MainScript.R

chirag

Thu Mar 03 12:05:39 2016

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('stringr')  
library(ggplot2)  
library(tidyr)  
rm(list = ls())  
setwd('C:/Users/chirag/Desktop/Kaggle/sf-salaries-release-2015-12-21-03-21-32/output')  
list.files()

## [1] "database.sqlite" "group.csv"   
## [3] "hashes.txt" "JobTitlePaySummary.csv"  
## [5] "MainScript.html" "MainScript.R"   
## [7] "MainScript.spin.R" "MainScript.spin.Rmd"   
## [9] "Salaries.csv" "SingleCount.csv"

#Need to check how to figure out the na string  
  
x <- read.csv('Salaries.csv',na=c('Not Provided',''),stringsAsFactors = FALSE)  
  
str(x)

## 'data.frame': 148654 obs. of 13 variables:  
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ EmployeeName : chr "NATHANIEL FORD" "GARY JIMENEZ" "ALBERT PARDINI" "CHRISTOPHER CHONG" ...  
## $ JobTitle : chr "GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY" "CAPTAIN III (POLICE DEPARTMENT)" "CAPTAIN III (POLICE DEPARTMENT)" "WIRE ROPE CABLE MAINTENANCE MECHANIC" ...  
## $ BasePay : num 167411 155966 212739 77916 134402 ...  
## $ OvertimePay : num 0 245132 106088 56121 9737 ...  
## $ OtherPay : num 400184 137811 16453 198307 182235 ...  
## $ Benefits : num NA NA NA NA NA NA NA NA NA NA ...  
## $ TotalPay : num 567595 538909 335280 332344 326373 ...  
## $ TotalPayBenefits: num 567595 538909 335280 332344 326373 ...  
## $ Year : int 2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...  
## $ Notes : logi NA NA NA NA NA NA ...  
## $ Agency : chr "San Francisco" "San Francisco" "San Francisco" "San Francisco" ...  
## $ Status : chr NA NA NA NA ...

#change names to lower case  
x$EmployeeName <- str\_to\_lower(x$EmployeeName)  
x$JobTitle <- str\_to\_lower(x$JobTitle)  
  
#Findout the non numeric columns  
non\_numeric\_columns <- names(x)[!sapply(x, is.numeric)]  
  
#Get the summary of numeric column  
  
x %>% select(-one\_of(non\_numeric\_columns)) %>% summary()

## Id BasePay OvertimePay OtherPay   
## Min. : 1 Min. : -166 Min. : -0.01 Min. : -7058.6   
## 1st Qu.: 37164 1st Qu.: 33588 1st Qu.: 0.00 1st Qu.: 0.0   
## Median : 74328 Median : 65007 Median : 0.00 Median : 811.3   
## Mean : 74328 Mean : 66325 Mean : 5066.06 Mean : 3648.8   
## 3rd Qu.:111491 3rd Qu.: 94691 3rd Qu.: 4658.18 3rd Qu.: 4236.1   
## Max. :148654 Max. :319275 Max. :245131.88 Max. :400184.2   
## NA's :609 NA's :4 NA's :4   
## Benefits TotalPay TotalPayBenefits Year   
## Min. : -33.89 Min. : -618.1 Min. : -618.1 Min. :2011   
## 1st Qu.:11535.40 1st Qu.: 36169.0 1st Qu.: 44065.7 1st Qu.:2012   
## Median :28628.62 Median : 71426.6 Median : 92404.1 Median :2013   
## Mean :25007.89 Mean : 74768.3 Mean : 93692.6 Mean :2013   
## 3rd Qu.:35566.86 3rd Qu.:105839.1 3rd Qu.:132876.5 3rd Qu.:2014   
## Max. :96570.66 Max. :567595.4 Max. :567595.4 Max. :2014   
## NA's :36163

#Get unique values in each numeric columns  
x %>% select(-one\_of(non\_numeric\_columns)) %>% summarise\_each(funs(length(unique(.))))

