Project 2

Cost of Living : Catherine Nader, Julia Joseph

1. Introduction:

We chose one dataset from Kaggle (https://www.kaggle.com/datasets/orhankaramancode/city-quality-of-life-dataset) which contains information on housing, cost of living, safety, healthcare, education, environmental quality, economy, taxation, internet access, leisure and culture, and outdoors. The values are based on the cost of living in different cities. Each row represents a city. Of the 21 variables, there are 18 numerical variables and 3 categorical variables. There are 266 total observations and we would like to explore the correlation between certain numerical variables and how they correlate with one another. The outcome variable is the cost of living.

Some trends we expect to see is we expect there to be a positive correlation with housing and cost of living. Additionally, we also expect to see a positive correlation between venture capitalism and start-ups as well as education and cost of living. Finally, we expect that as the score for most of the predictors go up, that the cost of living will go up. This is because we expect the numbers to follow that higher quality infrastructure will generally cost more.

Research Questions:

1. Is there a positive correlation between venture capital and start-ups?
2. Does housing impact the cost of living?

Let's first load the `tidyverse` package that contains `tidyr`, `dplyr` and `ggplot2`, and `tidytext`, and `textdata`, which we will use in this report.

```
# Load packages
library(tidyverse)
library(tidytext)
library(textdata)
library(readr)
library(dplyr)
library(cluster)
library(ggcorrplot)
library(ggplot2)
library(ade4)
library(factoextra)
library(plotROC)
library(caret)
library(rpart)
library(rpart.plot)
```

```
# loading in dataset from csv files from Kaggle
costofliving <- read_csv("cityquality.csv")
costofliving
```

```
## # A tibble: 266 x 21
##     ...1 UA_Name UA_Co~1 UA_Co~2 Housing Cost ~3 Start~4 Ventu~5 Trave~6 Commute
##    <dbl> <chr>   <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1     0 Aarhus  Denmark Europe     6.13    4.02    2.83    2.51   3.54     6.31
```

```
##  2     1 Adelai~ Austra~ Oceania    6.31    4.69    3.14    2.64   1.78     5.34
##  3     2 Albuqu~ New Me~ North ~    7.26    6.06    3.77    1.49   1.46     5.06
##  4     3 Almaty  Kazakh~ Asia       9.28    9.33    2.46    0      4.59     5.87
##  5     4 Amster~ Nether~ Europe     3.05    3.82    7.97    6.11   8.32     6.12
##  6     5 Anchor~ Alaska  North ~    5.43    3.14    2.79    0      1.74     4.72
##  7     6 Andorra Andorra Europe     3.97    0       1       0      0.5      0
##  8     7 Ankara  Turkey  Asia       9.93    9.12    3.97    0      2.05     5.29
##  9     8 Ashevi~ North ~ North ~    5.86    5.31    3.54    0      1.21     1.36
## 10     9 Asunci~ Paragu~ South ~    9.23    9.30    2.21    0      0.645    4.97
## # ... with 256 more rows, 11 more variables: `Business Freedom` <dbl>,
## #   Safety <dbl>, Healthcare <dbl>, Education <dbl>,
## #   `Environmental Quality` <dbl>, Economy <dbl>, Taxation <dbl>,
## #   `Internet Access` <dbl>, `Leisure & Culture` <dbl>, Tolerance <dbl>,
## #   Outdoors <dbl>, and abbreviated variable names 1: UA_Country,
## #   2: UA_Continent, 3: `Cost of Living`, 4: Startups, 5: `Venture Capital`,
## #   6: `Travel Connectivity`
```

Tidying

We decided to drop UA_Name, UA_Continent, and UA_Country since these were all categorical variables. The reason behind this is we just wanted to focus on the relationships between the numerical variables and how certain costs of living affect one another. Before dropping certain variables and all the possible NAs in our data set, there were 21 columns and 266 rows. After dropping the three categorical variables, there were 18 columns and still 266 rows. Once this was done, we decided to make it easier throughout our project to clean up some of the variable's names by taking out the back tick marks. Once this was done, we checked whether this worked using the head function.

```
nrow(costofliving) # finding how many rows there are
```

```
## [1] 266
```

```
ncol(costofliving) # finding how many columns there are
```

```
## [1] 21
```

```
# dropping variables and possible NA values
costofliving_drop <- subset(costofliving,
                            select = -c(UA_Name, UA_Continent, UA_Country)) %>%
  drop_na()

nrow(costofliving_drop) # finding how many rows there are after selected dropping
```

```
## [1] 266
```

```
ncol(costofliving_drop) # finding how many columns there are after selected dropping
```

```
## [1] 18
```

```
# renamed certain variables
costofliving_drop <- costofliving_drop %>%
  rename(CostofLiving = `Cost of Living`, VentureCapital = `Venture Capital`, TravelConnectivity = `Trav

# made sure tidying worked
head(costofliving_drop)
```

```
## # A tibble: 6 x 18
##    ...1 Housing CostofL~1 Start~2 Ventu~3 Trave~4 Commute Busin~5 Safety Healt~6
##   <dbl>   <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
```

```
## 1     0    6.13      4.02    2.83    2.51    3.54    6.31    9.94   9.62    8.70
## 2     1    6.31      4.69    3.14    2.64    1.78    5.34    9.40   7.93    7.94
## 3     2    7.26      6.06    3.77    1.49    1.46    5.06    8.67   1.34    6.43
## 4     3    9.28      9.33    2.46    0       4.59    5.87    5.57   7.31    4.55
## 5     4    3.05      3.82    7.97    6.11    8.32    6.12    8.84   8.50    7.91
## 6     5    5.43      3.14    2.79    0       1.74    4.72    8.67   3.47    6.06
## # ... with 8 more variables: Education <dbl>, env_quality <dbl>, Economy <dbl>,
## #   Taxation <dbl>, internet_access <dbl>, leisure_culture <dbl>,
## #   Tolerance <dbl>, Outdoors <dbl>, and abbreviated variable names
## #   1: CostofLiving, 2: Startups, 3: VentureCapital, 4: TravelConnectivity,
## #   5: BusinessFreedom, 6: Healthcare
```

2. Exploratory Data Analysis:

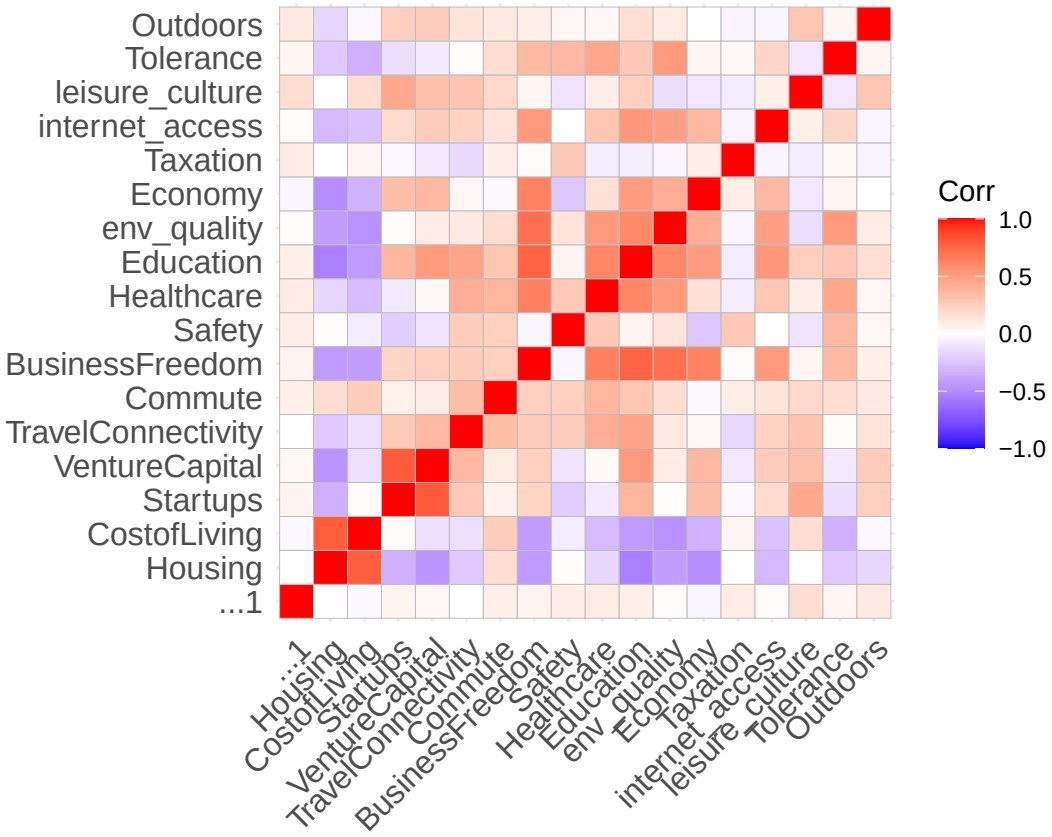
Let's explore the correlation matrix. The most correlated variables are the cost of living and housing this is because these two variables are colored in deep red which indicates almost a 1 to 1 correlation. Another highly correlated relationship is between venture capital and start-ups. This is highly correlated because you can see that these two variables' relationship with each other is very red which indicates that there is a highly correlated relationship with one another. Additionally, the least correlated is environmental quality and cost of living as is indicated by a very purple-colored square. The purple-colored square represents a -1 correlation which means that they are fair from being correlated. Education and Housing also have a nearly -1 correlation as it is shown by a very purple square in the correlation matrix.

```r
# computes the correlation matrix between the columns
costofliving_matrix <- cor(costofliving_drop,
# missing values are handled using pairwise deletion
                           use = "pairwise.complete.obs")

costofliving_matrix
```

