State Anxiety Analysis on Crime Replication Dataset

Names: BEUMER kiki, HUET alexandre, JOSEPH yudita

Thu 30 May 2024

Introduction

In this report, we aim to carry out an in depth analysis on a crime replication dataset obtained on Kaggle. Initially, we had decided to use a dataset on credit card transactions in India. However, this data is not normally distributed. The difference between its distribution and the normal distribution was too large to ignore. Finally, we chose a dataset describing information regarding people's psychological states associated with crime; looking into their levels of trait anxiety and state anxiety in different demographic factors. Given the sensitivity of the data, this dataset is randomly generated in R.

Definition of state anxiety: state anxiety reflects the psychological and physiological transient reactions directly related to adverse situations in a specific moment. In contrast, the term trait anxiety refers to a trait of personality, describing individual differences related to a tendency to present state anxiety.

Understanding the Dataset

We start off by loading our dataset as follows.

```r
crime<-read.csv(file = "H:\\Documents\\CYTech\\Data Analysis\\Project\\crime_replication.csv")
```

```r
View(crime)
```

The dataset contains the following variables:

- **Participant**: An identifier for each individual in the dataset.
- **Age**: The values range between 17 and 85 years old.
- **State_Anxiety**: Measured on a scale from 0 to 5, generated with a normal distribution (mean = 2.1, standard deviation = 0.47).
- **Trait_Anxiety**: Measured on a scale from 0 to 5, generated with a normal distribution (mean = 2.5, standard deviation = 0.56). Trait anxiety refers to a consistent tendency to experience anxiety across various situations.
- **Fear_of_Crime**: Measured on a scale from 0 to 5, generated with a normal distribution (mean = 2.6, standard deviation = 0.65).
- **Crime_Victim**: Indicates whether a participant has experienced being a victim of crime, with options "Yes" or "No."
- **Sex**: "Male" or "Female."
- **Country**: Categorized as "England," "Scotland," "Wales," or "NI" (Northern Ireland).

Key components of the analysis:

1. Comparison of State Anxiety Between Countries: We will split the dataset into two categories England and Scotland to compare the levels of state anxiety between these two countries.
2. Gender-Based Anxiety Comparison: Between England and Scotland, we will compare state anxiety levels between males and females to investigate if there is a significant difference based on gender.
3. Overall Correlation Analysis: We will perform different correlation analysis to examine the relationships between all variables.

Data Exploration

Dimensionality of the dataset:

```r
dim(crime)
```

```
## [1] 500   8
```

Finding the datatypes:

```r
str(crime)
```

```
## ’data.frame':    500 obs. of  8 variables:
##  $ participant  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age          : int  55 31 81 68 34 47 83 68 74 78 ...
##  $ sex          : chr  "Female" "Male" "Female" "Male" ...
##  $ crime_victim : chr  "Yes" "No" "Yes" "No" ...
##  $ state_anxiety: num  1.71 2.86 1.53 2.18 1.87 ...
##  $ trait_anxiety: num  2.9 2.19 2.55 3.17 2.38 ...
##  $ fear_of_crime: num  2.04 1.89 2.58 3.53 2.2 ...
##  $ country      : chr  "Scotland" "England" "England" "NI" ...
```

Summary statistics:

```r
summary(crime)
```

```
##   participant         age             sex              crime_victim
##  Min.   :  1.0   Min.   :17.00   Length:500         Length:500
##  1st Qu.:125.8   1st Qu.:35.00   Class :character   Class :character
##  Median :250.5   Median :51.00   Mode  :character   Mode  :character
##  Mean   :250.5   Mean   :51.58
##  3rd Qu.:375.2   3rd Qu.:69.00
##  Max.   :500.0   Max.   :85.00
##  state_anxiety    trait_anxiety   fear_of_crime      country
##  Min.   :0.5668   Min.   :0.658   Min.   :0.6424   Length:500
##  1st Qu.:1.7947   1st Qu.:2.130   1st Qu.:2.1697   Class :character
##  Median :2.0951   Median :2.513   Median :2.6005   Mode  :character
##  Mean   :2.1067   Mean   :2.501   Mean   :2.6210
##  3rd Qu.:2.4332   3rd Qu.:2.918   3rd Qu.:3.1078
##  Max.   :3.4442   Max.   :4.581   Max.   :4.5138
```

To retrieve additional summary statistics of numerical variables: age, state_anxiety, trait_anxiety, fear_of_crime Interquartile range (IQR)

```r
IQR(crime$age)
```

```
## [1] 34
```

```r
IQR(crime$state_anxiety)
```

```
## [1] 0.6385284
```

```r
IQR(crime$trait_anxiety)
```

```
## [1] 0.7879342
```

```r
IQR(crime$fear_of_crime)
```

```
## [1] 0.9381367
```

Variance

```r
var(crime$age)
```

```
## [1] 385.4545
```

```r
var(crime$state_anxiety)
```

```
## [1] 0.2228619
```

```r
var(crime$trait_anxiety)
```

```
## [1] 0.3385982
```

```r
var(crime$fear_of_crime)
```

```
## [1] 0.4021306
```

Standard deviation (SD)

```r
sd(crime$age)
```

```
## [1] 19.633
```

```r
sd(crime$state_anxiety)
```

```
## [1] 0.4720825
```

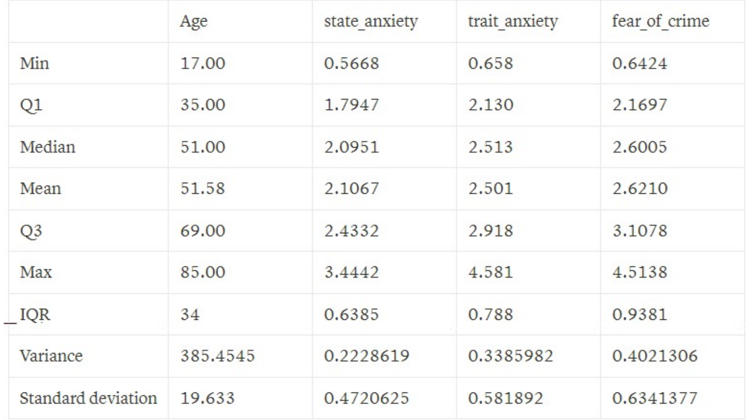
```r
sd(crime$trait_anxiety)
```

```
## [1] 0.581892
```

```r
sd(crime$fear_of_crime)
```

```
## [1] 0.6341377
```

We have created a table in which we display all the summary statistics in order to have a good overview.



