Black_Friday_Sales_Analysis.R

Bibob

Fri May 18 11:45:20 2018

```
                                                # ANALYSIS OF BLACK FRIDAY SALES

# Authors      : Bibobra Alabrah
# Project Goal: To Understand the Customers Purchase Behavior
# Date         : 04/20/2018

# Set Work Directory
setwd("C:/Data Analysis Projects/Black Friday Sales Analysis")

# Load Packages
library(data.table)
library(DataExplorer) # For initial exploratory data analysis
library(dplyr) # For data manipulation
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(xda) # For Exploratory data analysis
library(ggplot2)
library(vcd)
```

```
## Loading required package: grid
```

```
library(rpart)

# Load The Dataset
blackf_data <- fread("sales.csv")

# STEP 1: DATA PROFILING

# Basic Statistics: The following questions will be explored:

    # A. What is the size of the dataset?
object.size(blackf_data) # The data size is 37.6 MB
```

```
## 37641504 bytes
```

```
    # B. How many rows and columns are there in the dataset?
dim(blackf_data) # 550068 rows and 12 columns
```

```
## [1] 550068     12
```

```
    # C. What does my dataset look like?
head(blackf_data, 10)
```

```
##     User_ID Product_ID Gender   Age Occupation City_Category
##  1: 1000001  P00069042      F  0-17         10             A
##  2: 1000001  P00248942      F  0-17         10             A
##  3: 1000001  P00087842      F  0-17         10             A
##  4: 1000001  P00085442      F  0-17         10             A
##  5: 1000002  P00285442      M   55+         16             C
##  6: 1000003  P00193542      M 26-35         15             A
##  7: 1000004  P00184942      M 46-50          7             B
##  8: 1000004  P00346142      M 46-50          7             B
##  9: 1000004   P0097242      M 46-50          7             B
## 10: 1000005  P00274942      M 26-35         20             A
##     Stay_In_Current_City_Years Marital_Status Product_Category_1
##  1:                          2              0                 3
##  2:                          2              0                 1
##  3:                          2              0                12
##  4:                          2              0                12
##  5:                         4+              0                 8
##  6:                          3              0                 1
##  7:                          2              1                 1
##  8:                          2              1                 1
##  9:                          2              1                 1
## 10:                          1              1                 8
##     Product_Category_2 Product_Category_3 Purchase
##  1:                 NA                 NA     8370
##  2:                  6                 14    15200
##  3:                 NA                 NA     1422
##  4:                 14                 NA     1057
##  5:                 NA                 NA     7969
##  6:                  2                 NA    15227
##  7:                  8                 17    19215
##  8:                 15                 NA    15854
##  9:                 16                 NA    15686
## 10:                 NA                 NA     7871
```

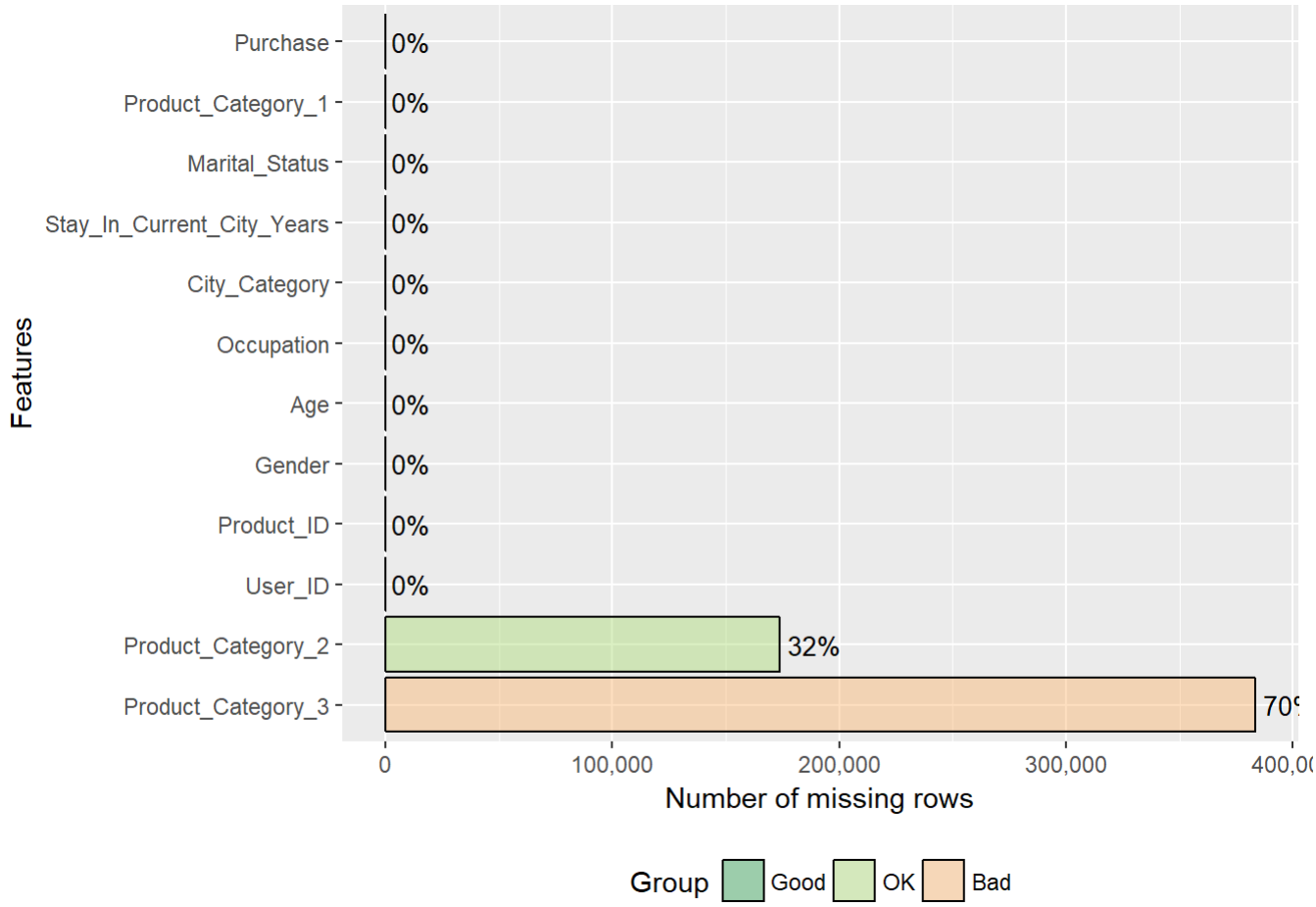
```
# From the first 10 rows it is evident that, the same user ID is repeated the number of times pu
rchases were made
# The Age variable is reported as a range of values
#There are 3 city categories namely A, B, & C
# The stay in current ranges from 1 - 4 years
# The customers comprised of both married (1) and singles (0)
# There are 3 different product categories available to customers

    # D. What is the structure of the data?
str(blackf_data) # Text
```

```
## Classes 'data.table' and 'data.frame':   550068 obs. of  12 variables:
##  $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1
000004 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>
```

```
plot_str(blackf_data) # Network Graph
# The dataset are made up of factors and integers
# Marital status should probably be a factor and not an integer, hence, it must be converted
# Purchase, age, and Stay in current city should be numeric
# Product ID, gender should be a factor

    # E. Are there any missing values?
sapply(blackf_data, function(x) sum(is.na(x)))
```

```
##                    User_ID                 Product_ID
##                          0                          0
##                     Gender                        Age
##                          0                          0
##                 Occupation              City_Category
##                          0                          0
## Stay_In_Current_City_Years            Marital_Status
##                          0                          0
##         Product_Category_1         Product_Category_2
##                          0                     173638
##         Product_Category_3                   Purchase
##                     383247                          0
```

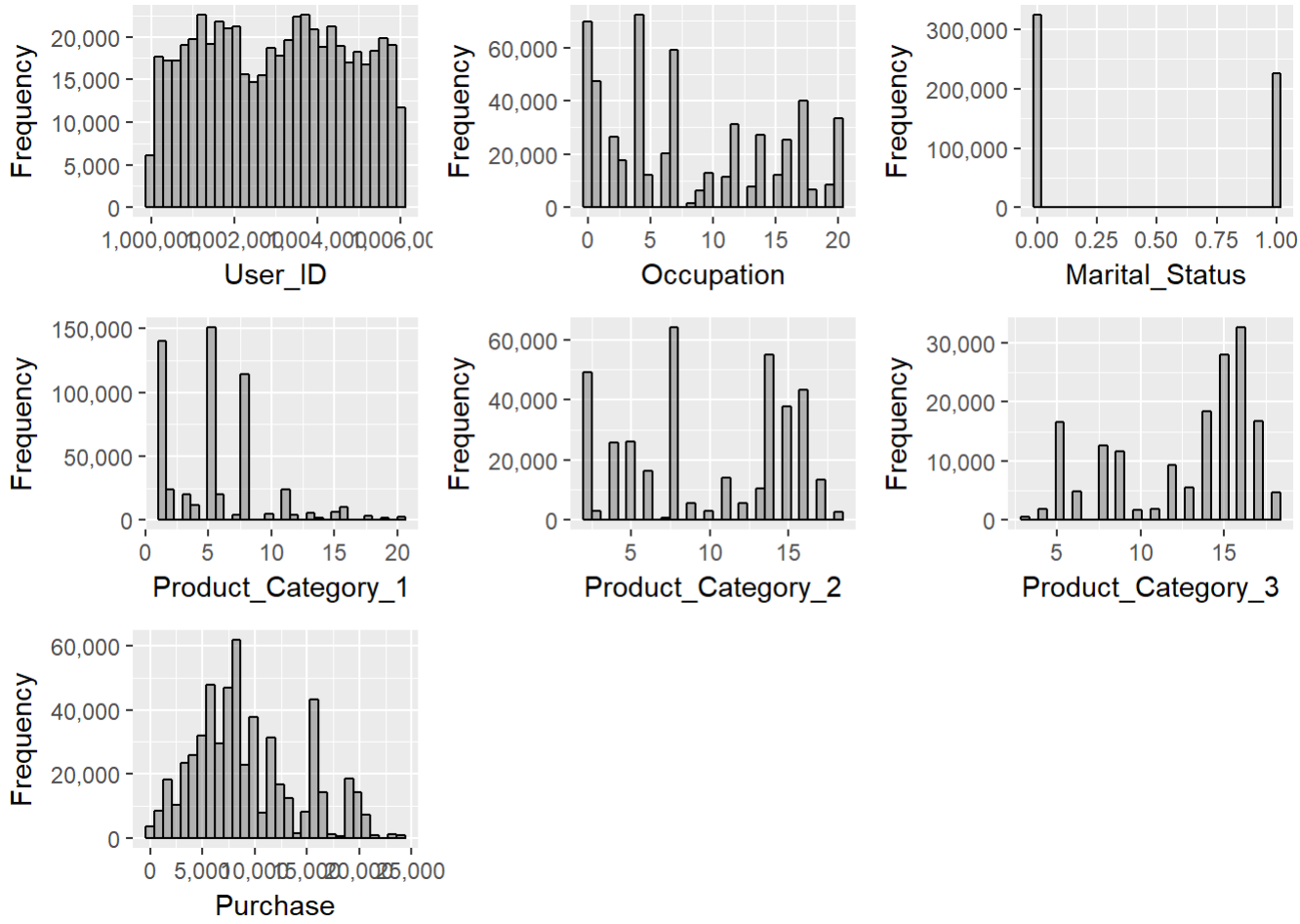
```
plot_missing(blackf_data)
```




```
    # F. What is the data distribution? i.e. what are the continous and discrete features?
# Continous Features
  #Did the married visited the store more?
prop.table(table(blackf_data$Marital_Status))
```

```
##
##        0        1
## 0.590347 0.409653
```

```
plot_histogram(blackf_data)
```




```
# Discrete Features

  # How many male and female customers?
prop.table(table(blackf_data$Gender))
```

```
##
##         F         M
## 0.2468949 0.7531051
```

```
  # Customers age group
prop.table(table(blackf_data$Age))
```

```
##
##       0-17      18-25      26-35      36-45      46-50      51-55
## 0.02745479 0.18117760 0.39919974 0.19999891 0.08308246 0.06999316
##        55+
## 0.03909335
```

```
  # Which city category has the highest customers?
prop.table(table(blackf_data$City_Category))
```

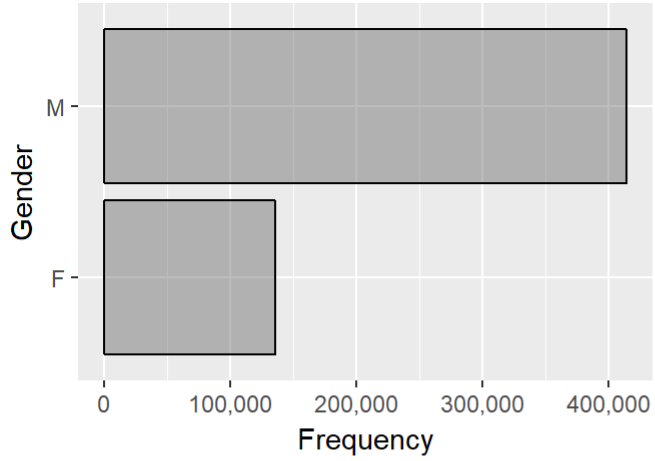
```
##
##         A         B         C
## 0.2685486 0.4202626 0.3111888
```

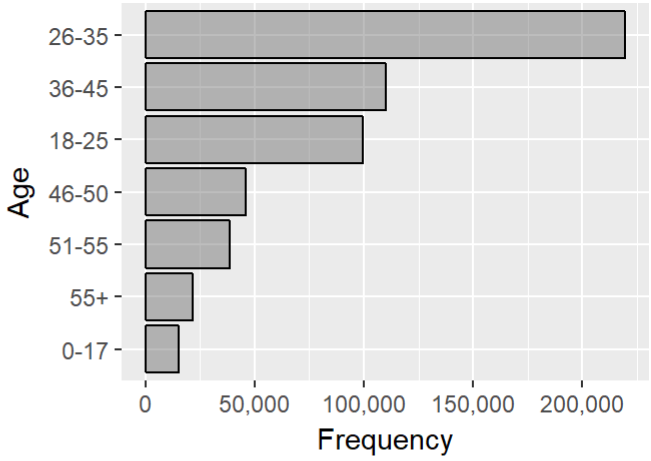
```
  # What effect does the duration of stay in current has on the store visit?
prop.table(table(blackf_data$Stay_In_Current_City_Years))
```

