Transport Data Science

Transportation Networks, Population Distribution and Air Quality: A Data Driven Exploration of Interconnected Systems in the UK

Sathyasri Sudhakar

Introduction

The United Kingdom has a diverse and dynamic population distribution and intricate transportation networks for both railways and roadways(Regt et al. 2019). This report aims to comprehend the many transport networks, identify the busiest hubs, ascertain the population density in these locations, and establish a connection between them and the air quality in that region.

This report is structured as follows: we first look into an overview of the different transportation networks, population distribution and air quality. Following this different data exploratory analysis is done to understand the data better and come up with findings to conclude the analysis.

Input Data and Cleaning

Importing of different Libraries

To import the different data sets and work with them, various libraries were loaded, after which different shape files and data files were imported.

```
packages <- c('ggplot2','maps','osmdata','sf','tidyverse','ggmap','giscoR','ggfx','readxl','writexl',
              'grid','dbplyr','viridis','tmap','ggnewscale','dplyr','sp','openair','gridExtra')
for (i in 1:length(packages)){
  library(packages[i], character.only=TRUE)}
```

```
## Data (c) OpenStreetMap contributors, ODbL 1.0. https://www.openstreetmap.org/copyright
```

```
## Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::map()    masks maps::map()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## i Google's Terms of Service: <https://mapsplatform.google.com>
##   Stadia Maps' Terms of Service: <https://stadiamaps.com/terms-of-service/>
##   OpenStreetMap's Tile Usage Policy: <https://operations.osmfoundation.org/policies/tiles/>
## i Please cite ggmap if you use it! Use `citation("ggmap")` for details.
##
## Attaching package: 'dbplyr'
##
```

```
##
## The following objects are masked from 'package:dplyr':
##
##     ident, sql
##
##
## Loading required package: viridisLite
##
##
## Attaching package: 'viridis'
##
##
## The following object is masked from 'package:maps':
##
##     unemp
##
##
## Breaking News: tmap 3.x is retiring. Please test v4, e.g. with
## remotes::install_github('r-tmap/tmap')
##
##
## Attaching package: 'gridExtra'
##
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

Importing of different shapefiles and datasets

Different shape files that were imported are:

- UK - The shapefile from IGISMap (Acharya 2018) contains the UK's boundaries or main outline.
- UK_railroads - shapefile from OSDataHub (osdatahub.os.uk n.d.) contains information on the rail route for the entire UK.
- London_roads - shapefile from London Datastore (Datastore n.d.) contains the London road map of all the major and minor roads.
- London_tube - shapefile got from Cornell Bower University's Database [Networks (2014)](Core-Periphery 2019), contains London's subways rail route.
- UK_lad - shapefile from ONS (Portalx n.d.) contains boundary data of the local authority districts in the UK.
- UK_major_roads - shapefile got OS OpenRoads (Survey 2021), containing a route map containing all the major roads in the UK.

```
UK<-st_read("united_kingdom_Country_Boundary_level_1.shp",quiet=TRUE)
UK_railroutes<-st_read("GBR_rails.shp",quiet=TRUE)
London_roads<-st_read("gis_osm_landuse_a_free_1.shp",quiet=TRUE)
London_tube<-st_read("London Train Lines.shp",quiet=TRUE)
UK_lad<-read_sf("LAD_DEC_2022_UK_BUC_V2.shp",quiet=TRUE)
UK_major_roads<-st_read("MRDB_2022_published.shp",quiet=TRUE)
```

Different Data files that were imported are:

- Stations - This .csv dataset was got from Github (Wheatley 2024). It contains information on the different stations present in the UK and their locations.
- pass_data — This .xlsx dataset, obtained from the Office of Rail and Road Transport in the UK (Rail and Road 2023), contains information on the number of journeys made by passengers in the UK by train.
- passenger_count- This .xlsx dataset, obtained from the Office of Rail and Road Transport (Rail and Road 2023), contains information on the total number of entries and exits for all stations in the UK.

- UK_popden - This .xls dataset, obtained from the Office of National Statistics (National Statistics 2022), contains information on the population count in each location across the UK.
- UK_Air_2022 - holds information on air pollution in the UK in 2022, obtained from the open-air dataset.(Carslaw and Ropkins 2012)
- UK_roads_data—A .csv dataset obtained from the Department of Transport UK database (Dft.gov.uk 2022) contains information on the number of vehicle movements per road in the UK.
- poll_air_gg - This .xlsx dataset was obtained from the Department of Transport UK database (Gov. 2022), which contains information on the different greenhouse gases present in the atmosphere. The based_trans — This .xlsx dataset was obtained from the Department of Transport UK database (Gov. 2022), which contains information on the greenhouse gases emitted by different transportation systems in the UK.
- UK_poll_LAD — This .csv dataset was from the Department of Transport UK database (Gov. 2022), which contains information on the greenhouse gas emissions by each Local authority district in the UK.

