Homework 1

Andey Nunes, MS
Jordan Hilton
additional team member name(s) here

January 14, 2019

Document Setup

The first step for this week is to set up the R Markdown document options.

Next step, load the data sets for the homework and summarize.

```
catalog <- read_excel("catalog.xls")
customers <- read_excel("customers.xls")
order_lines <- read_excel("order_lines.xlsx")
```

```
## New names:
## * `` -> `..2`
```

```
orders <- read_excel("orders.xls")
```

Summary tables

```
catalog_summary <- summary(catalog)
glimpse(catalog)
```

```
## Observations: 761
## Variables: 7
## $ id            <dbl> 446, 455, 445, 444, 443, 442, 438, 439, 440, 441...
## $ product_code  <chr> "G79761", "plastic", "G75329", "G75328", "G75231...
## $ catalog_price <dbl> 9.9, 0.0, 11.9, 10.9, 12.9, 11.9, 9.5, 6.0, 6.0,...
## $ category1     <chr> "accessories", NA, "fishing", "fillet", "fillet"...
## $ manufact_id   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ vendor_id     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ name          <chr> "Exchange-A-Blade Sheath for 7 inch saw", "Plast...
```

```
pander(catalog_summary, caption = "catalog summary table")
```

Table 1: catalog summary table (continued below)

id	product_code	catalog_price	category1
Min. : 307	Length:761	Min. : 0	Length:761
1st Qu.: 525	Class :character	1st Qu.: 18	Class :character
Median : 728	Mode :character	Median : 34	Mode :character
Mean : 725	NA	Mean : 49	NA
3rd Qu.: 930	NA	3rd Qu.: 57	NA
Max. :1125	NA	Max. :654	NA

manufact_id	vendor_id	name
Min. :0.0	Min. :0.0	Length:761

manufact_id	vendor_id	name
1st Qu.:1.0	1st Qu.:1.0	Class :character
Median :1.0	Median :1.0	Mode :character
Mean :1.2	Mean :1.2	NA
3rd Qu.:1.0	3rd Qu.:1.0	NA
Max. :8.0	Max. :8.0	NA

```
head(catalog)
```

```
## # A tibble: 6 x 7
##      id product_code catalog_price category1  manufact_id vendor_id name
##   <dbl> <chr>                <dbl> <chr>            <dbl>     <dbl> <chr>
## 1   446 G79761                9.95 accessori~           1         1 Exchan~
## 2   455 plastic               0    <NA>                 1         1 Plasti~
## 3   445 G75329               12.0  fishing              1         1 Silver~
## 4   444 G75328               11.0  fillet               1         1 Silver~
## 5   443 G75231               13.0  fillet               1         1 "Gator~
## 6   442 G75230               12.0  fillet               1         1 "Gator~
```

```
customers_summary <- summary(customers)
glimpse(customers)
```

```
## Observations: 22,070
## Variables: 10
## $ cust_id     <dbl> 20696, 15465, 19830, 25532, 16044, 32394, 29572, 3...
## $ merchant_id <dbl> 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ firstName   <chr> "Kristina", "Paige", "Sherri", "Gretchen", "Karen"...
## $ lastName    <chr> "Chung", "Chen", "Melton", "Hill", "Puckett", "Son...
## $ bt_city     <chr> "Piedmont", "Cincinnati", "Shelbyville", "North ri...
## $ bt_state    <chr> "OK", "OH", "TN", "AZ", "ON", "OR", "GA", "VA", "K...
## $ bt_country  <chr> "United States", "United States", "United States",...
## $ bt_zip      <chr> "73078", "45227", "37160", "86052", "K8H 2X3", "97...
## $ cc_type     <chr> "Visa", "Visa", "Mastercard", "Visa", "Visa", "Mas...
## $ custcode    <chr> "P20696", "G15465", "P19830", "G25532", "G16044", ...
```

```
pander(customers_summary, caption = "customers summary table")
```

Table 3: customers summary table (continued below)

cust_id	merchant_id	firstName	lastName
Min. :10000	Min. :1.00	Length:22070	Length:22070
1st Qu.:15930	1st Qu.:1.00	Class :character	Class :character
Median :21448	Median :1.00	Mode :character	Mode :character
Mean :21408	Mean :1.05	NA	NA
3rd Qu.:26965	3rd Qu.:1.00	NA	NA
Max. :32482	Max. :2.00	NA	NA