	Age	state_anxiety	trait_anxiety	fear_of_crime
Min	17.00	0.5668	0.658	0.6424
Q1	35.00	1.7947	2.130	2.1697
Median	51.00	2.0951	2.513	2.6005
Mean	51.58	2.1067	2.501	2.6210
Q3	69.00	2.4332	2.918	3.1078
Max	85.00	3.4442	4.581	4.5138
_ IQR	34	0.6385	0.788	0.9381
Variance	385.4545	0.2228619	0.3385982	0.4021306
Standard deviation	19.633	0.4720625	0.581892	0.6341377

Data Cleaning

To clean the data we performed the following steps: 1. Handling Missing Values: Decide on strategies to handle missing data. However, there are no missing values.

```r
sum(is.na(crime))
```

```
## [1] 0
```

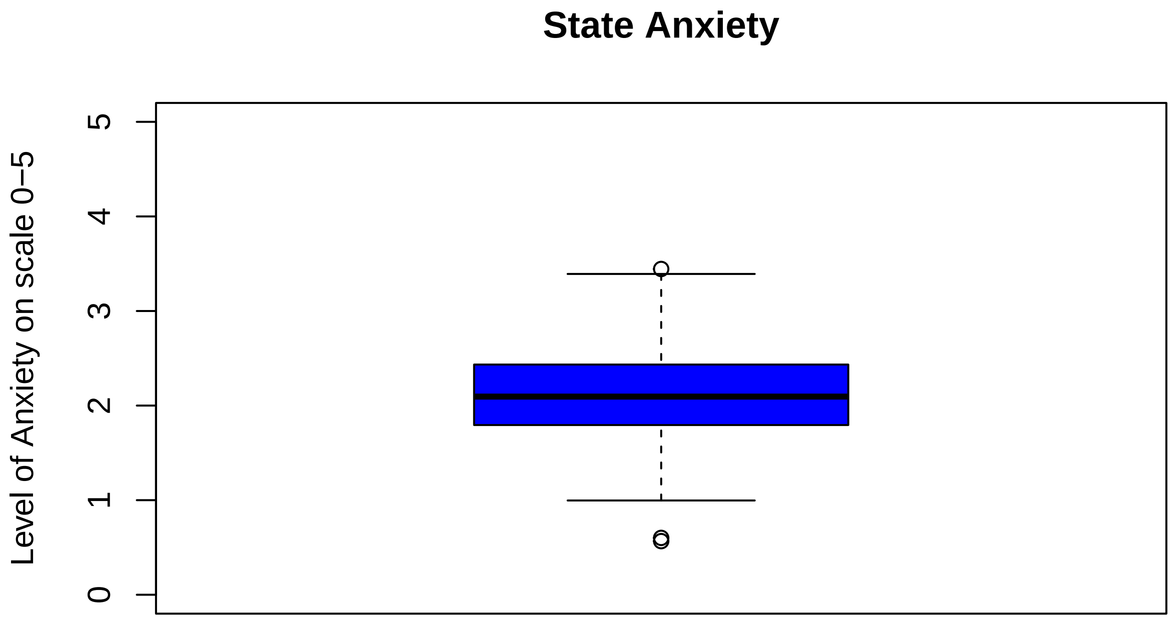
2. Checking for duplicates: There are no duplicates.

```r
sum(duplicated(crime))
```

```
## [1] 0
```

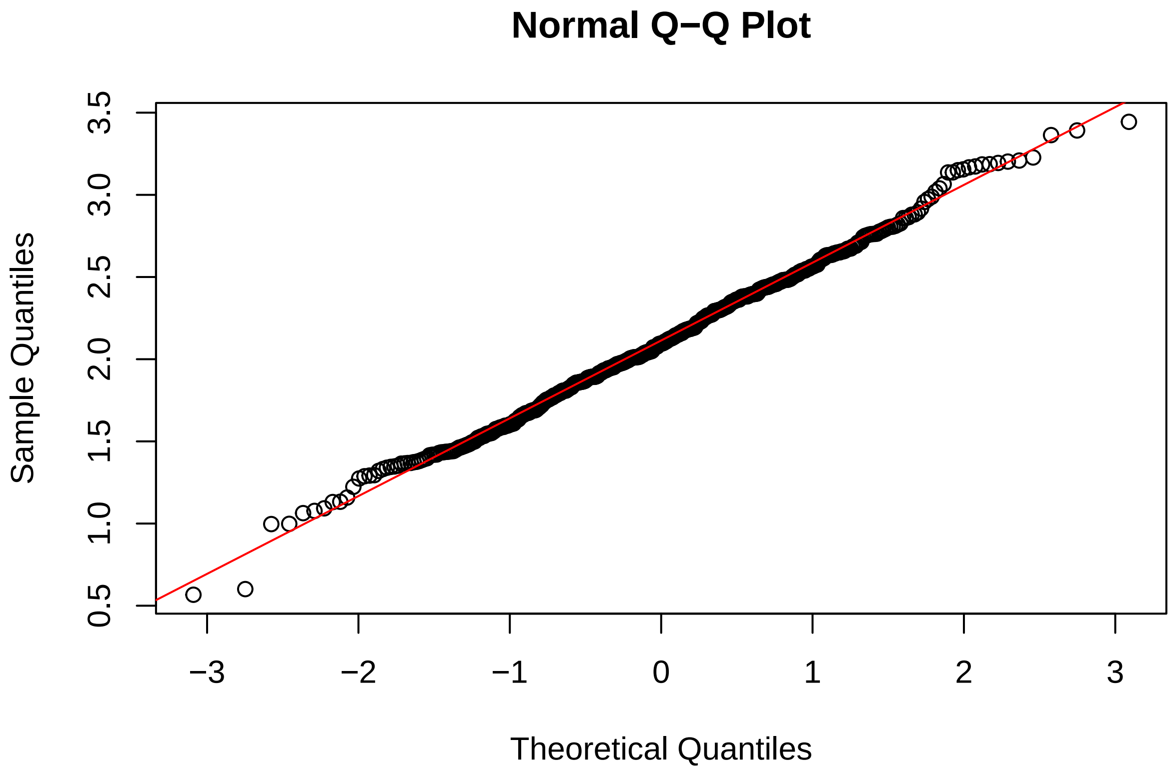
3. Outlier Detection: Identify and decide how we will handle outliers. The boxplot shows that there are a few outliers, but these do not change our summary statistic significantly. Moreover,they are within our scale (0-5).

```r
boxplot(crime$state_anxiety, col='blue', names='Anxiety',
        main='State Anxiety', ylab='Level of Anxiety on scale 0-5', ylim=c(0, 5))
```



