Contents

1 Packages

```
package <- c("readxl","tidyverse","ggplot2","dplyr","MASS")
lapply(package, library, character.only = TRUE)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
##
## Attaching package: 'MASS'
##
##
## The following object is masked from 'package:dplyr':
##
##     select
## [[1]]
## [1] "readxl"    "stats"     "graphics"  "grDevices" "utils"     "datasets" 
## [7] "methods"   "base"     
## 
## [[2]]
##  [1] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
##  [7] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "readxl"    "stats"    
## [13] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"     
## 
## [[3]]
##  [1] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
##  [7] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "readxl"    "stats"    
## [13] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"     
## 
## [[4]]
##  [1] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
##  [7] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "readxl"    "stats"    
## [13] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"     
## 
## [[5]]
##  [1] "MASS"      "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"    
##  [7] "readr"     "tidyr"     "tibble"    "ggplot2"   "tidyverse" "readxl"   
## [13] "stats"     "graphics"  "grDevices" "utils"     "datasets"  "methods"  
## [19] "base"
```

```
#lapply(package, install.packages, character.only = TRUE) # uncomment this to hellp in installtion oacp
```

2 Data importation

```
coffee <- read_excel("~/Data/coffee_shop_survey.xlsx")
head(coffee)
```

```
## # A tibble: 6 x 9
##   Customer_ID   Age Gender Visit_Frequency Favorite_Product Satisfaction_Score
##   <chr>       <dbl> <chr>            <dbl> <chr>                         <dbl>
## 1 CUST001        56 Male                 1 Sandwich                          2
## 2 CUST002        46 Male                 2 Sandwich                          1
## 3 CUST003        32 Male                 6 Pastry                            5
## 4 CUST004        60 Female               2 Pastry                            4
## 5 CUST005        25 Female               3 Tea                               2
## 6 CUST006        38 Female               7 Pastry                            1
## # i 3 more variables: `Time_Spent (min)` <dbl>, Loyalty_Member <chr>,
## #   Would_Recommend <chr>
```

```
summary(coffee)
```

```
##  Customer_ID             Age           Gender          Visit_Frequency
##  Length:100         Min.   :18.00   Length:100         Min.   :0.00   
##  Class :character   1st Qu.:30.50   Class :character   1st Qu.:1.00   
##  Mode  :character   Median :41.00   Mode  :character   Median :3.00   
##                      Mean   :40.88                      Mean   :3.51
##                      3rd Qu.:53.25                      3rd Qu.:6.00
##                      Max.   :64.00                      Max.   :7.00
##  Favorite_Product   Satisfaction_Score Time_Spent (min) Loyalty_Member
##  Length:100         Min.   :1.00       Min.   : 5.00    Length:100
##  Class  :character   1st Qu.:2.00       1st Qu.:21.75    Class  :character
##  Mode   :character   Median :3.00       Median :33.50    Mode   :character
##                      Mean   :3.16       Mean   :33.36
##                      3rd Qu.:4.25       3rd Qu.:45.00
##                      Max.   :5.00       Max.   :60.00
##  Would_Recommend
##  Length:100
##  Class  :character
##  Mode   :character
##
##
##
```

3 Descriptive Analysis Questions

3.1 What is the average age of customers visiting the coffee shop?

```
mean(coffee$Age)
```

```
## [1] 40.88
```

3.2 What is the gender distribution among respondents?

```
table(coffee$Gender)
```

```
##
## Female   Male  Other
##     43     49      8
```

3.3 What are the most and least popular products?

```
coffee %>%
  group_by(Favorite_Product) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 4 x 2
##   Favorite_Product count
##   <chr>            <int>
## 1 Tea                 27
## 2 Pastry              25
## 3 Coffee              24
## 4 Sandwich            24
```

