# Assignment 1 - Group 8

Elia Di Gregorio and Robert Auerbach

2024-03-31

## Contents



## Exercise A

The dependent variable `medv` shows the median value of owner-occupied homes in $1000s. We chose the following covariates for our linear model: 1) `crim`: per capita crime rate by town. 2) `zn`: proportion of residential land zoned for lots over 25,000 sq.ft. 3) `indus`: proportion of non-retail business acres per town. 4) `chas`: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). 5) `nox`: nitrogen oxides concentration (parts per 10 million).

Task: Create a function that takes your dependent variable and the covariates as inputs, and return a list with: – OLS point estimates for the intercept, slope parameters, and the error variance. – Suitable test statistics with corresponding p-values for the relevant coefficients. – Intervals of the coefficients for a confidence level of 95%.

```
pp_pred <- function(dependent_variable, covariates) {
  data <- data.frame(dependent_variable, covariates)
  model <- lm(dependent_variable ~ ., data)
  coefficients <- coef(model)
  se <- summary(model)$coefficients[, "Std. Error"]
  t_values <- coefficients / se
  p_values <- 2 * pt(abs(t_values), df = df.residual(model), lower.tail = FALSE)
  conf_int <- confint.default(model)
  results <- data.frame(
    "Coefficients" = coefficients,
    "Standard Errors" = se,
    "t-values" = t_values,
    "p-values" = p_values,
    "Confidence Intervals (95%)" = conf_int
```

```
  )

  return(results)
}
```

Call the function and print list:

```
result <- property_price_prediction(dependent_variable, covariates)
print(result)
```

```
##             Coefficients Standard.Errors  t.values     p.values
## (Intercept)  29.48994059      2.22434765 13.257793 1.365091e-34
## crim         -0.21851904      0.04389177 -4.978588 8.831134e-07
## zn            0.05511047      0.01743801  3.160365 1.671221e-03
## indus        -0.38348055      0.07944258 -4.827141 1.843157e-06
## chas          7.02622266      1.33711744  5.254754 2.198513e-07
## nox          -5.42465902      4.69550757 -1.155287 2.485247e-01
##             Confidence.Intervals..95...2.5.. Confidence.Intervals..95...97.5..
## (Intercept)                      25.13029931                       33.84958187
## crim                             -0.30454533                       -0.13249275
## zn                                0.02093261                        0.08928834
## indus                            -0.53918515                       -0.22777594
## chas                              4.40552064                        9.64692469
## nox                             -14.62768474                        3.77836670
```

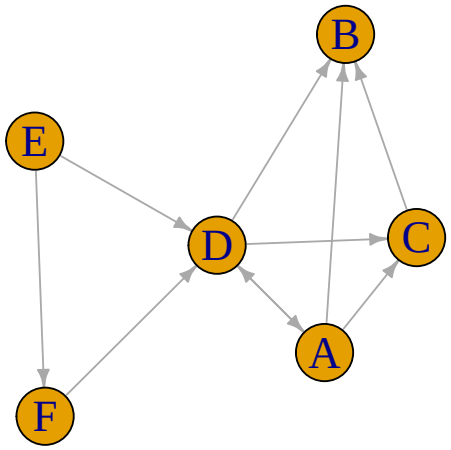
## Exercise B

Task: Come up with some network of interest, vaguely related to some real-world example (describe very briefly), with at least six agents and ten edges between them.

Our real world example is one of a supply chain network where there is rarely any reciprocity: construction of airplanes. The main vertices is the firm constructing the airplane (B), where all manufactured parts end up (high in-degree, low out-degree). On the other hand, among manufactures of airplane parts little reciprocity might occur as they need specialized part to finish their own parts (D and A).

The adjacency matrix in R.can be drawn as follows:

```
plot(g, edge.arrow.size = 0.5, vertex.label.cex = 1.5, vertex.size = 30)
```




```
print(adj_matrix)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## 5 x 5 sparse Matrix of class "dgCMatrix"
##   A B C D E
## A . 1 1 1 .
## B . . . . .
## C . 1 . . .
## D 1 1 1 . .
## E . . . 1 .
```

Who are the most and least central agents in the network? Name, explain, and try to quantify different notions of centrality.

We quantified the in- and out-degree of centrality. The in-degree shows how many links are directed towards are node, while the out-degree shows how many links are directed from a node (to another). A node from which a link goes out to another is considered a supplier, a node which is receiving links is considered a buyer. The degree_table shows how many links go in and out from each node. This way we can quantify the most central buyer and supplier in our network. According to the table B and D have the most links directed towards them while E has none directed towards it. Hence, B and C are the most central "buyers", E the least. Also, A and D have the most links directed towards others while B has none directed towards others. Hence A and D are the most central "suppliers" and B the least.

```
degree_table
```

```
##     A B C D E F
## In  1 3 2 3 0 1
## Out 3 0 1 3 2 1
```

How would centrality change if you considered a row-normalized network instead?

