## TASK 1: Data Quality Assessment

### Assessment of data quality and completeness in preparation for analysis

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(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(readxl)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.6 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.4 ✓ forcats 0.5.1  
## ✓ readr 2.1.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x lubridate::as.difftime() masks base::as.difftime()  
## x lubridate::date() masks base::date()  
## x dplyr::filter() masks stats::filter()  
## x lubridate::intersect() masks base::intersect()  
## x dplyr::lag() masks stats::lag()  
## x lubridate::setdiff() masks base::setdiff()  
## x lubridate::union() masks base::union()

getwd()

## [1] "/Users/srushtidesai/Desktop"

#Reading the each file separately  
df1<-read\_excel("KPMG\_VI/KPMG\_VI\_New\_raw\_data\_update\_final.xlsx",sheet="Transactions")  
df2<-read\_excel("KPMG\_VI/KPMG\_VI\_New\_raw\_data\_update\_final.xlsx",sheet="NewCustomerList")

## New names:  
## \* `` -> ...17  
## \* `` -> ...18  
## \* `` -> ...19  
## \* `` -> ...20  
## \* `` -> ...21

df3<-read\_excel("KPMG\_VI/KPMG\_VI\_New\_raw\_data\_update\_final.xlsx",sheet="CustomerDemographic")  
df4<-read\_excel("KPMG\_VI/KPMG\_VI\_New\_raw\_data\_update\_final.xlsx",sheet="CustomerAddress")

### Exploring Transactions Dataset

head(df1)

## # A tibble: 6 × 13  
## transaction\_id product\_id customer\_id transaction\_date online\_order  
## <dbl> <dbl> <dbl> <dttm> <lgl>   
## 1 1 2 2950 2017-02-25 00:00:00 FALSE   
## 2 2 3 3120 2017-05-21 00:00:00 TRUE   
## 3 3 37 402 2017-10-16 00:00:00 FALSE   
## 4 4 88 3135 2017-08-31 00:00:00 FALSE   
## 5 5 78 787 2017-10-01 00:00:00 TRUE   
## 6 6 25 2339 2017-03-08 00:00:00 TRUE   
## # … with 8 more variables: order\_status <chr>, brand <chr>, product\_line <chr>,  
## # product\_class <chr>, product\_size <chr>, list\_price <dbl>,  
## # standard\_cost <dbl>, product\_first\_sold\_date <dbl>

colnames(df1)

## [1] "transaction\_id" "product\_id"   
## [3] "customer\_id" "transaction\_date"   
## [5] "online\_order" "order\_status"   
## [7] "brand" "product\_line"   
## [9] "product\_class" "product\_size"   
## [11] "list\_price" "standard\_cost"   
## [13] "product\_first\_sold\_date"

glimpse(df1)

## Rows: 20,000  
## Columns: 13  
## $ transaction\_id <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,…  
## $ product\_id <dbl> 2, 3, 37, 88, 78, 25, 22, 15, 67, 12, 5, 61, 3…  
## $ customer\_id <dbl> 2950, 3120, 402, 3135, 787, 2339, 1542, 2459, …  
## $ transaction\_date <dttm> 2017-02-25, 2017-05-21, 2017-10-16, 2017-08-3…  
## $ online\_order <lgl> FALSE, TRUE, FALSE, FALSE, TRUE, TRUE, TRUE, F…  
## $ order\_status <chr> "Approved", "Approved", "Approved", "Approved"…  
## $ brand <chr> "Solex", "Trek Bicycles", "OHM Cycles", "Norco…  
## $ product\_line <chr> "Standard", "Standard", "Standard", "Standard"…  
## $ product\_class <chr> "medium", "medium", "low", "medium", "medium",…  
## $ product\_size <chr> "medium", "large", "medium", "medium", "large"…  
## $ list\_price <dbl> 71.49, 2091.47, 1793.43, 1198.46, 1765.30, 153…  
## $ standard\_cost <dbl> 53.62, 388.92, 248.82, 381.10, 709.48, 829.65,…  
## $ product\_first\_sold\_date <dbl> 41245, 41701, 36361, 36145, 42226, 39031, 3416…

str(df1)