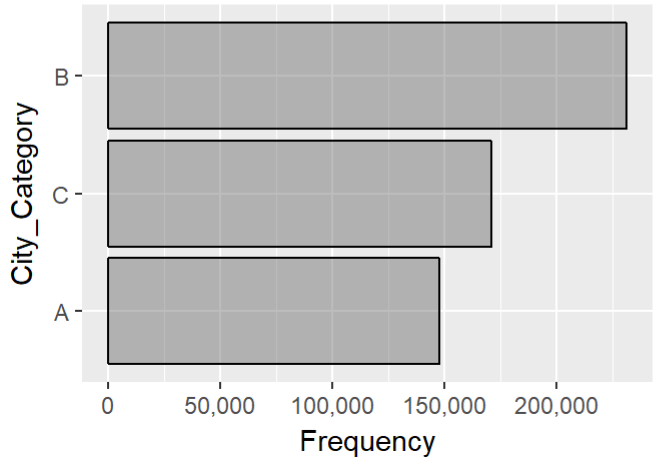
```
##
##         0         1         2         3        4+
## 0.1352524 0.3523583 0.1851371 0.1732240 0.1540282
```

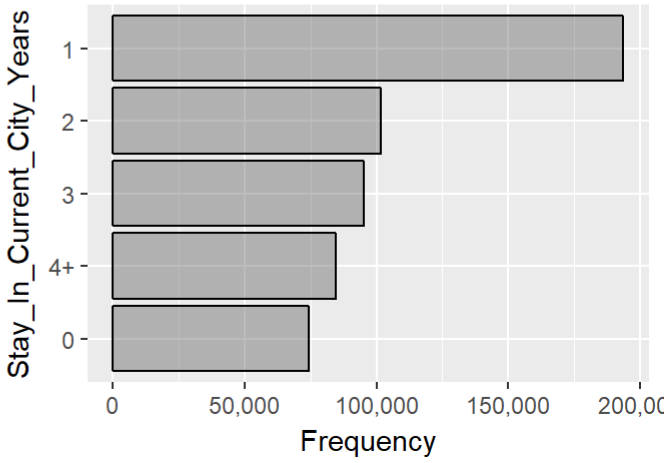
```
plot_bar(blackf_data)
```

```
## 1 columns ignored with more than 50 categories.
## Product_ID: 3631 categories
```

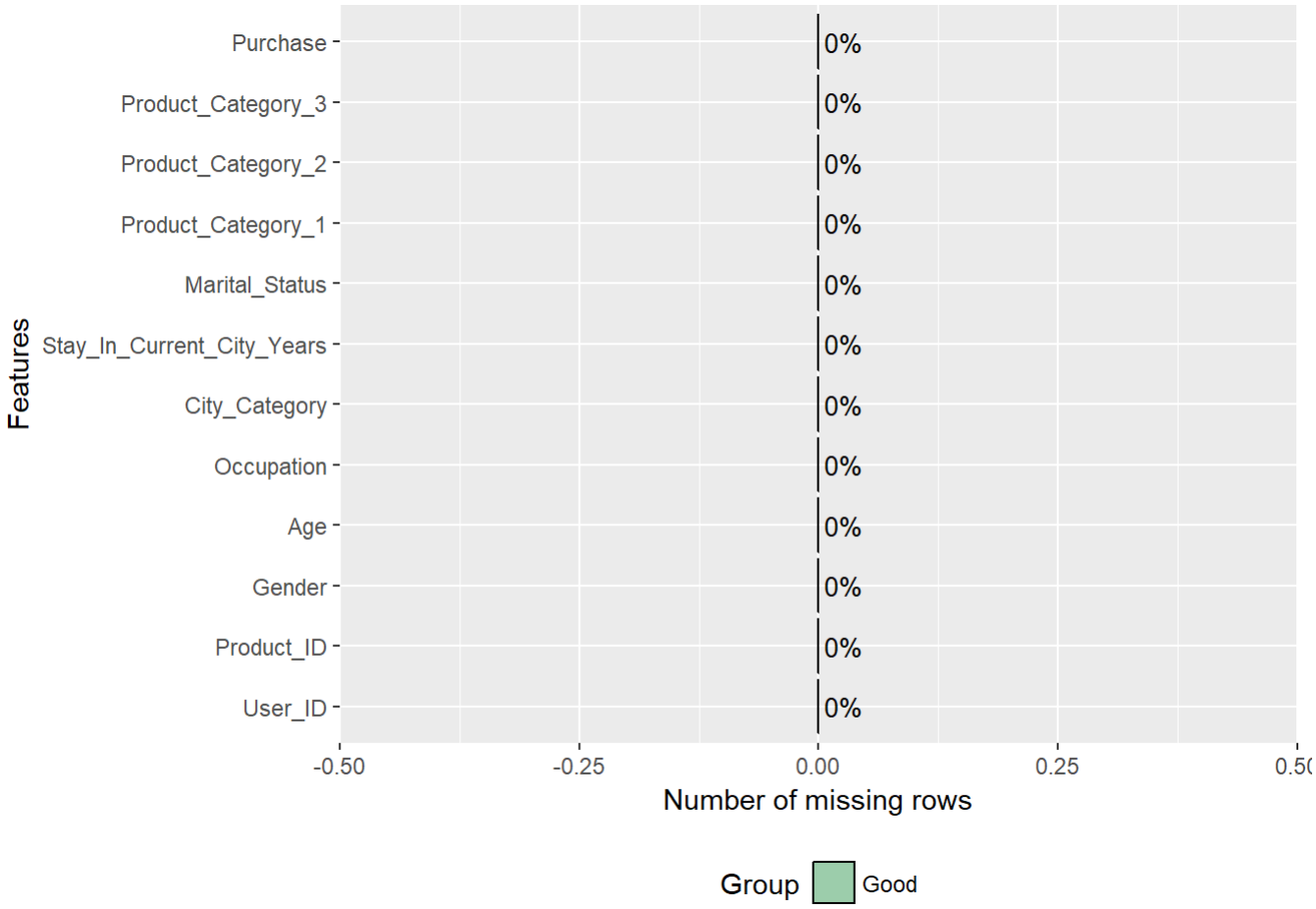








```
# STEP 2: DATA CLEANING AND TRANSFORMATION

# Transform the data types

blackf_data$User_ID <- as.factor(blackf_data$User_ID)
blackf_data$Product_ID <- as.factor(blackf_data$Product_ID)
blackf_data$Gender <- as.factor(if_else(blackf_data$Gender == 'M', 'Male', 'Female'))
blackf_data$Age <- as.factor(blackf_data$Age)
blackf_data$Occupation <- as.factor(blackf_data$Occupation)
blackf_data$City_Category <- as.factor(blackf_data$City_Category)
blackf_data$Stay_In_Current_City_Years <- as.factor(blackf_data$Stay_In_Current_City_Years)
blackf_data$Marital_Status <- as.factor(if_else(blackf_data$Marital_Status == 1, 'Married', 'Sin
gle'))
blackf_data$Product_Category_1 <- as.integer(blackf_data$Product_Category_1)
blackf_data$Product_Category_2 <- as.integer(blackf_data$Product_Category_2)
blackf_data$Product_Category_3 <- as.integer(blackf_data$Product_Category_3)
blackf_data$Purchase <- as.numeric(blackf_data$Purchase)

# Impute the missing values

fit <- rpart(Product_Category_2 ~ User_ID + Product_ID + Age + Gender,
                                data = blackf_data[!is.na(blackf_data$Product_Category_2),],
                                method = "anova")
blackf_data$Product_Category_2[is.na(blackf_data$Product_Category_2)] <-
  predict(fit, blackf_data[is.na(blackf_data$Product_Category_2),])

fit_1 <- rpart(Product_Category_3 ~ User_ID + Product_ID + Age + Gender,
             data = blackf_data[!is.na(blackf_data$Product_Category_3),],
             method = "anova")
blackf_data$Product_Category_3[is.na(blackf_data$Product_Category_3)] <-
  predict(fit_1, blackf_data[is.na(blackf_data$Product_Category_3),])

# Check for any missing values
plot_missing(blackf_data)
```




```
# Everything is now clean, hence, the analysis can now begin.

# STEP 3: EXPLORATORY DATA ANALYSIS

# How many unique User_IDs are there in the dataset?
length(unique(blackf_data$User_ID))# The store had 5891 customers
```

```
## [1] 5891
```

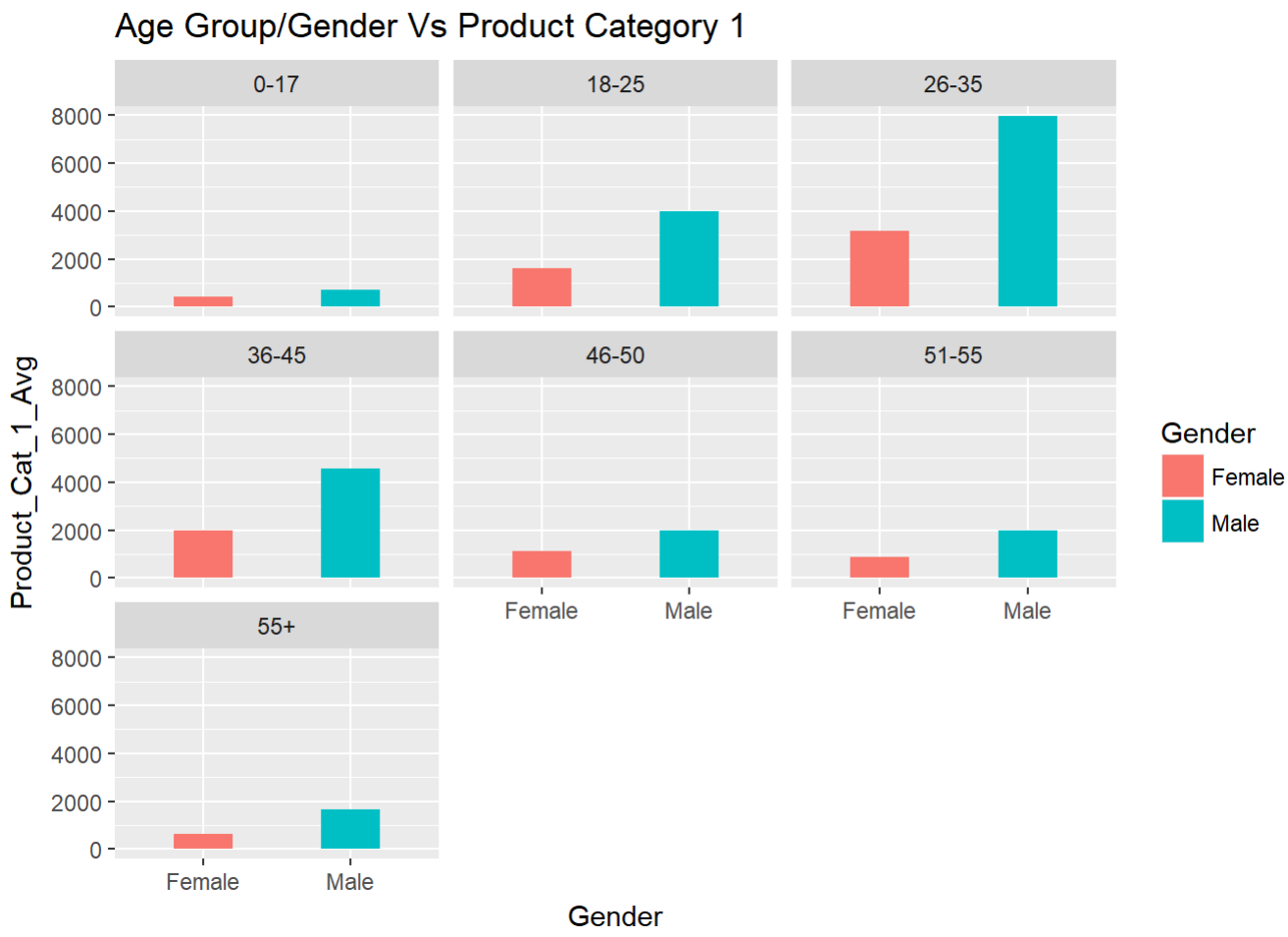
```
# How many items did each customer purchased?
Unique_UserID <- as.data.frame(table(blackf_data$User_ID))
names(Unique_UserID) <- c("User_ID", "Customer_Purchase_Count")
head(Unique_UserID)
```

```
##   User_ID Customer_Purchase_Count
## 1 1000001                      35
## 2 1000002                      77
## 3 1000003                      29
## 4 1000004                      14
## 5 1000005                     106
## 6 1000006                      47
```

