Analysing Business Establishments and Industry Distribution in Melbourne from 2002 to 2022.

Code

Alicia Zhao

Hide

```
        library(readxl)

        url <- "https://data.melbourne.vic.gov.au/api/explore/v2.1/catalo
g/datasets/business-establishments-per-block-by-anzsic/exports/xlsx?lang=e
n&timezone=Australia%2FSydney&use_labels=true"
        destfile <- "business-establishments-per-block-by-anzsic.xlsx"
        download.file(url, destfile, mode = "wb")

        anzsic <- read_excel(destfile)

        head(anzsic)
```

```
## # A tibble: 6 × 23
##   `Census year` `Block ID` `CLUE small area` `Accommodation and Food Se
rvices`
##   <chr>              <dbl> <chr>
<dbl>
## 1 2023                   5 Melbourne (CBD)
0
## 2 2023                  11 Melbourne (CBD)
23
## 3 2023                  16 Melbourne (CBD)
18
## 4 2023                  21 Melbourne (CBD)
18
## 5 2023                  25 Melbourne (CBD)
23
## 6 2023                  34 Melbourne (CBD)
32
## # ℹ 19 more variables: `Administrative and Support Services` <dbl>,
## #   `Agriculture, Forestry and Fishing` <dbl>,
## #   `Arts and Recreation Services` <dbl>, Construction <dbl>,
## #   `Education and Training` <dbl>,
## #   `Electricity, Gas, Water and Waste Services` <dbl>,
## #   `Financial and Insurance Services` <dbl>,
## #   `Health Care and Social Assistance` <dbl>, …
```

If the link doesn't work, that's another link to download xlsx file manually. https://data.melbourne.vic.gov.au/explore/dataset/business-establishments-per-block-by-anzsic/export/ (https://data.melbourne.vic.gov.au/explore/dataset/business-establishments-per-block-by-anzsic/export/)

First need to check whether there are missing values by column.

Hide

```
colSums(is.na(anzsic))
```

```
##                                      Census year
##                                                0
##                                         Block ID
##                                                0
##                                  CLUE small area
##                                                0
##                 Accommodation and Food Services
##                                                0
##             Administrative and Support Services
##                                                0
##               Agriculture, Forestry and Fishing
##                                                0
##                    Arts and Recreation Services
##                                                0
##                                     Construction
##                                                0
##                           Education and Training
##                                                0
##       Electricity, Gas, Water and Waste Services
##                                                0
##              Financial and Insurance Services
##                                                0
##               Health Care and Social Assistance
##                                                0
##         Information Media and Telecommunications
##                                                0
##                                    Manufacturing
##                                                0
##                                           Mining
##                                                0
##                                   Other Services
##                                                0
## Professional, Scientific and Technical Services
##                                                0
##                Public Administration and Safety
##                                                0
##         Rental, Hiring and Real Estate Services
##                                                0
##                                     Retail Trade
##                                                0
##               Transport, Postal and Warehousing
##                                                0
##                                  Wholesale Trade
##                                                0
##                    Total establishments in block
##                                                0
```

The result indicates that no missing value is in this dataset.

Hide

```
unique(anzsic$`CLUE small area`)
```

```
##  [1] "Melbourne (CBD)"                "Carlton"
##  [3] "Parkville"                      "North Melbourne"
##  [5] "West Melbourne (Residential)" "West Melbourne (Industrial)"
##  [7] "Kensington"                     "East Melbourne"
##  [9] "Melbourne (Remainder)"          "Southbank"
## [11] "Docklands"                      "Port Melbourne"
## [13] "South Yarra"                    "City of Melbourne (total)"
```

'City of Melbourne (total)' isn't part of the ANZSIC area, so we need to check it out.

Hide

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Hide

```
total_anz <- anzsic %>% filter(`CLUE small area` == "City of Melbourne (total)")
total_anz
```

```
## # A tibble: 22 × 23
##    `Census year` `Block ID` `CLUE small area`         Accommodation and
Food S…¹
##    <chr>              <dbl> <chr>
<dbl>
##  1 2021                   0 City of Melbourne (total)
2975
##  2 2016                   0 City of Melbourne (total)
3005
##  3 2015                   0 City of Melbourne (total)
2878
##  4 2014                   0 City of Melbourne (total)
2814
##  5 2010                   0 City of Melbourne (total)
2401
##  6 2009                   0 City of Melbourne (total)
2329
##  7 2007                   0 City of Melbourne (total)
2110
##  8 2004                   0 City of Melbourne (total)
1763
##  9 2003                   0 City of Melbourne (total)
1630
## 10 2022                   0 City of Melbourne (total)
2830
## # ℹ 12 more rows
## # ℹ abbreviated name: ¹`Accommodation and Food Services`
## # ℹ 19 more variables: `Administrative and Support Services` <dbl>,
## #   `Agriculture, Forestry and Fishing` <dbl>,
## #   `Arts and Recreation Services` <dbl>, Construction <dbl>,
## #   `Education and Training` <dbl>,
## #   `Electricity, Gas, Water and Waste Services` <dbl>, …
```

Those are summary rows from 2002 to 2022. Summary rows can be removed now, we can take it into account later.

Hide

```
anzsic <- anzsic %>%
  filter(`CLUE small area` != "City of Melbourne (total)")
```

Block ID is not very useful for our analysis, as it is divided into small city blocks by area, and the corresponding geographical location is confidential information. We can remove this column and calculate the total number of locations of establishment in different years and areas.

Hide

```
        anzsic_group_by <- anzsic %>%
          select(-`Block ID`) %>%
          group_by(`Census year`, `CLUE small area`) %>%
          summarise(across(where(is.numeric), sum), .groups = "drop") %>%
          rename(`Total establishments in area` = `Total establishments in
block`)

        head(anzsic_group_by)
```

```
## # A tibble: 6 × 22
##   `Census year` `CLUE small area`  Accommodation and Fo…¹ Administrativ
e and S…²
##   <chr>         <chr>                                <dbl>
<dbl>
## 1 2002          Carlton                                190
40
## 2 2002          Docklands                               23
6
## 3 2002          East Melbourne                          54
22
## 4 2002          Kensington                              12
1
## 5 2002          Melbourne (CBD)                        993
489
## 6 2002          Melbourne (Remain…                      24
24
## # ℹ abbreviated names: ¹`Accommodation and Food Services`,
## #   ²`Administrative and Support Services`
## # ℹ 18 more variables: `Agriculture, Forestry and Fishing` <dbl>,
## #   `Arts and Recreation Services` <dbl>, Construction <dbl>,
## #   `Education and Training` <dbl>,
## #   `Electricity, Gas, Water and Waste Services` <dbl>,
## #   `Financial and Insurance Services` <dbl>, …
```

Melbourne (CBD) area has a great number of establishment locations in all fields, therefore, distinguish between CBD and Non-CBD, and calculate the total number of establishment locations in different area types.

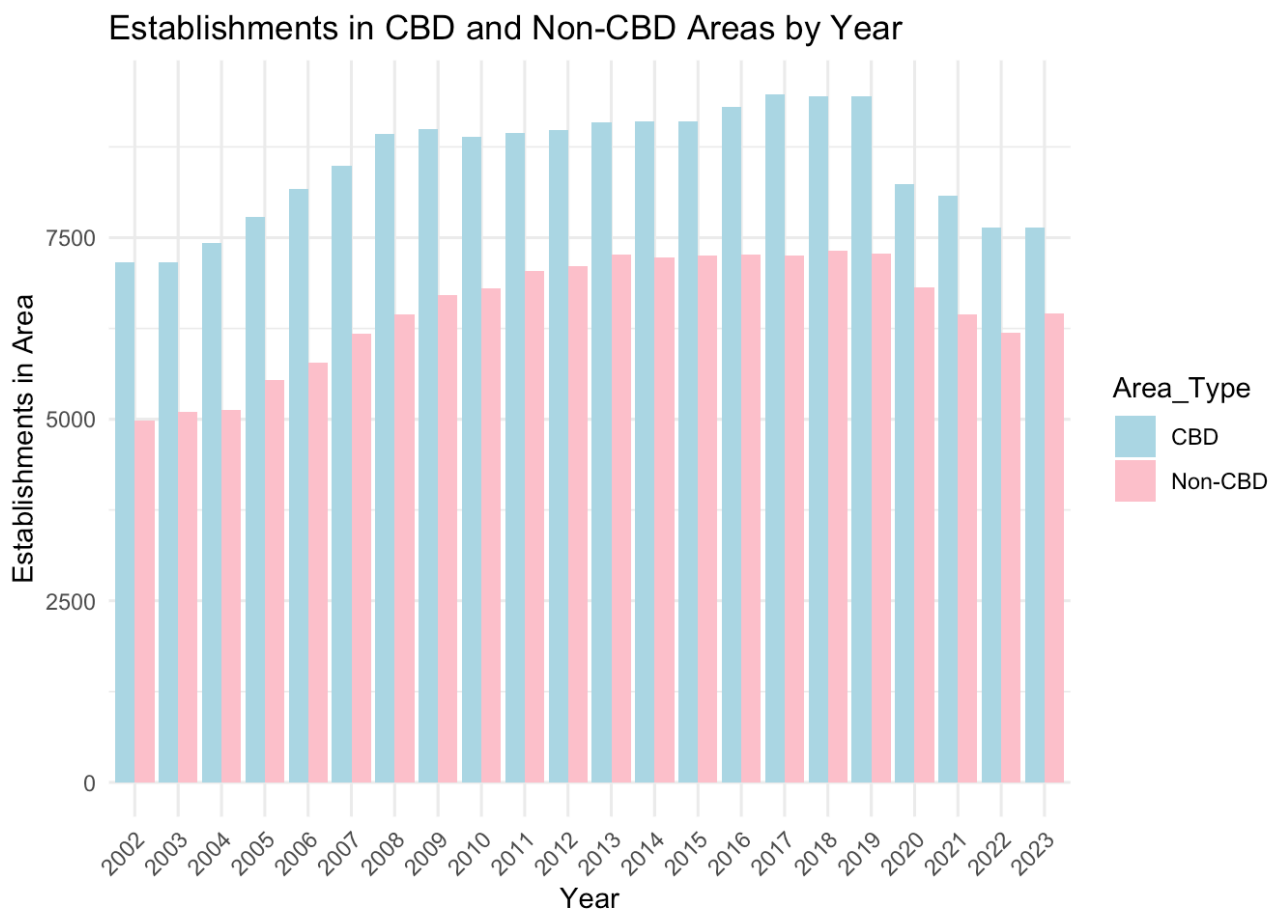
Hide

```
        anzsic_cbd <- anzsic_group_by %>%
          mutate(Area_Type = ifelse(`CLUE small area` == "Melbourne (CB
D)", "CBD", "Non-CBD"))

        anzsic_cbd <- anzsic_cbd %>%
          group_by(`Census year`, Area_Type) %>%
          summarise(`Total establishments in area` = sum(`Total establishm
ents in area`), .groups = "drop")
```

Make the plot

Hide

```
library(ggplot2)

ggplot(anzsic_cbd, aes(x = `Census year`, y = `Total establishments in area`, fill = Area_Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Establishments in CBD and Non-CBD Areas by Year",
       x = "Year",
       y = "Establishments in Area") +
  scale_fill_manual(values = c("CBD" = "lightblue", "Non-CBD" = "pink")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The number of establishment locations in the CBD is always greater than that in non-CBD areas, indicating that there are more business and employment opportunities in the CBD area. From the trend point of view, the number of establishments has been steadily increasing year by year. After 2019, due to Covid-19, the number of establishments began to decrease, especially in the CBD area.

Hide

```
colnames(anzsic_group_by)
```

```
##  [1] "Census year"
##  [2] "CLUE small area"
##  [3] "Accommodation and Food Services"
##  [4] "Administrative and Support Services"
##  [5] "Agriculture, Forestry and Fishing"
##  [6] "Arts and Recreation Services"
##  [7] "Construction"
##  [8] "Education and Training"
##  [9] "Electricity, Gas, Water and Waste Services"
## [10] "Financial and Insurance Services"
## [11] "Health Care and Social Assistance"
## [12] "Information Media and Telecommunications"
## [13] "Manufacturing"
## [14] "Mining"
## [15] "Other Services"
## [16] "Professional, Scientific and Technical Services"
## [17] "Public Administration and Safety"
## [18] "Rental, Hiring and Real Estate Services"
## [19] "Retail Trade"
## [20] "Transport, Postal and Warehousing"
## [21] "Wholesale Trade"
## [22] "Total establishments in area"
```

Now divide different industries into five main categories, recording their indices.

Industrial: Agriculture, Forestry and Fishing Construction Electricity, Gas, Water and Waste Services Manufacturing Transport, Postal and Warehousing
Wholesale Trade Mining

Entertainment: Arts and Recreation Services Accommodation and Food Services

Retail: Retail Trade

Institutional: Education and Training Health Care and Social Assistance

Commercial: Administrative and Support Services Financial and Insurance Services Information Media and Telecommunications
Other Services Rental, Hiring and Real Estate Services Public Administration and Safety Professional, Scientific and Technical Services

Hide

