2024-11-13

Contents

1 Question #1The Big Bang Theory

1.1 a. Compute the minimum and the maximum number of viewers.

```
#
viewers <- read_csv("C:/Users/admin/Desktop/MEM/     /   /BigBangTheory.csv",
col_types = cols(
    `Air Date` = col_character(),
    `Viewers (millions)` = col_double()
  ))
#
viewers1 <- viewers %>%
  select(`Viewers (millions)`) %>%
  pull()
#
min_viewers <- min(viewers1,na.rm = TRUE)
max_viewers <- max(viewers1,na.rm = TRUE)
#
cat("The minimum number of viewers:", min_viewers)
```

```
## The minimum number of viewers: 13.3
```

```
cat("The maximum number of viewers:", max_viewers)
```

```
## The maximum number of viewers: 16.5
```

1.2 b. Compute the mean, median, and mode.

```
#
mean_viewers <- mean(viewers1, na.rm = TRUE)
median_viewers <- median(viewers1, na.rm = TRUE)
#
viewers2<-table(viewers1)
viewers3<-which.max(viewers2)
mode_viewers <- as.numeric(names(viewers2)[viewers3])
#
cat("The mean number of viewers:", mean_viewers, "\n")
```

```
## The mean number of viewers: 15.04286
```

```
cat("The median number of viewers:", median_viewers, "\n")
```

```
## The median number of viewers: 15
```

```
cat("The mode of the number of viewers:", mode_viewers, "\n")
```

```
## The mode of the number of viewers: 13.6
```

1.3 c. Compute the first and third quartiles.

```
#
Q1 <- quantile(viewers1, probs = 0.25)
Q3 <- quantile(viewers1, probs = 0.75)
#
cat("the first quartile is:",Q1,"\n")
```

```
## the first quartile is: 14.1
```

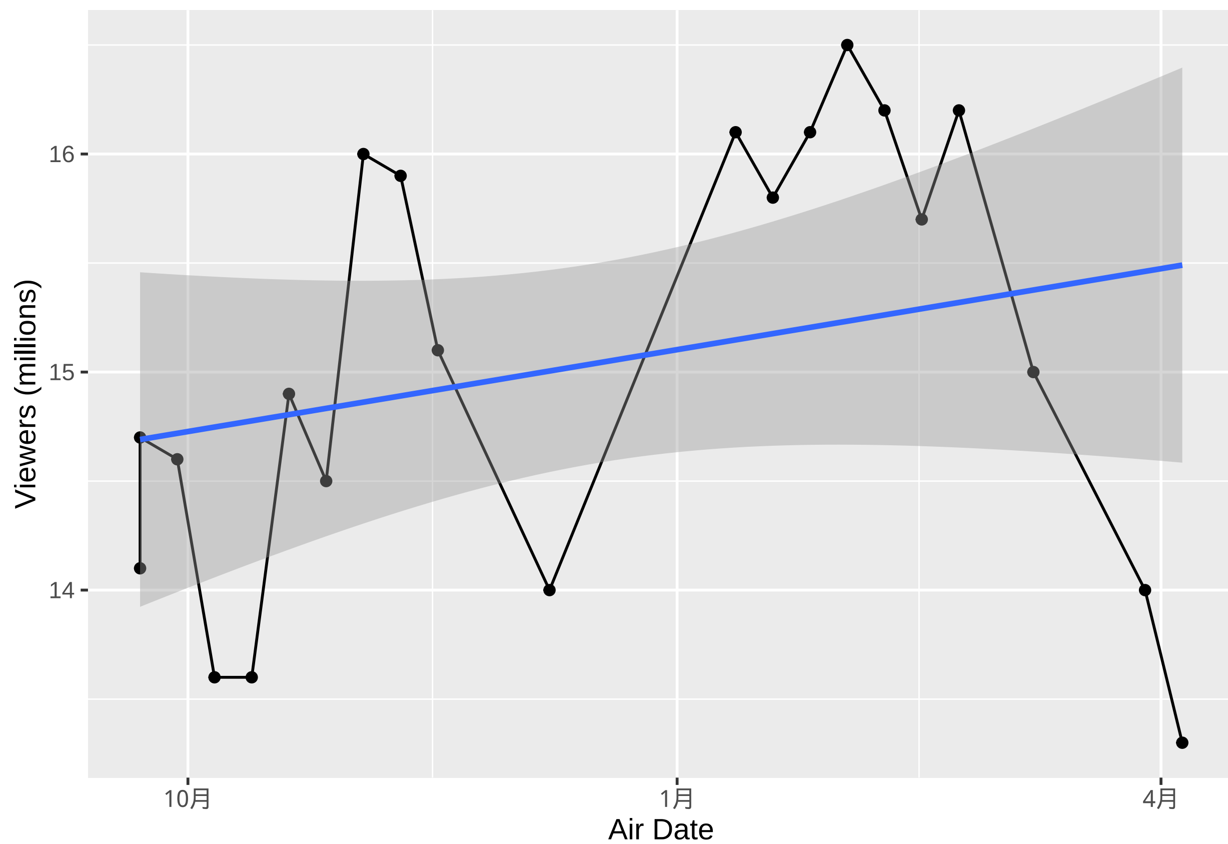
```
cat("the third quartile is:",Q3,"\n")
```

```
## the third quartile is: 16
```

1.4 d.has viewership grown or declined over the 2011–2012 season? Discuss.

```
#
  viewers_a <- read_csv("C:/Users/admin/Desktop/MEM/     /   /BigBangTheory.csv",
  col_types = cols(`Air Date` = col_date(format = "%B %d, %Y"),
                   `Viewers (millions)` = col_double()
    ))
# ggplot
  viewership_grown <- ggplot(data=viewers_a,aes(x = `Air Date` , y = `Viewers (millions)`)) +
    geom_point()+
    geom_line()+
    geom_smooth(method = "lm")
#
  print(viewership_grown)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



2 Question #2:NBAPlayerPts.

2.1 a. Show the frequency distribution.

