ISA 365 Homework 1

Last complied on May 17, 2024

1. Read in the "Houses.csv" data set. Show the `head()`. Use the option `stringsAsFactors = TRUE` in your `read.csv()` statement.

```
df<- read.csv("C:\\Users\\Cam\\OneDrive\\Desktop\\ISA365\\Houses.csv", stringsAsFactors = TRUE)
head(df)
```

```
##   Price SqFt Age Fireplace        Type
## 1   595 1605  34       Yes    Colonial
## 2   588 1263  37        No     Rancher
## 3   465 1864  29       Yes    Colonial
## 4   552 1253  43        No     Rancher
## 5   564 2726  17        No Split-Level
## 6   693 2996  15        No Split-Level
```

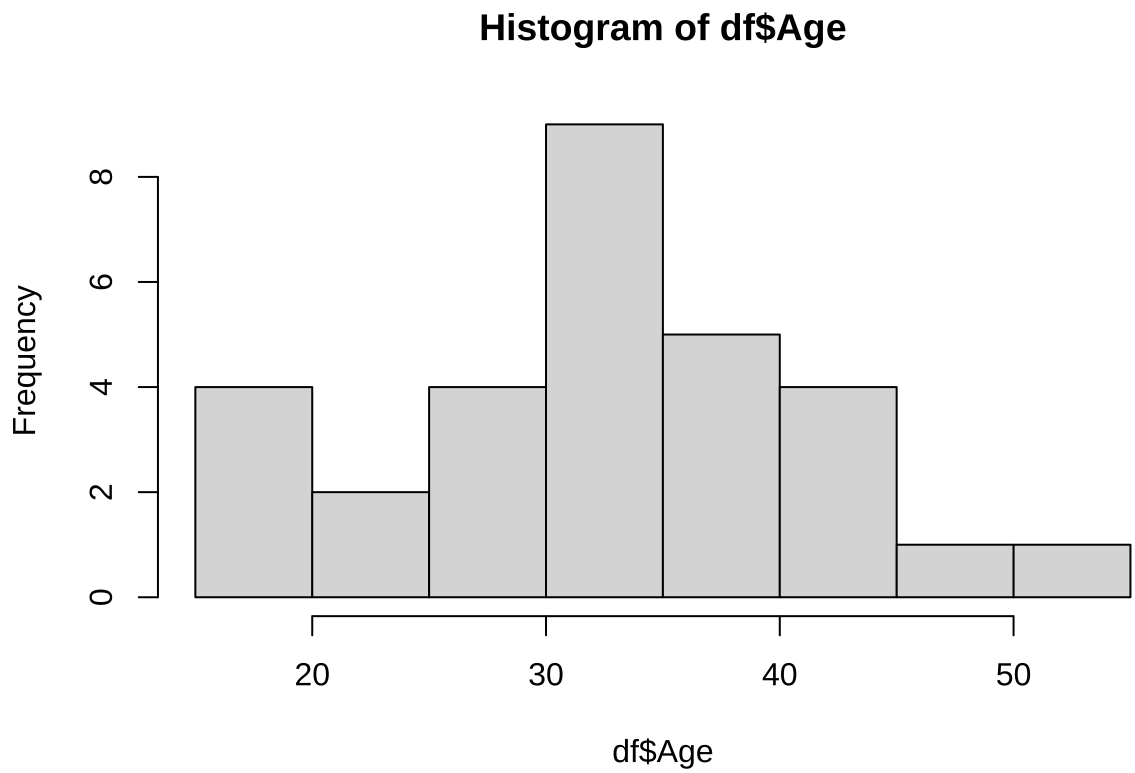
2. Make a table showing the mean and standard deviation of the Age variable.

```
library(tidyverse)

df_houses <- df |>
  summarize(mean_age = mean(Age), sd_age = sd(Age))

df_houses
```

```
##   mean_age   sd_age
## 1     32.9 9.060524
```

3. Make a histogram of the Age variable.

```
hist(df$Age)
```



4. Comment on the shape of the Age variable.

```r
summary(df$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   29.00   34.50   32.90   36.75   51.00
```

The shape of the Age variable suggests a unimodal distribution, meaning there's one predominant age range (30-35) for houses in the dataset. From this summary, we can see that the youngest house is 15 years old and the oldest house is 51 years old. This means the age variable ranges from 15 to 51, showing a spread of 36 years. This graph could be slightly skewed, but the skewness would be negligible if so.

5. What is necessary to show causation?

Causation requires strong evidence that one event directly leads to another. For example, this is expressed through the Big Three of Causality:

```
1. When X changed, Y also changed.  If X changes and Y doesn't change, then we cannot assert that X cau
```

```
2. X happened before Y. If X happens after Y, then X cannot cause Y.
```

```
3. Nothing else besides X changed systematically.
```

You cannot prove causality without an experiment.

6. Go to Miami University main landing page https://miamioh.edu/. Write down two or more tests you would like to make, for example I would like to see the slogan "Miami made. Forever prepared." changed out with a different one (you can't use this example).

A few tests I would like to make are the following:

```
1. Instead of having the hamburger button in the upper left hand corner that hides additional tabs/navi

2. I would like to see the tabs at the top of the screen (apply, visit, request info, etc.) made larger

3. Rather than having the video play button in the bottom right corner of the main picture, I would lik
```

7. Give two situations where experiments can't be conducted.

    1. They can't be conducted when it is impossible to run an experiment. For example, it's unethical to design an experiment where a group of people are asked to smoke to directly observe the impact of smoking on lung cancer. Instead, observational studies should be used.
    2. They can't be conducted when the population is too small. For example, conducting experiments on critically endangered species such as black rhinos, where the population size is extremely small, could potentially harm their very existence and is not viable. Again, observational studies should be used here instead.

8. Read the pdf "study 1.pdf". Is study 1 observational or experimental?

Observational

9. What is/are the responses measured in study 1?

Leakage, how easy it is to clean, temperature retention, other design features (comfort, durability)

10. What are the factors or treatments?

The factors or treatments (what we are changing) in this study are the different Stanley Tumbler dupes.

1. An A/B test results in 18 successes in 223 trials for version A and 26 successes in 287 trials for version B. What is the probability that the true proportion of version A is larger than the true proportion for version B. i.e. $P(p_a > p_b)$? Hint: Use the Beta-Binomial Model.

```
set.seed(100) # setting seed so the probability does not change each time code chunk is run
iter=100000


a=18+1
b=(223-18)+1
a1=26+1
b1=(287-26)+1
count<-c()
for (i in 1:iter){
A<-rbeta(1, a, b)
B<-rbeta(1, a1, b1)
count[i]<-ifelse(A>B, 1, 0)
```

```
}
pdiff<-sum(count)/(iter-1)
pdiff
```

```
## [1] 0.3578636
```

Booking.com ran a test in July of 2019 to evaluate the addition of logos next to its booking options (see below):

The company ran the experiment for two weeks and among the various response was the average spend for each user. Describe the treatment and response for this experiment.

