BLACK FRIDAY SALES ANALYSIS

GROUP 9

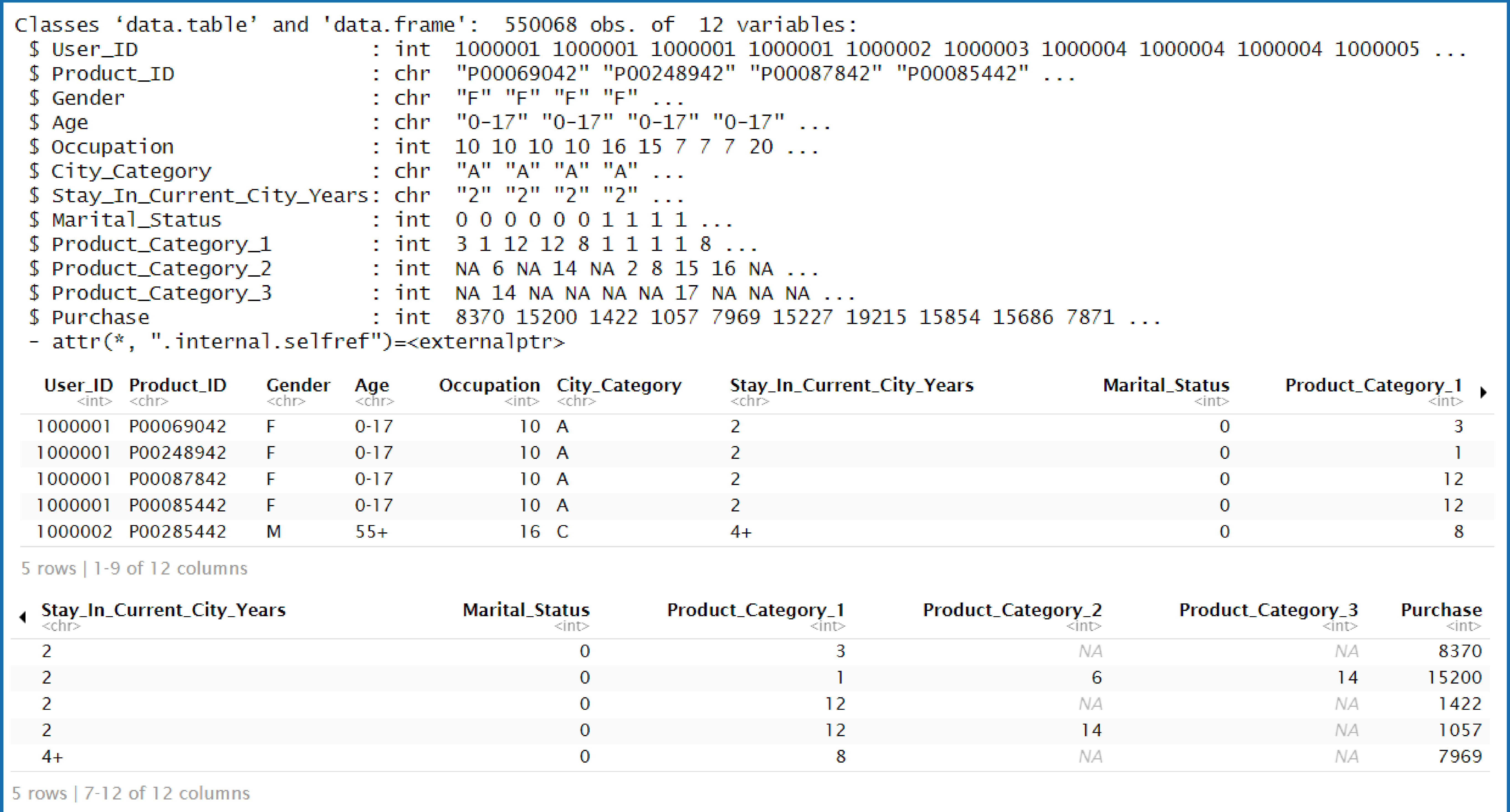
Steve Marcello Liem - 2602071410

Davin Edbert S. Halim - 2602067086

Felicia Andrea Tandoko - 2602059342

"ABC PRIVATE LIMITED" COMPANY

DATA DESCRIPTION



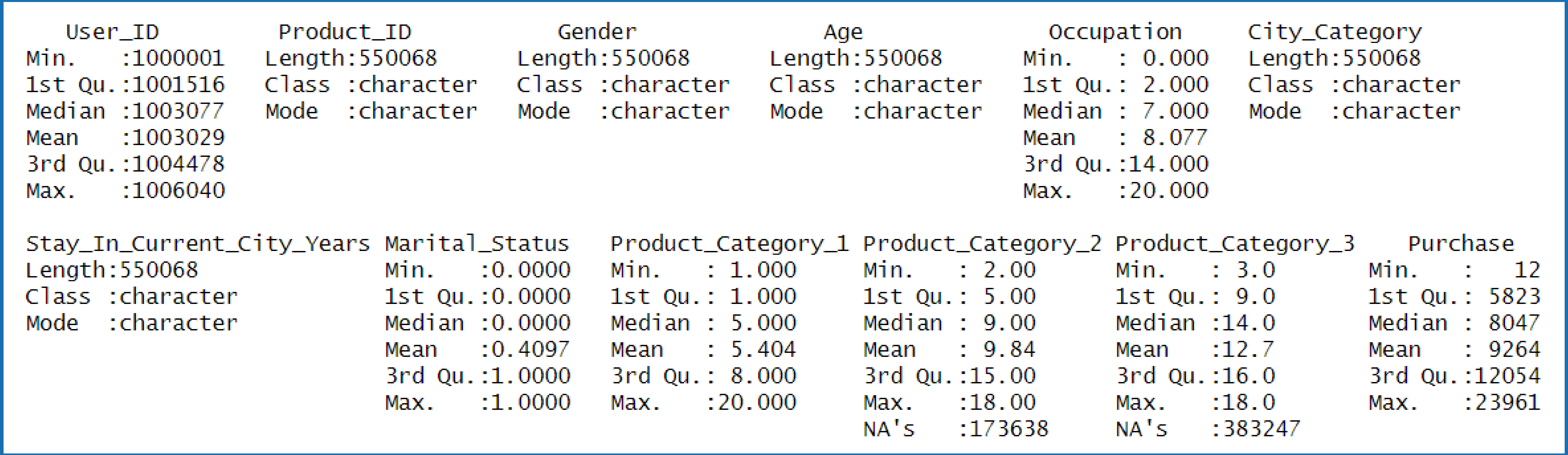
```
Classes ‘data.table’ and 'data.frame':  550068 obs. of  12 variables:
 $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1000004 1000004 1000005 ...
 $ Product_ID                : chr  "P00069042" "P00248942" "P00087842" "P00085442" ...
 $ Gender                    : chr  "F" "F" "F" "F" ...
 $ Age                       : chr  "0-17" "0-17" "0-17" "0-17" ...
 $ Occupation                : int  10 10 10 10 16 15 7 7 7 20 ...
 $ City_Category             : chr  "A" "A" "A" "A" ...
 $ Stay_In_Current_City_Years: chr  "2" "2" "2" "2" ...
 $ Marital_Status            : int  0 0 0 0 0 0 1 1 1 1 ...
 $ Product_Category_1        : int  3 1 12 12 8 1 1 1 1 8 ...
 $ Product_Category_2        : int  NA 6 NA 14 NA 2 8 15 16 NA ...
 $ Product_Category_3        : int  NA 14 NA NA NA NA 17 NA NA NA ...
 $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1
<int>	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	<int>	<int>
1000001	P00069042	F	0-17	10	A	2	0	3
1000001	P00248942	F	0-17	10	A	2	0	1
1000001	P00087842	F	0-17	10	A	2	0	12
1000001	P00085442	F	0-17	10	A	2	0	12
1000002	P00285442	M	55+	16	C	4+	0	8

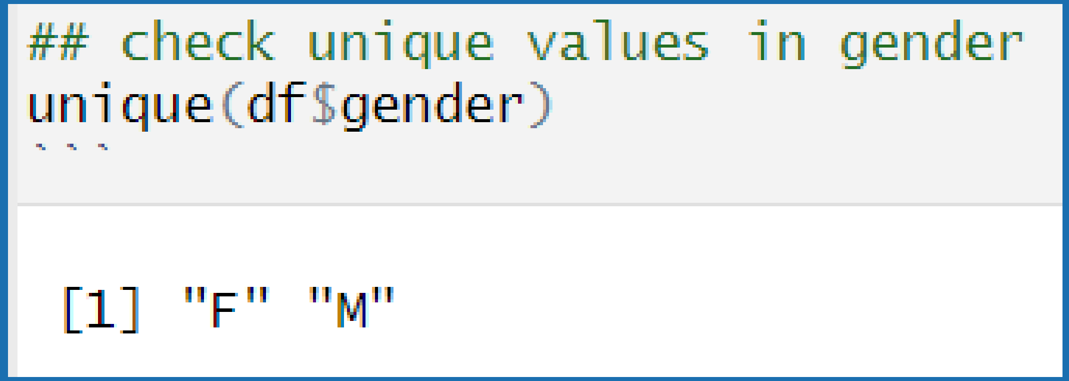
5 rows | 1-9 of 12 columns

Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
<chr>	<int>	<int>	<int>	<int>	<int>
2	0	3	NA	NA	8370
2	0	1	6	14	15200
2	0	12	NA	NA	1422
2	0	12	14	NA	1057
4+	0	8	NA	NA	7969

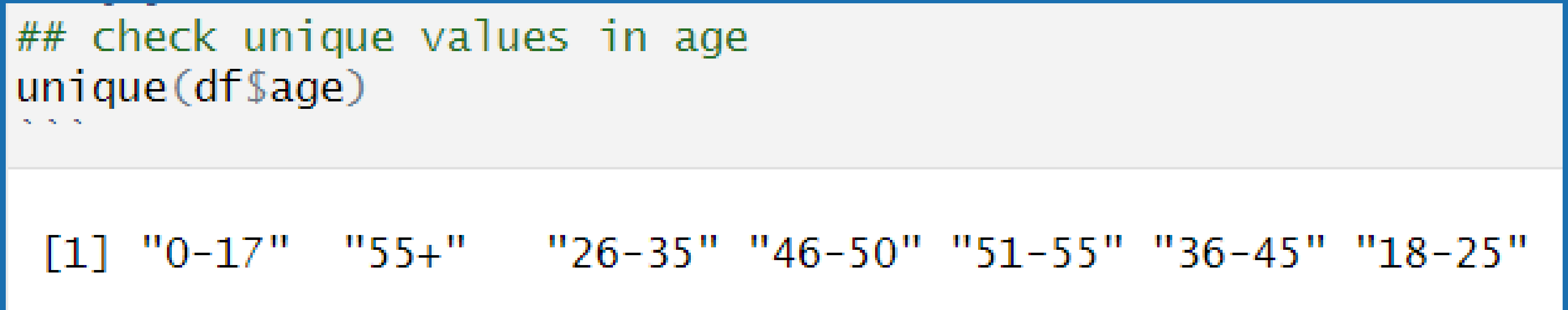
5 rows | 7-12 of 12 columns



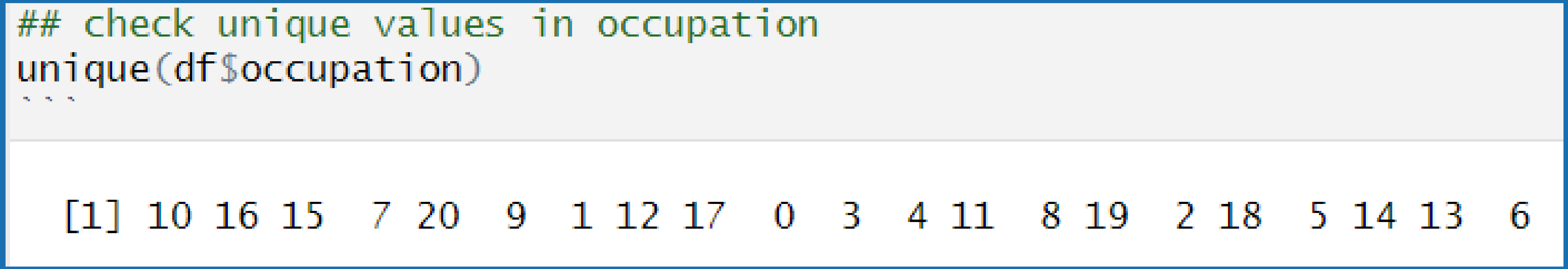
```
   User_ID         Product_ID           Gender              Age              Occupation     City_Category
Min.   :1000001   Length:550068      Length:550068      Length:550068      Min.   : 0.000   Length:550068
1st Qu.:1001516   Class :character   Class :character   Class :character   1st Qu.: 2.000   Class :character
Median :1003077   Mode  :character   Mode  :character   Mode  :character   Median : 7.000   Mode  :character
Mean   :1003029                                                            Mean   : 8.077
3rd Qu.:1004478                                                            3rd Qu.:14.000
Max.   :1006040                                                            Max.   :20.000