The Most popular product is Tea and leas product is Sandwich

4 Relationship & Comparison Questions

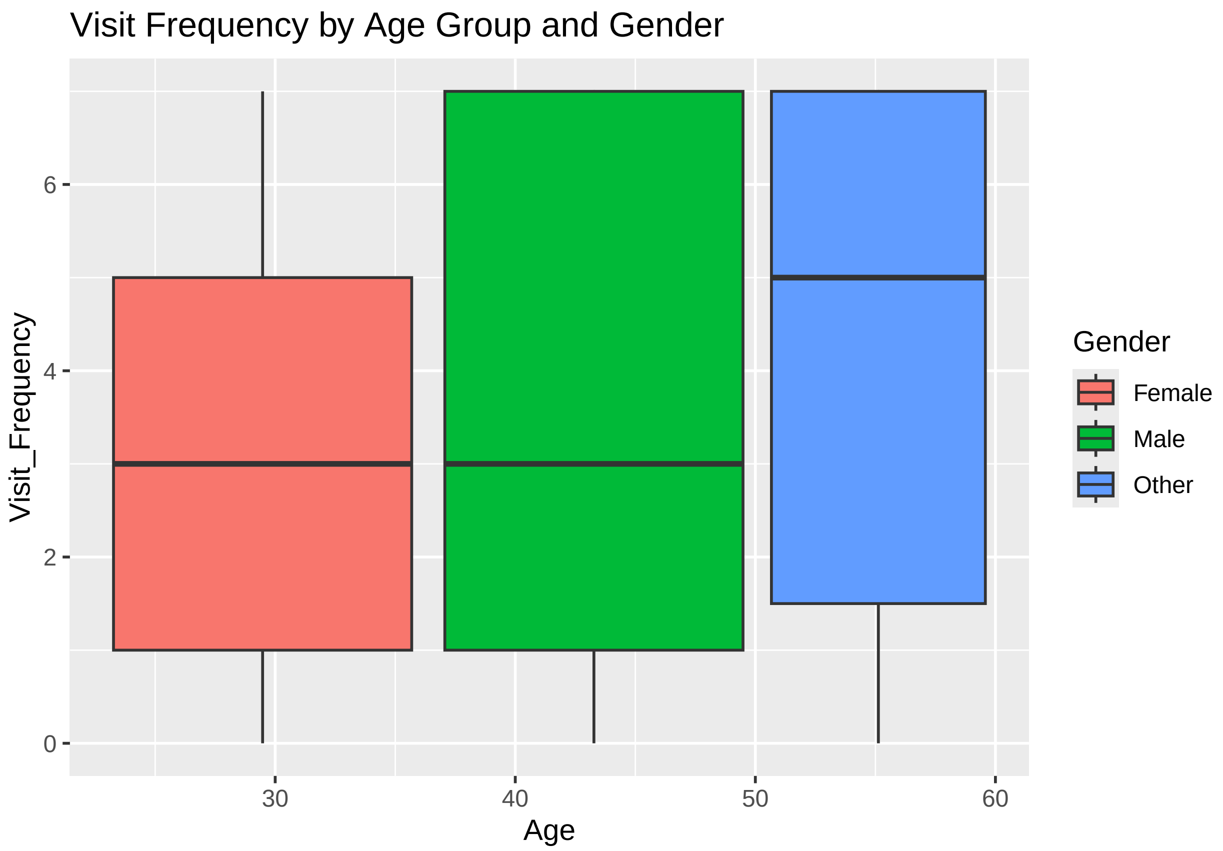
4.1 Does visit frequency differ by gender or age group?

Yes there is an increase in visits by age group with older age groups having a ahigher median in the box plot suggesting higher visits frequencies

```
coffee %>%
  group_by(Gender) %>%
  summarise(visit_mean = mean(Visit_Frequency),
            count = n())
```

```
## # A tibble: 3 x 3
##   Gender visit_mean count
##   <chr>       <dbl> <int>
## 1 Female       3.16    43
## 2 Male         3.71    49
## 3 Other        4.12     8
```

```
ggplot(coffee, aes(x = Age, y = Visit_Frequency, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Visit Frequency by Age Group and Gender")
```



4.2 Is there a relationship between visit frequency and satisfaction score?

The visit frequency returns a P-value less than 0.05, from this we fail to accept the null hyupothesis that there is a relationship between satisfaction score and visit frequency.

```
coffee %>%
  group_by(Satisfaction_Score) %>%
  summarise(count = n(),visit_mean = mean(Visit_Frequency))
```

```
## # A tibble: 5 x 3
##   Satisfaction_Score count visit_mean
##                <dbl> <int>      <dbl>
## 1                  1    16       3.5
## 2                  2    26       3.27
## 3                  3     9       4.22
## 4                  4    24       3.75
## 5                  5    25       3.28
```

```
model <- aov(Visit_Frequency ~ as.factor(Satisfaction_Score), data = coffee)
summary(model)
```

```
##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Satisfaction_Score)  4    8.8   2.195   0.313  0.869
## Residuals                     95  666.2   7.013
```