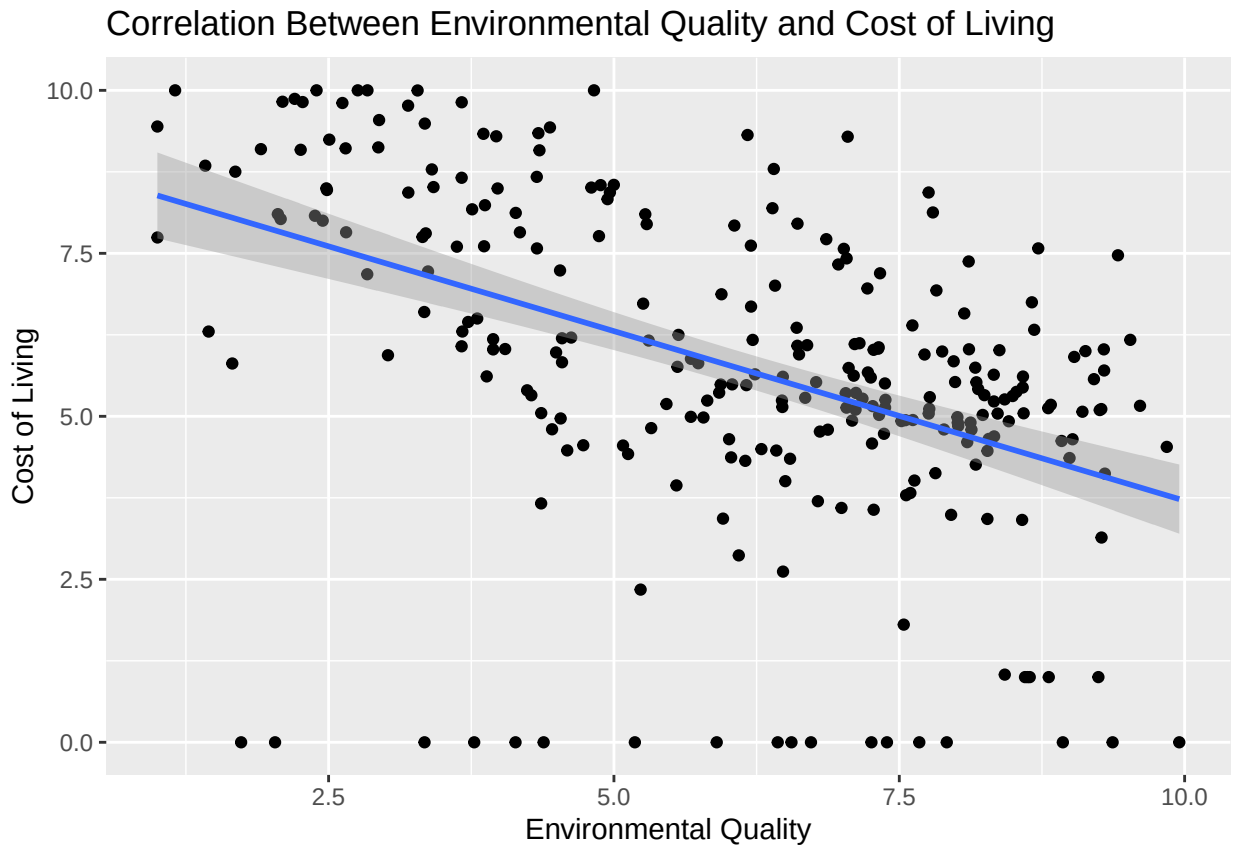
```
##                          ...1      Housing CostofLiving     Startups
## ...1               1.000000000 -0.009493552  -0.02777711  0.060980085
## Housing           -0.009493552  1.000000000   0.79457606 -0.344615583
## CostofLiving      -0.027777110  0.794576064   1.00000000  0.024625439
## Startups           0.060980085 -0.344615583   0.02462544  1.000000000
## VentureCapital     0.044530504 -0.462969086  -0.13226217  0.796213866
##                   VentureCapital TravelConnectivity     Commute
## ...1                  0.04453050       -0.001855841  0.07702965
## Housing              -0.46296909       -0.233816857  0.17600431
## CostofLiving         -0.13226217       -0.134216016  0.25875503
## Startups              0.79621387        0.276168207  0.07489593
## VentureCapital        1.00000000        0.368430813  0.09713324
##                   BusinessFreedom      Safety  Healthcare   Education
## ...1                   0.06321890  0.09128110  0.10195550  0.08315119
## Housing               -0.43498422  0.01174383 -0.16697888 -0.54943267
## CostofLiving          -0.42364535 -0.08033285 -0.29135136 -0.43019702
## Startups               0.21570233 -0.20940913 -0.09499766  0.37623875
## VentureCapital         0.25042940 -0.11973724  0.03285845  0.52415046
##                    env_quality      Economy     Taxation internet_access
## ...1                0.023296819 -0.041486727  0.099252321      0.02345355
## Housing            -0.417732953 -0.490244877 -0.006933262     -0.30174369
## CostofLiving       -0.467666653 -0.327044783  0.050610627     -0.25934850
## Startups            0.009754348  0.342310613 -0.025139530      0.19321226
## VentureCapital      0.103033214  0.373193620 -0.101221565      0.26630874
##                   leisure_culture   Tolerance     Outdoors
```

```
## ...1                  0.183161518  0.05292008  0.112169405
## Housing              -0.008200434 -0.23167091 -0.165504216
## CostofLiving          0.184745767 -0.33600616 -0.030578885
## Startups              0.446967928 -0.14209637  0.253848539
## VentureCapital        0.332003788 -0.09417587  0.269300567
##  [ reached getOption("max.print") -- omitted 13 rows ]
```

```r
# generates a correlation plot
#Graph 1
ggcorrplot(costofliving_matrix)
```

