Project Report 1 - Regression Analysis

Seda Ismail

1 Introduction

1.1 Formulation of the research question and description of the motivation

Background: The analysis aims to explore the relationship between several predictor variables, namely gender of a respondent, age category, desired work hours, and current work in months, with a dependent categorical variable indicating whether the interruption of work is longer than three months. Understanding the factors associated with extended work interruptions can provide valuable insights into workforce dynamics and labor market trends. By examining how these predictor variables contribute to the likelihood of experiencing a work interruption longer than three months, we can gain a deeper understanding of the underlying factors that influence work continuity.

Motivation: The motivation behind this analysis stems from the significance of work interruptions and their implications. Extended periods of work interruptions can have effects on individuals' career trajectories and financial stability. Furthermore, these interruptions can disrupt organizational productivity and workforce planning. By identifying the relationship between the predictor variables and the likelihood of experiencing a work interruption longer than three months, we can uncover potential patterns and associations that contribute to these interruptions.

Research Question: Is there a relationship between gender, age category, desired working hours and months in the current working place among workers with respect to the likelihood of an interruption of more than three months in Austria?

Aim of the study: The aim of this study is to investigate the relationship between gender, age category, desired_working_hours, current_work_in_months and it's effect on the interruption of work.

With the help of a regression analysis to determine whether any of these variables significantly predict interruption of work more than three months significantly.

1.2 Specification and Description of the Variables Used

Four additional variables to the chosen dependent variable are used as predictors. Two of these four independent variables are metric variables and two are categorical variables (one of which is polytomous).

1.2.1 Independent variables

Metric Variables

1) dwstd: Desired total working hours

2) dseitz: Current work since (in months)

Categorical variables

3) bsex: Gender

1 Männlich
2 Weiblich

4) balt5: 5 year age category until 85 (ordinal variable)

1.2.2 Dependent Variable

cdau Interruption of work longer than 3 months c6. (binary/logical variable) Here, the total duration of the interruption (including vacation) is of interest, not how long it has already lasted up to the reference week.

2 Data Collection and Description of the Sample

2.1 Type of data collection; circumstances of implementation (period, etc.).

The microcensus is a type of survey that collects information from a sample of households within a country or region. Type of data collection used for collecting microcensus data are summarized in the "atatmeth-Befragungsmethode" variable as follows:

- 1: Cati: includes those households that were completed in the in-house telephone studio
- 2: F2F: those households that were completed by means of personal interviews conducted by field interviewers.
- 3: Income Call:(s) are those interviews that would originally have been planned as F2F, but are nevertheless conducted in the telephone studio because the household calls in.
- 4: Neutraler Ausfall: Only applies if the person is a single-person household and is neither physically nor mentally able to conduct an interview and no one can actually provide information about this household, it will be treated as a neutral failure.
- 5: Selbstausfülle

Data is collected in the year 2012, which is described in the variable in the dataset as "ajahr-Referenzjahr"

2.2 Description of the data set (sample type, size, characteristics, level of measurement (Skalenniveaus) missing values, etc.).

The microcensus data for Austria refers to a specific survey conducted by Statistik Austria to gather detailed information about various aspects of the population and households. It provides valuable insights into demographic, social, economic, and housing characteristics of the Austrian population for the reference year 2012.

Level of Measurement (Skalenniveaus): The microcensus data set incorporates only metric variables. Where nominal and ordinal variables with categorical information such as gender, marital status, and occupation is coded as integers. Ordinal variables capture ordered categories, for example, educational levels. The rest of the variables, metric variable, are used for quantitative measures like working hours, employment since in months, and household size.

Load microcensus data and inspect

```r
microcensus <- read.csv("mz_2012_testdatensatz_070605.csv")
```

Get sample size

```r
cat("Microcensus dataset has", dim(microcensus)[1], "rows and", dim(microcensus)[2], "columns")
```

Microcensus dataset has 9287 rows and 216 columns

Get the microcensus data, where the dependent variable is defined.

```r
data <- microcensus[microcensus$cdau != -3, ]
```

Remove the row names

```r
rownames(data) <- NULL
```

2.3 Specification and Description of the Variables Used

2.3.1 Metric variables

3) dwstd: Desired total working hours (SPSS: Gewünschte Gesamtarbeitsstunden e9) and is measured on a ratio scale

4) dseitz: Current work since (in months) (SPSS: jetzige Arbeit seit (in Monaten)) and is measured on a ratio scale.

"dwstd" and "dseitz" are measured on a ratio scale because they have a true zero point, which allows them to have more meaningful ratio comparisons.

2.3.2 Categorical variables

3) bsex: Gender (SPSS: Geschlecht) and is a nominal variable after its transformation of the gender coding into categories.

4) balt5: 5-year age categories up to 85 (SPSS: 5-jahres Alterskategorien bis 85) is an ordinal variable after its transformation of the age categories.

5) cdau: (Dependent Variable) Interruption of work longer than 3 months (SPSS: Unterbrechung länger als 3 Monate c6) is a nominal variable and is binary.

Select only variables of study

```
selected_vars <- c("bsex", "balt5", "dwstd", "dseitz", "cdau")
data <- data[selected_vars]
```

The dimension of the dataset of this study is as follows.

```
dim(data)
```

```
## [1] 303   5
```

2.3.3 Handle Missing Values

Convert -3 to NA values

```
data[data == "-3"] <- NA
```

Calculate the total missing values for each variable and print out the results

```
missing_counts <- colSums(is.na(data))
print(missing_counts)
```

```
##   bsex  balt5  dwstd dseitz   cdau
##      0      0     11     11      0
```

Remove the rows with missing values

```
data <- na.omit(data)
```

Again calculate the total missing values and print out the results

```
missing_counts <- colSums(is.na(data))
print(missing_counts)
```

```
##   bsex  balt5  dwstd dseitz   cdau
##      0      0      0      0      0
```

There are no missing values in the data set.

2.3.4 Transform data

Rename columns

```
data <- dplyr::rename(data,
               gender = bsex,
               age_category = balt5,
               desired_working_hours = dwstd,
               current_work_in_months = dseitz,
               interruption_more_than_three_months = cdau)
```

View the structure of the dataset

```
str(data)
```

```
## 'data.frame':    292 obs. of  5 variables:
##  $ gender                             : int  1 2 2 1 1 2 1 1 1 2 ...
##  $ age_category                       : int  7 6 4 9 3 6 6 2 5 5 ...
##  $ desired_working_hours              : num  40 30 38.5 35 40 50 38.5 4 40 20 ...
##  $ current_work_in_months             : int  308 35 39 62 31 159 288 4 2 3 ...
##  $ interruption_more_than_three_months: int  2 2 2 2 2 2 2 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:11] 11 17 36 56 63 64 79 89 94 147 ...
##   ..- attr(*, "names")= chr [1:11] "11" "17" "36" "56" ...
```

- Gender (int) needs to be transformed to factor, where 1 represents "Male" and 2 represents "Female".
- Age category (int) needs to be transformed to factor, where levels 0-15 represent five year age categories.
- Interruption of work more than three months (int) needs to be transformed to factor, where 1 represents "yes" and 2 represents "no".

Transform "gender" variable

```
data[data$gender == 1,]$gender <- "male"
data[data$gender == 2,]$gender <- "female"
data$gender <- as.factor(data$gender)
```

View the frequency table for "age_category" variable

```
table(data$age_category)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12
##  9 21 29 26 31 34 50 47 36  5  1  3
```

As can be seen from the frequency table, there is no value for age_category 0 (0-14), 13 (75-79), 14 (80-85) and 15 (85+).

Transform "age_category" variable. Create new groups, with range of 15.

```
data <-data %>%
  mutate(age_category = case_when(
    age_category %in% c(1, 2, 3) ~ "15-29",
    age_category %in% c(4, 5, 6) ~ "30-44",
    age_category %in% c(7, 8, 9) ~ "45-59",
    age_category %in% c(10, 11, 12) ~ "60-74",
    ))
data$age_category <- as.factor(data$age_category)
```

View the frequency table for "age_category" variable with mutated age categories

```
table(data$age_category)
```

```
##
## 15-29 30-44 45-59 60-74
##    59    91   133     9
```

Transform “interruption_more_than_three_months” variable

```
data[data$interruption_more_than_three_months == 1,]$interruption_more_than_three_months <- "yes"
data[data$interruption_more_than_three_months == 2,]$interruption_more_than_three_months <- "no"
data$interruption_more_than_three_months <- as.factor(data$interruption_more_than_three_months)
```