## tibble [20,000 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ transaction\_id : num [1:20000] 1 2 3 4 5 6 7 8 9 10 ...  
## $ product\_id : num [1:20000] 2 3 37 88 78 25 22 15 67 12 ...  
## $ customer\_id : num [1:20000] 2950 3120 402 3135 787 ...  
## $ transaction\_date : POSIXct[1:20000], format: "2017-02-25" "2017-05-21" ...  
## $ online\_order : logi [1:20000] FALSE TRUE FALSE FALSE TRUE TRUE ...  
## $ order\_status : chr [1:20000] "Approved" "Approved" "Approved" "Approved" ...  
## $ brand : chr [1:20000] "Solex" "Trek Bicycles" "OHM Cycles" "Norco Bicycles" ...  
## $ product\_line : chr [1:20000] "Standard" "Standard" "Standard" "Standard" ...  
## $ product\_class : chr [1:20000] "medium" "medium" "low" "medium" ...  
## $ product\_size : chr [1:20000] "medium" "large" "medium" "medium" ...  
## $ list\_price : num [1:20000] 71.5 2091.5 1793.4 1198.5 1765.3 ...  
## $ standard\_cost : num [1:20000] 53.6 388.9 248.8 381.1 709.5 ...  
## $ product\_first\_sold\_date: num [1:20000] 41245 41701 36361 36145 42226 ...

summary(df1)

## transaction\_id product\_id customer\_id   
## Min. : 1 Min. : 0.00 Min. : 1.0   
## 1st Qu.: 5001 1st Qu.: 18.00 1st Qu.: 857.8   
## Median :10000 Median : 44.00 Median :1736.0   
## Mean :10000 Mean : 45.36 Mean :1738.2   
## 3rd Qu.:15000 3rd Qu.: 72.00 3rd Qu.:2613.0   
## Max. :20000 Max. :100.00 Max. :5034.0   
##   
## transaction\_date online\_order order\_status   
## Min. :2017-01-01 00:00:00 Mode :logical Length:20000   
## 1st Qu.:2017-04-01 00:00:00 FALSE:9811 Class :character   
## Median :2017-07-03 00:00:00 TRUE :9829 Mode :character   
## Mean :2017-07-01 14:08:05 NA's :360   
## 3rd Qu.:2017-10-02 00:00:00   
## Max. :2017-12-30 00:00:00   
##   
## brand product\_line product\_class product\_size   
## Length:20000 Length:20000 Length:20000 Length:20000   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## list\_price standard\_cost product\_first\_sold\_date  
## Min. : 12.01 Min. : 7.21 Min. :33259   
## 1st Qu.: 575.27 1st Qu.: 215.14 1st Qu.:35667   
## Median :1163.89 Median : 507.58 Median :38216   
## Mean :1107.83 Mean : 556.05 Mean :38200   
## 3rd Qu.:1635.30 3rd Qu.: 795.10 3rd Qu.:40672   
## Max. :2091.47 Max. :1759.85 Max. :42710   
## NA's :197 NA's :197

#Checking for null values for each column  
colSums(is.na(df1))

## transaction\_id product\_id customer\_id   
## 0 0 0   
## transaction\_date online\_order order\_status   
## 0 360 0   
## brand product\_line product\_class   
## 197 197 197   
## product\_size list\_price standard\_cost   
## 197 0 197   
## product\_first\_sold\_date   
## 197

#### There are missing values in 7 columns. They can be dropped or treated according to the nature of analysis

#Checking for duplicate values  
sum(duplicated(df1))

## [1] 0

#### There are no duplicate values.

#to count number of occurrences grouped by order\_status(approved and cancelled)  
table(df1$order\_status)

##   
## Approved Cancelled   
## 19821 179

#to count number of occurrences grouped by brand   
table(df1$brand)

##   
## Giant Bicycles Norco Bicycles OHM Cycles Solex Trek Bicycles   
## 3312 2910 3043 4253 2990   
## WeareA2B   
## 3295

#to count number of occurrences grouped by product\_line  
table(df1$product\_line)

##   
## Mountain Road Standard Touring   
## 423 3970 14176 1234

#to count number of occurrences grouped by product\_class  
table(df1$product\_class)

##   
## high low medium   
## 3013 2964 13826

#to count number of occurrences grouped by product\_size  
table(df1$product\_size)

##   
## large medium small   
## 3976 12990 2837

#convert product\_first\_sold\_date from numeric to date format  
df1$product\_first\_sold\_date<-as.Date(as.numeric(df1$product\_first\_sold\_date,na.rm=TRUE), origin = "1899-12-30")