2. What is the treatment and what is the response for the Booking.com experiment?

The treatment in this experiment was the addition of logos next to its booking options on the website (accommodation, flights, flight and hotel), while the response was the average spend for each user. I assume the experiment was conducted by showing one group of users the booking options with logos and showing another group of users the booking options without the logos. Therefore, the response would be measured by comparing the average spend of users who were exposed to the logos with those who were not. This determines if the logos had any statistically significant effect on spending behavior.

3. Read in the data from "booking.csv". What is the average spend and standard deviation of spend for each version? I find it best to use the `stringsAsFactors=TRUE` argument in my `read.csv()` statement here.

```
df<- read.csv("C:\\Users\\Cam\\OneDrive\\Desktop\\ISA365\\booking.csv", stringsAsFactors = TRUE)
head(df)
```

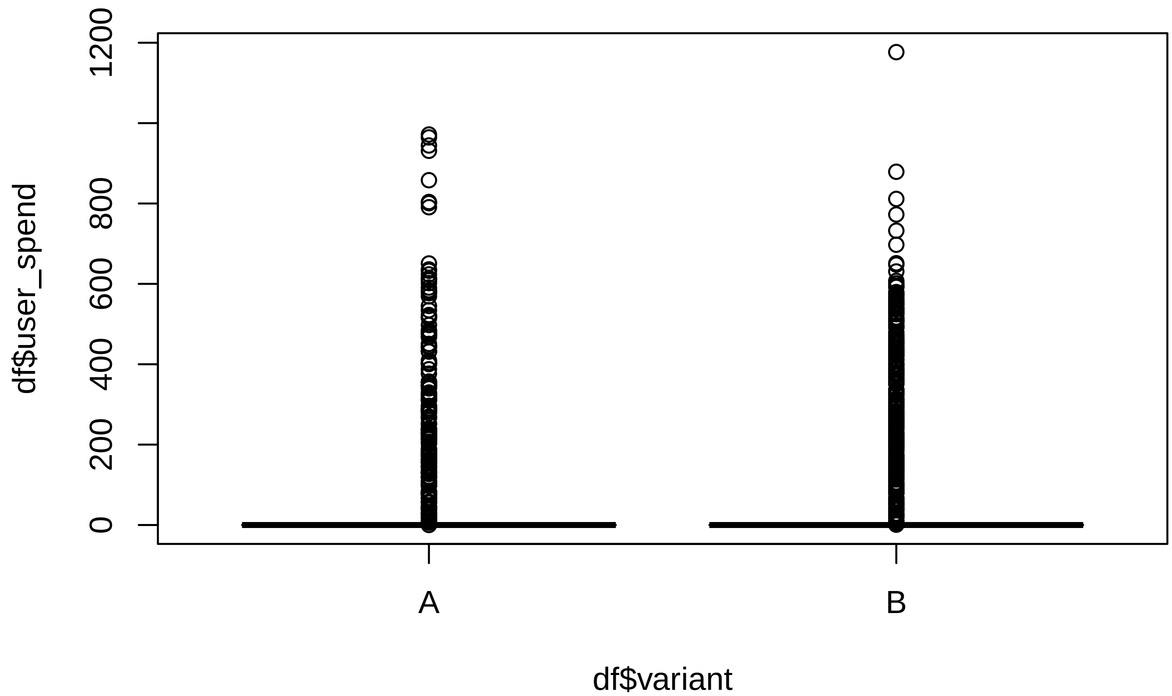
```
##   visit_id user_spend variant
## 1    51763          0       A
## 2    52436          0       A
## 3    50203          0       A
## 4    56289          0       A
## 5    14265          0       A
## 6    16182          0       A
```

```
library(tidyverse)
df |> group_by(variant) |> summarize("average spend"=mean(user_spend), "std spend"=sd(user_spend),"coun
```

```
## # A tibble: 2 x 4
##   variant `average spend` `std spend` count
##   <fct>             <dbl>       <dbl> <int>
## 1 A                  10.3        66.8  4889
## 2 B                  13.2        72.6  5630
```

4. Visualize the data by each variant. Use any graph you feel shows the distribution of spend the best. Comment on your findings.

```r
boxplot(df$user_spend~df$variant)
```



Based on this boxplot, it appears that the average spend amounts for both Variant A and Variant B are somewhat similar, with Variant B potentially having a slightly higher average spend. However, the boxplot alone makes it challenging to precisely determine the extent of this difference, so, while it does provide a visual representation of the spread of spend for each variant, additional analysis such as calculating summary statistics or conducting hypothesis tests would be needed to confidently quantify and interpret the difference in average spend between them.

5. Find a 95% confidence interval for the difference in average spend for each variant.

```r
t.test(df$user_spend~df$variant, var.equal = FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  df$user_spend by df$variant
## t = -2.1428, df = 10482, p-value = 0.03215
## alternative hypothesis: true difference in means between group A and group B is not equal to 0
## 95 percent confidence interval:
##  -5.5778297 -0.2482742
## sample estimates:
## mean in group A mean in group B 
##        10.31172        13.22478
```

6. Give a sentence explaining that interval so that anyone can understand what the test results are showing.

dont say "using the confidence interval, we can conclude..." just get right to the point

Using just the confidence interval, we can conclude that because 0 is not in the confidence interval, there is a difference between the average spend of version A and B. The test showed that the addition of logos next to the booking options significantly increased average spend per customer by about $2.91 per customer over booking options without logos (95% CI =($0.25, $5.58)).

7. For the version with the highest spend, what spend can booking.com expect to see if they implemented the best version from above? Use a 95% confidence interval? Hint: first use the `filter` command from `tidyverse` to make a data frame containing only the values from variant A. Then use `t.test` to perform a single sample test.

```r
library(tidyverse)

variant_b <- df |>
  filter(variant == "B")

t.test(variant_b$user_spend, var.equal = FALSE)
```

```
##
##  One Sample t-test
##
## data:  variant_b$user_spend
## t = 13.67, df = 5629, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  11.32825 15.12130
## sample estimates:
## mean of x
##  13.22478
```

The following problems you answered in the Homework 3 mechanics:

You manage a website where 10% of the visitors get to the checkout page with items in the cart, but only half of these actually complete the purchase. You want to test a new button that, you hope, will increase the percentage of actual purchases. The new button is "A" and the current button is "B". To keep the math simple, you'd like to see a 10% lift (improvement) at the usual power (0.80) and significance (0.05) levels. You want half the visitors who get to the checkout page assigned to group "A" and the other half assigned to group "B". Here's the question: Do you assign visitors to "A" or "B" when they first get your website or when they actually get to the checkout page?