3 Descriptive Analysis

View the summary of dataset

```
summary(data)
```

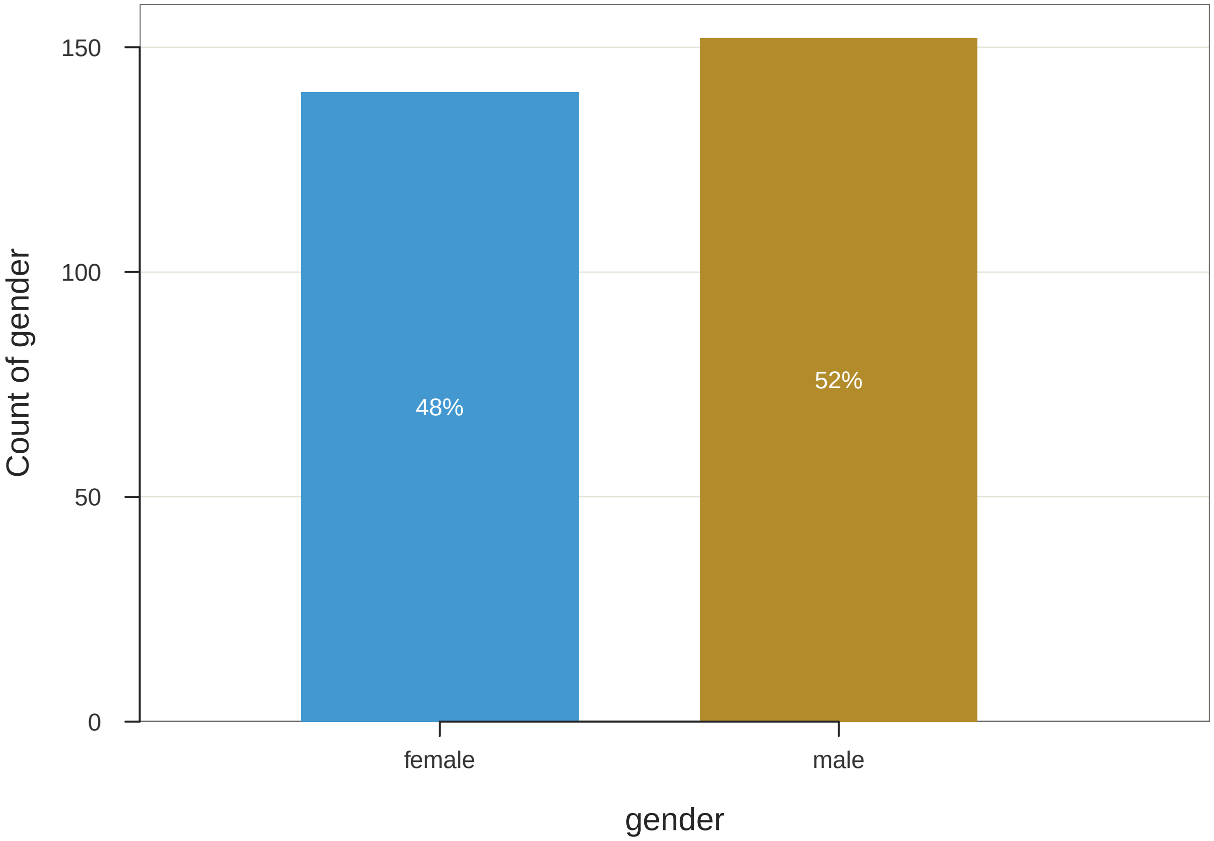
```
##     gender    age_category desired_working_hours current_work_in_months
##  female:140   15-29: 59    Min.   : 4.00         Min.   :  0.00
##  male  :152   30-44: 91    1st Qu.:30.00         1st Qu.: 27.75
##               45-59:133    Median :40.00         Median : 91.00
##               60-74:  9    Mean   :36.01         Mean   :133.73
##                            3rd Qu.:40.00         3rd Qu.:203.75
##                            Max.   :70.00         Max.   :570.00
##  interruption_more_than_three_months
##  no :289
##  yes:  3
##
##
##
##
```

3.1 Univariate Analysis

3.1.1 “gender” variable formerly “bsex” variable of study

Create bar chart for gender

```
BarChart(gender, data=data)
```



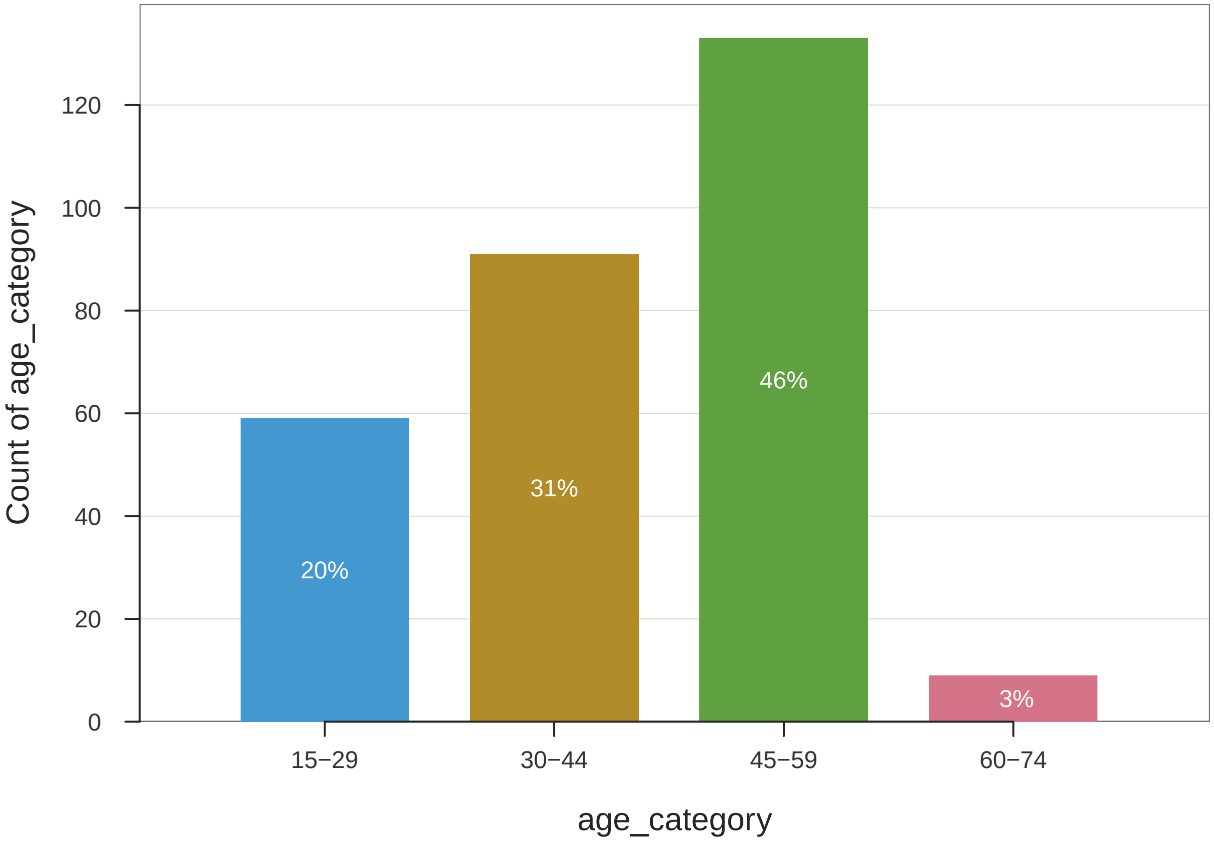
```
## >>> Suggestions
## BarChart(gender, horiz=TRUE)  # horizontal bar chart
## BarChart(gender, fill="reds")  # red bars of varying lightness
## PieChart(gender)  # doughnut (ring) chart
## Plot(gender)  # bubble plot
## Plot(gender, stat="count")  # lollipop plot
##
## --- gender ---
##
## Missing Values: 0
##
##                 female   male     Total
## Frequencies:       140    152       292
## Proportions:     0.479  0.521     1.000
##
## Chi-squared test of null hypothesis of equal probabilities
##   Chisq = 0.493, df = 1, p-value = 0.483
```

- The sample consists of 303 respondents, with 146 females and 157 males.
- The distribution of the sample in terms of gender can be considered as balanced, with slightly more percentage of males with 51.8% than percentage of females with 48.2%.

3.1.2 “age_category” formerly “balt5” variable of study

Create bar chart for age category

```
BarChart(age_category, data=data)
```



```
## >>> Suggestions
## BarChart(age_category, horiz=TRUE)  # horizontal bar chart
## BarChart(age_category, fill="reds")  # red bars of varying lightness
## PieChart(age_category)  # doughnut (ring) chart
## Plot(age_category)  # bubble plot
## Plot(age_category, stat="count")  # lollipop plot
##
## --- age_category ---
##
## Missing Values: 0
##
##                 15-29  30-44  45-59  60-74     Total
## Frequencies:       59     91    133      9       292
## Proportions:    0.202  0.312  0.455  0.031     1.000
##
## Chi-squared test of null hypothesis of equal probabilities
##   Chisq = 112.548, df = 3, p-value = 0.000
```

- The distribution is not uniform across all age groups. The highest proportion is in the 45-59 age group with 46%, followed by the 30-44 age group.
- The 15-29 and 60-74 age groups have lower representations.
- Overall, the sample appears to have a diverse age distribution, with a notable presence of respondents in their pre-retirement years, as well as a smaller representation of younger respondents who are newly entering the labor force.

3.1.3 "desired_working_hours" formerly "dwstd" variable of study

View the summary of the weekly_work_hours variable

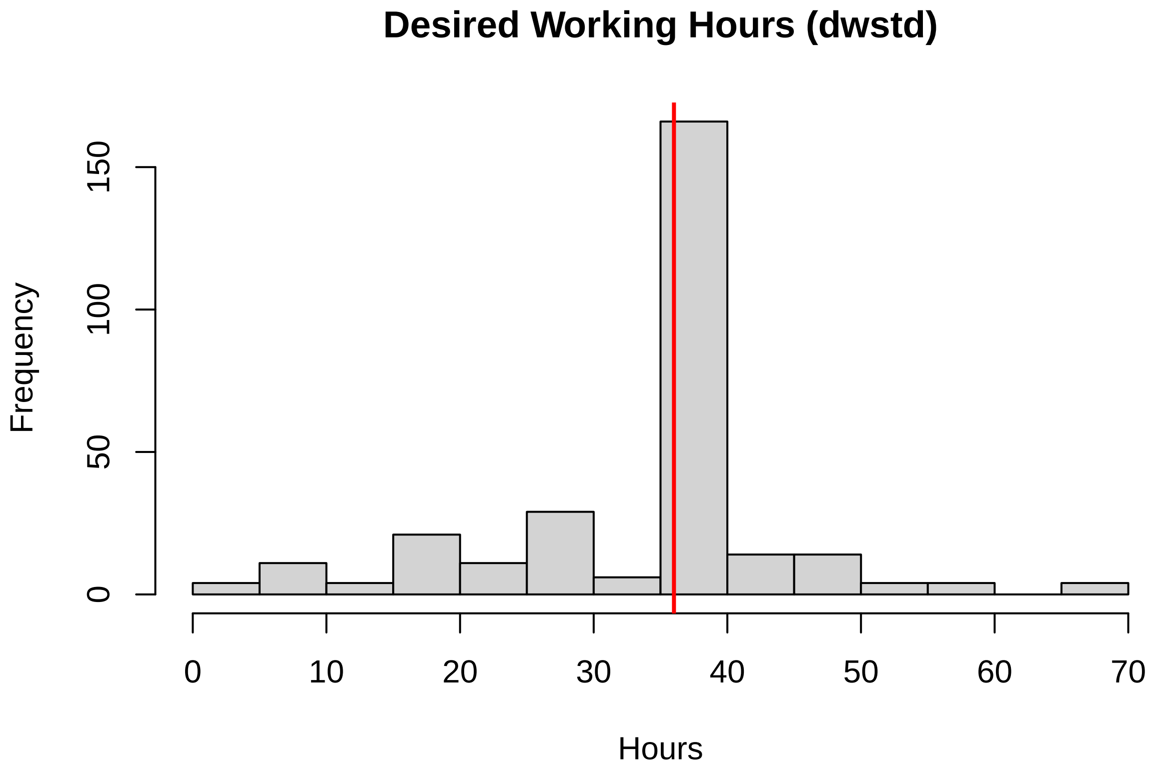
```
summary(data$desired_working_hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   30.00   40.00   36.01   40.00   70.00
```

- The minimum value is 4.00, which is the smallest hours of desired working hours.
- 1st Quartile: 25% of the data falls below 30.00 hours.
- The average desired working hours are approximately 36.17 hours, which might differ after the removel of outliers.
- 3rd Quartile: 75% of the data falls below 40.00 hours.
- The maximum value is 70 hours.
- The 3.Quartile and median is equal.

