Task 3 (Group Task) — Multivariate Data Analysis

Group Members:
Chloi Chatzinota (16317604)
Abhi Ashokbhai Thumar (16627134)
Yash Anand Kummar (16349731)
Maulikkumar Parmar (16421406)

Part I

Create a lookup table (From Country to Region)

To create a lookup from County (CTY24CD) to Region (RGN24CD), we first join the County lookup file to the Region lookup file using `LAD24CD as the common variable` . Then we check that each county appears only once and keep the country and region codes and names. The result is a tibble that maps every county to its region.

```
library(tidyverse)
library(GGally)
library(readr)

#Load datasets
imd_data= read_csv("imd2025_group.csv")
lad_to_county =
read_csv("Local_Authority_District_to_County_(December_2024)_Lookup_in_EN
.csv")
lad_to_region =
read_csv("Local_Authority_District_to_Region_(December_2024)_Lookup_in_EN
.csv")

 #create county to region lookup
county_region_lookup = lad_to_county %>%
  left_join(lad_to_region %>% select(LAD24CD, RGN24CD, RGN24NM),
    by = "LAD24CD") %>%
  distinct(CTY24CD, .keep_all = TRUE) %>%
  select(CTY24CD, CTY24NM, RGN24NM) %>%
```

Output:

```
 CTY24CD   CTY24NM         RGN24NM
   <chr>     <chr>           <chr>
 1 E10000003 Cambridgeshire  East of England
 2 E10000007 Derbyshire      East Midlands
 3 E10000008 Devon           South West
 4 E10000011 East Sussex     South East
 5 E10000012 Essex           East of England
 6 E10000013 Gloucestershire South West
 7 E10000014 Hampshire       South East
 8 E10000015 Hertfordshire   East of England
 9 E10000025 Oxfordshire     South East
10 E10000028 Staffordshire   West Midlands
# ℹ 19 more rows
```

To fill the countries from the IMD data set that have missing Region values, we first join the IMD data set to the lad to country file to add country codes to each district. Then we join again to the Country to Region lookup we created. If an IMD row has a missing region, we replace it with the region name coming from the lookup. This ensures that every district has a complete Region value.

```
#join LAD to region, join counties, final region=lad region
  imd_with_region = imd_data %>%
  left_join(lad_to_region %>% select(LAD24CD, RGN24NM),
            by="LAD24CD") %>%
  rename(Region_LAD = RGN24NM) %>%
  left_join(county_region_lookup %>% rename(LAD24CD = CTY24CD,
Region_County = RGN24NM),
            by="LAD24CD") %>%
  mutate(Region = coalesce(Region, Region_LAD, Region_County)) %>%
  select(-Region_LAD, -Region_County)
```

Now that all Region values are completed, we group the dataset by Region and count how many districts belong to each region. This produces a summary table giving the number of districts in each region.

```
#summary table
district_count_by_region = imd_with_region%>%
  group_by(Region) %>%
  summarise(number_of_districts = n()) %>%
  arrange(desc(number_of_districts))

district_count_by_region
```

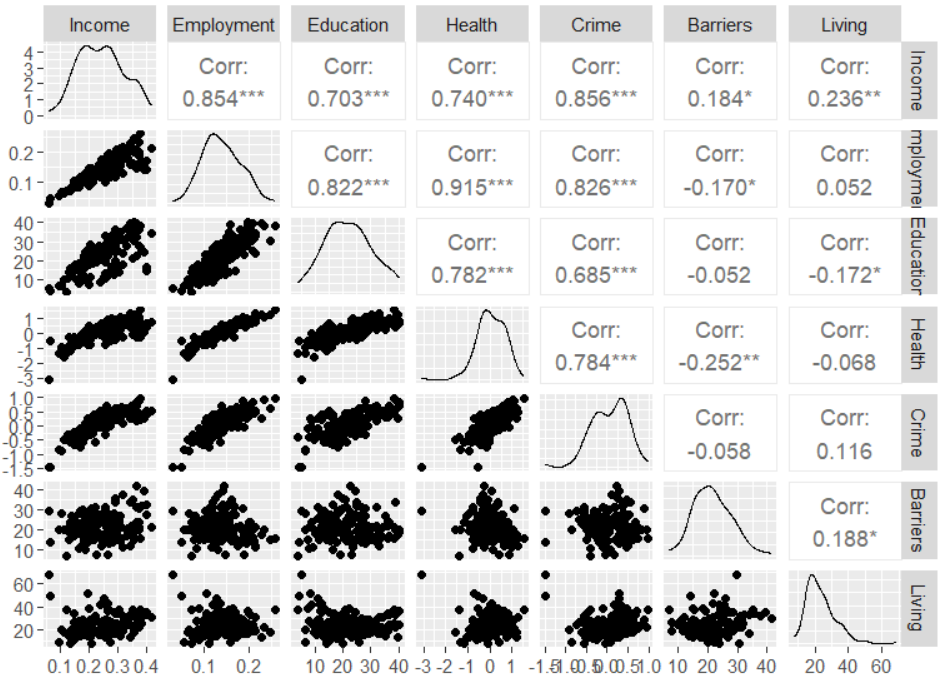
Output:

```
Region                    number_of_districts
  <chr>                                 <int>
1 London                                   33
2 North West                               24
3 South East                               19
4 South West                               15
5 Yorkshire and The Humber                 15
6 West Midlands                            14
7 North East                               12
8 East of England                          11
9 East Midlands                            10
```

Next, we build a scatter matrix using ggpairs(), to explore relationships between the IMD domains. That will help us to visually identify which domains are strongly related.

```
domains = imd_with_region %>%
  select(Income, Employment, Education, Health, Crime, Barriers, Living)
ggpairs(domains, color='Region')
```

Output:



The scatter matrix allows us to see how the seven IMD domains are related to each other. The strongest relationships are between Income, Employment and Education, all showing correlation coefficients above 0.70. Health also shows strong correlation with Employment and Income, indicating that poorer economics and employment conditions are closely linked to worse health outcomes. Crime shows a similarly strong positive relationship with Income and Employment, implying that areas with high deprivation tend to also have higher crime levels. In contrast, Barriers and Living show weak or near zero correlations with most domains, making them appear more independent compared with the income-employment-education cluster.