```
stations<-read.csv("stations.csv")
pass_data<- read_excel("Passenger_Journeys_UK.xlsx",sheet="1220_Passenger_journeys")
passengers_count<-read_excel("Enteries_Exit.xlsx")
UK_popden<-read_excel("ukpopestimatesmid2021on2021geographyfinal.xlsx",sheet='MYE2 - Persons')
UK_Air_2022<-importUKAQ(year = 2022,source = "aurn",data_type = "annual",pollutant = "all",
    hc = FALSE,meta = TRUE,meteo = TRUE,ratified=FALSE,to_narrow=FALSE,verbose=FALSE,progress=TRUE)
UK_roads_data<-read_csv("UK_roads_data.csv")
```

```
## Rows: 157992 Columns: 35
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): Direction_of_travel, Region_name, Region_ons_code, Local_authorit...
## dbl  (24): Count_point_id, Year, hour, Region_id, Local_authority_id, Eastin...
## date  (1): Count_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
poll_air_gg<-read_excel("Air poll due to transport.xlsx", sheet="Greenhouse gases")
based_trans<-read_excel("env0201.xlsx", sheet="1")
UK_poll_LAD<-read_csv("2005-21-local-authority-ghg-emissions-csv-dataset-update-060723.csv",
                      show_col_types = FALSE)
```

Cleaning and Transforming the data for further Analysis

Each data set undergoes different data cleaning and transformations based on the requirement. The process and transformations for each data set are mentioned below.

1. pass_data - Basic transformations that would help make the visualisation of the number of tickets sold yearly much better.

```
pass_25<-pass_data[-(1:126), , drop = FALSE]
pass_25$Year=substr(pass_25$`Time period`,5,9)
pass_25<-within(pass_25, rm(`Time period`))
pass_25$`Total Journeys` <- as.numeric(pass_25$`Total Journeys`)
```

2. passengers_count - Multiple transformations have been done to this data set, as it contains much information. The transformations done are:

- #1 First, I merged the stations shape file and the passengers count per station based on the crsCode, as every station has a unique crsCode, and this was one of the common columns in both.
- #2 Categorising the number of tickets purchased, the maximum number of tickets purchased at a station was 85 million.

  - #2a First, the dataset was divided into categories, each with a range of 5 million records. The first and second categories had 2491 and 47 records present in each, which had to be divided further as shown in Figure 1.
  - #2b The first and second categories formed from the main dataset were further divided into categories, each with a range of 0.5 million.
  - #2c Removed the category one and category two records from the main dataset, which now holds information on the stations that have issued more than 10 million tickets.
  - #2d After categorising the records into category 1 in the previous step, the new category one formed contained about 1795 records, which had to be broken down further.
  - #2e The new set of categories made have a range of 50 thousand.
- #3 Combined all the data by updating the category values accordingly.
- #4 Created a new dataset to hold the station details on which category of 25 or more.

```
knitr::include_graphics("IMG_0082.jpg")
```

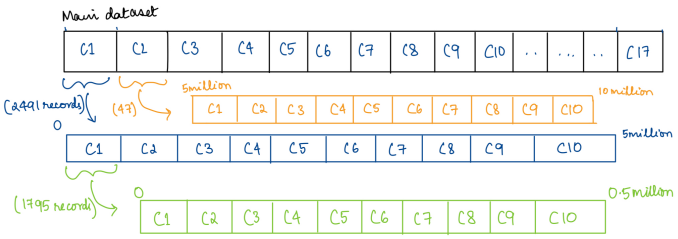

Figure 1: How the flow of categorisation works

```
#1
final_count<-merge(x = passengers_count, y = stations, by.x = "crsCode", by.y = "crsCode")
final_count$`Total` <- as.numeric(final_count$`Total`)
```

```
## Warning: NAs introduced by coercion
```

```
options(scipen = 100, digits = 4)
#2
a<-seq(from=0, to=85000000, by=5000000)
final_count$Categories <- findInterval(final_count$Total, a)
#2a
Category_1<-final_count[final_count$Categories==1,]
Category_2<-final_count[final_count$Categories==2,]
a1<-seq(from=0, to=5000000, by=500000)
Category_1$Categories <- findInterval(Category_1$Total, a1)
a2<-seq(from=5000000, to=10000000, by=500000)
Category_2$Categories <- findInterval(Category_2$Total, a2)
#2b
Category_others<-final_count[final_count$Categories!=1,]
Category_others<-Category_others[Category_others$Categories!=2,]
#2c
Low1<-Category_1[Category_1$Categories==1,]
Category_1<-Category_1[Category_1$Categories!=1,]
a11<-seq(from=0, to=500000, by=50000)
Low1$Categories<- findInterval(Low1$Total, a11)
#3
Category_1$Categories<-Category_1$Categories+8
Category_2$Categories<-Category_2$Categories+18
Category_others$Categories<-Category_others$Categories+26
final_cat_tic<-bind_rows(Low1, Category_1, Category_2, Category_others)
final_cat_tic<-na.omit(final_cat_tic)
```

```
#4
new_names<-final_cat_tic[final_cat_tic$Categories>=25,]
```

3. UK_popden - Multiple transformations were done to this dataset in order to get a clear visualisation of the data. The transformations done are:

- #1 Had to remove a single row that contained the entire country's information. (Overall Count)
- #2 Again, like the previous code and categorising the number of tickets sold per station, the same was done based on population density to categorise the cities in local authorities. There were different categories, each having a range of 0.5 million.
- #3 We noticed that the number of cities with a population between 0 and 0.5 million is very high (372), so we further categorised them (again, similar to the previous code set) and then removed the category 1 data from the original data.
- #4 In the further categorising part, the range used is 50 thousand.
- #5 Merging the two data sets with the UK_lad shape file and then concatenating them together.