```
#
NBA_player <- read_csv("C:/Users/admin/Desktop/MEM/     /   /NBAPlayerPts.csv")
```

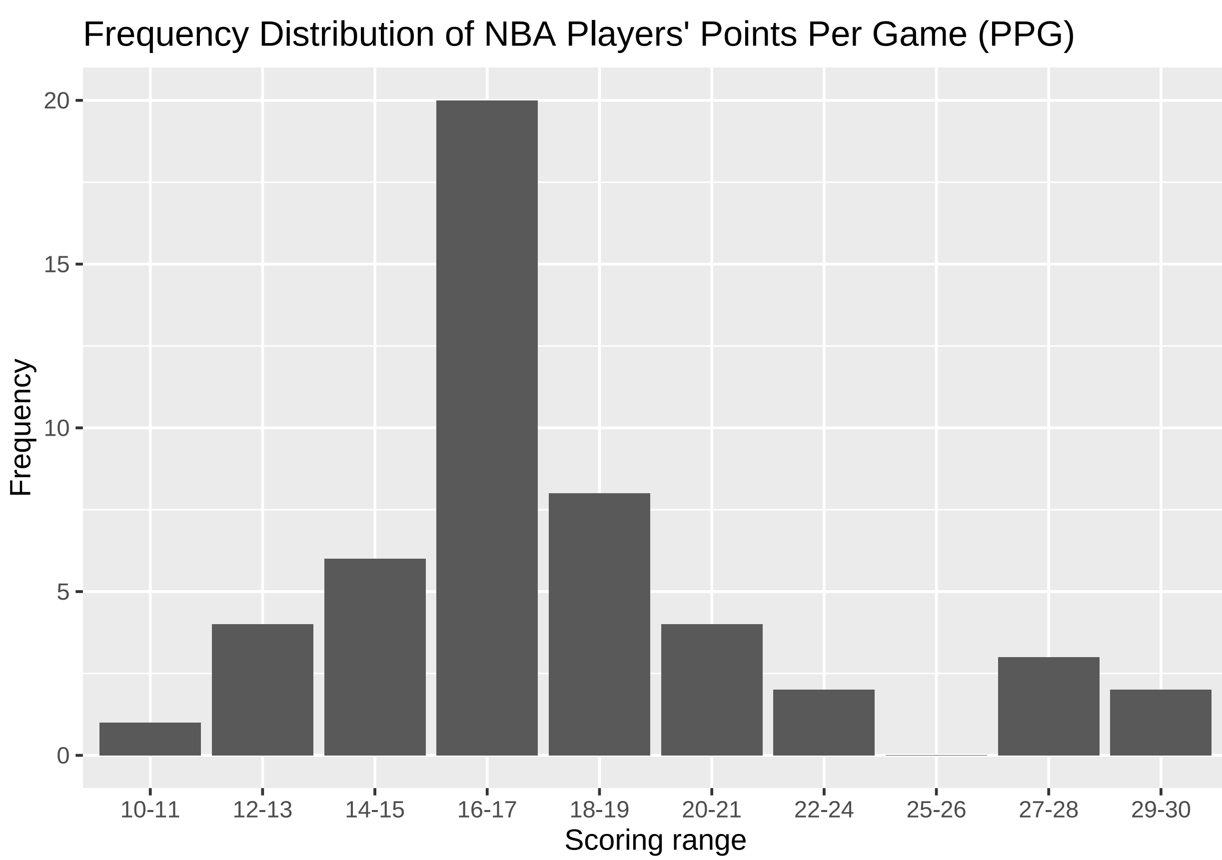
```
## Rows: 50 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Player
## dbl (2): Rank, PPG
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#  PPG     NA
if (!"PPG" %in% names(NBA_player) || any(is.na(NBA_player$PPG))) {
  stop("PPG don't have NA")
}
#
breaks <- seq(10, 30, by = 2)
```

```r
#  PPG
classified_NBA <- cut(NBA_player$PPG, breaks = breaks, labels = c("10-11", "12-13", "14-15", "16-17", "
NBA_freq_table <- table(classified_NBA)
#
NBA_freq_df <- as.data.frame(NBA_freq_table)
colnames(NBA_freq_df) <- c("PPG_Range", "Frequency")
#
NBA_Freq <- ggplot(data = NBA_freq_df, aes(x = PPG_Range, y = Frequency)) +
  geom_histogram(stat = "identity") +
 labs(title = "Frequency Distribution of NBA Players' Points Per Game (PPG)",
       x = "Scoring range",
       y = "Frequency")
```

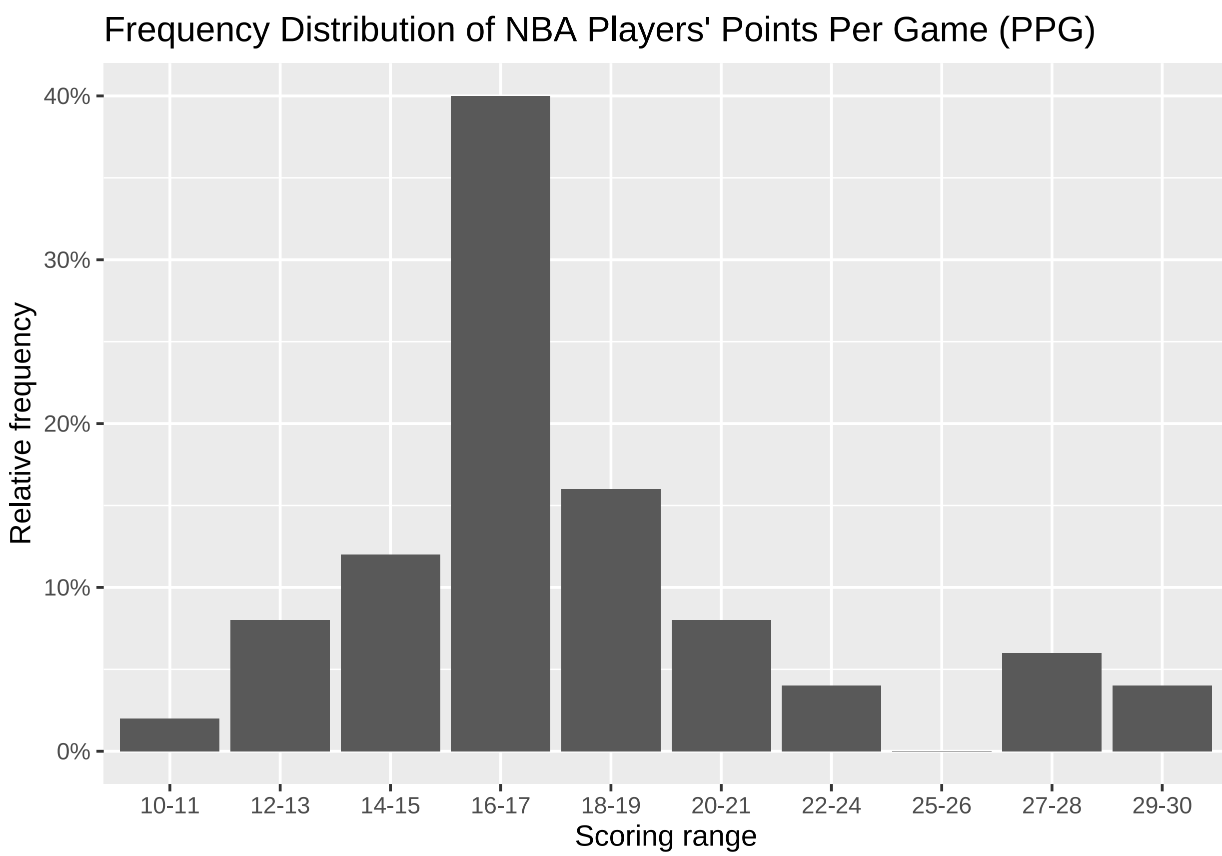
```
## Warning in geom_histogram(stat = "identity"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

```r
#
print(NBA_Freq)
```