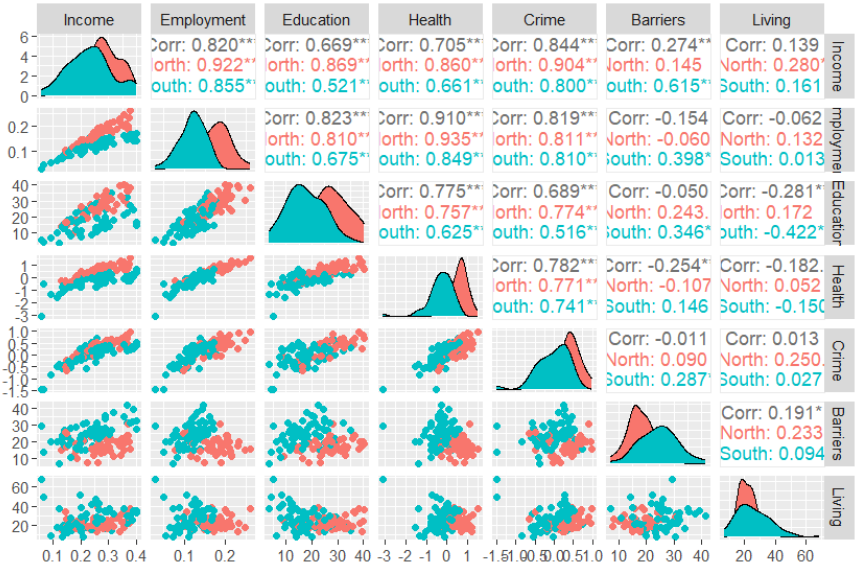
Part II

The original dataset did not contain a North/South classification, so we created a simple lookup table that has regions based on the map provided in the question. The north includes North East, North West and Yorkshire and The Humber while the South includes South East, South West, London and East of England. Then we used the function join to attach this grouping to every district.

```
library(tidyverse)
region_lookup <- tibble(
  Region = c(
    "North East",
    "North West",
    "Yorkshire and The Humber",
    "South East",
    "South West",
    "London",
    "East of England" ),
  area_group = c(
    "North",
    "North",
    "North",
    "South",
    "South",
    "South",
    "South")
)
north_south_data = imd_with_region %>%
  inner_join(region_lookup, by = "Region")
ggpairs(north_south_data,
        columns =
c("Income","Employment","Education","Health","Crime","Barriers","Living")
, aes(color = area_group))

ggplot(north_south_data, aes(Income, Employment, color = area_group)) +
geom_point(size=3) + theme_minimal()
```

Task 3 (Group Task) — Multivariate Data Analysis

Output:

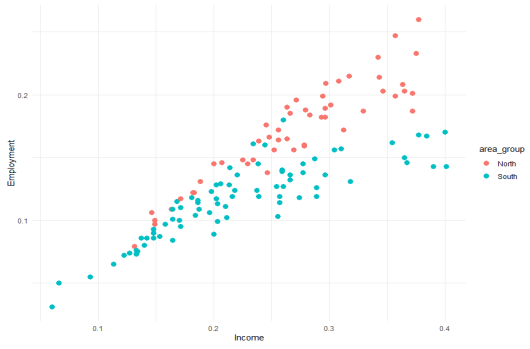


The scatter matrix shows the relationships between the seven IMD domains for the North (red) and South (blue). The points are colored by region and by that we can visually inspect which variables show the separation between the two groups. Most pairs of variables showed a lot of overlap between red and blue points but the combination involving Employment and Income showed the clearest separation.

In the Employment plots, the North (red) points are mostly higher and form a different shape compared to the blue points, which are generally lower. The same pattern appears in the Income plots. When Employment and Income are plotted together, the red and blue points tend to occupy different regions of the graph, with the North generally having higher deprivation values and the South having lower values.

Because Employment and Income show the strongest visual difference between North and South across multiple panels in the scatterplot matrix, these two variables appear most suitable for predicting whether a district belongs to the North or South.

Conclusion:



After looking at the scatter plot we saw the Income and Employment showed the clearest separation. But using correlation as the only criteria in order to determine the result would be insufficient so we used the OLSRR library to be specific the ols step best subset, in order to avoid overfitting and got to know that income and employment showed the best two predictors in order to compare North and South. So to conclude which is the best two predictors

we did an AIC check on two sets of predictors Employment + Health (highest correlation) and Income + Employment (best predictor using OLSRR) and the output showed that income and employment had lowest AIC value and so we have come to the conclusion that Income and Employment fits the best.

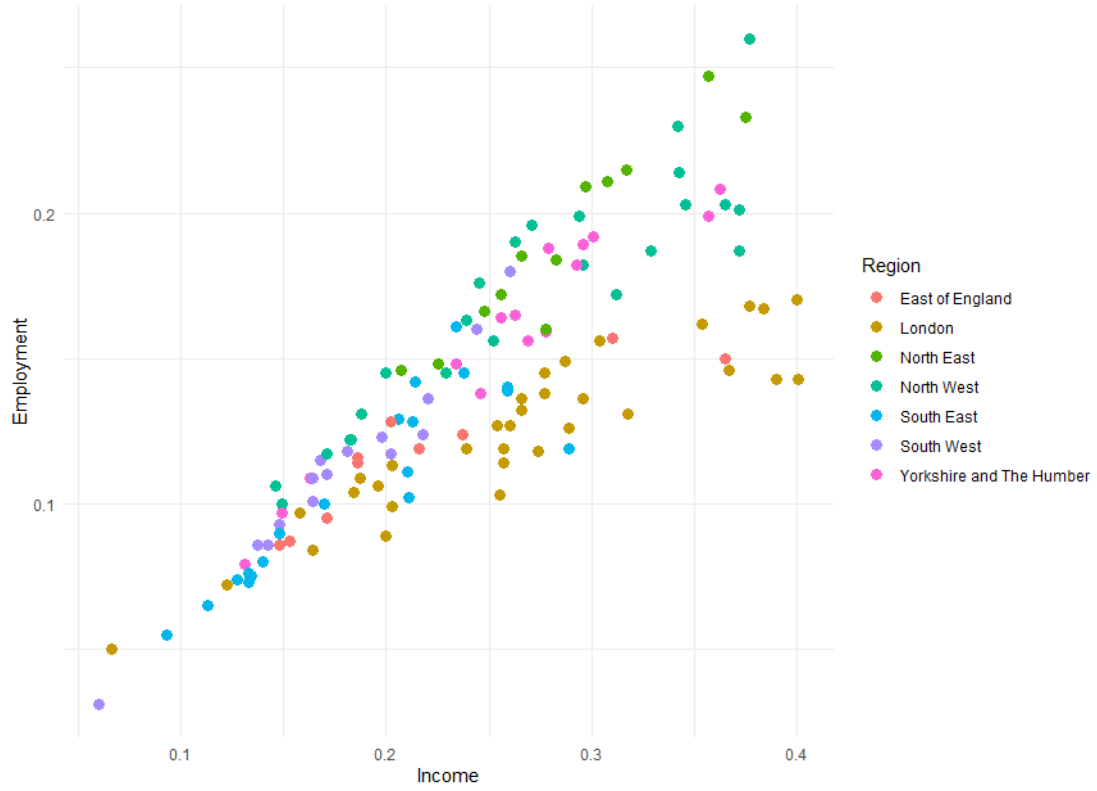
```
a <- north_south_data%>% select(-LAD24CD,-LAD24NM,-Rank,-CTY24NM,-
area_group,-Region)
full_model <- lm(Overall ~ ., data = a)
best <- ols_step_best_subset(full_model)
best



aic1 <- lm(Overall ~ Income+Employment, data = north_south_data)
summary(aic1)
AIC(aic1)



aic2 <- lm(Overall ~ Employment+Health, data = north_south_data)
summary(aic2)
AIC(aic2)
```