```
        # Industrial columns by index
        industrial_indices <- c(5, 7, 9, 13, 14, 20, 21)
        # Entertainment columns by index
        entertainment_indices <- c(3, 6)
        # Retail columns by index
        retail_indices <- 19
        # Institutional columns by index
        institutional_indices <- c(8, 11)
        # Commercial columns by index
        commercial_indices <- c(4, 10, 12, 15, 16, 17, 18)

        anzsic_category <- anzsic_group_by %>%
          # Group the data by row
          rowwise() %>%
          mutate(
            Industrial = sum(c_across(all_of(industrial_indices))),
            Entertainment = sum(c_across(all_of(entertainment_indices))),
            Retail = sum(c_across(all_of(retail_indices))),
            Institutional = sum(c_across(all_of(institutional_indices))),
            Commercial = sum(c_across(all_of(commercial_indices)))
          ) %>%
          # Select relevant columns to create the final table
          select(`Census year`, `CLUE small area`, Industrial, Entertainme
nt, Retail, Institutional, Commercial)

        head(anzsic_category)
```

```
## # A tibble: 6 × 7
## # Rowwise:
##   `Census year` `CLUE small area`  Industrial Entertainment Retail Inst
itutional
##   <chr>         <chr>                   <dbl>         <dbl>  <dbl>
<dbl>
## 1 2002          Carlton                    91           221    166
163
## 2 2002          Docklands                  56            32     11
4
## 3 2002          East Melbourne             33            96     16
165
## 4 2002          Kensington                 56            72      7
11
## 5 2002          Melbourne (CBD)           705          1085   1317
475
## 6 2002          Melbourne (Remain…         35            60     17
62
## # ℹ 1 more variable: Commercial <dbl>
```

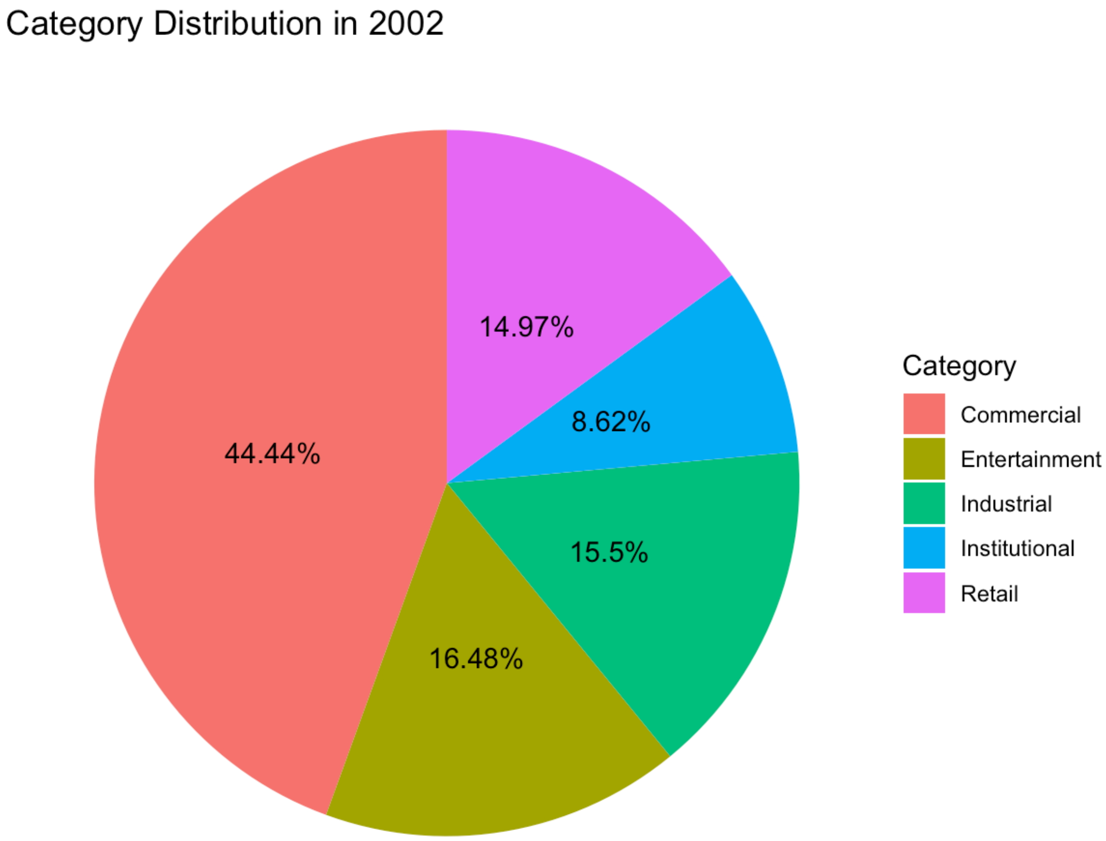
The result indicates the number of establishment locations by categories. Then choose year 2002, 2012 and 2022, draw 3 pie charts to compare the differences.

Hide

```
        anzsic_filtered <- anzsic_category %>%
          filter(`Census year` %in% c(2002, 2012, 2022)) %>%
          pivot_longer(cols = Industrial:Commercial, names_to = "Categor
y", values_to = "Total_count") %>%
          # Summarize data by Year and Category, and calculate percentages
          group_by(`Census year`, Category) %>%
          summarise(Total_count = sum(Total_count), .groups = "drop") %>%
          group_by(`Census year`) %>%
          mutate(Percentage = (Total_count / sum(Total_count)) * 100)
```

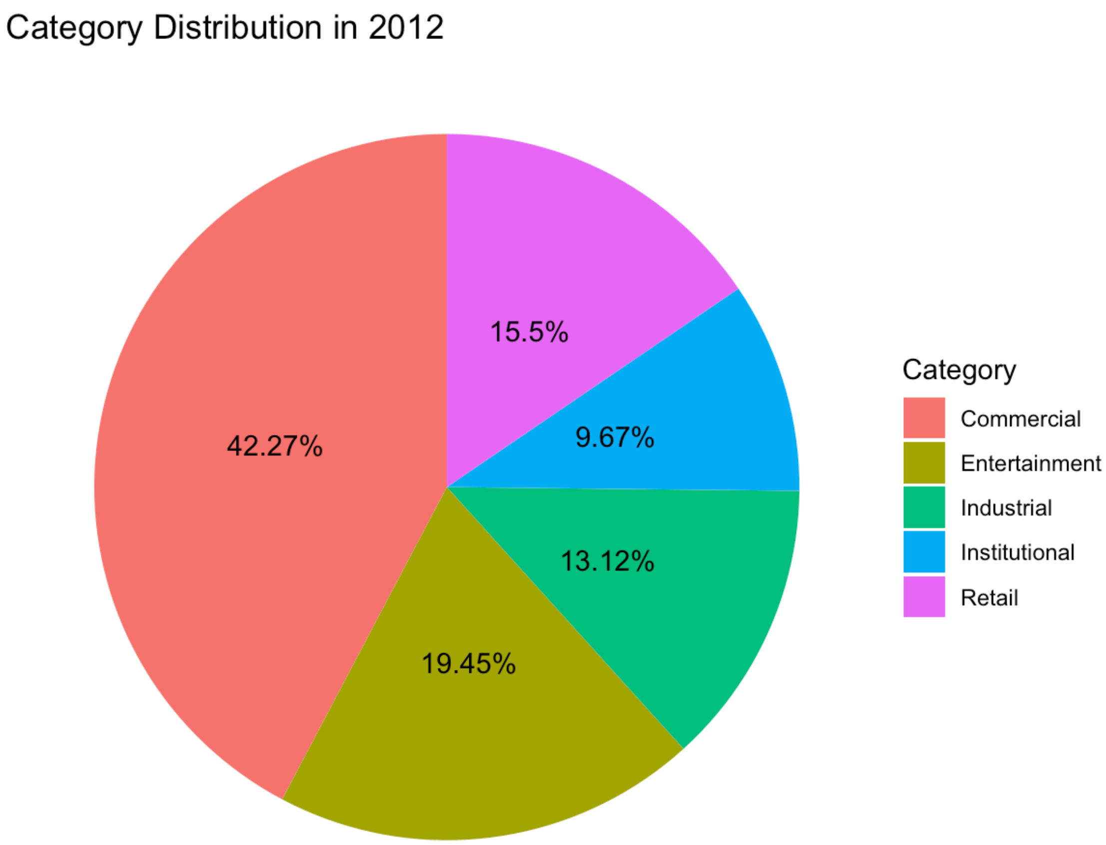
Hide

```
        # Create Pie Charts for 2002, 2012, and 2022
        # Pie chart for 2002
        anzsic_2002 <- ggplot(anzsic_filtered %>% filter(`Census year` ==
2002), aes(x = "", y = Percentage, fill = Category)) +
          geom_bar(stat = "identity", width = 1) +
          coord_polar("y", start = 0) +
          labs(title = "Category Distribution in 2002", x = NULL, y = NUL
L) +
          geom_text(aes(label = paste0(round(Percentage, 2), "%")),
                    position = position_stack(vjust = 0.5)) +
          theme_void() +
          theme(legend.position = "right")

        # Pie chart for 2012
        anzsic_2012 <- ggplot(anzsic_filtered %>% filter(`Census year` ==
2012), aes(x = "", y = Percentage, fill = Category)) +
          geom_bar(stat = "identity", width = 1) +
          coord_polar("y", start = 0) +
          labs(title = "Category Distribution in 2012", x = NULL, y = NUL
L) +
          geom_text(aes(label = paste0(round(Percentage, 2), "%")),
                    position = position_stack(vjust = 0.5)) +
          theme_void() +
          theme(legend.position = "right")

        # Pie chart for 2022
        anzsic_2022 <- ggplot(anzsic_filtered %>% filter(`Census year` ==
2022), aes(x = "", y = Percentage, fill = Category)) +
          geom_bar(stat = "identity", width = 1) +
          coord_polar("y", start = 0) +
          labs(title = "Category Distribution in 2022", x = NULL, y = NUL
L) +
          geom_text(aes(label = paste0(round(Percentage, 2), "%")),
                    position = position_stack(vjust = 0.5)) +
          theme_void() +
          theme(legend.position = "right")