We calculate the sums of each row of the adjacency matrix and then divide the addjacency matrix by the row sums. Again, we compute the centrality measures of in- and out-degree. As expected (based on the slides) the out-degree of the agents are equalized to 1. Also, the most central buyers are now B and D instead of B and C. Clearly the row-normalization leads to a distortion.

```
degree_table_norm
```

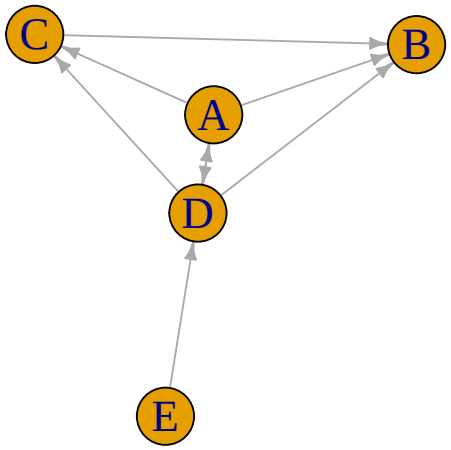
```
##             A        B         C        D E   F
## In  0.3333333 1.666667 0.6666667 1.833333 0 0.5
## Out 1.0000000 0.000000 1.0000000 1.000000 1 1.0
```

How would the network change if you removed or added a specific agent? We removed agent "F" and repeated the steps from above.

The most central buyer is now B alone, E still is the least central buyer. The most central supplier is still A and D, the least is still B.

We also compared the reciprocity and transitivity of both networks. The reciprocity is higher with the agent F removed, which is due to the fact that the network has a higher share of reciprocated links relative to the amount of possible reciprocated links. Also, the so called clustering coefficient (Transitivity) is higher in the Network with fewer agents as there is a 10% higher probability that vertices are connected. This is especially true for a network with low reciprocity such as ours.

```
plot(g1, edge.arrow.size = 0.5, vertex.label.cex = 1.5, vertex.size = 30)
```




```
degree_table_g1
```

```
##     A B C D E
## In  1 3 2 2 0
## Out 3 0 1 3 1
```

```
rectrans_table
```

```
##              Network Complete Network Removed
## Reciprocity        0.2000000            0.25
## Transivity         0.7142857            0.80
```

# Exercise C

In this exercise, we will work with spatial projections of Europe's NUTS-2 regions. Europe counts 331 subregions, however, for the sake of visualization, we excluded overseas territories and focused on the main continental area.

## Different projections

We accessed its spatial data directly from the GISCO source via `get_eurostat_spatial()` function. By default this function loads the EPSG-4326 projection of the map, corresponding to the World Geodetic System 1984 ensemble (WGS84). It is based on a geocentric datum, meaning it defines the Earth's shape as an ellipsoid (a flattened sphere) rather than a perfect sphere.

```
Europe <- get_eurostat_geospatial(
  resolution = "01",
  nuts_level = 2,
  year = 2021) %>%
  filter(!NUTS_ID %in% c("FRY1", "FRY2", "FRY3", "FRY4", "FRY5", "FRZZ",
                         "PT20", "PT30", "PTZZ", "ES70", "ESZZ", "NOOB", "NOZZ"))
```

```
## Extracting data using giscoR package, please report issues on https://github.com/rOpenGov/giscoR/issu
```

```
st_crs(Europe)
```

```
## Coordinate Reference System:
##   User input: EPSG:4326
##   wkt:
## GEOGCRS["WGS 84",
##     ENSEMBLE["World Geodetic System 1984 ensemble",
##         MEMBER["World Geodetic System 1984 (Transit)"],
##         MEMBER["World Geodetic System 1984 (G730)"],
##         MEMBER["World Geodetic System 1984 (G873)"],
##         MEMBER["World Geodetic System 1984 (G1150)"],
##         MEMBER["World Geodetic System 1984 (G1674)"],
##         MEMBER["World Geodetic System 1984 (G1762)"],
##         MEMBER["World Geodetic System 1984 (G2139)"],
##         ELLIPSOID["WGS 84",6378137,298.257223563,
##             LENGTHUNIT["metre",1]],
##         ENSEMBLEACCURACY[2.0]],
##     PRIMEM["Greenwich",0,
##         ANGLEUNIT["degree",0.0174532925199433]],
##     CS[ellipsoidal,2],
##         AXIS["geodetic latitude (Lat)",north,
##             ORDER[1],
##             ANGLEUNIT["degree",0.0174532925199433]],
##         AXIS["geodetic longitude (Lon)",east,
##             ORDER[2],
##             ANGLEUNIT["degree",0.0174532925199433]],
##     USAGE[
##         SCOPE["Horizontal component of 3D system."],
##         AREA["World."],
##         BBOX[-90,-180,90,180]],
##     ID["EPSG",4326]]
```

