# Data Analytics Mini Project 4

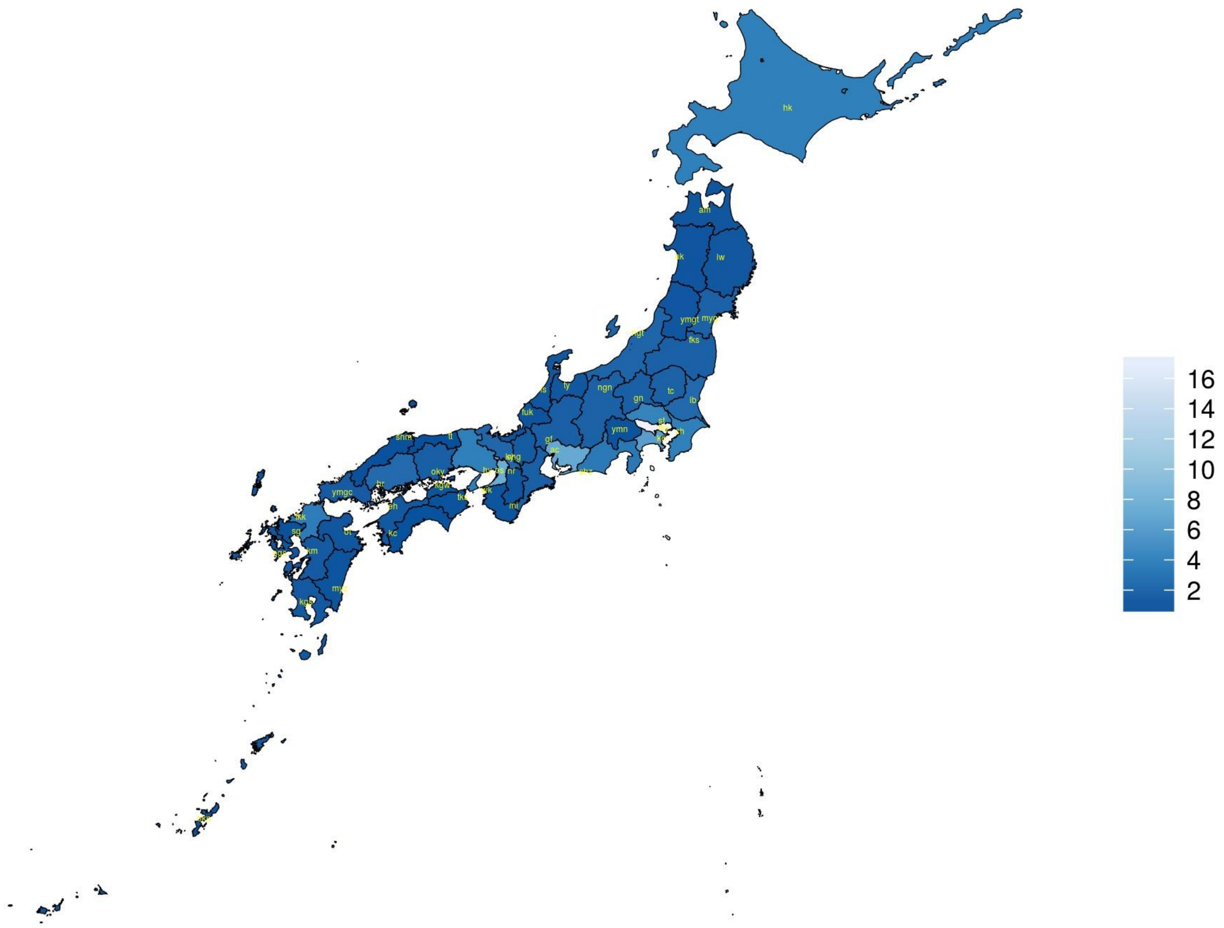
## M.Tech- Group 9.2

**Name of the Student:** Vijay Deshpande
**Student ID:** 201761003
**Role:q2, q3**

**Name of the Student:** Jayant Mahawar
**Student ID:** 201761004
**Role:q1, q4**

**Exercise 1**:
**Use R to make a map of the states/provinces/regions showing GDP or income inequality for countries according to your groups**