        # Display the pie charts
        anzsic_2002
```



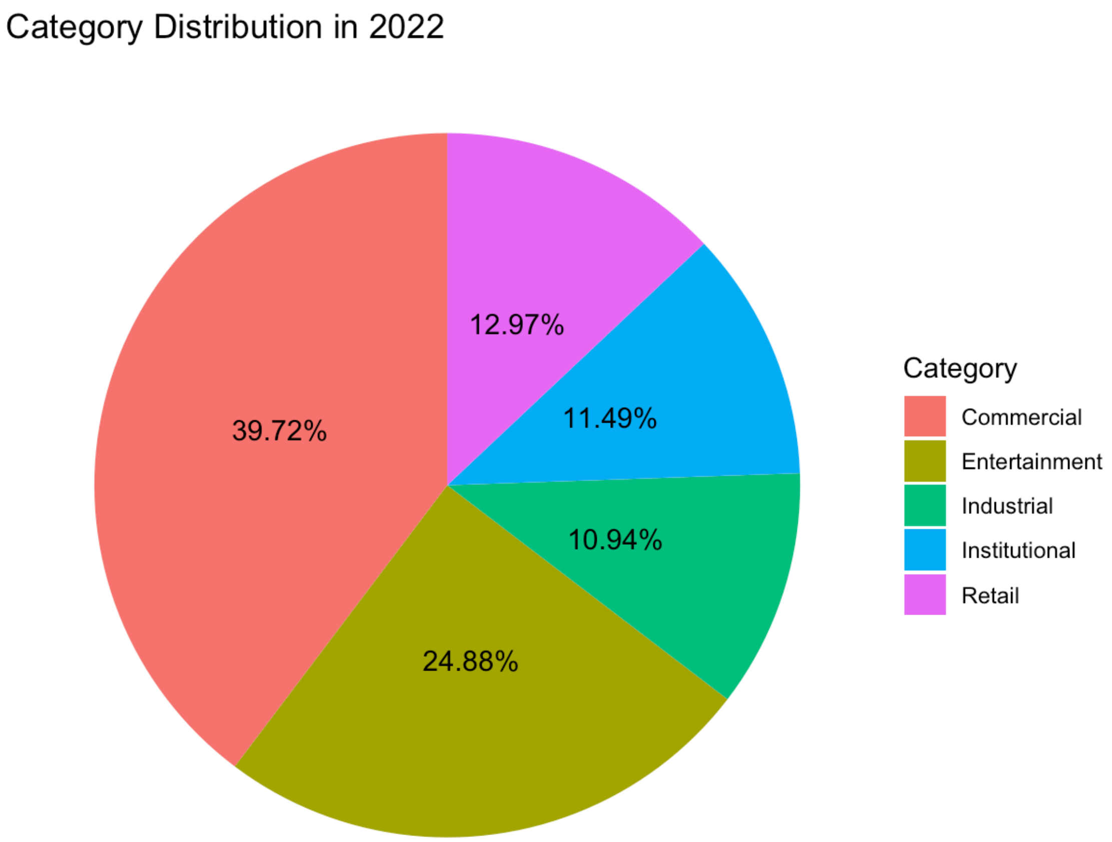
Hide

```
anzsic_2012
```



Hide

```
anzsic_2022
```

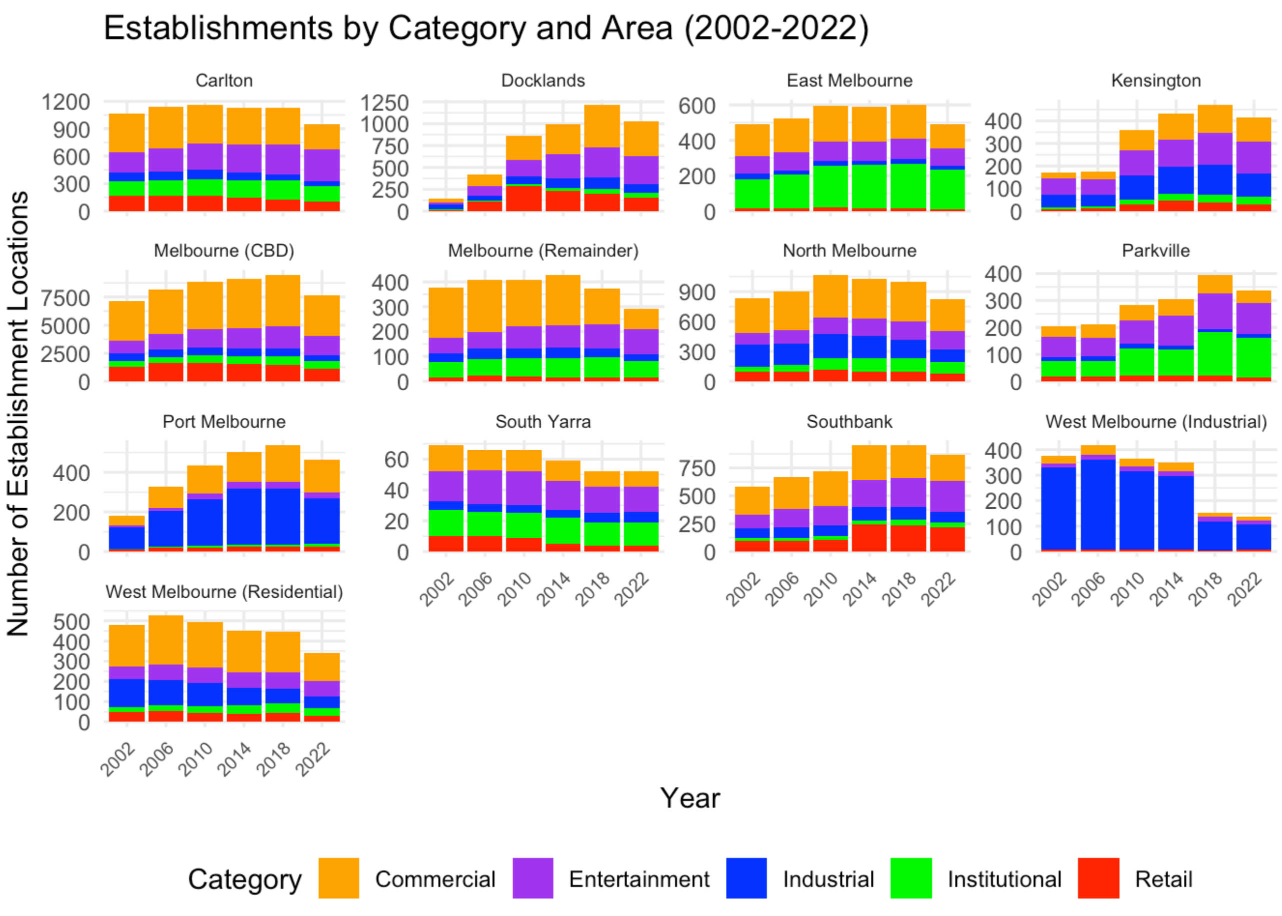


* The commercial category consistently remains the largest, but it shows a gradual decline in its share over time, dropping from 44.44% in 2002 to 39.71% in 2022. * The entertainment category shows a consistent increase over the three periods, growing from 16.48% in 2002 to 24.88% in 2022. * The industrial category is steadily declining, dropping from 15.5% in 2002 to 10.95% in 2022. * The institutional category has shown moderate growth over time, rising from 8.62% in 2002 to 11.49% in 2022. * The retail category shows a slight fluctuation, increasing slightly between 2002 and 2012, but then declining to 12.97% in 2022.

Now consider about areas' trends, choose year 2002, 2006, 2010, 2014, 2018 and 2022 to do the research.

Hide

```
        anzsic_area <- anzsic_category %>%
          filter(`Census year` %in% c(2002, 2006, 2010, 2014, 2018, 2022))
%>%
          pivot_longer(cols = Industrial:Commercial, names_to = "Categor
y", values_to = "Establishments")

        # Create a faceted bar plot to show categories over time for each
area
        ggplot(anzsic_area, aes(x = `Census year`, y = Establishments, fil
l = Category)) +
          geom_bar(stat = "identity") +
          facet_wrap(~`CLUE small area`, nrow = 4, scales = "free_y") +
          labs(title = "Establishments by Category and Area (2002-2022)",
               x = "Year",
               y = "Number of Establishment Locations",
               fill = "Category") +
          theme_minimal() +
          scale_fill_manual(values = c("orange", "purple", "blue", "gree
n", "red")) +
          theme(axis.text.x = element_text(angle = 45, hjust = 1, size =
7)) +
          theme(legend.position = "bottom") +
          theme(strip.text = element_text(size = 7))
```



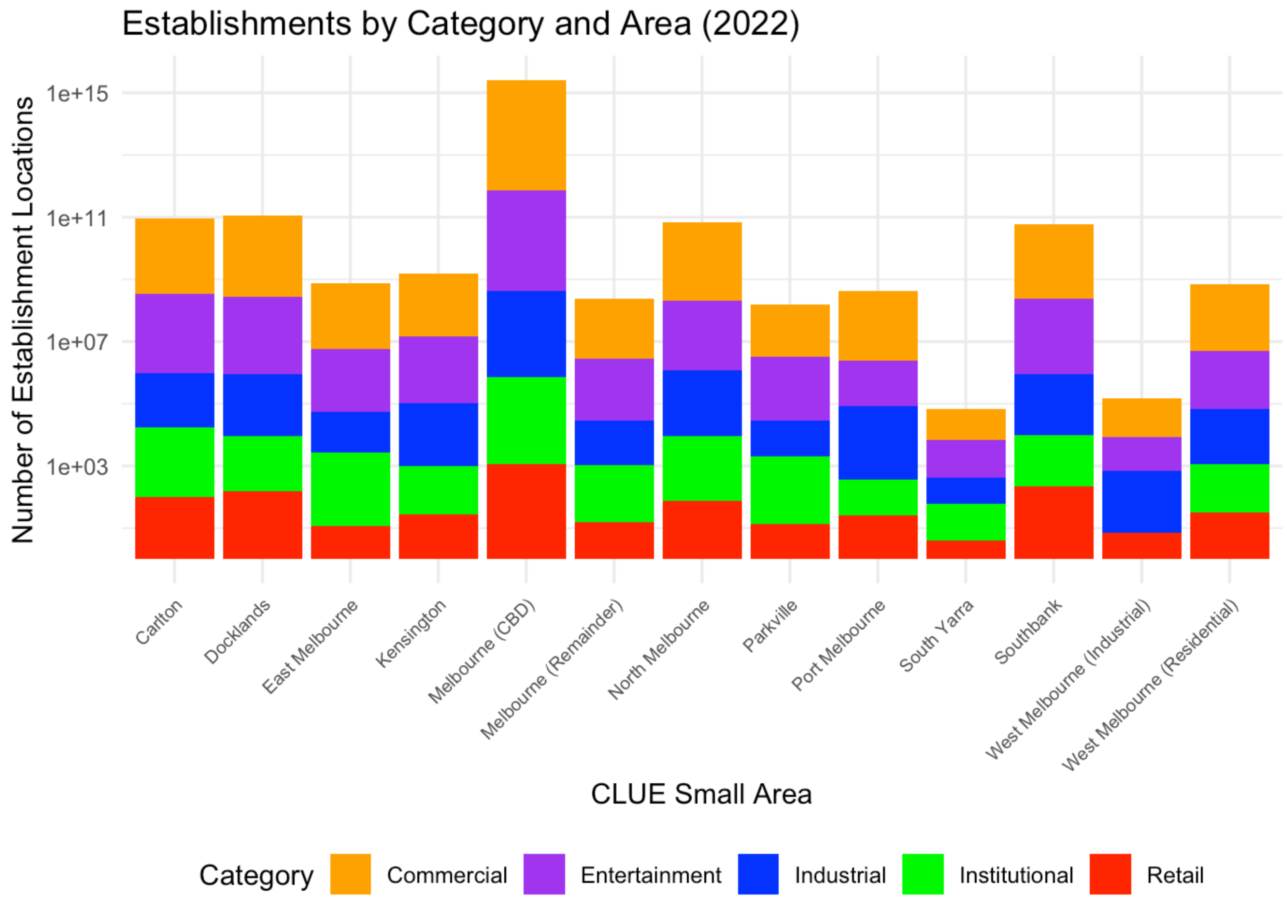
* Melbourne (CBD) consistently has the largest number of establishments compared to other areas. The distribution of categories remains relatively stable over time, with Commercial and Entertainment sectors making up a large portion of the total establishments. * Areas like Docklands, Port Melbourne, Kensington have experienced significant growth and establishment diversity over

the last two decades, with a focus on Commercial and Entertainment sectors. * There is a notable decline in Industrial establishments, especially in West Melbourne. * Other areas such as South Yarra, and Carlton have maintained a stable number of establishments with no dramatic shifts in category distribution.

Then only focus on the year 2022 as the most recent data we can get. Create a stacked bar plot for all areas with catogories of 2022.

Hide

```
        anzsic_category %>%
          filter(`Census year` == 2022) %>%
          pivot_longer(cols = Industrial:Commercial, names_to = "Categor
y", values_to = "Establishments") %>%
          filter(Establishments > 0) %>%
          ggplot(aes(x = `CLUE small area`, y = Establishments, fill = Cat
egory)) +
          geom_bar(stat = "identity", position = "stack") +
          labs(title = "Establishments by Category and Area (2022)",
               x = "CLUE Small Area",
               y = "Number of Establishment Locations",
               fill = "Category") +
          theme_minimal() +
          scale_fill_manual(values = c("orange", "purple", "blue", "gree
n", "red")) +
          theme(axis.text.x = element_text(angle = 45, hjust = 1, size =
7)) +
          theme(legend.position = "bottom") +
          scale_y_log10() # Apply log10 scaling, reduce the differences be
tween data
```



In most areas, Commercial and Entertainment establishments are the dominant categories, particularly in high-density areas like the Melbourne CBD and Docklands.

Areas like Kensington, North Melbourne, and Parkville show more balanced distributions across categories.

I wonder which areas have similar patterns, so cluster analysis is very important. Before we start, scale the data to reduce the impact of the much larger values in the CBD.

Hide

```
anzsic_2022 <- anzsic_category %>%
  filter(`Census year` == 2022) %>%
  select(-`Census year`)

scaled_anzsic_2022 <- anzsic_2022 %>%
  select(-`CLUE small area`)

scaled_anzsic_2022 <- scale(scaled_anzsic_2022)
scaled_anzsic_2022
```

```
##        Industrial Entertainment      Retail Institutional  Commercial
##  [1,] -0.39452260    0.16808043 -0.12320825    0.28379148 -0.15612644
##  [2,] -0.12878199    0.13504042  0.03696247   -0.35484901 -0.02999633
##  [3,] -0.63368915   -0.36276229 -0.42019146    0.56319669 -0.30222716
##  [4,] -0.08227738   -0.26804761 -0.36680122   -0.49740269 -0.33165751
##  [5,]  3.05346183    3.22758471  3.25038760    3.07214145  3.30299191
##  [6,] -0.58718454   -0.36496496 -0.40684390   -0.32633827 -0.35583245
##  [7,]  0.09045402   -0.18654894 -0.20329361   -0.03552876 -0.10882765
##  [8,] -0.68019376   -0.33192495 -0.41351768    0.14123780 -0.39367149
##  [9,]  0.74816203   -0.51915164 -0.37681189   -0.61144563 -0.27174571
## [10,] -0.72669836   -0.54778631 -0.44688658   -0.61144563 -0.43361269
## [11,] -0.16199956    0.02490708  0.26387100   -0.44608336 -0.19606764
## [12,] -0.10220793   -0.55659698 -0.43687591   -0.69697784 -0.42625510
## [13,] -0.39452260   -0.41782896 -0.35679055   -0.48029624 -0.29697173
## attr(,"scaled:center")
##    Industrial Entertainment        Retail Institutional    Commercial
##      116.3846      264.6923      137.9231      122.2308      422.5385
## attr(,"scaled:scale")
##    Industrial Entertainment        Retail Institutional    Commercial
##      150.5227      453.9951      299.6802      175.3725      951.3985
```

However, CBD still have much larger values than others, log tranformation has used to reduce values.

Hide