Create the histogram for desired working hours

```
with(data, hist(desired_working_hours, main="Desired Working Hours (dwstd)", xlab="Hours"))
abline(v = mean(data$desired_working_hours), col = "red", lwd = 2)
```



Desired working hours mean lie around the value of 36.01 and the data follows a leptokurtic curve and peaking in the middle.

Get the highest frequencies of common desired working hours

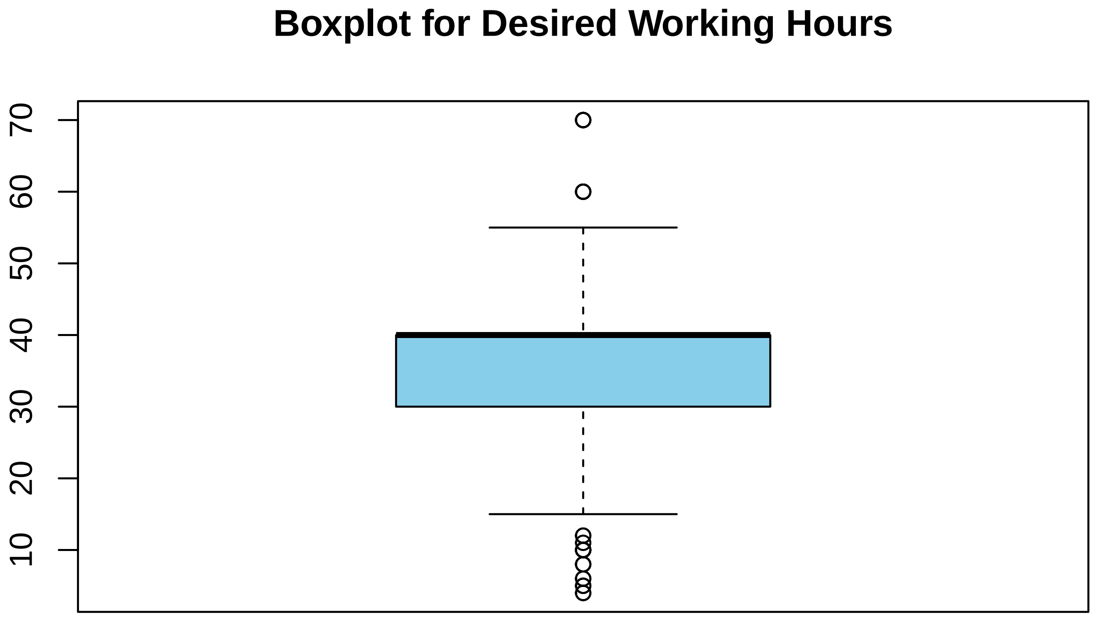
```
tab <- table(data$desired_working_hours)
highest_five <- head(sort(tab, decreasing = TRUE), 5)
highest_five
```

```
##
##   40 38.5   30   20   50
##  115   39   26   19   13
```

Highest desired working hours is 40 hours and second highest is 38.5 followed by 30 and 20 hours.

Create the boxplot for desired working hours

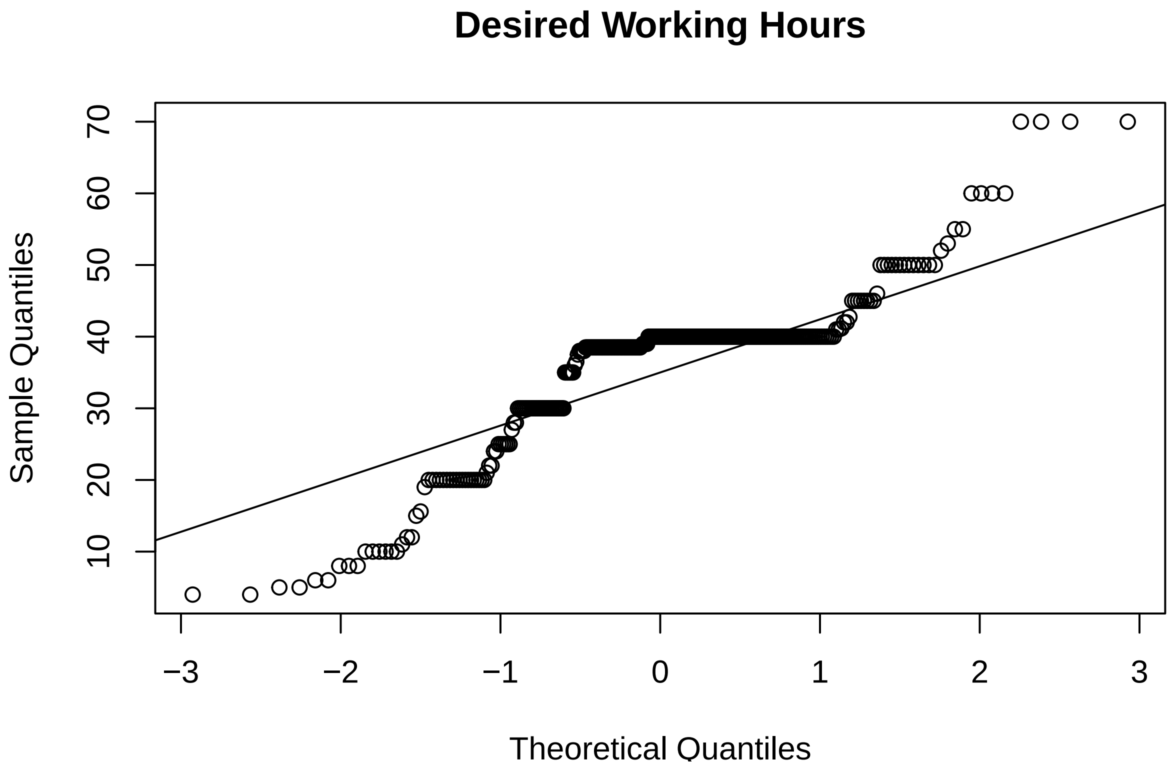
```
boxplot(data$desired_working_hours, main='Boxplot for Desired Working Hours', col='Sky Blue')
```



Boxplot for desired_working_hours show there are multiple outliers for the variable and median is equal to the Q3. The box's length is relatively small with a value of 10. (3.Quartil-1.Quartil) Upper whisker value: 40 + 1.5*10 = 55 Lower whisker value: 30 - 1.5*10) = 15.

Create a QQ plot to compare "desired_working_hours" to a theoretical normal distribution

```
qqnorm(data[, "desired_working_hours"], main = "Desired Working Hours"); qqline(data[, "desired_working_
```



Desired working hours does not follow the normal distribution as it can be seen from the plot.

3.1.4 "current_work_in_months" formerly "dseitz" variable of study

View the summary of the current_work_in_months variable

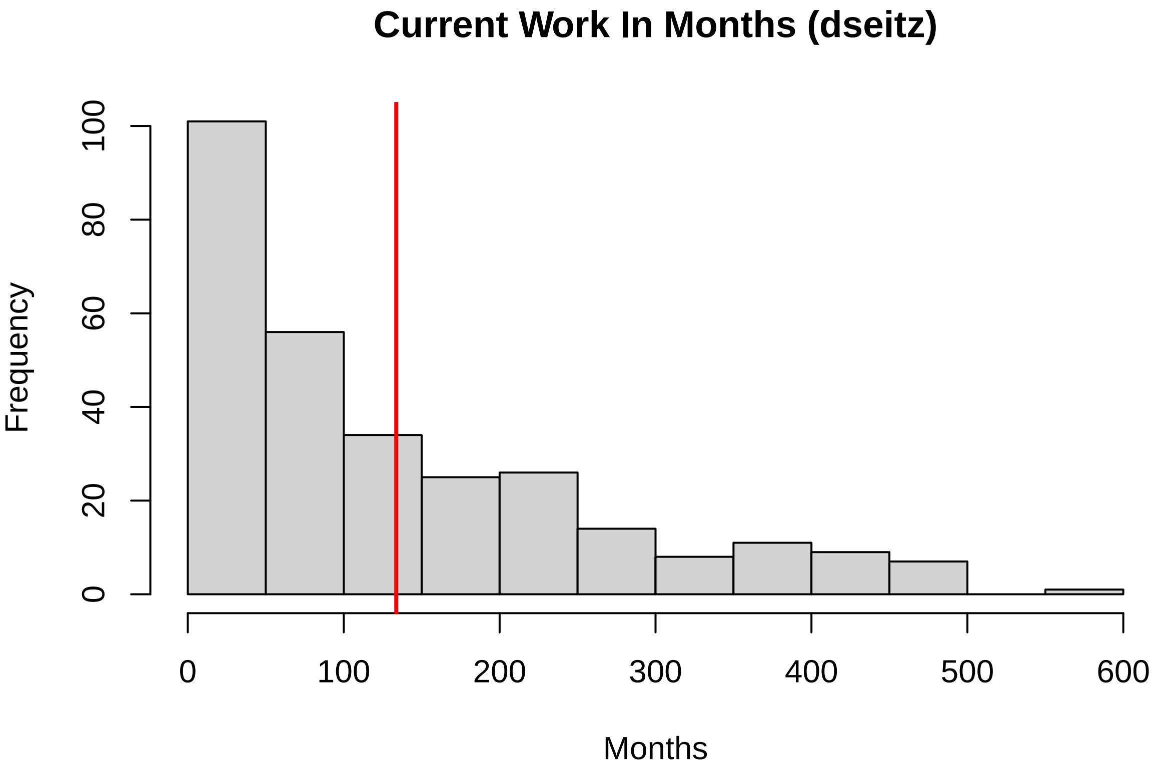
```
summary(data$current_work_in_months)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   27.75   91.00  133.73  203.75  570.00 
```

