Section A

Part 1 Exploratory Data Analysis (EDA)

Undertake an exploratory analysis of the variables in the survey dataset. Illustrate your analysis with suitable plots. Communicate what you have learnt from the EDA about the data to be analysed

1.1 Introduction

Within part 1 of this report survey data relating to different aspects of subjective wellbeing and demographic questions will be analysed. The data used within this section can be found below in **Table 1**.

Table 1. Description of Variables Used Within Study

Variable	Description	Coding
Case	Case number	-
MCZ_1	Overall, how satisfied are you with your life nowadays?	Likert scale (1-10)
MCZ_2	Overall, to what extent do you feel things you do in your life are worthwhile?	Likert scale (1-10)
MCZ_3	Overall, how happy did you feel yesterday?	Likert scale (1-10)
MCZ_4	Overall, how anxious did you feel yesterday	Likert scale (1-10)
Qhealthr8	How is your health in general	(1 = very good, 2 = good, 3 = fair, 4 = bad, 5 = very bad)
rsex	Sex of respondent	(1 = male, 2 = female)
agex	Grouped age	(1 = 16 to 24, 2 = 25 to 44, 3 = 45 to 54, 4 = 55 to 64, 5 = 65 to 74, 6 = 75 and over)
Marstat3r9	Marital status	(1 = married/cohabiting, 2 = single, 3 = widowed/divorced/separated)
Highed4	Highest level of qualification	(1 = degree of equivalent, 2 = below degree level, 3 = other, 4 = no qualifications)
ethnicity	Ethnicity	(1 = white, 2 = other)
Dvilo3a	Employment status according to international Labour Organisation (ILO) definition	(1 = employed, 2 = unemployed, 3 = economically inactive)
ftptwk	Full-time or part-time work	(1 = full-time, 2 = part-time)
Nsecac3	NS-SEC	(1 = managerial and professional occupations, 2 = intermediate occupations, 3 = routine and manual occupations, 4 = not classified)
gora	Government Office Region	(1 = North East, 2 = North West, 3 = Yorkshire and the Humber, 4 = East Midlands, 5 = West Midlands, 6 = Eastern, 7 = London, 8 = South East, 9 = South West, 10 = Wales, 11 = Scotland)

1.2 Data Preparation

```r
# loading required libraries
library(tidyverse)
library(moments)
library(patchwork)
library(corrplot)
library(PerformanceAnalytics)
# Importing Survey
survey_data <- read.delim("survey.txt", header = TRUE)
```

1.2.1 Missing Values

Before the EDA was carried out the data was prepared. This first included the identification and removal of missing values. The data was found to contain 215 missing values.

```
# Checking for missing values
is.na(survey_data)
sum(is.na(survey_data))
# Removing rows with missing
survey_data <- drop_na(survey_data)
```

1.2.2 Removal of -2 and -1

Some of the variables were identified as containing observations coded as '-1 = Do Not Know' and '-2 = Refused to Answer' this coding may cause errors in the analyses. These problem observations were identified and then removed from the data.

```
# checking which columns have
# redundant -2 or -1 coding
# Looking at min value
summary(survey_data)

# Removal of -1 and -2 observations
# MCZ_1
survey_data <- subset(survey_data, MCZ_1!=-1 & MCZ_1!=-2)

# MCZ_2
survey_data <- subset(survey_data, MCZ_2!=-1 & MCZ_2!=-2)

# MCZ_3
survey_data <- subset(survey_data, MCZ_3!=-1 & MCZ_3!=-2)

# MCZ_4
survey_data <- subset(survey_data, MCZ_4!=-1 & MCZ_4!=-2)
```

1.3 EDA

1.3.1 Numerical Summaries and Visualisations

Within this study the variables MCZ 1 - 4 will be treated as numeric (not ordinal). These summary statistics can be found below in **Table 2**, the variables MCZ 1-3 all share similar distributions with means between 7.42-7.71. This can be better observed in **Figure 1** and **Figure 2**, displaying the histogram and overlaid density plot.

In the context of the data these variables relate to questions regarding current happiness and satisfaction, this may indicate the surveyed popular are quite satisfied and happy with their lives. This is further supported by the mean for MCZ 4, which relates to how anxious the participant felt the day before. With a relatively low mean of 3.43 indicating low levels of anxiety on average.

```r
# Creating function to calculate stats
summary_stats = function(x){
  summary = summary(x)
  std = c('STD',sd(x))
  var = c('VAR',var(x))
  skew = c('SKEW',skewness(x))
  kurt = c('KURT',kurtosis(x))
  return(c(summary, std, var, skew, kurt))
}

# calculating summary stats for MCZ variables
# MCZ_1
summary_stats(survey_data$MCZ_1)
# MCZ_2
summary_stats(survey_data$MCZ_2)
# MCZ_3
summary_stats(survey_data$MCZ_3)
# MCZ_4
summary_stats(survey_data$MCZ_4)

# Distribution visualisations
# Numerical
# MCZ_1
mcz_1 <- ggplot(survey_data, aes(x = MCZ_1)) +
  geom_histogram()

# MCZ_2
mcz_2 <- ggplot(survey_data, aes(x = MCZ_2)) +
  geom_histogram()

# MCZ_3
mcz_3 <- ggplot(survey_data, aes(x = MCZ_3)) +
  geom_histogram()

# MCZ_4
mcz_4 <- ggplot(survey_data, aes(x = MCZ_4)) +
  geom_histogram()

mcz_1 + mcz_2 + mcz_3 + mcz_4


# density plot
density <- ggplot()
density <- density + geom_density(data=survey_data, aes(MCZ_1,
colour = "MCZ_1"))
density <- density + geom_density(data=survey_data, aes(MCZ_2,
colour = "MCZ_2"),  size=1.2)
density <- density + geom_density(data=survey_data, aes(MCZ_3,
colour = "MCZ_3"),  size = 1)
```

```
density <- density + geom_density(data=survey_data, aes(MCZ_4,
colour = "MCZ_4"))

density + scale_color_brewer(palette = "Set1") + theme_minimal() +
  theme(legend.spacing.y = unit(4.0, 'cm'),
        legend.text = element_text(size=10),
        legend.key.size = unit(3,"line")) +
  xlab("MCZ Score") +
  guides(colour=guide_legend(title="MCZ Variable", title.position =
"left"))
```