```
scaled_2022_log <- log1p(scaled_anzsic_2022)
scaled_2022_log
```

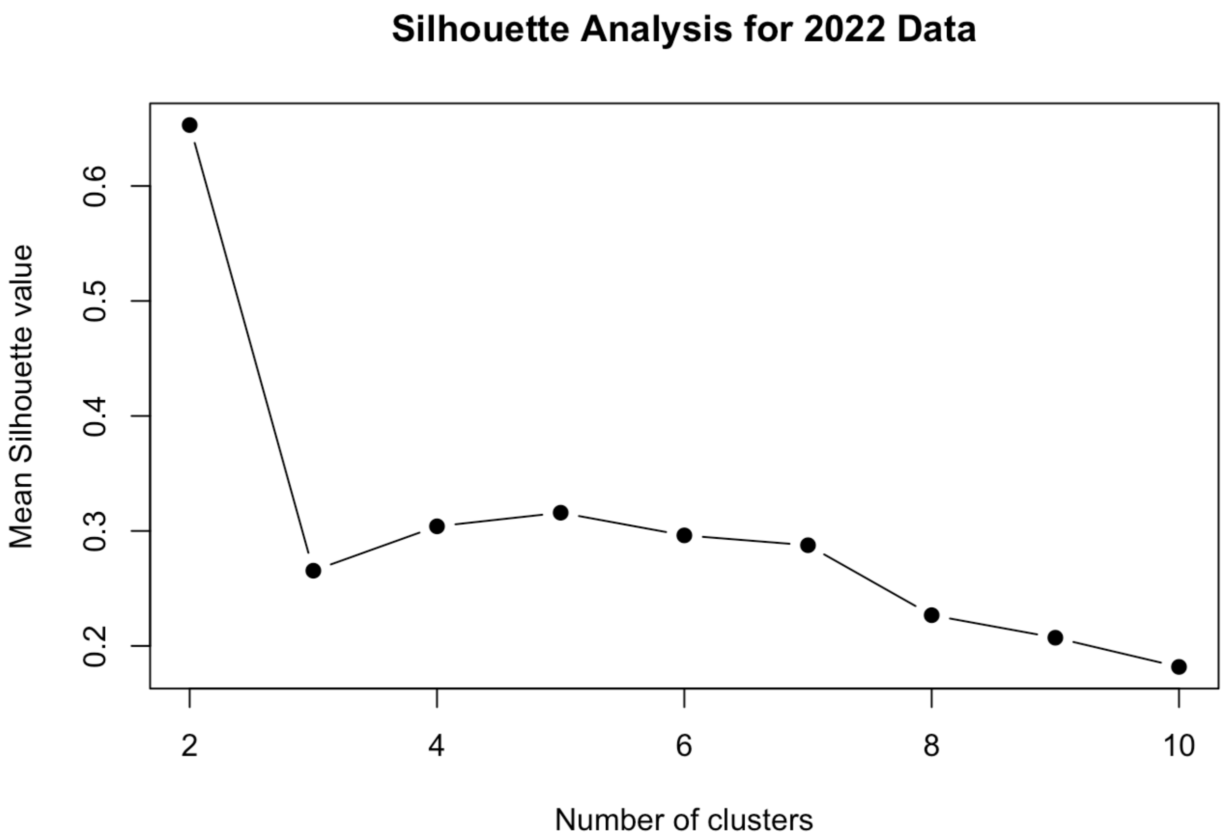
```
##        Industrial Entertainment      Retail Institutional  Commercial
##  [1,] -0.50173804    0.15536174 -0.13148577    0.24981779 -0.16975261
##  [2,] -0.13786303    0.12666827  0.03629574   -0.43827089 -0.03045542
##  [3,] -1.00427299   -0.45061252 -0.54505734    0.44673289 -0.35986167
##  [4,] -0.08586009   -0.31203981 -0.45697088   -0.68796600 -0.40295453
##  [5,]  1.39957129    1.44163084  1.44701018    1.40416902  1.45931057
##  [6,] -0.88475462   -0.45407510 -0.52229768   -0.39502718 -0.43979642
##  [7,]  0.08659414   -0.20646951 -0.22726906   -0.03617527 -0.11521743
##  [8,] -1.14003995   -0.40335477 -0.53361276    0.13211346 -0.50033334
##  [9,]  0.55856497   -0.73220332 -0.47290687   -0.94532217 -0.31710499
## [10,] -1.29717919   -0.79360044 -0.59219221   -0.94532217 -0.56847714
## [11,] -0.17673666    0.02460195  0.23417923   -0.59074107 -0.21824015
## [12,] -0.10781678   -0.81327616 -0.57425527   -1.19394933 -0.55557041
## [13,] -0.50173804   -0.54099100 -0.44128487   -0.65449633 -0.35235818
## attr(,"scaled:center")
##    Industrial Entertainment        Retail Institutional    Commercial
##      116.3846      264.6923      137.9231      122.2308      422.5385
## attr(,"scaled:scale")
##    Industrial Entertainment        Retail Institutional    Commercial
##      150.5227      453.9951      299.6802      175.3725      951.3985
```

Define number of clusters, using 2 methods.

Method 1: Silhouette score

Hide

```
        library(cluster)

        silhouette_vals <- c()

        for (i in 2:10) {
          kmeans_result <- kmeans(scaled_2022_log, centers = i, iter.max =
10, nstart = 10)
          silhouette_val <- silhouette(kmeans_result$cluster, dist(scaled_
2022_log))
          silhouette_vals <- c(silhouette_vals, mean(silhouette_val[, 3]))
# Mean silhouette value for each cluster
        }

        plot(2:10, silhouette_vals, type = "b", pch = 19, xlab = "Number o
f clusters",
             ylab = "Mean Silhouette value", main = "Silhouette Analysis f
or 2022 Data")
```

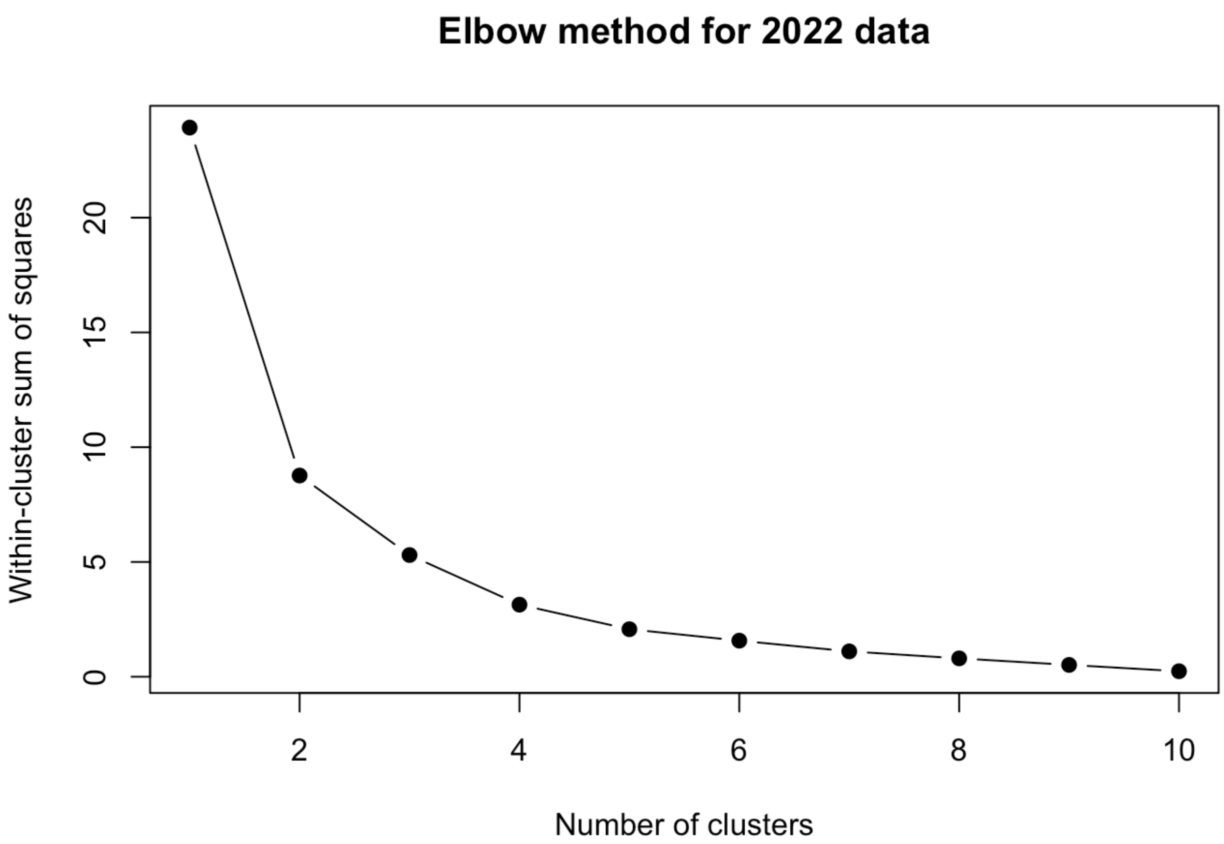


The graph highlights 3 clusters are the best choice.

Method 2: Elbow method

Hide

```
        wss <- c()
        for (i in 1:10) {
          kmeans_result <- kmeans(scaled_2022_log, centers = i, iter.max =
10, nstart = 10)
          wss[i] <- kmeans_result$tot.withinss
        }