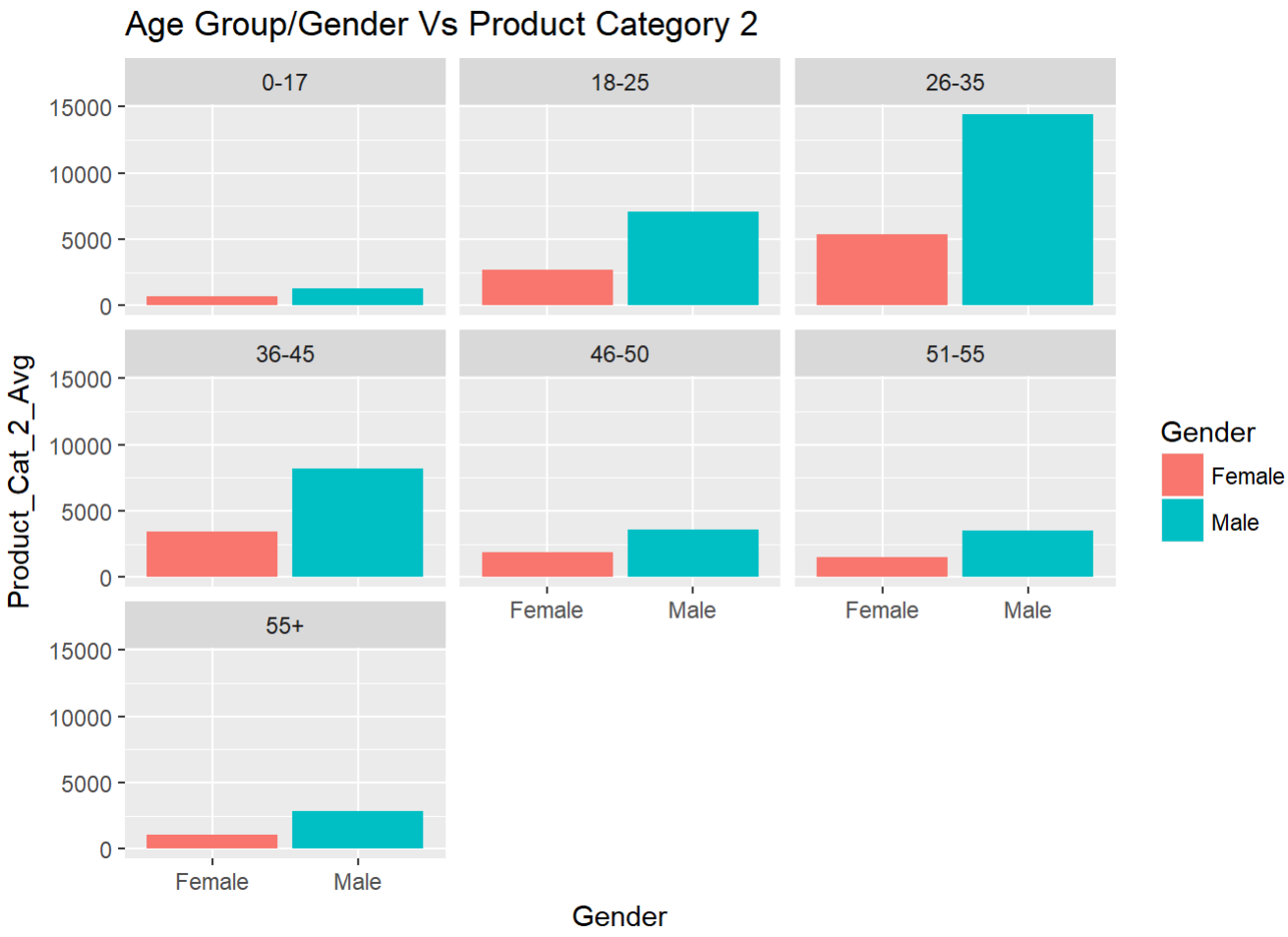
```
# Due to the large dataset, the average values were used for this analysis (using dplyr's chaini
ng method)
new_data <- blackf_data %>%
  group_by(User_ID, Age, Gender, Occupation, City_Category, Stay_In_Current_City_Years, Marital_
Status) %>%
  summarise_each(funs(mean), Product_Category_1, Product_Category_2, Product_Category_3, Purchas
e)
```

```
## `summarise_each()` is deprecated.
## Use `summarise_all()`, `summarise_at()` or `summarise_if()` instead.
## To map `funs` over a selection of variables, use `summarise_at()`
```

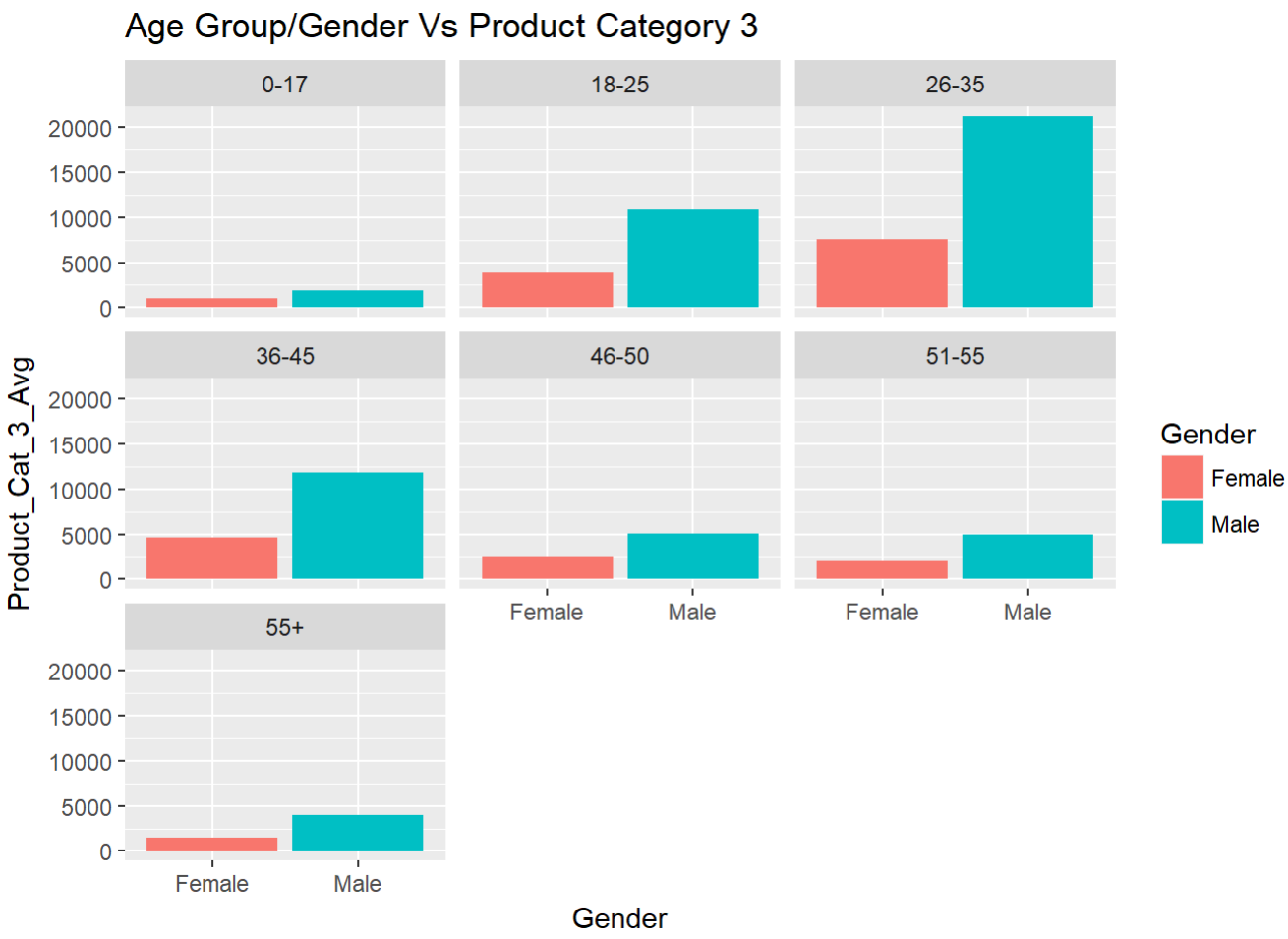
```
# Rename the average values accordingly
colnames(new_data)[8] <- "Product_Cat_1_Avg"
colnames(new_data)[9] <- "Product_Cat_2_Avg"
colnames(new_data)[10] <- "Product_Cat_3_Avg"
colnames(new_data)[11] <- "Avg_Purchase_Amount"

# Explore the age and gender variables versus the product categories and the purchase amount

# 1. Which Age group/gender had the highest purchase by product category?

# Product Category 1:
ggplot(new_data, aes(Gender, Product_Cat_1_Avg, fill = Gender)) + geom_col(width = 0.4) + facet_
wrap(~ Age) +
  labs(title = "Age Group/Gender Vs Product Category 1")
```



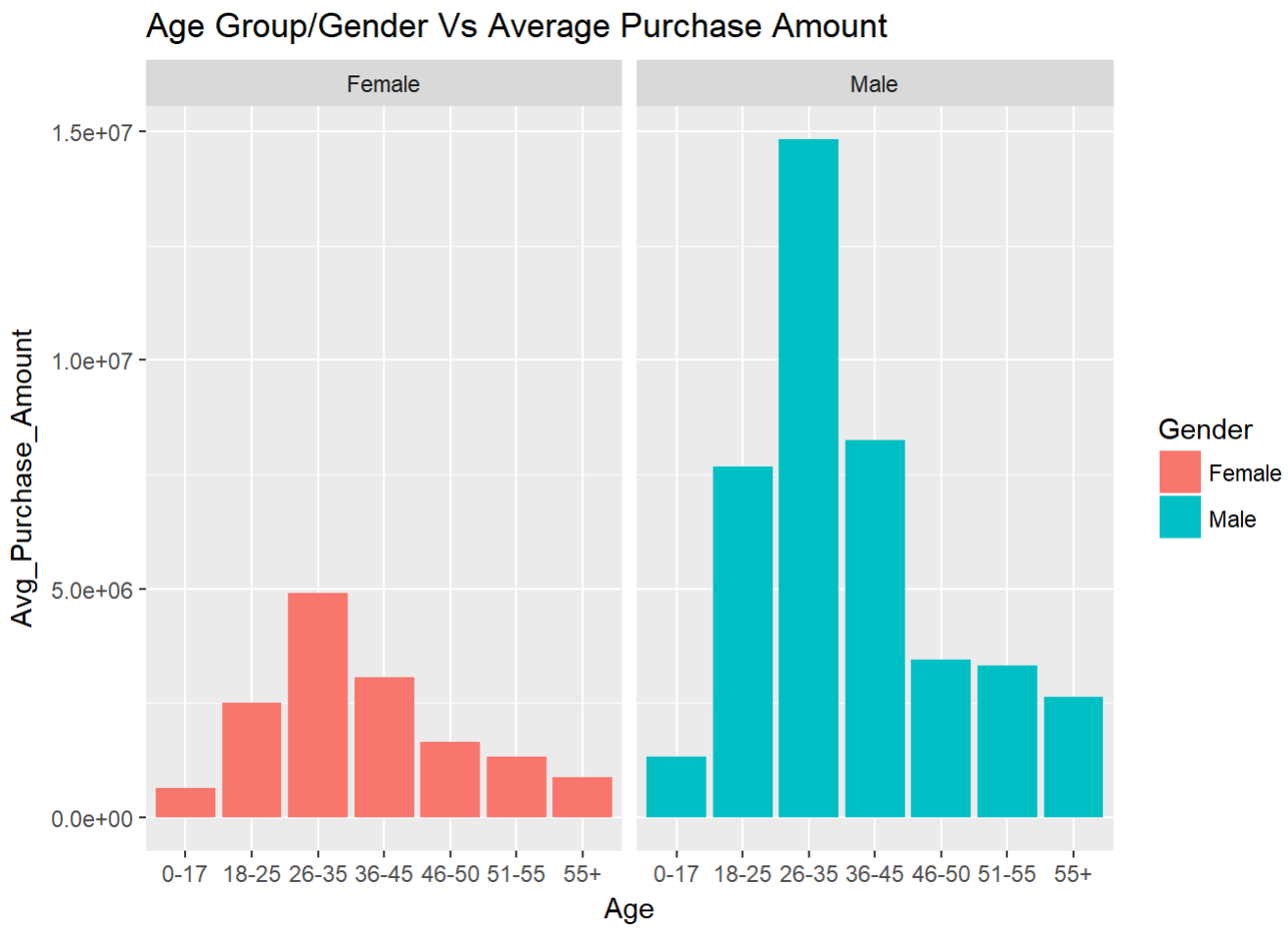

```
# Product Category 2:
ggplot(new_data, aes(Gender, Product_Cat_2_Avg, fill = Gender)) + geom_col() + facet_wrap(~ Age)
 +
  labs(title = "Age Group/Gender Vs Product Category 2")
```



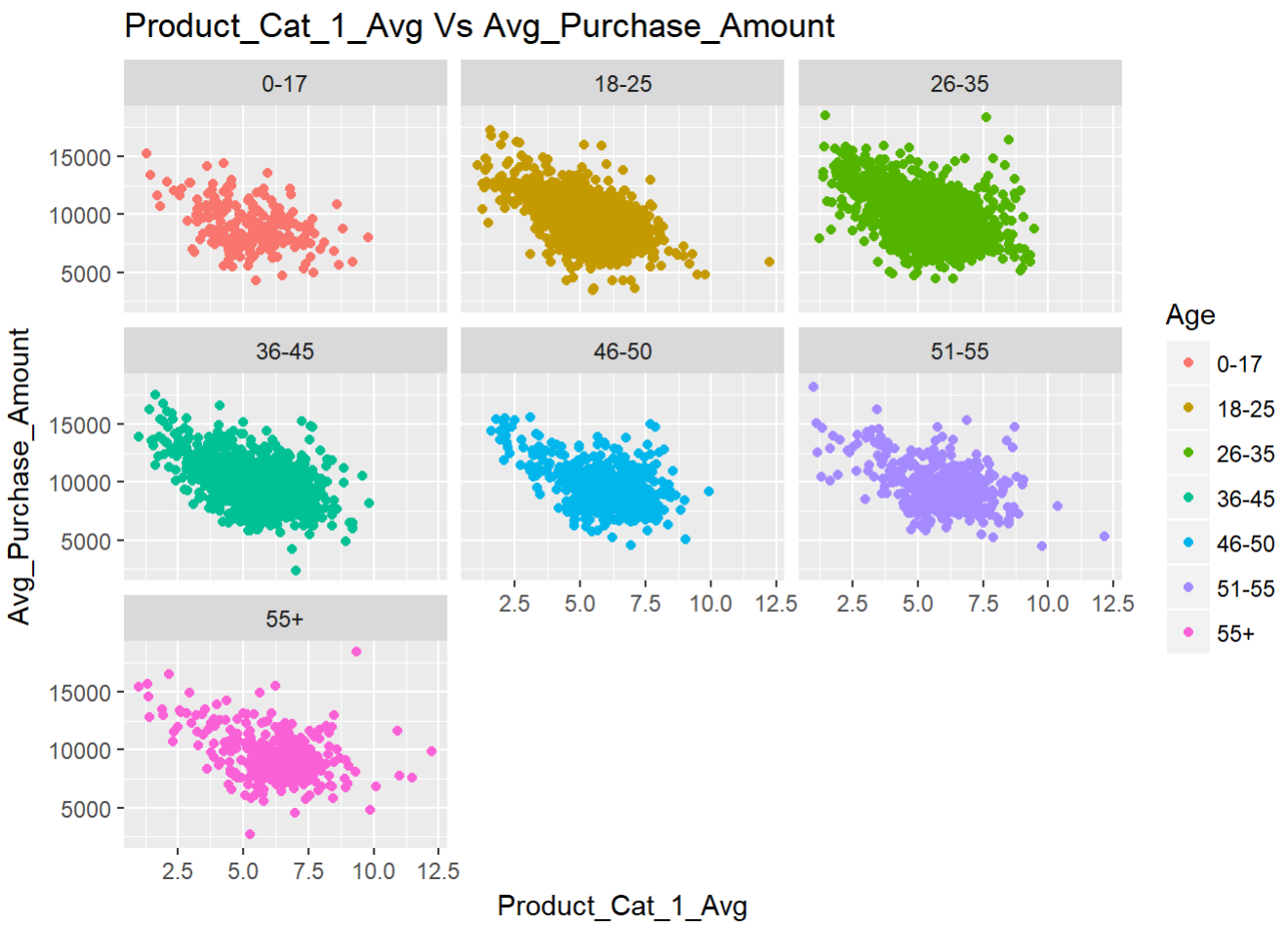

```
# Product Category 3:
ggplot(new_data, aes(Gender, Product_Cat_3_Avg, fill = Gender)) + geom_col() + facet_wrap(~ Age)
 +
  labs(title = "Age Group/Gender Vs Product Category 3")
```




```
# Age group versus Average purchase amount
ggplot(new_data, aes(Age, Avg_Purchase_Amount, fill = Gender)) + geom_col() + facet_wrap(~ Gende
r) +
  labs(title = "Age Group/Gender Vs Average Purchase Amount")
```