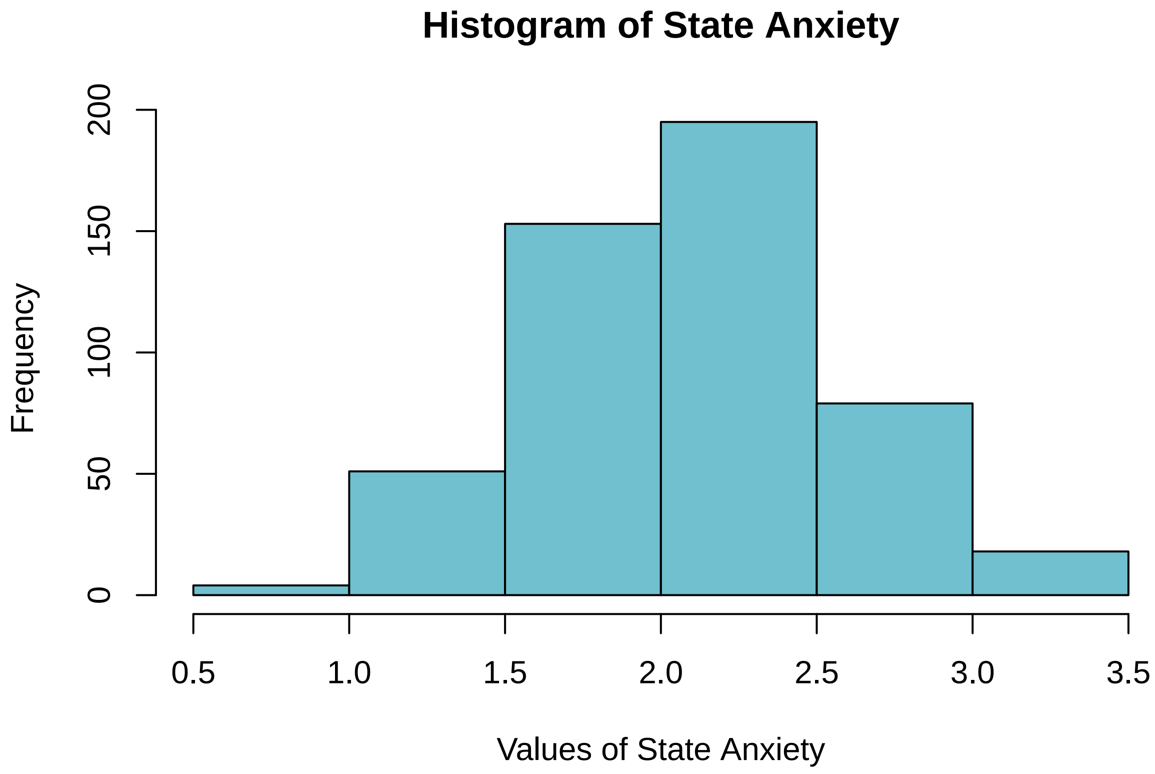
4. Data Transformation: Normalize or standardize data if necessary, especially for variables that will be used in hypothesis testing. From the QQ plot and the histogram we can conclude that the state anxiety of our sample is normally distributed.

```
qqnorm(crime$state_anxiety)
qqline(crime$state_anxiety,col="red")
```



5. Visualizing the data with a histogram:

```
hist(crime$state_anxiety,
     col='#70c0cf',           # Set the color of bars to blue
     main='Histogram of State Anxiety',     # Set the main title of the histogram
     xlab='Values of State Anxiety',        # Set the label for the x-axis
     ylab='Frequency',     # Set the label for the y-axis
     border='black',       # Set the color of the border of the bars
     )
```

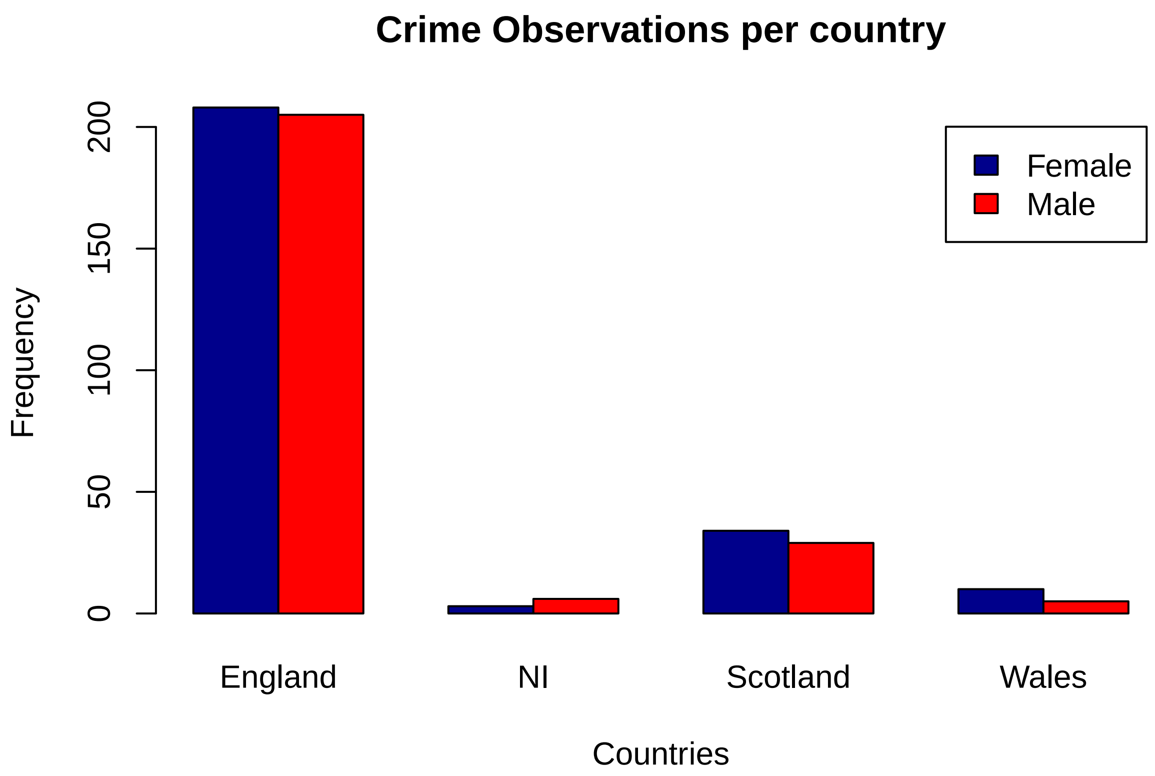


Our data was already normally distributed, so we do not have to normalize it. We will standardize the data while doing the z-test.