        plot(1:10, wss, type = "b", pch = 19, xlab = "Number of clusters",
             ylab = "Within-cluster sum of squares", main = "Elbow method
for 2022 data")
```



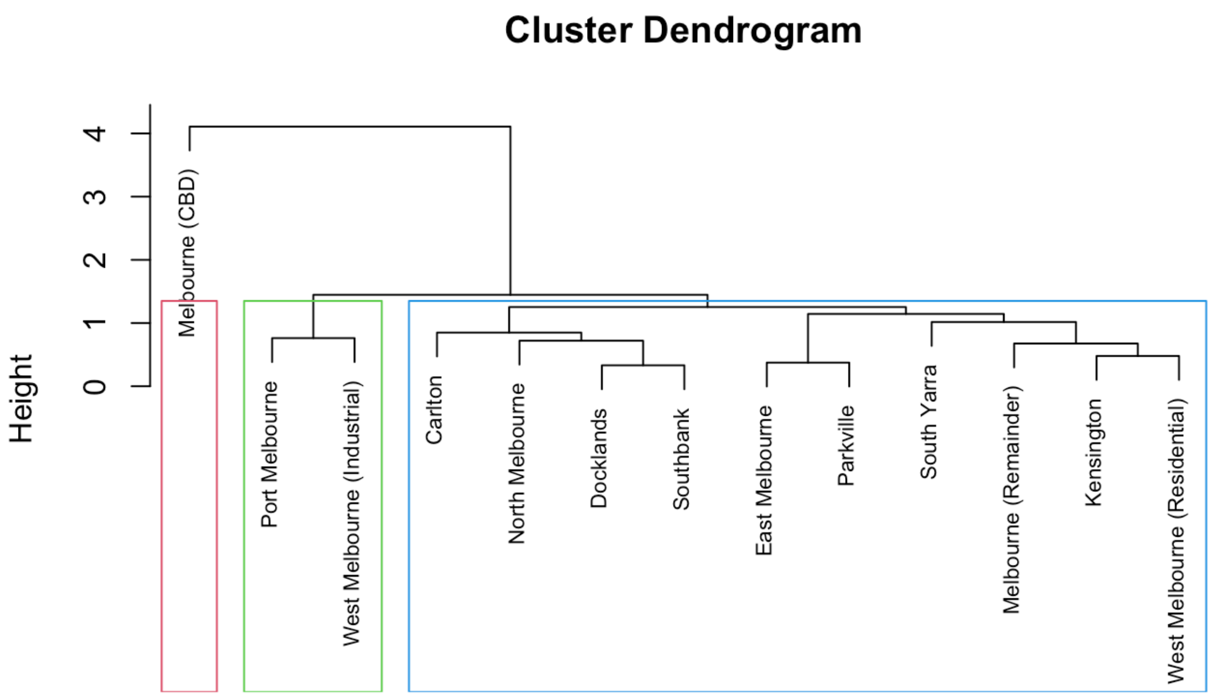
The graph highlights 2 clusters are the best choice.

However, Melbourne (CBD) has larger values and will always be splited as one cluster, so I choose 3 clusters for analyzing.

Method 1: Hierarchical clusters

Hide

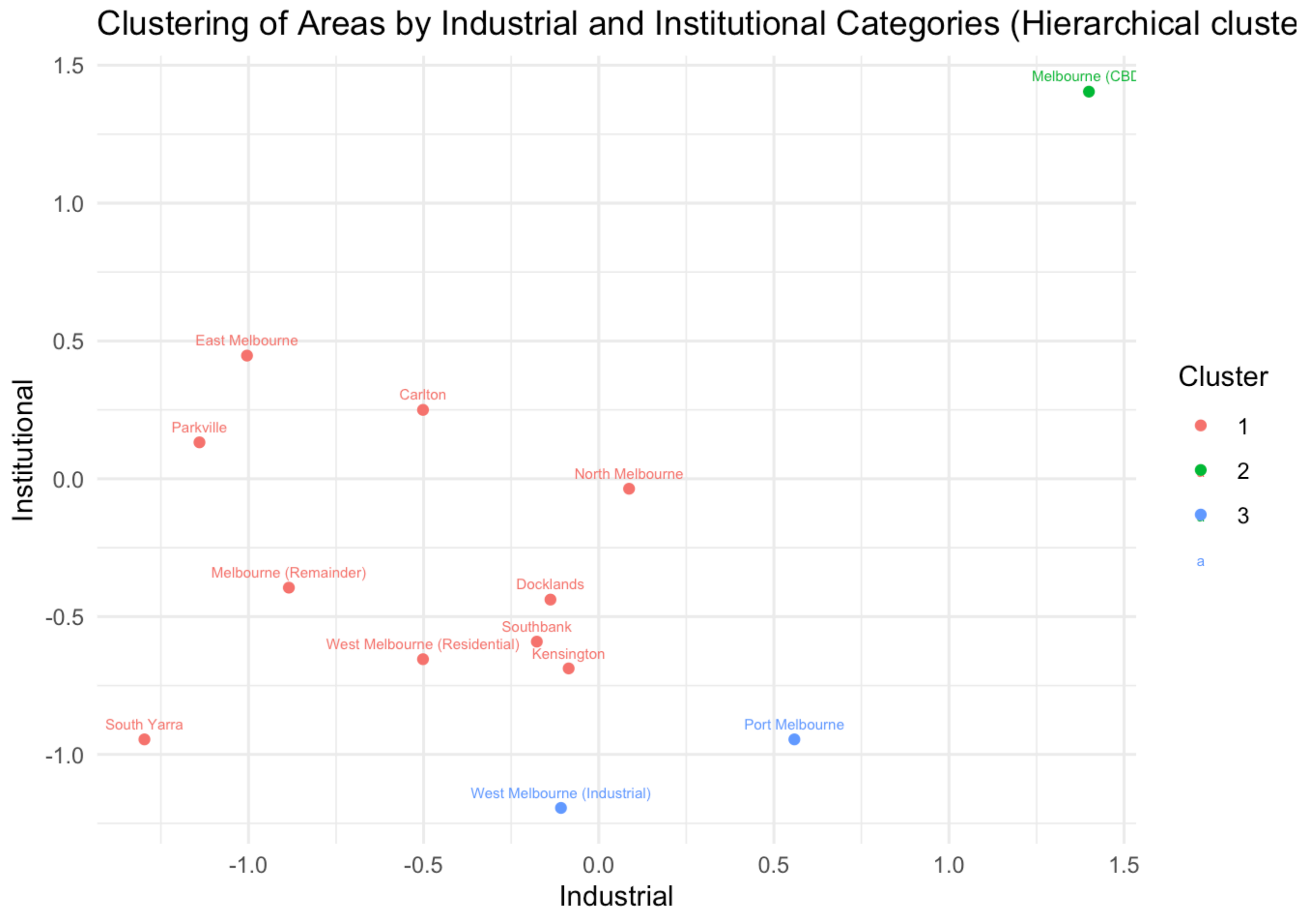
```
        clusters <- hclust(dist(scaled_2022_log), method = 'average')
        plot(clusters, labels = anzsic_2022$`CLUE small area`, cex = 0.7,
xlab = "", sub = "")
        rect.hclust(clusters, k = 3, border = 2:4)
```



Use Industrial and Institutional variables as examples to analyze differences between clusters

Hide

```
        clusterCut <- cutree(clusters, k = 3)

        ggplot(as.data.frame(scaled_2022_log), aes(x = Industrial, y = Ins
titutional, color = as.factor(clusterCut))) +
          geom_point() +
          geom_text(aes(label = anzsic_2022$`CLUE small area`), vjust = -
1, size = 2) +
          labs(title = "Clustering of Areas by Industrial and Institutiona
l Categories (Hierarchical clusters)",
               x = "Industrial", y = "Institutional", color = "Cluster") +
          theme_minimal()
```



Cluster 2: Melbourne (CBD) Cluster 3: West Melbourne (Industrial), Port Melbourne Cluster 1: Other areas

- Most of Cluster 1 have a more balanced or moderate level of institutional and industrial establishments, without one dominating over the other.
- Cluster 2 forms its own distinct cluster due to its high concentration of both Institutional and Industrial establishments, which sets it apart from all other areas.
- Cluster 3 is in a separate cluster due to their higher focus on industrial activity and lower institutional presence.

Hide

```
        # Calculate silhouette score
        cluster_hier_numeric <- as.numeric(as.character(clusterCut))
        silhouette_hierarchical <- silhouette(cluster_hier_numeric, dist(s
caled_2022_log))
        mean_silhouette_hierarchical <- mean(silhouette_hierarchical[, 3])

        mean_silhouette_hierarchical
```

```
## [1] 0.232817
```

Method 2: k-means

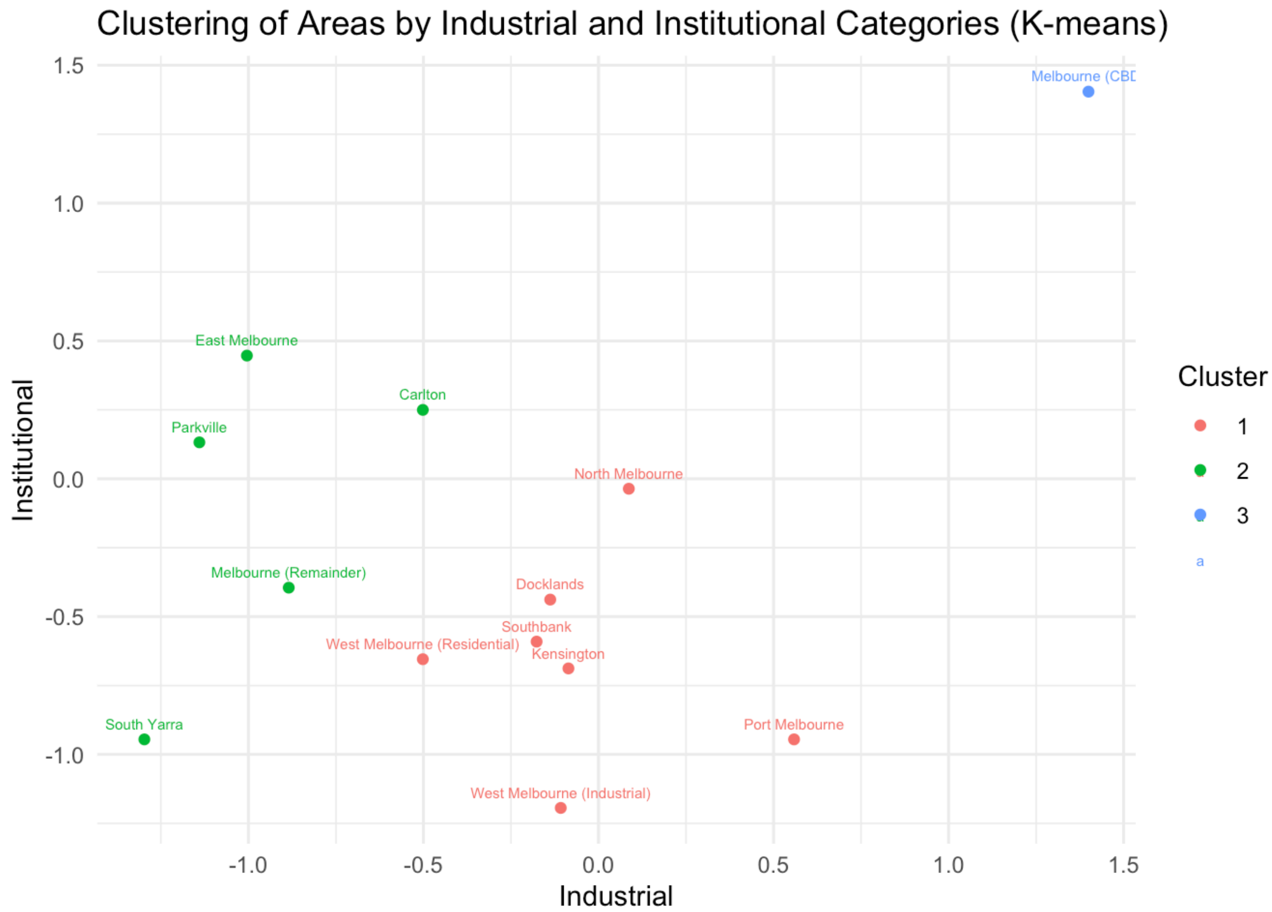
Hide

```
        set.seed(1234)
        cluster_k <- kmeans(scaled_2022_log, 3, iter.max = 10, nstart = 1
0)
        # Create a data frame to save area name and cluster number
        area_with_clusters <- data.frame(Area = anzsic_2022$`CLUE small ar
ea`, Cluster = as.factor(cluster_k$cluster))

        area_with_clusters
```

```
##                            Area Cluster
## 1                       Carlton       2
## 2                     Docklands       1
## 3                East Melbourne       2
## 4                    Kensington       1
## 5              Melbourne (CBD)       3
## 6        Melbourne (Remainder)       2
## 7               North Melbourne       1
## 8                     Parkville       2
## 9                Port Melbourne       1
## 10                  South Yarra       2
## 11                    Southbank       1
## 12  West Melbourne (Industrial)       1
## 13 West Melbourne (Residential)       1
```

Hide

```
        cluster_k$cluster <- as.factor(cluster_k$cluster)

        ggplot(as.data.frame(scaled_2022_log), aes(x = Industrial, y = Ins
titutional, color = cluster_k$cluster)) +
          geom_point() +
          geom_text(aes(label = anzsic_2022$`CLUE small area`), vjust = -
1, size = 2) +
          ggtitle("Clustering of Areas by Industrial and Institutional Cat
egories (K-means)") +
          labs(color = "Cluster") +
          theme_minimal()
```



* The majority of areas in Cluster 1 have relatively lower industrial and institutional activities * Carlton, Parkville, East Melbourne, and similar areas form a group of institutional-heavy areas. * The Melbourne CBD is distinct, having significantly more industrial and institutional establishments compared to other areas

Hide

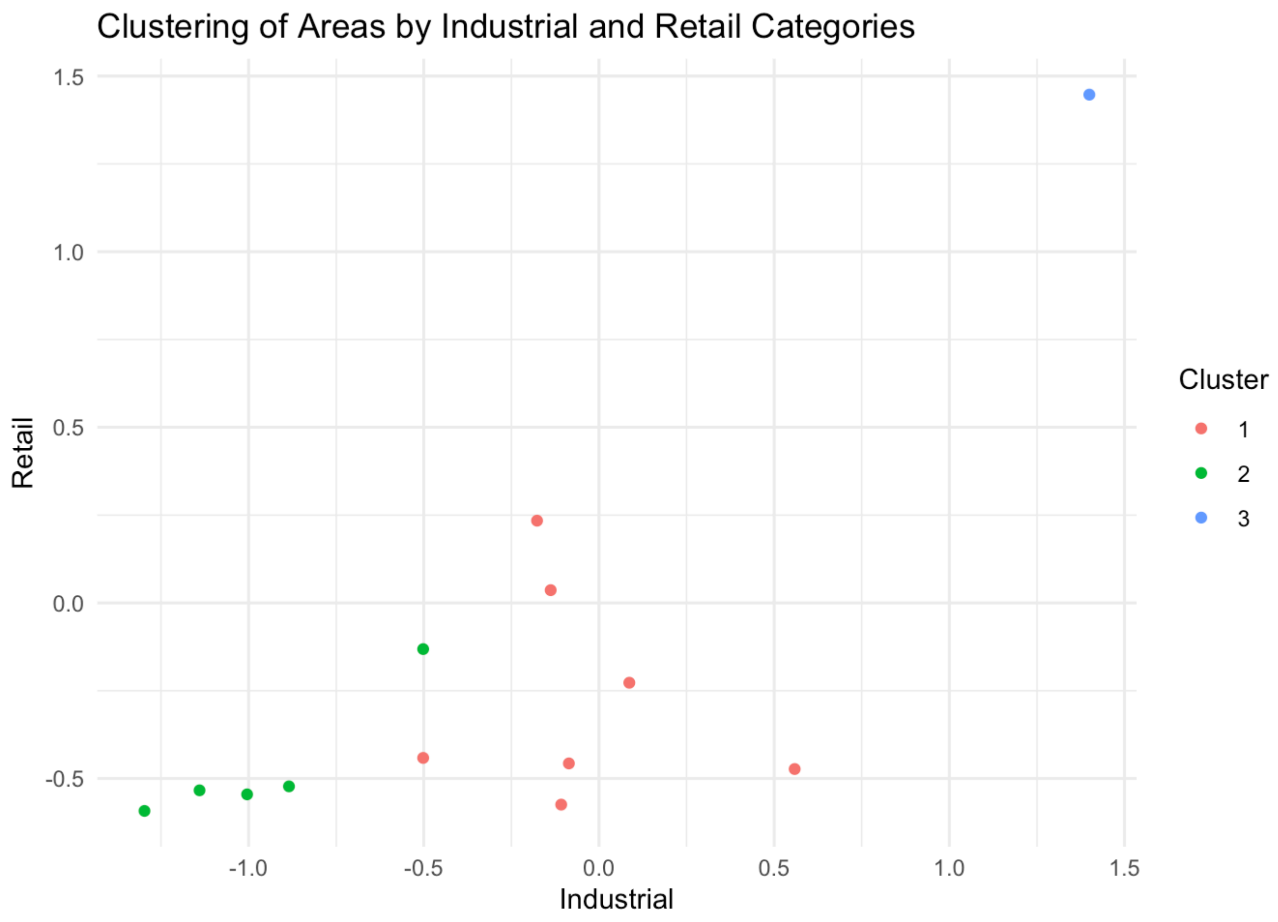
```
        # Calculate silhouette score
        cluster_kmeans_numeric <- as.numeric(as.character(cluster_k$cluste
r))
        silhouette_kmeans <- silhouette(cluster_kmeans_numeric, dist(scale
d_2022_log))
        mean_silhouette_kmeans <- mean(silhouette_kmeans[, 3])