#to count number of occurrences grouped by product\_first\_size\_sold\_date  
table(df1$product\_first\_sold\_date)

##   
## 1991-01-21 1991-05-06 1991-07-10 1991-08-05 1991-11-07 1991-11-10 1992-10-02   
## 213 203 201 181 193 179 234   
## 1992-10-11 1993-04-12 1993-04-20 1993-05-26 1993-06-23 1993-07-15 1993-07-20   
## 211 186 179 206 184 207 205   
## 1993-10-02 1994-07-12 1994-08-10 1994-09-09 1995-10-24 1995-12-19 1996-04-05   
## 193 200 206 162 209 206 189   
## 1996-11-09 1997-01-25 1997-02-09 1997-05-10 1997-08-25 1997-10-04 1998-12-16   
## 191 206 206 195 213 181 202   
## 1998-12-17 1999-06-23 1999-07-20 1999-07-26 1999-12-04 2000-05-22 2000-11-03   
## 201 202 201 205 198 181 189   
## 2001-11-25 2002-03-22 2002-08-31 2002-10-10 2003-01-05 2003-02-07 2003-02-16   
## 196 189 186 193 179 163 186   
## 2003-03-18 2003-07-21 2003-08-05 2003-09-09 2003-09-10 2004-01-16 2004-07-25   
## 203 227 191 207 189 215 189   
## 2004-08-07 2004-08-17 2004-09-28 2004-12-18 2005-05-10 2005-08-09 2005-10-22   
## 210 220 201 189 219 203 213   
## 2005-12-07 2006-02-02 2006-05-22 2006-10-01 2006-11-10 2007-08-04 2007-12-11   
## 191 202 194 171 203 212 208   
## 2008-03-19 2009-03-08 2009-04-12 2010-05-05 2010-06-07 2010-08-20 2010-11-05   
## 213 222 199 184 218 203 211   
## 2011-01-10 2011-03-16 2011-04-16 2011-05-07 2011-05-09 2011-08-24 2011-08-29   
## 199 207 182 193 176 193 192   
## 2012-04-10 2012-05-18 2012-06-04 2012-09-15 2012-12-02 2013-03-12 2013-06-09   
## 215 203 229 192 205 202 213   
## 2013-09-16 2014-03-03 2014-07-28 2014-10-10 2015-04-11 2015-05-21 2015-06-17   
## 202 212 169 166 193 177 187   
## 2015-08-02 2015-08-10 2015-10-18 2016-02-04 2016-03-29 2016-07-09 2016-11-14   
## 193 205 191 168 206 216 211   
## 2016-11-22 2016-12-06   
## 182 207

### Exploring New Customer List Dataset

head(df2)

## # A tibble: 6 × 23  
## first\_name last\_name gender past\_3\_years\_bik… DOB job\_title job\_industry\_ca…  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 Chickie Brister Male 86 1957… General … Manufacturing   
## 2 Morly Genery Male 69 1970… Structur… Property   
## 3 Ardelis Forrester Female 10 27269 Senior C… Financial Servi…  
## 4 Lucine Stutt Female 64 1979… Account … Manufacturing   
## 5 Melinda Hadlee Female 34 1965… Financia… Financial Servi…  
## 6 Druci Brandli Female 39 1951… Assistan… Entertainment   
## # … with 16 more variables: wealth\_segment <chr>, deceased\_indicator <chr>,  
## # owns\_car <chr>, tenure <dbl>, address <chr>, postcode <chr>, state <chr>,  
## # country <chr>, property\_valuation <chr>, ...17 <dbl>, ...18 <dbl>,  
## # ...19 <dbl>, ...20 <dbl>, ...21 <dbl>, Rank <dbl>, Value <dbl>

colnames(df2)

## [1] "first\_name" "last\_name"   
## [3] "gender" "past\_3\_years\_bike\_related\_purchases"  
## [5] "DOB" "job\_title"   
## [7] "job\_industry\_category" "wealth\_segment"   
## [9] "deceased\_indicator" "owns\_car"   
## [11] "tenure" "address"   
## [13] "postcode" "state"   
## [15] "country" "property\_valuation"   
## [17] "...17" "...18"   
## [19] "...19" "...20"   
## [21] "...21" "Rank"   
## [23] "Value"

glimpse(df2)