Table 2. Summary Statistics for the MCZ Variables (1-4)

Variable	Mean	Median	1^{st} quartile	3^{rd} quartile	Min	Max	STD	VAR	Skew	Kurtosis
MCZ 1	7.42	8	7	9	0	10	1.83	3.33	-0.92	1.02
MCZ 2	7.71	8	7	9	0	10	1.73	2.99	-1.02	1.53
MCZ 3	7.55	8	7	9	0	10	2.16	4.67	-1.17	1.25
MCZ 4	3.43	3	1	6	0	10	3.03	9.20	0.50	-0.97

Figure 1. Histograms for the MCZ Variables (MCZ 1-4)

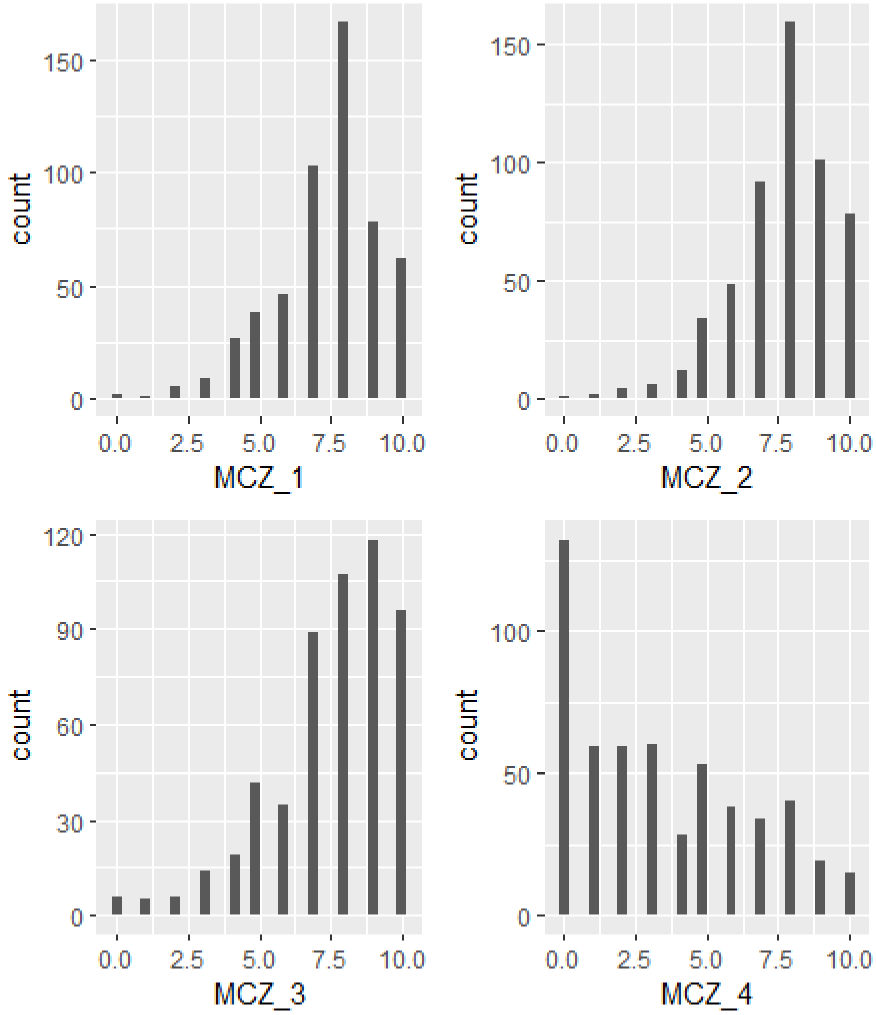
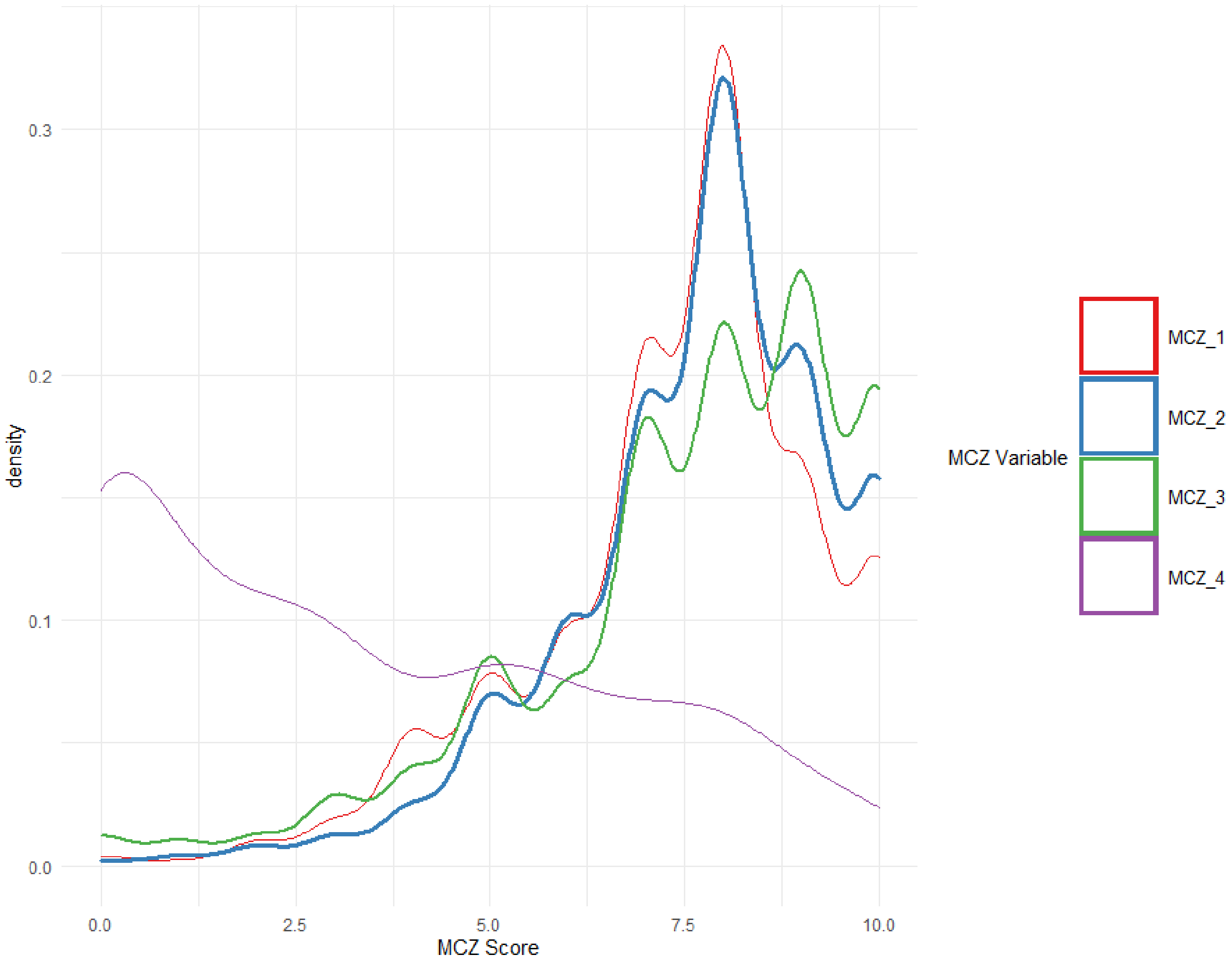