Output:

```
                        Best Subsets Regression
--------------------------------------------------------------------------
Model Index    Predictors
--------------------------------------------------------------------------
     1         Employment
     2         Income Employment
     3         Income Employment Living
     4         Income Employment Education Living
     5         Income Employment Education Barriers Living
     6         Income Employment Education Health Barriers Living
     7         Income Employment Education Health Crime Barriers Living
--------------------------------------------------------------------------
> AIC(aic1)
[1] 507.5568

> AIC(aic2)
[1] 620.174
```

Comment any particular districts that are difficult to predict:

```
ggplot(north_south_data, aes(Income, Employment, color = area_group)) +
geom_point(size=3) + theme_minimal()
```



Although most districts follow the general pattern, a few do not. Examples typically found in South West and sometimes East of England:

- These appear closer to the Northern cluster on the Income–Employment plot.
- They would be mis-classified as “North” if using only these 2 variables.

Some areas in Yorkshire and The Humber show:

- Lower Employment deprivation
- A more “Southern-like” position in the plot

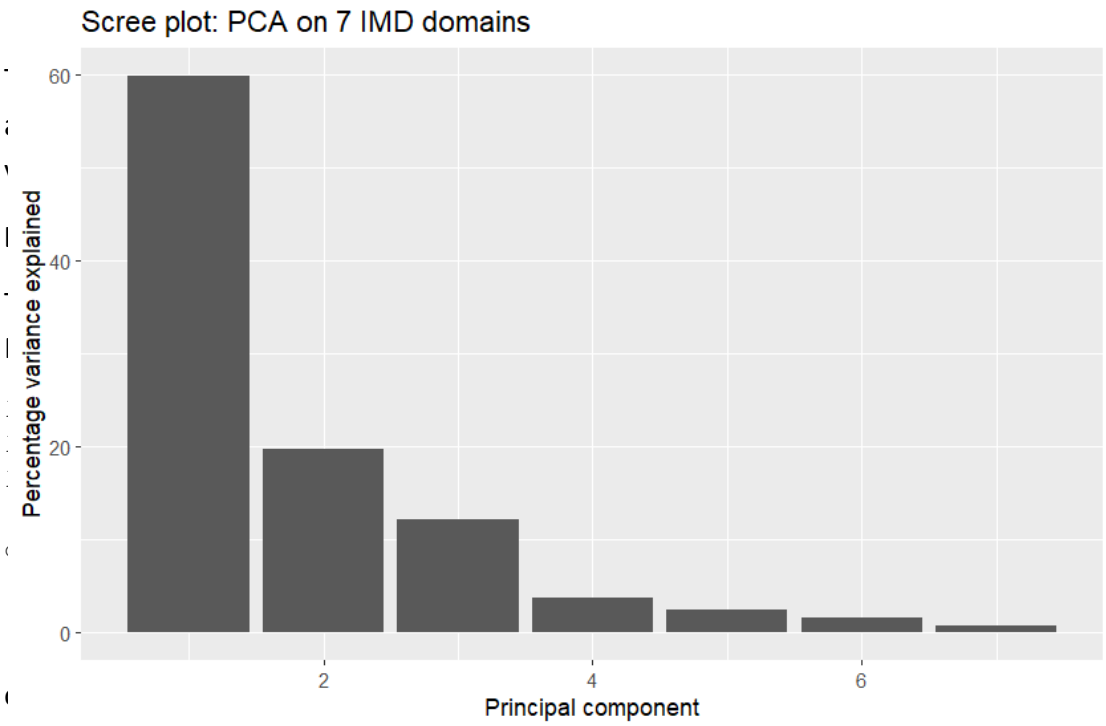
These districts sit on the boundary between clusters.