Stay_In_Current_City_Years Marital_Status   Product_Category_1 Product_Category_2 Product_Category_3    Purchase
Length:550068              Min.   :0.0000   Min.   : 1.000     Min.   : 2.00      Min.   : 3.0       Min.   :   12
Class :character           1st Qu.:0.0000   1st Qu.: 1.000     1st Qu.: 5.00      1st Qu.: 9.0       1st Qu.: 5823
Mode  :character           Median :0.0000   Median : 5.000     Median : 9.00      Median :14.0       Median : 8047
                           Mean   :0.4097   Mean   : 5.404     Mean   : 9.84      Mean   :12.7       Mean   : 9264
                           3rd Qu.:1.0000   3rd Qu.: 8.000     3rd Qu.:15.00      3rd Qu.:16.0       3rd Qu.:12054
                           Max.   :1.0000   Max.   :20.000     Max.   :18.00      Max.   :18.0       Max.   :23961
                                                               NA's   :173638     NA's   :383247
```



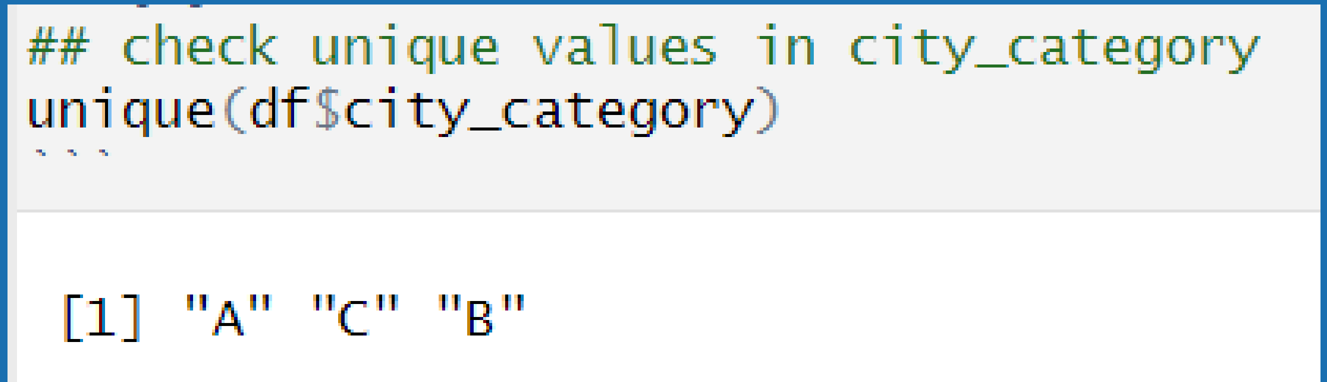
~~~
## check unique values in gender
unique(df$gender)
```

 [1] "F" "M"
~~~



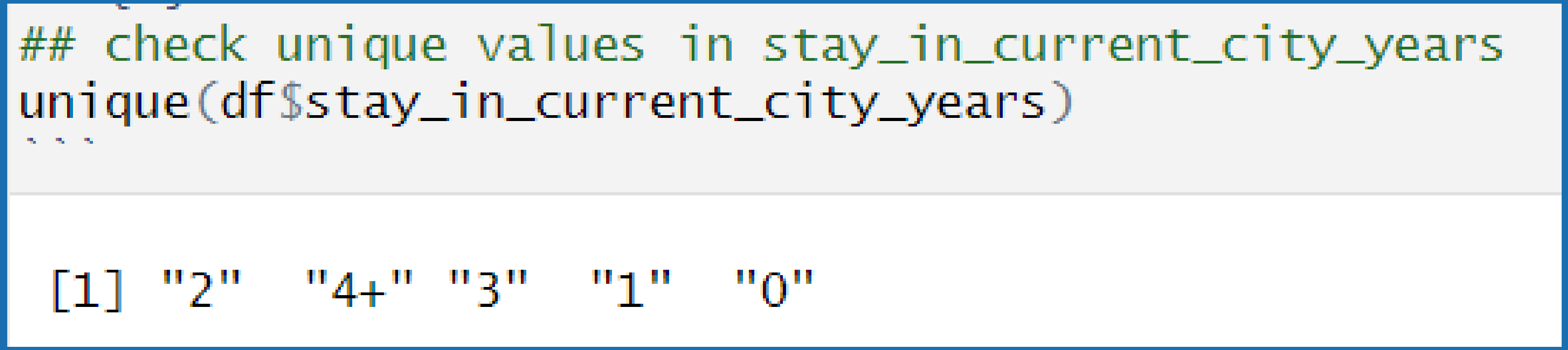
~~~
## check unique values in age
unique(df$age)
```

 [1] "0-17"  "55+"   "26-35" "46-50" "51-55" "36-45" "18-25"
~~~



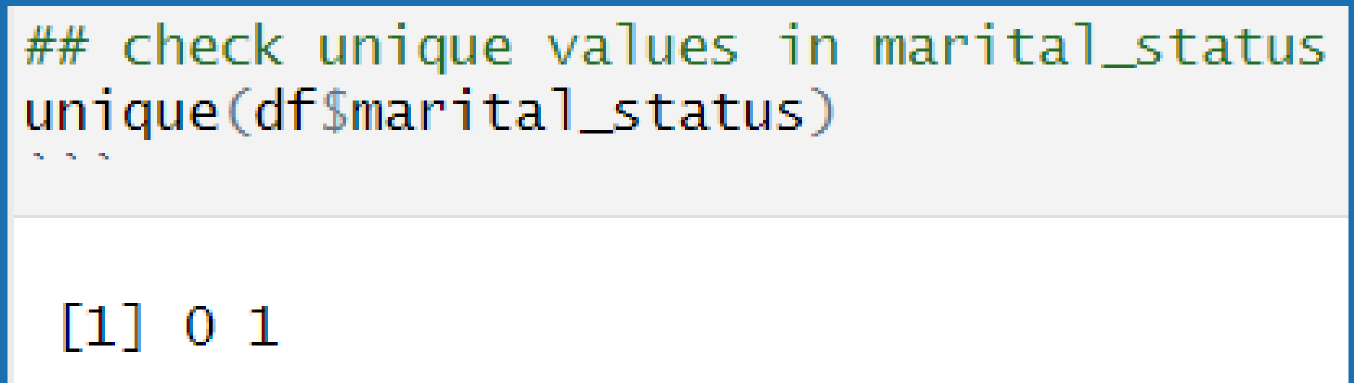
~~~
## check unique values in occupation
unique(df$occupation)
```

  [1] 10 16 15  7 20  9  1 12 17  0  3  4 11  8 19  2 18  5 14 13  6
~~~



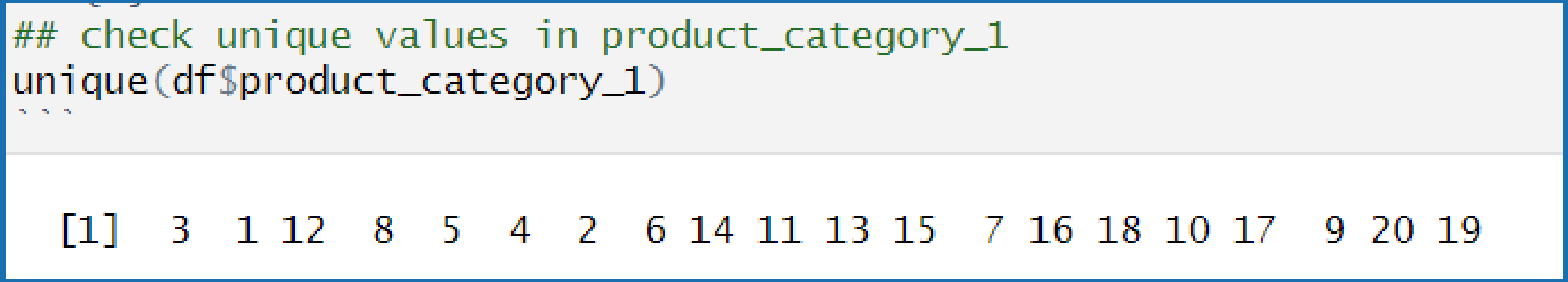
~~~
## check unique values in city_category
unique(df$city_category)
```

 [1] "A" "C" "B"
~~~



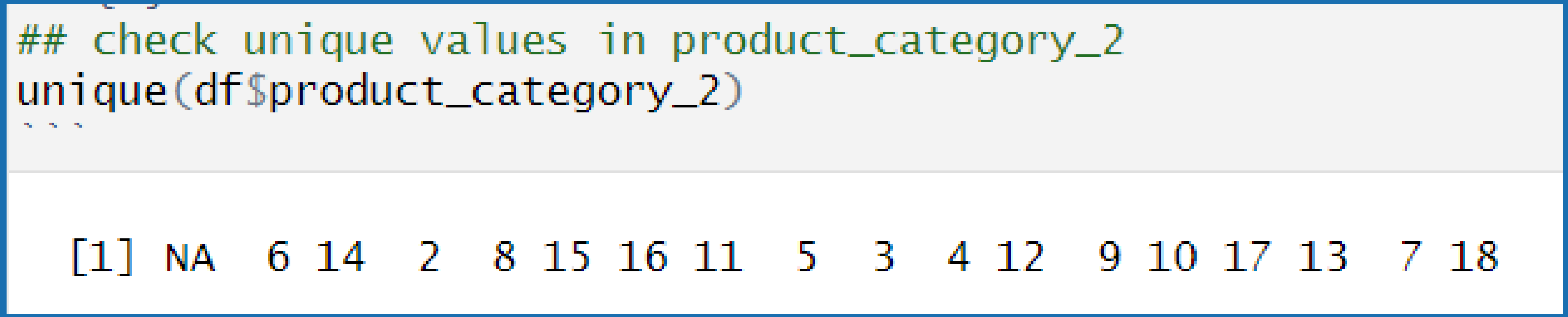
~~~
## check unique values in stay_in_current_city_years
unique(df$stay_in_current_city_years)
```

 [1] "2"  "4+" "3"  "1"  "0"
~~~



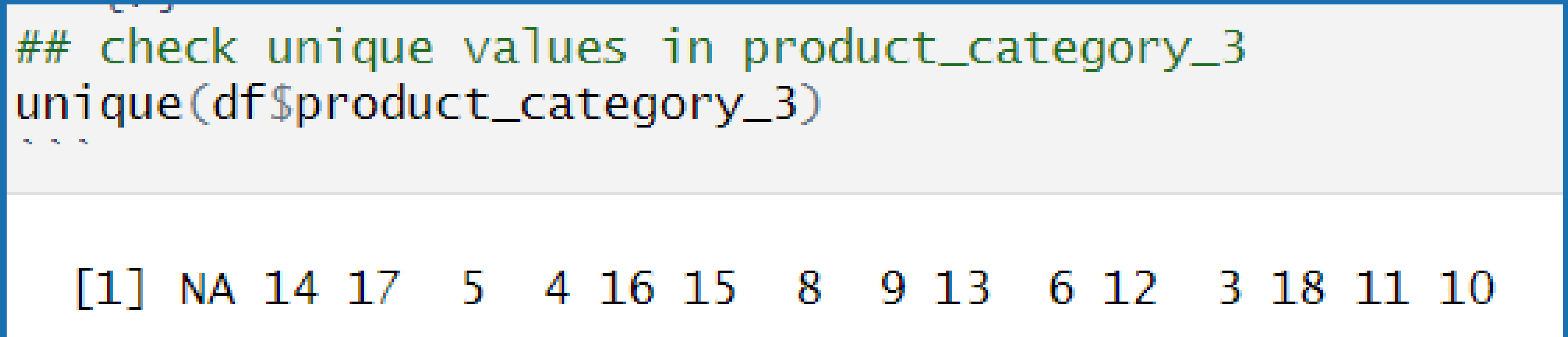
~~~
## check unique values in marital_status
unique(df$marital_status)
```