Figure 2. Overlaid Density Plots for the MCZ Variables (MCZ 1- 4)



1.3.2 Categorical Summary and Visualisation

For the remaining categorical variables, the summaries are presented in the form of the counts and proportions per category observed. These can be found below in **Table's 3-4**. A visualisation of the distribution per category can be found below in **Figure 3**.

From these summaries some conclusions regarding the surveyed population can be made:

- The proportion of male and female is almost equal
- 89% of those surveyed rated their health in general as either very good or good
- The largest proportion of participants were in the 25 to 44 age group
- 89% of the participants were white

```
# for categorical (frequencies)
# Counts
for(i in 6:15){
  print(table(survey_data[i]))
}

# proportions
for(i in 6:15){
  print(signif((prop.table(table(survey_data[i]))), digits = 2) *
100)
}
survey_data$QHealthr <- as.factor(survey_data$QHealthr)
survey_data$RSEX <- as.factor(survey_data$RSEX)
survey_data$AGEX <- as.factor(survey_data$AGEX)
survey_data$marstat3r <- as.factor(survey_data$marstat3r)
survey_data$highed4 <- as.factor(survey_data$highed4)
survey_data$Ethnicity <- as.factor(survey_data$Ethnicity)
survey_data$DVILO3a <- as.factor(survey_data$DVILO3a)
survey_data$FtPtWk <- as.factor(survey_data$FtPtWk)
survey_data$NSECAC3 <- as.factor(survey_data$NSECAC3)
survey_data$GorA <- as.factor(survey_data$GorA)
# Categorical Visualizations
# Bar Charts
# qhealthr8
qhealthr8 <- ggplot(survey_data, aes(x=QHealthr)) +
  geom_bar()
# rsex
rsex <- ggplot(survey_data, aes(x = RSEX)) +
  geom_bar()
# agex
agex <- ggplot(survey_data, aes(x = AGEX)) +
  geom_bar()
# marstat3r9
marstat3r9 <- ggplot(survey_data, aes(x = marstat3r)) +
  geom_bar()
# highed4
highed4 <- ggplot(survey_data, aes(x = highed4)) +
  geom_bar()
# ethnicity
ethnicity <- ggplot(survey_data, aes(x = Ethnicity)) +
  geom_bar()
# dvilo3a
dvilo3a <- ggplot(survey_data, aes(x = DVILO3a)) +
  geom_bar()
# ftptwk
ftptwk <- ggplot(survey_data, aes(x = FtPtWk)) +
  geom_bar()

# nsecac3
nsecac3 <- ggplot(survey_data, aes(x = NSECAC3)) +
  geom_bar()

# gora
gora <- ggplot(survey_data, aes(x = GorA)) +
  geom_bar()

qhealthr8 + rsex + agex + marstat3r9 +
  highed4 + ethnicity + dvilo3a +
  ftptwk + nsecac3 + gora
```