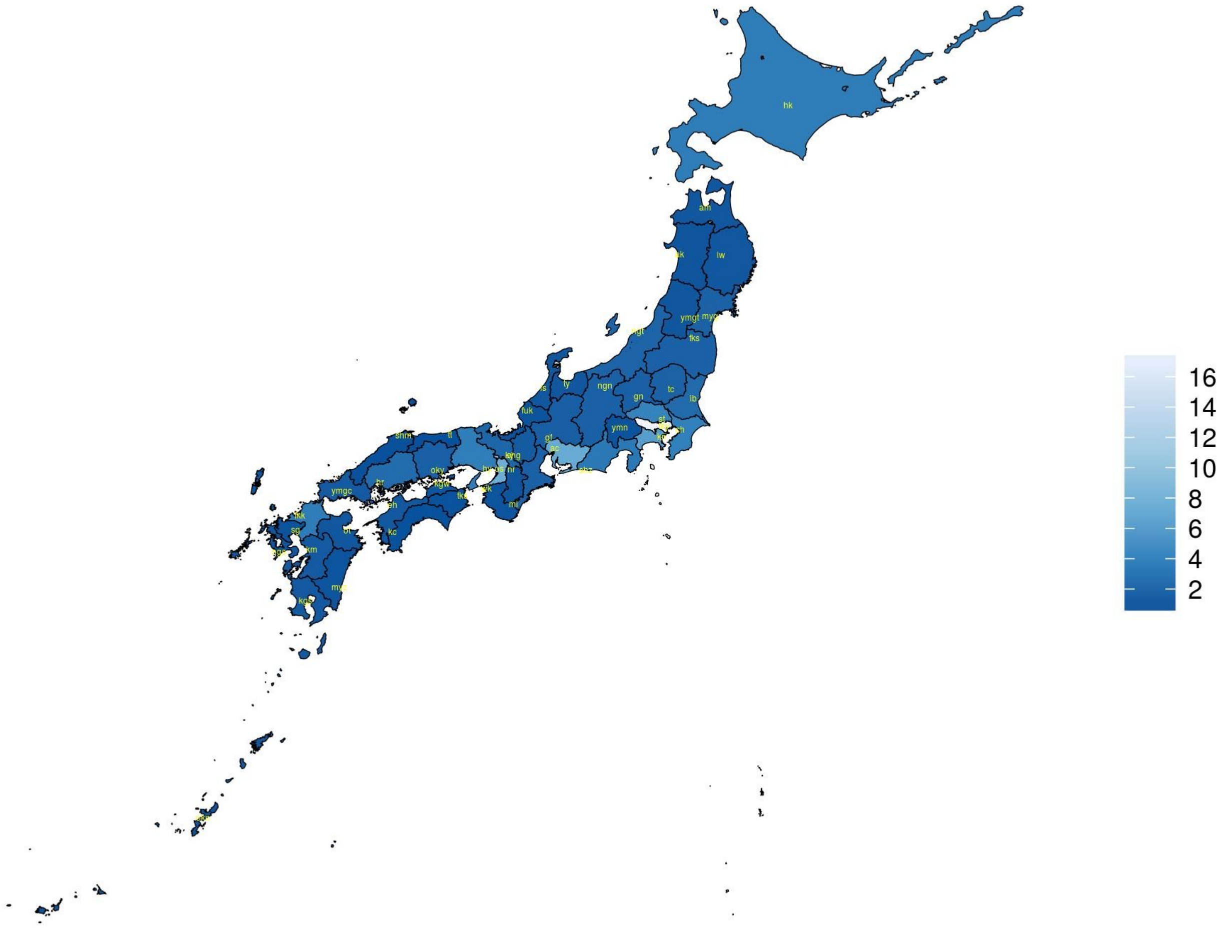
Figure 1: Gross Domestic Product of top 1% earners in 2005




**Inferences:**

Tokyo has the highest contribution to GDP (91086300 ¥)in the year 2005
Tottori has the least contribution to GDP (1999200 ¥) in the year 2005
Total GDP of Japan was 513560500 ¥ .