## Rows: 1,000  
## Columns: 23  
## $ first\_name <chr> "Chickie", "Morly", "Ardelis", "Lu…  
## $ last\_name <chr> "Brister", "Genery", "Forrester", …  
## $ gender <chr> "Male", "Male", "Female", "Female"…  
## $ past\_3\_years\_bike\_related\_purchases <chr> "86", "69", "10", "64", "34", "39"…  
## $ DOB <chr> "1957-07-12", "1970-03-22", "27269…  
## $ job\_title <chr> "General Manager", "Structural Eng…  
## $ job\_industry\_category <chr> "Manufacturing", "Property", "Fina…  
## $ wealth\_segment <chr> "Mass Customer", "Mass Customer", …  
## $ deceased\_indicator <chr> "N", "N", "N", "N", "N", "N", "N",…  
## $ owns\_car <chr> "Yes", "No", "No", "Yes", "No", "Y…  
## $ tenure <dbl> 14, 16, 10, 5, 19, 22, 8, 10, 5, 1…  
## $ address <chr> "45 Shopko Center", "14 Mccormick …  
## $ postcode <chr> "4500", "2113", "3505", "4814", "2…  
## $ state <chr> "QLD", "NSW", "VIC", "QLD", "NSW",…  
## $ country <chr> "Australia", "Australia", "Austral…  
## $ property\_valuation <chr> "6", "11", "5", "1", "9", "7", "7"…  
## $ ...17 <dbl> 0.82, 0.65, 0.75, 0.65, 0.60, 0.88…  
## $ ...18 <dbl> 1.0250, 0.6500, 0.7500, 0.8125, 0.…  
## $ ...19 <dbl> 1.281250, 0.812500, 0.750000, 0.81…  
## $ ...20 <dbl> 1.089063, 0.690625, 0.750000, 0.81…  
## $ ...21 <dbl> 1, 1, 1, 4, 4, 6, 6, 8, 8, 10, 10,…  
## $ Rank <dbl> 1, 1, 1, 4, 4, 6, 6, 8, 8, 10, 10,…  
## $ Value <dbl> 1.718750, 1.718750, 1.718750, 1.70…

str(df2)

## tibble [1,000 × 23] (S3: tbl\_df/tbl/data.frame)  
## $ first\_name : chr [1:1000] "Chickie" "Morly" "Ardelis" "Lucine" ...  
## $ last\_name : chr [1:1000] "Brister" "Genery" "Forrester" "Stutt" ...  
## $ gender : chr [1:1000] "Male" "Male" "Female" "Female" ...  
## $ past\_3\_years\_bike\_related\_purchases: chr [1:1000] "86" "69" "10" "64" ...  
## $ DOB : chr [1:1000] "1957-07-12" "1970-03-22" "27269" "1979-01-28" ...  
## $ job\_title : chr [1:1000] "General Manager" "Structural Engineer" "Senior Cost Accountant" "Account Representative III" ...  
## $ job\_industry\_category : chr [1:1000] "Manufacturing" "Property" "Financial Services" "Manufacturing" ...  
## $ wealth\_segment : chr [1:1000] "Mass Customer" "Mass Customer" "Affluent Customer" "Affluent Customer" ...  
## $ deceased\_indicator : chr [1:1000] "N" "N" "N" "N" ...  
## $ owns\_car : chr [1:1000] "Yes" "No" "No" "Yes" ...  
## $ tenure : num [1:1000] 14 16 10 5 19 22 8 10 5 17 ...  
## $ address : chr [1:1000] "45 Shopko Center" "14 Mccormick Park" "5 Colorado Crossing" "207 Annamark Plaza" ...  
## $ postcode : chr [1:1000] "4500" "2113" "3505" "4814" ...  
## $ state : chr [1:1000] "QLD" "NSW" "VIC" "QLD" ...  
## $ country : chr [1:1000] "Australia" "Australia" "Australia" "Australia" ...  
## $ property\_valuation : chr [1:1000] "6" "11" "5" "1" ...  
## $ ...17 : num [1:1000] 0.82 0.65 0.75 0.65 0.6 0.88 0.88 0.66 0.4 0.94 ...  
## $ ...18 : num [1:1000] 1.025 0.65 0.75 0.812 0.6 ...  
## $ ...19 : num [1:1000] 1.281 0.812 0.75 0.812 0.75 ...  
## $ ...20 : num [1:1000] 1.089 0.691 0.75 0.812 0.75 ...  
## $ ...21 : num [1:1000] 1 1 1 4 4 6 6 8 8 10 ...  
## $ Rank : num [1:1000] 1 1 1 4 4 6 6 8 8 10 ...  
## $ Value : num [1:1000] 1.72 1.72 1.72 1.7 1.7 ...