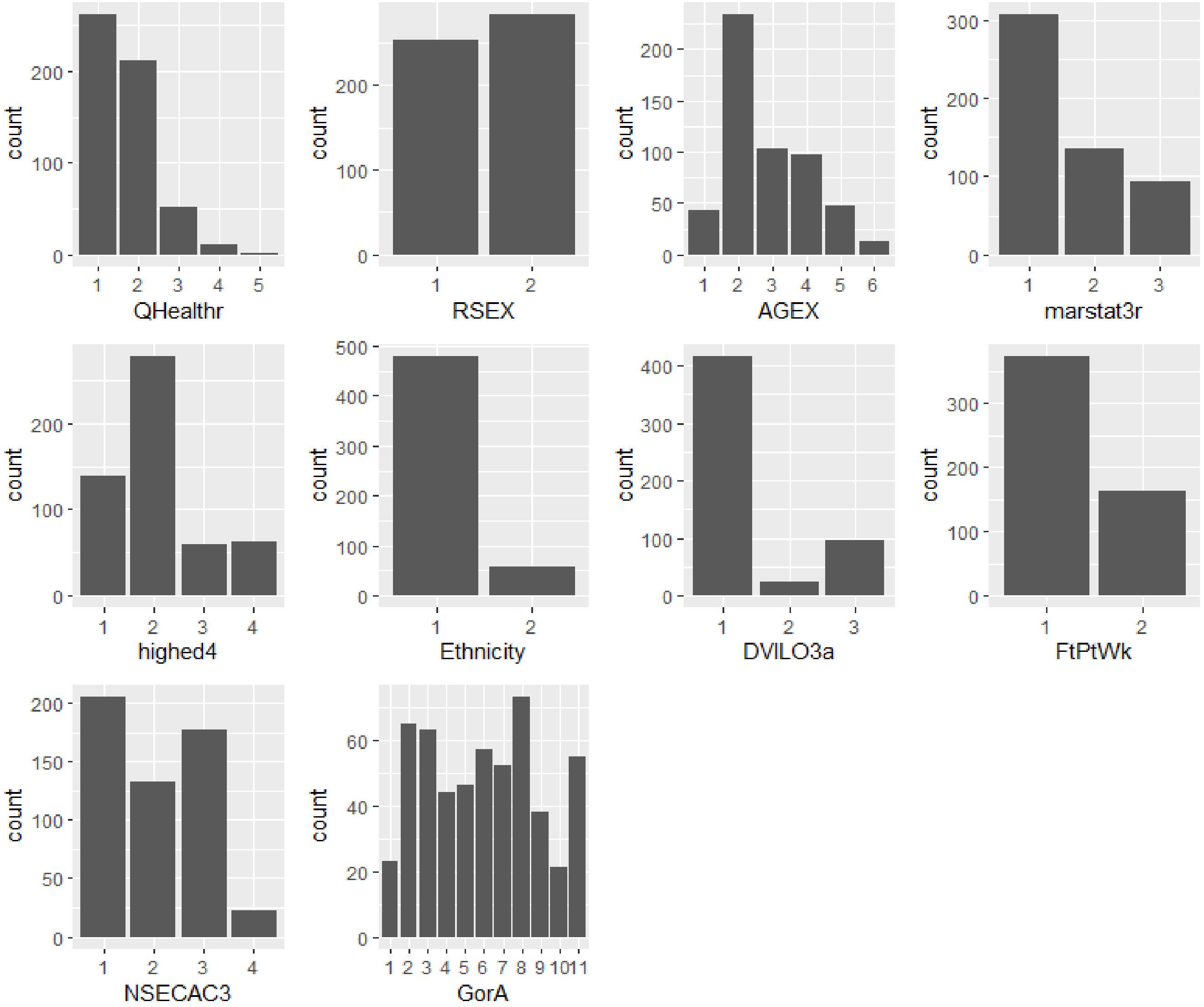
Table 3. Categorical Summary (Counts)

	Category										
Variable	1	2	3	4	5	6	7	8	9	10	11
qhealthr8	262	213	51	11	1	-	-	-	-	-	-
rsex	255	283	-	-	-	-	-	-	-	-	-
agex	43	234	103	98	48	12	-	-	-	-	-
marstat3r9	308	136	94	-	-	-	-	-	-		-
highed4	138	278	60	62	-	-	-	-	-	-	-
ethnicity	481	57	-	-	-	-	-	-	-	-	-
dvilo3a	418	24	96	-	-	-	-	-	-	-	-
ftptwk	375	163	-	-	-	-	-	-	-	-	-
nsecac3	206	132	177	23	-	-	-	-	-	-	-
gora	23	65	63	44	46	58	52	73	38	21	55

Table 4 Categorical Summary (Proportions)

	Category (%)										
Variable	1	2	3	4	5	6	7	8	9	10	11
qhealthr8	49	40	9.50	2	0.19	-	-	-	-	-	-
rsex	47	53	-	-	-	-	-	-	-	-	-
agex	8	43	19	18	8.9	2.2	-	-	-	-	-
marstat3r9	57	25	17	-	-	-	-	-	-		-
Highed4	26	52	11	12	-	-	-	-	-	-	-
ethnicity	89	11	-	-	-	-	-	-	-	-	-
dvilo3a	78	4.5	18	-	-	-	-	-	-	-	-
ftptwk	70	30	-	-	-	-	-	-	-	-	-
nsecac3	38	25	33	4.3	-	-	-	-	-	-	-
gora	4.3	12	12	8.2	8.6	11	9.7	14	7.1	3.9	10

Figure 3. Numerical Variable Visualisation of Counts



1.3.2 Exploration of MCZ Scores and Ethnicity and Highest Level of Qualification

Through conducting the EDA an interesting pattern was identified between the respondent's ethnicity, highest level of qualification, and the MCZ scores. To explore this the means for each MCZ score for each level of qualification and the two ethnicities "White" and "Other" were calculated. **Table 5** and **Figure 4** display these patterns. First, the ethnicities coded as "other" show lower mean scores for MCZ 1-3 and a higher mean score for MCZ 4 for almost all qualification levels. This may suggest that "other" ethnicities have lower life satisfaction and higher anxiety compared to those of white ethnicity.

