter_model,bike[c("Temperature","Humidity_pe
rcent")]))) +
  stat_qq()+
  stat_qq_line()+
  ggtitle("error normality check")
```




```
glue("looks ok")
```

```
## looks ok
```

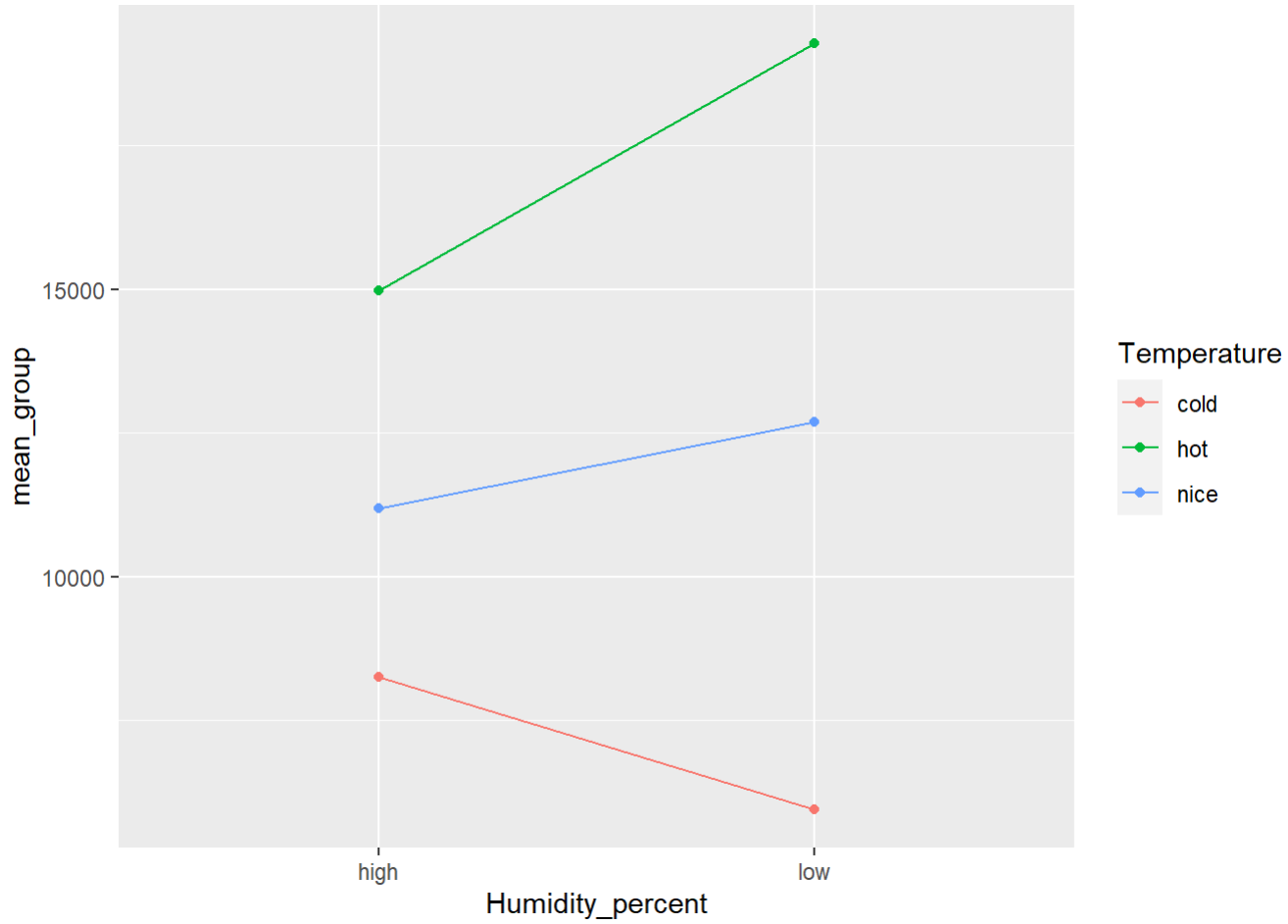
```
glue("we also assume independece but we cant check it")
```