To use another projection and/or CRS we employed the `st_transform()` function and recurred to the Lambert Azimuthal Equal Area Projection. It preserve the relative sizes of areas on the Earth's surface. This

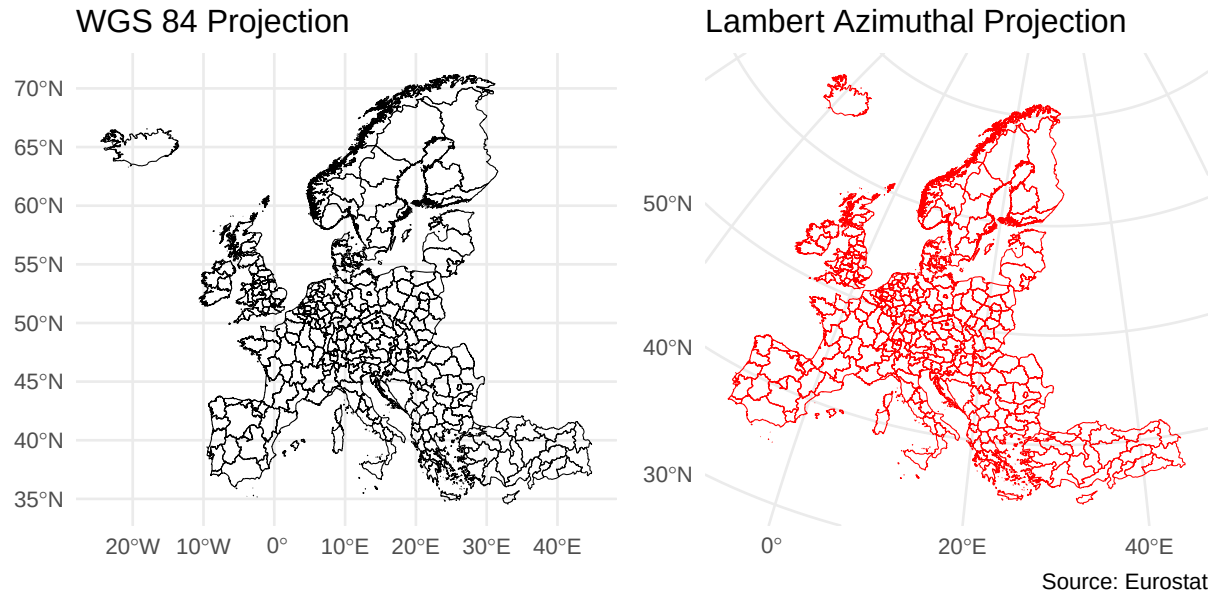
means that areas on the map are represented accurately in relation to each other in terms of size, making it suitable for thematic mapping and spatial analysis. However it projects the Earth's surface onto a plane tangent to a specific point (the center of the projection). The difference in projections is highlighted by the following map:

```
plot_Europe <- ggplot() +
  geom_sf(data = Europe, color = "black", fill = NA) +
  theme_minimal() +
  labs(title = "WGS 84 Projection", size = 12)

Europe_laea <- st_transform(Europe, crs = "+proj=laea +lat_0=45 +lon_0=30")
plot_laea <- ggplot() +
  geom_sf(data = Europe_laea, color = "red", fill = NA) +
  theme_minimal() +
  labs(title = "Lambert Azimuthal Projection", size = 12)

EU_plot_1 <- plot_Europe + plot_laea +
  plot_layout(ncol = 2) +
  labs(caption="Source: Eurostat")

plot(EU_plot_1)
```




## Map and Dataset

The dataset used was `tgs00111`: Nights spent at tourist accommodation establishments by NUTS 2 regions (here the Metadata for consultation). The dataset covers internal tourism, in other words tourism flows within the country (domestic tourism) or from abroad to destinations in the country (inbound tourism) for the year 2022. We further enhanced the dataset with two additional columns with the share of domestic and foreign tourist overnights (continuous scale), and one column for a factor variable based on the condition

that if the domestic tourism share of overnight stay is greater than 50%, the tourist is labeled as "Domestic Tourist"; otherwise, they are labeled as "Foreign Tourist".

```
data_nightstay <- get_eurostat("tgs00111",
                               time_format = "raw",
                               filters = list(
                                 TIME_PERIOD = "2022"
                               )) %>%
  merge(.,Europe, by = "geo", all=TRUE) %>%
  filter(!NUTS_ID %in% c("FRY1", "FRY2", "FRY3", "FRY4", "FRY5", "FRZZ",
                         "PT20", "PT30", "PTZZ", "ES70", "ESZZ", "NOOB", "NOZZ")) %>%
  pivot_wider(., names_from = c_resid, values_from = values) %>%
  mutate(DOM_SHARE = DOM/TOTAL,
         FOR_SHARE = FOR/TOTAL,
         TURIST = factor(ifelse(DOM_SHARE > 0.5, 0, 1), levels = c(1, 0), labels = c("Foreign Tourist",
```

```
## Dataset query already saved in cache_list.json...
```