2.2 b. Show the relative frequency distribution.

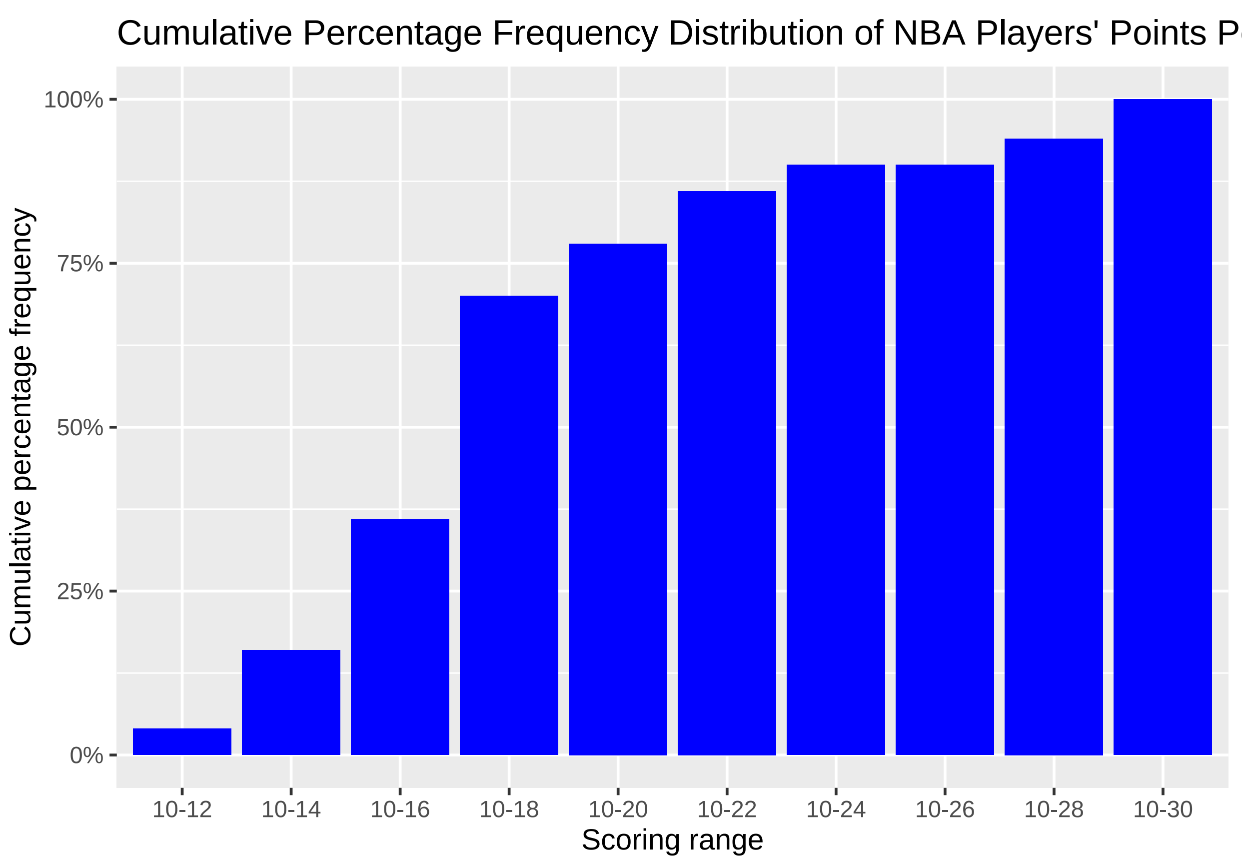
```r
#
NBA_relative_Fre <- NBA_freq_df$Frequency / sum(NBA_freq_df$Frequency)
#
NBA_relative_Fre_bar <- ggplot(data = NBA_freq_df, aes(x = PPG_Range, y = NBA_relative_Fre)) +
  geom_bar(stat = "identity") +
  labs(title = "Frequency Distribution of NBA Players' Points Per Game (PPG)",
       x = "Scoring range",
       y = "Relative frequency") +
  scale_y_continuous(labels = scales::percent_format()) #      y
#
print(NBA_relative_Fre_bar)
```



2.3 c. Show the cumulative percent frequency distribution.