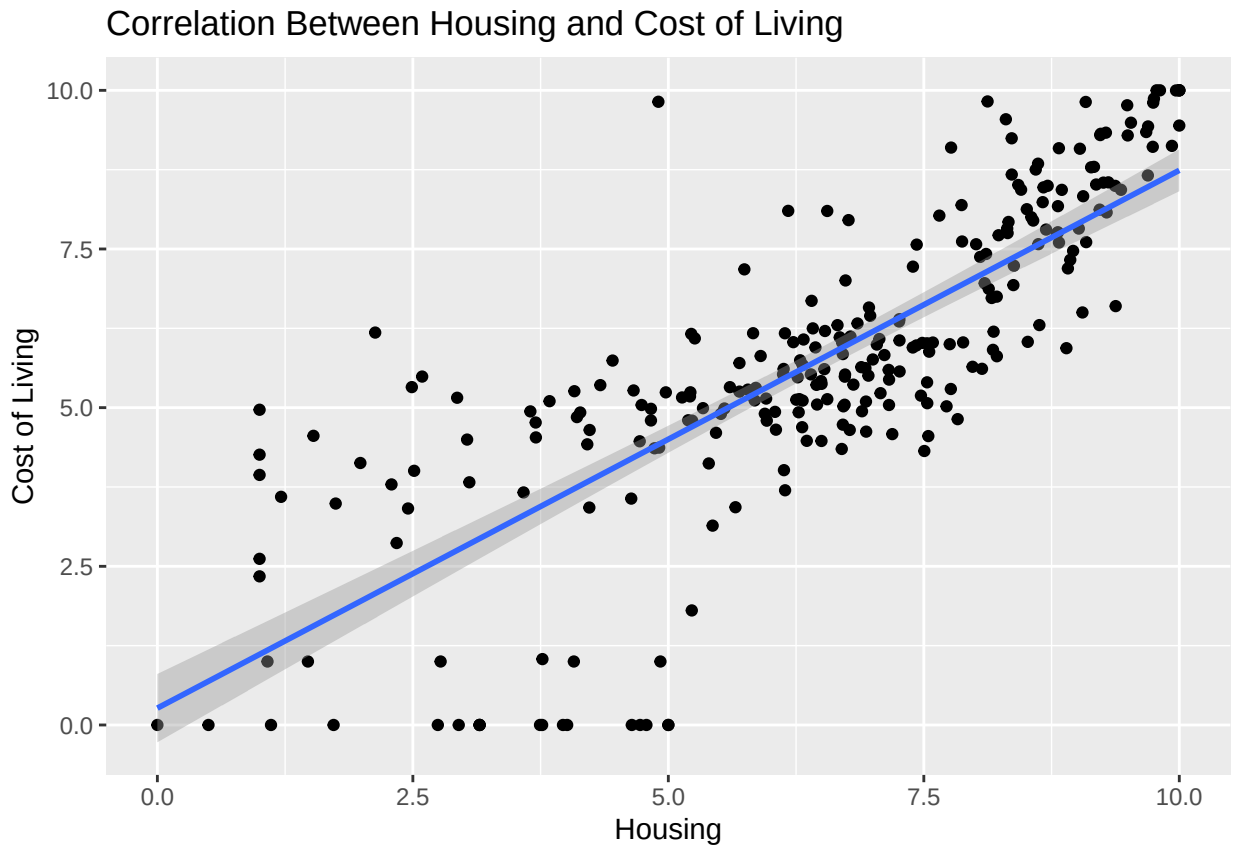



```r
# create scatterplot with line of best fit
# Graph 2
ggplot(costofliving_drop, aes(x = env_quality, y = CostofLiving)) +
  geom_point() +
  geom_smooth(method = "lm") + labs(title = "Correlation Between Environmental Quality and Cost of Livi
```



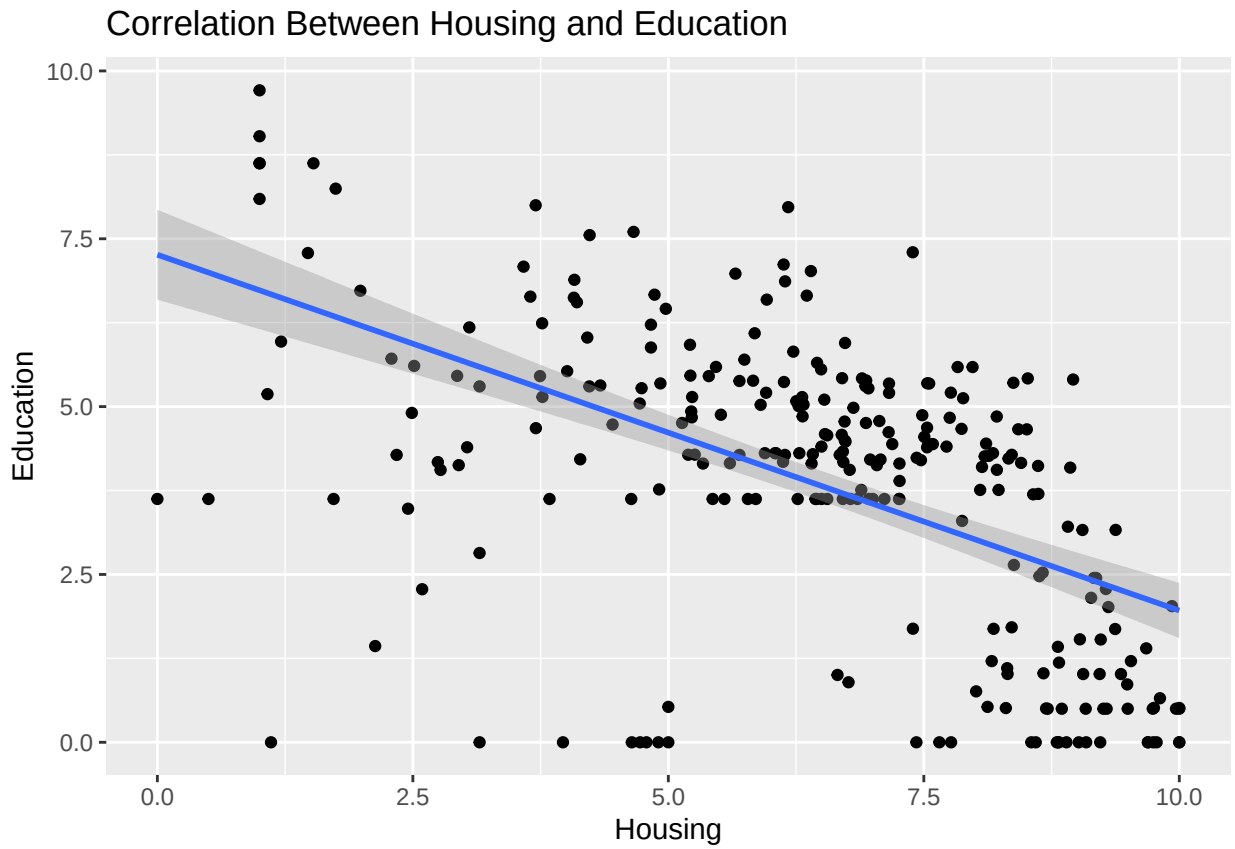
As said earlier, the correlation between environmental quality and cost of living is negatively correlated and is proven to show through a negatively linear trend. This negative correlation can be observed through a negatively sloped linear trend in a scatterplot of the two variables. This indicates that as one variable increases, the other variable tends to decrease. This could be because in a city, there is more pollution, but there is a higher cost of living due to the area.

```
# create scatterplot with line of best fit
# Graph 3
ggplot(costofliving_drop, aes(x = Housing, y = CostofLiving)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = 'Correlation Between Housing and Cost of Living', x = "Housing", y = 'Cost of Living')
```