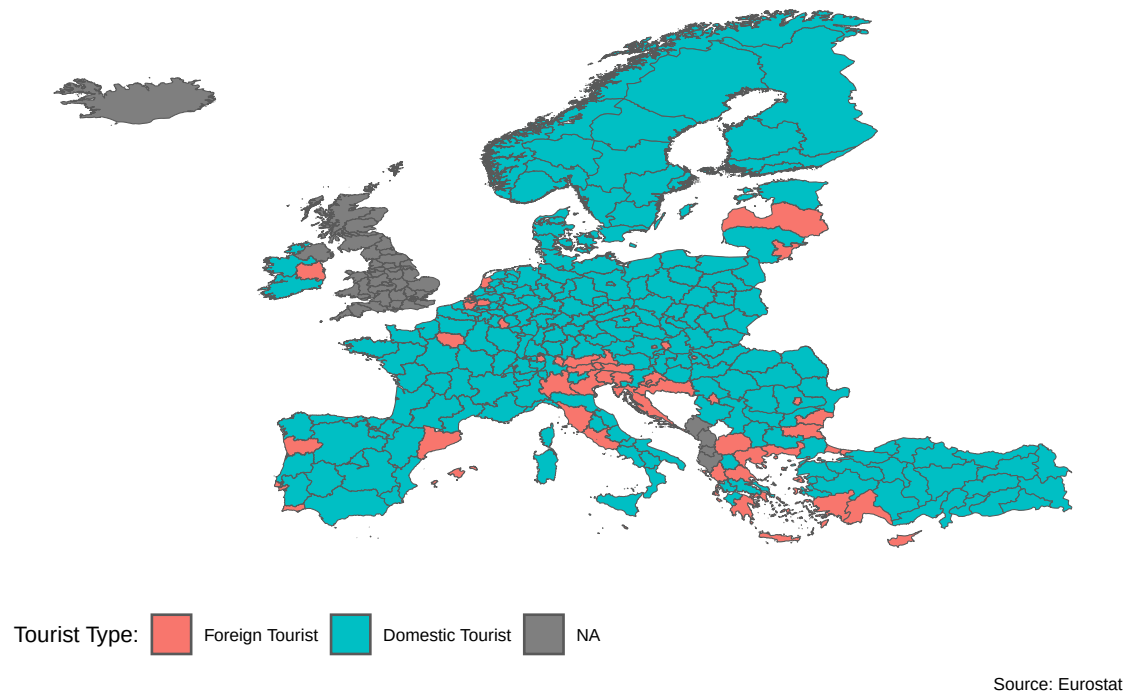
```
## Reading cache file C:\Users\eliad\AppData\Local\Temp\RtmpWYs5sQ/eurostat/d8af3781cdc4d6a8477f61ecf40
```

```
## Table  tgs00111  read from cache file:  C:\Users\eliad\AppData\Local\Temp\RtmpWYs5sQ/eurostat/d8af378
```

By plotting the latter variable, we were able to group together regions based on whether the majority of overnight stays in hotels were by domestic tourists or foreign tourists.

```
EU_plot_2 <- ggplot(data_nightstay) +
  geom_sf(aes(fill = TURIST, geometry = geometry)) +
  theme_map() +
  labs(x = NULL, y = NULL,
       title = "Distribution of Tourist Overnight-Stay by Origin",
       subtitle = "Europe - NUTS2 Level",
       caption = "Source: Eurostat")+
  guides(fill = guide_legend(title = "Tourist Type:")) +
  theme(legend.position = "bottom",
        plot.title = element_text(face = "bold"))

plot(EU_plot_2)
```




Not surprisingly, these where the regions with the most famous "Instagrammable" destinations such as:

- Italy: Venice, Milan, Florence, or Rome,
- Spain: Barcelona and the Balearic Islands,
- France: Paris,
- Austria: ski hubs in the Alpine regions and Vienna,
- the capital region of Brussels, the Netherlands, Portugal, Ireland, and many other Central and Eastern European (CEE) countries,
- the coastal regions of Greece and Turkey (Antalya and Bodrum).