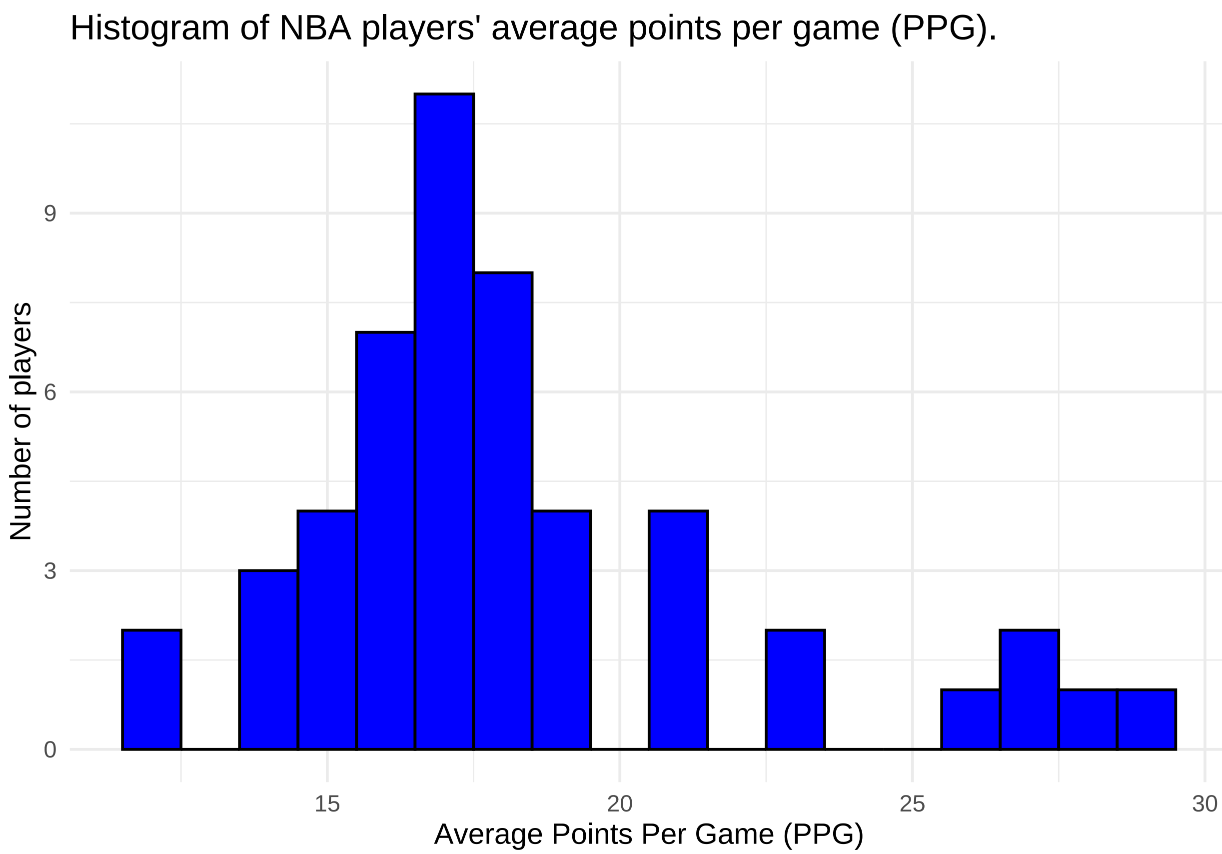
```r
breaks <- c(10)
#   PPG
classified_NBA1<- cut(NBA_player$PPG, breaks = breaks, labels = c("10-12", "10-14", "10-16", "10-18", "
#
NBA_freq_table1 <- table(classified_NBA1)
#
NBA_freq_df1 <- as.data.frame(NBA_freq_table1)
colnames(NBA_freq_df1) <- c("PPG_Range", "Frequency")
#
```

```r
NBA_freq_df1 <- NBA_freq_df1 %>%
  mutate(cumulative_Frequency = cumsum(Frequency),
         cumulative_Percent = cumulative_Frequency / sum(Frequency) )
#
NBA_cumulative_Fre_bar <- ggplot(data = NBA_freq_df1, aes(x = PPG_Range, y = cumulative_Percent)) +
  geom_bar(stat = "identity",fill="blue") +
  labs(title = "Cumulative Percentage Frequency Distribution of NBA Players' Points Per Game",
       x = "Scoring range",
       y = "Cumulative percentage frequency")+
  scale_y_continuous(labels = scales::percent_format()) #      y
#
print(NBA_cumulative_Fre_bar)
```



2.4 d. Develop a histogram for the average number of points scored per game.

```r
#
ggplot(NBA_player, aes(x = PPG)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(title = "Histogram of NBA players' average points per game (PPG).",
       x = "Average Points Per Game (PPG)",
       y = "Number of players") +
  theme_minimal()
```



2.5 e. Do the data appear to be skewed? Explain.

```
#  NBA_player$PPG
skewness_value <- skewness(summary(NBA_player $ PPG))
#
cat("Skewness value ", skewness_value, "\n")
```

```
## Skewness value  0.6687701
```

```
#
if (abs(skewness_value) > 0.5) {
  cat("The data appears to be skewed.\n")
} else {
  cat("The data appears to be symmetrical.\n")
}
```

```
## The data appears to be skewed.
```

2.6 f. What percentage of the players averaged at least 20 points per game?

```
# NBA_player    PPG   20
players_num <- NBA_player[ NBA_player$PPG >= 20,]
#      20
players_20 <- nrow( players_num )
#  NBA_player
players_all<- nrow( NBA_player )
#       20
cat(" The percentage of the players averaged at least 20 points per game ", round(players_20/players_all
```

```
##  The percentage of the players averaged at least 20 points per game  22 %
```

3 Question #3: A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.

3.1 a. How large was the sample used in this survey?

```
A <- 20
B <- 500
n <- (B / A)^2
print(n)
```

```
## [1] 625
```

3.2 b. What is the probability that the point estimate was within ±25 of the population mean?

```
line1 <- -25 / A
line2 <- 25 / A
#          line2     line1
C <- pnorm(line2) - pnorm(line1)
print(C)
```

```
## [1] 0.7887005
```

4 Question #4: Young Professional magazine.

4.1 a. Develop appropriate descriptive statistics to summarize the data.

```
#
Professinal <- read_csv("C:/Users/admin/Desktop/MEM/     /   /Professional.csv")%>%
 rename( age = Age,
         gender = `Gender`,
```

```r
    real_estate = `Real Estate Purchases?`,
    investments = `Value of Investments ($)`,
    num_trans = `Number of Transactions`,
    has_broadband = `Broadband Access?`,
    income = `Household Income ($)`,
    have_children = `Have Children?`) %>%
  select(age:have_children) %>%
  mutate(across(where(is.character), as.factor))
```

```
## New names:
## Rows: 410 Columns: 14
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (5): Gender, Real Estate Purchases?, Broadband Access?, Have Children?, ... dbl
## (4): Age, Value of Investments ($), Number of Transactions, Household In... lgl
## (5): ...9, ...11, ...12, ...13, ...14
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```r
#
skimr::skim(Professinal) %>%
  kable() %>%
  kable_styling()
```

skim_type	skim_variable	n_missing	complete_rate	factor.ordered	factor.n_unique	factor.top_counts	nu
factor	gender	0	1	FALSE	2	Mal: 229, Fem: 181	
factor	real_estate	0	1	FALSE	2	No: 229, Yes: 181	
factor	has_broadband	0	1	FALSE	2	Yes: 256, No: 154	
factor	have_children	0	1	FALSE	2	Yes: 219, No: 191	
numeric	age	0	1	NA	NA	NA	
numeric	investments	0	1	NA	NA	NA	28
numeric	num_trans	0	1	NA	NA	NA	
numeric	income	0	1	NA	NA	NA	74