4.3 Do loyalty members spend more time at the coffee shop than non-members?

From a tabulated outlook loyalty memer spend more time int shop than non-loyal members. From a t.test calcluation we have an extremely low pvalue but greater than 0.05 proving that there is minimal relaionship betewwen loyalty members and time spent in the shop. the high mean from tabulated data could there fore be conluded that its as a rsult of high numbers or another random factor.

```
coffee %>%
  group_by(Loyalty_Member) %>%
  summarise(count = n(), meantime = mean(`Time_Spent (min)`))
```

```
## # A tibble: 2 x 3
##   Loyalty_Member count meantime
##   <chr>          <int>    <dbl>
## 1 No                45     30.5
## 2 Yes               55     35.7
```

```
t.test(`Time_Spent (min)` ~ Loyalty_Member, data = coffee)
```

```
##
##  Welch Two Sample t-test
##
## data:  Time_Spent (min) by Loyalty_Member
## t = -1.7744, df = 95.64, p-value = 0.07917
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -10.8888843   0.6100964
## sample estimates:
##  mean in group No mean in group Yes
##          30.53333          35.67273
```

4.4 Are customers who would recommend the coffee shop more likely to be loyal members?

From the P-value its lesser than 0.05 hence : There is a significant association between loyalty membership and whether a customer would recommend the coffee shop.

```
coffee %>%
  group_by(Loyalty_Member,Would_Recommend) %>%
  count(Loyalty_Member,Would_Recommend)
```

```
## # A tibble: 4 x 3
## # Groups:   Loyalty_Member, Would_Recommend [4]
##   Loyalty_Member Would_Recommend     n
##   <chr>          <chr>           <int>
## 1 No             No                  7
## 2 No             Yes                38
## 3 Yes            No                 20
## 4 Yes            Yes                35
```

```
# Remove rows with NA in either Recommend or Loyalty_Member
cleaned_data <- coffee %>%
  filter(!is.na(Would_Recommend), !is.na(Loyalty_Member))

# Create the contingency table
table_data <- table(cleaned_data$Would_Recommend, cleaned_data$Loyalty_Member)

# Add meaningful row and column names
dimnames(table_data) <- list(
  "Recommendation" = c("Would Not Recommend", "Would Recommend"),
  "Loyalty Membership" = c("Not a Loyalty Member", "Loyalty Member")
)

head(table_data)
```

```
##                       Loyalty Membership
## Recommendation         Not a Loyalty Member Loyalty Member
##   Would Not Recommend                     7             20
##   Would Recommend                        38             35
```

```
# Run the chi-squared test
chisq.test(table_data)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_data
## X-squared = 4.4325, df = 1, p-value = 0.03526
```

5 Insightful/Advanced Questions

5.1 What factors (age, loyalty status, product preference) are associated with higher satisfaction scores?

- *No factor (Age, Loyalty Status, Product Preference) is statistically significantly associated with satisfaction score in this sample.*
- *However, there is a positive trend: customers who prefer Pastry, Sandwich, or Tea show higher odds of better satisfaction than those preferring Coffee.*
- *Loyalty Members actually have slightly lower satisfaction, but again, not significantly.*

```
# Model fitting
model <- MASS::polr(as.factor(Satisfaction_Score) ~ Age + Loyalty_Member + Favorite_Product, data = cof

# Summary of model
summary(model)
```

```
## Call:
## MASS::polr(formula = as.factor(Satisfaction_Score) ~ Age + Loyalty_Member +
##     Favorite_Product, data = coffee, Hess = TRUE)
##
## Coefficients:
##                              Value Std. Error  t value
## Age                       0.001169     0.0129  0.09064
## Loyalty_MemberYes        -0.332438     0.3654 -0.90976
## Favorite_ProductPastry    0.605953     0.5099  1.18828
## Favorite_ProductSandwich  0.711631     0.5202  1.36811
## Favorite_ProductTea       0.740303     0.4955  1.49403
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -1.3176  0.6914    -1.9056
## 2|3  0.0414  0.6718     0.0616
## 3|4  0.4125  0.6702     0.6156
## 4|5  1.5010  0.6832     2.1970
##
## Residual Deviance: 306.3518
## AIC: 324.3518
```

5.2 Is there a difference in satisfaction scores between customers who prefer coffee and those who prefer pastries?

```
unique(coffee$Favorite_Product)
```

```
## [1] "Sandwich" "Pastry"   "Tea"      "Coffee"
```

```
coffee_filtered <- coffee %>%
  filter(Favorite_Product %in% c("Coffee", "Pastry"))

t.test(Satisfaction_Score ~ Favorite_Product, data = coffee_filtered)
```

```
##
##  Welch Two Sample t-test
##
## data:  Satisfaction_Score by Favorite_Product
## t = -1.183, df = 46.929, p-value = 0.2427
## alternative hypothesis: true difference in means between group Coffee and group Pastry is not equal t
## 95 percent confidence interval:
##  -1.3232661  0.3432661
## sample estimates:
## mean in group Coffee mean in group Pastry
##                 2.75                 3.24
```