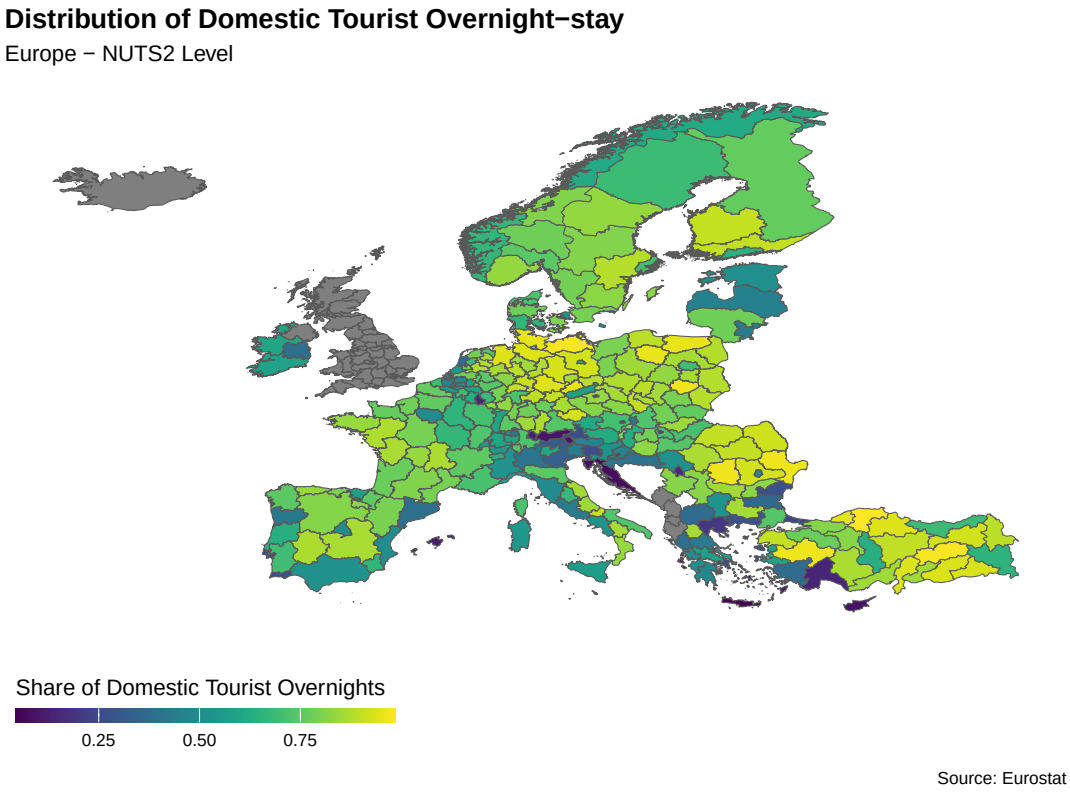
Overall, the plot conveys a picture of the main touristic hotspots of Europe where tourists from abroad mostly arrive, However it might be biased as it does not consider other touristic accomodations such as Airbnb, mostly used by young Europeans.

The plot for the continuous scale variable focused more on the domestic dimension of tourism and represented the share of domestic tourist overnight stays in different Turkish regions.

```
EU_plot_3 <- ggplot(data_nightstay) +
  geom_sf(aes(fill = DOM_SHARE, geometry = geometry)) +
  theme_map() +
  labs(x = NULL, y = NULL,
       title = "Distribution of Domestic Tourist Overnight-stay",
       subtitle = "Europe - NUTS2 Level",
       caption = "Source: Eurostat") +
  theme(legend.position = "bottom",
        plot.title = element_text(face = "bold")) +
  scale_fill_viridis(option = "viridis",
                     direction = 1,
                     name = "Share of Domestic Tourist Overnights",
                     guide = guide_colorbar(direction = "horizontal",
```

```
                                        barheight = unit(2, units = "mm"),
                                        barwidth = unit(54, units = "mm"),
                                        draw.ulim = TRUE,
                                        title.position = "top"))

plot(EU_plot_3)
```




Besides stressing the point made by the previous map, this time we get more insights on the distribution of Domestic tourists in Europe. Germany appears to see a higher flow of in-country movement in its Bundesländer. So does some Italian regions of the eastern coast, which are notably more difficult to reach via international connections like Abrruzzi and Molise or Calabria. Similarly, we notice a smaller share of foreigners in lateral CEE countries and Turkey, where the inland regions are mostly hosting domestic tourists.

## Storing visualizations

There are two conceptually different ways to store visualizations: raster-based and vector-based formats.

Raster-based formats like PNG and JPEG store images as grids of pixels, making them suitable for complex color gradients and detailed images. However, they are resolution-dependent, which means they can lose quality when scaled up.

On the other hand, vector-based formats such as SVG and PDF store image data using mathematical formulas to define shapes, lines, and colors. This makes them ideal for visualizations with geometric shapes, charts, maps, and illustrations where scalability and high-quality printing are crucial. Unlike raster formats, vector graphics are resolution-independent and can be scaled without loss of quality.