If you assign treatment when visitors first get to the website, then the difference you want to detect is a change from 5.0% to 5.5%. How many visitors need to come to the website so that this test can be conducted?

Follow on to 1: If you assign treatment when visitors finally get to the checkout page, then the difference you want to detect is a change from 50.0% to 55.0%. How many visitors need to come to the website so that this test can be conducted?

1. Which option would you prefer? Why?

- I would prefer the second option, which is assigning treatment when visitors finally get to the checkout page, because it directly assesses the impact of the button change. Considering our goal is to "increase the percentage of actual purchases," I believe option two presents the critical point of decision-making in the purchase process, where users are at the last step in deciding whether to continue/complete their purchase. Also, even the end of the prompt specifies assigning visitors to group A and B once they reach the checkout page, which supports this reasoning.

As part of a class assignment in an experimental design course, a student wants to test two different pots to determine which one boils water faster, steel bottom or copper bottom. In order to do a sample size calculation, he needs an estimate of the standard deviation. To do so, he took one of the pots and boiled water in it six times. The times to a rolling boil were 133, 148, 137, 142, 145 and 150 seconds. What sample size should he use to detect a difference of 15 seconds? By the way, if you want to create a vector of times use this: `times<-c(133, 148, 137, 142, 145, 150)` The following questions refer to this experiment.

2. What sample size should he use to detect a difference of 5 seconds? Use power=0.8 and α=0.05.

- He should use a sample size of 28 to detect a difference of 5 seconds.
- Also, to answer the other question in the prompt, he should use a sample size of 5 to detect a difference of 15 seconds.

```
times<-c(133, 148, 137, 142, 145, 150)
sd_time<-sd(times)

# sample size to detect difference of 5 seconds
power.t.test(n=NULL, delta=5, sd=sd_time, sig.level = 0.05, power=0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 27.80558
##           delta = 5
##              sd = 6.534524
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# sample size to detect difference of 15 seconds
power.t.test(n=NULL, delta=15, sd=sd_time, sig.level = 0.05, power=0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 4.195191
##           delta = 15
##              sd = 6.534524
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

3. What is the treatment this experiment?

- The treatment in this experiment is the type of pot (steel bottom or copper bottom).

4. Refer to 4: What is the response in this experiment?

- The response in this experiment is the time in seconds it takes for the water to reach a rolling boil.

5. Refer to 4: This describes the student's pilot experiment to help determine standard deviation. Criticize this pilot experiment.

There are a number of issues with the student's pilot experiment to determine standard deviation. They are:

- Selection Bias - No where in the experiment did it say that the pot was randomly selected, which means there could be bias. For example, if the pot is newer or cleaner than the other pots, the results might not be representative.
- Small Sample Size - Conducting the experiment only six times provides a very small sample size, as it may not capture the natural variability in boiling times.
- Single Pot Type - Only one type of pot (steel or copper bottom) was used, so the estimate of the standard deviation only reflects the variability of boiling times for that specific type of pot and not the entire "population". To generalize the findings, it would be necessary to test both types of pots. This is the big one
- Other Isues - The experiment does not account for potential variability in things like room temperature, starting water temperature, pot placement on the stove, and stove heat level.

6. Design a boiling water experiment. Describe the procedure for experimentation from start to finish. How much water will be used? Where will you get the water? How will "boiling" be determined? How will the temperature be measured? How will the time be measured? Should these all be done on the same burner? Back to Back? On different days? why would this matter? You can use bullets or steps if this helps.

From start to finish, my boiling water experiment would be structured as follows:

- The treatment and response will be the same as the student's experiment described above, where the treatment is the the type of pot (steel bottom or copper bottom) and the response is the time in seconds it takes the water to boil.
- How much water will be used?

A standardized amount of water will be used for consistency, so 1 liter.

- Where will you get the water?

Distilled water from Kroger will be used to eliminate any impurities that might affect the boiling point.

- How will "boiling" be determined?

Boiling will be determined by the water temperature reaching 212°F and further confirmed by the bubbles rising to the surface of the pot (indicating a phase change).

- How will the temperature be measured?

A digital thermometer with a probe will be used to measure the temperature of the water. While it is submerged, I will ensure that it is not touching the bottom of the pot.

- How will the time be measured?

A digital timer will be used to measure the time it takes for the water to reach the boiling point. It will be started when the heat is applied to the water and stopped when boiling is achieved.

- Should these all be done on the same burner? Back to Back? On different days? why would this matter?

All experiments will be conducted on the same burner to maintain consistency in heating, and they will be performed back to back on the same day to minimize variations due to external factors like room temperature and any air pressure changes. However, there will be a brief cool-down period between trials to allow the pot and thermometer to return to room temperature. The cool-down period is especially important for trials that involve pots of the same type in back to back runs, as randomization will be used to assign the order of treatments (pot types), so there is a possibility a steel bottom pot will be used a few consecutive times. Also, to avoid the threat to internal validity (instrumentation in this case), the same thermometer and person will be used to take the measurements.

Dont do it on back to back days

7. For the boiling water experiment, assume you will replicate, repeat the treatment, 5 times (boil water in each pot 5 different times). Design a randomized order for this experiment where your treatments are "steel" and "copper".

```
if(require(dplyr)==FALSE) install.packages('dplyr')
if(require(tibble)==FALSE) install.packages('tibble')

treatments_df <- tibble::tibble(
  Treatment = rep(c("steel", "copper"), each = 5)
)

set.seed(13)
random_df <- treatments_df |>
  dplyr::sample_n(size = nrow(treatments_df), replace = FALSE) |>
  dplyr::mutate(Trial = row_number())


print(random_df)
```

```
## # A tibble: 10 x 2
##    Treatment Trial
##    <chr>     <int>
##  1 copper        1
```

```
## 2 steel         2
## 3 copper        3
## 4 steel         4
## 5 steel         5
## 6 copper        6
## 7 copper        7
## 8 steel         8
## 9 copper        9
## 10 steel       10
```

The file "ttvariables.csv" contains 1000 observations on each of 23 variables. Y is a response variable corresponding to an experiment on X where 0 denotes control and 1 denotes treatment which was randomly assigned. The rest of the "Z" variables are covariates. Use this file to answer the following questions.

1. Read in the data and find the mean and standard deviation for Y at each value of X.