## Id BasePay OvertimePay OtherPay Benefits TotalPay TotalPayBenefits  
## 1 148654 109490 65999 83226 98466 138486 142098  
## Year  
## 1 4

#Get na values in each columns  
x %>% select(-one\_of(non\_numeric\_columns)) %>% summarise\_each(funs(sum(is.na(.))))

## Id BasePay OvertimePay OtherPay Benefits TotalPay TotalPayBenefits Year  
## 1 0 609 4 4 36163 0 0 0

#Get unique values in each non numeric columns  
x %>% select(one\_of(non\_numeric\_columns)) %>% summarise\_each(funs(length(unique(.))))

## EmployeeName JobTitle Notes Agency Status  
## 1 80460 1637 1 1 3

#removing the Notes and Agency columns as they just have a single value  
x <- x %>% select(-Notes,-Agency)  
  
str(x)

## 'data.frame': 148654 obs. of 11 variables:  
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ EmployeeName : chr "nathaniel ford" "gary jimenez" "albert pardini" "christopher chong" ...  
## $ JobTitle : chr "general manager-metropolitan transit authority" "captain iii (police department)" "captain iii (police department)" "wire rope cable maintenance mechanic" ...  
## $ BasePay : num 167411 155966 212739 77916 134402 ...  
## $ OvertimePay : num 0 245132 106088 56121 9737 ...  
## $ OtherPay : num 400184 137811 16453 198307 182235 ...  
## $ Benefits : num NA NA NA NA NA NA NA NA NA NA ...  
## $ TotalPay : num 567595 538909 335280 332344 326373 ...  
## $ TotalPayBenefits: num 567595 538909 335280 332344 326373 ...  
## $ Year : int 2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...  
## $ Status : chr NA NA NA NA ...

#Get na values in each columns  
x %>% summarise\_each(funs(sum(is.na(.))))

## Id EmployeeName JobTitle BasePay OvertimePay OtherPay Benefits TotalPay  
## 1 0 2 0 609 4 4 36163 0  
## TotalPayBenefits Year Status  
## 1 0 0 110535

x %>% filter(!is.na(Status)) %>% group\_by(Year) %>% summarise(mean(TotalPay))

## Source: local data frame [1 x 2]  
##   
## Year mean(TotalPay)  
## (int) (dbl)  
## 1 2014 75471.84

unique((x %>% filter(!is.na(Status)))$Year)

## [1] 2014

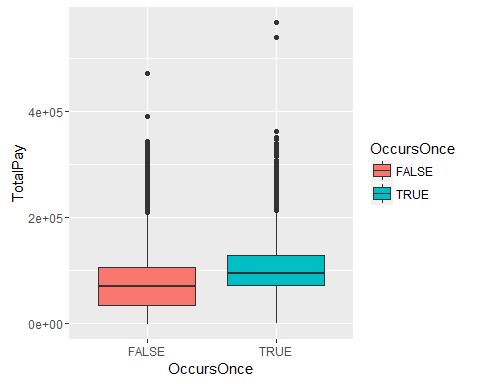
#Studying Job titles  
length(unique(x$JobTitle))

## [1] 1637

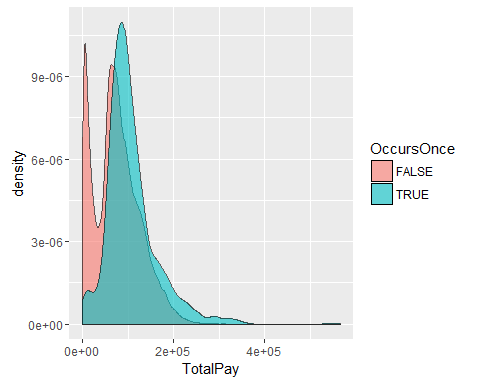
x %>%   
 group\_by(JobTitle) %>%   
 summarise(freq=n()) %>%   
 mutate(OccursOnce = freq==1) %>%  
 group\_by(OccursOnce) %>%   
 summarise(count=n())