Table 4: Table continues below

bt_city	bt_state	bt_country	bt_zip
Length:22070	Length:22070	Length:22070	Length:22070
Class :character	Class :character	Class :character	Class :character

bt_city	bt_state	bt_country	bt_zip
Mode :character	Mode :character	Mode :character	Mode :character
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA

cc_type	custcode
Length:22070	Length:22070
Class :character	Class :character
Mode :character	Mode :character
NA	NA
NA	NA
NA	NA

```
head(customers)
```

```
## # A tibble: 6 x 10
##   cust_id merchant_id firstName lastName bt_city bt_state bt_country bt_zip
##     <dbl>       <dbl> <chr>     <chr>    <chr>   <chr>    <chr>      <chr>
## 1   20696           2 Kristina  Chung    Piedmo~ OK       United St~ 73078
## 2   15465           1 Paige     Chen     Cincin~ OH       United St~ 45227
## 3   19830           2 Sherri    Melton   Shelby~ TN       United St~ 37160
## 4   25532           1 Gretchen  Hill     North ~ AZ       United St~ 86052
## 5   16044           1 Karen     Puckett  Petawa~ ON       Canada     K8H 2~
## 6   32394           1 Patrick   Song     Winche~ OR       United St~ 97495
## # ... with 2 more variables: cc_type <chr>, custcode <chr>
```

```
order_lines_summary <- summary(order_lines)
glimpse(order_lines)
```

```
## Observations: 1,356
## Variables: 2
## $ `Sum of Shipped Total` <chr> "Row Labels", "411", "Multi-PlierÂ® 800...
## $ `..2`                  <chr> "Total", "27507.100000000122", "27507.1...
```

```
pander(order_lines_summary, caption = "order_lines summary table")
```

Table 6: order_lines summary table

Sum of Shipped Total	..2
Length:1356	Length:1356
Class :character	Class :character
Mode :character	Mode :character

```
head(order_lines)
```

```
## # A tibble: 6 x 2
##   `Sum of Shipped Total`     ..2
##   <chr>                      <chr>
## 1 Row Labels                 Total
## 2 411                        27507.100000000122
## 3 Multi-PlierÂ® 800 - Legend 27507.100000000122
## 4 757                         21591.649999999994
## 5 LMFâ„¢ II Infantry - Black 21591.649999999994
## 6 395                         20355.900000000009
```

```
orders_summary <- summary(orders)
glimpse(orders)
```

```
## Observations: 23,256
## Variables: 18
## $ order_id          <dbl> 14035, 14034, 14033, 14032, 14031, 14030, 14...
## $ merchant_id       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ order_date        <dttm> 2003-10-17, 2003-10-16, 2003-10-16, 2003-10...
## $ po_number         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ cust_id           <dbl> 10034, 10033, 10032, 10031, 10030, 10029, 10...
## $ order_status      <chr> "S", "S", "S", "S", "S", "S", "S", "S", "S",...
## $ ship_method       <chr> "GND", "3DS", "GND", "GND", "3DS", "1DA", "G...
## $ items_amount      <dbl> 58.9, 8.9, 50.0, 11.9, 9.9, 109.9, 23.9, 40....
## $ amt_bracket       <chr> "C", "A", "B", "B", "A", "D", "B", "B", "A",...
## $ total_weight      <dbl> 2.3, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.2, 1.0,...
## $ total_ship        <dbl> 5.5, 9.0, 5.2, 5.4, 9.0, 27.3, 5.3, 6.1, 5.4...
## $ total_hand        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ total_tax         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ total_amount      <dbl> 64, 18, 55, 17, 19, 137, 29, 46, 15, 23, 29,...
## $ order_status_date <dttm> 2003-10-17, 2003-10-17, 2003-10-17, 2003-10...
## $ send_inv_to_bill  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ coupon_code       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ spec_instr        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
```

```
pander(orders_summary, caption = "orders summary table")
```