        mean_silhouette_kmeans
```

```
## [1] 0.2654368
```

K-means is a better method in this case. Will use k-means to get similar patterns for other variables

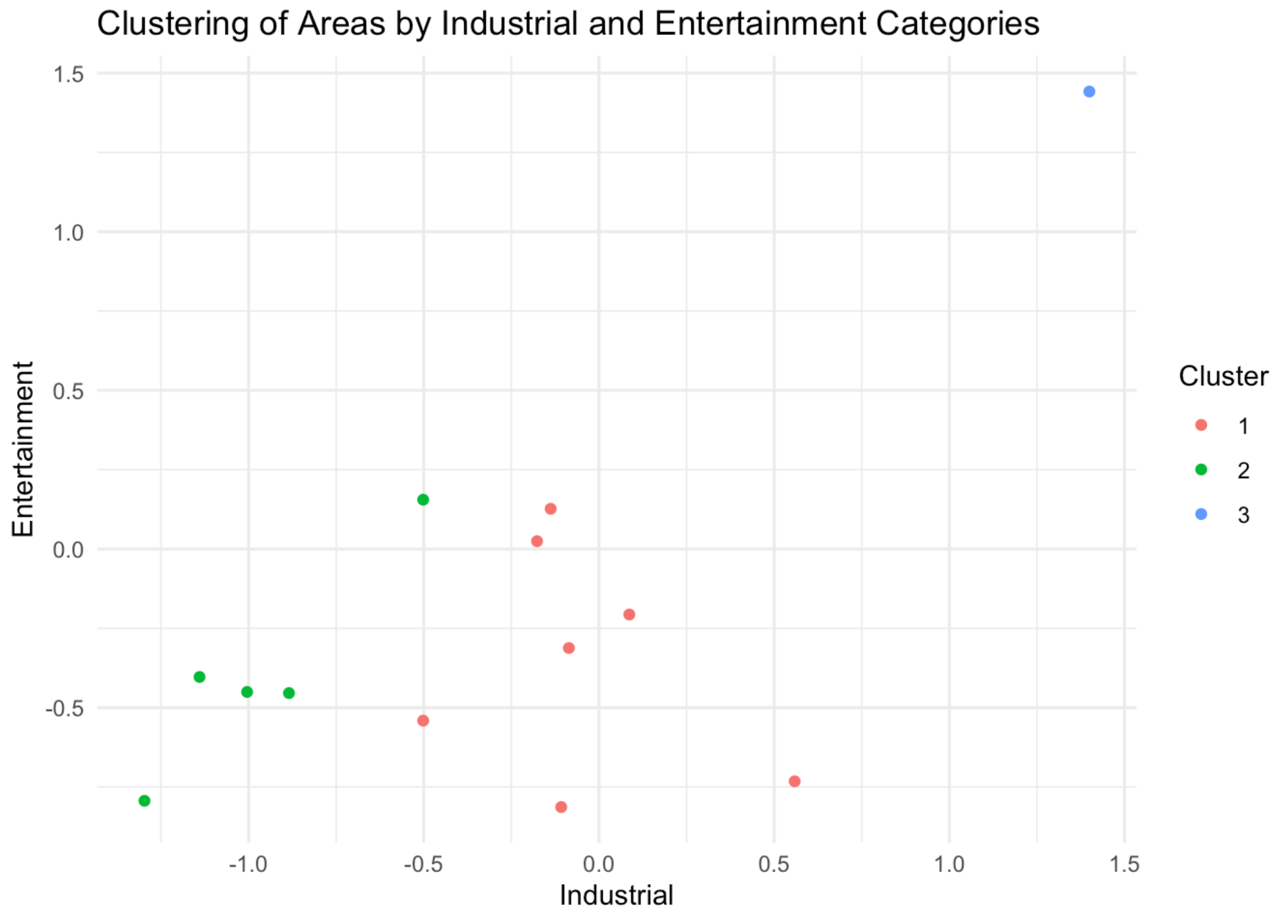
Hide

```
    ggplot(as.data.frame(scaled_2022_log), aes(x = Industrial, y = Retail, color = cluster_k$cluster)) +
      geom_point() +
      ggtitle("Clustering of Areas by Industrial and Retail Categories") +
      labs(color = "Cluster") +
      theme_minimal()
```



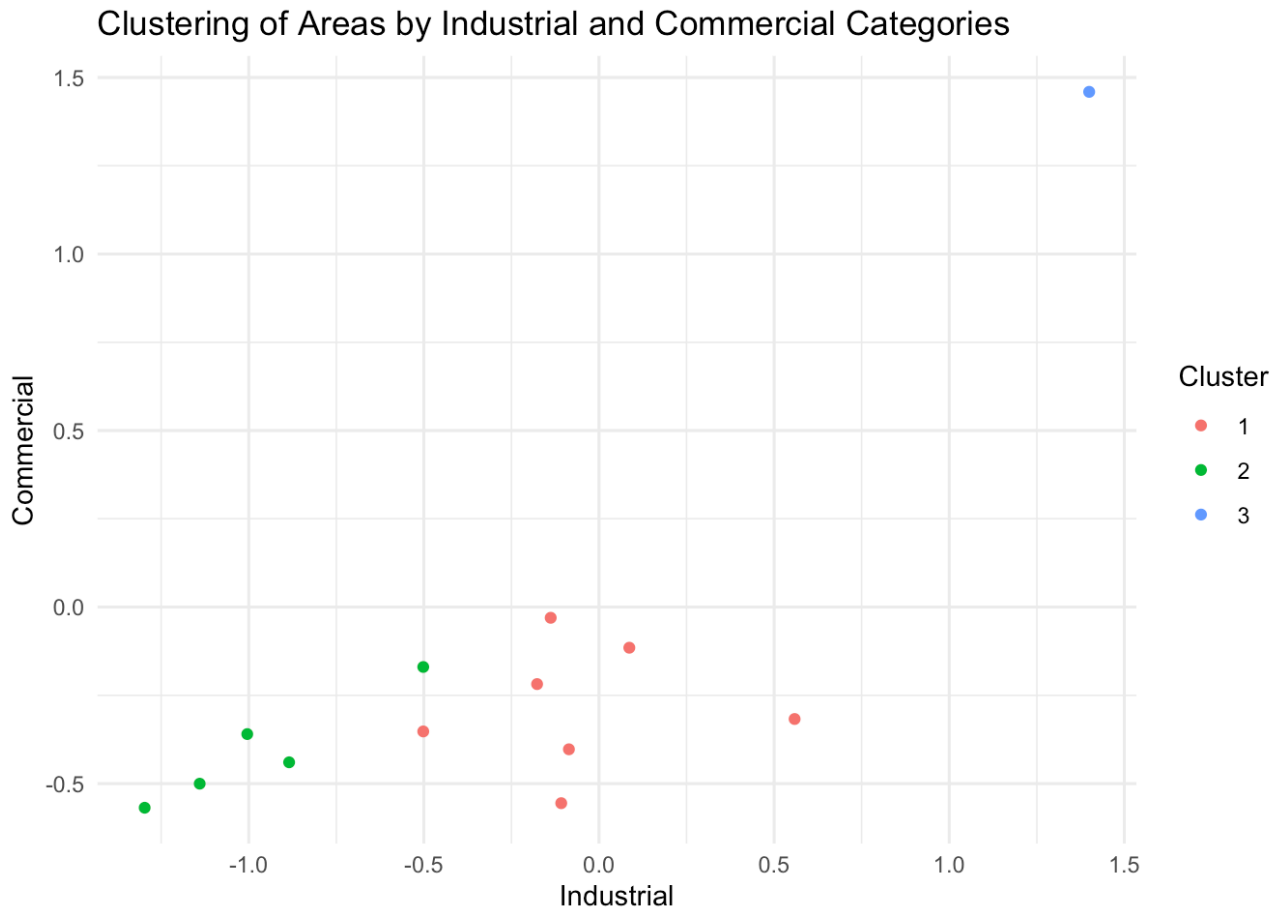
Hide

```
    ggplot(as.data.frame(scaled_2022_log), aes(x = Industrial, y = Entertainment, color = cluster_k$cluster)) +
      geom_point() +
      ggtitle("Clustering of Areas by Industrial and Entertainment Categories") +
      labs(color = "Cluster") +
      theme_minimal()
```



Hide

```
        ggplot(as.data.frame(scaled_2022_log), aes(x = Industrial, y = Com
mercial, color = cluster_k$cluster)) +
          geom_point() +
          ggtitle("Clustering of Areas by Industrial and Commercial Catego
ries") +
          labs(color = "Cluster") +
          theme_minimal()
```



Conclusion: * Cluster 1 represents mixed-use areas with moderate to low establishment activity across all categories. * Cluster 2 consists of institutional or residential areas with moderate activity in Retail, Entertainment, or Commercial but with relatively lower Industrial activity. * Cluster 3, with the Melbourne CBD, reflects the area's dominant role in all categories, acting as the hub for Industrial, Retail, Entertainment, and Commercial establishments.

Make predictions for the future 5 years.

Hide

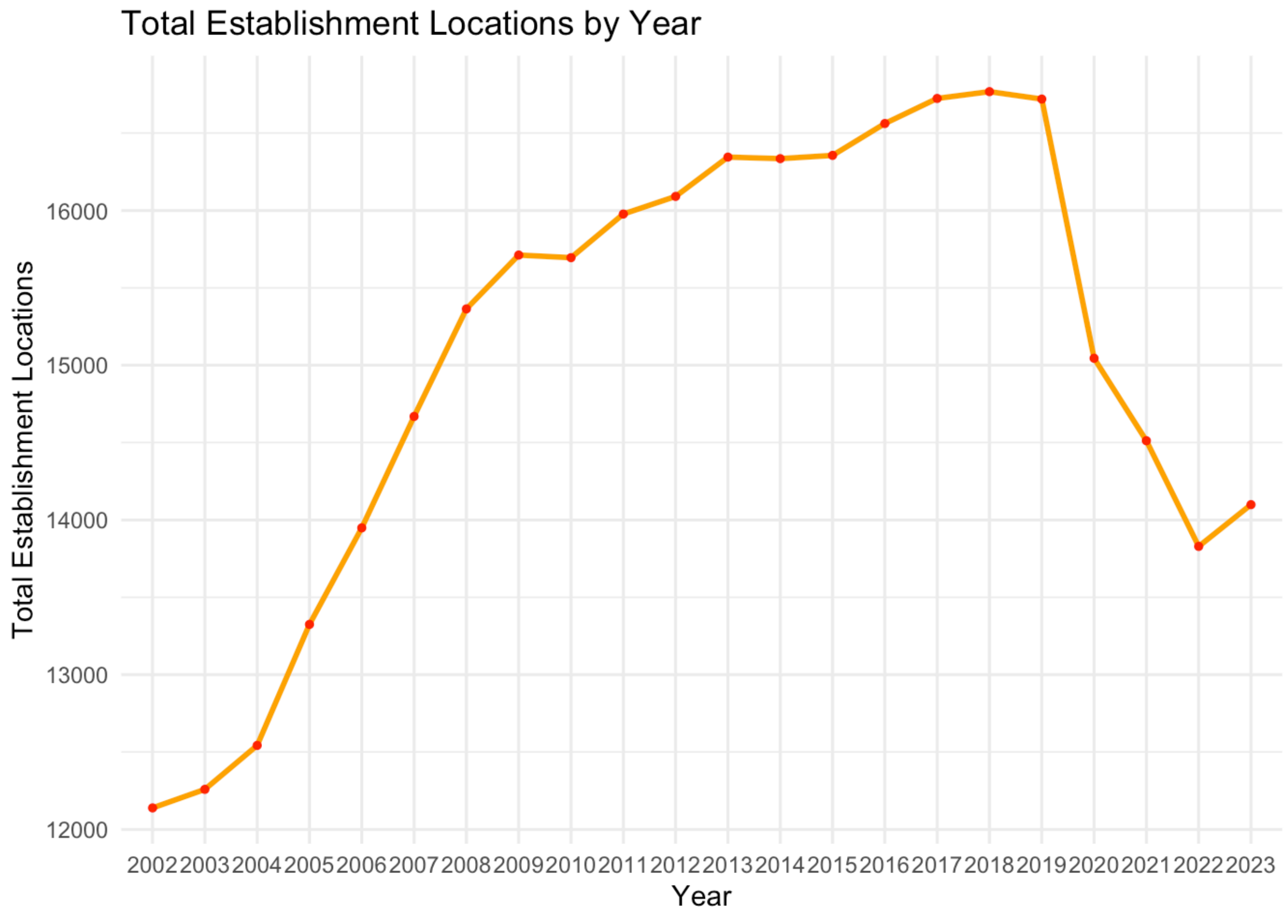
```
total_anz <- total_anz %>%
  select(-`Block ID`)

