BIN381\_Project\_Milestone 1\_Markdown

Group F

2024-10-05

## R Markdown

## Data Quality

# Read the dataset into the dataframe "customers"  
customers <- read.csv("CustData2.csv")  
  
  
# Missing Values  
sum(is.na(customers$Column1))

## [1] 0

sum(customers$Last.Name=="")

## [1] 6

sum(customers$First.Name=="")

## [1] 6

sum(customers$Middle.Initial=="")

## [1] 59056

sum(customers$Title=="")

## [1] 6

sum(customers$Department.Name=="")

## [1] 6

sum(is.na(customers$Annual.Salary))

## [1] 6

sum(is.na(customers$Gross.Pay.Last.Paycheck))

## [1] 6

sum(is.na(customers$Gross.Year.To.Date))

## [1] 6

sum(is.na(customers$Gross.Year.To.Date...FRS.Contribution))

## [1] 6

sum(is.na(customers$year\_of\_birth))

## [1] 0

sum(customers$marital\_status=="")

## [1] 60795

sum(customers$street\_address=="")

## [1] 0

sum(is.na(customers$postal\_code))

## [1] 0

sum(customers$city=="")

## [1] 0

sum(customers$State=="")

## [1] 0

sum(customers$Province=="")

## [1] 120613

sum(is.na(customers$Country\_id))

## [1] 0

sum(customers$phone\_number=="")

## [1] 0

sum(customers$email=="")

## [1] 0

sum(customers$Education=="")

## [1] 0

sum(customers$Occupation=="")

## [1] 0

sum(is.na(customers$household\_size))

## [1] 0

sum(is.na(customers$yrs\_residence))

## [1] 0

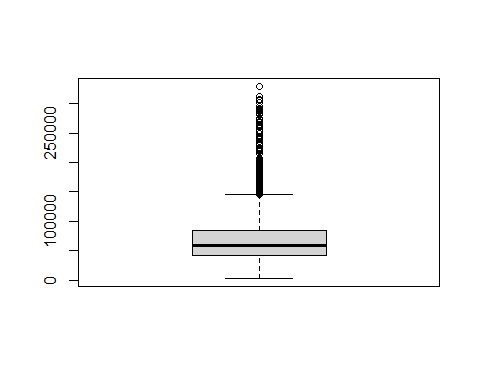
# Identify duplicate rows / records and insert into the dataframe "duplicated\_rows"  
duplicated\_rows <- customers[duplicated(customers), ]  
# Display "duplicated\_rows" and all its records  
print(duplicated\_rows)

## [1] Column1 Last.Name   
## [3] First.Name Middle.Initial   
## [5] Title Department.Name   
## [7] Annual.Salary Gross.Pay.Last.Paycheck   
## [9] Gross.Year.To.Date Gross.Year.To.Date...FRS.Contribution  
## [11] year\_of\_birth marital\_status   
## [13] street\_address postal\_code   
## [15] city State   
## [17] Province Country\_id   
## [19] phone\_number email   
## [21] Education Occupation   
## [23] household\_size yrs\_residence   
## <0 rows> (or 0-length row.names)

# Count the number of duplicate rows  
sum(duplicated\_rows)

## [1] 0

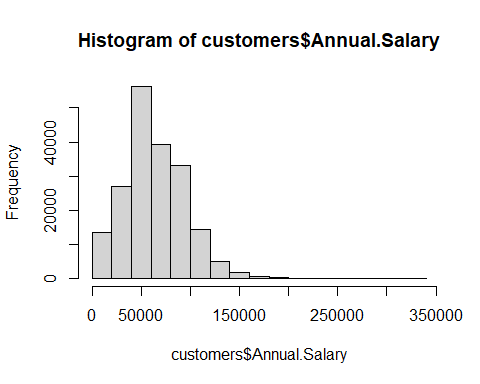
# Outliers shown in box plot of the "Annual.Salary"  
boxplot(customers$Annual.Salary)