For visualizations created using R and ggplot2, which inherently produce vector graphics, it is recommended to save them in vector-based formats like SVG or PDF. These formats maintain sharpness and clarity when scaled to any size, making them suitable for presentations, printing, and high-resolution displays. SVG is

particularly useful for web-based graphics and interactive visualizations, while PDF is excellent for high-quality printing and cross-platform compatibility.

```
# Example on how to save the above-displayed plots:
ggsave("EU_plot_1.svg", plot = EU_plot_1, path = "./plot", device = "svg", width = 30, height = 15, uni
ggsave("EU_plot_2.svg", plot = EU_plot_2, path = "./plot", device = "svg", width = 30, height = 15, uni
ggsave("EU_plot_3.svg", plot = EU_plot_3, path = "./plot", device = "svg", width = 30, height = 15, uni
```

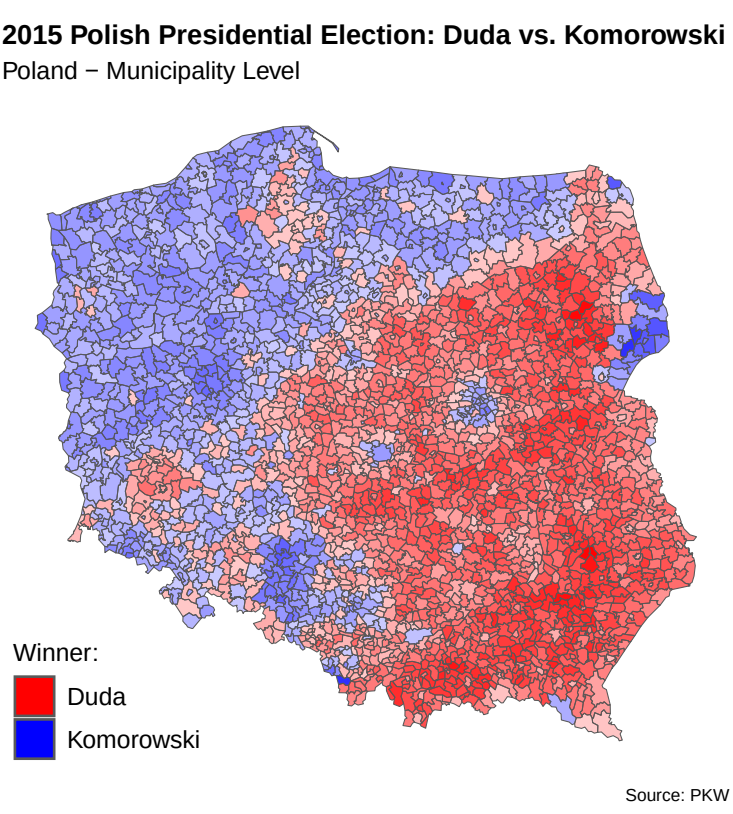
# Exercise D

## Comparison of support for Komorowski and Duda

```
df <- pol_pres15 %>%
  mutate(Winner = factor(ifelse(II_Duda_share > 0.5, 1, 0), levels = c(1, 0), labels = c("Duda", "Komor
         Majority_votes_brkdwn =ifelse(II_Duda_share > 0.5, II_Duda_share, II_Komorowski_share))

PL_plot_1 <- ggplot(df) +
  geom_sf(aes(fill = Winner, alpha = Majority_votes_brkdwn, geometry = geometry)) +
  scale_fill_manual(values = c("Duda" = "red", "Komorowski" = "blue")) +
  scale_alpha(range = c(0.2, 1), guide = "none") +
  theme_map() +
  labs(x = NULL, y = NULL,
       title = "2015 Polish Presidential Election: Duda vs. Komorowski",
       subtitle = "Poland - Municipality Level",
       caption = "Source: PKW") +
  guides(fill=guide_legend(title = "Winner:")) +
  theme(
    plot.title = element_text(size = 11, face = "bold"),
    plot.subtitle = element_text(size = 10),
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 10))

plot(PL_plot_1)
```




The provided R code generates a thematic map representing the outcomes of the second round of the 2015

Polish presidential elections at the municipality level. The data is grouped based on whether the winning candidate in each municipality was Duda or Komorowski. The plot uses different shades of red or blue (representing Duda or Komorowski, respectively) to indicate varying levels of majority support for the respective candidate.

The transparency of the colors also varies to reflect the degree of majority/support for the winning candidate in each municipality. Darker shades indicate a higher level of majority/support, while lighter shades represent a lower level of majority/support. This additional dimension helps visually emphasize areas where the winning candidate had a more significant lead in terms of voter share.