Table 7: orders summary table (continued below)

order_id	merchant_id	order_date	po_number
Min. :14000	Min. :1.00	Min. :2003-10-10 00:00:00	Length:23256
1st Qu.:20134	1st Qu.:1.00	1st Qu.:2006-04-28 00:00:00	Class :character
Median :25948	Median :1.00	Median :2007-07-02 00:00:00	Mode :character
Mean :25918	Mean :1.05	Mean :2007-08-11 16:51:42	NA
3rd Qu.:31761	3rd Qu.:1.00	3rd Qu.:2008-12-19 00:00:00	NA
Max. :37575	Max. :2.00	Max. :2011-01-21 00:00:00	NA

Table 8: Table continues below

cust_id	order_status	ship_method	items_amount
Min. : 0	Length:23256	Length:23256	Min. : 0
1st Qu.:15778	Class :character	Class :character	1st Qu.: 28
Median :21302	Mode :character	Mode :character	Median : 48
Mean :21295	NA	NA	Mean : 73
3rd Qu.:26849	NA	NA	3rd Qu.: 80
Max. :32482	NA	NA	Max. :9590

Table 9: Table continues below

amt_bracket	total_weight	total_ship	total_hand	total_tax
Length:23256	Min. : 0	Min. : 0	Min. :0	Min. :0
Class :character	1st Qu.: 1	1st Qu.: 7	1st Qu.:0	1st Qu.:0
Mode :character	Median : 2	Median : 8	Median :0	Median :0
NA	Mean : 3	Mean : 11	Mean :0	Mean :0
NA	3rd Qu.: 3	3rd Qu.: 10	3rd Qu.:0	3rd Qu.:0
NA	Max. :483	Max. :631	Max. :0	Max. :0

Table 10: Table continues below

total_amount	order_status_date	send_inv_to_bill	coupon_code
Min. : 6	Min. :2003-10-10 00:00:00	Min. :0.00	Mode:logical
1st Qu.: 36	1st Qu.:2006-05-30 18:00:00	1st Qu.:0.00	NA's:23256
Median : 57	Median :2007-07-12 00:00:00	Median :0.00	NA
Mean : 84	Mean :2007-08-21 21:51:27	Mean :0.05	NA
3rd Qu.: 94	3rd Qu.:2008-12-26 00:00:00	3rd Qu.:0.00	NA
Max. :9590	Max. :2011-01-21 00:00:00	Max. :1.00	NA

spec_instr
Mode:logical
NA's:23256
NA
NA
NA
NA

```r
head(orders)
```

```
## # A tibble: 6 x 18
##   order_id merchant_id order_date          po_number cust_id order_status
##      <dbl>       <dbl> <dttm>              <chr>       <dbl> <chr>
## 1    14035           1 2003-10-17 00:00:00 <NA>        10034 S
## 2    14034           1 2003-10-16 00:00:00 <NA>        10033 S
## 3    14033           1 2003-10-16 00:00:00 <NA>        10032 S
## 4    14032           1 2003-10-16 00:00:00 <NA>        10031 S
## 5    14031           1 2003-10-16 00:00:00 <NA>        10030 S
## 6    14030           1 2003-10-16 00:00:00 <NA>        10029 S
## # ... with 12 more variables: ship_method <chr>, items_amount <dbl>,
## #   amt_bracket <chr>, total_weight <dbl>, total_ship <dbl>,
## #   total_hand <dbl>, total_tax <dbl>, total_amount <dbl>,
## #   order_status_date <dttm>, send_inv_to_bill <dbl>, coupon_code <lgl>,
## #   spec_instr <lgl>
```

column names (variables) | assign a type: "question", "answer", or "link" | variable class | count missing values | range = max - min |

This section is for building some custom functions that will come in handy later

```
countNA <- function(x) {sum(is.na(x)) }
get_range <- function(x) {ifelse(is.numeric(x), diff(range(x)), NA)}

# This function creates the generic structure for the tables in Part B. The only content is the variabl
make_partBtable <- function(x){
   df <- tibble(variable_name = names(x),
                    variable_type = NA, # assign one of: "question", "answer", "link"
                    variable_class = map(x, class),
                    count_missing = map_int(x, countNA),
                    variable_range = map_dbl(x, get_range))

   return(df)

}
```