# Variable Type Validation  
str(customers)

## 'data.frame': 191323 obs. of 24 variables:  
## $ Column1 : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Last.Name : chr "ALBERT" "ARGUELLO" "TUCKER" "DELL" ...  
## $ First.Name : chr "JESSICA" "ADRIAN" "KEVIN" "JAMES" ...  
## $ Middle.Initial : chr "M" "A" "K" "A" ...  
## $ Title : chr "CORRECTIONAL OFFICER" "POLICE OFFICER" "CORRECTIONAL OFFICER" "WASTE SCALE OPERATOR" ...  
## $ Department.Name : chr "CORRECTIONS & REHABILITATION" "POLICE" "CORRECTIONS & REHABILITATION" "SOLID WASTE MANAGEMENT" ...  
## $ Annual.Salary : num 54620 65250 62394 37735 64386 ...  
## $ Gross.Pay.Last.Paycheck : num 2502 3468 4514 1562 6666 ...  
## $ Gross.Year.To.Date : num 48025 57932 49968 35470 132851 ...  
## $ Gross.Year.To.Date...FRS.Contribution: num 46617 56223 48501 34433 128949 ...  
## $ year\_of\_birth : int 1976 1964 1942 1977 1949 1950 1946 1978 1949 1951 ...  
## $ marital\_status : chr "married" "" "single" "married" ...  
## $ street\_address : chr "27 North Sagadahoc Boulevard" "37 West Geneva Street" "47 Toa Alta Road" "47 South Kanabec Road" ...  
## $ postal\_code : int 60332 55406 34077 72996 67644 83786 52773 37400 71349 55056 ...  
## $ city : chr "Ede" "Hoofddorp" "Schimmert" "Scheveningen" ...  
## $ State : chr "Gelderland" "Noord" "Limburg" "Zuid" ...  
## $ Province : chr "" "Holland" "" "Holland" ...  
## $ Country\_id : int 52770 52770 52770 52770 52775 52782 52775 52782 52770 52789 ...  
## $ phone\_number : chr "519-236-6123" "327-194-5008" "288-613-9676" "222-269-1259" ...  
## $ email : chr "Ruddy@company.com" "Ruddy@company.com" "Ruddy@company.com" "Ruddy@company.com" ...  
## $ Education : chr "Masters" "Masters" "Masters" "Masters" ...  
## $ Occupation : chr "Prof." "Prof." "Prof." "Prof." ...  
## $ household\_size : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ yrs\_residence : int 4 4 4 4 4 4 4 4 4 4 ...

# Data Distribution  
hist(customers$Annual.Salary)



# Unique records within attributes that contain "characters"  
length(unique(customers$Last.Name))

## [1] 10917

length(unique(customers$First.Name))

## [1] 7235

length(unique(customers$Middle.Initial))

## [1] 27

length(unique(customers$Title))

## [1] 2291

length(unique(customers$Department.Name))

## [1] 43

length(unique(customers$marital\_status))

## [1] 12

length(unique(customers$street\_address))

## [1] 50945

length(unique(customers$city))

## [1] 614

length(unique(customers$State))

## [1] 142

length(unique(customers$Province))

## [1] 31

length(unique(customers$phone\_number))

## [1] 51000

length(unique(customers$email))

## [1] 1699

length(unique(customers$Education))

## [1] 3

length(unique(customers$Occupation))

## [1] 4

# Display the unique values of each attribute  
unique(customers$Department.Name)

## [1] "CORRECTIONS & REHABILITATION"   
## [2] "POLICE"   
## [3] "SOLID WASTE MANAGEMENT"   
## [4] "TRANSPORTATION AND PUBLIC WORKS"   
## [5] "WATER AND SEWER"   
## [6] "SEAPORT"   
## [7] "PARKS, RECREATION AND OPEN SPACES"   
## [8] "COMMUNITY ACTION AND HUMAN SERVICES"   
## [9] "INTERNAL SERVICES"   
## [10] "AVIATION"   
## [11] "OFFICE OF THE MAYOR"   
## [12] "CAREERSOURCE SOUTH FLORIDA"   
## [13] "FINANCE"   
## [14] "TRANSPORTATION PLANNING ORGANIZATION"   
## [15] "FIRE RESCUE"   
## [16] "PROPERTY APPRAISER"   
## [17] "CLERK OF COURTS"   
## [18] "CULTURAL AFFAIRS"   
## [19] "COMMUNICATIONS DEPARTMENT"   
## [20] "ANIMAL SERVICES"   
## [21] "JUVENILE SERVICES"   
## [22] "STATE ATTORNEY OFFICE"   
## [23] "INFORMATION TECHNOLOGY DEPARTMENT"   
## [24] "REGULATORY AND ECONOMIC RESOURCES"   
## [25] "INSPECTOR GENERAL"   
## [26] "LIBRARY"   
## [27] ""   
## [28] "MEDICAL EXAMINER"   
## [29] "PUBLIC HOUSING AND COMMUNITY DEVELOPMENT"  
## [30] "JUDICIAL ADMINISTRATION"   
## [31] "LEGAL AID"   
## [32] "HUMAN RESOURCES"   
## [33] "COMMISSION ON ETHICS & PUBLIC TRUST"   
## [34] "MIAMI-DADE ECONOMIC ADVOCACY TRUST"   
## [35] "MANAGEMENT AND BUDGET"   
## [36] "HOMELESS TRUST"   
## [37] "BOARD OF COUNTY COMMISSIONERS"   
## [38] "ELECTIONS"   
## [39] "COUNTY ATTORNEY"   
## [40] "AUDIT AND MANAGEMENT SERVICES"   
## [41] "CITIZENS' INDEPENDENT TRANSPORTION TRUST"  
## [42] "LAW LIBRARY"   
## [43] "PUBLIC HEALTH TRUST SUPPORT"