Separating the Data

The following barplot shows more detailed information about the countries and sex and their level of state anxiety.

```
counts <- table(crime$sex , crime$country)
barplot(counts, main="Crime Observations per country",
  xlab="Countries", col=c("darkblue","red"),
  ylab= "Frequency",
  legend = rownames(counts), beside=TRUE)
```



We have chosen to compare England with Scotland in the hypothesis testing, because these two countries give us the most observations to compare. Both samples have large (>30), so we can perform the two sample z-test. We separate the data set into two subsets. England and Scotland. We exclude Wales and Northern Ireland, because there is very little observations available for these countries.

```
england <- subset(crime, country == "England")
mean(england$state_anxiety)
```

```
## [1] 2.11919
```

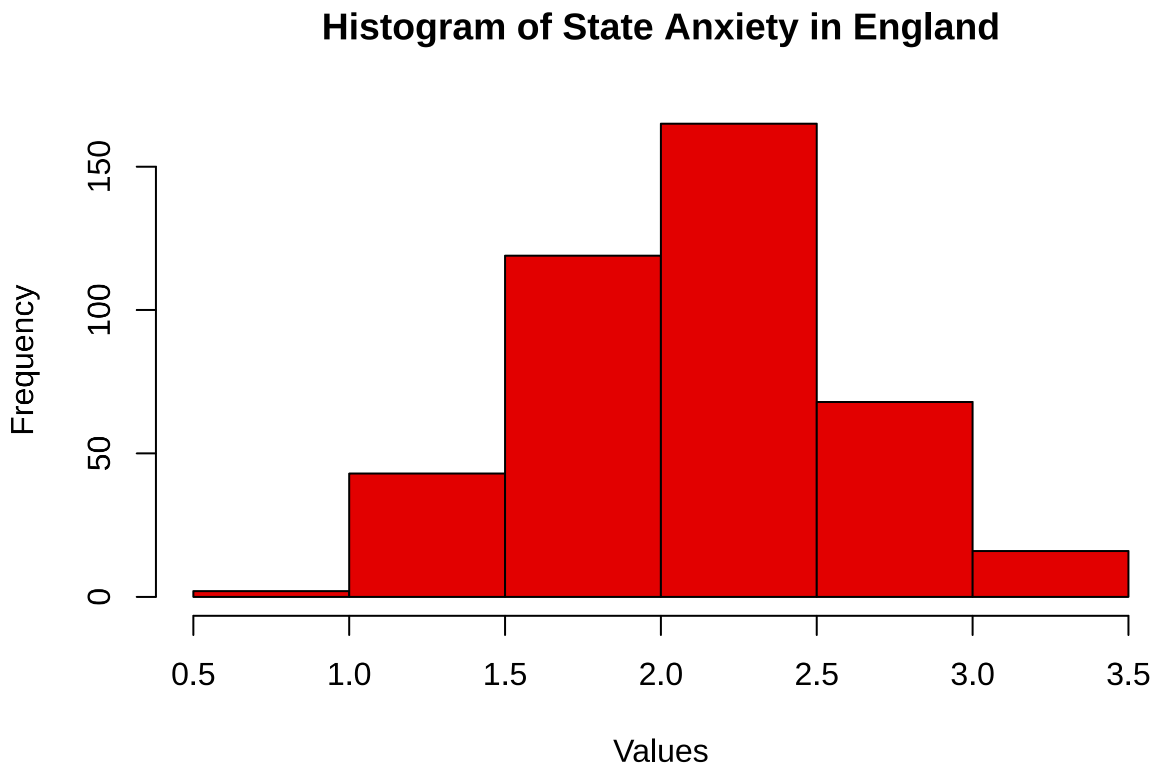
```
scotland <- subset(crime, country == "Scotland")
mean(scotland$state_anxiety)
```

```
## [1] 1.976453
```

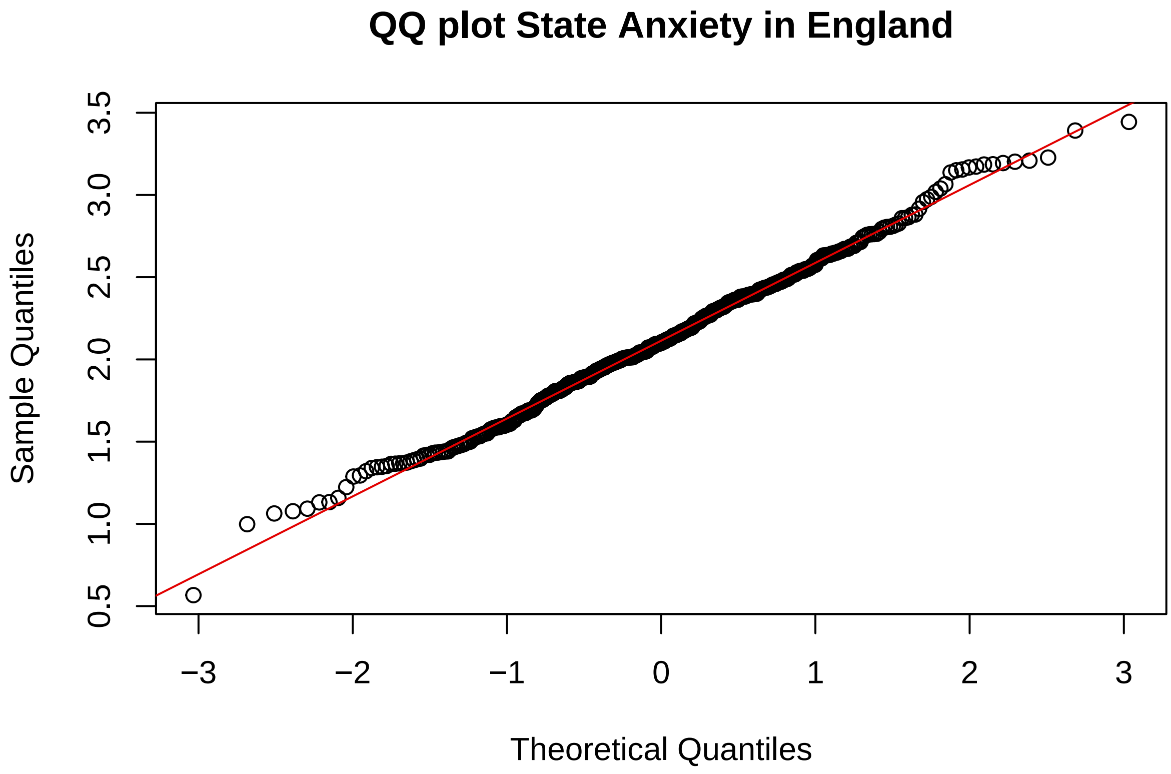
Data Visualization of England and Scotland