```
#1
UK_popden<-UK_popden[UK_popden$Geography != 'Country',]
#2
b<-seq(from=0, to=7000000, by=500000)
UK_popden$Categories <- findInterval(UK_popden$`All ages`, b)
#3
Category_1_pop<-UK_popden[UK_popden$Categories==1,]
UK_popden<-UK_popden[UK_popden$Categories !=1,]
#4
b1<-seq(from=0, to=500000, by=50000)
Category_1_pop$Categories <- findInterval(Category_1_pop$`All ages`, b1)
#5
UK_popden$Categories<-UK_popden$Categories+9
final_data<-merge(x = UK_lad, y = UK_popden, by.x = "LAD22CD", by.y = "Code")
final_cat_1<-merge(x=UK_lad, y=Category_1_pop, by.x = "LAD22CD", by.y = "Code")
bothdfs <- bind_rows(final_data,final_cat_1)
bothdfs$Categories<-as.factor(bothdfs$Categories)
```

4. UK_roads_data - There are multiple transformations done to this dataset, such as:

- #1 We reduced the data to only 2022, with the road type being major, as this data set contains information from 2010 for all major and minor roads in the UK.
- #2 Based on Count point(they are unique values for each count point), Year, date, hour, we add the S & N values and W & E values for the All cars and the All HGVs, to get the total on that route in that hour.
- #3 Based on Count point, Year, and date, we add all the values obtained in Step #2 to make a new column that shows the number of cars and HGVs on that date in that route.
- #4 Based on the month, we add and find the average for the number of vehicles on that route.
- #5 We left join the data found in step #4 with the UK major roads data to plot the data more easily.

```
#1
UK_roads_data<-UK_roads_data[UK_roads_data$Year==2022,]
UK_roads_data<-UK_roads_data[UK_roads_data$`Road_type`=='Major',]
#2
new_df <- UK_roads_data %>%
  group_by(Year, Count_point_id,Latitude, Longitude, Road_name, Count_date, hour) %>%
  summarise(
    S_sum = sum(ifelse(Direction_of_travel == "S", All_motor_vehicles, 0), na.rm = TRUE),
    N_sum = sum(ifelse(Direction_of_travel == "N", All_motor_vehicles, 0), na.rm = TRUE),
    W_sum = sum(ifelse(Direction_of_travel == "W", All_motor_vehicles, 0), na.rm = TRUE),
    E_sum = sum(ifelse(Direction_of_travel == "E", All_motor_vehicles, 0), na.rm = TRUE),
```

```
    .groups = 'drop')
new_df <- new_df %>% mutate(Total_NS = S_sum + N_sum)
new_df <- new_df %>% mutate(Total_WE = W_sum + E_sum)
new_df$Overall<-new_df$Total_NS+new_df$Total_WE
#3
Step2_df <- new_df %>%
  group_by(Count_point_id, Count_date,Latitude, Longitude, Road_name) %>%
  summarise(Average_Overall = mean(Overall, na.rm = TRUE), .groups = 'drop')
Step2_df$Month<-substr(Step2_df$Count_date, 6,7)
Step2_df <- Step2_df %>%
  group_by(Count_point_id, Month,Latitude, Longitude, Road_name) %>%
  summarise(Average_Overall_mon = mean(Average_Overall, na.rm = TRUE), .groups = 'drop')
#4
Step3 <- Step2_df %>%
  group_by(Count_point_id, Road_name, Latitude, Longitude) %>%
  summarise(Average_Overall = mean(Average_Overall_mon, na.rm = TRUE), .groups = 'drop')
#5
joined_data <- UK_major_roads %>%
  left_join(Step3, by = c("RoadNumber" = "Road_name"))
```

```
## Warning in sf_column %in% names(g): Detected an unexpected many-to-many relationship between ‘x‘ and ‘y‘.
## i Row 1 of ‘x‘ matches multiple rows in ‘y‘.
## i Row 769 of ‘y‘ matches multiple rows in ‘x‘.
## i If a many-to-many relationship is expected, set ‘relationship =
##   "many-to-many"‘ to silence this warning.
```

```
joined_data<-na.omit(joined_data)
```

5. poll_air_gg - This data is made into a pivot table for easier visualisation of time series data.

```
df_long <- poll_air_gg %>% pivot_longer(cols = -Pollutant, names_to = "Year", values_to = "Value")
```

6. based_trans - This data is made into a pivot table for easier visualisation of time series data.

```
based_trans <- based_trans %>% pivot_longer(cols = -`Transport type and mode`, names_to = "Year",
                                            values_to = "Value")
```

7. UK_poll_LAD - A few transformations were made to this dataset to understand the CO2 emissions by each local authority district in the UK. The transformations done are:

- #1 Reduced the data only to 2021 as it held values for 2021 to 2022.
- #2 Based on the Local Authority code, we summed up all the CO2 emissions due to various causes and made a new data frame.
- #3 Finally, I merged the UK_Lad shapefile with this new dataframe to make plotting the dataset much easier.

```
#1
UK_poll_LAD<-UK_poll_LAD[UK_poll_LAD$`Calendar Year`==2021,]
#2
UK_poll_sum <- UK_poll_LAD %>%
  group_by(`Local Authority Code`) %>%
  summarise(TotalValues=sum(`CO2 emissions within the scope of influence of LAs (kt CO2e)`,
                            na.rm=TRUE))
#3
final_data_poll<-merge(x = UK_lad, y = UK_poll_sum, by.x = "LAD22CD", by.y = "Local Authority Code")
```

Exploratory Data Analysis and Discussions

1. Understanding the number of Train journeys over the years in the UK

Understanding trends in this data would help one find how public transport, such as trains, has been affected over the years. From an environmental perspective, it gives us an idea of how the number of cars on the road is decreasing due to more people using the trains, reducing the amount of carbon dioxide in the air.