unique(customers$marital\_status)

## [1] "married" "" "single" "divorced" "widow" "Divorc."   
## [7] "NeverM" "Married" "Separ." "Mabsent" "Widowed" "Mar-AF"

unique(customers$Province)

## [1] "" "Holland"   
## [3] "Greater Manchester" "West Midlands"   
## [5] "Wuerttemberg" "Westfalen"   
## [7] "Roussillon" "County Antr"   
## [9] "Brabant" "West Yorkshire"   
## [11] "Oxfordshire" "de France"   
## [13] "South Glamorgan" "Avon"   
## [15] "Norfolk" "Greater London"   
## [17] "MI" "Alpes Cote d'Azur"   
## [19] "Alpes" "Pfalz"   
## [21] "Holstein" "Vorpommern"   
## [23] "VT" "Pyrenees"   
## [25] "NJ" "Anhalt"   
## [27] "Highlands and Islands" "Languedoc-Roussillon"   
## [29] "MN" "Provence-Alpes-Cote d'Azur"  
## [31] "IL"

unique(customers$Education)

## [1] "Masters" "Bach." "HS-grad"

unique(customers$Occupation)

## [1] "Prof." "Sales" "Cleric." "Exec."

## Data Visualisation

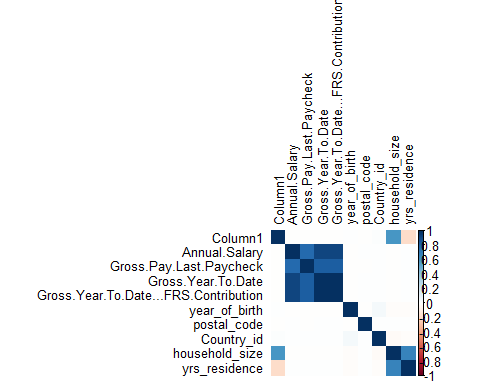
library(ggplot2)  
library(reshape2)  
library(corrplot)

## corrplot 0.94 loaded

custData <- read.csv('CustData2.csv')  
  
str(custData)

## 'data.frame': 191323 obs. of 24 variables:  
## $ Column1 : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Last.Name : chr "ALBERT" "ARGUELLO" "TUCKER" "DELL" ...  
## $ First.Name : chr "JESSICA" "ADRIAN" "KEVIN" "JAMES" ...  
## $ Middle.Initial : chr "M" "A" "K" "A" ...  
## $ Title : chr "CORRECTIONAL OFFICER" "POLICE OFFICER" "CORRECTIONAL OFFICER" "WASTE SCALE OPERATOR" ...  
## $ Department.Name : chr "CORRECTIONS & REHABILITATION" "POLICE" "CORRECTIONS & REHABILITATION" "SOLID WASTE MANAGEMENT" ...  
## $ Annual.Salary : num 54620 65250 62394 37735 64386 ...  
## $ Gross.Pay.Last.Paycheck : num 2502 3468 4514 1562 6666 ...  
## $ Gross.Year.To.Date : num 48025 57932 49968 35470 132851 ...  
## $ Gross.Year.To.Date...FRS.Contribution: num 46617 56223 48501 34433 128949 ...  
## $ year\_of\_birth : int 1976 1964 1942 1977 1949 1950 1946 1978 1949 1951 ...  
## $ marital\_status : chr "married" "" "single" "married" ...  
## $ street\_address : chr "27 North Sagadahoc Boulevard" "37 West Geneva Street" "47 Toa Alta Road" "47 South Kanabec Road" ...  
## $ postal\_code : int 60332 55406 34077 72996 67644 83786 52773 37400 71349 55056 ...  
## $ city : chr "Ede" "Hoofddorp" "Schimmert" "Scheveningen" ...  
## $ State : chr "Gelderland" "Noord" "Limburg" "Zuid" ...  
## $ Province : chr "" "Holland" "" "Holland" ...  
## $ Country\_id : int 52770 52770 52770 52770 52775 52782 52775 52782 52770 52789 ...  
## $ phone\_number : chr "519-236-6123" "327-194-5008" "288-613-9676" "222-269-1259" ...  
## $ email : chr "Ruddy@company.com" "Ruddy@company.com" "Ruddy@company.com" "Ruddy@company.com" ...  
## $ Education : chr "Masters" "Masters" "Masters" "Masters" ...  
## $ Occupation : chr "Prof." "Prof." "Prof." "Prof." ...  
## $ household\_size : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ yrs\_residence : int 4 4 4 4 4 4 4 4 4 4 ...