```r
df = readr::read_csv("ttvariables.csv")

library(tidyverse)

mean_sd <- df |>
  group_by(X) |>
  summarize(Mean_Y = mean(Y), SD_Y = sd(Y))

mean_sd
```

```
## # A tibble: 2 x 3
##       X Mean_Y  SD_Y
##   <dbl>  <dbl> <dbl>
## 1     0  10.1   4.22
## 2     1   9.83  4.05
```

2. The treatment had no effect. Perform a test to verify this.

```r
t.test(df$Y~df$X)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Y by df$X
## t = 0.94903, df = 996.34, p-value = 0.3428
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.2653141  0.7622821
## sample estimates:
## mean in group 0 mean in group 1
##       10.077652        9.829168
```

3. The variables Z1 through Z10 are binary covariates. Change them so they are correctly coded in r.

```r
library(dplyr)

df <- df |>
  dplyr::mutate(
    Z1 = factor(Z1),
    Z2 = factor(Z2),
    Z3 = factor(Z3),
    Z4 = factor(Z4),
    Z5 = factor(Z5),
    Z6 = factor(Z6),
    Z7 = factor(Z7),
    Z8 = factor(Z8),
    Z9 = factor(Z9),
    Z10 = factor(Z10)
  )
```

4. Test to see if subgrouping by covariate Z1 shows a different conclusion about the treatment and response from above? Test for both the "1" group and one for the "0" group.

```r
library(tidyverse)
sub_a<-df |> filter(Z1==1)

t.test (sub_a$Y~sub_a$X)
```

```
##
##  Welch Two Sample t-test
##
## data:  sub_a$Y by sub_a$X
## t = -1.8498, df = 43.506, p-value = 0.07114
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -4.6890354  0.2015816
## sample estimates:
## mean in group 0 mean in group 1
##        8.981304       11.225031
```

```r
sub_b<-df |> filter(Z1==0)

t.test (sub_b$Y~sub_b$X)
```

```
##
##  Welch Two Sample t-test
##
## data:  sub_b$Y by sub_b$X
## t = 1.4806, df = 942.63, p-value = 0.139
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.1291371  0.9227199
## sample estimates:
## mean in group 0 mean in group 1
##       10.130516        9.733724
```

*There is no difference in the means of the response variable Y between the control and treatment groups when Z1==1 (95% CI:-4.6890354, 0.2015816).

There is no difference in the means of the response variable Y between the control and treatment groups when Z1==0 (95% CI:-0.1291371 0.9227199).

5. Test to see if subgrouping by covariate Z2 has shows a different conclusion about the treatment and response from above? Test for both the "1" group and one for the "0" group.

```
library(tidyverse)
sub_c<-df |> filter(Z2==1)

t.test (sub_c$Y~sub_c$X)
```

```
##
##  Welch Two Sample t-test
##
## data:  sub_c$Y by sub_c$X
## t = -1.1995, df = 110.94, p-value = 0.2329
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.2722362  0.5586687
## sample estimates:
## mean in group 0 mean in group 1
##        9.641426       10.498210
```

```
sub_d<-df |> filter(Z2==0)

t.test (sub_d$Y~sub_d$X)
```

```
##
##  Welch Two Sample t-test
##
## data:  sub_d$Y by sub_d$X
## t = 1.4099, df = 881.48, p-value = 0.1589
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.1552627  0.9472730
## sample estimates:
## mean in group 0 mean in group 1
##       10.130469        9.734463
```

*There is no difference in the means of the response variable Y between the control and treatment groups when Z2==1 (95% CI:-2.2722362, 0.5586687).

There is no difference in the means of the response variable Y between the control and treatment groups when Z2==0 (95% CI:-0.1552627, 0.9472730).

11. Which subgroups show an effect?

*The following two subgroups showed an effect: sub_g (where it was filtered only to include when Z4==1) and sub_n (where it was filtered only to include when Z7==0).

12. which tests that you did above are causal? Why or why not?

*Each of the tests above are causal because the treatments were randomly assigned in each test.

13. The variables Z11-Z20 are continuous. Let's break down Z11 into three groups, 0<0.2, 0.2<Z11<0.8 and Z11>0.8. Then we will use those groups and preform a t-test for each group. There is code below. Explain what the first three lines of code are doing. Then write your conclusion about the subgroups.

```
A<-(df$Z11< 0.2)
B<- (df$Z11 > 0.2) & (df$Z11 <.8)
C<- (df$Z11 > 0.8)
t.test(df$Y[A]~df$X[A])
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Y[A] by df$X[A]
## t = 0.91659, df = 201.42, p-value = 0.3605
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.6437895  1.7622136
## sample estimates:
## mean in group 0 mean in group 1
##       10.480920        9.921708
```

```
t.test(df$Y[B]~df$X[B])
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Y[B] by df$X[B]
## t = 0.17948, df = 573.7, p-value = 0.8576
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.6060929  0.7280046
## sample estimates:
## mean in group 0 mean in group 1
##        10.06314        10.00219
```

```
t.test(df$Y[C]~df$X[C])
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Y[C] by df$X[C]
## t = 1.2344, df = 178.05, p-value = 0.2187
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.4162889  1.8071430
## sample estimates:
## mean in group 0 mean in group 1
##        9.748981        9.053554
```

*A: (df$Z11< 0.2): This line creates a logical vector A where each element is TRUE only if the corresponding value of Z11 is less than 0.2.

B: This line creates a logical vector B where each element is TRUE only if the corresponding value of Z11 is greater than 0.2 and less than 0.8.

C: This line creates a logical vector C where each element is TRUE only if the corresponding value of Z11 is greater than 0.8.

The tests reveal that there is no difference in the means of the response variable Y between the control and treatment groups within each of the three subgroups defined by Z11 (95% CI:-0.6437895, 1.7622136; -0.6060929 0.7280046; -0.4162889 1.8071430).

14. Your boss, who is an MU grad, calls you into his office. He has checked the source of the referrals to the landing page. He shows you the following:

"So version A does better on each source, but does worse overall? What is the meaning of this? Should we implement version A or version B?". How do you answer your boss?

*This is an example of Simpson's paradox, where the results of an analysis of aggregated data are the reverse of subgroup results. This happens when the subgroups are different sizes and lurking variables related to both the treatment and response exist. Knowing this, I would lean towards recommending version B. Implementing version B allows for a more holistic approach that considers the overall performance across all sources, potentially avoiding the pitfalls of Simpson's Paradox and providing a more reliable indication of performance.