- The minimum value is 0, which is the smallest hours of current work in months, this might be an outlier or the respondent might have not finished the first months on the new job.
- 1st Quartile: 25% of the data falls below 27 months, which is 2.25 years on the current job.
- The average months on the current work are approximately 133.9 months, nearly 11.15 years on the current job.
- 3rd Quartile: 75% of the data falls below 209 months, nearly 17.41 years on the current job.
- The maximum value is 570 hours, which is exactly 47.5 years. This respondent highly likely to be a person who stayed at the same job over the year, presumably getting closer to the retirement.

Create the histogram for weekly_work_hours

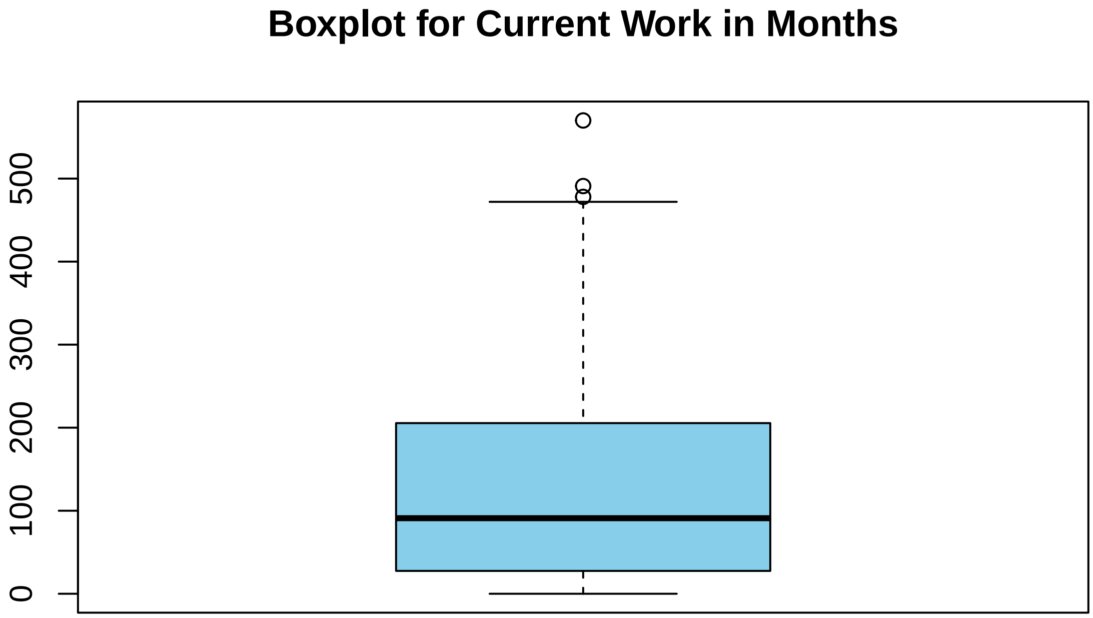
```
with(data, hist(current_work_in_months, main="Current Work In Months (dseitz)", xlab="Months"))
abline(v = mean(data$current_work_in_months), col = "red", lwd = 2)
```



- Mean months on the current work place is 133.73, which is roughly 11 years.
- Current Work in Months is right-skewed meaning that the months are more frequent around the lower end of the x-axis. The distribution extends more to the right (toward larger months) compared to the left (smaller months). Histogram suggests that the majority of data points are clustered at lower values, with a tail extending toward higher values.

Create the boxplot for desired working hours

```
boxplot(data$current_work_in_months, main='Boxplot for Current Work in Months',col='Sky Blue')
```



```
# Boxplot(~ data$current_work_in_months)
```

- Boxplot shows that the outliers are only after the upper extreme (whisker) and that it is right-skewed.
- There are three outliers to be seen from the boxplot. As mentioned before they are likely to be people who stayed at the same job for a long time. The box length is relatively big with a value of 182 (3.Quartil-1.Quartil) 209-27=182 Upper whisker value: 209 + 1.5*182 = 482 Lower whisker value: 0

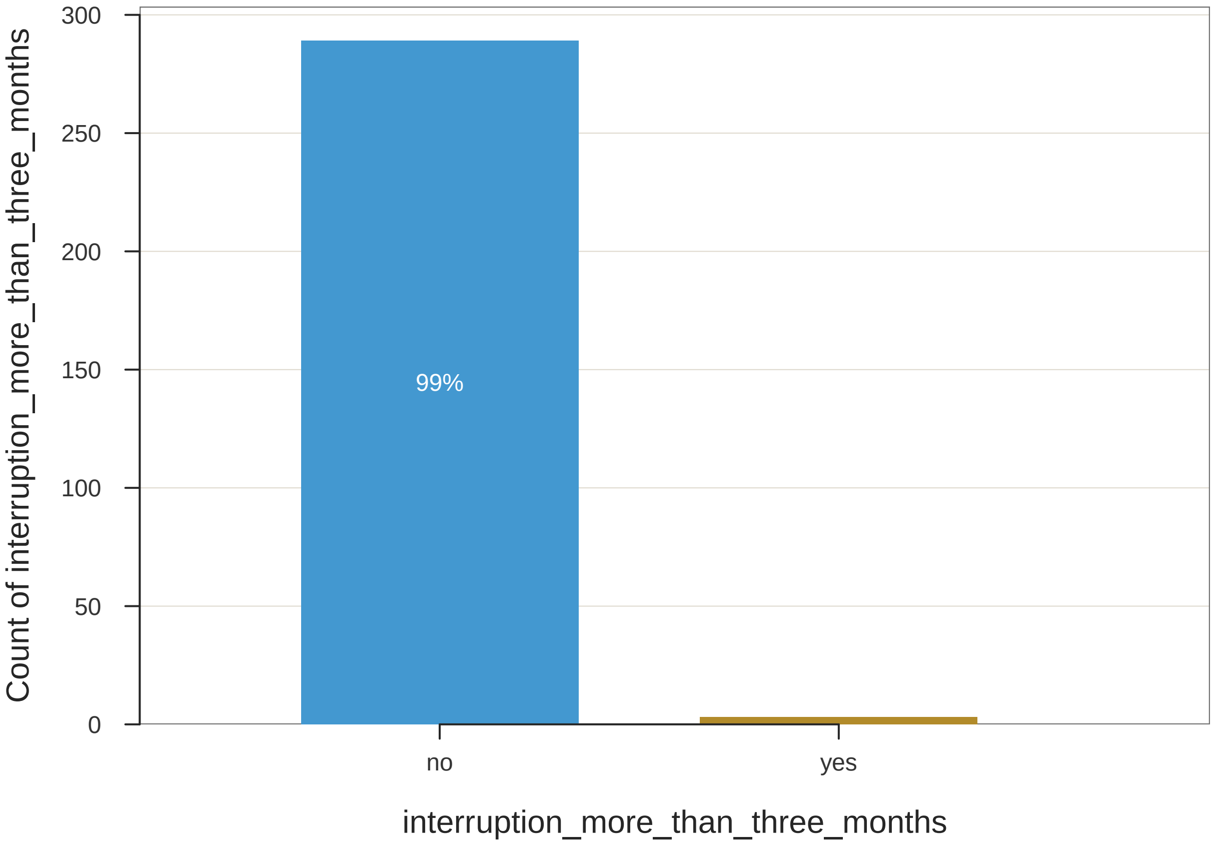
View the frequency table

```
table(data$interruption_more_than_three_months)
```

```
##
##  no yes
## 289   3
```

View the barchart for interruption_more_than_three_months variable

```
BarChart(interruption_more_than_three_months, data=data)
```



```
## >>> Suggestions
## BarChart(interruption_more_than_three_months, horiz=TRUE)  # horizontal bar chart
## BarChart(interruption_more_than_three_months, fill="reds")  # red bars of varying lightness
## PieChart(interruption_more_than_three_months)  # doughnut (ring) chart
## Plot(interruption_more_than_three_months)  # bubble plot
## Plot(interruption_more_than_three_months, stat="count")  # lollipop plot
##
## --- interruption_more_than_three_months ---
##
## Missing Values: 0
##
##                     no    yes     Total
## Frequencies:       289      3       292
## Proportions:     0.990  0.010     1.000
##
## Chi-squared test of null hypothesis of equal probabilities
##   Chisq = 280.123, df = 1, p-value = 0.000
```

There are 3 respondents, who had more than 3 months of interruption from work and the rest of the respondent had not have interruption from work more than 3 months.