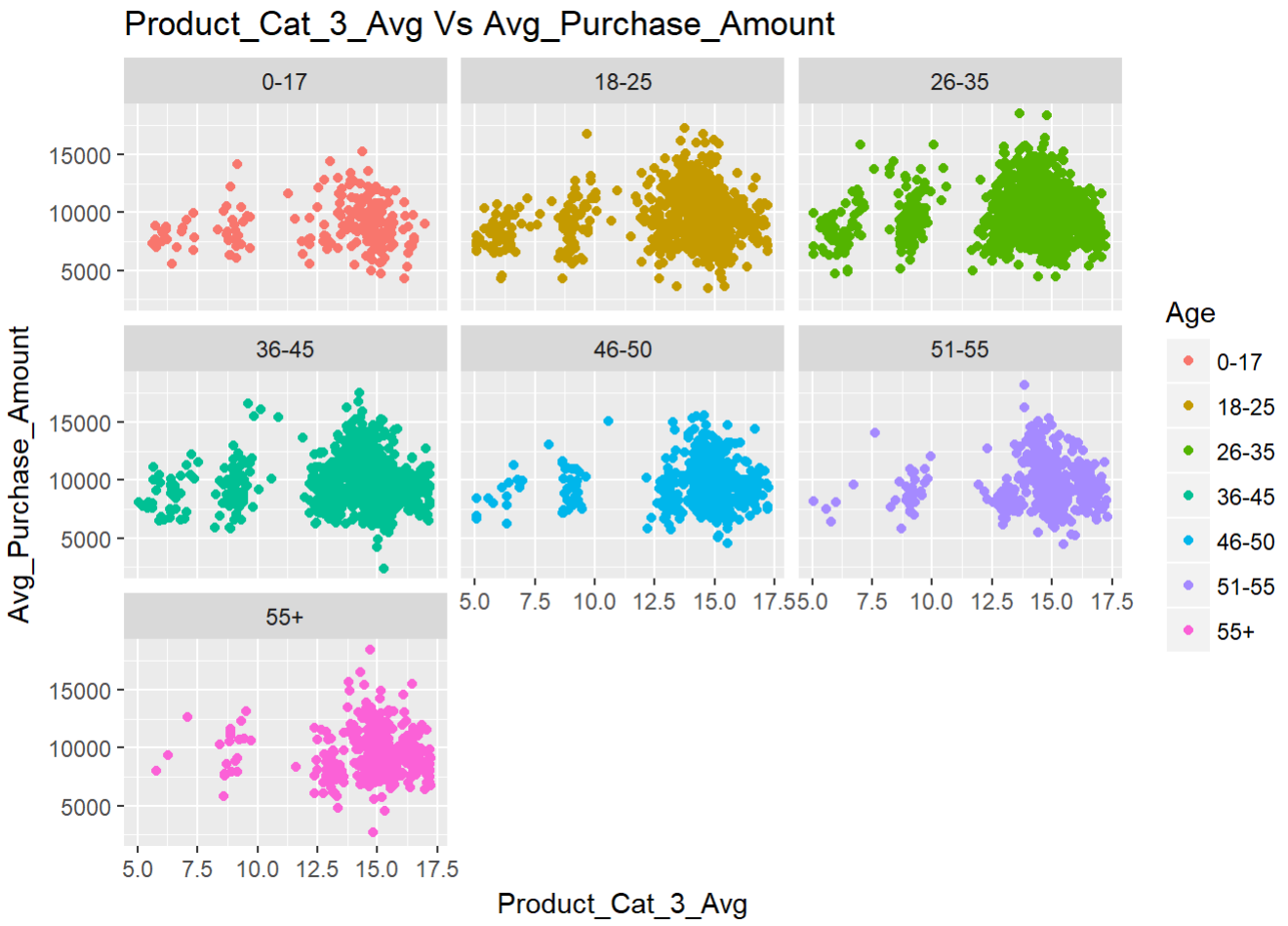



```
# 2. Which product category raked in the most money wrapped with age?
# Product Category 1:
ggplot(new_data, aes(Product_Cat_1_Avg, Avg_Purchase_Amount, color = Age)) + geom_point() + face
t_wrap(~ Age) +
  labs(title = "Product_Cat_1_Avg Vs Avg_Purchase_Amount")
```

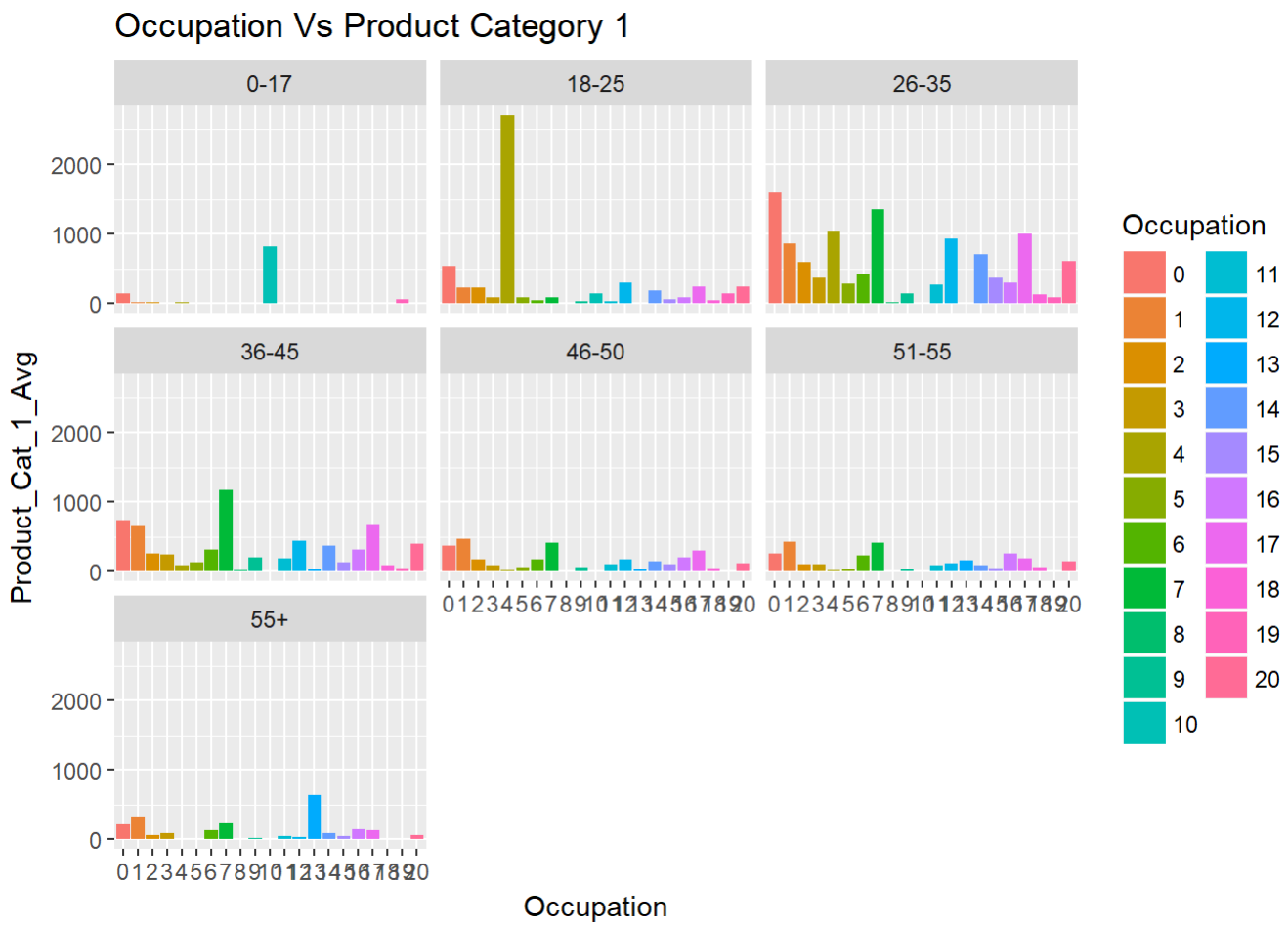



```
# Product Category 2:
ggplot(new_data, aes(Product_Cat_2_Avg, Avg_Purchase_Amount, color = Age)) + geom_point() + face
t_wrap(~ Age) +
  labs(title = "Product_Cat_2_Avg Vs Avg_Purchase_Amount")
```

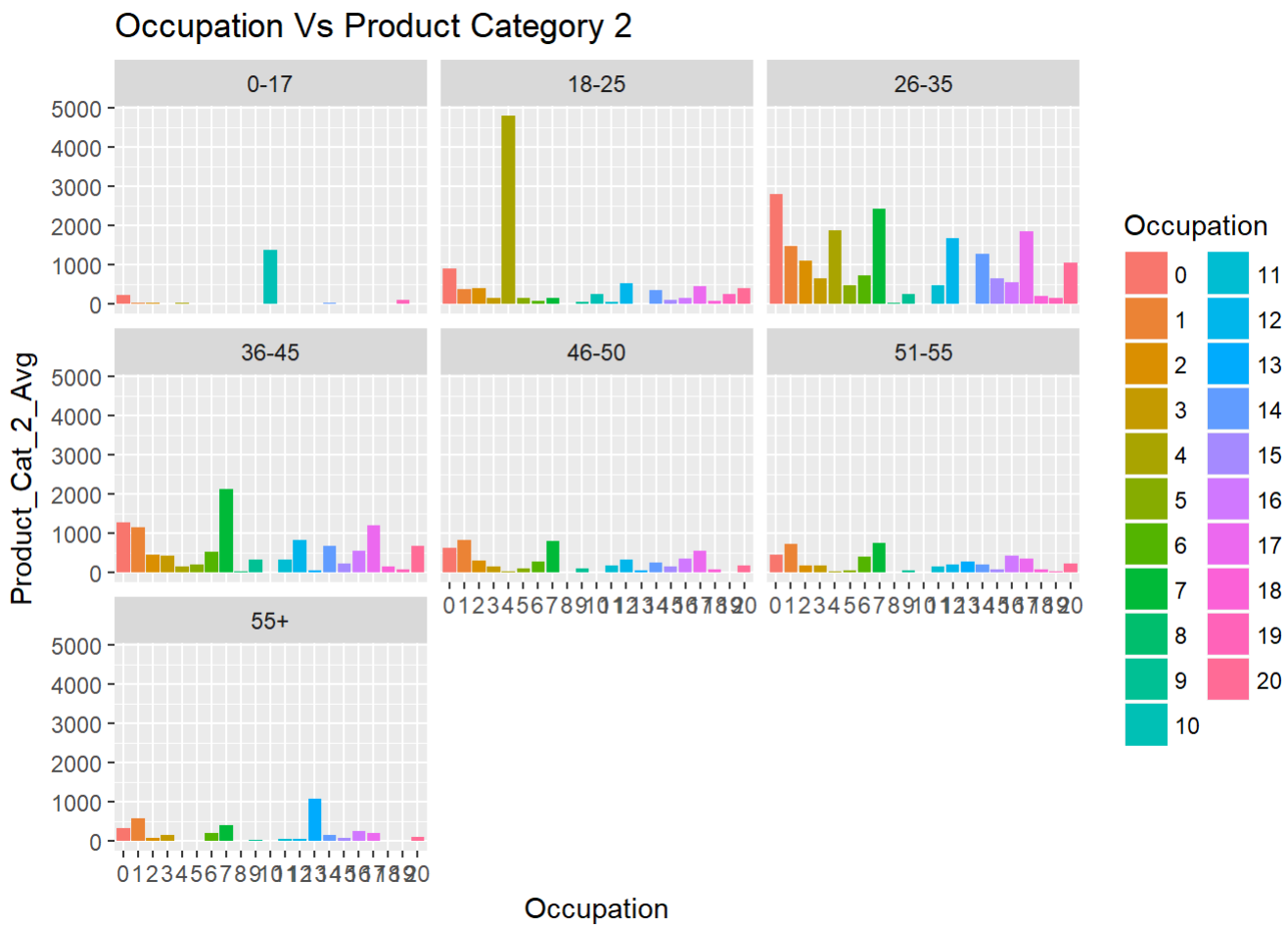

```
# Product Category 3:
ggplot(new_data, aes(Product_Cat_3_Avg, Avg_Purchase_Amount, color = Age)) + geom_point() + face
t_wrap(~ Age) +
  labs(title = "Product_Cat_3_Avg Vs Avg_Purchase_Amount")
```



