Statistical Modelling in R

Name: Ishika Shah **Roll No.:42**

INPUT:

```
#FOOD INSPECTIONS DATASET

chicago <- read.csv(file.choose())

#1. EDA
head(chicago)
dim(chicago)
tail(chicago)

#2. First 6 records

head(chicago, 6)

#3. last 10 records

tail(chicago, 10)

#4. dimension of dataset

dim(chicago)

#5. Structure of dataset

str(chicago)
```

#6. Names of columns

```
colnames(chicago)
```

#7. Summary for numerical columns

```
sapply(chicago, function(x) is.numeric(x))

summary(chicago$Inspection.ID)
summary(chicago$License..)
summary(chicago$Zip)
summary(chicago$Latitude)
summary(chicago$Longitude)
```

#alt

```
for(i in 1:22) {
  if(is.numeric(chicago[,i])== TRUE) {
    print(colnames(chicago)[i])
    ans = summary(chicago[,i])
    print(ans)
  }
}
```

#8. display column with index number 3

```
chicago[,3]
```

#9. select the rows with license number between 20l to 30l

```
chicago[(chicago$License..>=200000) & (chicago$License..<=300000),]
```

#10. select the records with Dba line = "blueline"

```
chicago[(chicago$DBA.Name == "BlueLine"),]
```

#11. Display only inspection id and facility type

```
chicago[,c("Inspection.ID", "Facility.Type")]
```

#12. Show dba name where risk is high

```
chicago[chicago$Risk == "Risk 1 (High)", c("DBA.Name")]
data.frame("Restaurant Name" = chicago[chicago$Risk == "Risk 1 (High)", c("DBA.Name")])
```

#13. 1st 2 columns where result of inspection is fail

```
fail <- filter(chicago, chicago$Results == "Fail")
head(fail, 2)
chicago[chicago$Results == "Fail", c(1,2)]
```

#14. Total number of rows in dataframe

```
nrow(chicago)
```

#15. Number of missing values columnwise

```
missing <- sapply(chicago, function(x) sum(is.na(x)))
missing
```

#16. Delete the columns which only have null values ##

```
#No such columns

#17. Histogram of latitude and longitude columns ##

par(mfrow = c(2,2))
ggplot(chicago, aes(Latitude)) + geom_histogram()
ggplot(chicago, aes(Longitude)) + geom_histogram()

#18. Replace NA values in latitude and longitude with mean ##

chicago[is.na(chicago$Latitude) == TRUE, "Latitude"] = mean(chicago$Latitude, na.rm = TRUE)
chicago[is.na(chicago$Longitude) == TRUE, "Longitude"] = mean(chicago$Longitude, na.rm = TRUE)

#19 Print the total number of duplicated values column-wise.##

library(dplyr)
sapply(chicago, function(x) count(duplicated(x)))

#20. Display the count of PASS results grouped by type of risk.

chicago %>%
  group_by(Risk) %>%
  summarise(count(filter(chicago, chicago$Results == "Pass")))

#21. Separate the inspection date into 3 (day ,month and year) columns.

library(tidyverse)
dates = separate(chicago, Inspection.Date, into = c("Day", "Month", "Year"), sep = "/")
chicago = dates
colnames(chicago)
```