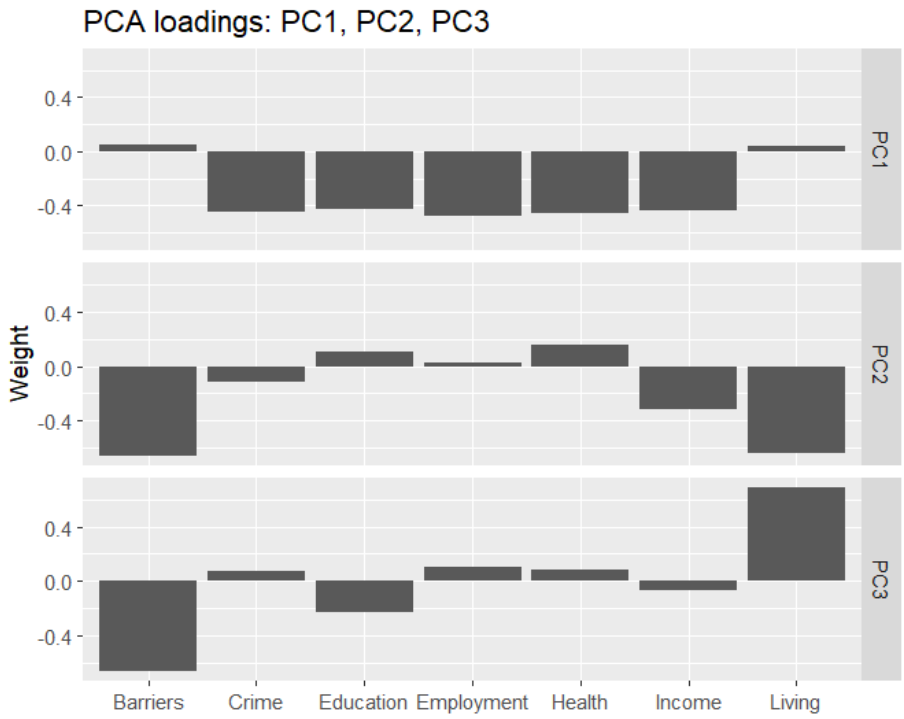
Part III

a.

```
library(tidyverse)
library(GGally)
library(ggfortify)
library(ggplot2)
imd = read_csv('imd2025_individual.csv')

imd_domains = imd %>% select(Income, Employment, Education, Health,
Crime, Barriers, Living)

pca_results = prcomp(imd_domains, scale. = TRUE)

The scree plot shows how much of the total variance each principal
component explains.

# PCA scree plot
variance = (pca_results$sdev)^2
variance_explained = variance / sum(variance)
ggplot(NULL, aes(x = 1:7, y = 100 * variance_explained)) +
  geom_col() +
  xlab("Principal component") + ylab("Percentage variance explained") +
  ggtitle("Scree plot: PCA on 7 IMD domains")
```

Output:





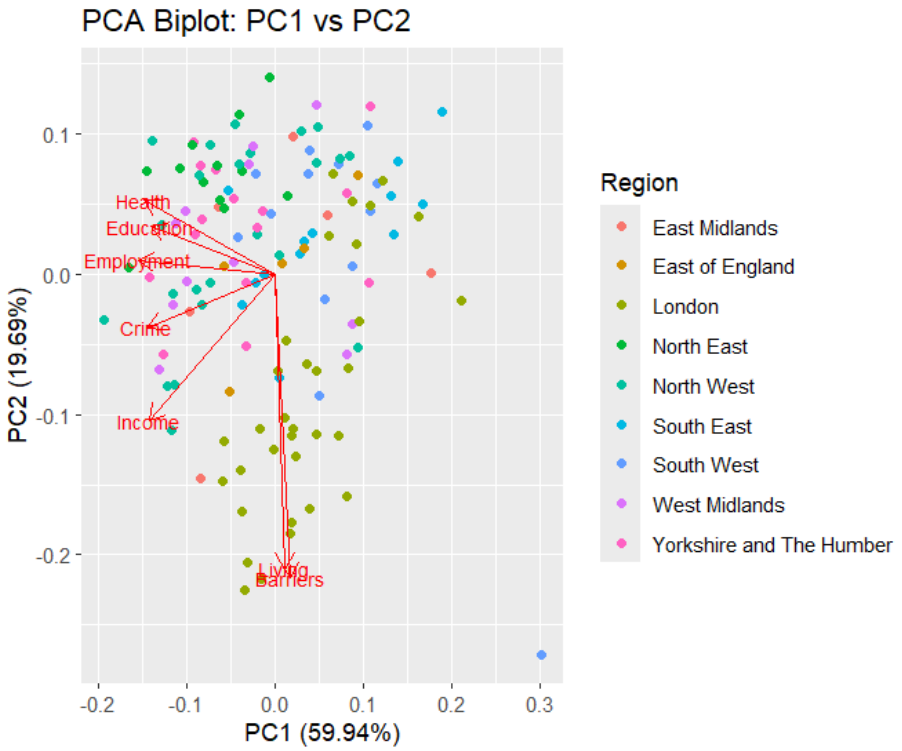
For PC1, Income, Employment, Education, Health and Crime all have large similar loadings, for PC2, Barriers and Living have the largest loadings and PC3 contrasts Barriers and Living, having opposite sign loadings so it represents a Living vs Barriers contrast.

Biplot: PC1 vs PC2

The PC1 vs PC2 biplot shows points and arrows to understand how IMD domains combine.

```
#PC1 vs PC2 biplot
autoplot(pca_results,
         data = imd,
         colour = "Region",
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 3) +
  ggtitle("PCA Biplot: PC1 vs PC2")
```

Output:

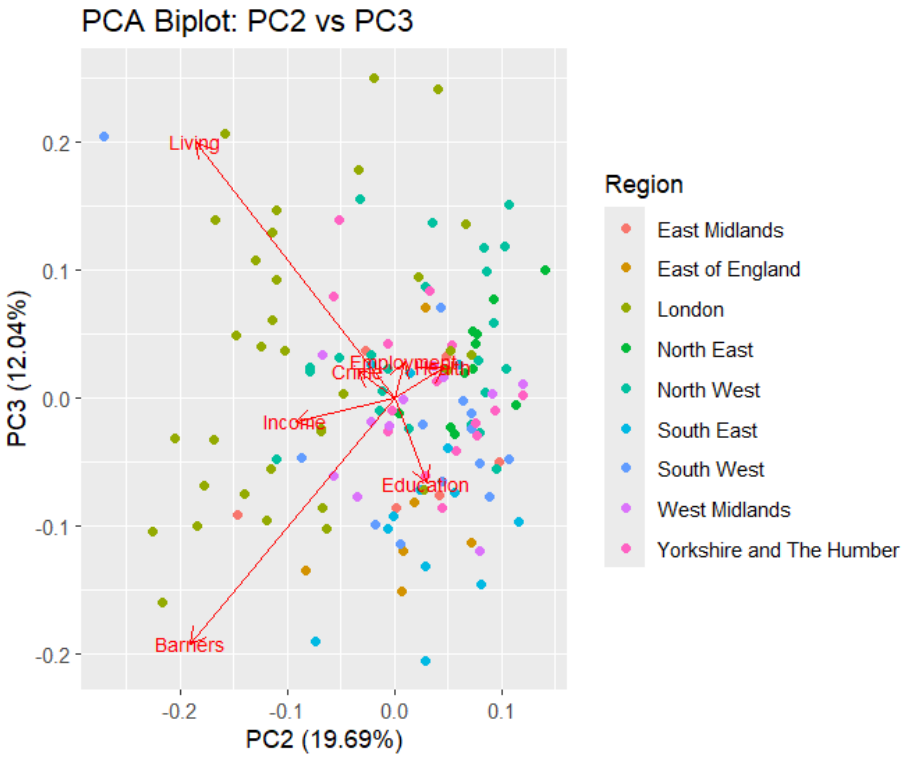


The arrows for Income, Employment, Education, Health and Crime all point in a similar direction, showing they are strongly correlated and define PC1. Barriers and Living point in a different direction and mainly define PC2. The coloring by region, shows that some regions tend to collect in particular areas of PCA space, meaning regional patterns are present in the IMD profiles.