#These columns are numeric:  
numericColumns <- custData[,sapply(custData, is.numeric)]  
################################################  
#Correlation Matrix:  
corMatrix <- cor(numericColumns, use = 'complete.obs')  
corrplot(corMatrix, method = 'color', tl.col = 'black', tl.cex = 0.8)



################################################  
#Pair Plots:  
  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(ggplot2)  
  
selectedData <- custData[, c("Annual.Salary", "year\_of\_birth", "Gross.Year.To.Date...FRS.Contribution", "Gross.Year.To.Date")]  
  
ggpairs(selectedData,   
 title = "Pair Plots:",  
 lower = list(continuous = "smooth"),   
 diag = list(continuous = "densityDiag"))

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_density()`).

## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 6 rows containing missing values  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 6 rows containing missing values  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 6 rows containing missing values

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 6 rows containing missing values  
## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 6 rows containing missing values

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_density()`).

## Warning in ggally\_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 6 rows containing missing values

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

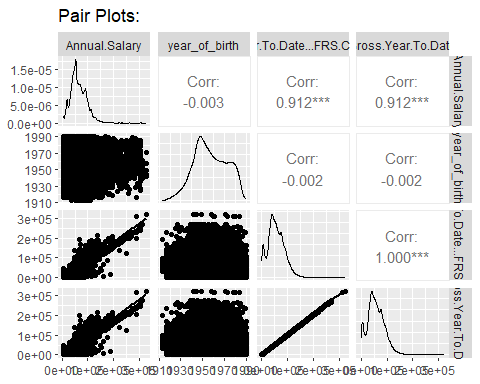
## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_density()`).

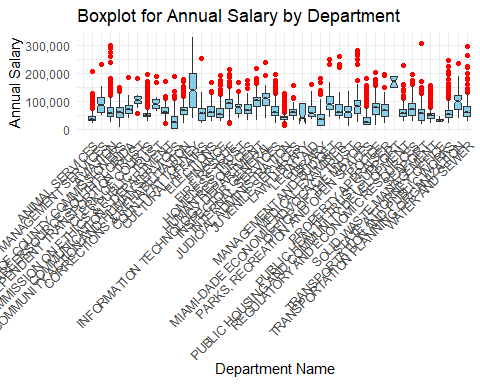


###############################################  
#Boxplots  
  
library(ggplot2)  
library(scales)  
  
ggplot(custData, aes(x = Department.Name, y = Annual.Salary)) +  
 geom\_boxplot(fill = 'skyblue', outlier.colour = 'red', outlier.shape = 16, notch = TRUE) +  
 labs(title = "Boxplot for Annual Salary by Department", x = 'Department Name', y = 'Annual Salary') +  
 scale\_y\_continuous(labels = comma) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).