4.2 b. Develop 95% confidence intervals for the mean age and household income of subscribers.

```r
#
Professinal_sum <- Professinal %>%
  summarise(
    MeanAge = mean(age, na.rm = TRUE),
    SDAge = sd(age, na.rm = TRUE),
    MeanHouseholdIncome = mean(income, na.rm = TRUE),
```

```r
    SDHouseholdIncome = sd(income, na.rm = TRUE)
  )
print(Professinal_sum)
```

```
## # A tibble: 1 x 4
##   MeanAge SDAge MeanHouseholdIncome SDHouseholdIncome
##     <dbl> <dbl>               <dbl>             <dbl>
## 1    30.1  4.02              74460.            34818.
```

```r
#  95%
Age1 <- with(Professinal, t.test(age)$conf.int)
Household_income1 <- with(Professinal, t.test(income)$conf.int)
#
cat("95% confidence interval for the mean age:\n", Age1, "\n")
```

```
## 95% confidence interval for the mean age:
##  29.72153 30.50286
```

```r
cat("95% confidence interval for the mean household income:\n", Household_income1, "\n")
```

```
## 95% confidence interval for the mean household income:
##  71079.26 77839.77
```

4.3 c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.

```r
#
broadband_access <- mean(Professinal$has_broadband == "Yes")
children_have <- mean(Professinal$have_children == "Yes")
#  95%
broadband1 <- prop.test(sum(Professinal$has_broadband == "Yes"), nrow(Professinal), conf.level = 0.95)$c
children1 <- prop.test(sum(Professinal$have_children == "Yes"), nrow(Professinal), conf.level = 0.95)$co
#
cat("95% Confidence Interval for Broadband Access Proportion:", broadband1, "\n")
```

```
## 95% Confidence Interval for Broadband Access Proportion: 0.5753252 0.6710862
```

```r
cat("95% Confidence Interval for Having Children Proportion:", children1, "\n")
```

```
## 95% Confidence Interval for Having Children Proportion: 0.4845521 0.5830908
```

4.4 d. Would Young Professional be a good advertising outlet for online brokers? Justify your conclusion with statistical data.

```r
young_professionals <- subset(Professinal, age >= 25 & age <= 40)
young_professionals$has_broadband <- as.numeric(young_professionals$has_broadband == "Yes")
young_professionals$have_children <- as.numeric(young_professionals$have_children == "Yes")
broadband2 <- mean(young_professionals$has_broadband)
children2 <- mean(young_professionals$have_children)
if (broadband2 > 0.4 & children2 > 0.4) {
  cat("According to statistical data, the young professional demographic has a high proportion of broadb
} else {
  cat("According to statistical data, the young professional demographic may not have a sufficiently hig
}
```

```
## According to statistical data, the young professional demographic has a high proportion of broadband
```

4.5 e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children?

Based on the known information that subscribers have a relatively low average age and a high proportion of them have young children, it can be inferred that this magazine targeted at young professionals is likely an appropriate advertising platform. Since these subscribers, as young parents or guardians, may have a relatively high demand for educational software and computer games for children, the answer is affirmative: this magazine is indeed a good advertising venue for companies selling educational software and computer games for young children.

4.6 f. Comment on the types of articles you believe would be of interest to readers of Young Professional.

The reader base of "Young Professionals" primarily consists of young professionals who typically possess high educational backgrounds, professional qualities, and aspirations for a better life. Based on the characteristics of this group, the following is an analysis of the types of articles they may find interesting: career development, technology and innovation, investment and financial management, lifestyle and health, as well as family and parenting.

5 Question #5:Quality Associate, Inc.

5.1 a. Conduct a hypothesis test for each sample at the .01 level of significance and determine what action, if any, should be taken. Provide the p-value for each test.

```r
#
Quality <- read_csv("C:/Users/admin/Desktop/MEM/    /   /Quality.csv")
```

```
## Rows: 30 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): Sample 1, Sample 2, Sample 3, Sample 4
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#
Quality1 <- as.matrix(Quality)
#
Quality_mean <- mean(Quality1)
#t
results <- list()
for (i in 1:ncol(Quality1)) {
  Quality2 <- Quality1[, i]
 t_test <- t.test(Quality2, mu = Quality_mean)
  results[[i]] <- list(
    p_value = t_test$p.value,
    action = ifelse(t_test$p.value < 0.01, "Take action", "Take no action")
  )
}
#
for (i in 1:length(results)) {
  cat("Sample", i, ":\n")
  cat("p-value:", results[[i]]$p_value, "\n")
  cat("Action:", results[[i]]$action, "\n\n")
}
```

```
## Sample 1 :
## p-value: 0.4508455
## Action: Take no action
##
## Sample 2 :
## p-value: 0.3373273
## Action: Take no action
##
## Sample 3 :
## p-value: 0.01275056
## Action: Take no action
##
## Sample 4 :
## p-value: 0.02090816
## Action: Take no action
```

5.2 b. compute the standard deviation for each of the four samples. does the assumption of .21 for the population standard deviation appear reasonable?

```r
#
sd_sample1 <- sd(Quality$`Sample 1`)
sd_sample2 <- sd(Quality$`Sample 2`)
sd_sample3 <- sd(Quality$`Sample 3`)
sd_sample4 <- sd(Quality$`Sample 4`)
#
cat("Standard deviation of Sample 1:", sd_sample1, "\n")
```

```
## Standard deviation of Sample 1: 0.220356
```

```r
cat("Standard deviation of Sample 2:", sd_sample2, "\n")
```

```
## Standard deviation of Sample 2: 0.220356
```

```r
cat("Standard deviation of Sample 3:", sd_sample3, "\n")
```

```
## Standard deviation of Sample 3: 0.2071706
```

```r
cat("Standard deviation of Sample 4:", sd_sample4, "\n")
```

```
## Standard deviation of Sample 4: 0.206109
```

The standard deviation is not significantly different from 0.21, suggesting that the hypothesis may be reasonable.