head(total_anz)
```

```
## # A tibble: 6 × 22
##   `Census year` `CLUE small area`  Accommodation and Fo…¹ Administrativ
e and S…²
##   <chr>         <chr>                                <dbl>
<dbl>
## 1 2021          City of Melbourne…                    2975
460
## 2 2016          City of Melbourne…                    3005
646
## 3 2015          City of Melbourne…                    2878
647
## 4 2014          City of Melbourne…                    2814
668
## 5 2010          City of Melbourne…                    2401
731
## 6 2009          City of Melbourne…                    2329
756
## # ℹ abbreviated names: ¹`Accommodation and Food Services`,
## #   ²`Administrative and Support Services`
## # ℹ 18 more variables: `Agriculture, Forestry and Fishing` <dbl>,
## #   `Arts and Recreation Services` <dbl>, Construction <dbl>,
## #   `Education and Training` <dbl>,
## #   `Electricity, Gas, Water and Waste Services` <dbl>,
## #   `Financial and Insurance Services` <dbl>, …
```

Hide

```
        ggplot(data = total_anz, aes(x = `Census year`, y = `Total establi
shments in block`, group = 1)) +
          geom_line(color = "orange", size = 1) +
          geom_point(color = "red", size = 1) +
          labs(title = "Total Establishment Locations by Year",
               x = "Year",
               y = "Total Establishment Locations") +
          theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.
4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning w
as
## generated.
```



The number of establishment locations grew steadily from 2002 to 2019 but experienced a sharp decline starting in 2020, likely due to the impact of the COVID-19 pandemic.

Hide

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

Hide

```
library(forecast)
```

First sort the data by year and create a time series starts from 2002. Using 3 methods to predict the trend for next 5 years.

Hide

```
total_anz <- total_anz[order(total_anz$`Census year`), ]
total_anz_ts <- ts(total_anz$`Total establishments in block`, start = c(2002), frequency = 1)
```

Method 1: Linear Regression Model.

Hide

```
# time series is yearly, therefore, no seasonal impact
linear_anz <- tslm(total_anz_ts ~ trend)
accuracy(linear_anz)
```

```
##                         ME     RMSE      MAE        MPE     MAPE     MAS
E
## Training set -3.72066e-13 1272.293 1176.149 -0.7620243 8.027796 3.12964
1
##                   ACF1
## Training set 0.8105473
```

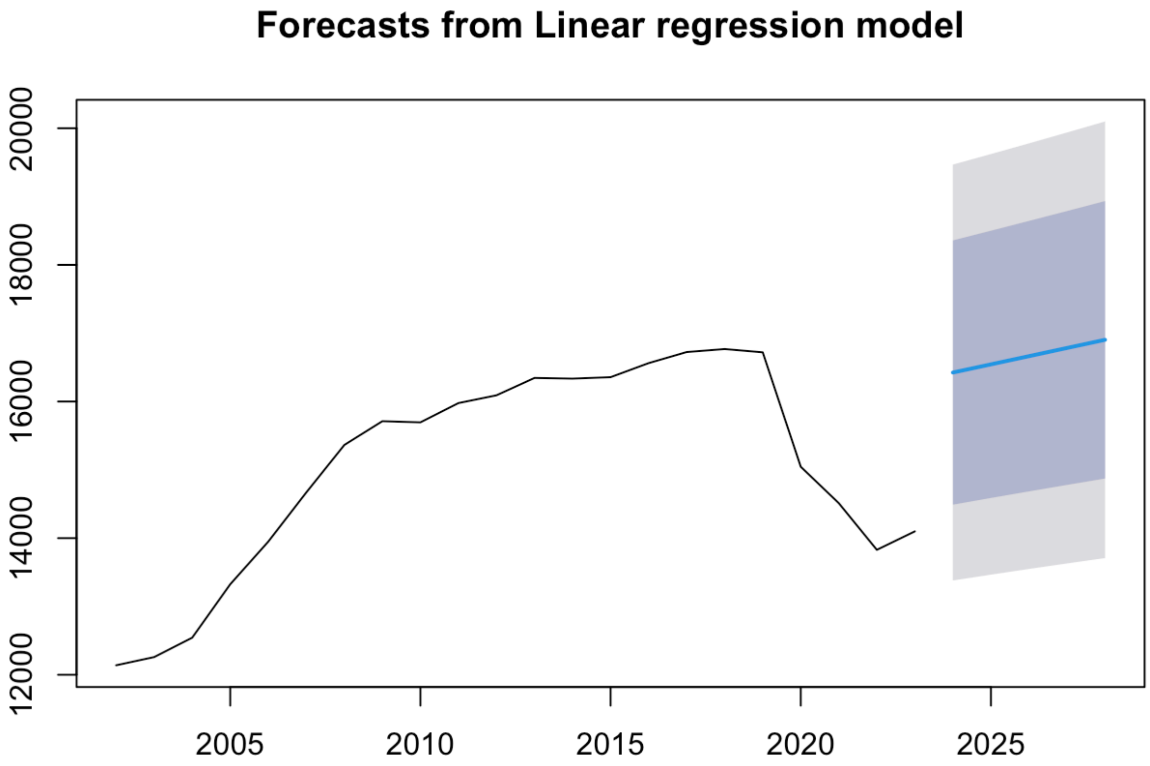
Hide

```
forecast(linear_anz, h=5)
```

```
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2024       16424.18 14491.06 18357.31 13381.62 19466.74
## 2025       16544.00 14589.07 18498.93 13467.13 19620.87
## 2026       16663.82 14685.54 18642.09 13550.20 19777.44
## 2027       16783.64 14780.52 18786.75 13630.92 19936.35
## 2028       16903.45 14874.06 18932.85 13709.38 20097.53
```

Hide

```
plot(forecast(linear_anz, h=5))
```



The linear regression forecast shows a steady trend for the number of establishment locations over the next five years.

Method 2: ETS Model

Hide

```
ets_anz <- ets(total_anz_ts)
accuracy(ets_anz)
```

```
##                    ME     RMSE      MAE      MPE     MAPE      MASE
ACF1
## Training set 89.88857 525.6733 359.5396 0.623651 2.460266 0.9567069 0.5
120304
```

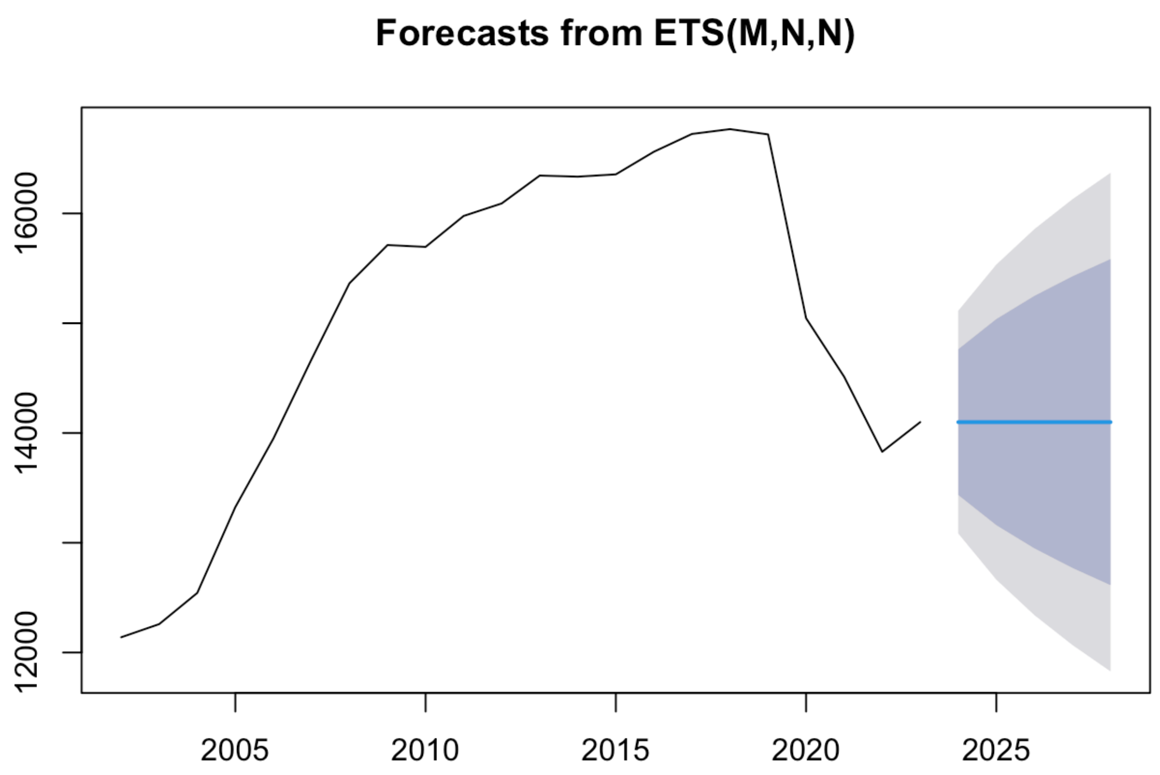
Hide

```
forecast(ets_anz,h=5)
```

```
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2024       14098.97 13436.02 14761.92 13085.08 15112.87
## 2025       14098.97 13161.15 15036.80 12664.70 15533.25
## 2026       14098.97 12950.01 15247.93 12341.79 15856.16
## 2027       14098.97 12771.83 15426.11 12069.28 16128.66
## 2028       14098.97 12614.69 15583.25 11828.96 16368.99
```

Hide

```
plot(forecast(ets_anz, h=5))
```



The ETS model forecast indicates a continued decline in the number of total establishment locations over the next five years, following the sharp drop observed in recent years.

Method 3: Auto.arima Model

Hide

```
arima_anz <- auto.arima(total_anz_ts)
summary(arima_anz)
```

```
## Series: total_anz_ts
## ARIMA(0,2,1)
##
## Coefficients:
##           ma1
##       -0.5750
## s.e.   0.2815
##
## sigma^2 = 248391:  log likelihood = -152.29
## AIC=308.59   AICc=309.29   BIC=310.58
##
## Training set error measures:
##                    ME     RMSE      MAE        MPE     MAPE      MASE
## Training set -48.48194 463.1623 262.4461 -0.2448277 1.766468 0.6983486
##                    ACF1
## Training set 0.03847962
```

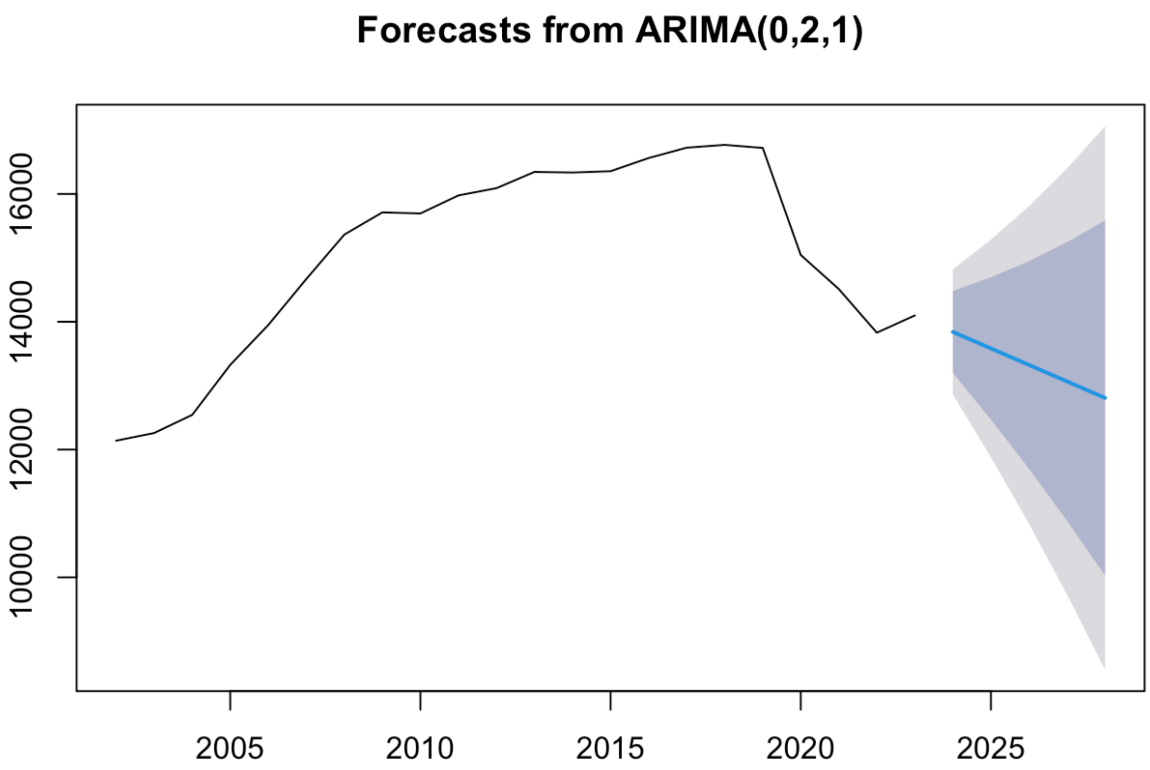
Hide