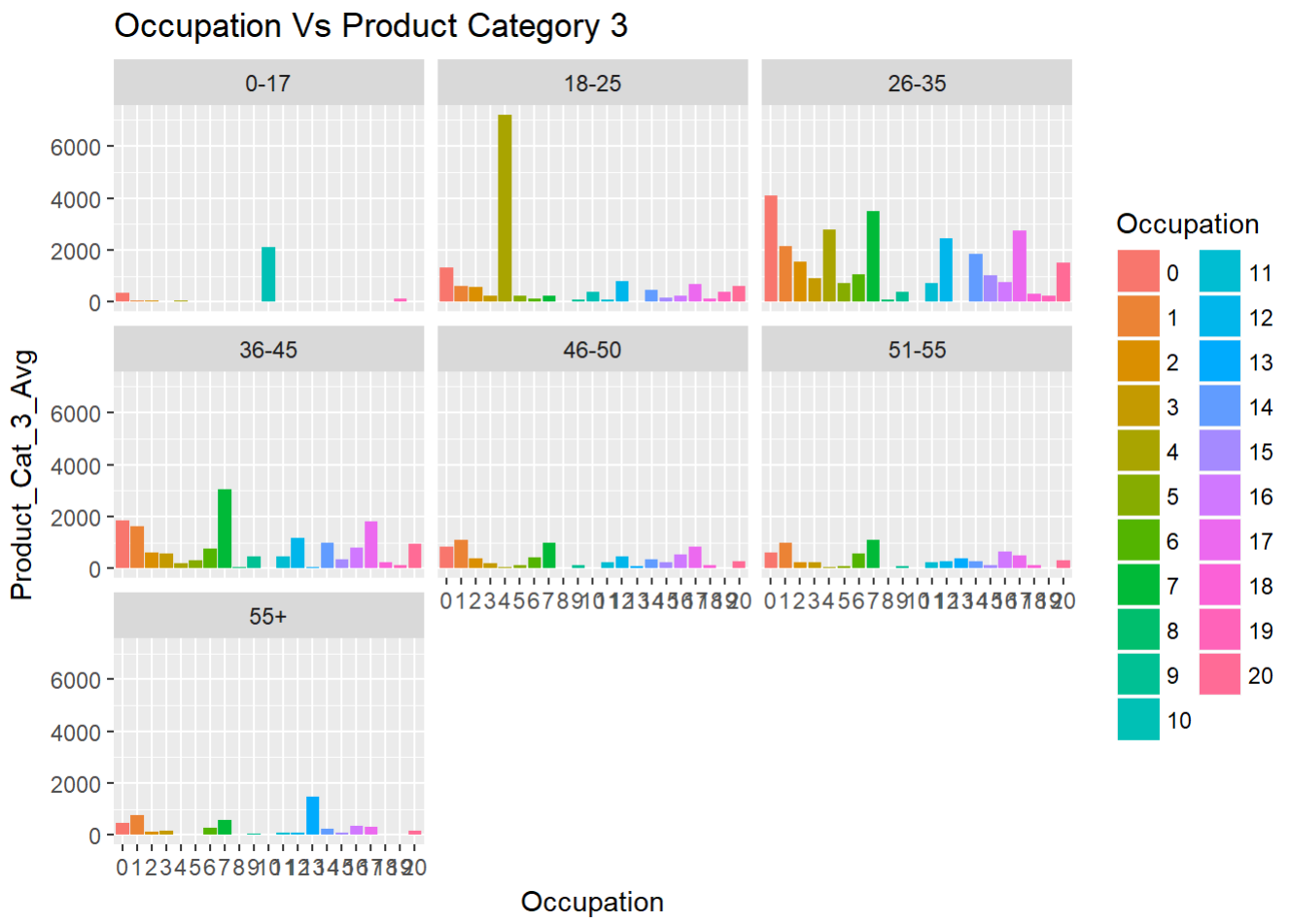

```
# Explore the occupation variable versus the product categories and the purchase amount

# 1. Which occupation had more influence on product purchase?
# Product Category 1:
ggplot(new_data, aes(Occupation, Product_Cat_1_Avg, fill = Occupation)) + geom_col() +  facet_wr
ap(~ Age) +
  labs(title = "Occupation Vs Product Category 1")
```



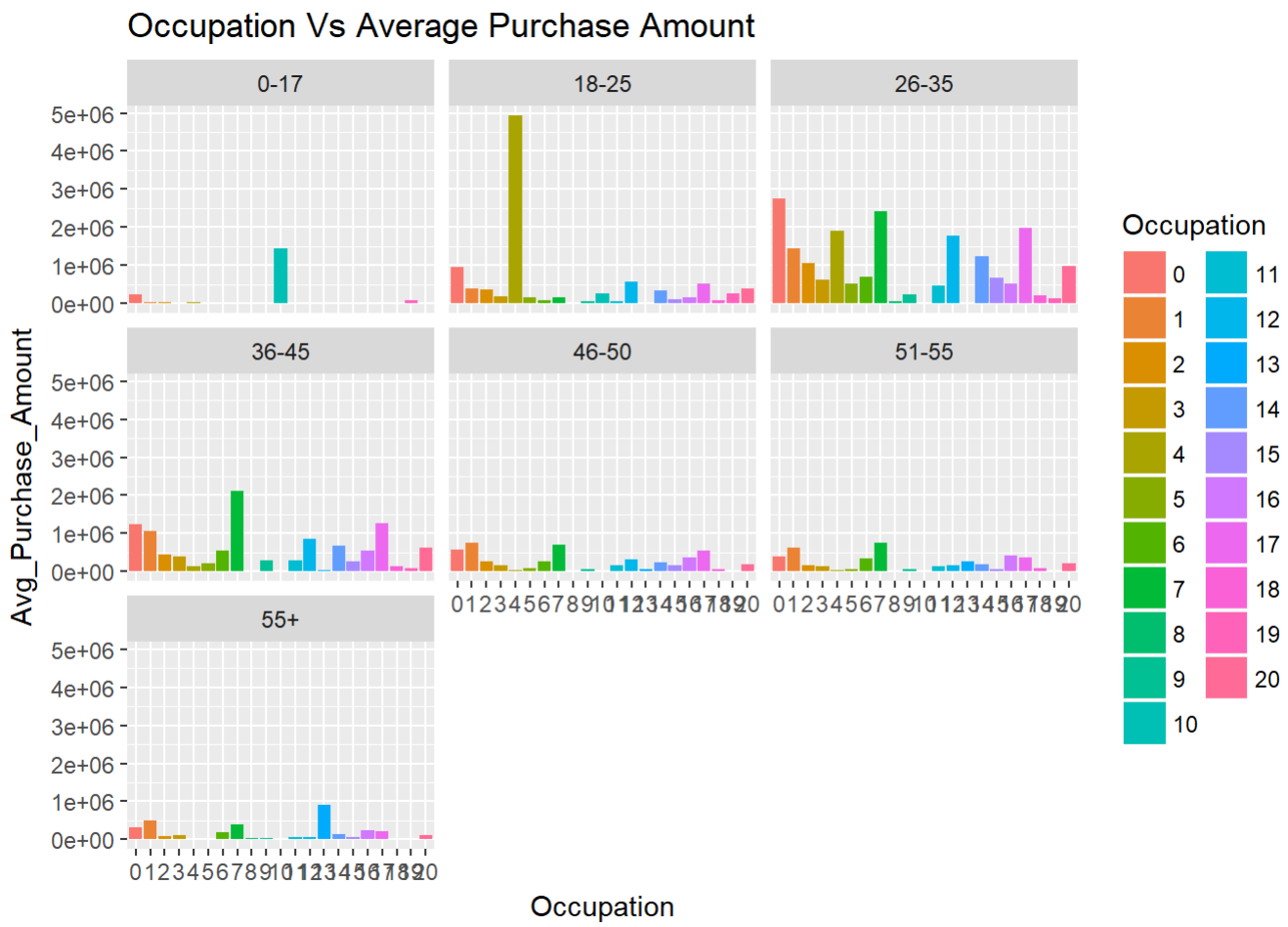

```
# Product Category 2:
ggplot(new_data, aes(Occupation, Product_Cat_2_Avg, fill = Occupation)) + geom_col() +  facet_wr
ap(~ Age) +
  labs(title = "Occupation Vs Product Category 2")
```



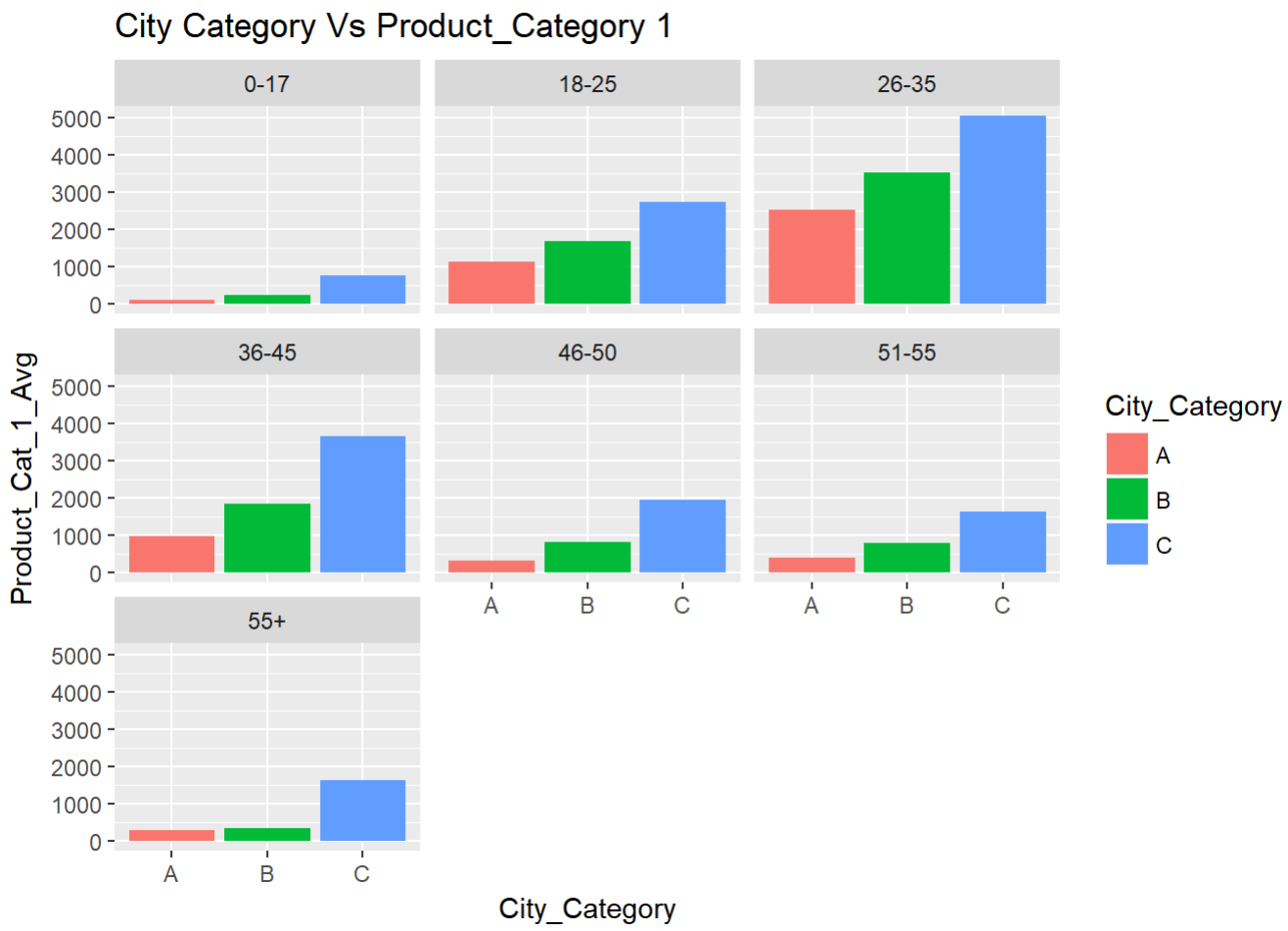

```
# Product Category 3:
ggplot(new_data, aes(Occupation, Product_Cat_3_Avg, fill = Occupation)) + geom_col() +  facet_wr
ap(~ Age) +
  labs(title = "Occupation Vs Product Category 3")
```



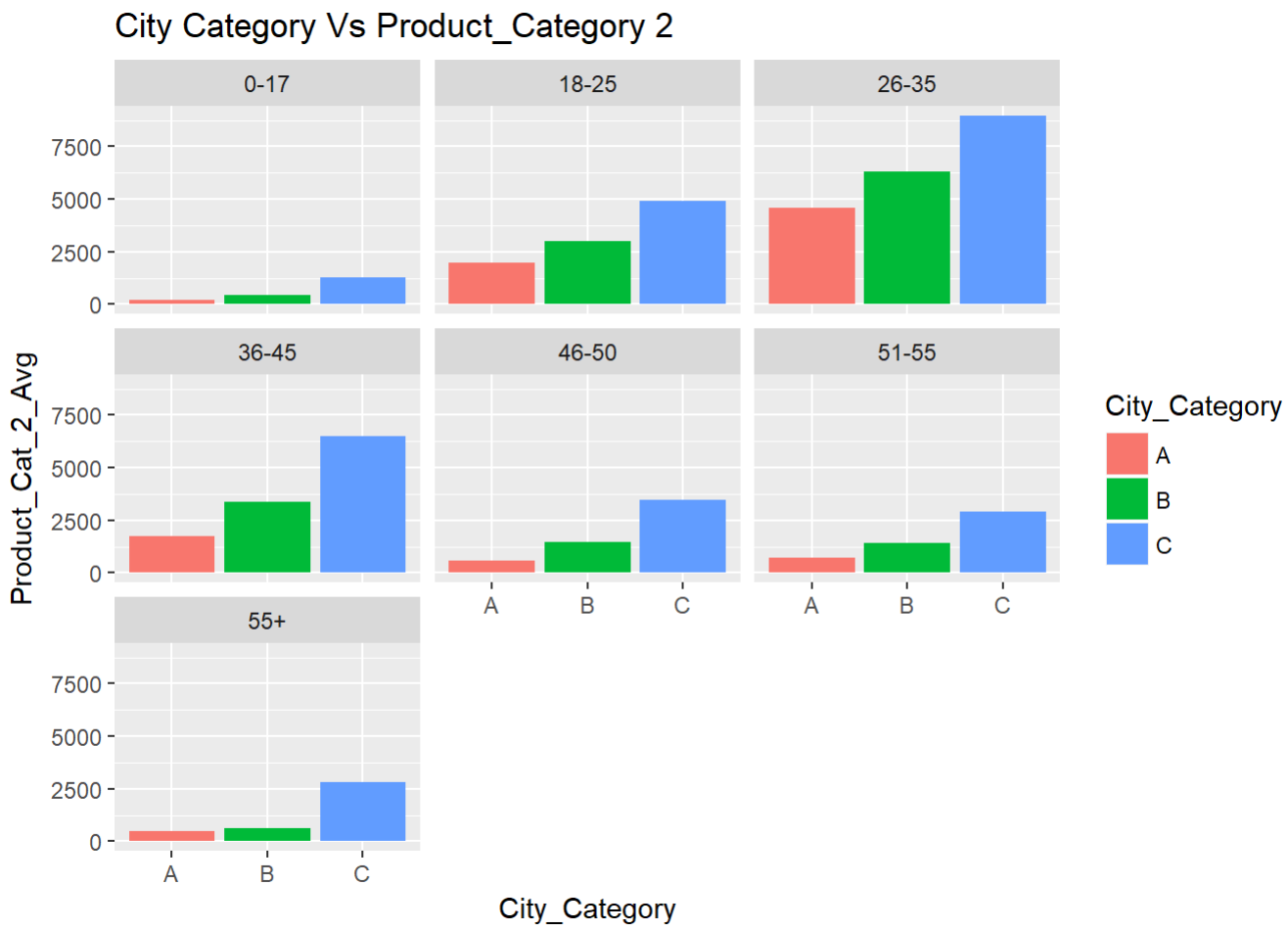

```
# Which occupation spent the most money?
ggplot(new_data, aes(Occupation, Avg_Purchase_Amount, fill = Occupation)) + geom_col() +  facet_
wrap(~ Age) +
  labs(title = "Occupation Vs Average Purchase Amount")
```