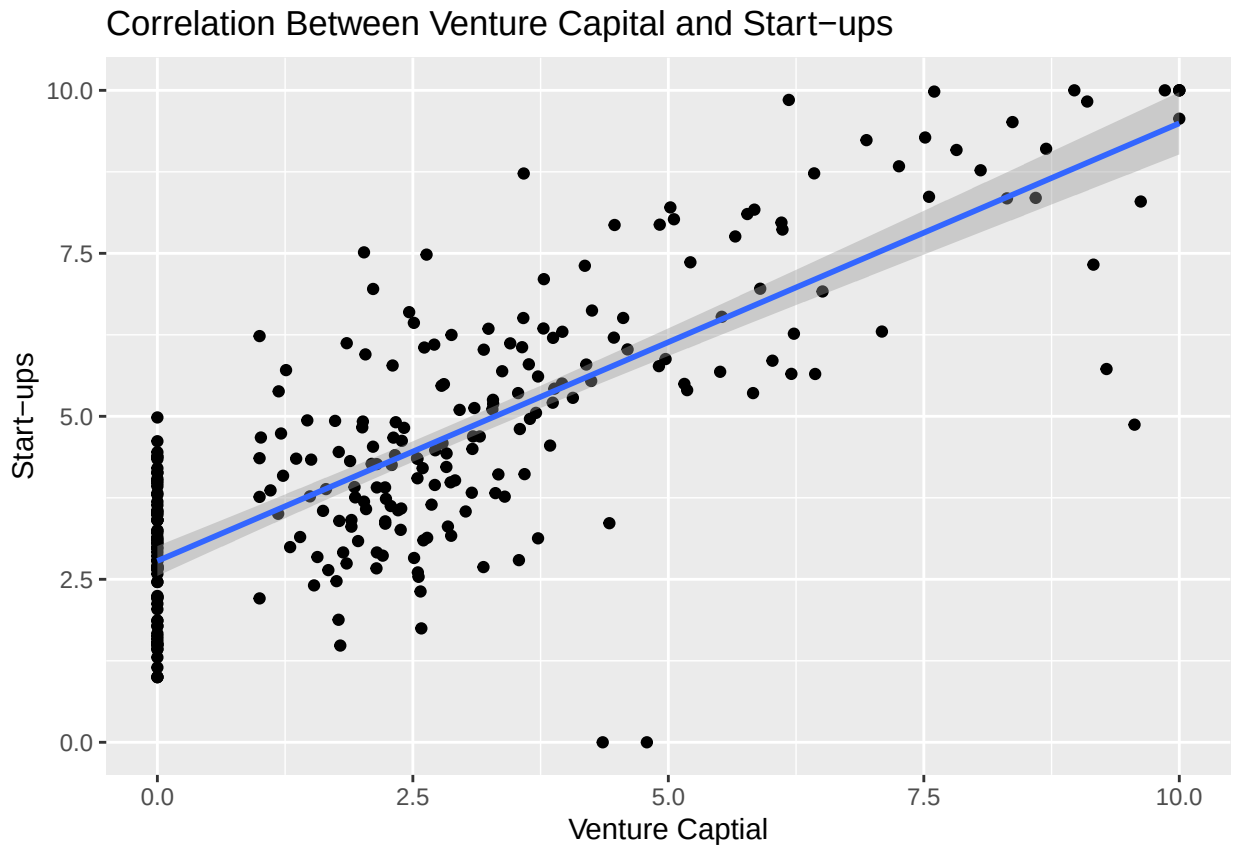
Cost of living and housing a positively correlated, which was previously shown on the correlation matrix. When this is put in a scatterplot with the line of best fit, it shows a positive slope which proves that these two variables are highly correlated with one another. In other words, as the cost of living increases, so does the cost of housing. In the case of cost of living and housing, the positive correlation between the two variable suggest that as high housing costs can increase the cost of living because expenses tend to increase like utilities and electricity.

```
# create scatterplot with line of best fit
# Graph 4
ggplot(costofliving_drop, aes(x = Housing, y = Education)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = 'Correlation Between Housing and Education', x = "Housing", y = "Education")
```



As seen earlier, the correlation matrix showed a negative correlation between housing and education. When this was made into a scatterplot with the line of best fit, there is sufficient evidence to back up the correlation matrix that there is a negative slope between the two variables. A negative correlation between housing and education implies that as one variable increases, the other variable decreases and vice versa. This could imply that as the cost of living rises the quality of education gets lower. It is important to note that correlation does not mean causation. Just because there is a negative correlation does not mean it's because the one variable causes the other.

```
# create scatterplot with line of best fit
# Graph 5
ggplot(costofliving_drop, aes(x = VentureCapital, y = Startups)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = 'Correlation Between Venture Capital and Start-ups', x = "Venture Captial", y = "Start-u
```



The relationship between venture capital and start-ups is a positive slope. This confirms that there is a positive correlation between the two variables. As venture capital increases, so does start-ups. This is shown by the scatterplot and line of best fit. This suggests that in areas where there is more venture capital available, there tend to be more start-ups. Although correlation does not mean causation this could be a plausible reason as to why there is a positive trend between the two variables.

3. Prediction and Cross-Validation:

We performed linear regression on the model by fitting it to the entire dataset and generating predictions of the model. For our evaluation metric, we used RMSE(Root Mean Squared Error) which essentially calculated how much error is in the model's prediction compared to the actual values. Additionally, we calculated how R^2 which is how much variance in the dataset label can be explained by the predictors we chose. Our average RMSE was 1.370307 which is relatively good as the lower the number the better. Our R2 was 0.678 which means that 67.8% of the variance in our data can be explained by the predictors in the model.

```
fit_lin <- lm(`CostofLiving` ~ `Housing` + `Education` + `env_quality` + `BusinessFreedom` + `Economy` +

costofliving_drop %>%
  mutate(predictions = predict(fit_lin)) %>%
  select(`CostofLiving` , `Housing` , `Education` , `env_quality` , `BusinessFreedom` , `Economy` , `To
```

```
## # A tibble: 266 x 7
##    CostofLiving Housing Education env_quality BusinessFreedom Economy Tolerance
##           <dbl>   <dbl>     <dbl>       <dbl>           <dbl>   <dbl>     <dbl>
##  1         4.02    6.13      5.37        7.63            9.94    4.89      9.74
##  2         4.69    6.31      5.14        8.33            9.40    6.07      7.82
##  3         6.06    7.26      4.15        7.32            8.67    6.51      7.03
```

```
##  4         9.33    9.28      2.28        3.86            5.57    5.27      6.54
##  5         3.82    3.05      6.18        7.60            8.84    5.05      8.37
##  6         3.14    5.43      3.62        9.27            8.67    6.51      7.09
##  7         0       3.97      0           7.26            0       0         8.70
##  8         9.12    9.93      2.03        2.94            5.95    4.09      4.49
##  9         5.31    5.86      3.62        8.49            8.67    6.51      7.73
## 10         9.30    9.23      0           3.97            3.68    4.11      7.05
## # ... with 256 more rows
```

```
# Take a look at the model summary
summary(fit_lin)
```

```
##
## Call:
## lm(formula = CostofLiving ~ Housing + Education + env_quality +
##     BusinessFreedom + Economy + Tolerance, data = costofliving_drop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9516 -0.5796  0.0695  0.7729  4.8437
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.20273    0.69190   1.738  0.08335 .
## Housing          0.87222    0.04775  18.267  < 2e-16 ***
## Education        0.19576    0.06415   3.052  0.00251 **
## env_quality     -0.14677    0.06072  -2.417  0.01634 *
## BusinessFreedom -0.21201    0.07968  -2.661  0.00828 **
## Economy          0.25343    0.08002   3.167  0.00172 **
## Tolerance       -0.10070    0.06023  -1.672  0.09574 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.389 on 259 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.678
## F-statistic: 93.98 on 6 and 259 DF,  p-value: < 2.2e-16
```

```
# Calculate RMSE of regression model
sqrt(mean(resid(fit_lin)^2))
```

```
## [1] 1.370307
```