3.2 Bivariate Analysis

3.2.1 "gender" and "interruption_more_than_three_months"

Create the contingency table for the variables gender and interruption_more_than_three_months

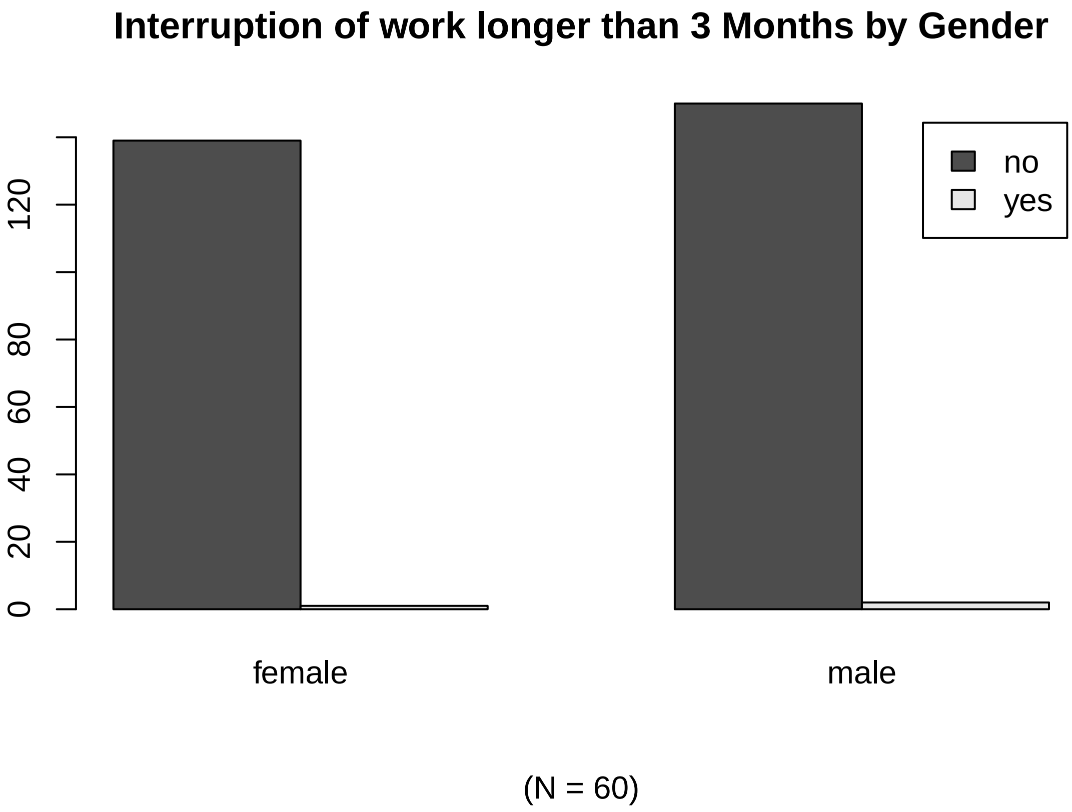
```
tab <- with(data, table(interruption_more_than_three_months, gender))
addmargins(tab)
```

```
##                                    gender
## interruption_more_than_three_months female male Sum
##                                 no     139  150 289
##                                 yes      1    2   3
##                                 Sum    140  152 292
```

- The contingency table shows that among 292 individuals in the dataset, 140 are female and 152 are male. From the three respondents who had interruption of work more than three months are male, whereas the respondent is a woman

View the barplot for the variables “gender” and “interruption_more_than_three_months”

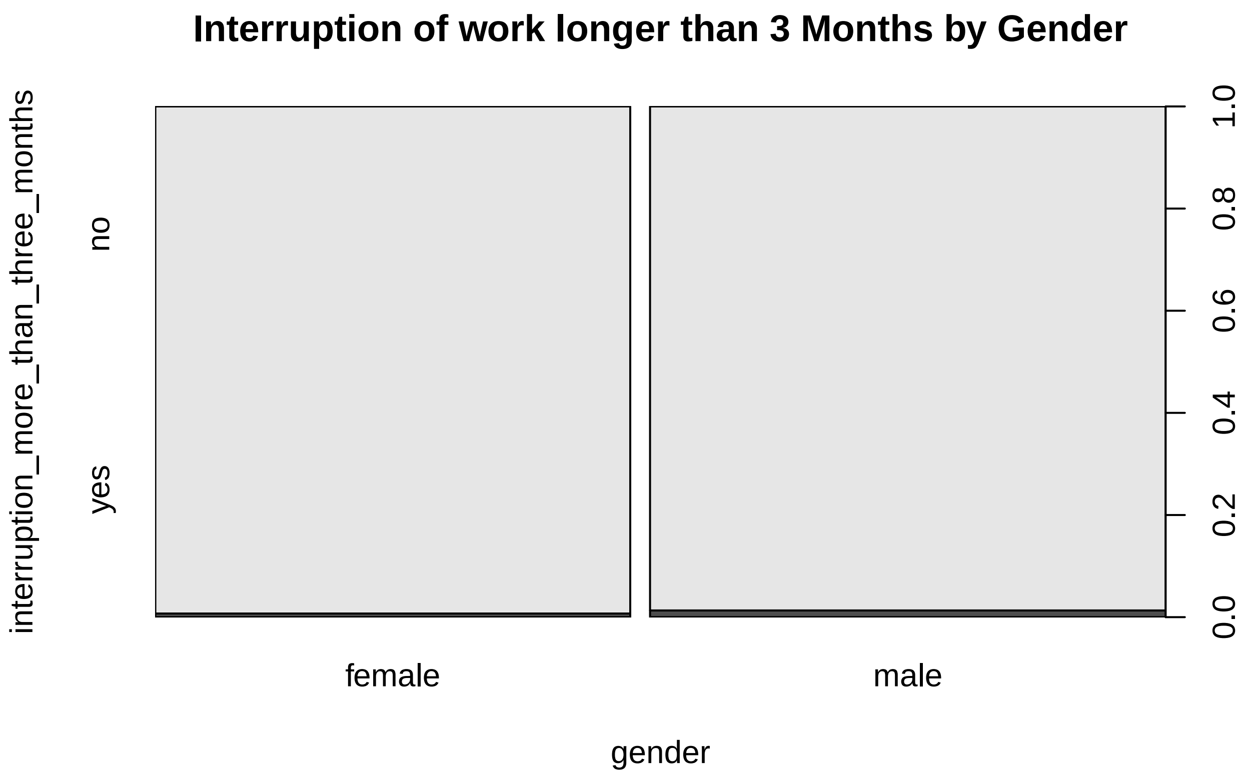
```
barplot(tab, beside = TRUE, legend = TRUE,
        main = "Interruption of work longer than 3 Months by Gender", sub = "(N = 60)")
```



This a representation of interruption of work longer than 3 Months by gender.

For a binary dependent variable, the spinogram is suitable, view the spine plot for the variables “gender” and “interruption_more_than_three_months”

```
spineplot(interruption_more_than_three_months ~ gender, data = data,
          main = "Interruption of work longer than 3 Months by Gender")
```



The x-axis shows the explanatory variable (gender). From the proportion in the x-axis it is to see that male respondents are slighty has more area of the bars. Additionally, the proportion of the males who experienced interruption of work longer than three months is higher than the females, this can be seen in the y-axis.

3.2.2 "age_category" and "interruption_more_than_three_months"

Create the contingency table for the variables age_category and interruption_more_than_three_months

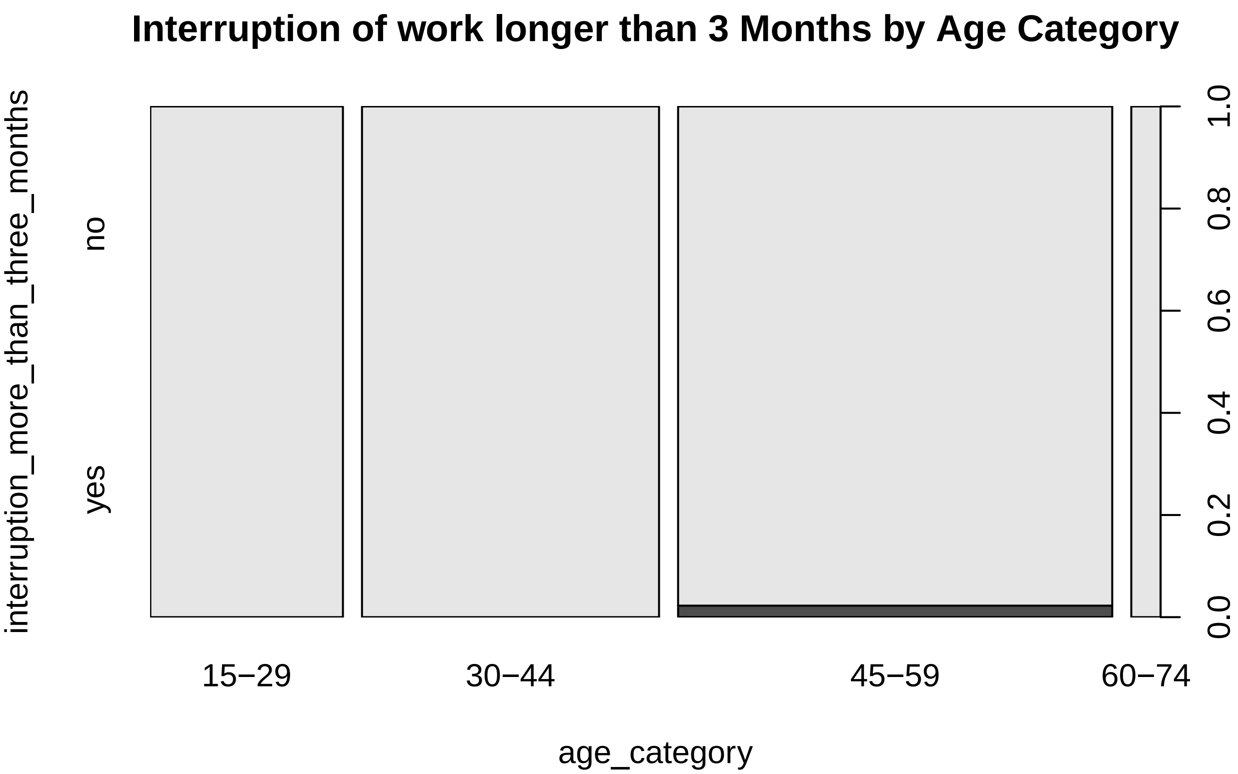
```
tab = with(data, table(interruption_more_than_three_months, age_category))
tab
```

```
##                                    age_category
## interruption_more_than_three_months 15-29 30-44 45-59 60-74
##                                 no     59    91   130     9
##                                 yes     0     0     3     0
```

The contingency table shows, in the age categories "15-29", "30-44", and "60-74" there are no individuals who have experienced the interruption (0). On the other hand, all the individuals who have experienced interruption of work longer than three months belong to the age category group "45-59". Hence this can be indication to there is significat relationship between age category and interruption of work longer than three months.

