Workshop 2.2

Veerasak Kritsanapraphan

12/18/2019

## HR Analysis

library(ggplot2)  
library(tidyr)  
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(purrr)  
hremployee <- read.csv("WA\_Fn-UseC\_-HR-Employee-Attrition.csv",   
 header=TRUE)  
hremployee$Education = as.factor(hremployee$Education)  
hremployee$EnvironmentSatisfaction = as.factor(hremployee$EnvironmentSatisfaction)  
hremployee$JobInvolvement = as.factor(hremployee$JobInvolvement)  
hremployee$JobLevel = as.factor(hremployee$JobLevel)  
hremployee$JobSatisfaction = as.factor(hremployee$JobSatisfaction)  
hremployee$RelationshipSatisfaction = as.factor(hremployee$RelationshipSatisfaction)  
hremployee$StockOptionLevel = as.factor(hremployee$StockOptionLevel)  
hremployee$WorkLifeBalance = as.factor(hremployee$WorkLifeBalance)  
  
str(hremployee)

## 'data.frame': 1470 obs. of 35 variables:  
## $ Age : int 41 49 37 33 27 32 59 30 38 36 ...  
## $ Attrition : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...  
## $ BusinessTravel : Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 3 2 3 2 3 2 3 3 2 3 ...  
## $ DailyRate : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...  
## $ Department : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...  
## $ DistanceFromHome : int 1 8 2 3 2 2 3 24 23 27 ...  
## $ Education : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 4 1 2 3 1 3 3 ...  
## $ EducationField : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...  
## $ EmployeeCount : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ EmployeeNumber : int 1 2 4 5 7 8 10 11 12 13 ...  
## $ EnvironmentSatisfaction : Factor w/ 4 levels "1","2","3","4": 2 3 4 4 1 4 3 4 4 3 ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...  
## $ HourlyRate : int 94 61 92 56 40 79 81 67 44 94 ...  
## $ JobInvolvement : Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3 4 3 2 3 ...  
## $ JobLevel : Factor w/ 5 levels "1","2","3","4",..: 2 2 1 1 1 1 1 1 3 2 ...  
## $ JobRole : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1 ...  
## $ JobSatisfaction : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3 ...  
## $ MaritalStatus : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...  
## $ MonthlyIncome : int 5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...  
## $ MonthlyRate : int 19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...  
## $ NumCompaniesWorked : int 8 1 6 1 9 0 4 1 0 6 ...  
## $ Over18 : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ OverTime : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...  
## $ PercentSalaryHike : int 11 23 15 11 12 13 20 22 21 13 ...  
## $ PerformanceRating : int 3 4 3 3 3 3 4 4 4 3 ...  
## $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 4 2 3 4 3 1 2 2 2 ...  
## $ StandardHours : int 80 80 80 80 80 80 80 80 80 80 ...  
## $ StockOptionLevel : Factor w/ 4 levels "0","1","2","3": 1 2 1 1 2 1 4 2 1 3 ...  
## $ TotalWorkingYears : int 8 10 7 8 6 8 12 1 10 17 ...  
## $ TrainingTimesLastYear : int 0 3 3 3 3 2 3 2 2 3 ...  
## $ WorkLifeBalance : Factor w/ 4 levels "1","2","3","4": 1 3 3 3 3 2 2 3 3 2 ...  
## $ YearsAtCompany : int 6 10 0 8 2 7 1 1 9 7 ...  
## $ YearsInCurrentRole : int 4 7 0 7 2 7 0 0 7 7 ...  
## $ YearsSinceLastPromotion : int 0 1 0 3 2 3 0 0 1 7 ...  
## $ YearsWithCurrManager : int 5 7 0 0 2 6 0 0 8 7 ...

table(hremployee$Attrition)

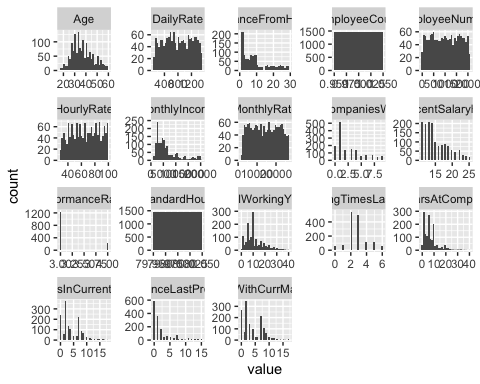
##   
## No Yes   
## 1233 237

## Exploratory Data Analysis

### Plot Single Variable - Histogram

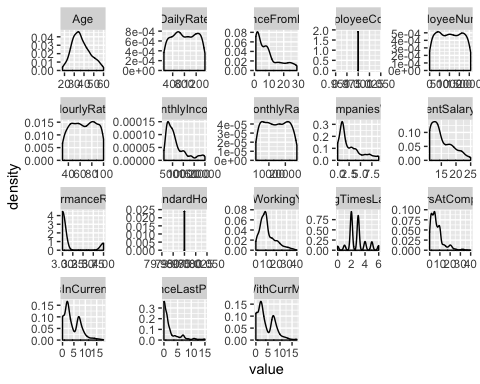
hremployee %>%  
 keep(is.numeric) %>% # Keep only numeric columns  
 gather() %>% # Convert to key-value pairs  
 ggplot(aes(value)) + # Plot the values  
 facet\_wrap(~ key, scales = "free") + # In separate panels  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



### Plot Single Variable - Density Curve

hremployee %>%  
 keep(is.numeric) %>% # Keep only numeric columns  
 gather() %>% # Convert to key-value pairs  
 ggplot(aes(value)) + # Plot the values  
 facet\_wrap(~ key, scales = "free") + # In separate panels  
 geom\_density()



### Plot Single Variable - Bar Chart

hremployee %>%  
 keep(is.factor) %>% # Keep only numeric columns  
 gather() %>% # Convert to key-value pairs  
 ggplot(aes(value)) + # Plot the values  
 facet\_wrap(~ key, scales = "free") + # In separate panels  
 geom\_bar()

## Warning: attributes are not identical across measure variables;  
## they will be dropped