```
df=readr::read_csv("snacks.csv")
df$City=as.factor(df$City)
df$Campaign=as.factor(df$Campaign)
df$Description=as.factor(df$Description)
```

A company wishes to test the market for a new snack. The factor of interest is the advertising campaign. This is taking place in Europe so the product description on the package is in four different languages. Four types of campaigns are considered, and four different descriptions are considered. The various test campaigns were conducted in four cities, 1, 2, 3, and 4. Sales in thousands of dollars were recorded. See page 187 in your text for a description of the factor levels. The data is contained in the "snacks.csv" file. Note, the most common use of an LSD is with two blocking factors, but this does not necessarily have to be the case.

1. What are the factor(s)?

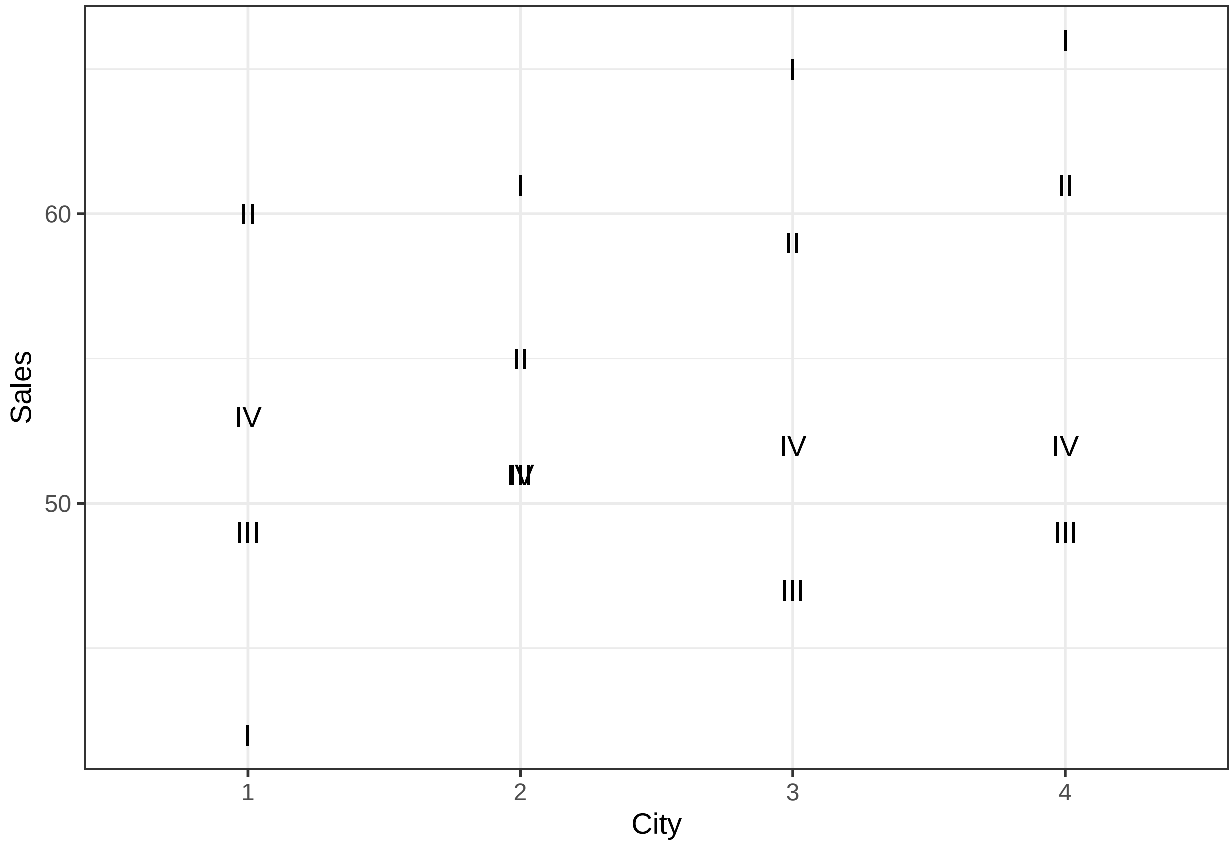
- The advertising campaign and product description

2. What are the blocking factor(s)?

- The city

3. Visualize the results of this design. Hint: look at the Designing Tests for Small Samples Notes

```
library(tidyverse)