We would like to know what the distribution of the England group and the Scotland group is respectively.
England:

```
hist(england$state_anxiety,
     col='#E20000',            # Set the color of bars to blue
     main='Histogram of State Anxiety in England',      # Set the main title of the histogram
     xlab='Values',         # Set the label for the x-axis
     ylab='Frequency',      # Set the label for the y-axis
     border='black',        # Set the color of the border of the bars
     #breaks=seq(0, 60, by=5)  # Set the breaks for the histogram bins
     )
```



```
qqnorm(england$state_anxiety, main='QQ plot State Anxiety in England')
qqline(crime$state_anxiety,col='#E20000')
```



Summary statistics England:

```
summary(england$state_anxiety)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5668  1.8087  2.1003  2.1192  2.4352  3.4442
```

```
IQR(england$state_anxiety)
```

```
## [1] 0.6265297
```

```
var(england$state_anxiety)
```

```
## [1] 0.2223119
```

```
sd(england$state_anxiety)
```

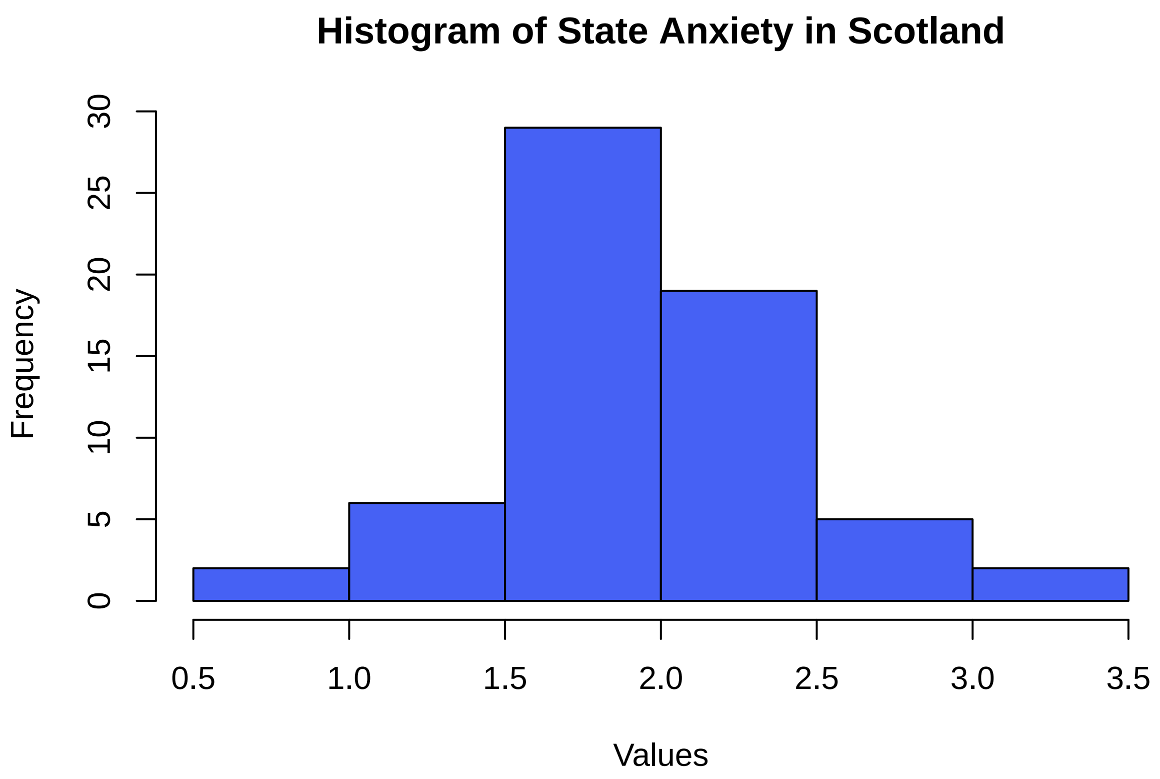
```
## [1] 0.4714996
```

```
nrow(england)
```