Secondly, it can be seen the greatest difference in means in general is seen in the "no qualification" category between white and other ethnicities for MCZ 1- 3. With white ethnicities score on average 2 whole points higher. This suggests that when both ethnicities have no qualifications, those who are white report significantly higher happiness and highlights an area for further exploration.

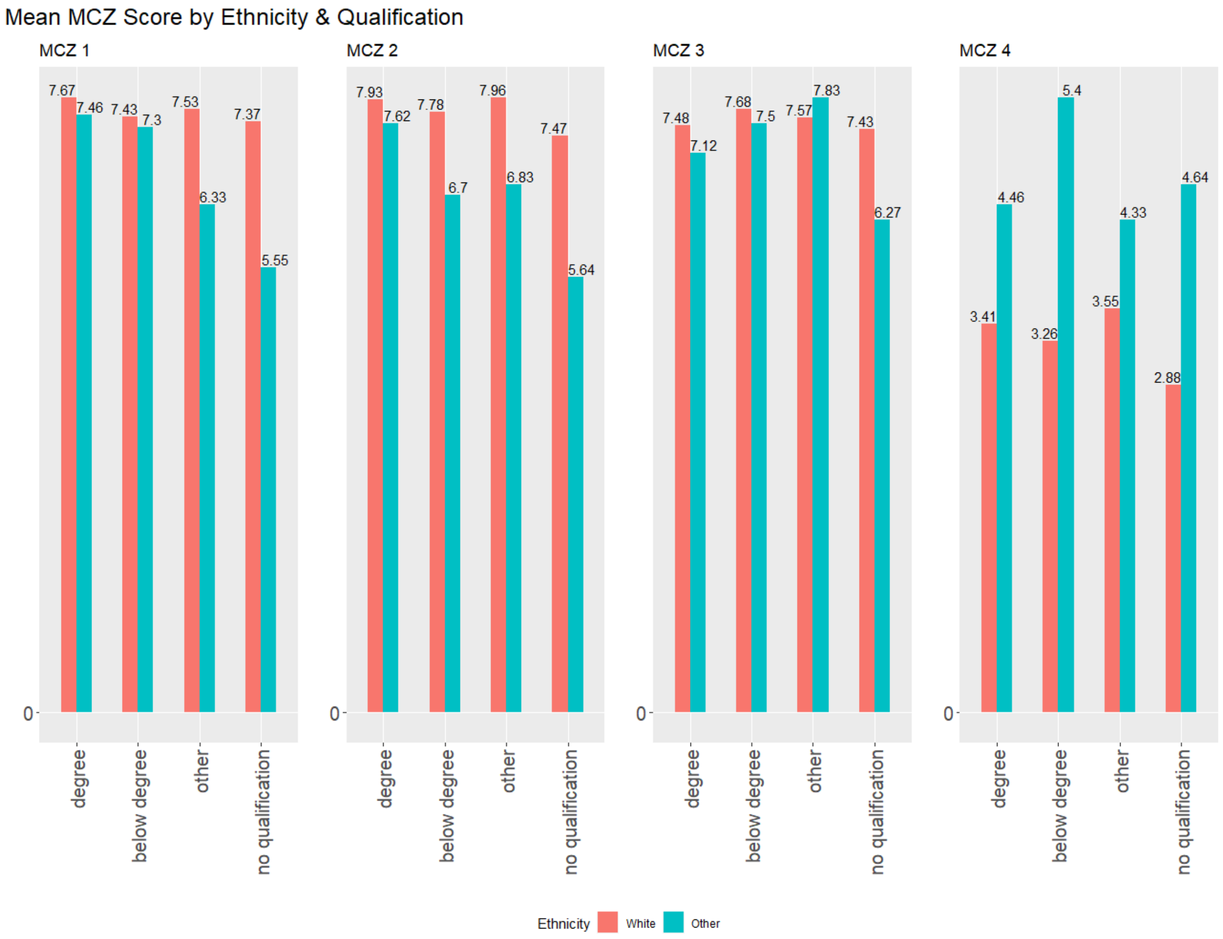
Although it should be noted that the "other" ethnicities have significantly less representation in this study, making up just 11% of the population.

```
# MCZ and qualification and ethnicity
plotdata.EQ <- survey_data %>%
  group_by(highed4, Ethnicity) %>%
  summarise(mean_mcz1 = round(mean(MCZ_1),2),
            mean_mcz2 = round(mean(MCZ_2),2),
            mean_mcz3 = round(mean(MCZ_3),2),
            mean_mcz4 = round(mean(MCZ_4),2))
print.data.frame(plotdata.EQ)

# visualising

p1 <- ggplot(plotdata.EQ, aes(x=factor(highed4,
                                    labels = c("degree",
                                               "below degree",
"other", "no qualification")),
                           y=mean_mcz1, fill=Ethnicity)) +
  geom_bar(stat="identity", position = position_dodge(width = 0.5),
width=0.5) +
  geom_text(aes(label=round(mean_mcz1, 2)),
            vjust=-0.25,
            position = position_dodge(width=0.9)) +
  scale_y_continuous(breaks = 0) +
  labs(title = "MCZ 1",
       x="",
       y="") +
  scale_fill_discrete(name = "Ethnicity", labels = c("White",
"Other"))


p2 <- ggplot(plotdata.EQ, aes(x=factor(highed4,
                                    labels = c("degree",
                                               "below degree",
"other", "no qualification")),
                           y=mean_mcz2, fill = Ethnicity)) +
  geom_bar(stat="identity", position = position_dodge(width = 0.5),
width=0.5) +
  geom_text(aes(label=round(mean_mcz2, 2)),
            vjust=-0.25,
            position = position_dodge(width=0.9)) +
  scale_y_continuous(breaks = 0) +
  labs(title = "MCZ 2",
       x="",
       y="") +
  scale_fill_discrete(name = "Ethnicity", labels = c("White",
"Other"))

p3 <- ggplot(plotdata.EQ, aes(x=factor(highed4,
                                    labels = c("degree",
                                               "below degree",
"other", "no qualification")),
                           y=mean_mcz3, fill = Ethnicity)) +
  geom_bar(stat="identity", position = position_dodge(width = 0.5),
width=0.5) +
  geom_text(aes(label=round(mean_mcz3, 2)),
            vjust=-0.25,
            position = position_dodge(width=0.9)) +
  scale_y_continuous(breaks = 0) +
  labs(title = "MCZ 3",
       x="",
       y="") +
  scale_fill_discrete(name = "Ethnicity", labels = c("White",
"Other"))

p4 <- ggplot(plotdata.EQ, aes(x=factor(highed4,
                                       labels = c("degree",
                                                  "below degree",
"other", "no qualification")),
                              y=mean_mcz4, fill = Ethnicity)) +
  geom_bar(stat="identity", position = position_dodge(width = 0.5),
width=0.5) +
  geom_text(aes(label=round(mean_mcz4, 2)),
            vjust=-0.25,
            position = position_dodge(width=0.9)) +
  scale_y_continuous(breaks = 0) +
  labs(title = "MCZ 4",
       x="",
       y="") +
  scale_fill_discrete(name = "Ethnicity", labels = c("White",
"Other"))


patchwork <- p1  + p2 + p3 + p4 +
  plot_layout(ncol = 4, guides = 'collect') & theme(legend.position
= "bottom") +
  theme(axis.text = element_text(size=15),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust =
1))

patchwork + plot_annotation(
  title = "Mean MCZ Score by Ethnicity & Qualification",
  theme = theme(plot.title = element_text(size=18))
)
```