A "spinogram" is used to visualize the proportions depending on the age category:

```
spineplot(interruption_more_than_three_months ~ age_category,
          data = data,
          main = "Interruption of work longer than 3 Months by Age Category")
```



- In each bar (i.e. in each age category), the proportion of respondents who had interruption of work longer than 3 months are marked in black.
- Considering the density is therefore simply proportional to the width of the bar, we can see that the largest proportion is the respondents with age_category 45-59, where also all the respondents with interruption of work longer than three months belong.
- The proportion of respondents who had interruption of work longer than 3 months is only in the 45-59 category.

3.2.3 "desired_working_hours" and "interruption_more_than_three_months"

Provide summary statistics for two variables

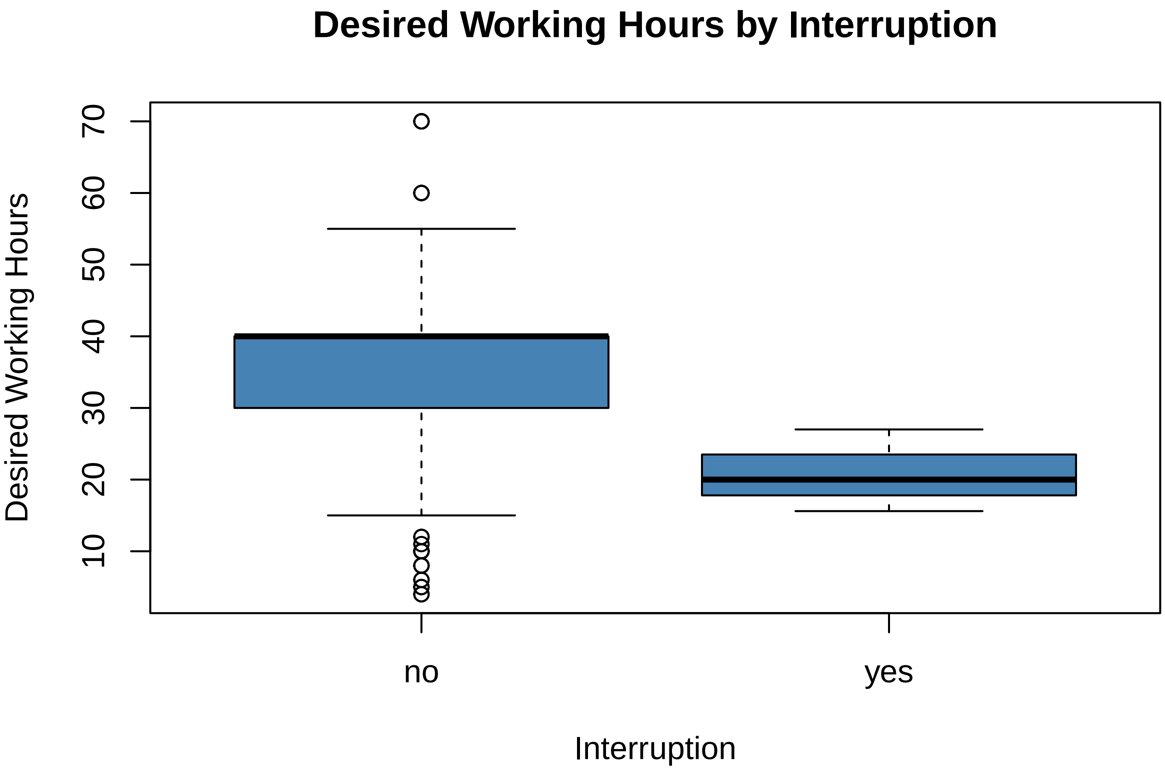
```
data %>%
  group_by(interruption_more_than_three_months) %>%
  summarise(
    count = n(),
    mean = mean(desired_working_hours)
  )
```

```
## # A tibble: 2 x 3
##   interruption_more_than_three_months count  mean
##   <fct>                               <int> <dbl>
## 1 no                                    289  36.2
## 2 yes                                     3  20.9
```

It can be seen that respondents with interruption of work longer than three months have lower mean value when it comes to desired working hours. This can be an indication to respondents with lower desired working hours, tend to experience interruption of work longer than three months.

View boxplot for variables "desired_working_hours" and "interruption_more_than_three_months"

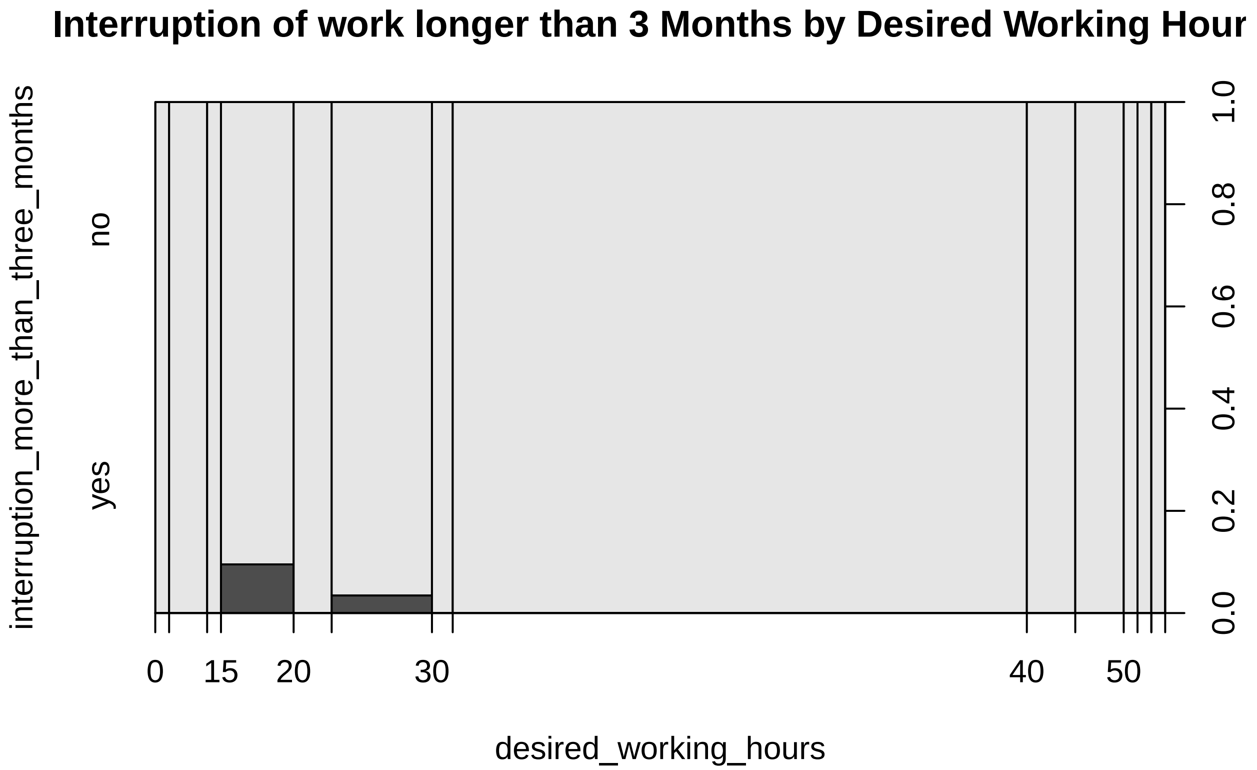
```r
boxplot(desired_working_hours~interruption_more_than_three_months,
        data=data,
        main="Desired Working Hours by Interruption",
        xlab="Interruption",
        ylab="Desired Working Hours",
        col="steelblue",
        border="black"
)
```



* Mean desired working hours in the interruption with “no” answer, is significantly higher than the mean score in the group, with “yes” answer. It can also be seen that the desired working hours for the “yes” answered group have less variability than the “no” group.

A “spinogram” is used to visualize the proportions depending on the desired working hours:

```r
spineplot(interruption_more_than_three_months ~ desired_working_hours,
          data = data,
          main = "Interruption of work longer than 3 Months by Desired Working Hours")
```



The taller segment 10-20 would correspond to the desired working hours with a higher proportion of individuals who experienced an interruption of work longer than 3 months than the other segment.