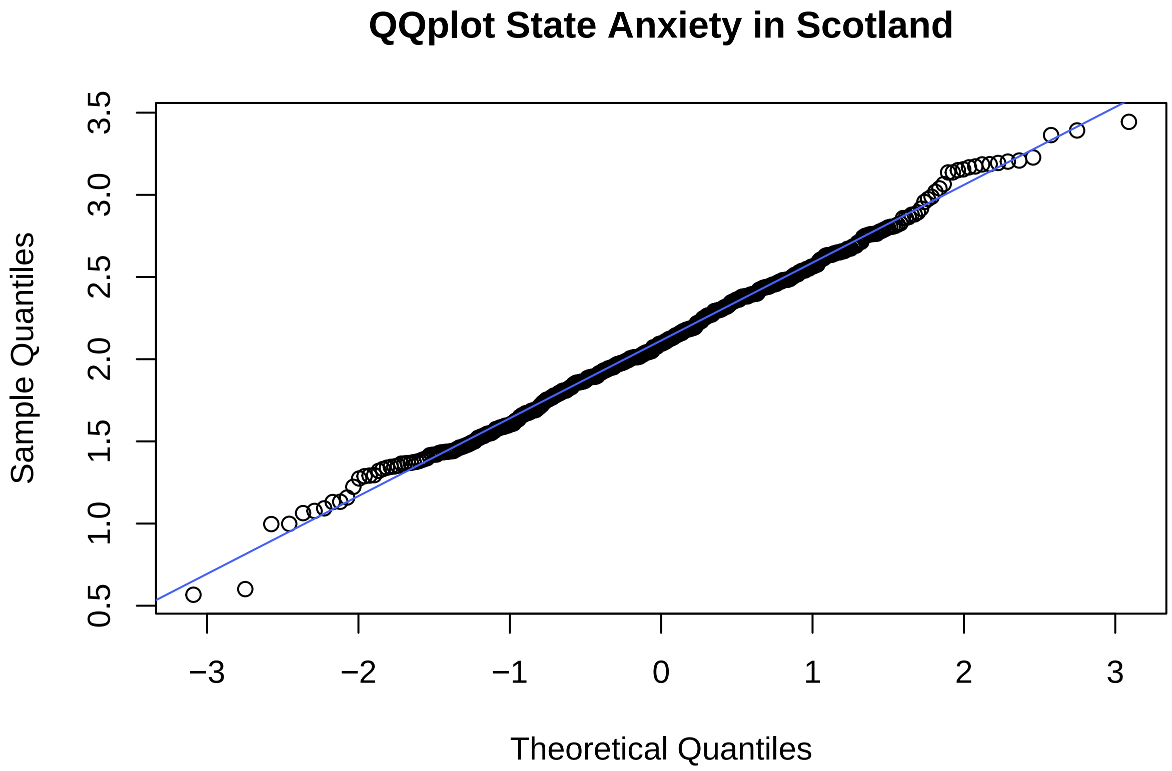
```
## [1] 413
```

Scotland:

```r
hist(scotland$state_anxiety,
     col='#4661F4',           # Set the color of bars to blue
     main='Histogram of State Anxiety in Scotland',     # Set the main title of the histogram
     xlab='Values',         # Set the label for the x-axis
     ylab='Frequency',      # Set the label for the y-axis
     border='black',        # Set the color of the border of the bars
     #breaks=seq(0, 60, by=5)  # Set the breaks for the histogram bins
     )
```



```r
qqnorm(crime$state_anxiety, main='QQplot State Anxiety in Scotland')
qqline(crime$state_anxiety,col='#4661F4')
```



Summary statistics Scotland:

```
summary(scotland$state_anxiety)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6011  1.6818  1.9443  1.9765  2.2717  3.3636
```

```
IQR(scotland$state_anxiety)
```

```
## [1] 0.5899529
```

```
var(scotland$state_anxiety)
```

```
## [1] 0.2330864
```

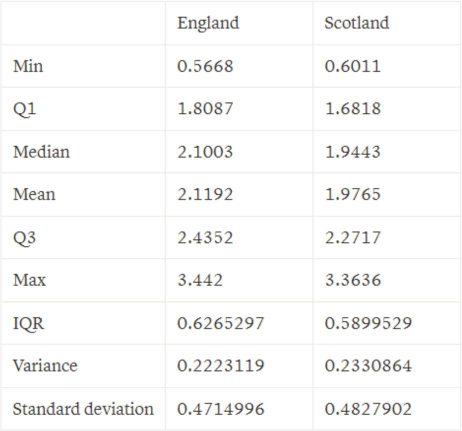
```
sd(scotland$state_anxiety)
```

```
## [1] 0.4827902
```

```
nrow(scotland)
```

```
## [1] 63
```

The code shown above give us the following information about England and Scotland.



	England	Scotland
Min	0.5668	0.6011
Q1	1.8087	1.6818
Median	2.1003	1.9443
Mean	2.1192	1.9765
Q3	2.4352	2.2717
Max	3.442	3.3636
IQR	0.6265297	0.5899529
Variance	0.2223119	0.2330864
Standard deviation	0.4714996	0.4827902

Hypothesis Testing

We will now use the obtained subsets (England, Scotland), to perform a hypothesis test. We would like to do an independent two-sample z-test, because the sample size in large and the groups are independent of each other.

Question: *Is there a higher state anxiety in England or in Scotland? Test for* $\alpha = 0.05$ Null and Alternative Hypothesis: H_0: The mean of the state_anxiety in England is equal to the mean of the state anxiety in Scotland. H_a: The mean of the state_anxiety in England is **not** equal to the mean of the state anxiety in Scotland.

Performing the Z independent two sample test:

```r
#group 1= england
#group 2= scotland

z_test_two_sample <- function(sample1, sample2, population_sd1, population_sd2) {
  n1 <- length(sample1)
  n2 <- length(sample2)
  mean1 <- mean(sample1)
  mean2 <- mean(sample2)
  pooled_sd <- sqrt(population_sd1^2/n1 + population_sd2^2/n2)
  z <- (mean1 - mean2) / pooled_sd
  p_value <- 2 * pnorm(-abs(z))  # Two-tailed test
  return(list(z = z, p_value = p_value))
}

z_test_two_sample(england$state_anxiety,scotland$state_anxiety,0.471, 0.4828)
```

```
## $z
## [1] 2.192826
##
## $p_value
## [1] 0.02831989
```

$p < \alpha \iff 0.0283 < 0.05$ We reject the hypothesis; there is a significant difference between the mean of state anxiety in England and that of Scotland. We can say that the mean state anxiety in England is higher than that of Scotland. $\mu_{eng} = 2.1192$ $\mu_{scot} = 1.9765$

Correlation and Regression Model

We would like to find a correlation between the variables. We start by checking the correlation between age and state anxiety. We can see below that there is no linear correlation between the two.

```
cor(crime$age, crime$state_anxiety)
```

```
## [1] -0.03249081
```

We therefore try to find a correlation between age and the fear of crime and the correlation between the state anxiety and the fear of crime. We finally made a correlation matrix, because we were curious to see if any variables are correlated. None of these calculations show a strong or even a weak correlation.

```
cor(crime$age, crime$fear_of_crime)
```

```
## [1] -0.01519112
```

```
cor(crime$state_anxiety, crime$fear_of_crime)
```

```
## [1] 0.03441303
```