summary(df2)

## first\_name last\_name gender   
## Length:1000 Length:1000 Length:1000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## past\_3\_years\_bike\_related\_purchases DOB job\_title   
## Length:1000 Length:1000 Length:1000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## job\_industry\_category wealth\_segment deceased\_indicator owns\_car   
## Length:1000 Length:1000 Length:1000 Length:1000   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## tenure address postcode state   
## Min. : 0.00 Length:1000 Length:1000 Length:1000   
## 1st Qu.: 7.00 Class :character Class :character Class :character   
## Median :11.00 Mode :character Mode :character Mode :character   
## Mean :11.39   
## 3rd Qu.:15.00   
## Max. :22.00   
## country property\_valuation ...17 ...18   
## Length:1000 Length:1000 Min. :0.4000 Min. :0.4000   
## Class :character Class :character 1st Qu.:0.5875 1st Qu.:0.6500   
## Mode :character Mode :character Median :0.7500 Median :0.8337   
## Mean :0.7534 Mean :0.8460   
## 3rd Qu.:0.9225 3rd Qu.:1.0300   
## Max. :1.1000 Max. :1.3750   
## ...19 ...20 ...21 Rank   
## Min. :0.4000 Min. :0.3400 Min. : 1.0 Min. : 1.0   
## 1st Qu.:0.7125 1st Qu.:0.6587 1st Qu.: 250.0 1st Qu.: 250.0   
## Median :0.9300 Median :0.8500 Median : 500.0 Median : 500.0   
## Mean :0.9515 Mean :0.8790 Mean : 498.8 Mean : 498.8   
## 3rd Qu.:1.1625 3rd Qu.:1.0731 3rd Qu.: 750.2 3rd Qu.: 750.2   
## Max. :1.7188 Max. :1.7188 Max. :1000.0 Max. :1000.0   
## Value   
## Min. :0.3400   
## 1st Qu.:0.6495   
## Median :0.8600   
## Mean :0.8817   
## 3rd Qu.:1.0750   
## Max. :1.7188

#Dropping the unnamed columns from the datasets  
df2 <- df2 %>% select(-...17,-...18,-...19,-...20,-...21)

colnames(df2)

## [1] "first\_name" "last\_name"   
## [3] "gender" "past\_3\_years\_bike\_related\_purchases"  
## [5] "DOB" "job\_title"   
## [7] "job\_industry\_category" "wealth\_segment"   
## [9] "deceased\_indicator" "owns\_car"   
## [11] "tenure" "address"   
## [13] "postcode" "state"   
## [15] "country" "property\_valuation"   
## [17] "Rank" "Value"

#Checking for null values for each column  
colSums(is.na(df2))

## first\_name last\_name   
## 0 29   
## gender past\_3\_years\_bike\_related\_purchases   
## 0 0   
## DOB job\_title   
## 17 106   
## job\_industry\_category wealth\_segment   
## 0 0   
## deceased\_indicator owns\_car   
## 0 0   
## tenure address   
## 0 0   
## postcode state   
## 0 0   
## country property\_valuation   
## 0 0   
## Rank Value   
## 0 0

#### There are missing values in 3 columns. They can be dropped or treated according to the nature of analysis

#Checking for duplicate values  
sum(duplicated(df2))

## [1] 0

#### There are no duplicate values.

#to count number of occurrences grouped by gender  
table(df2$gender)

##   
## Female Male U   
## 513 470 17

#### There are 17 rows with unknown/unspecified gender.

#to count number of occurrences grouped by job\_industry\_category  
table(df2$job\_industry\_category)

##   
## Argiculture Entertainment Financial Services Health   
## 26 37 203 152   
## IT Manufacturing n/a Property   
## 51 199 165 64   
## Retail Telecommunications   
## 78 25

#to count number of occurrences grouped by wealth\_segment  
table(df2$wealth\_segment)

##   
## Affluent Customer High Net Worth Mass Customer   
## 241 251 508

#to count number of occurrences grouped by state  
table(df2$state)

##   
## NSW QLD VIC   
## 506 228 266

#to count number of occurrences grouped by owns\_car  
table(df2$owns\_car)

