

Personal Project - Data Gaji

Agnes Septilia

23/10/2021

Here I'm practicing EDA with the data set of Data Gaji (Salary Data). Let's start with loading the library and take a look on the dataset.

```
library(readxl)
library(scales)
library(tidyverse)
library(dplyr)
library(descr)

# take a look on the dataset
salary <- read_xlsx("Data Gaji 2.xlsx")
glimpse(salary)

## Rows: 111
## Columns: 12
## $ `Masa Perolehan Awal`      <dbl> 1, 8, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1~
## $ `Masa Perolehan Akhir`     <dbl> 12, 12, 12, 12, 12, 12, 12, 4, 12, 12, 12~
## $ NIP                       <chr> "168.2-012", "280.8-484", "126.8-014"~
## $ `Jenis Kelamin`           <chr> "M", "M", "M", "M", "M", "F", "M", "M~
## $ `Status PTKP`             <chr> "TK", "TK", "K", "TK", "K", "TK", "TK~
## $ `Jumlah Tanggungan`       <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0~
## $ `Gaji Pokok dan Tunjangan Tetap` <dbl> 21316200, 20140223, 18300000, 2131620~
## $ `Tunjangan lain (Variabel)` <dbl> 10604813, 0, 12166831, 2600212, 10134~
## $ `JKK & JKM & BPJS Kesehatan` <dbl> 182661.1, 231113.6, 3318460.2, 183613~
## $ `THR dan Bonus`           <dbl> 2126320, 0, 1422000, 4114862, 2120000~
## $ `Tunjangan PPH`           <dbl> 0, 0, 38204, 0, 0, 0, 0, 102000, 8220~
## $ `Jumlah Penghasilan Bruto` <dbl> 34229994, 20371337, 35245495, 2986741~
```

Task 1 : We want to check whether any duplicate data based on NIP

```
salary %>% count(NIP)
```

```
## # A tibble: 111 x 2
##   NIP      n
##   <chr> <int>
## 1 002.1-231    1
## 2 010.2-504    1
## 3 016.4-483    1
## 4 018.1-242    1
## 5 018.8-425    1
## 6 020.8-408    1
## 7 020.8-442    1
```

```
## 8 022.2-016      1
## 9 026.4-242      1
## 10 028.1-012     1
## # ... with 101 more rows
```

Result on Task 1 : Total data of NIP is equal with total row in dataframe.

So there's no duplicate data (person) on the data.

Task 2 : Check Turnover rate

```
resign <- salary %>%
  select(`Masa Perolehan Awal`, `Masa Perolehan Akhir`) %>%
  mutate(Resigned = ifelse(`Masa Perolehan Akhir` == 12, 0, 1)) %>%
  group_by(Resigned) %>%
  count() %>%
  pull(n)

turnover_rate <- (resign[2] / sum(resign) * 100)
cat(sprintf("Turnover rate is %.0f%%", turnover_rate, '%'))

## Turnover rate is 12%
```

Task 3 : Make new column PTKP amount

Here's the basic rule of PTKP amount: - if Status PTKP = TK -> 54,000,000 - if Status PTKP = K -> 58,500,000 - then add each with Jumlah Tanggungan * 4,500,000

```
salary <- salary %>%
  mutate(PTKP_Amount = ifelse(`Status PTKP` == 'TK', (54000000 + `Jumlah
Tanggungan` * 4500000),
                                (58500000 + `Jumlah Tanggungan` * 4500000)))
%>%
  relocate(PTKP_Amount, .after = `Jumlah Tanggungan`)
glimpse(salary)

## Rows: 111
## Columns: 13
## $ `Masa Perolehan Awal`      <dbl> 1, 8, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1~
## $ `Masa Perolehan Akhir`     <dbl> 12, 12, 12, 12, 12, 12, 4, 12, 12, 12~
## $ NIP                        <chr> "168.2-012", "280.8-484", "126.8-014"~
## $ `Jenis Kelamin`           <chr> "M", "M", "M", "M", "M", "F", "M", "M~
## $ `Status PTKP`             <chr> "TK", "TK", "K", "TK", "K", "TK", "TK~
## $ `Jumlah Tanggungan`       <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0~
## $ PTKP_Amount               <dbl> 54000000, 54000000, 72000000, 5400000~
## $ `Gaji Pokok dan Tunjangan Tetap` <dbl> 21316200, 20140223, 18300000, 2131620~
## $ `Tunjangan lain (Variabel)` <dbl> 10604813, 0, 12166831, 2600212, 10134~
## $ `JKK & JKM & BPJS Kesehatan` <dbl> 182661.1, 231113.6, 3318460.2, 183613~
## $ `THR dan Bonus`          <dbl> 2126320, 0, 1422000, 4114862, 2120000~
```

```
## $ `Tunjangan PPh` <dbl> 0, 0, 38204, 0, 0, 0, 0, 102000, 8220~
## $ `Jumlah Penghasilan Bruto` <dbl> 34229994, 20371337, 35245495, 2986741~
```

Task 4 : Make new column PTKP_to_Bruto, to check whether Jumlah Penghasilan Bruto exceed PTKP or not.

The column will have value: 'Under PTKP' if Bruto <= PTKP, and 'Over PTKP' if otherwise.