Table 5. MCZ 1-4 Mean Scores for Each Level of Qualification and Both Ethnicities

		Mean			
Highest level of qualification	Ethnicity	MCZ 1	MCZ 2	MCZ 3	MCZ 4
Degree	White	7.67	7.93	7.48	3.41
	Other	7.46	7.62	7.12	4.46
Below Degree	White	7.43	7.78	7.68	3.26
	Other	7.30	6.70	7.50	5.40
Other	White	7.53	7.96	7.57	3.55
	Other	6.33	6.83	7.83	4.33
No Qualification	White	7.37	7.47	7.43	2.88
	Other	5.55	5.64	6.27	4.64

Figure 4. Bar Plot of MCZ 1-4 Mean Scores for Both Ethnicities



Part 2 Linear regression

Use the variables from qhealthr8 to gora as features (potential predictor variables) in a model for mcz_1 using a suitable supervised learning approach. Describe what your model shows about the influence of the different predictors included within it on the outcome variable

2.1 Introduction

As the MCZ variables are treated as numeric the supervised learning approach chosen was linear regression. The aim of this section is to fit a linear model using MCZ 1, a measure of life satisfaction, as the dependent variable and the variables from qhealthr8 to gora as the potential predictor variables.

2.2 Feature Selection Via Best Subset Selection

```
# loading libraries
library(olsrr)
library(tidyverse)
library(caret)
library(car)
```

The variables MCZ 2-4 were removed from the data frame as these variables will not be used within this analysis.

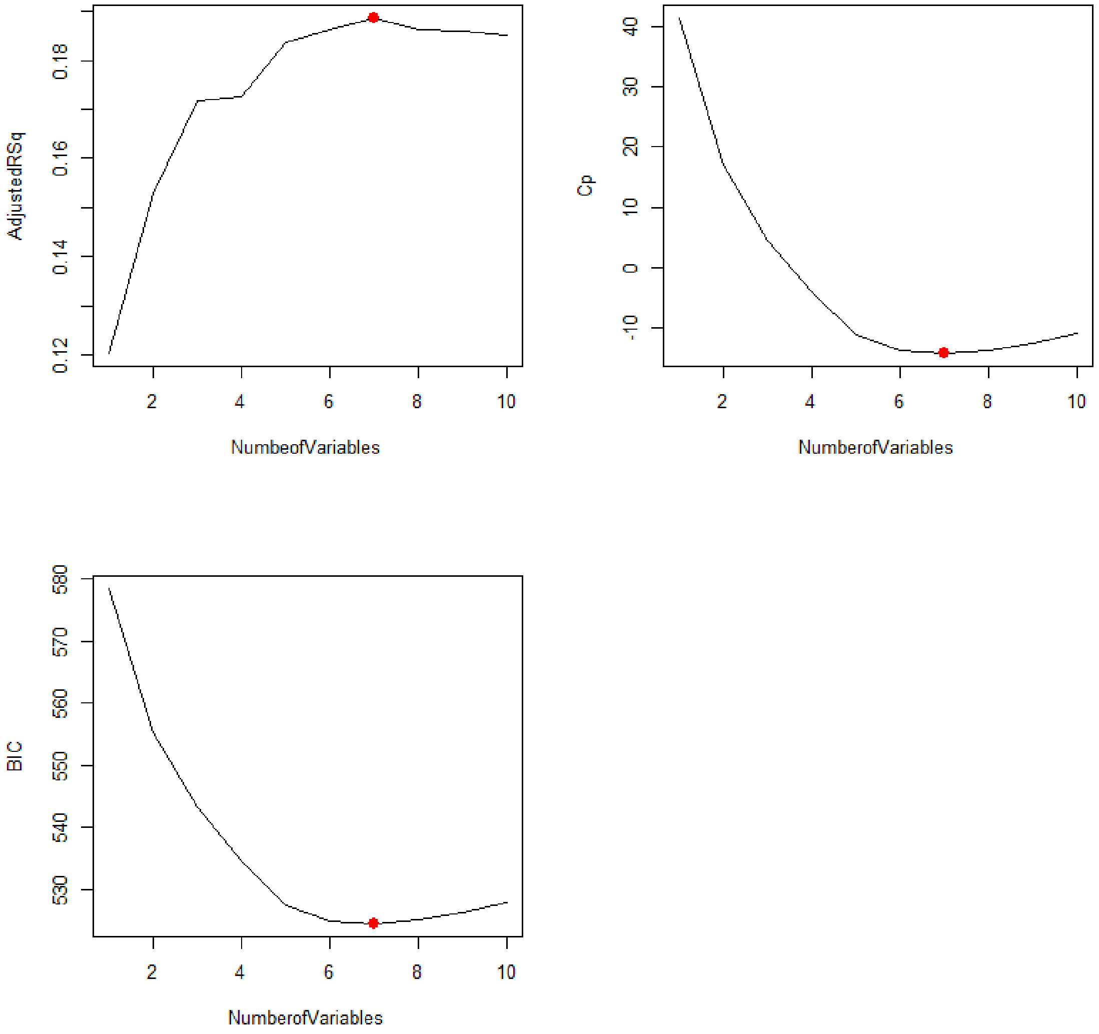
```
# dropping mcz2-4 (3-5)
survey_df <- survey_data[-c(1,3:5)]
str(survey_df)
```