##   
## No Yes   
## 507 493

#to count number of occurrences grouped by deceased\_indicator  
table(df2$deceased\_indicator)

##   
## N   
## 1000

### Exploring Customer Demographic Dataset

head(df3)

## # A tibble: 6 × 13  
## customer\_id first\_name last\_name gender past\_3\_years\_bike… DOB job\_title  
## <dbl> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 1 Laraine Medendorp F 93 19644 Executiv…  
## 2 2 Eli Bockman Male 81 29571 Administ…  
## 3 3 Arlin Dearle Male 61 19744 Recruiti…  
## 4 4 Talbot <NA> Male 33 22557 <NA>   
## 5 5 Sheila-kathryn Calton Female 56 28258 Senior E…  
## 6 6 Curr Duckhouse Male 35 24366 <NA>   
## # … with 6 more variables: job\_industry\_category <chr>, wealth\_segment <chr>,  
## # deceased\_indicator <chr>, default <chr>, owns\_car <chr>, tenure <dbl>

colnames(df3)

## [1] "customer\_id" "first\_name"   
## [3] "last\_name" "gender"   
## [5] "past\_3\_years\_bike\_related\_purchases" "DOB"   
## [7] "job\_title" "job\_industry\_category"   
## [9] "wealth\_segment" "deceased\_indicator"   
## [11] "default" "owns\_car"   
## [13] "tenure"

glimpse(df3)

## Rows: 4,000  
## Columns: 13  
## $ customer\_id <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,…  
## $ first\_name <chr> "Laraine", "Eli", "Arlin", "Talbot…  
## $ last\_name <chr> "Medendorp", "Bockman", "Dearle", …  
## $ gender <chr> "F", "Male", "Male", "Male", "Fema…  
## $ past\_3\_years\_bike\_related\_purchases <dbl> 93, 81, 61, 33, 56, 35, 6, 31, 97,…  
## $ DOB <chr> "19644", "29571", "19744", "22557"…  
## $ job\_title <chr> "Executive Secretary", "Administra…  
## $ job\_industry\_category <chr> "Health", "Financial Services", "P…  
## $ wealth\_segment <chr> "Mass Customer", "Mass Customer", …  
## $ deceased\_indicator <chr> "N", "N", "N", "N", "N", "N", "N",…  
## $ default <chr> "\"'", "<script>alert('hi')</scrip…  
## $ owns\_car <chr> "Yes", "Yes", "Yes", "No", "Yes", …  
## $ tenure <dbl> 11, 16, 15, 7, 8, 13, 11, 7, 8, 20…

str(df3)

## tibble [4,000 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ customer\_id : num [1:4000] 1 2 3 4 5 6 7 8 9 10 ...  
## $ first\_name : chr [1:4000] "Laraine" "Eli" "Arlin" "Talbot" ...  
## $ last\_name : chr [1:4000] "Medendorp" "Bockman" "Dearle" NA ...  
## $ gender : chr [1:4000] "F" "Male" "Male" "Male" ...  
## $ past\_3\_years\_bike\_related\_purchases: num [1:4000] 93 81 61 33 56 35 6 31 97 49 ...  
## $ DOB : chr [1:4000] "19644" "29571" "19744" "22557" ...  
## $ job\_title : chr [1:4000] "Executive Secretary" "Administrative Officer" "Recruiting Manager" NA ...  
## $ job\_industry\_category : chr [1:4000] "Health" "Financial Services" "Property" "IT" ...  
## $ wealth\_segment : chr [1:4000] "Mass Customer" "Mass Customer" "Mass Customer" "Mass Customer" ...  
## $ deceased\_indicator : chr [1:4000] "N" "N" "N" "N" ...  
## $ default : chr [1:4000] "\"'" "<script>alert('hi')</script>" "43132" "() { \_; } >\_[$($())] { touch /tmp/blns.shellshock2.fail; }" ...  
## $ owns\_car : chr [1:4000] "Yes" "Yes" "Yes" "No" ...  
## $ tenure : num [1:4000] 11 16 15 7 8 13 11 7 8 20 ...

summary(df3)