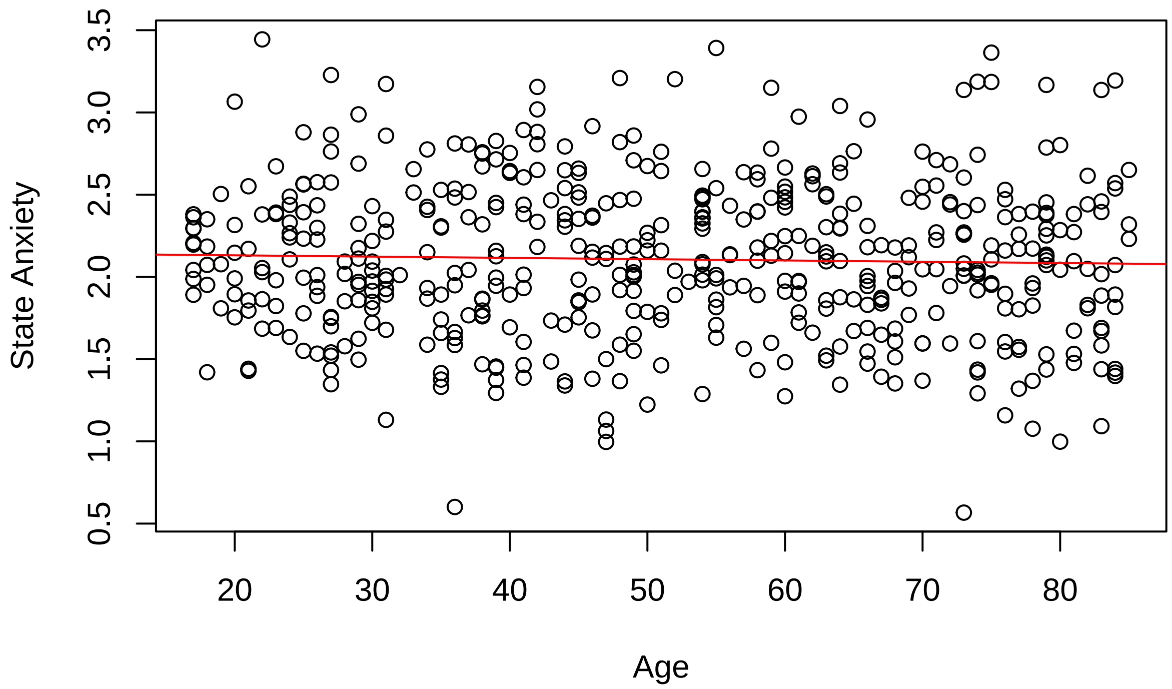
```
sub_crime = subset(crime, select = c(age, state_anxiety, trait_anxiety, fear_of_crime ))

#correlation matrix
cor_matrix=cor(sub_crime)
print(cor_matrix)
```

```
##                        age state_anxiety trait_anxiety fear_of_crime
## age            1.000000000  -0.032490813  -0.004443002  -0.015191125
## state_anxiety -0.032490813   1.000000000   0.007360813   0.034413029
## trait_anxiety -0.004443002   0.007360813   1.000000000  -0.006889861
## fear_of_crime -0.015191125   0.034413029  -0.006889861   1.000000000
```

To come back to the variables age and state anxiety, we create a best fit line for these two variables.

```
model <- lm(crime$state_anxiety ~ crime$age, data = crime)
plot(crime$state_anxiety ~ crime$age, data = crime, xlab="Age", ylab="State Anxiety")
abline(model, col = "red")
```



```
summary(model)
```

```
##
## Call:
## lm(formula = crime$state_anxiety ~ crime$age, data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52314 -0.31677 -0.00877  0.31462  1.31440
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1469615  0.0594284  36.127   <2e-16 ***
## crime$age   -0.0007813  0.0010769  -0.725    0.469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4723 on 498 degrees of freedom
## Multiple R-squared:  0.001056,   Adjusted R-squared:  -0.0009503
## F-statistic: 0.5263 on 1 and 498 DF,  p-value: 0.4685
```

Formula for best fit line: $\hat{y} = 2.147 - 0.00078x$ The slope is almost 0, which means that an increase in age does not increase nor decrease the state anxiety. The y intercept is 2.147.

Chi square test

Question: *Can we conclude that there is a relationship between the sex and whether you are a crime victim or not? Test for* $\alpha = 0.05$ Null and Alternative Hypothesis:

H_0: There is no relationship between sex and whether you are a crime victim.

H_1: There is a relationship between sex and whether you are a crime victim.

Observed table:

```
table(crime$sex, crime$crime_victim)
```

```
##
##            No Yes
##   Female 142 113
##   Male   128 117
```

Chi square test:

```
result <- chisq.test(crime$sex, crime$crime_victim , correct=FALSE)
result
```

```
##
##  Pearson's Chi-squared test
##
## data:  crime$sex and crime$crime_victim
## X-squared = 0.59573, df = 1, p-value = 0.4402
```

$p > \alpha \iff 0.4402 > 0.05$ shows that we can not reject the null hypothesis. There is no relationship between the two variables/ the variables are independent.

Expected table:

```
result$expected
```

```
##           crime$crime_victim
## crime$sex    No   Yes
##    Female 137.7 117.3
##    Male   132.3 112.7
```

PCA

We have 4 numerical variables, so it is better to simplify and plot the components with the highest or most significant variance. With the PCA we can reduce the dimensions of our dataset and with that we are able to simplify our analysis.

```
normalized <- scale(sub_crime)
pca_data <- princomp(normalized)
summary(pca_data)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4
## Standard deviation     1.026458 1.0010184 0.9898940 0.9779854
## Proportion of Variance 0.263932 0.2510115 0.2454634 0.2395931
## Cumulative Proportion  0.263932 0.5149435 0.7604069 1.0000000
```

```
pca_data$loadings[,1:3]
```

```
##                    Comp.1      Comp.2    Comp.3
## age             0.53123731  0.17725546 0.7359353
## state_anxiety -0.65072341 -0.06321215 0.0969710
## trait_anxiety -0.06166301 -0.91544467 0.3563185
## fear_of_crime -0.53902100  0.35573274 0.5674795
```

```
#install.packages("factoextra")
library(factoextra)
```