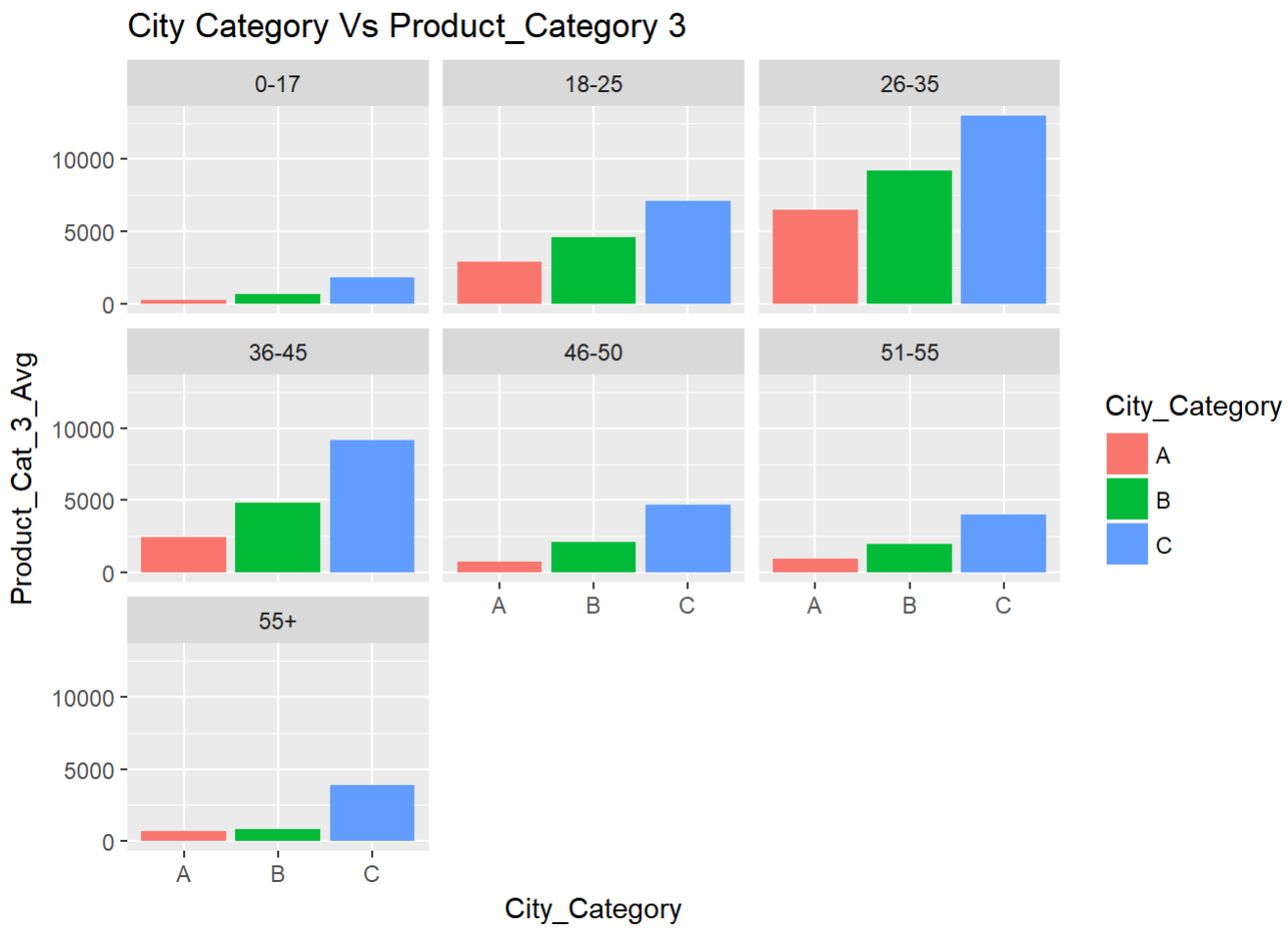

```
# Explore the city category variable versus the product categories and the purchase amount

# Product category 1:
ggplot(new_data, aes(City_Category, Product_Cat_1_Avg, fill = City_Category)) + geom_col() +  fa
cet_wrap(~ Age) +
  labs(title = "City Category Vs Product_Category 1")
```



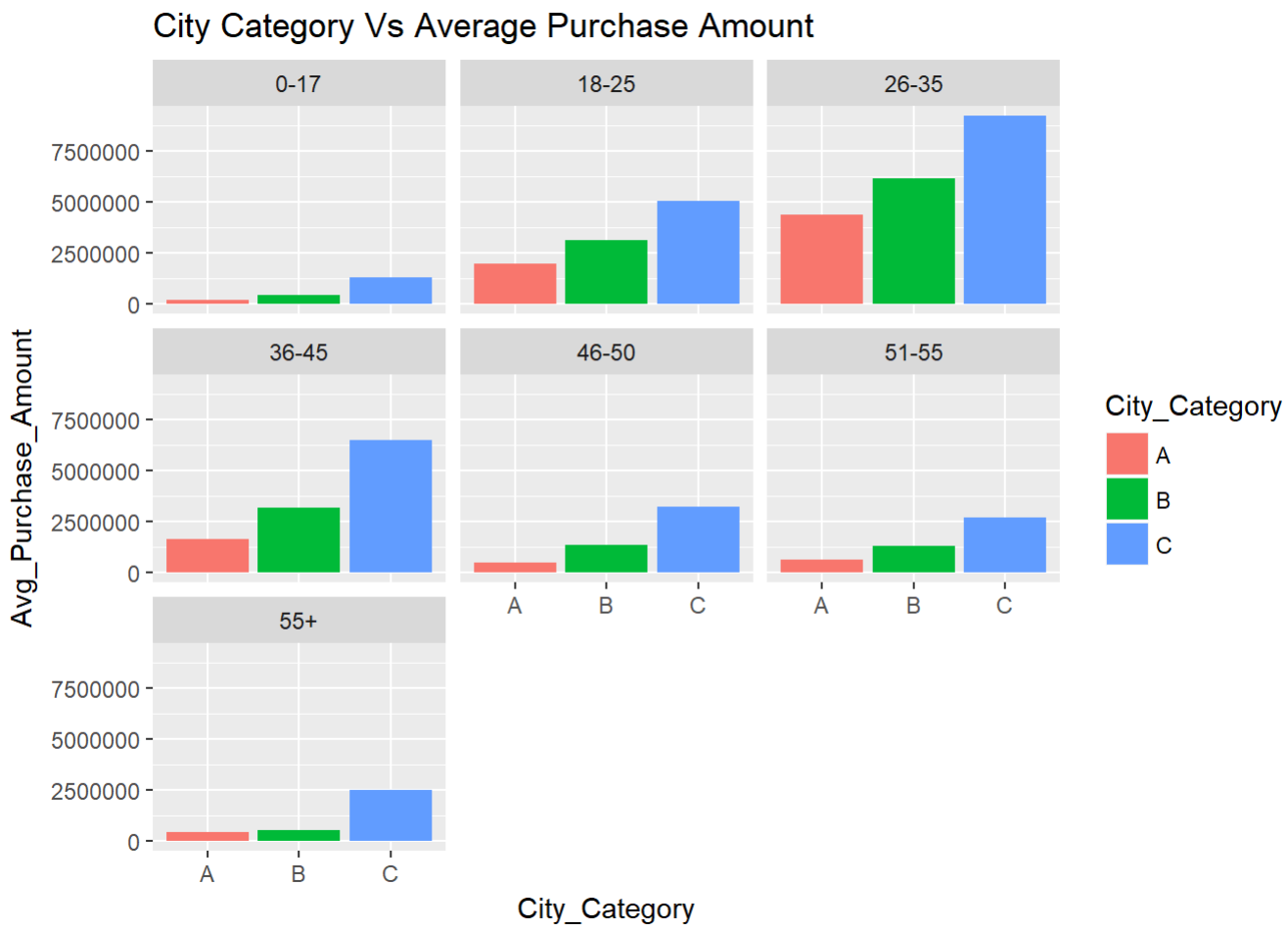

```
# Product Category 2:
ggplot(new_data, aes(City_Category, Product_Cat_2_Avg, fill = City_Category)) + geom_col() +  fa
cet_wrap(~ Age) +
  labs(title = "City Category Vs Product_Category 2")
```



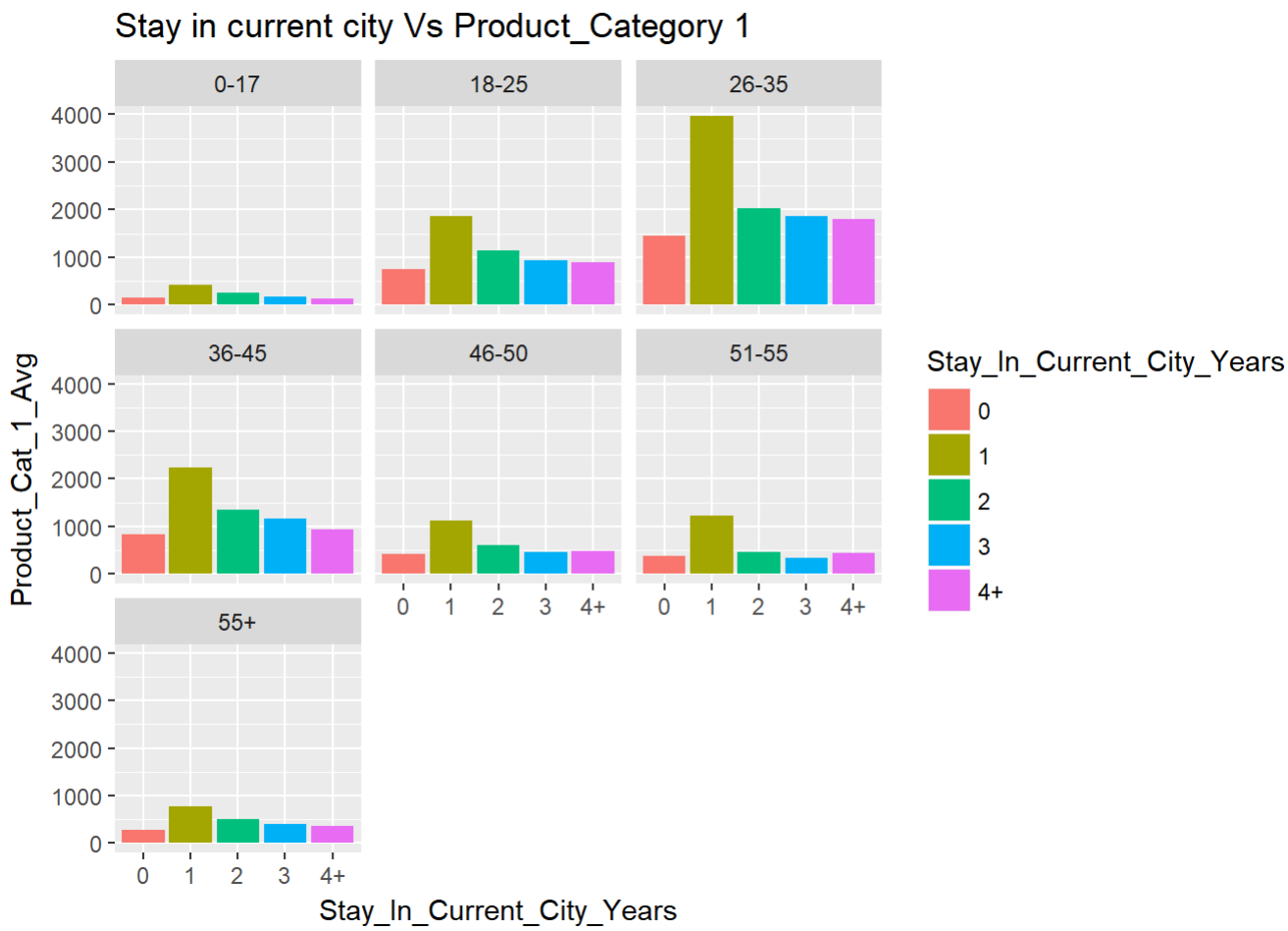

```
# Product Category 3:
ggplot(new_data, aes(City_Category, Product_Cat_3_Avg, fill = City_Category)) + geom_col() +  fa
cet_wrap(~ Age) +
  labs(title = "City Category Vs Product_Category 3")
```