```
## we also assume independece but we cant check it
```

```
Rented_Bikes_mean <- bike %>%
  group_by(Temperature,Humidity_percent) %>%
  summarize(mean_group = mean(Rented_Bikes))
```

```
## `summarise()` regrouping output by 'Temperature' (override with `.groups` argument)
```

```
Rented_Bikes_mean %>%
  ggplot() +
  aes(x = Humidity_percent,y = mean_group, color = Temperature)+
  geom_line(aes(group = Temperature))+
  geom_point()
```




```
glue("we can see in this graph that the lines are not parallel meaning we have a different slope
for each interaction so the interaction does have an effect")
```

```
## we can see in this graph that the lines are not parallel meaning we have a different slope fo
r each interaction so the interaction does have an effect
```

Q.5

שאלה 5

בשאלה זו נשתמש בנתונים בשם bike הזמינים כקובץ csv במודל.

אלו נתונים יומיים על השכרת אופניים בעיר סיאול שבדרום קוריאה, אשר נאספו במהלך פברואר ומרץ של 2018. כל שורה מייצגת יום אחר, והעמודות הנדרשות לנו בשאלה זו הן -

- Date - תאריך
- Rented_Bikes - כמה אופניים הושכרו באותו היום (נומרי)
- Humidity_percent - מה היה אחוז הלחות הממוצע באותו היום (גבוה או נמוך)
- Wind_speed - מהירות הרוח הממוצעת באותו היום (נומרי)
- Rainfall - האם ירד גשם באותו היום (כן/לא)
- Snowfall - האם ירד שלג באותו היום (כן/לא)
- Solar_Radiation - הקרינה הסולארית הממוצעת באותו היום (נומרי)
- Visibility - הראות שהייתה באותו היום (נומרי)

נרצה לבחון את הנתונים ולחפש קשרים מעניינים. לכן נרצה לבצע 12 מבחני t test שונים על הנתונים הללו -

- בידקו האם מס' האופניים שהושכרו (Rented_Bikes) שונה בין ימים עם לעומת בלי גשם (Rainfall).
- בידקו האם מס' האופניים שהושכרו (Rented_Bikes) שונה בין ימים עם לעומת בלי שלג (Snowfall).
- בידקו האם מס' האופניים שהושכרו (Rented_Bikes) שונה בין ימים עם לחות (Humidity_percent) גבוהה לעומת נמוכה.
- בידקו האם מהירות הרוח (Wind_speed) שונה בין בין ימים עם לעומת בלי גשם (Rainfall).
- בידקו האם מהירות הרוח (Wind_speed) שונה בין בין ימים עם לעומת בלי שלג (Snowfall).
- בידקו האם מהירות הרוח (Wind_speed) שונה בין בין ימים עם לחות (Humidity_percent) גבוהה לעומת נמוכה.
- בידקו האם הקרינה הממוצעת (Solar_Radiation) שונה בין ימים עם לעומת בלי גשם (Rainfall).
- בידקו האם הקרינה הממוצעת (Solar_Radiation) שונה בין ימים עם לעומת בלי שלג (Snowfall).
- בידקו האם הקרינה הממוצעת (Solar_Radiation) שונה בין ימים עם לחות (Humidity_percent) גבוהה לעומת נמוכה.
- בידקו האם הראות (Visibility) שונה בין ימים עם לעומת בלי גשם (Rainfall).
- בידקו האם הראות (Visibility) שונה בין ימים עם לעומת בלי שלג (Snowfall).
- בידקו האם הראות (Visibility) שונה בין ימים עם לחות (Humidity_percent) גבוהה לעומת נמוכה.

מכיוון שביצענו 12 השוואות שונות, נרצה לתקן להשוואות מרובות.

א. מה יצאו הpvalues לפני תיקון להשוואות מרובות?

ב. תקנו ע"י שיטת בנימיני-הוכברג בעזרת הפונקציה p.adjust שהוצגה בכיתה.
מה יצאו הadjusted pvalues? מה המסקנות עבור $\alpha = 0.05$?

ג. תקנו ע"י שיטת בנימיני-הוכברג בעזרת קוד שכתבתם בעצמכם באופן "ידני".
מה יצאו הadjusted pvalues? מה המסקנות עבור $\alpha = 0.05$?

ד. תקנו ע"י שיטת בונפרוני. מה המסקנות עבור $\alpha = 0.05$?

ה. הסבירו במילים על מה שולטת שיטת בנימיני-הוכברג שולטת ועל מה שולטת שיטת בופנרוני.

a