 [1] 0 1
~~~



~~~
## check unique values in product_category_1
unique(df$product_category_1)
```

  [1]  3  1 12  8  5  4  2  6 14 11 13 15  7 16 18 10 17  9 20 19
~~~

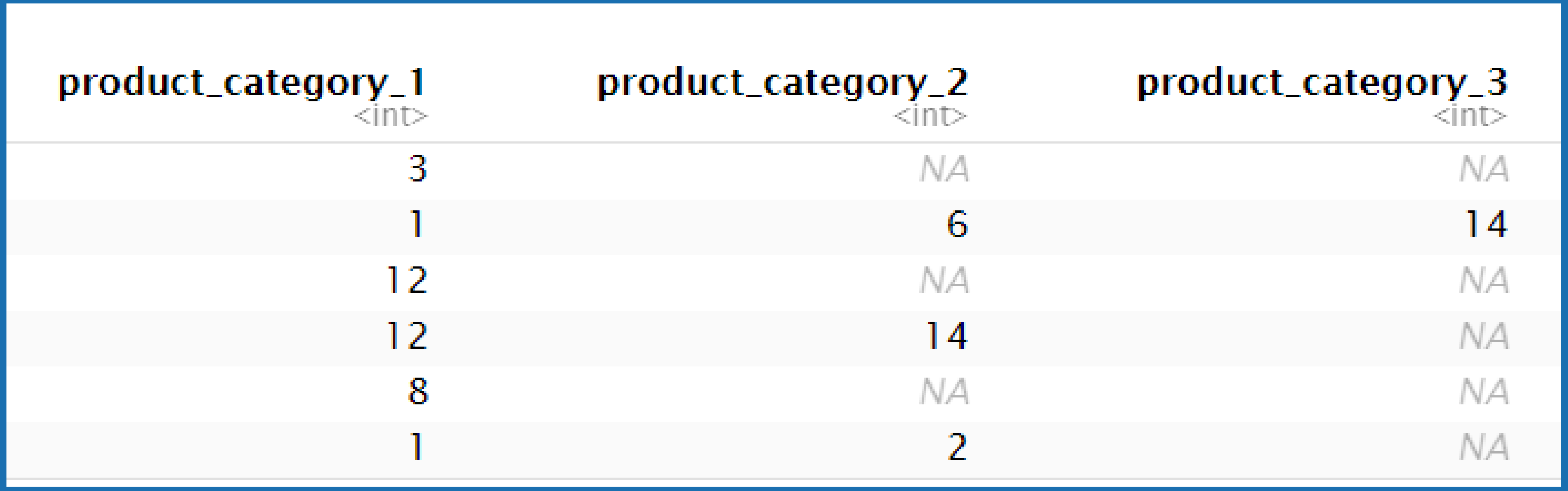


~~~
## check unique values in product_category_2
unique(df$product_category_2)
```

  [1] NA  6 14  2  8 15 16 11  5  3  4 12  9 10 17 13  7 18
~~~

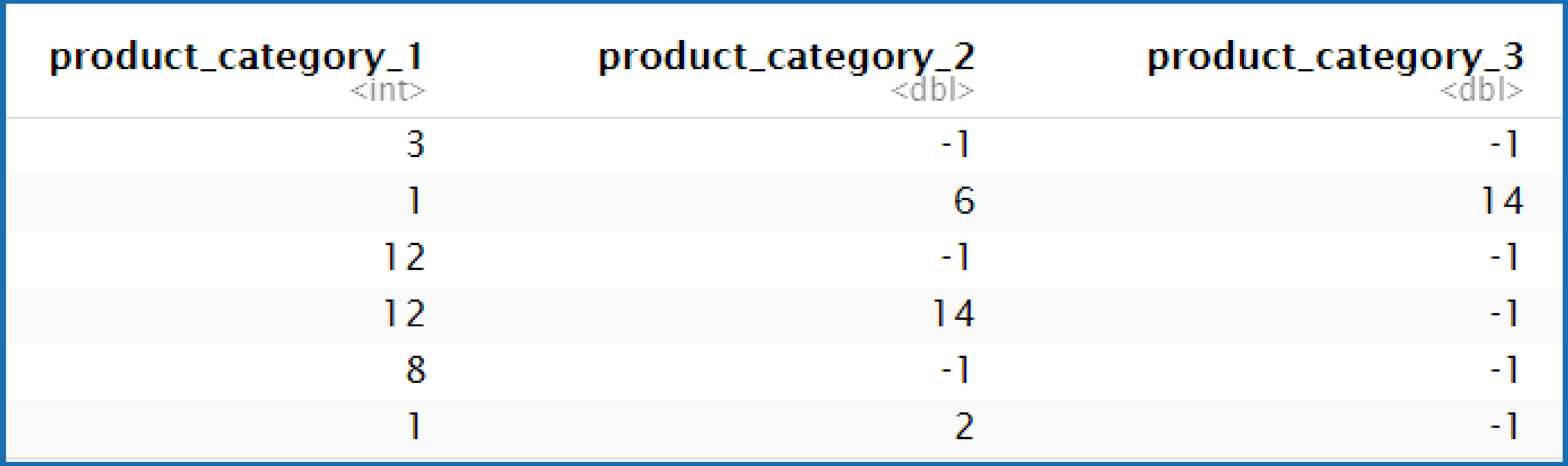


~~~
## check unique values in product_category_3
unique(df$product_category_3)
```

  [1] NA 14 17  5  4 16 15  8  9 13  6 12  3 18 11 10
~~~

DATA CLEANING

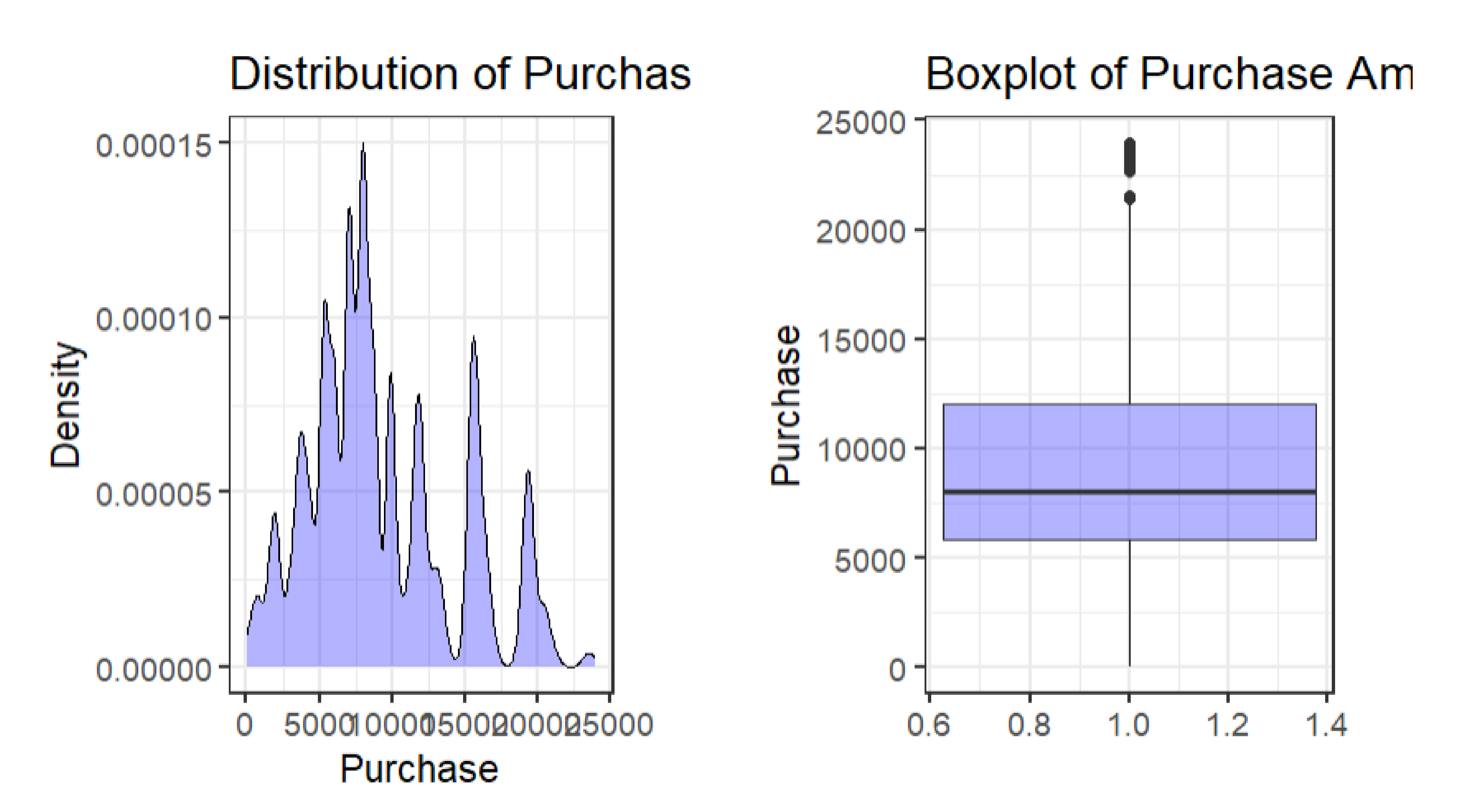


product_category_1	product_category_2	product_category_3
<int>	<int>	<int>
3	NA	NA
1	6	14
12	NA	NA
12	14	NA
8	NA	NA
1	2	NA

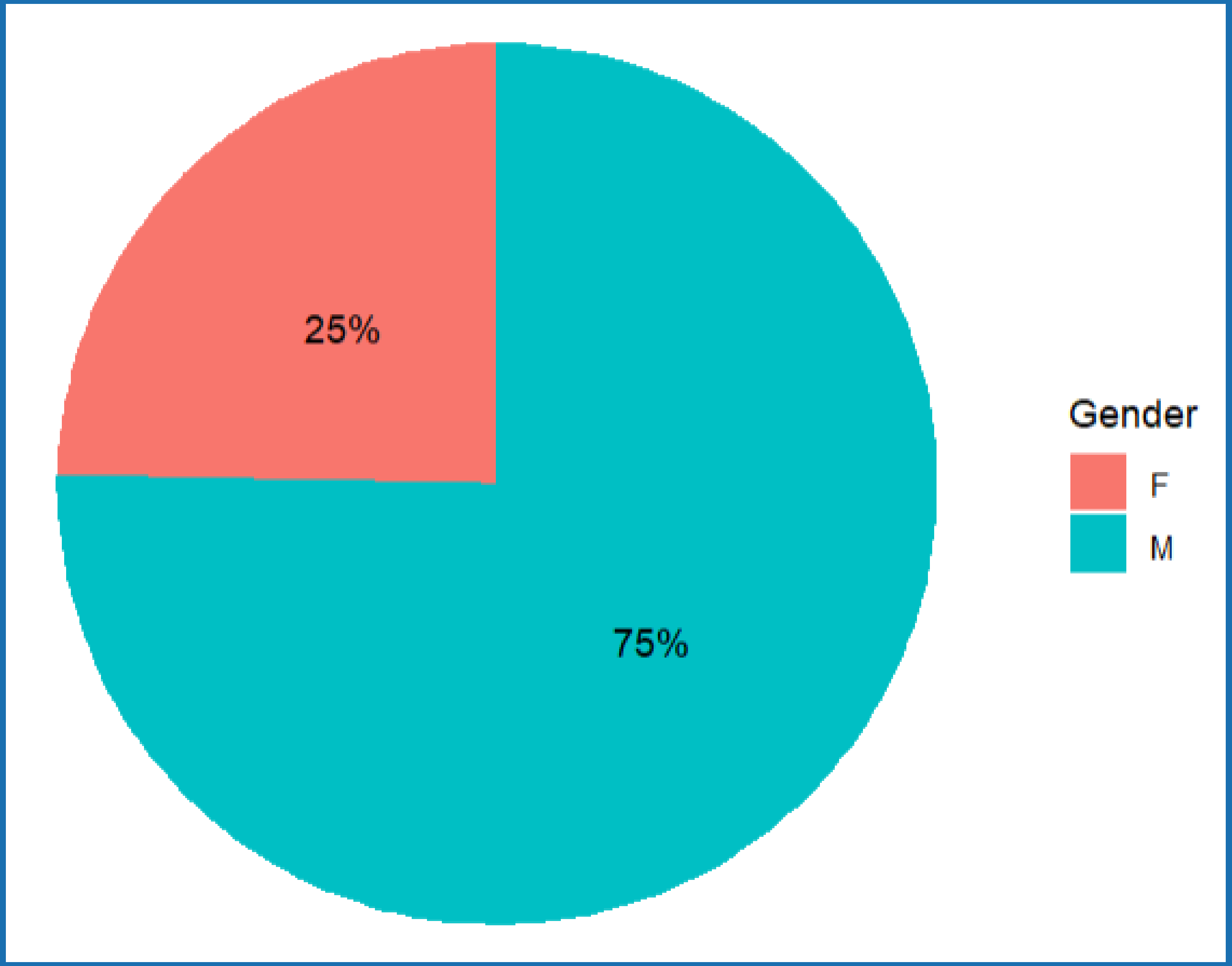


product_category_1	product_category_2	product_category_3
<int>	<dbl>	<dbl>
3	-1	-1
1	6	14
12	-1	-1
12	14	-1
8	-1	-1
1	2	-1