```
plot1<-ggplot(pass_25, aes(x=Year, y=`Total Journeys`)) +
  geom_bar(stat = "identity",width=0.4)+
  geom_line(stat = "identity", group = 1, colour = "red", linetype='dashed')+
  geom_point(mapping = aes(x =Year, y =`Total Journeys` ),color='red',cex=1)+
  ggtitle("Total Number of Train Journeys in the UK")+
  labs(y = "Total Journey in millions", x = "Year")+
  theme(axis.text.x = element_text(size = 6),plot.title = element_text(size = 8),
        axis.title.x = element_text(size = 7), axis.title.y = element_text(size = 7))
plot1
```

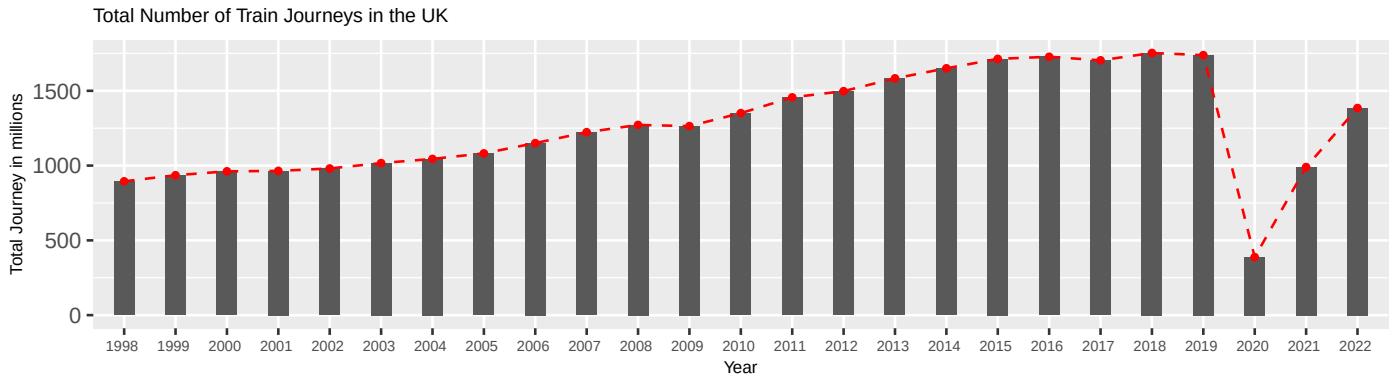

Figure 2: Overall Usage of Trains over the years

Figure 2 shows the overall number of passengers who have taken a journey by train in the last 25 years. Ridership has steadily increased over the years until 2020, when COVID-19 and lock down were announced. After that, ridership has steeply increased again, almost reaching the same level as it was before COVID by 2024.

2. Understanding the population density and the Tickets sold in that place

Locations with high population density often have a high demand for public transport. We want to see whether the population density at a place affects the number of tickets sold at the railway stations in it.

```
size_rule<-theme(legend.key.size =unit(0.5, "cm"),legend.text =element_text(size =5),
                 legend.title =element_text(size = 5), plot.title = element_text(size = 8))
plot2<-ggplot()+
  geom_sf(data = UK,aes(color = "UK"), lwd = 0.4)+
  geom_sf(data = UK_railroutes,aes(color = "Railroads"), lwd = 0.25)+
  geom_point(data = stations,aes(x = long, y = lat, color ='Stations'), cex = 0.5, alpha = 0.4)+
  scale_color_manual(values =c("UK" = "grey", "Stations" = "blue", "Railroads" = "black"),
                     name = "Legend",breaks =c("UK","Stations","Railroads"),
                     labels =c("UK","Stations","Railroads"))+
  new_scale_color()+
  geom_point(data = final_cat_tic,aes(x=long, y=lat,color = Categories, size=Categories),alpha=0.3)+
  scale_color_gradientn(colors=c("lightblue","green","orange", "red") )+
  geom_text(data=new_names,aes(x=long, y =lat, label =`Station name`), size = 1,
            check_overlap = TRUE)+coord_sf()+theme_void()+labs(color="Tickets Purschased Category")+
  size_rule
```

```
color_vector <-colorRampPalette(c("green4", "white", "violetred2"))(12)

plot3<-ggplot()+ geom_sf(data=UK, color="grey",lwd=0.4)+
  geom_sf(data=bothdfs,aes(fill =`Categories`))+scale_fill_manual(values = color_vector)+
  coord_sf()+theme_void()+size_rule+
  labs(fill="Categories for population", title='Population Density in the UK')+
  theme(plot.title = element_text(size = 8))

plot2<-plot2+labs(title ='UK map showing the stations with the most tickets bought')+
  guides(size='none')