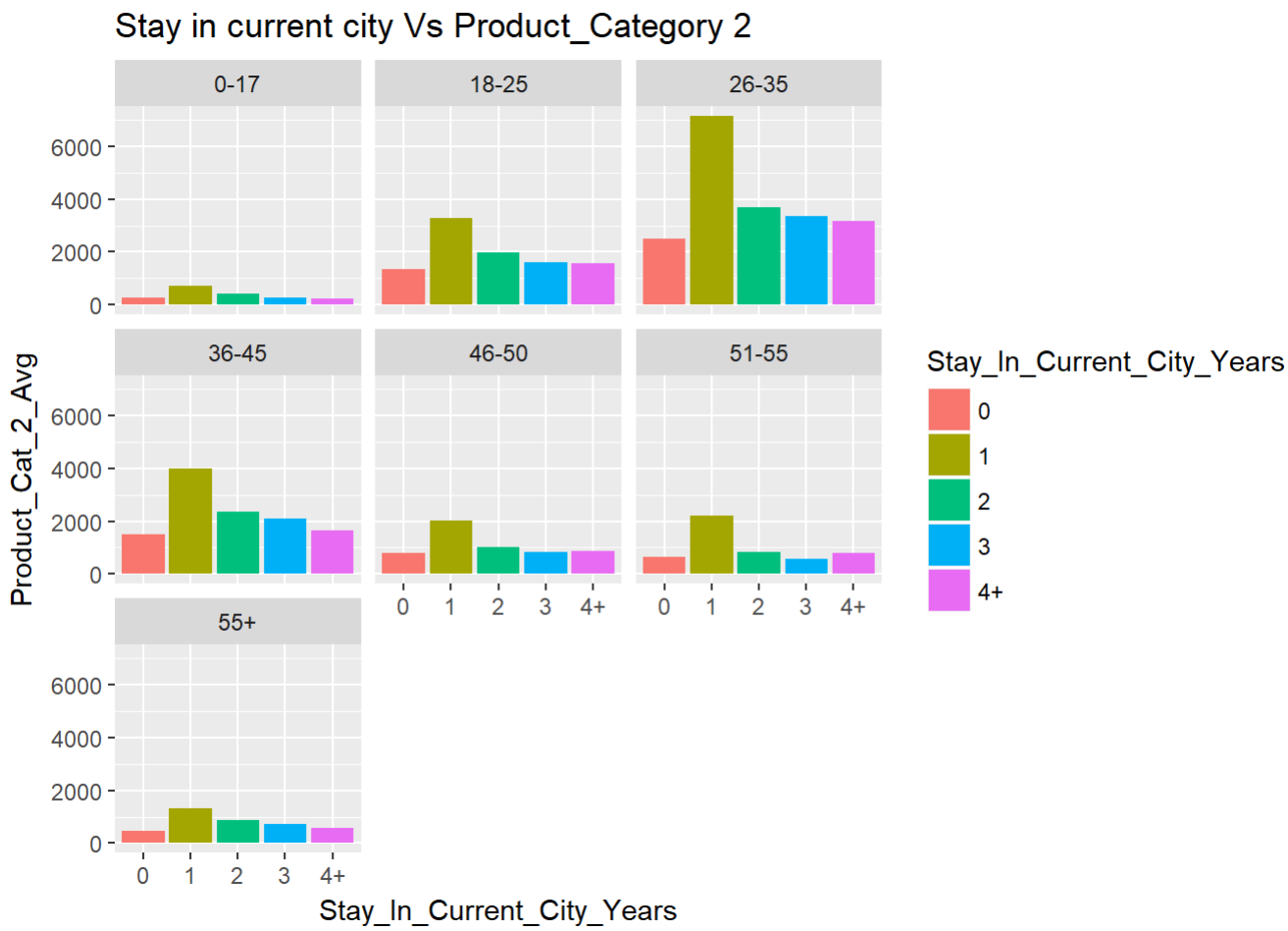

```
# City Category versus Average purchase amount
ggplot(new_data, aes(City_Category, Avg_Purchase_Amount, fill = City_Category)) + geom_col() +
 facet_wrap(~ Age) +
  labs(title = "City Category Vs Average Purchase Amount")
```



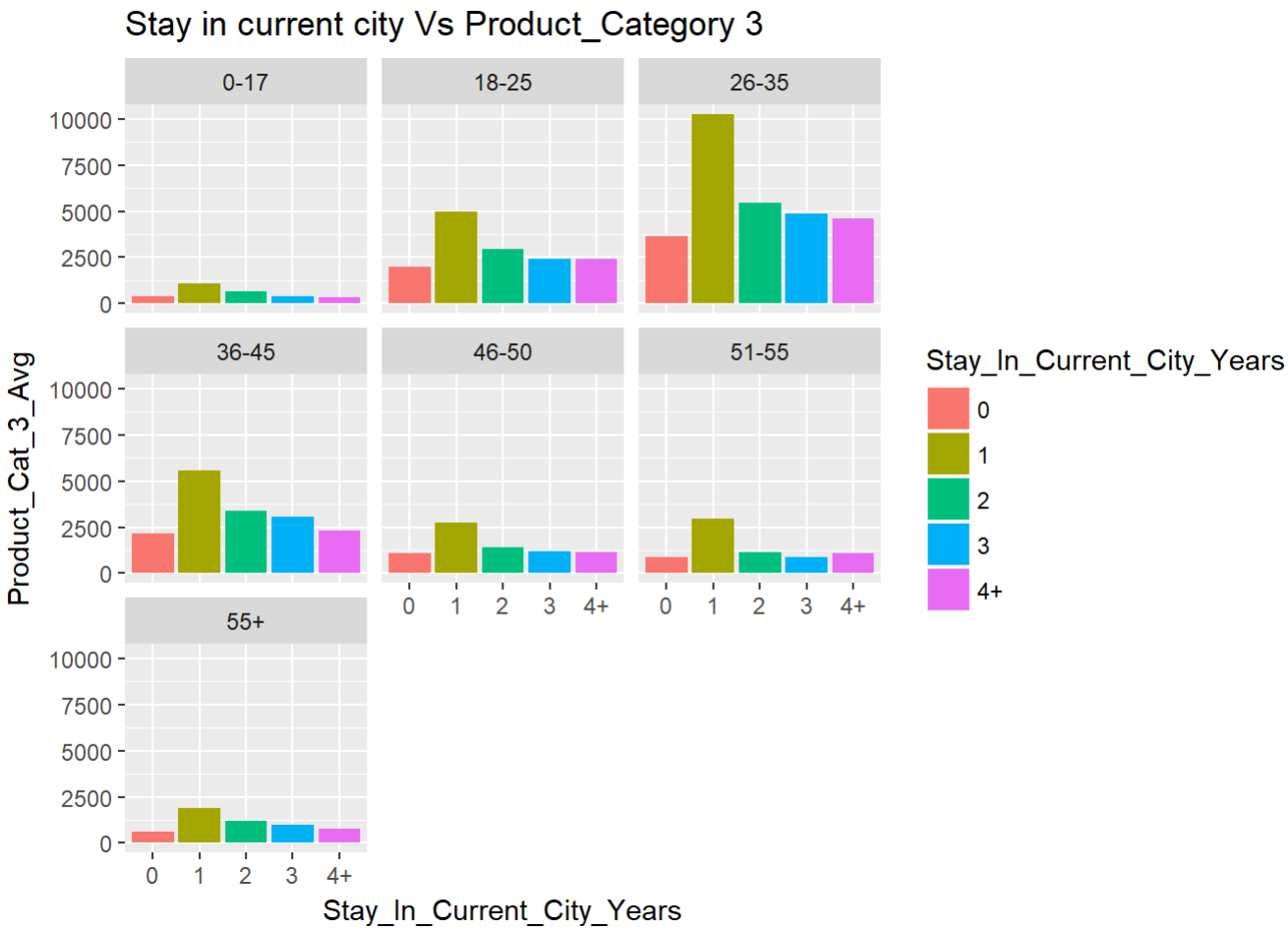

```
# Explore the Stay in current city variable versus the product categories and the purchase amoun
t

# Product Category 1:
ggplot(new_data, aes(Stay_In_Current_City_Years, Product_Cat_1_Avg, fill = Stay_In_Current_City_
Years)) + geom_col() +
  facet_wrap(~ Age) + labs(title = "Stay in current city Vs Product_Category 1")
```



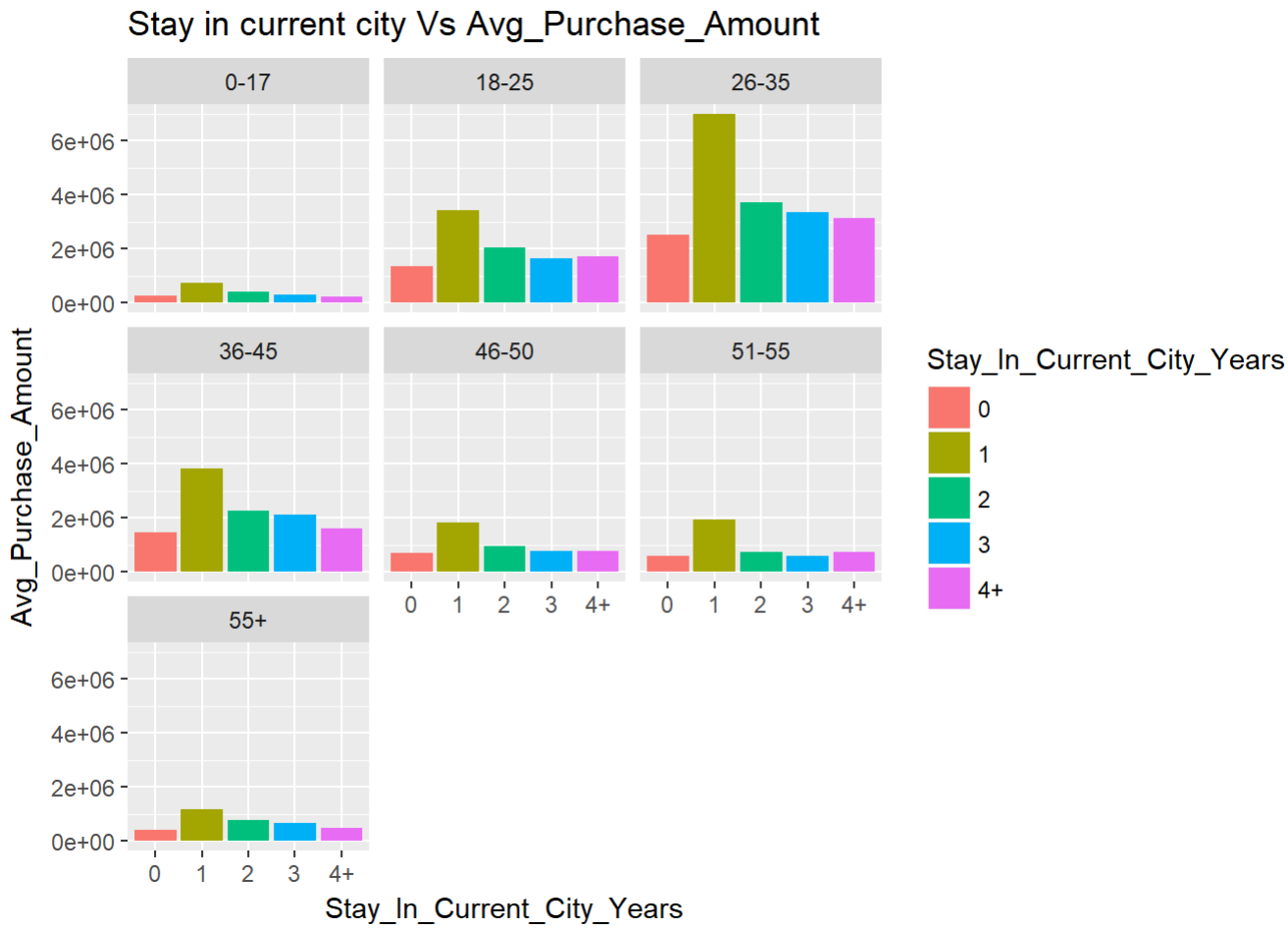

```
# Product category 2:
ggplot(new_data, aes(Stay_In_Current_City_Years, Product_Cat_2_Avg, fill = Stay_In_Current_City_
Years)) + geom_col() +
  facet_wrap(~ Age) + labs(title = "Stay in current city Vs Product_Category 2")
```



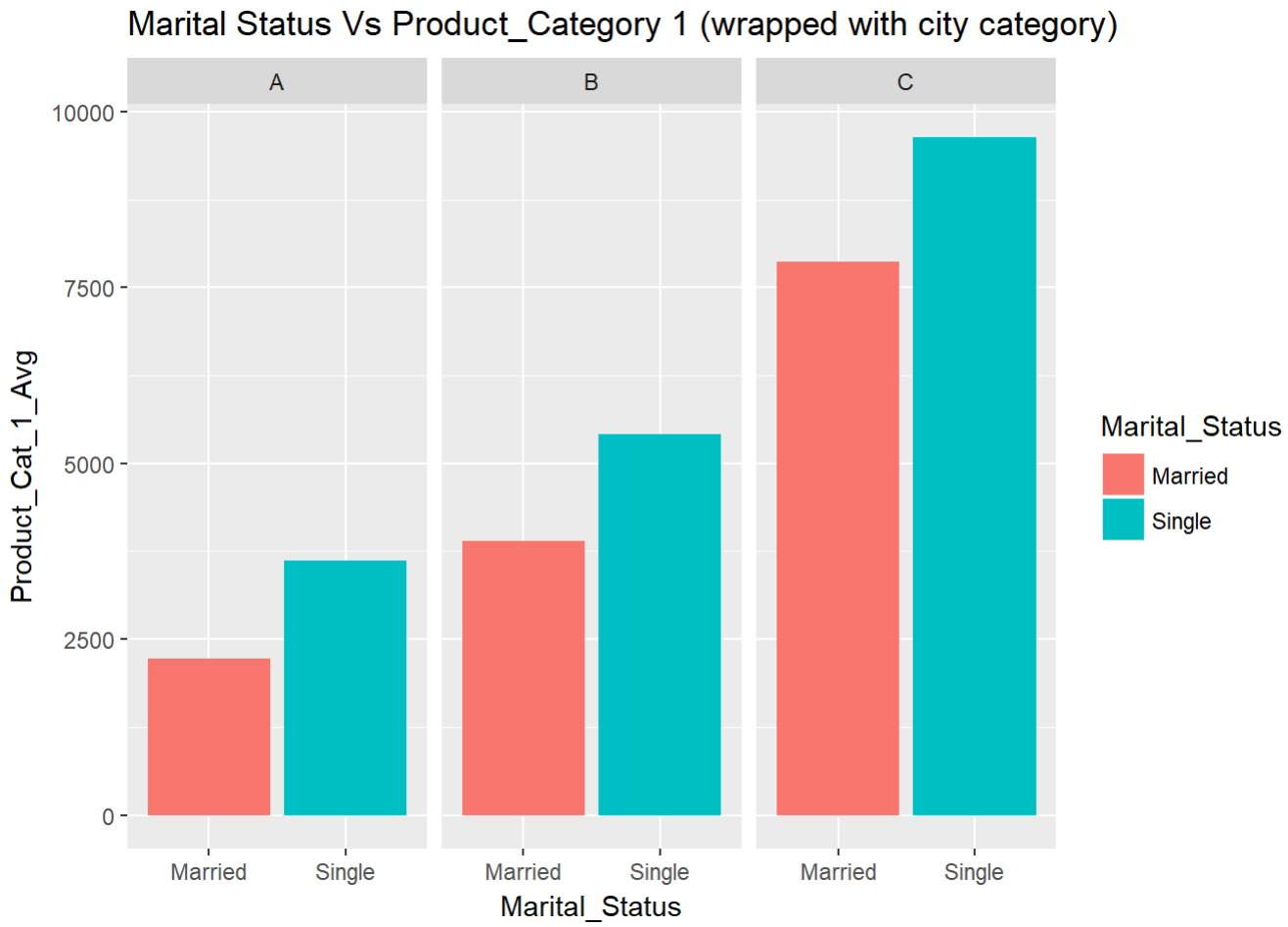

```
# Product category 3:
ggplot(new_data, aes(Stay_In_Current_City_Years, Product_Cat_3_Avg, fill = Stay_In_Current_City_
Years)) + geom_col() +
  facet_wrap(~ Age) + labs(title = "Stay in current city Vs Product_Category 3")
```