EXPLORATORY DATA ANALYSIS

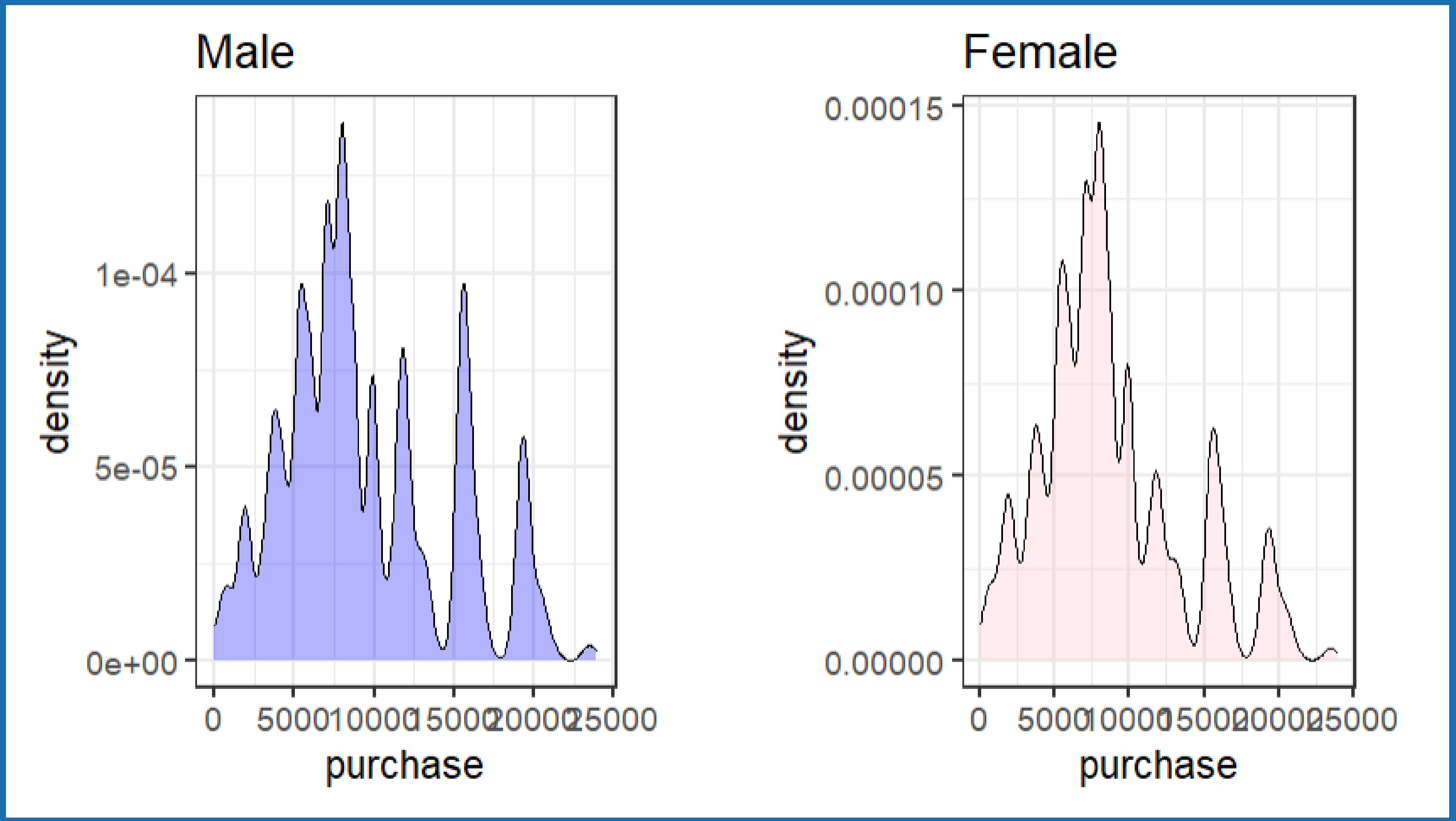


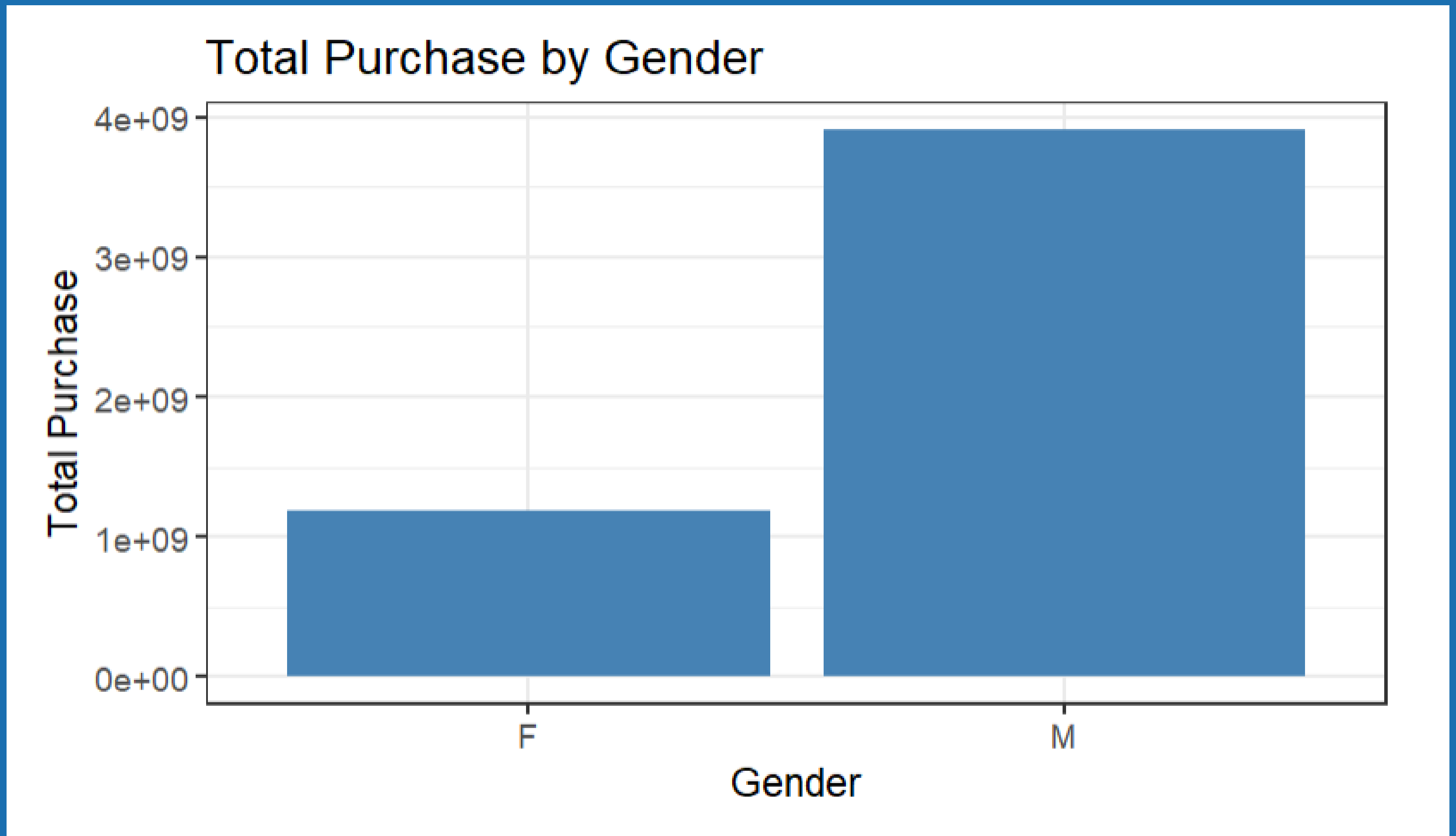
GENDER VS PURCHASE



MALE CUSTOMER VISITED THE STORE MORE THAN FEMALE CUSTOMER

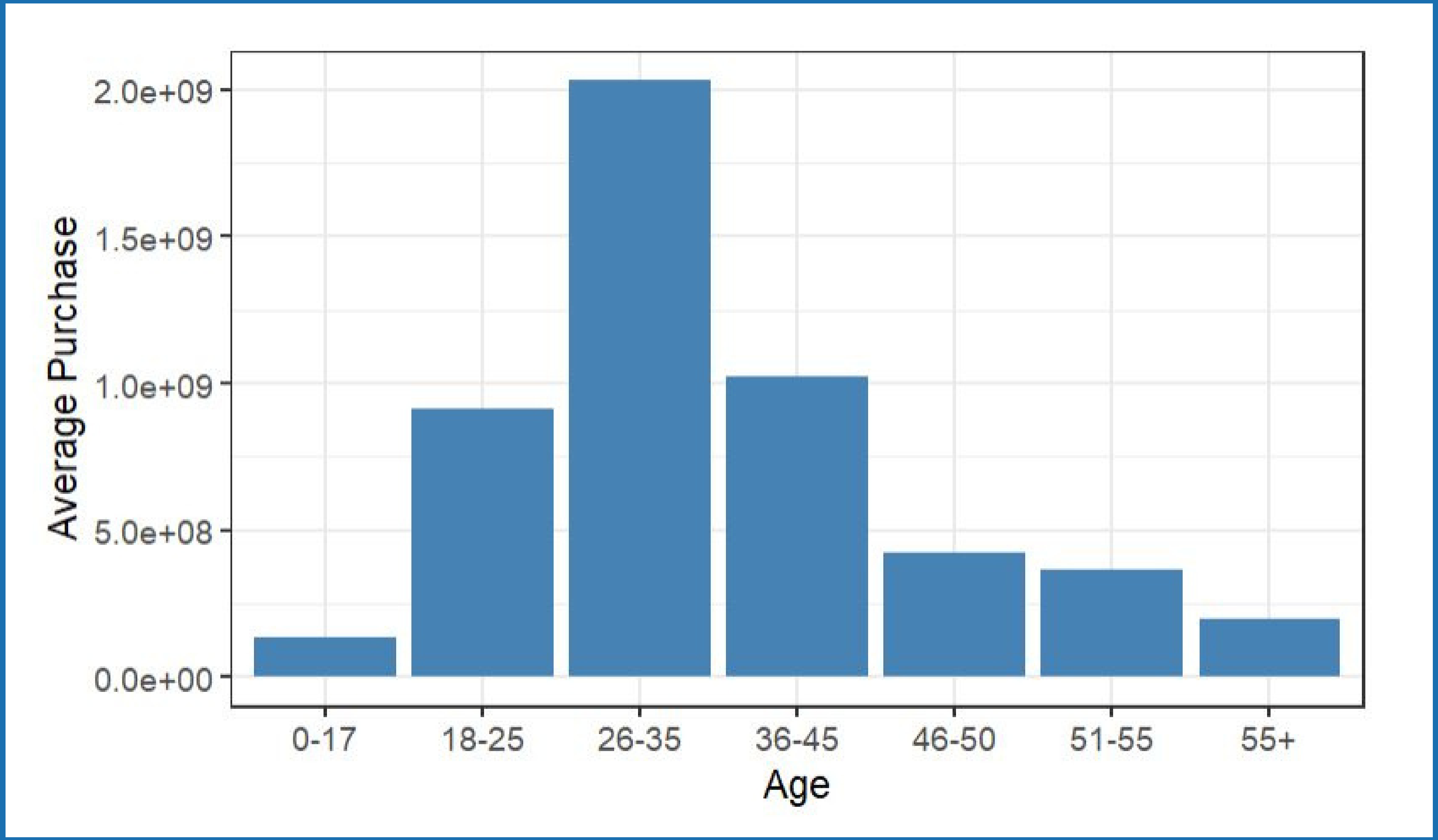
GENDER VS PURCHASE





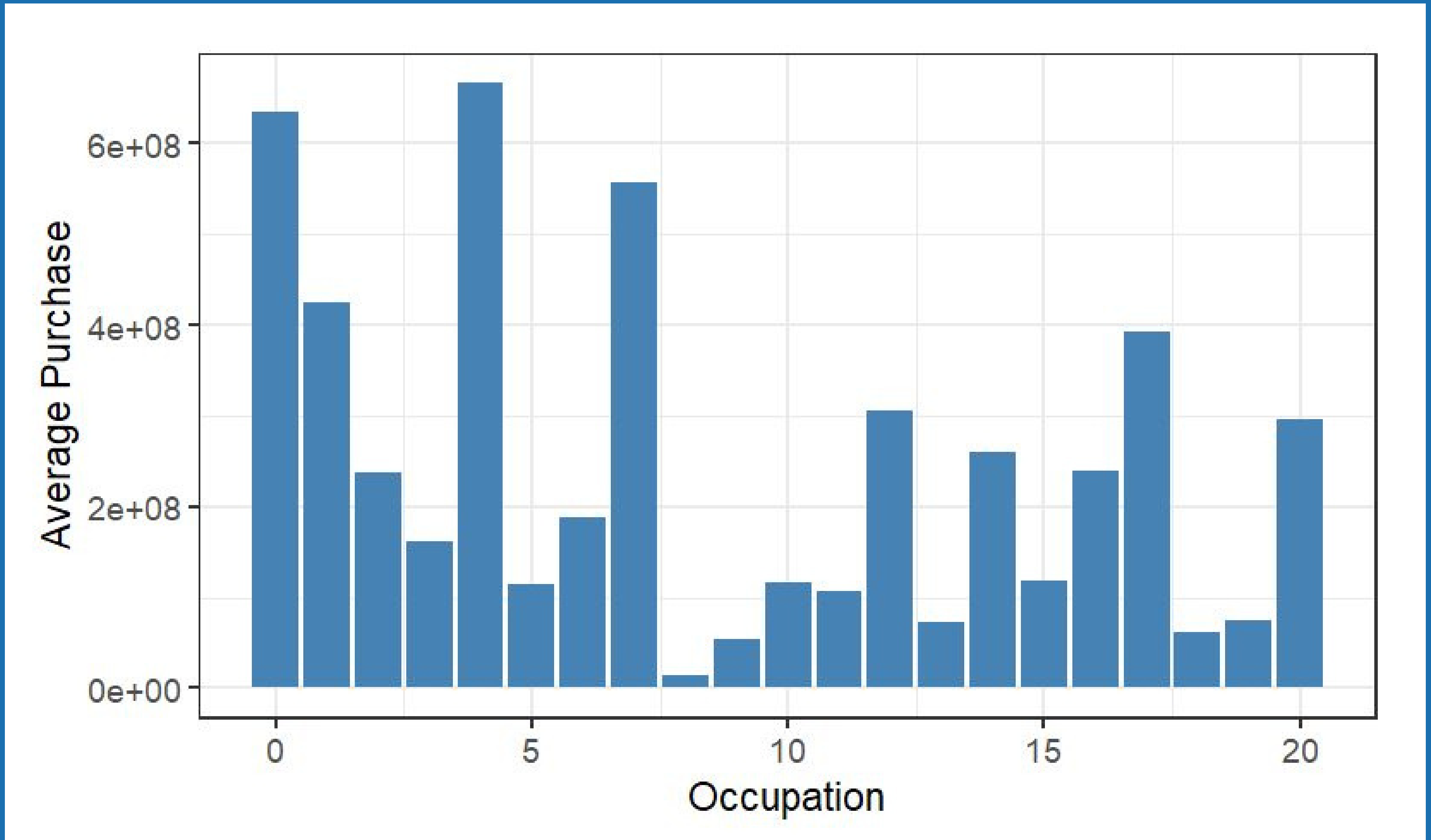
