Task-02

Import necessary libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(knitr)
```

Loading dataset

```
business_data <- read.csv("../data/businesses.csv")
review_data <- read.csv("../data/reviews.csv")
```

```
business_data[business_data == ""] <- NA
review_data[review_data == ""] <- NA
```

```
# Define the names of columns you will use
business_cols <- c("business_id", "state", "business_avg_stars", "review_count")

# use necessary columns
business_data <- business_data[, (colnames(business_data) %in% business_cols)]
```

```
# Define the names of columns you will use
review_cols <- c("user_id", "business_id")

# use necessary columns
review_data <- review_data[, (colnames(review_data) %in% review_cols)]
```

NA Handling

```
colSums(is.na(business_data)) # checking NA value for each columns on business_data
```

```
##        business_id              state business_avg_stars       review_count
##                  1                580                  0                  0
```

```
colSums(is.na(review_data)) # checking NA value for each columns on review_data
```

```
##     user_id business_id
##        5829        5834
```

In the *business_data*, there are missing values in certain columns, including the *business_id* column, which is crucial for deeper analysis. Therefore, I decided to remove rows that contain missing values.

In the *review_data*, there are many missing values in both columns. However, only the *business_id* column

is essential for the analysis, while the *user_id* column is only used to count the number of users per business. Thus, I filled the missing values in the *user_id* column with 'UNK' and removed the rows where the *business_id* is missing.

```
business_data = na.omit(business_data) # remove every row with NA value
```

```
colSums(is.na(business_data)) # checking NA value for each columns on business_data
```

```
##        business_id              state business_avg_stars       review_count
##                  0                  0                  0                  0
```

```
review_data$user_id[is.na(review_data$user_id)] <- "UNK" # Fill missing values in the user_id column wi
review_data = na.omit(review_data) # Remove remaining missing values in the business_id column
```

```
colSums(is.na(review_data)) # checking NA value for each columns on review_data
```

```
##     user_id business_id
##           0           0
```

Analysis

Merge business_data with review_data and calculate the number of users for each business

```
# Count the number of unique user_id values for each business_id in review_data
review_summary <- review_data %>%
  group_by(business_id) %>%
  summarise(user_count = n_distinct(user_id))

# Merge the review_summary with business_data
business_data <- business_data %>%
  left_join(review_summary, by = "business_id")

head(business_data) # display 6 rows of business_data
```

```
##   business_id state business_avg_stars review_count user_count
## 1         b_0    NV                 2.5          351          5
## 2         b_1    KY                 4.8          267         12
## 3         b_2    PA                 3.9          397         10
## 4         b_3    CA                 3.4           54         16
## 5         b_4    GA                 1.6          278         11
## 6         b_5    DC                 1.6          320          8
```

```
colSums(is.na(business_data)) # checking NA value for each columns on business_data
```

```
##        business_id              state business_avg_stars       review_count
##                  0                  0                  0                  0
##         user_count
##                  0
```

Group the data by state and calculate the average business rating, the total number of reviews for all businesses in each state, the total number of users associated with all businesses in each state, and the number of businesses in each state

```
# Grouping data by state
summary_table <- business_data %>%
  group_by(state) %>%
  summarise(
    average_review_star = mean(business_avg_stars), # calculating avg review stars for each business of
    number_of_review = sum(review_count), # calculating total of review for each business of every stat
```

```
    user_count_total = sum(user_count), # calculating total of user for each business of every state
    business_count = n(), # calculating number of business for each state
    .groups = 'drop' # Remove the grouping structure after summarise
  )

# visualize table using kable
kable(summary_table, # using summary table
      caption = "Business summary for each State",
      align = 'lcccc') # 'l' for left alignment, 'c' for center alignment.
```

Table 1: Business summary for each State

state	average_review_star	number_of_review	user_count_total	business_count
AK	3.011230	95487	3450	374
AL	2.960548	97669	3532	365
AR	2.878378	91862	3472	370
AZ	2.964524	99369	3597	389
CA	3.016011	90071	3526	381
CO	3.037736	99167	3445	371
CT	2.932869	93236	3352	359
DC	3.046615	97794	3614	384
DE	3.056111	90450	3383	360
FL	2.995568	89556	3339	361
GA	2.945161	89997	3502	372
HI	3.083668	103074	3712	398
IA	2.922955	94616	3499	379
ID	2.949014	93780	3324	355
IL	3.009375	87414	3282	352
IN	3.067823	82594	3028	317
KS	3.041547	87991	3246	349
KY	3.061600	94581	3456	375
LA	3.098667	92830	3493	375
MA	3.013490	84995	3278	341
MD	2.921607	95441	3340	361
ME	3.035556	96133	3374	360
MI	3.055041	88334	3335	367
MN	2.862670	91652	3511	367
MO	3.038931	101459	3726	393
MS	3.006989	95686	3460	372
MT	2.965013	96613	3617	383
NC	3.024384	92341	3501	365
ND	2.999747	97197	3783	395
NE	3.043465	84505	3189	329
NH	3.014286	91424	3330	357
NJ	2.982432	94931	3444	370
NM	2.980101	104148	3648	397
NV	2.943195	83237	3091	338
NY	2.914714	92117	3436	367
OH	2.991282	96702	3596	390
OK	2.950000	98213	3744	388
OR	2.955107	94528	3537	372
PA	3.097933	102410	3665	387
RI	2.899443	93876	3332	359

state	average_review_star	number_of_review	user_count_total	business_count
SC	2.911488	100514	3605	383
SD	3.081818	98699	3651	385
TN	2.989722	93455	3320	360
TX	3.093316	98212	3629	389
UT	3.010141	88482	3155	355
VA	2.829730	93294	3482	370
VT	3.032386	88576	3319	352
WA	3.017268	96017	3746	388
WI	3.003188	87708	3268	345
WV	2.982552	95741	3671	384
WY	2.941918	91879	3337	365

```
summary(summary_table) # basic statistic of summary_table
```

```
##     state           average_review_star number_of_review user_count_total
##  Length:51          Min.   :2.830       Min.   : 82594   Min.   :3028
##  Class :character   1st Qu.:2.950       1st Qu.: 90260   1st Qu.:3334
##  Mode  :character   Median :3.003       Median : 93876   Median :3460
##                     Mean   :2.993       Mean   : 93727   Mean   :3458
##                     3rd Qu.:3.038       3rd Qu.: 96950   3rd Qu.:3601
##                     Max.   :3.099       Max.   :104148   Max.   :3783
##  business_count
##  Min.   :317.0
##  1st Qu.:359.5
##  Median :370.0
##  Mean   :369.0
##  3rd Qu.:383.5
##  Max.   :398.0
```