```
bike <- read_csv("bike.csv")
```

```
## Parsed with column specification:
## cols(
##   Date = col_date(format = ""),
##   Rented_Bikes = col_double(),
##   Humidity_percent = col_character(),
##   Temperature = col_character(),
##   Wind_speed = col_double(),
##   Rainfall = col_character(),
##   Snowfall = col_character(),
##   Holiday = col_character(),
##   Functioning.Day = col_character(),
##   Visibility = col_double(),
##   Solar_Radiation = col_double()
## )
```

```
bike <- bike %>%
  mutate(across(c(Rainfall,Snowfall,Humidity_percent),as.factor))

t_test_pval <- function(formula){
  #checks equality assumption and based on that conducts the test
  f_test_pv <- var.test(formula)$p.value
  if (f_test_pv > 0.05){
    return(t.test(formula)$p.value)
  }
  return(t.test(formula)$p.value)
}
bike_tibl <- tibble(values = rep(c("Rented_Bikes","Wind_speed","Solar_Radiation","Visibility"),e
ach = 3),
       groups = rep(c("Rainfall","Snowfall","Humidity_percent"),4))
bike_tibl <- bike_tibl %>%
  rowwise() %>%
  mutate(pval = t_test_pval(pull(bike,values)~pull(bike,groups)))
glue("pvalues before correction are:")
```

```
## pvalues before correction are:
```

```
bike_tibl
```

```
## # A tibble: 12 x 3
## # Rowwise:
##    values          groups                 pval
##    <chr>           <chr>                 <dbl>
##  1 Rented_Bikes    Rainfall         0.00343
##  2 Rented_Bikes    Snowfall         0.000105
##  3 Rented_Bikes    Humidity_percent 0.0219
##  4 Wind_speed      Rainfall         0.694
##  5 Wind_speed      Snowfall         0.291
##  6 Wind_speed      Humidity_percent 0.659
##  7 Solar_Radiation Rainfall         0.0247
##  8 Solar_Radiation Snowfall         0.741
##  9 Solar_Radiation Humidity_percent 0.439
## 10 Visibility      Rainfall         0.0238
## 11 Visibility      Snowfall         0.784
## 12 Visibility      Humidity_percent 0.00000404
```

b

```
bike_tibl$pval_adjust_BH = p.adjust(bike_tibl$pval,method = "BH")
bike_tibl$BH_conc <- ifelse(bike_tibl$pval_adjust_BH<0.05,"Reject","don't reject")
bike_tibl
```

```
## # A tibble: 12 x 5
## # Rowwise:
##    values          groups                 pval pval_adjust_BH BH_conc
##    <chr>           <chr>                 <dbl>          <dbl> <chr>
##  1 Rented_Bikes    Rainfall         0.00343         0.0137    Reject
##  2 Rented_Bikes    Snowfall         0.000105        0.000627  Reject
##  3 Rented_Bikes    Humidity_percent 0.0219          0.0494    Reject
##  4 Wind_speed      Rainfall         0.694           0.784     don't reject
##  5 Wind_speed      Snowfall         0.291           0.499     don't reject
##  6 Wind_speed      Humidity_percent 0.659           0.784     don't reject
##  7 Solar_Radiation Rainfall         0.0247          0.0494    Reject
##  8 Solar_Radiation Snowfall         0.741           0.784     don't reject
##  9 Solar_Radiation Humidity_percent 0.439           0.659     don't reject
## 10 Visibility      Rainfall         0.0238          0.0494    Reject
## 11 Visibility      Snowfall         0.784           0.784     don't reject
## 12 Visibility      Humidity_percent 0.00000404      0.0000485 Reject
```

c