MALE CUSTOMERS MAKE MORE PURCHASES THAN FEMALE CUSTOMERS

AGE VS PURCHASE



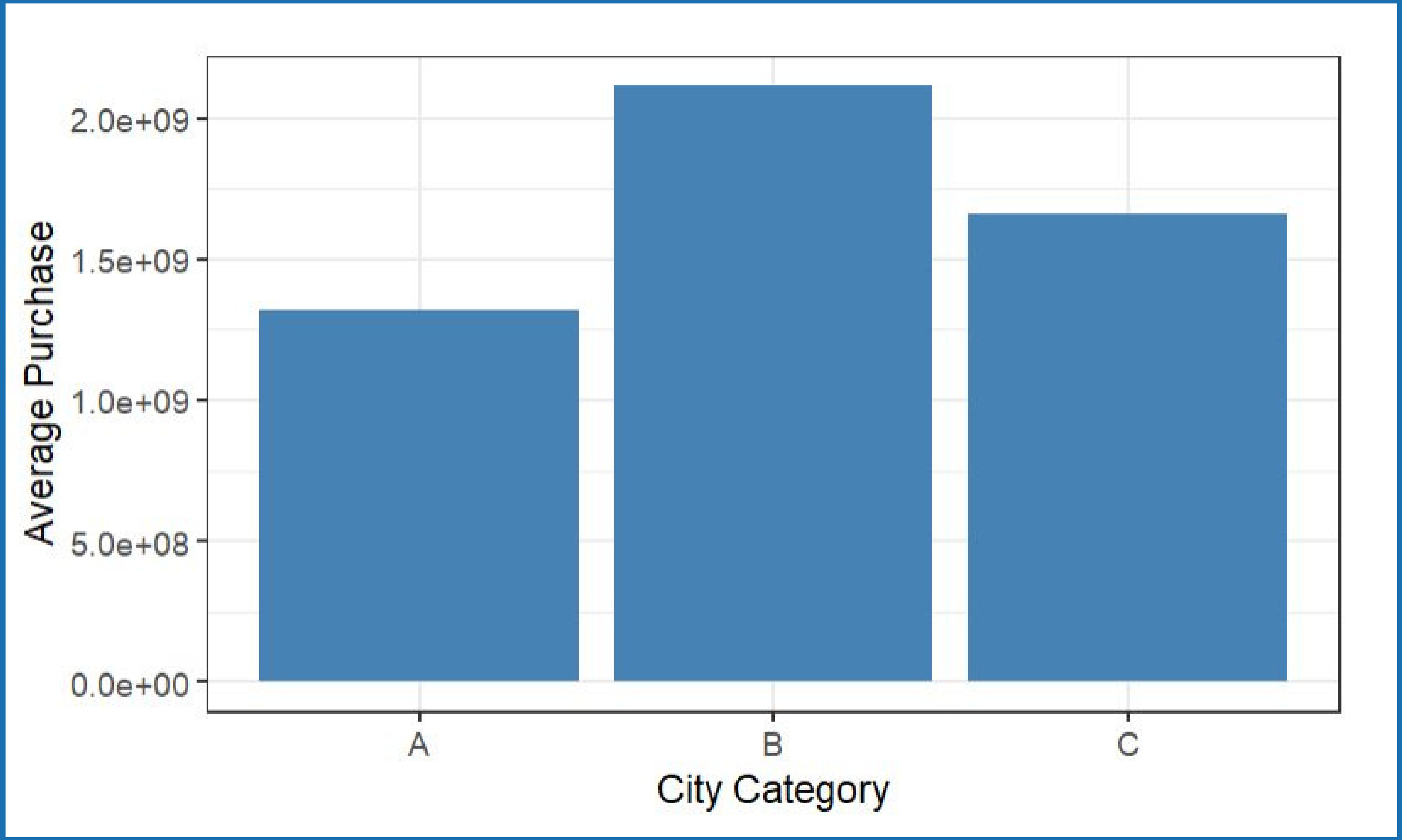
INFORMATION IS OBTAINED THAT THE AGE RANGE OF 26-35 HAS THE MOST PURCHASES

OCCUPATION VS PURCHASE



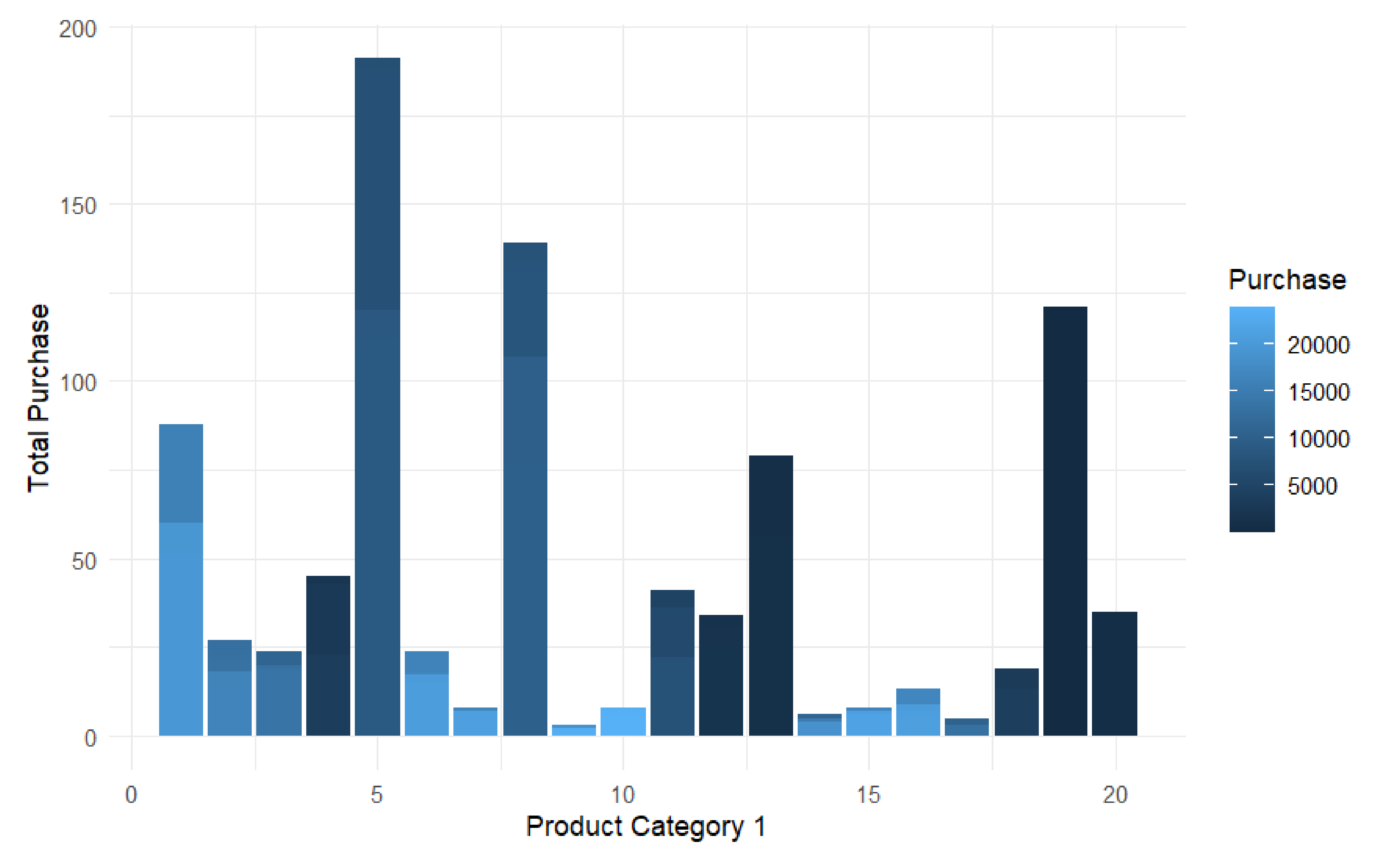
INFORMATION IS OBTAINED THAT OCCUPATION "4" MAKES THE MOST PURCHASES