3.2.4 "current_work_in_months" and "interruption_more_than_three_months"

Provide summary statistics for two variables

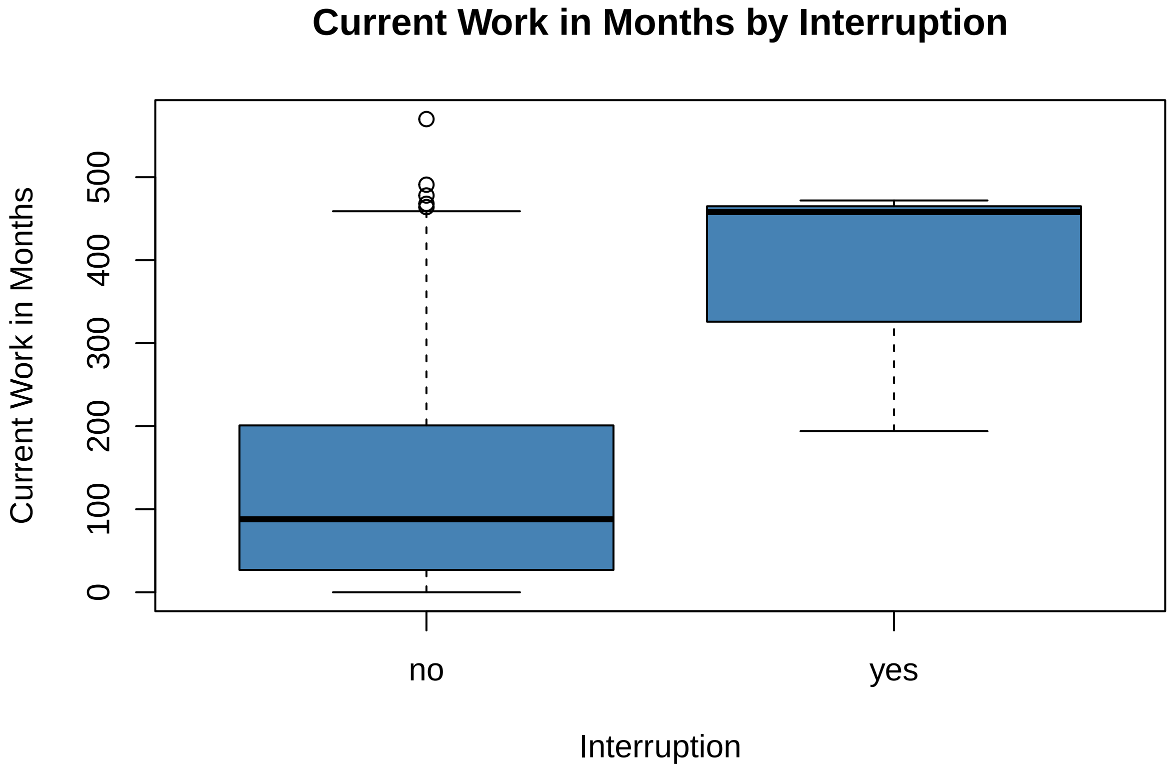
```
data %>%
  group_by(interruption_more_than_three_months) %>%
  summarise(
    count = n(),
    mean = mean(current_work_in_months)
    )
```

```
## # A tibble: 2 x 3
##   interruption_more_than_three_months count  mean
##   <fct>                               <int> <dbl>
## 1 no                                    289  131.
## 2 yes                                     3  375.
```

It can be seen that respondents with interruption of work longer than three months have much higher mean value when it comes to months at the current work. This can be an indication to respondents with higher months at the current work, tend to experience interruption of work longer than three months.

View boxplot for variables "current_work_in_months" and "interruption_more_than_three_months"

```
boxplot(current_work_in_months ~ interruption_more_than_three_months,
        data = data,
        main = "Current Work in Months by Interruption",
        xlab = "Interruption",
        ylab = "Current Work in Months",
        col = " steelblue",
        border = "black"
)
```



- Mean current work in months in the interruption with “no” answer, is significantly lower than the mean score in the group, with “yes” answer. It can also be seen that the current work in months for the “yes” answered group have more variability than the “no” group.
- There are additionally outliers to be seen in the “no” group, this can be due highly experienced group of people with no interruption of work longer than three months.

3.3 Multivariate Analysis

3.3.1 Logistic regression model to see joint influences of predictors on the dependent variable

```
model <- glm(interruption_more_than_three_months ~ desired_working_hours + current_work_in_months + age_
             data = data, family = binomial)

summary(model)
```

```
##
## Call:
## glm(formula = interruption_more_than_three_months ~ desired_working_hours +
##     current_work_in_months + age_category + gender, family = binomial,
##     data = data)
##
## Coefficients:
##                          Estimate   Std. Error z value Pr(>|z|)
## (Intercept)            -18.857542  5124.213585  -0.004   0.9971
## desired_working_hours   -0.165396     0.067118  -2.464   0.0137 *
## current_work_in_months   0.011995     0.007833   1.531   0.1257
## age_category30-44         -0.347017  6815.317292   0.000   1.0000
## age_category45-59         15.987991  5124.214188   0.003   0.9975
## age_category60-74         -6.328073 13862.827147   0.000   0.9996
## gendermale                 0.762096     1.952531   0.390   0.6963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.438  on 291  degrees of freedom
## Residual deviance: 15.733  on 285  degrees of freedom
## AIC: 29.733
##
## Number of Fisher Scoring iterations: 21
```

- The coefficient for desired_working_hours variable is -0.165396. It has a relatively small standard error and a statistically significant p-value (0.0137, marked with '*'). A negative coefficient suggests that as desired working hours increase, the odds of experiencing an interruption decrease.
- The coefficient for "gendermale" is 0.762096, indicating that being male is associated with higher odds of experiencing an interruption of work longer than three months. However, it has a relatively large standard error and also the p-value (0.6963) is not statistically significant.
- Null deviance represents the deviance of a model with no predictors (only the intercept), which is 33.438. Whereas, residual devianc represents the deviance for the model with all the predictors. In this case, the residual deviance is 15.733, which is lower than the null deviance. This suggests that the model with all the predictors explains more of the variation in the data compared to a null model.
- The final model is: Log(Interruption > 3 Months) = -18.857542 - 0.165396 * desired_working_hours
- In summary, the above logistic regression model suggests that "desired_working_hours" is a statistically significant predictor, whereas higher desired working hours are to have lower odds of experiencing an interruption. Other predictors, such as "current_work_in_months," "age_category," and "gender," do not appear to be statistically significant in explaining likelihood of experiencing interruptions. Indicated by the reduction in deviance and the AIC suggests that the model with the predictor performs better than a null model.

3.3.2 Logistic regression with age_categrory

Generalized linear model (GLM) with the function is used

```
model = glm((interruption_more_than_three_months == "yes") ~ age_category, data = data, family = binomia
summary(model)
```

```
##
## Call:
## glm(formula = (interruption_more_than_three_months == "yes") ~
##     age_category, family = binomial, data = data)
##
## Coefficients:
##                             Estimate          Std. Error z value Pr(>|z|)
## (Intercept)         -21.5660685223333  3805.7392623904070  -0.006    0.995
## age_category30-44    -0.0000000005286  4886.1181502719646   0.000    1.000
## age_category45-59    17.7971463605459  3805.7393071945353   0.005    0.996
## age_category60-74    -0.0000000004949 10460.9766555454080   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 33.438  on 291  degrees of freedom
## Residual deviance: 28.682  on 288  degrees of freedom
## AIC: 36.682
##
## Number of Fisher Scoring iterations: 20
```

- This logistic regression model suggests that the “age_category” variable is not a statistically significant predictor of the likelihood of experiencing interruptions of work. The model’s coefficients for age categories are not statistically significant, and the intercept (reference group “15.29”) is also not significant.

Change the reference group to “45-59”:

```
data$age_category_r <- relevel(data$age_category, ref=3)
levels(data$age_category_r)
```

```
## [1] "45-59" "15-29" "30-44" "60-74"
```

Fit a logistic regression model to newly changed age_category

```
model1 <- glm(interruption_more_than_three_months ~ age_category_r, data = data, family = 'binomial')
summary(model1)
```

```
##
## Call:
## glm(formula = interruption_more_than_three_months ~ age_category_r,
##     family = "binomial", data = data)
##
## Coefficients:
##                      Estimate Std. Error z value       Pr(>|z|)
## (Intercept)            -3.769      0.584  -6.454 0.000000000109 ***
## age_category_r15-29   -17.797   3805.739  -0.005          0.996
## age_category_r30-44   -17.797   3064.392  -0.006          0.995
## age_category_r60-74   -17.797   9744.146  -0.002          0.999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.438  on 291  degrees of freedom
## Residual deviance: 28.682  on 288  degrees of freedom
## AIC: 36.682
##
## Number of Fisher Scoring iterations: 20
```

- However, after changing the reference group, the intercept becomes statistically significant. Age category “15-29” serves as the reference category, the logistic regression model suggests that none of the other age categories “15-29”, “30-44” and “60-74” have a statistically significant effect on the likelihood to experience interruptions.
- The estimated intercept is approximately -3.769 and is aelatively small standard error also highly statistically significant ($p < 0.001$). The intercept represents the log-odds of experiencing an interruption in the reference category “45-59”. In this output, only the intercept is statistically significant.
- Intercept: the log-odds of interruption of work longer than three months for someone aged “45-59” (reference group) is -3.769. Thus, the baseline chance of interruption of work longer than three months for a respondent with age_category “45-59” is 0.02307692, or 2 in 100.

```r
exp(coef(model1)[1])
```

```
## (Intercept)
##  0.02307692
```

4 Regression Analysis

4.1 Stepwise Logistic Regression

For the regression anlysis the stepwise logistic regression technique has been chosen for building a logistic model that will iteratively selects or deselects predictors based on their statistical significance.