5.3 c. compute limits for the sample mean x around =12 such that, as long as a new sample mean is within those limits, the process will be considered to be operating satisfactorily. if x exceeds the upper limit or if x is below the lower limit, corrective action will be taken. these limits are referred to as upper and lower control limits for quality control purposes.

```r
#
n <- length(Quality)
#
Quality1 <- as.matrix(Quality)
#
x_bar <- mean(Quality1)
sigma <- sd(Quality1)
#
alpha <- 0.05
mu <- 12
#  z
z <- qnorm(1 - alpha/2)
sigma <- sd(Quality1)
#
UCL <- mu + z * sigma / sqrt(n)
LCL <- mu - z * sigma / sqrt(n)
#
control_limits <- c(LCL, UCL)
print(control_limits)
```

```
## [1] 11.78136 12.21864
```

5.4 d.discuss the implications of changing the level of significance to a larger value. what mistake or error could increase if the level of significance is increased?

An increase in the probability of a Type I error (false positive) leads to a more lenient threshold for rejecting the null hypothesis (H). This makes the statistical test more sensitive to subtle differences in the sample data. It may imply that we are more willing to bear the risk of rejecting the null hypothesis, resulting in a decrease in our confidence in the outcome.

6 Question #6:Occupancy

6.1 a. Estimate the proportion of units rented during the first week of March 2007 and the first week of March 2008.

```
#
Occupancy <- read_csv("C:/Users/admin/Desktop/MEM/     /   /Occupancy.csv",skip=1)%>%
#
rename(mar_2007 = `March 2007`, mar_2008 = `March 2008`) %>%
#
mutate(across(is.character,as.factor))
```

```
## Rows: 200 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): March 2007, March 2008
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(is.character, as.factor)`.
## Caused by warning:
## ! Use of bare predicate functions was deprecated in tidyselect 1.1.0.
## i Please use wrap predicates in `where()` instead.
##   # Was:
##   data %>% select(is.character)
##
##   # Now:
##   data %>% select(where(is.character))
```

```
#
week_2007 <- sum(Occupancy$mar_2007 == "Yes") / length(Occupancy$mar_2007)
week_2008 <- sum(Occupancy$mar_2008 %in% c("Yes"))/150
#
print(week_2007)
```

```
## [1] 0.35
```

```
print(week_2008)
```

```
## [1] 0.4666667
```

6.2 b. Provide a 95% confidence interval for the difference in proportions.

```
confidence_week <- qnorm(0.975) * sqrt(week_2007*(1-week_2007)/200 + week_2008*(1-week_2008)/150)
print(confidence_week)
```

```
## [1] 0.1036515
```

6.3 c. On the basis of your findings, does it appear March rental rates for 2008 will be up from those a year earlier?

Yes, the interval does not contain zero, which indicates that we should reject the null hypothesis (i.e., the hypothesis that there is no significant difference in rental rates between the two periods). In statistics, if the confidence interval for the difference between two proportions does not contain zero, we generally consider these two proportions to be statistically significantly different. Therefore, based on the result of this confidence interval, we can infer that rental rates in March 2008 have increased compared to those a year earlier.

7 Question #7Air Force Training Program

7.1 a. use appropriate descriptive statistics to summarize the training time data for each method. what similarities or differences do you observe from the sample data?

```
Training <- read_csv("C:/Users/admin/Desktop/MEM/     /   /Training.csv")
```

```
## Rows: 61 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Current, Proposed
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
skimr::skim(Training) %>%
  kable() %>%
  kable_styling()
```

skim_type	skim_variable	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	nu
numeric	Current	0	1	75.06557	3.944907	65	72	
numeric	Proposed	0	1	75.42623	2.506385	69	74	

7.2 b. Comment on any difference between the population means for the two methods. Discuss your findings.

```r
t.test(Training$Current,Training$Proposed)
```

```
##
##  Welch Two Sample t-test
##
## data:  Training$Current and Training$Proposed
## t = -0.60268, df = 101.65, p-value = 0.5481
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5476613  0.8263498
## sample estimates:
## mean of x mean of y
##  75.06557  75.42623
```

7.3 c. compute the standard deviation and variance for each training method. conduct a hypothesis test about the equality of population variances for the two training methods. Discuss your findings.

```r
map(Training,sd)
```

```
## $Current
## [1] 3.944907
##
## $Proposed
## [1] 2.506385
```

```r
map(Training,var)
```

```
## $Current
## [1] 15.5623
##
## $Proposed
## [1] 6.281967
```

```r
var.test(Training$Current,Training$Proposed)
```

```
##
##  F test to compare two variances
##
## data:  Training$Current and Training$Proposed
## F = 2.4773, num df = 60, denom df = 60, p-value = 0.000578
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.486267 4.129135
## sample estimates:
## ratio of variances
##           2.477296
```

7.4 d. what conclusion can you reach about any differences between the two methods? what is your recommendation? explain.

It can only reflect the central tendency of a dataset, but cannot provide information about the degree of dispersion of the data.Standard deviation and variance can provide important information about the degree of dispersion of the data. A larger standard deviation indicates greater differences between data points and the mean; a larger variance indicates a greater degree of deviation of data points from the mean.

7.5 e. can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future?

To determine whether the two programs offer similar or differing amounts of learning, such analysis should be conducted prior to making the final decision to adopt the proposed method. Additionally, gathering user preferences and experiences is also crucial.