The graph displays the prediction error as well as the line of best fit. In the graph, the residuals are relatively evenly dispersed above and below the line of best fit. The predictions are colored in orange whereas the actual values are in black. The line of best fit has a slight negative slope which means there is a negative correlation between the cost of living and the other predictors in our model. This is surprising as we expected the cost of living to increase as the rating of our predictors increased.

```
# residuals
costofliving_drop %>%
  mutate(residuals = resid(fit_lin)) %>%
  select(`CostofLiving` , `Housing` , `Education` , `env_quality` , `BusinessFreedom` , `Economy` , `To
```

```
## # A tibble: 266 x 7
##    CostofLiving Housing Education env_quality BusinessFreedom Economy Tolerance
##           <dbl>   <dbl>     <dbl>       <dbl>           <dbl>   <dbl>     <dbl>
##  1         4.02    6.13      5.37        7.63            9.94    4.89      9.74
```

```
##  2         4.69    6.31      5.14        8.33          9.40    6.07      7.82
##  3         6.06    7.26      4.15        7.32          8.67    6.51      7.03
##  4         9.33    9.28      2.28        3.86          5.57    5.27      6.54
##  5         3.82    3.05      6.18        7.60          8.84    5.05      8.37
##  6         3.14    5.43      3.62        9.27          8.67    6.51      7.09
##  7         0       3.97      0           7.26          0       0         8.70
##  8         9.12    9.93      2.03        2.94          5.95    4.09      4.49
##  9         5.31    5.86      3.62        8.49          8.67    6.51      7.73
## 10         9.30    9.23      0           3.97          3.68    4.11      7.05
## # ... with 256 more rows
```

```
#Plot Linear Regression

costofliving_drop %>%
  # save predictions
  mutate(predictions = predict(fit_lin)) %>%
  # use a ggplot to represent the relationship
  ggplot(aes(x =  `Housing` + `Education` + `env_quality` + `BusinessFreedom` + `Economy` + `Tolerance`
  # add the linear model
  geom_smooth(method = "lm", se = FALSE) +
  # add residuals = vertical segments from observations to predictions
  geom_segment(aes(xend = `Housing` + `Education` + `env_quality` + `BusinessFreedom` + `Economy` + `To
  # display the observed data (on top of the line and segments)
  geom_point() +
  # display the predictions
  geom_point(aes(y = predictions), color = "orange")
```

We performed cross-validation on our linear regression model to see how well our model performs on new data. The Average RMSE calculated was 1.393569 while the average R-squared was 0.6815018. Our RMSE was higher that the model that was trained on the entire dataset which means our model performed a bit worse when given new data. Our R^2 remained about the same which means that 67.56% of the variance in our data can be explained by the predictors in our model.

```
# cross Validation Linear Regression
num_folds <- 5

# Create a data partition for cross-validation
folds <- createFolds(costofliving_drop$CostofLiving, k = num_folds)

# Initialize a vector to store the RMSE for each fold
perf_k <- numeric(num_folds)

# Loop through each fold
for (i in 1:num_folds) {
  # Extract the training and testing subsets for the current fold
  train_subset <- costofliving_drop[-folds[[i]], ]
  test_subset <- costofliving_drop[folds[[i]], ]

  # Fit the linear regression model on the training subset
  fit <- lm(CostofLiving ~ Housing + Education + env_quality + BusinessFreedom + Economy + Tolerance, da

  # Make predictions on the testing subset
  predictions <- predict(fit, newdata = test_subset)

  # Calculate the RMSE for the current fold
  perf_k[i] <- sqrt(mean((test_subset$CostofLiving - predictions)^2))
}

# Print the RMSE for each fold
print(perf_k)
```

```
## [1] 1.489806 1.210141 1.831336 1.269919 1.294690
```

```
sum(mean(perf_k))
```

```
## [1] 1.419178
```

For our second classification model, we chose to do KNN. Since our response variable was numerical we couldn't calculate the ROC and compute the AUC. Instead, we used the RMSE as our evaluation metric. The average RMSE was 5.702439 which means this model was worse than the linear regression model. This model's performance differed from training the whole dataset and cross validation. The KNN model that trained on the whole dataset had an average RMSE of 0 and thus perfectly classifies everything while the KNN model that performed cross validation and this was tested against new data had an average RMSE 5.703589. This suggests that this model is prone to overfitting.

```
# choose number of folds
k = 5
# randomly order rows in the dataset
data <- costofliving_drop[sample(nrow(costofliving_drop)), ]
# create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)
# initialize a vector to keep track of the performance
perf_k <- NULL
```

```
for(i in 1:k){
train <- data[folds != i, ] # all observations except in fold i
test <- data[folds == i, ] # observations in fold i
# train model on train set (all but fold i)
costofliving_kNN <- knn3(CostofLiving ~ Housing + Education + env_quality + BusinessFreedom + Economy +
# kNN with k = 10
predict_i <- data.frame(
  predictions = predict(costofliving_kNN, newdata = test, type = "class"),
  CostofLiving = test$CostofLiving)
# compute the root mean squared error (RMSE) for fold i
perf_k[i] <- sqrt(mean(test$CostofLiving - predict(costofliving_kNN, newdata = test)[,1]))^2
}
perf_k
```

```
## [1] 5.778722 5.993660 5.682717 5.031981 6.042000
```

```
sum(mean(perf_k))
```

```
## [1] 5.705816
```

```
#KNN model
# Perform k-NN on numerical data with a numerical label
costofliving_kNN <- knn3(CostofLiving ~ Housing + Education + env_quality + BusinessFreedom + Economy +

# Generate predictions
costofliving_predictions <- data.frame(predict(costofliving_kNN, costofliving_drop))
costofliving_predictions$CostofLiving <- costofliving_drop$CostofLiving

# Calculate RMSE
costofliving_predictions$error <- (costofliving_predictions$CostofLiving - costofliving_drop$CostofLivi
mse <- mean(costofliving_predictions$error)
rmse <- sqrt(mse)

# Print RMSE
cat("RMSE:", rmse, "\n")
```

```
## RMSE: 0
```

We performed cross-validation 5 times on our kNN model and got an average RMSE of 1.193047 which is similar to the RMSE of the linear regression model. The average R^2 was 0.6671175 which means that 66.7% of the variation in the Cost of Living can be explained by the predictors.

```
#ASK WHY IT ONLY PRINTS OUT THREE VALUES
# cross-Validation KNN
 train_control <- trainControl(method = "cv", number = 5)

 # train the k-NN model with cross-validation
 knncv <- train(`CostofLiving` ~ `Housing` + `Education` + `env_quality` + `BusinessFreedom` + `Economy

 # calculate the RMSE with cross-validation
 rmse_cv <- knncv$results$RMSE
 rmse_cv
```

```
## [1] 1.489061 1.464953 1.482490
```

```
 knncv$results
```

```
##   k     RMSE  Rsquared       MAE    RMSESD RsquaredSD     MAESD
## 1 5 1.489061 0.6517759 1.0133391 0.3825636 0.09554686 0.2126584
## 2 7 1.464953 0.6665961 0.9918612 0.3864515 0.09960819 0.2305731
## 3 9 1.482490 0.6608501 1.0022288 0.3878300 0.09547601 0.2528379
```

```
# print the RMSE with cross-validation
cat("RMSE with cross-validation:", rmse_cv, "\n")
```

```
## RMSE with cross-validation: 1.489061 1.464953 1.48249
```