ggplot(df, aes(x=City, y=Sales, group=Campaign))+geom_text(aes(label=Campaign))+theme_bw()
```



4. Run the analysis to check and see if the advertisements had any effect on the sales. Make sure the degrees of freedom are correct! You do not have to do the follow up analysis here.

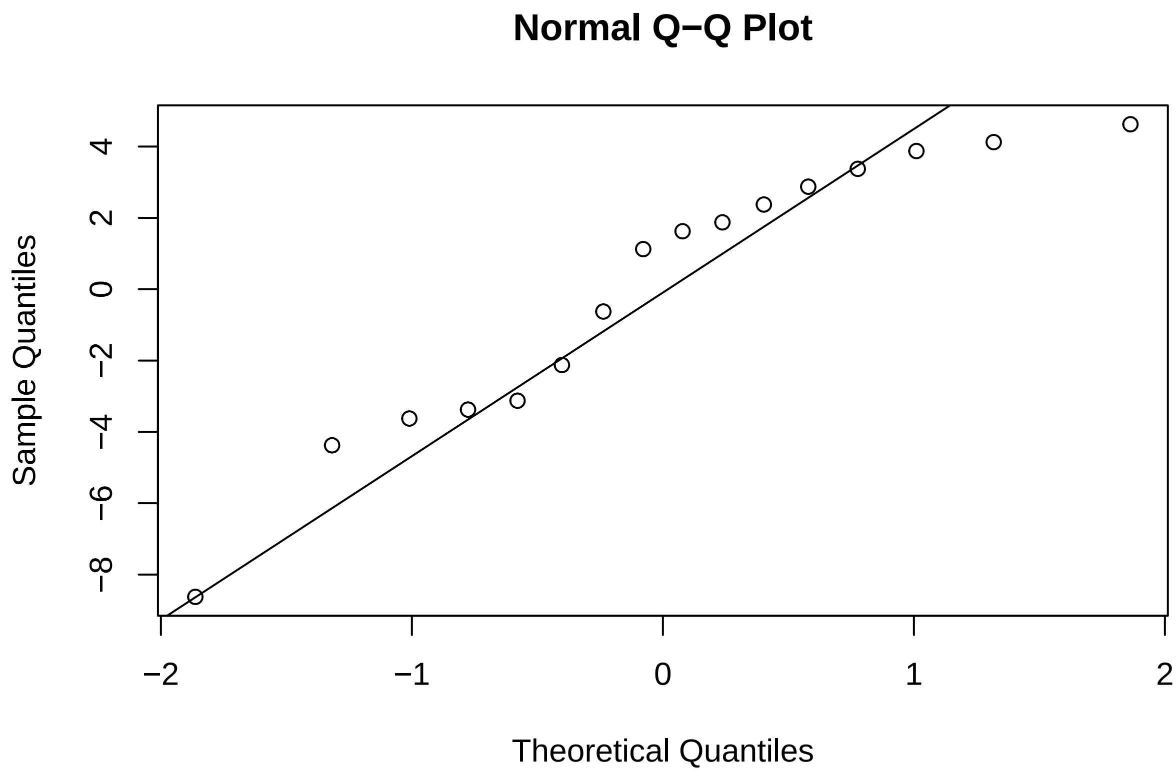
```
mod=aov(Sales~Campaign+City+Description, data=df)
summary(mod)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Campaign     3 282.19   94.06   2.579  0.149
## City         3  80.19   26.73   0.733  0.569
## Description  3 108.69   36.23   0.993  0.457
## Residuals    6 218.88   36.48
```

- The type of advertising campaign does not have an effect on sales (p value = 0.149).
- The degrees of freedom in this analysis are correct, as city needed to be coded as a factor.

5. Check the assumption that the residuals are normally distributed. Make the plot and comment on it.

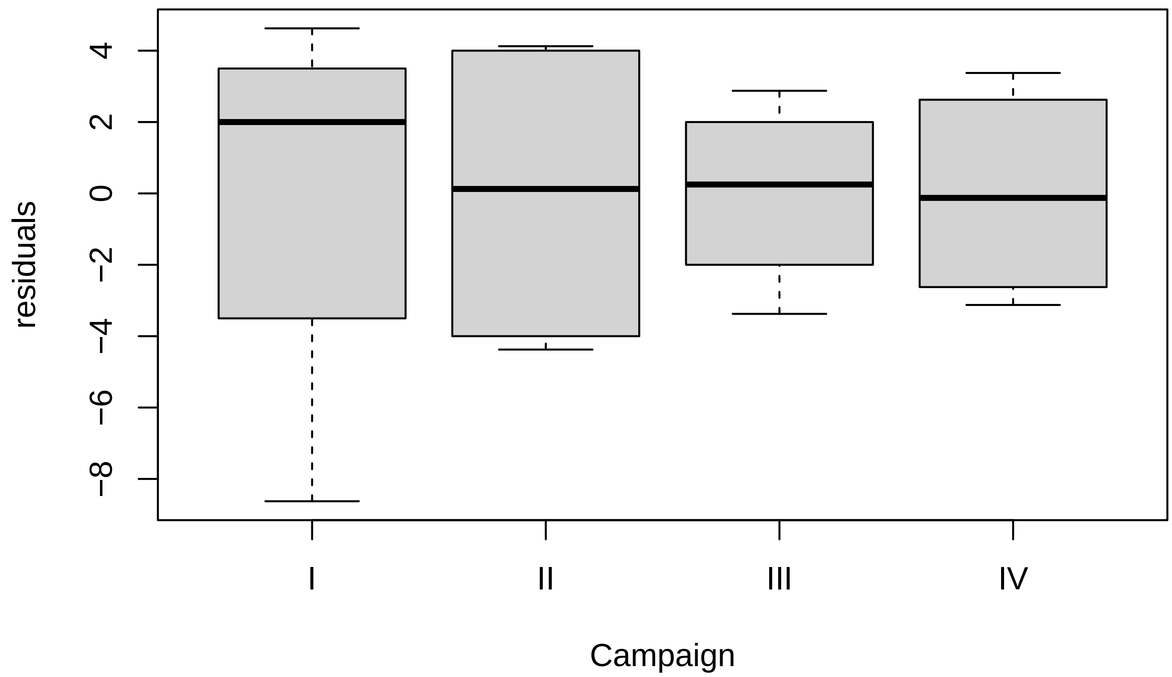
```
qqnorm(mod$residuals)
qqline(mod$residuals)
```



- Considering the few samples in the dataset, this plot shows the points relatively following the diagonal line, which means the residuals are normally distributed. Therefore, the assumption that the probability distribution of epsilon is normally distributed is supported.

6. Check to see if the residuals have a constant variance. You only have to look at the treatment factor. Make the plot and comment on it.

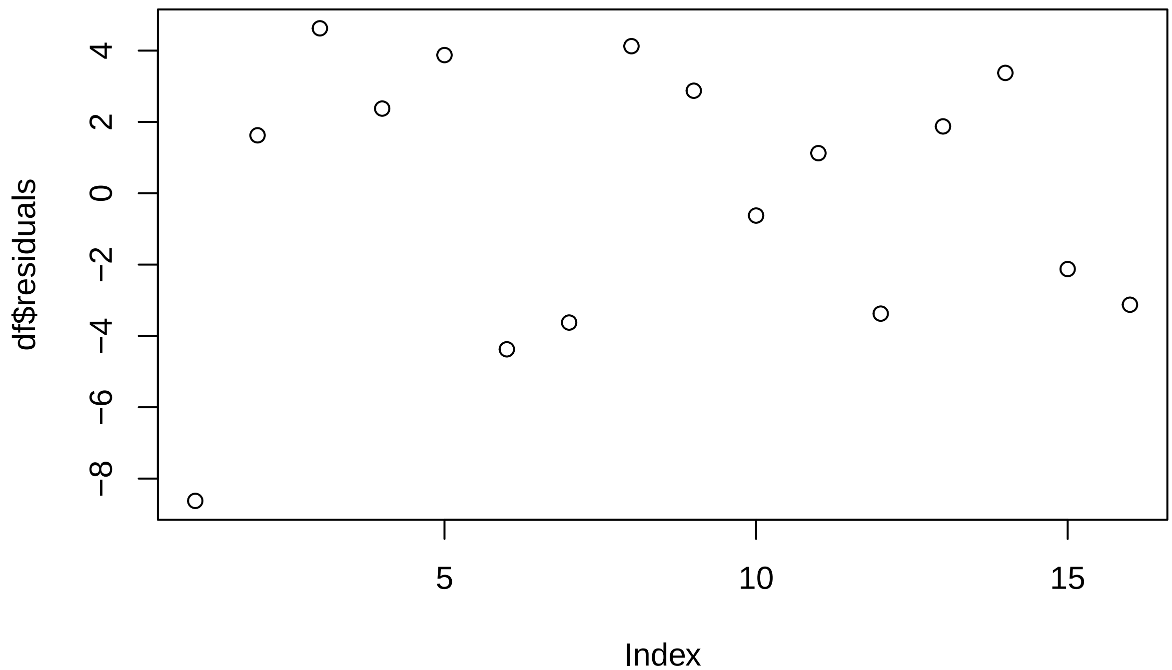
```
df$residuals=mod$residuals
boxplot(residuals~Campaign, data=df)
```



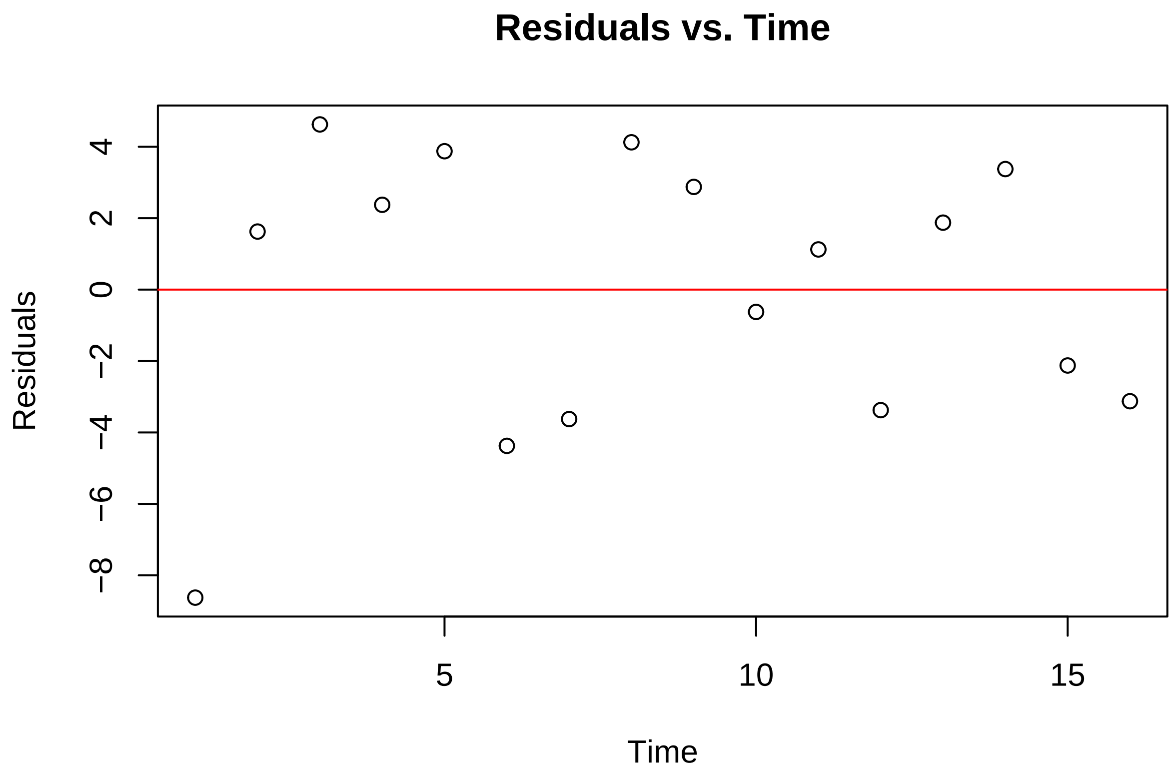
- The spread of the residuals is relatively consistent across each level of the advertising campaign, with no obvious patterns or trends. Therefore, the assumption of constant variance is met.

7. Assuming that the data is listed in the order in which is collected, plot the residuals vs. time to see if there is a trend. Make the plot and comment on it.