## customer\_id first\_name last\_name gender   
## Min. : 1 Length:4000 Length:4000 Length:4000   
## 1st Qu.:1001 Class :character Class :character Class :character   
## Median :2000 Mode :character Mode :character Mode :character   
## Mean :2000   
## 3rd Qu.:3000   
## Max. :4000   
##   
## past\_3\_years\_bike\_related\_purchases DOB job\_title   
## Min. : 0.00 Length:4000 Length:4000   
## 1st Qu.:24.00 Class :character Class :character   
## Median :48.00 Mode :character Mode :character   
## Mean :48.89   
## 3rd Qu.:73.00   
## Max. :99.00   
##   
## job\_industry\_category wealth\_segment deceased\_indicator default   
## Length:4000 Length:4000 Length:4000 Length:4000   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## owns\_car tenure   
## Length:4000 Min. : 1.00   
## Class :character 1st Qu.: 6.00   
## Mode :character Median :11.00   
## Mean :10.66   
## 3rd Qu.:15.00   
## Max. :22.00   
## NA's :87

#Checking for null values for each column  
colSums(is.na(df3))

## customer\_id first\_name   
## 0 0   
## last\_name gender   
## 125 0   
## past\_3\_years\_bike\_related\_purchases DOB   
## 0 87   
## job\_title job\_industry\_category   
## 506 0   
## wealth\_segment deceased\_indicator   
## 0 0   
## default owns\_car   
## 240 0   
## tenure   
## 87

#### There are missing values in 5 columns. They can be dropped or treated according to the nature of analysis

#Checking for duplicate values  
sum(duplicated(df3))

## [1] 0

#### There are no duplicate values.

#to count number of occurrences grouped by gender  
table(df3$gender)

##   
## F Femal Female M Male U   
## 1 1 2037 1 1872 88

#### Categories are not correctly named. So we have to rename the categories.

#Rename the categories  
df3 <- df3 %>% mutate(gender = recode(gender, Femal = "Female",  
 Female = "Female",  
 F = "Female",  
 M = "Male",  
 Male = "Male"))

#to count number of occurrences grouped by gender  
table(df3$gender)

##   
## Female Male U   
## 2039 1873 88

#### There are 88 rows with unknown/unspecified gender.

#to count number of occurrences grouped by past\_3\_years\_bike\_related\_purchases  
table(df3$past\_3\_years\_bike\_related\_purchases)

##   
## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25   
## 38 35 50 30 37 33 45 35 28 39 43 48 42 47 40 37 56 42 42 56 54 30 43 46 39 42   
## 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51   
## 33 48 31 41 41 34 41 49 44 43 38 49 49 39 40 36 41 33 49 29 39 45 46 41 37 42   
## 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77   
## 32 49 39 35 34 35 29 44 40 29 37 40 41 38 38 54 42 43 41 43 48 48 35 32 41 38   
## 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99   
## 34 39 49 43 44 43 33 27 27 35 37 38 42 37 24 46 36 27 46 42 47 44

#Convert the datatype of DOB column from numeric to date format   
df3$DOB<-as.Date(as.numeric(df3$DOB,na.rm=TRUE), origin = "1899-12-30")

## Warning in as.Date(as.numeric(df3$DOB, na.rm = TRUE), origin = "1899-12-30"):  
## NAs introduced by coercion

#to count number of occurrences grouped by DOB  
df3.1 <- table(df3$DOB)  
head(df3.1)

##   
## 1931-10-23 1935-08-22 1940-09-22 1943-08-11 1944-01-24 1953-08-09   
## 1 1 1 1 1 1

#to count number of occurrences grouped by job\_title  
df3.2 <- table(df3$job\_title)  
head(df3.2)

##   
## Account Coordinator Account Executive   
## 31 33   
## Account Representative I Account Representative II   
## 10 4   
## Account Representative III Account Representative IV   
## 6 12

#to count number of occurrences grouped by job\_industry\_category  
table(df3$job\_industry\_category)

##   
## Argiculture Entertainment Financial Services Health   
## 113 136 774 602   
## IT Manufacturing n/a Property   
## 223 799 656 267   
## Retail Telecommunications   
## 358 72

#to count number of occurrences grouped by wealth\_segment  
table(df3$wealth\_segment)

##   
## Affluent Customer High Net Worth Mass Customer   
## 979 1021 2000

#to count number of occurrences grouped by deceased\_indicator  
table(df3$deceased\_indicator)

##   
## N Y   
## 3998 2

#### The values are inconsistent, hence dropping the default column

#Dropping the default column from the datasets  
df3 <- df3 %>% select(-default)

colnames(df3)