Best subset was used as a method of feature selection. However, as regsubset() from the leaps library does not work on categorical data for feature selection, in this analysis ols_step_best_subset() from the olsrr library will be used. From the outputs of best subset selection, the Adjusted R-squared, Mallows Cp (Cp), and Bayesian Information Criterion (BIC) were plotted. These are metrics used in model selection, the aim is to maximise the Adjusted R-squared and minimise the prediction error (Cp and BIC). These can be found below in **Figure 5**. Best subset selection identified the 7-variable model as the potential best subset.

```
# fitting model with all predictors
model <- lm(MCZ_1 ~., data=survey_df)
# using ols_step_best_subset() to conduct best subset
subset <- ols_step_best_subset(model)

# plotting the best subset outputs
par(mfrow=c(2,2))

# adjusted R2
plot(subset$adjr, xlab = "NumbeofVariable", ylab = "AdjustedRSq",
     type = "l")
which.max(subset$adjr)
#7
points(7, subset$adjr[7], col="red", cex=2, pch=20)
```

```
#cp
plot(subset$cp, xlab = "NumberofVariables", ylab = "Cp",
     type = "l")
which.min(subset$cp)
# 7
points(7, subset$cp[7], col="red", cex = 2, pch=20)

# BIC
plot(subset$sbic, xlab = "NumberofVariables", ylab = "BIC",
     type = "l")
which.min(subset$sbic)
#7
points(7, subset$sbic[7], col="red", cex=2, pch=20)
```

Figure 5. Best Subset Selection Output



2.3 Model Selection Via k-fold cross-validation

Best subset selection offers a viable method for feature selection, however, the metrics for model selections (R2, BIC Cp) are calculated on the training that data, which has been used to fit the model and may lead to overfitting.

k-fold cross-validation was used on each of the models created by the best subset selection. The k-fold cross-validation measures the model's ability to predict 'out of sample' by partitioning the data into training and test data. This method produces an Averaged Root Mean Squared Error (RMSE) for each model which describes the prediction error within each model, the lower the RMSE the better the model. The output of this can be found below in **Table 6**. This method suggests that the model containing 3 variables is the most optimal model. This model was chosen despite best subset selections output, as it is clear the addition of extra variables increases the RMSE.

The variables of the 3 variable model were QHealthR, AGE, and DVILo3a.

```r
# using k-fold cross-validation
# helper function 1
get_model_formula <- function(id, object, outcome) {
  predictors <- object$predictors[id]
  predictors <- gsub(" ", "+", predictors)
  as.formula(paste0(outcome, "~", predictors))
}

# Helper function 2
get_cv_error <- function(model.formula, data){
  set.seed(1)
  train.control <- trainControl(method = "repeatedcv", number = 10,
repeats = 5)
  cv <- train(model.formula, data=data, method="lm",
              trControl=train.control)
  round(cv$results$RMSE, 2)
}

# generating errors for each best subset model
model.ids <- 1:10
cv.errors <- map(model.ids, get_model_formula, subset, "MCZ_1") %>%
  map(get_cv_error, data=survey_df) %>%
  unlist()
cv.errors

# identifying lowest errror
which.min(cv.errors)
```

Table 6. RMSE Output for The k-fold cross-validation Method

Model	RMSE
1	1.72
2	1.70
3	1.69
4	1.71
5	1.70
6	1.70
7	1.70
8	1.71
9	1.72
10	1.72

2.4 Model Interpretation

2.4.1 Model Summary

The output for this linear model can be found below in **Table 7**. These outputs suggest an estimated 17% of the variation in the dependent variable 'MCZ 1', is explained by the selected explanatory variables using adjusted R-square. Additionally, the significance of the reported F value of <2.2e-16 suggests this model provides a better fit to the data compared to the null model and at least one variable is a significant explainer.

```r
# fitting the model
final.model <- lm(MCZ_1 ~ QHealthr + AGEX + DVILO3a, data=survey_df)

# analysing model
summary(final.model)
```

Table 7. Linear Regression Model Summary Output

Variable	Unstandardized coefficient	Significance
Constant (intercept)	7.74	<2e-16
QHealthr2	-0.48	0.002
QHealthr3	-1.20	7.05e-06
QHealthr4	-3.95	5.73e-13
QHealthr5	-3.67	0.030
AGEX2	-0.0030	0.99
AGEX3	-0.24	0.44
AGEX4	0.47	0.12
AGEX5	0.71	0.053
AGEX6	1.67	0.0030
DVILo3a2	-1.32	0.00020
DVILo3a3	-0.070	0.75
R-Square	Adjusted R-square	Sig.
0.19	0.17	<2.2e-16

2.4.2 Coefficient Interpretation

As the explanatory variables are categorical the interpretation of the coefficients must be made in reference to the ‘reference’ category of each variable. This is the category that is not included in the model.