Homework Questions

<<<<<<< HEAD ### Part B: Specific Questions

In an effort to code more efficiently I've defined a function to produce each table, however, I ran into a problem with the `variable_class` column. Compare the tables below with the class reported in the summary/glimpse tables above and you will see.

```
catalog_table <- make_partBtable(catalog)
pander(catalog_table, caption = "Catalog Data Table Details")
```

Table 12: Catalog Data Table Details

variable_name	variable_type	variable_class	count_missing	variable_range
id	NA	numeric	0	818
product_code	NA	character	1	NA
catalog_price	NA	numeric	0	654
category1	NA	character	645	NA
manufact_id	NA	numeric	0	8
vendor_id	NA	numeric	0	8
name	NA	character	1	NA

```
customers_table <- make_partBtable(customers)
pander(customers_table, caption = "Customers Data Table Details")
```

Table 13: Customers Data Table Details

variable_name	variable_type	variable_class	count_missing	variable_range
cust_id	NA	numeric	0	22482
merchant_id	NA	numeric	0	1
firstName	NA	character	12070	NA
lastName	NA	character	12070	NA
bt_city	NA	character	1	NA
bt_state	NA	character	137	NA
bt_country	NA	character	0	NA
bt_zip	NA	character	0	NA
cc_type	NA	character	0	NA

variable_name	variable_type	variable_class	count_missing	variable_range
custcode	NA	character	0	NA

```r
order_lines_table <- make_partBtable(order_lines)
pander(order_lines_table, caption = "Order_lines Data Table Details")
```

Table 14: Order_lines Data Table Details (continued below)

variable_name	variable_type	variable_class	count_missing	variable_range
Sum of Shipped Total	NA	character	0	NA
..2	NA	character	0	NA

```r
orders_table <- make_partBtable(orders)
pander(orders_table, caption = "Orders Data Table Details")
```

```
## Warning in `[<-.data.frame`(`*tmp*`, , j, value = list(order_id =
## "numeric", : provided 18 variables to replace 1 variables
```

Table 16: Orders Data Table Details (continued below)

variable_name	variable_type	variable_class	count_missing
order_id	NA	numeric	0
merchant_id	NA	numeric	0
order_date	NA	numeric	0
po_number	NA	numeric	22742
cust_id	NA	numeric	0
order_status	NA	numeric	0
ship_method	NA	numeric	186
items_amount	NA	numeric	0
amt_bracket	NA	numeric	0
total_weight	NA	numeric	0
total_ship	NA	numeric	0
total_hand	NA	numeric	0
total_tax	NA	numeric	0
total_amount	NA	numeric	0
order_status_date	NA	numeric	0
send_inv_to_bill	NA	numeric	0
coupon_code	NA	numeric	23256
spec_instr	NA	numeric	23256

variable_range
23575
1

variable_range
NA
NA
32482
NA
NA
9590
NA
483
631.3
0
0
9584
NA
1
NA
NA

========

For question B

```
### this function finds the number of NAs for each column
sapply(catalog, function(y) sum(is.na(y)))
```

```
##            id  product_code catalog_price     category1   manufact_id 
##             0             1             0           645             0 
##     vendor_id          name 
##             0             1
```

```
### note that only one row has a blank value for product code or name, find out which that is
which(is.na(catalog[,2]))
```

```
## [1] 267
```

```
catalog[267,]
```

```
## # A tibble: 1 x 7
##      id product_code catalog_price category1 manufact_id vendor_id name 
##   <dbl> <chr>                <dbl> <chr>           <dbl>     <dbl> <chr>
## 1   596 <NA>                     0 <NA>                1         1 <NA>
```

```
### load our table of answers about the catlog and display it
cataloganswer<-read_excel("cataloganswer.xlsx")
pander(cataloganswer)
```

Field	Q/A/L	Data Type	Nulls
id	Link	Integer	0
product_code	Link	Text	1
catalog_price	Answer	Currency	0
category1	Question	Text	645
manufact_id	Question	Integer	0
vendor_id	Question	Integer	0

Field	Q/A/L	Data Type	Nulls
name	Answer	Text	1

```
#>>>>>>> ae8ab7bbb4f6e6eb6429a38c088e54c7b85037b0
```

References