Note: In real calculation, there will be element of reduction over Bruto before it was compared to PTKP. But here, we assume the reduction element is none.

```
salary <- salary %>%
  mutate(PTKP_to_Bruto = ifelse(PTKP_Amount <= `Jumlah Penghasilan Bruto`,
    'Under PTKP', 'Over PTKP'))
glimpse(salary)

## Rows: 111
## Columns: 14
## $ `Masa Perolehan Awal` <dbl> 1, 8, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1~
## $ `Masa Perolehan Akhir` <dbl> 12, 12, 12, 12, 12, 12, 4, 12, 12, 12, 12~
## $ NIP <chr> "168.2-012", "280.8-484", "126.8-014"~
## $ `Jenis Kelamin` <chr> "M", "M", "M", "M", "M", "F", "M", "M~
## $ `Status PTKP` <chr> "TK", "TK", "K", "TK", "K", "TK", "TK~
## $ `Jumlah Tanggungan` <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0~
## $ PTKP_Amount <dbl> 54000000, 54000000, 72000000, 54000000~
## $ `Gaji Pokok dan Tunjangan Tetap` <dbl> 21316200, 20140223, 18300000, 2131620~
## $ `Tunjangan lain (Variabel)` <dbl> 10604813, 0, 12166831, 2600212, 10134~
## $ `JKK & JKM & BPJS Kesehatan` <dbl> 182661.1, 231113.6, 3318460.2, 183613~
## $ `THR dan Bonus` <dbl> 2126320, 0, 1422000, 4114862, 2120000~
## $ `Tunjangan PPh` <dbl> 0, 0, 38204, 0, 0, 0, 0, 102000, 8220~
## $ `Jumlah Penghasilan Bruto` <dbl> 34229994, 20371337, 35245495, 2986741~
## $ PTKP_to_Bruto <chr> "Over PTKP", "Over PTKP", "Over PTKP"~
```

Task 5 : Single, Married, or Divorced?

Make new column called Marriage_Status' with below condition: - ifStatus PTKP== TK andJumlah Tanggungan== 0 -> Then Single - ifStatus PTKP== K -> Then Married - ifStatus PTKP== TK andJumlah Tanggungan` != 0 -> Then Divorced

Note: Of course, not all TK/1/2/3 are divorced in real life. So the category here is only for practice.

```
salary <- salary %>%
  mutate(Marriage_Status = case_when(`Status PTKP` == 'K' ~ 'Married',
    `Status PTKP` == 'TK' & `Jumlah
Tanggungan` == 0 ~ 'Single',
    TRUE ~ 'Divorced')) %>%
```

```

relocate(Marriage_Status, .after = `Jumlah Tanggungan`)
glimpse(salary)

## Rows: 111
## Columns: 15
## $ `Masa Perolehan Awal`      <dbl> 1, 8, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1~
## $ `Masa Perolehan Akhir`     <dbl> 12, 12, 12, 12, 12, 12, 4, 12, 12, 12~
## $ NIP                       <chr> "168.2-012", "280.8-484", "126.8-014"~
## $ `Jenis Kelamin`           <chr> "M", "M", "M", "M", "M", "F", "M", "M~
## $ `Status PTKP`             <chr> "TK", "TK", "K", "TK", "K", "TK", "TK~
## $ `Jumlah Tanggungan`       <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0~
## $ Marriage_Status            <chr> "Single", "Single", "Married", "Singl~
## $ PTKP_Amount               <dbl> 54000000, 54000000, 72000000, 5400000~
## $ `Gaji Pokok dan Tunjangan Tetap` <dbl> 21316200, 20140223, 18300000, 2131620~
## $ `Tunjangan lain (Variabel)` <dbl> 10604813, 0, 12166831, 2600212, 10134~
## $ `JKK & JKM & BPJS Kesehatan` <dbl> 182661.1, 231113.6, 3318460.2, 183613~
## $ `THR dan Bonus`           <dbl> 2126320, 0, 1422000, 4114862, 2120000~
## $ `Tunjangan PPh`           <dbl> 0, 0, 38204, 0, 0, 0, 0, 102000, 8220~
## $ `Jumlah Penghasilan Bruto` <dbl> 34229994, 20371337, 35245495, 2986741~
## $ PTKP_to_Bruto             <chr> "Over PTKP", "Over PTKP", "Over PTKP"~

```

Task 6 : How much Single that makes over 100jt per year ?

```

rich_single <- salary %>%
  filter (Marriage_Status == 'Single', `Jumlah Penghasilan Bruto` >
100000000) %>%
  count()
cat(sprintf("There are %d person which are single and make over 100million
per year", rich_single$n))

## There are 10 person which are single and make over 100million per year

```

Task 7 : Visualize the correlation between gender and salary using Dot Plot

gender -> using column Jenis Kelamin salary -> using column Gaji Pokok dan Tunjangan Tetap

```

salary %>%
  ggplot(aes(x=`Jenis Kelamin`,
             y = `Gaji Pokok dan Tunjangan Tetap`/1000000,
             color = `Jenis Kelamin`)) +
  geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 0.5) +
  labs(title = "Salary distribution based on gender",
       y = "Fixed Salary in million IDR",
       x = "Gender",
       col = "Gender")

```



```
## M               1      25      21      47
##               0.021    0.532    0.447    0.423
##               0.009    0.225    0.189
## -----
## Total           1      25      85     111
## =====
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## -----
## Chi^2 = 46.23379      d.f. = 2      p = 9.13e-11
```

Result on Task 8 : p value is less than 0.05 (alpha), so there is correlation between Gender and Marriage Status.