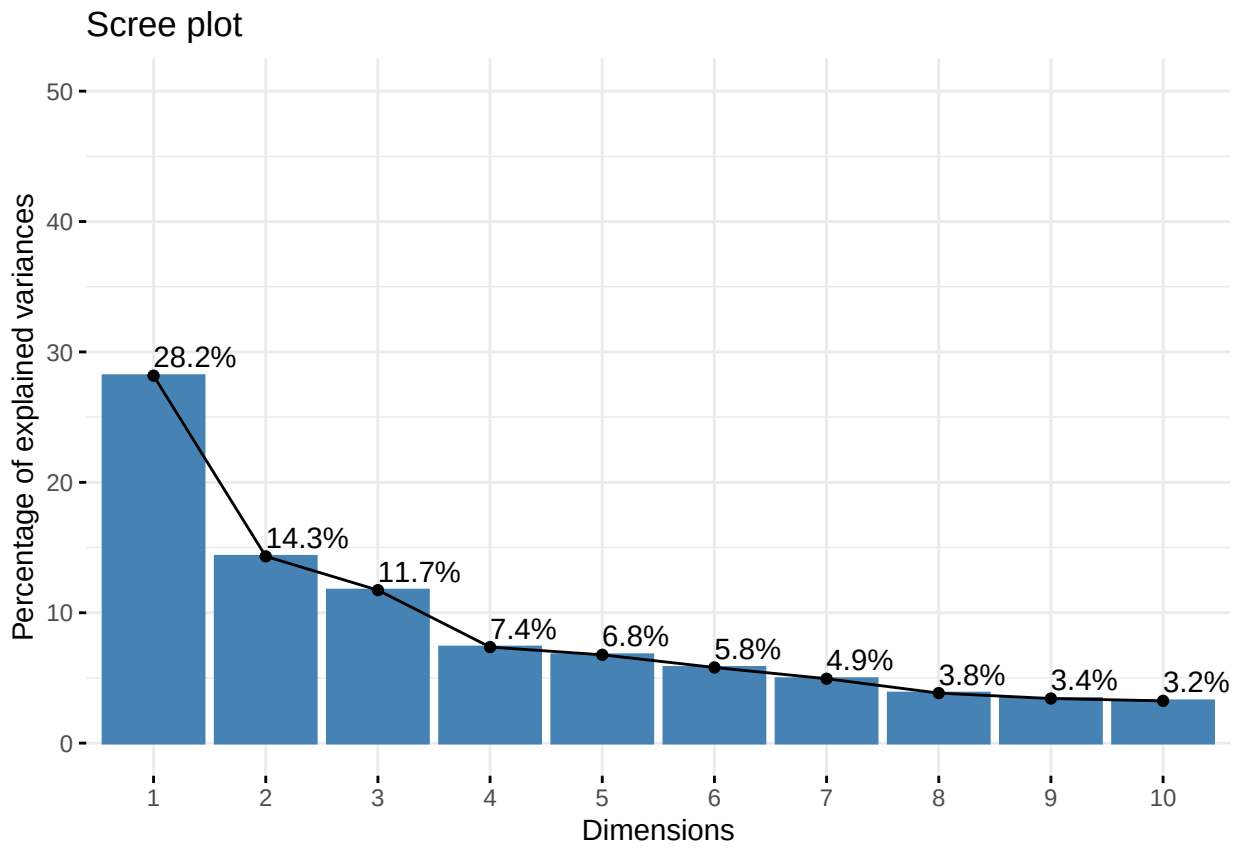
```
print(sum(mean(rmse_cv)))
```

```
## [1] 1.478835
```

4. Dimensionality reduction:

The first line scales the data, which standardizes the variables to have a mean of zero and a standard deviation of one. This is done so that variables with larger magnitudes do not dominate the PCA results. The first graph shows the proportion of variance explained by each principal component. We saw that we needed to keep the first 6 variables that explain about 80% of the variance in our data. We found out that 42.5% of the variation in our Cost of Living can be explained by the first two principal components.

```
# scale the cost of living data
costofliving_scaled <- scale(costofliving_drop)
# run principal component analysis
pca <- prcomp(costofliving_scaled)
# visualize eigenvalues
fviz_eig(pca, addlabels = TRUE, ylim = c(0, 50))
```



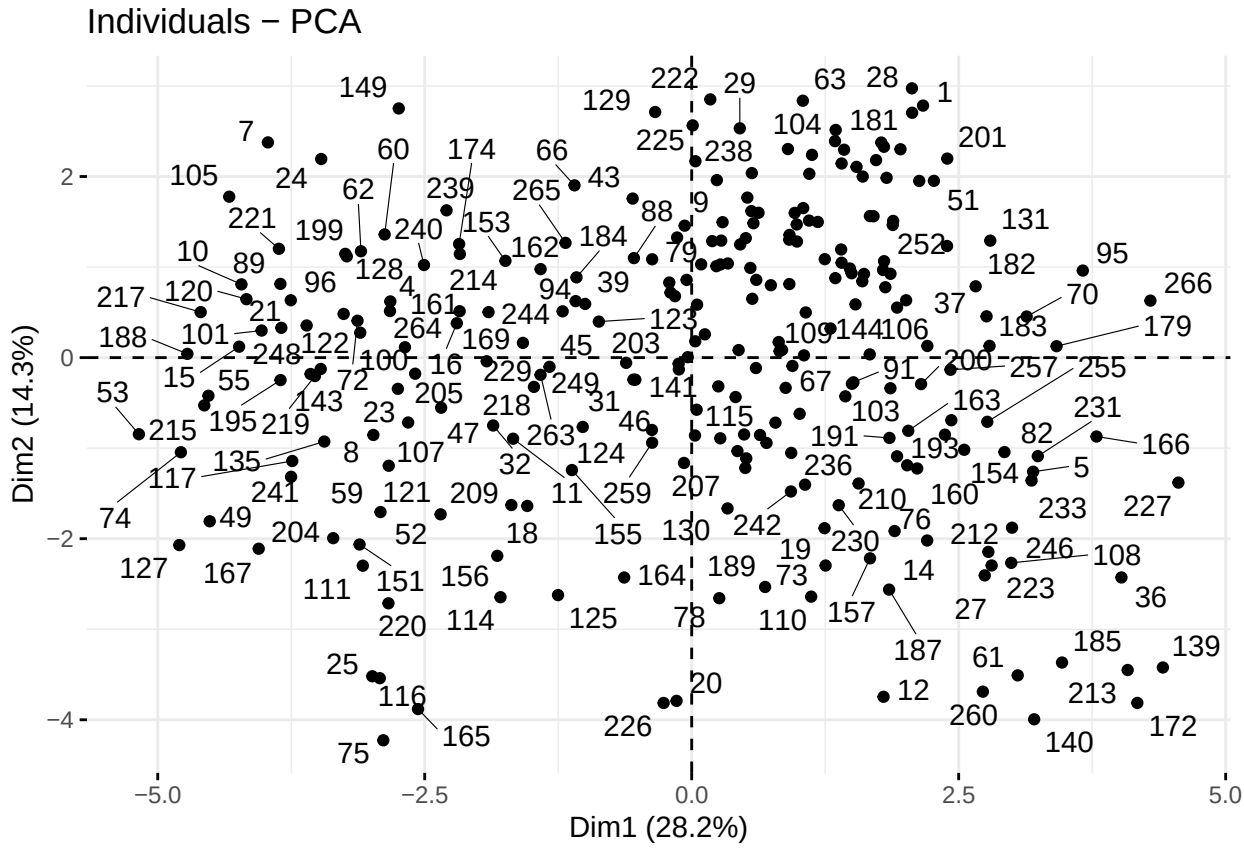
```r
# extract the PCA loadings
get_pca_var(pca)$coord %>% as.data.frame %>%
  arrange(Dim.1) %>% select(Dim.1) # arrange by first dimension
```

```
##                         Dim.1
## Housing            -0.69148070
## CostofLiving       -0.58321742
## Taxation           -0.05684455
## Safety              0.08988676
## ...5                0.09089716
## leisure_culture     0.17181712
## Outdoors            0.20868406
## Commute             0.25904183
## Startups            0.40657155
## Tolerance           0.42754898
## TravelConnectivity  0.47128217
## VentureCapital      0.53279310
## Economy             0.61246219
## internet_access     0.62090950
## Healthcare          0.63230169
## env_quality         0.73325643
## BusinessFreedom     0.84888015
## Education           0.90637863
```

```r
get_pca_var(pca)$coord %>% as.data.frame %>%
  arrange(Dim.2) %>% select(Dim.2) # arrange by second dimension
```

```
##                         Dim.2
## Startups           -0.77585436
## VentureCapital     -0.68516449
## leisure_culture    -0.59200769
## Outdoors           -0.31430974
## CostofLiving       -0.28661878
## TravelConnectivity -0.21947227
## Economy            -0.11269757
## Education          -0.07872737
## ...9               -0.04118900
## Commute             0.03516601
## internet_access     0.03826738
## Housing             0.08079540
## BusinessFreedom     0.15900973
## Taxation            0.16913303
## Healthcare          0.38489634
## env_quality         0.40725614
## Safety              0.45483410
## Tolerance           0.56824117
```

```r
# create a scatter plot of the first two principal components
fviz_pca_ind(pca,
             repel = TRUE)
```




```
pca
```

```
## Standard deviations (1, .., p=18):
##  [1] 2.2521849 1.6057620 1.4532471 1.1515289 1.1041088 1.0219617 0.9428881
##  [8] 0.8309555 0.7846671 0.7637976 0.6513832 0.5646369 0.5392822 0.5216836
## [15] 0.4329047 0.3926486 0.3515637 0.3128882
##
## Rotation (n x k) = (18 x 18):
##                              PC1         PC2         PC3         PC4         PC5
## ...1                0.04035955 -0.02565075  0.16293999 -0.28590727  0.25102596
## Housing            -0.30702661  0.05031592  0.30815577  0.35927496  0.09791213
## CostofLiving       -0.25895628 -0.17849394  0.31931598  0.33787678  0.23392937
## Startups            0.18052317 -0.48316896 -0.01145011 -0.06784330  0.15148755
## VentureCapital      0.23656721 -0.42669119 -0.02272899 -0.10756117  0.03226744
##                              PC6          PC7         PC8         PC9
## ...1               -0.605664918  0.565292017 -0.02089921  0.35345645
## Housing            -0.115978576 -0.055504923  0.04891101  0.07227806
## CostofLiving       -0.022424804 -0.165171996  0.12462646  0.11973603
## Startups            0.091633941 -0.061929785  0.22949630  0.09782465
## VentureCapital      0.175683724 -0.023970725  0.18685868  0.28687311
##                             PC10        PC11          PC12         PC13
## ...1               -0.02626055  0.01602766 -0.0840694915  0.001058995
## Housing            -0.05443446 -0.34633950  0.0322879167 -0.048440388
## CostofLiving        0.01359432 -0.31219641 -0.0689540934  0.041049423
## Startups            0.26104576 -0.14426533  0.0456002322 -0.045367193
## VentureCapital      0.15983983 -0.07978222  0.2467910475 -0.249859897
##                             PC14        PC15         PC16        PC17        PC18
```

```
## ...1               -0.06545619  0.04078017 -0.003540195  0.02094457  0.01547139
## Housing            -0.03994015  0.01591422  0.115890284 -0.43029050 -0.56453289
## CostofLiving       -0.07751474  0.25629761 -0.030129329  0.37620341  0.50433833
## Startups           -0.27794623 -0.52296579  0.134414170  0.30629193 -0.27583035
## VentureCapital      0.17519984  0.28388084 -0.269154505 -0.48627799  0.15731548
##  [ reached getOption("max.print") -- omitted 13 rows ]
```