CITY CATEGORY VS PURCHASE



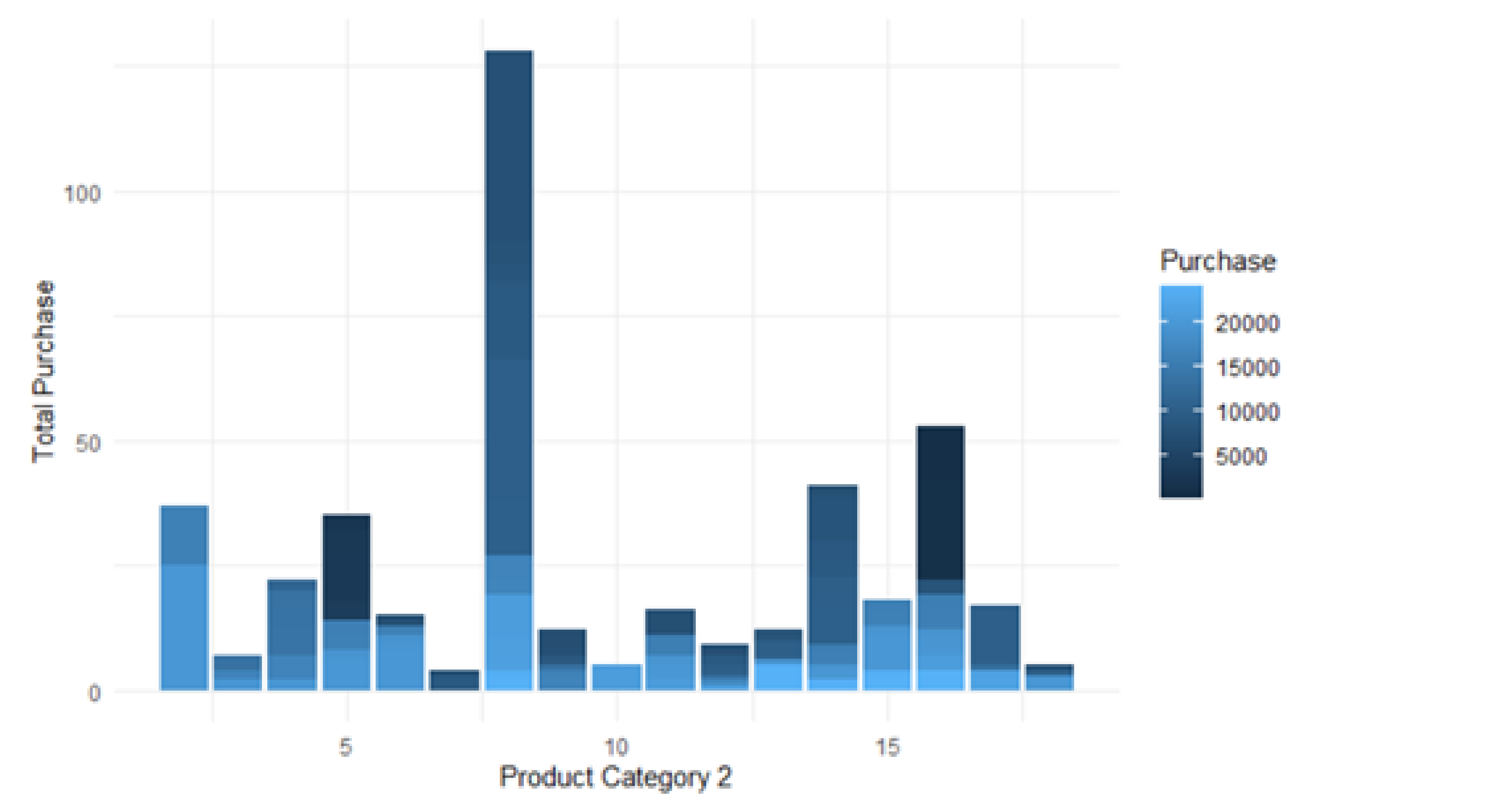
CITY CATEGORY B IS WHERE MOST OF THE CUSTOMERS COME FROM

PRODUCT CATEGORY 1 VS PURCHASE



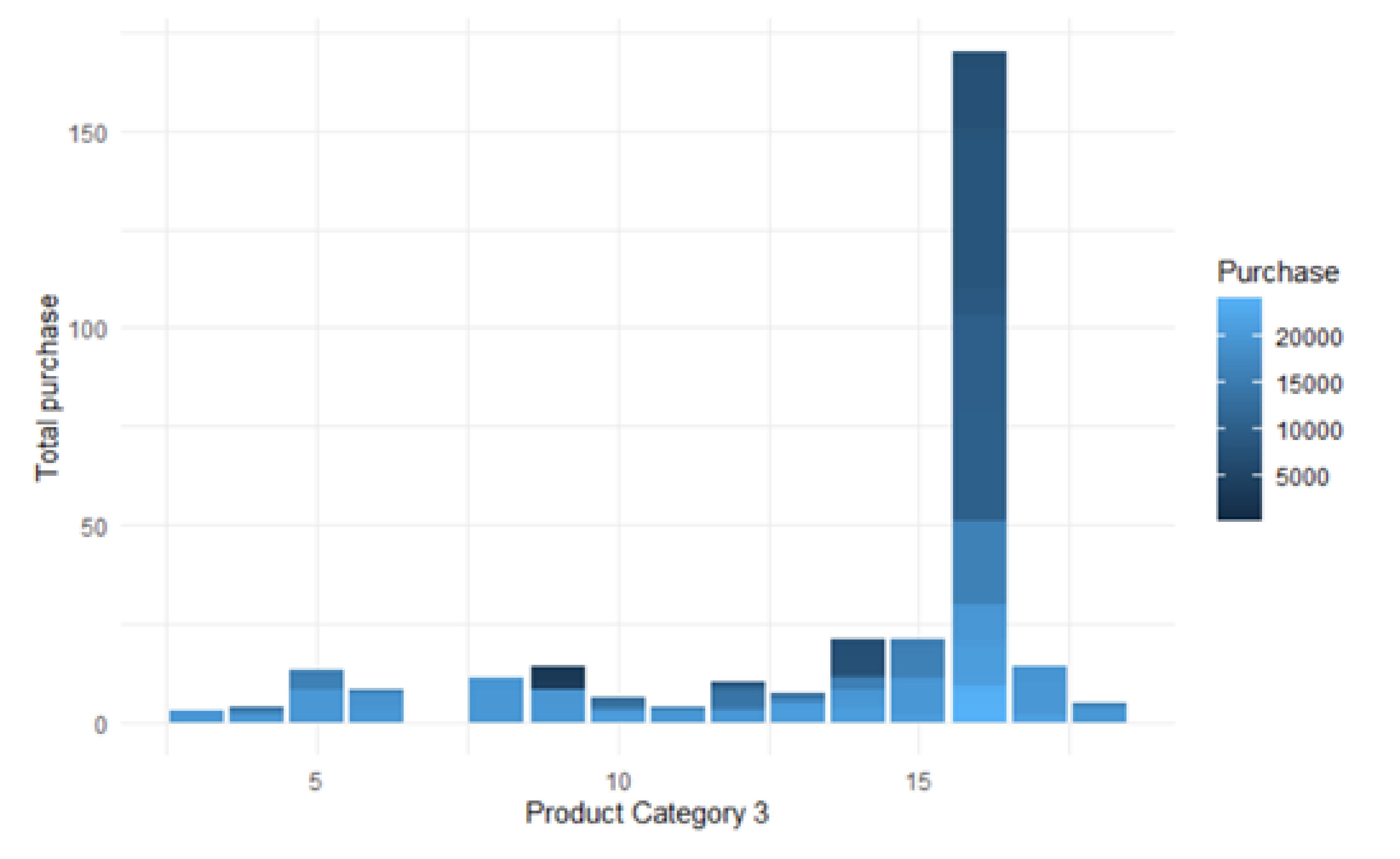
IN PRODUCT CATEGORY 1, PRODUCT "5" HAS THE MOST SALES FOLLOWED BY PRODUCT "8" AND "19"