## [1] "customer\_id" "first\_name"   
## [3] "last\_name" "gender"   
## [5] "past\_3\_years\_bike\_related\_purchases" "DOB"   
## [7] "job\_title" "job\_industry\_category"   
## [9] "wealth\_segment" "deceased\_indicator"   
## [11] "owns\_car" "tenure"

#to count number of occurrences grouped by owns\_car  
table(df3$owns\_car)

##   
## No Yes   
## 1976 2024

#to count number of occurrences grouped by tenure  
table(df3$tenure)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20   
## 166 150 160 191 228 192 235 211 200 218 221 202 191 200 179 215 182 208 159 96   
## 21 22   
## 54 55

### Exploring Customer Address Dataset

head(df4)

## # A tibble: 6 × 6  
## customer\_id address postcode state country property\_valuati…  
## <dbl> <chr> <dbl> <chr> <chr> <dbl>  
## 1 1 060 Morning Avenue 2016 New South… Austral… 10  
## 2 2 6 Meadow Vale Court 2153 New South… Austral… 10  
## 3 4 0 Holy Cross Court 4211 QLD Austral… 9  
## 4 5 17979 Del Mar Point 2448 New South… Austral… 4  
## 5 6 9 Oakridge Court 3216 VIC Austral… 9  
## 6 7 4 Delaware Trail 2210 New South… Austral… 9

colnames(df4)

## [1] "customer\_id" "address" "postcode"   
## [4] "state" "country" "property\_valuation"

glimpse(df4)

## Rows: 3,999  
## Columns: 6  
## $ customer\_id <dbl> 1, 2, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17,…  
## $ address <chr> "060 Morning Avenue", "6 Meadow Vale Court", "0 Hol…  
## $ postcode <dbl> 2016, 2153, 4211, 2448, 3216, 2210, 2650, 2023, 304…  
## $ state <chr> "New South Wales", "New South Wales", "QLD", "New S…  
## $ country <chr> "Australia", "Australia", "Australia", "Australia",…  
## $ property\_valuation <dbl> 10, 10, 9, 4, 9, 9, 4, 12, 8, 4, 6, 8, 9, 4, 9, 12,…

str(df4)

## tibble [3,999 × 6] (S3: tbl\_df/tbl/data.frame)  
## $ customer\_id : num [1:3999] 1 2 4 5 6 7 8 9 11 12 ...  
## $ address : chr [1:3999] "060 Morning Avenue" "6 Meadow Vale Court" "0 Holy Cross Court" "17979 Del Mar Point" ...  
## $ postcode : num [1:3999] 2016 2153 4211 2448 3216 ...  
## $ state : chr [1:3999] "New South Wales" "New South Wales" "QLD" "New South Wales" ...  
## $ country : chr [1:3999] "Australia" "Australia" "Australia" "Australia" ...  
## $ property\_valuation: num [1:3999] 10 10 9 4 9 9 4 12 8 4 ...

summary(df4)

## customer\_id address postcode state   
## Min. : 1 Length:3999 Min. :2000 Length:3999   
## 1st Qu.:1004 Class :character 1st Qu.:2200 Class :character   
## Median :2004 Mode :character Median :2768 Mode :character   
## Mean :2004 Mean :2986   
## 3rd Qu.:3004 3rd Qu.:3750   
## Max. :4003 Max. :4883   
## country property\_valuation  
## Length:3999 Min. : 1.000   
## Class :character 1st Qu.: 6.000   
## Mode :character Median : 8.000   
## Mean : 7.514   
## 3rd Qu.:10.000   
## Max. :12.000

#Checking for null values for each column  
colSums(is.na(df4))

## customer\_id address postcode state   
## 0 0 0 0   
## country property\_valuation   
## 0 0

#### There are no null values.

#Checking for duplicate values  
sum(duplicated(df4))

## [1] 0

#### There are no duplicate values.

#to count number of occurrences grouped by postcode  
df4.1 <- table(df4$postcode)  
head(df4.1)

##   
## 2000 2007 2008 2009 2010 2011   
## 8 3 1 4 12 4

#to count number of occurrences grouped by state  
table(df4$state)

##   
## New South Wales NSW QLD VIC Victoria   
## 86 2054 838 939 82

#to count number of occurrences grouped by country  
table(df4$country)

##   
## Australia   
## 3999

#to count number of occurrences grouped by property\_valuation  
table(df4$property\_valuation)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12   
## 154 143 186 214 225 238 493 646 647 577 281 195