grid.arrange(plot3, plot2, ncol=2)
```

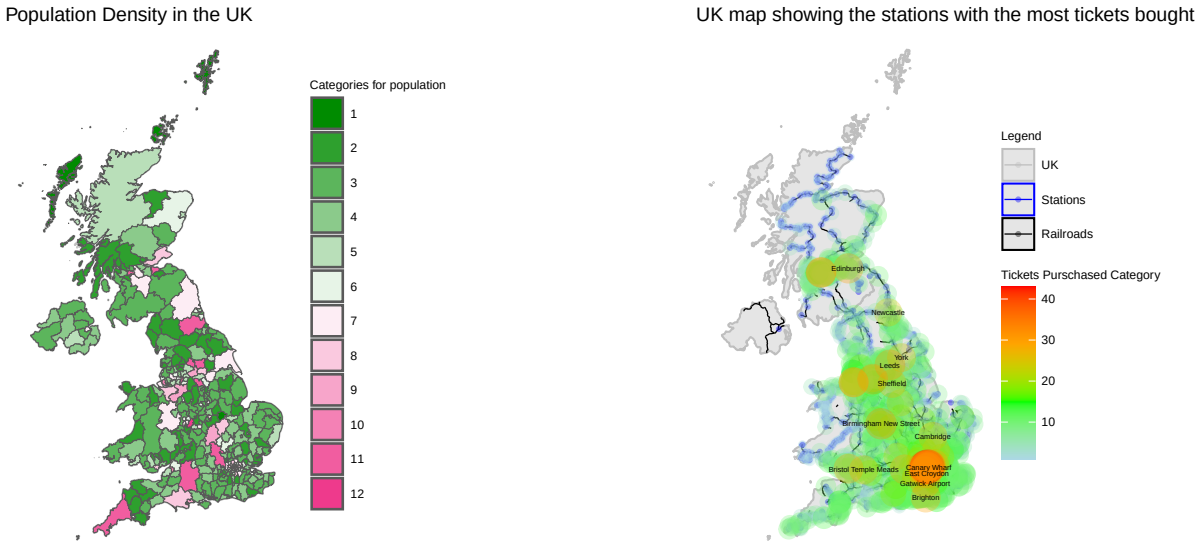

Figure 3: Population Density and Tickets bought at stations in the UK

Figure 3 shows the population density and the main stations where tickets were bought in the UK. These two plots give us an idea of whether the population density affects the number of tickets bought in that area. From both these plots, if we look at the London region or the mid-UK region, we see a lot of pink or light pink local authority districts in the population density plot, indicating that they have a high population density compared to the rest of the LADs. Most tickets have also been purchased in this location based on the plot on the left. They are indicating that the population and tickets bought are directly proportional to each other.(Balcombe et al. 2004)

3.Network Analysis on the Train ticket data to understand the busiest hubs in the UK

This helps us find the busiest hubs in the UK, where trains are used the most compared to the other locations. This analysis would also help improve the accuracy of timetables in these locations, leading to improved performance and the ability to maximize rail capacity.

```
plot4<-plot2+coord_sf(xlim =c(-3.75, -1), ylim = c(54, 52))+
      labs(title = "Showing the main hubs in Mid UK region")+guides(size='none')
```

```
## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.
```

```
plot5<-ggplot()+
  geom_sf(data=UK, aes(color="UK"),lwd=0.4) +
  geom_sf(data = London_roads,aes(color = "London_roads"),lwd=0.5, fill="gray70") +
```

```
  geom_sf(data = London_tube,aes(color = "London_tube"), lwd=0.1)+
  geom_sf(data = UK_railroutes,aes(color = "Railroads"), lwd=0.25) +
  geom_point(data = stations,aes(x=long, y=lat,color='Stations'),cex=0.5, alpha=0.4)+
  scale_color_manual(values = c("UK" = "grey", "Stations" = "blue", "Railroads" = "black",
      "London_roads"="honeydew4","London_tube"="gold"),name = "Legend",
      breaks = c("UK", "Stations","Railroads","London_roads","London_tube"),
      labels = c("UK", "Stations", "Railroads","London_roads","London_tube")) + new_scale_color()+
  geom_point(data = final_cat_tic, aes(x=long, y=lat,color = Categories, size=Categories),alpha=0.3)+
  scale_color_gradientn(colors=c("lightblue","green","orange", "red") )+
  geom_text(data=new_names, aes(x=long, y =lat, label =`Station name`), size=1, check_overlap=TRUE)+
  coord_sf(xlim=c(-0.2,0.0),ylim =c(51.45,51.57))+
  theme_void()+
  labs(title="Showing the main hubs in London",color="Tickets Purschased")+size_rule+
  guides(size='none')

grid.arrange(plot4,plot5,ncol=2)
```

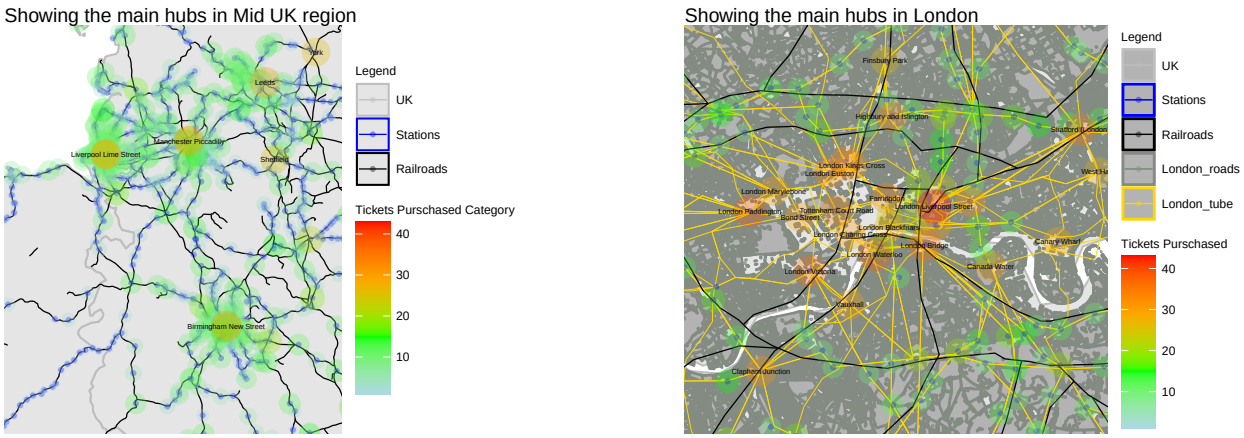

Figure 4: Finding the busiest Hubs in UK

Figure 4 zooms into the mid-region and London region as they contain the most red and orange circles, which indicate the location of the stations where the majority of the tickets were bought. This plot helps us determine the busiest hubs, such as Liverpool Street station in London, London Waterloo and others. It is indicating that the London Region has the majority of the stations with the highest count of tickets sold, which also tells us that it is one of the busiest hubs in the UK.

4. Understanding the different air pollutants in the air and the roadways across the UK

This analysis will help us investigate if there is a correlation between the density of the roadways, indicated by the number of vehicles that travel on them, and the air pollution levels. Also, by looking into the pollutants in the air and their locations, one can ascertain the hotspots for air pollution in the UK.