```
plot(df$residuals)
```




```
df <- df |>
  dplyr::mutate(Time = 1:nrow(df))

df <- df |>
  dplyr::mutate(Residuals = residuals(mod))

plot(df$Time, df$Residuals, xlab = "Time", ylab = "Residuals", main = "Residuals vs. Time")
abline(h = 0, col = "red")  # Add a horizontal line at y=0 for reference
```



- The residuals appear to be randomly scattered around the horizontal line at y=0 without any systematic pattern, which suggests that there is no time-related trend affecting the residuals.

8. If appropriate run a follow up TukeyHSD test on the treatment. If not appropriate state why.

- Considering the results of the ANOVA test ran above, where we found that the type of advertising campaign does not have an effect on sales (p value = 0.149), a follow up TukeyHSD test is not appropriate. Doing so could lead to misinterpretations or incorrect conclusions.

Miami is running an experiment attempting to entice more students to eat on campus. They sent out mailers with a different offer to students. One offer was "bring a friend for free" another was "take a to-go bag" and lastly "50% off one meal". They measured the spend on campus dining for each offer. They know that different years (First Year, Second Year, Third Year, Fourth Year+) on-campus eating habits are different, but they would like to make a conclusion about the offer that appeals to all students, regardless of their year in school. The data is contained in the file "miami.csv".

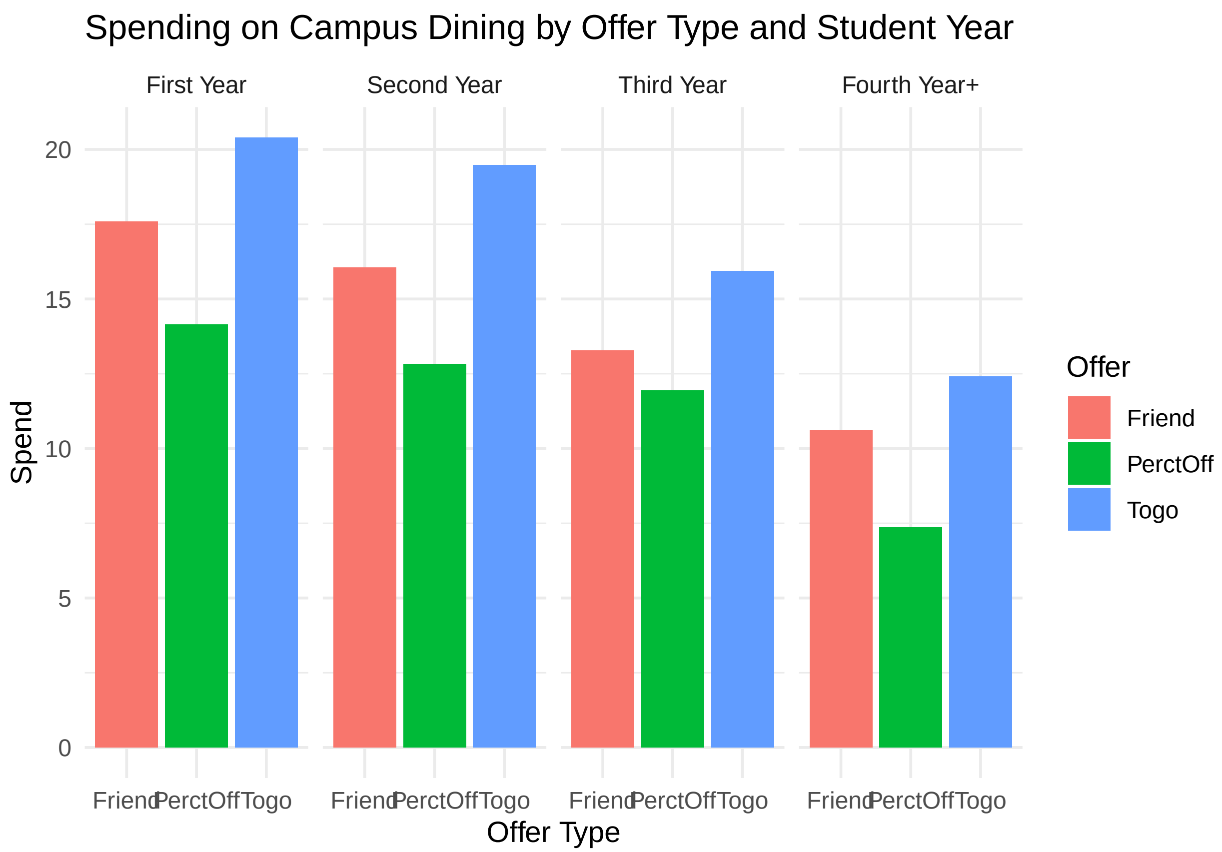
9. Visualize the results of the experiment. Make sure you get the order correct on a graph.