```
forecast(arima_anz, h = 5)
```

```
##      Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
## 2024       13840.85 13202.14 14479.56 12864.027 14817.67
## 2025       13582.70 12470.80 14694.61 11882.188 15283.21
## 2026       13324.55 11702.05 14947.05 10843.149 15805.95
## 2027       13066.40 10888.36 15244.44  9735.380 16397.42
## 2028       12808.25 10030.17 15586.33  8559.549 17056.95
```

Hide

```
plot(forecast(arima_anz, h = 5))
```



The ARIMA model shows a continued decline in total establishment locations, with the point forecast decreasing steadily each year from 2023 to 2027.

Hide

```
# Compare 3 methods' AIC
AIC(linear_anz, ets_anz, arima_anz)
```

```
## Warning in AIC.default(linear_anz, ets_anz, arima_anz): models are not
all
## fitted to the same number of observations
```

```
##            df      AIC
## linear_anz  3 382.9706
## ets_anz     3 349.1803
## arima_anz   2 308.5871
```

auto.arima has the smallest AIC, which means the prediction is more reliable. So we choose this prediction method for the following analysis.

In order to predict the trend of different industries, create a new data frame with 5 categories.

Hide

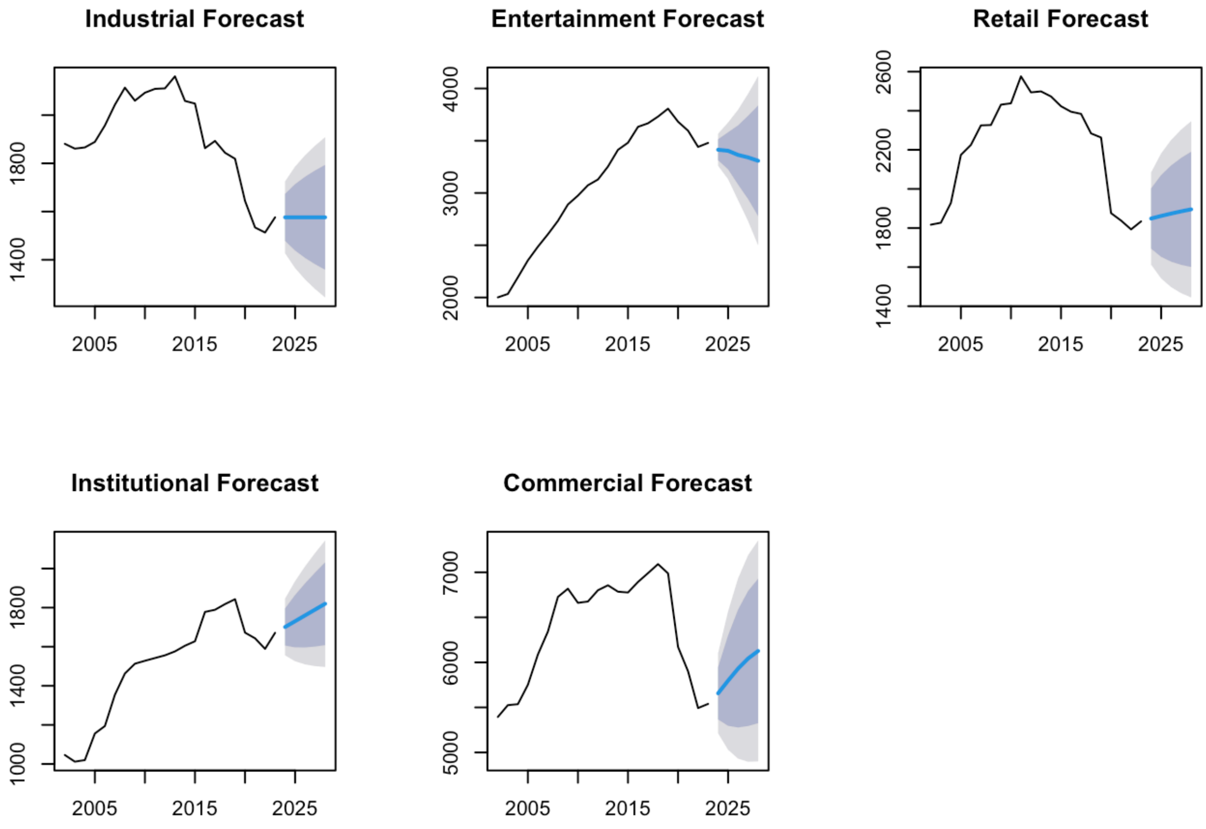
```
        total_cat <- setNames(data.frame(
          Year = total_anz$`Census year`,
          Industrial = rowSums(total_anz[, c(5, 7, 9, 13, 14, 20, 21)]),
          Entertainment = rowSums(total_anz[, c(3, 6)]),
          Retail = total_anz[, 19],
          Institutional = rowSums(total_anz[, c(8, 11)]),
          Commercial = rowSums(total_anz[, c(4, 10, 12, 15, 16, 17, 18)])
        ), c("Year", "Industrial", "Entertainment", "Retail", "Institution
al", "Commercial"))
```

Hide

```
        categories <- c("Industrial", "Entertainment", "Retail", "Institut
ional", "Commercial")

        # Create time series for each category and store them in a list
        ts_list <- lapply(categories, function(cat) ts(total_cat[[cat]], s
tart = min(total_cat$Year), frequency = 1))

        # Fit ARIMA models for each time series and store the models in a
list
        arima_models <- lapply(ts_list, auto.arima)

        # Forecast for the next 5 years for each category and store the fo
recasts in a list
        forecasts <- lapply(arima_models, forecast, h = 5)

        # Assign names to the forecasts
        names(forecasts) <- categories
```

Hide

```
        # Define plot titles
        titles <- c("Industrial Forecast", "Entertainment Forecast", "Reta
il Forecast",
                    "Institutional Forecast", "Commercial Forecast")

        par(mfrow = c(2, 3))

        # Use a loop to plot each forecast
        for (i in 1:length(forecasts)) {
          plot(forecasts[[i]], main = titles[i])
        }

        par(mfrow = c(1, 1))
```



The forecasts indicate an overall decline in most sectors, including industrial, entertainment, and retail establishments, reflecting a potential shift in economic or business conditions. The institutional sector shows moderate stability with slight fluctuations, while the commercial sector demonstrates a more resilient trend, with a potential slight recovery after recent declines.

Now only focus on Commercial category.

Hide

```
forecasts$Commercial
```

```
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2024       5656.548 5367.474 5945.621 5214.448 6098.647
## 2025       5797.459 5296.213 6298.705 5030.870 6564.048
## 2026       5931.644 5278.224 6585.064 4932.324 6930.963
## 2027       6043.295 5294.415 6792.175 4897.981 7188.609
## 2028       6126.771 5325.501 6928.040 4901.335 7352.207
```

The Commercial sector shows a moderate recovery after a period of decline, with forecasts indicating steady growth from 2023 to 2027.

There are 2 columns may related to IT jobs which IT students may be interested in, called “Information Media and Telecommunications” and “Professional, Scientific and Technical Services”.

Hide

```
        # Create time series for each column
        Info_media_ts <- ts(total_anz[, 12], start = min(total_anz$`Census
year`), frequency = 1)
        tech_serv_ts <- ts(total_anz[, 16], start = min(total_anz$`Census
year`), frequency = 1)

        # Fit ARIMA models to both time series
        Info_media <- auto.arima(Info_media_ts)
        tech_serv <- auto.arima(tech_serv_ts)

        # Forecast for the next 5 years for both columns
        Info_media_forecast <- forecast(Info_media, h = 5)
        tech_serv_forecast <- forecast(tech_serv, h = 5)

        Info_media_forecast
```

```
##      Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
## 2024       202.4317 185.7928 219.0705 176.98479 227.8785
## 2025       202.1086 171.1600 233.0573 154.77677 249.4405
## 2026       201.9251 157.7869 246.0632 134.42156 269.4286
## 2027       201.8207 145.7839 257.8575 116.11982 287.5216
## 2028       201.7614 135.0159 268.5069  99.68303 303.8398
```

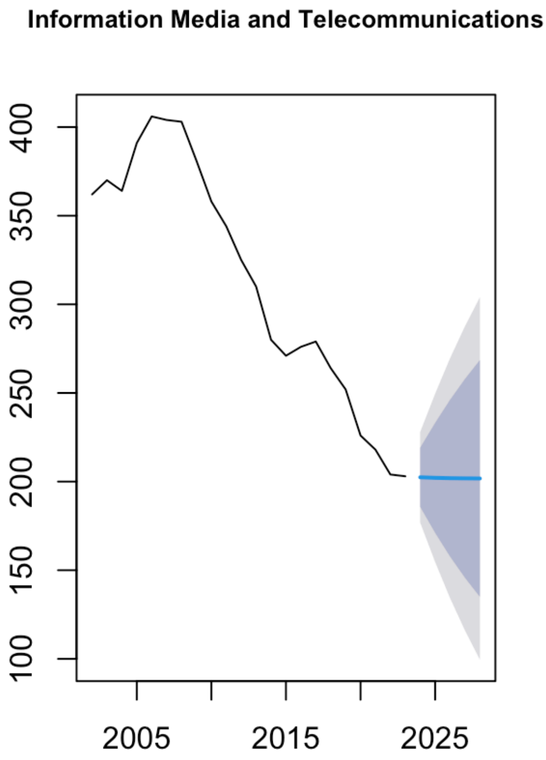
Hide

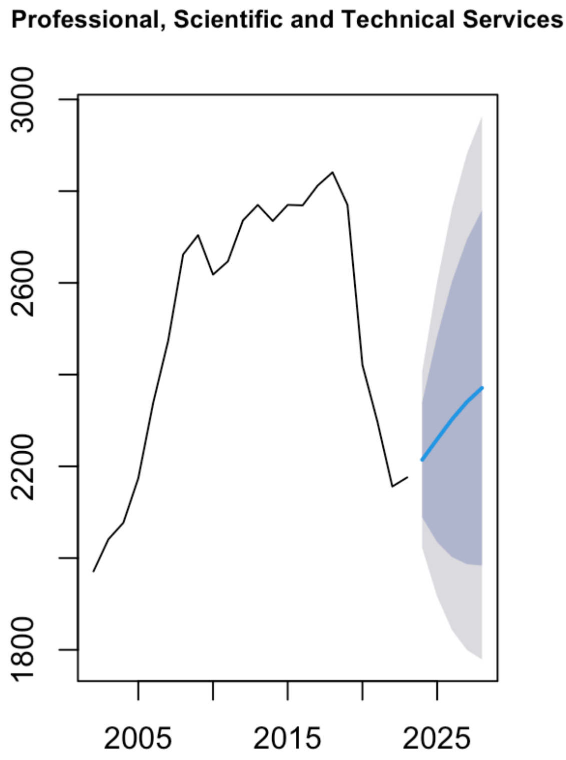
```
        tech_serv_forecast
```

```
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2024       2214.284 2089.034 2339.535 2022.730 2405.839
## 2025       2259.108 2035.196 2483.021 1916.664 2601.553
## 2026       2302.736 2002.200 2603.273 1843.105 2762.367
## 2027       2340.698 1986.972 2694.423 1799.721 2881.675
## 2028       2370.975 1983.940 2758.009 1779.057 2962.893
```

Hide

```
        # Plot the forecasts
        par(mfrow = c(1, 2))
        plot(Info_media_forecast, main = "Information Media and Telecommun
ications", cex.main = 0.8)
        plot(tech_serv_forecast, main = "Professional, Scientific and Tech
nical Services", cex.main = 0.8)
```





Hide

```
par(mfrow = c(1, 1))
```

For Information Media and Telecommunications, the forecast shows a steady decline, with the industry expected to continue shrinking gradually over the next few years. In contrast, Professional, Scientific, and Technical Services exhibits a sharper and more volatile decline. The steeper drop and wider confidence interval suggest more instability in this industry, with a potential for greater variability in the forecast outcomes.