Biplot for PC2 vs PC3

```
#PC2 vs PC3 biplot
autoplot(pca_results,
         data = imd,
         x = 2, y = 3,
         colour = "Region",
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 3) +
  ggtitle("PCA Biplot: PC2 vs PC3")
```

Output:



This biplot focuses on PC2 and PC3. PC2 is still determined by Barriers and Living, which have the largest arrows on the axis. PC3 computes a contrast between Living and Barriers, with the one to point in a positive direction and the other in a negative direction. Region coloring again allows us to see whether certain regions tend to have similar profiles on these dimensions.

Interpretation of PC1, PC2, PC3

PC1 has large similar loadings for Income, Employment, Education, Health and Crime. These domains move together and form a single major dimension, representing overall socio-economic and health deprivation

PC2 is dominated by Barriers and Living, forming a separate dimension from PC1, representing housing accessibility and environmental quality.

PC3 positions Barriers and Living in opposite directions.

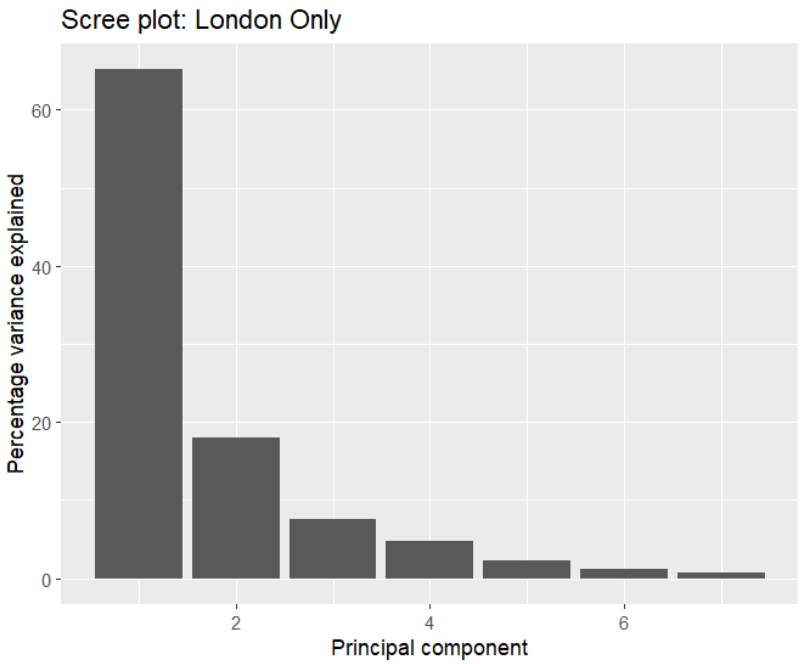
Effect of Region

When the points are colored by Region in the biplots, we can see that local authorities from the same Region often appear close to each other. This means regions tend to have similar IMD patterns. Overall, the plot shows that Region has an effect on where the areas appear in the PCA space.

b.

```
#London only with seven IMD domains
imd_london = imd %>% filter(Region == "London")

imd_lon_domains = imd_london %>%
  select(Income, Employment, Education, Health, Crime, Barriers, Living)

# PCA for London only
pca_london = prcomp(imd_lon_domains, scale. = TRUE)

# Scree plot
lon_var = (pca_london$sdev)^2
lon_var_explained = lon_var / sum(lon_var)
ggplot(NULL, aes(x = 1:7, y = lon_var_explained *100)) +
  geom_col() +
  xlab("Principal component") +
  ylab("Percentage variance explained") +
  ggtitle("Scree plot: London Only")
```

Output:

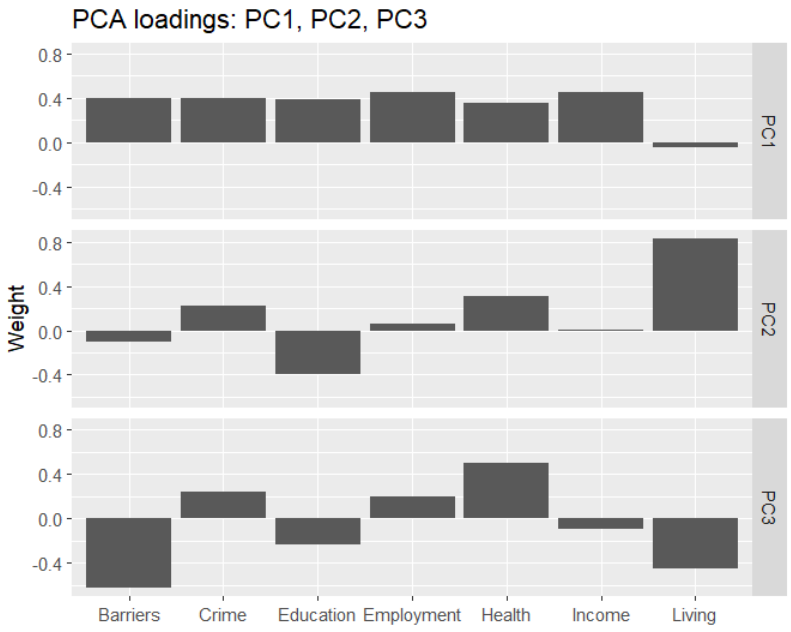


The scree plot for London shows that PC1-PC3 explains more than 90% of the total variation.