We chose to do K-means clustering with the Healthcare and Education variable. First, we scaled the data and created a visualization of the silhouette score to determine the optimal number of clusters for these two variables. The silhouette width was the largest (0.5) for k = 2 so we visualized k-means clustering with 2 clusters. The first cluster contains cities with low-high healthcare scores and below average education scores with a centroid of a Healthcare score of -1.009 and an education score of -1.248. The second cluster contains cities with average-high education and healthcare scores with a centroid of a Healthcare score of 0.442 and an education score of 0.546. We also created basic summary statistics for each cluster in original units with all the predictors. The highest difference in centers between cluster 1 and cluster 2 was education which makes sense since education is one of the most important predictors of the cost of living.

5. Clustering:

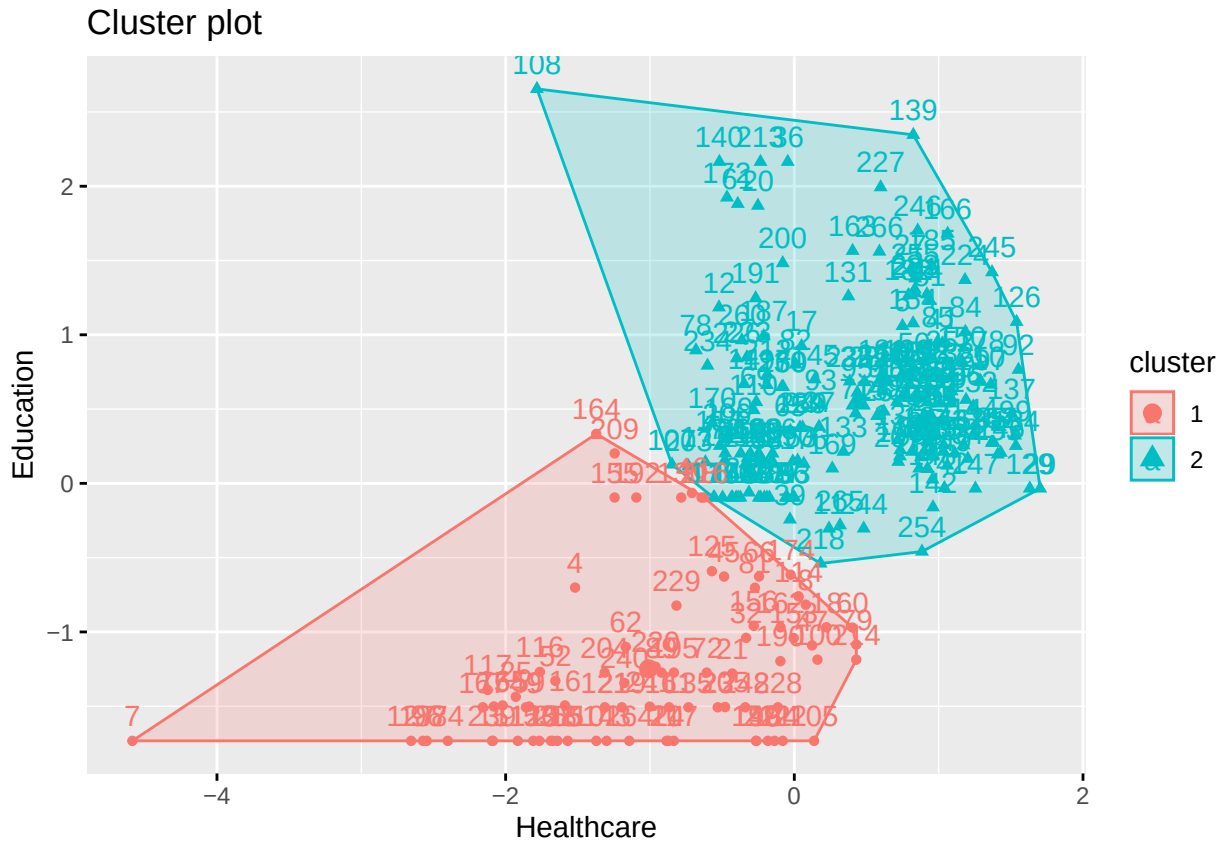
```
#k-means Clustering
costofliving_scaled_2 <- costofliving_drop %>%
# select the Healthcare and Education columns from the data frame
  select(Healthcare, Education) %>%
# scale data
  scale

kmeans_results <- costofliving_scaled_2 %>%
  kmeans(centers = 2) # centers sets the number of clusters to find

# take a look at the resulting object
kmeans_results
```

```
## K-means clustering with 2 clusters of sizes 81, 185
##
## Cluster means:
##   Healthcare Education
## 1 -1.0098284 -1.247688
## 2  0.4421411  0.546285
##
## Clustering vector:
##   [1] 2 2 2 1 2 2 1 1 2 1 2 2 2 2 1 1 2 1 2 2 1 2 1 1 1 2 2 2 2 2 2 1 2 2 2 2 2
##  [38] 2 2 2 2 2 2 2 1 1 1 2 1 2 2 1 1 2 1 2 2 2 1 1 2 1 2 2 2 1 2 1 2 2 2 1 2 1
##  [75] 1 2 2 2 1 2 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 1
##  [ reached getOption("max.print") -- omitted 166 entries ]
##
## Within cluster sum of squares by cluster:
## [1]  87.12639 142.80437
##  (between_SS / total_SS =  56.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

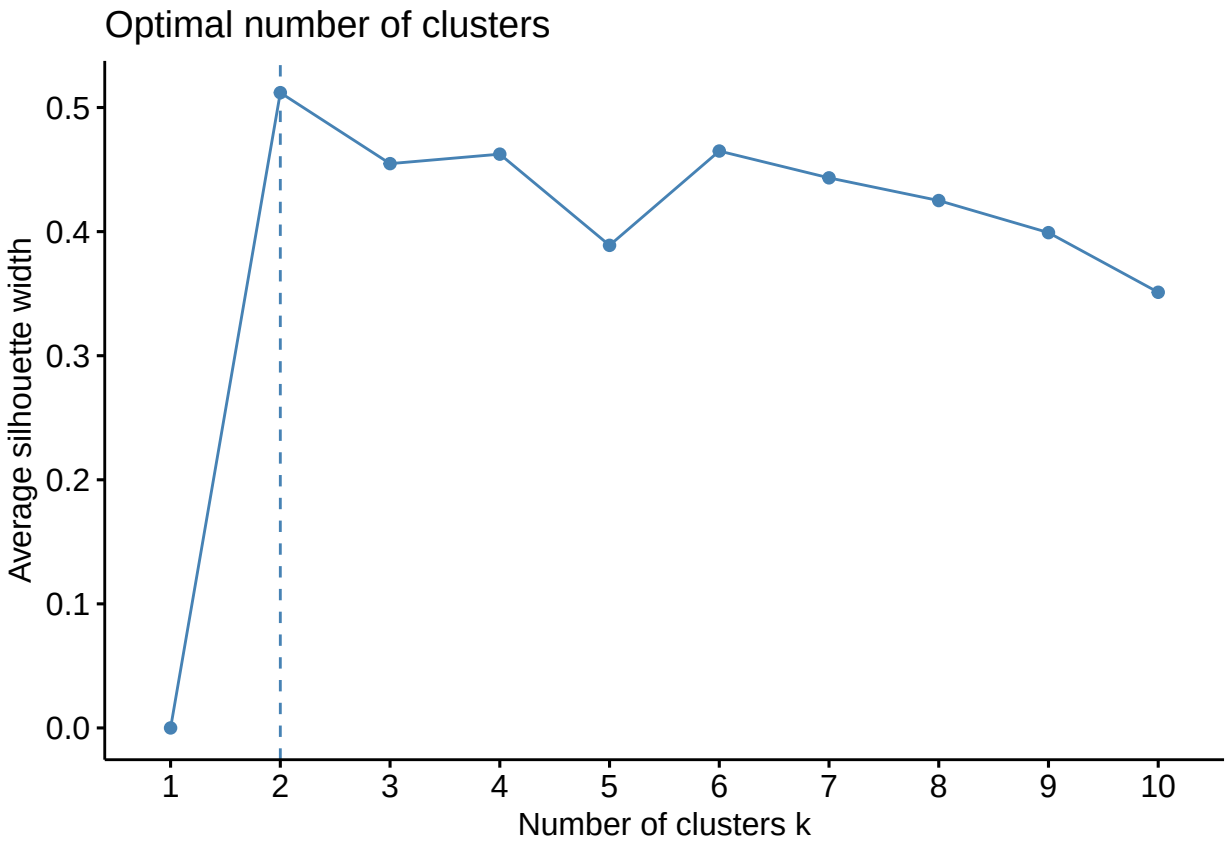
```
# gives the coordinates of the centers of the clusters
kmeans_results$centers
```

