FDP – R for Data Science

Session 4: Statistics using R

Introduction

Before we attempt to describe data, we need to make sure the data is in the right format. This means

- Making sure all the data is contained in a data frame (or in a vector if it's a single variable)
- Ensuring that all the variables are of the correct type
- Checking that the values are all processed correctly

```
library(dplyr)
library(ggplot2)

path <- "C:/Users/Admin/Desktop/FDP_R/Placement_Data_FDP.csv" # copy applicable path
placement <- read.csv(path, stringsAsFactors = T)
str(placement)
colnames(placement)
placement$sl_no <- NULL # remove sl.no column
# split the dataset
placementnum <- select(placement, ends_with("_p"), salary)
placementcat <- select(placement, -(ends_with("_p")), -salary)
levels(placement$status)
unique(placement$status) # alternative
placedset <- filter(placement, status == "Placed")
placedset <- na.omit(placement) # alternative
notplacedset <- subset(placement, status == "Not Placed")

# basic exploration for NAs in the dataset
class(placement)
anyNA(placement) # check for NAs
is.na(placement)
colSums(is.na(placement))
lapply(placement, anyNA)
```

Descriptive Analytics

Descriptive analytics is about finding "what has happened" by summarizing the data using innovative methods and analysing the past data using simple queries. Analysing past data can provide insights that can assist organisations to take appropriate decisions. Primary objective of descriptive analytics is simple

comprehension of data using data summarisation, basic statistical measures and visualisation.
Statistics is basically the study of what causes variability in the data. Descriptive statistics such as measures of central tendency, measures of variation and measures of shape provide useful insights.

Measures of central tendency
Measures of central tendency are the measures that are used for describing the data using a single value. Mean, Median and Mode are the three commonly used measures to compare different data sets. Percentile, Decile and Quantile are frequently used to identify the position of the observation in the data set.

Measures of variation
While the measures of central tendency yield information about the center or middle part of a data, measures of variability describe the spread or dispersion of a set of data. Using the measures of variation in conjunction with measures of central tendency makes possible a more complete numerical description of data. Range, Inter Quartile Range, Mean Absolute Deviation, Variance and Standard Deviation are common measures of variability.

Measures of shape – Skewness and Kurtosis
Skewness is a measure of symmetry or lack of symmetry. A data set is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. A value of zero for the measure indicates the data is symmetrical.
Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis value of 3 indicates standard normal distribution. The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is obtained by subtracting the value 3 from the kurtosis value.

Note: You can refer to any standard book on Statistics for a detailed explanation to all the above measures

```
# summary statistics
# measures of central tendency
mean(placement$degree_p) # Describing center of the data
placement %>% group_by(degree_t) %>% summarise(mean(degree_p))
group_by(placement, degree_t, status) %>% summarise(mean(degree_p))
aggregate(degree_p ~ status + degree_t, placement, FUN = mean) # alternative
sapply(placementnum, mean) # add na.rm if necessary
apply(placementnum, 2, mean) # other apply family function. 1 indicates rows, 2 indicates columns
lapply(placementnum, mean)
```

```
tapply(placement$degree_p,placement$gender, mean)
median(placement$salary, na.rm = T) # Dealing with missing values
quantile(placement$salary, na.rm = T)
quantile(placement$salary, 0.25, na.rm = T) # Customized quantiles
fivenum(placement$salary, na.rm = T) # same as quantile for odd no. of observations

# The core package in R doesn't have a function for calculating the mode
lsr::modeOf(placement$salary)

# other means
colMeans(placementnum)
rowMeans(placementnum[ , -6])

# skewness and kurtosis
psych::skew(placementnum) # value close to 0 indicates data is symmetrical
colnames(placementnum)
ggplot(placementnum, aes(salary)) + geom_histogram(binwidth = 25000, na.rm = T)
ggplot(placementnum, aes(mba_p)) + geom_histogram(binwidth = 5)

psych::kurtosi(placementnum) # negative-flat, positive-pointy, zero-just enough pointy
ggplot(placementnum, aes(ssc_p)) + geom_histogram(bins = 10)
ggplot(placementnum, aes(hsc_p)) + geom_histogram(bins = 10)

# mean v/s median
mean(placement$salary, na.rm = T)
median(placement$salary, na.rm = T)
mean(placement$salary, trim = 0.10, na.rm = T) #Trimmed mean. Trimmed from each end

# measures of variation
min(placement$mba_p)
max(placement$mba_p)
range(placement$mba_p)
IQR(placement$salary, na.rm = T)
sd(placement$degree_p) # this uses denominator n-1
var(placement$degree_p) # this uses denominator n-1
mad(placement$degree_p) # median absolute deviation

# Summarizing a variable
summary(placement$mba_p)
# Summarizing a complete dataset
summary(placementnum)
summary(placement)
psych::describe(placementnum))
```

```
# Describing Categories
table(placementcat$status)
sapply(placementcat, table)
table(placementcat$specialisation, placementcat$status) # Creating a two-way table
mytab <- with(placementcat, table(specialisation, status)) # alternative
mytab
class(mytab)
addmargins(mytab)
prop.table(mytab) # proportion based on total number
prop.table(mytab, margin = 1) # proportions over rows and columns
# visualisation
df_mytab <- as.data.frame(mytab)
df_mytab
ggplot(df_mytab, aes(x=specialisation, y = Freq)) +
  geom_bar(aes(fill = status), stat = "identity")
# Note - when heights of the bars have to represent values (freq) in the data, use stat="identity" and map a value to the y aesthetic
ggplot(placement, aes(specialisation)) + geom_bar(aes(fill=status)) # directly from the file

tab3 <- xtabs(~gender+specialisation+status, placement) # 3-way crosstabs
tab3
ftable(tab3)
```

Measure of association

Correlation is a measure of the strength and direction of relationship that exists between two random variables and is measured using correlation coefficient. Correlation is only an association relationship and not a causal relationship.

```
# Tracking correlations
cor(placement$degree_p, placement$mba_p)
ggplot(placement, aes(degree_p, mba_p)) + geom_point()
corcomplete <- cor(placementnum) # correlations for multiple variables
corcomplete
corcomplete["ssc_p", "mba_p"]
placementnum %>% cor()  # alternative
# Dealing with missing values
cor(placement$mba_p, placement$salary, use="complete.obs")
cor(placementnum, use = "pairwise.complete.obs")
```