Loading Plot

```
# Loadings
pca_london$rotation[,1:3]

loadings = as.data.frame(pca_london$rotation[,1:3])
loadings$Symbol = row.names(loadings)
loadings= gather(loadings, key='Component', value='Weight', -Symbol)

ggplot(loadings, aes(x = Symbol, y = Weight)) +
  geom_bar(stat='identity') +
  facet_grid(Component ~ .) +
  xlab("") + ggtitle("PCA loadings: PC1, PC2, PC3")
```

Output:

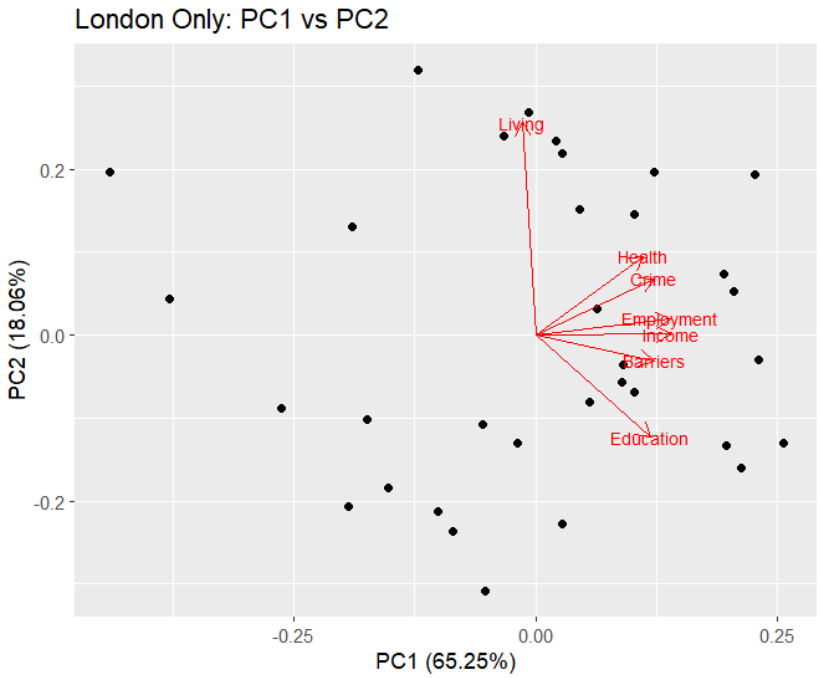


For PC1, all seven domains have positive loadings and are quite similar in size, meaning that PC1 represents a general deprivation pattern across London. London's PC1 shows a more balanced contribution from all domains unlike the national PCA. For PC2, Living has a very large positive loading, while Barriers is slightly negative. PC3 has Living and Barriers strongly negative with Health strongly positive. This means PC3 contrasts the Living/Barriers domains with Health, capturing smaller but meaningful differences with London.

Biplot PC1 vs PC2

```
# Biplot PC1 vs PC2
autoplot(pca_london,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 3) +
  ggtitle("London Only: PC1 vs PC2")
```

Output:



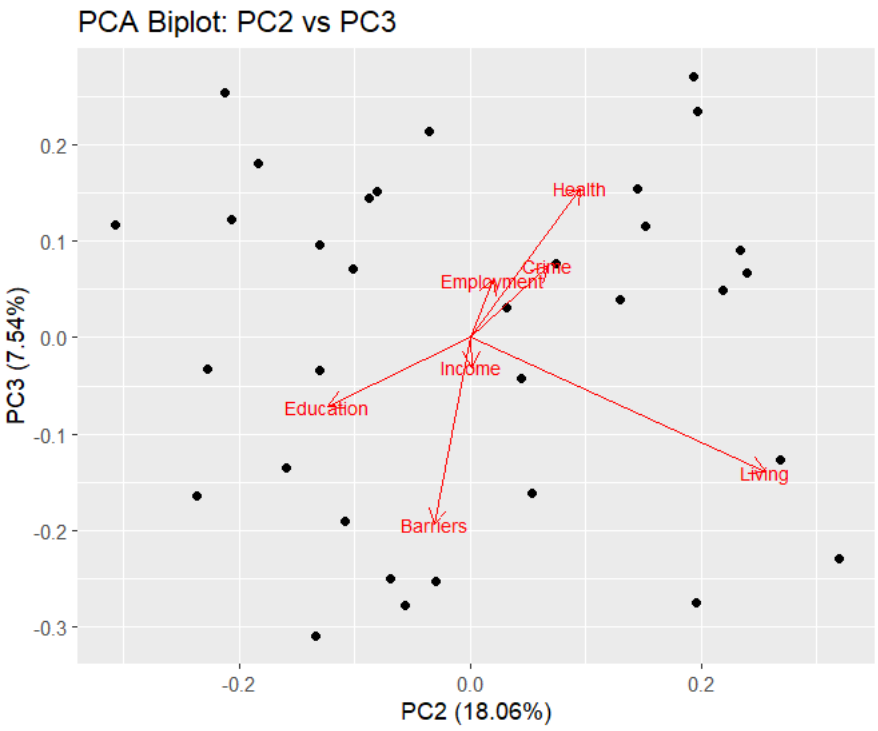
In the London only biplot, Income, Employment, Crime, Health, Education point in very similar directions along PC1, confirming that PC1 represents general deprivation. The Living points almost straight upward along PC2, showing that PC2 is strongly driven by variation in the Living within London.

The spread of points is narrower than the national plot, indication that London boroughs are more similar to each other overall.

Biplot PC2 vs PC3

```
#PC2 vs PC3 biplot
autoplot(pca_london,
         x = 2, y = 3,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 3) +
  ggtitle("PCA Biplot: PC2 vs PC3"))
```

Output:



In the London only PC2 vs PC3 biplot, the Living domain has a strong positive loading on PC2 and a strong negative loading on PC3, while Barriers has a negative loading on PC3. Income, Employment, Crime, Health, Education gather closely together near the center, which means they contribute much less to these components.

Overall

When comparing the PCA results for all English regions with the PCA using only London boroughs, the main difference is how the components are structured. In the London only PCA, PC1 still represents a general deprivation dimension, but the loadings for all seven IMD domains are more balanced.

Part IV:

Cluster Analysis using IMD domains

a.

```
library(tidyverse)
imd_domains = imd %>%
  select(Income, Employment, Education, Health, Crime, Barriers, Living)
```