Define the base model (intercept-only)

```r
base.model <- glm(interruption_more_than_three_months ~ 1, data = data[1:5], family = binomial)
```

Define the scope model (full model)

```r
scope.model <- glm(interruption_more_than_three_months ~ ., data = data[1:5], family = binomial)
```

Perform stepwise logistic regression

```r
step.model <- stepAIC(base.model,
                      direction = "both",
                      scope = list(upper=scope.model),
                      trace = FALSE)
```

Summarize the final selected model

```r
summary(step.model)
```

```
##
## Call:
## glm(formula = interruption_more_than_three_months ~ current_work_in_months +
##     desired_working_hours, family = binomial, data = data[1:5])
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -4.011951   1.838994  -2.182   0.0291 *
## current_work_in_months  0.010217   0.004155   2.459   0.0139 *
## desired_working_hours  -0.101853   0.050251  -2.027   0.0427 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.438  on 291  degrees of freedom
## Residual deviance: 21.057  on 289  degrees of freedom
## AIC: 27.057
##
## Number of Fisher Scoring iterations: 9
```

Calculate Odds Ratios

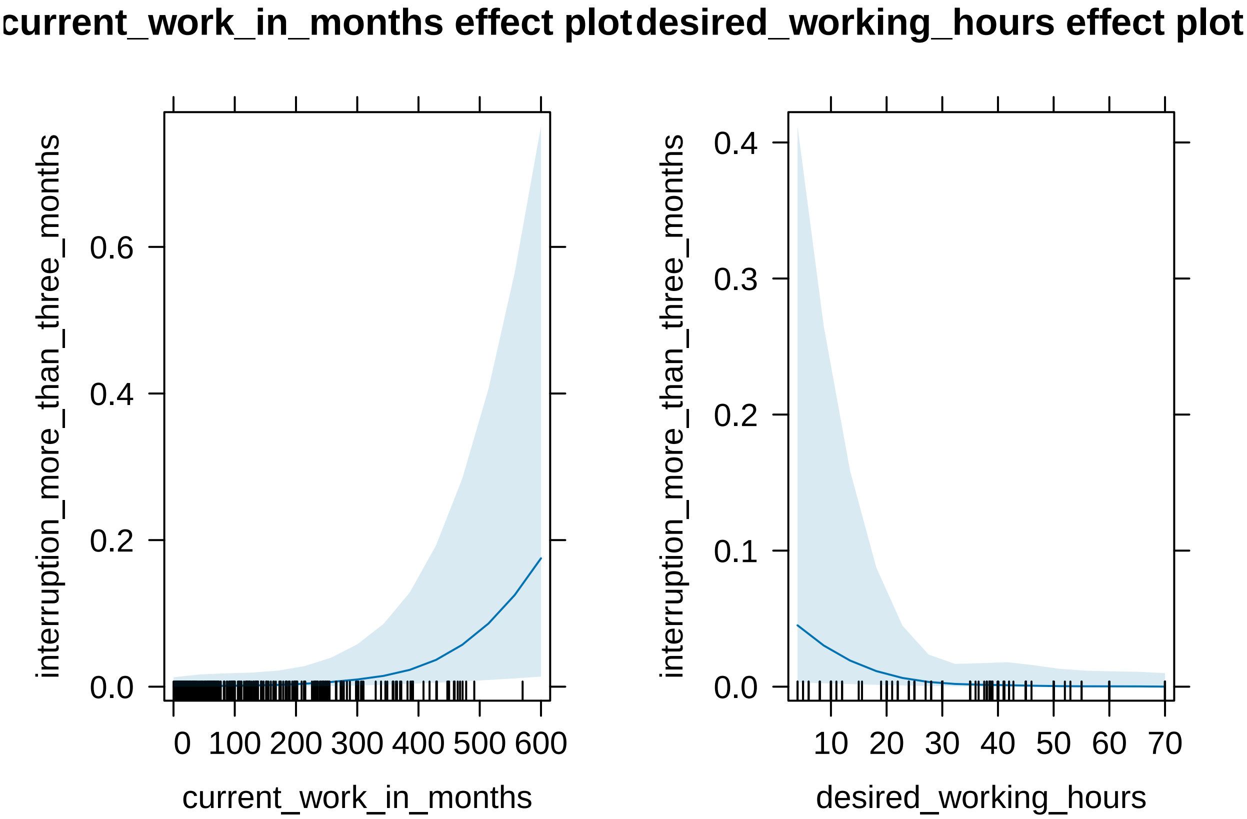
```r
odd.ratios <- exp(step.model$coefficients)
round(odd.ratios, 3)
```

```
##            (Intercept) current_work_in_months  desired_working_hours
##                      0.018                      1.010                      0.903
```

- The logistic regression analysis conducted in both stepwise direction aimed suggests that the "current_work_in_months" and "desired_working_hours" variables are statistically significant predictors of the binary response variable "interruption_more_than_three_months."
- The coefficient estimate for the intercept was found to be -4.011951.
- current_work_in_months is one of the predictor variables, and has an estimated coefficient of 0.010217. This coefficient represents the change in the log-odds of the response variable for a one-unit increase in the "current_work_in_months" variable while holding other variables constant.
- desired_working_hours is another predictor variable with an estimated coefficient of -0.101853.
- Log(interruption_more_than_three_months) = -4.011951 + 0.010217 * current_work_in_months - 0.101853 * desired_working_hours
- In summary, the stepwise logistic regression on both ways suggests that the "current_work_in_months" and "desired_working_hours" variables are statistically significant predictors of the binary response variable "interruption_more_than_three_months."

View the all effects plot for the step wise

```r
plot(allEffects(step.model), type = "response")
```



It is also to see from the alleEffects plot that likelihood to experience interruption of work longer than three months increases as the months in the current work increases. In desired_working_hours variable it is opposite, namely likelihood to experience interruption of work longer than three months decreases as the months in the desired working hours increase.

Analyze the 4 main effects and 6 pairwise interactions among these variables with logistic regression as well.

```r
binlrfit <- glm(interruption_more_than_three_months ~ .^2, data = data[1:5], family = binomial)
summary(binlrfit)
```

```
## 
## Call:
## glm(formula = interruption_more_than_three_months ~ .^2, family = binomial, 
##     data = data[1:5])
## 
## Coefficients:
##                                                     Estimate     Std. Error
## (Intercept)                                    -115.8834797 158090.6456492
## gendermale                                       91.4021516 108611.5629265
## age_category30-44                                89.9944573 219535.5198101
## age_category45-59                               111.3785287 158090.6459195
## age_category60-74                                93.7302289 191467.5231595
## desired_working_hours                             2.2970615   3785.3119275
## current_work_in_months                           -0.2766975   1039.3230599
## gendermale:age_category30-44                    -30.2954406 127773.3064508
## gendermale:age_category45-59                   -168.6219579 116928.1610422
## gendermale:age_category60-74                   -122.4704180 210376.8210008
## gendermale:desired_working_hours                 -2.3777290    875.1648835
## gendermale:current_work_in_months                 0.3197881    111.3967323
## age_category30-44:desired_working_hours          -2.1902358   5549.5998387
## age_category45-59:desired_working_hours          -2.2955524   3785.3119372
## age_category60-74:desired_working_hours          -2.2998771   7013.4683631
## age_category30-44:current_work_in_months          0.0649526   1078.0410060
## age_category45-59:current_work_in_months          0.2858808   1039.3230593
## age_category60-74:current_work_in_months          0.2547249   1394.8688943
## desired_working_hours:current_work_in_months     -0.0002385      0.0007693
##                                              z value Pr(>|z|)
## (Intercept)                                   -0.001    0.999
## gendermale                                     0.001    0.999
## age_category30-44                              0.000    1.000
## age_category45-59                              0.001    0.999
## age_category60-74                              0.000    1.000
## desired_working_hours                          0.001    1.000
## current_work_in_months                         0.000    1.000
## gendermale:age_category30-44                   0.000    1.000
## gendermale:age_category45-59                  -0.001    0.999
## gendermale:age_category60-74                  -0.001    1.000
## gendermale:desired_working_hours              -0.003    0.998
## gendermale:current_work_in_months              0.003    0.998
## age_category30-44:desired_working_hours        0.000    1.000
## age_category45-59:desired_working_hours       -0.001    1.000
## age_category60-74:desired_working_hours        0.000    1.000
## age_category30-44:current_work_in_months       0.000    1.000
## age_category45-59:current_work_in_months       0.000    1.000
## age_category60-74:current_work_in_months       0.000    1.000
## desired_working_hours:current_work_in_months  -0.310    0.756
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 33.438  on 291  degrees of freedom
## Residual deviance:  9.889  on 273  degrees of freedom
## AIC: 47.889
##
## Number of Fisher Scoring iterations: 24
```

All the coefficients in the model are not statistically significant (p-values close to 1.000).

5 Conclusion & Critical Evaluation

- The variables for the study are used as from of the microcensus data provided Statistik Austria. Multiple steps including transforming data to more convenient names, handling the NA values are performed.
- After the initial steps of preparing the data, a descriptive analysis is performed to individual and in pairs, to identify the relationships/pattern that exist in the dataset. In the descriptive analysis part data univariate data analysis is performed for each individual variable of study. The bivariate and multivariate analysis of the variables of have included to gain insights to individual and joint influences of the predictors on the dependent variable. Joint influence of predictor variables are investigated by simple logistic regression models.
- Logistic regression model done with all the predictors on the dependent variable, suggests that "desired_working_hours" is a statistically significant predictorl. Other predictors, such as "current_work_in_months," "age_category," and "gender," do not appear to be statistically significant in explaining likelihood of experiencing interruptions.
- Another simple logistic regression model done with the age_category variable shows the reference group "45-59" is statistically significant ($p < 0.001$). This has been done after the potential relationship observed in the bivariate analysis. The baseline chance of interruption of work longer than three months for a respondent with age_category "45-59" is 0.02307692, or 2 in 100.
- Complete regression analysis for by extracting and adding predictors manually has been done with step-wise logistic regression method by using "45-59" stepAIC() function from the "MASS" package. The results of selection of predictors in both variables shows us that both "current_work_in_months" and "desired_working_hours" variables are statistically significant predictors of the binary response variable "interruption_more_than_three_months." with Log(interruption_more_than_three_months) = -4.011951 + 0.010217 * current_work_in_months - 0.101853 * desired_working_hours
- The answer to the research question " Is there a relationship between gender, age category, desired working hours and months in the current working place among workers with respect to the likelihood of an interruption of more than three months in Austria?" is as follows: according to the step wise logistic regression the current_work_in_months and desired_working_hours was found to have a statistically significant relationship with the likelihood of an interruption of more than three months among workers in Austria. For every one-unit increase in the number of months worked in the current position, the log-odds of experiencing an interruption of more than three months increase by 0.010217. For every one-unit increase in desired working hours, the log-odds of experiencing an interruption of more than three months decrease by 0.101853. This can be implication that respondents who desire to work more hours may have a lower likelihood of experiencing an interruption of more than three months.
- The primary challenge encountered with the dataset was the limited availability of respondents who provided responses to the survey question related to the dependent variable
- While the analysis provides insights into the specific dataset used, generalizing the findings should be done carefully. The microcensus dataset is limited to Austria, and also the specific characteristics and dynamics of employment data, may not apply for other countries.
- The analysis raises one of the following question: Could the inclusion of additional variables related to employment-related characteristics or the choice of variables would provide more insights about the interruption of work longer than three months?