```
df1=readr::read_csv("miami.csv")


library(dplyr)
library(tidyr)
df1_long <- df1 |>
```

```
  pivot_longer(
    cols = c(Friend, Togo, PerctOff),
    names_to = "Offer",
    values_to = "Spend"
  )

df1_long$Year=as.factor(df1_long$Year)
df1_long$Offer=as.factor(df1_long$Offer)
```

use box plots instead, refer to her soultion in GAP Analysis

```
library(tidyverse)
df1_long$Year=factor(df1_long$Year, levels=c("First Year", "Second Year", "Third Year", "Fourth Year+"))
ggplot(df1_long, aes(x = Offer, y = Spend, fill = Offer)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(~Year) +
  labs(title = "Spending on Campus Dining by Offer Type and Student Year",
       x = "Offer Type",
       y = "Spend") +
  theme_minimal()
```



10. Provide a numerical summary of the data that shows the spend by treatment.

```r
library(tidyverse)
df1_long |>  group_by(Offer) |> summarize("average spend"=mean(Spend), "std spend"=sd(Spend),"count"= n
```

```
## # A tibble: 3 x 4
##   Offer    `average spend` `std spend` count
##   <fct>              <dbl>       <dbl> <int>
## 1 Friend              6.50        4.38   800
## 2 PerctOff            3.01        4.12   800
## 3 Togo                3.82        5.76   800
```

```r
mod<-aov(Spend~Offer+Year, data=df1_long)
summary(mod)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Offer          2   5344    2672   148.3 <2e-16 ***
## Year           3  12295    4098   227.5 <2e-16 ***
## Residuals   2394  43132      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(mod)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Spend ~ Offer + Year, data = df1_long)
##
## $Offer
##                       diff        lwr       upr     p adj
## PerctOff-Friend -3.4933041 -3.9910181 -2.995590 0.0000000
## Togo-Friend     -2.6780529 -3.1757668 -2.180339 0.0000000
## Togo-PerctOff    0.8152513  0.3175373  1.312965 0.0003691
##
## $Year
##                               diff       lwr        upr     p adj
## Second Year-First Year    0.192972 -0.437043  0.8229869 0.8601945
## Third Year-First Year    -2.647135 -3.277150 -2.0171197 0.0000000
## Fourth Year+-First Year  -5.357358 -5.987373 -4.7273435 0.0000000
## Third Year-Second Year   -2.840107 -3.470122 -2.2100917 0.0000000
## Fourth Year+-Second Year -5.550330 -6.180345 -4.9203155 0.0000000
## Fourth Year+-Third Year  -2.710224 -3.340239 -2.0802088 0.0000000
```

11. Kroger is testing out different produce layouts in it's 5 pre-designated testing stores. These stores are chosen based on the the demographics of their customers. They will test produce layout configurations 1-5 during 5 months of testing. Produce an appropriate design for this scenario and show it below. Make sure the treatments are in a random order.

```r
library(agricolae)

tmts<-c("Layout 1", "Layout 2", "Layout 3", "Layout 4", "Layout 5")
```

```r
outdesign<-design.lsd(tmts, seed=45, randomization = TRUE) #creates design and lists a bunch of propert

lsd<-outdesign$book #outdesign$book is the actual design matrix from our "outdesign" object


levels(lsd$row)<-c("Month 1", "Month 2", "Month 3", "Month 4", "Month 5") #relabel the rows

levels(lsd$col)<-c("Store 1", "Store 2", "Store 3", "Store 4", "Store 5") #relabel the columns

lsd #notice everything is coded correctly
```

```
##    plots     row     col     tmts
## 1    101 Month 1 Store 1 Layout 3
## 2    102 Month 1 Store 2 Layout 4
## 3    103 Month 1 Store 3 Layout 2
## 4    104 Month 1 Store 4 Layout 5
## 5    105 Month 1 Store 5 Layout 1
## 6    201 Month 2 Store 1 Layout 1
## 7    202 Month 2 Store 2 Layout 2
## 8    203 Month 2 Store 3 Layout 5
## 9    204 Month 2 Store 4 Layout 3
## 10   205 Month 2 Store 5 Layout 4
## 11   301 Month 3 Store 1 Layout 4
## 12   302 Month 3 Store 2 Layout 5
## 13   303 Month 3 Store 3 Layout 3
## 14   304 Month 3 Store 4 Layout 1
## 15   305 Month 3 Store 5 Layout 2
## 16   401 Month 4 Store 1 Layout 2
## 17   402 Month 4 Store 2 Layout 3
## 18   403 Month 4 Store 3 Layout 1
## 19   404 Month 4 Store 4 Layout 4
## 20   405 Month 4 Store 5 Layout 5
## 21   501 Month 5 Store 1 Layout 5
## 22   502 Month 5 Store 2 Layout 1
## 23   503 Month 5 Store 3 Layout 4
## 24   504 Month 5 Store 4 Layout 2
## 25   505 Month 5 Store 5 Layout 3
```

```r
## Let's simulate a response and analyze the design.

lsd$y <- rnorm(nrow(lsd), mean = 25, sd = 5)

# Analyze the design
mod <- aov(y ~ row + col + tmts, data = lsd)
summary(mod)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## row          4  51.07   12.77   0.730  0.589
## col          4  51.67   12.92   0.738  0.584
## tmts         4 136.22   34.06   1.947  0.167
## Residuals   12 209.91   17.49
```