```
adj_pv <- function(vec,alpha,index){
  bool_vec = vec<=alpha
  M <- length(bool_vec)
  if (index > 1) {bool_vec[1:index-1] = FALSE}
  if (sum(bool_vec) == 0){return(min(vec[index:M]))}
  nex_rej <- min(which(bool_vec))
  return(vec[nex_rej])
}
my_bh <- function(vec,alpha = 0.05){
  ord <- order(vec)
  sorted_vec <- sort(vec)
  M <- length(vec)
  k <- 1:M
  corrected_alphas <- k/M*alpha
  corrected_pvals <- c()
  for (i in k){
    corrected_pvals[i] <- adj_pv(sorted_vec*M/k,alpha,i)
  }
  return(ifelse(corrected_pvals >1 ,1 ,corrected_pvals)[order(ord)])
}

bike_tibl$pval_adjust_my_BH = my_bh(bike_tibl$pval)
bike_tibl$my_BH_conc <- ifelse(bike_tibl$pval_adjust_my_BH<0.05,"Reject","don't reject")
bike_tibl %>%
  dplyr::select(values,groups,pval_adjust_my_BH,my_BH_conc)
```

```
## # A tibble: 12 x 4
## # Rowwise:
##    values          groups           pval_adjust_my_BH my_BH_conc
##    <chr>           <chr>                        <dbl> <chr>
##  1 Rented_Bikes    Rainfall                  0.0137    Reject
##  2 Rented_Bikes    Snowfall                  0.000627  Reject
##  3 Rented_Bikes    Humidity_percent          0.0494    Reject
##  4 Wind_speed      Rainfall                  0.784     don't reject
##  5 Wind_speed      Snowfall                  0.499     don't reject
##  6 Wind_speed      Humidity_percent          0.784     don't reject
##  7 Solar_Radiation Rainfall                  0.0494    Reject
##  8 Solar_Radiation Snowfall                  0.784     don't reject
##  9 Solar_Radiation Humidity_percent          0.659     don't reject
## 10 Visibility      Rainfall                  0.0494    Reject
## 11 Visibility      Snowfall                  0.784     don't reject
## 12 Visibility      Humidity_percent          0.0000485 Reject
```

d

```
bonf_pval <- bike_tibl$pval* length(bike_tibl$pval)
bike_tibl$p_adjusted_bonf <- ifelse(bonf_pval>1,1,bonf_pval)
bike_tibl$bonf_conc <- ifelse(bike_tibl$p_adjusted_bonf<0.05,"Reject","don't reject")
bike_tibl %>%
  dplyr::select(values,groups,p_adjusted_bonf,bonf_conc)
```

```
## # A tibble: 12 x 4
## # Rowwise:
##    values          groups           p_adjusted_bonf bonf_conc
##    <chr>           <chr>                      <dbl> <chr>
##  1 Rented_Bikes    Rainfall               0.0411    Reject
##  2 Rented_Bikes    Snowfall               0.00125   Reject
##  3 Rented_Bikes    Humidity_percent       0.262     don't reject
##  4 Wind_speed      Rainfall               1         don't reject
##  5 Wind_speed      Snowfall               1         don't reject
##  6 Wind_speed      Humidity_percent       1         don't reject
##  7 Solar_Radiation Rainfall               0.296     don't reject
##  8 Solar_Radiation Snowfall               1         don't reject
##  9 Solar_Radiation Humidity_percent       1         don't reject
## 10 Visibility      Rainfall               0.286     don't reject
## 11 Visibility      Snowfall               1         don't reject
## 12 Visibility      Humidity_percent       0.0000485 Reject
```

בינימיני הוכברג שולטת על ה

FDR

כלומר על אחוז ההשערות שדחינו אך לא היינו אמורים לדחות

לעומת זאת בונפריני שומר על כך שלא נעשה כלל טעות מסוג ראשון בסיכוי אלפא