```
plot6<-ggplot() +
  geom_sf(data=UK, aes(color="UK"),lwd=0.4) +
  geom_sf(data = UK_lad,aes(color = "UK_lad"),lwd=0.1)+
  geom_point(data=UK_Air_2022, aes(x = longitude, y = latitude,size=o3,color='o3'),alpha=0.4)+
  geom_point(data=UK_Air_2022, aes(x = longitude, y = latitude,size=pm10,color='pm10'),alpha=0.4)+
  geom_point(data=UK_Air_2022, aes(x = longitude, y = latitude,size=no2,color='no2'),alpha=0.4)+
  geom_point(data=UK_Air_2022, aes(x = longitude, y = latitude,size=so2,color='so2'),alpha=0.4)+
  scale_color_manual(values=c("UK"="grey","UK_lad"="black","o3" = "wheat3", "pm10" = "red",
      "no2" = "yellow","so2"="orange"),name="Legend",breaks=c("UK","UK_lad","o3", "pm10","no2","so2")
      ,labels=c("UK","UK_lad","Ozone o3", "Particulate matter pm10","Nitrous oxide no2",
                "Sulphur dioxide so2")) + coord_sf(xlim = c(-8.2,2))+theme_void()+guides(size='none')+
```

```
  labs(title='Air pollulants measured in 2022')+size_rule

plot7<-ggplot() +
  geom_sf(data=UK, color="grey",lwd=0.4) +
  geom_sf(data = joined_data, aes(color = Average_Overall),lwd=0.3) +
  scale_color_gradientn(colors=c("lightgreen", "gold", "orange", "red") )+
  labs(color="Average Number of cars yearly",
       title='Average number of vehicles traveling on that route')+
  coord_sf()+theme_void()+size_rule

grid.arrange(plot6,plot7,ncol=2)
```

```
## Warning: Removed 96 rows containing missing values (‘geom_point()‘).

## Warning: Removed 60 rows containing missing values (‘geom_point()‘).

## Warning: Removed 13 rows containing missing values (‘geom_point()‘).

## Warning: Removed 142 rows containing missing values (‘geom_point()‘).
```

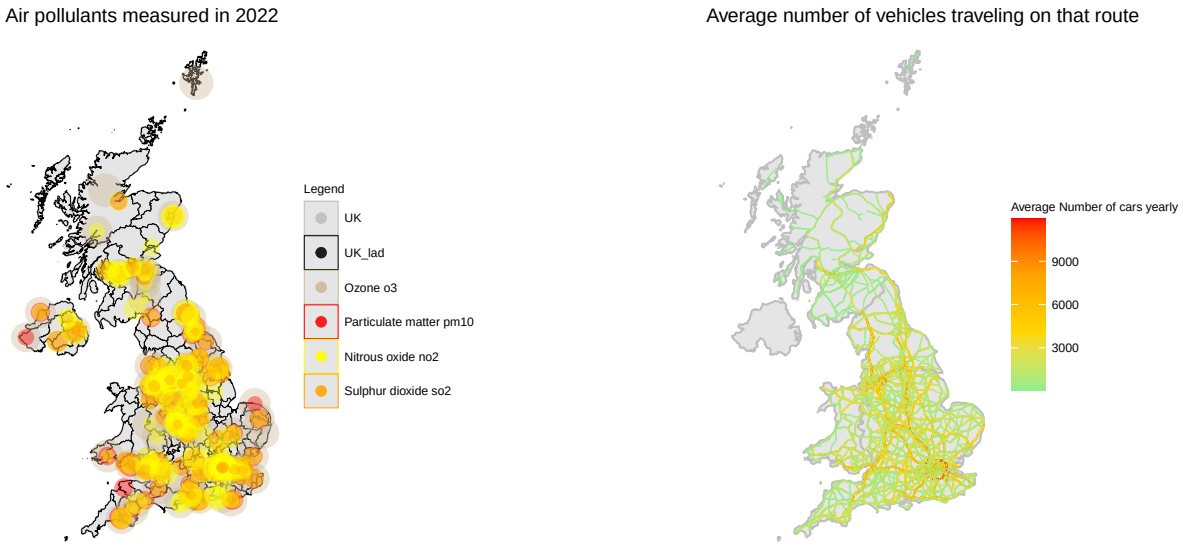

Figure 5: Pollutants in measured in Air and Average Traffic on road

In Figure 5, the plot on the left shows different greenhouse gases measured in the air in 2022, including their locations. Ozone indicated in wheat colour absorbs radiation from the sun and acts like a strong greenhouse gas, which contributes to altering evaporation, cloud formation, and atmospheric circulation. Looking into pm or particulate matter, indicated by red- a combination of many chemicals and high exposure to these particles can lead to respiratory diseases. Nitrous oxide, one of the three main greenhouse gases seen abundantly on the map, along with Sulphur dioxide, can change soil chemistry and affect sensitive habitats(Sofianopoulou et al. 2019). The plot on the right indicates the average number of vehicles that go on that major road yearly. A detailed analysis of how much greenhouse gases are produced by the transportation system is done in the next part, as it would give us a clearer understanding.

5. Understanding the amount of greenhouse gases produced and how different transports have contributed

Understanding the different amounts of greenhouse gases present in the atmosphere is important as it reveals how different rules and regulations governing the reduction of greenhouse gases are playing out. It is also important to know how emissions caused by different transportation means have affected greenhouse gas emissions (CO2) over the years.(McKinnon 2007)