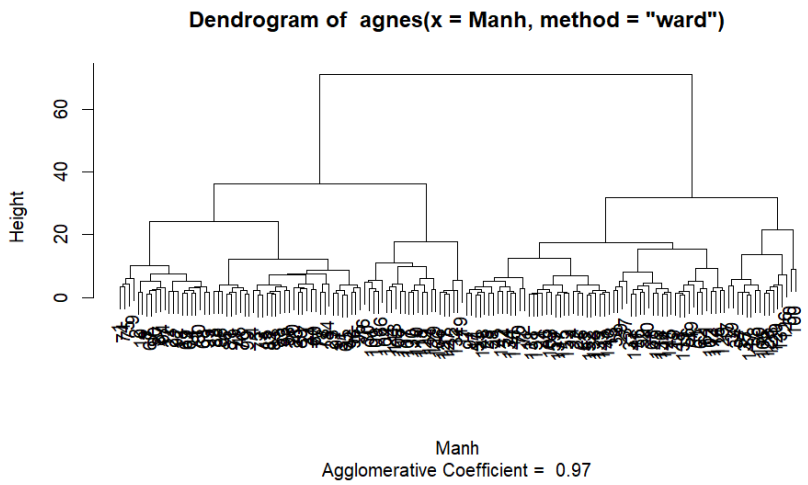
We will compute four clusterings: Euclidean + Ward, Euclidean+ Single linkage, Manhattan + Ward and Manhattan+ Single linkage. Each will give us an agglomerative coefficient, a measure of how well the data fit the clustering, the higher the better.

```
library(cluster)
#EUCLIDEAN AND SINGLE
Eucl = dist((domains_t), method='euclidean')
cluster_results= agnes ( Eucl, method='single')
cluster_results
ac_e_single = cluster_results$ac
#AC = 0.43

#MANHATTAN AND WARD
Manh = dist((domains_t), method='manhattan')
cluster_results= agnes ( Manh, method='ward')
cluster_results
ac_m_ward = cluster_results$ac
#AC = 0.63



#MANHATTAN AND SINGLE
Manh = dist((domains_t), method='manhattan')
cluster_results= agnes ( Manh, method='single')
ac_m_single = cluster_results$ac
cluster_results
#AC = 0.47

#EUCLIDEAN AND WARD
Eucl = dist((domains_t), method='euclidean')
cluster_results= agnes ( Eucl, method='ward')
cluster_results
ac_e_ward = cluster_results$ac
#AC = 0.64
```

Dendogram:

```
#dendogram

#EUCLIDEAN AND WARD
Eucl = dist((domains_t), method='euclidean')
cluster_results= agnes ( Eucl, method='ward')
cluster_results
ac_e_ward = cluster_results$ac
plot(cluster_results,which.plots=2)
#AC=0.64
```

Output:



This dendogram displays the clustering of all displays based on their seven IMD domains scores. Because there are many districts the labels are packed and individual names cannot by interpreted but the overall structure of the tree can still be understood. Districts that merge at low highs are very similar in their deprivation profiles, while districts that join at higher branches are more different. The use of Manhattan distance with Ward's method produces a clear, well-balanced cluster structure with an excellent agglomerative coefficient (0.97), indicating that the data fits this clustering extremely well. The dendrogram suggests that districts naturally form several distinct groups with similar patterns of deprivation.

```
#comparison table
comparison_table = data.frame(
  Distance = c("Euclidean", "Euclidean", "Manhattan", "Manhattan"),
  Method   = c("Single", "Ward", "Single", "Ward"),
  Agglomerative_Coefficient = c(ac_e_single, ac_e_ward, ac_m_single,
ac_m_ward))

comparison_table
```

Output:

```
comparison_table
   Distance Method Agglomerative_Coefficient
1 Euclidean Single                 0.8344735
2 Euclidean   Ward                 0.9645544
3 Manhattan Single                 0.8025332
4 Manhattan   Ward                 0.9705309
```

b.

```
# Transpose to cluster IMD domains
imd_domains = imd %>%
  select(Income, Employment, Education, Health, Crime, Barriers, Living)
domains_t = t(scale(imd_domains))
```

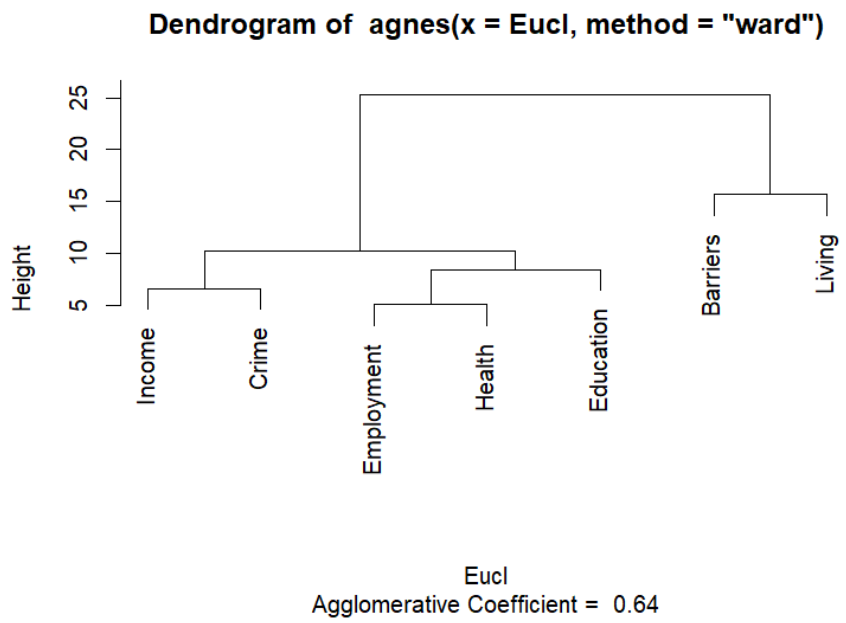
We will compute four clusterings: Euclidean + Ward, Euclidean+ Single linkage, Manhattan + Ward and Manhattan+ Single linkage. Each will give us an agglomerative coefficient, a measure of how well the data fit the clustering, the higher the better.

```
#EUCLIDEAN AND SINGLE
Eucl = dist((domains_t), method='euclidean')
cluster_results= agnes ( Eucl, method='single')
cluster_results
ac_e_single = cluster_results$ac
#AC = 0.43

#MANHATTAN AND WARD
Manh = dist((domains_t), method='manhattan')
cluster_results= agnes ( Manh, method='ward')
cluster_results
ac_m_ward = cluster_results$ac
#AC = 0.63



#MANHATTAN AND SINGLE
Manh = dist((domains_t), method='manhattan')
cluster_results= agnes ( Manh, method='single')
```

```
ac_m_single = cluster_results$ac
cluster_results
#AC = 0.47

#EUCLIDEAN AND WARD
Eucl = dist((domains_t), method='euclidean')
cluster_results= agnes ( Eucl, method='ward')
cluster_results
ac_e_ward = cluster_results$ac
#AC = 0.64
```

Dendogram:

```
#dendogram

#EUCLIDEAN AND WARD
Eucl = dist((domains_t), method='euclidean')
cluster_results= agnes ( Eucl, method='ward')
cluster_results
ac_e_ward = cluster_results$ac
plot(cluster_results,which.plots=2)
#AC=0.64
```