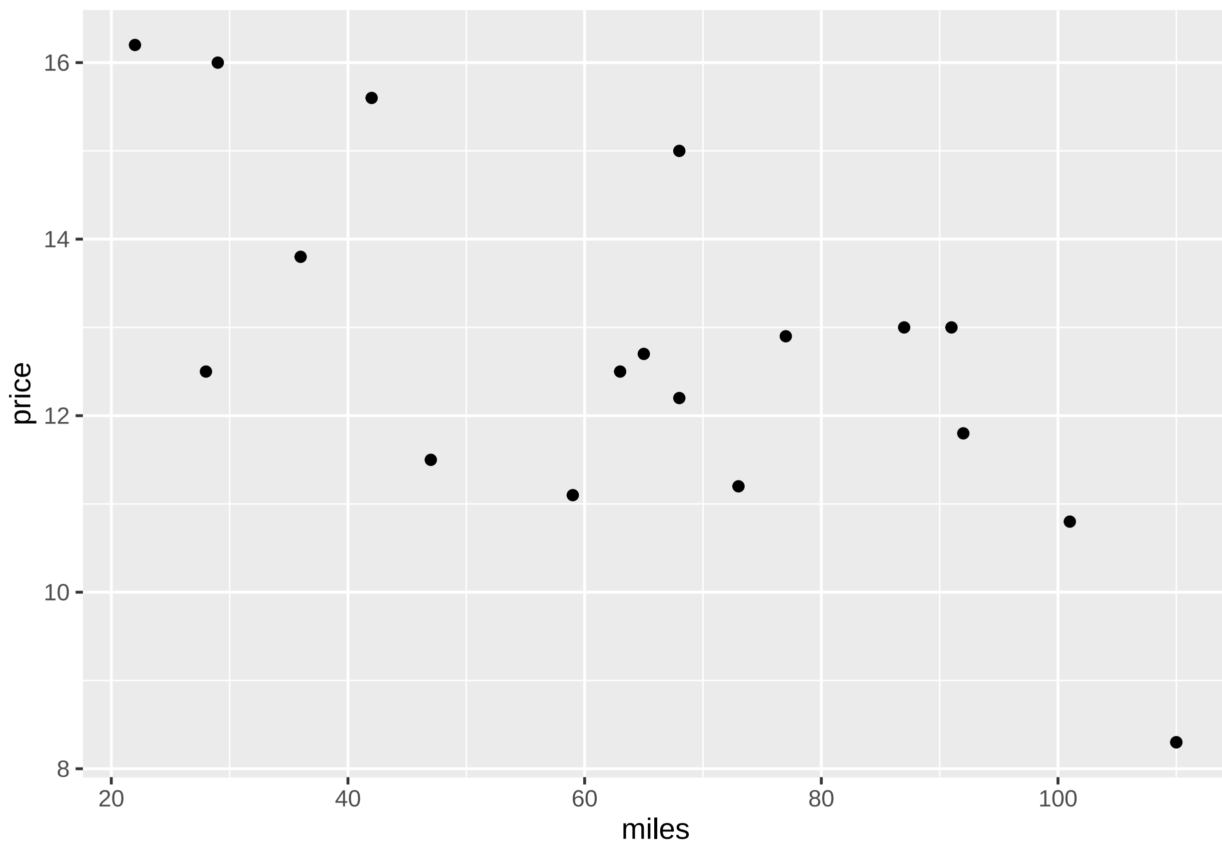
8 Question #8: The Toyota Camry

8.1 a. Develop a scatter diagram with the car mileage on the horizontal axis and the price on the vertical axis.

```
#
Camry <- read_csv("C:/Users/admin/Desktop/MEM/     /   /Camry.csv") %>%
  rename(miles = `Miles (1000s)`,
         price = `Price ($1000s)`)
```

```
## Rows: 19 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Miles (1000s), Price ($1000s)
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#
  Camry %>%
  ggplot() +
  geom_point(aes(miles,price))
```



8.2 b. what does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

The relationship between the two variables can be approximated by a straight line that slopes downwards, indicating a negative correlation as the points on the scatter plot roughly follow this downward-sloping line.

8.3 c. Develop the estimated regression equation that could be used to predict the price ($1000s) given the miles (1000s).

```
lm_camry <- lm(price ~ miles, data = Camry)
summary(lm_camry)
```

```
##
## Call:
## lm(formula = price ~ miles, data = Camry)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32408 -1.34194  0.05055  1.12898  2.52687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 16.46976    0.94876  17.359 2.99e-12 ***
## miles       -0.05877    0.01319  -4.455 0.000348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.541 on 17 degrees of freedom
## Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115
## F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

8.4 d. Test for a significant relationship at the .05 level of significance.

```r
Camry1 <- cor.test(Camry$miles, Camry$price)
print(Camry1)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Camry$miles and Camry$price
## t = -4.4552, df = 17, p-value = 0.0003475
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8910894 -0.4196015
## sample estimates:
##        cor
## -0.7339328
```

8.5 e. Did the estimated regression equation provide a good fit? Explain.

The estimated regression equation provides a statistically significant fit, explaining over half of the price variation, and the model parameters are also statistically significant.

8.6 f. Provide an interpretation for the slope of the estimated regression equation.

The slope represents the average expected change in the dependent variable when the independent variable increases by one unit. If the slope is positive, there exists a positive correlation between the two variables. If the slope is negative, it indicates a negative correlation between the two variables. The larger the absolute value of the slope, the greater the average change in the dependent variable for each unit increase in the independent variable.

8.7 g. Suppose that you are considering purchasing a previously owned 2007 Camry that has been driven 60,000 miles. Using the estimated regression equation developed in part (c), predict the price for this car. Is this the price you would offer the seller.

The predicted price for a 2007 Camry with 60,000 miles, based on the regression equation, is $17,617 while $17,617 is a useful estimate based on mileage, you should consider these additional factors before making an offer to the seller. It may be wise to conduct further research, inspect the car thoroughly, and negotiate based on all relevant information.

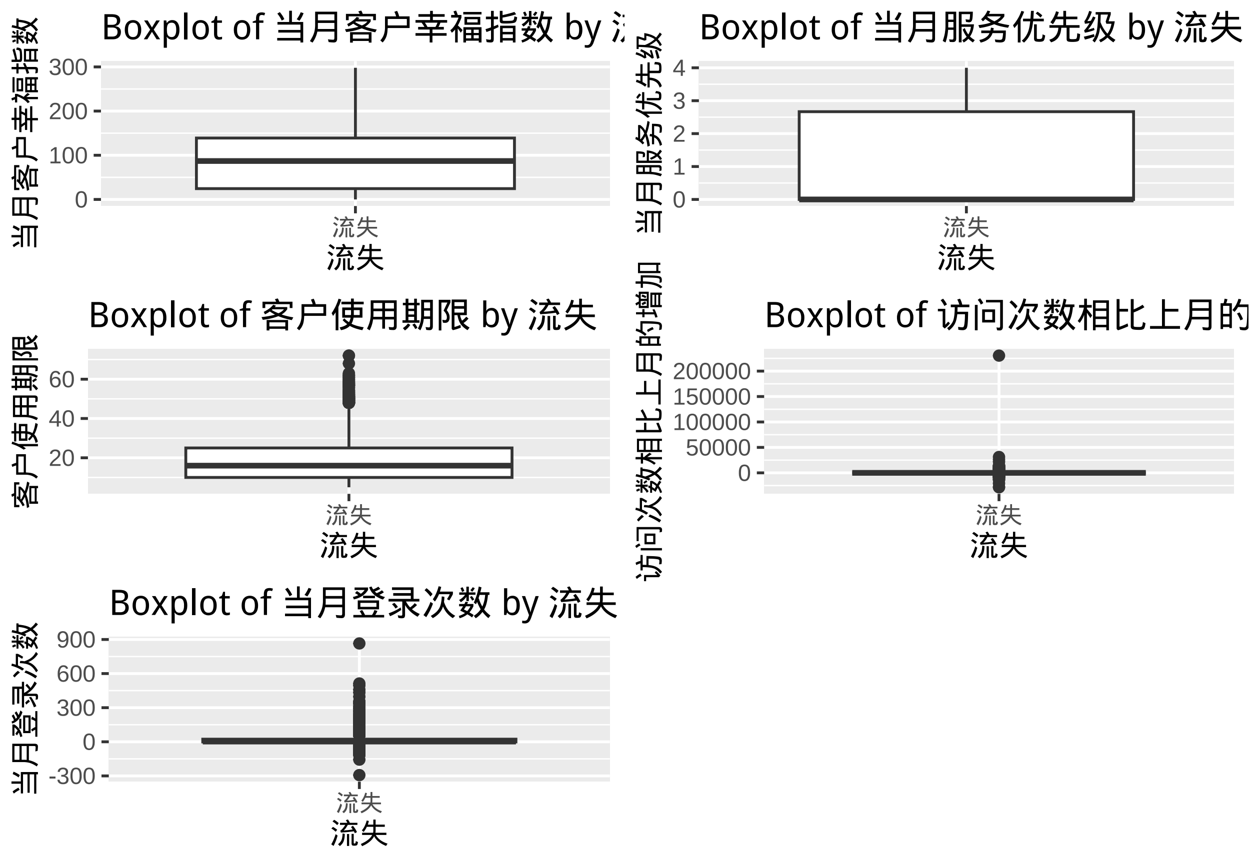
9 Question #9:

9.1 a.