PRODUCT CATEGORY 2 VS PURCHASE

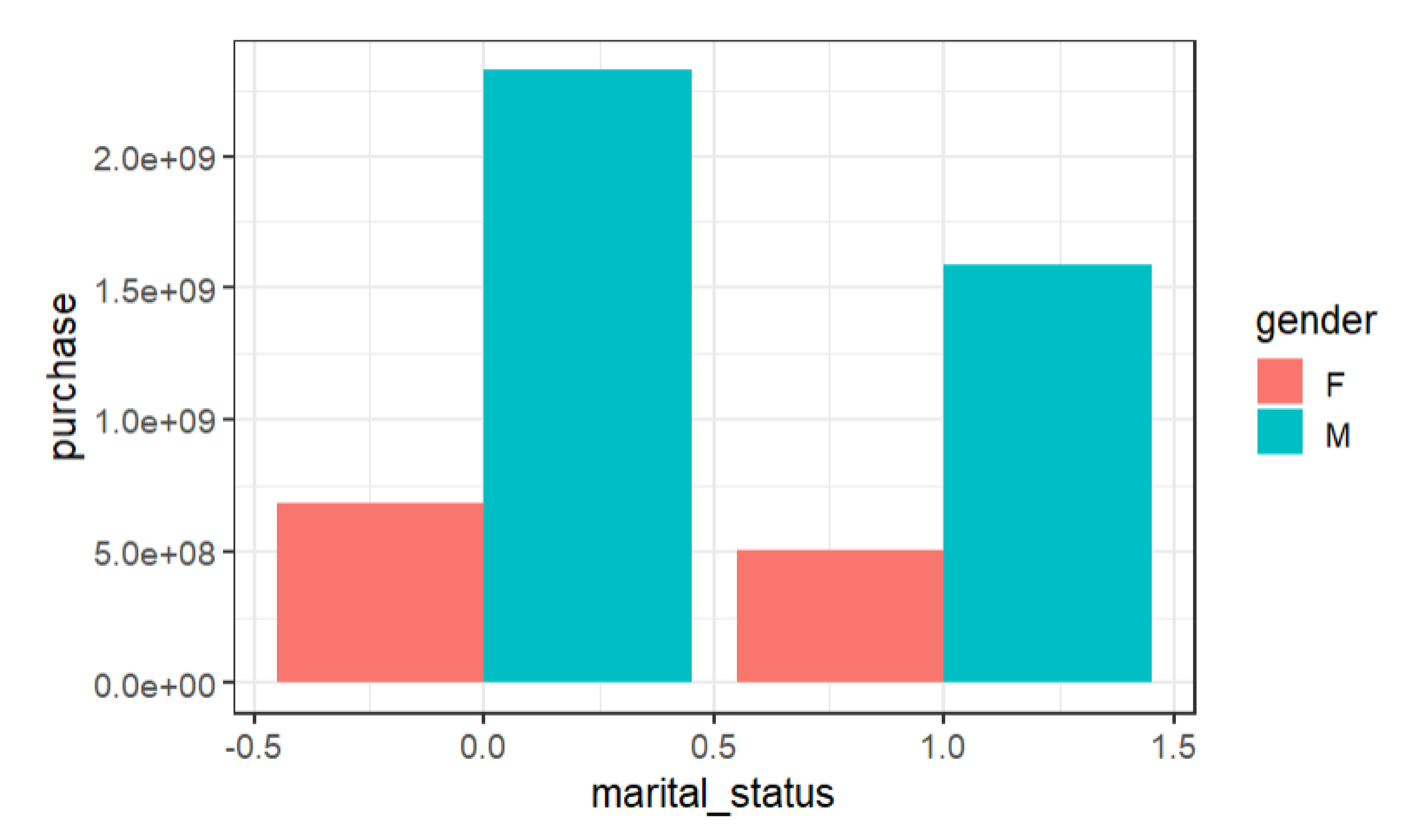


IN PRODUCT CATEGORY 2, PRODUCT "8" IS THE MOST PURCHASED PRODUCT

PRODUCT CATEGORY 3 VS PURCHASE

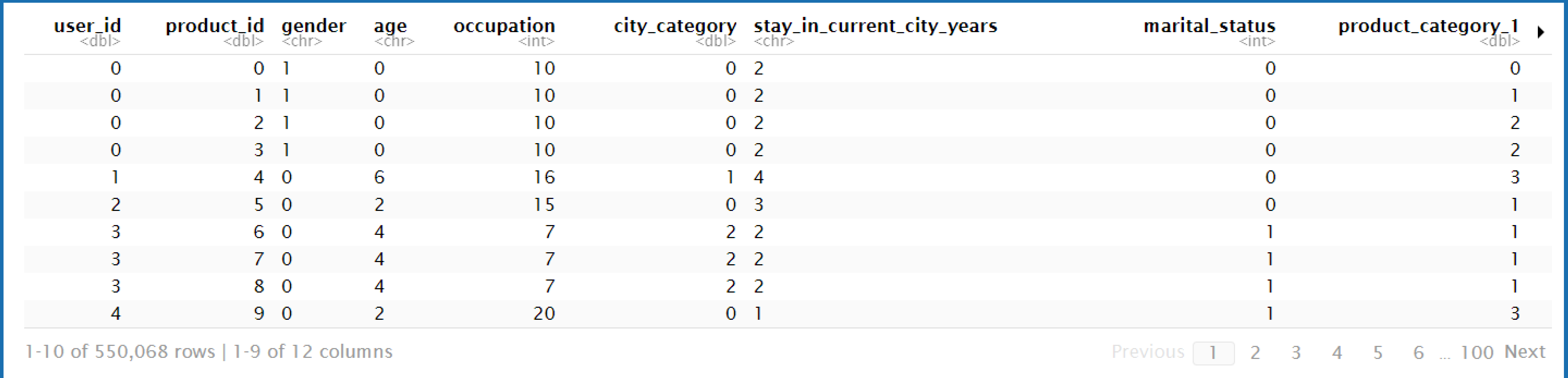


IN PRODUCT CATEGORY 3, PRODUCT "16" IS THE MOST PURCHASED PRODUCT



MARRIED AND UNMARRIED PEOPLE DO NOT HAVE A SIGNIFICANT DIFFERENCE

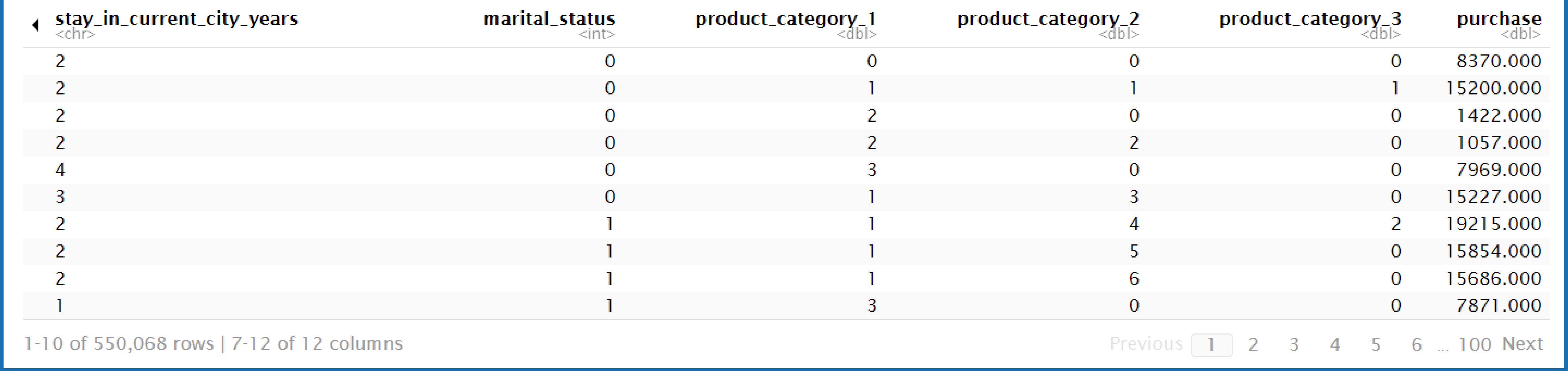
DATA ENCODING



user_id	product_id	gender	age	occupation	city_category	stay_in_current_city_years	marital_status	product_category_1
<dbl>	<dbl>	<chr>	<chr>	<int>	<dbl>	<chr>	<int>	<dbl>
0	0	1	0	10	0	2	0	0
0	1	1	0	10	0	2	0	1
0	2	1	0	10	0	2	0	2
0	3	1	0	10	0	2	0	2
1	4	0	6	16	1	4	0	3
2	5	0	2	15	0	3	0	1
3	6	0	4	7	2	2	1	1
3	7	0	4	7	2	2	1	1
3	8	0	4	7	2	2	1	1
4	9	0	2	20	0	1	1	3

1-10 of 550,068 rows | 1-9 of 12 columns

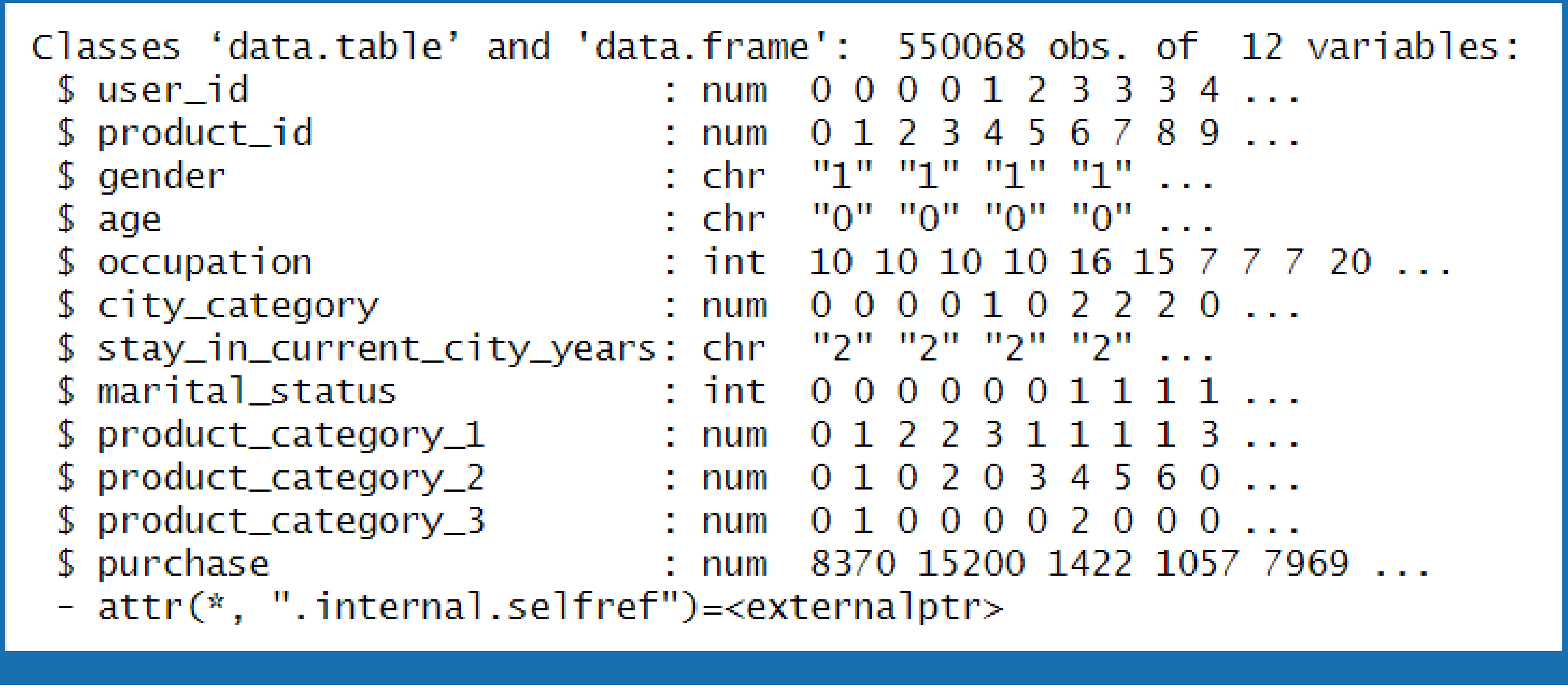
Previous 1 2 3 4 5 6 ... 100 Next



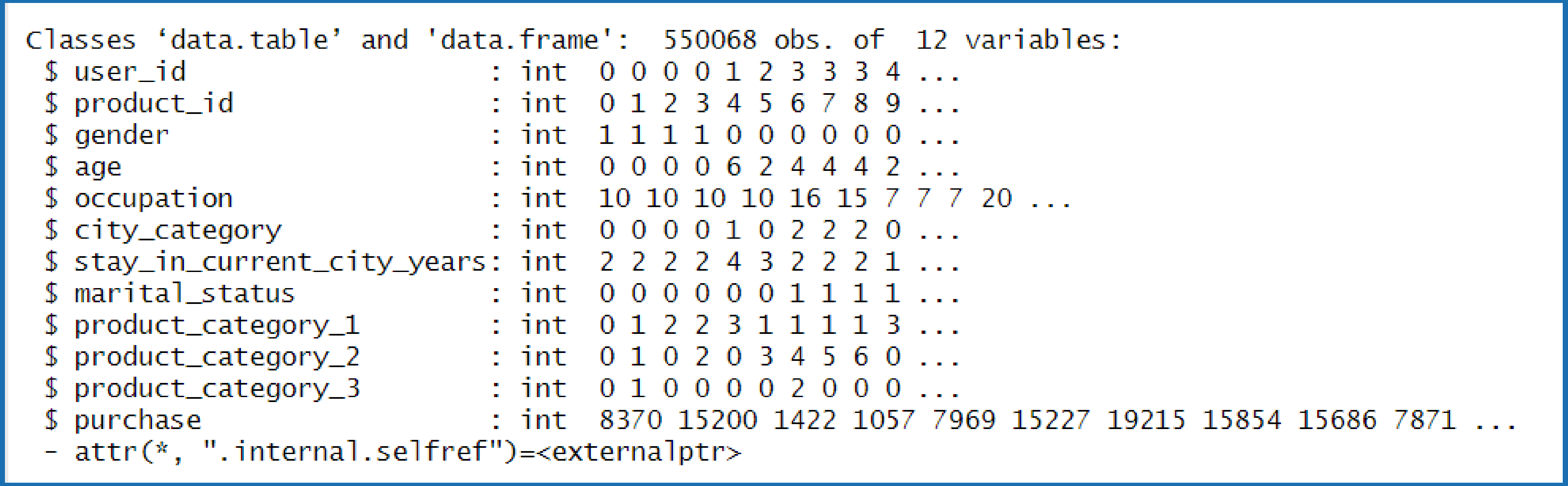
stay_in_current_city_years	marital_status	product_category_1	product_category_2	product_category_3	purchase
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
2	0	0	0	0	8370.000
2	0	1	1	1	15200.000
2	0	2	0	0	1422.000
2	0	2	2	0	1057.000
4	0	3	0	0	7969.000
3	0	1	3	0	15227.000
2	1	1	4	2	19215.000
2	1	1	5	0	15854.000
2	1	1	6	0	15686.000
1	1	3	0	0	7871.000