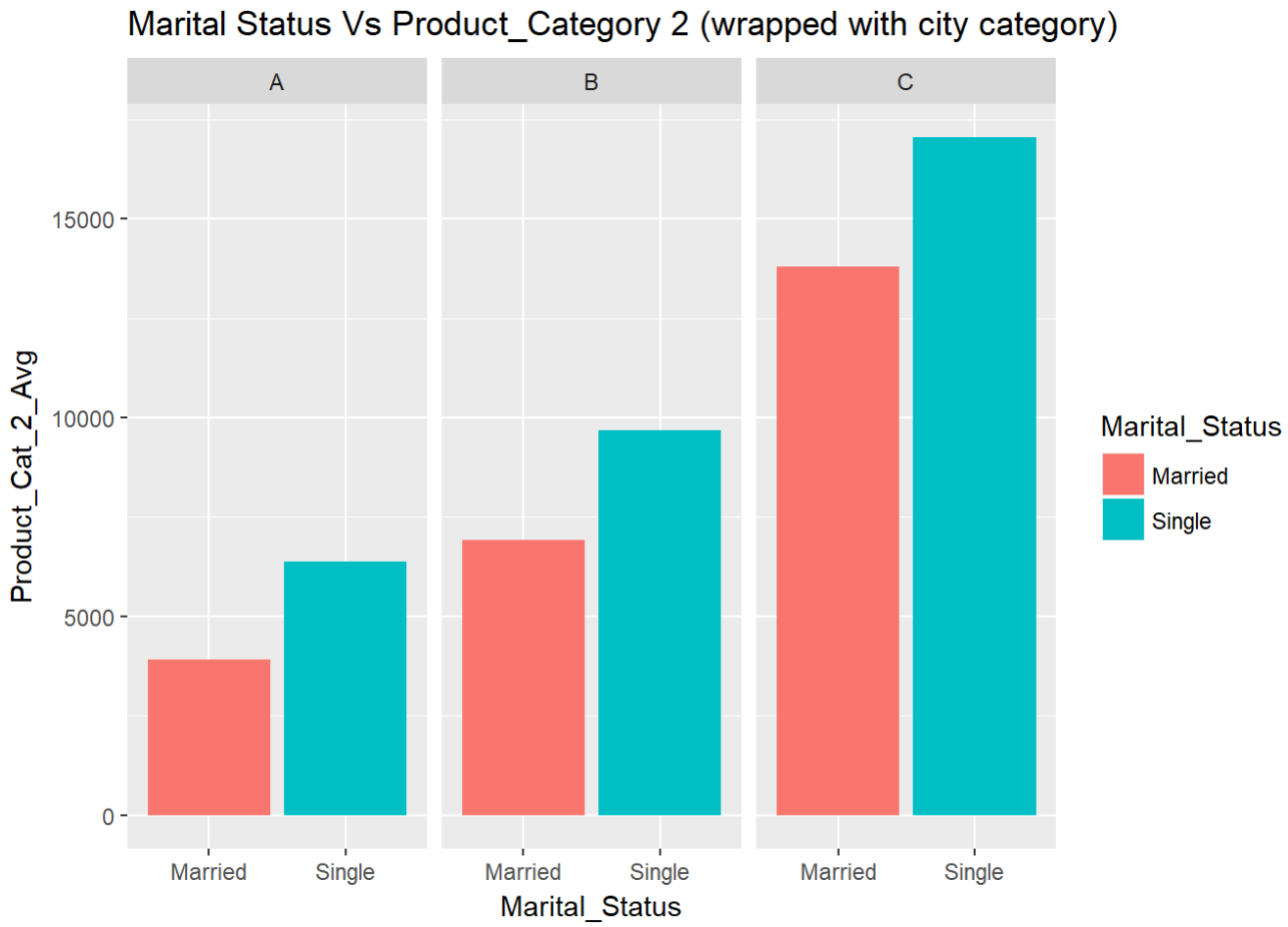

```
# Stay in current city versus Average purchase amount
ggplot(new_data, aes(Stay_In_Current_City_Years, Avg_Purchase_Amount, fill = Stay_In_Current_Cit
y_Years)) + geom_col() +
  facet_wrap(~ Age) + labs(title = "Stay in current city Vs Avg_Purchase_Amount")
```



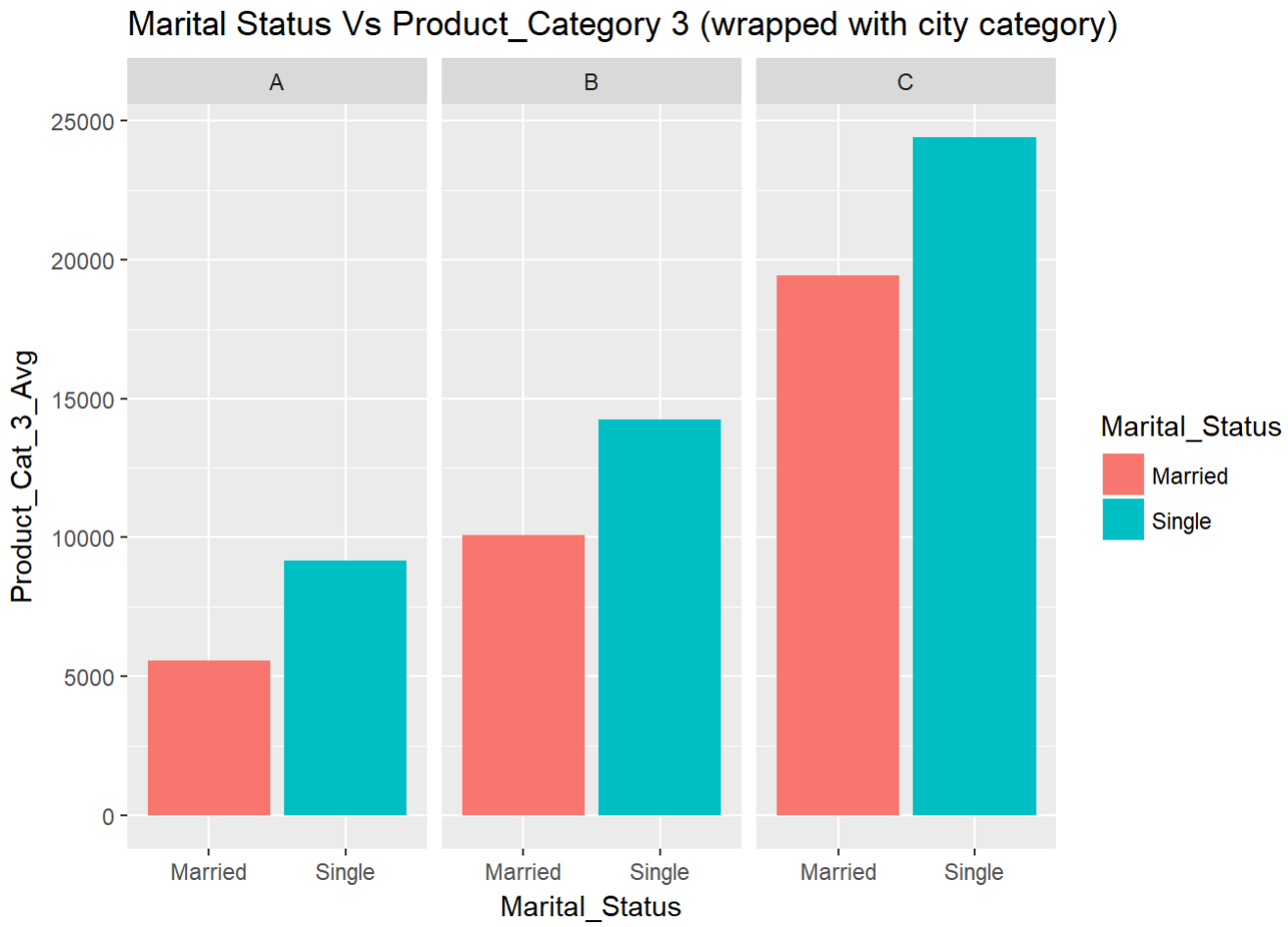

```
# Explore the marital status variable.

# Product Category 1:
ggplot(new_data, aes(Marital_Status, Product_Cat_1_Avg, fill = Marital_Status)) + geom_col() +
  facet_wrap(~ City_Category) + labs(title = "Marital Status Vs Product_Category 1 (wrapped with
 city category)")
```



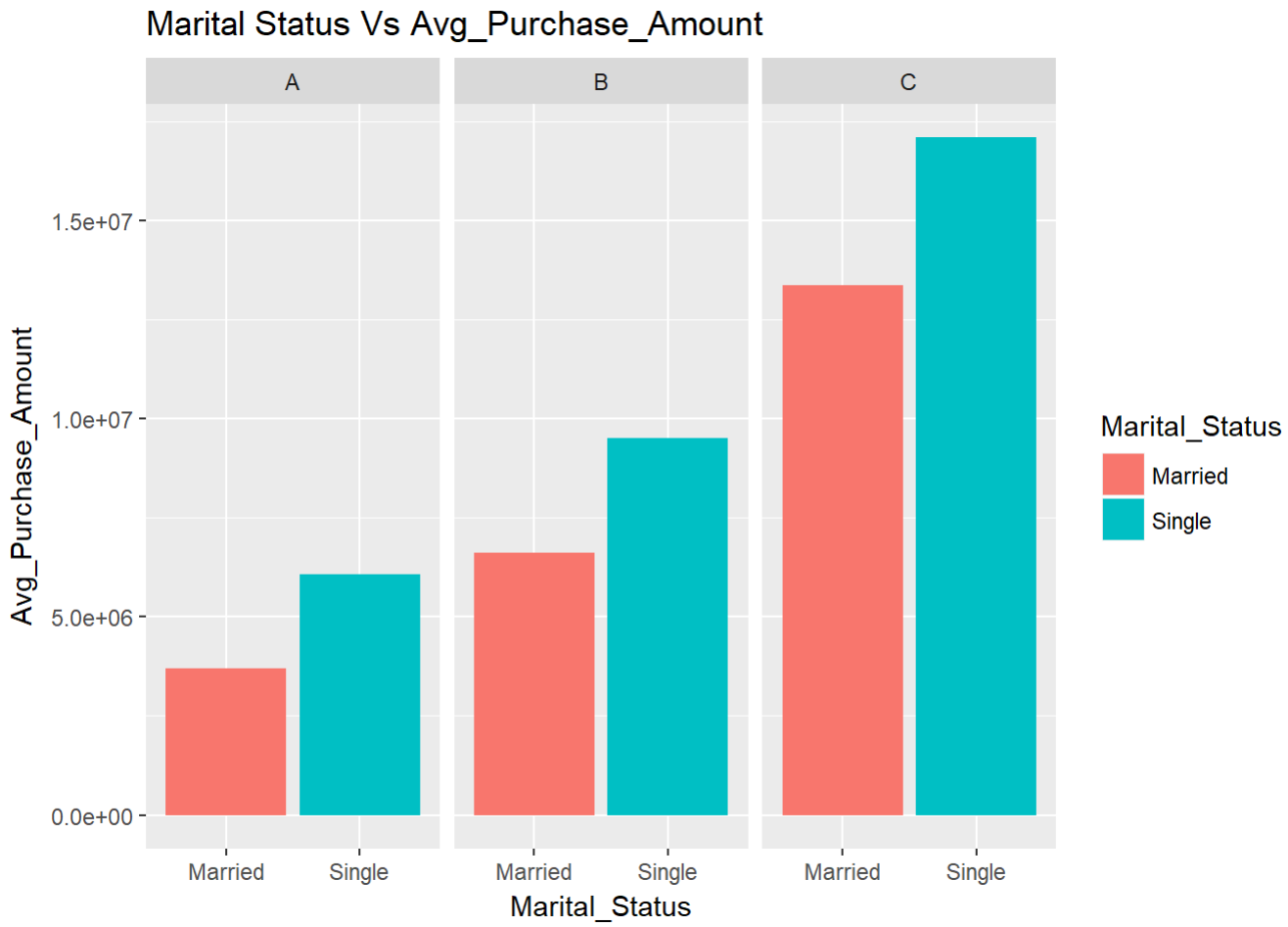

```
# Product Category 2:
ggplot(new_data, aes(Marital_Status, Product_Cat_2_Avg, fill = Marital_Status)) + geom_col() +
  facet_wrap(~ City_Category) + labs(title = "Marital Status Vs Product_Category 2 (wrapped with
 city category)")
```




```
# Product Category 3:
ggplot(new_data, aes(Marital_Status, Product_Cat_3_Avg, fill = Marital_Status)) + geom_col() +
  facet_wrap(~ City_Category) + labs(title = "Marital Status Vs Product_Category 3 (wrapped with
 city category)")
```




```
# Marital status versus Average purchase amount
ggplot(new_data, aes(Marital_Status, Avg_Purchase_Amount, fill = Marital_Status)) + geom_col() +

  facet_wrap(~ City_Category) + labs(title = "Marital Status Vs Avg_Purchase_Amount")
```




```
# CONCLUSION

# 1. THERE WERE MORE MALES (75%) CUSTOMERS THAN FEMALES (25%)

# 2. THE 3 MAJOR CUSTOMERS CLASSIFIED UNDER THE AGE GROUP VARIABLE ARE
      # A. 26-35 YEARS OLD --- 40%
      # B. 36-45 YEARS OLD --- 20%
      # C. 18-25 YEARS OLD --- 18%
    # THE LEAST WAS THE 0 -17 YEARS OLD WITH JUST 3%, FOLLOWING THAT CLOSELY WAS THE 55+ WITH
 4%,
    # THE 46-50 AND 51-55 WERE CLOSELY MATCHED WITH 8% AND 7% RESPECTIVELY

# 3. CITY CATEGORY B HAD THE MOST CUSTOMERS(42%), FOLLOWED BY C (31%), AND LASTLY A (27%)

# 4. pEOPLE WHO HAD STAYED IN THE CITY FOR 1 YEAR CONSTITUTED MAJORITY OF THE CUSTOMERS

# 5. sINGLES MADE UP 59 % WHILE MARRIED MADE UP 41% OF THE CUSTOMERS

# 6. MORE MALES THAN FEMALES PURCHASED ALL 3 PRODUCT CATEGORIES, ALTHOUGH FOR THE 0-17 YEARS OL
D, IT WAS CLOSELY MATCHED

# 7. AS EXPECTED 26-35 YEAR OLDS SPENT MORE MONEY

# 8. IN ALL THE CITY CATEGORIES, THE SINGLES SPENT THE MOST MONEY

# 9. FOR THE PRODUCT CATEGORIES WITH THE HIGHEST PURCHASE, THE PLOTS PRESENT THEM IN DETAILS
```