## Notch went outside hinges  
## ℹ Do you want `notch = FALSE`?

## Notch went outside hinges  
## ℹ Do you want `notch = FALSE`?

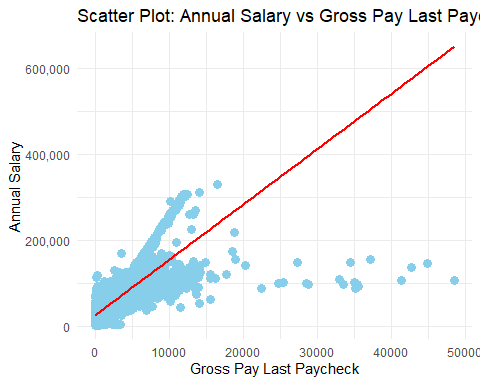


###############################################  
#Scatter Plots  
  
library(ggplot2)  
library(scales)  
  
# Scatter plot for Annual Salary vs Gross Pay Last Paycheck  
ggplot(custData, aes(x = Gross.Pay.Last.Paycheck, y = Annual.Salary)) +  
 geom\_point(color = "skyblue", size = 3, alpha = 0.6) +  
 geom\_smooth(method = "lm", color = "red", se = FALSE) +  
 labs(title = "Scatter Plot: Annual Salary vs Gross Pay Last Paycheck",   
 x = "Gross Pay Last Paycheck",   
 y = "Annual Salary") +  
 scale\_y\_continuous(labels = comma) +   
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

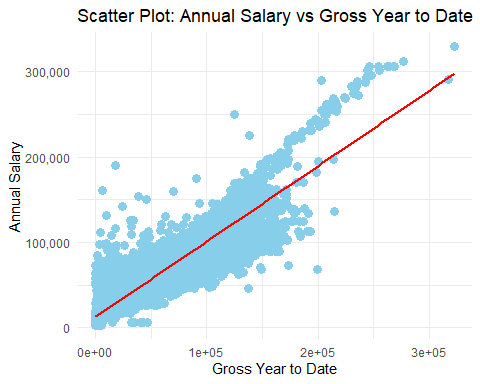
## Warning: Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



library(ggplot2)  
library(scales)  
# Scatter plot for Annual Salary vs Gross Year to Date  
ggplot(custData, aes(x = Gross.Year.To.Date, y = Annual.Salary)) +  
 geom\_point(color = "skyblue", size = 3, alpha = 0.6) +  
 geom\_smooth(method = "lm", color = "red", se = FALSE) +  
 labs(title = "Scatter Plot: Annual Salary vs Gross Year to Date",   
 x = "Gross Year to Date",   
 y = "Annual Salary") +  
 scale\_y\_continuous(labels = comma) +   
 theme\_minimal()

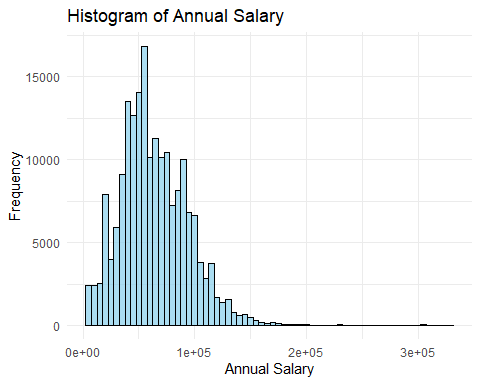
## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 6 rows containing non-finite outside the scale range (`stat\_smooth()`).  
## Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).

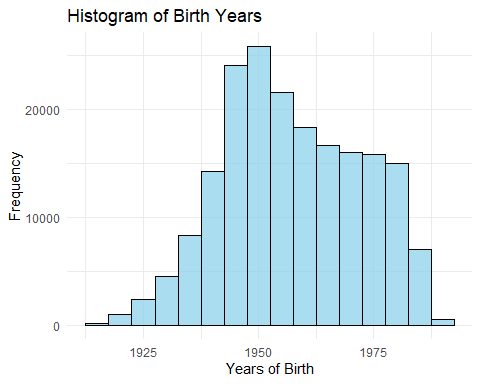


################################################  
#Histogram  
  
#Histogram for Frequency of Annual Salary  
ggplot(custData, aes(x = Annual.Salary)) +  
 geom\_histogram(binwidth = 5000, fill = "skyblue", color = "black", alpha = 0.7) +  
 labs(title = "Histogram of Annual Salary",   
 x = "Annual Salary",   
 y = "Frequency") +  
 theme\_minimal()

## Warning: Removed 6 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



#Histogram for Distribution of Birth Years  
ggplot(custData, aes(x = year\_of\_birth)) +  
 geom\_histogram(binwidth = 5, fill = "skyblue", color = "black", alpha = 0.7) +  
 labs(title = "Histogram of Birth Years",   
 x = "Years of Birth",   
 y = "Frequency") +  
 theme\_minimal()



.