## Postal voting envelopes anomalies

Three types of anomalies are identified:

- `anomaly_invalid`: Checks if there are more invalid voting papers than postal voting envelopes received in either round of voting.
- `anomaly_spelling`: Identifies anomalies related to discrepancies in the number of invalid voting papers and specific errors on postal voting envelopes, such as missing declarations, signatures, voting envelopes, or signs of envelope opening.
- `anomaly_count`: Detects anomalies in the count of voting envelopes placed in the ballot box versus the number of voting papers taken from envelopes in either round of voting.

```
data <- pol_pres15 %>%
  mutate(anomaly_invalid = ifelse(I_invalid_voting_papers > I_postal_voting_envelopes_received | II_inva
         anomaly_spelling = ifelse(I_invalid_voting_papers > 0 & I_PVE_of_which_no_declaration == 0 & I_
                                   II_invalid_voting_papers > 0 & II_PVE_of_which_no_declaration == 0
         anomaly_count = ifelse(I_voting_envelopes_placed_in_ballot_box != I_of_which_voting_papers_take
         anomaly = ifelse(anomaly_invalid == 1 | anomaly_spelling == 1 | anomaly_count == 1, anomaly_inv
  mutate(anomaly_type = case_when(
    anomaly_invalid == 0 & anomaly_spelling == 0 & anomaly_count == 0 ~ "None",
    anomaly_invalid == 1 & anomaly_spelling == 0 & anomaly_count == 0 ~ "Invalid Anomaly",
    anomaly_invalid == 0 & anomaly_spelling == 1 & anomaly_count == 0 ~ "Spelling Anomaly",
    anomaly_invalid == 0 & anomaly_spelling == 0 & anomaly_count == 1 ~ "Extraction Anomaly",
    anomaly_invalid == 1 & anomaly_spelling == 1 & anomaly_count == 0 ~ "Invalid + Spelling Anomaly",
    anomaly_invalid == 1 & anomaly_spelling == 0 & anomaly_count == 1 ~ "Invalid + Extraction Anomaly",
    anomaly_invalid == 0 & anomaly_spelling == 1 & anomaly_count == 1 ~ "Extraction + Spelling Anomaly"
    TRUE ~ "Other"
  ))

# Need x and y coordinates for point, thus convert the geometry column to sf object
data_sf <- st_as_sf(data, wkt = "geometry")
centroid <- st_centroid(data_sf)
```

```
## Warning: st_centroid assumes attributes are constant over geometries
```

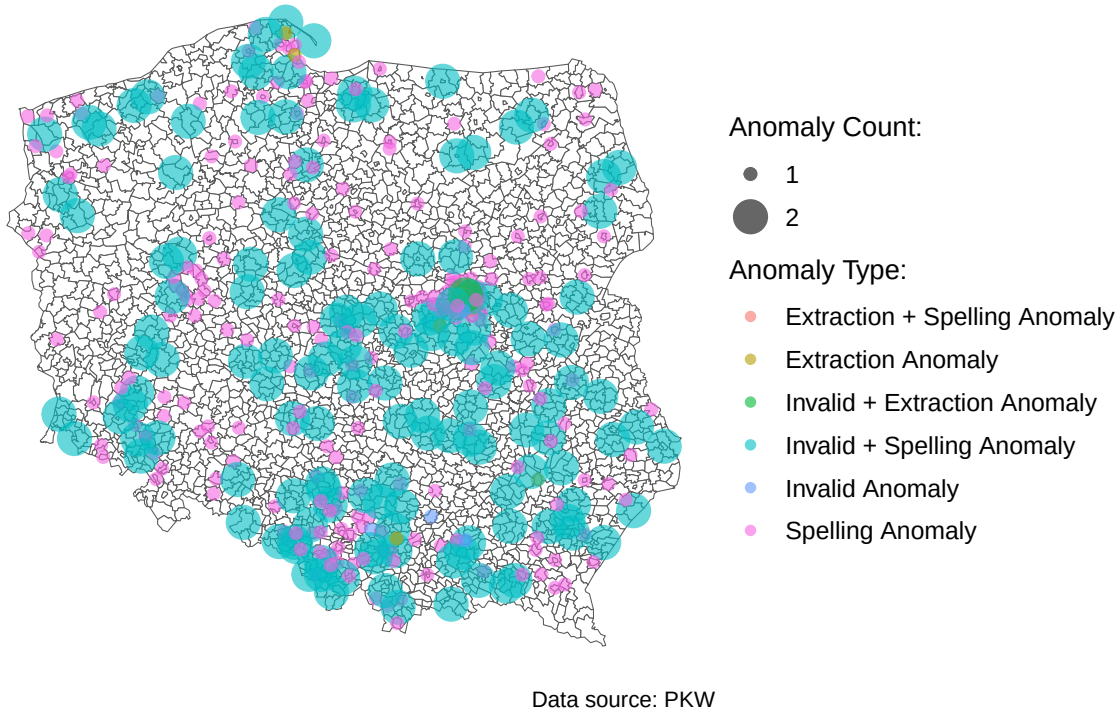
```
data_with_centroid <- cbind(data, st_coordinates(centroid))

PL_plot_2 <- ggplot(data_with_centroid) +
  geom_sf(fill = "white") +
  geom_point(data = subset(data_with_centroid, anomaly > 0), aes(x = X, y = Y, color = as.factor(anomaly
  theme_void() +
  labs(title = "2015 Polish Presidential Election: anomalies in PVE",
       subtitle = "Poland - Municipality Level",
       caption = "Data source: PKW",
       color = "Anomaly Type:",
```

```
       size = "Anomaly Count:") +
  theme(
    plot.title = element_text(size = 11, face = "bold"),
    legend.text = element_text(size = 10))

plot(PL_plot_2)
```

```
## Warning: Using size for a discrete variable is not advised.
```