```
##   Healthcare Education
## 1 -1.0098284 -1.247688
## 2  0.4421411  0.546285
```

```
# creates a visualization of the k-means clustering results
fviz_cluster(kmeans_results, data = costofliving_scaled_2)
```




```
# visualizes the silhouette score for different numbers of clusters
fviz_nbclust(costofliving_scaled_2, kmeans, method = "silhouette")
```




```
# create basic summary statistics for each cluster in original units
costofliving_drop %>%
  select_if(is.numeric) %>%
  na.omit %>%
  mutate(cluster = as.factor(kmeans_results$cluster)) %>%
  group_by(cluster) %>%
  summarize_all(mean)
```

```
## # A tibble: 2 x 19
##   cluster  ...1 Housing CostofL~1 Start~2 Ventu~3 Trave~4 Commute Busin~5 Safety
##   <fct>   <dbl>   <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>
## 1 1         123.    7.93      7.33    4.10    1.52    2.40    4.07    4.99   6.98
## 2 2         136.    5.83      5.05    4.81    3.22    3.91    4.88    8.36   6.97
## # ... with 9 more variables: Healthcare <dbl>, Education <dbl>,
## #   env_quality <dbl>, Economy <dbl>, Taxation <dbl>, internet_access <dbl>,
## #   leisure_culture <dbl>, Tolerance <dbl>, Outdoors <dbl>, and abbreviated
## #   variable names 1: CostofLiving, 2: Startups, 3: VentureCapital,
## #   4: TravelConnectivity, 5: BusinessFreedom
```

```
#optimal number of clusters is 2 since it contains the largest silhouette width
```

6. Discussion

Our first research question was, "Is there a positive correlation between venture capital and start-ups?". We expected there to be a positive trend between the two variables because venture capital firms can provide the necessary funding for start-ups which helps these businesses grow. Venture capital can also provide start-ups with strategies to improve their business. Venture capital can provide financial and strategic

support. According to Graph 5, it is shown that there is a positive correlation between the two variables. The scatterplot indicates that this positive linear fit indicates a positive correlation, which supports our hypothesis.

Our second research question was "Does housing impact the cost of living?". We expected that housing would impact the cost of living since typically a higher quality housing costs more. According to Graph 3 there is a strong positive correlation between housing and cost of living which supports our hypothesis.

During our clustering we saw that if housing values were split based on clusters of education and healthcare, there would be two housing clusters surrounding 5.82 and 7.92 respectively.There were lower variable scores in the first cluster than the second cluster which indicates that some variables correlated with each other amongst this dataset. This further supports our hypothesis that housing affects cost of living. We can also see that in the correlation matrix that the majority of boxes are a shade of red which supports our hypothesis that most of our predictors' scores will increase as the cost of living increases.

Some challenges throughout this project were finding the right amount of data sets with numerical data. This was challenging because we initially had a lot of observations in our data, but decided to find a different data set since the number of observations was hard on the programming tool. Once, we found a new data set, it was challenging to perform a 5-fold cross-validation for our kNN model because it was only printing out 3 values when we had put 5 in the code chunk. However, once we got help from our teaching assistant and professor, we ended up getting 5 different fold values, which was relieving. Although dimensionality reduction and clustering were not overly difficult, they did require a significant amount of time and experimentation to achieve the results we wanted.

This project helped us gain insight and an understanding of how different models work. Both of us were involved in selecting data, data analysis, prediction and cross-validation, dimensionality reduction, and clustering. We mostly worked together remotely, or in-person. Individually, both of us came up with a research question. Finally, we collaborated on the explanations for our data finding, introduction, and conclusion. We would like to thank our lovely professor, out TA's and our classmates for helping us with the project.

7. Formatting:

Comment your code, write full sentences, and knit your file!

```
##                                                 sysname
##                                                 "Linux"
##                                                 release
##                                       "5.15.0-67-generic"
##                                                 version
## "#74~20.04.1-Ubuntu SMP Wed Feb 22 14:52:34 UTC 2023"
##                                                nodename
##                               "educcomp02.ccbb.utexas.edu"
##                                                 machine
##                                                "x86_64"
##                                                   login
##                                               "unknown"
##                                                    user
##                                               "jaj4598"
##                                          effective_user
##                                               "jaj4598"
```