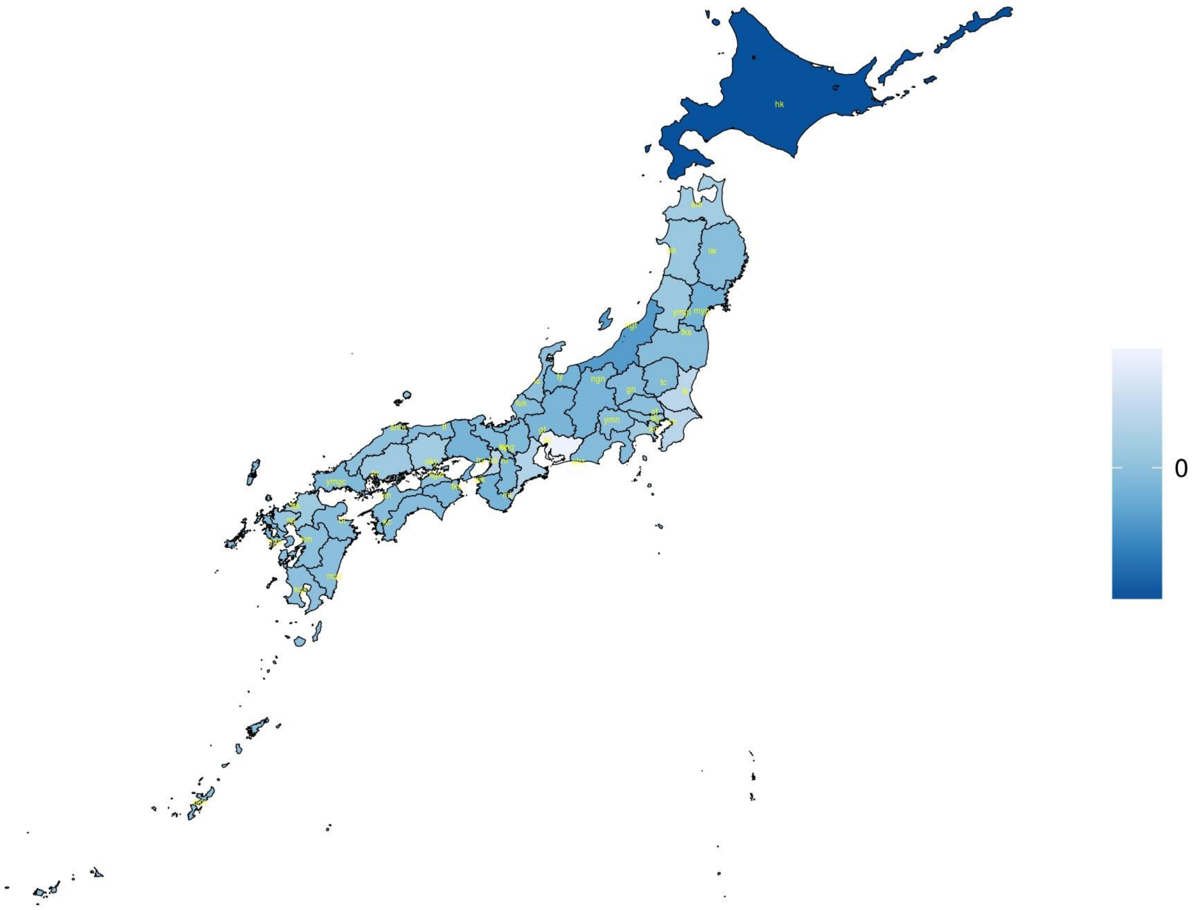
Figure 2: Gross Domestic Product of top 1% earners in 2007




**Inferences:**

Tokyo has the highest contribution to GDP (92300500 ¥)in the year  2007
Tottori has the least contribution to GDP (1999200 ¥) in the year 2007
Total GDP of Japan was 520249100 ¥.

Figure 3: Change in Gross Domestic Product, 2007 - 2005




**Inferences**

Highest decrease in the GDP  was observed by Hokkaido which is 0.2%
2005 = 19290100 ¥
2007 = 18458400 ¥

Highest Increase in the GDP was observed by  Aichi which is 0.2%.
2005 = 35756100 ¥
2007 = 37171900 ¥

There is no change in the contribution percentage of GDP by Kagoshima which is 1.05%.

**Exercise 2**:
**The following table shows the Myers-Briggs personality preferences for a random sample of 389 past computer science graduates (from a University) in the listed professions**

| Occupation | Personality preference Type | |
|---|---|---|
| | E | I |
| Faculty | 62 | 45 |
| Data Scientist | 56 | 81 |
| Entrepreneur | 94 | 51 |

**Determine if there is any association between the listed occupations and the personality preferences at 5% level of significance.**

**Solution :**

**Chi-square Test statistic**

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left\{Obs(i,j) - \widehat{Exp}(i,j)\right\}^2}{\widehat{Exp}(i,j)}.$$

$H_0$ : Personality and occupations are not Associated
$H_A$ : Personality and occupations are Associated.

Performing chi-square test at 5 % level of significance
We get **p-value** = 0.0002031
The observed p-value is very small compared to level of significance ( **α =0.05** ) this gives a weak evidence that values are not associated.

**Exercise 3: Makers of generic drugs are required by the FDA to show that the extent of absorption of their drug in blood does not differ significantly from the "brand-name" drug that they imitate. To show this, 20 healthy nonsmoking male subjects are selected. For each subject, one of the two drugs is randomly chosen and given first. Then after a washout period, the other drug is given. In both the cases, the absorption of the drug in the blood is measured. The dataset stored in the file medicine.txt gives the measurements taken on 20 subjects. Do the drugs differ significantly in absorption?**

**Solution:**

$H_0$ Drugs do not differ significantly in absorption.
$H_A$ Drugs differ significantly in absorption.

Performing chi-square test at 5 % level of significance
We get **p-value** = 2.2e-16

The observed p-value is very small compared to level of significance ( **α =0.05** ) this gives a weak evidence that drugs differ significantly in absorption.