Anomalies are represented as points (`geom_point()`) on the map, with each point's color indicating the type of anomaly and size representing the how many anomalies types were overall detected.In this way, one can highlight areas where specific types of anomalies or their combinations were observed more frequently.

By visually exploring anomalies in the handling of voting materials, it appears that this was a widespread problem, involving several municipalities, although no clear pattern can be detected.

Notice that out of the three anomalies, the one about spelling is the "weak" one. Despite postal voting envelopes were return with a declaration, and a sealed voting envelope, the count of invalid PVE is still higher. It is not clear though, whether this is due to some further unrecorded issues (e.g. the misspelling of the candidate's name, thus the label for this variable). Further material, information or context should be validate the theory.

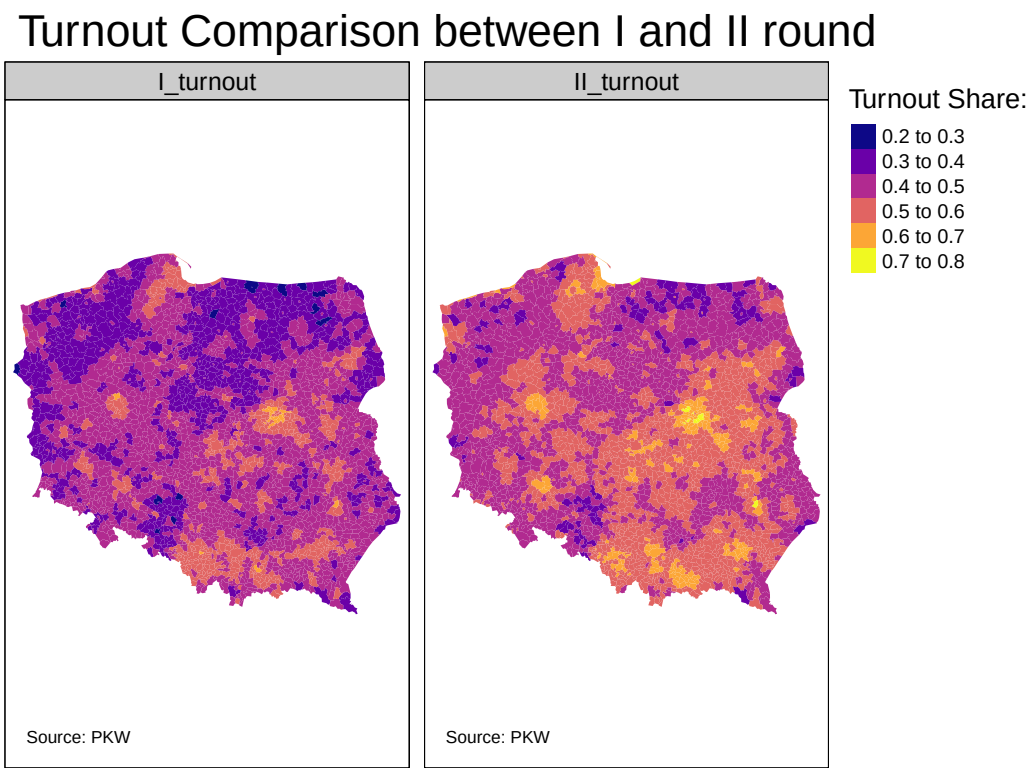
## Turnout for each election round

For the final visualization, employing `tm_shape()` function, Turnout share in the first and second round of the presidential election was plotted. We made use of the `tm_facets` specification to combine in a single plot both heatmaps.

```r
ds <- pol_pres15 %>%
  pivot_longer(cols = c(I_turnout, II_turnout), names_to = "election", values_to = "turnout")

PL_plot_3 <- tm_shape(ds) +
  tm_fill(col = "turnout", title = "Turnout Share:", palette = "plasma" ) +
  tm_borders(invisible()) +
  tm_facets(by = "election", free.scales = FALSE) +
  tm_layout(main.title = "Turnout Comparison between I and II round",
            title.position = c("center", "top"),
            frame = TRUE) +
  tm_credits("Source: PKW", position = "left")

PL_plot_3
```




Overall, it appears that in the second round there was a higher public involvement. In rural areas this was the less stronger than in the large urban areas, where participation peaked as far as reaching almost 80%.