This dendrogram shows how the seven IMD domains cluster together based on their similarity across districts. These are two clear groups. Income, Crime, Employment, Health merge at low heights, meaning these domains behave similarly across local authorities. In contrast, Barriers and Living form their own separate cluster, joining the others only at much higher height. The agglomerative coefficient (0.64) is moderate, meaning the clustering structure is present but not extremely strong.

```
#comparison table
comparison_table = data.frame(
  Distance = c("Euclidean", "Euclidean", "Manhattan", "Manhattan"),
  Method   = c("Single", "Ward", "Single", "Ward"),
  Agglomerative_Coefficient = c(ac_e_single, ac_e_ward, ac_m_single,
ac_m_ward))
comparison_table
```

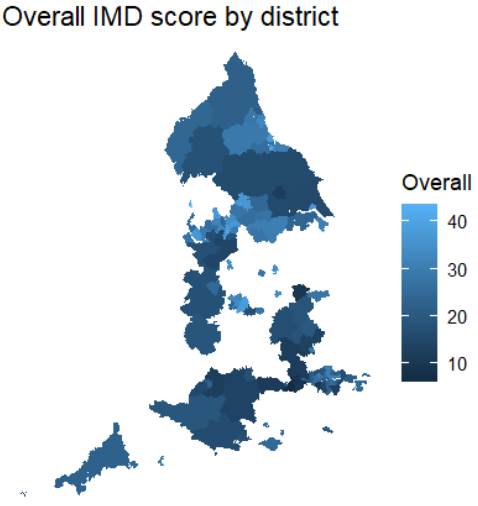
Output:

```
comparison_table
   Distance Method Agglomerative_Coefficient
1 Euclidean Single                 0.4389104
2 Euclidean   Ward                 0.6420693
3 Manhattan Single                 0.4704886
4 Manhattan   Ward                 0.6377784
```

Part V

In this section we investigate the spatial distribution using choropleth maps. We first map the Overall IMD score, which provides a direct view of how deprivation varies across districts. We then map PC1 and PC2, the first two principal components from part 3, in order to compare their spatial patterns with the Overall IMD map. Mapping these variables that PC1 and PC2 includes, allow us to see how different types of deprivation are geographically distributed.

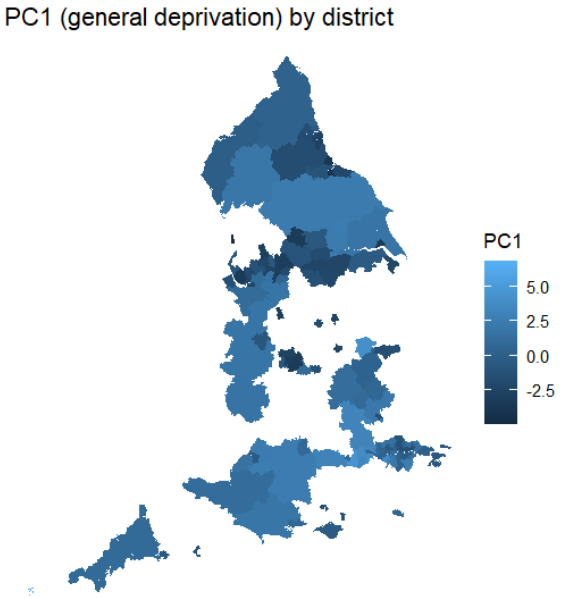
```
library (ggplot2)
library (sf)
library (tidyverse)
districts_map =
st_read("C:/Users/Chloe/Desktop/R/LAD2024/LAD_DEC_24_UK_BFC.shp")

pca_results = prcomp(imd %>% select(Income, Employment, Education,
Health, Crime, Barriers, Living), scale. = TRUE)

imd = imd %>%
  mutate(
    PC1 = pca_results$x[, 1],
    PC2 = pca_results$x[, 2])

map_data = districts_map %>%
  left_join(imd, by = "LAD24CD")



ggplot(map_data) +
  geom_sf(aes(fill = Overall), colour = NA) +
  ggtitle("Overall IMD score by district") +
  theme_void()
```



Task 3 (Group Task) — Multivariate Data Analysis

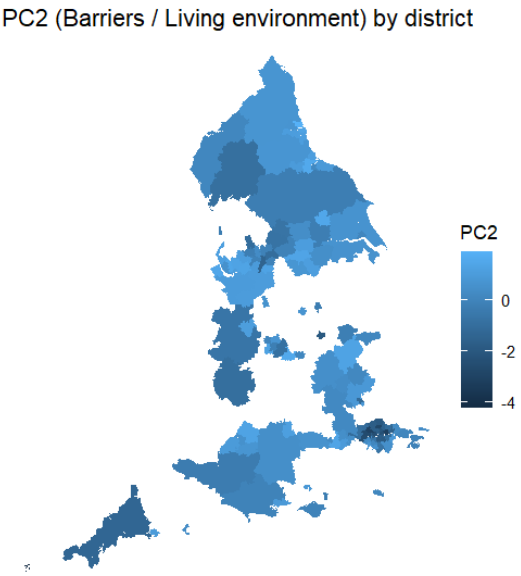
The overall IMD map shows clear spatial clustering of deprivation. Higher deprivation districts appear mainly in the North of England, the Midlands and some coastal areas while lower deprivation districts are concentrated in the South and South East. This suggests that deprivation is strongly geographically patterned rather than randomly distributed.

```
ggplot(map_data) +
  geom_sf(aes(fill = PC1), colour = NA) +
  ggtitle("PC1 (general deprivation) by district") +
  theme_void()
```



The PC1 map closely looks like the Overall IMD map. Districts with high Overall IMD are also high on PC1, especially in the North and North West.

```
ggplot(map_data) +
  geom_sf(aes(fill = PC2), colour = NA) +
  ggtitle("PC2 (Barriers / Living environment) by district") +
  theme_void()
```



The PC2 map shows a very different pattern from both Overall IMD and PC1. High PC2 values appear in coastal areas, particularly in the South West and parts of East England. This reflects that PC2 measures a separate dimension of deprivation related to Barriers and the Living, which does not follow the typical North-South deprivation divide.

Comparing the three maps shows that general deprivation (PC1) aligns strongly with the overall IMD geography, while PC2 highlights a different pattern. This demonstrates that deprivation is multi-dimensional and different aspects of deprivation affect different parts of the country.

Reference:

Wickham, H., Navarro, D. & Pedersen, T. L. (2023). *ggplot2: Elegant Graphics for Data Analysis* (3rd ed.). Chapter 6 — Maps, Section 6.2: Simple features (sf). Available at: https://ggplot2-book.org/maps.html#sec-simplefeatures