```
plot8<-ggplot(df_long, aes(x = as.numeric(Year), y = Value, color = Pollutant)) + geom_line() +
  labs(x = "Year", y = "Value", title ="Greenhouse gases over the years in the UK" ) +
  theme_minimal() + size_rule
plot9<-ggplot(based_trans, aes(x=as.numeric(Year), y = Value,color =`Transport type and mode`))+
  geom_line()+labs(x="Year",y="Value",title="CO2 emission from different transportation systems")+
  theme_minimal()+theme(legend.text=element_text(size=4),legend.title=element_text(size=4),
            axis.title.x=element_text(size=7),axis.title.y=element_text(size=7))+ size_rule
grid.arrange(plot8,plot9,ncol=2)
```

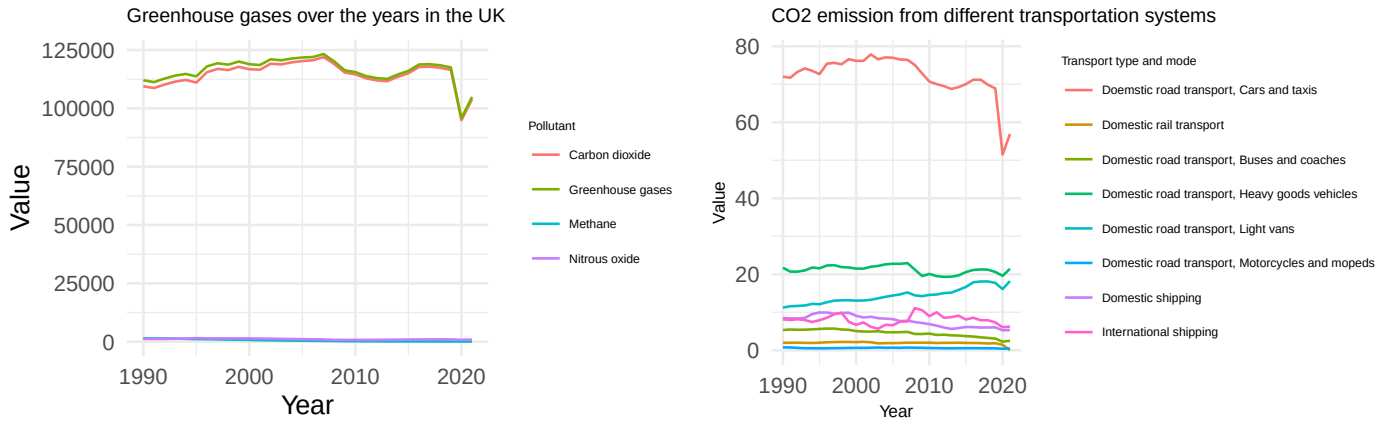

Figure 6: About greenhouse gas emissions

In Figure 6, the plot on the left shows that the amount of carbon dioxide emissions over the years has increased compared to other greenhouse gases. The main reason for the increase in carbon dioxide emissions over the years is the burning of fossil fuels. Fossil fuels, on an elemental level, are made up of carbon, and when on combustion, they mix with oxygen and form carbon dioxide; over the years, almost a lot of transport modes have used fossil fuels for fuel. If you look at the plot on the right, it indicates that road transport has the highest amount of carbon dioxide emissions compared to the rest. There has been a decrease since 2010 due to the implementation of different environmentally friendly fuels and electric vehicles to reduce carbon dioxide emissions.

6. Greenhouse Gas Emission based on LAD comparison with the population density and the road traffic

This analysis would help identify areas with high greenhouse emissions. These could be areas with high population density and heavy road traffic. Once these areas are identified, targeted interventions can be implemented to reduce emissions.

```
color_vector <-colorRampPalette(c("gold", "white", "red"))(10)
plot10<-ggplot()+ geom_sf(data=UK, color="grey",lwd=0.4)+theme_void()+
  geom_sf(data=final_data_poll,aes(fill =`TotalValues`))+scale_fill_gradientn(colors  = color_vector)+
  coord_sf()+labs(fill='Total Emissions',title='CO2 Emissions based on LADs')+size_rule
plot11<-plot3 + geom_sf(data = joined_data, aes(color = Average_Overall),lwd=0.3) +
  scale_color_gradientn(colors=c("lightgreen", "gold", "orange", "red") )+
  labs(color="Average Number of cars yearly",title='Population density and the traffic on that route')+
  coord_sf()+theme_void()+size_rule+theme(legend.key.size =unit(0.25, "cm"),legend.text=element_text(size=3),
```