## Source: local data frame [2 x 2]  
##   
## OccursOnce count  
## (lgl) (int)  
## 1 FALSE 1487  
## 2 TRUE 150

#plot the totalPay for unique as well as duplicated job titles  
x %>%  
 mutate(OccursOnce=!duplicated(JobTitle)) %>%  
 ggplot(aes(x=OccursOnce,y=TotalPay)) + geom\_boxplot(aes(fill=OccursOnce))



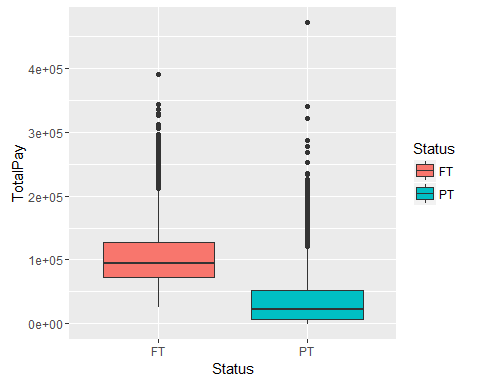
x %>%  
 mutate(OccursOnce=!duplicated(JobTitle)) %>%  
 ggplot(aes(x=TotalPay)) + geom\_density(aes(fill=OccursOnce),alpha=0.6)



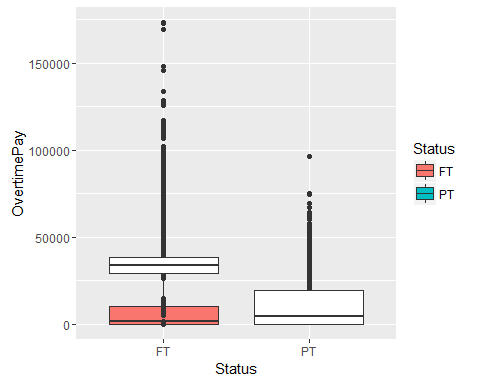
#understanding for what years we have the values of status variable  
x %>%  
 group\_by(Status,Year) %>%  
 summarise(Frequency = n()) %>%  
 spread(Year,Frequency)

## Source: local data frame [3 x 5]  
## Groups: Status [3]  
##   
## Status 2011 2012 2013 2014  
## (chr) (int) (int) (int) (int)  
## 1 FT NA NA NA 22334  
## 2 PT NA NA NA 15785  
## 3 NA 36159 36766 37606 4

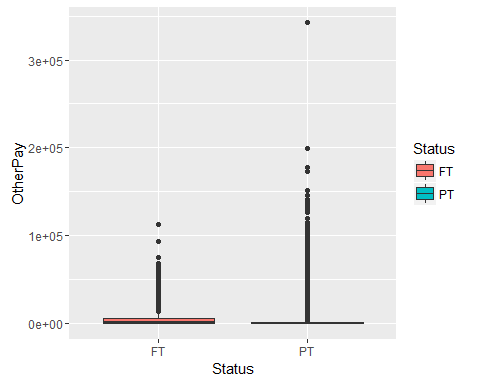
#analysing total pay of FT and PT employees  
x %>%   
 filter(!is.na(Status)) %>%   
 ggplot(aes(x=Status,y=TotalPay)) +geom\_boxplot(aes(fill=Status))



#analysing other benifits for FT and Pt employee  
x %>%   
 filter(!is.na(Status)) %>%   
 ggplot(aes(x=Status)) +geom\_boxplot(aes(y=OvertimePay,fill=Status)) +  
 geom\_boxplot(aes(y=Benefits))



#  
x %>%   
 filter(!is.na(Status)) %>%   
 ggplot(aes(x=Status)) +geom\_boxplot(aes(y=OtherPay,fill=Status))



x %>%  
 group\_by(Status,Year) %>%  
 summarise(Frequency = n()) %>%  
 spread(Year,Frequency)

## Source: local data frame [3 x 5]  
## Groups: Status [3]  
##   
## Status 2011 2012 2013 2014  
## (chr) (int) (int) (int) (int)  
## 1 FT NA NA NA 22334  
## 2 PT NA NA NA 15785  
## 3 NA 36159 36766 37606 4