```
#
WE <- read_excel("C:/Users/admin/Desktop/MEM/     /   /WE.xlsx")
  plot_boxplot <- function(data, x_var, y_var) {
  ggplot(data, aes_string(x = factor(x_var), y = y_var)) +
    geom_boxplot() +
    labs(title = paste("Boxplot of", y_var, "by", x_var),
         x = x_var,
         y = y_var)
  }
plot_list <- list(
  plot_boxplot(WE, " ", "     "),
  plot_boxplot(WE, " ", "    "),
  plot_boxplot(WE, " ", "   "),
  plot_boxplot(WE, " ", "      "),
  plot_boxplot(WE, " ", "   ")
)
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
grid.arrange(grobs = plot_list, ncol = 2)
```



9.2 b.

```
#
t.test(     ~   , data = WE)
```

```
##
##  Welch Two Sample t-test
##
## data:        by
## t = 7.6242, df = 369.36, p-value = 2.097e-13
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  18.79956 31.86737
## sample estimates:
## mean in group 0 mean in group 1
##        88.60591        63.27245
```

```
#p-value = 2.097e-13 p   0.05             95%   [18.79956,31.86737]     0
#
t.test(     ~   , data = WE)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:       by
## t = 5.1428, df = 373.13, p-value = 4.381e-07
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.2038355 0.4562009
## sample estimates:
## mean in group 0 mean in group 1
##       0.8295759       0.4995577
```

```
#p-value = 4.381e-07 p   0.05
#
t.test(    ~   , data = WE)
```

```
##
##  Welch Two Sample t-test
##
## data:       by
## t = -2.9811, df = 379.9, p-value = 0.003057
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.5461200 -0.5223121
## sample estimates:
## mean in group 0 mean in group 1
##        18.81873        20.35294
```

```
#p-value = 0.003057 p   0.05
```

9.3 c. ” “ a b

```
#
model <- glm( ~      +      +    , family = binomial(), data = WE)
summary(model)
```

```
##
## Call:
## glm(formula =   ~      +      +
##         , family = binomial(), data = WE)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.771237   0.114435 -24.217  < 2e-16 ***
##      -0.006936   0.001076  -6.444 1.17e-10 ***
##        -0.082358   0.055273  -1.490    0.136
##           0.021643   0.004777   4.531 5.88e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
```

```
## Residual deviance: 2482.7  on 6343  degrees of freedom
## AIC: 2490.7
##
## Number of Fisher Scoring iterations: 6
```

9.4 d. =0 100 ID

```r
#
data_non_churn <- WE %>% filter(  == 1)

#
predictions <- predict(model, newdata = data_non_churn, type = "response")

#
data_non_churn$predictions <- predictions

#
sorted_customers <- data.frame( ID = data_non_churn$ ID,    = data_non_churn$predictions) %>%
  arrange(desc(  ))

#  100  ID
top_100_ids <- sorted_customers %>% head(100) %>% select( ID)

#  100  ID
print(top_100_ids)
```

```
##       ID
## 2       60
## 4       94
## 112   1363
## 7      156
## 117   1488
## 5      105
## 113   1405
## 114   1456
## 176   2296
## 163   2011
## 101   1069
## 195   2653
## 177   2316
## 166   2082
## 145   1823
## 116   1473
## 193   2636
## 159   1987
## 165   2077
## 191   2624
## 110   1303
## 170   2166
## 149   1871
## 169   2120
## 208   2922
```

```
## 167   2084
## 155   1926
## 180   2371
## 209   2928
## 218   3092
## 122   1563
## 185   2521
## 211   2951
## 120   1532
## 134   1711
## 181   2413
## 129   1672
## 143   1803
## 190   2616
## 81     891
## 88     945
## 89     947
## 90     948
## 127   1659
## 207   2902
## 82     896
## 83     904
## 87     938
## 205   2835
## 11     227
## 94     979
## 13     257
## 20     300
## 21     317
## 22     319
## 24     335
## 28     363
## 31     371
## 49     523
## 52     543
## 53     548
## 71     787
## 105   1214
## 138   1760
## 225   3228
## 228   3267
## 229   3312
## 230   3313
## 285   4483
## 287   4500
## 61     640
## 215   3050
## 223   3163
## 226   3235
## 233   3349
## 265   4171
## 37     412
## 174   2212
## 248   3772
```

```
## 284    4482
## 132    1696
## 19      299
## 36      402
## 146    1831
## 97     1021
## 172    2189
## 85      930
## 234    3363
## 239    3569
## 241    3604
## 256    3978
## 263    4156
## 219    3117
## 186    2529
## 271    4273
## 30      369
## 162    2003
## 133    1709
## 141    1782
## 270    4263
```