QHealthr – How is your health in general?

Reference category: ‘1 = very good’

Interpretation of the QHealthr variable suggest, controlling for the other variables in the model respectively, respondents who scored their health in general as ‘2’, ‘3’, ‘4’, and ‘5’all have respectively 0.48, 1.20, 3.95, and 3.67 decrease in expected MCZ 1 score compared to respondents who scored their health as ‘1’. This may suggest that as general health decreases (i.e. a number closer to 5), respondents satisfaction with life also decreases (i.e. an MCZ 1 number closer to 1). All observed differences were significant.

AGEX – Grouped Age

Reference category: ‘1 = 16 to 24’

Interpretation of the AGEX variable suggest, controlling for the other variables in the model respectively, respondents in the ‘2’ and ‘3’ age group have respectively 0.0030 and 0.24 decrease in expected MCZ 1 score compared to respondents in the ‘1’ age group. This suggests respondents aged 25-44 and 45-54 have an expected life satisfaction score lower than respondents aged 16-24. However, these observed differences were not significant.

Whereas respondents in the ‘4’, ‘5’, and ‘6’ age group have respectively 0.47, 0.71, and 1.67 increase in expected MCZ 1 score compared to respondents in the ‘1’ age group. This suggest respondents aged 55-64, 65-76, and 75+ have an expected life satisfaction score higher than respondents aged 16-24. However, only the different between age group 6 and 1 were significant.

Dvilo3a – Employment status according to international labour organisation

Reference category: ‘1 = employed’

Interpretation of the Dvilo3a variable suggest, controlling for the other variables in the model respectively, respondents in the ‘2’ and ‘3’ employment status group both have respectively 1.32 and 0.070 decrease in expected MCZ 1 score compared to respondents in the ‘1’ employment status group. This suggests respondents who work full time have an expected life satisfaction score higher than both respondents who are unemployed and respondents who are economically inactive. However only the observed difference between group ‘2’, unemployed, and group ‘1’ was significant.

Part 2 Prediction

Use your model from part 2 to predict mcz_1 for the cases in eval.data. Use this information to assess the quality of your model

3.1 Introduction

Within this section the model produced in part 2 will be used to predict values for MCZ 1 using a new dataset of unseen observations. The accuracy of these predictions will be assessed using the R2, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and a quantification of the error rate.

3.2 Model Predictions

The new dataset first had to be processed and cleaned in order to match the original training data.

```r
#loading in predict set
eval_data <- read.delim("eval.txt", header = TRUE)

# Checking for missing values
is.na(eval_data)
sum(is.na(eval_data))

# removing missing
eval_data <- drop_na(eval_data)


# removing mcz's
str(eval_data)
eval_df <- eval_data[-c(1,3:5)]
str(eval_df)

# checking which variables have -1 and -2 coding
summary(eval_df)
#Qhealth
# removing rows with this coding
remove2 <- which(eval_df$QHealthr == -2)

eval_df <- eval_df[-remove2,]
summary(eval_df)

# Recoding -1 and -2 as 11 in MCZ_1

# MCZ_1
eval_df$MCZ_1[eval_df$MCZ_1<0] <- 11
eval_df$MCZ_1

# creating factors in predict set
eval_df$QHealthr <- as.factor(eval_df$QHealthr)
eval_df$RSEX <- as.factor(eval_df$RSEX)
eval_df$AGEX <- as.factor(eval_df$AGEX)
eval_df$marstat3r <- as.factor(eval_df$marstat3r)
eval_df$highed4 <- as.factor(eval_df$highed4)
eval_df$Ethnicity <- as.factor(eval_df$Ethnicity)
```

```
eval_df$DVILO3a <- as.factor(eval_df$DVILO3a)
eval_df$FtPtWk <- as.factor(eval_df$FtPtWk)
eval_df$NSECAC3 <- as.factor(eval_df$NSECAC3)
eval_df$GorA <- as.factor(eval_df$GorA)
```

The model is then used to predict values for MCZ 1 based on the new observations for the independent variables. The accuracy results of these predictions can be found below in **Table 8**. As can be seen the RMSE, representing the average difference between the observed and predicted outcome values, is 1.64, representing an error rate of 21.8% which can be considered relatively low. The MAE, and MSE are both relatively low which further suggests the model performed well in prediction of MCZ 1. As the MAE are in the units of the outcome variable, we can conclude that the model's prediction on average is 1.25 away from the actual value.

However, the R-square (R2), representing the correlation between the observed outcome values and the predicted values of 0.14 is quite low. To further assess why this low value occurred model diagnostics should be performed on the linear model to identify assumption violations.

```
# making predictions
predicted <- predict(final.model, newdata = eval_df)

# computing accruacy metrics of the models predictions
data.frame(R2 = R2(predicted, eval_df$MCZ_1),
           RMSE = RMSE(predicted, eval_df$MCZ_1),
           MAE = MAE(predicted, eval_df$MCZ_1),
           MSE = mean((eval_df$MCZ_1-predicted)^2),
           error.rate = (RMSE(predicted, eval_df$MCZ_1) /
mean(eval_df$MCZ_1)))
```

Table 8. Model Prediction Accuracy Metrics Output

R2	RMSE	MAE	MSE	Error Rate (%)
0.14	1.64	1.25	2.70	21.8