**Exercise 4**:

**This landmark experiment in genetics investigated whether, for a certain kind of sweet pea plant, the traits “flower color" and “pollen grain type" are inherited independently or not. Flower color can be purple (P) or red (R), but the purple color is dominant) and Grain type can be Long (L) or Round (R), but long grain type is dominant. According to Mendel's law of independent segregation, the genes for these two traits segregate independently and the “flower color" and “pollen grain type" combinations-P&L, P&R, R&L and R&R are expected in 9:3:3:1 ratio. The following are the observed frequencies for each combination when the experiment was carried out on 256 sweet pea plant.**

| “flower color" and “pollen grain type" combinations | Observed Frequencies |
|---|---|
| P&L | 177 |
| P&R | 15 |
| R&L | 15 |
| R&R | 49 |

**Write your conclusion at 5% level of significance.**

**Solution :**

$H_0$ The proportion of PL to PR to RL to RR is 9 : 3 : 3 :1.
$H_A$ The proportion of PL to PR to RL to RR is different from 9 : 3 : 3 :1

Chi-square test of goodness of fit is used to find out how the observed value of a given phenomena is different from the expected value.

Performing chi-square test at 5 % level of significance
We get **p-value** = 2.2e-16
The observed p-value is very small compared to level of significance ( **α =0.05** ) this gives a weak evidence that the proportion of PL to PR to RL to RR is different from 9 : 3 : 3 :1.

**R Code**

```
library(Rcpp)
library(raster) # to get map shape file
library(ggplot2) # for plotting and miscellaneuous things
library(ggmap) # for plotting
library(plyr) # for merging datasets
library(scales) # to get nice looking legends
library(maps)
library(mapdata)
library(httr)
jp.df <- map_data("japan")
colnames(jp.df)[5] <- "Prefecture"
jp.df$Prefecture <- tolower(jp.df$Prefecture)
jp.dat <- read.table("R/JP_GDP.csv", header = T, sep = ",")
jp.dat$Prefecture <- tolower(jp.dat$Prefecture)
jp.dat <- jp.dat[jp.dat$Year == 2007 | jp.dat$Year == 2005, c("Year", "Prefecture", "GDP_per")]
jp.df <- join(jp.df, jp.dat, by = "Prefecture", type = "inner")
jp.2007 <- jp.df[jp.df$Year == 2007,]
jp.2005 <- jp.df[jp.df$Year == 2005,]
jp.diff <- jp.2007
jp.diff$GDP_per <- jp.2007$GDP_per - jp.2005$GDP_per
brks.2007 <- seq(14, 32, by = 2)
brks.2005 <- seq(12, 28, by = 2)
brks.diff <- seq(-10, 8, by = 2)
title1 <- "Figure 2: Gross Domestic Product of top 1% earners in 2007"
title2 <- "Figure 1: Gross Domestic Product of top 1% earners in 2005"
title3 <- "Figure 3: Change in Gross Domestic Product, 2007 - 2005"
prefectures <- read.table("R/Japan_GeoData.csv", header = T, sep = ",")
p <- function(data, brks, title) {
  ggp <- ggplot() + geom_polygon(data = data, aes(x = long, y = lat, group = group,
                                    fill = GDP_per), color = "black", size = 0.15) +
scale_fill_distiller(palette = "Blues",
                                                                  breaks = brks) +
theme_nothing(legend = TRUE) + labs(title = title, fill = "") +
    geom_text(data = prefectures, aes(x = lon, y = lat, label =
abbreviate(Prefecture,minlength=2)),colour="yellow", size = 1)
  return(ggp)
}
ggsave(p(jp.2007, brks.2007, title1), height = 5, width = 6, file = "japan_map_2007.jpg")
ggsave(p(jp.2005, brks.2005, title2), height = 5, width = 6, file = "japan_map_2005.jpg")
ggsave(p(jp.diff, brks.diff, title3), height = 5, width = 6, file = "japan_map_diff.jpg")
```

**2)**

```
> x <- c(62, 45, 56, 81, 94, 51)
> xm <- matrix(x, byrow=T, ncol=2)
> xm
     [,1] [,2]
[1,]   62   45
[2,]   56   81
[3,]   94   51
> chisq.test(xm)

         Pearson's Chi-squared test

data:  xm
X-squared = 17.003, df = 2, p-value = 0.0002031
```

**3)**

```
mydata = "/home/vijay/Downloads/medicine.txt"
> ydata<-read.delim(mydata, row.names = 1)
> chisq.test(ydata)


         Pearson's Chi-squared test

data:  ydata
X-squared = 4699.4, df = 19, p-value < 2.2e-16
```

**4)**

```
> chisq.test(c(177,15,15,49),p=c(9,3,3,1)/16)

         Chi-squared test for given probabilities

data:  c(177, 15, 15, 49)
X-squared = 121, df = 3, p-value < 2.2e-16
```