```
## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.
```

```
grid.arrange(plot10, plot11, ncol=2)
```

From Figure 7 the left plot shows the CO2 emissions based on Local Authority districts, in this the areas with white to red shades indicate that they have high emissions. The plot on the right shows the population density and the traffic on these routes. Comparing the two maps, it appears that areas with higher population density and more traffic tend to have higher CO2 emissions. Thereby suggesting a correlation between population density, traffic and CO2 emissions.

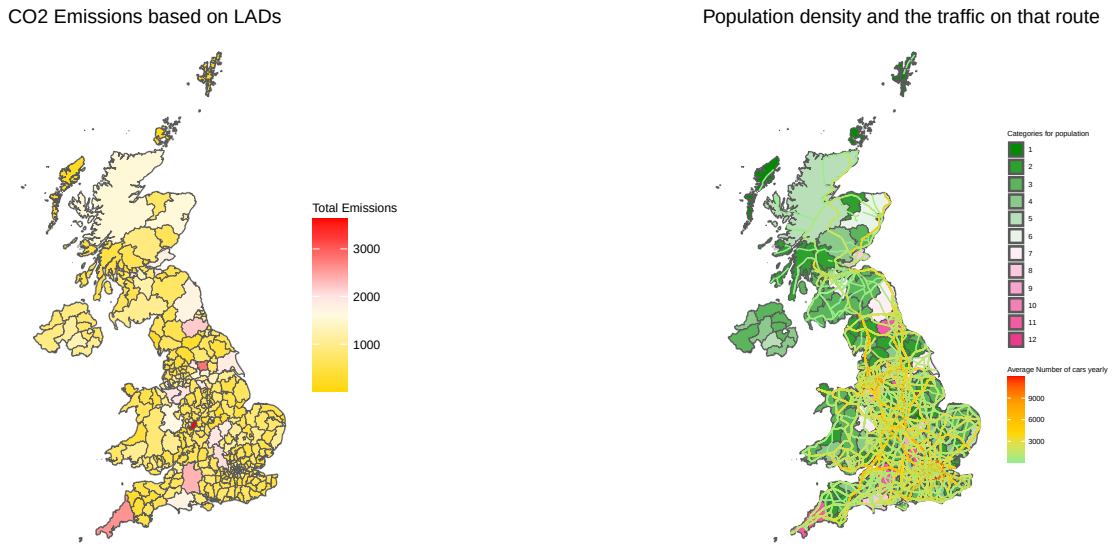

Figure 7: CO2 Emissions from LADs in 2021-2022 and Average traffic

Conclusion and Future Work

This study has provided a significant correlation between population density and air quality, suggesting that areas with higher population density and traffic tend to have poorer air quality.The first part of the study, which analysed train usage, showed that train routes with the busiest hubs often align with areas of high population density, suggesting that trains are a crucial mode of transport in these areas. The second part of the study examined vehicle movement data. It was found that areas with heavy vehicle traffic also tend to have higher levels of air pollution due to road transport being one of the biggest contributory factors to greenhouse gas emissions, which was confirmed in the last part. It overall highlights the need for sustainable transport solutions and effective air pollution control measures, particularly in areas with high population density and heavy traffic.

Future work could include a more detailed analysis of other modes of transport, such as buses, bicycles which could provide a more comprehensive picture of transport networks and their impact on air quality.

References

Acharya, J. 2018. "Download United Kingdom Administrative Boundary Shapefiles - Countries." 2018. https://www.igismap.com/download-united-kingdom-administrative-boundary-shapefiles-countries-regions-counties-unitary-authorities-wards/.

Balcombe, Richard, Roger Mackett, Neil Paulley, John Preston, Jeremy Shires, Helena Titheridge, Mark Wardman, and Peter White. 2004. "The Demand for Public Transport: A Practical Guide."

Carslaw, D. C., and K. Ropkins. 2012. "Openair — an r Package for Air Quality Data Analysis." *Environmental Modelling & Software.* 27-28: 52–61.

Core-Periphery, Detecting. 2019. "Structure in Spatial Networks. Junteng Jia and Austin r." In *Benson. Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM), 2019.*

Datastore, London. n.d. "Statistical GIS Boundary Files for London – London Datastore." n.d. https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london.

Dft.gov.uk. 2022. "Road Traffic Statistics - Download Data." 2022. https://roadtraffic.dft.gov.uk/downloads.

Gov., Uk. 2022. "Transport and Environment Statistics 2022. GOV.UK." 2022. https://www.gov.uk/government/statistics/transport-and-environment-statistics-2022/transport-and-environment-statistics-2022.

McKinnon, Alan. 2007. "CO2 Emissions from Freight Transport: An Analysis of UK Data." In *Logistics Research Network-2007 Conference Global Supply Chains: Developing Skills, Capabilities and Networks.* Citeseer.

National Statistics, Office for. 2022. "Estimates of the Population for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics. Ons.gov.uk." 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland.

Networks, Core-Periphery Structure in. 2014. "M. Puck Rombach, Mason a. Porter, James h." *Fowler, and Peter J. Mucha. SIAM Journal on Applied Mathematics.*

osdatahub.os.uk. n.d. "OS Data Hub." n.d. https://osdatahub.os.uk/.

Portalx, Open Geography. n.d. "Local Authority Districts (December 2022) Boundaries UK BUC. Geoportal.statistics.gov.uk." n.d. https://geoportal.statistics.gov.uk/datasets/ons::local-authority-districts-december-2022-boundaries-uk-buc-2/explore?location=55.044823.

Rail, Office of, and Road. 2023. "Passenger Rail Usage | ORR Data Portal. Orr.gov.uk." 2023. https://dataportal.orr.gov.uk/statistics/usage/passenger-rail-usage/.

Regt, Robin de, Christian von Ferber, Yurij Holovatch, and Mykola Lebovka. 2019. "Public Transportation in Great Britain Viewed as a Complex Network." *Transportmetrica A: Transport Science* 15 (2): 722–48.

Sofianopoulou, Eleni, Stephen Kaptoge, Stefan Gräf, Charaka Hadinnapola, Carmen M Treacy, Colin Church, Gerry Coghlan, et al. 2019. "Traffic Exposures, Air Pollution and Outcomes in Pulmonary Arterial Hypertension: A UK Cohort Study Analysis." *European Respiratory Journal* 53 (5).

Survey, O. 2021. "OS Open Roads. Www.data.gov.uk." 2021. https://www.data.gov.uk/dataset/65bf62c8-eae0-4475-9c16-a2e81afcbdb0/os-open-roads.

Wheatley, D. 2024. "Davwheat/Uk-Railway-Stations. GitHub." 2024. https://github.com/davwheat/uk-railway-stations?tab=readme-ov-file.