1-10 of 550,068 rows | 7-12 of 12 columns

Previous 1 2 3 4 5 6 ... 100 Next

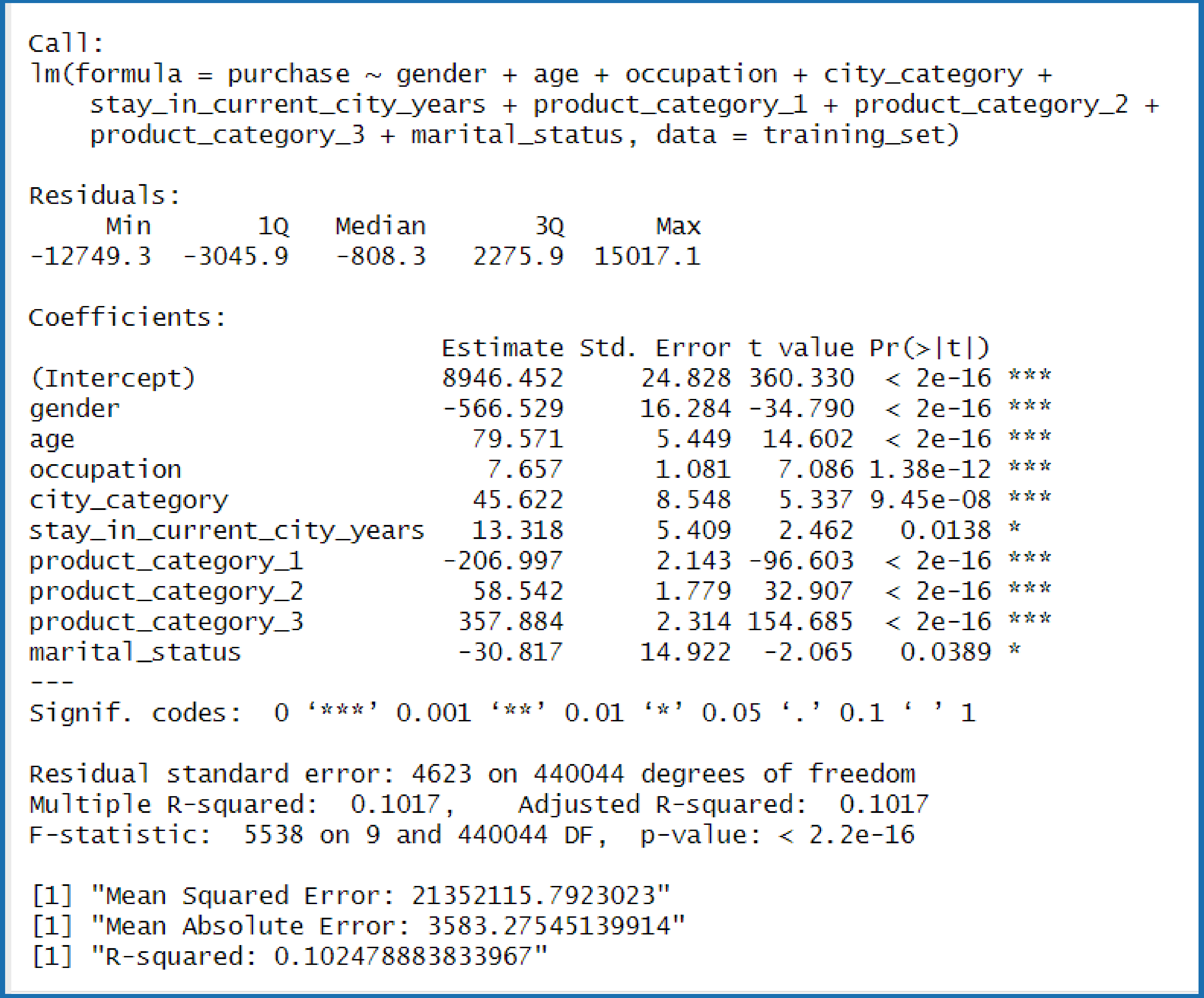


```
Classes ‘data.table’ and 'data.frame':  550068 obs. of  12 variables:
 $ user_id                 : num  0 0 0 0 1 2 3 3 3 4 ...
 $ product_id              : num  0 1 2 3 4 5 6 7 8 9 ...
 $ gender                  : chr  "1" "1" "1" "1" ...
 $ age                     : chr  "0" "0" "0" "0" ...
 $ occupation              : int  10 10 10 10 16 15 7 7 7 20 ...
 $ city_category           : num  0 0 0 0 1 0 2 2 2 0 ...
 $ stay_in_current_city_years: chr  "2" "2" "2" "2" ...
 $ marital_status          : int  0 0 0 0 0 0 1 1 1 1 ...
 $ product_category_1      : num  0 1 2 2 3 1 1 1 1 3 ...
 $ product_category_2      : num  0 1 0 2 0 3 4 5 6 0 ...
 $ product_category_3      : num  0 1 0 0 0 0 2 0 0 0 ...
 $ purchase                : num  8370 15200 1422 1057 7969 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

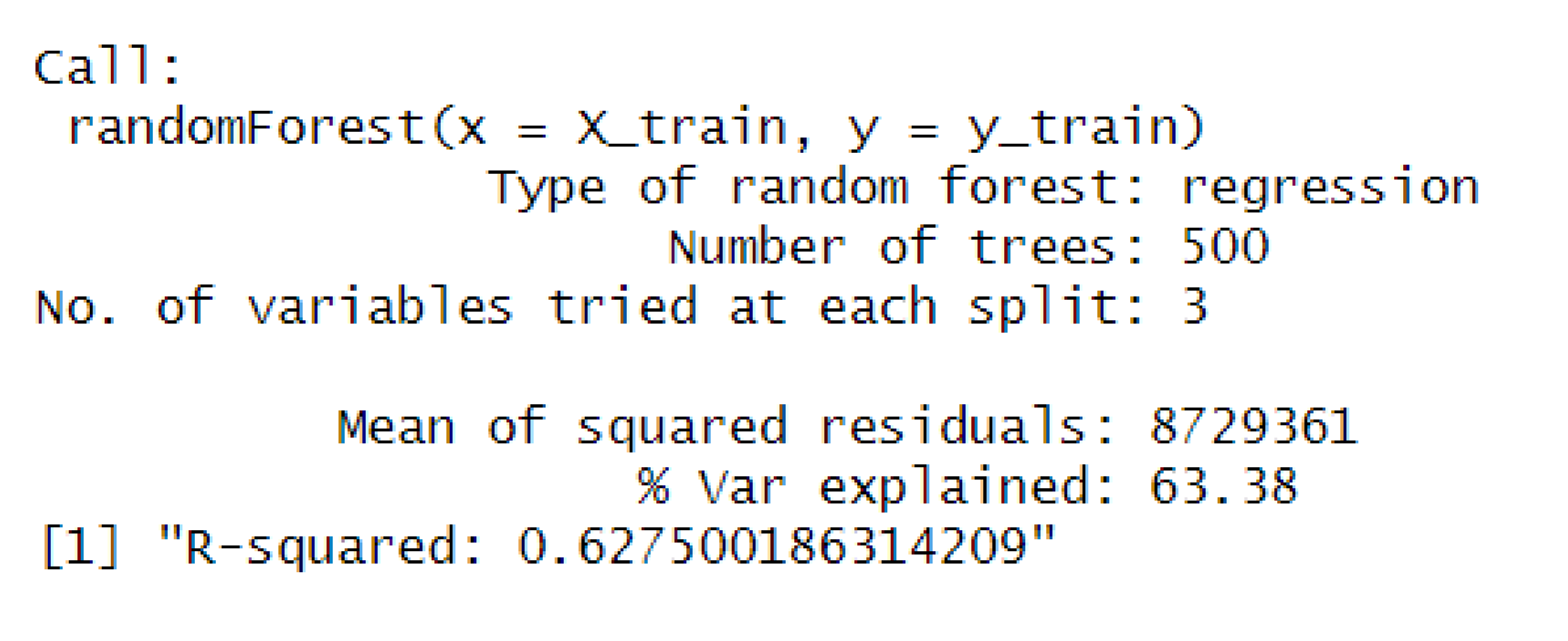


```
Classes ‘data.table’ and 'data.frame':  550068 obs. of  12 variables:
 $ user_id                 : int  0 0 0 0 1 2 3 3 3 4 ...
 $ product_id              : int  0 1 2 3 4 5 6 7 8 9 ...
 $ gender                  : int  1 1 1 1 0 0 0 0 0 0 ...
 $ age                     : int  0 0 0 0 6 2 4 4 4 2 ...
 $ occupation              : int  10 10 10 10 16 15 7 7 7 20 ...
 $ city_category           : int  0 0 0 0 1 0 2 2 2 0 ...
 $ stay_in_current_city_years: int  2 2 2 2 4 3 2 2 2 1 ...
 $ marital_status          : int  0 0 0 0 0 0 1 1 1 1 ...
 $ product_category_1      : int  0 1 2 2 3 1 1 1 1 3 ...
 $ product_category_2      : int  0 1 0 2 0 3 4 5 6 0 ...
 $ product_category_3      : int  0 1 0 0 0 0 2 0 0 0 ...
 $ purchase                : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

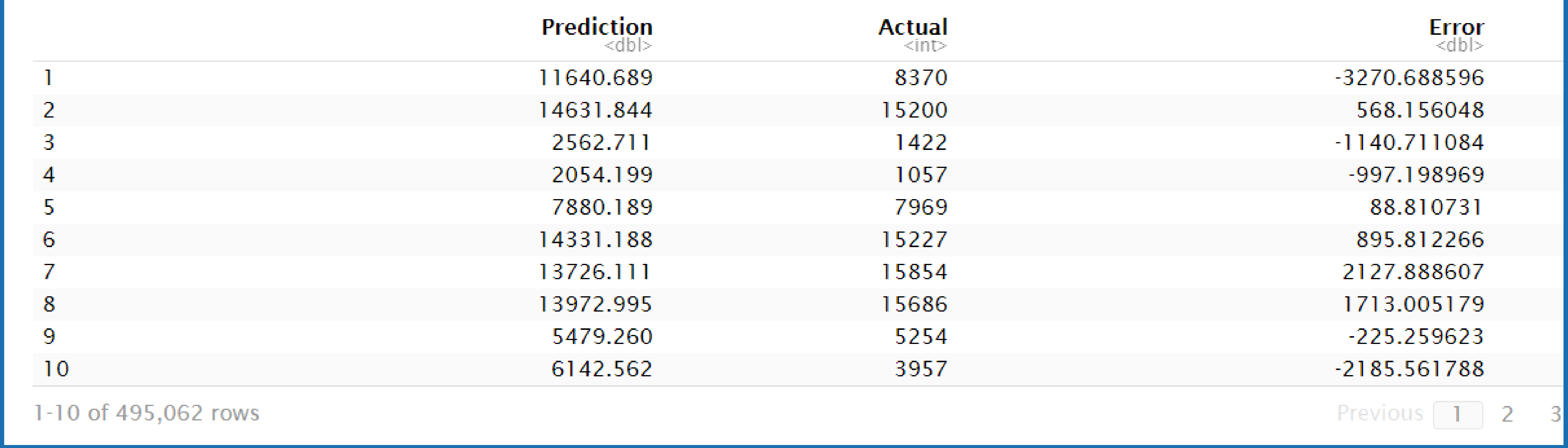
PREDICTIVE MODELING



```
Call:
lm(formula = purchase ~ gender + age + occupation + city_category +
    stay_in_current_city_years + product_category_1 + product_category_2 +
    product_category_3 + marital_status, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-12749.3  -3045.9   -808.3   2275.9  15017.1

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                8946.452     24.828 360.330  < 2e-16 ***
gender                     -566.529     16.284 -34.790  < 2e-16 ***
age                          79.571      5.449  14.602  < 2e-16 ***
occupation                    7.657      1.081   7.086 1.38e-12 ***
city_category                45.622      8.548   5.337 9.45e-08 ***
stay_in_current_city_years   13.318      5.409   2.462   0.0138 *
product_category_1         -206.997      2.143 -96.603  < 2e-16 ***
product_category_2           58.542      1.779  32.907  < 2e-16 ***
product_category_3          357.884      2.314 154.685  < 2e-16 ***
marital_status              -30.817     14.922  -2.065   0.0389 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4623 on 440044 degrees of freedom
Multiple R-squared:  0.1017,    Adjusted R-squared:  0.1017
F-statistic:  5538 on 9 and 440044 DF,  p-value: < 2.2e-16

[1] "Mean Squared Error: 21352115.7923023"
[1] "Mean Absolute Error: 3583.27545139914"
[1] "R-squared: 0.102478883833967"
```



```
Call:
 randomForest(x = X_train, y = y_train) 
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 8729361
                    % Var explained: 63.38
[1] "R-squared: 0.627500186314209"
```



	Prediction <dbl>	Actual <int>	Error <dbl>
1	11640.689	8370	-3270.688596
2	14631.844	15200	568.156048
3	2562.711	1422	-1140.711084
4	2054.199	1057	-997.198969
5	7880.189	7969	88.810731
6	14331.188	15227	895.812266
7	13726.111	15854	2127.888607
8	13972.995	15686	1713.005179
9	5479.260	5254	-225.259623
10	6142.562	3957	-2185.561788

1-10 of 495,062 rows