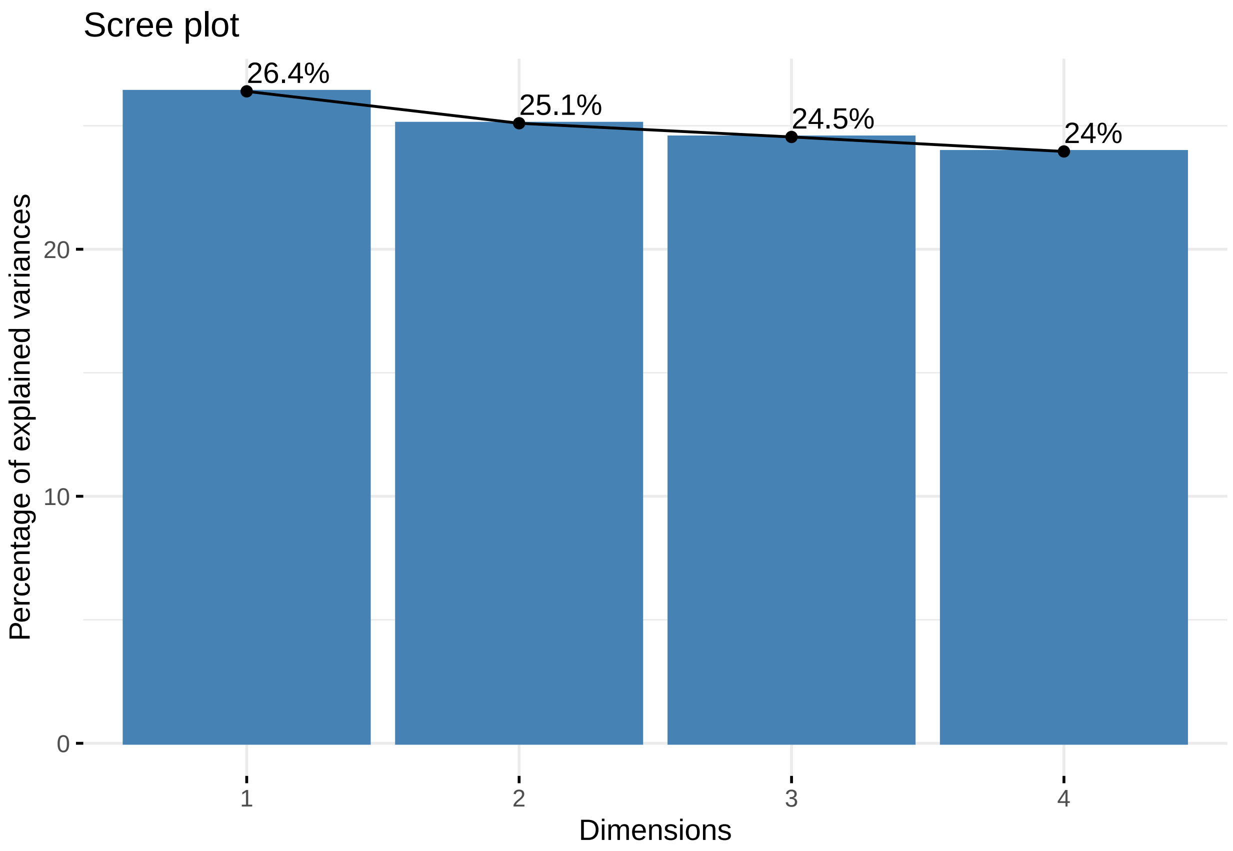
```
## Warning: package 'factoextra' was built under R version 4.3.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.3

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_eig(pca_data, addlabels = TRUE)
```



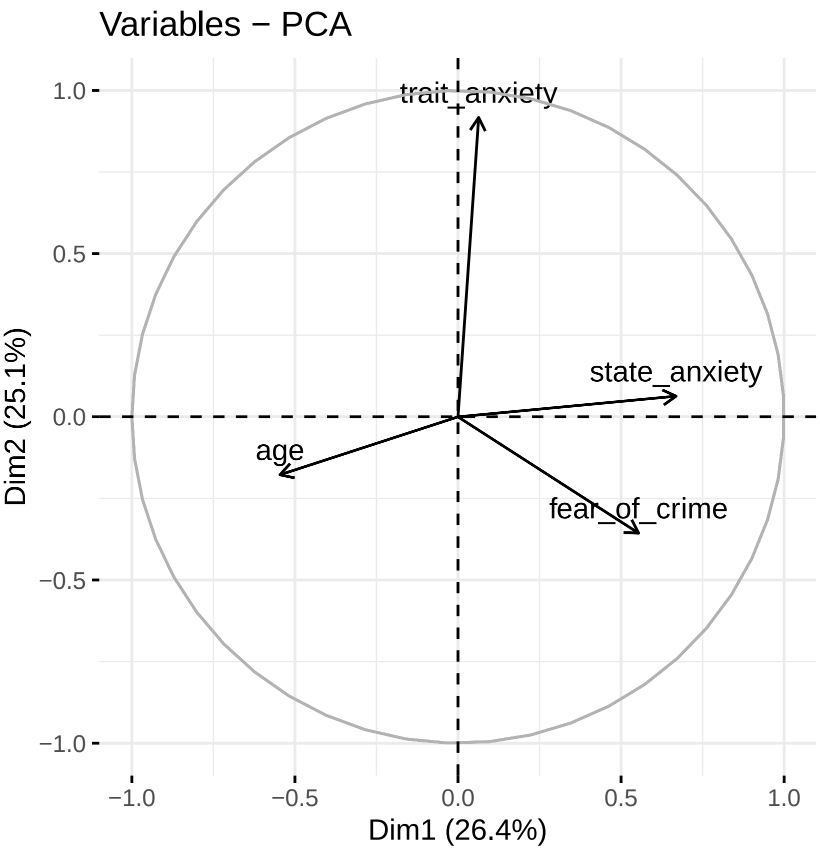
We need at least 85% to sufficiently represent the variances of the data. This would mean in our case that we would have to include PC1, PC2, PC3 and PC4. Therefore reducing the dimensions, would not be an accurate representation of the dataset.

Despite that we still tried to plot the PCA biplot with PC1 and PC2. From the biplot we can see that relation between state anxiety and fear of crime show a relatively strong positive relation.

```
library(FactoMineR)
```

```
## Warning: package ’FactoMineR’ was built under R version 4.3.3
```

```
pca_result <- PCA(normalized, graph = FALSE)
fviz_pca_var(pca_result, col.var = "black")
```



Conclusion

In this analysis, we did not find any noteworthy or interesting correlations among the variables. However, we did discover that the average state anxiety is higher in England than in Scotland. We examined the data and applied various tests that we learned in class.